1. Introduction

Problem Statement:

The aim of this project work is analyze the data set received from the medical practitioner in order to identify whether the person either male or female will have a heart risk disease after ten years given the set of variables. The project work also identifies the most important variables which are the main causes of heart attacks and suggests the appropriate solution to reduce the risk of heart attack in the next ten years.

Need of the Study:

The study of the causes or factors responsible for the heart attack is necessary, so that the propensity of either male or female to heart attacks can be controlled which increases the life expectancy of the individuals.

This study will also be helpful to the hospitals and medical/pharmaceutical industries as well as government's "Ministry of Health Affairs" department to analyze what are the main factors for heart attacks so that accordingly they can manufacture drugs and organize campaigns or awareness programs.

Understanding Business/Social Opportunity:

During the study of causes/factors responsible for heart attack, I believe that it is a social opportunity to inform the people about the health concerns and the major disadvantages related to smoking cigarettes, high cholesterol level, high blood pressure etc. Meanwhile, arranging awareness program informing the benefits of regular exercises and good food habits.

It is also the business opportunity for the pharmaceutical industries and health instructors/ dieticians to create wealth by manufacturing drugs and suggesting proper diet and exercises to their customers.

2. Data Report

- The data was collected from the medical practitioner for the purpose of study and building appropriate model to predict the risk of heart attack in the next ten years.

Table-1: Variable Description

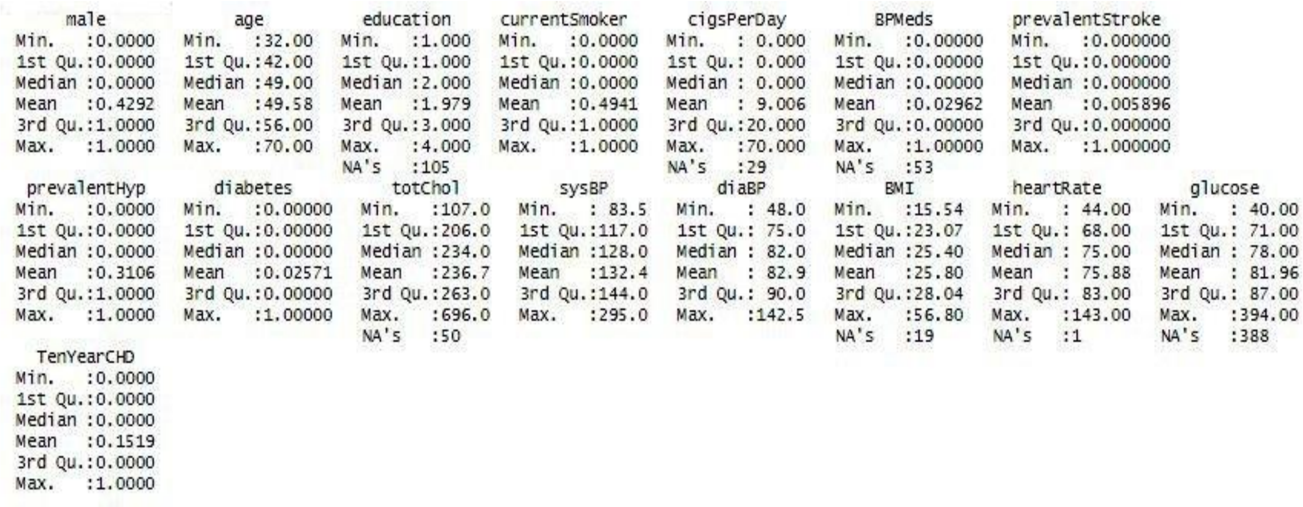
Category of variable	**Name of the column**	**Meaning**	**Type of variable**
Demographic	sex	It represents the sex of the patient whether male or female ("1" means Male, "0" means Female)	Nominal
	age	It represents the age of the patient	Continuous
	education	It represents the education level ("1" means High School, "2" means Bachelors, "3" means Masters and "4" means Professional/PhD)	Continuous

Behavioral	currentSmoker	It represents whether the person is a current smoker or not (“1” means Yes, “0” means No)	Nominal
	cigsPerDay	It represents average number of cigarettes that a person smoke in a day	Continuous
Medical(history)	BPMeds	It indicates whether the patient was under blood pressure medication or not (“1” means Yes, “0” means No)	Nominal
	prevalentStroke	It indicates that whether the patient had stokes or not (“1” means Yes, “2” means No)	Nominal
	prevalentHyp	It indicates that whether the patient was hypertensive or not (“1” means Yes, “2” means No)	Nominal
	diabetes	It indicates whether the patient had diabetes or not (“1” means Yes, “2” means No)	Nominal
Medical(current)	totChol	Represents total cholesterol level of the patient	Continuous
	sysBP	Represents systolic blood pressure of the patient	Continuous
	diaBP	Represents diastolic blood pressure of the patient	Continuous
	BMI	Represents body mass index of the patient	Continuous
	heartRate	Represents the heart rate of the patient	Continuous
	glucose	Represents glucose level in the patient	Continuous
Predict variable (desired target)	TenYearCHD	Represents 10 year coronary heart risk disease (“1” means Yes, “2” means No).	Nominal

3. Exploratory Data Analysis

Summary:

- Age is a continuous variable with minimum value of 32 and maximum value of 70. It has mean of 49.58.
- The data set contains the “NA” values for the columns “educations”, “cigsPerDay”, “BPMeds”, “totChol”, “BMI”, “heartRate”, “glucose”.
- totChol has a minimum value of 107 and maximum value of 696. It also has a mean value of 236.7.
- glucose has a minimum value of 40 and a maximum value of 394. The mean value of glucose is 81.96.
- heartRate is also found to be a very important variable as it has large variation between minimum and maximum value. The minimum value is 44 and maximum value is 143. The mean value of heart rate is 75.88.
- diaBP is also a variable of consideration. It has a minimum value of 48 and maximum value of 142.5. The mean value is 82.9.



```
      male             age          education     currentSmoker      cigsPerDay        BPMeds         prevalentStroke
 Min.   :0.0000   Min.   :32.00   Min.   :1.000   Min.   :0.0000   Min.   : 0.000   Min.   :0.00000   Min.   :0.000000
 1st Qu.:0.0000   1st Qu.:42.00   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.: 0.000   1st Qu.:0.00000   1st Qu.:0.000000
 Median :0.0000   Median :49.00   Median :2.000   Median :0.0000   Median : 0.000   Median :0.00000   Median :0.000000
 Mean   :0.4292   Mean   :49.58   Mean   :1.979   Mean   :0.4941   Mean   : 9.006   Mean   :0.02962   Mean   :0.005896
 3rd Qu.:1.0000   3rd Qu.:56.00   3rd Qu.:3.000   3rd Qu.:1.0000   3rd Qu.:20.000   3rd Qu.:0.00000   3rd Qu.:0.000000
 Max.   :1.0000   Max.   :70.00   Max.   :4.000   Max.   :1.0000   Max.   :70.000   Max.   :1.00000   Max.   :1.000000
                                  NA's   :105                      NA's   :29       NA's   :53
  prevalentHyp       diabetes         totChol          sysBP           diaBP            BMI          heartRate         glucose
 Min.   :0.0000   Min.   :0.00000   Min.   :107.0   Min.   : 83.5   Min.   : 48.0   Min.   :15.54   Min.   : 44.00   Min.   : 40.00
 1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:206.0   1st Qu.:117.0   1st Qu.: 75.0   1st Qu.:23.07   1st Qu.: 68.00   1st Qu.: 71.00
 Median :0.0000   Median :0.00000   Median :234.0   Median :128.0   Median : 82.0   Median :25.40   Median : 75.00   Median : 78.00
 Mean   :0.3106   Mean   :0.02571   Mean   :236.7   Mean   :132.4   Mean   : 82.9   Mean   :25.80   Mean   : 75.88   Mean   : 81.96
 3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:263.0   3rd Qu.:144.0   3rd Qu.: 90.0   3rd Qu.:28.04   3rd Qu.: 83.00   3rd Qu.: 87.00
 Max.   :1.0000   Max.   :1.00000   Max.   :696.0   Max.   :295.0   Max.   :142.5   Max.   :56.80   Max.   :143.00   Max.   :394.00
                                    NA's   :50                                      NA's   :19      NA's   :1        NA's   :388
   TenYearCHD
 Min.   :0.0000
 1st Qu.:0.0000
 Median :0.0000
 Mean   :0.1519
 3rd Qu.:0.0000
 Max.   :1.0000
```

Fig.1: Summary of the data set

Univariate Analysis:

- Variable age has a right skewed distribution (Fig.2).
- Variable BMI also has right skewness (Fig.2).
- Variable cigsPerDay has highly right skewed distribution (Fig.2).
- Variable diaBP has normal distribution (Fig.2).
- Variable glucose is also highly right skewed (Fig.3).
- Variable heartRate also has right skewed distribution (Fig.3).
- Variable sysBP has right skewness (Fig.3).
- Variable totChol is also right skewed (Fig.3).

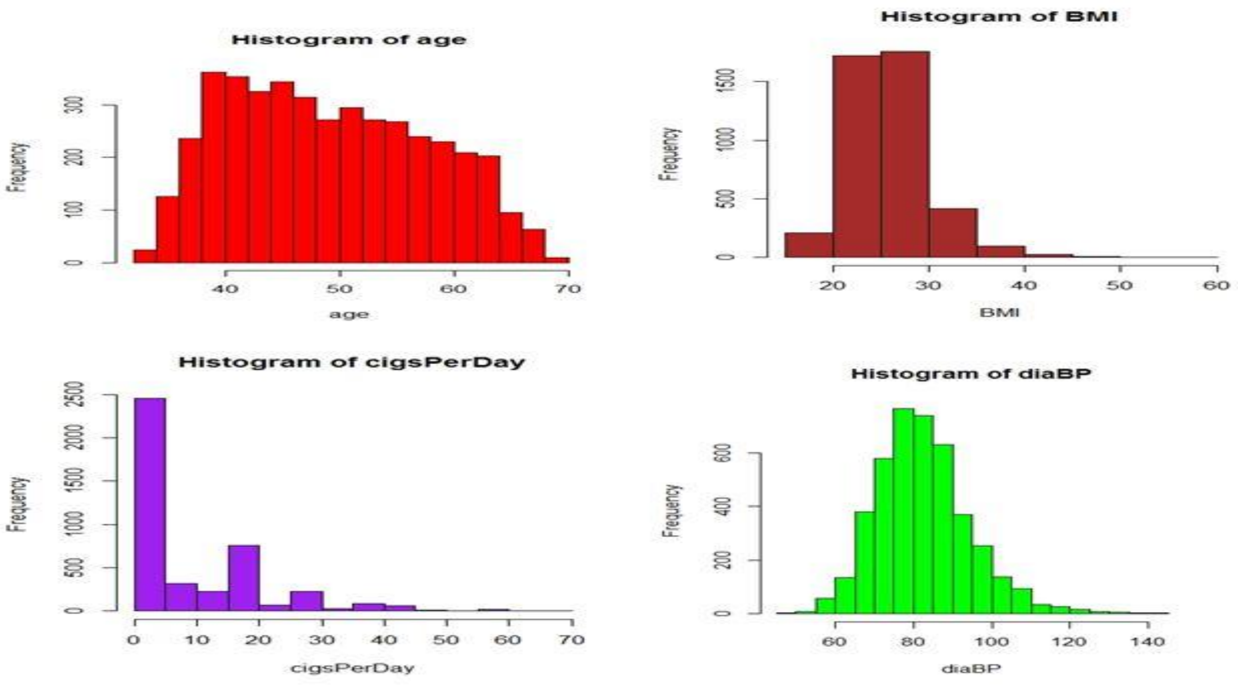


Fig.2: Histogram

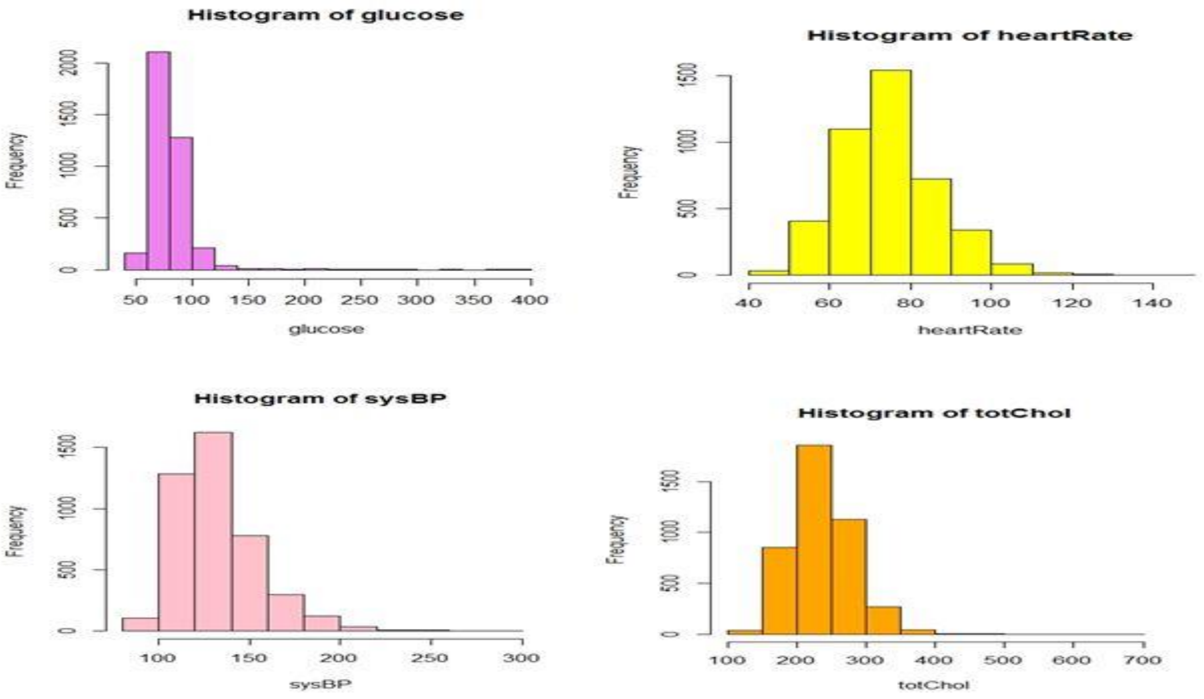


Fig.3: Histogram

Bivariate Analysis:

- Variable “age” is a continuous variable. It is not normally distributed over the entire range of distribution (Fig.4).
- Education v/s age plot shows that the number of patients who have high school education are less then the patients who have Bachelors, Masters and Professioal/PhD (Fig.4).
- The variable “totChol” has a right skewed distribution and it can also be seen from Fig.4 that the patients who are under blood pressure medication or diabetic have relatively higher level of total cholesterol.
- “sysBP” is a right skewed distribution. It can be seen (Fig.4) that the patients who are smokers and hypertensive have high systolic blood pressure.
- “diaBP” is a right skewed distribution. It can be seen (Fig.4) that the patients who are smokers and hypertensive have high diasystolic blood pressure.
- The variable BMI is a right skewed distribution. The scatter plot between BMI and diaBP shows that the patients with high diastolic blood pressure have high body mass index (Fig.4).
- The plot between BMI and sysBP shows that patients with high BMI have high value of systolic blood pressure (Fig.4).
- The plot between “sysBP and “diaBp” indicates that the patients who are having high diastolic blood pressure also have high stolic blood pressure (Fig.4).

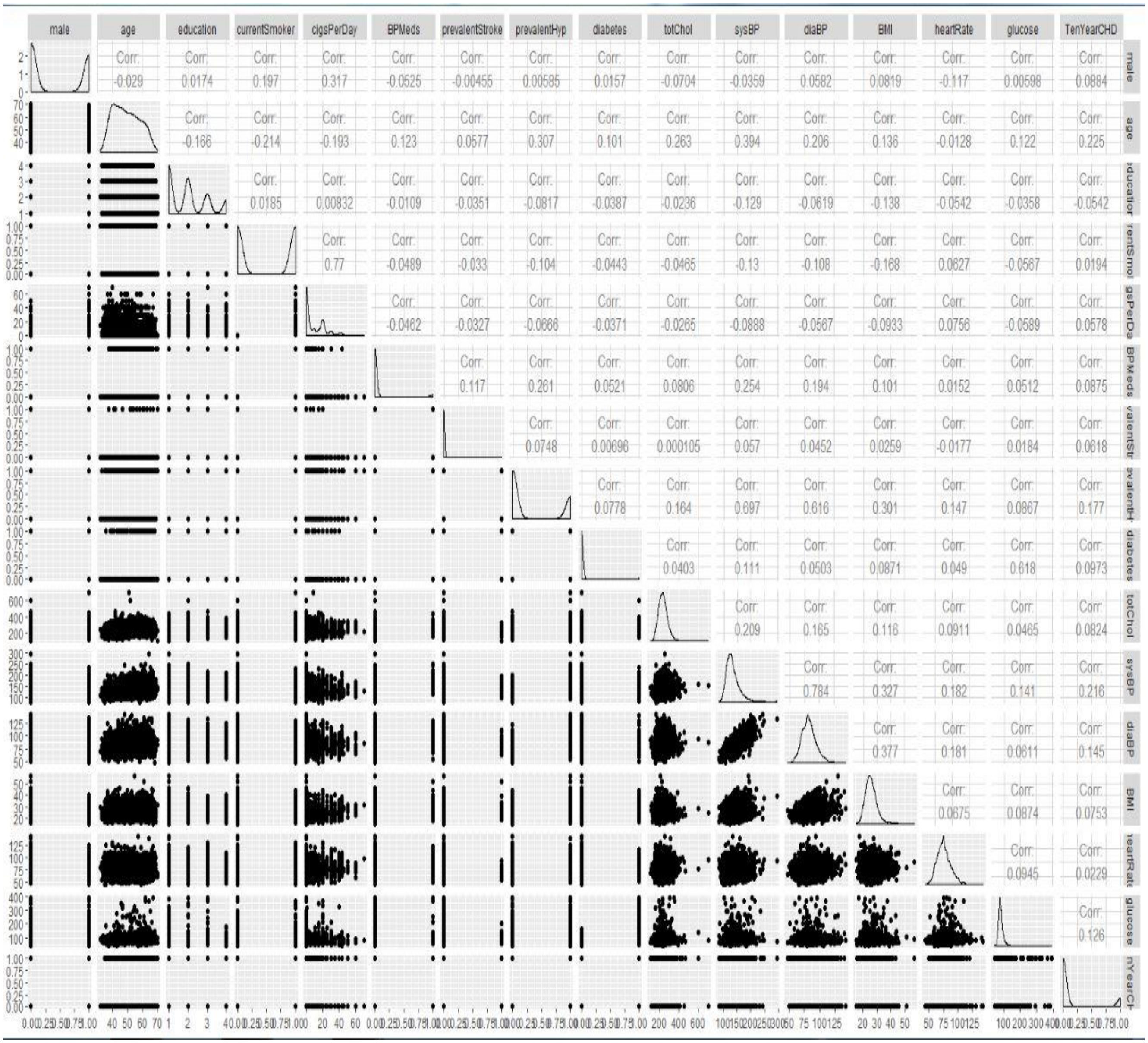


Fig.4: Bivariate Analysis graphs

Correlation:

- For correlation plot, we are considering only the continuous variables viz. totChol, sysBP, diaBP, BMI, heartRate, glucose, age, education and cigsPerDay.
- The correlation plot is shown in the Fig.5.
- From Fig.5 it can be seen that
  - Variables sysBP and diaBP has 78% correlation.
  - Variables sysBP and age has 39% correlation.
  - Variables and age has 26% correlation

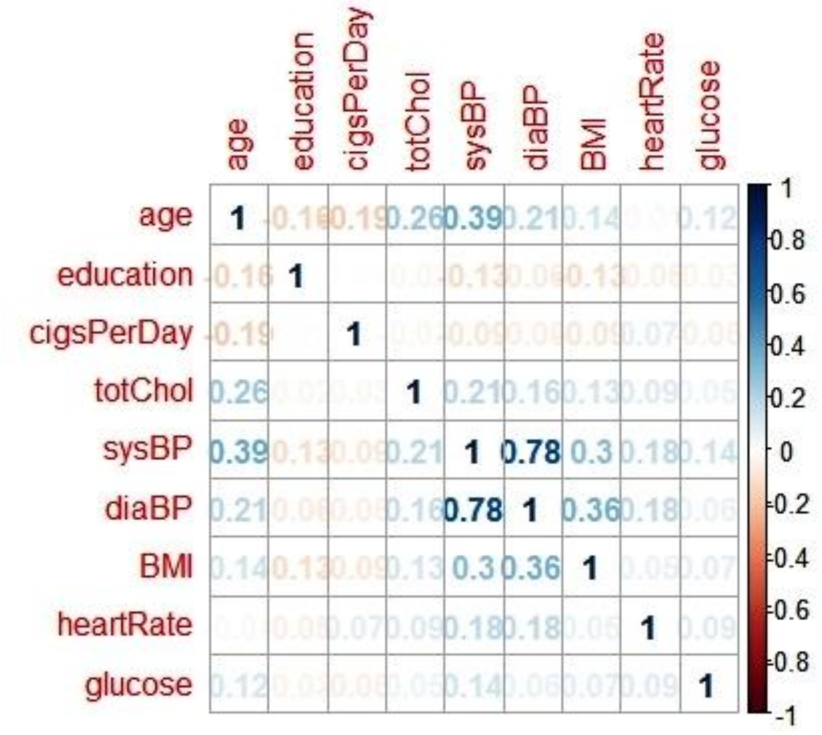


Fig.5: Correlation plot

Removal of Unwanted Variable:

- There are no unwanted variable present in the data set which can be removed.

Missing Value Treatment:

- Plot of missing value is shown in Fig.6
- Variable heartRate has 0.02% of missing values.
- Variable BMI has 0.45% of missing values.
- Variable cigsPerDay contains 0.68% of missing value.
- Variable totChol contains 1.18% of missing values.
- Variable BPMeds contains 1.25% of missing values.
- Variable education contains 2.48% of missing values.
- Variable glucose contains 9.15% of missing values.
- In order to have a good model building, it is required to treat these missing values.
- For treatment of missing values, imputation method is deployed.
- Since the variable education does not contain any outlier (Fig.9), therefore for its missing value treatment, imputation by mean is done.
- Rest all the other variables has outliers therefore to treat them imputation by median is done.
- After imputing the missing values, again the plot of missing values is done to verify whether or not any missing value exist in the data set.
- The Fig.7 shows that there is no missing value in the data set after imputation.

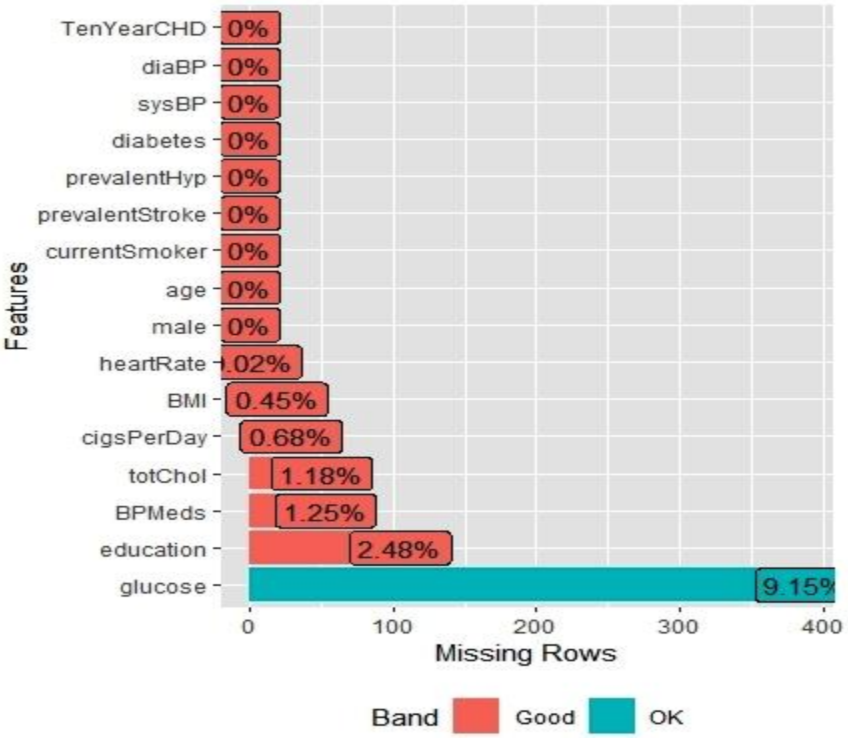


Fig.6: Missing Value Plot before imputation

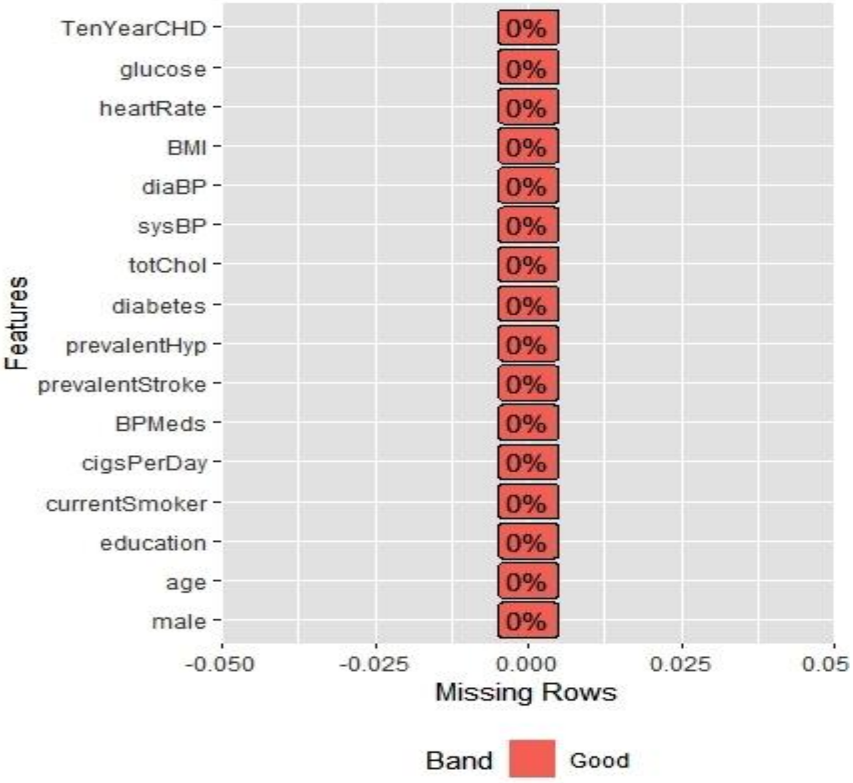


Fig.7: Missing value plot after imputation

Outlier Treatment:

- The boxplot of each variable is plotted as shown in Fig.8, Fig.9 and Fig.10.
- It shows that the variables BMI, cigsPerDay, diaBP, glucose, heartRate, sysBP and totChol contains the outlier.
- The treatment of outliers involves replacing the outliers by capped values meaning replacing those values which lie outside the lower limit by 5th percentile and those which lie above the upper limit by the value of 95th percentile.
- After outlier treatment, boxplot of all the variables BMI, cigsPerDay, diaBP, glucose, heartRate, sysBP and totChol is plotted again to verify whether the outliers exist in the data set or not (Fig.11)

