Declaration on Plagiarism

Name:	Naveen Garaga Krishnamurthy
Student Number:	19210263
Programme:	MSc. In Computing (Data Analytics)
Module Code:	CA682
Assignment Title:	Data Visualisation
Submission Date:	15 Dec 2019
Module Coordinator:	Dr Suzanne Little

I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. I have read and understood the Assignment Regulations. I have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged, and the source cited are identified in the assignment references. This assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study.

I have read and understood the referencing guidelines found at http://www.dcu.ie/info/regulations/plagiarism.shtml, https://www4.dcu.ie/students/az/plagiarism and/or recommended in the assignment guidelines

Name: Naveen Garaga Krishnamurthy Date: 15-12-2019

FIFA 19 Dataset Analysis

Abstract
The story that is visualized here is the analysis of the FIFA19 game data set containing information of various clubs, countries and football players. FIFA 19 is a game developed by Electronic Arts as part of their FIFA series and is officially licenced and endorsed by the International Federation of Association (FIFA). The conclusion that is derived from the analysis and visualization is that footballers exhibit prime form and work rate between the age 23-29 and are highest valued then. The Striker (ST) position in football is the most sought after and Spain seems to house most of the football worlds MVPs.

Dataset
The data set was taken from data set repository www.Kaggle.com, the name of the dataset is "FIFA 19 complete player dataset". The data for this was scraped from "https://sofifa.com/players" and put in a single csv file by the contributor. The size of the data is about 9 MB with 18207 rows and 88 columns. Data types present is a combination of numerical data, links, text, date, text and numeric data combined and NULL values. Looking at the data the variety aspect of big data is considered present in the data set as the number of columns present in the dataset is high and there are different data types present as well.

Data Exploration, Processing, Cleaning
The dataset was in csv file format, in order to prepare the dataset as per my visualization needs, python code was written to filter out the required data. Python provides libraries which aid in the complete process of cleaning, processing and visualizing raw data to depict a meaningful story to the end user. Exploratory Data analysis technique was applied first in order to find the shape of the dataset, next null values were removed as they might tamper with the charts visualized. Values were also imputed in some columns where the number of missing values were less rather than dropping the entire row and loosing data because of a single column. Clubbing technique was used in order to generate new data fields on which the charts were graphed. Pythons pandas library is used to generate data frames on which data manipulation is done such that it is easy and effective to plot graphs using the seaborn and plotly libraries of python. Data analysis done helped to plot graphs based on 'single variate', 'bivariate', 'multivariate' concepts of exploratory data analysis. Depending on this suitable column from the dataset were picked like the 'Position' attribute which is used to plot a single variate graph which describes the total number of the players playing in various positions. Age of a football player is an important aspect for a club manager as it is seen that the work rate and form dips as the player ages hence the 'age' and 'work rate' column is chosen. To show which country boast the most valuable player, the 'value' of the player in the market and his 'ratings' as per FIFA 19 game is plotted.

Visualisation
Figure 1 indicates a single variate graph of the exploratory analysis done on the dataset. The position field plotted on the x-axis is a categorical value which shows the different positions that a player can play at. The choice of the graph is a bar chart which shows the number of players playing in the positions shown the x-axis. The background colour was chosen as light blue to keep it in contrast with the dark colours applied for the bars of the bar chart.

Figure 2 indicates the bivariate relationship between a quantitative numerical value age and the corresponding work rate (categorical value) for it. The chart type is a box-plot graph which gives the information that players between the age of 23-29 have good work rate both halves of the game. A dark background with a light-coloured boxplot is used to make the quadrants and the outliers more visible and catchier.
Figure 3 is a multivariate relationship graph between the value and the ratings of the players of the top 3 countries in the dataset. It is a graph that maps numerical (value and ratings) and categorical (country) values. The chart shows that majority of players from England dwell in the lower value and ratings category. Figure 1 and 3 were plotted using the seaborn module of python and figure 2 used the graph object module of the plotly package of python. The charts are chosen in order to represent the 3 kinds of relationship that can exist between the attributes of a dataset when doing exploratory data analysis. This plays a vital role in providing clarity on what kind of data are we looking at and how models can be derived from it. Python libraries like pandas, seaborn, Markdown, matplotlib pyplot, plotly graph objects libraries are used.

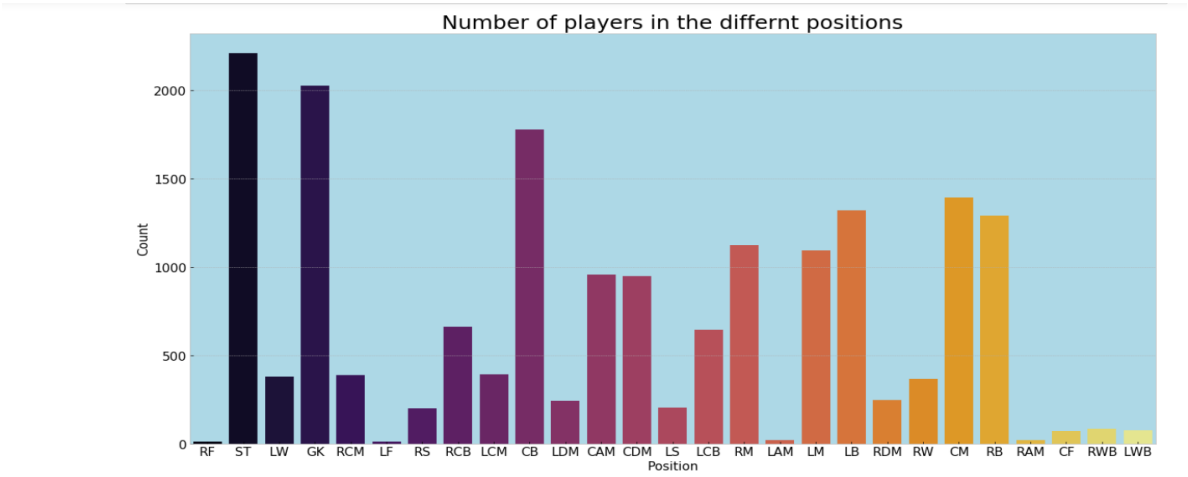

Figure 1

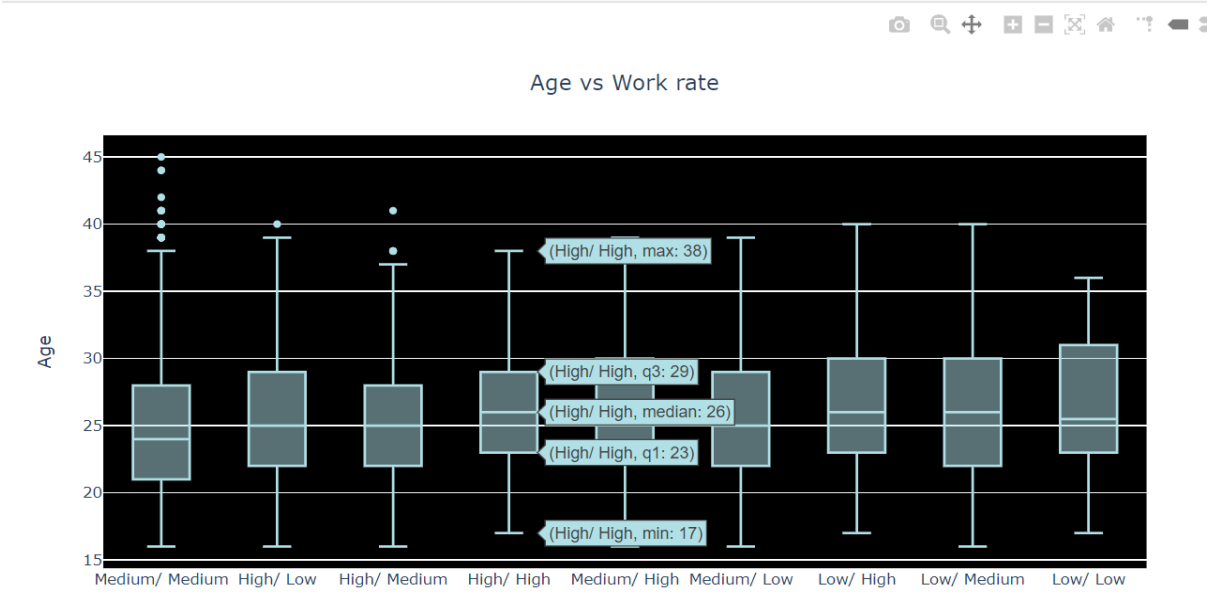

Figure 2

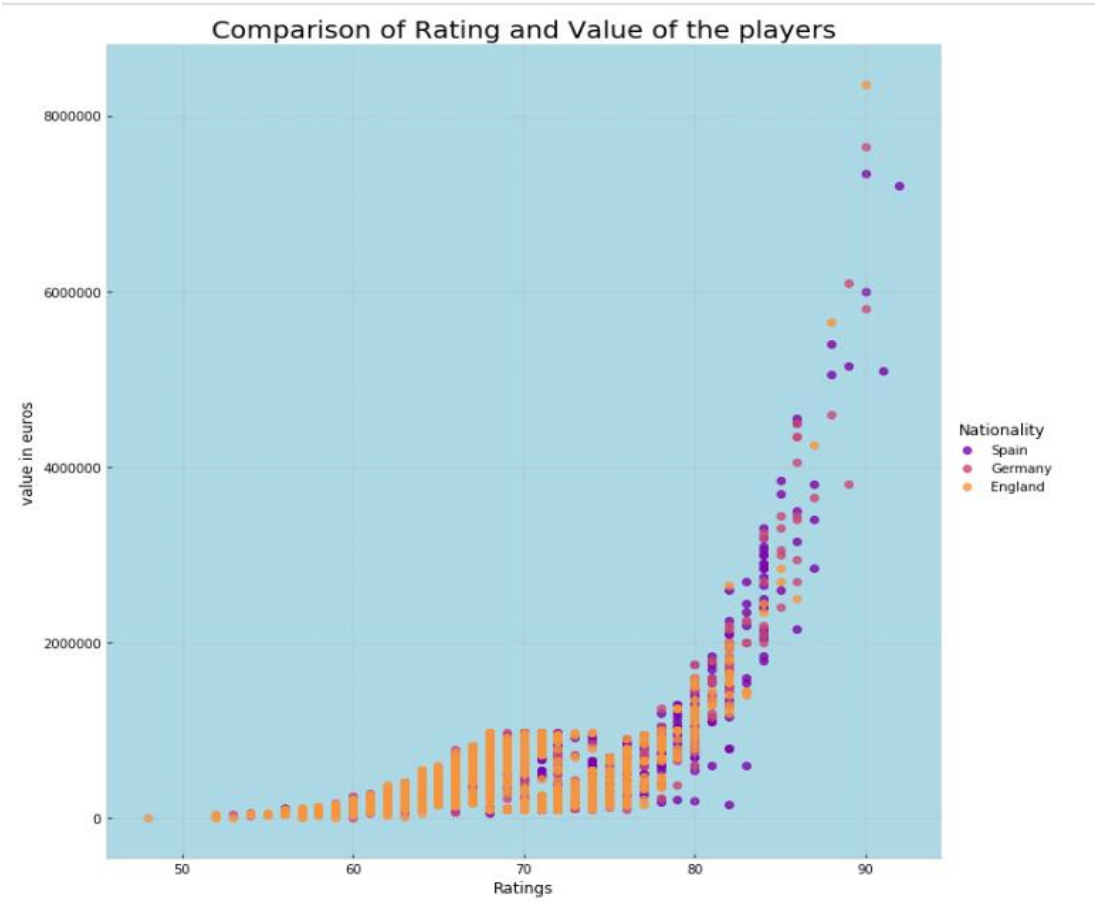

Figure 3

Conclusion

The analysis of the data set along with the visualization provides insight on how data needs to be collected, cleaned and the explored to derive meaningful patterns from the data set. It provides ways to communicate complex data in the form of beautiful graphs which convey the message better than a verbal explanation. From the above data set visualization, we see that in the footballing world the prime time of a footballer is his upper 20's. England though being one of the most famous footballing countries and having the best clubs in the world has its distribution of players in the lower value vs ratings index. This maybe because the clubs house players of various nationality and players from England are playing elsewhere. The position of striker is the most sought after by footballers as shown by the bar graph in figure 1. The field where I feel I could have improved on would be the graph plotting as I wanted more dynamic graphs to be plotted with several dimensions involved. Achieving a multivariable radial graph was tried and was not successful in coding it, would have wanted to achieve that. Wanted to implement more colours and styling in the graphs plotted and was partially successful in it.

References

https://towardsdatascience.com/data-visualization-using-seaborn-fc24db95a850
https://www.youtube.com/watch?v=a9UrKTVEeZA
https://www.kaggle.com/gpreda/plotly-tutorial-120-years-of-olympic-games