Title
Predicting Who Will Cover the Spread in NFL Games

Author
Patel, Ajay Rakesh

Publication Date
2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Predicting Who Will

Cover the Spread in NFL Games

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Applied Statistics

by

Ajay Rakesh Patel

2023

ABSTRACT OF THE THESIS

Predicting Who Will
Cover the Spread in NFL Games

by

Ajay Rakesh Patel
Master of Science in Applied Statistics
University of California, Los Angeles, 2023
Professor Frederic R. Paik Schoenberg, Chair

Sports betting can be lucrative for some while unfavorable for others, but what if you could tilt the odds in your favor? This thesis helps uncover whether or not machine learning can accurately predict who will cover the spread in NFL games. I investigated which combination of game statistics, box scores, power ranks, and Elo ratings would yield the best results. Additionally, I tested three different machine learning models with varying levels of interpretability and predictability. In the end, I found that I can correctly predict who will cover the spread enough times to slightly tilt the odds in my favor.

The thesis of Ajay Rakesh Patel is approved.

Miles Chen

Michael Tsiang

Ying Nian Wu

Frederic R. Paik Schoenberg, Committee Chair

University of California, Los Angeles

2023

To my family and friends

that supported me along the way.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

Sports betting has become increasingly popular over the last few years. As of January 2023, there are 36 states that offer some legal form of sports betting[1]. With plenty of sportsbooks, various ways to bet, and an abundance of games to bet on, the opportunity to get rich quick has never been so prevalent. In the NFL, betting on which team will win a game is not so difficult, but when a bettor has to determine if Team A will beat Team B by a certain number of points, betting can become more challenging.

At the start of each week, sportsbooks set the spread for every NFL game. For the Super Bowl, BetMGM set the initial spread at 1.5 points in favor of the Kansas City Chiefs[2]. This means BetMGM predicts the Kansas City Chiefs will beat the Philadelphia Eagles by 1.5 points. If the Chiefs win the Super Bowl by 2 or more points, the Chiefs will have covered the spread. However, if the Eagles win or lose by 1 point, the Eagles will have covered the spread. From here, a bettor has to determine the outcome of the game and place their bet accordingly.

Throughout the week, sportsbooks update the spread based on people's bets. If more people bet on the favorite than on the underdog, sportsbooks increase the spread. This encourages future bettors to bet on the underdog. If more people bet on the underdog than on the favorite, sportsbooks decrease the spread which encourages future bettors to bet on the favorite. For each game, sportsbooks want the wagers of bets on both teams to be equal, or in other words, sportsbooks want the total amount of money bet on the underdog to be

[1] https://www.cbssports.com/general/news/u-s-sports-betting-heres-where-all-50-states-stand-on-legalizing-sports-gambling-betting-mobile-bets/

[2] https://sports.betmgm.com/en/blog/nfl/super-bowl-spread-open-bm07/

equal to the total amount of money bet on the favorite[3]. Let's say Person A bet on the favorite to cover the spread, and Person B bet on the underdog. Both bettors had 10 to 11 odds, meaning both bettors gambled \$11 to win \$10. Note, sportsbooks always offer 10 to 11 odds when betting on the spread. Now, let's say the favorite covered the spread. In this example, the sportsbook received \$22 total. Person A gets their \$11 back for winning the bet plus \$10 from Person B, who lost the bet. The sportsbook keeps the extra \$1 as profit[4]. By keeping the wagers equal, the losers of the bet payout the winners, and a sportsbook keeps all the profits at no risk. After BetMGM set the initial spread at 1.5 points in favor of the Chiefs, people wagered more money on the Eagles to cover the spread than on the Chiefs. Two days later, the Eagles emerged as 1.5 point favorites for the Super Bowl. Note, when a bettor places their bet, the spread remains fixed. It will stay the same even when sportsbooks update the spread for future bettors.

Now, it may seem that if you can correctly pick who will cover the spread with 51% or 52% accuracy, that would be enough to make a profit in the long run. However, that is not the case because of the payout structure. Let's assume you are given 10 to 11 odds and both teams have a 50% chance of covering the spread ($p = 0.5$). The Expected Value (EV) of your bet is:

$$EV = p \times Payout - (1 - p) \times Amount\ Gambled$$

$$EV = 0.5 \times \$10 - (1 - 0.5) \times \$11$$

$$EV = -0.50$$

This means you are expected to lose money. So, how often does a bettor have to be correct to break even? Let's solve for p from the equation above with 10 to 11 odds. In this scenario, p is the proportion of time a bettor correctly picks who will cover the spread.

[3] https://medium.com/@PhilipAndrews/how-a-sportsbook-stays-in-business-72e5f3b5243d

[4] https://sports.betmgm.com/en/blog/how-to-read-nfl-betting-odds-online-spread-over-under-money-line/

$$0 = p \times \$10 - (1 - p) \times \$11$$

$$p = \frac{\$11}{(\$11 + \$10)}$$

$$p = 0.5238$$

Thus, to break even, a bettor must pick who will cover the spread with at least 52.38% accuracy[5]. Some people have a knack for picking which team or teams will cover the spread each week. Others might attribute their winnings or losses to skill, knowledge, and/or luck. But, can we obtain 53% accuracy and beat Las Vegas in the long run with machine learning? If so, what variables are needed to create a successful model? Can a simple linear regression model perform as well as complex models? In this paper, I will collect data from different resources, combine the data together, determine which variables are best, and find the optimal model to predict whether the underdog or the favorite will cover the spread across all NFL games.

[5] `https://medium.com/the-intelligent-sports-wagerer/why-52-4-is-the-most-important-percentage-in-sports-gambling-16ade8003c04`

CHAPTER 2

Data Collection

To predict whether the underdog or the favorite will cover the spread, I used 4 main sources of data: the historical spreads for all NFL games, ESPN box scores and team statistics, ESPN power ranks, and FiveThirtyEight Elo ratings. For each source, I collected the weekly data on every team from the 2016 season through the 2021 season. Then, I merged all four datasets together into 1 dataset. In the final dataset, each observation is a game, and I have the box scores, team statistics, power ranks, and Elo ratings for both the underdog team and the favorite team.

2.1 Spread

First, I needed the historical spreads of each NFL game dating back to 2016. Unfortunately, individual sportsbooks keep their historical spreads private. However, Sports Odds History records the historical closing spreads for each NFL game[6]. The spread listed on their website is the average of the closing spreads listed on 6 different sportsbooks (PointsBet, BetMGM, Caesars, Unibet, DraftKings, and FanDuel). In addition to the average spread, Sports Odds History also records which team covered the spread, the score, the average Over/Under, as well as the day, date, and time of the games.

[6] `https://www.sportsoddshistory.com/nfl-game-season/?y=2016`

2.2 ESPN

In addition to the historical spreads, I collected game data from ESPN. I scraped the number of wins, the amount of points scored in each quarter[7], and the post game team statistics. The postgame team statistics include the number of first downs, passing first downs, rushing first downs, first downs by penalty, plays, total yards, passing yards, rushing yards, passing attempts, rushing attempts, interceptions, sacks, yards lost on sacks, penalties, and defensive / special teams touchdowns as well as the third down, fourth down, and red zone conversion rate[8].

2.3 Power Ranks

Additionally, I gathered the weekly power ranks for each team from ESPN. Every week, a panel of more than 80 writers, editors, and TV personalities rank the NFL teams from 1 to 32[9]. The weekly power ranks help assess how good a team is and evaluate particular aspects of each game. For example, if a team jumps out to an early lead, the opposing team will likely have to pass the ball more to get the score closer. After the game, the team statistics may indicate the opposing team is a great passing team and an ineffective running team. But in reality, this assumption may not be true. In the previous weeks, the opposing team could have had a more balanced passing yards to rushing yards ratio. Therefore, the opposing team's power rank might stay the same that week.

In addition, the weekly power ranks can help quantify how healthy a team is and each team's momentum. Teams with more injuries are more likely to lose their upcoming game because less experienced players will have to be on the field more. Thus, this could shift their power rank down. And on the flip side, a team that has won a couple games in a row should see a shift up in their power rank. The hope is that each team's power rank will

[7] `https://www.espn.com/nfl/scoreboard/_/week/1/year/2016/`

[8] `https://www.espn.com/nfl/matchup/_/gameId/400874484`

[9] `https://www.espn.com/nfl/story/_/id/18395280/nfl-2016-final-regular-season-power-rankings-new-england-patriots-dallas-cowboys-pittsburgh-steelers`

provide additional predictability beyond the numeric team statistics.

2.4 Elo Ratings

Lastly, I gathered FiveThirtyEight's Elo ratings for each team and quarterback[10]. In general, the Elo rating system calculates the relative skill of players in zero sum games such as chess. FiveThirtyEight adapted the method to assign each team an Elo rating, give each team a probability of winning their upcoming game, and assess how well each team's quarterback has been playing throughout the season[11]. All of these metrics can help predict who will cover the spread.

In FiveThirtyEight's Elo ratings, each team started with a score of 1500, and after each game, a team's rating changed based on the game's outcome and margin of victory. Once a season is over, a team's ending Elo rating serves as their initial Elo rating for the following season, but their rating is shifted one-third closer to the league average Elo rating. This methodology accounts for the NFL draft, free agency, and other offseason acquisitions.

Over the years, FiveThirtyEight has added features to adjust for home-field advantage, how many rest days each team has had, and which quarterback will be starting in a given game. After calculating the Elo ratings for each team, each team is assigned a probability of winning their upcoming game with the formula below:

$P(A) = \frac{1}{10^{\frac{-EloDiff}{400}}+1}$, where *EloDiff* is the difference in Elo ratings between Team A and Team B.

For quarterback play, FiveThirtyEight created the following formula to assign each quarterback a value:

Value = -2.2 × *Passing Attempts* + 3.7 × (*Completions* + $\frac{PassingYards}{5}$) + 11.3 × *Passing TDs* + 14.1 × *Interceptions* - 8 × *Times Sacked* - 1.1 × *Rush Attempts* + 0.6 × *Rushing Yards* + 15.9 × *Rushing TDs*

[10] https://github.com/fivethirtyeight/data/tree/master/nfl-elo

[11] https://fivethirtyeight.com/methodology/how-our-nfl-predictions-work/

Every week, each quarterback value is adjusted based on how good the opponent defense is. Lastly, the quarterback value is calculated over a 10 game window with the following formula:

$$\textit{New Rating} = 0.9 \times \textit{Old Rating} + 0.1 \times \textit{Value}$$

CHAPTER 3

Feature Engineering

Initially, I planned to use each team's game statistics, power rank, and Elo ratings from the previous week as variables for predicting the current week's outcome. After further thought, I decided against this approach. Due to the variance in a team's performance and opponents from week to week, I needed variables describing how teams perform over time. So, I used a four week rolling average of all my variables to predict the current week's outcome[12]. Thus, my first predictions start in week 5 because I used weeks 1 through 4 to calculate the rolling averages. For week 6, I generated the rolling averages with weeks 2 through 5. This continued until the season ended. For each new season, my first predictions start with week 5. Data from the end of a season does not contribute to the following season. Ultimately, this method reduced the total number of observations in my dataset from 1424 to 1142 observations.

Rolling Average	Predicting For
Weeks 1 - 4, 2016	Week 5, 2016
Weeks 2 - 5, 2016	Week 6, 2016
...	...
Weeks 13 - 16, 2016	Week 17, 2016
Weeks 1 - 4, 2017	Week 5, 2017
...	...

Table 3.1: Rolling Average Calculation

After calculating the four week rolling average, I noticed some variables did not have

[12] `https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.rolling.html`

normally distributed values - specifically, variables with low frequency counts such as, the *Underdog's Number of First Downs by Penalty* and the *Underdog's Number of Interceptions Thrown*. Instead of using the rolling average of these variables, I converted them into binary categorical variables. So, the *Underdog's Number of First Downs by Penalty* became *Did the Underdog Have At Least 8 First Downs by Penalty in the Last Four Games?* and the *Underdog's Number of Interceptions Thrown* became *Did the Underdog Throw At Least 4 Interceptions in the Last Four Games?* During the conversion process, I made the frequency count of each category as equal as possible.

Lastly, I created 2 additional variables - one called *Underdog Opponent Strength* and the other called *Favorite Opponent Strength*. Both use the same calculation - the only difference being one is for the underdog and the other is for the favorite. I defined *Opponent Strength* as the average number of wins a team's previous four opponents had prior to playing that particular team. For example, this week, Team A is playing Team B. In the past 4 weeks, Team A played Teams C, D, E, and F. Prior to Team A's matchup with each team, let's say Team C had 5 wins, Team D had 6 wins, Team E had 9 wins, and team F had 8 wins. This means the average number of wins among Team A's opponents before playing Team A is 7. The hope is that this variable will provide some insight into how difficult it has been for Team A to cover the spread the last 4 weeks.

CHAPTER 4

Exploratory Data Analysis

Now, I will analyze some of the variables in my final dataset. For histograms and boxplots of the remaining variables, please refer to the Appendix.

4.1 Who Covered the Spread

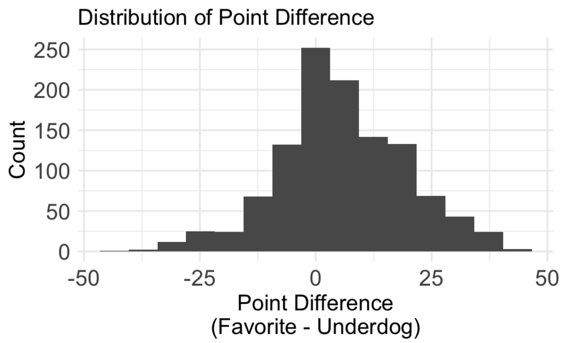
According to Table 4.1 below, we see that the favorite covered the spread nearly just as much as the underdog since 2016. This could also indicate that Las Vegas sportsbooks are nearly perfect at setting the spread.

Favorite	Underdog
567	575

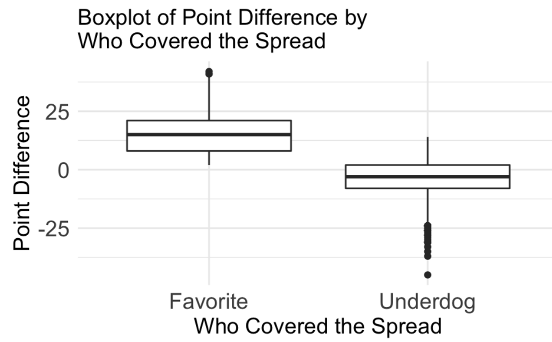
Table 4.1: Who Covered the Spread

4.2 Point Difference

In addition to nearly 50% of the favored teams covering the spread, we see in Figure 4.1 below, that the point difference in games between the favorite and underdog is approximately normally distributed. We see that when the underdog covered the spread, they only won the game about 25% of the time. When the favorite covered the spread, the point difference was greater than 3 touchdowns, or 21 points, nearly 25% of the time.



(a) Distribution of Point Difference

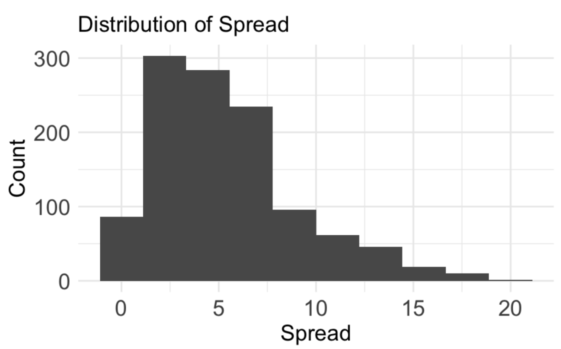


(b) Boxplot of Point Difference by Who Covered the Spread

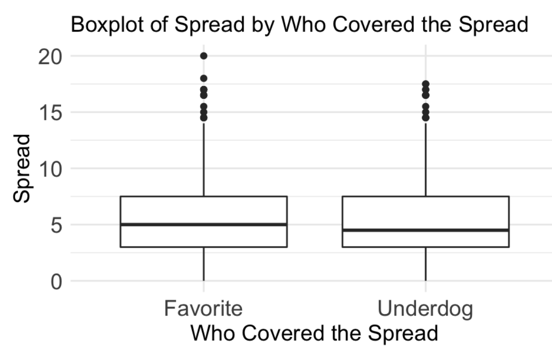
Figure 4.1: Point Difference

4.3 Spread

According to the histogram on the left, we see that sportsbooks usually set the spread between 0 and 7 points. This makes sense because, in Figure 4.1, the point difference is frequently between -7 and 7 points. According to the boxplots below, the favorite and the underdog covered the spread equally despite what sportsbooks set the spread at.



(a) Distribution of Spread

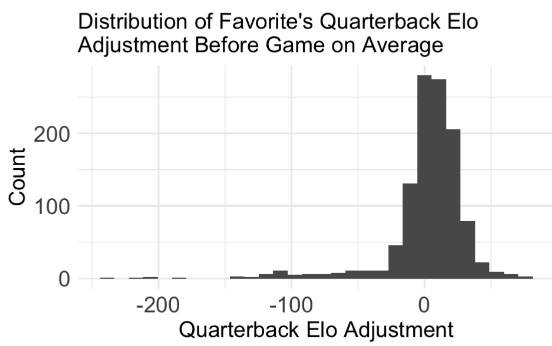


(b) Boxplot of Set Spread by Who Covered the Spread

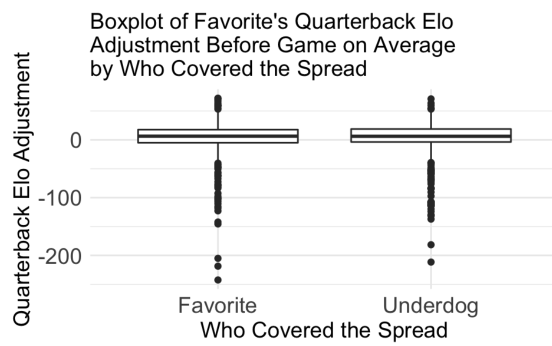
Figure 4.2: Spread

4.4 Favorite's Quarterback Elo Adjustment

Each week, FiveThirtyEight assigned each team an Elo Rating. Then, FiveThirtyEight adjusted the ratings with different metrics - one being how a team's quarterback has performed in the last 10 games. According to Figure 4.3, FiveThirtyEight adjusted the favorite team's quarterback Elo rating up, on average. On occasion, however, FiveThirtyEight adjusted the rating downward by more than 100 points. This occurred when the favorite team ruled the typical starting quarterback out for the upcoming game. Again, we see that this adjustment did not make a significant impact on which team covered the spread.



(a) Distribution of Favorite's Quarterback Elo Adjustment

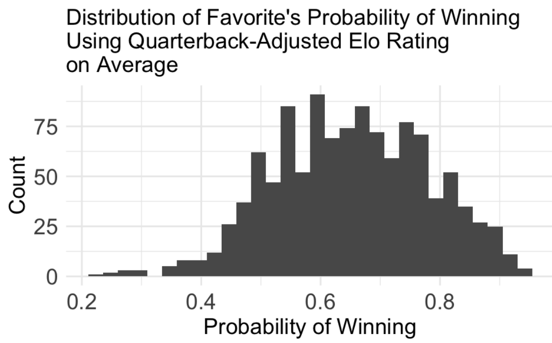


(b) Boxplot of Favorite's Quarterback Elo Adjustment by Who Covered the Spread

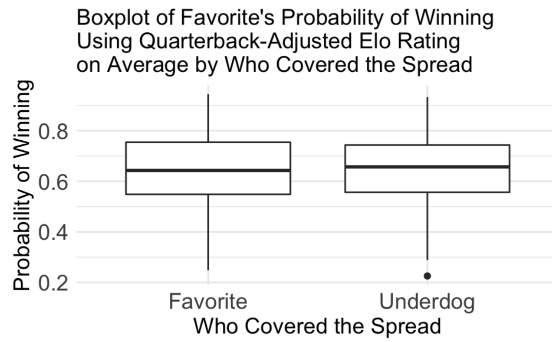
Figure 4.3: Favorite's Quarterback Elo Rating Adjustment

4.5 Favorite's Probability of Winning Using Quarterback-Adjusted Elo Rating

The favorite's probability of winning using the quarterback-adjusted Elo rating, on average, appears roughly normally distributed. Interestingly, even though sportsbooks have deemed these teams as the favorites, FiveThirtyEight still assigned some teams a probability of winning less than 50%, and sometimes, as low as 30%, on average. Once more, there is not a noticeable difference between who covered the spread and the favorite's assigned probability of winning, on average.



