International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 /doi:10.14445/22315381/IJETT-V70I1P203

Volume 70 Issue 1, 24-34, January, 2022

Role of Exploratory Analytics and Visualization in Heart Disease Prediction

Lijetha.C. Jaffrin[1], Dr.J. Visumathi[2], Dr.G.Umarani Srikanth[3]

[1]*Assistant Professor, Department of IT, Veltech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Chennai, India*
[2]*Professor, Department of CSE, Veltech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Chennai, India*
[3]*Professor, Department of CSE, Bharath Institute of Higher Education and Research Chennai, India*

[1]lijethacjaffrin@veltech.edu.in, [2]drvisumathij@veltech.edu.in, [3]umarani.cse@bharathuniv.ac.in

Abstract *- Data Analysis is carried out to discover useful knowledge from the dataset and to drive quick and better decisions. It is also used to increase the efficiency of the work. Exploratory Data analysis is the first phase in Data Analysis. It is a method to understand the data and summarize the main features in the dataset by analyzing the data. It is also used for the visual representation of data. Visualization includes line plot, subplot, pair plot, violin plot, joint plot, swarm plot, Histograms, Box plot, Scatter plot. In this paper, Exploratory Data Analysis is done using python and implemented in Spyder IDE. Univariate analysis, bivariate analysis, multivariate analysis, and dimensionality reduction have been done on variables in the heart dataset. Different types of graphs have been plotted using the python Seaborn library to analyze the heart dataset. The primary objective is to get a more explained sight for which sort of traits might be a more critical sign of approaching heart disease.*

Keywords - *Exploratory Data Analysis, Visualization, Spyder IDE, python, Seaborn*

I. INTRODUCTION

Data analysts apply exploratory data analysis techniques to explore, evaluate, and summarize the main characteristics of datasets, frequently consuming data visualization approaches. Exploratory Data Analysis is a procedure of exploring data and mining insights or key features of the data. It is a way of getting an overview of the quality and nature of the information available before studying it in more detail. The prime motive of EDA is to increase perception into a data set, uncover underlying arrangements, extract important variables, handle missing values of the dataset, manage the outliers, eliminate identical data, encode the categorical variables, normalize and scale. It is used for showing what the data reveals before the task of modeling. It is difficult to go through the whole spreadsheet and look at the numbers in each column. It may be overwhelming to develop perceptions by observing the plain numbers. Exploratory data analysis techniques have been created as a utility in this situation. EDA centers on observing the characteristics of a dataset before deciding what to do with that dataset. This enables us to gain in-depth knowledge of the variables in datasets and their relationships.

In this paper, the heart dataset has been considered. Exploratory data analysis has been carried out using Seaborn libraries of python and implemented in Spyder IDE. Different types of plots such as categorical plots, distribution plots, matrix plots, grid plots, Regression plots, advanced plots have been plotted with the heart dataset. Exploratory data analysis has been done on the heart dataset to summarize the main characteristics of the dataset, find associations among the features to better understand patterns in data, detect abnormal events.

II. LITERATURE SURVEY

Heart disease is widely regarded as the most lethal disease in human survival. Physical bodily weakness, inappropriate breathing, swollen feet are the major signs of heart disease. The approaches are necessary for detecting difficult heart illnesses that carry a significant risk of affecting human life. Early detection of heart disease is critical for protecting the heart from serious hazards and reducing heart-related disorders. Furthermore, due to human interaction, diagnosis is time-consuming and costly. Invading approaches are used to identify cardiac disorders based on a patient's medical history. Exploratory Data analysis is used to scrutinize the data, visualize and summarize the main features in the dataset. From the data visualization using various plots, the subset of important variables can be selected from the dataset to forecast the occurrence of heart disease.

Kabitaet. al [1] have done exploratory analytics on data where data elaboration was completed in row and column format. Data analysis has been done using python, which is an object-oriented, obvious, collaborative programming language and non-proprietary with built-in libraries like pandas, Matplotlib, seaborn, etc. Different charts and parameters have been used to examine Amazon survey data sets-comprising assessments of electronic data items.

Tejas et al. [2] enlightened the process of dimensionality reduction to achieve EDA. This paper sketches practices of EDA as well as accepted methodologies to attain dimensionality reduction. Mostly, two methods are described by the pictorial depiction of the data, namely Principal Component Analysis (PCA) and t- Distributed Stochastic Neighbor Embedding (t-SNE). The authors also explained the way by which a high-dimensional data set could be envisioned into lower dimensions, recollecting the implicit structure of the data. The merits and downsides of each step related to this system were discussed by comparing the visuals provided in two ways.

John et al. [3] familiarize the principal heuristics and computational tools of Exploratory Data Analysis, divergences it with Confirmatory Data Analysis and exploratory measurements in broad-spectrum. EDA procedures are demonstrated with emotional data. Variations in statistical training are suggested to integrate these tools.

Aindrila et al. [4] performed deep exploration of an industrial dataset and identified added exploratory requirements from huge datasets. A wide-ranging study of the latest innovations in the growing arena of exploratory data analytics is presented. 50 academic and non-academic pictorial data exploration tools based on six ultimate stages of exploratory data analytics were investigated. The authors also looked at the extent of fulfillment of supplemental requirements for using data exploration tools to evaluate large datasets. Lastly, they identified and presented a collection of research prospects in pictorial exploratory data analysis.

R. IndraKumari et al. [5] have taken the risk aspects of heart disease, projected using K-means algorithm, and performed exploration using widely accessible heart disease data. The dataset consists of 209 records with 8 attributes such as age, chest pain type, blood pressure, blood glucose level, ECG in rest, heart rate, and four types of chest pain. To foretell the occurrence of heart disease, the K-means clustering algorithm was deployed along with a data analytics and visualization tool. The author deliberates pre-processing approaches, classifier performances, and assessment metrics. In the outcome section, the pictured data indicates that the forecast is precise.

III. MOTIVATION AND JUSTIFICATION OF WORK

The motivation of this work is to support the review of data prior to making any assumptions. It can assist in the discovery of obvious errors, as well as a better understanding of data designs, the discovery of outliers or infrequent events, and the discovery of fascinating relationships among features. It also motivates to double-check for missing data and other errors to gain a comprehensive understanding of the data set and its underlying structure.

