Sports Analytics: Final Project Proposal

Bryce Dean

Evan Callaghan

November 5, 2021

Abstract

Across the United States, there is a strong push to get people involved in mobile and in-person sports betting. This industry is growing at an enormous rate, and so is the amount of money that bettors are risking every week. Is there a way to take advantage of this gambling? Can the average bettor make money in this industry? What if, instead of looking at a game with a "gut feeling," there is a strong and dependable analytical approach to placing bets. This project aims to analyze NFL matchup data and build predictive models to offer an analytical approach to sports betting. It focuses on the three main areas for American bettors: money line bets, spread bets, and total over/under bets.

Project Introduction

In-person and mobile sports betting is becoming increasingly apparent across the United States. The landmark Supreme Court decision in 2018 that lifted the federal ban of sports gambling opened the doors for states to establish and regulate their own sports betting markets. Advertisements for sports betting sites or sportsbooks appear in virtually every televised sporting event, and billboards line the streets with the odds for the upcoming big game. There is clearly a strong push from casinos and sportsbooks to get people involved in the industry and bet actively. These organizers have a lot to gain from the average gambler. Already making millions of dollars every year, casinos and sportsbooks will not slow down as the popularity of sports betting continues to rise.

Now, what if there is a better betting strategy? What if, instead of looking at a game with a "gut feeling," there is a strong and dependable analytical approach to placing bets. Instead of seeing only a risk, we can attack bets with a solid strategy and windup making money. This is the basis for our final project. Looking at NFL matchup data and individualized team statistics, we want to build a regression model to predict how many points each team will score in a particular game. This prediction opens the door to the most popular kind of bets in football: money line bets, spread bets, and total over/under bets. We want to create a strong and consistent analytical method for approaching a bet, as opposed to having a "gut feeling" about a particular team or proposed score.

Data Description

The final NFL matchup data used for this project was aggregated from multiple sources using web-scraping techniques. The original data set, spreadspoke_scores.csv, was obtained from Kaggle, an online community for data scientists and provider of public data sets. To complete the final data set, several tables were scraped from "TeamRankings," a publicly available repository of sports statistics and data. All data sets and tables were aggregated to create the final data set. Then, we performed data cleaning and variable engineering to produce a data set ready for analysis. The final NFL matchup data set consists of 2,923 rows and 43 columns with no missing values. Each row represents an NFL game played between the 2010 and 2020 seasons.

The variable descriptions for the final data sets are displayed below:

Variable Name	Data Type	Description
schedule_date	Date	Date of the game (mm/dd/yyyy)
schedule_season	Categorical	NFL season year (2010, 2011, …, 2020)

schedule_week	Categorical	NFL season week (1, 2, … , Conference, Super Bowl)
schedule_playoff	Categorical	Whether or not the game is a playoff game
team_home/away	Categorical	Team name
score_home/away	Numerical	Number of points scored in the game
stadium	Categorical	Name of the game stadium
stadium_neutral	Categorical	Whether or not the stadium is at a neutral site
team_home/away_ppg	Numerical	Points per game for the given season
team_home/away_oppg	Numerical	Opposition points per game for the given season
team_home/away_ypg	Numerical	Yards per game for the given season
team_home/away_oypg	Numerical	Opponent yards per game for the given season
team_home/away_fdpg	Numerical	First downs per game for the given season
team_home/away_ofdpg	Numerical	Opponent first downs per game for the given season
team_home/away_tapg	Numerical	Takeaways per game for the given season
team_home/away_tpg	Numerical	Turnovers per game for the given season
team_home/away_pypg	Numerical	Penalty yards per game for the given season
team_home/away_opypg	Numerical	Opponent penalty yards per game for the given season
team_home/away_asm	Numerical	Average scoring margin for the given season
over_under_line	Numerical	Set over-under line for the given game
real_over_under_line	Numerical	Total points scored for the given game
over_under	Categorical	Whether or not the total score was over the line
over_under_push	Categorical	Whether or not the total score was equal to the line
favorite_team	Categorical	Name of team favored to win
favorite_team_score	Numerical	Points scored by the team favored to win
underdog_team	Categorical	Name of team not favored to win
underdog_team_score	Numerical	Points scored by the team not favored to win
spread_favorite	Numerical	Points the favored team is expected to win by
favorite_team_win	Categorical	Whether or not the favored team won the given game
favorite_team_cover	Categorical	Whether or not the favored team won by more than the spread

