UNIVERSITY OF CALIFORNIA,
IRVINE

AN EPIDEMIOLOGIC APPROACH
FOR USING SOCIAL MEDIA IN PUBLIC HEALTH INTERVENTIONS

THESIS

Submitted in partial satisfaction of the requirements for the degree of

MASTER OF SCIENCE

In Epidemiology

By

John Paul Schomberg

Thesis Committee:
Professor Hoda Anton Culver Chair
Professor Gillian Hayes
Professor Ralph Delfino

2015

TABLE OF CONTENTS

List of Figures

Page

List of Tables

Acknowledgements

I must convey my gratitude and great respect for my committee chair, Professor Hoda Anton-Culver, who had the wisdom to know when my imagination should be reined in and when it should be set free. I shall forever be indebted for the unwavering support and encouragement provided over the course of my studies at UCI. I shall always emulate the spirit and energy that Dr. Anton-Culver channels into her everyday work as a teacher, researcher, and leader.

I would like to thank my committee members, Professor Ralph Delfino and Professor Gillian Hayes, whose contributions during my thesis defense allowed me to identify when more is less and when less is more.

Thank you to Professor Marc Strassburg of University of California Los Angeles, who first introduced me to the science and art of Epidemiology, to Professor Fred Hagigi who first started me on the path of graduate study and to Professor Becky Yano for telling me the path has no end.

A special thanks to all of the hard working San Franciscans and New Yorkers who love to write and share their dining experiences with their neighbors and the world, without you this study would not have been possible.

ABSTRACT OF THE THESIS

An Epidemiologic Approach for Using Social Media in Public Health Interventions

By
John Paul Schomberg

Master's of Science in Epidemiology

University of California Irvine, 2015

Professor Hoda Anton-Culver Chair

Participation in social media particularly in urban centers is growing rapidly. Understanding how information in social media can modify public health behaviors and how social media can be mined to make meaningful public health intervention shall be highly useful as social media use expands. The specific focus of this thesis is to describe how social media on Yelp.com can be mined to gain meaningful public health surveillance that is predictive of real world health code violation. The Second aim of this thesis is to survey an urban area with a high concentration of Yelp users to identify how Yelp use and value of information on Yelp can modify the health behavior of restaurant selection and modify odds of food borne illness.

In our analysis of the predictive power of social media data mined from Yelp.com we found that keywords like "vomit" and "DIRTY" were predictive of substandard health code rating (<80) with Odds Ratio of (45.4), and (3.68) respectively. The logistic regression model used had Sensitivity, Specificity, Positive Predictive Value, and Area under the Receiver Operator Curve of .72, .44, .61, and .78 respectively. Our Survey of an urban area with a high concentration of Yelp users found that those Yelp users that valued Yelp’s measurement of quality “Stars” the most had increased odds of reported food borne illness (1.01-2.54). We also found that despite Yelp.com’s “partnership” with public health officials in San Francisco and their agreement to present public health data on Yelp.com only 10% of respondents knew public health data was posted for restaurants on Yelp.com.

Our results show us that knowledge of health code violations like employee hand washing and presence of vermin decreased respondent desire to select restaurant more than knowledge of health code rating. This is important to note as yelp only presents health code ratings along the restaurants on its site. The findings of the analysis conducted in this thesis allow public health officials to improve the effectiveness of surveillance of restaurants for food borne illness risk factors, and improve partnerships with social media companies like yelp.com to better communicate public health findings and change public health behaviors.

Chapter 1:

Public Health Inspection Enhancement, Individual Misinformation Use and Misuse of Social Media in Public Health Intervention

Introduction:

Communicating the safety of restaurants, food trucks, coffee shops and bars to the public is a federal mandate that must be followed by public health departments across the nation. It is important that this information be delivered with clarity and accuracy in a way that effectively modifies public behavior. It is important that communication alters the behavior of proprietors/restaurant employees preparing food and the larger public making the decision to consume food. Currently restaurants receive inspections once to twice a year over a very brief period of time 20-30 minutes. The restaurant industry is growing quickly at an average annual rate of 4.9% per year. At the same time, public sector has faced increasing cutbacks in recent years due to budgetary restrictions. This means that public health inspectors are faced with an increasing workload with the same or reduced staff. Nationally there are reports of increases in the incidence of food borne illnesses reported at the emergency department from 2012 to 2015 of Shigella and E.coli organisms and endotoxin. There have been recent studies that attempt to use existing public data to create predictive models that help identify restaurants that may create a public health risk. However, these models rely on existing data and do not incorporate new data sources or provide increased surveillance coverage. As the restaurant industry grows thoughtful solutions must be considered to increase surveillance and acquire new information sources that allow inspectors to better identify those restaurants that present the greatest risk to public health, and to move quickly to prevent epidemics in the cases where highly predictive risk factors are reported.

Once information has been acquired from surveillance and analyzed it must be presented to the public to change the behavior of restaurant employees and to change the public health behavior of restaurant selection. Newer methods of communicating public health risk to the public such as letter grades posted outside restaurants have been associated with decreases in the rates of emergency room admissions for food borne illness in those areas. If the letter grades are causal in the reduction of food borne illness cases then it is unclear if the effect is due to the effect on customer behavior or the effect on vendor behavior. Health departments of many municipalities have elected to make public health inspection data on department of public health websites. Unfortunately this information is not readily accessed by the public. Private entities that publicize restaurant inspection scores online have also been associated with improvement in restaurant inspection scores. This reflects the effect of such websites on improving vendor behavior. It is unknown what effect this has on the public behavior of restaurant selection. The single study on publication of public health information on a private website was conducted on a small website that only posted data on restaurants in Salt Lake City Utah. There is no study that examines the effect of public health data displayed on social media sites designed for restaurant review like Yelp.com or UrbanSpoon.com. It is unknown how public health information is interpreted when placed alongside ratings and other data aggregated across many reviewers.

Recent research is beginning to show potential uses of social media for the use of public health surveillance. Google, Twitter, and Craigslist, have all been shown to be useful in tracking flu, public health sentiment, and prevalence of practices that increase risk for STD transmission. When used creatively social media can be put to meaningful public health use. The specific focus for this thesis is to provide evidence that Social media like yelp.com can be put to the meaningful use of enhancing the system of screening restaurants for practices that increase risk of food borne illness, and a validation of yelp.com in its use as a tool for distributing health inspector's findings.

Screening conducted by public health agencies may include inspecting rental homes or units for lead, pests, or unlivable conditions, restaurants for unsanitary food preparation practices, and service establishments such as salon, barbershops, and masseuse for unsafe practices. Such inspections may also be conducted at clinics and hospitals as well. As the ubiquity of technology that can capture and broadcast unique experience and observations increases so does our ability to use such information for public good. As the number of social media sites increase, the use of such data sources becomes limited only by the imagination and skill of the scientist making use of it. In a large urban metropolis the number of people taking part in social media observations is large, and easily exceeds the survey capabilities of any public health department. I propose that public health departments harness the information hidden within social media to improve public health screening and generation of more rapid alerts to emerging crises that may affect public health. Harnessing the engine of data science will strengthen and enrich the efforts of our often underpowered and underfunded public health departments.

The appeal of social media is that it can provide a large sample of the populous in a short period of time at exceedingly low cost. The caveat is that the design of data collection methods and analysis must be thoughtfully constructed to ensure that any inference drawn from such data is likely to be valid. Public health officials can make use of the wide surveillance coverage provided by social media on Yelp.com to improve their current surveillance. Municipalities where use of social media is highest like San Francisco would serve as an ideal place to mine social media like Yelp.com for public health surveillance. Predictive models created from Yelp data (tags, and reviews) can be validated against the observed inspection data provided by San Francisco Department of public health. Such a system could be used to offer additional information to inspectors to rank restaurants according to their public health risk, and to serve as an alert to inspectors when specific keywords are reported within a Yelp review.

Currently the ability of public health officials to alert the public to potential food borne illness threats is limited to what is reported by the public or identified upon annual or biannual inspection. This approach is severely limited by proportion of the public actually reporting food borne illness to public health officials and the coverage of surveillance provided by health inspectors. Public health alerts for food borne illness can be improved by identifying the most effective way to communicate public health findings through social media. The website Yelp has adopted a data format that allows public health officials to upload health inspection data to the Yelp website. However, there is currently no method to measure the effectiveness of the means used to display public health data on Yelp. The first city to adopt Yelp Local Inspector Value Entry System (LIVES) was San Francisco; by surveying individuals in San Francisco most likely to use Yelp we can identify the features of public health data that are most effective in determining an individual's decision to eat food at a given restaurant. This analysis will show what types of information display are most effective in influencing public health behavior. Results will allow public health officials to assess the effectiveness of the Yelp.com public health data display system, and it will allow yelp executives to reevaluate their current display of public health data.

In this paper I examine how social media may be used to enhance detection of restaurants in need of inspection, and then identify how the public's interpretation and value placed onYelp.com data may actually lead to an increase in food borne illness risk. When viewed through an analytical lens public health data can help officials find threats earlier, and possibly prevent epidemics from spreading. Unfortunately the methods that may be employed by public health officials are not readily available to the public. This may lead the public to be influenced by a single data point available on yelp.com such as number of stars or number of reviews. It is difficult to assess all yelp.com information on a given restaurant in a systematic way. A large and growing number of people use Yelp.com to guide their decision of where to eat at a given restaurant. Yelp's popularity and large user base has engendered a large degree of trust in the quality of restaurants that receive favorable ratings on Yelp.com. Unfortunately the taste of food and quality of dining experience are not always correlated with food that is safe for public consumption.

Public health screening of businesses and institutions providing personal services to the public requires large resources in terms of labor hours, staff, and expertise. Current coverage provided by public health screening is provided 1-2 times per year at most. The surveillance provided by such screening is inadequate, evidenced by lapses in compliance with public health code on an increasing annual basis. We also know that screening even when provided may not adequately catch all cases of noncompliance. It is clear that additional measures that can improve or supplement screening of businesses or institutions accessed by the public at a low cost are needed.

Social media is a vast resource that can be put to the meaningful purpose of enhancing public health screening and generation of public health alerts. Social media has become a term that is generally recognized to reference data on the internet generated and utilized by a large group of people.   The appeal of such a resource is that it can provide a large sample of the populous in a short period of time at exceedingly low cost. The caveat is that the design of data collection methods and analysis must be thoughtfully constructed to ensure that any inference drawn from such data is likely to be valid.

Currently the ability of public health officials to alert the public to potential food borne illness threats is limited to what is reported by the public or identified upon annual or biannual inspection. This approach is severely limited by proportion of the public actually reporting food borne illness to public health officials and the coverage of surveillance provided by health inspectors. Public health alerts for food borne illness can be improved by identifying the most effective way to communicate public health findings through social media. Currently the website Yelp has adopted a data format that allows public health officials to upload health inspection data to the Yelp website. However, there is currently no method to measure the effectiveness of the means used to display public health data on Yelp. The first city to adopt Yelp Local Inspector Value Entry System (LIVES) was San Francisco; by surveying individuals in San Francisco most likely to use Yelp we can identify the features of public health data that are most effective in determining an individual's decision to eat food at a given restaurant. This analysis will show what types of information display are most effective in influencing public health behavior. Results will allow public health officials to assess the effectiveness of Yelp public health data display system, and it will allow yelp executives to reevaluate their display of public health data.

Specific Aims:

1. To validate the use of keywords and tags found in yelp.com in predicting sub-standard health code rating/health code violation. Using the gold standard of health code rating/violation set by the San Francisco Department of Public Health.
2. To determine which keywords and tags are significantly predictive of substandard health code rating/health code violation.
3. To determine the effect of public health data displayed on Yelp.com on informing the public and influencing public health behavior of selection of restaurant with substandard health rating.
4. To examine the association of public health data and social media data value and food borne illness risk.
5. To measure the impact of each year of Yelp LIVES formatting on restaurant health code rating, and vermin and hand washing health code violations

Chapter 2

Prediction of Restaurant Health Code Violation Using Yelp

Introduction:

Food pathogens cause 9.4 million food borne illnesses in the United States each year.[1] In recent years the prevalence of certain food borne organism contaminants such as Shigella, Vibrio, and shiga-toxin producing Escheria coli have been on the rise [2]. Outbreaks of foodborne illness in the United States have increased from 675 in 2009 to 852 in 2010[3]. Restaurants have been connected to food borne illness outbreaks in the past however the ability to detect outbreaks is limited by variability in reporting of foodborne illness. Health Departments face a burdensome task of assessing a growing number of restaurants while confronted with funding challenges. In the last year, the San Francisco, CA Department of Public Health conducted inspections of 7,000 restaurants and other food serving establishments one to three times a year[4][5], while the growth of the restaurant industry increased at a rate of 4.9% per year.[6] Further contributing to the problem, risk of contracting food borne illness is also increasing on an annual basis for specific food pathogens.[3] The majority of cases of foodborne illness are traced back to food served at restaurants[7]. Health code violations such as sick employees participating in food preparation, and lack of employee hygiene are particularly difficult to detect given limited surveillance coverage available to local health departments.[8]

Annual health inspections only capture a small window of time and may not accurately reflect the true practices of an establishment. The coverage provided by health inspection covers less than 1% of a restaurant's annual operation time. Given that at most the San Francisco Department of Public Health will inspect a restaurant three times per year. Current risk ranking dictates that restaurants that receive a favorable health code rating (> 80 out of 100 points) will not be inspected again for the rest of the year. Additionally, while restaurants receiving suboptimal health department scores are more likely to be connected to food borne illness outbreaks, suboptimal scores currently predict a minority of restaurant related outbreaks. [2] Foodborne illness can be transmitted through multiple routes, including food handler practices, contaminated food products or equipment, contamination by vermin, poor employee hygiene, or malfunctioning sanitation or food processing equipment. There are many ways that establishments may contribute to the propagation of food borne illness, and not all of these risks can be detected by health inspectors in a short period of time[7]. Furthermore, the food production environment is very dynamic and changes with time[8,9]. Defining the utility of new epidemiologic surveillance tools that improve coverage of surveillance that use alternative means of detection is a useful first step in improving and monitoring restaurant practices to achieve greater public health.

Recently, there has been an increased interest in the use of social media sites like Facebook, Twitter, Craigslist, Google, and Wikipedia, to conduct different types of public health surveillance.[10–24] The use of social media to enhance public health surveillance is a new approach, that is beginning to gain momentum.[25] In a recent work by Generous et. al examining use of Wikipedia access logs to track global disease incidence, four

challenges are set forth to academicians that wish to identify approaches to use social media to track health measures in the real world[26]. The challenges to be met are; "openness, breadth, transferability, and forecasting". The authors state that open source data and open source code are necessary to ensure that achievements can be built upon by third parties, the authors also state that models should demonstrate the ability to be adapted from one disease context to another. In other words models should offer some degree of exportability. Transferability refers to the ability to use models in places where incidence data does not exist due to lack of tracking or an inability to access information. In other words, robust models that do not require new training set data when transported to new locations. Ability to provide forecasting and "nowcasting" of disease incidence is also set as a challenge by the authors. When referring to forecasting the authors also state that; "models should provide not only estimates of the current state of the world — *nowcasts* — but also *forecasts* of its future state". When describing the work in our current study we shall show how our study meets or does not need to meet these four challenges.

Yelp.com is a social media site where individuals may go to freely write and post reviews of restaurants they have personally experienced. Reviewers may also assign a star rating to a restaurant that denotes the reviewer's personal opinion of the restaurant's quality. This star rating is also aggregated into a composite star rating that reflects the average opinion of yelp reviewers. In addition to stars yelp also tags each restaurant with data on how many reviews have been written on the restaurant, the aggregated "expensiveness" of the restaurant which is measured by a range of 1-3 dollar signs. Reviews are scored by users according to whether they are "cool", and or "funny" and or "useful". Yelp uses a proprietary algorithm to rank the order in which reviews are viewed for any given restaurant. Yelp also applies filters on reviews to remove reviews that are not based on personal experience, incorporate spam or irrelevant hyperlinks, or are written by reviewers with few reviews on Yelp and few connections (friends) that are also Yelp reviewers.

In Yelp.com, restaurant reviews and informative tags for restaurants can be accessed publicly through Yelp's application programming interface (API) or by using web extraction tools via programming languages such as R, Python, Ruby or Perl which are all open source. Web extraction of Yelp data by account holders is referred to as allowable use in the Yelp.com terms of use agreement. Yelp.com accounts are free to create and may be used with very few restrictions. [23] In this way Foodborne illness-related keywords in free text review fields can be identified, tagged, and tracked spatially and temporally. Such methods provide epidemiologists with key surveillance data needed to identify clusters of at risk restaurants that could alert public health officials to potential food borne illness risks, allowing them to prioritize restaurant inspection of high risk institutions before assessing others. Such a system also allows for increased coverage of criteria that would trigger inspection. For example, multiple reports of food borne illness symptoms within a given time period, visualization of vermin, or employee hygiene breaches. This system could also be used to track prevalence of high risk restaurants in a given area and or period of time.

However, before these techniques can be employed it is necessary to better understand/validate the method's potential for prediction of suboptimal health inspection scores/health code violations. The first step to understanding this potential is defining the predictive power of this surveillance tool. A recent study in Mortality and Morbidity Weekly Report (MMWR) found that Yelp data in New York City could be used to identify food borne illness outbreaks.[24] While this method of surveillance was able to identify three food borne illness outbreaks that had gone unreported, it also required additional staff time to "read reviews and send emails" and "interview Yelpers." Additional services of a food borne illness epidemiologist were also required. This method of surveillance, while effective, may be beyond the financial means of many public health departments. It is important to note that the focus of the New York study was specifically that of outbreak surveillance. Our study focuses upon surveillance of a related but different subject which is surveillance of restaurant health practices. In this context surveillance of restaurant risk factors may be of even greater value than that of surveillance on outbreaks. This is because surveillance of risk factors allows for interventions to be taken that may in turn prevent outbreaks from occurring. Of course if resources are available, then both approaches could be used in adjunct with traditional inspection methods.

This study validates a method of web-based surveillance similar to that conducted in New York City by the New York Health Department and the CDC[28]. Our method is different in that we are trying to detect health code violations that increase risk of food borne illness transmission. In this paper we aim specifically to show that our method is able to form robust predictions of prevalence of health code violation, identify restaurants with high risk of health code violation, and validate increased surveillance coverage by using free text and tags created by reviewers on Yelp.com. Yelp.com is a website devoted to providing reviews on local stores, restaurants and services. Our predictions represent a snapshot of restaurant practices weighted towards the present day. Yelp's proprietary algorithm pushes the most recent and relevant reviews to the first page of reviews of a restaurant. This selection of relevant reviews also allows our predictive model o be based upon the most recent reviews provided by Yelp reviewers. Our predictive model achieved all aims and was validated as a new tool that can be adopted by public health officials to improve surveillance scope and risk factor detection. Our study took advantage of the existing standard of the health code rating to measure the effectiveness of the model. Furthermore, our study models prevalence of health code violation over a three year period while the MMWR report covered only one year. Our approach can easily be adopted by many health departments without additional expenditure in terms of time and staff. Our model is created with the aim of improving detection of restaurant risk factors which is very different from detecting food borne illness outbreak detection. By providing an improved scope of detection inspectors may make more informed choices on which restaurants present the greatest risk and have the greatest need for re-inspection.

Methods:

The approach tested in this study is predicting health code ratings and health code violations. Health code violations are measurements of the risk a restaurant poses to public safety based on observed deficiencies during an annual health inspection. Health code violations may be cited if sick employees are working, vermin or signs of vermin are observed, improper use of sanitation equipment or unhygienic behavior is observed. A health code rating is assigned by a health inspector after reviewing all hazard critical control points in a food serving establishment (restaurants, food truck, sidewalk vendors). The health code rating is set based on the number and severity of health code violations a restaurant incurs. Restaurants are divided into three categories by SFDPH category one restaurants are graded two to three times each year and receive scores ranging from 0-100. This study is focused solely upon prediction of scores for category one restaurants. Our approach sets out to achieve the goal of improving detection of restaurants that are public safety risks with little added burden in terms of time/cost. This is achieved by detecting keywords and tags via yelp reviews that may be related to deficiencies in restaurant procedure or practice that may result in health code violation/citation. Reporting food poisoning or symptoms of food poisoning after eating at an establishment could be predictive of a variety of health code violations however the means of detecting this deficiency is very different from the surveillance employed by health inspectors. We attempt to employ and validate such an approach in this study.

Coding:

Coding of model terms was based upon the author’s knowledge of the scoring rubric employed by San Francisco health inspectors when conducting restaurant health inspections. Keyword selection was first based upon relation to Hazard analysis and critical control points on which restaurants are graded. The Authors then selected terms that were correlated with the low health code rating when deciding upon inclusion or exclusion in the model. Terms could also be directly related to aspects of the fourteen possible health code violations a restaurant in San Francisco could receive:

(Overview of Inspection Requirements San Francisco Department of Health 2015

- All walls, floors and ceilings must be clean and intact without large cracks or holes.
- All foods must be stored 6" off the floor to facilitate cleaning and sweeping of floors and to prevent vermin harborage.
- No vermin (rodents, insects or other pests) infestation upon the premises.
- All food storage must be arranged to prevent cross contamination. Foods are to be stored to prevent possible contamination from hazardous materials (i.e. bleaches, cleaning liquids, etc.)
- All equipment used in daily operations is to be in good running order. All storage areas and shelving must be clean.
- There shall be sufficient regular refuse collection to prevent garbage problems (overfilled receptacles causing garbage accumulation problems).
- All food service workers shall exhibit good personal hygiene and work habits (i.e. good health for the worker, cleanliness of outer garments, proper food handling, etc.)
- All establishments serving food shall have an employed Certified Food Handler to comply with AB1978 (Campbell Bill)
- All food facilities are to comply with the Labor Code Sec. 6404.5 which prohibits smoking in enclosed workplaces.

Author review of yelp reviews across different yelp reviewers and restaurants also proved to be useful in identifying keywords that would be predictive of low health code rating. Authors restricted reading to only the first part of the Yelp dataset to prevent bias in model creation.

Data Extraction:

A web extraction program was created that would extract review data and aggregated tags from the Yelp website. Data were extracted from Yelp reviews on Chinese Restaurants in San Francisco.   Restaurants in San Francisco were chosen due to the unique embedding of Yelp’s Local Inspector Value Entry Specification (LIVES)[25] formatted health score data in restaurant pages in the San Francisco section of the Yelp website. LIVES is a format that allows public health inspection data to be inserted into the corresponding restaurant page on the Yelp website. This allowed for review data and public health data to be extracted simultaneously via web extraction tools. Chinese restaurants were chosen as a pilot study specifically due to their greater reported prevalence of health code violation in the San Francisco Department of Health database with a prevalence of 25% in Chinese restaurants vs. 7% prevalence in other restaurants.[5] Selection of high risk populations when testing a surveillance or screening tool is a standard method in conducting such cross sectional studies. By using this high-risk population we were able to validate the usefulness of our screening tool under favorable circumstances. Using this high-risk population to better measure the effectiveness of our model aided in testing model sensitivity, specificity and area under the Receiver Operator Characteristic (ROC) curve. To validate the generalizability of this study we followed up our pilot study with a random sample that was reflective of all areas and cuisine types in the city. Extracted data were parsed and analyzed using R statistical programming language.[26][27] After extraction, Pilot study restaurants were randomly separated into two datasets of 220 restaurants each. These datasets were used for training and validation of the model predicting substandard health code rating (Health code rating <80). This model was also employed upon the larger sample of San Francisco Restaurants that represented all areas and cuisine types in the city. This sample consisted of 1,543 San Francisco restaurants. Restaurants were selected from the list of restaurants in the public dataset of restaurant inspections found at https://data.sfgov.org/Health-and-Social-Services/Restaurant-Scores/stya-26eb. Restaurants that were not classified as "category one" were excluded from analysis.

To further validate the exportability the same model was also employed on a sample of all restaurants in New York City. This sample did not exclude any restaurants based upon restaurant cuisine type. In order to construct our New York City Dataset we extracted our usual tags and keywords from Yelp.com website pages for New York City. We then merged data from our sample with health score data found on https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/xx67-kt59. Once we had constructed a dataset representing 755 restaurants in New York City and we had parsed all the same keywords that we had analyzed in our pilot study, We were able to apply the exact same logistic regression model that we created from Yelp reviews on Chinese restaurants in San Francisco, and apply that model to data created by Yelp reviewers in New York City. This analysis allowed us to further validate the ability of our model to "now cast" the health behaviors of restaurant owners and employees in areas outside of San Francisco that may have different slang/linguistic variation, regional memes, and place different value on food borne illness information. Furthermore, by evaluating the robustness of our model in New York we are able to display our model's transferability in events where restaurant inspection data may not be available.

Model Creation:

The purpose of this predictive model was to classify restaurants as above or below the health code rating <81 threshold in San Francisco and >14 for New York City. The San Francisco Department of Public Health (SFDPH) uses this threshold to set the frequency a restaurant is inspected on an annual basis. We used a logistic regression model based upon aggregated tags, keywords, and a combination of aggregated tags and keywords. Tags included: number of reviews, price (which is depicted by the number of dollar sign symbols), Page Rank/"Usefullness" (The page a review appeared on) and number of stars (average number of stars assigned by Yelp Reviewers). Keywords of positive and negative weight were also used in the model to predict a health code rating less than 81 (Positive and negative keywords displayed in table 2)

Keywords were selected based upon their correlation with restaurant health code rating. Keywords were measured by the frequency they appeared in the body of reviews related to the given restaurant. Keyword selection was also based upon relation to health inspection scoring rubric and author experience with language/slang used in yelp reviews. The most predictive models included keywords and tags combined. (Results displayed in table 1). As is common with language there were many keywords that were highly correlated. Using a robust number of keywords did enhance our model's predictive power however the high degree of co-linearity made it difficult to interpret the effect of specific keywords.

Therefore, we used principle components analysis to address this co-linearity and combined the variability of keywords into three dimensions which explained the majority of variability in our model while at the same time enhancing model interpretation. Dimension 1 was related to keywords describing vermin like mice, roach, spider, and rat. Dimension 2 was related to foodborne illness symptoms (nausea, vomiting, diarrhea, etc.) and the physical environment of the restaurant (humid, smelly, clean, dirty, etc.), and the behavior of employees (rude, pushy, employee, courteous, etc.). Finally, dimension 3 combined the overall sentiment of the reviewers in regards to the restaurant (affordable, the best, I love,etc.). This approach is of particular benefit to public health officials because it would inform inspectors not only which restaurants would be likely to be "substandard" but it would also predict the category (Vermin, Employee Behavior, Foodborne Illness, and Physical Environment) of observations that were powering the prediction. This information could cue inspectors to what to look for on future inspections. Please see table 3 for Confidence intervals and Odds Ratios of Odds of substandard rating for all three dimensions of covariates.

Table 1(Model Fits Across Datasets)

	AIC	BIC	LRT *
Model Tags and Keyword	994.079	1284.06	9.96x10-8
Model Keyword	1010.34	1280.81	*
Model Tags	1010.34	1029.05	*
Model Tags and Keyword(ALL SF)	159.54	390.54	*
Model Tags and Keyword (ALL NYC)	423.4	694.56	*

Table1. * Likelihood ratio test is comparing Tags and Keyword & Tag model only. Keyword and Keywords & Tags could not be compared due to different numbers of observations. Differences in AIC and BIC highlight different model aspects. BIC highlights TAGS as the driver of prediction in the model. AIC identifies the improved predictive power when using keywords with tags. LRT also identifies this relationship. Note that Model fit improves dramatically when applied to data with greater degree of heterogeneity. This is seen when applied to datasets representing the entirety of San Francisco and the Entirety of New York City.

Table 2

Keyword	Negative Correlation with Positive Health Code Rating	Keyword	Positive Correlation with Positive Health Code Rating
Vomiting	-0.163164396	Dishes	0.113003645
Truck	-0.156643603	Clean	0.10058891
Humid	-0.079674732	Recommend	0.092272018
high quality	-0.039595829	Excellent	0.089751973
food poisoning	-0.036498186	Affordable	0.087436841
Employees	-0.026620598	Delicious	0.077566414
		Service	0.074604752
		Fuck	0.068001779
		Fish	0.066839608
		Favorite	0.066197252
		Fabulous	0.065465611
		Ache	0.062553019
		Craving	0.052296809
		professional	0.051127693
		Pushy	0.050162638

Table 2. An analysis of correlation between tags and keywords was used to decide which terms would be included in the model. A correlation cutoff of .05 was used for inclusion in the model. Unless the authors strongly believed the keyword would be useful in the model despite low correlation.). A liberal cut off point was used to include as many predictors as possible. Correlation is specific to Pilot Study Training data which excluded all but Chinese restaurants.

