In [6]:
```python
import pandas as pd
```

In [7]:
```python
pd.read_csv?
```

In [8]:
```python
titanic_df = pd.read_csv('titanic1.csv')
```

In [9]:
```python
type(titanic_df)
```

Out[9]: pandas.core.frame.DataFrame

In [10]:
```python
titanic_df.head
```

Out[10]:
```
<bound method NDFrame.head of      PassengerId  Survived  Pclass  \
0              1         0       3
1              2         1       1
2              3         1       3
3              4         1       1
4              5         0       3
..           ...       ...     ...
886          887         0       2
887          888         1       1
888          889         0       3
889          890         1       1
890          891         0       3

                                                  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
..                                                 ...     ...   ...    ...
886                              Montvila, Rev. Juozas    male  27.0      0
887                       Graham, Miss. Margaret Edith  female  19.0      0
888           Johnston, Miss. Catherine Helen "Carrie"  female   NaN      1
889                              Behr, Mr. Karl Howell    male  26.0      0
890                                Dooley, Mr. Patrick    male  32.0      0

     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
..     ...               ...      ...   ...      ...
886      0            211536  13.0000   NaN        S
887      0            112053  30.0000   B42        S
888      2        W./C. 6607  23.4500   NaN        S
889      0            111369  30.0000  C148        C
890      0            370376   7.7500   NaN        Q

[891 rows x 12 columns]>
```

In [11]:
```python
titanic_df.describe(include=['object'])
```

Out[11]:

	Name	Sex	Ticket	Cabin	Embarked
count	891	891	891	204	889
unique	891	2	681	147	3
top	Braund, Mr. Owen Harris	male	347082	B96 B98	S
freq	1	577	7	4	644

In [12]:
```
titanic_df.describe()
```

Out[12]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	747.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.751232	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.318104	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	21.000000	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

In [13]:
```
titanic_df.shape
```

Out[13]: (891, 12)

In [14]:
```
titanic_df.iloc[:,:]
```

Out[14]:

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

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin
				Henry							
...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN

891 rows × 12 columns

In [15]:
```
titanic_df.iloc[0:5,0:3]
```

Out[15]:

	PassengerId	Survived	Pclass
0	1	0	3
1	2	1	1
2	3	1	3
3	4	1	1
4	5	0	3

In [16]:
```
titanic_df.iloc[:4,:3]
```

Out[16]:

	PassengerId	Survived	Pclass
0	1	0	3
1	2	1	1
2	3	1	3
3	4	1	1

In [17]:
```
titanic_df.iloc[-1:,:-2]
```

Out[17]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.75

In [18]:
```python
titanic_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          747 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 [19]:
```python
titanic_df.dtypes
```

Out[19]:
```
PassengerId      int64
Survived         int64
Pclass           int64
Name            object
Sex             object
Age            float64
SibSp            int64
Parch            int64
Ticket          object
Fare           float64
Cabin           object
Embarked        object
dtype: object
```

In [20]:
```python
titanic_df['Survived'][0:10]
```

Out[20]:
```
0    0
1    1
2    1
3    1
4    0
5    0
6    0
7    0
8    1
9    1
Name: Survived, dtype: int64
```

In [21]:
```python
titanic_df.Survived.head()
titanic_df['Survived'].head()
```

Out[21]:
```
0    0
1    1
2    1
3    1
4    0
Name: Survived, dtype: int64
```

In [22]:
```python
titanic_df['Survived'][0:5]
```

Out[22]:
```
0    0
1    1
2    1
3    1
4    0
Name: Survived, dtype: int64
```

In [23]:
```python
titanic_df.Survived.value_counts(normalize=True)
titanic_df.Survived.value_counts(normalize=False)
```

Out[23]:
```
0    549
1    342
Name: Survived, dtype: int64
```

In [24]:
```python
titanic_df['Survived'].value_counts()
```

Out[24]:
```
0    549
1    342
Name: Survived, dtype: int64
```

In [25]:
```python
titanic_df.Survived.value_counts(normalize=True)
titanic_df['Survived'].value_counts(normalize=True)
```

Out[25]:
```
0    0.616162
1    0.383838
Name: Survived, dtype: float64
```

In [26]:
```python
titanic_df[['Survived','Age']].iloc[:5,:]
```

Out[26]:

	Survived	Age
0	0	22.0
1	1	38.0
2	1	26.0
3	1	35.0
4	0	35.0

In [27]:
```python
pd.crosstab(titanic_df.Sex,
            titanic_df.Survived,margins=True)
```

Out[27]:

Survived	0	1	All
Sex			
female	81	233	314
male	468	109	577

Survived	0	1	All
Sex			
All	549	342	891

In [28]:
```python
pd.crosstab(titanic_df.Sex,
            titanic_df.Survived,margins=False)
```

Out[28]:

Survived	0	1
Sex		
female	81	233
male	468	109

In [29]:
```python
pd.crosstab(titanic_df.Sex,
            titanic_df.Survived,normalize='columns')
```

Out[29]:

Survived	0	1
Sex		
female	0.147541	0.681287
male	0.852459	0.318713

In [30]:
```python
pd.crosstab(titanic_df.Sex,
            titanic_df.Survived,normalize="index")
```

Out[30]:

Survived	0	1
Sex		
female	0.257962	0.742038
male	0.811092	0.188908

In [31]:
```python
pd.crosstab(titanic_df.Sex,
            titanic_df.Survived,normalize='all')
```

Out[31]:

Survived	0	1
Sex		
female	0.090909	0.261504
male	0.525253	0.122334

In [32]:
```python
titanic_df[(titanic_df.Age <=5)].shape
```

Out[32]: (44, 12)

In [33]:
```python
titanic_df[(titanic_df.Age <=5)]
```

Out[33]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cal
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.00	3	1	349909	21.0750	N
10	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.00	1	1	PP 9549	16.7000	
16	17	0	3	Rice, Master. Eugene	male	2.00	4	1	382652	29.1250	N
43	44	1	2	Laroche, Miss. Simonne Marie Anne Andree	female	3.00	1	2	SC/Paris 2123	41.5792	N
58	59	1	2	West, Miss. Constance Mirium	female	5.00	1	2	C.A. 34651	27.7500	N
63	64	0	3	Skoog, Master. Harald	male	4.00	3	2	347088	27.9000	N
78	79	1	2	Caldwell, Master. Alden Gates	male	0.83	0	2	248738	29.0000	N
119	120	0	3	Andersson, Miss. Ellis Anna Maria	female	2.00	4	2	347082	31.2750	N
164	165	0	3	Panula, Master. Eino Viljami	male	1.00	4	1	3101295	39.6875	N
171	172	0	3	Rice, Master. Arthur	male	4.00	4	1	382652	29.1250	N
172	173	1	3	Johnson, Miss. Eleanor Ileen	female	1.00	1	1	347742	11.1333	N
183	184	1	2	Becker, Master. Richard F	male	1.00	2	1	230136	39.0000	
184	185	1	3	Kink-Heilmann, Miss. Luise Gretchen	female	4.00	0	2	315153	22.0250	N
193	194	1	2	Navratil, Master. Michel M	male	3.00	1	1	230080	26.0000	
205	206	0	3	Strom, Miss. Telma Matilda	female	2.00	0	1	347054	10.4625	

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cal
233	234	1	3	Asplund, Miss. Lillian Gertrud	female	5.00	4	2	347077	31.3875	N
261	262	1	3	Asplund, Master. Edvin Rojj Felix	male	3.00	4	2	347077	31.3875	N
297	298	0	1	Allison, Miss. Helen Loraine	female	2.00	1	2	113781	151.5500	((
305	306	1	1	Allison, Master. Hudson Trevor	male	0.92	1	2	113781	151.5500	((
340	341	1	2	Navratil, Master. Edmond Roger	male	2.00	1	1	230080	26.0000	
348	349	1	3	Coutts, Master. William Loch "William"	male	3.00	1	1	C.A. 37671	15.9000	N
374	375	0	3	Palsson, Miss. Stina Viola	female	3.00	3	1	349909	21.0750	N
381	382	1	3	Nakid, Miss. Maria ("Mary")	female	1.00	0	2	2653	15.7417	N
386	387	0	3	Goodwin, Master. Sidney Leonard	male	1.00	5	2	CA 2144	46.9000	N
407	408	1	2	Richards, Master. William Rowe	male	3.00	1	1	29106	18.7500	N
445	446	1	1	Dodge, Master. Washington	male	4.00	0	2	33638	81.8583	/
448	449	1	3	Baclini, Miss. Marie Catherine	female	5.00	2	1	2666	19.2583	N
469	470	1	3	Baclini, Miss. Helene Barbara	female	0.75	2	1	2666	19.2583	N
479	480	1	3	Hirvonen, Miss. Hildur E	female	2.00	0	1	3101298	12.2875	N
530	531	1	2	Quick, Miss. Phyllis May	female	2.00	1	1	26360	26.0000	N

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cal
618	619	1	2	Becker, Miss. Marion Louise	female	4.00	2	1	230136	39.0000	
642	643	0	3	Skoog, Miss. Margit Elizabeth	female	2.00	3	2	347088	27.9000	N
644	645	1	3	Baclini, Miss. Eugenie	female	0.75	2	1	2666	19.2583	N
691	692	1	3	Karun, Miss. Manca	female	4.00	0	1	349256	13.4167	N
750	751	1	2	Wells, Miss. Joan	female	4.00	1	1	29103	23.0000	N
755	756	1	2	Hamalainen, Master. Viljo	male	0.67	1	1	250649	14.5000	N
777	778	1	3	Emanuel, Miss. Virginia Ethel	female	5.00	0	0	364516	12.4750	N
788	789	1	3	Dean, Master. Bertram Vere	male	1.00	1	2	C.A. 2315	20.5750	N
803	804	1	3	Thomas, Master. Assad Alexander	male	0.42	0	1	2625	8.5167	N
824	825	0	3	Panula, Master. Urho Abraham	male	2.00	4	1	3101295	39.6875	N
827	828	1	2	Mallet, Master. Andre	male	1.00	0	2	S.C./PARIS 2079	37.0042	N
831	832	1	2	Richards, Master. George Sibley	male	0.83	1	1	29106	18.7500	N
850	851	0	3	Andersson, Master. Sigvard Harald Elias	male	4.00	4	2	347082	31.2750	N
869	870	1	3	Johnson, Master. Harold Theodor	male	4.00	1	1	347742	11.1333	N

UNIQUE VALUE

In [34]:
```python
titanic_df.Sex.unique()
```

Out[34]:
```
array(['male', 'female'], dtype=object)
```

WORKING WITH NA VALUE

In [35]:
```python
titanic_df.Embarked.value_counts(dropna=False)
```

Out[35]:
```
S      644
C      168
Q       77
NaN      2
Name: Embarked, dtype: int64
```

In [36]:
```python
titanic_df.Embarked.value_counts(dropna=True)
```

Out[36]:
```
S    644
C    168
Q     77
Name: Embarked, dtype: int64
```

In [37]:
```python
titanic_df.dropna?
```

In [38]:
```python
titanic_df.Embarked.dropna(inplace=True)
```

In [39]:
```python
titanic_df.Embarked.dropna(inplace=False)
```

Out[39]:
```
0      S
1      C
2      S
3      S
4      S
      ..
886    S
887    S
888    S
889    C
890    Q
Name: Embarked, Length: 889, dtype: object
```

In [40]:
```python
titanic_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          747 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 [41]:
```python
clean_titanic_df=titanic_df.dropna(how='all')
len(clean_titanic_df)
```

Out[41]: 891

In [42]:
```python
clean_titanic_df=titanic_df.dropna()
clean_titanic_df.shape
```

Out[42]: (187, 12)

Apply filter criteria and Select columns based

In [43]:
```python
titanic_df[(titanic_df. Age <=5)&
           (titanic_df.Survived==1)][['Age','Sex','Pclass']].head()
```

Out[43]:

	Age	Sex	Pclass
10	4.00	female	3
43	3.00	female	2
58	5.00	female	2
78	0.83	male	2
172	1.00	female	3

In [44]:
```python
titanic_df[titanic_df.Age.isnull()][['Age','Sex','Pclass']][0:5]
```

Out[44]:

	Age	Sex	Pclass
168	NaN	male	1
176	NaN	male	3
180	NaN	female	3
181	NaN	male	2
185	NaN	male	1

In [45]:
```python
titanic_df[-titanic_df.Age.isnull()][['Age','Sex','Pclass']][0:5]
```

Out[45]:

	Age	Sex	Pclass
0	22.0	male	3
1	38.0	female	1

	Age	Sex	Pclass
2	26.0	female	3
3	35.0	female	1
4	35.0	male	3

Groupby and Aggregate

In [49]:
```python
titanic_df.groupby(['Pclass','Sex'])['Age'].mean().reset_index()
```

Out[49]:

	Pclass	Sex	Age
0	1	female	34.563218
1	1	male	41.081748
2	2	female	28.722973
3	2	male	30.533300
4	3	female	22.259091
5	3	male	26.953919

In [50]:
```python
pclass_gender_age_df = titanic_df.groupby(['Pclass',
                                           'Sex'])['Age'].mean().reset_index()

pclass_gender_age_df
```

Out[50]:

	Pclass	Sex	Age
0	1	female	34.563218
1	1	male	41.081748
2	2	female	28.722973
3	2	male	30.533300
4	3	female	22.259091
5	3	male	26.953919

In [51]:
```python
pclass_gender_sur_df = titanic_df.groupby( ['Pclass',
                                            "Sex"] )["Survived"].sum().reset_index()

pclass_gender_sur_df
```

Out[51]:

	Pclass	Sex	Survived
0	1	female	91
1	1	male	45
2	2	female	70
3	2	male	17
4	3	female	72

	Pclass	Sex	Survived
5	3	male	47

Merge Dataframe

In [52]:
```python
pclass_gender_merge_df = pclass_gender_age_df.merge(
                         pclass_gender_sur_df, on = ['Pclass','Sex'])

pclass_gender_merge_df
```

Out[52]:

	Pclass	Sex	Age	Survived
0	1	female	34.563218	91
1	1	male	41.081748	45
2	2	female	28.722973	70
3	2	male	30.533300	17
4	3	female	22.259091	72
5	3	male	26.953919	47

Sort values

In [53]:
```python
pclass_gender_merge_df.to_json('tojason')
```

In [55]:
```python
pclass_gender_merge_df.sort_values("Survived")
```

Out[55]:

	Pclass	Sex	Age	Survived
3	2	male	30.533300	17
1	1	male	41.081748	45
5	3	male	26.953919	47
2	2	female	28.722973	70
4	3	female	22.259091	72
0	1	female	34.563218	91

In [56]:
```python
pclass_gender_merge_df.sort_values("Survived", ascending=False)
```

Out[56]:

	Pclass	Sex	Age	Survived
0	1	female	34.563218	91
4	3	female	22.259091	72
2	2	female	28.722973	70
5	3	male	26.953919	47
1	1	male	41.081748	45

Pclass	Sex	Age	Survived
3	2	male	30.533300	17

User defined functions

```
In [57]: def find_na(name):
             return titanic_df[name].value_counts(dropna=False)
```

```
In [58]: def find_unique(name):
             return titanic_df[name].unique()
```

```
In [60]: def cross_tab(x,y):
             return pd.crosstab(titanic_df[x],titanic_df[y])
```

```
In [61]: cross_tab('Sex','Pclass')
```

Out[61]:

Pclass	1	2	3
Sex			
female	94	76	144
male	122	108	347

Get the categorical and numeric features

```
In [62]: import numpy as np

         cat_feat_list = [x for x in titanic_df.select_dtypes(include = np.object)]
         cat_feat_list

         num_feat_list = [x for x in titanic_df.select_dtypes(include = np.number)]
         num_feat_list
```

```
C:\Users\artik\AppData\Local\Temp/ipykernel_29064/3860435171.py:3: DeprecationWarnin
g: `np.object` is a deprecated alias for the builtin `object`. To silence this warnin
g, use `object` by itself. Doing this will not modify any behavior and is safe.
Deprecated in NumPy 1.20; for more details and guidance: https://numpy.org/devdocs/re
lease/1.20.0-notes.html#deprecations
  cat_feat_list = [x for x in titanic_df.select_dtypes(include = np.object)]
```

Out[62]: ['PassengerId', 'Survived', 'Pclass', 'Age', 'SibSp', 'Parch', 'Fare']

```
In [64]: cat_feat = ['Sex','Survived']

         for c in cat_feat:
             find_na(c)
             find_unique(c)
```

Lambda function with map

In [65]:
```python
titanic_df['gender'] = titanic_df.Sex.map(lambda x: int(x =='male'))
titanic_df.iloc[0:4,]
```

Out[65]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	

In [ ]: