RESEARCH
www.commprac.com
ISSN 1462 2815

UNVEILING THE HIDDEN SECRETS: EXPLORING CARDIOVASCULAR DISEASE WITH DATA ANALYSIS USING R PROGRAMMING LANGUAGE

Rahbre Islam [1], Safdar Tanweer [1], Tabrez Nafis [1], Imran Hussain [1]
Shahzad Niwazi Qurashi [2]* and Javed Azimi [1]

[1] Department of Computer Science & Engineering, School of Engineering Sciences & Technology, Jamia Hamdard University, New Delhi, India.
[2] Department of Health Informatics, College of Public Health and Tropical Medicine, Jazan University, Kingdom of Saudi Arabia.
*Corresponding Author Email: squrashi@jazanu.edu.sa, ORCID ID: 0000-0002-9258-0473

DOI: 10.5281/zenodo.10065888

Abstract

EDA is an important and robust approach to summarizing data, visualizing key features, and assisting in building a prediction model. The paper finds different graphical methods, including bar plot, box plot, histogram, scatter plot, and correlation analysis to fetch the vital features, characteristics, and relationships among variables. The primary aim of this research was to perform exploratory data analysis using R Studio, specifically on a real-time dataset that has key information about patients diagnosed with cardiovascular disease. Moreover, the study was intended to trace the key attributes responsible for cardiovascular disease and design an intelligent prognosis model using machine learning techniques. To accomplish this research, an open-source software “R” was used where lots of inbuilt libraries are available to play with the data for analysis, graphical presentation, etc. A few packages like ggplot, dplyr, and many more are available in the R language. The findings and insights play a significant role in the early and effective detection of CVD. Using R packages such as dplyr, tidyr, and ggplot2, efficiently processed and visualized the data, facilitating meaningful conclusions. Visual representations such as boxplots, histograms, crosstabs, and scatterplots provided valuable information about patterns in the data set. Furthermore, the inclusion of larger data sets and advanced machine learning algorithms my help to develop predictive models of CVD.

Keywords: Exploratory Data Analysis (EDA), “R” language, Library, ggplot, dplyr, Visualization.

I. INTRODUCTION

In today's scenario, the world is experiencing a vast amount of data to be analyzed so as to land on the final spot to get well-informed data to design a model. A graphical presentation tells a lot of numerical data in one go. To accomplish the desired outcome, we used an R programming language that offers an alternative technique to apply for any application, especially in statistical presentations.

EDA is an approach that summarizes vital and crucial details of the data by fetching its important features and focusing on visualizing it with appropriate representation. It also emphasizes filtering assumptions and insights required for building a model.

Moreover, EDA instantly describes the dataset with these properties, number of rows, columns, missing data, data and its types as well as preview. It also provides a glimpse of noisy data, missing data, invalid data, and its types. The EDA visualizes different statistical distributions and patterns of data, for example, histograms, bar charts, box plots, correlations among variables, etc. EDA is a classical way of thinking where we use an appropriate model and express procedures [1].

EDA mechanism has been applied to the cardiovascular disease dataset so as to explore vital attributes responsible for the disease. This exploration will assist in building a CVD prediction model using machine learning (ML) algorithms.

II. RELATED WORK

Ghosha *et al.* [2] evaluated various kinds of data exploration tools and techniques for analysis. Whereas, John T. Behrens [3] used many graphical techniques to find the difference between classical and exploratory data analysis. A study on the SME sector of bank lending in Bhutan was conducted by Wangmo [4]. Another similar type of study was conducted in Ghana by Matthew Ntow-Gyamfi *et al.* [5] who studied loan default and credit risk among banks in Ghana. Furthermore, Francis *et al.* [6] conducted an EDA for loan prediction and applied machine learning techniques to assess the nature of the clients. Ulaga *et al.* [7] did an EDA analysis using R to predict loan privilege for clients applying the Random Forest (RF) algorithm. Konopka *et al.* [8] also performed an EDA of clinical data and developed a procedure to explore the multidimensional nature of data. In the literature, the use of EDA techniques gives a glimpse of model building using ML algorithms, so the prediction of heart disease may be incorporated by applying the same.

III. CARDIOVASCULAR DISEASE (HEART DISEASE)

The fast-paced and stressful nature of modern-day routines can cause stress and anxiety in the population. Additionally, the increasing number of individuals with obesity and smoking addiction has resulted in a surge of diseases, including heart problems, cancer, and other related conditions [9]. Prediction of diseases like cancer, heart disease, or other chronic ailments is a significant challenge due to the variations in any individual's blood pressure, heart rate, cholesterol, etc. Any abnormality in blood circulation can result in cardiovascular disease which includes heart-beat disorders, blood vessel diseases, and vascular diseases of the brain or any other vital organ. These symptoms, overall, show a green signal of severe health conditions [10].

The World Health Organization (WHO) reports that every year, approximately 18 million people die globally from heart disease and its complications. Moreover, heart attacks and strokes account for more than 80% of cardiovascular disease (CVD) related deaths [11].

The early and effective detection of heart disease is crucial for taking preventative measures to lower its toll. The majority of cardiac ailments can be prevented by addressing lifestyle and behavioral risk factors such as poor eating habits, physical inactivity, smoking, and alcohol intake. These risk factors can lead to various symptoms like obesity, hypertension, hyperglycemia (elevated blood sugar levels), and hyperlipidemia (increased blood lipids) [12].

The invasive diagnostic techniques are expensive, time-consuming, uncomfortable, and may produce false results due to human error. Therefore, there is a need for a non-invasive method that can diagnose cardiac disease quickly and at a lower cost. Machine learning (ML) is one of the promising techniques that can provide better and faster results.

IV. WORK ON REAL-TIME DATASET

In this study, the vitals from a real-time dataset of heart-disease patients were taken from Rohilkhand Hospital, Shahjahanpur, (U.P., India). This dataset has 28 attributes i.e. Age, Sex, Blood Pressure (High or Low), diabetes, etc. Based upon these features, the Target variable considers whether a patient has heart disease or not, where '0' represents no heart disease while '1' represents heart disease. Using the str() function, the nature, and structure of the dataset were analyzed and then further proceeded for exploration based on its nature. The structure of the dataset is shown in Figure 1.

Age	Ht.m2.	Wt.kg.	BMI .kg.m2.	SBP	DBP	HR	PP	RBP	chol	MHR	OPK	CPT	FBS	RES	EX	slope	VCA	THA	Physical_Act	Smoking	Alcohol Drinking	HTN	Family. History.of. CVD	Stress	Sex	Diabeties	Target
int	num	num	num	int	int	int	int	int	int	int	num	in	int	in	int	int	int	int	int	int	int	int	int	int	factorial	int	factotial

Fig 1: Structure of dataset

To analyze the above dataset with a possible set of options, first of all, the CSV file was imported using the read.csv () function in the R environment, then a variable was assigned to this CSV file and converted into a data frame. This dataset has 28 columns with different attributes. Additionally, the structure and class of the dataset were observed with the str() function. The head (data frame) function showed 5-6 rows of data, giving a glimpse of the whole dataset. Figure 2 represents the heart disease dataset.

Age	Ht.m2.	Wt.kg.	BMI..kg.m2.	SBP	DBP	HR	PP	RBP	chol	MHR	OPK	CPT	FBS	RES	EX	slope	VCA	THA
62	1.675	57	21.13405	110	65	90	45	160	164	145	6.2	0	0	1	1	2	0	2

Physical_Act	Smoking	Alcohol.Drinking	HTN	Family..History.of.CVD	Stress	Sex	Diabeties	Target
1	0	0	0	0	2	M	3	1

Fig 2: Record of heart disease dataset

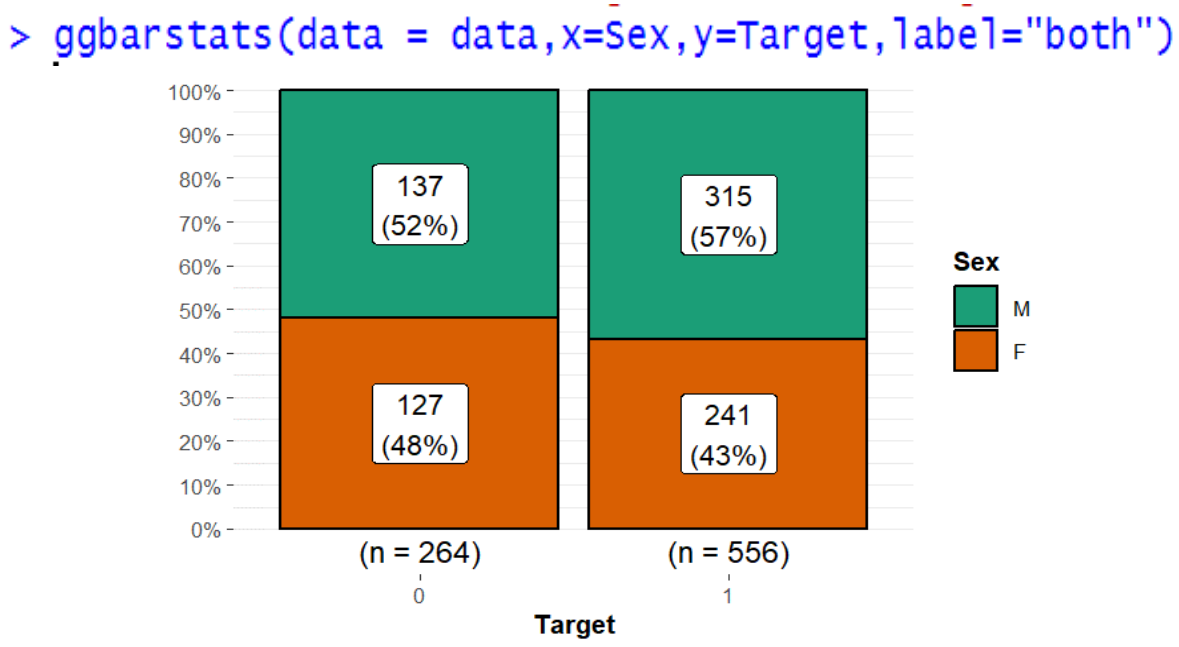

Fig 3: Gender Vs CVD

The above R studio code (Figure 3) shows the overall graphical representation of the patient dataset based on gender with heart disease.

V. EDA TECHNIQUES

EDA plays a significant role in data analysis which involves cleaning the data and using visualizations to fetch insights with summarization. EDA employs interactive displays and graphics, emphasizing model building and addressing measurement issues. It can be a graphical or non-graphical representation and focuses on both univariate and multivariate data. Furthermore, it helps to understand data before modeling and dealing with large datasets. The available techniques like bar charts, histograms, box plots, scatter plots, and violin plots are mostly used for visualization purposes and they help in data exploration, pattern recognition, and well-informed decision-making [13].

EDA can be classified into two sections, descriptive statistical technique and graphical techniques. The descriptive technique goes through both univariate and multivariate statistical techniques, whereas the graphical section possesses various visualization techniques [14].

These EDA techniques are used to explore data, understand the patterns hidden in the data, and provide clues to trace existing relationships between the variables. Overall, the statistical and graphical techniques of EDA produce insights from the data that help in building a sophisticated model.

VI. EDA IN "R"

In the present study, a simple language "R" was used for data analyses. It is an open-source, free-to-use interpreted language that has a galaxy of in-built libraries for data analysis and can communicate with other third-party languages. The availability of library functionality in "R" provides an efficient visualization process that helps in a better understanding of the information [15].

VII. PACKAGES IN "R"

A collection of powerful and advanced packages for data analysis is available in "R" that are used to clean, transform, and analyze CSV format files as well. The accessible packages are as follows:

dplyr: It facilitates data manipulation and transformation processes through data filtering, selecting, arranging, creating groups, and summarizing data easily with great precision.

tidyr: This package is used for data tidying as it provides functions like gather() and spread() to reshape and transform data between wide and long formats.

ggplot2: A robust package for data visualization and it offers a flexible grammar of graphics approach to create high-quality, customizable plots.

data.table: To handle large datasets, this package provides an efficient way to work. It offers fast manipulation and operation. It also supports features like joins aggregations, and data reshaping.

readr: A package for efficient reading of rectangular data files. It provides functions like read_csv(), read_tsv(), and read_delim() that are faster and more user-friendly alternatives to base R's read.table().

caret: A comprehensive package used to develop prediction models applying Machine Learning (ML) techniques. It also provides an interface for training, testing, and evaluating models, along with different pre-processing and feature-selection techniques.

tidyverse: A collection of packages like dplyr, tidyr, ggplot2, and more. The tidyverse emphasizes clean and consistent data manipulation for high-quality visualization.

VII. DESCRIPTIVE STATISTICAL TECHNIQUES AND GRAPHICAL EDA TECHNIQUES

Graphical Exploratory Data Analysis is a complement to non-graphical EDA that is used to analyze and summarize statistical features keeping a view of the key [15].

A. Histograms: It is a graphical representation of the distributed data that shows the frequency of data items in successive order with numerical intervals having equal size. In a histogram, the horizontal axis represents independent variables while the vertical axis represents dependent variables. The data appears in a rectangular shape of variables having a color of shade [16].

B. **Graphical representation of skewness feature:** The Histogram represents the location, the spread, the skewness, the presence of outliers of the data, and the presence of multiple modes in the data [17]. The symmetric distribution displays a histogram where the two halves are mirror images of each other, while the non-symmetric distribution doesn't have a mirror image. Skewed distributions often show one tail that is significantly longer or expanded compared to the other. A "skewed right" distributed data is on the right side with a long tail [17].

In this study, the histogram was plotted based on one of the continuous variables “cholesterols” of the heart disease dataset. ggplot() function has been applied via the pipe operator of R language that is useful to construct a plot of an object. The histogram of cholesterol with skewness is shown in Figure 5.

```
> data%>%ggplot(data = data,mapping = aes(x=chol))+
+   geom_histogram(bins = 50,fill="green")+
+   geom_freqpoly(binwidth=3,color="red")+theme_light()
```

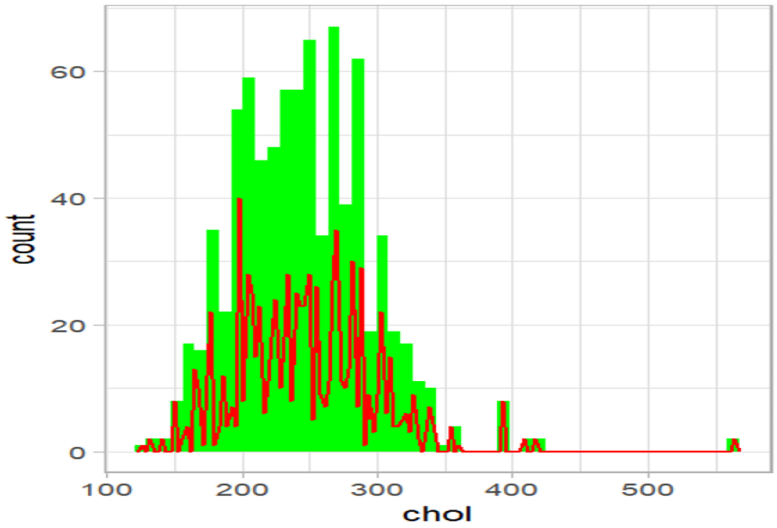

Fig 4: Histogram of Cholesterol with skewness

The above diagram reveals that “chol” is slightly skewed to the right, it can be verified by showing the numerical values of the same. For this, we can apply the “R code” to summarize the “chol” variable.

variables	min	Q1	mean	median	Q3	max	zero	minus	outlier
chol	126.00000	207.7500	245.887805	243.00000	279.00000	564.00000	0	0	14

The above summary also confirms that skewness is due to the variations in the distance. In this case, the distance between the 3rd quartile and the maximum value is 285 (564.0-279.0), and the distance between the 1st quartile and the minimum is 81.8 (207.8-126.0). To cross-check the above values, we can use a boxplot with the same data.

C. Box plots

A five-numbered summary of data distribution with graphical representation, its ranges are minimum, maximum, sample, median, and the first and the third [18]. Box plots are considered to be a superior graphical tool for identifying outliers compared to histograms.

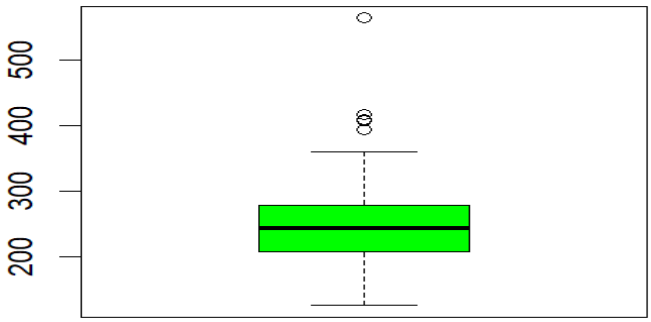

Fig 5: Boxplot of Cholesterol

In the above boxplot some unusual points are shown that points are nothing but an outlier so, it is reconfirmed that our data are not normally distributed.

D. Outlier

An observation that appears to be inconsistent with the other one. In real-world scenarios, outliers can arise from various causes, equipment malfunctions, day-to-day variations, batch inconsistencies, anomalous input conditions, and warm-up effects [17]. It also shows the visual points falling more than 1.50 times the Inter-Quartile Range from either edge of the confined area of a box plot [19].

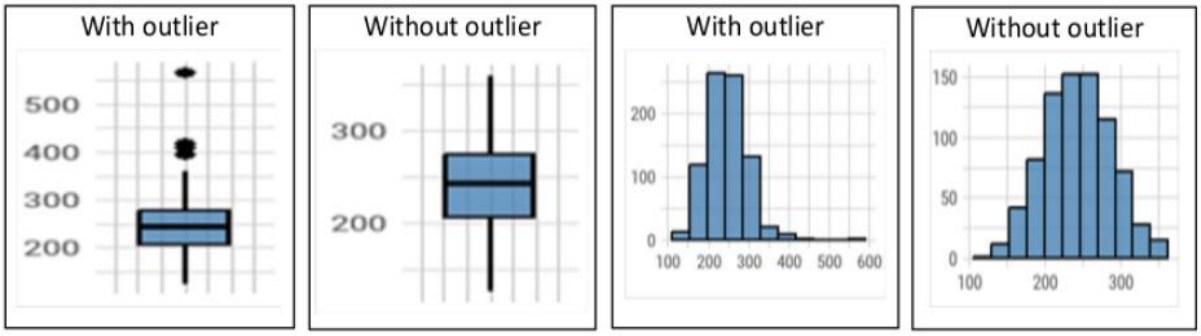

Fig 6: Representation of "chol" variable of heart dataset with outlier and without outlier

E. Quantile-Quantile plots

Another well-featured plot in the case of continuous variables is the Quantile-Quantile plot which is used to test whether the variables are well-suited to a normal distribution or not [36].

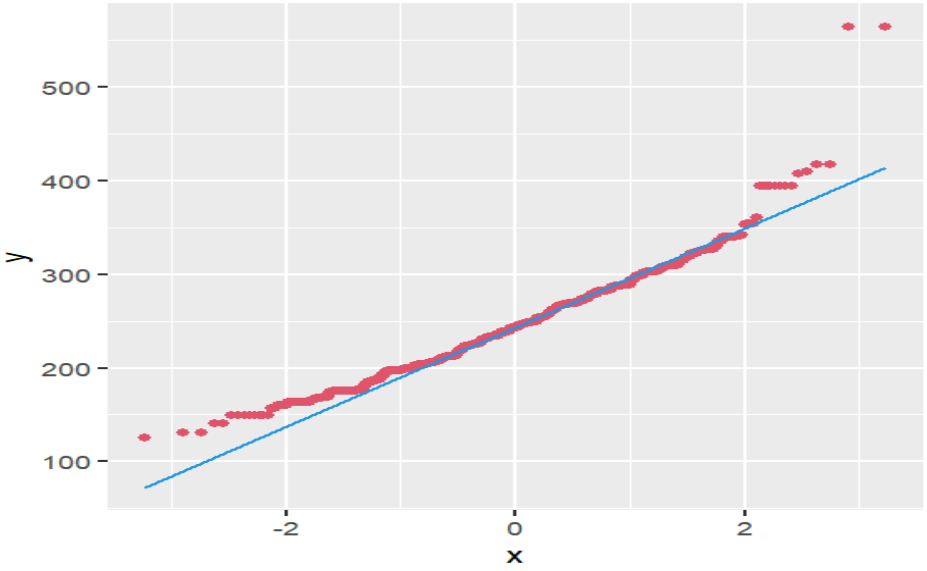

Fig 7: Quantile-Quantile plots

In the above graphical representation, many data points are significantly distant from the normal distribution line, it could be interpreted that our data does not follow a normal distribution.

F. Cross-Tabulation

A contingency table or crosstab is used to summarize and analyze the association or relationship among categorical variables. It is a way to judge the distribution of variables and identify if any relationships or patterns exist between them [20].

Total Observations in Table: 820

data$Sex	data$Target 0	data$Target 1	Row Total
F	127	241	368
	0.345	0.655	0.449
	0.481	0.433	
	0.155	0.294	
M	137	315	452
	0.303	0.697	0.551
	0.519	0.567	
	0.167	0.384	
Column Total	264	556	820
	0.322	0.678	

Fig 8: Cross-table of two categorical variables, Sex and Target

It has been shown in Figure 5, that 65.5% of females have heart diseases while 69.7% of males suffer from heart diseases.

G. Marginal Plot

It is basically a scatter plot with additional information that is shown in the margins by plotting the same. Following R studio code shown below, a marginal plot was generated by using ggplo() and dplyr properties of R.

```
p<-data%>%ggplot(aes(x=Age,y=chol))+
+ geom_point(color='red')+
+ geom_smooth(se=0,method = 'lm')
> ggMarginal(p,fill='blue',type = 'histogram')
```

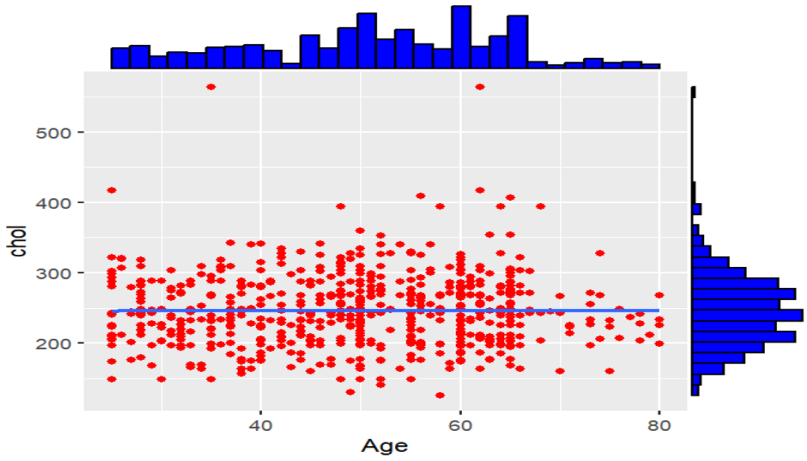

Fig 9: Marginal plot of "chol" and "Age" variables

In marginal plots the x- and y-axes show the indicators of non-normal or unusual data as both the axes are showing margins in the form of bars.

H. Bar Plot

It uses horizontal or vertical columns to represent data values for different categories or groups. A longer bar indicates higher values, allowing easy comparison of a single variable across multiple groups [21].

Following the below-mentioned R studio code, a bar graph was generated that displayed ages of male and female patients with heart disease.

```
> db$Age_Cat_Sex_M_with_disease <-
+ ifelse((db$Age>=25) & (db$Age<=30)& (db$Sex=='M')&(db$Target==1), "25-30",
+ ifelse((db$Age>=31) & (db$Age<=35)& (db$Sex=='M')&(db$Target==1), "31-35",
+ ifelse((db$Age>=36) & (db$Age<=40)& (db$Sex=='M')&(db$Target==1), "36-40",
+ ifelse((db$Age>=41) & (db$Age<=45)& (db$Sex=='M')&(db$Target==1), "41-45",
+ ifelse((db$Age>=46) & (db$Age<=50)& (db$Sex=='M')&(db$Target==1), "46-50",
+ ifelse((db$Age>=51) & (db$Age<=55)& (db$Sex=='M')&(db$Target==1), "51-55",
+ ifelse((db$Age>=56) & (db$Age<=60)& (db$Sex=='M')&(db$Target==1), "56-60",
+ ifelse((db$Age>=61) & (db$Age<=65)& (db$Sex=='M')&(db$Target==1), "61-65",
+ ifelse((db$Age>=66) & (db$Age<=70)& (db$Sex=='M')&(db$Target==1), "66-70",
+ ifelse((db$Age>=71) & (db$Age<=75)& (db$Sex=='M')&(db$Target==1), "71-75",
+ ifelse((db$age>=76) & (db$Age<=80)& (db$Sex=='M')&(db$Target==1), "76-80","80+")))))))))))
> table(db$Age_Cat_Sex_M_with_disease)

25-30 31-35 36-40 41-45 46-50 51-55 56-60 61-65 66-70 71-75
   26    18    26    18    45    41    50    70    11     9
```

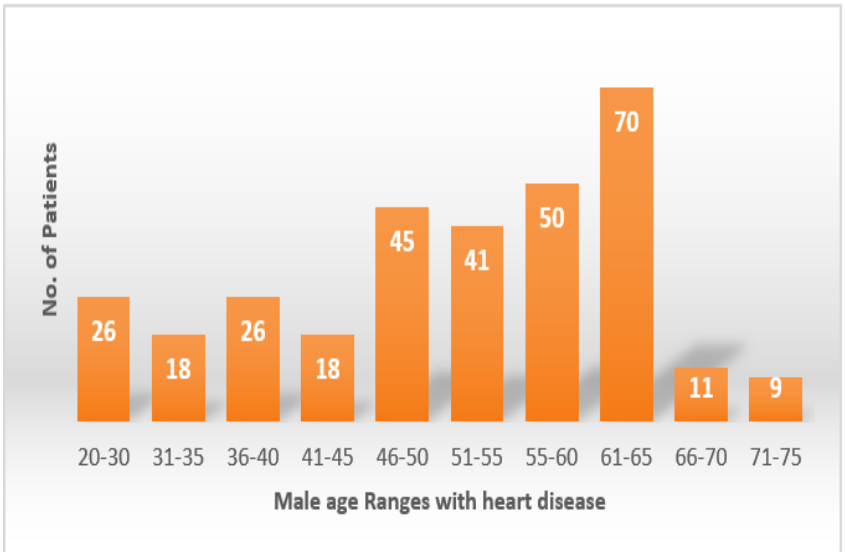

Fig 10: Bar graph of Male age Ranges with heart disease

```
> db$Age_Cat_Sex_F_with_disease<-
+ ifelse((db$Age>=25) & (db$Age<=30)& (db$Sex=='F')&(db$Target==1), "25-30",
+ ifelse((db$Age>=31) & (db$Age<=35)& (db$Sex=='F')&(db$Target==1), "31-35",
+ ifelse((db$Age>=36) & (db$Age<=40)& (db$Sex=='F')&(db$Target==1), "36-40",
+ ifelse((db$Age>=41) & (db$Age<=45)& (db$Sex=='F')&(db$Target==1), "41-45",
+ ifelse((db$Age>=46) & (db$Age<=50)& (db$Sex=='F')&(db$Target==1), "46-50",
+ ifelse((db$Age>=51) & (db$Age<=55)& (db$Sex=='F')&(db$Target==1), "51-55",
+ ifelse((db$Age>=56) & (db$Age<=60)& (db$Sex=='F')&(db$Target==1), "56-60",
+ ifelse((db$Age>=61) & (db$Age<=65)& (db$Sex=='F')&(db$Target==1), "61-65",
+ ifelse((db$Age>=66) & (db$Age<=70)& (db$Sex=='F')&(db$Target==1), "66-70",
+ ifelse((db$Age>=71) & (db$Age<=75)& (db$Sex=='F')&(db$Target==1), "71-75",
+ ifelse((db$age>=76) & (db$Age<=80)& (db$Sex=='F')&(db$Target==1), "76-80","80+")))))))))))
> table(db$Age_Cat_Sex_F_with_disease)

25-30 31-35 36-40 41-45 46-50 51-55 56-60 61-65 66-70 71-75
   15    18    20    32    52    23    28    30     8     5
```

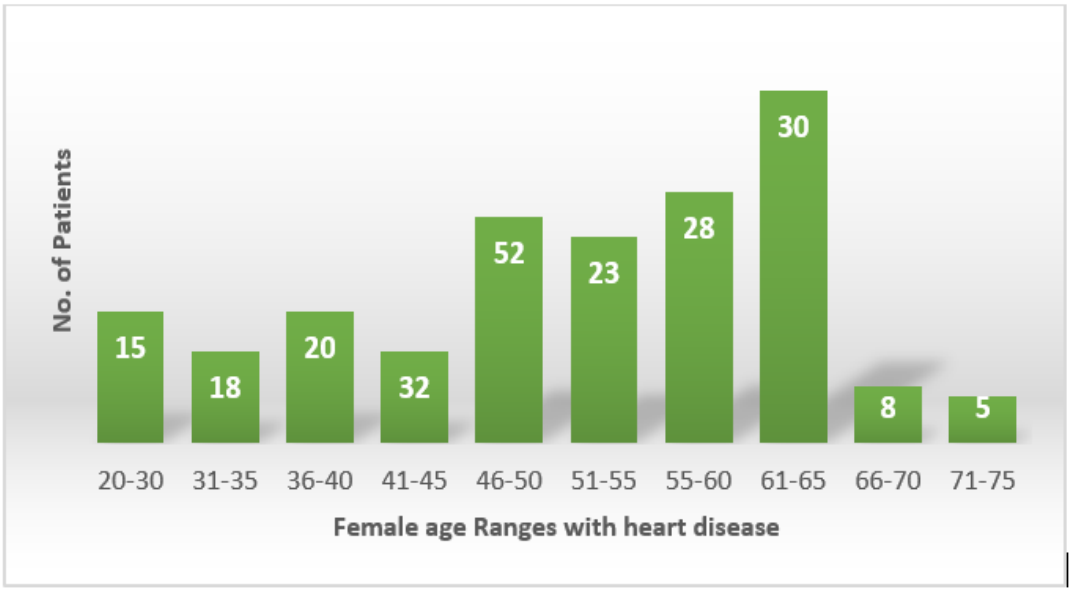

Fig 11: Bar graph of Female age Ranges with heart disease

I. Scatter plots

This plot utilizes Cartesian coordinates to demonstrate the relationship between two variables in the dataset. The X and Y coordinates represent the values of the variables, and the data is represented as a collection of points on the plot [21].

```
> ggplot(data=my_data,aes(x=Age,y=chol))+
+  geom_point()+geom_smooth(formula = y~poly(x,2),
method = "lm",se=T,level=0.95)
```

The above-mentioned R studio code generates the following scatter plot with a linear regression line in blue color.

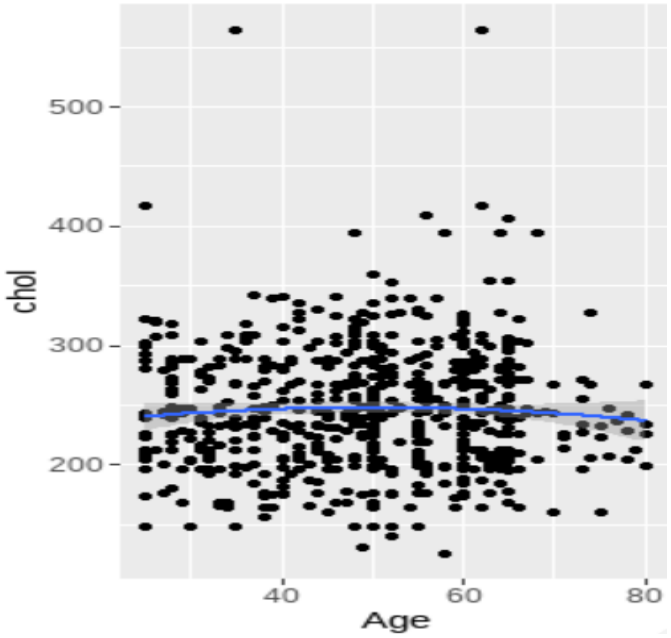

Fig 12: Linear regression for correlation between age and cholesterol level

This plot has a linear regression line with standard error and confidence with an interval of 0.95%. Here ages between 40 and 60 have more cholesterol levels of 200 to 300.

VIII. CORRELATION

A statistical technique used to examine a linear relationship between two continuous variables. It helps to assess a linear association between variables. A small ‘r’ is used to represent the relationship. It is also known as the correlation coefficient and falls in the range from -1 to +1[22].

A. Pearson's Correlation Coefficient

Pearson's correlation coefficient, represented by 'ρ' for a population and 'r' for a sample indicates a linear relationship between two variables. It is best suitable when both variables follow normal distribution properties. This coefficient can be influenced by extreme values, which may weaken the strength of the relationship. So, it is not a best practice to use when one of the variables is not following the normal distribution rule [22].

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} \qquad \text{Eq.01}$$

Here r is Pearson’s correlation coefficient, x is a value in the first set of data, y is a value in the second set of data and n is to total number of values.

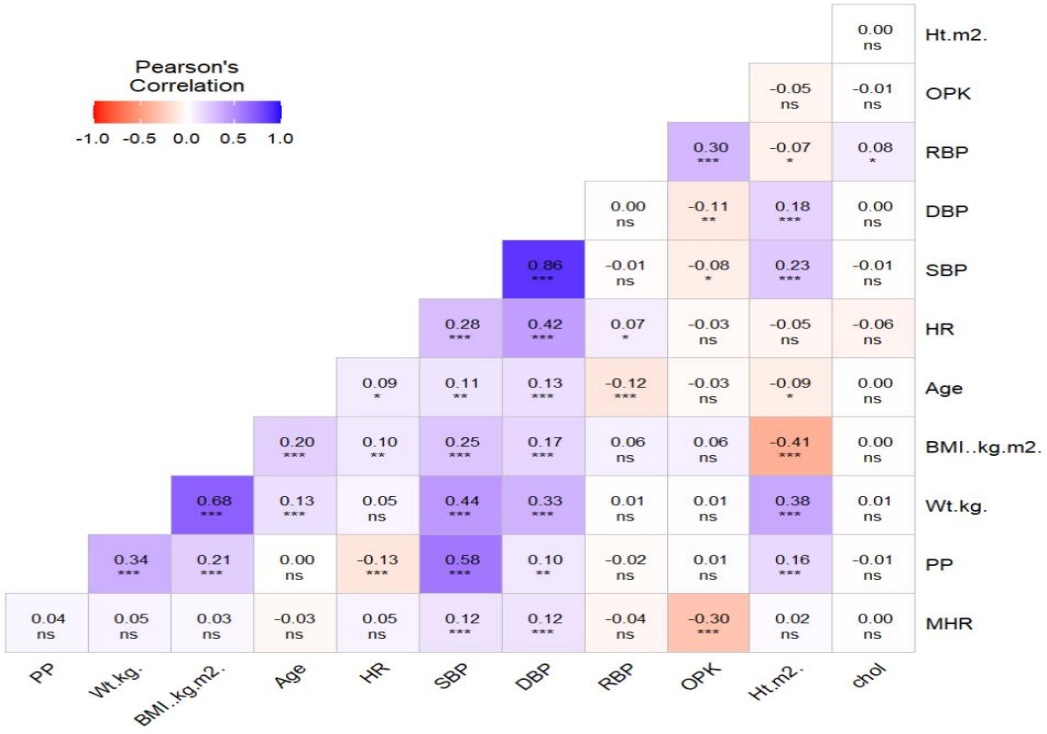

Fig 13: Pearson correlation among various variables

As observed in Figure 13, the correlation of continuous variables with star sign *** represents a strong relationship among the variables which may be useful in disease prediction.

B. Spearman's rank Correlation Coefficient

Correlation coefficients 'ρs' for a population and 'rs' for a sample of the population are special measurements that describe how the two variables are related to each other. They give some indication regarding the strength and direction of the relationship, even if it is not a straight line. These measures are suitable when one or both variables are not evenly distributed or are ranked in order. They are also reliable and not easily affected by extreme values [23].

$$r_R = 1 - \frac{6\Sigma_i {d_i}^2}{n(n^2-1)} \qquad \text{Eq.02}$$

Here 'n' denotes the number of data points of the two variables, and 'di' denotes the difference in ranks of the 'ith' element. The Spearman Coefficient 'ρ' takes a value between +1 to -1 where a 'ρ' value of +1 means it shows a perfect association of ranks, a 'ρ' value of '0' means no association of ranks, and a 'ρ' value of '-1' means a perfect negative association between ranks.

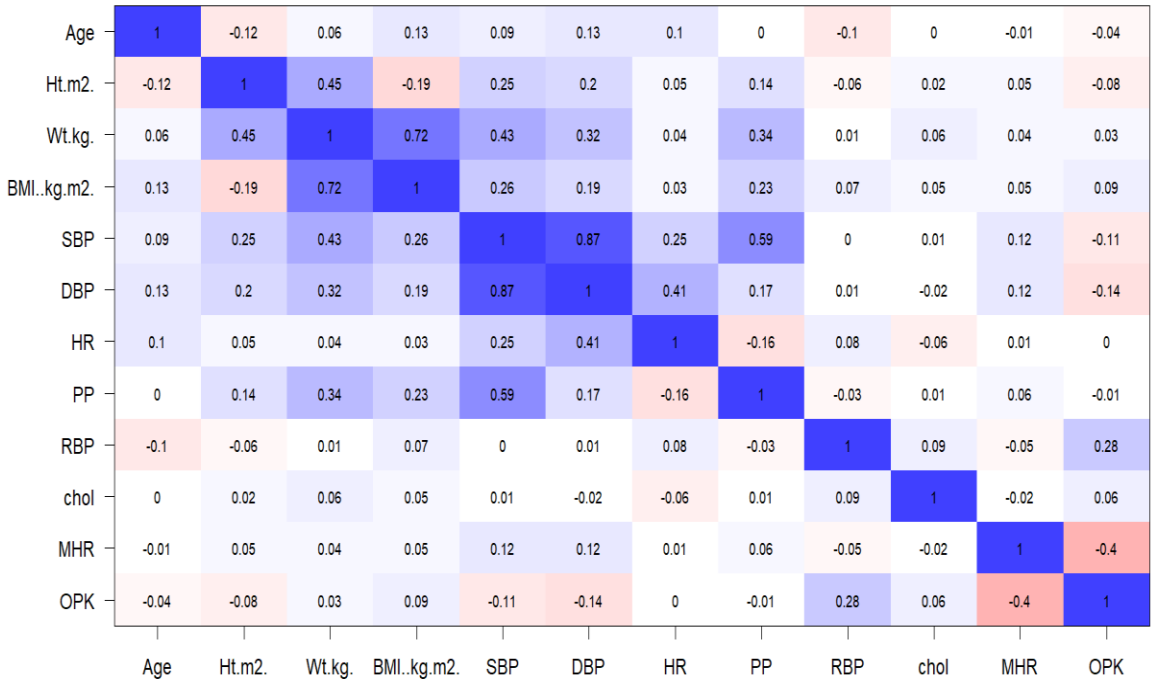

Fig 14: Spearman's rank Correlation Coefficient of continuous variables

IX. RESULT AND ANALYSIS WITH SOME EXAMPLES

The main aim of this study was to study cardiovascular disease data and understand the different key factors that affect it. An "R" programming language has been used for analyzing and visualizing datasets. We examined various factors in the dataset and visualized them through graphs to see how they relate to the results and how can they help to develop a machine learning model. The graphs obtained in the study help us to see the patterns and trends in the data more clearly.

X. SOME OF THE FINDINGS BASED ON THE ABOVE-SAID DATASET ARE GIVEN BELOW:

A. What is the correlation between age and heart disease?

Solution: To get the result we used "cor", an inbuilt function of R studio, and got a correlation coefficient of 0.07067206 which shows that there is a positive correlation between "age" and "heart disease".

```
> correlation <- cor(data$Age,data$Target)
> print(correlation)
[1] 0.07067206
```

B. Show the number of patients having heart disease or not where ages are >=50 years and cholesterol level is >=100.

Solution: - To get the result we used "dlookr" package of R studio which is generally used for data diagnosis.

The R code was applied for the same and found that 339 patients have heart disease while 139 patients do not have heart disease.

```
>age_chol<-my_data%>%
+ filter(Age>=50,chol>=100,Target==1)%>%
+ glimpse()
> age_chol<-my_data%>%
+ filter(Age>=50,chol>=100,Target==0)%>%
+ glimpse()
```

C. Does Age and cholesterol have a significant impact on heart disease?

Solution: - Yes both age and cholesterol have an impact on heart disease. To evaluate the effect of cholesterol and age on heart disease, we applied an analysis of variance (ANOVA) test and found that p-values are 0.0426, and 0.0281 which are less than 0.05. So null hypothesis (H0) can be rejected.

```
>result<-aov (Target~Age+chol, data=data)
> summary(result)
             Df Sum Sq Mean Sq F value Pr(>F)
Age           1   0.89  0.8940   4.125 0.0426 *
chol          1   1.05  1.0481   4.836 0.0281 *
Residuals   817 177.06  0.2167
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
```

XI. FUTURE SCOPE

The present study also contains a few limitations that can be studied in future studies. Here, we applied and used exploratory data analysis only. Additionally, we can apply different machine learning algorithms with the existing dataset after analyzing it and can develop a succinct predictive model that can perfectly forecast the statistics of future chronic disease. In addition to this, we can establish a sound correlation analysis by analyzing the relationship between two continuous variables.

XII. CONCLUSION

This study highlights the importance of exploratory data analysis (EDA) in understanding key features of data in model development. Using the “R” programming language and visualization techniques, EDA helps us to obtain valuable information about a dataset, including its distribution, structure, and relationships among variables. Potential risk factors for heart disease and intriguing associations have been revealed between certain variables. Using R packages such as dplyr, tidyr, and ggplot2, efficiently processed and visualized the data, facilitating meaningful conclusions. Visual representations such as boxplots, histograms, crosstabs, and scatterplots provided valuable information about patterns in the data set.

In the future, the inclusion of larger data sets and advanced machine learning algorithms may help to develop predictive models of CVD. Moreover, examining the correlation between continuous variables and other risk factors can provide additional information for the analysis of heart disease.

References

1. Morgenthaler, S. (2009). Exploratory data analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *1*(1), 33-44.
2. Ghosh, A., Nashaat, M., Miller, J., Quader, S., & Marston, C. (2018). A comprehensive review of tools for exploratory analysis of tabular industrial datasets. *Visual Informatics*, *2*(4), 235-253.
3. Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological methods*, *2*(2), 131.
4. Wangmo, C. (2017). An Exploratory Study on Bank Lending To SME Sector In Bhutan. *International Journal of Scientific & Technology Research*, *6*(1), 1-23.
5. Ntow-Gyamfi, M., & Boateng, S. (2013). Credit risk and loan default among Ghanaian banks: An exploratory study. *Management Science Letters*, *3*(3), 753-762.
6. Jency, X. F., Sumathi, V. P., & Sri, J. S. (2018). An exploratory data analysis for loan prediction based on nature of the clients. *International Journal of Recent Technology and Engineering (IJRTE)*, *7*(4), 17-23.
7. Konopka, B. M., Lwow, F., Owczarz, M., & Łaczmański, Ł. (2018). Exploratory data analysis of a clinical study group: Development of a procedure for exploring multidimensional data. *PloS one*, *13*(8), e0201950.
8. Bogumil M. Konopka, Felicja Lwow, Magdalena Owczarz, Łukasz Łaczmański, (2018) "Exploratory data analysis of a clinical study group: development of a procedure for exploring multidimensional data. *PLOS ONE*. doi.org/10.1371/journal.pone.0201950
9. Pouriyeh, S., Vahid, S., Sannino, G., De Pietro, G., Arabnia, H., & Gutierrez, J. (2017, July). A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease. In *2017 IEEE symposium on computers and communications (ISCC)* (pp. 204-207). IEEE.
10. Azmi, J., Arif, M., Nafis, M. T., Alam, M. A., Tanweer, S., & Wang, G. (2022). A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data. *Medical Engineering & Physics*, *105*, 103825.
11. https://www.who.int/health-topics/cardiovascular-diseases#tab
12. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases
13. Vignesh, K., & Nagaraj, P. (2022, February). Analyzing the Nutritional Facts in Mc. Donald's Menu Items Using Exploratory Data Analysis in R. In *International Conference on Emerging Technologies in Computer Engineering* (pp. 573-583). Cham: Springer International Publishing.
14. Pearson, R. K. (2018). *Exploratory data analysis using R*. CRC Press.
15. Ioannidis, Y. (2003, January). The history of histograms (abridged). In *Proceedings 2003 VLDB Conference* (pp. 19-30). Morgan Kaufmann.
16. https://www.itl.nist.gov/div898/handbook/
17. Williamson, D. F., Parker, R. A., & Kendrick, J. S. (1989). The box plot: a simple visual method to interpret data. *Annals of internal medicine*, *110*(11), 916-921.
18. Kumar, R. V. (2021). Exploratory Data Analysis using R & RStudio.
19. Kaliyadan, F., & Kulkarni, V. (2019). Types of variables, descriptive statistics, and sample size. *Indian dermatology online journal*, *10*(1), 82.
20. Slutsky, D. J. (2014). The effective use of graphs. *Journal of wrist surgery*, *3*(02), 067-068.
21. Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi medical journal*, *24*(3), 69-71.
22. Xiao, C., Ye, J., Esteves, R. M., & Rong, C. (2016). Using Spearman's correlation coefficients for exploratory data analysis on big dataset. *Concurrency and Computation: Practice and Experience*, *28*(14), 3866-3878.