Exploratory Analysis

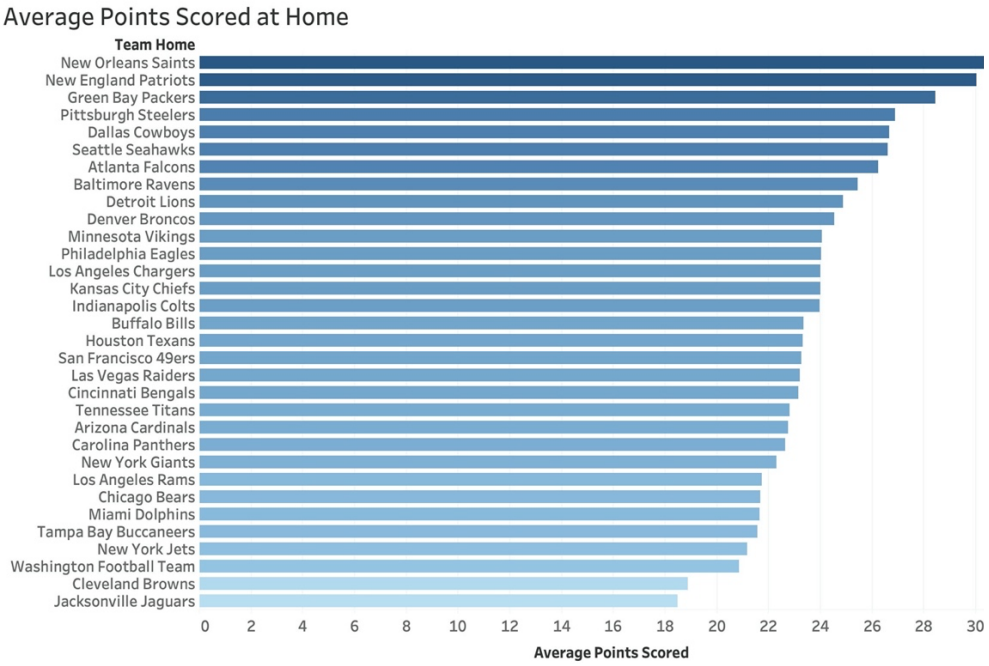

Figure 1:
The displayed bar chart ranks the 32 NFL teams in order of points per home game. We can see that the New Orleans Saints and the New England Patriots lead the way with just over 30 points per home game with the Cleveland Browns and Jacksonville Jaguars at the bottom of the rankings with just under 20 points per home game.

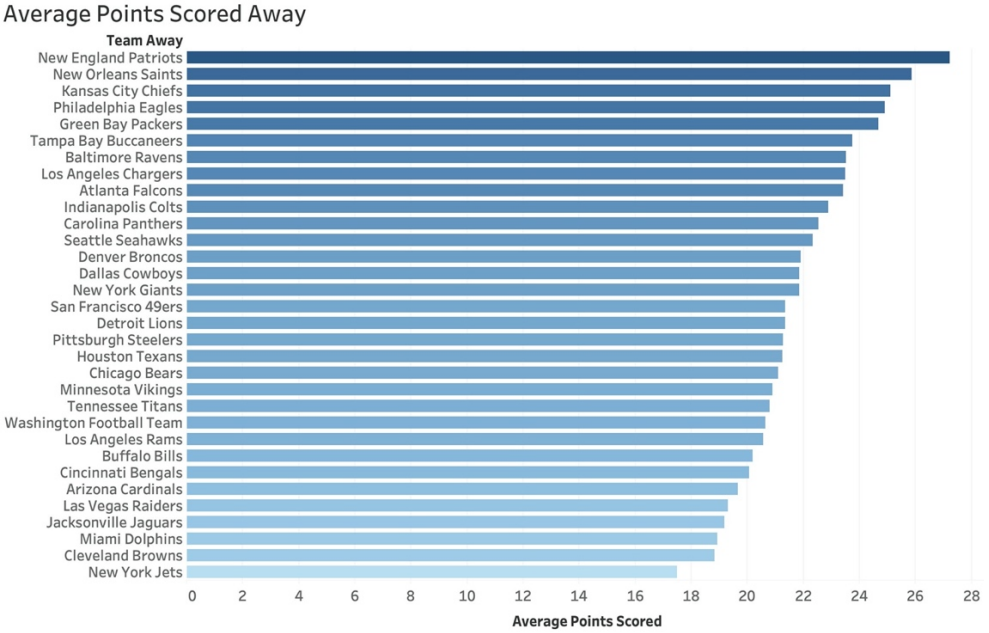

Figure 2:
Like *Figure 1,* this bar chart shows the average points scored for away games. Again, we see the New Orleans Saints and New England Patriots at the top of the rankings with around 26 or 27 points scored per away game. At the bottom of the rankings, we have the Cleveland Browns and the New York Jets, average around 18 points per away game.

