Logistic Regression

Titanic dataset

Data Loaded

In [190…
```
import pandas as pd
```

In [191…
```
df=pd.read_csv('C:/Users/vaitheeswaran/OneDrive/Documents/titanic1.csv')
```

In [192…
```
df.head()
```

Out[192…

	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

Preprocessing the data

In [193…
```
#Preprocessing
df.nunique()
```

Out[193…
```
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
```

In [194…
```
df.isnull().sum()
```

Out[194…
```
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
```

In [195…
```
df['Age'].mean()
```

Out[195…
```
29.69911764705882
```

In [196…
```
df['Age'].fillna(df['Age'].mean(),inplace=True)
```

In [197…
```
df
```

Out[197…

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.000000	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.000000	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.000000	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.000000	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.000000	0	0	373450	8.0500	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.000000	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.000000	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	29.699118	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.000000	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.000000	0	0	370376	7.7500	NaN	Q

891 rows × 12 columns

In [198…
```
df.isnull().sum()
```

Out[198…
```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age              0
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

In [199…
```
df['Embarked'].unique()
```

Out[199…
```
array(['S', 'C', 'Q', nan], dtype=object)
```

In [200…
```
df.tail()
```

Out[200…

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
886	887	0	2	Montvila, Rev. Juozas	male	27.000000	0	0	211536	13.00	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.000000	0	0	112053	30.00	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	29.699118	1	2	W./C. 6607	23.45	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.000000	0	0	111369	30.00	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.000000	0	0	370376	7.75	NaN	Q

In [201…
```
df['Sex'].replace({
    0:'male',
    1:'female'
},inplace=True)
```

In [202…
```
df.head()
```

Out[202…

	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 [203…
```
df.groupby('Embarked')['Embarked'].count()
```

Out[203…
```
Embarked
C    168
Q     77
S    644
Name: Embarked, dtype: int64
```

In [204…
```
df['Embarked'].fillna(644,inplace=True)
```

In [205…
```
df.head()
```

Out[205…

	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 [206…
```
df.drop(columns=['Cabin'],axis=1,inplace=True)
```

In [207…
```
df.head()
```

Out[207…

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

In [208…
```
df.isnull().sum()
```

Out[208…
```
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
Fare           0
Embarked       0
dtype: int64
```

In [209…
```
df.tail()
```

Out[209…

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
886	887	0	2	Montvila, Rev. Juozas	male	27.000000	0	0	211536	13.00	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.000000	0	0	112053	30.00	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	29.699118	1	2	W./C. 6607	23.45	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.000000	0	0	111369	30.00	C
890	891	0	3	Dooley, Mr. Patrick	male	32.000000	0	0	370376	7.75	Q

In [210…
```
import seaborn as sns
```

In [211…
```
df.shape
```

Out[211… (891, 11)

In [212…
```
df.describe()
```

Out[212…

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	891.000000	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	13.002015	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	22.000000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	29.699118	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	35.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

In [213…
```
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
```

In [214…
```
df.isnull().sum()
```

Out[214…
```
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
```

