ANALYSIS

European Soccer Data Analysis

The study attempt to analyze the European Soccer dataset, which has more than 25,000 matches and above 10,000 players for European professional soccer seasons from 2008 to 2016. In addition, the study focused on the analytical process includes the steps for exploring and cleaning our dataset, for predicting individual and team performances using python packages..

To accomplish the objective of analysis, two questions were posed and Matplotlib is used to visualize the answers. First question pointed to finding the team which scored most goals over the time period. In this venture to investigate a solution for the question, we merged two table data which gives information about the teams and goals scored into one single table and then sorted the resultant table based on the goals scored in the descending order. Besides, created a pie chart using matplotlib which illustrates the top ten teams. Second question remarks to find out the relation between ball control and free kick accuracy. Also we have to identify the most preferred foot of players related with free kick accuracy. To solve this second part we created a graph that displays each instance of free kick accuracy percentage and the number of players who prefer right foot and left foot, distinctly.

And the results obtained from the study was mainly compared with the displayed graphs. And according to the comparison the result and conclusion were made. In this we analyzed a large dataset of European soccer and answered the related two questions. For completing the procedures, we used packages like pandas, numpy and matplotlib for visualizing the data.

Data wrangling phase

The primary purpose of data wrangling can be described as getting data in coherent shape. In other words, it is making raw data usable.

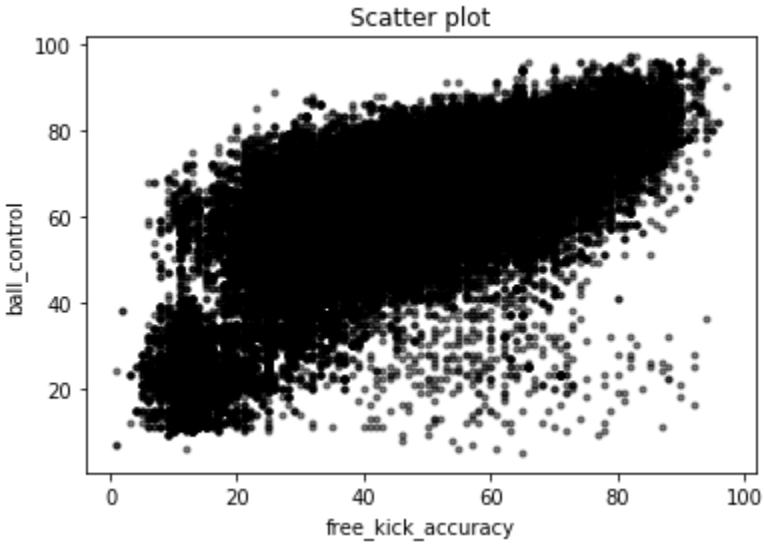
Here, we got a sql database with several related dataset. But the information's were distributed across different datasets. So we select required columns from the table using sql queries and made new pandas dataframe . data merging also done for getting relevant data together. Another

import step in data wrangling phase is checking cleanliness of dataset. In our case we found some null values, duplicated values and dropped the values .

1. Team which scored most goals over the time period

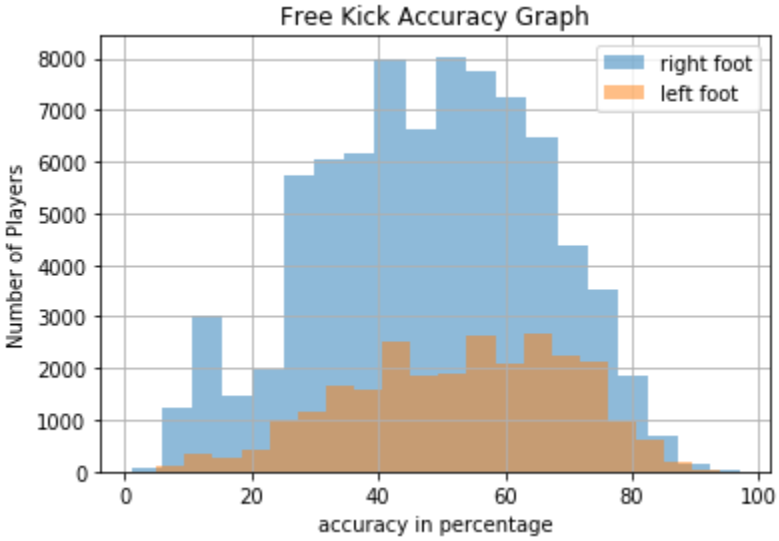
From the pie-chart, we found that the FC Barcelona won more games over the time period and it also gives information about the percentage of goals scored by the top ten teams

2. Relation between the ball control & free kick accuracy



Here, we concluded a positive correlation of free kick accuracy and ball control of players and clearly plotted using a scatter plot.

and most preferred foot of players related to free kick accuracy



In the second part of this question we created a graph that displays, in every instance of free kick accuracy percentage the number of players who prefer right foot is very much higher than players who prefer left foot.

And thus the result obtained from the dataset analysis has details of more than 25,000 matches and 10,000 plus players. Since, it has some null values and duplicate values in the dataset and the dataset contains much irrelevant data and misses relevant data like number of goals and number of wins etc. So the most time consuming phase in this analysis is Data Wrangling, because the whole data is spread across several tables and we have to merge data from different tables for particular uses.