(a) Distribution of Favorite's Probability of Winning Using Quarterback-Adjusted Elo Rating on Average

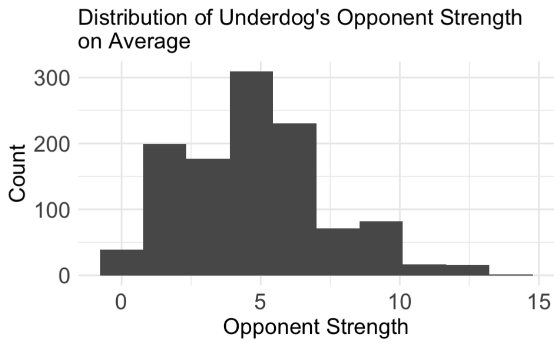


(b) Boxplot of Favorite's Probability of Winning Using Quarterback-Adjusted Elo Rating on Average by Who Covered the Spread

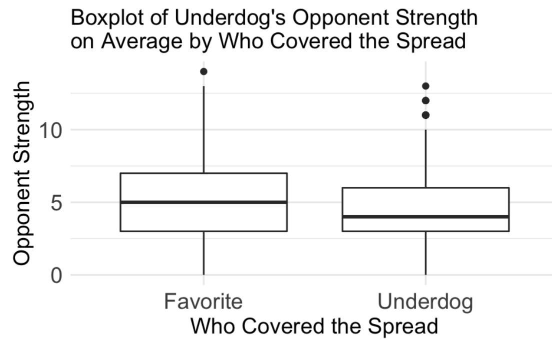
Figure 4.4: Favorite's Probability of Winning Using Quarterback-Adjusted Elo Rating

4.6 Underdog's Opponent Strength

Over the previous four games, the underdog's opponents have had at least 5 wins on average. This makes sense because teams' win-loss records are closer together than they are distinct, especially in the middle of the season. Each year, only a handful of teams have a dominating record. According to the boxplots, the favored team covered the spread more often when the underdog's opponent strength was greater than 10 wins, on average.



(a) Distribution of Underdog's Opponent Strength



(b) Boxplot of Underdog's Opponent Strength by Who Covered the Spread

Figure 4.5: Underdog's Opponent Strength

CHAPTER 5

Methodology

In this section, I will discuss the approach taken to predict who covered the spread in each NFL game. First, I will discuss how I split the data into a training set, validation set, and test set. Then, I will share which variable I used as the response variable during the modeling process. Lastly, I will explain the different feature sets I created, the different machine learning models considered, and how I evaluated the feature sets and models together.

5.1 Train-Validation-Test Split

First, I used data from the 2021 season as my test set. Data from the 2021 season was completely unseen during the training process. It was not used until I found the optimal model and feature set. Then, I randomly split 80% of the remaining data from the 2016-2020 seasons into a training set and standardized each column[13]. I used the other random 20% as a validation set. Next, I standardized the validation set and test set with the training set's variable means and standard deviations[14]. During training, I used 5-fold cross validation to evaluate my models and feature sets[15]. Lastly, I re-evaluated the models with the validation set and checked if the models overfit the training data.

[13] `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html`

[14] `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html`

[15] `https://scikit-learn.org/stable/modules/cross_validation.html`

5.2 Response Variable and Evaluation

Initially, I started with the binary variable, *Who Covered the Spread* as the response variable. However, I did not have much success with my models. The cross validation accuracy plateaued at 52%. So, I switched the response variable to the point difference (*Point Difference = Favorite Team Score - Underdog Team Score*). Then, I compared the spread to the predicted point difference for each game. If the predicted point difference was less than the spread, I predicted the Underdog would cover the spread. Otherwise, I predicted the Favorite would cover the spread. Plus, with a numeric response variable, I could compare the predictions to the true point difference, what the spread was set at for each game, and potentially find a subset of games to bet on.

Since the response variable is numeric, I used RMSE as my metric to assess model and feature set performance during cross validation[16]. But first, I redefined the RMSE formula. By default, the residuals would have been calculated as *Observed Point Difference - Predicted Point Difference.* This would not help someone decide which team to bet on because technically, one would not know the observed point difference until after the game. At that point, a bettor cannot place bets on the underdog or the favorite. Instead, I redefined the residuals as *Spread - Predicted Point Difference.* Both the spread and predicted point difference are known before a game takes place, allowing one to place a bet on the underdog or favorite.

5.3 Feature Selection

In the training data, there are 98 variables measuring the underdog's and favorite's performance over the previous four weeks. It would not have been optimal to fit models with all of the variables available. Instead, I investigated which set of variables were the best combination of predictors. In total, I tested six combinations of variables. Each feature set

[16] `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html`

attempted to eliminate noisy variables and only keep important ones. Hopefully then, the models would detect patterns between the remaining variables and who covered the spread.

1. Remove Correlated Variables

If variables had greater than 0.5 correlation with another variable, I kept one variable amongst the group of correlated variables for this feature set. Ultimately, 42 variables remained after eliminating the correlated variables.

2. Top Correlated Variables with Response

I calculated the correlation between the point difference and each predictor variable. Then, I kept the top 30% of the most correlated variables with the response variable for this feature set. In this case, the top correlated variables had a correlation between $|0.079|$ and $|0.296|$.

3. Remove Redundant Variables

Among the variables, some linear dependencies exist. For example, the sum of *Passing Yards* and *Rushing Yards* is equivalent to *Total Yards*. Thus, it is not necessary to keep all three variables. In this case, I only kept *Passing yards* and *Rushing yards*. I went through the variables, subset them down to a set of non linearly dependent variables, and removed variables that contained repetitive information for this feature set.

For each feature set listed above, I created an additional feature set that contained each variable, X_i, and X_i^2. During the exploratory data analysis, the quadratic terms revealed small differences in who covered the spread. Therefore, there was reason to believe including the quadratic terms could help with predictions. Thus, I created the three additional feature sets. Please refer to the Appendix to see which variables are in each feature set.

5.4 Models

After creating the feature sets above, I tested three different models: Elastic Net Linear Regression, Random Forest, and XGBoost. Typically, Linear Regression models are more interpretable than Random Forests and XGBoosts. However, Random Forests and XGBoosts are more predictive. I tested all three models with each feature set to see which combination of interpretability, predictability, and feature set would yield the best results.

Each model had at least one hyperparameter that needed tuning. For the Elastic Net Linear Regression model, I tuned the penalty term (Alpha) and the L1-to-L2 regularization ratio with each feature set[17]. For the Random Forest, I tuned the number of trees, the maximum depth of the tree, the minimum number of observations needed to split each node, and the minimum number of observations needed to be considered a leaf node[18]. Lastly, for the XGBoost, I tuned the same hyperparameters as the Random Forest as well as the learning rate[19]. I tuned the hyperparameters with 10-fold cross validated Grid Search[20]. Then, I refit the model, with the optimal hyperparameters, on the entire training set.

The following table describes the range of values used to tune each hyperparameter in the models:

Hyperparameter	Elastic Net	Random Forest	XGBoost
Alpha	[1, 2, 3]		
L1-to-L2 Ratio	[0.1, 0.2, ... , 0.9]		
No. of Trees		[250, 500, 750]	[250, 350, 450]
Max Depth		[None, 2, 3]	[None, 2, 3]
Min. Samples Split		[2, 3, 4]	[2, 3, 4]
Min. Samples Leaf		[1, 2, 3]	[1, 2, 3]
Learning Rate			[0.1, 0.01, 0.001]

Table 5.1: Hyperparameters

[17] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.html

[18] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

[19] https://scikit-learn.org/stable/auto_examples/ensemble/plot_gradient_boosting_regression.html

[20] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

CHAPTER 6

Results and Discussion

After hours of tuning each feature set and model combination, I obtained the 5-fold cross validation RMSE values and validation set accuracies listed in Table 6.1.

Feature Set	Metric	Elastic Net	Random Forest	XGBoost
Remove Correlated	Training Set Cross Validation	1.1404	1.5578	0.3358
Variables	Validation Set Accuracy	0.4628	0.4628	0.4787
Top Correlated Variables	Training Set Cross Validation	1.4727	0.9545	0.1552
	Validation Set Accuracy	0.4840	0.5106	0.5745
Remove Redundant	Training Set Cross Validation	1.1404	1.4397	2.2805
Variables	Validation Set Accuracy	0.4840	0.5106	0.4415
Remove Correlated	Training Set Cross Validation	1.1826	1.1454	0.3362
Variables (X_i & X_i^2)	Validation Set Accuracy	0.4940	0.4947	0.4787
Top Correlated Variables	Training Set Cross Validation	1.1597	0.7974	0.1551
(X_i & X_i^2)	Validation Set Accuracy	0.5053	0.5319	0.5851
Remove Redundant	Training Set Cross Validation	1.1059	1.1528	2.2805
Variables (X_i & X_i^2)	Validation Set Accuracy	0.4840	0.5372	0.4415

Table 6.1: Training Set and Validation Set Results

Many of the feature set and model combinations yielded RMSE values within two points of the set spread. Four out of the six XGBoost models have a cross-validated RMSE within half a point of the set spread. Despite the promising training set cross-validation results, many of the models underperformed on the validation set in terms of accuracy. All of the ElasticNet Regression models have accuracies less than 50%, and only two Random Forest Models obtained 53% accuracy on the validation set. However, two of the XGBoost models performed beyond expectations, both achieving greater than 57% accuracy. In fact, the

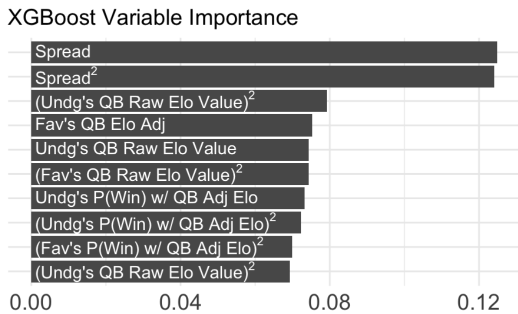
XGBoost model with the Top Correlated Variables (X_i & X_i^2) earned 58.5% accuracy and had the smallest RMSE among all models and feature set combinations. Table 6.2 has the optimal hyperparameters for this model.

No. of Trees	Max Depth	Min. Samples Split	Min. Samples Leaf	Learning Rate
350	3	2	1	0.01

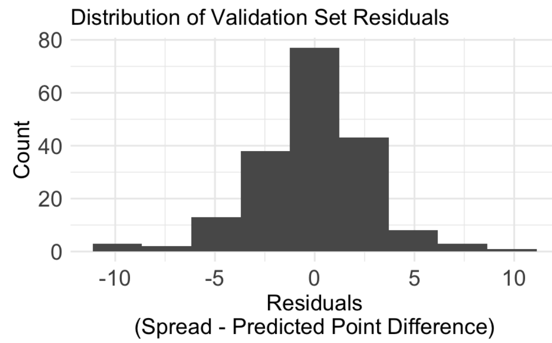
Table 6.2: Hyperparameter Results

According to Figure 6.1, the most important variables are the current week's spread and the (current week's spread)^2. Surprisingly, the top ten most important variables exclude power rankings, opponent strength, and all variables from the ESPN game statistics. The other top variables are all from FiveThirtyEight's Elo Ratings. The XGBoost model emphasized the underdog's and favorite's quarterback Elo values, quarterback's Elo rating adjustment, and each team's probability of winning using the quarterback-adjusted Elo rating.

According to histogram below, the residuals on the validation set from this model are normally distributed and centered around 0. As mentioned before, if the predicted point difference is greater than what the spread was set at for a particular game, then the model predicts the favorite would have covered the spread. Otherwise, the model predicts the underdog would have covered the spread. So, for all games in the validation set, when the predicted point difference for a game is within one point of the set spread, the model is 59% accurate. The model is 60.87% accurate when the predicted point difference for a game is within two points of the set spread. According to Table 6.3, as the predicted point difference deviates farther from the set spread, the model consistently stays between 57% and 58% accurate on the validation set.



(a) XGBoost Variable Importances



(b) Distribution of Validation Set Residuals

Figure 6.1: XGBoost Results

Spread - Predicted Point Difference	Correct	Incorrect	Accuracy
[-1, 1]	36	25	0.5902
[-2, 2]	70	45	0.6087
[-3, 3]	81	58	0.5827
[-4, 4]	93	70	0.5706
[-5, 5]	99	73	0.5756
[-6, 6]	101	76	0.5706
[-7, 7]	105	77	0.5769
[-8, 8]	106	77	0.5792
[-9, 9]	107	77	0.5815
[-10, 10]	109	78	0.5829
[-11, 11]	110	78	0.5851

Table 6.3: Spread - Predicted Point Difference, Validation Set

If the spread for a game is set at 3 points and the model predicts the point difference for the game will be 2 points, then the *Spread - Predicted Point Difference* is between [-1, 1].

Because of the consistent performance on the validation set, I proceeded forward with this model on the test set. For the entire 2021 season, I predicted who covered the spread with 53.65% accuracy. This beats the break even percentage of 52.38%, meaning the model would have been profitable for the 2021 season. If we recalculate the Expected Value (EV)

of a bet with $p = 0.5365$, we see that at 10 to 11 odds, each bet is expected to make \$0.2665.

$$EV = p \times Payout - (1-p) \times Amount\ Gambled$$

$$EV = 0.5365 \times \$10 - (1 - 0.5365) \times \$11$$

$$EV = 0.2665$$

According to the confusion matrix below, we see that the model correctly predicted who covered the spread 110 times compared to the 95 incorrect predictions on the test set.

		Prediction	
		Favorite	Underdog
Truth	Favorite	43	60
	Underdog	35	67

Table 6.4: Confusion Matrix, Test Set

When betting on the spread, sportsbooks always offer to 10 to 11 odds. If a bettor gambled \$11 on every game, the bettor would have profited \$55 for the entire season with the following formula:

$$Profit = (AG + P) \times N_{CP} - AG \times N$$

where AG is the amount gambled, P is the payout, N_{CP} is the number of times a bettor correctly predicted who would cover the spread, and N is the total number of games bet on.

$$Profit = (\$11 + \$10) \times 110 - \$11 \times 205$$

$$Profit = \$2310 - \$2255$$

$$Profit = \$55$$

In the above example, a bettor would have gambled \$2255 for the entire season. Rather than betting on every game, a bettor could have taken a more conservative approach and gambled less money. On the validation set, the model performed well when the predicted point difference is within one point of the set spread. In 2021, if a bettor only gambled on games where the predicted point difference is within one point of the spread, the bettor would have correctly picked who covered the spread 61.54% of the time. With 40 correct predictions compared to 25 incorrect predictions, the bettor would have profited \$125.

$$Profit = (\$11 + \$10) \times 40 - \$11 \times 65$$

$$Profit = \$840 - \$715$$

$$Profit = \$125$$

Spread - Predicted Point Difference	Correct	Incorrect	Accuracy
[-1, 1]	40	25	0.6154
[-2, 2]	59	52	0.5315
[-3, 3]	84	75	0.5283
[-4, 4]	95	82	0.5367
[-5, 5]	101	92	0.5233
[-6, 6]	105	92	0.533
[-7, 7]	108	94	0.5347
[-8, 8]	110	94	0.5392
[-9, 9]	110	94	0.5392
[-10, 10]	110	95	0.5366

Table 6.5: Spread - Predicted Point Difference, Test Set

If the spread for a game is set at 3 points and the model predicts the point difference for the game will be 2 points, then the *Spread - Predicted Point Difference* is between [-1, 1].

Now, let's say we have a person who is new to sports betting and has \$100 to gamble. It would be unwise to bet all \$100 on one game. Instead, the bettor could follow the

Kelly Criterion for betting[21]. The Kelly Criterion determines the optimal bet size using the following formula:

$$prop = p - odds \times (1 - p)$$

where *prop* is the proportion of the amount of money available to gamble, p is the probability of the gambler correctly picking who will cover the spread, and *odds* is defined as $\frac{AmountGambled}{Profit}$. Let's say the bettor only wants to bet on games where the predicted point difference is within one point of the spread because the model is 59.02% accurate for these games on the validation set. Hence, $p = 0.5902$. At 10 to 11 odds, the formula above would become:

$$prop = 0.5902 - \frac{11}{10} \times (1 - 0.5902)$$

$$prop = 0.5902 - 1.1 \times 0.4098$$

$$prop = 0.13942$$

This means on a given bet, the bettor should only place 13.942% of the amount available to gamble. Since the bettor has \$100 available and each bet should only be \$13.94, the bettor can place bets on 7 different games where the model predicts the point difference will be within one point of the spread. According to Table 6.3, we would expect the model to correctly predict who covers the spread in 59.02% of these games. At 10 to 11 odds, we would bet \$13.94 for the chance to win \$12.67. Theoretically, if we correctly predicted who covered the spread in the 4 of the 7 games, the bettor would have profited \$8.86.

$$Profit = (\$13.94 + \$12.67) \times 4 - \$13.94 \times 7$$

$$Profit = \$106.44 - \$97.58$$

$$Profit = \$8.86$$

[21] `https://en.wikipedia.org/wiki/Kelly_criterion`

Table 6.6 shows how well the model performed for each team on the test set. When the Tampa Bay Buccaneers were the favorite, the model correctly predicted they would cover the spread 5 times and made 0 incorrect predictions. In 4 of these games, the model's predicted point difference was within one point of the spread. Again, if the bettor had $100 to gamble, the Kelly Criterion would have suggested to bet $13.94 on each game. From these 4 games, the bettor would have profited $50.68 in total. On the test set, we see that the model predicted well for the Dallas Cowboys when they were the favorite and for the Las Vegas Raiders when they were the underdog.

	Model's Record When			
	Team Was Favorite And Model Predicted		Team Was Underdog And Model Predicted	
Team	Team Would Cover Spread	Team Would Not Cover Spread	Team Would Cover Spread	Team Would Not Cover Spread
Arizona Cardinals	2-2	4-1	2-0	0-2
Atlanta Falcons	0-1	1-2	2-2	3-1
Baltimore Ravens	1-3	4-1	2-1	0-1
Buffalo Bills	3-2	3-2	2-1	0-0
Carolina Panthers	0-2	3-0	0-3	3-2
Chicago Bears	0-1	0-1	3-5	2-1
Cincinnati Bengals	1-1	2-3	1-1	0-3
Cleveland Browns	0-2	4-1	2-4	0-0
Dallas Cowboys	5-0	3-3	0-0	1-1
Denver Broncos	2-1	3-0	2-4	0-1
Detroit Lions	0-0	0-0	4-3	1-5
Green Bay Packers	0-0	4-5	1-0	0-2
Houston Texans	0-1	0-0	3-1	5-2
Indianapolis Colts	2-1	3-2	4-0	1-0
Jacksonville Jaguars	0-1	0-0	2-5	3-2
Kansas City Chiefs	4-4	2-3	0-0	0-0
Las Vegas Raiders	0-1	2-2	5-1	2-0
Los Angeles Chargers	0-2	4-4	0-0	2-1
Los Angeles Rams	0-2	4-5	1-0	0-0
Miami Dolphins	3-0	2-1	2-2	1-1
Minnesota Vikings	0-2	1-4	2-2	1-1
New England Patriots	3-0	3-3	2-2	0-0
New Orleans Saints	2-0	2-2	3-3	1-0
New York Giants	0-0	0-0	3-4	5-1
New York Jets	1-0	0-0	3-3	5-1
Philadelphia Eagles	3-0	2-2	2-2	1-1
Pittsburgh Steelers	0-2	1-2	3-2	1-2
San Francisco 49ers	4-2	2-2	2-0	1-0
Seattle Seahawks	1-2	0-2	3-4	0-1
Tampa Bay Buccaneers	5-0	5-3	0-0	0-0
Tennessee Titans	1-0	3-2	4-1	1-1
Washington Commanders	0-0	0-2	2-4	3-2

Table 6.6: 2021 Betting Results, Test Set

When the Arizona Cardinals were the favorite, the model predicted the Cardinals would cover the spread 4 times. The model was correct 2 times and incorrect 2 times. Again, when the Arizona Cardinals were the favorite, the model predicted they would not cover the spread 5 times. The model was correct 4 times and incorrect 1 time.

CHAPTER 7

Conclusion and Limitations

In conclusion, it is possible to accurately predict who will cover the spread with machine learning. Generally, the more predictive models outperformed the more interpretable models. I found that using the top correlated variables, each variables' quadratic term, and the four week rolling average of the variables offered the best predictability. In fact, the XGBoost model with this feature set correctly predicted who would cover the spread 53.65% of the time on the test set. In addition, when the predicted point difference is within one point of the set spread, the model determined who would cover the spread with an accuracy of 61.54%. Despite the promising results with the 2021 season, this does not guarantee the model would have been profitable this past season nor is it a guarantee it will be profitable for the 2023 season without retraining, revalidating, and retesting. Advanced analytical metrics and data at the player level could help improve accuracy. Also, accuracy might improve if I compared my predictions to one sportsbook rather than the average of six sportsbooks. Overall, there is reason to believe a bettor could beat Las Vegas in the long run with machine learning.

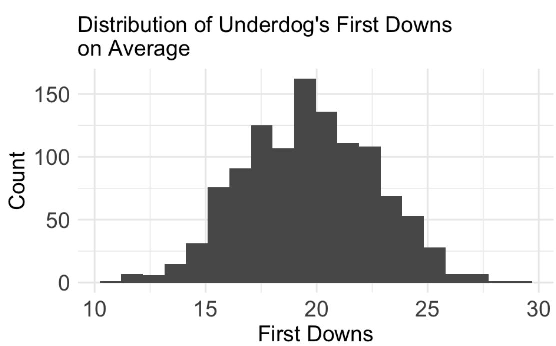
CHAPTER 8

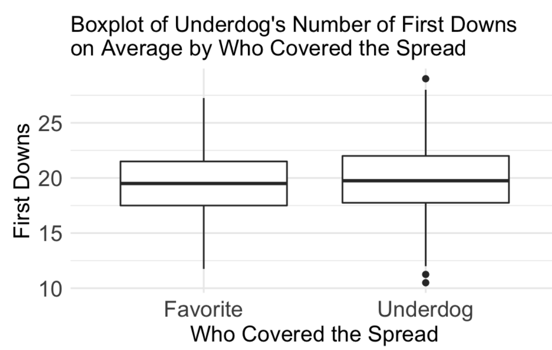
Appendix

8.1 Underdog Numeric Variables

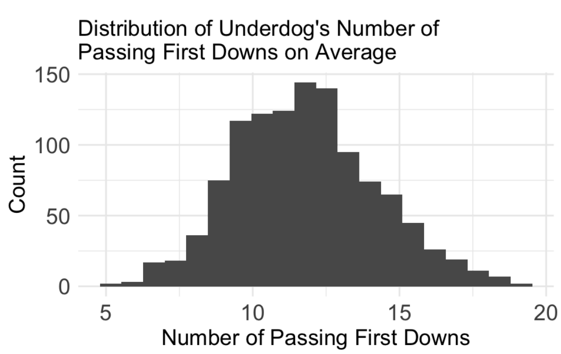
The following section contains histograms and boxplots of the remaining numeric variables for the underdog. In all cases, the boxplots show little to no difference in who covered the spread.

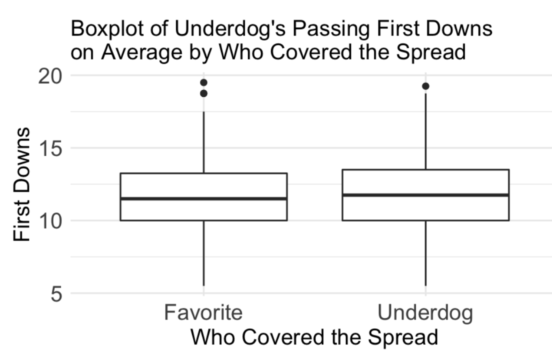
1. Underdog's Number of First Downs, On Average, From Last 4 Games



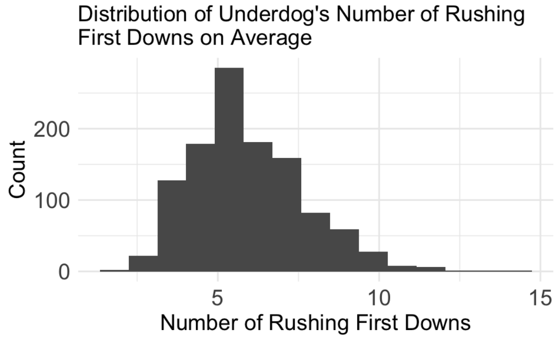


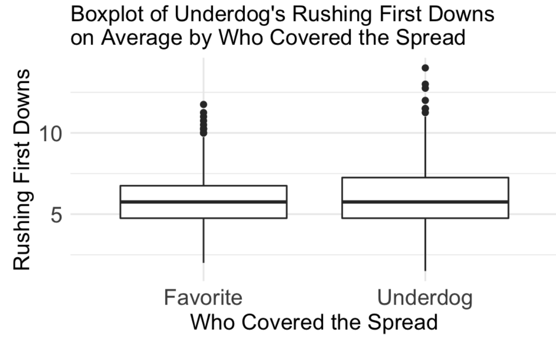
2. Underdog's Number of Passing First Downs, On Average, From Last 4 Games



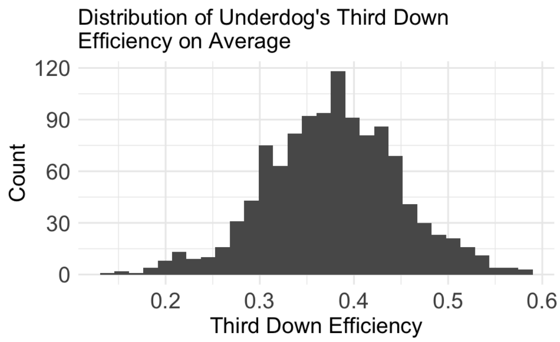


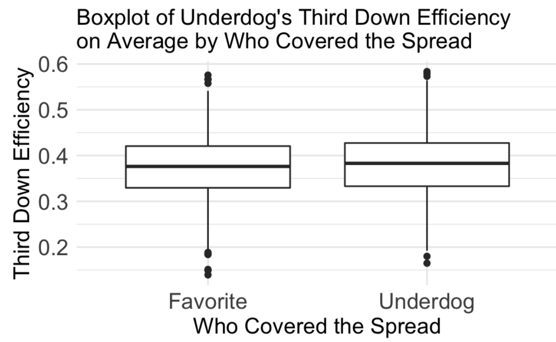
3. Underdog's Number of Rushing First Downs, On Average, From Last 4 Games



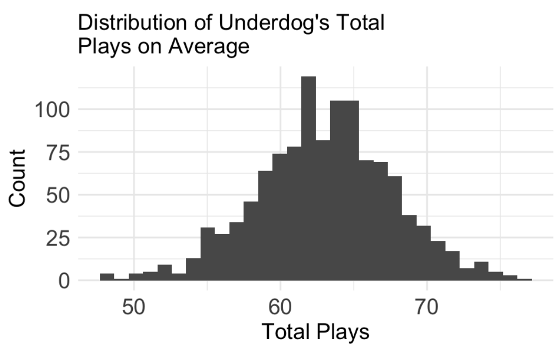


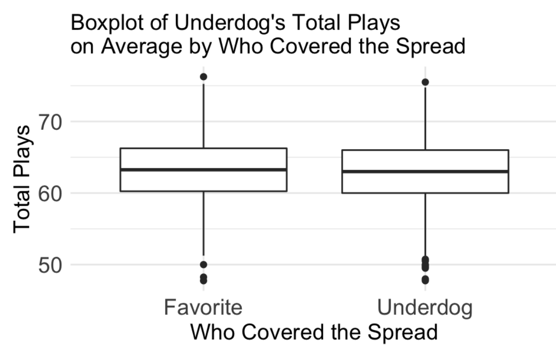
4. Underdog's Third Down Efficiency, On Average, From Last 4 Games



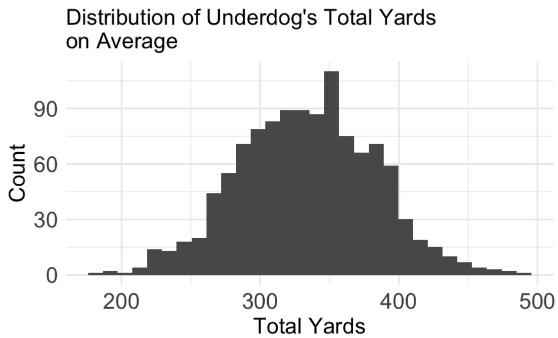


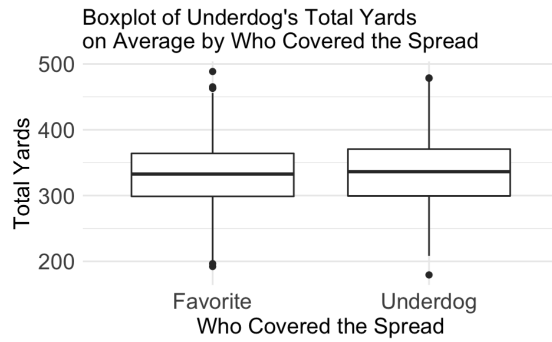
5. Underdog's Total Plays, On Average, From Last 4 Games



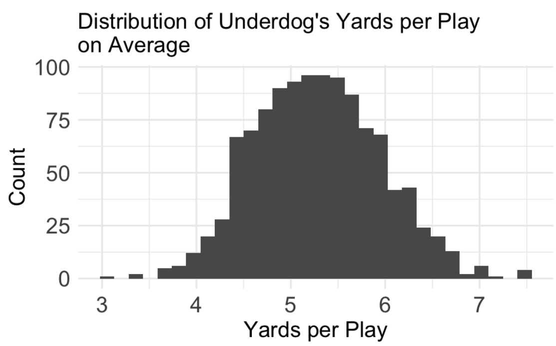


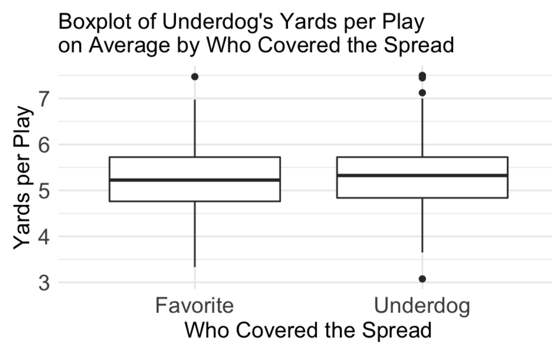
6. Underdog's Total Yards, On Average, From Last 4 Games



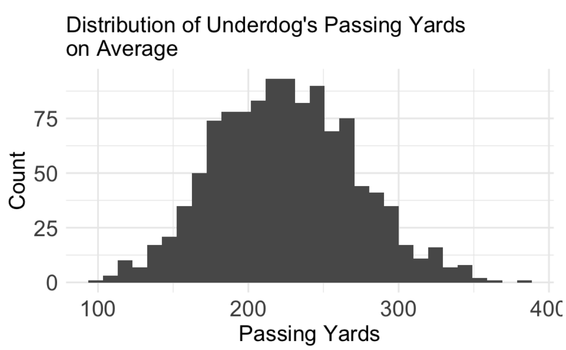


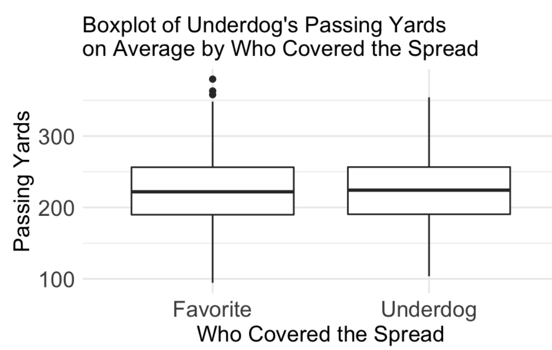
7. Underdog's Yards Per Play, On Average, From Last 4 Games



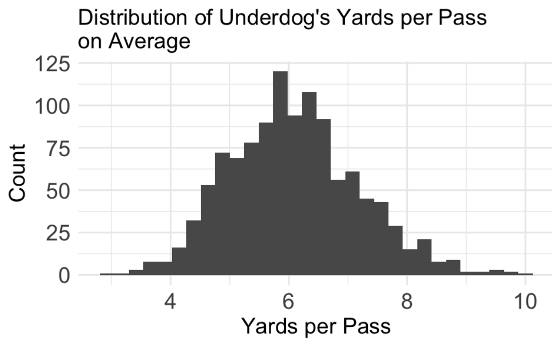


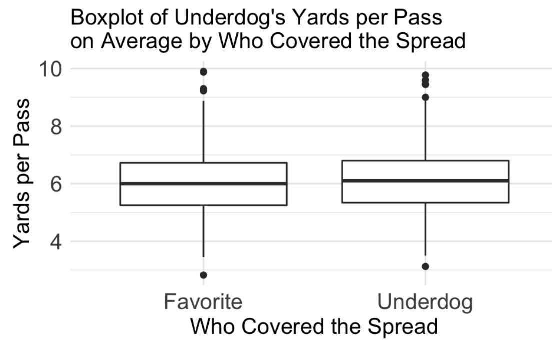
8. Underdog's Passing Yards, On Average, From Last 4 Games



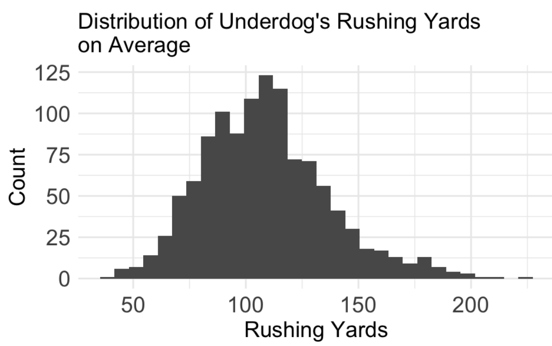


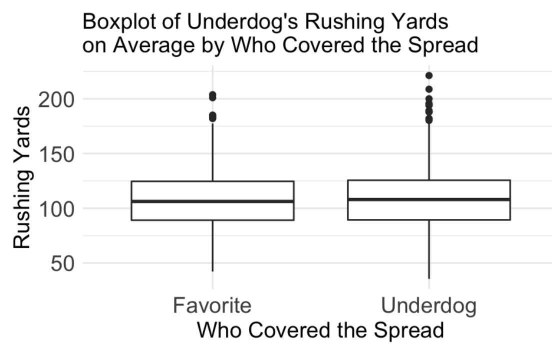
9. Underdog's Yards Per Pass, On Average, From Last 4 Games



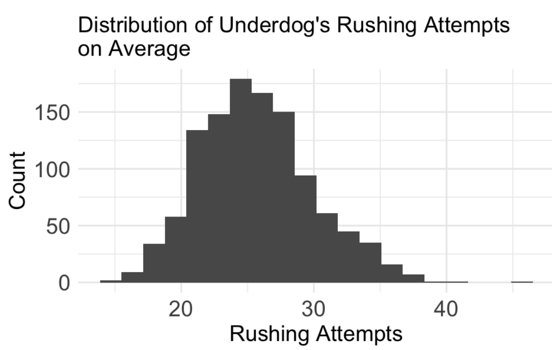


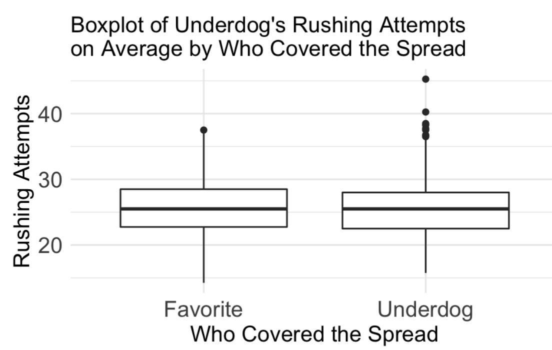
10. Underdog’s Rushing Yards, On Average, From Last 4 Games



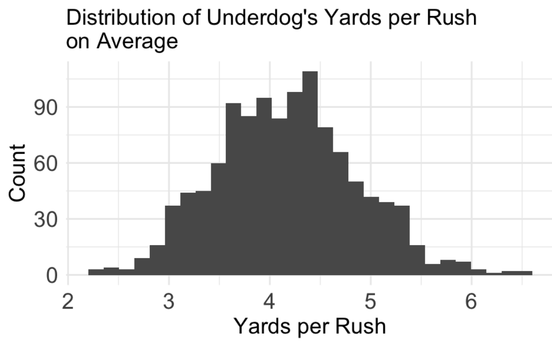


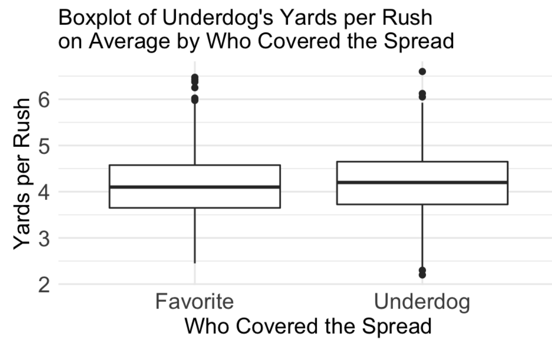
11. Underdog’s Rushing Attempts, On Average, From Last 4 Games



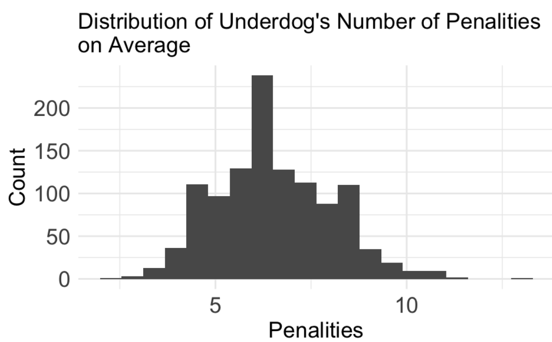


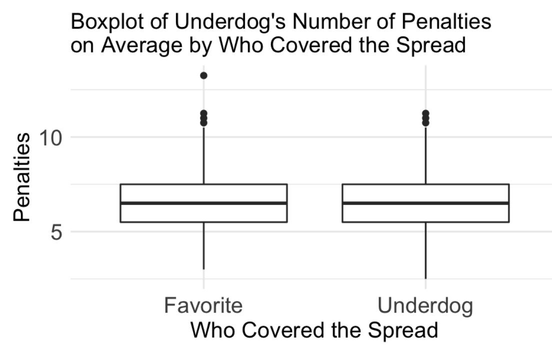
12. Underdog’s Yards Per Rush, On Average, From Last 4 Games



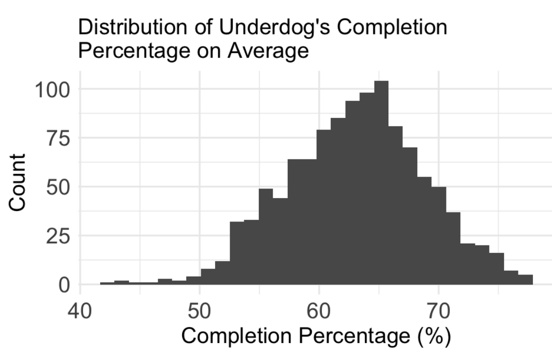


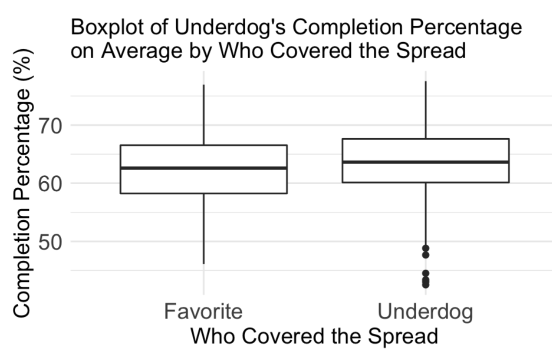
13. Underdog's Number of Penalties, On Average, From Last 4 Games



