Welcome to Olympics Data Analysis

REPORT

- We all know that Olympics is the one of the most popular sports event and It was first begin in 1896 which was began in Athens , Greece.
- The 5 Rings in this logo represents the five continents: Europe, Africa, Asia, the Americas and Oceania.
- In this popular sports event numerous countries participate for different games.

There is **summer** and **winter** Olympics games and in winter Olympics (and just to give you the headshot :which is happen after every 4 years in which players practice on snow and ice) and in this analysis I have shown you visualization on the basis of winter and summer.

In today's analysis we are going to do a Exploratory Data Analysis using Python on this 'Olympics Data' to analyze and visualize.

So, in order to do analysis onto this we are going to take several steps below:

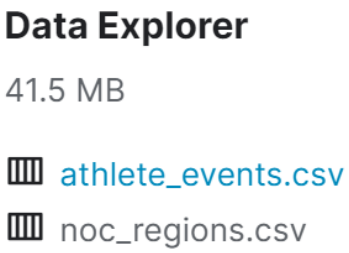
I am going to do Analysis from 1896-2016

- Data Collection
- Take Overview Of Your Data(By Observation)
- Data Preprocessing and Data Manipulation
- Data Visualization

DATA COLLECTION

- So, I am going collect my data from kaggle platform and the name of that dataset is 120 years of Olympic History from 1896 to 2016 (RIO-DE-JENERIO).

- **And we have this dataset in two parts one is athletics_events and second on is noc_regions. So, we have to merge this for our analysis.**
- **This datasets are in the form of csv file that is comma separated value.**



Take Overview Of Your Data(By Observation):

- In this step you have to observe the data by just seeing and take a overview of your data that is what you have to do actually.

Data Preprocessing and Data Manipulation:

- So, in this step we are going to merge our datasets so, that we can do our analysis in a better way.

```
#Join the dataframes or Merge the dataframe

ath=ath.merge(region_df,on='NOC',how='left')
```

```
ath.head()
```

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal	region	notes
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NaN	China	NaN
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	NaN	China	NaN

- Renaming the column ‘region’ into ‘Region’

```
#Here there is a problem which is many of the column
#whose name is start with lower case so
#we have to convert that into upper case
ath.rename(columns={'region':'Region','notes':'Notes'},inplace=True)
```

```
ath.head()
```

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal	Region	Notes
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NaN	China	NaN

- SO Here we have some duplicate values so we have to remove them by calling function called

```
ath.drop_duplicates(subset=['Team', 'NOC', 'Games', 'Year', 'City', 'Sport', 'Event', 'Medal'],inplace=True)
```

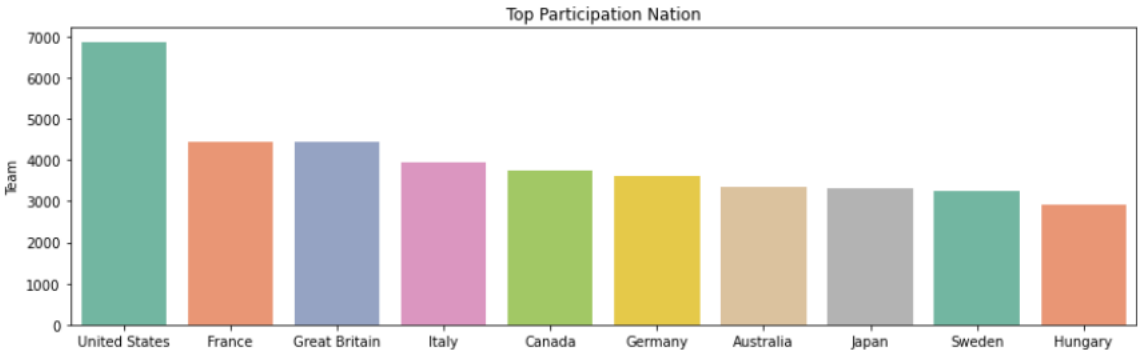
Data Visualization:

- In this we are visualizing our data by plotting bar graphs and distributions.

