Ben Onderick

10-10-2023

Data Wrangling Final Project

NFL Data Project

1. **Introduction:**
   Being a successful team in the National Football League (NFL) is a hard feat to achieve. With the level of talent that exists in the NFL, making the playoffs and let alone winning a game each week is a hard task alone. For this project, I want to explore the NFL and the playoffs in the last decade. My analysis will include overviews of teams over the years and looking at the statistics of these teams. I hope to be able to use these high-level statistics, to determine which team was the most successful during the decade, and why they were and how they got there.

2. **Data:**
   My data is comprised from two sources. The first is a dataset from Kaggle containing statistics for each team in a season. The second is playoff information for each season from Wikipedia.

   The first dataset from Kaggle contains a whole bunch of statistics related to each team in each season. The statistics focus on offensive statistics for each team in each season, some defensive statistics and penalty and turnover statistics. The data didn't need a whole lot of cleaning, I just added a location column, along with the conference and division.

   For my second source, I scraped some data about the playoffs for each season from Wikipedia. The data I collected shows the top regular season seed for each conference, the playoff champion for each conference and the super bowl winner. This data came with glossary markers for teams that were in the Wikipedia table. For example, some of the team names in the column would look like "Dallas Cowboys[ag]". I removed these markers.

   When I merged the two datasets, I merged them by year. I made it more applicable for an analysis by turning the playoff information into a binary column. If a team was an AFC champion or an NFC top seed for example, there is a 1 in the column indicating that. If not, there is a 0. This makes it so that we can count the number of instances that teams, divisions, locations, etc. have achieved certain playoff feats. This concluded the cleaning and transforming of my data.

   Data Dictionary

Column	**Type**	**Source**	**Description**
year	Date	Kaggle	Year of season
team_code	Text	Kaggle	Unique identifier for each team
team	Text	Kaggle	Team location and name
team_name	Text	Kaggle	Team name
location	Text	User added	Location of team
conference	Text	User added	Conference of team
division	Team	User added	Division of team
wins	Numeric	Kaggle	Number of wins in the season
losses	Numeric	Kaggle	Number of losses in the season
points_for	Numeric	Kaggle	Points scored in the season
yards	Numeric	Kaggle	Yards gained in the season
plays	Numeric	Kaggle	Plays ran in the season
yards_per_play	Numeric	Kaggle	Yards per play in the season
turnovers	Numeric	Kaggle	Number of turnovers in the season
passing_yards	Numeric	Kaggle	Passing yards accumulated in the season
passing_td	Numeric	Kaggle	Amount of passing touchdowns scored in the season
rushing_yards	Numeric	Kaggle	Rushing yards accumulated in the season
rushing_td	Numeric	Kaggle	Amount of rushing touchdowns scored in the season
penalties	Numeric	Kaggle	Number of penalties in the season
penalty_yards	Numeric	Kaggle	Penalty yards accumulated in the season
afc_top_seed	Binary	Wikipedia	Top playoff seed in the AFC

			conference after the regular season
nfc_top_seed	Binary	Wikipedia	Top playoff seed in the NFC conference after the regular season
afc_champion	Binary	Wikipedia	The team who won the AFC side of the playoff bracket
nfc_champion	Binary	Wikipedia	The team who won the NFC side of the playoff bracket
super_bowl_champion	Binary	Wikipedia	Team that wins the Superbowl

3. **Analysis**

3.1 Postseason summary statistics

For my first analysis, I wanted to first get an idea of who were winners and who were losers during the decade that my data is from. The first indicator, which is the most important one, super bowl wins, is the main way that success is determined for teams. The whole purpose of playing 18 regular season games and going to the playoffs is to win the big one. To do this, I started by taking my final data frame, and refining it down to the five playoff indicator columns. In this new data frame, I kept the team name, the conference, and then created 5 new columns, called the same thing, but just as a total of the number of times the team achieved said column. I then created a graph to easily display this, figure 1 down below.

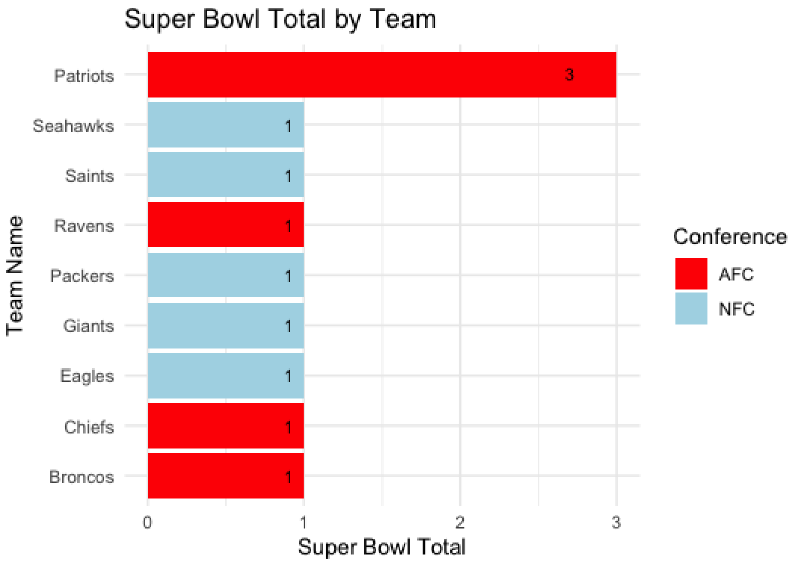


Figure 1 Super Bowl Wins

Clearly, the Patriots are the leading team with 3, and all other teams behind them have 1. I also added in the conference as the color, to see if a particular conference was dominating the super bowl, but it seems to be an almost even split with the NFC having 5 and the AFC having 4. So, we can see the super bowl wins, but what is the super bowl record for these teams. The bottom graphs show the number of times that teams have been champions of their respective conference, meaning they play in the Superbowl. Based on the charts below, we can get a pretty good idea of who were the good teams, and who maybe a few one hit wonder or lucky seasons. The Patriots are the clear leader in the club house, going to 5 Super Bowls and winning 3 of them. Some teams were not so lucky, teams like the 49ers and the Steelers, who made it to the Super Bowl, but didn’t win any. This analysis has really sparked my interest in the Patriots, my next analysis will look at what statistics are the most important for winning.

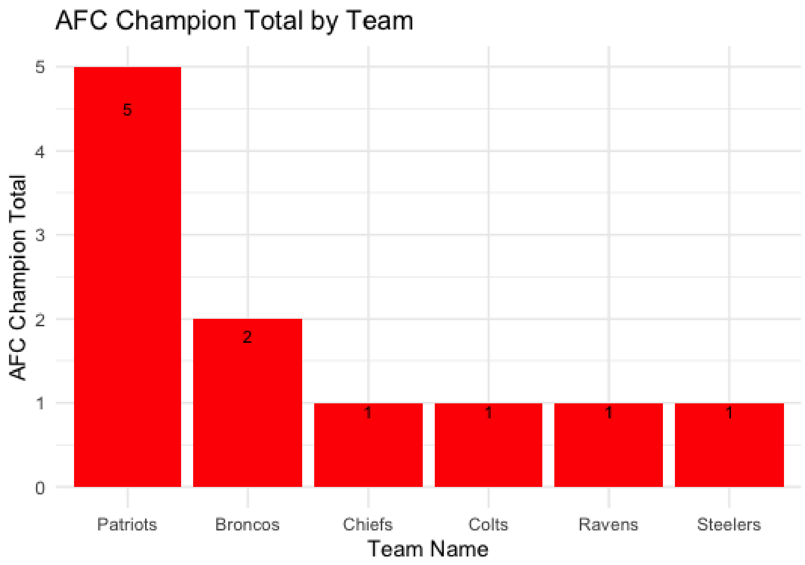


Figure 2 AFC Championship Appearances

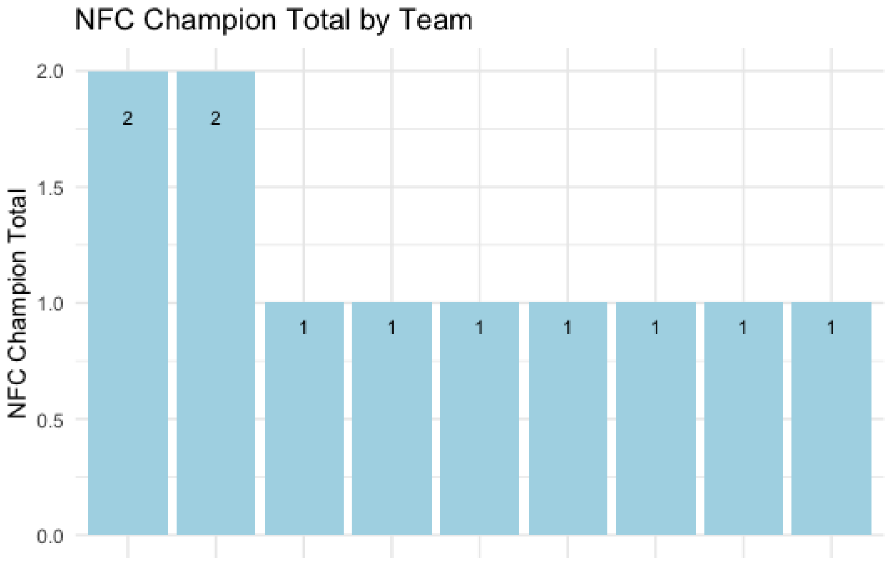
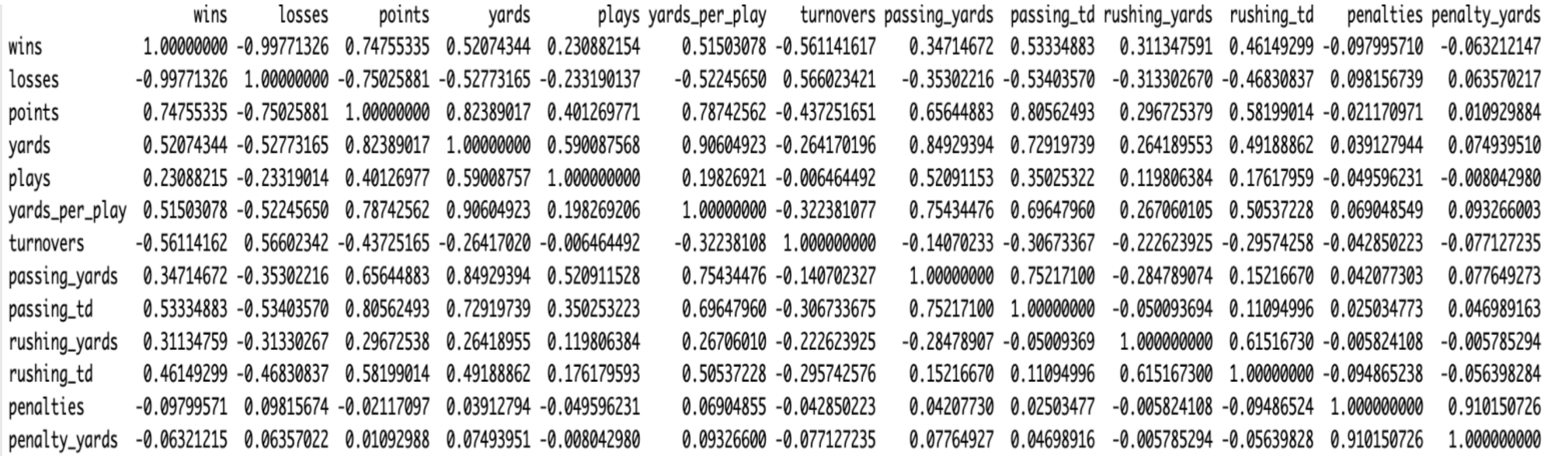


Figure 3 NFC Championship Appearances

3.2 Correlation between statistics

For the next piece of analysis, I want to see what stats work well together. To lead to success, what stats mean the most for wins? What statistics can affect a team negatively and hinder their success? To do this, I did a basic correlation, just standard to see if anything sticks out to me.



```
                      wins      losses      points       yards        plays yards_per_play    turnovers passing_yards  passing_td rushing_yards  rushing_td    penalties penalty_yards
wins            1.00000000 -0.99771326  0.74755335  0.52074344  0.230882154     0.51503078 -0.561141617    0.34714672  0.53334883   0.311347591  0.46149299 -0.097995710  -0.063212147
losses         -0.99771326  1.00000000 -0.75025881 -0.52773165 -0.233190137    -0.52245650  0.566023421   -0.35302216 -0.53403570  -0.313302670 -0.46830837  0.098156739   0.063570217
points          0.74755335 -0.75025881  1.00000000  0.82389017  0.401269771     0.78742562 -0.437251651    0.65644883  0.80562493   0.296725379  0.58199014 -0.021170971   0.010929884
yards           0.52074344 -0.52773165  0.82389017  1.00000000  0.590087568     0.90604923 -0.264170196    0.84929394  0.72919739   0.264189553  0.49188862  0.039127944   0.074939510
plays           0.23088215 -0.23319014  0.40126977  0.59008757  1.000000000     0.19826921 -0.006464492    0.52091153  0.35025322   0.119806384  0.17617959 -0.049596231  -0.008042980
yards_per_play  0.51503078 -0.52245650  0.78742562  0.90604923  0.198269206     1.00000000 -0.322381077    0.75434476  0.69647960   0.267060105  0.50537228  0.069048549   0.093266003
turnovers      -0.56114162  0.56602342 -0.43725165 -0.26417020 -0.006464492    -0.32238108  1.000000000   -0.14070233 -0.30673367  -0.222623925 -0.29574258 -0.042850223  -0.077127235
passing_yards   0.34714672 -0.35302216  0.65644883  0.84929394  0.520911528     0.75434476 -0.140702327    1.00000000  0.75217100  -0.284789074  0.15216670  0.042077303   0.077649273
passing_td      0.53334883 -0.53403570  0.80562493  0.72919739  0.350253223     0.69647960 -0.306733675    0.75217100  1.00000000  -0.050093694  0.11094996  0.025034773   0.046989163
rushing_yards   0.31134759 -0.31330267  0.29672538  0.26418955  0.119806384     0.26706010 -0.222623925   -0.28478907 -0.05009369   1.000000000  0.61516730 -0.005824108  -0.005785294
rushing_td      0.46149299 -0.46830837  0.58199014  0.49188862  0.176179593     0.50537228 -0.295742576    0.15216670  0.11094996   0.615167300  1.00000000 -0.094865238  -0.056398284
penalties      -0.09799571  0.09815674 -0.02117097  0.03912794 -0.049596231     0.06904855 -0.042850223    0.04207730  0.02503477  -0.005824108 -0.09486524  1.000000000   0.910150726
penalty_yards  -0.06321215  0.06357022  0.01092988  0.07493951 -0.008042980     0.09326600 -0.077127235    0.07764927  0.04698916  -0.005785294 -0.05639828  0.910150726   1.000000000
```

There is some good information here, but first we can get rid of some of the obvious ones that will be irrelevant. We can first ignore all of the dark blue diagonal line down the center of the chart. This is just when every statistic is matched up with each other, for example, wins and wins, which is equal to one. We can also ignore the two dark red boxes near the top of the chart showing wins and losses. We can ignore this because there two fields are exact opposites of each other, you can't have both in one game. Now that we have the useless and obvious correlations out of the way, lets get into the interesting statistics. The highest correlations. For this, I made a heatmap, so it is easier to tell which correlations are high or low, rather than searching the table. You can see this in figure 4 below.

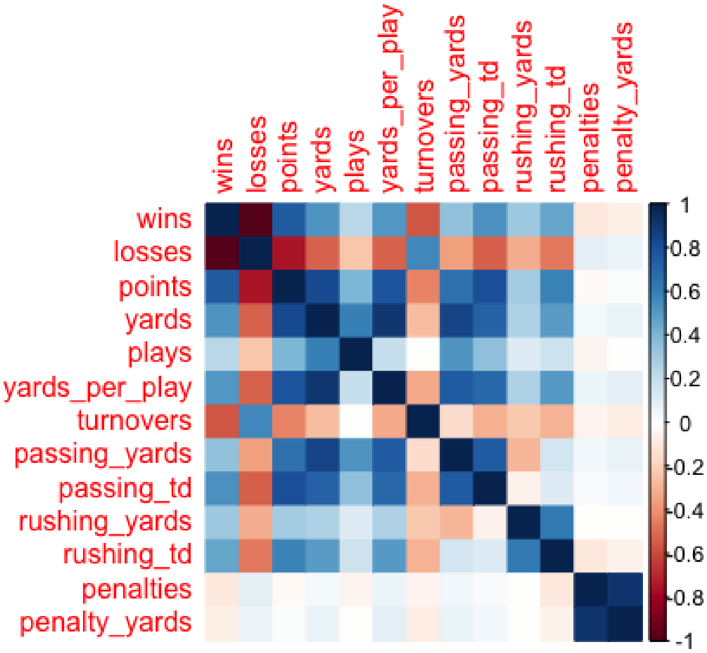


Figure 4 Correlations

Based on figure 4 above, we can tell a few interesting things about the data and statistics. There are several statistics with relationships here that can prove a team's success. Obviously, stats that work off one another are going to be a high correlation as well. For example, yards per play and yards are correlated because yards per play comes from the total amount of yards. One interesting correlation I found was the correlation between wins and the different touchdowns. It looks like passing touchdowns has a higher correlation to wins than rushing touchdowns, since it is a shade darker. This means that passing touchdowns is slightly more important to getting wins than rushing yards. It is also the same for the other way around with losses. It looks like passing touchdowns have a slightly larger impact on a team's ability to lose than rushing touchdowns do. It is also interesting that the number of yards per play appears to impact a team's ability to win more than both of the touchdowns do. Now that we know the relationship between statistics and what leads to success, let's see the leaders in some of these categories and see if that makes sense for the playoff results for the years.

3.3 Leaders in highly correlated statistics

For my last analysis, I am going to look at the leaders for each of the main offensive statistics, passing and rushing yards, and passing and rushing touchdowns. There clearly seem to be the most important statistics when determining success in the NFL. I want to see if the leaders in their categories are the same ones that lead the playoff statistic charts from earlier in the report. I am going to start with the passing statistics. First, we will look at passing statistics.

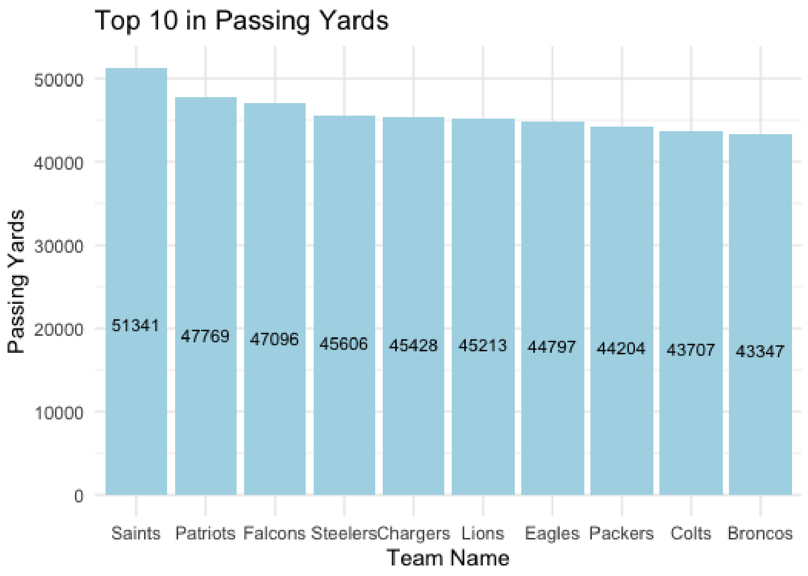


Figure 5 Top 10 in Passing Yards

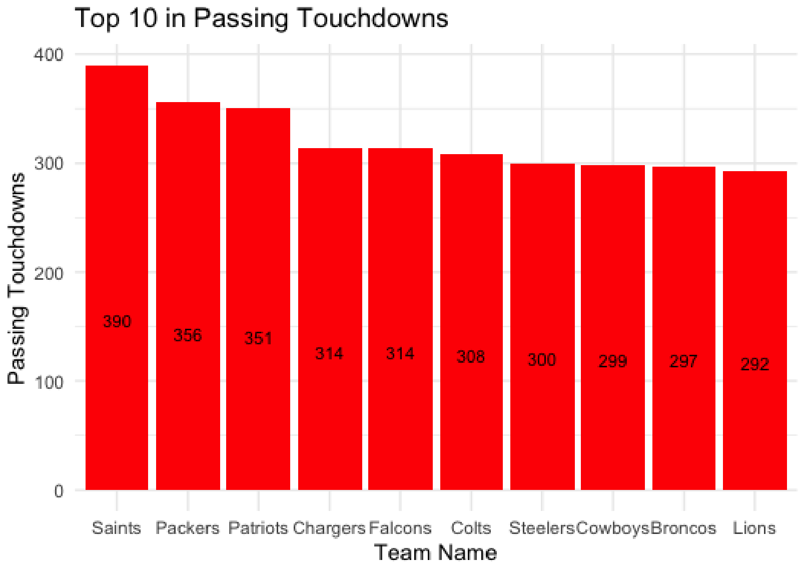


Figure 6 Top 10 Passing Touchdowns

Figure five and figure six contain the top 10 teams in two categories, the number of passing yards and touchdowns for the entire decade. A few things stick out, the Saints are number one in each of the categories. They had 51,314 passing yards and 390 passing touchdowns during this span. The Patriots and Falcons followed for yards, and the Packers and Patriots followed for yards. These two graphs and the trends make sense, the more passing yards means the more touchdowns, and vice versa. The correlation chart about proves this as well.

It is also interesting since both all three of those teams won a super bowl except for the Falcons. Let's see if the trend follows for the rushing statistics.

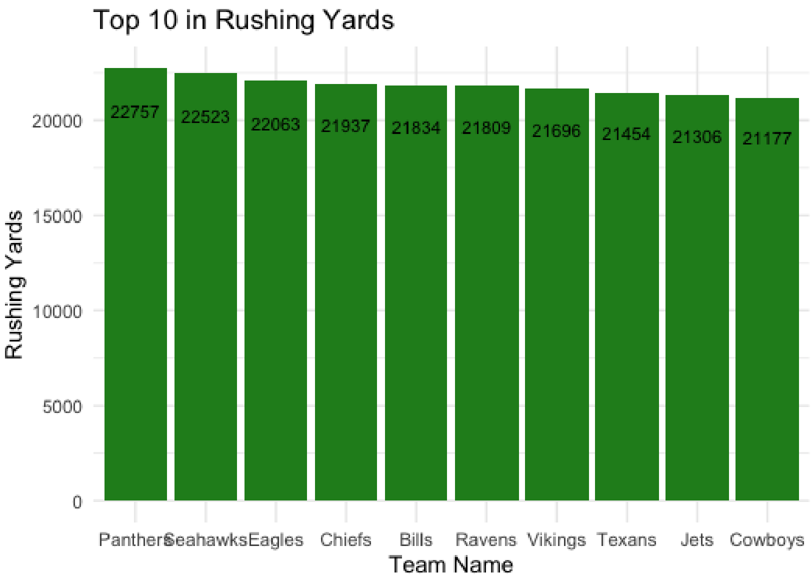


Figure 7 Top 10 Rushing Yards

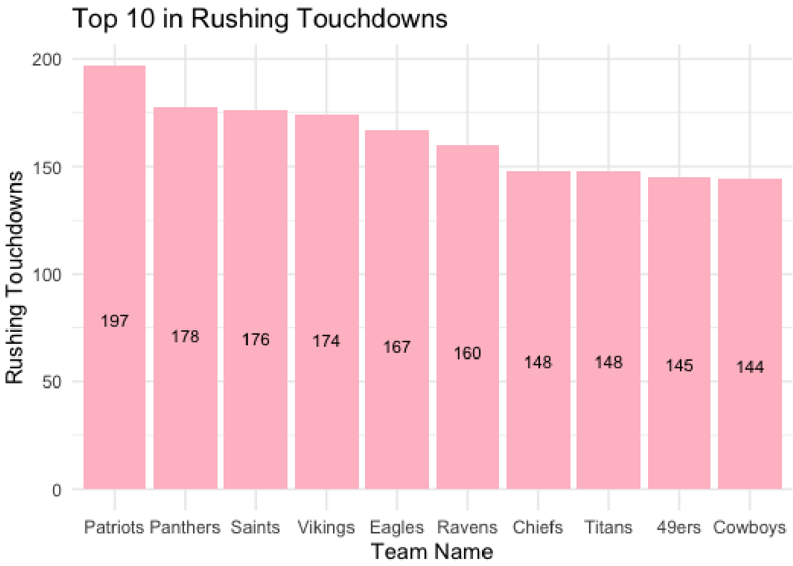


Figure 8 Top 10 Rushing Touchdowns

Figure 7 and figure 8 show the top 10 teams in both rushing yards and rushing touchdowns for the entire decade. We see similar teams near the top of each graph. The patriots lead in both rushing yards and rushing touchdowns. They had 22,757 rushing yards and 197 rushing touchdowns. They beat out the second-place panthers by 19 touchdowns. While there is a split for the top teams here, one team doesn't dominate both categories, Patriots lead the rushing, and the Saints lead the passing. While looking at this we can come to our conclusion on the best team of the decade.

4. Conclusion

After all this analysis, data searching, and graphing, we are able to determine a favorite for the best team in the NFL during the 2010's. I believe that the best team was the New England Patriots, While the statistics were very close, the Patriots were consistently in the top 5 for the statistics and finished there in the decade. The real main reason, which is complimented nicely by the statistics, are the playoff wins. Winning 3 Superbowl's and going there 50% of the time is something no other team came close to doing. There are obviously other factors that I didn't discuss, rosters, schedules, injuries, etc. With the statistics provided and my analysis, the Patriots are clearly the best team of the decade.