Piyusha Supe.

Experiment - 09 (Group A)

- Title of the assignment - Data Visualization II.

- Problem Statement -

  (1) Use the in built dataset 'titanic' as used in previous. plot a box plot for the distribution of age with respect to each gender along with the information about whether they survived or not. (Column names - 'sex' and 'age')

  (2). Write observations on the inference from above statistics.

- Objective : Students should be able to perform the data visualization using python on any open source dataset

- Prerequisite: 1. Basic of python.
  2. Seaborn library, Concept of data visualization.

- Contents of theory - 1. Exploratory data analysis.
  2. Univariate analysis.

- THEORY :

1) Exploratory Data analysis :-

- There are various techniques to understand the data, and the basic need is the knowledge of Numpy, for mathematical operations and pandas for data manipulation.

2] Univariate Analysis -

- Univariate analysis is the simplest form of analysis where we explore a single variable.
- Univariate analysis is performed to describe the data in a better way. We perform it on numerical and categorical variables differently because plotting uses different plots.

3] Categorical data -

- A variable data that has text based information is referred to as categorical variable. Now following are various plots which can be used for visualizing catagorical data.

(a). Count Plot -

- Count plot is basically a count of frequency plot in form of a bar graph.
- It plots the count of each catagory in a separate bar.
- When we use the pandas value counts function on any column.. It is the same visual form of the value counts function. In our data target variable is survived and it is categorical so plot a countplot of this.

```
sns. Countplot (data ['Survived']
plt. show().
```

(b). Pie Chart -

- Pie chart is also the same as countplot, only gives us additional information about the percentage of presence of each category in data means which category is getting how much weightage in data.

Now we check about the sex column, what is a percentage of male and female members traveling.

```
data ['Sex']. value_counts (). plot (kind = "pie" , auto pct = "%.2f")
```

4] Numerical data -

Analyzing numerical data is important because understanding the distribution of variables helps to further process the data. Most of the time, we will find much inconsistency with numerical data.

(1) Histogram -
A histogram is a value distribution plot of numerical columns. Basically creates bins in various ranges in values and plots it where we can visualize how values are distributed.

(2) Distplot - Improved version of histogram. It gives us (KDE) Kernel Density Estimation.

(3). Boxplot - It plots a fire number summary to get

- Median - middle value in series after sorting.
- Percentile - Any number of values present before percentile so it explains total of 50 values below 25th percentile.
- Minimum and maximum -

IQR = Q3 - Q1

Lower_ boundary = Q1 - 1.5 * IQR

Upper_ boundary = Q3 + 1.5 * IQR.

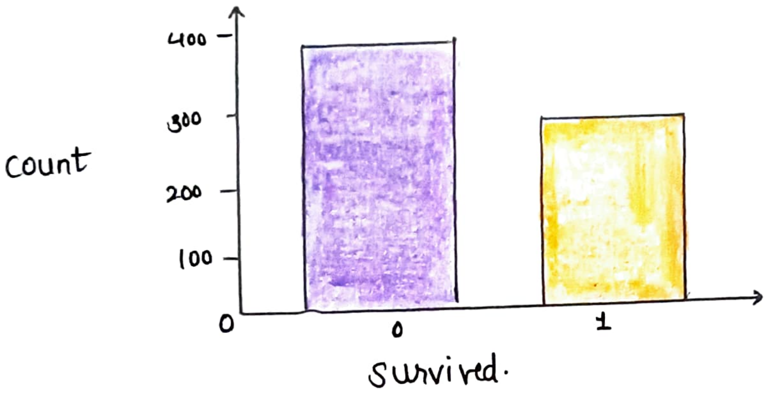
5] Multivariate Analysis -

Relationship between 2 variables is shown by multivariate analysis.

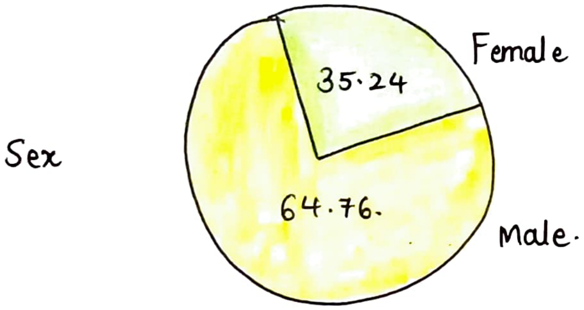
(1) Scatter plot -
To plot relationship between two numerical variables scatter plot is a simple plot to show that.

```
sns. scatterplot( tips ["total_ bill"], tips ["tip"]).
```

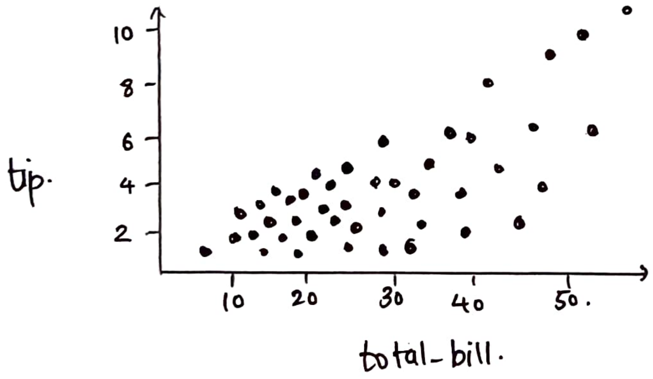
COUNT PLOT.



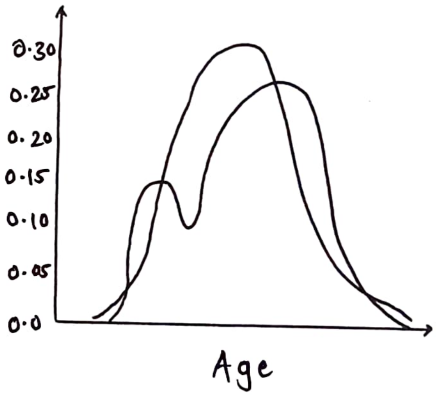
PIE PLOT (PIE CHART)



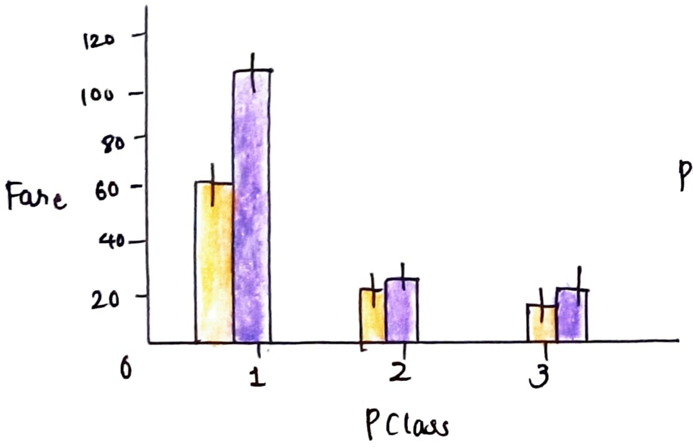
SCATTER PLOT.



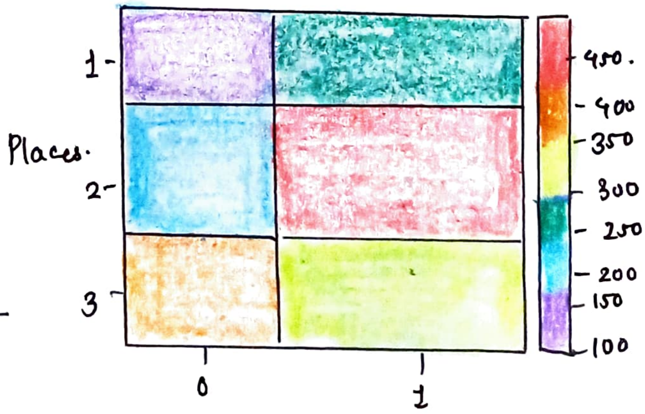
DIST PLOT.



BAR PLOT.



HEAT MAP.



* Multivariate Analysis with scatter plot –

```
sns. scatterplot (tips ["total-bills"], tips ["tip"], hue =tips
            ["sex"]).
```

6] Numerical and categorical –

If one variable is numerical and other is categorical then there are various plots that we can use for this.

a) Bar plot – Simple plot which can be used to plot categorical variable on x-axis and numerical variable on y-axis.

```
sns.barplot (data ['Pclass'], data ['Age'])
```

Similarly, we can use multivariate analysis using box plot, distplot, bar plot for

7) Categorical and categorical –

(a) Heat map –

If you have ever used a crosstab function of pandas then Heatmap is a similar visual representation of that only. It basically shows presence of one category. concerning another category is present in the dataset.

```
pd. crosstab (data ['Pclass'], data ['Survived'])
```

Survived	0	1
Pclass		
1	80	136
2	97	87
3	372	119.

Now with heatmap - sns. heatmap (pd. crosstab (data ['Pclass']
, data ['Survived'])]

(b) Cluster map -

We can also use a cluster map to understand the relationship between two categorical variables. A cluster map basically plots a dendrogram that shows the categories of similar behavious. together.

```
sns. cluster map (pd. crosstab (data ['Parch'],
                  data ['Survived']);
```

* **Conclusion:** In this way we have explored the functions of the python library for data preprocessing. Data Wrangling Techniques and How to handle missing values on Dataset. with visualization.

* * * * * * *