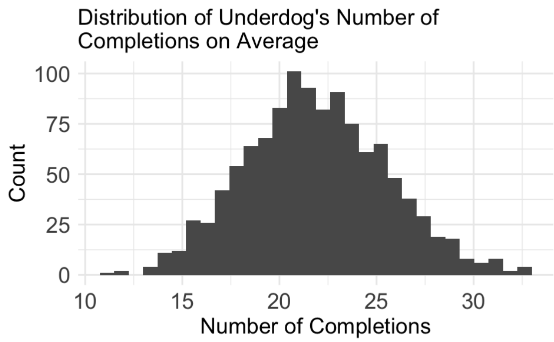


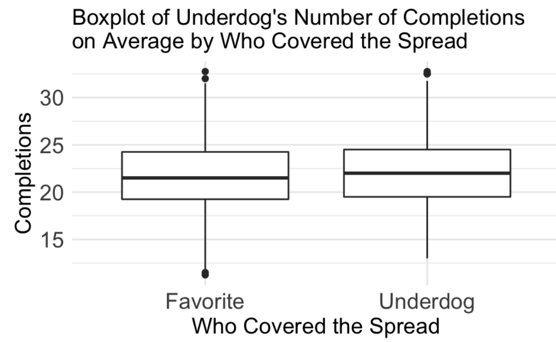
14. Underdog's Completion Percentage, On Average, From Last 4 Games



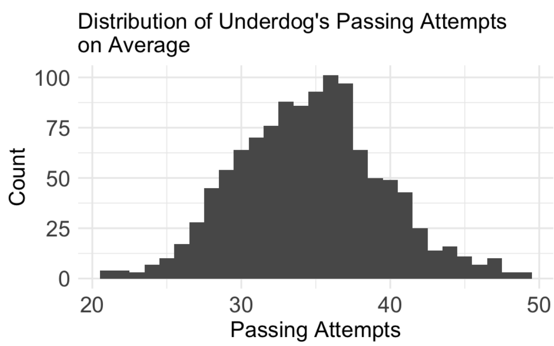


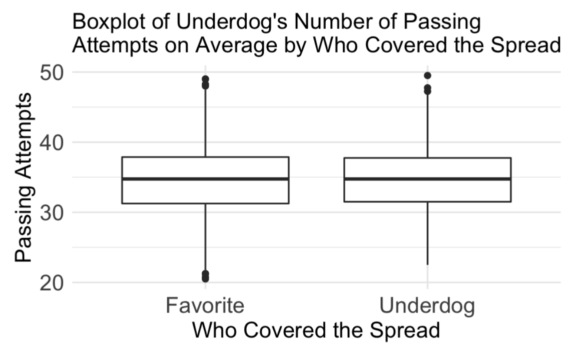
15. Underdog's Number of Pass Completions, On Average, From Last 4 Games



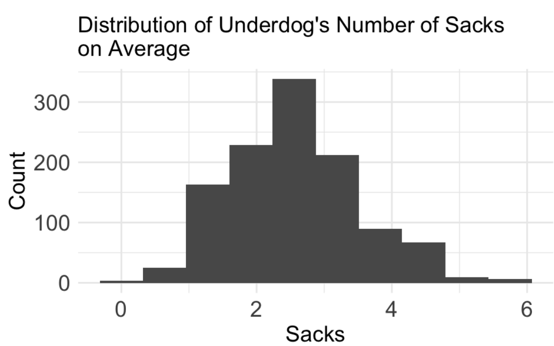


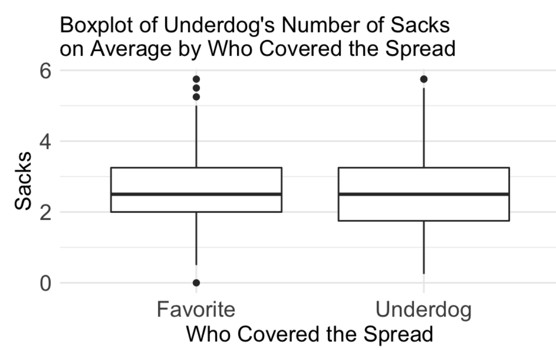
16. Underdog's Number of Pass Attempts, On Average, From Last 4 Games



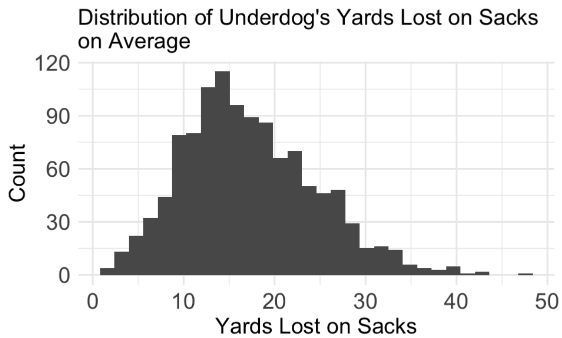


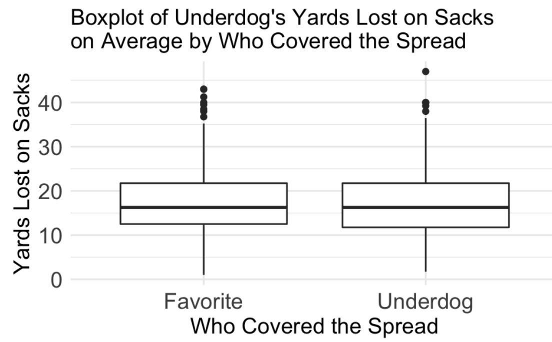
17. Underdog's Number of Sacks, On Average, From Last 4 Games



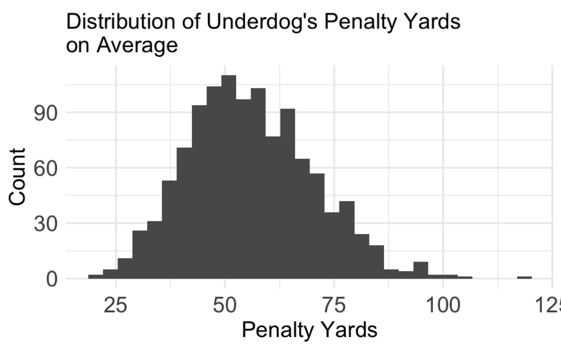


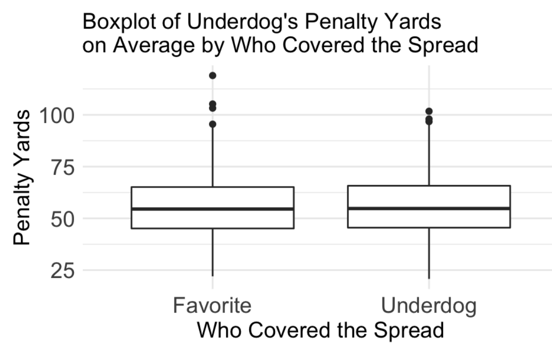
18. Underdog's Yards Lost on Sacks, On Average, From Last 4 Games



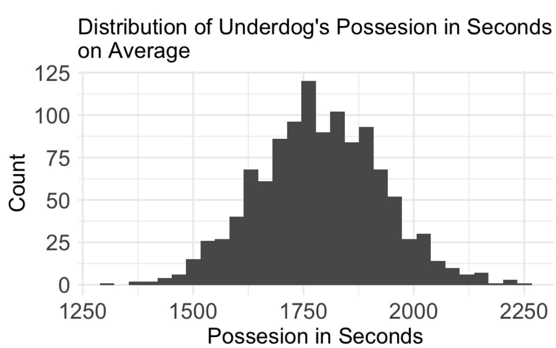


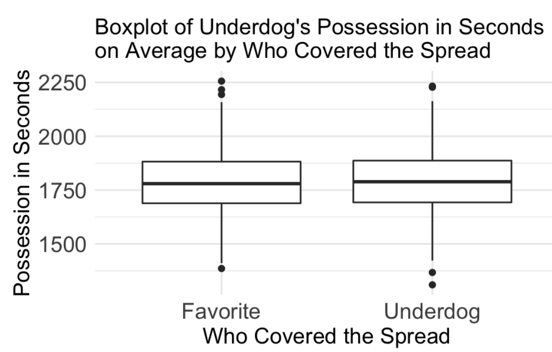
19. Underdog’s Penalty Yards, On Average, From Last 4 Games



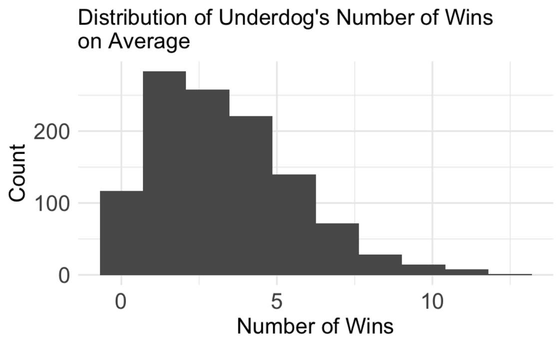


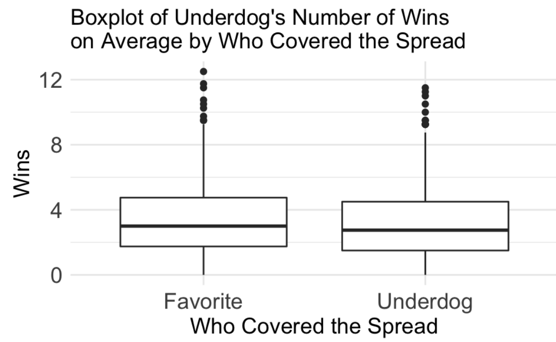
20. Underdog’s Time of Possession (Seconds), On Average, From Last 4 Games



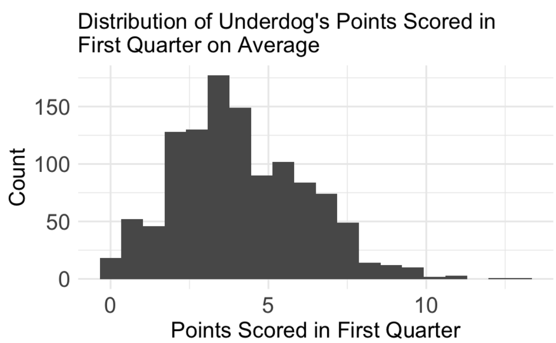


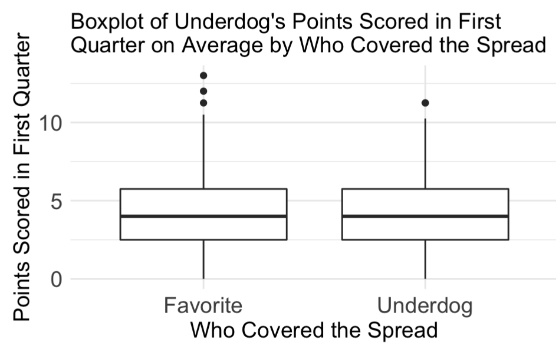
21. Underdog’s Number of Wins, On Average, From Last 4 Games



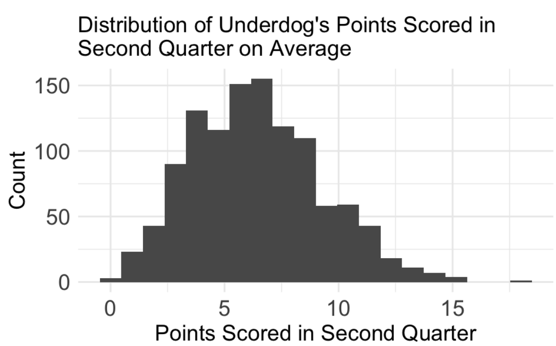


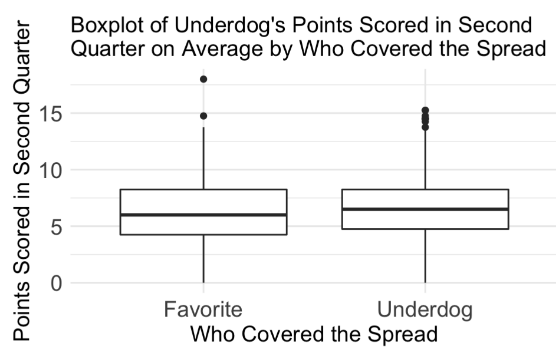
22. Underdog's Points Scored in First Quarter, On Average, From Last 4 Games



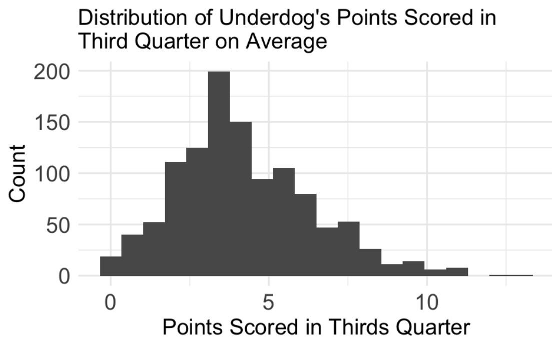


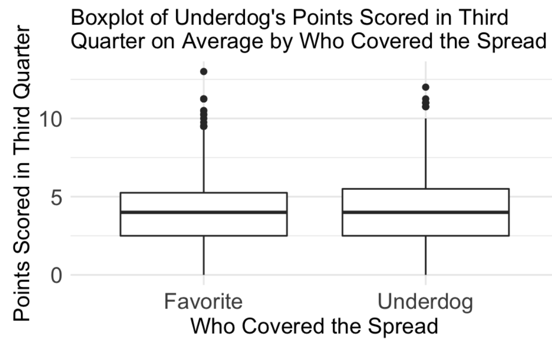
23. Underdog's Points Scored in Second Quarter, On Average, From Last 4 Games



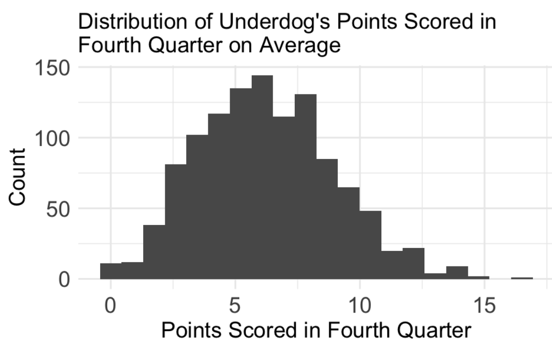


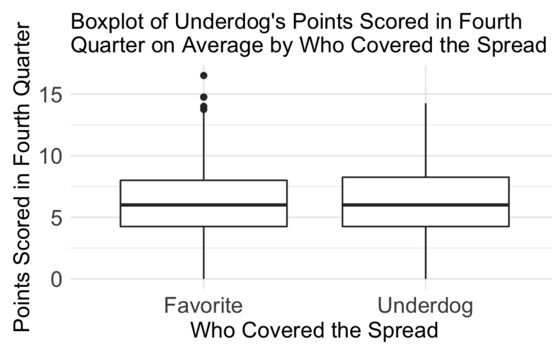
24. Underdog's Points Scored in Third Quarter, On Average, From Last 4 Games



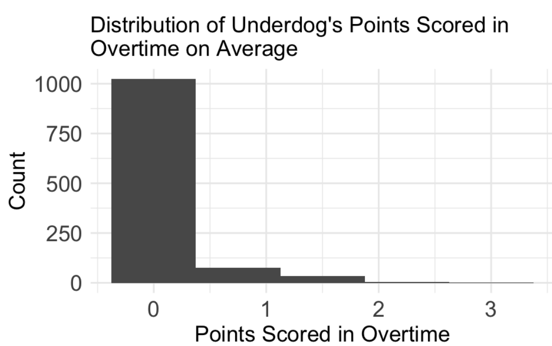


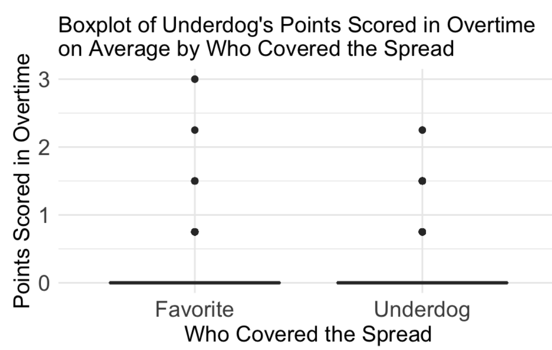
25. Underdog's Points Scored in Fourth Quarter, On Average, From Last 4 Games



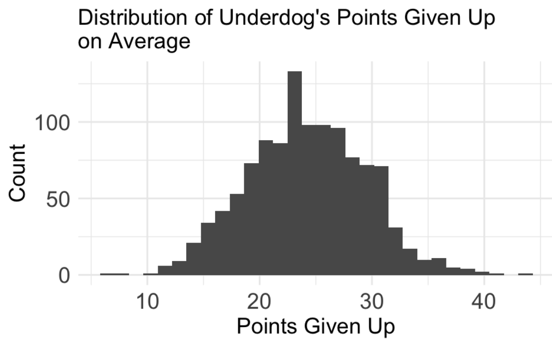


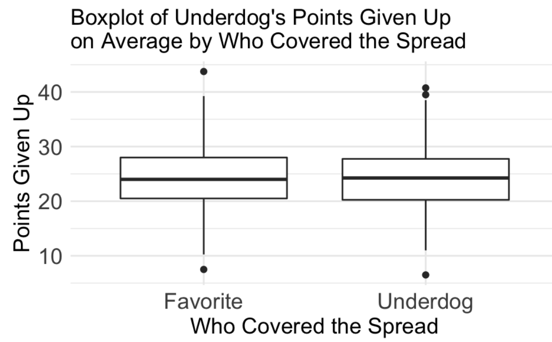
26. Underdog's Points Scored in Overtime, On Average, From Last 4 Games



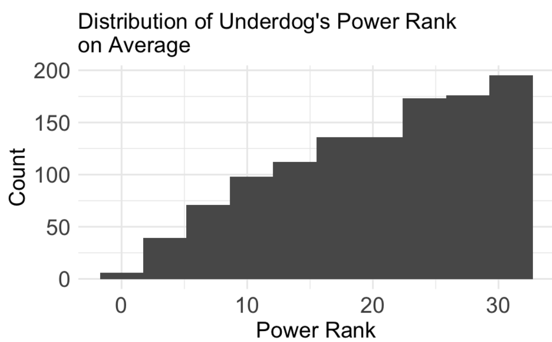


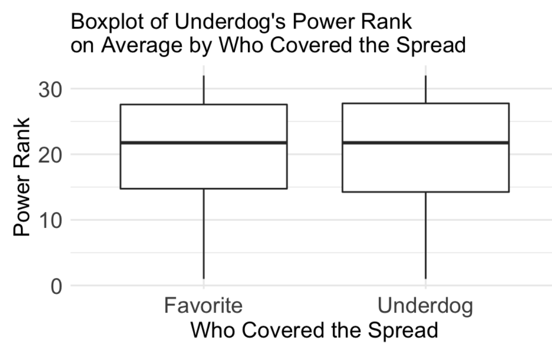
27. Underdog's Points Allowed, On Average, From Last 4 Games



