assignment-03-smartbridge

September 20, 2023

0.0.1 Name : Maguluri Venkata Siva Rama Krishna

0.0.2 Reg no : 21BCE9322

0.0.3 Assignment - 3

0.0.4 Importing Libraries

[4]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

0.0.5 Importing DataSet

[5]:
```
df = pd.read_csv('D:\Smartbridge_Externship\Titanic-Dataset.csv')
```

[6]:
```
df
```

[6]:
```
     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]
```

[7]:

```
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   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
```

[8]:

```
df.tail()
```

[8]:

```
     PassengerId  Survived  Pclass                                      Name  \
886          887         0       2                     Montvila, Rev. Juozas
887          888         1       1              Graham, Miss. Margaret Edith
888          889         0       3  Johnston, Miss. Catherine Helen "Carrie"
889          890         1       1                     Behr, Mr. Karl Howell
890          891         0       3                       Dooley, Mr. Patrick

        Sex   Age  SibSp  Parch      Ticket   Fare Cabin Embarked
886    male  27.0      0      0      211536  13.00   NaN        S
887  female  19.0      0      0      112053  30.00   B42        S
888  female   NaN      1      2  W./C. 6607  23.45   NaN        S
889    male  26.0      0      0      111369  30.00  C148        C
890    male  32.0      0      0      370376   7.75   NaN        Q
```

[9]:

```python
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
```

[10]:

```python
df.describe()
```

[10]:

```
       PassengerId    Survived      Pclass         Age       SibSp  \
count   891.000000  891.000000  891.000000  714.000000  891.000000
mean    446.000000    0.383838    2.308642   29.699118    0.523008
std     257.353842    0.486592    0.836071   14.526497    1.102743
min       1.000000    0.000000    1.000000    0.420000    0.000000
25%     223.500000    0.000000    2.000000   20.125000    0.000000
50%     446.000000    0.000000    3.000000   28.000000    0.000000
75%     668.500000    1.000000    3.000000   38.000000    1.000000
max     891.000000    1.000000    3.000000   80.000000    8.000000
```

```
            Parch        Fare
count  891.000000  891.000000
mean     0.381594   32.204208
std      0.806057   49.693429
min      0.000000    0.000000
25%      0.000000    7.910400
50%      0.000000   14.454200
75%      0.000000   31.000000
max      6.000000  512.329200
```

[11]:

```
df.shape
```

[11]: (891, 12)

[12]:

```
numeric_df = df.select_dtypes(include=['number'])
correlation_matrix = numeric_df.corr()
print(correlation_matrix)
```

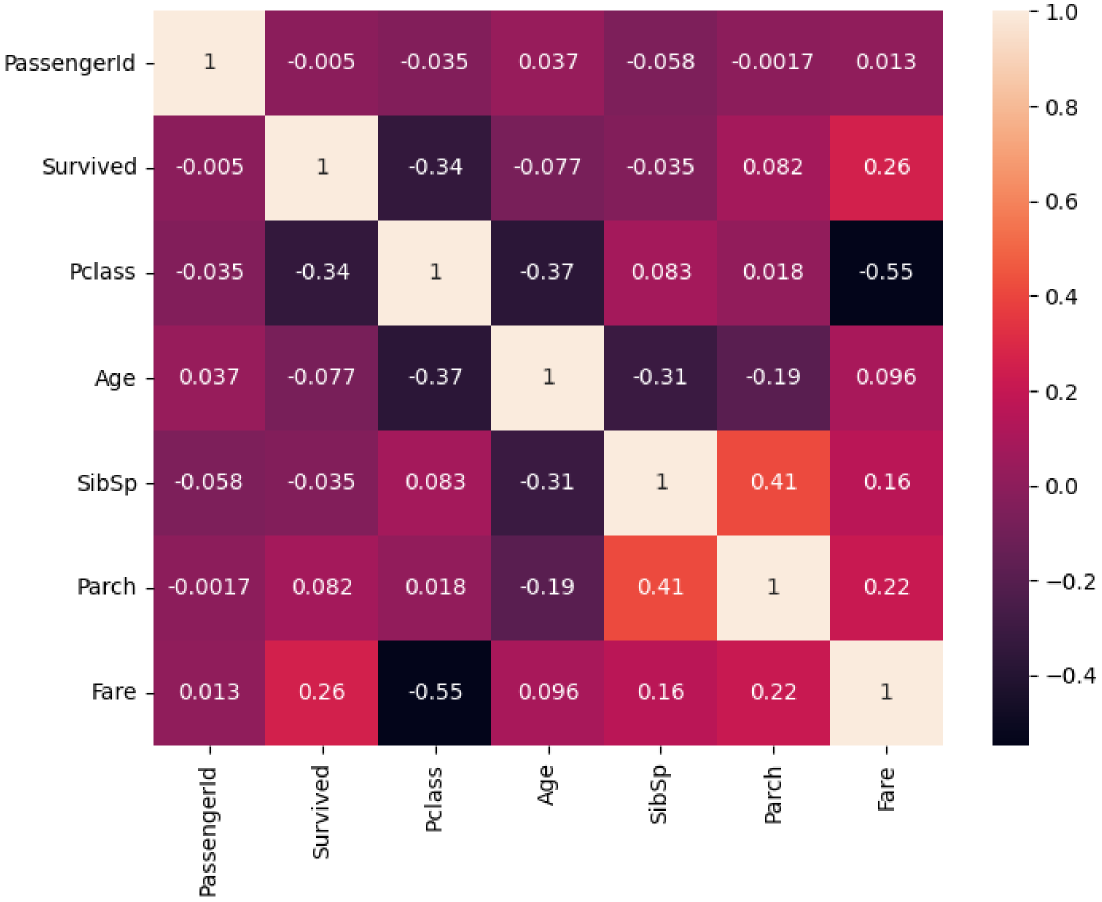
```
             PassengerId  Survived    Pclass       Age     SibSp     Parch  \
PassengerId     1.000000 -0.005007 -0.035144  0.036847 -0.057527 -0.001652
Survived       -0.005007  1.000000 -0.338481 -0.077221 -0.035322  0.081629
Pclass         -0.035144 -0.338481  1.000000 -0.369226  0.083081  0.018443
Age             0.036847 -0.077221 -0.369226  1.000000 -0.308247 -0.189119
SibSp          -0.057527 -0.035322  0.083081 -0.308247  1.000000  0.414838
Parch          -0.001652  0.081629  0.018443 -0.189119  0.414838  1.000000
Fare            0.012658  0.257307 -0.549500  0.096067  0.159651  0.216225

                 Fare
PassengerId  0.012658
Survived     0.257307
Pclass      -0.549500
Age          0.096067
SibSp        0.159651
Parch        0.216225
Fare         1.000000
```

