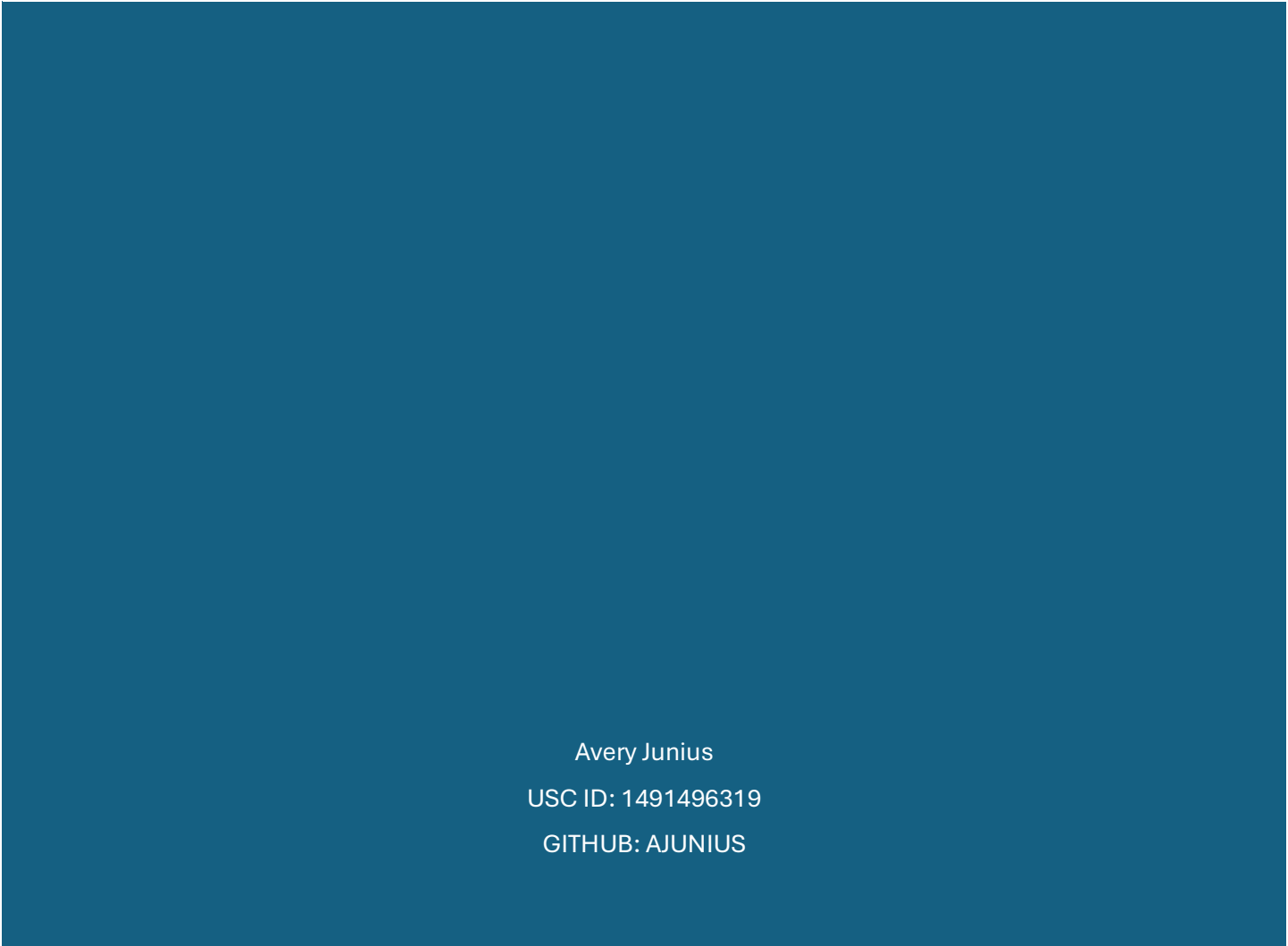


ANALYSIS OF NFL QUARTERBACK PERFORMANCE ON TEAM SUCCESS



Avery Junius
USC ID: 1491496319
GITHUB: AJUNIUS

Table of Contents

Introduction.....	2
Metrics Used.....	2
Data Collection	3
Data Source	3
Approach.....	3
Data Collected.....	3
Deviations, Challenges and Resolutions.....	3
Data Cleaning	4
Data Changes	4
Raw Data Format.....	4
Processed Data Format	5
Impact of the Cleaning Process.....	5
Data Analysis & Visualizations	5
Analysis Techniques.....	5
Findings	6
Visualizations	7
Data Visualization Types.....	7
Explanation of Visualizations	7
Conclusion	8
Observations	8
Impact of Findings	9
Future Work	9
Improvements to the Project.....	9
Data Source Evaluation	9

Introduction

The quarterback position for American football is one of the most difficult positions in all of sports. The game of football is similar to a chess match in which a plethora of variables present themselves before and during each play in which the quarterback must adapt to.

This analysis aims to answer three questions utilizing publicly available data from ProFootballReference.com:

1. Which quarterback performance metrics have the highest correlation with team success?
2. How do playoff teams compare statistically to non-playoff teams across various quarterback metrics?
3. Do dual-threat quarterbacks and their rushing statistics contribute significantly to quarterback performance evaluation?

Metrics Used

1. **Passing Yards:** The total number of yards gained by the quarterback through completed passes during the season.
2. **Passing Touchdowns (TDs):** The total number of touchdowns scored by the quarterback through passing.
3. **Interceptions:** The total number of quarterback's passes that were intercepted by the opposing team.
4. **Quarterback Rating (Rating):** A statistical measure developed by ESPN in 2011 that incorporates all a quarterback's contributions to winning, including how he impacts the game with passes, rushing, turnovers, and penalties to evaluate quarterback performance.
5. **4th Quarter Comebacks (4QC):** The total number of times the quarterback has led to a successful offensive drive in the 4th quarter to tie or take the lead in the game.
6. **Game-Winning Drives (GWD):** The total number of times the quarterback has orchestrated an offensive drive resulting in the game-winning score.
7. **Rushing Yards:** The total number of yards gained by the quarterback while running with the ball.
8. **Rushing Touchdowns:** The total number of touchdowns scored by the quarterback through rushing.
9. **Total Yards:** The sum of passing and rushing yards, representing the overall yardage contribution of the quarterback.
10. **Total Touchdowns (Total TDs):** The combined total of passing and rushing touchdowns.
11. **Passing TD to Interception Ratio:** The ratio of passing touchdowns to interceptions, reflecting the quarterback's efficiency in scoring versus turnovers.

These metrics provide a comprehensive view of a quarterback's contributions, combining traditional passing statistics with rushing performance and situational success indicators like 4QC and GWD.

Data Collection

Data Source

Data was collected from *Pro Football Reference*, which is a comprehensive source for current and historical NFL players and teams. Three main sources, corresponding to passing stats, rushing stats, and NFL team standing tables were used for data extraction.