This work is justified by the use of Exploratory Data Analysis on the heart dataset. Data scientists use data visualization techniques to evaluate and investigate data sets and describe their primary properties. Exploratory Data Analysis extracts averages, mean, minimum, and maximum values, among other things, to gain a better understanding of variables. It also identifies data errors, outliers, and missing values and visualizes data in graphs like box plots, scatters plots, and histograms to find trends.

IV. CONTRIBUTIONS

The main contributions in this paper were:

1. As an initial input, the system used a disease dataset from the Kaggle website. Because they contain a lot of information about the patients' health care and general statistics, heart disease datasets are used for analysis.
2. In this work, univariate analysis, bivariate analysis, multivariate analysis, and dimensionality reduction were performed on variables in the heart dataset. A selection of essential variables was chosen from the original variables in the dataset using the dimensionality reduction technique.

V. OUTLINE OF THE PAPER

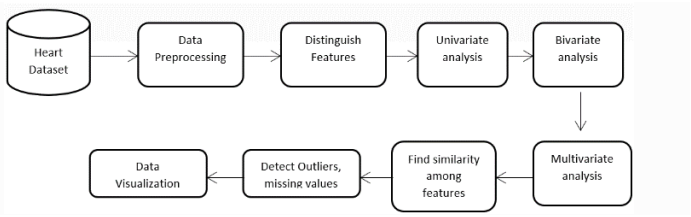

Fig1: Overview of System Methodology

VI. ORGANIZATION OF THE PAPER

The paper was systematized as below, Sect. 1 represents the introduction, Sect. 1.1 represents the literature survey, while Sect. 2 proposes the system methodology, Sect. 3 represents the experimental design, Sect. 4 represents results, and the final Sect. 5 proposes a conclusion with references.

VII. SYSTEM METHODOLOGY

Exploratory Data Analysis was a method of data inquiry that was used for handling missing values [10], identifying, removing outliers, selecting important variables, and deriving the relationships among the variables. It produces a graphical summary of the dataset as output [14],[15].

Dataset for heart disease was taken in csv format. This heart dataset was taken from the Kaggle UCI repository [6]. It consists of 16 variables, including one target variable, 'TenYearCHD'. With the help of EDA, graphical summaries are produced as output. The first phase in EDA is data description. The data Description phase describes the shape of the dataset, shows the data type of a particular variable, summarizes the dataset, frequency distribution of a variable. The second phase in EDA is visualizing the various variables in the dataset using the seaborn python library.

The various techniques of EDA include univariate analysis, bivariate analysis, multivariate analysis, dimensionality reduction. Univariate analysis [7],[11] analyzes single variables in the dataset such as male, age, TenYearCHD, current smokers and displays the frequency

distribution of that variable. Bivariate analysis [11] analyzes any two variables in the dataset such as male and TenYearCHD, male and heart rate achieved, male and diaBP, prevalent Hypertension and diaBP, etc., and displays the relationship between the two variables. Multivariate analysis analyzes more than two variables in the dataset and maps the relationships between different variables in the dataset. Dimensionality Reduction allows us to understand the various fields in the dataset, select important variables, and allow processing of reduced volume of data. Pair plot is a dimensionality reduction technique that analyzes heart rate achieved, Prevalent Hypertension, and Ten-year CHD.

VIII. EXPERIMENTAL DESIGN

Following are the steps for conducting exploratory data analysis.

A. Data Description:

The Shape of the dataset can be checked by using the panda's df. shape() function in python.The shape property returns a tuple representing the dimensionality of the DataFrame. The format of shape would be (rows, columns) as shown in TableI.

TABLE I
THE SHAPE OF THE HEART DATASET

Shape of	Rows	Columns
Dataset	4239	16

Now it is observed that the dataset contains 4239 instances and 16 variables. The dataset can be previewed by using df. head().The pandas head() function is used to get the first n rows, as shown in Fig. 2.

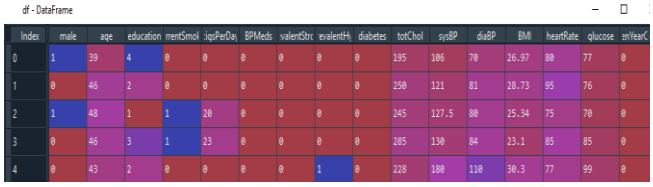

Fig2: Preview of the dataset

The summary of the dataset can be viewed by df.info () function of Pandas in python as shown in Fig. 3.

```
RangeIndex: 4239 entries, 0 to 4238
Data columns (total 16 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   male             4239 non-null   int64
 1   age              4239 non-null   int64
 2   education        4134 non-null   float64
 3   currentSmoker    4239 non-null   int64
 4   cigsPerDay       4210 non-null   float64
 5   BPMeds           4187 non-null   float64
 6   prevalentStroke  4239 non-null   int64
 7   prevalentHyp     4239 non-null   int64
 8   diabetes         4239 non-null   int64
 9   totChol          4189 non-null   float64
 10  sysBP            4239 non-null   float64
 11  diaBP            4239 non-null   float64
 12  BMI              4220 non-null   float64
 13  heartRate        4238 non-null   float64
 14  glucose          3851 non-null   float64
 15  TenYearCHD       4239 non-null   int64
```

Fig 3: Summary of the dataset

The data type of a particular column can be viewed by df. dtypesas are shown in Fig. 4.

```
male                int64
age                 int64
education         float64
currentSmoker       int64
cigsPerDay        float64
BPMeds            float64
prevalentStroke     int64
prevalentHyp        int64
diabetes            int64
totChol           float64
sysBP             float64
diaBP             float64
BMI               float64
heartRate         float64
glucose           float64
TenYearCHD          int64
dtype: object
```

Fig4: Data types of variables in the dataset

The Frequency distribution of TenYearCHD variable can be obtained by df['TenYearCHD'].value_counts() as shown in Fig. 5.

```
0    3595
1     644
Name: TenYearCHD, dtype: int64
```

Fig 5: Preview of the dataset

B. Data Exploration [18]

a) Variable Importance

From the dataset, input, output, datatype, and category of all features were identified [18].

b) Univariate Analysis

It looked only at a variable at a time in the data [18]. If there were continuous variables, properties such as central tendency and measure of dispersion were explored in the previous stage, and histograms, count, and boxplots were used to represent various statistical metrics. If there were categorical variables, a frequency table with the count and frequency of each category's data was created, and bar chat was used to represent them**.**

c) Bivariate Analysis

The relationship between two variables [18]was depicted using scatterplots, bar plot, Box plot, Bee swarm plot.

d) Multivariate analysis

Multivariate statistics were a subset of statistics that involved observing and analyzing multiple outcome variables at the same time [18]. Here the relationship between multiple variables was depicted using box plot, violin plot, Heat map, pair plot, facet grid.

C. Handling Missing values

Missing values in a dataset could lead to incorrect predictions, which could be risky, specifically in healthcare datasets [20]. As a result, null values in the dataset might be recognized and dealt with appropriately. Missing values were common in health databases when they were collected. These variables could be removed, however, depending on the

number of missing values, and this might result in a significant reduction in the dataset size. Here the missing values were identified using a heat map.

D. Outliers Detection

Outliers that did not fit with the rest of the data might be identified since they might skew the results of data analysis [18]. They might appear in medical datasets due to errors in data entry, instrument measurements, studies, and data processing. To find them, techniques like quantile plots and boxplots were used. They could be eliminated or imputed once they had been found. They had been discovered from the original data as the outlier values were attributable to error and were very small in number. Here they were identified from the dataset using Scatter plot, box plot, etc.

E. Visualization with Python

Data Visualization is the demonstration of data in an illustrative format. It supports people in recognizing the importance of data by tersing a massive amount of data in the easiest format and helps communicate information clearly and effectively. Thus, it becomes easier to grasp difficult concepts. Pandas, numpy, matplotlib, and seaborn libraries are used for data visualization and analysis.

a) Categorical Plots

1) Histogram:

The histogram is a type of graph which shows frequency distributions [13]. It is used for the visual representation of discrete data. It is used for summarizing numeric data by presenting the number of data points that come in various ranges of values. There will not be any spacing between the bars. It provides a visual representation of a huge amount of data that are difficult to understand in a table format or excel format. It is used to understand how the output of a process relates to targets. The first step in plotting a histogram is to decide how the process can be measured and what data should be collected. Once the histogram is developed, the data can be analyzed with regard to specifications. The matplotlib. pyplot.hist() function is used to calculate and generate a histogram of x. In this paper, the histogram is used to visualize the frequency distribution of the age variable, as shown in Fig. 6.

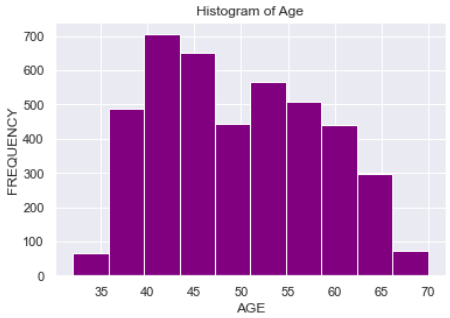

Fig6: Univariate analysis using Histogram

2) Bar Plot:

A bar plot is helpful in aggregating the categorical data based on some methods, and by default, it's the mean. A categorical variable is chosen on the x-axis, and a continuous variable is chosen on the y-axis. It creates a plot by taking a mean of the categorical column. Here two variables are taken for analysis, such as male and heart rate, male and Ten Year CHD, as shown in Fig. 7a and 7b.

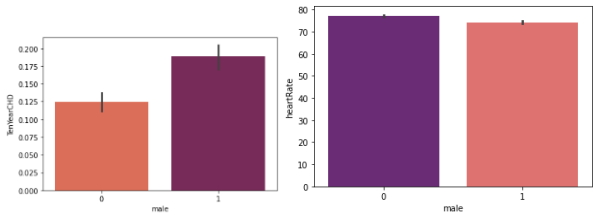

Fig7a: Bivariate analysis using bar plot **Fig7b: Bivariate analysis using bar plot**

3) Count plot

In Python, the Seaborn count plot () is a solution to create a data analysis graph. A count plot[9] is the type of bar graph for categorical data. It solely indicates the number of occurrences of a variable based on the type of category. It is one of the plots that Seaborn can create. Seaborn is a unit in Python that is built on top of matplotlib that is designed for statistical plotting. It makes statistical plots more gorgeous by incorporating beautiful default styles and color palettes. Here count plot is used to show the number of males suffering from heart illness [13], as shown in Fig. 8.

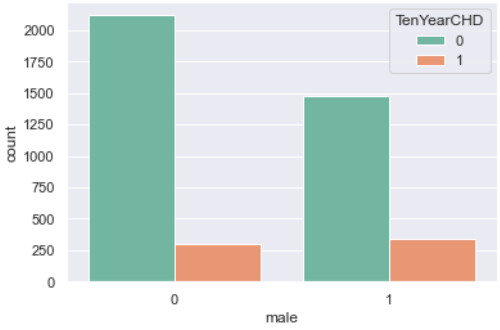

Fig8:Categorical and Univariate analysis using Count Plot

4) Violin Plot

Violin Plot is a function of the python Seaborn data visualization library. It shows the detailed distribution of different numeric variables and their probability density. It is also used to compare different variable categories. It shows several numeric data across one or more categorical variables. Here, violin plot is used to plot the numeric variables Systolic blood pressure, cholesterol level, Diastolic blood pressure against categorical variable Prevalent Hypertension variable as shown in Fig. 9a, 9b.

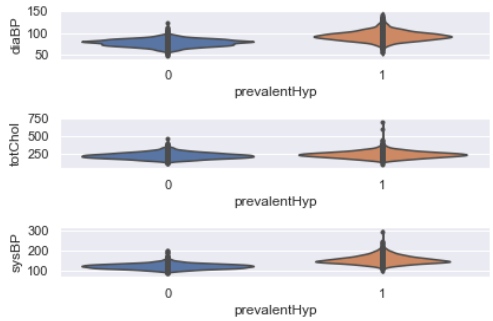

Fig 9a: Multivariate analysis using Violin plot

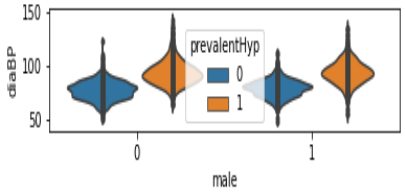

Fig9b: Multivariate analysis using Violin plot

5) Box Plot

Box Plot is the pictorial depiction of a group of numerical data through their quartiles. It is used to visualize distributions. It summarizes the dataset as Minimum, First Quartile, Median (Second Quartile), Third Quartile, Maximum. A box is drawn that connects the innermost two quartiles, and a horizontal line is drawn at a median position that falls within the box. Then another set of lines are drawn at some distance from the inner box referred to as "maximum" and "minimum" value of data, and then values falling outside maximum" and "minimum" values are referred to as outliers and plotted as distinct points. Bivariate analysis is done between variables male and diaBP, male and heartRate as shown in Fig. 10a, 10b. Multivariate Analysis is done between variables male, diaBP, TenYearCHD using Box plot to understand the presence of heart disease in males with different diastolic Blood Pressure levels as shown in Fig. 10c.

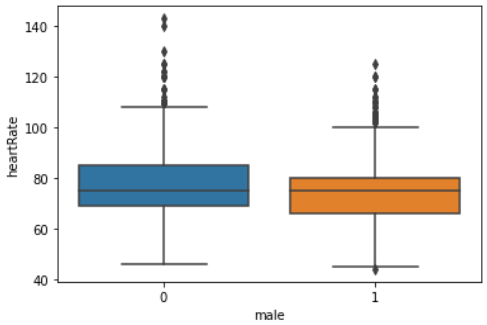

Fig 10a: Bivariate analysis using Box plot

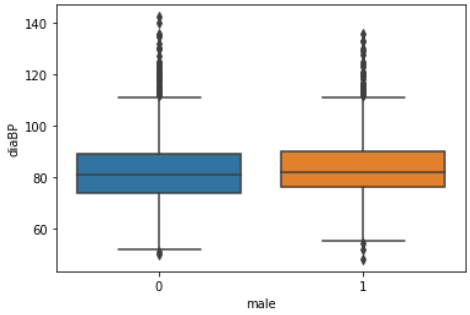

Fig10b: Bivariate analysis using Box plot

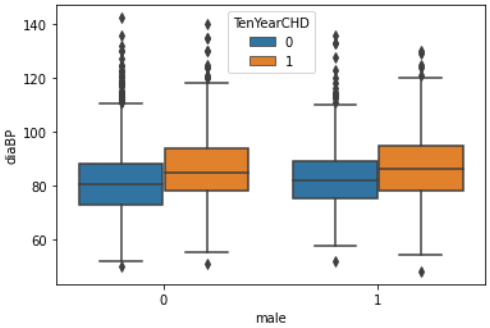

Fig10c: Multivariate analysis using Box plot

b) Advanced Plots

1) Swarm plot

Swarm Plots are also called bee-swarm plots. It spreads out the data points of the variable automatically and allows you to view each observation. It shows each point while heaping with similar values. The points are adjusted to avoid overlapping. It helps to represent a better representation of the distribution of values. The First swarm plot, as in Fig. 11a, shows the classification of heart defects in males and females. The second swarm plot, as in Fig. 11b, shows the classification of heart failures in males and females with Prevalent Hypertension types.

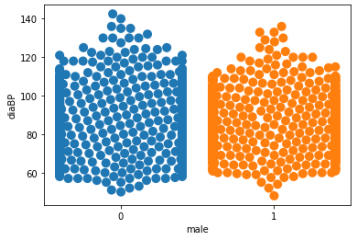

Fig11a: Bivariate analysis using BeeSwarm plot for classifying common kinds of heart defects in males and females

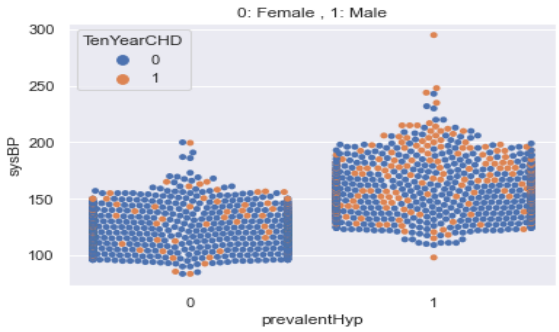

Fig 11b: Bivariate analysis using Bee Swarm plot to understand the heart failures in males and females with Prevalent Hypertension types

2) Strip Plot

A strip plot is ultimately a scatter plot where the *x*-axis signifies a categorical variable. It is used for applying a small random jitter value to each data point such that the parting between points comes to be perfect. Here advanced analysis was done between the variables age and prevalent Hyp to understand the prevalent hypertension types for all age groups, as shown in Fig. 12.

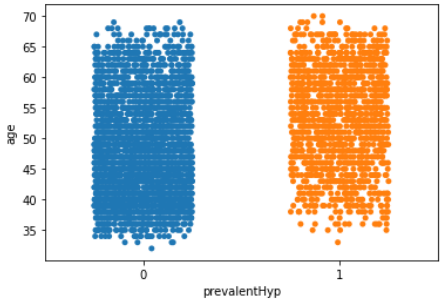

Fig12: Advanced analysis using strip plot

c) Regression Plot

1) Line plot

A line plot is a chart used to show the association between X and Y on a different axis. It depicts the association between continuous and categorical variables in a continuous format. It is easy to understand the instabilities and patterns in this data. Solid lines are used to keep the line plot concise and different colors are used for each variable to analyze information. It is used to project beyond this data. Here it is used to depict the correlation between age and Systolic blood pressure in the first line plot, as shown in Fig. 13a. The second line plot depicts the correlation between Systolic blood pressure and Age and the effect of Prevalent Hypertension, as shown in Fig. 13b.

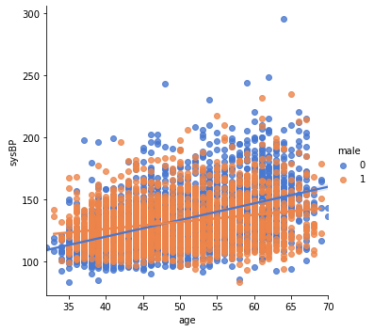

Fig13a: Regression analysis using a line plot

Correlation Coefficient: 0.3939041565320508

There is a positive correlation between age and Systolic blood pressure. Females have a greater chance of high blood pressure than males after the age of 45.

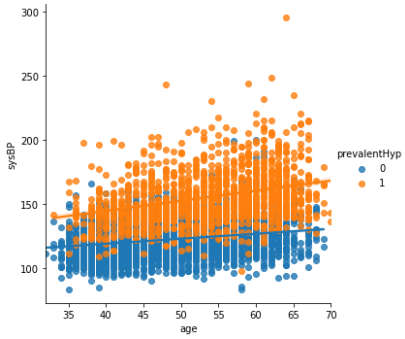

Fig13b: Regression analysis between Systolic blood pressure and Age using a line plot

Regression model between Systolic blood pressure and Age and effect of Prevalent Hypertension.

d) Matrix Plots

1) Heatmap

One of the built-in functions in the Python Seaborn library is the heatmap. Heatmap [9] is a data visualization method plotted in rectangular form as a matrix [19]. The aim of the heatmap is to provide a visual summary of data points through colored maps. It adds more visualization to the graph by representing different shades of the same color for each data point. The higher value in the plot indicates the darker shades. Any pattern in the data can be predicted by observing each color variation [12],[17].

Here darker shades in the last column depict that target variable 'TenYearCHD' in the heart dataset has a higher correlation with Systolic blood pressure, Prevalent Hypertension, Diastolic blood pressure, and Glucose level than other variables as shown in Fig. 14a.

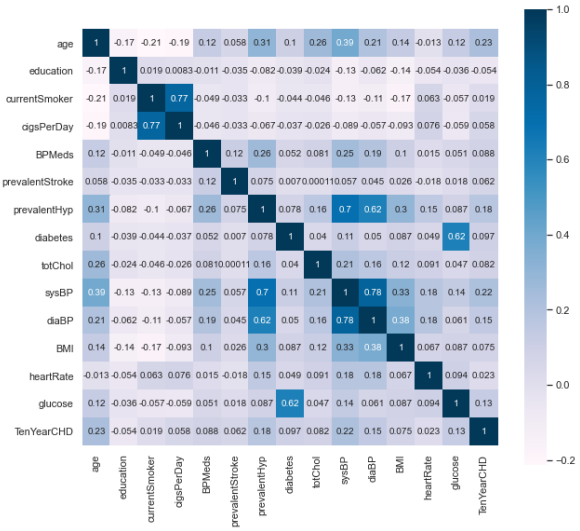

Fig14a: Multivariate Analysis using heat Map

The target variable 'TenYearCHD' has a higher correlation with Systolic blood pressure, Prevalent Hypertension, Diastolic blood pressure, and Glucose level than other variables, as shown in Fig. 14b.

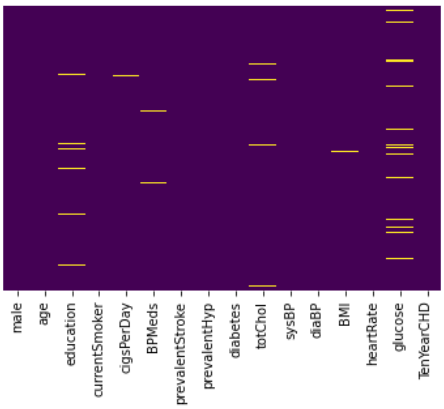

Fig14b: Missing value identification using heat Map

A heat map is also used to understand the missing value patterns. The yellow dashes in the plot represent missing values. Thus, it makes tasks effortless to identify the missing values.

2) Cluster Map

Cluster maps use Hierarchical clustering to form different clusters. It is used to group some features according to their similarity [16]. The x-label and y-label are the same, and they are grouped according to their similarity. The flow-chart-like structure at the left and top illustrate their degree of similarity. Here multiple variables are analyzed, and they are grouped in the form of clusters, as shown in Fig. 15.

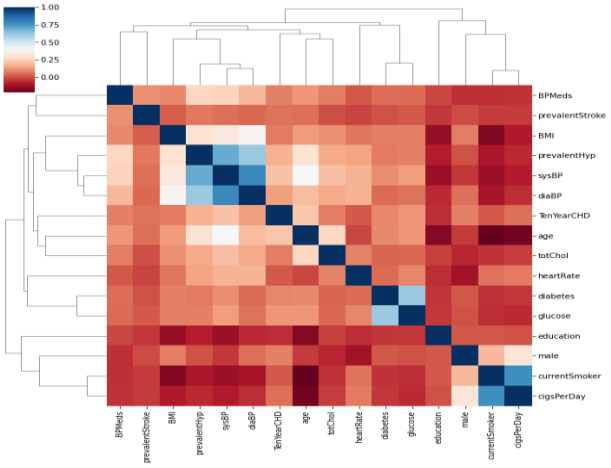

Fig15:Hierarchical clustering using cluster Map

e) Distribution Plots

1) Joint Plot

A joint plot is used to show univariate and bivariate graphs together. It merges three plots. One plot illustrates how the dependent variable(Y) fluctuates within the dependent variable(X). Another plot is positioned horizontally at the top, and it illustrates the dispersal of an independent variable(X). The third plot is positioned on the right boundary of the graph, and it shows the distribution of the dependent variable(Y). In this paper, a Joint plot is used to depict contours for different cholesterol levels for various age groups, different Systolic Blood Pressure, Diastolic Blood Pressure levels, and heart rate for various age groups, as shown in Fig. 16a, 16b, 16c, and 16d.

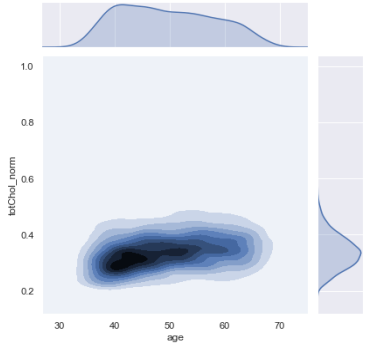

Fig 16a: Contours for cholesterol using joint plot

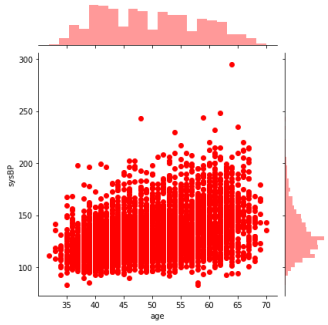

Fig16b: SysBP levels using joint plot

Joint Plot showing Contours for different cholesterol levels for various age groups

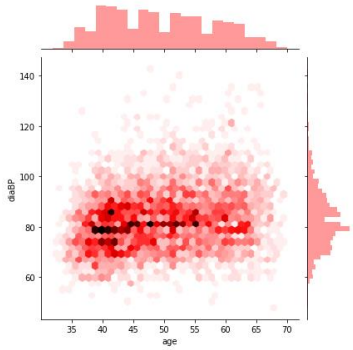

Fig16c: DiaBP levels using joint plot

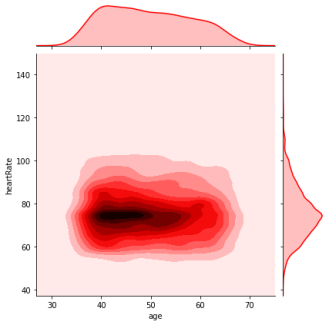

Fig16d: Heart Rate analysis using Joint plot

2) Scatter Plot

Scatter Plot is a cause analysis tool that is used to perceive the association between variables. It is also referred to as a scatter chart or scatter diagram [8]. It is a graph that uses points to signify the values of two dissimilar variables. Individual data point values are considered as the location of each dot on the horizontal and vertical axis. Based on the proximity of dots, it is used to divide data points into groups. It also identifies any unexpected gaps in the data and if there are any outlier points. If the variables are correlated, the dots come along a line. Here scatter plot predicts the presence or absence of heart disease by comparing cholesterol levels for different age groups, as shown in Fig. 17.

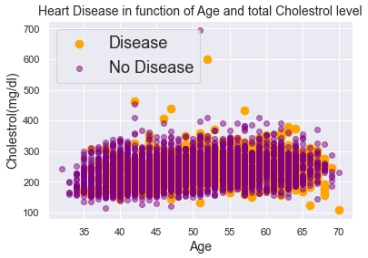

Fig17:Cause analysis using scatter plot

3) PairPlot

pair plot() is a function of python Seaborn Library. It is the easiest way to envision the relation between features, the pair plot method plots all pair associations at once. It considers all variables in the dataset and plots every variable column against every other variable column. It shows pairwise variable associations. Thus, it leads to the matrix of every column against every other column. Here, three variables in the heart dataset have been analyzed, and a pair plot has been shown for those three variables heart rate achieved, Prevalent Hypertension and the target variable 'TenYearCHD' as shown in Fig. 18.

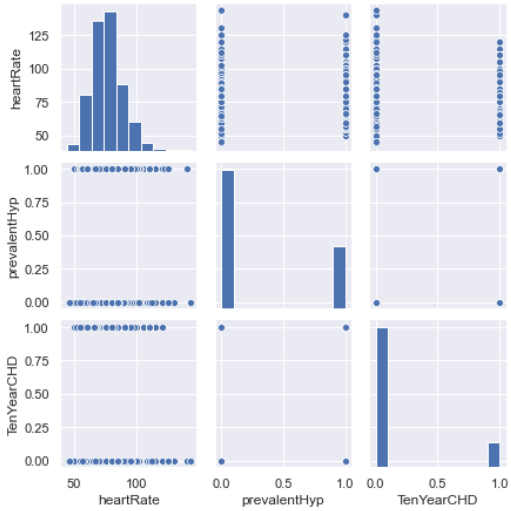

Fig18:Multivariate analysis using Pairplot

4) Rug plot

Rug plot considers any single column and depicts the values in the dataset as dash marks on an axis. It creates dashes all across the plot. Here the bin count is taken as the count of dashes. In this paper, the Rug plot is used to analyze the achieved heart rate of patients, as shown in Fig. 19a. It is also used here to analyze the glucose level of patients, as shown in Fig. 19b.

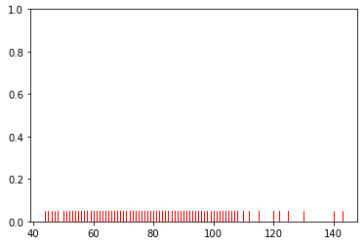

Fig19a: Rugplot for a heart rate of patients

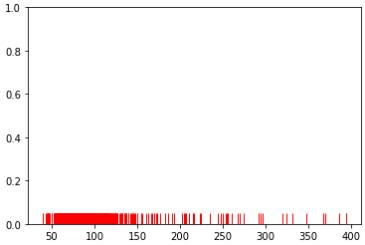

Fig19b: Rugplot for glucose level of patients

5) Sub-plot

Sub-plot is used to plot multiple plots in one figure, as shown in Fig. 20. It is used to compare different views of data side by side. These subplots might be insets, grids of plots, or other complex layouts. Subplots ()is a function in the pyplot module of the matplotlib library.

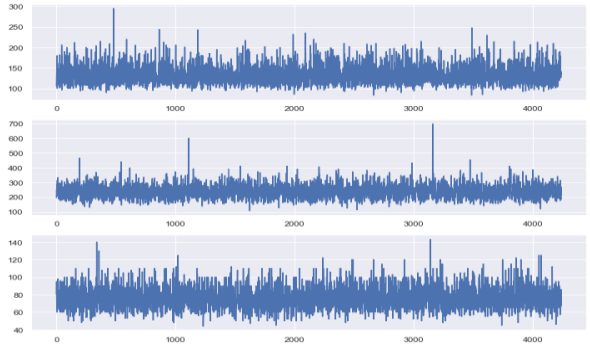

Fig20: Subplot

f) Grid plots

1) Facet Grid

Facet Grid depicts the distribution of one variable and the associations among multiple variables distinctly within subcategories of the dataset using multiple panels. It can be drawn with three dimensions: row, col, and hue. The first two dimensions are related to the resulting array of axes. Hue is the third dimension, where dissimilar levels are plotted with dissimilar colors. Here Facet Grid is plotted for the age distribution of males and females in all types of Prevalent Hypertension variables. So, four plots have been generated as, shown in Fig. 21.

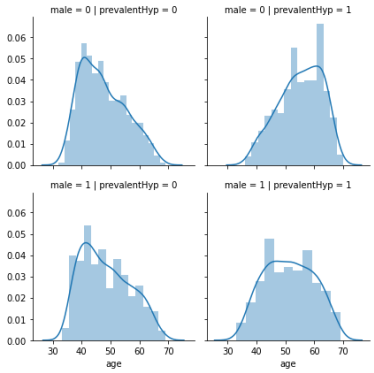

Fig21:Multivariate analysis using Facet Grid

IX.RESULTS AND DISCUSSION

Univariate analysis was done using Count Plot shows the number of males suffering from heart illness. Violin plot performs multivariate analysis on numeric variables such as Systolic blood pressure, cholesterol level, Diastolic blood pressure against categorical variable Prevalent Hypertension variable. Box plot performs bivariate analysis and shows the presence of heart disease in males with different diastolic Blood Pressure levels. BeeSwarm plot performs bivariate analysis and shows the heart failures in males and females with Prevalent Hypertension types. Strip plot performs analysis between the variables age and prevalentHyp and shows the severity of prevalent hypertension types for all age groups. Regression analysis was done using a line plot illustrates that there was a positive correlation between age and Systolic blood pressure and concludes that females have a greater chance of high blood pressure than males after the age of 45. Multivariate Analysis using Heat maps shows that the target feature'TenYearCHD' has a higher correlation with Systolic blood pressure, Prevalent Hypertension, Diastolic blood pressure, and Glucose level than other variables. Cluster map grouped the features according to their similarity. Joint depicted contours for different cholesterol levels for various age groups, the severity of different Systolic Blood Pressure, Diastolic Blood Pressure levels, and heart rate for various age groups. Scatter plots predicted the presence or absence of heart disease by comparing cholesterol levels for different age groups. Pair plot performs Multivariate analysis on three variables heart rate achieved, Prevalent Hypertension, and the target variable 'TenYearCHD'. Rug plot was used to analyze the achieved heart rate and the glucose level of patients. Facet Grid was plotted for the age distribution of males and females in all types of Prevalent Hypertension variables.

Table IIbelow shows bivariate analysis between variables sex and Heart Rate achieved. It reveals the average heart rate achieved by males and females. Heart rate is one of the important variables for heart disease prediction.

TABLE II
BIVARIATE ANALYSIS-SEX AND HEART RATE

Sex/ Heartrate	HeartRate Min (bpm)	HeartRate Max (bpm)
Males	42	98
Females	45	110

Table III. below shows a bivariate analysis between variables sex and diastolic blood pressure. It reveals that the diastolic blood pressure of males and females is higher than 90 mm Hg, which represents high blood pressure.

TABLE III
BIVARIATE ANALYSIS-SEX AND DIABP

Sex/ diaBP	diaBP(mm Hg)	Description
Males	110	hypertension
Females	112	hypertension

Multivariate Analysis using Heat maps shows that the Systolic blood pressure, Prevalent Hypertension, Diastolic blood pressure, heart rate, cholesterol, and Glucose level have

a higher correlation with the target feature. 'TenYearCHD' is shown in Table IV.

TABLE IV
MULTIVARIATE ANALYSIS

	Prevalent Hyp	totChol	sysBP	diaBP	heart Rate	Glucose	Ten Year CHD
Prevalent Hyp	1	0.16	0.7	0.62	0.15	0.087	0.18
totChol	0.16	1	0.21	0.16	0.091	0.047	0.082
sysBP	0.7	0.21	1	0.78	0.18	0.14	0.22
diaBP	0.62	0.16	0.78	1	0.18	0.061	0.15
heart Rate	0.15	0.091	0.18	0.18	1	0.094	0.023
glucose	0.087	0.047	0.14	0.061	0.094	1	0.13
Ten Year CHD	0.18	0.082	0.22	0.15	0.023	0.13	1

The features such as prevalentHyp, sysBP, diaBP, heartRate, totChol, glucose have been concluded as the most important subset of variables to forecast the existence of heart disease in patients. The correlation coefficients of these variables with the target variable 'TenYearCHD' are shown in Table V.

TABLE V
THE SUBSET OF VARIABLES HAVING A HIGH CORRELATION WITH THE TARGET VARIABLE

	TenYearCHD
prevalentHyp	0.18
totChol	0.082
sysBP	0.22
diaBP	0.15
heartRate	0.023
glucose	0.13

From the data visualization using various plots, it was shown that the subset of important variables [21] had been selected from the original variables in the dataset. Among 16 features from the dataset, the features such as prevalentHyp, sysBP, diaBP, heartRate, totChol, glucose have been selected as an important subset of variables to forecast the occurrence of heart disease of patients [22] of various age groups. The exploratory data resulted in a visually attractive and precise grouping experience.

X. CONCLUSION

In this paper, exploratory data analysis has been done on the heart dataset. Exploratory data analysis has been carried out using Seaborn libraries of python and implemented in Spyder IDE. Different types of plots such as categorical plots, distribution plots, matrix plots, grid plots, Regression plots, advanced plots have been plotted with the heart dataset. Different features from the heart dataset have been considered, plotted, and derived outcomes. Analysis has been done on the heart dataset to summarize the main characteristics of the dataset, find associations among the features to better understand patterns in data, detect abnormal events in the data. In this paper, univariate analysis, bivariate analysis, multivariate analysis, and dimensionality reduction have been done on variables in the heart dataset. With the help of the dimensionality reduction technique, a subset of important variables has been selected from the original variables in the dataset. Among 16 variables from the dataset, the variables such as prevalentHyp, sysBP, diaBP, heartRate, totChol, glucose have been selected as an important subset of variables to foresee the occurrence of heart disease in the patients of various age groups.

REFERENCES

[1] Kabita Sahoo, Abhaya Kumar Samal, Jitendra Pramanik, Subhendu Kumar Pani, Exploratory Data Analysis using Python, International Journal of Innovative Technology and Exploring Engineering(IJITEE) ISSN: 2278-3075, 8(12) (2019).

[2] Tejas Nanaware, Prashant Mahajan, Ravi Chandak, Pratik Deshpande, Prof. Mahendra Patil, Exploratory Data Analysis Using Dimension Reduction, Exploratory Data Analysis Using Dimension Reduction, IOSR Journal of Engineering (IOSRJEN) ISSN (e): 2250-3021, ISSN (p): 2278-8719, 2 81-84

[3] John T. Behrens, Principles and Procedures of Exploratory Data Analysis, Psychological Methods, 2(2) (1997)131-160.

[4] Aindrila Ghosh, Mona Nashaat, James Miller, Shaikh Quader, and Chad Marston, A Comprehensive Review of Tools for Exploratory Analysis of Tabular Industrial Datasets, Visual Informatics, 2(4) (2018) 235-253

[5] R. Indrakumari, T. Poongodi and Soumya Ranjan Jena Heart Disease Prediction using Exploratory Data Analysis, Procedia Computer Science, 173 (2020) 130–139.

[6] https://www.kaggle.com/ronitf/heart-disease-uci

[7] https://rstudio-pubs-static.s3.amazonaws.com/517828_8a3252fe1f604347845517e735a518fe.html

[8] Rony Chowdhury Ripan, Iqbal H. Sarker, Md. Hasan Furhad, Md Musfique Anwar, and Mohammed Moshiul Hoque, An Effective Heart Disease Prediction Model based on Machine Learning Techniques, Preprints 2020, 2020110744 (doi: 10.20944/preprints202011. 0744.v1)

[9] Shah Apeksha, Ahirrao Swati, Pandya Sharnil, Kotecha Ketan, Rathod Suresh, Smart Cardiac Framework for an Early Detection of Cardiac Arrest Condition and Risk, Frontiers in Public Health,https://doi.org/10.3389/fpubh.2021.762303

[10] Imran Chowdhury Dipto, Tanzila Islam, H M Mostafizur Rahman, Md Ashiqur Rahman, Comparison of Different Machine Learning Algorithms for the Prediction of Coronary Artery Disease, Journal of Data Analysis and Information Processing, 8 (2020) 41-68, DOI: 10.4236/jdaip.2020.82003

[11] Jinglin Peng, Weiyuan Wu, Brandon Lockhart, Song Bian, Jing Nathan Yan, Linghao Xu, Zhixuan Chi, Jeffrey M. Rzeszotarski, Jiannan Wang, DataPrep.EDA: Task-Centric Exploratory Data Analysis for Statistical Modeling in Python, arXiv:2104.00841v2 [cs.DB] 10 Apr, (2021).

[12] Sumaya Habib, Maisha Binte Moin, Sujana Aziz, Kalyan Banik, Hossain Arif, Heart Failure Risk Prediction and Medicine Recommendation using Exploratory Data Analysis", 1st International Conference on Advances in Science, Engineering and Robotics Technology , (2019). 978-1-7281-3445-1, IEEE.

[13] Ching-seh (Mike) Wu, Mustafa Badshah, Vishwa Bhagwat, Heart Disease Prediction Using Data Mining Techniques, In Proceedings of 2019 2nd International Conference on Data Science and Information Technology (DSIT'19). Seoul, Korea, (2019) 5 pages. https://doi.org/10.1145/3352411.3352413.

[14] Dr T Lalitha, & Didwania, R., Future Prediction of Heart Disease through Exploratory Analysis of Data. SPAST Abstracts, 1(01) (2021). https://spast.org/techrep/article/view/324.

[15] H. Agrawal, J. Chandiwala, S. Agrawal and Y. Goyal, "Heart Failure Prediction using Machine Learning with Exploratory Data Analysis, *2021 International Conference on Intelligent Technologies (CONIT)*, (2021) 1-6, doi: 10.1109/CONIT51480.2021.9498561.

[16] Dr. P.K.A. Chitra, Dr. P. Udaykumar, Exploratory Analysis to Predict Heart Disease Occurrence through machine Learning Approaches, International Journal of Advanced Science and Technology, 29(9) (2020) 2702-2709.

[17] Akella, Aravind, and Sudheer Akella, Machine learning algorithms for predicting coronary artery disease: efforts toward an open-source solution, Future science OA , FSO698. 29 Mar., 7(6) (2021) doi:10.2144/fsoa-2020-0206.

[18] Owk Mrudula, A.Mary Sowjanya, Understanding Clinical Data using Exploratory Analysis, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, 8(5) (2020).

[19] Dhai Eddine Salhi, Abdelkamel Tari, and M-Tahar Kechadi, Using Machine Learning for Heart Disease Prediction, In book: Advances in Computing Systems and Applications (2021) 70-81. February 2021.DOI:10.1007/978-3-030-69418-0_7.

[20] Rishabh Magar, Rohan Memane, Suraj Raut, Prof. V. S. Rupnar, Heart Disease Prediction Using Machine Learning, Journal of Emerging Technologies and Innovative Research (JETIR), 7(6) (2020).

[21] M. Thangamani, R. Vijayalakshmi, M. Ganthimathi, M. Ranjitha, P. Malarkodi, S. Nallusamy, Efficient Classification of Heart Disease using KMeans Clustering Algorithm, International Journal of Engineering Trends and Technology., 68(12) (2020) 48-53. ISSN: 2231 – 5381 /doi:10.14445/22315381/IJETT-V68I12P209.

[22] Aman, Rajender Singh Chhillar, Disease Predictive Models for Healthcare by using Data Mining Techniques: State of the Art, International Journal of Engineering Trends and Technology, 68(10) (2020) 52-57. ISSN: 2231 – 5381 /doi:10.14445/22315381/IJETT-V68I10P209.