MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR

(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

Skill Based Mini Project Report

on

DATA VISUALIZATION USING PYTHON

Submitted By:

Aman Dwivedi

0901CS213D01

Faculty Mentor:

Submitted to:

Dr. RANJEET KUMAR SINGH

ASSISTANT PROFESSOR

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE

GWALIOR - 474005 (MP) est. 1957

Jan-Jun- 2022

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR

(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

CERTIFICATE

This is certified that **Aman Dwivedi** (0901CS213D01) has submitted the project report titled **Data Visualization using python** under the mentorship of **Dr. Ranjeet Kumar Singh**, in partial fulfilment of the requirement for the award of degree of Bachelor of Technology in Computer Science and Engineering from Madhav Institute of Technology and Science, Gwalior.

Dr. Ranjeet Kumar Singh
Faculty Mentor
Assistant Professor
Computer Science and Engineering

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR

(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

DECLARATION

I hereby declare that the work being presented in this project report, for the partial fulfilment of requirement for the award of the degree of Bachelor of Technology in Computer Science and Engineering at Madhav Institute of Technology & Science, Gwalior is an authenticated and originalrecord of my work under the mentorship of **Dr. Ranjeet Kumar Singh**, **Assistant professor**, **Computer Science and Engineering.**

I declare that I have not submitted the matter embodied in this report for the award of any degree ordiploma anywhere else.

Aman Dwivedi
0901CS213D01
2nd Year, IV sem
Computer Science and Engineering

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR

(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

ACKNOWLEDGEMENT

The full semester project has proved to be pivotal to my career. I am thankful to my institute, **Madhav Institute of Technology and Science** to allow me to continue my disciplinary/interdisciplinary project as a curriculum requirement, under the provisions of the Flexible Curriculum Scheme (based on the AICTE Model Curriculum 2018), approved by the Academic Council of the institute. I extend my gratitude to the Director of the institute, **Dr. R. K. Pandit** and Dean Academics, **Dr. Manjaree Pandit** for this.

I would sincerely like to thank my department, **Department of Computer Science and Engineering, for allowing** me to explore this project. I humbly thank **Dr. Manish Dixit**, Professor and Head, Department of Computer Science and Engineering, for his continued support during the course of this engagement, which eased the process and formalities involved.

I am sincerely thankful to my faculty mentors. I am grateful to the guidance of **Dr. Ranjeet Kumar Singh,** Assistant Professor, Computer Science and Engineering, for his continued support and guidance throughout the project. I am also very thankful to the faculty and staff of the department.

Aman Dwivedi
0901CS213D01
2nd Year, IV sem
Computer Science and Engineering

TABLE OF CONTENTS

TITLE	PAGE NO.
Abstract	6
Chapter 1: Introduction	7
Chapter 2: Requirements	8
2.1 Hardware Requirements 2.2 Software Requirements	
Chapter 3: Problem Statement	9
Chapter 4: Appendices	10-18
Chapter 5: Result	19
Chapter 6:Conclusion	19
References	20

ABSTRACT

This project is showing visualization of big datasets using python libraries. Data visualization involves presenting data in graphical or pictorial form which makes the information easy to under stand. It helps to explain facts and determine courses of action. It will benefit any field of study that requires innovative ways of presenting large, complex information. In this project, we are visualizing dataset with different graphs likeScatter plot, Histogram, jointPlot, barplot, Strip Plot Swarn Plot, Displot, Boxplot & Count Plot. The main objective of this project is to learn these visualization techniques so that it will be helpful to understand the data. We know the data visualization are widely used in many applications like Healthcare industries, Military, Finance Industries and many other as well.

INTRODUCTION

Data visualization is a graphical representation of any data or information.Visual elements such as charts, graphs and maps are the few data visualization tools that provide the viewers with an easy and accessible way of understanding the represented information. In this world governed by Big Data, data visualization enables you or decision-makers of any enterprise or industry to look into analytical reports and understand concepts that might otherwise be difficult to grasp.

there has been the need for displaying massive amounts of data in a way that is easily accessible and understandable. Organizations generate data every day. As a result, the amount of data available on the Webhas increased dramatically. It is difficult for users to visualize, explore, and use this enormous data. The ability to visualize data is crucial to scientific research. Today, computers can be used to process large amounts of data. Data visualization is concerned with the design, and application of computer generated graphical representation of the data. It provides effective data representation of data originating from different sources. This enables decision makers to see analytics in visual form and makes it easy for them to make sense of the data.

The main goal of the project was to understand & make sense of the data by visualizing data .To gain maximal benefit from learning you can try each graph plotting on your dataset. It might be used Creationof a dashboard to visualize patients' history can help an existing or a new doctor understand a patient's condition. In case of emergency, it could provide quicker according to disease.

SOFTWARE & HARDWARE REQUIREMENTS

Hardware Environment :

- Processor: x86 or x64
- RAM : 512 MB (minimum), 1 GB (recommended)
- Hard disc: up to 3 GB of free space may be required

Development Environment :

- Any web based IDE such as Google colab or Jupyter notebook.
- Visual Studio Code (optional text-editor)
- If you want to use Reporting or Business Intelligence controls, it is necessary to have one of the IDE – Visual Studio 2010+ in the machine.

PROBLEM STATEMENT

In data visualization, data is abstracted and summarized. Spatial variables such as position, size, and shape represent key elements in the data. A visualization system should perform a data reduction, transform and project the original dataset on a screen. It should visualize results in the form of charts graphs and present results in user friendly way.So, In industries we usually have large dataset and its not possible to tellin first sight that what algorithm will be applicable so that it will give good result. So ,In order to draw meaningful insights from the dataset or to understand dataset, we need to visualize the dataset. There are many plotting graphs, in which we can plot our dataset like Scatter plot, Histogram jointplot, barplot, strip plot, swarn plot, displot, boxplot & countplot through we can visualize the data.

APPENDICES

The following is the partial / subset of the code. Code of some module(s) have been wilfully supressed.

Import library

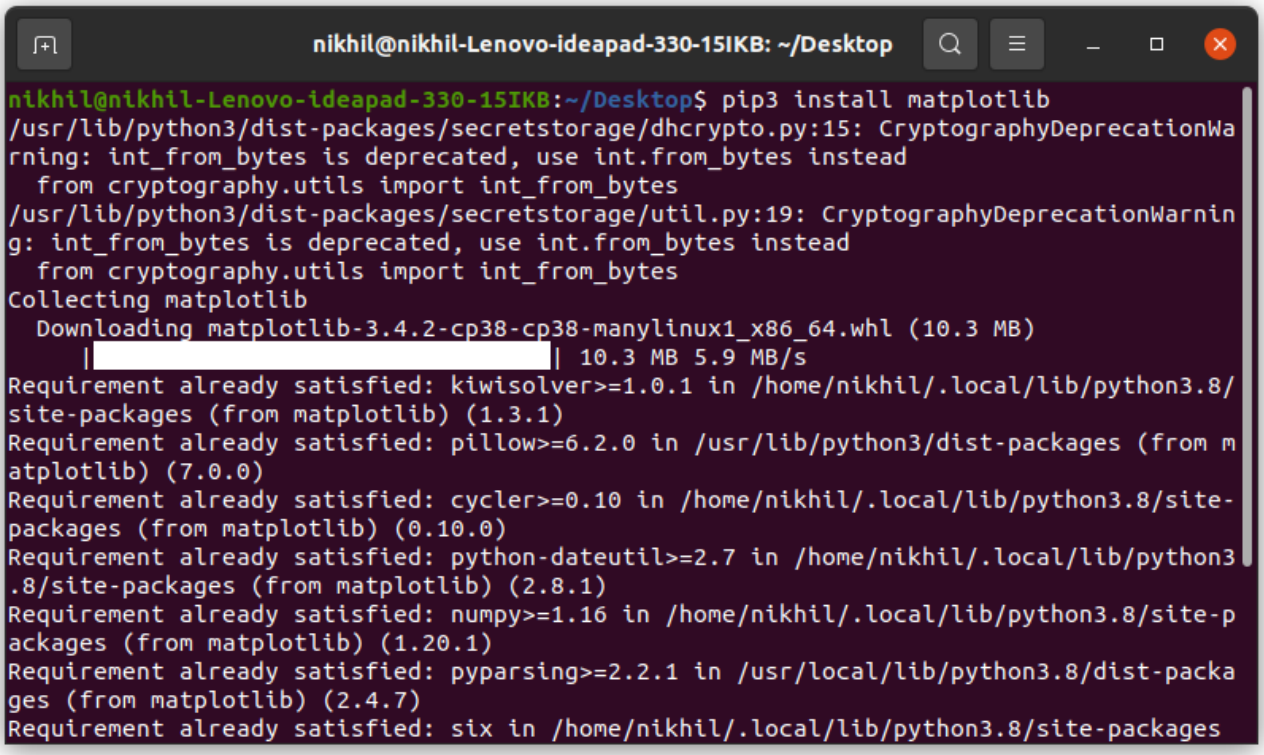
```
import pandas as pd


# reading the database
data = pd.read_csv("tips.csv")

# printing the top 10 rows
display(data.head(10))
```

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4
5	25.29	4.71	Male	No	Sun	Dinner	4
6	8.77	2.00	Male	No	Sun	Dinner	2
7	26.88	3.12	Male	No	Sun	Dinner	4
8	15.04	1.96	Male	No	Sun	Dinner	2
9	14.78	3.23	Male	No	Sun	Dinner	2

Matplotlib

Matplotlib is an easy-to-use, low-level data visualization library that is built on NumPy arrays. It consists of various plots like scatter plot, line plot, histogram, etc. Matplotlib provides a lot of flexibility.
To install this type the below command in the terminal.

```
pip install matplotlib
```

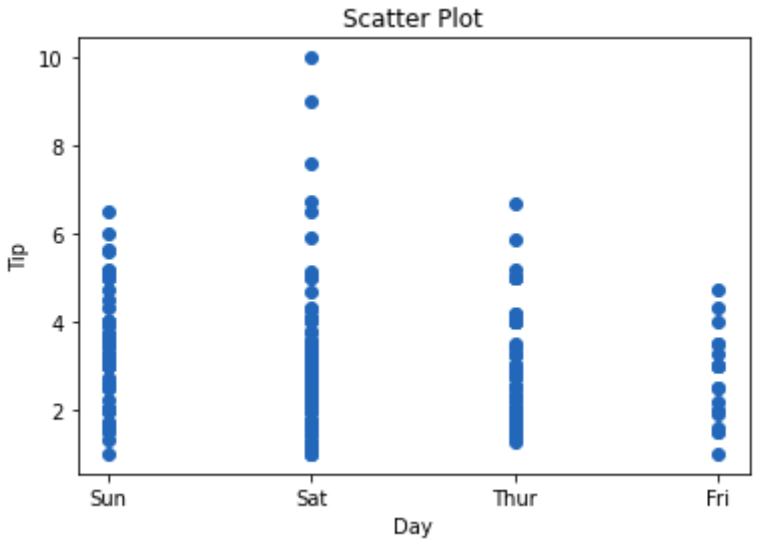


Scatter Plot

Scatter plots are used to observe relationships between variables and uses dots to represent the relationship between them. The **scatter()** method in the matplotlib library is used to draw a scatter plot.

```
import pandas as pd
import matplotlib.pyplot as plt


# reading the database
data = pd.read_csv("tips.csv")

# Scatter plot with day against tip
plt.scatter(data['day'], data['tip'])

# Adding Title to the Plot
plt.title("Scatter Plot")

# Setting the X and Y labels
plt.xlabel('Day')
plt.ylabel('Tip')

plt.show()
```

output

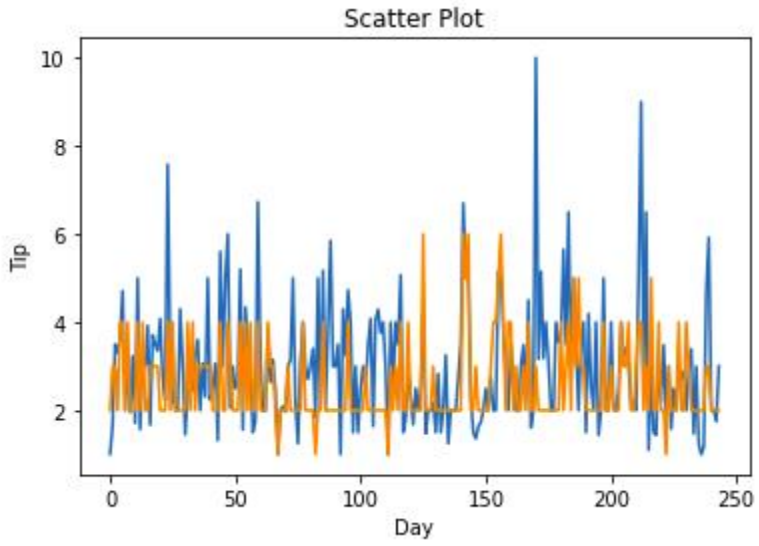


Line Chart

Line Chart is used to represent a relationship between two data X and Y on a different axis. It is plotted using the **plot()** function.

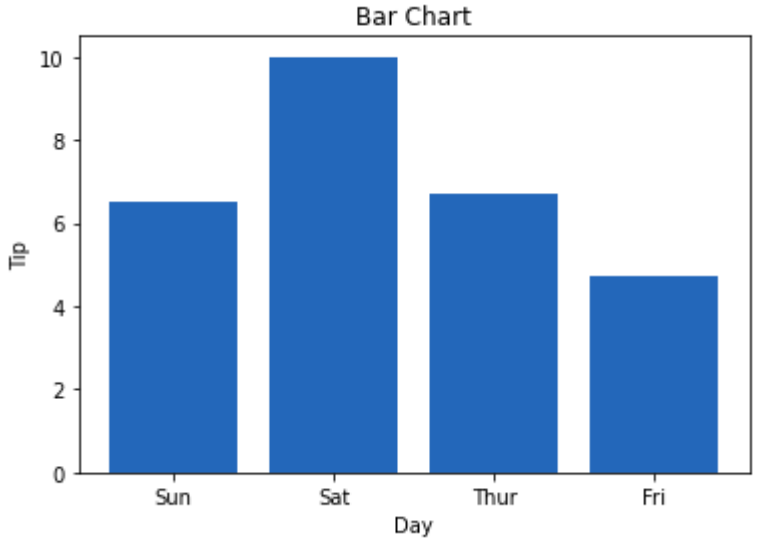
```
import pandas as pd
import matplotlib.pyplot as plt


# reading the database
data = pd.read_csv("tips.csv")

# Scatter plot with day against tip
plt.plot(data['tip'])
plt.plot(data['size'])

# Adding Title to the Plot
plt.title("Scatter Plot")

# Setting the X and Y labels
plt.xlabel('Day')
plt.ylabel('Tip')

plt.show()
```

OUTPUT:-



Bar Chart

A bar plot or bar chart is a graph that represents the category of data with rectangular bars with lengths and heights that is proportional to the values which they represent. It can be created using the **bar()** method.



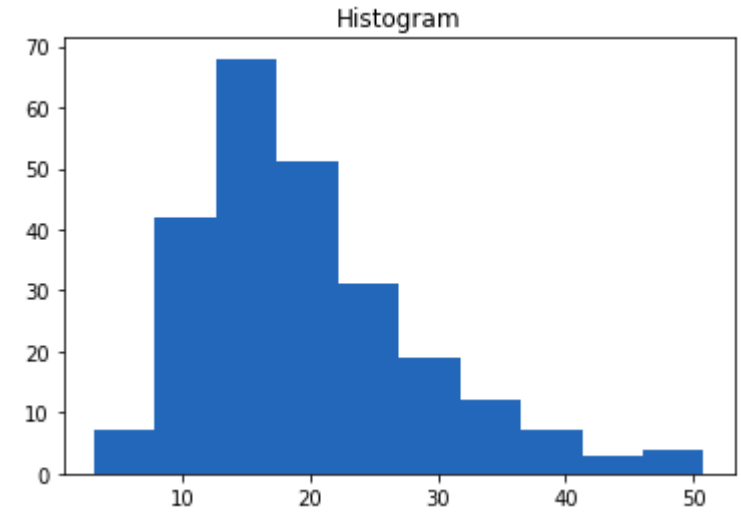
Histogram

A histogram is basically used to represent data in the form of some groups. It is a type of bar plot where the X-axis represents the bin ranges while the Y-axis gives information about frequency. The **hist()** function is used to compute and create a histogram.

import pandas as pd
import matplotlib.pyplot as plt

```
# reading the database
data = pd.read_csv("tips.csv")

# histogram of total_bills
plt.hist(data['total_bill'])

plt.title("Histogram")

# Adding the legends
plt.show()
```

OUTPUT:

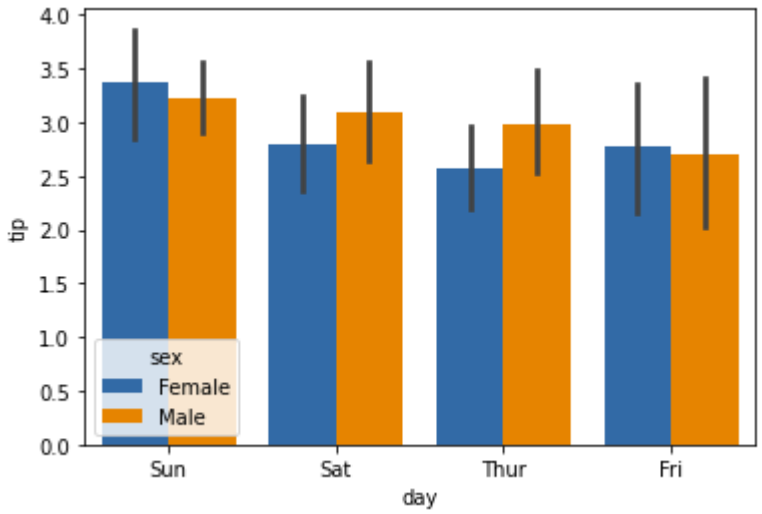


Bar Plot

Bar Plot in Seaborn can be created using the **barplot()** method.

```
# importing packages
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd


# reading the database
data = pd.read_csv("tips.csv")

sns.barplot(x='day',y='tip', data=data,
                        hue='sex')

plt.show()
```

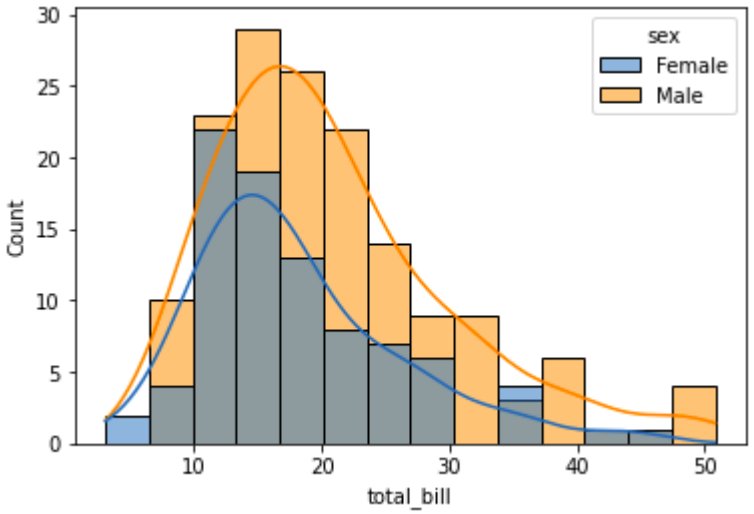
OUTPUT:



Histogram

The histogram in Seaborn can be plotted using the **histplot()** function.

```
# importing packages
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd


# reading the database
data = pd.read_csv("tips.csv")

sns.histplot(x='total_bill', data=data, kde=True, hue='sex')

plt.show()
```



RESULT

Data visualization are generally used for the graphical representation of data or information. Data are converted into visual form so that it can be analyzed to solve problems. This representation helps in identifying the shape of the data and to draw meaningful insights from it. Further, before applying any regression algorithm, you need to visualize the dataset to know, which algorithm will be applicable to the dataset. So ,by visualizing dataset ,it will help us to understand the dataset.

CONCLUSION

Data visualization is the process of representing data in a graphical or pictorial way in a clear and effective manner. It has emerged as a powerful and widely applicable tool for analyzing and interpreting large and complex data. It has become a quick, easy means of conveying concepts in a universal format. It must communicate complex ideas with clarity, accuracy, and efficiency. These benefits have allowed data visualization to be useful in many fields of study.

REFERENCES

1. https://www.ibm.com/topics/data-visualization
2. https://www.google.com/amp/s/www.geeksforgeeks.org/data-visualization-with-python/amp/