UNIVERSITY OF CALIFORNIA

Los Angeles

2012 U.S. Vehicle Analysis

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Statistics

by

Ho Yeung Michael Lam

2012

ABSTRACT OF THE THESIS

2012 U.S. Vehicle Analysis

by

Ho Yeung Michael Lam

Master of Science in Statistics

University of California, Los Angeles, 2012

Professor Ying Nian Wu, Chair

By 2010, the United States has about 250 millions registered vehicles - roughly 800 motor vehicles per 1000 people. According to a report by Environmental Defense in 2006, American cars are responsible for nearly half of the greenhouse gases emitted by automobiles globally, despite the fact that the nation's vehicles only made up just 30% of the total cars in use in the worldwide. Intuitively, the idea of replacing low fuel-efficient vehicles with high fuel-efficient ones could significantly decrease the emission of greenhouse gases that contribute to global warming. Nowadays, the U.S. government has always been encouraging people to purchase clean-energy cars; in addition, both American and foreign automobile companies

emphasize energy conservation over engine performance. Based on a large pool of data on 2012 car and light truck models' features as well as their relative emissions and fuel economy information, this thesis is focused on investigating the vehicles' emission and fuel-efficiency performances with the assistance of using different blocking factors such as vehicle make and class level. The aim of this thesis is to help the people ease the process of choosing the relative cleanest and most fuel-efficient vehicle that meet their own needs.

The thesis of Ho Yeung Michael Lam is approved.

Nicolas Christou

Robert L. Gould

Hong Quan Xu

Ying Nian Wu, Committee Chair

University of California, Los Angeles

2012

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

Global warming has become a rising problem and one of the most complicated issues facing the world leaders. Global warming refers to the rising average temperature of the surface of the Earth. This average temperature has been rising particularly in a fast pace since 1980, making a 0.8 degree Celsius increase in the Earth's average surface temperature. Global warming will lead to severe consequences in the ecosystem and many other areas. One of the main causes to the climate change is the excess emission of the greenhouse gases - gases that trap heat in the Earth's atmosphere. As one of the main greenhouse gases, carbon dioxide emits to the atmosphere through both natural processes and human activities. Yet, human activities are indeed the main cause to global warming. The emission of carbon dioxide reached a peak record in 2010, and one of the main contributors of the emission of carbon dioxide are the vehicles on the road.

The report by Environmental Defense in 2006 stated that U.S. automobiles and light trucks were responsible for almost half of all the greenhouse gases emitted by automobiles globally. Hence, reducing greenhouse gas emissions from U.S. automobiles would be one of the most efficient ways for slowing global warming. As gasoline is still the most dependent fuel on our planet, drivers have no choice but to continue to purchase gasoline from gas station. However, drivers could reduce the number of miles they drive and replace low fuel-efficient cars with more environmentally friendly ones.

Most people who do not have particular interest in a specific car make would question themselves when buying a new car. Which vehicle class best suits my need? What choice of cars can I get with this budget? What are the pros and cons of domestic and foreign cars? How much money could I save on gasoline per year if I choose an eco friendly car?

In this thesis, we are trying to analyze this vehicle data with various statistical methods, hoping to find helpful information for people to ease their process of choosing the relative cleanest and most fuel-efficient vehicle that meet their own needs. Besides, domestic versus foreign and the luxury levels of a car make are some of the important blocking factors in our later analysis. With the additional information on the average manufacturer's suggested retail price (MSRP) for each of the vehicle models, we are also trying to explore the relationships between MSRP and other variables with regression analysis. Last but not least, the supplementary Vehicle Eliminator Excel Template, which utilizes the latest 2010 Excel advanced techniques, provides a platform for people to instantly access the 2012 vehicle data, and to easily acquire and compare their target vehicles by interactively choosing their own preferences.

CHAPTER 2

Description of the Data

The source of the data is from the official website of the U.S. Environmental Protection Agency (EPA), which includes all the 2012 vehicle model ratings based on emissions and fuel economy. The original data contains 1952 observations; each observation represents a unique automobile. However, there are 69 observations missing their MPG information, since these are mostly the new 2012 models that have not yet been tested on their MPG and greenhouse gas score. The final data contains 1883 observations and 22 variables after data cleaning. The variables include vehicle make, model, class, country of origin, fuel type, miles per gallon, environmental score, luxury level, and MSRP etc.

2.1 Vehicle Class

There are a total 9 categories for the vehicle class variable. Yet, this thesis is focused on the following 6 vehicle classes due to their relative high market demand: small car, midsize car, large car, pickup, SUV, and minivan. From Table 2.1, we can see the distribution of vehicle class in the data.

Vehicle Class	Total
Cars	
Midsize car	270
Small car	798
Large car	154
Light-duty trucks	
Minivan	22
SUV	501
Pickup	138
Grand Total	**1883**

Table 2.1: Distribution of vehicle class

2.2 MPG

MPG stands for miles per gallon, which is commonly adopted by the United States. There are a total of 3 MPG related variables in the data; however, this thesis is focused on only the overall combined city/highway MPG estimate (Cmb MPG). Hence, MPG refers to combined MPG and will be rounded up to the nearest integer in the rest of this thesis.

Vehicle Class	Average MPG
Cars	
Midsize car	25
Small car	24
Large car	19
Light-duty trucks	
Minivan	21
SUV	20
Pickup	16

Table 2.2: Average MPG by vehicle class

Table 2.2 shows the list of MPG by vehicle class, suggesting that small/midsize cars seem to have a relatively higher average MPG than the light-duty trucks.

2.3 Environmental Score

The environmental score has two main parts: air pollution score and greenhouse gas score. The score is ranging from 0 to 10, where 0 is the worst and 10 is the best. The score is given based on the emission levels and fuel economy values. If a vehicle scores well on both Air Pollution and Greenhouse Gas Score, it receives the SmartWay designation.

2.2.1 Air Pollution Score (APS)

The Air Pollution Score reflects how a vehicle's emission of health-damaging and smog-forming airborne pollutants contributes to air pollution. Nitrogen oxides, carbon monoxide, and particulate matter are some main examples of these pollutants. This score is scaling from 0 to 10. Vehicle that scores a 10 are the cleanest, saying that they do not emit any of these pollutants. Yet, these pollutants do not contain any greenhouse gases such as carbon dioxide. By the way, the average Air Pollution Scores will be rounded up to the nearest integer in the rest of this thesis.

2.2.2 Greenhouse Gas Score (GGS)

The Greenhouse Gas Score reflects the amount of a vehicle's emission of greenhouse gases. Carbon dioxide, methane, and nitrous oxide are some main examples of greenhouse gases. Again, this score is scaling from 0 to 10. Vehicle that scores a 10 represents the lowest emission of greenhouse gases. By the way, the average Greenhouse Gas Scores will be rounded up to the nearest integer in the rest of this thesis.

2.2.3 SmartWay

Those vehicles that have combined Air Pollution and Greenhouse Gas Scores that place them in the top 20% tier of environmental performers are granted the SmartWay designation.

2.4 Automakers

There are a total of 43 makes recorded in this dataset. Let us take a look of the 2011/2012 car makes distribution in the United States, which is listed in Table 2.1 as below.

From Table 2.3, the colored makes produce the majority of car models. The country of origin of these makes include the United States (red), Japan (blue), Germany (green), and South Korea (purple).

Make	Total	Make	Total
ACURA	22	*LAMBORGHINI*	10
ASTON MARTIN	22	*LAND ROVER*	12
AUDI	64	*LEXUS*	40
BENTLEY	12	*LINCOLN*	24
BMW	176	*MASERATI*	6
BUGATTI	2	*MAZDA*	48
BUICK	20	*McLAREN*	2
CADILLAC	20	*MERCEDES-BENZ*	106
CHEVROLET	100	*MINI*	62
CHRYSLER	17	*MITSUBISHI*	49
CODA	2	*NISSAN*	114
DODGE	34	*PORSCHE*	118
FIAT	6	*ROLLS-ROYCE*	12
FISKER	2	*SAAB*	16
FORD	101	*SCION*	10
GMC	57	*SMART*	4
HONDA	61	*SUBARU*	48
HYUNDAI	70	*SUZUKI*	34
INFINITI	40	*TOYOTA*	110
JAGUAR	18	*VOLKSWAGEN*	79
JEEP	32	*VOLVO*	35
KIA	66	**Grand Total**	**1883**

Table 2.3: Distribution of 2011/2012 car makes

As we take a deeper look into the distribution of the car brand's country of origin, Table 2.4 suggests that four of the biggest automobile manufacturers are from countries: Japan, United States, Germany, and South Korea.

Country of Origin	Total	%
FRANCE	2	0.09%
GERMANY	**559**	**29.69%**
ITALY	22	1.17%
JAPAN	**576**	**30.59%**
SOUTH KOREA	**136**	**7.22%**
SWEDEN	35	1.86%
U.K.	128	6.80%
U.S.	**425**	**22.57%**
Grand Total	**1883**	**100%**

Table 2.4: Distribution of the car make's country of origin

2.5 Fuel Type

As shown in Table 2.5, most vehicles rely on gasoline, while there is a small portion of vehicles rely on diesel, electricity, and ethanol fuel mixtures (ethanol/gasoline). Due to the dominance of gasoline vehicles (90.8%), gasoline and non-gasoline vehicles analysis are done separately in the later chapter of this thesis.

Fuel Type	Total	%
Diesel	28	1.5%
Electricity	6	0.3%
Electricity/Gasoline	1	0.1%
Ethanol/Gas	139	7.4%
Gasoline	**1709**	**90.8%**
Total	**1883**	**100%**

Table 2.5: Distribution of the fuel type

2.6 Luxury vs. Non-Luxury

The variable “Luxury” determines whether its automaker is a luxury brand or not. From Table 2.6 shows the following car makes are generally considered as luxury:

ACURA	*CHRYLSER*	*MASERATI*
ASTON MARTIN	*FISKER*	*McLAREN*
AUDI	*INFINITI*	*MERCEDES-BENZ*
BENTLEY	*JAGUAR*	*PORSCHE*
BMW	*LAMBORGHINI*	*ROLLS-ROYCE*
BUGATTI	*LAND ROVER*	*SAAB*
BUICK	*LEXUS*	
CADILLAC	*LINCOLN*	

Table 2.6: List of luxury car brands

In addition, Table 2.7 shows how both non-luxury and luxury vehicles are distributed among different vehicle classes. There are only a small portion of minivan models. The data contains no luxury pickup models. There are plenty of luxury small car models, while the rest of the luxury models are almost equally split among large cars, midsize cars, and SUVs.

	large car	**midsize car**	**minivan**	**pickup**	**small car**	**SUV**	**Total**
Non-Luxury	49	147	21	138	395	372	1122
Luxury	105	123	1	0	403	129	761
Total	**154**	**270**	**22**	**138**	**798**	**501**	**1883**

Table 2.7: Distribution of luxury/non-luxury vehicles by vehicle class

CHAPTER 3

Explanatory Data Analysis

This vehicle data consists of thousands of observations, each observation represents a unique vehicle. In this case, explanatory data analysis is used to help us better understand the data and explore some subtle trends and findings.

First of all, in order to grasp a better picture of the sales market share by vehicle class, Table 3.1 shows the unit sales by vehicle class in April 2012 compared to that of April 2011. The table also shows the % change from year-to-date 2011 compared to year-to-date 2012.

	Apr-12	% Chg from Apr-11	YTD 2012	% Chg from YTD 2011
Cars	632,129	**3.2**	2,475,918	**14.9**
Midsize	323,095	**8.3**	1,245,484	**19.5**
Small	224,415	**-0.4**	899,269	**14.1**
Luxury	84,013	**5.9**	327,691	**11.9**
Large	606	**-93.7**	3,474	**-89.2**
Light-duty trucks	552,318	**1.3**	2,176,025	**5.5**
Pickup	145,100	**4.1**	595,532	**10**
Cross-over	232,635	**-4**	915,056	**-0.6**
Minivan	73,079	**8.3**	264,236	**11.2**
Midsize SUV	55,122	**1.5**	221,357	**12.8**
Large SUV	17,593	**-0.1**	67,773	**-5.9**
Small SUV	16,244	**26.1**	60,067	**26.3**
Luxury SUV	12,545	**12.7**	52,004	**10.5**
Total SUV/Cross-over	334,139	**-1.2**	1,316,257	**2.6**
Total SUV	101,504	**5.8**	401,201	**10.6**
Total Cross-over	232,635	**-4**	915,056	**-0.6**

Source: www.online.wsj.com

Table 3.1: Unit sales by vehicle class

As we can see from Table 3.1, midsize cars and small cars dominate the majority of the car sales; while pickups, crossovers and SUVs dominate the majority of the light-duty trucks sales. Meanwhile, large car sales in 2012 experience a sharp drop of 89.2% compared to that in 2011. The drop in large car and large SUV sales plus the increase in midsize and small car sales may suggest that more and more people favor small/mid-sized vehicles over large ones.

Yet, further analysis by each of the vehicle class is necessary. In addition, analysis of luxury cars is carried out separately due to the discrepancy in marketing strategies between luxury and non-luxury car makes. Besides, comparisons between domestic and foreign makes are carried out by different manners. Last but not least, all the analyses performed in this chapter are done with gasoline type vehicles only, except for the last section in this chapter where the analysis of alternative fuel vehicles is performed.

3.1 Small car

As suggested in Table 3.1, the small car sales market is still expanding in a relative fast pace. Moreover, Table 2.1 points out that the small-sized car occupies the majority of the 2011/2012 models. By eliminating the luxury cars, we are interested in how good are the emission and fuel economy performances within each of the non-luxury car makes.

By combining the Air Pollution Scores and the Greenhouse Gas Scores, the aggregated sums are compared to the average MPG. As we can see from Figure 3.1, it seems that there exists a positively correlated relationship between the average MPG as well as sum of the emission scores within each of the car makes. That is, as sum of the emission scores increase, so does the average miles per gallon.

Furthermore, it seems that the domestic cars generally have lower average MPG and average Greenhouse Gas Scores compared to that of foreign cars, while their average Air Pollution Scores are on a pretty satisfactory level. This may suggest that even though domestic small cars have good controls on the air pollutants emission, they are not doing well enough in fuel-efficiency and controlling the emission of greenhouse gases.