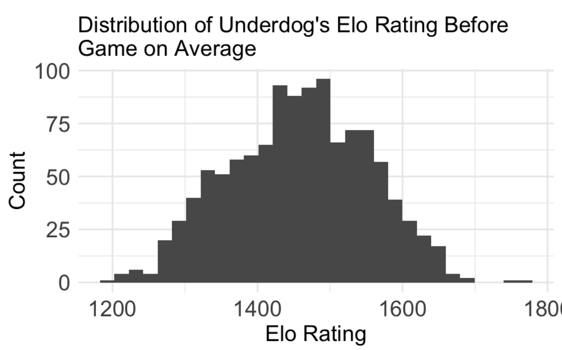


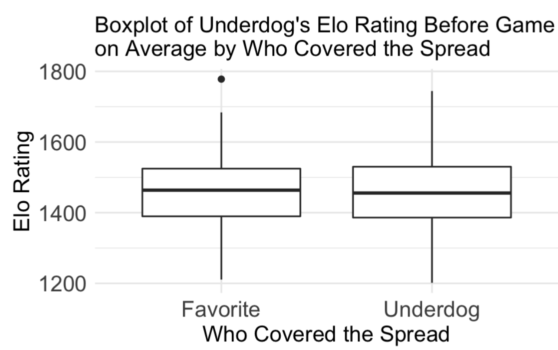
28. Underdog’s Power Rank, On Average, From Last 4 Games



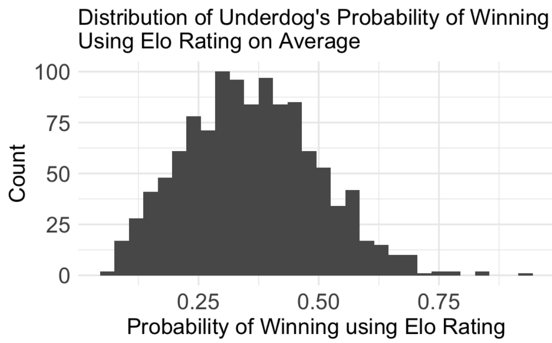


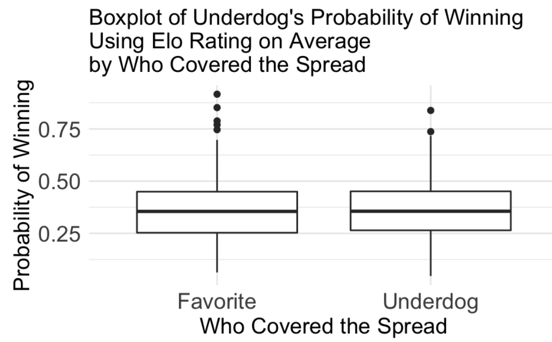
29. Underdog’s Elo Rating Before Game, On Average, From Last 4 Games



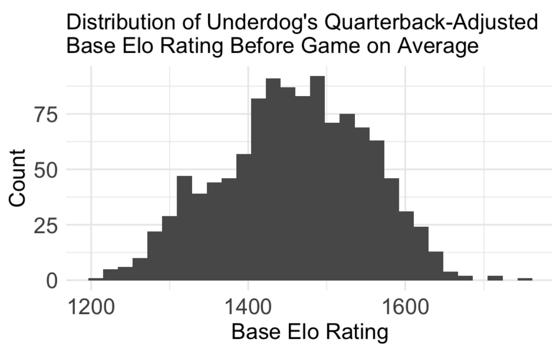


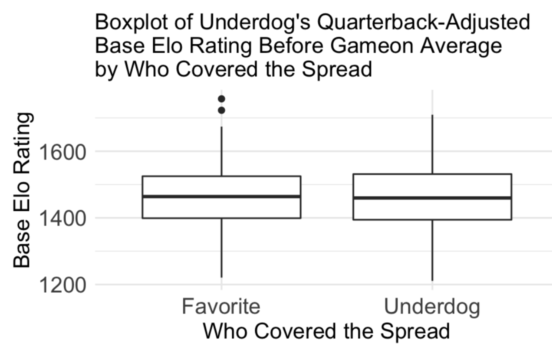
30. Underdog’s Probability of Winning Using Elo Rating, On Average, From Last 4 Games



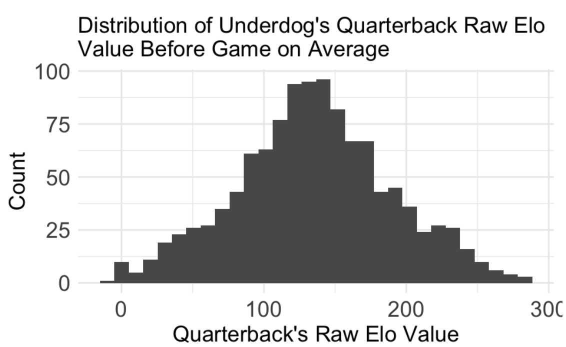


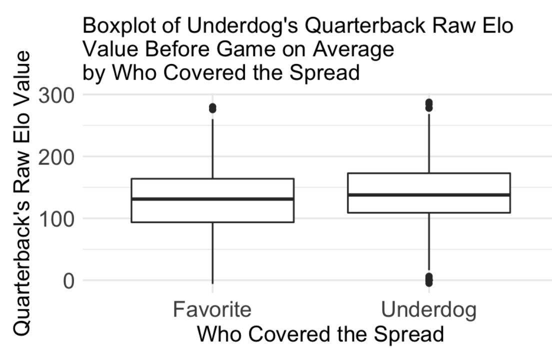
31. Underdog's Quarterback-Adjusted Elo Rating Before Game, On Average, From Last 4 Games



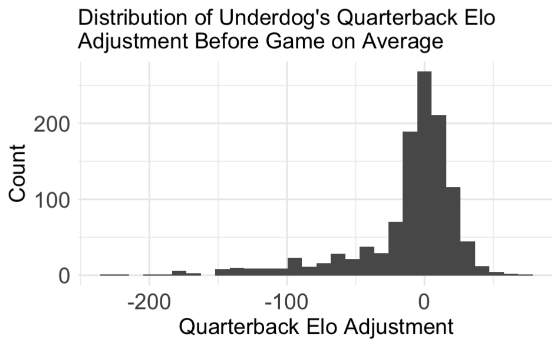


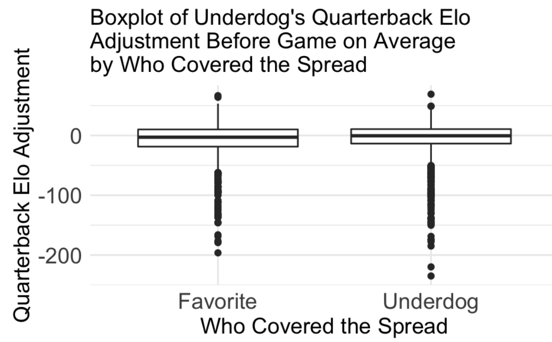
32. Underdog's Quarterback Raw Elo Value Before Game, On Average, From Last 4 Games



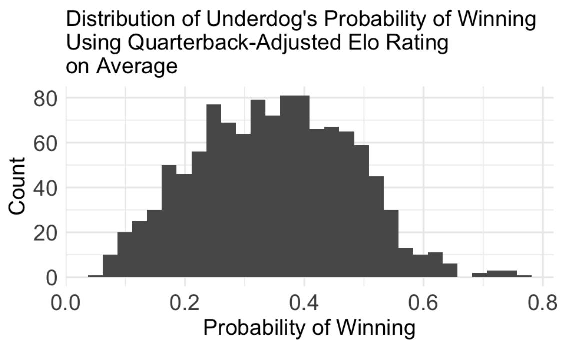


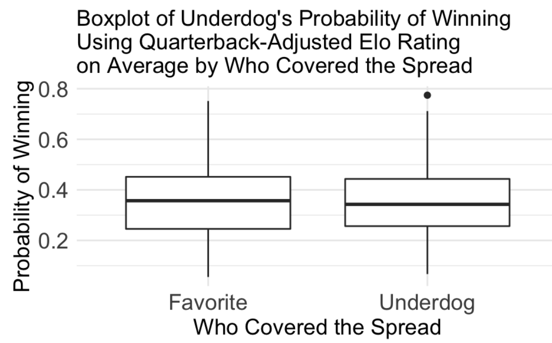
33. Underdog's Quarterback Elo Adjustment Before Game, On Average, From Last 4 Games



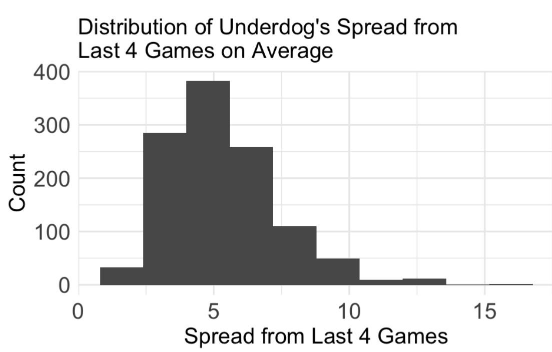


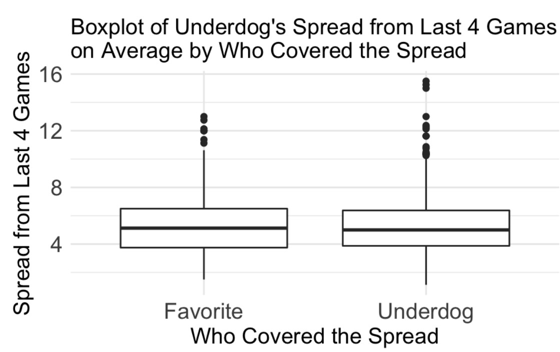
34. Underdog's Probability of Winning Using Quarterback-Adjusted Elo Rating, On Average, From Last 4 Games



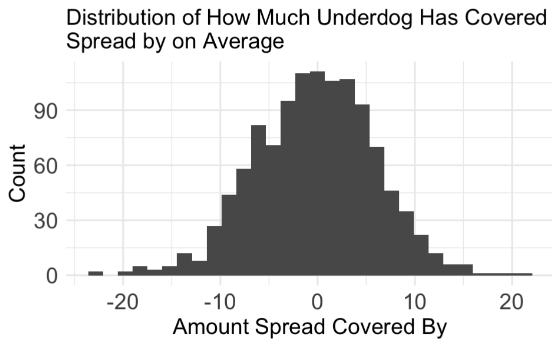


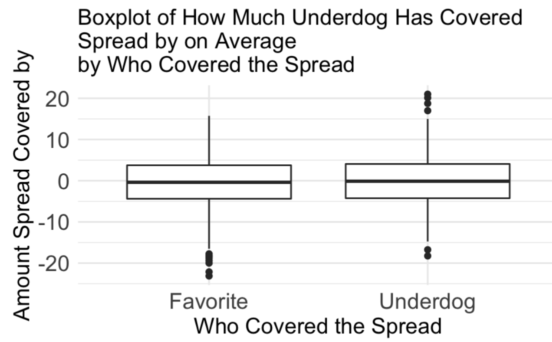
35. Underdog's Spread, On Average, From Last 4 Games



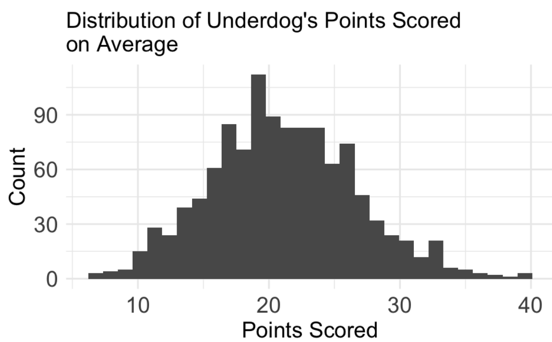


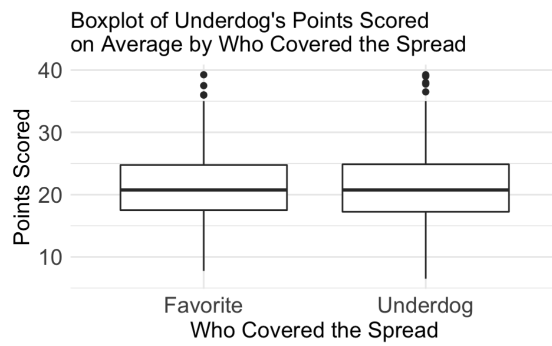
36. How Much Underdog Has Covered Spread by, On Average, From Last 4 Games





37. Underdog's Points Scored, On Average, From Last 4 Games





8.2 Underdog Categorical Variables

The following section contains tables of the remaining categorical variables for the underdog. Despite some categories having discernible counts, these variables provided little to no predictability in determining who covered the spread.

1. Underdog's Number of Wins From Last 4 Games

Number of Wins	Count
0	189
1	359
2	351
3	193
4	50

	Who Covered Spread	
Number of Wins	Favorite	Underdog
0	96	93
1	171	188
2	188	163
3	89	104
4	23	27

2. Number of Times Underdog Covered Spread From Last 4 Games

Number of Times Underdog Covered Spread	Count
0	88
1	348
2	449
3	225
4	32

	Who Covered Spread Current Week	
Number of Times Underdog Covered Spread	Favorite	Underdog
0	43	45
1	162	186
2	236	213
3	111	114
4	15	17

3. Number of Times Underdog Was Favorite in Last 4 Games

Number of Times Underdog Was Favorite	Count
0	269
1	332
2	272
3	183
4	86

	Who Covered Spread	
Number of Times Underdog Was Favorite	Favorite	Underdog
0	141	128
1	173	159
2	119	153
3	86	97
4	48	38

4. Underdog's Number of First Downs by Penalty in Last 4 Games

Number of First Downs	Count
0 - 7	561
8+	581

	Who Covered Spread	
Number of First Downs	Favorite	Underdog
0 - 7	281	280
8+	286	295

5. Did Underdog Attempt to Convert on Fourth Down in Last 4 Games

Attempted 4th Down	Count
Yes	1005
No	137

	Who Covered Spread	
Attempted 4th Down	Favorite	Underdog
Yes	491	514
No	76	61

6. Underdog's Number of Offensive Drives in Last 4 Games

Total Drives	Count
0 - 44	535
45+	607

	Who Covered Spread	
Total Drives	Favorite	Underdog
0 - 44	245	290
45+	322	285

7. Underdog's Number of Interceptions Thrown in Last 4 Games

Interceptions	Count
0 - 3	620
4+	522

	Who Covered Spread	
Interceptions	Favorite	Underdog
0 - 3	300	320
4+	267	255

8. Underdog's Red Zone Efficiency in Last 4 Games

Efficiency	Count
0 - 54%	585
54%+	557

	Who Covered Spread	
Efficiency	Favorite	Underdog
0 - 54%	301	284
54%+	266	291

9. Underdog's Number of Turnovers in Last 4 Games

Turnovers	Count
0 - 5	540
6+	602

	Who Covered Spread	
Turnovers	Favorite	Underdog
0 - 5	253	287
6+	314	288

10. Underdog's Number of Fumbles Lost in Last 4 Games

Fumbles Lost	Count
0 - 2	684
3+	458

	Who Covered Spread	
Fumbles Lost	Favorite	Underdog
0 - 2	321	363
3+	246	212

11. Underdog's Number of Defensive or Special Teams Touchdowns in Last 4 Games

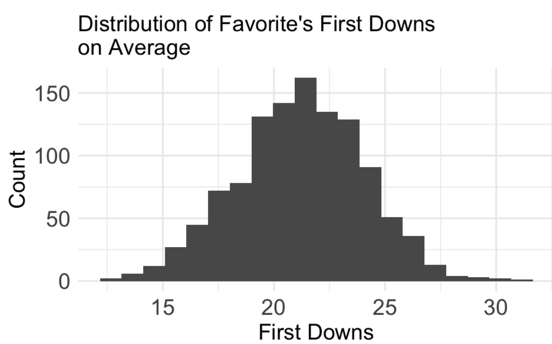
Touchdowns	Count
0	660
1+	482

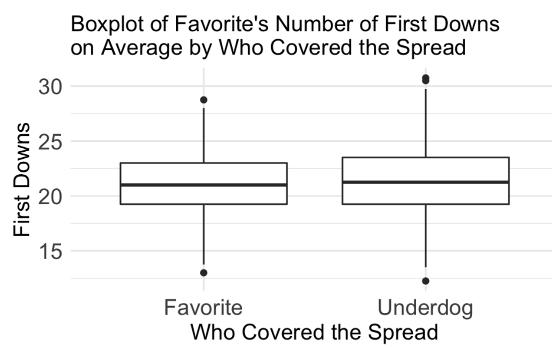
	Who Covered Spread	
Touchdowns	Favorite	Underdog
0	310	350
1+	257	225

8.3 Favorite Numeric Variables

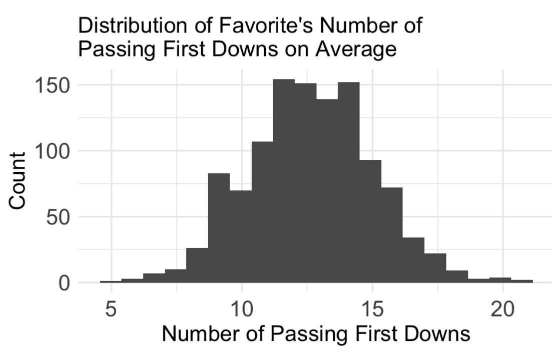
The following section contains histograms and boxplots of the remaining numeric variables for the favorite. In all cases, the boxplots show little to no difference in who covered the spread.

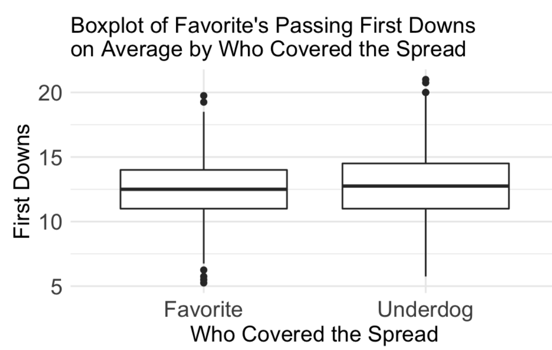
1. Favorite's Number of First Downs, On Average, From Last 4 Games



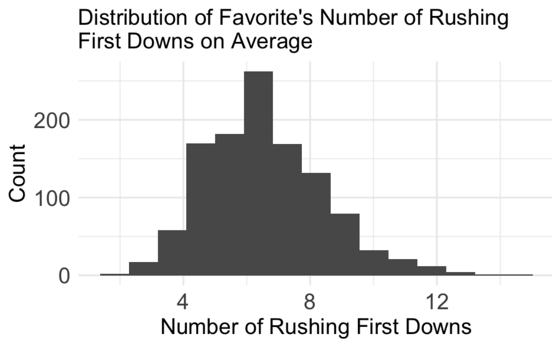


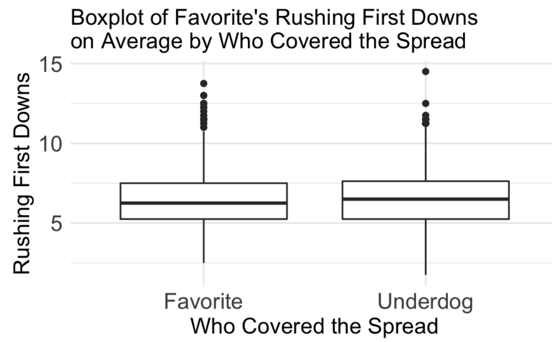
2. Favorite's Number of Passing First Downs, On Average, From Last 4 Games



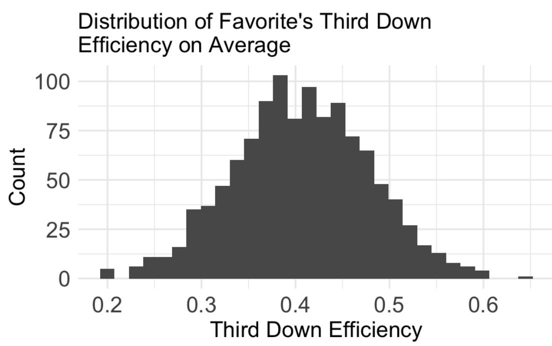


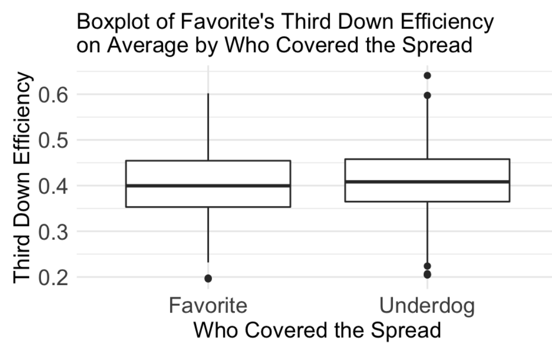
3. Favorite's Number of Rushing First Downs, On Average, From Last 4 Games



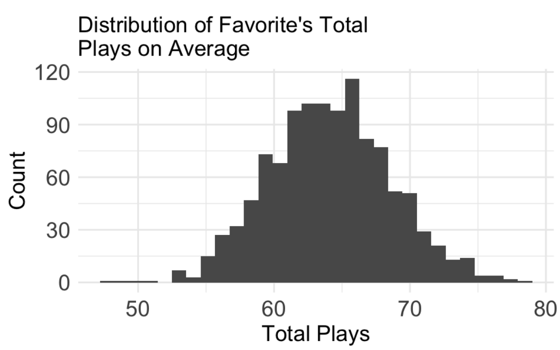


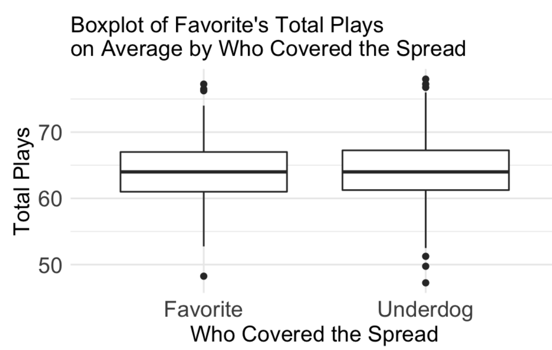
4. Favorite's Third Down Efficiency, On Average, From Last 4 Games



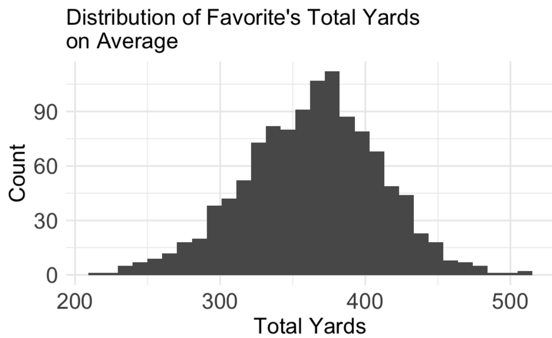


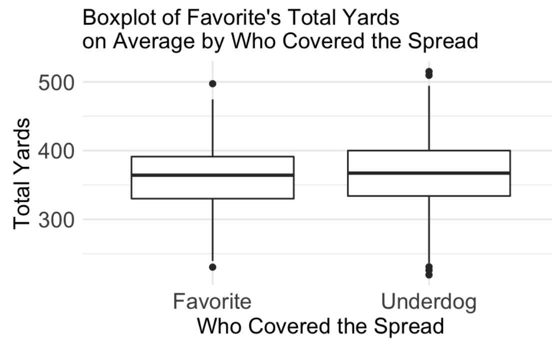
5. Favorite's Total Plays, On Average, From Last 4 Games



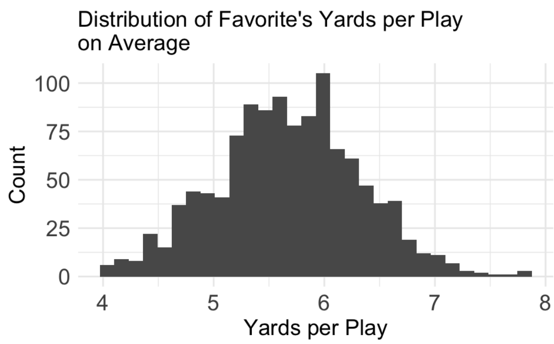


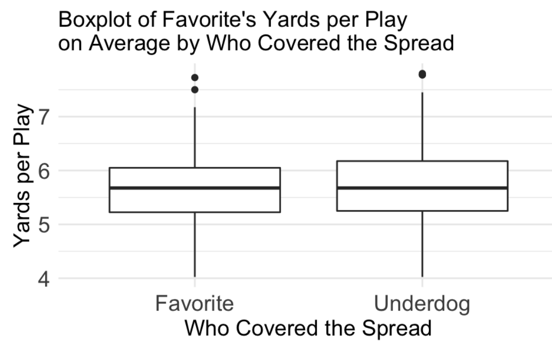
6. Favorite's Total Yards, On Average, From Last 4 Games



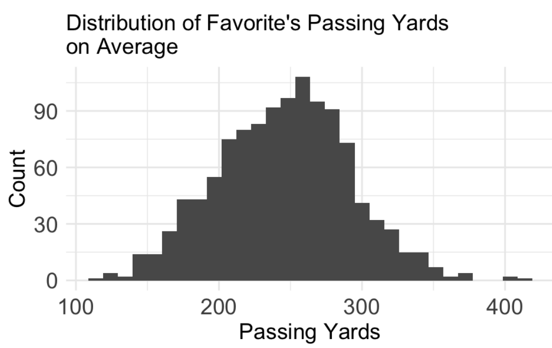


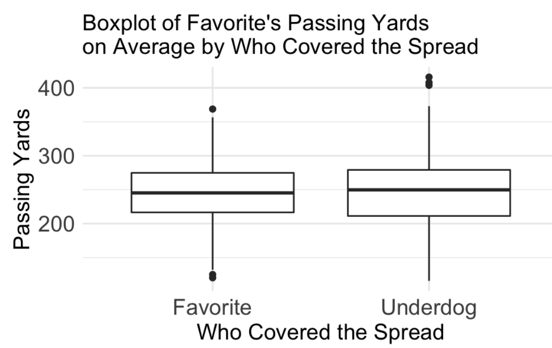
7. Favorite's Yards Per Play, On Average, From Last 4 Games



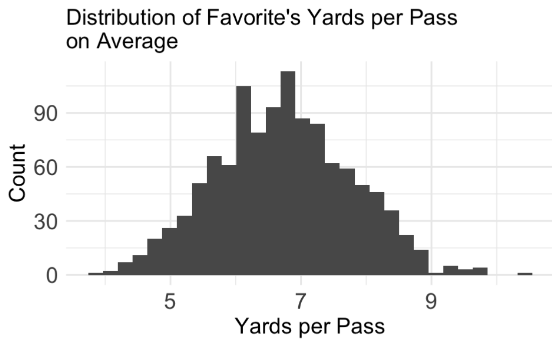


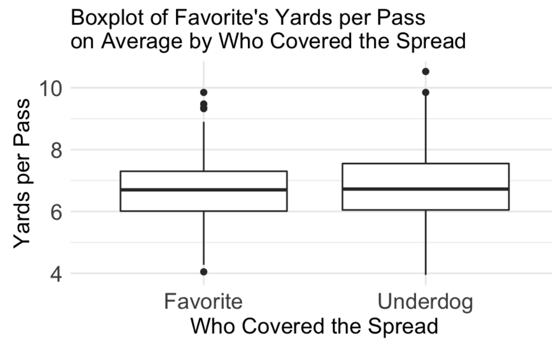
8. Favorite's Passing Yards, On Average, From Last 4 Games



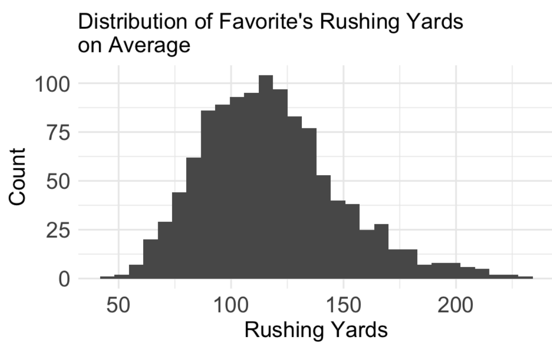


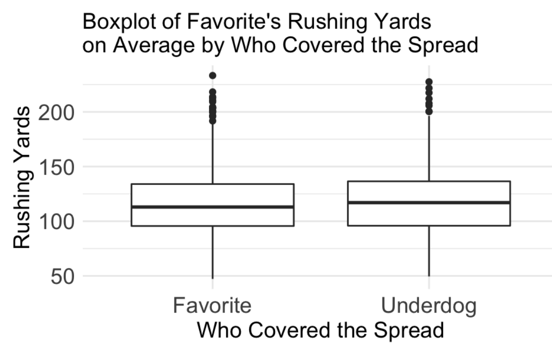
9. Favorite's Yards Per Pass, On Average, From Last 4 Games



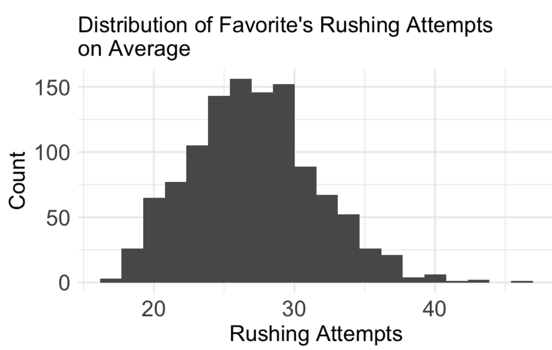


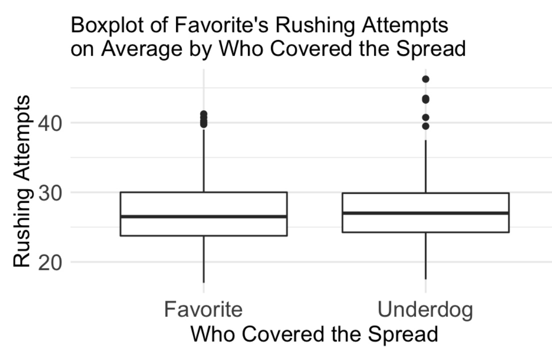
10. Favorite's Rushing Yards, On Average, From Last 4 Games



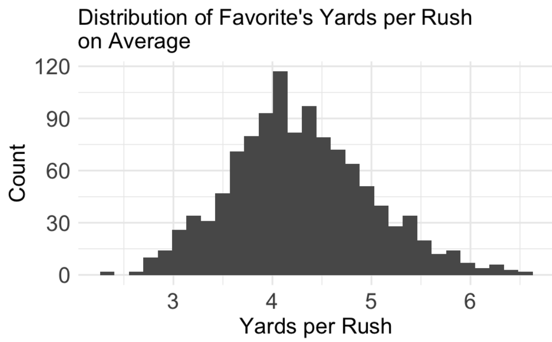


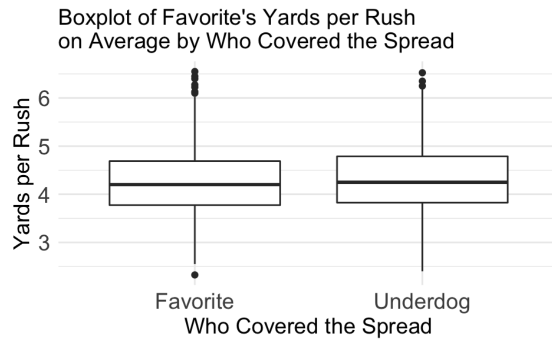
11. Favorite's Rushing Attempts, On Average, From Last 4 Games



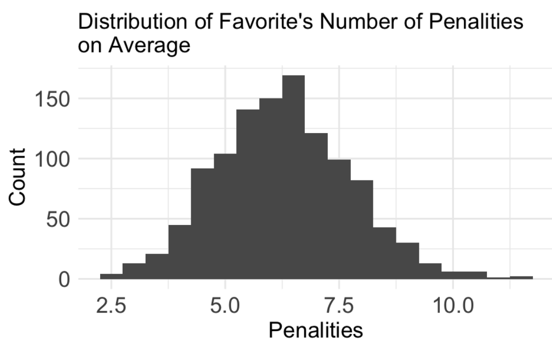


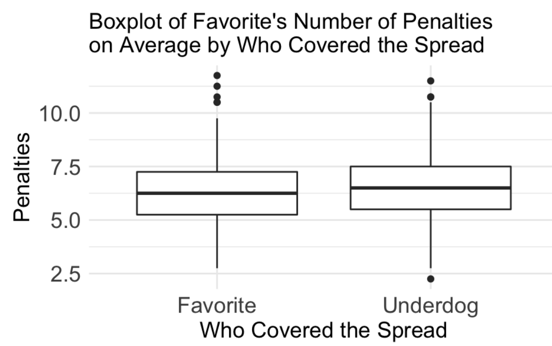
12. Favorite's Yards Per Rush, On Average, From Last 4 Games



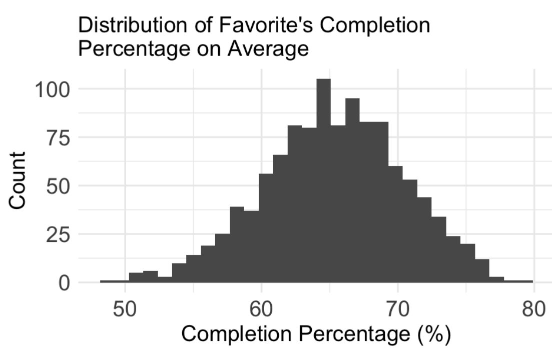


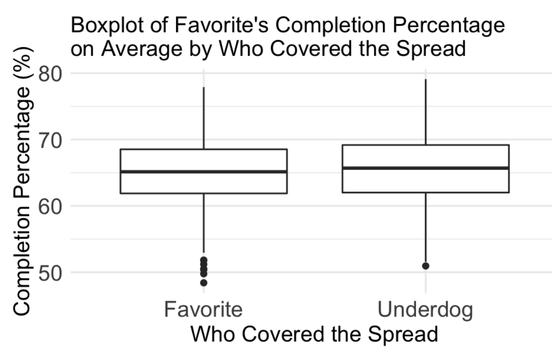
13. Favorite’s Number of Penalties, On Average, From Last 4 Games



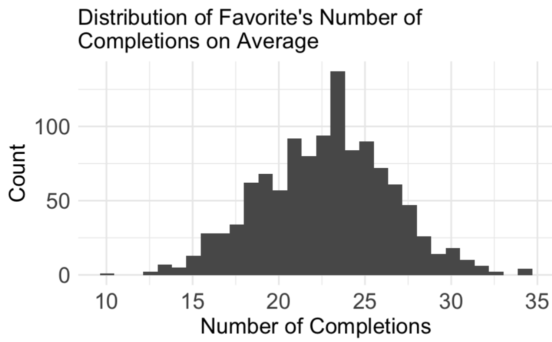


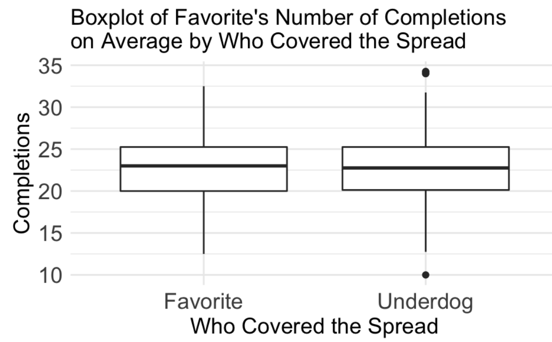
14. Favorite’s Completion Percentage, On Average, From Last 4 Games



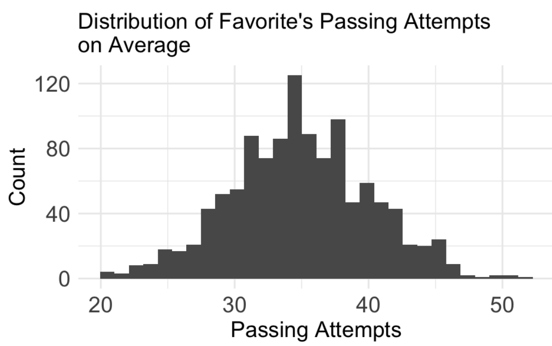


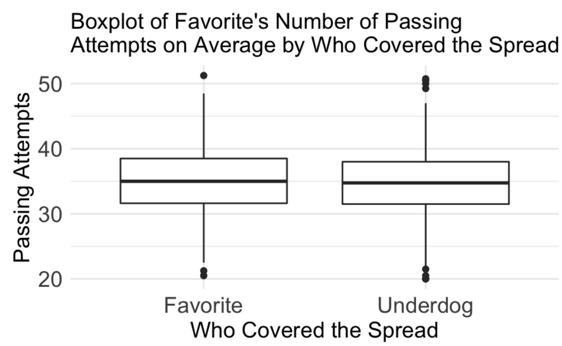
15. Favorite’s Number of Pass Completions, On Average, From Last 4 Games



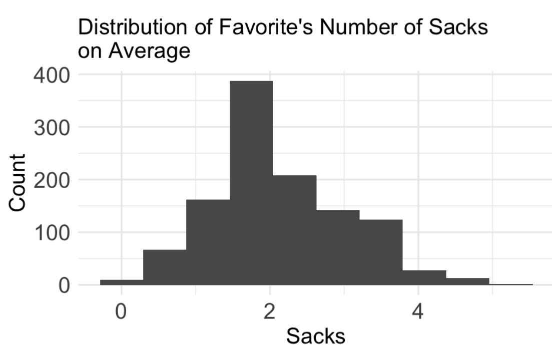


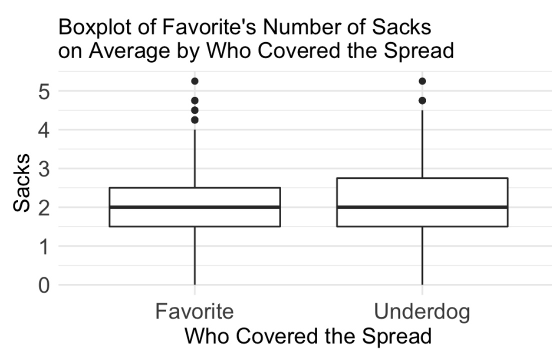
16. Favorite's Number of Pass Attempts, On Average, From Last 4 Games



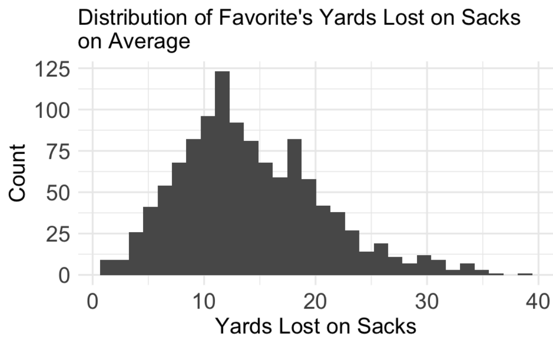


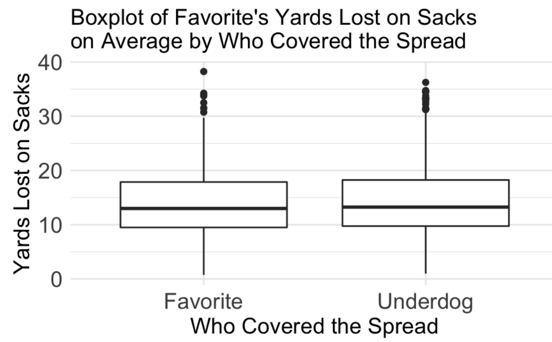
17. Favorite's Number of Sacks, On Average, From Last 4 Games



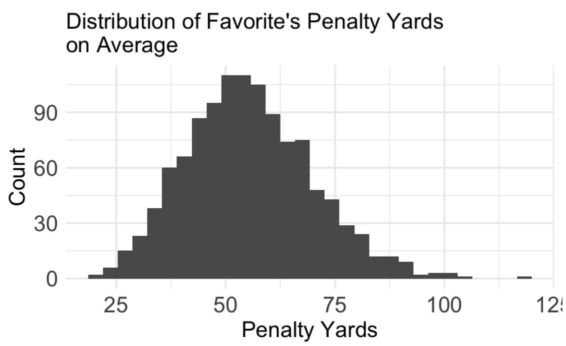


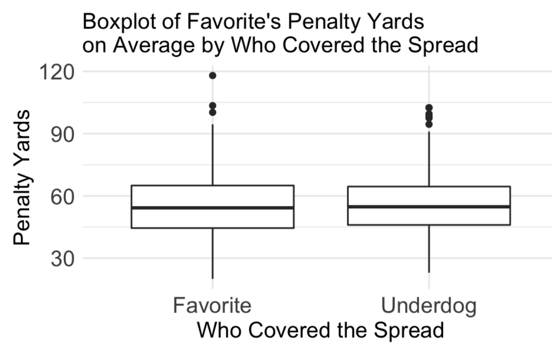
18. Favorite's Yards Lost on Sacks, On Average, From Last 4 Games



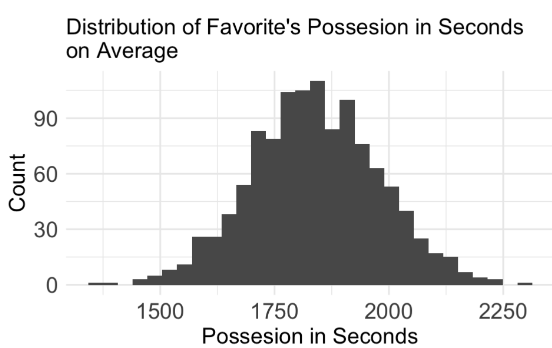


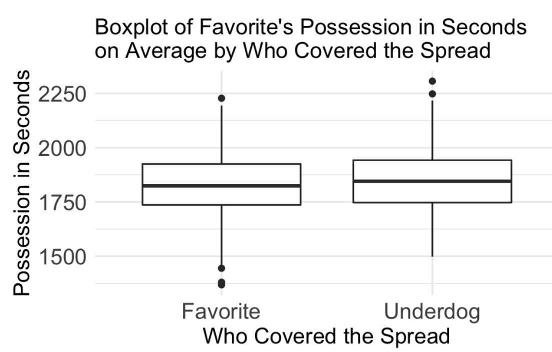
19. Favorite's Penalty Yards, On Average, From Last 4 Games



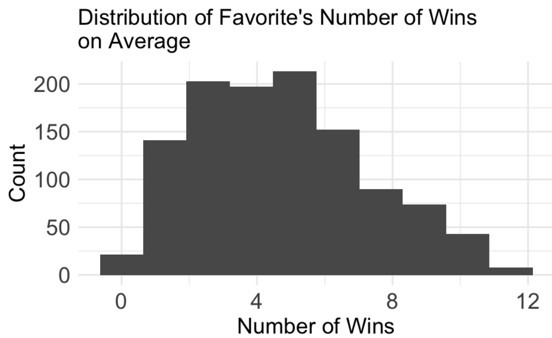


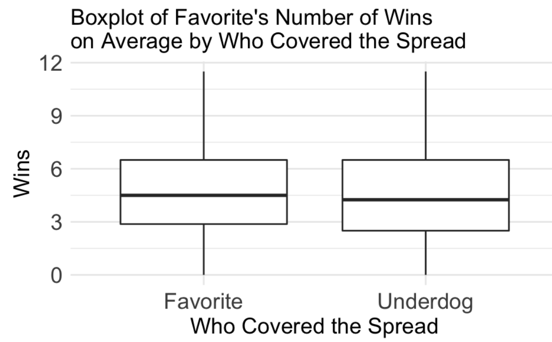
20. Favorite's Time of Possession (Seconds), On Average, From Last 4 Games



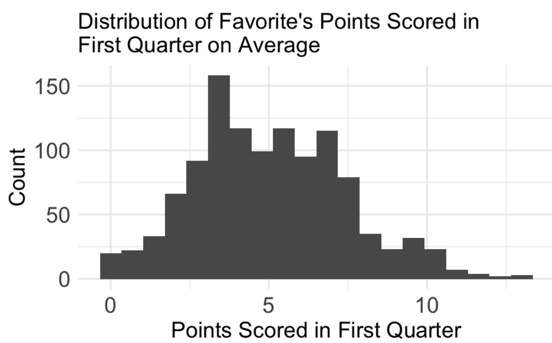


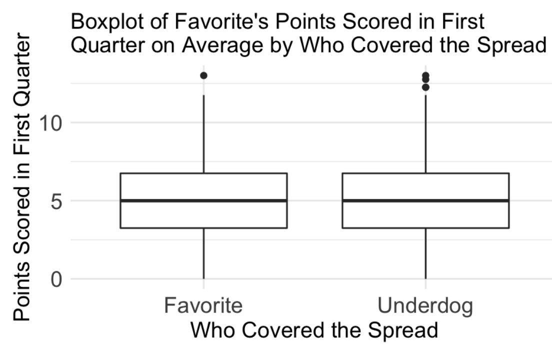
21. Favorite's Number of Wins, On Average, From Last 4 Games



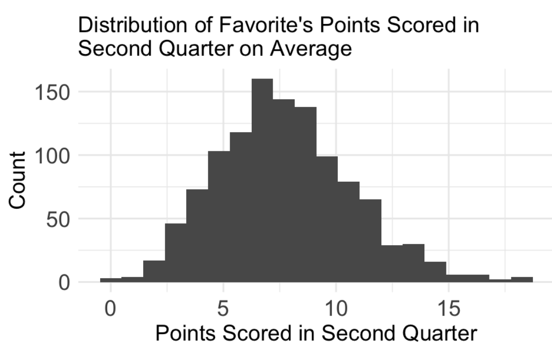


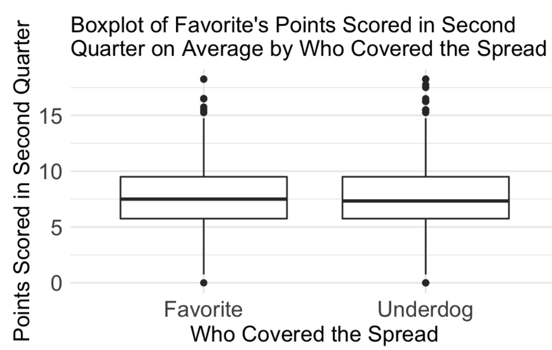
22. Favorite's Points Scored in First Quarter, On Average, From Last 4 Games



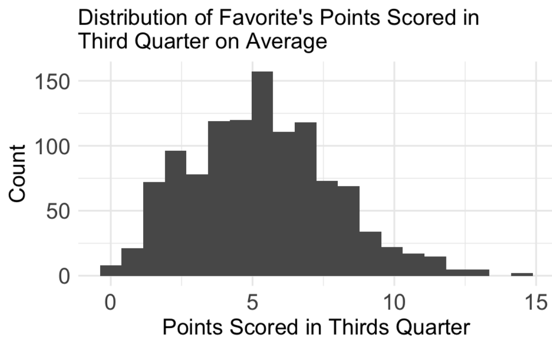


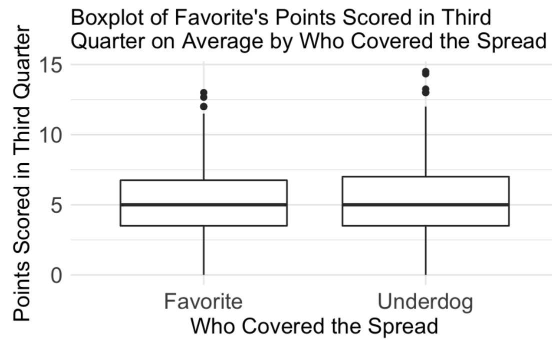
23. Favorite's Points Scored in Second Quarter, On Average, From Last 4 Games



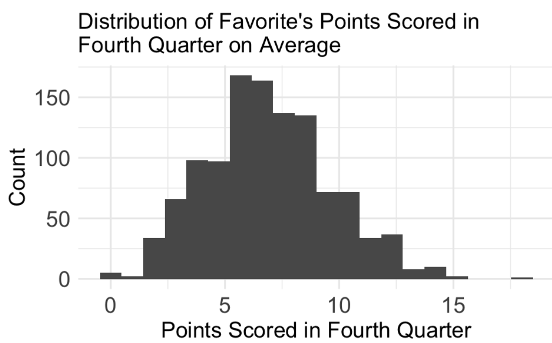


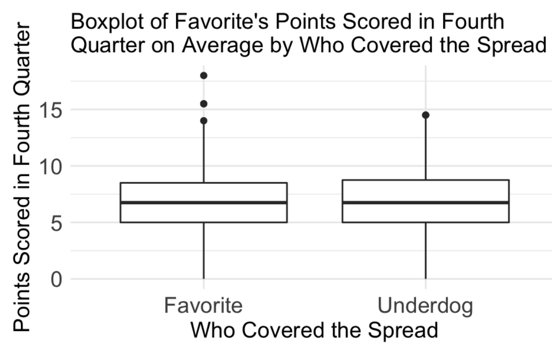
24. Favorite's Points Scored in Third Quarter, On Average, From Last 4 Games



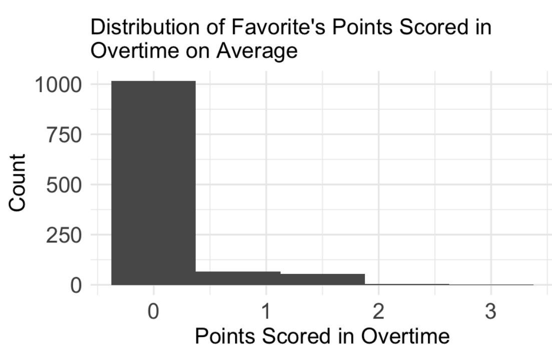


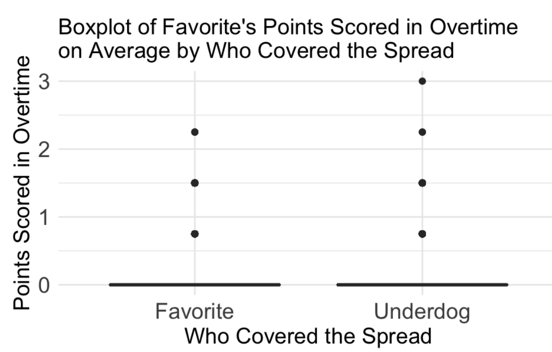
25. Favorite's Points Scored in Fourth Quarter, On Average, From Last 4 Games



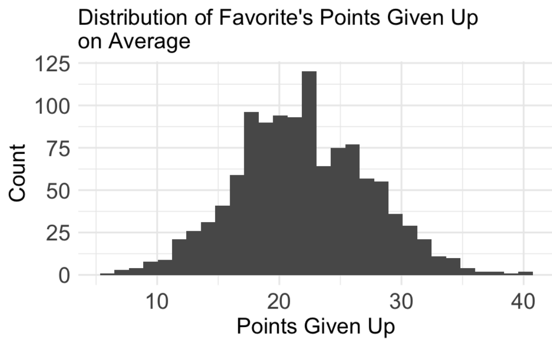


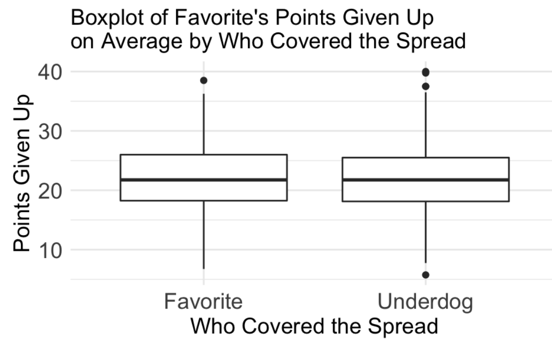
26. Favorite's Points Scored in Overtime, On Average, From Last 4 Games



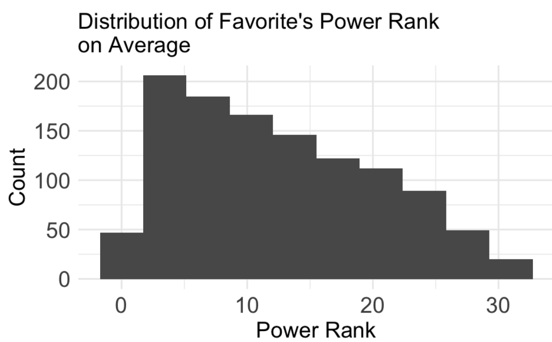


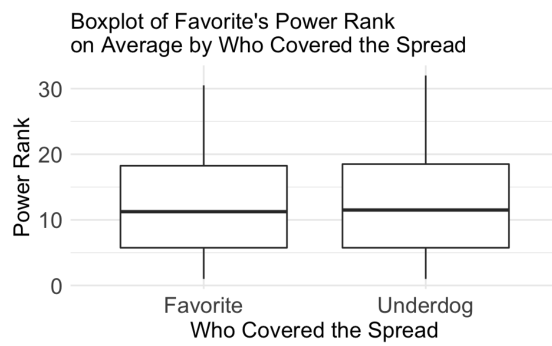
27. Favorite's Points Allowed, On Average, From Last 4 Games



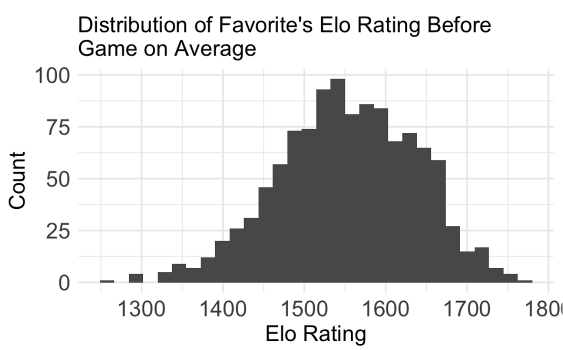


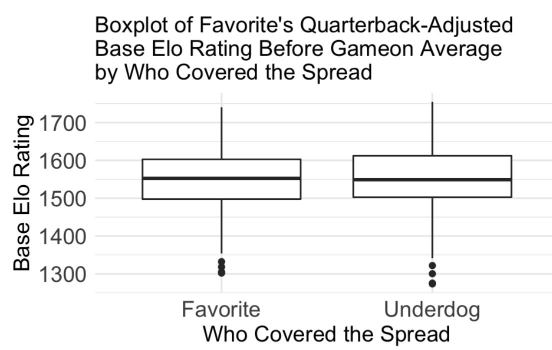
28. Favorite's Power Rank, On Average, From Last 4 Games



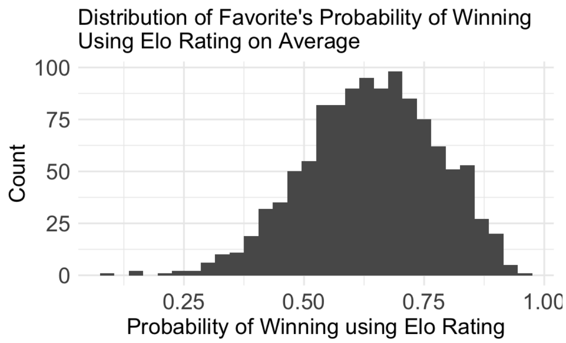


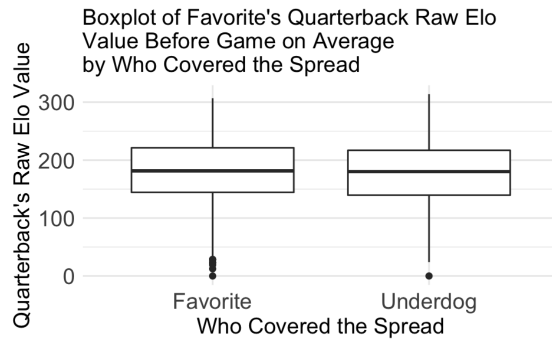
29. Favorite's Elo Rating Before Game, On Average, From Last 4 Games



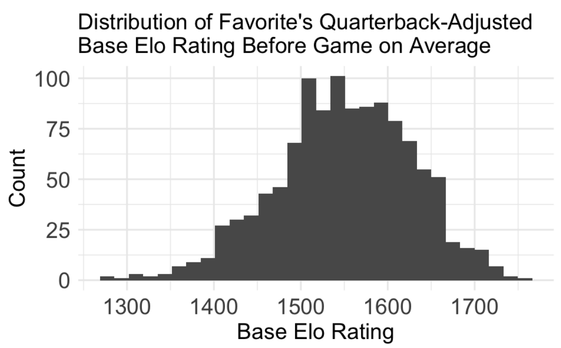


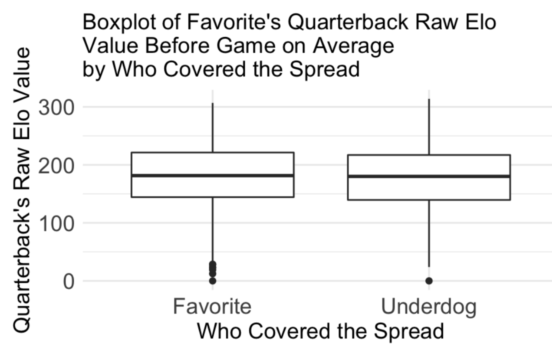
30. Favorite's Probability of Winning Using Elo Rating, On Average, From Last 4 Games



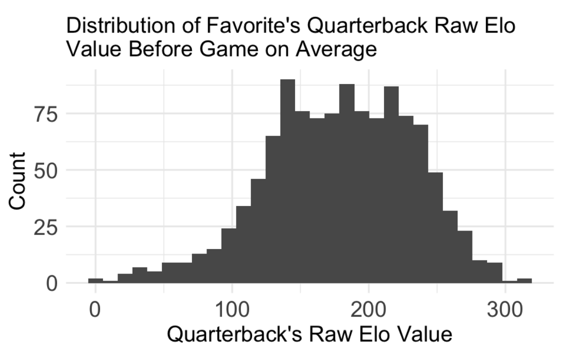


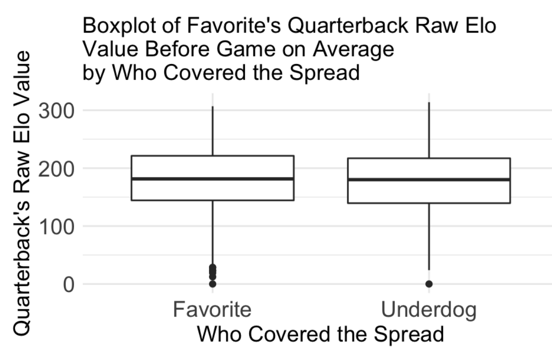
31. Favorite's Quarterback-Adjusted Elo Rating Before Game, On Average, From Last 4 Games



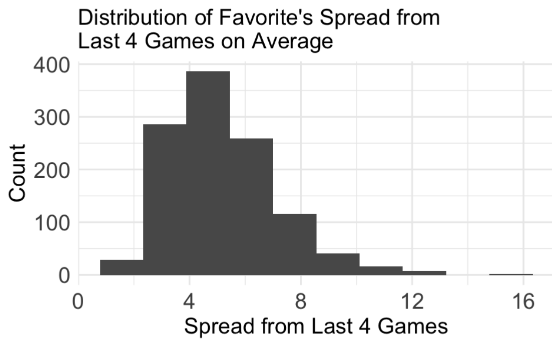


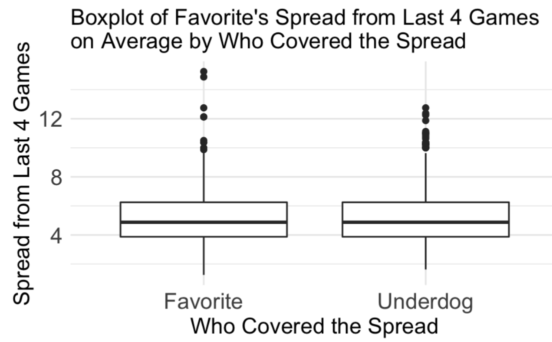
32. Favorite's Quarterback Raw Elo Value Before Game, On Average, From Last 4 Games



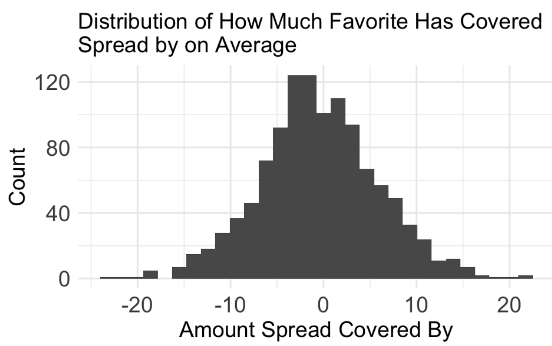


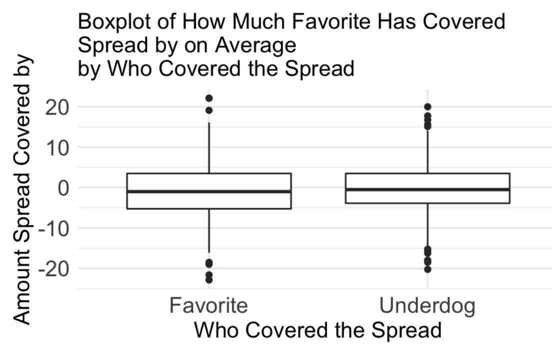
33. Favorite's Spread, On Average, From Last 4 Games



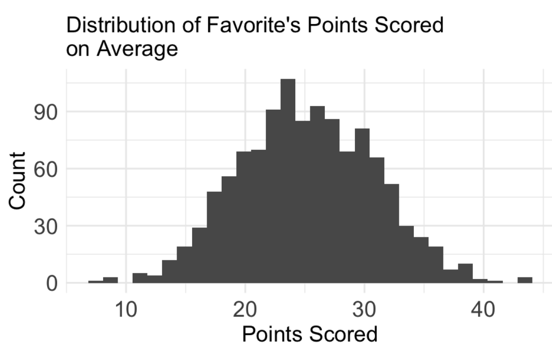


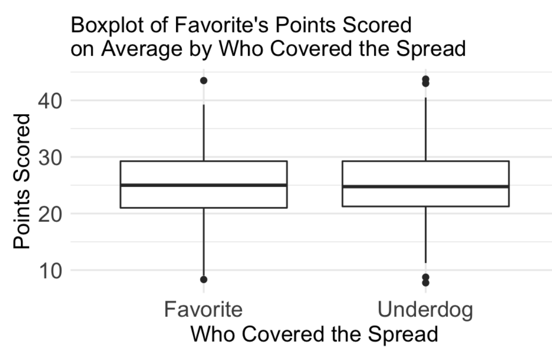
34. How Much Favorite Has Covered Spread by, On Average, From Last 4 Games



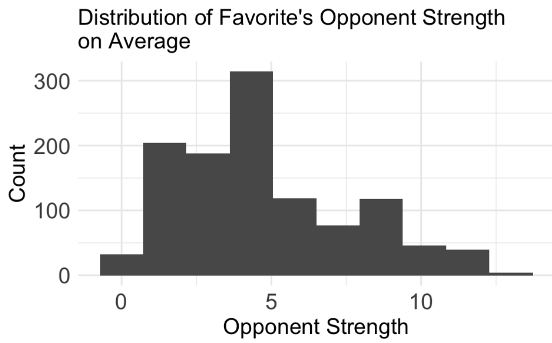


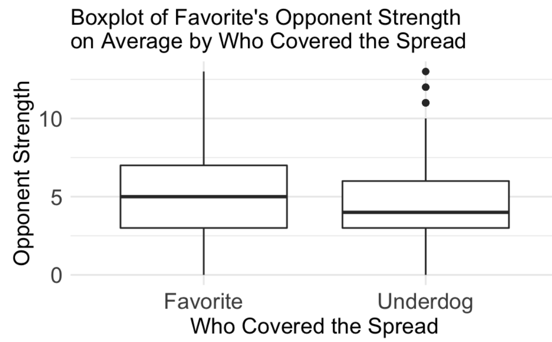
35. Favorite's Points Scored, On Average, From Last 4 Games





36. Favorite's Opponent Strength





8.4 Favorite Categorical Variables

The following section contains tables of the remaining categorical variables for the favorite. Despite some categories having discernible counts, these variables provided little to no predictability in determining who covered the spread.

1. Favorite's Number of Wins From Last 4 Games

Number of Wins	Count
0	47
1	193
2	356
3	372
4	174

	Who Covered Spread	
Number of Wins	Favorite	Underdog
0	25	22
1	95	98
2	183	173
3	173	199
4	91	83

2. Number of Times Favorite Covered Spread From Last 4 Games

Number of Times Favorite Covered Spread	Count
0	59
1	239
2	442
3	306
4	96

Number of Times Favorite Covered Spread	Who Covered Spread Current Week: Favorite	Underdog
0	33	26
1	117	122
2	226	216
3	142	164
4	49	47

3. Number of Times Favorite Was Favorite in Last 4 Games

Number of Times Favorite Was Favorite	Count
0	76
1	186
2	286
3	340
4	254

Number of Times Favorite Was Favorite	Who Covered Spread: Favorite	Underdog
0	37	39
1	89	97
2	147	139
3	171	169
4	123	131

4. Favorite's Number of First Downs by Penalty in Last 4 Games

Number of First Downs	Count
0 - 6	566
7+	576

Number of First Downs	Who Covered Spread: Favorite	Underdog
0 - 6	285	281
7+	282	294

5. Did Favorite Attempt to Convert on Fourth Down in Last 4 Games

Attempted 4th Down	Count
Yes	1001
No	141

Attempted 4th Down	Who Covered Spread: Favorite	Underdog
Yes	488	513
No	79	62

6. Favorite's Number of Offensive Drives in Last 4 Games

Total Drives	Count
0 - 44	540
45+	602

	Who Covered Spread	
Total Drives	Favorite	Underdog
0 - 44	253	287
45+	314	288

7. Favorite's Number of Interceptions Thrown in Last 4 Games

Interceptions	Count
0 - 2	526
3+	616

	Who Covered Spread	
Interceptions	Favorite	Underdog
0 - 2	260	266
3+	307	309

8. Favorite's Red Zone Efficiency in Last 4 Games

Efficiency	Count
0 - 60%	587
60%+	555

	Who Covered Spread	
Efficiency	Favorite	Underdog
0 - 60%	286	301
60%+	281	274

9. Favorite's Number of Turnovers in Last 4 Games

Turnovers	Count
0 - 4	529
5+	613

	Who Covered Spread	
Turnovers	Favorite	Underdog
0 - 4	257	272
5+	310	303

10. Favorite's Number of Fumbles Lost in Last 4 Games

Fumbles Lost	Count
0 - 1	465
2+	677

	Who Covered Spread	
Fumbles Lost	Favorite	Underdog
0 - 1	219	246
2+	348	329

11. Favorite's Number of Defensive or Special Teams Touchdowns in Last 4 Games

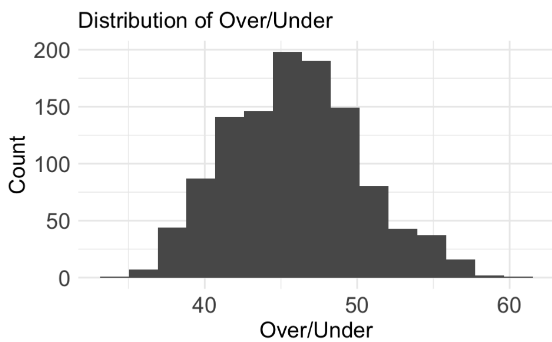
Touchdowns	Count
0	597
1+	545

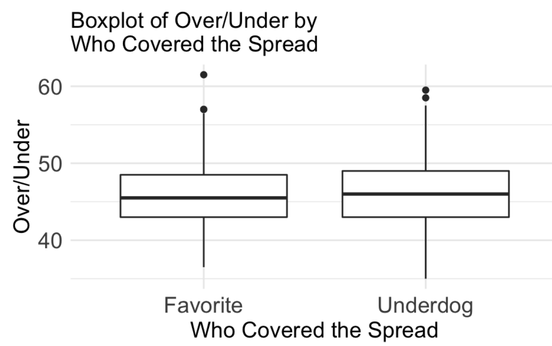
	Who Covered Spread	
Touchdowns	Favorite	Underdog
0	286	311
1+	281	264

8.5 Miscellaneous Numeric Variables

1. Over/Under, On Average, From Last 4 Games

The over/under is the total amount of points sportsbooks believe will be scored, on average, by both teams prior to the game. There is no discernible difference in who covered the spread based on the over/under value.





8.6 Feature Set Variables

1. Remove Correlated Variables

If variables had greater than 0.5 correlation with another variable, only one variable amongst the group of correlated variables was kept for this feature set. The following is a list of the variables that remained:

- The Current Week's Spread
- The Current Week's Over/Under
- Favorite's Opponent Strength on Average

- Underdog's Opponent Strength on Average
- Underdog's Number of First Downs on Average
- Underdog's Number of Rushing First Downs on Average
- Underdog's Number of Penalties on Average
- Underdog's Number of Sacks on Average
- Underdog's Number of Wins on Average
- Underdog's Number of Points in the First Quarter on Average
- Underdog's Number of Points in the Third Quarter on Average
- Underdog's Number of Points in the Fourth Quarter on Average
- Underdog's Number of Points in OT on Average
- Underdog's Number of Points Allowed on Average
- Favorite's Number of First Downs on Average
- Favorite's Number of Rushing First Downs on Average
- Favorite's Number of Penalties on Average
- Favorite's Number of Sacks on Average
- Favorite's Number of Points in the First Quarter on Average
- Favorite's Number of Points in the Second Quarter on Average
- Favorite's Number of Points in the Third Quarter on Average
- Favorite's Number of Points in the Fourth Quarter on Average
- Favorite's Number of Points in OT on Average
- Favorite's Number of Points Allowed on Average
- Underdog's Spread Last 4 Games on Average
- How Much the Underdog Covered the Spread by in Last 4 Games on Average
- Favorite's Spread Last 4 Games on Average

- How Much the Favorite Covered the Spread by in Last 4 Games on Average
- Underdog's Number of First Downs by Penalty in Last 4 Games
- Did Underdog Attempt to Convert on Fourth Down in Last 4 Games
- Underdog's Number of Offensive Drives in Last 4 Games
- Underdog's Number of Interceptions Thrown in Last 4 Games
- Underdog's Red Zone Efficiency in Last 4 Games
- Underdog's Number of Fumbles Lost in Last 4 Games
- Underdog's Number of Defensive or Special Teams Touchdowns in Last 4 Games
- Favorite's Number of First Downs by Penalty in Last 4 Games
- Did Favorite Attempt to Convert on Fourth Down in Last 4 Games
- Favorite's Number of Offensive Drives in Last 4 Games
- Favorite's Number of Interceptions Thrown in Last 4 Games
- Favorite's Red Zone Efficiency in Last 4 Games
- Favorite's Number of Fumbles Lost in Last 4 Games
- Favorite's Number of Defensive or Special Teams Touchdowns in Last 4 Games

2. Top Correlated Variables

The following variables are among the top 30% of the most correlated variables with the response variable:

- Underdog's Yards Per Pass on Average
- Underdog's Quarterback Elo Adjustment on Average
- Underdog's Third Down Efficiency on Average
- Underdog's Number of Sacks on Average
- Underdog's Number of Wins From Last 4 Games
- Favorite's Quarterback Elo Adjustment on Average

- Favorite's Number of Points Scored in Third Quarter on Average
- Favorite's Number of Wins From Last 4 Games
- Underdog's Elo Rating Before Game From Last 4 Games on Average
- Underdog's Power Rank on Average
- Underdog's Yards Lost on Sacks on Average
- Favorite's Number of Sacks on Average
- Number of Times Favorite Was Favorite in Last 4 Games
- Number of Times Underdog Was Favorite in Last 4 Games
- Underdog's Quarterback-Adjusted Elo Rating Before Game From Last 4 Games on Average
- Underdog's Number of First Downs on Average
- Favorite's Yards Lost on Sacks on Average
- Underdog's Completion Percentage on Average
- Favorite's Quarterback Raw Elo Value Before Game From Last 4 Games on Average
- Favorite's Number of Wins on Average
- Favorite's Quarterback-Adjusted Elo Rating Before Game From Last 4 Games on Average
- Favorite's Elo Rating Before Game From Last 4 Games on Average
- Favorite's Power Rank on Average
- Underdog's Quarterback Raw Elo Value Before Game From Last 4 Games on Average
- Favorite's Probability of Winning Using Quarterback-Adjusted Elo Rating on Average
- Underdog's Probability of Winning Using Quarterback-Adjusted Elo Rating on Average
- Underdog's Probability of Winning Using Quarterback-Adjusted Elo Rating on Average
- Favorite's Probability of Winning Using Quarterback-Adjusted Elo Rating on Average
- The Current Week's Spread

3. Remove Redundant Variables

After subsetting all the variables down to a set of non linearly dependent variables and removing variables that contained repetitive information, these are the variables that remained:

- The Current Week's Spread
- Favorite's Opponent Strength on Average
- Underdog's Opponent Strength on Average
- Underdog's Number of Passing First Downs on Average
- Underdog's Number of Rushing First Downs on Average
- Underdog's Third Down Efficiency on Average
- Underdog's Number of Passing Yards on Average
- Underdog's Yards Per Pass on Average
- Underdog's Number of Rushing Yards on Average
- Underdog's Yards Per Rush on Average
- Underdog's Number of Penalties on Average
- Underdog's Completion Percentage on Average
- Underdog's Number of Pass Attempts on Average
- Underdog's Number of Sacks on Average
- Underdog's Yards Lost on Sacks on Average
- Underdog's Number of Penalty Yards on Average
- Underdog's Time on Possession in Seconds on Average
- Underdog's Number of Wins on Average
- Underdog's Power Rank on Average
- Underdog's Number of Wins From Last 4 Games

- Favorite's Number of Passing First Downs on Average
- Favorite's Number of Rushing First Downs on Average
- Favorite's Third Down Efficiency on Average
- Favorite's Number of Passing Yards on Average
- Favorite's Yards Per Pass on Average
- Favorite's Number of Rushing Yards on Average
- Favorite's Yards Per Rush on Average
- Favorite's Number of Penalties on Average
- Favorite's Completion Percentage on Average
- Favorite's Number of Pass Attempts on Average
- Favorite's Number of Sacks on Average
- Favorite's Yards Lost on Sacks on Average
- Favorite's Number of Penalty Yards on Average
- Favorite's Time on Possession in Seconds on Average
- Favorite's Number of Wins on Average
- Favorite's Power Rank on Average
- Favorite's Number of Wins From Last 4 Games
- Underdog's Elo Rating Before Game From Last 4 Games on Average
- Underdog's Quarterback-Adjusted Elo Rating Before Game From Last 4 Games on Average
- Favorite's Elo Rating Before Game From Last 4 Games on Average
- Favorite's Quarterback-Adjusted Elo Rating Before Game From Last 4 Games on Average
- Underdog's Spread Last 4 Games on Average
- How Much the Underdog Covered the Spread by in Last 4 Games on Average

- How Many Times the Underdog Covered the Spread in Last 4 Games
- Number of Times Underdog Was Favorite in Last 4 Games
- Favorite's Spread Last 4 Games on Average
- How Much the Favorite Covered the Spread by in Last 4 Games on Average
- How Many Times the Favorite Covered the Spread in Last 4 Games
- Number of Times Favorite Was Favorite in Last 4 Games
- Underdog's Red Zone Efficiency in Last 4 Games
- Underdog's Number of Turnovers in Last 4 Games
- Favorite's Red Zone Efficiency in Last 4 Games
- Favorite's Number of Turnovers in Last 4 Games

CHAPTER 9

References

[1] Feb 22, CBS Sports Staff. "U.S. Sports Betting: Here's Where All 50 States Stand on Legalizing Sports Gambling, Betting Mobile Bets." *CBSSports.com*, 22 Feb. 2023, `https://www.cbssports.com/general/news/u-s-sports-betting-heres-where-all-50-states-stand-on-legalizing-sports-gambling-betting-mobile-bets/`.

[2] BetMGM. "What Did the Super Bowl Spread Open at?" *BetMGM*, 31 Jan. 2023, `https://sports.betmgm.com/en/blog/nfl/super-bowl-spread-open-bm07/`.

[3] Andrews, Phil. "How a Sportsbook Stays in Business." *Medium*, Medium, 18 Sept. 2018, `https://medium.com/@PhilipAndrews/how-a-sportsbook-stays-in-business-72e5f3b5243d`.

[4] BetMGM. "How Do NFL Odds Work? How to Read NFL Betting Odds Online." *BetMGM*, 2 Aug. 2022, `https://sports.betmgm.com/en/blog/how-to-read-nfl-betting-odds-online-spread-over-under-money-line/`.

[5] Culver, John V. "Why 52.4% Is the Most Important Percentage in Sports Gambling." *Medium*, Medium, 20 Oct. 2018, `https://medium.com/the-intelligent-sports-wagerer/why-52-4-is-the-most-important-percentage-in-sports-gambling-16ade8003c04`.

[6] SportsOddsHistory. "Historical NFL Game Odds." *SportsOddsHistory*, `https://www.sportsoddshistory.com/nfl-game-season/?y=2016`.

[7] ESPN. "NFL Scoreboard." *ESPN*, ESPN Internet Ventures, `https://www.espn.com/nfl/scoreboard/_/week/1/year/2016/`.

[8] ESPN. "Panthers vs. Broncos - NFL Team Statistics - September 8, 2016." *ESPN*, ESPN Internet Ventures, https://www.espn.com/nfl/matchup/_/gameId/400874484.

[9] ESPN. "2016 Final Regular-Season Power Rankings: Patriots Rise to No. 1." *ESPN*, ESPN Internet Ventures, 3 Jan. 2017, https://www.espn.com/nfl/story/_/id/18395280/nfl-2016-final-regular-season-power-rankings-new-england-patriots-dallas-cowboys-pittsburgh-steelers.

[10] Boice, Jay. "NFL Elo." *GitHub*, FiveThirtyEight, 14 Oct. 2021, https://github.com/fivethirtyeight/data/tree/master/nfl-elo.

[11] Silver, Nate. "How Our NFL Predictions Work." *FiveThirtyEight*, FiveThirtyEight, 4 Sept. 2014, https://fivethirtyeight.com/methodology/how-our-nfl-predictions-work/.

[12] pandas. "Pandas.Dataframe.rolling." *Pandas.DataFrame.rolling - Pandas 1.5.3 Documentation*, https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.rolling.html.

[13] scikit-learn. "sklearn.model_selection.train_test_split." *scikit-learn*, https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html.

[14] scikit-learn. "sklearn.preprocessing.StandardScaler." *scikit-learn*, https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html.

[15] scikit-learn. "3.1. Cross-Validation: Evaluating Estimator Performance." *scikit-learn*, https://scikit-learn.org/stable/modules/cross_validation.html.

[16] scikit-learn. "sklearn.metrics.mean_squared_error." *scikit-learn*, https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html.

[17] scikit-learn. "sklearn.linear_model.ElasticNet." *scikit-learn*, https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.html.

[18] scikit-learn. "sklearn.ensemble.RandomForestRegressor." *scikit-learn*, https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html.

[19] scikit-learn. "Gradient Boosting Regression." *scikit-learn*, https://scikit-learn.org/stable/auto_examples/ensemble/plot_gradient_boosting_regression.html.

[20] scikit-learn. "Sklearn.model_selection.GridSearchCV." *scikit-learn*, https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html.

[21] Wikipedia. "Kelly Criterion." *Wikipedia*, Wikimedia Foundation, https://en.wikipedia.org/wiki/Kelly_criterion.