Fifa Data Visualization

A COURSE PROJECT REPORT

Submitted by

SOUMIL BALDOTA (RA2011027010141)
YUVRAJ NUGHAAL (RA2011027010138)

Under the guidance of

Dr. M. PRAKASH

In partial fulfilment for the Course

of

Machine Learning I (18CSE392T)

in

DEPARTMENT OF DATA SCIENCE AND BUSINESS SYSTEMS

SCHOOL OF COMPUTING
COLLEGE OF ENGINEERING AND TECHNOLOGY
SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
(Deemed to be University u/s 3 of UGC Act, 1956)
KATTANKULATHUR - 603 203
November, 2022

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

(Under Section 3 of UGC Act, 1956)

BONAFIDE CERTIFICATE

Certified that this mini project titled **"FIFA Data Visualization"** is the bonafide work of **Soumil Baldota (RA2011027010141), Yuvraj Nughaal (RA2011027010138),** who carried out the project work under my supervision.

21/11/2022

SUPERVISOR
Dr. M. PRAKASH
Associate Professor
Department of Data Science and Business Systems
SRM Institute of Science and Technology
Kattankulathur – 603 203

M. Lakshmi

HEAD OF THE DEPARTMENT
Dr. M. LAKSHMI
Professor & Head
Department of Data Science and Business Systems
SRM Institute of Science and Technology
Kattankulathur – 603 203

ABSTRACT

Founded in 1904 to provide unity among national soccer associations, the Federation Internationale de Football Association (FIFA) boasts 209 members, rivaling that of the United Nations, and is arguably the most prestigious sports organization in the world. In this Project we will do some analysis on the matches and records of FIFA with Python.

ACKNOWLEDGEMENT

We express our heartfelt thanks to our honorable **Vice Chancellor Dr. C. MUTHAMIZHCHELVAN**, for being the beacon in all our endeavors.

We would like to express my warmth of gratitude to our **Registrar Dr. S. Ponnusamy,** for his encouragement

We express our profound gratitude to our **Dean (College of Engineering and Technology) Dr. T. V.Gopal,** for bringing out novelty in all executions.

We would like to express my heartfelt thanks to Chairperson, School of Computing **Dr. Revathi Venkataraman,** for imparting confidence to complete my course project

We wish to express my sincere thanks to **Course Audit Professor** and **Course Coordinator** for their constant encouragement and support.

We are highly thankful to my Course project Faculty Dr. **M. Prakash, Associate Professor, Department of Data Science and Business Systems,** for his assistance, timely suggestion and guidance throughout the duration of this course project.

We extend our gratitude to our **HoD, Dr. M. Lakshmi, Professor, Department of Data Science and Business Systems,** and my Departmental colleagues for their Support.

Finally, we thank our parents and friends near and dear ones who directly and indirectly contributed to the successful completion of our project. Above all, I thank the almighty for showering his blessings on me to complete my Course project.

TABLE OF CONTENTS

1. INTRODUCTION

1.1 Scenario Description

Matches and records from the Federation Internationale de Football Association(2019) World Cup. The FIFA19 dataset contains all key attributes, and features of male football players around the world. In this analysis, We focus on major attributes to see relationship between common variables such as;

We also try to answer questions like.

- Comparison of preferred foot over the different players
- Plotting a pie chart to represent share of international reputation
- Different positions acquired by the players
- Comparing the players' Wages
- Skill Moves of Players
- Height of Players
- To show Different body weight of the players participating in the FIFA 2019
- To show Different Work rate of the players participating in the FIFA 2019
- To show Different Speciality Score of the players participating in the FIFA 2019
- To show Different potential scores of the players participating in the FIFA 2019
- To show Different nations participating in the FIFA 2019
- Every Nations' Player and their Weights
- Finding the the popular clubs around the globe
- Distribution of Wages in some Popular clubs
- Comparing the performance of left-footed and right-footed footballers

Overall it is an interesting dataset with a lot of dimension, We will however be focusing on dimensions I consider common and most important.

2. LITERATURE SURVEY

M. Burch, G. Wallner, S. L. Angelescu and P. Lakatos, "Visual Analysis of FIFA World Cup Data," *2020 24th International Conference Information Visualisation (IV)*, 2020, pp. 114-119, doi: 10.1109/IV51561.2020.00028.
Abstract: Soccer is one of the most popular sports in the world, played by thousands of professionals and amateurs every week. Consequently, it is no surprise that it generates an enormous amount of data. In today's data-driven world it is essential to find an optimal, self-explanatory, way to present the data in a way to be able to derive visual patterns that relate to the underlying data patterns. In this paper, we describe an interactive visualization for analyzing soccer data and identifying patterns, correlations, and insights. We illustrate the usefulness of our approach, especially targeted towards non-visualization experts, by applying it to World Cup data and by discussing potential use cases.
URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9373171&isnumber=9373072

Jose Luis Sotomayor Malqui, Noemí Maritza Lapa Romero, Rafael Garcia, Hande Alemdar, João L.D. Comba,
How do soccer teams coordinate consecutive passes? A visual analytics system for analysing the complexity of passing sequences using soccer flow motifs,
Computers & Graphics,
Volume 84,
2019,
Pages 122-133,
ISSN 0097-8493,
https://doi.org/10.1016/j.cag.2019.08.010.
(https://www.sciencedirect.com/science/article/pii/S0097849319301384)

3. REQUIREMENTS

3.1 Requirement Analysis

Python :

Python is an easy to learn, powerful programming language. It has efficient high-level data structures and a simple but effective approach to object-oriented programming. Python’s elegant syntax and dynamic typing, together with its interpreted nature, make it an ideal language for scripting and rapid application development in many areas on most platforms. The Python interpreter and the extensive standard library are freely available in source or binary form for all major platforms from the Python Web site, https://www.python.org/, and may be freely distributed. The same site also contains distributions of and pointers to many free third party Python modules, programs and tools, and additional documentation.

Seaborn :

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.For a brief introduction to the ideas behind the library, you can read the introductory notes or the paper. Visit the installation page to see how you can download the package and get started with it. You can browse the example gallery to see some of the things that you can do with seaborn, and then check out the tutorials or API reference to find out how.

Pandas :

pandas aims to be the fundamental high-level building block for doing practical, real world data analysis in Python. Additionally, it has the broader goal of becoming the most powerful and flexible open source data analysis / manipulation tool available in any language.

Matplotlib :

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.

