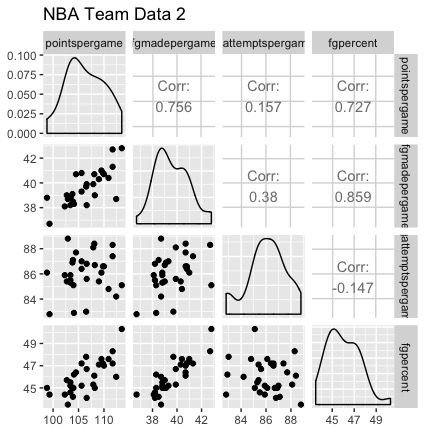
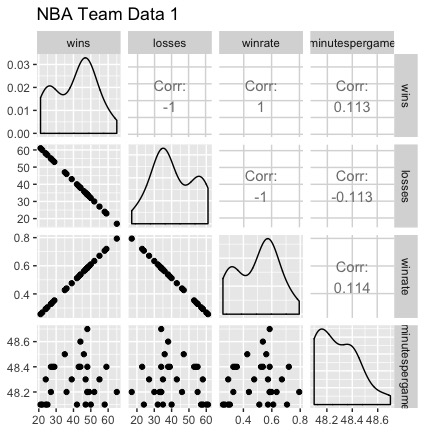
# Exploratory Analysis of NBA Betting and Vegas Accuracy

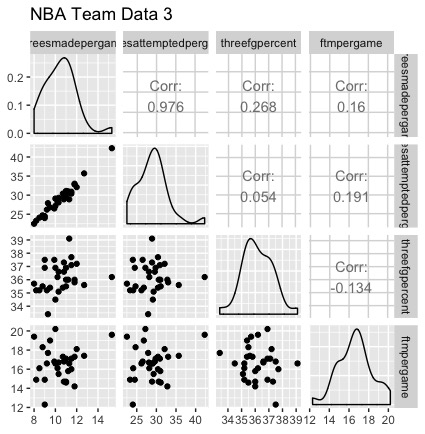
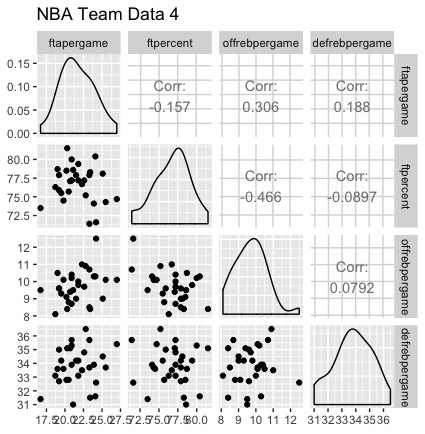
This project is an analysis of NBA team data, historical betting data, Vegas predictions and identifying how to win as a gambler using the information available to identify the best opportunities to make money. The initial plan of action in this project was to identify the information gap between my own predicted NBA lines and Vegas’ lines. The steps to this involved working with a dataset of historical predictions and the end results of NBA games from 2017-18, wrangling the data to get each game’s own information in one row, and lastly create my own model for prediction and test against Vegas to determine the gap. There has been some change made as the project has progressed and the primary focus in now an exploratory analysis into Vegas historical accuracy and where trends can be identified to take advantage of as a bettor.

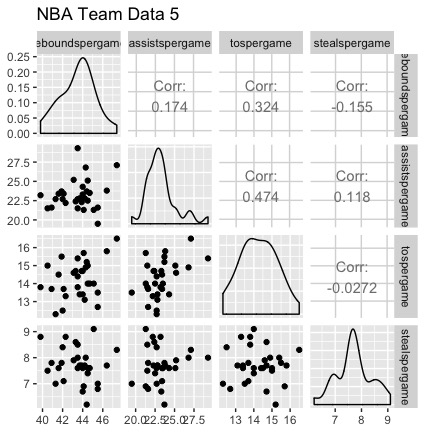
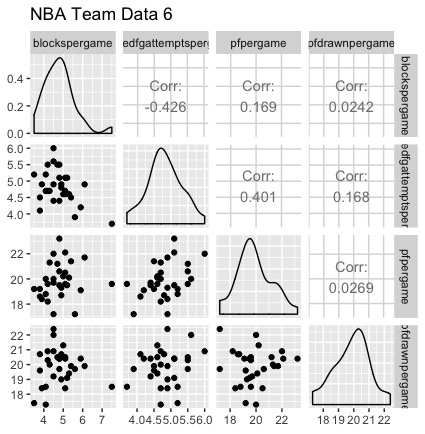
Exploring the data: I was able to find game data on the Vegas predictions and end result of those same games from the 2017-18 season. The dataset (all\_games) used, covers the beginning of the season on October 17, 2017 until March 16th and observes 1031 games. The limitation to this dataset is that it does not include all of the games that occurred for the entire season, as when this project began, the season was still going on. Some of the most important pieces of data used for this analysis included the teams, dates, predicted lines from Vegas for both spread and over/under, and finally the end results of the games. Using this information, I was able to manipulate the data to create more columns for easier analysis including the difference in the predicted values and actual results for both the spread and over/under.

