EXPERIMENT 1.1

Name- SANSKAR AGRAWAL **UID- 20BCS5914**
Branch- CSE **Section/Group- 806 B**
Semester- 5th **Date of Performance-28/08/2022**
Subject Name_ Machine Learning Lab
Subject Code- 20CSP-317

AIM -EXPLORATORY DATA ANALYSIS (EDA).

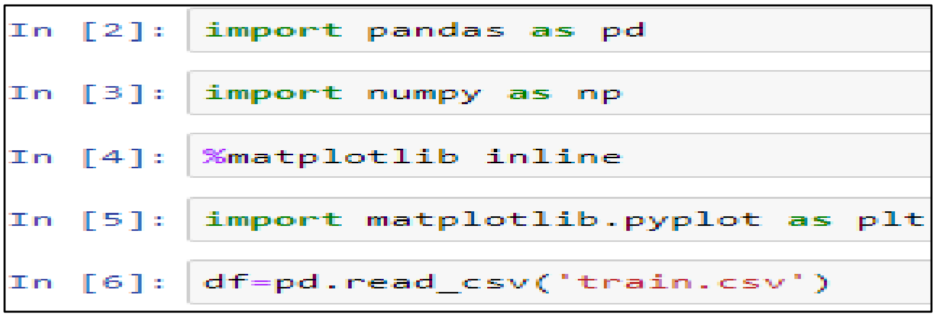
OBJECTIVE –To Understand the data i.e., Data is clean , it doesn’t have any null values , missing values , remove noise , identify variables in dataset and relationship between variables to conclude the values.

S/W Requirement: - VS Code or Jupyter Notebook

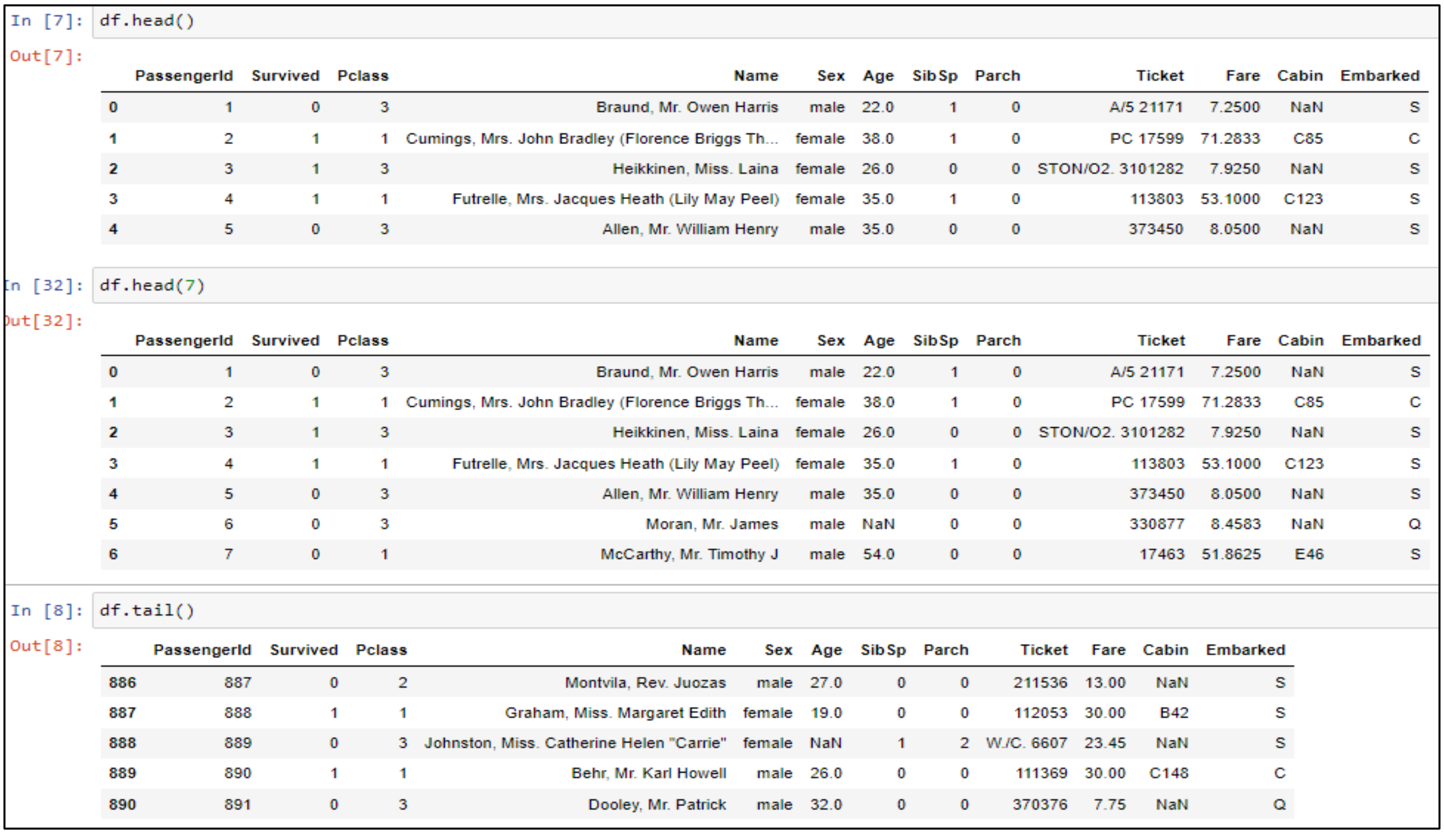
INPUT AND OUTPUT –

Importing Libraries: -

```
import pandas as pd
import numpy as np
%matplotlib inline
import matplotlib.pyplot as plt
df = pd.read_csv('train.csv')
df.head()
df.head(7)
df.tail()
df.info()
df.describe()
```



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

In [3]: import numpy as np

In [4]: %matplotlib inline

In [5]: import matplotlib.pyplot as plt

In [6]: df=pd.read_csv('train.csv')
```



In [7]: `df.head()`

Out[7]:

	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

In [32]: `df.head(7)`

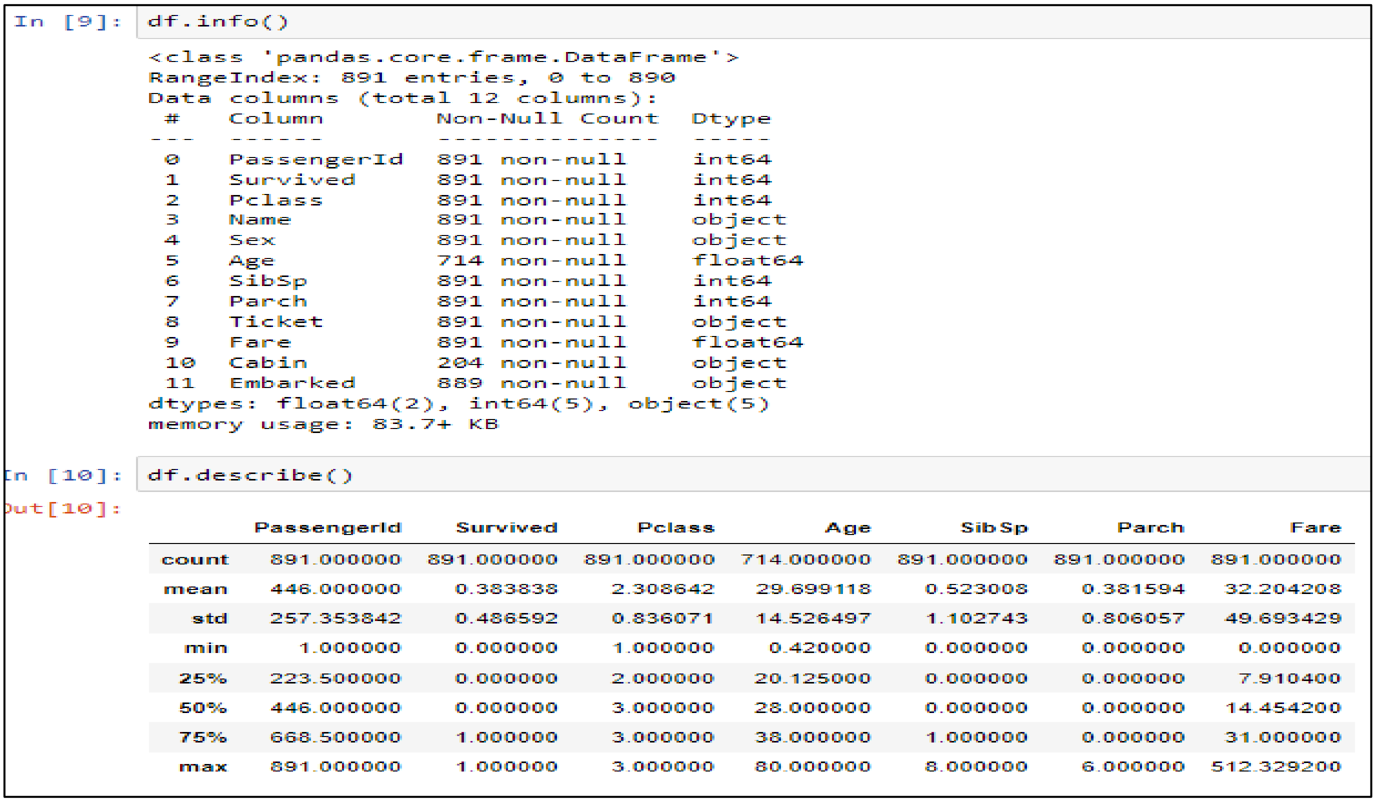
Out[32]:

	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
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S

In [8]: `df.tail()`

Out[8]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.00	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.00	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.45	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.00	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.75	NaN	Q



In [9]: `df.info()`

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  891 non-null    int64  
 1   Survived     891 non-null    int64  
 2   Pclass       891 non-null    int64  
 3   Name         891 non-null    object 
 4   Sex          891 non-null    object 
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64  
 7   Parch        891 non-null    int64  
 8   Ticket       891 non-null    object 
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object 
 11  Embarked     889 non-null    object 
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

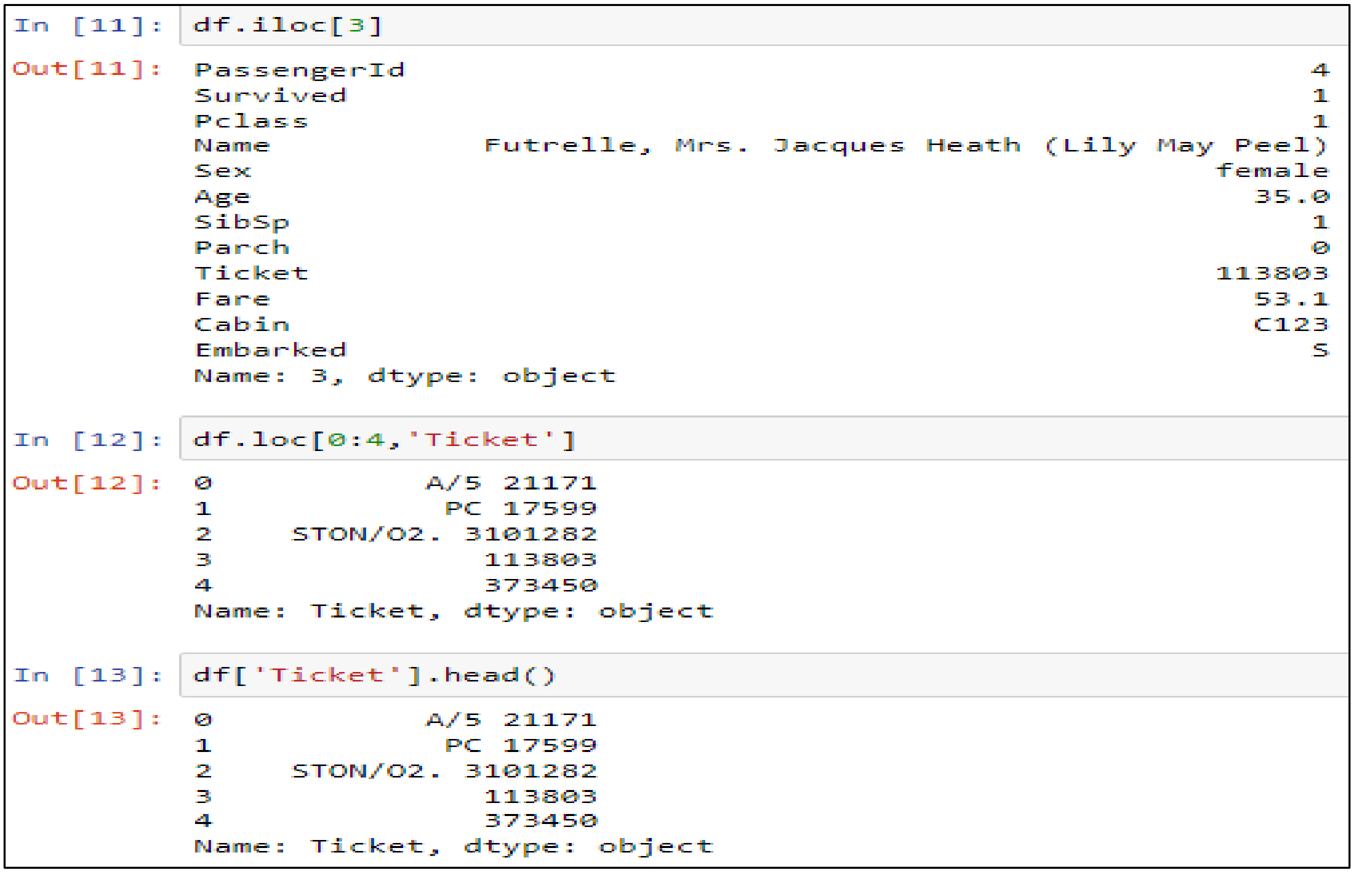
In [10]: `df.describe()`

Out[10]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

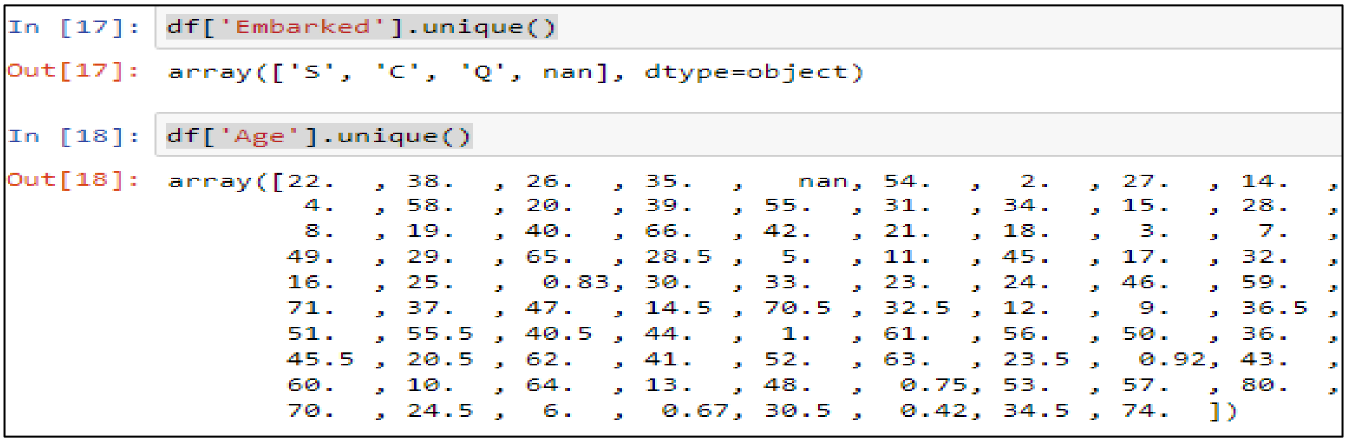
Indexing: -

```
df.iloc[3]
df.loc[0:4,'Ticket']
df['Ticket'].head()
```



```
In [11]: df.iloc[3]

Out[11]: PassengerId                                                  4
         Survived                                                     1
         Pclass                                                       1
         Name           Futrelle, Mrs. Jacques Heath (Lily May Peel)
         Sex                                                     female
         Age                                                       35.0
         SibSp                                                        1
         Parch                                                        0
         Ticket                                                  113803
         Fare                                                      53.1
         Cabin                                                     C123
         Embarked                                                     S
         Name: 3, dtype: object

In [12]: df.loc[0:4,'Ticket']

Out[12]: 0           A/5 21171
         1            PC 17599
         2    STON/O2. 3101282
         3              113803
         4              373450
         Name: Ticket, dtype: object

In [13]: df['Ticket'].head()

Out[13]: 0           A/5 21171
         1            PC 17599
         2    STON/O2. 3101282
         3              113803
         4              373450
         Name: Ticket, dtype: object
```

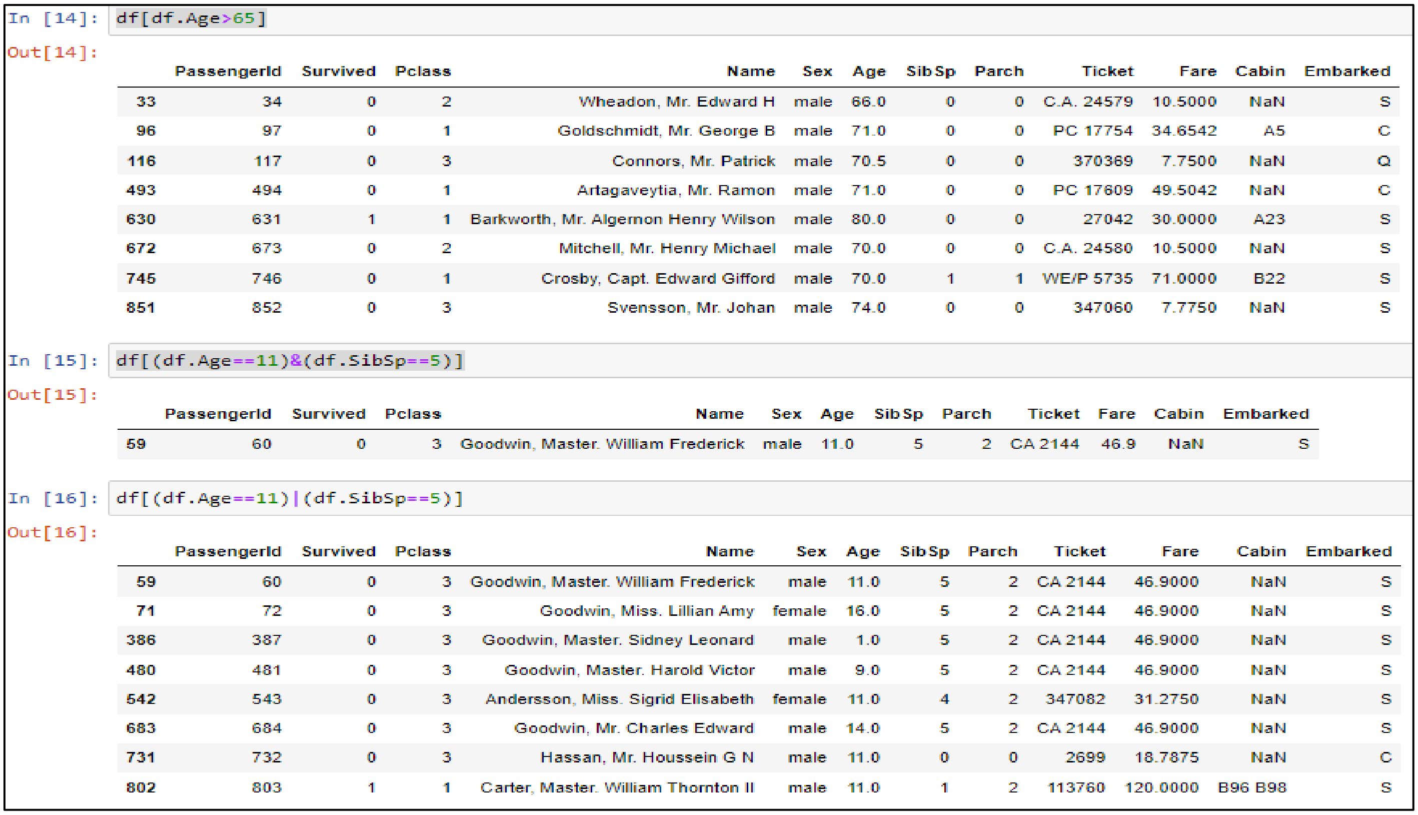
Distinct Elements: -



```
In [17]: df['Embarked'].unique()

Out[17]: array(['S', 'C', 'Q', nan], dtype=object)

In [18]: df['Age'].unique()

Out[18]: array([22.  , 38.  , 26.  , 35.  ,   nan, 54.  ,  2.  , 27.  , 14.  ,
                 4.  , 58.  , 20.  , 39.  , 55.  , 31.  , 34.  , 15.  , 28.  ,
                 8.  , 19.  , 40.  , 66.  , 42.  , 21.  , 18.  ,  3.  ,  7.  ,
                49.  , 29.  , 65.  , 28.5 ,  5.  , 11.  , 45.  , 17.  , 32.  ,
                16.  , 25.  ,  0.83, 30.  , 33.  , 23.  , 24.  , 46.  , 59.  ,
                71.  , 37.  , 47.  , 14.5 , 70.5 , 32.5 , 12.  ,  9.  , 36.5 ,
                51.  , 55.5 , 40.5 , 44.  ,  1.  , 61.  , 56.  , 50.  , 36.  ,
                45.5 , 20.5 , 62.  , 41.  , 52.  , 63.  , 23.5 ,  0.92, 43.  ,
                60.  , 10.  , 64.  , 13.  , 48.  ,  0.75, 53.  , 57.  , 80.  ,
                70.  , 24.5 ,  6.  ,  0.67, 30.5 ,  0.42, 34.5 , 74.  ])
```

Selections: -

df[df.Age>65]
df[(df.Age==11)&(df.SibSp==5)]
df[(df.Age==11)|(df.SibSp==5)]



```
In [14]: df[df.Age>65]
Out[14]:
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
33	34	0	2	Wheadon, Mr. Edward H	male	66.0	0	0	C.A. 24579	10.5000	NaN	S
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
493	494	0	1	Artagaveytia, Mr. Ramon	male	71.0	0	0	PC 17609	49.5042	NaN	C
630	631	1	1	Barkworth, Mr. Algernon Henry Wilson	male	80.0	0	0	27042	30.0000	A23	S
672	673	0	2	Mitchell, Mr. Henry Michael	male	70.0	0	0	C.A. 24580	10.5000	NaN	S
745	746	0	1	Crosby, Capt. Edward Gifford	male	70.0	1	1	WE/P 5735	71.0000	B22	S
851	852	0	3	Svensson, Mr. Johan	male	74.0	0	0	347060	7.7750	NaN	S

```
In [15]: df[(df.Age==11)&(df.SibSp==5)]
Out[15]:
```

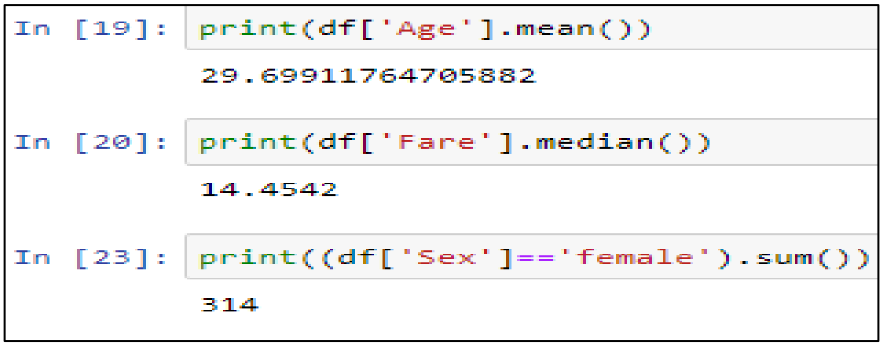
	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
59	60	0	3	Goodwin, Master. William Frederick	male	11.0	5	2	CA 2144	46.9	NaN	S

```
In [16]: df[(df.Age==11)|(df.SibSp==5)]
Out[16]:
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
59	60	0	3	Goodwin, Master. William Frederick	male	11.0	5	2	CA 2144	46.9000	NaN	S
71	72	0	3	Goodwin, Miss. Lillian Amy	female	16.0	5	2	CA 2144	46.9000	NaN	S
386	387	0	3	Goodwin, Master. Sidney Leonard	male	1.0	5	2	CA 2144	46.9000	NaN	S
480	481	0	3	Goodwin, Master. Harold Victor	male	9.0	5	2	CA 2144	46.9000	NaN	S
542	543	0	3	Andersson, Miss. Sigrid Elisabeth	female	11.0	4	2	347082	31.2750	NaN	S
683	684	0	3	Goodwin, Mr. Charles Edward	male	14.0	5	2	CA 2144	46.9000	NaN	S
731	732	0	3	Hassan, Mr. Houssein G N	male	11.0	0	0	2699	18.7875	NaN	C
802	803	1	1	Carter, Master. William Thornton II	male	11.0	1	2	113760	120.0000	B96 B98	S

Missing values find and treatment: -

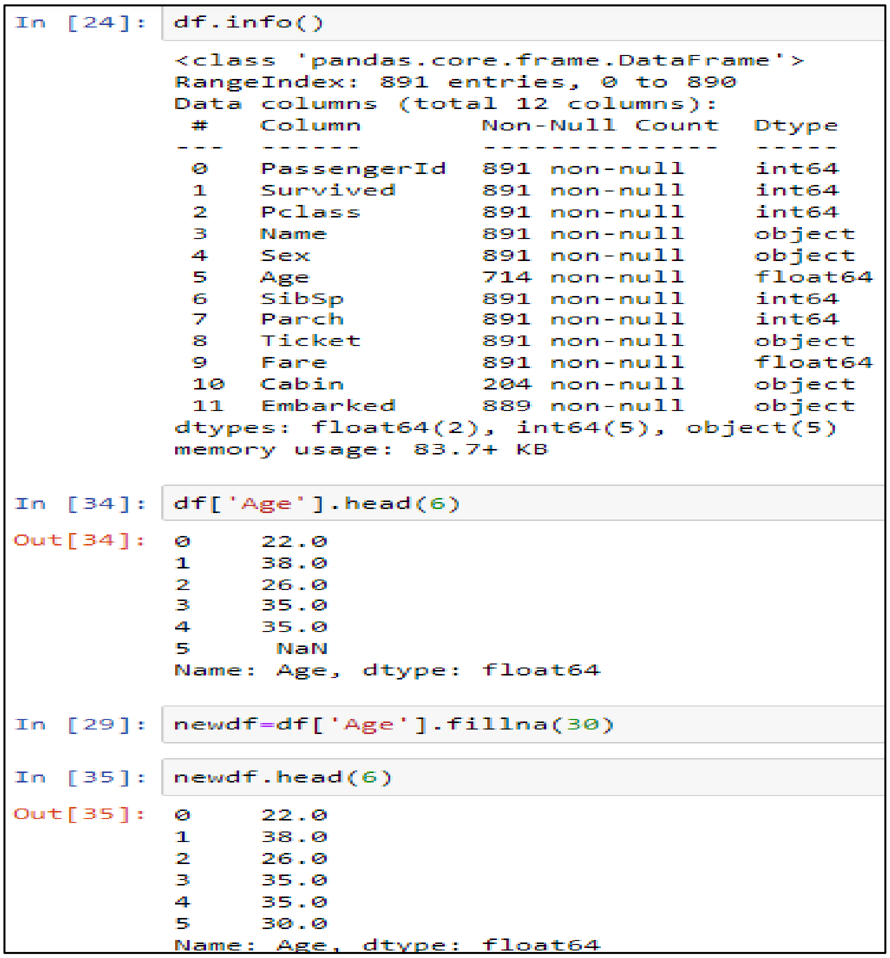
print(df['Age'].mean())
print(df['Fare'].median())
print((df['Sex']=='female').sum())



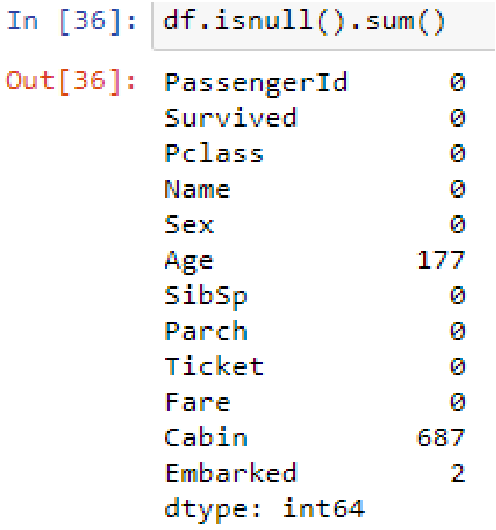
```
In [19]: print(df['Age'].mean())
         29.69911764705882

In [20]: print(df['Fare'].median())
         14.4542

In [23]: print((df['Sex']=='female').sum())
         314
```

Missing Data: -

df[df.Age>65]
df[(df.Age==11)&(df.SibSp==5)]
df[(df.Age==11)|(df.SibSp==5)]



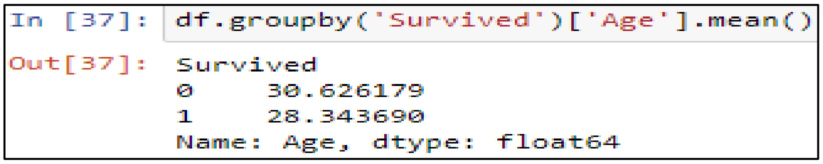
```
In [24]: df.info()

         <class 'pandas.core.frame.DataFrame'>
         RangeIndex: 891 entries, 0 to 890
         Data columns (total 12 columns):
          #   Column       Non-Null Count  Dtype
         ---  ------       --------------  -----
          0   PassengerId  891 non-null    int64
          1   Survived     891 non-null    int64
          2   Pclass       891 non-null    int64
          3   Name         891 non-null    object
          4   Sex          891 non-null    object
          5   Age          714 non-null    float64
          6   SibSp        891 non-null    int64
          7   Parch        891 non-null    int64
          8   Ticket       891 non-null    object
          9   Fare         891 non-null    float64
          10  Cabin        204 non-null    object
          11  Embarked     889 non-null    object
         dtypes: float64(2), int64(5), object(5)
         memory usage: 83.7+ KB

In [34]: df['Age'].head(6)

Out[34]: 0    22.0
         1    38.0
         2    26.0
         3    35.0
         4    35.0
         5     NaN
         Name: Age, dtype: float64

In [29]: newdf=df['Age'].fillna(30)

In [35]: newdf.head(6)

Out[35]: 0    22.0
         1    38.0
         2    26.0
         3    35.0
         4    35.0
         5    30.0
         Name: Age, dtype: float64
```



```
In [36]: df.isnull().sum()

Out[36]: 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
```

Groupby: -

df[df.Age>65]

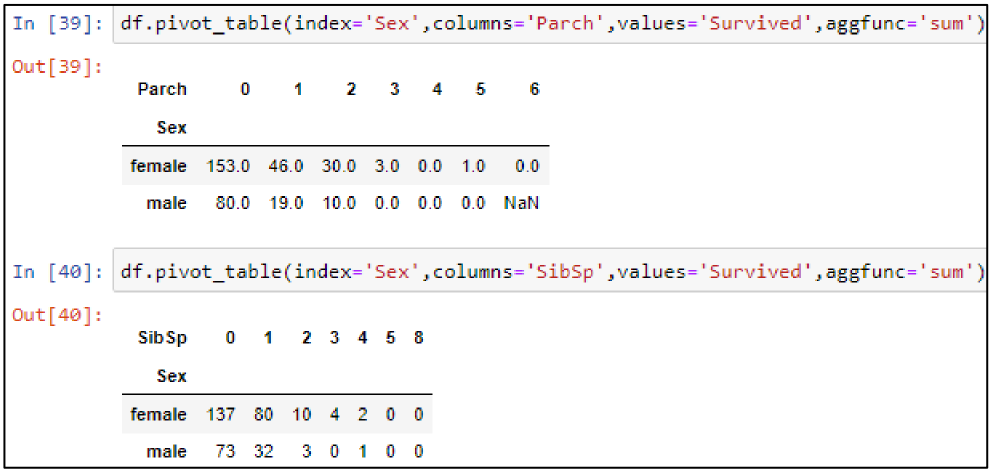


```
In [37]: df.groupby('Survived')['Age'].mean()
Out[37]: Survived
         0    30.626179
         1    28.343690
         Name: Age, dtype: float64
```

Missing Data: -

df. pivot_table(index='Sex’, columns='Parch',values='Survived',aggfunc='sum')

df. pivot_table(index='Sex’, columns='SibSp',values='Survived',aggfunc='sum')



```
In [39]: df.pivot_table(index='Sex',columns='Parch',values='Survived',aggfunc='sum')
Out[39]:
```

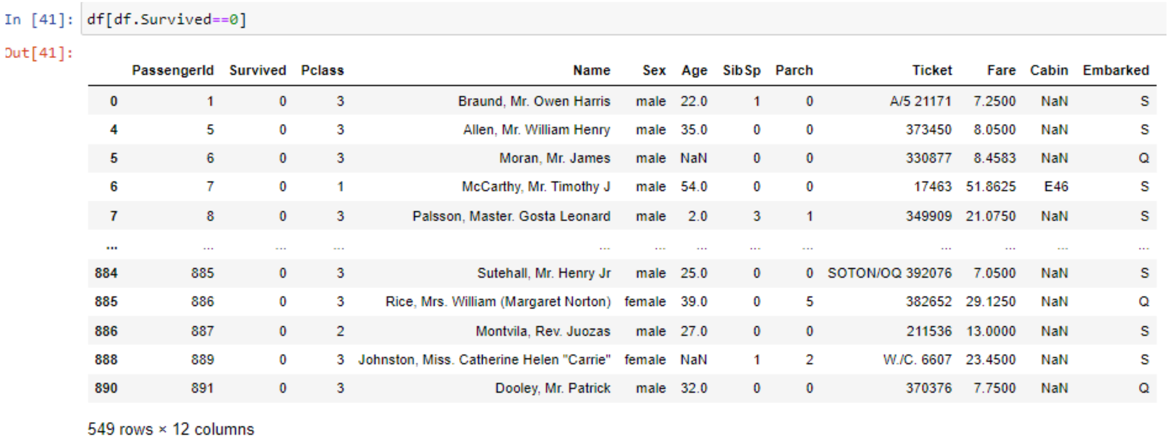
Parch	0	1	2	3	4	5	6
Sex							
female	153.0	46.0	30.0	3.0	0.0	1.0	0.0
male	80.0	19.0	10.0	0.0	0.0	0.0	NaN

```
In [40]: df.pivot_table(index='Sex',columns='SibSp',values='Survived',aggfunc='sum')
Out[40]:
```

SibSp	0	1	2	3	4	5	8
Sex							
female	137	80	10	4	2	0	0
male	73	32	3	0	1	0	0

Exercises:

select passengers that died

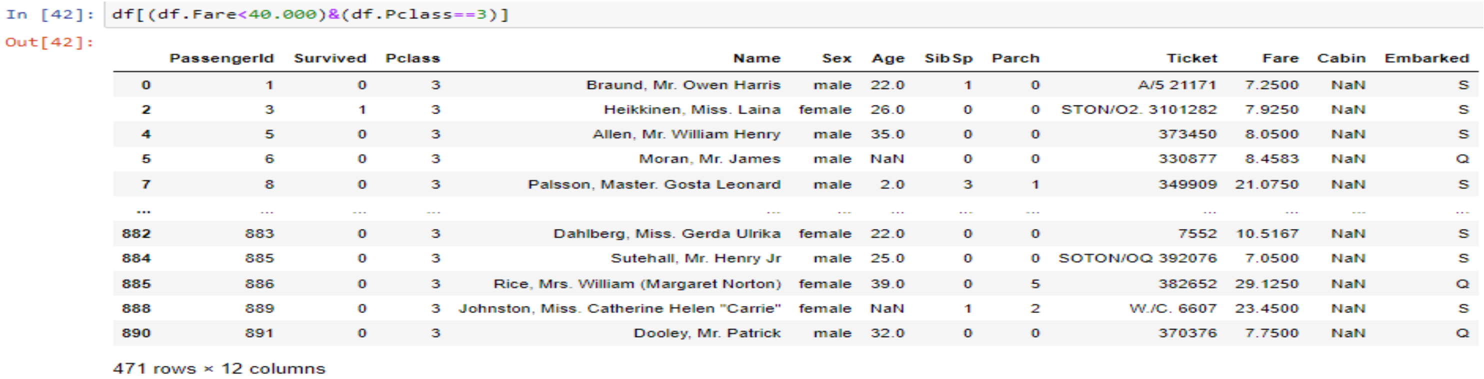


```
In [41]: df[df.Survived==0]
Out[41]:
```

	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
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
884	885	0	3	Sutehall, Mr. Henry Jr	male	25.0	0	0	SOTON/OQ 392076	7.0500	NaN	S
885	886	0	3	Rice, Mrs. William (Margaret Norton)	female	39.0	0	5	382652	29.1250	NaN	Q
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

549 rows × 12 columns

select passengers who paid less than 40.000 and were in third class

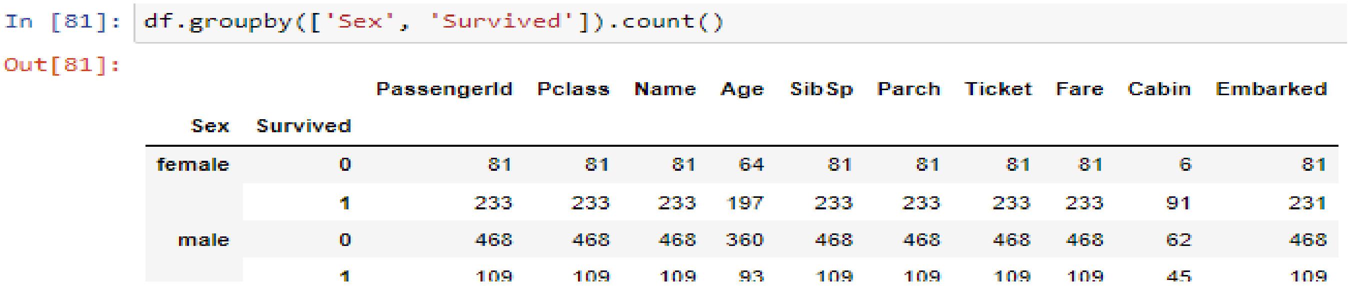


```
In [42]: df[(df.Fare<40.000)&(df.Pclass==3)]
Out[42]:
```

	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
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
882	883	0	3	Dahlberg, Miss. Gerda Ulrika	female	22.0	0	0	7552	10.5167	NaN	S
884	885	0	3	Sutehall, Mr. Henry Jr	male	25.0	0	0	SOTON/OQ 392076	7.0500	NaN	S
885	886	0	3	Rice, Mrs. William (Margaret Norton)	female	39.0	0	5	382652	29.1250	NaN	Q
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

471 rows × 12 columns

count the number of survived and dead per each gender



```
In [81]: df.groupby(['Sex', 'Survived']).count()
Out[81]:
```

		PassengerId	Pclass	Name	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
Sex	Survived										
female	0	81	81	81	64	81	81	81	81	6	81
	1	233	233	233	197	233	233	233	233	91	231
male	0	468	468	468	360	468	468	468	468	62	468
	1	109	109	109	93	109	109	109	109	45	109

Learning outcomes (What I have learnt) -

1. Identify the faulty points so that we can clean the data.
2. How to deal with missing values of variables (Columns) in dataset.
3. To Deal with Outliers.
4. To find Relationship between different variables and map different type of Graphs.

Evaluation Grid (To be created as per the SOP and Assessment guidelines by the faculty):

Sr. No.	Parameters	Marks Obtained	Maximum Marks
1.			
2.			
3.			
4.			