Page No.
Date

Name:- Rohini Janardan Devkar
PRN no:- 72030818G
Roll no:- 23272
Class :- TE2 (comp)
DSBDA pr-9.

Practical No:-9
Data Visualization II

Aim:- 1. Use the inbuilt dataset 'titanic' as used in the above problem. Plot a box plot for 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 the above statistics.

Theory:-

* Data Visualization:-

Data Visualization is a field in data analysis that deals with visual representation of data. It graphically plots data and is an effective way to communicate inferences from data.

With pictures, maps and graphs, the human mind has an easier time processing and understanding any given data.

Python offers several plotting libraries, namely Matplotlib, Seaborn and many other such data visualization packages with different features for creating informative, customized and appealing plots to present data in the most simple and effective way.

* Benefits of Data Visualization:-

1) It promotes improved absorption of business information.

2) With the help of data visualization, decision - makers can easily understand how the data is being interproted to determine business variations.

3) A large amount of data is handled and is visualized to establish patterns in the data. Many meaningful insights and the evidence behind the data can be used to establish a business goal.

4) Visualizing the data helps managers to achieve growth and use the new pattern trends found in business strategies.

* Python Libraries:-

1) Seaborn:-

When you read the official documentation on seaborn, it is defined as the data visualization library based on Matplotlib that provides a high-level interface for drawing attractive and informative statistical graphics. Putting it simply, seaborn is an extension of Matplotlib with advanced features.

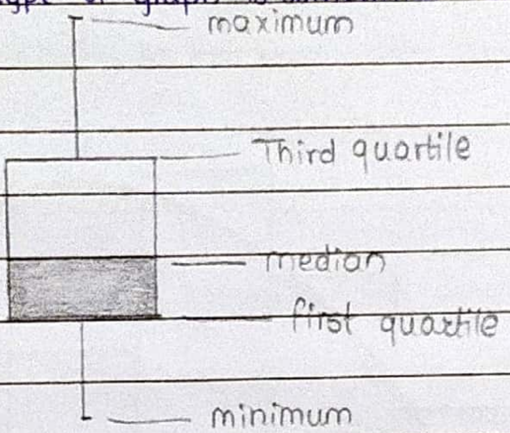
2) Matplotlib:-

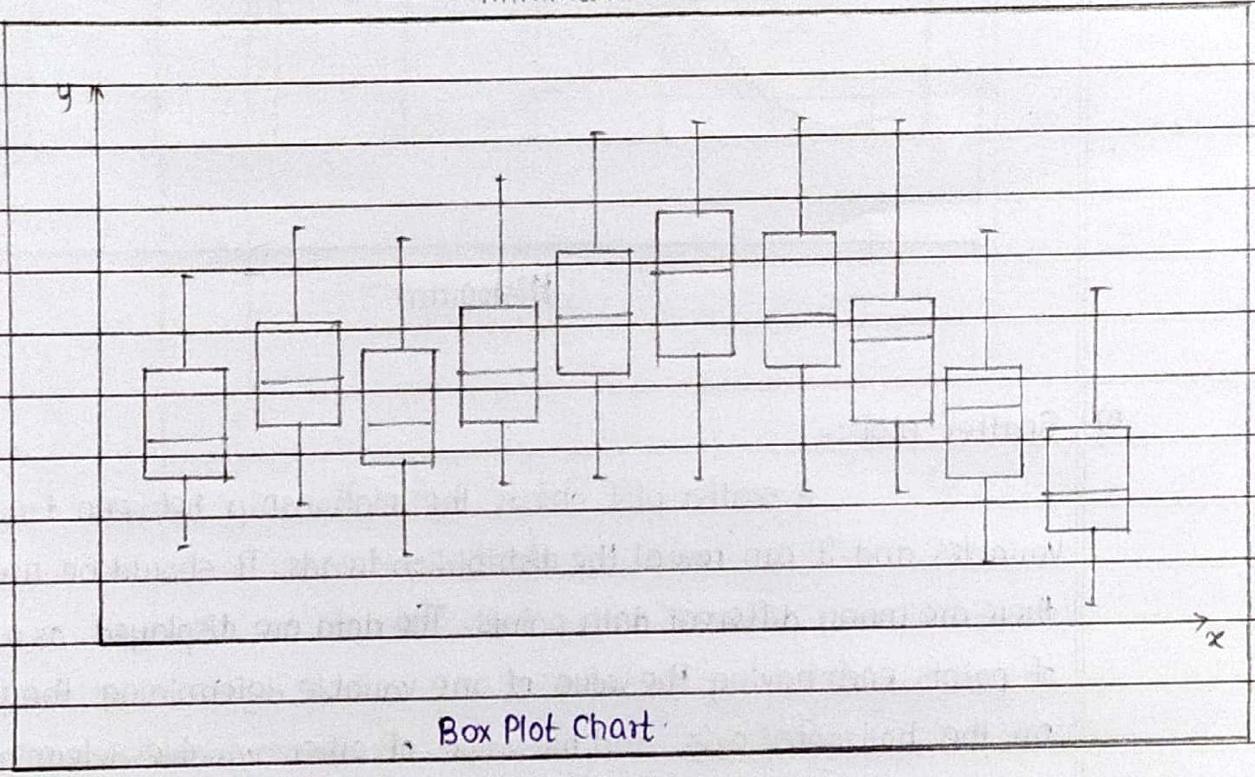
This is undoubtedly my favourite and a quintessential python library. You can create stories with the data visualized with Matplotlib. Another library from the SciPy stack, Matplotlib plots 2D figures.

* Types of Graphs:-

1) Box plot chart:-

A box plot chart is a graphical representation of statistical data based of the minimum, first quartile, median, third quartile and maximum. The term extending ~~from the top~~ "box plot" comes from the fact the graph looks like a rectangle with lines extending from the top and bottom. Because of the extending lines, this type of graph is sometimes called a box-and-whisker plot.

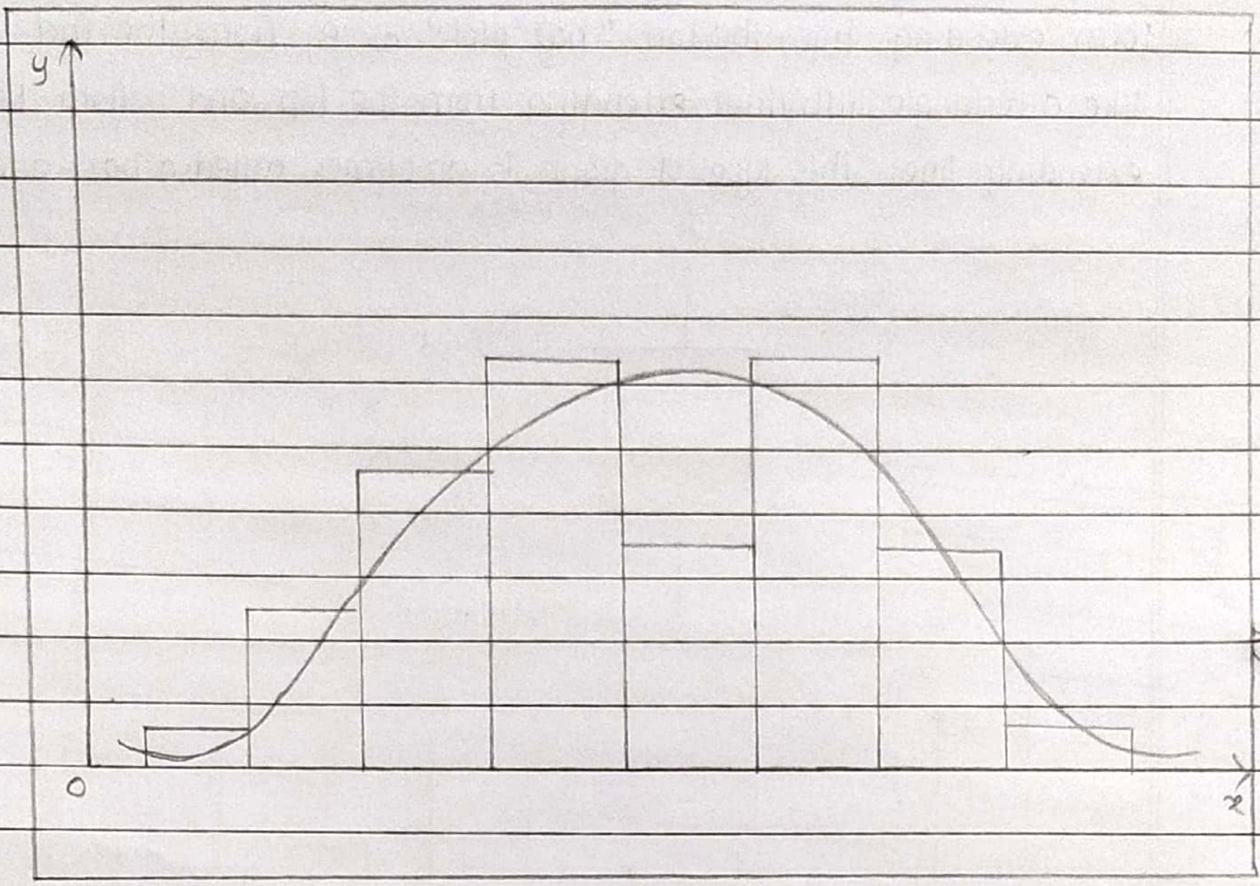




Box Plot Chart

2) Histogram:-

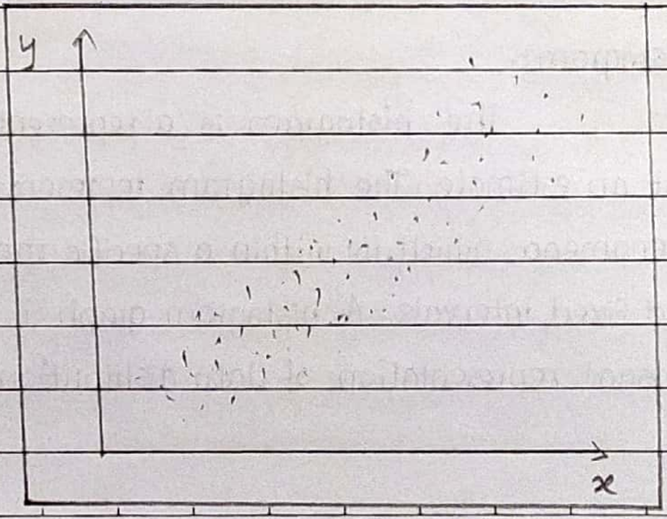
The histogram is a representation of the numerical data, not accurate but an estimate. The histogram represents the frequency of occurence of specific phenomena which lie within a specific range of values and arranged in consecutive and fixed intervals. A histogram graph is a popular graphing tool that provides a visual representation of data distribution.



Histogram

3) Scatter plot :-

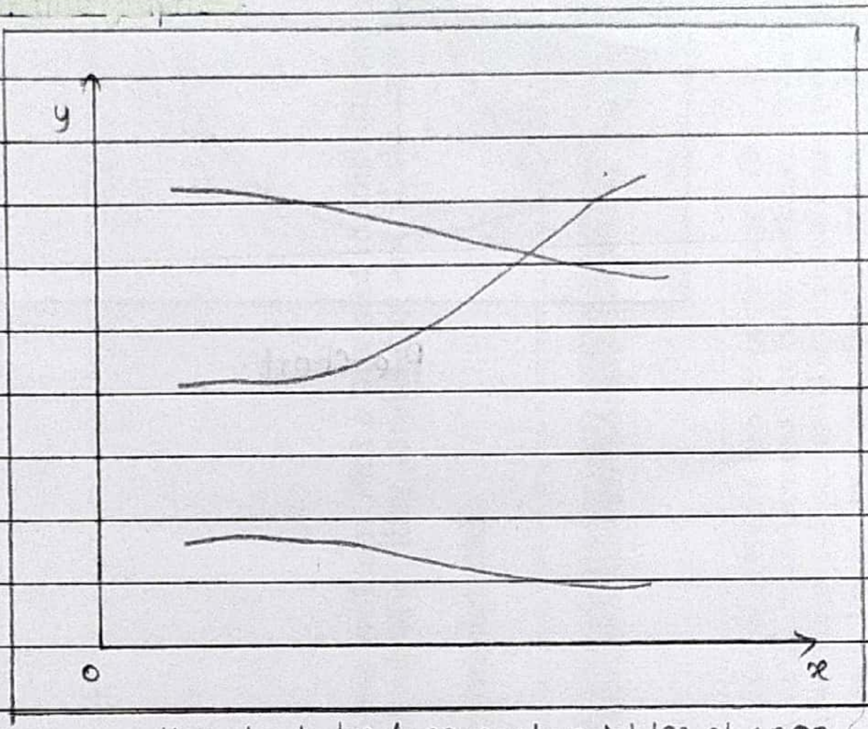
A scatter plot shows the relationship between two different variables and it can reveal the distribution trends. It should be used when there are many different data points. The data are displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of other variable determining the position on the vertical axis.



Scatter Plot chart.

4) Line graph :-

A line graph graphically displays data changes continuously over time. Each line graph consists of points that connect data to show a trend. Line graphs have x and y-axis. In the most cases, time is distributed on the horizontal axis.



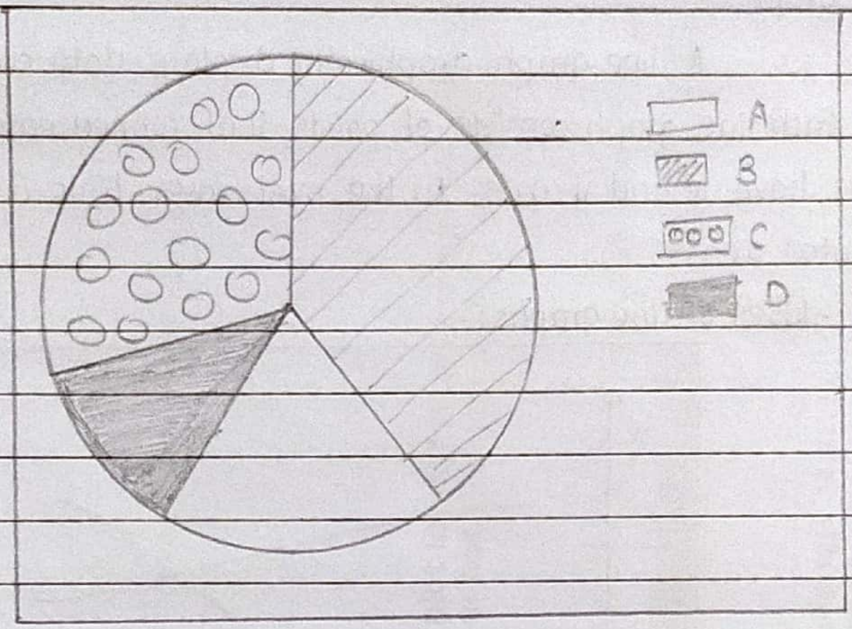
line charts track several variables at onee

Line graph.

5) Pie Chart :-

When it comes to statistical types of graphs and charts, the pie chart has a crucial place and meaning. It displays data and statistics in an easy-to-understand 'pie-slice' format and illustrates numerical proportion. The larger a slice is the bigger portion of the total quantity it represents.

When you want to create and represent the composition of something, it uses a pie chart. To show percentages or proportional data.



Pie Chart.

Data Science And Big Data Analytics Practical - 9

===

Name:- Rohini Devkar

Roll no:- 23272

Prn no:- 72030818G

Class :- TE-2 (COMPUTER)

===

Problem Statement:-

Data Visualization II

Use the inbuilt dataset 'titanic' as used in the above problem. Plot a box plot for distribution of age with respect to each gender along with the information about whether they survived or not. (Column names : 'sex' and 'age')

Write observations on the inference from the above statistics.

===

In [1]:

```
pip install seaborn
```

```
Requirement already satisfied: seaborn in c:\users\lenovo\anaconda3\lib\site-packages (0.11.2)
Requirement already satisfied: scipy>=1.0 in c:\users\lenovo\anaconda3\lib\site-packages (from seaborn) (1.7.1)
Requirement already satisfied: numpy>=1.15 in c:\users\lenovo\anaconda3\lib\site-packages (from seaborn) (1.20.3)
Requirement already satisfied: pandas>=0.23 in c:\users\lenovo\anaconda3\lib\site-packages (from seaborn) (1.3.4)
Requirement already satisfied: matplotlib>=2.2 in c:\users\lenovo\anaconda3\lib\site-packages (from seaborn) (3.4.3)
Requirement already satisfied: cycler>=0.10 in c:\users\lenovo\anaconda3\lib\site-packages (from matplotlib>=2.2->seabor
n) (0.10.0)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\lenovo\anaconda3\lib\site-packages (from matplotlib>=2.2-
>seaborn) (2.8.2)
Requirement already satisfied: pyparsing>=2.2.1 in c:\users\lenovo\anaconda3\lib\site-packages (from matplotlib>=2.2->sea
born) (3.0.4)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\lenovo\anaconda3\lib\site-packages (from matplotlib>=2.2->se
aborn) (1.3.1)
Requirement already satisfied: pillow>=6.2.0 in c:\users\lenovo\anaconda3\lib\site-packages (from matplotlib>=2.2->seabor
n) (8.4.0)
Requirement already satisfied: six in c:\users\lenovo\anaconda3\lib\site-packages (from cycler>=0.10->matplotlib>=2.2->se
aborn) (1.16.0)
Requirement already satisfied: pytz>=2017.3 in c:\users\lenovo\anaconda3\lib\site-packages (from pandas>=0.23->seaborn)
(2021.3)
Note: you may need to restart the kernel to use updated packages.
```

In [2]:

```
conda install seaborn
```

```
Collecting package metadata (current_repodata.json): ...working... done
Solving environment: ...working... done

# All requested packages already installed.


Note: you may need to restart the kernel to use updated packages.
```

In [3]:

```
import pandas as pd
import numpy as np

import matplotlib.pyplot as plt
import seaborn as sns

dataset = sns.load_dataset('titanic')

dataset.head()
```

Out[3]:

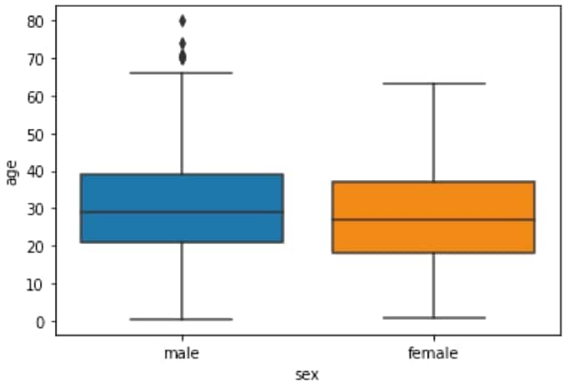
	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN	Southampton	no	False
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton	yes	True
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton	yes	False
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN	Southampton	no	True

In [4]:

```
sns.boxplot(x='sex', y='age', data=dataset)
```

Out[4]:

```
<AxesSubplot:xlabel='sex', ylabel='age'>
```



The first quartile starts at around 5 and ends at 22 which means that 25% of the passengers are aged between 5 and 25.

The second quartile starts at around 23 and ends at around 32 which means that 25% of the passengers are aged between 23 and 32.

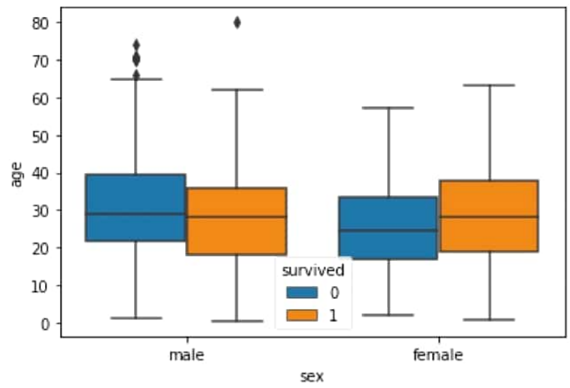
Similarly, the third quartile starts and ends between 34 and 42, hence 25% passengers are aged within this range and finally the fourth or last quartile starts at 43 and ends around 65.

In [5]:

```
sns.boxplot(x='sex', y='age', data=dataset, hue="survived")
```

Out[5]:

```
<AxesSubplot:xlabel='sex', ylabel='age'>
```



Now in addition to the information about the age of each gender, the distribution of the passengers who survived.

For instance, that among the male passengers, on average more younger people survived as compared to the older ones.

Similarly, that the variation among the age of female passengers who did not survive is much greater than the age of the surviving female passengers.

Page No.
Date

* Conclusion:-

Seaborn is an advanced data visualization library built on top of Matplotlib library. In this practical, we looked at how we can draw distributional and categorical plots using Seaborn library. We implemented the box plot for distribution of age with respect to each gender along with the information about whether they survived or not.