[13]:

```
plt.subplots(figsize=(8,6))
sns.heatmap(correlation_matrix,annot=True)
```

[13]: <Axes: >



0.0.6 Checking and Handling Null Values

[14]:
```
df.isnull().any()
```

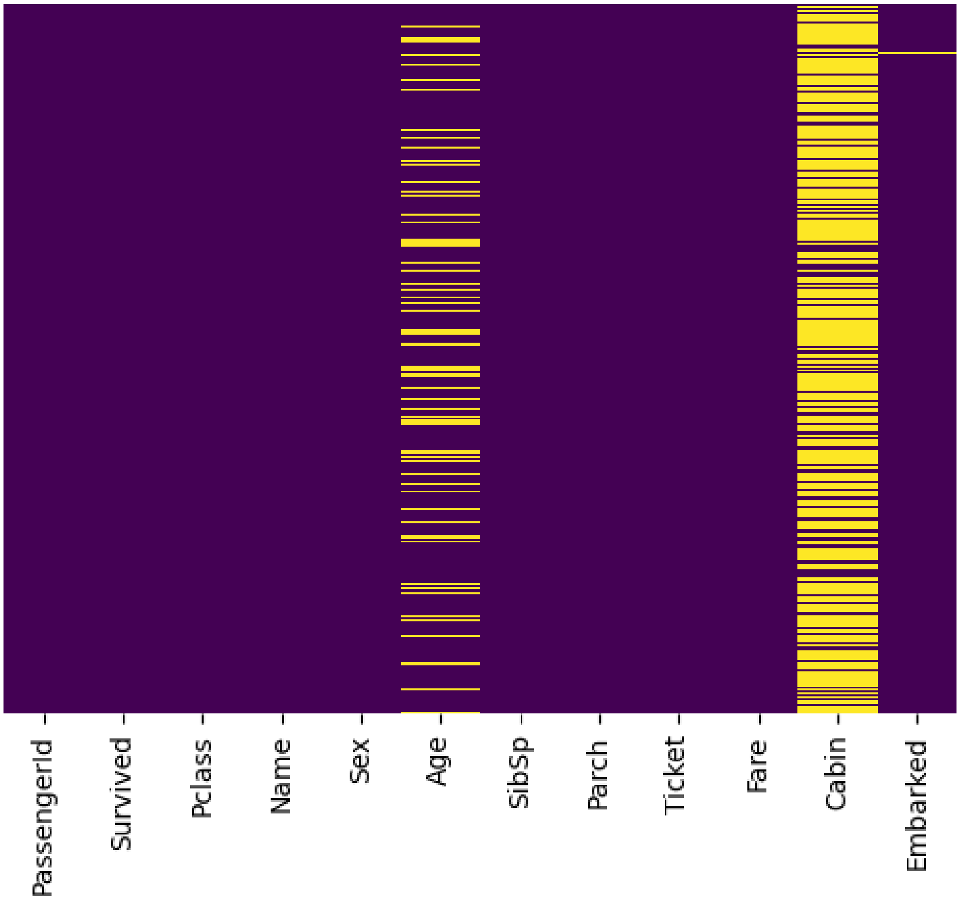
[14]:
```
PassengerId    False
Survived       False
Pclass         False
Name           False
Sex            False
Age             True
SibSp          False
Parch          False
Ticket         False
Fare           False
Cabin           True
Embarked        True
dtype: bool
```

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

```
[15]: 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
```

```
[16]: #Heatmap Representation of null values.
      sns.heatmap(df.isnull(),yticklabels=False,cbar=False,cmap='viridis')
```

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



[17]:

```
numeric_columns = df.select_dtypes(include=['number']).columns
df[numeric_columns] = df[numeric_columns].fillna(df[numeric_columns].mean())

# Fill missing values in the "Embarked" column with the mode
df["Embarked"].fillna(df["Embarked"].mode()[0], inplace=True)

print(df.isnull().sum())
print("\n")
df.head()
```

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age              0
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         0
dtype: int64
```

[17]:

```
   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
```

[18]:

```
df.drop(labels = ["Cabin", "Name"], axis=1).head()
```

[18]:

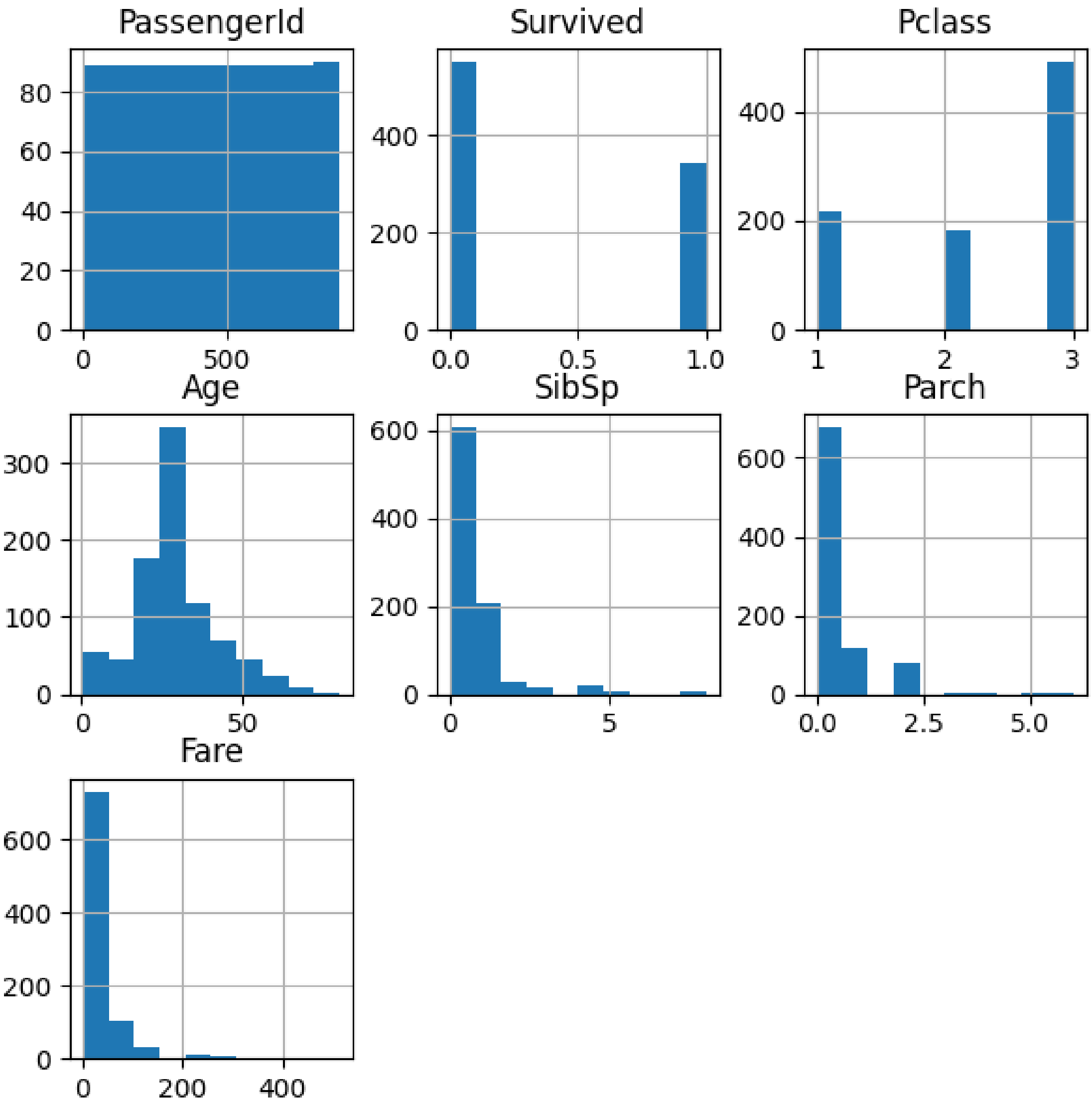
```
   PassengerId  Survived  Pclass     Sex   Age  SibSp  Parch  \
0            1         0       3    male  22.0      1      0
1            2         1       1  female  38.0      1      0
2            3         1       3  female  26.0      0      0
3            4         1       1  female  35.0      1      0
4            5         0       3    male  35.0      0      0