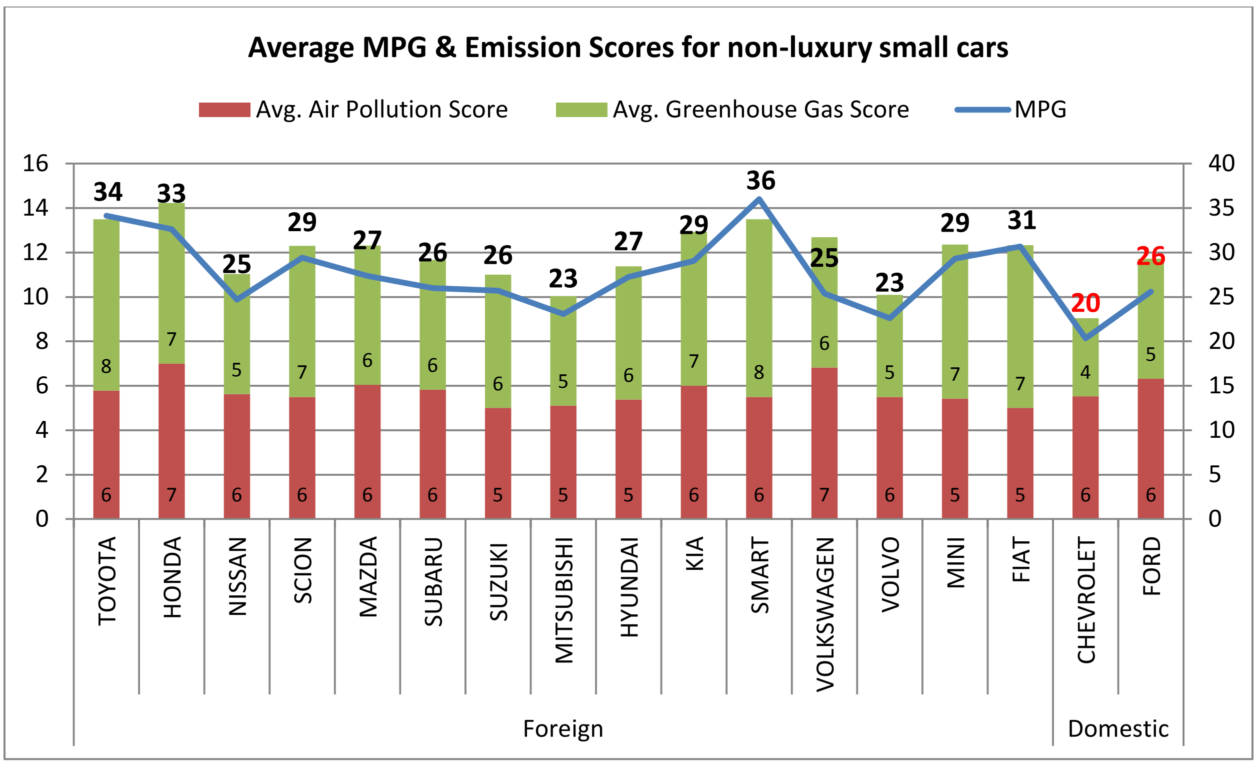


Figure 3.1: Average MPG & Emission Scores for non-luxury small cars

Country of Origin	Makes	No. of unique models	SmartWay	Proportion
GERMANY	*SMART*	2	2	100%
GERMANY	*VOLKSWAGEN*	7	6	86%
ITALY	*FIAT*	1	1	100%
JAPAN	*SUZUKI*	4	1	25%
JAPAN	*HONDA*	6	6	100%
JAPAN	NISSAN	5	2	40%
JAPAN	*SCION*	3	3	100%
JAPAN	*SUBARU*	1	1	100%
JAPAN	*TOYOTA*	3	3	100%
JAPAN	*MAZDA*	3	3	100%
JAPAN	*MITSUBISHI*	5	2	40%
SOUTH KOREA	*HYUNDAI*	2	1	50%
SOUTH KOREA	*KIA*	2	2	100%
SWEDEN	*VOLVO*	3	0	0%
UK	*MINI*	16	16	100%
US	*FORD*	7	4	57%
US	*CHEVROLET*	5	3	60%
US	*CODA*	1	1	100%

Table 3.2: Proportion of SmartWay designated models by non-luxury makes

Table 3.2 shows a clear picture of how good the non-luxury makes are doing at achieving SmartWay designations for each of their models. It seems that almost all makes are doing a good job except Volvo.

Let us take a look of how the emission and fuel economy performances within each of the luxury car makes are doing.

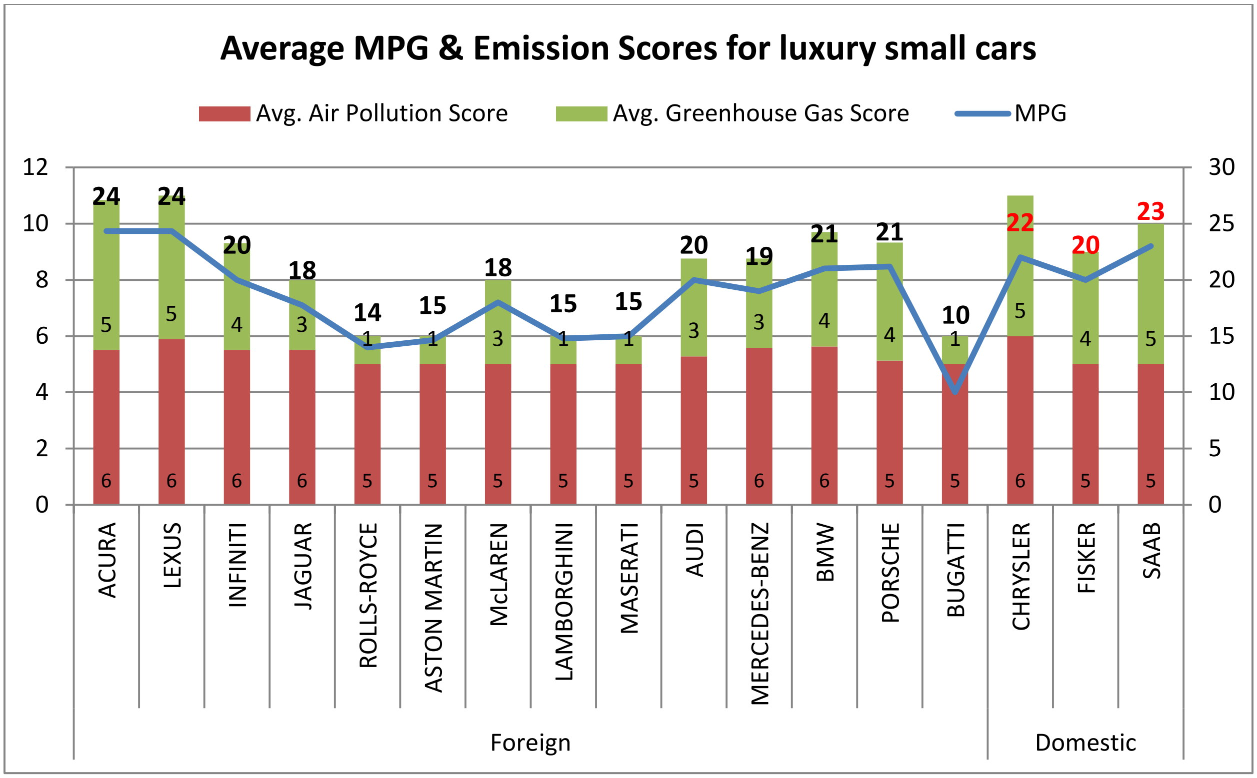


Figure 3.2: Average MPG & Emission Scores for luxury small cars

From Figure 3.2, there is also an obvious trend between the average MPG and the aggregated sums of emission scores. It seems that both luxury and non-luxury small cars share the similar kind of relationship between emission and fuel economy performances. For luxury small cars, domestic cars are doing a pretty good job in both MPG and emission sectors compared to the foreign rivals.

Country of Origin	Make	No. of unique models	SmartWay	Proportion
FRANCE	*BUGATTI*	1	0	0%
GERMANY	***AUDI***	**11**	**5**	**45%**
GERMANY	***BMW***	**18**	**5**	**28%**
GERMANY	*MERCEDES-BENZ*	20	1	5%
GERMANY	*PORSCHE*	25	0	0%
GERMANY	*ROLLS-ROYCE*	2	0	0%
ITALY	*LAMBORGHINI*	3	0	0%
ITALY	*MASERATI*	2	0	0%
JAPAN	***ACURA***	**1**	**1**	**100%**
JAPAN	*INFINITI*	3	0	0%
JAPAN	***LEXUS***	**8**	**3**	**38%**
UK	*ASTON MARTIN*	7	0	0%
UK	*BENTLEY*	3	0	0%
UK	*JAGUAR*	2	0	0%
US	***BUICK***	**1**	**1**	**100%**
US	*CHRYSLER*	1	0	0%
US	*FISKER*	1	0	0%
US	***SAAB***	**3**	**2**	**67%**

Table 3.3: Proportion of SmartWay designated models by luxury makes (small car)

Table 3.3 shows a clear picture of how good the luxury makes are doing at achieving SmartWay designations for each of their models. It seems that Acura, Buick, Saab, Audi, Lexus, and BMW are selling relatively cleaner vehicles than the other rival luxury makes.

3.2 Midsize car

As we see in Table 3.1, midsize car is one of the top selling vehicle classes in the 2012 car sales market. The year-to-date 2012 sales has a nearly 20% growth compared to that in 2011. Let us take a look of how good are the emission and fuel economy performances within each of the non-luxury car makes.

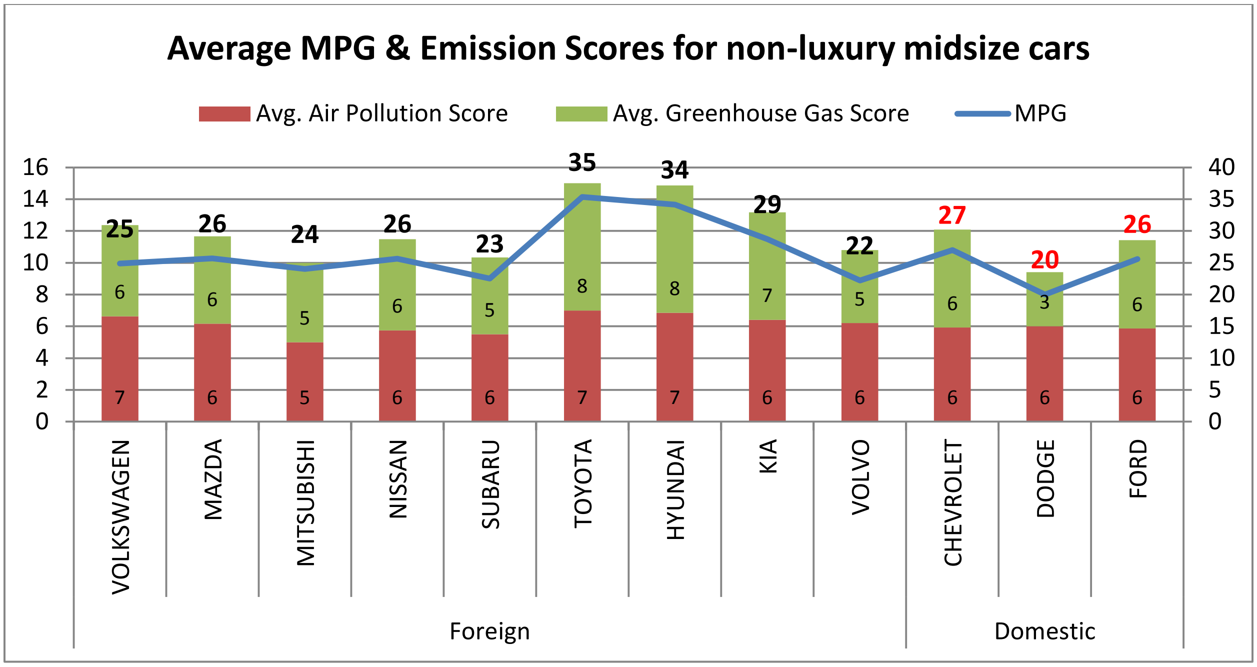


Figure 3.3: Average MPG & Emission Scores for non-luxury midsize cars

Country of Origin	Make	No. of unique models	SmartWay	Proportion
GERMANY	***VOLKSWAGEN***	**1**	**1**	**100%**
JAPAN	***MAZDA***	**3**	**2**	**67%**
JAPAN	*MITSUBISHI*	1	0	0%
JAPAN	***NISSAN***	**4**	**3**	**75%**
JAPAN	***SUBARU***	**1**	**1**	**100%**
JAPAN	***TOYOTA***	**4**	**3**	**75%**
SOUTH KOREA	***HYUNDAI***	**2**	**2**	**100%**
SOUTH KOREA	***KIA***	**4**	**4**	**100%**
SWEDEN	*VOLVO*	1	0	0%
US	***CHEVROLET***	**3**	**3**	**100%**
US	*CHRYSLER*	1	0	0%
US	*DODGE*	3	0	0%
US	***FORD***	**4**	**3**	**75%**

Table 3.4: Proportion of SmartWay designated models by non-luxury makes (midsize car)

From Figure 3.3, there is a clear trend between the average MPG and the aggregated sums of emission scores. Toyota and Hyundai are doing particular good at both fuel-efficiency and low emission of greenhouse gases, while Dodge seems to perform badly. As we look at Table 3.4, it seems that Mitsubushi, Volvo, Chrylser and Dodge are not selling enough SmartWay designated midsize cars.