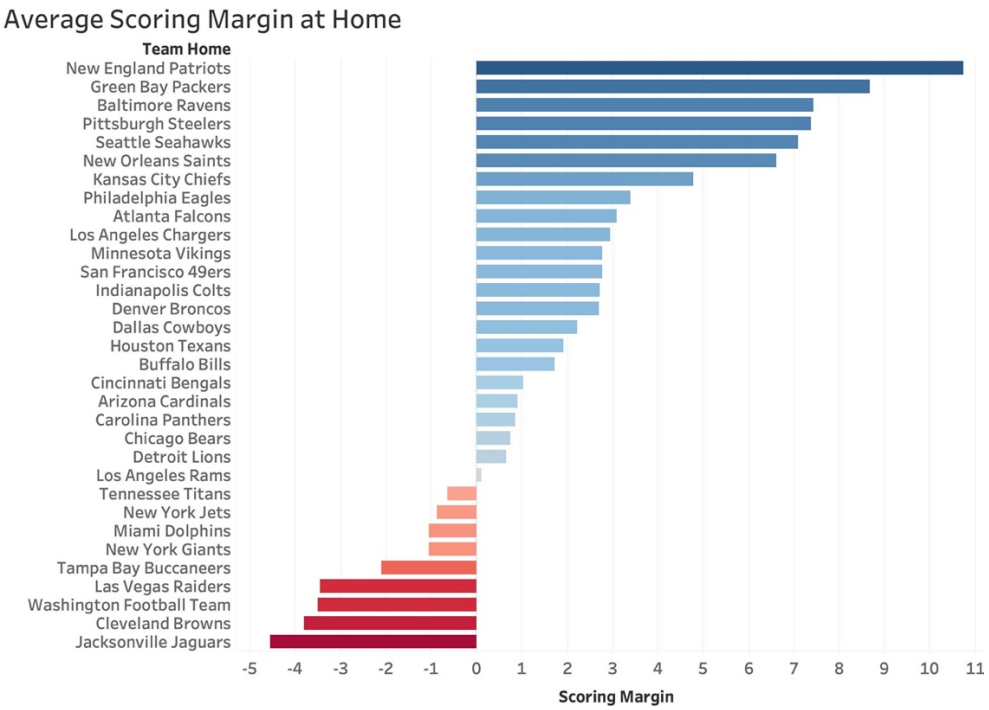

Figure 3:

This bar chart shows the average scoring margin for home games between the 2010 and 2020 NFL seasons. A positive value indicates a winning average, and a negative value indicates a losing average. We see that the New England Patriots defeated their opponents by, on average, more than 10 points at home games. On the other end, we see that the Jacksonville Jaguars lost to their opponents by, on average, more than four points at home games.

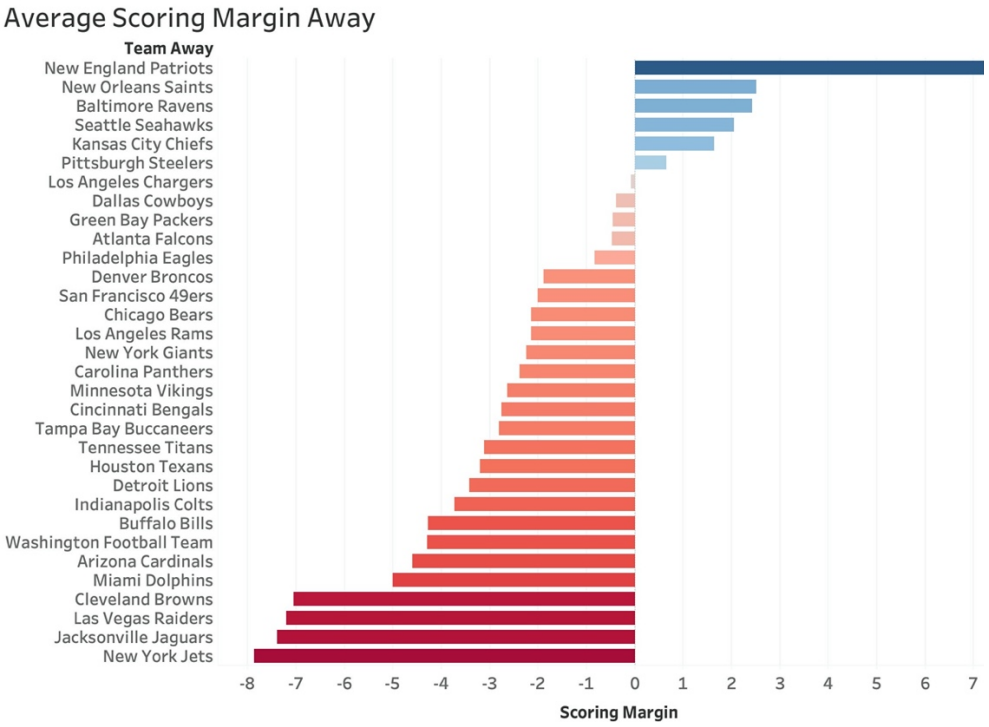

Figure 4:

Like Figure 3, the bar chart shows the average scoring margin for away games. Again, we see the Patriots leading the way, defeating their opponent by seven points, on average, for away games. At the bottom, we see the Jets, losing to their opponent by about eight points on average for away games. Notice that more teams have a positive scoring margin at home, while having a negative scoring margin on the road. Assuming that the strength of opponents at home is to equal to away, this could indicate some sort of home-field advantage.

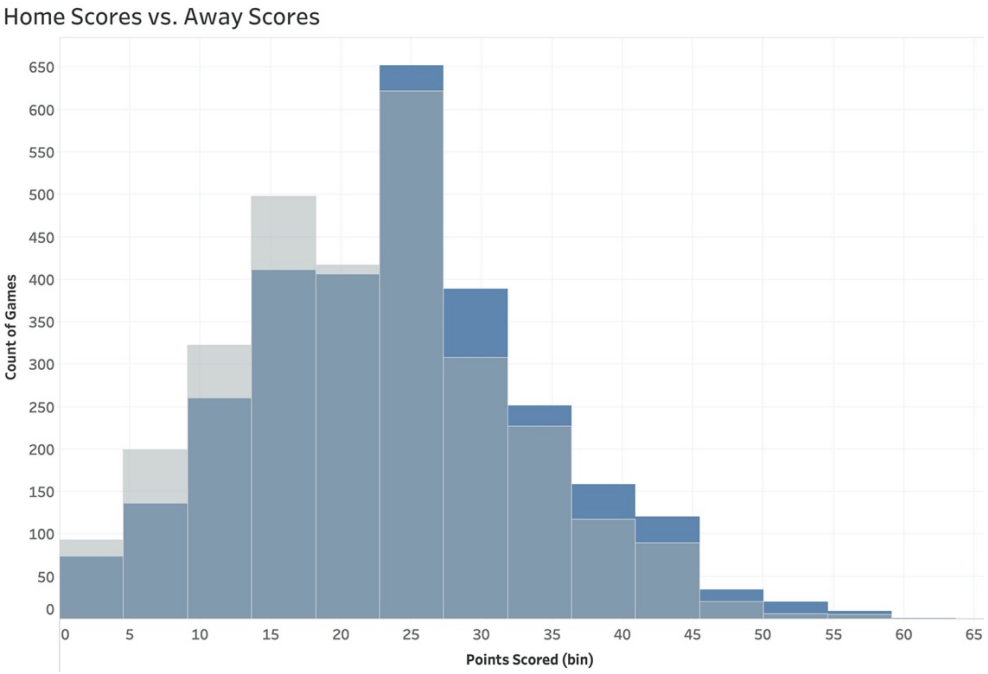

Figure 5:
The histogram shows points scored for the home team (blue) vs. the away team (grey). We see more home teams with higher scores, and more away teams with lower scores. This is supported as the mean score for the home team is about 23.9 while the mean score for the away team is about 21.8. This shows that playing at home might bring a competitive advantage to the game.

