EXPERIMENT -3

AIM

To perform EDA on the given data set.

Explanation

The primary aim with exploratory analysis is to examine the data for distribution, outliers and anomalies to direct specific testing of your hypothesis.

ALGORITHM

STEP 1:

Import the required packages(pandas,numpy,seaborn).

STEP 2:

Read the given csv file.

STEP 3:

Convert the file into a dataframe and get information of the data.

STEP 4:

Remove the non numerical data columns using drop() method.

STEP 5:

Replace the null values using (.fillna).

STEP 6:

returns object containing counts of unique values using (value_counts()).

STEP 7:

Plot the counts in the form of Histogram or Bar Graph.

STEP 8:

find the pairwise correlation of all columns in the dataframe(.corr()).

STEP 9:

Save the final data set into the file.

```
import pandas as pd
import numpy as np
import seaborn as sns
```

```
df=pd.read_csv("titanic_dataset.csv")
```

```
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
```

```
df.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th	female	38.0	1	0	PC 17599	71.

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

```
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
```

```
df.drop("Cabin",axis=1,inplace=True)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 11 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  Embarked     889 non-null    object 
dtypes: float64(2), int64(5), object(4)
memory usage: 76.7+ KB
```

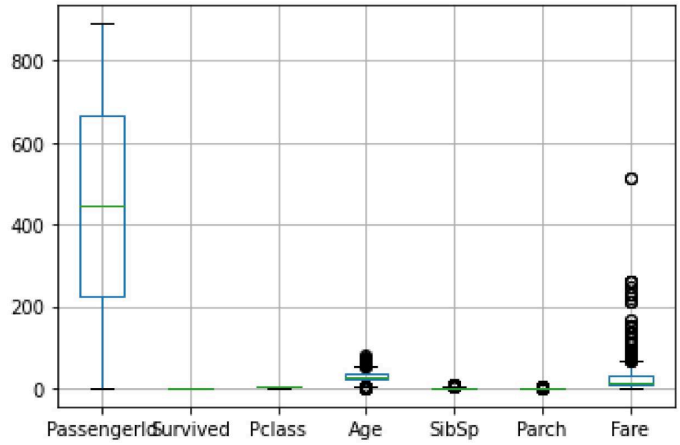
```
df.isnull().sum()
```

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Embarked         2
dtype: int64
```

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

```
df.boxplot()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f597af952d0>
```




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

```
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
Fare           0
Embarked       2
dtype: int64
```

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

```
df["Embarked"].value_counts()
```

```
S    646
C    168
Q     77
Name: Embarked, dtype: int64
```

```
df["Pclass"].value_counts()
```

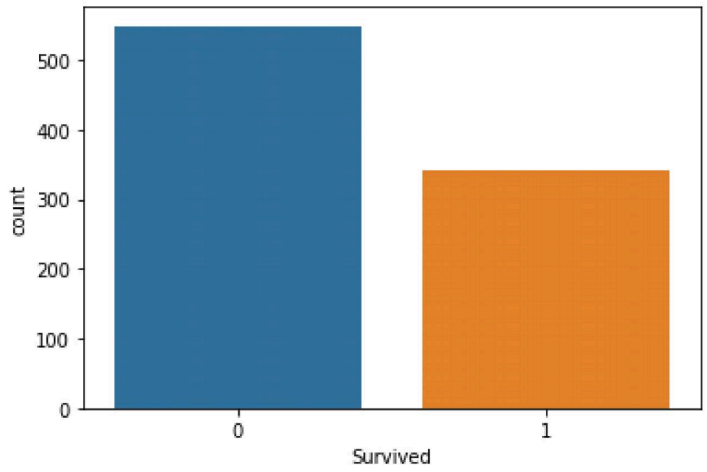
```
3    491
1    216
2    184
Name: Pclass, dtype: int64
```

```
df["Survived"].value_counts()
```

```
0    549
1    342
Name: Survived, dtype: int64
```

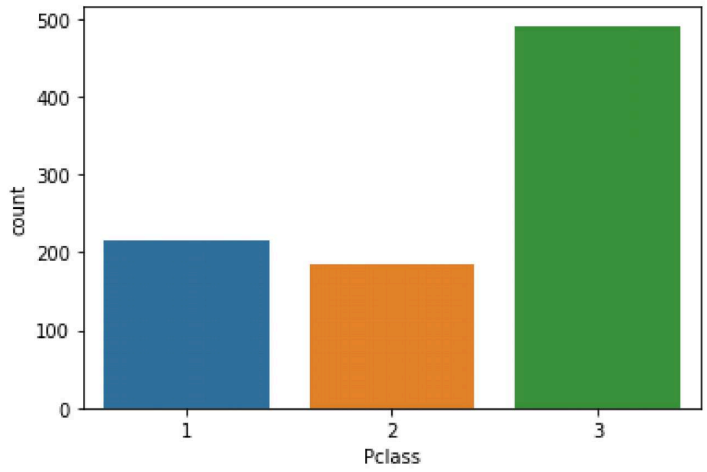
```
sns.countplot(x="Survived",data=df)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f597a97c650>
```



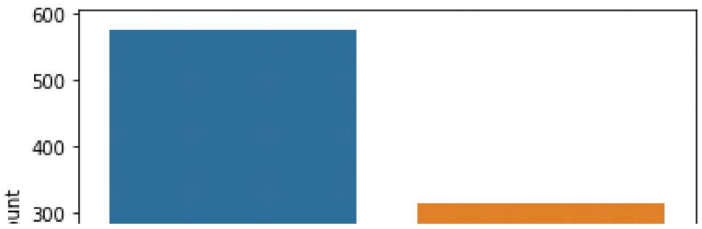

```
sns.countplot(x="Pclass",data=df)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f597a940610>
```




```
sns.countplot(x="Sex",data=df)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f597afd6710>
```

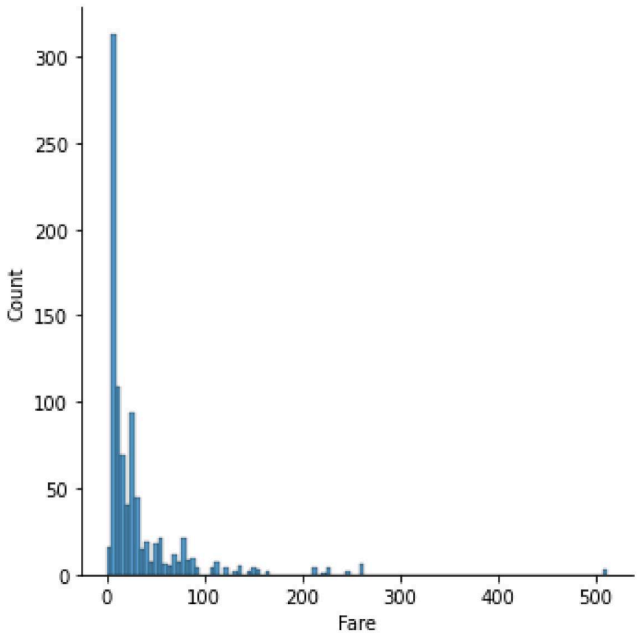



```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 11 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          891 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  Embarked     891 non-null    object 
dtypes: float64(2), int64(5), object(4)
memory usage: 76.7+ KB
```

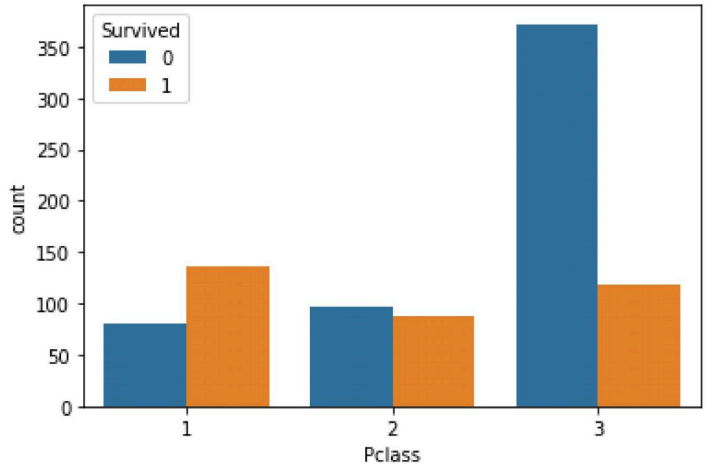
```
sns.displot(df["Fare"])
```

```
<seaborn.axisgrid.FacetGrid at 0x7f597ae074d0>
```



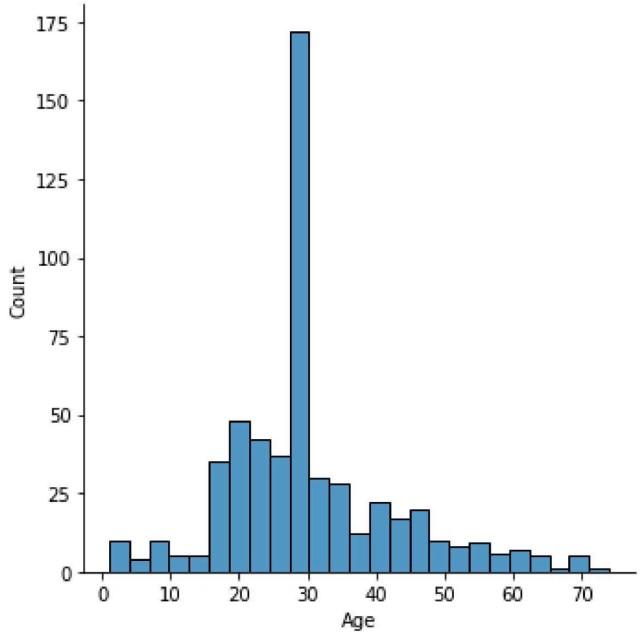
```
sns.countplot(x="Pclass",hue="Survived",data=df)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f5977e65c10>
```




```
sns.displot(df[df["Survived"]==0]["Age"])
```

```
<seaborn.axisgrid.FacetGrid at 0x7f5977e31e50>
```




```
pd.crosstab(df["Pclass"],df["Survived"])
```

Survived	0	1
Pclass		
1	80	136
2	97	87
3	372	119

```
pd.crosstab(df["Sex"],df["Survived"])
```

Survived	0	1
Sex		
female	81	233
male	468	109

```
df.corr()
```

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Far
PassengerId	1.000000	-0.005007	-0.035144	0.034212	-0.057527	-0.001652	0.01265
Survived	-0.005007	1.000000	-0.338481	-0.064910	-0.035322	0.081629	0.25730
Pclass	-0.035144	-0.338481	1.000000	-0.339898	0.083081	0.018443	-0.54950
Age	0.034212	-0.064910	-0.339898	1.000000	-0.233296	-0.172482	0.09668
SibSp	-0.057527	-0.035322	0.083081	-0.233296	1.000000	0.414838	0.15965
Parch	-0.001652	0.081629	0.018443	-0.172482	0.414838	1.000000	0.21622
Fare	0.012658	0.257307	-0.549500	0.096688	0.159651	0.216225	1.00000

```
sns.heatmap(df.corr(),annot=True)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f5977e61d10>
```