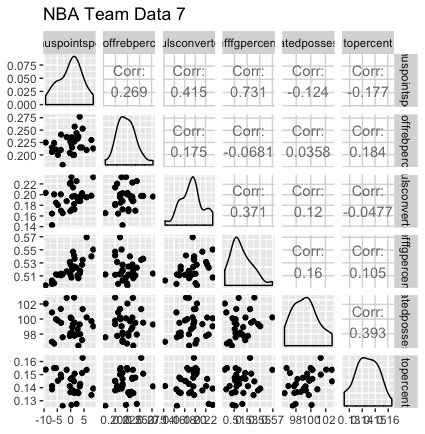
The second portion of data that I used in this analysis was from the past season NBA teams data from the 2017-18, focused around their averages in a variety of stat lines. The most important data piece in this dataset was the average points margin per game. This can be used as a good foundation for estimating ending margin between two teams.

I decided to look externally and research further for what other data was available that could be useful in prediction and came across an NBA theory based upon only a certain set of stats for each team. Using this information, I calculated these stats for each team for use in prediction.





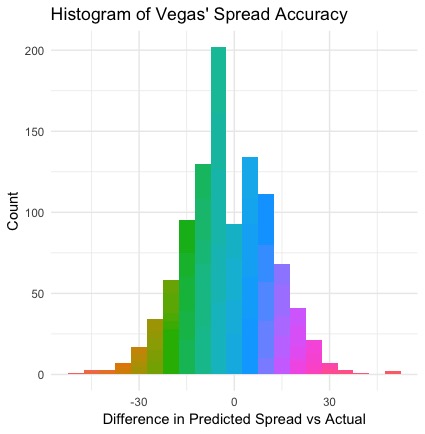


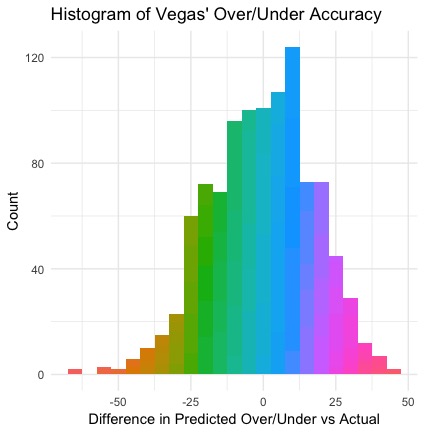


Professional gamblers and the general public alike can utilize the information throughout this analysis for guidance in the upcoming NBA season. The entire goal of a gambler is to identify spots where the line that Vegas predicted for a game does not make sense in the current situation. There are a variety of ways that this can be accomplished and the exploration intends to expose as many of those possibilities for success as a gambler.

There were a variety of questions that I wanted to answer throughout this analysis. The most obvious of the group, just how accurate is Vegas in comparison with what actually happened in each game? Can I identify the information gap between what the data says the lines should be and what Vegas predicted? Were there certain time frames that Vegas was more accurate than others, or did they become more accurate as the season went on? Given that there are numerous times that Vegas is not even close on a prediction, is it possible to identify those situations before they occur to take advantage of them? Is it more advantageous for the bettor to focus more on spread lines or over/under lines? These are some of the questions that I wanted answers to, as I know that this information can be used to make the best decisions possible moving forward to make money.

As with almost anything, there were certain questions that I couldn’t answer or complete with the data I had available. Most notably, can I reverse engineer Vegas’ model? Obviously getting as close as possible is any gamblers dream, but what information besides just historical data and stats can be used to determine their lines? This opens doors to many other ideas outside of this analysis such as real time information and many times information that isn’t available to the general public.

Initial findings show just how accurate Vegas can be, on the surface. Diving into the actual results from the games in this dataset, the average predicted spread for the home teams by Vegas was -2.59. What this means is that the home team will win by 2.59 points on average across all games in this dataset. The actual results show that the average ending spread of these games was -2.12 for the home team. This means that Vegas, on average, is within just .47 points per game in regards to spread.



Additionally, the average predicted over/under by Vegas for the observed games was 212.19, just .39 points lower than the actual ending over/under total for all games, 212.58.

Yes, you’re reading this correctly, on average, Vegas is within .5 point on both the home team spread and over/under for each observed game.

There is positive news to this as well! Averages can be misleading. Taking a closer look at the data, there are 420 occurrences where Vegas’ home spread prediction was incorrect by 10 points or more, or 40.74% of all observed situations. Diving into the over/under, there were 429 occurrences that Vegas’ predicted line was incorrect by 15 points or more, or 41.61% of the time. Given that Vegas can be painfully incorrect, at such an alarming rate, is it possible to identify when those situations are going to occur using just the data available? This gives gamblers hope, that maybe Vegas just isn’t as accurate as they are led on to be, and some money can be made.

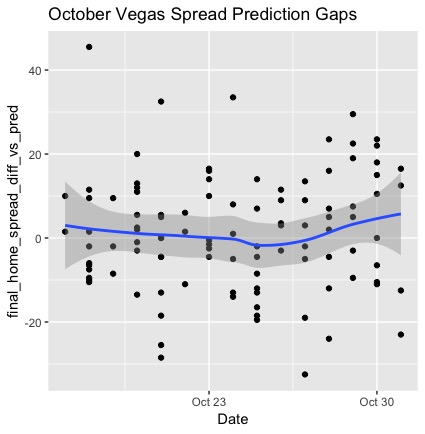
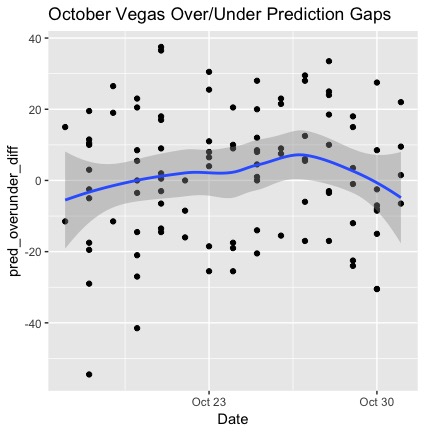
The next step after the high level basics was to build a model that would predict what the plus/minus points per game of each team would be, based upon their stats. This would be used as a guideline to predict the spread of a game between any two teams. The process began with building a simple linear model based upon the stats of each team that were already available in the dataset (nbateams). To determine the accuracy of the models, I used the multiple R-squared value and adjusted R-squared values. R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression. The key limitation: R-squared *cannot* determine whether the coefficient estimates and predictions are biased, which is why you must assess the residual plots. In general, the higher the R-squared, the better the model fits your data.

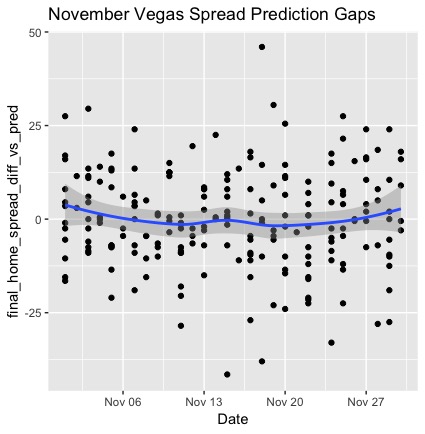
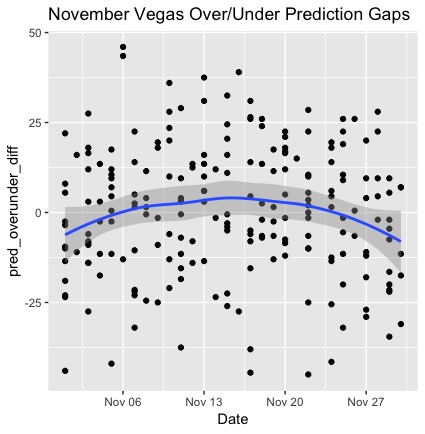
The first linear model for predicting plus/minus points per game per team resulted in multiple R-squared value of .8932 and an adjusted R-squared value of .8064. This biggest issue with this model was that there was a lot of noise in the data as just 5 statistical categories showed significance of the 13. After reducing the variables impacting the plus/minus points per game per team to the 5 variables that showed significance, that model resulted in a multiple R-squared value of .7179 and adjusted R-squared value of .6591.

