NBA Capstone Project
Jeffrey Ebert

Evaluating Game Trends in the NBA

The purpose of this project is to evaluate what are the most important factors in determining if an NBA team will win a specific game, and answering questions about underlying trends that take place throughout a season. The intended audience of a report like this is not necessarily people within an NBA front office looking for analytical insight that cannot be seen just by watching the games take place on the court, although it could be. Its focus is geared more for the average fan who is looking to understand the nuances and breakdowns of important determinants of how a game is won, besides the biggest guys on the court being able to shoot, dunk, dribble and pass the best. Specifically, it tailors to the NBA fan who is not adverse to risk, and is interested in understanding what should be focused on and known for gambling purposes. This is observed through the lens of predicting individual games based on statistics and recent trends, as well as observing season to season data that could be beneficial for long term betting endeavors.

With many states having recently passed bills to legalize sports betting, and more reviewing the prospects of doing so, gambling on sports is becoming more popular and mainstream that it ever has before. In states such as New Jersey, millions of tax dollars are already being generated from legalized gambling. Major sports media outlets are starting to regularly provide information about betting, such as the lines for games, in addition to the usual information about a game. This is because there is such high demand for this information. Betting on sports is nothing new, with illegal and underground gambling circuits having existed about as long as professional sports leagues have. But with increased regularization and its expanded legality the stigma is continually lessening and there are many companies and individuals looking to make a business off selling valuable sports-related information to gamblers that they can use to their advantage in making some money.

The reason that states can create so much tax revenue from legalized gambling, and places such as Las Vegas have made so much money off of sports betting is simply because it is difficult and the odds are stacked against you. But that does not stop a large number of people from trying to get rich off of sports betting or people who are just trying to make a few bucks and have some fun at the same time. This growing demographic has paved the way for the need of information based on sports gambling and statistical trends. While this project solely focuses on games within the National Basketball Association, these types of analyses could be

applied to many of the most common professional sports leagues that the gambling public wagers on throughout the year.

This project will take an exploratory approach into evaluating the most important trends and factors in determining what makes for a winner in the NBA. it evaluates data taken from the past four complete NBA seasons taken from 2014-2018, as well as the statistics of those games and other associated data for those games. All of the scores used for the data were taken using a web scraper from espn.com. It includes all regular season games from the 2014-2015, 2015-2016, 2016-2017, and 2017-2018 NBA seasons. The data was cleaned to ensure the accuracy of all scores, teams and dates of the games. The other data uploaded for the project came from separate csv files that include statistics and the points spreads of all of these games.

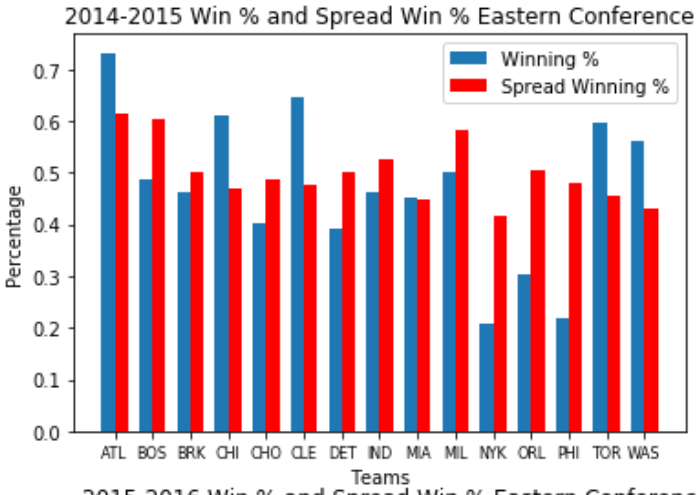
The following analysis is broken down into sections on data storytelling and an exploratory analysis on inferential statistics and machine learning. The data story investigates seasonal trends within the NBA and tries to discover underlying features that arise during an NBA season and what role they can have in determining the success or failures of teams in certain situations. The next section on inferential statistics and machine learning focuses more on the statistics of teams throughout the season, and evaluates which are the most significant regarding a team winning a game. Many different variations of models were created and adapted in order to try and find a combination of values that gives the highest predictive power towards any NBA game played.

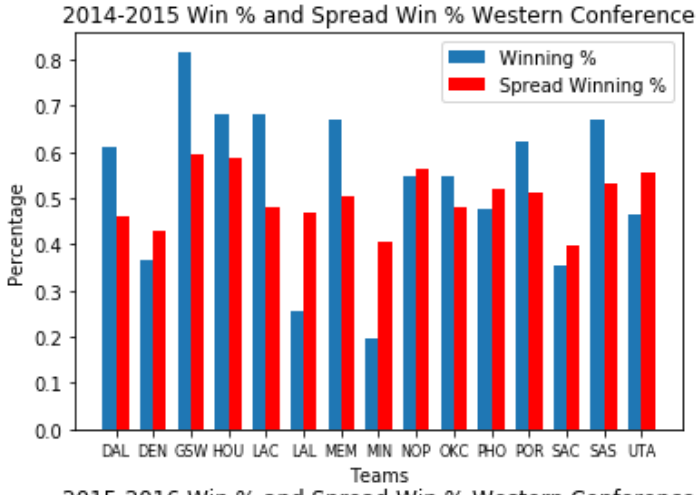
Data Storytelling

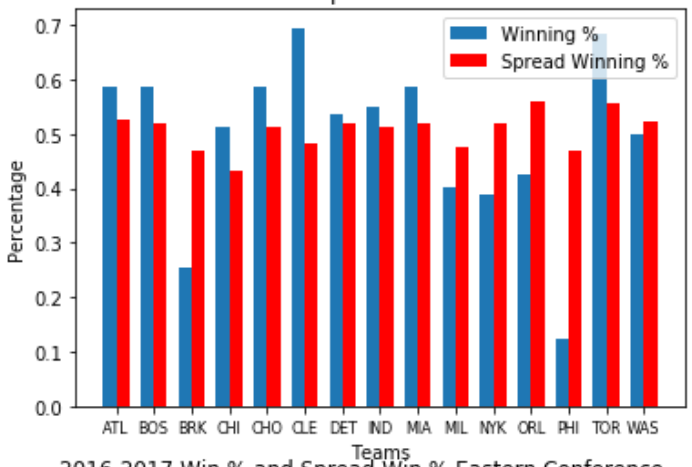
The objective of the analysis in this project is to evaluate various yearly trends for teams playing in the NBA, finding patterns and factors that could be important determinants if a team wins a certain game. Most of the trends focus on one regular season's length, and relate to how often and when a team wins a game compared to how the spreads and indicated odds predict that team to win or lose. There are many variables that can be examined to observe these seasonal patterns, and multitudes of ways to discover insight on all of them, such as by breaking them down more by team, opponent, point in the season and more. What information about how a team has performed, along with who they are playing and the context of the game can yield strong predictive value in evaluating the outcome of a game? It will be important knowing which factors are important and which ones are not as significant as we may believe.

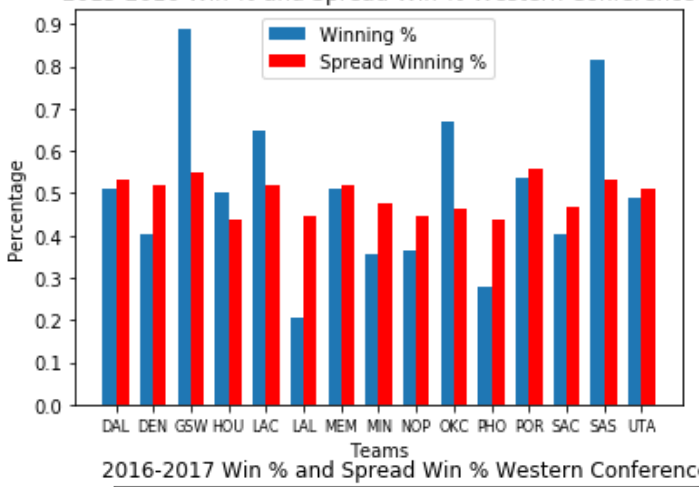
For fully understanding the analysis below, it is crucial to know the terminology. A points spread or line in sports betting is a figure that is set by oddsmakers (such as Las Vegas) in order to provide an associated mark for what they estimate a team will win or lose by. This is to provide betting interest for both teams in a game, as there is often a team that is deemed clearly better than the other. The favorite's spread is designated with a "-" sign and the number that follows indicates the amount of points they would need to win by in order to cover the spread. An underdog's spread is marked with a "+" sign the number after indicates the amount of points a team can lose by and still cover their spread. There are many factors that go into the line for each game, and it is the role of the oddsmakers to create a fair line that generates betting interest for both sides, while the bettor is trying to pick the team that will perform better on their spread when wagering against the spread.

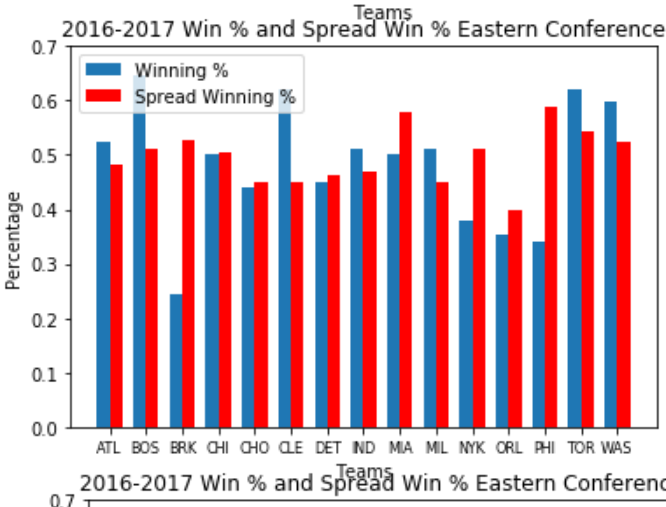
Many people think that the best teams are the best at covering the spread, but this may not always be true. It is significant to know the true relationship between being a winning team and performing better against the spread. Below are graphs by conference for each season of a team's winning percentage compared to their winning percentage at covering their spread:

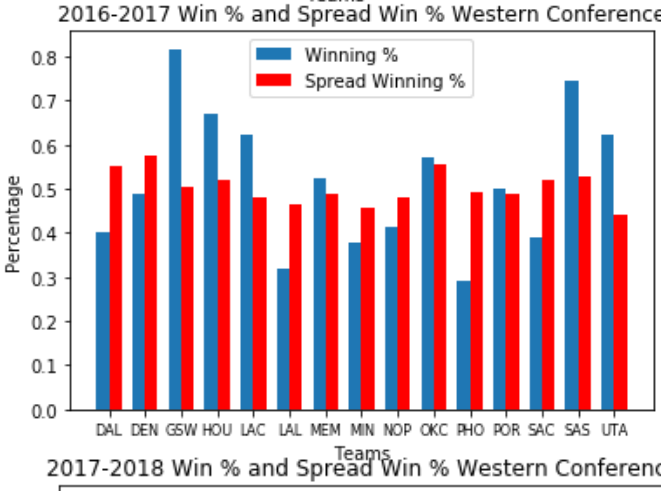

2014-2015 Win % and Spread Win % Eastern Conference
Winning %
Spread Winning %
Percentage
Teams
ATL BOS BRK CHI CHO CLE DET IND MIA MIL NYK ORL PHI TOR WAS

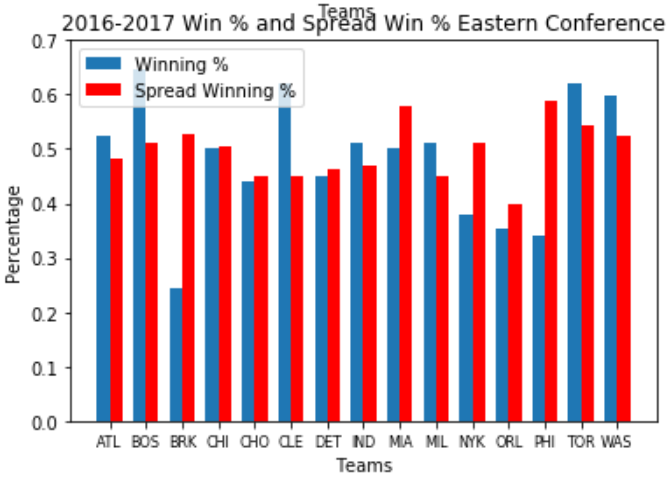

2014-2015 Win % and Spread Win % Western Conference
Winning %
Spread Winning %
Percentage
Teams
DAL DEN GSW HOU LAC LAL MEM MIN NOP OKC PHO POR SAC SAS UTA

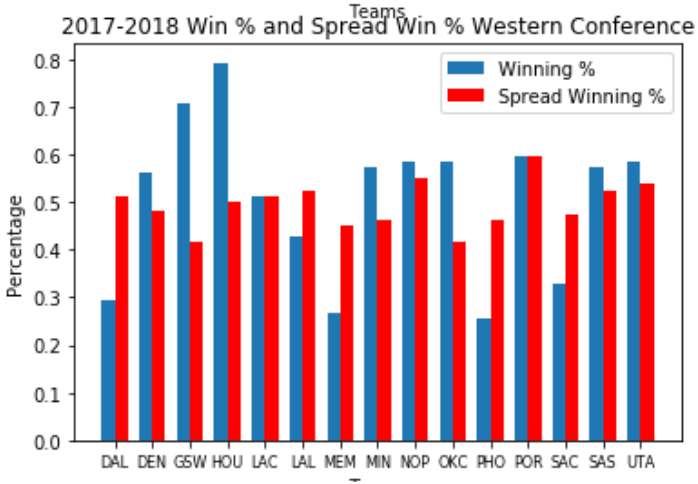

2015-2016 Win % and Spread Win % Eastern Conference
Winning %
Spread Winning %
Percentage
Teams
ATL BOS BRK CHI CHO CLE DET IND MIA MIL NYK ORL PHI TOR WAS


2015-2016 Win % and Spread Win % Western Conference
Winning %
Spread Winning %
Percentage
Teams
DAL DEN GSW HOU LAC LAL MEM MIN NOP OKC PHO POR SAC SAS UTA


2016-2017 Win % and Spread Win % Eastern Conference
Winning %
Spread Winning %
Percentage
Teams
ATL BOS BRK CHI CHO CLE DET IND MIA MIL NYK ORL PHI TOR WAS


2016-2017 Win % and Spread Win % Western Conference
Winning %
Spread Winning %
Percentage
Teams
DAL DEN GSW HOU LAC LAL MEM MIN NOP OKC PHO POR SAC SAS UTA

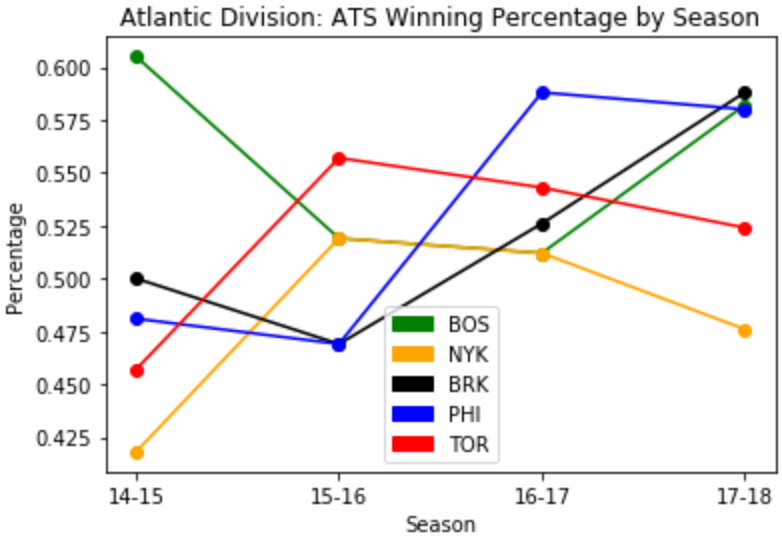

2016-2017 Win % and Spread Win % Eastern Conference
Winning %
Spread Winning %
Percentage
Teams
ATL BOS BRK CHI CHO CLE DET IND MIA MIL NYK ORL PHI TOR WAS

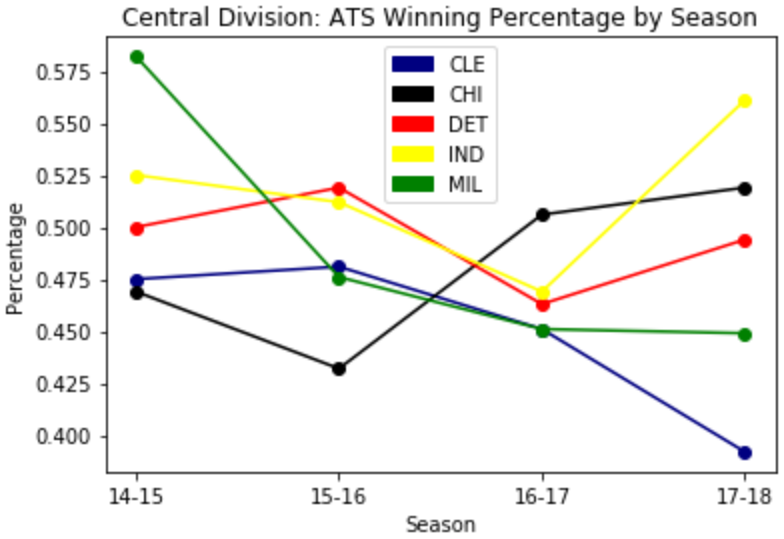

2017-2018 Win % and Spread Win % Western Conference
Winning %
Spread Winning %
Percentage
DAL DEN GSW HOU LAC LAL MEM MIN NOP OKC PHO POR SAC SAS UTA

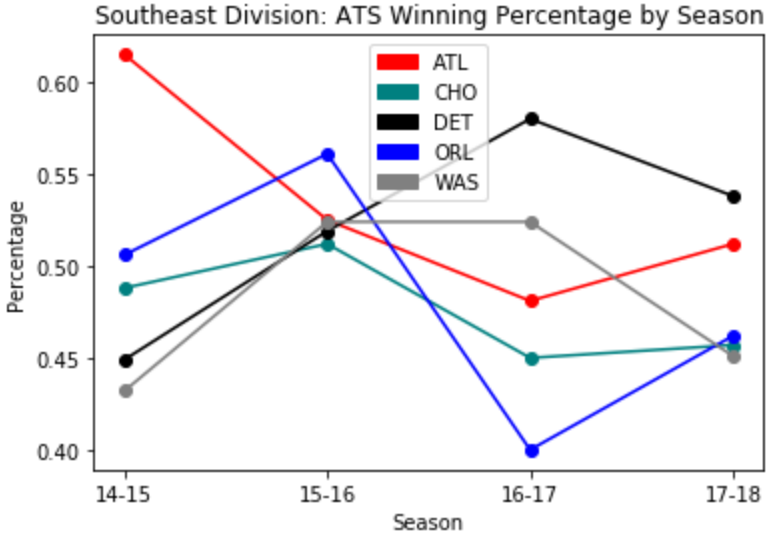
These eight graphs were separated by season in order to view how a team's winning percentage and their winning percentage against the spread (ATS) are related to each other each year, and if there are any patterns to these relationships. The graphs were separated by conference in order to improve readability in the graphs and not over clutter them. Splitting up the teams by conference is a natural decision since it is a sensible way to group the teams together and most games throughout an NBA season are played by teams within the same conference.

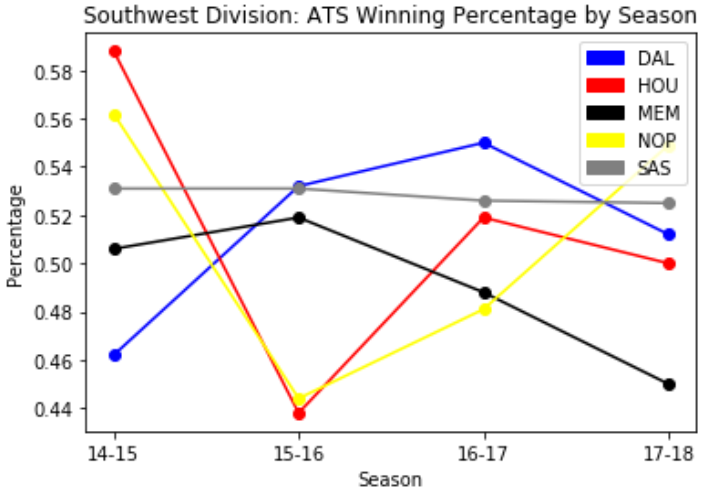
It is clear that yearly winning percentages have more variability than their percentages ATS because oddsmakers are trying to curb the spreads to make them as even as possible, which is why spread percentages hover closer to 50%. But, what is the relationship between a team's winning percentage and their ATS percentage? Based on the graphs, it appears most team's percentages are close to one another, and an extreme winning percentage can yield an extreme ATS percentage. For example in the '14-'15 season, ATL and GSW had the highest winning percentages and highest ATS percentages in their conference. However each season has a few outliers. In the '14-'15 season, PHI had the second worst record in the Eastern Conference, but there ATS % was still close to 50% and in the middle of the pack that year. In the '17-'18 season, GSW, HOU, and TOR had the best records in the league, but had huge disparities in their ATS percentages, which were average. This shows that just because a team is one of the best or worst in the league doesn't mean they will necessarily be the most or least profitable ATS throughout the season. There are many possible reasons why this could be, and it would be best to pay close attention to a team throughout a season to fully understand why. Common reasons could include high winning teams winning in close contests, and not blowing out their opponents like could be expected by the best teams, thus not covering the spread. This applies in the opposite for more frequent losing teams where they are losing often in close games and covering the point spread. There is an old sports betting adage that goes "Good teams win, great teams cover."

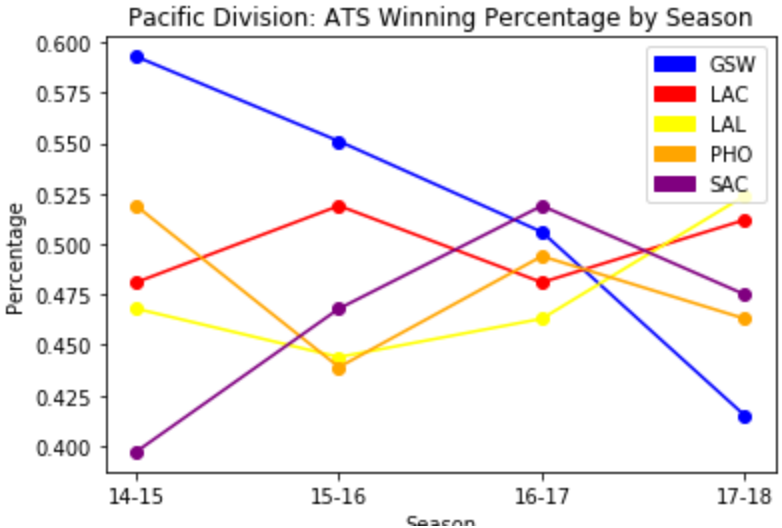
We can look how these season ATS percentages vary from year to year, and see if certain teams demonstrate patterns from one season to the next:

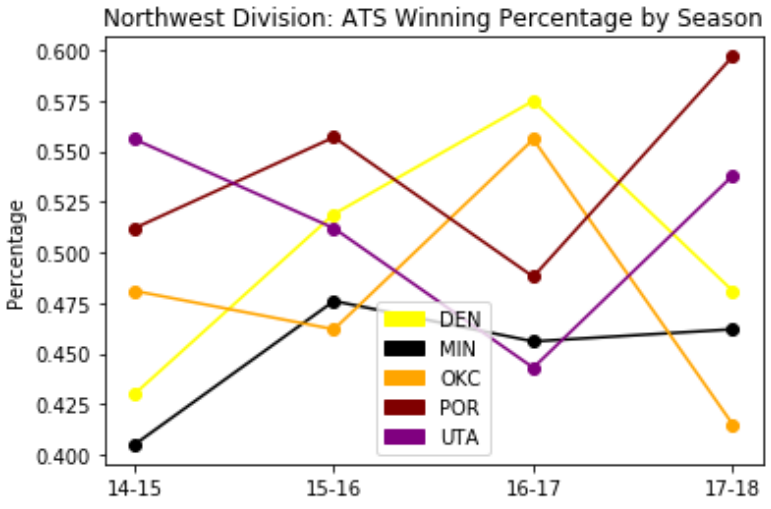








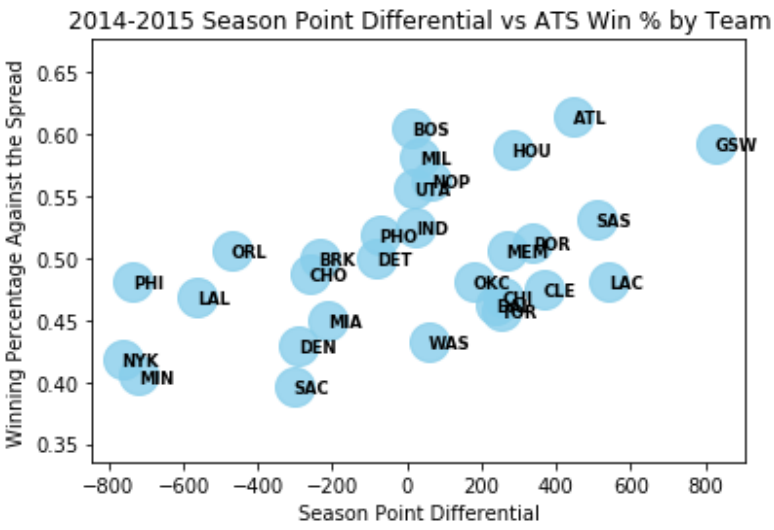


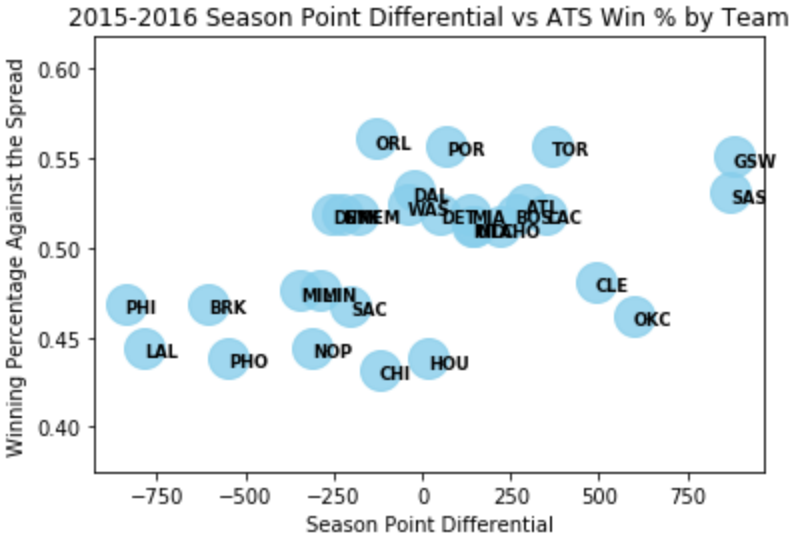


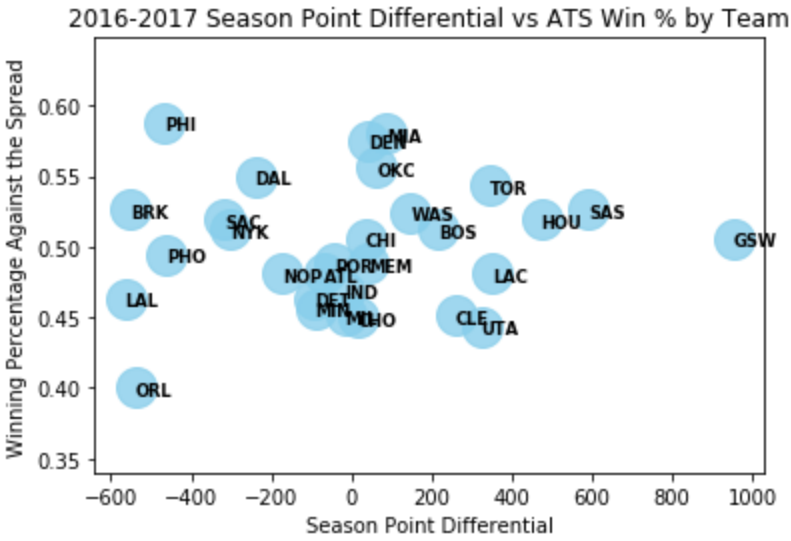
These six graphs represent a team's winning percentage against the spread (ATS) by season, broken down by division in order to maintain readability. The purpose of these is to detect any possible patterns between team's ATS winning percentage by year, and see if there is any relationship. From these graphs, it does not look like a previous season's ATS percentage is indicative of the next season's. There are very few instances a team had the highest ATS percentage in their division for consecutive seasons. Interestingly there are two teams that had ATS percentages of over .500 all four seasons. Those teams are the Spurs and Celtics, two of the most consistently winning teams over the last few years. There are also just two teams that have had an ATS percentage of under .500 for each season. Those teams are the Cavaliers and the Timberwolves. However, there are instances where the lines on the graphs are relatively even, meaning a consistent performance in season to season ATS performance. Previously mentioned as one of the only teams to finish about .500 every year, the Spurs ATS has barely deviated, hovering around .53 each season. The Lakers had a three year stretch with a consistent mark from .44 to .47, where they were one of the most frequent losers in basketball. The Timberwolves had a similar losing stretch to the Lakers, and the Raptors had a 3 year consistent stretch as well where they were among the top teams in the Eastern Conference each year, and had ATS percentages ranging from .524 to .557.

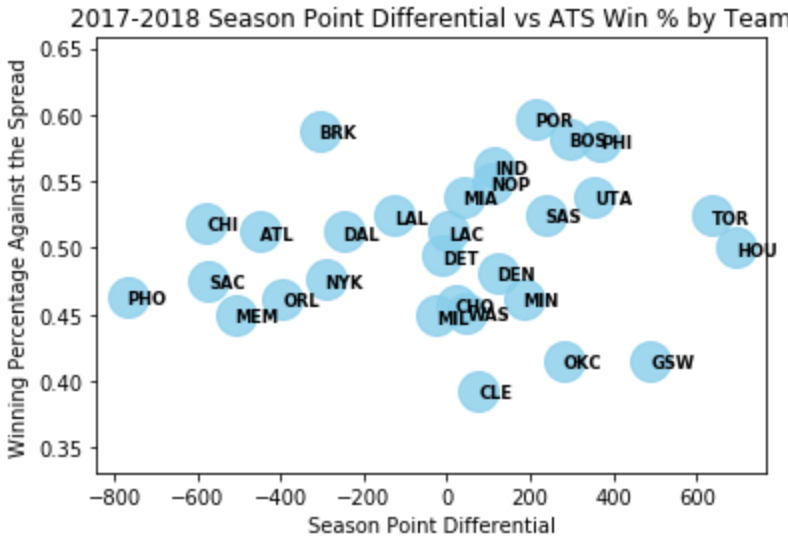
Even if there are some longer term trends within the data here, it is hard to say that ATS performance in one season has a large effect on the next, albeit some exceptions depending on the team.

An important measure for a team throughout the season is its point differential. If a team has a high point differential they are often winning their games by a lot of points and losing in close games. If a team has a low differential they are likely getting blown out often and not winning their games by as many points on average. These next four plots represent a team's season points differential vs their winning percentage against the spread:





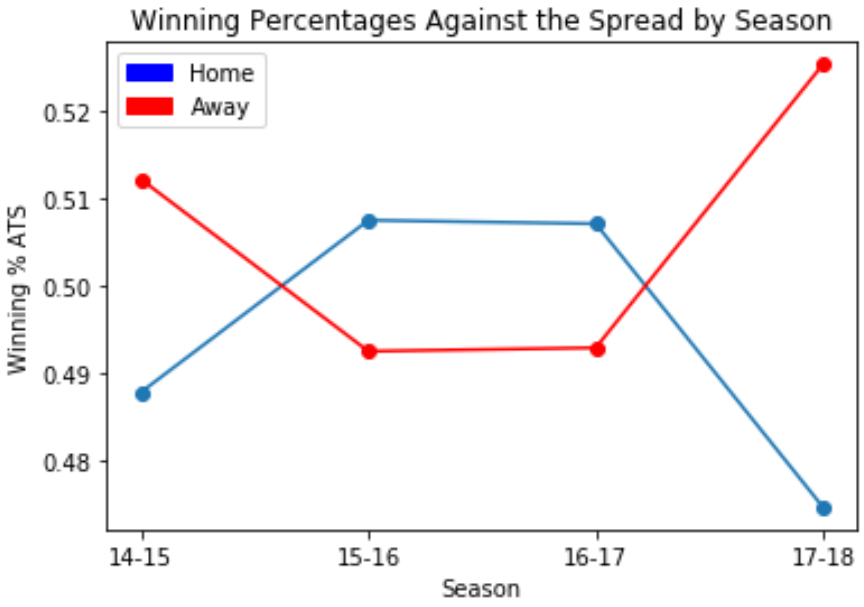




These plots representing a team's season points differential against their against the spread (ATS) winning percentage generally follow a pattern of appearing to be positively linearly correlated. This makes sense because a team that wins by a lot of points is likely to win more games by covering the spread and vice versa for teams that lose by a lot of points. There are exceptions to this trend though, with many coming in the 2016-2017 season, which looks noticeably different from the other plots. In this plot, five of the eight teams with the season's

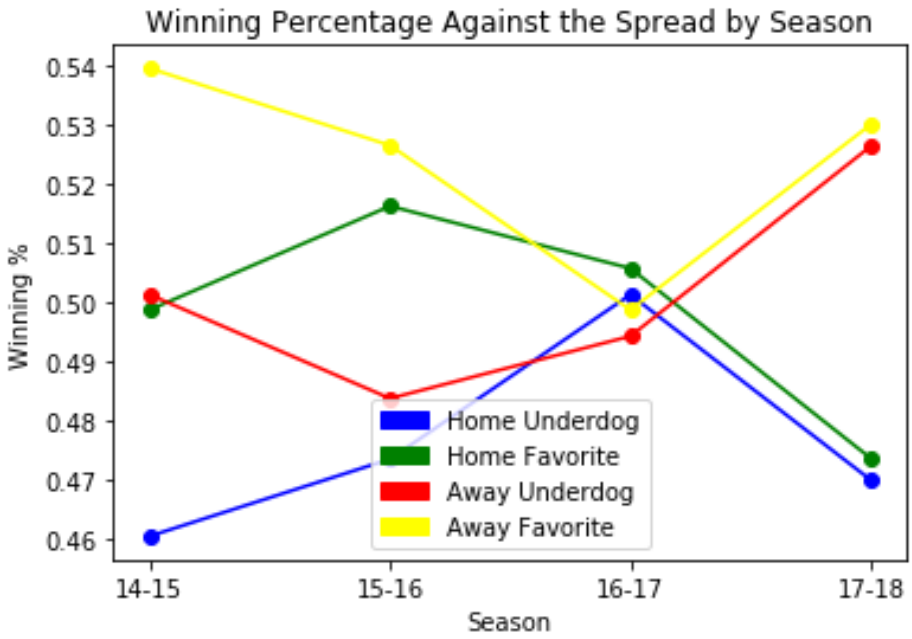
lowest differentials actually had an ATS winning percentage of greater than .500. As a bettor it could be very profitable to bet on teams having large disparities in their points differential. If this is the case, the spread would likely reflect that, but an underlying feature could be a team's recent bulk of games point differential or their differential facing similar opponents. Evaluating a team's specific points margin for certain games or the whole season could often indicate who will win a game.

From 2014-2018, in the regular season home teams won their games over 58% of the time. The home court advantage is something that may bias potential bettors into believing that home teams convert against the spread at a high rate as well.



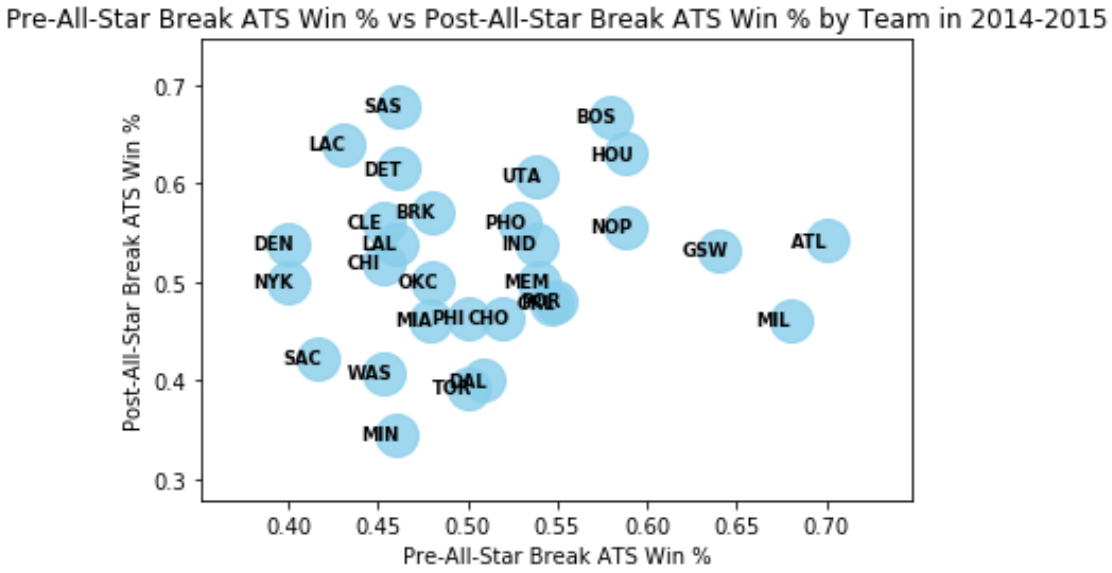
Based on this graph, there is no advantage when playing either at home or on the road for a team in covering the spread for. This makes sense because spread values factor in the difference in points associated with playing home or away games. All of these values are very close to 50%, with the most deviation in the last season with road teams covering the spread almost 53% of the time. It is difficult to say what this can be attributed to without looking at the data on a game by game basis. Although home/away splits as a whole for the NBA does not

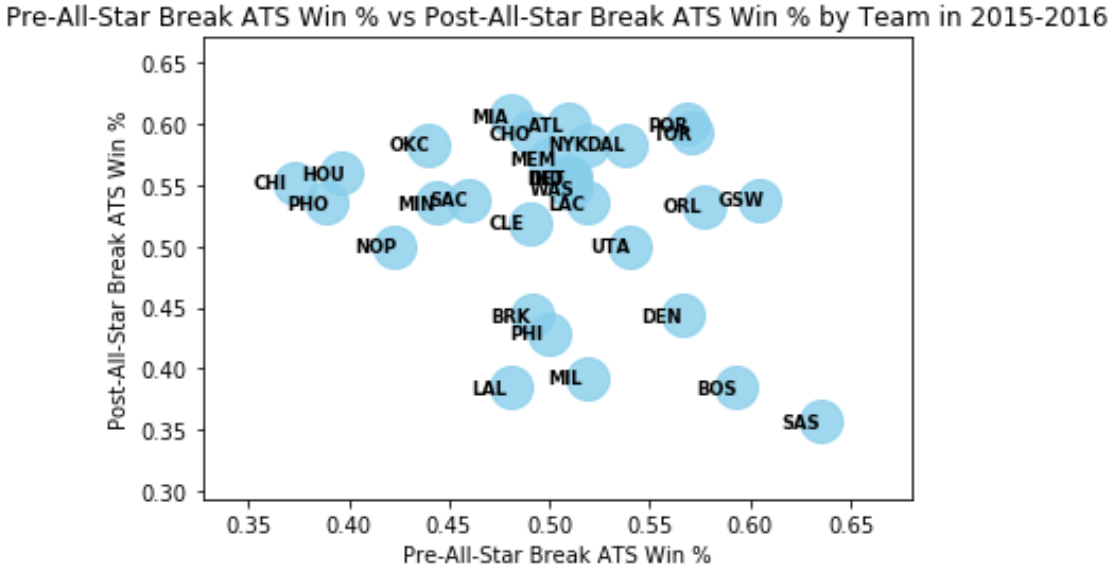
provide significant value in telling us when teams will cover, that does not mean certain teams do not show important trends in covering the spread based on whether they are in their home stadium or on the road. We can break this data down further to view teams seasonal trends at home or away as a favorite or an underdog.

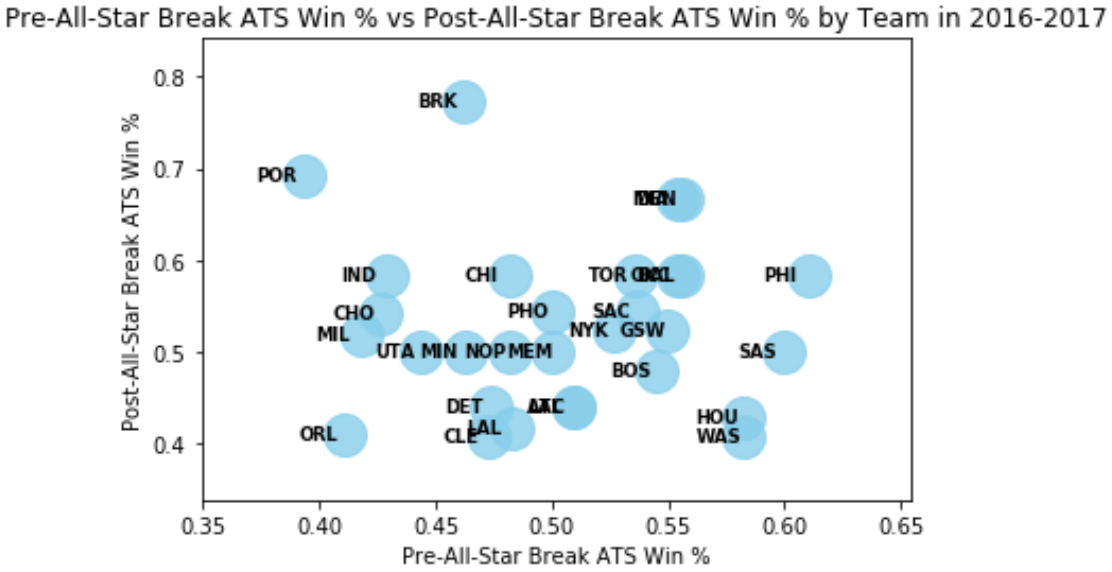


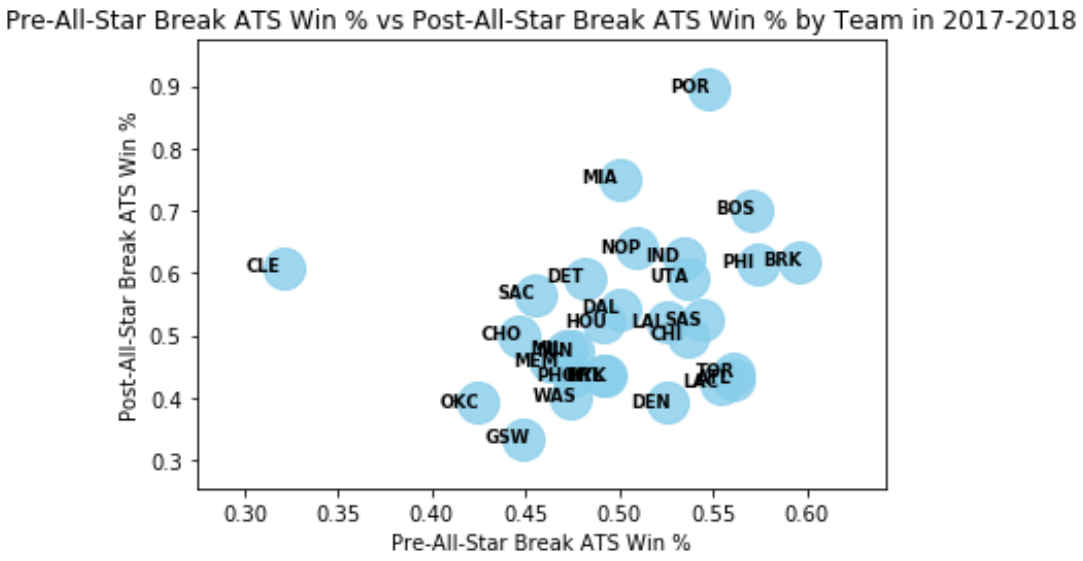
Because the spread is intended to level the playing field, it is expected that these season averages hover around 50%. For the most part this is true, but there is some fluctuation in these values. The greatest outlier is in the 2014-2015 season, where road favorites covered the spread at a 54% rate and thus home underdogs covered at just a 46% rate. The blue line debunks the popular theory that there is value in “home dogs”. With three of the four years as under 50%, it would not have been smart to consistently wager on underdogs playing at home. This does not mean home dogs never have value, but in order to bet them successfully one would need to know the circumstances that they perform best under. While these differences in values are interesting and can likely be explained by in-season data, these values could not help a potential gambler in predicting the result against the spread because there is no clear pattern.

If a team is doing well or poorly throughout the year, how often has their past performance been indicative of their future play? For purposes of actually wagering on NBA games, it could be of more value to see how these ATS percentages vary from the beginning to the end of the season. A natural point to compare these percentages in the NBA season would be the All-Star break. If a team performs well ATS at home, away, or both, before the All-Star break, it would be interesting to see if a team's performance remains consistent for the end of the season. Below are graphs comparing team's pre and post ATS winning percentages:





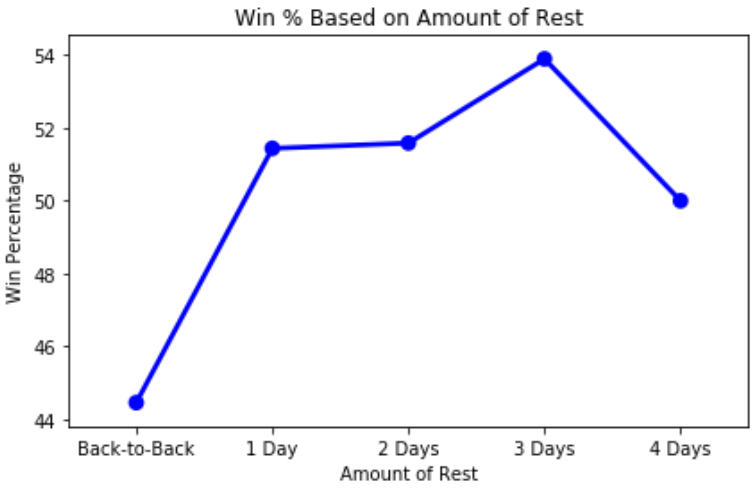


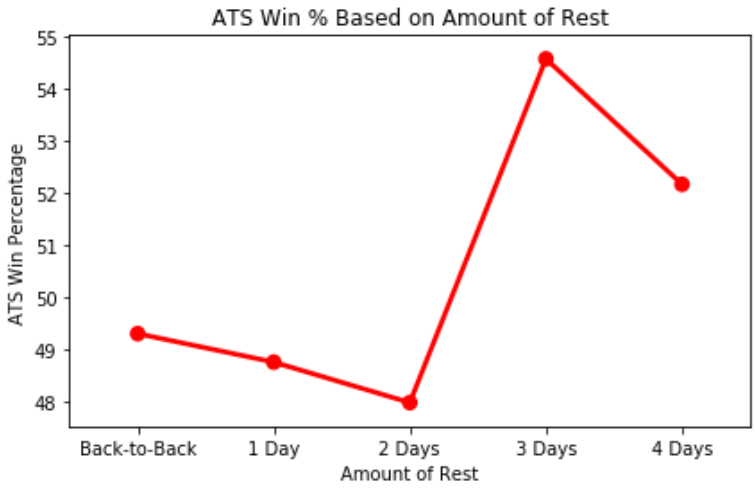


Based on the data, there is a fair split of teams that have consistent results from before to after the break and those that change their ATS fates dramatically. The most obvious outlier comes in the 2017-2018 season with Cleveland. They had by far the lowest ATS percentage before the break but raised that to over 60% in the second half of the year. Results like this are possible because of oddsmakers curbing their lines more appropriately, or changing them more

favorably if the betting public is often wagering against them. Or it could just be a case of teams reverting to the norm. There could be many factors why some team's remain consistent, and some teams do not. If a team is good and fighting for a playoff spot or better playoff positioning they would likely have consistent ATS percentages before and after the break. But if a team on the fringe of making the playoffs starts to scuffle, they could start performing worse as their incentive to win goes down throughout the year. In any case like this, a team by team, game by game analysis of these splits would be best to truly uncover what is happening, but these overall trends are important in understanding what generally takes place for ATS performances throughout the year.

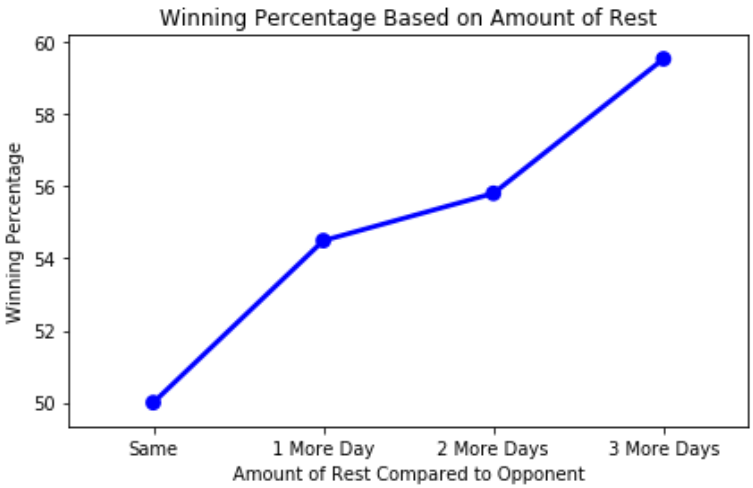
One variable for NBA games that is determined before anyone even touches a basketball during the season is the schedule. With an 82 game season, injuries and traveling it can be difficult for a team to be ready to play at their full potential. The upcoming analysis will take a look at how the amount of rest of teams and their opponents get between games impacts performance:

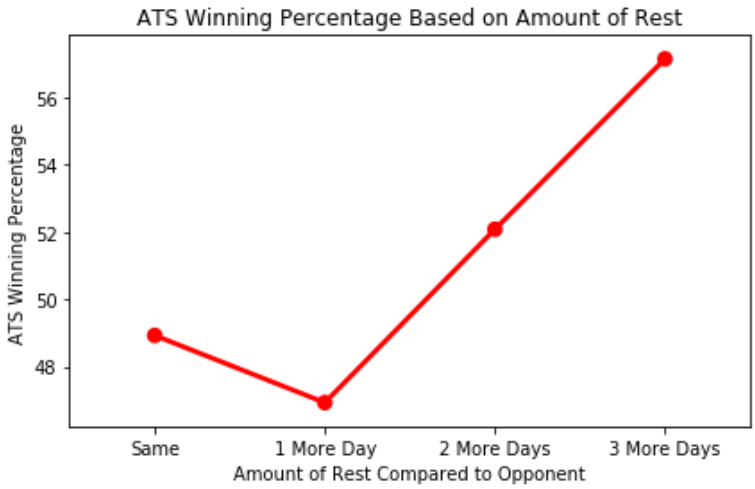




These two graphs vary dramatically, with the regular winning percentage based on rest before more intuitive. There teams playing on a back-to-back lose almost 44% of those games, and that percentage increases to around 50% and above for normal rest days. The ATS graph is a bit more confusing, with 0-2 days of rest being under 50% and the only outlier is 3 days of rest

above 54% ATS performance. It could be interesting to see how a team's rest compared to their opponent's rest impacts these percentages.





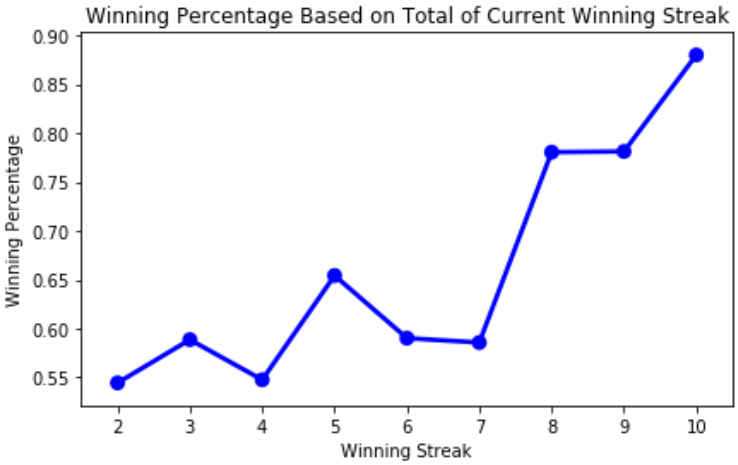
These graphs are what we would expect in terms of the win percentages rising, and unlike the previous graphs, this ATS graph rises similarly to the regular winning percentage. It is clear having more rest than your opponent is a strong indicator of whether or not a team will win a game, especially as the rest differential increases. For ATS percentages, having 1 more day of rest actually has a less than 50% win rate which is interesting, but as the differential goes up ATS percentages steadily rise. It should be noted that these percentages might be slightly less than perceived because it is including ties, and they are not counted as ATS wins, even though they are not losses either.

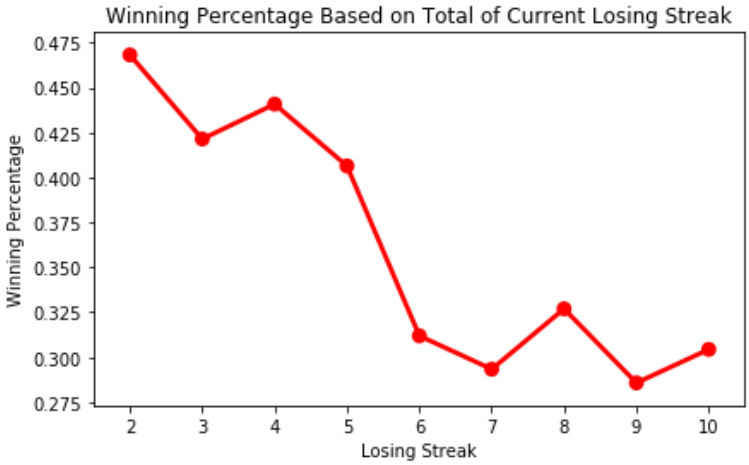
Another interesting factor determined by a team's schedule is when they play a team in consecutive games. Does their performance in the first game have any influence in the next game vs that team?

- When a team wins the first game of a back-to-back vs the same opponent, they win the second game at a 52.1% rate.
- When a team loses the first game of a back-to-back vs the same opponent, they win the second game at a 46.5% rate.

- When a team loses the first game of a back-to-back vs the same opponent at home, they win the second game on the road at a 48.7% rate.
- When a team loses the first game of a back-to-back vs the same opponent on the road, they win the second game at home at a 43.8% rate.

These percentages are a bit surprising. In the second game of a back-to-back vs the same opponent, if a team wins the first one they are more likely to win the next one and complete the sweep, and if they lose the first one they are even more likely to lose the next game,



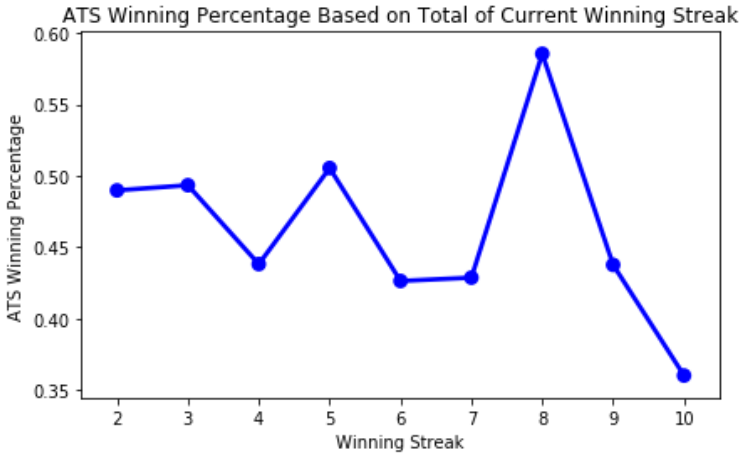


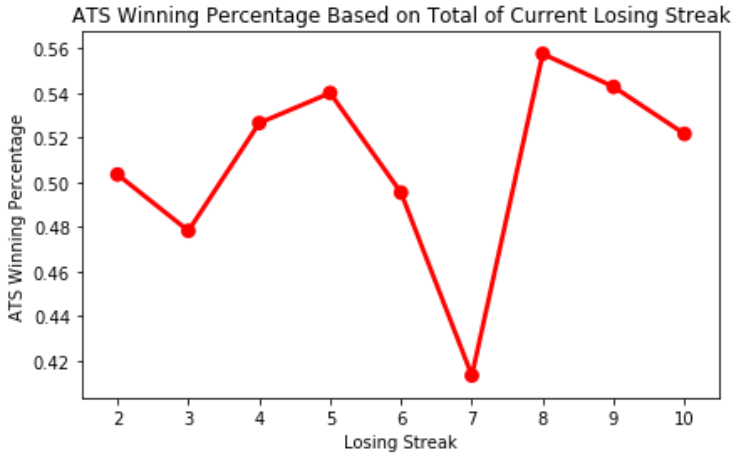
debunking the "revenge game" theory that a team is not as likely to have the same result in these back-to-back situations. Maybe even more surprising is that when a team loses the first game at home as opposed to away, they are more likely to win the second game on the road than the second game at home. These percentages are relatively close to 50% as it is hard to take definitive insight on them for betting purposes, other than knowing not to bet a team just because they won or lost vs the same opponent they just played. However, it could be worthwhile to see which teams perform best in these scenarios across various seasons.

Another potentially valuable measure to investigate is how does a team's recent performance affect their winning percentages. Below will be analysis of team performance based on their winning and losing streaks as well as previous 10 game performance.

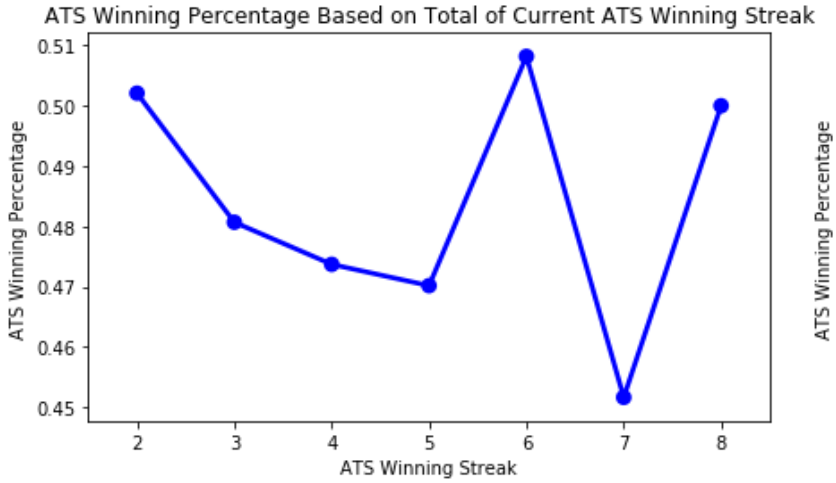
These trends make sense, as a team's winning and losing streaks increase, they become more and less likely to win the next game. This shows the importance of how a team's most recent very good or bad performance does impact the next game for a team. If a team is on a

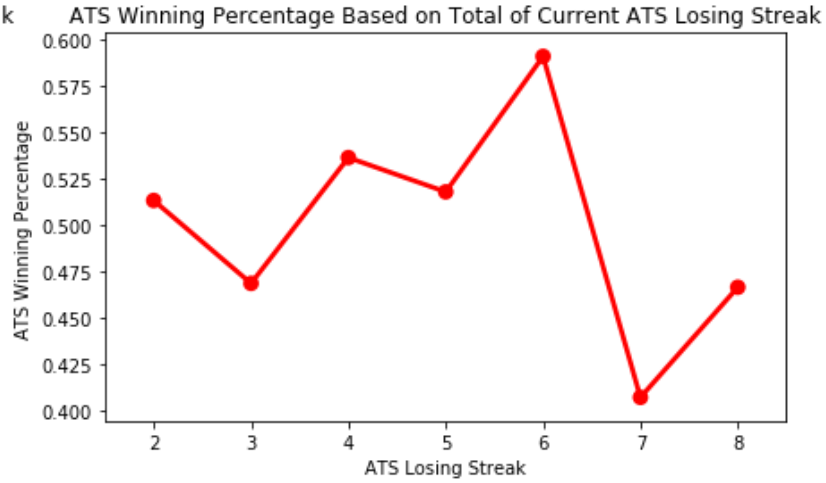
streak, they are more likely to continue that streak than to break it in the most current game. Let's see if winning streaks have any noticeable effect on team's ATS performances:





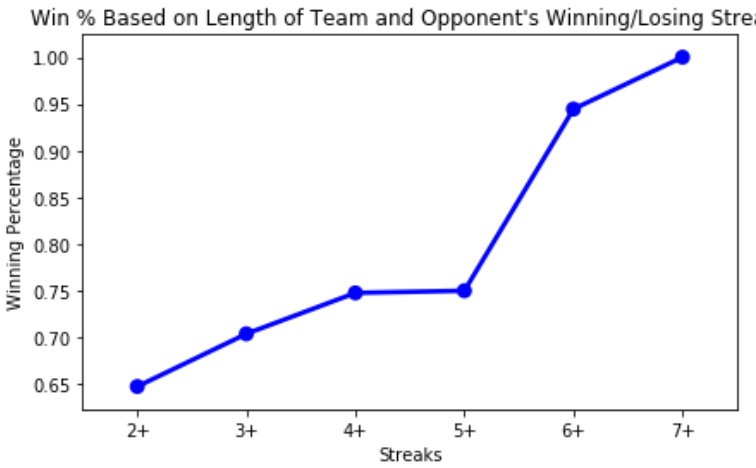
There is no pattern to these results, and it is likely a team's actual streak has nothing to do with whether or not they will perform any better or worse against the spread. This can be because a team's recent performance is factored into the line of a game. It could be possible that a team's streak against the spread could be indicative of their upcoming performance.

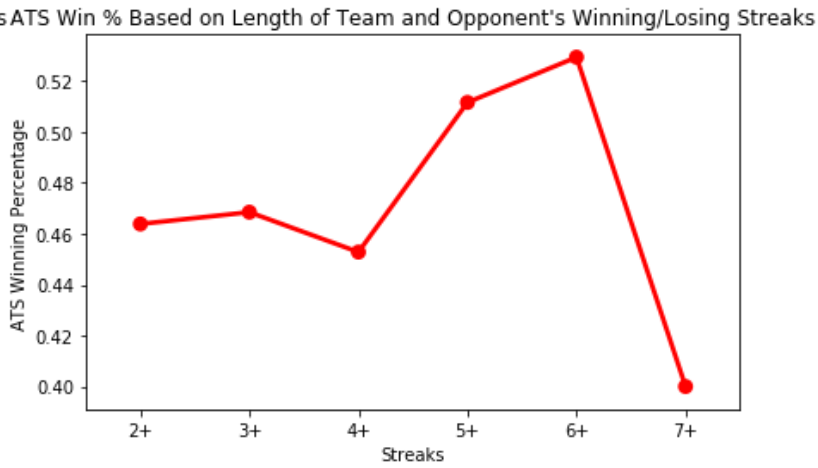




It looks difficult to make sense of these trends, as they fluctuate greatly. It does not appear that a team's most recent and extreme performances against the spread will impact whether or

not a team will cover or not in their next game. If a team is consistently not covering, oddsmakers could The next analysis will evaluate how teams streaking in opposite directions perform against one another.

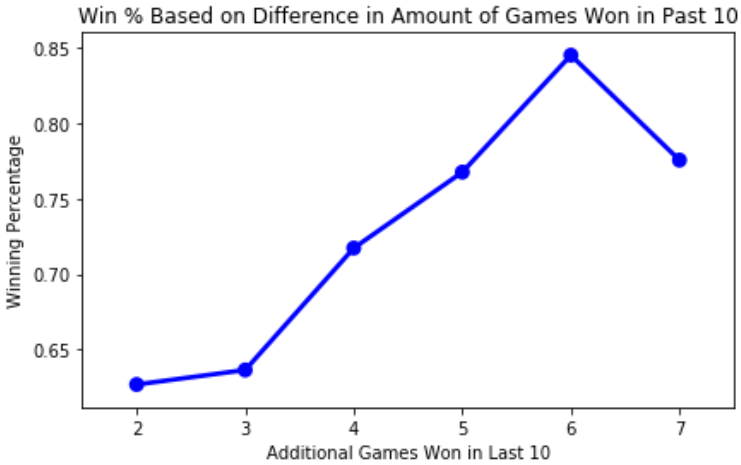


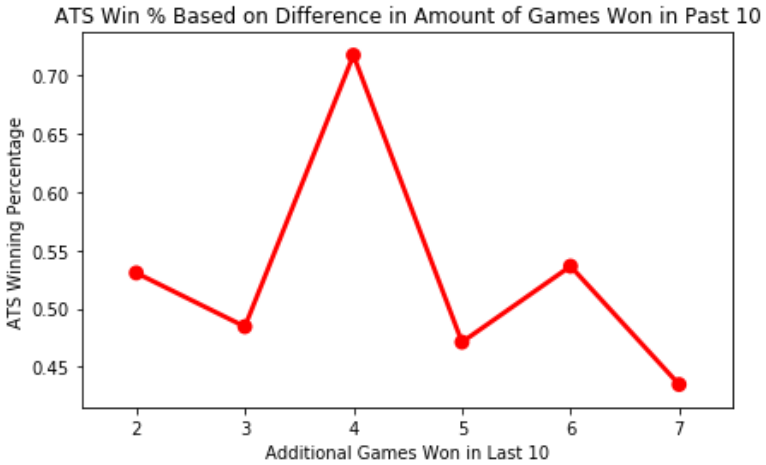


The graph on the left for regular winning percentage is largely what we would expect, with the winning percentage consistently rising as the length of opponent's winning and losing streaks increase. Teams that are hot playing teams that are cold are more likely to win. As these streaks increase the percentages become more dramatic. There were 10 games from 2014-2018 where a team with a winning streak of 7 or more played a team with a losing streak of 7 games or more and they won each time. The graph for ATS winning percentages is not as straightforward. It remains under 50% for streaks of 2 or more to 4 or more. Even though those teams are winning the games at roughly a 65-75% mark, they do not cover the majority of the time. This could be a result from oddsmakers factoring in the diverging performance of teams and having bettors overvalue these marks. Although as the streaks increase to 5 and 6 the percentage goes over 50%, it does not by much, and then drops significantly at 7. There is no true value for insight on evaluating ATS performance based on these streaks, but there is in just solely predicting a winner.

Another measure of a team's performance is how they have performed in their last 10 games, a good sample of how teams compare to one another in their most recent stretches. A team's last 10 games will include their streaks as well as breaks in those streaks, or just show if

of late they have been good, bad or average. These recent stretches will be compared to their opponent's and we will see how teams who have been winning more often of late fare against teams who have been losing more of late and vice versa.





The graph on winning percentage increases greatly based on team's previous 10 game winning percentages. This makes sense, as teams that are hot should win often against struggling ones. But, these results are extreme and do look valuable in predicting winners. The same cannot be said about ATS performances, as this graph varies greatly. A team's recent bulk of work is likely a factor in the line of the game, and is not a good variable to include when making a bet on the spread. A deeper analysis on this could prove important based on other variables within this trend such as if a team is an underdog or on the road. It could be that if a team that has done well recently is playing a team that hasn't but is still an underdog could be a worthwhile gamble. Of course this does not tell us everything about how a team is actually performing recently, as it doesn't give us insight on the type of competition they have faced, their stats, or their point margins in these games. But, the last 10 game record is a fair measure of how teams with varying degrees of their most recent performance compete against one another, and it appears to have strong correlation based on the left graph.

The analysis performed in this section has given us insight as to what are some critical factors and what are not as significant in trying to predict the winner of a game just based off of solely these trends. These visualizations have provided an understanding of how different types of winning percentages and metrics that are important in the scope of gambling are related with one another and when certain team win specific games. We have seen how variables such as a team's rest, their current steaks, and their given spread can help influence prediction, and when analyzed the correct way could be beneficial information for bettors to know. There have also been false predictors that may seem important but are not as significant as we may have previously thought such as where a team is playing and if they are facing the same opponent in a row.

Exploratory Data Analysis - Inferential Statistics & Machine Learning

The next section will evaluate which game statistics are most correlated with winning throughout a season. We will take a look at which statistics that occur during a specific game are the most important for a team winning a game, as well as which average statistics of a team coming into a game are the most associated with winning.

	win	FGM	FGA	FG%	3PM	3PA	3P%	FTM	FTA	FT%	OREB	REB	AST	STL	BLK	TO	PF
win	1	0.38	-0.042	0.45	0.24	0.033	0.32	0.14	0.11	0.11	-0.04	0.26	0.31	0.14	0.17	-0.12	-0.11
FGM	0.38	1	0.44	0.78	0.33	0.11	0.36	-0.18	-0.19	0.0049	-0.00092	0.08	0.64	0.098	0.062	-0.15	0.051
FGA	-0.042	0.44	1	-0.22	0.092	0.26	-0.14	-0.22	-0.22	-0.037	0.51	0.42	0.2	0.12	0.037	-0.28	0.074
FG%	0.45	0.78	-0.22	1	0.29	-0.066	0.49	-0.036	-0.047	0.031	-0.35	-0.21	0.55	0.021	0.042	0.028	0.0061
3PM	0.24	0.33	0.092	0.29	1	0.75	0.69	-0.099	-0.12	0.028	-0.12	-0.013	0.42	0.0081	-0.0021	-0.0058	0.033
3PA	0.033	0.11	0.26	-0.066	0.75	1	0.068	-0.092	-0.099	0.0072	0.017	0.09	0.23	0.047	-0.0056	-0.012	0.047
3P%	0.32	0.36	-0.14	0.49	0.69	0.068	1	-0.05	-0.066	0.033	-0.2	-0.12	0.37	-0.029	0.00076	0.0051	0.0054
FTM	0.14	-0.18	-0.22	-0.036	-0.099	-0.092	-0.05	1	0.92	0.33	0.052	0.063	-0.15	0.039	0.013	0.023	0.2
FTA	0.11	-0.19	-0.22	-0.047	-0.12	-0.099	-0.066	0.92	1	-0.029	0.092	0.091	-0.17	0.051	0.016	0.028	0.21
FT%	0.11	0.0049	-0.037	0.031	0.028	0.0072	0.033	0.33	-0.029	1	-0.09	-0.05	0.015	-0.018	0.0033	-0.01	0.00027
OREB	-0.04	-0.00092	0.51	-0.35	-0.12	0.017	-0.2	0.052	0.092	-0.09	1	0.56	-0.12	0.035	0.0078	0.045	0.046
REB	0.26	0.08	0.42	-0.21	-0.013	0.09	-0.12	0.063	0.091	-0.05	0.56	1	0.014	-0.11	0.17	0.13	0.005
AST	0.31	0.64	0.2	0.55	0.42	0.23	0.37	-0.15	-0.17	0.015	-0.12	0.014	1	0.094	0.088	-0.032	-0.0049
STL	0.14	0.098	0.12	0.021	0.0081	0.047	-0.029	0.039	0.051	-0.018	0.035	-0.11	0.094	1	0.0054	0.13	0.028
BLK	0.17	0.062	0.037	0.042	-0.0021	-0.0056	0.00076	0.013	0.016	0.0033	0.0078	0.17	0.088	0.0054	1	0.033	-0.011
TO	-0.12	-0.15	-0.28	0.028	-0.0058	-0.012	0.0051	0.023	0.028	-0.01	0.045	0.13	-0.032	0.13	0.033	1	0.17
PF	-0.11	0.051	0.074	0.0061	0.033	0.047	0.0054	0.2	0.21	0.00027	0.046	0.005	-0.0049	0.028	-0.011	0.17	1

This correlation matrix represents how all statistical variables are related to one another during a certain game. The top row or leftmost column is of the most importance to us because it shows the relationship between a certain statistic and a win for a team. The statistics that are most correlated with a win are field goal percentage (FG%), field goals made (FGM), three point percentage (3P%) and assists (AST). These are directly related to scoring and higher marks would determine a more efficient team. The statistics with the smallest correlations are personal fouls (PF) with a negative correlation, free throw attempts (FTA), and (FT%). It is surprising that these free throw related values are seen as the least indicative of a win here, but could be because they just account for one point each opposed to field goals and threes, which are two and three points respectively.

After testing different models to predict wins based off these statistics, the gradient boosting classifier performed the best, with a 83.64% success rate. This shows how teams can learn what to focus on in order to win more games just by looking at basic statistics alone. This is no big secret in the NBA. Because we do not know the stats of a game before it happens, we need to analyze how the average statistics of teams can predict performance and which indicators are the best and worst at determining wins.

Of course we do not know the statistics of a game before it happens, so in order to try and predict wins before the game starts, we could use a team's average statistics up to that point in the season.

	win	A_FGM	A_FGA	A_FG%	A_3PM	A_3PA	A_3P%	A_FTM	A_FTA	A_FT%	A_OREB	A_REB	A_AST	A_STL	A_BLK	A_TO	A_PF
win	1	0.11	-0.013	0.15	0.1	0.079	0.095	0.033	0.016	0.051	-0.024	0.042	0.12	0.065	0.066	-0.048	-0.055
A_FGM	0.11	1	0.59	0.77	0.38	0.27	0.42	-0.085	-0.16	0.19	-0.032	0.25	0.67	0.21	0.17	-0.0069	-0.092
A_FGA	-0.013	0.59	1	-0.068	0.24	0.29	-0.043	-0.14	-0.14	-0.017	0.37	0.54	0.23	0.11	0.019	-0.028	0.027
A_FG%	0.15	0.77	-0.068	1	0.28	0.098	0.56	0.0025	-0.092	0.25	-0.33	-0.11	0.63	0.16	0.19	0.0072	-0.14
A_3PM	0.1	0.38	0.24	0.28	1	0.94	0.52	-0.086	-0.12	0.092	-0.26	0.063	0.38	0.12	0.027	0.088	-0.073
A_3PA	0.079	0.27	0.29	0.098	0.94	1	0.22	-0.064	-0.069	0.024	-0.21	0.086	0.27	0.12	0.0043	0.11	-0.03
A_3P%	0.095	0.42	-0.043	0.56	0.52	0.22	1	-0.11	-0.19	0.2	-0.24	-0.045	0.41	0.044	0.064	-0.017	-0.13
A_FTM	0.033	-0.085	-0.14	0.0025	-0.086	-0.064	-0.11	1	0.92	0.28	0.17	0.14	-0.14	0.063	0.025	0.0094	0.23
A_FTA	0.016	-0.16	-0.14	-0.092	-0.12	-0.069	-0.19	0.92	1	-0.098	0.26	0.19	-0.2	0.064	0.021	0.068	0.25
A_FT%	0.051	0.19	-0.017	0.25	0.092	0.024	0.2	0.28	-0.098	1	-0.23	-0.13	0.16	0.021	0.022	-0.15	-0.048
A_OREB	-0.024	-0.032	0.37	-0.33	-0.26	-0.21	-0.24	0.17	0.26	-0.23	1	0.57	-0.22	0.0024	-0.0049	0.13	0.15
A_REB	0.042	0.25	0.54	-0.11	0.063	0.086	-0.045	0.14	0.19	-0.13	0.57	1	0.043	-0.19	0.23	0.18	-0.046
A_AST	0.12	0.67	0.23	0.63	0.38	0.27	0.41	-0.14	-0.2	0.16	-0.22	0.043	1	0.26	0.25	0.11	-0.14
A_STL	0.065	0.21	0.11	0.16	0.12	0.12	0.044	0.063	0.064	0.021	0.0024	-0.19	0.26	1	0.033	0.26	0.16
A_BLK	0.066	0.17	0.019	0.19	0.027	0.0043	0.064	0.025	0.021	0.022	-0.0049	0.23	0.25	0.033	1	0.11	-0.0068
A_TO	-0.048	-0.0069	-0.028	0.0072	0.088	0.11	-0.017	0.0094	0.068	-0.15	0.13	0.18	0.11	0.26	0.11	1	0.35
A_PF	-0.055	-0.092	0.027	-0.14	-0.073	-0.03	-0.13	0.23	0.25	-0.048	0.15	-0.046	-0.14	0.16	-0.0068	0.35	1

This matrix represents how all of the average statistical variables of teams are related to one another throughout a season. The top row or leftmost column is of the most importance to us because it shows the relationship between a certain average statistic and a win for that team. It appears that the statistics that are most correlated with a win are average field goal percentage (A_FG%) and assists per game (A_AST). These are again directly related to scoring and higher marks would determine a more efficient basketball team. The average statistics with the smallest correlations are free throw attempts per game (FTA) and field goal attempts per game (A_FGA). It is possible that the values representing team attempts for a certain type of shot have the least significant because it does not matter how many shots they take, and what is important how many they make and how many the other team make and attempt compared to their totals.

None of the models solely based on a team's average statistics up to that point in the season performed very well. The K Nearest Neighbors model with a k value of 80 performed best giving a score of close to a 60% correct prediction right. This could be because a team's stats early in the year were not indicative of their performance, or just because the model is lacking crucial information.

The team's statistics are not the only important numerical features occurring in the game, mainly because they are facing an opponent that will have its own strengths and weaknesses. Which team statistics compared to their opponents are significant in terms of winning and which aren't? Below we will evaluate a couple of highly popular basketball statistics.

One important statistic often referenced during games is field goal percentage (FG%), which is equal to the percentage of made shots by a team during gameplay (not including free throws)

Null hypothesis (H0): There is no significant relationship between having a higher average field goal percentage than your opponent and winning

Alternative Hypothesis (H1): There is a significant relationship between having a higher average field goal percentage than your opponent and winning

The result of a one sample t-test yielded:

Ttest_1sampResult(statistic=13.903051918643442, pvalue=4.0362161127095995e-43)

Based on these values and our very small p-value of close to zero, we reject the null hypothesis that having a higher average field goal percentage than your opponent is not significant towards winning. Logistically this makes sense as teams that shoot better compared to their opponent would be thought to having an advantage. Teams with a higher average FG% win at a 59.79% rate, which seems very high.

Another important game statistic is offensive rebounds (OREB). This is equal to the amount of times a team gets the rebound after their own missed shot.

Null hypothesis (H0): There is no significant relationship between averaging more offensive rebounds than your opponent and winning

Alternative Hypothesis (H1): There is a significant relationship between averaging more offensive rebounds than your opponent and winning

The result of a one sample t-test yielded:

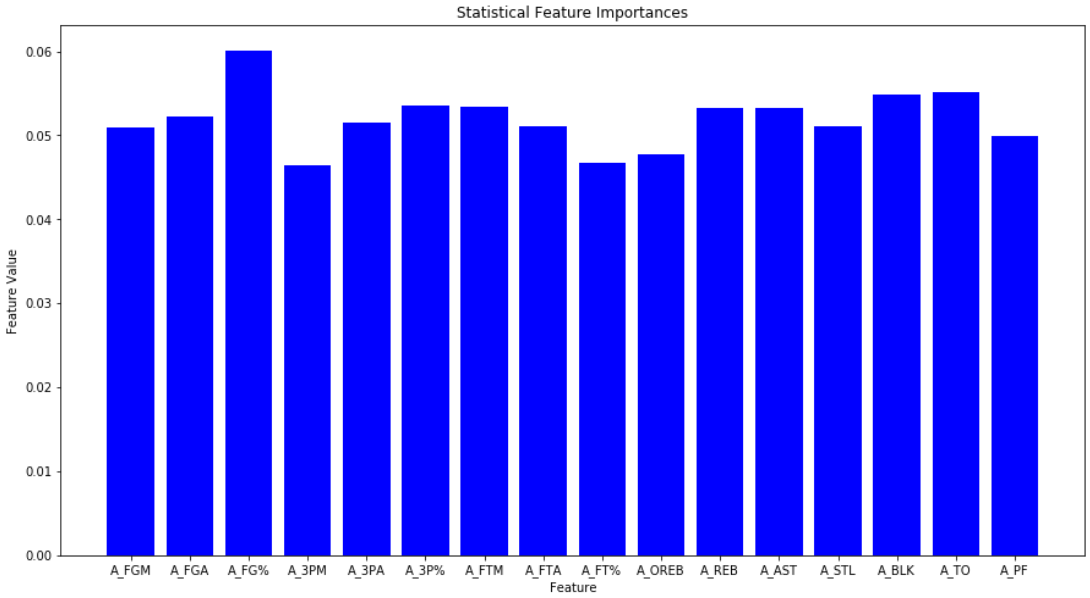
Ttest_1sampResult(statistic=-2.8510962776697903, pvalue=0.004375360137271079)

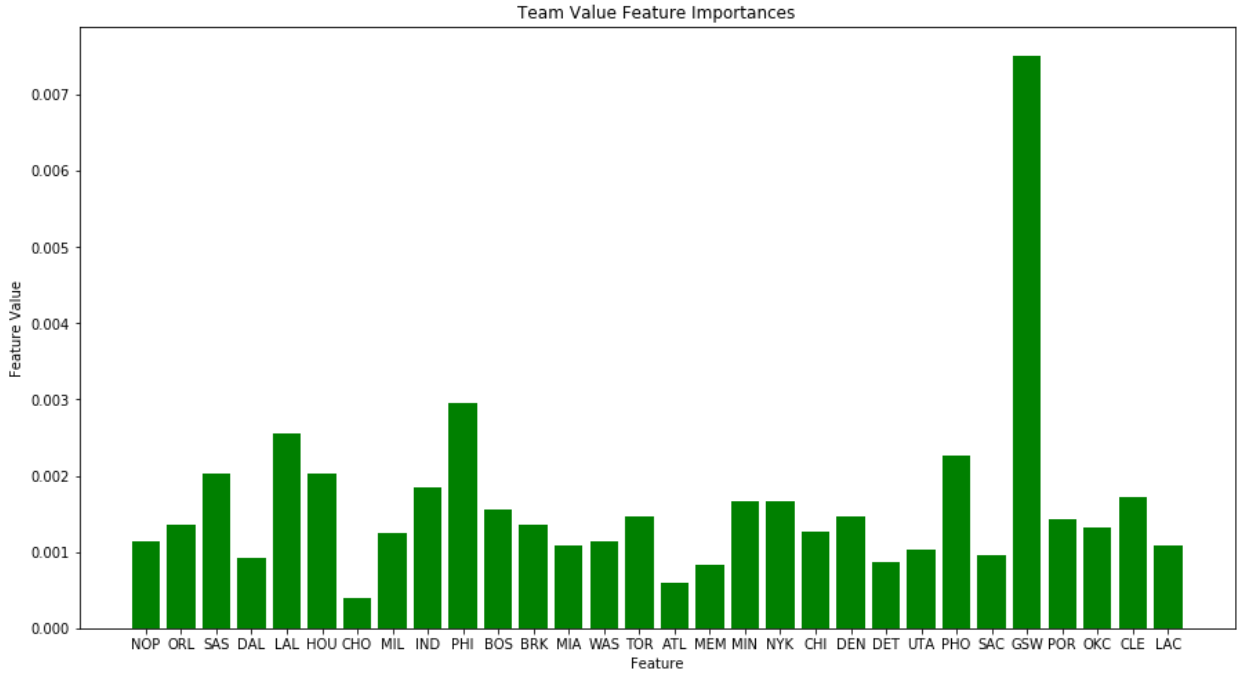
Based on these values and our p-value of .004, we can not reject the null hypothesis that averaging more offensive rebounds than your opponent is not significant towards winning. This makes sense as teams that average more offensive rebounds coming in compared to their opponent only win 47.95% of the time. This could be because they are getting more offensive rebounds since they don't shoot as well as other teams and miss more shots.

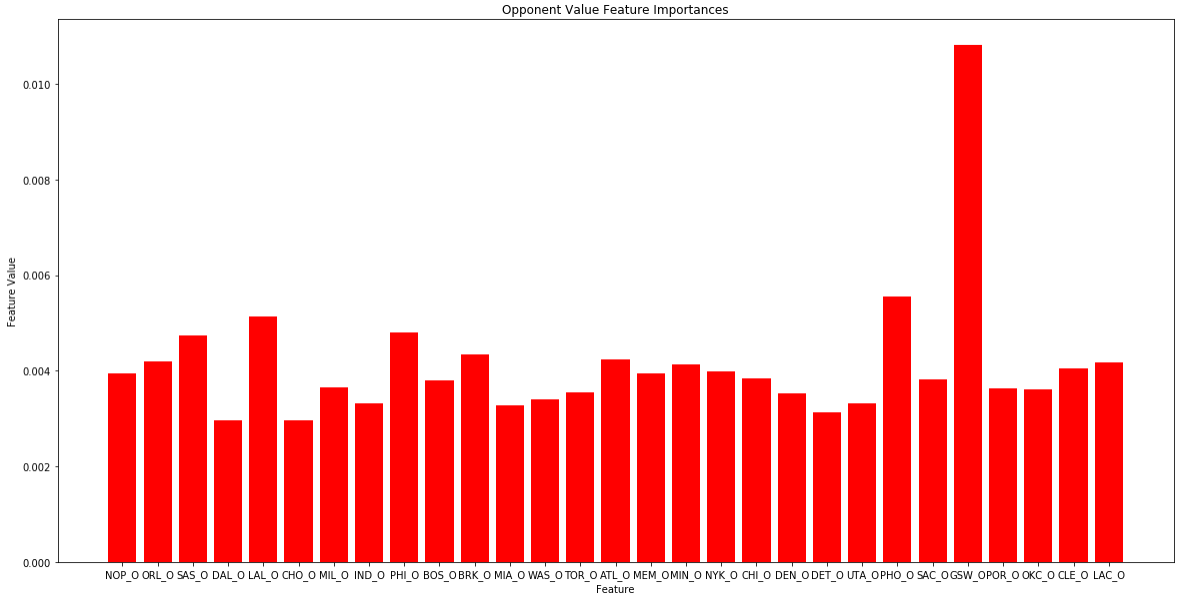
Next we will evaluate more models and see how much predictive value the game statistics along with other variables have in forecasting a winner, and which combinations of factors can

give the best results. The main machine learning algorithms tested to predict these win rates are K Nearest Neighbors, Random Forest and Gradient Boosting. We will see which consistently works the best under what specifications. All models referenced above and below are built using 80% of the game data as training data and 20% as test data, which are the games being predicted on by the models.

It is also important to note which teams are playing. If the teams playing are known, the model should forecast results better by understanding which teams perform well against other certain types of teams based off of their season statistics. Below are the feature importance values of the random forest model for these variables:







The Warriors (GSW) have by far highest feature importance for a team and when playing as an opponent. This is likely because their statistics diverge most from other teams because of their record scoring production, along with winning among the most games in the NBA each year it is easier to predict when the Warriors are going to win than it is for other teams. They are the only team with a greater value than any statistic. Average field goal percentage has the highest importance here, which is consistent to what we have seen earlier. On average the opponent's feature importance has more value than which team is playing. This is interesting and shows the stats on their own are not very telling of who will win, but combining that information with how others teams perform vs similar teams with those statistical trends yields a more accurate predictive model. Despite this, the random forest model accurately predicted a win 57.46% of the time while the gradient boosting model had a 62.86% accuracy rate.

The next iterations of the model inputted the home variable which denotes if a team is playing at home or on the road. The random forest rate increased to 57.56%. The confusion matrix for the test data is as follows:

```
[634, 327]
[498, 485]
```

These values are consistent with the other models used for Random Forest Classification, where it does not perform very accurately. Wins are predicted more accurately than losses leading to more Type 2 error. But, there are more false negatives than true negatives predicted, showing the random forest is not a good model to follow. The Gradient Boosting classification with the home variable inputted correctly predicted a win 65.28% of the time. The confusion matrix is as follows:

```
[620, 341]
[334, 649]
```

This matrix shows very consistent predictions, which is in line with the highest prediction rate recorded yet. There are very close to as many wins and losses predicted accurately, meaning there are almost an identical amount of false negatives as false positive values.

Using these variables, others models were also created based on the portion of the season as opposed to the full season which we have been working with. These sections include before and after the all star break, and all games besides a team's first 10 games played in a season. I theorized that a model using statistics that don't include early season data would lead to statistics that are more indicative of a team's true performance and yield better results. However, all of these iterations ended up having less predictive value than using all season data, meaning the more information about a team during the season the better we can forecast wins, regardless of when the games are played.

The last analysis conducted altered the statistics chosen for the model. After factoring in various combinations of statistical variables to incorporate and exclude, it was found that only removing average personal fouls per game (A_PF) increased performance. All other variations of which statistics to use lessened the model's prediction rates. Fouls not being significant in terms of determining a winning team in a game is in line what we have seen this section in the correlation matrices.

Formulating a model to determine the result of any game just by using basic game statistics for teams that was accurate close to 66% of the time shows the value in these predictive measures and can give valuable insight for someone looking to make money by gambling on NBA games using the data to their advantage. These models can be broken down even further to include more advanced statistics, player statistics, time of the season, rest, weight recent play, and more to forecast a certain game. Like we saw with the Warriors feature importance, it could also be beneficial to separate the models by just individual team data to see when one team wins games. There are countless amounts of variations on different types of machine learning algorithms that can be explored to find the best measures of game prediction, and this analysis is truly just the beginning.

There are so many variables that do and do not factor into an NBA team winning a game on any given day. When looking at the context of a whole season, it can be even more difficult to find what trends matter and which would be foolish to follow. However, this analysis has shown that are some definitive factors that more often than not have predictive value. Although this is valuable, you can't say that it necessarily is for every game, which could be problematic for someone actually trying to wager on games. This project has some flaws in evaluating what could be the truest trends throughout a long stretch of time in the NBA. Only the four most recent full regular seasons were analyzed, and even though that is almost 5000 games played, the NBA has been in continuation since before 1950. Also, basketball has a myriad of more advanced stats that can be analyzed and have been proven to show importance in evaluating player and game data. The statistics used here are just some of the most fundamental ones applied in basketball, and do not even take into account individual player statistics. There are also many qualitative variables that come into play that you can't just stick a data figure too. Coaching, team chemistry, and simple things such as how a star player is feeling that day all can have a major impact on a game, so going by the numbers is not always an exact science.

In summation, it is difficult to know what exactly to look at in order to predict game results, but by looking at the correct data there is a lot to learn about the nuances within an NBA game and

season for betting purposes, understanding the important variables in what makes of a winning team, and yearly trends for teams and the league as a whole. At the end of the day, sometimes basketball just comes down to the randomness of how the ball bounces on the rim, but that doesn't mean we can't analyze it and break it down to a game of numbers as much as possible.

****

For a more in depth-look at the analysis conducted the link to the project code is below:

https://github.com/jebert10/NBA-Capstone-Project/blob/master/NBA_Capstone_Code.ipynb