MS-E2112 Multivariate Statistical Analysis

Final Project Report

Student: Khoa Lai (732255)

Motivation

As a student major in Statistics, a data analyzing enthusiast and a football lover, I wonder what are the underlying characteristics attributing to the entertainment of a football match, and how it could be found and explained statistically. In addition, the movement of the ball and the skill of the player has always intrigued my curiosity since I was a kid, and I would like to understand other attributes of the match given the past league dataset. To utilize the content learned from this course and to satisfy my curiosity, I have conducted a multivariate statistical analysis project using the most recent Premier League dataset 2018/2019.

In this project, I use the dataset collected and distributed by http://football-data.co.uk, a site providing historical results and odds to help football betting enthusiasts analyze many years of data quickly and efficiently to gain an edge over the bookmaker. The project aims at *finding latent relationships among the attributes of a football match, predicting to some acceptable extent the total number of goals in a match given available measurements.* The author conducts his analysis independently, and therefore the result might be different from others' production.

Research Questions

Concretely, this project aims at providing the fully explained, detailed answers for the following questions:

(1) To which extent are the different attributes correlated to each other?
(2) Is it possible to explain the variances in the metrics with fewer components? If yes, what could those components be and how much could they explain the variances?
(3) Is there a similarity between the group of matches? If yes, what are these similarities?
(4) What are the necessary models or algorithms that help classify the matches into appropriate groups of similarity?

Data Description

In this project, the author uses the results of the most recent Premier League season 2018/2019. The data has 380 rows and 13 columns. Each row corresponds to a match of two teams (home team and away team). Each column corresponds to a specific attribute of the match.

The dataset contains the following attributes:

- **“Matches”**: names of the home and away teams, in the format “HomeTeam v AwayTeam”.
- **“HTHG", “HTAG"**: half time goals of the home and away team.
- **“HTR"**: half time results, 1, if the home team won, -1 if it lost, 0 for a draw.
- **“HS", “AS":** number of shots by the home and the away team.
- **“HST", “AST"**: number of shots on target by the home and the away team.
- **“HF", “AF"**: number of fouls by the home and away team.
- **“HC", “AC"**: number of corners of the home and away team.
- **“FTG”:** total number of goals in the match.

Data Analysis

Univariate Data Analysis

To have an understanding of the distribution of the data, the author plots the histogram of all attributes:

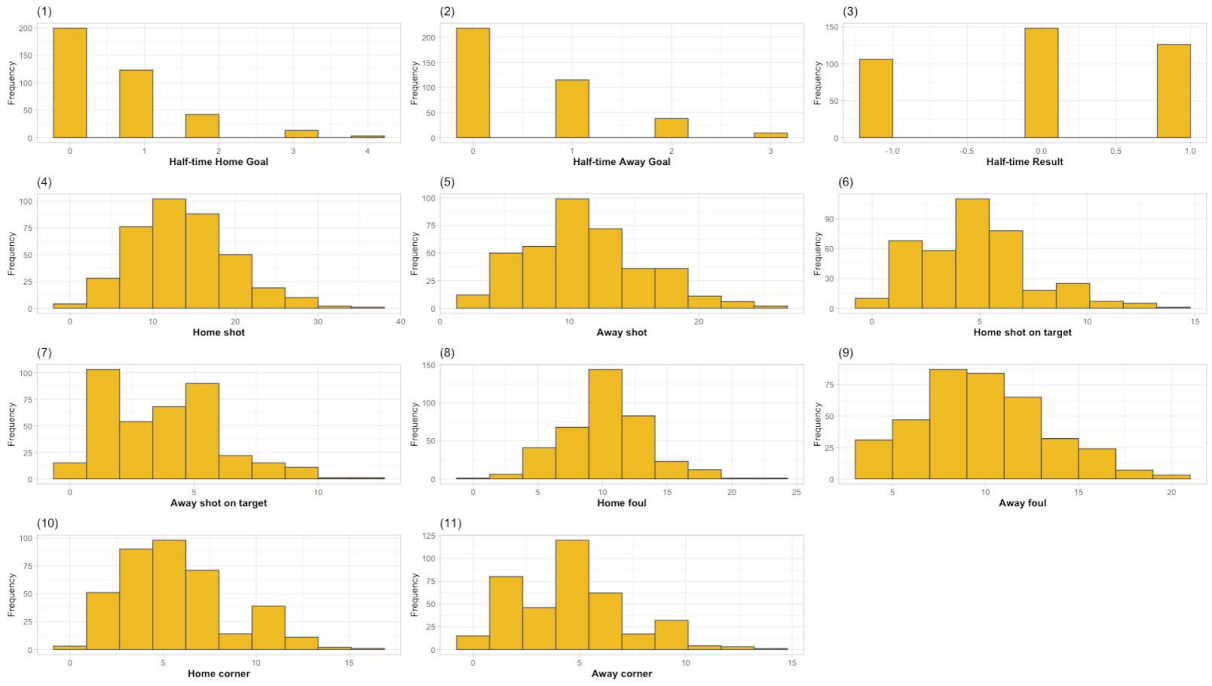


Figure 1: The distribution of attributes

Observations:

- In general, the home team plays better in a match against the away team.
- Most of the matches have no goal at half-time.
- There are no matches in which the Away Team leads by more than 3 goals at half-time, while there are 3 matches the Home Team shows the domination by leading 4 goals at half-time.

- The Home Foul seems to follow the normal distribution with mean 10.

Bivariate Data Analysis

Theoretical Background

	Skewness	Kurtosis	Jarque-Bera Test
Description	***The skewness*** describes the deviation of the data from symmetry. In other words, the skewness tests by how much the **overall shape** of a distribution deviates from the shape of the normal distribution. The skewness of the normal distribution is 0.	**Kurtosis** is a measure of **how differently shaped are the tails of a distribution** as compared to the tails of the normal distribution. While skewness focuses on the overall shape, Kurtosis focuses on the tail shape.	**The Jarque-Bera test** is a goodness-of-fit test of **whether sample data have the skewness and kurtosis matching a normal distribution.**
Test statistic	$\hat{\gamma} = \frac{m_3}{s^3}$ where $m_3 = \frac{1}{n}\sum_1^n (x_i - \bar{x})^3$	$\hat{k} = \frac{m_4}{s^4} - 3$ where $m_4 = \frac{1}{n}\sum_1^n (x_i - \bar{x})^4$	BS = $n(\frac{\hat{\gamma}^2}{6} + \frac{\hat{k}^2}{24})$, where $\hat{\gamma}$ is the sample skewness coefficient and $\hat{k}$ is the sample kurtosis coefficient.
Interpretation	If the skewness coefficient $\hat{\gamma} > 0$, then the distribution is *skewed to the right (positively skewed)*. The distribution has a long-right tail and the mass of the distribution is concentrated on the left. If $\hat{\gamma} < 0$, then the distribution is *skewed to the left (negatively skewed)*. The distribution has a long-left tail and the mass of the distribution is concentrated on the right.	A random variable *with normal distribution has kurtosis value 0.* If the kurtosis value is k > 0, then the distribution has *heavier tails than the normal distribution.* If k < 0, then the distribution has *lighter tails than the normal distribution.*	If the observed skewness or kurtosis values differ significantly from the skewness and/or kurtosis values of the normal distribution (0 and 0), *the test statistic gets large values.* If *n* is large, then under H_0 the test statistic BS *follows approximately* X_2^2 *distribution.* The expected value of the test statistic under H_0 is approximately 2 and **large values** of the test statistic suggests *that the null hypothesis* H_0 *is false.*

Figure 2: The description, test statistic and interpretation of descriptive statistics

The R's fitdistrplus has a built-in function namely **descdis**t which computes descriptive parameters of the distribution of our dataset:

	HTHG	HTAG	HTR	HS	AS	HST	AST	HF	AF	HC	AC
Min	0	0	-1	0	2	0	0	0	3	0	0
Max	4	3	1	36	25	14	12	23	21	16	14
Median	0	0	0	14	11	5	4	10	10	5	4
Mean	0.678	0.573	0.052	14.13	11.14	4.778	3.92	10.15	10.3	5.7	4.55
Estimated SD	0.86	0.767	0.78	5.855	4.654	2.677	2.283	3.29	3.5	2.97	2.73
Estimated skewness	1.3	1.21	-0.092	0.542	0.463	0.582	0.488	0.232	0.295	0.47	0.58
Estimated kurtosis	4.43	3.82	1.648	3.36	2.856	3.1	2.912	3.484	2.922	2.84	3.02
Jarque-Bera Test	< 2.2e-16	< 2.2e-16	4.178e-07	3.844e-05	0.0009502	2.256e-05	0.0005095	0.0334	0.05981	0.0006608	2.538e-05

Figure 3: Summary Statistics and Jarque-Bera Test

According to the table, some observations could be made:

- The HS min is 0 => there exists a match in which the Home Team *could not shot at least once.* The match is BOU v CHE. In contrast, the AS min is 2 indicates that all of the matches witness the Away Team be able to shot at least 2 times.
- The HF min is 0 => there exists a match in which the Home Team *did not commit any foul against the Away Team.* The match is WOL v BRI. In contrast, the AF min is 3 indicates that all of the matches witness the Away Team committed at least 3 fouls.
- The sample skewness coefficient estimates the population skewness. All variables except the HTR has positive coefficients, indicating the distribution of those variables are *skewed to the right (positively skewed).*
- All variables except the HTR have skewness value slightly > 0, indicating that *their distributions are slightly skewed to the right*. The HTR skewness value is -0.092, indicating that *its distribution is slightly skewed to the left (almost matches the shape of the normal distribution).*
- All variables have kurtosis value > 0, indicating that their distributions has *heavier tails than the normal distribution.*
- All variables except the AF have p-value for the Jarque-Bera Test significantly < 0.05, meaning that the null hypothesis that the data is normally distributed is rejected. For the AF (Away Foul), given the graph, the AF seems to follow the normal distribution, although it is not totally clear cut.

Multivariate Data Analysis

Multivariate Correlation

Theoretical Background

Correlation is defined as a measure of the *linear* relationship between two quantitative variables. When the values of one variable increase as the values of the other increase, this is known as *positive correlation* and vice versa.

Correlation Analysis

Understanding the correlation between the features is a common practice that helps to facilitate and validate the further multivariate analysis.

The R's **corrplot** package allows building and customizing the correlation matrix directly from the dataset:

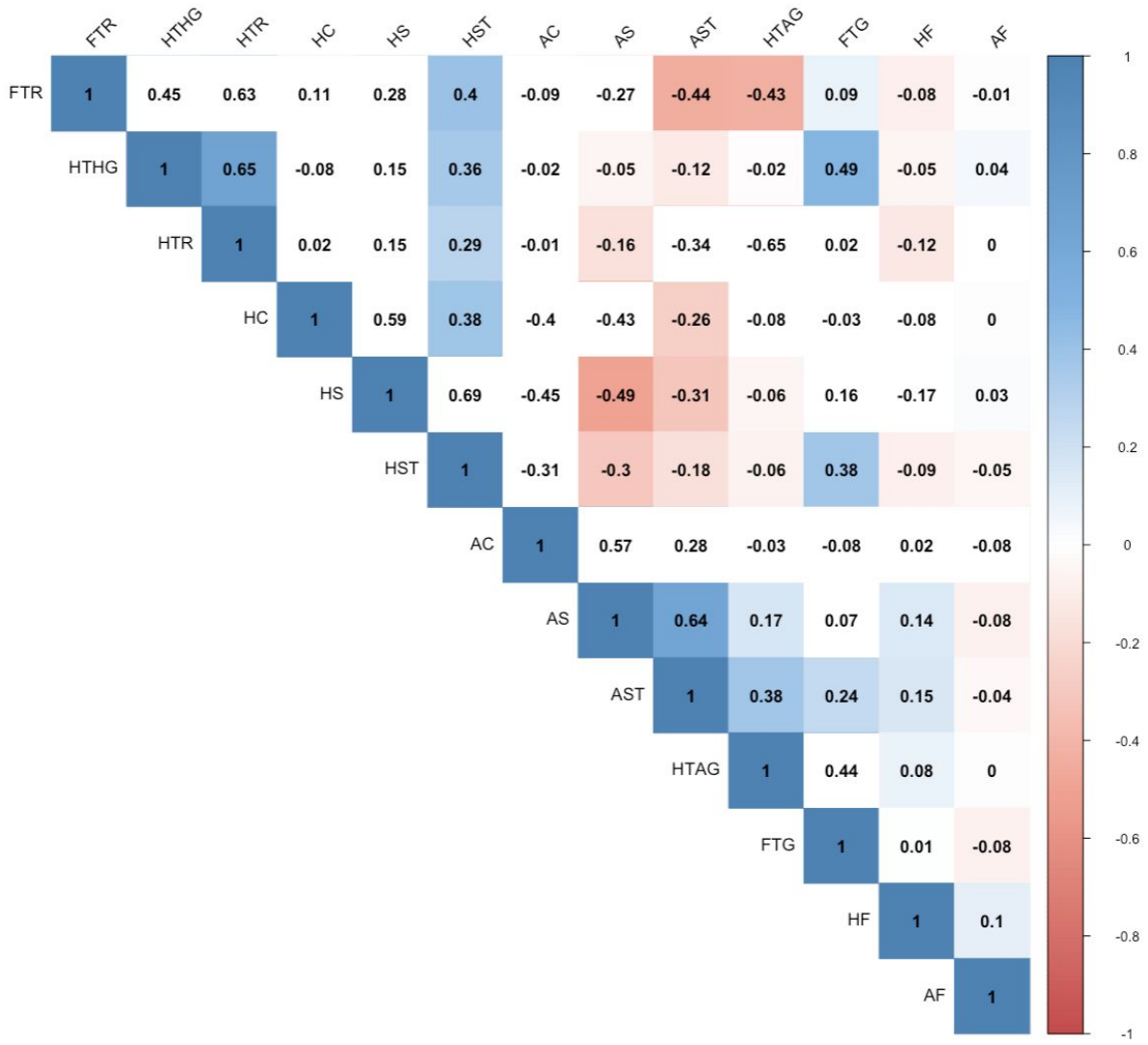


Figure 4: Correlation Matrix

Here, the author would use the cut-off value of 0.4 to indicate significant dependencies between the variables. The choice of the cut-off value depends on the task and the interpretation drawn from it.

Observations:

- The HST has a positive correlation with the FTR (0.4). Interestingly, the correlation coefficient between the AST and the FTR is negative, ***indicating that in general, the Away Team has a low chance-goal conversion rate.***
- Both the coefficient between HTHG and FTG and between HTAG and FTG are significantly positive (0.49 and 0.44) that ***for every 1 unit increase in HTHG, there is a 0.49 increase in the HTHG, 0.44 increase in the HTAG.***
- There is a negative correlation between the HTAG and FTR (-0.43) and in contrast, a positive correlation between the HTHG and FTR (0.45). Both provide strong evidence for the argument that ***the Home Team was likely to score during the second half of the match and won eventually,***

irrespective of the team being temporarily led during the first half of the match.

- If the Home Team was winning after the first half, it is likely that it secures the winning during the second half and wins the match eventually. It is shown by the positive correlation between the HTHG and FTR.

Principal Component Analysis

Theoretical Background

Principal Component Analysis (PCA) looks for a few linear combinations of *p* variables, losing in the process as little information as possible. More precisely, PCA transformation is an orthogonal linear transformation that transforms a p-variate random vector to a new coordinate system such that, the obtained new variables are uncorrelated, and the greatest possible variance lies on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on (Pauliina, 2020).

PCA Analysis

The dataset contains 13 features and plotting the data in its raw format makes it difficult to visualize the variation present. Thus, the author uses PCA to reduce the dimensionality of the data and tries to find components which represent the most variation of the data.

The R's prcomp package allows performing and summarizing the result of PCA using the summary built-in function:

```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10
Standard deviation     1.9356 1.4575 1.4033 1.05864 0.95500 0.91266 0.74917 0.69569 0.62480 0.55767
Proportion of Variance 0.2882 0.1634 0.1515 0.08621 0.07016 0.06407 0.04317 0.03723 0.03003 0.02392
Cumulative Proportion  0.2882 0.4516 0.6031 0.68928 0.75944 0.82351 0.86668 0.90391 0.93394 0.95787
                         PC11   PC12    PC13
Standard deviation     0.49837 0.4645 0.28910
Proportion of Variance 0.01911 0.0166 0.00643
Cumulative Proportion  0.97697 0.9936 1.00000
```

Figure 5: PCA descriptive summary

The author obtained 13 principal components, namely PC1-13. Each of these explains a percentage of the total variation in the dataset. By checking the Cumulative Proportion, we could see that PC1 to PC5 explains about 76% of the total variance.

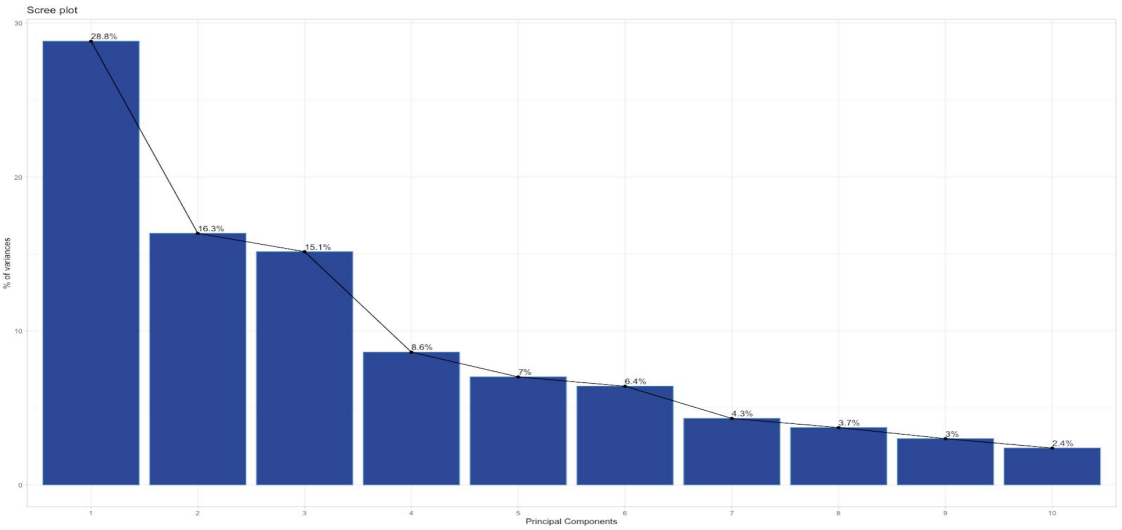


Figure 7: Scree Plot of the PC1-13 explained variance

Visualize the scores of the observations with respect to the first 4 components:

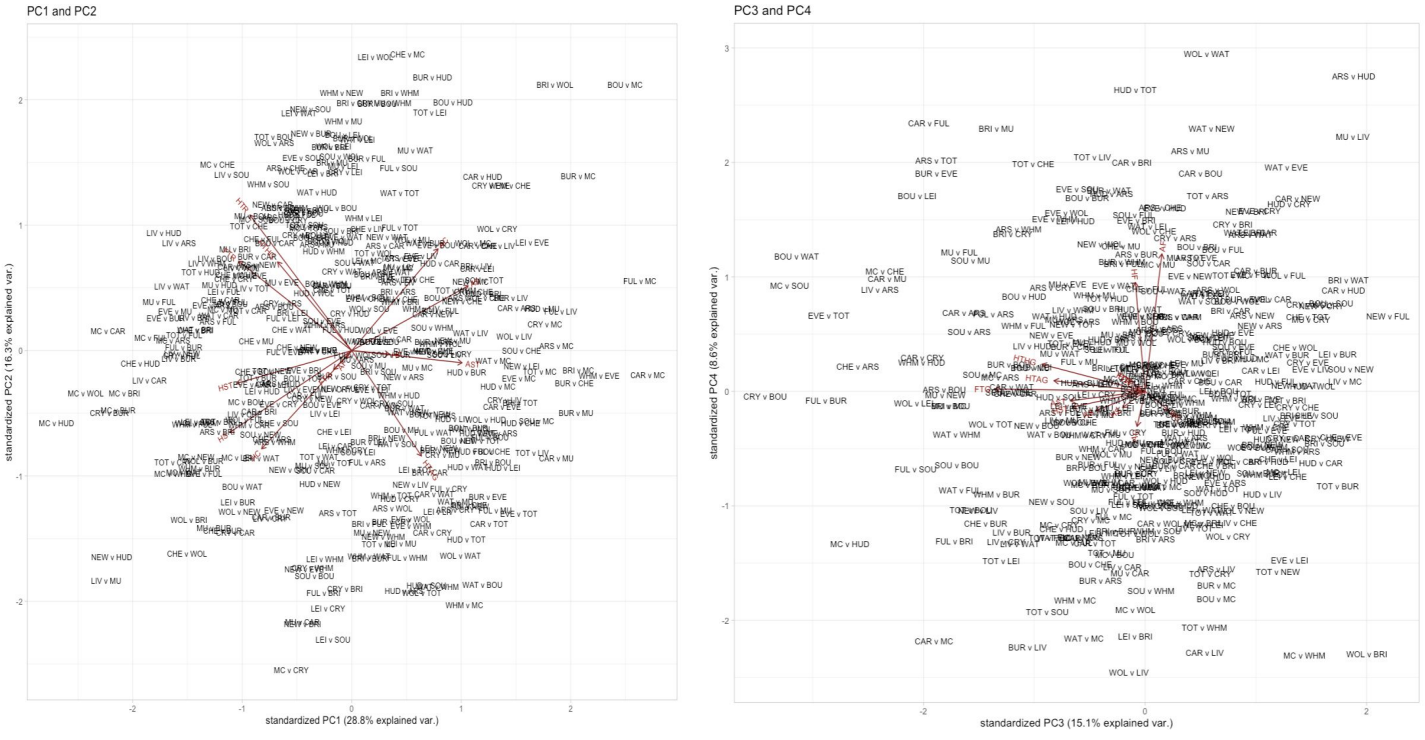


Figure 8: Biplot of the first 4 PCA components

To have a clearer understanding of the image, please refer to the section **Appendix.**

Interpretation:

- In the first two PCA biplot, we can see that on the left axis, there are mostly matches in which the strong team is the Home Team ***(Manchester City, Chelsea, Liverpool, Tottenham, Arsenal, Leicester City)***. A football lover would easily recognize these teams are always the dominating teams in the Premier League, and they are more likely to win the match against other teams.

- On the upper-left axis, *there are positive values for HTR, HTHG, HST* showing that in those matches ***the strong Home Team dominates from the beginning of the matches, it shot more, led by the half-time and won the final match***.
    + On the lower-left axis, *there is a group of negative values for HST, HS, HC, FTG and AF* indicating that in those matches ***the strong Home Team dominates the game by figures such as HST and HC, but the Away Team did not concede the goal easily and it tried to prevent the Home Team from scoring through committing the foul. The renowned strong Home Team still eventually won the match, but it faced difficulty and resistance from the Away Team.***
    + On the upper and lower right axis, these are matches in which the strong team played as the Away Team. *There is a group of correlating features such as AST, HTHG, AS, AC,* indicating that ***the strong team also played better without the support of the home ground.*** Thus, it can be inferred that ***the strong team is mostly unaffected and won the game regardless of where it played.***

- In the third and fourth PCA biplot, we can see that ***there is a diversity of teams participating in the match***, unlike one team being the strong team as of the first two PCA plot. On the left axis, there is a group of positive values for HTHG, HTAG, FTG, AST, HST, HS, AS and on the upper axis, there is a group of highly correlated and positive values for HF and AF. This ***indicates the harshness of the matches.*** Most of the matches that are in those groups have a high number of shots and a significant number of fouls committed by both sides. From a football watcher's perspective, those ***indicate several characteristics of the match, for example, both of the team are equal in strength, trying their best to prevent the other from scoring by committing fouls, shoot as much as the other and thus are not seeing complete domination in the match.***

K-means Clustering

Theoretical Background

The basic idea behind k-means clustering consists of defining clusters so that the total intra-cluster variation (known as total within-cluster variation) is minimized.

The author uses the Hartigan-Wong algorithm (Hartigan and Wong 1979), which defines the total within-cluster variation as the sum of squared distances Euclidean distances between items and the corresponding centroid:

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

- x_i design a data point belonging to the cluster C_k
- μ_k is the mean value of the points assigned to the cluster C_k

Each observation (x_i) is assigned to a given cluster such that the sum of squares (SS) distance of the observation to their assigned cluster centers u_k is a minimum. We define the total within-cluster variation as follow:

K-Means Clustering

$$tot.withinss = \sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

One problem with K-Means Clustering is that the number of clusters (k) must be set before we start the algorithm, so in order to know the optimal number of clusters, we need to use several different values of k and examine the differences in the results.

The author uses the **Elbow method** to determine the optimal number of clusters. The choice of method depends on the person conducting the analysis, and the final result might vary between different methods.

The results suggest that **3 is the optimal number of clusters** as it appears to be the bend in the knee (or elbow):

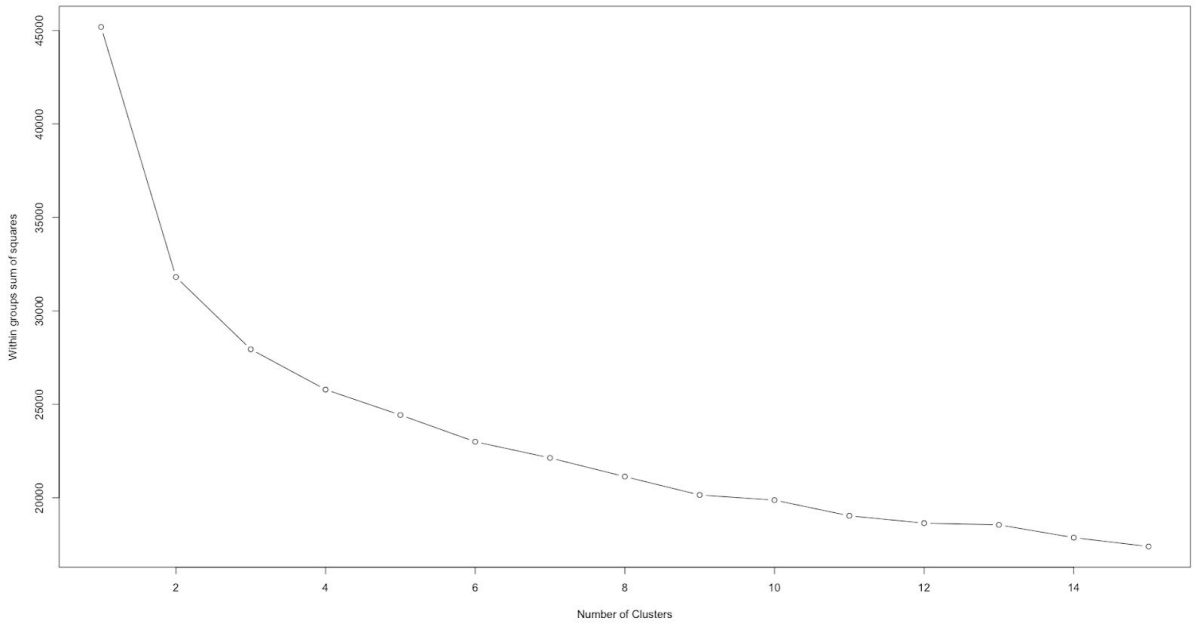


Figure 9: Within-group of sum of squares (WSSS) vs Number of clusters

Figure 10: k=3 clusters plot.

Observations:

- On the left cluster, ***these are matches that the home team wins in which the home team dominates the match through the high number of on-target shot and away-corner.*** In general, this ***consists of the matches in which the strong***

team plays as the Home Team, and thus shows the domination through the figures and secure eventual winning.

- On the right cluster, ***these are matches that the away team wins in which the away team dominates the match through the high number of on-target shot and away-corner.*** We could verify that in this case, ***the strong team plays against the Home Team as the Away Team***, for example, BRI v CHE, BUR v EVE, SOU v MC, etc.
- On the middle cluster, ***these are matches that the result is not predictable based on the given match figures. In some of the matches, the away team wins in which the home team dominates the match, shots more than the away team but then be concealed and lost against the away team.*** From a football lover's perspective, this ***occurs in the case when two teams are equal in strength***, for example, CHE v MC, WOL v CRY.

Based on the observations, we could draw reasonable explanations about the cluster of groups of similarity, and it verifies that the k-means model is able to sufficiently classify the matches.

Conclusion

Based on the analysis, the author evaluates the research questions and obtains the following conclusion. Each number represents the corresponding research question raised in the beginning of the report.

(1) By using the correlation matrix, the author is able to find the correlation between the variables. The cut-off value for the significant dependency between variables above 0.4 is reasonable to obtain the interpretation, and thus, proves itself to be a good choice in this task.

(2) It is possible to explain the variances with fewer components. Principal Component Analysis (PCA) shows that by using the first 4 principal components, we obtained about 76% of the total variance of the data. The components are also used for further clustering analysis.

(3) There are 3 groups of similarity for the recorded matches. Those are highly divided, interpretable, and descriptive by looking into the cluster plot.

(4) The author uses the K-Means clustering method to classify the similar matches into corresponding groups. The use of the optimal number of clusters is determined by the Elbow method, and the choice of k = 3 helps to classify and draws the distinct difference between the groups.

Critical Evaluation of the Analysis

The author performs the analysis on his knowledge about different multivariate statistical analysis methods. During the analysis, the author consults materials, for example, the lecture slide, the Internet to help facilitate the R-code and the explanation. However, the analysis explanation is written by his own opinion and interpretation. As the

abovementioned, it is likely that the result of the findings is different from other, since the choice of the parameter might be different and thus affects the final result.

It is highly recommended that one tries with different parameters to tune the accuracy of the used methods, and to see more findings could be found and explained.

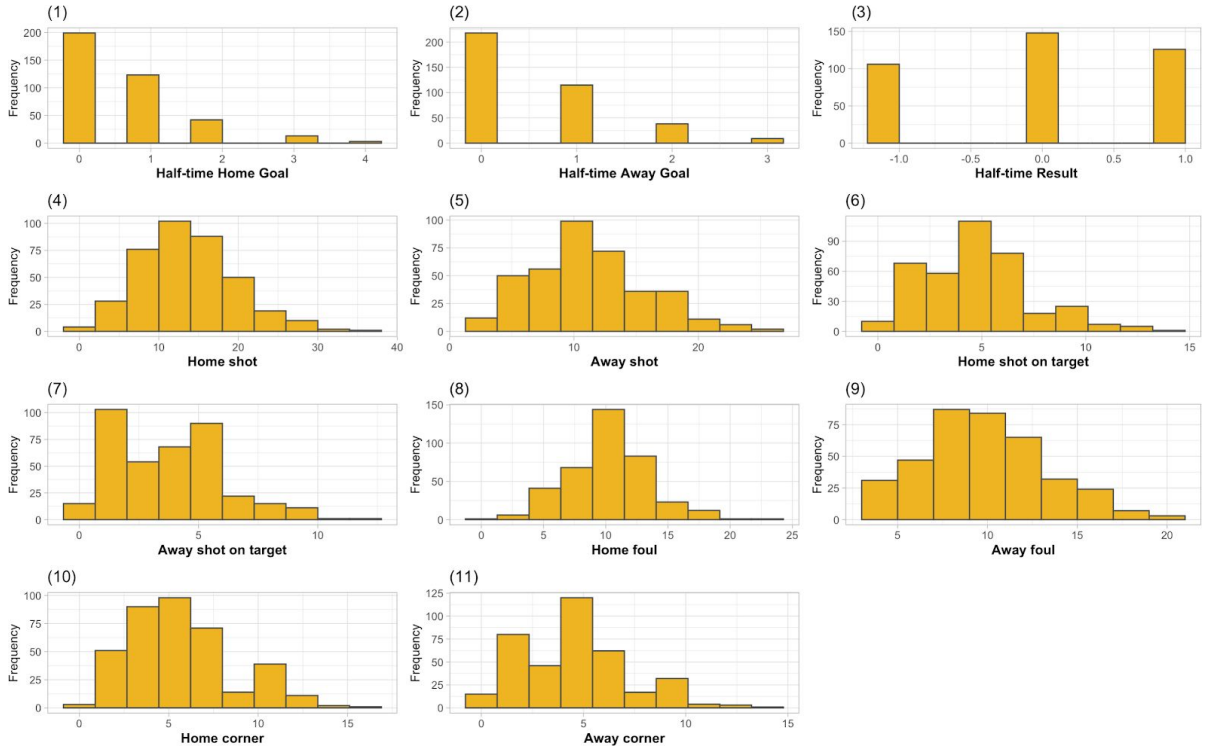
References

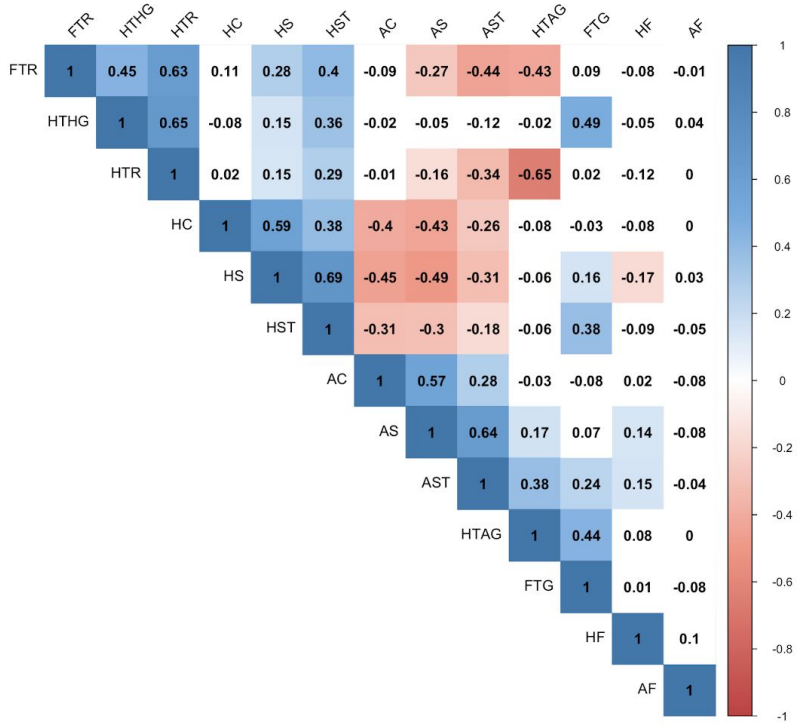
Pauliina, 2020. “Lecture 2: Principal Component Analysis”. Lecture slide. Page 4. Available from: https://mycourses.aalto.fi/pluginfile.php/1151573/mod_resource/content/1/2020Mult2.pdf. [Accessed 26th March 2020].

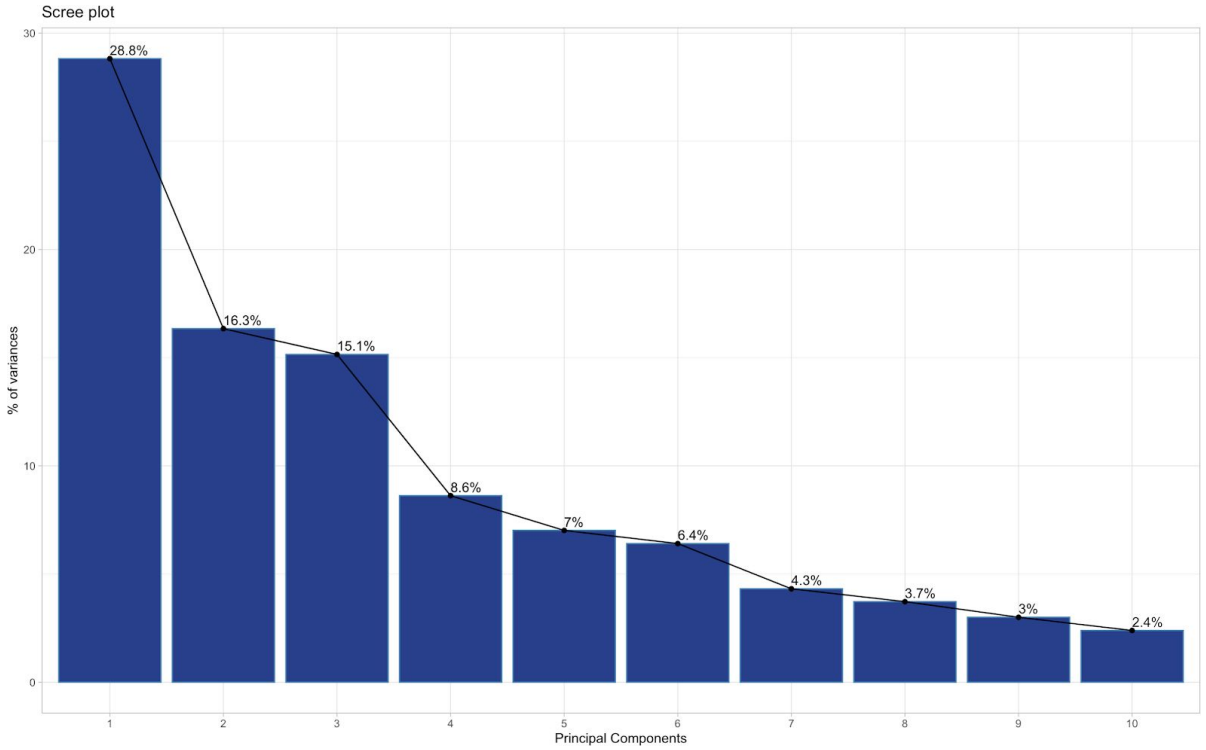
Hartigan, JA, and MA Wong. 1979. “Algorithm AS 136: A K-means clustering algorithm.”

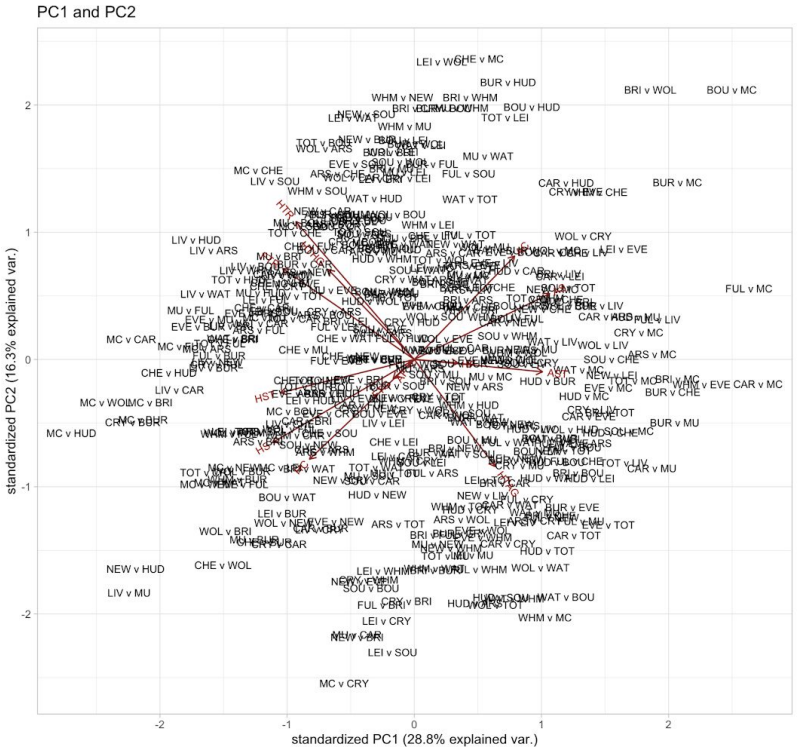
Applied Statistics. Royal Statistical Society, 100–108.

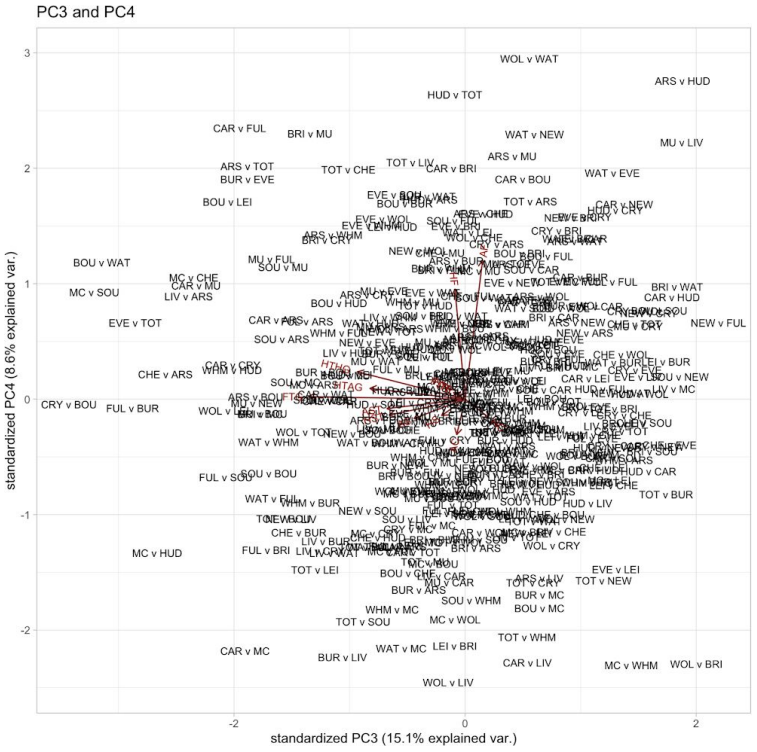
Appendix

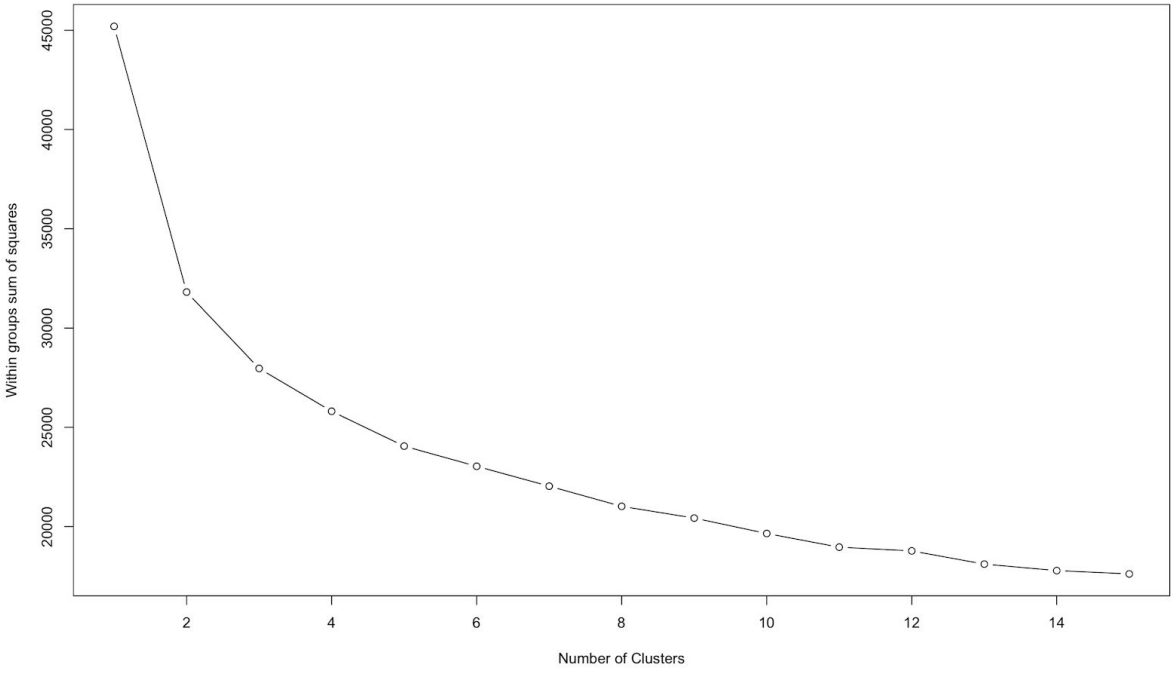

(1)
Frequency
Half-time Home Goal
(2)
Frequency
Half-time Away Goal
(3)
Frequency
Half-time Result
(4)
Frequency
Home shot
(5)
Frequency
Away shot
(6)
Frequency
Home shot on target
(7)
Frequency
Away shot on target
(8)
Frequency
Home foul
(9)
Frequency
Away foul
(10)
Frequency
Home corner
(11)
Frequency
Away corner

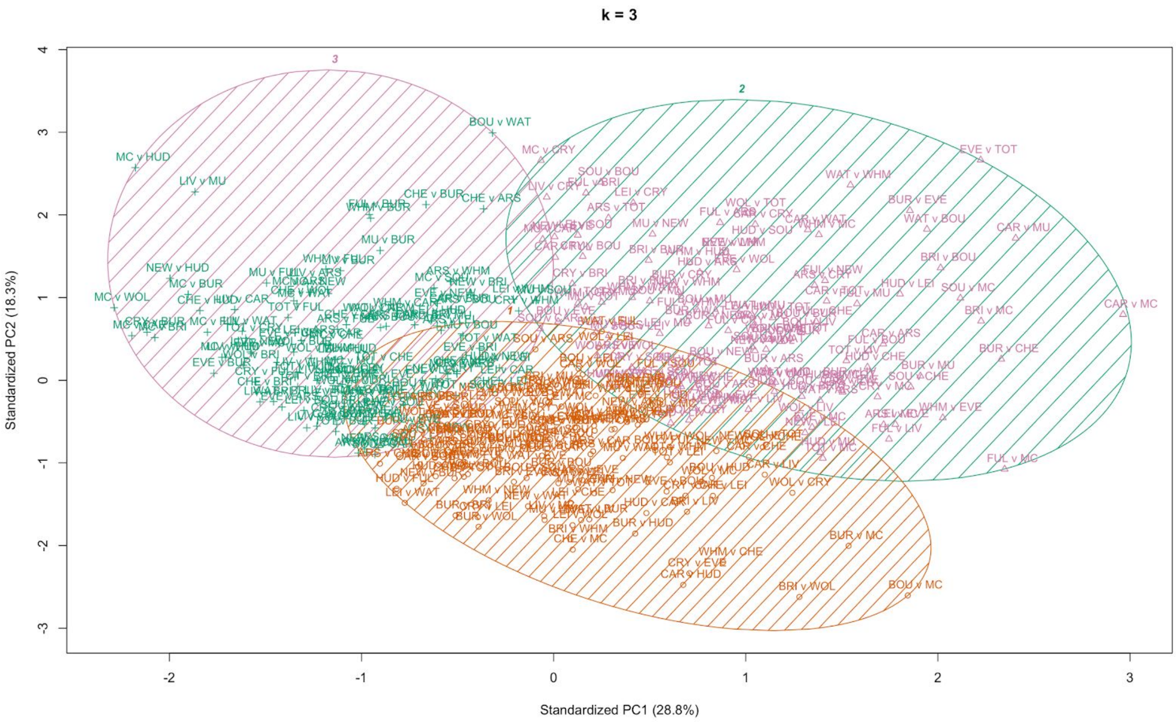

FTR HTHG HTR HC HS HST AC AS AST HTAG FTG HF AF
FTR 1 0.45 0.63 0.11 0.28 0.4 -0.09 -0.27 -0.44 -0.43 0.09 -0.08 -0.01
HTHG 1 0.65 -0.08 0.15 0.36 -0.02 -0.05 -0.12 -0.02 0.49 -0.05 0.04
HTR 1 0.02 0.15 0.29 -0.01 -0.16 -0.34 -0.65 0.02 -0.12 0
HC 1 0.59 0.38 -0.4 -0.43 -0.26 -0.08 -0.03 -0.08 0
HS 1 0.69 -0.45 -0.49 -0.31 -0.06 0.16 -0.17 0.03
HST 1 -0.31 -0.3 -0.18 -0.06 0.38 -0.09 -0.05
AC 1 0.57 0.28 -0.03 -0.08 0.02 -0.08
AS 1 0.64 0.17 0.07 0.14 -0.08
AST 1 0.38 0.24 0.15 -0.04
HTAG 1 0.44 0.08 0
FTG 1 0.01 -0.08
HF 1 0.1
AF 1
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1


Scree plot
% of variances
Principal Components
28.8%
16.3%
15.1%
8.6%
7%
6.4%
4.3%
3.7%
3%
2.4%


PC1 and PC2
standardized PC2 (16.3% explained var.)
standardized PC1 (28.8% explained var.)


PC3 and PC4
standardized PC4 (8.6% explained var.)
standardized PC3 (15.1% explained var.)
WOL v WAT
ARS v HUD
HUD v TOT
CAR v FUL
BRI v MU
WAT v NEW
MU v LIV
ARS v MU
TOT v LIV
ARS v TOT
TOT v CHE
CAR v BRI
BUR v EVE
CAR v BOU
WAT v EVE
BOU v LEI
TOT v ARS
BOU v WAT
MU v FUL
SOU v MU
BRI v FUL
MC v CHE
CAR v MU
MC v SOU
LIV v ARS
BRI v WAT
CAR v HUD
EVE v TOT
NEW v FUL
CHE v ARS
CRY v BOU
FUL v BUR
WAT v WHM
FUL v SOU
SOU v BOU
TOT v BUR
WAT v FUL
WHM v BUR
NEW v SOU
MC v HUD
FUL v BRI
TOT v LEI
SOU v WHM
WHM v MC
TOT v SOU
MC v WOL
TOT v WHM
CAR v MC
WAT v MC
LEI v BRI
BUR v LIV
CAR v LIV
MC v WHM
WOL v BRI
WOL v LIV


Within groups sum of squares
Number of Clusters
20000
25000
30000
35000
40000
45000
2
4
6
8
10
12
14



k = 3
Standardized PC1 (28.8%)
Standardized PC2 (18.3%)