DSC 630 Predictive Analytics

Final Project Milestone 3

Michael Loos

Predictive Analytics to Find an Edge in Sports Gambling

**Abstract**

Legal sports gambling is spreading across the country after the Supreme Court ruled The Professional and Amateur Sports Protection Act (PASPA) unconstitutional. This has opened the door to betting on college football games. Bettors and sports books seek to gain an advantage in the betting market. Historical data was used to predict the scores of college football games to help improve betting profitability by winning bets over 52.4% of the time. Major statistics for each team were used in regression models to predict the score of each team. The findings indicate that previous scores, game statistics, and team talent are major contributors to points scored in a game. This allows us to more accurately predict the final score of each game and increase the number of successful bets.

**Background**

The are many reasons one could have for watching sports. Some watch just because they love the game. Most are likely a fan of a particular team and watch that team frequently. Others, might find themselves watching a game because they have a money on the line. While sports gambling has been legal in Las Vegas and the rest of Nevada since 1949, only a few other states were able to offer gambling on sports for quite some time (Licata 2019). This changed on May 14, 2018, when the Supreme Court ruled in favor of the states, declaring The Professional and Amateur Sports Protection Act (PASPA) unconstitutional (Liptak & Draper 2018). This gave individual states the ability to legalize sports gambling. As of February 3rd, 2021, 19 states have legalized sports gambling, with several other states pending for reasons such as finalizing final rules and details in the legislative process (Butler 2021). While sports gambling has been illegal in most states, Americans clearly have an appetite for betting on sports. The American Gaming Association’s current estimate for illegal sports betting is $150 billion per year (Wasiolek 2020). Legalization of sports gambling across the country presents a huge opportunity for sports books, gamblers, and states through taxation.

**Problem Statement**

As new sportsbooks open up across the country, it will be important to both sides of each bet, the better and the sportsbook/house, to have an understanding of each bet and look for any possible advantages. The most common types of bets are: money line, a bet against the point spread, and betting the total (or over/under). A money line bet is simply picking which team you expect to win a given game/match. If that team wins, yo win. Betting on the favored team will win less money, while betting on the underdog will pay out more. Betting against the spread is a handicap bet where a spread is set, say 7 points. If you bet on the favorite, that team would have to win by more than 7 points for the bet to be a winner. Conversely, if you bet on the underdog, the bet is a winner as long as the team loses by less than 7 points (or wins). A 7-point win by the favorite would be a push and the better would get their original money back. A total bet is a bet against the total number of points score. A bet on the over would be betting more points are scored than the set number and a bet on the under would be betting fewer points are scored than the set number. For most bets against the spread or total, the payout will be close to 1:1, but not quite. One of the most common odds for bets against the spread or total is -110. This means to win $100, you must bet $110. This means you must win 52.4% of your bets to break even. It seems plausible that, over a large volume, anyone should be able to win at least 50% of their bets. This extra 2.4% is what requires solving. As a sportsbook, you need to be certain that you set your odds well enough to keep bettors under a 52.4% win-rate to maintain a profit. As a bettor, you need to find an edge to win more than 52.4% of your bets and earn a profit. Is it possible for this to be done using current and historical team data to improve bet selection? The goal of this project is to use this data and machine learning to test different models in order to correctly predict over 52.4% of bets, allowing for a profit.

**Methods**

Data from the 2019 college football season will be used. Data will be pulled from the API on CollegeFootballData.com. This website provides many different data sets, ranging from game results to team statistics to betting information. Key fields will include, but are not limited to:

* Week: week of season
* Home Team: home team of game
* Away Team: visiting team of game
* Home Points: total points scored by home team
* Away Points: total points score by away team
* Over Under: set line for point total of game
* Spread: point spread of game

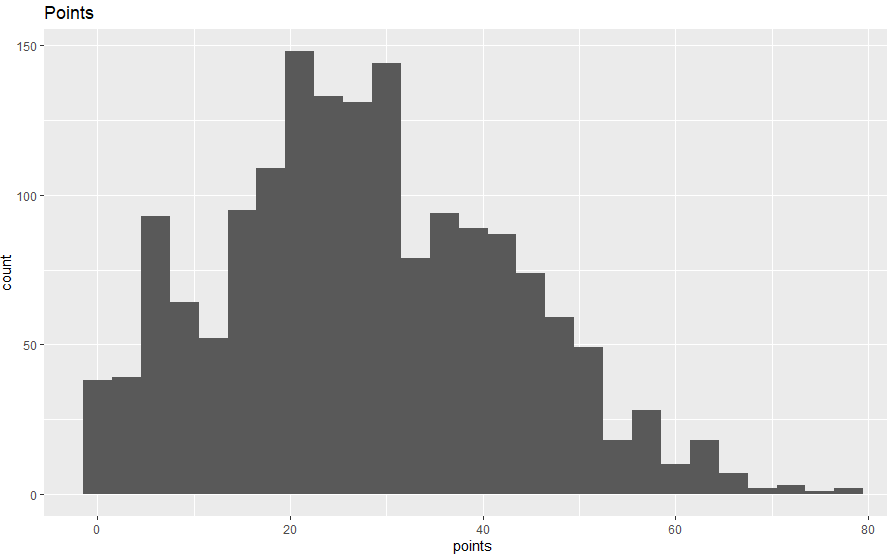
Games from the 2019 college football season were used to train and test the models, with a goal of accurately predicting the score of each game. This enables us to correctly predict the winner of each game, the winner against the spread, and the point total for the game. If we can confidently predict scores that range on one side of the spread and or total in either direction, that allows us to predict the outright winner, winner against the spread, and the over or under of the points total. This is considered a successful outcome.

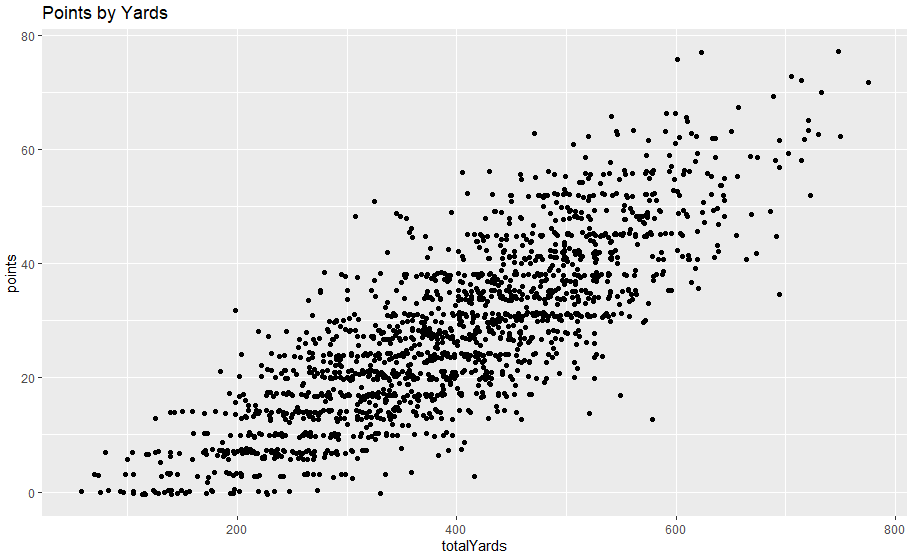
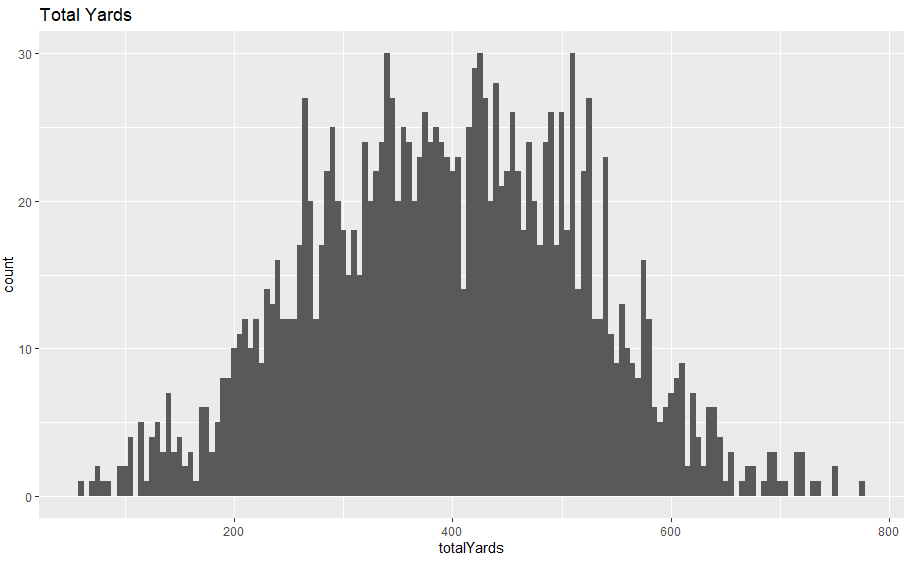
**Analysis**

