In [474…

```
import pandas as pd
import numpy as np
import seaborn as sns
import sklearn.metrics
```

In [475…

```
data = pd.read_csv("./Titanic-Dataset.csv")
```

In [476…

```
data.head()
```

Out[476]:

	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 [477…

```
data.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
```

Cabin,Age and Embarked have null values.Data imputation is uncharacteristic for this dataset.So it is better to drop the Cabin column and the rows with any null values in other two columns.

In [478…

```
data.describe()
```

Out[478]:

	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

In [479…

```
data.nunique()
```

Out[479]:

```
PassengerId    891
Survived         2
Pclass           3
Name           891
Sex              2
Age             88
SibSp            7
Parch            7
Ticket         681
Fare           248
Cabin          147
Embarked         3
dtype: int64
```

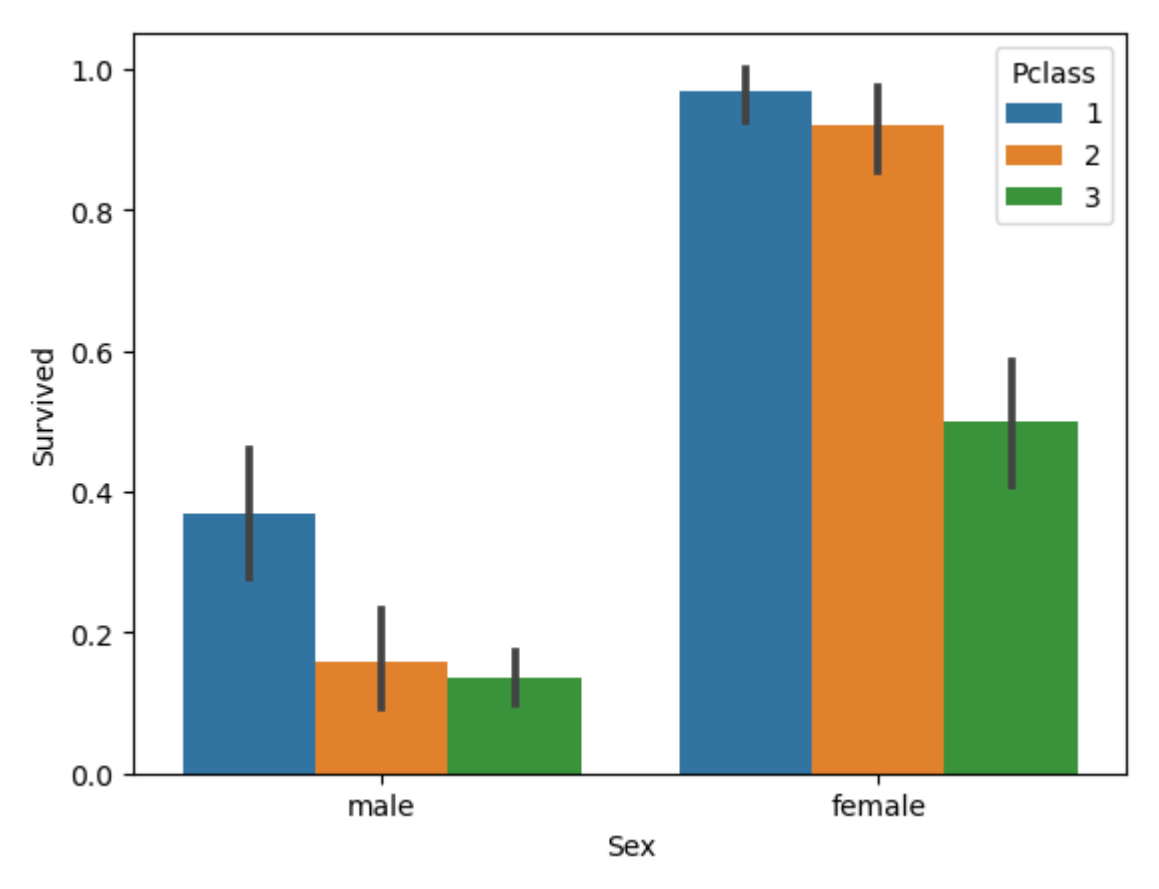
The three remaning obejct typed features are name,sex,ticket and embarked, out of which,Sex has 2 categories and Embarked has 3 catergories.

In [480…

```
sns.barplot(data=data,x='Sex',y='Survived',hue='Pclass')
```

Out[480]:

```
<Axes: xlabel='Sex', ylabel='Survived'>
```



We see that females had a overall more chance of survival than males.

Males from class 1 had a lot more chance of survival than other two classes.

Females from class 3 had a lot less chance of survival than other two classes.

In [481…

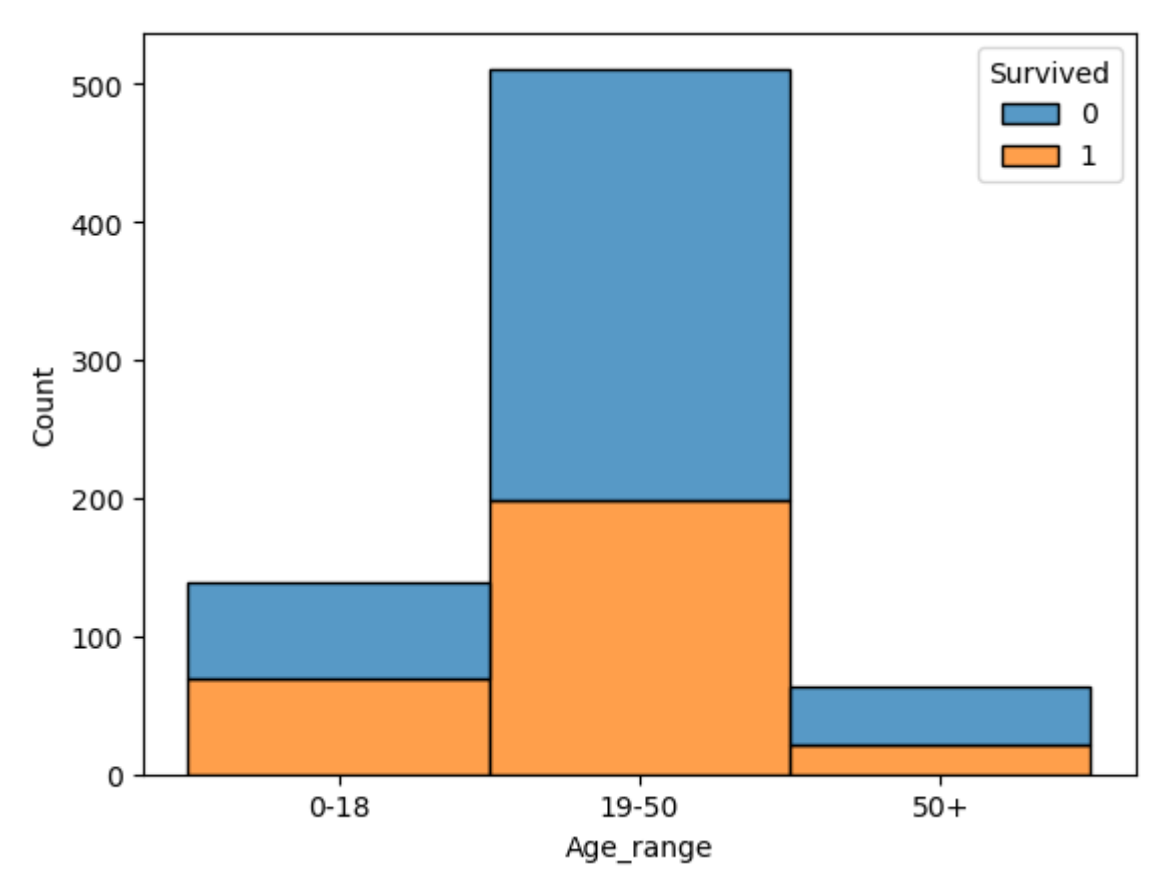
```
data['Age_range'] = pd.cut(data['Age'], bins=[0, 18, 50, 100], labels=['0-18', '19-50', '50+'])
```

In [482…

```
sns.histplot(data=data,x="Age_range",hue="Survived",multiple="stack")
```

Out[482]:

```
<Axes: xlabel='Age_range', ylabel='Count'>
```



Around half the children survived.

Almost no senior citizen survived.

Less than half the middle aged persons survived.

In [483…

```
data['Alone'] = data['SibSp'] + data['Parch']
```

In [484…

```
data.head()
```

Out[484]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Age_range	Alone
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S	19-50	1
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C	19-50	1
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S	19-50	0
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S	19-50	1
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S	19-50	0

In [485…

```
j=0
for i in data.loc[:,"Alone"]:
    if(i!=0):
        data.loc[j,"Alone"] =1
    j+=1
```

In [486…

```
data['Southampton'] = 0
data['Queenstown']  = 0
data['Cherbourg']   = 0
```

In [487…

```
j=0
for i in data.loc[:,"Embarked"]:
    if(i=="S"):
        data.loc[j,"Southampton"] =1
        j+=1
    elif(i=='Q'):
        data.loc[j,"Queenstown"] =1
        j+=1
    else:
        data.loc[j,"Cherbourg"] =1
        j+=1
```

In [488…

```
data['Male'] = 0
data['Female']  = 0
```

In [489…

```
data
```

Out[489]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Age_range	Alone	Southampton	Queenstown	Cherbourg	Male	Female
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S	19-50	1	1	0	0	0	0
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C	19-50	1	0	0	1	0	0
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S	19-50	0	1	0	0	0	0
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S	19-50	1	1	0	0	0	0
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S	19-50	0	1	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S	19-50	0	1	0	0	0	0
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S	19-50	0	1	0	0	0	0
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S	NaN	1	1	0	0	0	0
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C	19-50	0	0	0	1	0	0
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q	19-50	0	0	1	0	0	0

891 rows × 19 columns

In [490…

```
j=0
for i in data.loc[:,"Sex"]:
    if(i=="male"):
        data.iloc[j,-2] =1
        j+=1
    else:
        data.iloc[j,-1] =1
        j+=1
```

In [491…

```
data.dropna(axis=0,subset=["Embarked","Age"],inplace=True)
```

In [492…

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 712 entries, 0 to 890
Data columns (total 19 columns):
 #   Column       Non-Null Count  Dtype   
---  ------       --------------  -----   
 0   PassengerId  712 non-null    int64   
 1   Survived     712 non-null    int64   
 2   Pclass       712 non-null    int64   
 3   Name         712 non-null    object  
 4   Sex          712 non-null    object  
 5   Age          712 non-null    float64 
 6   SibSp        712 non-null    int64   
 7   Parch        712 non-null    int64   
 8   Ticket       712 non-null    object  
 9   Fare         712 non-null    float64 
 10  Cabin        183 non-null    object  
 11  Embarked     712 non-null    object  
 12  Age_range    712 non-null    category
 13  Alone        712 non-null    int64   
 14  Southampton  712 non-null    int64   
 15  Queenstown   712 non-null    int64   
 16  Cherbourg    712 non-null    int64   
 17  Male         712 non-null    int64   
 18  Female       712 non-null    int64   
dtypes: category(1), float64(2), int64(11), object(5)
memory usage: 106.5+ KB
```

In [493…

```
data.drop(['PassengerId','Name','Ticket','Cabin','Age_range','SibSp','Parch','Embarked',"Sex"],axis=1,inplace=True)
```

In [494…

```
data
```

Out[494]:

	Survived	Pclass	Age	Fare	Alone	Southampton	Queenstown	Cherbourg	Male	Female
0	0	3	22.0	7.2500	1	1	0	0	1	0
1	1	1	38.0	71.2833	1	0	0	1	0	1
2	1	3	26.0	7.9250	0	1	0	0	0	1
3	1	1	35.0	53.1000	1	1	0	0	0	1
4	0	3	35.0	8.0500	0	1	0	0	1	0
...	...	...	...	...	...	...	...	...	...	...
885	0	3	39.0	29.1250	1	0	1	0	0	1
886	0	2	27.0	13.0000	0	1	0	0	1	0
887	1	1	19.0	30.0000	0	1	0	0	0	1
889	1	1	26.0	30.0000	0	0	0	1	1	0
890	0	3	32.0	7.7500	0	0	1	0	1	0

712 rows × 10 columns

All columns except Fare and Age have scaled values.

In [495…

```
print('Skewness measure of Age: ',data['Age'].skew())
print('Skewness measure of Fare: ',data['Fare'].skew())
```

```
Skewness measure of Age:  0.38888061787537076
Skewness measure of Fare:  4.667008753630038
```

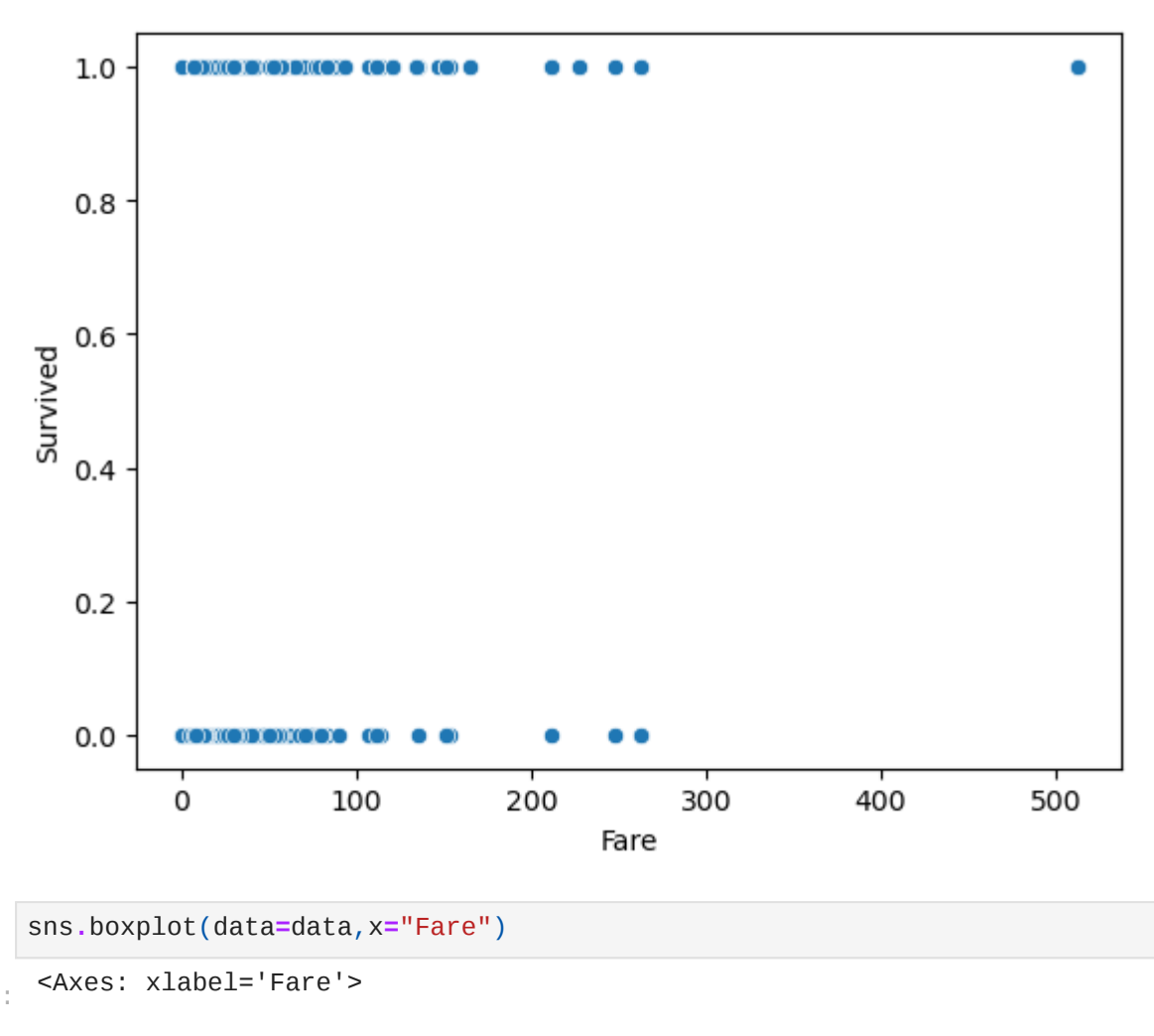
We can see that there are outliers in the FARE column

In [496…

```
sns.scatterplot(data=data,x='Fare',y="Survived")
```

Out[496]:

```
<Axes: xlabel='Fare', ylabel='Survived'>
```

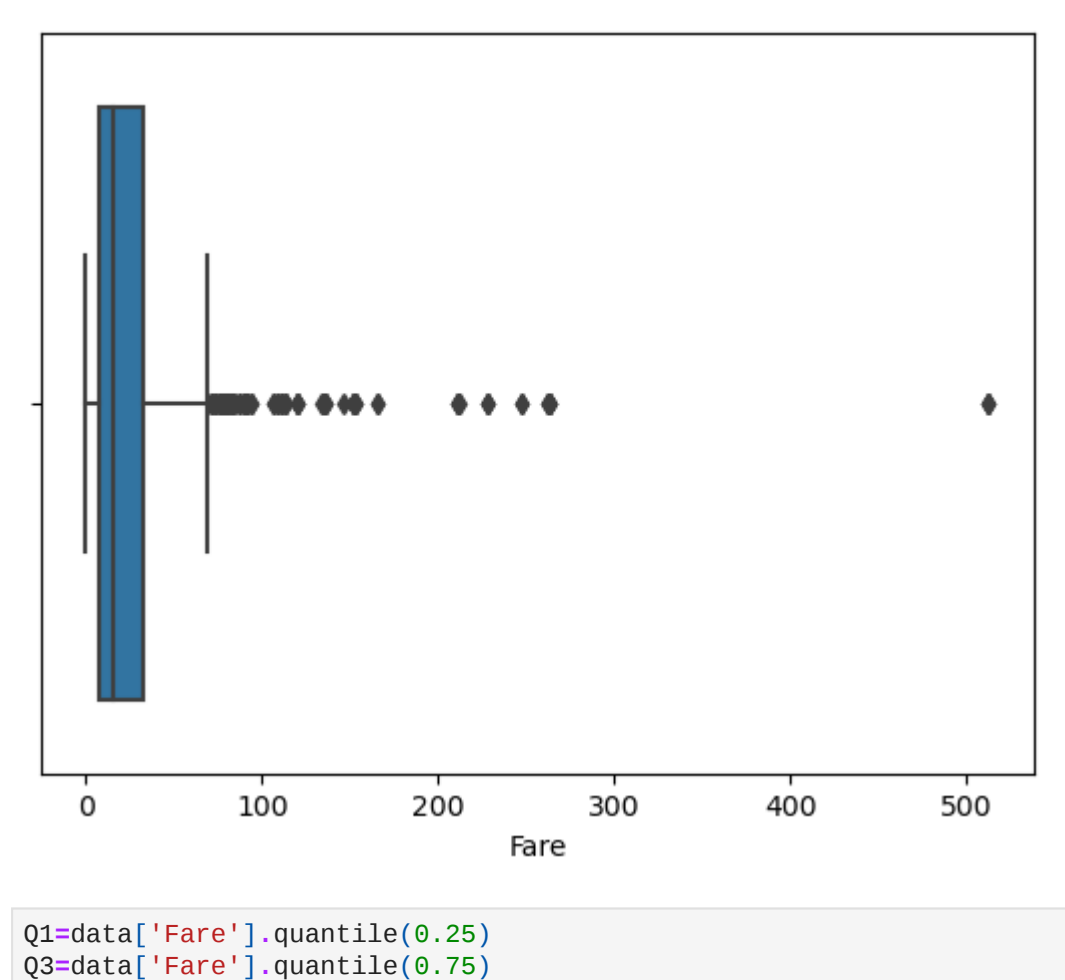


In [497…

