**EXPERIMENT-1**

**Student’s Name: Pratham Mittal UID: 20BCS1040**

**Section/Group: 707\_WM\_B Subject Code: 20CSP-317**

**Subject Name: ML Lab Date of performance:28/8/2022**

**Branch: BE CSE Semester:5th**

**Aim:** To perform EDA analysis on the given dataset.

**Objective:** To do Exploratory Data Analysis.

**Software/Hardware Requirements:** Windows 7 & above version.

**Tools to be used:**

1. Anaconda Jupyter Notebook
2. Kaggle

**Introduction**

Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.

**Dataset**

The data has been split into two groups:

* training set (train.csv)
* test set (test.csv)

The training set should be used to build your machine learning models. For the training set, we provide the outcome (also known as the “ground truth”) for each passenger. The model will be based on “features” like passengers’ gender and class.

The titanic dataset we have taken in to account for the EDA. Train.csv will contain the details of a subset of the passengers on board (891 to be exact) and importantly, will reveal whether they survived or not, also known as the “ground truth”.

**Code:**

**First few commands are to know about the dataset like the columns, rows and characteristics of the dataset.**

import pandas as pd

df=pd.read\_csv("titanic-train.csv")

# The above command help to read files from the user into variable for analysis.

df.head() # This function returns the first n rows for the object based on position.

data.nunique() # Count number of distinct elements in specified axis.

df.describe() # The describe() function applies basic statistical computations on the dataset like extreme values, count of data points standard deviation, etc

df.shape # We will use shape parameter to get the shape of the dataset.

df.isnull()

df.isnull().sum()

# We will check if our data contains any missing values or not. Missing values can occur when no information is provided for one or more items or for a whole unit. We will use the isnull() method.

data # To print the data in the module in the notebook.

df.value\_counts("PassengerId") # Return a Series containing counts of unique rows in the Data Frame.

data = df.drop\_duplicates(subset ="PassengerId",) # Pandas drop\_duplicates() method helps in removing duplicates from the data frame.

**The next few commands are to plot the various graphs and visualize data in the form of different type of plots**

# importing packages

import seaborn as sns

import matplotlib.pyplot as plt

sns.countplot(x='PassengerId', data=df, )

plt.show()

#Data Visualization

sns.scatterplot(x='PassengerId', y='Age', hue='gender', data=df, )

plt.show()

sns.scatterplot(x='PassengerId', y='Ticket', hue='gender', data=df, )

# Placing Legend outside the Figure

plt.legend(bbox\_to\_anchor=(1, 1), loc=2)

plt.show()

sns.pairplot(df.drop(['PassengerId'], axis = 1),

hue='Age', height=2)

#Histograms

fig, axes = plt.subplots(2, 2, figsize=(10,10))

axes[0,0].set\_title("PassengerId")

axes[0,0].hist(df['gender'], bins=7)

axes[0,1].set\_title("PassengerId")

axes[0,1].hist(df['Age'], bins=5);

axes[1,0].set\_title("PassengerId")

axes[1,0].hist(df['Ticket'], bins=6);

axes[1,1].set\_title("PassengerId")

axes[1,1].hist(df['Fare'], bins=6);

sns.pairplot(df)

data.corr(method='pearson')

#correlation

sns.heatmap(df.corr(method='pearson').drop(

['PassengerId'], axis=1).drop(['PassengerId'], axis=0),

annot = True);

plt.show()

def graph(y):

sns.boxplot(x="PassengerId", y=y, data=df)

plt.figure(figsize=(10,10))

# Adding the subplot at the specified

# grid position

plt.subplot(224)

graph('Age')

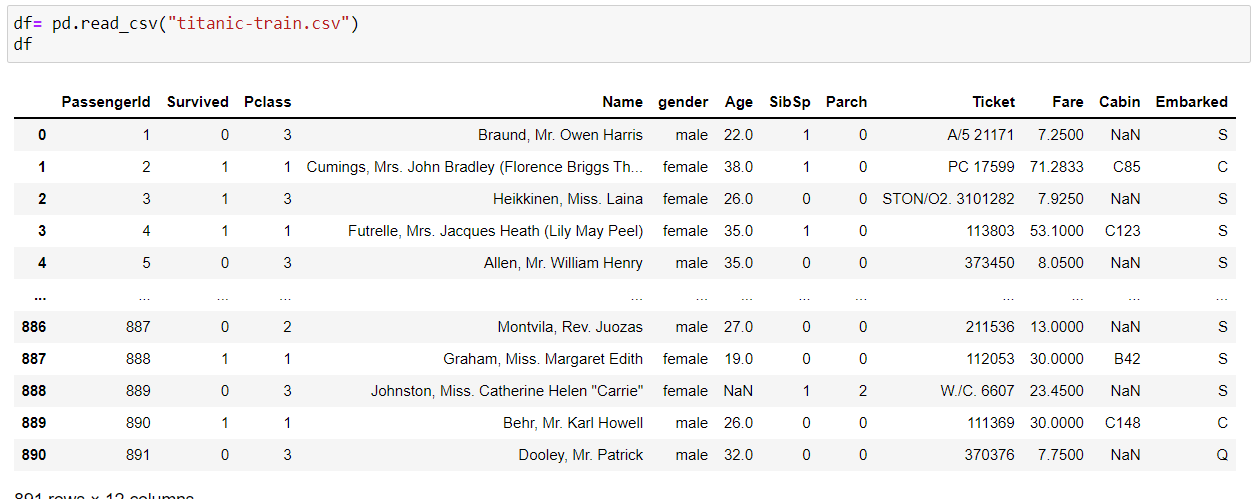
plt.show()

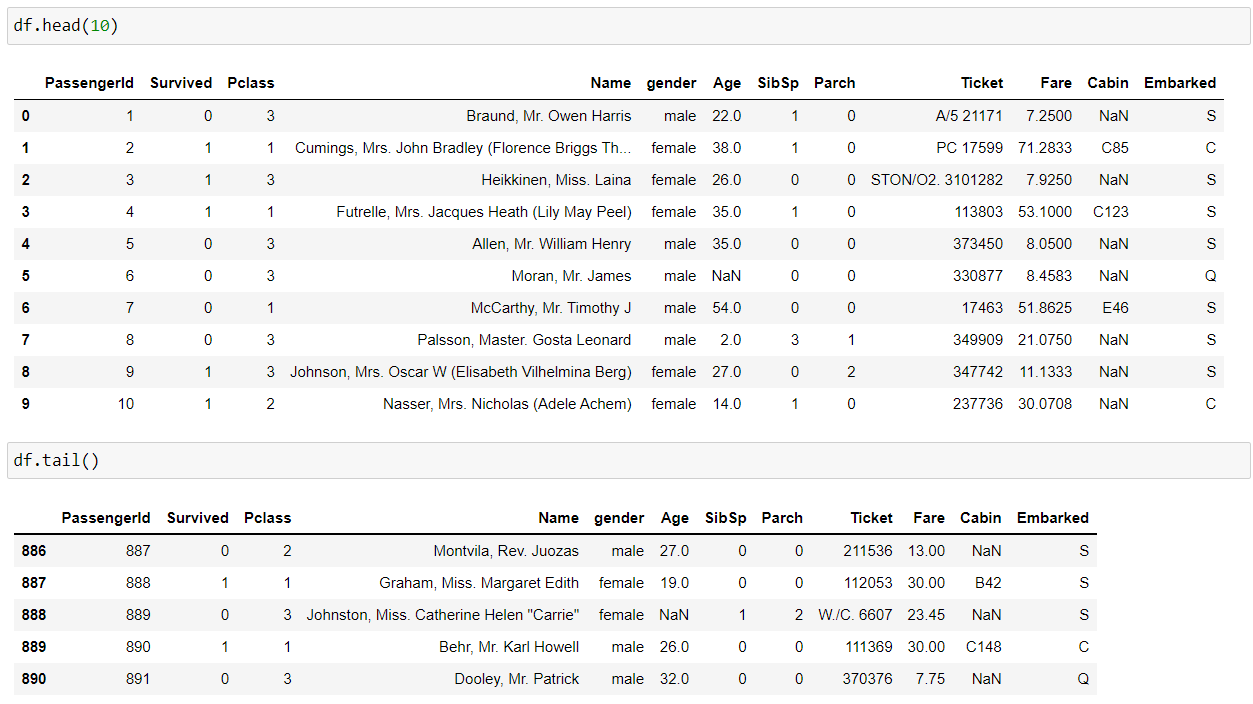
passenger=df.drop("gender",axis=1)

passenger

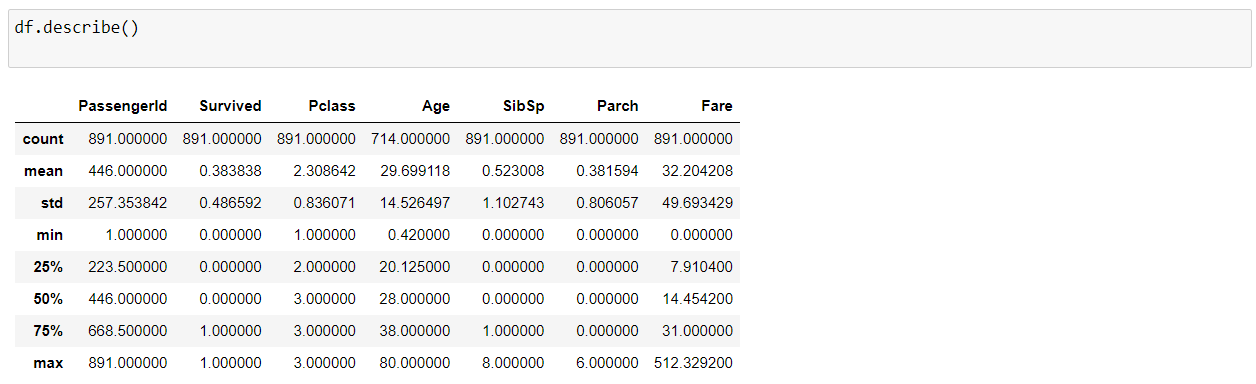
sns.relplot(x ="PassengerId",y ="Ticket",hue= "Age", data= passenger)

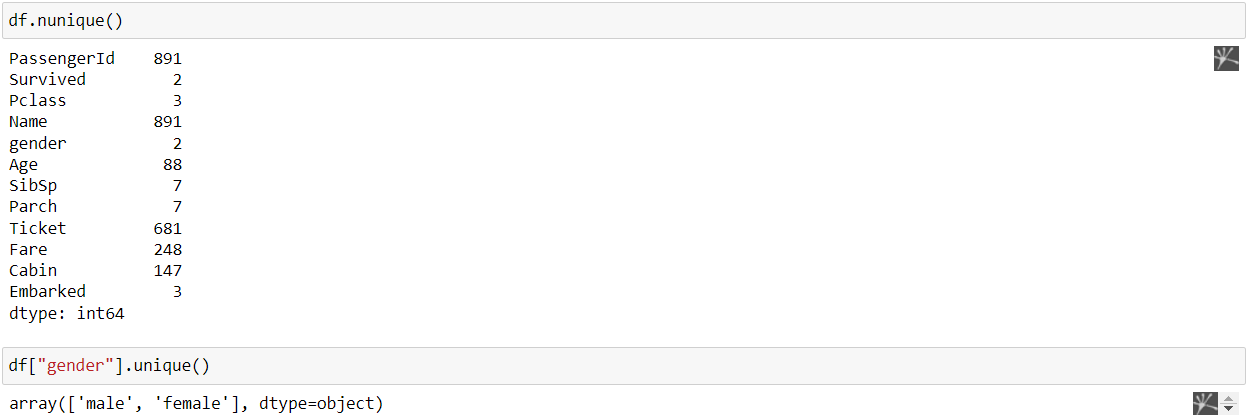
**Output:**



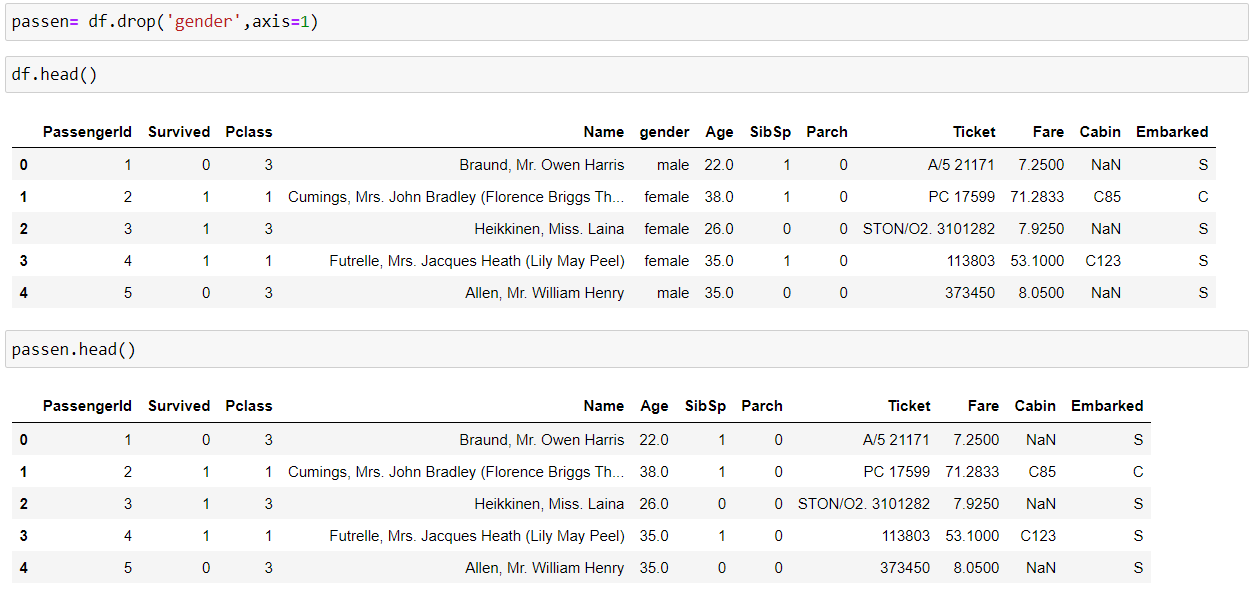


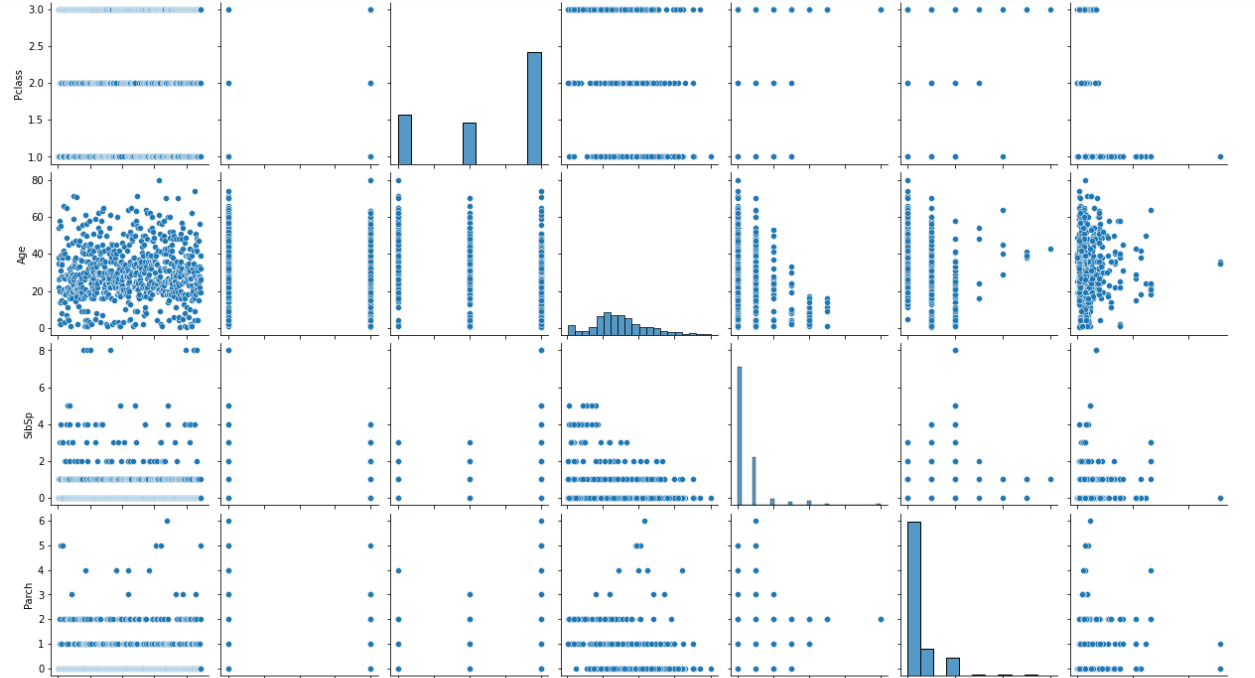


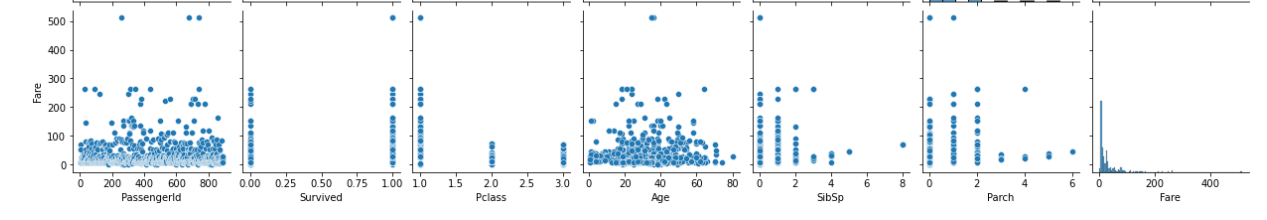


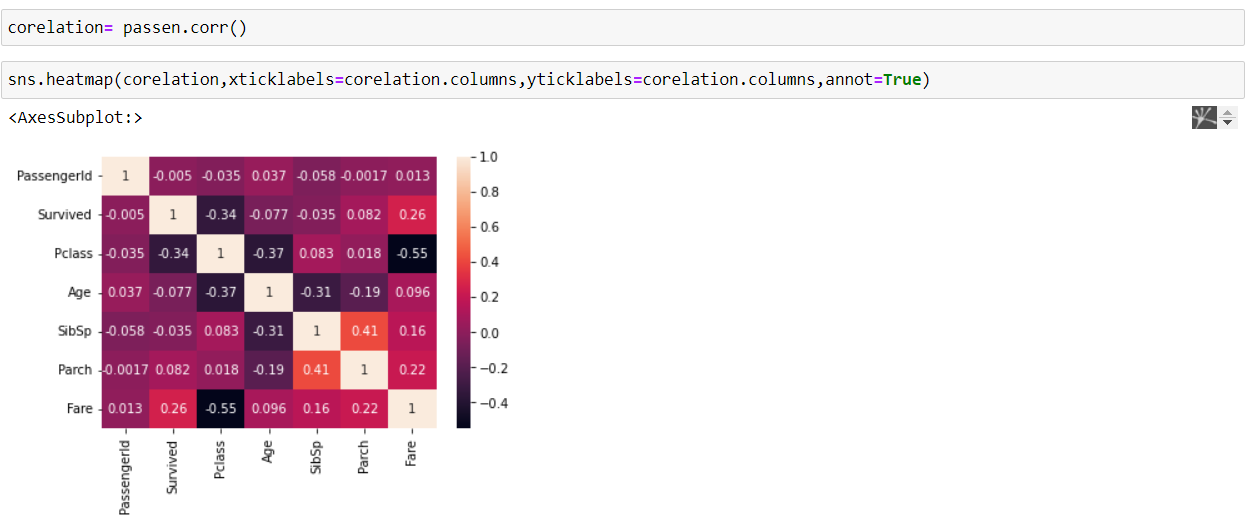


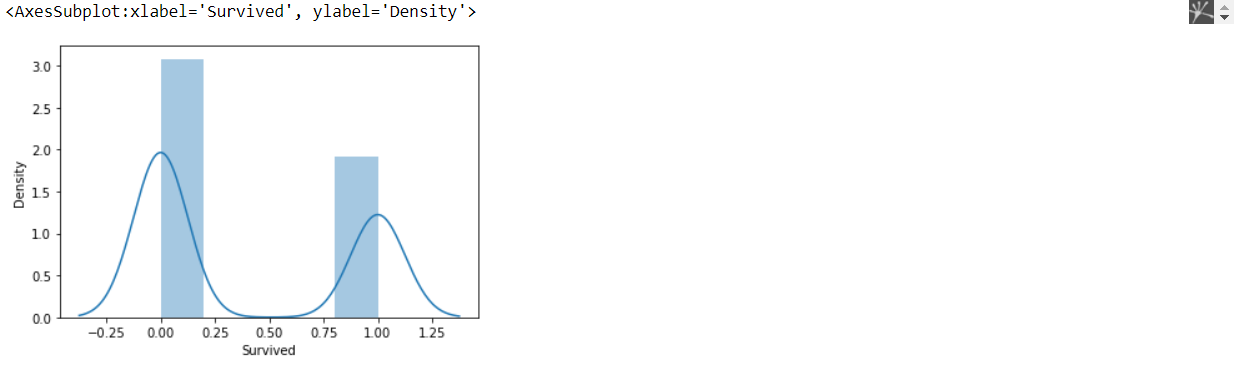
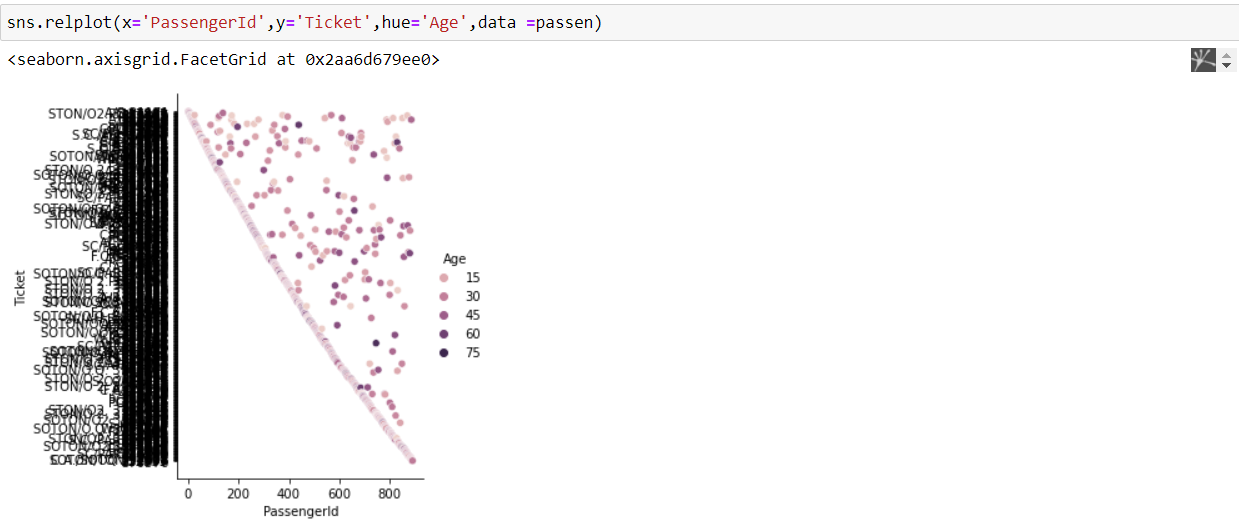
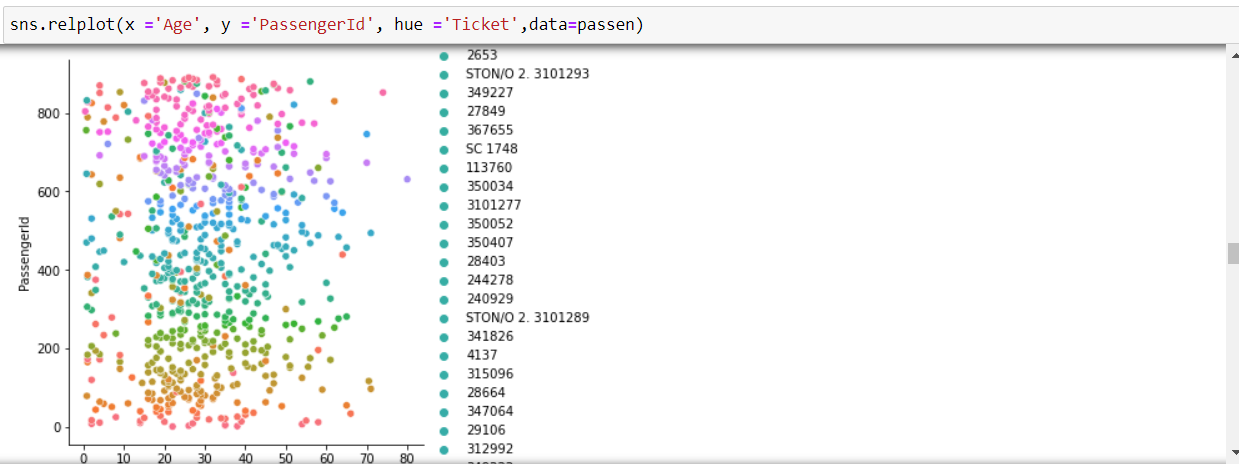












**Learning Outcomes:**

1.We learnt about how to perform EDA on any dataset given to us.

2.We learnt about the characteristics of the dataset that is about the various rows and columns with all the information stored in the dataset.

3.We learnt about how we use different functions and techniques to extract information from data and create plots and also take data in tabulated form.