EDA

IMPORTING LIBRARIES

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import sklearn
import warnings
warnings.filterwarnings('ignore')
```

IMPORTING DATASET

In [2]:
```
df=pd.read_csv("C:/Users/Yug/Downloads/titanic/train.csv")
```

DATA PREPROCESSING AND VISUALIZATION

In [3]:
```
df.head(6)
```

Out[3]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabi
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	Na
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C8
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	Na
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C12
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	Na
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	Na

In [4]:
```
df.shape
```

Out[4]:
```
(891, 12)
```

In [5]:
```
df.size
```

Out[5]:
```
10692
```

In [6]:
```
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 [7]:
```
df.dtypes
```

Out[7]:
```
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 [8]:
```
df.columns
```

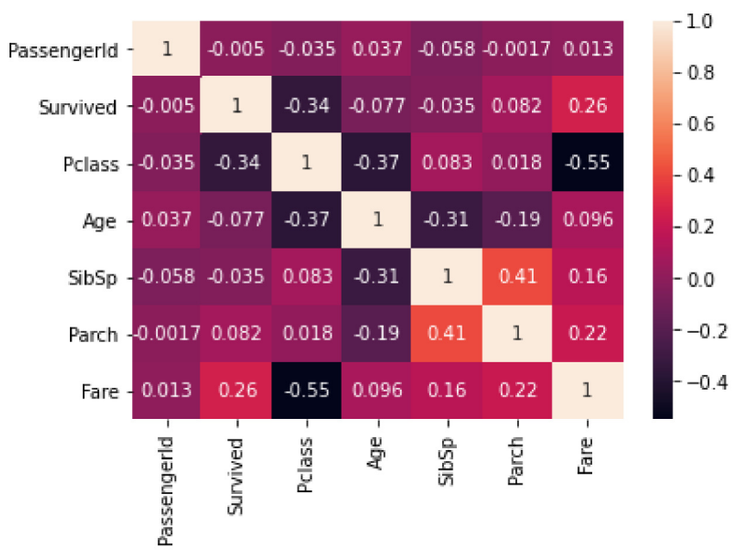
Out[8]:
```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
       'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],
      dtype='object')
```

In [9]:
```
df.select_dtypes(include='object').columns #RETURNS CATEGORICAL COLUMNS
```

Out[9]:
```
Index(['Name', 'Sex', 'Ticket', 'Cabin', 'Embarked'], dtype='object')
```

In [10]:
```
sns.heatmap(df.corr(),annot=True)
```

Out[10]:
```
<AxesSubplot:>
```



In [11]:
```
df.value_counts()                                    #COUNTS DISTINCT VALUES PRESENT IN A PA
```

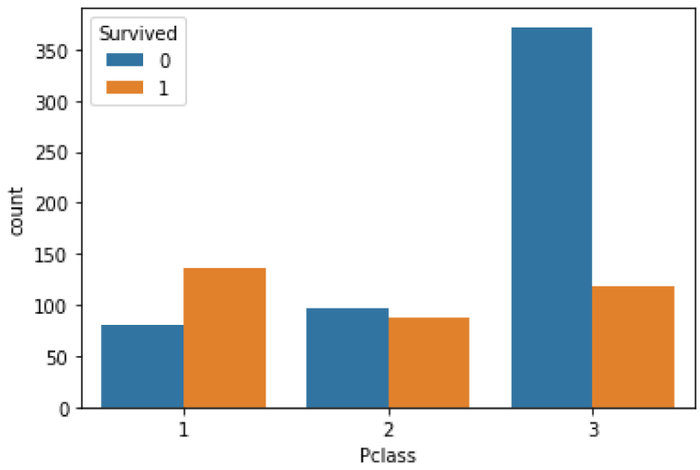
Out[11]:
```
PassengerId  Survived  Pclass  Name
Sex     Age   SibSp  Parch  Ticket    Fare      Cabin  Embarked
890          1         1       Behr, Mr. Karl Howell
male    26.0  0      0      111369    30.0000   C148   C           1
337          0         1       Pears, Mr. Thomas Clinton
male    29.0  1      0      113776    66.6000   C2     S           1
332          0         1       Partner, Mr. Austen
male    45.5  0      0      113043    28.5000   C124   S           1
330          1         1       Hippach, Miss. Jean Gertrude
female  16.0  0      1      111361    57.9792   B18    C           1
328          1         2       Ball, Mrs. (Ada E Hall)
female  36.0  0      0      28551     13.0000   D      S           1

..
584          0         1       Ross, Mr. John Hugo
male    36.0  0      0      13049     40.1250   A10    C           1
582          1         1       Thayer, Mrs. John Borland (Marian Longstreth Mor
ris)  female  39.0  1      1      17421     110.8833  C68    C           1
578          1         1       Silvey, Mrs. William Baird (Alice Munger)
female  39.0  1      0      13507     55.9000   E44    S           1
573          1         1       Flynn, Mr. John Irwin ("Irving")
male    36.0  0      0      PC 17474  26.3875   E25    S           1
2            1         1       Cumings, Mrs. John Bradley (Florence Briggs Thay
er)   female  38.0  1      0      PC 17599  71.2833   C85    C           1
Length: 183, dtype: int64
```

In [12]:
```
df['Pclass'].value_counts(sort=False)
```

Out[12]:
```
1    216
2    184
3    491
Name: Pclass, dtype: int64
```

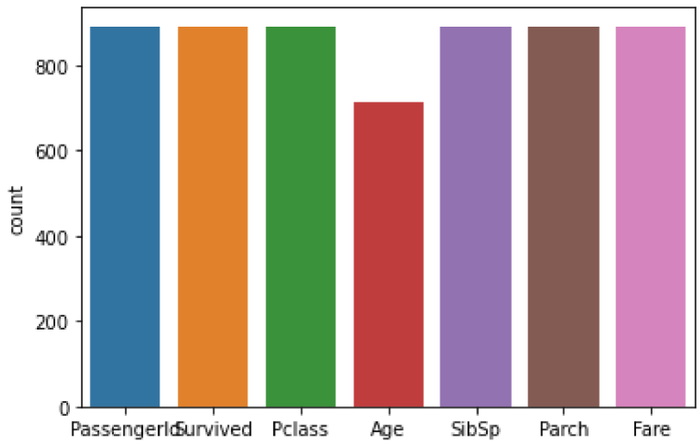
In [13]:
```
sns.countplot(x='Pclass',hue='Survived',data=df)
```

Out[13]:
```
<AxesSubplot:xlabel='Pclass', ylabel='count'>
```



In [14]:
```
sns.countplot(data=df)
```

Out[14]: <AxesSubplot:ylabel='count'>



MISSING VALUES

In [15]:
```
df.isna()                #CHECKING NULL VALUES (FALSE= NOT NULL)
```

Out[15]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Emb
0	False	False	False	False	False	False	False	False	False	False	True	
1	False	False	False	False	False	False	False	False	False	False	False	
2	False	False	False	False	False	False	False	False	False	False	True	
3	False	False	False	False	False	False	False	False	False	False	False	
4	False	False	False	False	False	False	False	False	False	False	True	
...	...	...	...	...	...	...	...	...	...	...	...	
886	False	False	False	False	False	False	False	False	False	False	True	
887	False	False	False	False	False	False	False	False	False	False	False	
888	False	False	False	False	False	True	False	False	False	False	True	
889	False	False	False	False	False	False	False	False	False	False	False	
890	False	False	False	False	False	False	False	False	False	False	True	

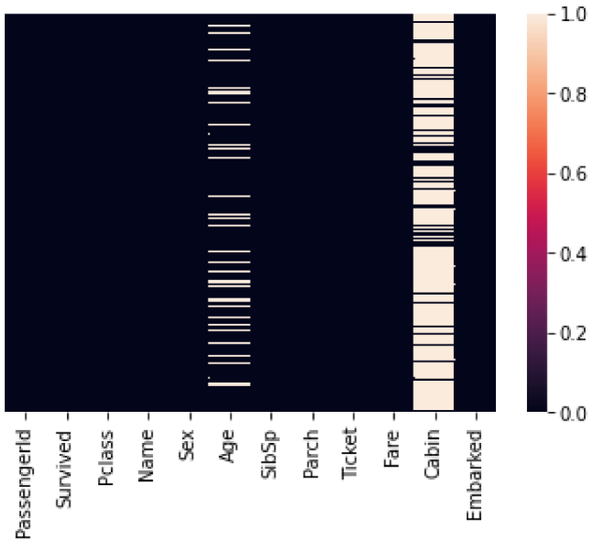
891 rows × 12 columns

In [16]:
```
df.isna().sum()
```

Out[16]:
```
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 [17]:
```
sns.heatmap(df.isnull(),yticklabels=False)    #(single color in a column represer
```

Out[17]: <AxesSubplot:>



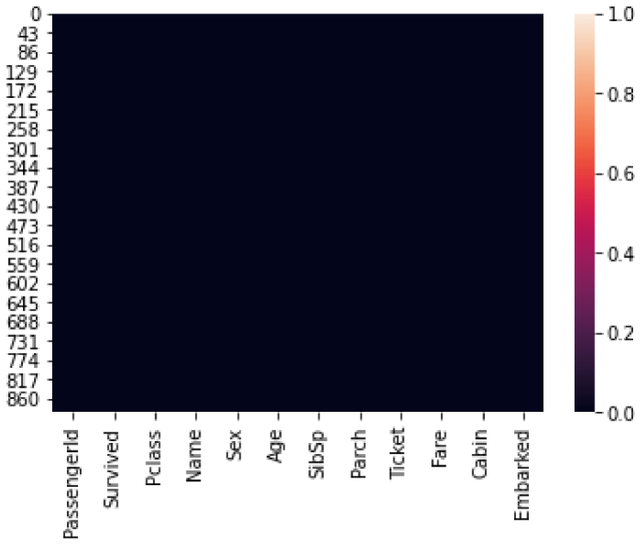
In [18]:
```
df=df.fillna(
{'Age':df['Age'].interpolate(),
 'Cabin':df['Cabin'].fillna(method='bfill'),
 'Embarked':df['Embarked'].replace(np.NaN,'S')

})                                                    #treating the null values
```

In [19]:
```
df.isna().sum()
```

Out[19]:
```
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
Fare           0
Cabin          1
Embarked       0
dtype: int64
```

In [20]:
```
sns.heatmap(df.isna())    #no null values now
```

Out[20]: <AxesSubplot:>



DUPLICATE VALUES

In [21]:
```
df.duplicated()
```

Out[21]:
```
0      False
1      False
2      False
3      False
4      False
       ...
886    False
887    False
888    False
889    False
890    False
Length: 891, dtype: bool
```

In [22]:
```
df.duplicated().sum()     #no duplicate value present
```

Out[22]: 0

In [23]:
```
df.describe(include='object')     #basic statistics
```

Out[23]:

	Name	Sex	Ticket	Cabin	Embarked
count	891	891	891	890	891
unique	891	2	681	147	3
top	Murdlin, Mr. Joseph	male	347082	C78	S
freq	1	577	7	33	646

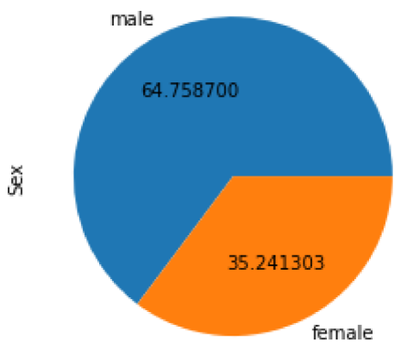
PLOTTING

In [24]:
```
def piechart(x):
    return x.value_counts().plot(kind='pie',autopct='%2f')

def countplot(x1):
    return sns.countplot(x=x1,data=df)

def histogram(x):
    return plt.hist(x,bins=10)

def distplot(x):
    return sns.distplot(x)

def boxplot(x):
    return sns.boxplot(x)
```

1. SEX

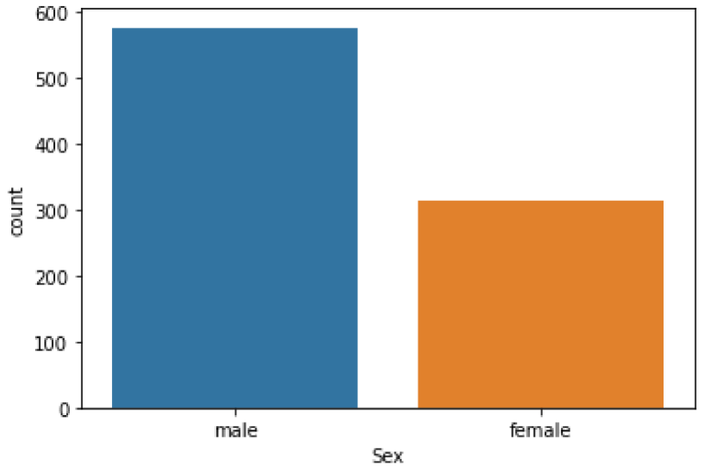
In [25]:
```
piechart(df['Sex'])
```

Out[25]: <AxesSubplot:ylabel='Sex'>



In [26]:
```
countplot(df['Sex'])
```

Out[26]: <AxesSubplot:xlabel='Sex', ylabel='count'>



In [27]:
```
histogram(df['Sex'])
```

Out[27]:
```
(array([577.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   0., 314.]),
 array([0. , 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1. ]),
 <BarContainer object of 10 artists>)
```

OBSERVATION : MALES > FEMALES (ALMOST DOUBLE)

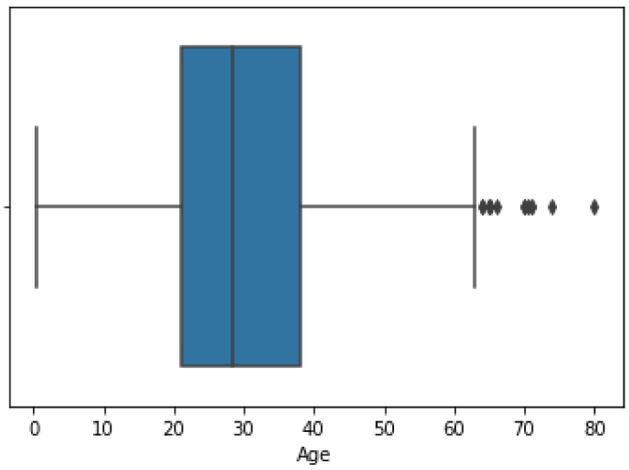
In [28]:
```
df.columns
```

Out[28]:
```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
       'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],
      dtype='object')
```

2. AGE

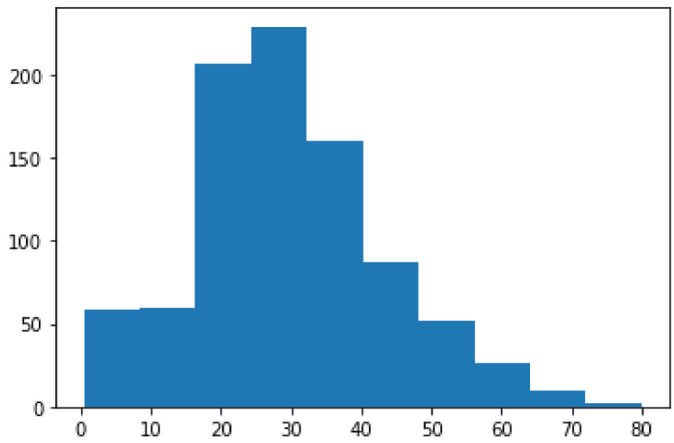
In [29]:
```
boxplot(df['Age'])
```

Out[29]:
```
<AxesSubplot:xlabel='Age'>
```



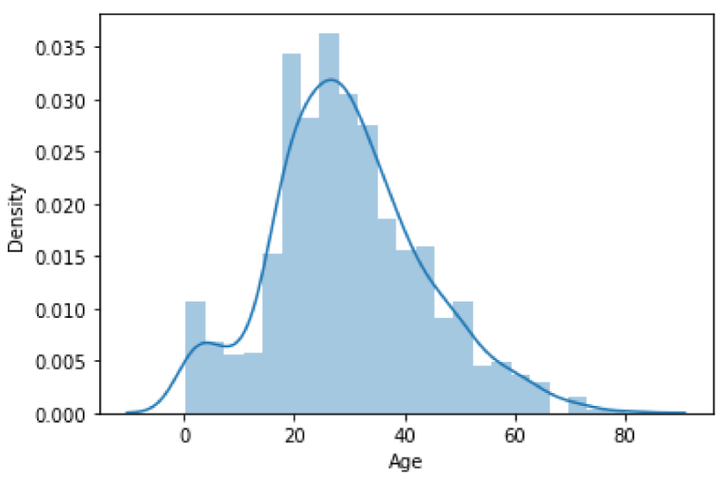
In [30]:
```
histogram(df['Age'])
```

Out[30]:
```
(array([ 58.,  60., 207., 229., 160.,  87.,  52.,  26.,  10.,   2.]),
 array([ 0.42 ,  8.378, 16.336, 24.294, 32.252, 40.21 , 48.168, 56.126,
        64.084, 72.042, 80.   ]),
 <BarContainer object of 10 artists>)
```



In [31]:
```
distplot(df['Age'])
```

Out[31]:
```
<AxesSubplot:xlabel='Age', ylabel='Density'>
```

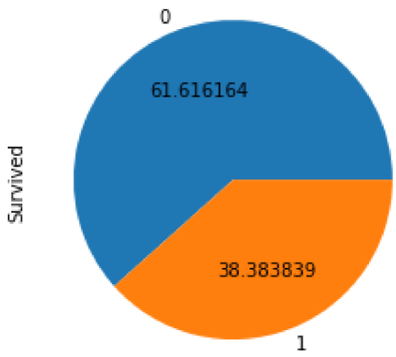


OBSERVATION : In titanic most of the people were in the range 20-40 and very few people were above 65. so age>65 is considered as outliers

3. Survived

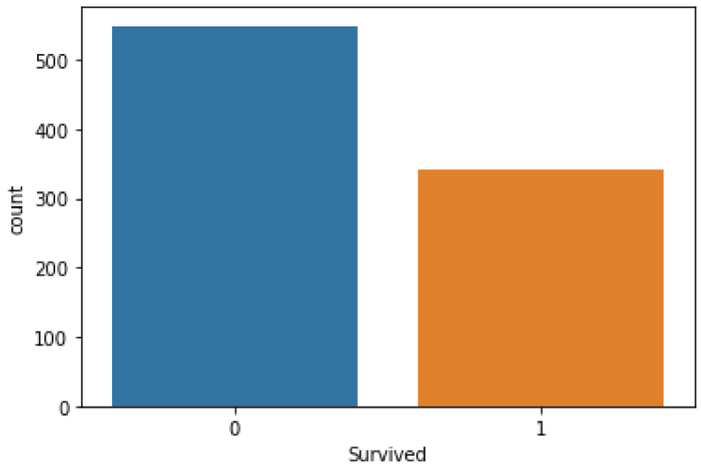
In [32]:
```
piechart(df['Survived'])
```

Out[32]: <AxesSubplot:ylabel='Survived'>



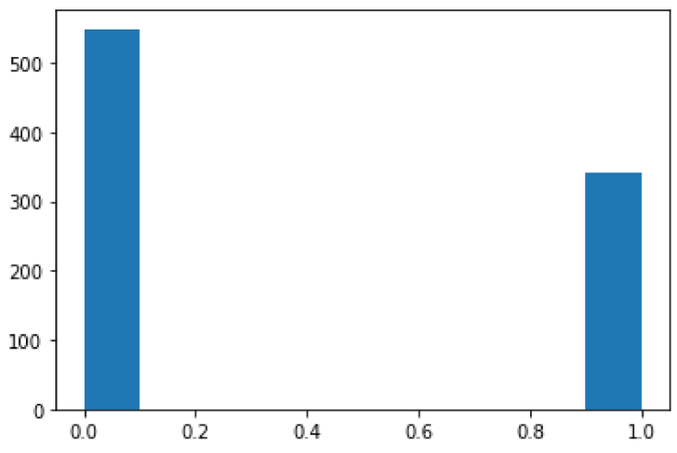
In [33]:
```
countplot(df['Survived'])
```

Out[33]: <AxesSubplot:xlabel='Survived', ylabel='count'>



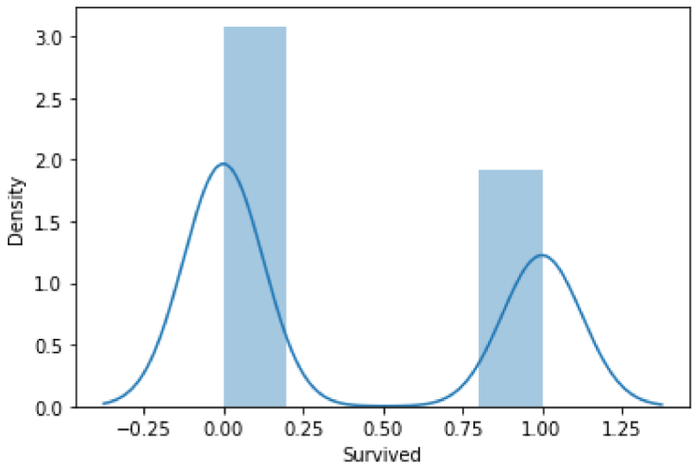
In [34]:
```
histogram(df['Survived'])
```

Out[34]:
```
(array([549.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   0., 342.]),
 array([0. , 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1. ]),
 <BarContainer object of 10 artists>)
```



In [35]:
```
distplot(df['Survived'])
```

Out[35]:
```
<AxesSubplot:xlabel='Survived', ylabel='Density'>
```

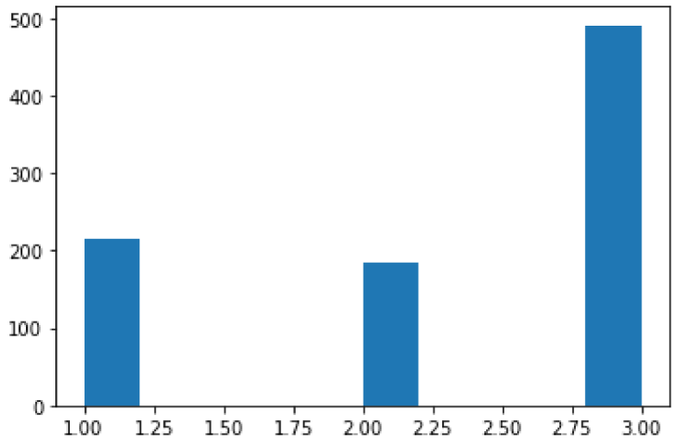


OBSERVATION : More people died and very few ratio of people were able to save themselves.

4. Pclass

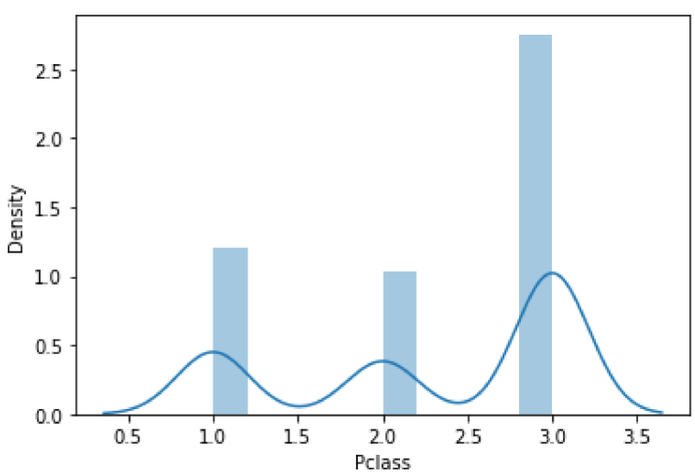
In [36]:
```
histogram(df['Pclass'])
```

Out[36]:
```
(array([216.,   0.,   0.,   0.,   0., 184.,   0.,   0.,   0., 491.]),
 array([1. , 1.2, 1.4, 1.6, 1.8, 2. , 2.2, 2.4, 2.6, 2.8, 3. ]),
 <BarContainer object of 10 artists>)
```



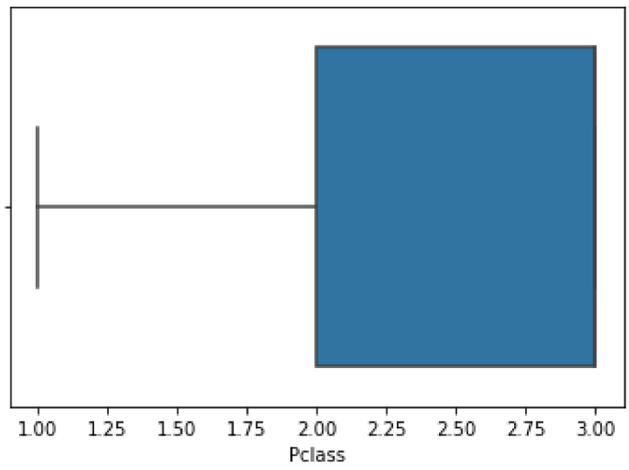
In [37]:
```
distplot(df['Pclass'])
```

Out[37]:
```
<AxesSubplot:xlabel='Pclass', ylabel='Density'>
```



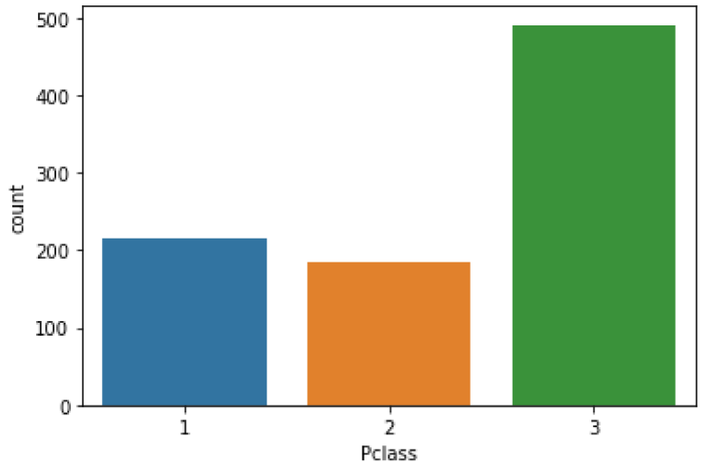
In [38]:
```
boxplot(df['Pclass'])
```

Out[38]: <AxesSubplot:xlabel='Pclass'>



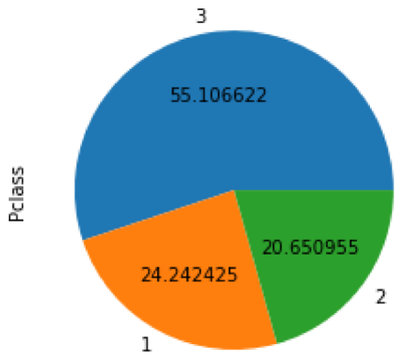
In [39]:
```
countplot(df['Pclass'])
```

Out[39]: <AxesSubplot:xlabel='Pclass', ylabel='count'>



In [40]:
```
piechart(df['Pclass'])
```

Out[40]: <AxesSubplot:ylabel='Pclass'>

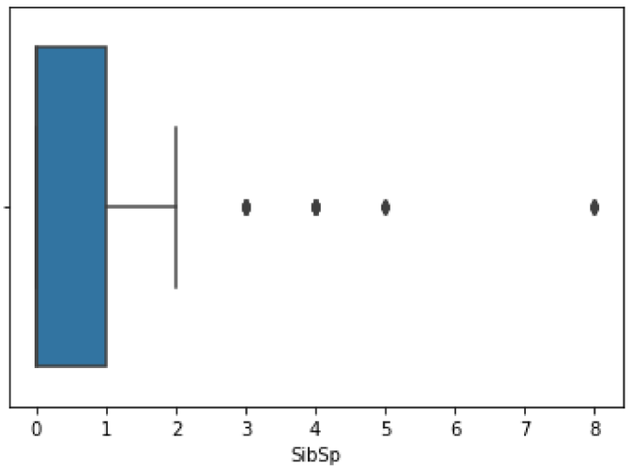


OBSERVATION : MOST OF THE PEOPLE BELONGS TO 3RD CLASS IN TITANIC.

4. SibSp

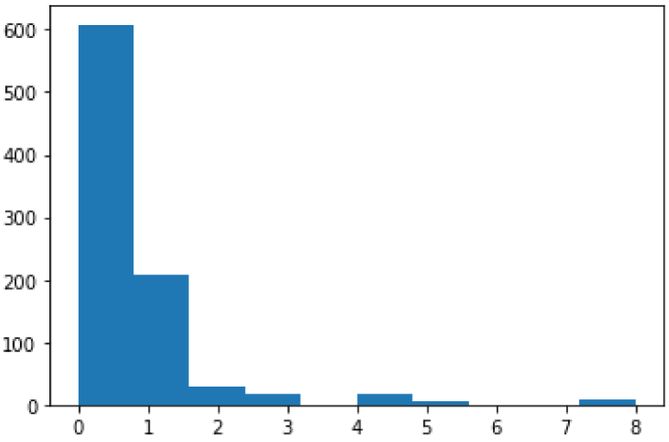
In [41]:
```
boxplot(df['SibSp'])
```

Out[41]:
```
<AxesSubplot:xlabel='SibSp'>
```



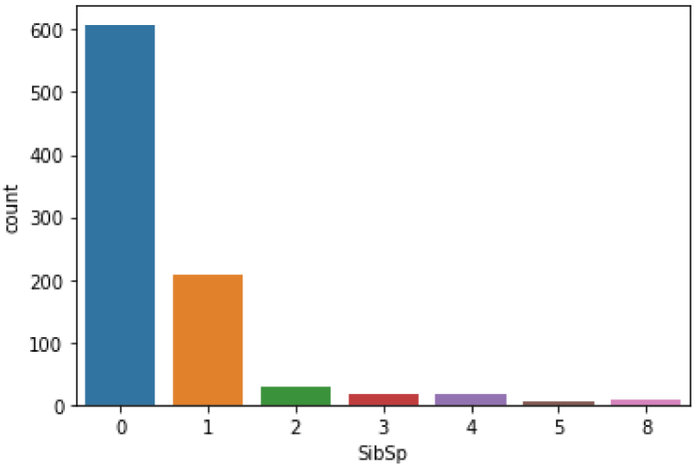
In [42]:
```
histogram(df['SibSp'])
```

Out[42]:
```
(array([608., 209.,  28.,  16.,   0.,  18.,   5.,   0.,   0.,   7.]),
 array([0. , 0.8, 1.6, 2.4, 3.2, 4. , 4.8, 5.6, 6.4, 7.2, 8. ]),
 <BarContainer object of 10 artists>)
```



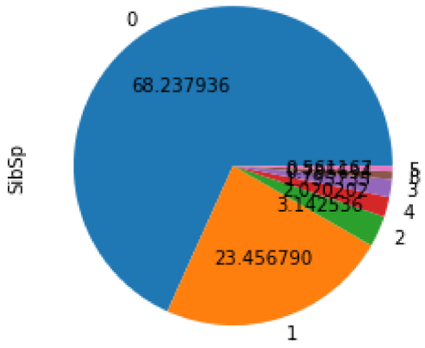
In [43]:
```
countplot(df['SibSp'])
```

Out[43]:
```
<AxesSubplot:xlabel='SibSp', ylabel='count'>
```



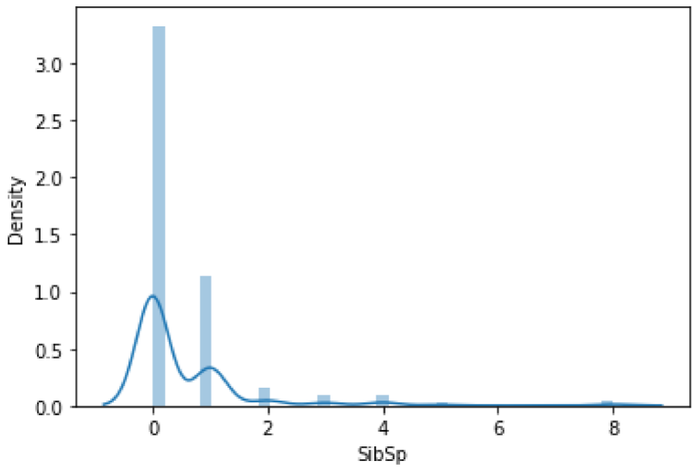
In [44]:
```
piechart(df['SibSp'])
```

Out[44]:
```
<AxesSubplot:ylabel='SibSp'>
```



In [45]:
```
distplot(df['SibSp'])
```

Out[45]:
```
<AxesSubplot:xlabel='SibSp', ylabel='Density'>
```

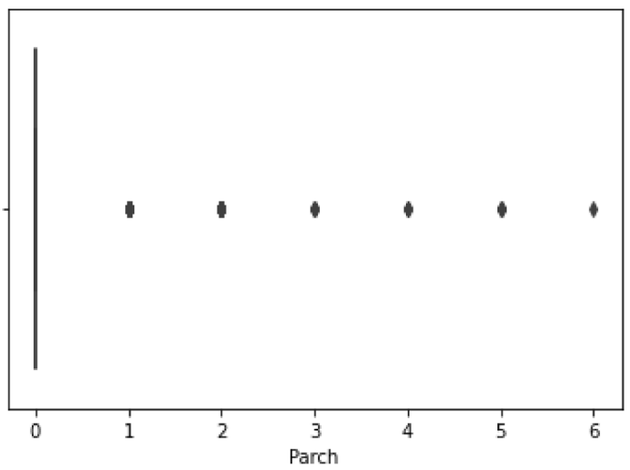


OBSERVATION : Most of the people who were in titanic were either alone or have 1 sibling along with them.

4. Parch

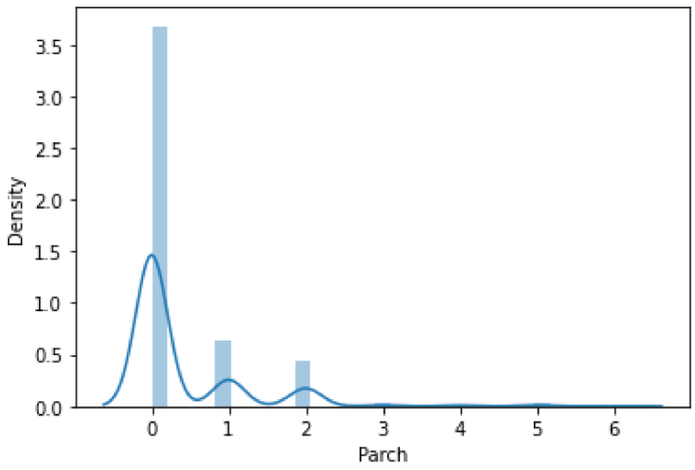
In [46]:
```
boxplot(df['Parch'])
```

Out[46]:
```
<AxesSubplot:xlabel='Parch'>
```



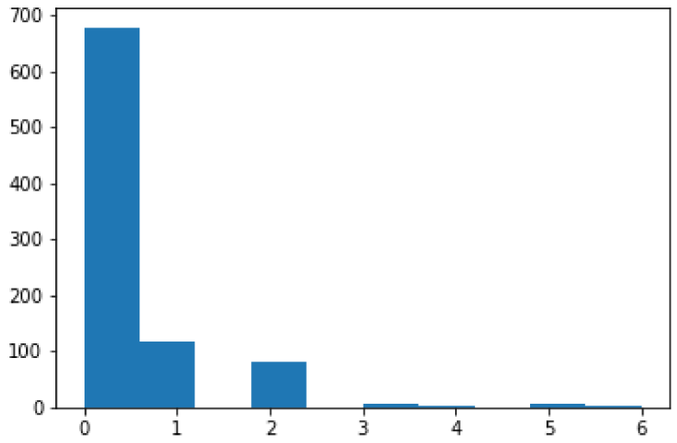
In [47]:
```
distplot(df['Parch'])
```

Out[47]:
```
<AxesSubplot:xlabel='Parch', ylabel='Density'>
```



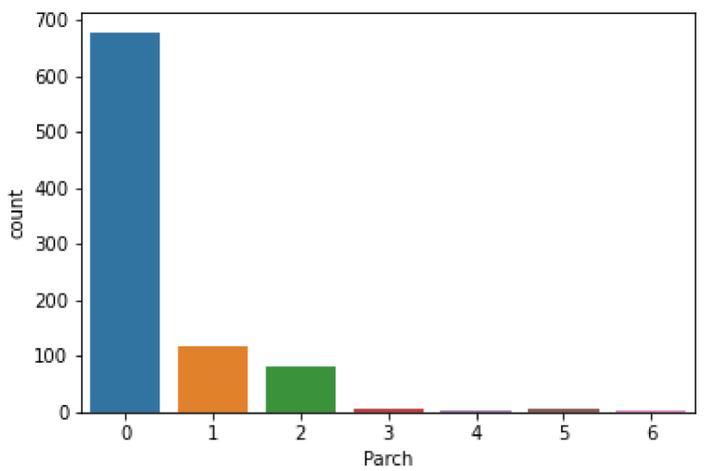
In [48]:
```
histogram(df['Parch'])
```

Out[48]:
```
(array([678., 118.,   0.,  80.,   0.,   5.,   4.,   0.,   5.,   1.]),
 array([0. , 0.6, 1.2, 1.8, 2.4, 3. , 3.6, 4.2, 4.8, 5.4, 6. ]),
 <BarContainer object of 10 artists>)
```



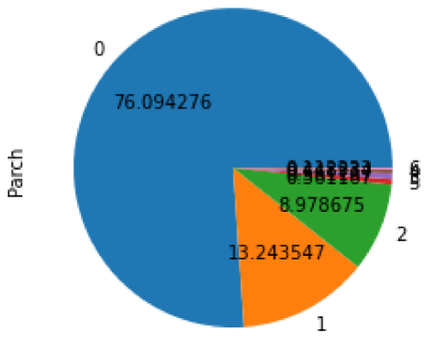
In [49]:
```
countplot(df['Parch'])
```

Out[49]:
```
<AxesSubplot:xlabel='Parch', ylabel='count'>
```



In [50]:
```
piechart(df['Parch'])
```

Out[50]: <AxesSubplot:ylabel='Parch'>

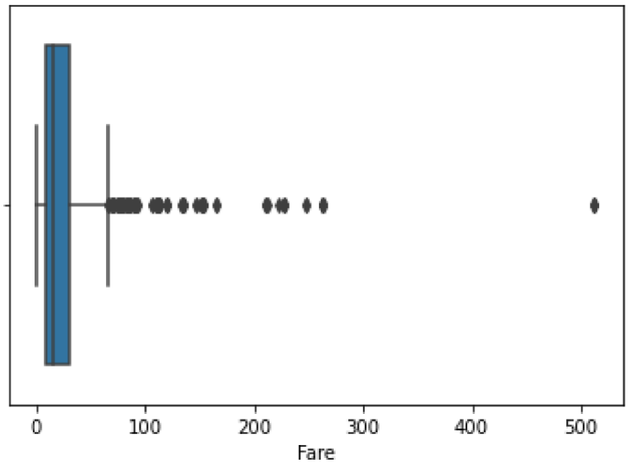


OBSERVATION: Only 24% of the people were there with their children in titanic .

5. 'Fare'

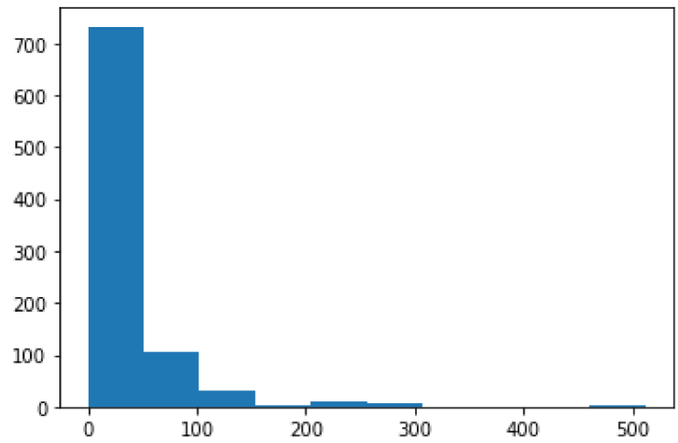
In [51]:
```
boxplot(df['Fare'])
```

Out[51]: <AxesSubplot:xlabel='Fare'>



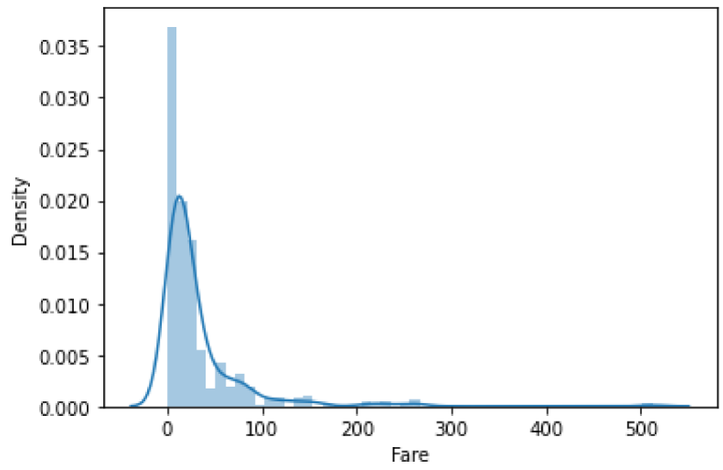
In [52]:
```
histogram(df['Fare'])
```

Out[52]:
```
(array([732., 106.,  31.,   2.,  11.,   6.,   0.,   0.,   0.,   3.]),
 array([  0.     ,  51.23292, 102.46584, 153.69876, 204.93168, 256.1646 ,
        307.39752, 358.63044, 409.86336, 461.09628, 512.3292 ]),
 <BarContainer object of 10 artists>)
```



In [53]:
```
distplot(df['Fare'])
```

Out[53]:
```
<AxesSubplot:xlabel='Fare', ylabel='Density'>
```



In [54]:
```
df['Fare'].max()
```

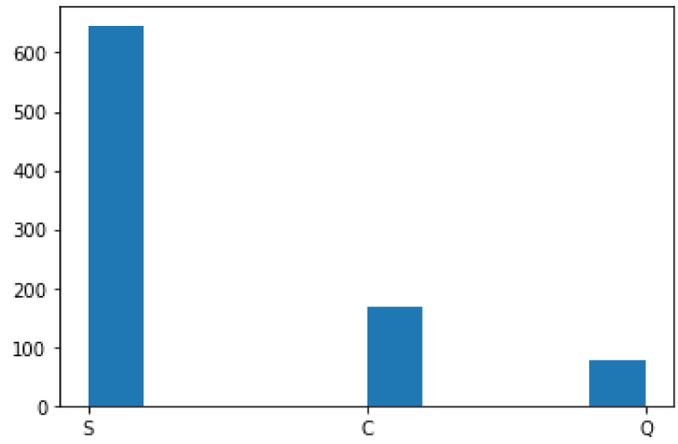
Out[54]:
```
512.3292
```

OBSERVATION : FARE OF TITANIC MAINLY LIES IN THE RANGE 0-100 .FARE ABOVE 100 WAS CONSIDERED AS OUTLIER (MAX FARE- 512.3)

6. 'Embarked'

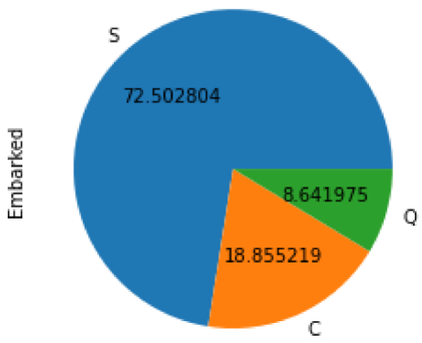
In [55]:
```
histogram(df['Embarked'])
```

Out[55]:
```
(array([646.,   0.,   0.,   0.,   0., 168.,   0.,   0.,   0.,  77.]),
 array([0. , 0.2, 0.4, 0.6, 0.8, 1. , 1.2, 1.4, 1.6, 1.8, 2. ]),
 <BarContainer object of 10 artists>)
```



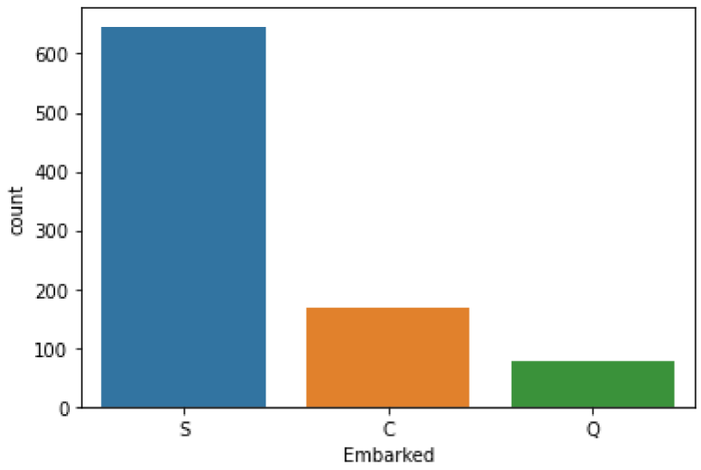
In [56]:
```
piechart(df['Embarked'])
```

Out[56]:
```
<AxesSubplot:ylabel='Embarked'>
```



In [57]:
```
countplot(df['Embarked'])
```

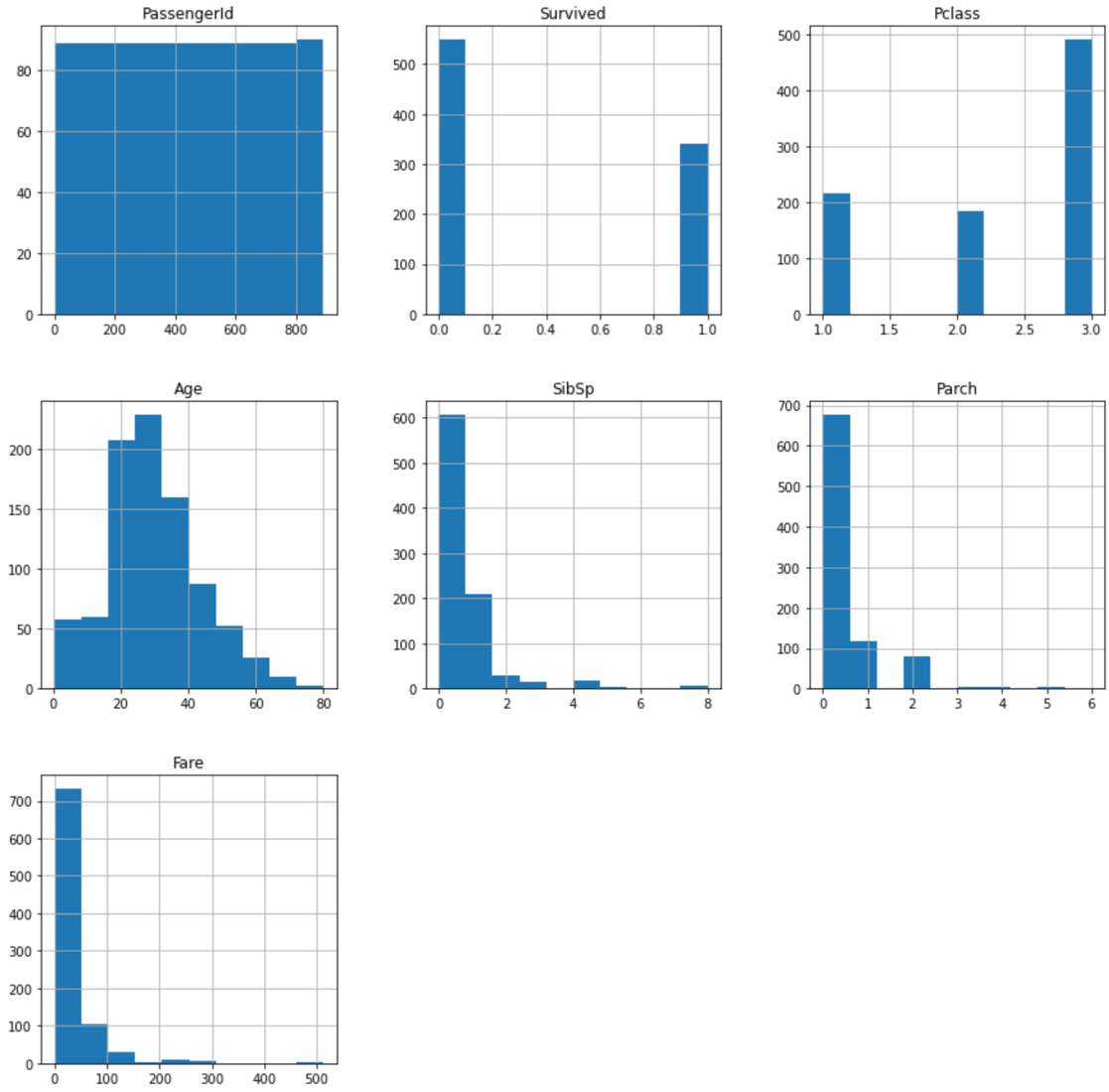
Out[57]:
```
<AxesSubplot:xlabel='Embarked', ylabel='count'>
```



OBSERVATION : S>C (ALMOST 4 TIMES) AND S>Q (ALMOST 9 TIMES)

In [58]:
```
df.hist(figsize=(15,15))
```

Out[58]:
```
array([[<AxesSubplot:title={'center':'PassengerId'}>,
        <AxesSubplot:title={'center':'Survived'}>,
        <AxesSubplot:title={'center':'Pclass'}>],
       [<AxesSubplot:title={'center':'Age'}>,
        <AxesSubplot:title={'center':'SibSp'}>,
        <AxesSubplot:title={'center':'Parch'}>],
       [<AxesSubplot:title={'center':'Fare'}>, <AxesSubplot:>,
        <AxesSubplot:>]], dtype=object)
```

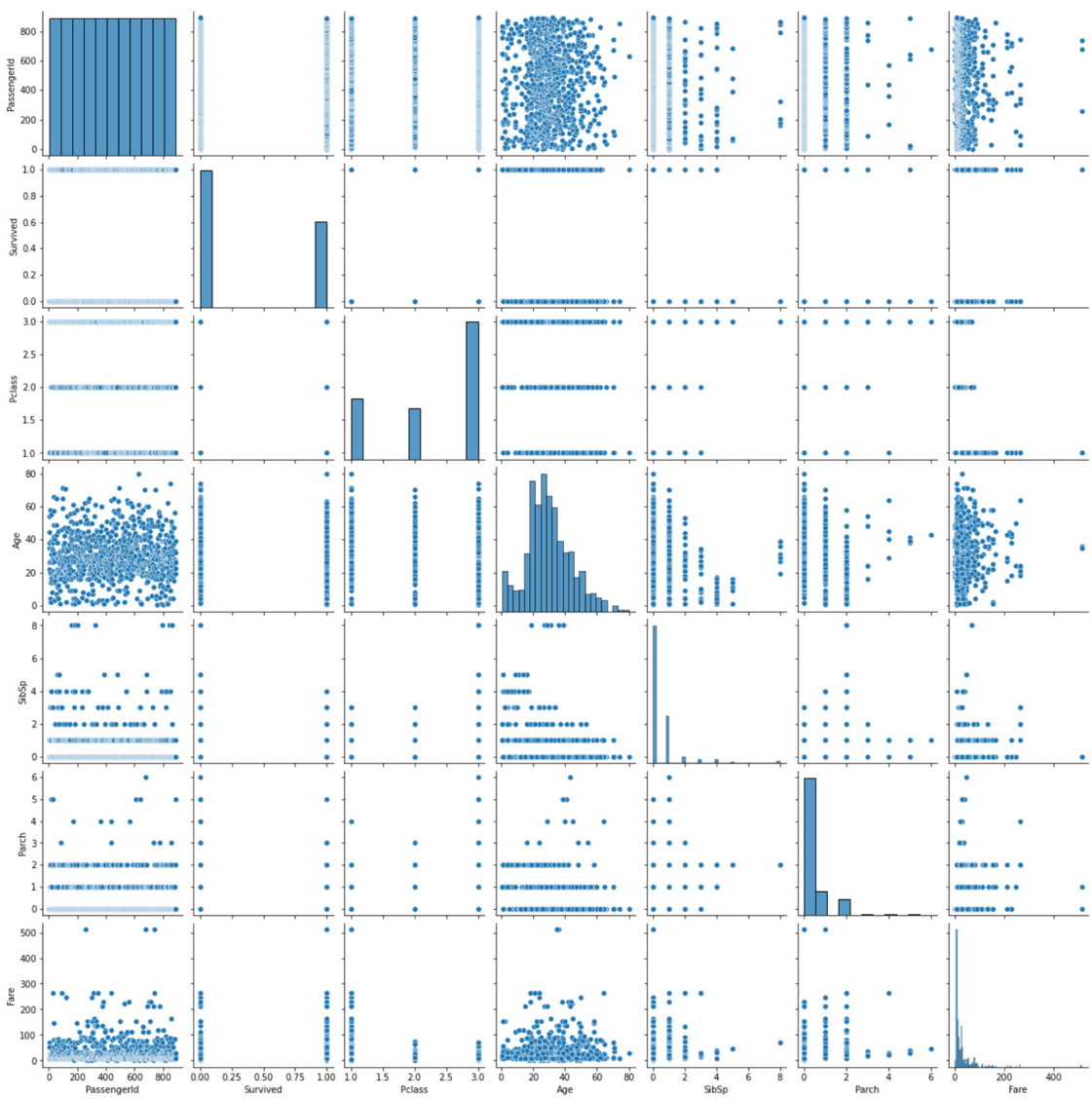


BIVARIATE / MULTIVARIATE ANALYSIS

In [59]:
```
def scatterplot(x,y):
    return sns.scatterplot(x,y)   #num-num

def barplot(x,y):
    return sns.barplot(x,y)       #num-cat

def distplot(x,y):
    return sns.distplot(x,y)     #num-cat

def heatmap(x,y):
    return sns.heatmap(x,y,annot=True)     #cat-cat

def lineplot(x,y):
    return sns.lineplot(x,y)   #num-num

def clustermap(x,y):
    return sns.clustermap(pd.crosstab(x,y),annot=True) #cat-cat

def boxplot(x,y):
    return sns.boxplot(x,y)
```

In [60]:
```
sns.pairplot(df)
```

Out[60]:
```
<seaborn.axisgrid.PairGrid at 0x1b4e818c880>
```



In [61]:
```
df.select_dtypes(include='object').columns
```

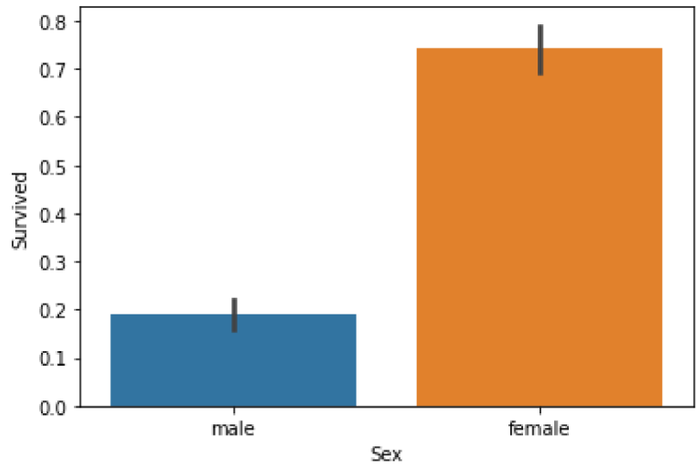
Out[61]:
```
Index(['Name', 'Sex', 'Ticket', 'Cabin', 'Embarked'], dtype='object')
```

In [62]:
```
df.select_dtypes(include=['int64','float64']).columns
```

Out[62]:
```
Index(['PassengerId', 'Survived', 'Pclass', 'Age', 'SibSp', 'Parch', 'Fare'], d
type='object')
```

In [63]:
```
barplot(df['Sex'],df['Survived'])
```

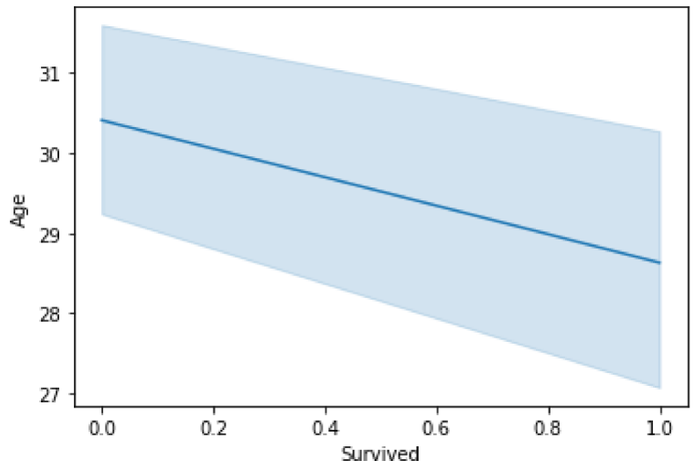
Out[63]: <AxesSubplot:xlabel='Sex', ylabel='Survived'>



OBSERVATION : SURVIVAL RATE OF FEMALE IS MORE THAN MALE.

In [64]:
```
lineplot(df['Survived'],df['Age'])
```

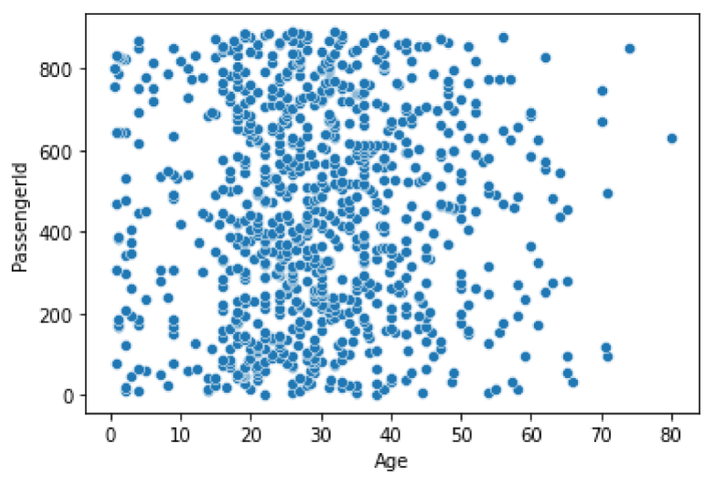
Out[64]: <AxesSubplot:xlabel='Survived', ylabel='Age'>



OBSERVATION : SURVIVAL RATE OF YOUNGER PEOPLE IS MORE THAN OLDER CROWD.

In [65]:
```
scatterplot(df['Age'],df['PassengerId'])
```

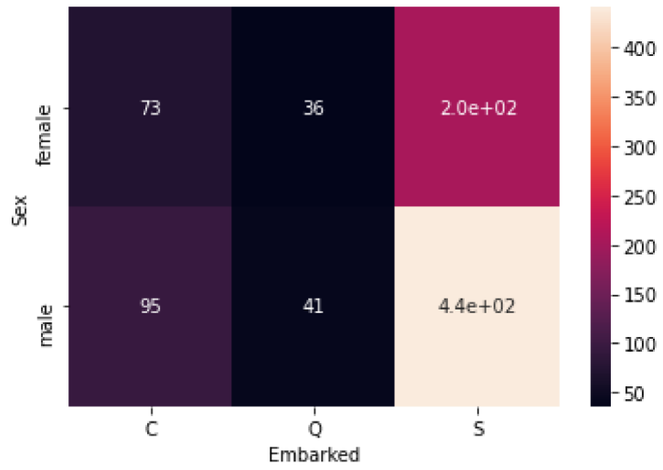
Out[65]: <AxesSubplot:xlabel='Age', ylabel='PassengerId'>



OBSERVATION : people of age 20-50 were travelling more.

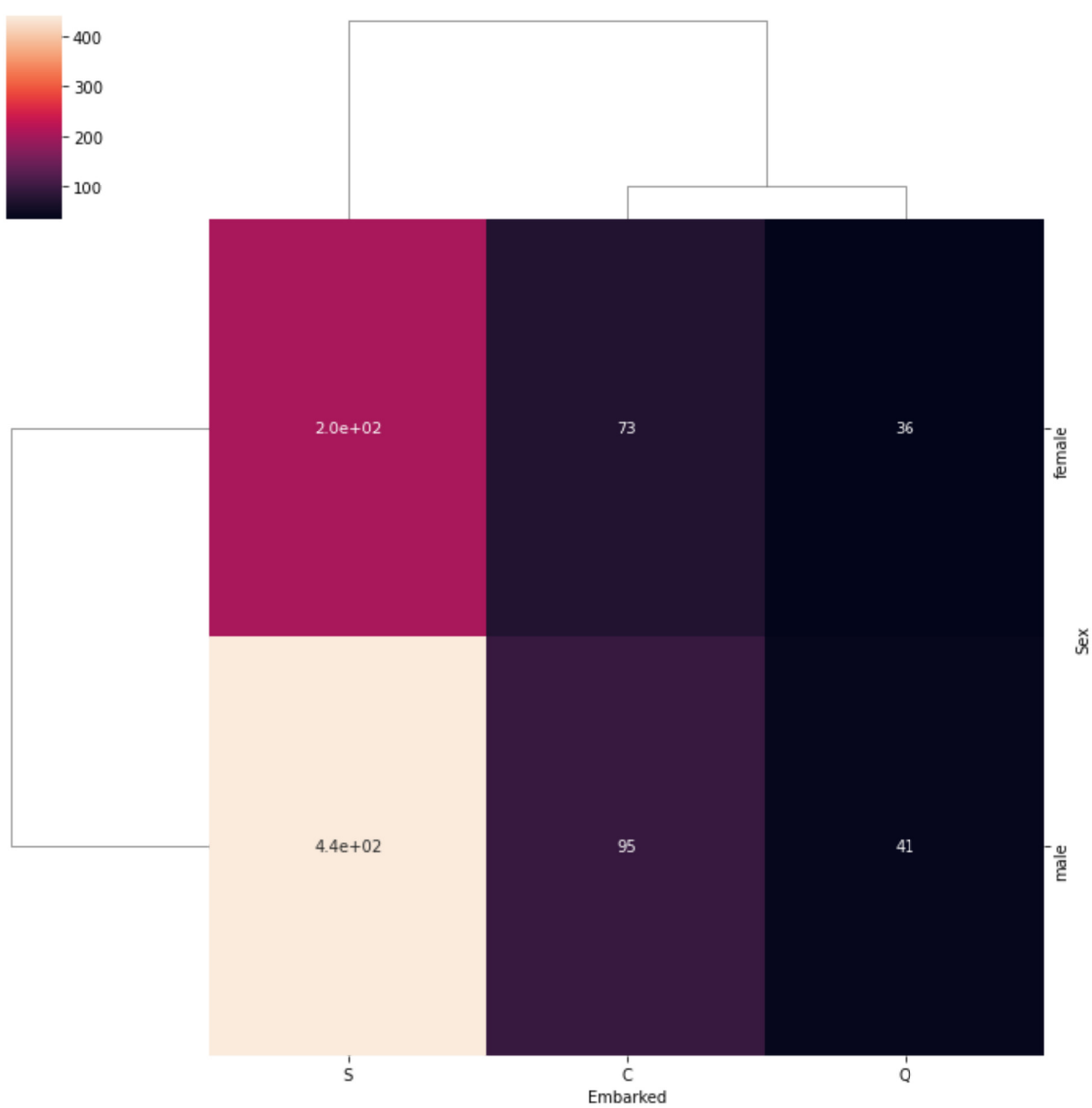
In [66]:
```
sns.heatmap(pd.crosstab(df['Sex'],df['Embarked']),annot=True)
```

Out[66]: <AxesSubplot:xlabel='Embarked', ylabel='Sex'>



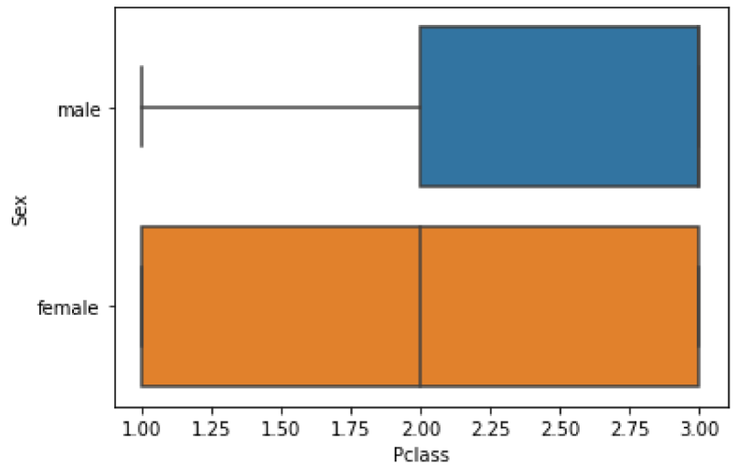
In [67]:
```
clustermap(df['Sex'],df['Embarked'])
```

Out[67]: <seaborn.matrix.ClusterGrid at 0x1b4eb3c8fa0>



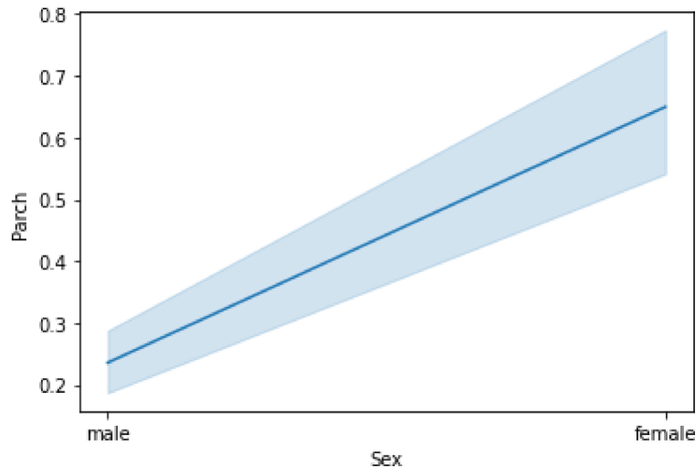
In [68]:
```
boxplot(df['Pclass'],df['Sex'])
```

Out[68]: <AxesSubplot:xlabel='Pclass', ylabel='Sex'>



In [69]:
```
lineplot(df['Sex'],df['Parch'])
```

Out[69]: <AxesSubplot:xlabel='Sex', ylabel='Parch'>



Observation: on an average, there are more females than males along with their child .