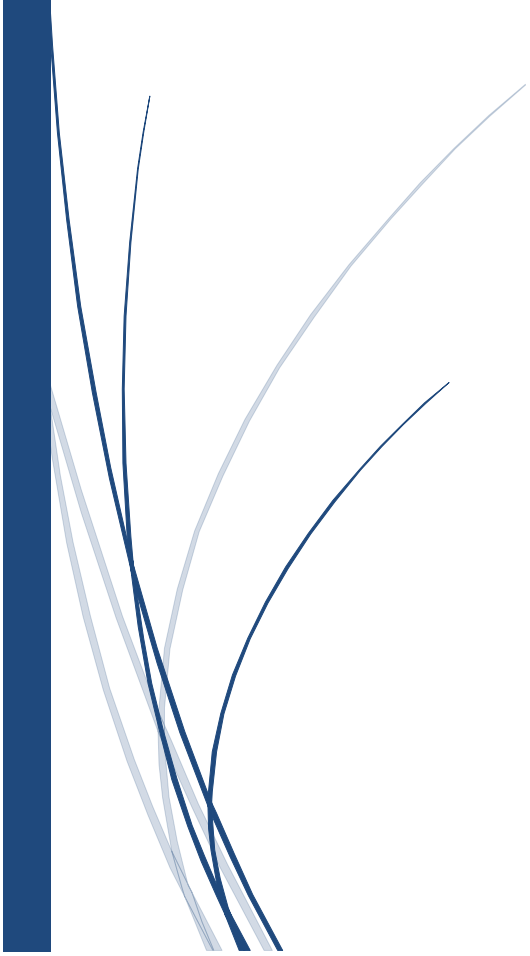
[Date]

CWC-2019 Data Analysis

IT606 – Programming Lab(Python)



Jenil Doshi

202118034

Table Of Contents

1. Problem Definition:

Exploratory Data Analysis And Statistical Analysis With Cricket World-Cup-2019 Dataset With DataCleaning Process.

2. Explanation:

Cricket is a bat and ball game played between two teams, 11 players each, on a field which has a rectangular 22-yard-long pitch in the center. The game is played by 120 million players world-wide making it the second most popular sport in the world. The purpose of the game is to score more runs than your opposing team.

And the Cricket World-Cup-2019 is in few Months.

This Project Aims To Strategize in the Upcoming Cricket World-Cup-2019 and using the data of the 2012-2018 ODI Data. Giving Different Strategies of Pitch, Playing-11, Batsmen, Teams to Target, etc.

3.Work Done In Project:

A) Importing Libraries:

- I have imported Numpy library as np which helps us to working with arrays. It is also useful in area of mathematical and statistical domain.
- I have imported Pandas library as pd which helps us in analysis part. It is comparatively easy to use tool for analysis.
- I have imported Matplotlib as plt which helps us in visualization of data by graphical representation.
- I have imported Seaborn as sns which helps us in getting statistical graphical analysis such as correlation.
- I used Numpy,Pandas,Matplotlib More in comparison to Seaborn.

B) Reading And Understanding Dataset:

- We are reading a .csv files named "Batsmen", "Bowler", "WC_Players", "Ground_Data","ODI_Match_Results" by using pandas pd.read().
- First we see the content of dataset using batsmen.head().
- We see the dimension of dataset using shape function.
- We use dtypes function to see the datatype of the columns.
- By using describe() function we get general statistical overview of the dataset.

C) Data Cleaning

- First we will see how many null elements we have in columns as well as in rows by using isnull() and sum with axis 0 and 1.
- Then we will drop unnecessary columns by drop () function .
- We then see how much percentage nullness columns has so we remove the colums as TDNB and DNB.
- In the end we just remove the duplicate elements by keeping first as preference in drop_duplicate() function.

4. Data Analysis:

1. **Team India's Most Wins**
    - **Code :** We are going to display India's Most Wins Against Countries by value_count() function counting India's Wins.
    - **Result :** India has won most Against Sri-Lanka.

2. **Team India's Most Losses**
    - **Code :** We are going to display India's Most Losses Against Countries by value_count() function counting India's Losses.
    - **Result :** India has lost most Against Australia.

3. **Batsmen Average Higher Than 40**
    - **Code :**We are going to display Batsmen who has the Average Higher than 40 by filtering the data in England Pitches.
    - **Result :** Imam-ul-Haq is the Batsmen with the Highest Average of 117.

4. **Top 5 Batsmen in Strike Rate**
    - **Code :** We are going to display Top-5 Batsmen who has Strike Rate by filtering the data in England Pitches and Using head().
    - **Result :** Evin Lewis has the Highest Strike Rate of 131.57.

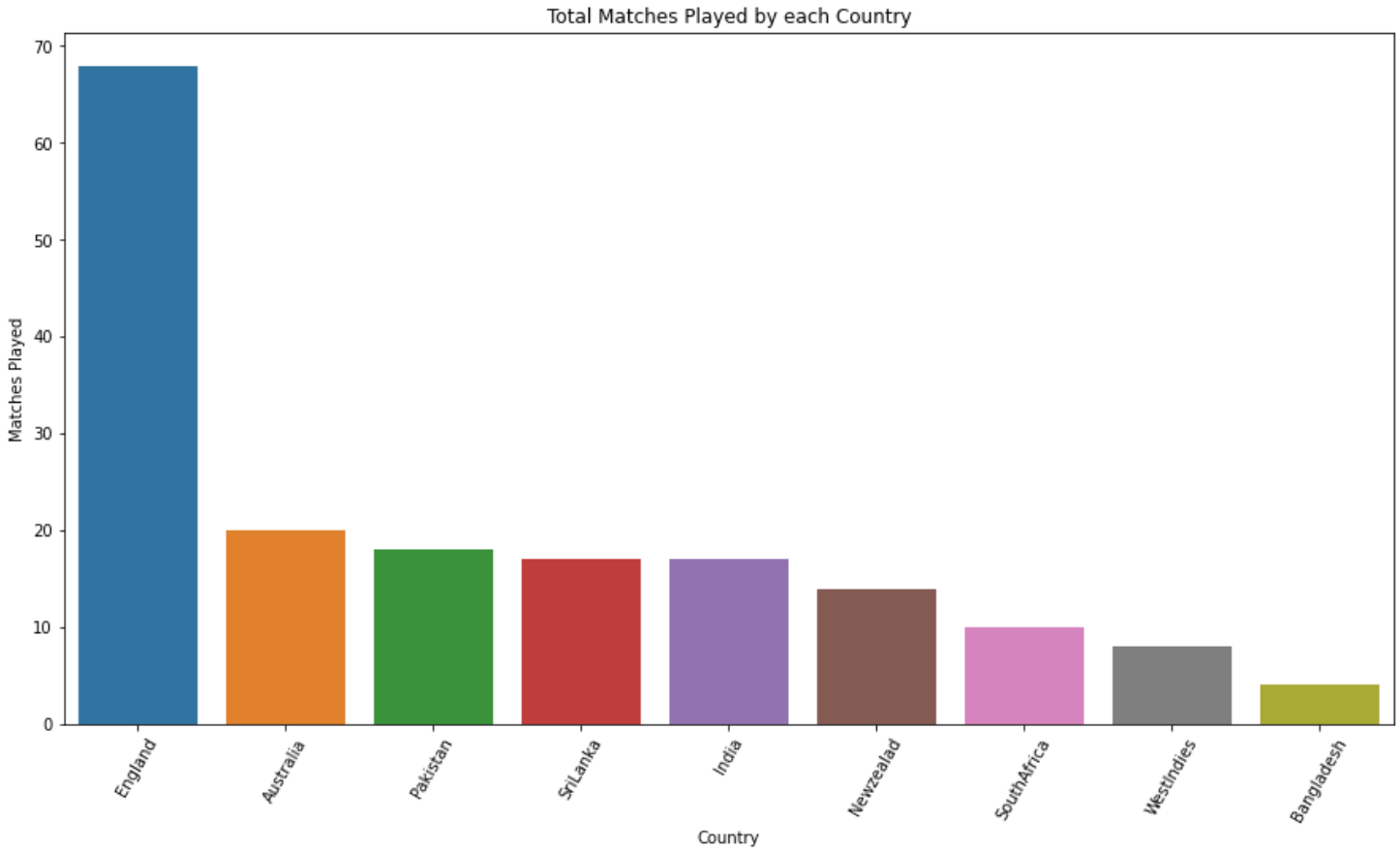
5. **Top 10 Bowlers in Average**
    - **Code :** We are going to display Top-10 Bowlers who has Best Average by filtering the data in England Pitches and Using head().
    - **Result :**Kuldeep Yadav has the Best Average of 16.4.

6. **Top 10 Bowlers in Wickets**
    - **Code :** We are going to display Top-10 Bowlers who has Most Wickets by filtering the data in England Pitches and Using head().
    - **Result :**Adil Rashid has the Most number of Wickets 72.

7. **Top 10 Bowlers in Strike Rate**
    - **Code :** We are going to display Top-10 Bowlers who has Best Strike Rate by filtering the data in England Pitches and Using head().
    - **Result :** Kuldeep Yadav has the Best Strike Rate of 20.

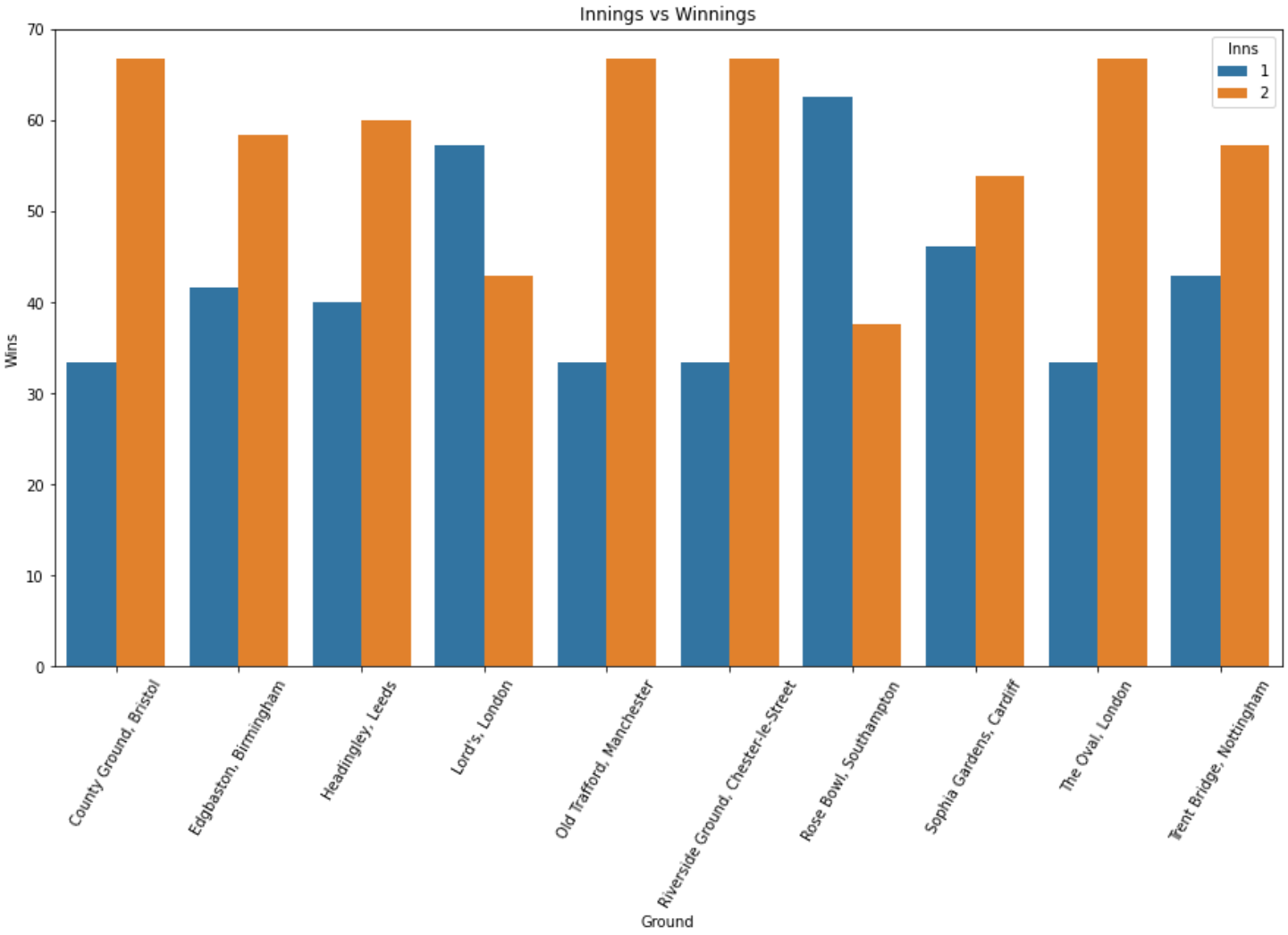
5. Data Visualization

1. **Barplot for Countries Played Matches in England.**
    - **Code :** We use barplot() function by which we will get vertical bar graph with number of matches played by every team in England.



    - **Result :**As England is the Home Team most number of matches would be played by England.

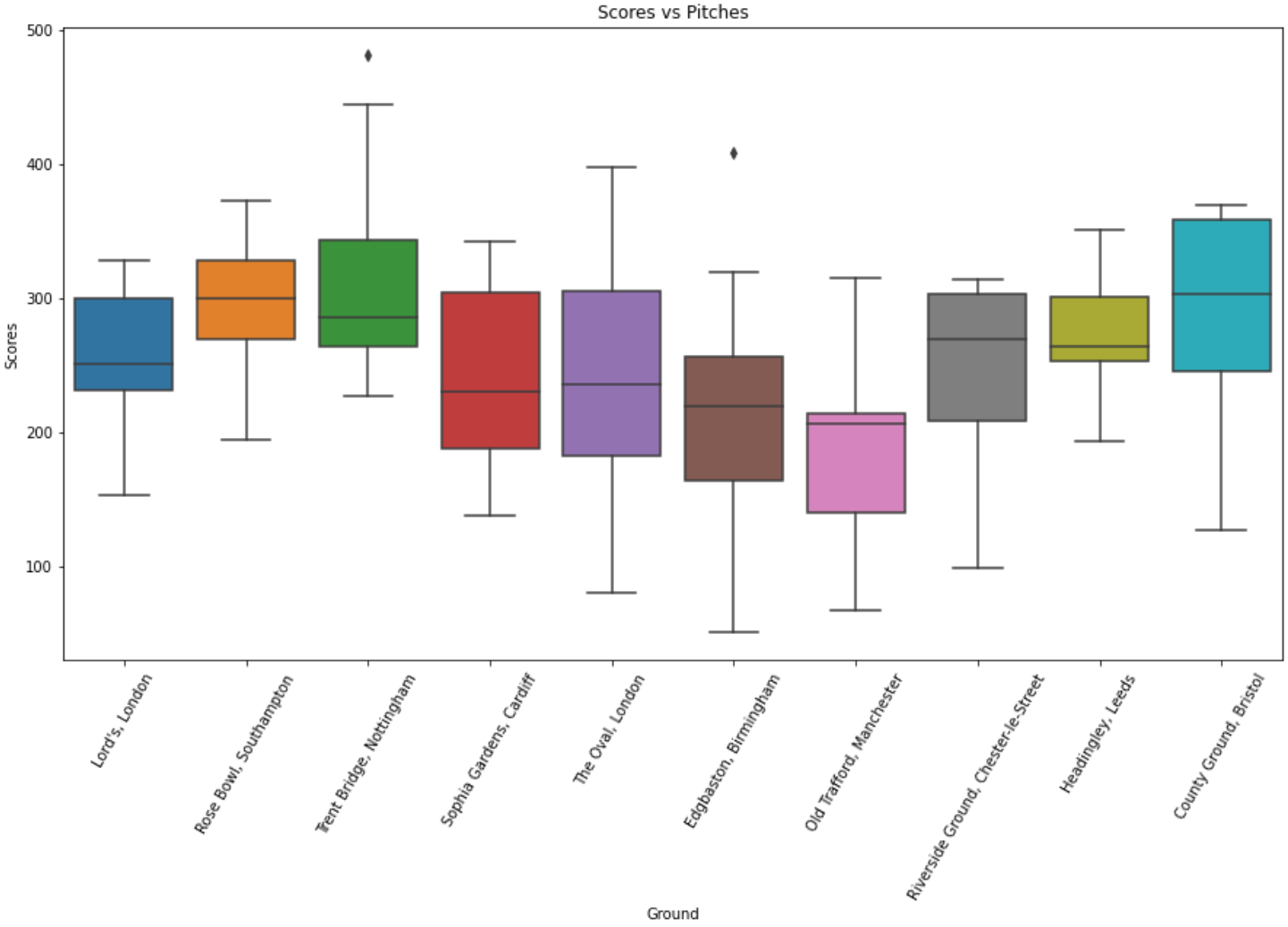
2. **Barplot of Dividing them in pitches and see the Winning Percentage of Innings.**
    - **Code :** We use values_count() and barplot() function by whichwe will plot a bar graph of Innings VS Winnings Regarding Different Pitches of England.



