Name: Rohan Vinayak Chaudhari

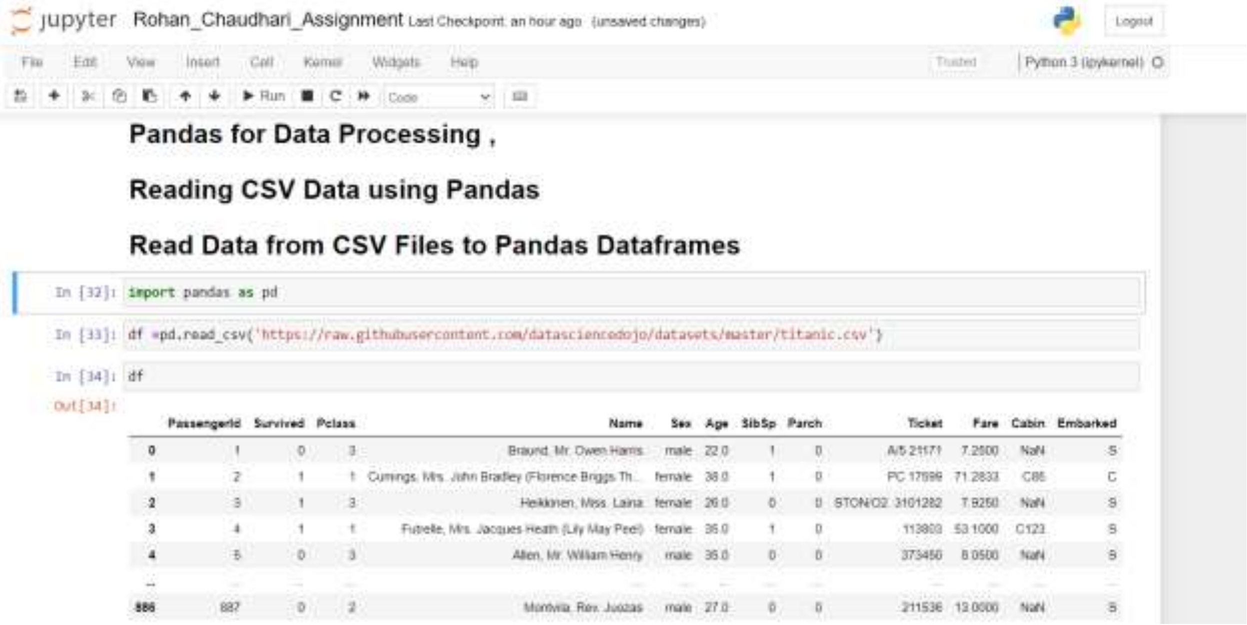
Batch: Data Engineering

Date:01/02/2024

Topic: Python

Solution:

1.Python:



Pandas for Data Processing ,

Reading CSV Data using Pandas

Read Data from CSV Files to Pandas Dataframes

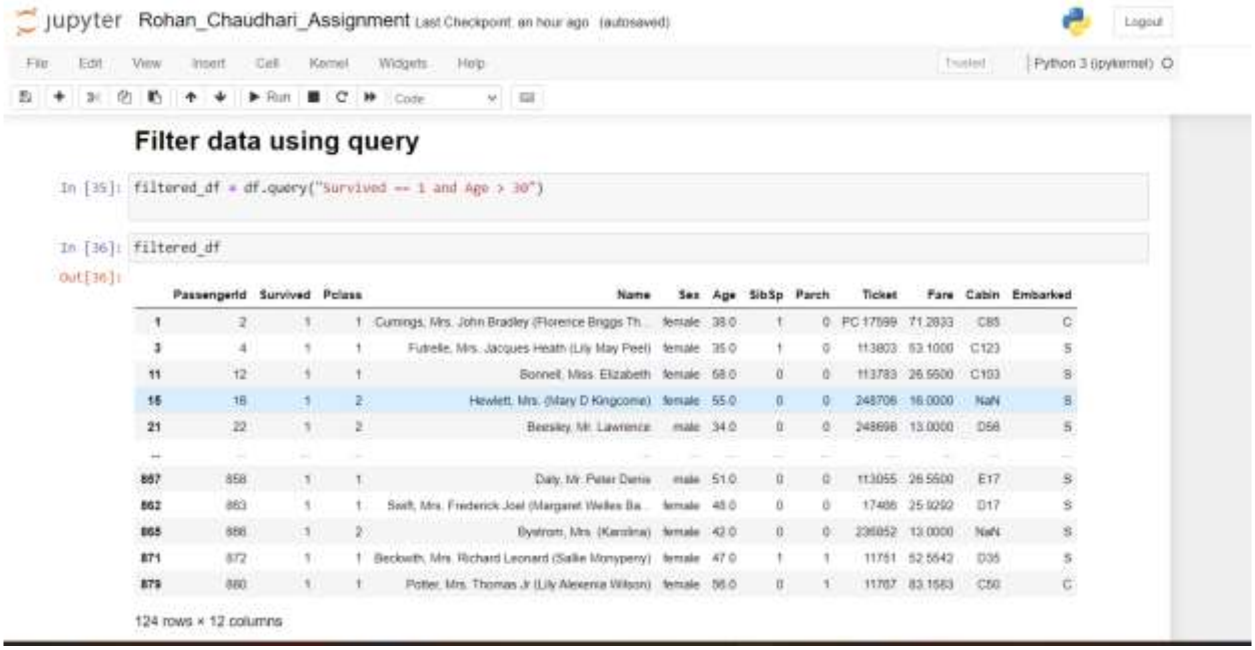
```
In [32]: import pandas as pd
```

```
In [33]: df =pd.read_csv('https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv')
```

```
In [34]: df
```

Out[34]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S


jupyter Rohan_Chaudhari_Assignment Last Checkpoint: an hour ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

Filter data using query

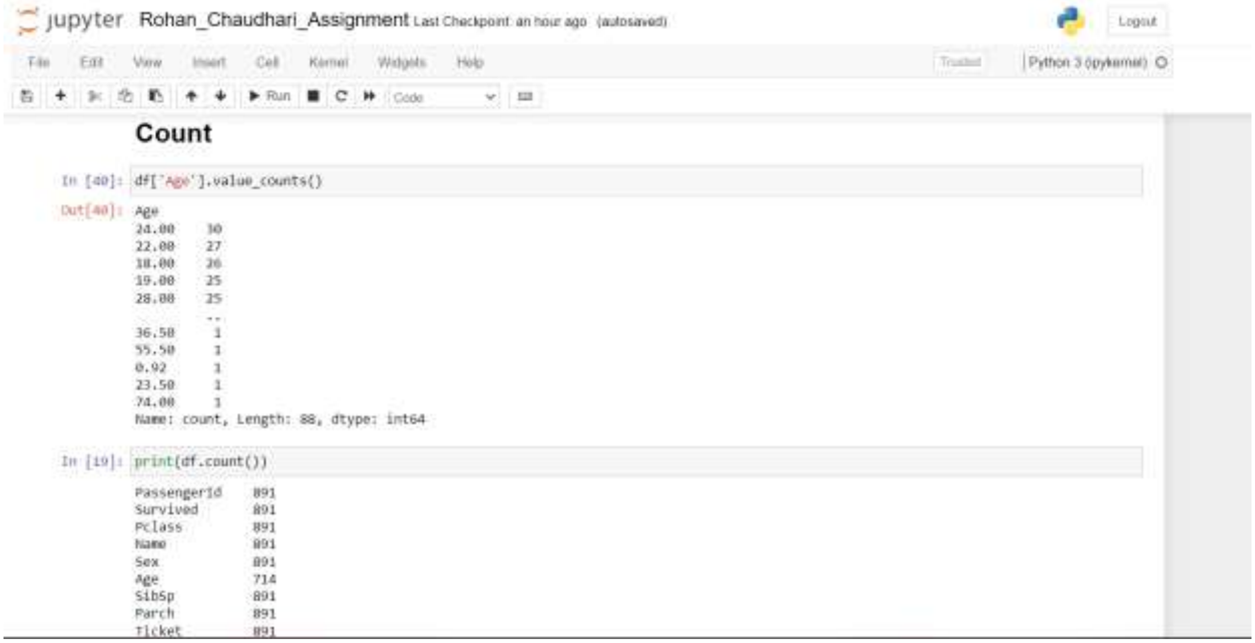
```
In [35]: filtered_df = df.query("Survived == 1 and Age > 30")
```

```
In [36]: filtered_df
```

Out[36]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
11	12	1	1	Bonnell, Miss. Elizabeth	female	58.0	0	0	113783	26.5500	C103	S
15	16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55.0	0	0	248706	16.0000	NaN	S
21	22	1	2	Beesley, Mr. Lawrence	male	34.0	0	0	248698	13.0000	D56	S
...	...	...	...	...	...	...	...	...	...	...	...	...
857	858	1	1	Daly, Mr. Peter Denis	male	51.0	0	0	113055	26.5500	E17	S
862	863	1	1	Swift, Mrs. Frederick Joel (Margaret Welles Ba...	female	48.0	0	0	17466	25.9292	D17	S
865	866	1	2	Bystrom, Mrs. (Karolina)	female	42.0	0	0	236852	13.0000	NaN	S
871	872	1	1	Beckwith, Mrs. Richard Leonard (Sallie Monypeny)	female	47.0	1	1	11751	52.5542	D35	S
879	880	1	1	Potter, Mrs. Thomas Jr (Lily Alexenia Wilson)	female	56.0	0	1	11767	83.1583	C50	C

124 rows × 12 columns


jupyter Rohan_Chaudhari_Assignment Last Checkpoint: an hour ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

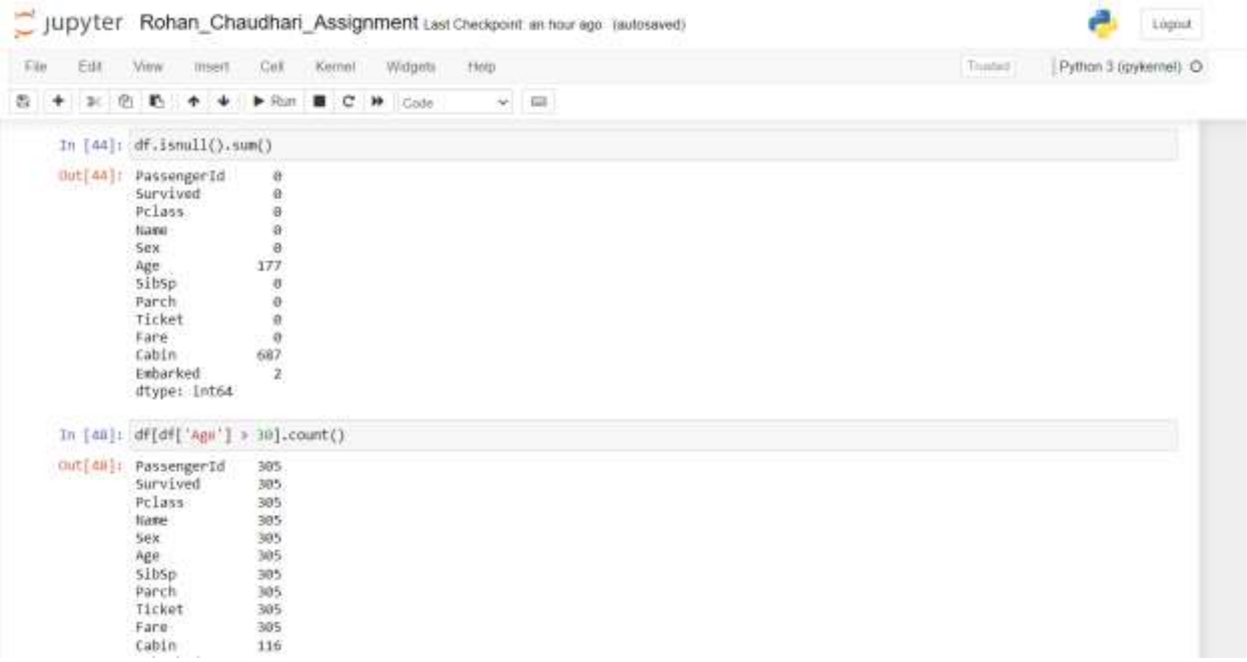
Count

```
In [40]: df['Age'].value_counts()
```

```
Out[40]: Age
24.00    30
22.00    27
18.00    26
19.00    25
28.00    25
         ..
36.50     1
55.50     1
0.92      1
23.50     1
74.00     1
Name: count, Length: 88, dtype: int64
```

```
In [19]: print(df.count())
```

```
PassengerId    891
Survived       891
Pclass         891
Name           891
Sex            891
Age            714
SibSp          891
Parch          891
Ticket         891
```

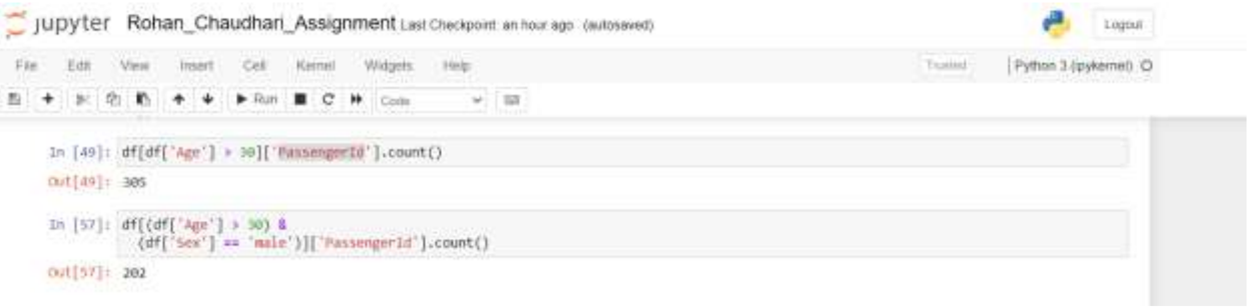

jupyter Rohan_Chaudhari_Assignment Last Checkpoint: an hour ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel) ○
▶ Run
Code

```
In [44]: df.isnull().sum()

Out[44]: PassengerId      0
         Survived         0
         Pclass           0
         Name             0
         Sex              0
         Age            177
         SibSp            0
         Parch            0
         Ticket           0
         Fare             0
         Cabin          687
         Embarked         2
         dtype: int64

In [48]: df[df['Age'] > 30].count()

Out[48]: PassengerId    305
         Survived       305
         Pclass         305
         Name           305
         Sex            305
         Age            305
         SibSp          305
         Parch          305
         Ticket         305
         Fare           305
         Cabin          116
```

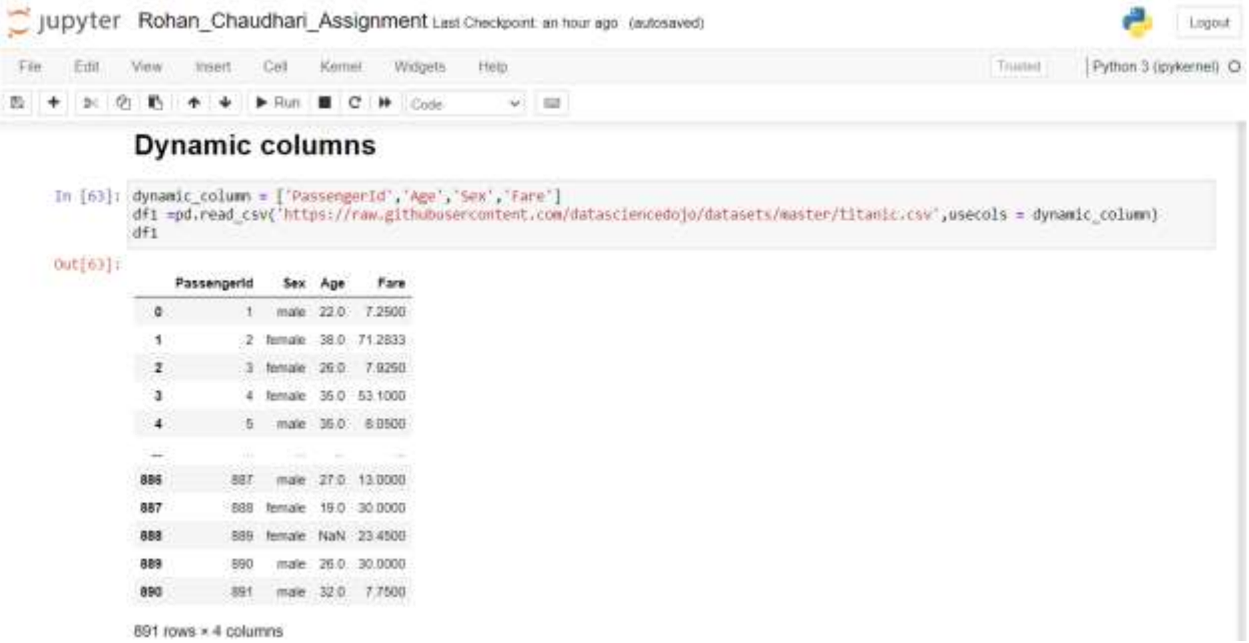

jupyter Rohan_Chaudhari_Assignment Last Checkpoint: an hour ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel) ○
▶ Run
Code

```
In [49]: df[df['Age'] > 30]['PassengerId'].count()

Out[49]: 305

In [57]: df[(df['Age'] > 30) &
           (df['Sex'] == 'male')]['PassengerId'].count()

Out[57]: 202
```



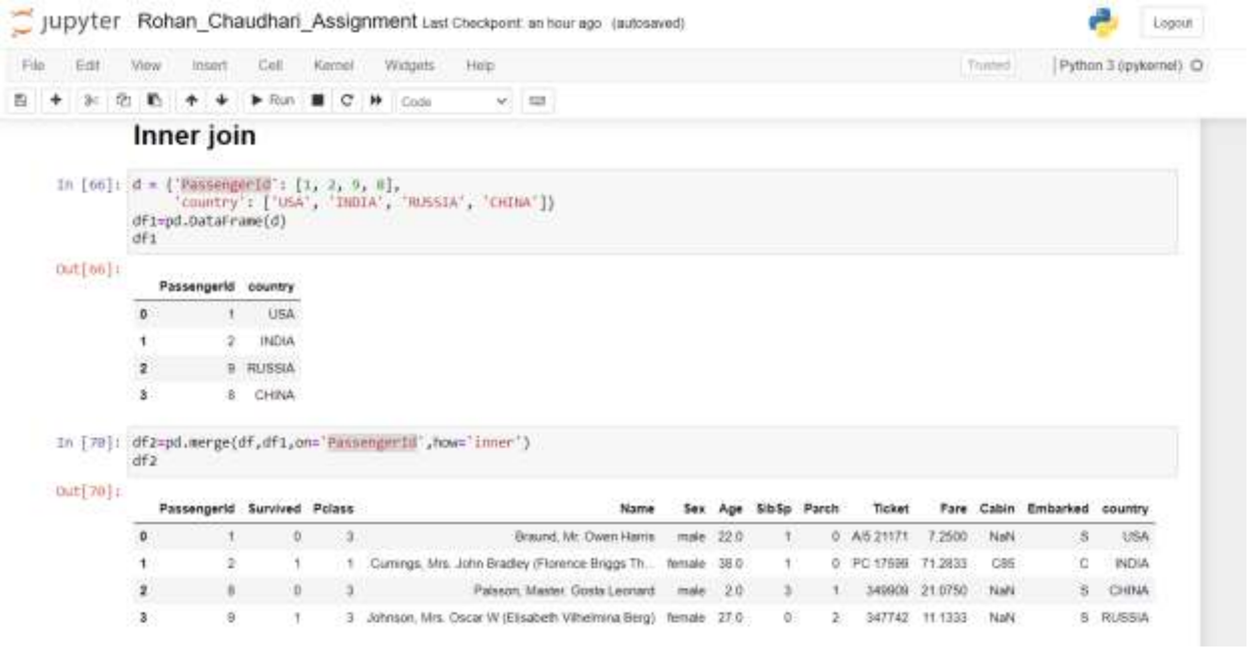
Rohan_Chaudhari_Assignment Last Checkpoint: an hour ago (autosaved)

Dynamic columns

```
In [63]: dynamic_column = ['PassengerId','Age','Sex','Fare']
         df1 =pd.read_csv('https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv',usecols = dynamic_column)
         df1
```

```
Out[63]:
         PassengerId     Sex   Age     Fare
    0              1    male  22.0   7.2500
    1              2  female  38.0  71.2833
    2              3  female  26.0   7.9250
    3              4  female  35.0  53.1000
    4              5    male  35.0   8.0500
  ...            ...     ...   ...      ...
  886            887    male  27.0  13.0000
  887            888  female  19.0  30.0000
  888            889  female   NaN  23.4500
  889            890    male  26.0  30.0000
  890            891    male  32.0   7.7500

891 rows × 4 columns
```



Rohan_Chaudhari_Assignment Last Checkpoint: an hour ago (autosaved)

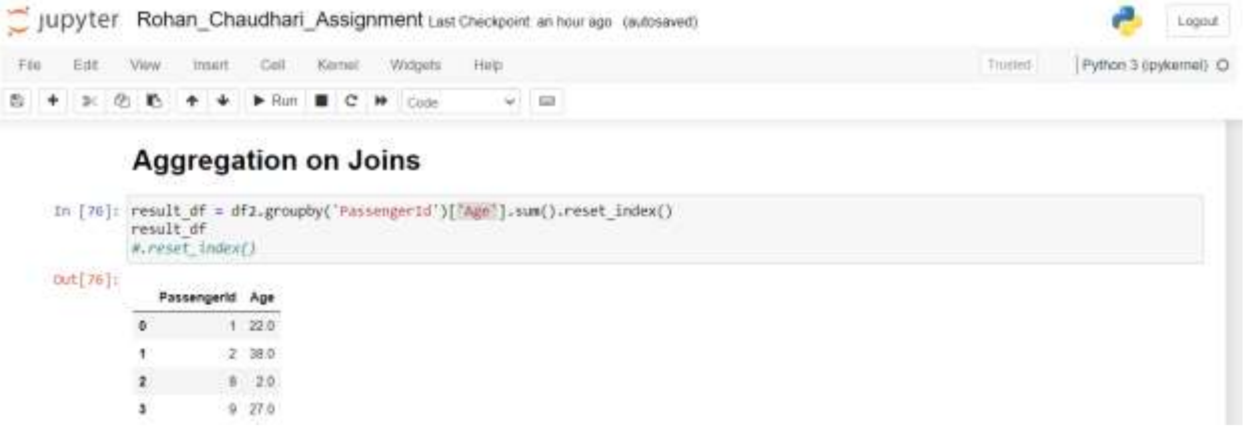
Inner join

```
In [66]: d = {'PassengerId': [1, 2, 9, 8],
              'country': ['USA', 'INDIA', 'RUSSIA', 'CHINA']}
         df1=pd.DataFrame(d)
         df1
```

```
Out[66]:
     PassengerId  country
  0            1      USA
  1            2    INDIA
  2            9   RUSSIA
  3            8    CHINA
```

```
In [70]: df2=pd.merge(df,df1,on='PassengerId',how='inner')
         df2
```

```
Out[70]:
     PassengerId  Survived  Pclass                                              Name     Sex   Age  SibSp  Parch     Ticket     Fare  Cabin  Embarked  country
  0            1         0       3                           Braund, Mr. Owen Harris    male  22.0      1      0  A/5 21171   7.2500    NaN         S      USA
  1            2         1       1  Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0      1      0   PC 17599  71.2833    C85         C    INDIA
  2            8         0       3                    Palsson, Master. Gosta Leonard    male   2.0      3      1     349909  21.0750    NaN         S    CHINA
  3            9         1       3  Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)  female  27.0      0      2     347742  11.1333    NaN         S   RUSSIA
```

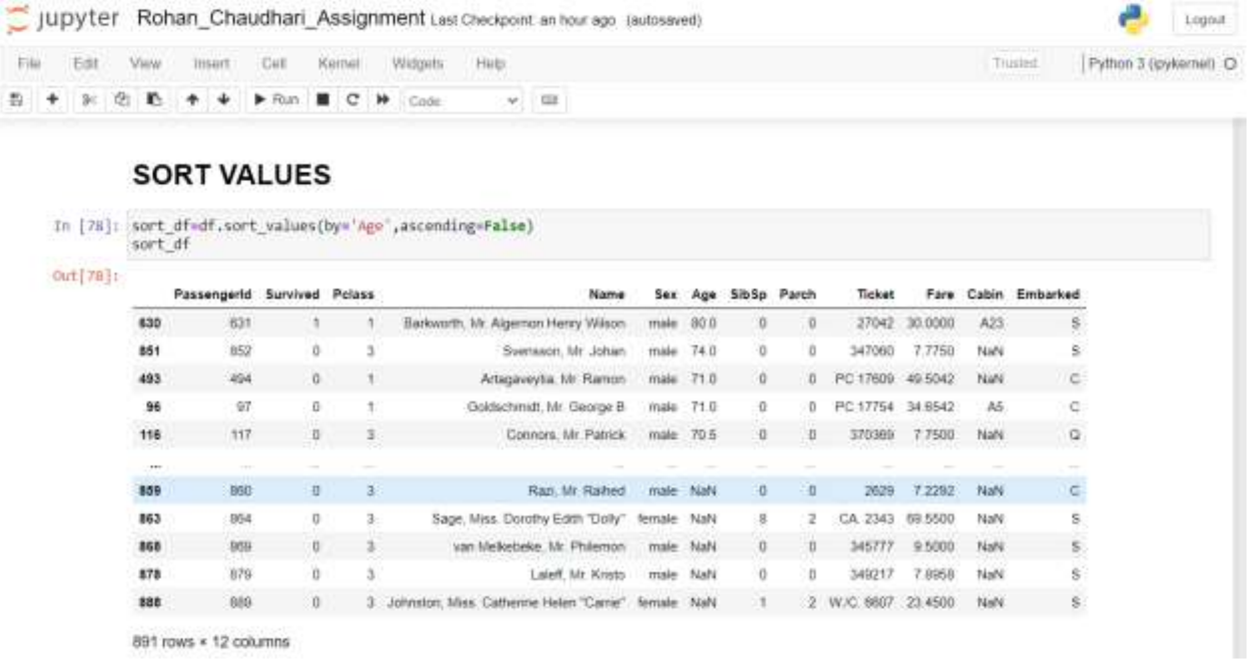


Aggregation on Joins

```
In [76]: result_df = df2.groupby('PassengerId')['Age'].sum().reset_index()
         result_df
         #.reset_index()
```

Out[76]:

	PassengerId	Age
0	1	22.0
1	2	38.0
2	8	2.0
3	9	27.0



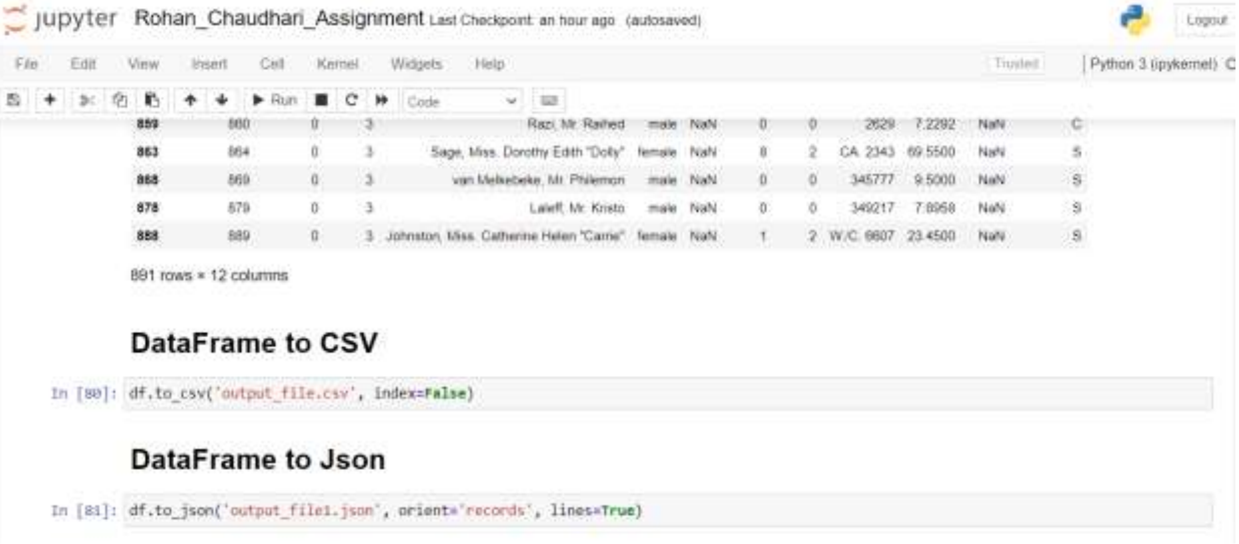
SORT VALUES

```
In [78]: sort_df=df.sort_values(by='Age',ascending=False)
         sort_df
```

Out[78]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
630	631	1	1	Barkworth, Mr. Algernon Henry Wilson	male	80.0	0	0	27042	30.0000	A23	S
851	852	0	3	Svensson, Mr. Johan	male	74.0	0	0	347060	7.7750	NaN	S
493	494	0	1	Artagaveytia, Mr. Ramon	male	71.0	0	0	PC 17609	49.5042	NaN	C
96	97	0	1	Goldschmidt, Mr. George B	male	71.0	0	0	PC 17754	34.6542	A5	C
116	117	0	3	Connors, Mr. Patrick	male	70.5	0	0	370369	7.7500	NaN	Q
...	...	...	...	...	...	...	...	...	...	...	...	...
859	860	0	3	Razi, Mr. Raihed	male	NaN	0	0	2629	7.2292	NaN	C
863	864	0	3	Sage, Miss. Dorothy Edith "Dolly"	female	NaN	8	2	CA. 2343	69.5500	NaN	S
868	869	0	3	van Melkebeke, Mr. Philemon	male	NaN	0	0	345777	9.5000	NaN	S
878	879	0	3	Laleff, Mr. Kristo	male	NaN	0	0	349217	7.8958	NaN	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S

891 rows × 12 columns



859	860	0	3	Razi, Mr. Raihed	male	NaN	0	0	2629	7.2292	NaN	C
863	864	0	3	Sage, Miss. Dorothy Edith "Dolly"	female	NaN	8	2	CA. 2343	69.5500	NaN	S
868	869	0	3	van Melkebeke, Mr. Philemon	male	NaN	0	0	345777	9.5000	NaN	S
878	879	0	3	Laleff, Mr. Kristo	male	NaN	0	0	349217	7.8958	NaN	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S

891 rows × 12 columns

DataFrame to CSV

```
In [80]: df.to_csv('output_file.csv', index=False)
```

DataFrame to Json

```
In [81]: df.to_json('output_file1.json', orient='records', lines=True)
```