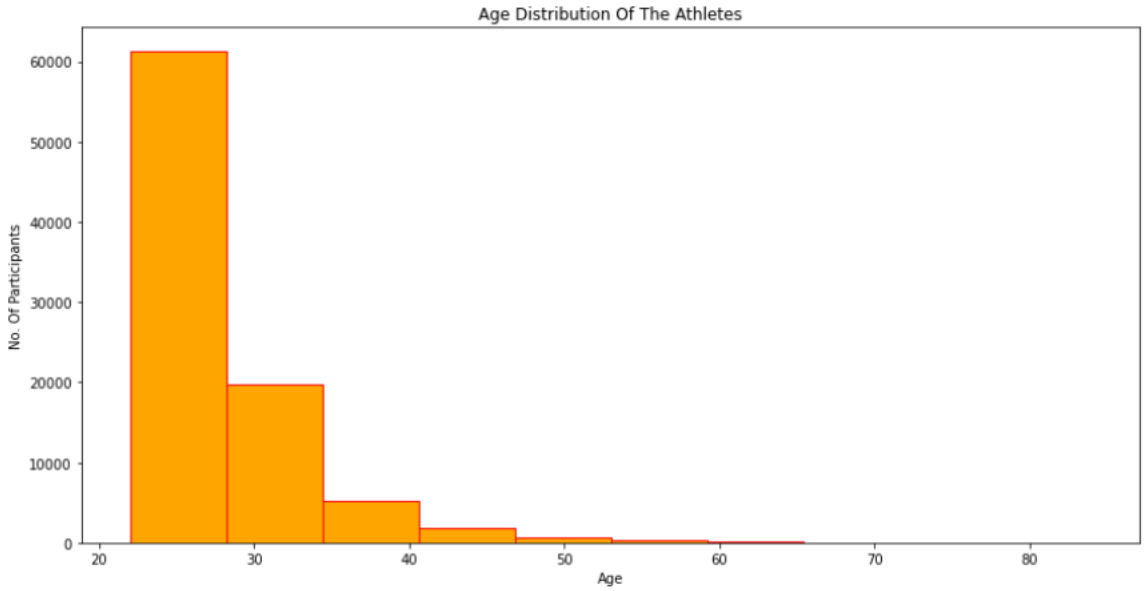
```
In [48]: #Let's see the statistical behaviour of our data
         #by applying describe() function
         #But this will us only analysis about numeric columns
         ath.describe()
```

Out[48]:

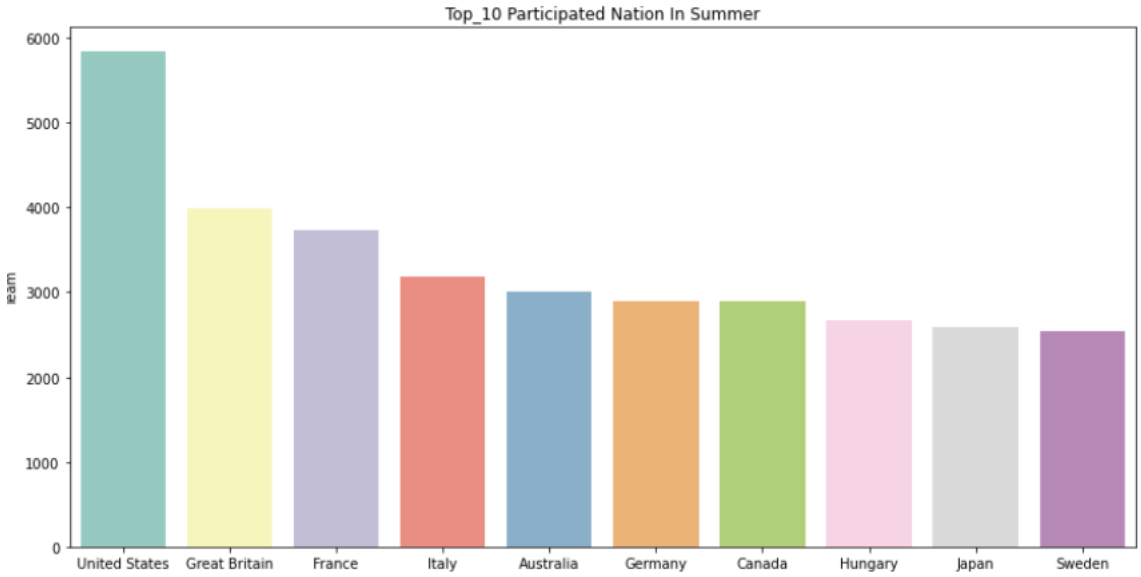
	ID	Age	Height	Weight	Year
count	124634.000000	121282.000000	102110.00000	101331.000000	124634.000000
mean	54363.256559	25.257375	174.76825	70.499033	1982.168766
std	38477.267803	5.987941	10.05985	14.905281	28.146116
min	1.000000	11.000000	127.00000	28.000000	1896.000000
25%	20326.250000	21.000000	168.00000	60.000000	1968.000000
50%	48102.000000	24.000000	175.00000	69.000000	1988.000000
75%	84618.750000	28.000000	182.00000	78.000000	2004.000000
max	135560.000000	84.000000	218.00000	214.000000	2016.000000



