Project Proposal Temlate

Data Analysis and Project Proposal for Olympics Dataset

05/2020

Volkan K.

I have chosed the Olympics Dataset. The datasets are in csv format and I am more comfortable reading csv's to pandas DataFrame. For this reason I selected this dataset.

a. I have downloaded the csv files from coursera web page.

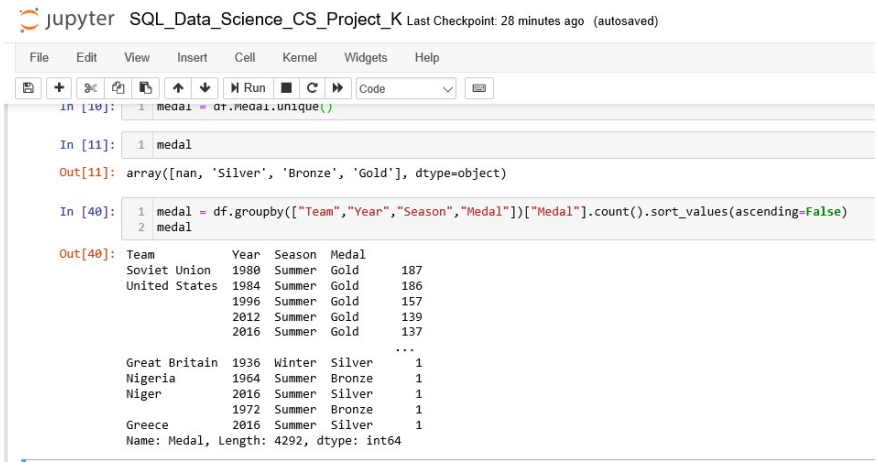
b. I have imported necessery libraries for reading the data set with pd.read_csv() into pandas dataframe

b. First things first, so I have did the EDA(Exploratory Data Analysis) to understand and get some sense of the data.

c. Then I used some visualization to see the relationship or trends between the columns.

d. Datatypes of the columns were good, so there was no need to convert them.

3. Reading and initial exploration of data.

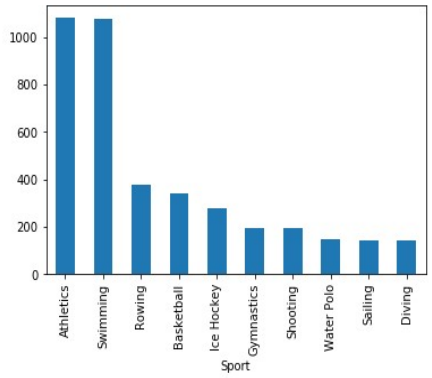


Statistics regarding USA:

```
[14]: USA_sport_grouped = df.loc[df.NOC == 'USA'].groupby('Sport')['Medal'].count().sort_values(ascending=False)
      USA_sport_grouped
```

...

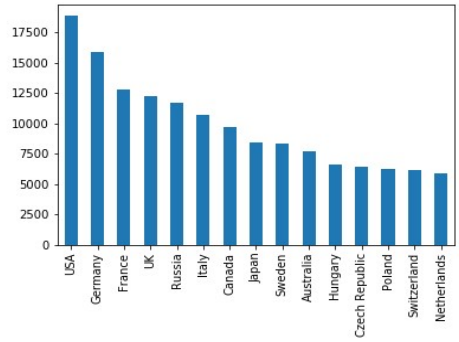
```
[18]: USA_sport_grouped[:10].plot(kind="bar")
      plt.show()
```



It is clear that in Athletics and Swimming USA is much better than other branches.

```
[21]: country_names = df["region"]
```

```
[23]: medal_counts = country_names.value_counts()

      # Print top 15 countries ranked by medals
      #print(medal_counts.head(15))
      medal_counts[:15].plot(kind="bar")
      plt.show()
```



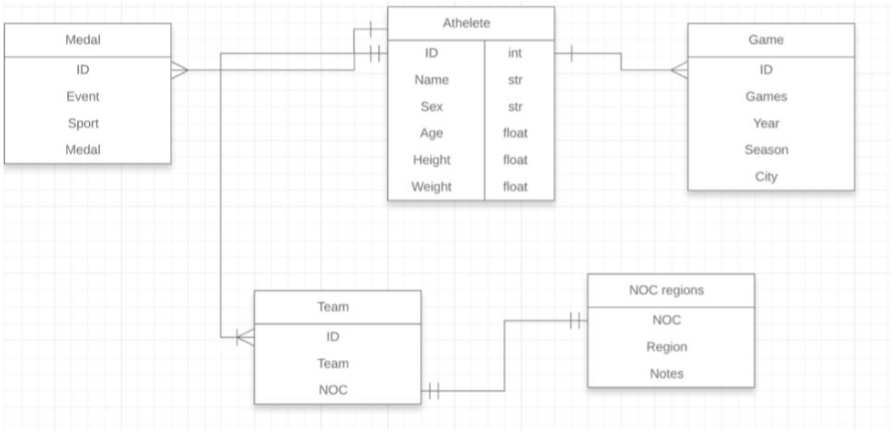
First 15 countries according to their total medal counts

[24]:
```
# Construct the pivot table: counted
counted = df.pivot_table(index="NOC", columns="Medal", values="Name", aggfunc="count")

# Create the new column: counted['totals']
counted['totals'] = counted.sum(axis="columns")

# Sort counted by the 'totals' column
counted = counted.sort_values("totals", ascending=False)

# Print the top 15 rows of counted
print(counted.head(15))
```

```
Medal  Bronze    Gold  Silver  totals
NOC
USA    1358.0  2638.0  1641.0  5637.0
URS     689.0  1082.0   732.0  2503.0
GER     746.0   745.0   674.0  2165.0
GBR     651.0   678.0   739.0  2068.0
FRA     666.0   501.0   610.0  1777.0
ITA     531.0   575.0   531.0  1637.0
SWE     535.0   479.0   522.0  1536.0
CAN     451.0   463.0   438.0  1352.0
AUS     517.0   348.0   455.0  1320.0
RUS     408.0   390.0   367.0  1165.0
HUN     371.0   432.0   332.0  1135.0
NED     413.0   287.0   340.0  1040.0
NOR     294.0   378.0   361.0  1033.0
GDR     281.0   397.0   327.0  1005.0
CHN     292.0   350.0   347.0   989.0
```

Countries by Medal counts.
Last Column is the total number of other three columns.

4. ERD to show the relationships of the data.



Anybody who wants to explore olympics data set is a **potential audience** for this Project. This Project can help them to understand the processes of how to make an exploratory data analysis and getting some insights of the data.

6. Did the learner create 2-3 questions that they want to answer with the data?

1. Which region won the most medals in total? How many?

2. Which region have the highest medal winning ratio(number of medal winners compared with participated atheletes)?

3. What's the average participation times of an athelete?

4. Who have participated the most Olympic games?

5. How many times have they won?

7. Did the learner provide 2-3 assumptions about the data?

1. I assume that womens were not involved at the very first years of the olympics. I will try to find when was the first year that womens are began to attend to the olympics.

2. I assume that host country/countries had shown better performance than the years that they were not the host of the games.

8. Did the learner describe in 5-6 sentences their approach to prove or disprove their hypotheses?

Approach

1. To test my hypothesis, I'll be using primarily average metrics.

2. I will separate the data into two to see the difference of male and female participance.

3. I will count the medals for every country to be able to measure the performance of the countries.

4. I will try to find the host countries and try to measure their performance for both while hosting and not hosting the olympics.

5. I will try to find the first participation of female athletes to olympics.

Summary of the different descriptive statistics:

1. There is increase in number of medals won from 1896 Olympics to 2012 Summer olympics.
2. One of the reasons will definitely be increase in participating countries, more sports and also number of athletes.
3. One thing clearly stands out is richest and developed countries have won the highest number of medals
4. They are still winning majority of the medals.
5. I have looked at these interesting facts while I try to find interesting facts about the data sets.

Some key points:

1. I have found out that, there is not any female athletes before 2000. With this information I have found the answer to one of my hypothesis in week 1.
2. Athletes have won more medals in London.

3. USA has won the highest number of gold medals when compared to other medals and almost equal percent of silver and bronze medals.
4. in London athletes have won more medals and 4 cities have hosted the olympics more than once. In these London is the only city which hosted the olympics 3 times. Los Angeles, Paris and Athens have hosted twice.
5. Great Britain has won medals in all the games from 1896-2012. USA is by far the most highest medal winner and followed by Soviet Union and Great Britain.
6. Nearly 50% of all the medals won by Americans are Gold.
7. Almost 75% of the medals are won by men