Project 1

„Analysis of Linear ordering, Cluster analysis and Multidimensional scaling on NFL players example”

Author: Bartłomiej Jamiołkowski

Cracow 18.10.2021

Table of Contents

1. Introduction

Sport has always been very popular in society as it can be surprising and evoke a lot of extreme emotions. An example is American football, "*a contact team sport in which two teams of eleven each aim to score more points.*" One of the most popular sports in the United States is the subject of numerous statistical analyzes, which makes it a good candidate to be selected as the main topic in the project.

1.1 Project goal

The project is carried out with a view to presenting selected methods of data analysis in the areas of: linear ordering, cluster analysis and multidimensional scaling. For this purpose, a set of 30 observations of NFL (National Football League) players collected up to October 2021 is used. The project is implemented using the RStudio environment.

1.2 Data sources

- *NFL.com | Official Site of the National Football League* – official website of the NFL for each player's current soccer stats. Information on players' predispositions and detailed measurements of their movements during the game are published.
- *The Football Database: Football Statistics and History* – NFL player statistics database providing additional information, for example on the number of matches won or lost by a player.

2. Data analysis and description

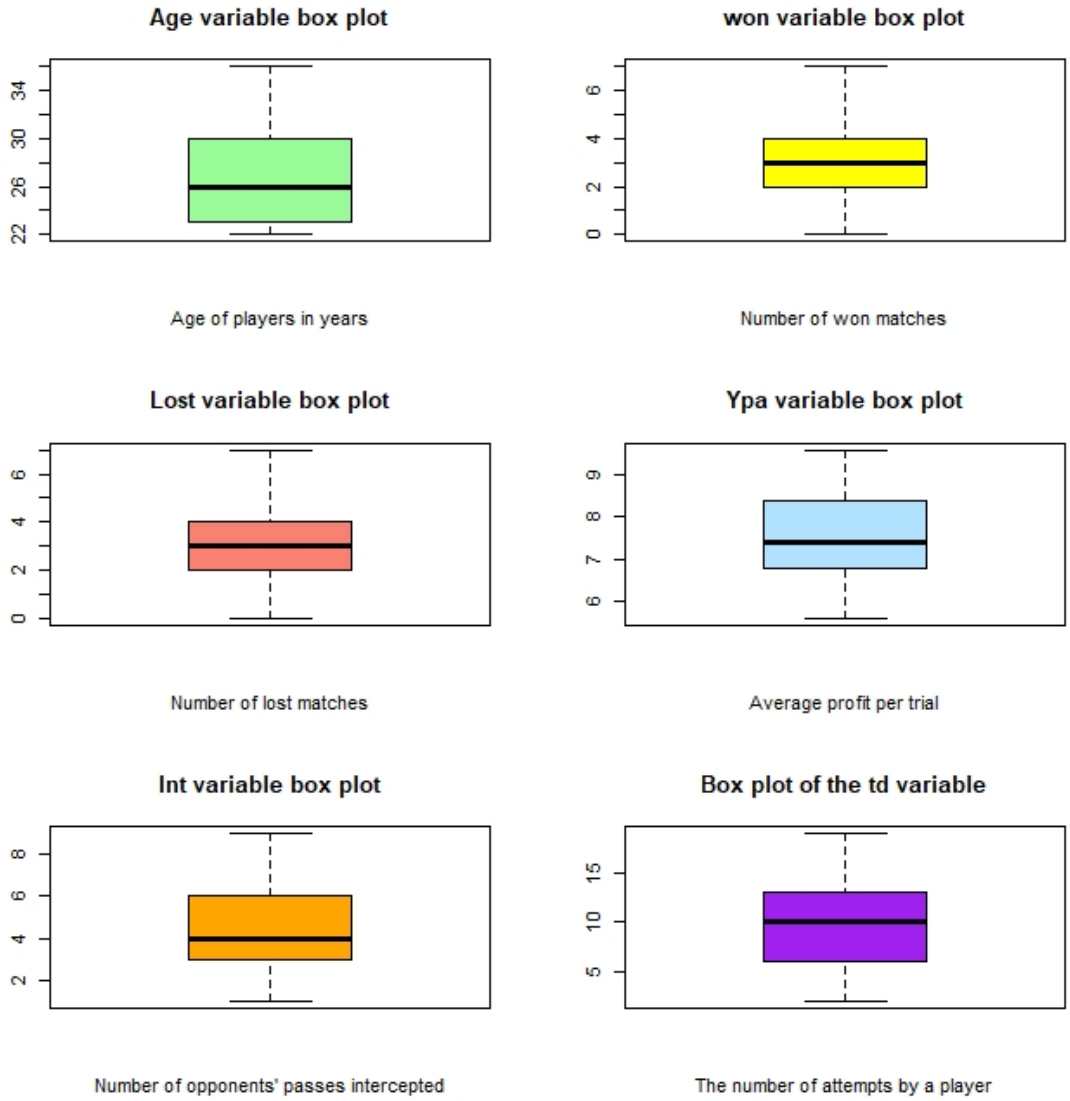
The following variables apply to the project:

- *player* – variable containing the name and surname of a given player;
- *age* – information about the age of the player;
- *won* – number of matches won with the featured person;
- *lost* – number of matches lost by the player;
- *ypa* – average profit on trial of 8.2 yards [1];
- *int* – number of opponents' passes intercepted;
- *td* – number of touchdowns by the player pushing the ball into the opponent's end zone.

