Lab Report – 2A

CSE564: Visualization

Visual analysis of football players from FIFA 2022 database

Hritam Basak

Aim: The project aims to teach several important aspects of data visualization. First, we learn to do basic dimension reduction and data visualization with PCA using Interactive Scree Plot and PCA- based Biplot. We also use the PCA components < user selected dimensionality index and obtain 4 attributes with highest PCA loadings. Plot the attributes using a scatterplot matrix using k means clustering. Additionally, we visualize the best number of clusters from the property of the data for better K-Means Clustering.

Description of Dataset: The dataset has been scrapped from publicly available sofifa website, cleaned, and made public on Kaggle. The datasets provided include the players data for the Career Mode from FIFA 22. This data allows multiple comparisons for different players across 100+ attributes in the game.

The dataset contains information about 19,239 players and their 121 attributes. The dataset has been cleaned and formatted for the specific requirement of this project. Among these 26 attributes are selected for 550 players (randomly selected players). The attributes can be distributed in the following categories:

Numerical Data:

a. "age": Age of player in Fifa 22 game.
b. "overall": The overall rating of the player in Fifa 22. A good player has a high (>80 rating).
c. "height_cm": The height of a player. It is positively correlated to player's heading skills.
d. "weight_kg": The weight of player in kg.
e. "pace": Indicates how fast a player can run. Interestingly some of the older players had very high pace values.
f. "shooting": Shooting skills and accuracy of a player. Generally, a striker has high shooting skills, whereas defenders and goalkeepers do not have that.
g. "passing": How good is the passing (both ground and lofted pass). Midfielders are known for their passing skills.
h. "dribbling": How good the player can dribble. Midfielders and strikers are known as the best dribblers.
i. "defending": Defending skill of a player. Needless to mention, defenders have very high value, as compared to strikers.
j. "physic": Physical contact attribute of a player. Muscular players are having high score.
k. "attacking_crossing": Crossing skill of a player. Generally, wingers have a high value.
l. "attacking_heading_accuracy": How accurate is a player's heading. Surprisingly defenders have somewhat high value in heading, as compared to the strikers, as heading is a part of defending as well.
m. "attacking_short_passing": Skill of short pass and move. Key attribute for hole-players.
n. "skill_dribbling": How likely is a player to dribble past a defender. Also relates to number of different dribbling tricks.
o. "skill_curve": A high value of this attribute indicates that a player can curl the ball well. Iconic players like David Beckham have very high value.
p. "power_shot_power": Attributes to the generated power while shooting of a player.
q. "power_jumping": How good can a player jump. Strikers and defenders have good jumping skills.
r. "mentality_aggression": This indicates how likely is a player to engage in attacking build-up.

s. “mentality_interceptions”: This indicates how likely a player can intercept a ball. Key skill for defensive midfielders and defenders.
t. “mentality_vision”: This is a very unique skill of some players, who can map the game really well, and has ability to create chances by providing through balls from interpreting the opponent’s movements.
u. “mentality_penalties”: How good penalty taker is the player.
v. “wage_eur”: Wage of a player in euro in their current clubs. Generally international players have very high wages.

After cleaning the initial data using a python script (Preprocess_data.ipynb), the final data is saved in CSV format in the file final_data.csv.

Why did I think this dataset was interesting?: FIFA 22 player data contains a vast amount of information on the performance and attributes of individual players in the game. By analyzing this data, we can gain insights into player characteristics, strengths, weaknesses, and tendencies, which can be useful for a variety of projects, such as:

1. Building a winning team: By analyzing player data, we can identify players who have the skills and attributes needed to complement one another and build a strong team.

2. Developing player ratings: FIFA player ratings are determined by a complex algorithm that takes into account a wide range of factors, including player performance data. Analyzing this data can help us understand how player ratings are calculated and how they can be improved.

3. Predicting game outcomes: By analyzing player data and team statistics, we can make more accurate predictions about the outcome of FIFA matches.

4. Improving player performance: Coaches and players can use FIFA player data to identify areas of weakness and develop strategies for improving player performance.

Overall, FIFA 22 player data is a valuable resource for anyone interested in understanding the game of soccer and the factors that contribute to player and team success.

Implementation:

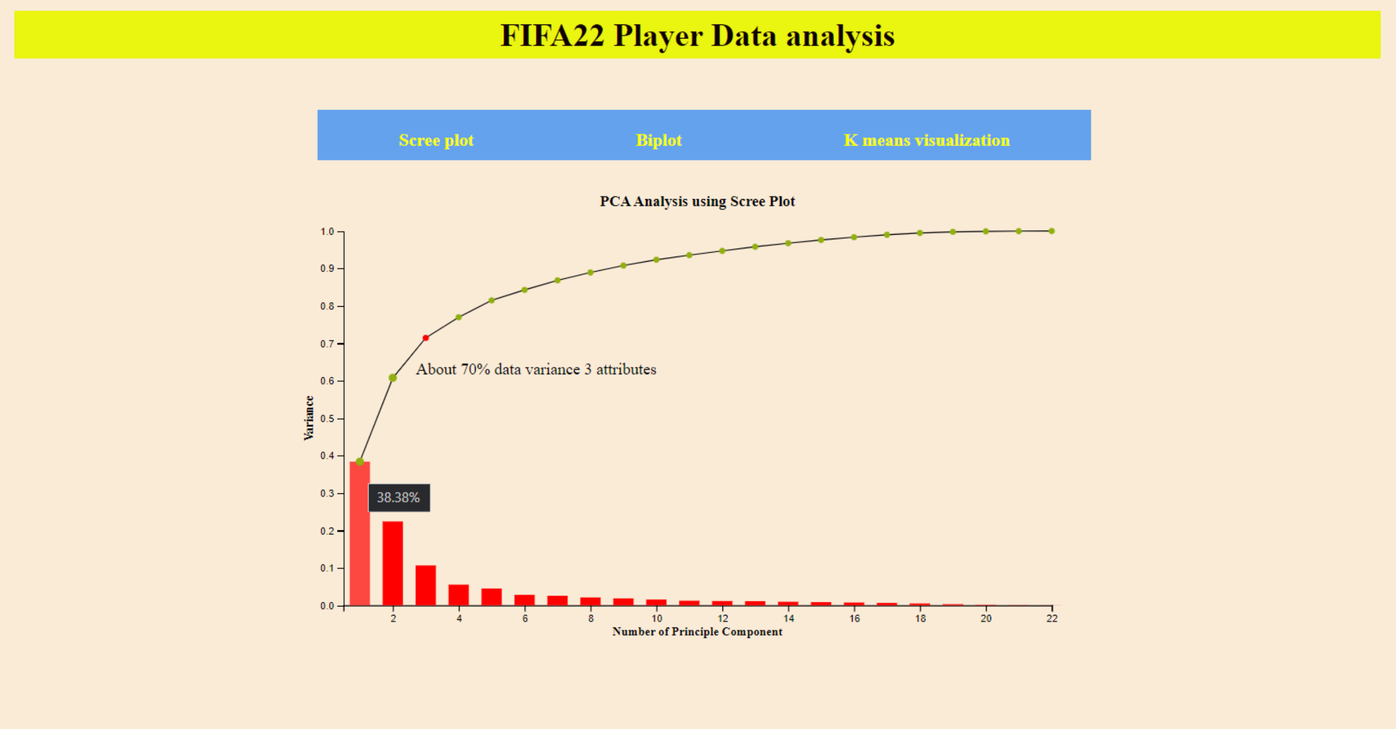
The front end of the application is made using React and d3. Each element on the application has been divided into component, molecules and atoms depending upon their usage to ensure modularity and code reusability. To run the application, navigate into the run the python file: *application.py* from the command prompt/terminal using the following command:

```
!python application.py
```

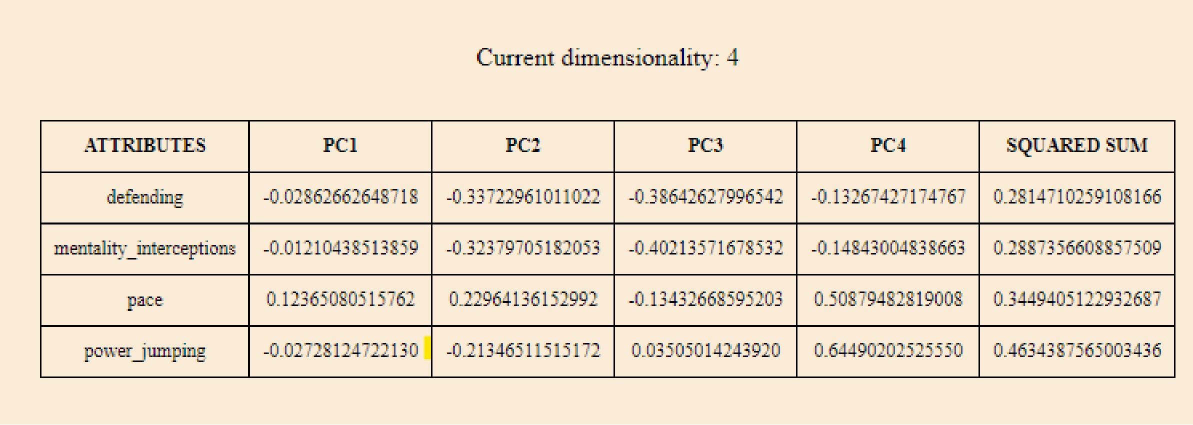
It provides a link: http://127.0.0.1:8000/ where the application is hosted. Upon clicking the link, the user is redirected to the homepage.

The homepage of the application page displays the heading of “FIFA22 Player Data analysis” with all the available functionalities: Scree Plot, Bi-Plot and K-Means Visualization.

A few snapshots of the webpage are attached as follows:

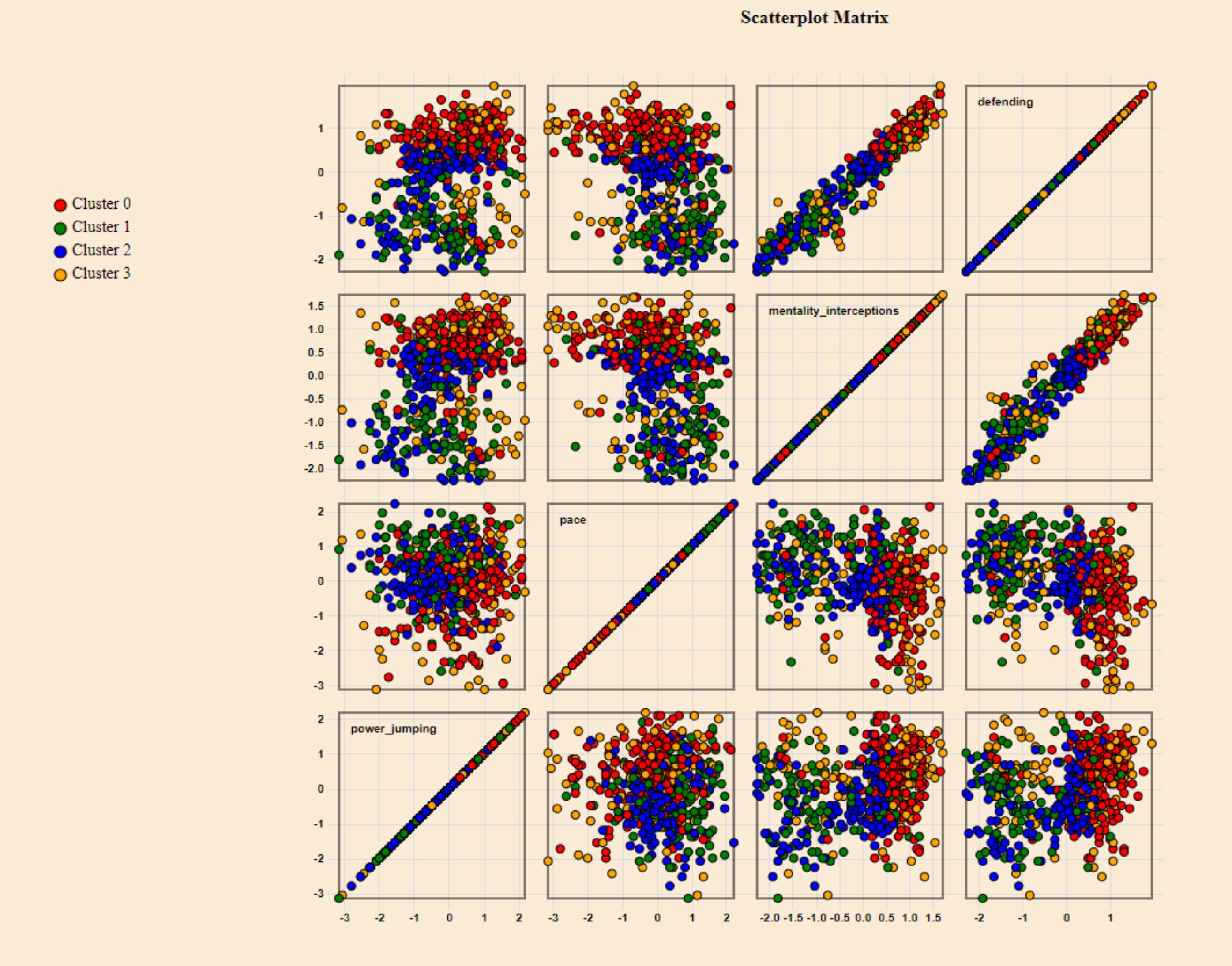

FIFA22 Player Data analysis
Scree plot
Biplot
K means visualization
PCA Analysis using Scree Plot
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Variance
About 70% data variance 3 attributes
38.38%
2
4
6
8
10
12
14
16
18
20
22
Number of Principle Component

If the user selects a scree plot, a scree plot is plotted of the various Principal Components VS the Explainable Variance. A line chart (of screeplot) shows the cumulative variance. The cumulative explainable variance if all the principal components are selected is always equal to 1 or 100%. A user can hover over the bars or the points of the scree plot to see the explainable variance as a pop-up text. Doing so also changes the color of the bar to red to give it a highlight effect.

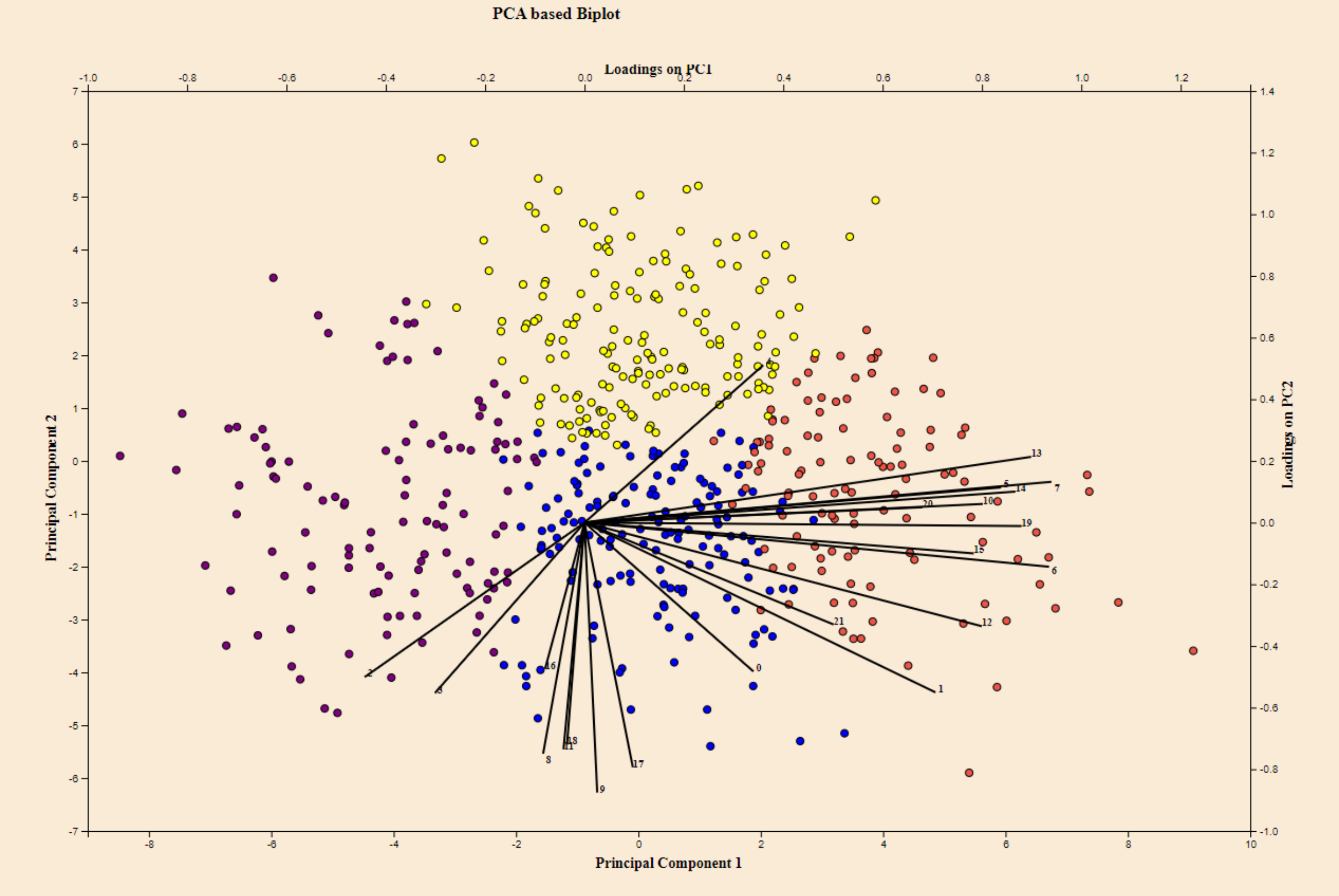


Current dimensionality: 4

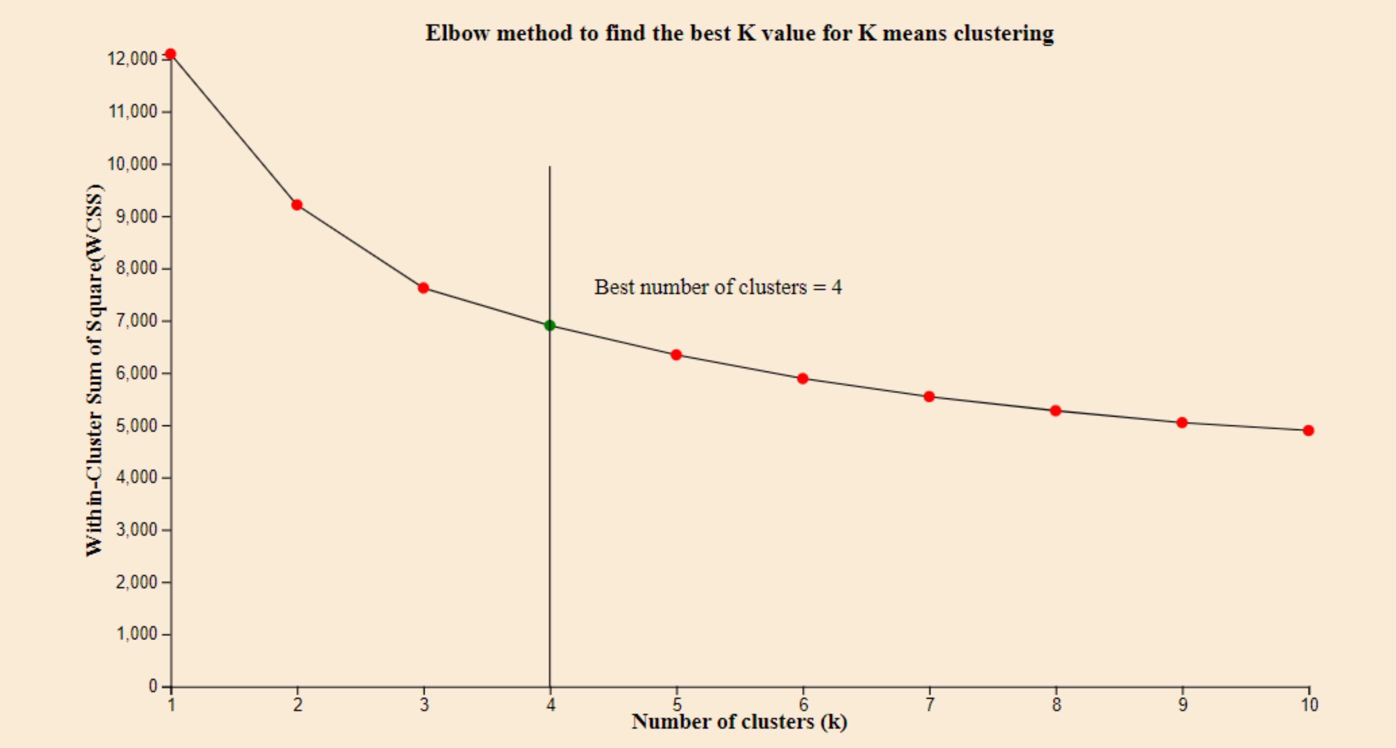
ATTRIBUTES	PC1	PC2	PC3	PC4	SQUARED SUM
defending	-0.02862662648718	-0.33722961011022	-0.38642627996542	-0.13267427174767	0.2814710259108166
mentality_interceptions	-0.01210438513859	-0.32379705182053	-0.40213571678532	-0.14843004838663	0.2887356608857509
pace	0.12365080515762	0.22964136152992	-0.13432668595203	0.50879482819008	0.3449405122932687
power_jumping	-0.02728124722130	-0.21346511515172	0.03505014243920	0.64490202525550	0.4634387565003436



A user can click on any bar of the principal component to select the dimensionality index. Additionally, the dimensionality index is used to select the number of principal components used for calculating the sum of squares of loading. The 4 attributes/features with the maximum sum of square of their PCA components (loadings) are selected to plot the scatterplot matrix.



If the user clicks on the Biplot button of the dashboard, a Biplot is plotted. The plot is between principal component 1 vs principal component 2. The lines on the graph show the various attributes/ features.



Finally, if the user selects K-Means option, it provides a visualization of WCSS vs number of clusters. By analysing this plot, we can understand the best value of k = 4, i.e. number of optimal cluster = 4.