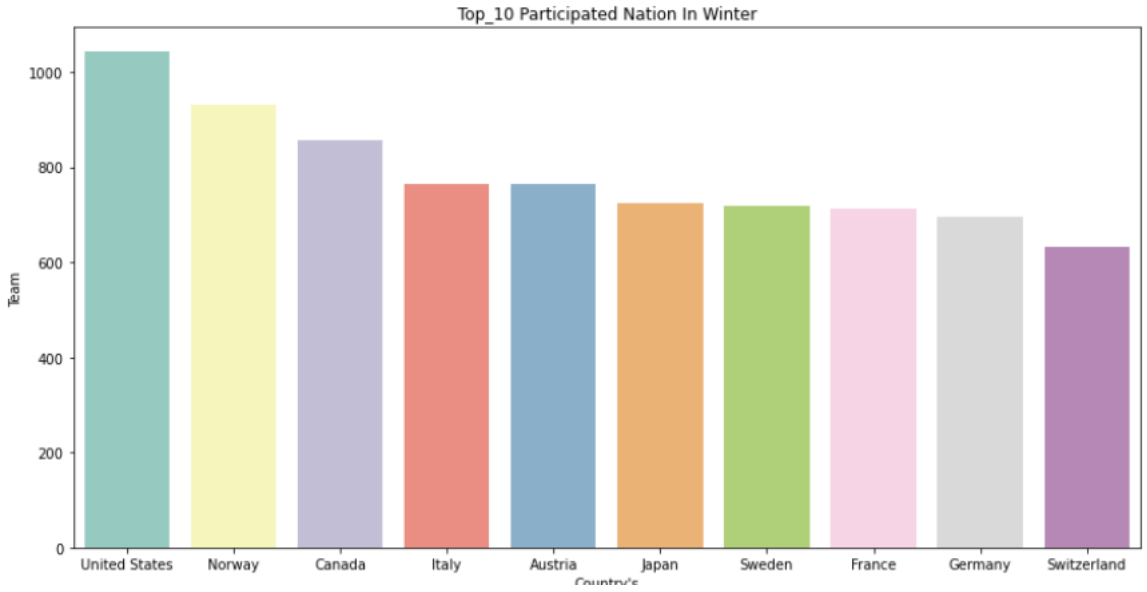
- Here, are the top participated nations and teams participated onto this.
- You can see in this bar graph United Nation ‘s Team Participation Rate is high and ranked as first.
- Outcome is UNITED NATION HAS TOP PARTICIPATION RATE.



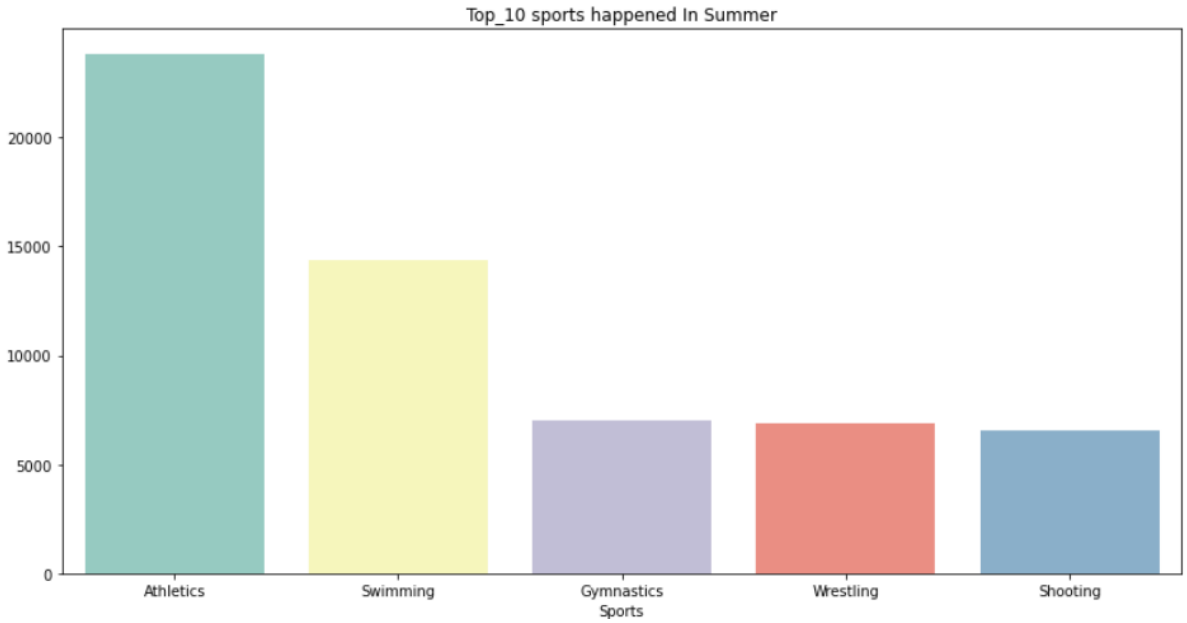
- Bar graph of the ***athletes*** participation whose age is greater than 21 years
- In this you can clearly see between 21 years-30 years there is maximum participation.
- Outcome ***21 YEARS-30 YEARS HIGHEST PARTICIPATION RATE***.



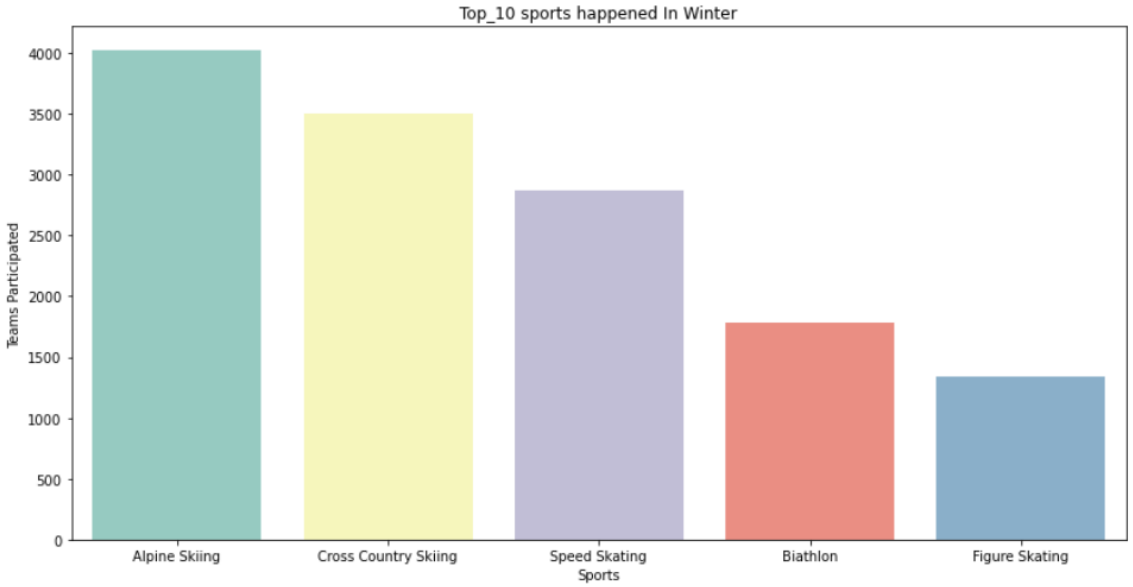
- Here , again united states is the nation who is maximum participation in summer.
- Outcome is ***UNITED NATION HAS TOP PARTICIPATION RATE*** in ***SUMMER*** as well.



- Here , again united states is the nation who is maximum participation in winter.
- Outcome is ***UNITED NATION HAS TOP PARTICIPATION RATE*** in ***WINTER*** as well.



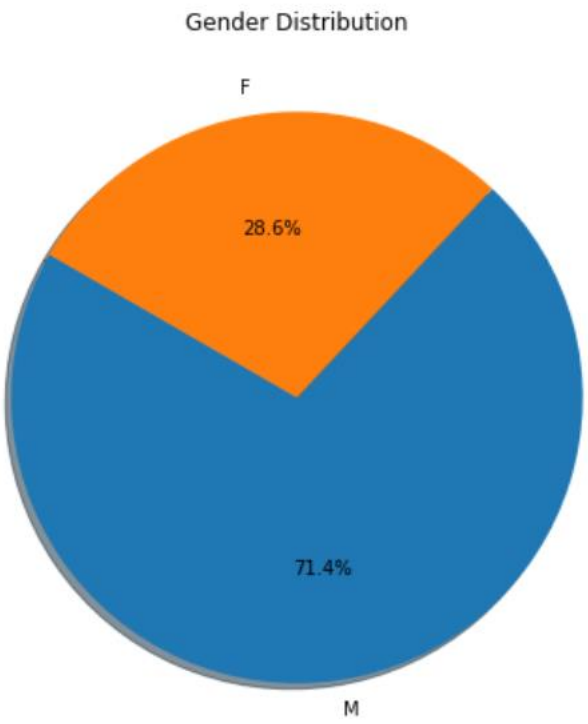
- *Athletics* was the sport which was held maximum times in SUMMER.
- And number of the teams participated in athletics also very high.



- Alpine Skiing was the sport which was held maximum times in WINTER.
- And number of the teams participated in *ALPINE SKIING* also very high.

```
#Male and Female Participants
gender_counts=ath.Sex.value_counts()
gender_counts
```

```
M    88949
F    35685
Name: Sex, dtype: int64
```

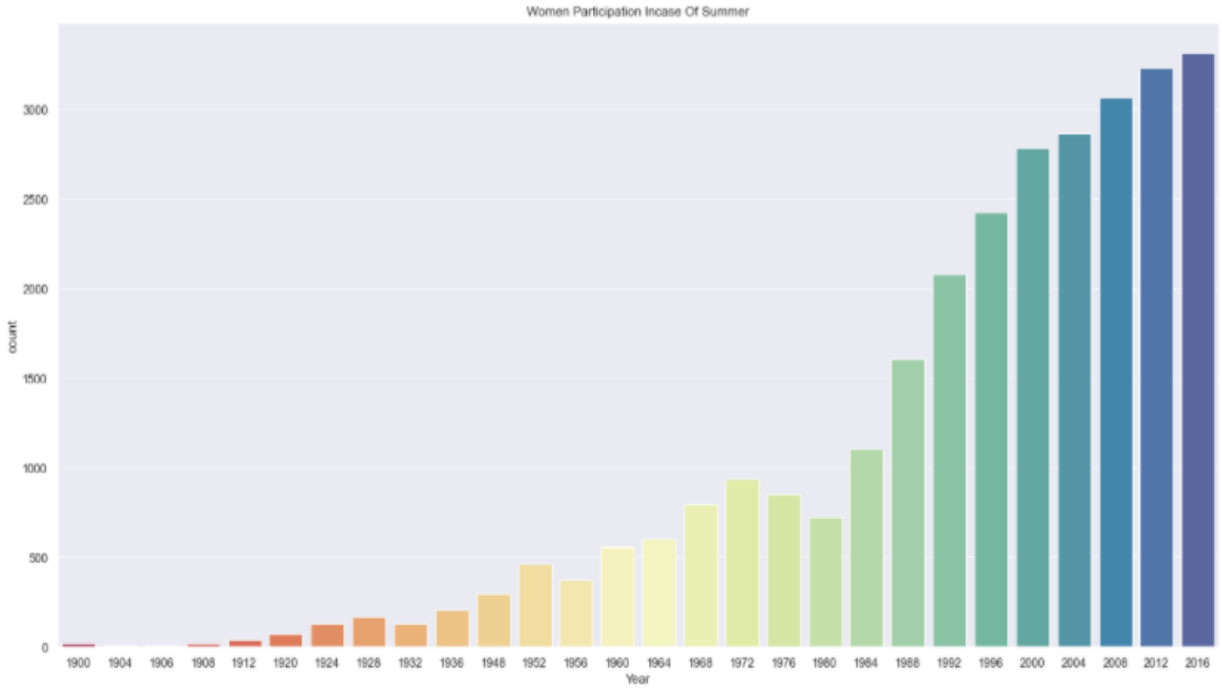


	Year	Sex
0	1900	18
1	1904	9
2	1906	8
3	1908	18
4	1912	39

- ***Female*** Participation In ***Each Olympics***
- Incase of ***Summer*** season.

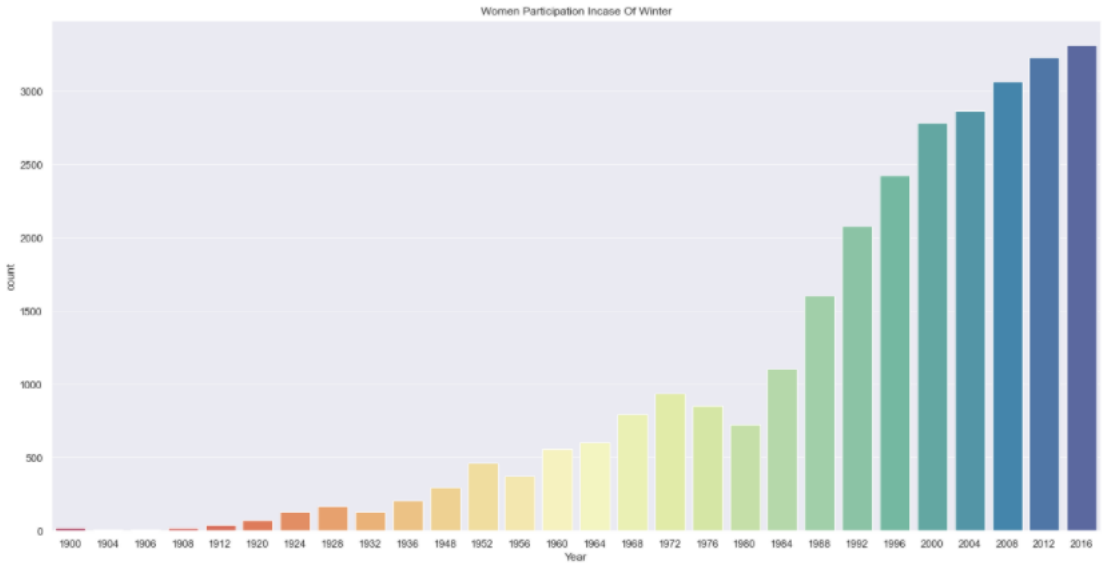
	Year	Sex
0	1924	11
1	1928	16
2	1932	11
3	1936	41
4	1948	61

- ***Female*** Participation In ***Each Olympics***
- Incase of ***Winter*** season.

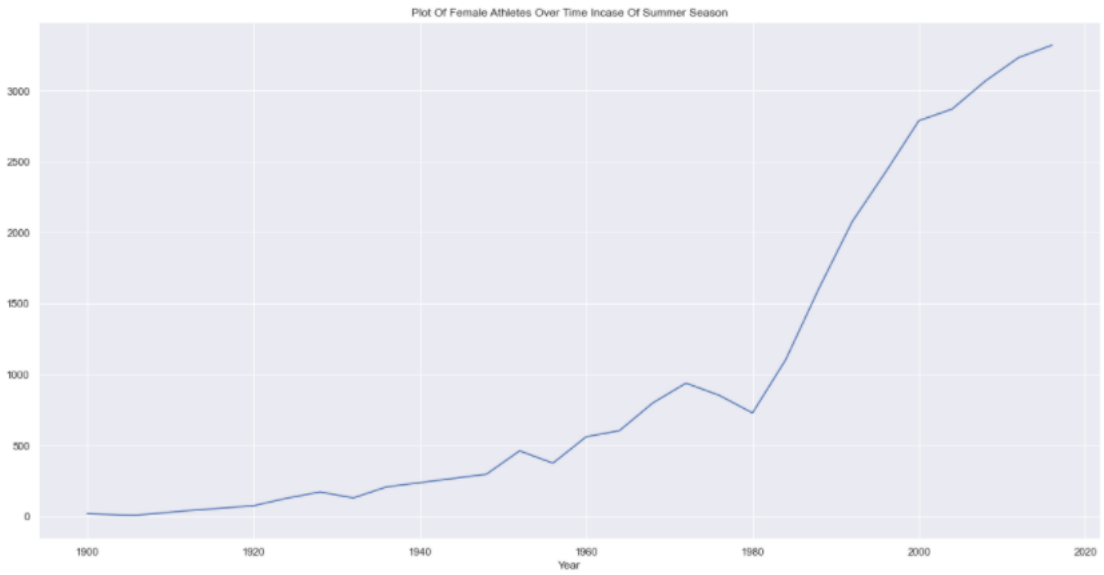


- ***Countplot*** of Women Participation ***Incase Of Summer*** Season

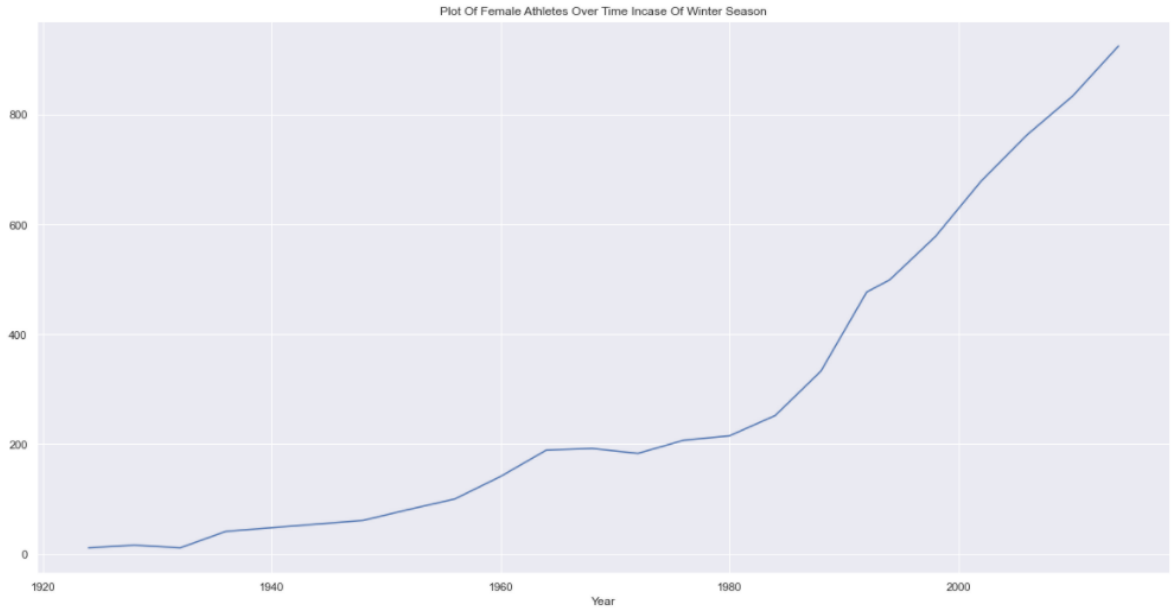
- ***Highest*** in ***2016***



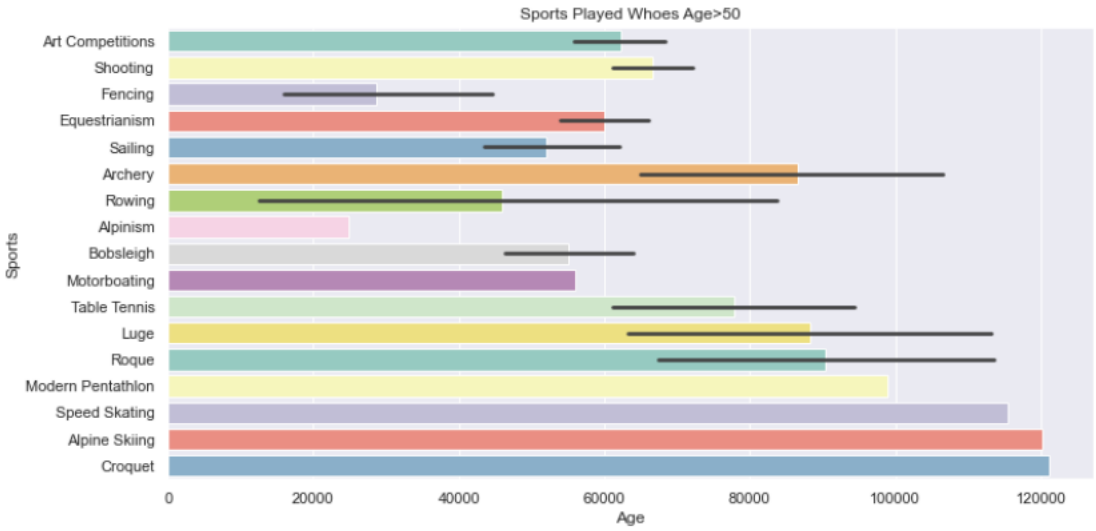
- ***Countplot*** of Women Participation ***Incase Of Winter*** Season also same
- ***Which is Highest*** in ***2016.***



- ***Plot Of Female*** Athletes Over Time ***Incase Of Summer*** Season
- Highest in 2016



- ***<u>Plot Of Female</u>*** Athletes Over Time ***<u>Incase Of Winter</u>*** Season
- Highest in 2016.



- Sports Played ***<u>Whoes Age>50</u>***
- **<u>Croquet</u>** ***<u>highest played sports</u>*** whose age is over 50years .

```
# Athletes over 50 years age won how may golds
gold_guy['ID'][gold_guy['Age']>50].count()
```

```
32
```

- Here, there are ***32 athletes*** who ***won gold*** and there age is ***above 50 years.***

```
gold_sports=gold_guy['Sport'][gold_guy['Age']>50]
```

```
gold_sports
```

```
4527          Equestrianism
11772         Equestrianism
11774         Equestrianism
21014         Equestrianism
24889              Alpinism
45910               Sailing
53872         Equestrianism
66723     Art Competitions
67359                Roque
74732         Equestrianism
79304     Art Competitions
80376              Shooting
82440         Equestrianism
89929              Shooting
91417              Shooting
91419              Shooting
93119         Equestrianism
94212               Archery
94465     Art Competitions
95285               Sailing
106834        Equestrianism
111213        Equestrianism
113681    Art Competitions
114293             Shooting
116051    Art Competitions
```

- Here are the ***athletes whose age*** is **more than *50*** years won ***GOLD*** in these major sports.

So Our Olympic Data Analysis Finished I Hope It Was Informative and Insights Full.

Report By:

Amit kr. Upadhyay,

Phone no.

9560085010

THANK YOU