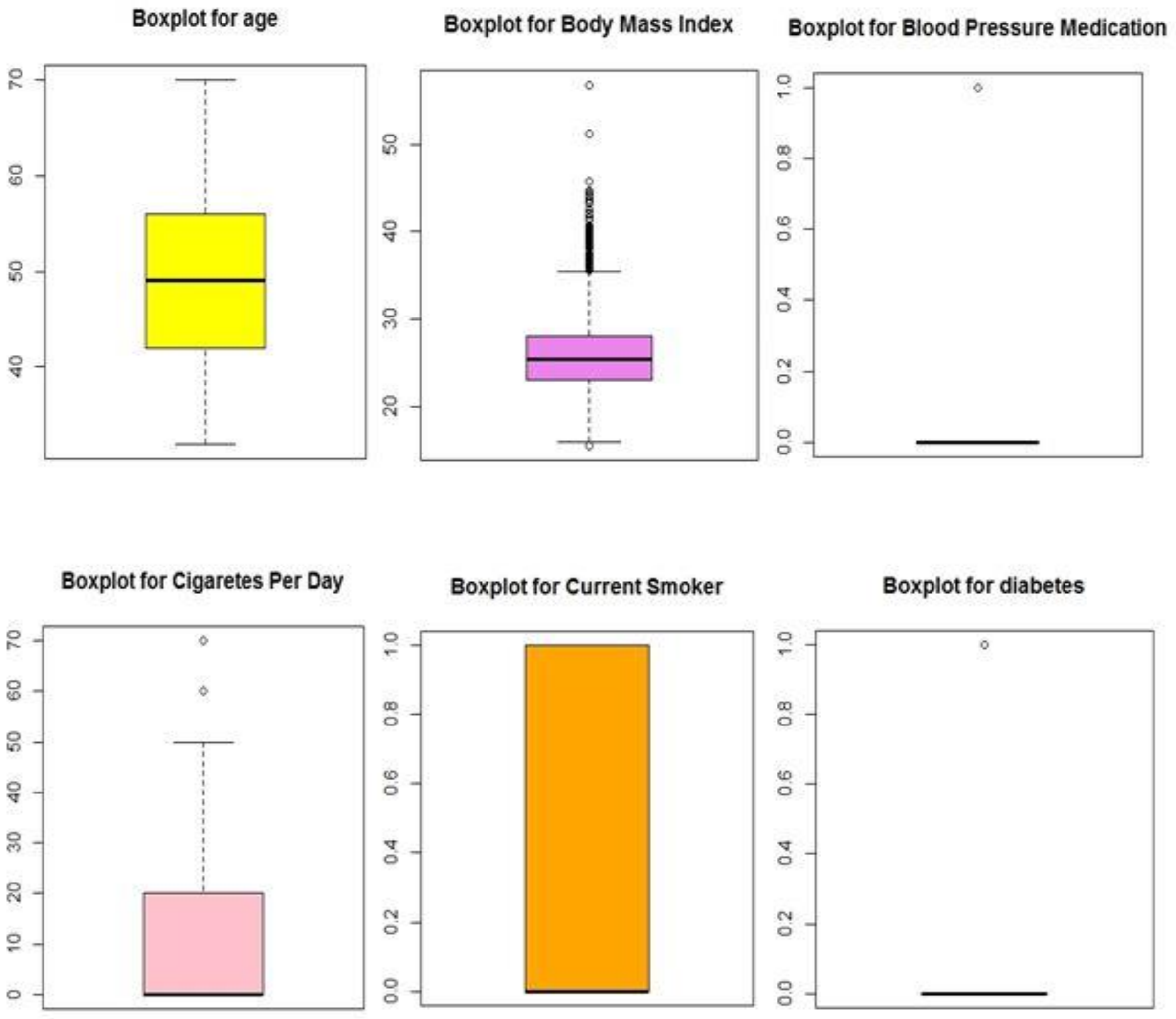
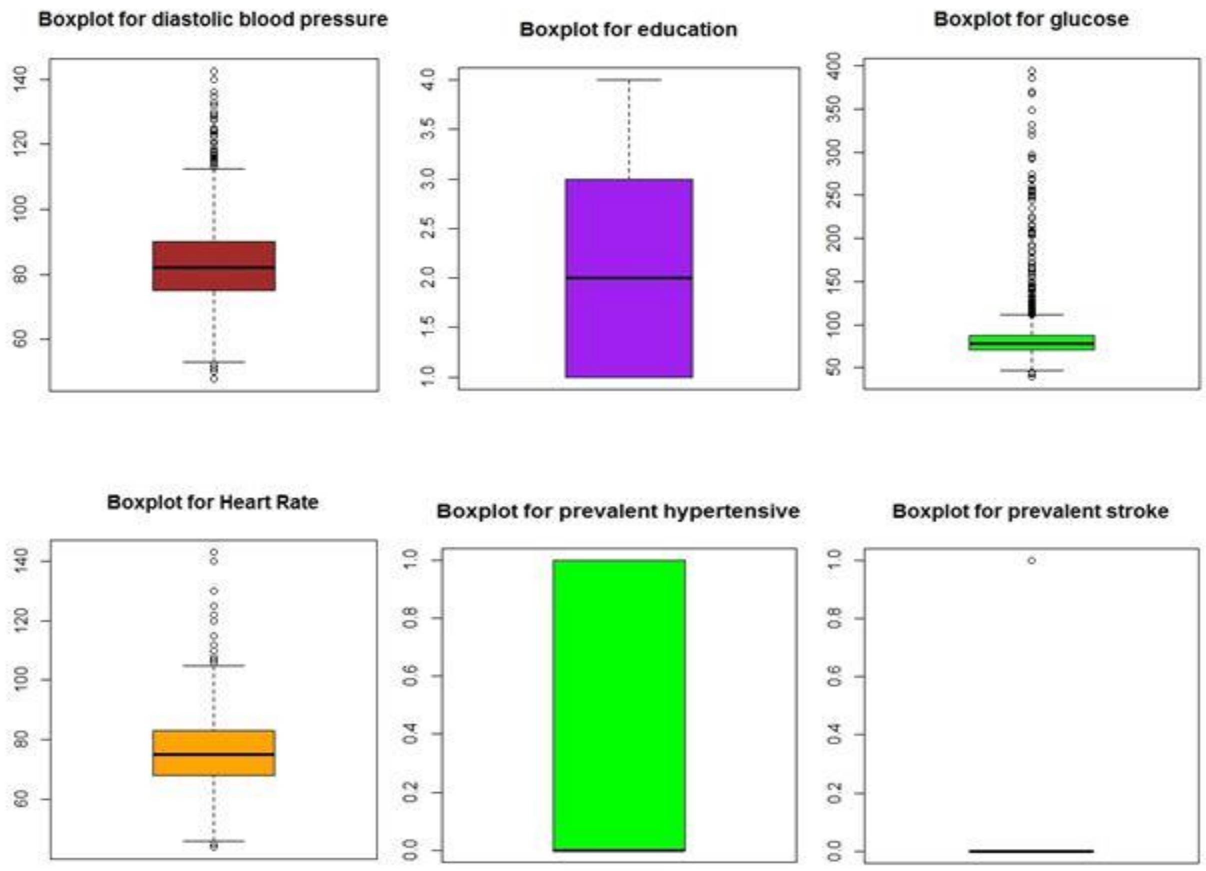


Fig.8: Boxplot before outlier treatment


Boxplot for diastolic blood pressure

Boxplot for education

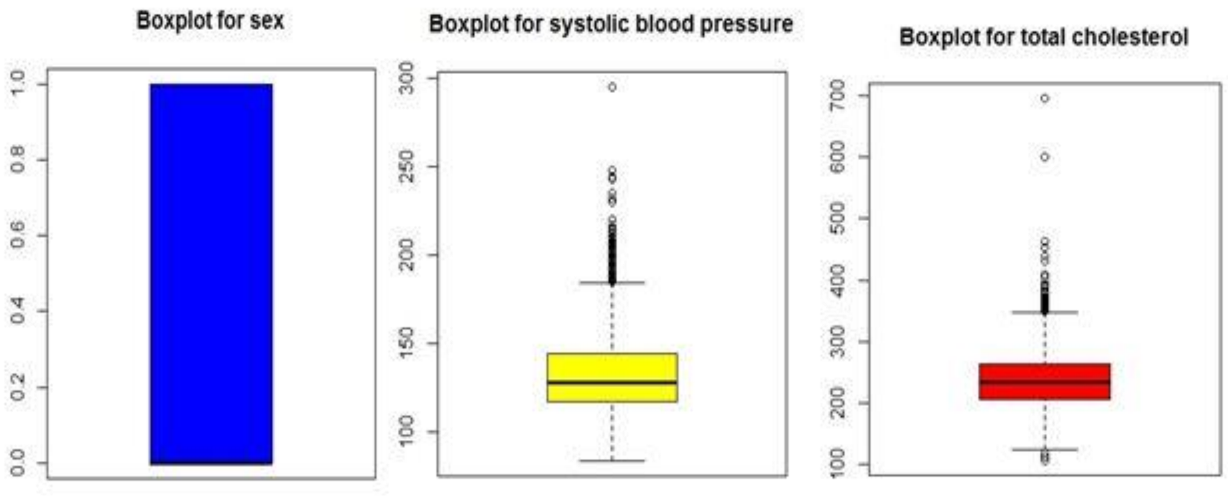
Boxplot for glucose

Boxplot for Heart Rate

Boxplot for prevalent hypertensive

Boxplot for prevalent stroke

Fig.9: Boxplot before outlier treatment


Boxplot for sex

Boxplot for systolic blood pressure

Boxplot for total cholesterol

Fig.10: Boxplot before outlier treatment

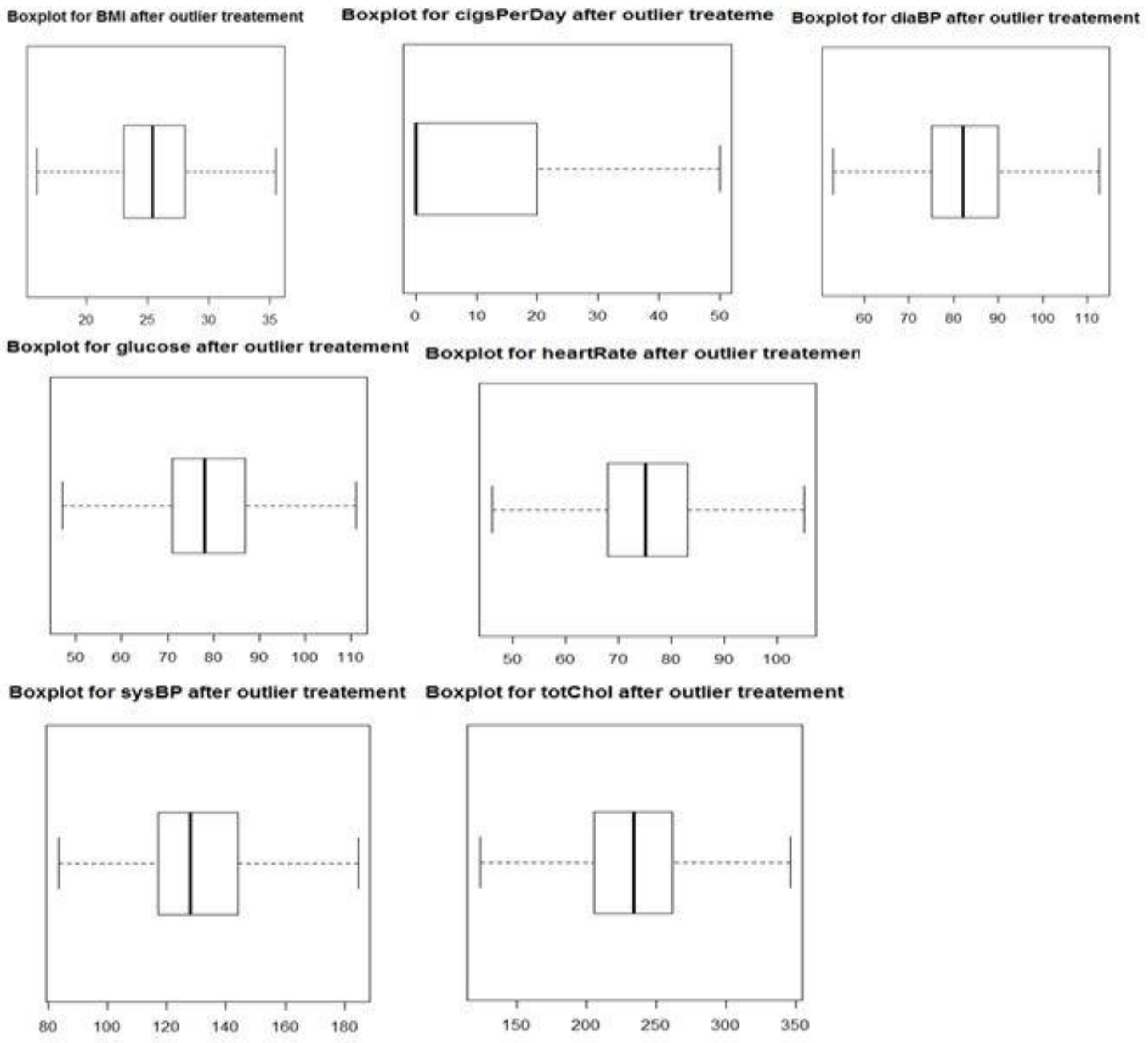


Fig.11: Boxplot after outlier treatment

Variable Transformation:

- Variable scaling is need to be done because the variables like age, cigsPerDay, totChol, sysBP, diaBP, heartRate, glucose, BMI are continuous variables whose minimum and maximum value ranges upto 100 or 200.
- On the other hand, rest other variables are factor variables having output “1” or “0”.
- Therefore it is necessary to bring all the variables at the same scale to avoid any kind of biasing in the model building.
- The variables are scaled with mean = 0 and standard deviation = 1.

Addition of New Variable:

- No new variable is added to the given data set.

4. Insights from Exploratory Data Analysis:

Data Balancing:

- Exploring the target variable i.e TenYearCHD in the original data set received from medical practitioner. It is seen that the data is distributed in 84:20 ratio meaning that the data is imbalanced (Table-2). The data set has a total of 4240 rows.

Table-2: Original distribution of data

TenYearCHD	
0	1
3596	644

- In order to balance the data, SMOTE analysis is deployed. It breaks the data set into 70:30 ratio, which in our scenario looks good (Table-3). SMOTE contains a total data set of 68264 rows.

Table-3: Distribution of data after SMOTE

TenYearCHD	
0	1
48300	19964

Insights from Clustering:

- Since the data set contains large number of data therefore kmeans clustering is used to cluster the similar rows.
- NbClust package is used to identify the exact number of cluster.
- NbClust library shows that the best number of clusters is 2. (Fig.14)
- Number of cluster plot is shown in Fig.12.
- Cluster plot is shown in Fig.13.
- From Fig.15, it can be seen that the mean values of all the Medical (current) variables are negative in cluster-1.
- The mean values of all the Medical (current) variables are positive in cluster-2 (Fig.15).
- Hence, it is concluded that if a patient belongs to cluster-2, he/she has the high probability of getting a heart attack.

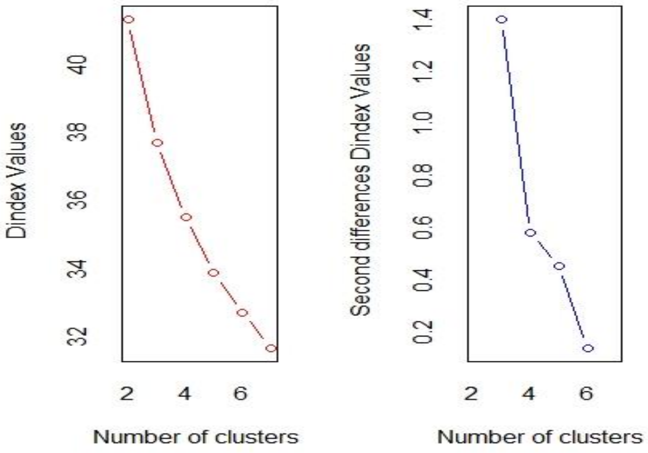
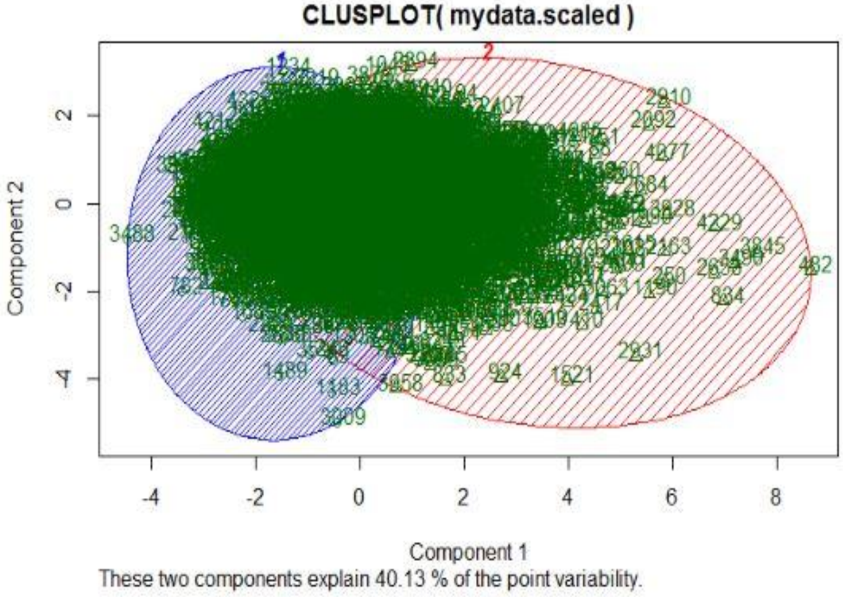
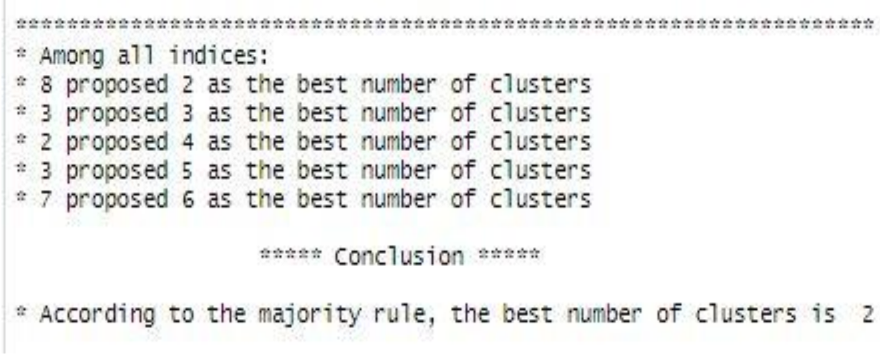


Fig.12: Number of Cluster



These two components explain 40.13 % of the point variability.

Fig.13: Cluster plot

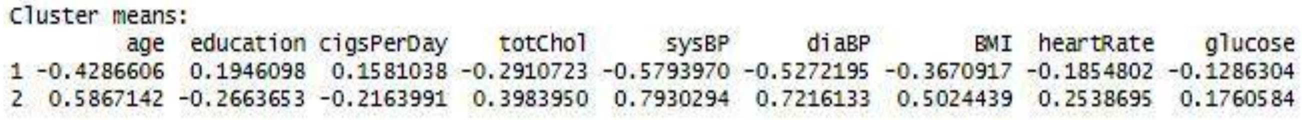


```
*******************************************************************
* Among all indices:
* 8 proposed 2 as the best number of clusters
* 3 proposed 3 as the best number of clusters
* 2 proposed 4 as the best number of clusters
* 3 proposed 5 as the best number of clusters
* 7 proposed 6 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  2
```

Fig.14: NbClust Output



```
Cluster means:
         age  education cigsPerDay    totChol      sysBP      diaBP        BMI  heartRate    glucose
1 -0.4286606  0.1946098  0.1581038 -0.2910723 -0.5793970 -0.5272195 -0.3670917 -0.1854802 -0.1286304
2  0.5867142 -0.2663653 -0.2163991  0.3983950  0.7930294  0.7216133  0.5024439  0.2538695  0.1760584
```

Fig.15: Cluster means

Any Other Insights:

Demographic Variable

- Demographic variable are male, age and education.
- Peoples who are aged above 50 years have high cholesterol level and high value of systolic blood pressure as well as diastolic blood pressure. Hence they are most likely to get heart attack after ten years.
- Education contains 2.48% of the NA values.
- Age is 26% correlated with totChol.
- Age has 39% correlation with sysBP.

Behavioral Variable

- Variables that belong to this group are currentSmoker and cigsPerDay.
- From EDA, it is concluded that the person who smoke on an average of 9 cigarettes per day have higher chances of getting a heart attack.
- Key variables responsible for heart attack when cigsPerDay is considered are heartRate, sysBP and diaBP.

Medical History

- If a patient is under BP medication then prevalentHyp is most responsible variable for getting a heart attack.
- If a patient is diabetic, he/she is most to likely to get heart attack. The important variables are prevalentHyp and prevalentStroke.

Medical (current)

- A patient who is having a high value of sysBP is more likely to have high value of diaBP.
- Patient with high BMI has high heart rate and glucose level.

Recommendation

- Patients with high level of glucose and cholesterol are advised to do exercise regularly.
- Such patients are also advised to maintain the level of sugar by following a proper diet chart prescribed by the dietician.
- Patients who are frequent smoker and diabetic, they are advised to walk daily atleast 5km.
- Patients who are having medical history are advised to take the medicines properly on time and consult with the doctors about their health.