Introduction

American football is a team sport in which two teams of 11 players compete to score points by advancing the ball into the opponent's end zone through passing or running plays. During play, one team is on offense, while the other team is on defense and each team typically has separate offensive and defensive units which swap when a change of possession occurs. Like in regular football (what they might call soccer), statistical analysis is now a large aspect of managing teams. As such, lots of data is collected and deeply analysed to optimise team composition and strategy. In particular, the performance of the key offensive positions of quarter backs (QB), wide receivers (WR), running backs (RB) and tight ends (TE) are of importance because they are deemed to be the most influential in the outcome of games.

In this report, data is scraped from the statistics website pro-football-reference.com and MDS is performed to analyse similarities and patterns among key offensive players using performance based statistics collected in the 105th season of the national football league (NFL) (September 5, 2024 - January 5, 2025) so that data driven decisions can be made.

Data

Pro-Football-Reference (PFR) is a well known database for yearly American football statistics from the NFL. That data is primarily collected from official game logs provided by the league and contains a huge variety of statistics. As we are interested in the performance of key offensive players (QBs, WRs, RBs and TEs), we need to isolate relevant statistics that are available for these players. The main statistics used by analysts when analysing these offensive players are passing, rushing, receiving and fumble statistics as these are the direct measurements of the offensive unit. From each of these statistical categories, a select few metrics, commonly used by fantasy football players and professional scouts, were chosen and are presented in table (1). This data can be found on PFR and was extracted using an `R` package called `profootballref` which gives statistics for 623 players (N = 623).

Cleaning and Preprocessing

A common challenge in data analysis is the non-homogeneity of data and the presence of outliers, both of which can distort results. Ensuring high-quality data is therefore essential.

One issue encountered was missing statistics for some players. To ensure consistency, players with missing values were removed from the analysis. While ordinal MDS can accommodate missing data, we chose to restrict the analysis to complete cases due to the dataset's large size. Another concern was players with few appearances, as their statistics might not be representative of their overall performance. To address this, only players who participated in more than 10 games were included in the analysis. Also, we see that some features of the dataset might not directly measure or relate to the performance of a player such as age or rank. Furthermore, some data points are derived metrics based on other considered datapoints such as rushing yards per attempt and if we considered such features it might skew the results by overemphasising particular qualities. Finally, we observe that a few entries even to this point have fully 0 values for each variable and that some players have identical stats because they are low. This could be ascribed to many reasons such as error in collection, error in communication or that they genuinely might have 0 in their stats. To fix all these issues, the dataset was pruned and we are finally left with the 15 variables depicted in table (1) and 363 data points.

Additionally, the dataset contained features on different scales, such as count variables (e.g., completions) and continuous variables (e.g., passing yards). To control the impact, each variable was standardised by subtracting the mean and dividing by the standard deviation. Since many features are count-based and right-skewed, other transformations may be more appropriate in future analyses.

Analysis

Using the `MASS` library, we can conduct MDS. First, to check the euclidean-ness and to estimate the dimensionality of the data, we can look at the eigenvalues of the B matrix and we see that there are 3 larger eigenvalues of `2.450596e+03 1.160883e+03 1.035883e+03` and the rest are small and close to 0 as shown in figure 1. As such we see that the data is approximately Euclidean but there are a few non-negative eigenvalues and that k = 3 is a good cut off for capturing the variance. As such, the 3 dimensional classical MDS was conducted using `cmdscale` and plotted in figure (2). We can see some very clear clusters forming, and QBs are all similar to each other and RBs the same. What we notice is that WRs and TEs overlap quite a bit. This is expected because in practice WRs and TEs serve very similar roles of catching passes from the QB.

Since the chosen statistics include a mix of count variables which are skewed, we do not expect the data to naturally have a Euclidean structure. Another metric which might serve as a good measure of dissimilarity is the Canberra metric because it is particularly useful for data where relative differences matter more than absolute differences. Unlike Euclidean distance, which is dominated by large-magnitude features, Canberra distance scales differences proportionally, making it better when comparing players with vastly different statistics. Additionally, since many performance metrics in the dataset are right-skewed count variables, Canberra distance prevents features with large numerical ranges (e.g., passing yards) from overpowering smaller-scale statistics (e.g., fumbles). One draw back is that canberra metric might amplify differences with low magnitudes, however, while not ideal, it is acceptable given the drastic differences in player statistics.

Hence, Canberra distance was calculated using the `dist` function. Because we removed players with 0 for all their stats and players with identical stats, we can conduct ordinal MDS using the dissimilarity matrix calculated using Canberra distance. Then, using the `isoMDS` function, ordinal MDS was done with 3 target dimensions and it converged in 17 iterations with a stress value of 10.364621. The results are plotted in figure (3). We observe a similar clustering of players based on the position they play and that members of each cluster seem to be closer using oMDS than with cMDS with Euclidean distance. As these results are hard to interpret due to the 3d nature of the data, ordinal scaling was done in 2 dimensions and converged in 12 steps with stress of 14.432823. In addition to the MDS dimensions, an additional dimension of fantasy position rank can be added. Fantasy Position Rank is a metric that ranks players relative to others at the same position based on their fantasy football performance. A lower rank indicates a better (more valuable) player among his peers. This is plotted in figure (4). In addition to being grouped by positions, we postulate that MDS dimension 2 is correlated to the rank of the player among the position in fantasy football as we see higher points (indicating higher rank) in the positive direction on MDS dimension 2 and lower points (indicating lower rank) in the negative direction of this axis.

Interestingly, lower-ranked players from different positions cluster together, suggesting that statistical similarities among underperforming players make their positions less distinguishable. This is likely due to insufficient playtime, as low-ranked players typically have fewer game-time opportunities to accumulate meaningful statistics. Although we filtered players based on the number of games played, future analyses should consider additional filtering based on total playing time or snap counts.

In summary, we observe clustering based on position and rank, which are both variables not directly included in the dataset.

Application

With these insights, we can make certain data driven decisions such as finding players with higher position rank than their neighbours in figure (4). This is useful as we potentially identify players that have a performance better than their rank. These players might be undervalued as value might be strongly correlated to rank and hence would be good purchases or hires for team managers looking to obtain good talent for a good price.

Additionally, we can identify players who play a specific position but are close to players who play another position. This is useful because we can identify players who might be more suited to play another position than the one they currently play, allowing managers to optimise positioning and thus overall performance of the team.

Limitations and Future Analysis

While there are interesting findings in this report. A heuristic improvement we can make is to have more tailored scaling regimes for each variable, rather than the current z score standardisation technique. This is because of the right skewed nature of count variables which might be distributed in a Poisson manner.

Additionally, all these variables might be correlated to the overall team performance and players on teams with stronger immeasurable qualities like player chemistry and strategy strength might lead to superior statistics of the player that is not directly related to their strength. Hence, some correction can be attempted on each of the players statistics accounting for team performance.

Table 1 Summary of Dataset Variables and Inclusion in MDS Analysis

Variable Name	Type	Included?	Missing (#)	Category
Rank	Count	No	0	Metadata
Player Name	Categorical	No	0	Metadata
Team	Categorical	No	0	Metadata
Position	Categorical	No	0	Metadata
Age	Continuous	No	0	General
Games Played	Count	Yes	0	Games
Games Started	Count	Yes	0	Games
Passes Completed	Count	Yes	0	Passing
Passes Attempted	Count	Yes	0	Passing
Passing Yards	Continuous	Yes	0	Passing
Passing Touchdowns	Count	Yes	0	Passing
Interceptions Thrown	Count	Yes	0	Passing
Rushing Attempts	Count	Yes	0	Rushing
Rushing Yards	Continuous	Yes	0	Rushing
Rushing Yards per Attempt	Continuous	No	157	Rushing
Rushing Touchdowns	Count	Yes	0	Rushing
Receiving Targets	Count	Yes	0	Receiving
Receptions	Count	Yes	0	Receiving
Receiving Yards	Continuous	Yes	0	Receiving
Receiving Yards per Reception	Continuous	No	37	Receiving
Receiving Touchdowns	Count	Yes	0	Receiving
Fumbles	Count	Yes	0	Fumbles
Fumbles Lost	Count	Yes	0	Fumbles
Total Touchdowns	Count	Yes	0	Scoring
Two-Point Conversions Made	Count	No	330	Scoring
Two-Point Conversion Passes	Count	No	359	Scoring
Fantasy Points (Standard)	Continuous	No	11	Fantasy
Fantasy Points (PPR)	Continuous	No	10	Fantasy
DraftKings Fantasy Points	Continuous	No	10	Fantasy
FanDuel Fantasy Points	Continuous	No	10	Fantasy
Value-Based Drafting Score	Continuous	No	299	Fantasy
Fantasy Position Rank	Count	No	0	Fantasy
Fantasy Overall Rank	Count	No	294	Fantasy

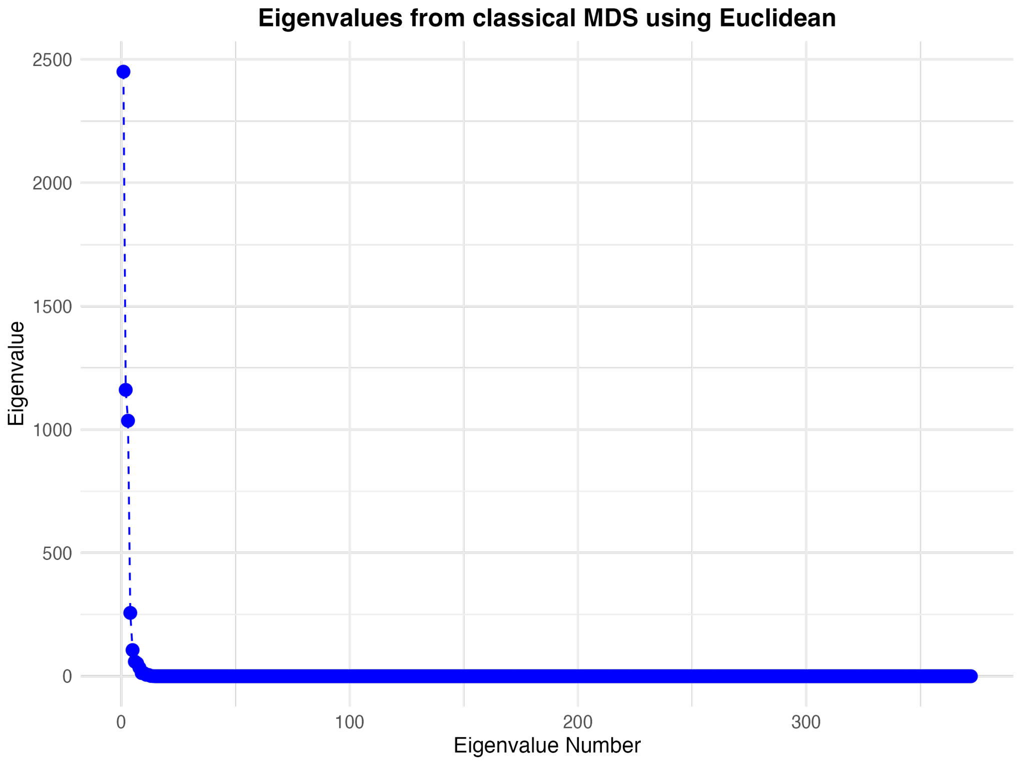

Figure 1 Eigenvalues of B matrix obtained through classical MDS using Euclidean distance. We observe 3 dominating eigenvalues suggesting 3 is a good target dimension for MDS. Additionally, as most eigenvalues are greater than 0, it suggests that the data is close to Euclidean, but not completely.

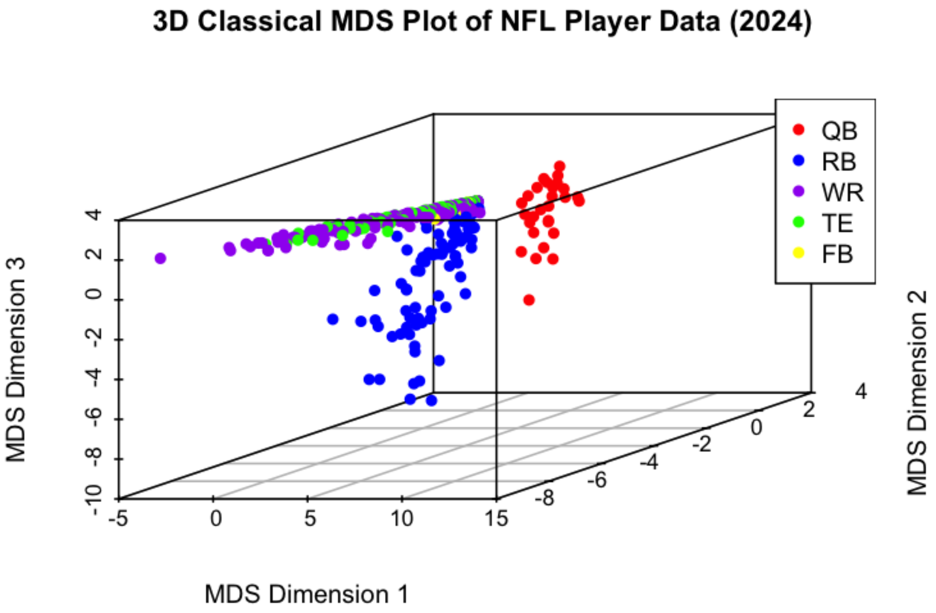

Figure 2 We observe 3 separate clusters for classical MDS done using Euclidean distance, correctly identifying the similarity among players of similar positions. We also observe a convergence of players with different roles which is discussed later.

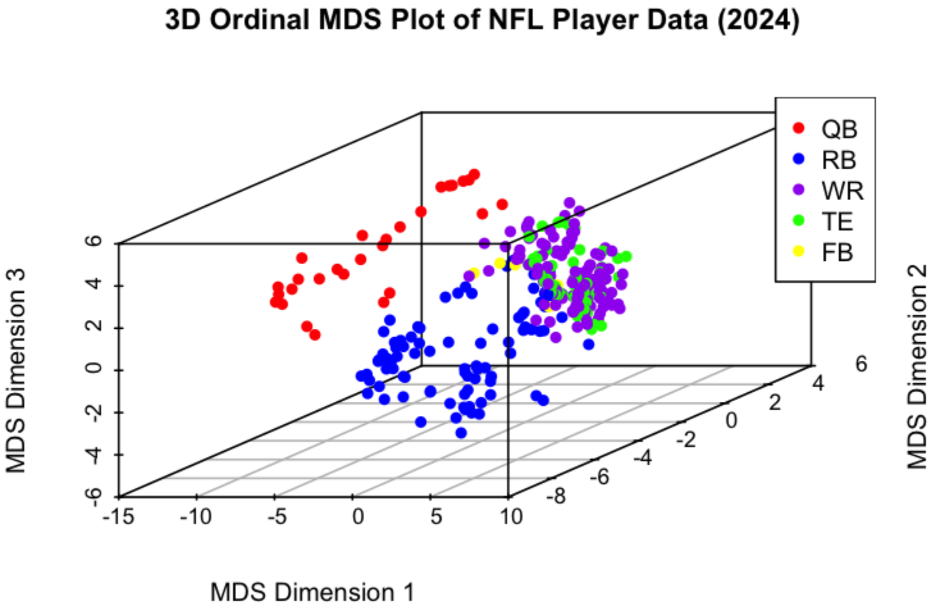

Figure 3 Three-dimensional ordinal MDS using Canberra distance. The structure is similar to classical MDS but better accounts for non-Euclidean relationships.

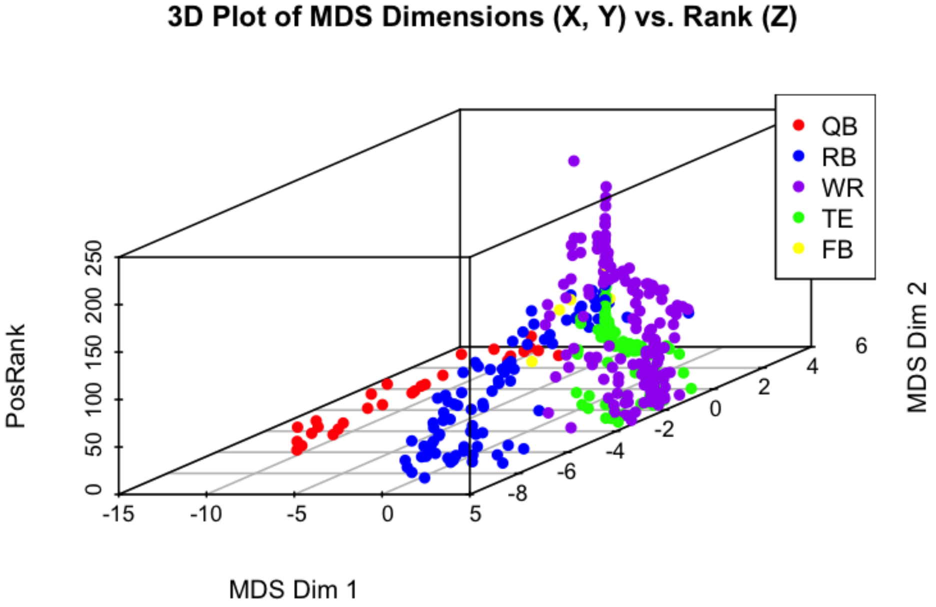

Figure 4 Two-dimensional ordinal MDS with fantasy position rank as an additional variable. Higher-ranked players cluster together, suggesting MDS dimension 2 correlates with player ranking.