Statistics For Data Science Datathon UE20CS203

Name : Sneha Saravanan
SRN: PES1UG20CS721
Section:L
Dataset : Olympics

About the Data :

The given dataset gives us information about the olympics medallists . It has a set of rows and columns from their ID to the number of medals won .

What Needs to be done with the Data :

First the dataset has to be studied thoroughly . Post that the dataset should be cleaned off null values and missing values .
The dataset was then visualised for one column specified to identify the outliers . And later on the dataset was manipulated to visualise graphs as necessary .

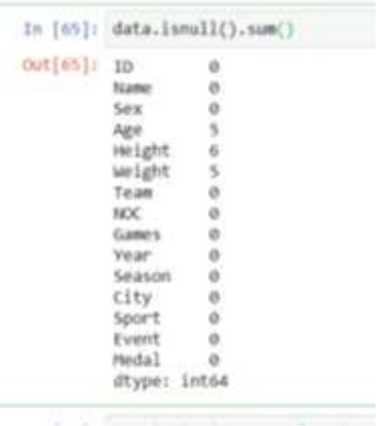
Libraries Used :

1. Numpy
2. Pandas
3. Seaborn
4. Matplotlib

Introductory Tasks :

1. Clean your dataset and fill any missing values in numeric column with the mean

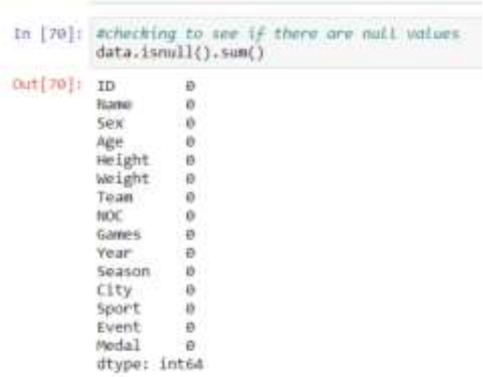
Output : The picture below contains the number of null values in the column before data cleaning was done .



```
In [65]: data.isnull().sum()

Out[65]: ID        0
         Name      0
         Sex       0
         Age       5
         Height    6
         Weight    5
         Team      0
         NOC       0
         Games     0
         Year      0
         Season    0
         City      0
         Sport     0
         Event     0
         Medal     0
         dtype: int64
```

The mean of the column values was found as required and then filled in the place of null values .

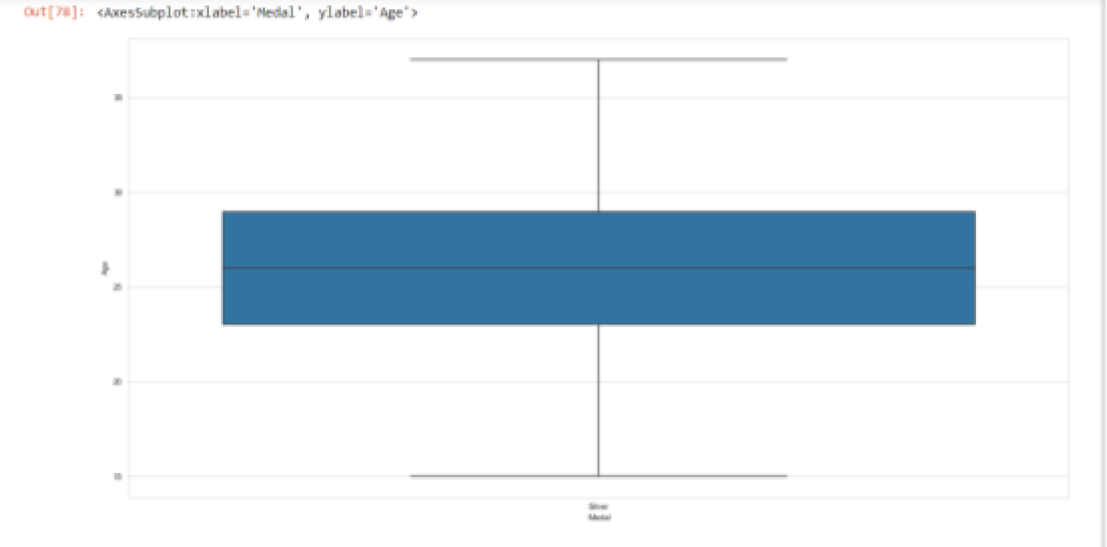


```
In [70]: #checking to see if there are null values
         data.isnull().sum()

Out[70]: ID        0
         Name      0
         Sex       0
         Age       0
         Height    0
         Weight    0
         Team      0
         NOC       0
         Games     0
         Year      0
         Season    0
         City      0
         Sport     0
         Event     0
         Medal     0
         dtype: int64
```

2. Visualize the age distribution for silver medallists

Output :

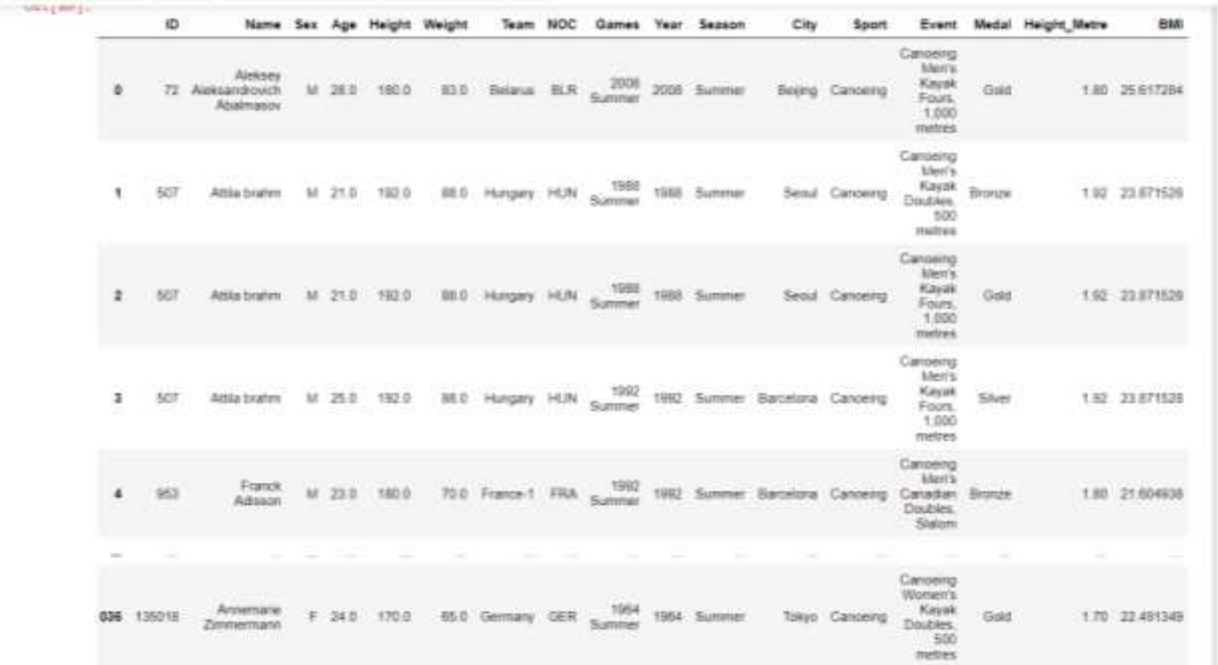


```
Out[78]: <AxesSubplot:xlabel='Medal', ylabel='Age'>
```

3. Create column BMI and calculate the same for each athlete

   Formula used = weight/height*height

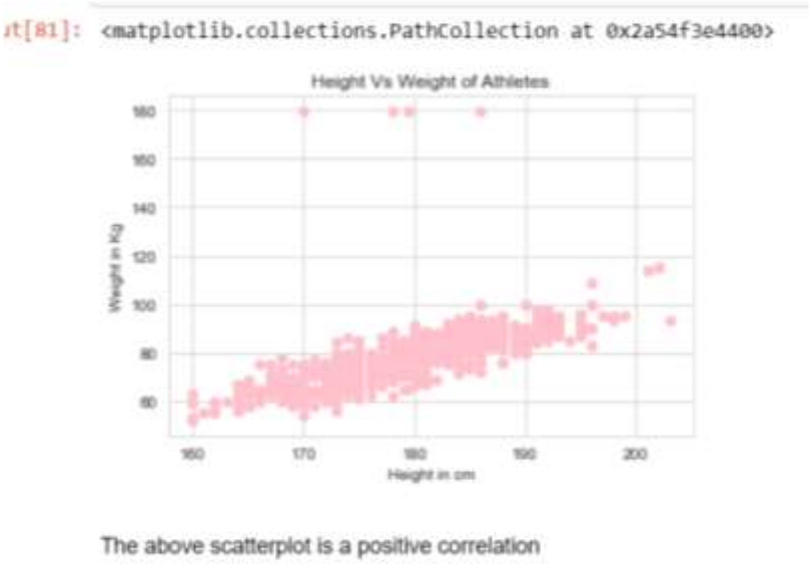
Output :



	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal	Height_Metre	BMI
0	72	Aleksey Aleksandrovich Abalmasov	M	28.0	180.0	83.0	Belarus	BLR	2008 Summer	2008	Summer	Beijing	Canoeing	Canoeing Men's Kayak Fours, 1,000 metres	Gold	1.80	25.617284
1	507	Attila Ibrahim	M	21.0	192.0	88.0	Hungary	HUN	1988 Summer	1988	Summer	Seoul	Canoeing	Canoeing Men's Kayak Doubles, 500 metres	Bronze	1.92	23.871528
2	507	Attila Ibrahim	M	21.0	192.0	88.0	Hungary	HUN	1988 Summer	1988	Summer	Seoul	Canoeing	Canoeing Men's Kayak Fours, 1,000 metres	Gold	1.92	23.871528
3	507	Attila Ibrahim	M	25.0	192.0	88.0	Hungary	HUN	1992 Summer	1992	Summer	Barcelona	Canoeing	Canoeing Men's Kayak Fours, 1,000 metres	Silver	1.92	23.871528
4	953	Franck Adisson	M	23.0	180.0	70.0	France-1	FRA	1992 Summer	1992	Summer	Barcelona	Canoeing	Canoeing Men's Canadian Doubles, Slalom	Bronze	1.80	21.604938
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
036	135018	Annemarie Zimmermann	F	24.0	170.0	65.0	Germany	GER	1964 Summer	1964	Summer	Tokyo	Canoeing	Canoeing Women's Kayak Doubles, 500 metres	Gold	1.70	22.491349

4. Height vs Weight Visualisation

Output :

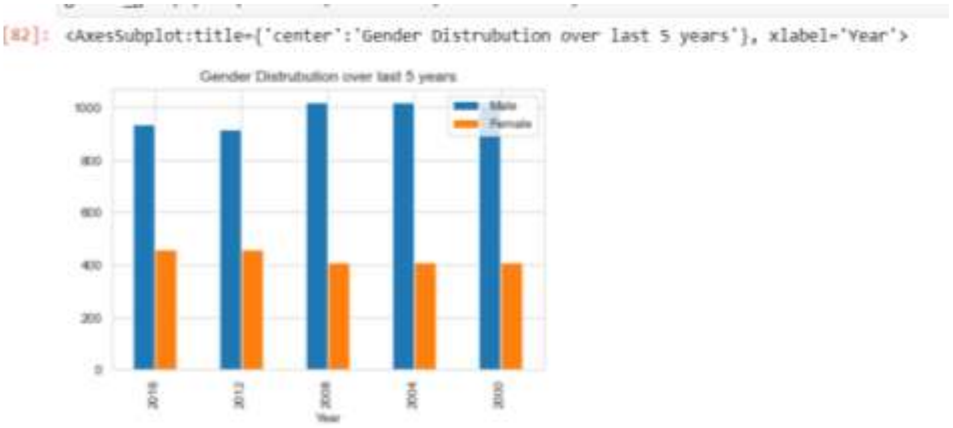


```
ıt[81]: <matplotlib.collections.PathCollection at 0x2a54f3e4400>
```

The above scatterplot is a positive correlation

5. Gender visualisation

Output :

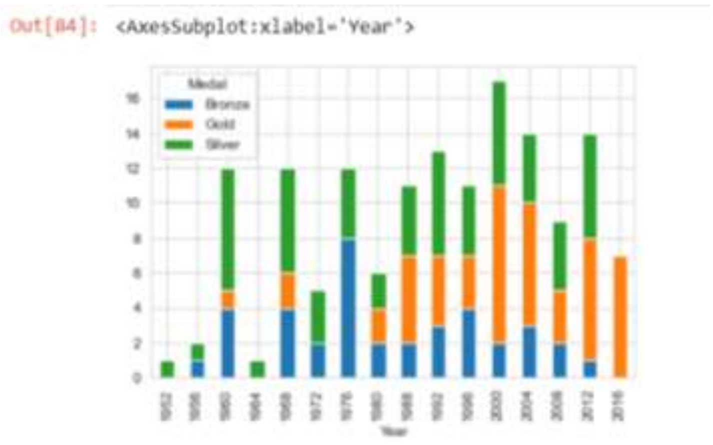


```
[82]: <AxesSubplot:title={'center':'Gender Distrubution over last 5 years'}, xlabel='Year'>
```

Task Questions :

1. Counting teams with maximum year of participation and visualising the number of medals

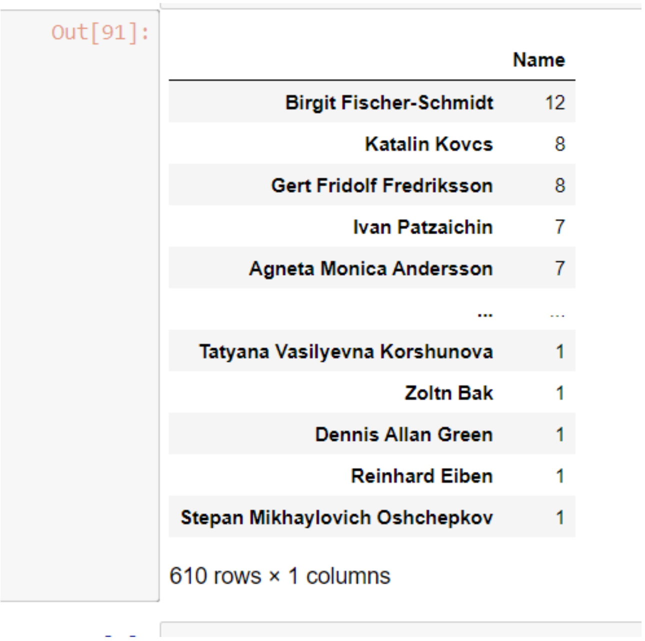
Output :



```
Out[84]: <AxesSubplot:xlabel='Year'>
```

2. Creating a new dataset and calculating medal frequency

Output :



Out[91]:

	Name
Birgit Fischer-Schmidt	12
Katalin Kovcs	8
Gert Fridolf Fredriksson	8
Ivan Patzaichin	7
Agneta Monica Andersson	7
...	...
Tatyana Vasilyevna Korshunova	1
Zoltn Bak	1
Dennis Allan Green	1
Reinhard Eiben	1
Stepan Mikhaylovich Oshchepkov	1

610 rows × 1 columns