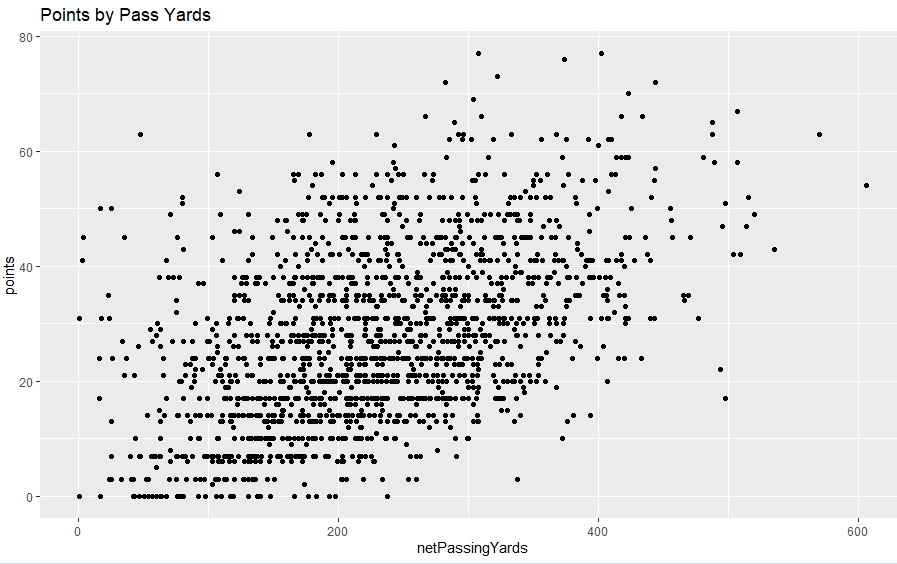
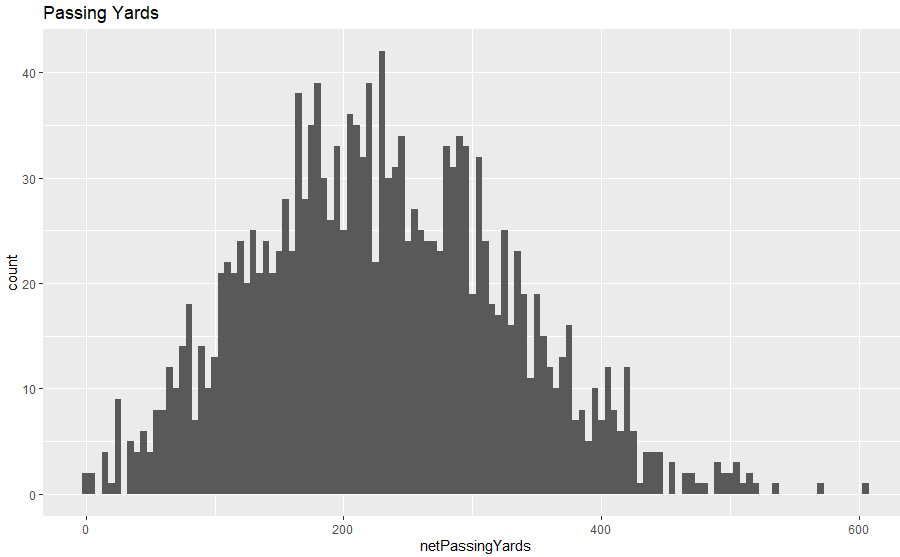
The analysis was done with Python and R. Python was used to gather and combine the data. R was used to conduct Exploratory Data Analysis (EDA), statistical analysis, and create regression models. The analysis began by pulling several data sets from the API and combining them. Initial EDA was done on the data to discover variables of interest. In this step, we attempted to find variables that had a strong correlation with points scored, which would help accurately predict scores. After completion of EDA, variables were selected to incorporate into the linear regression models. Once the variables were selected, multiple models were trained on data from weeks 1-10 of the 2019 season to predict the scores of both teams in each game.

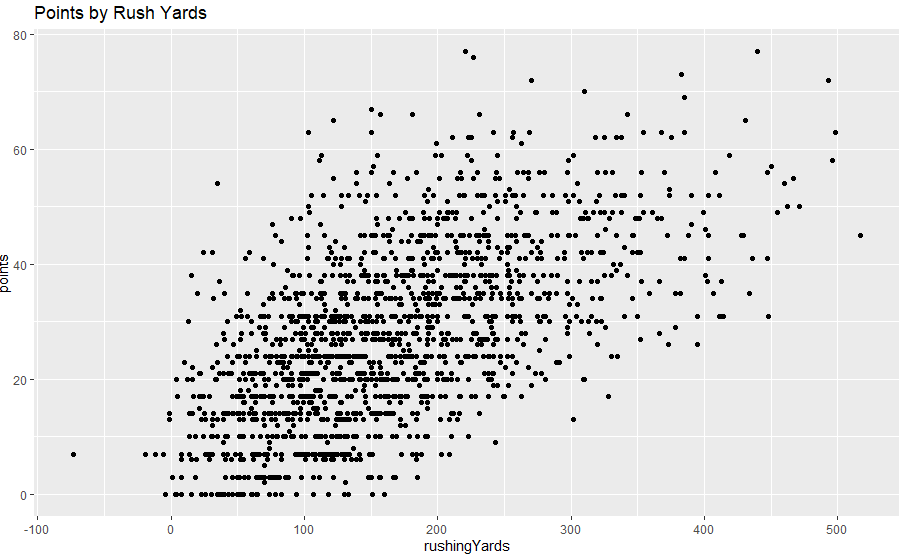
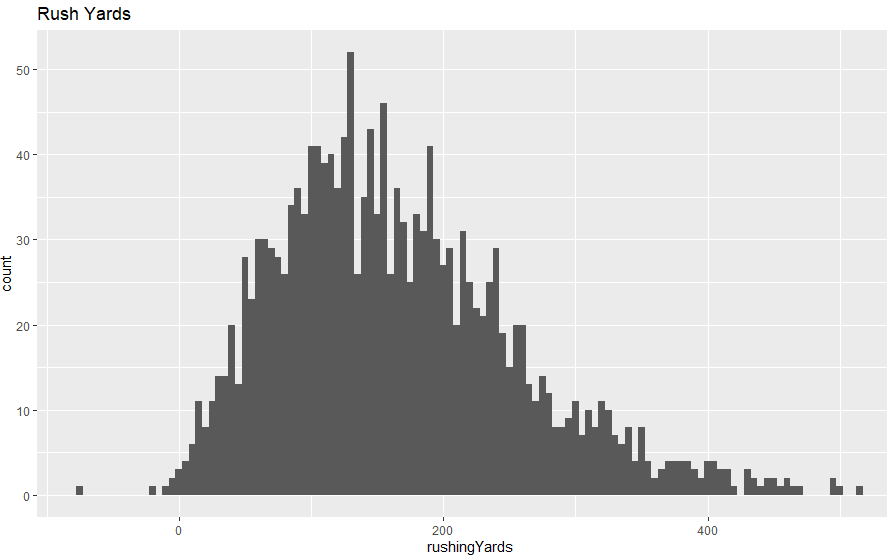
**Results**

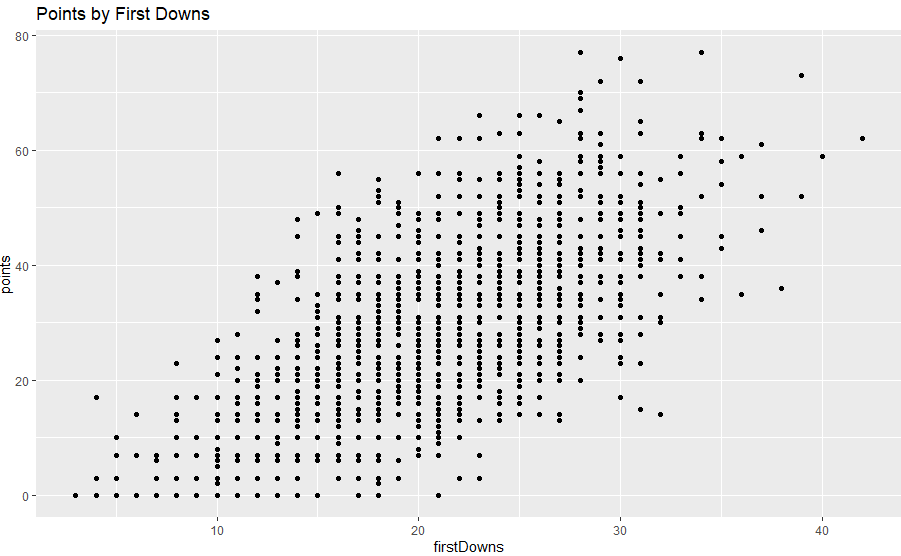
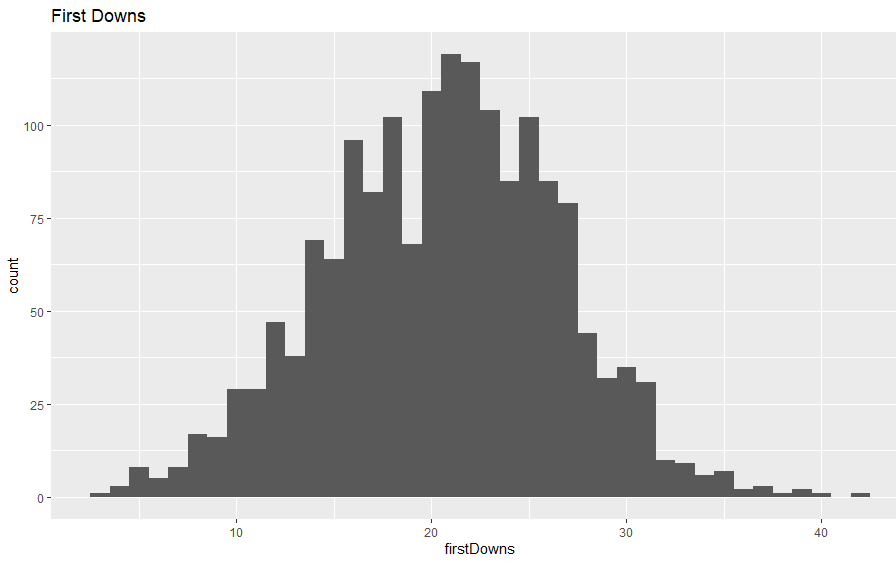
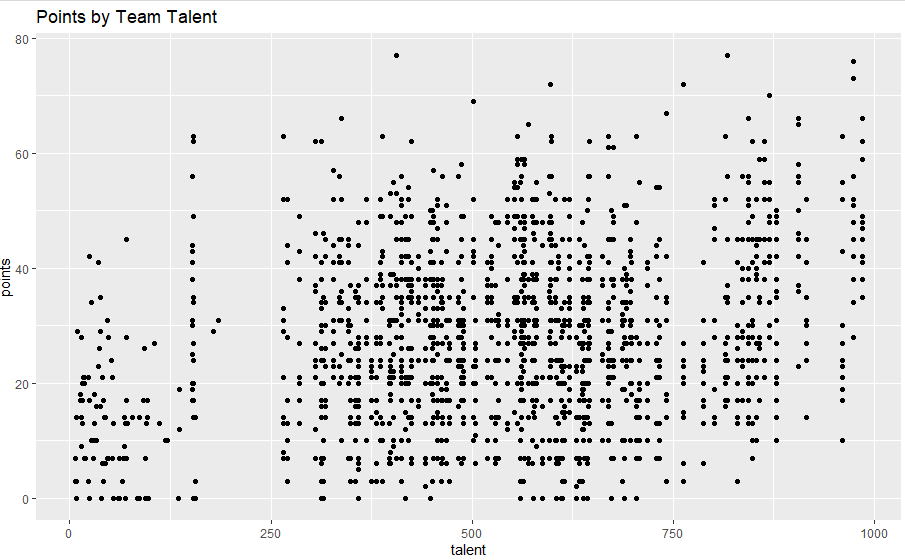
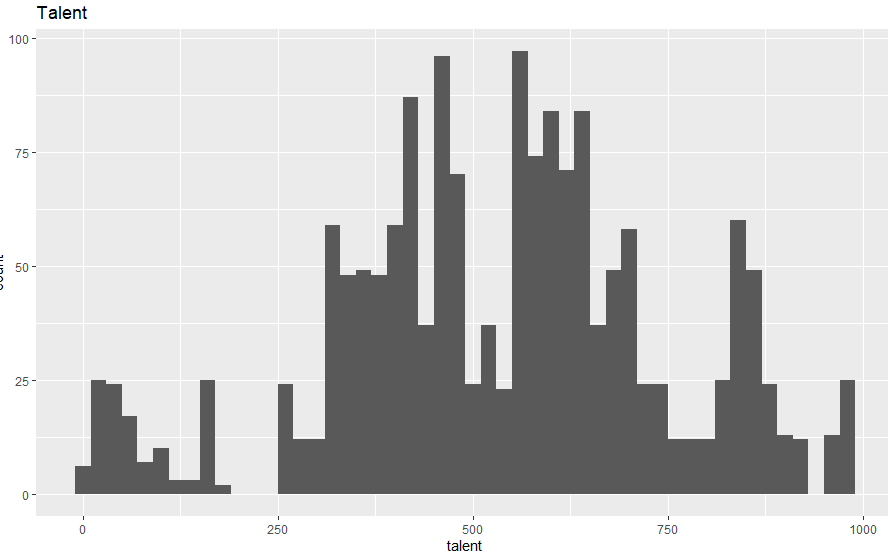
**EDA**

Several variables were considered for the models. Variables ultimately tested in the model include: Total Yards, Passing Yards, Rushing Yards, First Downs, Turnovers, Defensive TDs (touchdowns), Sacks, Kick Return Yards, Kick Return TDs, Team Talent, Home Field Advantage, and 4-week moving average of the same statistics.  
  
As seen above, our target variable, team points, has an approximately normal distribution.

