MIS 548 Milestone Report
Group 9: Zachary Newland, Siddarth Gopalakrishnan, Sean Valerga, Pooja Veenu, Xiao Liang
Date: December 18th, 2023
Dataset Selection: NFL play-by-play data

Description:
NFL play-by-play dataset provides an overview of player stats among players, injuries, QBR, depth charts, contracts, and snap counts. Data such as player stats can relate to completions, passing yards, sacks, fumbles, rushing yards, rushing first downs, etc. There is a high variation of data available that may rely heavily on pandas to summarize the dataset. This is accessed through the nflreadr package and provides data for each season selected.

The NFL play-by-play dataset is organized into three tables: players, plays, and teams, featuring meticulously structured data. Each table boasts an extensive collection, exceeding one thousand records for every variable. The variables exhibit diverse data types, including integers, floats, objects, and more. After the data cleaning, our dataset will exhibit uniform table sizes, devoid of any missing values.

Our project is dedicated to refining the NFL play-by-play dataset and performing a comprehensive analysis to address specific business problems. As we progress through this endeavor, our objective is to cultivate an understanding of the intricacies involved in the data wrangling process. Moreover, we aim to proficiently implement Python code for essential tasks, including data cleaning, preprocessing, and thorough analysis.

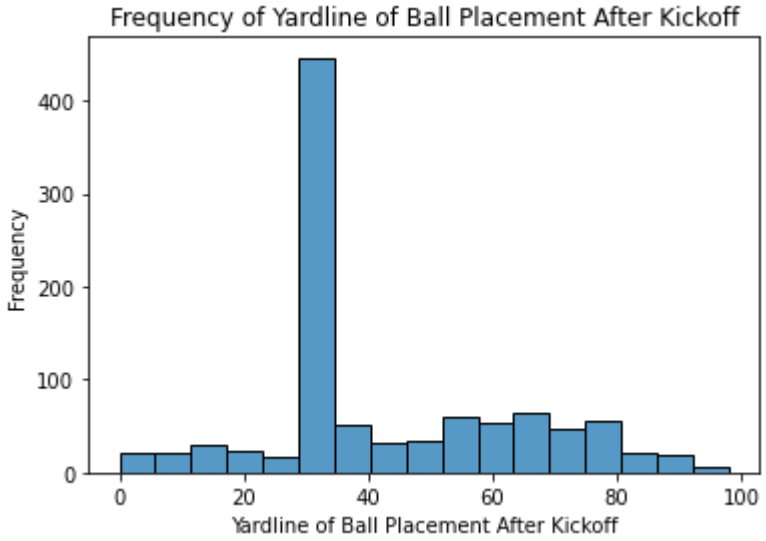
Initial Considerations:
Player and plays datasets are both susceptible to holding missing values. The following table indicates missing values among the Player dataset that will either require dropping or adusting:

Player Dataset	
Column	**Missing Values**
middleName	138
status	306
gsisId	1
homeTown	33
jerseyNumber	29

Further analysis within the play datasets illustrates a greater concern of missing variables. Among the 306 unique columns within the play dataset, there are 238 columns missing values compared to the 68 columns that contain full observations. There will be further analysis to determine feature importance, dropped columns, and input regarding dropping missing values or filling missing data with appropriate information.

Aside from the inconsistencies among integrity among specific feature columns, the dataset may be prone to outliers that should be further evaluated within the analysis. For example, the dataset plays provide observations among ball placement against yard lines for teams over the years

1999 – 2017. If plotting frequency of occurrences, we can see there is a significant number of outliers observed among the 25 to 30-yard line:



The frequency of this occurrence is likely contributed to the fair catch rule which results in a ball placement at the 25-yard line to prevent player injury. Further observations may illustrate similar findings dependent on the NFL rules and result in a higher frequency of occurrences.

Business Problem:

Over the course of cleaning each dataset, there are a few research questions the team would like to further pursue as areas of interest. The following are future considerations to explore as scope is expanded upon:

- Are there any correlations between player attributes (e.g., height, weight) and performance?
- Do higher play counts during a drive result in touchdown scores?
    - Can scoring a touchdown, field goal, or safety be classified and predicted based on feature selections.

Milestones:

Sprint	Task	Deliverables
Dec 18th - Dec 23rd	- Identify missing data outliers and inconsistencies	Code and document
Dec 26th - Dec 29th	- Data cleaning and transformation	Code and document
Jan 2nd - Jan 7th	- Data integration - Handling missing data - Exploratory data analysis	Code and document
Jan 8th - Jan 13th	- Create 12 minutes presentation video. - Finish 5 pages summary report.	Code, document, Slides and Presentation