assignment-3

September 20, 2023

0.1 Abhijeet Bhardwaj

[1]:
```
# import libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

[2]:
```
# create dataframe
df=pd.read_csv("Titanic-Dataset.csv")
df.head()
```

[2]:
```
   PassengerId  Survived  Pclass
0            1         0       3  \
1            2         1       1
2            3         1       3
3            4         1       1
4            5         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

   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
```

[3]:
```
#Checking for null values
df.isnull().any()
```

[3]:
```
PassengerId    False
Survived       False
```

```
Pclass         False
Name           False
Sex            False
Age             True
SibSp          False
Parch          False
Ticket         False
Fare           False
Cabin           True
Embarked        True
dtype: bool
```

[4]:
```
df["Age"].fillna(df["Age"].mean(),inplace=True)
df.head()
```

[4]:
```
   PassengerId  Survived  Pclass
0            1         0       3  \
1            2         1       1
2            3         1       3
3            4         1       1
4            5         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

   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
```

[7]:
```
# filling embarked value
df["Embarked"].fillna(df["Embarked"].mode()[0],inplace=True)

#filling cabin null values
df["Cabin"].fillna(df["Cabin"].mode()[1],inplace=True)
df.head()
```

[7]:
```
   PassengerId  Survived  Pclass
0            1         0       3  \
1            2         1       1
2            3         1       3
```

```
3            4         1       1
4            5         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

   Parch            Ticket     Fare        Cabin Embarked
0      0         A/5 21171   7.2500  C23 C25 C27        S
1      0          PC 17599  71.2833          C85        C
2      0  STON/O2. 3101282   7.9250  C23 C25 C27        S
3      0            113803  53.1000         C123        S
4      0            373450   8.0500  C23 C25 C27        S
```

[8]:
```
df.isnull().sum()
```

[8]:
```
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
Fare           0
Cabin          0
Embarked       0
dtype: int64
```

[10]:
```
df.isnull().any()
```

[10]:
```
PassengerId    False
Survived       False
Pclass         False
Name           False
Sex            False
Age            False
SibSp          False
Parch          False
Ticket         False
Fare           False
Cabin          False
Embarked       False
```

```
dtype: bool
```

the above value shows there is no more null values

0.1.1 Data Visualization

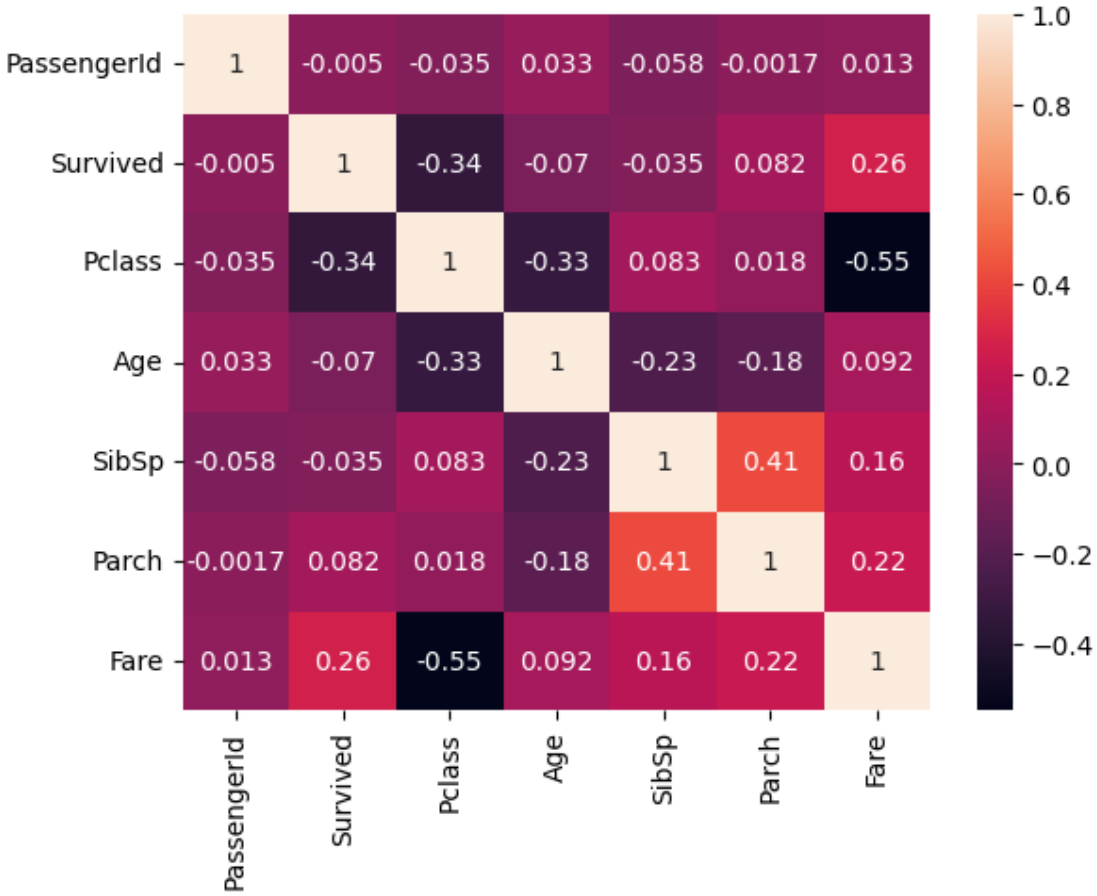
[14]:
```
test_data = df.drop(columns=['Name', 'Sex', 'Embarked','Ticket', 'Cabin'])
corr=test_data.corr()
corr
```

[14]:
```
             PassengerId  Survived    Pclass       Age     SibSp     Parch  \
PassengerId     1.000000 -0.005007 -0.035144  0.033207 -0.057527 -0.001652   \
Survived       -0.005007  1.000000 -0.338481 -0.069809 -0.035322  0.081629
Pclass         -0.035144 -0.338481  1.000000 -0.331339  0.083081  0.018443
Age             0.033207 -0.069809 -0.331339  1.000000 -0.232625 -0.179191
SibSp          -0.057527 -0.035322  0.083081 -0.232625  1.000000  0.414838
Parch          -0.001652  0.081629  0.018443 -0.179191  0.414838  1.000000
Fare            0.012658  0.257307 -0.549500  0.091566  0.159651  0.216225

                 Fare
PassengerId  0.012658
Survived     0.257307
Pclass      -0.549500
Age          0.091566
SibSp        0.159651
Parch        0.216225
Fare         1.000000
```

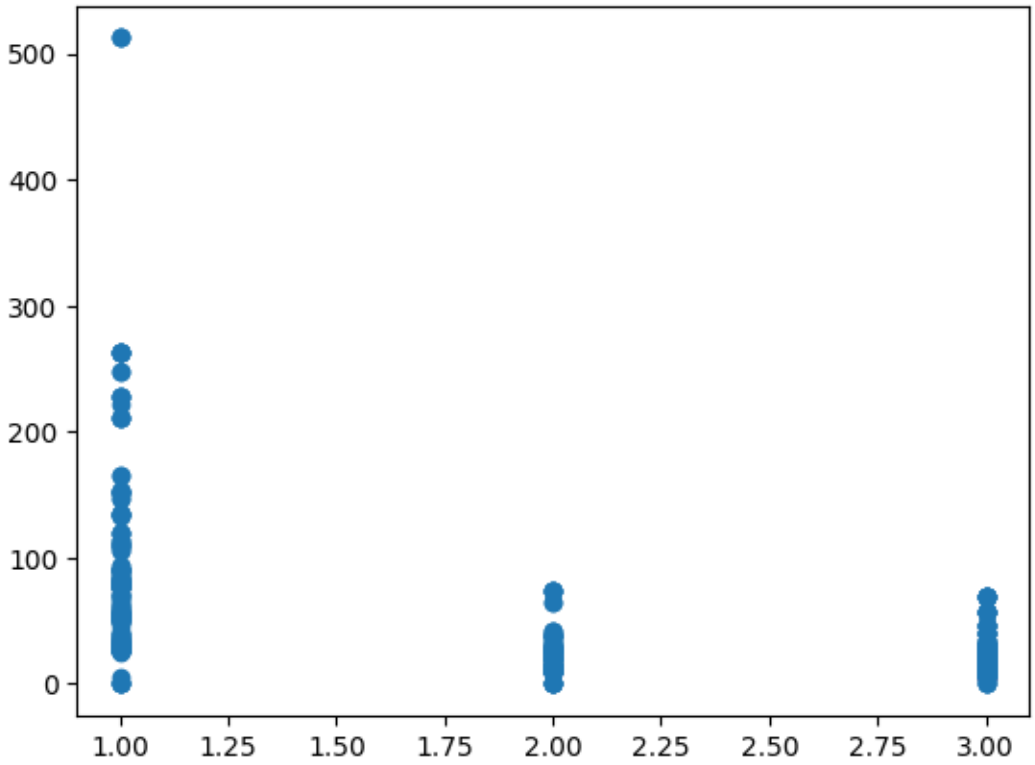
[16]:
```
sns.heatmap(corr,annot=True)
```

[16]:
```
<Axes: >
```



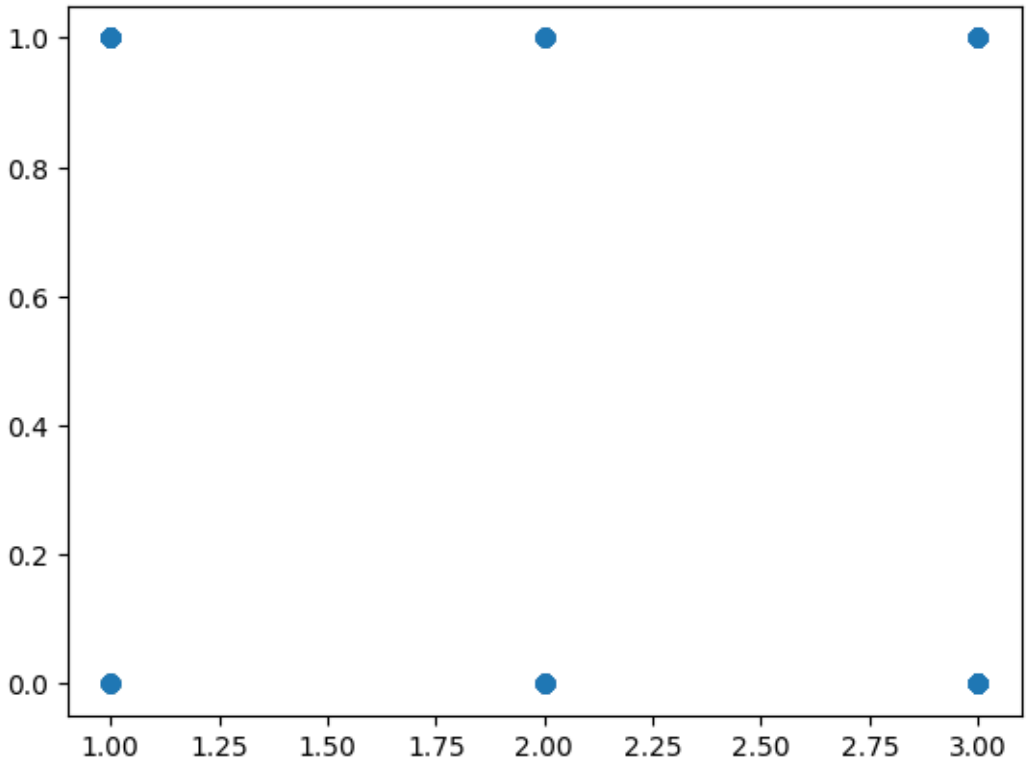
[26]:
```
# we see the fare is closely related to class
plt.scatter(df["Pclass"],df["Fare"])
```

[26]: <matplotlib.collections.PathCollection at 0x16ce602b0>



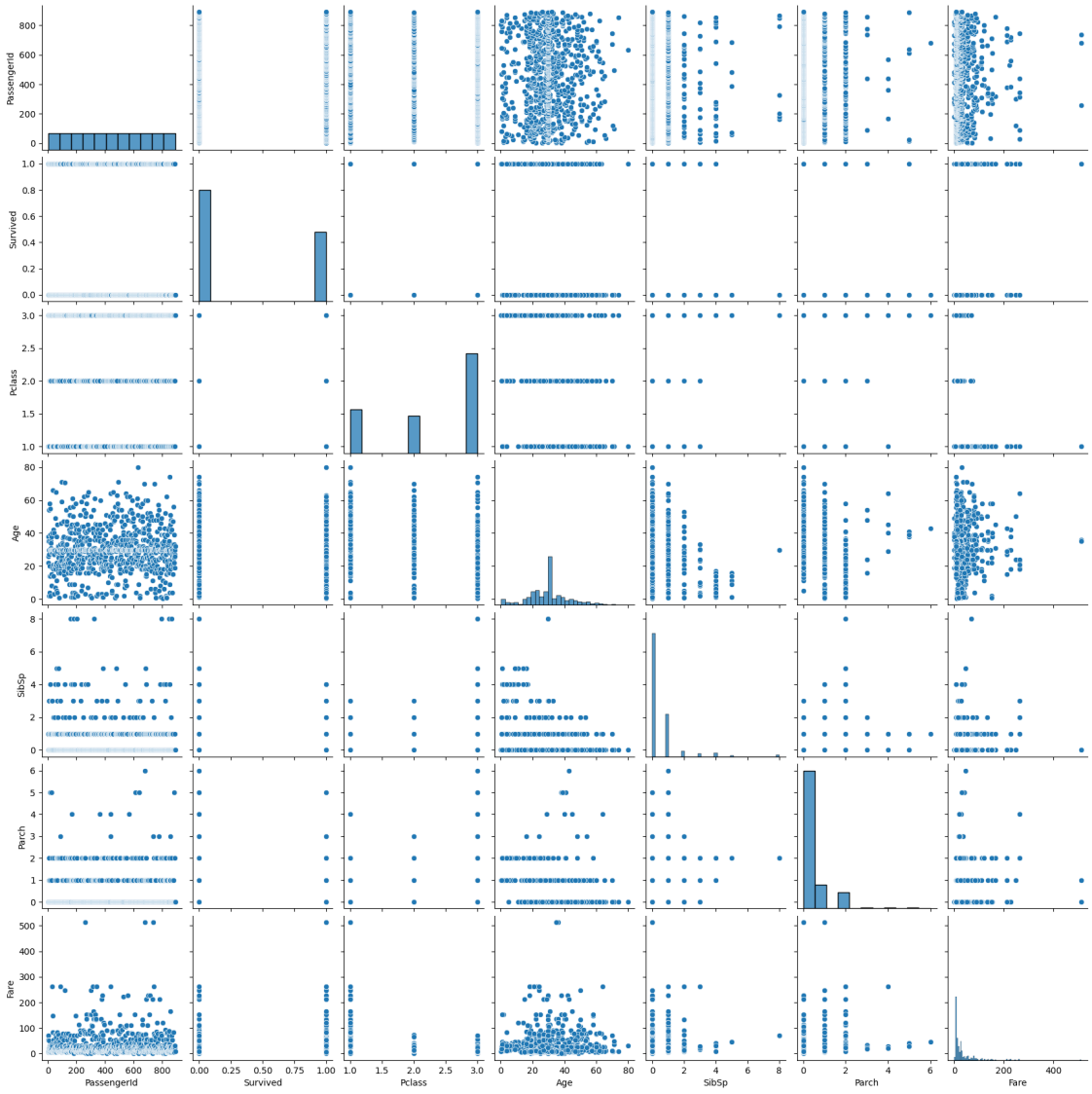
[27]:
```
# let's check if passengers with heigh class ticket survived.
plt.scatter(df["Pclass"], df["Survived"])
```

[27]: <matplotlib.collections.PathCollection at 0x16cec2bc0>



[28]:
```
sns.pairplot(df)
```

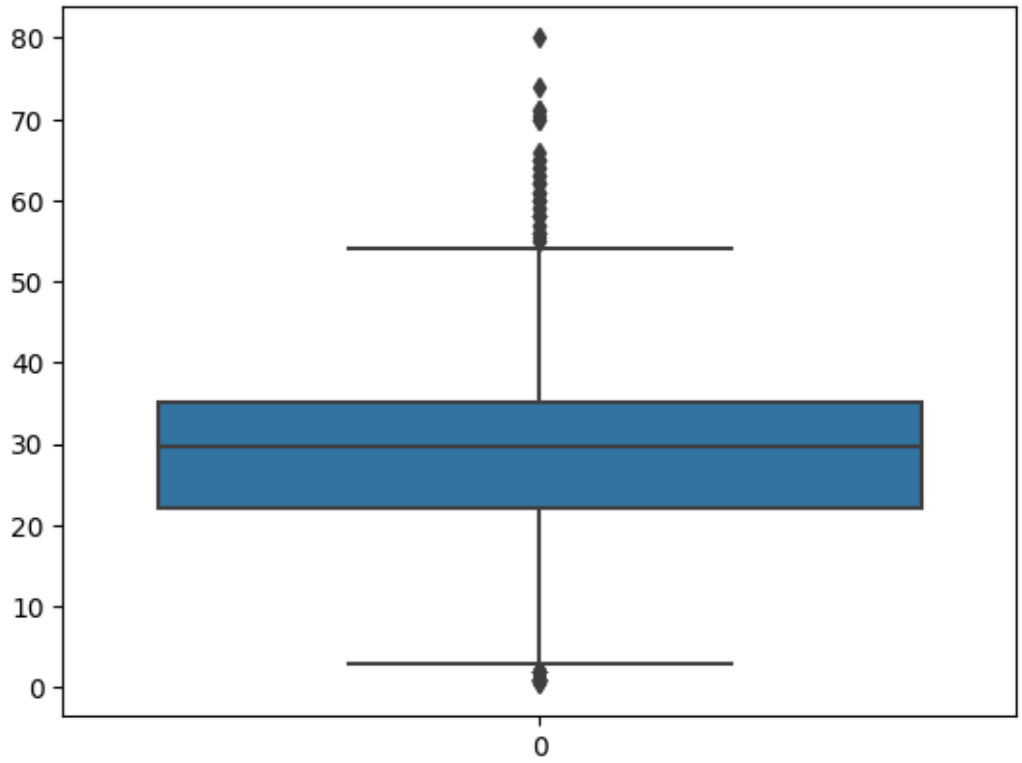
[28]: <seaborn.axisgrid.PairGrid at 0x16cf06380>



0.1.2 Outlier Detection

[30]:
```
sns.boxplot(df.Age)
```

[30]: <Axes: >



[32]:
```
q1=df.Age.quantile(0.25)
q1
```

[32]: 22.0

[33]:
```
q2=df.Age.quantile(0.50)
q2
```

[33]: 29.69911764705882

[34]:
```
q3=df.Age.quantile(0.75)
q3
```

[34]: 35.0

[35]:
```
IQR=q3-q1
```

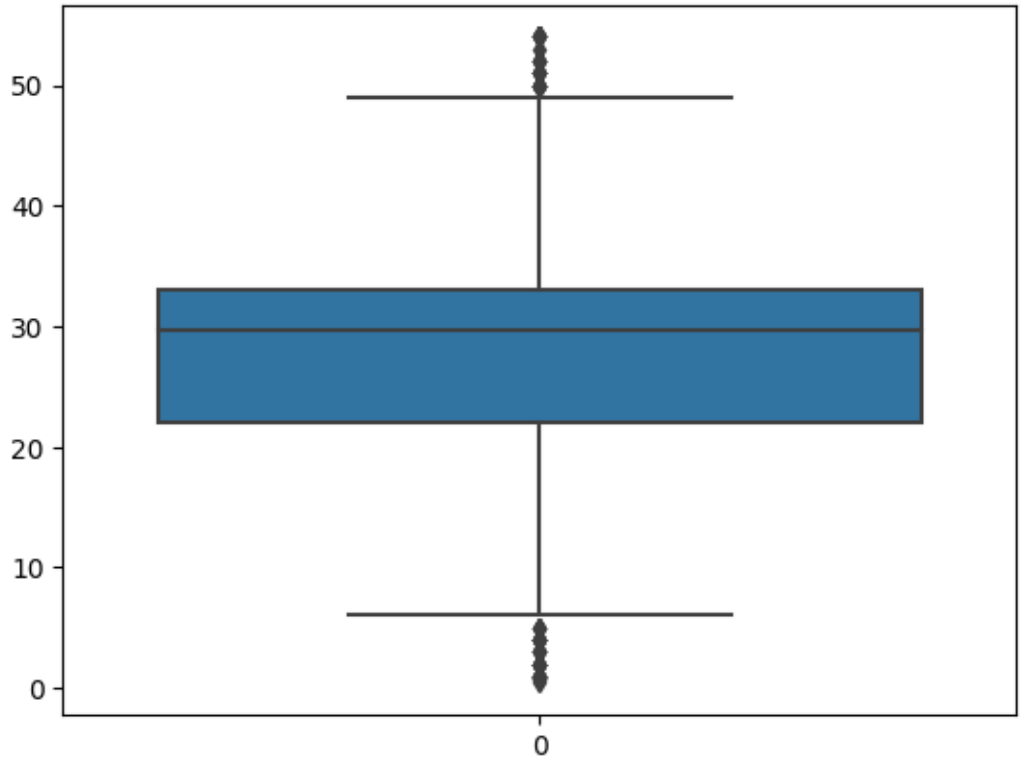
[36]:
```
upper_limit=q3+1.5*IQR
upper_limit
```

[36]: 54.5

[37]:
```
df['Age']=np.where(df['Age']>upper_limit,30,df['Age'])
```

[38]:
```
sns.boxplot(df.Age)
```

[38]:
```
<Axes: >
```



0.1.3 Splitting Dependent and Independent variables

[39]:
```
#datset.iloc[rows,column]
x=df.iloc[:,3:13]
y=df.iloc[:,1:2]
```

[40]:
```
y.head()
```

[40]:

	Survived
0	0
1	1
2	1
3	1
4	0

[41]:
```
x.head()
```

[41]:
```
                                                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

   Parch            Ticket     Fare        Cabin Embarked
0      0         A/5 21171   7.2500  C23 C25 C27        S
1      0          PC 17599  71.2833          C85        C
2      0  STON/O2. 3101282   7.9250  C23 C25 C27        S
3      0            113803  53.1000         C123        S
4      0            373450   8.0500  C23 C25 C27        S
```

0.1.4 Perform Encoding

[42]:
```
from sklearn.preprocessing import LabelEncoder
le=LabelEncoder()
x["Sex"]=le.fit_transform(x["Sex"])
x["Sex"]
```

[42]:
```
0      1
1      0
2      0
3      0
4      1
      ..
886    1
887    0
888    0
889    1
890    1
Name: Sex, Length: 891, dtype: int64
```

[43]:
```
x.head()
```

[43]:
```
                                                Name  Sex   Age  SibSp  Parch
0                            Braund, Mr. Owen Harris    1  22.0      1      0  \
1  Cumings, Mrs. John Bradley (Florence Briggs Th…    0  38.0      1      0
2                             Heikkinen, Miss. Laina    0  26.0      0      0
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)    0  35.0      1      0
4                           Allen, Mr. William Henry    1  35.0      0      0

             Ticket     Fare        Cabin Embarked
0         A/5 21171   7.2500  C23 C25 C27        S
1          PC 17599  71.2833          C85        C
2  STON/O2. 3101282   7.9250  C23 C25 C27        S
```

```
3            113803  53.1000         C123        S
4            373450   8.0500  C23 C25 C27        S
```

[44]:
```
x.Embarked.value_counts()
```

[44]:
```
Embarked
S    646
C    168
Q     77
Name: count, dtype: int64
```

0.1.5 One hot encoding on geography column

[45]:
```
Embarked=pd.get_dummies(x["Embarked"],drop_first=True)
Embarked
```

[45]:
```
         Q      S
0    False   True
1    False  False
2    False   True
3    False   True
4    False   True
..     …    …
886  False   True
887  False   True
888  False   True
889  False  False
890   True  False

[891 rows x 2 columns]
```

[46]:
```
x=pd.concat([x,Embarked],axis=1)
```

[47]:
```
x.head()
```

[47]:
```
                                                Name  Sex   Age  SibSp  Parch  \
0                            Braund, Mr. Owen Harris    1  22.0      1      0  \
1  Cumings, Mrs. John Bradley (Florence Briggs Th…    0  38.0      1      0
2                             Heikkinen, Miss. Laina    0  26.0      0      0
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)    0  35.0      1      0
4                           Allen, Mr. William Henry    1  35.0      0      0

             Ticket     Fare        Cabin Embarked      Q      S
0         A/5 21171   7.2500  C23 C25 C27        S  False   True
1          PC 17599  71.2833          C85        C  False  False
2  STON/O2. 3101282   7.9250  C23 C25 C27        S  False   True
3            113803  53.1000         C123        S  False   True
```

```
4            373450   8.0500  C23 C25 C27        S  False   True
```

[48]:
```
x.drop(["Embarked"],axis=1,inplace=True)
x.head(6)
```

[48]:
```
                                                 Name  Sex        Age  SibSp  \
0                             Braund, Mr. Owen Harris    1  22.000000      1
1  Cumings, Mrs. John Bradley (Florence Briggs Th…    0  38.000000      1
2                              Heikkinen, Miss. Laina    0  26.000000      0
3        Futrelle, Mrs. Jacques Heath (Lily May Peel)    0  35.000000      1
4                            Allen, Mr. William Henry    1  35.000000      0
5                                    Moran, Mr. James    1  29.699118      0

   Parch            Ticket     Fare        Cabin      Q      S
0      0         A/5 21171   7.2500  C23 C25 C27  False   True
1      0          PC 17599  71.2833          C85  False  False
2      0  STON/O2. 3101282   7.9250  C23 C25 C27  False   True
3      0            113803  53.1000         C123  False   True
4      0            373450   8.0500  C23 C25 C27  False   True
5      0            330877   8.4583  C23 C25 C27   True  False
```

0.2 Splitting Data into Train and Test

[49]:
```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.5,random_state=0)
x_train.shape,x_test.shape,y_train.shape,y_test.shape
```

[49]:
```
((445, 10), (446, 10), (445, 1), (446, 1))
```

[ ]:
```

```