Assignment 3 - Pandas DataFrame

Part 1 (75 Points evenly distributed to first three questions)

The following code loads the olympics dataset (olympics.csv), which was derived from the Wikipedia entry on [All Time Olympic Games Medals (https://en.wikipedia.org/wiki/All-time_Olympic_Games_medal_table)](https://en.wikipedia.org/wiki/All-time_Olympic_Games_medal_table), and does some basic data cleaning.

The columns are organized as # of Summer games, Summer medals, # of Winter games, Winter medals, total # number of games, total # of medals. Use this dataset to answer the questions below.

In [2]:
```
import pandas as pd

df = pd.read_csv('olympics.csv', index_col=0, skiprows=1)

for col in df.columns:
    if col[:2]=='01':
        df.rename(columns={col:'Gold'+col[4:]}, inplace=True)
    if col[:2]=='02':
        df.rename(columns={col:'Silver'+col[4:]}, inplace=True)
    if col[:2]=='03':
        df.rename(columns={col:'Bronze'+col[4:]}, inplace=True)
    if col[:1]=='№':
        df.rename(columns={col:'#'+col[1:]}, inplace=True)

names_ids = df.index.str.split('\s\(') # split the index by '('

df.index = names_ids.str[0] # the [0] element is the country name (new index)
df['ID'] = names_ids.str[1].str[:3] # the [1] element is the abbreviation or ID (

df = df.drop('Totals')
df.head()
```

Out[2]:

	# Summer	Gold	Silver	Bronze	Total	# Winter	Gold.1	Silver.1	Bronze.1	Total.1	Gam
Afghanistan	13	0	0	2	2	0	0	0	0	0	
Algeria	12	5	2	8	15	3	0	0	0	0	
Argentina	23	18	24	28	70	18	0	0	0	0	
Armenia	5	1	2	9	12	6	0	0	0	0	
Australasia	2	3	4	5	12	0	0	0	0	0	

Question 1

What is the first country in df?

This function should return a Series.

In [4]:
```
# You should write your whole answer within the function provided.
def answer_zero():
    # This function should return the row for the first country, which is a Serie
    return df.iloc[0]

# You can examine what your function returns by calling it in the cell.
answer_zero()
```

Out[4]:
```
# Summer           13
Gold                0
Silver              0
Bronze              2
Total               2
# Winter           0
Gold.1              0
Silver.1            0
Bronze.1            0
Total.1             0
# Games            13
Gold.2              0
Silver.2            0
Bronze.2            2
Combined total      2
ID                AFG
Name: Afghanistan, dtype: object
```

Question 2

In summer games, which nation has won the most gold medals?

This function should return a single string value.

In [5]:
```
def answer_one():
     return df.sort_values(by = "Gold", ascending = False).iloc[0,1:2]
    #return df.Gold.max() - However, this is only returning the max value and not


answer_one()
```

Out[5]:
```
Gold    976
Name: United States, dtype: object
```

Question 3

Which nation had the biggest difference on gold medal counts? (between their summer and winter)

This function should return a single string value.

In [25]:
```
def answer_one():
    df['Gold dif'] = df['Gold'] - df['Gold.1']
    return df.sort_values(by = "Gold dif", ascending = False).iloc[0:1,-1]

answer_one()
```

Out[25]:
```
United States    880
Name: Gold dif, dtype: int64
```

PART 2

Question 4 (25 Points for first two questions + 25 Points Bonus for last two questions)

We will look at the publicly available airline data in this question similar to flights.csv that we covered in class practices. However, in this assignment, you are given 6 months of seperate data along with lookup tables for carriers and airports. Please apply data exploration and pre-procesing techniques and provide your answers for the following questions.

Questions:

1. What carrier has flown the 1st most number of flights? How many?
2. Which airport has the 3rd most delays?
3. What is the most popular day of the week to travel?
4. What is the 1st and 5th most flown route?

Hints:

1- pd.concat(list) list=[A,B,C...] e.g. A= pd.read_csv("1.csv",encoding='utf-8')

2-please leverage from pandas dataframe features including groupby(...).size()... groupby(...).sum().sort_values(....)

3- Dont forget to consider cancelled flights

4- Try to create a new column for "route"

Dataset Details: Dataset name: On-Time Performance, Lookup Table: Carrier Lookup, Lookup Table: Airport Lookup

In [16]:
```
#Importing the Data
import pandas as pd

df1 = pd.read_csv('1.csv',encoding='utf-8')
df2 = pd.read_csv('2.csv',encoding='utf-8')
df3 = pd.read_csv('3.csv',encoding='utf-8')
df4 = pd.read_csv('4.csv',encoding='utf-8')
df5 = pd.read_csv('5.csv',encoding='utf-8')
df6 = pd.read_csv('6.csv',encoding='utf-8')
frames=[df1,df2,df3,df4,df5,df6]
df_f = pd.concat(frames)
```

In [22]:
```
result_df = df_f[df_f['CANCELLED'] < 1]
result_df.groupby('UNIQUE_CARRIER').size().nlargest(1)
```

Out[22]:
```
UNIQUE_CARRIER
WN    568904
dtype: int64
```

In [27]:
```
departure_del = result_df[rslt_df['DEP_DELAY_NEW']>0]
departure_del.groupby('ORIGIN').size().nlargest(3).iloc[2:]
```

Out[27]:
```
ORIGIN
DEN    58710
dtype: int64
```