4. DATA SET DESCRIPTION

This dataset consists of fo all the details of the players who played in the FIFA World Cup and their final value at which they were sold to the club(target variable).

Provenance:

Sources:

The data is officially collected from record of the players who participated in the world cup matches.

Collection Methodology:

This dataset consists of fo all the details of the players who played in the FIFA World Cup and their final value at which they were sold to the club(target variable).

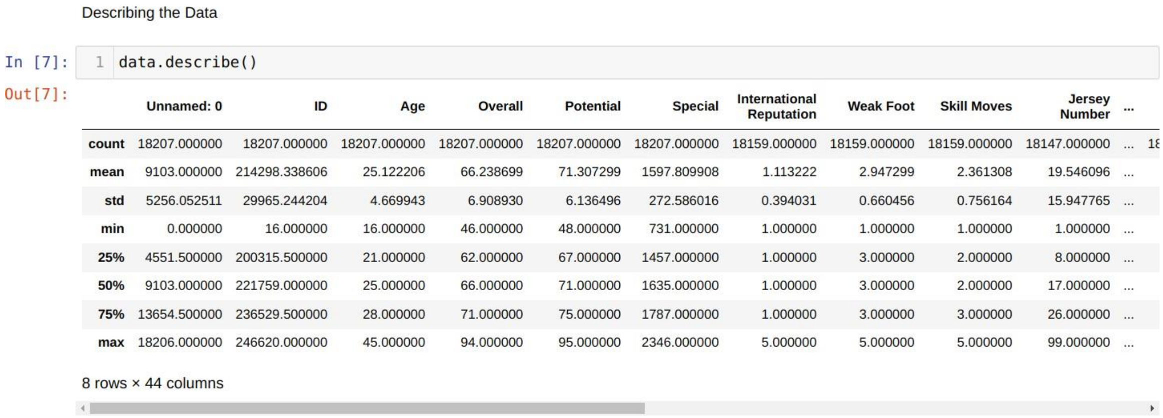
License

CC0: Public Domain

Size :

This Dataset contains data of more than 18000 players. With 89 columns/Fields.

Description:



Describing the Data

```
In [7]:   1 data.describe()
```

Out[7]:

	Unnamed: 0	ID	Age	Overall	Potential	Special	International Reputation	Weak Foot	Skill Moves	Jersey Number	...	
count	18207.000000	18207.000000	18207.000000	18207.000000	18207.000000	18207.000000	18159.000000	18159.000000	18159.000000	18147.000000	...	18
mean	9103.000000	214298.338606	25.122206	66.238699	71.307299	1597.809908	1.113222	2.947299	2.361308	19.546096	...	
std	5256.052511	29965.244204	4.669943	6.908930	6.136496	272.586016	0.394031	0.660456	0.756164	15.947765	...	
min	0.000000	16.000000	16.000000	46.000000	48.000000	731.000000	1.000000	1.000000	1.000000	1.000000	...	
25%	4551.500000	200315.500000	21.000000	62.000000	67.000000	1457.000000	1.000000	3.000000	2.000000	8.000000	...	
50%	9103.000000	221759.000000	25.000000	66.000000	71.000000	1635.000000	1.000000	3.000000	2.000000	17.000000	...	
75%	13654.500000	236529.500000	28.000000	71.000000	75.000000	1787.000000	1.000000	3.000000	3.000000	26.000000	...	
max	18206.000000	246620.000000	45.000000	94.000000	95.000000	2346.000000	5.000000	5.000000	5.000000	99.000000	...	

8 rows × 44 columns

5. METHOD/ALGORITHM/MODEL USED

Methods Used:

plt.pie :

Plot a pie chart. Make a pie chart of array x. The fractional area of each wedge is given by x/sum(x). The wedges are plotted counterclockwise, by default starting from the x-axis.

sns.distplot:

Flexibly plot a univariate distribution of observations.This function combines the matplotlib hist function (with automatic calculation of a good default bin size) with the seaborn kdeplot() and rugplot() functions. It can also fit scipy.stats distributions and plot the estimated PDF over the data.

sns.countplot:

Show the counts of observations in each categorical bin using bars. A count plot can be thought of as a histogram across a categorical, instead of quantitative, variable. The basic API and options are identical to those for `barplot()`, so you can compare counts across nested variables.

Input data can be passed in a variety of formats, including:

- Vectors of data represented as lists, numpy arrays, or pandas Series objects passed directly to the `x`, `y`, and/or `hue` parameters.
- A “long-form” DataFrame, in which case the `x`, `y`, and `hue` variables will determine how the data are plotted.
- A “wide-form” DataFrame, such that each numeric column will be plotted.
- An array or list of vectors.

sns.violinplot:

Draw a combination of boxplot and kernel density estimate. A violin plot plays a similar role as a box and whisker plot. It shows the distribution of quantitative data across several levels of one (or more) categorical variables such that those distributions can be compared. Unlike a box plot, in which all of the plot components correspond to actual datapoints, the violin plot features a kernel density estimation of the underlying distribution. This can be an effective and attractive way to show multiple distributions of data at once, but keep in mind that the estimation procedure is influenced by the sample size, and violins for relatively small samples might look misleadingly smooth.

sns.boxplot:

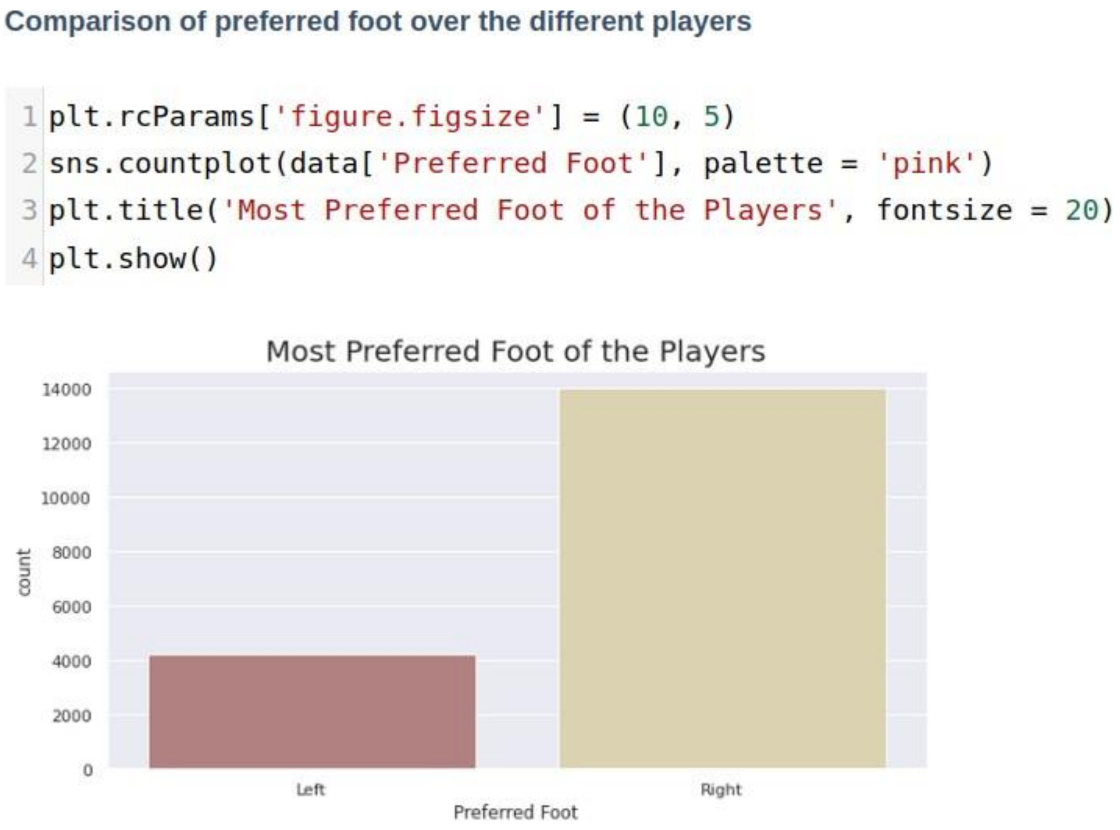
Draw a box plot to show distributions with respect to categories. A box plot (or box-and-whisker plot) shows the distribution of quantitative data in a way that facilitates comparisons between variables or across levels of a categorical variable. The box shows the quartiles of the dataset while the whiskers extend to show the rest of the distribution, except for points that are determined to be "outliers" using a method that is a function of the inter-quartile range.

sns.lmplot:

This function combines `regplot()` and `FacetGrid`. It is intended as a convenient interface to fit regression models across conditional subsets of a dataset. When thinking about how to assign variables to different facets, a general rule is that it makes sense to use `hue` for the most important comparison, followed by `col` and `row`. However, always think about your particular dataset and the goals of the visualization you are creating. There are a number of mutually exclusive options for estimating the regression model. See the tutorial for more information. The parameters to this function span most of the options in `FacetGrid`, although there may be occasional cases where you will want to use that class and `regplot()` directly.

6. RESULTS AND DISCUSSION

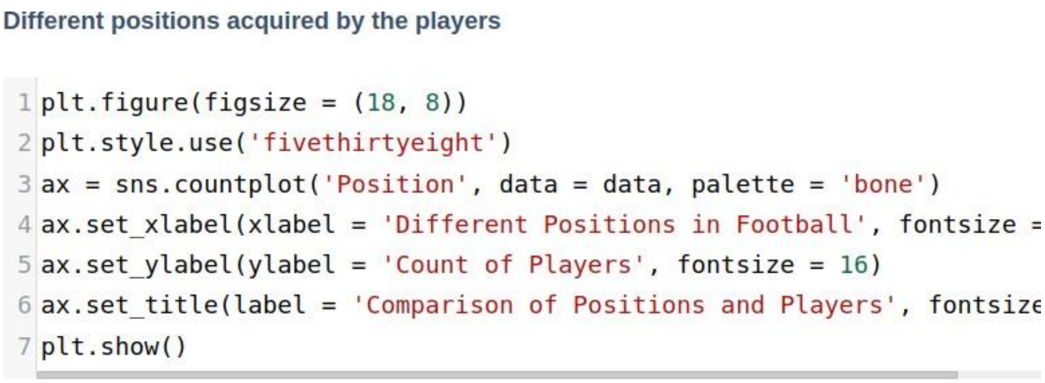
The Data Clearly shows a greater number of players being right footed.



Comparison of preferred foot over the different players

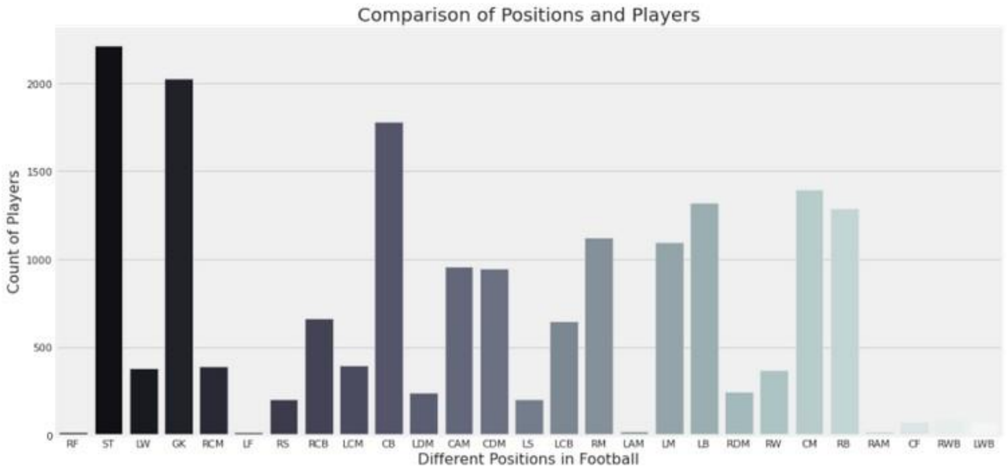
```
plt.rcParams['figure.figsize'] = (10, 5)
sns.countplot(data['Preferred Foot'], palette = 'pink')
plt.title('Most Preferred Foot of the Players', fontsize = 20)
plt.show()
```

The Data depicts a large number of strikers and Mid fielders.

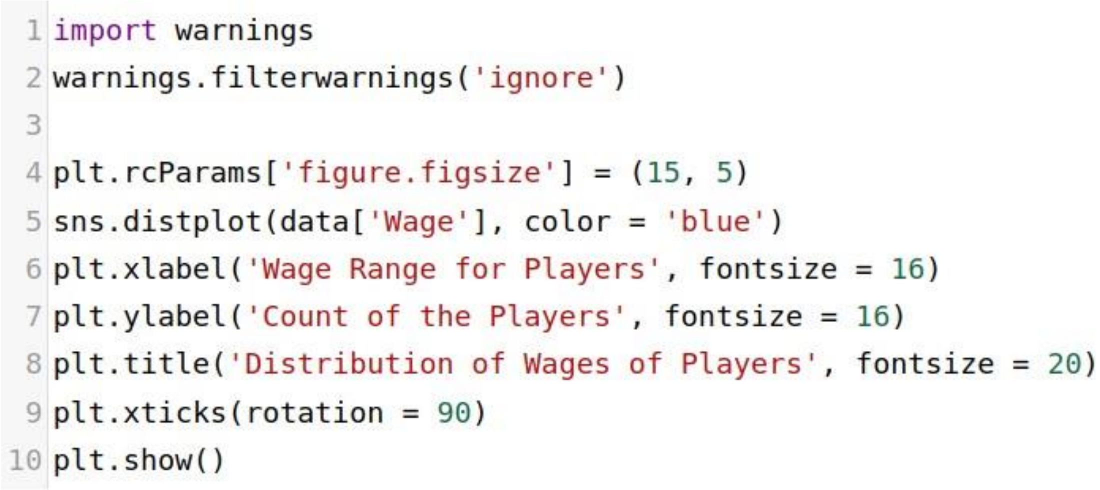


Different positions acquired by the players

```
plt.figure(figsize = (18, 8))
plt.style.use('fivethirtyeight')
ax = sns.countplot('Position', data = data, palette = 'bone')
ax.set_xlabel(xlabel = 'Different Positions in Football', fontsize =
ax.set_ylabel(ylabel = 'Count of Players', fontsize = 16)
ax.set_title(label = 'Comparison of Positions and Players', fontsize
plt.show()
```

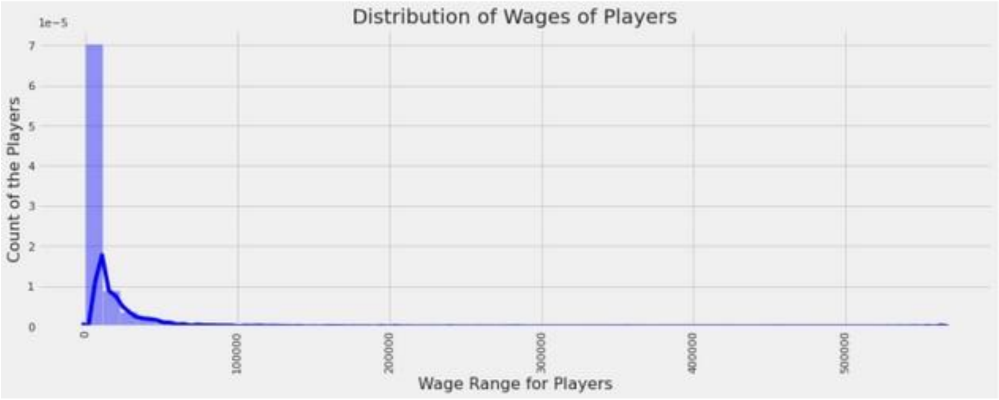


The Data shows a sharp normal distribution of Wages.



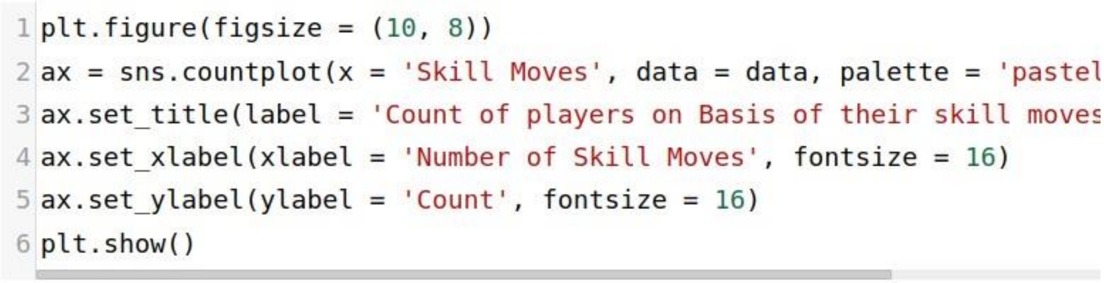
```python
import warnings
warnings.filterwarnings('ignore')

plt.rcParams['figure.figsize'] = (15, 5)
sns.distplot(data['Wage'], color = 'blue')
plt.xlabel('Wage Range for Players', fontsize = 16)
plt.ylabel('Count of the Players', fontsize = 16)
plt.title('Distribution of Wages of Players', fontsize = 20)
plt.xticks(rotation = 90)
plt.show()
```

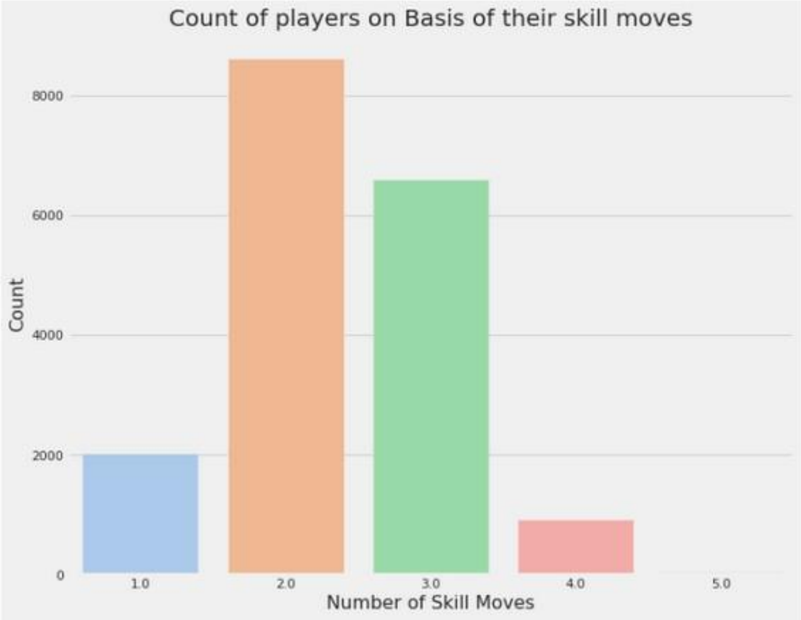


The data shows 2-3 skill moves in majority of players

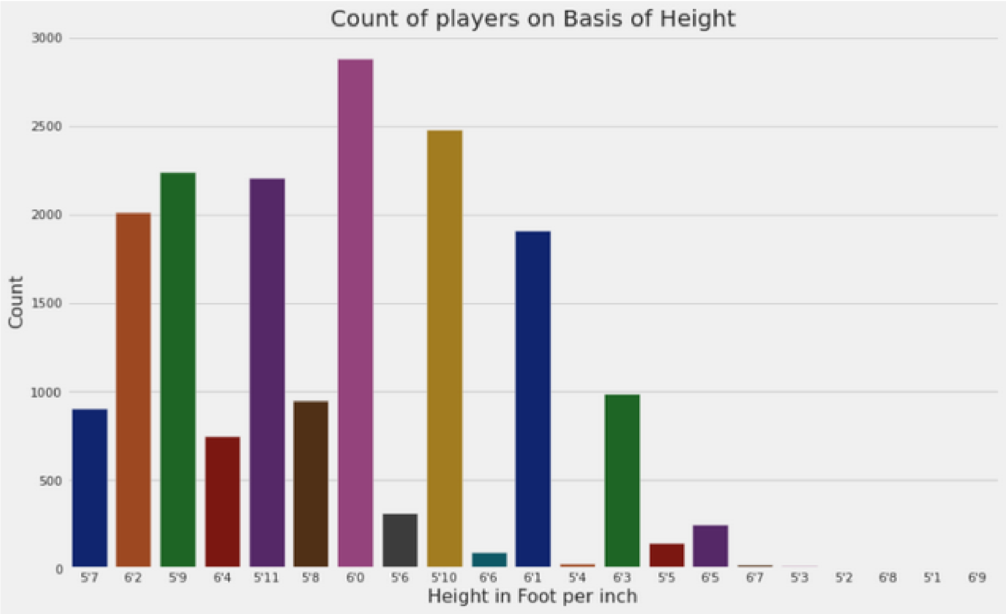
Skill Moves of Players



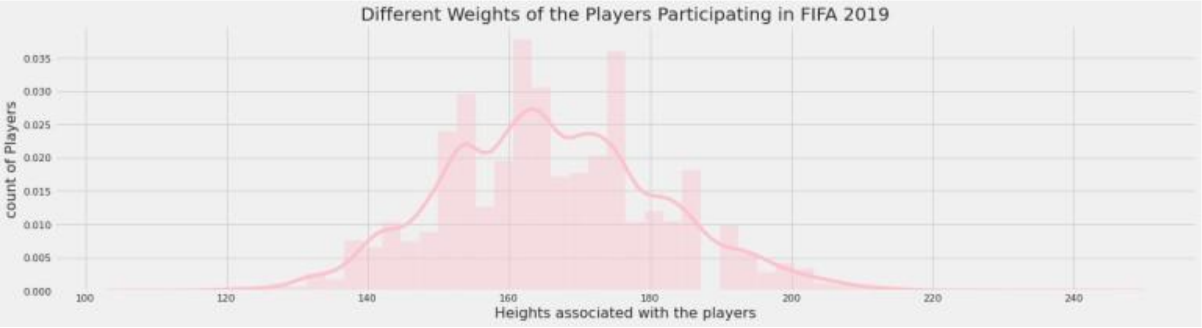
```
plt.figure(figsize = (10, 8))
ax = sns.countplot(x = 'Skill Moves', data = data, palette = 'pastel
ax.set_title(label = 'Count of players on Basis of their skill moves
ax.set_xlabel(xlabel = 'Number of Skill Moves', fontsize = 16)
ax.set_ylabel(ylabel = 'Count', fontsize = 16)
plt.show()
```



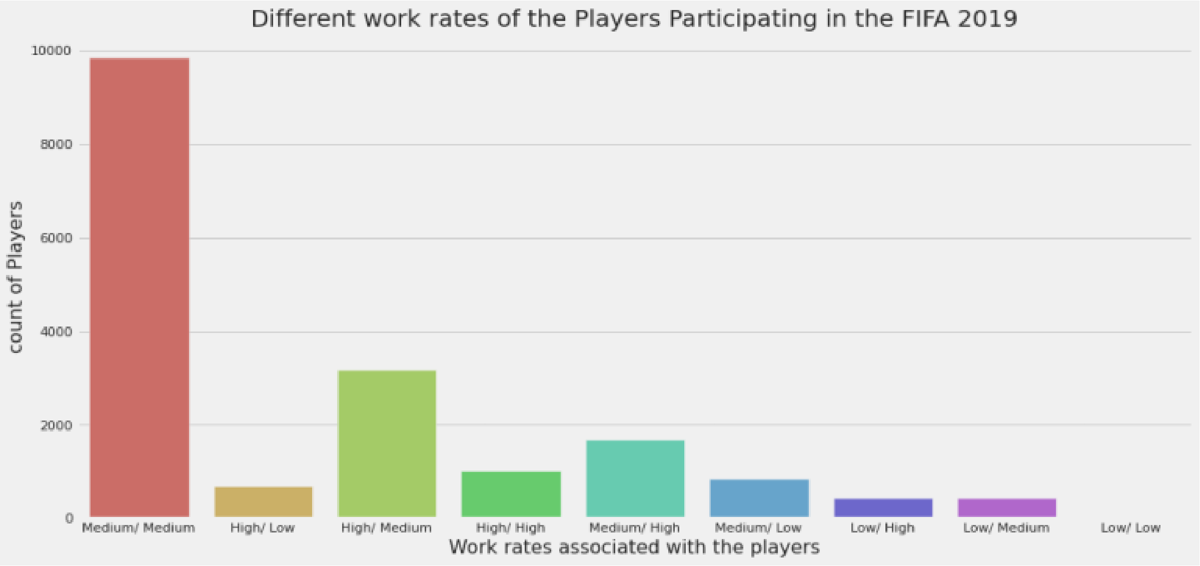
Count of heights of Players



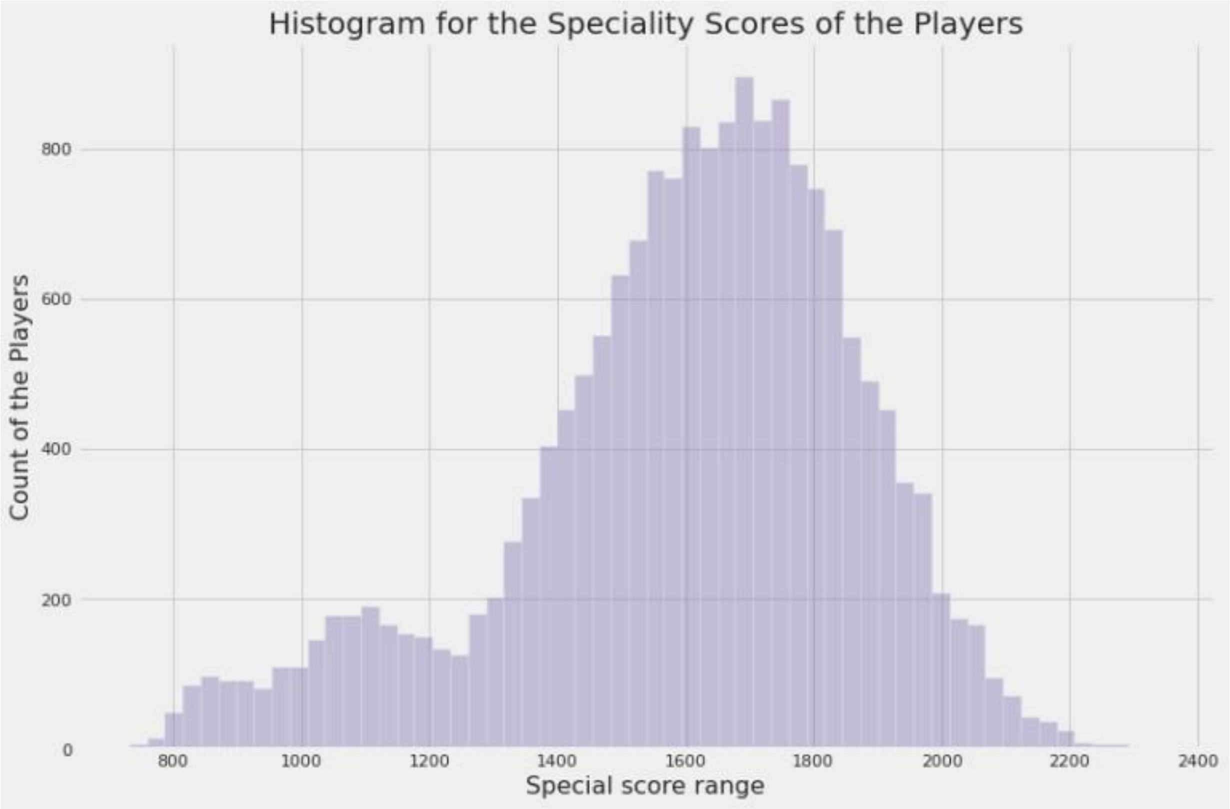
Weight Distribution of Different Players



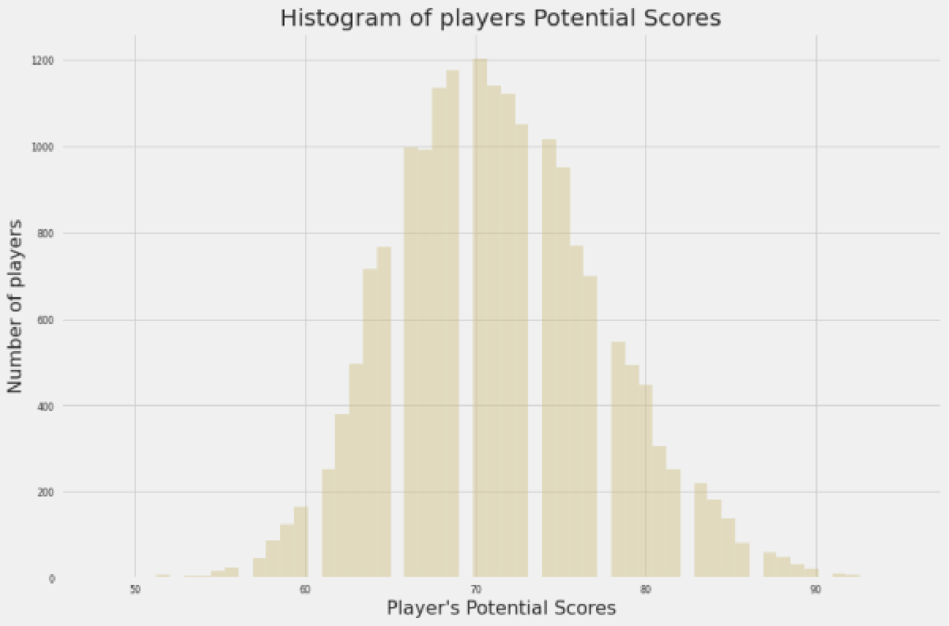
Work Rates of Different Players



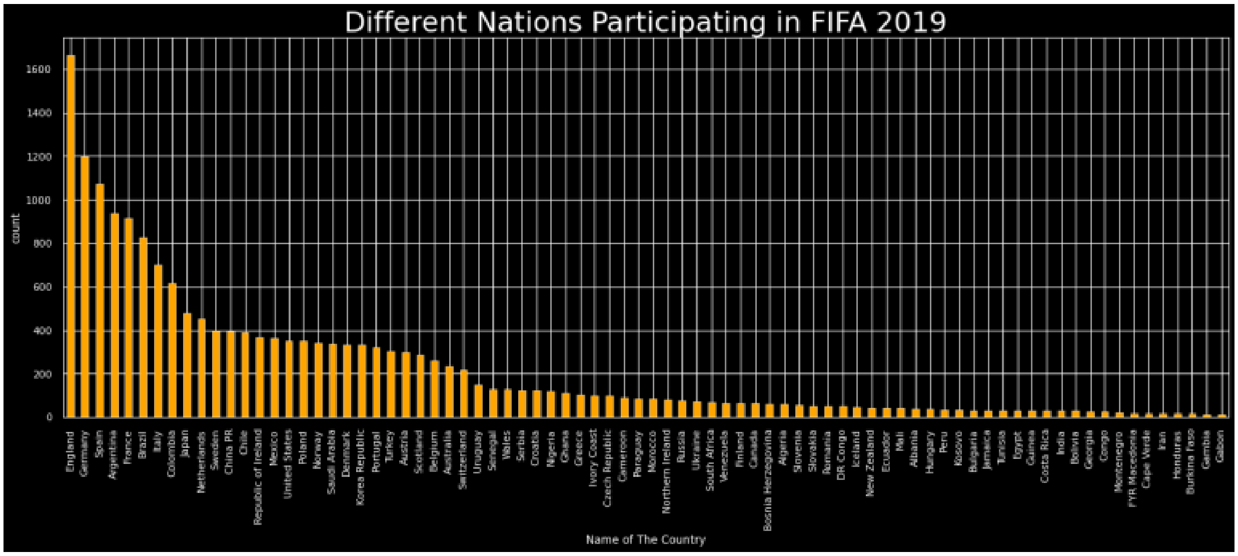
To show Different Speciality Score of the players



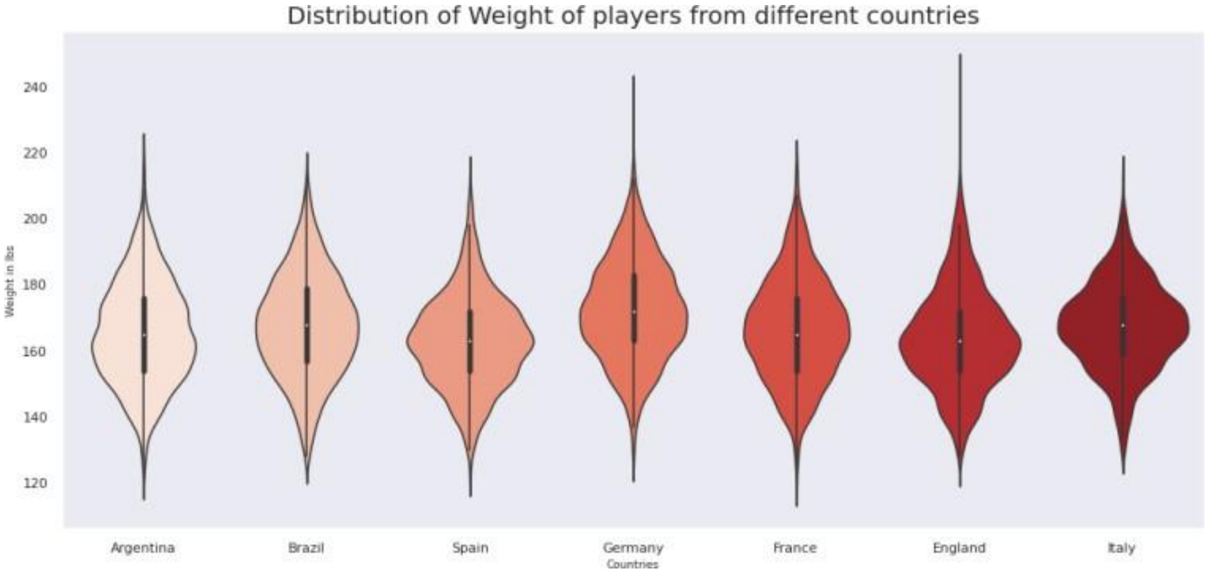
To show Different potential scores of the players



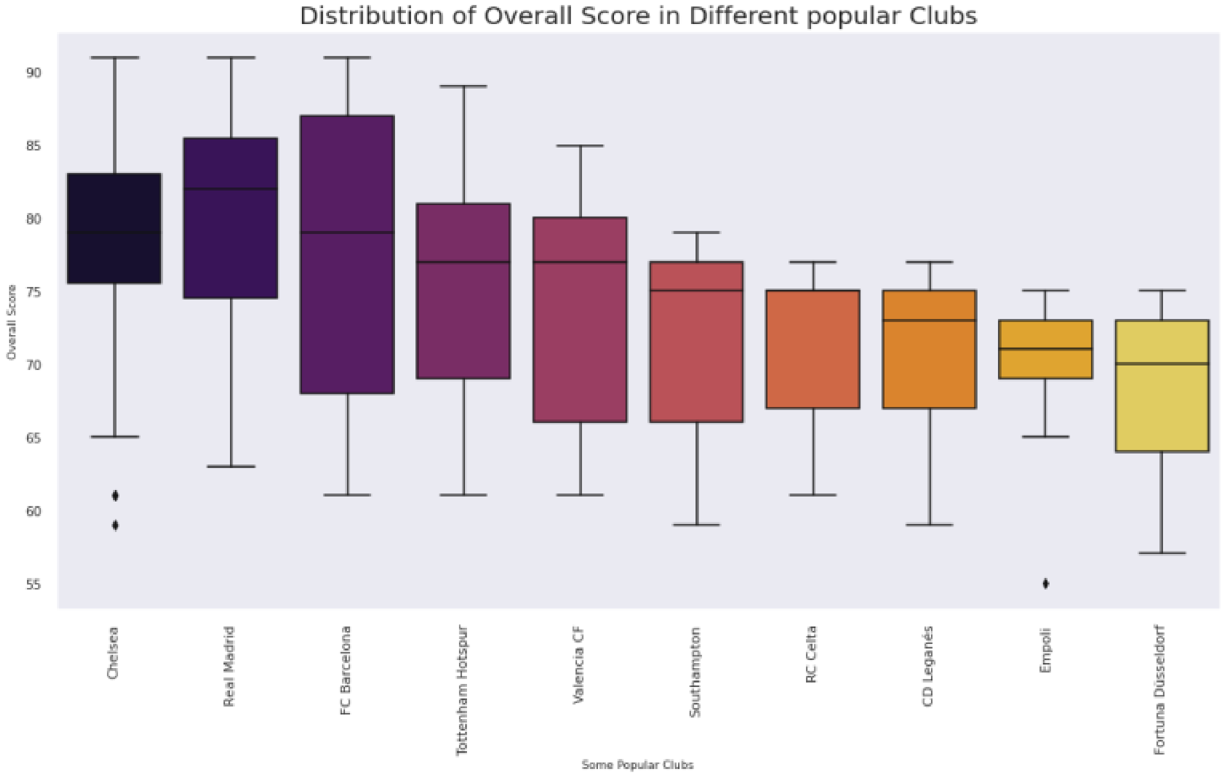
To show Different nations participating



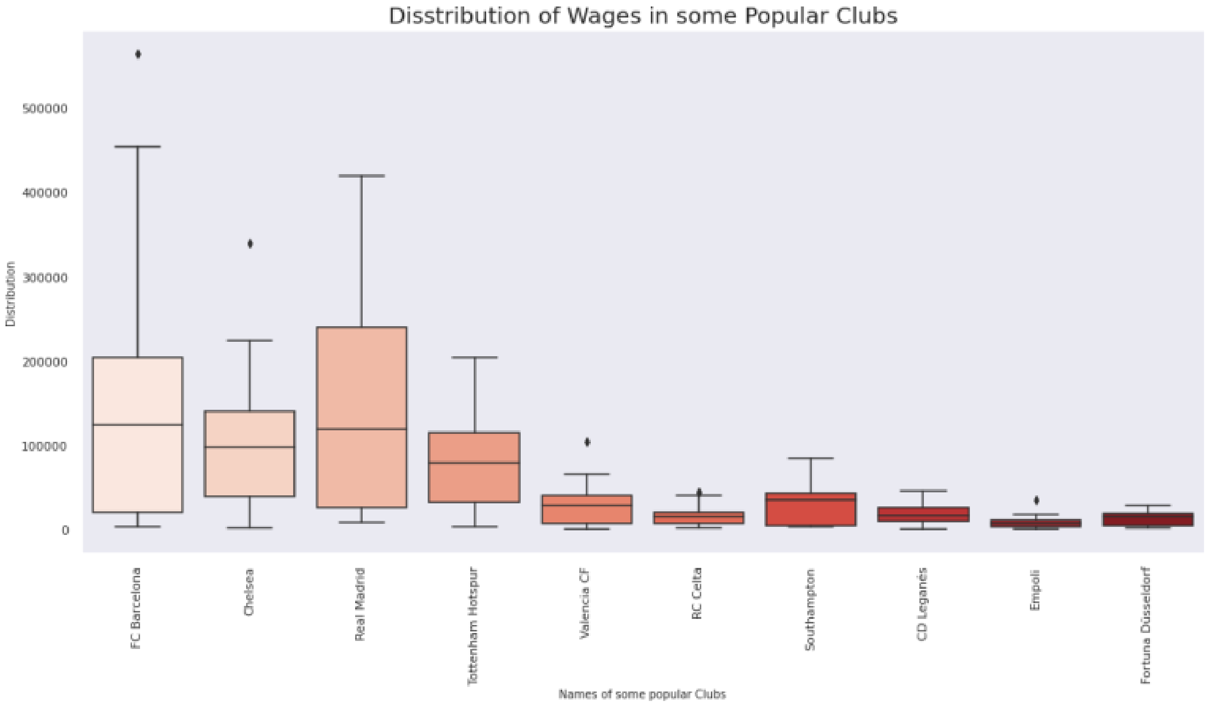
Every Nations’ Player and their Weights



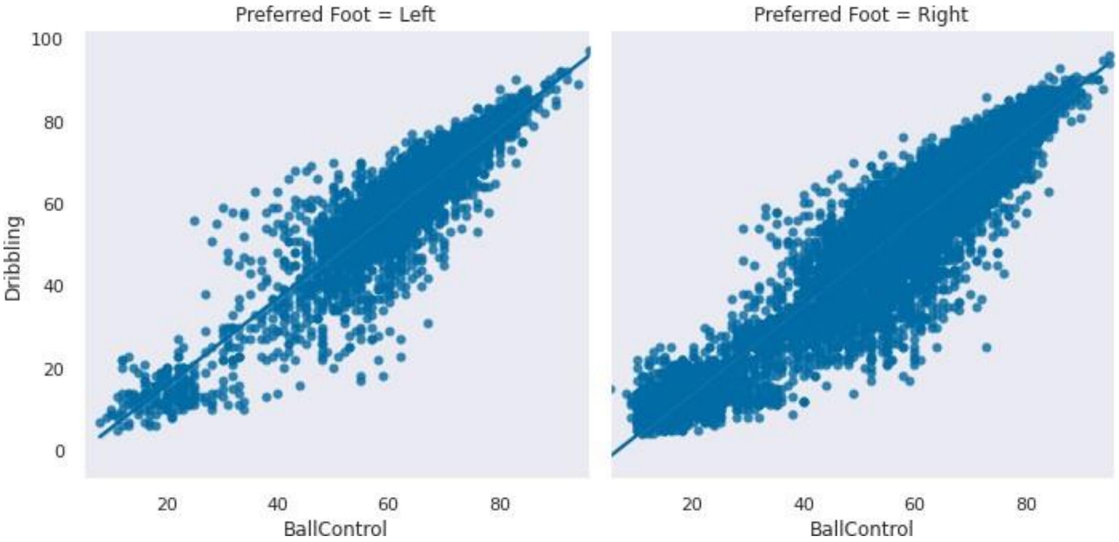
Finding the the popular clubs around the globe



Distribution of Wages in some Popular clubs



Comparing the performance of left-footed and right-footed footballers



7. CONCLUSION AND FUTURE ENHANCEMENT

The Results conclude the Exploratory Data Analysis of The FIFA 2019 dataset. Further Analysis / Future Analysis could include comparision of the FIFA 2019 Dataset to previous datasets regarding the past FIFA world cups.

REFERENCES

M. Burch, G. Wallner, S. L. Angelescu and P. Lakatos, "Visual Analysis of FIFA World Cup Data," *2020 24th International Conference Information Visualisation (IV)*, 2020, pp. 114-119, doi: 10.1109/IV51561.2020.00028.

Abstract: Soccer is one of the most popular sports in the world, played by thousands of professionals and amateurs every week. Consequently, it is no surprise that it generates an enormous amount of data. In today's data-driven world it is essential to find an optimal, self-explanatory, way to present the data in a way to be able to derive visual patterns that relate to the underlying data patterns. In this paper, we describe an interactive visualization for analyzing soccer data and identifying patterns, correlations, and insights. We illustrate the usefulness of our approach, especially targeted towards non-visualization experts, by applying it to World Cup data and by discussing potential use cases.

URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9373171&isnumber=9373072

Jose Luis Sotomayor Malqui, Noemí Maritza Lapa Romero, Rafael Garcia, Hande Alemdar, João L.D. Comba,

How do soccer teams coordinate consecutive passes? A visual analytics system for analysing the complexity of passing sequences using soccer flow motifs,

Computers & Graphics,

Volume 84,

2019,

Pages 122-133,

ISSN 0097-8493,

https://doi.org/10.1016/j.cag.2019.08.010.

(https://www.sciencedirect.com/science/article/pii/S0097849319301384)