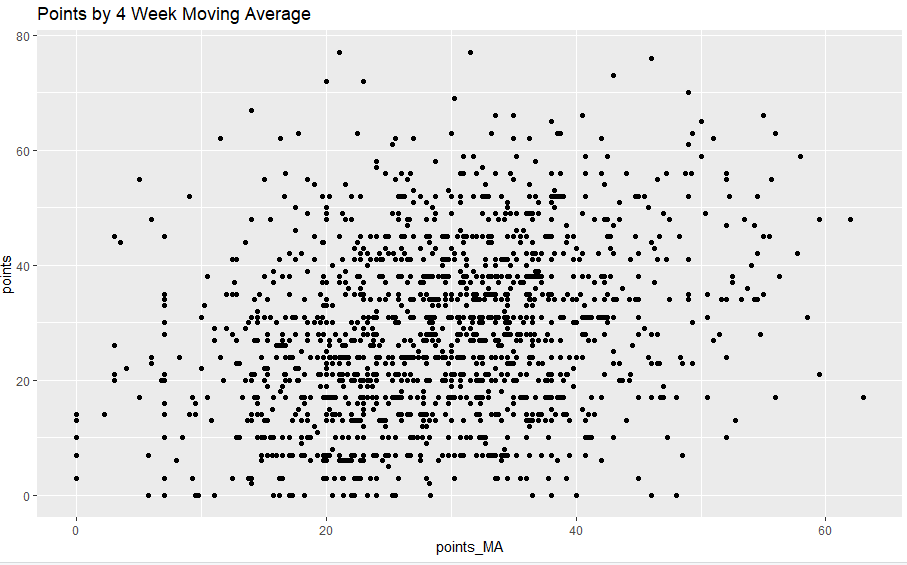
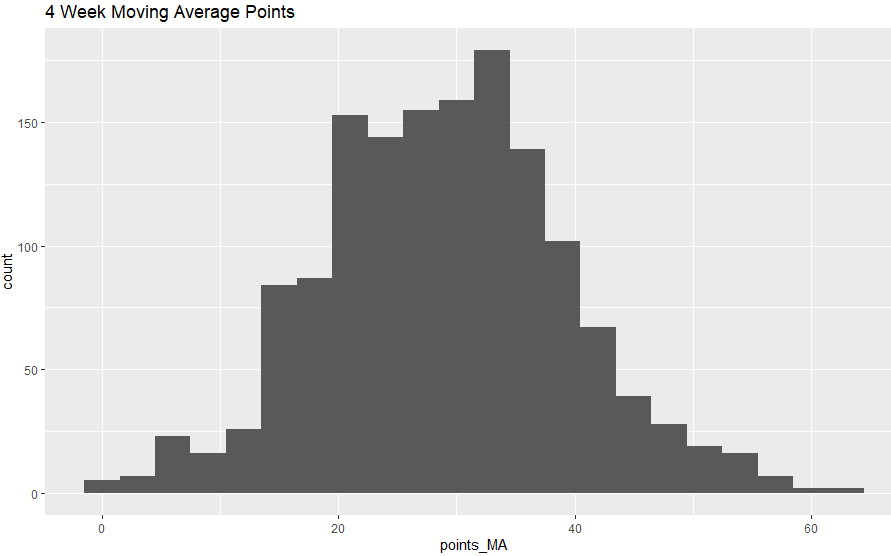
  
Total yards is approximately normally distributed and has a strong positive correlation with Points Scored.

  
Total passing yards is approximately normally distributed and appears have a positive correlation with Points Scored, although not very strong.

  
Total rush yards is approximately normally distributed, perhaps there is some skew, and appears to a positive correlation with Points Scored.

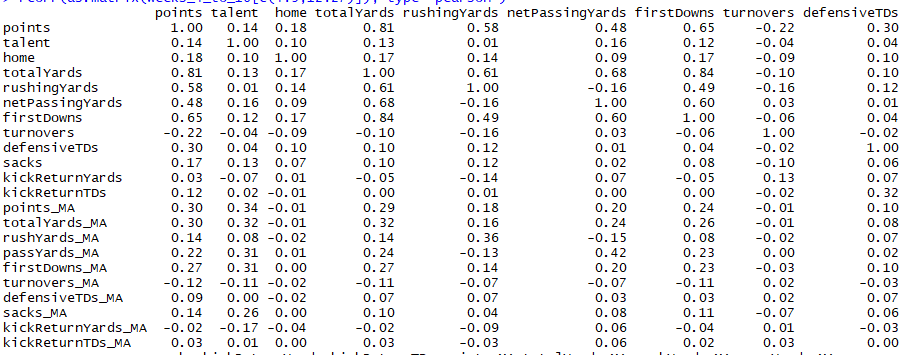
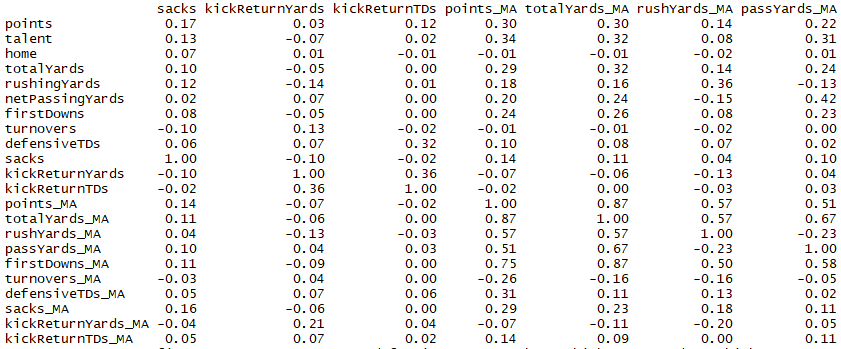
  
First downs is approximately normally distributed and has a positive correlation with Points Scored.  


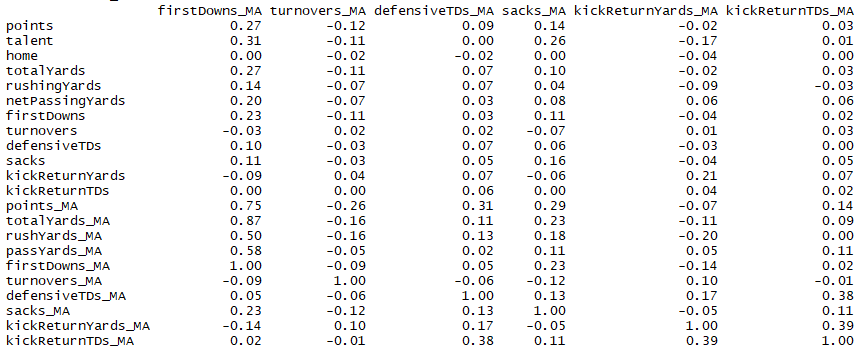
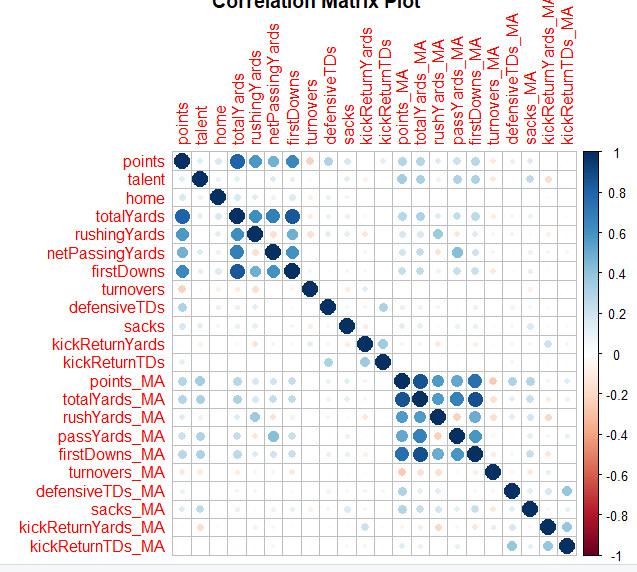
Team Talent has a general normal shape but appears to be bimodal. May possibly have a very slight correlation with Points Scored.

  
Team 4-week moving average of Points Scored is approximately normally distributed and may have a correlation with Points Scored.

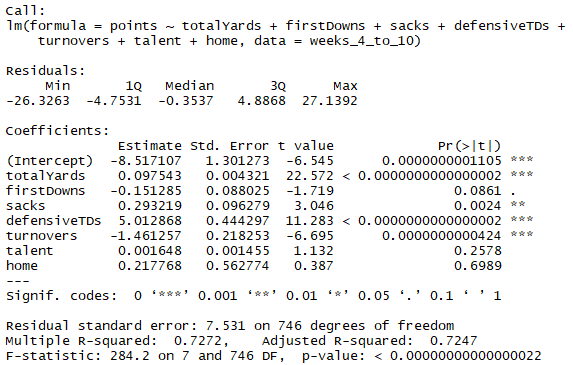
All variables above appear to at least be decent candidates for our first models.

**Correlation**

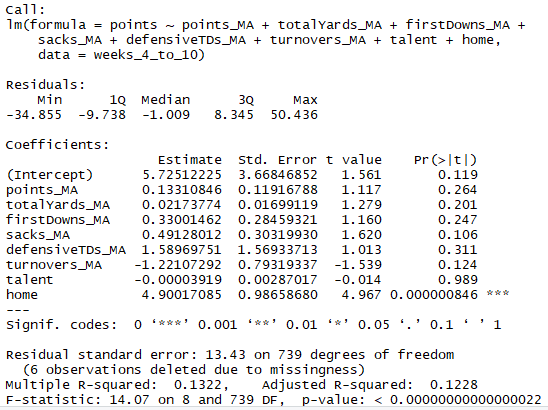
  


Most of our candidate variables have positive correlations with points scored, although most correlations are quite weak. A couple variables, turnovers and moving average for kick return yards have negative correlations. As seen during the EDA process, the correlation between Points Scored and Total Yards is quite strong.