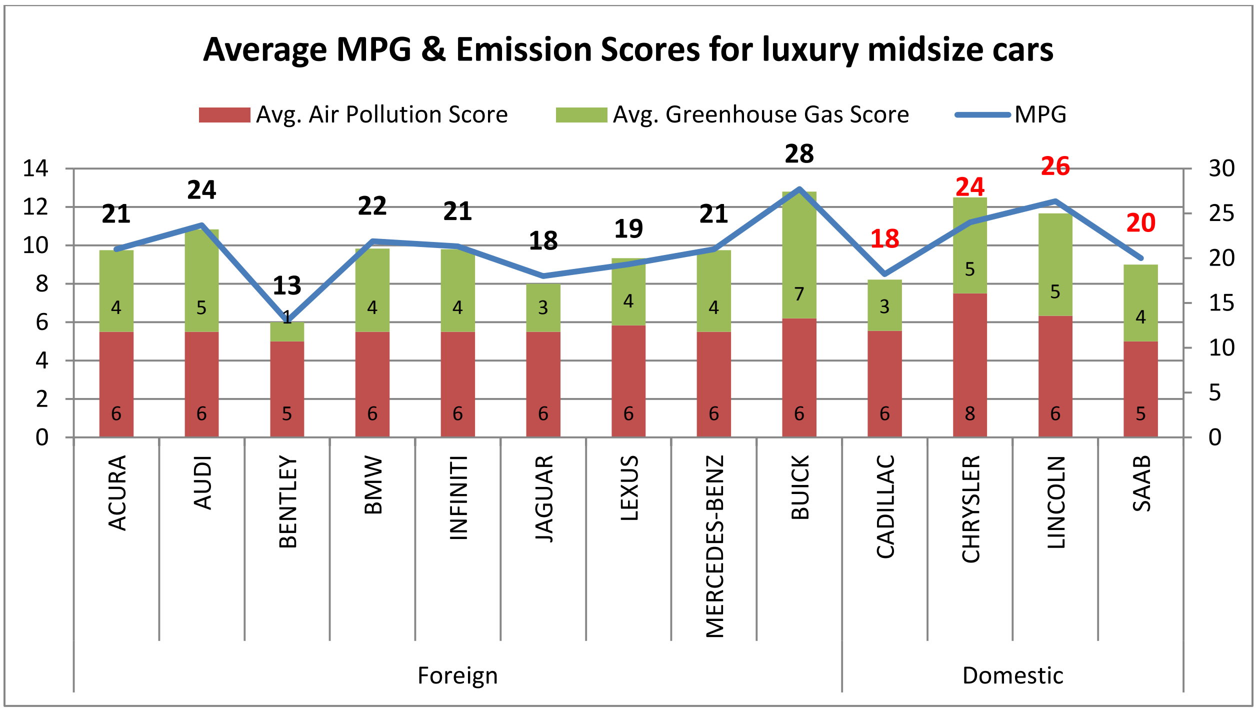


Figure 3.4: Average MPG & Emission Scores for luxury midsize cars

Again, Figure 3.4 shows a clear trend between the line and the bar. In the luxury midsize car sector, Chrysler performs particular well on low emission of air pollutants. Buick performs well in both areas. It seems that Bentley does badly in terms of both MPG as well as emission of greenhouse gas score. Overall, the variances of the fuel-efficiency performance for both foreign and domestic cars are pretty large.

Country of Origin	Make	No. of unique models	SmartWay	Proportion
GERMANY	***AUDI***	**3**	**1**	**33%**
GERMANY	***BMW***	**4**	**1**	**25%**
GERMANY	*MERCEDES-BENZ*	4	0	0%
JAPAN	*ACURA*	2	0	0%
JAPAN	***INFINITI***	**9**	**1**	**11%**
JAPAN	*LEXUS*	4	0	0%
UK	*BENTLEY*	2	0	0%
UK	*JAGUAR*	1	0	0%
US	***BUICK***	**2**	**2**	**100%**
US	*CADILLAC*	1	0	0%
US	*CHRYSLER*	1	0	0%
US	***LINCOLN***	**2**	**1**	**50%**
US	*SAAB*	1	0	0%

Table 3.5: Proportion of SmartWay designated models by luxury makes (midsize car)

From Table 3.5, it seems that Audi, BMW, Infiniti, Buick, and Lincoln provide options of relative cleaner and more fuel-efficient vehicles than the other luxury car makes.

3.3 Large car

It seems that the American no longer have a huge interest in large cars the recent years. As shown in Table 3.1, both of the sales and market share of large car dropped significantly. Yet, there are still a total of 154 large car models in this dataset. Let us take a look which automakers produce large cars and their relative emission and fuel-economy performances.

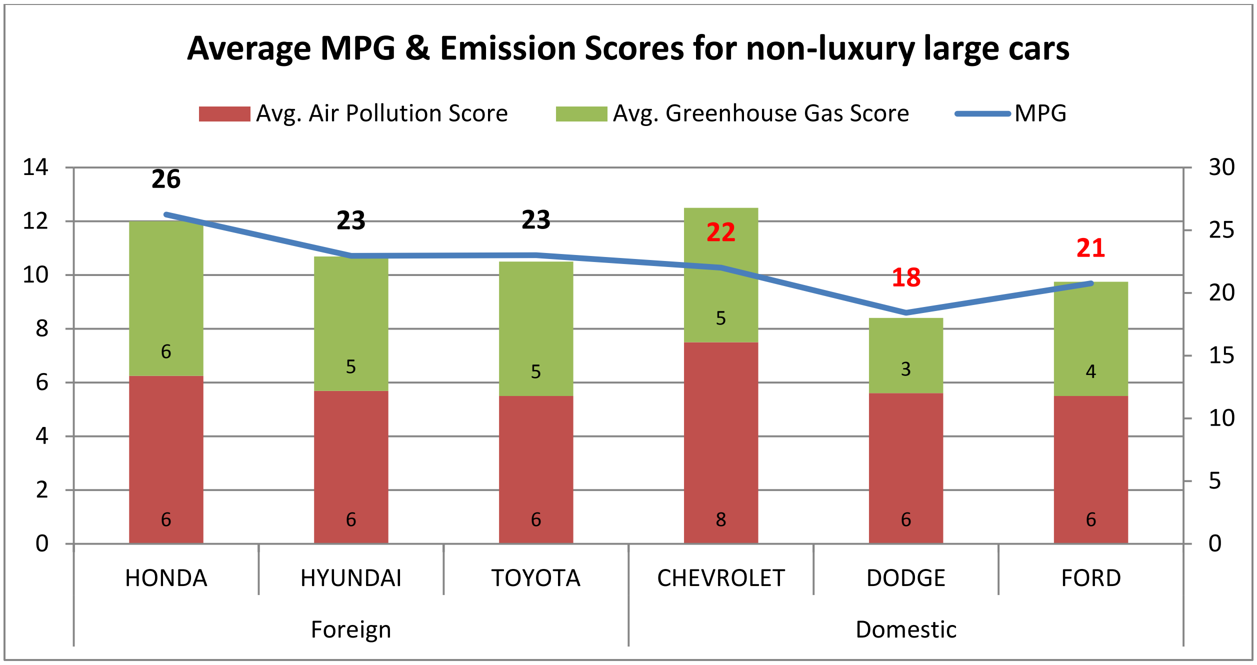


Figure 3.5: Average MPG & Emission Scores for non-luxury large cars

From Figure 3.5, Chevrolet's large cars have the lowest emission of air pollutants. However, the overall average MPGs of domestic large cars are lower than that of foreign large cars. Besides, Dodge has the lowest Greenhouse Gas Score as well as average MPG. It seems that foreign makes are doing better than domestic makes in terms of the large car sector. Table 3.6 validates that foreign makes are selling relative cleaner and more fuel-efficient cars.

Country of Origin	Make	No. of unique models	SmartWay	Proportion
JAPAN	***HONDA***	**1**	**1**	**100%**
JAPAN	*TOYOTA*	1	0	0%
SOUTH KOREA	***HYUNDAI***	**5**	**1**	**20%**
US	*CHEVROLET*	1	0	0%
US	*DODGE*	2	0	0%
US	*FORD*	1	0	0%

Table 3.6: Proportion of SmartWay designated models by non-luxury makes (large car)

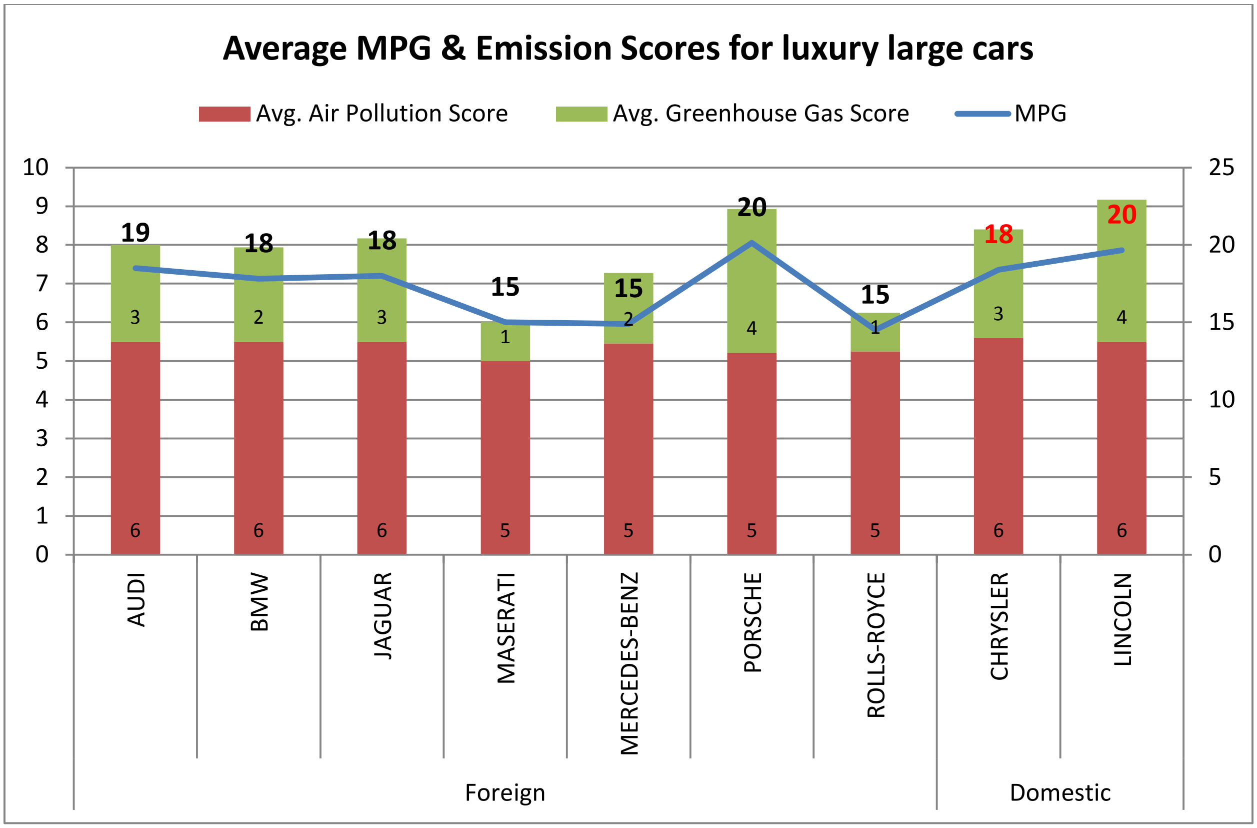


Figure 3.6: Average MPG & Emission Scores for luxury large cars

Country of Origin	Make	No. of unique models	SmartWay	Proportion
GERMANY	*AUDI*	1	0	0%
GERMANY	*BMW*	10	0	0%
GERMANY	*MERCEDES-BENZ*	10	0	0%
GERMANY	***PORSCHE***	**7**	**1**	**14%**
GERMANY	*ROLLS-ROYCE*	4	0	0%
ITALY	*MASERATI*	1	0	0%
UK	*JAGUAR*	2	0	0%
US	*CHRYSLER*	2	0	0%
US	*LINCOLN*	1	0	0%

Table 3.7: Proportion of SmartWay designated models by luxury makes (large car)

By comparing Figure 3.5 and 3.6, we can see that domestic luxury and non-luxury large cars have similar average MPGs and emission scores. However, foreign non-luxury large cars are doing way better foreign luxury ones. Porsche has the best ratings among the foreign makes, while Maserati, Mercedes-Benz, and Rolls-Royce all have low ratings. As only a few large car models can earn the SmartWay Designation, Table 3.6 and 3.7 suggest that large car may not be an eco-friendly vehicle class.

3.4 Pickup

Pickup trucks occupy a significant portion of the light-duty truck sales market. One interesting fact is that none of the luxury car makes is selling luxury pickup trucks.

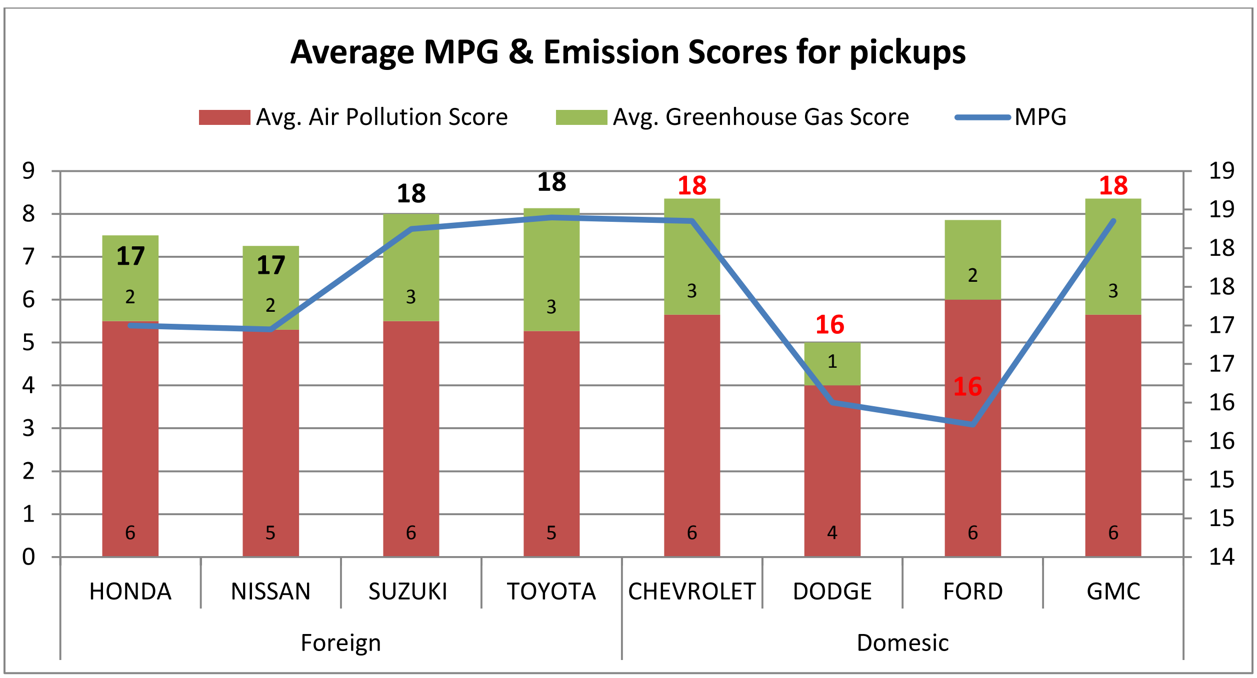


Figure 3.7: Average MPG & Emission Scores for pickups

Country of Origin	Make	No. of unique models	SmartWay	Proportion
JAPAN	*HONDA*	1	0	0%
JAPAN	*NISSAN*	2	0	0%
JAPAN	*SUZUKI*	1	0	0%
JAPAN	*TOYOTA*	2	0	0%
US	*CHEVROLET*	3	0	0%
US	*DODGE*	1	0	0%
US	*FORD*	2	0	0%
US	*GMC*	3	0	0%

Table 3.8: Proportion of SmartWay designated models by non-luxury makes (pickup)

From Figure 3.7, we can see that both foreign and domestic pickups have pretty low average MPG and emission ratings. Dodge has particular low emission scores. Table 3.8 justifies our findings and suggests that pickup is not an environmentally friendly vehicle class.

