Unsupervised Learning - FIFA 23

DTSA 5510
Final Project

Project Agenda

1. Project Introduction
2. Data Cleaning
3. Exploratory Data Analysis
4. Cluster and Classification
5. Conclusion

1. Project Introduction

Dataset

- FIFA23 1st Edition Player Database
- From Kaggle Website
- Players' Information and Attributes

Goal

Explore the positions of players
based on their attributes
and accurately predict them.

2.2 Cleaning Steps

Feature Selection

```
df1 = df[['ID', 'Name', 'Club', 'BestPosition',
          'Overall', 'PaceTotal', 'ShootingTotal'
          'PassingTotal', 'DribblingTotal',
          'DefendingTotal', 'PhysicalityTotal']]
print(df1.shape)
```

```
(18539, 11)
```

Delete Duplicates

```
duplicate_rows = df1[df1.duplicated()]
duplicate_rows.shape
```

```
(119, 11)
```

Checking Missing Value

```
missing_values = df1.isna().sum()
missing_values.info()
```

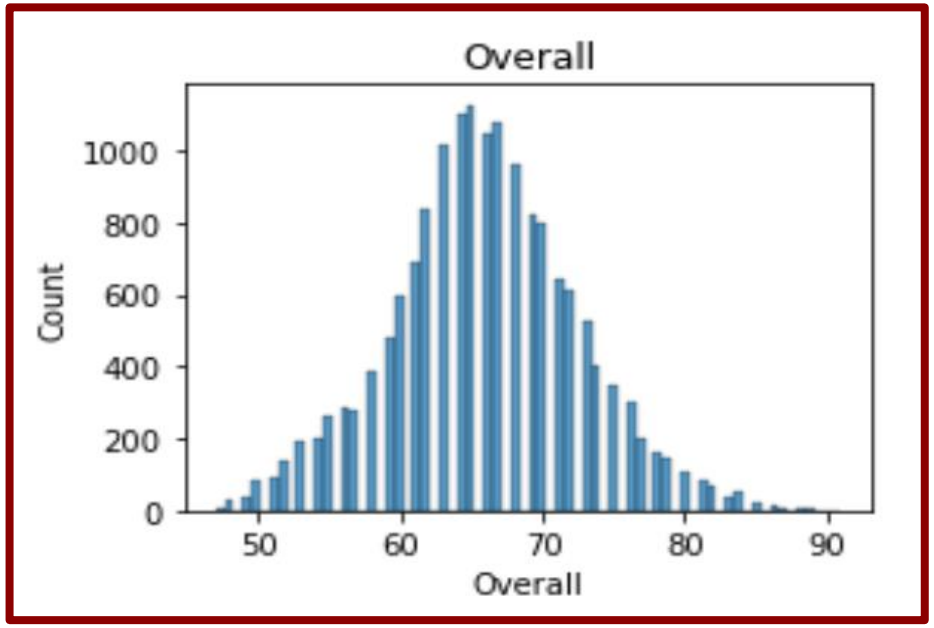
```
<class 'pandas.core.series.Series'>
Index: 11 entries, ID to PhysicalityTotal
Series name: None
Non-Null Count  Dtype
--------------  -----
11 non-null     int64
dtypes: int64(1)
memory usage: 176.0+ bytes
```

Subset Selection

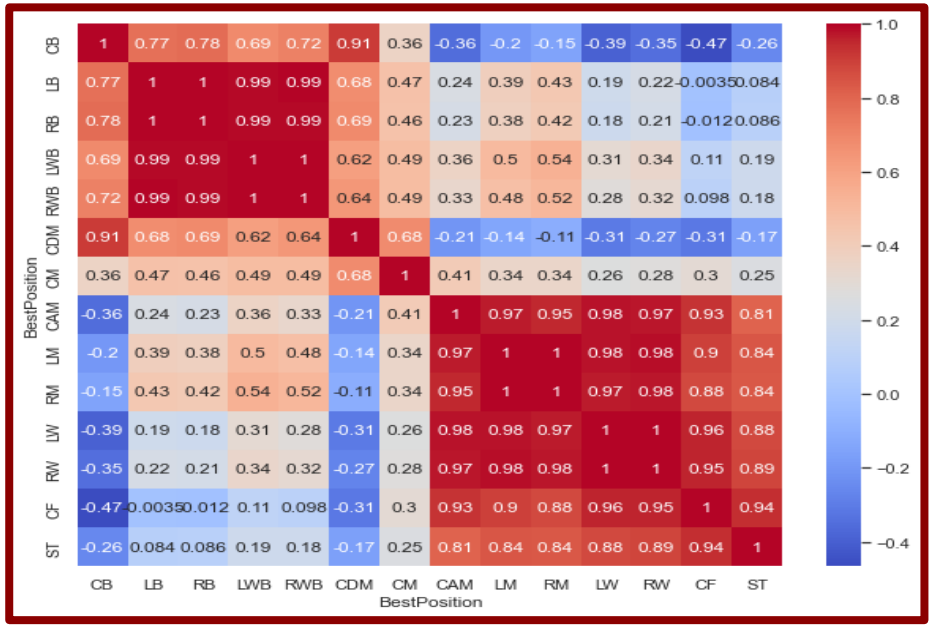
```
# Remove Goalkeeper
df3 = df2[df2['BestPosition'] != 'GK']
print(df3.shape)
```

```
(16367, 11)
```

3. Exploratory Data Analysis



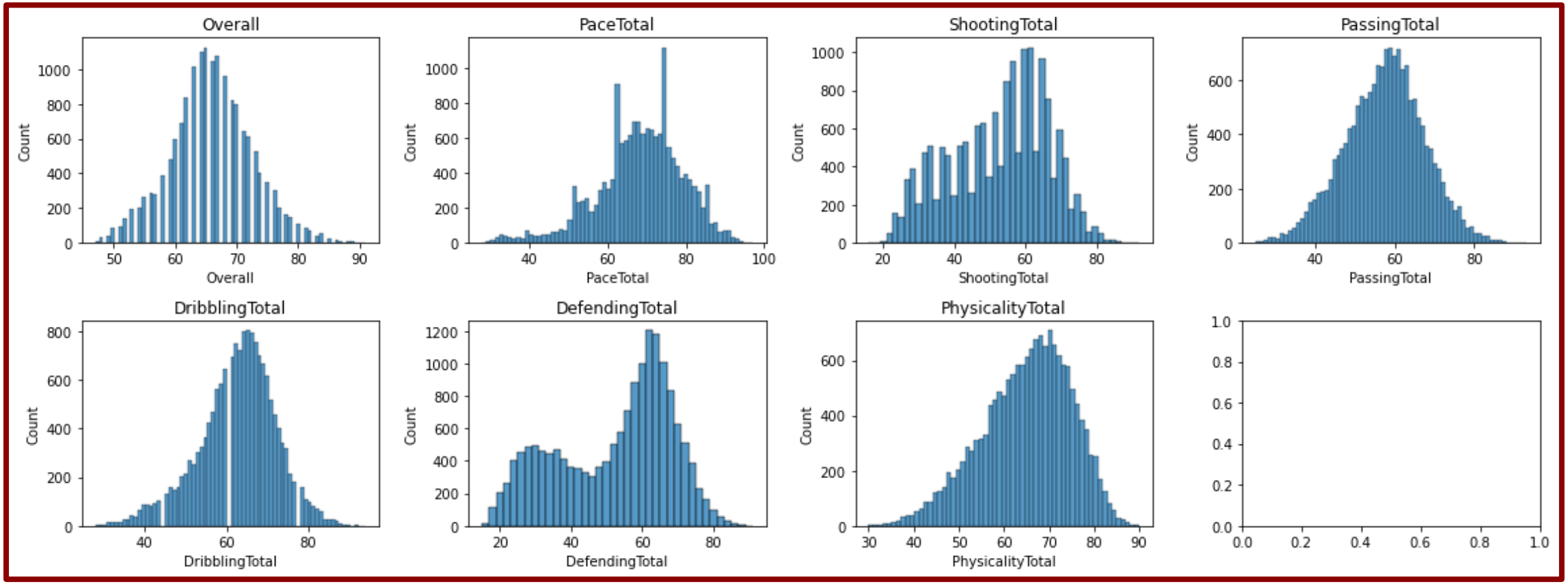
Distribution



Correlation

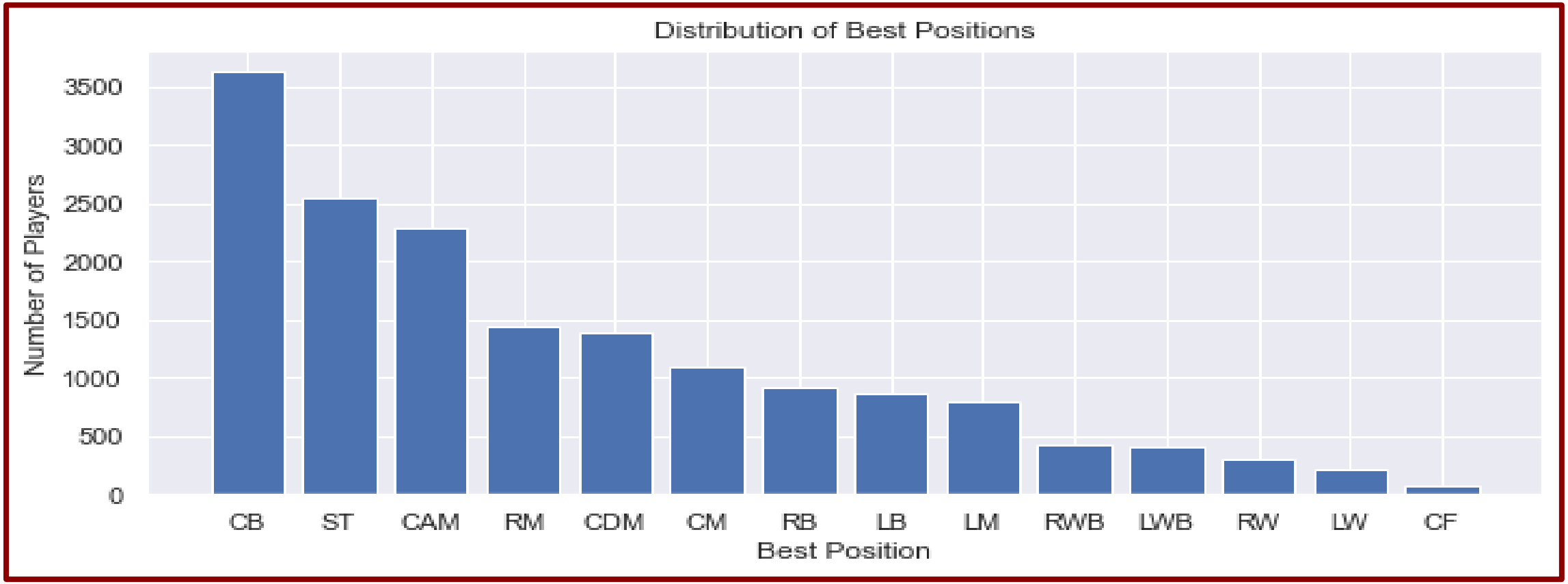
3.1 Attribute Distribution

- **General attribute is very likely to be able to predict the position.**

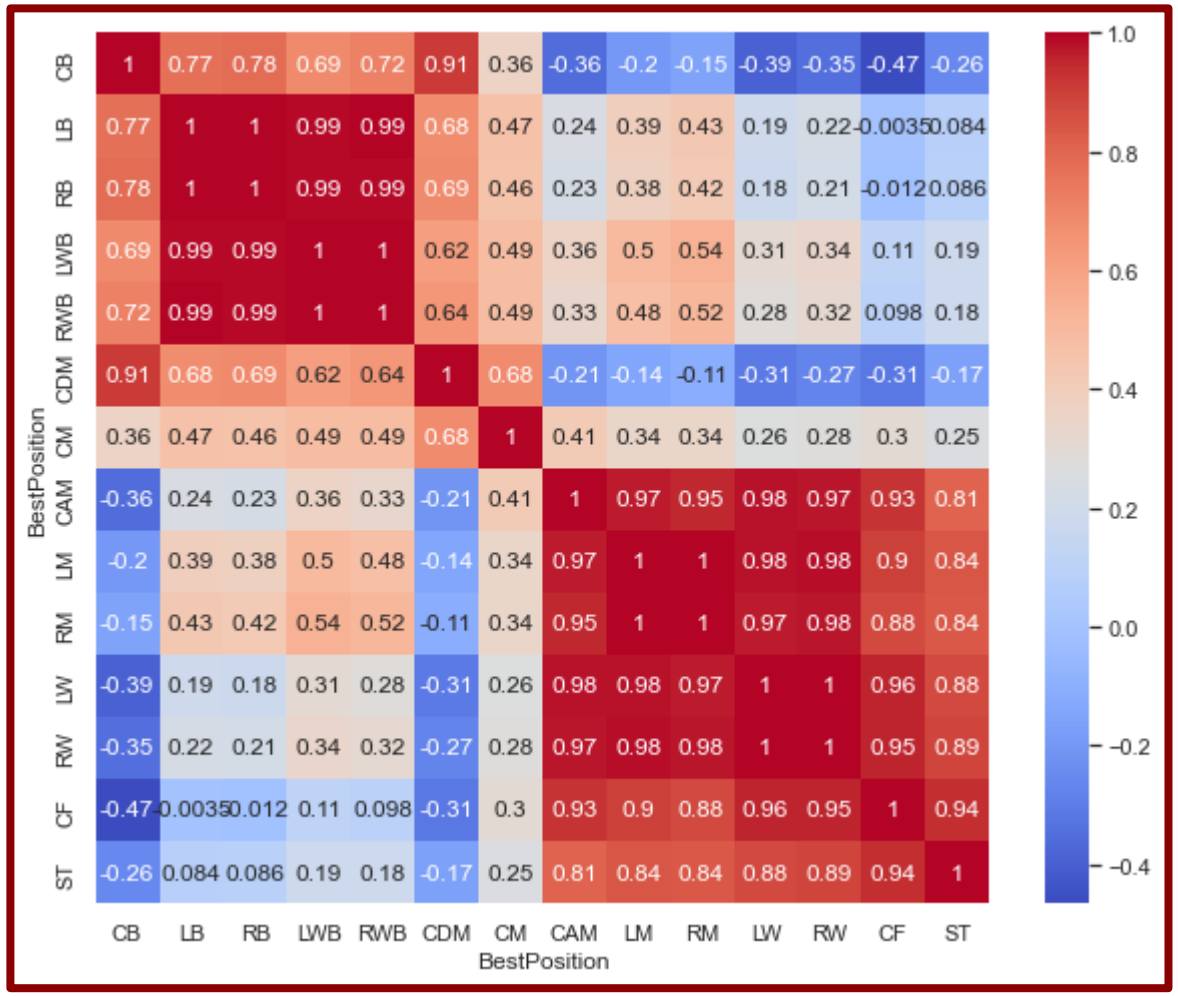


3.2 Position Distribution

- **Based on the understanding of the real world, many positions on the soccer field are similar to each other.**



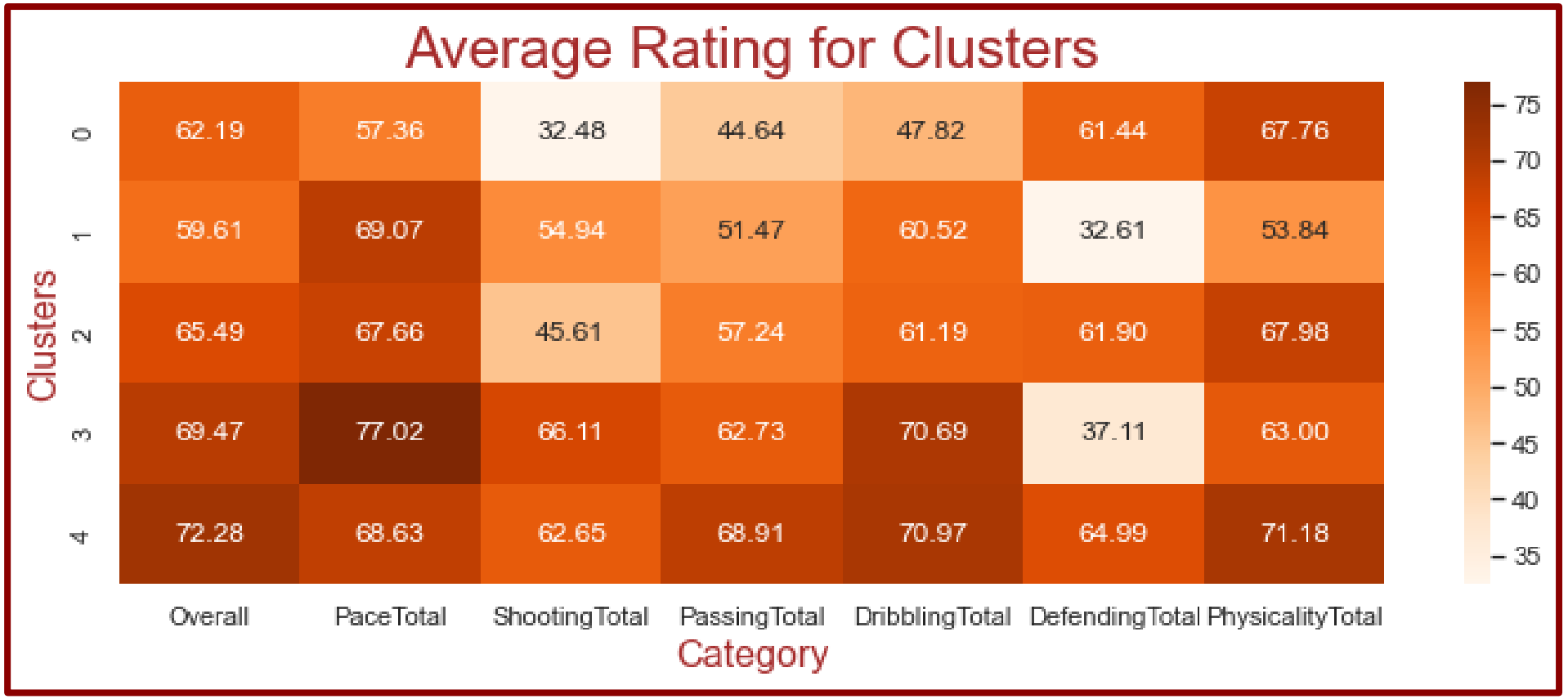
3.3 Position Correlation



- Heat Map shows the strong correlation between positions.
- Positions should be clustered by the nature of similarities within the dataset.

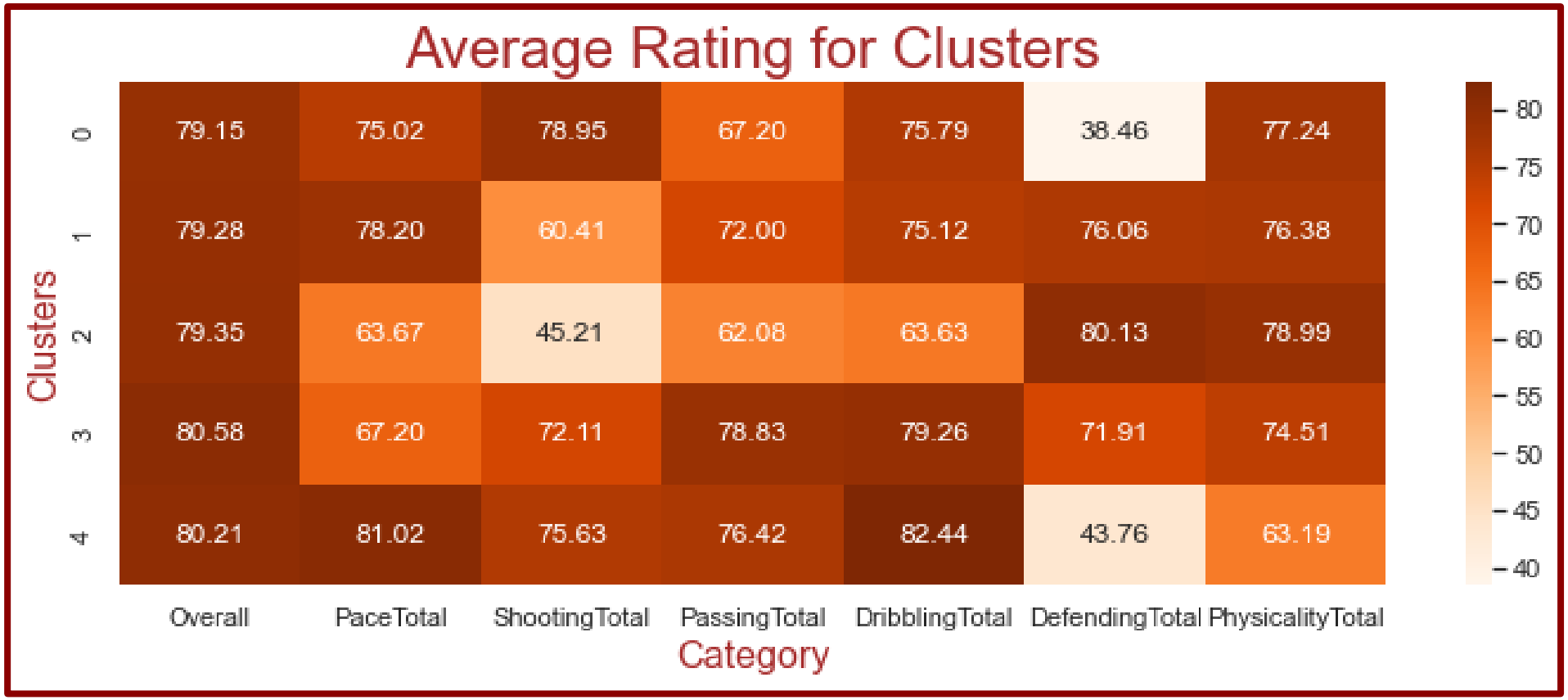
4.1 First K-means Clustering

- The significant difference in the average Overall Rating among the five clusters of players indicates that:
- K-means algorithm prioritizes the classification of players with different overall levels rather than grouping them based on the distribution of their attributes.



4.2 Second K-means Clustering

- Narrow Down the number of players to Top 1000.
- K-means algorithm perfectly grouped the players into five categories, and each category of players has different tendencies in their attributes.



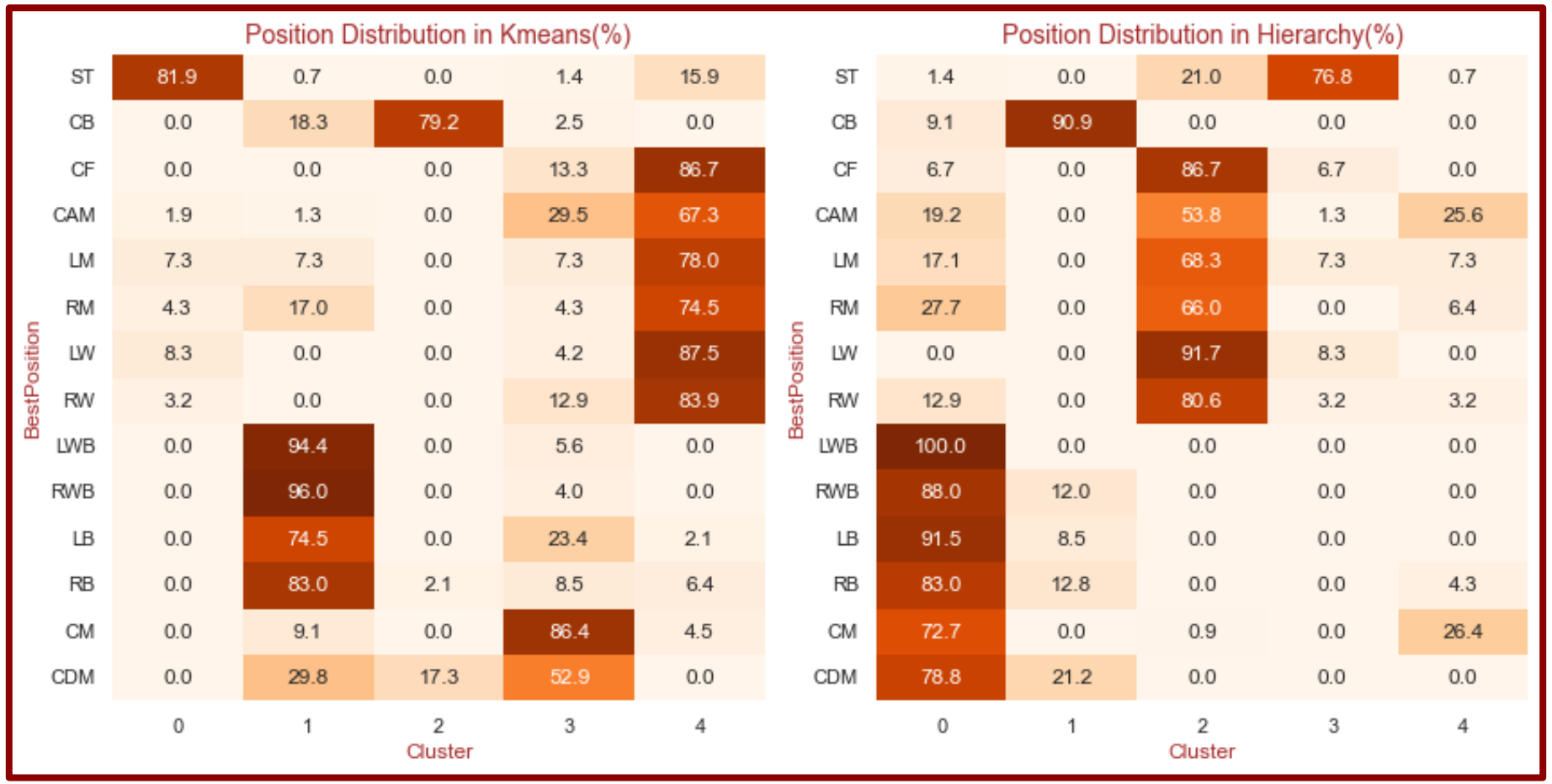
4.3 Double Checking

Name	Kmeans
A. Rebić	0
O. Giroud	0
Z. Ibrahimović	0
D. Origi	0
D. Calabria	1
P. Kalulu	1
S. Dest	1
T. Hernández	1
S. Kjær	2
F. Tomori	2
I. Bennacer	3
S. Tonali	3
R. Krunić	3
C. De Ketelaere	3
Rafael Leão	4
A. Florenzi	4
A. Saelemaekers	4
Junior Messias	4
Brahim	4

Name	Kmeans
R. Lukaku	0
L. Martínez	0
E. Džeko	0
M. Darmian	1
D. Dumfries	1
R. Gosens	1
D. D'Ambrosio	1
F. Dimarco	1
M. Škriniar	2
F. Acerbi	2
S. de Vrij	2
A. Bastoni	2
H. Çalhanoğlu	3
N. Barella	3
M. Brozović	3
H. Mkhitaryan	4
J. Correa	4

4.4 Compare to Hierarchy Clustering

- In K-means clustering, the positions of players in each cluster are very focused and concentrated.
- In Hierarchy, Cluster 0 and 2 intersect with each other, and cluster 5 has no meaning.



4.5 Label Clusters

>>cluster0 = Striker

>>cluster1 = Wing-defender

>>cluster2 = Center-back

>>cluster3 = Mid-field

>>cluster4 = Wing-attacker

4.6 Random Forest Classifier

- **Random Forest achieved an overall accuracy of 94%.**
- **For each position, the precision and recall were both over 90%.**

```
best parameters: {'max_depth': 10, 'n_estimators': 200}
best score: 0.9337500000000001
              precision    recall  f1-score   support

  Center-back       0.91      1.00      0.95        29
   Mid-field        0.98      0.93      0.95        55
     Striker        0.96      0.93      0.95        28
Wing-attacker       0.90      0.98      0.94        45
Wing-defender       0.97      0.91      0.94        43

    accuracy                            0.94       200
   macro avg        0.94      0.95      0.95       200
weighted avg        0.95      0.94      0.95       200
```

4.7 AdaBoost Classifier

- **Adaboost did not perform well on this dataset.**
- **There was a large prediction bias for wing players, and there were significant misclassifications compared to forwards.**

```
best parameters: {'learning_rate': 0.5, 'n_estimators': 300}
best score: 0.85875
```

	precision	recall	f1-score	support
Center-back	0.89	0.83	0.86	29
Mid-field	0.88	0.89	0.88	55
Striker	1.00	0.75	0.86	28
Wing-attacker	0.84	0.96	0.90	45
Wing-defender	0.84	0.88	0.86	43
accuracy			0.88	200
macro avg	0.89	0.86	0.87	200
weighted avg	0.88	0.88	0.87	200

4.6 Gradient Boosting Classifier

- The overall prediction accuracy of Gradient Boosting reached 94%.
- However, the accuracy for wing-defender positions is relatively low.

```
best parameters: {'learning_rate': 1, 'max_depth': 5, 'n_estimators': 100}
best score: 0.92875
              precision    recall  f1-score   support

  Center-back       0.97      1.00      0.98        29
    Mid-field       0.96      0.87      0.91        55
      Striker       1.00      0.89      0.94        28
Wing-attacker       0.94      1.00      0.97        45
Wing-defender       0.85      0.93      0.89        43

     accuracy                           0.94       200
    macro avg       0.94      0.94      0.94       200
 weighted avg       0.94      0.94      0.93       200
```

4.8 Model Summary

Clustering	**K-means performs better to explore the relationship between player attributes and positions.**
Classification	**Random Forest and Gradient Boosting successfully predicted the results of K-means clustering.**

5. Final Conclusion

Data	**The attributes of players in FIFA23 are consistent with the real world, and we can use classification to predict and prove the rationality of clustering.**
Method	**The natural distribution of the dataset may affect our analysis, but through continuous attempts and elimination, there is always a way to achieve or approach the goal.**

Unsupervised Learning - FIFA 23

Thank you for watching!