Table 3

	Estimate	2.50%	97.50%
Reviewer Sentiment Towards Restaurant	0.907	0.822	0.977
Physical Environment and Vermin	0.978	0.849	1.093
Food Borne Illness Related Symptoms	1.122	0.998	1.260

Table 3. An analysis of the Dimensions of combined Keywords in predicting odds of substandard health code rating using a logistic regression model with Substandard rating as the response. See Supplemental figure one for a factor plot depicting these three dimensions.

The first 220 restaurants in the pilot dataset (excluding all restaurants but Chinese restaurants in San Francisco) were used for model training. The second part of the dataset also consisting of 220 restaurants serves to validate the final model sensitivity, specificity and positive predictive value. An ROC Curve was constructed to allow analysis of the area under the ROC curve/predictive power of the model. (See figure 1 below) ROC analysis provides a measure of the probability that our model will rank a restaurant with substandard health rating below a restaurant with adequate health rating. For San Francisco data Health code rating less than 80 was the binary outcome variable being predicted by our model. This cut-off was used because it is the threshold at which the San Francisco Department of Public Health increases the number of annual inspections from 1-2 to 2-3 inspections per year. For New York City, the cut off of > 14 was used because the scale and direction of health code ratings differ from that of San Francisco. [4] New York City health code ratings included in our study ranged between 1-100, a score >14 is the cut off used by health officials to categorize a restaurant as having food safety compliance below the top level. The positive predictive value was plotted using our validation dataset, and SF, and NYC datasets to visualize the effect of different thresholds on identifying restaurants with high numbers of health code violations. (See figure 2 below)

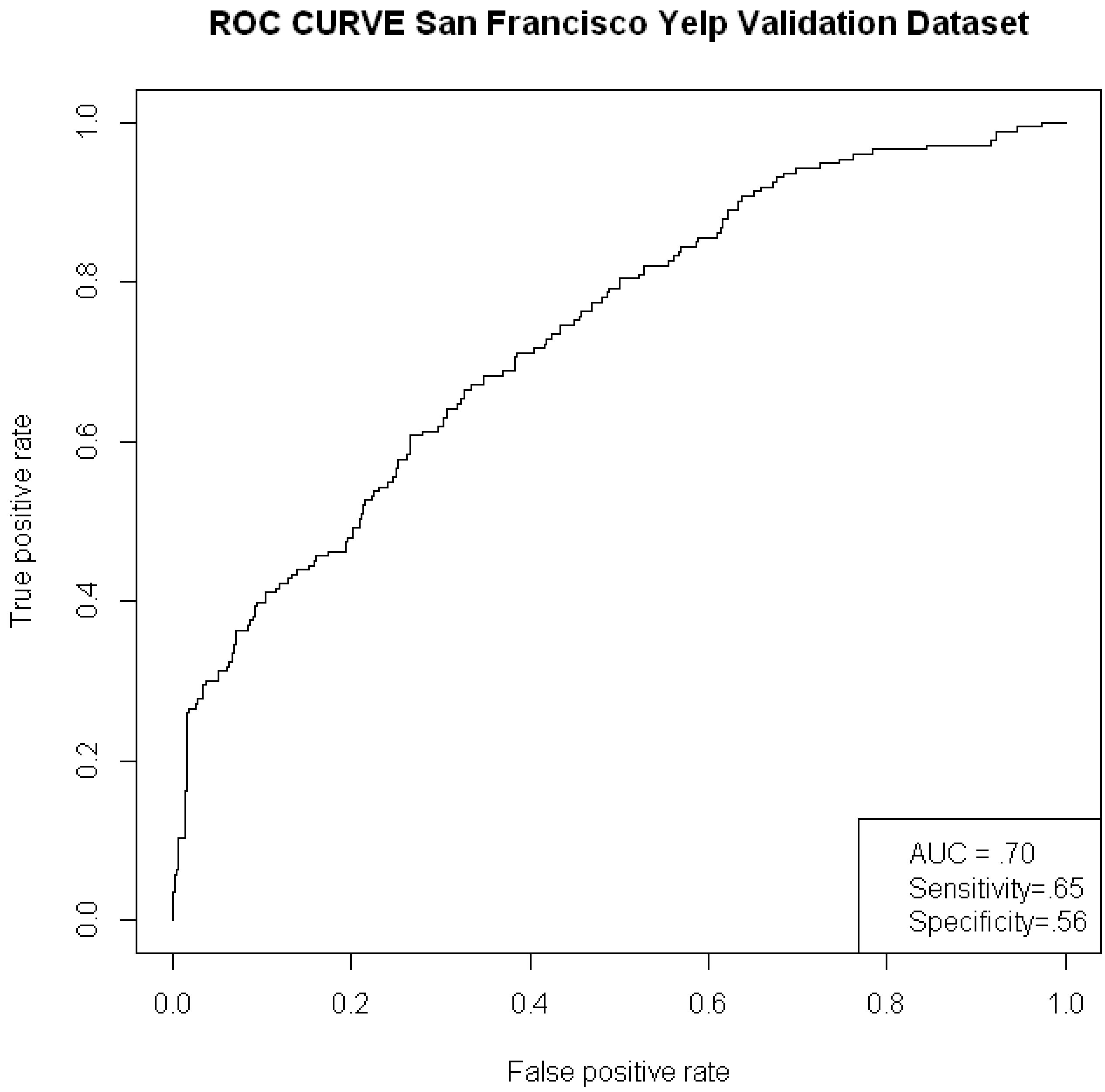


Figure 1. Extra Receiver Operator Curve using validation dataset created using Yelp data compiled from 220 San Francisco Restaurants.

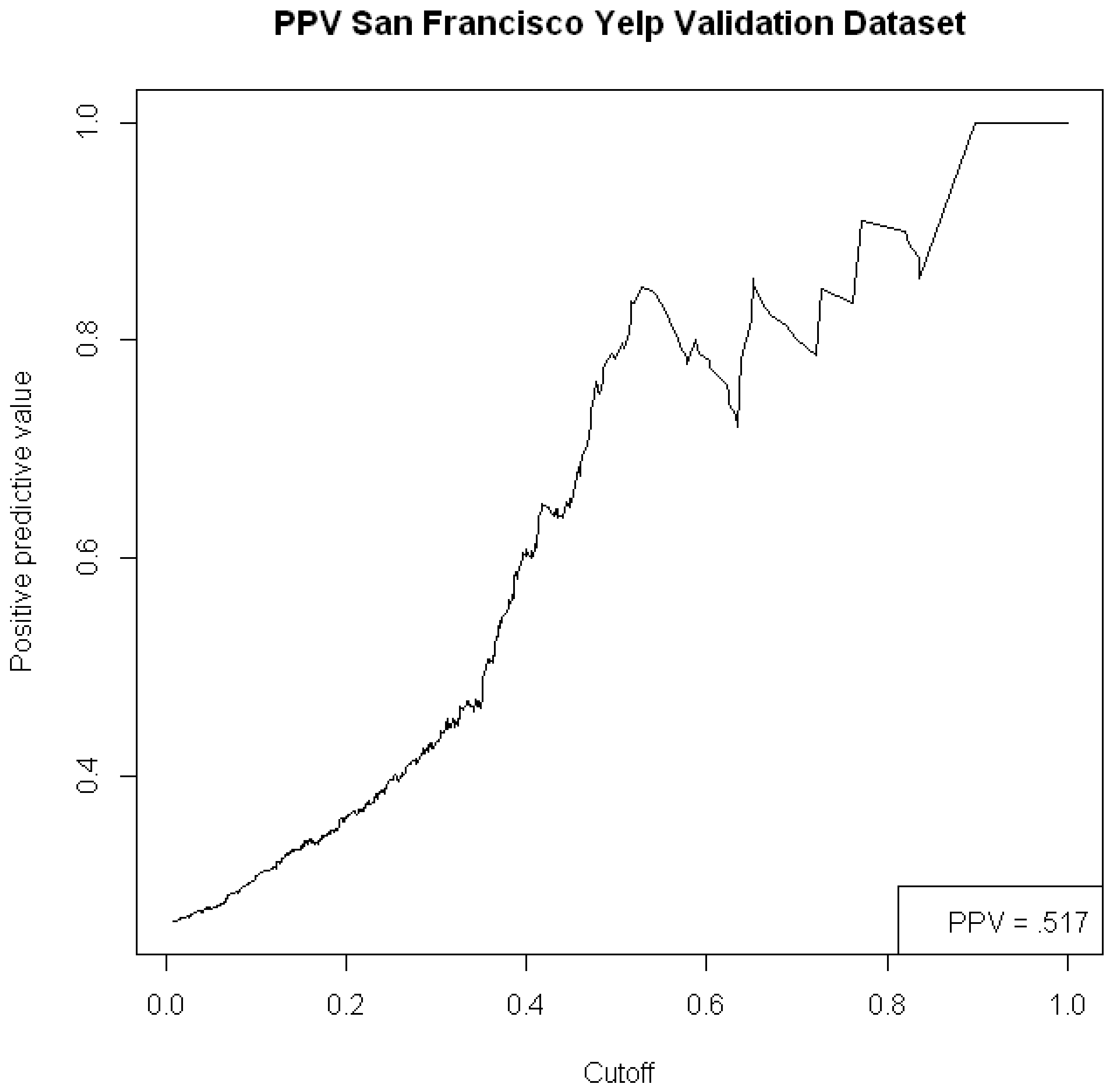


Figure 2. Positive Predictive Value using validation dataset created using Yelp data compiled from 220 San Francisco Restaurants.

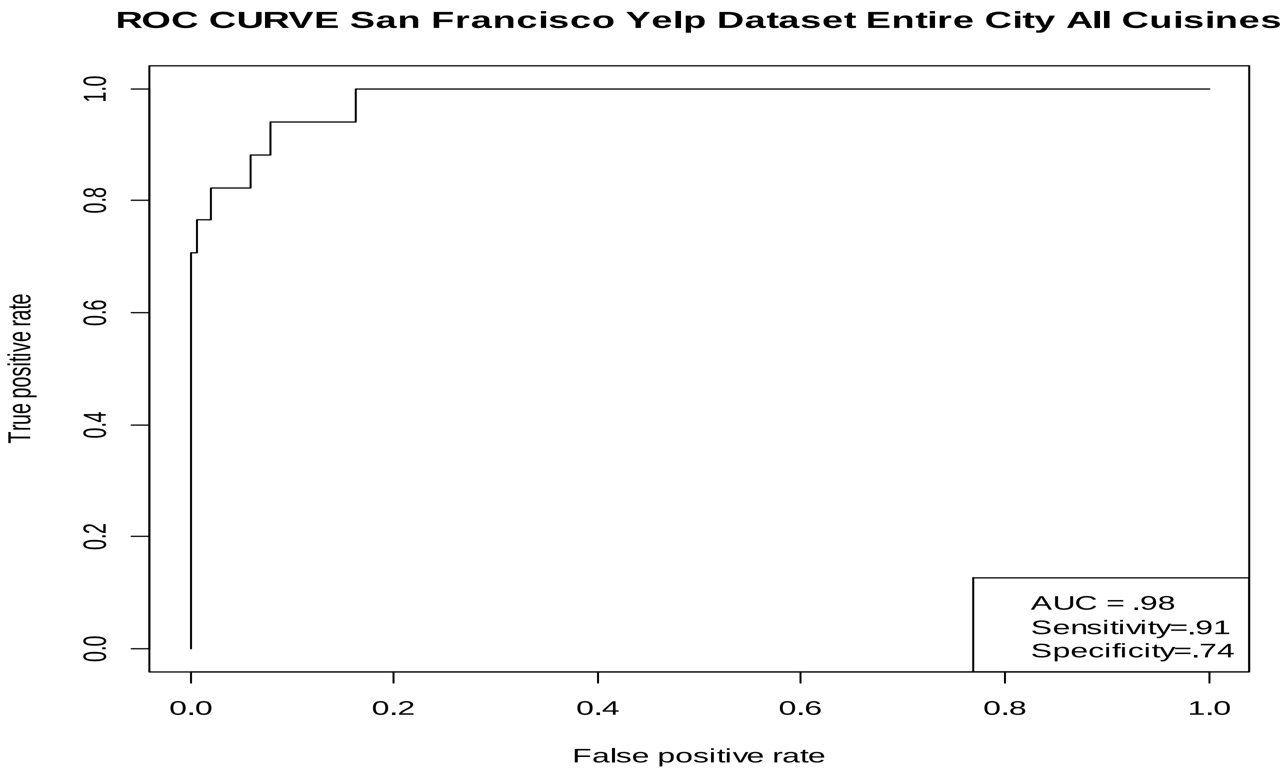


Figure 3. Receiver Operator Curve created using Yelp data compiled from 1,543 San Francisco Restaurants including all cuisine types.

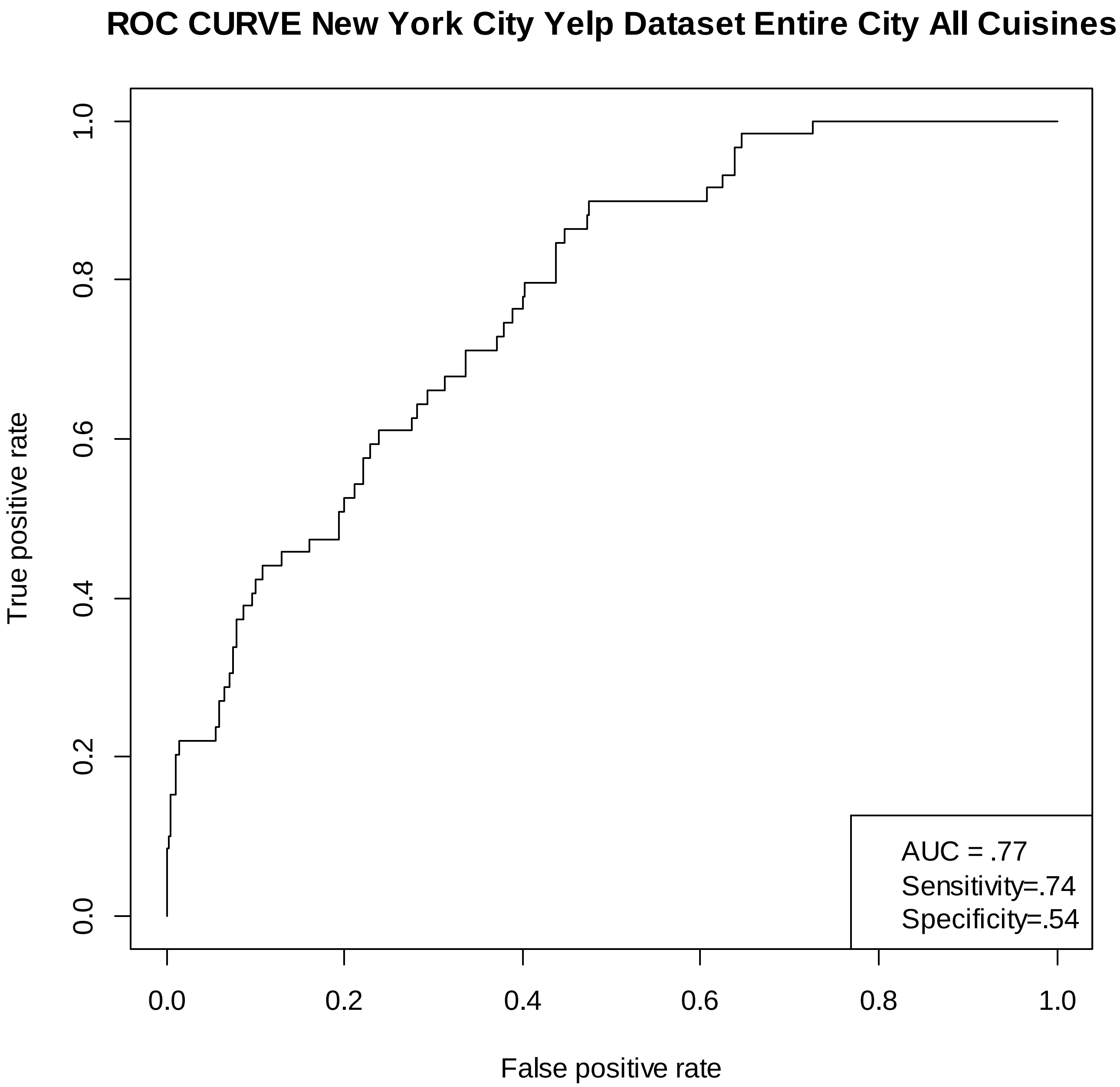


Figure 4. Receiver Operator Curve created using Yelp data compiled from 745 New York City Restaurants including all cuisine types.

Yelp uses a proprietary algorithm to order reviews. The more "useful" a review is (according to the proprietary Yelp algorithm) the higher that review will be placed among other reviews for the restaurant. This means the most "useful" reviews will be placed on the first page with less "useful" reviews being placed on the following pages. It is important to note that Yelp users have the ability to comment on usefulness of reviews. It is possible that the user rating of "usefulness" is used for Yelp's proprietary review ranking algorithm. If this is the case then if the majority of Yelp users believed that reviews discussing risks for food borne illness were important or useful then those reviews would gain better rank than others. Under this assumption, if only the "best" reviews were used for model creation then the model would become better at predicting restaurant health code rating. We divided groups of reviews according to which page they appeared on as a means of ordering reviews by "usefulness" as determined by the proprietary Yelp algorithm. We did this in an attempt to address whether our predictive model performed better using only reviews that were placed on the first page of reviews by Yelp. Including the usefulness term in the model allowed us to test the power of Yelp's proprietary algorithm without needing to know how it was constructed. Inclusion of "usefulness" in our model would also allow us to measure the importance that Yelp users were placing on reviews that contained keywords correlated with negative health rating. This is operating under the assumption that user review rating strongly influences the ranking algorithm.

Using predicted classifications generated by our logistic regression model we were able to compare the prevalence of low health code rating between real world observations and the predictive model. This was done to evaluate whether our models prediction of prevalence closely matched real world prevalence values generated by SFDPH. This part of the study was done one year after initial data collection using SFDPH data collected across inspection dates in 2014. Inspections were summed across months to more easily visualize the correlation between real and predicted values. The authors did not set out to determine temporality of review data. Review data was not collected past the last day of 2014. The specific aim of this part of the study was to further validate the fit of the predictive model and its reliability in identifying incidence rates of substandard restaurant scores across time. (See Figure 5, 6).

Simulation Methods:

In order to further validate our model we chose to use bootstrap simulation to better define the predictive power (in terms of positive predictive value, and area under the receiver operator characteristic curve (AUC). Simulation of AUC and PPV was used because it provided a useful measure of the possible variation in our prediction metrics. All simulation was conducted across 10000 iterations of the dataset using a 200 restaurant sample size. AUC was measured across different simulated parameters such as sample size, and page rank/"usefulness". Page rank was defined as the page number a review appeared on for a particular restaurant. If a restaurant had three pages of reviews then reviews occurring on the first page would have a page rank of one while reviews occurring on the second page would have a page rank of two and so on. First page reviews were of interest because Yelp appears to order their reviews in part by user rating. Our simulation allowed us to measure whether Yelp's proprietary algorithm for ranking reviews acted as a filter collecting more predictive reviews towards the beginning of a restaurant's listed reviews.

To test the power of page rank/"usefulness". A 10000 iteration loop was opened running analysis on the area under the ROC curve on random samples of the data using a sample size equal to 200 restaurants extracted from al restaurants observed. A logistic regression model was created inside each loop of the simulation. This model running on simulated data would then yield ; sensitivity, specificity, positive predictive value, and AUC using the R "ROCR" module for each iteration of the dataset.[28] Collected predictive measures were plotted via curve, histogram, and scatter plot to visualize the predictive power across simulated sample sizes.

We then examined simulated AUC and PPV at different levels of page rank by combining both parts of the real dataset and simulating a "page rank/usefulness" spectrum where a dataset transitions from 100% "high page rank data" to 100% low page rank data(reviews occurring on first page). This was done by randomly sampling from two datasets at different proportions within a loop, and combining the two proportions at each iteration to make one dataset for which AUC was calculated. The proportion of "low page rank" reviews changed from .5% low page rank reviews to 100% low page rank reviews creating a "Yelp Page Rank/Usefulness Index". This Index allowed us to visualize the effect of page rank/page number on the predictive power of specific reviews.

Results:

Keywords were defined as whether or not a keyword appeared at least once within a review. Keywords and tags that were not significant predictors of health code violations were left in the model if they added to the model's ability to predict low health code rating for restaurants within the study. The average number of Stars assigned to a restaurant as well as five keywords were significant predictors of low health code rating (Please See Table 4).

Table 4

Variable Name	OR	0.025	0.975
STAR(tag)	0.64	0.43	0.96
I love(keyword)	0.05	0.00	0.42
Affordable(keyword)	0.1	0.02	0.36
Microwave(keyword)	0.08	0.01	0.79
Vomit(keyword)	45.4	1.34	273.00
Dirty(keyword)	2.21	1.43	3.68

Table 4. Odds Ratio and 95% Confidence Interval of Odds Ratio are listed above. Table is limited to significant predictors. Additional terms that were highly predictive but not identified as significant due to co-linearity were not listed in this table.

Table 5 (Significant Predictors in New York City Model)

Variable Name	OR	0.025	0.975
recommend(keyword)	0.67	0.45	0.94
I found a	7.72	1.16	45.09

Table 5. Odds Ratio and 95% Confidence Interval of Odds Ratio are listed above. Table is limited to significant predictors. Additional terms that were highly predictive but not identified as significant due to co-linearity were not listed in this table.

The logistic regression model was evaluated in terms of sensitivity, AUC, and PPV. We see from our results in the simulated datasets that the interval of AUC went from .5 to 1 as the number of restaurants using only first page reviews ranged from 0.5%-25% of restaurants. While the range of AUC narrowed to .75-.9 when first page reviews were included in 75%-100% restaurants. (Please See Figure 5) We also measured the effect of sample size on Prediction Model AUC simulating sample sizes randomly drawn from our original sample we saw that there was no increase in AUC after a 300 restaurant sample size was reached. Sample sizes were simulated from 1-1000 restaurants.

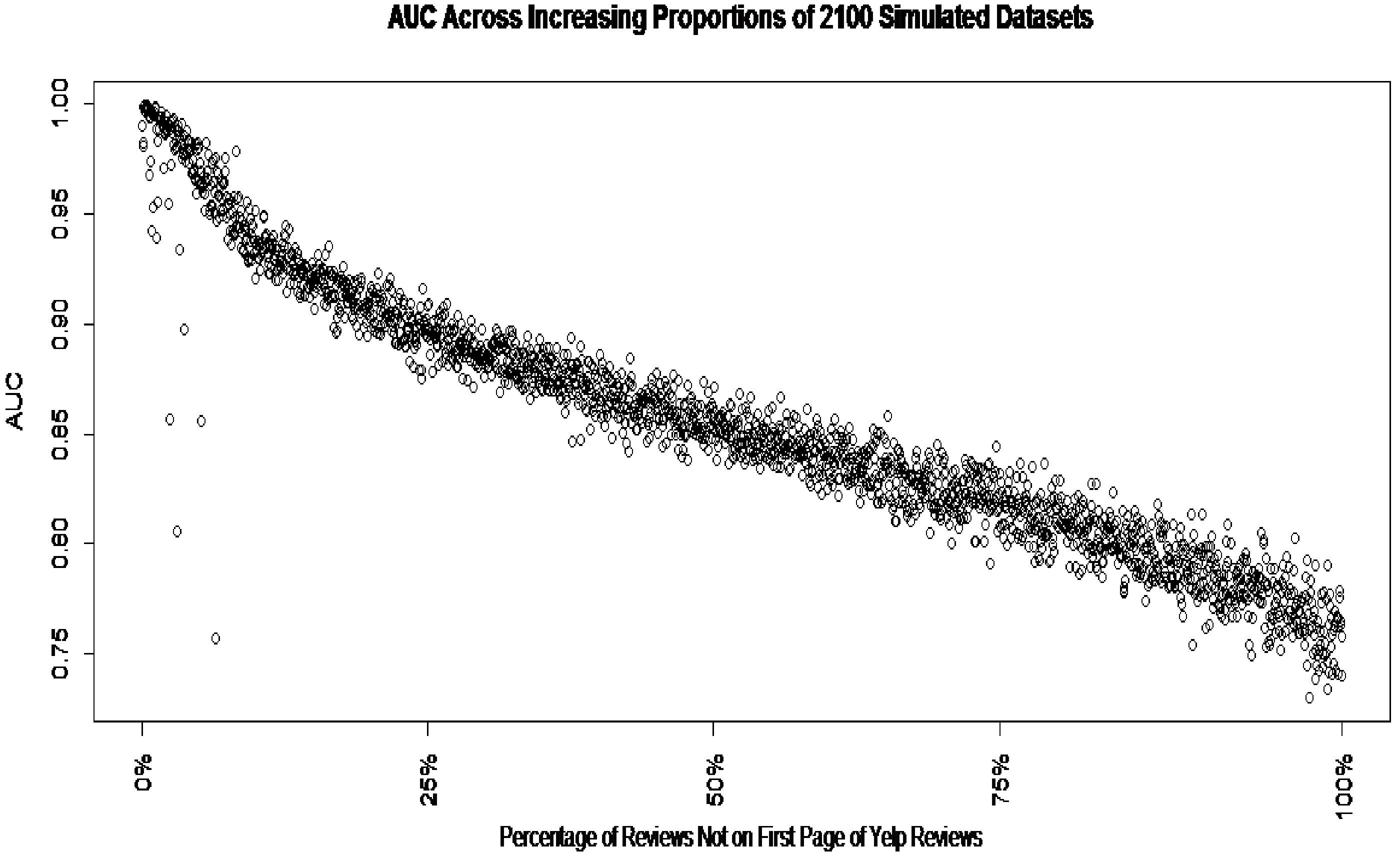


Figure 5. Plot of AUC across simulated datasets with varying proportions of first page reviews using the validation data set.

The AUC was .79 for the first (Training) part of the dataset and .7 for the second (Validation) part of the dataset (Please See Figure 1). This means that our model accurately separated the poorly rated restaurants from the highly rated restaurants 70% of the time. The simulated dataset had a mean of .78 AUC. Thus, across 10000 simulations our model was able to discern which restaurant would be poorly rated in 78% of restaurants on average. Measurement of model effectiveness in terms of prediction can be seen below. Simulated data were generated using 200 restaurant samples each represented by a minimum of 40 reviews. No page rank restrictions were placed on these data. (Please See Table 6 Below).

Table 6	AUC	Sensitivity	Specificity	PPV	Prevalence
Training Data	0.79	0.7	0.58	0.5	0.25
Validation Data	0.7	0.65	0.56	0.5	0.33
Simulated Data 10000 iterations	0.78	0.72	0.44	0.61	0.29
Sample of All San Francisco Restaurants with no cuisine exclusion	.98	.91	.74	.29	.10
Sample of All New York City Restaurants with no cuisine exclusion	.77	.74	.54	.25	.12

Training Data refers to the first part of the dataset used for model creation. Validation Data refers to the second part of the dataset used for validation purposes. Simulated data is a 200 restaurant random sampling and analysis repeated over 10000 iterations using the complete dataset from the pilot study. "Prevalence" Refers to the prevalence of restaurants with low health code rating in the specific dataset.

The prevalence of low health code rating for restaurants on average in San Francisco is roughly 7%[5] while the prevalence of low health code rating for Chinese restaurants is approximately 25% (the prevalence of low health code rating in our sample). The positive predictive value (averaged over simulations) was .61 with a standard deviation of +-.052. This indicates that the model improves the identification of restaurants with low health code rating with a 36% greater probability than chance alone.

Using bootstrap simulation our direct measurement of PPV was .5 for both parts of the simulated dataset and .61 for the simulated dataset (Please See Figure 2) The prevalence observed within our sample which was biased towards a high occurrence of substandard health code ratings (HCR<80 occurred between 26-33% of restaurants across 1st and 2nd parts of the dataset). The dataset representative of the entirety of areas and cuisine types in San Francisco produced had a PPV of .29 which was nearly three times higher than the prevalence of low health code rating in our sample which was .10. Sensitivity for this sample was very high at 91% and a specificity of 71% yielding an AUC of 98* meaning our model was able to rank and differentiate between adequately and inadequately performing restaurant in our sample 98% of the time.

The dataset representative of the entirety of areas and cuisine types in New York City produced a PPV of .25 which is over double the likelihood of detecting a substandard restaurant by chance alone. The sensitivity of the model in detecting substandard restaurants in new York city was 74% and specificity was 54% Yielding an AUC of 77% this meant that using the sample representative of New York City Yelp Reviews our model was able to rank and differentiate between adequately and inadequately performing restaurants 77% of the time.