- **Result :** Every Pitch (except Lord's,London and Rose Bowl,Southampton) has a better record in winning the game by bowling first.

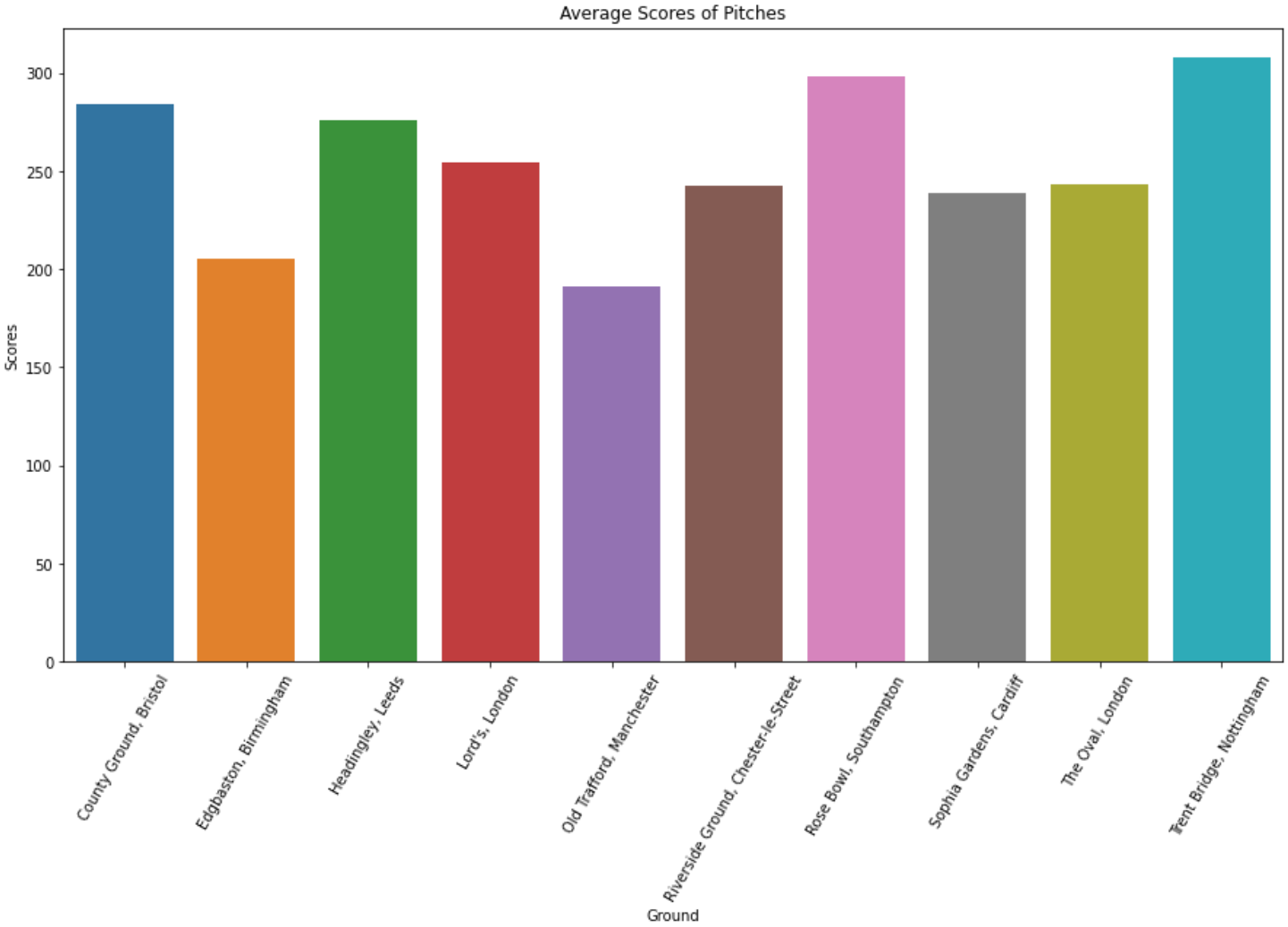
3. Boxplot of Scores Against Pitches

- **Code :** We use boxplot() function by which wewill plot a box graph of Scores Regarding Pitches where the range will appear and if you bowl/bat first you would know at which total you should restrict/score against a Team.



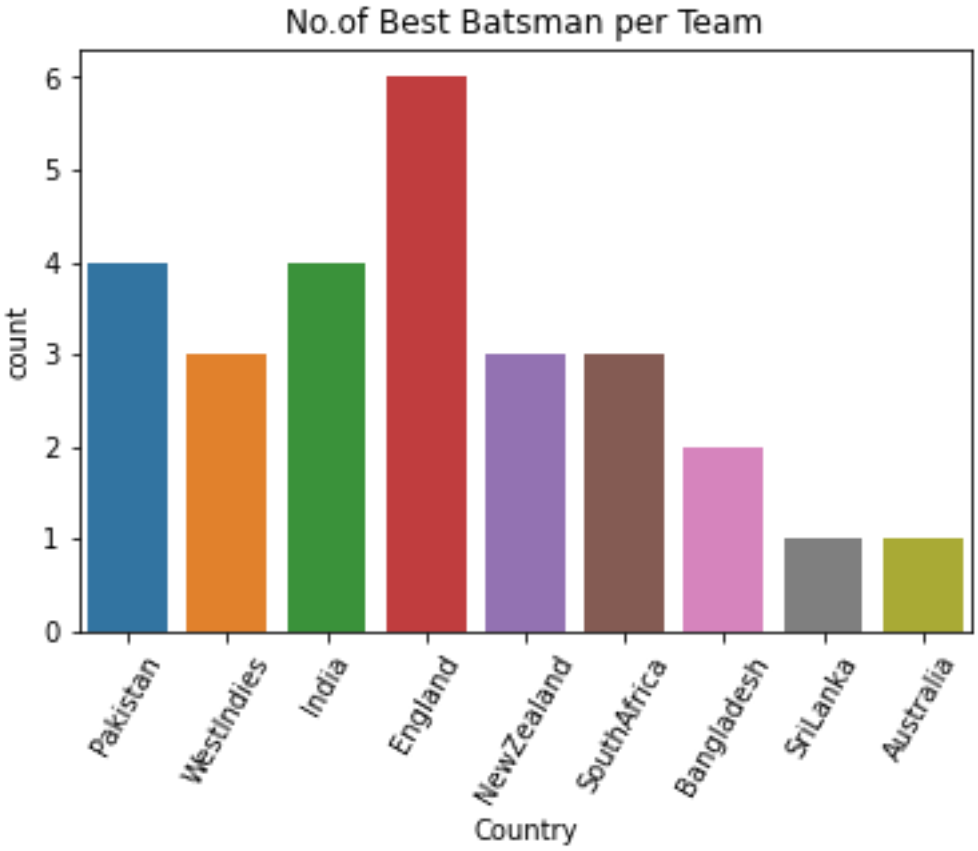
- **Result :** Trent Bridge, Nottingham, County Ground, Bristol and Headingley, Leeds are going to be high scoring matches and Old Trafford, Manchester ,Sophia Gardens, Cardiff ,Edgbaston, Birminghim are low scoring matches.

4. **Barplot of Average Scores in Pitches.**
    - **Code :** We use groupby() , sort_values() and barplot() to plot the average scores of pitches with respect to the grounds in england.



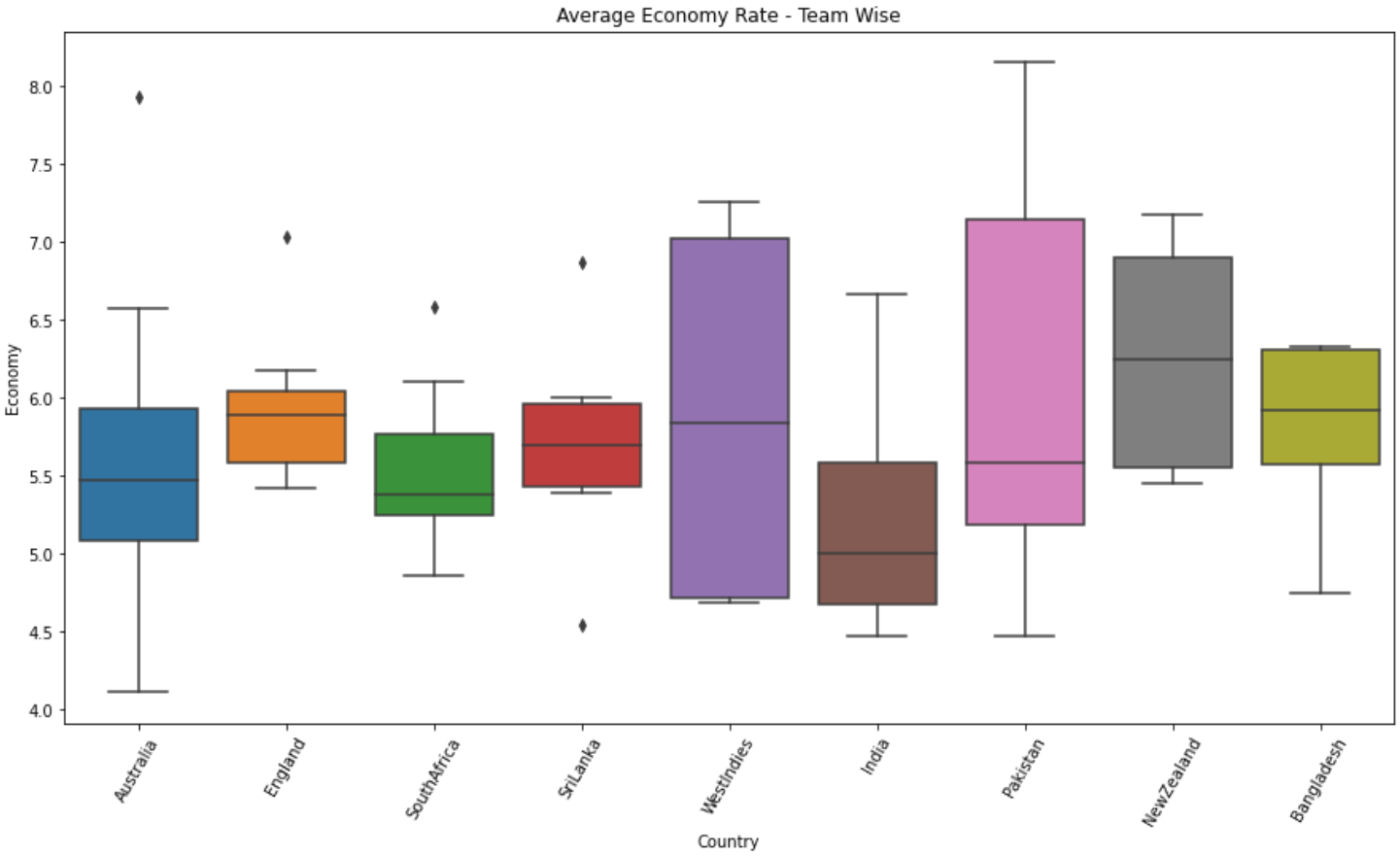
- **Result :** Most pitches are high scoring grounds so you should carry an extra bowler if you know bowlers are going have a bad day.

5. **Countplot of Number of Best Batsmen in a Team**
    - **Code :**We filter the data with batsmen having average greater than 40 and having a good stike rate by using count() and countplox() we can differentiate them in their respective countries.



- **Result :**England has most number of best batsmen.

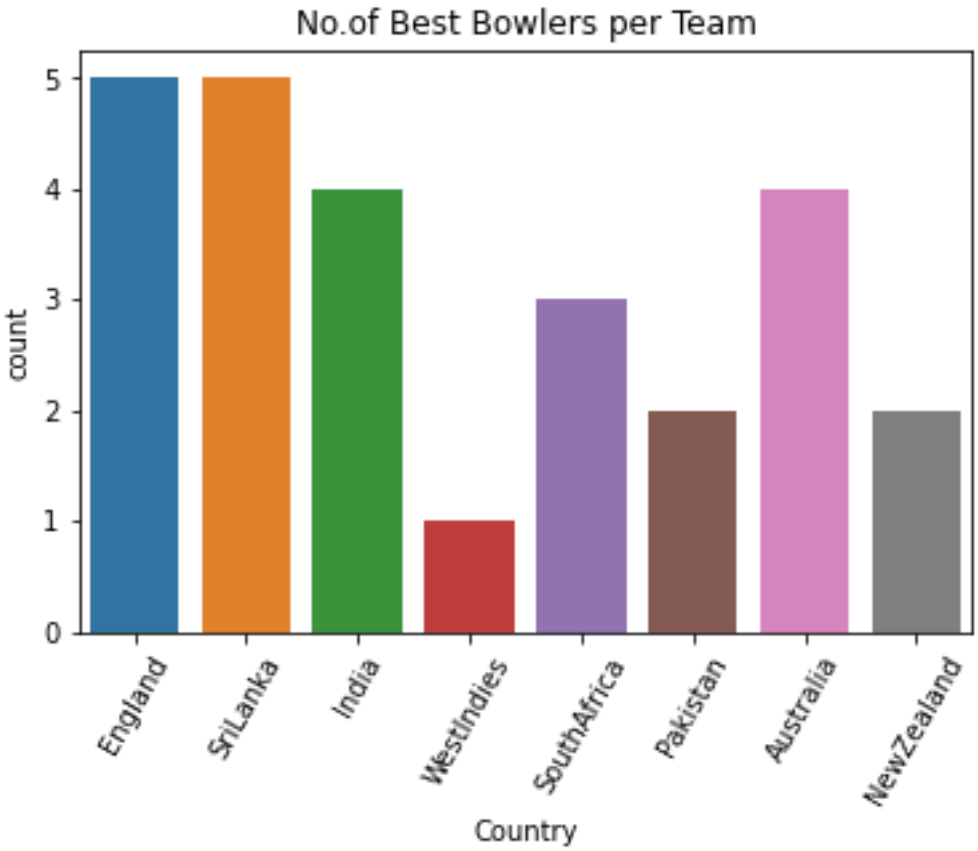
6. **Boxplot of Average Economy Regarding Teams.**
    - **Code:** We plot a boxplot of Average Economy Rate of every team where we know which teams should be targeted.



Result : Batsmen can target High Economy Rate Teams like Pakistan, New Zealand, West Indies and India, South Africa, Australia have good Economy Rates.

- .

7. **Countplot of Number of Best Bowler in a Team**
    - **Code :** We filter the data with bowler having average lower than 40 , stike rate lower than 40 and economy lower than 7.5 by using count() and countplox() we can differentiate them in their respective countries.
    -



- **Result :**England and Sri-Lanka has the Most number of Best Bowlers in English Conditions.

6.Conclusion and future work:

- The analysis done in the project is to brought out details in every possible scenario to provide useful information regarding the CWC-2019 Strategies.
- The quantity and range of the ODI Data is large. As analyst we have too many choices to choose which one is suitable for your interest.
- By the analysis we can also predict which team can win the worldcup-2019.
- We could do some analysis for All-Rounders also.
- We can also find the data of a team regarding their batting(Average Score,Run-Rate).

7.Learning from project:

- Learning the python is the first step toward the world of data analysis.
- There is so many library such as numpy, pandas for getting details of your data in addition we have matplotlib , seaborn to visualize the data in specific manner.
- This project definitely helps me to get out of my comfort and push the programming skills to another level.

8. Bibliography :

https://pandas.pydata.org/docs/user_guide/index.html

https://www.w3schools.com/python/numpy/numpy_intro.asp

https://matplotlib.org/

https://seaborn.pydata.org/