REPORT II: Player's performance in Premier League (Fifa22)

Usama **ZAFAR** Bishwash **NEUPANE** Kirubel **WELDENTENSAY** Nethaji **RUTHIRAVEL**

Objective:

- Perform statistical analysis of a selected dataset using unsupervised learning techniques
- Perform Kmeans clustering in the original space and reduced dimension following PCA
- Use linear model and develop insights/inferences from the analysis

Dataset:

The dataset used for the analysis purpose is FIFA 22 available on Kaggle. As the dataset is large (more than 19000 players with 110 features), data cleaning is done in the original dataset. Players playing in **English Premier League**, having an **overall rating of more than 80** are considered **excluding the goalkeepers**. The goalkeepers are removed from consideration since they are too different from the field players who are defensive by nature and they can be easily clustered in a separate group (**only 10% of goalkeepers** in the initial sample size).

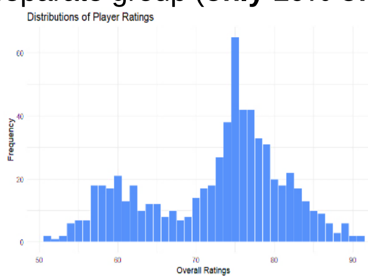

Figure 1(a) Distribution of overall player ratings

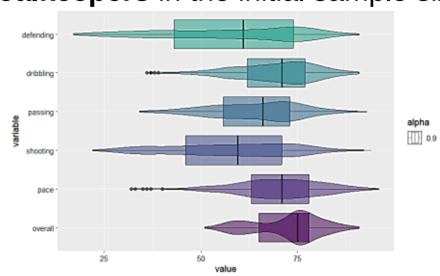

Figure 1(b) Violin plot for a few of the selected features

```
## [1] "pace"                          "shooting"
## [3] "passing"                       "dribbling"
## [5] "defending"                     "physic"
## [7] "attacking_crossing"            "attacking_finishing"
## [9] "attacking_heading_accuracy"    "attacking_short_passing"
## [11] "attacking_volleys"            "skill_dribbling"
## [13] "skill_curve"                  "skill_fk_accuracy"
## [15] "skill_long_passing"           "skill_ball_control"
## [17] "movement_acceleration"        "movement_sprint_speed"
## [19] "movement_agility"             "movement_reactions"
## [21] "movement_balance"             "power_shot_power"
## [23] "power_jumping"                "power_stamina"
## [25] "power_strength"               "power_long_shots"
## [27] "mentality_aggression"         "mentality_interceptions"
## [29] "mentality_positioning"        "mentality_vision"
## [31] "mentality_penalties"          "mentality_composure"
## [33] "defending_marking_awareness"  "defending_standing_tackle"
## [35] "defending_sliding_tackle"
```

Figure 1(c) Features based on which clustering is possible

Feature Discussion and Clustering Idea:

The dataset contains characteristics that can be used to cluster players into different groups: attacking, defensive, and midfield players. Age, potential, salary, and a few other features are excluded as they only stratify the players making the clustering meaningless. Distribution of overall rating and the violin plots **(Figure 1(a) and 1(b))** reveal that some characteristics have a **bimodal distribution**, indicating that clustering is feasible with the selected features. **Figure 1(c)** shows the list of features that are used to cluster the players. As there are a bunch of features a **minimal analysis** is reasonable. For this purpose, two data frames have been created, the first with features used for analysis and the second one with the categorical values with relevant information about the players that is used to interpret the result later on.

k-means in the Original Space

overall	shooting	passing	dribbling	defending	physic
Min. :80.00	Min. :30.0	Min. :52.00	Min. :57.00	Min. :33.00	Min. :57.00
1st Qu.:81.00	1st Qu.:59.5	1st Qu.:72.00	1st Qu.:76.00	1st Qu.:47.75	1st Qu.:69.75
Median :83.00	Median :72.5	Median :76.00	Median :80.50	Median :72.50	Median :75.00
Mean :83.34	Mean :69.0	Mean :76.03	Mean :79.34	Mean :65.62	Mean :74.48
3rd Qu.:85.00	3rd Qu.:79.0	3rd Qu.:81.00	3rd Qu.:84.00	3rd Qu.:81.00	3rd Qu.:80.00
Max. :91.00	Max. :94.0	Max. :93.00	Max. :91.00	Max. :91.00	Max. :88.00

Figure 2 Summary of the few features

After selecting relevant features from a dataset (42 variables), the k-means algorithm was applied to the original space without scaling as the selected features are on the same scale. The **Within-Cluster Sum of Squares (WSS)** graph is plotted to determine the optimal number of clusters, which is found to be **4**. The algorithm resulted in the formation of 4 clusters: **cluster 1 for attacking midfielders** (such as Kevin De Bruyn and Bruno Fernandes), **cluster 2 for defensive midfielders** (like Thomas Partey and Fernandino), **cluster 3 for defensive players** (such as Raphael Varane and Ruben Dias), and **cluster 4 for attackers/strikers** (such as Cristiano Ronaldo and Son Heung Min).

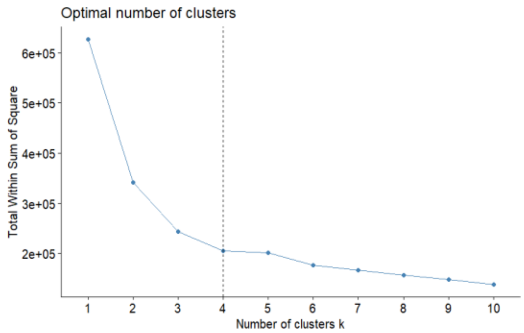

Figure 3(a) Optimal Number of Clusters

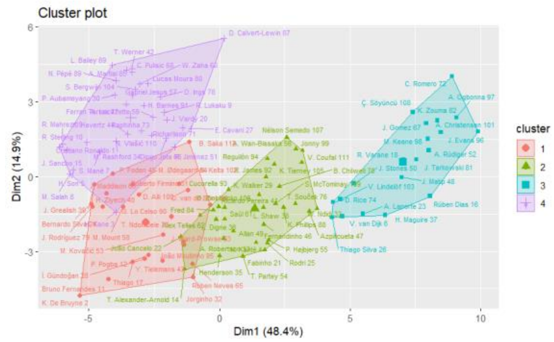

Figure 3(b) Clustering on original dimension

Principal Component Analysis and Dimensionality Reduction

The dimensionality of the feature space is **reduced (from around 35 to 5)** using the principal component analysis to explore the data on the factor plane. Analyzing **the factor plane**, it can be seen that the **first component(x-axis) is about the attacking** (features like dribbling, attacking finishing, power shots, movement balance, and attacking finishing), **the second one (y-axis) is about defense** (defending, mentality aggression, defending sliding tackle). Also, it is easy to see that **good midfielders tend to exhibit qualities that are closer to the y=x line**. It can be seen in the **PCA graph of individuals**, the **best attackers** end up on the **right end of the x-axis** (M. Salah, Son, Cristiano Ronaldo), the **best defenders on the y-axis** (Thiago Silva, Maguire, Stones), and the **best midfielders on the farther end of the y=x line** (K. De Bryune, Bruno Fernandes).

Looking at the **scree plot** to see how well the two components plane **describes the variance in the dataset**, it is around **65% which is decent** and indicates interpretation and inferences made using this planes make sense.

