DATA 607 - Final Project Presentation

Julian Adames-Ng

2024-12-11

1. Introduction

Initial Project Proposal:

For my project, I aim to analyze NFL statistics to uncover insights into player performance and game outcomes. The motivation stems from the growing interest in data-driven sports analytics and its impact on player evaluation and fan engagement. I also have had growing interest in NFL data as someone who was not into sports growing up. My data will be sourced from public NFL datasets, scraped directly from websites such as ESPN.com, The Football Database, and directly from NFL.com. I will try to find CSV files and other sources of data. The workflow will involve data acquisition, cleaning, and transformation (e.g., converting wide to long formats for analysis), followed by statistical analysis and visualization to highlight trends and validate my conclusions. I also plan to use topics not covered in class, such as predictive modeling for player performance.

Motivation:

My motivation is to investigate NFL team and player performance by analyzing game outcomes, weather data, and player stats. The goal is to see patterns in team success, win rates, and average yards gained in different weather conditions, like extreme or ideal weather.

The process includes cleaning and organizing the data (like updating old team names) and grouping weather conditions into categories such as "Rain" or "Cold." Using tools like regression analysis, scatterplots, and violin plots, the results show that while extreme weather can have some impact, factors like team strength and player skill are more important. These findings suggest future studies could include more variables to better understand what drives performance.

Data Sources:

(1) NFL Team and Weather Stats: I downloaded a CSV file with NFL statistics dating back to 1960 on the following website:

https://nflsavant.com/about.php.

The raw data can be found on my Github in the following url:

https://raw.githubusercontent.com/JAdames27/DATA-607---Data-Acquisition-and-Management/refs/heads/main/DATA%20607%20-%20Final%20Project/weather_20131231.csv

(2) NFL Player Data Obtained from the nflreadr package in R

Loading the Data:

The following code clears the R environment and loads essential libraries like nflreadr and tidyverse for data handling. It then imports weather and player data from CSV files into separate dataframes. Finally, it displays the structure of both datasets to examine their contents and variable types.

```
## -- Attaching core tidyverse packages --------------------
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.o
```

```
##             id            home_team home_score
## 1 196009230ram    Los Angeles Rams         21 St. Louis
```

Creating a Winner Column

A new column, winner, is created to indicate which team won each game by comparing home and away scores. If the game was a tie, it is assigned “Tie” in the column. Rows with ties are then filtered out to focus only on games with clear winners.

```
##              id           home_team home_score
## 1 196009230ram    Los Angeles Rams         21 St. Louis
## 2 196009240dal      Dallas Cowboys         28 Pittsburgh
## 3 196009250gnb   Green Bay Packers         14       Chic
## 4 196009250sfo San Francisco 49ers         19     New Yo
## 5 196009250clt     Baltimore Colts         20 Washington
## 6 196009250phi Philadelphia Eagles         24    Clevela
##   temperature wind_chill humidity wind_mph
## 1          66         NA      78%        8
## 2          72         NA      80%       16
## 3          60         NA      76%       13
## 4          72         NA      44%       10
## 5          62         NA      80%        9
## 6          61         NA      77%        9
```

Mapping Outdated Team Names

This chunk defines a mapping of old team names to their current names for standardization. It then updates the home_team and away_team columns in the dataset using the mapping. This ensures that all references to teams are consistent across the dataset.

```
##              id           home_team home_score
## 1 196009230ram    Los Angeles Rams         21 St. Louis
## 2 196009240dal      Dallas Cowboys         28 Pittsburgh
## 3 196009250gnb   Green Bay Packers         14       Chic
## 4 196009250sfo San Francisco 49ers         19     New Yo
## 5 196009250clt     Baltimore Colts         20 Washington
## 6 196009250phi Philadelphia Eagles         24    Clevela
##   temperature wind_chill humidity wind_mph
## 1          66         NA      78%        8
## 2          72         NA      80%       16
## 3          60         NA      76%       13
## 4          72         NA      44%       10
## 5          62         NA      80%        9
## 6          61         NA      77%        9
```

Pairing Team to Home City

A lookup table is created to pair each team with their home location. The dataset is updated by joining this location data with the existing home_team information. This provides geographical context for each game based on where it was played.

```
##              id           home_team home_score
## 1 196009230ram    Los Angeles Rams         21     Arizon
## 2 196009240dal      Dallas Cowboys         28   Pittsbur
## 3 196009250gnb   Green Bay Packers         14         Ch
## 4 196009250sfo San Francisco 49ers         19       New 
## 5 196009250clt  Indianapolis Colts         20 Washingto
## 6 196009250phi Philadelphia Eagles         24      Cleve
##   temperature wind_chill humidity wind_mph
## 1          66         NA      78%        8
## 2          72         NA      80%       16
## 3          60         NA      76%       13
## 4          72         NA      44%       10
## 5          62         NA      80%        9
## 6          61         NA      77%        9
```

Seasonal Football Data Analysis

The following code organizes football game data by season, adjusting for games in January and February belonging to the previous year, and calculates season-level summaries including total games, average home score, and average away score.

```
## # A tibble: 6 x 4
##   season total_games average_home_score average_away_sco
##    <dbl>       <int>              <dbl>              <db
## 1   1960          74               22.6               19
## 2   1961          96               23.1               20
## 3   1962          95               21.8               22
## 4   1963          94               22.8               21
## 5   1964          93               22.7               21
## 6   1965          98               22.7               23
```

Team Win Calculation by Season

This code processes game data to assign each game to a specific season, determine the winning team, and calculate the total wins for each team by season, excluding ties. It ensures all teams are included in the dataset, even those with no wins, and outputs a comprehensive summary of team performance across seasons.

```
## # A tibble: 6 x 3
##   season team                total_wins
##    <dbl> <chr>                    <int>
## 1   1960 Los Angeles Rams             4
## 2   1960 Arizona Cardinals            6
## 3   1960 Dallas Cowboys               0
## 4   1960 Pittsburgh Steelers          5
## 5   1960 Green Bay Packers            8
## 6   1960 Chicago Bears                5
```

Team Performance Metrics by Season

This code calculates the total games played by each team (both home and away) for every season, combines it with the total wins data, and computes each team's win rate. The results are organized by season and team, providing a comprehensive overview of team performance metrics.

```
## # A tibble: 6 x 5
##   season team              total_wins total_played win_r
##    <dbl> <chr>                  <int>        <int>    <d
## 1   1960 Arizona Cardinals          6           11    0.
## 2   1960 Chicago Bears              5           11    0.
## 3   1960 Cleveland Browns           8           11    0.
## 4   1960 Dallas Cowboys             0           11    0
## 5   1960 Detroit Lions              7           12    0.
## 6   1960 Green Bay Packers          8           13    0.
```

Game Statistics by Temperature Range

This code defines a function to filter games based on specified temperature and wind speed ranges, categorizing them into extreme weather games and ideal weather games (moderate conditions). The filtered datasets provide insights into game characteristics under varying weather conditions.

```
##              id            home_team home_score
## 1 196010090was Washington Commanders         26       Dal
## 2 196010230gnb     Green Bay Packers         41 San Fran
## 3 196010230det         Detroit Lions         30  Indiana
## 4 196010300was Washington Commanders         10    Cleve
## 5 196010300nyg       New York Giants         13   Arizon
## 6 196010300dal         Dallas Cowboys          7  Indiana
##   temperature wind_chill humidity wind_mph
## 1          59         NA      85%        1
## 2          49         NA      71%       21
## 3          53         NA      60%       22
## 4          54         NA      56%        3
## 5          54         NA      52%       27
```

Team Performance in Extreme Weather Games

This code processes games played in extreme weather conditions to determine the winning team and calculate total wins for each team by season. It ensures that all teams, even those without wins, are included in the dataset and organizes the results by season and team for a clear overview of performance under extreme conditions.

```
## # A tibble: 6 x 3
##   season team              total_wins
##    <dbl> <chr>                  <int>
## 1   1960 Arizona Cardinals          1
## 2   1960 Chicago Bears              0
## 3   1960 Cleveland Browns           3
## 4   1960 Dallas Cowboys             0
## 5   1960 Detroit Lions              3
## 6   1960 Green Bay Packers          1
```

Team Performance in Ideal Weather Games

The code below processes games played under ideal weather conditions to determine the winning team and calculate total wins for each team by season, ensuring all teams are included, even those with no wins.

```
## # A tibble: 6 x 3
##   season team                  total_wins
##    <dbl> <chr>                      <int>
## 1   1960 Arizona Cardinals              1
## 2   1960 Chicago Bears                  0
## 3   1960 Cleveland Browns               3
## 4   1960 Dallas Cowboys                 0
## 5   1960 Detroit Lions                  3
## 6   1960 Green Bay Packers              1
```

```
## # A tibble: 6 x 3
##   season team                  total_wins
##    <dbl> <chr>                      <int>
## 1   1960 Arizona Cardinals              5
```

Performance Metrics

This code calculates the total games played by each team in extreme weather conditions and combines it with total wins to compute win rates for each season. Missing values for wins or games played are replaced with zeros, ensuring completeness, and the data is sorted by season and team for clarity. The result provides a comprehensive overview of team performance in extreme weather.

```
## # A tibble: 6 x 5
##   season team                 total_wins total_played win_r
##    <dbl> <chr>                     <int>        <int>    <d
## 1   1960 Arizona Cardinals             1            1
## 2   1960 Chicago Bears                 0            2
## 3   1960 Cleveland Browns              3            3
## 4   1960 Dallas Cowboys                0            3
## 5   1960 Detroit Lions                 3            3
## 6   1960 Green Bay Packers             1            2
```

Ideal Weather Team Performance Summary

The process calculates the total games played by each team under ideal weather conditions and combines it with win data to determine seasonal win rates. Missing values are replaced with zeros for completeness, and the results are sorted by season and team for clarity. The final output highlights team performance in favorable weather scenarios.

```
## # A tibble: 6 x 5
##   season team                 total_wins total_played win_r
##    <dbl> <chr>                     <int>        <int>    <d
## 1   1960 Arizona Cardinals             1            1
## 2   1960 Chicago Bears                 0            2
## 3   1960 Cleveland Browns              3            3
## 4   1960 Dallas Cowboys                0            3
## 5   1960 Detroit Lions                 3            3
## 6   1960 Green Bay Packers             1            2
```

```
## # A tibble: 6 x 5
##   season team                 total_wins total_played win_r
```

Average Win Rate Calculation Summary

This process calculates the average win rate for each team across all seasons using data from extreme weather games. Missing values are ignored during computation, and the results are saved to a CSV file for further analysis. The output provides a clear summary of team performance trends in extreme conditions.

```
## # A tibble: 6 x 2
##   team              mean_win_rate
##   <chr>                     <dbl>
## 1 Arizona Cardinals         0.362
## 2 Atlanta Falcons           0.253
## 3 Baltimore Ravens          0.561
## 4 Buffalo Bills             0.457
## 5 Carolina Panthers         0.559
## 6 Chicago Bears             0.480
```

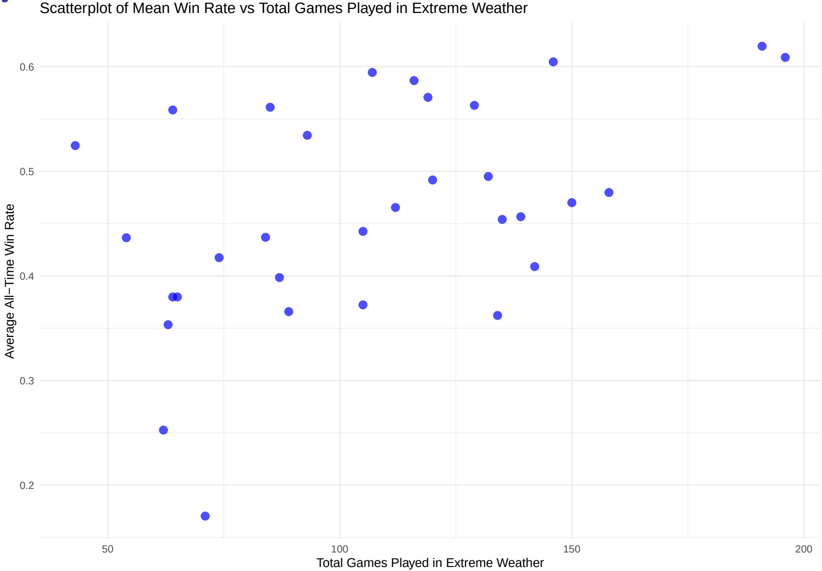
Lifetime Performance in Extreme Weather Summary

This analysis aggregates each team's total games played, total wins, and average win rate across all seasons in extreme weather conditions. It also calculates an all-time win rate by dividing total wins by total games played, providing a comprehensive overview of long-term performance under challenging conditions.

```
## # A tibble: 6 x 4
##   team               total_games total_wins avg_win_rate
##   <chr>                    <int>      <int>        <dbl>
## 1 Arizona Cardinals          134         50        0.362
## 2 Atlanta Falcons             62         20        0.253
## 3 Baltimore Ravens            85         54        0.561
## 4 Buffalo Bills              139         63        0.457
## 5 Carolina Panthers           64         36        0.559
## 6 Chicago Bears              158         82        0.480
```

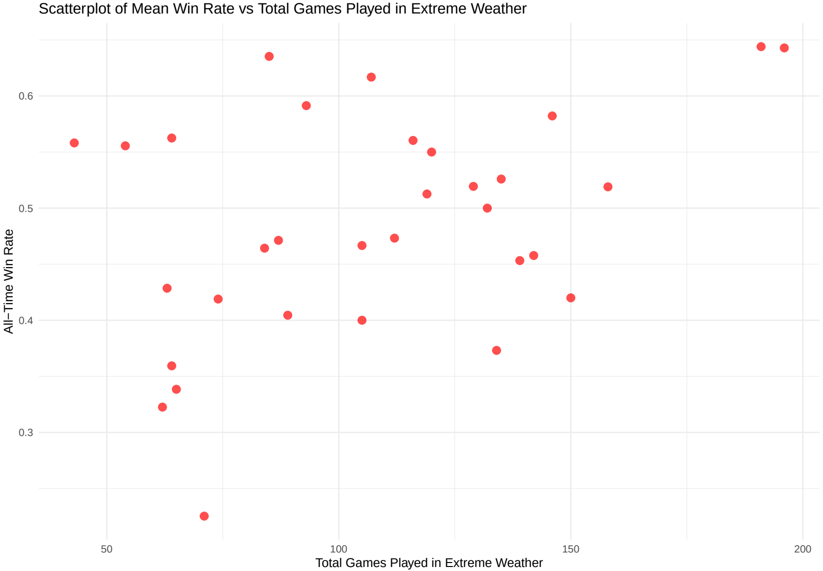
```
## # A tibble: 6 x 5
##   team               total_games total_wins avg_win_rate
##   <chr>                    <int>      <int>        <dbl>
```

Average Seasonal Win Rate vs Total Number of Games Played in Extreme Weather



##

All-Time Win Rate vs Total Number of Games Played in Extreme Weather



##

Conclusion

Part 1

Of the two models, the one comparing Mean Win Rate and Total Games Played in Extreme Weather is the stronger model with a higher adjusted R-squared, better predictor significance, and a good overall fit, while matching the latter model in accuracy. The results show a modest positive relationship between games played in extreme weather and average win rate, but the low R-squared suggests other factors, like team strength or weather severity, may play a bigger role. Adding more variables could improve the model's effectiveness.

Second Data Analysis - Data Set 2 (Using NFLfastR and NFLreadR Libraries)

Quarterback Stats Data Compilation

This script uses the nflfastR package to load play-by-play data for NFL seasons from 2000 to 2022 and filters it for quarterback-specific stats. It loops through each season, processes the data to extract relevant columns (e.g., passing yards, touchdowns, interceptions), and combines the results into a single dataframe. The final dataset provides a comprehensive view of quarterback performance across multiple seasons.

```
## Loading data for season: 2000
## Loading data for season: 2001
## Loading data for season: 2002
## Loading data for season: 2003
## Loading data for season: 2004
## Loading data for season: 2005
## Loading data for season: 2006
## Loading data for season: 2007
## Loading data for season: 2008
## Loading data for season: 2009
## Loading data for season: 2010
```

Processing and Cleaning Play-by-Play Data

This section loads and cleans play-by-play data for each season, filtering for relevant columns and quarterback-specific plays, while categorizing weather conditions.

```
## Processing data for season: 2000
## Processing data for season: 2001
## Processing data for season: 2002
## Processing data for season: 2003
## Processing data for season: 2004
## Processing data for season: 2005
## Processing data for season: 2006
## Processing data for season: 2007
## Processing data for season: 2008
## Processing data for season: 2009
## Processing data for season: 2010
## Processing data for season: 2011
## Processing data for season: 2012
## Processing data for season: 2013
## Processing data for season: 2014
```

Analyzing Play Types and Quarterback Performance

Analyzes the relationship between play types, quarterback performance, and weather conditions, categorizing plays based on weather types.

```
## Analyzing play types and quarterbacks for season: 2000
## Analyzing play types and quarterbacks for season: 2001
## Analyzing play types and quarterbacks for season: 2002
## Analyzing play types and quarterbacks for season: 2003
## Analyzing play types and quarterbacks for season: 2004
## Analyzing play types and quarterbacks for season: 2005
## Analyzing play types and quarterbacks for season: 2006
## Analyzing play types and quarterbacks for season: 2007
## Analyzing play types and quarterbacks for season: 2008
## Analyzing play types and quarterbacks for season: 2009
## Analyzing play types and quarterbacks for season: 2010
## Analyzing play types and quarterbacks for season: 2011
## Analyzing play types and quarterbacks for season: 2012
## Analyzing play types and quarterbacks for season: 2013
## Analyzing play types and quarterbacks for season: 2014
```

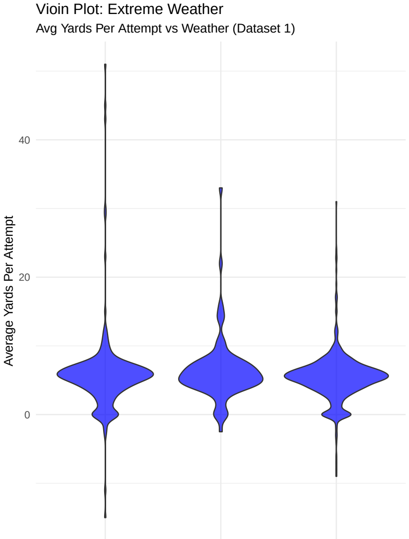
Regression Analysis and Scatterplots

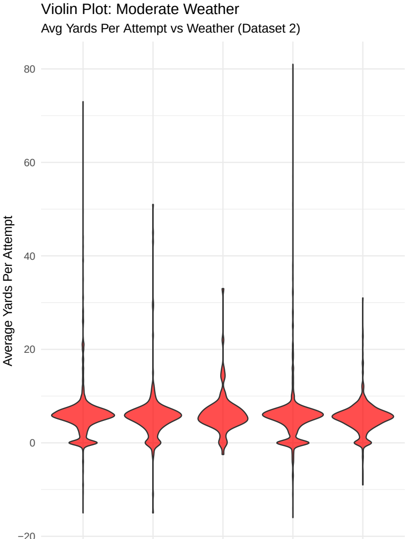
Performs regression analysis to explore the impact of weather conditions on average yards per attempt, complemented by scatterplots with regression lines for visualization.

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine

##
## Call:
## lm(formula = avg_yards_per_attempt ~ weather_condition,
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.756  -1.528  -0.008   1.213  45.244
##
## Coefficients:
```

Violin Plots for Weather Impact

The code below creates violin plots to compare the distributions of average yards per attempt across weather conditions, highlighting differences between datasets for extreme and moderate weather.


Vioin Plot: Extreme Weather
Avg Yards Per Attempt vs Weather (Dataset 1)


Violin Plot: Moderate Weather
Avg Yards Per Attempt vs Weather (Dataset 2)

Conclusion

Part 2

The violin plots show that weather conditions generally have minimal impact on average yards per attempt, as the distributions overlap significantly across categories. Rain stands out slightly, with lower performance compared to other conditions, which aligns with earlier regression results showing a modest negative effect. Overall, weather seems to play a secondary role, with other factors like team or player performance likely having a greater influence. The regression results confirm that weather has little impact on average yards per attempt, with both models showing very low explanatory power (adjusted R-squared values near zero). While Rain in Dataset 2 shows a small, statistically significant negative effect, other weather conditions are not significant. This aligns with the violin plots, highlighting that performance is likely influenced more by other factors like team or player quality than by weather. These findings are consistent with the conclusion that while extreme weather may modestly impact win rates, the overall effect of weather is secondary to other factors, suggesting the need for additional variables in future models. Although there was no

Challenges and Solutions

While working on this project, I encountered several challenges, including managing large datasets, integrating data sources, and optimizing R code. Working with play-by-play data across multiple NFL seasons and combining it with weather data resulted in long run times. To address this, I filtered out irrelevant data early on, cached intermediate results, and used tools like dplyr to streamline processing. Finding and merging compatible weather data also proved difficult due to mismatched formats, but I resolved this by standardizing fields such as team names and weather descriptions.