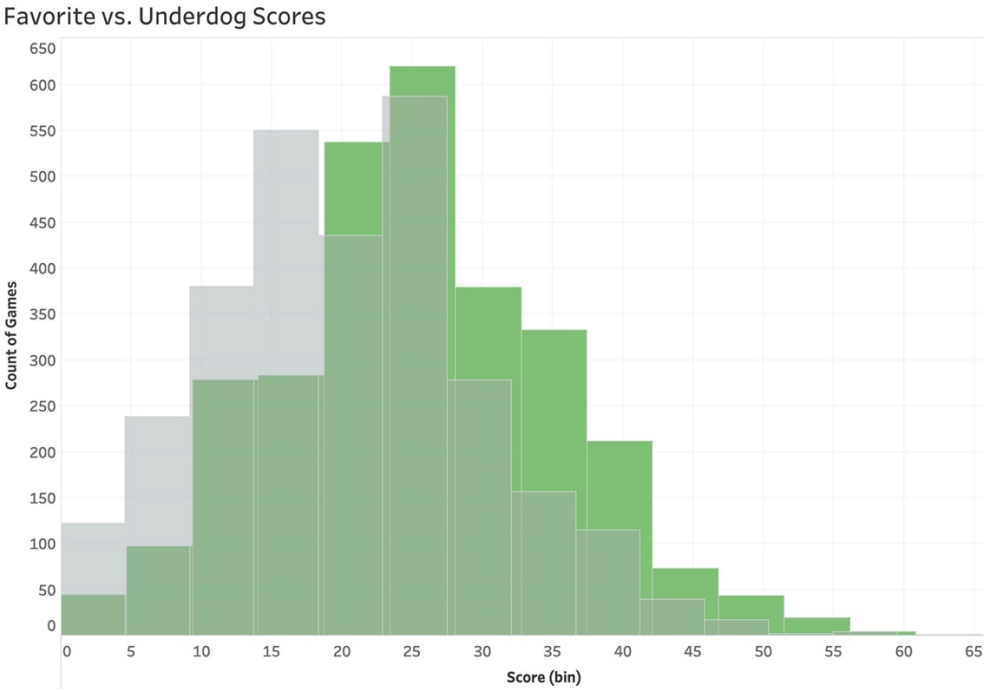

Figure 6:
The histogram shows points scored for the favored team (green) vs. the underdog team (grey). We see a significant difference between the points scored by the teams. The favorite team scores, on average, 25.4 points, while the underdog team scores, on average, 20.3 points. This is a difference of 5.1. The average favored spread (how much the favored team is expected to win by) is 5.3. This shows the accuracy of the oddsmakers.

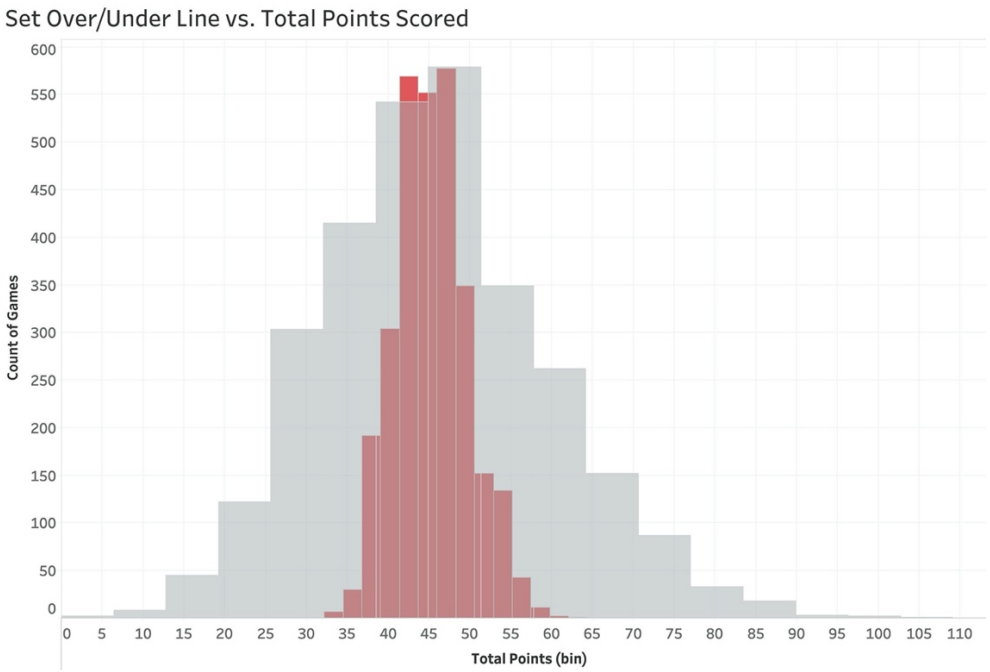

Figure 7:
This histogram displays the set over/under score (red) and the actual total amount of points scored in the game (grey). It is clear that oddsmakers do not stray far from the mean value of 45.3 points. The distribution has a standard deviation of 4.4 points. On the other hand, the total points scored has a mean of 45.7 points and a standard deviation of 14. This is a very wide dispersion compared to the set over/under line and could represent a potential vulnerability for the oddsmakers.

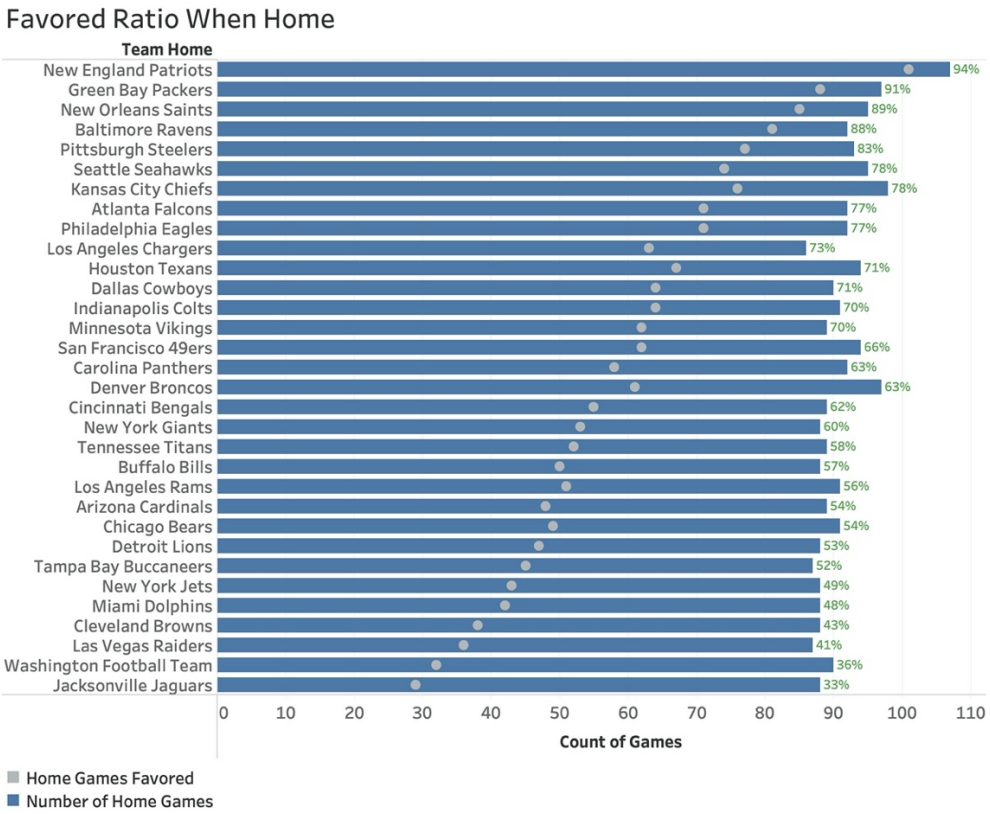

Figure 8:
The blue bar shows the total number of home games played and the grey dot represents the number of times each team was favored to win. The teams are ranked in order of home favored rate, the proportion of favored home games to number of home games. The Patriots lead the way with a 94% home favored rate and the Jacksonville Jaguars are at the bottom of the rankings, being favored only 33% of the time for home games. Only six teams are favored less than 50% of the time at home.

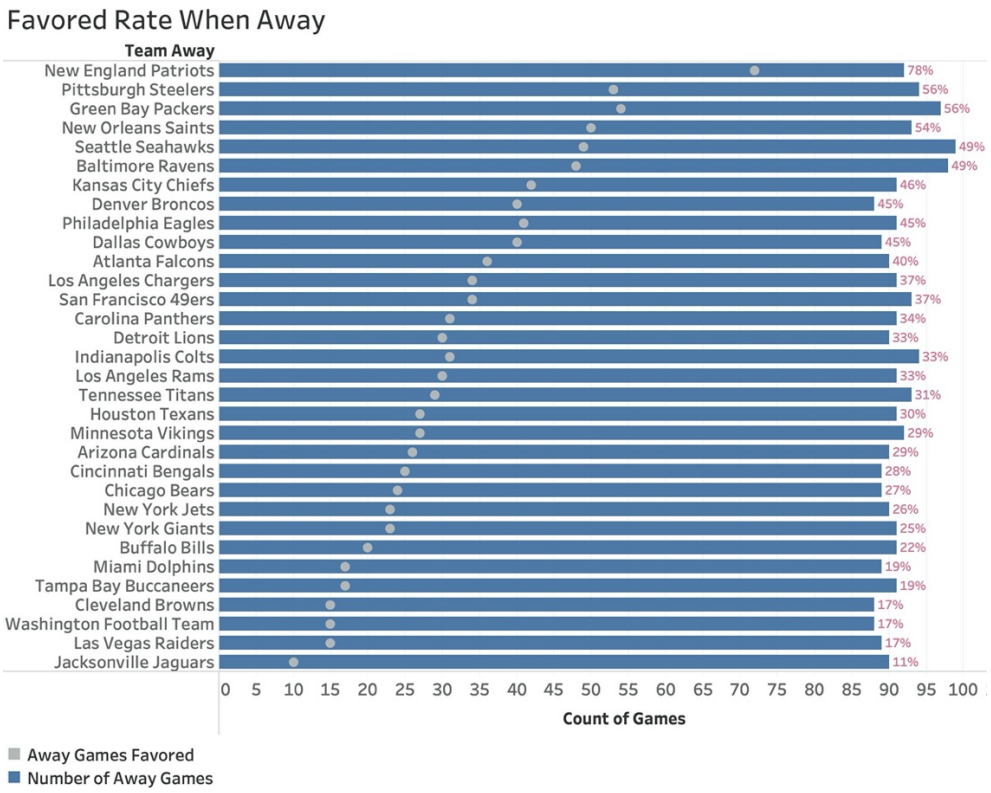

Figure 9:

Similar to the previous chart, the blue bar shows the total number of away games played and the grey dot represents the number of times each team was favored to win. The teams are ranked in order of away favored rate, the proportion of favored away games to number of away games. Again, the Patriots lead the way with a 78% away favored rate and the Jacksonville Jaguars are at the bottom of the rankings, being favored only 11% of the time for away games. Only four team are favored more than 50% of the time on the road.

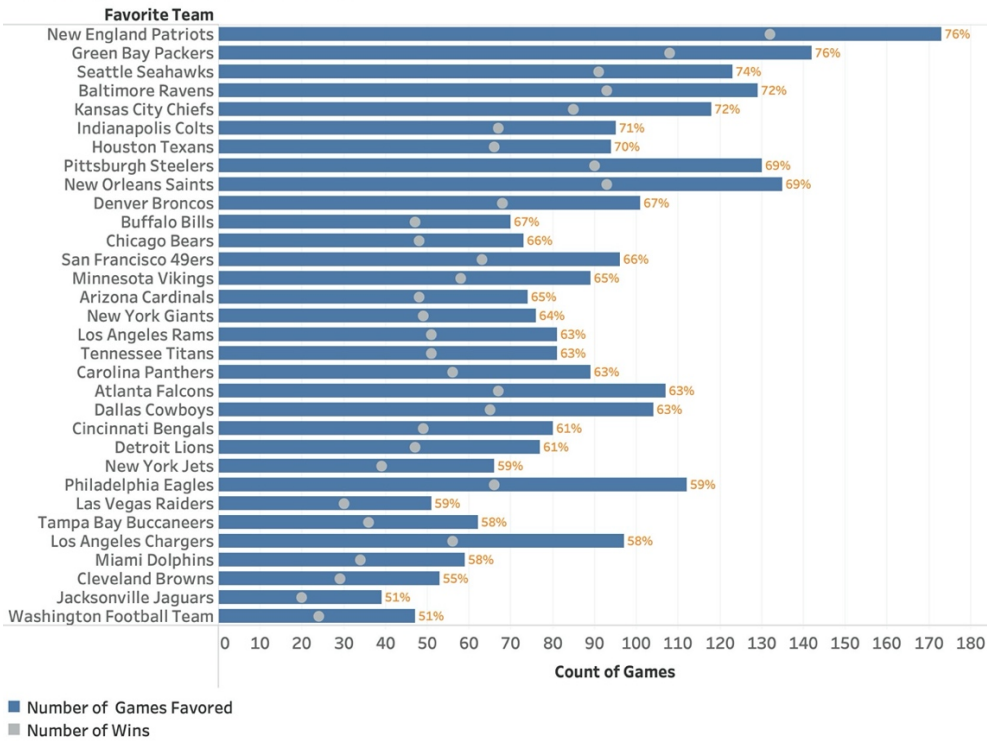

Figure 10:

The blue bar represents the number of games favored and the grey dot indicates how many of the favored games the given team won. The teams are ranked in order of favored winning rate, the proportion of favored wins to the total number of games favored. We see that the Patriots and Packers won about 76% of the games that they were favored to win, the best proportion in the NFL from 2010 to 2020. At the other end, the Jaguars and the Football Team only won about 51% of the games they were favored in.

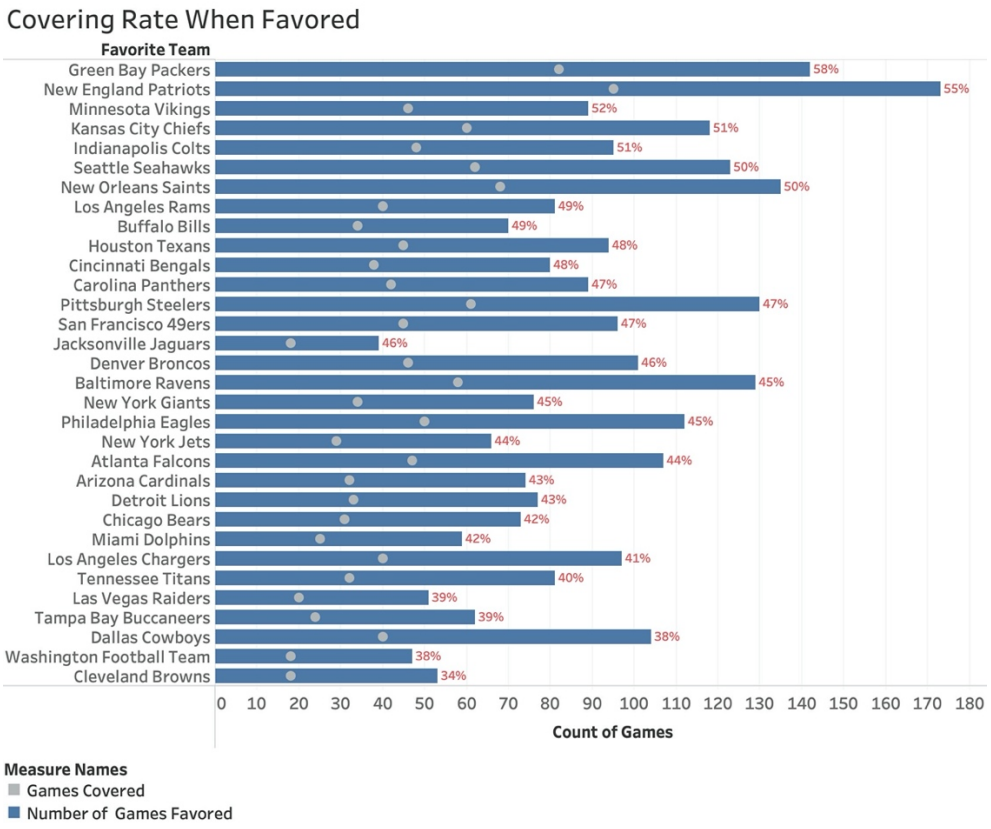

Figure 11:
The blue bar represents the number of games favored and the grey dot indicates how many of the favored games the given team covered the projected spread. The teams are ranked in order of favored covering rate, the proportion of times covering the spread to the total number of games favored. The Packers covered about 58% of the games they were favored, while the Browns covered about 34% of the games they were favored. The covering rate values are centered around 50%, indicating the strength of the oddsmakers to predict the spread value.

Next Steps

After completing the initial exploratory analysis, we have a much better understanding of our data and how to complete our goals moving forward. In the next stage of this project, we will be building and evaluating several regression models to predict the score_home and score_away variables. With this information, we have an analytical approach to placing money line bets, spread bets, and total over/under bets for NFL games. After the evaluation stage, the selected model will be rolled-out on new data: NFL matchups from this season. We will then see how the model performed upon implementation and see if this analytical approach offers an improved betting strategy in the long run.

References

Data set available at: https://www.kaggle.com/tobycrabtree/nfl-scores-and-betting-data?select=spreadspoke_scores.csv

Web scraped tables available at: https://www.teamrankings.com/nfl/team-stats/