DATA SCIENCE PROJECT

Topic: Olympic Games Analysis



**Points included in Project:**

**1.Reading DataSet**

**2.Passing the object in dataset**

**3.Printing dataset**

**4.Check/count missing or Null values from the dataset**

**5.If any null values exits then filling them.**

**6.Performing data visualization .**

**1.Data Visvualiztion Libraries Used:**

**The data visvualization library used in the project are matplotlib,seaborn and pandas.**

**1)Matplotlib:**

Matplotlib is a Python 2D plotting library for data visualization built on Numpy arrays and designed to work with the broader SciPy stack. It produces quality figures in a variety of formats and interactive environments across platforms.Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib uses an object oriented API to embed plots in Python applications.

**2)Seaborn**

Seaborn is a library for making statistical graphics in Python. It builds on top of [matplotlib](https://matplotlib.org/) and integrates closely with [pandas](https://pandas.pydata.org/) data structures.

**3)Pandas**

[Pandas](http://pandas.pydata.org/) is a Python library for data analysis.

Pandas makes it simple to do many of the time consuming, repetitive tasks associated with working with data, including:

* Data cleansing
* Data fill
* Data normalization
* Merges and joins
* Data visualization
* Statistical analysis
* Data inspection
* Loading and saving data
* And much more

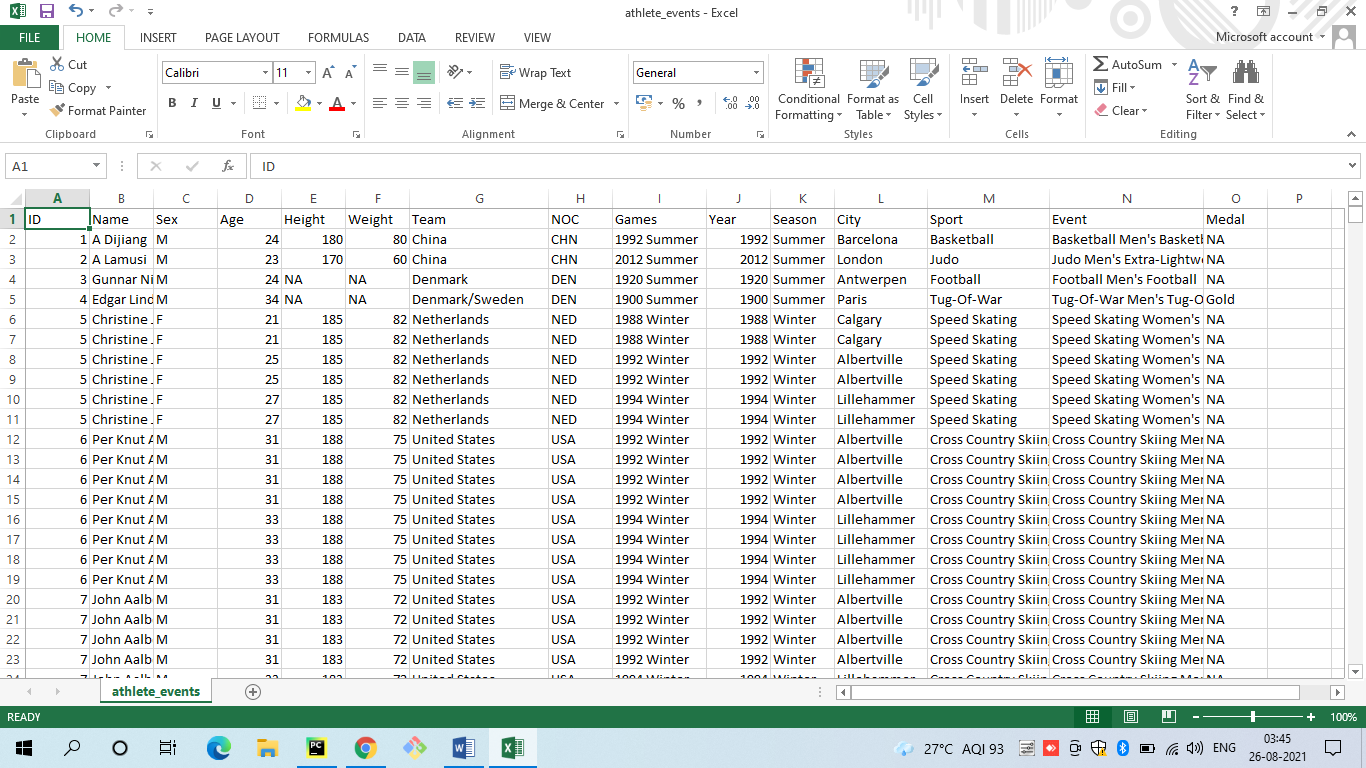
**2.Dataset Used:**

There are two data sets used in the project. Both the datasets give information regarding the participants in the Olympic games till the year 2020.The datasets are through Kaggle.com.

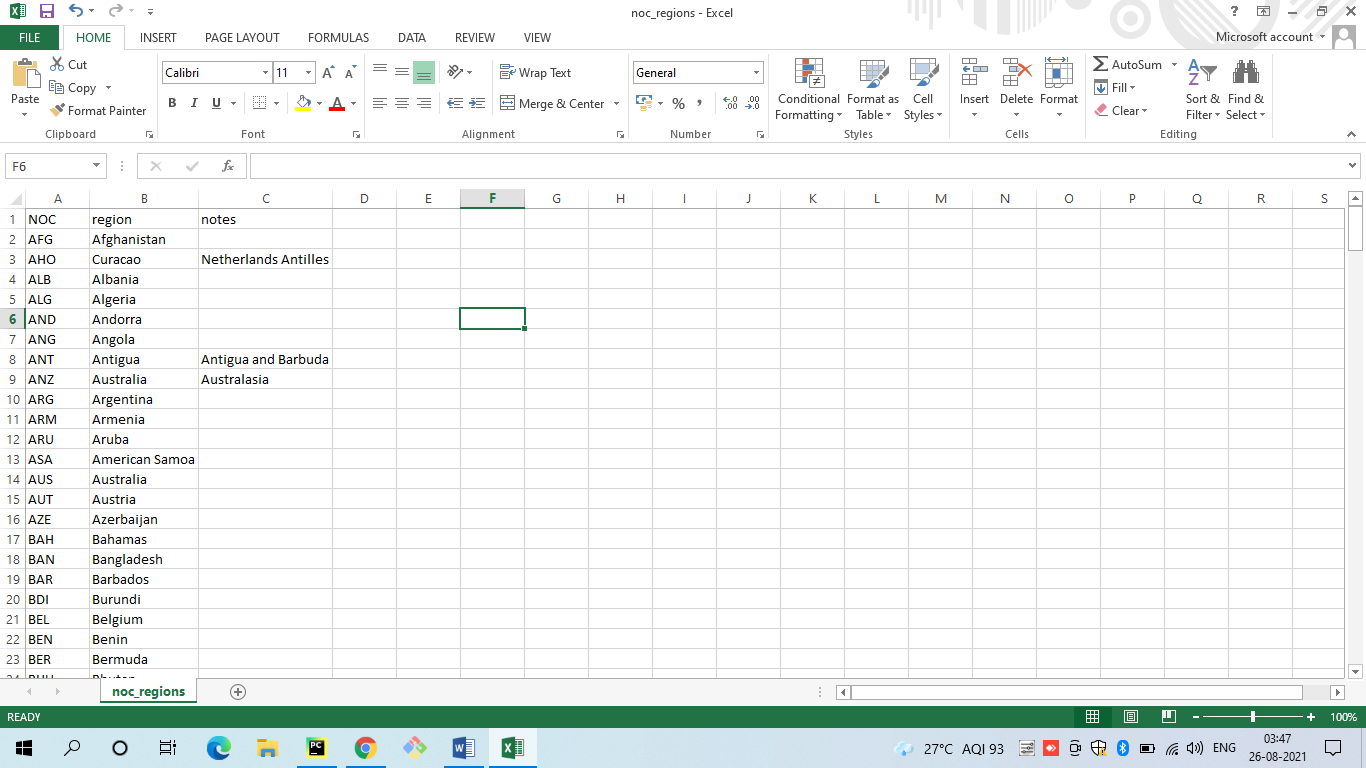
Dataset\_Link: <https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results>

Both the datasets were present in a zip file which were later extracted and copied to the folder containing the projects python file.

**Dataset athlete\_events:**



**Dataset regions**:

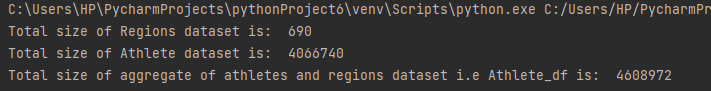


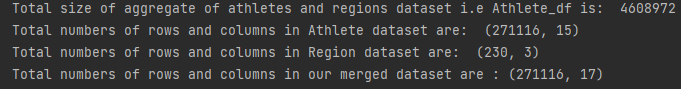
**4.Code and Outputs:**

import numpy as np  
import pandas as pd  
import matplotlib.pyplot as plt  
import seaborn as sns  
#importing all the required libraries  
  
athletes = pd.read\_csv('C:/Users/HP/Desktop/Project/athlete\_events.csv')  
regions = pd.read\_csv('C:/Users/HP/Desktop/Project/noc\_regions.csv')  
  
athletes\_df = athletes.merge(regions,how='left',on='NOC')#merging both the datasets since they have common column NOC  
#print(athletes\_df.head())  
begin=athletes\_df.head()  
#Size of the dataset!!  
print("Total size of Regions dataset is: ",regions.size)  
print("Total size of Athlete dataset is: ",athletes.size)  
print("Total size of aggregate of athletes and regions dataset i.e Athlete\_df is: ",athletes\_df.size)  
#Shape of the dataset!!  
print("Total numbers of rows and columns in Athlete dataset are: ",athletes.shape)  
print("Total numbers of rows and columns in Region dataset are: ",regions.shape)  
print("Total numbers of rows and columns in our merged dataset are :",athletes\_df.shape)  
#Shape of the dataset!!  
print("The null values in datasets are:",athletes\_df.isnull().sum())  
#Dealing with null value of column "Age"  
print("The null value in Age column are:",athletes\_df.Age.isnull())  
print("The null value in Age column are:",athletes\_df["Age"].fillna(athletes\_df.Age.mode()).head(20))  
#Dealing with null value of column "Weight"  
print("The null value in Weight column are:",athletes\_df.Weight.isnull())  
print("After filling the age columns null values with average of age:",athletes\_df["Weight"].fillna(athletes\_df.Weight.mean()))  
#Dealing with null value of column "Height"  
print(athletes\_df.Height.head(20))  
print("The null value in Height column are:",athletes\_df.Height.isnull())  
print("After filling the Height column :",athletes\_df["Height"].fillna(athletes\_df.Height.mean()).head(20))  
#Dealing with null value of column "Medal"  
print("The null value in Medal column are:",athletes\_df.Medal.isnull())  
print("The null value after filling with padding method in Medal column are:",athletes\_df["Medal"].fillna(method ='pad').head(20))  
#Dealing with null value of column "Region"  
print("The null value in region column are:",athletes\_df.region.isnull())  
print("After filling the region column :",athletes\_df["region"].fillna(athletes\_df.region.mode()).head(20))  
#Dealing with null value of column "Notes"  
print("The null value in notes column are:",athletes\_df.notes.isnull())  
print("After filling the notes column :",athletes\_df["notes"].fillna(0).head(20))

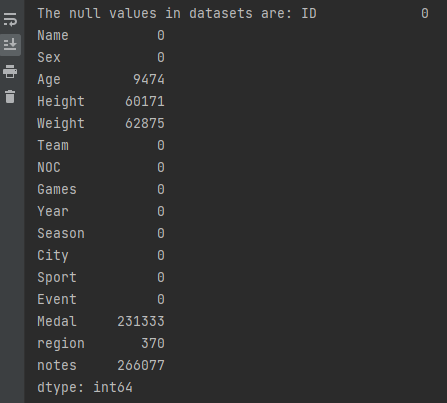
**=>**

**Size:**

**Shape:**

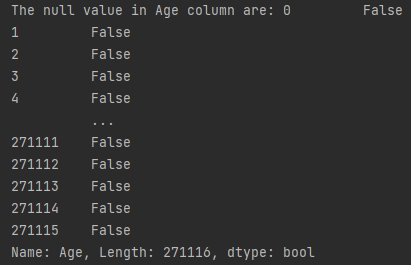
****

**Counting Null value:**

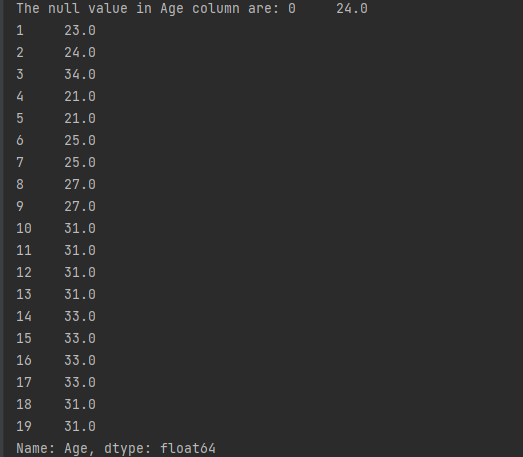
****

**(Here we can see that in the whole dataset**

**The column of age, weight,height,Medal,region ,notes is having null values)**

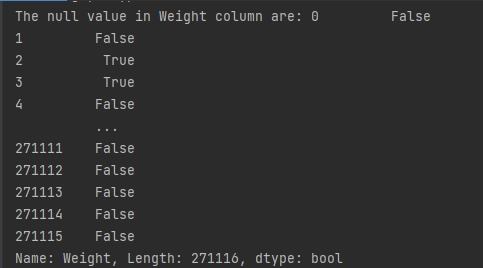
****

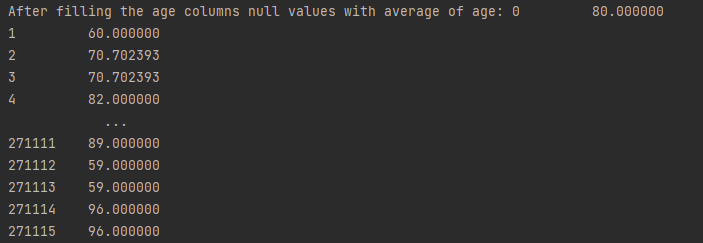
**After filling values null values in Age Column**

****

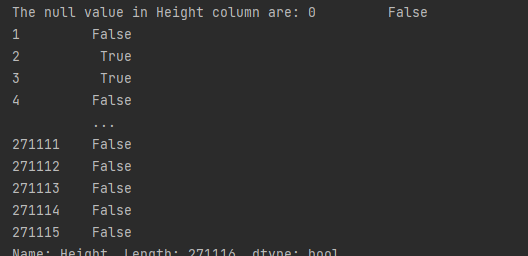
**(only the starting 20 values are printed to see better excecution of fillna method)**

**Checking null values of Weight Column**

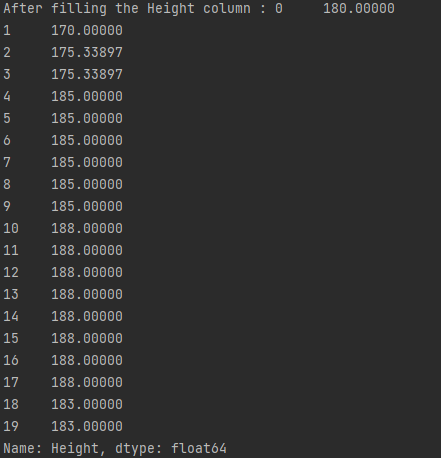
****

**After filling values null values in Weight Column **

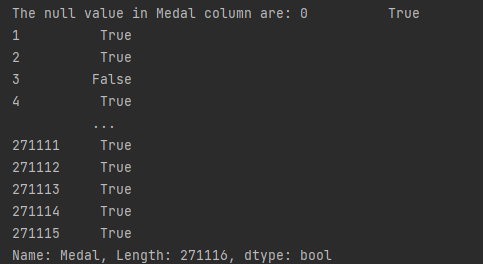
**Checking null values of Height Column**

****

**After filling values null values in Height Column**

****

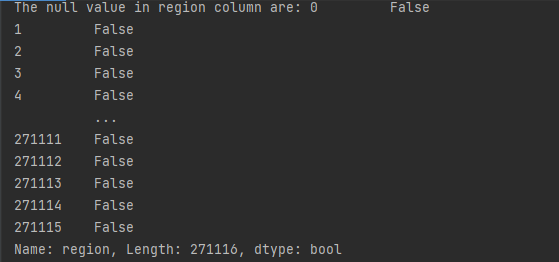
**Checking null values of Medal Column**

****

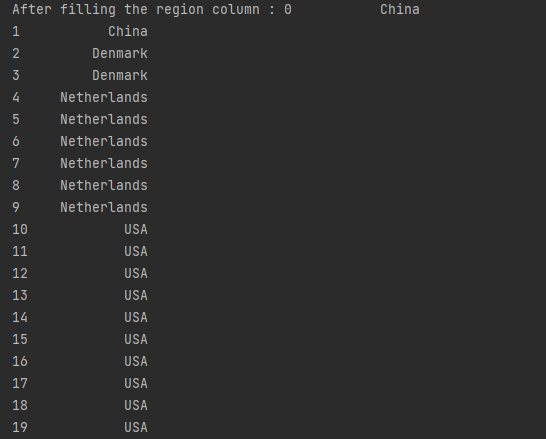
**After filling values null values in Medal Column**

****

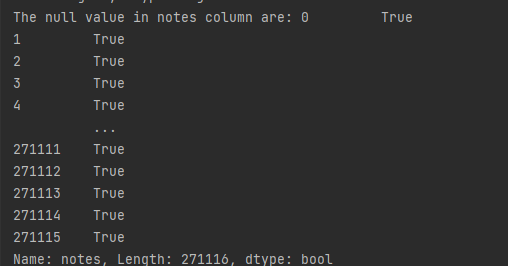
**Checking null values of Region Column**

****

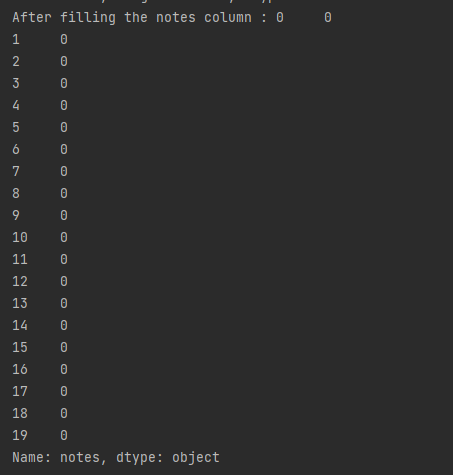
**After filling values null values in Region Column**

****

**Checking null values of Notes Column**

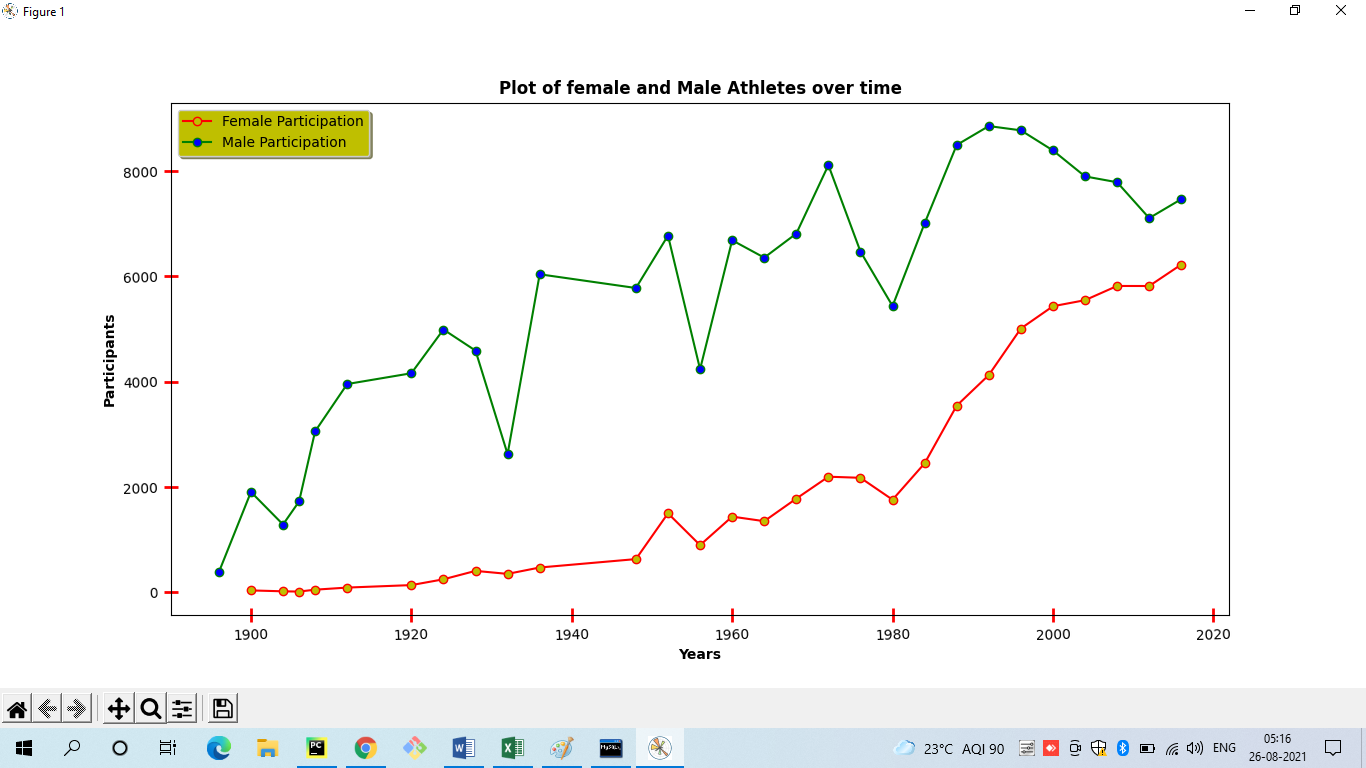
****

**After filling values null values in Medal Column**

****

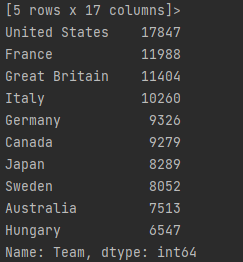
Womply=athletes\_df[(athletes\_df.Sex == 'F') & (athletes\_df.Season == 'Summer')]  
Menply=athletes\_df[(athletes\_df.Sex == 'M') & (athletes\_df.Season == 'Summer')]  
Men = Menply.groupby('Year')['Sex'].value\_counts()  
part = Womply.groupby('Year')['Sex'].value\_counts()  
plt.figure(figsize=(12,6))  
part.loc[:,'F'].plot(label="Female Participation",color="r" ,marker="o",markerfacecolor='y')  
Men.loc[:,'M'].plot(label="Male Participation",color="g" ,marker="o",markerfacecolor='b')  
plt.ylabel("Participants",fontweight = 'bold')  
plt.xlabel("Years",fontweight = 'bold')  
plt.title("Plot of female and Male Athletes over time",fontweight = 'bold')  
plt.tick\_params(axis = 'both',width=2,length = 10, direction ='inout',color ='r',pad =4 ,rotation = 2)  
plt.legend(facecolor="y",shadow=True)  
plt.show()

**=>**

****

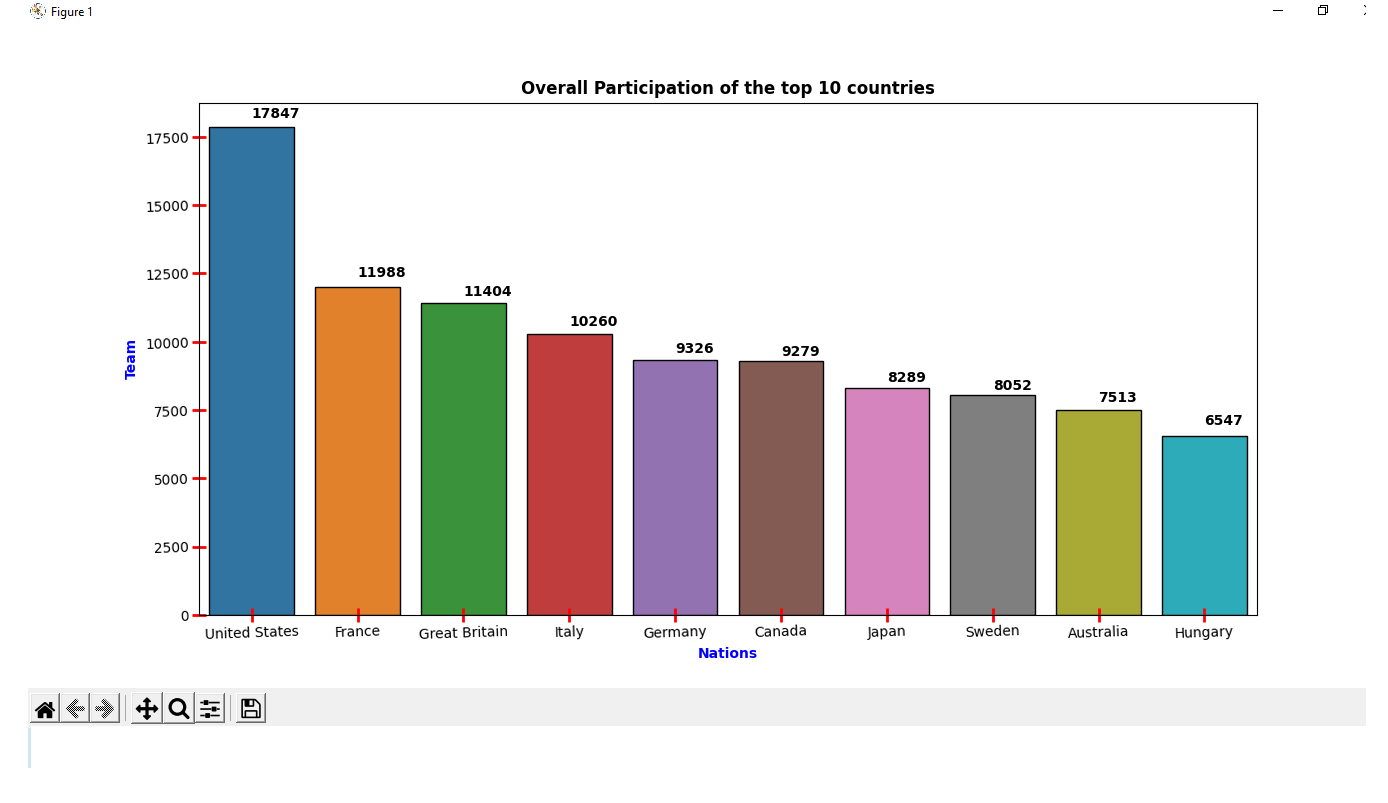
top\_10\_countries= athletes\_df.Team.value\_counts().sort\_values(ascending=False).head(10)#most number of participants participating  
print(top\_10\_countries)

**=>**

****

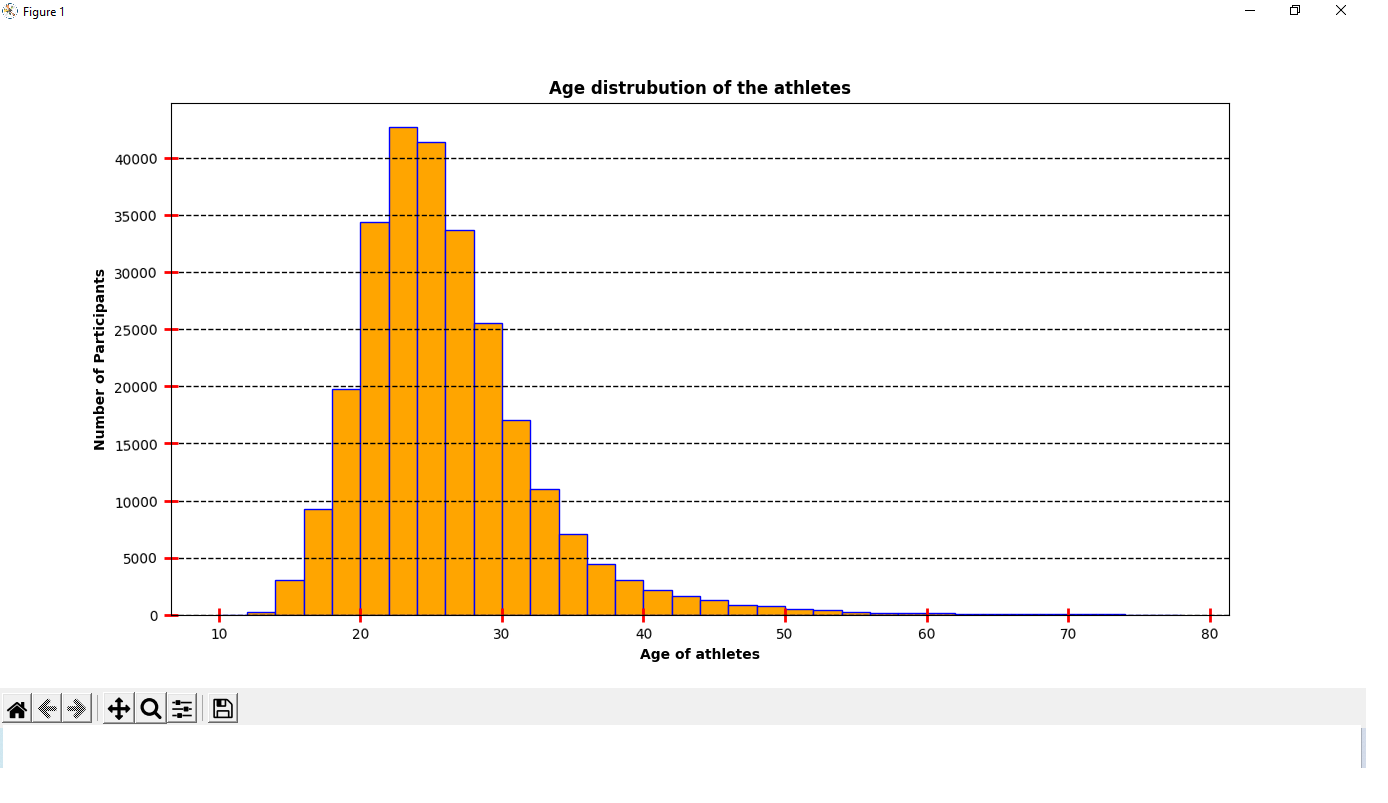
plt.figure(figsize=(12,6))  
plt.tick\_params(axis ='both',length = 10, direction='inout',color ='r',pad =2 ,rotation = 2,width=2)  
plt.xlabel('Nations',color='blue',fontweight = 'bold')  
plt.ylabel('Participation',color='blue',fontweight = 'bold')  
plt.title("Overall Participation of the top 10 countries",fontweight = 'bold')  
value1= plt.annotate('17847',xy=(0,1.793),xytext=(0,1.818e+04),fontsize=10,fontweight = 'bold')  
value2= plt.annotate('11988',xy=(9,6.76),xytext=(1,1.236e+04),fontsize=10,fontweight = 'bold')  
value3= plt.annotate('11404',xy=(2,1.06),xytext=(2,1.167e+04),fontsize=10,fontweight = 'bold')  
value4= plt.annotate('10260',xy=(3,2.57),xytext=(3,1.057e+04),fontsize=10,fontweight = 'bold')  
value5= plt.annotate('9326',xy=(4,2.54),xytext=(4,9.58e+03),fontsize=10,fontweight = 'bold')  
value6= plt.annotate('9279',xy=(5,0.62),xytext=(5,9.47e+03),fontsize=10,fontweight = 'bold')  
value7= plt.annotate('8289',xy=(6,-2.79),xytext=(6,8.52e+03),fontsize=10,fontweight = 'bold')  
value8= plt.annotate('8052',xy=(1,-0.81),xytext=(7,8.23e+03),fontsize=10,fontweight = 'bold')  
value9= plt.annotate('7513',xy=(2,-0.58),xytext=(8,7.79e+03),fontsize=10,fontweight = 'bold')  
value10= plt.annotate('6547',xy=(2,-0.58),xytext=(9,6.95e+03),fontsize=10,fontweight = 'bold')  
#plt.grid(axis='y',color = 'black', linestyle = '--', linewidth = 1)  
plt.tick\_params(axis = 'both',width=2,length = 10, direction ='inout',color ='g',pad =4 ,rotation = 2)  
sns.barplot(x=top\_10\_countries.index,y=top\_10\_countries,edgecolor ='black')  
plt.show()

**=>**



#histogram for age of athletes  
plt.figure(figsize=(12,6))  
plt.title("Age distrubution of the athletes",fontweight = 'bold')  
plt.xlabel('Age of athletes',fontweight = 'bold')  
plt.ylabel('Number of Participants',fontweight = 'bold')  
plt.grid(axis='y',color = 'black', linestyle = '--', linewidth = 1)  
plt.hist(athletes\_df.Age, bins=np.arange(10,80,2) ,color='orange',edgecolor ='blue')  
plt.tick\_params(axis = 'both',width=2,length = 10, direction ='inout',color ='r',pad =4 ,rotation = 2)  
plt.show()

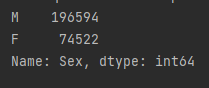
**=>**



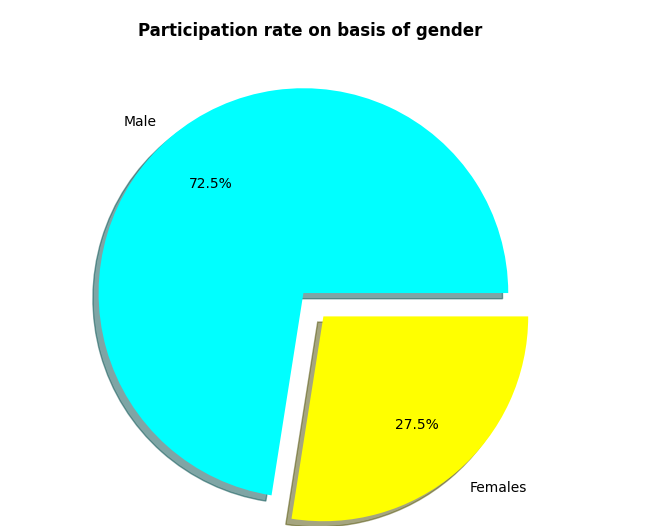
winter\_sports=athletes\_df[athletes\_df.Season=='Winter'].Sport.unique()#So olympics has winter olympic and summer olympic  
#so to mention particular games in winter olympics we use this  
print(winter\_sports)  
  
summer\_sports=athletes\_df[athletes\_df.Season=='Summer'].Sport.unique()  
print(summer\_sports)

**=> **

gender\_counts= athletes\_df.Sex.value\_counts()#to count total number of Male athletes and female athletes  
print(gender\_counts)

🡺****

plt.figure(figsize=(12,6))  
mylabels=['Male','Females']  
col=['cyan','yellow']  
myexplode = [0.05,0.1]  
plt.title("Participation rate on basis of gender",fontweight = 'bold')  
plt.pie(gender\_counts,labels=mylabels,autopct=('%1.1f%%'),startangle=0,colors=col,center=(0,0),  
 rotatelabels=False,pctdistance=0.7,shadow=True,explode=myexplode)  
plt.show()

****

**DataVisualization conclusion**

**1)Line Graph:**

The line graphs indicates the participation number of Male in Female athletes in the Olympic Games since 1900 till the year 2020. One can see that the participation of both, the male and female athletes has increased since 1900. Also one can see that the rate of male participation is considerably more than the female participants.

**2)Pie plot:**

The pie chart indicates the overall of participation of athletes on the basis of gender.It is observed that the participation of Male atletes is 72.5% whereas the participation of Female athletes is 27.5%. This shows that the number of Male athletes taking part in the Olympic games is more than that of Female.

**3)Histogram:**

The histogram shows the distribution of athletes participating in the Olympic games across the age groups they belong.It is observed that most of the athletes participating belong to the age interval of 20-30. Apart from this ,the age groups from 15-20 and 30-40 also have moderate number of athlete participation.It is also observed that there are very few athletes which belong to the age group of 40-80,hence we can conclude that this age group involves least participants.

**4)Bar Graph:**

The Bar graph Shows the top 10 nations with maximum number of athletes participating in the Olympic games. The countries United states, France ,Great Britain, Italy ,Germany,Canada,Japan,Sweden,Australia and Hungary grab there position in this criteria. Here among these top 10 countries the United States rank on the 1st position with the most participation of athletes since 1900’s. Hungary holds the 10th position with the least participation among the top 10 countries.

**Conclusion: Successfully executed Data frame related operations and visualized data related to Olympic games.**