Green University Of Bangladesh

Department Of Computer Science and Engineering (CSE)

Faculty of Sciences and Engineering

Semester: (Fall, Year: 2023), B.Sc. in CSE (DAY)

LAB REPORT NO - 04

Course Title: Data Mining Lab

Course Code: CSE-436 **Section:** D2

Lab Experiment Name: Statistical Analysis and Data Visualization using Python

Student Details

	Name	ID
1	Shamim Ahmed	201902067

Lab Date : 20th October 2023

Submission Date : 27th October 2023

Course Teacher Name : Rezwanul Haque

Lab Report Status

Mark:..... **Signature:.....**

Comments:..... **Date:.....**

1 INTRODUCTION

In this lab report we are going develop a Colab using all of the data processing. We will Choose an appropriate dataset from kaggle containing NULL and garbage values then we will then show the effect of all visualizing techniques with proper documentation in the notebook and will create two dataframe and show their correlation using Pearson's correlation and visualize with appropriate plot.

2 OBJECTIVE

The aim of this lab is to interpret data more easily and visualize the dataset effectively in data mining, machine learning and other data science tasks.To visualise raw data more effectively we are going to know different techniques like dispot,boxplot and different data visualisation techniques.

3 DATASET

For this lab I use the "titanic" dataset from Kaggle which is a well-known dataset. For the environment I used the Colab platform.

4 IMPLEMENTATION

4.1 Load the dataset

First of all I download the "Titanic" dataset and load the the dataset on df

```
import pandas as pd
import numpy as np

# Load the dataset
df = pd.read_csv('/content/titanic.csv')

print("Original Dataset:")
print(df.head())
```

Listing 1: Importing the librarys and Loading the dataset

4.2 Checking for NULL and garbage values

Then to checks for NULL values in the DataFrame I used the isnull() method, and then sums up the number of NULL values in each column using the sum() function.

```
# 1. Checking for NULL and garbage values
print("\nChecking for NULL and garbage values:")
```

```
print(df.isnull().sum())
```

Listing 2: Checking for NULL and garbage values

4.3 Distribution Plot

Here I use distribution plots . This plot gives us a combination of both probability density functions(pdf) and histogram in a single plot.

```
import seaborn as sns
sns.distplot(df['Age'], bins=10, kde=True, rug=False)
```

Listing 3: Distribution Plot

4.4 Box Plot

With the help of a box plot,I determine the Interquartile range(IQR) where maximum details of the data will be present.

```
sns.catplot(x="Sex", y="Age", kind="box", data=df)
```

Listing 4: Box Plot

4.5 Violin Plot

Here plotting a violin plot using the DataFrame df, with the x-axis representing the 'Sex' column and the y-axis representing the 'Age' column. The argument size=6 is likely specifying the size of the plot.

```
sns.violinplot(x="Sex",y="Age",data=df,size=6)
```

Listing 5: Violin Plot

4.6 Bar Plot

Here I use the Seaborn library in Python to create a categorical plot (catplot) in the form of a bar plot. It seems to be plotting the relationship between the 'Sex' and 'Survived' columns from the DataFrame df

```
sns.catplot(x="Sex", y="Survived", kind="bar", data = df)
```

Listing 6: Bar Plot

4.7 Scatter Plot

Here utilizing the Seaborn library I create a joint plot. It plotting the relationship between the 'Age' and 'Fare' columns from the DataFrame df.

```
sns.jointplot(x="Age", y="Fare", data=df)
```

Listing 7: Scatter Plot

4.8 Pair Plot

Here using the Seaborn library I create a pair plot, which is a grid of pairwise relationships in a dataset. It appears to be plotting the relationships between all numerical columns in the DataFrame df

```
sns.pairplot(df)
```

Listing 8: Pair Plot

4.9 Multivariate Analysis

Here I creating a categorical plot using the Seaborn library. It visualizing the relationship between the categorical variables 'Sex' and 'Survived' while using the 'Pclass' variable as a hue. The plot type specified is a bar plot.

```
sns.catplot(x="Sex", y="Survived", hue="Pclass", kind="bar", data =df)
```

Listing 9: Multivariate Analysis

4.10 Pearson's Correlation

Finally to calculating the correlation between the 'Age' and 'Survived' columns in the DataFrame df, The Pearson correlation coefficient measures the linear relationship between two variables and has a value between -1 and 1, where 1 indicates a strong positive correlation, -1 indicates a strong negative correlation, and 0 indicates no linear correlation. The Spearman correlation coefficient assesses the monotonic relationship between two variables and has a value between -1 and 1, similar to the Pearson coefficient.

```
print(df["Age"].corr(df["Survived"]))
print(df["Age"].corr(df["Survived"], method='spearman'))
```

Listing 10: Pearson's Correlation

5 OUTPUT

```
Original Dataset:
   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
```

Figure 1: Loading the dataset

```
Checking for NULL and garbage values:
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
```

Figure 2: Checking for NULL and garbage values

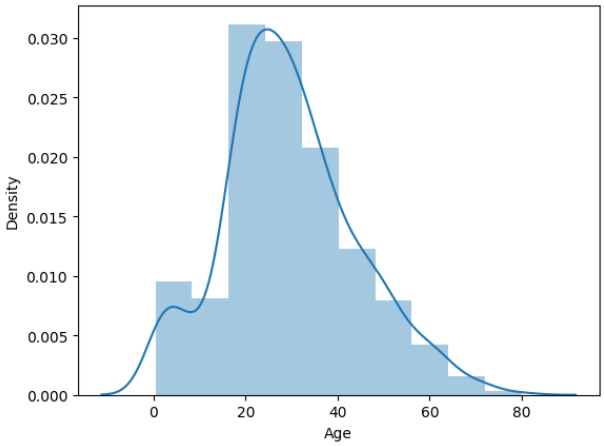

Figure 3: Distribution Plot

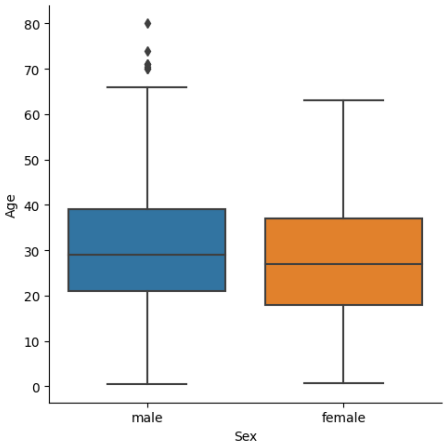

Figure 4: Box Plot

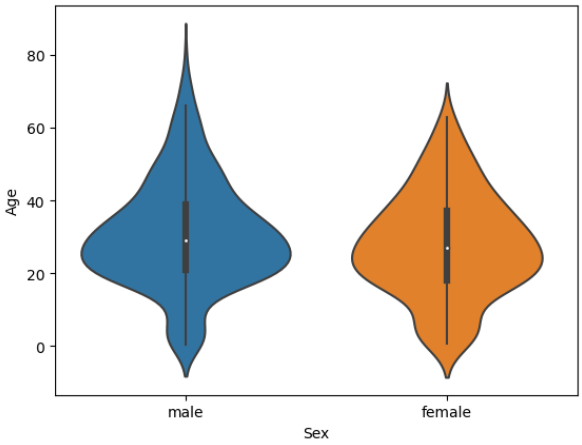

Figure 5: Violin Plot

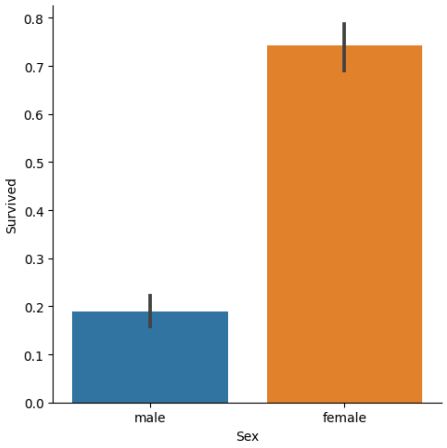

Figure 6: Bar Plot

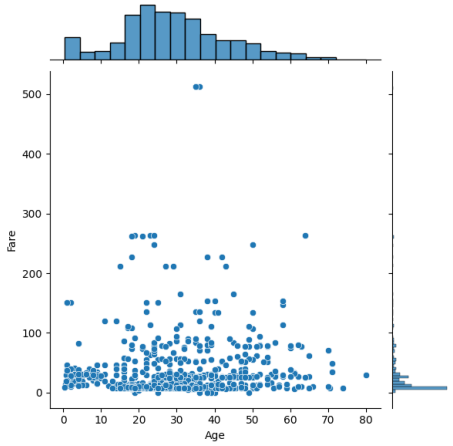

Figure 7: Scatter Plot

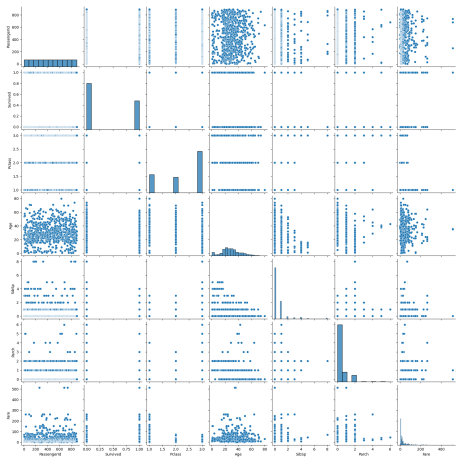

Figure 8: Pair Plot

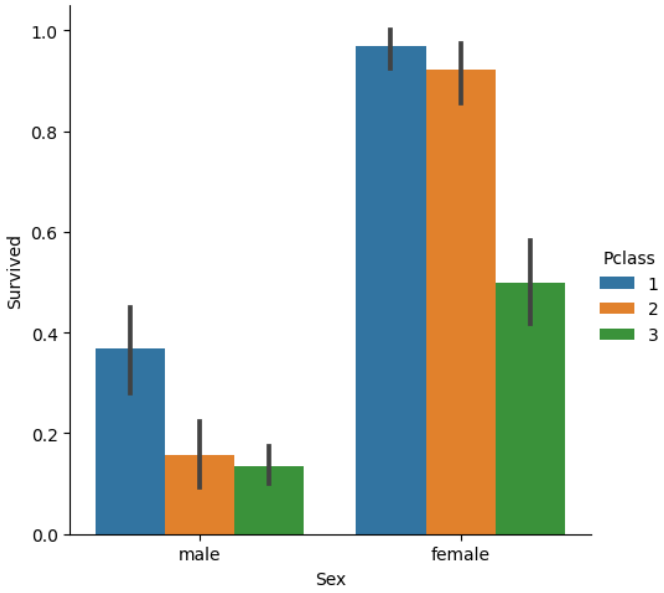

Figure 9: Multivariate Analysis

```
[ ]  print(df["Age"].corr(df["Survived"]))
     print(df["Age"].corr(df["Survived"], method='spearman'))

     -0.07722109457217768
     -0.05256530004469449
```

Figure 10: Multivariate Analysis

6 DISCUSSION & ANALYSIS

In this lab report, I perform data checks, visualization, and correlation analysis on titanic dataset.Firstly begin by checking for missing values and assessing data integrity. Subsequently, then generate visualizations including violin plots, categorical bar plots, joint plots, and pair plots, which aid in understanding data distributions and potential correlations between variables. Finally, I computes both Pearson and Spearman correlation coefficients to quantify the relationships between specific variable These operations help ensure data quality, display relationships between variables, and assess correlations, contributing to a comprehensive understanding of the dataset's characteristics and potential insights.