

TITANIC DATASET

```
## import the data
df= pd.read_csv("train.csv")
df.head()
```

	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

Data contains 12 columns

df.columns


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

[] df.dtypes

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


Column details and datatype of each column

 df.describe()



	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

It shows that some column has missing values we need to deal with this

 df.describe()



	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

It shows that some column has missing values we need to deal with this

```
df.isnull().sum()
```

	0
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

Count of missing values. So column Age and Cabin and embarked has missing value



```
## drop irrelevant column such as name , ticket and the column Cabin contain >75% missing value so drop that column
```

```
drop_cols = ['Name', 'Ticket', 'Cabin']  
df.drop(drop_cols, axis=1, inplace=True)  
|
```

drop irrelevant column such as
name , ticket and the column
Cabin contain >75% missing
value so drop that column



```
## drop irrelevant column such as name , ticket and the column Cabin contain >75% missing value so drop that column
```

```
drop_cols = ['Name', 'Ticket', 'Cabin']  
df.drop(drop_cols, axis=1, inplace=True)  
|
```

drop irrelevant column such as name , ticket and the column Cabin contain >75% missing value so drop that column and filling age missing value with median and embarked with mode



```
df['Age']=df['Age'].fillna(df['Age'].median())
```

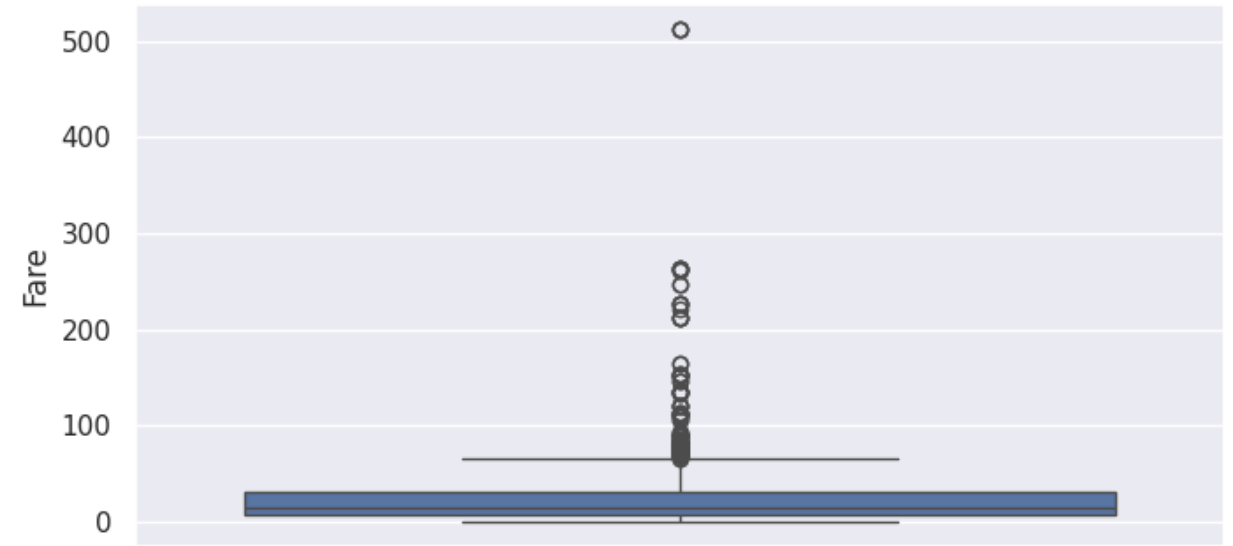
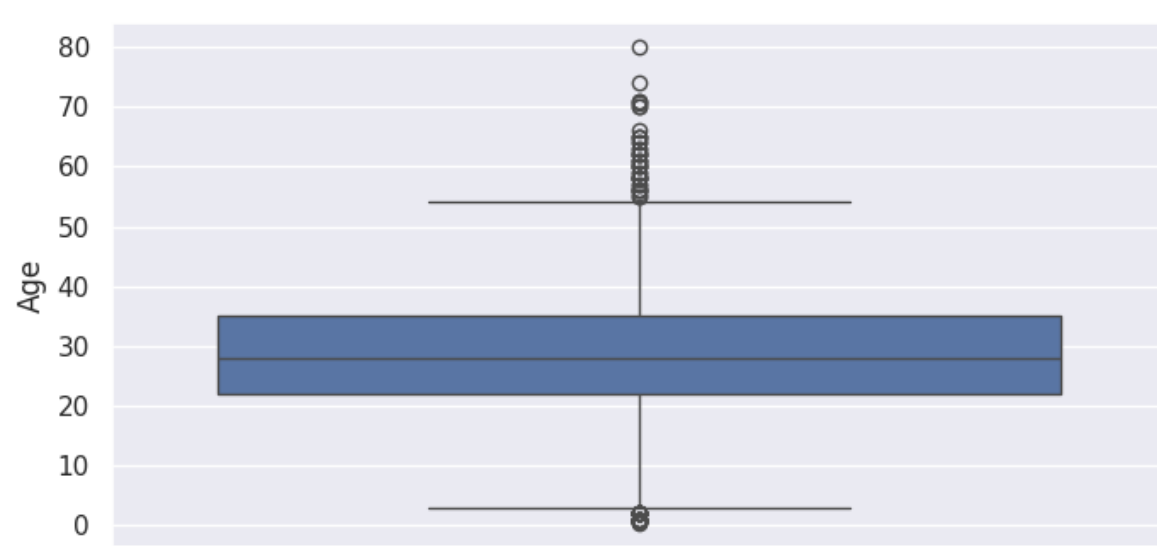
```
[ ] df['Embarked']=df['Embarked'].fillna(df['Embarked'].mode()[0])
```

```
[ ] df.shape
```



```
(891, 9)
```

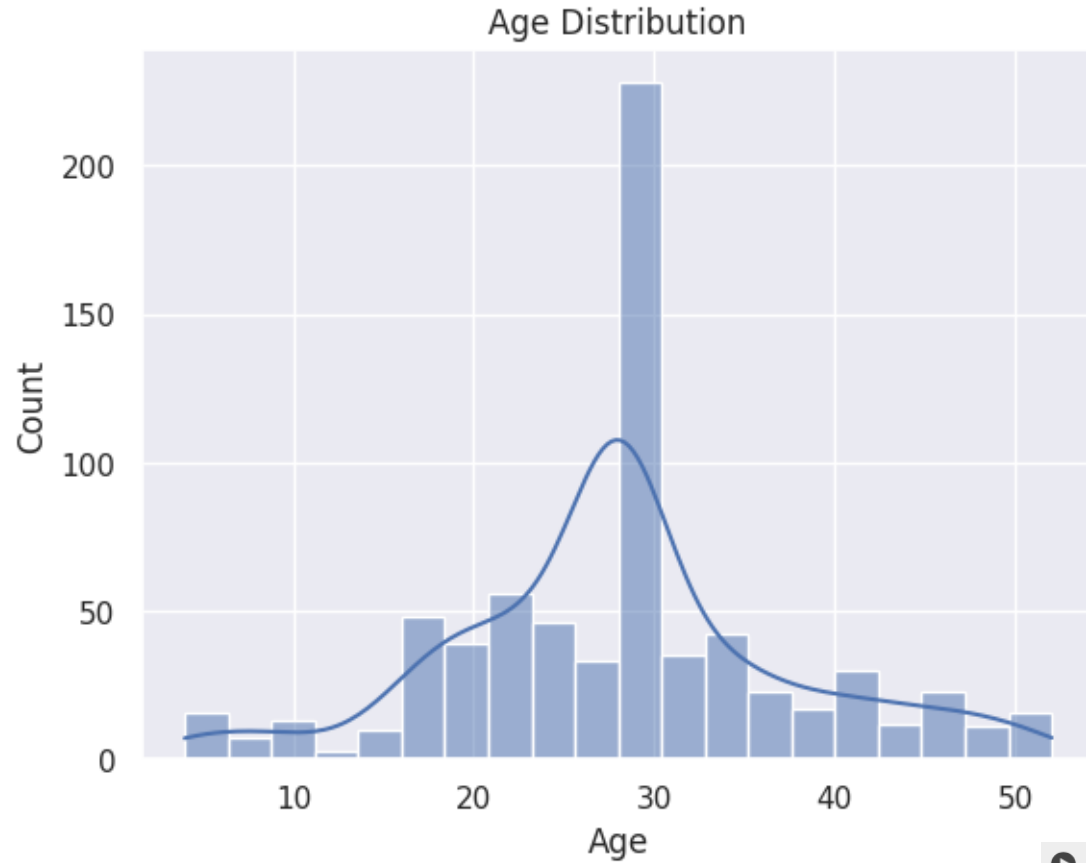
Box Plot: Find the outliers in Age and Fare column needs to remove it



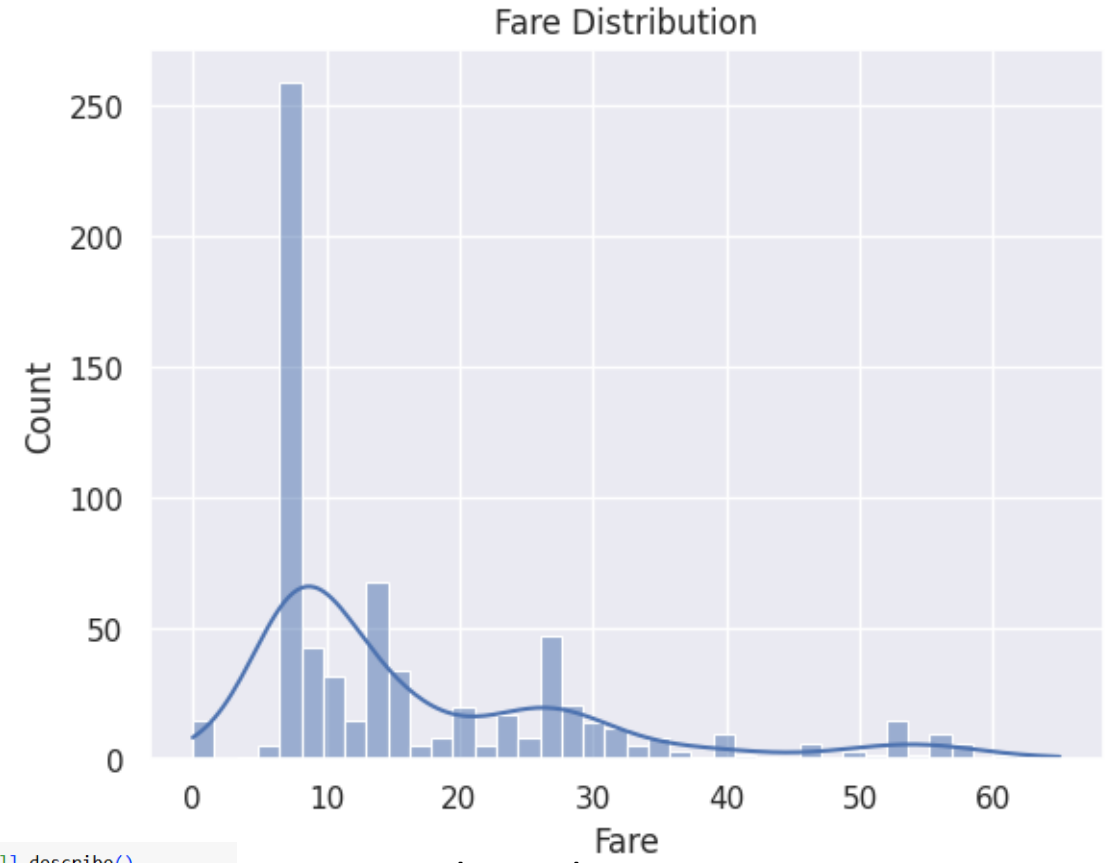
After the outliers removal in Age and Fare column needs to remove it



Histogram Plot for Age and Column



It is symmetric



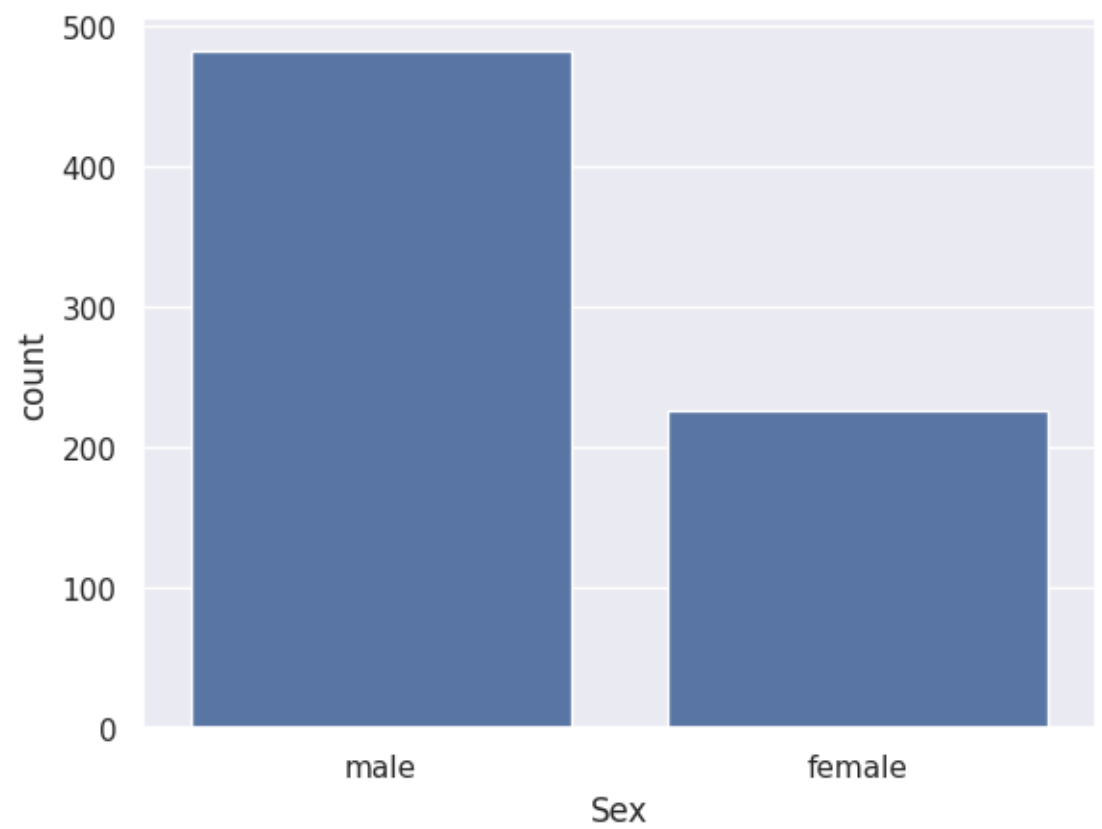
It is skewed

```
df[['Age', 'Fare']].describe()
```

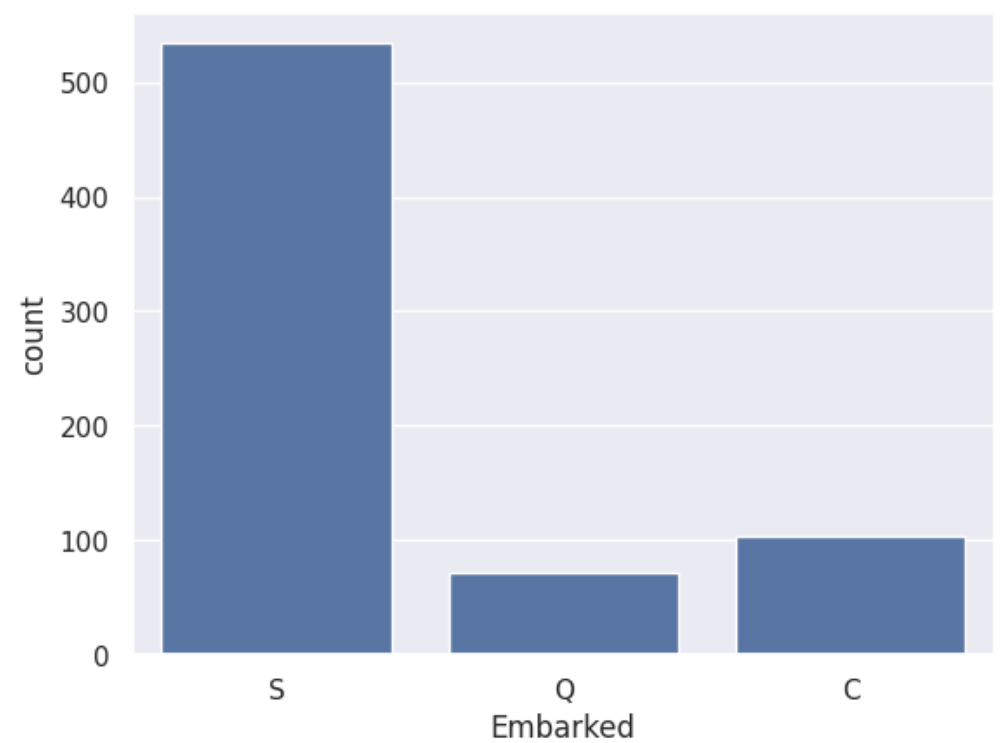


	Age	Fare
count	708.00000	708.000000
mean	28.05226	17.135092
std	9.50620	13.414455
min	4.00000	0.000000
25%	22.00000	7.879200
50%	28.00000	11.241700
75%	32.00000	25.496900
max	52.00000	65.000000

Count Plot for Passengers

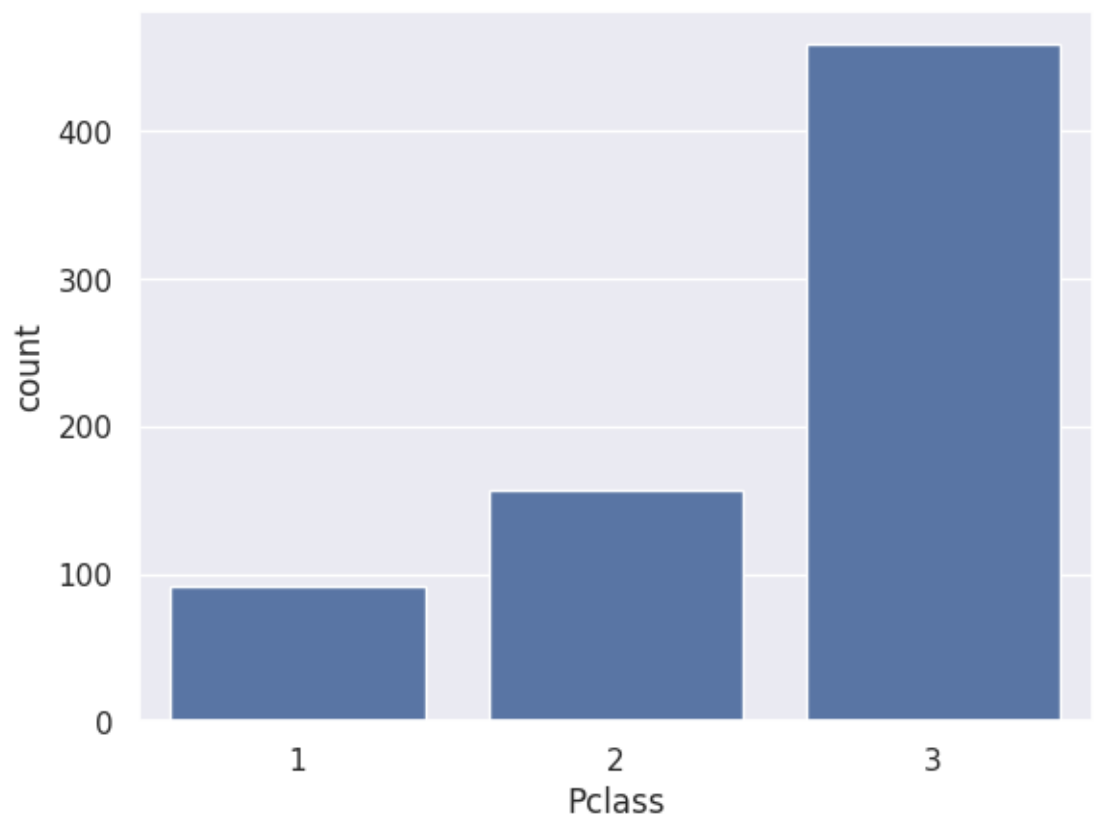


Number of male passenger is more

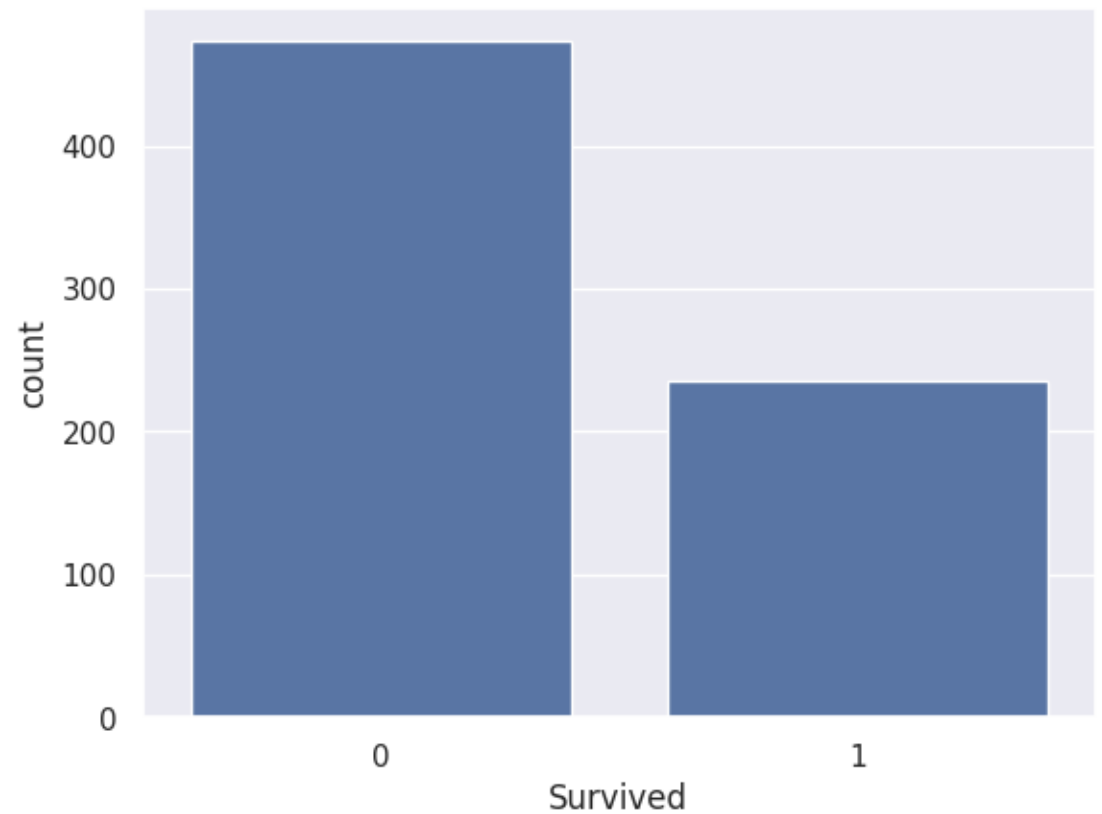


Majority Passenger embarked from S

Count Plot for Passengers

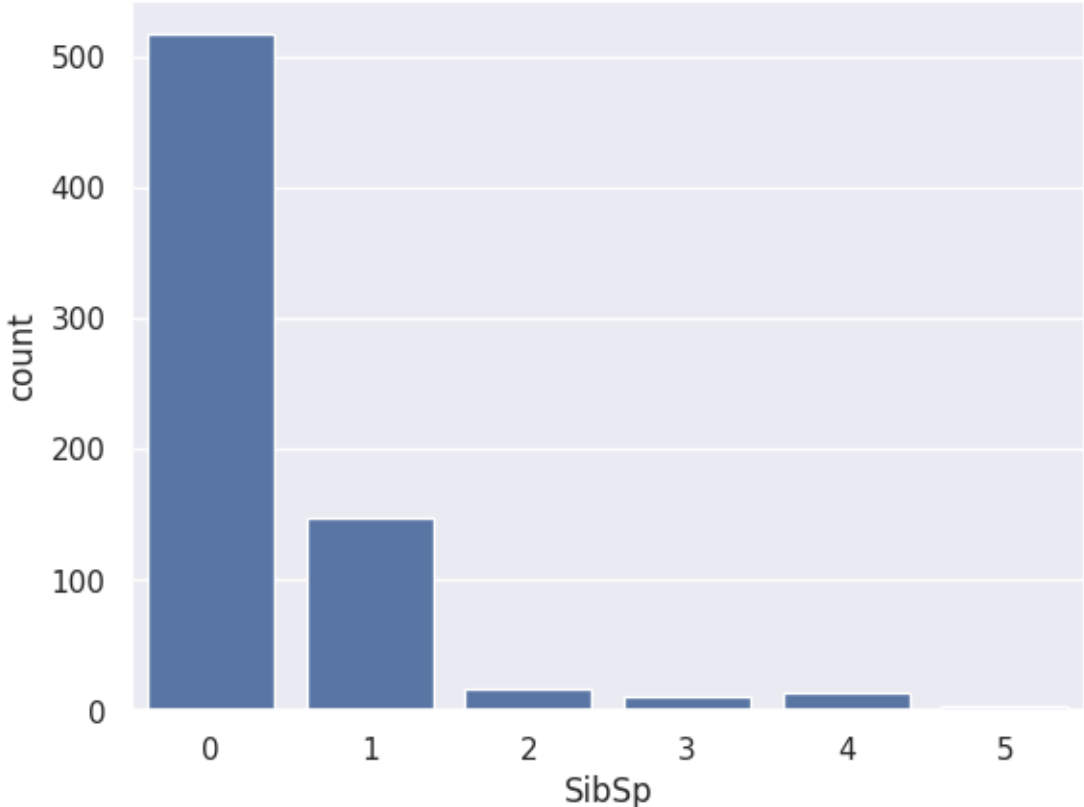


More people are in 3rd class

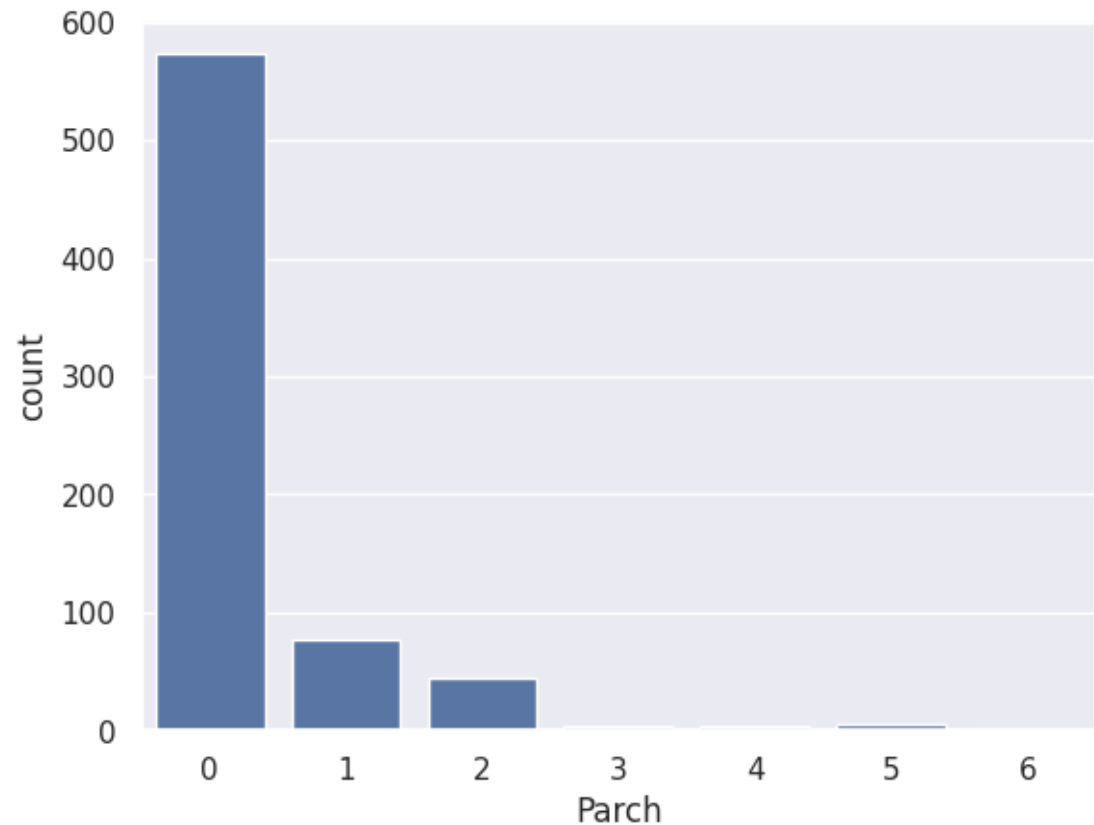


Number of people survived is less

Count Plot for Passengers

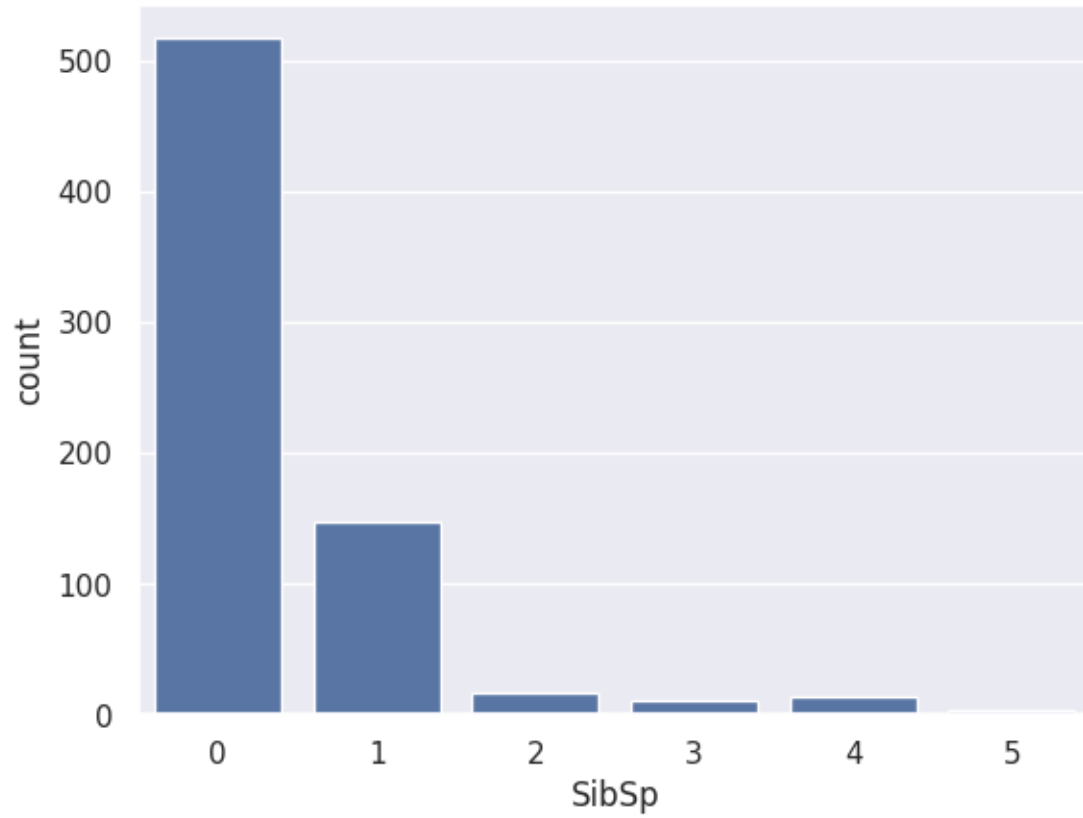


More people are solo travelers

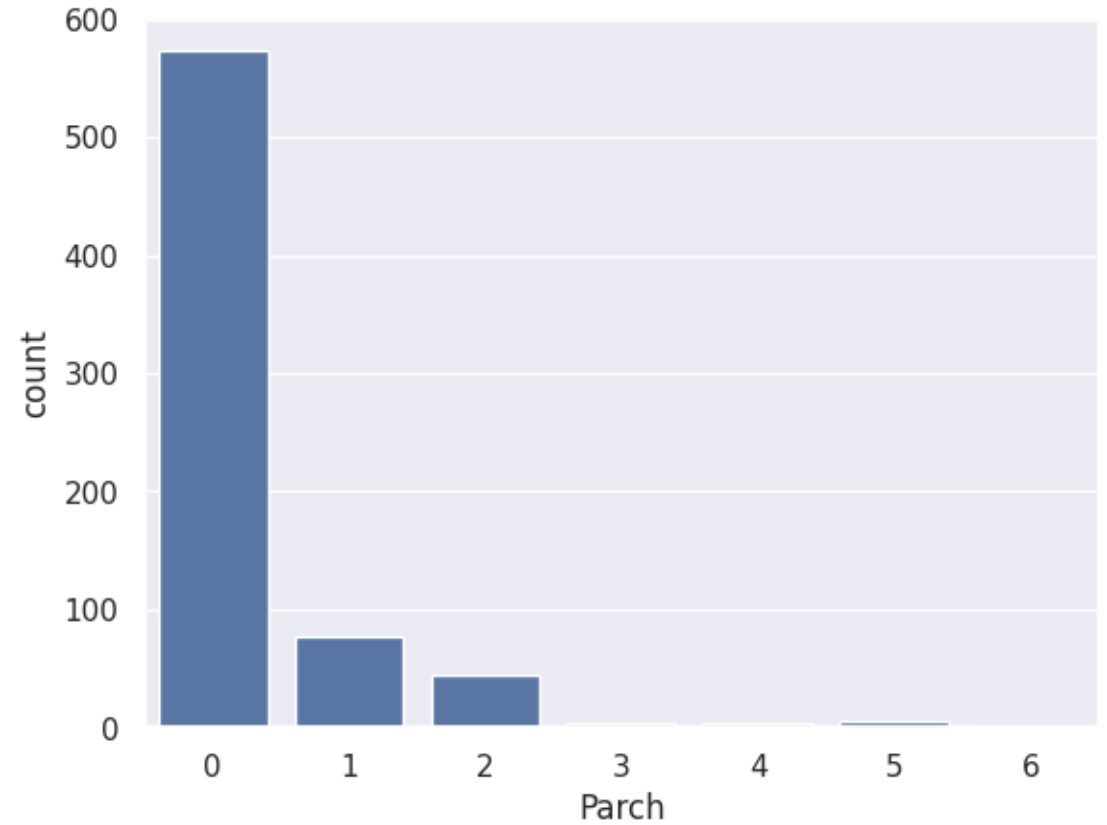


Less number of people are parched

Count Plot for Passengers

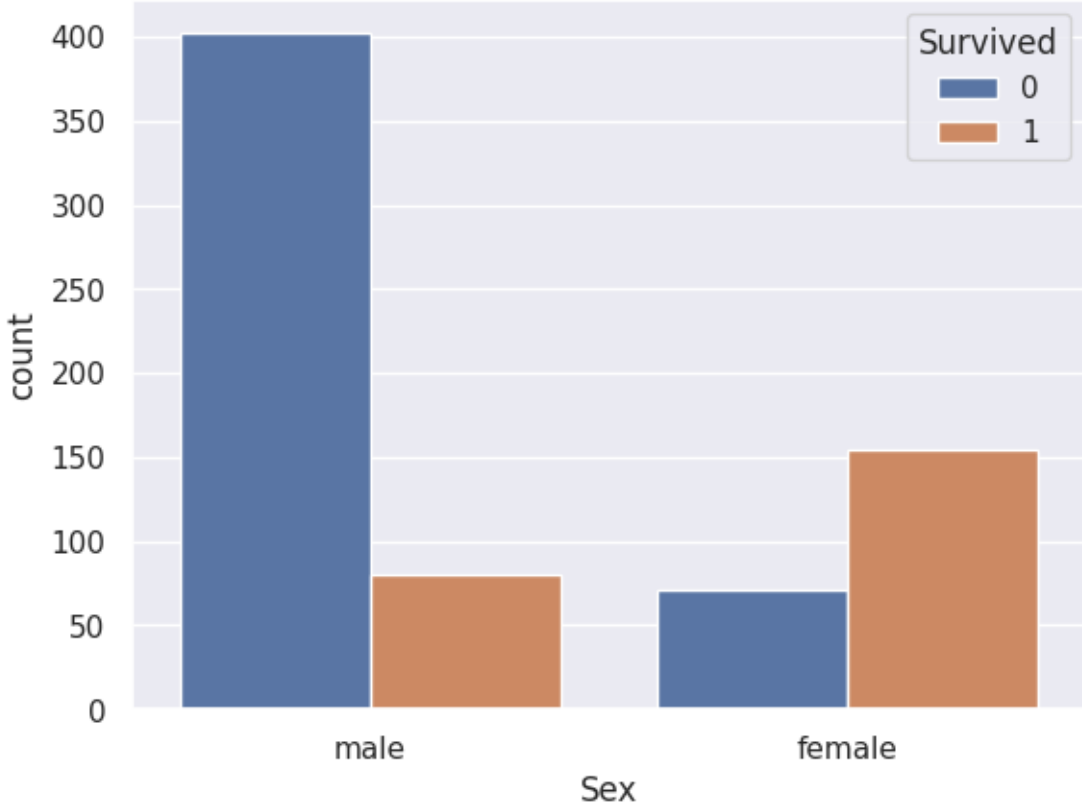


More people are solo
travelers

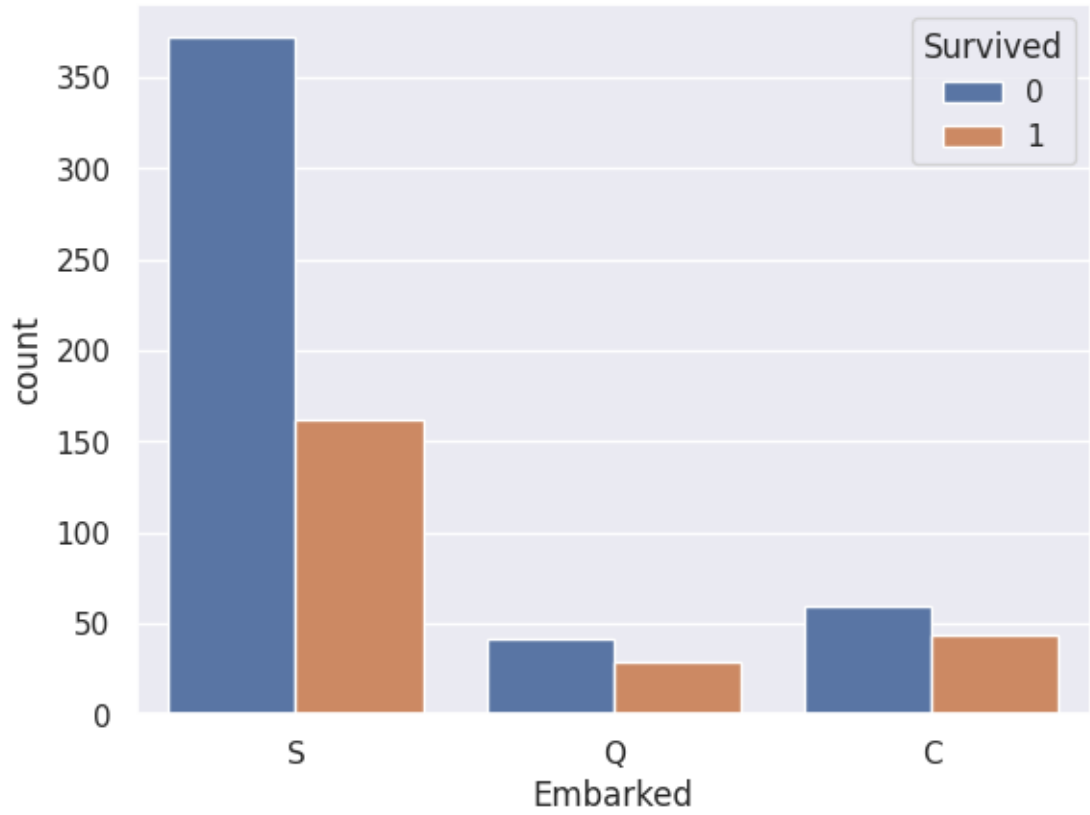


Less number of people
are parched

Count Plot for Passengers

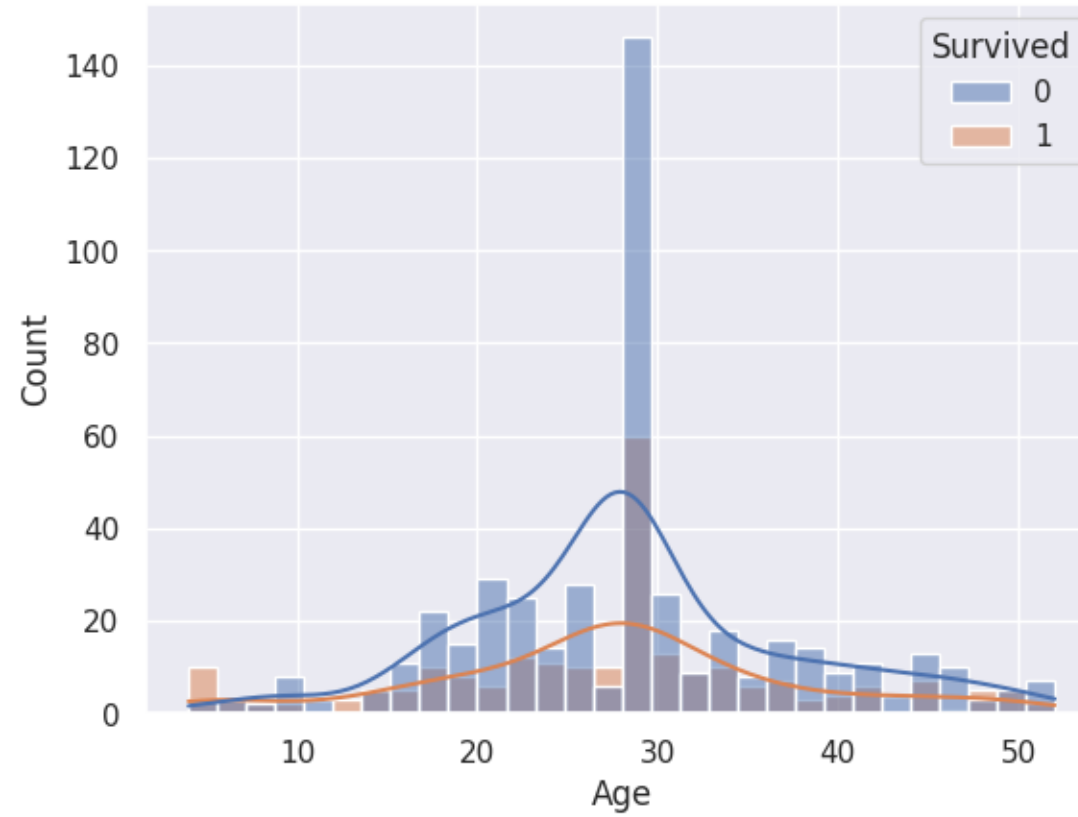


Chances of female passenger survival is more



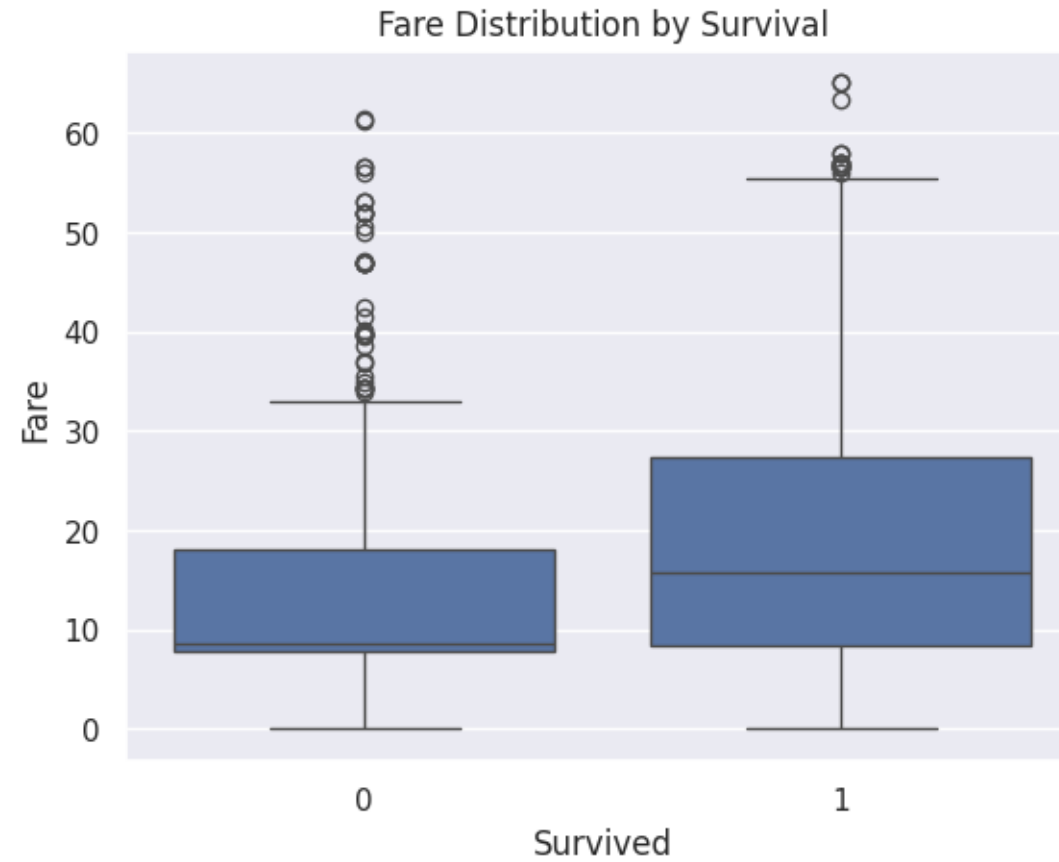
C embarked has more chances of survival

Histogram Plot for Passengers



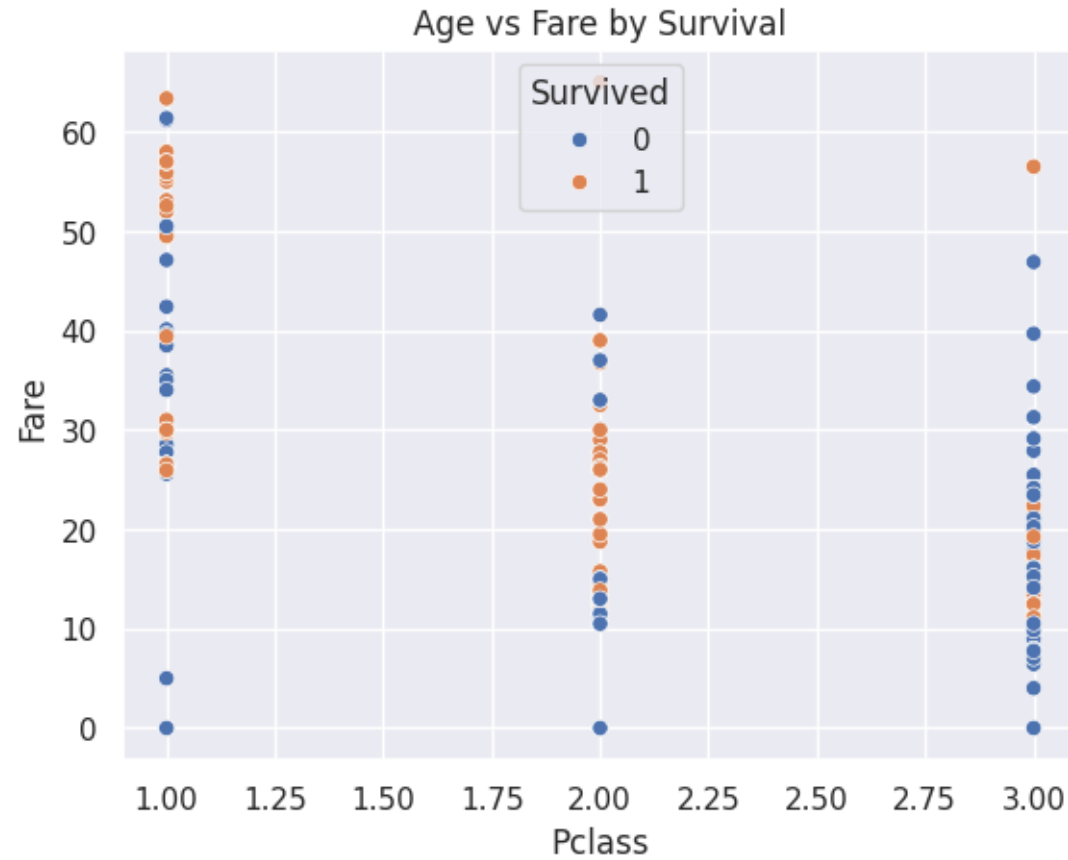
Survived people are less

Box Plot for Passengers



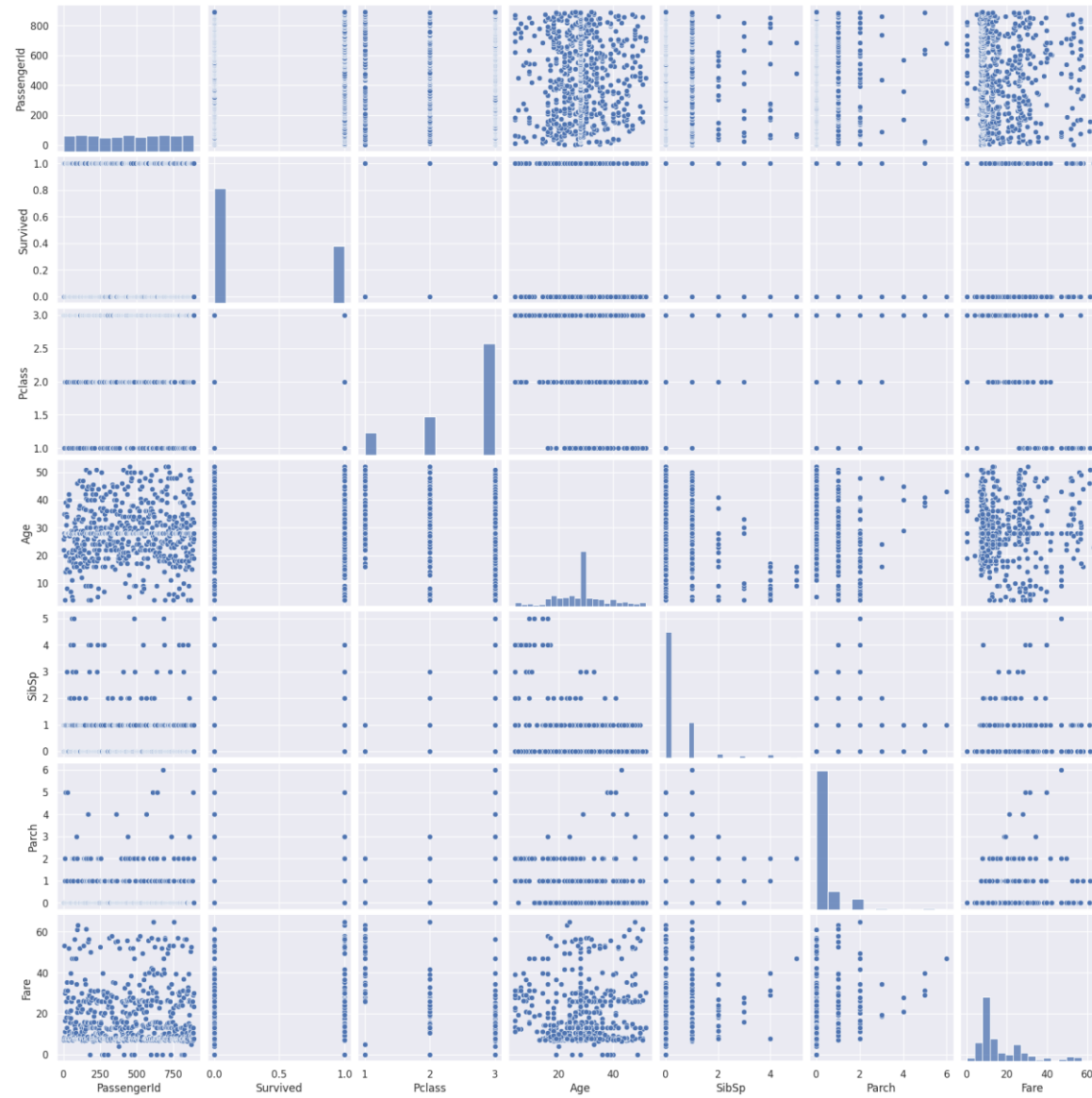
Survived people are those
who pay more fare

Scatter plot for Passengers

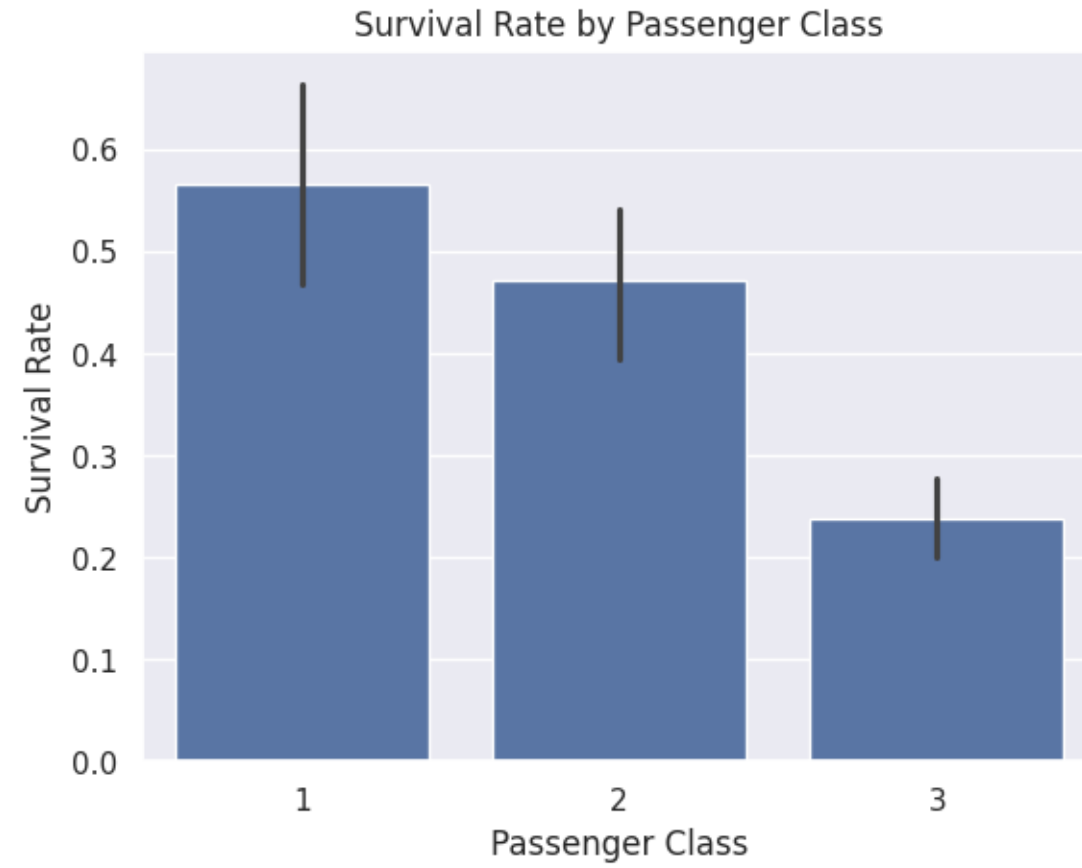


3rd class passenger are
less survived

Pair plot for Passengers



Bar plot for Passengers



Heat map for features