[1] https://www.dummies.com/sports/football/offense/how-to-read-a-quarterbacks-statistics/ [29.10.2021]

2.1 Investigating the distribution of variables and looking for outliers

Box plots can be used to detect outliers. The graphic below shows that there are no outliers. Moreover, one can notice the lack of normal distributions of the analyzed variables.



2.2 Basic Statistics

- summary():

```
   player              age             won             lost            ypa             int             td        
 Length:30          Min.   :22.00   Min.   :0.0   Min.   :0.000   Min.   :5.600   Min.   :1.000   Min.   : 2.000  
 Class :character   1st Qu.:23.25   1st Qu.:2.0   1st Qu.:2.000   1st Qu.:6.875   1st Qu.:3.250   1st Qu.: 6.250  
 Mode  :character   Median :26.00   Median :3.0   Median :3.000   Median :7.400   Median :4.000   Median :10.000  
                    Mean   :26.73   Mean   :2.9   Mean   :3.133   Mean   :7.507   Mean   :4.733   Mean   : 9.967  
                    3rd Qu.:29.50   3rd Qu.:4.0   3rd Qu.:4.000   3rd Qu.:8.225   3rd Qu.:6.000   3rd Qu.:13.000  
                    Max.   :36.00   Max.   :7.0   Max.   :7.000   Max.   :9.600   Max.   :9.000   Max.   :19.000
```

Among the results obtained, some values are worth paying attention to. The age of the players ranges from 22 to 36 years, with the average being around 27 years old. The number of matches

won and lost ranges from 0 to 7, with the average player's team succeeding 0.233 less often than losing. YPA is the number of yards gained in the trial. At the time of measurements in 2021, it is in the range of 5.6 - 9.6 yards. The number of intercepted opponents' passes is in the range of 1-9, with the information that the average player intercepts about 5 times during the measurements. One of the most interesting statistics is the one concerning the number of touchdowns on the opponent's field (TD). The record holder does it 19 times, while the weakest player in this category does it only twice.

- Skewness:

```
> skewness(data_nfl$age)
[1] 0.6146645
> skewness(data_nfl$won)
[1] 0.1403262
> skewness(data_nfl$lost)
[1] 0.3119484
> skewness(data_nfl$ypa)
[1] 0.3390957
> skewness(data_nfl$int)
[1] 0.200911
> skewness(data_nfl$td)
[1] 0.2138462
>
```

The presented data show that practically all variables have a right-hand asymmetry, which means that most of the observations are to the right of the mean.

- Kurtosis:

```
> kurtosis(data_nfl$age)
[1] -0.8274696
> kurtosis(data_nfl$won)
[1] -0.783287
> kurtosis(data_nfl$lost)
[1] -0.1345954
> kurtosis(data_nfl$ypa)
[1] -0.5972258
> kurtosis(data_nfl$int)
[1] -0.8819289
> kurtosis(data_nfl$td)
[1] -1.119612
```

All presented kurtoses are negative, which proves the platokurtic distributions of the considered variables.

2.3 Coefficient of Variation

The coefficient of variation is a classic measure of the differentiation of a feature, calculated as the quotient of the standard deviation and the mean. The desired values in the design are

numbers greater than 0.1 (10%). The graphic below shows that the analyzed variables meet this criterion required in the later methods.

```
> cv_age = sd(data_nfl$age)/mean(data_nfl$age)
> print(cv_age)
[1] 0.1483088
> cv_won = sd(data_nfl$won)/mean(data_nfl$won)
> print(cv_won)
[1] 0.642566
> cv_lost = sd(data_nfl$lost)/mean(data_nfl$lost)
> print(cv_lost)
[1] 0.4795176
> cv_ypa = sd(data_nfl$ypa)/mean(data_nfl$ypa)
> print(cv_ypa)
[1] 0.1294847
> cv_int = sd(data_nfl$int)/mean(data_nfl$int)
> print(cv_int)
[1] 0.4866382
> cv_td = sd(data_nfl$td)/mean(data_nfl$td)
> print(cv_td)
[1] 0.4782806
```

2.4 Corelations

	age	won	lost	ypa	int	td
age	1.00	0.17	-0.31	0.26	-0.53	0.21
won	0.17	1.00	-0.73	0.72	0.07	0.74
lost	-0.31	-0.73	1.00	-0.59	0.27	-0.49
ypa	0.26	0.72	-0.59	1.00	-0.14	0.62
int	-0.53	0.07	0.27	-0.14	1.00	0.06
td	0.21	0.74	-0.49	0.62	0.06	1.00

The presented correlation matrix shows which variables can be more closely related to each other. Noteworthy is the td-won pair with a fairly strong dependence of 0.74, indicating that the win depends on the number of touchdowns. Another interesting moderate correlation is that of 0.62 td-ypa, suggesting that the more yards a player moves, the greater the number of touchdowns. The consequence of this is a strong correlation of 0.72 ypa-won, indicating a direct relationship between a player's movement and winning.

3. Linear ordering

The purpose of linear ordering is to order objects, in this case players, from best to worst according to a synthetic variable. Its value is determined by selected features represented by variables divided into:

- stimulants: won, td, int, ypa;
- destimulants: lost;
- nominants: age – the desired value is the mean age of 27.

It is advisable to replace the variables with stimulants before using them in different methods.

3.1 Template methods

3.1.1 Hellwig method

Through numerous transformations and activities, the player ranking is obtained. Matthew Stafford is in first place, while negative players are in the last positions.

	Player	Hellwig
1	Matthew Stafford	1.00000000
2	Patrick Mahomes	0.99476384
3	Kyler Murray	0.97905535
4	Joe Burrow	0.97905535
5	Dak Prescott	0.95287455
6	Josh Allen	0.91622141
7	Justin Herbert	0.86909596
8	Kirk Cousins	0.81149818
9	Jameis Winston	0.81149818
10	Derek Carr	0.74342808
11	Matt Ryan	0.74342808
12	Teddy Bridgewater	0.74342808
13	Carson Wentz	0.66488565
14	Jalen Hurts	0.57587091
15	Lamar Jackson	0.57587091
16	Russell Wilson	0.57587091
17	Mac Jones	0.47638383
18	Jared Goff	0.36642444
19	Sam Darnold	0.24599272
20	Ryan Tannehill	0.24599272
21	Trevor Lawrence	0.24599272
22	Tua Tagovailoa	0.24599272
23	Baker Mayfield	0.11508868
24	Jimmy Garoppolo	0.11508868
25	Daniel Jones	-0.02628769
26	Davis Mills	-0.02628769
27	Zach Wilson	-0.17813637
28	Jacoby Brissett	-0.17813637
29	Geno Smith	-0.34045739
30	Justin Fields	-0.51325072

>

3.2 Templateless methods

3.2.1 Standardized sum method

One of the patternless methods that creates a ranking based on an indicator in the interval [0-1]. Having the printouts of this method and Hellwig's method, the results are compared. When analyzing both graphics, it can be noticed that although the order in the rankings is a bit different, the leading players, such as Kyler Murray or Matthew Stafford, are still in the top 4.

The same is with the worse players at the bottom of the list, such as Jacoby Brissett in the last 3.

	Player	indicator
1	Kyler Murray	1.00000000
2	Joe Burrow	0.86469408
3	Matthew Stafford	0.80369876
4	Dak Prescott	0.78833936
5	Jameis Winston	0.70565063
6	Patrick Mahomes	0.68844056
7	Derek Carr	0.62916691
8	Lamar Jackson	0.60488312
9	Josh Allen	0.51912697
10	Justin Herbert	0.46734656
11	Teddy Bridgewater	0.44298197
12	Ryan Tannehill	0.44131731
13	Baker Mayfield	0.41769904
14	Russell Wilson	0.34780046
15	Sam Darnold	0.33011689
16	Carson Wentz	0.31740106
17	Mac Jones	0.31027135
18	Kirk Cousins	0.28884853
19	Matt Ryan	0.27744218
20	Jimmy Garoppolo	0.25386735
21	Jared Goff	0.23430222
22	Tua Tagovailoa	0.17236373
23	Trevor Lawrence	0.16845574
24	Jalen Hurts	0.15970056
25	Zach Wilson	0.12981957
26	Justin Fields	0.11733195
27	Daniel Jones	0.10415136
28	Geno Smith	0.01803256
29	Davis Mills	0.01794737
30	Jacoby Brissett	0.00000000

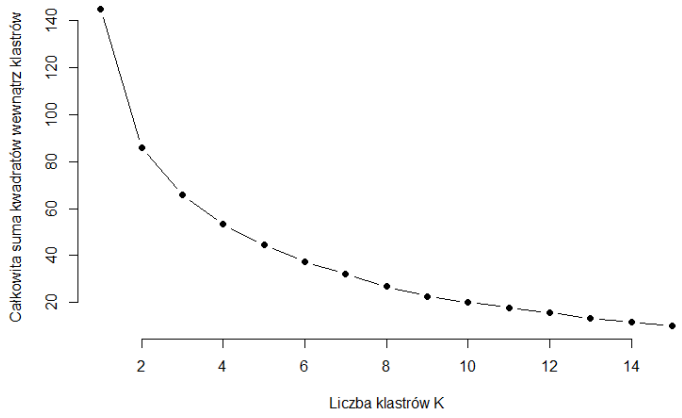
4. Cluster analysis

4.1 Division methods

4.1.1 Choosing the optimal number of clusters

Choosing the number of clusters in a cluster analysis is an important step due to how many clusters will be visualized in each method. You can define k clusters yourself or with dedicated methods. One of them is the Elbow method, which bases its operation on "plotting the explained variability as a function of the number of clusters and selecting the curve elbow as the number of clusters to use.*"*[2]

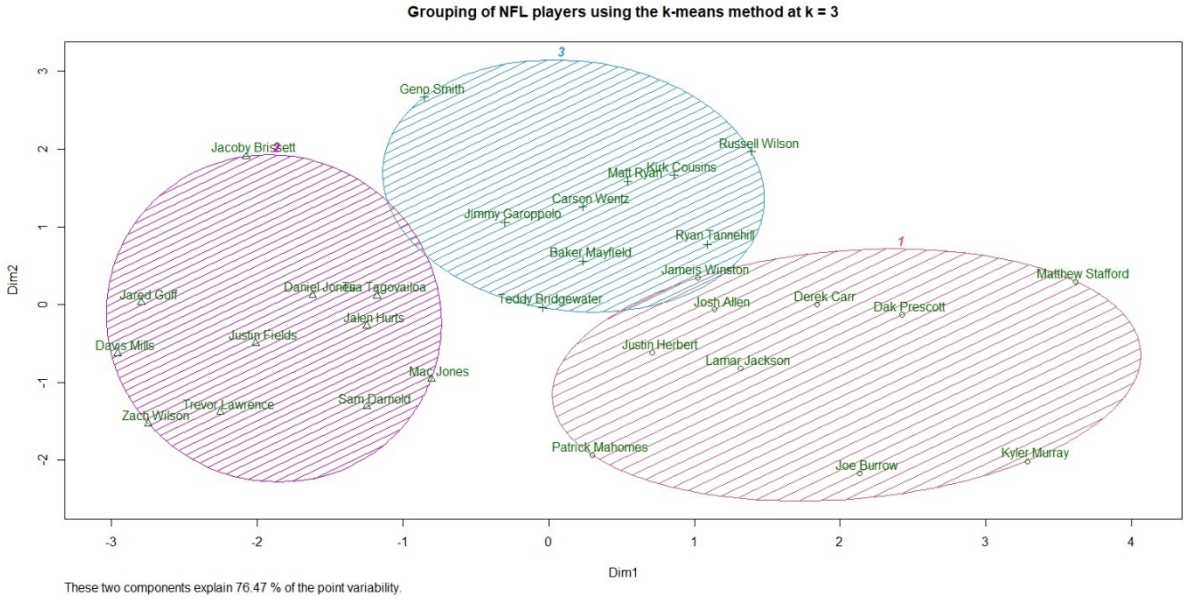
[2] https://en.wikipedia.org/wiki/Elbow_method_(clustering) [02.11.2021]



When analyzing the graph, it can be seen that the considered k may have a value of 2 or 3. Later analyzes show why 3 is chosen as the number of clusters k = 3.

4.1.2 K-means method

One of the most frequently used partitioning methods is the k-means method. After standardizing the scale () and checking that the values of all variables are less than 3 (the rule of three sigmas), a visualization of the clusters of NFL players is created using the kmeans () function.



On the basis of the obtained: graph, data and previous rankings, it is possible to distinguish the division into 3 main clusters. The purple cluster represents the weak players at the time of trading, the blue cluster contains the average players dropping out, and the red cluster contains the best players. This is evidenced by the statistics of individual clusters.

```
> summary(cluster1)
   players1             age1            won1           lost1            ypa1            int1            td1       
 Length:10          Min.   :22.00   Min.   :0.0   Min.   :3.0   Min.   :6.10   Min.   :4.00   Min.   : 2.00  
 Class :character   1st Qu.:22.25   1st Qu.:1.0   1st Qu.:4.0   1st Qu.:6.50   1st Qu.:4.50   1st Qu.: 5.00  
 Mode  :character   Median :23.00   Median :1.5   Median :5.0   Median :6.80   Median :6.00   Median : 7.00  
                    Mean   :23.30   Mean   :1.5   Mean   :4.6   Mean   :6.75   Mean   :6.20   Mean   : 6.40  
                    3rd Qu.:23.75   3rd Qu.:2.0   3rd Qu.:5.0   3rd Qu.:7.10   3rd Qu.:7.75   3rd Qu.: 7.75  
                    Max.   :27.00   Max.   :3.0   Max.   :7.0   Max.   :7.20   Max.   :9.00   Max.   :10.00  
> summary(cluster2)
   players2             age2            won2           lost2            ypa2            int2            td2       
 Length:9           Min.   :26.00   Min.   :0.000   Min.   :2   Min.   :6.800   Min.   :1.000   Min.   : 3.000  
 Class :character   1st Qu.:28.00   1st Qu.:2.000   1st Qu.:3   1st Qu.:7.400   1st Qu.:1.000   1st Qu.: 6.000  
 Mode  :character   Median :31.00   Median :3.000   Median :3   Median :7.600   Median :3.000   Median :10.000  
                    Mean   :30.78   Mean   :2.667   Mean   :3   Mean   :7.744   Mean   :2.889   Mean   : 8.556  
                    3rd Qu.:33.00   3rd Qu.:3.000   3rd Qu.:3   3rd Qu.:7.700   3rd Qu.:4.000   3rd Qu.:12.000  
                    Max.   :36.00   Max.   :5.000   Max.   :4   Max.   :9.600   Max.   :5.000   Max.   :13.000  
> summary(cluster3)
   players3             age3            won3           lost3            ypa3            int3            td3       
 Length:10          Min.   :23.00   Min.   :3.0   Min.   :0.00   Min.   :7.200   Min.   :3.0   Min.   :10.00  
 Class :character   1st Qu.:24.00   1st Qu.:4.0   1st Qu.:1.25   1st Qu.:7.525   1st Qu.:4.0   1st Qu.:13.25  
 Mode  :character   Median :25.50   Median :5.0   Median :2.00   Median :8.450   Median :4.5   Median :15.50  
                    Mean   :26.40   Mean   :4.8   Mean   :1.80   Mean   :8.240   Mean   :5.2   Mean   :15.10  
                    3rd Qu.:27.75   3rd Qu.:5.0   3rd Qu.:2.00   3rd Qu.:8.900   3rd Qu.:6.5   3rd Qu.:17.00  
                    Max.   :33.00   Max.   :7.0   Max.   :4.00   Max.   :9.200   Max.   :9.0   Max.   :19.00  
> data2 = data.pfl
```

Cluster 1, the so-called purple is considered a cluster of the weakest players due to several aspects. The first is the lowest number of wins. The average win is 1.5 per player. At the same time, the average of lost matches is exceptionally high, amounting to 4.6 losses per player. What characterizes these players is that the average age is 23.30 years, which is little more than the minimum for the entire data set. This means that these people still have little experience at the time of measurement. Other interesting values are the number of yards covered and touchdowns. Both averages of these variables are the smallest among the 3 clusters.

Cluster 2, the so-called blue is the center of the oldest players. Of course, they are experienced, but they do not have this ability anymore, so they do not perform as sensational as the red group. This can be seen in the majority of averages whose values are between the values of the average of the weakest and the best group. In addition, they have the lowest number of intercepted opponent's passes.

Cluster 3, the so-called red represents the cluster of players with the best scores. These age players are at the optimal point in their career - around 26.40 years. Their results are very good. The average number of wins is 4.8 compared to 1.8 losses. The average yardage and touchdown values are the highest of all the clusters.

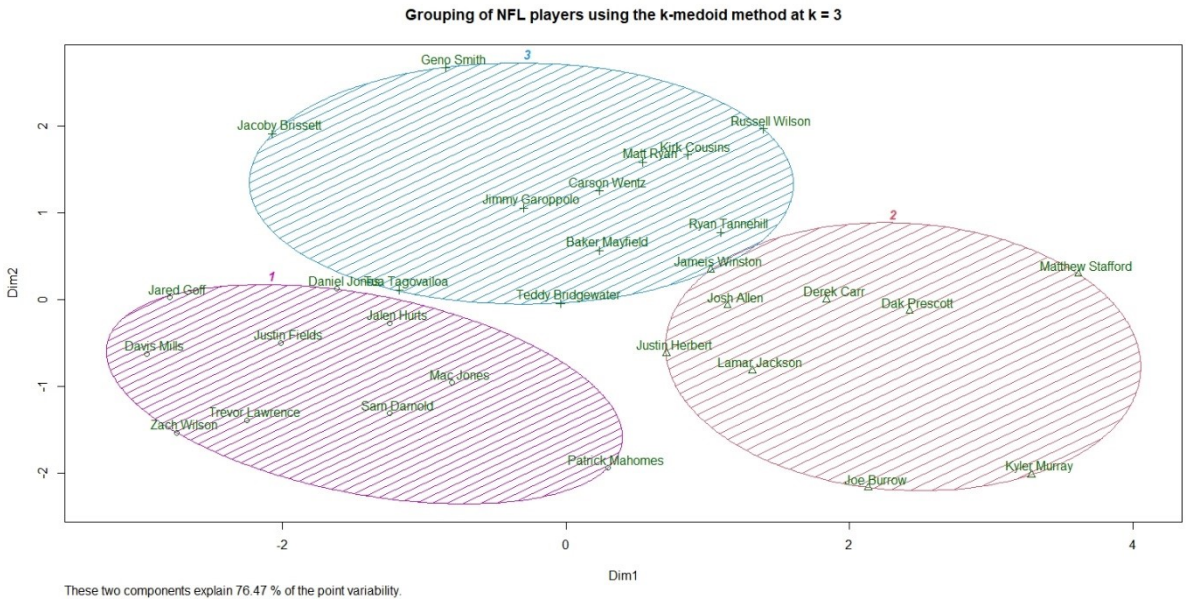
The leaders of the red group are Matt Stafford and Kyler Murray. Their extreme opposite is the anti-violet group leaders David Mills and Jared Goff. What makes these pairs drastically different is the win-lose ratio. For the first person it is 7: 0, while for the third named character it is 0: 7. This means that the further to the right a player is, the greater his effectiveness (wins

or touchdowns). Another interesting observation is the alignment of players with respect to the OY axis. The higher the player is, the smaller the number of captured throws by the opponent.

The graphic shows the generated information about 76.47% of the quality of the division of players.

4.1.3 K-medoid method

A method based on choosing the median over the coordinates as a focal point.

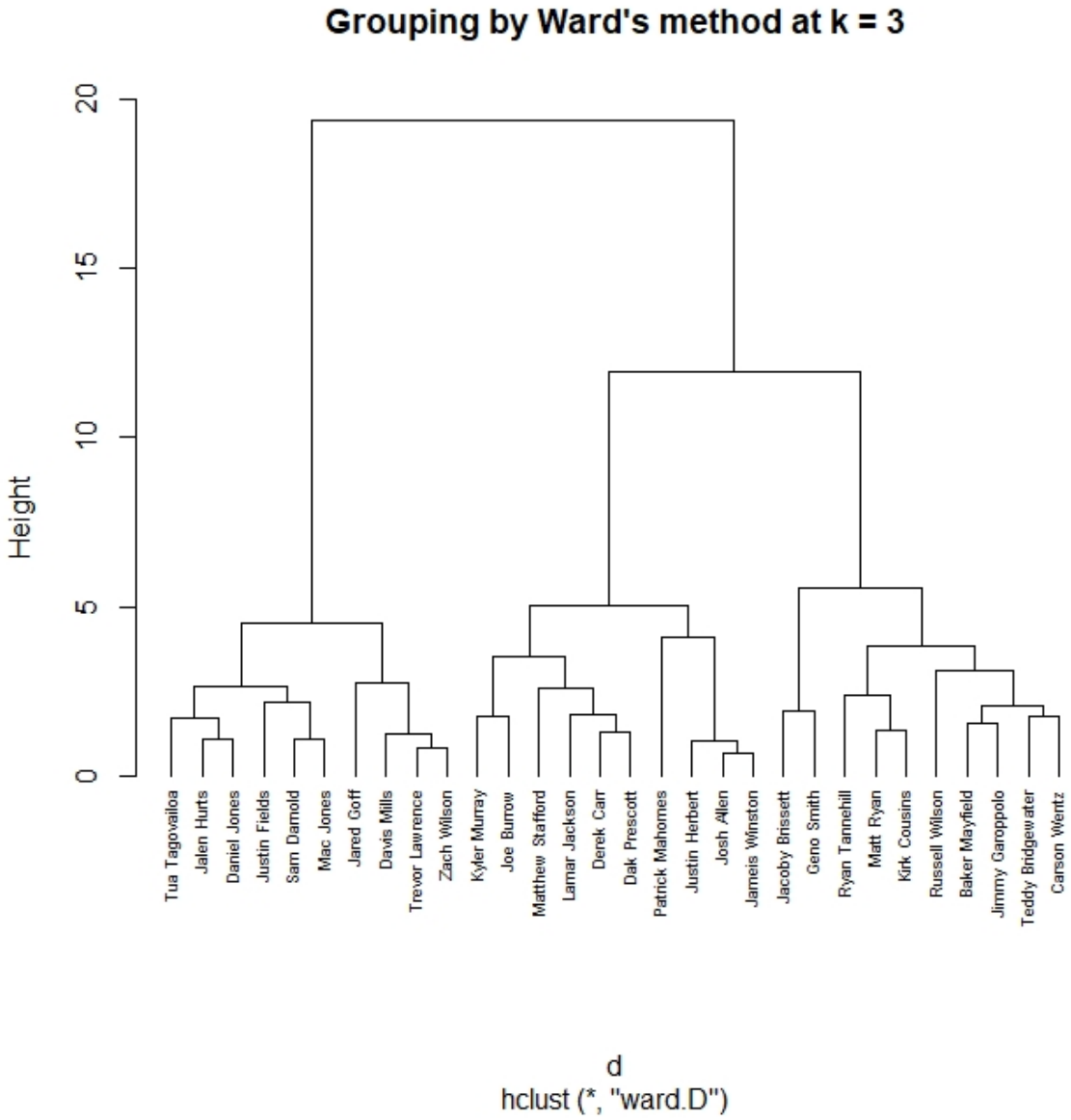


The generated k-medoid plot differs from the k-means plot in the size of the clusters, although there is still a division into the discussed clusters. It is unreasonable to re-analyze the position of players. The curiosity is certainly aroused by a few players who changed their group affiliation compared to the previous chart. These include Jacoby Brissett or Patrick Mahomes. By analyzing their statistics, one can formulate a conclusion that they do not fit into new groups, e.g. the number of interceptions or tryouts, and thus this method is worse than k-means in the case of these data.

4.2 Hierarchical grouping

4.2.1 Ward method

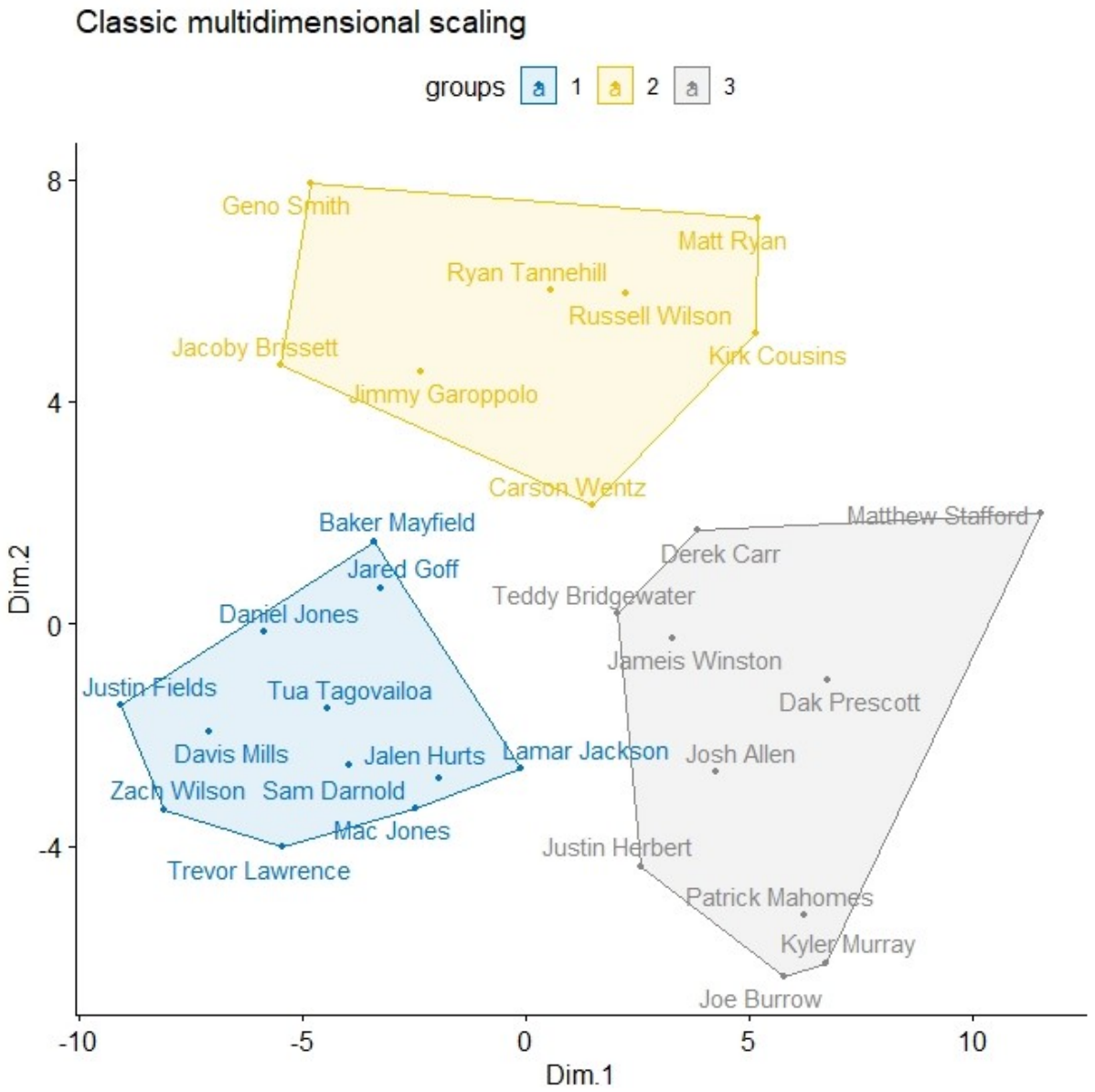
The Ward method is considered to be an effective grouping method as it shows the order of the connecting objects as well as the distances between them. The following dendrogram generated using the Euclid's method shows the described division into 3 clusters: worse players on the left, best in the middle, and average players on the right. K = 3 is well chosen because there is a big difference between the level k and k + 1.



5. Multidimensional scaling

5.1 The method of classical multidimensional scaling

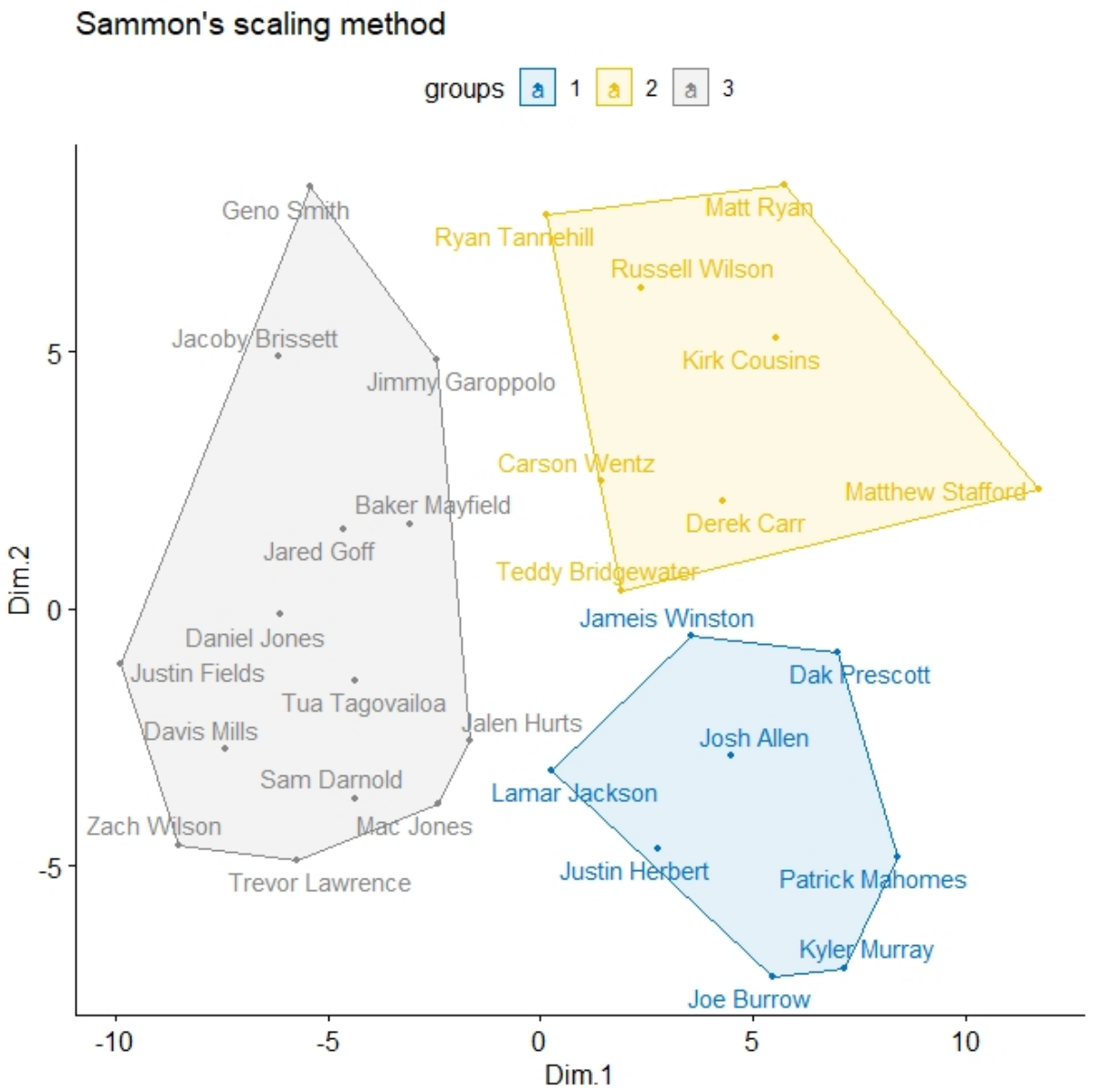
Multidimensional scaling is the method used to visualize n-dimensional objects in m-dimensional space. The k-means method is used to generate a 2-dimensional plot.



The presented chart shows three clusters of players. In the visualization, it is interesting to change the position of some players in relation to the previous graphics. When analyzing the arrangement, it can be concluded that the players were grouped less well, eg Derek Carr, whose key statistics differ significantly from the average group of average players.

5.2 Summon scaling method

An alternative method to classical multidimensional scaling is Summon's scaling method based on mapping small distances. Here, too, there are some minor changes when generating the graph using the k-means method.



6. Summary

By analyzing the obtained: visualizations, data and statistics, it is possible to formulate the view that the best method in this case turns out to be the k-means method. Which method is selected has a significant impact on the subsequent presentation of the results. The subject matter of the project itself encourages the exploration of these issues, and the presented aspects of American football can broaden the reader's knowledge about this quite exotic field of sport.