Team Members: Lance Regehr, Sebastian Cochran, Lucas Jensen

Data Wrangling Project Report

1. Introduction

Any avid NFL watcher recognizes that inclement weather has a major impact on the outcomes of games. From the 1967 "Ice Bowl" to the blustery 1982 AFC Championship game, the weather has impacted how different teams plan and play games. But what specific weather environments create specific game scripts? Does being outside greatly affect stats like kicking percentage and the amount of passing yards thrown? Are some weather factors more detrimental than others?

In this project, we used Pro Football Reference's in-game statistics and Open_Mateo's historical weather API to see if there is a correlation between inclement weather and in-game statistics, and whether being indoors and outdoors affects the play on the field.

2. Data

This project uses two primary sources of data: Pro Football Reference's in-game statistics for NFL games and Open-Mateo API to find the weather during the game at the specific stadium. We also briefly had to use GeoApify to convert the stadium address to coordinates.

2.1 Pro Football Reference Data

We collected data from Pro Football Reference (PFR), which contained all the games we wished to scrape data from. These games were the 129 most recent NFL games with a unique final score.

Pro Football Reference game links were found on one page, but the game data was found on 129 different web pages using the game links. We wrote a web crawling script to collect data across the 129 pages. We collected the Game ID, total passing yards, total rushing yards, average kicking percentage (field goals), date, start time, stadium, start time, and roof type. We then created lists to store these data points for specific games.

And within each game scraped, we accessed each stadium link to reach the specific stadium's page on the PFR site. In this stadium page, we scraped the address of the specific stadium and stored that data in a list.

After scraping, we realized that to get the proper weather data, we had to create a specific time stamp for our start time. In this process, we had to combine our start time and date features into one timestamp. We did this by splitting our start time data and using datetime library operations to manipulate the data points into one concise time stamp.

Pro Football Reference

2.2 GeoApify (address -> coordinates)

We had to use a secondary API (GeoApify) to convert the addresses we scraped from the stadium page into coordinates to plug into our primary API for historical weather. We had to do this because Open-Mateo only accepts coordinates and not addresses.

GeoApify

2.3 Open-Mateo Historical Weather API

Open-Mateo has an expansive collection of hourly weather data points from dew point to soil moisture. For our analysis, we gathered wind speed (10 m), total precipitation (rain + showers + snow), and temperature (Celsius) data into separate stored lists as we did with the Pro Football Reference datapoints.

For each NFL game, we entered the game's date into Open-Mateo API to receive the 24-hour weather report for the specific day and location. We used the start time of the game plus 3 hours to create the timespan the game took place. Within this 3-hour range, we calculated average wind speed, total precipitation, and average temperature for each NFL game.

Open-Mateo

2.4 Creating a Data frame from Stored Lists

With the stored lists of all our variables, we created an all_*model_data* data frame that will now be used to store our data to conveniently conduct our analysis. The web crawling script that we wrote is contained in the *Wrangling_Project.ipynb* in the project folder.

Column	Type	Source	Description
game	Text	Pro Football Reference	Unique identifier for each game

pass_yrds	Numeric	PFR	Total passing yards during game
rush_yrds	Numeric	PFR	Total rushing yards during game
fgp	Numeric	PFR	Field goal percentage during game
avg_temperature	Numeric	Open-Mateo	Average temperature during game
tot_precipitation	Numeric	Open-Mateo	Total precipitation during game
avg_wind_speed	Numeric	Open-Mateo	Average wind speed at 10m during game
date	Date	both	Timestamp of the start of the game
roof	Text	PFR	Whether the game was played in a dome, with a retractable roof(open or closed), or outdoors

3. Analysis

3.1 Inclement Weather and Rushing Yards

In our analysis, we wanted to find out how inclement weather affected how teams play games. The first variable we wanted to analyze was total rushing yards per game. Before using the model, we eliminated games played in domes and stadiums with retractable roofs, as weather does not play a factor. With this subset of data, we used a multi-linear regression model to determine the impact each weather variable had on our y variable, *rush_yrds*. To create the model, we used the *statsmodels* library. The library was also used to obtain a model summary, which included our r^2 value, variable p-values, and our coefficients. In this model, our coefficients were -.1407 for *tot_precipitation,* -1.046 for *avg_wind_speed*, and -0.652 for *avg_temperature*. Our intercept was 257.68. Our r^2 value was .02 and all our variables' p-values were greater than our alpha of .05, and thus statistically insignificant. All these metrics told us that our weather variables did not statistically make a significant impact on the variability of total rush yards for NFL games.

This lack of correlation was a little surprising to us. We had thought that an increase in precipitation might lead to an increase in rushing yards. In our experience, we have seen many teams run the ball more during rain or snow games. To further show the correlation between variables, we decided to plot our variables in a scatter plot as shown in *Figure 1*. *Figure 1* compares *rush_yrds*

and *tot_precipitation* and shows a slight negative correlation, which matches *tot_precipitation*'s small negative coefficient. This negative correlation could just be explaining that there is generally less offensive success when there is more precipitation, which would make sense.

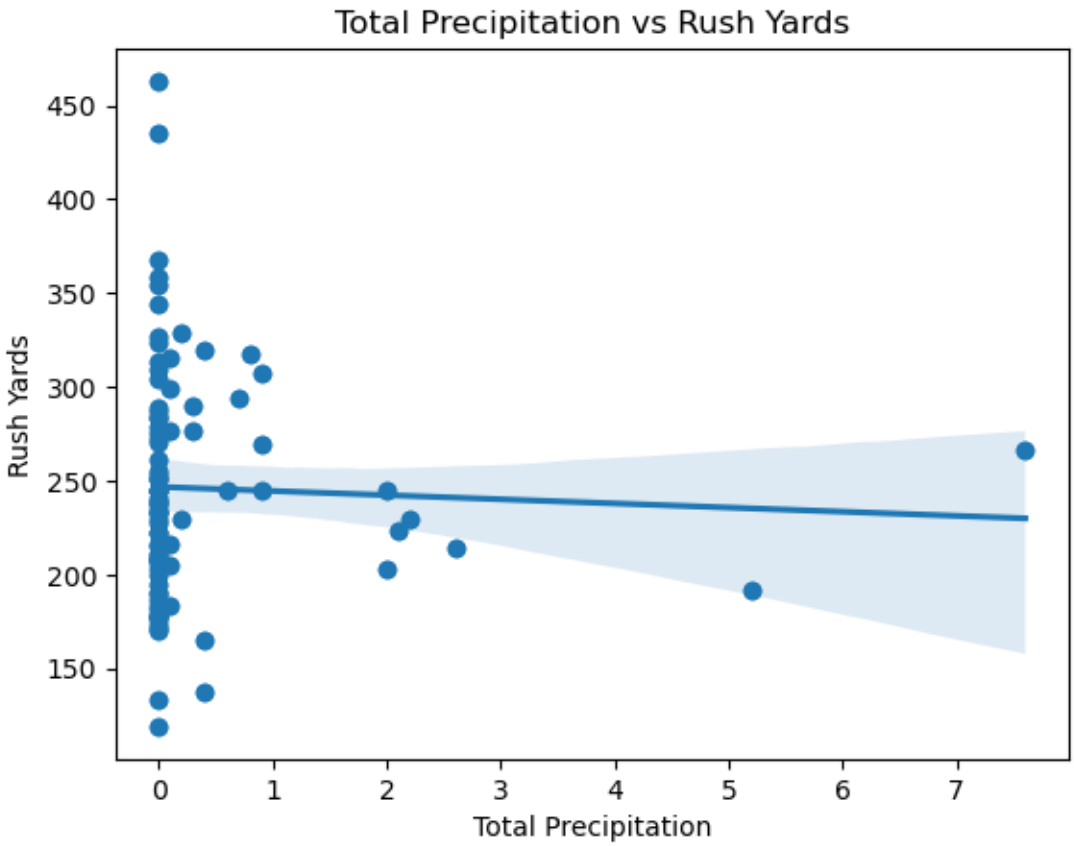

Figure 1

3.2 Field Goal Percentage: Indoor vs Outdoor

The next variable we analyzed was field goal percentage (*fgp*), the total amount of field goals made in a game divided by how many field goals were attempted. We were interested in finding out whether or not kickers performed better indoors or outdoors. Our hypothesis was that kickers performed worse outdoors than indoors. To test this, we created a subset of our *all_model_data that* only contained games in which a field goal was attempted. We then split the data into indoor games (where *roof* is not 'outdoors') and outdoor games (where *roof* is 'outdoors'). From there, we computed the mean for each data set. The mean of *fgp* for indoor games was .875, while for outdoor games it was .852. Although we found that, on average, the field percentage is less for outdoor games, we needed to properly test its statistical significance. To test this, we ran a t-test

using SciPy and found a p-value of .542. As the p-value was greater than our alpha of .05, we failed to reject our null hypothesis, and found that there is no significant difference in *fgp* between indoor and outdoor games. To properly visualize the distributions of our data subset, we utilized a box and whisker plot (seen in Figure 2). This figure shows the slight difference in means, but generally the same distribution. This supports our t-test findings. An interesting thing we found within this plot is that the worst field goal percentage game happened outdoors, where no kicker made a field goal with at least one attempt.

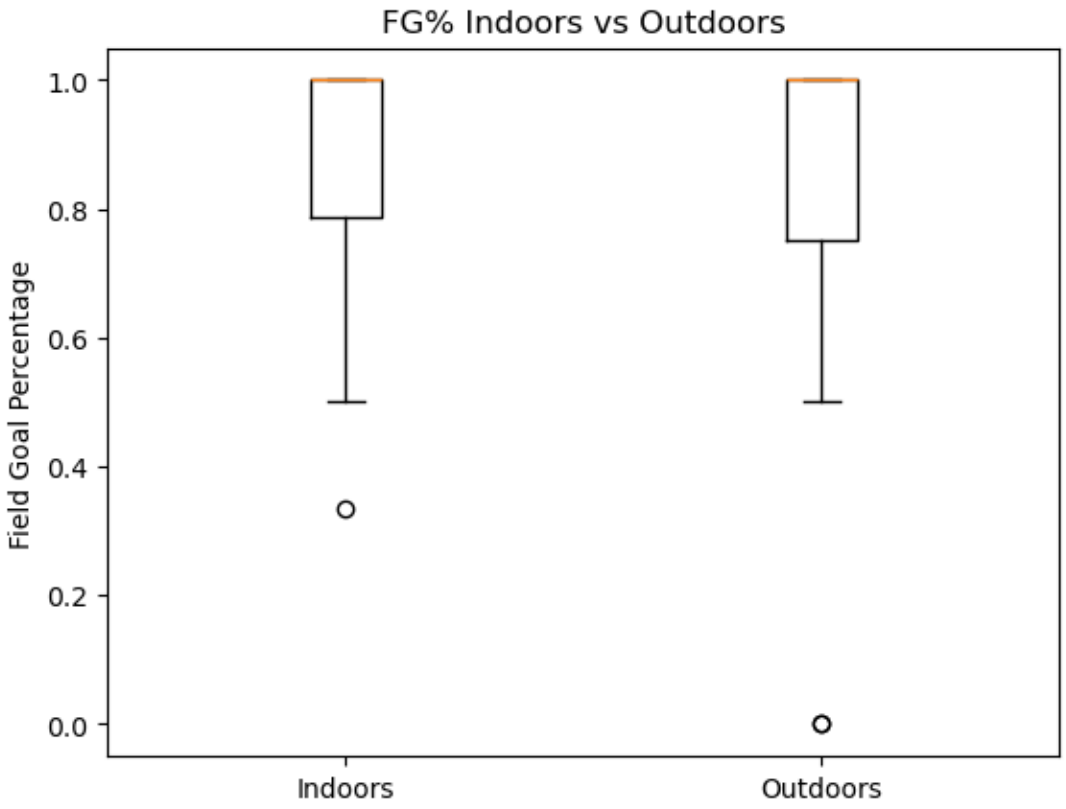

Figure 2

3.3 Passing Yards: *Indoor vs Outdoor*

Our final variable we wanted to investigate was total passing yards per game (*pass_yrds*). In this analysis, we were interested in how much roof type mattered in regard to total passing yards. To understand this better, we first had to create a dummy variable column *roof_dummy* in *all_model_data*. For each game, the value of *roof_dummy* is '1' if outdoors and '0' if not. We then ran a linear regression with *roof_dummy* as our only x variable and *pass_yrds* as our y. Our model had an R^2 of .049, with *roof_dummy* having a coefficient of -51.2284 and

a p-value of .011. Since *roof_dummy*'s p-value is less than our alpha of .05, whether a game is played outdoors or indoors has a statistically significant impact on the *pass_yrds* of that game. Specifically, if a game is played outdoors, based on our model, total passing yards will decrease by approximately 51 yards, holding all else constant. By our R^2 value, we can tell that approximately 5% of the variable *pass_yrds* can be explained by what roof the stadium has. To better visualize the difference in pass_yrds by roof, we used a box-and-whisker plot seen in Figure 3. This plot supports our model as we see the passing yards distribution shift lower when the game is outdoors. An ironic data point in this plot, though, is the fact that the highest passing yard total of our dataset occurred in an outdoor stadium. Upon investigating this data point, we found it was a matchup between the Cincinnati Bengals and the Chicago Bears, where the teams combined for 772 passing yards. Conditions were not too bad: approximately 7.7° C (45.9° F), 0.2 cm of precipitation, and 2.8 mph wind.

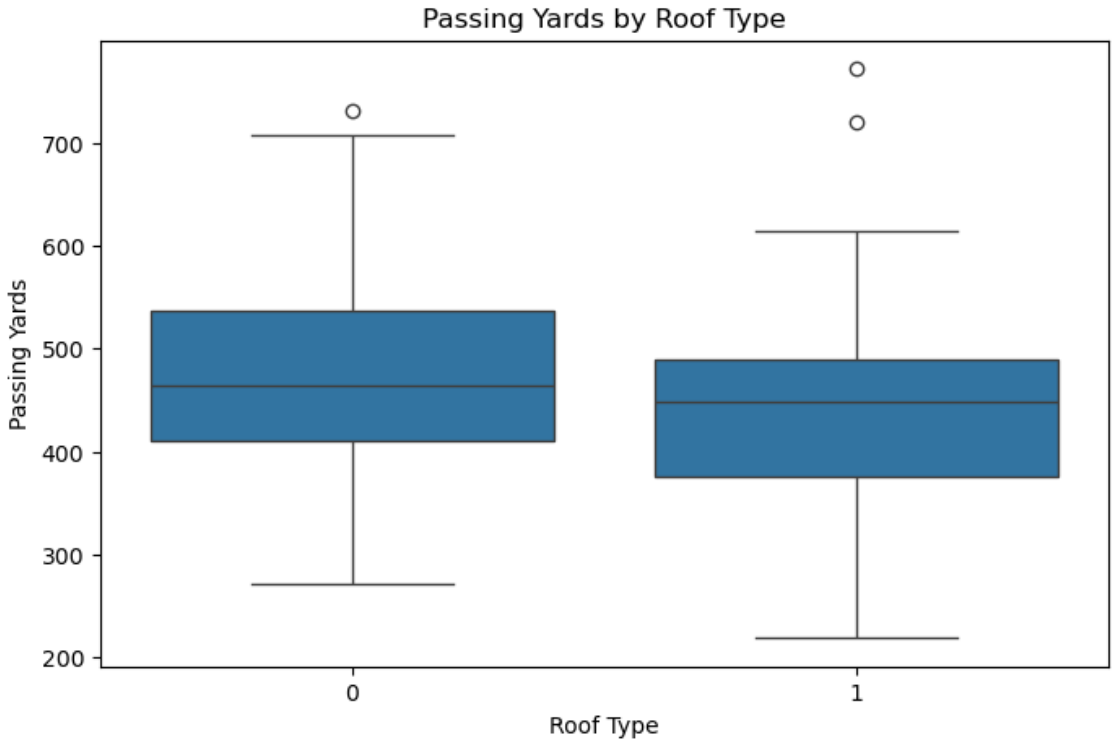

Figure 3

4. Conclusion

In this project, we analyzed three aspects of in-game statistics for NFL games: total passing yards, total rushing yards, and field goal percentage, and how outdoor conditions and weather affect each of them. In summary, from the analysis questions presented in our proposal, we found the following results.

1. *How does inclement weather affect the amount of total rushing yards?*
   Inclement weather does not make a statistically significant impact on the number of rushing yards in a game. By way of a multi-linear regression model, we found limited correlation and too large of p-values amongst our weather variables of precipitation, average wind speed, and average temperature for any influence to be statistically significant.

2. *Do kickers perform better at field goals in indoor or outdoor stadiums?*

   There is no statistically significant difference between the performance of kickers in indoor stadiums vs. outdoor stadiums. By way of a t-test, we found that the two subsets of field goal percentages were not significantly different.

3. *How are total passing yards affected in an outdoor stadium versus an indoor stadium?*

   Total passing yards are negatively impacted by being outdoors. Through a linear regression model, we found that when a game is being held outdoors, total passing yards decrease by approximately 50 yards, holding all else constant. The difference in roof type accounts for about 5% of the total passing yards variations.

This project has several limitations. Our models and tests were limited by the amount of data we collected. We were limited in the amount of data we could gather due to the length of time it took to scrape page by page within Pro Football Reference. We believe that with more data, we could create better and more significant insights. Our variables could have been better selected. After the analysis, we found that maybe rushing attempts and passing attempts were better measures of how teams play the game in inclement weather than total rushing and passing yards. More factors seem to affect the 'yards' measure than just general play calling. Something to take note of is the fact that we only scraped games with unique scores due to the convenience of having a whole list of game links on one page. Future work on this project could include editing the script so that we scrape

data from every single game within a large timeframe (like the last decade etc.), and we collect more descriptive data like passing and rushing attempts, or even look at how defense is affected. Our analysis with more games and better descriptive variables could lead to better insights with more robust models.