Approach

Separate scraping functions were implemented for passing and rushing statistics to ensure data completeness. *Main.py* is the program that accesses the data collection and cleaning functions.

The following Python libraries were utilized for web scraping techniques:

- *requests* – used to send HTTP requests to *Pro Football Reference* website and retrieve quarterback statistics, NFL team standings and playoff outcomes for specific years
- *BeautifulSoup* - used for parsing and navigating the HTML structure of the specified web page to extract specific data elements from the table
- *time* – used to add capability for delays of HTTP requests
- *random* – used to randomize the time delays between HTTP requests to avoid rate-limiting from data source
- *pandas* – used for data manipulation and storage. Extracted data from *Pro Football Reference* can be structured into data frames and exported into CSV files.

Data Collected

- Data spans quarterback performances from 2013, 2021, and 2022 NFL seasons who played at least 10 games for the respective season.
- Playoff results for each season were collected and appended to the quarterbacks' statistics.
- Total data collected includes statistics for 144 quarterbacks after data cleaning.

Deviations, Challenges and Resolutions

The original plan for this analysis was to select a single season and compare playoff quarterback performance to non-playoff quarterback performance. However, collecting data from a single season limits the strength of the hypothesis that great quarterback play is correlated with team success. The capability to scrape the same source for multiple years was implemented. Another aspect of this project that was approved upon was the inclusion of rushing statistics for quarterbacks. NFL defenses have become more complex, and the physical talent of defenders has improved. Quarterbacks that can affect the game with not only their passing ability, but also their running ability have seen an increase in value. The capability for scraping for rushing statistics and appending these to the quarterbacks' passing statistics was implemented as well. Below are additional challenges faced when developing the python programs for data scraping and analysis:

- **Challenge:** Inconsistent team names due to name and location changes for past seasons. This caused issues with determining playoff status for certain quarterback(e.g., *Washington Redskins* to *Washington Commanders* or *San Diego Chargers* to *Los Angeles Chargers*).

Resolution: Implemented a mapping dictionary to standardize team names. Historic team names were mapped to the current team's name for consistency.

- **Challenge:** Handling missing data for quarterback ratings.

Resolution: Players with missing quarterback rating (QBR) were excluded. These quarterbacks (3) were not eligible for the playoffs in those years. This data was missing due to the small sample size of the quarterback in that specific season.

- **Challenge:** Too many requests to data source (429 errors).

Resolution: Testing via my home network consistently led to rate-limiting errors where attempts to collect data from *Pro Football Reference* were blocked. The capability of delaying HTTP requests was implemented to alleviate this issue.

Data Cleaning

The *clean_data.py*, ensures the quality and consistency of quarterback statistics by handling missing values, standardizing formats, and preparing the dataset for analysis. This step was crucial due to the inconsistencies and irregularities in the raw data collected from *Pro Football Reference* HTML Tables.

Data Changes

To make the data usable for analysis, the following transformations were applied:

1. **Cleaning Text Data:**
 - Removed annotations (e.g., text within parentheses or brackets) using regular expressions.
 - Stripped extra whitespace and converted numeric strings to integers or floats. For instance, commas in large numbers were removed to allow for proper conversion.
 - Invalid or missing values were replaced with None to handle conversion errors.
2. **Standardizing Team Names:**
 - Team names were inconsistent, often using abbreviations or outdated names (e.g., "SDG" for "Los Angeles Chargers").
 - A mapping dictionary (*TEAM_NAME_MAPPING*), stored in the *config.py* file, was used to standardize team names. Any unmapped team names were logged for debugging.
 - Missing team names were matched with "Unknown" to maintain data integrity and avoid errors during the analysis phase.
3. **Handling Missing or Invalid Data:**
 - Rows with missing values for (Rating) were excluded from the dataset.
 - Missing or invalid values in other columns were replaced with default values to avoid removing potentially valuable data and present debugging opportunities.

Raw Data Format

The data was initially collected from *Pro Football Reference* as HTML tables. These tables included various irregularities:

- Columns with annotations or non-standard formatting.

- Numeric values stored as strings with commas.
- Team names in abbreviated format or historic formats.
- Missing data for certain metrics, such as quarterback ratings.

Processed Data Format

Following the cleaning of the data it was exported to a standardized CSV file. Key characteristics of the processed data format include:

- Cleaned and consistent numeric columns (e.g., Passing Yards, Passing TDs, Interceptions) converted to integers or floats.
- Standardized team names stored in a new column, Standardized Team.
- Rows with missing critical data (e.g., Rating) removed.
- All other missing or invalid values replaced with default or calculated values.

Impact of the Cleaning Process

The cleaning process had the following effects on the data:

1. **Improved Consistency:** Standardizing team names allowed for accurate grouping and comparison of players by playoff status.
2. **Enhanced Usability:** Cleaning numeric columns and handling missing values ensured that the dataset could be used directly for statistical analysis without further preprocessing.
3. **Impactful Data:** Limiting the data scraping to quarterbacks that have played at least 10 games, shrunk the amount the data pulled, showed meaningful correlations between data, and limited abnormalities.

These cleaning steps ensured that the data was reliable and ready for subsequent analysis, minimizing potential biases or inaccuracies introduced by inconsistencies in the raw data.

Data Analysis & Visualizations

Analysis Techniques

Data analysis consisted of summarizing the data collected from *Pro Football Reference*, correlating these statistics, and hypothesis testing to uncover insights into quarterback performance metrics and their correlation to team success. `Run_analysis_visulation.py` is the program that accesses the data analysis and visualization functions for the project.

1. **Descriptive Statistics:**
 - Summarizes quarterback performance metrics and exported to `descriptive_stats.csv`. By year, key statistics are compiled and calculated with a mean, standard deviation, minimum, and maximum values.
 - Passing Yards, Passing TDs, Interceptions, Rating, Rushing Yards, Total Yards, and Passing TD to INT Ratio were evaluated.
 - This data can be used to compare specific quarterback statistics with the pool of data collected.
2. **Correlation Analysis:**

- Correlation matrices were generated for each year to assess the relationships between performance metrics.
 - Visuals include heatmaps for these correlations which reveal strong relationships between key metrics.
3. **Playoff vs. Non-Playoff Comparisons:**
- T-tests compare the means of two separate groups. It evaluates whether the difference between playoff and non-playoff quarterbacks is likely due to random chance or if it reflects a true difference in the population. The difference is referred to as t-stat.
 - P-values show the probability of obtaining a certain result.
 - A p-value < 0.05: Indicates statistically significant evidence to reject the null hypothesis.
 - A p-value ≥ 0.05: Fails to reject the null hypothesis (no significant evidence of a difference).

Findings

- **Descriptive Statistics:**
 - Playoff quarterbacks generally had higher Passing Yards, Passing TDs, and Total Yards than the average compared to their non-playoff counterparts.
 - Passing TD to INT Ratio also tended to be higher among playoff quarterbacks, reflecting greater scoring efficiency and ball security .
- **Correlation Analysis:**
 - **Strong Correlations:**
 - **Passing Yards & Total Yards** -Consistently above 0.9. This is expected as a quarterback's primary method of gaining yards is via passing.
 - **Passing TDs & Total TDs** – Correlation shows that passing touchdowns are the primary method for a quarterback's scoring contribution.
 - **4QC (4th Quarter Comebacks) and GWD (Game-Winning Drives):** Moderate correlations with Rating, highlighting the significance of situational performance.
 - **Negative Correlation:**
 - **Interceptions** showed a negative correlation with Rating, confirming that higher interception rates reduce quarterback efficiency.
 - **Playoff vs. Non-Playoff Comparisons**
 - **Key Metrics with Significant Differences (p < 0.05):**
 - **Passing Yards (t_stat=4.345, p_value=0.000):** Playoff quarterbacks tend to achieve higher passing yard totals compared to non-playoff quarterbacks.
 - **Passing TDs (t_stat=5.500, p_value=0.000):** Playoff quarterbacks scoring significantly more touchdowns.
 - **Rating (t_stat=6.441, p_value=0.000):** Rating is a strong indicator of playoff success, with higher values associated with playoff quarterbacks.
 - **Rushing TDs (t_stat=2.685, p_value=0.009):** Indicates the value of dual-threat quarterbacks. Quarterbacks on playoff teams are more effective in scoring rushing touchdowns.
 - **Passing TD to INT Ratio (t_stat=4.003, p_value=0.000):** Playoff quarterbacks show greater efficiency with higher ratios, reinforcing the principle of limiting turnovers while maximizing scoring opportunities.
 - **Total Yards (t_stat=4.868, p_value=0.000) and Total TDs (t_stat=6.303, p_value=0.000):** Playoff quarterbacks contribute to

offensive efficiency regarding total yards and touchdowns more than their counterparts.

- **4QC (t_stat=2.789, p_value=0.006) and GWD (t_stat=2.742, p_value=0.007):** Playoff quarterbacks excel in clutch moments, driving their teams to success in crucial games.
- **Metrics Without Significant Differences ($p \geq 0.05$):**
 - **Interceptions (t_stat=-1.622, p_value=0.108):** No significant difference in interception rates between playoff and non-playoff quarterbacks.
 - **Rushing Yards (t_stat=1.491, p_value=0.140):** Playoff quarterbacks tended to rush for more yards, but the difference was not statistically significant.

Visualizations

Data Visualization Types

1. **Boxplots:** Used to compare the distributions of key metrics, such as Passing Yards and Passing TD to INT Ratio, between playoff and non-playoff quarterbacks.
 1. A **median line** is featured within each box, assists in comparing tendencies of playoff vs non-playoff teams. Datasets are divided into two halves
 2. The top and bottom edges of the box represent the **75th percentile (upper quartile)** and **25th percentile (lower quartile)** of the data, respectively.
 3. The distance between the top and bottom of the box is known as the **Interquartile range (IQR)**, measuring the spread of the middle 50% of the data. A larger IQR means there's increased variability within the data.
 4. The top and bottom lines are known as **whiskers** and show the minimum and maximum values within 1.5 times the IQR, capturing most of the data excluding outliers.
 5. Outliers
 1. The circle symbols plotted outside of the whiskers represent data that deviates greatly from the dataset it belongs to.
2. **Correlation Matrix Heatmaps:** Utilized for visualizing correlation matrices, highlighting the strength and direction of relationships between quarterback performance metrics.

Explanation of Visualizations

1. **Passing_Yds_Comparison.png: Fig. 1**
 - The boxplot shows that playoff quarterbacks tend to have higher median Passing Yards, with a wider distribution compared to non-playoff quarterbacks.
2. **Passing_TDs_Comparison.png: Fig. 2**
 - Quarterbacks on playoff teams throw more Passing TDs, with a statistically significant difference in distributions.
 - A non-playoff quarterback, Justin Herbert for the Los Angeles Chargers, was shown as an outlier, throwing 38 passing touchdowns. This specific datapoint is recorded in *qb_combined_stats_with_playoff_status.csv*. His team had the number 29 / 32 ranked defense in the league that year which is a likely a factor in missing the playoffs.
3. **Passing_TD_to_INT_Ratio_Comparison.png: Fig. 3**
 - Playoff quarterbacks show greater efficiency, reflected in higher TD to INT Ratios compared to their non-playoff counterparts. Less variance in each dataset indicates efficient scoring relative to turnovers is a strong indicator of team success.

- Outliers on each side exist in each dataset, suggesting extreme performances by some quarterbacks, either positively or negatively.
- 4. **4QC_Comparison.png and GWD_Comparison.png: Fig. 4-5**
 - Playoff quarterbacks demonstrate slightly higher median values for 4QC and GWD, suggesting better situational performance under pressure.
- 5. **Total_Yards.png and Total_TDs_Comparison.png: Fig. 6 -7**
 - Playoff quarterbacks significantly outperform in Total Yards and Total TDs, as evidenced by the higher median and broader distributions. These metrics emphasize their overall contribution to team success. The median line for Total TDs highlights the stark contrast between playoff and non-playoff quarterbacks.
- 6. **Rating_Comparison.png: Fig. 8**
 - The boxplot reveals a clear distinction in quarterback ratings between playoff and non-playoff quarterbacks. Playoff quarterbacks have significantly higher median ratings, with the upper quartile extending well beyond that of non-playoff quarterbacks.
 - The presence of an outlier in the playoff group indicates a lower-performing quarterback who reached the playoffs, potentially due to strong team support or exceptional defense.
- 7. **Correlation Matrix Heatmaps (2013, 2021, 2022) – Fig. 9 - 11**
 - **Strong Positive Correlations**
 - **Passing Yards & Total Yards:** Demonstrates that passing efficiency remains a critical driver of total offensive yardage. Passing yards is a primary method in which quarterbacks affect an offense.
 - **Passing TDs & Rating:** Indicates that successful scoring via passing correlates with higher ratings.
 - **Total Yards & Total TDs:** Highlights the overall offensive contributions of quarterbacks.
 - **Weak Correlations**
 - Interceptions showed inconsistency across season, lending to differences in team scheme and quarterback risk tolerance.
 - These matrices show stability year-to-year between the metrics extracted and the relationships between them.

Conclusion

Observations

This analysis give emphasis to the critical role quarterback performance plays in determining team success in the NFL. The findings reveal that playoff quarterbacks consistently outperform their non-playoff counterparts across several key metrics, including Passing Yards, Passing TDs, Rating, Total Yards, and Total TDs. 4th Quarter Comebacks (4QC) and Game-Winning Drives (GWD) are metrics that highlight the importance of situational performance, specifically in high-pressure moments. Furthermore, the Passing TD to INT Ratio emerged as a strong indicator of team success, reinforcing the importance of balancing scoring efficiency with turnover prevention. Turnovers provide the opposing team with increased opportunities to score within a game, putting the offense and team at a disadvantage.

The integration of rushing statistics, such as Rushing TDs, highlighted the growing significance of dual-threat quarterbacks who contribute not only through passing but also with their running ability. This trend reflects the evolving nature of the quarterback position becoming more dynamic in response to modern defensive complexities and improved physical abilities of defensive players.

Considering the above observations and recognizing there are multiple factors that can impact a football game such as offense, defense, and special teams the quarterback's performance is often the greatest influence on a given matchup. A quarterback's ability to excel in passing, handle pressure, and contribute as a dual-threat player significantly influences team success and subsequently playoff aspirations.

Impact of Findings

The findings have broader implications for evaluating quarterbacks and team-building strategies in professional football. By identifying the metrics most correlated with team success, teams can focus on scouting and developing quarterbacks who excel in these areas from the college level. Additionally, this analysis provides a statistical foundation for evaluating the impact of dual-threat quarterbacks, emphasizing the value of adaptability and versatility in today's NFL. Teams will increasingly value quarterbacks who can excel in multiple dimensions of the game.

Future Work

Improvements to the Project

Given more time, there are a few improvements I would implement for this project. First, I would expand the dataset to more seasons to strengthen the validity of findings and present more trends supporting the correlation between quarterback performance and team success. Secondly, I believe incorporating advanced football metrics such as Expected Points Added (EPA) or Completion Percentage Over Expected (CPOE) could provide deeper insights into quarterback performance. Additionally, adding context for other variables within a football game that affect quarterbacks such as defensive team rankings and offensive line protection would isolate a quarterback's individual impact. Lastly, I would research how to implement interactive visualizations for the data to be more engaging.

Data Source Evaluation

Pro Football Reference was a reliable, historic data source that is well structured for data scraping and manipulation. However, I would consider additional data sources such as NFL Next Gen to add additional context to a quarterback's passing profile such as time to pass and completion probability. Also, to address rate-limiting issues during data collection, I would explore the possibility of using an API if the capability is there. Rate-limiting hindered the development of my code since testing too much in succession would lead to my home network being blocked from extracting data.

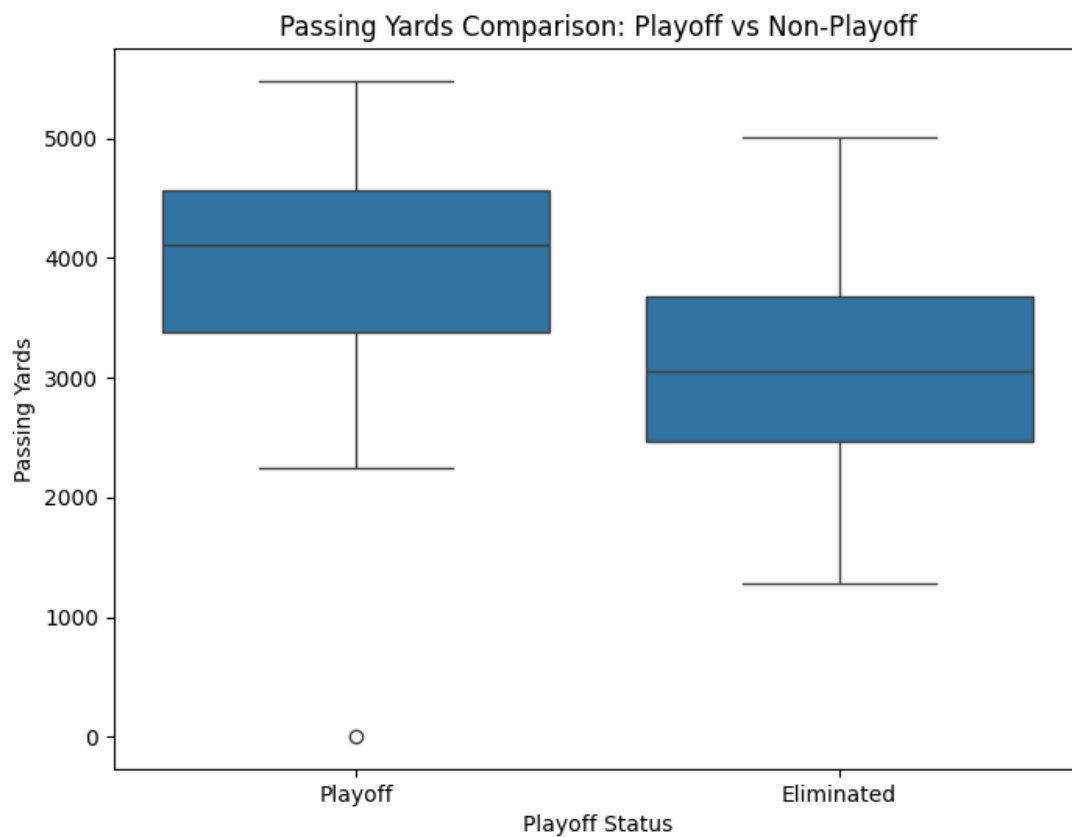


Figure 1

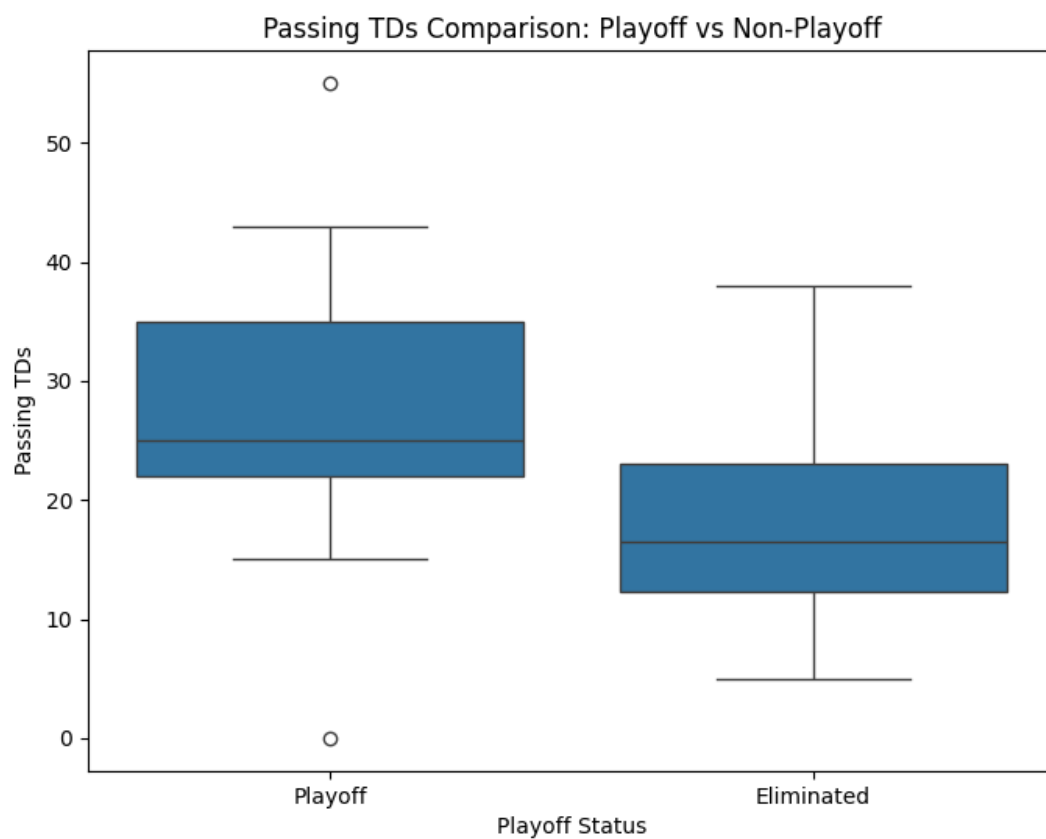


Figure 2

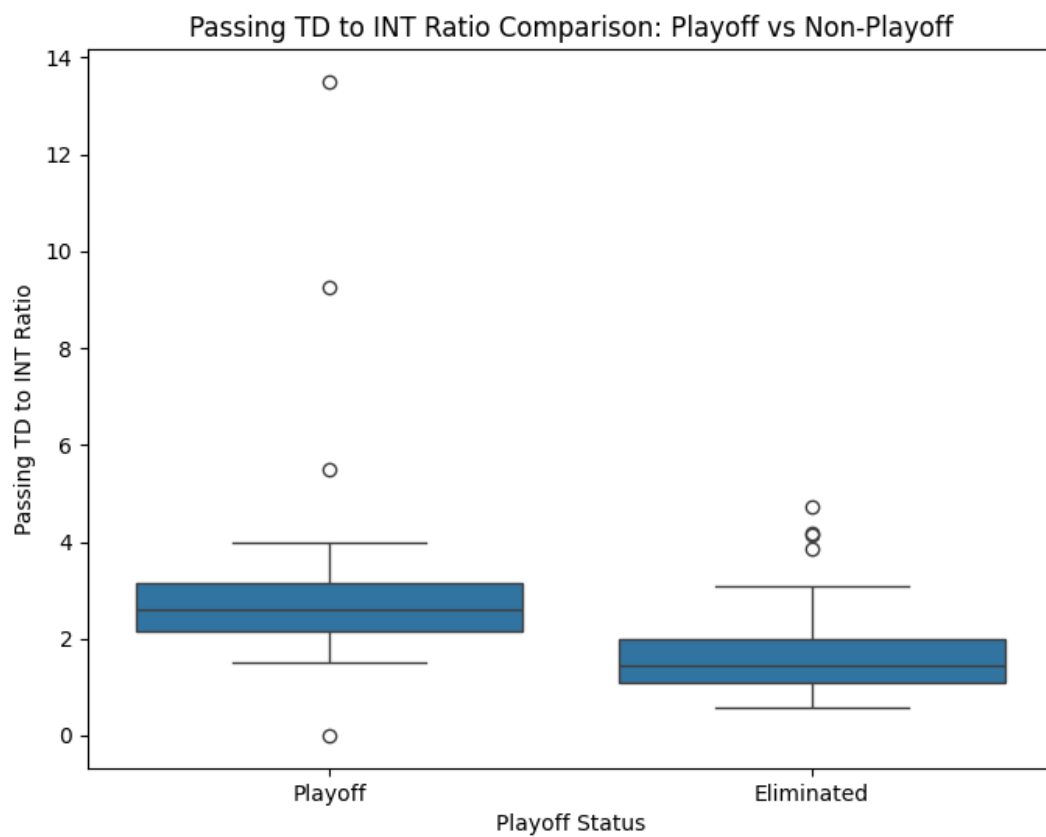


Figure 3

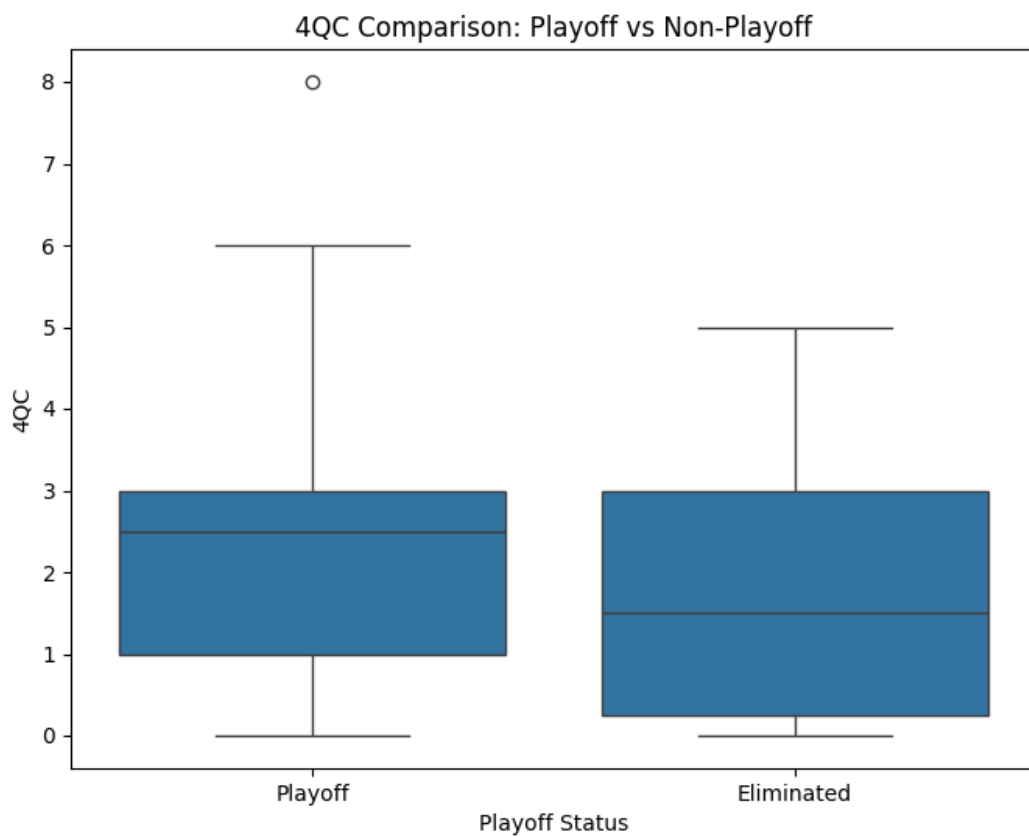


Figure 4

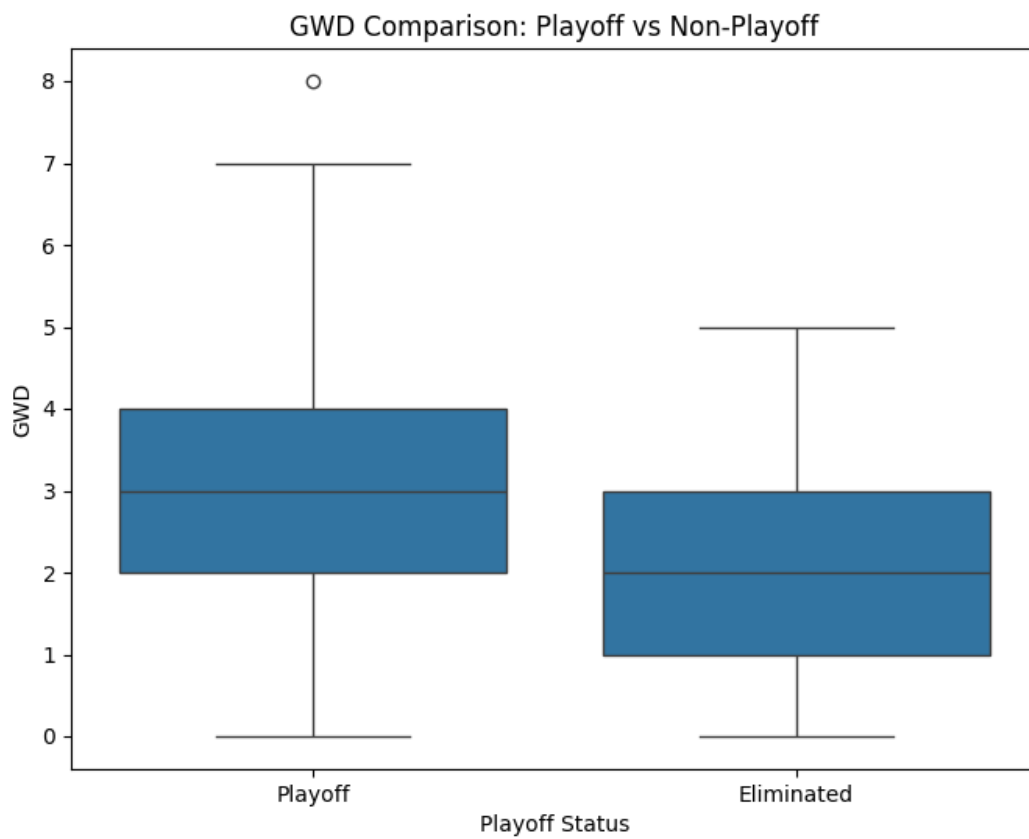


Figure 5

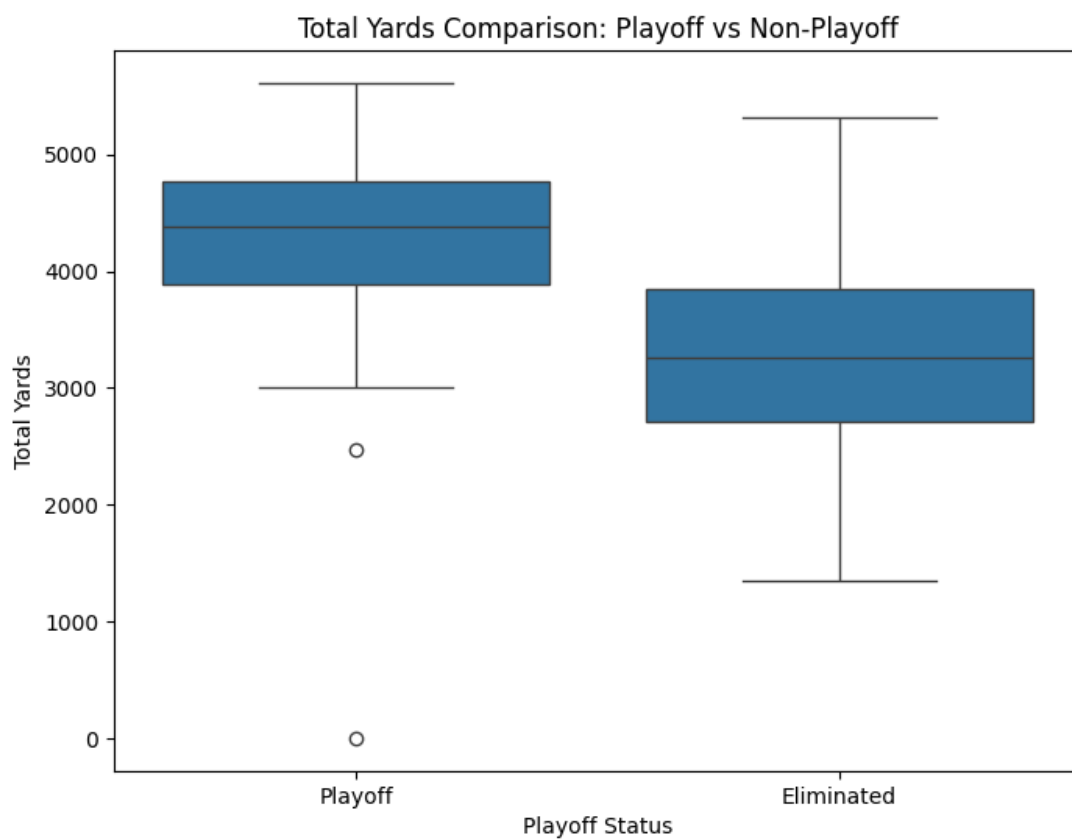


Figure 6

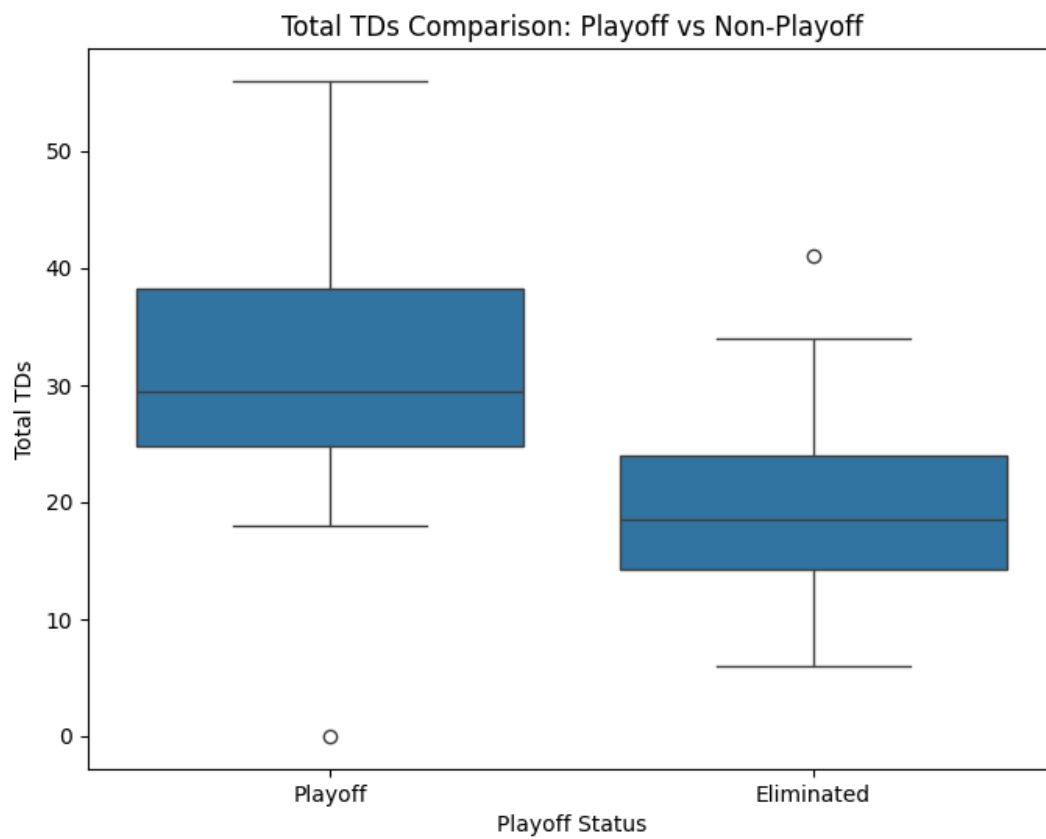


Figure 7

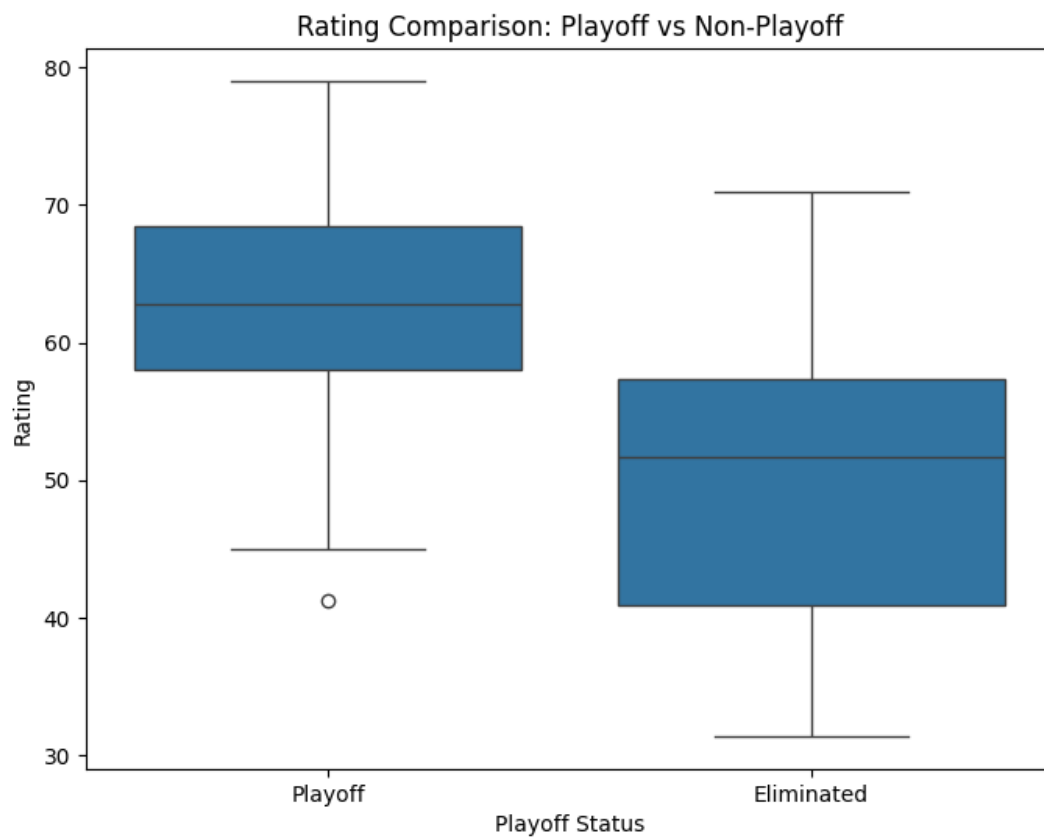


Figure 8

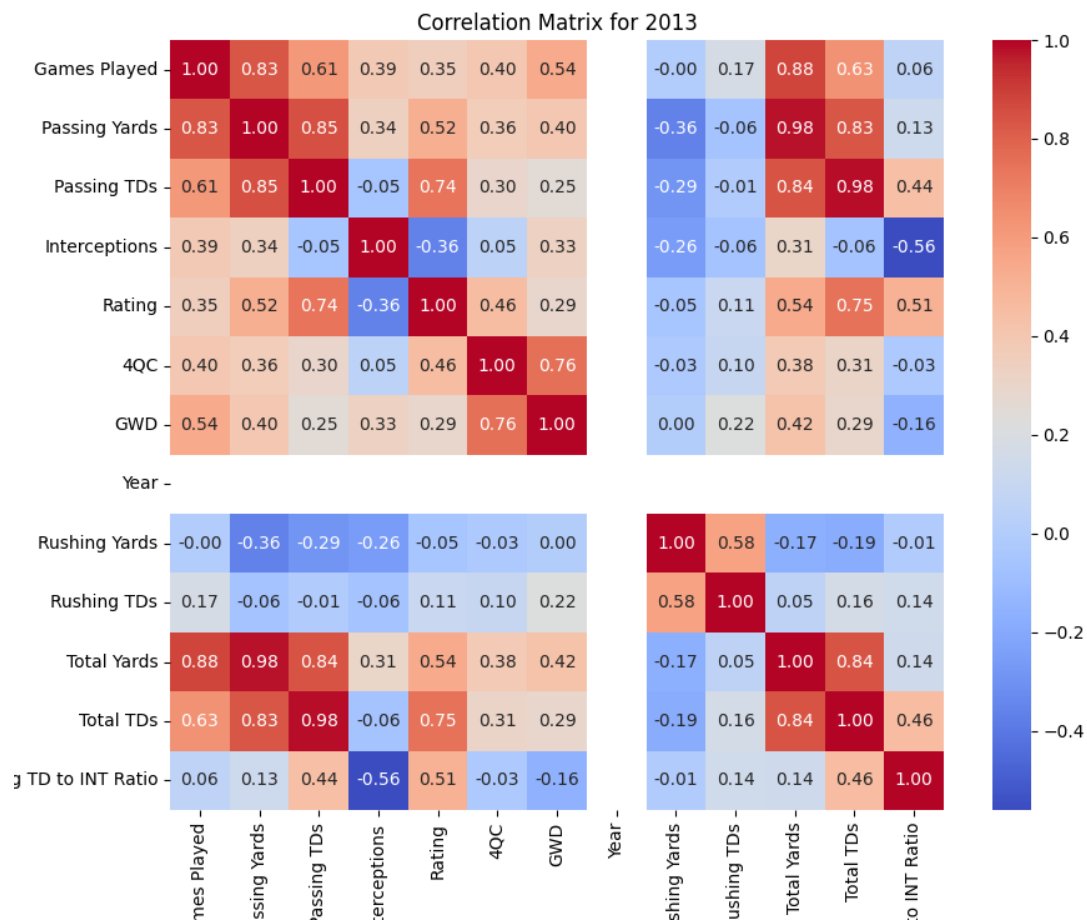


Figure 9

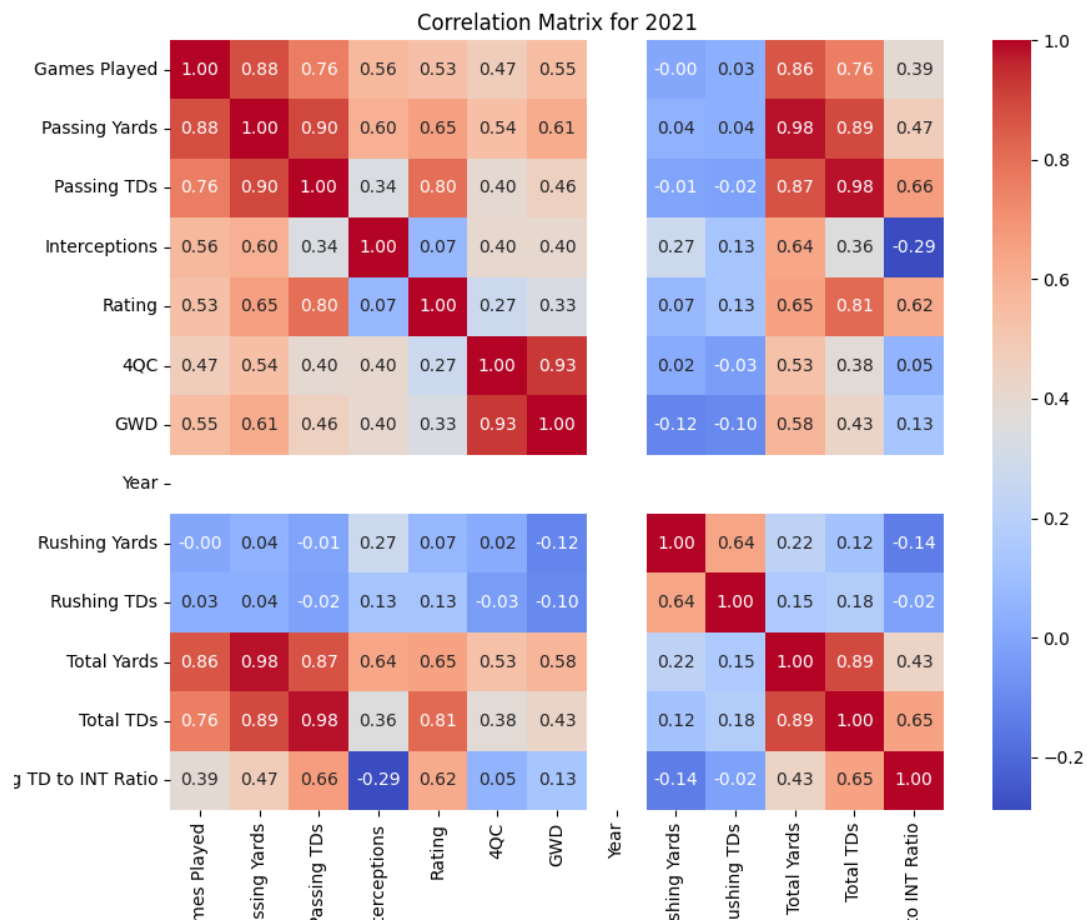


Figure 10

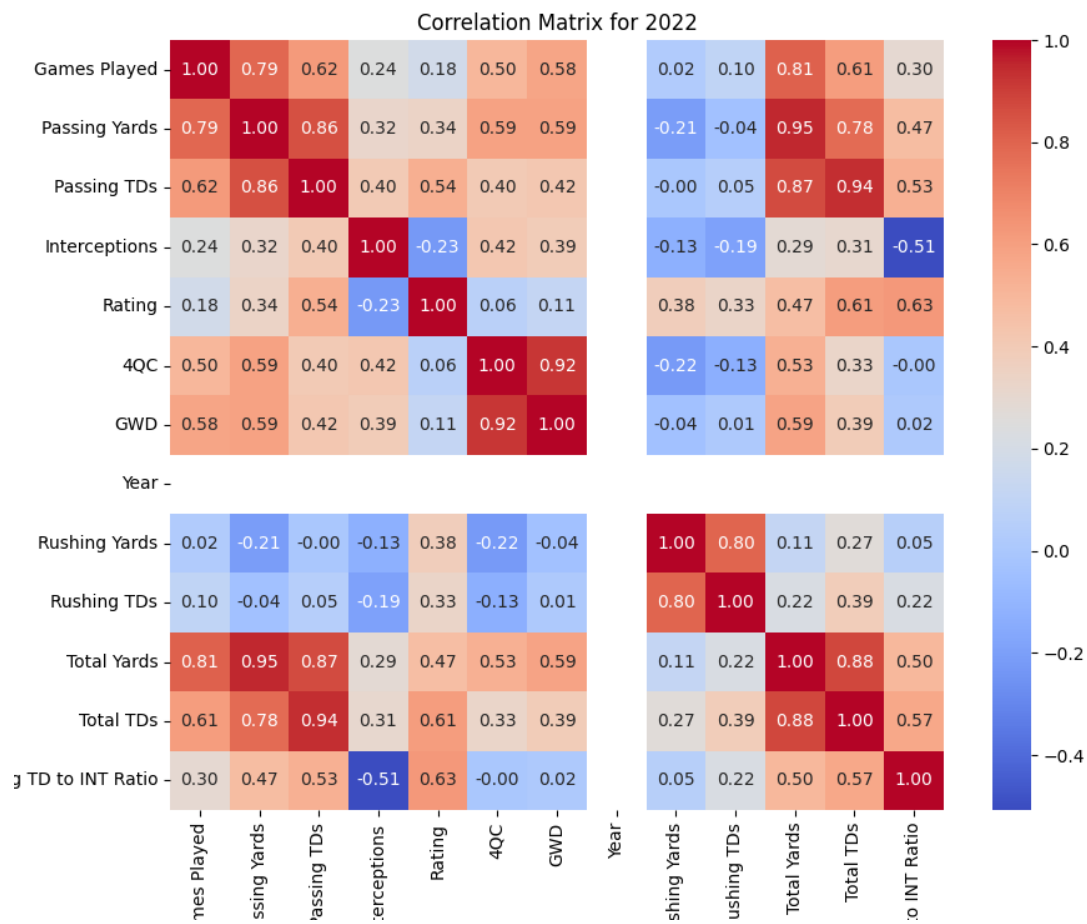


Figure 11