By collapsing predictions and real reports of low health rating across months we were able to plot the predicted and real prevalence of health code violations across a two year period. (See figure 3) The predictive model did predict a greater than observed prevalence in 30% of the months observed, a perfect match in 25% of the months observed and predicted less than the reported cases in 45% of the months observed. The Pearson R^2 showed that predicted and observed values were correlated at .759%. A Hosmer-Lemeshow (HL) test was utilized to identify goodness of fit of the model. The Pearson chi squared statistic produced by this HL test was 12.39 with a p-value of .259 providing no evidence of lack of fit. The Hosmer-Lemeshow test was used to account for the >10 covariates included in the predicted model. Measurement of goodness of fit along with analysis of predictive power via AUC for the validation dataset of the pilot study, the boot Strap simulation set, the dataset representing the entirety of San Francisco, and the dataset representing the entirety of New York City (Inclusive of all cuisine types) validated the robustness and exportability of this predictive model.

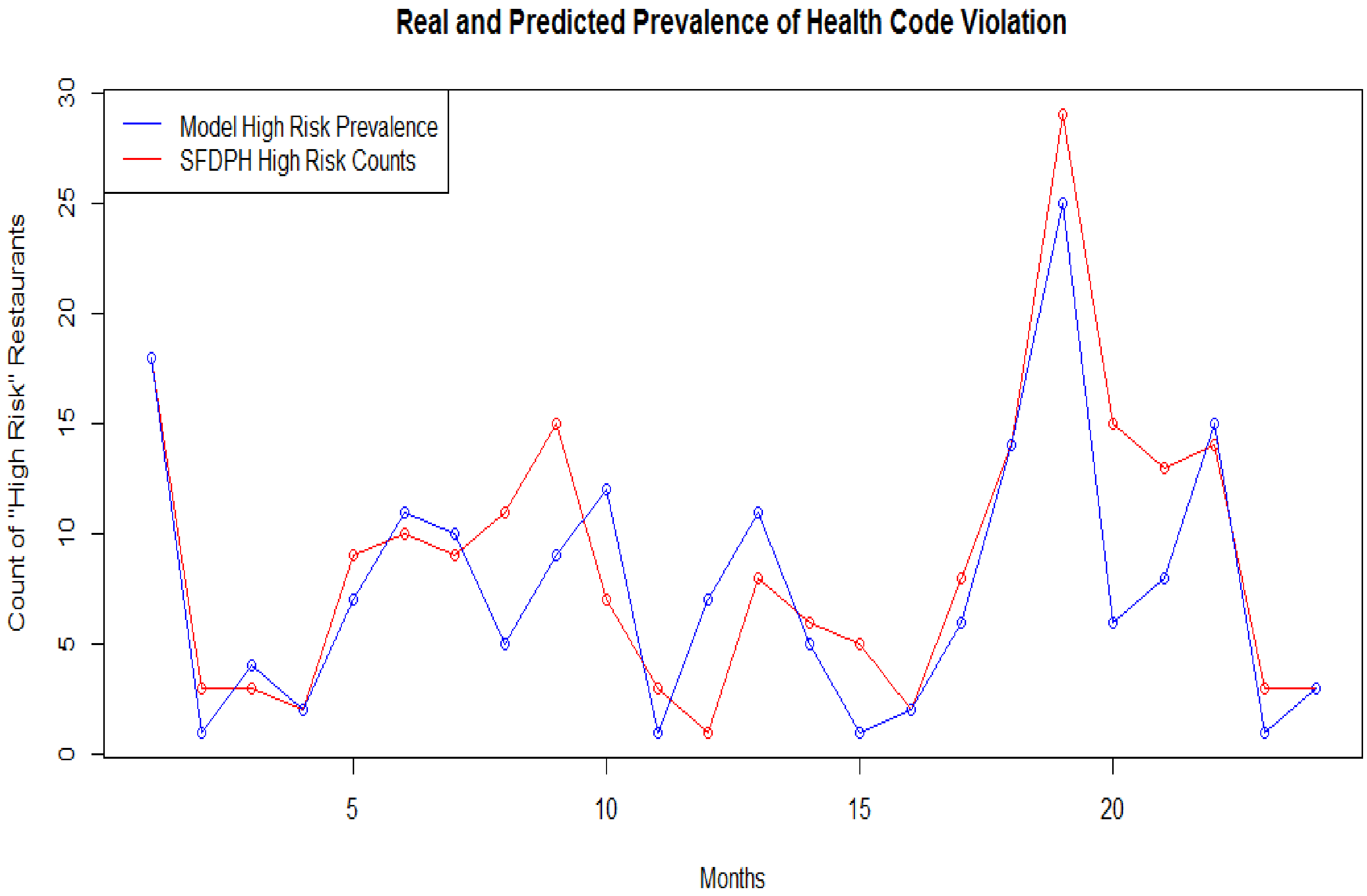


Figure 6. Plot of observed and predicted prevalence of low health code rating over a two year period from the beginning of 2013 to the end of 2014 using validation dataset for first year and all restaurants for second year. Blue Lines reflect predicted counts and red lines reflect the observed counts of restaurants with health code rating <80.

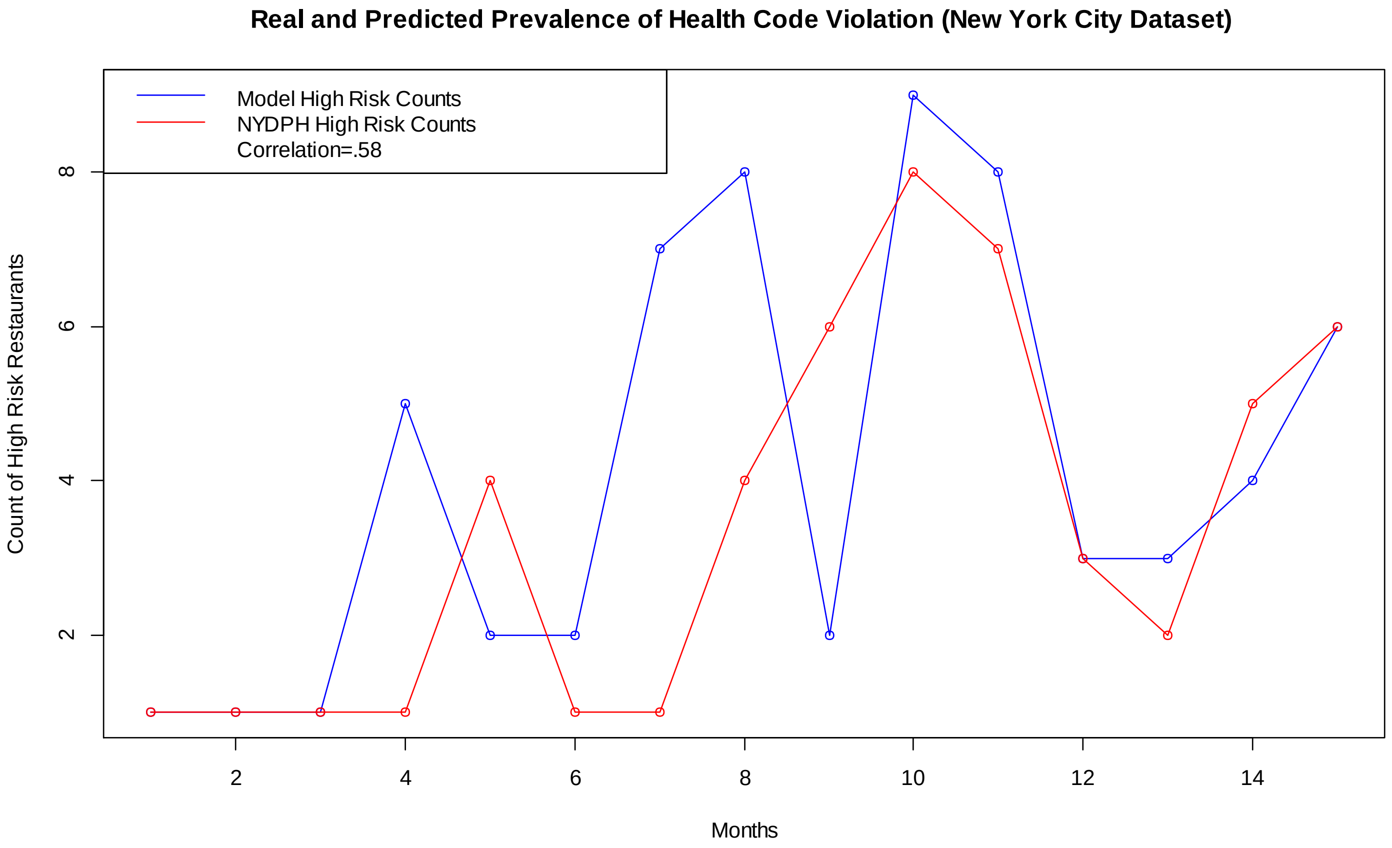


Figure 7. Plot of observed and predicted prevalence of low health code rating over a one year period from the beginning of 2014 to the end of 2015 using sample of all restaurants. Blue Lines reflect predicted counts and red lines reflect the observed counts of restaurants with health code rating <80.

Discussion:

There were several keywords that were significant predictors of health code violation. Unsurprisingly, aggregated Yelp "stars" are significant predictors of low health code rating. For each additional star added to a restaurant's average star rating the risk of low health code rating drops by approximately 36%. It is logical to assume that people who submit Yelp reviews who observe vermin, poor hygiene, or become sick after eating at an establishment are less likely to award stars than those who do not observe or experience these unpleasant things. "Affordable" and "microwave" were keywords that were significantly associated with decreased odds of low health code rating. Restaurants described as serving microwaveable products are most likely different types of food serving establishments with little to no prepping/food contact and therefore at lower risk for violation. Yelp reviewers that use a positive word like “affordable” have a good impression of the restaurant. Yelpers that perceive substandard hygiene, vermin, or those who become sick will not remark upon the affordability of the establishment. Those reviews that do remark upon negative circumstances tend to focus upon such information. In this context, there is a case why “affordable” could be predictive of an acceptable Health Code rating.

Negative keywords also predicted low health code ratings: the keyword “dirty” increased odds of low health code rating by 121% on average and “vomiting” increased odds of low health code rating by 45.4 times. The keyword “vomiting” is unique in that it is very significantly associated with low health code rating. Although this keyword occurred with low frequency, we see that restaurants with low health code rating have much greater likelihood of having reviews with the keyword “vomit”/”vomiting” than those without. This shows that single keywords can be powerful predictors of low health code rating. It is important to note that words like “pain,” “diarrhea,” “poisoning” and “ache” were highly correlated with the word “vomiting.” This multi-co-linearity (>.7) masked the significance of these words; however these terms are useful in that they do improve the accuracy of our predictive model.

Using Sample data collected from New York City Yelp reviews our model found different keywords that were predictive of substandard health code rating. These keywords were the keyword string “I found a” and the keyword “Recommend”. The keyword string “I found a” was used to detect statements related to signs of Vermin in restaurants, although the string of keywords could also detect lapses in employee hygiene such as “I found a hair in my soup” for example. Given that this keyword string was a positive predictor of a substandard health code rating it would appear that the keyword was used to denote something negative related to a restaurant the majority of the time. The keyword (Recommend) was also a significant predictor of substandard rating in New York Yelp Reviews. This makes sense in the context that those reviewers that recommend a restaurant are unlikely to make observations related to vermin, employee hygiene, or other elements pertinent to food safety.

Through simulation of datasets with different proportions of reviews with first page rank we can infer that the first page of review data by itself does the best job of predicting low health code rating. We can see that as additional pages of higher page rank reviews are added, the predictive power of the model decreases. Unfortunately, because Yelp's review rank algorithm is proprietary, we are unable to peer inside the black box to unveil why improved ranking of reviews improves their predictive power. We do know that Yelp allows users to click a box on a review if they believe it is "funny", "cool", or "useful". It is possible that this user data is incorporated into Yelp's review ranking algorithm. If so this would mean citizens of San Francisco who use Yelp may place high importance on reviews discussing keywords related to health code violations/health code rating. It is unknown if this importance would be similarly designated across other Yelp users in different geographic areas. If public health departments adopt a surveillance strategy based on user-generated content, it is important that models account for page rank to decrease the time for generating predictions.

By applying our predictive model to datasets representing the entirety of San Francisco and the entirety of New York City (in addition to validation datasets from our pilot study) we have validated that this specific model will work in a variety of geographic areas. Variations in dialect, slang, and local memes may have altered the effectiveness of this predictive model. However despite these variations our model was able to retain the power to discern substandard restaurants from their compliant counterparts 77% of the time in New York City. It is possible that speech used in Yelp reviews may not be similar to that found in other review sites, making this model only applicable to Yelp reviews. Public health workers may wish to become familiar with local Yelp speech/slang so they may populate the model with relevant keywords. In this way, area expertise can help to create a model specific to a linguistically unique geographic area. Customization of models to include keyword terms that reflect local slang and social memes may further enhance the predictive power. However in locales where there is an inability to provide such customization, our model will still be a useful adjunct to restaurant surveillance and can be used to rank a restaurants risk of violating health code regulations and possibly transmitting pathogens causing food borne illness In this context we see that the variability of keyword assignment that could be seen as a weakness is in fact a strength of this model in that it allows for (but does not require) increased localization of the predictive model.

A limitation of this approach is that it relies upon high participation by reviewers in the production of yelp reviews. Yelp participation is highest in large urban areas like San Francisco and New York City thus it should not be surprising that our model works well in these areas. As cities continue to grow and social media becomes more embedded in everyday use, the utility of this method will be likely to increase. It is important to note that the majority of Chinese restaurants studied were centered in the cultural enclave of "Chinatown" this is important in that it reflects that even in small geographic areas Yelp can be a powerful screening tool if there is support in terms of a user base within the local community.

Use of the methods outlined in this study can act as blueprint for those agencies that may wish to take the first step towards using crowdsourced surveillance. Our Study used open source programming languages R and Python to extract data from Yelp.com using a single user account and operating within the terms of use of that account. Since ownership of a Yelp account is a good that is publicly available we feel that both the data and programming languages of this study meet the “Openness” challenge set forth by Generous and colleagues. We feel that the code used for this project can easily be built upon by “third parties” to create even more robust models that may apply to other public health measures.

While our model focused specifically upon the measure of health code rating it is possible that this model could be used for more specific health code violations that are represented by health code rating. For instance vermin infestation, or employee hygiene citations could also be detected by this model. Although it is likely that increasing model precision would require further localization of the terms within the model. However, the potential for this type of surveillance to be used for public health measures beyond restaurant health code rating does show that this type of surveillance could be shown to meet the “breadth” challenge set forth by Generous and colleagues.

We do show that this model meets the “transferability” challenge in that it was shown to be robust when applied to samples generated from the entirety of restaurants in San Francisco and the entirety of restaurants in New York City. Furthermore we demonstrated that when a model has the benefit of localization (as with San Francisco) The model can achieve a high degree of predictive power as we achieved when our model yielded an AUC of 98% when applied to the heterogeneous sample extracted from all restaurants in San Francisco.

For this study, which focused on the prediction of public health behaviors practiced by restaurant owners and employees it does not necessarily make sense to forecast the behaviors of restaurants in the future. Instead the value of this model is that by extracting observations made by Yelp reviewers that are weighted towards the present day, we are able to “now cast” behaviors of restaurants outside of the normal window of inspection and or re-inspection. This increased coverage is most meaningful if the model could be used adjunctly when assigning risk to restaurants and using that risk to rank restaurants for re-inspection.

Future study including restaurants in multiple geographic areas or restaurants representing cultural enclaves would be warranted to further validate the findings we report here. However, since our sole purpose was to define the effectiveness of this surveillance approach, a small specifically-defined population and our larger follow-up studies were ideal for defining the effectiveness of our model. This study takes a step forward in identifying how social media data can have applications for the public health department of the metropolitan city.

Our Findings offer a first step towards the meaningful use of social media data in public health interventions. It is important to note that the San Francisco department of public health is currently participating in Yelp.com’s local inspector value entry system (LIVES) formatting system. This means that individuals writing reviews on Yelp.com will also have access to health code rating information on yelp.com. This system however remains unproven and usage statistics of public health data on Yelp.com are unknown. Health code violation data on Yelp.com can only be accessed through accessing hyperlinks that are not easily identified, and health code rating itself is on the periphery of restaurant pages on Yelp and may be difficult for users to detect. Without an evaluation of this new feature offered by Yelp we are unable to measure any inherent bias that it may create in our model. Furthermore, we did test this model in New York City which has not yet introduced Yelp LIVES formatting and we still identified a strong predictive power of our model.

Misattribution of food borne illness symptoms is a common finding in food borne illness outbreak investigation. We cannot say with certainty that all keywords related to food borne illness used in restaurant reviews represent foodborne illness that is causally linked with the specific restaurant for which the review was written. However, if we identify a high frequency of keywords related to foodborne illness (along with other predictive keywords and tags), in the body of reviews for a given restaurant then our results indicate that restaurant would have a 45 times greater odds on average of receiving health code violation/substandard health code rating. We observed this effect in samples generated from restaurant review and tags found in two large municipalities (New York City and San Francisco). Although we cannot expect all cities to have populations of Yelp reviewers and users that rival New York City and San Francisco today, as urban areas becomes denser and participation in social media becomes more common we expect the predictive power of this approach to continue to grow.

The approach outlined here is not something that will be easily replicated by the general public. However, this study offers proof that there is validity in this approach. It will be up to public health departments of large municipalities to build upon our methods and offer information to the public on the risks that have been identified by local citizens and aggregated via Yelp.com. There are biases that are inherent in the review data created by reviewers on Yelp.com, and by the proprietary algorithms applied by yelp.com. This study is the first of its kind to objectively evaluate this data source and measure its ability to predict substandard health code rating, and by using results of principal components analysis inform inspectors upon which aspects of reviewers' observations are driving the prediction.

It is important to note that the system of assigning health code violations themselves is not a perfect one. Health inspectors only review the operations of restaurants a few times a year[4][29]. While Yelp reviewers in San Francisco, New York City and other municipalities are creating hundreds to thousands of reviews for restaurants each and every year[23]. Yelp reviewers are also analyzing the safety of a restaurant in a far more direct way by actually consuming the food created by the restaurant. The real value of this method of surveillance is that not only does it catch elements that inspectors have already seen, but it also identifies what the inspectors have not yet seen, and may be unable to detect. One possible interpretation of prevalence prediction would be that our model was actually able to detect cases that the SFDPH could not. Indeed our model's lack of specificity may be showing that our model is detecting something that SFDPH inspectors cannot. Thus, we should not be overly conservative when judging the models, sensitivity, AUC, and positive predictive value, because the standard against which the model is judged is far from perfect.

Conclusion:

Mining publicly-available, crowd-sourced data to develop a surveillance method for tracking food borne illness risk factors gives health inspectors an improved ability to identify restaurants with greater odds of low health code ratings/violations outside of the normal inspection window. Our approach utilizes freely available data and utilizes open source software. Our code for data extraction and analysis is available in public repositories and may be built upon to increase the breadth and transferability of our model. Additionally, tracking clusters of food safety compliance and foodborne illness-related keywords in large, crowdsourced data sets improves traditional surveillance methods without substantially increasing costs. This study serves as a step forward in illustrating how social media data may be utilized for the benefit of public health.

Chapter 3

Impact of LIVES data in Yelp.com on Patron Restaurant Selection

Introduction:

Public health ratings for restaurants were created to enhance public safety and alert the public to possible danger. These ratings have been shown to be effective in significantly improving the practices of restaurants.[1–4] A study evaluating the impact of restaurant grade cards in Los Angeles found that there was a significant reduction in the number of food borne illness related emergency room visits since the introduction of the law mandating posting of restaurant grades in 1998[5]. New York and Los Angeles (Cities that have both adopted the card grading system) report 14% and 13% drops in food borne illness admissions since initiating the mandate that restaurants must display their health ratings in their window for all customers to see[6][5][7]. The success of this program has led public health officials in the San Francisco department of public health (which also has adopted posting of grade cards on restaurants) to expand this program from the doors of brick and mortar restaurants to the web pages of restaurant reviews on Yelp.com. However, it remains to be seen whether these postings will be equally effective in curtailing the public health behavior of selection of restaurants that present a public health risk. A study of online posting of restaurant grades in Salt lake city Utah found that restaurant scores improved significantly after posting grades online.[8] This provides evidence that online grade posting may influence patron restaurant selection and provide incentive for adoption of better food handling practices. However, this study did not show how customer perception of public health data may be altered when taken in alongside attractive and engaging information like Yelp.com picture, aggregated ratings, and reviews.

Yelp.com is a Social media site that allows people known colloquially as “Yelpers” to post reviews on businesses in a geographic region/locale[9]. People can use Yelp.com to obtain reviews and ratings on the quality of a restaurant. Information included in a restaurant’s page on Yelp.com would be price of meals, number of reviews, average number of stars, pictures of food served at a restaurant and all reviews posted for a given restaurant. Yelp.com users trust the ratings and reviews found on Yelp.com because they are generated from a large group of Yelp.com reviewers. It is not uncommon for restaurants in San Francisco to have thousands of reviews.   Even the reviews themselves are graded by users of Yelp.com and thus may be considered (by users) to have greater validity. Two years ago Yelp.com released a new feature "Local Inspector Value Entry System "(LIVES) formatting which allowed public health officials to post restaurant health ratings on Yelp.com in addition to patron reviews[10]. San Francisco was an early adopter of the (LIVES) format and other cities are beginning to follow San Francisco’s example. A possible benefit to public health officials is that their health ratings are now more accessible than ever before. The ultimate goal would be for negative health ratings to deter the public from eating at a given establishment. This creates an incentive for businesses to improve their practices, achieve an adequate score, and attract more customers[11]. However, public health officials must commit to formatting and sending their inspection data to Yelp.com on a regular basis. This would mean that public health departments may need to augment time budgeted to data management in order to keep up with Yelp.com reporting. Allocating staff to unfruitful assignments may negatively impact program performance in assessing public health risk[12,13]. Small business owners may feel that they are being penalized in a new way as the public access of this rating system may be more influential than displaying a Rating Card on a door or window. Currently there has been no publicly available validation of the LIVES reporting system examining the effect of public health data on Yelp.com user restaurant selection.

Currently in San Francisco all businesses receive a score from 1-100 that reflects a business's ability to adhere to public safety guidelines[11–13]. A health score less than 80 is considered substandard while a score greater than 80 is considered adequate. Those restaurants with a score less than 80 will receive an additional follow up visit after inspection to determine if the proper corrections have been carried out. Restaurant owners are required by law to present their health code rating in the door or window of their establishment so the score can be easily accessed by the public. Health code violations are publicly accessible on the San Francisco department of public health (SFDPH) website[16]. However violation data are not immediately available in restaurant doors or windows. Ratings and violations are directly correlated with restaurants with greater, and more severe violations receiving the lowest health scores[17].

LIVES data on Yelp.com presents a restaurant's public health score on a restaurant's page on Yelp.com. If a user wants to view health code violation information on Yelp.com they must click through two links to view the violations. At this point in time it is unknown how often people access public health data on Yelp.com. It is also unknown how people that use Yelp.com.com interpret the public health ratings and health code violations posted on Yelp.com. The goal of this study will be to measure perception of accessibility of public health data on Yelp.com, the degree to which negative public health data may deter Yelp.com users from selecting certain restaurants, and the degree to which positive information on Yelp.com may negate the effect of negative health information on restaurant selection, and finally the association between the valuation of public health data and reported food borne illness and Yelp.com valuation and reported food borne illness.

Methods:

An electronic survey of students and faculty at the University of California San Francisco, Berkley and San Francisco State University was conducted using the research electronic data capture (REDCAP) system[18]. This system was used to consent track and collect survey responses of 808 participants. 792of 808 survey responses were complete. 12,609 email addresses were collected from public directories using Python [19]web extraction tools. A 6% response rate was achieved and a representative sample with +-5% margin of error was generated. This survey tool and respondent recruitment letter received authorization from the University of California Irvine internal review board. The survey used consisted of 20 survey questions which were either multiple choice True False or fill in the blank. Standard demographic variables like age, gender, and race were collected from each survey respondent. Additional confounding variables like history of food borne illness, cuisine preference, history of Yelp.com usage and Yelp.com review writing, and frequency eating out were also adjusted for. The intent of this survey was to measure how Yelp.com information and public health information influenced patron's decisions to select restaurants. In order to achieve this measurement respondents were asked; under which Yelp.com rating (1-5 stars) a restaurant would be selected, and under which health rating (1-100 100 being the best) would a restaurant would be selected. Respondents were then asked if a Yelp.com rating or health rating were perfect would they ignore a substandard Yelp.com or substandard health rating? (Please see supplemental figure 1 for the full survey tool). The answers to these questions were used to determine the relationship between the value placed on Yelp information and public health information in Yelp and to assess the value placed on public health information by Yelp.com users, and finally to see whether the value placed on either information type was related to a reported history of food borne illness.

Study Population (Survey)

San Francisco is a minority majority city where the white racial group makes up the minority of the population[20]. However, Yelp.com users have their own unique demographics which the website surveillance site quantcast.com describes as Majority white. With 65% of our sample reporting being white our sample is reflective of the distribution of actual Yelp.com users as reported by quantcast.com[21]. However there is a greater proportion of users reporting other/mixed race and this may be reflective of the ethnic diversity found in San Francisco. This is reflective of an accurate representation of San Francisco Yelp.com users. We see a difference in racial representation between our sample distribution and the quantcast.com reported user base distribution because quantcast.com represents the entire U.S, user base while our sample merely represent the San Francisco user base. A bias of this sample is that there is a greater representation of those greater than 50 years of age compared to the user base reported by quantcast.com. Our survey was distributed to students, faculty and medical professionals at UCSF, SFSU, and Berkley. Medical professionals and university faculty who are older on average may have been more likely to complete the survey than students due to their interest and familiarity with research. Despite this difference, our sample represents each age group equally, and all statistical models are adjusted for age. Gender is also disproportionately female when compared to the San Francisco population but is reflective of Yelp.com user demographics reported by quantcast.com. There were several demographic groups that initially seemed disproportionately represented when compared to the populous of San Francisco, such as younger age groups, Asian race, women, and Yelp.com users. However, using the user demographics found on quantcast.com (a site used to describe user base of a website, and popularity of website) we found that this demographic data was representative of a typical Yelp.com user base. Unfortunately, we were unable to measure if other parameters we adjusted for such as cuisine preference or food borne illness history were also representative of Yelp.com users. Our sample did not include a large number of African American respondents and so may not be exportable to those user groups. (Distributions of all confounding variables included in the model can be found in table 1). It is also important to note that the majority (90%) of our sample reported using Yelp.com "occasionally" to "frequently". In this context our sample was representative of typical Yelp.com.com users/review writers in the San Francisco area.

(Please see supplemental material 2 for demographic distributions of San Francisco residents, and Yelp.com Users).