3.5 SUV

SUV has become popular in the United States since 1990s. There are different sizes of SUV; however, SUV refers to all kinds of SUV in this dataset. Let us take a look of how good are the emission and fuel economy performances of the SUVs within each of the non-luxury car makes.

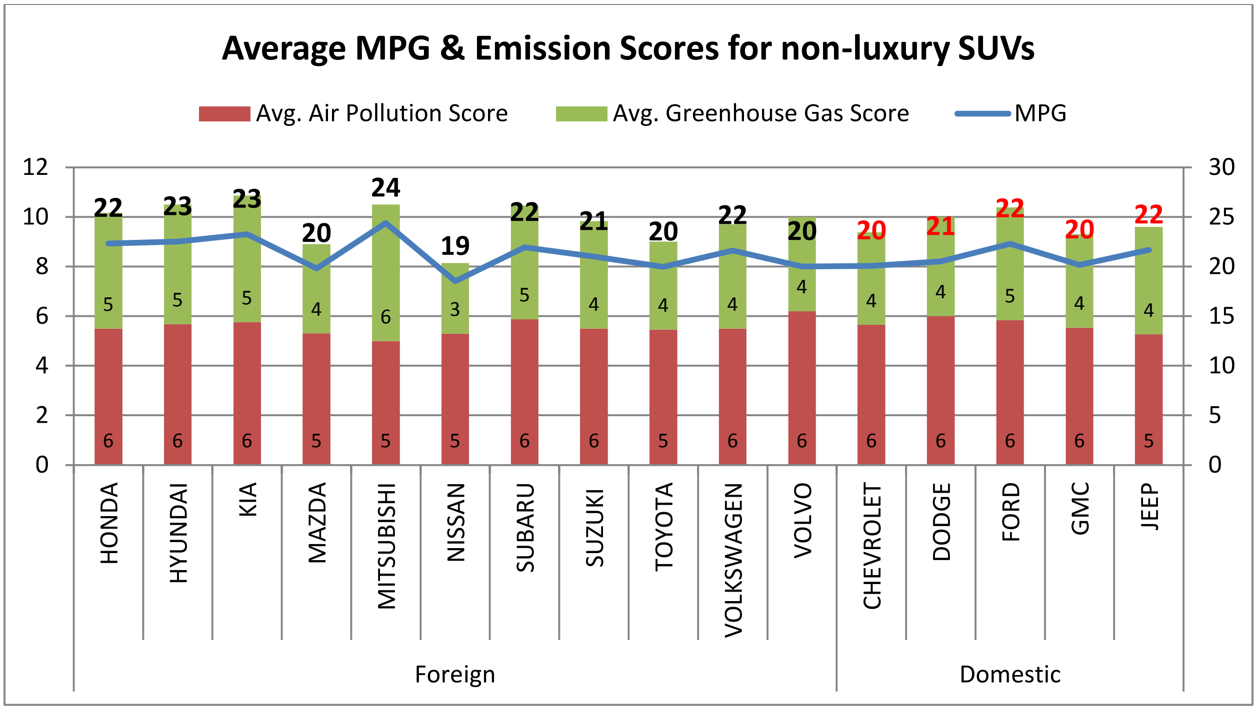


Figure 3.8: Average MPG & Emission Scores for non-luxury SUVs

Country of Origin	Make	No. of unique models	SmartWay	Proportion
GERMANY	*VOLKSWAGEN*	3	0	0%
JAPAN	***HONDA***	**3**	**1**	**33%**
JAPAN	*MAZDA*	2	0	0%
JAPAN	*MITSUBISHI*	2	0	0%
JAPAN	***NISSAN***	**6**	**1**	**17%**
JAPAN	*SUBARU*	3	0	0%
JAPAN	*SUZUKI*	1	0	0%
JAPAN	***TOYOTA***	**7**	**1**	**14%**
SOUTH KOREA	***HYUNDAI***	**3**	**1**	**33%**
SOUTH KOREA	***KIA***	**2**	**2**	**100%**
SWEDEN	*VOLVO*	3	0	0%
US	***CHEVROLET***	**5**	**1**	**20%**
US	*DODGE*	1	0	0%
US	***FORD***	**5**	**2**	**40%**
US	***GMC***	**7**	**1**	**14%**
US	*JEEP*	5	0	0%

Table 3.9: Proportion of SmartWay designated models by non-luxury makes (SUV)

From Figure 3.8, we can see that the average MPG and emission ratings for both domestic and foreign non-luxury SUVs are similar. When we look into Table 3.9, unlike pickup trucks, we notice that both foreign and domestic non-luxury car makes provide the choices of relative cleaner and more fuel-efficient SUV models.

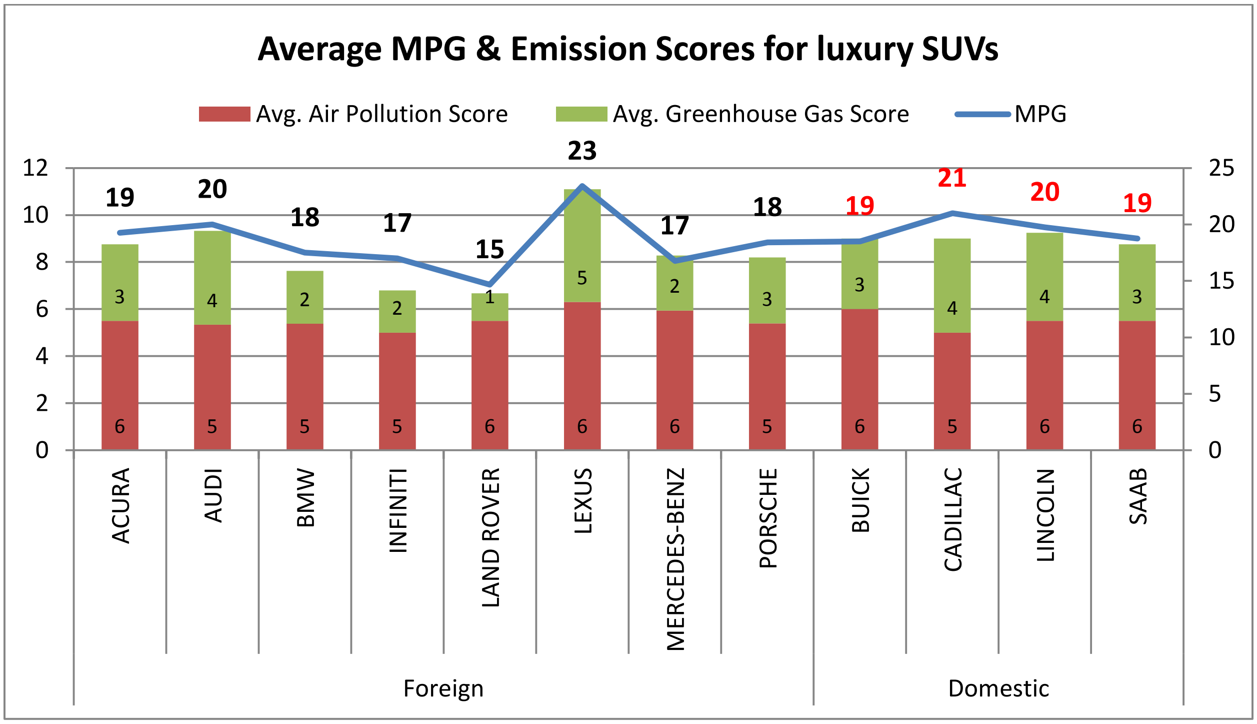


Figure 3.9: Average MPG & Emission Scores for luxury SUVs

From Figure 3.9, we can see that Lexus SUVs have the best performance on MPG and emission ratings, while Land Rover's SUVs are performing badly. Compared to the relative large variances in terms of average MPG and emission scores among foreign luxury makes, the variances of that of domestic luxury makes are relatively smaller.

Table 3.10 justifies our findings that Lexus is the only luxury make that provides choices of relative cleaner and more fuel-efficient SUV models. It also suggests that SUV may not be an eco-friendly vehicle class.

Country of Origin	Make	No. of unique models	SmartWay	Proportion
GERMANY	*AUDI*	2	0	0%
GERMANY	*BMW*	3	0	0%
GERMANY	*MERCEDES-BENZ*	6	0	0%
GERMANY	*PORSCHE*	4	0	0%
JAPAN	*ACURA*	3	0	0%
JAPAN	*INFINITI*	3	0	0%
JAPAN	***LEXUS***	**3**	**1**	**33%**
UK	*LAND ROVER*	4	0	0%
US	*BUICK*	1	0	0%
US	*CADILLAC*	1	0	0%
US	*LINCOLN*	2	0	0%
US	*SAAB*	1	0	0%

Table 3.10: Proportion of SmartWay designated models by luxury makes (SUV)

3.6 Minivan

From Table 3.1, we can see that the minivan sales are almost as many as SUV sales. It is one of the most popular family vans. The funny fact is that Chryslers is the only luxury make selling minivans. Let us take a look of how good are the emission and fuel economy performances of the minivans within each of the non-luxury car makes.

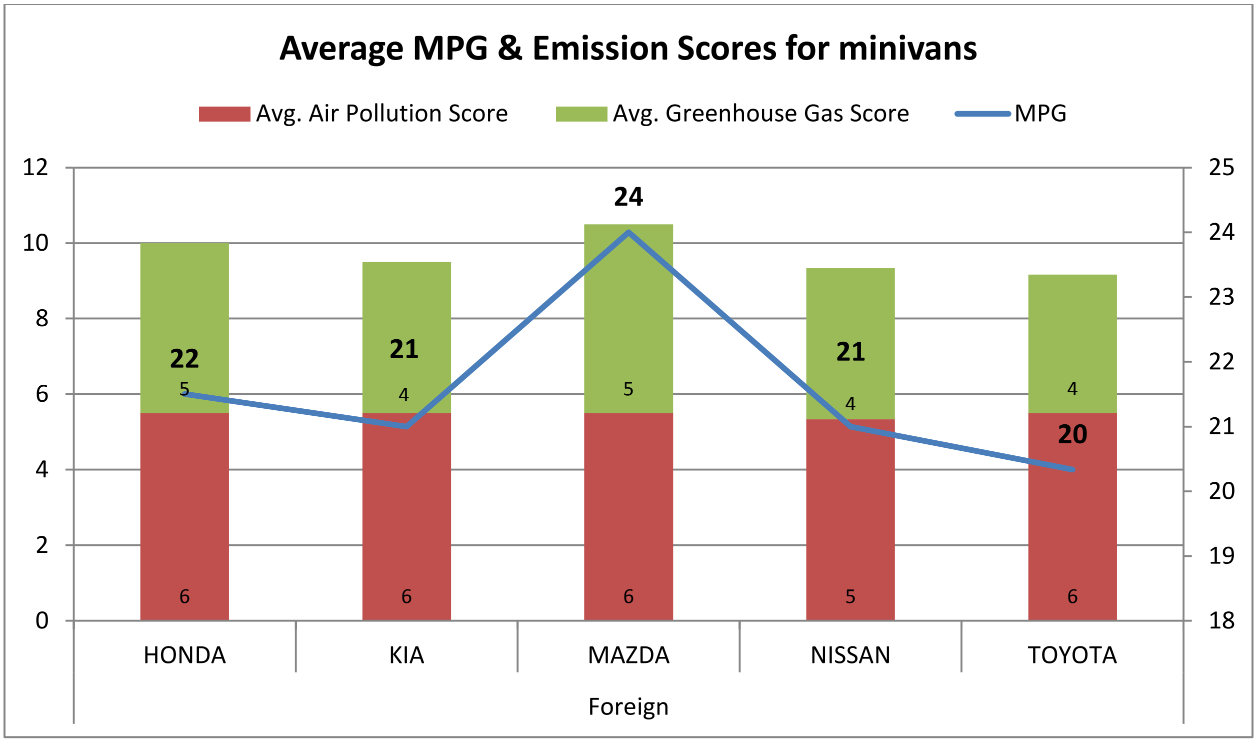


Figure 3.10: Average MPG & Emission Scores for minivans

Country of Origin	Make	No. of unique models	SmartWay	Proportion
JAPAN	*HONDA*	1	0	0%
JAPAN	*MAZDA*	1	0	0%
JAPAN	*NISSAN*	1	0	0%
JAPAN	*TOYOTA*	1	0	0%
SOUTH KOREA	*KIA*	1	0	0%

Table 3.11: Proportion of SmartWay designated models by makes (minivan)

Country of Origin	Make	Avg. MPG	Avg. Air Pollution Score	Avg. Greenhouse Gas Score	SmartWay
US	*CHRYSLERS*	14	6	4	No

Table 3.12: The only Chryslers luxury minivan model

From Figure 3.10, we can see that Mazda's minivan has the highest average MPG and emission scores. The minivan performances of other makes are more or less the same.

From both Table 3.11 and 3.12, none of the minivan models earn the SmartWay designation. It seems that minivan may not be an eco-friendly vehicle class.

3.7 Alternative fuel Vehicles

Gasoline is the most dependent fuel for vehicles. Yet, gasoline will be running out one day and people have to move to other alternatives. Scientists have been trying to find replacement energy to fossil fuels. In this data, alternative fuels to gasoline are diesel, electricity, and ethanol fuel mixtures (Ethanol/Gas). Note that we count the fuel type "electricity/gasoline" as electricity. Let us see how these alternative fuel vehicles perform on emission scores and fuel efficiency compared to those gasoline vehicles.

	Avg. MPG	**Avg. Air Pollution Score**	**Avg. Greenhouse Gas Score**
Non-Luxury			
Diesel	32	6	7
Electricity	95	10	10
Electricity/Gasoline	94	10	10
Ethanol/Gas	13	5	3
Gasoline	24	6	5
Luxury			
Diesel	22	5	4
Ethanol/Gas	14	5	3
Gasoline	20	5	3

Table 3.13: Emission and fuel efficiency performance by fuel types

From Table 3.13, electricity vehicles are undoubtedly the most cleanest and fuel efficient. Compared to gasoline vehicles, diesel vehicles have higher average MPG as well as emission scores. It seems that ethanol fuel mixtures have significant low performance on both emission scores and fuel efficiency.

3.7.1 Electric Vehicle

There are not many choices of electric vehicles, and there are only non-luxury midsize or small cars. The emission scores of all these cars are perfect, saying that they are absolutely environmentally friendly. Besides, the average MPG of electric cars is significantly higher than that of gasoline cars.

	MPG	Avg. Air Pollution Score	Avg. Greenhouse Gas Score
midsize car	**99**	**10**	**10**
NISSAN Leaf	99	10	10
small car	**93.6**	**10**	**10**
CHEVROLET Volt	94	10	10
CODA Coda	75	10	10
MITSUBISHI i-MiEV	112	10	10

Table 3.14: Emission and fuel efficiency performance of electric vehicles

3.7.2 Diesel Vehicle

From Table 3.15, we can see that there are only foreign diesel vehicles. It seems that diesel vehicles generally have a better fuel economy than gasoline vehicles. They also emit less air pollutants and greenhouse gases.

	MPG	Avg. Air Pollution Score	Avg. Greenhouse Gas Score
Non-Luxury	**32**	**6**	**7**
midsize car	**35**	**6**	**7**
VOLKSWAGEN Passat	35	6	7
small car	**34**	**6**	**7**
VOLKSWAGEN Golf	34	6	7
VOLKSWAGEN Jetta	34	6	7
SUV	**22**	**6**	**4**
VOLKSWAGEN Touareg	22	6	4
Luxury	**22**	**5**	**4**
large car	**25**	**6**	**5**
MERCEDES-BENZ S350 Bluetec 4Matic	25	6	5
midsize car	**25**	**6**	**5**
MERCEDES-BENZ E350 Bluetec	25	6	5
SUV	**21**	**5**	**3**
AUDI Q7	20	6	3
BMW X5	22	5	4
MERCEDES-BENZ GL350 Bluetec 4Matic	19	6	2
MERCEDES-BENZ ML350 Bluetec 4Matic	22	6	4
MERCEDES-BENZ R350 Bluetec 4Matic	20	6	3

Table 3.15: Emission and fuel efficiency performance of diesel vehicles

3.7.3 Ethanol fuel mixtures Vehicle

One may questions the fuel efficiency of ethanol vehicles on the road given such low average MPG compared to that of gasoline vehicles. It is because ethanol contains approximately 34% less energy per unit volume than gasoline, and fuel economy is directly proportional to the fuel's energy content. Ethanol fuel mixtures product are cheaper than gasoline, since they contain less gasoline content. However, Table 3.16 suggests that ethanol vehicles are not performing well enough in terms of emission scores.

	MPG	Avg. Air Pollution Score	Avg. Greenhouse Gas Score
Non-Luxury	**13**	**5**	**3**
large car	16	6	5
midsize car	16	6	5
minivan	14	6	4
pickup	12	5	2
SUV	13	5	3
Luxury	**14**	**5**	**3**
large car	17	6	5
midsize car	15	6	4
minivan	14	6	4
small car	13	5	3
SUV	12	5	2

Table 3.16: Emission and fuel efficiency performance of ethanol fuel mixtures vehicles

CHAPTER 4

Regression Analysis on MSRP

In this part of analysis, we are interested in investigating the factors that have significant effects on the average manufacturer's suggested retail price (MSRP) of each of the vehicle models. By eliminating the price variable in the previous analysis, we are allowed to focus on the vehicles' emission and fuel efficiency performance.

A few things have to do before beginning the regression analysis. First, since alternative fuel vehicles only count a very small portion toward the entire data, we decide to only include gasoline vehicles in the regression analysis. Secondly, since the MSRP data is separated from our original data, the average MSRP for each of the observations is based on unique pairs of model and cylinder. Hence, we want to narrow down our original data by grouping same model and cylinder as one unique observation. That is, if there are more than one observation that has exactly the same model and cylinder, all the relevant observations will be averaged out and become one unique observation. In this case, the final data for regression analysis is narrowed down to 459 observations.

On the other hand, the categorical variable "Origin" with 7 categories is chosen to replace the binary variable "Domestic" in the regression. As the binary variable can only show whether the vehicle is domestic or not, the categorical variable can better represent the origin of the vehicle. Note that Origin and Class are categorical variables with more than 2 levels. Luxury is a binary variable indicated by 0=non-luxury and 1=luxury. Now, let us take see which variables are considered in our first regression and its residual plots.

	Estimate	Std. Error	t value P	r(>|t|)
(Intercept)	1491763.8	91429	16.316	**< 2e-16**
Cyl	31250.2	3180	9.827	**< 2e-16**
MPG	469.9	2025	0.232	0.81662
APS	-16118.2	5566.4	-2.896	**0.00397**
GGS	8025.6	6455.2	1.243	0.21443
OriginGERMANY	-1532579.8	74171.4	-20.663	**< 2e-16**
OriginITALY	-1512492.1	78607.3	-19.241	**< 2e-16**
OriginJAPAN	-1555727.3	74875.1	-20.778	**< 2e-16**
OriginSOUTH KOREA	-1559768.5	76179.7	-20.475	**< 2e-16**
OriginSWEDEN	-1554590.9	78460.8	-19.814	**< 2e-16**
OriginUK	-1547935.6	74688.5	-20.725	**< 2e-16**
OriginUS	-1567863	74936.7	-20.923	**< 2e-16**
Luxury	8032.1	9793.8	0.82	0.41259
Classmidsize car	-23964.9	13620.1	-1.76	0.07918
Classminivan	-24046.1	31091.1	-0.773	0.4397
Classpickup	-30882.1	19860	-1.555	0.12067
Classsmall car	-17933.2	11809.7	-1.519	0.1296
ClassSUV	-28745.2	13103.9	-2.194	**0.02878**
Multiple R-squared: 0.7172			**Adjusted R-squared: 0.7063**	

Table 4.1: Summary table of the first regression model

As we see from the summary table, bold coefficients are the ones which are significant at the 5% significance level, while MPG and GGS are shown not significant. Adjusted R-squared is 70.63% which is not bad but it seems that there is some improvement space for the model. Let us check the how the residual plots are doing.

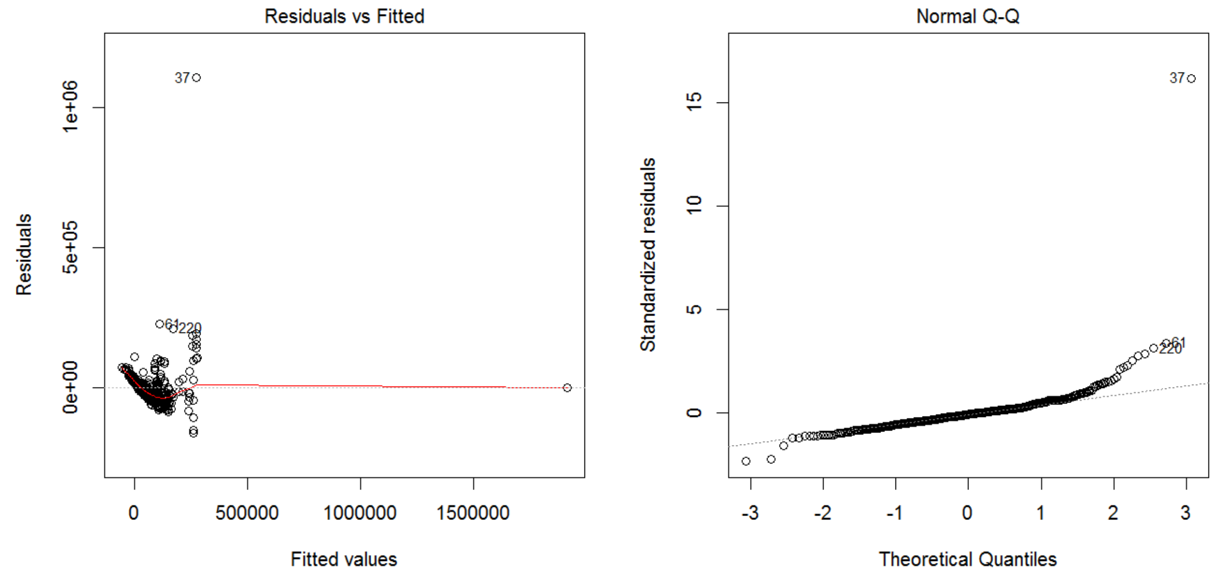


Figure 4.1: Residual plots of the first regression model

As we see from the residual plots in Figure 4.1, the residuals are not randomly scattered in the plot, indicating that the assumption of linearity is violated. Moreover, it seems that the residuals are not equally likely distributed along the 0 line (seems more on the negative side), indicating that the assumption of equal variance of the residuals may be violated. Besides, the normal Q-Q plot shows that there are heavy tails on both sides of the Q-Q line, indicating that the normality assumption is also violated.

From Figure 4.1, the residual plots also suggest that there are some potentially outliers (201, 37, 61, 220) exist. These outliers most likely represent extremely rare and expensive luxury vehicles. It might be a good idea to eliminate these observations in our second try of regression model.

On the other hand, there is suspicious presence of multicolinearity problem in the first regression model. That is, some of the predictor variables might be highly correlated with each other, making the estimate of one variable's impact on the dependent variable while controlling for the others tend to be less precise. Hence, it is necessary for us to check the correlations between the variables.

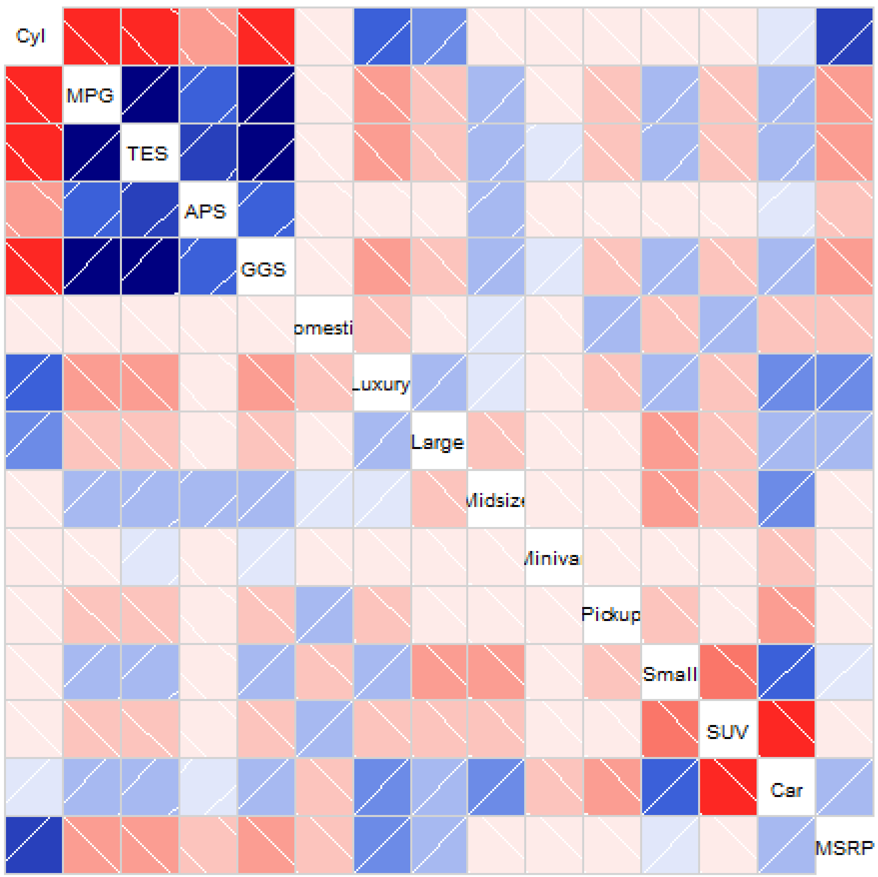


Figure 4.2: Correlogram of the data in the first regression model

From Figure 4.2, the correlogram helps us visualize the data in correlation matrices. The depth of the shading indicates the magnitude of the correlation. Blue indicates positive correlation while red indicates negative correlation. It seems that the variable MPG and GGS might have some potential multicolinearity problem. Variance inflation factors (VIF) test is carried out to further investigate this potential multicolinearity problem.

Cyl	MPG	APS	GGS	Origin	Luxury	Class
4.16	**12.12**	1.62	**16.54**	2.97	2.27	2.68

Table 4.2: VIF test of the first regression model

From Table 4.2, the VIF test suggests that there is major multicolinearity involving the variables MPG and GGS, since their respective VIF are greater than 5. This is not surprising considering that vehicles with higher fuel economy burn less fuel to travel the same distance. As MPG is higher, so does the GGS. This validates our findings in the correlogram in Figure 4.2 that the variables MPG and GGS have multicolinearity problem. In the following second try of our regression model, we decided to drop off GGS since it has the highest VIF indicating that it has the most severe multicolinearity problem.

From Table 4.3, we can see that almost all the variables are significant at the 5% significance level, despite some insignificant categorical levels and insignificant intercept. Yet, the adjusted R-squared is decreased to 65.05%. We have to further check with the residual plots whether the assumptions of linear regression are violated.

From Figure 4.3, the residuals are still not randomly scattered in the plot, indicating that the assumption of linearity is violated. Also, the residuals are not equally likely distributed along the 0 line (seems more on the negative side), indicating that the assumption of equal variance of the residuals may be violated. The normal Q-Q plot does not show any obvious improvement compared to that of Figure 4.1 after eliminating the potential outliers. Overall, the second regression shows only little improvements compared to the first one.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-44810.8	25605.8	-1.75	0.080813
Cyl	25148.1	1622.3	15.502	**< 2e-16**
MPG	2375.7	655.8	3.623	**0.000326**
APS	-12835	3317.3	-3.869	**0.000126**
OriginITALY	29851.2	16362	1.824	**0.068767**
OriginJAPAN	-25016.3	6278.9	-3.984	**7.93E-05**
OriginSOUTH KOREA	-24653.1	10414.3	-2.367	**0.018354**
OriginSWEDEN	-20924.3	15926.7	-1.314	0.189602
OriginUK	-16758.8	8354.1	-2.006	**0.045463**
OriginUS	-32243.1	6932.3	-4.651	**4.38E-06**
Luxury	10320.3	5851.4	1.764	0.078472
Classmidsize car	-16017.1	8180.3	-1.958	**0.050863**
Classminivan	-11721	18558.1	-0.632	0.527987
Classpickup	-22188.5	11463.6	-1.936	**0.053563**
Classsmall car	-5809.4	7081.6	-0.82	0.412461
ClassSUV	-17842.6	7709.4	-2.314	**0.021107**
Multiple R-squared: 0.662		**Adjusted R-squared:0.6505**		

Table 4.3: Summary table of the second regression model

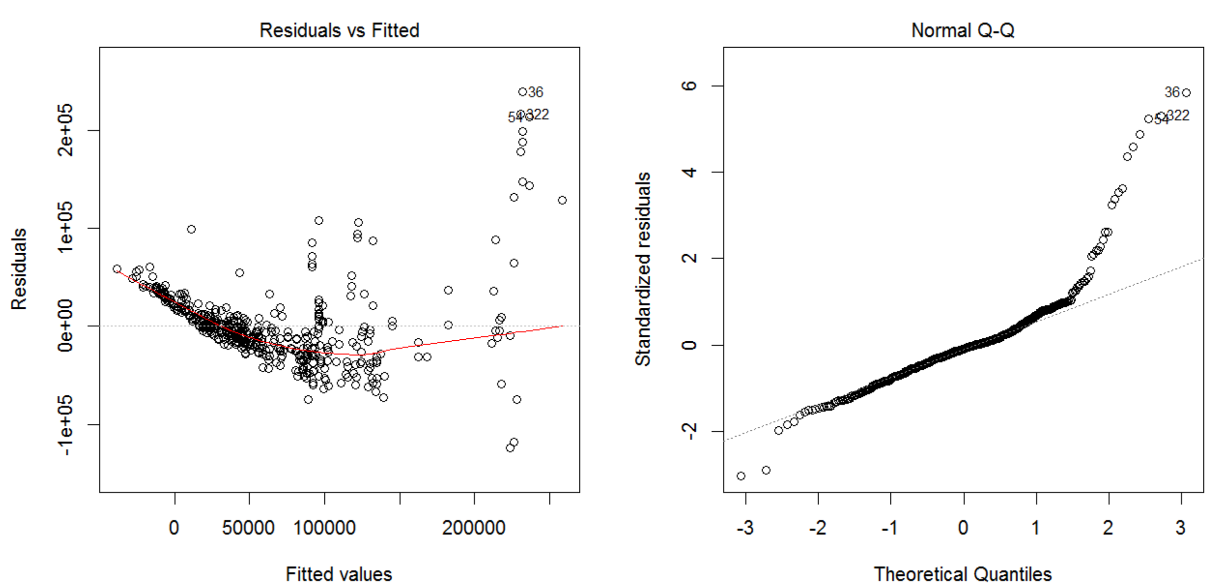


Figure 4.3: Residual plots of the second regression model

It seems that some transformations to the data are necessary. Before we do any transformations, let us take a look of the histograms of the variables to check whether they are normally distributed.

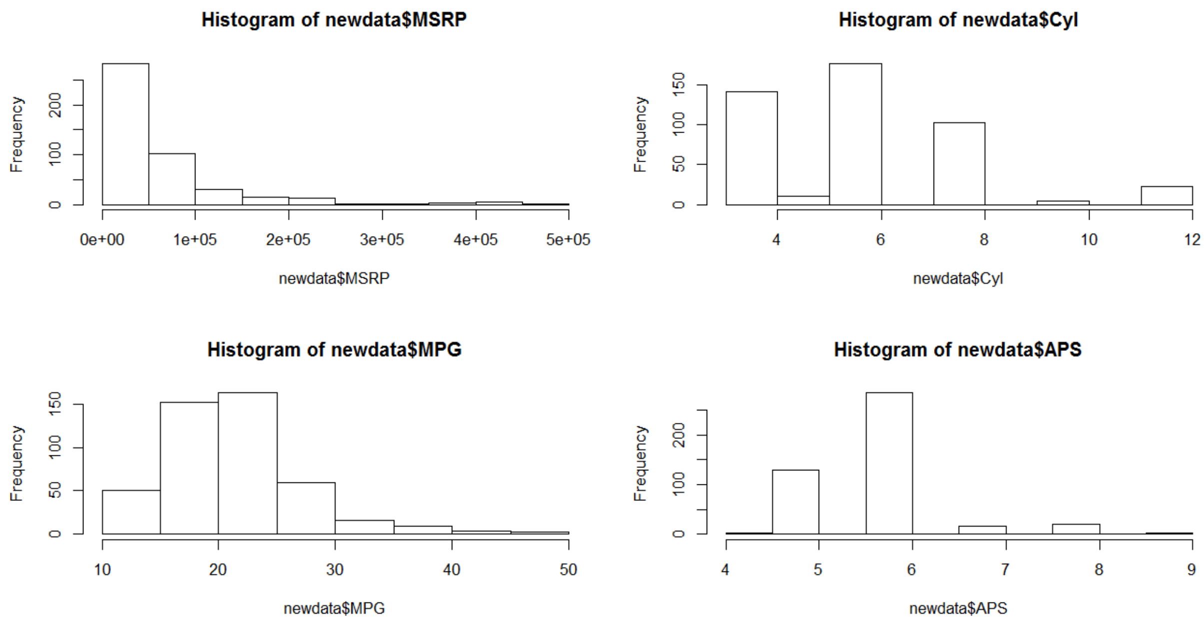


Figure 4.4: Histograms of MSRP, Cyl, MPG, and APS

From Figure 4.4, we can see that all histograms seem to be skewed to the right, indicating that some logarithmic transformation should be done on both dependent and independent variables.

From Figure 4.5, all histograms seem to be more normal than before. Hence, we accept the changes and make up our third regression model on the log-transformed MSRP.

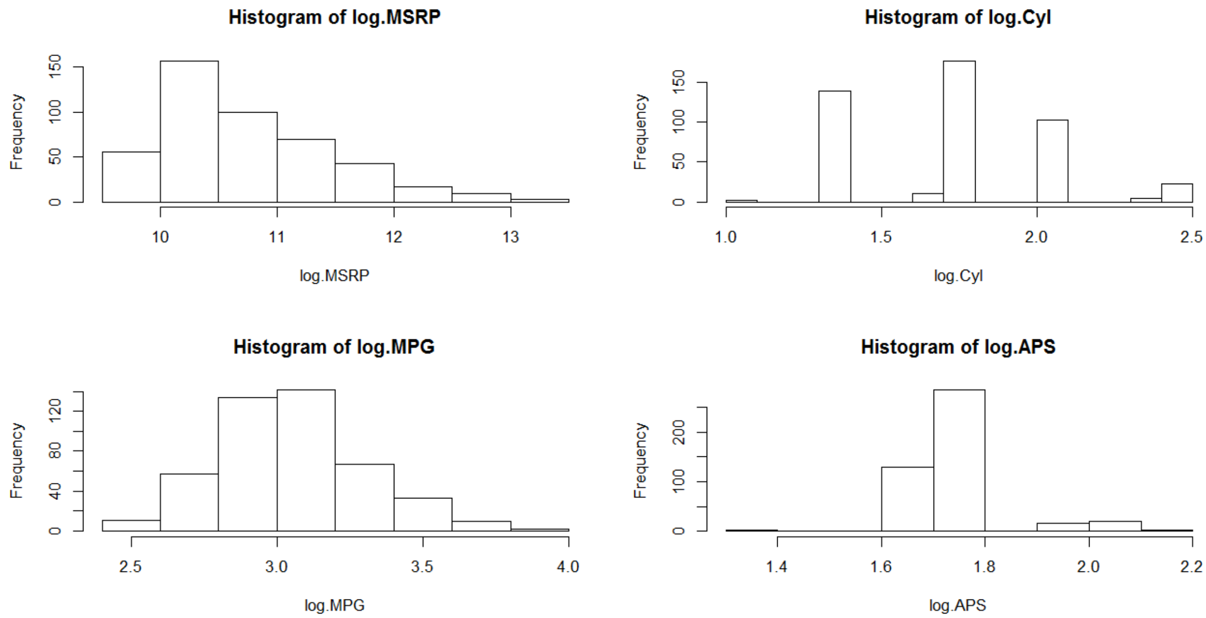


Figure 4.5: Histograms of MSRP, Cyl, MPG, and APS after log-transformation

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.26843	0.52796	19.449	**< 2e-16**
log.Cyl	1.09894	0.09632	11.41	**< 2e-16**
log.MPG	-0.15517	0.13515	-1.148	0.25154
log.APS	-0.50956	0.14919	-3.416	**0.000696**
OriginITALY	0.15757	0.12375	1.273	0.203573
OriginJAPAN	-0.32371	0.04751	-6.814	**3.15E-11**
OriginSOUTH KOREA	-0.33368	0.07887	-4.231	**2.84E-05**
OriginSWEDEN	-0.14928	0.12043	-1.24	0.215818
OriginUK	-0.07446	0.06283	-1.185	0.236566
OriginUS	-0.37967	0.0525	-7.232	**2.14E-12**
Luxury	0.41604	0.04453	9.342	**< 2e-16**
Classmidsize car	-0.13396	0.06165	-2.173	**0.030316**
Classminivan	-0.13381	0.14	-0.956	0.339708
Classpickup	-0.3062	0.08749	-3.5	**0.000514**
Classsmall car	-0.04521	0.05335	-0.847	0.397242
ClassSUV	-0.16535	0.05809	-2.847	**0.004627**
Multiple R-squared: 0.8187	**Adjusted R-squared: 0.8125**			

Table 4.4: Summary table of the third regression model (log-transformation)

From Table 4.4, we can see that almost all variables are significant at the 5% significance level, except for log(MPG) and some insignificant categorical levels, which is not bad. Also, the adjusted R-squared increases to 81.25%, which looks pretty good. Yet, further investigations on the residual plots are needed to check whether the assumptions of linear model hold.

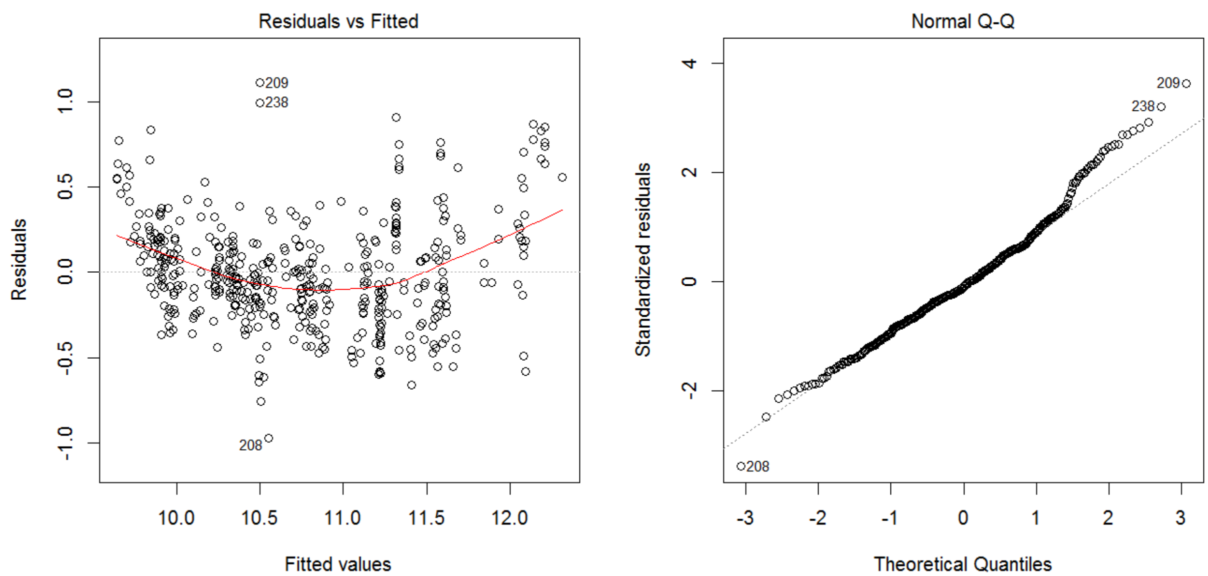


Figure 4.6: Residual plots of the third regression model (log-transformation)

From Figure 4.6, the residuals seem to be more likely randomly scattered than before log-transformation, indicating that linearity assumption may hold. Also, the residuals are somehow distributed along the 0 line (both positive and negative), indicating that the assumption of equal variance of the residuals may also hold. The normal Q-Q plot shows that most of the points fall on the Q-Q line despite that there is still a right tail, indicating that the normality assumption may hold. Overall, the third regression model after log-transformation of both dependent and independent variables shows some huge improvements compared to that before log-transformation.

log.Cyl	log.MPG	log.APS	Origin	Luxury	Class
4.11	**5.09**	1.5	2.54	2.28	2.49

Table 4.5: VIF test of the third regression model (log-transformation)

From Table 4.5, the VIF test still shows some mild multicolinearity involving the variable log(MPG), as the respective VIF is a little bit greater than 5. Together with the fact that the variable log(MPG) is not statistically significant in Table 4.4, we decided to drop off log(MPG) in the following final model.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.77236	0.30352	32.196	**< 2e-16**
log.Cyl	1.18231	0.0633	18.677	**< 2e-16**
log.APS	-0.58252	0.13502	-4.314	**1.98E-05**
OriginITALY	0.16039	0.12377	1.296	0.19571
OriginJAPAN	-0.32757	0.04741	-6.91	**1.71E-11**
OriginSOUTH KOREA	-0.3404	0.07869	-4.326	**1.88E-05**
OriginSWEDEN	-0.14309	0.12035	-1.189	0.23511
OriginUK	-0.07826	0.06276	-1.247	0.2131
OriginUS	-0.38131	0.0525	-7.263	**1.73E-12**
Luxury	0.42198	0.04425	9.537	**< 2e-16**
Classmidsize car	-0.13812	0.06157	-2.243	**0.02537**
Classminivan	-0.1188	0.13944	-0.852	0.39467
Classpickup	-0.27581	0.08343	-3.306	**0.00102**
Classsmall car	-0.05012	0.0532	-0.942	0.34663
ClassSUV	-0.14914	0.05637	-2.646	**0.00844**
Multiple R-squared: 0.8182	**Adjusted R-squared: 0.8124**			

Table 4.6: Summary table of the final model

From Table 4.6, we can see that all the variables are statistically significant at the 5% significance level, except for some insignificant categorical levels, which is acceptable and not a big deal to the model. The adjusted R-squared is 81.24%, indicating that 81.24% of the variance in the MSRP variable can be explained by the independent variables, which says that this model fits the data pretty well.

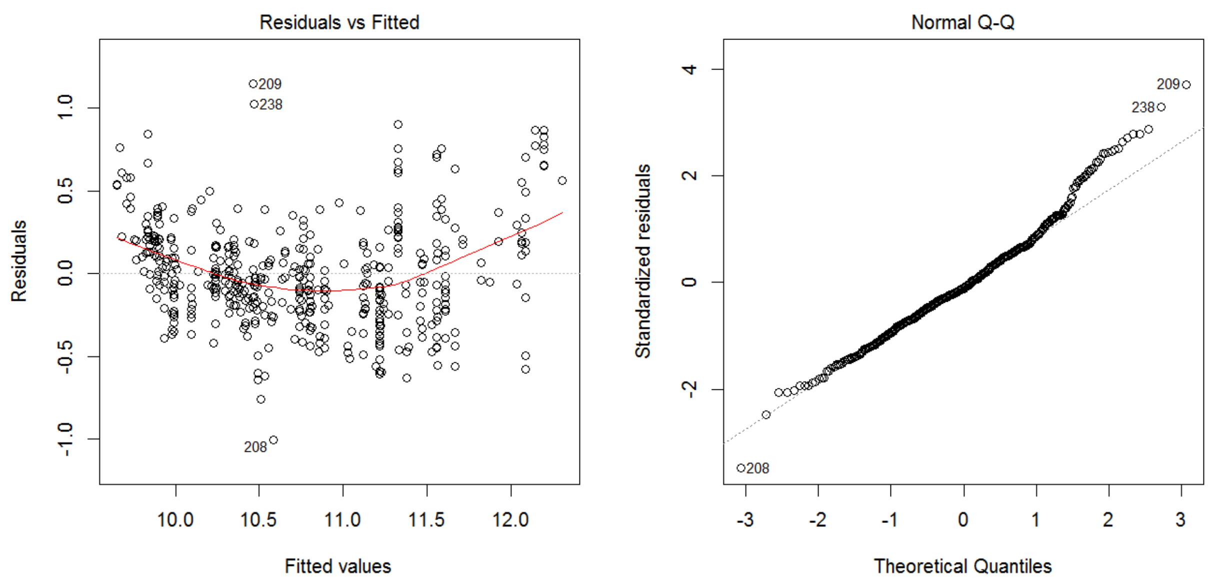


Figure 4.7: Residual plots of the final model

From Figure 4.7, it seems that there is no obvious trend in the residuals, indicating that linearity assumption may hold. Also, the residuals seem to quite evenly and constantly distributed along the 0 line, indicating that the assumption of equal variance of the residuals may also hold. Despite the residual plot and the normal Q-Q may suggest some potential outliers, we decide not to take them out of the regression model, as they are not considered as extreme outliers which are totally fine for a real-life data. The normal Q-Q plot shows that most of the points fall on the Q-Q line despite that there is still a little tail on the right hand side, indicating that the normality assumption may hold.

From Table 4.7, the VIF test shows that there is no problem of multicolinearity among the variables, as all VIF are smaller than 5. Overall, the assumptions of this regression model are held and the regression model is considered to be valid.

log.Cyl	log.APS	Origin	Luxury	Class
1.77	1.23	2.51	2.24	1.87

Table 4.7: VIF test of the final model

The final model is as follow:

$$\log MSRP = 9.77 + 1.18 \cdot \log Cyl - 0.58 \cdot \log APS + 0.16 \cdot O_1 - 0.33 \cdot O_2 - 0.34 \cdot O_3 - 0.14 \cdot O_4 - 0.08 \cdot O_5 - 0.38 \cdot O_6 + 0.42 \cdot Luxury - 0.14 \cdot C_1 - 0.19 \cdot C_2 - 0.28 \cdot C_3 - 0.05 \cdot C_4 - 0.15 \cdot C_5$$

where,

$Origin\ levels$: $O_1: Italty$; $O_2: Japan$; $O_3: South\ Korea$; $O_4: Sweden$; $O_5: UK$; $O_6: US$

$Origin\ level = France$, $when$ $O_1 = O_2 = O_3 = O_4 = O_5 = O_6 = 0$

$Class\ levels$: $C_1: Midsize$; $C_2: Minivan$; $C_3: Pickup$; $C_4: Small$; $C_5: SUV$

$Class\ level = Large$, $when$ $C_1 = C_2 = C_3 = C_4 = C_5 = 0$

If one is interested in predicting the MSRP for a non-luxury midsize car of a Japan make with an average air pollution score of 7 and 4 cylinder, the steps are shown below:

$$\log MSRP = 9.77 + 1.18 \cdot \log 4 - 0.58 \cdot \log 7 - 0.33 \cdot 1 - 0.14 \cdot 1$$

$$\log MSRP = 9.77 + 1.64 - 0.58 \cdot 1.13 - 0.33 \cdot 1 - 0.14 \cdot 1$$

$$\log MSRP = 9.77 + 1.64 - 0.58 \cdot 1.13 - 0.33 \cdot 1 - 0.14 \cdot 1$$

$$\log MSRP = 9.81$$

$$\text{MSRP} = e^{9.81} = 18286$$

Hence, the MSRP of this car with the preference mentioned above is $18,286 US dollars.

To interpret the coefficient estimates of the model with both sides log-transformed, for example:

1. 1% increase in Cyl is associated with 1.18% increase in MSRP, while holding other variables constant.
2. 1% increase in APS is associated with a 0.58% decrease in MSRP, while holding other variables constant.
3. O_2=1, which refers to a car of Japanese make, is associated with a 33% decrease in MSRP compared to that of French make (when all origin levels equal to 0), while holding other variables constant.
4. C_4=1, which refers to a small car, is associated with a 5% decrease in MSRP compared to a large car, while holding other variables constant.

Even though the final model seems to look pretty good at this moment, the residual plot shown in Figure 4.7 seems to show a quadratic pattern. As a result, we would like to add some quadratic terms to the model to see if they can improve this. Besides, we would like to consider a model without doing logarithm transformation on the dependent variables to see if the model looks similar as before. If so, we would go for the model without logarithm transformation on the dependent variables so as to make the interpretation of the model easier to interpret.

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	15.2529662	1.2280954	12.42	< 2e-16
A	-0.2162552	0.1602689	-1.349	0.177937
B	-0.1774927	0.0438017	-4.052	6.01E-05
C	-0.5961587	0.2369235	-2.516	0.012222
I(A^2)	0.021026	0.0051854	4.055	5.95E-05
I(B^2)	0.0004437	0.0006784	0.654	0.513466
I(C^2)	0.0279007	0.0198111	1.408	0.159748
OriginITALY	0.0330559	0.1105056	0.299	0.764982
OriginJAPAN	-0.3492721	0.042391	-8.239	2.08E-15
OriginSOUTH KOREA	-0.3545008	0.0692826	-5.117	4.68E-07
OriginSWEDEN	-0.0784023	0.1064379	-0.737	0.461764
OriginUK	-0.1009764	0.056125	-1.799	0.072694
OriginUS	-0.3922356	0.0462147	-8.487	3.40E-16
Luxury	0.4460823	0.0399812	11.157	< 2e-16
Classmidsize car	-0.1126954	0.0547426	-2.059	0.040126
Classminivan	-0.0807278	0.1235316	-0.653	0.513781
Classpickup	-0.2882856	0.0794983	-3.626	0.000322
Classsmall car	-0.0308974	0.0474835	-0.651	0.515588
ClassSUV	-0.1289378	0.0526834	-2.447	0.014785
A:B	0.0112382	0.0042986	2.614	0.009251
A:C	-0.0255554	0.0176604	-1.447	0.148609
B:C	0.0128815	0.0064687	1.991	0.047068
Multiple R-squared:0.8637		Adjusted R-squared: 0.8571		

Table 4.8: Summary table of the first quadratic model

Table 4.8 shows the summary table with a long list of dependent variables, where A refers to “Cyl”, B refers to “APS”, and C refers to “MPG”. The interaction term A:B represents the effects of average miles per gallon on log(MSRP) depend on the number of cylinder. Since there are a bunch of both significant and insignificant dependent variables, it is hard to tell the best performing model by trying all the different combinations of variables. In this case, utilizing stepwise regression with backward elimination method may help us search for the best model, with using Schwartz’s Bayesian Information Criterion (BIC) as objective criteria for model selection.

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	15.498159	0.791823	19.573	**< 2e-16**
A	-0.345526	0.109036	-3.169	**0.001637**
B	-0.186	0.028681	-6.485	**2.38E-10**
C	-0.576316	0.069158	-8.333	**1.01E-15**
I(A^2)	0.02229	0.004255	5.239	**2.51E-07**
OriginITALY	0.082457	0.110859	0.744	0.457394
OriginJAPAN	-0.39783	0.040629	-9.792	**< 2e-16**
OriginSOUTH KOREA	-0.378244	0.068116	-5.553	**4.86E-08**
OriginSWEDEN	-0.0923	0.106866	-0.864	0.388222
OriginUK	-0.100738	0.056564	-1.781	0.075605
OriginUS	-0.440576	0.044348	-9.935	**< 2e-16**
Luxury	0.465697	0.038275	12.167	**< 2e-16**
A:B	0.010127	0.002933	3.453	**0.000608**
B:C	0.020058	0.002714	7.392	**7.34E-13**
Multiple R-squared: 0.8566		Adjusted R-squared: 0.8524		

Table 4.9: Summary table of the quadratic model after using BIC backward elimination

From Table 4.9, we can see that most of the variables are significant, despite some insignificant categorical levels which is not a big deal. The R-squared 85.24% looks even better compared to the 81.24% obtained in our previous final model, suggesting that this model with the addition of quadratic terms and interaction effects fits the data better.

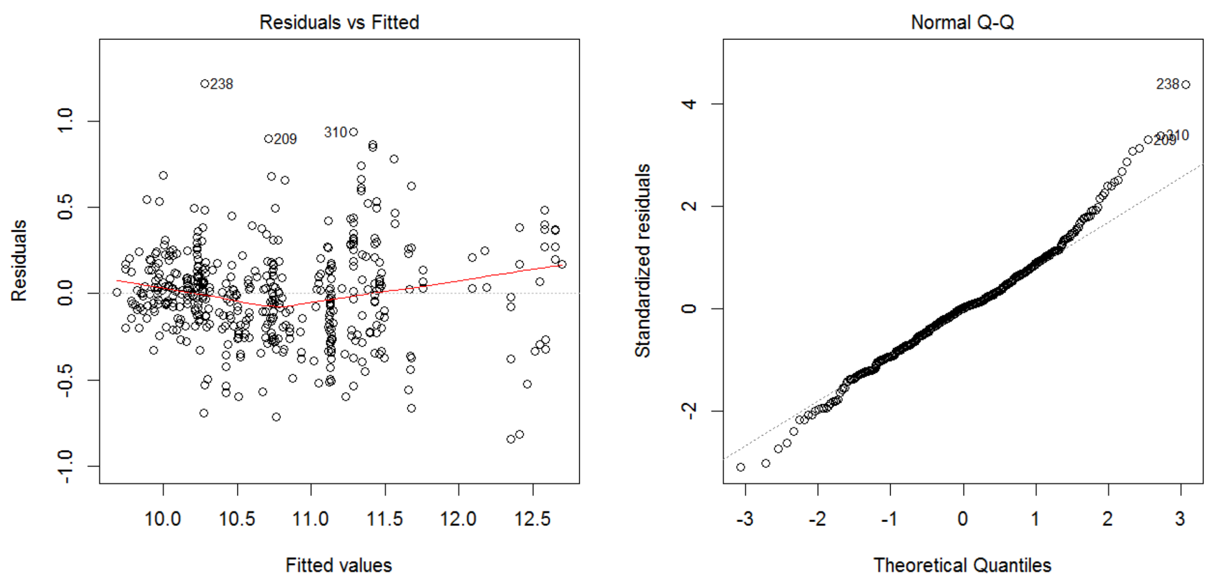


Figure 4.8: Residual plots of the quadratic model after using BIC backward elimination

From Figure 4.8, the residuals seem to have a more random pattern (red line less curved) compared to that of Figure 4.7, indicating an improvement in linearity. The residuals are evenly and distributed along the 0 line, indicating that the assumption of equal variance of the residuals may hold. The normal Q-Q plot shows that most of the points fall on the Q-Q line despite that there is still a little tail on the right hand side, indicating that the normality assumption may also hold.

A	B	C	I(A^2)	Origin	Luxury	A:B	B:C
280.96	159.35	15.34	92.28	2.29	2.14	25.12	105.08

Table 4.10: VIF test of quadratic model after using BIC backward elimination

The only question is the VIF table shown in Table 4.10, the VIF test suggests some multicolinearity problem. Yet, quadratic terms and interaction effects are often correlated with their respective original terms. The sign of large VIF can be explained by doing another regression with the means of A, B, C subtracted from original A, B, C respectively. The subtracted mean terms are indicated as “subA”, “subB”, and “subC”.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.702124	0.04321	247.678	**< 2e-16**
subA	0.148593	0.01572	9.452	**< 2e-16**
subB	-0.006597	0.005665	-1.164	0.244855
subC	-0.138649	0.022714	-6.104	**2.27E-09**
I(subA^2)	0.02229	0.004255	5.239	**2.51E-07**
OriginITALY	0.082457	0.110859	0.744	0.457394
OriginJAPAN	-0.39783	0.040629	-9.792	**< 2e-16**
OriginSOUTH KOREA	-0.378244	0.068116	-5.553	**4.86E-08**
OriginSWEDEN	-0.0923	0.106866	-0.864	0.388222
OriginUK	-0.100738	0.056564	-1.781	**0.075605**
OriginUS	-0.440576	0.044348	-9.935	**< 2e-16**
Luxury	0.465697	0.038275	12.167	**< 2e-16**
subA:subB	0.010127	0.002933	3.453	**0.000608**
subB:subC	0.020058	0.002714	7.392	**7.34E-13**
Multiple R-squared: 0.8566			Adjusted R-squared: 0.8524	

Table 4.11: Summary table of the quadratic model with the means subtracted

subA	subB	subC	I(subA^2)	Origin	Luxury	subA:subB	subB:subC
5.84	**6.22**	1.65	**5.56**	2.29	2.14	**8.26**	2.71

Table 4.12: VIF test of the quadratic model with the means subtracted

As we see from Table 4.12, despite the fact that some of the VIF values are slightly greater than 5, they are considered within the acceptance level compared to that of Table 4.10.

Since the two models in Table 4.9 and 4.11 are similar in essence, we confirm that the quadratic model in Table 4.9 is valid and its assumptions of regression model are held.

The new final quadratic model is as follow:

$$\log MSRP = 15.50 - 0.35 \cdot Cyl - 0.19 \cdot MPG - 0.58 \cdot APS + 0.02Cyl^2 + 0.08 \cdot O_1 - 0.40 \cdot O_2 - 0.38 \cdot O_3 - 0.09 \cdot O_4 - 010 \cdot O_5 - 0.44 \cdot O_6 + 0.47 \cdot Luxury + 0.01 \cdot Cyl * MPG + 0.02 \cdot MPG * APS$$

where,

$Origin\ levels\text{:}\ \ O_1\text{:}\,Italty\ \ ;\ \ O_2\text{:}\,Japan\ \ ;\ \ O_3\text{:}\,South\ Korea\ \ ;\ \ O_4\text{:}\,Sweden\ \ ;\ \ O_5\text{:}\,UK\ \ ;\ \ O_6\text{:}\,US$

$Origin\ level = France\ \ ,\ \ when\ \ O_1 = O_2 = O_3 = O_4 = O_5 = O_6 = 0$

$Cyl * MPG\text{:}\ \ interaction\ effects\ of\ MPG\ on \log(MSRP)\ \ depend\ on\ Cyl$

$MPG * APS\text{:}\ \ interaction\ effects\ of\ MPG\ on \log(MSRP)\ \ depend\ on\ APS$

CHAPTER 5

Vehicle Eliminator Excel Template

The idea of Vehicle Eliminator Excel Template is to provide an offline platform for people to easily access the 2012 vehicle data at any time. Without using sophisticated statistical methods or programming, the Vehicle Eliminator Excel Template demonstrates high-quality presentation of complex data by utilizing advanced 2010 Excel commands as well as data visualization techniques.

Vehicle Eliminator is to help the people ease the choosing process from a large pool of 2012 vehicle models and allow them to easily compare between vehicles with their own preferences. This excel template shares the same data pool with the thesis. It urges users to choose their own preferences step by step while automatically eliminating the unwanted vehicles simultaneously. The result table eventually generates a list of vehicle models based on the user's preferences.

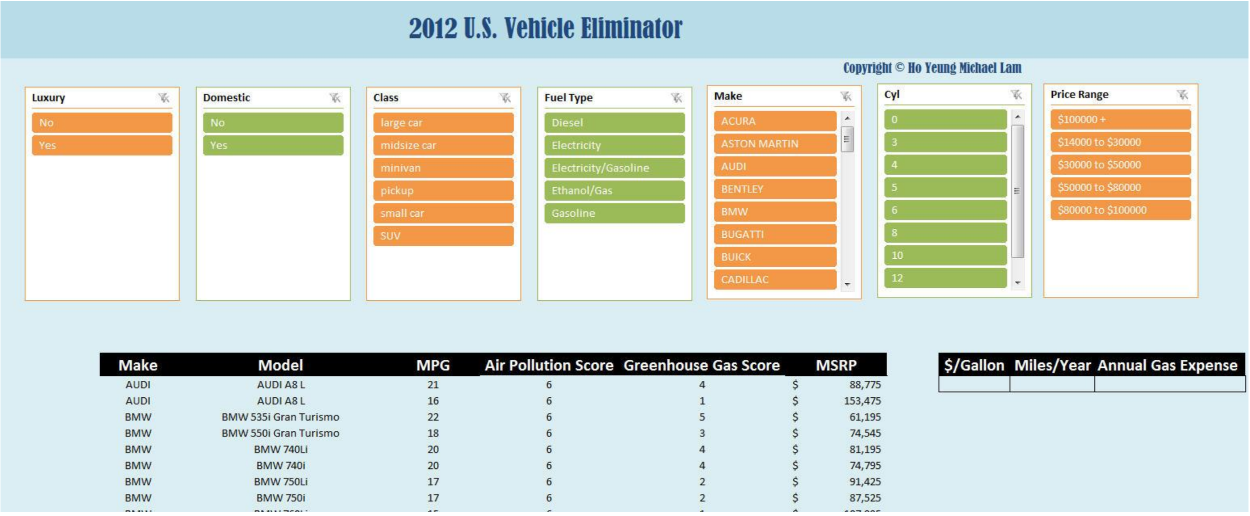
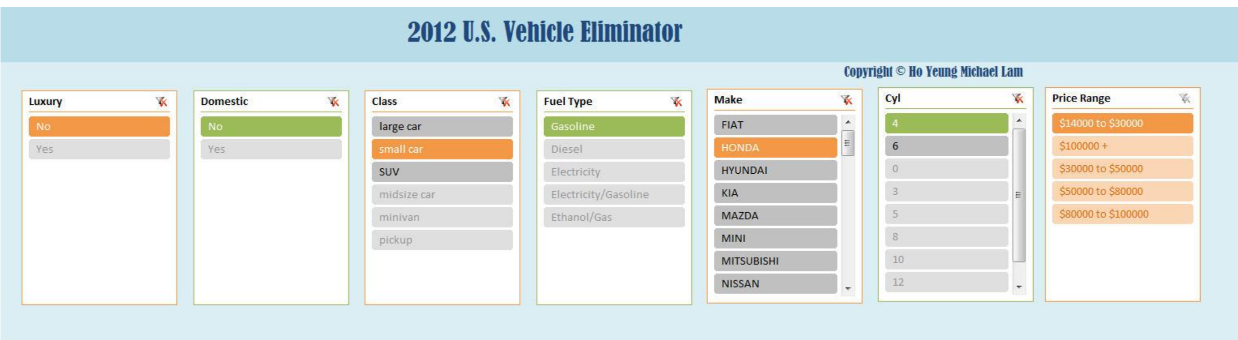


Figure 5.1: Vehicle Eliminator Snapshot #1

One of the most important techniques used in this template is Slicers, which is one of the latest techniques of Excel 2010. Slicers are user-friendly visual controls that associate with pivot tables, allowing users to instantly filter the data in an interactive way. From Figure 5.1, a snapshot of the template shows the Slicers which are indicated in orange and green color. The information shown in the result table includes the average combined local/highway miles per gallon (MPG), Air Pollution Score, Greenhouse Gas Score, and the average manufacturer's suggested retailed price (MSRP). There is also a little box on the right hand side of the result table, allowing the user to predict the respective gas expenses of the first car on the result table by manually inputting the cost of gas per gallon as well as the average miles driven per year.

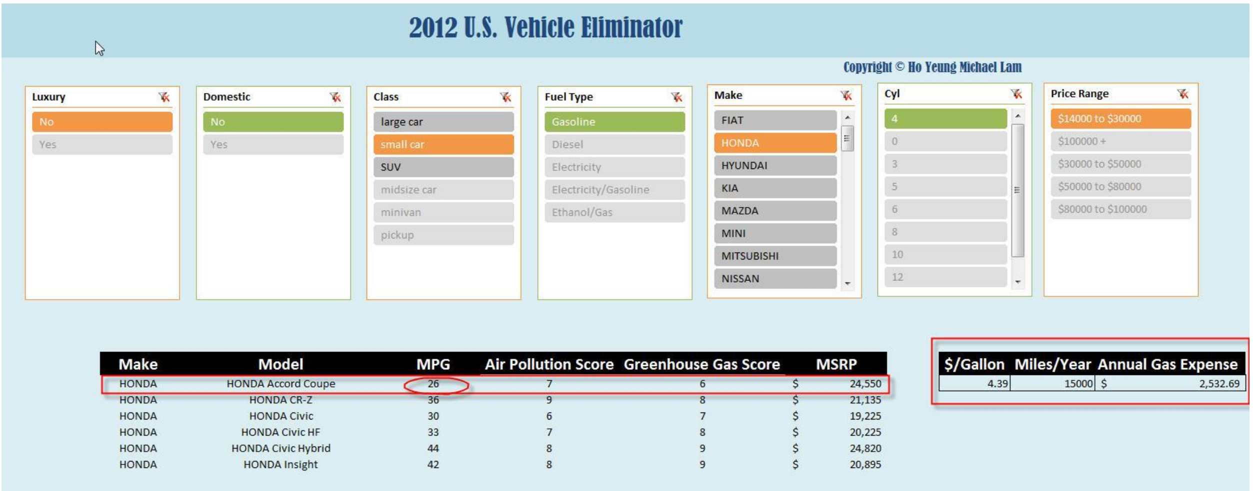


Make	Model	MPG	Air Pollution Score	Greenhouse Gas Score	MSRP
HONDA	HONDA Accord Coupe	26	7	6	$ 24,550
HONDA	HONDA CR-Z	36	9	8	$ 21,135
HONDA	HONDA Civic	30	6	7	$ 19,225
HONDA	HONDA Civic HF	33	7	8	$ 20,225
HONDA	HONDA Civic Hybrid	44	8	9	$ 24,820
HONDA	HONDA Insight	42	8	9	$ 20,895

Figure 5.2: Vehicle Eliminator Snapshot #2

Figure 5.2 shows the example of filtering the vehicle models with Slicers. The user is prompted to select their preferences in the Slicers from left to right. The result table lists all the vehicle models that meet the user's favor.

Figure 5.3 shows how the little box on the right predicts the annual gas expense. Simply inputting the cost of gas per gallon and the number of miles driven per year, the annual gas expense would be calculated based on the MPG of the first vehicle model listed on the result table.



Make	Model	MPG	Air Pollution Score	Greenhouse Gas Score	MSRP
HONDA	HONDA Accord Coupe	26	7	6	$ 24,550
HONDA	HONDA CR-Z	36	9	8	$ 21,135
HONDA	HONDA Civic	30	6	7	$ 19,225
HONDA	HONDA Civic HF	33	7	8	$ 20,225
HONDA	HONDA Civic Hybrid	44	8	9	$ 24,820
HONDA	HONDA Insight	42	8	9	$ 20,895

$/Gallon	Miles/Year	Annual Gas Expense
4.39	15000	$ 2,532.69

Figure 5.3: Vehicle Eliminator Snapshot #3

If anyone is interested in this Vehicle Eliminator Excel Template, the template is available upon email request. (Email: hoyeung.lam@stat.ucla.edu)

CHAPTER 6

Summary

1. Non-luxury vehicles are generally cleaner and more fuel efficient than luxury vehicles.
2. Among the non-luxury make sector, none of the vehicles from Dodge, Jeep, and Volvo earn a SmartWay designation, indicating that their vehicles are not excellent environmental performers.
3. Among the luxury make sector, domestic vehicles generally have higher fuel efficiency than foreign vehicles in all kinds of vehicle classes.
4. Among the non-luxury make sector, foreign vehicles are more environmentally friendly and fuel economy than domestic vehicles in all the vehicle classes except SUV.
5. The emissions of carbon dioxide and greenhouse gases vary by fuel type, since each fuel type contains a different amount of carbon. Among the alternative fuel types, electricity is the cleanest fuel, followed by diesel, and ethanol fuel mixtures.
6. MPG is highly correlated with Greenhouse Gas Score, this might due to vehicles with higher fuel economy burn less fuel to travel the same distance, and as a result less greenhouse gases are emitted. However, MPG is only fairly correlated with Air Pollution Score, the amount of air pollutants might be more related to the types of vehicle engine.
7. The linear regression model shows that the performance of the cylinder seems to have relative large effect on MSRP compared to other variables. It seems that MSRP tends to increase as the performance of the cylinder is greater. However, MPG is not considered in the model due to multicolinearity.

8. The final quadratic regression model suggests that the variable Cyl has a significant quadratic effect on MSRP. Besides, there are significant interaction effects between MPG and Cyl as well as MPG and APS, indicating that MSRP depends quite heavily on MPG.
9. The quadratic regression also suggests that the interaction effect between average miles per gallon and average air pollution score seems to have a slightly greater effect on MSRP than the interaction effect between average miles per gallon and numbers of cylinder.

REFERENCES

1. United States. Environmental Protection Agency. Office of Transportation and Air Quality. *Green Vehicle Guide*. Web. May 2012.
2. "U.S. Autos Account for Half of Global Warming Linked to Cars Worldwide." *About.com*. Ed. Larry West. Web. May 2012. <http://environment.about.com/od/globalwarming/a/autoemissions.htm>.
3. "World of Change: Global Temperatures." *Http://earthobservatory.nasa.gov*. National Aeronautics and Space Administration (NASA). Web. May 2012.
4. "Auto Sales." Rev. of *Cars & Light Trucks in the U.S.* The Wall Street Journal. Web. May 2012.
5. *Edmunds.com*. Edmunds. Web. May 2012.