**PROJECT REPORT ON**

**ANALYSIS ON INTERNATIONAL FOOTBALL RESULTS**

***In partial fulfillment of I st Semester***

**MSc Computer Science**

**Specialization in Soft Computing**

**Submitted By**

**Naheela PK**

**Sathyajith KP**

**Edwin U Kannanaikkal**

****

**DEPARTMENT OF COMPUTER APPLICATIONS**

**COCHIN UNIVERSITY OF SCIENCE AND TECHNOLOGY**

**COCHIN-22**

**NOVEMBER 2019**

**DEPARTMENT OF COMPUTER APPLICATIONS**

**UNIVERSITY OF SCIENCE AND TECHNOLOGY COCHIN-22**

****

**BONAFIDE CERTIFICATE**

This is to certify that the project report entitled “**Analysis on International Football Results**” is a bonafide record of the project work done by *Naheela PK, Sathyajith KP & Edwin U Kannanaikkal*, in the partial fulfillment of the requirements of the Ist semester MSc Computer Science specialization in soft computing at Department of Computer Applications, Cochin University of Science and Technology

**Head of the Department**

Department of Computer Applications

CUSAT

**External Examiner Internal Examiner**

Department of Computer Applications Department of Computer Applications

CUSAT CUSAT

**DECLARATION**

We hereby declare that the project work entitled with “**Analysis on International Football Results**” is submitted in the partial fulfillment of the requirements for the Ist semester, MSc Computer Science specialization in soft computing and it is a report of the original work done by us in the Department of Computer Applications, Cochin University of Science And Technology, Cochin**.**

Place: Kalamassery Naheela P K

Date: 25/11/2019 Sathyajith K P

Edwin U Kannanaikkal

**ACKNOWLEDGEMENT**

With great pleasure we hereby acknowledge that the help given to as by various individuals throughout the project/ the project itself is an acknowledgement to the inspiration driven and technical assistance contributed by many individuals.

We extend our sincere thanks to all the non-teaching staff for providing the necessary facilities and help. We thank Lord God, the almighty for His immeasurable blessing upon my life.

We are pleased our indebtedness to Dr. Sabu M K, Head of the Department, Department of Computer Applications, CUSAT for his gracious encouragement. And also, we are thankful to Ms. Jiby T C, our project guide for her support.

We are obliged to the teaching staff for being helpful and co-operative during the period of project. We extend our heartfelt thankful to parents, friends and well wishes for their support and timely help.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **No:** | **Title** | **Page no:** |
| 1. | Abstract | 6 |
| 2. | Introduction | 7 |
| 3. | Literature Review | 10 |
| 4. | System requirements | 11 |
| 5. | Design and Implementation | 12 |
| 6. | Result and Analysis | 26 |
| 7. | Conclusion | 28 |
| 8. | Bibliography | 29 |

**ABSTRACT**

Kaggle is the place to do data science projects. Our project was inspired by the Kaggle.com “[International football results from 1872 to 2019](https://www.kaggle.com/martj42/international-football-results-from-1872-to-2017)” dataset. From International football results we can predict the game trend of 1872 to 2019. Between these years there are 40,000+ matches have been played between countries in championships and friendly matches. In our project we visualized this data set for obtaining the following conclusions

1. Number of matches played in each year.
2. Best 10 teams of all time (Based on goal score).
3. Which teams dominated different periods of football:

(1870-1920, 1920-1970, 1970-2020)

1. What trends have there been in international football throughout the ages (home advantage)

From the analysis through various plotting methods we are able to determine the trends that changed over the years in International football

**INTRODUCTION**

Data Science to provide a stronger foundation in data visualization in Python. **Data Visualization** converts information into visuals. And through our affinity for visuals, we can more quickly and easily understand the information being represented.

Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed focus on quantitative descriptions and estimations of data. Data visualization provides an important suite of tools for gaining a qualitative understanding.

This can be helpful when exploring and getting to know a dataset and can help with identifying patterns, corrupt data, outliers, and much more. With a little domain knowledge, data visualizations can be used to express and demonstrate key relationships in plots and charts that are more visceral to yourself and stakeholders than measures of association or significance.

There are five key plots that we need to know well for basic data visualization.

* **Line Plot**

A line plot is generally used to present observations collected at regular intervals. The x-axis represents the regular interval, such as time. The y-axis shows the observations, ordered by the x-axis and connected by a line. A line plot can be created by calling the plot () function and passing the x-axis data for the regular interval, and y-axis for the observations.

* **Count Plot**

A count plot can be thought of a histogram across a categorical, instead of quantitative variable. The basic API and options are identical to those for barplot(). Input data can be passed as vectors of data represented as lists,

NumPy arrays, or pandas Series objects passed directly to the x, y and/or hue parameters

* **Scatter Plot**

A scatter plot (or ‘scatterplot’) is generally used to summarize the relationship between two paired data samples. Paired data samples mean that two measures were recorded for a given observation, such as the weight and height of a person. The x-axis represents observation values for the first sample, and the y-axis represents the observation values for the second sample. Each point on the plot represents a single observation.

* **Box Plot**

A Box Plot is standardized way of displaying the distribution of data based on five number summaries (“minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”). It can visualize about the outliers and what their values are. It can also tell us if the data is symmetrical, how tightly data is grouped and if and how your data skewed.

* **Bar Chart**

A bar chart is generally used to present relative quantities for multiple categories. The x-axis represents the categories and are spaced evenly. The y-axis represents the quantity for each category and is drawn as a bar from the baseline to the appropriate level on the y-axis. A bar chart can be created by calling the bar () function and passing the category names for the x-axis and the quantities for the y-axis. Bar charts can be useful for comparing multiple point quantities or estimations.

With a knowledge of these plots, you can quickly get a qualitative understanding of most data.

This project helps us to analyze the trends in International football that changed over the years. This project use various key aspects of football results and they are visually presented using graphs, tables and plots.

Data visualization is the representation of information in the form of chart diagram, pictures etc. It can also be used as a reporting tool. Visualizing different type of data and using data to build statistical or machine learning models. Quite simply, this is the must-have reference for scientific computing in python. For visualization we use;

* IPython & Jupyter: Provide Computational environment for data scientists using python
* NumPy: Includes the ndarray for efficient storage and manipulation of dense data arrays in python.
* Pandas: Features the Data Frame for efficient storage and manipulation of labeled columnar data in Python.
* Matplotlib: Includes capabilities for the flexible range of data visualization in Python.

**LITERATURE REVIEW**

‘International Football Results from 1872 to 2019’ is a quite large data set that available in the Kaggle platform. Currently the data set is on its 4th version by Mart Jürisoo. There are more than 60 kernels available that are contributed by various people. Among these kernels the most studied topics are

* Prediction of world cup winner
* Rivalries through pivot tables
* Euro 2020 qualifier predictions
* Trend changing over the years in International football

These topics are studied through various plotting methods available in python libraries like NumPy, matplotlib and seaborn. Most of the kernels are completed and some of them need much more contributions.

As we studied over these kernels we decided to do different analysis on the data set using the plotting available in python

**REQUIREMENT**

**SYSTEM REQUIRMENTS**

* **OS-Ubuntu 18.04**

Linux is a family of free and open-source software operating systems built around the Linux kernel. Ubuntu 18.04 LTS introduced new features such as colour emoji, a new To-Do app preinstalled in the default installation, [and added the option of a "Minimal Install" to the Ubuntu 18.04 LTS installer, which only installs a web browser and system tools.

* **Anaconda**

**Anaconda** is a free and open-source distribution of the Python and R programming languages for data science and machine learning applications (large-scale data processing, predictive analytics, scientific computing), that aims to simplify package management and deployment. Package versions are managed by the package management system *conda* The Anaconda distribution is used by over 6 million users and includes more than 250 popular data-science packages suitable for Windows, Linux, and MacOS.

* **Jupyter notebook**

Jupyter is a non-profit organization created to "develop open-source software, open-standards, and services for interactive computing across dozens of programming languages." Spun-off from IPython in 2014 by Fernando Pérez Project Jupyter supports execution environments in several dozen languages.

* **Pandas library**

In computer programming, **pandas** is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license. The name is derived from the term "panel data", an econometrics term for data sets that include observations over multiple time periods for the same individuals

**DESIGN AND IMPLEMENTATION**

**DATA VISUALIZATION**

# *This python3 environment comes with many helpful analytics libraries installed*

*# It is defined by the kaggle/python docker image: #http://github.com/kaggle/docker-python*

*# for example, here’s several helpful packages to load in*

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

%matplotlib inline

data = pd.read\_csv("C:/Users/Jithu/Desktop/results.csv")

data.head()



1. **Graph representing number of match played in each year:**

**Query:**

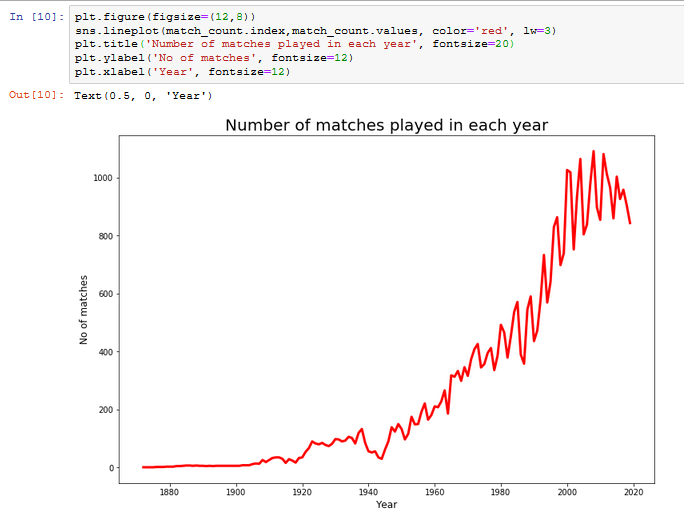
plt.figure(figsize=(12,8))

sns.lineplot(match\_count.index,match\_count.values,color='red', lw=3)

plt.title('Number of matches played in each year', fontsize=20)

plt.ylabel('No of matches', fontsize=12)

plt.xlabel('Year', fontsize=12**)**

****

1. **Graph representing best teams of all time (Best 10 teams):**

**Query:**

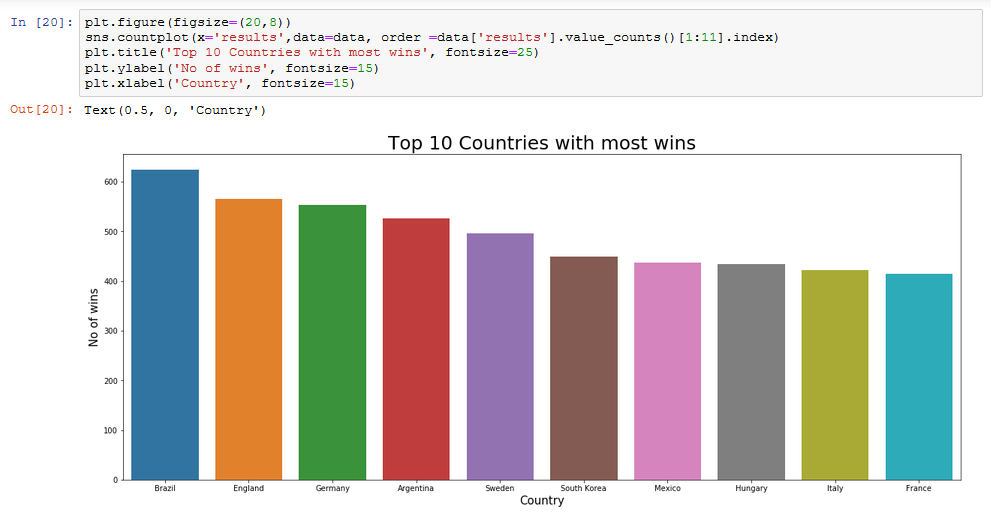
plt.figure(figsize=(20,8))

sns.countplot(x='results',data=data, order =data['results'].value\_counts()[1:11].index)

plt.title('Top 10 Countries with most wins', fontsize=25)

plt.ylabel('No of wins', fontsize=15)

plt.xlabel('Country', fontsize=15)

****

1. **Graph representing Which teams dominated in the different eras of football**

**Query:**

*#The visualization using CountPlot*

for i in range(1870,2019,50):

plt.figure(figsize=(10,4))

newdata= data[(data.year >= i)& (data.year <= i+50)]

sns.countplot(x='results',data=newdata , order =

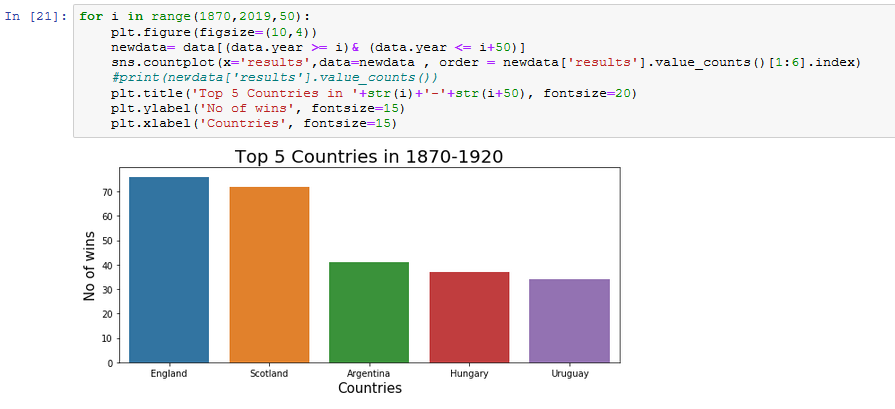
newdata['results'].value\_counts()[1:6].index)

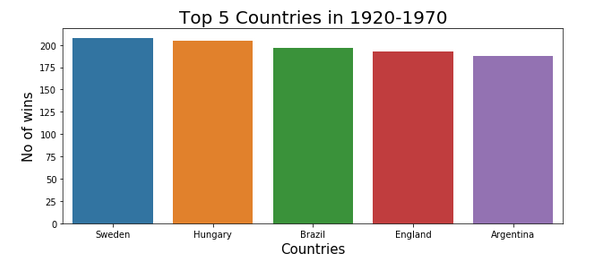
plt.title('Top 5 Countries in '+str(i)+'-'+str(i+50),

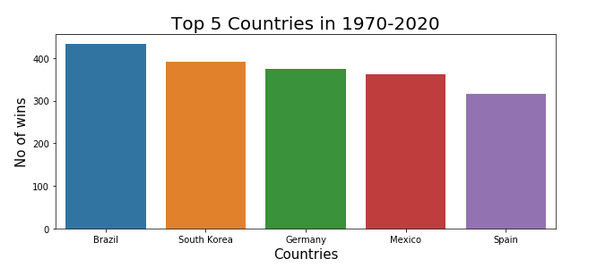
fontsize=20)

plt.ylabel('No of wins', fontsize=15)

plt.xlabel('Countries', fontsize=15)



****

****

**Query:**

*#The visualization using ScatterPlot*

for i in range(1870,2019,50):

plt.figure(figsize=(6,4))

newdata= data[(data.year >= i)& (data.year <= i+50)]

d = newdataset(newdata)

ax=sns.scatterplot(y="noofwins",

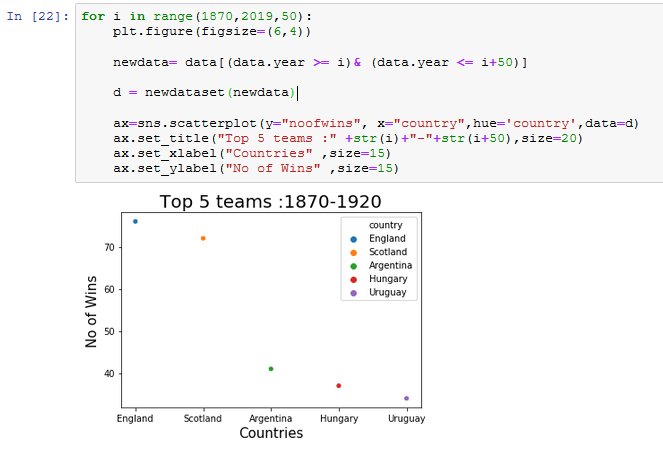
x="country",hue='country',data=d)

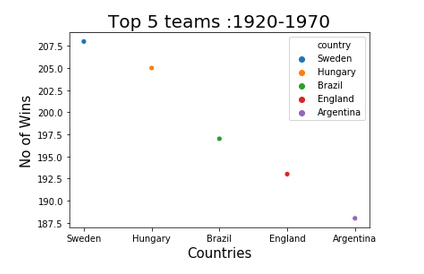
ax.set\_title("Top 5 teams :" +str(i)+"

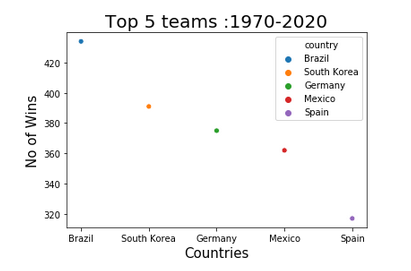
-"+str(i+50),size=20)

ax.set\_xlabel("Countries" ,size=15)

ax.set\_ylabel("No of Wins" ,size=15)







**Query:**

*#The visualization using BoxPlot*

for i in range(1870,2019,50):

plt.figure(figsize=(6,4))

newdata= data[(data.year >= i)& (data.year <=

i+50)]

d = newdataset(newdata)

ax = sns.boxplot(x="country", y="noofwins",

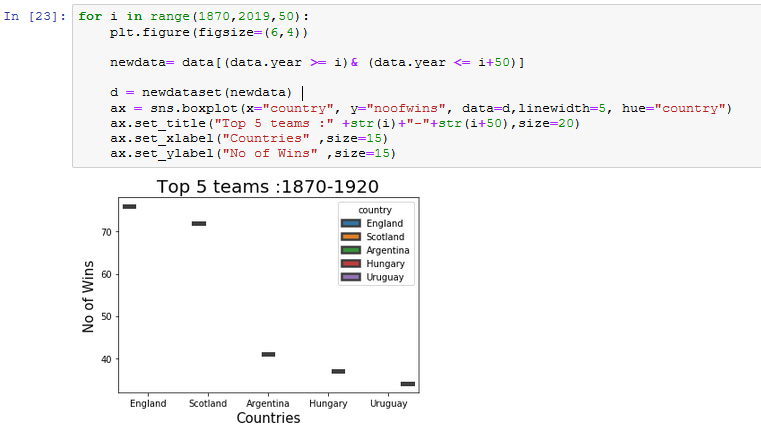
data=d,linewidth=5, hue="country")

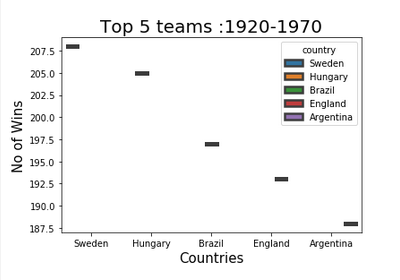
ax.set\_title("Top 5 teams :" +str(i)+"

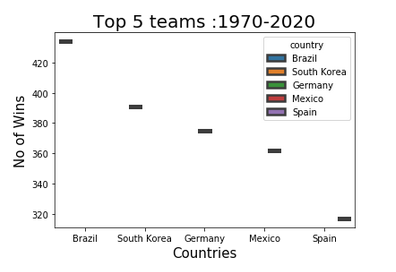
-"+str(i+50),size=20)

ax.set\_xlabel("Countries" ,size=15)

ax.set\_ylabel("No of Wins" ,size=15)







**Query:**

*#The visualization using StripPlot*

for i in range(1870,2019,50):

plt.figure(figsize=(6,4))

newdata= data[(data.year >= i)& (data.year <=

i+50)]

d = newdataset(newdata)

ax = sns. stripplot (x="country", y="noofwins",

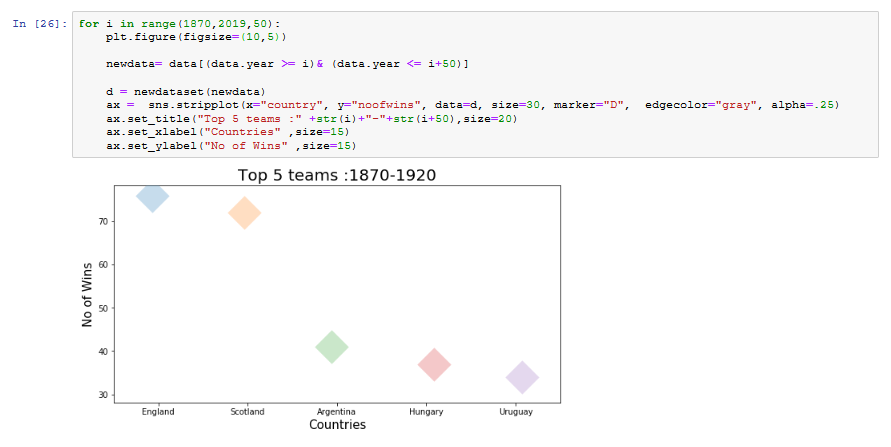
data=d,linewidth=5, hue="country")

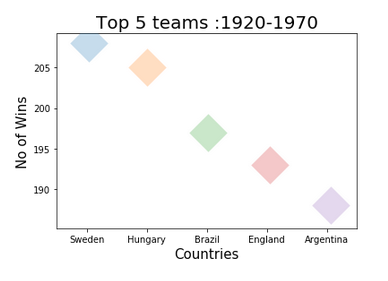
ax.set\_title("Top 5 teams :" +str(i)+"

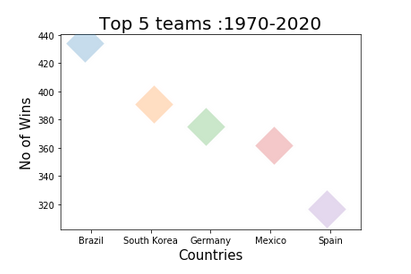
-"+str(i+50),size=20)

ax.set\_xlabel("Countries" ,size=15)

ax.set\_ylabel("No of Wins" ,size=15)







1. **Graph representing What trends have there been in international football throughout the ages – home advantage**

**Query:**

plt.figure(figsize=(7,6))

sns.countplot(x='home\_win',palette='rainbow',hue='neutral',data=data[data.home\_win != 'Tie'])

plt.title('Home vs Away wins', fontsize=20)

plt.ylabel('No of wins', fontsize=12)

plt.xlabel('Home or Away', fontsize=12)

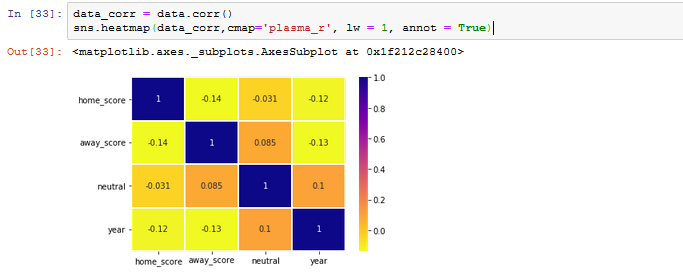
****

1. **Heat Map:** A heat map is a two-dimensional representation of data in which values are represented by colors. A simple heat map provides an immediate visual summary of information. More elaborate heat maps allow the viewer to understand complex data sets.

**Query:**

data\_corr = data.corr()

sns.heatmap(data\_corr,cmap='plasma\_r', lw = 1, annot = True)

****

**RESULT AND ANALYSIS**

* The **Line Plot** is used to visualize the number of matches that played over the years in International football. X axis represent the years and Y axis represent the number of matches. From the plot we can analyze that number of matches played in International football increases as year increases. The key point that we can conclude from this plot is number of matches increases in quick succession at recent centauries.
* The **Count Plot** is used to visualize the top 10 countries in the International football over these years. X axis represents the countries name and Y axis represents the number of wins in matches. From this plot we can analyze that ‘Brazil’ is the top country with most wins in International football and England, Germany, Argentina, Sweden, South Korea, Mexico, Hungary, Italy and France occurring the remaining positions in the graph respectively.
* Another **Count Plot** is used to visualize the countries that have most wins in International football in each era. X axis represent the team names and Y axis represents the no of wins. Each era categorized as 1870 – 1920, 1920 – 1970 and 1970 – 2020 respectively. In each plots England, Sweden and Brazil stands at the first position with the greatest number of wins respectively.
* The **Scatter Plot, Box Plot** and **Strip Plot** are used to analyze the performance of countries in each era and visualize the top 5 countries. In each plots England, Sweden and Brazil stands at the first position with the greatest number of wins respectively
* Another **Count Plot** is used to visualize the home match and away match wins over the years in International football. We also used the neutral venue wins in this plot for better visualizing. As we can analyze that home matches have always the advantage than away matches. And neutral venue also makes advantage to the home team
* The **Heat Map** is used for a comparative study between home score, away score, neutral and year of match. Over the years the home score and away score are fluctuating in a bias while neutral stands in a much steady way.

**CONCLUSION**

We have arrived at the end of our project into the field of data visualization.

In our project, we have presented a number of the most important theoretical and graphs involved in the design of visualization. As we have seen, designing an efficient and effective data visualization from our project. involves representing the data of interest, processing the data to extract relevant information for the problem at hand, designing a mapping of this information to a visual representation, rendering this representation, and combining all this information for an efficient data visualization.

**BIBLIOGRAPHY**

* Kaggle
* Python DataScience Hand Book
* Data visualization(Google) :

1. <https://www.kaggle.com/romabash/titanic-data-visualization>

2. <https://www.kaggle.com/konohayui/titanic-data-visualization-and-models>

3. <www.stacabuse.com/python-for-nlp-sentiment-analysis-with-scikit-learn/>