```
Fare           0
Embarked       0
dtype: int64
```

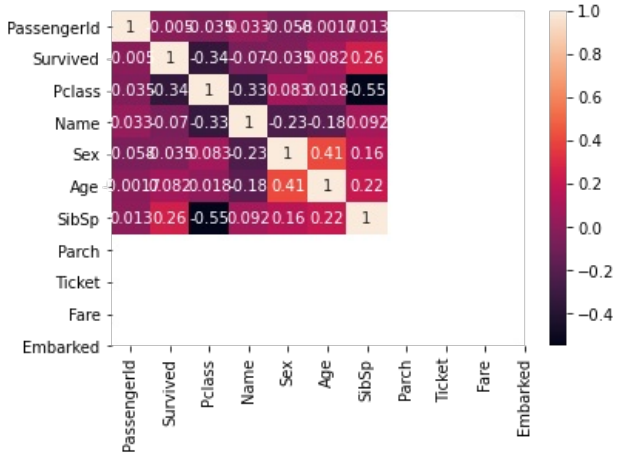
In [215…
```
df.corr()
```

Out[215…

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

In [216…
```
sns.heatmap(df.corr(),xticklabels=df.columns,yticklabels=df.columns,annot=True)
```

Out[216… <AxesSubplot:>



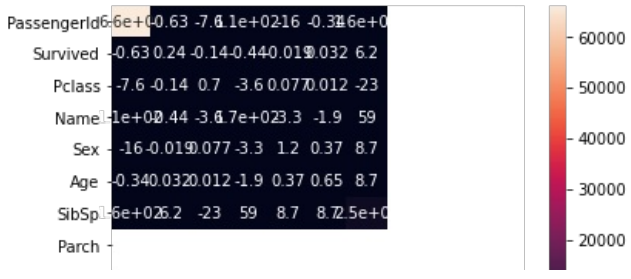
In [217…
```
df.cov()
```

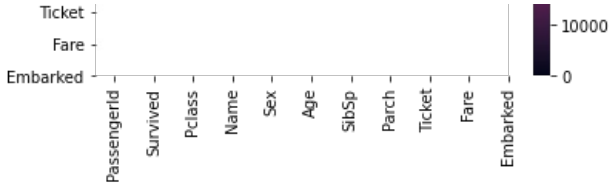
Out[217…

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
PassengerId	66231.000000	-0.626966	-7.561798	111.113042	-16.325843	-0.342697	161.883369
Survived	-0.626966	0.236772	-0.137703	-0.441656	-0.018954	0.032017	6.221787
Pclass	-7.561798	-0.137703	0.699015	-3.601855	0.076599	0.012429	-22.830196
Age	111.113042	-0.441656	-3.601855	169.052400	-3.335345	-1.877987	59.162200
SibSp	-16.325843	-0.018954	0.076599	-3.335345	1.216043	0.368739	8.748734
Parch	-0.342697	0.032017	0.012429	-1.877987	0.368739	0.649728	8.661052
Fare	161.883369	6.221787	-22.830196	59.162200	8.748734	8.661052	2469.436846

In [218…
```
sns.heatmap(df.cov(),xticklabels=df.columns,yticklabels=df.columns,annot=True)
```

Out[218… <AxesSubplot:>

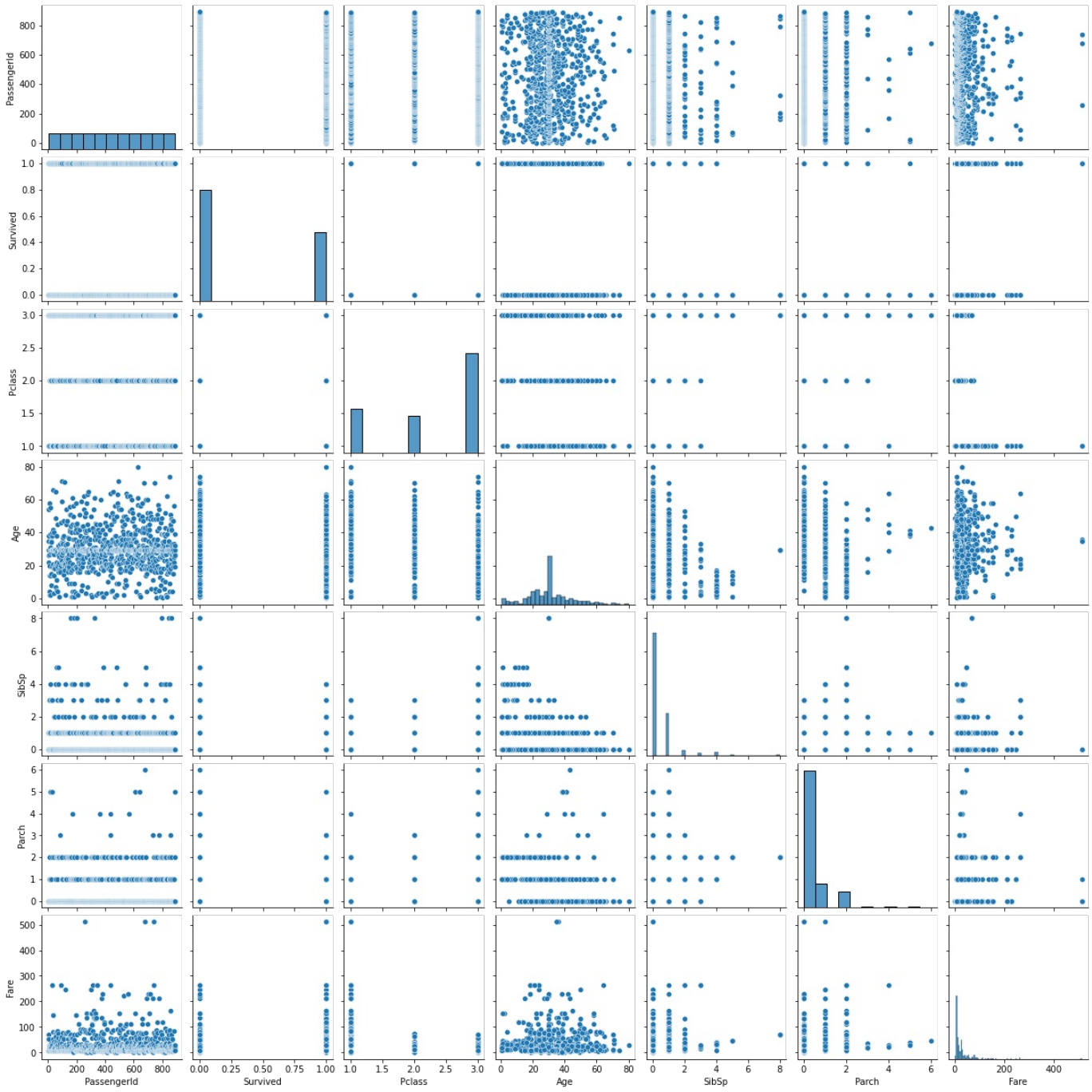




Exploratory Data Anlaysis

In [219…
```
