VISUALIZATION FINAL PROJECT REPORT
FIFA 2019 CLUB ANALYSIS

SIDDHI MUNDADA (112684006) SAKSHI GUPTA (112552239)

BACKGROUND

The dataset that we chose for our project is the FIFA 2019 Player Dataset. This dataset contains all the relevant information about European Football players that participated in FIFA World Cup 2019. After combining and filtering columns based on relevance to the project, we picked out the final 25 attributes which are:

- Name: Name of each player
- Age: Age of each player
- Nationality: Birth country of each player
- Club: Team that each player plays for
- Value: Net worth of a player
- Wage: Earnings of a player
- International Reputation: Score out of 5 for international standing of each player
- Jersey Number: Jersey Number of each player
- Position: Position in the field that each player plays at
- Preferred Foot: Active foot with which each player prefers to play
- Body Type: Body type of each player
- Joined: Year of debut of each player
- Height: Height of a player
- Weight: Weight of a player
- Crossing, Acceleration, SprintSpeed, Agility, Balance, ShotPower, Jumping, Stamina, Strength, Penalties: Scores out of 100 for each player based on some important playing attributes
- Overall: Average score out of 100 for all the available scores based on a scoring function by FIFA

Apart from this we used another dataset which is called football_project_data_encoded.csv which is basically the dataset after performing Label Encoding. The original dataset is called Football_MR.csv.

DATA CLEANING AND PREPROCESSING

These are some of the steps we took:

- Impute missing and nan values in our dataset with appropriate values
- Convert the datatypes of some columns
- Remove unnecessary symbols and characters from column values
- Rectify the incorrect entries for some column values

IDEA & APPROACH

We drew this idea from the entertainment (sports) domain, as it has been a never- ending discussion (with no proper conclusion) between die hard football fans on which football club seems to be the best. To put an end to this argument, we propose a dashboard to compare different football clubs. We plan to pick 15-20 football clubs and then analyze their player information using various visualizations on various features such as different score values, Overall Rating, International Reputation etc.

Our goal is that the dashboard will help a user to compare clubs which he/she wishes to, and then draw a conclusion on which club seems to be doing well based on the statistics and visualizations displayed on the dashboard. Our dashboard consists of an interactive single screen which does not involve any scrolling. All charts are visible at all times. We are using linked brushing i.e. if a user selects a subset of the data from a chart, then these selections will be reflected in all the other charts. So, a user can the consolidated data of all the clubs or just choose one or more clubs for comparison. We are using a good mix of standard (bar chart, scatter plot, pie chart) and non-standard visualizations (PCA, Parallel Coordinates).

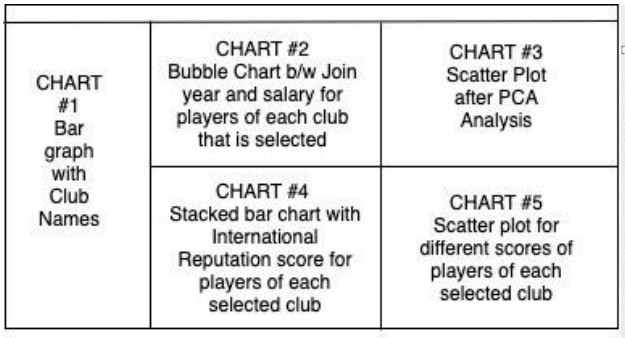


Figure 1: Preliminary Sketch of the Dashboard

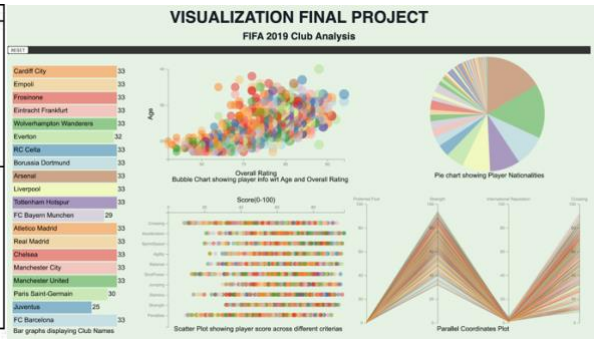
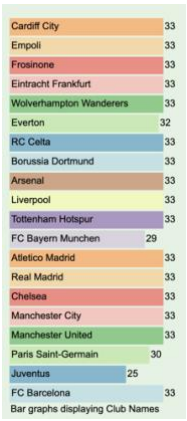


Figure 2: Actual Dashboard

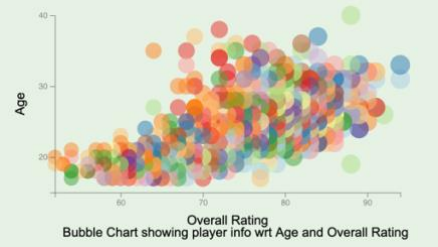
As you can see it is almost close to what we initially proposed with just a few changes in some of the charts as we later thought that other charts would be a better fit. Now, we will briefly explain each chart and the motivation behind using each chart.

CHART #1



In this section, we displayed a **horizontal bar graph** with the club names that we wish to compare. On clicking a particular bar in the graph, you will be able to see the information about the players playing for that particular club in the other four charts.
Motivation: We thought that a horizontal bar graph would take less space and display all the different club names in a concise and neat manner.

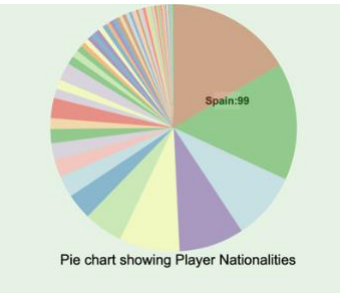
CHART #2



Bubble Chart showing player info wrt Age and Overall Rating

Here, we have plotted a **bubble chart** to represent the players in a particular club. On the X-axis we will have the **Age of players** and on the Y-axis, we will have the **Overall Rating.** We chose the axis values such that we can analyze which club has the youngest/oldest or the most experienced players. On hovering over the bubbles, the user will be able to see player information such as Name, Age, Nationality etc.
Motivation: It seems like a good chart for comparison against two important features and the number of bubbles also gives a good idea of the number of players in a particular club. The values on the axes give us a good idea about the age/experience of players.

CHART #3



Pie chart showing Player Nationalities

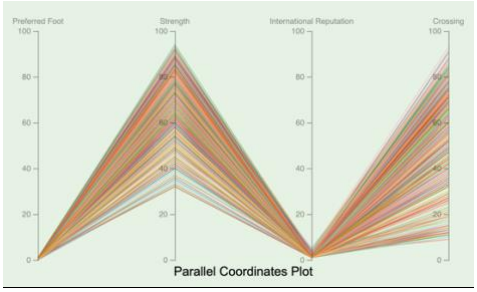
Here, we have a **pie chart** which displays the various **countries** that players come from along with the count of players on hover.
Motivation: A pie chart takes less space on our dashboard and gives us a clear idea of the proportion of players that come from different countries.

CHART #4

Scatter Plot showing player score across different criterias

In this **scatter plot**, we have plotted different scoring attributes such as crossing, dribbling etc on the Y-axis and have a range of 0-100 on the X-axis. Through this graph, the user can see which players of a club have higher skill sets (through their scores) which indirectly tells us about the success of a particular club.
Motivation: A scatterplot gives a clear idea of the distribution of scores on the scale of 0-100 and the way we have used multiple X-axis, it gives us a good tool for immediate comparison of different scores.

CHART #5



Here, we have a **parallel coordinate** plot for comparison between the **top 4 important attributes** we got after performing **PCA Analysis** on the data. The ordering has been done according to the loading values and the scales have been kept uniform for all 4 axes.
Motivation: This plot gives a good idea about the correlation between attributes and since we chose the top 4 important attributes it tells us about the relationship between those 4 columns in a nice manner.

IMPLEMENTATION

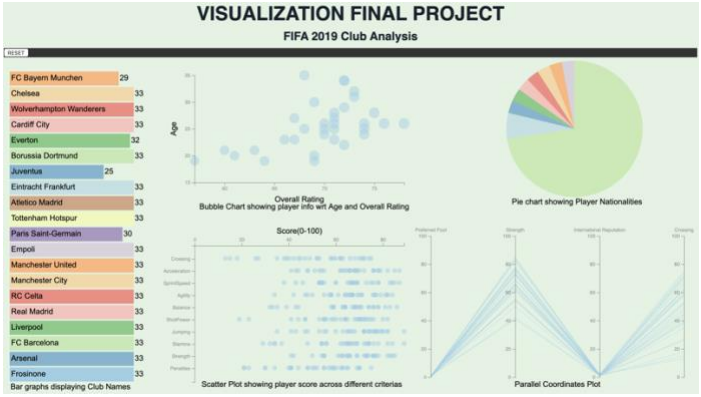
We have used HTML, CSS, JavaScript and python for this project.

1. In the html file (index.html) we have added all the components of the web page i.e. the heading, navigation bar and all the charts. We have divided our web page into 5 parts, one part for each chart. We have also added the JavaScript part in the html page under the script tag.
2. The CSS file(style.css) is used for the formatting and styling of all the components added on the web page.
3. The python file(app.py) is used for the backend processing of the data.
4. We have used Flask as the server
5. Our dashboard deals with two types of data: Consolidated data that has data of all the clubs and Data specific to each club selected through the bar graph. We needed this segregation for implementing brushing and linking and this is implemented by sending keywords through AJAX.

The web page consists of the following charts and functionalities:
CHART 1: Bar chart

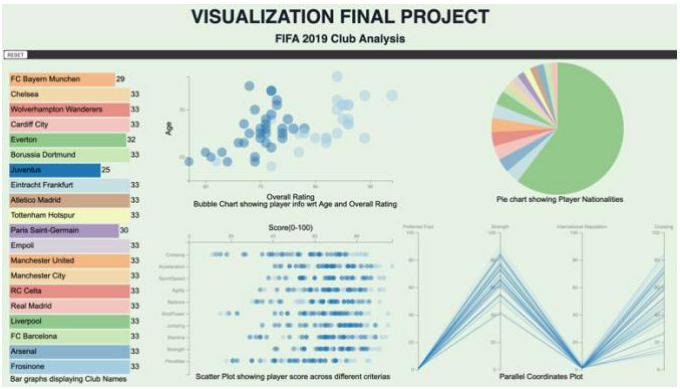
We have made a horizontal bar chart with the club names on the y axis and the frequency on the x axis. The count of each of the bar is displayed in front of it. We have used '**schemePaired'** as the color scheme.
We have added some on click functionality on the bars of the chart.
Functionalities on clicking any bar:

1. The **opacity** of the bar is initially kept 0.5 and on click function the opacity is changed to 1.
2. We have done **linking** in this chart, i.e. linked each of the chart wrt to club name.
   Every time a club is selected the other graphs changes according to it.
   For example, if we select club Frosinone all the other graphs will show the data related to only that club.



Single selection

We have also included the selection of multiple bars. Every time a bar is selected it is added to a string and then added to the URL .In the python part the data all the rows with those clubs will be added to the dataset and the updated data will be sent to all other charts .So if you now select Juventus, we will get the data of both the clubs and the clubs can be distinguished by the color and also we have added hover on each point with club name on it.



Multiple selection

The code snippet for the onclick function:

```
.on("click",function (d) {
    str=str+"$$"+d.A;
    console.log(str)
    d3.select(this).style("opacity", 1.0);
    d3.select("#scatter").select("svg").remove();
    button2("/scatter/"+str);
    d3.select("#bubble").select("svg").remove();
    button3("/bubble/"+str);
    d3.select("#stack").select("svg").remove();
    button4("/stack/"+str);
    d3.select("#parallel").select("svg").remove();
    button5("/parallel/"+str);
    });
```

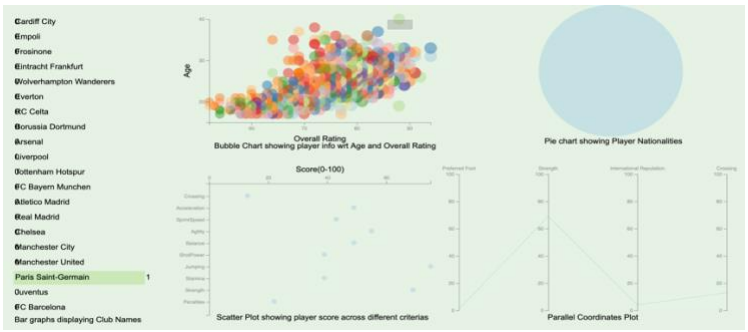
CHART 2: Bubble Chart

We have made a bubble chart with overall rating as the x axis and age as the y axis. Also, the radius of the each of the bubble is according to the overall rating. We have normalized the overall rating as it was quite high for the radius value.

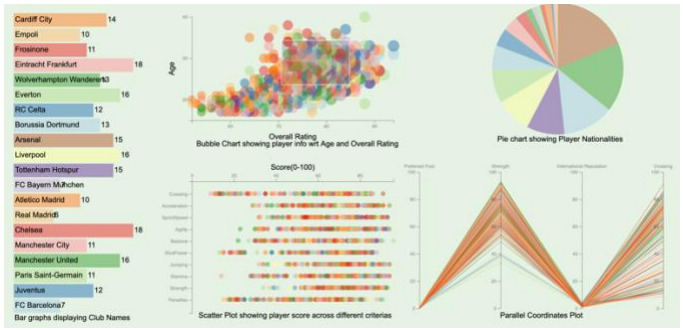
Functionalities:

1. On mouseover on any of the bubble it will show the data of the player like name, nationality, age and value.
2. The most important feature of the bubble chart is **brushing and linking.**
   If you select any number of points from the bubble chart, those changes will be reflected in all the other charts.

For e.g., If you select only one point:



On selecting multiple points:



The code snippet of brushing:
First, we need to get the coordinated of all the points in the window selected.

```
function highlightBrushedCircles() {
        if (d3.event.selection != null) {
            myCircle.attr("class", "non_brushed");
            var brush_coords = d3.brushSelection(this);
            myCircle.filter(function (){
                    var cx = d3.select(this).attr("cx"),
                        cy = d3.select(this).attr("cy");
                    return isBrushed(brush_coords, cx, cy);
                })
                .attr("class", "brushed");
        }
    }

    function displayTable() {
        if ( d3.event.selection) return;
        d3.select(this).call(brush.move, null);
        var d_brushed =  d3.selectAll(".brushed").data();

        if (d_brushed.length > 0) {
            clearTableRows();
            d_brushed.forEach(d_row => populateTableRow(d_row))
        } else {
            clearTableRows();
        }
    }
    var brush = d3.brush()
                  .on("brush", highlightBrushedCircles)
                  .on("end", displayTable);
    svg.append("g")
       .call(brush);

function clearTableRows() {

    hideTableColNames();
    d3.selectAll(".row_data").remove();
}

function isBrushed(brush_coords, cx, cy) {

     var x0 = brush_coords[0][0],
         x1 = brush_coords[1][0],
         y0 = brush_coords[0][1],
         y1 = brush_coords[1][1];

    return x0 <= cx && cx <= x1 && y0 <= cy && cy <= y1;
}
```

After getting the coordinates we need to update the data and send it to the other charts for updating:

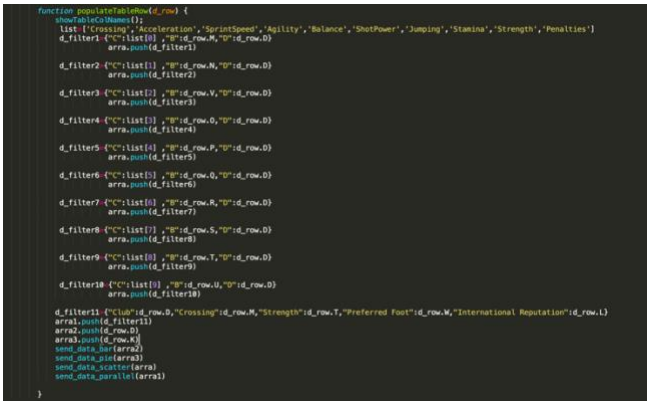
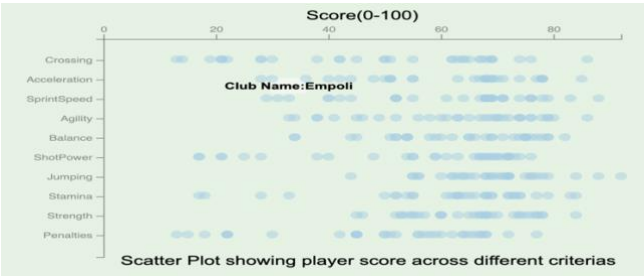


CHART 3: Scatter plot

We have made a scatter plot with scores on x axis and different score attributes like crossing, acceleration etc. on the y axis.

We have also added a mouse over with the club name on every point.
Now, all the attributes will have the same number of points which will represent a player. Every color will represent a different club.
So, you can see the value of crossing, acceleration, agility etc. for every player in this graph. It can also be filtered by using brushing and linking done in the bar and bubble chart. If you select any club from the bar chart you can see the data for the players for only that club.



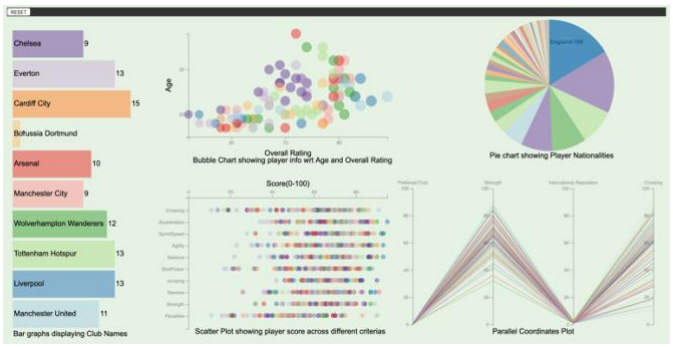
Scatter Plot showing player score across different criterias

CHART 4: Pie Chart

This chart tells us the distribution of countries wrt to the players. For example, in the following you can see that the England has 106 players. We have also added hover to show the name of the country and its counts. You can also see the distribution of any particular club by selecting that club from the bar chart. For example the 2nd image shows that country distribution of players from of arsenal club.

Functionality:**Linking**
On clicking any country from the pie chart ,

1. Opacity of that country will change to 1
2. The other graphs will change according to this as now the data will be filtered in python according to the country data.



In the figure you can see that on selecting England, bar chart shows the distribution of club wrt to England. The total number should be 106. The other plots also shows data of player from England. This is the code snippet of the onclick function:

```
.on("click",function (d) {
        d3.select(this).style("opacity", 0.8);
    d3.select("#bar").select("svg").remove();

            button1("/bar/"+d.data.A);
    d3.select("#scatter").select("svg").remove();

            button2("/scatter/"+d.data.A);
    d3.select("#bubble").select("svg").remove();

            button3("/bubble/"+d.data.A);
    d3.select("#parallel").select("svg").remove();

            button5("/parallel/"+d.data.A);

          })
```

CHART 5: Parallel Coordinate Plot

For this chart we firstly need to find the most relevant attributes from our dataset. We did this by using PCA and finding the top attributes with maximum square loading values.
Here is the code snippet of PCA and square loading matrix:

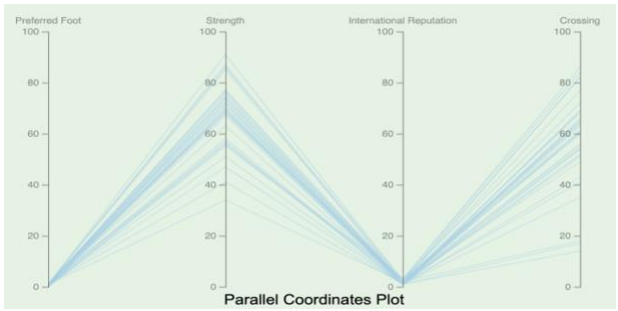
```
data=data[features]# Separating out the target
print (data)
# Standardizing the features
from sklearn import preprocessing

data_scaled = pd.DataFrame(preprocessing.scale(data),columns = data.columns)
# PCA
pca = PCA(n_components=4)
pca.fit_transform(data_scaled)

# Dump components relations with features:
pd.DataFrame(pca.components_,columns=data_scaled.columns,index = ['PC-1','PC-2','PC-3','PC-4'])
```

	Age	Value	International Reputation	Position	Preferred Foot	Body Type	Crossing	Acceleration	SprintSpeed	Agility	Balance	ShotPower	Jumping	Stamina	Stren
PC-1	0.076717	0.178618	0.191487	0.145709	-0.036119	-0.014228	0.343918	0.331021	0.320034	0.332430	0.285370	0.334426	0.100113	0.325682	0.038
PC-2	0.424964	0.222884	0.359436	-0.065504	0.083325	0.213421	-0.072010	-0.207799	-0.133260	-0.198971	-0.257813	0.026403	0.298701	0.035641	0.425
PC-3	0.177706	0.340803	0.386383	-0.291123	-0.080083	-0.393937	-0.018593	-0.062372	-0.194096	0.138432	0.156292	-0.174800	-0.248248	-0.179479	-0.443
PC-4	-0.039463	0.151004	0.045900	0.455183	0.825207	-0.203519	-0.141895	0.018854	0.011998	0.001896	-0.046123	-0.046011	-0.041315	-0.076362	-0.072

After doing that we got Preferred Foot, Crossing, International reputation and strength as the four attributes and plotted the parallel chart. We have added a hover to show which line corresponds to which country. Also, it is linked to the charts and it will change if any club is selected. On selecting Arsenal we get the following chart.

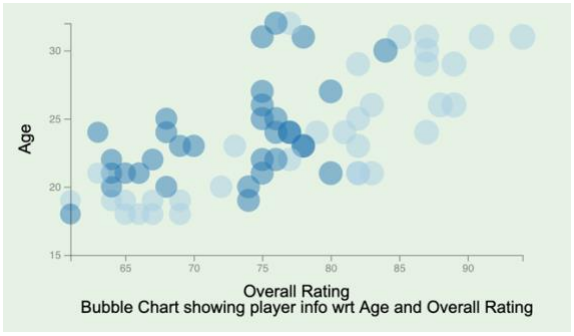


Parallel Coordinates Plot

In addition, we have also provided a reset button for the user to go back to the initial setting of consolidated data.

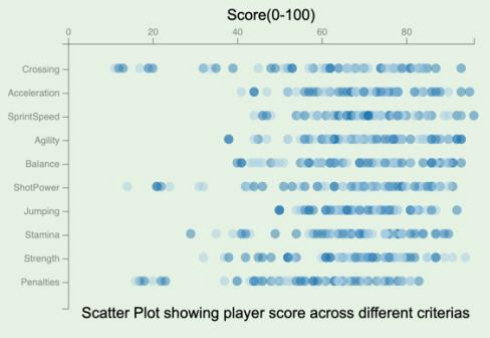
OBSERVATION & FINDINGS

1. From the horizontal bar graph, we found out roughly how many players play for each club.
2. From the image below you can see that the bubble chart gives us a clear idea that RC Celta has more players with lesser overall rating than FC Barcelona which is somewhat correct as FC Barcelona is a more successful team.

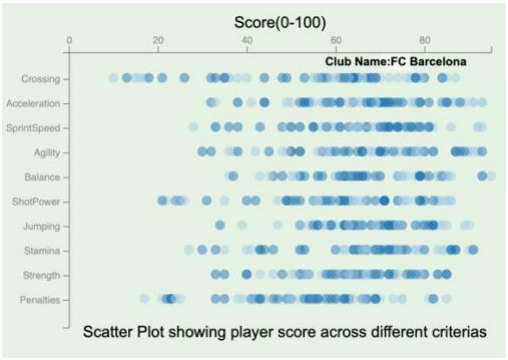


Bubble Chart showing player info wrt Age and Overall Rating

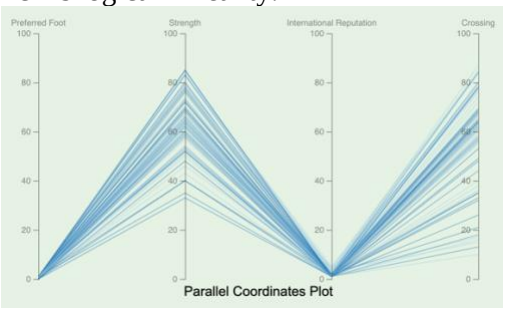
3. Also, from the same plot we could see ‘Lionel Messi’ and ‘Christiano Ronaldo’ as outliers as their overall ratings are much higher than other players which is true as they are well renowned players in their 30s.
4. In addition, while plotting this graph we experimented with different combinations of axes and on plotting Year of joining vs Value of a player we found an interesting outlier who had a very high salary, but the year of joining was 2018. On further inspection we found out that the outlier was ‘Ronaldo’ as he recently joined a new team ‘Juventus’ in 2018.
5. On observing the pie chart, we found out that a high majority of players come from European countries and very few players come from countries like United States, Canada and Mexico. This tells us about the craze and popularity of football in European countries in comparison to countries such as USA or Canada.
6. From the scatterplot of scores (Chart 4), we found out that two very famous rival teams- Manchester City and Manchester United are almost at par based on the score criteria’s which can be seen from the plot below.



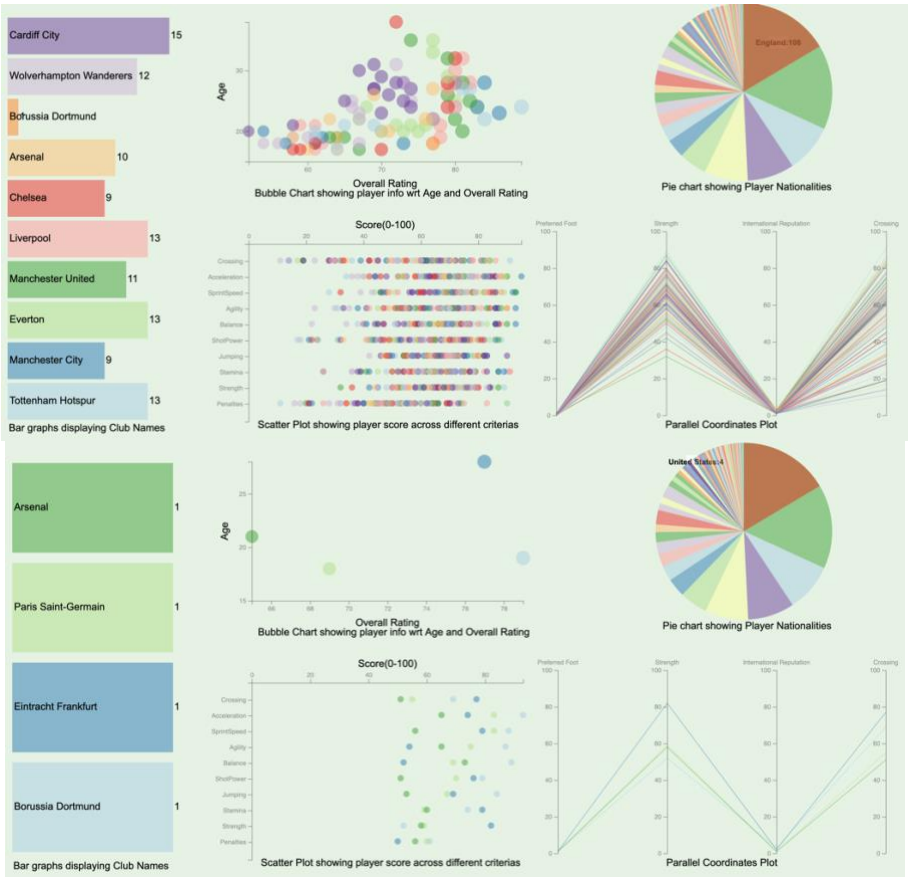
7. On the other hand, from the same scatterplot we could see that successful teams such as FC Barcelona had more players on the higher score side (right side, light blue circles) than comparatively less successful teams such as RC Celta.



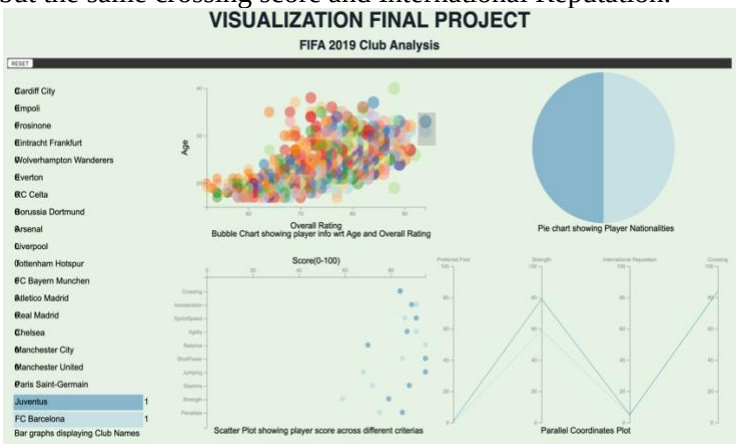
8. From the plot below (PC Plot), we could easily see the relationship between attributes such as Strength and International Reputation- A player with more strength score should have a higher International Reputation who in turn should have a high crossing score. All the attributes seem to have positive correlation which is logical in reality.



9. From the pie chart, we could also compare the contributions and player statistics of different countries such as England and USA.



10. On using brushing in the bubble plot and selecting the two outliers- ‘Lionel Messi’ and ‘Christiano Ronaldo’, we can make comparisons between the two players based on the different charts. From the pie chart we can see that Ronaldo comes from Argentina and Messi comes from Portugal and from the scatterplot of scores we can see that both have a mix of scores but they are mostly on the higher(right) side. From the PC Plot we can see that Ronaldo has a higher strength score than Messi but the same crossing score and International Reputation.



CONCLUSION

Our **interactive dashboard** helps any user to experiment with the data of different teams by seeing and analyzing different type of data on their screen. **Brushing and Linking** along with **Multiple Selections** helps to visualize the clubs together and separately which makes it easier to draw concrete conclusions and decide for themselves which club wins over all the others. For our own analysis we found **Real Madrid and FC Barcelona** to be the teams at par based on features such as player income, skill scores, preferred foot etc.

YOUTUBE LINK: https://youtu.be/QID7s6UE2ls