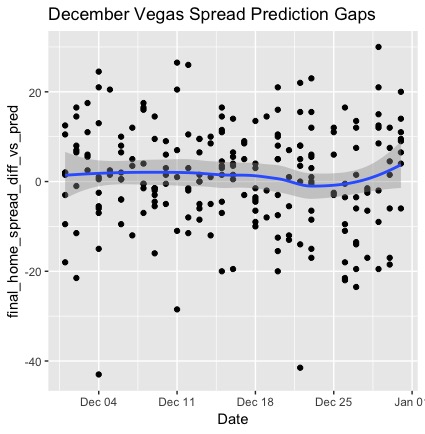
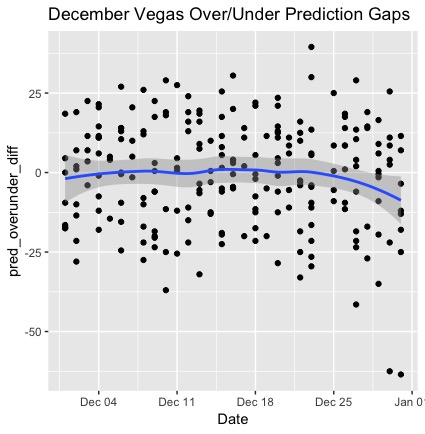
Realizing that this is the wrong direction in terms of accuracy, I dove into external research that has already been done about predicting spreads. The information I found helped me to develop 5 new variables based upon manipulating current stats. The variables I created were focused around offensive rebounding efficiency, converting on fouls drawn, true efficiency field goal percent, estimated possessions per game, and lastly turnover percentage.

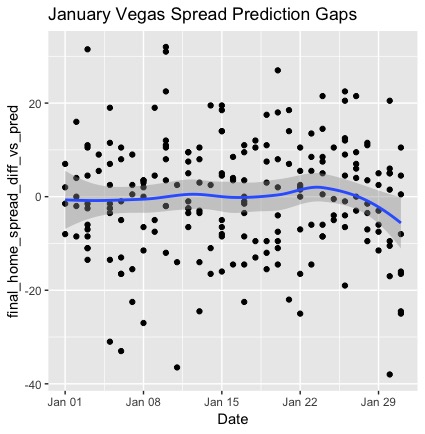
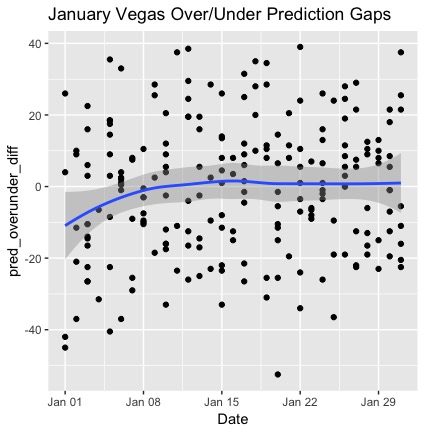
Hoping to increase my accuracy, I incorporated those variables into a new model that resulted in a multiple R-squared value of .9756 and adjusted R-squared value of .9356. Great accuracy on the surface, but there is also a lot of noise contributing to this model. The final model created was centered on just my manually created variables. The resulting values of multiple R-squared and adjusted R-squared were .7146 and .6689, respectively.

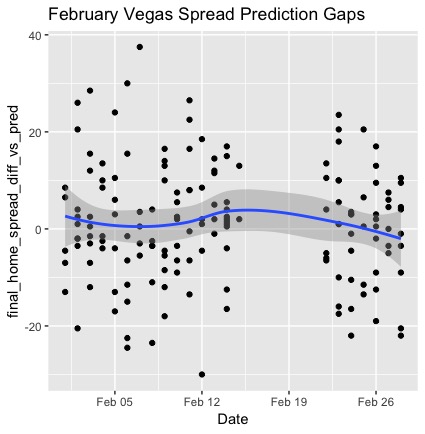
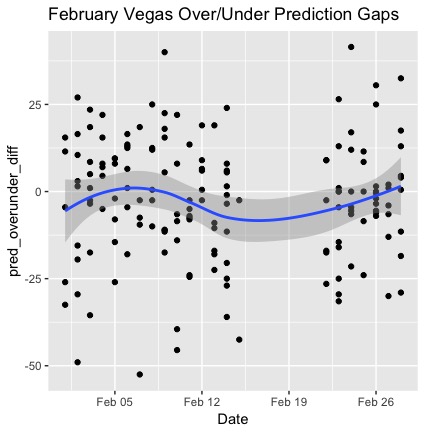
A further analysis of the Vegas’ accuracy over time in terms of both spread and over/under could help with noticing a time trend. A complete breakdown month over month by spread and over/under will provide gamblers with an idea as to the best months to increase unit size of bets or to look for better opportunities. A suggestion to gamblers using this data, it would appear that early in the season Vegas is among their least accurate months as they perhaps are trying to get a better understanding of teams’ performance and consistency. Another great opportunity for gamblers is in the month of February. A potential cause of decreased accuracy would be that the trade deadline is in early February. The deadline usually leads to a change in teams’ rosters, and Vegas appears to be their least accurate trying to adjust for these changes.

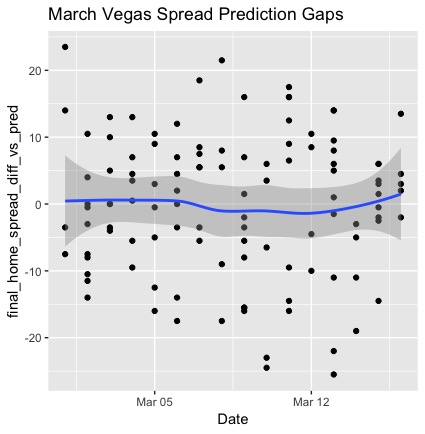
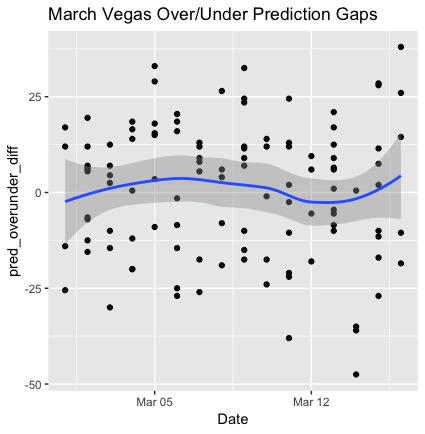




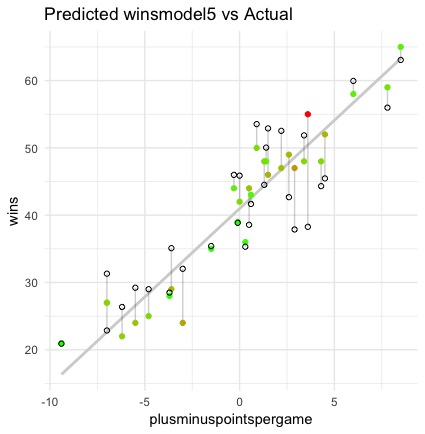




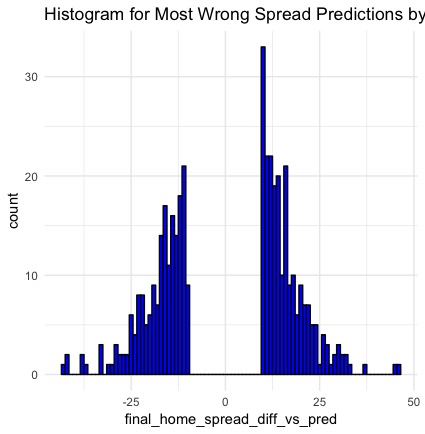


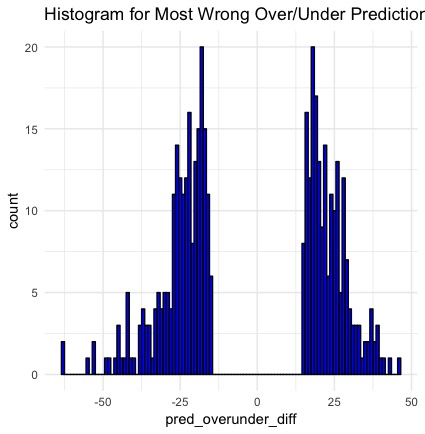


The next steps I took revolved around creating different models for prediction that would increase accuracy. I began with creating decision tree models. Through multiple variations of trees, there was no statistical increase in accuracy. Moving forward, the next learning model I aimed to create was a random forest model. This is built off of the same idea for decision tree models, but using a learning set and a test set to know which variables are most impactful, the model creates itself. After going through multiple models, the accuracy of the prediction model was still not as accurate as using the linear models. Curious, I did some external research in regards to prediction models for sports betting on basketball, more specifically NBA. The most interesting thing I had come across was that of the many different predictive models, the linear model was the most consistent for NBA prediction.

Discovering this new information, I decided to take on linear models as my primary tool for prediction. Aside from just looking at spread, I aimed to predict the wins total for each NBA team. The reason behind this is because before the NBA season begins every year, gamblers can bet on the over/under for total wins for each team. This wasn’t an initial direction I intended for this analysis, but this offers additional opportunity for bettors to bet in a different direction. The accuracy of the model for this past year’s results is as follows:

The accuracy of the linear predictive model for wins resulted in a multiple R-squared value of .8041 and an adjusted R-squared value of .7418. Given the input of certain stats of a team, or predicted stats, a gambler could use this model to predict the total wins of an NBA team next season.



A final spot I wanted to dive into were what I believe to be the best opportunities as a gambler, the times where Vegas was most wrong. Beginning with an exploration into the 420 times Vegas was wrong by 10 points or more on home spread prediction, 159 of the times the home spread was -5 or more. Taking a look at the teams with the best four records and the teams with the worst four records; those teams accounted for 117 of the 420 occurrences. A month over month breakdown of the spread biggest fails in Vegas prediction is as follows: October = 51, November = 91, December = 86, January = 85, February = 66, and March = 41.

The last place to look into are the times where Vegas was wrong by 15 points or more on over/under predictions; this happened 429 times. The first thing to note is that on average, the over/under prediction in the league is 212.1867 and the over/under prediction of these 429 occurrences is 211.9231. A month over month breakdown of the over/under biggest fails in Vegas prediction is as follows: October = 49, November = 87, December = 88, January = 97, February = 61, and March = 47.

While looking at the lower October and March numbers, please keep in mind that October data started on October 17th, 2017 and March data goes through March 16th, 2018.

Due to these opportunities not providing any noticeable trend to take advantage of as a gambler, there’s a final piece to dive into: the home underdog. There were 343 instances where the home team was predicted by Vegas to be an underdog, or a spread of greater than 0. Of these 343 instances, the average absolute value difference in spread prediction from the predicted Vegas value was 9.973761 points. An average of being wrong by almost 10 points on spread for home underdogs gives bettors a great opportunity to take make great money. The average spread predicted for the home team in these home underdog situations results in 4.962099.

After an in depth analysis, there are three takeaways that professionals and amateurs alike should use to make future bets and get an advantage on Vegas. The first is looking to take advantage of the learning curve Vegas goes through early in the season and right after the trade deadline. These two months provide the best opportunity for success, as Vegas’ average margin of error is greatest during these times. The second takeaway is to use the predictive total wins per team model for the upcoming 2019 season. This is an opportunity prior to the season to make accurate decisions on whether or not teams will win X number of games. Noticing the largest gaps in predicted using my model vs Vegas’ prediction will provide the best chance of success. The third and final takeaway is to focus on home team underdog situations, as Vegas tend to miss the mark the most in these instances.