Table 7

	San Francisco County	California
Population, 2014 estimate	852,469	38,802,500
Population, 2013 estimate	841,138	38,431,393
Population, 2010 (April 1) estimates base	805,195	37,254,503
Population, percent change - April 1, 2010 to July 1, 2014	5.90%	4.20%
Population, percent change - April 1, 2010 to July 1, 2013	4.50%	3.20%
Population, 2010	805,235	37,253,956
Persons under 5 years, percent, 2013	4.60%	6.50%
Persons under 18 years, percent, 2013	13.40%	23.90%
Persons 65 years and over, percent, 2013	14.20%	12.50%
Female persons, percent, 2013	49.10%	50.30%
White alone, percent, 2013 (a)	54.30%	73.50%
Black or African American alone, percent, 2013 (a)	6.00%	6.60%

American Indian and Alaska Native alone, percent, 2013 (a)	0.80%	1.70%
Asian alone, percent, 2013 (a)	34.40%	14.10%
Native Hawaiian and Other Pacific Islander alone, percent, 2013 (a)	0.50%	0.50%
Two or More Races, percent, 2013	4.10%	3.70%
Hispanic or Latino, percent, 2013 (b)	15.30%	38.40%
White alone, not Hispanic or Latino, percent, 2013	41.60%	39

Table7.Demographics of San Francisco

Table 8 Demographics of Yelp.com

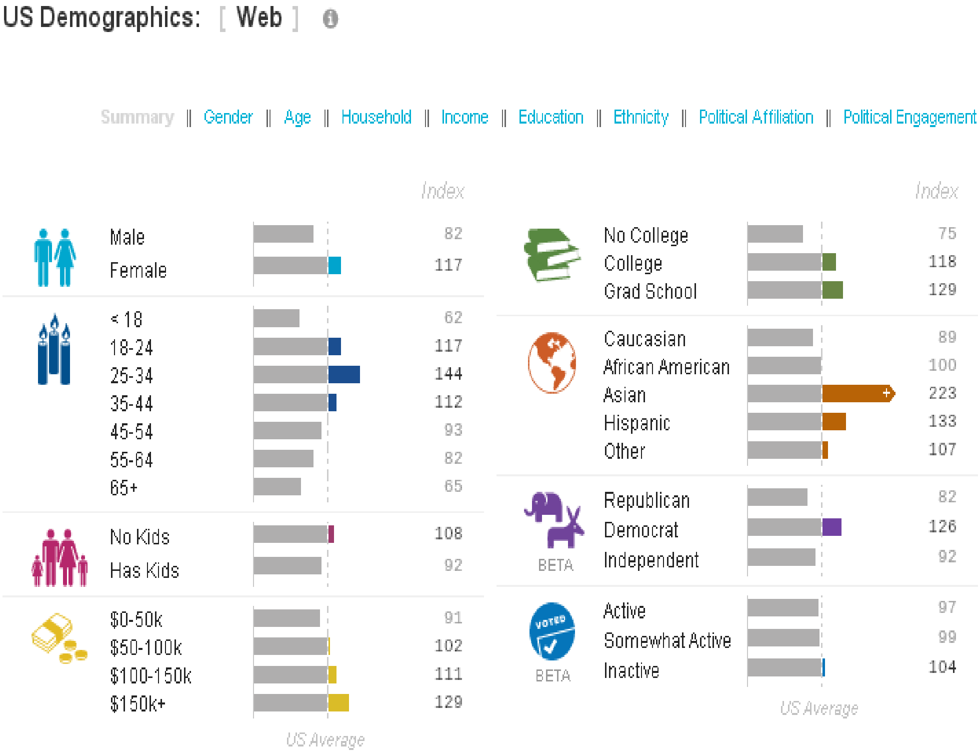


Table 8. Demographics of Yelp.com for United States Users using Quantcast Usership Index. Index Scores above or below 100 represent an above or below average usership when averaged over all userbases assessed by Quantcast.com.

Table 9

	Variables	Counts	Percentage
Age	18-29 Years Old	136	15.83%
	30-50 Years Old	390	45.40%
	>50 Years Old	325	37.83%
Gender	**Male**	295	34.34%
	Female	557	64.84%
Ethnicity	**White**	545	63.45%
	African American	29	3.38%
	Asian	179	20.84%
	Hispanic	70	8.15%
	Other	144	16.76%
Cuisine Preference	**Chinese**	408	47.50%
	Japanese	485	56.46%
	Mexican	569	66.24%
	Italian	429	49.94%
	French	234	27.24%
	Vietnamese	370	43.07%
	American	369	42.96%
	FastFood	78	9.08%
	Greek	281	32.71%
	Indian	422	49.13%
	Thai	521	60.65%
	Lebanese	137	15.95%
Frequency Dining Out	**Never**	2	0.23%

	Less than four times a month	321	37.37%
	2-4 times per week	450	52.39%
	4+ times per week	81	9.43%
Frequency writing Yelp reviews	Never	556	64.73%
	Occasionally	264	30.73%
	Frequently	12	1.40%
	Every Time I go out	2	0.23%
	I don't know	2	0.23%
Use Yelp to Select Restaurant	Never	85	9.90%
	Occasionally	410	47.73%
	Frequently	279	32.48%
	All of the time	81	9.43%
History of Food Borne Illness	This has never happened to me	330	38.42%
	This happened to me once	389	45.29%
	This happened to me more than once	218	25.38%
Knowledge of Yelps posting of Public Health Information	Yes	112	13.04%
	No	743	86.50%

Table9. Sample Demographics Assessed through E-mail Survey of Berkley, University California San Francisco, and San Francisco State University email address holders.

Exclusion Criteria:

We collected 859 responses from students and faculty at San Francisco State University, University of California San Francisco, and Berkley. Of the 859, 3 were excluded due to illogical values and 3 were excluded due to values that were in the top 1% of reported values for "number of Yelp reviews desired before eating at a restaurant" Responses were excluded if they were illogical or if values were in the highest 1% of the distribution of values reported. Exclusions were in number of reviews desired to eat at a restaurant and public health score required to eat at a restaurant. If the desired public health score reported was greater than 100 it was excluded as an illogical value as the max score awarded is 100. Restaurant reviews at 10,000 and 10^{17} were also removed because they were in the top .01% of reported data. 1 response was excluded due to a conflict of reporting to being an occasional Yelp review writer despite reporting never using Yelp. After all exclusions were performed 850 complete responses remained.

Power Analysis:

Power analysis of the logistic regression model was conducted by measuring the ability of the model to correctly identify significant covariates. This analysis was conducted using a range of sample sizes from 10 to 800 participants. Odds ratios were created artificially and the power to correctly identify those odds ratios with p<.05 was measured over 100 simulation runs for each sample size. Odds ratios tested were 1.25, 1.5 1.75 and 2.0. Power curves were plotted to visualize the ability of the logistic regression model to accurately detect a given odds ratio at a given sample size. Given our sample size of 850 we are able to detect odds ratios of 1.5 and greater with 90% to 100% accuracy. Unfortunately our survey did not receive enough responses to be able to accurately detect odds ratios less than 1.5. In terms of comparative statistics such as a t-test or chi squared statistic assuming a .05% alpha value and a sample size of 850 we have the ability to detect a significant difference between groups with 85% accuracy. That is if there is a significant difference between two groups of similar size in our sample then we will have the power to detect that difference 85% of the time. We also performed power analysis in R using the R (pwr) library. Our analysis showed us that when conducting analysis of variance between groups we could partition our results into 8 groups while maintaining a 90% power to detect a significance of .05% assuming a medium size effect of .25. Unfortunately to be able to detect a smaller effect of .1 a sample size of 400 would be required. It is for this reason that age and racial groups were repartitioned from three age groups to two age groups, and from 5 racial groups to two racial groups white and non-white. A measure of the amount of effort that respondents were willing to exert to obtain health code violation information was also converted from 5 variables to two. This repartitioning of data allows us to identify smaller effect size at the cost of sacrificing the ability to generalize data to a greater number of groups. (Please see figure 2 for a plot of all simulated power curves)

Statistical Methods:

Logistic regression models were created to identify the effect of Yelp.com.com information on restaurant selection, public health information on restaurant selection, and the interaction of Yelp.com and public health information on restaurant selection. Models were adjusted for confounding variables like; gender, age, and race. Respondents' personal definition of an adequate health and Yelp.com ratings were used for model adjustment. This meant that respondents were asked the number of Yelp.com stars they required to eat at a given restaurant and they were also asked the public health rating they required to eat at a given restaurant. Cuisine preferences that are associated with higher rates of food borne illness (like Chinese food and Fast Food) [14]were also adjusted for. Rates were assessed using a San Francisco Department of Public Health dataset listing all health ratings and health code violations over a three period. Gender, race and age were also included as standard adjustments. Frequency dining out, frequency of Yelp.com usage, and personal experience with food borne illness were also adjusted for to minimize bias in our results.

In addition to measuring the effect of different types of information on the health behavior of restaurant selection we also wanted to see what variables were predictive of actually accessing the public health information presented on Yelp.com. Currently Yelp.com presents health code rating data alongside reviews, stars and other information aggregated from Yelp.com reviewers. However, specific health code violation on why a restaurant has a low health code rating, the number of violations and the number of violations over time are all available after clicking through two links. It is important to note that this information is only accessible by clicking on the score itself and then on an additional link next to a list of inspection dates. Yelp.com.com currently offers no guidance on the availability of this information. Indeed, over 90% of respondents (the majority of which are Yelp.com users reported that they did not know that public health data was available on Yelp.com despite the information being available for over two years.

Measurement of this relationship was performed by taking the vector of answers to the question: Would you take extra steps to find public health information on Yelp.com? "a. NO b. I would click a link to find health code violation information c. I would click to links to find health code violation information. d. I would click more than two links to find health code violation information." By setting these responses on a scale where "a" represents the least effort and d. the most effort we can use a log linear model to assess how respondent's eating habits, response to Yelp.com information, and response to public health information are associated with their interest in accessing public health data. This portion of our survey analysis attempts to measure if public health data is considered "valuable" by those who use Yelp.com,, with "value" being represented by the amount of effort a Yelp.com user would exert to access public health information. (Please see figure one for a plot of frequency of Yelp.com usage against the "value" placed on public health information.)

The point of a public health rating system, of health inspections and assignment of health code violations to restaurants that are deficient in their adherence to state health codes is to protect the public from infectious disease and or adulterants that may be transmitted through foods. In this survey we assessed an individual's food borne illness history by asking "have you ever had an experience of food borne illness after eating at a restaurant?" Afterwards, we set the response to this question as a binary variable and adjusted for the same cofounders as our other models and accounted for cuisine preference[22,23][24], and frequency dining out, if the respondent writes Yelp.com reviews the frequency at which they compose those reviews and finally the value they place on Yelp.com rating (stars) and public health rating (health score). This analysis allowed us to identify if placing greater value on a public health rating or placing greater value on a Yelp.com rating significantly modified odds of food borne illness after adjusting for confounding covariates. In this way we can identify if valuation of Yelp.com when finding a restaurant or valuation of public health data when identifying a restaurant increases the risk of food borne illness. A limitation of this method of analysis is that it does not take into account the temporality of events. It could be that the historical food borne illness experience happened before Yelp.com usage and valuation of Yelp.com data is not associated with food borne illness. However, given the fact that Yelp.com.com has been in existence for over a decade we held the belief that overlap of food borne illness history with Yelp.com usage was likely. Given there is a likely overlap between food borne illness recall and yelp usage we believe that inferring temporality is justified.

Results:

An initial analysis found that there is a significant difference in Yelp Star Preference across Age, Race, Gender, and reported food borne illness history.(See table 8) Foodborne illness history was reported as recalling one on more instances of food borne illness. The significant difference between those reporting food borne illness history is of particular interest because food borne illness is the very outcome that the department of Public health is focused on preventing in communication of health code rating and health code violation via Yelp. It is possible that the significance observed could be due to correlation with other significant variables like age. However when age was adjusted for in a logistic regression model preference for Yelp stars when selecting a restaurant were still a significant predictor of history of one of more instances of food borne illness. (See table 9 below) .

Using logistic regression analysis we found that after adjusting for confounding variables preference for restaurants with greater public health rating was negatively associated with the response of selecting restaurants with poor health rating given they have a perfect (5 star) Yelp.com rating. We also found that preference for restaurants with a greater public health rating were significantly associated with a "yes" response when asked the opposing question "Would you select a restaurant with perfect health rating and less than your desired number of stars? (Please see table2 and three for the corresponding confidence intervals for each covariate in our logistic regression model.) Thus as respondents place greater value on public health rating the odds of willingness to sacrifice Yelp stars rating for greater public health rating increases. With our opposing model we find the opposite is also true as the respondent places greater value on Public health rating the odds that a respondent is willing to sacrifice the desired number of stars for a perfect health rating increases. It is important to note that in both models the number of Yelp Stars preferred by respondents was a factor that was marginally significant. It could be that this study lacked the statistical power necessary to identify this relationship. Indeed, a power analysis revealed that the sample size of this study allowed for the identification of odds ratios at 1.25 or less only 60% of the time.

For our logistic regression model measuring the degree of effort users were willing to exert in accessing public health information we found that highest Yelp.com use was most predictive of wanting to exert effort to access public health data. The scale of exertion used was; I am not willing to click on any links to access public health information, I will click on one link to access public health information, I will click on 2 links to access public health information. Race, age, and gender, were adjusted for in this model along with Yelp.com usage, frequency dining out, and experience as a Yelp.com reviewer. This means as a person reported greater levels of Yelp.com usage the odds of resistance to applying greater effort to accessing data within Yelp.com decreased. (Please see table 4 for confidence intervals of all covariates and their significance in predicting odds of effort exerted to access public health data on ordinal scale).

We also found that when adjusting for race, age, and gender, and setting food borne illness as a binary response, the valuation of Yelp.com stars greater than 2 significantly increased the odds of food borne illness. That is as respondents placed greater importance on Yelp.com stars the odds of a respondent reporting experiencing food borne illness were significantly increased. Frequency dining out was also a significant predictor of reporting a history of food borne illness. It is important to note that Frequency dining out and valuation of Yelp.com stars are only negligibly correlated (.053). It is unlikely that increased valuation of stars may be an indirect measurement of frequency dining out. Additionally, since this covariate is included in the logistic regression model along with the covariate “frequency dining out” we can accept that the valuation of Yelp.com stars is able to explain variance in the data that frequency of dining out is not. This result is consistent with our other results in that we see a decreased valuation of public health data in those individuals placing high value on Yelp.com stars. It is plausible that those that place high value on Yelp.com information and are willing to ignore public health warnings are at greater risk for contracting food borne illness. (Please See Table 5 for Confidence intervals of all covariates predicting and adjusting for history of food borne illness) (Please See Figure 1 for the distribution of the rate of food borne illness across all star preference groups)

When the response of food borne illness history was reset to a binary response with a history of multiple food borne illnesses coded as 1 and less than multiple instances coded as 0.   Using the same variables that we used in our model that defined food borne Illness history as any history of food borne illness we found that Frequency Dining out, age, Gender, and preference for Chinese, and Vietnamese cuisine were significantly associated with a history of multiple food borne illness occurrences. Unfortunately this model lacked the power to detect significant association with Yelp Stars that was detected in our model using a liberal definition of food borne illness history. In our conservative model we found that Chinese restaurant preference gave respondents significantly lower odds (.47-.96) of multiple instances of food borne illness. While preference for Vietnamese restaurants significantly increased odds for multiple instances of food borne illness (1.31-2.81). The significance of dining frequency alone remains constant across liberal and conservative models of food borne illness. (.704-.9598 liberal model) and (.6777 and .9637 conservative model). It is important to note that there is little to no correlation between frequency dining out and Yelp Star preference, and Frequency dining out and Public Health rating Preference. The correlations are -.049, and .039 respectively. In this context we should not assume that frequency dining out is an indirect measure of respondent preference for Yelp Stars or preference for public health rating in terms of predicting odds of food borne illness history.

In this analysis preference for Chinese cuisine was negatively associated with multiple instances of food borne illness. It is possible that that those individuals that preferred Chinese food were disproportionately representative of other groups that appeared protected from odds of multiple instances of food borne illness. For instance it could be that Groups stating a preference for Chinese food were more likely to be female, young, and dine out with greater frequency than those not reporting a preference.

Table 10

Variable	Characteristic	Public Health(Mean /Standard Deviation)	t-test statistic p-value	Yelp Stars(Mean /Standard Deviation)	t-test statistic p-value	Yelp Reviews (Mean /Standard Deviation)	t-test Statistic p-value
Race	White(n=522)	82.49/19.79		2.92/1.21		33.9/72.82	0.000472
	Minority(n=335)	83.39/18.97	0.5117	2.64/1.35	0.001766	18.19/45.7	
Age	18-29 Years Old (n=136)	79.7/25.72		2.44/1.51		11.27/19.63	
	30-50 Years Old(n=390)	84.99/8.48		3.04/0.917		50.99/113.43	
	50+ Years Old(n=325)	85.09/15.1	0.02919*	2.90/1.203	0.000257	25.58/47.23	0.00157
Gender	Female(n=557)	84.25/17.5		2.91/1.2		26.02/63.05	
	Male(n=295)	80.96/21.75	0.02611	2.44/1.44	0.000003	21.2/48.38	0.2172
FBI HX	No History FBI(n=330)	81.34/22.49		2.55/1.39		25.33/71.75	
	History FBI(n=527)	84.08/16.96	0.05994	2.83/1.24	0.000614	23.76/48.25	0.7266

*Comparison of 18-29 year old and 30+ years Old

Table 10

Characteristic	2.5%	97.5%
Chinese	0.675	1.274
FastFood	0.794	2.369
Mexican	0.852	1.655
Indian	0.543	1.048
Vietnamese	0.881	1.753
Thai	0.612	1.251
French	0.791	1.596
Italian	0.681	1.299
18-29 Years Old	0.308	0.6826
Male	0.679	1.272
White	0.496	0.9279
Frequency Dining Out	0.704	0.9598
Preference for Stars	1.01	2.534
Value of Public Health	0.9964	1.012
Reviews	0.996	1.001
Using Yelp	0.8844	2.534

Table8.Odds Ratio 95% Confidence Interval for Odds of Food Borne Illness History.

Table 11

Characteristic	2.75%	97.50%
Chinese Cuisine Preference	0.47129239	0.9645665
FastFood Cuisine Preference	0.76795094	2.3193488
Mexican Cuisine Preference	0.90423468	1.9052474
Indian Cuisine Preference	0.49543256	1.0231888
Vietnamese Cuisine Preference	1.31070622	2.8156176
Thai Cuisine Preference	0.53630070	1.1912226
French Cuisine Preference	0.69923378	1.5033119
Italian Cuisine Preference	0.82948301	1.7072417
18-29 Years old	0.37082415	0.9888238
Male	1.01007146	1.99577
White Race	0.74080516	1.4694031
Frequency Dining Out	0.67778802	0.9637431
Number of Yelp Star Preference	0.85375418	1.122859
Using Yelp	0.12476381	3.9243021
Public Health Rating Preference	0.98697048	1.0208049
Review Number Preference	0.99624057	1.0023842

Table9.Odds Ratio 95% Confidence Interval for Odds of Multiple Instances of Food Borne Illness.

Table 12

Characteristic	2.5	97.5
18-29 Years Old	0	NA
Male	0.156	48.55
White	0.002	7.714
History of Food Borne Illness	0.002	7.1

Frequency Dining Out	**0.0392**	7.813
Preference for Stars	**0.0241**	1.507
Yelp Use	**2.105**	831.6
Public Health Rate Preference	**0.977**	1.507

Table10. Odds Ratio 95% Confidence Interval for Odds of Extra Exertion to Access Public Health Data on Yelp.com

Conclusion:

Currently San Francisco Health department officials have opted to submit LIVES formatted data to Yelp.com so that public health ratings may be displayed on the Yelp.com website. The goal of public health ratings displayed on storefronts is to inform consumers on public health risk and to also give store owners an economic incentive to improve their practices. There has been no evaluation of the effectiveness ofYelp.com public health rating display on restaurant selection until now. Yelp's primary function is to inform users on local services, restaurants, businesses etc. This is dissimilar to those sites that exist for the sole purpose of informing the public on restaurant public health risk. First it is necessary to point out that the display of Yelp.com information is particularly attractive with various colors, pictures of reviewers, and icons that the user may click upon to engage with the Yelp.com community. Public health ratings are on the right side of the restaurant page displayed in black and white with no demarcation of whether ratings are deficient or sufficient. Health code violation information can only be accessed by navigating through two hyperlinks and there is no way for users to know of the existence of these pages by reading content on connected pages.

When Cigarette packages were shown to consumers without colorful packaging it was found that the public health warnings had greater effect[25,26]. One option that may improve visualization of public health data would be to adopt graphic labels next to those restaurants with severe health infractions such as vermin, or unhygienic employees. This method of public health warning has been shown to improve recall when used on cigarette packaging[27]. Public health data and Yelp.com data on Yelp.com may be perceived in the same way as cigarette packaging, strategies that reduce the attractiveness of Yelp information or strategies that make public health data more graphic or engaging should be reviewed. While this study does provide evidence that Yelp.com data draws attention away from public health data posted on Yelp.com, it may be possible that this is a simple reflection of the main focus of Yelp.com users on restaurant experience. Further studies with experimental design assessing perception of public health data and Yelp.com data in the presence and absence of attractive coloring and icon illustrations could better define this relationship.

Despite the fact that 90% of respondents reporting using Yelp.com occasionally to every time they go out to eat, and one third reporting writing reviews for Yelp.com, and despite the reporting of public health data on Yelp(San Francisco) for over two years only 10% of respondents reported knowledge of the public health ratings being displayed in Yelp.com. At time of survey public health information had been displayed on the site for over two years. When the question was asked: “If a restaurant had a perfect health rating but less than the desired number of stars would you go?” We found that respondents placing high importance of Yelp.com stars had significantly reduced odds of wanting to go to such a restaurant after adjusting for confounders. It may be, those individuals that invest time in Yelp.com and have positive dining experiences using Yelp.com trust the information that Yelp.com displays. This is not to say that Yelp.com or other social media sites cannot be mined for useful health information or have a positive impact on restaurant selection. There are many recent articles that report upon the possible value of such data [28–30].

Given that few individual's report being aware of public health data on Yelp.com despite the program's existence in san Francisco for over two years, and even if individuals were aware of such information, those that use Yelp.com the most would appear to be the least likely to let public health data influence their restaurant selection decision. Our results with the logistic regression model that measured the odds of exerting extra effort to access public health data was surprisingly associated with more frequent Yelp.com use. An interpretation of these results could be that those that use Yelp.com the most are least likely to be bothered by clicking through extra links because frequent users don't mind spending time using the Yelp.com website. A limitation of this portion of the survey may be that it is not truly detecting the value that frequent Yelp.com users place on public health data. Instead this portion of the survey may be merely measuring frequent Yelp.com user's indifference to spending more time on Yelp.com. It is important to note that there are some situations where even Yelp.com users place high value on public health data. For instance if vermin were observed in a restaurant there were fewer than 5% of respondents that replied they would be willing to eat at a restaurant if it had a positive Yelp.com rating (5-stars) Respondents also stated that if employees were observed working without washing their hands then only 105 of respondents would be willing to go to such a restaurant if it had 5 stars. This implies that there is a limit to which Yelp.com users will place their trust in Yelp.com. It is important to note that data on hand washing and signs of vermin can be viewed on Yelp.com however this data is placed on an obscure area of the website and no guidance is provided by Yelp.com on where this information may be found. If public health officials want their data to be effective in changing patron health behavior then the information that has the greatest effect (health code violations) must be placed in an area that is highly visible to the public much like the storefront door. Instead, this influential data has been relegated to the darkest corners of Yelp.com and only 10% of respondents report any awareness of the existence of any public health data on Yelp.com.com let alone anything as specific as health code violations.

The significant finding that valuation of Yelp.com stars is predictive of food borne illness experience does have limitations. We do not know with certainty that the food borne illness event happened within the same time period as the respondent's Yelp.com usage. However, since Yelp has been in existence over a decade it is likely that recall of a food borne illness episode would happen within the same time period as Yelp usage particularly with respondents 18-29 years old. Another limitation may be that Yelp Star valuation is merely an indirect measure of dining frequency. However, since Yelp.com star valuation is negligibly correlated with frequent dining at (.048) it is unlikely that Yelp.com stars are an indirect measurement of frequent dining. This strengthens the possibility that the valuation of Yelp.com stars may reflect the willingness to ignore poor public health ratings. Indeed, the results of our two other logistic regression analyses (Probability of Under valuing Yelp Stars, and Probability of under valuing Public health rating)are consistent with this argument. In this context it would seem that there is a risk to the sacrifice, or willful ignorance to public health data. Although Yelp stars did not predict multiple instances of food borne illness history this could be due to the fact that multiple instances of food borne illness are strongly correlated with older age while valuation of Yelp Stars is higher in younger patients. Indeed when we repartition the age factor and designate 30-50 as young we still found that age was a strong predictor for history of multiple food borne illnesses, and there was no significance in Yelp Star valuation in modifying odds of multiple instances of food borne illness. It may be that this survey lacks the power to detect the effect of Star valuation on multiple instances of food borne illness. Further study of this relationship with a larger sample may be warranted to better define this relationship.

This does not mean that free text within Yelp.com cannot hold useful data regarding public health. The CDC provided a valuable proof of this concept in parsing Yelp.com data to identify new food borne illness outbreaks[30]. Unfortunately when it comes to regular Yelp.com usage the public are unable to parse out meaningful data that might supplement the information provided by public health officials. Findings in Chapter 2 of this thesis on the predictive power of reviews and tags mined from Yelp.com also validate that Social media on Yelp.com can be used for a meaningful public health purpose. It falls to public health officials and Yelp.com.com executives to reevaluate how public health data can be displayed alongside reviews in a way that supplements the user's ability to select restaurants that are palatable and safe.

It is clear that a great proportion (90%) of Yelp.com users do not know about the presence of public health data on Yelp.com. This lack of knowledge could be attributable to the poor visualization of public health data on Yelp.com when contrasted with the presentation of Yelp.com's own data. It could also be due to a disinterest in Public health Users among Yelp Users in General. When a t-test was conducted examining difference in valuation of Yelp Stars, Yelp Reviews, and Public Health Rating there was a significant difference between those respondents that reported using Yelp and those that reported never using Yelp. These findings may illustrate a key difference in the value placed on Public Health Data by Yelp Users.

These findings are generalizable to overall public perception of public health ratings/ warnings in large diverse municipalities like San Francisco. If the San Francisco health department wishes to have greater impact on the public's decision to dine at certain restaurants then it may be necessary to open new discussions with Yelp.com executives regarding how public health data is accessed and displayed. Maintaining the current deployment of information is not an effective use of public time and resources. If officials want their data to be viewed on Yelp.com then perhaps displaying data in a manner that attracts as much information as the Yelp.com data such as health code violation data, or presenting the health scores or health code violations in a way that supplements rather than competes with public health data may be an effective stratagem to modify the public health behavior of restaurant selection.

Chapter 4

Impact of Yelp.com LIVES Formatting on San Francisco Health Department

Health Code Violations and Public Health Ratings

Introduction:

Yelp.com has included public health data for restaurants on it's website for the past three years. Yelp.com refers to the embedding of this data on it's website as Local Inspector Value Entry Specification or "LIVES" formatting Despite this collaboration with the local public health department running for the last three years a survey of San Francisco residents found that the majority of San Franciscans have no knowledge of this information existing on the Yelp.com website. Additionally those respondents placing a high value on Yelp.com information (value placed on Yelp.com stars) had higher rates of reported food borne illness history. Also, those respondents that valued Yelp.com data had lower odds of valuing public health data. These results do little to strengthen the point of the posting of Restaurant scores; that negative health ratings will dissuade patrons from visiting restaurants that pose a greater health risk.

It is possible that Yelp.com LIVES formatting is still having a positive effect on the behavior of restaurant management and employees. It is possible that publicizing negative scores is associated with changes in the behavior of employees and management in that they would become adherent to the regulations placed on restaurants by the San Francisco Health department. If the Yelp LIVES formatting were having a positive effect on the behavior of vendor then a valid case could be made for public investment in its maintenance.

Previous longitudinal studies have assessed the effect of the Restaurant grade card system on Emergency Room visits (Los Angeles), and Health Rating Scores (New York) (Toronto). These studies all came to similar conclusions that the advent of restaurant grading had a positive impact on improving either rates of food borne illness admissions in emergency department or improvement of health ratings assigned by public health department officials. However, these studies were confounded by other aspects of public health enforcement (increased fines, increased frequency of visits, and having visitations assigned by a risk prediction model.). San Francisco is unique in that is one of the first municipalities to adopt Yelp LIVES formatting. This study would be the first to evaluate the utility of the LIVES formatting in its impact on vendor behavior. This study is also unique in that it not only assesses the effect of Yelp LIVES formatting on Health Rating but it also evaluates the effect on health code violations such as signs of vermin and employee hand washing.

‘’Yelp.com and the San Francisco Public Health Department have partnered together to better communicate public heath ratings and violations. It is unknown if this reporting system will have a significant impact on the behavior of food vending institutions. If this partnership is to continue then a validation of the current system is necessary to make the case for the continued investment of public funds in this partnership. The goal of this study is to measure the association of health code violations, health ratings and the amount of time that has passed since the adoption of yelp.com LIVES formatting.

Study Population

35,256 observation on 5947 restaurants in San Francisco were retrieved from over 3 years’ worth of health inspection data collected by the San Francisco department of public health. Restaurants that were on record as residing within a zip code outside of 94102-94158 1322 restaurants were excluded from geospatial analysis as only zip codes with corresponding zip code tabulated areas could be utilized by mapping software. Observations that did not include zip codes but did include San Francisco business IDs were included in statistical analysis conducted via linear mixed effect models. Of the 35256 observations retrieved 28541 were utilized for geospatial analysis while the complete dataset was used with linear mixed effect model analysis excluding only those observations with missing data points.

Power Analysis

This study was equipped to detect with 80% power up to a 12% and greater change in inspection score with 5% probability of falsely rejecting the null hypothesis. While this degree of power does mean that there is a 20% probability that our study may fail to identify a significant association, it does mean that if an association is identified then we can say with 95% confidence that the association was not reached by chance. Power Analysis was conducted using the longpower module in R specifically designed for conducting power analysis on linear mixed effect models.

