Lab Report – 2B

CSE564: Visualization
Visual analysis of football players from FIFA 2022 database
Hritam Basak

ID: 114783055

Aim: The project aims to teach several important aspects of data visualization. First, we learn to construct MDS data plot using Euclidian distance using scatterplot (with points being clustered) and a MDS variable's plot using 1-|correlation| distance. Next, we aim to visualize the data in a Parallel Coordinates Plot (both numerical and categorical data) with user moveable axis ordering and polylines colored by cluster ID.

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.

w. "league_level": Indicates what is the current level of player's club league, ranging from 1 to 5. A higher level indicates player quality is better.

x. "nationality_rank": The international rank of player's country. Represents if the player is from a very reputed footballing country or not. Ranging from 1-10.

y. "club_contract_valid_until": The year until which the player is retained in his current club. Ranging from 2022 – 2027.

z. "work_rate": Defines if a player is hardworking or not. Ranges from 1-9, 1 indicating low work whereas 9 indicates very hardworking.

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: *app.py* from the command prompt/terminal using the following command:

!python app.py

It provides a link: http://127.0.0.1:5000/ where the application is hosted. Upon clicking the link, the user is redirected to the homepage.
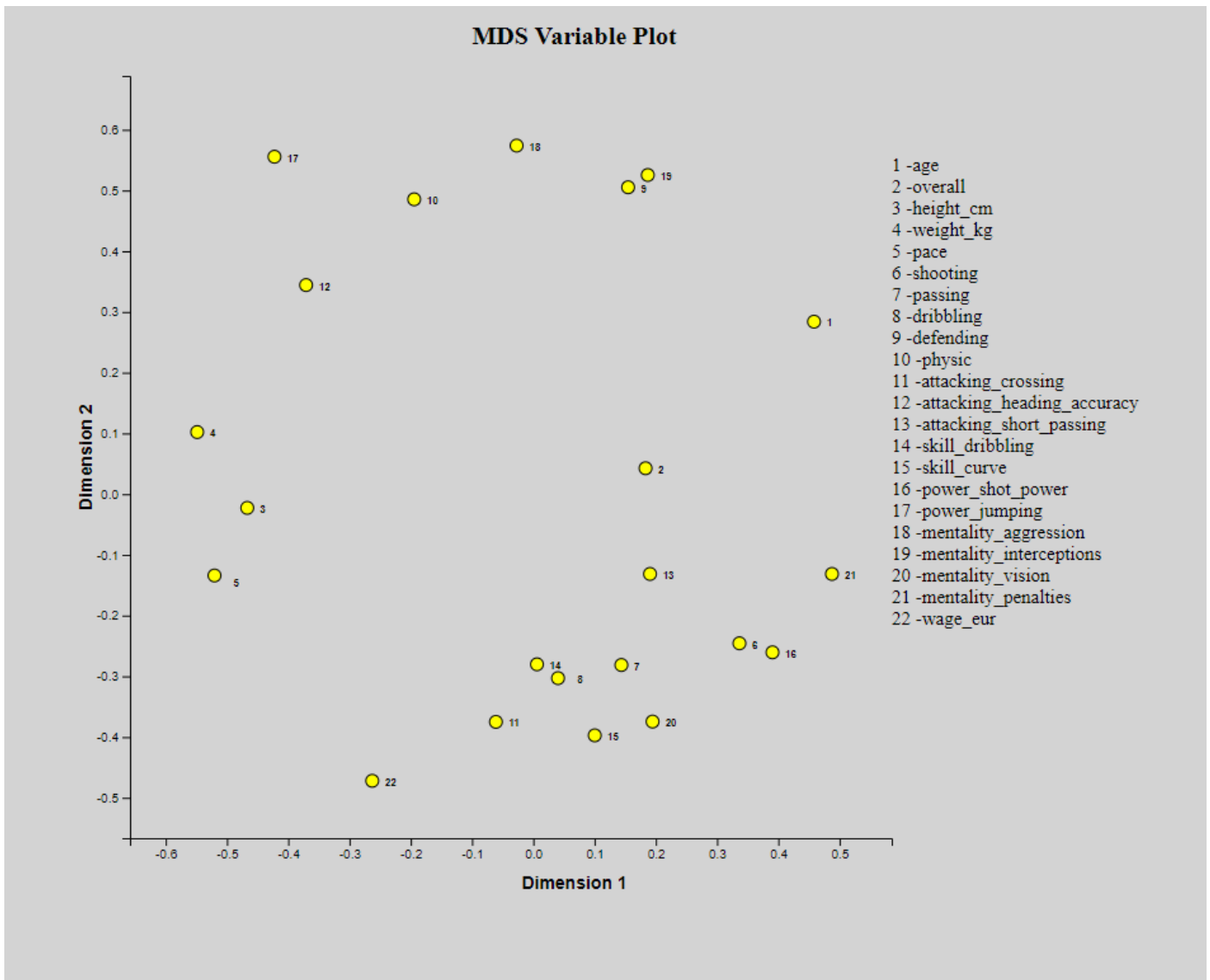
The homepage of the application page displays the heading of "FIFA22 Player Data analysis" with all the available functionalities: MDS Plot, MDS variable Plot and PCP plots.

A few snapshots of the webpage are attached as follows:
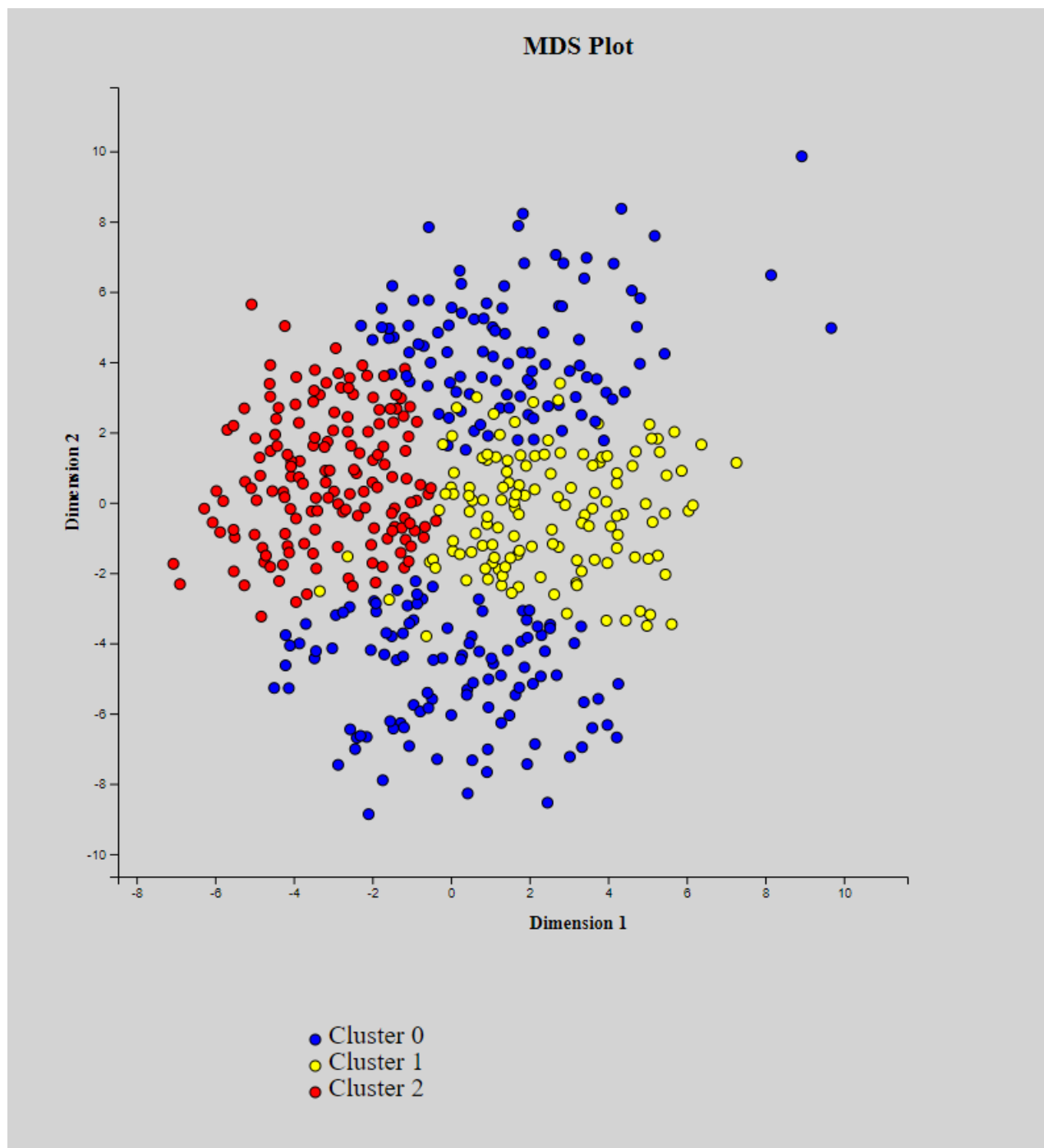
This MDS plot shows the plot of different attributes in the 2D space, which is a compact form of the data distribution in N-dimensional space. I have also shown k-means clustering for the data distribution with three different clusters, highlighted in three colors.

Now, moving to MDS variable plot, a scatterplot is plotted showing the exact same no of points as the features/attributes in the dataset. Each point has the feature name associated with it. This plot also has the feature that if the user selects points on the plot, they turn blue and the order of the selection is then used to plot the PCP for those plots selected.

MDS Variable Plot

1 -age
2 -overall
3 -height_cm
4 -weight_kg
5 -pace
6 -shooting
7 -passing
8 -dribbling
9 -defending
10 -physic
11 -attacking_crossing
12 -attacking_heading_accuracy
13 -attacking_short_passing
14 -skill_dribbling
15 -skill_curve
16 -power_shot_power
17 -power_jumping
18 -mentality_aggression
19 -mentality_interceptions
20 -mentality_vision
21 -mentality_penalties
22 -wage_eur

There is also a checkbox button to activate/deactivate the MDS-PCP mode. If a user plots a MDS Variables plot and selects the points on the plots according to the attributes, the order of the user selection is maintained, and the points turn blue. Now when the user plots a PCP and activates the MDS-PCP mode, a plot for the attributes selected in the MDS Variables plot are used for plotting a parallel coordinates plot.
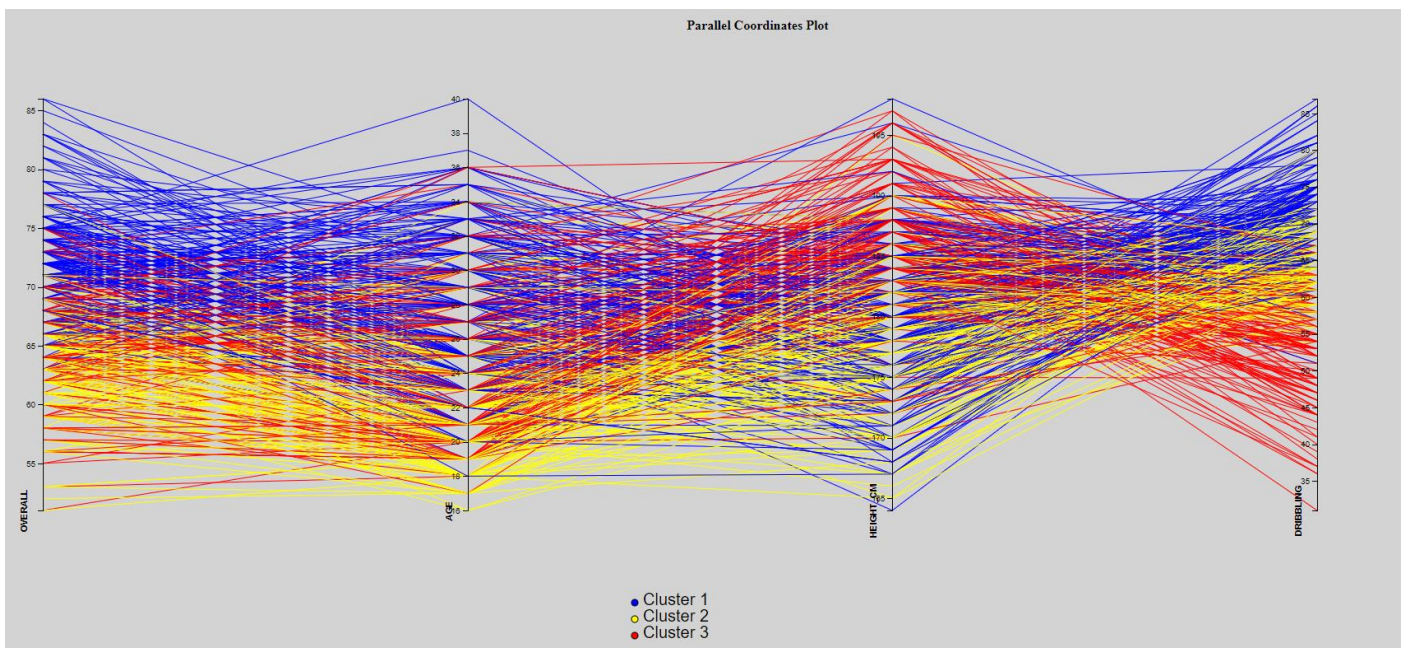
Interestingly, if the user selects points which are close in the MDS-variable plot (i.e. closely correlated), the lines in PCP are almost parallel, else if the points are far apart in MDS-variable space, the lines in PCP are unparallel.
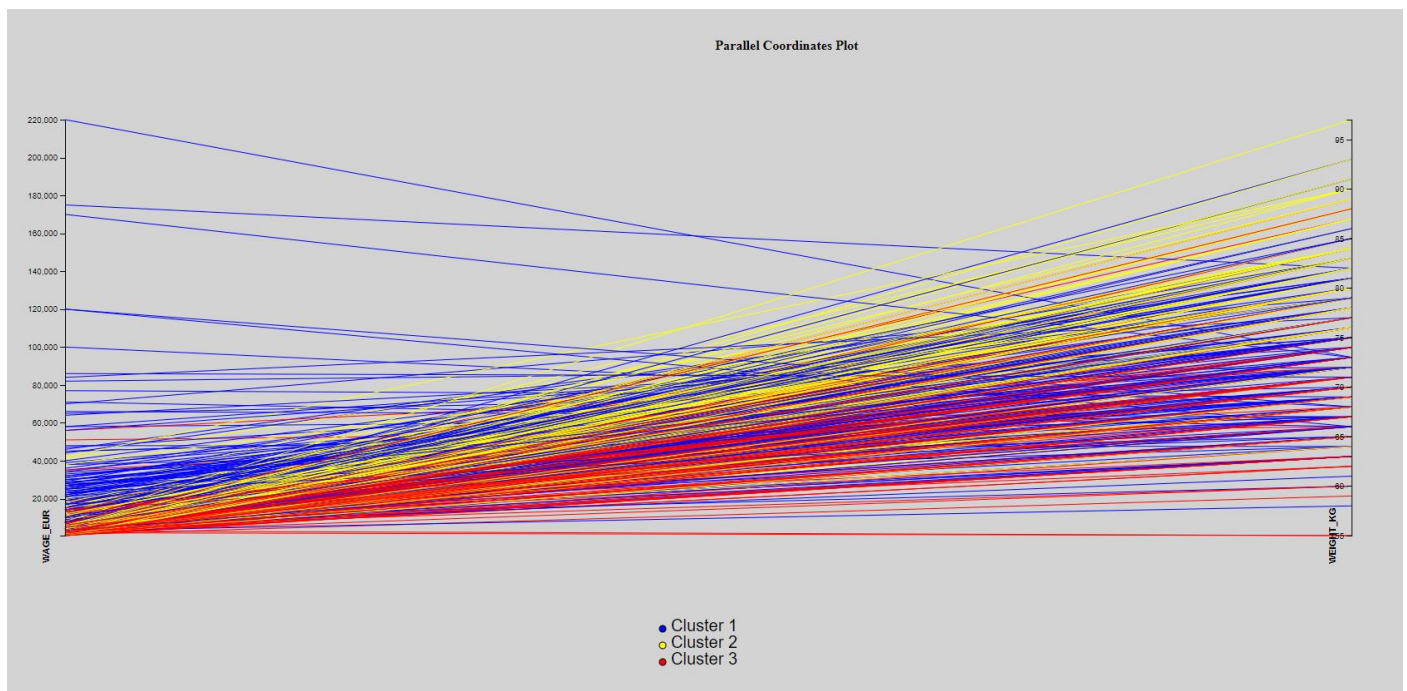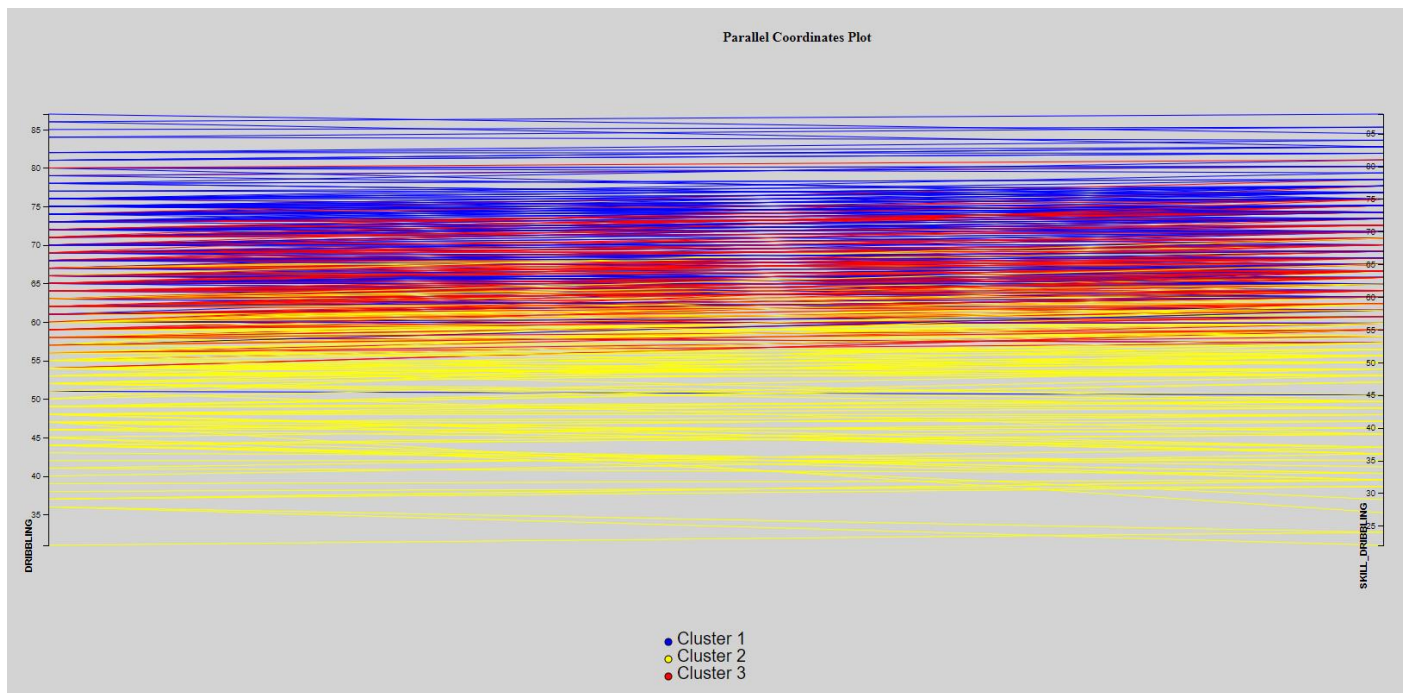
Additionally, the user can select some regions in the axis to filter out some data points in particular.
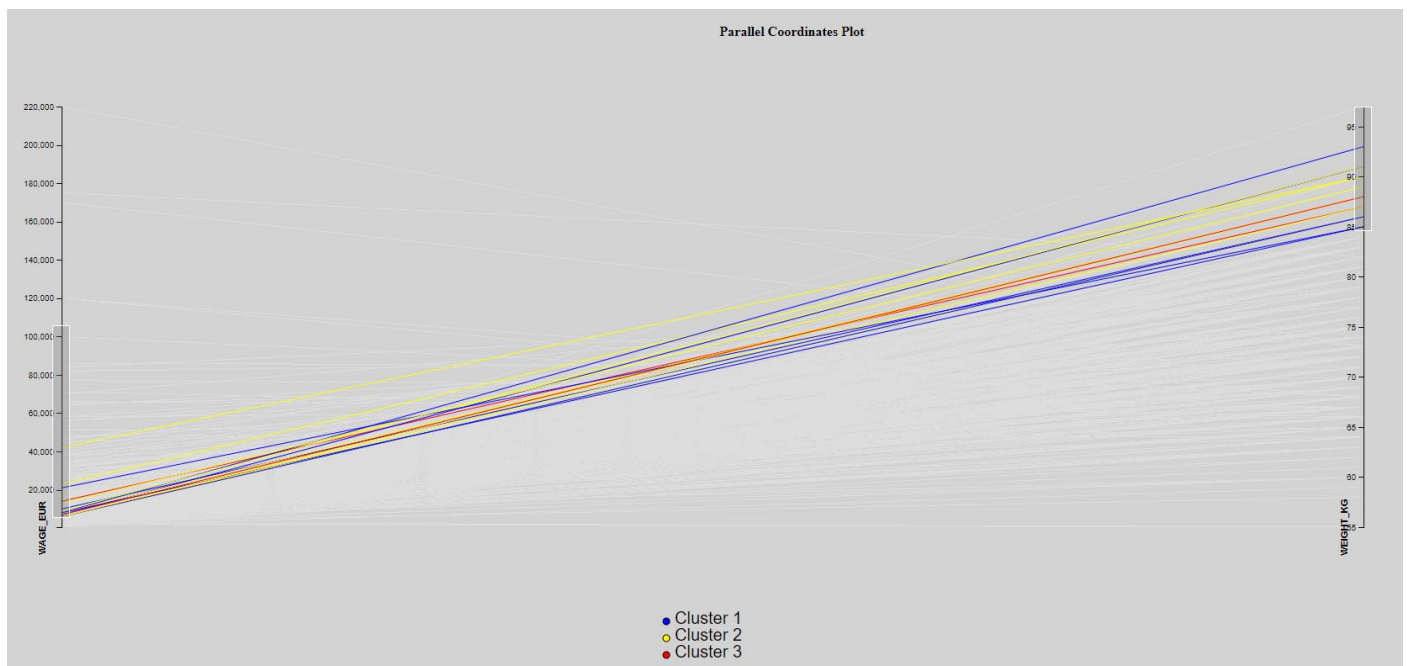
Finally, we plot the PCP plot of all the attributes.
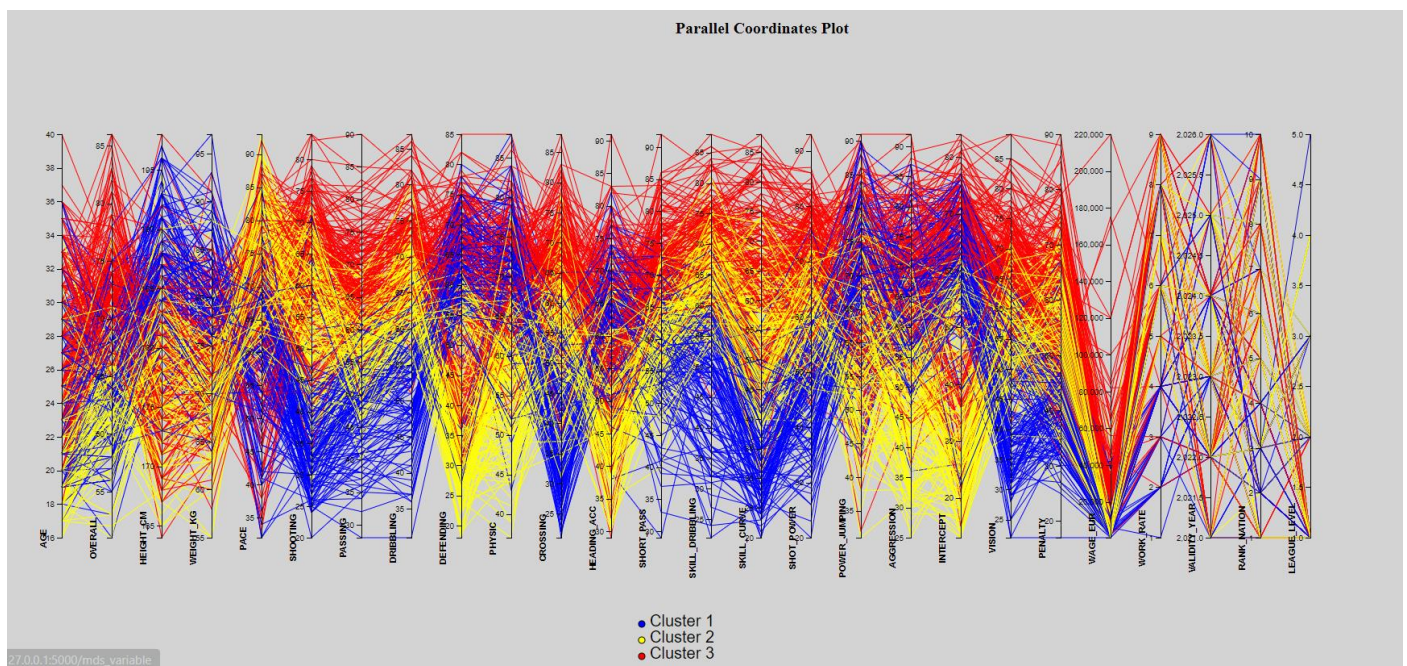
Here also, we have the brushing/filtering functionality:

Parallel Coordinates Plot

- Cluster 1
- Cluster 2
- Cluster 3