LAB 2

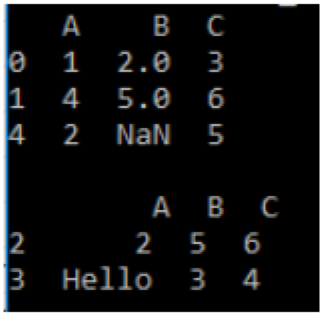
CS481 - Data Science

Note: Question # 2,3,6 do not need to be submitted and are just for practice

Question # 1 :

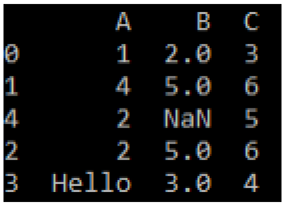
In this question, we will be taking input from different sources. Data may be incomplete and there may be missing values. Objective is to combine data and save it as one dataframe object. Do the following steps :

a. Read data1.csv and data2.csv as df1 and df2 dataFrames objects respectively. Use Pandas I/O libraries. Use column 0 as index for df1 and column 1 as index for df2. Print df1 and df2



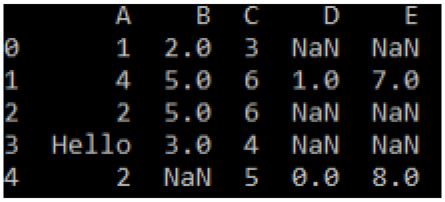
```
   A    B  C
0  1  2.0  3
1  4  5.0  6
4  2  NaN  5

       A  B  C
2      2  5  6
3  Hello  3  4
```

b. Modify (a) to concatenate df1 and df2 and assign it to df3. Print df3



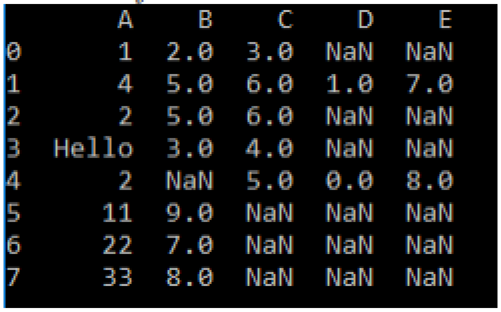
```
       A    B  C
0      1  2.0  3
1      4  5.0  6
4      2  NaN  5
2      2  5.0  6
3  Hello  3.0  4
```

c. Read data3.csv as df4 dataframe object and print df4 (not shown below). There are 2 new column 'D' and 'E' in this file. Merge df4 with df3 so that new dataframe (df5) has total 5 columns (A, B, C, D, E)



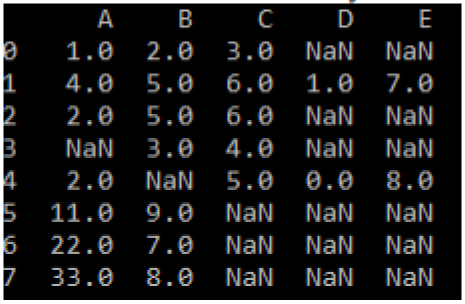
```
       A    B  C    D    E
0      1  2.0  3  NaN  NaN
1      4  5.0  6  1.0  7.0
2      2  5.0  6  NaN  NaN
3  Hello  3.0  4  NaN  NaN
4      2  NaN  5  0.0  8.0
```

d. Read data.json as df6 and concatenate with df5. Use df7 as name of dataframe



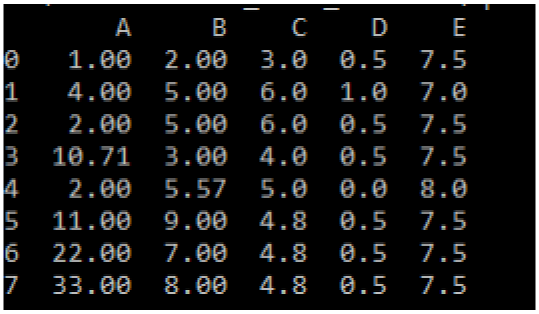
```
       A    B    C    D    E
0      1  2.0  3.0  NaN  NaN
1      4  5.0  6.0  1.0  7.0
2      2  5.0  6.0  NaN  NaN
3  Hello  3.0  4.0  NaN  NaN
4      2  NaN  5.0  0.0  8.0
5     11  9.0  NaN  NaN  NaN
6     22  7.0  NaN  NaN  NaN
7     33  8.0  NaN  NaN  NaN
```

e. Replace Hello with NaN. Make it as general as possible so that all strings in dataframe df7 automatically becomes NaN



```
      A    B    C    D    E
0   1.0  2.0  3.0  NaN  NaN
1   4.0  5.0  6.0  1.0  7.0
2   2.0  5.0  6.0  NaN  NaN
3   NaN  3.0  4.0  NaN  NaN
4   2.0  NaN  5.0  0.0  8.0
5  11.0  9.0  NaN  NaN  NaN
6  22.0  7.0  NaN  NaN  NaN
7  33.0  8.0  NaN  NaN  NaN
```

f. Replace NaN with mean values of the columns. Save the final dataframe as "newdata.csv"



```
       A     B    C    D    E
0   1.00  2.00  3.0  0.5  7.5
1   4.00  5.00  6.0  1.0  7.0
2   2.00  5.00  6.0  0.5  7.5
3  10.71  3.00  4.0  0.5  7.5
4   2.00  5.57  5.0  0.0  8.0
5  11.00  9.00  4.8  0.5  7.5
6  22.00  7.00  4.8  0.5  7.5
7  33.00  8.00  4.8  0.5  7.5
```

Question # 2:

The following website provides a very good exercise about real world data cleaning.
Follow the steps and complete the exercise below :
https://realpython.com/python-data-cleaning-numpy-pandas/

Question # 3:

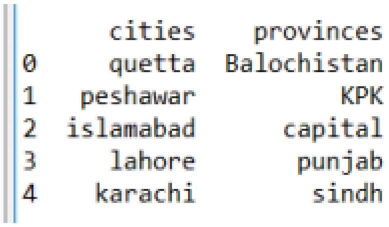
Finish Ten Minutes exercise at :

https://pandas.pydata.org/pandas-docs/stable/user_guide/10min.html

Question # 4:

Complete the following program

```
import pandas as pd
data = {'cities' : ['lahore','karachi',], 'provinces' : ['punjab','sindh']}

# store data as DataFrame object. Assign object name as frame1
frame1 = ______________________

# print frame
 _______________________


data2 = {"cities": ["islamabad","karachi","peshawar","quetta"], "provinces": ["capital","sindh",
"KPK","Balochistan"]}

# store data as DataFrame object. Assign object name as frame2
______________________________________


# combine both objects frame1 and frame2; without any duplicate rows and re-arrange all indexes
frame3 = ……………………………………. # combine frame1 and frame2
frame3 = ……………………………………… # remove duplicates rows
frame3 = …………………………………………# sort based on provinces
frame3 = …………………………………………# re-arrange all indexes
 ………………………………………………………. # print frame3
```

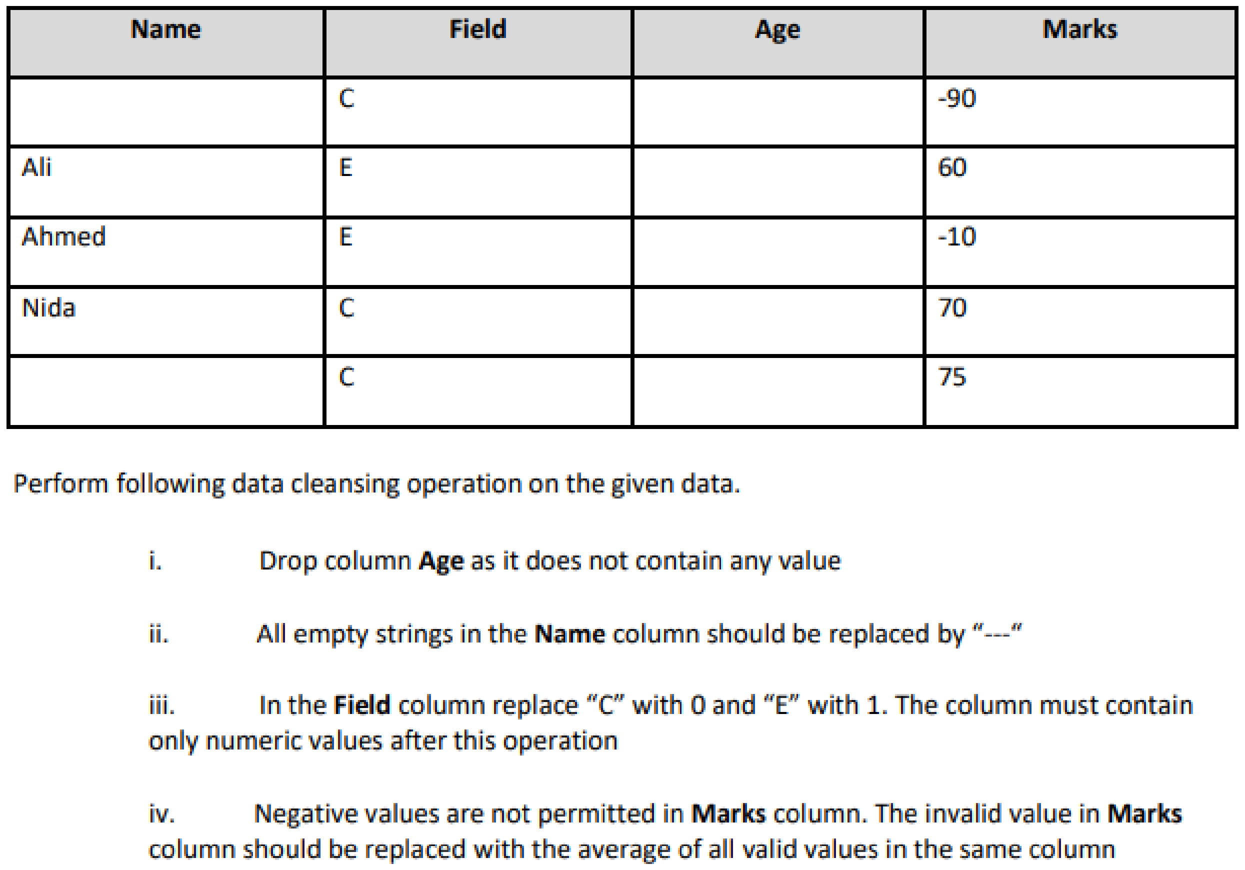


```
      cities    provinces
0     quetta  Balochistan
1   peshawar          KPK
2  islamabad      capital
3     lahore       punjab
4    karachi        sindh
```

Figure1: Screen shot of Final Output

Question # 5 :

Consider the following table :



Name	Field	Age	Marks
	C		-90
Ali	E		60
Ahmed	E		-10
Nida	C		70
	C		75

Perform following data cleansing operation on the given data.

i. Drop column **Age** as it does not contain any value

ii. All empty strings in the **Name** column should be replaced by "---"

iii. In the **Field** column replace "C" with 0 and "E" with 1. The column must contain only numeric values after this operation

iv. Negative values are not permitted in **Marks** column. The invalid value in **Marks** column should be replaced with the average of all valid values in the same column

Question # 6:

Finish Exploratory Data Analysis with Pandas from the following link :

https://www.kaggle.com/kashnitsky/topic-1-exploratory-data-analysis-with-pandas

https://www.kaggle.com/ekami66/detailed-exploratory-data-analysis-with-python