Methods

In order to establish the impact of Yelp LIVES formatting on restaurant health ratings and specific health code violation frequency we created three linear mixed effect models predicting health code rating, frequency of vermin related violations, and frequency of hand washing violations. We also divided zip code tabulated areas (ZCTA) of San Francisco into Septiles representing tiers of health code ratings and then color coded the map of San Francisco ZCTAs according to color coded tiers. We performed the same data visualization method using counts of vermin related violations and hand washing related violations using the proportion of violations related to vermin of hand washing (separate measurements) on a 1-100 scale. Once each ZCTA had a percentage assigned to it the choropleth module in R was used to place the percentage into a distribution and divide that distribution into septiles or a seven partition percentile. Septiles were color coded with greater blue intensity to visualize the proportion of variable of interest displayed in each zip code tabulated area. Rating and Violation distributions were plotted over a three year period to visualize increase or decrease in violation frequency over time for specific zip code tabulated areas. These visualizations were compared with similar visualizations conducted on income and racial demographics obtained from the U.S. Census

2012. These visualizations were used to better understand possible relationships between ZCTA and health code ratings/violations.

Statistical Methods

One linear mixed model was created to predict the continuous response of the variable "health code rating" which may be any whole number from 1-100. The distribution of values in the Score variable did not meet the assumption of normal distribution when scores were plotted via histogram. Scores were squared and log transformed to better fit the assumption of a normal distribution. This linear mixed effect model incorporated fixed effects for whether the inspection was scheduled or unscheduled, and whether the inspection occurred in a high, mid, or low income ZCTA, and for vermin related violations a seasonal fixed effect was also placed. Random intercepts were assumed for Business ID, and year of inspection. Linear mixed effect models were used to make estimates of the probability of binary outcome of health code violations related to vermin or hand washing. Adjustments were made for ZCTA income levels in LME models as were racial demographic scores of each ZCTA, seasonal variation was also adjusted for in the vermin violation model. Confidence intervals were set at .05%.

Results

We used R and *lme4* to perform a linear mixed effects analysis of the relationship between Year after introduction of Yelp.com LIVES formatting and health code rating. As fixed effects, we entered Season, Severity of Violation, Income level, median rent, proportion Hispanic, proportion white, proportion black, proportion Hispanic , and population of the zip code tabulated area the restaurant resides in and whether the inspection was a scheduled inspection or not. As random effects, we had intercepts for Business IDs and the year that the inspection was performed. We also created random slopes for the effect of Season on Year. Visual inspection of residual plots did not reveal any obvious deviations from homoscedasticity or normality. P-values were obtained by likelihood ratio tests of the full model with the effect in question against the model without the effect in question. Severity of Violation and type of inspection were terms excluded from models predicting probability of health code violation for violations related to hand washing and violations related to vermin.

After adjusting for confounders we found that Income of the zip code tabulated area from which a restaurant resided was a significant predictor of Health Code rating, and Vermin Infestation. Linear mixed effect model found that those restaurants residing in the top half of Income reported for zip code tabulated areas had a two point higher score on average when compared to lower income zip code tabulated areas. We also found that low income strata increased the probability of vermin related citation by 3.2%. We were unable to identify any covariates that were significant predictors of employee hand washing citations. It could be that our study was not powered to detect the significance of these predictors. There was no detectable affect for years post LIVES formatting adoption. That is we could not identify any significant effect that could be attributed to the inclusion of public health rating data on yelp over the last three years. We did include an interaction term between years and season covariates. We found that once seasonal variance and interaction had been accounted for the fixed effect of year had no significant effect on health code rating or vermin citation. Reporting Style and format of statistical reports were adapted from examples in “Budowinter LME tutorial”. (Please See Tables 12, and 13 for analysis of covariates in models predicting the public health inspection outcomes of interest.) (Please see Figures 5-12 for maps depicting partitioning of health code rating and vermin citation by zip code tabulated area)(Please Note the Hyperlink to the Google motion chart depicting health code rating and vermin citation with possible stratification by Race, Ethnicity, Per Capita Income, Median Rent, and population for each of the zip code tabulated areas in San Francisco.)

Table 14

Random.effects:	Groups	Name	Variance	Std.Dev.	Corr	
	Businessid	(Intercept)	38.45	6.20		
	month	(Intercept)	0.19	0.44		
	Season	(Intercept)	37.19	6.10		
	year	0.00	0.00	-1.00		
	Residual	27.61	5.25			
Number.of.obs:	15749,	groups:	business_id,	4327;	Month.12	Season.4
Fixed.effects:						
	Estimate	Std.Error	df	tvalue	Pr(>|t|)	
(Intercept)	6145.00	1364.00	4344.00	4.505	0.000007	***
Season	283.70	102.40	4257.00	2.771	0.005619	**
typeRoutine.Scheduled	-2.91	7.14	12400.00	-0.408	0.683344	
typeRoutine.Unscheduled	5.72	4.19	12670.00	1.367	0.171783	
year	0.05	0.13	2726.00	0.377	0.706115	
INCOMETRUE	1.14	0.34	3983.00	3.319	0.000911	***
Zip	-0.07	0.01	4084.00	-4.568	0.000005	***
percent_white	-0.15	0.11	3965.00	-1.339	0.180725	
percent_black	-0.06	0.12	3971.00	-0.528	0.597369	
percent_asian	-0.16	0.11	3975.00	-1.395	0.163187	
percent_hispanic	-0.12	0.11	3969.00	-1.105	0.269173	
median_rent	0.00	0.00	3998.00	4.611	0.000004	***
Season:year	-0.14	0.05	4259.00	-2.774	0.005564	**

Linear Mixed Effect Model with Response of Health Code Rating

(Summary Statistics of significant predictors of health code rating using linear mixed effect model)

Table 15

	Estimate	Std.error	.05 CI	.95 CI	df	tvalue	Pr(>|t|)	
(Intercept)	57.73	247.30	542.44	-426.98	4439.00	0.233	0.81542	
log(BOIL)	0.81	0.33	1.46	0.16	11480.00	2.442	0.01463	*
Season	269.20	99.24	463.71	74.69	4111.00	2.713	0.0067	**
year	0.01	0.12	0.25	-0.23	4437.00	0.102	0.91871	
Season:year	-0.13	0.05	-0.04	-0.23	4110.00	-2.7	0.00696	**
log(BOIL):Season	-0.15	0.10	0.04	-0.34	12500.00	-1.559	0.1191	

Summary Statistics for linear mixed model predicting health code rating linearly regressed estimates combining Social and Racial covariates used to prevent colinearity in model and enhance parsimony.

Table 16

(Summary Statistics of General Estimating Equation Covariates)

Coefficients:							
	Estimate	NaiveS.E.	Naivez	RobustS.E.	0.05	0.95	Robustz
(Intercept)	40.0564	4.48E+01	0.893253	5.47E+01	-67.0938	147.2067	0.732715
Season	-45.6722	1.40E+01	-3.26147	1.96E+01	-84.1689	-7.1755	-2.32533
INCOMETRUE	-0.0102	5.94E-03	-1.71421	5.91E-03	-0.0218	0.0014	-1.7239
Zip	-0.0004	4.71E-04	-0.78747	5.74E-04	-0.0015	0.0008	-0.64648
year	-0.0025	3.96E-03	-0.63598	5.62E-03	-0.0135	0.0085	-0.44822
Season:Zip	0.0005	1.46E-04	3.261619	2.06E-04	0.0001	0.0009	2.314694
Season:year	0.0004	1.56E-03	0.281805	2.16E-03	-0.0038	0.0047	0.203602

(General Estimating Equation for Vermin Model)

Table 17

Fixed	Effects: Estimate	Std.Error	CI 0.05	CI 0.975	df	t-value	Pr(>|t|)	
(Intercept)	3.87E+01	4.45E+01	126.00	-48.56	4.21E+03	0.87	0.38461	
INCOMETRUE	-1.03E-02	5.77E-03	0.0010	-0.0216	4.55E+03	-1.785	0.07426	.
Season	-4.54E+01	1.42E+01	-17.62	-73.24	5.49E+03	-3.203	0.00137	**
Zip	-3.69E-04	4.67E-04	0.00	0.00	6.26E+03	-0.791	0.42921	
year	-1.93E-03	4.06E-03	0.01	-0.01	7.90E+01	-0.474	0.63665	

Season:year	2.90E-04	1.68E-03	0.00	0.00	1.70E+01	0.173	0.86484	
Season:Zip	4.77E-04	1.47E-04	0.0008	0.0002	2.70E+04	3.236	0.00121	**

(Summary Statistics for linear mixed models predicting vermin citation)

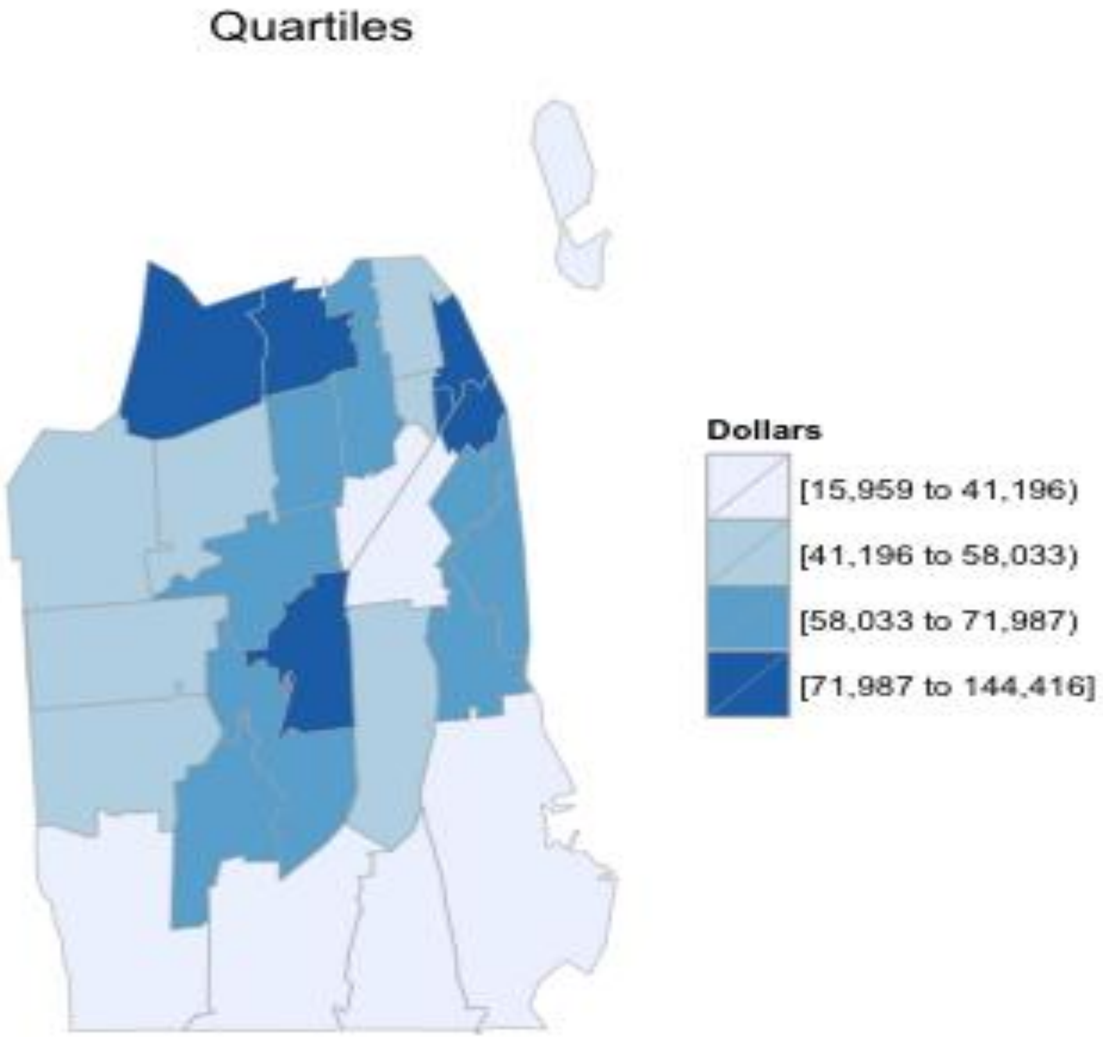


Figure 6

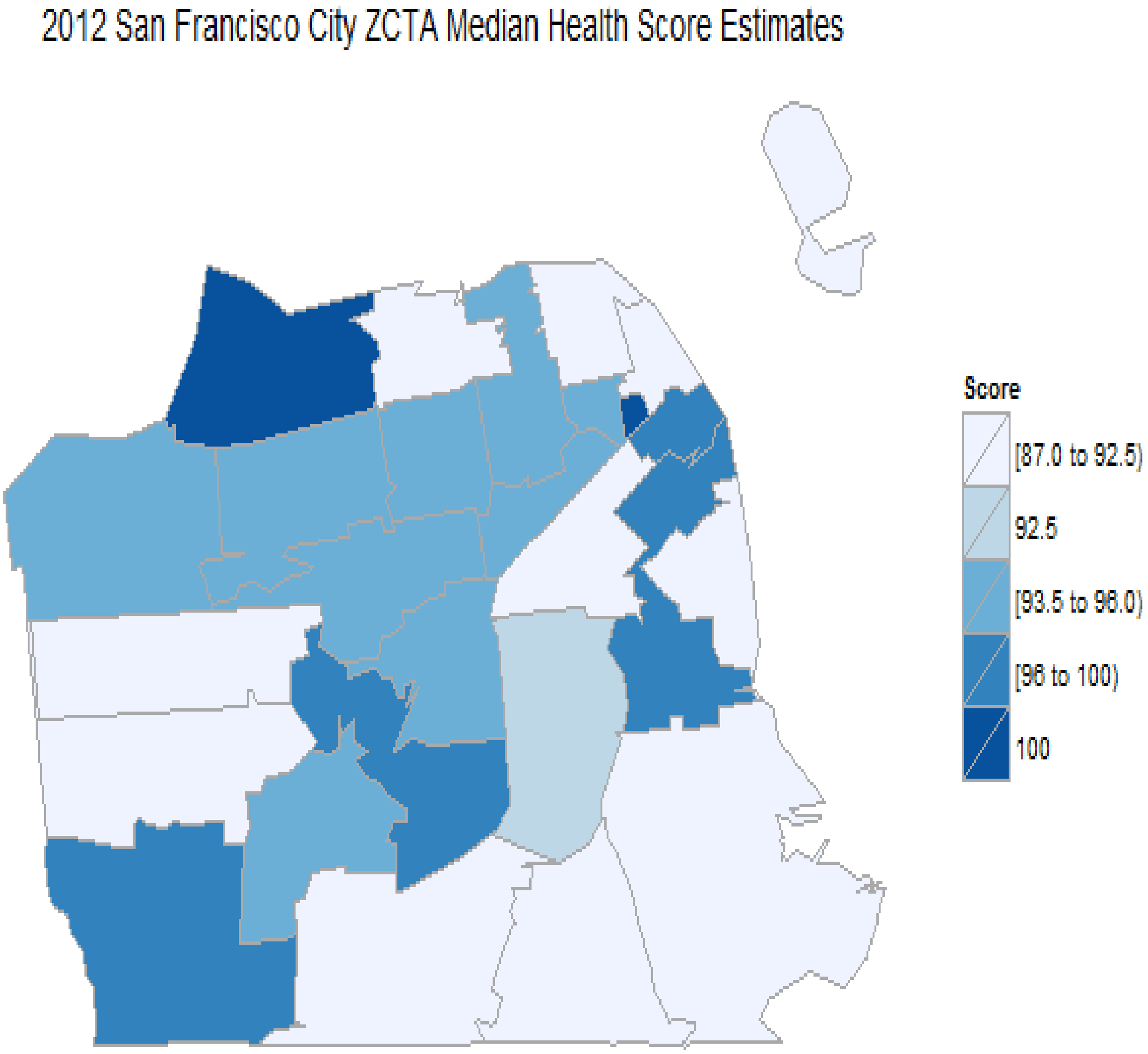

2012 San Francisco City ZCTA Median Health Score Estimates
Score
[87.0 to 92.5)
92.5
[93.5 to 96.0)
[96 to 100)
100

Figure 8

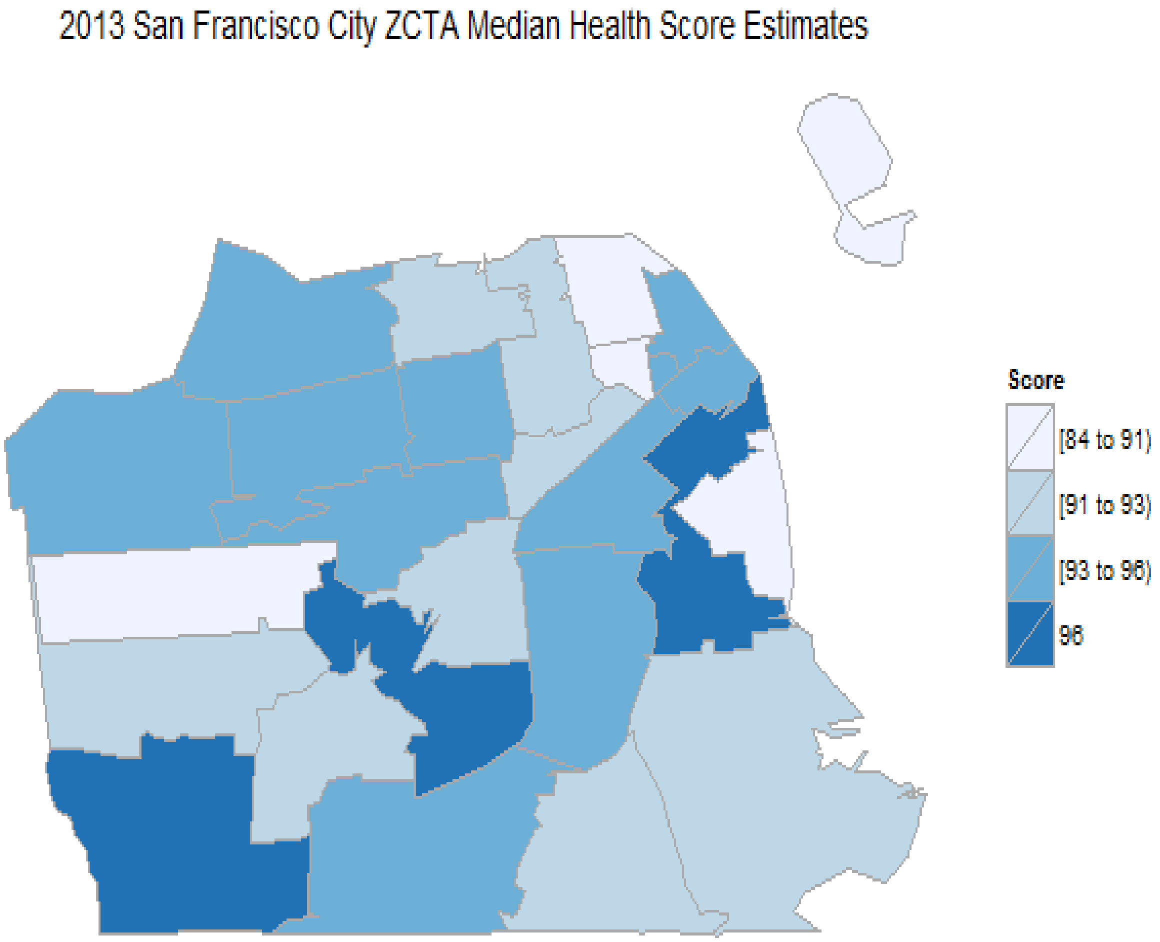

2013 San Francisco City ZCTA Median Health Score Estimates
Score
[84 to 91)
[91 to 93)
[93 to 96)
96

Figure 9

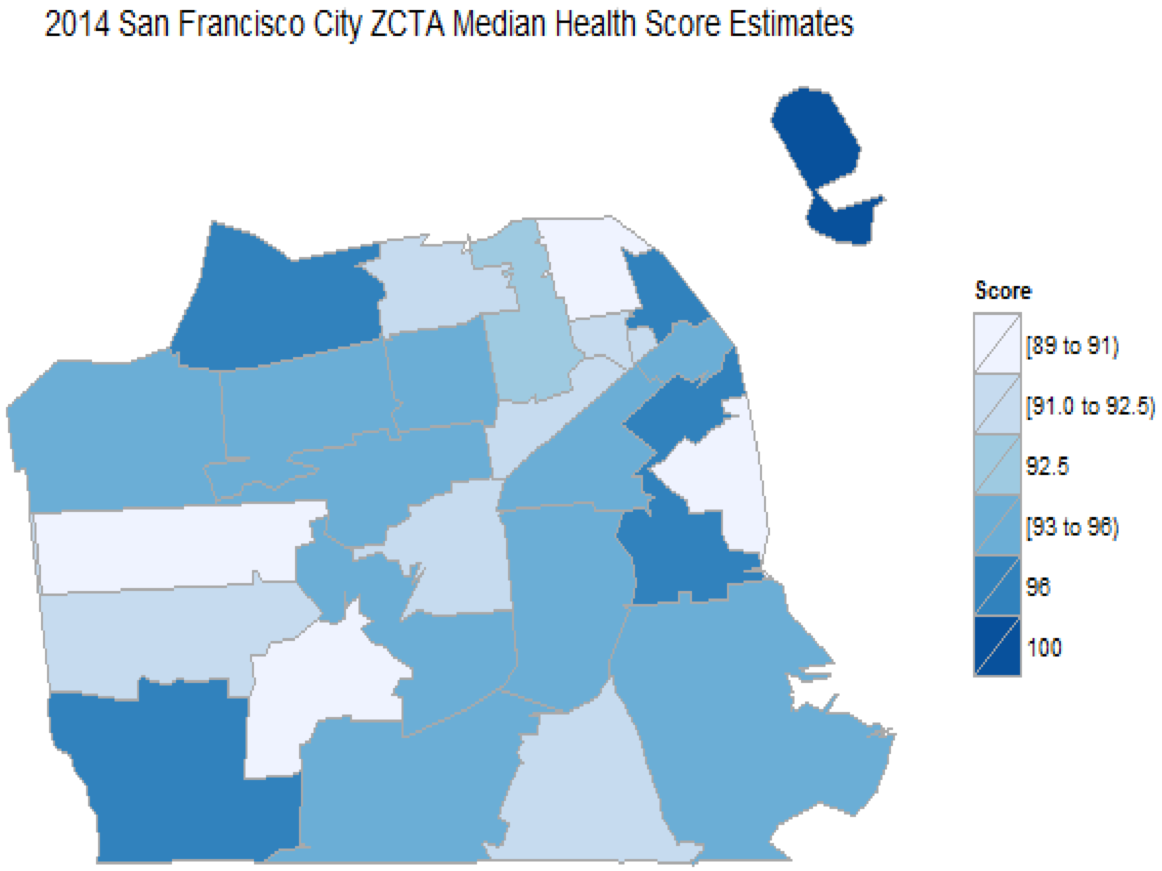

2014 San Francisco City ZCTA Median Health Score Estimates
Score
[89 to 91)
[91.0 to 92.5)
92.5
[93 to 96)
96
100

Figure 10

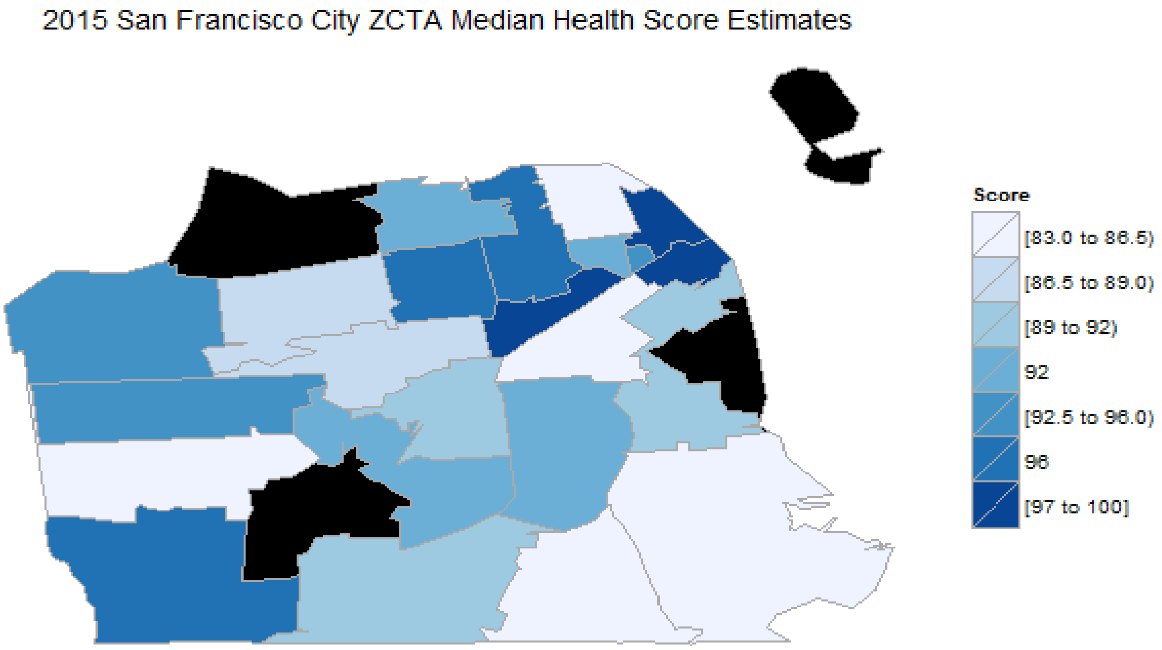


Figure 11

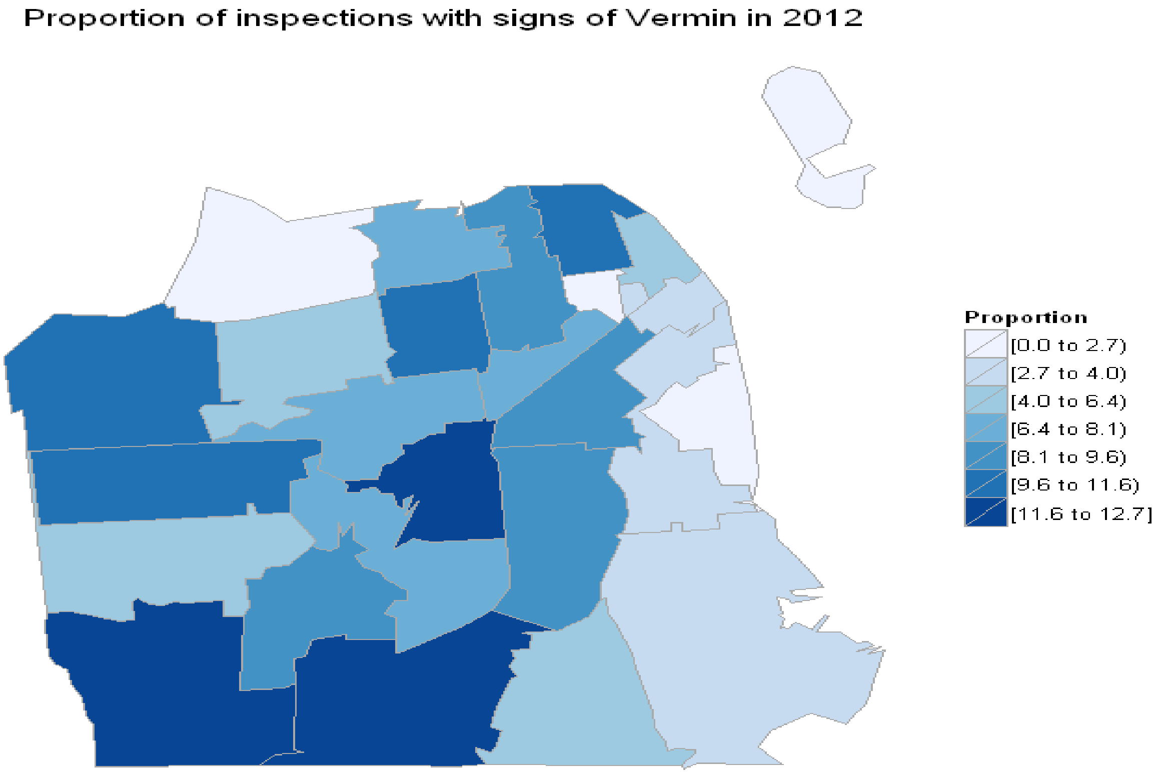
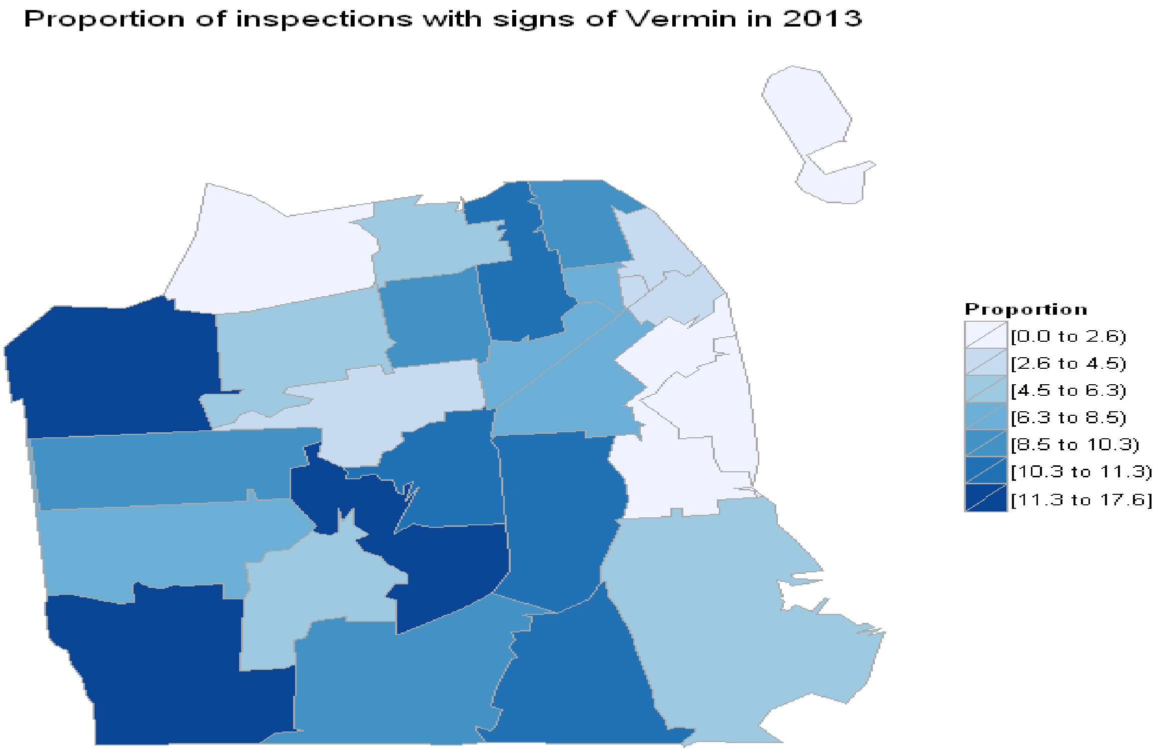
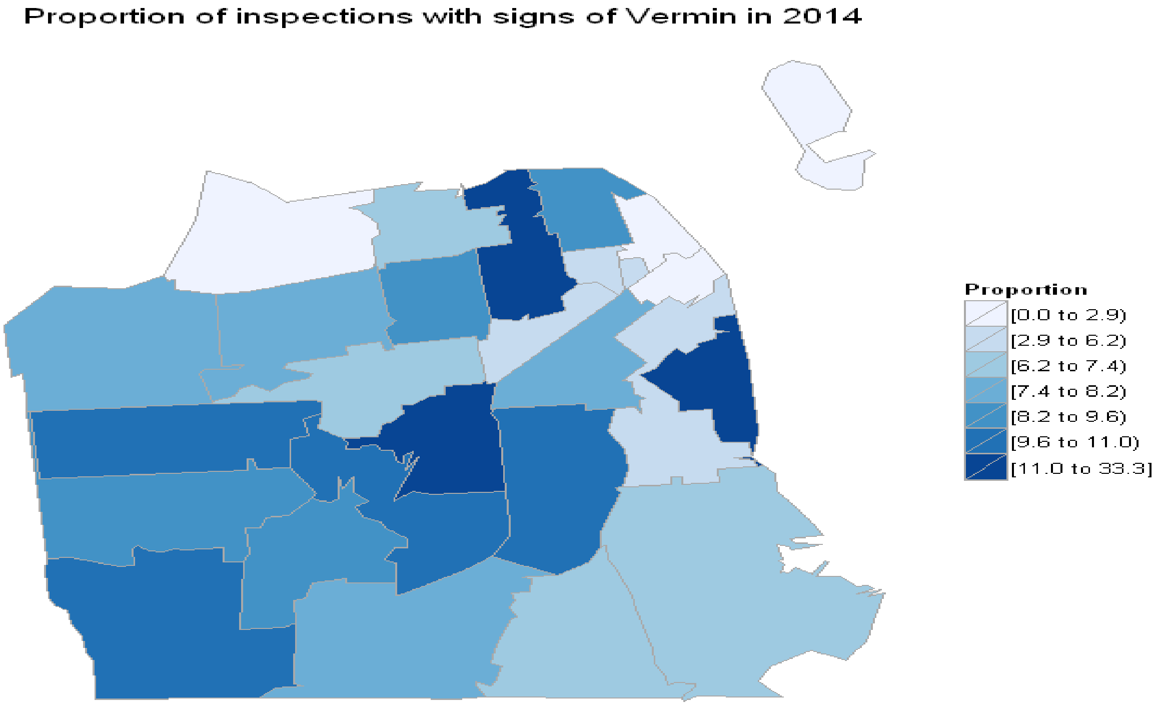


Figure 12



Proportion of inspections with signs of Vermin in 2013

Figure 13



Proportion of inspections with signs of Vermin in 2014

Figure 14

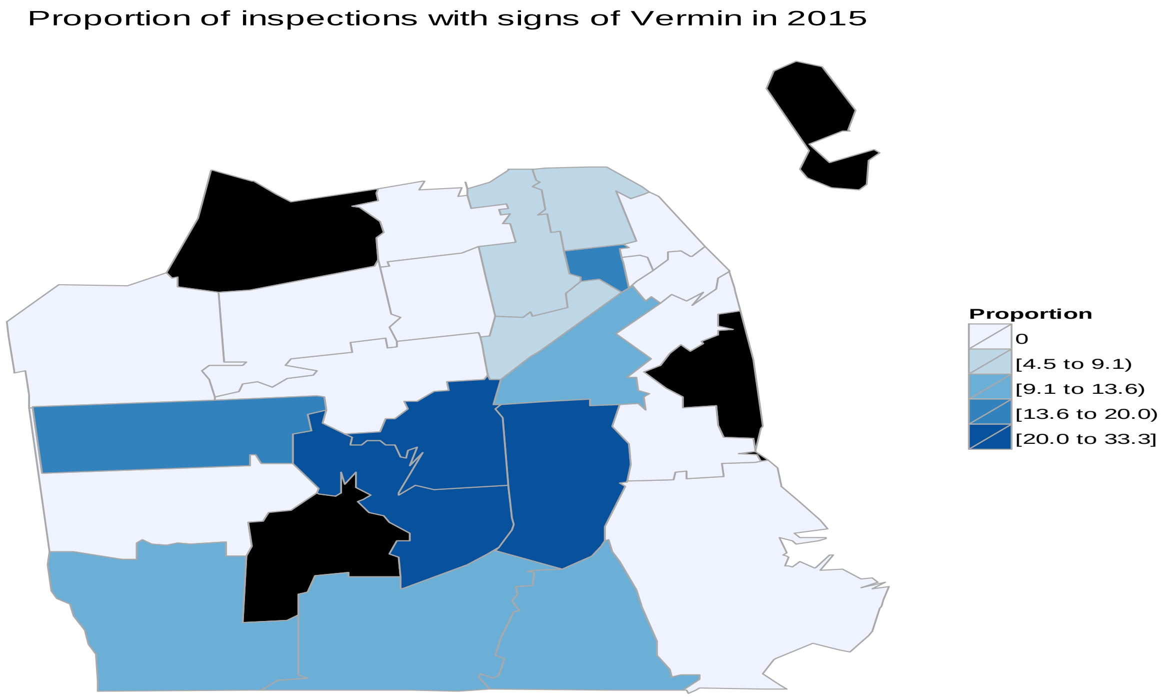


Figure 15

GOOGLE VIS MOTION CHART

https://docs.google.com/spreadsheets/d/1L8oA6HyD3KFMD_8tHi7Y3xcCTX__nZ3t1_zVsjxyDF0/pubchart?oid=1953588741&format=interactive

Figure 16

Discussion

Linear Mixed effect model analysis of San Francisco Department of Public Health Inspection data was unable to identify any significant effect of time variables (formatted in years) in explaining the variance in health scores, vermin citations, or employee hygiene citations after adjusting for socio economic and racial covariates of the zip code tabulated areas in which each restaurant resides. The single most predictive variable in our analysis was income of the zip code tabulated area in which a restaurant resides. Graphical analysis of zip code tabulated areas produced the same results. There was a large concordance between the septiles of income strata and septiles of health code rating. The three years that have passed since Health code ratings and health code violations were first posted on the website Yelp.com are not significantly associated with an increase in health code ratings, vermin violations or employee hygiene (hand washing violations. These results along with the previous analysis in chapter 3 suggest that the current system of communicating public health inspection findings via Yelp.com has not been effective in altering vendor behavior or in swaying the public behavior of high risk restaurant selection.

Additionally we have found that the problem of low health code rating is largely correlated with the socio economic status of the zip code tabulated area in which a restaurant resides. In addition to the seasonal variance in health code violations and low health code ratings which of course are correlated. Indeed, in areas where income is at the highest percentile we see that health code ratings rarely if ever fluctuate below the standard of 80. We have also found that those lower socio economic areas are more likely to Asian or Hispanic which poses the question of public health department's ability to communicate results in a language that is intelligible to local shop owners, and if it is possible that other language based biases may contribute to disproportionately lower scores in these ethnic enclaves.

Analysis of vermin violations also revealed that lower socio economic status of a zip code tabulated area is significantly predictive of an increased probability of restaurant vermin citation. It is unknown what favorable circumstances may exist that attract vermin in lower income zip code tabulated areas. It may be affordable or effective extermination services are difficult for lower income restaurant owners, it is also possible that higher income restaurants provide more food safety training to their employees which creates an environment less inviting to vermin infestation. We can also imagine that the type of foods served in lower cost establishments may somehow be more attractive to vermin and thus induce higher rate of infestation. These investigations are however beyond the scope of this research and would warrant further study in the future.

Hand washing citations occurred with greater frequency in zip code tabulated areas where Income was below the median for San Francisco. However the effect of socio economic status on employee hand washing is not as great as that observed on vermin citations. While an effect may exist, it could be that our study was not adequately powered to detect the size of the effect.

Unfortunately, this study had several limitations. The nature of this study was primarily ecological and so we can only draw inferences from the associations observed during our ecologic analysis. While we can observe a relationship between food borne illness risk factors and zip code tabulated areas with lower socio economic status our study is not designed to identify what underlying issues may be creating these disparities. Further investigation is required to determine if our results can be explained by racial inequality, linguistic barriers, or perhaps some other type of inspection bias.

This study did benefit from a fairly robust sample size in that over 5000 restaurants that had inspection data over a three year period. Our power analysis confirmed that our study would be able to detect a 12% or greater effect on health code rating using the sample described and conducting analysis using a linear mixed effect model. The time available for analysis was similar to time periods used for similar studies that supported letter grade programs in New York City and Los Angeles. Our Results were unable to show any association with time since introduction of Yelp LIVES formatting on Restaurant Health Scores, Employee Handwashing citations, or Vermin Citations after adjusting for confounders.

Conclusion

Based on our analysis we can provide no evidence validating a positive effect of the Yelp LIVES formatting system on restaurant health code rating, nor can we find any association between the period of time after LIVES formatting was introduced and reduction in rates of Vermin citations or employee hygiene (hand washing) citations. Rather we found that Key significant drivers were the socioeconomic status of the zip code tabulated area in which a restaurant resided. If public Health institutions wish to make meaningful interventions that result in positive impacts on vendor behavior then it would appear germane to address barriers to guideline compliance encountered by low income restaurant owners. Creating interventions that target the challenges of low income restaurant owners would be a good first step to reducing the rates of the food borne illness risk factors.

Chapter 5

Conclusion

This thesis has identified how Yelp social media can be mined to identify and rank restaurants at high risk of transmission of food borne illness using data previously unavailable to public health officials. This is also the first analysis of yelp free text data that examines the increased odds of health code violation associated with specific keywords in Yelp.com free text reviews. This discovery highlights that this surveillance approach can be utilized by public health departments to identify high risk restaurants that may not be identified using traditional surveillance techniques. Additionally in contrast with the CDC study examining the ability of Yelp data to identify new food borne illness outbreaks this study does not rely upon an additional workforce to run surveillance and analysis of yelp data. Indeed, we validate that analysis of Yelp.com data can be conducted programmatically and analysis results can be achieved daily in a span of hours. Opposed to less nimble surveillance methodologies reported in the New York Study conducted by the CDC. Furthermore the surveillance strategy outlined in the methods of this thesis are ideally suited for prevention of epidemics while the New York Public Health Department Study performed in conjunction with the CDC is focused upon identification of epidemics and root cause analysis of sources. Much like a tornado siren our system is able to create early warning at the first sign of high risk food borne illness keywords reported via Yelp. Free text reviews on Yelp.com can be parsed for meaningful information that is predictive of health code violation. Our analysis shows that the algorithm that orders Yelp reviews is biased in its selection of reviews related to food borne illness. This bias may be limited to San Francisco Yelp Data, but the analysis of this thesis shows that this bias can be exploited to identify the reviews that are most predictive of food borne illness. We were able to validate that this bias exists by plotting the improved AUC in simulated datasets that consisted of increasing proportions of “first page” reviews. The benefit of this finding is that it would considerably reduce the time needed to extract and analyze review data. In this way we are able to leverage the Yelp.com crowd and the value they place on reviews related to food borne illness. Additionally our survey of Yelp preferences did not identify any significant difference in public health information across user cuisine preference. An interpretation of these results would be that crowd interest in food borne illness would be equal across all cuisine types.

Social Media on Yelp.com may misinform public health behavior of restaurant selection, at the same time this social media data may be mined for meaningful information regarding the public health risk a restaurant presents. The key difference between public perception and our proposed methods of surveillance is the use of analytical tools that the general public does not have available. We use Python, and R programming languages to access and analyze the data available in Yelp. If public health officials were interested in allowing public access to data similar to that described in chapter two, such as the presence of high risk keywords in reviews or the frequency of high risk keywords, then the public health department could provide access to this data by making a browser extension or plugin that gave access to this data or by dealing directly with Yelp programmers to modify the presentation of Yelp data to include such points. Another approach would require that a caption be posted below each review stating that comments and statements made by Yelp.com or individual reviewers should not be interpreted as public health information. It would also be feasible for a lone programmer with sufficient skills to create a program that would automate a Yelp Review account for the sole purpose of creating Yelp reviews that contained data pertaining to Specific Health code Violations and Health Code Ratings. This merging of Yelp and Public Health Data could provide both public health and Yelp management valuable insight on user interest in public health data. Public Interest could be easily tracked using Yelp's User popularity Statistics. This is of course assuming Yelp.com would allow automated Yelping for public good of presenting public health data in a more informative manner. Such endeavors are beyond the scope of this thesis but could serve to further validate the principles outlined in chapter 3.

The fact that the majority of Survey respondents reported no knowledge of Yelps offering of public health information despite the existence of this partnership for over two years gives strong evidence that the current display of such data has been ineffective. This evidence is strengthened in light of results presented in chapter 3 that showed that across all questions related to restaurant selection Public health rating and Yelp stars were the strongest determining factors for selection of a restaurant. Additionally using a simple paired t-test we found that there was a significant difference in the number of stars preferred when selecting a restaurant after stratifying by age and food borne illness history. This relationship was only able to be identified in young white respondents as this was the largest racial subgroup in our survey. Unfortunately we did not have the statistical power to detect a relationship in minority groups responding to our survey. Additionally, we found that in addition to age and Race the number of stars preferred was significantly associated with reported food borne illness history. A possible conclusion is that respondents that value Yelp stars are less likely to value public health ratings, this may be because Yelp users are more likely to adopt a "foodie" lifestyle where they enjoy eating at a variety of dining establishments regardless of public health rating. A possible intervention would be to target "foodies" or those yelp users or Yelpers that dine out frequently by including public health messages disclaimers or warnings within Yelp reviews.

Yelp.com is a social media website with great potential to communicate public health messages to large groups of people. Specific data on Yelp.com can be influential in guiding customers' decision in selecting a restaurant. Unfortunately this guiding force is value placed on Yelp stars which is associated with personal devaluation of public health information and personal reports of food borne illness history. If public health officials wish to harness the power of Yelp to communicate with the public then they must modify their message to more closely emulate the content found on Yelp.com which we have shown is associated with modifying public health behavior of restaurant selection.

Our studies also show that the time period since the introduction of Yelp LIVES formatting also has little significance in the prediction of health code rating and citations. Unlike similar studies we are unable to show that Yelp public health scores are significant predictors of health code compliance. Unfortunately we are able to show that a significant amount of variance in health code rating can be accounted for by the socio economic status of neighborhood surrounding a given restaurant. These findings send a message to SFDPH officials that Socio economic status of neighborhoods surrounding communities is an important matter to consider when designing effective interventions. Yelp officials may also consider different ways they may want to enhance YELP LIVE formatting so it may be adopted by and engage a greater number of municipalities.

Social Media data can only be refined into meaningful information when thoughtful analytical tools are employed. When the public consumes this data outside the context of thoughtful analysis we see that there is the possibility that misinformation leading to negative health behavior (high risk restaurant selection) may occur. Public Health partnerships with companies profiting from social media collection and publication must either include clauses that warn the public when data should not be consumed as public health data or alternative data displays must be adopted to avoid misinterpretation or under value of public health information by social media users. Public officials would also benefit from improved surveillance if thoughtful analysis of such social media data were adopted. The size of the crowd participating in social media in San Francisco may be larger than other parts of America and therefore findings may not yet be exportable to all American municipalities. However, as we look toward the future and see that internet accessibility, affordability, and mobility are ever increasing and urban centers are becoming more populous then we shall expect the exportability of this to thesis expand and become applicable to the public health officials of tomorrow.

REFERENCES Chapter 2

1. Foodborne Diseases Active Surveillance Network FoodNet 2012 Surveillance Report. 2012.

2. Changes in incidence of laboratory-con f i rmed bacterial infections , US , 2013. 2013;(April 2014):2014.

3. States U, Diseases ZI, Diseases E. CDC Estimates of Foodborne Illness in the United States CDC 2011 Estimates. 2011;3–4.

4. Department of Public Health [Internet]. [cited 2014 Oct 9]. Available from: http://www.sfdph.org/dph/EH/Food/Inspections.asp

5. Restaurant inspection findings and violations in San Francisco 100-And (Afc Sushi @ Safeway, A.G. Ferrari Foods, Amc-Level, Afc Sushi @ Cala, Alcatraz Landing Cafe, ...). Food safety and Restaurant Scores [Internet]. [cited 2014 Oct 9]. Available from: http://www.city-data.com/sf-restaurants/index1.html

6. Restaurants enter 5th year of real sales growth | National Restaurant Association [Internet]. [cited 2014 Oct 9]. Available from: http://www.restaurant.org/News-Research/News/Restaurant-industry-enters-fifth-year-of-real-sale

7. Buchholz U, Run G, Kool JL, Fielding J, Mascola L. A risk-based restaurant inspection system in Los Angeles County. J Food Prot. 2002;65:367–72.

8. Todd ECD, Greig JD, Bartleson CA, Michaels BS. Outbreaks where food workers have been implicated in the spread of foodborne disease. Part 5. Sources of contamination and pathogen excretion from infected persons. J Food Prot. 2008;71:2582–95.

9. Petran RL, White BW, Hedberg CW. Using a Theoretical Predictive Tool for the Analysis of Recent Health Department Inspections at Outbreak Restaurants and Relation of This Information to Foodborne Illness Likelihood. Journal of Food Protection. 2012. p. 2016–27.

10. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009;457:1012–4.

11. Walcott BP, Nahed B V, Kahle KT, Redjal N, Coumans J-V. Determination of geographic variance in stroke prevalence using Internet search engine analytics. Neurosurg Focus. 2011;30:E19.

12. Polgreen PM, Chen Y, Pennock DM, Nelson FD. Using internet searches for influenza surveillance. Clin Infect Dis. 2008;47:1443–8.

13. Kang M, Zhong H, He J, Rutherford S, Yang F. Using Google Trends for Influenza Surveillance in South China. PLoS One. 2013;8.

14. Hulth A, Andersson Y, Hedlund KO, Andersson M. Eye-opening approach to norovirus surveillance. Emerging Infectious Diseases. 2010. p. 1319–21.

15. Dukic VM, David MZ, Lauderdale DS. Internet queries and methicillin-resistant staphylococcus aureus surveillance. Emerg Infect Dis. 2011;17:1068–70.

16. Desai R, Hall AJ, Lopman BA, Shimshoni Y, Rennick M, Efron N, et al. Norovirus disease surveillance using google internet query share data. Clin Infect Dis. 2012;55.

17. Cho S, Sohn CH, Jo MW, Shin SY, Lee JH, Ryoo SM, et al. Correlation between national influenza surveillance data and Google Trends in South Korea. PLoS One. 2013;8.

18. Chan EH, Sahai V, Conrad C, Brownstein JS. Using web search query data to monitor dengue epidemics: A new model for neglected tropical disease surveillance. PLoS Negl Trop Dis. 2011;5.

19. Carneiro HA, Mylonakis E. Google trends: a web-based tool for real-time surveillance of disease outbreaks. Clin Infect Dis. 2009;49:1557–64.

20. S. W, M. N. Use of internet search data to provide real time data on kidney stone disease in the United States [Internet]. Journal of Urology. 2011. p. e898–e899. Available from: http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed10&NEWS=N&AN=70378824

21. Althouse BM, Ng YY, Cummings DAT. Prediction of dengue incidence using search query surveillance. PLoS Negl Trop Dis. 2011;5.

22. Lindh J, Magnusson M, Grünewald M, Hulth A. Head Lice Surveillance on a Deregulated OTC-Sales Market: A Study Using Web Query Data. PLoS One. 2012;7.

23. Seifter A, Schwarzwalder A, Geis K, Aucott J. The utility of “Google Trends” for epidemiological research: Lyme disease as an example. Geospat Health. 2010;4:135–7.

24. Ayers JW, Althouse BM, Allem JP, Rosenquist JN, Ford DE. Seasonality in seeking mental health information on Google. Am J Prev Med. 2013;44:520–5.

25. Brownstein JS, Freifeld CC, Madoff LC. Digital disease detection--harnessing the Web for public health surveillance. N Engl J Med. 2009;360:2153–5, 2157.

26. Generous N, Fairchild G, Deshpande A, Del Valle SY, Priedhorsky R. Global disease monitoring and forecasting with Wikipedia. PLoS Comput Biol [Internet]. 2014 Nov 13 [cited 2015 Apr 24];10(11):e1003892. Available from: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003892

27. San Francisco Restaurants, Dentists, Bars, Beauty Salons, Doctors [Internet]. [cited 2014 Oct 9]. Available from: http://www.yelp.com/

28. Harrison C, Jorder M, Stern H, Stavinsky F, Reddy V, Hanson H. Using Online Reviews by Restaurant Patrons to Identify Unreported Cases of Foodborne Illness — New York City , 2012 – 2013. 2014;63(20):2012–3.

29. Local Inspector Value-Entry Specification | Yelp [Internet]. [cited 2014 Oct 9]. Available from: http://www.yelp.com/healthscores

30. Venables WN, Smith DM. An Introduction to R. 2014;1.

31. Strand MA. Package " rmetasim ."2014;

32. ROCR: Classifier Visualization in R [Internet]. [cited 2014 Oct 9]. Available from: http://rocr.bioinf.mpi-sb.mpg.de/

33. San Francisco Department of Public Health website home page [Internet]. [cited 2014 Oct 9]. Available from: http://www.sfdph.org/dph/default.asp

References Chapter 3:

1. Jin GZ, Leslie P. The Effect Of Information On Product Quality: Evidence From Restaurant Hygiene Grade Cards. Q J Econ [Internet]. 2003;118:409–51. Available from: http://www.mitpressjournals.org/doi/abs/10.1162/003355303321675428

2. Lee J-E, Nelson DC, Almanza BA. The Impact of Individual Health Inspectors on the Results of Restaurant Sanitation Inspections: Empirical Evidence. Journal of Hospitality Marketing & Management. 2010. p. 326–39.

3. Petran RL, White BW, Hedberg CW. Health Department Inspection Criteria More Likely To Be Associated with Outbreak Restaurants in Minnesota. Journal of Food Protection. 2012. p. 2007–15.

4. Jones TF, Pavlin BI, LaFleur BJ, Ingram LA, Schaffner W. Restaurant Inspection Scores and Foodborne Disease. Emerg Infect Dis. 2004;10:688–92.

5. Impact of Restaurant Hygiene Grade Cards on Foodborne-Disease Hospitalizations in Los Angeles County. [cited 2015 Mar 21]; Available from: http://kuafu.umd.edu/~ginger/research/JEH-final.pdf

6. Restaurant Grading in New York City at 18 months [Internet]. [cited 2015 Apr 2]. Available from: http://www.nyc.gov/html/doh/downloads/pdf/rii/restaurant-grading-18-month-report.pdf

7. Wong MR, McKelvey W, Ito K, Schiff C, Jacobson JB, Kass D. Impact of a letter-grade program on restaurant sanitary conditions and diner behavior in New York City. Am J Public Health [Internet]. American Public Health Association; 2015 Mar 9 [cited 2015 Apr 4];105(3):e81–7. Available from: http://ajph.aphapublications.org/doi/abs/10.2105/AJPH.2014.302404?url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org&rfr_dat=cr_pub%3Dpubmed

8. Waters a B, VanDerslice J, Porucznik C a, Kim J, DeLegge R, Durrant L. Impact of internet posting of restaurant inspection scores on critical violations. J Environ Health [Internet]. 2013;75:8–12. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23858661

9. San Francisco Restaurants, Dentists, Bars, Beauty Salons, Doctors [Internet]. [cited 2014 Oct 9]. Available from: http://www.yelp.com/

10. Local Inspector Value-Entry Specification | Yelp [Internet]. [cited 2014 Oct 9]. Available from: http://www.yelp.com/healthscores

11. Ho DE. Fudging the nudge: Information disclosure and restaurant grading. Yale Law J. 2012;122:574–688.

12. Realmuto L. State health agency workforce shortages. J Environ Health [Internet]. 2014 Mar [cited 2015 Apr 4];76(7):70. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24683942

13. Zablotsky Kufel JS, Resnick BA, Fox MA, McGready J, Yager JP, Burke TA. The Impact of Local Environmental Health Capacity on Foodborne Illness Morbidity in Maryland. Am J Public Health [Internet]. 2011 Aug [cited 2015 Apr 4];101(8):1495–500. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3134516&tool=pmcentrez&rendertype=abstract

14. Restaurant inspection findings and violations in San Francisco 100-And (Afc Sushi @ Safeway, A.G. Ferrari Foods, Amc-Level, Afc Sushi @ Cala, Alcatraz Landing Cafe, ...). Food safety and Restaurant Scores [Internet]. [cited 2014 Oct 9]. Available from: http://www.city-data.com/sf-restaurants/index1.html

15. Department of Public Health [Internet]. [cited 2014 Oct 9]. Available from: http://www.sfdph.org/dph/EH/Food/Inspections.asp

16. San Francisco Department of Public Health website home page [Internet]. [cited 2014 Oct 9]. Available from: http://www.sfdph.org/dph/default.asp

17. California Retail Food Code [Internet]. [cited 2015 Apr 2]. Available from: http://www.cdph.ca.gov/services/Documents/fdbRFC.pdf

18. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)-A metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009;42:377–81.

19. Welcome to Python.org [Internet]. [cited 2014 Oct 9]. Available from: https://www.python.org/

20. San Francisco County QuickFacts from the US Census Bureau [Internet]. [cited 2015 Apr 2]. Available from: http://quickfacts.census.gov/qfd/states/06/06075.html

21. Yelp.com Traffic and Demographic Statistics by Quantcast [Internet]. [cited 2015 Apr 2]. Available from: https://www.quantcast.com/yelp.com?country=US

22. Nawa Y, Hatz C, Blum J. Sushi Delights and Parasites: The Risk of Fishborne and Foodborne Parasitic Zoonoses in Asia. Clin Infect Dis [Internet]. 2005 Nov 1 [cited 2015 Apr 4];41(9):1297–303. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16206105

23. GORMLEY FJ, RAWAL N, LITTLE CL. Choose your menu wisely: cuisine-associated food-poisoning risks in restaurants in England and Wales. Epidemiol Infect [Internet]. 2011 Aug 19 [cited 2015 Apr 4];140(06):997–1007. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21854669

24. Liu P, Kwon J. The exploration of effects of Chinese cultural values on the attitudes and behaviors of Chinese restaurateurs toward food safety training. J Environ Health [Internet]. 2013 Jun [cited 2015 Apr 4];75(10):38–46. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23858664

25. Shankleman M, Sykes C, Mandeville KL, Di Costa S, Yarrow K. Standardised (plain) cigarette packaging increases attention to both text-based and graphical health warnings: experimental evidence. Public Health [Internet]. 2015 Jan [cited 2015 Apr 4];129(1):37–42. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4315810&tool=pmcentrez&rendertype=abstract

26. Agaku IT, Filippidis FT, Vardavas CI. Effectiveness of text versus pictorial health warning labels and predictors of support for plain packaging of tobacco products within the European Union. Eur Addict Res [Internet]. 2015 Jan [cited 2015 Mar 17];21(1):47–52. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25402440

27. Wang A-L, Lowen SB, Romer D, Giorno M, Langleben DD. Emotional reaction facilitates the brain and behavioural impact of graphic cigarette warning labels in smokers. Tob Control [Internet]. 2015 Jan 6 [cited 2015 Apr 4]; Available from: http://www.ncbi.nlm.nih.gov/pubmed/25564288

28. Harrison C, Jorder M, Stern H, Stavinsky F, Reddy V, Hanson H. Using Online Reviews by Restaurant Patrons to Identify Unreported Cases of Foodborne Illness — New York City , 2012 – 2013. 2014;63(20):2012–3.

29. Nsoesie EO, Kluberg SA, Brownstein JS. Online reports of foodborne illness capture foods implicated in official foodborne outbreak reports. Prev Med (Baltim) [Internet]. 2014 Oct [cited 2015 Mar 2];67:264–9. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4167574&tool=pmcentrez&rendertype=abstract

30. Health Department Use of Social Media to Identify Foodborne Illness — Chicago, Illinois, 2013–2014 [Internet]. [cited 2015 Apr 4]. Available from: http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6332a1.htm

Appendix A

Python Code for Web Extraction of Social Media Data

```
import csv

import urllib2

import mechanize

import cookielib

import re

# Browser


br = mechanize.Browser()


br.set_handle_robots(False)

br.set_handle_equiv(True)

br.set_handle_gzip(True)

br.set_handle_redirect(True)

br.set_handle_referer(True)


# Cookie Jar

cj = cookielib.LWPCookieJar()

br.set_cookiejar(cj)




br.addheaders = [('User-agent', 'Mozilla/5.0 (X11; U; Linux i686; en-US; rv:1.9.0.1) Gecko/2008071615
Fedora/3.0.1-1.fc9 Firefox/3.0.1')]

urls = csv.reader(open('C:\Python27\SFMEGA.csv'))

hits_out=open('C:\Python27\SFDATA2.txt','wb')
```

```
url_list = []
mywriter = csv.writer(hits_out,delimiter=',')
n =0

for url in urls:
    try:
        print (url)

        response = br.open(url[0])
        page = response.readlines()

        #mo = re.search('<p itemprop="description" lang="en"> (.*) </p>', page)

        data=[]
        hits=[]
        HCR=[]
        RC=[]
        Name=[]
        Raddress=[]
        star=[]
        Dollar=[]
        Date=[]
        Review=[]

        for line in page:
```

```
if'<span itemprop="streetAddress">' in line:
   Raddressline=line
   Raddress.append(line)



if'<meta itemprop="datePublished" content="' in line:
   Dateline=line
   Date.append(line)

if'<span itemprop="reviewCount">' in line:
   RCline=line
   RC.append(line)
if'out of 100' in line:
   HCRline=line
   HCR.append(line)

if '<span class="business-attribute price-range" itemprop="priceRange">' in line:
   Dollarline=line
   Dollar.append(line)




if'<p itemprop="description" lang="en">' in line:
    Reviewline=line
```

```
       Review.append(line)
    if'star rating" c' in line:
      starline=line
      star.append(line)
  mywriter.writerow([Raddress]+[HCR]+[Date]+[Dollar]+[RC]+[star]+[Review])
except Exception:
  continue
```

Appendix B

R Code used for Statistical Analysis

```
library(ROCR)


#want to check mean and sd at different levels of sample size?

#validate appropriate power to identify prevalence

#Identify prevalence of given sample at different simulated sample sizes

#Show distribution of ROC Curves at different sample sizes and thresholds

#Null No difference in Predictive power at different levels of usefulness?.



#Read in 1st half Partial 2nd Half and Full 2nd Half of Yelp Dataset



Yelp<-read.csv("C:\\Python27\\Yelp Simulation.csv") #1st Half of Dataset

Yelp2<-read.csv("C:\\Python27\\Yelp Simulation2.5.csv") #2nd Half of Dataset




Yelp3<-read.csv("C:\\Python27\\SFDATA3.csv") #Combined Dataset


YelpS<-read.csv("C:\\Python27\\Yelp Simulation_STAR.csv")

Yelp2S<-read.csv("C:\\Python27\\Yelp Simulation2_STAR.csv")
```

```
Yelp3S<-read.csv("C:\\Python27\\SFDATA3_STAR.csv")


#Create Subsets using only most useful reviews
Yelp$Response[Yelp$Response==0]=2
Yelp$Response[Yelp$Response==1]=0
Yelp$Response[Yelp$Response==2]=1
YelpS$Response[Yelp2$Response==0]=2
YelpS$Response[Yelp2$Response==1]=0
YelpS$Response[YelpS$Response==2]=1

YelpS$response

YELPS<- YelpS[ which(YelpS$Usefulness<='2'),]
head(YELP)
YELP2S<- Yelp2S[ which(Yelp2S$Usefulness<='2'),]
head(YELP2)
YELP3S<- Yelp3S[ which(Yelp3S$Usefulness<='2'),]
head(YELP3)
dim(YELP3)
dim(Yelp3S)
#look at data
cor(Yelp3S$HCR, use="all.obs", method="pearson" )
YCorr <- cor(t(as.numeric(unlist(Yelp3S[1:10]))),method="spearman")
cor(Yelp3S$Response, Yelp3S$dollar)
```

```
summary(Yelp)

library(rpart)


head(yelp)


#Make Models referencing different subsets


YM <- glm( Response ~ Stars+RC+dollar+ilove+oldschool+pushy+pool+affordable+Christ+stench+employees+humid+septic+Jesus+hell+dishes+thebest+highquality+adorable+fabulous+craving+favorite+excellent+service+recommend+professional+delicious+washhands+burnt+ache+pain+cigarette+asshole+awful+rotten+bathroom+toilet+puke+fuck+microwaved+shit+bitch+sucks+mold+mice+spider+exclaim+filthy+roach+DIRTY+Ifounda+clean+diarrhea+vomiting+foodpoisoning+dirty+truck+sick+stomach+hospital+fish+nausea+terrible+horrible, data=Yelp, family=binomial,na.action = na.exclude )

YM2<-glm( Response ~ Usefulness+RC+dollar+modKW+tagmod,data=Yelp, family=binomial,na.action = na.exclude )

YM3 <- glm( Response ~ Usefulness+RC+dollar+ilove+oldschool+pushy+pool+affordable+Christ+stench+employees+humid+septic+Jesus+hell+dishes+thebest+highquality+adorable+fabulous+craving+favorite+excellent+service+recommend+professional+delicious+washhands+burnt+ache+pain+cigarette+asshole+awful+rotten+bathroom+toilet+puke+fuck+microwaved+shit+bitch+sucks+mold+mice+spider+exclaim+filthy+roach+DIRTY+Ifounda+clean+diarrhea+vomiting+foodpoisoning+dirty+truck+sick+stomach+hospital+fish+nausea+terrible+horrible, data=Yelp3, family=binomial,na.action = na.exclude )

YMS <- glm( Response ~ dollar+RC+Star+ilove+oldschool+pushy+pool+affordable+Christ+stench+employees+humid+septic+Jesus+hell+dishes+thebest+highquality+adorable+fabulous+craving+favorite+excellent+service+recommend+professional+delicious+washhands+burnt+ache+pain+cigarette+asshole+awful+rotten+bathroom+toilet+puke+fuck+microwaved+shit+bitch+sucks+mold+mice+spider+exclaim+filthy+roach+DIRTY+Ifounda+clean+diarrhea+vomiting+foodpoisoning+dirty+truck+sick+stomach+hospital+fish+nausea+terrible+horrible, data=YelpS, family=binomial,na.action = na.exclude )

YM2S<-glm( YelpS$Response ~ Usefulness+RC+dollar+modKW+Star+tagmod,data=YelpS, family=binomial,na.action = na.exclude )

YM3S <- glm( Response ~dollar+STAR+ilove+oldschool+pushy+pool+affordable+Christ+stench+employees+humid+septic+Jesus+
hell+dishes+thebest+highquality+adorable+fabulous+craving+favorite+excellent+service+recommend+professional+delicious+washhands+burnt+ache+pain+cigarette+asshole+awful+rotten+bathroom+toilet+puke+fuck+microwaved+shit+bitch+sucks+mold+mice+spider+exclaim+filthy+roach+DIRTY+Ifounda+clean+diarrhea+vomiting+foodpoisoning+dirty+truck+sick+stomach+hospital+fish+nausea+terrible+horrible, data=YELP3S, family=binomial,na.action = na.exclude )

Test<-glm(Response ~ RC+dollar+STAR+foodpoisoning+affordable+pushy, data=Yelp3S, family=binomial,na.action=na.exclude)

length(YELP3S$Response=1)

library(rpart)

YR=rpart(HCR>80~STAR+Usefulness+dollar+ilove+oldschool+pushy+pool+affordable+Christ+stench+employees+humid+septic+Jesus+hell+dishes+thebest+highquality+adorable+fabulous+craving+favorite+excellent+service+recommend+professional+delicious+washhands+burnt+ache+pain+cigarette+asshole+awful+rotten+bathroom+toilet+puke+fuck+microwaved+shit+bitch+sucks+mold+mice+spider+exclaim+filthy+roach+DIRTY+Ifounda+clean+diarrhea+vomiting+foodpoisoning+dirty+truck+sick+stomach+hospital+fish+nausea+terrible+horrible, data=Yelp3S)


par(mar = rep(1, 7))

plot(YR, uniform = TRUE, branch = .6, compress = FALSE, margin = 0.1,main="Analysis of Yelp Data via Classification Tree")

text(YR, all = TRUE, use.n = FALSE, fancy = TRUE, cex= 0.9)

?rpart

?rpartplot

library(rpart.plot)

?rpart.plot

data(ptitanic)

Yelptree=rpart(HCR>80 ~STAR+dollar+ilove+oldschool+pushy+pool+affordable+Christ+stench+employees+humid+septic+Jesus+hell+dishes+thebest+highquality+adorable+fabulous+craving+favorite+excellent+service+recommend+professional+delicious+washhands+burnt+ache+pain+cigarette+asshole+awful+rotten+bathroom+toilet+puke+fuck+microwaved+shit+bitch+sucks+mold+mice+spider+exclaim+filthy+roach+DIRTY+Ifounda+clean+diarrhea+vomiting+foodpoisoning+dirty+truck+sick+stomach+hospital+fish+nausea+terrible+horrible, data=Yelp3S,cp=.02)
```

```
                    # cp=.02 because want small tree for demo


rpart.plot(YR, main="Yelp Data Classification Tree",  under = TRUE,type=1,extra=1, faclen=0)
text(YR, all = TRUE, use.n = FALSE, fancy = TRUE, cex= 0.9)




summary(Test)
exp(Test$coef)
anova(YM,YMS)
anova(YM2S)
anova(YMS)
#Look at models
summary(YM)
summary(YMS)
exp(YM$coef)
exp(confint(YM))
summary(YM2)
summary(YM3)
exp(YMS$coef))
exp(YMS$coef)
summary(YM2S)
exp(confint(YM2S))
exp(YM2S$coef)
```

```
summary(YM3S)
exp(confint(YMS))
exp(YM3S$coef)
exp(confint(YM3S))
Yelp3S$Star


##Now Check predictive power without looping ie simulated Using data with Stars data:
# Fitted probabilities:
#I decide on YMS first
probs = fitted(YMS)
# Proportion in sample with low HCR
mean(YELPS$HCR<80,na.rm=TRUE)
# Use 0.264 as cutoff to calculate sensitivity and specificity:
# pred prob > 0.26 -> 1; <= 0.26 -> 0
preds = as.numeric(probs>0.33625)
table(YELPS$Response,preds)
#look at sens spec
library(ROCR)
pred.obj = prediction(probs,YELP3S$Response)
plot(performance(pred.obj,"tpr","fpr"),main="ROC CURVE San Francisco Yelp Dataset 1ST HALF")
plot(performance(pred.obj,"ppv"),main="PPV San Francisco Yelp Dataset 1st HALF",ylim=c(0,1))
performance(pred.obj,"auc")
PPV=(performance(pred.obj,"ppv")@y.values)
```

```
mean(as.numeric(unlist(PPV)),na.rm=TRUE)


##Now Check predictive power without looping Using data without Stars data on first half:
# Fitted probabilities:

probs = fitted(YM)
# Proportion in sample with low HCR
mean(Yelp$HCR<80)
# Use 0.325 as cutoff to calculate sensitivity and specificity:
# pred prob > 0.325 -> 1; <= 0.325 -> 0
preds = as.numeric(probs>0.2678)
table(Yelp$Response,preds)
#look at sens spec
library(ROCR)
pred.obj = prediction(probs,Yelp$Response)
plot(performance(pred.obj,"tpr","fpr"),main="ROC CURVE San Francisco Yelp Validation Dataset")
legend("bottomright",legend=c( paste("AUC = .70"),
paste("Sensitivity=.65"),
paste("Specificity=.56")))

plot(performance(pred.obj,"ppv"),main="PPV San Francisco Yelp Validation Dataset")
performance(pred.obj,"auc")
legend("bottomright",legend=c( paste("PPV = .517")))
```

```
performance(pred.obj,"spec")
SENS=performance(pred.obj,"sens")@y.values
mean(as.numeric(unlist(SENS)))
SPEC=performance(pred.obj,"spec")@y.values
mean(as.numeric(unlist(SPEC)))

PPV=(performance(pred.obj,"ppv")@y.values)
mean(as.numeric(unlist(PPV)),na.rm=TRUE)



##Now Check predictive power without looping ie simulated Using data with Stars data on second half:
# Fitted probabilities:
#I decide on YMS first
probs = fitted(YMS)
# Proportion in sample with low HCR
mean(YELPS$HCR<80)
# Use 0.325 as cutoff to calculate sensitivity and specificity:
# pred prob > 0.325 -> 1; <= 0.325 -> 0
preds = as.numeric(probs>0.3253)
table(YelpS$Response,preds)
#look at sens spec
```

```
library(ROCR)
pred.obj = prediction(probs,YelpS$Response)
plot(performance(pred.obj,"tpr","fpr"),main="ROC CURVE San Francisco Yelp Reviews First Dataset")
legend("topright",legend=c( paste("AUC = .785"),
paste("Sensitivity=.719"),
paste("Specificity=.444")))


plot(performance(pred.obj,"ppv"),main="PPV San Francisco Yelp Dataset Complete")
performance(pred.obj,"auc")
PPV=(performance(pred.obj,"ppv")@y.values)
mean(as.numeric(unlist(PPV)),na.rm=TRUE)



##Now Check predictive power without looping ie simulated Using data with Stars data on second half:
# Fitted probabilities:
#I decide on YM3S first
probs = fitted(YM3S)
length(probs)
# Proportion in sample with low HCR
mean(Yelp3S$HCR<80,na.rm=TRUE)
# Use 0.3488 as cutoff to calculate sensitivity and specificity:
# pred prob > 0.325 -> 1; <= 0.325 -> 0
```

```
preds = as.numeric(probs>0.25)

length(preds)

table(YELP3$Response,preds)

length(YELP3$S$Response)

#look at sens spec

library(ROCR)

pred.obj = prediction(probs,YELP3$Response)

plot(performance(pred.obj,"tpr","fpr"),main="ROC CURVE San Francisco Yelp Dataset Complete")

performance(pred.obj,"auc")

plot(performance(pred.obj,"ppv"),main="PPV San Francisco Yelp Dataset Complete")

PPV=(performance(pred.obj,"ppv")@y.values)

mean(as.numeric(unlist(PPV)),na.rm=TRUE)
```

```
# Fitted probabilities:

#I decide on Test first

probs = fitted(YM)

# Proportion in sample with low HCR

mean(Yelp$HCR<80)

# Use 0.264 as cutoff to calculate sensitivity and specificity:

# pred prob > 0.26 -> 1; <= 0.26 -> 0

preds = as.numeric(probs>0.267)
```

```
table(Yelp$Response,preds)
#look at sens spec
library(ROCR)
pred.obj = prediction(probs,Yelp$Response)
plot(performance(pred.obj,"tpr","fpr"),main="ROC CURVE San Francisco Yelp Dataset Complete")
plot(performance(pred.obj,"ppv"),main="PPV San Francisco Yelp Dataset Complete")
performance(pred.obj,"auc")
PPV=(performance(pred.obj,"ppv")@y.values)
mean(as.numeric(unlist(PPV)),na.rm=TRUE)



?sample
?performance
#Start simulation using same analysis of dataset using variables calculated within excel

for (i in 1:1000) { #i is the counter variable – the exact name doesn’t matter.  The { opens the loop

bottle=Yelp[sample(nrow(Yelp), 260,replace=TRUE), ]


water=sum(bottle$Sensitivity)
oil=sum(bottle$Specificity)
wine=sum(bottle$Response)
FP=169-oil
```

```
FN=91-water
sens[i]=water/wine
spec[i]=oil/(260-wine)
ppv[i]=water/(water+FP)
sens
spec
FP
FN
ppv
} #closes the loop.  The counter i does not need to be incremented (i.e., it is not necessary to say i=i+1)
#emptyvar[i]<-somefunction #assigns the value to position or column i in the empty variable
length(Yelp2)

#Use for second dataset
for (n in 1:1000) {#i is the counter variable – the exact name doesn’t matter.  The { opens the loop

bottle2=Yelp2[sample(nrow(Yelp2),70,replace=FALSE), ]


water2=sum(bottle2$Sensitivity)
oil2=sum(bottle2$Specificity)
wine2=sum(bottle2$Response)
sens2[n]=water2/wine2
spec2[n]=oil2/(260-wine2)
sens2
spec2


}
mean(sens2)
mean(spec2)
hist(sens2, type="l", xlab="i")
hist(spec2, type="l", xlab="i")





plot(spec, type="l", xlab="i")
SENS=sens>.7
SPEC=spec>.65
sum(SENS)
sum(SPEC)
#744/1000
#487/1000
?rnorm

sd(sens)
mean(sens)
sd(spec)
mean(spec)
```

```
#Use same method with complete second half of dataset

for (n in 1:1000) { #i is the counter variable – the exact name doesn’t matter.  The { opens the loop

bottle3=Yelp3[sample(nrow(Yelp3), 40,replace=TRUE), ]


water3=sum(bottle3$Sensitivity)
oil3=sum(bottle3$Specificity)
wine3=sum(bottle3$Response)
sens3=water3/wine3
spec3=oil3/(260-wine3)
sens3
spec3

}
hist(sens3, type="l", xlab="i")
hist(sens3, type="l", xlab="i")
###########################################################################################################
#TEST MODEL PREDICTIVE POWER USING ROCR
#This approach will use logistic regression model will repeat  for each model using same code just changing out dataset name.

library(ROCR)
```

```
for (n in 1:3000) {#i is the counter variable – the exact name doesn't matter.  The { opens the loop


bottleM=YelpS[sample(nrow(YelpS),200,replace=FALSE), ]

YMM <- glm( Response ~
RC+dollar+ilove+oldschool+pushy+pool+affordable+Christ+stench+employees+humid+septic+Jesus+hell
+dishes+thebest+highquality+adorable+fabulous+craving+favorite+excellent+service+recommend+prof
essional+delicious+washhands+burnt+ache+pain+cigarette+asshole+awful+rotten+bathroom+toilet+puk
e+fuck+microwaved+shit+bitch+sucks+mold+mice+spider+exclaim+filthy+roach+DIRTY+Ifounda+clean+
diarrhea+vomiting+foodpoisoning+dirty+truck+sick+stomach+hospital+fish+nausea+terrible+horrible,
data=bottleM, family=binomial,na.action = na.exclude )

probs = fitted(YMM)

Yelpmean=mean(YMM$HCR<80)

preds = as.numeric(probs>Yelpmean)

pred.obj = prediction(probs,bottleM$Response)

sensi=performance(pred.obj,"tpr")

speci=performance(pred.obj,"fpr")

#accu=performance(pred.obj,"acc")

#AUC<-performance(pred.obj,"auc")@y.values

AUC[n]<-performance(pred.obj,"auc")@y.values


ppv=performance(pred.obj,"ppv")@y.values

#PPV=mean(as.numeric(unlist(ppv)),na.rm = TRUE)

PPV[n]=mean(as.numeric(unlist(ppv)),na.rm = TRUE)


}

AUC

auc=mean(as.numeric(unlist(AUC)),na.rm=TRUE)

auc
```

```
plotauc=(as.numeric(unlist(AUC)))

plot(plotauc)

hist(plotauc,main="Distribution of AUC Across First Half of Dataset")

sdauc=sd(as.numeric(unlist(AUC)),na.rm=TRUE)

sdauc

mean(PPV)

sd(PPV)

plot(PPV)

hist(PPV)

mean(auc)

mean(as.numeric(auc))

median(as.numeric(auc))

sd(as.numeric(auc))

ppv=performance(pred.obj,"ppv")@y.values

PPVmean=mean(as.numeric(unlist(ppv)),na.rm = TRUE)

plot(PPV)

sd(as.numeric(unlist(ppv)),na.rm = TRUE)

plot(as.numeric(unlist(ppv)),na.rm = TRUE)

hist(as.numeric(unlist(ppv)),na.rm = TRUE)


#Plots will show us distribution of PPV and AUC

#####################################################################################
##################

#TEST MODEL PREDICTIVE POWER USING ROCR Test on Complete Dataset

#This approach will use logistic regression model will repeat  for each model using same code just
changing out dataset name.
```

```
library(ROCR)
for (n in 1:10000) {#i is the counter variable – the exact name doesn’t matter.  The { opens the loop


bottleM=Yelp3S[sample(nrow(Yelp3S),200,replace=TRUE), ]
YMM <- glm( Response ~
RC+dollar+ilove+oldschool+pushy+pool+affordable+Christ+stench+employees+humid+septic+Jesus+hell
+dishes+thebest+highquality+adorable+fabulous+craving+favorite+excellent+service+recommend+prof
essional+delicious+washhands+burnt+ache+pain+cigarette+asshole+awful+rotten+bathroom+toilet+puk
e+fuck+microwaved+shit+bitch+sucks+mold+mice+spider+exclaim+filthy+roach+DIRTY+Ifounda+clean+
diarrhea+vomiting+foodpoisoning+dirty+truck+sick+stomach+hospital+fish+nausea+terrible+horrible,
data=bottleM, family=binomial,na.action = na.exclude )
probs = fitted(YMM)
Yelpmean=mean(YMM$HCR<80)
preds = as.numeric(probs>Yelpmean)
pred.obj = prediction(probs,bottleM$Response)
SENSI[n]=performance(pred.obj,"tpr")@y.values
SPECI[n]=performance(pred.obj,"fpr")@y.values
#accu=performance(pred.obj,"acc")
#AUC<-performance(pred.obj,"auc")@y.values
AUC[n]<-performance(pred.obj,"auc")@y.values
ppv=performance(pred.obj,"ppv")@y.values
#PPV=mean(as.numeric(unlist(ppv)),na.rm = TRUE)
PPV[n]=mean(as.numeric(unlist(ppv)),na.rm = TRUE)


}
SENSI
sensi=mean(as.numeric(unlist(SENSI)),na.rm=TRUE)
```

```
sensi

SPECI
speci=mean(as.numeric(unlist(SPECI)),na.rm=TRUE)
speci

AUC
auc=mean(as.numeric(unlist(AUC)),na.rm=TRUE)
auc
plotauc=(as.numeric(unlist(AUC)))
plot(plotauc,main="Distribution of AUC across 10000 Simulated Datasets Using Complete Dataset")
legend("bottomright",legend=c( paste("AUC = .785"),
 paste("Sensitivity=.719"),
paste("Specificity=.444")))
hist(plotauc,main="Distribution of AUC Across 10000 Simulated Datasets,
Using Complete Yelp Data")
legend("topright",legend=c( paste("AUC = .785"),
 paste("Sensitivity=.719"),
paste("Specificity=.444")))

sdauc=sd(as.numeric(unlist(AUC)),na.rm=TRUE)
sdauc
mean(PPV)
sd(PPV)
plot(PPV)
```

```
hist(PPV,main="Distribution of PPV Across 10000 Simulated Datasets,
Using Complete Yelp Dataset")
legend("bottomright",legend=c( paste("Mean = .564"),
 paste("Std Dev=.052")))
mean(auc)
mean(as.numeric(auc))
median(as.numeric(auc))
sd(as.numeric(auc))
ppv=performance(pred.obj,"ppv")@y.values
PPVmean=mean(as.numeric(unlist(ppv)),na.rm = TRUE)
PPVmean
plot(PPV)
sd(as.numeric(unlist(ppv)),na.rm = TRUE)
plot(as.numeric(unlist(ppv)),na.rm = TRUE)
hist(as.numeric(unlist(ppv)),na.rm = TRUE)

#Plots will show us distribution of PPV and AUC

########################################################################################
#############################################
#TEST MODEL PREDICTIVE POWER USING ROCR USING Second Half Data
library(ROCR)
for (n in 1:1000) {#i is the counter variable – the exact name doesn’t matter.  The { opens the loop

bottleM=YELP3S[sample(nrow(YELP3),200,replace=TRUE), ]
YMM <- glm( Response ~
RC+dollar+STAR+ilove+oldschool+pushy+pool+affordable+Christ+stench+employees+humid+septic+Jesus+hell+dishes+thebest+highquality+adorable+fabulous+craving+favorite+excellent+service+recommend+professional+delicious+washhands+burnt+ache+pain+cigarette+asshole+awful+rotten+bathroom+toilet+puke+fuck+microwaved+shit+bitch+sucks+mold+mice+spider+exclaim+filthy+roach+DIRTY+Ifounda+clean+diarrhea+vomiting+foodpoisoning+dirty+truck+sick+stomach+hospital+fish+nausea+terrible+horrible, data=bottleM, family=binomial,na.action = na.exclude )

probs = fitted(YMM)

Yelpmean=mean(YMM$HCR<80)

preds = as.numeric(probs>Yelpmean)

pred.obj = prediction(probs,bottleM$Response)

sensi=performance(pred.obj,"tpr")

speci=performance(pred.obj,"fpr")

#accu=performance(pred.obj,"acc")

#AUC<-performance(pred.obj,"auc")@y.values

AUC[n]<-performance(pred.obj,"auc")@y.values


ppv=performance(pred.obj,"ppv")@y.values

#PPV=mean(as.numeric(unlist(ppv)),na.rm = TRUE)

PPV[n]=mean(as.numeric(unlist(ppv)),na.rm = TRUE)



}

AUC

auc=mean(as.numeric(unlist(AUC)),na.rm=TRUE)

auc

aucmedian=median(as.numeric(unlist(AUC)),na.rm=TRUE)

aucmedian

plotauc=(as.numeric(unlist(AUC)))
```

```
plot(plotauc)

hist(plotauc)

sdauc=sd(as.numeric(unlist(AUC)),na.rm=TRUE)

sdauc

PPV

mean(PPV)

median(PPV)

sd(PPV)

plot(PPV)

hist(PPV)

mean(auc)

mean(as.numeric(auc))

median(as.numeric(auc))

sd(as.numeric(auc))

ppv=performance(pred.obj,"ppv")@y.values

PPVmean=mean(as.numeric(unlist(ppv)),na.rm = TRUE)

PPVmean

plot(PPV,main="Positive Predictive Values Across 10,000 Simulated Datasets",ylim=c(0.2,1))

legend("bottomright",legend=c( paste("PPV Mean = .618"),

 paste("PPV Std Dev=.052")))


sd(as.numeric(unlist(ppv)),na.rm = TRUE)

plot(as.numeric(unlist(ppv)),na.rm = TRUE)

hist(as.numeric(unlist(ppv)),na.rm = TRUE)

###########################################################################################
###############
```

```
#Want to test how auc changes with usefulness

for (n in 1:10) {

for (i in 1:10000){


bottle1=Yelp[sample(nrow(Yelp),0+i,replace=TRUE), ]

bottle2=Yelp3[sample(nrow(Yelp3),10000-i,replace=TRUE), ]

bottle3=rbind(bottle1,setNames(bottle2,names(bottle1)))




YMM <- glm( Response ~
RC+dollar+ilove+oldschool+pushy+pool+affordable+Christ+stench+employees+humid+septic+Jesus+hell
+dishes+thebest+highquality+adorable+fabulous+craving+favorite+excellent+service+recommend+prof
essional+delicious+washhands+burnt+ache+pain+cigarette+asshole+awful+rotten+bathroom+toilet+puk
e+fuck+microwaved+shit+bitch+sucks+mold+mice+spider+exclaim+filthy+roach+DIRTY+Ifounda+clean+
diarrhea+vomiting+foodpoisoning+dirty+truck+sick+stomach+hospital+fish+nausea+terrible+horrible,
data=bottle3, family=binomial,na.action = na.exclude )

probs = fitted(YMM)

Yelpmean=mean(YMM$HCR<80)

preds = as.numeric(probs>Yelpmean)

pred.obj = prediction(probs,bottle3$Response)

#sensi=performance(pred.obj,"tpr")

#speci=performance(pred.obj,"fpr")

#accu=performance(pred.obj,"acc")

auc<-performance(pred.obj,"auc")@y.values


AUC[i]=mean(as.numeric(auc))
```

```
}

AUC=mean(as.numeric(auc))


AUC

summary(AUC)

mean(AUC)

plot(AUC,main="AUC Across Increasing Proportions of 2100 Simulated Datasets", xlab="Percentage of
Reviews Not on First Page of Yelp Reviews",xaxt="n")

axis(1, at=c("0","2500","5000","7500","10000"), labels=c("0%","25%","50%","75%","100%"),
col.axis="black", las=2)

legend("bottomright",legend=c( paste("Simulated AUC across 1-100% of reviews occurring on first
page")))




hist(AUC)

plot(as.numeric(auc))

hist(as.numeric(auc))

########################################################################################
####

#Want to test increase in AUC as sample size increases


for (n in 1:10) {#i is the counter variable – the exact name doesn’t matter.  The { opens the loop

for (i in 50:1000) {

bottleM=Yelp[sample(nrow(Yelp3S),i,replace=TRUE), ]

YMM <- glm( Response ~
RC+dollar+ilove+oldschool+pushy+pool+affordable+Christ+stench+employees+humid+septic+Jesus+hell
+dishes+thebest+highquality+adorable+fabulous+craving+favorite+excellent+service+recommend+prof
essional+delicious+washhands+burnt+ache+pain+cigarette+asshole+awful+rotten+bathroom+toilet+puk
e+fuck+microwaved+shit+bitch+sucks+mold+mice+spider+exclaim+filthy+roach+DIRTY+Ifounda+clean+
diarrhea+vomiting+foodpoisoning+dirty+truck+sick+stomach+hospital+fish+nausea+terrible+horrible,
data=bottleM, family=binomial,na.action = na.exclude )

probs = fitted(YMM)

Yelpmean=mean(YMM$HCR<80)

preds = as.numeric(probs>Yelpmean)

pred.obj = prediction(probs,bottleM$Response)

SENSI=performance(pred.obj,"tpr")

SPECI=performance(pred.obj,"fpr")

#accu=performance(pred.obj,"acc")

auc<-performance(pred.obj,"auc")@y.values

AUC[n][i]=mean(as.numeric(auc))


}

}

AUC

plot(as.numeric(unlist(AUC)),ylim=c(0.65,1),main="AUC Across Different Sample Sizes")

PPV=mean(as.numeric(unlist(PPV)),na.rm = TRUE)


#########################################################################################
######

#Want to test how auc changes with stars work in progress


YELPS1 <- Yelp3S[ which(Yelp3S$STAR <='1'),]

YELPS1.5 <- Yelp3S[ which(Yelp3S$STAR<='1.5'),]

YELPS2 <- Yelp3S[ which(Yelp3S$STAR<='2'),]
```

```
YELPS2.5 <- Yelp3S[ which(Yelp3S$STAR<='2.5'),]
YELPS3 <- Yelp3S[ which(Yelp3S$STAR<='3'),]
YELPS3.5 <- Yelp3S[ which(Yelp3S$STAR<='3.5'),]
YELPS4 <- Yelp3S[ which(Yelp3S$STAR<='4'),]
YELPS4.5 <- Yelp3S[ which(Yelp3S$STAR<='4.5'),]
YELPS5 <- Yelp3S[ which(Yelp3S$STAR='5'),]
for (n in 1:100) {
for (i in 1:210){
bottle1=YELPS2[sample(nrow(YELPS1),0+i,replace=TRUE), ]
bottle1.5=YELPS3[sample(nrow(YELPS1.5),210-(1*i),replace=TRUE), ]
bottle2=YELPS4[sample(nrow(YELPS2),210-(2*i),replace=TRUE), ]
bottle2.5=YELPS5[sample(nrow(YELPS2.5),210-(3*i),replace=TRUE), ]
bottle3=YELPS6[sample(nrow(YELPS3),210-(4*i),replace=TRUE), ]
bottle3.5=YELPS3.5[sample(nrow(YELPS3.5),210-(5*i),replace=TRUE), ]
bottle4=YELPS4[sample(nrow(YELPS4),210-(5*i),replace=TRUE), ]


bottleK=rbind(bottle2,bottle3,bottle4,bottle5,bottle6,bottle7)


}


YMM <- glm( Response ~
RC+dollar+SS+ilove+oldschool+pushy+pool+affordable+Christ+stench+employees+humid+septic+Jesus+hell+dishes+thebest+highquality+adorable+fabulous+craving+favorite+excellent+service+recommend+professional+delicious+washhands+burnt+ache+pain+cigarette+asshole+awful+rotten+bathroom+toilet+puke+fuck+microwaved+shit+bitch+sucks+mold+mice+spider+exclaim+filthy+roach+DIRTY+Ifounda+cle
an+diarrhea+vomiting+foodpoisoning+dirty+truck+sick+stomach+hospital+fish+nausea+terrible+horrible, data=Yelp3S, family=binomial,na.action = na.exclude )

probs = fitted(YMM)

Yelpmean=mean(YMM$HCR<80)

preds = as.numeric(probs>Yelpmean)

pred.obj = prediction(probs,bottle3$Response)

#sensi=performance(pred.obj,"tpr")

#speci=performance(pred.obj,"fpr")

#accu=performance(pred.obj,"acc")

auc<-performance(pred.obj,"auc")@y.values


AUC[i]=mean(as.numeric(auc))

}

AUC

plot(AUC)

hist(AUC)

plot(as.numeric(auc))

hist(as.numeric(auc))

plot(Yelp3S$HCR,probs)
```

Appendix C

Power Analysis Chapter 3

```
#Yelp IMPPACT Public Health Data  POWER ANALYSIS
library(pwr)
pwr.anova.test(k=4,f=.25,sig.level=.05,n=100)


pwr.t.test(n=425,sig.level=.05,alternative="greater",power=0.8)


pwr.t.test(n = , d =.2 , sig.level =.05 , power =.90, type = c("two.sample"))


pwr.t2n.test(n1 = , n2=400 , d =.5 , sig.level =.05, power =.95 )


for (i in 10:800) {


nn <- i
 runs <- 100
intercept <- log(9)
odds.ratio <- 2
beta <- log(odds.ratio)
proportion  <-  replicate(
 n = runs,
expr = {

xtest <- rnorm(nn)

linpred <- intercept + (xtest * beta)

prob <- exp(linpred)/(1 + exp(linpred))

runis <- runif(length(xtest),0,1)

ytest <- ifelse(runis < prob,1,0)

prop <- length(which(ytest <= 0.5))/length(ytest)

}

)



result <-  replicate(

 n = runs,

 expr = {

xtest <- rnorm(nn)

linpred <- intercept + (xtest * beta)

prob <- exp(linpred)/(1 + exp(linpred))

runis <- runif(length(xtest),0,1)

ytest <- ifelse(runis < prob,1,0)

summary(model <- glm(ytest ~ xtest,  family = "binomial"))$coefficients[2,4] < .05                    }            )


RESULT[i]=sum(result)


}

RESULTS=na.omit(RESULT)
```

```
plot(RESULTS)




OR1.25=RESULTS

OR1.5=RESULTS

OR1.75=RESULTS

OR2.0=RESULTS



PLOTALL=cbind(OR1.25,OR1.5,OR1.75,OR2.0)




plot(OR2.0, col="red", ylab="Power",xlab="Sample Size", main="Power Curves by Sample Size and
Odds Ratio")

points(OR1.75, col="blue")

points(OR1.5, col="green")

points(OR1.25, col="violet")

legend("bottomright", c("OR 2.0", "OR 1.75", "OR 1.5", "OR 1.25"), col = c("red", "blue",
"green","violet"),

    text.col = "black", lty = c(1, 1, 1,1), pch = c(21, 21, 21,21))


write.csv(PLOTALL,"C:\\Python27\\PHYPOWER.csv")
```

```
plot(OR2.0, col="red", ylab="Power",xlab="Sample Size", main="Power Curves by Sample Size and
Odds Ratio")

points(OR1.75, col="blue")

points(OR1.5, col="green")

points(OR1.25, col="violet")

legend("bottomright", c("OR 2.0", "OR 1.75", "OR 1.5", "OR 1.25"), col = c("red", "blue",
"green","violet"),

      text.col = "black", lty = c(1, 1, 1,1), pch = c(21, 21, 21,21))


plot(OR2.0, col="red", ylab="Power",xlab="Sample Size", main="Power Curves by Sample Size and
Odds Ratio",type="line")

points(OR1.75, col="blue", type="line")

points(OR1.5, col="green", type="line")

points(OR1.25, col="violet", type="line")

legend("bottomright", c("OR 2.0", "OR 1.75", "OR 1.5", "OR 1.25"), col = c("red", "blue",
"green","violet"),

      text.col = "black", lty = c(1, 1, 1,1))




RESULTS1=RESULTS

RESULTS2=RESULTS

RESULTS3=RESULTS

RESULTS4=RESULTS

RESULTS5=RESULTS
```

```
plot(RESULTS1, col="red")

points(RESULTS2, col="blue")

points(RESULTS3, col="green")

points(RESULTS4, col="orange")

points(RESULTS5, col="violet")

power
```

Appendix D

R Code for Statistical Analysis Chapter 3

```
library(MASS)
as.number <- function(x) {as.numeric(levels(x))[x]}

YPH=read.csv("C:\\Users\\John\\Dropbox\\Yelp\\YPHI.csv")
YPH=read.csv("C:\\Documents and Settings\\jschombe\\My Documents\\Dropbox\\Yelp\\YPHI.csv")
YPH=read.csv("C:\\Users\\John\\Dropbox\\Yelp\\PHY.csv")
YPH=read.csv("C:\\Documents and Settings\\jschombe\\My Documents\\Dropbox\\Yelp\\PHY.csv")


sum(YPH$NotWhite)
summary(YPH$NotWhite)
sum(YPH$Older)
summary(YPH$Older)
sum(YPH$HADFBI)
summary(YPH$HADFBI)
head(YPH)


# independent 2-group t-test
t.test(FBIY$STARS,FBIN$STARS) # where y1 and y2 are numeric
t.test(FBIYYW$STARS,FBINYW$STARS) # where y1 and y2 are numeric
```

```
t.test(FBIYYB$Phrate,FBINYB$Phrate) # where y1 and y2 are numeric
t.test(FBIYYW$Phrate,FBINYW$Phrate) # where y1 and y2 are numeric
t.test(FBIYYB$STARS,FBINYB$STARS) # where y1 and y2 are numeric
t.test(FBIYYW$STARS,FBINYW$STARS) # where y1 and y2 are numeric
t.test(FBIYYB$Reviews,FBINYB$Reviews) # where y1 and y2 are numeric
t.test(FBIYYW$Reviews,FBINYW$Reviews) # where y1 and y2 are numeric

chisq.test(mytable)


hist(FBIYYB$Phrate)
hist(FBINYB$Phrate)

summary(FBIYYB)
summary(FBINYB)
summary(FBINYW)
summary(FBIYYW)

mytable <- table(YPH$Gender, YPH$Older, YPH$RACE)
#DIFFERENCE in RACE
#DIFFERENCE in Older


ftable(mytable)
```

```
FBIY=as.data.frame(subset(YPH, YPH$HADFBI==TRUE))

FBIN=as.data.frame(subset(YPH, YPH$HADFBI==FALSE))


FBIYO=as.data.frame(subset(FBIY, FBIY$Older==TRUE))

FBIYY=as.data.frame(subset(FBIY, FBIY$Older==FALSE))

FBINO=as.data.frame(subset(FBIN, FBIN$HADFBI==TRUE))

FBINY=as.data.frame(subset(FBIN, FBIN$HADFBI==FALSE))


FBIYOW=as.data.frame(subset(FBIYO, FBIYO$RACE=='White'))

FBIYOB=as.data.frame(subset(FBIYO, FBIYO$RACE=='Minority'))

FBIYYW=as.data.frame(subset(FBIYY, FBIYY$RACE=='White'))

FBIYYB=as.data.frame(subset(FBIYY, FBIYY$RACE=='Minority'))

FBINOW=as.data.frame(subset(FBINO, FBINO$RACE=='White'))

FBINOB=as.data.frame(subset(FBINO, FBINO$RACE=='Minority'))

FBINYW=as.data.frame(subset(FBINY, FBINY$RACE=='White'))

FBINYB=as.data.frame(subset(FBINY, FBINY$RACE=='Minority'))


YOUNG=as.data.frame(subset(YPH, YPH$Older=='TRUE'))

OLD=as.data.frame(subset(YPH, YPH$Older=='FALSE'))

summary(OLD)


nrow(FBIYOW)

nrow(FBIYOB)

nrow(FBIYYW)

nrow(FBIYYB)
nrow(FBINOW)

nrow(FBINOB)

nrow(FBINYW)

nrow(FBINYB)



USERS=as.data.frame(subset(YPH, YPH$Useyelp=='1'))

NUSERS=as.data.frame(subset(YPH, YPH$Useyelp=='0'))


head(YPH)

head(USERS)

head(NUSERS)

nrow(USERS)

nrow(NUSERS)


S1=as.data.frame(subset(YPH, YPH$STARS=='1'))

S2=as.data.frame(subset(YPH, YPH$STARS=='2'))

S3=as.data.frame(subset(YPH, YPH$STARS=='3'))

S4=as.data.frame(subset(YPH, YPH$STARS=='4'))

S5=as.data.frame(subset(YPH, YPH$STARS=='5'))


mean(as.number(S1$phrate))

mean(as.number(S2$phrate))

mean(as.number(unlist(S3$phrate)))

mean(as.number(S4$phrate))
mean(as.number(S5$phrate))


plot(as.number(YPH$STARS),as.number(YPH$phrate))


plot(mean(as.number(S1$phrate)),mean(as.number(S2$phrate)))#,mean(as.number(S3$phrate)),mean(as.number(S4$phrate)),mean(as.number(S5$phrate)))


YTAB <- xtabs(~Age+Gender+RACE2, data=USERS)

NTAB <- xtabs(~Age+Gender+RACE2, data=NUSERS)

YTAB

NTAB

prop.table(YTAB, 1) # row percentages

prop.table(YTAB, 2) # column percentages

prop.table(NTAB, 1) # row percentages

prop.table(NTAB, 2) # column percentages

library(MASS)



plot(YPH$RACE)


library(mass)


plot(YPH$Gender)


plot(YPH$Age)
plot(YPH$effort)

head(YPH)

summary(YPH)

ncol(YPH)

nrow(YPH)

 YPH$Gender=factor(YPH$What.is.Your.Gender.)

head(YPH$Gender)

Male=sum(YPH$Gender=="Male")
Female=sum(YPH$Gender=="Female")
mytable <- table(iris$Species)
 lbls <- paste(names(YPH$Gender), "\n", YPH$Gender, sep="")
 pie(factor(YPH$Gender))

head(YPH$Gender)

slices <- c(Male,Female)
 lbls <- c("Male", "Female")
 pie(slices, labels = lbls, main="Pie Chart of Participation by Gender")
```

```
mod1<-glm(fivestar ~
Older+Gender+RACE+STARG2+as.numeric(FREQ)+as.numeric(Phrate)+as.numeric(Phrate),data=YPH,
family=binomial,na.action = na.exclude )

summary(mod1)


mod1y<-glm(fivestar ~ Gender+
RACE+HADFBI+as.numeric(FREQ)+as.numeric(STARS)+as.numeric(Phrate)+as.numeric(Reviews),data=
YOUNG, family=binomial,na.action = na.exclude )

summary(mod1y)

summary(YOUNG)


mod1o<-glm(fivestar ~ Gender+
RACE+HADFBI+as.numeric(FREQ)+as.numeric(STARS)+as.numeric(Phrate)+as.numeric(Reviews)+as.nu
meric(FREQ)*as.numeric(STARS),data=OLD, family=binomial,na.action = na.exclude )

summary(mod1o)

exp(confint(mod1o))


mod2<-glm(perfhealth ~ Hispanic+Older+Gender+
RACE+FBIHX+Chinese+as.numeric(FREQ)+as.numeric(STARS)+YELPYN+as.numeric(Phrate)+as.numeric
(Reviews)+as.numeric(STARS)*as.numeric(FREQ),data=YPH, family=binomial,na.action = na.exclude )

summary(mod2)


mod2y<-glm(perfhealth ~ Hispanic+Gender+
RACE+FBIHX+Chinese+as.numeric(FREQ)+as.numeric(STARS)+YELPYN+as.numeric(Phrate),data=YOUN
G, family=binomial,na.action = na.exclude )

summary(mod2y)
```

```
mod3<-glm(Violyn ~ Older+Gender+
RACE+HADFBI++as.numeric(FREQ)+as.numeric(STARS)+YELPYN+as.numeric(Phrate),data=YPH,
family=binomial,na.action = na.exclude )

summary(mod3)



mod4=glm(HADFBI~
Chinese+FastFood+Mexican+Indian+Vietnamese+Thai+French+Italian+Older+Gender+
RACE+as.numeric(FREQ)+as.numeric(STARS)+YELPYN+as.numeric(Phrate)+as.numeric(Reviews),data=
YPH, family=binomial,na.action = na.exclude )

summary(mod4)



mod2<-glm(perfhealth ~ as.factor(Age)+Gender+
White+Asian+Black+Latino+as.factor(FT)+FBIHX+Chinese+Fast.Food+Italian+French+as.numeric(FREQ)
+as.numeric(STARS)+Useyelp,data=YPH, family=binomial,na.action = na.exclude )

summary(mod2)

mod3<-glm(perfhealth ~ as.factor(Age)+as.factor(Gender)+
White+Asian+Black+Latino+as.factor(FT)+FBIHX+Chinese+Fast.Food+Italian+French+as.numeric(FREQ)
+as.numeric(STARS)+Useyelp,data=YPH, family=binomial,na.action = na.exclude )

summary(mod2)

fbimod<-glm(fbi ~ as.factor(Age)+as.factor(Gender)+
White+Asian+Black+Latino+as.factor(FT)+Chinese+Fast.Food+Italian+French+as.numeric(FREQ)+Useye
lp+as.factor(vermin)+as.factor(washhand)+as.factor(perfhealth)+as.number(fivestar)+as.number(STA
RS)+as.number(phrate),data=YPH, family=binomial,na.action = na.exclude )

summary(fbimod)

exp(confint(fbimod))



fbimod2<-glm(fbi ~ as.factor(Age)+as.factor(Gender)+
RACE2+as.factor(FT)+Chinese+Fast.Food+Italian+French+Japanese+Indian+Thai+Lebanese+Mexican+V
ietnamese+Greek+as.numeric(FREQ)+Useyelp+Yelper+Yelps+Yelps*as.numeric(FREQ)+as.number(STA
RS)+as.number(phrate)+as.numeric(FREQ)+as.number(STARS)*RACE2,data=YPH,
family=binomial,na.action = na.exclude)

summary(fbimod2)
```

```
exp(confint(fbimod2))



FBIMOD=glm(HADFBI~
Older+Gender+RACE+STARS+YELPYN+as.numeric(Phrate)+YELPERYN+as.numeric(Reviews),data=YPH,
family=binomial,na.action=na.exclude)

summary(FBIMOD)

exp(confint(FBIMOD))


exp(confint(mod1))

exp(confint(mod2))



exp(confint(mod))

attach(YPH)

detach(YPH)

PLRY=subset(YPH, YPH$EFFORT>=1)

YPOLR <- polr(as.factor(EFFORT) ~as.factor(Age)+as.factor(Gender)+
White+Asian+Black+Latino+Useyelp+as.numeric(STARS), data = PLRY)

summary(YPOLR)

exp(confint(YPOLR))

#Test Parralel slopes assumption required by proportional odds logistic regression model


Y0=glm(I(as.numeric(EFFORT) >= 0) ~ as.factor(Age)+as.factor(Gender)+
White+Asian+Black+Latino+Useyelp+as.numeric(STARS), family="binomial", data = YPH)
```

```
Y1=glm(I(as.numeric(EFFORT) >= 1) ~ as.factor(Age)+as.factor(Gender)+
White+Asian+Black+Latino+Useyelp+as.numeric(STARS), family="binomial", data = YPH)


Y2=glm(I(as.numeric(EFFORT) >= 2) ~ as.factor(Age)+as.factor(Gender)+
White+Asian+Black+Latino+Useyelp+as.numeric(STARS), family="binomial", data = YPH)


Y3=glm(I(as.numeric(EFFORT) >= 3) ~ as.factor(Age)+as.factor(Gender)+
White+Asian+Black+Latino+Useyelp+as.numeric(STARS), family="binomial", data = YPH)


exp(confint(Y0))

exp(confint(Y1))

exp(confint(Y2))

exp(confint(Y3))


FBIm <- polr(as.factor(FBI) ~as.factor(Age)+as.factor(Gender)+
White+Asian+Black+Latino+Useyelp+as.numeric(STARS)+vermin+washhand+perfhealth+fivestar, data
= YPH, na.action = na.exclude)

summary(FBIm)

exp(confint(FBIm))
```

Appendix E

Survey Instrument Assessing Public Health Impact of Social Media Data

What is your age range?

Total Count (N)	Missing	Unique
851	7 (0.8%)	3

Counts/frequency: 18-29 (136, 16.0%), 30-50 (390, 45.8%), >50 (325, 38.2%)

What is Your Gender?

Total Count (N)	Missing	Unique
852	6 (0.7%)	2

Counts/frequency: Male (295, 34.6%), Female (557, 65.4%)

What is your Ethnicity?

Total Count (N)	Missing	Unique
837	21 (2.4%)	1

Counts/frequency: White (545, 65.1%), African American (29, 3.5%), Asian (179, 21.4%), Hispanic (70, 8.4%), Other (44, 5.3%)

Do you enjoy eating food on Food Trucks?

Total Count (N)	Missing	Unique
847	11 (1.3%)	2

Counts/frequency: Yes (613, 72.4%), No (234, 27.6%)

What types of cuisine do you most enjoy?

Total Count (N)	Missing	Unique
842	16 (1.9%)	1

Counts/frequency: Chinese (408, 48.5%), Japanese (485, 57.6%), Mexican (569, 67.6%), Italian (429, 51.0%), French (234, 27.8%), Vietnamese (370, 43.9%), American (369, 43.8%), Fast Food (78, 9.3%), Greek (281, 33.4%), Indian (422, 50.1%), Thai(521, 61.9%), Lebanese (137, 16.3%)

How frequently do you go out to eat?

Total Count (N)	Missing	Unique
854	4 (0.5%)	4

Counts/frequency: never (2, 0.2%), less than 4 times per month (321, 37.6%), 2-4 times per week (450, 52.7%), 4+ times per week (81, 9.5%)

Do you write reviews for Yelp?

Total Count (N)	Missing	Unique
834	24 (2.8%)	4

Counts/frequency: Never (556, 66.7%), Occasionally (264, 31.7%), Frequently (12, 1.4%), Every time I go out (2, 0.2%)

When going out to eat how often do you use the website Yelp to choose the restaurant you will eat at?

Total Count (N)	Missing	Unique
857	1 (0.1%)	5

Counts/frequency: I don't know (2, 0.2%), Never (85, 9.9%), Occasionally (410, 47.8%), Frequently (279, 32.6%), All of the time(81, 9.5%)

Have you ever eaten food from a restaurant that made you sick?

Total Count (N)	Missing	Unique
857	1 (0.1%)	3

Counts/frequency: This has never happened to me (330, 38.5%), This happened to me once (309, 36.1%), This happened to me more than once (218, 25.4%)

Do you know if restaurants in San Francisco have their public health inspection ratings and health code Violations posted on Yelp.com?

Total Count (N)	Missing	Unique
855	3 (0.3%)	2

Counts/frequency: Yes (112, 13.1%), No (743, 86.9%)

For you to choose to eat at a restaurant how many Yelp stars would a restaurant need?(1-5)

Total Count (N)	Missing	Unique
855	3 (0.3%)	6

Counts/frequency: Stars are not important (134, 15.7%), I will eat at a restaurant with 1 star (10, 1.2%), I will only eat at a restaurant with two stars or more (24, 2.8%), I will only eat at a restaurant with three stars or more (460, 53.8%), I will only eat at a restaurant with four stars or more (220, 25.7%), I only eat at five star restaurants (7, 0.8%)

For you to choose to eat at a restaurant how many Yelp reviews would a restaurant need?(enter Number)

Total Count (N)	Missing	Unique	Min	Max	Mean	StDev	Sum	Percentile						
								0.05	0.10	0.25	0.50 Median	0.75	0.90	0.95
852	6 (0.7	33	0.	100,000,000,0	117,370,89	3,425,943,54	100,000,000,0	0.	0.	2.	5.00	20.	50.	100

	%)		00	00,000.00	2,054.83	9,136.42	30,712.00	00	00	00		00	00	.00

Lowest values: 0, 0, 0, 0, 0

Highest values: 500, 500, 1000, 10000, 100000000000000

For you to choose to eat at a restaurant what public health rating would a restaurant need?(1-100)

Total Count (N)	Missing	Unique	Min	Max	Mean	StDev	Sum	Percentile						
								0.05	0.10	0.25	0.50 Median	0.75	0.90	0.95
850	8 (0.9%)	34	0.00	0.00	82.94	19.29	70,500.00	50.00	75.00	80.00	90.00	90.00	95.00	100.00

Lowest values: 0, 0, 0, 0, 0

Highest values: 100, 100, 100, 100, x

If a restaurant had the necessary number of stars but less that your desired health rating would you go?

Total Count (N)	Missing	Unique
848	10 (1.2%)	2

Counts/frequency: Yes (254, 30.0%), No (594, 70.0%)

If a restaurant had a perfect health rating but a star less than the your desired yelp rating would you go?

Total Count (N)	Missing	Unique
846	12 (1.4%)	2

Counts/frequency: Yes (443, 52.4%), No (403, 47.6%)

If a restaurant had an average of 5 stars on Yelp (the best rating) but less than 80 (deficient) health code rating would you go?

Total Count (N)	Missing	Unique
848	10 (1.2%)	2

Counts/frequency: Yes (242, 28.5%), No (606, 71.5%)

If a restaurant had a perfect health rating but only one star on Yelp would you go?

Total Count (N)	Missing	Unique
847	11 (1.3%)	2

Counts/frequency: Yes (169, 20.0%), No (678, 80.0%)

Would you go to a restaurant where employees were observed not washing their hands if the restaurant had a perfect Yelp rating (5 stars)?

Total Count (N)	Missing	Unique
849	9 (1.0%)	2

Counts/frequency: Yes (115, 13.5%), No (734, 86.5%)

If a restaurant had a perfect yelp rating and signs of vermin (rats, mice, insects) had been identified would you go?

Total Count (N)	Missing	Unique
850	8 (0.9%)	2

Counts/frequency: Yes (39, 4.6%), No (811, 95.4%)

If health code violation information were available on Yelp would you make an extra effort to access it?

Total Count (N)	Missing	Unique
847	11 (1.3%)	4

Counts/frequency: No I am only interested in reviews and other information found on yelp (40, 4.7%), No I want the process as easy as possible I will not click an extra link to view a list of restaurant health code violations (109, 12.9%), Yes I would click an extra link to see a list of health code violations (541, 63.9%), Yes I would click two or more links to see a list of health code violations (157, 18.5%)

Complete?

Total Count (N)	Missing	Unique
858	0 (0.0%)	2

Counts/frequency: Incomplete (19, 2.2%), Unverified (0, 0.0%), Complete (839, 97.8%)

`