NFL Success Predictors: Home-Field Advantage and Playoff Analytics

Max Enabnit, Greg Fagan, Brian Karstens, Christian Kleronomos, Luke Rubin

1. Executive Summary:

In the subsequent pages, our two models predicting (1) whether an NFL team makes the playoffs or (2) predicting if the home or away team wins a game are detailed. Using a dataset with statistics from the last 22 NFL seasons we were able to make 2 classifier models that will accurately predict select outcomes in the 2024-2025 season and beyond. The model that we have created has been designed to be used as an additive model in addition to current data projection metrics to help improve money line and futures predictions line setting for FanDuel Sportsbook. Key findings in the models reveal most offensive stats and some key defensive stats allow for accurate predictions of playoff or winner status prior to games or seasons starting. The only associated costs with this model would be to allocate time to running the models with additional data—on top of your current models—and any associated costs with running these new models, whether it's paying employees or server costs. These additional costs can be counteracted by profits from parlay and prop futures bets.

2. Problem Description:

2.1 Background:

The NFL is a highly competitive league where teams' performances fluctuate each season, influenced by various game-level and team-level factors. Accurately predicting game winners and playoff qualifications is valuable for sports betting and strategic decision-making by teams and analysts. Using historical data and statistical methods provides an opportunity to discover patterns that improve predictions. In the United States, adults were predicted to place $35 billion in NFL bets alone, including $14 billion in money line bets and $1.5 billion in futures bets.

2.2 Business Goal and Data Mining Goal:

Business Goal: To support sports betting strategies by predicting home/away game winners and identifying teams likely to qualify for the playoffs in the 2025-2026 season, allowing you to set lines increasing your profitability.

Data Mining Goal: Create predictive models leveraging historical data to perform supervised classification, determining whether the home or away team will win a game, and to predict whether a team will qualify for the playoffs.

3. Data Description:

3.1 Data:

The dataset we chose comes from https://www.kaggle.com/datasets/cviaxmiwnptr/nfl-team-stats-20022019-espn/data. This data is scraped directly from ESPN, compiled of 61 columns and 5930 rows from the 2002-2003 season through the 2023-2024 season. All the rows have information from the "Team Stats" page directly on ESPN and include games from the regular season and playoffs. We used data starting from the 2006-2007 season through the 2023-2024 season due to structurally missing data in the redzone completion column prior to the 2006-2007 season. Following this reduction, we were left with 61 columns and 4861 rows. For one of our predictive models, we derived new variables by calculating the difference between home and away values for all statistical features. This combined variable simplified identifying predictors for home or away wins.

See appendix items i. and ii. for data dictionaries

3.2 Exploratory Analysis:

To determine which features would be most useful to our models and to understand the relationships between the data, we created a number of graphs and charts. First, we made a bar chart that highlighted the importance of each feature in predicting game outcomes, with passing yards and third-down completions as key predictors (appendix iii.). Next, we created a bar chart that highlighted the correlation between various different game features (appendix iv.). This illustrated that there was data leakage with the winner variable, but that several variables including rushing attempts, redzone completions, and first downs positively correlated with winning games and variables such as interceptions, fourth down attempts, and sacks were negative correlated with wins, leading on average to more losses with higher counts. After making our correlation graph, we created a number of box plots. These illustrated that first downs (appendix v.), a higher number of home penalty yards (appendix vi.), and yardage totals (appendix vii.) were good indicators of whether or not a team would make the playoffs and had a positive correlation with teams making the playoffs. Number of sacks (times sacked on offense) (appendix viii.) was also a good indicator of whether or not a team would make the playoffs with a negative correlation of teams making the playoffs. Next, we created bar charts for win rate % (appendix ix.) and average points scored (appendix x.) for home and away teams. We found that these had correlation with wins but determined that these may have been instances of data leakage and chose to emit the features from our models. We also made a bar chart breakdown of the average number of first downs by each method (passing, rushing, and penalty) for home teams and away teams (appendix xi.). We determined that while the averages were very similar, first downs for each category were a valuable feature to include. Lastly, we created a line chart to examine total points scored per season (appendix xii.). This chart illustrated the fluctuations in points scored per season and that points scored were not a good predictor for our models.

3.3 Data Pre-processing:

When first examining the data and creating the models we made sure to get rid of data leakage variables such as the Winner when we are trying to predict the winning team. We went through every variable making sure there was no leakage along with making sure correlations were not too high but still leaving enough to create a flexible model. Other leakage variables included points_scored for home and away. We then feature engineered variables to create a few different columns in the datasets. These included our Differences columns where we took the home minus away for all the stats to create a uniform variable for all the stats. For predicting playoffs, we aggregated the dataset using python to sort it by team and year and created a column made_playoffs which was Yes or No, we then used that as the target variable in our model. Some other variables we created include Winner because the dataset only included the scores from the teams so to create a target variable we created that column.

4. Data Mining Solution:

4.1 Models:

Model 1 involves taking the differences between the home and away features from the dataset. For example, we took passing yards home minus passing yards away, which gave us the difference between the two. So, a positive differences feature means that the home team had better stats for that feature. This model's data is manually split up instead of using a random sampler. The training set consisted of 2006-2018 data and the testing set consisted of 2019-2023 data. It is important to manually split the data to prevent data from, for example, 2006 appearing in both training and testing sets. We decided to test our target variable, winner, with four different classification models: logistic regression, decision tree, gradient boosting, and random forest.

Model 2 is another classification problem, predicting whether a team would make the playoffs. Since this is a classification task, we used logistic regression, decision trees, gradient boosting, and random forest models. For this model, we included data from 2002 to 2006, which had been excluded in Model 1. However, structurally missing redzone data was excluded, even though it could have been a strong predictor in the finalized model. We also excluded playoff statistics, as they explain playoff performance rather than help predict playoff qualification. Similar to Model 1, we manually split the data into a training set (2002–2018) and a test set (2019–2023) to ensure that data from the same season were not split between the sets. To prepare the dataset for this model, we aggregated team statistics for each season and created a column indicating whether the team made the playoffs based on these stats.

4.2 Performance Evaluation:

Model 1 Performance:

Cross Validation Model:	**AUC:**	**CA:**	**F1:**	**Prec:**	**Recall:**	**MCC:**
Logistic Regression	0.973	0.916	0.926	0.923	0.929	0.829
Decision Tree	0.839	0.826	0.847	0.842	0.851	0.645
Gradient Boosting	0.960	0.893	0.906	0.902	0.909	0.782
Random Forest	0.922	0.841	0.862	0.846	0.878	0.675

Training Error Model:	**AUC:**	**CA:**	**F1:**	**Prec:**	**Recall:**	**MCC:**
Logistic Regression	0.974	0.921	0.930	0.926	0.934	0.838
Decision Tree	0.983	0.936	0.944	0.931	0.957	0.869
Gradient Boosting	0.990	0.946	0.953	0.949	0.957	0.891
Random Forest	0.931	0.846	0.867	0.848	0.886	0.685

Testing Error Model:	**AUC:**	**CA:**	**F1:**	**Prec:**	**Recall:**	**MCC:**
Logistic Regression	0.968	0.900	0.907	0.907	0.907	0.799
Decision Tree	0.834	0.813	0.830	0.809	0.852	0.624
Gradient Boosting	0.955	0.880	0.890	0.875	0.905	0.758
Random Forest	0.910	0.817	0.834	0.809	0.860	0.631

Comparing the cross validation and training tables above, the decision tree and gradient boosting models are potentially overfitting because the training AUC is higher than the cross-validation AUC by a large margin. The random forest model works well, but the logistic regression model is the best fit since the cross validation and training AUC are both high and are about the same.

Model 2 Performance:

Cross Validation Model:	**AUC:**	**CA:**	**F1:**	**Prec:**	**Recall:**	**MCC:**
Logistic Regression	0.885	0.818	0.744	0.787	0.706	0.606
Decision Tree	0.736	0.737	0.636	0.661	0.613	0.432
Gradient Boosting	0.853	0.783	0.699	0.729	0.672	0.531
Random Forest	0.869	0.790	0.669	0.821	0.564	0.543

Training Error Model:	**AUC:**	**CA:**	**F1:**	**Prec:**	**Recall:**	**MCC:**
Logistic Regression	0.905	0.831	0.760	0.811	0.716	0.633
Decision Tree	0.990	0.945	0.926	0.935	0.917	0.882
Gradient Boosting	1.000	0.996	0.995	1.000	0.990	0.992
Random Forest	0.922	0.831	0.736	0.889	0.627	0.637

Testing Error Model:	**AUC:**	**CA:**	**F1:**	**Prec:**	**Recall:**	**MCC:**
Logistic Regression	0.862	0.787	0.734	0.783	0.691	0.561
Decision Tree	0.772	0.756	0.719	0.704	0.735	0.504
Gradient Boosting	0.875	0.787	0.746	0.758	0.735	0.564
Random Forest	0.841	0.775	0.705	0.796	0.632	0.536

Looking at the cross-validation and training error models area under the curve values, logistic regression has the highest value with the smallest margin of overfitting. The decision tree model and gradient boosting models also demonstrate big differences in the AUC values demonstrating overfitting. While random forest overfits, it would likely be a second-choice model as the margin is not as high.

5. Conclusion:

5.1 Recommendations:

FanDuel can improve its betting lines by using home-field advantage insights, as home teams often perform better in key areas like passing yards, third-down conversions, and red zone efficiency. This can help set more accurate lines, especially for close games. FanDuel should also add more prop and parlay bets focused on important factors like penalties, turnovers, and third-down conversions. This would engage users more and create more opportunities for betting. Lastly, FanDuel should hire us so we can work together on creating more accurate models and further increase their profitability.

5.2 Limitations:

The biggest limitation we ran into was the lack of data. The lack of data was mostly when aggregating on a season basis predicting whether a team made the playoffs because there are just limited total amounts of teams that have made the playoffs. As the years go on the model predicting playoffs will continue to get more accurate. Another limitation was making sure we were using the data in the right way to not cause overfitting or data leakage.

5.3 Future Work:

The main idea for improvement for our 2 main models is to gain access to larger databases. The models we have include 23 numerical features and 2 categorical features which is enough to create accurate models on those data points but with more features we can improve it more. We also tried to create a 3rd model in which we could predict the team's winning percentage based on the game averages in features such as yards per game, first downs per game etc. For this we found another dataset that had almost the same variables but was already aggregated by team by season and included win percentage. We imported everything to orange and started to create different models. We found that linear regression gave us an R2 0.66 which was our best model. We did not do anything else with this model as we were continuing to improve our other 2 models but, in the future, having this as another additive model is something we can focus on.

Appendix

Item i.

Model #1 Data Dictionary predicting home or away winning

Attribute:	Type:	Description:
winner	categorical	shows whether the winner of the game is home or away
season	categorical	the season that specific game data is from
first_downs	numerical	the difference between home minus away total first downs
passing_first_downs	numerical	the difference between home minus away passing first downs
rushing_fist_downs	numerical	the difference between home minus away in rushing first downs
penalty_first_downs	numerical	the difference between home minus away first downs gained from penalties
third_down_comp	numerical	the difference between home minus away 3rd down conversions
third_down_att	numerical	the difference between home minus away third down attempts
fourth_down_comp	numerical	the difference between home minus away 4th down conversions
plays	numerical	the difference between home minus away number of plays
yards	numerical	the difference between home minus away amount of total yards
pass_comp	numerical	the difference between home minus away number of passes completed
pass_att	numerical	the difference between home minus away passing attempts
pass_yards	numerical	the difference between home minus away passing yards
sacks	numerical	the difference between home minus away number of sacks

sack_yards	numerical	the difference between home minus away sack yards gained or lost
rush_att	numerical	the difference between home minus away number of rushes attempted
rush_yards	numerical	the difference between home minus away rushing yards
penalties	numerical	the difference between home minus away penalties committed
penalty_yards	numerical	the difference between home minus away penalty yards
redzone_comp	numerical	the difference between home minus away passing completions in the redzone
redzone_att	numerical	the difference between home minus away pass attempts in the redzone
interceptions	numerical	the difference between home minus away interceptions thrown
fumbles	numerical	the difference between home minus away fumbles committed
drives	numerical	the difference between home minus away in number of drives

Item ii.

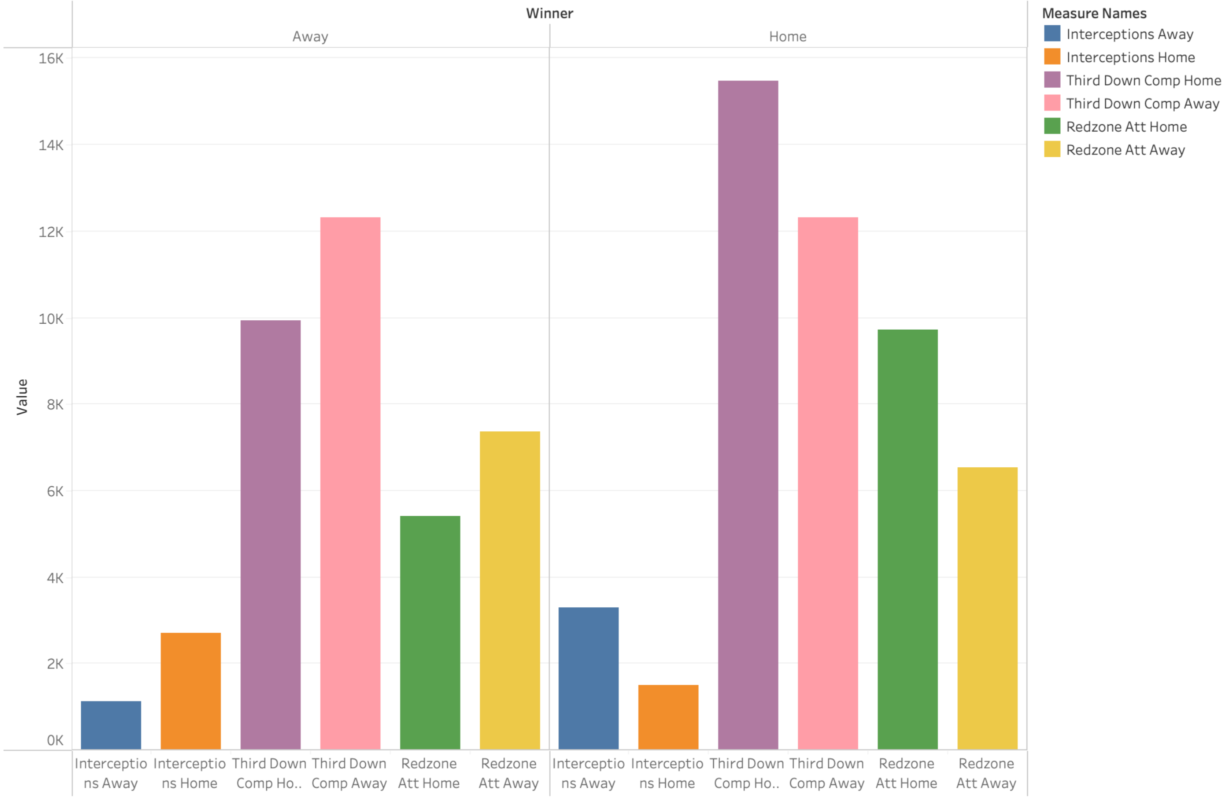
Model #2 Data Dictionary

Attribute:	Type:	Description:
made_playoffs	categorical	yes or no if that team made the playoffs that year or not
season	categorical	what season or year those teams' stats are from
yards_away	numerical	total amount of yards given up
yards_home	numerical	total amount of yards gained
pass_yards_away	numerical	total amount of passing yards given up
pass_yards_home	numerical	total amount of passing yards gained

sack_yards_away	numerical	total amount of sack yards given up
sack_yards_home	numerical	total amount of sack yards gained
rush_yards_away	numerical	total amount of rush yards given up
rush_yards_home	numerical	total amount of rush yards gained
pen_yards_away	numerical	total number of penalty yards assessed against the other team
pen_yards_home	numerical	total number of penalty yards assessed to the team
pen_num_away	numerical	total amount of penalties called against the other team
pen_num_home	numerical	total amount of penalties called against the team
fumbles_away	numerical	total amount of fumbles lost by the other team
fumbles_home	numerical	total amount of fumbles gained by the team
interceptions_away	numerical	total amount of interceptions thrown by the other teams
interceptions_home	numerical	total amount of interceptions thrown

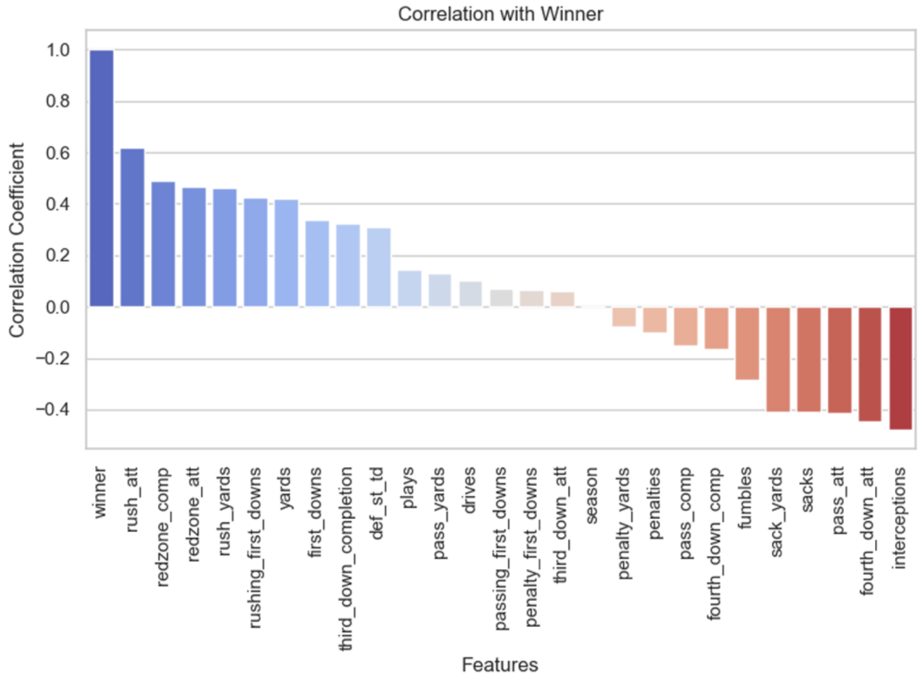
Item iii.

Teams with More Offensive Production and Less Turnovers Win

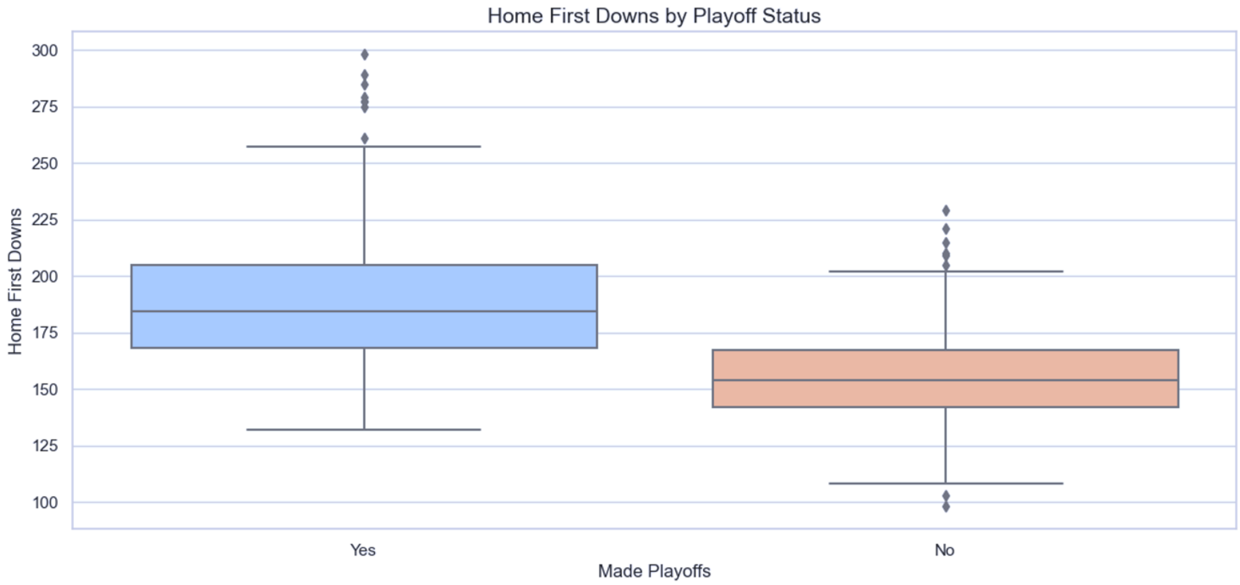


Interceptions Away, Interceptions Home, Third Down Comp Home, Third Down Comp Away, Redzone Att Home and Redzone Att Away for each Winner. Color shows details about Interceptions Away, Interceptions Home, Third Down Comp Home, Third Down Comp Away, Redzone Att Home and Redzone Att Away. The view is filtered on Winner, which keeps Away and Home.

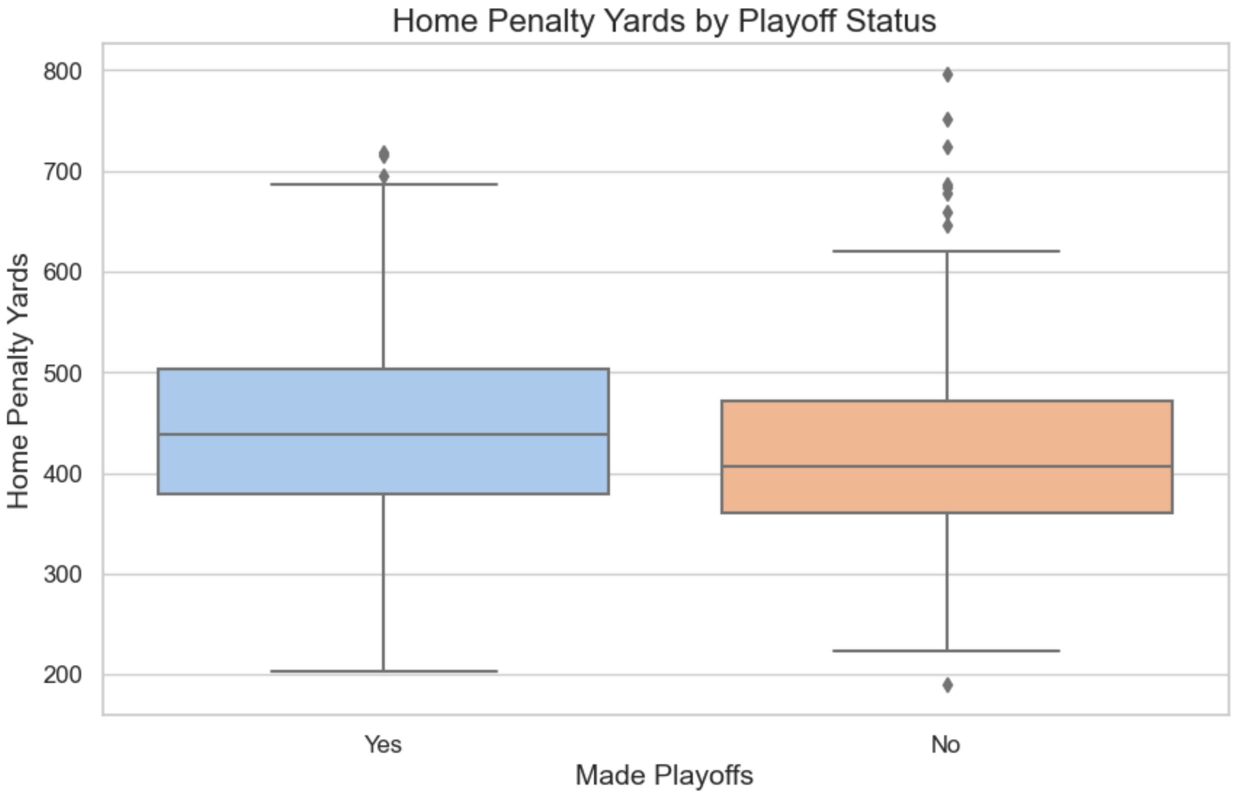
Item iv.



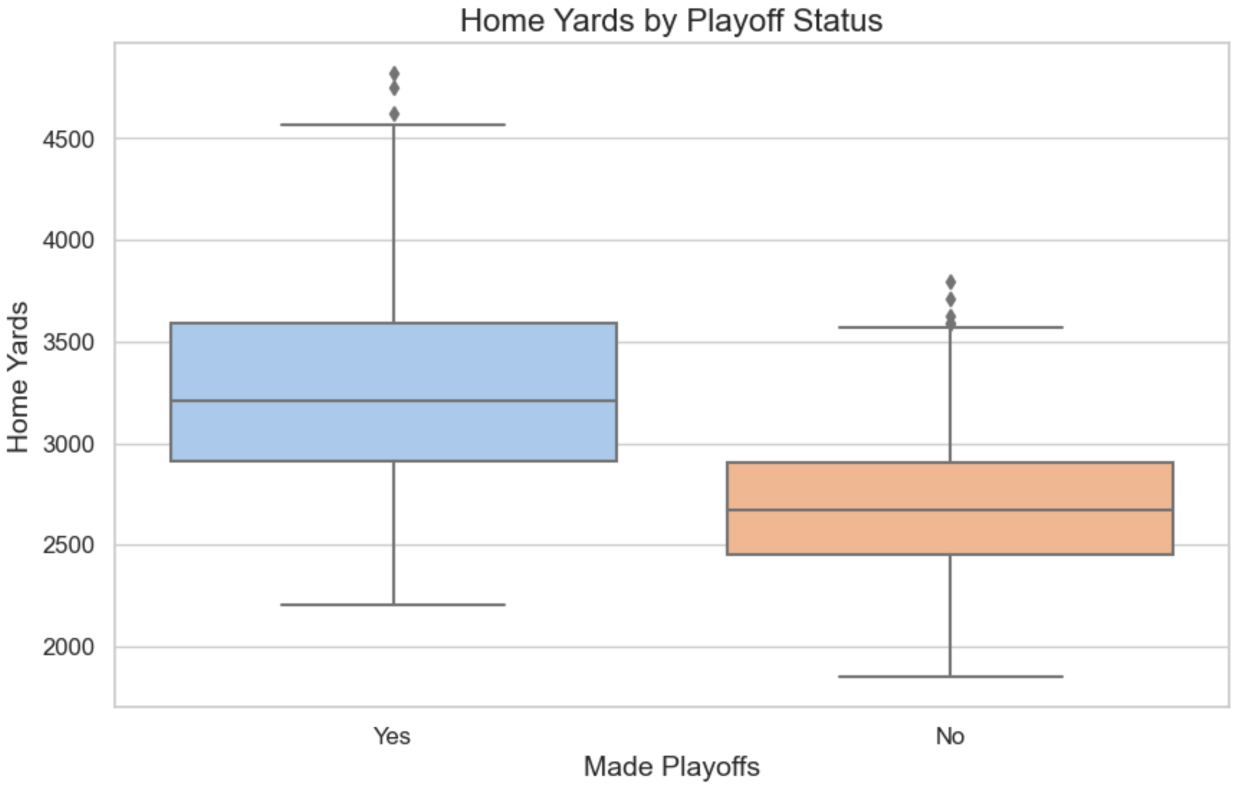
Item v.



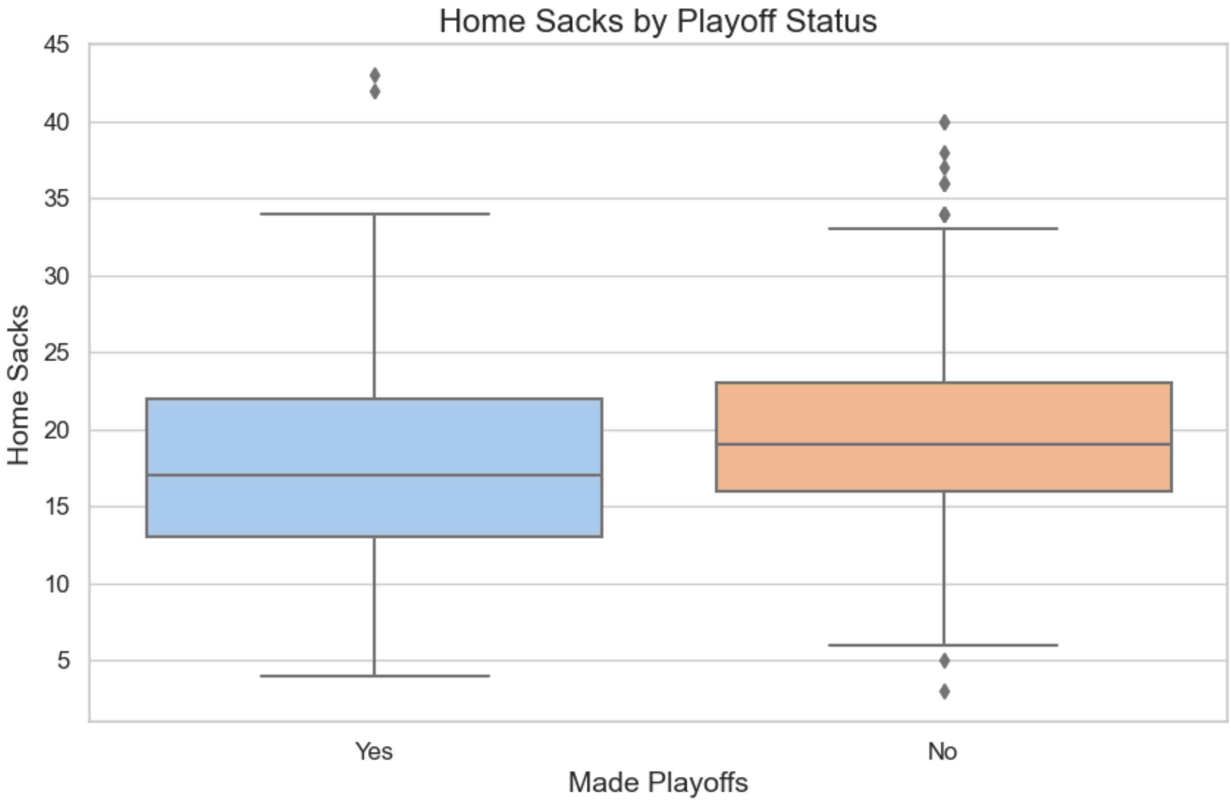
Item vi.



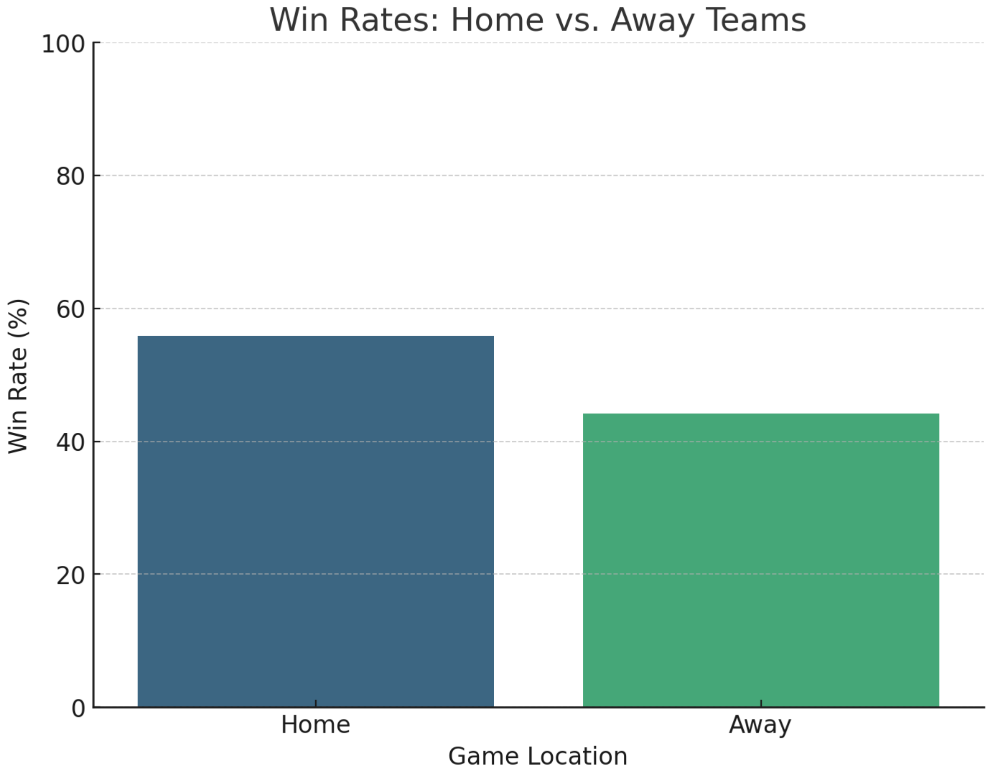
Item vii.


Home Yards by Playoff Status
Home Yards
4500
4000
3500
3000
2500
2000
Yes
No
Made Playoffs

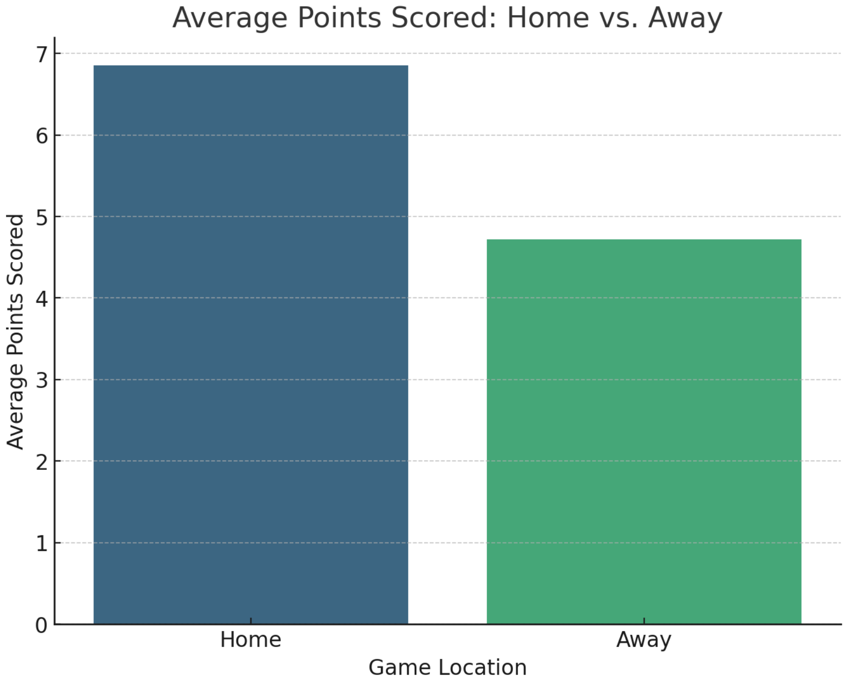
Item viii.


Home Sacks by Playoff Status
45
40
35
30
25
20
15
10
5
Home Sacks
Yes
No
Made Playoffs

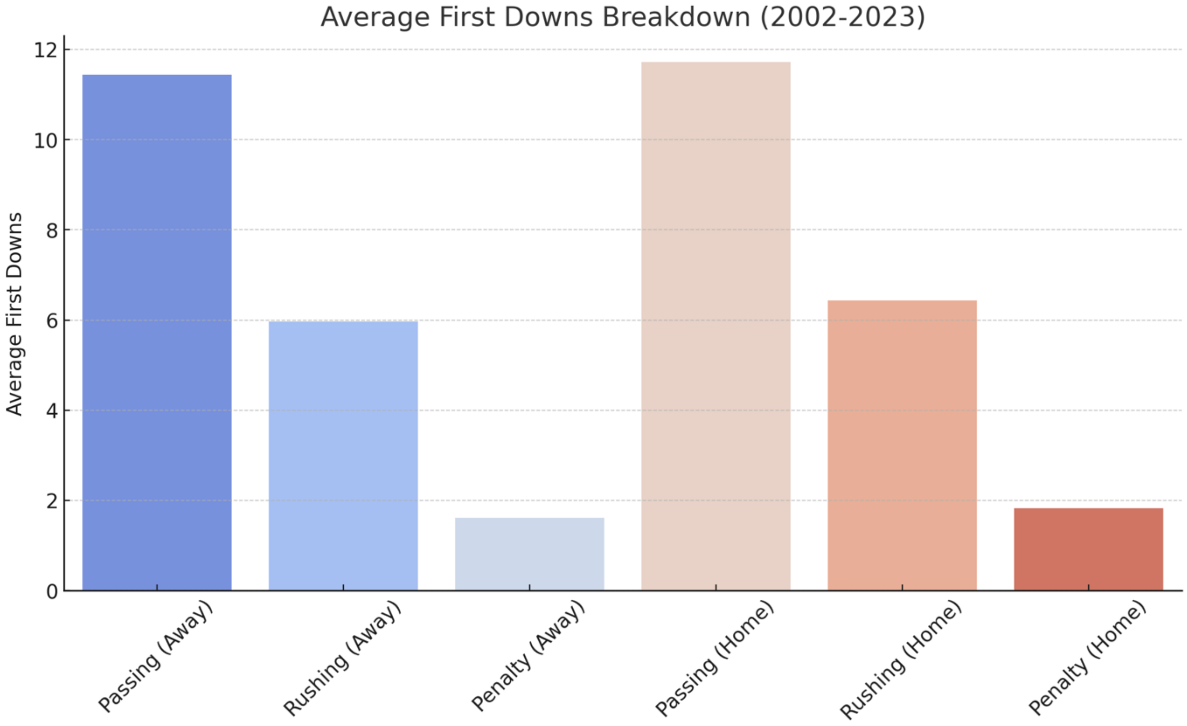
Item ix.


Win Rates: Home vs. Away Teams
100
80
60
40
20
0
Win Rate (%)
Home
Away
Game Location

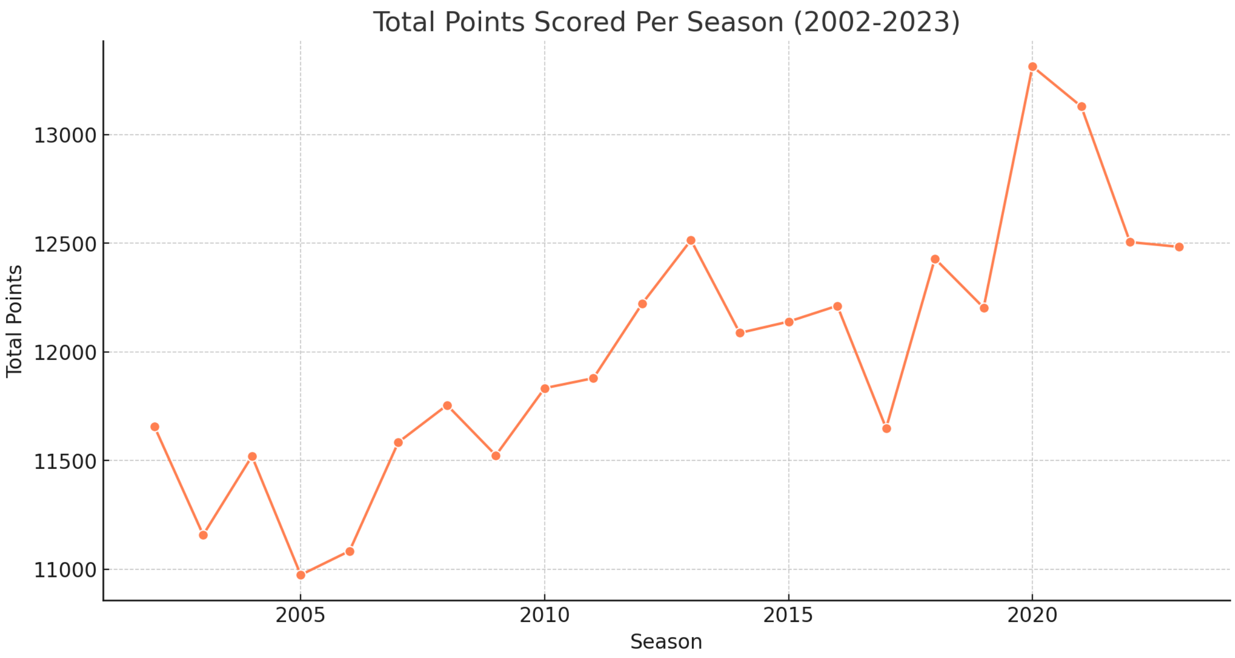
Item x.



Item xi.



Item xii.


Total Points Scored Per Season (2002-2023)
Total Points
13000
12500
12000
11500
11000
2005
2010
2015
2020
Season