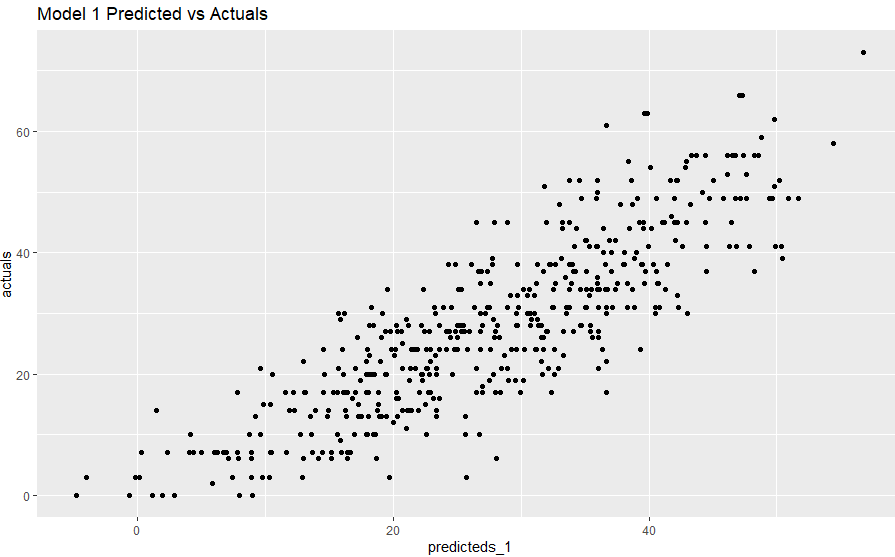
**Regression**  
Both linear regression models were trained on data from weeks 4-10 of the season. As seen below, the first model is statistically significant and can explain 72.7% of the variance in points scored for a team each week. Total Yards, Sacks, Defensive TDs, and Turnovers are statistically significant variables. We also have some counterintuitive results. For example, an increase first downs decreases the total points scored (although not statistically significant at 0.05).   


The second model was based on moving averages and is also statistically significant, but with a much lower R-squared value, only explaining 13% of the variance in points scored for a team each week. Home field advantage is the only significant variable.

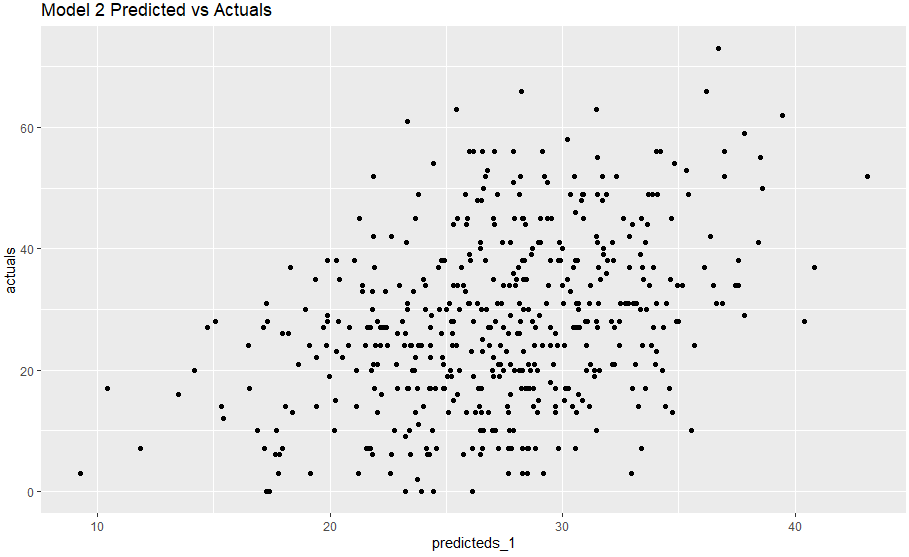


**Predictions**

While directional, the accuracy of the first model can certainly improve. The model seems to overestimate the number of points scored.



The second model is far less accurate than the first.



**Discussion and Conclusion**

Our first model has identified several attributes that contribute to total points scored. The moving average model was much worse than I would have thought. One major problem with the first model is data leak. The model is using data from the current game to predict the final score. This is not the goal of our model. Our model needs to take historic data to predict the score of the current game. This is where I had hoped the Moving Average model would be more successful as it used the game statistics from the previous four games. Additional variables need to be considered, along with the possibility of transforming the current variables to improve correlation and predictive power. The final output of the model needs to be used for both home and away teams and compared against the spread of each game to determine our ultimate success.

References

Licata, A. (2019, August 02). 42 states have or are moving towards legalizing sports betting – here are the states where sports betting is legal. Retrieved from [https://www.businessinsider.com/states-where-sports-betting-legal-usa-2019-7#nevada-8](https://www.businessinsider.com/states-where-sports-betting-legal-usa-2019-7" \l "nevada-8)

Liptak, A., & Draper, K. (2018, May 14). Supreme Court Ruling Favors Sports Betting. Retrieved from https://www.nytimes.com/2018/05/14/us/politics/supreme-court-sports-betting-new-jersey.html#:~:text=The law the decision overturned,college and professional basketball star.

Butler, R. (2021, February 04). Sports Betting Legalization Tracker: Cuomo Wants Lottery Model in New York. Retrieved March 27, 2021, from https://www.actionnetwork.com/news/legal-sports-betting-united-states-projections

Wasiolek, D. (2020, October 09). U.S. Sports Betting Worth a Wager for Investors. Retrieved from https://www.morningstar.com/articles/1003994/us-sports-betting-worth-a-wager-for-investors