import libraries

In [12]:
```
import pandas as ps
```

Load the Dataset

In [13]:
```
url = 'https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.cs
data = pd.read_csv(url)
```

Inspect the Data

In [14]:
```
# View the first few rows
print(data.head())
```

```
   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  
```

In [15]:
```
# Check for missing values
print(data.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
```

In [16]:
```
# Check data types
print(data.dtypes)
```

```
PassengerId      int64
Survived         int64
Pclass           int64
Name            object
Sex             object
Age            float64
SibSp            int64
Parch            int64
Ticket          object
Fare           float64
Cabin           object
Embarked        object
dtype: object
```

Handle Missing Values

In [17]:

```
from sklearn.impute import SimpleImputer

# Define numerical and categorical features
numeric_features = ['Age', 'Fare']
categorical_features = ['Embarked', 'Sex', 'Pclass']

# Impute missing values for numerical features with median
numeric_imputer = SimpleImputer(strategy='median')
data[numeric_features] = numeric_imputer.fit_transform(data[numeric_features])

# Impute missing values for categorical features with the most frequent value
categorical_imputer = SimpleImputer(strategy='most_frequent')
data[categorical_features] = categorical_imputer.fit_transform(data[categorical_fea

# Verify there are no missing values left
print(data.isnull().sum())
```

```
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
```

Encode Categorical Variables

In [18]:

```
from sklearn.preprocessing import OneHotEncoder

# One-Hot Encode categorical features
encoder = OneHotEncoder(handle_unknown='ignore')
encoded_categories = encoder.fit_transform(data[categorical_features])

# Convert encoded categories to a DataFrame
encoded_df = pd.DataFrame(encoded_categories.toarray(), columns=encoder.get_feature

# Drop the original categorical features and concatenate the encoded ones
data = data.drop(columns=categorical_features)
data = pd.concat([data, encoded_df], axis=1)
```

Scale Numerical Features

In [19]:

```python
from sklearn.preprocessing import StandardScaler

# Standardize numerical features
scaler = StandardScaler()
data[numeric_features] = scaler.fit_transform(data[numeric_features])

# View the preprocessed data
print(data.head())
```

```
   PassengerId  Survived                                               Name  \
0            1         0                            Braund, Mr. Owen Harris
1            2         1  Cumings, Mrs. John Bradley (Florence Briggs Th...
2            3         1                             Heikkinen, Miss. Laina
3            4         1       Futrelle, Mrs. Jacques Heath (Lily May Peel)
4            5         0                           Allen, Mr. William Henry

        Age  SibSp  Parch            Ticket      Fare Cabin  Embarked_C  \
0 -0.565736      1      0         A/5 21171 -0.502445   NaN         0.0
1  0.663861      1      0          PC 17599  0.786845   C85         1.0
2 -0.258337      0      0  STON/O2. 3101282 -0.488854   NaN         0.0
3  0.433312      1      0            113803  0.420730  C123         0.0
4  0.433312      0      0            373450 -0.486337   NaN         0.0

   Embarked_Q  Embarked_S  Sex_female  Sex_male  Pclass_1  Pclass_2  Pclass_3
0         0.0         1.0         0.0       1.0       0.0       0.0       1.0
1         0.0         0.0         1.0       0.0       1.0       0.0       0.0
2         0.0         1.0         1.0       0.0       0.0       0.0       1.0
3         0.0         1.0         1.0       0.0       1.0       0.0       0.0
4         0.0         1.0         0.0       1.0       0.0       0.0       1.0
```

Split Features and Target Variable

In [20]:

```python
# Separate target variable and features
X = data.drop(columns=['Survived', 'Name', 'Ticket', 'Cabin'])
y = data['Survived']

# Verify the shapes of X and y
print(X.shape, y.shape)
```

```
(891, 13) (891,)
```

In [ ]: