Data Wrangling Final Project Report

NFL Stats and Win Percentage

Dakota Wernimont and Ahmed Abbas

1. Introduction

Many fans question what parts of football are most important in order to win games, some organizations also struggle to answer the same question. Exactly what stats of football are most impactful to the win percentage of a team? What stats correlate with each other? Being able to see exactly how impactful each stat is on if a team will win a game or not is a question we want to answer. How much of an impact does score percentage have on win percentage? Sacks or penalties taken? Yards per play? Do big plays lead to a higher point differential? Yards per play? How do negative plays like turnovers and penalties affect win percentage?

In this project, we plan to use NFL stats directly from nfl.com along with a dataset from Kaggle to see what correlation there is between certain stats. We are going to merge and clean the data until we create a data dictionary with the variables we desire.

2. Data

This project uses two main sources of data: the NFL's website on passing stats per team by season and NFL team data CSV file from Kaggle that contains basic team stats by season.

2.1 Passing Stats by Team per Season

We collected this data directly from the NFL's website. To begin, this data frame contained 17 different columns with a total of 448 rows containing the years 2010-2023 that we scraped.

The NFL's data format was very easy to work with, along with the URL used. We wrote a scraping code that will change the year by 1 after every scrape and will go through a total of 14 seasons of NFL data. We decided not to do any initial cleaning with the columns and kept them all, and waited until we merged the other source of data to decide what we wanted to answer. Some minor problems we experienced with the scraping were name changes,

especially with the Washington Commanders. We decided to keep the name Commanders consistent across the team information so we wrote a line of code that will replace all other names they had (Team, Redskins, Football Team) and used the replace function to have them all changed to commanders. Along with this the 49ers were sometimes referred to as Niners, so we just replaced all those with 49ers to ensure we got all the data.

This information from our scraping is stored in a CSV file called "nfl_passing_stats.csv" and is also in our GitHub dictionary.

2.1 Team Data by Team per Season

Through Kaggle, we found a data file that contained many team stats ranging from the years 2003 – 2023.

To make the data easy to work with through Jupyter I deleted all information for the years 2003 – 2009. At this point, we were still not focused on the columns, instead making sure the merge was perfect.

To achieve an accurate merge, we first had to change the team name as this file formatted the team as New England Patriots, when we wanted just Patriots to match the other data frame. So we split and just kept the last team (thankfully, all NFL teams are only 1 word). Then this file didn't have the team names capitalized so we also had to capitalize the start of every team name. Finally, we had to do the same we did with the first data frame to keep the team's name consistent (just made the same changes to Commanders and 49ers).

The raw file was saved under "team_stats_2003_2023.csv", but after I made some changes in Excel, it was renamed to "team_stats_2013_2023.csv". The final file with all the changes kept that name.

2.2 Combining Passing Stats and Team Data

The next step that we took after we had our data saved as data frames with matching columns, we wanted to merge we performed a merge. We saved this as merged_df and didn't on Year and Team. This was not only matched by teams but also the exact year that the

information happened to the team, meaning that our data frames were merged perfectly and matched up.

This is when we began the deep cleaning of our dataset to prepare our final_df. The main problem that we experienced was that there was an uneven number of games that was very hard to work with, as the NFL just added another game not too many seasons ago, and there was a tragic accident to 2 teams a couple of years ago that made them play 1 less game. We took the columns penalties, turnovers, 20+ plays, 40+ plays, and sacks. With these, we made all of them on a per-game basis by just dividing them by the number of games the team played that season to get a per-game stat and remove any bias or inaccurate information that may skew our data. However, for the big plays column we decided to add 20+ and 40+ plays together and divide them by games, as this would work great, as 40+ plays would be counted twice, still keeping them impactful and more impactful than 20+ plays.

The final step we took was to drop all columns that were not going to be used, were already converted, or was going to skew some of our information. Since our datasets were so large we dropped about 25 columns.

Table 1 Data Dictionary

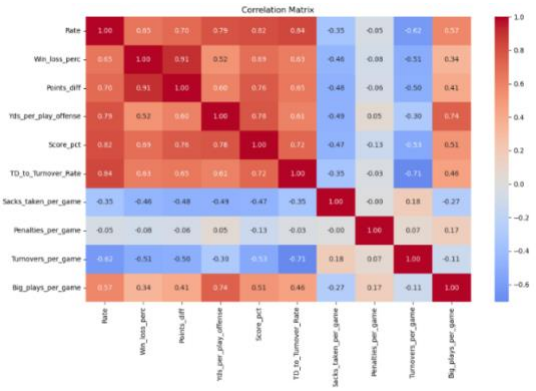
Column	Type	Source	Description
Year	Date	Both	Year of season played
Team	Text	Both	Team name for the season
Rate	Numeric	NFL.com	Team QBR for season
Win_loss_perc	Numeric	Kaggle Stats	Win to loss ratio for a given team
Points_diff	Numeric	Kaggle Stats	Point differential for a given team
Yds_per_play_offense	Numeric	Kaggle Stats	Yards gained per play

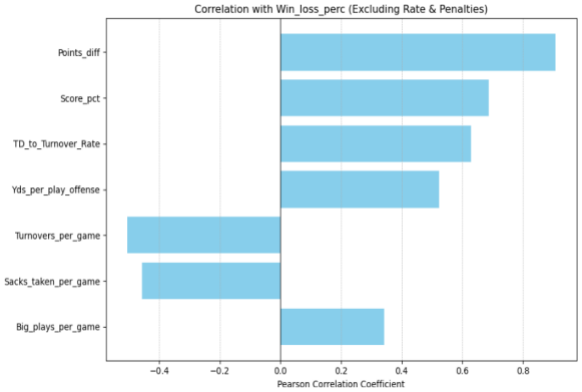
Score_pct	Numeric	Kaggle Stats	% to score points on a offensive poss.
TD_to_Turnover_Rate	Numeric	NFL.com	Rate of touchdown per turnovers
Sacks_taken_per_game	Numeric	NFL.com	Avg sacks taken per game by season
Penalties_per_game	Numeric	Kaggle Stats	Avg penalties per game by season
Turnovers_per_game	Numeric	Kaggle Stats	Avg turnovers per game by season
Big_plays_per_game	Numeric	NFL.com	Avg big plays per game by season

3. Analysis

3.1 Team Stats and Win Percentage

We wanted to run this analysis so we were able to see an overview of exactly how much weight each variable has when it comes to predicting win percentage. The first thing we ran was a simple hypothesis test on each of the variables compared to win-loss percentage. Through this, we discovered that each variable had a 0 p-value, showing a large amount of significance besides penalties per game. This variable had a .0894 P-Value, showing that there is not much significance when trying to predict our target so we are going to drop this variable.





We then wanted to test the correlation between the variables to know what variables might need to be dropped due to high correlation. Below is a heatmap of the correlations.

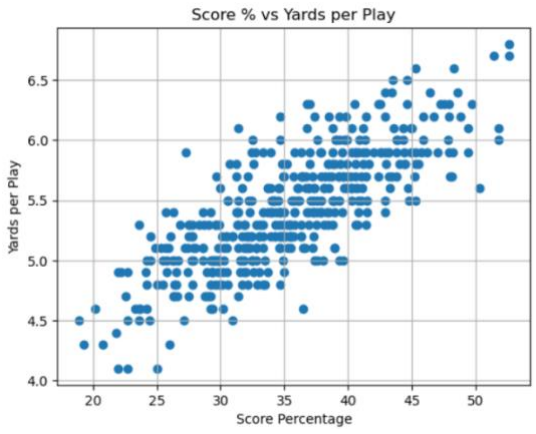
After analyzing the matrix, we decided that the only variable that needed to be dropped was the Rate variable. As it had 2 different correlations over .8, with another at .79. This makes sense because the correlation with a QB and how well they play directly affect how much they score, yards per play, and their TD to Turnover ratio of the team.

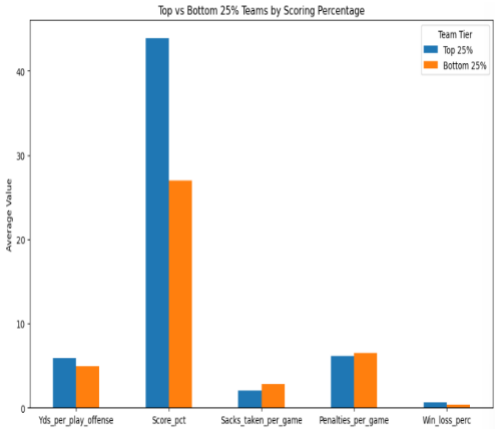
The 2nd figure was our final correlation graph. We thought this would be a very useful graph to create due as it easily shows how much correlation each variable has to win and loss percentage. Discovering which variables lower the percentage and which raise it, and roughly how much is the main overall question that we wanted to discover from this dataset. Now with the variables exactly as we want them, no correlation, and cleaned; we are able to analyze the data. The equation for predicting win and loss percentage is:

```
Multiple Linear Regression Equation:
Win_loss_perc = 0.662 + 0.002 × Points_diff - 0.017 × Yds_per_play_offense - 0.001 × Score_pct + 0.026 × TD_to_Turnover_Rate - 0.012 × Sacks_taken_per_g
ame - 0.016 × Turnovers_per_game - 0.005 × Big_plays_per_game
```

3.2 Score Percentage and Win Pct / Yards per Play / Sacks Taken / Penalties

One variable that really stood out to use was the teams scoring percentage for a given drive. We decided to run some tests to see which variables change the most based off scoring percentage.





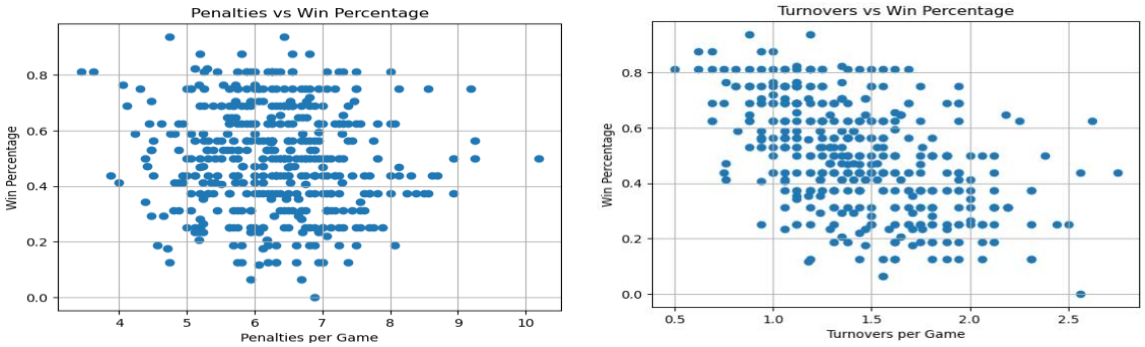
This is one example of a simple bivariate graph that we created. This shows the relationship between scoring percentage and yards per play, we created a similar graph for the 3 other variables above. However, we wanted to take this a step further by creating a graph that

divides the teams into the top and bottom 25% to see what really makes a team thrive and see what variables lead to teams doing poorly and doing good.

The figure above on the right compares the top and bottom 25% teams to see how important each variable is. The top 25% of teams have a much higher win-loss percentage, which is consistent to our assumptions. This bar graph allows us to see how important score percentage is and how teams with a higher score percentage are much more likely to be a top 25% team in the league. Sacks and penalties per game look to be very minor in deciding if a team is on the top or bottom quartile, as they are very close and there aren't many values for each of them.

3.3 Turnovers / Penalties and Win Percentage

After taking a deeper dive into the positive correlations and the positive plays in the game of football, we decided to look into how negative plays directly affect win-loss percentage.



Our hypothesis test showed that penalties per game had no major significance when predicating win-loss percentage which is clearly shown in the graph, but the turnovers per game has a clear negative relationship when coming to win-loss percentage. The next steps we took was comparing both of these models using linear, Ridge, and Lasso regressions.

Model	R Squared	MSE
Linear Regression	0.309317	0.024933
Ridge Regression	0.308140	0.024975
Lasso Regression	-0.004384	0.036257

Through the comparison of different models, Linear regression performed the best out of all the models, while Ridge had nearly identical results. The linear regression explained nearly 31% of the variance, Lasso performed very poorly, which is most likely due to overregularization. If we were to add more features, we would likely increase the accuracy, but this alone tells us that this current model is important when predicant win percentages.

3.4 *Predicting Win-Loss Percentage*

	Team	Year	Actual	Forecast
0	49ers	2022	0.765	0.409
2	Bears	2022	0.176	0.388
4	Bengals	2022	0.750	0.368
6	Bills	2022	0.813	0.719
8	Broncos	2022	0.294	0.383

Through the use of exponential smoothing to forecast a NFL team's win-loss for the last 2 season of our data set. For each team, we trained the model on all previous seasons and predicted the most recent two, then compared the predicted values to the actual results. While some forecasts were close (Bills), others showed significant discrepancies (49ers and Bengals), suggesting that recent team changes or variability in performance can limit the accuracy of simple time series models. Overall, the analysis highlights that while historical trends can offer insights, they may not fully capture the dynamic nature of team performance in the NFL. It would be much more beneficial to include other variables to use to predict and train the model to build a fully reliable predictor for W/L percentage.

4. Conclusion

In this project, we analyzed both passing and basic team stats, comparing them to win percentage and also to each other. We took a deep dive into correlation to ensure we only have impactful variables while also running many different statistics, different machine learning techniques, and also a forecasting example.

1. What variables are the most important to a team's win-loss percentage? If we wanted a model with no correlation or bias what would it include?
   The most important variables for when it comes to a team's win loss percentage would be point differential, scoring percentage, and TD to turnover rate. With the main focus to reduce turnovers per game as it is the largest negative value. The best model with all this data would include: point differential, scoring percentage, TD to Turnover rate, yards per offensive play, turnovers per game, sacks taken per game, and big plays per game.
2. How does scoring percentage affect win percentage, yards per play, sacks taken per game and penalties per game?
   Scoring percentage has a very consistent and direct correlation with yards per play and win percentage. An increase in scoring percentage will also increase both variables. The correlation for sacks taken per game has a very negative relationship, proven in our graphs, meaning an increase in scoring percentage

will lower sacks per game. The relationship with penalties per game does not have a noticeable relationship, as it is nearly symmetrical.

3. How do negative plays like penalties and turnovers per game have a direct impact on win-loss percentage?

   Negative plays like penalties and turnovers per game have a noticeable impact on win-loss percentage. Based on the different models, these two factors together explain about 31% of the variation in team performance, indicating a moderate relationship. This suggests that teams with fewer turnovers and penalties generally tend to win more games. The more a team avoids these mistakes, the better its chances of success.

4. Can we predict the win percentage of a team using history of win-loss percentage?

   Yes, we can predict a team's win percentage using historical performance, and our analysis using Exponential Smoothing shows that past trends can offer some predictive power. However, the accuracy varies by team—some predictions were close, while others had large errors. This suggests that while historical data is useful, it may not fully account for recent changes like roster moves, injuries, or coaching changes that significantly impact team performance.

This project demonstrates the power of data wrangling and statistical analysis in understanding the key drivers behind NFL team success. By cleaning, merging, and analyzing multiple datasets, we identified which statistics most influence a team's win-loss percentage, and we built models to evaluate both current season predictors and historical forecasting.

While some variables like turnovers, scoring percentage, and point differential proved to be consistently impactful, others, such as penalties, showed limited predictive power. Our modeling also showed that although simple historical models can offer some insight, they are not fully reliable on their own due to the changing dynamics of NFL teams. Using player-level data, injury reports, and advanced metrics could significantly improve model performance. Overall, this project highlights how data analytics can be a valuable tool for teams, analysts, and fans alike in breaking down what truly leads to winning in the NFL.