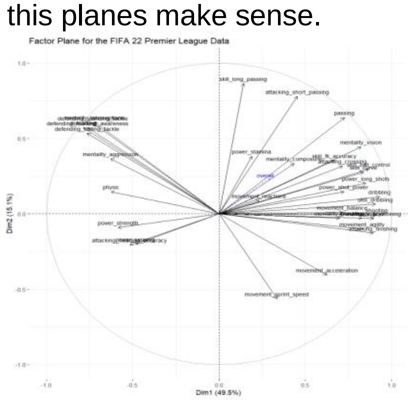

Figure 4(a) Factor Plane

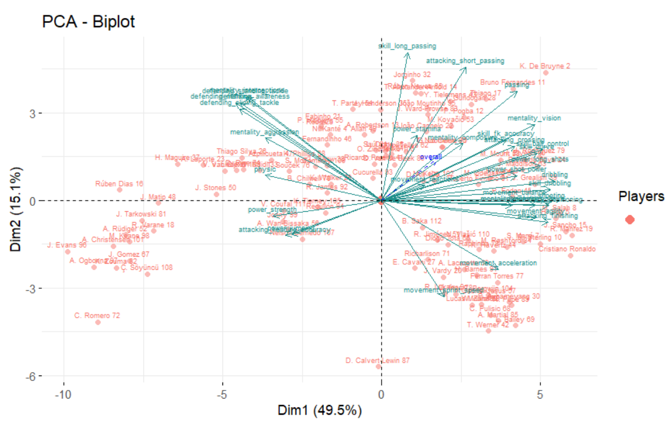

Figure 4(b) PCA Biplot

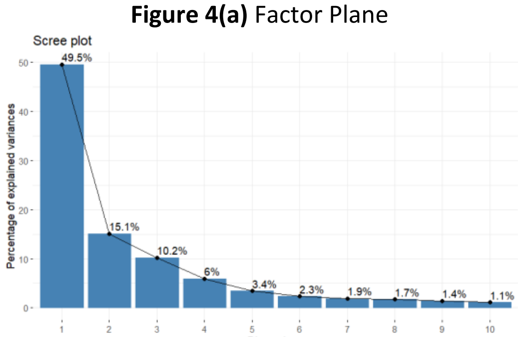

Figure 4(c) Scree Plot

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
shooting	0.9141771	-0.0064744	0.3285974	0.0000640	0.0961080
passing	0.7264262	0.6394850	-0.0459907	-0.0465993	0.0881217
dribbling	0.9316159	0.1132483	-0.1072490	0.1504085	-0.1576981
defending	-0.7853845	0.5716969	-0.0701256	0.1600101	0.0189975
physic	-0.6257763	0.1465741	0.6065795	0.3340613	0.0902804
attacking_crossing	0.7161563	0.3167059	-0.2034897	0.1861484	0.2604805
attacking_finishing	0.8945097	-0.1275912	0.2937005	-0.0150552	-0.0038267

Figure 4(d) PCA Coordinates

k-means in Reduced Space

After the dimensionality reduction, k-means clustering is done by extracting the PCA coordinates of the reduced dimensions. Just like clustering done in the original space, the optimal number of clusters using the elbow method is identified (4), and clusters were visualized. Cluster 1 consists of all the defensive players, cluster 2 consists of strikers, cluster 3 consists of defensive midfielders, and attacking fullbacks and Cluster 4 consists of attacking midfielders.

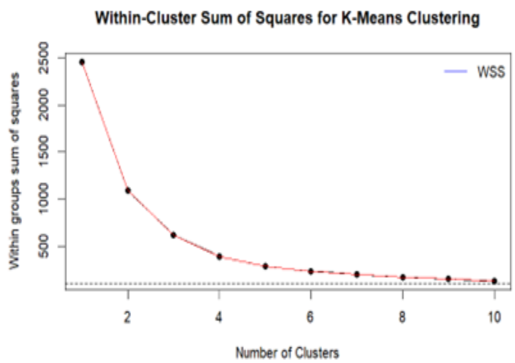

Figure 5(a) Optimal Number of CLusters

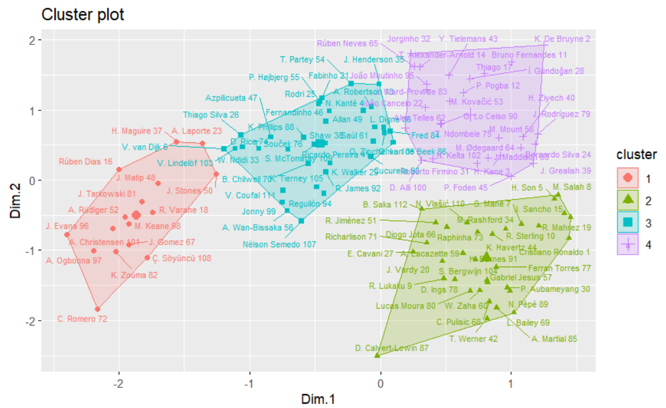

Figure 5(b) Clustering in Reduced Dimension

Comparison between k-means clustering (original and reduced space)

To compare the performance of clustering models in original and reduced dimensions, inertia, clustering visualization, and the number of players on each cluster are taken into consideration for both models.

- **The inertia of k-mean clustering** is a measure of how well the data points in each cluster are grouped. Lower Inertia indicates better clustering. In this case, the inertia for the original dimension is much larger **(205447.8)** than the inertia for the reduced dimension after PCA **(394.6016).** This indicates that the clustering performance of the k-means algorithm is much better in the reduced dimension after PCA. This is because PCA has reduced the dimensionality of the data, removed redundant or noisy features, and focused on the most important ones.
- **Visualization:** It can be visualized from **Figure 3(b)** that there are four clusters but clusters 1,3 and 4 are **overlapped meaning** the clusters are **not well separated**. However, in **Figure 5(b)** after the dimension reduction, the clusters **are well separated** and there are no overlaps as well. This indicates that k-means after PCA is more accurate to identify the underlying structure of the data in the reduced dimension space and assign data points to their respective clusters. Also, PCA has been successful in capturing the most important and distinctive features of the data set that allow for a clear separation between clusters.
- **Cluster Numbers:**

k-means model	**Defenders**	**Strikers**	**Attack. Midfield**	**Def. Midfield**
Original Space	20	33	26	33
Reduced Space	35	33	28	16

After performing the dimensionality reduction, the clustering of players has significantly improved. In the original space, there were 33 defensive midfield players, but this was reduced to 16 after dimensionality reduction. Additionally, the number of defenders increased from 20 to 35 as previously, the left back and full-backs were clustered as defensive midfield players, despite having more features of defensive players. This demonstrates the significance of dimensionality reduction in data analysis and machine learning.

Linear Regression

A simple linear regression multimodal is taken into consideration for analyzing the relationship between a **response/dependent variable (y= value_euro)** and its interaction with one or more **independent variables (x= {age, height, weight, overall, pace, shooting, passing, dribbling, defending, physic})**.

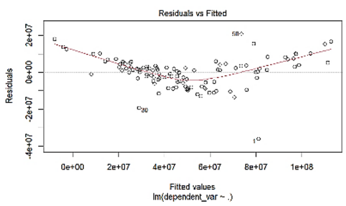

Figure 6(a) Residual vs Fitted

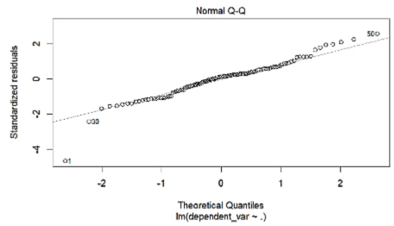

Figure 6(b) Q-Q plot

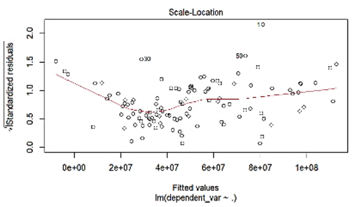

Figure 6(c) Scale-Location

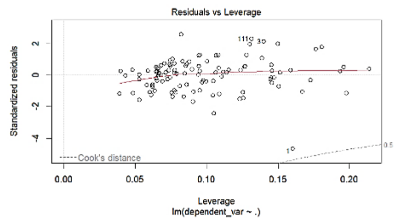

Figure 6(d) Residual vs Levrage

```
Call:
lm(formula = dependent_var ~ ., data = data_lm2)

Residuals:
     Min       1Q   Median       3Q      Max
-36062391 -4953981  1162101  4345130 20911816

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -571016647   45359407 -12.589  < 2e-16 ***
age           -4331004     279510 -15.495  < 2e-16 ***
height_cm       -93104     228751  -0.407  0.68486
weight_kg      -278217     178054  -1.563  0.12129
overall        9941471     396431  25.077  < 2e-16 ***
pace           -312591     113825  -2.746  0.00714 **
shooting       -104981     137293  -0.765  0.44626
passing         118025     238209   0.495  0.62135
dribbling      -259775     292282  -0.889  0.37623
defending      -235420     107907  -2.182  0.03145 *
physic           30799     202180   0.152  0.87923
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8463000 on 101 degrees of freedom
Multiple R-squared:  0.9141,	Adjusted R-squared:  0.9056
F-statistic: 107.5 on 10 and 101 DF,  p-value: < 2.2e-16
```

Figure 6(e) Linear Model Summary

Based on the summary of the linear regression model, it can be concluded that the market value (in euros) of players is significantly **related to age and overall rating**. This means that **younger players** and those with **higher overall ratings** tend to have higher market values. Additionally, the pace and defending ratings are also statistically significant, but to a lesser extent than age and overall rating. However, features such as weight, shooting, passing, dribbling, and physics do not appear to have a significant relationship with the market value of players in the Premier League. The residual standard error gives a difference between the observed value and the predicted values. In this case, **the residual standard error is 8.46 million euros**, which means that the model's prediction is off by around 8.46 million euros and it seems agreeable as the young players with a rating at around 90 have a market value of more than 100 million.

Multiple R squared value measures how well the independent variables in the model explain the variation in the dependent variable. In this case, the **multiple R-squared is 0.9141**, which means the independent variables in the **model explain around 91.41%** of the variation in the dependent variables. The **adjusted R-squared value** penalizes the inclusion of unnecessary independent variables that do not improve the model's performance. The adjusted R-squared value is **0.9056**, which is slightly lower than the multiple R-squared, indicating that some of the independent variables do not add much to the model's performance.

A high F-statistic and a low value (less than 0.05) indicate the model's fit is statistically significant. In this case, the F-statistics is 107.5, and the p-value is less than 2.2e-16, which means the **overall fit of the model is highly significant.**

Conclusion

All the objectives defined in the first section of this report have been achieved and the report is made more concise and to the point. Logistics regression is not considered here as it is optional and also, the report seemed to exceed the limit. The learning experience for this course has been great so far. Starting from the basics of inferential statistics and hypothesis testing, statistical modeling, and ending up with the machine learning frameworks: mathematical foundation and practical application through the project/lab sessions in R has been a good exploration curve. Even more, when applying mathematical concepts through a project on the selected dataset for analysis and logical reasoning has helped to clear doubts and enhanced the learning ability.