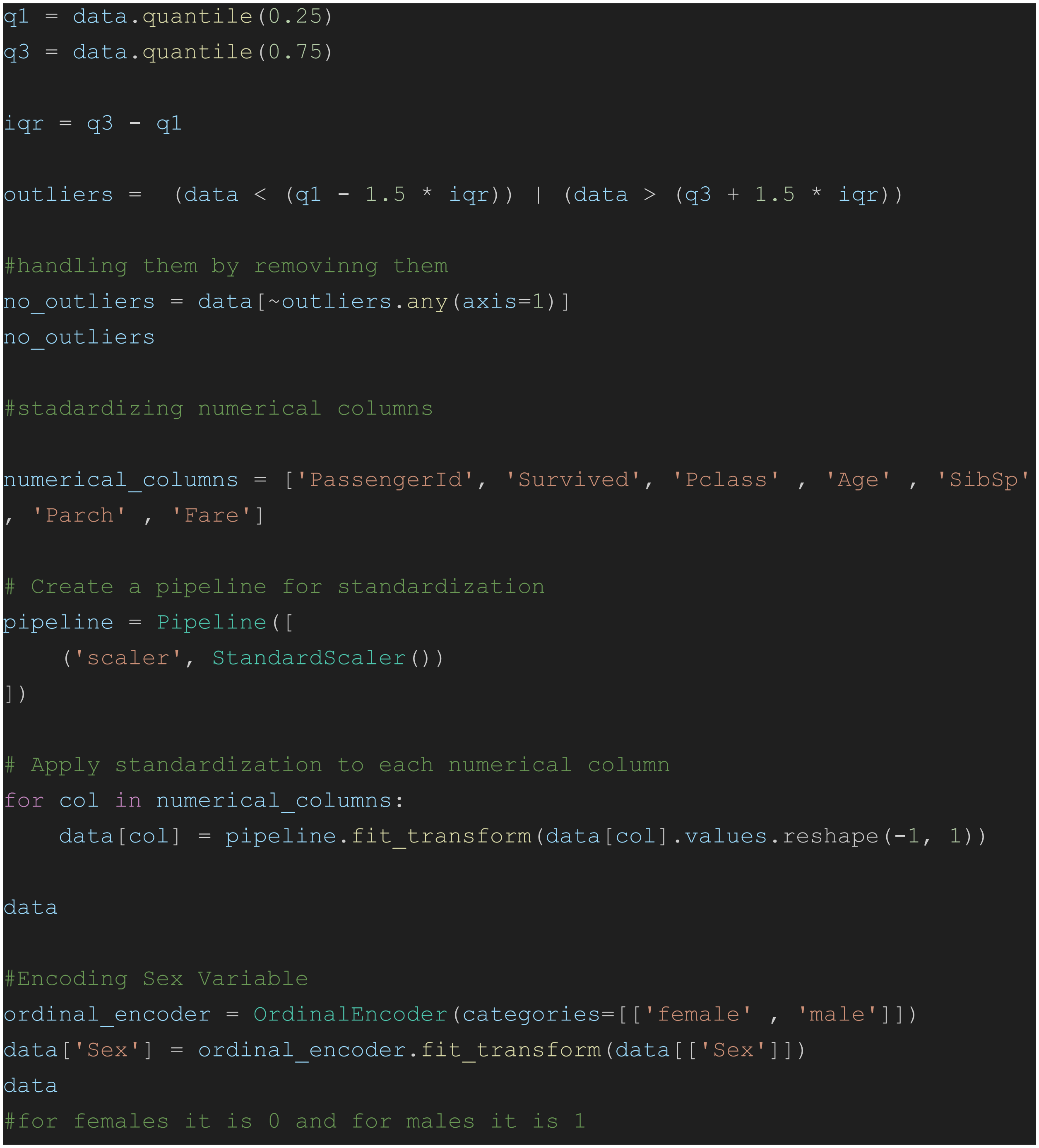
```
# -*- coding: utf-8 -*-
"""Alaiba_Nawaz_Day1.ipynb

Automatically generated by Colaboratory.

Original file is located at

https://colab.research.google.com/drive/1bR68rs_JAH9CMmo1Sz9Jff7joTnlM3OJ
"""

import pandas as pd
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import OrdinalEncoder

#loading data in dataframe
df1 = pd.read_csv("/content/test.csv")
df2 = pd.read_csv("/content/train.csv")

#combining training ad testing data
data = pd.concat([df1,df2])

#getting informmation about data
data.info()

#dropppping duplicates
data.drop_duplicates(inplace = True)
data

#dropping missing values
data.dropna(inplace = True)
data

#check and handle outlier

#plotting box plot to check for outlier
sns.boxplot(data=data)

#finding outliers
```



```
q1 = data.quantile(0.25)
q3 = data.quantile(0.75)

iqr = q3 - q1

outliers =  (data < (q1 - 1.5 * iqr)) | (data > (q3 + 1.5 * iqr))

#handling them by removinng them
no_outliers = data[~outliers.any(axis=1)]
no_outliers

#stadardizing numerical columns

numerical_columns = ['PassengerId', 'Survived', 'Pclass' , 'Age' , 'SibSp'
, 'Parch' , 'Fare']

# Create a pipeline for standardization
pipeline = Pipeline([
    ('scaler', StandardScaler())
])

# Apply standardization to each numerical column
for col in numerical_columns:
    data[col] = pipeline.fit_transform(data[col].values.reshape(-1, 1))

data

#Encoding Sex Variable
ordinal_encoder = OrdinalEncoder(categories=[['female' , 'male']])
data['Sex'] = ordinal_encoder.fit_transform(data[['Sex']])
data
#for females it is 0 and for males it is 1
```