NFL Sabermetrics

By Daniel Reiter

TABLE OF CONTENTS

1. <u>Proposal</u>

To all NFL owners, thank you for the time to discuss my proposal. As we all know, hiring a general manager or coach is not an easy task. With how competitive the league is, you really only get a couple years before deciding if the people in power are the right ones. But could there be a better way to predict long-term success? My proposal is to go deep into NFL statistics from the last few years and see what statistical trends bettered each team's chances of success and could possibly create a predictive model.

There is a fine margin for success in the league. A few plays here or there can make the difference between finishing 10-6 or 6-10. The league really is that competitive. And one of the issues that face teams is their coaching philosophy. Some are big believers in running the ball while others strive for a dynamic passing attack. The question is how much does either really impact the chances of a team's success? Is the total amount of rushing yards more or less important than the amount a team averages per game? Should teams focus more on passing yards than pass attempts? And let's not forget about the defense. What statistics should they focus on when crafting their schemes? Are sacks more important than yards given up? Does creating turnovers give a bigger boost to a team's winning percentage? Also, how important is a team's success on fourth down? Should they go for it more often? All of these are possible questions that could be answered by this deep dive into NFL statistics.

My method for getting all this information is to go team-by-team and game-by-game from 2015 to 2017 and see what statistics really gave each team the better chance of winning. While there are only 32 teams, each is constructed very differently with personnel as well as coaching schemes so there can be some issues with a simple look at the NFL as a whole. You need to go through each game for each team to really see what led to their success and failure. And the reason to keep it within only 3 years is how much turnover occurs in the league. Teams change drastically each season when it comes to personnel. While winning teams may make small changes, losing ones can overhaul their entire roster. It's important to keep the seasons close together to better the chances of obtaining a more consistent dataset.

And when it comes to datasets, there are plenty out there to help dive into all of these statistics with the main one being the Stattleship API. It has game-by-game statistics for each team that will give me the main information I need. There will also be more feature-specific statistics in the project that focus on where a team is by a certain point. For example, there could be a column for average rushing yards through the first half of the year or median amount of points against for each team going into the final stretch of the season. These features and others will assist in making this project more successful.

When all is said and done, the purpose of this project is to see what statistics truly measure a team's success as well as figure out how much they can predict a club's future and help you as owners make better decisions when hiring a coach or general manager. I will have my code on my GitHub repository, a report that describes the process I went through in obtaining and wrangling the data as well as what I discovered plus a PowerPoint presentation to discuss it. There is certainly a need for this type of research as everyone in the NFL looks to get a leg up on their competition. Yet no matter how much these datasets can help, there is still no accounting for the human element. There will always be unforeseen setbacks for each team like a bad injury or a horrible decision by a referee. But despite these possible difficulties, this project could have a major impact on the way we look at the NFL.

2. Data Wrangling Report

The dataset that I was wrangling from is an API called Stattleship. It has a wide assortment of sports statistics from all the major American sports leagues including the NFL. Whether you want to query by season, by game, by team or by player, Stattleship has the info you need for any project like this. And one thing that made it more convenient was the ability to import the API as a package to use on my Jupyter Notebook to make querying data simpler than from a standard API.

For this project, I needed to get the game logs for each team so I could go through each game from the 2015-2016 season through the 2017-2018 season. In exploring the API, I found most things were labeled conveniently and there really was no missing data. When a team did not achieve one of the features, it gets a “0” (or “False” for Boolean features) instead of “NaN” making it simpler to deal with missing values since they do not need to be dropped or replaced. Also, while there are some outliers, they are standard for a normal NFL season and should still be included because of how much uncertainty there is in the league. In fact outliers are necessary in a project like this because I am trying to find what gives a team the best chance to win and those outliers can have a major impact on a team’s ability to win.

However, one of the biggest issues with using Stattleship is that the API limits how much it will display for a query. So, for example, if I try to query all the stats from one season, I will only get about 40 games worth of statistics when in fact there are many more games. This means that in order to get the game logs that I want, I needed to query many times since there are 32 teams with 16 games for each of the three seasons I am looking at. Clearly, I could not query individually like that so I needed to work on functions and For Loops to obtain the data.

After exploring different ways to query the data and what indexes are used to get certain information, I worked on a function that would help transfer the game names to game slugs. To do that, I made a list that split up the game’s name by spaces, several dictionaries for the key indexes that needed to be changed (including a tuple to make the season slug) and returned them all together. Later on, I would have to update the time dictionary as some game times were minimally different than expected as well as got rid of excess spaces that were in the names especially for games that occurred in January.

The next step was critical, making For Loops to loop through each team and make individual queries so I could get the names for each game. I made two separate lists: one had the three season slugs; the other had 32 team slugs (and thankfully, despite the Rams and Chargers moving, their slugs stayed consistent throughout each season). After that I used a For Loop that did a specialized zip on the seasons and teams to make them tuples that then made individual queries for each combination. These were all added to a list, which was then flattened and looped into another list in order to go through each index to get the game information from each team for each season. Lastly, I extracted the game names from this newer list into another new list. Now I had a list with all of the game names and the next step was getting the statistical information for each one.

After moving the game names to another list, I began creating a For Loop that would go through each game and grab the specific features I wanted. After making an empty list, I began looping through each game using the function I made earlier to transfer the names into slugs and query the data. I chose which features I wanted to include and created a Series to add them to, which would be concatenated to the list at the end of the loop. One issue that came up, however, was that the statistics are split up by the road and home team which means not only did I need to find the right list slicing for each side and use an If statement but also double up the features so each stat was included for the Road and Home team. Another issue was finding out the result of the game and which team won or lost since that was an important distinction. To do that, I defined another function (outside of the For Loop) that would give the names for the winners and losers as well as let me know which team won, the away team or home team. Also, in the rare cases that may arise, I was able to include if the teams tied. The function was then used inside the For Loop to get the winning and losing teams as well as the rest of the features to create a list, which was then transformed into a DataFrame.

Finally, because the For Loop goes through each game for each team, I needed to drop the duplicates in the DataFrame so each game is only listed once. Using the pandas' method "drop_duplicates," the DataFrame dropped all of the duplicate rows. I also filled the "NaNs" that occurred in most of the "Name of Tying Teams" column as well as the "Name of Winning Teams" and "Name of Losing Teams" columns for when a tie was achieved. None of this really impacts the data but I just wanted it to look cleaner. Now I had a DataFrame with clearly labeled statistics for each game for the last three seasons. To wrap it all up, I saved it to a CSV, with the delimiter being tabs, so that the data could easily be extracted moving forward.

3. **Data Story**

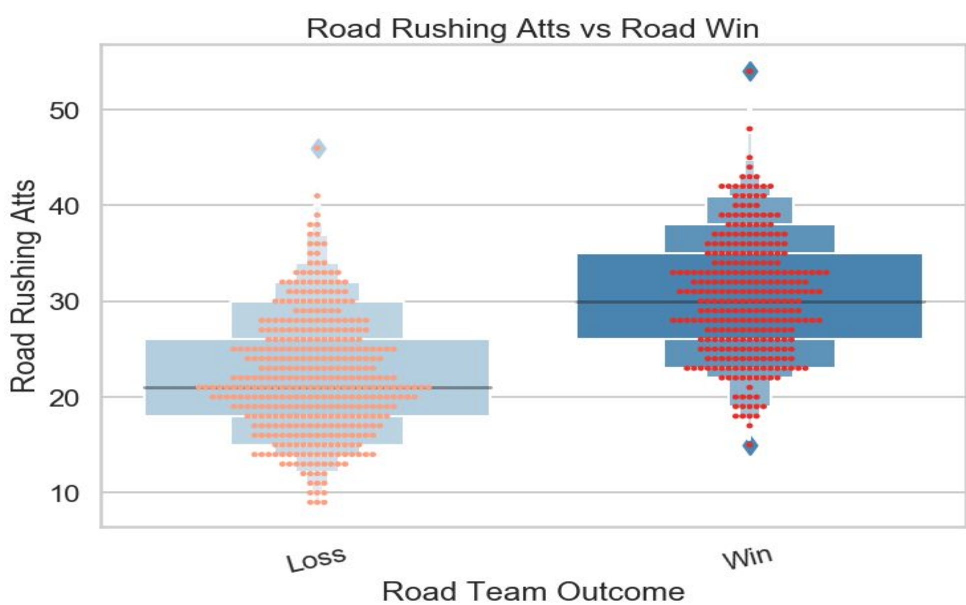
When it comes to plots regarding NFL games, there are a few important things to keep in mind. First off, while plenty of statistics have a strong impact on winning, I wanted to focus on certain trends and special impacts. Things like Total Points and Total Yards will always help your chances overall but that is too obvious. While there are many plots to discuss, I focused on the ones that caught my eye the most.

Also, as most of the Bar Plots will show, the Road and Home teams tend to start at different points when it comes to winning percentage. Road teams will usually start at lower percentages while Home teams start at higher percentages. This means the significance of a statistic impacting the two separate teams is focused on how much it increases their chances from the starting point to the end point. Many of the aspects that explain why Road teams start low are hard to quantify (traveling, unfamiliar hotels, lower quality locker rooms, etc.) but know that this is expected.

Finally, remember that these Bar Plots show the mean or calculated average so not every situation will end in the same result. With all of that out of the way, it is time to focus on plots that stood out to me for how they impacted a team's chances of winning.

Plot #1

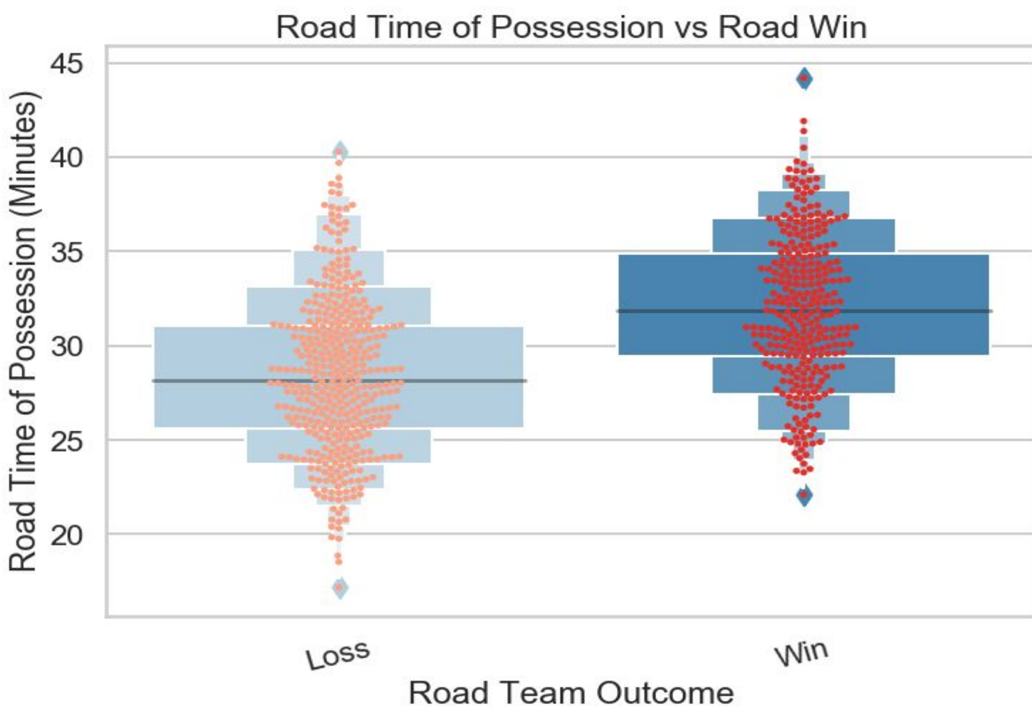
Although the NFL has become known as a passing league, many still say you need to run the ball to be a true title contender and win the Super Bowl. The question then is how much should you run the football? Many think about rushing yards as the true statistic and while there is an upward trend for winning and rushing yards, I am not sure that is the real key. Instead just attempting to run the ball is more important especially for away teams.



When looking at the above plot, we can see the median is almost 10 attempts different between winning and losing. There may be outliers for losing away teams who run the ball a lot but overall there is a significant increase for winning teams. The amount of attempts to run the ball has a very strong impact on the away team's ability to win. If the Road team runs it over 30 times, they have a very good chance of winning.

Plot #2

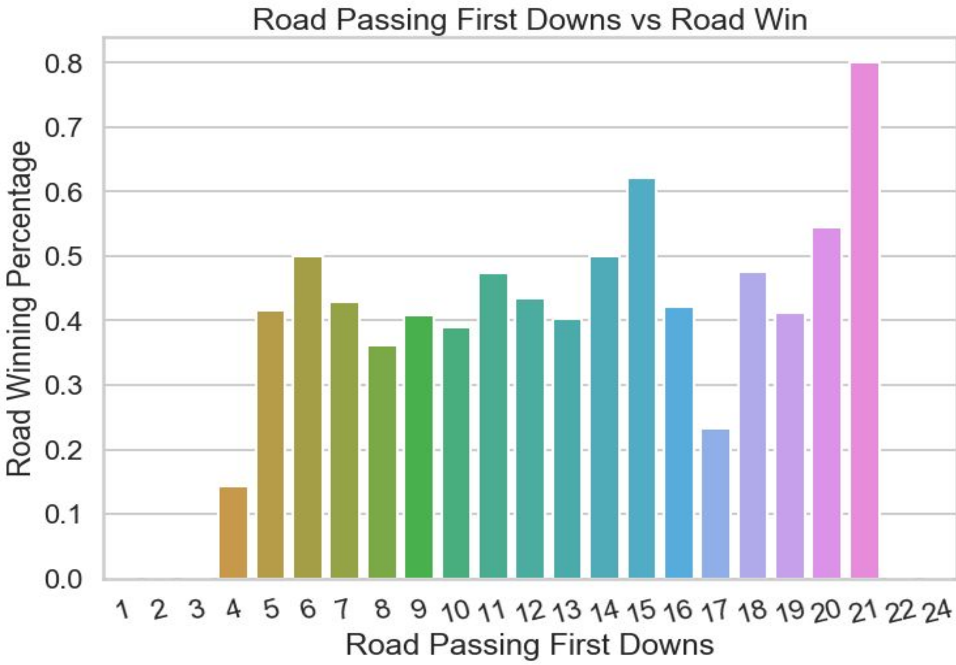
As we saw in the previous plot, it is clear running the football is key. But it is more than just about scoring. In fact, some of the biggest reasons teams focus on rushing include wearing down the defense as well as controlling field position. Another important reason for rushing the ball is increasing your time of possession.



For Road teams we see another big increase in the medians for time of possession. In losing efforts, the median is around 28 minutes while for winning efforts we see a median closer to about 32 minutes. Although the difference between the two times may not seem too drastic, in such a fast-paced sport every second is crucial plus, with a normal game taking 60 minutes to complete (unless they go to overtime), controlling the ball for over half of the game is huge. Clearly, away teams gain a noticeable advantage when they can control the clock.

Plot #3

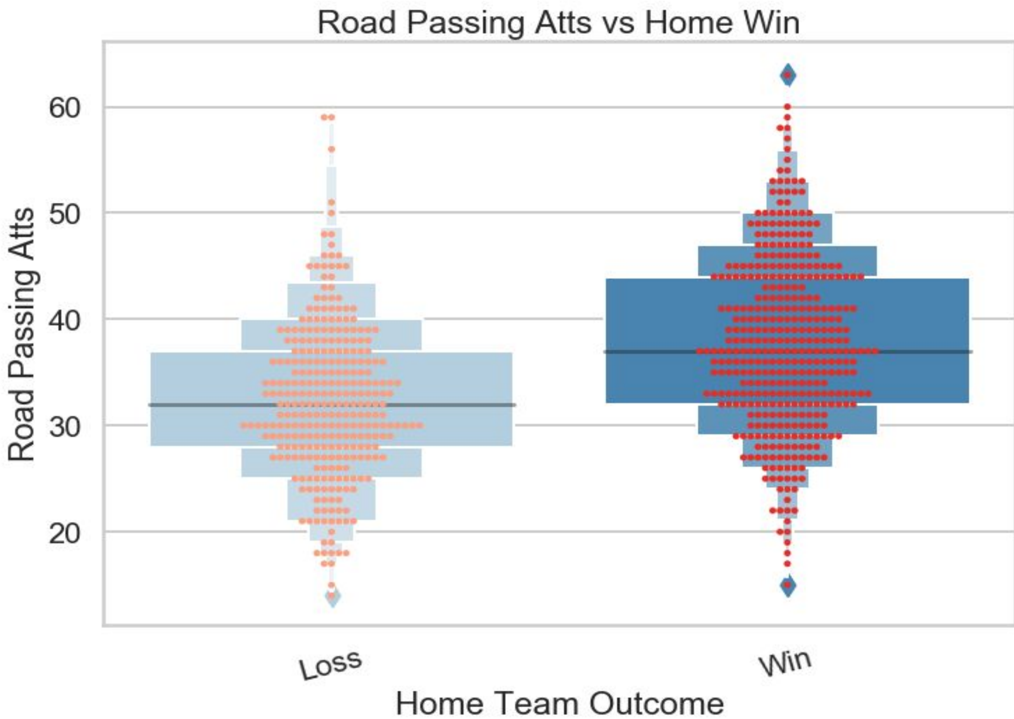
While I have focused a lot on Road teams running the football, that does not mean they should not pass the ball. Most of the top offenses have a good variety of running and passing plays that keep the defense honest. But is there a smart situation where a team should pass? For away teams, passing for First Downs is immensely helpful.



Starting off at low percentages for a low amount of passing First Downs, we notice an upward trend as the Road team passes for more of them. The middle area is consistently around 40 to 50 percent but that is a huge increase especially with how hard it is to win on the road. When you start getting close to 20 passing First Downs, the chances of winning increase dramatically before some surprising dips. While the late drops are likely outliers, we can see, along with the two prior plots, that away teams really do better when they control the football. Passing for First Downs keeps drives going and limits the amount of times the Home team gets the ball. If more Road teams consistently took their time driving down the field, this plot could have even more of an upward trend.

Plot #4

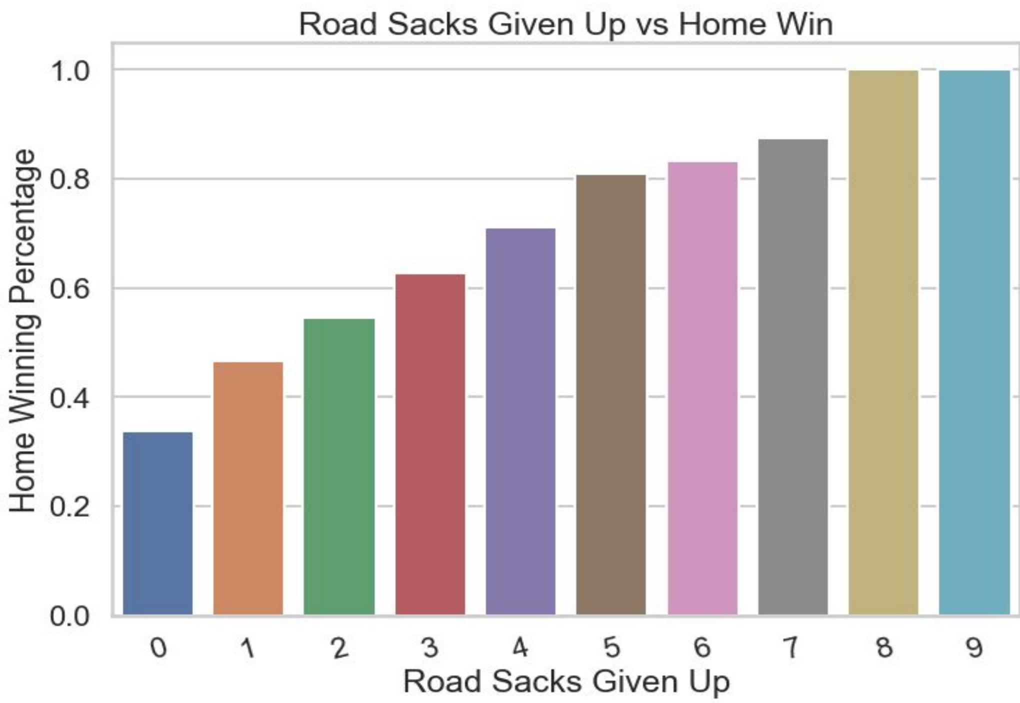
After a lot of talk about the Road teams, it is time to look at what helps Home teams win. While their chances of winning are already higher just by playing at home, there are certain trends that catch the eye. The previous plots focused on Road teams controlling the ball so what do Home teams do in response? As the upcoming plot will show, there is a reason Road teams should focus on running the football.



As a Road team passes the football more, the Home team's chances actually increase with a median that is higher in winning situations. There may be outliers where the Home team actually lost when the away team attempted a lot of passes but, overall, we see a definite advantage for the Home team in these situations. This further strengthens why Road teams need to control the football as well as why the team at home may want more of a high scoring game. If they make the away team pass it more because they are behind, the Home team has a better chance of winning.

Plot #5

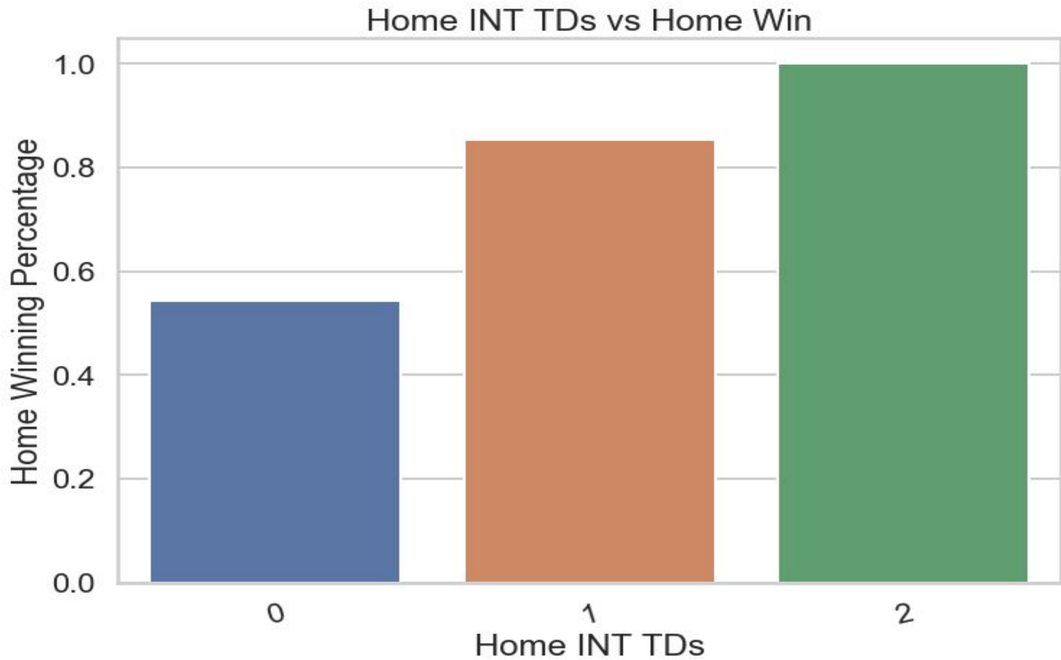
Looking at defenses, we already know basic stats like points and yards are things you do not want to give up. Yet defenses still have a hard time in the NFL today because of how the rules make it easier for the offense. So teams need to know what is important to focus on when coming up with defensive schemes. One aspect that stands out, especially for Home teams, is getting to the opposing team's quarterback.



As we can see if the Home team does not get a single sack, their chances of winning are around 35%. But with each sack they obtain, there is a significant bump in their chances of winning. For the low numbers, we see about a 10% increase each time before it lessens a bit in extreme situations. And although a win percentage of 100% for 8 and 9 sacks may seem odd to some, that would make sense in extreme outliers like those because of their rarity. What is important when it comes to sacks is that Home teams see a consistent upward trend for each sack they record making rushing and getting to the quarterback an important key for their defensive schemes.

Plot #6

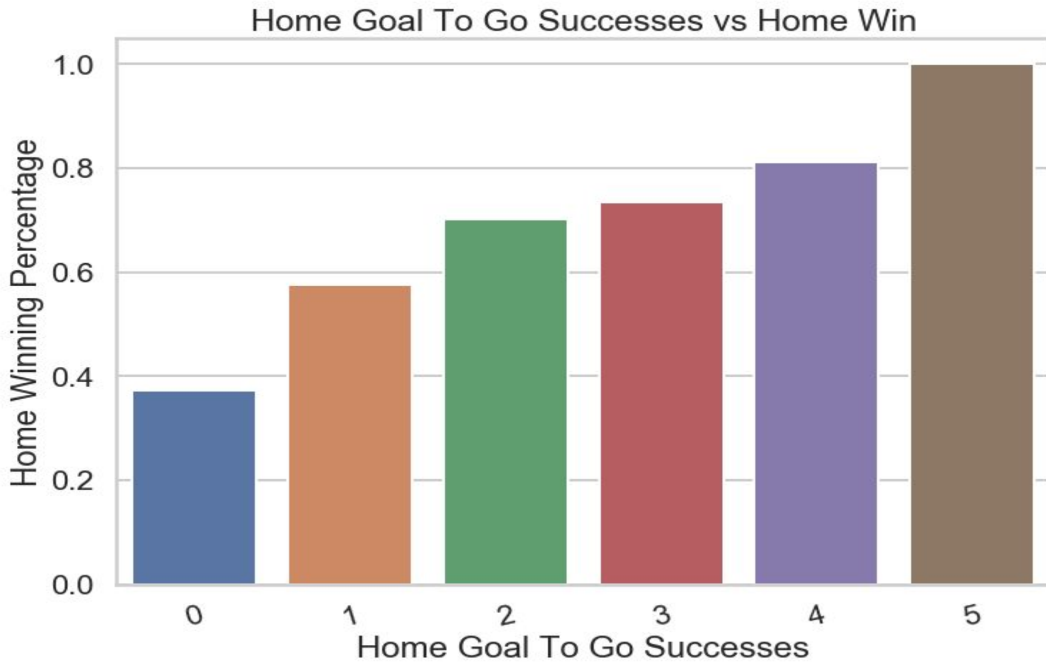
But what happens if you do not get to the quarterback? We saw previously that the more a team attempts to pass the ball, the lesser their chances are of winning. While part of that may be from sacks, another aspect is turning the ball over. Turnovers are always an important factor in an NFL game and, as expected, the more you turn it over, the more likely you are to lose. But if you are at home and you can turn the other team over, what is the best thing to do with the change in possession? The answer: score immediately.



Some may think it is obvious that scoring a touchdown when returning an interception is going to help your team. But I do not think many realize just how beneficial it is. When a Home team returns just 1 interception for a touchdown, their chances of winning increase by around 30%. To have one play increase your chances of winning by that much is stunning. And although 2 of these plays seem to guarantee a victory, those are rare situations and it would make sense that a Home team would win in those cases. What is important to notice is how just one poor play by the Road team can drastically increase the Home team's chance of winning.

Plot #7

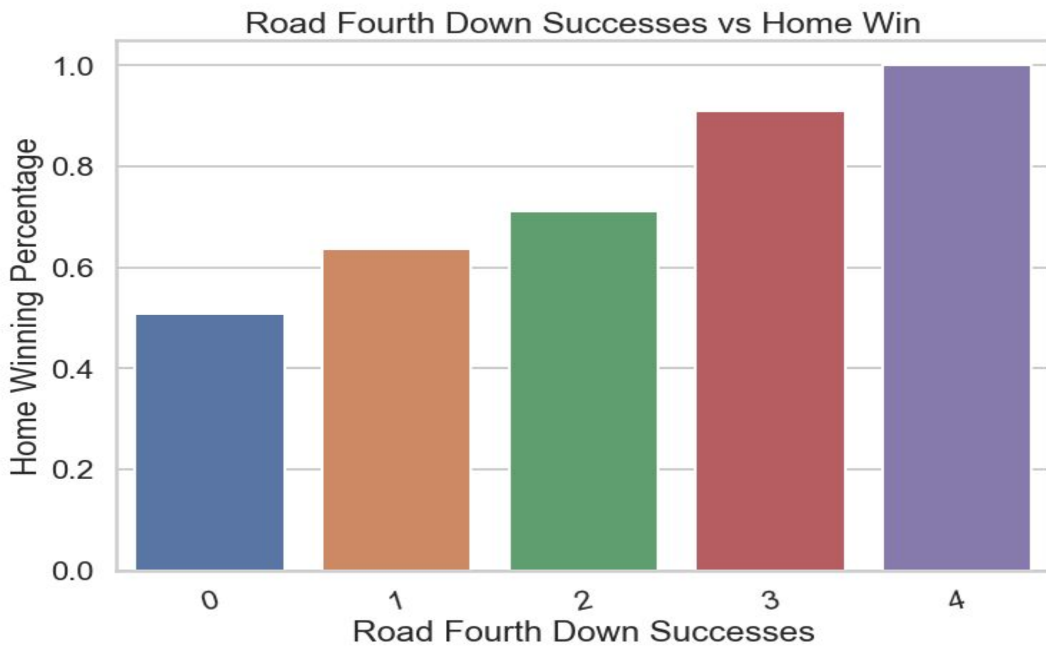
Now that I have talked a lot about the defensive successes for Home teams, it is time to look at an interesting statistic for their offenses. Like I mentioned plenty of times, obvious stats are always going to see better chances of winning (even for away teams) so it is important to look at the more overlooked statistics. There are plenty of ways to score in a game as well as a variety in types of scoring situations. Big plays will always get attention but there is something to be said about finishing when close to the goal line. And for Home teams, the more they succeed in these situations the better their chances of winning become.



In case the vernacular is confusing, a Goal To Go success is when a team has a down where the line to gain is the end zone (like First and Goal, Second and Goal, etc.) and they score. With Home teams there is a big increase just from 0 to 1 successes as a team's win percentage increases from less than 40% to closer to 60%. And after that, it just keeps rising. This is important for away teams to keep in mind as well because when their defense is in this situation, they need to keep the Home team out of the end zone. Otherwise, it is just a consistent upward trend for each success for the Home team.

Plot #8

Finally, to put it all together, it is important to look at how offenses and defenses can be used together to better a team's winning percentage. While the above stats can definitely see aspects of both sides, there was one stat that perfectly encapsulated this collaboration. When teams get to Fourth Down, they have their choices of either going for it, punting it or kicking a Field Goal. Each is used for different situations usually focused on field location and score. But many think being aggressive and going for Fourth Down is the way to go even though failure results in turning the ball over. Yet when going through the data, teams succeeding on Fourth Down is not beneficial especially for teams on the road.



For each Fourth Down success that Road team accumulates, the better the Home team's chances of winning. That means each time the away team either keeps the drive alive or scores in these situations, their opponent is more likely to win. This statistic is very much impacted by both sides of the ball because for the defense it means they are forcing the opposing team to Fourth Downs while, since the Road team is going for it, the offense for the Home team has likely gained a lead which is why their opponent feels the need to take this chance. While many fans like aggressive play calling, in Fourth Down situations the aggressiveness tends to hurt teams especially for those on the road.

Conclusion

If there is something to learn from these specific plots it is a general strategy of how both Road and Home teams should play in an NFL game. Road teams would be best to focus on controlling the ball by running the football, passing when needed and keeping the ball out of their opponent's hands. Home teams, on the other side, likely prefer high scoring affairs where they can force the away team into passing situations. That way they can get after the quarterback, force turnovers as well as make their opponent go for Fourth Down while they take advantage of their own key scoring plays.

Now keep in mind, just because I may have focused on one team's chances of winning in certain plots, it does not mean that the other does not see their chances of winning increase in the same situations. In fact there were similar plots for each team in plenty of situations including statistics that were not utilized here. What I focused on were the ones that seemed most profound and had the biggest impact as well as most exponential increase in winning over time. And as mentioned in the beginning, many of these plots show the averages so the expected result may not happen every time but these are statistics that I want to dive into more to find deeper meaning into how much of an impact they truly have on winning.

4. <u>Inferential Statistical Analysis</u>

When it came to the inferential statistical analysis of the dataset, I not only wanted to look at the previously intriguing plots from the Data Story but also some of the more standard statistics people focus on especially yards and points. For stats like those, I felt t-tests were the best to use because I would be comparing continuous variables (that could go very high) with discrete, binary variables (win or loss). The test that was used the most throughout my analysis was chi-squared which also compared continuous variables with the discrete and binary variables but on a smaller scale. With those tests, I dug deeper with filtering around the median and seeing how that changed the p-values as well as also comparing with a t-test. Lastly, for certain continuous variables, I used a Pearson correlation test to see if there was any predictive power between them. For all tests, the null hypothesis was that the statistic had no impact on a team's chance of winning while the alternative hypothesis was that the statistic had an impact on a team's chance of winning; alpha was set at 0.05. With all of that out of the way, here are some of my findings from the inferential statistics analysis.

For the stats I focused on in the Data Story, many tended to be statistically significant. The t-tests for Road Rushing Attempts vs. Road Win as well as Road Time of Possession vs. Road Win each had very low p-values. Others such as Road Passing Attempts vs. Home Win, Road Sacks Given Up vs. Home Win and Home Goal To Go Successes vs. Home Win also showed low p-values in all the tests they underwent including t-tests, chi-squared and chi-squared filtered above and below the median. All therefore were statistically significant and had impacts on a team's chance of winning.

However, Road Passing First Downs vs. Road Win was proven not to be statistically significant as all tests yielded a high p-value meaning passing for First Downs did not have an impact on the away team's chance of winning. When it came to Home INT TDs vs. Home Win as well as Road Fourth Down Successes vs. Home Win, both proved statistically significant with low p-values when it came to t-tests and standard chi-squared but differed when they were filtered by medians. They still showed low p-values above the median but yielded exceedingly high p-values below the median. This may be that because of how infrequently these events occur. With many data points being around 0 or 1 below the median, these statistics would not be significant and therefore not have an impact on the Home team winning. For both of those cases, they may be statistically significant overall but are not quite as impactful as other stats.

One of the more interesting things I found when it came to standard, popular statistics is although rushing and total yards did impact their individual teams' chances of winning, passing yards did not. The t-tests yielded high p-values, which really challenges this idea of how the NFL has become a big passing league. While passing attempts may have an impact, overall passing yards is not statistically significant or impactful meaning teams may want to avoid completely focusing on their passing game. Along those same lines, the Pearson correlation tests between Road Total Yards and Home Total Yards as well as Road Points and Home

Points resulted in high p-values which means they are not statistically significant. Because they are not predictive of each other, these stats also negate the idea of a shootout.

For turnovers, both interceptions and fumbles are statistically significant and impact either team's chances (both the team giving it away and the team taking it away) of winning overall especially when it came to t-tests and standard chi-squared tests. For interceptions there is also a negative correlation between interceptions thrown from each team with low p-values. This means that they are statistically significant and can be predictive of each other. Seemingly this relationship would strengthen the idea of a sloppy game, however fumbles tell a different story. First off, when the chi-squared test was filtered by medians, the p-values were very high below the median for lost fumbles, which may tie into what happened with Home INT TDs and Road Fourth Down Successes. Because certain games may have zero or one fumble lost, they may not be statistically significant or have an impact on a team winning. Also, the p-value was high for Road Fumbles Lost and Home Fumbles Lost when they were put through a Pearson correlation test, which means they were not statistically significant, predictive or promote the idea of a sloppy game. Interceptions appear to be more consistently impactful on a team winning compared to fumbles.

Overall the inferential statistics told a mixed story about what stats are significant and which are not. Most of the ones that were focused on in the Data Story did appear to impact the chances of a team winning outside of Road Passing First Downs as well as the below the medians of Home INT TDs and Road Fourth Down Successes. The thing that stood out the most to me, however, was how passing yards, for either team, was not statistically significant. As everyone talks about how focused the NFL has become on the passing game, it actually appears that running the ball has more of an impact on a team winning the game. And couple that with the lack of significance for correlations between Total Yards as well as Total Points and this image of shootout NFL games seems less clear. When it comes to turnovers, both interceptions and fumbles are statistically significant but the former shows more consistency along with more predictive power between away and home teams. This leads more to the idea of a sloppy game, which could be a more likely outcome when teams get pass-happy. Couple this with the Data Story and winning in the NFL looks to be a lot less about the high-powered passing games that many assume about the league.

5. Machine Learning Analysis

When it came to the machine learning aspect of this project, my initial focus was on a Logistic Regression model. Although many of these statistics deal with continuous variables, the results are essentially binary; teams either win or lose (I did not really consider teams tying because of the rarity of its occurrence). And since the clubs were split between Road and Home teams, I essentially did the same model twice with each focusing on one of the types of teams. Overall it was a very effective model for both types but did come with some issues especially high log losses. For comparison, I also put the data through a Random Forest model to further study the impact of these features. While these models also came with high log losses, they showed high amounts of accuracy and were very informative about the predictor variables in the data.

For the Logistic Regression model, I set the variable “X” to be all of the features in the data that were continuous numbers, meaning I left out data such as wins, losses, week number, etc. I also decided to leave out points scored for either team because that is too predictive; clearly the more points a team scores the more likely they are to win. The “y” variable was either Road Win or Home Win depending on which version of the model I was using. The data was then split into training and testing sets with the test size being about 20% of the data. In order to avoid overfitting, I put the training data through a GridSearchCV with 5-fold cross validation that included the Logistic Regression model as well as parameter “C” that used the logspace method from NumPy. From there, I fitted the training data and was able to find the best score and best parameter. Lastly, I used the predict method on the testing data for “X” and got the accuracy score from that prediction compared with the testing data for “y.”

For Road Win, I was able to get an accuracy score and AUC score around 96%, which is very high. After running a Confusion Matrix, 133 out of 138 labels were correctly classified and a Classification Report showed the average precision and recall were both at 96%. When it came to the Home Win model, the accuracy score and AUC score were around 91% which is also high, but not quite as good as the Road Win model. For the Confusion Matrix and Classification Report, 126 out of 138 labels were correctly classified along with average precision and recall equal to the accuracy score. While the Home Win model was not as strong, it still showed very high accuracy when it came to predicting winning.

To dig deeper, I wanted to see the features that had the highest predictive power. In order to do that, I created another Logistic Regression model for each team with C set to the best parameter for each model. After fitting the training data, I found their coefficients and set up a pandas DataFrame that sorted each feature by its impact on the model. When it came to the values, the strongest could be either positive or negative; it was about the overall impact on the model. Each showed essentially the same values just with the negative ones being positive for the other team and vice versa. The most common types of statistics at the top were scoring plays (like touchdowns and field goals) and turnovers while statistics near the bottom included all types of yards (rushing, passing and total) as well as time of possession and special teams’ plays. Much of this was as expected but really emphasized how unimportant yards are. According to this model, winning is more about scoring and turnovers then just racking up as many yards as possible no matter which type of offense a team is running. And based on these features, it makes sense which games were mislabeled as they tended to have higher amounts of touchdowns and field goals made.

As I mentioned earlier, a Random Forest model was also used and took less time to set up. Because of its nature, I did not need to run GridSearchCV but set my number of estimators to 1,000 and fit the training data. For the Road team version, I received an accuracy score as well as AUC score around 92%, 127 out of 138 labels correctly classified and the same average precision and recall as the accuracy score. For the Home team model, the accuracy score was around 86%, AUC score around 85%, 118 out of 138 labels correctly classified as well as an average precision around 85% and an average recall around 86%. So, yet again, the Home team model was weaker than the Road team model. And while both showed significantly high accuracy, the Random Forest models were not as strong as those from Logistic Regression.

Feature importance was much simpler to display this time around and also all of the values were positive. While the two teams had similar results, they were a bit different from the features highlighted by Logistic Regression. Some of the more common, stronger features were kickoffs, rushing attempts, total touchdowns and time of possession while weaker features included scores off turnovers, safeties and special teams' scores. This model seemed to emphasize scoring but in a slightly different manner. While touchdowns were important, kicking the ball off was important too which has to do with the amount of times a team scores plus rushing attempts and time of possession emphasize controlling the ball. The weaker features were more about infrequent scoring plays that came from outside of the offense. Once again, the importance of these features led to the mislabeled data as higher amounts of kickoffs and rushing attempts were common in the incorrect labeling. This of course was very different from the Logistic Regression features that focused more on all scoring plays and turnovers with less focus on how a team went down the field. However, both showed less inclination towards the impact of special teams.

Between the two models, there is a very high amount of accuracy, which is a good thing for many situations. However, with large log losses, you can be punished greatly for being confident and wrong. And as seen by the features, each model focused on different predictive variables, which is part of what led to different scores. There is definitely more preprocessing that could be done to help fix some of these issues but overall, because of their high accuracy scores, both models would do a good job of taking the statistics inputted and predicting whether the Road team or Home team would win, which was the main focus of the project. And together they indicate the different ways a team wins with one more focused on overall scoring and avoiding turnovers while the other is more about how a team scores along with added preference placed on running as well as controlling the football. They both perfectly symbolize the different coaching styles seen throughout the NFL.

6. Conclusion

After all of the data wrangling, visualization, statistical analysis and modeling, I have found some trends in the NFL that are important to winning. The one thing that stands out the most is how unimportant passing is as everything showed it had little impact on winning. In the plots, teams did worse when they passed more while the inferential statistics showed high p-values for Road and Home Passing Yards. Also, the relationship between Road and Home Total Yards was not statistically significant. Finally, even the machine learning models placed little importance on passing the ball except for touchdowns which makes sense since points have the biggest impact on a team winning.

When it came to most other statistics (especially defense), there were some interesting trends. The plots showed more success when racking up sack yards as well as scoring off turnovers while the inferential statistical analysis found turnovers to be statistically significant especially the relationship with INTs Given Up for each team. Yet, the two types of models placed different impacts on defense. The Logistic Regression ones placed a lot of importance on scoring off turnovers while the Random Forest models displayed little impact from turnovers. And for special teams, their impact was as expected both in plotting as well as machine learning.

Through all of these statistical methods, I was able to create two models for each team that all had high accuracy scores but also high log losses. While I think the accuracy is what is most important for this specific project, it leads into the flaws and what needs to be done next in order to take this project to the next level. When I started this, the plan was to make the project more about each individual team because of how different their personnel may be. But through the data wrangling process with the Stattleship API, I could really only split it between Road and Home teams. Although these were still helpful, and important, when it came to analyzing the data, my original plan was for this project to be more of a team-by-team analysis. Therefore some of the specialized features I mentioned in the proposal did not end up being created.

Another aspect that could help bring this project even deeper would be looking at more than just the past three seasons. While that kept the rules similar, there could be more to how these statistics impacted winning over a longer stretch of time. Plus, there could be analysis on how certain rules changed the game as well as how teams evolved over time. Also, with a more robust data source, I could look at more situational decisions. Right now, the analysis is more of an overall look at how statistics impact winning in a general scale. On average, these should help you win but in certain conditions these averages as well as impacts might not be as strong. Conditions like field position, score, time, weather, injuries or down number are all great examples of situations that could impact some of these stats I have focused on. Lastly, there are even more statistics that could be used moving forward. Although I doubt their impact on winning would be that strong, they could still be looked into.

Unfortunately I did have issues with the API that did not bring the full dataset I had expected. The process took a very long time and should have reached the expected number of games (768) but despite all of my best efforts, the highest amount of games I could obtain was 686. As frustrating as this was, that is still a large enough dataset to make real analysis on. The missing games are unlikely to cause too much of a difference to the analysis I made since they likely displayed similar trends. However, this flaw is definitely worth noting.

While parts of the project did not go as planned, I still feel there were some really interesting patterns and insights gained from the data. Despite the NFL being known as a passing league, it turns out passing is not very helpful to winning. And for defenses, giving up the fewest points is not necessarily the answer; turnovers and sacks can have bigger impacts for a team. The machine learning models illustrated how much accuracy can be gained by these statistics and their impact on a team winning, even if they focused on different features. No matter which strategy a team focuses on, no statistic will ever guarantee victory. But through all of the analysis, we can see which aspects of the game can truly give your team the best chance to win.