sns.pairplot(df)
```

Out[219…
```
<seaborn.axisgrid.PairGrid at 0x1d7ed45f460>
```

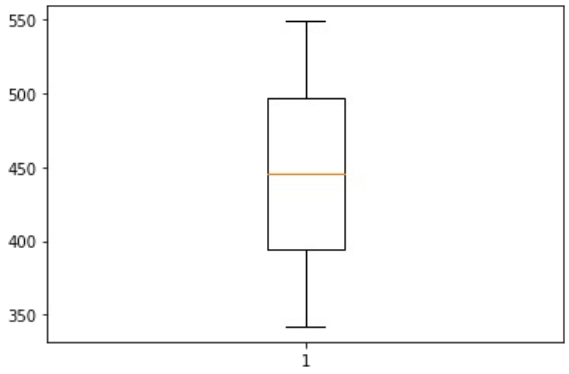


In [220…
```
df1=df.groupby('Survived')['Survived'].count()
```

In [221…
```
import matplotlib.pyplot as plt
```

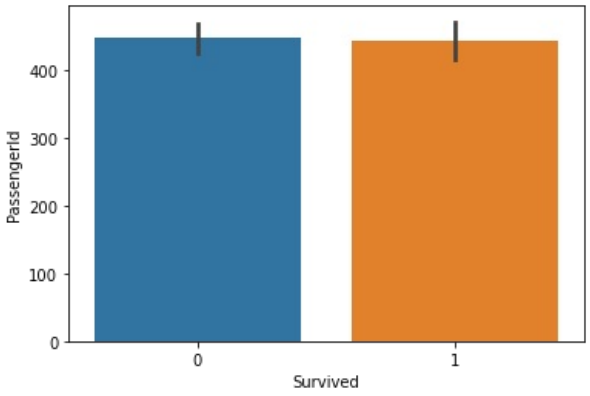
In [222…
```
plt.boxplot(df1)
```

Out[222…
```
{'whiskers': [<matplotlib.lines.Line2D at 0x1d7efddc700>,
  <matplotlib.lines.Line2D at 0x1d7efddca60>],
 'caps': [<matplotlib.lines.Line2D at 0x1d7efddcdc0>,
  <matplotlib.lines.Line2D at 0x1d7f03d5160>],
 'boxes': [<matplotlib.lines.Line2D at 0x1d7efddc400>],
 'medians': [<matplotlib.lines.Line2D at 0x1d7f03d54c0>],
 'fliers': [<matplotlib.lines.Line2D at 0x1d7f03d5820>],
 'means': []}
```



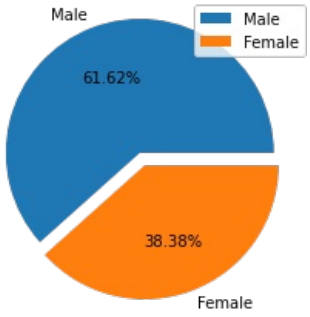
In [223…
```
sns.barplot(x=df['Survived'],y=df['PassengerId'])
```

Out[223…
```
<AxesSubplot:xlabel='Survived', ylabel='PassengerId'>
```



In [224…
```
plt.pie(df1,labels=['Male','Female'],explode=[0.1,0],autopct='%1.2f%%')
plt.legend()
```

Out[224…
```
<matplotlib.legend.Legend at 0x1d7f0421640>
```



In [225…
```
df1
```

Out[225…
```
Survived
0    549
1    342
Name: Survived, dtype: int64
```

In [226…
```
df2=df.groupby('Survived')['Sex'].count()
```

In [227…
```
df2
```

Out[227…
```
Survived
0    549
1    342
Name: Sex, dtype: int64
```

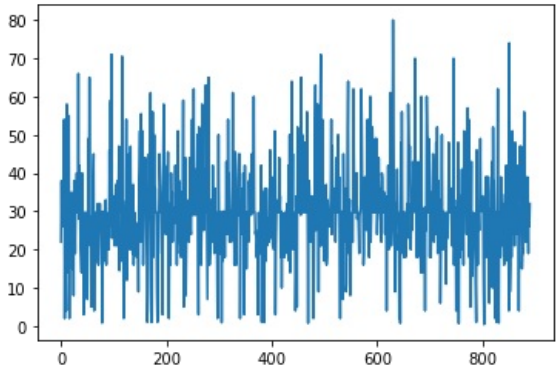
In [228…
```
df.tail()
```

Out[228…

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
886	887	0	2	Montvila, Rev. Juozas	male	27.000000	0	0	211536	13.00	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.000000	0	0	112053	30.00	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	29.699118	1	2	W./C. 6607	23.45	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.000000	0	0	111369	30.00	C
890	891	0	3	Dooley, Mr. Patrick	male	32.000000	0	0	370376	7.75	Q

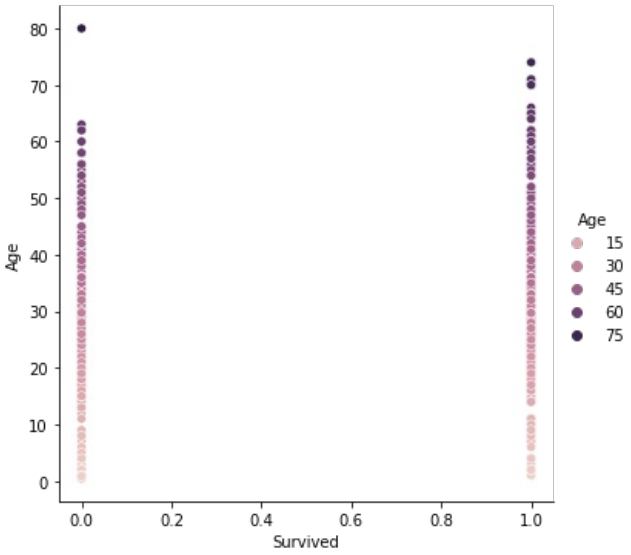
In [229…
```
plt.plot(df['Age'])
```

Out[229…
```
[<matplotlib.lines.Line2D at 0x1d7f04aa910>]
```



In [230…
```
sns.relplot(x=(df['Survived']==0),y='Age',hue='Age',data=df)
```

Out[230…
```
<seaborn.axisgrid.FacetGrid at 0x1d7f04cb6d0>
```

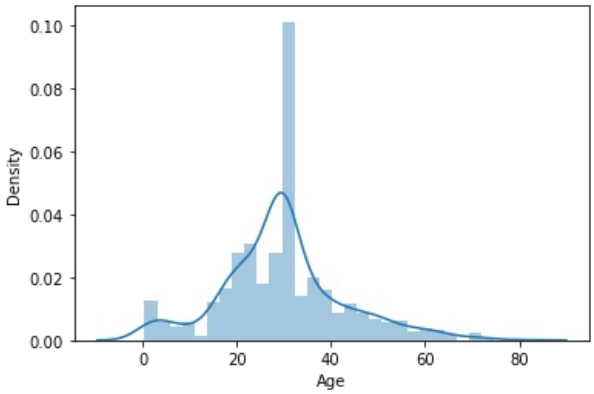


In [231…
```
df3=df.groupby('Pclass')[['Survived','Sex']].count()
```

In [232…
```
sns.distplot(df['Age'])
```

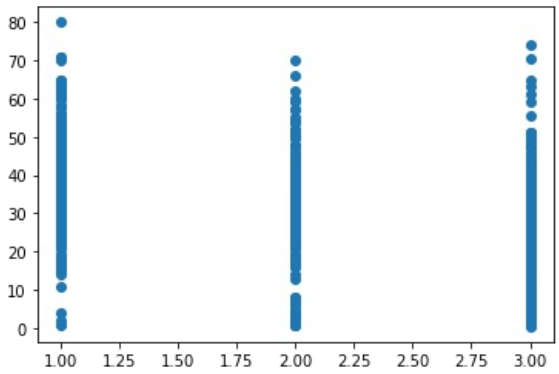
```
C:\Users\vaitheeswaran\anaconda3\lib\site-packages\seaborn\distributions.py:2557: FutureWarning: `distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