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

0.0.7 Data Visualization.

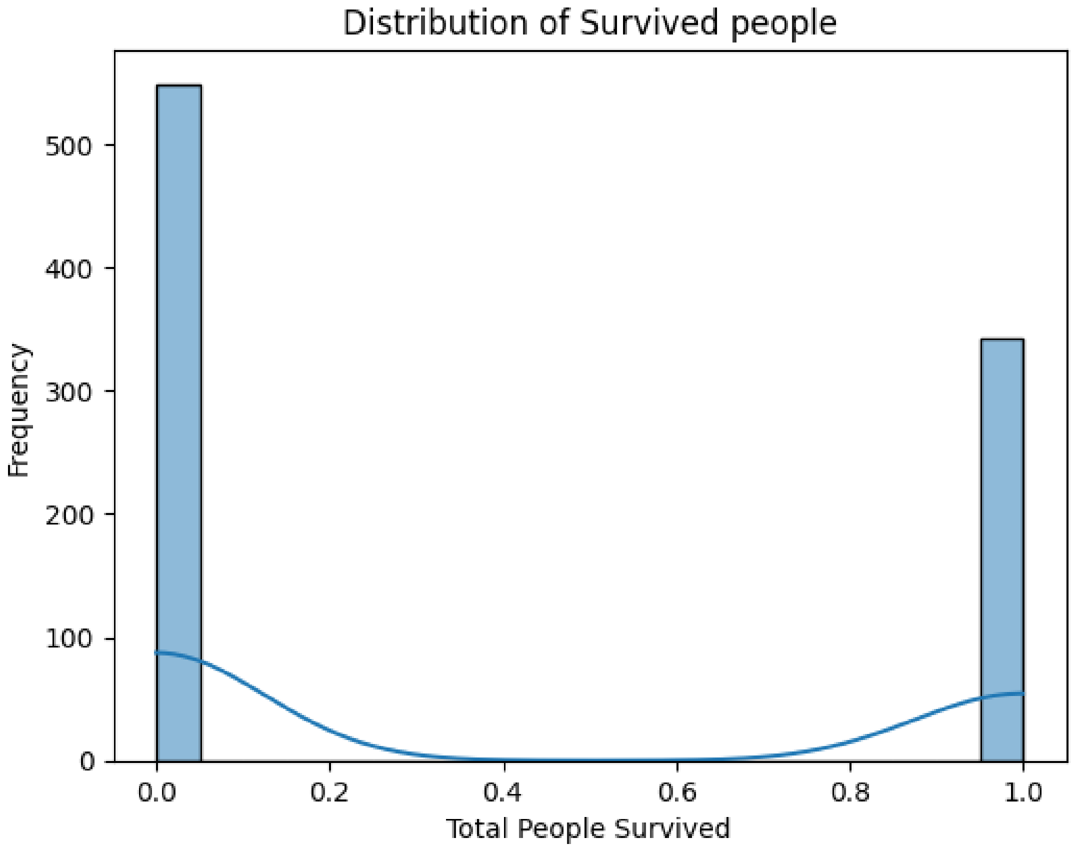
[19]:

```
df.hist(figsize=(7,7))
plt.show()
```



[20]:

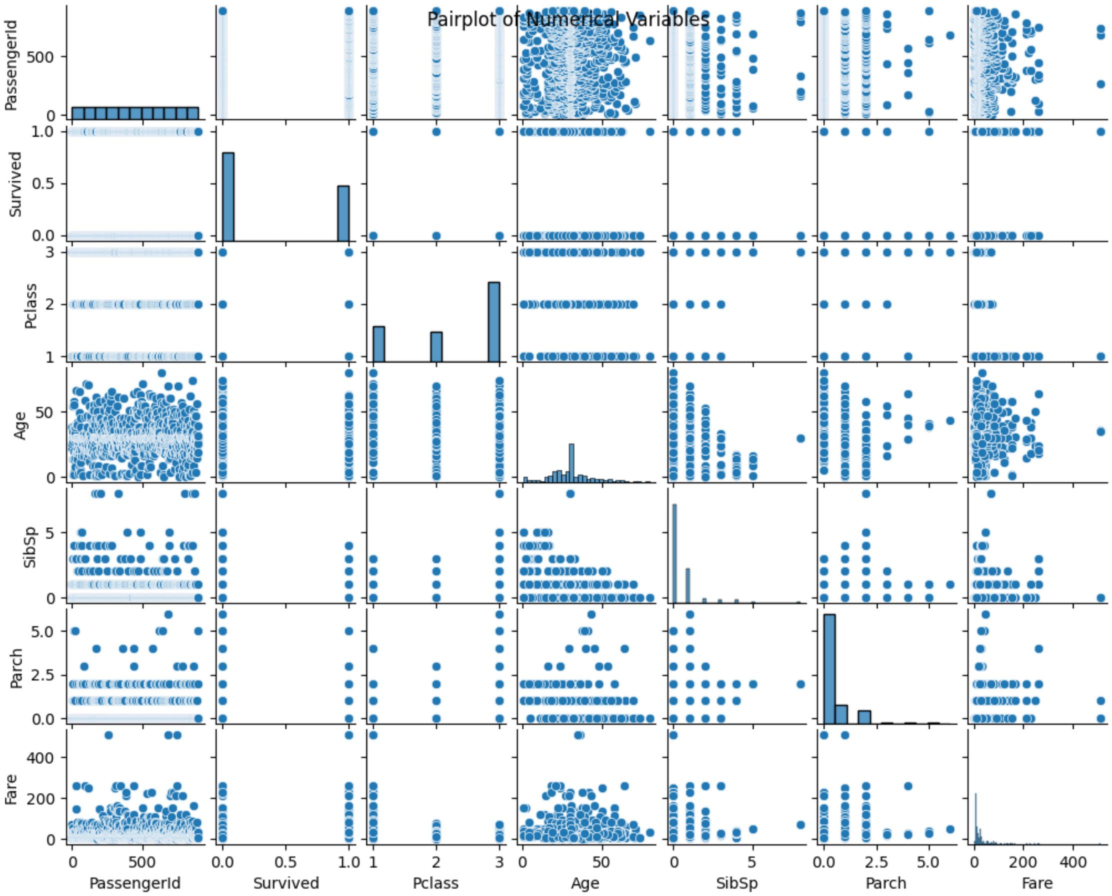
```
sns.histplot(df['Survived'], bins=20, kde=True)
plt.title('Distribution of Survived people')
plt.xlabel('Total People Survived')
plt.ylabel('Frequency')
plt.show()
```



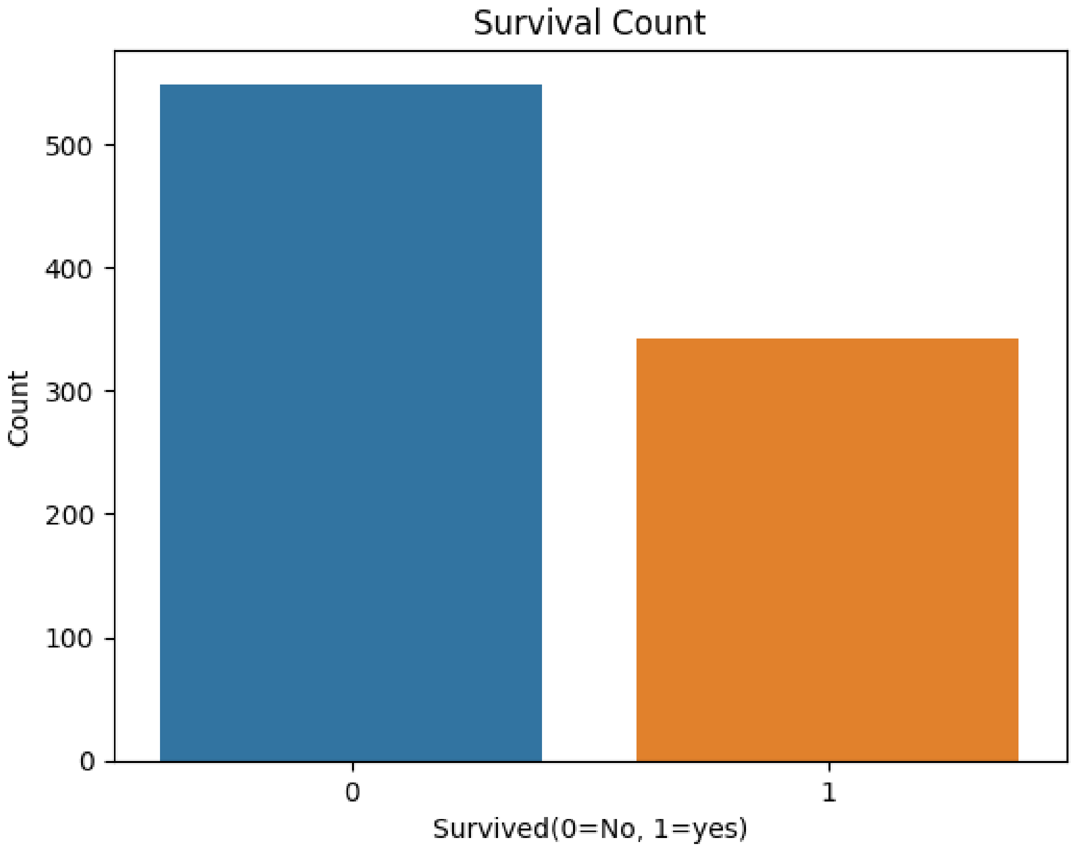
[21]:

```
g = sns.pairplot(df)
g.fig.set_size_inches(10,8)
plt.suptitle('Pairplot of Numerical Variables')
plt.show()
```

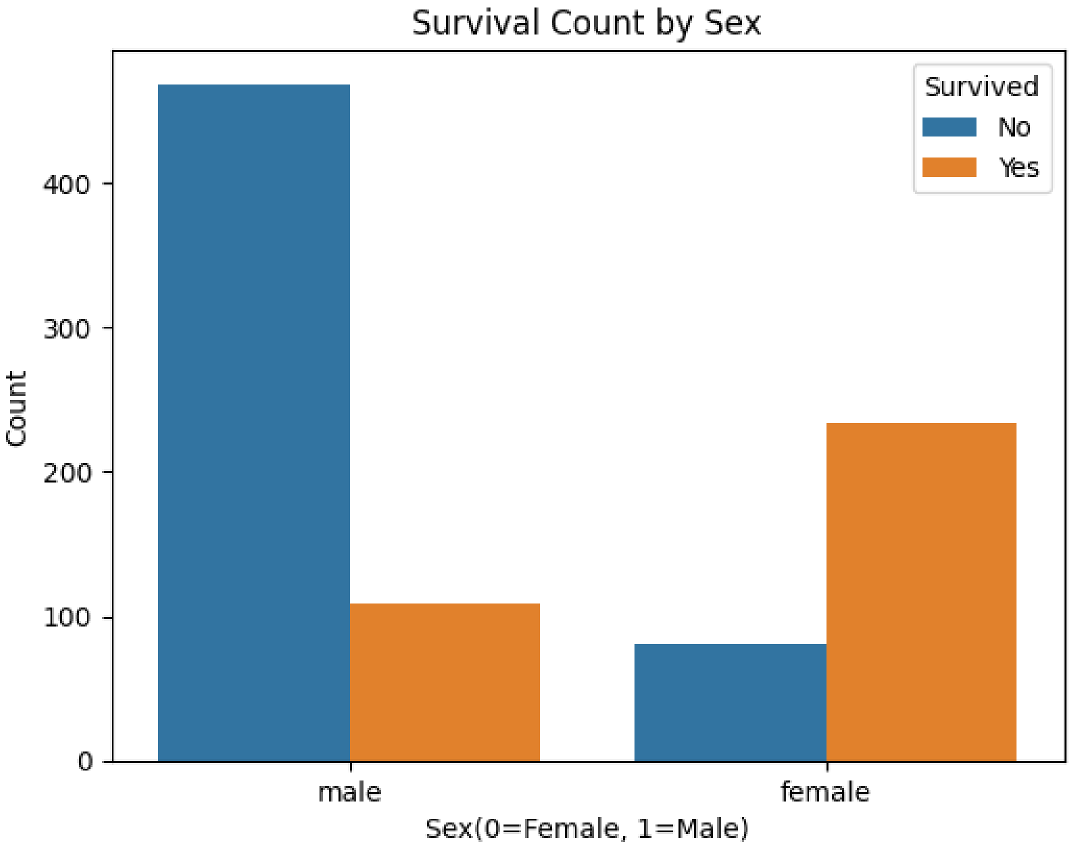
```
c:\Users\sivar\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\axisgrid.py:118: UserWarning: The figure layout has changed to
tight
  self._figure.tight_layout(*args, **kwargs)
```



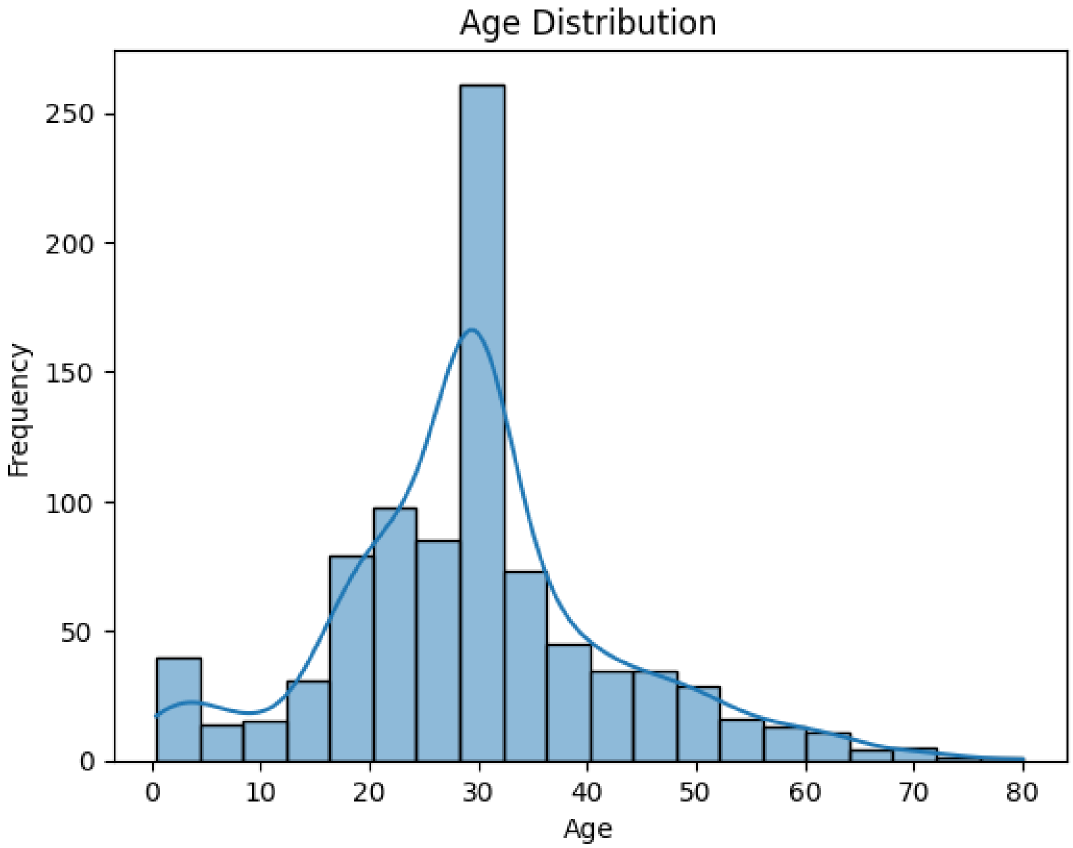
[22]:
```
sns.countplot(x='Survived', data=df)
plt.title('Survival Count')
plt.xlabel('Survived(0=No, 1=yes)')
plt.ylabel('Count')
plt.show()
```



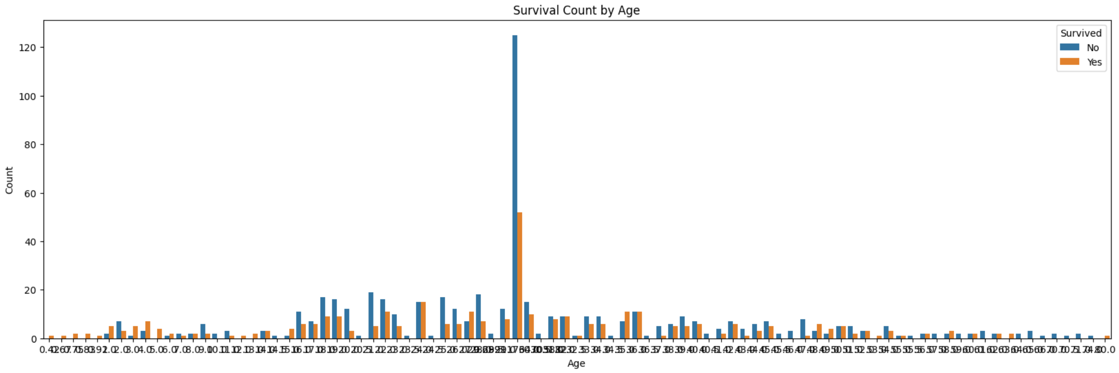
[23]:
```
sns.countplot(x='Sex', hue='Survived', data=df)
plt.title('Survival Count by Sex')
plt.xlabel('Sex(0=Female, 1=Male)')
plt.ylabel('Count')
plt.legend(title='Survived', labels=['No', 'Yes'])
plt.show()
```



```
[24]: sns.histplot(df['Age'], bins=20, kde=True)
      plt.title('Age Distribution')
      plt.xlabel('Age')
      plt.ylabel('Frequency')
      plt.show()
```



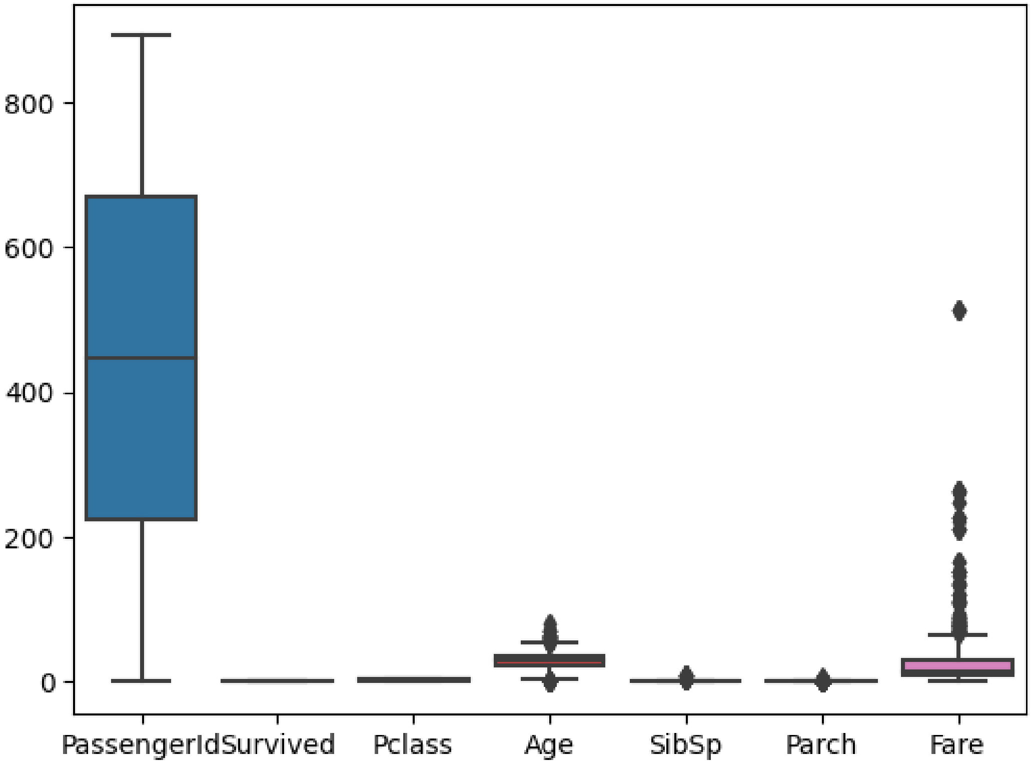
[25]:
```
plt.figure(figsize=(20,6))
sns.countplot(x='Age', hue='Survived', data=df)
plt.title('Survival Count by Age')
plt.xlabel('Age')
plt.ylabel('Count')
plt.legend(title='Survived', labels=['No', 'Yes'])
plt.show()
```



0.0.8 Outlier Detection

[26]:
```
sns.boxplot(df)
```

[26]:
```
<Axes: >
```



[27]:
```
q1 = df.Age.quantile(0.25)
q3 = df.Age.quantile(0.75)
print(q1)
print(q3)
```

```
22.0
35.0
```

[28]:
```
IQR = q3-q1
print(IQR)
```

```
13.0
```

[29]:
```
ul = q3+1.5*IQR
print(ul)
```

```
54.5
```

[30]:

```
l1 = q1-1.5*IQR
print(l1)
```

```
2.5
```

[31]:

```
numeric_columns = df.select_dtypes(include=['number']).columns
df[numeric_columns].median()
```

[31]:

```
PassengerId    446.000000
Survived         0.000000
Pclass           3.000000
Age             29.699118
SibSp            0.000000
Parch            0.000000
Fare            14.454200
dtype: float64
```

[32]:

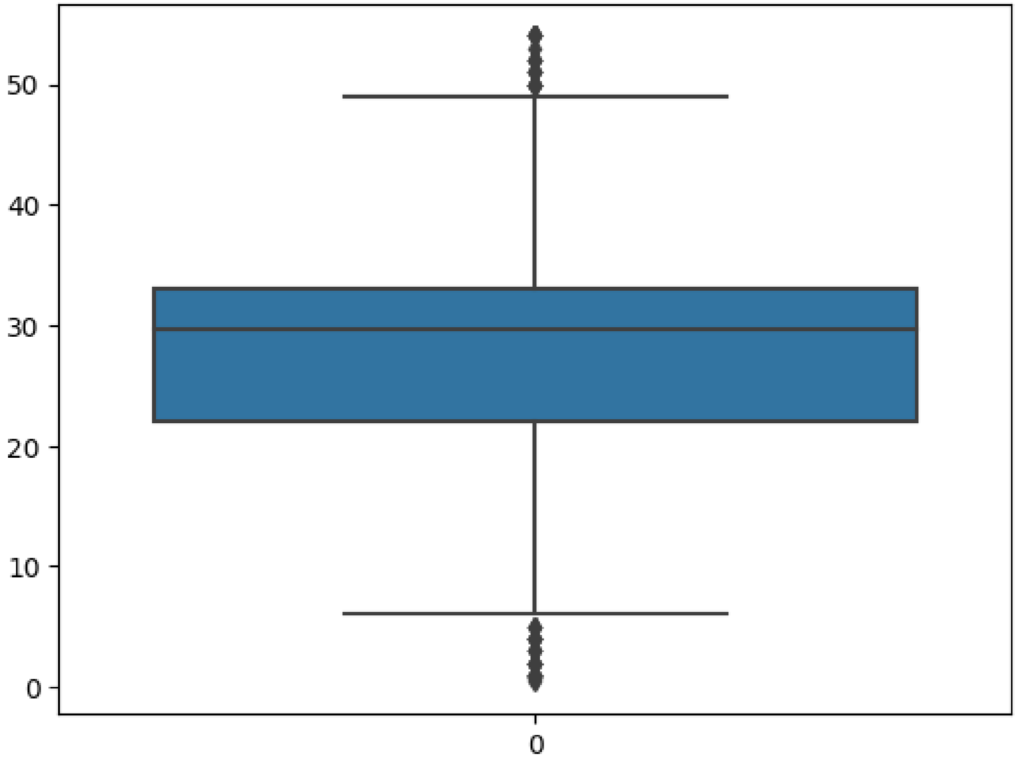
```
df = df[df.Age<ul]
```

[33]:

```
sns.boxplot(df.Age)
```

[33]:

```
<Axes: >
```



[34]:

```
q1 = df.SibSp.quantile(0.25)
q3 = df.SibSp.quantile(0.75)
print(q1)
print(q3)
```

```
0.0
1.0
```

[35]:

```
IQR = q3-q1
print(IQR)
```

```
1.0
```

[36]:

```
ul = q3+1.5*IQR
print(ul)
```

```
2.5
```

[37]:

```
l1 = q1-1.5*IQR
print(l1)
```

```
-1.5
```

[38]:

```
numeric_columns = df.select_dtypes(include=['number']).columns
df[numeric_columns].median()
```

[38]:

```
PassengerId    444.000000
Survived         0.000000
Pclass           3.000000
Age             29.699118
SibSp            0.000000
Parch            0.000000
Fare            14.108300
dtype: float64
```

[39]:

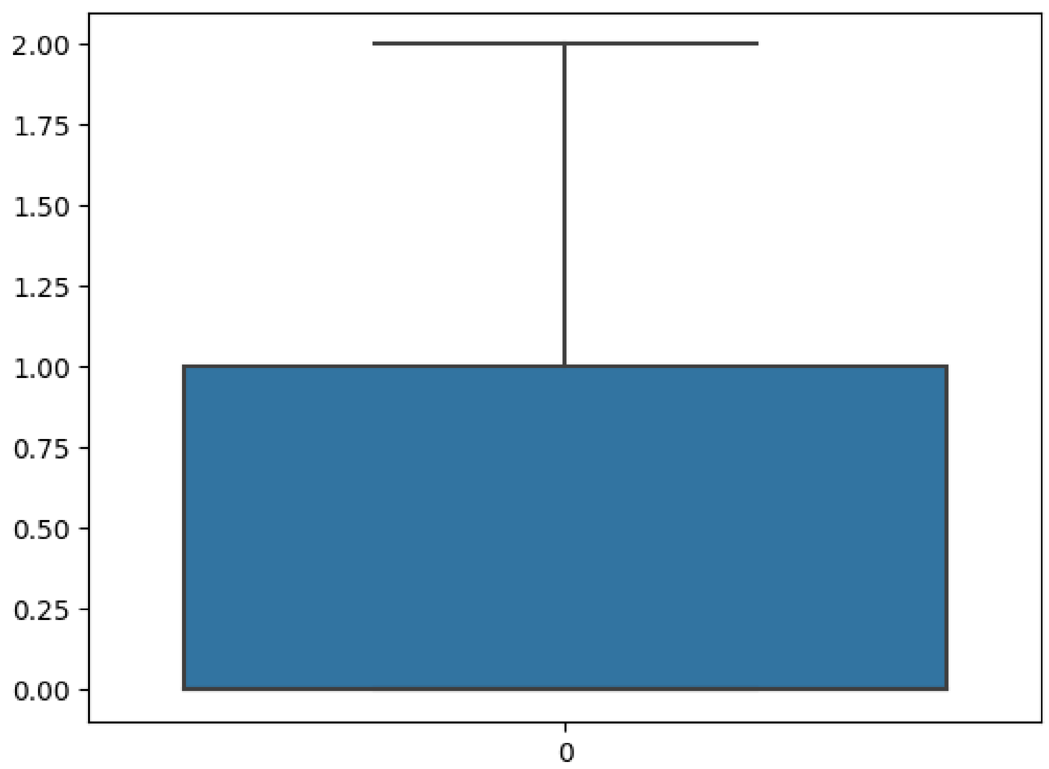
```
df = df[df.SibSp<ul]
```

[40]:

```
sns.boxplot(df.SibSp)
```

[40]:

```
<Axes: >
```



[41]:
```
q1 = df.Fare.quantile(0.25)
q3 = df.Fare.quantile(0.75)
print(q1)
print(q3)
```

```
7.8958
27.825
```

[42]:
```
IQR = q3-q1
print(IQR)
```

```
19.929199999999998
```

[43]:
```
ul = q3+1.5*IQR
print(ul)
```

```
57.7188
```

[44]:
```
l1 = q1-1.5*IQR
print(l1)
```

```
-21.997999999999998
```

[45]:

```
numeric_columns = df.select_dtypes(include=['number']).columns
df[numeric_columns].median()
```

[45]:

```
PassengerId    450.000000
Survived         0.000000
Pclass           3.000000
Age             29.699118
SibSp            0.000000
Parch            0.000000
Fare            13.000000
dtype: float64
```

[46]:

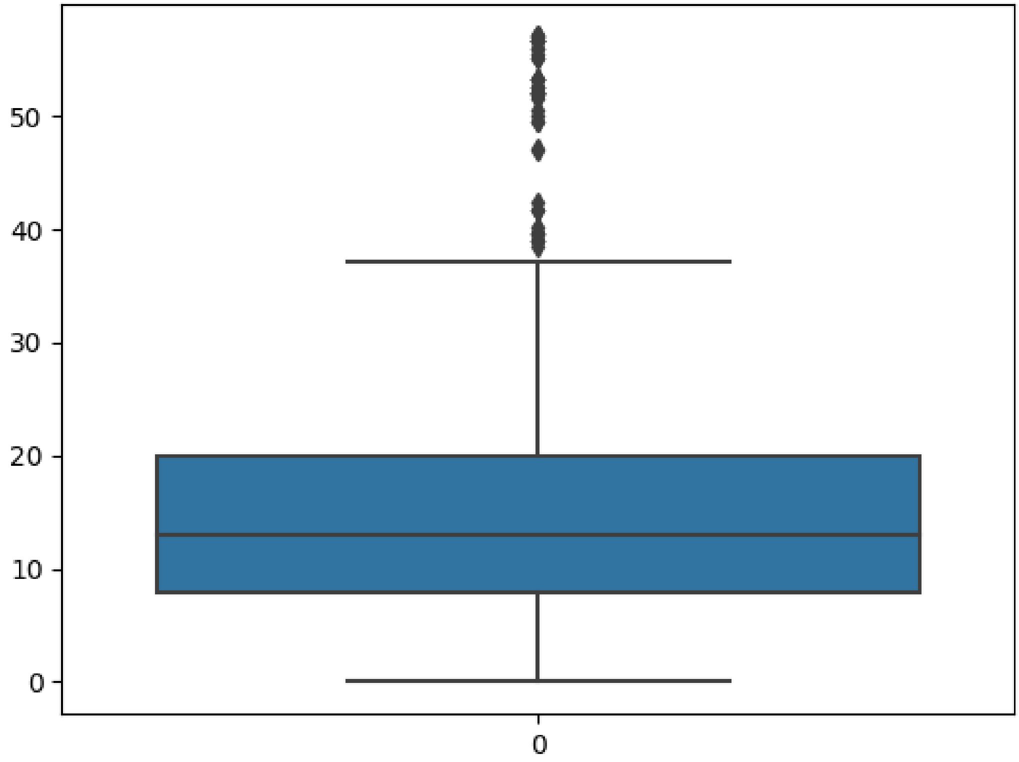
```
df['Fare'] = np.where(df['Fare']>ul,13,df['Fare'])
```

[47]:

```
sns.boxplot(df.Fare)
```

[47]:

```
<Axes: >
```



[48]:

```
q1 = df.Parch.quantile(0.25)
q3 = df.Parch.quantile(0.75)
print(q1)
print(q3)
```

```
0.0
0.0
```

[49]:

```
IQR = q3-q1
print(IQR)
```

```
0.0
```

[50]:

```
ul = q3+1.5*IQR
print(ul)
```

```
0.0
```

[51]:

```
l1 = q1-1.5*IQR
print(l1)
```

```
0.0
```

[52]:

```
numeric_columns = df.select_dtypes(include=['number']).columns
df[numeric_columns].median()
```

[52]:

```
PassengerId    450.000000
Survived         0.000000
Pclass           3.000000
Age             29.699118
SibSp            0.000000
Parch            0.000000
Fare            13.000000
dtype: float64
```

[53]:

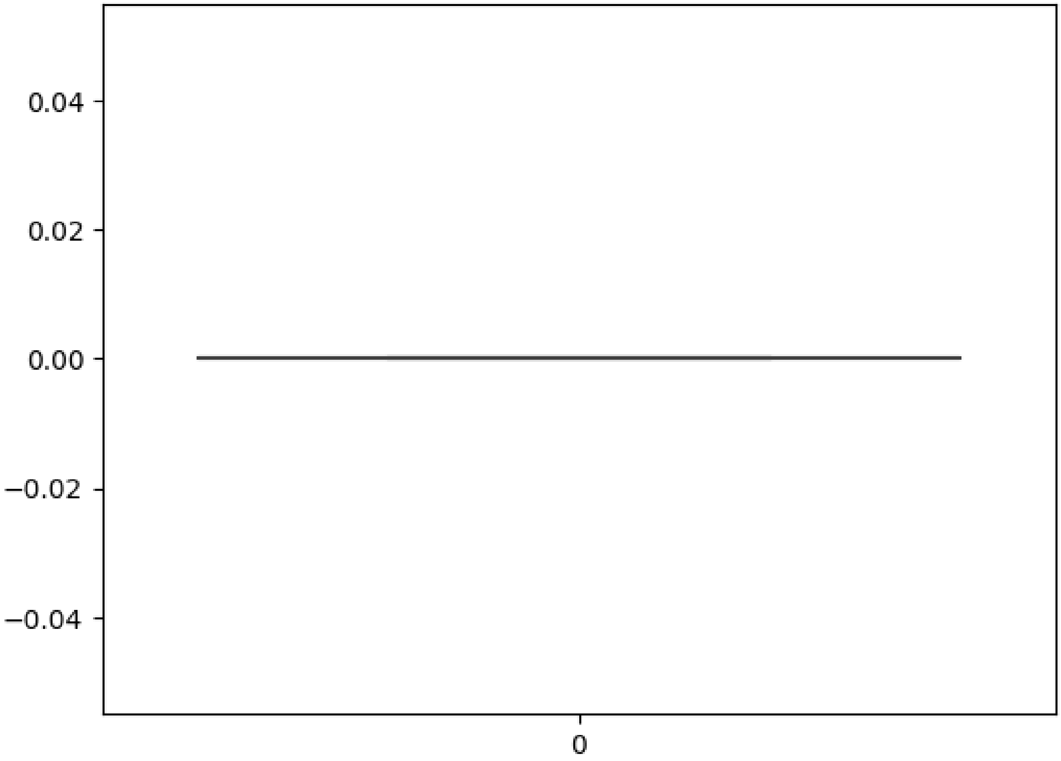
```
df['Parch'] = np.where(df['Parch']>ul,0,df['Parch'])
```

[54]:

```
sns.boxplot(df.Parch)
```

[54]:

```
<Axes: >
```



0.0.9 Splitting Dependent and independent Variables

[55]:
```
df.head()
```

[55]:
```
   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   7.250   NaN        S
1      0          PC 17599  13.000   C85        C
2      0  STON/O2. 3101282   7.925   NaN        S
```

```
3      0              113803  53.100  C123         S
4      0              373450   8.050   NaN         S
```

[56]:

```
df.Pclass.value_counts()
```

[56]:

```
Pclass
3    442
1    186
2    175
Name: count, dtype: int64
```

[57]:

```
x=df.
 ↪drop(columns=["PassengerId","Survived","Name","Parch","Ticket","Cabin"],axis=1)
x.head()
```

[57]:

```
   Pclass     Sex   Age  SibSp    Fare Embarked
0       3    male  22.0      1   7.250        S
1       1  female  38.0      1  13.000        C
2       3  female  26.0      0   7.925        S
3       1  female  35.0      1  53.100        S
4       3    male  35.0      0   8.050        S
```

[58]:

```
type(x)
```

[58]:

```
pandas.core.frame.DataFrame
```

[59]:

```
y=df["Survived"]
y.head()
```

[59]:

```
0    0
1    1
2    1
3    1
4    0
Name: Survived, dtype: int64
```

[60]:

```
type(y)
```

[60]:

```
pandas.core.series.Series
```

0.0.10 Encoding

[61]:

```
x.head()
```

[61]:

```
   Pclass     Sex   Age  SibSp    Fare Embarked
0       3    male  22.0      1   7.250        S
1       1  female  38.0      1  13.000        C
```

```
2        3  female  26.0      0   7.925        S
3        1  female  35.0      1  53.100        S
4        3    male  35.0      0   8.050        S
```

[62]:
```
from sklearn.preprocessing import LabelEncoder
le=LabelEncoder()
```

[63]:
```
x["Sex"]=le.fit_transform(x["Sex"])
x["Embarked"]=le.fit_transform(x["Embarked"])
```

[64]:
```
x.head()
```

[64]:
```
   Pclass  Sex   Age  SibSp    Fare  Embarked
0       3    1  22.0      1   7.250         2
1       1    0  38.0      1  13.000         0
2       3    0  26.0      0   7.925         2
3       1    0  35.0      1  53.100         2
4       3    1  35.0      0   8.050         2
```

[65]:
```
print(le.classes_)
```

```
['C' 'Q' 'S']
```

[66]:
```
mapping=dict(zip(le.classes_,range(len(le.classes_))))
mapping
```

[66]:
```
{'C': 0, 'Q': 1, 'S': 2}
```

0.0.11 Feature Scaling

[67]:
```
from sklearn.preprocessing import MinMaxScaler
ms=MinMaxScaler()
```

[68]:
```
X_Scaled=ms.fit_transform(x)
```

[69]:
```
X_Scaled=pd.DataFrame(ms.fit_transform(x),columns=x.columns)
```

[70]:
```
X_Scaled.head()
```

[70]:
```
   Pclass  Sex       Age  SibSp      Fare  Embarked
0     1.0  1.0  0.402762    0.5  0.127193       1.0
1     0.0  0.0  0.701381    0.5  0.228070       0.0
2     1.0  0.0  0.477417    0.0  0.139035       1.0
3     0.0  0.0  0.645390    0.5  0.931579       1.0
4     1.0  1.0  0.645390    0.0  0.141228       1.0
```

0.0.12 Train test split

[71]:

```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test = train_test_split(X_Scaled,y,test_size =0.2,random_state =0)
```

[72]:

```
print(x_train.shape,x_test.shape,y_train.shape,y_test.shape)
```

```
(642, 6) (161, 6) (642,) (161,)
```

[ ]: