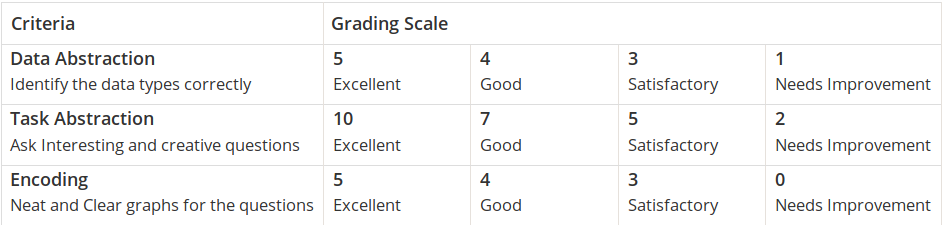
**J COMPONENT REVIEW -1 DOCUMENT**

**NAME – ADITYA DEWANI**

**REG NO – 16BCE2319**

**SLOT-C2**

****

**DATASET :**

**Fifa world cup data from 1930 to 2014**

**WorldCups.csv**

**WorldCupMatches.csv**

**WorldCupPlayers.csv**

**DATA ABSTRACTION**

**Data types:**

There are certain categorical data as

1. country

2.winners

3.runnerups

4.team name

5.player name

6.city name

There are certain numerical data : (All are discrete data).

1. year

2. goals scored

3. match id

4. attendance

5.match played

**( Reduction of dataset to useful data set )**

**# taking dataset into variables :**

matches = pd.read\_csv(r"WorldCupMatches.csv")

cups = pd.read\_csv(r"WorldCups.csv")

**# calculating null values :**

matches.isnull().sum()

**# setting the style of plots :**

sns.set\_style("darkgrid")

**# removing duplicate rows for match id coloumn :**

matches = matches.drop\_duplicates(subset="MatchID",keep="first")

**# removing null values :**

matches = matches[matches["Year"].notnull()]

**# using only the coloumns that are needed and removing the other coloumns :**

new\_matches=pd.read\_csv("WorldCupMatches.csv", usecols=['Year','Datetime','Stadium','City','Home Team Name','Home Team Goals','Away Team Goals','Away Team Name','Attendance','MatchID','Home Team Initials','Away Team Initials'])

**# no data abstraction in WorlsCups dataset as all the coloumns are nessesary and all rows have values**

**# WorlsCupPlayers Dataset**

**# reading dataset into variables :**

players= pd.read\_csv("WorldCupPlayers.csv")

# **using only the necessary coloumns :**

new\_players=pd.read\_csv("WorldCupPlayers.csv", usecols= ['Team Initials','Coach Name','Player Name','Position'])

**TASK ABSTARCION :**

The main task of this project is to visualise and interpret the dataset of football worldcup from 1930 to 2014 and hence deduce the change in trends of various attributes .

Various tasks have been identified and the data is visualised accordingly.

1. What is the total no of goals scored in respective years . Does it follow any increasing or decreasing trend ?

2. What is the trend in the attendence of the people watching the worldcup in stadium ?

3. Who are the teams with most number of victories ?

4. How many times a particular country has been winner or runner up (top 10 countries) ?

5. Which team has played most matches till now ?

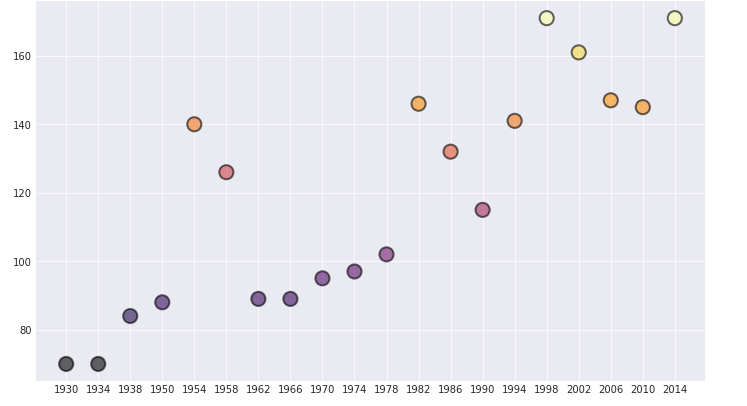
6. What is the outcome of Win for home team and away team ?

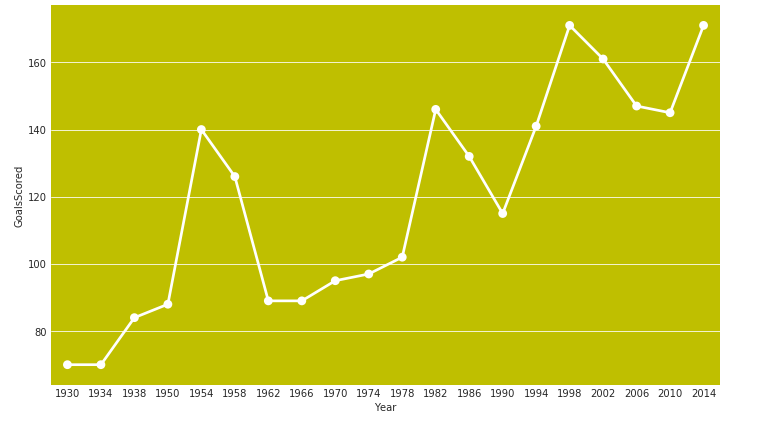
7. Which team has highest number of Fifa Goals ?

8. Interaction between teams for a particular year ?

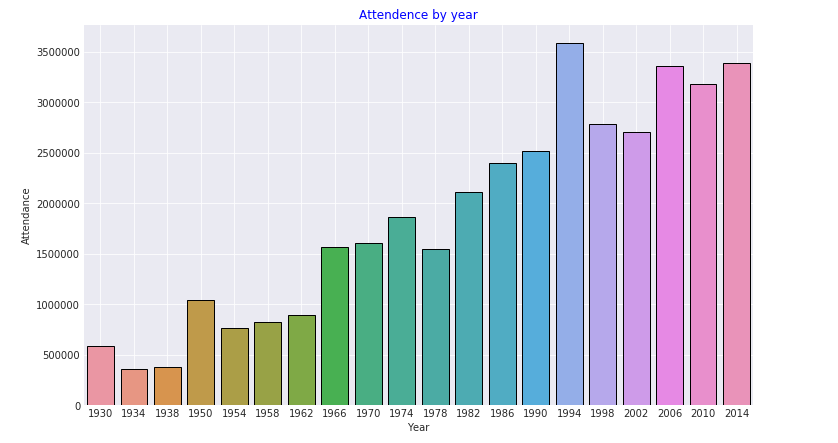
Answers to above questions by visulaisation :

1.





2.



**( Reduction of dataset to useful data set )**

**# taking dataset into variables :**

matches = pd.read\_csv(r"WorldCupMatches.csv")

cups = pd.read\_csv(r"WorldCups.csv")

**# calculating null values :**

matches.isnull().sum()

**# setting the style of plots :**

sns.set\_style("darkgrid")

**# removing duplicate rows for match id coloumn :**

matches = matches.drop\_duplicates(subset="MatchID",keep="first")

**# removing null values :**

matches = matches[matches["Year"].notnull()]

**# using only the coloumns that are needed and removing the other coloumns :**

new\_matches=pd.read\_csv("WorldCupMatches.csv", usecols=['Year','Datetime','Stadium','City','Home Team Name','Home Team Goals','Away Team Goals','Away Team Name','Attendance','MatchID','Home Team Initials','Away Team Initials'])

**# no data abstraction in WorlsCups dataset as all the coloumns are nessesary and all rows have values**

**# WorlsCupPlayers Dataset**

**# reading dataset into variables :**

players= pd.read\_csv("WorldCupPlayers.csv")

# **using only the necessary coloumns :**

new\_players=pd.read\_csv("WorldCupPlayers.csv", usecols= ['Team Initials','Coach Name','Player Name','Position'])