TOPIC 5

ANONYMOUS DATASET VISUALIZATION AND INSIGHTS:

- I made project on Google Collab. So, I started with importing NumPy, Pandas and Matplotlib libraries. NumPy because it is the basic foundation of pandas and matplotlib, while pandas due to its multipurpose usage on series, and datasets. Matplotlib as it is one of the best libraries there to make graphs, to do mathematical application.

- Then, I imported excel sheet in Google Collab, this was achieved by uploading it first on my Google Drive.

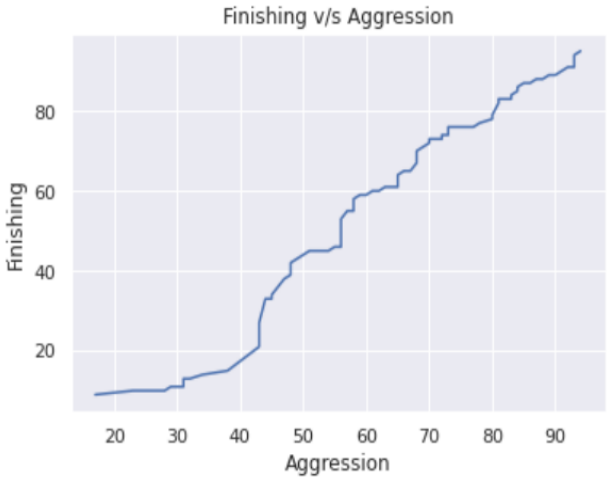
- Now I had Excel sheet in my dataset, I used few functions on it like describe and head. Describe helps us to get overview about our data.

- After this I made a list only of players which are not goalkeepers, a list of top 50 players.

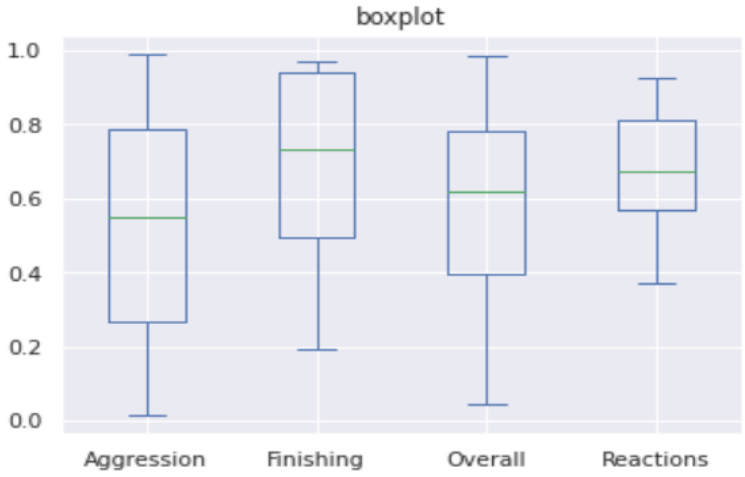
- Now I started analysing data, and plotted graphs based on them.

- Above graph shows us how important reaction time is in football, and it affects players overall performance by significant amount.

- Another graph is plotted between attributes of player's such as Finishing and Aggression clearly showing linear relationship between them which means if a player is said to be more aggressive than other, then we can mostly say that he is good finisher also.



- Now I plotted Boxplot of Aggression, Finishing, Overall and Reactions. Boxplot helps us in visualising data distributions such as we can compare two sets of data on grounds of variance, percentile, mean and outliers in them.



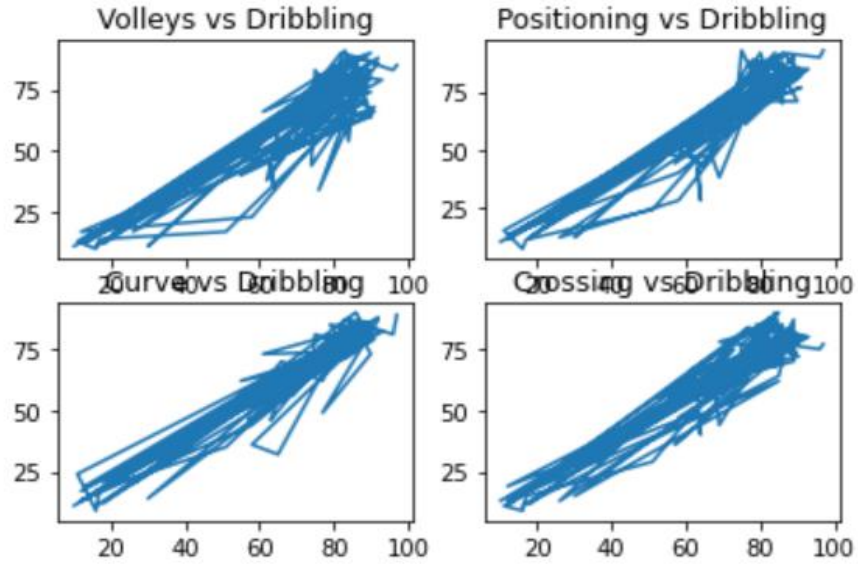
- For outliers' removal, we can find 25th percentile(q1) and 75th percentile(q2) of a data and difference between them is called as IQR (Inter Quartile Range), and we can say that data lying between range of (q1-1.5*iqr, q3+1.5*iqr) are not outliers and other we can remove. Removing outliers helps us in having more accurate data distributions.

CORELATION:

- After doing this, I studied about Corelation & Covariance. Covariance is used to determine to how much extent 2 random variables are connected while Corelation is a statistical measure which lies between [-1,1] and tells us about how strongly two random variables are related. It has a value of 0 if on plotting multiple values of those two random variables we get a parallel line to x-axis, it gives negative value if those 2 random values are inversely related while positive if both values are linearly related. The more the value of Corelation, more we can say that those two variables are strongly connected.

- If we have two strongly related variables we can remove them, from our dataset as we can use our other data to find removed one.

- I calculated Corelation table for our dataset and after some analysis, I came to know about Dribbling is having strong relationship with many other attributes of player such as Long Passing, Short Passing, Curve, Crossing, Penalties, Ball Control and Free kick accuracy. I plotted a graph also to verify these conclusions.



- We can conclude from above graphs that our conclusion from Corelation is correct.

LDA REDUCTION TECHNIQUE:

- Linear Discriminant Analysis is used to focus on maximising the seperability among known categories.

- Our main focus in LDA is to maximise the distance between means and minimise the variation(scatter).

- For high numbers of data we choose two axes to represent them on those 2 axes, whereas if we do not do that then let's say we have 10,000 different data and then, we will need 10,000 axes to represent those data's, but with help of LDA technique we can achieve the same in 2 axes only.

CONCLUSION:

- We can conclude that Data Analysis is a very important aspect of almost every field such as in this particular example also it helps us in finding a player's

weakness, strengths and tells which attribute is more important. Also it tells us about which data is more strongly related with other.

TOOLS USED:

- Google Collab, Excel, Pandas Library, Matplotlib Library

ACKNOWLEDGEMENTS:

- FreeCodeCamp online tutorial on Data Science.
- StatQuest is also a Youtube channel which helped a lot.
- GeeksForGeeks, StackOverflow, YourDataTeacher and TowardsDataScience.com were used in case of doubts.

Google Collab Link:
https://colab.research.google.com/drive/1Zz5DNvTP0u8tlnAK0KOOR8XyD3UrCynI?usp=sharing