**Data Analysis and Visualisation of European Soccer Dataset**

Abstract (max 200 words)

European Soccer has been one of the most followed sports in the world, which hosts a lot of leagues and has been home for many great players the world as ever seen, analysing such a rich dataset to find some insight and visualize the results

For this assignment, I am exploring the European Soccer dataset from Kaggle. I am trying to answer the following questions.

* Comparison of different attributes of Lionel Messi and Cristiano Ronaldo in order to conclude who is superior?
* Finding out which leagues are most popular in Europe, top teams in each league, top players and their best attributes?

To visualize these questions, I have used an interactive radar plot to show attributes of each player and a hover function to get most of the information from the graph, and to visualize question 2, I have used a Sunburst graph to create a parent child relationship between different attributes so the user can drill down into data to inspect more. I have used Python and Jupyter Notebook for the purpose of data cleaning and visualizations.

Messi excels in few attributes but Ronaldo has a slightly more overall rating, which makes him win the fight on paper

**1. Dataset [½ page]**

Dataset used for this assignment is European Soccer Dataset from kaggle[1], format of the dataset is an SQLite Database which consists of 7 tables, size of the whole data set is 300 MB, few of the tables are as follows:

Dataset 1: Player

* Number of rows: 11100
* Number of Columns: 7

Dataset 2: Player\_Attributes

* Number of rows: 184000
* Number of Columns: 42

Dataset 3: Match

* Number of rows: 26000
* Number of Columns: 115

**Player table** consists of various attributes related to each player and a unique ID which is linked to other tables. Few of the attributes are player\_api\_id, player\_name, birthday, height, weight etc.

**Player\_Attributes** table consists about 42 attributes (Potential, Crossing, Short pass etc.) which rates each player to those attributes.

**Team and Team\_Attributes** tables consists of 2000 rows and 30 columns combined which gives information about each team and their mentality during the game. Few of the attributes are defencePressure, PlaySpeed, PlayPositioning, DefenderLine etc.

**Match** table is the main table where all the data is linked to, it has 26000 rows and 155 columns (match\_api\_id, home\_team\_api\_id, away\_team\_api\_id, home\_team\_goal, away\_team\_goal etc.) and describes about each player involved in that particular game and his respective country, team, season and league he belongs to and linked with ID to each field.

**League and Country** tables consists of various league played in Europe and all the teams taking part in those leagues. Attributes are id, country\_id - to link country and league table and name

In aspects of big data, European soccer dataset contains volume and variety.

**2. Data Exploration, Processing, Cleaning and/or Integration [½ page]**

In order to produce visualization with respect to the questions, I had to read the data from the SQLite file and create multiple data frames to store the data. In order to answer my first question I had to merge multiple data sources (**Player and Player\_Attributes, League and match** ) to generate “**MergedPlayer.csv**” and “**data.csv**” files. In order to do that I had to perform a full outer join on “**player\_api\_id**” and “**country\_id**” attribute between datasets respectively and once the dataset was merged I had to clean the dataset and fill out the missing values as well as drop few columns which were not needed for my analysis.

As there were numerous occurrence for each Player in different matches, I had to sort the name field and club all the attributes of each individual player using their mean value and recalculate the attribute values, further more I had to drop more columns as we are comparing skills of players hence their height, weight, BMI, age and few more attributes which are not related to player skills were filtered.

For the second question, **country and league data** had to be merged on “**id**” field using full outer join as well as **league and match data** had to be merged on “**country\_id**” to create a bigger table containing all the details of matches played between teams, players and there attributes, leagues in which each team participated and a parent child relationship had to be developed from this values.

Once the data was merged and cleaned, I had achieved a subset from the original dataset containing attributes that had to be visualized

What did you need to do to prepare the dataset(s) to create your graph/chart?

How did you choose the attributes to visualise?

**3. Visualisation** [½-1 page]

Screenshot or image of visualisation

Explain your choice of chart or graph type - what relationship or data type are you showing?

Design choices - justify your use of colour, shapes, marks, layout, structure, font, labels

Any interactivity or animation and how it helps answer your question

List of tools or libraries used

4. Conclusion [½ page]

Critically analyse the outcome of your visualisation.

Were there aspects that you think could be improved upon?

Were there effects or functionality that you were technically unable to achieve?

References

Include any citation of the dataset

Include links to any tutorial or example that contributed significantly to your work

Include any articles or web resources supporting your design choices