** Bansilal Ramnath Agarwal Charitable Trust’s**

**Vishwakarma Institute of Information Technology, Pune-48**

**(An Autonomous Institute affiliated to Savitribai Phule Pune University)**

**Department of Computer Science and Engineering (Artificial Intelligence)**

**LAB SUBMISSION**

**Data Science and Machine Learning**

**CAUA22201**

*Submitted by:*

*Avdhoot Nilesh Pimparkar,*

*PRN: 22210178,*

*Roll Number: 281054*

*Second Year*

*Semester II Academic Year 2023-24*

**Assignment: 2**

Aim: To perform the following operations using R/Python on the data sets:

a) Compute and display summary statistics for each feature available in the dataset. (e.g.

minimum value, maximum value, mean, range, standard deviation, variance and

percentiles)

b) Data Visualization-Create a histogram for each feature in the dataset to illustrate the

feature distributions.

c) Data cleaning, Data integration, Data transformation, Data model building (e.g.

Classification)

Theory:

Summary statistics provide a concise overview of the key characteristics of the dataset. These statistics include measures such as minimum and maximum values, mean, range, standard deviation, variance, and percentiles. The minimum and maximum values represent the smallest and largest values in the dataset, respectively. The mean is the average value of the dataset, providing a measure of central tendency. The range is the difference between the maximum and minimum values, indicating the spread of the data. Standard deviation and variance quantify the dispersion of data points around the mean, with higher values indicating greater variability. Percentiles divide the dataset into 100 equal parts, allowing us to understand the distribution of data across different percentiles.

Data visualization through histograms offers insights into the distribution of individual features within the dataset. Histograms represent the frequency distribution of numerical data by dividing the data into bins or intervals and plotting the frequency of observations within each bin. By visualizing the distribution of features, we can identify patterns, skewness, and outliers within the data. Histograms provide a visual representation of the data's shape, central tendency, and spread, allowing for a better understanding of its underlying characteristics.

Data cleaning, integration, transformation, and model building are essential steps in the data analysis process. Data cleaning involves identifying and handling missing values, outliers, and inconsistencies to ensure data quality and integrity. Data integration combines multiple datasets into a single, unified dataset, facilitating comprehensive analysis. Data transformation includes preprocessing steps such as feature scaling, encoding categorical variables, and creating new features to prepare the data for modelling. Finally, data model building involves selecting an appropriate machine learning algorithm, splitting the data into training and testing sets, training the model on the training data, and evaluating its performance on the testing data. These steps collectively enable the development of robust and accurate predictive models for classification tasks, aiding in decision-making and problem-solving in various domains.

Results:

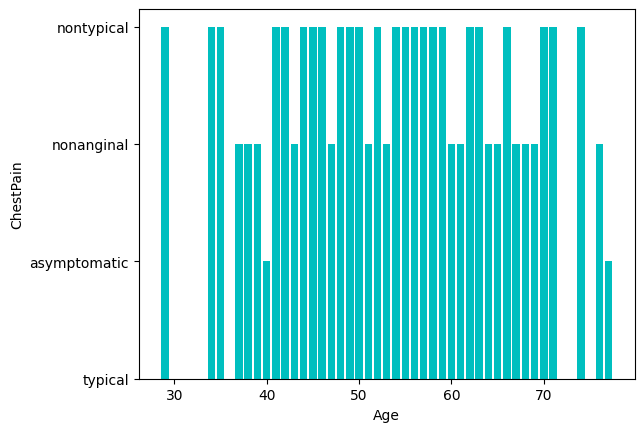
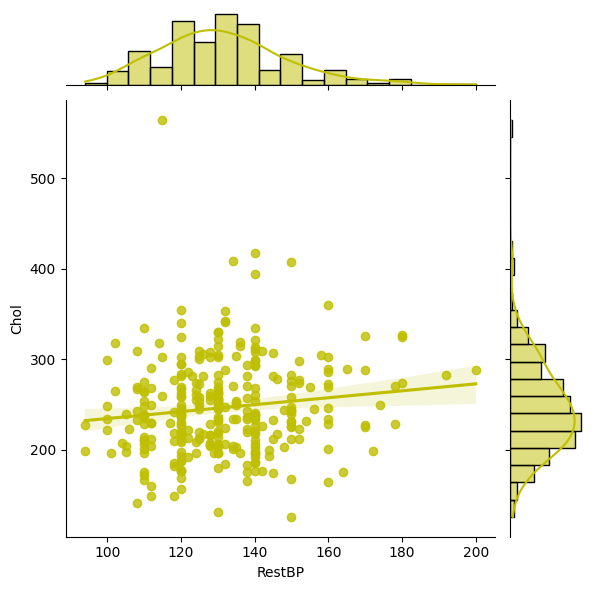
 

Fig. Bar Chart Fig. Jointplot

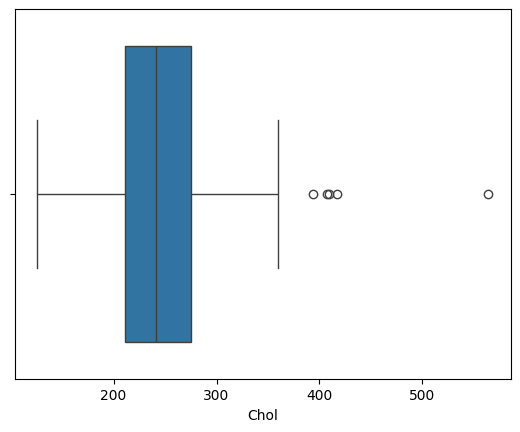
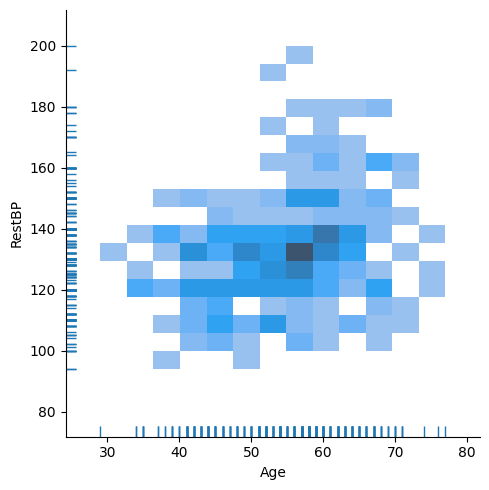
 

Fig. Boxplot Fig. Displot

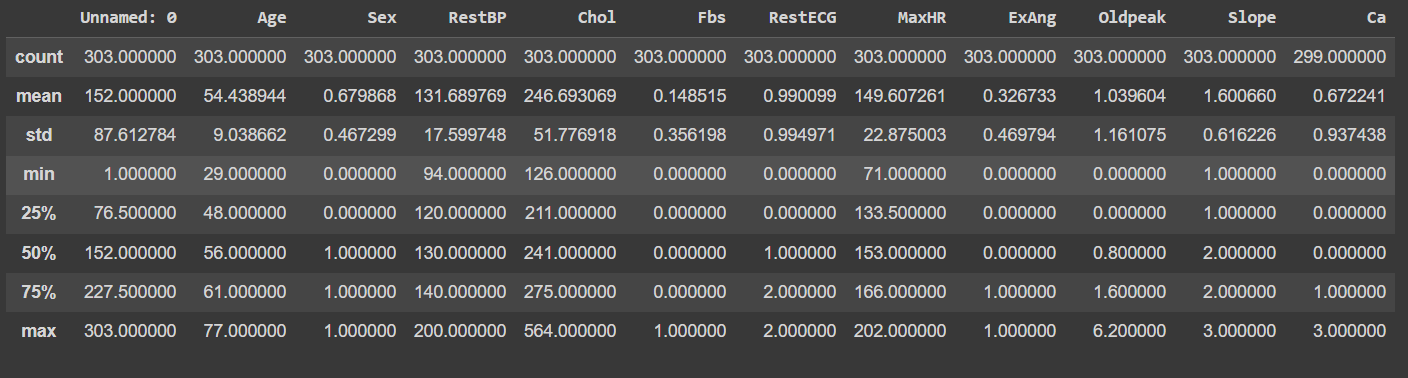


Fig. Brief Description about Dataset

Conclusion:

In this assignment, we were able to analyse the dataset and visualize it by using various charts and plots.