```
sns.boxplot(data=data,x="Fare")
```

Out[497]:

```
<Axes: xlabel='Fare'>
```



In [498…

```
Q1=data['Fare'].quantile(0.25)
Q3=data['Fare'].quantile(0.75)
IQR = Q3 - Q1
Minimum_Fare = Q1 - (1.5*IQR)
Maximum_Fare = Q3 + (1.5*IQR)
```

Here we can see the outliers at about Fare>Maximum_Fare

In [499…

```
data[data['Fare']<Minimum_Fare]
```

Out[499]:

Survived	Pclass	Age	Fare	Alone	Southampton	Queenstown	Cherbourg	Male	Female

In [500…

```
data[data['Fare']>Maximum_Fare]
```

Out[500]:

	Survived	Pclass	Age	Fare	Alone	Southampton	Queenstown	Cherbourg	Male	Female
1	1	1	38.0	71.2833	1	0	0	1	0	1
27	0	1	19.0	263.0000	1	1	0	0	1	0
34	0	1	28.0	82.1708	1	0	0	1	1	0
52	1	1	49.0	76.7292	1	0	0	1	0	1
62	0	1	45.0	83.4750	1	1	0	0	1	0
...	...	...	...	...	...	...	...	...	...	...
802	1	1	11.0	120.0000	1	1	0	0	1	0
820	1	1	52.0	93.5000	1	1	0	0	0	1
835	1	1	39.0	83.1583	1	0	0	1	0	1
856	1	1	45.0	164.8667	1	1	0	0	0	1
879	1	1	56.0	83.1583	1	0	0	1	0	1

95 rows × 10 columns

In [501…

```
data.drop(data[data.Fare > Maximum_Fare].index, inplace=True)
```

In [502…

```
print('Skewness measure of Fare: ',data['Fare'].skew())
```

```
Skewness measure of Fare:  1.4089935925930184
```

Outlier have been dealt with.

In [503…

```
from sklearn.preprocessing import MinMaxScaler
mm = MinMaxScaler()
```

In [536…

```
scaled_data = pd.DataFrame(mm.fit_transform(data),columns=data.columns)
```

In [538…

```
X = scaled_data.iloc[:,1:]
y = scaled_data[['Survived']]
```

In [539…

```
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test = train_test_split(X,y,random_state=13,test_size=0.2)
```

In [540…

```
X_train
```

Out[540]:

	Pclass	Age	Fare	Alone	Southampton	Queenstown	Cherbourg	Male	Female
480	1.0	0.321438	0.113336	1.0	1.0	0.0	0.0	1.0	0.0
462	1.0	0.535059	0.676768	1.0	1.0	0.0	0.0	0.0	1.0
106	0.5	0.453380	0.375180	1.0	1.0	0.0	0.0	1.0	0.0
298	0.5	0.522493	0.375180	1.0	1.0	0.0	0.0	0.0	1.0
114	1.0	0.371701	0.116162	0.0	1.0	0.0	0.0	1.0	0.0
...	...	...	...	...	...	...	...	...	...
153	0.0	0.497361	0.447330	0.0	0.0	0.0	1.0	1.0	0.0
528	0.5	0.673285	0.331890	1.0	1.0	0.0	0.0	0.0	1.0
74	1.0	0.258608	0.110390	0.0	1.0	0.0	0.0	0.0	1.0
176	0.5	0.359135	0.151515	0.0	1.0	0.0	0.0	1.0	0.0
338	1.0	0.258608	0.104618	0.0	1.0	0.0	0.0	1.0	0.0

493 rows × 9 columns

In [541…

```
y_train
```

Out[541]:

	Survived
480	0.0
462	0.0
106	0.0
298	1.0
114	0.0
...	...
153	1.0
528	1.0
74	1.0
176	0.0
338	0.0

493 rows × 1 columns

In [ ]: