Capstone Project - Final Report
Coronary Heart Disease Study - 2020

Karthik Paranthaman

University of Texas

Contents

Acknowledgement

I take this opportunity to express my sincere gratitude to my capstone mentor Mr. Prashant Verma for providing invaluable guidance throughout the capstone project. I specially want to thank all my lecturers starting from Mr.Theo, Dr.Daniel Mitchell,Dr.P.K Viswananthan,Prof. Kumar Muthuraman and Dr.Abinanda Sarkar whose lectures were highly informative and interesting. Furthermore, my graditude also goes to instructors Mr.Shaam Ramamoorthy and Mr. Vivekanand R in helping me to add Tableau into my data visulaisations skills.

My special thanks goes to Ms. Neha, Mr.Amit, Mr. Michelle, Ms. Sayanthini, Ms. Swetha and all the mentors who taught me throughout this course to make this a fruitful journey.

Last but not the least, I thank the administrative staff of Great Learning, Ms. Nandhini who helped me joining this program and Mr. Prateek being a wonderful course coordinator who patiently answered all my queries and gave me maximum support to complete PG DSBA course from University Of Texas.

Word thank you is not enough to say to my family members who were with me from the begining till now – My lovely mother, My wise Father and My little sister who gave nothing but love and motivation.

Thank you all.

Karthik

Executive Summary

Aim: This document disscusses about the model to predict Coronary Heart Disease (CHD) Risk on people, based on 16 attributes. Various prediction models were built aiming to support healthcare professionals,medical researchers, health ministry decision makers and anyone who is interested in acuring knowledge in risk of CHD. Built models were compared and evaluated interms of validity, accuracy and thier predicting power inorder to select the best model. Furthermore, document is aimed to provide statistical related prevalence, causes and recommendations to overcome the risk of getting CHD based on the identified patterns among the 4000 odd test candidates in the given dataset "Coronary Heart Disease Risk.csv".

Materials & methods: Exploratory data analysis is conducted on the given Farmingham Heart Study dataset. Final dataset was built for prediction after impurity analysis such as missing data and outlier analysis etc were carried on. Machine Algorithm (ML) such as Logistic Regression, Naïve Bayes, K-Nearest Neighbours and data ensemble techniques such as Bagging and XG-Boost were built. R-Studio 4.0 was used to do data manipulation and model building.

Results: EDA-The entire datset were analysed and clustered into three different groups such as "Demographic Analysis" ," Behavioural Analysis" and "Patients health Anlaysis".

Model Prediction - Sensitivity and AUC of each models were considered as prime deciders when it comes to selection of best model. It was interesting to note that all models delivered almost an equal true positive rate (~ 85%).Hence, the algorithm that yielded with a best sensitivity rate (84.9%) and AUC (73.2%) was observed through Logistic Regression ML algorithm.

Conclusions: Demographic Analysis – Majority of the dataset has been dominated by female gender (55.36%).Patients with good educational background (4 years) show a very low risk rate (12%) of getting CHD. **Behavioural Analysis** – Among the risk test patients it was found that majority of them are smokers which supports the Farmingham Heart Study "modyfying risk factors" outcome. **Patient's Health Analysis** -cluster helped to identify the test patients who have a healthy lifestyle by investigating on their historical medical information such as blood pressure medicine intake and experience of stroke, hypertension and diabetes.

Furthermore, this document consists of useful recommendations for the potential stakeholders such as officials from health care sectors, company management who is interested on staff long term health and well being and to all the ML engineers involved in end consumer product development.

1.0 Introduction

Coronary Heart Disease (CHD) is one of the leading cause of deaths worldwide. Approximately, 3.8 million men and 3.4 million women die every year due to CHD. Researches on CHD show statrtiling statistics that developed countries like Europe accounts about 95 million deaths per annum.

However, this disease has been identified as modern epidemic by the World Health Organization (WHO) and it had been diagnoised at United States patients at early 1920s and 1930 in Britain.Furthermore, other countries with advanced economy also started to recognize CHD patients among their popoulations. Now even the developing countries are catching up with CHD. However, USA,Australia,Canada and New-zealand started to show a flat curve in the mortality rate during late 1980. The reason for this decline is not known but WHO took reactive measures by starting an awareness program called MONICA "(Multinational monitoring of trends and determinants in cardiovascular diseases)" to control the risk of CHDs. Evidently, constant monitoring and taking ncessary measures to keep risk of CHD under control has become a social responsibility.

1.2 Background Theory

1.2.1 Anatomy & Plaque Development

Heart is a constant working muscle which pumps blood throughout the body with about 100,000 beats per day. Have you realised why we breath heavily after running or jumping or doing any physical activities?. Here's why, heart is responsible to supply oxygen thorugh blood to the respective muscle motor systems to get them activated to complete a particular physical task. Once the work gets completed the blood pump rate of the heart will be brought back to the normal value. In order to gain the oxygen that being utilised at a task, will be abosrbed into the blood stream through heavy breathing, which is supported by the exapnsive nature of the lungs. However, the heart also made out of muscle tissues which require oxygen for that muscle to work efficeiently. If a heart is sliced into two and viewed at a cross section, it can be seen that the blood flows inside heart will not get diffused through the interior and the exterior heart muscle due to its thickness. That's why coronary arteries make sure to supply oxygenated blood into the heart muscle. Coronary arteris are divided into two categories based on their position of location. Coronary arteris which supply pure blood to right heart muscles are known as right coronary arteris and arteris which supply blood to the left muscle is known as left coronary descending arteries.

In general blood stream contains various particles such as choloestrol, white blood cells, and red blood cells. As I mentioned in the beginning of this section, heart starts its job the moment a life was born and constanly pumps blood thorugh out the life with various rates. There are certain occasion blood needs to be pumped against the gravity to the brain. The blood vessels goes under very high strain due to the fluctuated blood pressure and high shear stress damages the interior blood vessel walls. Increased fatty food intake and sendentray lifestyle will rise the LDL in the blood stream and these LDLs develope a sedimentation at the damaged walls cavity. White Blood Cells (WBC) tries to clear these deposit and during this process WBC dies. This process is called atheroscleroris. The combination of choloestral and dead WBCs sedimentation called plaque and over the period it becomes clinically significant.

CHDs can be classified into two. One of it is Stable Agina. This is due to the normal palque development disrupting the blood flow rate to activate a particular heart muscle motor system to pump blood to complete a necessary physical task. Slight or severe pain with tightness in the chest could be experienced by the patient depending on the size of the plaque and the difficulty of the activity. A greek term *"Angina Pestoris"* is used in the medical industry as symptoms to identify this diesease. *("Angina"- tightness* and *"Pestoris" – Chest)*. In result the particular heart muscle becomes hypoxic. The pain will not be felt when

there is no necessitiy of high blood rate. In other words at rest the pain will not be incuded, hence this type of CHD is known as reproduceable.

The second type of CHD is knonwn as acute coronary syndrome. This syndrom has two subdivision of unstable agina and heart attack. Unstable agina syndrome is caused due to the spontaneous rupture of the plaque. Once, the the size of the plaque gets developed and due to the sudden rise in blood pressure the part of the plaque gets detached from the blood vessel and flaps inside the blood arteries causing a disruption to the regular blood flow. The theromogenic nature of the plaque will attract blood cells around the plaque after rupture and becomes a blood clog in the vessel. Reduced supply of blood flow into down stream heart muscle will induce a pain even when patient is at rest. Lack of blood flow will make the particular muscle hypoxic. In addition to that thermogenic nature of the palque creates a high shear stress at blood vessel.

The heart attack happens when blood clog knownsas thorombus gets completely detached from the flapping or attached plaque and mixed into the blood stream. This seperation process of thoroumbus is known as embolization. Embolous float downstream and block the artery down stream completley as the diameter of the blood vessel of the aretery gets smaller as it moves down. Patient will experience a severe pain around the chest and he/she must put under medication within 20 minutes to save that heart muscle from permanently dying.

1.2.2 CHD Risk Factors

"So, what causes Coronary Heart Disease in humans?...". In 1940 Framin Heart Study was conducted and various risk factors were identified. "Non-modyfying " and "modyfying" factors are two types of risk factors.

Non-Modifiable CHD Risk Factors:

Non-modyfiable factors like are risk factors that cannot be controlled by change in life style modification or medication.

"Family History" the genetic history influences the deveopement of plaque in arteris.

"Ethnicity" – Test subject used in the study have proven that african community have very high risk of having CHD.

"Age" – After a certain age risk of having CHD increases.

"Gender" – Men have high risk than women. This findings were supported by interesting behavioural arguments such as most men live very unhealthy lifestyle and women have tendency of listening and following doctors and healthcare professionals advice. Inaddition, to that estrogen hormone also rdeuces the risk of CHD.

Modifiable CHD Risk Factors:

Modyfing risk factors can be kept under control by following certain healthy tips. Idenfied modyfing risk factors are *"high level of LDL","hypertension","smoking","diabetes","physical activities"* , *"physical activity"* and *"drug consumption".*

"High LDL"- Triglycerides or certain types of LDL fats increase plaque formation in the damaged blood vessels.

"Hypertension"- High blood pressure damages the blood vessels and this will create opportunity for LDL to form a sedimentation.

"Smoking"- Another research on tobacoo consumption tells that cigarate contains 400 different toxic chemicals. Ingesting these toxic materials into our blood stream will damage the blood vessels. As discussed before,damaged blood vessel are more prone for plaque formation.

"Diabetes"- The risk of getting CHD is 2-3 times higher among the diabetes than the non-diabetes.CHD is responsible of 30% to 50% of death in developed countries.

"Physical activity"- Regular physical activity will stabilize the glucose level and the LDL will be mobilized and burnt. This makes the overall choloestral profile better. Another interesting factor is physically active person will be less likely to be a smoker.

"Drug consumption"- Cocaine and Amphatamine conusumption causes the coronary arteries to clamp down and compromises on downstream blood flow. This clamp down of coronary arteris is called vasospasm.

2.0 Assumptions

All the non-measureable information like "smoking habit" and "experiencing a stroke in the past "and "Undergoing medication for certain clinically significant syndromes" are true and accurate. There are certain missing values were observed. No missing value treatment algorithms were applied due to the uniqueness of every test patients. Therefore, only test patients with all information were considered for this analysis by protecting the purity of the dataset.

Furthermore, measureable information such as glucose and cholesterol checkups were done followed by proper fasting blood sugar methodology recommended by doctors in order collect accurate information.

2.1 Data Collection

All patients were tested on the same day to build a model to predict a risk of CHD on patients in 10 years from the time measurements were taken. Furthermore, all the information have been collected directly from patients' health record from the same hospital.

3.0 Materials & Methods

Preliminary EDA was carried out using R-Studio 4.0 to understand the given dataset. Furthermore, to get an in-depth understanding of the problem, test patients were categorized based on their demography, behavior and medical history. The upcoming data preparation chapter discusses steps taken. The data was split in to 70:30 ratio and used the larger dataset to train the model. Also, the prediction power of each models were investigated against test dataset.

Secondly, comparative methodological study was carried out between the famous supervised classification ML algorithms- Logistic Regression, Naïve Bayes and KNN. In addition to that bagging and XG-boost data ensemble methods were also put to test in building models in R-studio 4.0 platform.

Thirdly, the built models were evaluated and compared using following parameters: accuracy, sensitivity, specificity, concordance & discordance intervals, area under the curve (AUC) and KS-Statistics.

3.1 Data Preparation

A descriptive data analysis was conducted by investigating the nature of the variables: mean, median, range, absolute & relative frequencies and quantiles. The attribute values names were converted into a meaningful names and their class types were corrected as shown below.

Attributes	Class
Gender	Factor
Age	Numerical
Education	Factor
Smoker	Factor
Ciggarattes_per_Day	Numerical
BP_Medication	Factor
Prevalent_Stroke	Factor
Prevalent_Hypertension	Factor
Diabetes	Factor
Total_Cholestrol	Numerical
Systolic_BP	Numerical
Diastolic_BP	Numerical
BMI	Numerical
Heart_Rate	Numerical
Glucose	Numerical
10_Years_CHD	Factor

As mentioned in the "Assumptions" section of this document, a "complete.case" function was used to filter observations with 100% information. This is mainly due to the outcome of the missing value analysis have shown no pattern. All missing values are completely at random (MCAR) and to make prediction with pure data.

The attribute selection was done in model building using "stepAIC" function from MASS library and "optimal cutoff" value was calculated using InformationValue library.

3.2 Software Used

R-studio 4.0 was used for all analysis. Following are the libraries used for ML algorithm building:

ALGORITHMS	Library Function
Logistic Regression	caTools
KNN	ISLR
Naïve Bayes	e1071
Bagging	ipred
XG Boosting	xgboost

Refer to the *Appendix* for the complete code.

4.0 Results & Discussion

4.1 Univariate Analysis

The table below illustrates the summarize version of the univariate analysis on each attribute from the given dataset.

Categorical	Categories	Number	%				Missing
Gender	Male	1623	44.36				0
	Female	2035	55.36				
Education (Years)	1	1526	41.72				105
	2	1101	30.1				
	3	608	16.62				
	4	423	11.56				
Smoking Habit	Smoker	1789	48.91				0
	Non-Smoker	1869	51.09				
BP Medication	Yes	111	3.03				53
	No	3547	96.97				
Prevalent Stroke	Yes	21	0.57				0
	No	3637	99.43				
Prevalent Hypertension	Yes	1140	31.16				0
	No	2518	68.84				
Diabetes	Yes	99	2.71				0
	No	3559	97.29				
Continuous	**Mean**	**Median**	**Max**	**Min**	**Q1**	**Q3**	**Missing**
Age	49.55	49	70	32	42	56	0
Cigarettes per Day	9.03	0	70	0			29
Total Cholesterol (mg/dl)	236.85	234	600	113	206	263	50
Systolic BP (mg of Hg)	132	128	295.0	83.5	117	143.88	0
Diastolic BP (mg of Hg)	82.9	82.9	142.5	48.0	75	90	0
BMI	25.78	25.38	56.80	15.54	23.08	28.04	19
Heart Rate (BPM)	75.73	75	143	44	68	82	1
Glucose (mg/dl)	81.85	76	394	40	87	71	388
Predicted Variable	**Categories**	**Number**	**%**				**Missing**
10 Years CHD	Risk	557	15.22				0
	Safe	3101	84.78				

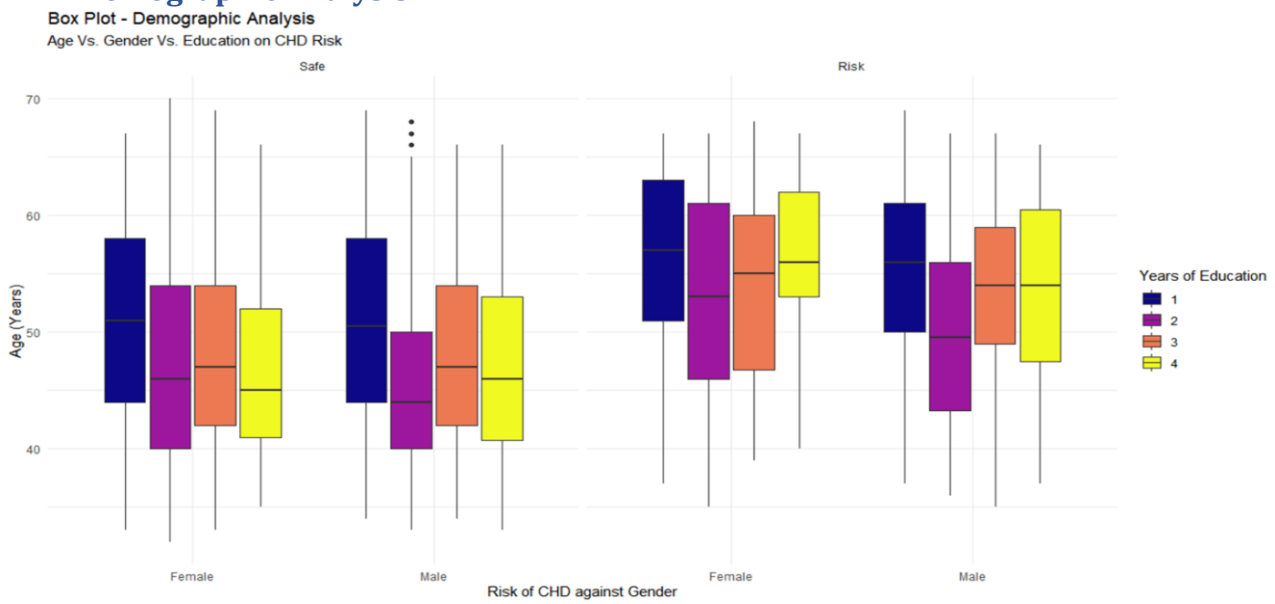
It can be seen that majority of the test patients are female and with at least 1 year of educational background. Referring to the "*Framingham Study*" outcomes that smoking habit been one of the prime causes of the CHD, the given dataset is almost a balanced dataset if we divide them based on their smoking habit. This is an indication at which there are various other parameters influencing the predicted variable "10 Year CHD". Hence, the variables will be grouped and analyzed in the upcoming chapters.

The most of the missing values were observed in glucose level information (388). The second most variable contains missing values is the education level of the patients (105). However, missing value analysis were conducted and found to be completely at random. Refer to the A*ppendix - B – Missing Value Analysis.*

4.2 Multivariate Analysis

The given attributes have been categorized into different clusters based on their nature. Attributes Age, Education and Gender are grouped and titled as Demographic Information. Smoking Habit and Cigarettes per day attributes are grouped as Behavioral Information. Rest of the attributes BP medication, Prevalent stroke, Prevalent hypertension, Diabetes, Total Cholesterol, Systolic BP, Diastolic BP, BMI, Heart Rate and Glucose are grouped as Patients Health Information.

4.2.1 Demographic Analysis



Male and female patients age distribution is plotted through their educational history and clustered based on the risk factor of getting a CHD in ten years' time.

It can be observed that safe female test patients with 2 and 4 years of educational background have their mean age equal. In addition to that, males with CHD risk with 3 and 4 years educational background have equal mean age. Furthermore, both gender risky patients have higher mean age than the safe patients. Hence, this supports "*Framingham Study*" at which the age is a non-modifiable factor which increases the chances of getting CHD as the patients gets older.

The following table illustrates the risky female average age based on the acquired level of education.

Educational Level	Risky Female Average Age
1	56.38636
2	53.35385
3	53.675
4	56.69231

It's interesting to note that all of the potential CHD female patients with all four level education have the mean age more than 50. However, the below hypothesis will be verified with the help of a chi-square:

H_o : Mean age at all level of educational background among risky female are EQUAL.

H_A : Mean age at all level of educational background among risky female are NOT EQUAL.

PARAMETERS	VALUE
df	3
p-value	0.9826

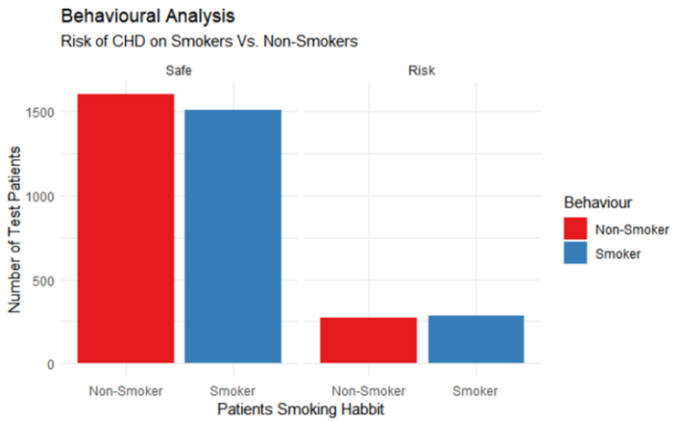
Since, the p-value is more than the 5% alpha value the null hypothesis can be accepted. Therefore, the average mean age of risky female from four different level of education is equal.

4.2.2 Behavioral Analysis

Patients CHD risk factor is analyzed against the independent variables such as Smoking Habit and Cigarettes per Day. Below bar chart tells about the CHD risk factors on test patients who have a habit of smoking and avoiding cigarettes.

Classification	Risk	Safe
Overall	557 (15.23%)	3101 (84.77%)
Smoker	285 (7.79%)	1504 (41.12%)
Non-Smoker	272 (7.44%)	1597 (43.66%)

The overall CHD risk patient is about 15.23%. However, 7.44% of non-smokers have risk of having CHD in 10 years from the time of testing.

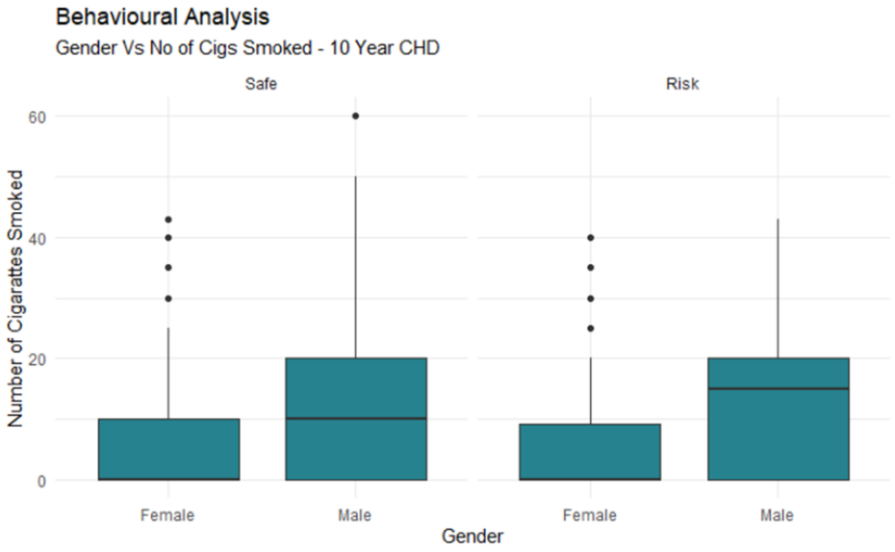


It is somewhat relaxed to know that the total percentage of safe nonsmoking patients are higher than the CHD risk smokers by 0.35%.

This finding supports the *"Framingham Heart Study"* that smoking is one of the major cause of damaging the blood vessels which could potentially expediate the development of palque in the coronary arteries.

4.2.2.1 GENDER BASED BEHAVIOURAL ANALYSIS:

The second level of slicing is conducted by adding gender into the behavioral analysis. The box plot below indicates the spread of number of cigarettes smoked by test patients.



Interesting note that 3rd quartile of both risky and safe test patients among both gender have same number of cigarettes been smoked. As per the distribution the male test patients are dominating both on safe / risk factors of CHD. Hence, the average cigarettes smoked is calculated and tabulated below.

Male Patients	Average Cigarettes Smoked
Safe	13.16945
Risk	14.57003

The Hypothesis testing using Chi-square on the equality of the average cigarettes smoked.

H_o : Average number of cigarettes smoked by males with SAFE and RISK of getting CHD are EQUAL.

H_A : Average number of cigarettes smoked by males with SAFE and RISK of getting CHD are NOT EQUAL.

PARAMETERS	VALUE
df	1
p-value	0.7903

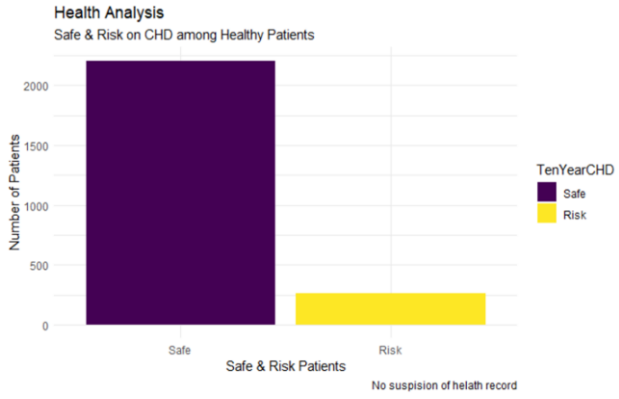
As per the p-value of hypothesis testing a null hypothesis cannot be rejected. Hence, the average number of cigarettes smoked by CHD safe male and CHD risk male are 79.03% equal value.

4.2.3 Patients Health Analysis

Patient health cluster is divided into two sub categories one is "Patients Medical History" which includes BP medication, Prevalent stroke, Prevalent hypertension and Diabetes attributes. The second cluster is created based on the present status of the test report such as Total cholesterol, Systolic BP, Diastolic BP and BMI.

I assumed that patients with no historical record of taking BP medication, stroke, hypertension and diabetes are considered to be living a healthy life style. Below table talks about the patients risk factor of getting CHD in 10 years among who live a healthy lifestyle.

Parameters	SAFE from CHD	Risk of CHD
Total Dataset	3101 (84.77%)	557 (15.23%)
Healthy	2206 (60.31%)	260 (7.11%)



Therefore, out of 84% of the patients with no risk of getting CHD in 10 years. At which 60% of them have a clear medical history.

Hence, this proves that aforementioned patient's historical data have big influence in determining the CHD risk factor.

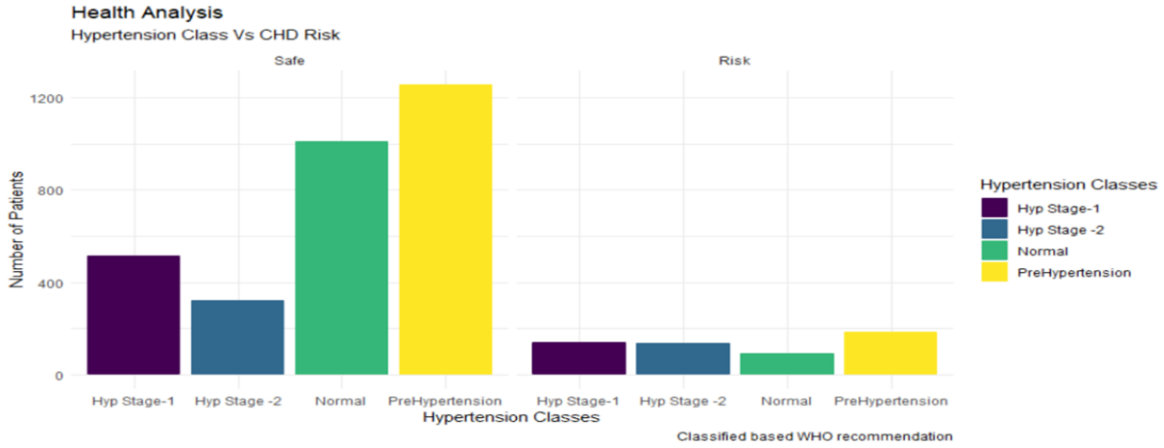
Bar plot illustrates the visual representation of the findings.

WHO states that healthy patient's systolic and diastolic heart pressure must be less than 120/80 mm of Hg. Anything above is known as symptoms of hypertension. The below chart is recommended range and classification of the stages of hypertension based on the measured systolic and diastolic blood pressure.

Category	Systolic Blood Pressure (mm of Hg)	Diastolic Blood Pressure (mm of Hg)
Normal	<120	<80
Prehypertension	120-139	80-90
Hypertension		
Stage - 1	140-159	90-99
Stage - 2	≥160	≥100

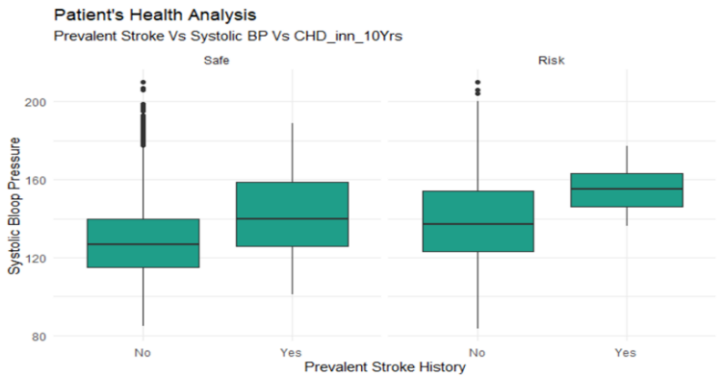
Below table illustrates the count of each classes of hypertension from the given dataset. Majority seems to be safe patients with prehypertension stage. At the same time 93 patients with below 120/80 mm of Hg systolic BP and diastolic BP also have been recognized as potential CHD patients. This is probably the influence of other attributes in the dataset.

CHD Risk Status	Normal	Pre-Hypertension	Hypertension Stage-1	Hypertension Stage-2
Risk	93	186	140	138
Safe	1011	1256	514	320



Bar plot shows the graphical representation of the potentail CHD and healthy test subjects with various hypertenstion.

4.2.3.1 Analysis on Prevalent Stroke, Systolic BP and 10_Years_CHD:



In this section certain assumptions needs to be taken with the importance of the variables. From the patient's medical history, the attributes like BP medication and prevalent hypertension and diabetes are good indicators of CHD. However, the prevalent stroke medical historical data reveal information on the presence of "thoroumbous" in the down side arteries.

This also increases the probability of various other plaque sedimentation in the upstream blood arteries. Furthermore, the systolic blood pressure of the heart is directly involved and disturbed by the plaque formation in the heart muscle arteries. Hence, investigating the possibility of not getting the coronary heart disease in 10 years' time against the aforementioned variable may give useful insights to proceed the analysis.

Patient's Health Analysis	Average Systolic Blood Pressure
No Prevalent Stroke with a Risk	139.8663
Prevalent Stroke and Safe	142.3182

The mean systolic blood pressure of safe patients with historical stroke experience seems equal to the inexperienced stroke patients who have been highlighted as CHD risk patient's systolic blood pressure. However, this finding needs to be verified with the help of hypothesis testing.

H_o: Mean systolic BP of Risk / Safe with historical stroke experience/inexperience patients have EQUAL value.

H_A : Mean systolic BP of Risk / Safe with historical stroke experience/inexperience patients are NOT EQUAL.

Parameters	Value
df	1
p-value	0.884

Above table shows that null hypothesis can be accepted with the p-value of 0.884.

4.3 Model Building

Each of the Machine Learning algorithm have demonstrated different types of predicting power. The summary and a comparison on test dataset is shown in the table below.

PARAMETER	LOGISTIC REGn	KNN	NB	BAGGING	XGBOOST
Accuracy	84.78%	85.14%	84.78%	84.41%	84.87%
Sensitivity	84.9%	85.48%	84.97%	84.91%	84.95%
Specificity	50.00%	62.5%	50.00%	30%	66.67%
Concordance	73.2%	51.25%	63.72%	48.7%	67.97%
Discordance	26.8%	48.75%	36.28%	51.3%	32.03%
AUC	73.2%	61.57%	65.55%	67.08%	67.97%
KS	35.62%	16%	24.4%	31.65%	29.34%

- **Accuracy** and **sensitivity** of the all models fluctuate at a decent score of around 84.5 %.
- **Specificity** of the model indicates the percentage of misclassification of safe patients into a risk patient's category. Highest specificity score 66.67% is obtained through XG boosting algorithm.
- **Concordance** and **Discordance** interval states how good the selected optimal cutoff value against the model classification. Logistic regression shows a high score of 73.2% of concordance interval compared to other models. XG boost algorithm have also returned a noticeable score of concordance interval (i.e.: 67.97%)
- **AUC** parameter tells the classification power of the model. As per the above comparison, only logistic regression have showed a highest score of 73.2% which is considered to be FAIR in statistics. Moreover, all other models have returned an AUC value between 60% - 70% which is an indication of a poor model.(i.e.: Bagging & XG Boost)
- **KS** statistics more than 50% is considered to be a good model. None of the model have a KS value more than 50%. Once, again the logistic regression model have scored the most KS value which is about 35.6%.

Based on the overall findings it can be said that logistic regression and XG boost have performed pretty well. Evidently, logistic regression have overpowered XG boost in all the parameters except the specificity. Considering the seriousness of the topic, I have put sensitivity and AUC-the global criterion in evaluating the predicting power of the model is the AUC, as prime deciders in selection of the best model. Therefore, I recommend the customer to proceed with the logistic regression ML algorithm when it comes to end product development and deployment.

5.0 Conclusion

As stated in the previous project notes it can be seen that the test patients who were safe from getting CHD in 10 years and do not smoke dominates the risky smokers. This "behavioral analysis" outcome on the dataset supports the *"Framingham Heart Study"* that smoking is one of the root cause of getting coronary heart diseases.

It can be found from "patient's health analysis" outcome that test candidates with no historical record of BP medication, stroke, hypertension and diabetes are assumed to be living a healthy lifestyle. Evidently, the dataset is dominated by safe patients with healthy life style. Furthermore, it's understood based on the dataset that diastolic BP and systolic BP are closely related with potential of developing plaque in the arteries. Therefore, the sedentary life style must be changed and constant physical activities must be carried on.

6.0 Recommendations

6.1 For Stake holders:

As found that, smokers are prone to develop coronary heart disease. Studies state startling results on consumption of tobacco among active people are less likely to smoke. Hence, companies with sedentary life style staff must take initiatives to create opportunities to do physical activities by providing subsidiary gym facilities.

Government must create awareness among citizens by building free fitness parks for public. Singapore government's free physical park concept at each living complex can be taken as a very good example for encouraging the people to live actively and to protect them from such coronary heart diseases.

6.2 For Data Collection Team:

The reliability and the quality of the data needs to be verified. It can be seen that some of the information collected such as glucose level seems very extreme for a human being to have without not been hospitalized.

The number of observations will also decides the predicting power of the ML models. Hence, data collection team could redesign the data collection strategy to gather more number of accurate test patient's information to improve the accuracy, sensitivity and specificity of models.

Furthermore, it's worth considering including other independent attributes such as LDL –cholesterol, physical activity and drug consumption related information in analysis. Especially, the patient's family history and ethnicity could offer different perspective to the investigation at which an early *Framingham study* shows high chances of CHD among African test subjects. Hence, incorporating genetically related parameters to the study is recommended.

Appendix – A

EDA Analysis

In order to understand the given dataset, a univariate analysis is carried on all attributes after removing all the missing values. That will bring down the dimension of the dataset to be 3658 rows with same number of 16 attributes of columns.

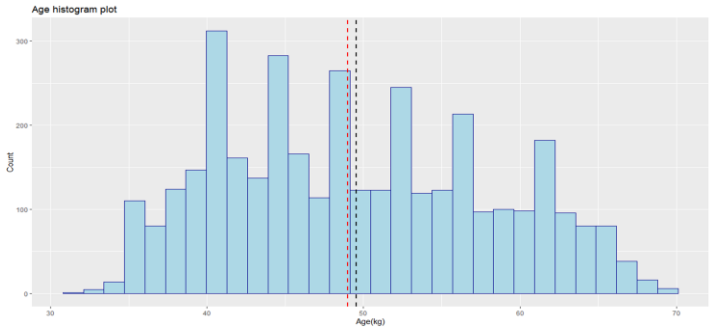
GENDER:

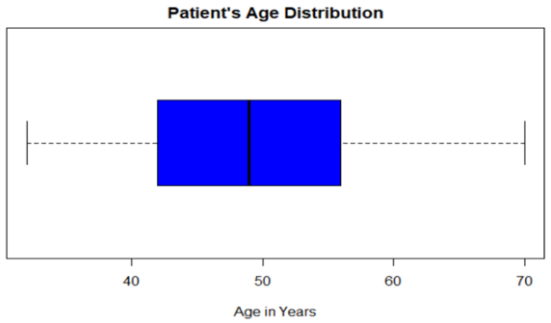
Total number males in the given dataset to be 1623 and female to be 2035.

GENDER	Count	% of Total
Male	1623	44.36
Female	2035	55.36

AGE:

As you can see the 5 number summary of the "Age" variable have a mean of 49.55 and median of 49 while the minimum value being 32 and maximum being 70. By theory if the mean is greater than the median, would make the distribution to be the Right Skewed. However, the difference between mean and median is not very noticeable (i.e: 0.55). In the histogram plot shown below it can be said that "Age" variable observation shows peaks between 40 to 50 age group with the count approximately of 270 to 320. Let's analyze the outliers by doing an outlier examinations.





As observed from the "Age" boxplot, indicates that there are no outliers. The first quantile value is 42. This indicates that 25% of the age group is around 42 and 75% of the age group is below 56.

EDUCATION:

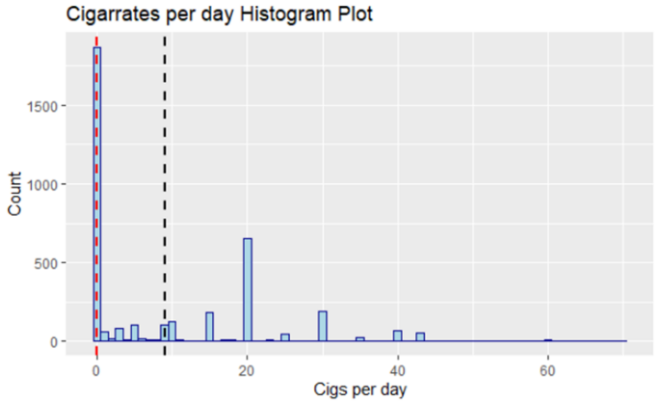
Patient's years of education level can be identified in the dataset as 1, 2, 3 and 4 years. Hence, the attribute class was changed to be a factor. Majority of the patients (i.e: ~42%) have 1 year of educational background while approximately 12% have 4 years educational history.

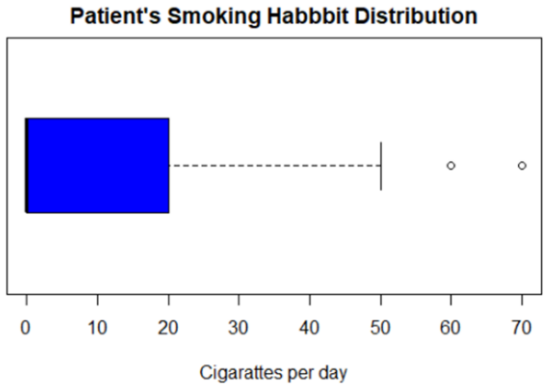
EDUCATION LEVEL (YEARS)	Count	% of Total
1	1526	41.72
2	1101	30.1
3	608	16.62
4	423	11.56

The average time spent on education by each patient given in this dataset on education is about 1.9 years.

CIGS PER DAY:

Average number of cigarettes smoked by the test subjects is around 9.03 and the median lies at 0. The range on number of cigarettes smoked per day is from 0 to 70. Information on smoking rate of the patients do not follow any patterns. Right spectrum consist of 10 outliers and they are 60, 60, 60, 60, 60, 60, 60, 60, 70 and 60.





Top 25% of the test subjects are "nonsmokers".

CURRENT SMOKER:

The original binary classification has been converted into "Smoker" and "Non-Smoker" for easy interpretation. Majority of the dataset contains non-smokers.

BP MEDICATION	Count	% of Total
SMOKER	1789	48.91
NON-SMOKER	1869	51.09

BP MEDICATION:

Patient's medical history on their BP is tabulated in the table below. It can be stated that approximately 97% of the patients have no blood pressure medication history.

BP MEDICATION	Count	% of Total
YES	111	3.03
NO	3547	96.97

PREVALENT STROKE:

Historically, minute percentage of the patients have had stroke in the past. It's a positive sign that given the dataset is filled with healthy personals. This also indicates an imbalance dataset considering the patient's stroke history.

Experienced Stroke	Count	% of Total
YES	21	0.57
NO	3637	99.43

PREVALENT HYPERTENSION:

Patient's hypertensive nature have been illustrated in the table below. Approximately 69% of the patients were found to be hypertension.

Hypertensive	Count	% of Total
YES	1140	31.16
NO	2518	68.84

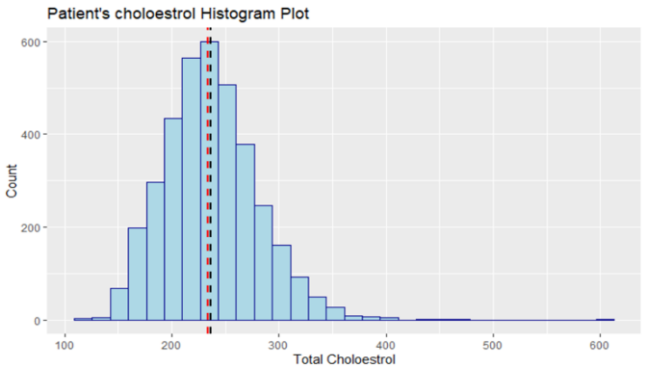
DIABETES:

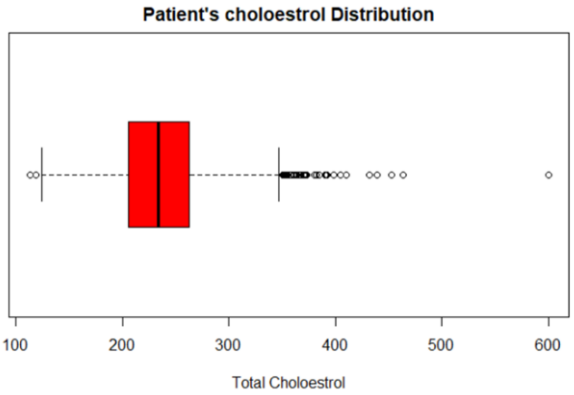
Based on the World Health organization's report people with diabetes has risen from 108 million in 1980 to 422 million in 2014. However, based on the given dataset most of the patients are very healthy. Evidently, only 2.71% patients were diabetes.

Diabetes	Count	% of Total
YES	99	2.71
NO	3559	97.29

TOTAL CHOLOESTROL:

Total cholesterol level on the test patients ranges from 113 to 600 mg/deciliter. Average total cholesterol monitored on the patient is about 236.85 mg/deciliter and the center value of the ascending observations is at 234 mg/deciliter. Top 25% of the patient's total cholesterol is below 206 mg/deciliter and top 75% of the patients total cholesterol level to be less than 263 mg/deciliter.



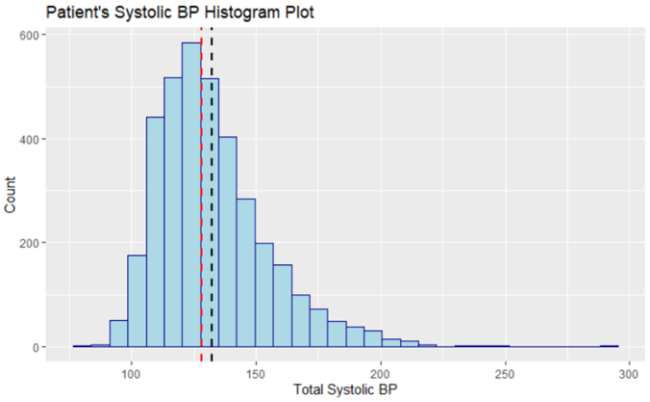


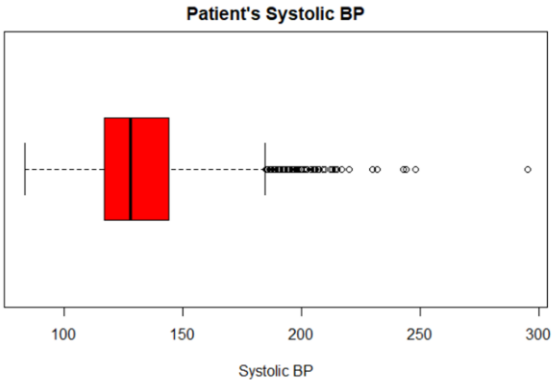
Left spectrum outliers of total cholesterol are 113 mg/deciliter and 119 mg/deciliter. Total observed 44 right spectrum outliers are displayed below.

```
464 352 368 370 439 398 355 353 360 352 600 392 358
391 410 372 366 410 390 405 359 350 380 355 390 371
350 354 382 364 367 352 432 351 363 382 361 453 352
410 350 358 373 385
```

SYSTOLIC BLOOD PRESSURE:

Observations of Systolic BP values from 83.5 to 295.0 mg of Hg. The Mean of Systolic BP 132.37 mm of Hg have a lesser median value of 128 mm of Hg. This makes the skewness of the distribution to dominate on the right. The 25% of the top patients have systolic BP around 117 mm of Hg and top 75% of the patients with the value below 143.88 mm of Hg.



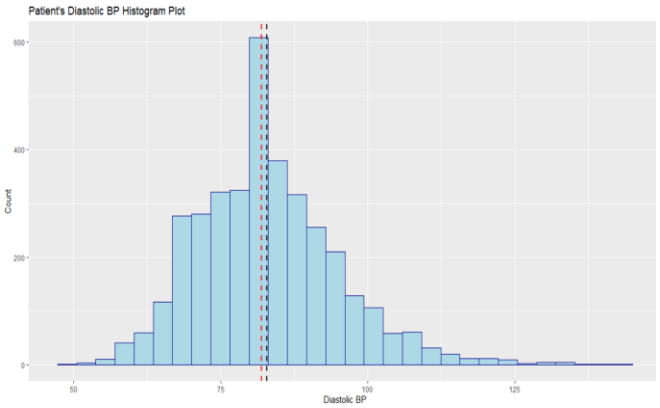


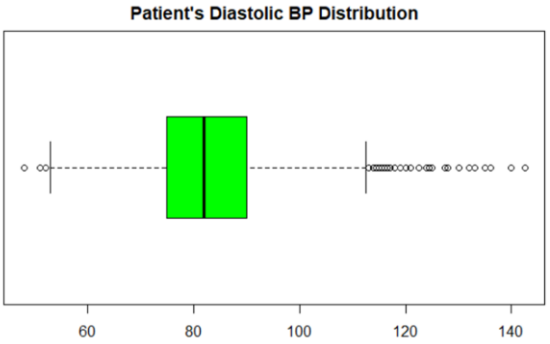
110 points were identified as outliers at the right side of the distribution.

```
206.0 190.0 200.0 187.0 212.0 191.0 200.0 189.0 197.5 189.0 204.0 215.0 197.0 209.0
295.0 189.0 185.0 220.0 205.5 186.0 192.0 185.0 200.0 244.0 213.0 206.0 199.0 198.0
206.0 201.0 189.0 243.0 187.5 199.0 186.5 186.0 204.0 217.0 196.0 193.0 187.0 196.0
189.0 196.0 190.0 202.0 195.0 200.0 232.0 191.0 184.5 188.0 205.0 185.0 220.0 210.0
193.0 188.5 192.0 199.0 197.5 190.0 195.0 210.0 184.5 202.5 191.0 190.0 210.0 197.0
198.0 190.0 204.0 207.5 191.0 195.0 198.0 197.0 186.5 193.0 196.0 199.5 193.0 195.0
248.0 196.0 202.0 185.0 230.0 197.0 189.0 214.0 196.0 215.0 192.5 187.0 194.0 207.0
185.5 213.0 192.5 192.5 200.0 187.0 190.0 206.0 210.0 195.0 188.0 190.0
```

DIASTOLIC BLOOD PRESSURE:

Diastolic blood pressure ranges from 48.0 to 142.5 mg of Hg. Mean and median diastolic BP are almost equal to value of 82.9 mg of Hg. However, the wider range has created a skewness and imbalance in the dataset. Outlier analysis states that there are outliers at both ends of the spectrum.



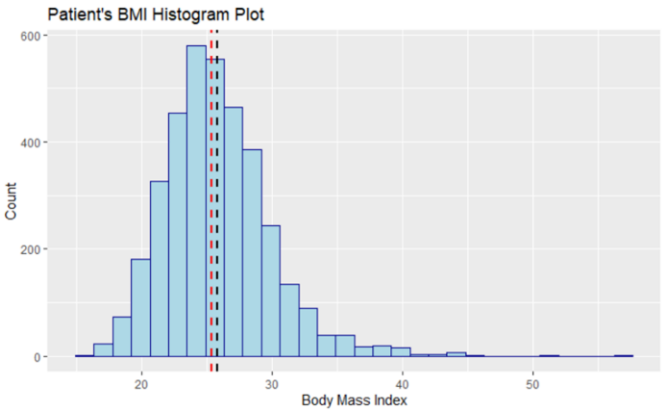


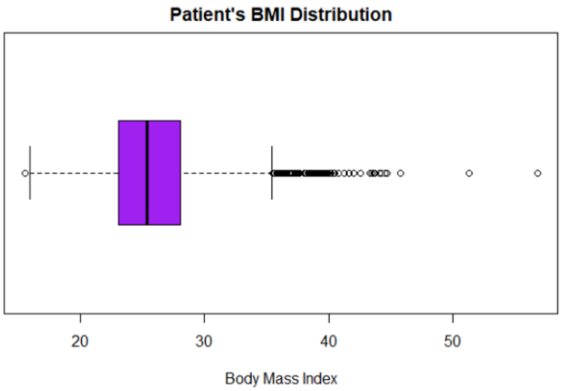
Top 25% of the patients diastolic BP is below 75 mm of Hg and top 75% of the diastolic BP Value is below 90 mm of Hg. The 65 right spectrum outliers are displayed below:

```
121.0 114.0 124.5 122.5 120.0 118.0 120.0 133.0 135.0 117.0 121.0 114.0 118.0 114.5
140.0 124.0 115.0 115.0 142.5 116.5 118.0 116.0 120.0 119.0 118.0 132.0 124.0 120.0
114.0 136.0 120.0 128.0 120.0 115.0 114.0 125.0 130.0 113.0 117.0 130.0 135.0 115.0
114.0 118.0 113.0 118.0 113.0 120.0 130.0 119.0 124.0 121.0 118.0 113.0 113.0 130.0
122.5 115.5 133.0 115.0 125.0 125.0 116.0 127.5 130.0
```

The 4 left spectrum outliers are 51, 52, 52 and 48 mm of Hg.

BMI:





Observed patients Body Mass Index values ranges from 15.54 to 56.80. The average value of the overall dataset is 25.78 while the center value of the ascending order observations is 25.38. Statistical theory states that if both mean and median are equal the distribution follows a normal spread feature. However, the histogram plot shows a nice bell curve centered around 25 with an extended tail on the right will impact the predicting power of the model on patients with high BMI value. These imbalance nature can be corrected in the upcoming chapters by applying a transformation techniques.

Based on the outlier analysis, right spectrum have 84 outliers and left spectrum have only one outlier value. The list of outlier values are listed below.

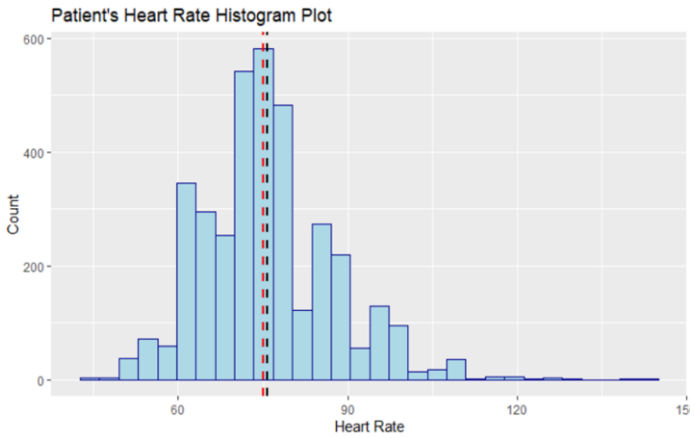
Right Spectrum Outliers:

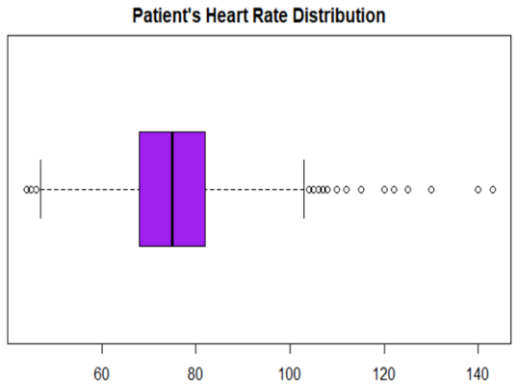
```
38.53 40.11 45.80 38.46 40.52 36.81 38.39 42.00 44.27 36.29 38.14 39.88 35.58 36.11
38.82 37.41 36.62 37.48 44.09 43.30 43.69 42.53 36.21 36.52 38.88 35.99 35.85 38.75
44.55 39.64 36.46 38.06 35.78 38.43 36.04 44.71 38.54 39.04 38.42 39.53 35.62 43.48
36.91 39.08 39.69 36.65 39.82 36.79 37.02 35.96 35.53 37.38 36.54 56.80 38.11 40.21
37.15 39.40 40.81 38.61 36.01 38.31 35.68 37.10 38.96 39.94 39.94 39.22 41.29 40.08
36.12 36.18 41.61 37.62 40.38 37.58 51.28 38.94 37.30 41.66 38.17 36.07 39.17 43.67
```

Left Spectrum got only one outlier 15.54. First quantile shows that 25% of the patients BMI records circulates through 23.08 and top 75% of the patients have values below 28.04.

HEART RATE:

The range of patient's heart rate is from 44 to 143 BPM. The histogram indicates peaks at around 75BPM and the bell curve is formed around that point indicates a normally distributed nature. This is confirmed by the almost equal to each other of mean value 75.73 BPM and median value 75 BPM of the heart rate attributes observations. However, the slight right skewness is observed through the histogram plot due to the wide range had induced a variance.





The heart rate distribution box plot indicates outlier values outside the both maximum and minimum whiskers. 76 outliers were observed at the right spectrum while only 4 outliers were detected at the left spectrum of distribution.

Following are the right spectrum outlier values.

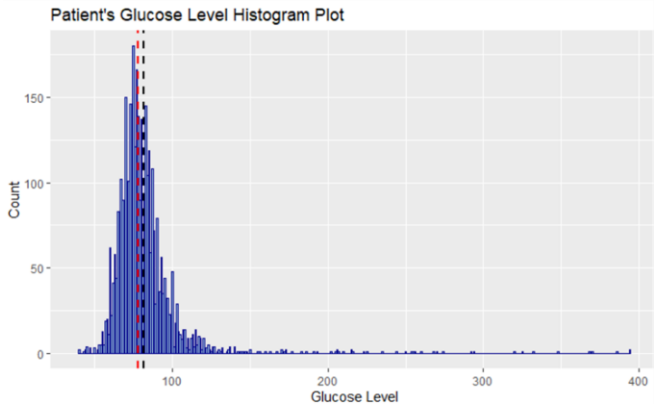
```
110 110 140 130 104 105 110 105 108 105 110 106 110 104 110 107 108 110 108 110 110 112
125 110 105 110 110 105 110 110 110 105 110 110 110 122 110 110 106 110 110 107 107 120
120 108 120 115 110 120 105 105 143 110 110 120 110 115 115 104 105 105 107 105 115 120
110 115 110 122 110 108 110 125 125 112
```

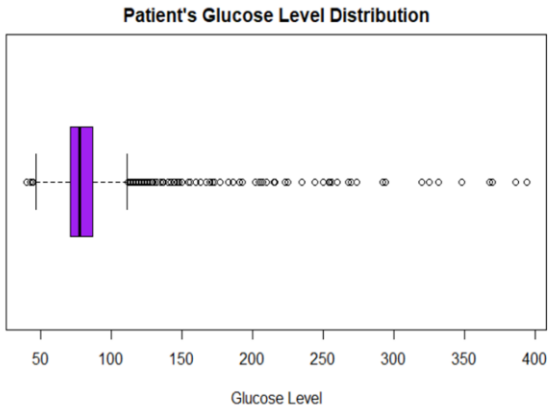
The left spectrum outliers are 44, 45, 45 and 46 BPM.

The Q1 value indicates that 25% of the people having heart around 68 BPM while top 75% of the patient's heart rate is around 82 BPM. This small range confirms the densed peaks shown in the histogram from 60 to 90 BPM.

GLUCOSE LEVEL:

Measured glucose level in the dataset ranges from 40 to 394mg/deciliter. Mean glucose level is 81.85 mg/deciliter and median is 76 mg/deciliter. Median below mean is a sign of right skew distribution. Then again the variance due to wide range of dataset have created this effect. Outlier analysis indicates that both spectrum contain outlier values.





Q3 quantile is 87 mg/deciliter and Q1 is 71 mg/deciliter. The right spectrum outliers are given below.

```
113 225 215 202 126 120 117 132 150 120 113 115 117 113 140 112 118 113 114 160 117 115
123 145 126 118 117 120 122 137 127 205 114 115 113 118 130 118 112 120 112 216 163 113
113 112 144 116 121 172 140 124 112 126 186 223 117 325 156 268 120 122 117 274 292 118
114 112 116 114 127 120 115 115 118 255 123 136 123 206 127 131 148 120 118 132 113 173
118 126 115 206 140 386 127 121 155 215 150 147 117 123 170 115 112 112 320 132 140 170
137 254 394 394 124 270 244 130 183 115 142 137 117 119 167 113 135 207 115 129 115 112
137 115 177 119 250 136 113 117 116 294 115 123 125 332 115 368 348 122 116 370 173 120
117 193 191 256 235 115 210 118 113 120 116 260
```

The left spectrum outliers are 45, 45, 40, 44, 43,44,45,45 and 40 mg/deciliter.

CHD Risk Patients

This is the dependent variable of having a risk of getting CHD in 10 years. Approximately 85% of the test subjects from the dataset are healthy and do not have any risk of getting a coronary heart disease in 10 years from the time testing was conducted, provided that the same life style and dietary patterns were followed.

CHD Risk	Count	% of Total
YES	557	15.22
NO	3101	84.78

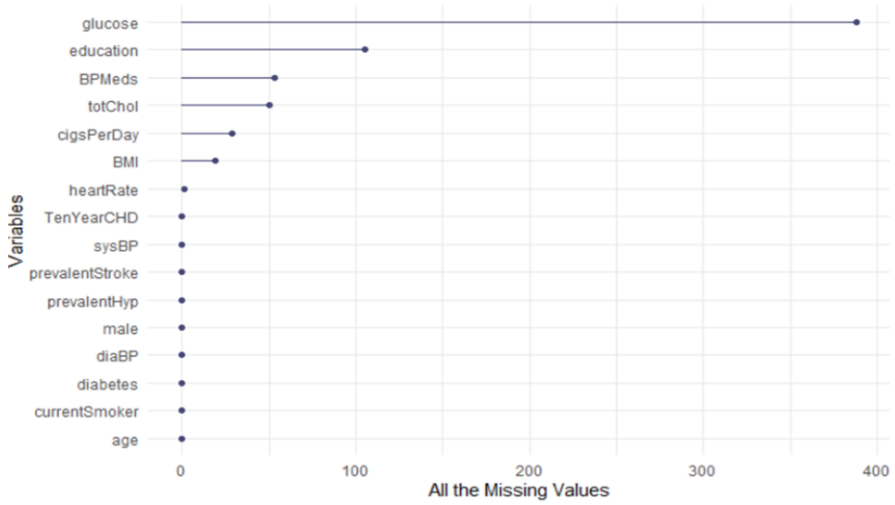
Appendix – B

Missing Value Analysis

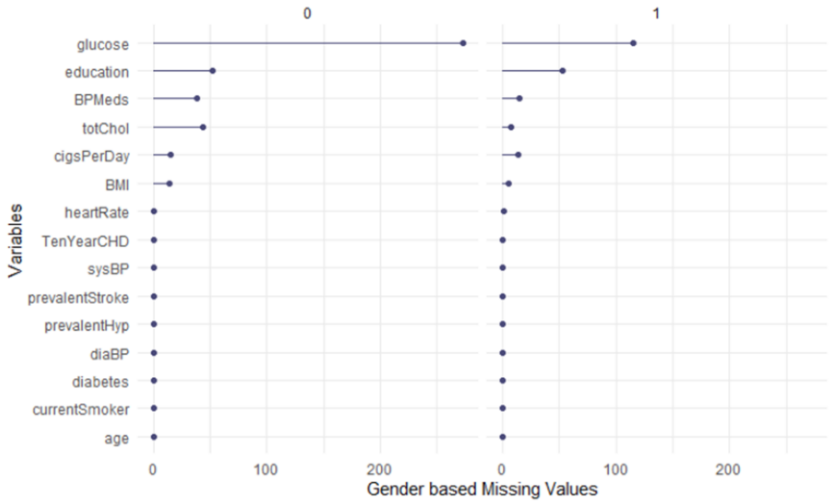
The dataset contains of 645 individual information have been found to be missing in the original dataset.

ATTRIBUTES	NUMBER OF MISSING VALUES
Glucose	388
Education	105
BPMeds	53
totChol	50
CigsPerDay	29
BMI	19
Heart Rate	1

The below diagram illustrates a graphical representation of the number of missing values found at each attributes.

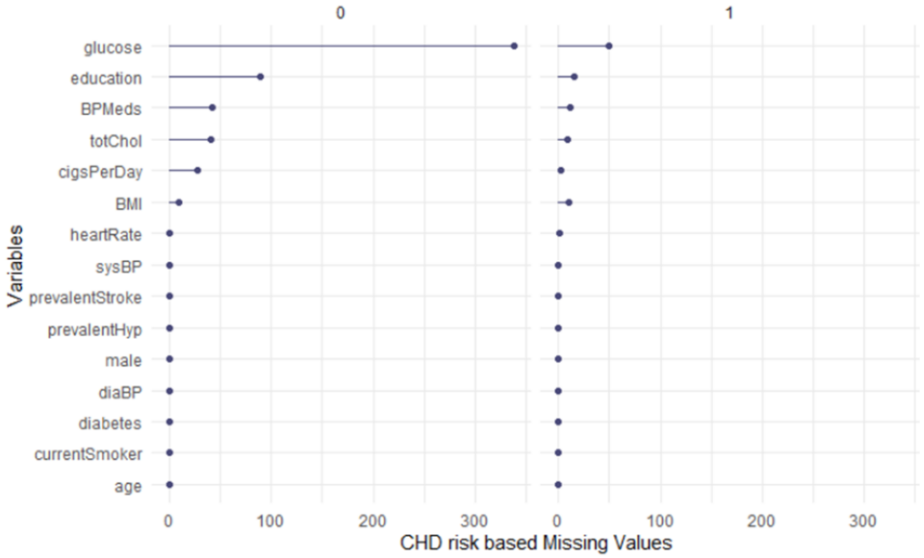


Let's analyze whether any patterns can be found in the missing values. Therefore, a gender based missing value classification of graphical representation is shown below. It can be seen that "Education" and "Number of cigarattes" smoked per day attribute missing values show an equal split among female and male patients.

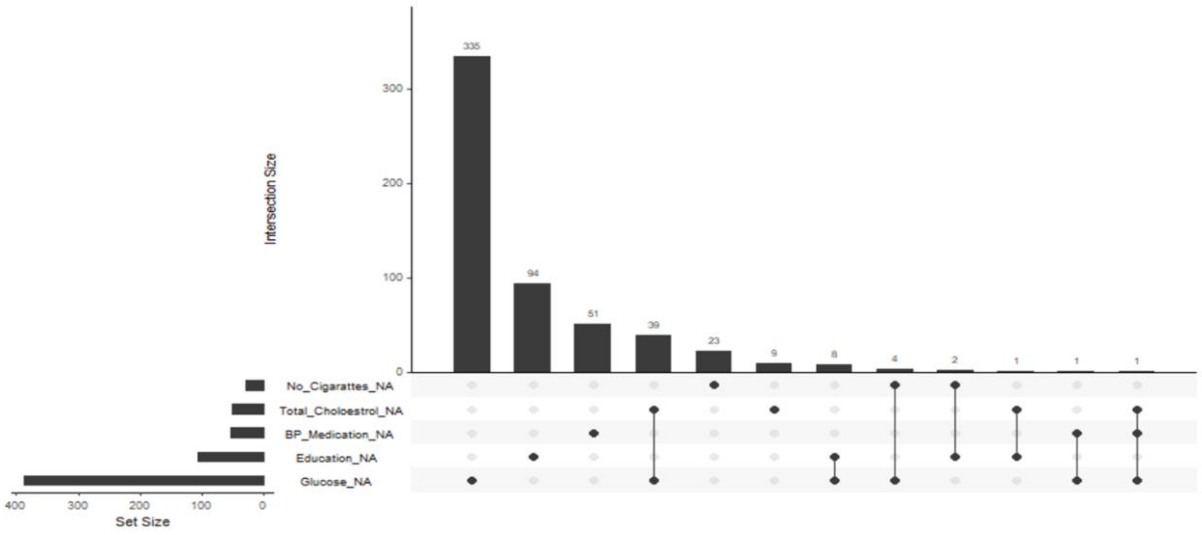


Therefore, it can be said that most of the male patient's records have been maintained properly.

Below graph shows CHD risk based classification on the missing values. It's clear that there is no evidence of any pattern on missing values between risky CHD and safe test subjects.



The bar plot below states that the missing data combination and an overall combinations of attributes observed.



Based on the bar plot above an attribute glucose contains the most number of missing values of 335 by itself. Addionally, most number of combinations were related to glucose the observed number of missing values combinations are very small and insignificant.

MISSING VALUE PATTERN IDENTIFICATION

The missing value pattern identification process have been carried out in two stages. First approach is to see the probability of replacing the numerical column observations based on their rate of missing. Second approach is to identify the possibility of replacing missing values in categorical variables.

APPROACH – 1: Missing Values in Numerical Variables.

The percentage of missing values on each attributes must be analyzed against each test subjects in the data set and the variables. In this analysis, only numerical values have been taken into consideration. The following table is the percentage of numerical independent variables with the missing values in their distribution.

VARIABLES	% OF NA
Glucose	9.15%
Total Cholesterol	1.18%
No of Cigarattes per day	0.68%
BMI	0.45%
Heart Rate	0.24%

Glucose variable have more than 9% of missing data and all other numerical variable such as "No of Cigarattes", "Total-Cholesterol", "BMI" and "Heart Rate" have missing data less than 1.2%.

The following table illustrates the percentage of missing values contributed by each test subjects. 390 observations have 12.5% of missing data. 47 people were with 25% missing values and only ONE test subject have contributed for 37.5% missing values.

NUMBER OF PATIENTS	% OF NA
3802	0
390	12.5%
47	25%
1	37.5%

Furthermore, since, this is a research on health and wellbeing of a community, an error or any sort of a false imputation will cause severe biasness in the prediction. Hence, any rows and column more than 5% of missing data will not be replaced. Surprisingly, filtered out dataset with the 5% benchmark shows no evidence of any missing values on the numerical columns. Hence, there is no need of any imputation algorithm to be applied replace NA entries.

APPROACH – 2: Missing Values in Categorical Variables.

The table below illustrates the percentage of missing values of categorical variables. Both patient's educational and BP medical history have contributed very less of an inaccurate data. In addition to that, both variables does not carry any significant correlation with any other variable to find a pattern to predict the missing values.

VARIABLES	% OF NA
Education	2.5%
BP_Medication	1.25%

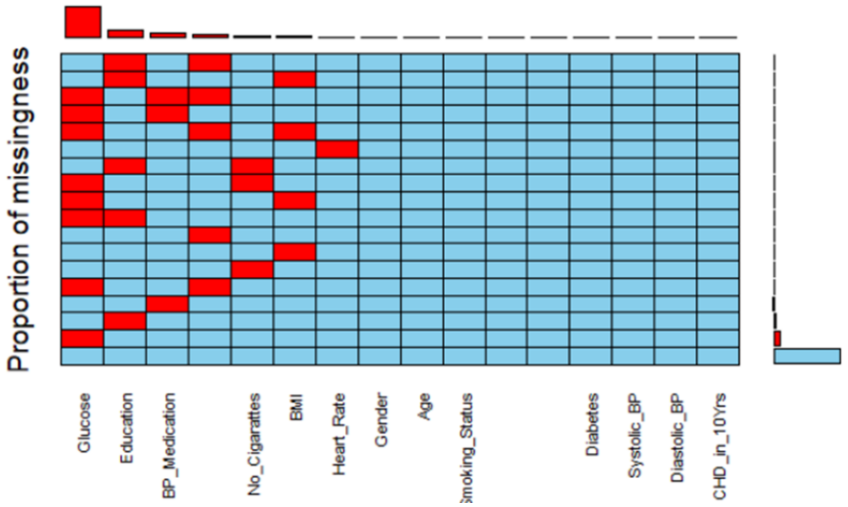
By considering the sensitivity of this research, it was decided that the replacement of missing data using imputation algorithm have been avoided at all possible occasion. Therefore, missing data of the categorical attributes have been removed without dilating the original dataset.

MISSING VALUE VISUAL ANALYSIS.

The glucose and education attributes are contributing the majority of the missing values. Hence, the investigation of NA entries will be centered on these two attributes.

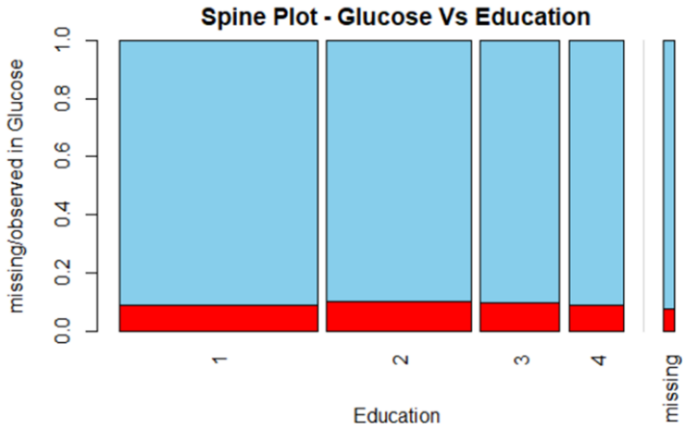
Aggregation Plot:

Moreover, it's worth investigating the presence of missing data with a support of visualization tools. The aggregation plot below shows the overall spread of the data with the missing values. The blue cells indicate the observed value and red cells indicate the missing values. The y axis shows the presence/absence based on each patients. X-axis indicates the proportion of presence/absence of data based on attributes.



Spinogram Plot:

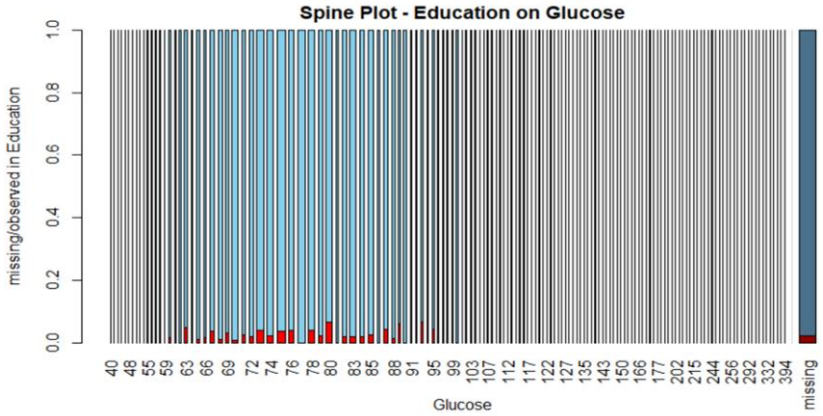
The spinogram plot is most suitable to investigate the impact of the numerical variable among other categorical variable.



Spinogram above states that all missing values in RED and available clean NA_CHD in BLUE. The width of each bar denotes the frequency of the levels while the right most bar indicates the missing proportion for the entire dataset of glucose attribute. It's very interesting that all levels of education shows almost an equal of amount patient's glucose information are missing.

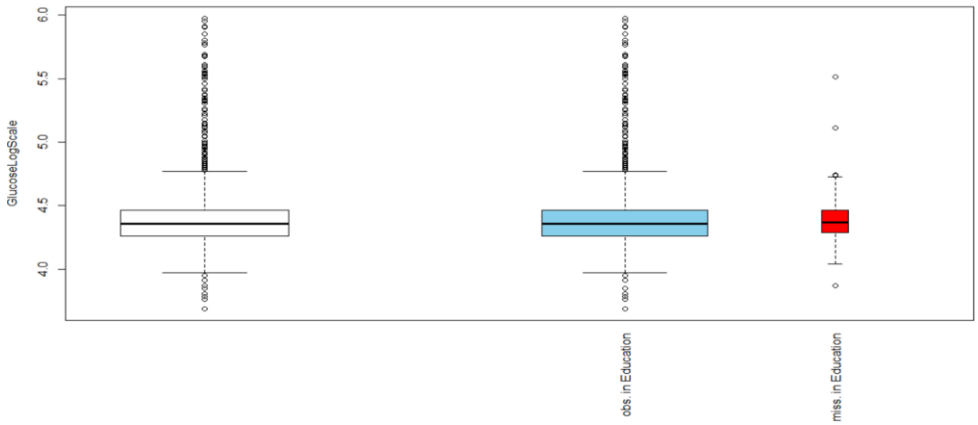
Spine Plot:

This plot is very useful to research missing value data from categorical variable on numerical data. Based on the uni-variate analysis from previous chapters on glucose it was found that observed values range from 40 to 394 mg/deciliter. Furthermore, Min IQR of glucose is at 47 mg/deciliter and Max IQR of glucose was found to be at 111 mg/deciliters. The spine plot below states that patient's educational history have been missed for glucose values between 57 to 103 mg/deciliters only.



Parallel Box Plot:

In the parallel boxplot analysis the entire dataset is divided into 2 box plot. The left most boxplot illustrates the glucose distribution as given in the original dataset. Second blue box plot is the distribution of the glucose omitting observations of missing values of education variables. The red right most box plot is the removed missing value distribution. It can be seen that both first and second box plot looks identical in size tells us that removal of missing values have not created an impact on the glucose distribution.



MISSING VALUE PATTERN TREATMENT

In conclusion it can be said that missing values does not follow any pattern and they purely are MCAR (Missing Completely At Random) in nature. Hence, further analysis can be carried out after applying complete case analysis. After removing the missing value the cleaned dataset takes the dimensions of 3658 observations with all 16 attributes

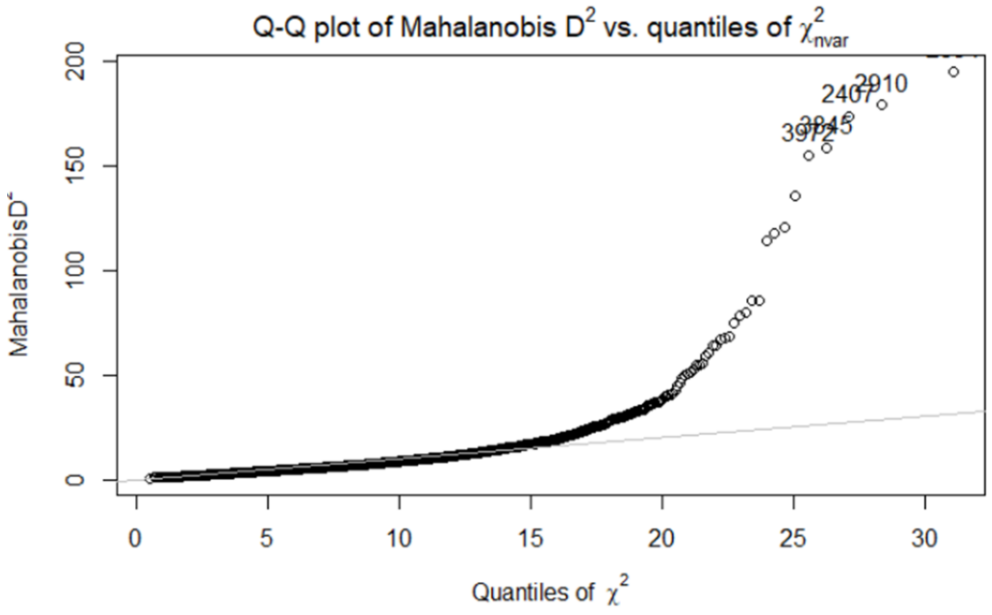
Appendix – C

Outlier Analysis

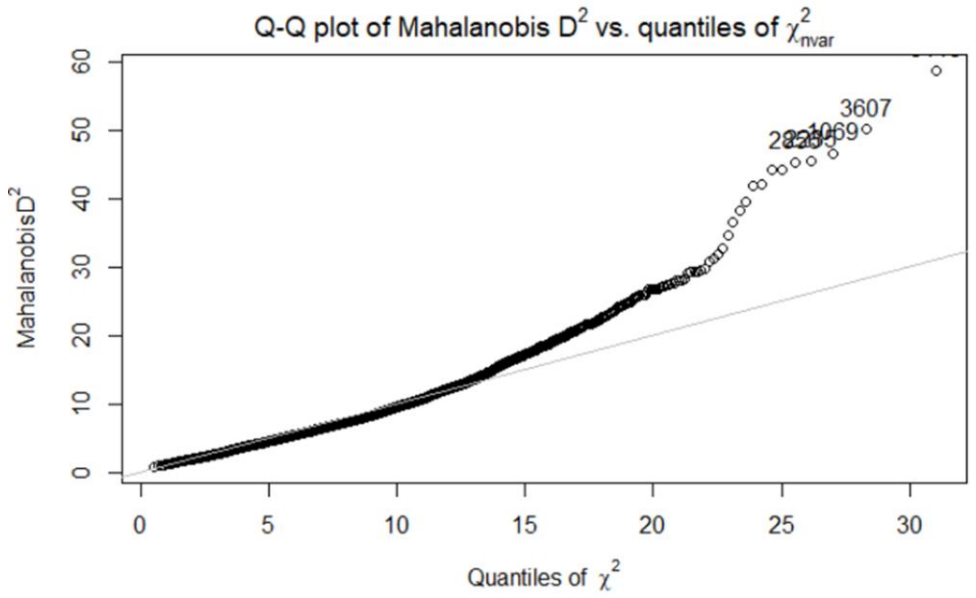
The dataset contains 15 independent variable at which following 7 variables contains outliers. Outliers heavily influence the mean of the distribution, hence it's worth analyzing and treating them to avoid any bias conclusion.

ATTRIBUTES	NUMBER OF OUTLIERS		RANGE
	LOW	HIGH	
CIGS PER DAY	0	10	0-70
TOTAL CHOLOESTROL	2	44	113 – 600 mg/dl
SYSTOLIC BLOOD PRESSURE	0	110	83.5 – 295 mg of Hg
DIASTOLIC BLOOD PRESSURE	4	65	48.0-142.5 mg of Hg
BMI	1	84	15.54-56.80
HEART RATE	4	76	44 – 143 BPM
GLUCOSE LEVEL	9	166	40-394 mg/dl

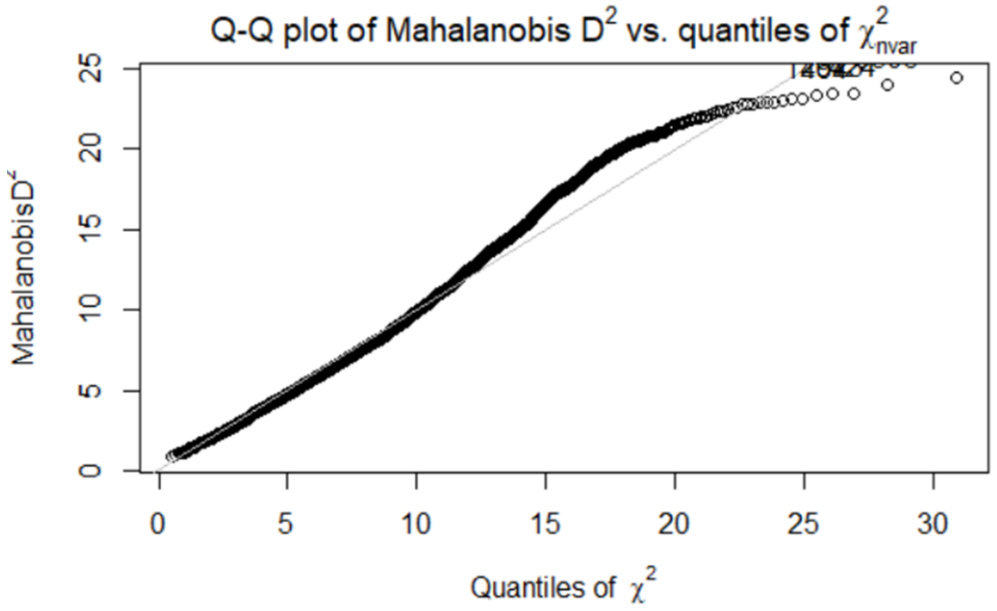
The above table clearly states that there are numerous outliers can be observed in the above attributes. Hence, Mahalanobis Distance were calculated between the mean of the distribution and the extreme values. Before we proceed with the mahalanobis analysis, it's worth plotting the Q-Q plot to understand the current spread of the distribution.



The above QQ plot clearly states the extreme values are highly deviated from the best of fit line. The below graph indicates the QQ plot after applying mahaolonobis outlier detection technique on the dataset.



The above QQ plot shows that mahalanobis technique application have improved the spread created by the extreme outliers. In mahalanobis distance calculation mean is calculated with the outlier itself. Hence, that outlier is influencing the calculation of each point from the mean. In order to overcome this indirect biasness Minimum Covariance Determinant (MCD) and mean can be calculated among the most centered 75% of the data(i.e: without extreme values). New mahalanobis distance can be calculated with the new mean and new covariance. The below QQ plot is the fined tuned model.



The above graph is a clear evidence of the new dataset with the treated extreme values.