Assignment 3 15th Sept

September 21, 2023

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

[2]:
```
df = pd.read_csv('Titanic-Dataset.csv')
```

[3]:
```
df
```

[3]:
```
     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]
```

[4]:
```
df.isnull().any()
```

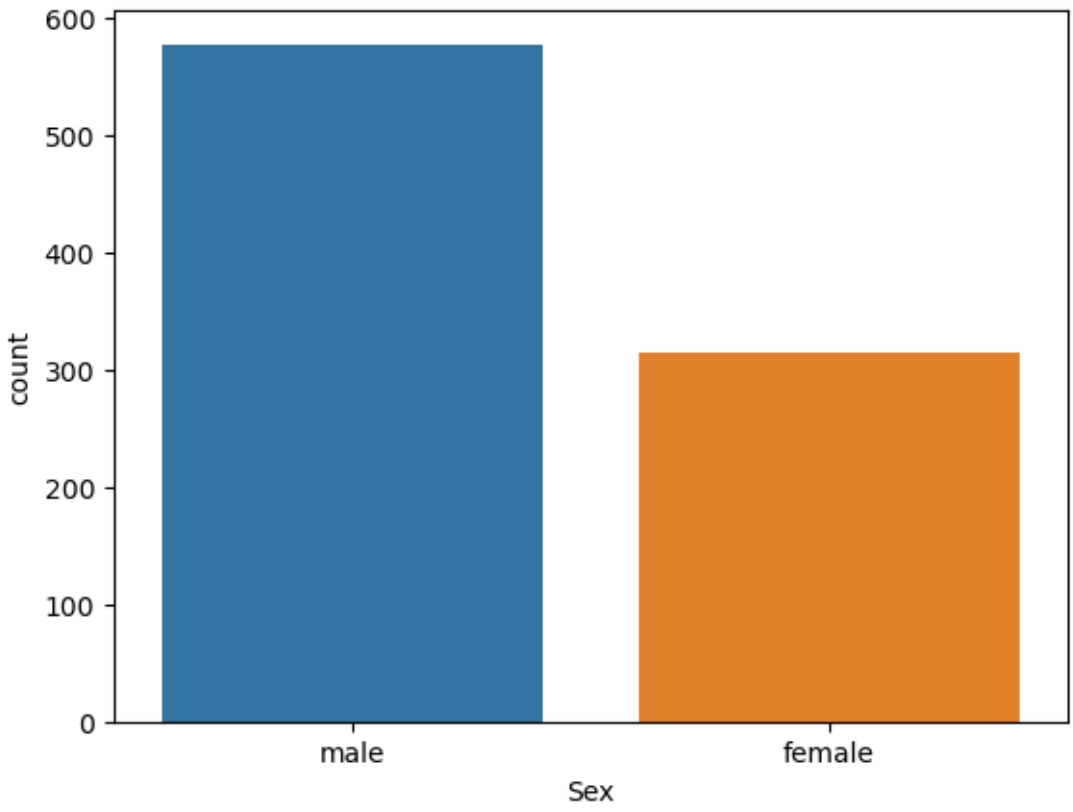
[4]:
```
PassengerId    False
Survived       False
Pclass         False
Name           False
Sex            False
Age             True
SibSp          False
Parch          False
Ticket         False
Fare           False
Cabin           True
Embarked        True
dtype: bool
```

[5]:
```
df.isnull().sum()
```

[5]:
```
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
```

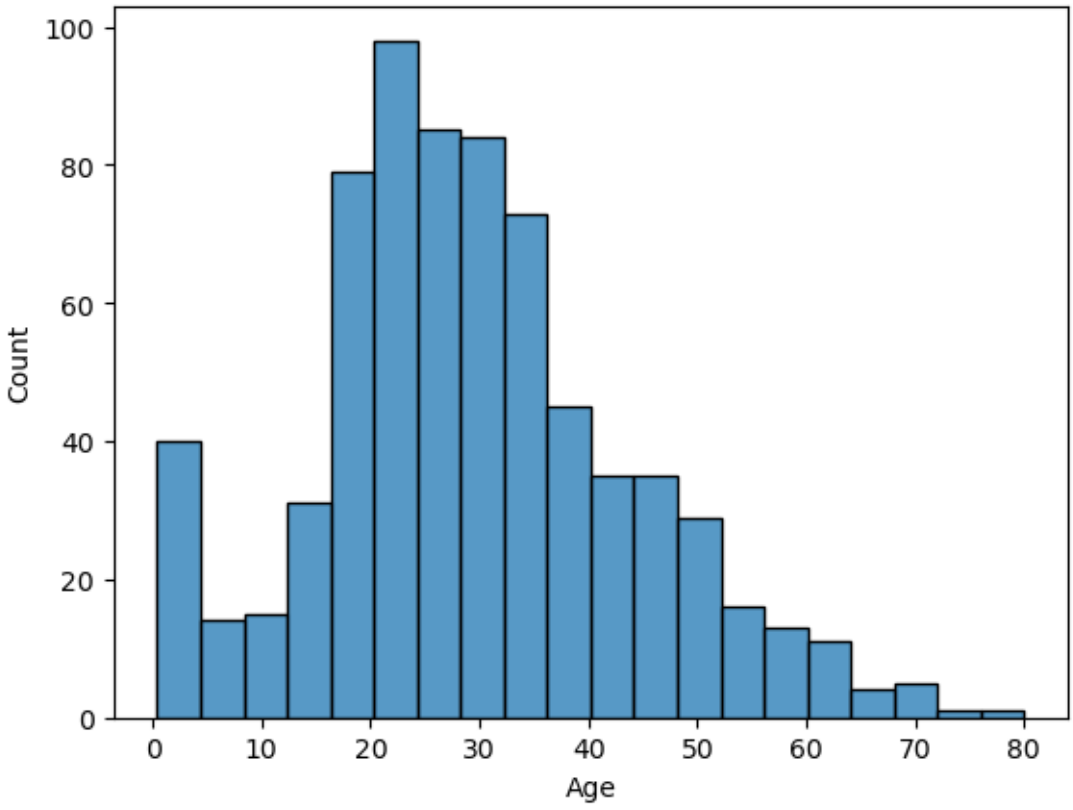
[6]:
```
sns.countplot(x='Sex', data=df)
```

[6]:
```
<Axes: xlabel='Sex', ylabel='count'>
```



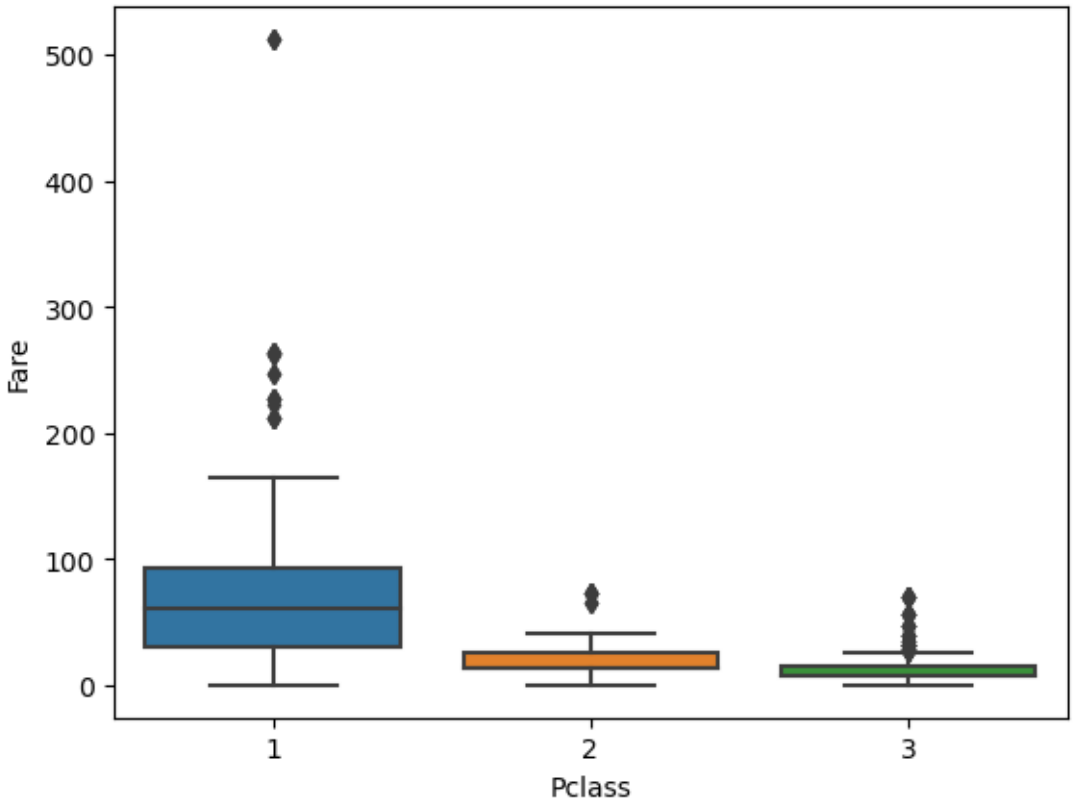
[7]:
```
sns.histplot(df['Age'])
```

[7]: <Axes: xlabel='Age', ylabel='Count'>



[8]:
```
sns.boxplot(x='Pclass', y='Fare', data=df)
```

[8]: <Axes: xlabel='Pclass', ylabel='Fare'>



[10]:
```
Q1 = df['Age'].quantile(0.25)
Q3 = df['Age'].quantile(0.75)
IQR = Q3 - Q1
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
```

[11]:
```
X = df.drop('Survived', axis=1)
y = df['Survived']
```

[12]:
```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder, StandardScaler
```

[21]:
```
label_encoder = LabelEncoder()
X['Sex'] = label_encoder.fit_transform(X['Sex'])
print(X['Sex'])
```

```
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
```

[20]:
```
scaler = StandardScaler()
X['Age'] = scaler.fit_transform(X[['Age']])
print(X['Age'])
```

```
0     -0.530377
1      0.571831
2     -0.254825
3      0.365167
4      0.365167
         …
886   -0.185937
887   -0.737041
888         NaN
889   -0.254825
890    0.158503
Name: Age, Length: 891, dtype: float64
```

[29]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,␣
 ↪random_state=0)
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[29], line 1
----> 1 X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2 ␣
 ↪random_state=0)

NameError: name 'Y' is not defined
```

[27]:
```
X_train
```

[27]:
```
     PassengerId  Pclass                                          Name  \
140          141       3                     Boulos, Mrs. Joseph (Sultana)
439          440       2          Kvillner, Mr. Johan Henrik Johannesson
817          818       2                              Mallet, Mr. Albert
378          379       3                             Betros, Mr. Tannous
491          492       3                             Windelov, Mr. Einar
..           …     …                                               …
835          836       1                  Compton, Miss. Sara Rebecca
192          193       3  Andersen-Jensen, Miss. Carla Christine Nielsine
629          630       3                      O'Connell, Mr. Patrick D
```

```
559          560       3     de Messemaeker, Mrs. Guillaume Joseph (Emma)
684          685       2                 Brown, Mr. Thomas William Solomon

     Sex       Age  SibSp  Parch            Ticket     Fare Cabin Embarked
140    0       NaN      0      2              2678  15.2458   NaN        C
439    1  0.089615      0      0       C.A. 18723  10.5000   NaN        S
817    1  0.089615      1      1   S.C./PARIS 2079  37.0042   NaN        C
378    1 -0.668153      0      0              2648   4.0125   NaN        C
491    1 -0.599265      0      0  SOTON/OQ 3101317   7.2500   NaN        S
..   …       …    …    …               …       …   …       …
835    0  0.640719      1      1          PC 17756  83.1583   E49        C
192    0 -0.737041      1      0            350046   7.8542   NaN        S
629    1       NaN      0      0            334912   7.7333   NaN        Q
559    0  0.434055      1      0            345572  17.4000   NaN        S
684    1  2.087366      1      1             29750  39.0000   NaN        S

[712 rows x 11 columns]
```

[30]:
```
print(X_train.shape)
print(X_test.shape)
print(y_train.shape)
print(y_test.shape)
```

```
(712, 11)
(179, 11)
(712,)
(179,)
```