NFL Game Outcomes and Predictive Modeling

Nicholas V Romano

nromano@bellarmine.edu

6 February 2024

Exploratory Data Analysis Report

Introduction

The goal of this project is to develop machine learning models to predict the outcomes of NFL Games. An appropriate dataset had to be gathered with that goal in mind. Initially, I was open to the idea of extracting game data from the pro-football-reference page on the Sports-Reference website. However, I reached out to a few people who worked at Sports-Reference to see if I could possibly get permission to use the data that they provided on their site. But I got mixed responses in the answers that I received. With that idea running into a roadblock, I pivoted my approach and started searching for other available sources of data.

In searching for a new source of data for this project, I came across a dataset titled "NFL Team Stats 2002 – Feb. 2023 (ESPN)" on Kaggle that contains team stats for NFL games spanning from 2002-2023. The creator of this dataset pulled the team stats page for each NFL game provided on ESPN's website. Looking at this dataset, it contains similar statistics from NFL games that I was hoping to extract from the pro-football-reference page. The unique thing about this dataset is that it provides stats for each NFL game split by home and away team for each NFL game and contains the score for each game as well. The creator does note a few problems in the dataset: (1) there were a total of 3 games over the span of the dataset that weren't able to be pulled from ESPN; and (2) the statistic of redzone conversions were missing prior to the 2006-2007 NFL season.

Dataset Description

Within the "NFL Team Stats 2002-2023" dataset are team statistics for both the home and away teams in all but 3 NFL games starting in the 2002-2003 NFL season and ending in the 2022-2023 NFL season. The total number of games played are 5641 – it is important to note that this number includes both regular season and playoff games. For each game, the dataset contains a total of 39 columns including information pertaining to: the date on which the game was played, the two teams that were playing, the score of the game, and 17 statistics for both the home and away teams. To differentiate between the 17 statistics for the home and away team, the column names have '_away' or '_home' at the end of the column name to indicate which team statistic that the column belongs to.

Before going into detail and explaining what each column represents it is important to point out that initially there were 12 columns (6 statistics for both the home and away team) that were essentially two statistics in one that were separated by a dash. Instead of having these numerical values being treated as text, the columns were

split into two appropriately name columns that only containing the numerical values and dropping the dash. For example, initially the dataset contained information on the number of pass completions and pass attempts in one column cmp_att_away and cmp_att_home. These two columns were split into four columns passing_completions_away, passing_attempts_away, passing_completions_home, and passing_attempts_away, then the original comp_att_away and cmp_att_home columns were dropped. This correction of columns was performed on the following columns: third_downs_away, third_downs_home, fourth_downs_away, fourth_downs_home, sacks_away, sacks_home, penalties_away, penalties_home, redzone_away, and redzone_home. Furthermore, the dataset was rearranged so that the away team statistics were shown first then the home team statistics were shown. The decision to rearrange the dataset was due to the original dataset have each statistic in order but alternate between the away and home teams. Lastly, a game_result column was added to the end of the dataset with one of three possible values: 0 indicating that the away team won, 1 indicating that the home team won, 2 indicating that there was a tie. These few changes resulted in the dataset going from 39 columns to 52 but maintained the number of rows at 5641.

The following two tables provide insight into the data provided within the dataset. Table 1 provides basic information on all 52 columns, including the column names and a short explanation of the what the data within the column represents. Table 2 provides the more detailed insight into the columns by providing the column name, the data type, variable type, and range of the values within the column, and the percentage of missing values within the column.

Column Name	Description
date	The date on which the corresponding NFL game was played on
away	The away team's name
home	The home team's name
first_downs_away	The number of first downs gained by the away team's offense
third_down_conversions_away	The number of first downs gained on third down by the away team's offense
third_down_opportunities_away	The number of third downs faced by the away team's offense
fourth_down_conversions_away	The number of first downs gained on fourth down by the away team's offense
fourth_down_conversion_attempts_away	The number of fourth downs conversion attempts by the away team's offense
passing_completions_away	The number of offensive passing plays that resulted in a catch for the away team
passing_attempts_away	The number of offensive passing plays ran by the away team
passing_yards_away	The number of yards gained by the away team's offense as a result of passing plays
rushing_attempts_away	The number of offensive rushing plays ran by the away team
rushing_yards_away	The number of yards gained by the away team's offense as a result of running plays
total_yards_away	The total number of yards gained by the away team's offense
times_sacked_away	The number of sacks given up by the away team's offense
sack_yards_lost_away	The number of yards lost by the away team's offense due to sacks
fumbles_away	The number of fumbles lost by the away team's offense
int_away	The number of interceptions thrown by the away team's offense
turnovers_away	The number of turnovers/giveaways given up by the away team's offense
team_penalties_away	The number of penalties committed by the away team
penalty_yards_away	The number of yards lost by the number of penalties committed by the away team
redzone_conversions_away	The number of offensive drives which included at least one offensive play that started between the opposing team's 20-yard line and the opposing team's goal line and resulted in an offensive score for the away team.
redzone_opportunities_away	The number of offensive drives which included at least one offensive play that started between the opposing team's 20-yard line and opposing team's goal line for the away team
def_st_td_away	The number of touchdowns scored by the away team's defense and special team unit
drives_away	The number of offensive drives by the away team
possession_away	The amount of time that the away team spent on offense throughout the game
first_downs_home	The number of first downs gained by the home team's offense
third_down_conversions_home	The number of first downs gained on third down by the away home offense
third_down_opportunities_home	The number of third downs faced by the home team's offense
fourth_down_conversions_home	The number of first downs gained on fourth down by the home team's offense
fourth_down_conversion_attempts_home	The number of fourth downs conversion attempts by the home team's offense
passing_completions_home	The number of offensive passing plays that resulted in a catch for the home team
passing_attempts_home	The number of offensive passing plays ran by the home team
passing_yards_home	The number of yards gained by the home team's offense as a result of passing plays
rushing_attempts_home	The number of offensive rushing plays ran by the home team
rushing_yards_home	The number of yards gained by the home team's offense as a result of running plays
total_yards_home	The total number of yards gained by the home team's offense
times_sacked_home	The number of sacks given up by the home team's offense
sack_yards_lost_home	The number of yards lost by the home team's offense due to sacks
fumbles_home	The number of fumbles lost by the home team's offense
int_home	The number of interceptions thrown by the home team's offense
turnovers_home	The number of turnovers/giveaways given up by the home team's offense
team_penalties_home	The number of penalties committed by the home team
penalty_yards_home	The number of yards lost by the number of penalties committed by the home team
redzone_conversions_home	The number of offensive drives which included at least one offensive play that started between the opposing team's 20-yard line and the opposing team's goal line and resulted in an offensive score for the home team.
redzone_opportunities_home	The number of offensive drives which included at least one offensive play that started between the opposing team's 20-yard line and opposing team's goal line for the home team
def_st_td_home	The number of touchdowns scored by the home team's defense and special team unit
drives_home	The number of offensive drives by the home team
possession_home	The amount of time that the home team spent on offense throughout the game
score_away	The number of points scored by the away team
score_home	The number of points scored by the home team
game_result	The outcome of the game

Table 1: Description of each column in the dataset including the column name and a short description of what the values represent within the column.

Column Name	Data Type	Variable Type	Range Of Values	Percentage Of Missing Data
date	datetime64	interval	2002-09-05 to 2023-02-12	0%
away	object	Categorical	One of the 32 NFL team names. However, t he home and away teams cannot have the s ame value since a team cannot play itself.	0%
home	object	Categorical		0%
first_downs_away	int64	ratio	Between 3 and 37	0%
third_down_conversions_away	int64	ratio	Between 0 and 15	0%
third_down_opportunities_away	int64	ratio	Between 5 and 24	0%
fourth_down_conversions_away	int64	ratio	Between 0 and 6	0%
fourth_down_conversion_attempts_away	int64	ratio	Between 0 and 7	0%
passing_completions_away	int64	ratio	Between 1 and 43	0%
passing_attempts_away	int64	ratio	Between 3 and 67	0%
passing_yards_away	int64	interval	Between -7 and 516	0%
rushing_attempts_away	int64	ratio	Between 6 and 57	0%
rushing_yards_away	int64	interval	Between -18 and 404	0%
total_yards_away	int64	interval	Between 26 and 643	0%
times_sacked_away	int64	ratio	Between 0 and 12	0%
sack_yards_lost_away	int64	interval	Between 0 and 91	0%
fumbles_away	int64	ratio	Between 0 and 5	0%
int_away	int64	ratio	Between 0 and 6	0%
turnovers_away	int64	ratio	Between 0 and 8	0%
team_penalties_away	int64	ratio	Between 0 and 23	0%
penalty_yards_away	int64	interval	Between 0 and 200	0%
redzone_conversions_away	int64	ratio	Between 0 and 6	0%
redzone_opportunities_away	int64	ratio	Between 0 and 15	0%
def_st_td_away	int64	ratio	Between 0 and 6	0%
drives_away	int64	ratio	Between 0 and 26	0%
possession_away	float	ratio	Between 14.8833 and 47.1333	0%
first_downs_home	int64	ratio	Between 3 and 40	0%
third_down_conversions_home	int64	ratio	Between 0 and 15	0%
third_down_opportunities_home	int64	ratio	Between 5 and 24	0%
fourth_down_conversions_home	int64	ratio	Between 0 and 5	0%
fourth_down_conversion_attempts_home	int64	ratio	Between 0 and 6	0%
passing_completions_home	int64	ratio	Between 1 and 47	0%
passing_attempts_home	int64	ratio	Between 7 and 70	0%
passing_yards_home	int64	interval	Between 6 and 522	0%
rushing_attempts_home	int64	ratio	Between 5 and 60	0%
rushing_yards_home	int64	interval	Between -3 and 378	0%
total_yards_home	int64	interval	Between 77 and 653	0%
times_sacked_home	int64	ratio	Between 0 and 11	0%
sack_yards_lost_home	int64	interval	Between 0 and 79	0%
fumbles_home	int64	ratio	Between 0 and 4	0%
int_home	int64	ratio	Between 0 and 6	0%
turnovers_home	int64	ratio	Between 0 and 7	0%
team_penalties_home	int64	ratio	Between 0 and 20	0%
penalty_yards_home	int64	interval	Between 0 and 182	0%
redzone_conversions_home	int64	ratio	Between 0 and 7	0%
redzone_opportunities_home	int64	ratio	Between 0 and 14	0%
def_st_td_home	int64	ratio	Between 0 and 6	0%
drives_home	int64	ratio	Between 0 and 25	0%
possession_home	float	ratio	Between 14.25 and 46.35	0%
score_away	int64	ratio	Between 0 and 59	0%
score_home	int64	ratio	Between 0 and 62	0%
game_result	int64	nominal	0,1,2	0%

Table 2: Description of each column in the dataset including column name, data type, variable type, data range and percent of missing data within that specific column.

Dataset Statistics

One of the more important things to do with a dataset, is to get to know what you are working with. This includes looking into the statistics of the dataset. One of the first things evaluated in terms of the dataset statistics creating a correlation matrix for the columns containing numerical data. The correlation matrix shown in **Figure 1** (page 7) was overlayed with a heatmap using the seaborn library in Python. The darker shade of red indicates that two columns have a stronger correlation between their values. The lighter the shade of red indicates that the two columns have a negative correlation between their values. In the correlation matrix heatmap in **Figure 1,** the maroon diagonal are columns have a correlation of 1 with itself which makes sense. However, when you take a glance at the heatmap you see two large dark square highlighted regions (one in the top left corner, the other in the middle right). This indicates that there are a handful of team statistics that are highly correlated with one another, which is something to take note of when advancing to the modeling stage. It turns out that these columns are the first down, third down, fourth down, and offensive passing, rushing and total yard columns for both the home and away teams. This makes sense that these columns are positively correlated given that the more yards, either it be through passing or rushing, the more third and fourth downs a team is going to make which in turn leads to more first downs. Furthermore, we can see that time of possession is strongly correlated to these values as well since the longer a team's offense is on the field the more plays an offense can run and yards an offense can gain.

After looking creating the correlation matrix, the next step was gaining further insight into the individual columns. This was done by creating a data summary table, and a value count/frequency for each column. The data summary table includes the number of entries within a column, the mean value within a column, the standard deviation of the values of a column, the minimum and maximum value within a column, and the 25-50-75 percentile values within the column. **Table 3: Data Statistics for Numerical Valued Columns** (page 8) is the data summary output for the numerical columns within the dataset. This table provides insight into the range and variability of the various columns within the dataset. The columns that have the largest variability are the yardage statistics. This is because in comparison to a statistic like first downs gained, there are a larger range of values in yards gained that an offense can gain. So, statistics like penalty yards, passing yards, rushing yards, and total yards columns have more variability in them then other columns.

When it comes to the data value count and frequency tables, these tables help gather insight into the bigger picture of each column. These tables give you the change to see how common values are in relation to the rest of the

column. In creating the data value count and frequency tables there were a few tables that stood out from the rest. The first was the home and away team columns. Since this dataset contains game data for both regular season and post season games, not every team is going to be playing the same number of games. A team like the New England Patriots, who year in and year out practically made deep playoff runs, are going to be playing more games than a team like the Detroit Lions or Cleveland Browns, who struggle to make post season appearances on a regular basis.

Another column of note was the redzone conversions for both the home and away teams. For both teams, 0 redzone conversions (redzone scores) are 10% more frequent than the next common number. Sure, there could be many games where either the home team, away team, or both teams failed to score while in the redzone. But in downloading the dataset, there was a cautionary note that ESPN had inaccuracies in record keeping for redzone conversions for the early years of this data. This could be a column that is dropped in the future for fear of inaccurate data. On a similar note, out of the 5641 games, the game result column indicates that 14 games ended with a tie. This corresponds to around 0.2% of all games played during this time. With this in mind, there is the option to drop the 14 values or find a way to better encode ties within the data.

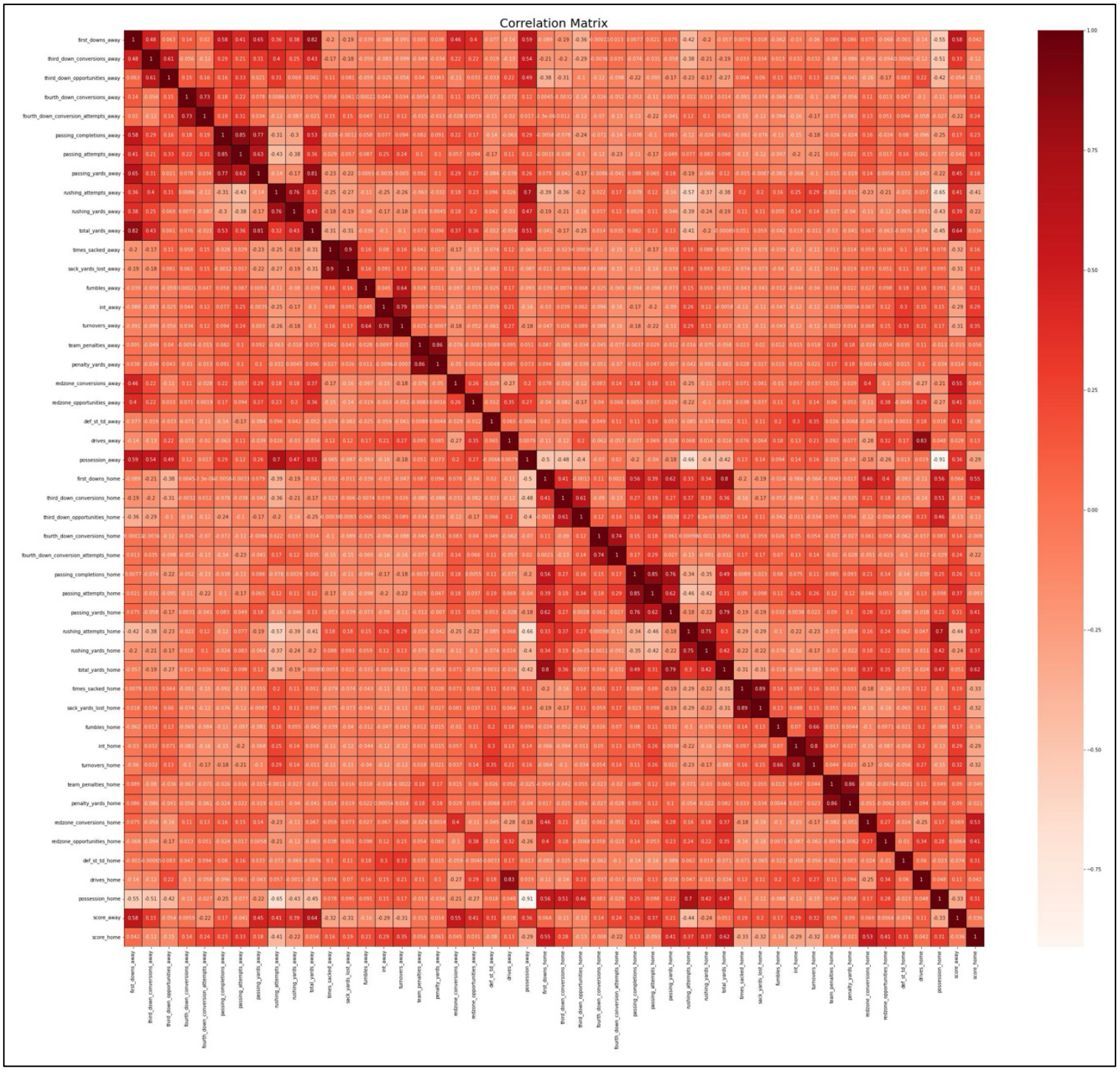

Figure 1: Correlation Matrix Heatmap

Column Name	count	mean	std	min	25%	50%	75%	max
first_downs_away	5641	19.033505	5.072675	3.000000	15.000000	19.000000	22.000000	37.000000
third_down_conversions_away	5641	5.067009	2.253157	0.000000	3.000000	5.000000	7.000000	15.000000
third_down_opportunities_away	5641	13.262010	2.520580	5.000000	12.000000	13.000000	15.000000	24.000000
fourth_down_conversions_away	5641	0.528098	0.770212	0.000000	0.000000	0.000000	1.000000	6.000000
fourth_down_conversion_attempts_away	5641	1.066832	1.117067	0.000000	0.000000	1.000000	2.000000	7.000000
passing_completions_away	5641	20.905690	6.159261	1.000000	17.000000	21.000000	25.000000	43.000000
passing_attempts_away	5641	34.113278	8.523087	3.000000	28.000000	34.000000	40.000000	67.000000
passing_yards_away	5641	221.456834	79.092931	-7.000000	165.000000	218.000000	275.000000	516.00000
rushing_attempts_away	5641	26.608225	7.858818	6.000000	21.000000	26.000000	32.000000	57.000000
rushing_yards_away	5641	110.913313	50.863886	-18.00000	74.000000	104.000000	140.000000	404.00000
total_yards_away	5641	332.370147	86.276865	26.000000	273.000000	332.000000	391.000000	643.00000
times_sacked_away	5641	2.358093	1.741476	0.000000	1.000000	2.000000	3.000000	12.000000
sack_yards_lost_away	5641	15.385038	12.702123	0.000000	6.000000	13.000000	23.000000	91.000000
fumbles_away	5641	0.632512	0.805697	0.000000	0.000000	0.000000	1.000000	5.000000
int_away	5641	0.947882	1.015367	0.000000	0.000000	1.000000	2.000000	6.000000
turnovers_away	5641	1.580394	1.324335	0.000000	1.000000	1.000000	2.000000	8.000000
team_penalties_away	5641	6.483248	2.816764	0.000000	4.000000	6.000000	8.000000	23.000000
penalty_yards_away	5641	54.241978	27.065657	0.000000	35.000000	50.000000	70.000000	200.00000
redzone_conversions_away	5641	1.257756	1.239082	0.000000	0.000000	1.000000	2.000000	6.000000
redzone_opportunities_away	5641	3.246233	1.959111	0.000000	2.000000	3.000000	4.000000	15.000000
def_st_td_away	5641	0.318029	0.771783	0.000000	0.000000	0.000000	0.000000	6.000000
drives_away	5641	12.209537	2.580283	0.000000	11.000000	12.000000	13.000000	26.000000
possession_away	5641	30.013550	4.586094	14.883333	26.800000	29.983333	33.100000	47.133333
first_downs_home	5641	19.968445	4.973625	3.000000	17.000000	20.000000	23.000000	40.000000
third_down_conversions_home	5641	5.236837	2.250594	0.000000	4.000000	5.000000	7.000000	15.000000
third_down_opportunities_home	5641	13.147137	2.577380	5.000000	11.000000	13.000000	15.000000	24.000000
fourth_down_conversions_home	5641	0.493530	0.738267	0.000000	0.000000	0.000000	1.000000	5.000000
fourth_down_conversion_attempts_home	5641	0.980677	1.072950	0.000000	0.000000	1.000000	2.000000	6.000000
passing_completions_home	5641	21.051409	6.094834	1.000000	17.000000	21.000000	25.000000	47.000000
passing_attempts_home	5641	33.783726	8.526244	7.000000	28.000000	33.000000	39.000000	70.000000
passing_yards_home	5641	227.349052	77.613961	6.000000	173.000000	222.000000	277.000000	522.00000
rushing_attempts_home	5641	27.670626	7.840204	5.000000	22.000000	27.000000	33.000000	60.000000
rushing_yards_home	5641	117.976245	52.349052	-3.000000	81.000000	111.000000	149.000000	378.00000
total_yards_home	5641	345.325297	83.446688	77.000000	288.000000	344.000000	401.000000	653.00000
times_sacked_home	5641	2.216274	1.683574	0.000000	1.000000	2.000000	3.000000	11.000000
sack_yards_lost_home	5641	14.437157	12.253151	0.000000	5.000000	12.000000	21.000000	79.000000
fumbles_home	5641	0.630207	0.792989	0.000000	0.000000	0.000000	1.000000	4.000000
int_home	5641	0.891509	0.997302	0.000000	0.000000	1.000000	1.000000	6.000000
turnovers_home	5641	1.521716	1.316938	0.000000	1.000000	1.000000	2.000000	7.000000
team_penalties_home	5641	6.084914	2.687848	0.000000	4.000000	6.000000	8.000000	20.000000
penalty_yards_home	5641	50.916504	25.846237	0.000000	32.000000	48.000000	66.000000	182.00000
redzone_conversions_home	5641	1.411275	1.341498	0.000000	0.000000	1.000000	2.000000	7.000000
redzone_opportunities_home	5641	3.483248	1.954930	0.000000	2.000000	3.000000	5.000000	14.000000
def_st_td_home	5641	0.334692	0.793826	0.000000	0.000000	0.000000	0.000000	6.000000
drives_home	5641	12.145364	2.489866	0.000000	11.000000	12.000000	13.000000	25.000000
possession_home	5641	30.423985	4.548237	14.750000	27.333333	30.466667	33.516667	46.350000
score_away	5641	21.139337	10.072014	0.000000	14.000000	21.000000	28.000000	59.000000
score_home	5641	23.398511	10.270509	0.000000	16.000000	23.000000	30.000000	62.000000

Table 3: Data Statistics For Numerical Valued Columns

Dataset Graphical Exploration

Another way of getting a good grasp of the data at hand is to visualize the data. Creating visualizations for the data assist in picking up trends or corrections that you might not see by staring at numbers. The visualizations created for this dataset include histograms, boxplots, bar charts and line graphs. Histograms and boxplots were used to getting a better grasp at the shape of the data, if the data is skewed one way or the other, or if there are any outliers. Bar charts and line graphs were used in evaluating trends.

When it came to looking at the shape of the data, most of the histograms that were created seemed to indicate that the data within the columns were symmetric. The only columns that contain data that appear skewed are the redzone, turnover, sack, and defensive and special team touchdown statistics. These columns appear to be heavily skewed. Given previous concerns for the redzone statistic columns, this gives more credence to dropping it. On the other hand, since turnovers, sacks, and defensive and special teams touchdowns do not occur in large numbers in comparison to other statistics and are integer based data it makes sense that these graphs are skewed. In terms of the boxplots for the columns, it appeared that almost all of the boxplots had some number of outliers involved. For example, in the offensive passing efficiency boxplots the outliers begin to show around the 450 mark. However, knowing that there have been a handful of quarterbacks that have thrown for over 450 passing yards in a single game these outliers are not a concern from my viewpoint since the number of 450+ passing yard games is relatively small in comparison to the rest of the data.

When it comes to trends in the data, this is where bar graphs and line graphs come in play. For bar graphs, I generated three grids of bar graphs comparing the different team statistics for both the home and away teams by game result. These graphs are Statistics_by_Game_Outcome1.jpeg, Statistics_by_Game_Outcome2.jpeg, and Statistics_by_Game_Outcome3.jpeg files. In creating these graphs, it appears that the winning team had improved stats across the board in comparison to the losing team. So, the home team had better stats in home wins then in home losses. These suggests the possibility of the idea that one could predict the outcome based on home and away team statistics. For line graphs, I generated graphs looking at average year-to-year statistics. The reason for this is that there is the notion that the NFL has become more pass heavy in more recent years. To evaluate this claim I generated a line graph that compared passing and rushing attempts for both the home and away teams. **Figure 2: Average Play Types Ran per Game per Year** is the graph generated. Looking at this graph, one can see a settle positive trend in passing attempts per game and a settle negative trend in rushing attempts over the years. This gives

the notion of an evolving game of football. This may indicate that I might need to look at a shorter time frame than 20 seasons.

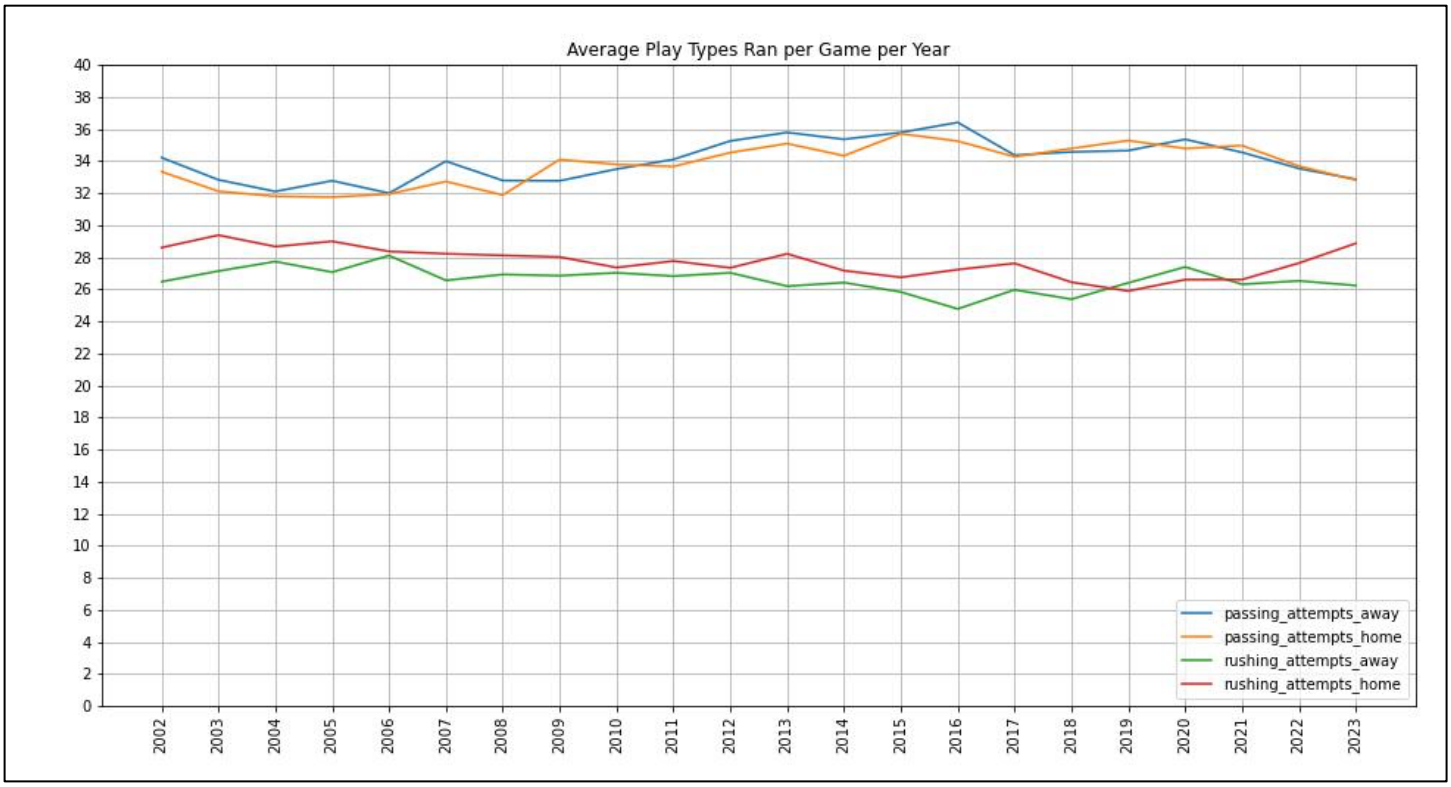

Figure 2: Average Play Types Ran per Game per Year

Summary of Findings

While performing the exploratory data analysis, a few things of note were found, some of which have already been discussed. First is the game_result column, which I personally added to the dataset based on the relationship between the score_away and score_home columns. However, in getting the value count and frequency table for this column, it came to my attention that this dataset had 14 games resulting in a tie. A possible solution is to drop these rows. Another concern was the team representation which is due to the dataset containing both regular season and playoff games. The options are to either keep the postseason playoff games in the dataset and lose a chunk of the dataset, or leave it in. Finally, there is the shift toward pass heavy offenses in recent years. A possible solution would be to look at say 10-12 years then 20 to better represent the move toward pass heavy offenses. These are the things to consider in the next steps of this project.

References

cviaxmiwnptr. NFL Team Stats 2002 - Feb. 2023 (ESPN). *Kaggle*. Retrieved 18. Jan 2024 from https://www.kaggle.com/datasets/cviaxmiwnptr/nfl-team-stats-20022019-espn