Topic 6 - Python for Data Analysis : Data Preprocessing with Pandas

Join

Using two of the dataframes below, answer the questions :

```
df1 = pd.DataFrame({'key': ['K0', 'K1', 'K2', 'K3', 'K4', 'K5'],
                    'A': ['A0', 'A1', 'A2', 'A3', 'A4', 'A5']})

df2=pd.DataFrame({'key': ['K0', 'K1', 'K2'],
                      'B': ['B0', 'B1', 'B2']})
```

```
#import libraries here
import pandas as pd
```

Create a dataframe (df3) which consists both dataframe inner-joined by column "Key" (5 pts)

```
#code goes here
df3=pd.merge(df1,df2,on='key',how='inner')
df3
```

```
  key   A   B
0  K0  A0  B0
1  K1  A1  B1
2  K2  A2  B2
```

Create a dataframe (df4) which consists of the result of df1 and df2 left-joined by column "Key" (10pts)

```
#code goes here
df4=pd.merge(df1,df2,on='key',how='left')
df4
```

```
  key   A    B
0  K0  A0   B0
1  K1  A1   B1
2  K2  A2   B2
3  K3  A3  NaN
4  K4  A4  NaN
5  K5  A5  NaN
```

1. Replace key "K2" into "K3" on df1

2. Left join it with df2

(15pts)

```
#code goes here
df1['key']=df1.replace({'K2':'K3'})
pd.merge(df1,df2,on='key',how='left')
```

```
  key   A    B
0  K0  A0   B0
1  K1  A1   B1
2  K3  A2  NaN
3  K3  A3  NaN
4  K4  A4  NaN
5  K5  A5  NaN
```

Dataframe Processing

Create a dictionary and convert it to a dataframe called "Customer_df" (15 pts)

name=['Anna','Jason','Cindy']

age=['22','23','21']

gender=['F','M','F']

```
#code goes here
dictionary = {
    'name' : ['Anna','Jason','Cindy'],
    'age' : ['22','23','21'],
    'gender' :  ['F','M','F']
}

Customer_df=pd.DataFrame(dictionary)
Customer_df
```

```
    name age gender
0   Anna  22      F
1  Jason  23      M
2  Cindy  21      F
```

Delete the "age" column (5 pts)

```
#code goes here
Customer_df.drop('age',axis=1)
```

```
    name gender
0   Anna      F
1  Jason      M
2  Cindy      F
```

Rename the "name" column into "first name" (10 pts)

```
#code goes here
Customer_df.rename(columns={'name':'first name'})
```

```
  first name age gender
0       Anna  22      F
1      Jason  23      M
2      Cindy  21      F
```

Sort the dataframe by age, descending (10 pts)

```
#code goes here
Customer_df.sort_values('age',ascending=False)
```

```
    name age gender
1  Jason  23      M
0   Anna  22      F
2  Cindy  21      F
```

Run the code below for data source and answer the following questions

```
import pandas as pd
df=pd.read_csv('http://bit.ly/kaggletrain')
```

Show top 10 rows of the dataset (5 pts)

```
#code goes here
df.head(10)
```

```
   PassengerId  Survived  Pclass  \
0            1         0       3
1            2         1       1
2            3         1       3
3            4         1       1
4            5         0       3
5            6         0       3
6            7         0       1
7            8         0       3
8            9         1       3
9           10         1       2

                                                 Name     Sex   Age
SibSp  \
0                             Braund, Mr. Owen Harris    male  22.0
1
1  Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0
1
2                              Heikkinen, Miss. Laina  female  26.0
0
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0
1
4                            Allen, Mr. William Henry    male  35.0
0
5                                    Moran, Mr. James    male   NaN
0
6                             McCarthy, Mr. Timothy J    male  54.0
0
7                     Palsson, Master. Gosta Leonard    male   2.0
3
8  Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)  female  27.0
0
9                Nasser, Mrs. Nicholas (Adele Achem)  female  14.0
```

```
1
```

```
   Parch            Ticket     Fare Cabin Embarked
0      0         A/5 21171   7.2500   NaN        S
1      0          PC 17599  71.2833   C85        C
2      0  STON/O2. 3101282   7.9250   NaN        S
3      0            113803  53.1000  C123        S
4      0            373450   8.0500   NaN        S
5      0            330877   8.4583   NaN        Q
6      0             17463  51.8625   E46        S
7      1            349909  21.0750   NaN        S
8      2            347742  11.1333   NaN        S
9      0            237736  30.0708   NaN        C
```

Using data aggregation in pandas , Answer the question below:

Show the total amount of passengers who survived and not survived (10pts)

```
#code goes here
df.groupby('Survived')['Survived'].count()
```

```
Survived
0    549
1    342
Name: Survived, dtype: int64
```

Show all male survivors from titanic incident (10 pts)

```
#code goes here
df[df.Sex.isin(['male'])].merge(df[df.Survived.isin([1])])
```

```
     PassengerId  Survived  Pclass                                Name
Sex  \
0             18         1       2     Williams, Mr. Charles Eugene
male
1             22         1       2            Beesley, Mr. Lawrence
male
2             24         1       1     Sloper, Mr. William Thompson
male
3             37         1       3                  Mamee, Mr. Hanna
male
4             56         1       1                Woolner, Mr. Hugh
male
..           ...       ...     ...                               ...
...
104          839         1       3                  Chip, Mr. Chang
male
105          840         1       1              Marechal, Mr. Pierre
male
106          858         1       1           Daly, Mr. Peter Denis
male
```

```
107          870         1       3  Johnson, Master. Harold Theodor
male
108          890         1       1            Behr, Mr. Karl Howell
male

      Age  SibSp  Parch  Ticket     Fare Cabin Embarked
0     NaN      0      0  244373  13.0000   NaN        S
1    34.0      0      0  248698  13.0000   D56        S
2    28.0      0      0  113788  35.5000    A6        S
3     NaN      0      0    2677   7.2292   NaN        C
4     NaN      0      0   19947  35.5000   C52        S
..    ...    ...    ...     ...      ...   ...      ...
104  32.0      0      0    1601  56.4958   NaN        S
105   NaN      0      0   11774  29.7000   C47        C
106  51.0      0      0  113055  26.5500   E17        S
107   4.0      1      1  347742  11.1333   NaN        S
108  26.0      0      0  111369  30.0000  C148        C

[109 rows x 12 columns]
```

Show the average fare paid based on gender (10 pts)

```
#code goes here
df.groupby('Sex')['Fare'].mean()
```

```
Sex
female    44.479818
male      25.523893
Name: Fare, dtype: float64
```

Show the total amount of passengers who survived and not survived based on gender (10pts)

```
#code goes here
df.groupby(['Survived','Sex'])['Survived'].count()
```

```
Survived  Sex
0         female     81
          male      468
1         female    233
          male      109
Name: Survived, dtype: int64
```