Out[232…
```
<AxesSubplot:xlabel='Age', ylabel='Density'>
```



In [233…
```
plt.scatter(df['Pclass'],df['Age'])
```

Out[233…
```
<matplotlib.collections.PathCollection at 0x1d7f15f8a30>
```



In [234…
```
dd=df.groupby('Age')[['Survived','Sex']].count()
```

In [235…
```
dd
```

Out[235…

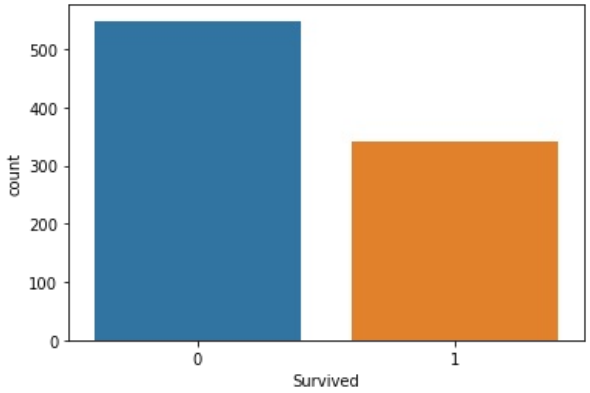
Age	Survived	Sex
0.42	1	1
0.67	1	1
0.75	2	2
0.83	2	2
0.92	1	1
...	...	...
70.00	2	2
70.50	1	1
71.00	2	2
74.00	1	1
80.00	1	1

89 rows × 2 columns

In [236…
```
sns.countplot(x='Survived',data=df)
```

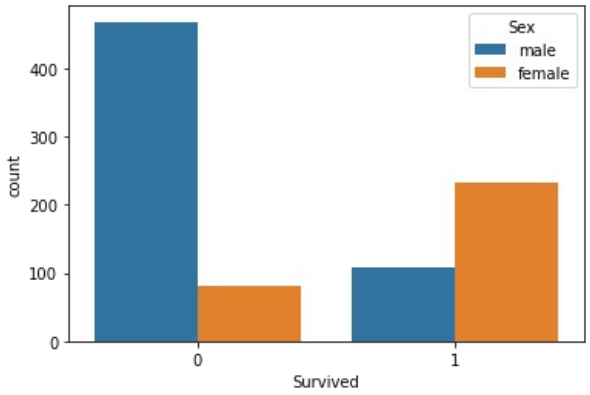
```
sns.countplot(x='Survived',data=df)
```

Out[236… <AxesSubplot:xlabel='Survived', ylabel='count'>



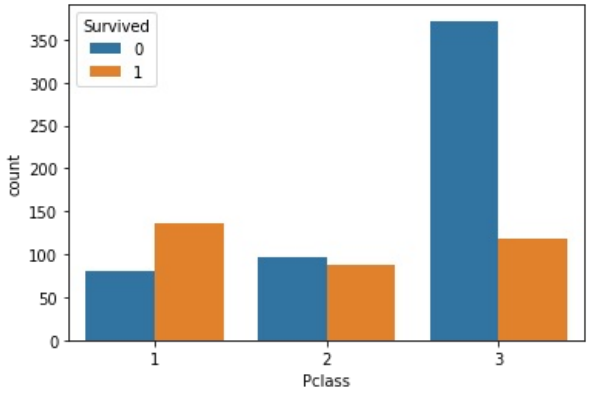
In [237…
```
sns.countplot(x='Survived',hue='Sex',data=df)
```

Out[237… <AxesSubplot:xlabel='Survived', ylabel='count'>



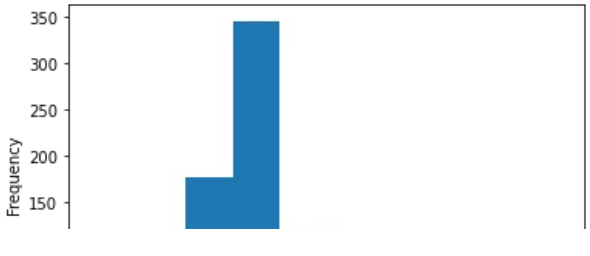
In [238…
```
sns.countplot(x='Pclass',hue='Survived',data=df)
```

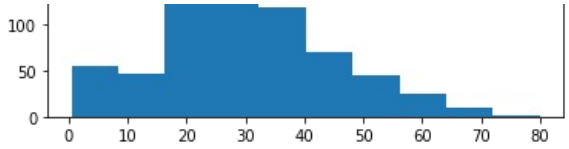
Out[238… <AxesSubplot:xlabel='Pclass', ylabel='count'>



In [239…
```
df['Age'].plot.hist()
```

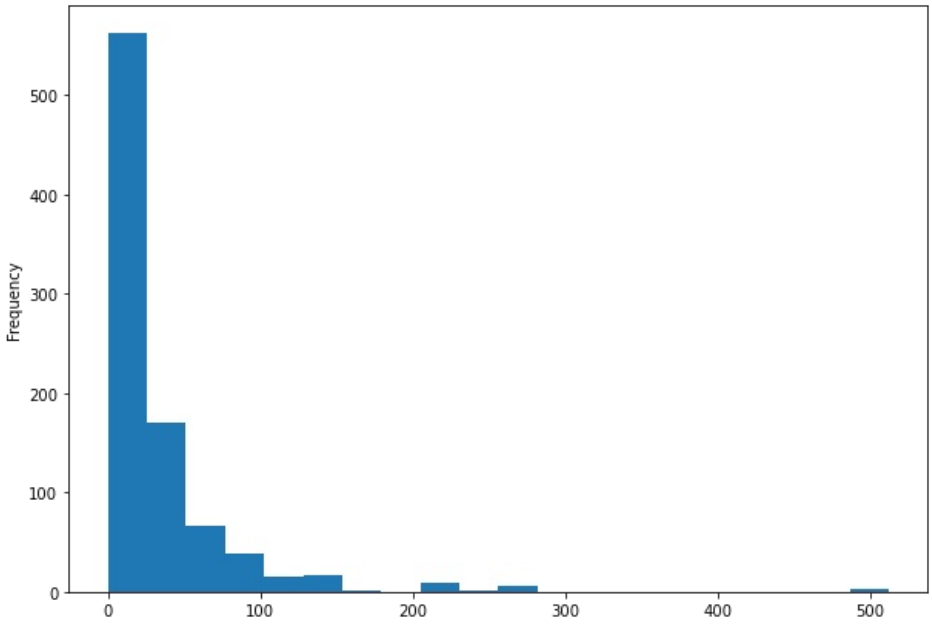
Out[239… <AxesSubplot:ylabel='Frequency'>





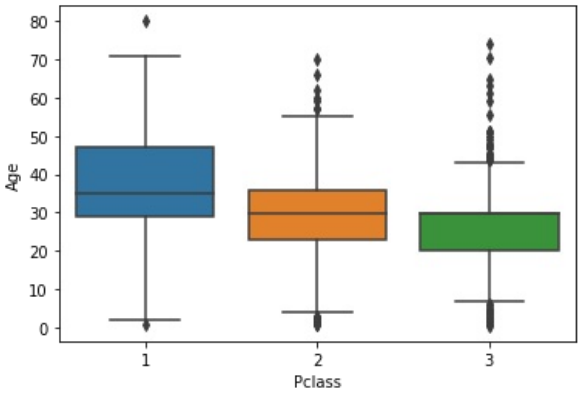
In [240…
```
df['Fare'].plot.hist(bins=20,figsize=(10,7))
```

Out[240…
```
<AxesSubplot:ylabel='Frequency'>
```



In [241…
```
sns.boxplot(x='Pclass',y='Age',data=df)
```

Out[241…
```
<AxesSubplot:xlabel='Pclass', ylabel='Age'>
```



In [242…
```
sex=pd.get_dummies(df['Sex'],drop_first=True)
sex
```

Out[242…

	male
0	1
1	0
2	0
3	0
4	1
...	...
886	1
887	0

888	0
889	1
890	1

891 rows × 1 columns

In [243…
```
embark=pd.get_dummies(df['Embarked'])
```

In [244…
```
Pcl=pd.get_dummies(df['Pclass'])
```

In [245…
```
df.head()
```

Out[245…

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

In [246…
```
df['Sex'].replace({
    'male':0,
    'female':1
},inplace=True)
```

In [247…
```
df
```

Out[247…

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
0	1	0	3	Braund, Mr. Owen Harris	0	22.000000	1	0	A/5 21171	7.2500	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	1	38.000000	1	0	PC 17599	71.2833	C
2	3	1	3	Heikkinen, Miss. Laina	1	26.000000	0	0	STON/O2. 3101282	7.9250	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	1	35.000000	1	0	113803	53.1000	S
4	5	0	3	Allen, Mr. William Henry	0	35.000000	0	0	373450	8.0500	S
...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	0	27.000000	0	0	211536	13.0000	S
887	888	1	1	Graham, Miss. Margaret Edith	1	19.000000	0	0	112053	30.0000	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	1	29.699118	1	2	W./C. 6607	23.4500	S
889	890	1	1	Behr, Mr. Karl Howell	0	26.000000	0	0	111369	30.0000	C
890	891	0	3	Dooley, Mr. Patrick	0	32.000000	0	0	370376	7.7500	Q

891 rows × 11 columns

In [248…
```
df.head()
```

Out[248…

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

In [249…
```
df.drop(['Pclass'],axis=1,inplace=True)
```

Training and Testing data

In [250…
```
## Train And testing the data
x=df[['Sex','Age','SibSp','Parch','Fare']]
y=df['Survived']
```

In [251…
```
from sklearn.model_selection import train_test_split
```

In [252…
```
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.3,random_state=1)
```

In [253…
```
from sklearn.linear_model import LogisticRegression
```

In [254…
```
lr=LogisticRegression()
```

In [255…
```
lr.fit(x_train,y_train)
```

Out[255… LogisticRegression()

In [256…
```
ypred=lr.predict(x_test)
```

Evaluation

In [257…
```
from sklearn.metrics import classification_report
```

In [258…
```
classification_report(y_test,ypred)
```

Out[258…
```
'              precision    recall  f1-score   support\n\n           0       0.75      0.86      0.80       153\n
1       0.77      0.62      0.69       115\n\n    accuracy                           0.76       268\n   macro avg
0.76      0.74      0.74       268\nweighted avg       0.76      0.76      0.75       268\n'
```

In [259…
```
from sklearn.metrics import confusion_matrix,accuracy_score
```

In [260…
```
print("Confusion Matrix:",confusion_matrix(y_test,ypred))
```

```
Confusion Matrix: [[132  21]
 [ 44  71]]
```

In [261…
```
print("Accuracy:",accuracy_score(y_test,ypred))
```

```
Accuracy: 0.7574626865671642
```

In [ ]: