ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645
Indian Journal of Science and Technology, *Vol 12(39), DOI: 10.17485/ijst/2019/v12i39/145565, October 2019*

Analysis of FIFA World Cup Data Set

Mansi Sharma[1,2,*], Palak Mittal[1,2], Nidhi Garg[1,2] and Dr. Prateek Jain[1,2]

[1]Department of Computer Science & Engineering, Manav Rachna International Institute of Research & Studies, aridabad, India; mansi261198@gmail.com, palakmittal2109@gmail.com, nidhi.fet@mriu.edu.in, prateek.jain@accendere.co.in
[2]Accendere CL Educate Ltd, India

Abstract

Background/objectives: To analyze a data set related to FIFA World Cup using a suitable method. **Methods/statistical analysis:** In this study we have taken up the data sets of the FIFA World Cup and analyzed them using Python and R programming. The analysis focused on: a) which team conceded a greater number of goals than they scored; b) the percentage of goals scored in the First Half, Second Half, Extra Time, and Penalty Shootout; the c) highest average attendance in a particular stage of the match. **Findings:** Python seems to be an emerging programming language and is thriving to a great extent. Due to its advantages of easy-to-learn syntax, improved readability, object-oriented programming support, integration support, and extensive libraries, this language is adaptable in many fields and hence increasing its applications. **Improvements/applications:**

Keywords: Big Data, Analytics, Data Set, API, Machine Learning

1. Introduction

In the data era, sizeable quantities of statistics have come to be reachable to decision makers. Big data refers to data sets that are now not only big, but additionally high in range and velocity, which makes them challenging to be taken care of using normal tools and techniques. Due to the speedy boom of such data, options need to be studied and supplied to take care of and extract price and expertise from these data sets. Furthermore, decision makers want to be in a position to obtain treasured insights from such varied and unexpectedly changing data. Such fee can be furnished using huge records analytics, which is the utility of advanced analytics techniques on big data. There are a number of tools that can be used for storing and analyzing data. Some of the popular tools for storing data are:

1. Apache Hadoop: It can be used to store enormous amounts of data in a cluster. It is a Java-based framework. It can run in parallel on a cluster and is capable of allowing users to process data across all nodes. This provides replication of data resulting in high availability of data.
2. Hive: It's a distributed data management for Hadoop. It can be used for data mining purposes as it supports query operations, such as Hive SQL, for accessing the big data.
3. Apache Cassandra: It is a No SQL database. It is scalable and has high-performance distributed database to handle large amounts of data. We can store and retrieve data other than tabular relations with the help of a No SQL database. The qualities of this database are that it is schema-free, has a simple API, is consistent, supports easy replication, and can handle large amounts of data.[1]

Some of the popular tools for analyzing data are:

1. Rapid Miner: Rapid Miner can include any number of information source types, which include Microsoft SQL, Sybase, IBM SPSS, Excel, Oracle, My SQL, Access, Tera Data, IBM DB2, Ingress, and Dbase. The tool is very effective and can generate analysis

*Author for correspondence

primarily based on real-life record transformation settings.
2. Tableau Public: It's an intuitive and simple tool that offers interesting insights by data visualisation. One can inspect a hypothesis, discover the data, and cross-check their insights.
3. Jupyter Notebook: It is an accessible tool for performing end-to-end data science workflows – information cleansing, statistical modelling, building and training machine learning models, and visualising data.[2,3]

In this study we used Jupyter Notebook for analysing data.

2. Data Sets

A data set is a collection of similar and related data or information. It is organised for better accessibility of an entity. Data sets are used for data analytics as they provide related information in a united form. It can be structured or unstructured. Structured data sets mean that it is structured in a proper way like a tabular data set; such data sets contain information in the form of tables with rows, column, cells; hence, when we talk about data set being structured, it means that is arranged in some predefined model or format. Contrary to structured data sets, an unstructured data set is not arranged in some predefined format such as tables and is text-heavy, containing facts, numbers, and other information.

Depending upon the work, a type of data set is chosen that suits the requirements better. In this paper, a structured data set is taken in consideration for further work containing data in a tabular form with rows and columns.

In this study we used a data set related to the FIFA World Cup; it is a structured data set in the form of rows and columns.

The data set contains data about Goals, Matches, Players, and Teams which in turn have further fields, that is, columns of different data types.

1. Goals contains data about player name in the field, called Player Name; Id Match containing id assigned to match; Team Name having team names, for example, Russia, UK, Portugal, Opposition Team names, Goal Keeper names, player shirt number, and number of minutes.
2. Matches contains Id Match, Home Team, Away Team, Attendance, Match Day, Stage, Home Team Tactics, Away Team Tactics, Penalty Score for both the teams, Stadium Name having names of stadium where match was held, Temperature, Humidity, Wind Speed, Winner of the respective matches.
3. Players contain Id Player for each player, Name, Team Name they are associated with, Birth Date, Weight, Height, and Goals.
4. Teams contain Id Team, Team Name, Coach Name associated with teams, Coach Country signifies the country of coach.

2.1. Importance of Data Analytics

While analysing data sets, it is important to define the objectives so that further steps become clearer. Analysis lets us pose questions about data. For questioning data, it is important to have data collection on which further operations will be carried out.

After the above steps, "Data Wrangling" comes into picture. Data wrangling or data munging is the process of raw data cleansing and conversion so that further operations become easier to carry on and then the conclusions can be drawn from the results.[4]

Today, data has become the backbone of all research in almost every field. Research and analysis is no more limited to just the area of sciences, but has grown to be a part of businesses – startups and established organisations, government works and more.[5]

3. R for Data Analytics

When it comes to the tools used in data analytics, R has been ruling the area for quite some times. In fact, before Python, R was the language for data science and still is for many organisations. Most of the data science-related work has been done in R language. R runs on many modern phones, tablets and game console systems, but R has its own limitations. Let us discuss some of them.

3.1. Limitations of R

Not every language is perfect and so is R. Below are listed some of the limitations of R:

1. R is very similar to S language; in fact, it is an implementation of S language and hence it is based on a technology which itself is 50 years old.
2. Its object must be generally stored in physical memory; it is a part of scoping rules of language.

3. R's functionality is based on consumer demand and (voluntary) user contributions. If no one feels like implementing your favorite method, then it becomes your own job to implement it yourself or get someone who can[6] (Figure 1).

4. Results

We are making three analyses from the FIFA World Cup data set. These analyses are:

- which teams conceded a greater number of goals than they scored;
- the percentage of goals scored in the First Half, Second Half, Extra Time, and Penalty Shootout;
- the highest average attendance in a particular stage of the match.

4.1. Analysis 1

To identify which teams conceded a greater number of goals than they scored, we plotted a bar chart of goals conceded and goals scored by all the teams that participated in the World Cup. A pictorial representation of the goals conceded and scored by individual teams is shown in Figure 2.

4.2. Analysis 2

To calculate the percentage of goals scored in the first half, second half, extra time, and penalty shootout of the matches, we plotted a pie chart to represent the respective percentages of goals scored in the first half, second half, extra time, and penalty shootout by all teams. The orange part represents the first half; the blue part represents the second half; the red part represents the extra time, and the green part represents the penalty shootout (Figure 3).

According to the graph, 33.3% goals were done in first half, 51.8% in second half, 15% in extra time, and 13.3% in penalty shootout.

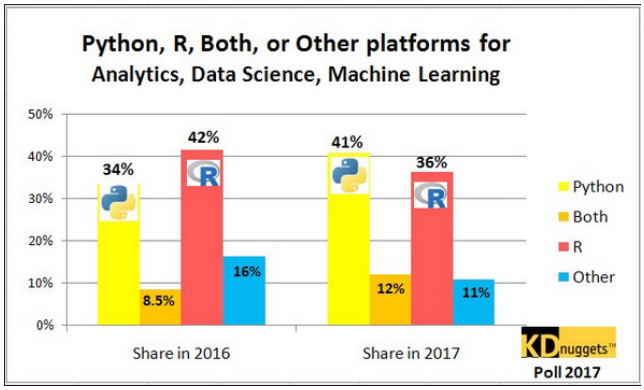

Figure 1. Comparing Python and R in 2016 and 2017.[7]

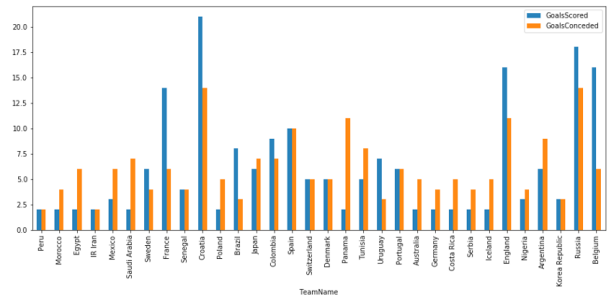

Figure 2. Teams that conceded a greater number of goals than they scored.

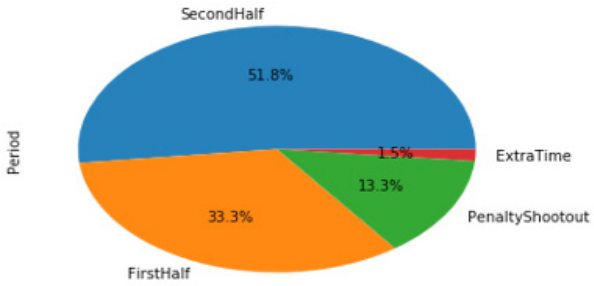

Figure 3. Pie chart representation for analysis 2.

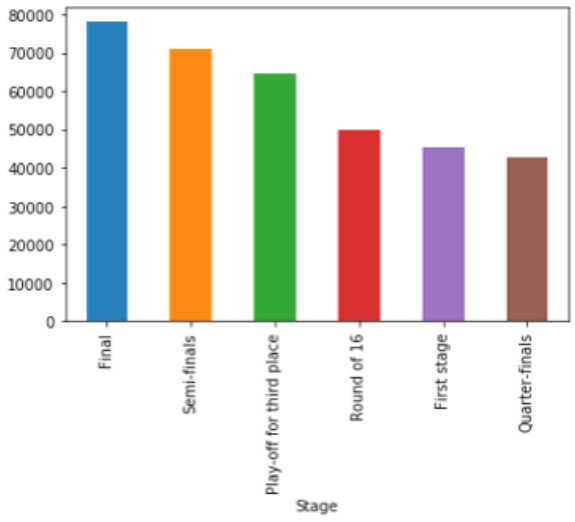

Figure 4. Bar chart for analysis 3.

4.3. Analysis 3

To find the highest average attendance in a particular stage of the match, we considered six stages, namely, first stage, round of 16, play-off for third place, quarter-final, semi-final, and final (Figure 4).

As shown in the graph, the finals had the maximum average attendance of 78011.000, and the quarter-finals had the least average attendance of 42617.500.

5. Conclusion

In recent times, Python seems to be an emerging programming language and is thriving to a great extent. Due to its advantages of easy-to-learn syntax, improved readability, object-oriented programming support, integration support, and extensive libraries, this language is adaptable in many fields and hence increasing its applications. In this paper, we have taken up the data sets of the FIFA World Cup and analyzed them using Python and R programming to identify which team conceded a greater number of goals than they scored; the percentage of goals scored in the First Half, Second Half, Extra Time, and Penalty Shootout; and the highest average attendance in a particular stage of the match.

References

1. Mittal P, Sharma M, Jain P. A detailed study of security and privacy concerns in big data. Int J Appl Eng Res. 2018;13:7406–11
2. Paul Z, Eaton C. Understanding big data: analytics for enterprise class Hadoop and streaming data. McGraw-Hill Osborne Media;2011. P. 1–166.
3. Philip R. Big data analytics. TDWI best practices report, fourth quarter. 2011;19(4):1–34.
4. Karthik K. Trends in big data analytics. J Parallel Distr Com. 2014;74(7):2561–73.
5. Chong Ho Y. Exploratory data analysis. Methods. 1977;2:131–60.
6. Gdata: various R programming tools for data manipulation. [cited 2005 May]. https://www.researchgate.net/publication/275646231_gdata_Various_R_programming_tools_for_data_manipulation.
7. The art of R programming: a tour of statistical software design. [cited 2012 Apr]. https://www.researchgate.net/publication/254296013_The_Art_of_R_Programming_A_Tour_of_Statistical_Software_Design_by_Norman_Matloff.