Department of Computer Engineering

**Academic Year:** 2022-2023 **Semester:** VIII

**Subject:** Applied Data Science **Class / Division:** BE/CMPN/B

**Name:** Domingos J Pinto **Roll Number:** 007

**Date: Seat Number:** BE/CMPN/B

**Experiment No.: 4**

**Explore data visualization techniques.**

**Aim :** Explore data visualization techniques.**.**

**I-THEORY**

## What are data visualizations?

Simply put, [data visualizations](https://www.sisense.com/glossary/data-visualization/) allow humans to explore data in many different ways and see patterns and insights that would not be possible when looking at the raw form. Humans crave narrative and visualizations allow us to pull a story out of our stores of data.

The phrase “A picture is worth a thousand words” is expressly true when turning huge piles of data into images a viewer can actually understand and derive meaning from. Children’s storybooks contain lots of images, but very few words. As kids, we don’t know many words, but the visuals allow us to easily understand the story.

In our modern digital world, we have [huge amounts of data](https://www.sisense.com/blog/delivering-more-impactful-insights-from-your-cloud-data/) all around us. Data scientists and ML engineers get most of the data they deal with data in a structured or unstructured data format, however, it’s difficult for humans to understand and analyze this. Data visualizations (or graphical representations of data) are vital for understanding the data. They help users explore data through visual elements like charts, graphs, plots, maps, and other visualizations.

## Different types of exploratory data analysis

In every dataset, we have many **variables**(also called features, input-variables, or independent-variables) and **target/output variables**(also known as labels, dependent-variables, classes, or class-labels). The data scientist’s job is to completely understand each feature individually and the relationship between different features. The goal is to get ready the dataset for ML algorithms implementation.

We have three methods for exploratory data analysis:

### Univariate analysis

In the univariate analysis, each variable is analyzed individually. It will get us to the complete statistical data for each feature. There are a variety of data visualization techniques for univariate analysis, including **Box Plot, Histogram, PDF, CDF**

### Bivariate analysis

Bivariate analysis is performed to find the relationship between each feature with the target variable.  Data visualization techniques for bivariate analysis are **Scatter Plot**and **Heatmap**

### Multivariate Analysis

As the name signifies, multivariate analysis is performed to understand the relationship between different features of the dataset. One of the main multivariate analysis data visualization techniques is the **Pair Plot**.

## Data Visualization in Python

There are a wide array of libraries you can use to create Python data visualizations, including **Matplotlib, seaborn, Plotly**, and others. A Python data visualization helps a user understand data in a variety of ways: Distribution, mean, median, outlier, skewness, correlation, and spread measurements.

### Bar Plot

A bar plot is a plot that presents categorical data with rectangular bars. The length or height of bars is proportional to the frequency of the category. We can count the values of various categories using bar plots.

### Pie Chart

Pie Chart is a circular chart that uses pie slices to show the relative size of data. The arc length of each pie slice is proportional to the quantity it represents. It works beautifully on categorical values. There are different variants of pie charts available.

### Box-plot

Box-plot gives us a five-number summary of any variable: the minimum, maximum, the sample median, the first and third quartile. Box-plot helps in measuring two observations:

1. Skewness of distribution  
2. Outliers (Outliers comes outside the box-plot)

### Histogram and PDF

A histogram is a graphical representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable. Histogram basically represents the number of points that exist for each bin(range of values). PDF is a **Probability Density Function** which is basically smoothening of the histogram.

### Scatter Plots

A scatter plot is a plot that shows the relationship between two variables of a data set.

### Heat Map

A heatmap is a graphical representation of data in which data values are represented as colors. It uses color in order to communicate the correlation between two variables. Values are between -1 to 1. 1 denotes perfect positive correlation. 0 means no correlation and -1 means the highest negative correlation.

### Line chart

The line chart represents a series of data points connected by a straight line. It is generally used to visualize data that changes over time.

**III Implementation-**

**Dataset :** Tipping data

**Description :** It is a dataset of tips received by a waiter over the period of few months The dataset contains 244 records. There are 7 variables in this dataset such as tip, bill, day, time etc

**Code:**

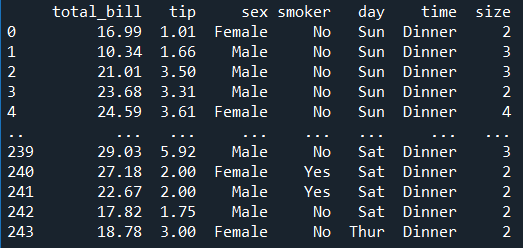
#Load the required libraries and data

import seaborn as sns

import plotly.express as px

df = px.data.tips()

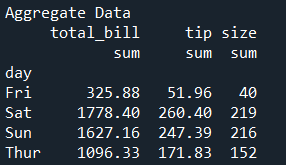
print(df)



print("\n\nAggregate Data")

aggrdf = df.groupby('day').agg({'total\_bill':['sum'],'tip':['sum'],'size':['sum']})

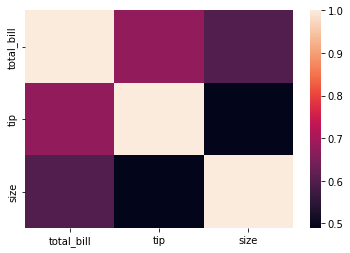
print(aggrdf)



#Data Visualization

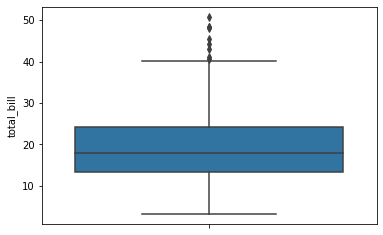
#1) Correlation plot

sns.heatmap(df.corr())



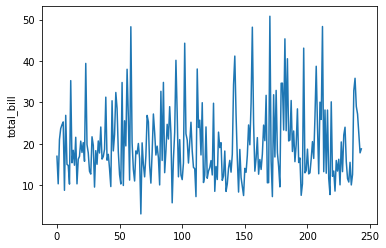
#2) Boxplot

sns.boxplot(y=df['total\_bill'],data=df)



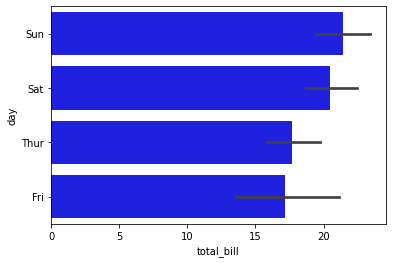
#3) Lineplot

sns.lineplot(data=df['total\_bill'])



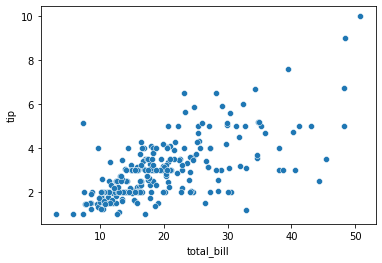
#4) Bar Graph

sns.barplot(y=df['day'],x=df['total\_bill'],data=df,color='b')



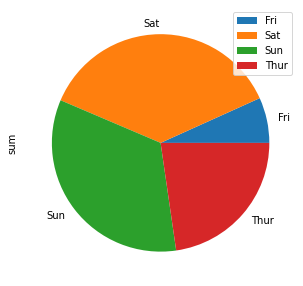
#5) Scatterplot

sns.scatterplot(x=df['total\_bill'], y=df["tip"])



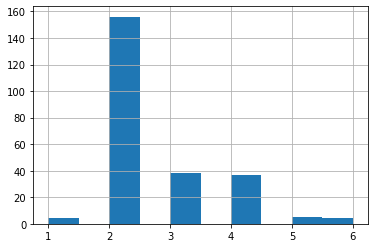
#6) Pie Diagram

plot = aggrdf.plot.pie(y='total\_bill', figsize=(5, 5))



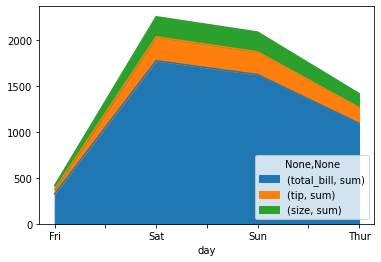
#7) Histogram

hist = df['size'].hist()



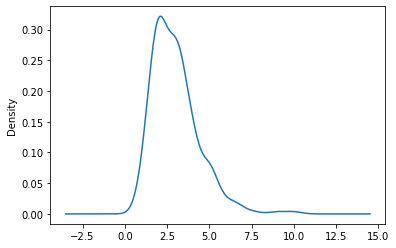
#8) Area Plot

ax = aggrdf.plot.area()



#9) Density Plot

ad = df['tip'].plot.kde()



#10) Multiple Linear Regression

g = sns.lmplot(data=df,x="total\_bill", y="tip", hue="day")

