Predict the likelihood of a patient being diagnosed with Coronary Heart Disease.

Capstone Project Report

By: ***Chinmay R Govilkar***

Dated: ***15th March 2020***

❒ Capstone project lifecycle:

The capstone project report is submitted in parts; each part will include a pre-defined set of activities. So, I'll briefly define the outline and the scope of our analysis. The report is submitted in three phases (submission notes); the breakdown of each phase is as follows:

Submission Notes I	Submission Notes II	Submission Notes III
Define the problem	Detailed data exploration	Preparing the data for modelling
Need of the problem	Missing data identification & treatment	Building multiple models
Inspecting the data	Outlier identification & treatment	Refining the model
Presenting hypothesis	Variable transformation	Selection of the best suited model
Initial data exploration	Data pre-processing	Evaluating and interpreting the model
	Machine learning approach	Business insights & recommendations
12th January 2020	**2nd February 2020**	**23rd February 2020**

❒ Table of contents:

Project Introduction:

1. Problem background

Coronary heart disease (**CHD**) also known as **Coronary artery disease** (**CAD**) involves the reduction of blood flow to the heart due to build-up of plaque in the arteries. It is the most common of the heart diseases.

Risk factors include high blood pressure, smoking, diabetes, lack of exercise, obesity, high blood cholesterol, poor diet, depression, excessive alcohol and others.

The term "coronary heart disease" is often branched under the term "cardiovascular disease". Cardiovascular disease generally refers to conditions that involve narrowed or blocked blood vessels that can lead to a heart attack, chest pain (angina) or stroke. Other heart conditions, such as those that affect your heart's muscle, valves or rhythm, also are considered forms of heart disease. Now that we have established a sound background, let us define the problem and the need for it to be solved in the following section.

2. Define the problem

We have to predict the probability of a patient being diagnosed with coronary heart disease in the next 10 years. We have been provided with a set of variables or factors, on the basis of which we will predict whether or not the characteristics of the patient are benign or malignant enough to be diagnosed with a coronary heart disease. The dataset is broadly classified into the following categories.

- Demographics of the patients
- Behavioral data of the patients
- Past medical measurements of the patients
- Current medical measurements of the patients
- Dependent or the outcome variable

Demographics of the patient will assist us in identifying how contributing factors act differently as the age, sex and locality of a patient changes. It is believed that men and women will exhibit different characteristics and thus the parameters will differ from one another.

The behavioral data will help us understand and segment the patients better. Patients who do and do not smoke will exhibit varying risk to heart diseases. The segmentation of patients is important as it will help us select an optimal model approach.

The current medical measurements will help us lay a foundation to various factors, which can contribute to our dependent variable, while past will help us to correctly predict the outcome.

Talking about the outcome variable, the most important question is deciding on a decision threshold value. The algorithms which have probabilistic values, define the probability of predicting the observed class correctly. So, setting a decision threshold value basically is highly domain specific.

Here the consequences of setting a low threshold value means predicting that a patient will be diagnosed with CHD even when he/she may not exhibit characteristics of it, which is a blunder. Therefore, this decision threshold must be chosen wisely and will be clarified during further discussion.

3. Need for the problem

Heart disease is one of the biggest causes of mortality among the population of the world. Prediction of cardiovascular disease is regarded as one of the most important subjects in the section of clinical data analysis. The amount of data in the healthcare industry is huge. Data mining turns the large collection of raw healthcare data into information that can help to make informed decisions and predictions. According to a reference, heart disease proves to be the leading cause of death for both women and men. The report states the following:

In 2015, CHD affected 110 million people and resulted in 8.9 million deaths. It makes up 15.6% of all deaths, making it the most common cause of death globally. The Registrar General of India reported that CHD led to 17% of total deaths and 26% of adult deaths in 2001-2003, which increased to 23% of total and 32% of adult deaths in 2010-2013; rates were higher among men than women of a given age.

This conclusion confirms that CHD is a major concern to be dealt with. In addition, CHD is difficult to identify because of several contributory risk factors such as high blood pressure, smoking, diabetes, lack of exercise, obesity, high blood cholesterol, poor diet, depression, excessive alcohol other factors. The high number of contributors makes it very difficult to efficiently correlate between the factors, thus has become an overhead to continue the diagnose manually. Therefore, in recent times we relied on sustainable and highly efficient techniques such as Machine Learning (ML).

4. Business opportunity

It is also important to understand the business opportunities of ML in the healthcare industry. ML already is lending a hand in healthcare by analyzing thousands of data points, which can lay a foundation for diagnosing a plethora of diseases. ML is extremely efficient and reliable as it is based on statistics and relevant past data. The ML models then uncover insights and patterns which would have been extremely difficult to correlate manually and makes accurate predictions, thus enabling us to first detect the possibility and then prevent it to all extent.

In this project, I will be applying ML approaches for classifying whether a patient will or will not be diagnosed with CHD in the upcoming decade. Various ML models will be built and tested; the most optimal performance model will be then selected for evaluation. After the evaluation, a list of recommendations and interpretations from the model will be documented. We have been provided with a dataset with some parameters. The following section will help us understand our data better.

5. Variables (Visual inspection of data)

Independent variable names	Old Data type	Need for conversion	New data type (levels)	Classified as
Male	Integer	Yes	Factor (2)	Demographic
Age	Integer	No	Integer	Demographic
Education	Integer	Yes	Factor (4)	Demographic
Current smoker	Integer	Yes	Factor (2)	Behavioural
Cigarettes per day	Integer	No	Integer	Behavioural

BP medication	Integer	Yes	Factor (2)	Medical history
Prevalent stroke	Integer	Yes	Factor (2)	Medical history
Prevalent hypertension	Integer	Yes	Factor (2)	Medical history
Diabetes	Integer	Yes	Factor (2)	Medical history
Total cholesterol	Integer	No	Integer	Medical current
Systolic BP	Numeric	No	Numeric	Medical current
Diastolic BP	Numeric	No	Numeric	Medical current
Body mass index (BMI)	Numeric	No	Numeric	Medical current
Heartrate	Integer	No	Integer	Medical current
Glucose	Integer	No	Integer	Medical current
Dependent variable name	**Old Data type**	**Need for conversion**	**New data type (levels)**	**Classified as**
Ten-year CHD	Integer	Yes	Factor (2)	Outcome variable

As observed from the table above, there are a total of 16 variables, with 15 independent or predictor variables and 1 dependent or the outcome variable. As mentioned in the earlier sections, the variables have been classified into 4 broad categories, each of which serve a specific purpose.

The dataset only includes a few variables out of many possible by which we will be predicting our outcome variable. However, we will be listing a few more in our hypothesis and assumptions section. Now, let us describe our included variables and also assess why were these variables included in our study in the below table.

Variable name	Short description	Reason for inclusion
	Demographic	
Male	Whether patient is a male or female 0 – Female patient 1 – Male patient	Gender is an important determinant. Males and post-menopausal females exhibit higher risk of CHD.
Age	Age of the patient in years	Age is the most important risk factor in developing heart diseases, with approximately a tripling of risk with each decade.
Education	Highest education of the patient 1 – High School 2 – Bachelor's 3 – Master's 4 – PhD	Education is indirectly linked in patient's risk assessment of factors. An educated person is aware of associated risks.

Behavioural		
Current smoker	Whether patient currently smokes 0 – Non-smoker 1 – Smoker	Smoking is one of the top causes for plaque formation in arteries.
Cigarettes/day	If yes, how many in one single day	Important determinant for time-frame assessment.
Medical history		
BP medication	Whether patient is on BP medication 0 – Not on BP medication 1 – Active on BP medication	Blood pressure is directly related to your arteries. If on medication, approach for treatment needs further alteration.
Prevalent stroke	Whether patient had a heart stroke 0 – No sign of previous stroke 1 – Had a previous stroke	Stroke is a mild attack, in which oxygen to heart is blocked. It means that heart is already treated once, thus would mean added caution.
Prev. hypertension	Whether patient had hypertension 0 – No sign of previous hypertension 1 – Had hypertension previously	Again, related to BP. Arteries are persistently elevated. It is altogether a combination of several factors, hence is critical.
Diabetes	Whether patient had diabetes 0 – Did not have diabetes 1 – Had diabetes previously	It is again a combination of BP, sugar levels, obesity or insulin deficiency.

Medical current		
Total cholesterol	Total cholesterol level of the patient	High cholesterol levels mean depositing lipids in blood vessels, thereby contracting and increasing risk of a heart disease.
Systolic BP	Systolic blood pressure of the patient	The pressure when your heart beats or pumps in blood to your body. It's a generic requirement for any clinical diagnosis.
Diastolic BP	Diastolic blood pressure of the patient	The pressure when your heart rests between beats or relaxes. It's again a generic requirement for any clinical diagnosis.
BMI	Body mass index of the patient	It determines whether your height and weight are in proportion. Interlinked with obesity and cholesterol.
Heart rate	Measured heart rate of the patient	Also known as pulse, is a measure which shows your resting heart rate (bpm). Adult range is 60-70 bpm.

Glucose	Measured glucose level of the patient	Normal HBA1C blood sugar level is less than 5.7 and is interlinked with diabetes when the measure overshoots the mark of 6.5.
Outcome variable		
Ten-year CHD	Decade risk of coronary heart disease 0 – No risk of CHD in next decade 1 – Risk of CHD in next decade	To predict if a patient will be diagnosed with CHD on the basis of above-mentioned factors.

6. Hypothesis and assumptions

- The dataset should include many other variables, which are excluded in our analysis. Chest-pain type, serum cholesterol (LDL or HDL), resting ECG, exercise and thalassemia are few variables excluded from dataset, which needed to be included.
- Heart rate of the patient is "resting heart rate" and measured under ideal conditions.
- The ideal blood pressure range of (120-80) is suited for adults and tolerance of high BP than normal is fine for patients aged 60 and above.
- BMI ideal ranges are different for males and females and also different w.r.t localities around the world.
- Total cholesterol includes the serum cholesterol levels, as low density lipoprotein (LDL) is a risk as it narrows the arteries, whereas high density lipoprotein (HDL) actually reduces the risk of a heart stroke.
- An ex-smoker level is excluded, meaning patient either is currently smoking or never smoked previously, effect of previous smoking is not considered in our analysis.
- Cigarettes per day is correctly recorded for each patient with no tolerance of error.
- Patient correctly justifies his/her education and is vary of the potent risks.
- If patient is female and above a certain age, is considered post-menopausal.

Exploratory data analysis & Data pre-processing:

The EDA is the next step in an analytics project after the problem is defined, the need of the problem is well explained and the listing of hypothesis and assumptions is complete. It involves multiple tasks which primarily are concerned with data inspection, analysis and manipulation. The process includes analysis, identification, imputation, transformation, balancing and visualisation. The following tasks will be completed in this section:

- Univariate analysis & plots – continuous variables
- Bi-variate analysis & plots – continuous variables
- Skewness & kurtosis of the dataset
- Missing value identification & treatment
- Outlier identification & treatment
- Variable transformation
- Addition or removal of the variables
- Scaling & balancing of the dataset
- Univariate analysis & plots – categorical variables
- Bi-variate analysis & plots – categorical variables
- Preparation of dataset for modelling

a) Univariate analysis of continuous variables:

1.1. Five point summary:

I will be analysing all the continuous variables on the basis of their distribution and spread. The measures of central tendencies includes mean, median, standard deviation, range and IQR.
The following table shows the distribution and spread of all continuous variables:

Variable	*Mean*	*Median*	*Std. Dev.*	*Minimum*	*Maximum*	*Range*	*IQR*
Age	49.58	49	8.57	32	70	28	14
Cigarettes/day	9	0	11.92	0	70	70	20
Total cholesterol	236.69	234	44.59	107	696	589	57
Systolic BP	132.35	128	22.03	83.5	295	211.5	27
Diastolic BP	82.89	82	11.91	48	142.5	94.5	15
BMI	25.80	25.4	4.07	15.54	56.80	41.26	5
Heartrate	75.87	75	12.02	44	143	99	15
Glucose	81.96	78	23.95	40	394	354	16

We can infer from the above table that,

Mean and median are in almost all cases in close proximity to each other. It could mean that the outliers, if any are not having an extreme impact on the central tendencies. However one exception stands out, for no. of cigarettes per day, median is 0, meaning the non-smokers are skewing the data, i.e. more than 50% of our data exhibit data of non-smokers.

- Median age of our sample population is 49 years, it is fair to conclude that for a patient to be diagnosed with CHD, 49 years is the average age, where symptoms start showing up.
- Total cholesterol at 234, also indicates that patients are extremely close to the safe threshold, and to reduce the risk of CHD, a level below the average is recommended.
- The blood pressure levels show no signs of risk. Both levels exhibit safe thresholds.
- BMI again averages out to 25, which is irrespective of gender, not a sign of obesity.
- Patients also exhibit safe heart rates as a whole, synchronises with the median age.
- Glucose on the other hand is way below the general accepted high level above 140, which certainly aligns with BMI and BP.

The standard deviation also is not varying to a huge extent. So we may or may not have to scale the variables because visual inspection showed total cholesterol values could skew the dataset.

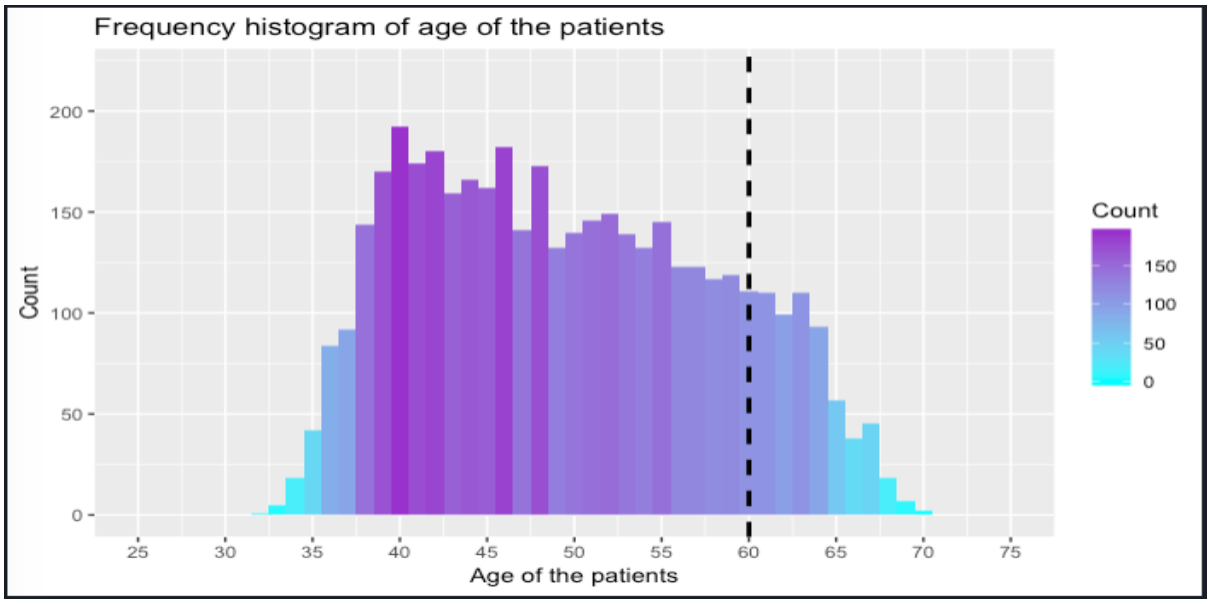
The range of the variables however shows a significant fluctuation. The IQR however again shows no spikes in the rise or decline in values.

1.2. Histograms with frequency plots

Histogram is one of the most important visualisation providing an insight on the spread, frequency or density of the variable distribution. Here I will be showing both frequency and density histograms. Each histogram will provide us with the range of values as observed in our five point summary, the count of observations falling under each range segment and finally I have marked a partition (ab-line) at a range segment, a value beyond which depicts a critical or a severe level of proportion.

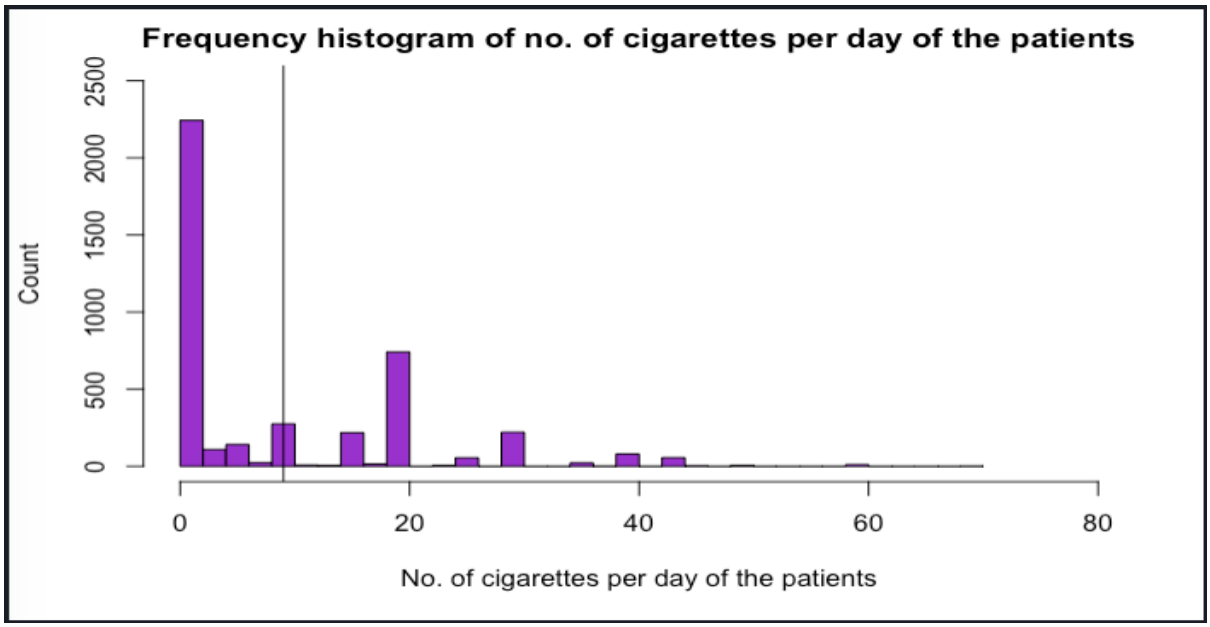
We will now plot frequency histograms for each of our continuous variables:

Age of the patients



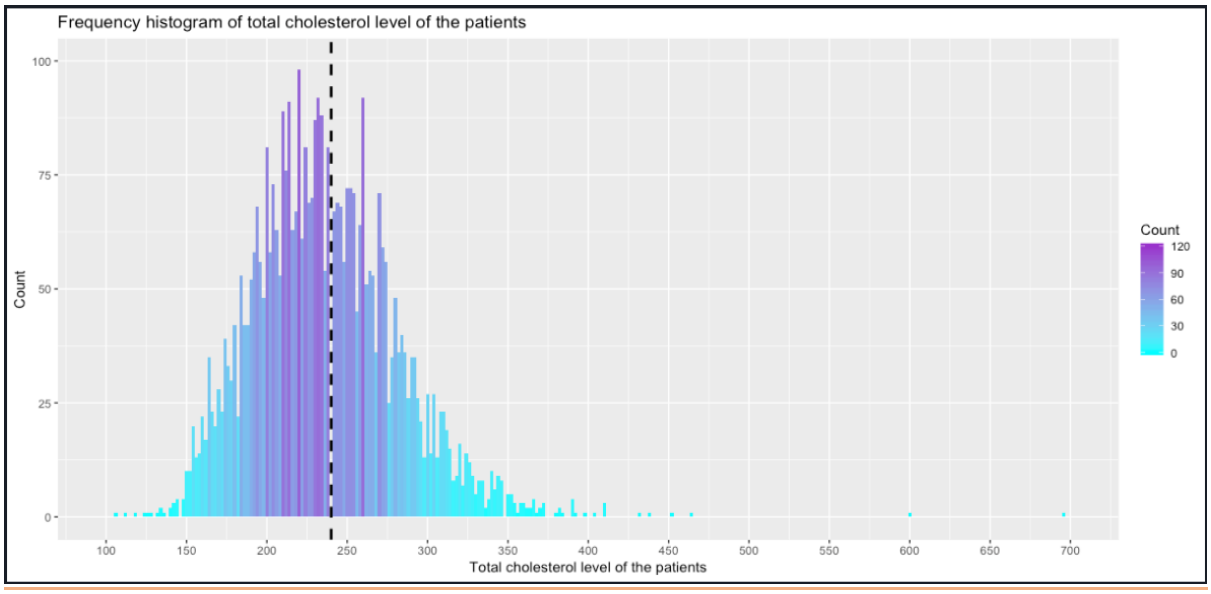
As we infer from the histogram and previously calculated five point summary, the maximum number of patients are within the age group of 39-48 years, which is evident from the dark coloured bars. The frequency of patients starts to decrease from 50 years. I have de-marked the histogram at 60 years old, stating an age beyond that signifies as senior citizen and the standard safe levels for each of the parameters do not hold true to them.

Number of cigarettes per day



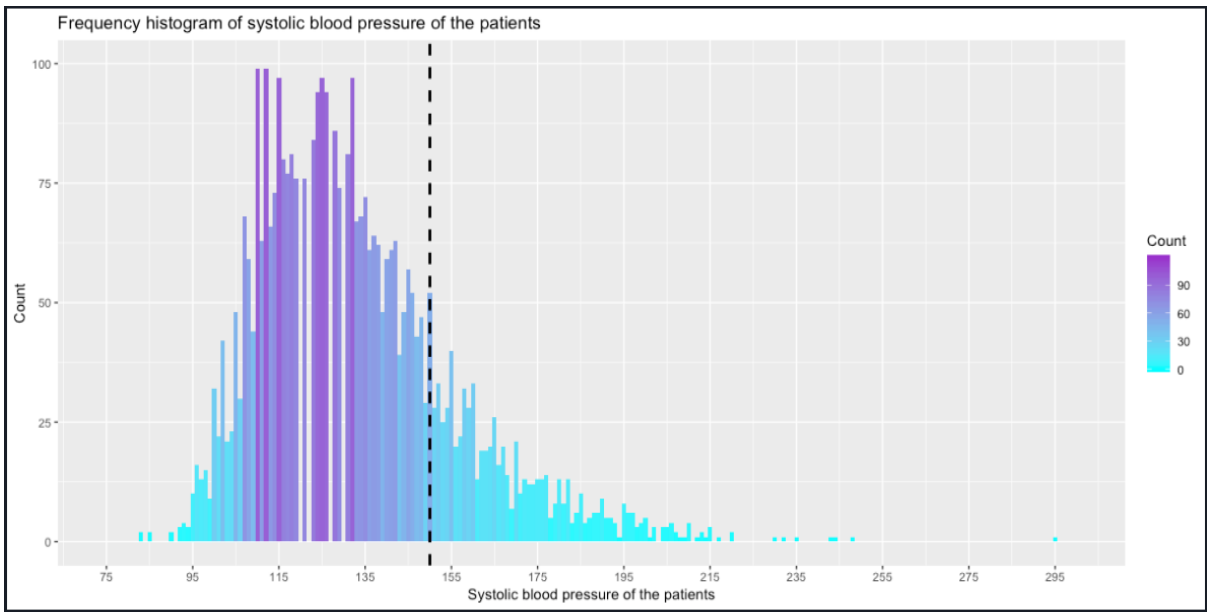
As we infer from the histogram and previously calculated five point summary, a striking difference is the count of non-smokers. As stated above, more than 50% of our patients do not smoke, which is clearly evident with non-smokers count at 2158 (~51%). The remainder of the patients have number of cigarettes per day values ranging from 1-70. The demarcation here is the mean value of number of cigarettes, which is 9. A value higher than 9 suggests that risk contributing to the diagnosis of CHD increases significantly higher.

Total cholesterol level



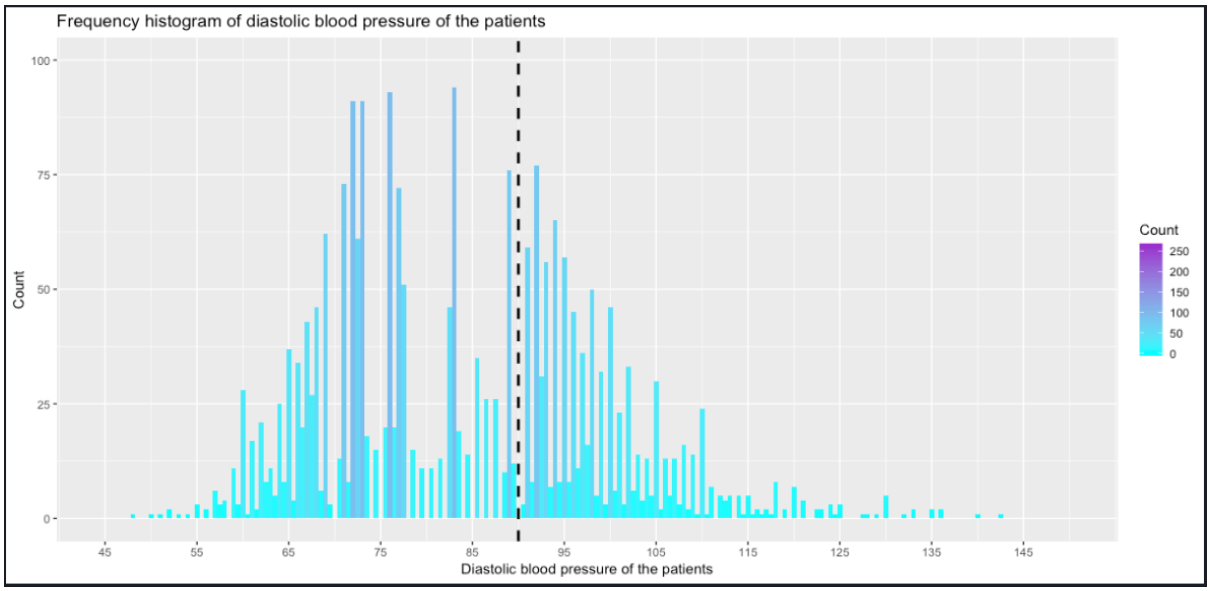
As we infer from the histogram and previously calculated five point summary, the distribution of cholesterol levels is high compared to previous histograms. Here the maximum number of patients have a cholesterol level range of 180-260 and we have de-marked the histogram at 240. Again most of the patients are just under the risk limit, but there also exist patients who are well over the threshold limit.

Systolic blood pressure



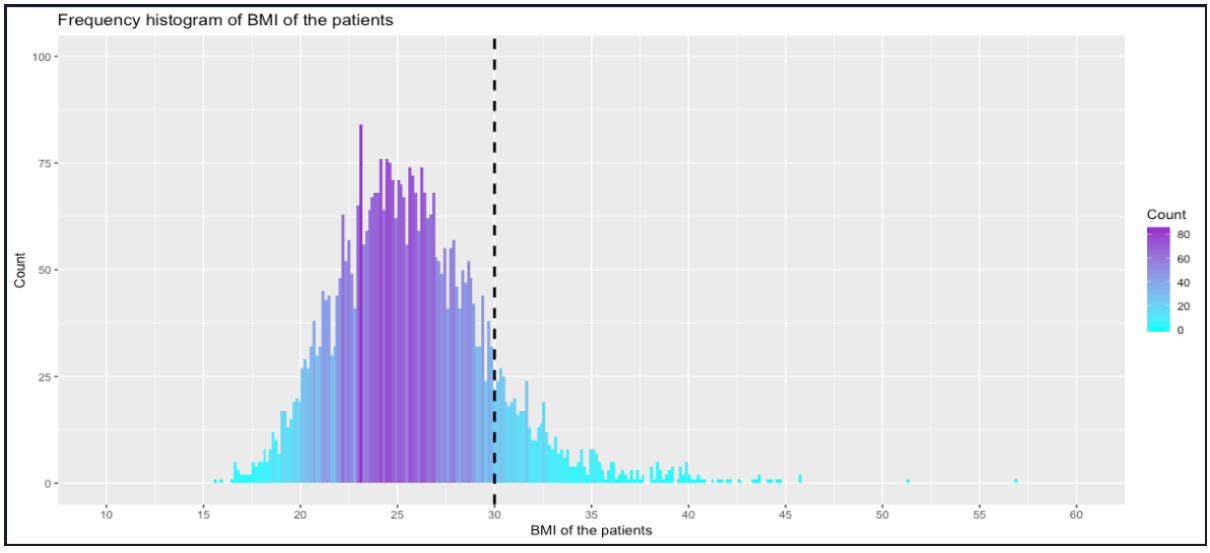
As we infer from the histogram and previously calculated five point summary, the median is at 128, wherein we observe maximum spikes. Here we have kept the threshold a little higher around 145 as the median age is 49 and blood pressure tends to be on the higher side as the age increases. Some of the recorded values are extreme (ranging above 180-200), which potentially could be recording errors and if not, then definite outliers. Some detailed inspection suggest that BP values ranging above 200, also show an average age group of 58-65 years.

Diastolic blood pressure



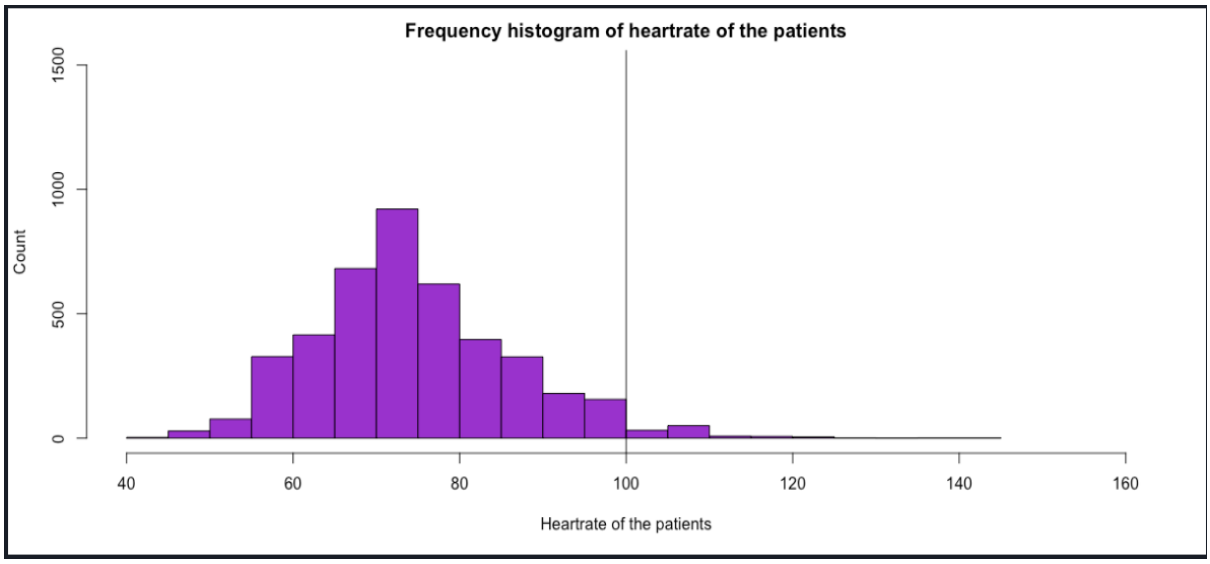
As we infer from the histogram and previously calculated five point summary, the median is at 82. This distribution is slightly spread out compared to the systolic one, and one interesting inference is that the frequency is not clustered around the mean and the median as observed in all other histograms. Also we have used a slightly tough threshold demarcation for diastolic, because compared to diastolic, systolic blood pressure has high tolerance levels, which do not combine to have a benign effect. Values above 90, are considered as a risk to be potentially be diagnosed with CHD, and here we see handful who are ranging from 90-105, which is still an acceptable value, but very few patients with values above 105.

Body mass index (BMI)



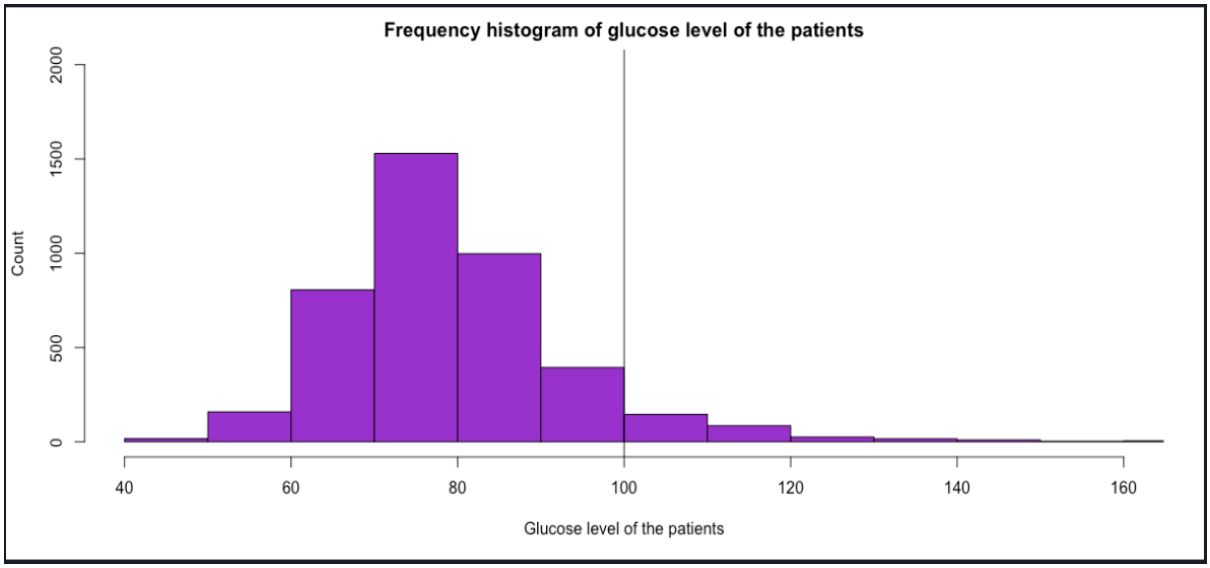
BMI is calculated as (weight in kg/height^2 in meters) and indicates how a patient is in terms of body weight and structure. BMI values vary with gender, accepted fit thresholds are 22 F, 25 M. Accepted underweight thresholds are 19 F, 22 M and accepted obese thresholds are 28 F and 31 M. Here we observe that most patients are under 30, handful from 31-35 and very few above 35. This indicates healthy patient behaviour and not a risk as to being diagnosed with CHD.

Heart rate



As we infer from the histogram and the five point summary, the median heart rate is 75. We observe distinct spikes from 60-100 with high count of patients. However a heart rate of 50 or below is either an extremely high risk or the patient is a highly trained athlete. We will consider the earlier possibility and thus have demarcated at 100. An acceptable range for an adult ranges between 60-100, therefore a value beyond that is considered as a high risk especially for CHD.

Glucose level



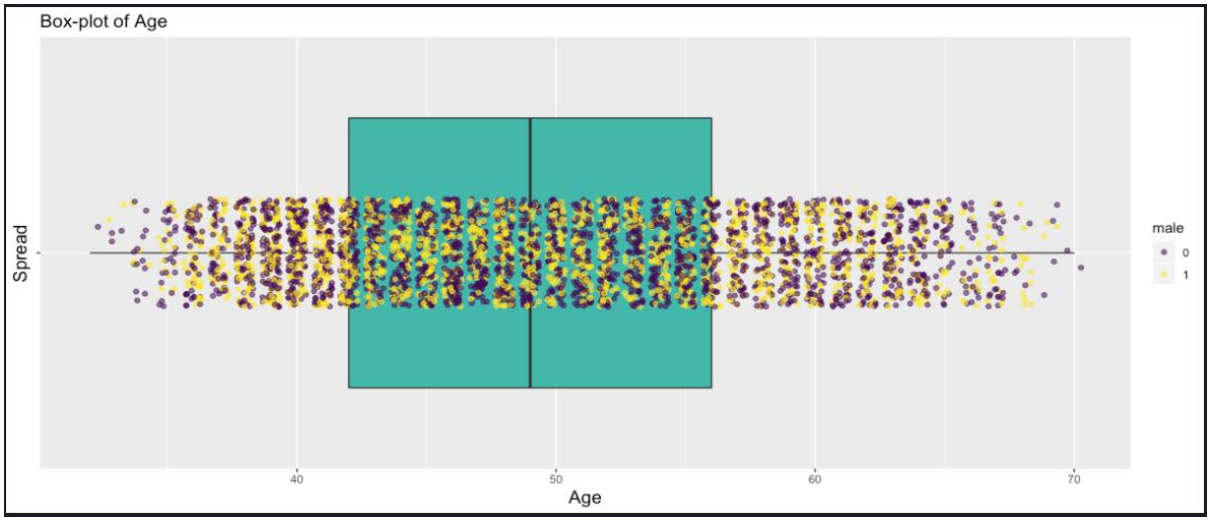
As we infer from the histogram and the five point summary, the median glucose level is 78. Similar to what we observed with heart rate, the frequency of patients is very high with glucose level ranging from 60-90 and the frequency is reduced drastically after 100. Now we have demarked the data at 100 which is recently considered as a threshold value. However initially a value of 140 was considered a risk. As our patients do not exhibit extreme range values, I decided to go with 100 as the safe limit. Now glucose is somewhat related to sugar levels and therefore to BMI and diabetes, If we dig deeper, our data did not show signs of obesity among our patients barring a few exceptions, the glucose levels certainly confirm our hypothesis regarding the same.

1.3. Box-plots with jitter plots

Box plots are statistical plots which define the five point summary i.e. minimum, maximum value, the lower (25%) and upper (75%) of IQR and also the median (50%). Box-plots are one of the ways to detect outliers statistically. The “Tukey’s method”, which statistically calculates the presence of outlier values by using the formula : less than Q1 – 1.5 * IQR and Q3 + 1.5 * IQR. Thus box-plot enables the outlier identification by plotting extreme values by a different colour.

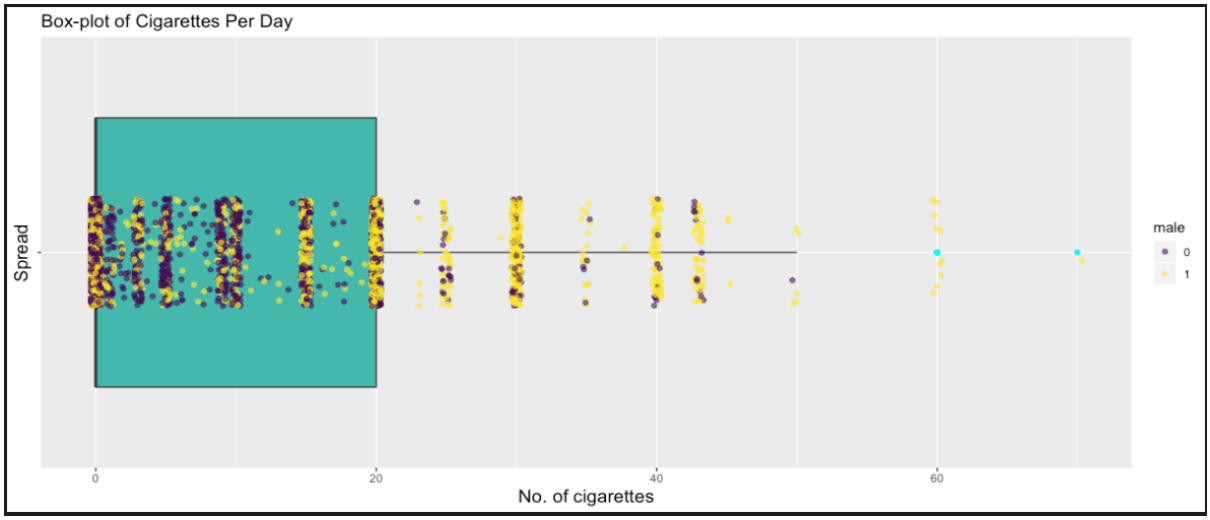
I have superimposed jitter points on the box-plot, as box-plots only depicts the spread and it can at times lead to incomplete inferences. Thus jitter will add noise to our box-plot and spread the points randomly, thereby enabling us to infer to our conclusions in a more detailed approach.

Age



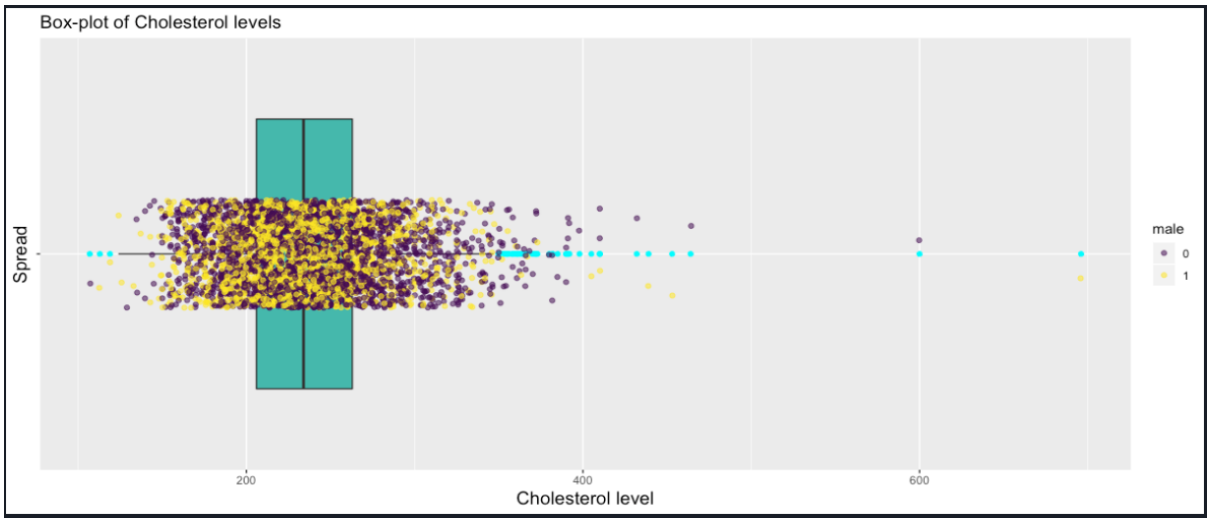
The box defines the five point summary with dark-centred line as median. The jitter points depict the spread of our data points by revealing details which are not obvious. The legend shows the jitter points are grouped by gender and it easily summarises the age-range of our patients, which is now evident by the count of female and male patients.

Cigarettes per day



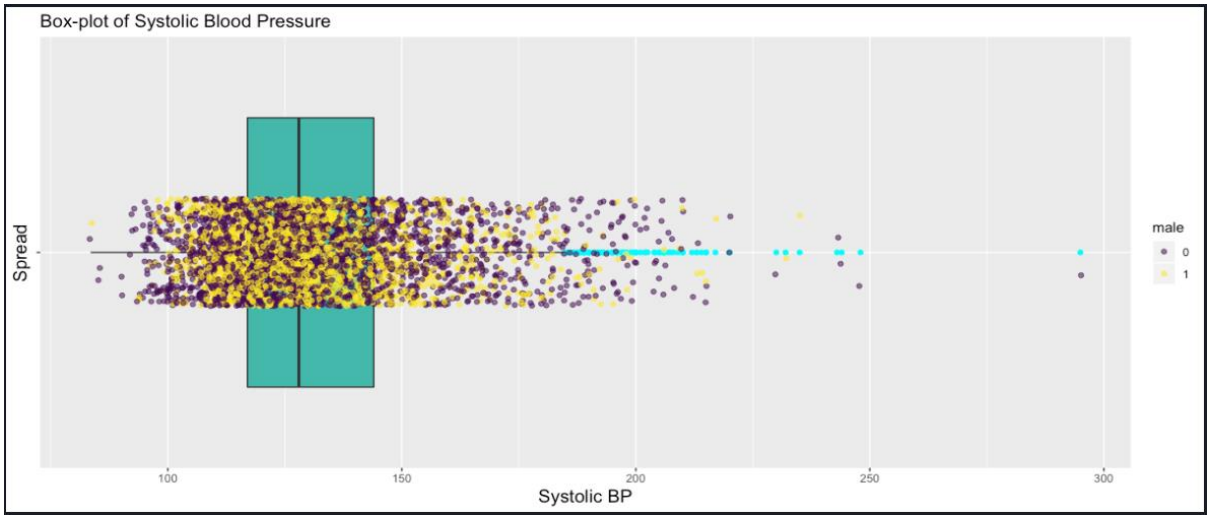
As we infer from the above plot, we have previously established that median is affected as more than 50% of our patients are non-smokers, hence we have the median marker at zero. There appears to be a clear demarcation of colour, meaning males are more likely to smoke compared to females, which however does not come as a surprise. The other observation is that we clearly see clusters at multiples of 5, meaning patients who smoke, it is easier to keep a track of number of smoked cigarettes. Important are points to extreme right, marked in cyan, are our outlier values.

Total cholesterol



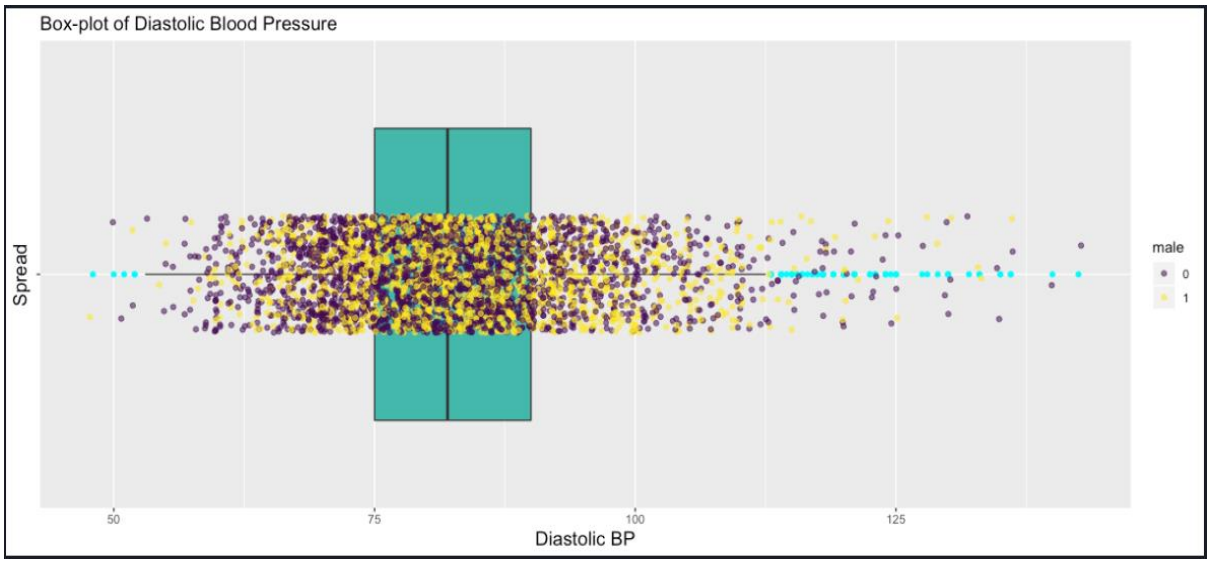
Here we observe that the range of our observations has drastically shifted. The IQR box is very narrow, stating that very few observations actually lie between the 75th and 25th percentiles. The number of outlier values have also risen. There is no significant inference that is evident w.r.t total cholesterol level of patients grouped by gender.

Systolic blood pressure



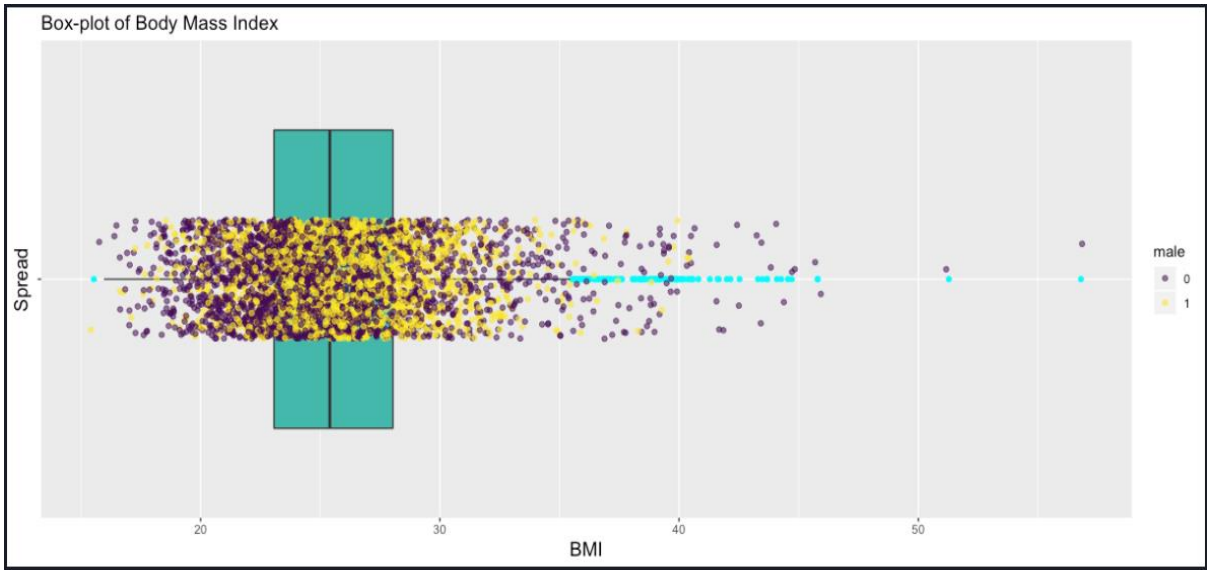
As we infer from the above plot, the median lies around 125, which is perfectly normal range for systolic blood pressure. A congested cluster of male patients hovers around 120-130 while the female patients are widespread. The outliers also are plenty with values larger than 180. Blood pressure is a subjective parameter and systolic has a high tolerance level w.r.t values beyond and below the threshold of 120.

Diastolic blood pressure



As we infer from the above plot, the median value is just around 80, again a perfectly normal metric for diastolic BP. Unlike systolic, diastolic blood pressure does not have a high tolerance for outlier values. Values ranging significantly above or below the threshold possess an extremely high risk. Thus we see a higher spread below 80 and majority between 81-100.

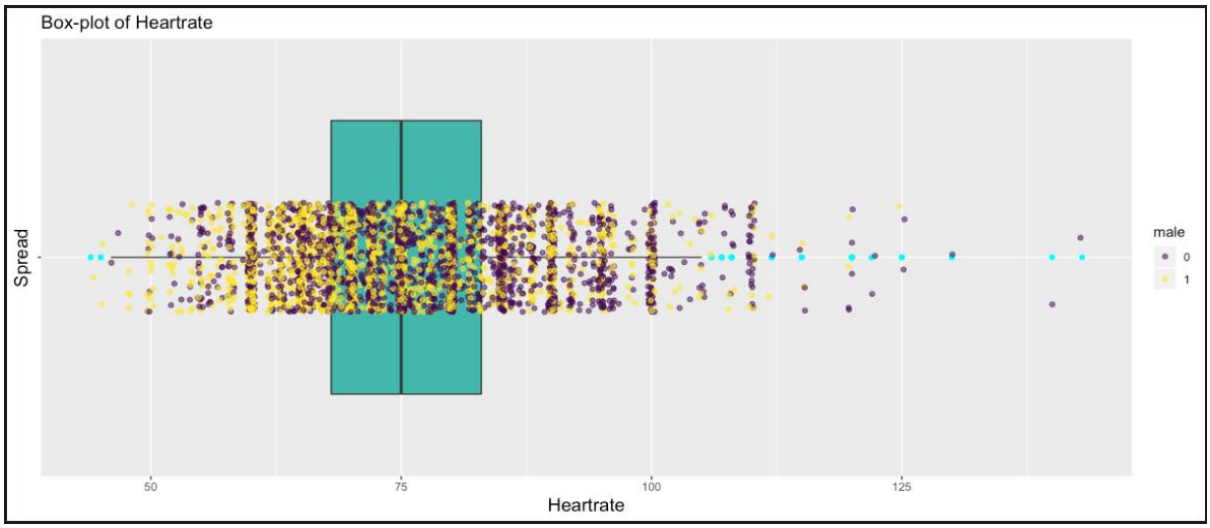
Body mass index



Body mass index abbreviated as BMI is a measurement of obesity. It considers height and weight of the patient and gives a metric which states whether a value corresponds to underweight, ideal weight or overweight.

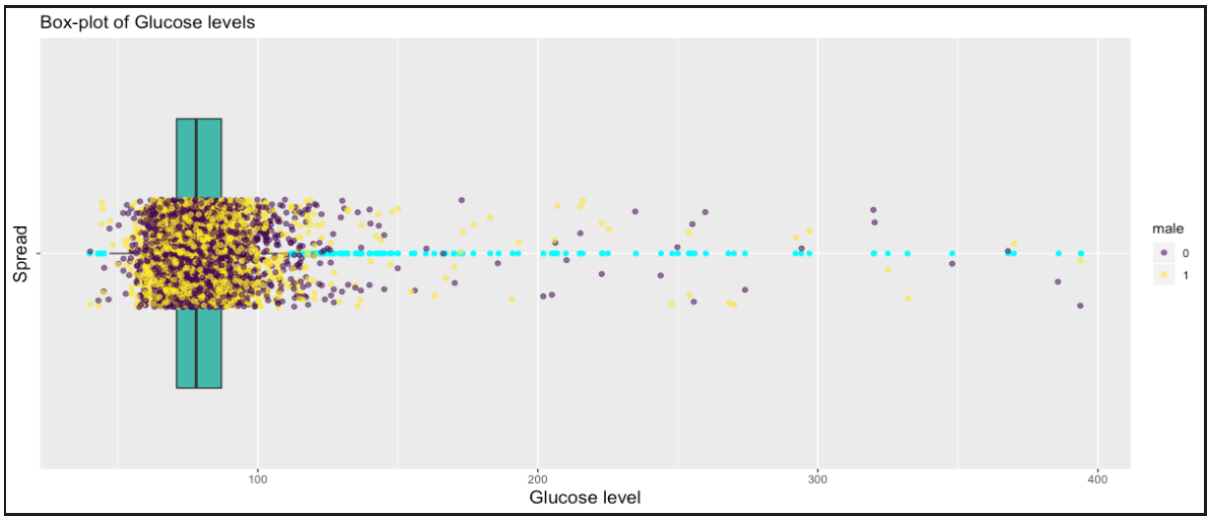
The ideal range for female is 22 and for male is 25. Hence we observe a cluster of yellow points corresponding to males from 24-27 and purple points from 19-23.
The overweight range for female is 24 and above and for male is 28 and above. Interesting is we observe a mixture of purple and yellow points well above our overweight ranges, meaning that these observations are in turn obese and possess a high risk of heart diseases.

Heartrate



Heartrate commonly known as a pulse is a metric which calculates the rate at which our heart beats and is often in terms of a unit minute time frame, i.e. if a heartrate is 65, it infers that it is 65 bpm where bpm refers to beats per minute. It may appear that higher the bpm, more healthy is the body, however it is not true.
Either a lower or a higher value is a risk, the ideal range for an adult is 60-75 bpm measured in resting position. A slightly elevated pulse is acceptable and normal if high intensity activity was performed prior to measurement. Only professional athletes have heartrates lower than 50. Here we observe that the measurements start around 50 and go as high as 125+, but the median is at 75, which is of sorts ideal.

Glucose



As we infer from the above plot, many of the observations have an extremely wide spread. The IQR is also very narrow, in addition to a congested cluster of observations below 100. The observations above 140 definitely exhibit a high risk of being diagnosed with CHD. However we have a high number of observations above 140, a high glucose level does not necessarily mean death risk, patients have survived with glucose levels close to 400, but it has a direct impact on the body parts and movement. We infer that generally, women are showing a high number of cases with extreme glucose levels.

b) Categorical variable univariate analysis:

I will be analysing all the categorical variables on the basis of their frequency and distribution. The measures include mode, count and percentage count.

The following table shows the frequency and distribution of all categorical variables:

Variable name	*Mode*	*Count (0)*	*Count(1)*	*% count(0)*	*% count(1)*
Male	0	2420	1820	57.07	42.93
Current smoker	0	2145	2095	50.58	49.42
BP medication	0	4063	124	95.82	2.92
Prevalent stroke	0	4215	25	99.41	0.59
Prevalent hypertension	0	2923	1317	68.93	31.07
Diabetes	0	4131	109	97.43	2.57
Ten year CHD	0	3596	644	84.81	15.19

We can infer from the above table that, the majority class in each of the above mentioned variables is 0. The distribution or the balance of the variable is in some cases balanced and in the remaining highly skewed. We will need to balance the variables with an appropriate balancing technique for better model performance, which we will accomplish in the submission notes II section. The education variable unlike the others has 4 levels, so lets us check its distribution below:

Variable	*Mode*	*Count(1)*	*Count(2)*	*Count(3)*	*Count(4)*
Education	***1***	1720	1253	689	473
		% count(1)	***% count(2)***	***% count(3)***	***% count(4)***
		40.56	29.55	16.25	11.15

The above table has a mode of 1, i.e. majority of our patients have only studied till high school, which may not be a positive sign, but we will conclude our hypothesis once the modelling is complete. The second majority class is of patients who have graduated. There are however patients who have further completed masters and PhD's.

I already have performed the exploration of our categorical variables on the statistical count and percentage fronts. Let us list the visualisations we will be plotting:

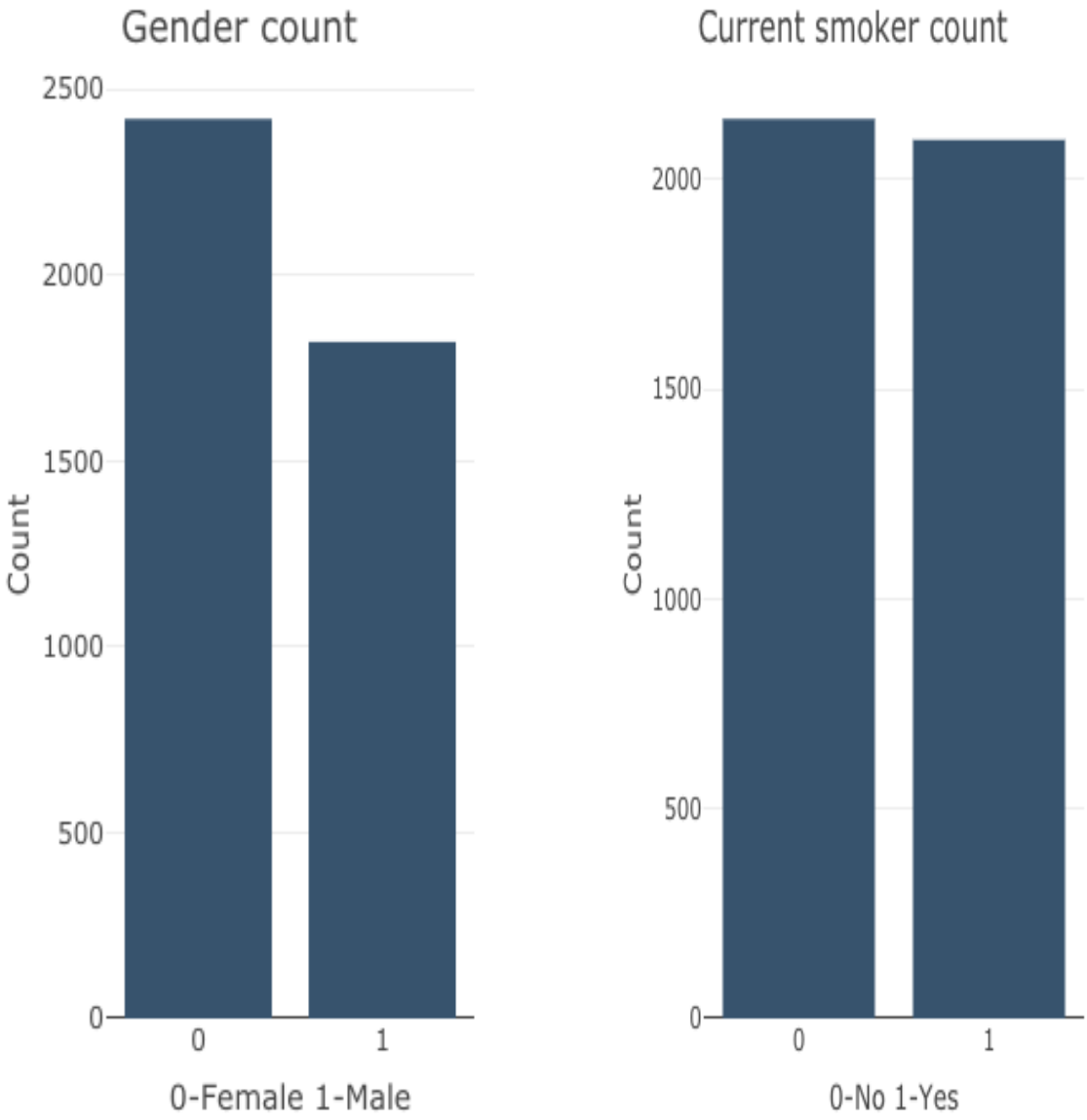
1. **Categorical**

- Category count
- Category percentage
- Bar plots
- Chi-square tests

2. **Continuous**

- Histograms
- Box-plots
- Jitter plots
- QQ-normal plots
- QQ-line plots
- Correlation chart

Male and current smoker count (Relationship between variables)

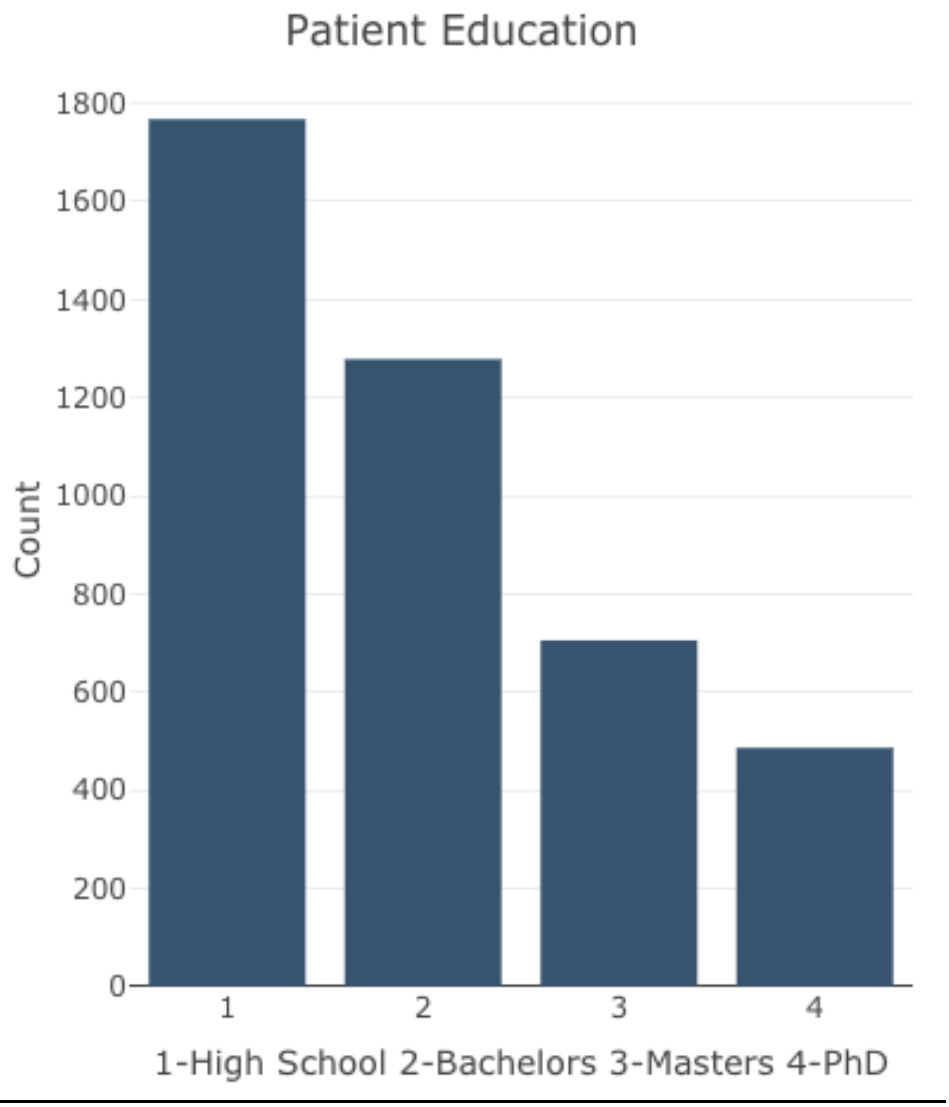


The bar plots display the count of patients grouped by gender and the count of patients who are currently smoking. As the labels display, 0 refers to a female patient while 1 refers to as a male patient. In context of the smoking bar-plot, 0 refers to a non-smoker and 1 refers to a smoker.

As we infer from the gender bar plot that our dataset has more number of female patients. This is an important find because we can relate the other parameters to whether it is associated with a male or a female patient. More so males are more likely to be diagnosed with a heart disease compared to females. Only post-menopausal females exhibit risk equal to males for being diagnosed with CHD.

The number of smokers and non-smokers is a direct indication of symptoms, effect of which is compounded at every step to contribution of diagnosis of heart disease. We have an almost perfectly balanced distribution of currently smoking and non-smoking patients. An important correlation here is smoking patients to have risk of hypertension, high blood pressure and high cholesterol and glucose levels.

Education (Relationship between variables)

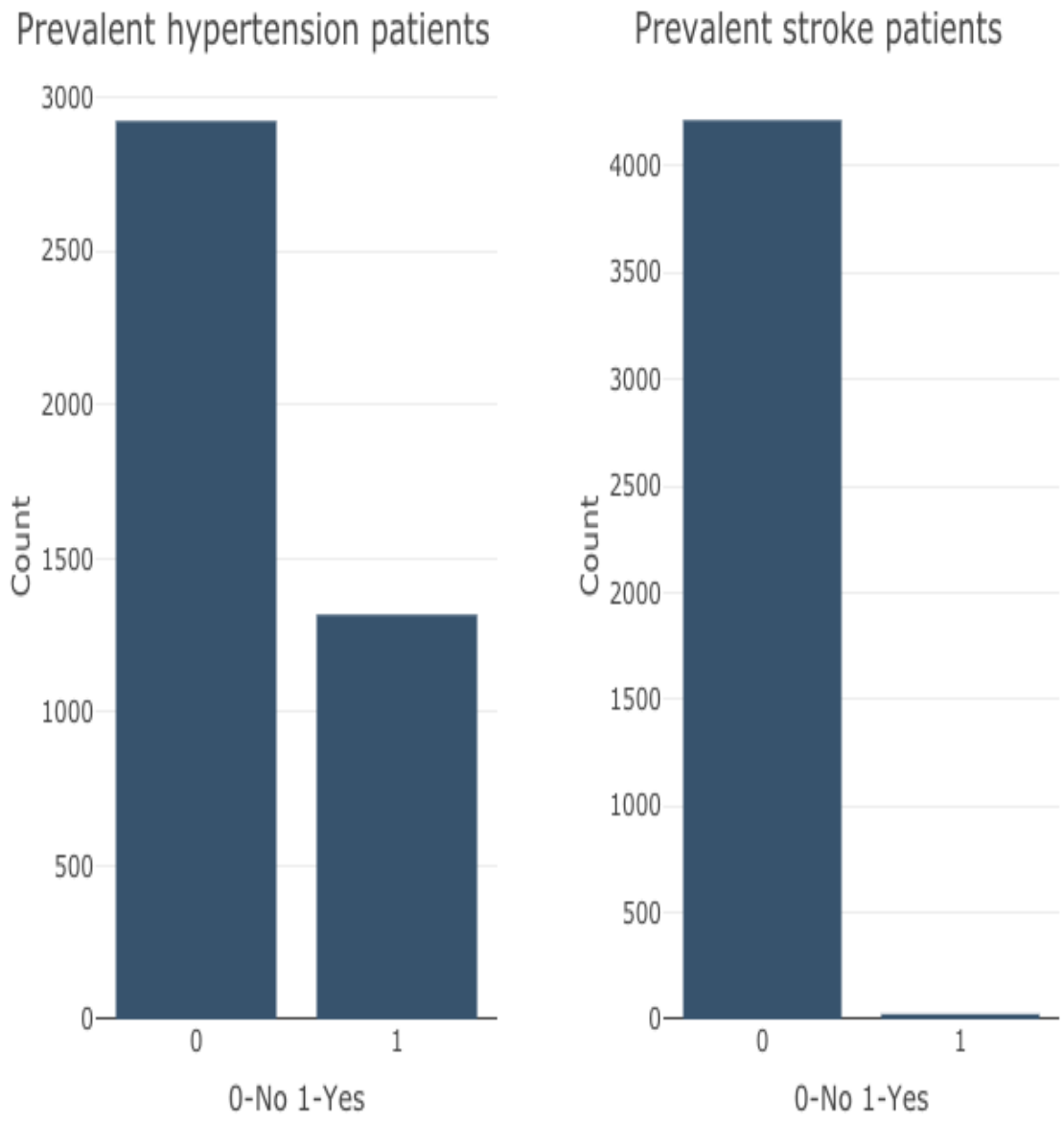


Perhaps, out of all the independent variables, education may appear to be the least important in contribution to our dependent variable of being able to be diagnosed with CHD in the next 10 years. Irrelevant or not, I will try and relate its somewhat significance to our outcome.

The above bar-plot shows the education levels of our patients. It is categorized as 4 broad sub-classes as 1-high school education completed, 2-bachelor degree completed, 3-master's degree completed and 4-PhD completed. Now each of these education levels have a corresponding civic sense, common-knowledge and ability to assess a scenario outcome characteristics. Instead of going with the small details, I will try to correlate it with our outcome variable.

It is true, that the characteristics we mentioned above need not hold true for the particular education level, but I will assume an ideal scenario where-in these characteristics are followed. A well educated person say, bachelor degree completed, is bound to know his/her social and health limits. It also has a direct impact on the adaptability in case of an unforeseen circumstance. Perhaps overqualification could also indicate its downside, meaning, a well-educated person, living an above average lifestyle might indulge in few parameters because of the status-quotient and so on.

Prevalent hypertension and stroke (Relationship between variables)

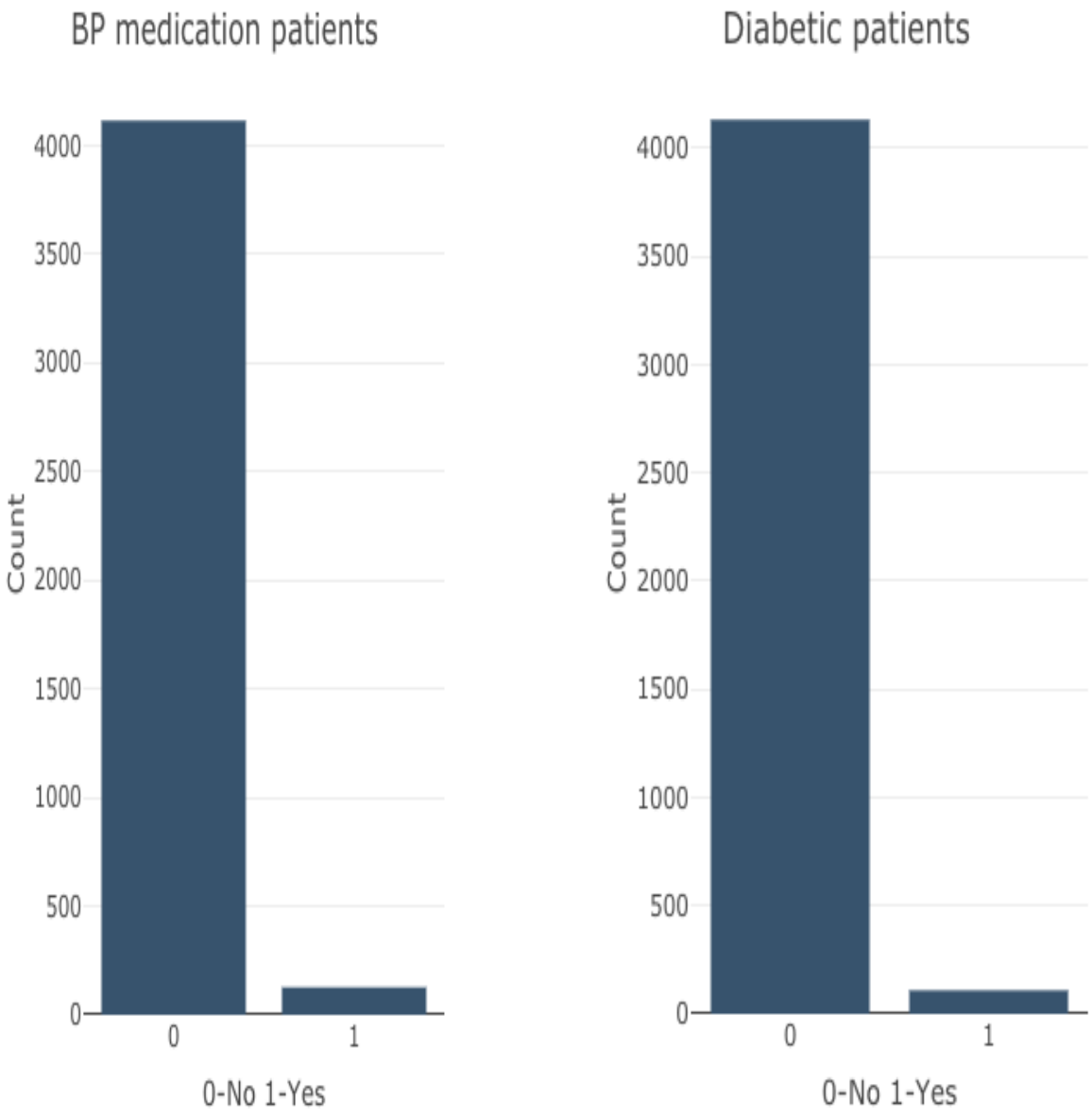


The above two bar-plots show the count of patients who previously had hypertension or who had suffered a stroke. A stroke is a milder form of heart attack, where the arteries are contracted and the blood pressure shoots up making people inefficient and incapable of doing things. It generally occurs if patients are heavy smokers or consume alcohol in heavy proportions and also who tend to be overweight and obese.

We infer from the data, that approximately 25% of the patients had been previously diagnosed with hypertension. Hypertension is another name for high blood pressure and it is common in patients who are overweight, obese and have high salt intake, are smokers etc. It is maybe the first warning that if not properly cautioned, can cause trouble to the heart. Thus it is not a highly concerning.

On the other hand, we see that there are only 25 patients who had a previous stroke. As previously mentioned, this is the latter stage in hypertension, when the BP is persistently elevated, the formation of plaque begins and starts contracting arteries. So we may establish that smoking, BMI and BP medication are 3 most significant factors leading up to heart diseases.

BP medication and diabetes (Relationship between variables)



Here we have the bar-plots of patients on BP medication and patients who are diabetic. Diabetes is a severe stage of high sugar and glucose levels in the body. It develops in 2 stages, in type 1, immune system cells attack the pancreas, where insulin is made and in type 2, the body becomes resistant to insulin. It predominantly deals with sugar levels and symptoms are mostly based on the eating habits. However relating it back to heart, increase in sugar levels means increase in body fat, also the lipids which are formed in the arteries and obstruct the flow of blood, causing elevated BP levels.

Picking off from where we left, elevated BP levels are the first sign of hypertension. Thus it is important to know whether a patient was prescribed to take a BP medication. The treatment of such patient will differ and adjusted to the new findings.

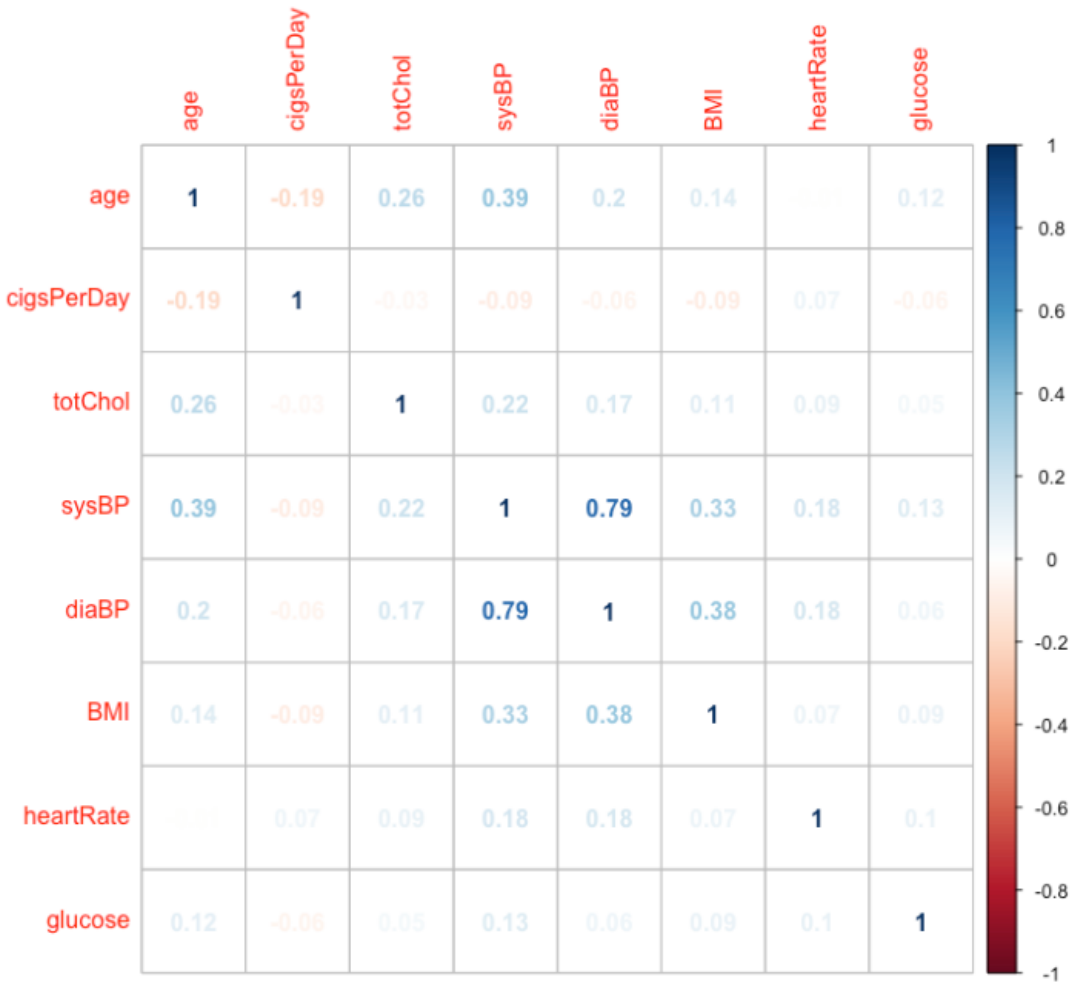
In both the bar-plots we have approximately 100 patients who are on BP medication and who have been diagnosed with diabetes. It means that there exists an imbalance in the dataset, which needs to be fixed, before we model the data, which we will look in greater detail in the subsequent section.

c) Continuous variable bivariate analysis:

In bivariate analysis of continuous variables, I'll be exploring the correlation between variables and also check if the variables show skewness and kurtosis. I'll first implement the correlation matrix to our continuous variables and also plot the correlation plot and correlation chart.
The correlation matrix of the continuous variables:

	age	cigsPerDay	totChol	sysBP	diaBP	BMI	heartRate	glucose
age	1.00000000	-0.19195913	0.26387335	0.3902480	0.20378296	0.13639588	-0.00689402	0.11911522
cigsPerDay	-0.19195913	1.00000000	-0.03119569	-0.0947884	-0.05915929	-0.09296323	0.06915221	-0.05769427
totChol	0.26387335	-0.03119569	1.00000000	0.2157637	0.16824608	0.11436650	0.09412776	0.04667768
sysBP	0.39024801	-0.09478840	0.21576368	1.0000000	0.78509962	0.32772352	0.18247080	0.13299106
diaBP	0.20378296	-0.05915929	0.16824608	0.7850996	1.00000000	0.38052898	0.17708767	0.06014756
BMI	0.13639588	-0.09296323	0.11436650	0.3277235	0.38052898	1.00000000	0.07141530	0.08882778
heartRate	-0.00689402	0.06915221	0.09412776	0.1824708	0.17708767	0.07141530	1.00000000	0.09722613
glucose	0.11911522	-0.05769427	0.04667768	0.1329911	0.06014756	0.08882778	0.09722613	1.00000000

As we infer from the matrix above, there is no sign of extremely high correlation except for systolic and diastolic blood pressures. This correlation is pretty obvious because it is a part-wise measurement of the same variable "blood pressure". Beyond this obvious correlation, there is no evidence of high correlation or for that matter multi-collinearity. However we will check multi-collinearity in detail in our submission notes II. A correlation plot of the above matrix is as follows:

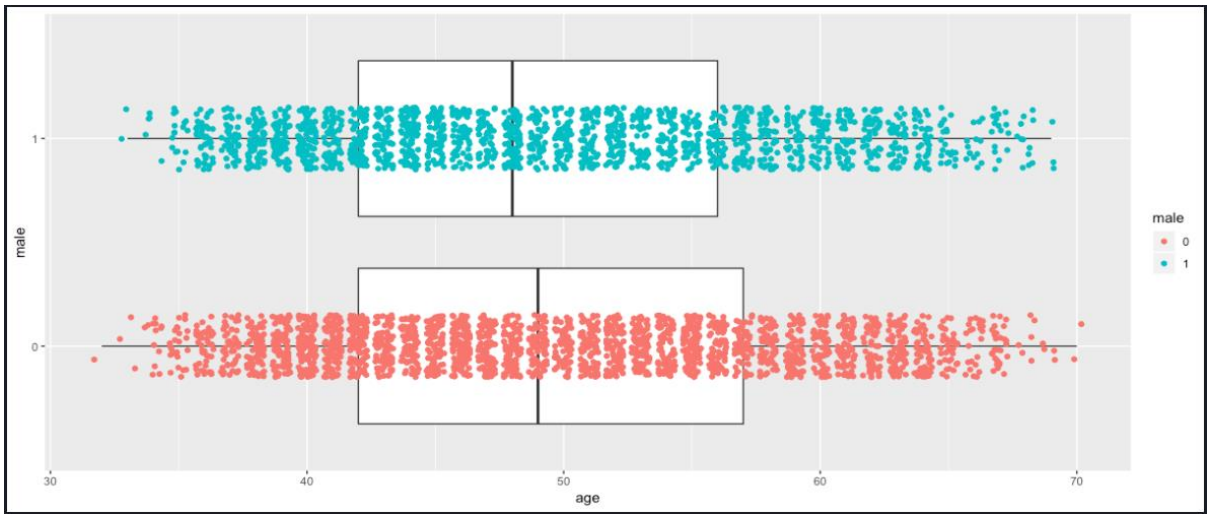


This plot is much easier to interpret. It is clearly evident what we observed in the matrix above.

Relationship between variables

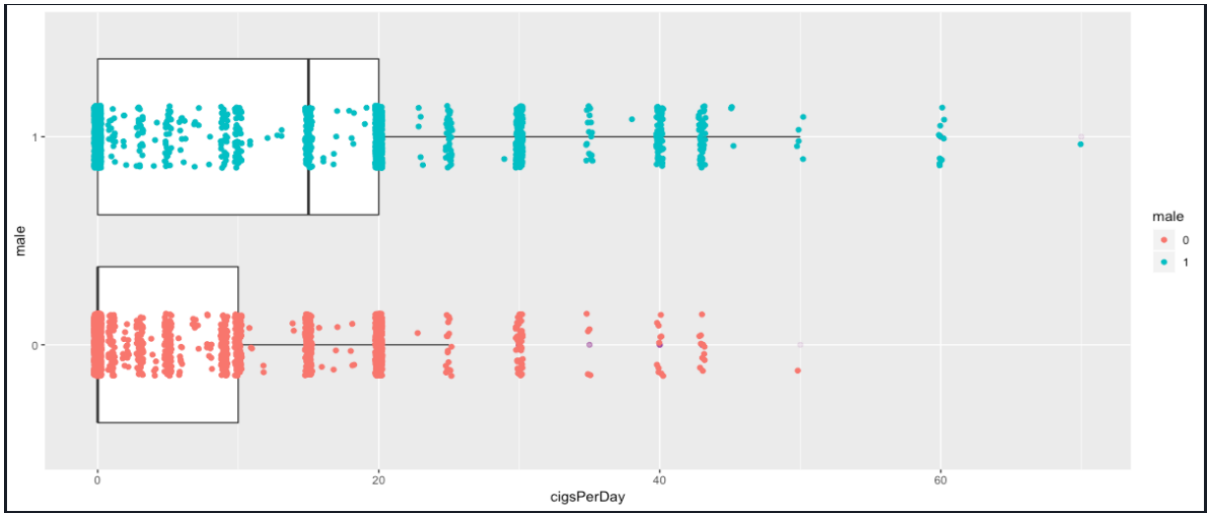
I will now explore two continuous variables with one categorical variable to further dive as to what insights we may find. I will start with bi-variate box-plots.

Male ~ Age



The above plot shows the spread of age grouped by gender. The categorical variable gender in this case has been distinctly coloured to differentiate the spread of data. As the age variable is generic, there are no significant inferences from the same.

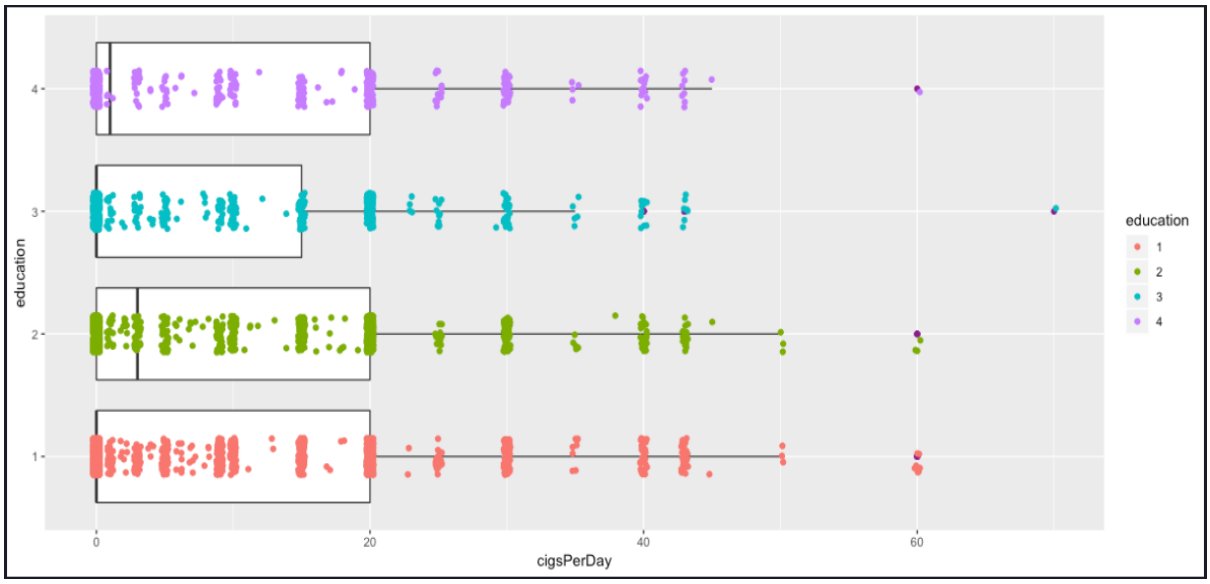
Male ~ cigarettes per day



The above plot shows the spread of number of cigarettes per day grouped by gender. As previously established, we know that number of cigarettes grouped by gender is a significant statistic. We can infer from the plot that, males are more likely to smoke higher number of cigarettes in a day compared to females. However we do observe high number for females in few cases.

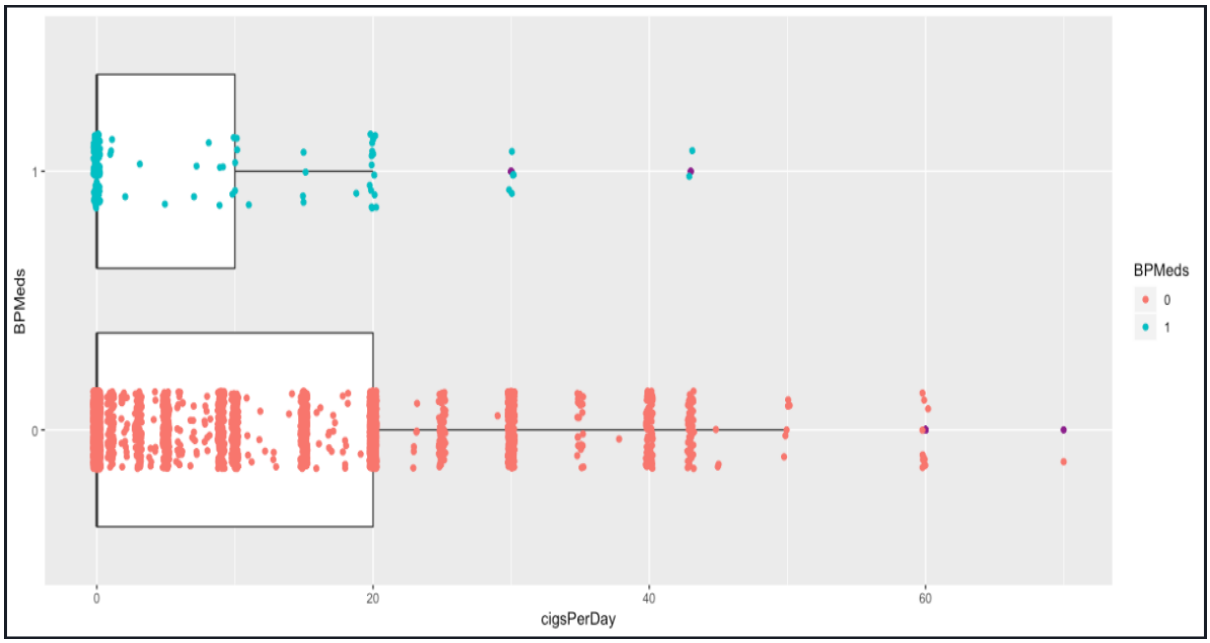
We had demarcated the box plot at mean value i.e. 9 cigarettes per day, however we can observe that there are a lot of observations which are higher than 9 for both males as well as females. It may prove our hypothesis incorrect, that education makes a person more vary about the parameters.

Education ~ Cigarettes per day



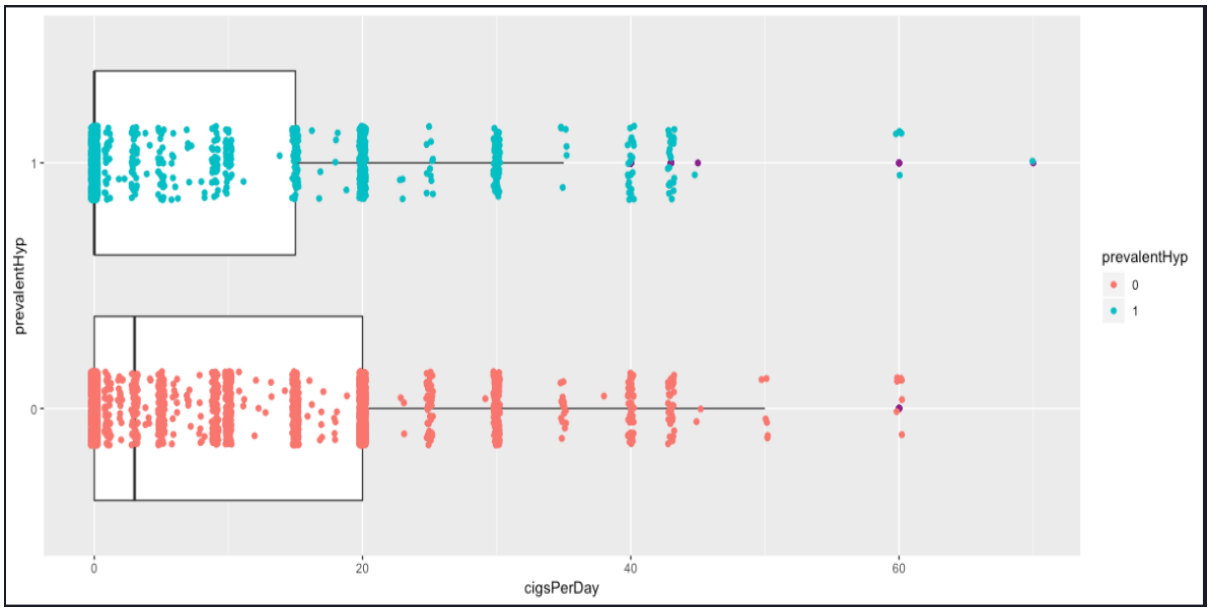
The above plot gathers the same hypothesis and displays the number of cigarettes smoked by a patient grouped by the highest level of education attained. As expected we do see a high number of cigarettes smoked for patients who have attained either only a high school or a bachelor degree and the number starts to decrease as the education level increases. However there is no significant evidence, which will arrive at a safe conclusion, hence as speculated, hypothesis proves incorrect.

BP medication ~ cigarettes per day



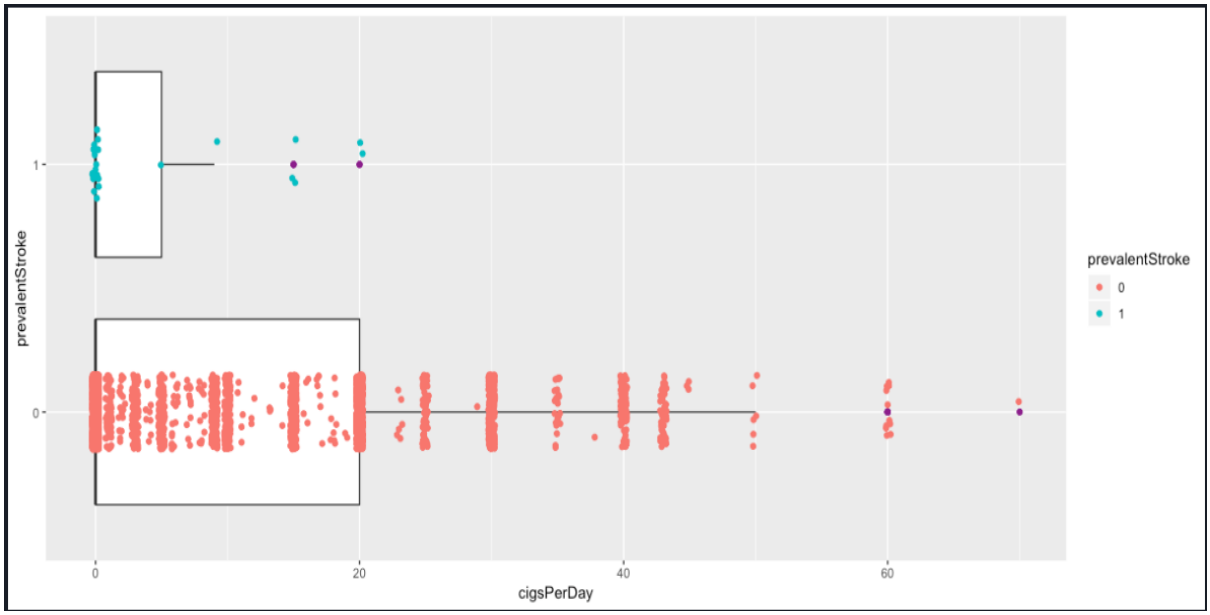
The above plot shows the correlation between patients who are on BP medication and the out of those who are currently smoking, how many cigarettes do they smoke in a day. In short, 1 refers to patients who are on BP medication, and 0 refers to patients who are not on BP medication. We see a significant shift, patients who are on BP medication have low number cigarettes smoked compared to the patients who are not prescribed to take any BP medication.

Prevalent hypertension ~ cigarettes per day



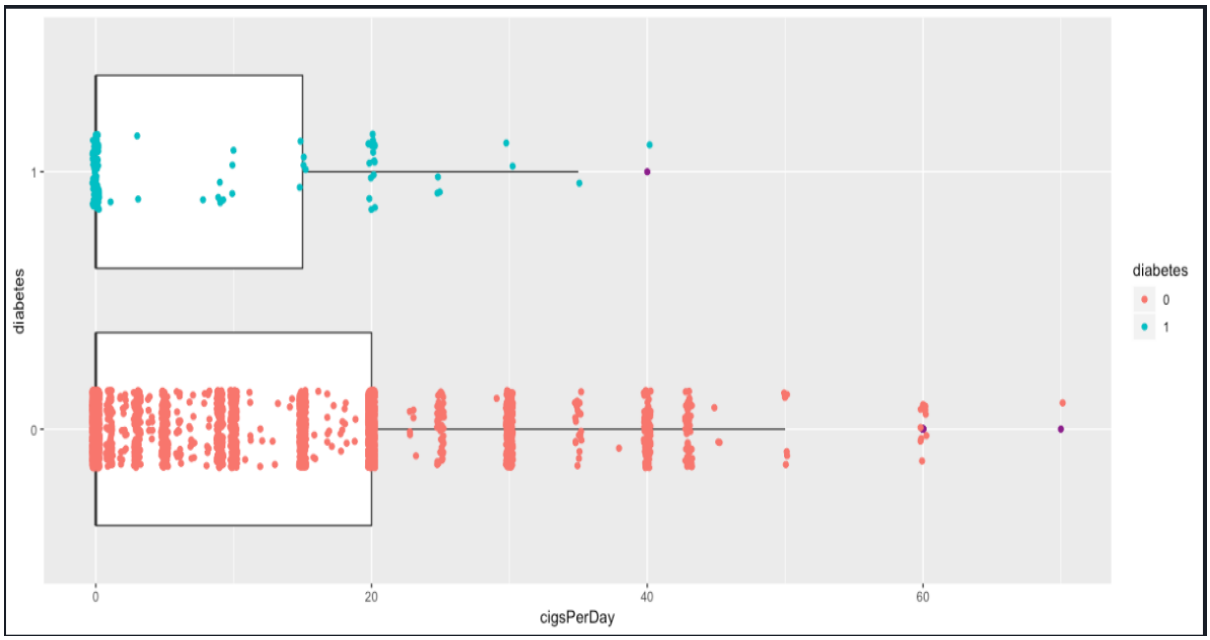
The above plot shows the correlation between patients who have been diagnosed with prevalent hypertension or high blood pressure and the patients who are currently smoking, how many cigarettes do they smoke. One interesting observation is that there is hardly any difference in the number of cigarettes the patients smoke despite they had hypertension previously. It may only in one case, that patients do not treat hypertension as a threat and continue to smoke.

Prevalent stroke ~ cigarettes per day



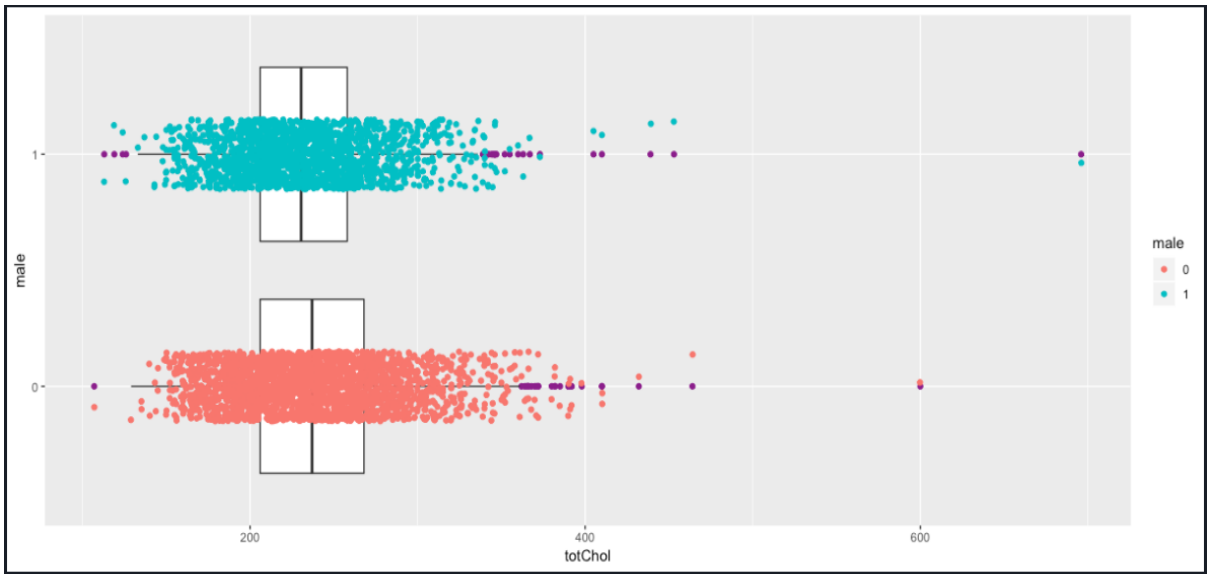
The above plot shows the correlation between patients who previously had a stroke and the number of cigarettes the patients smoke in a day. This is perhaps the most obvious inference. It is evident that patients who had a stroke previously, have realised it as a high threat and have stopped smoking, whereas in patients who did not have a stroke continue to smoke as they used to. A high impact, but a disappointing trend, that patients give up on smoking only in the 11th hour.

Diabetes ~ cigarettes per day



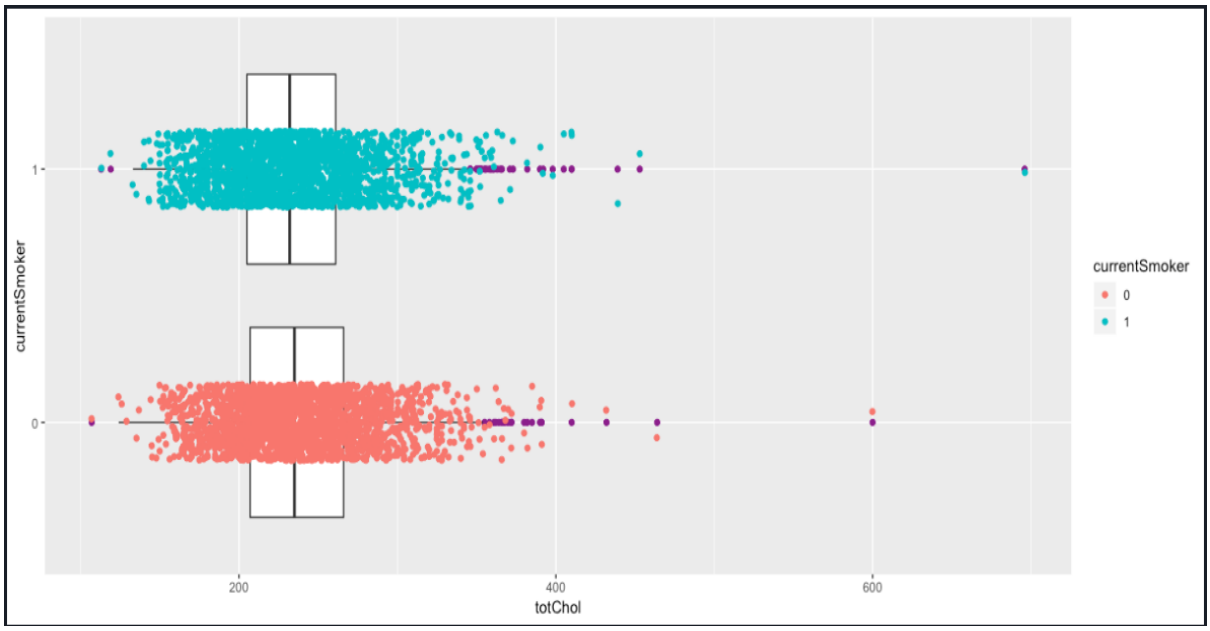
The above plot shows the correlation between patients who have been diagnosed with diabetes and the number of cigarettes smoked by patients. We observe that despite high BMI values, very few patients have actually been diagnosed with diabetes. Out of those patients, there are few patients who still have not given up on the smoking and a similar pattern is observed where the patients have not been diagnosed with diabetes at all.

Male ~ total cholesterol



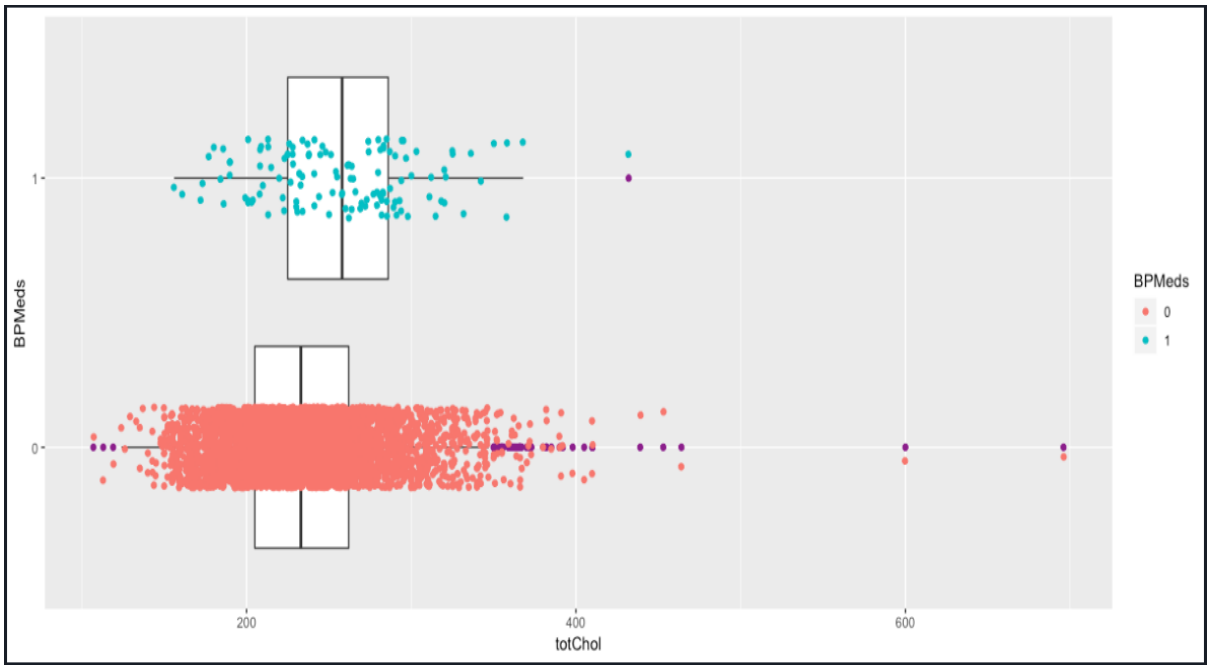
The above plot shows correlation between total cholesterol levels of the patients w.r.t gender. As we know, 0 refers for a female patient and 1 refers for a male patient. The spread for cholesterol levels for both males and females is around 240, surprisingly it is slightly lower for males. The high total cholesterol level is however an incorrect indication of a contributing factor, as it is further categorized in low density and high density cholesterol, where high density is beneficial to us, and low density is harmful.

Current smoker ~ total cholesterol



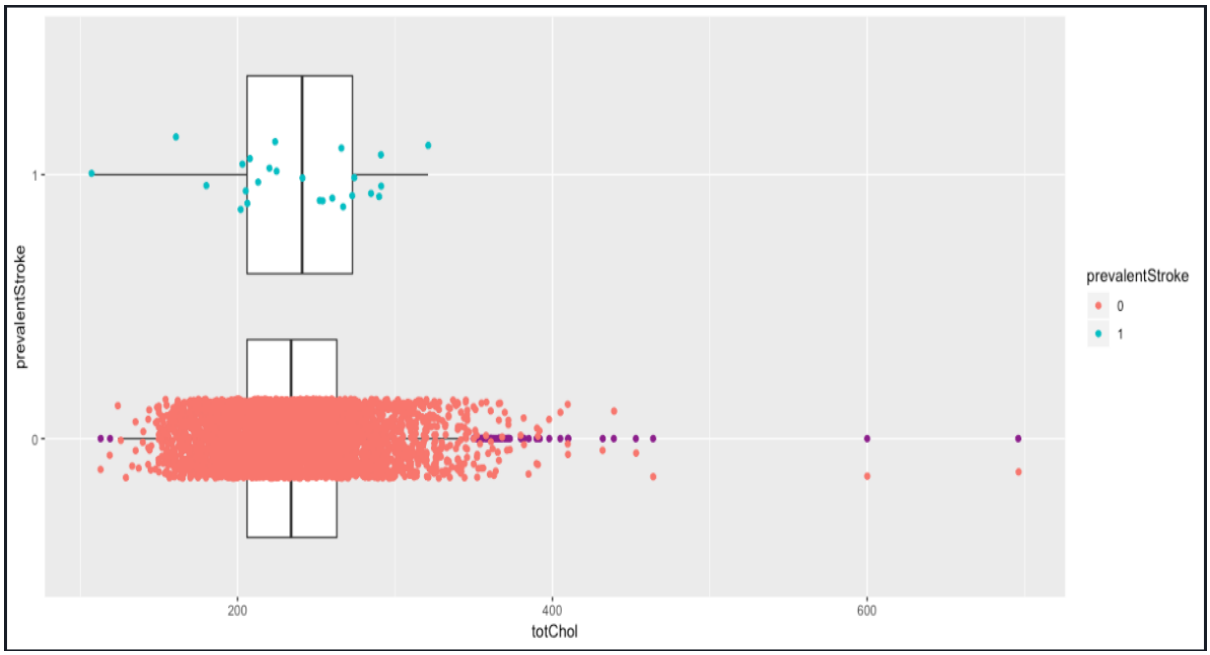
The above plot shows the correlation between whether a patient is a current smoker and the total cholesterol level of the patient. According to the plot, there is absolutely no evident difference in the pattern of total cholesterol level, whether or not a patient smokes, hence we can rule out the correlation between the total cholesterol and current smoker.

BP medication ~ total cholesterol



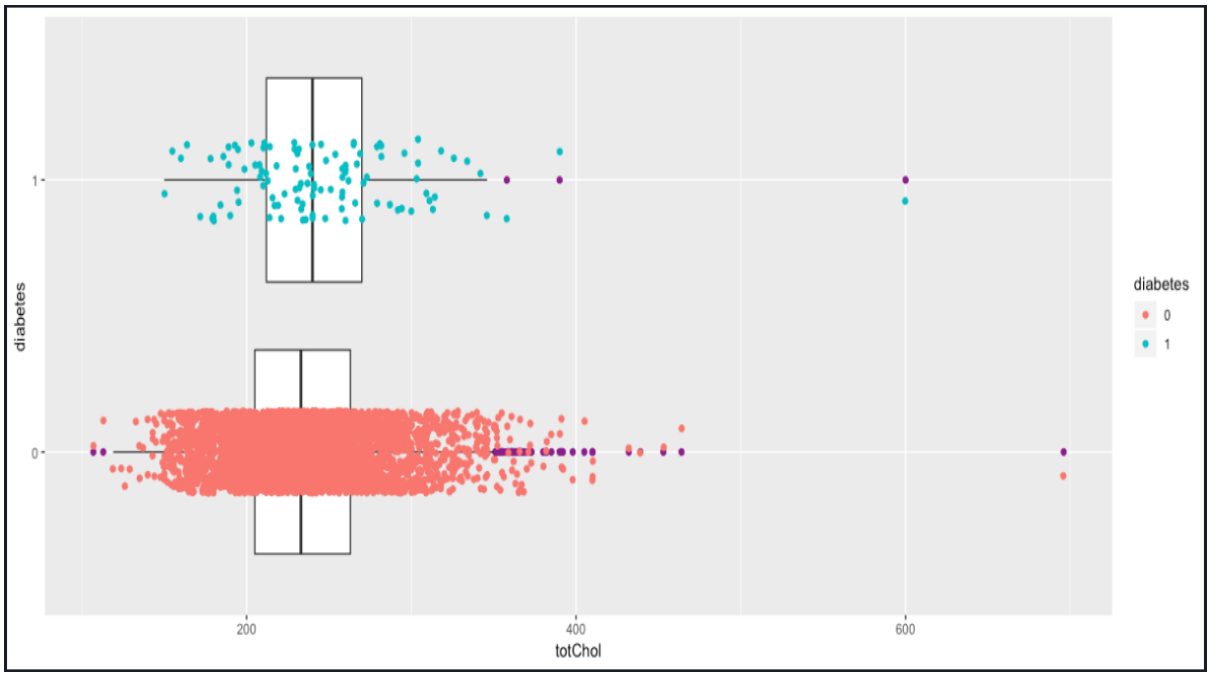
The above plot shows the correlation between whether or not a patient is prescribed for BP medication and the total cholesterol level of the patient. We observe a drastic change, the cholesterol levels for patients who are not on BP medication is lower compared to those who are on BP medication. This is an interesting find as the patients who are on BP medication are more likely to have a controlled intake on diet, ideally lowering the cholesterol levels.

Prevalent stroke ~ total cholesterol



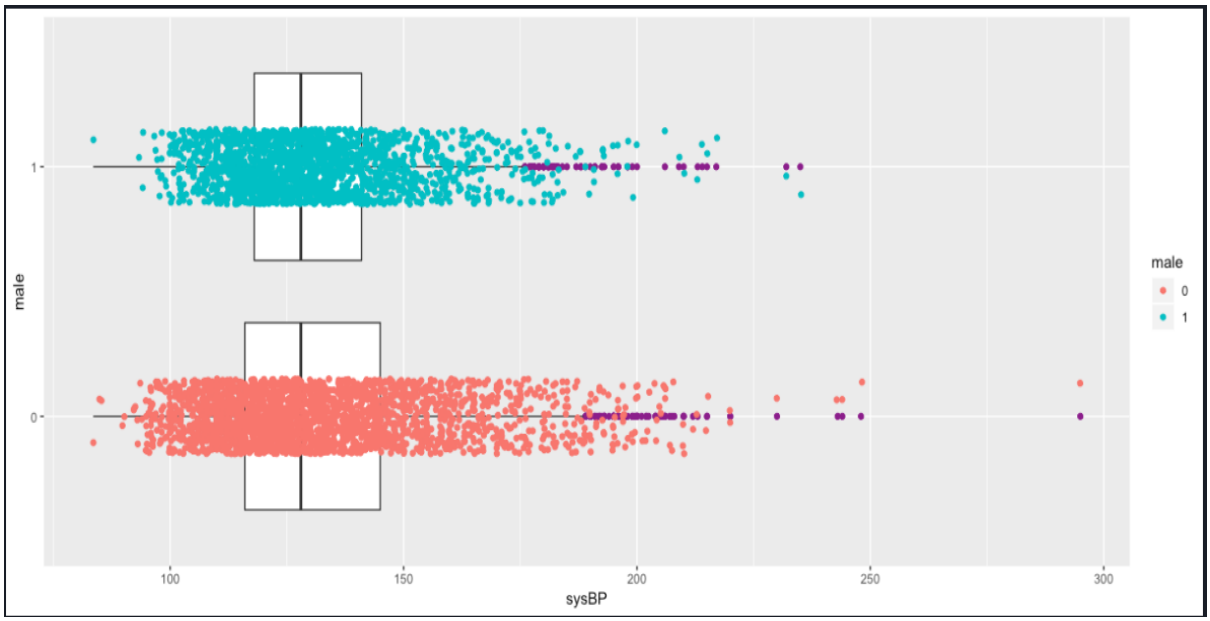
The above plot shows the correlation between the patients who had a stroke previously and the total cholesterol level. We observe that there exist a huge number of patients who have a high total cholesterol level, who did not have a stroke previously. However the number is sparse, when it comes to cholesterol level of patients who had a stroke. The finding is somewhat obvious.

Diabetes ~ total cholesterol



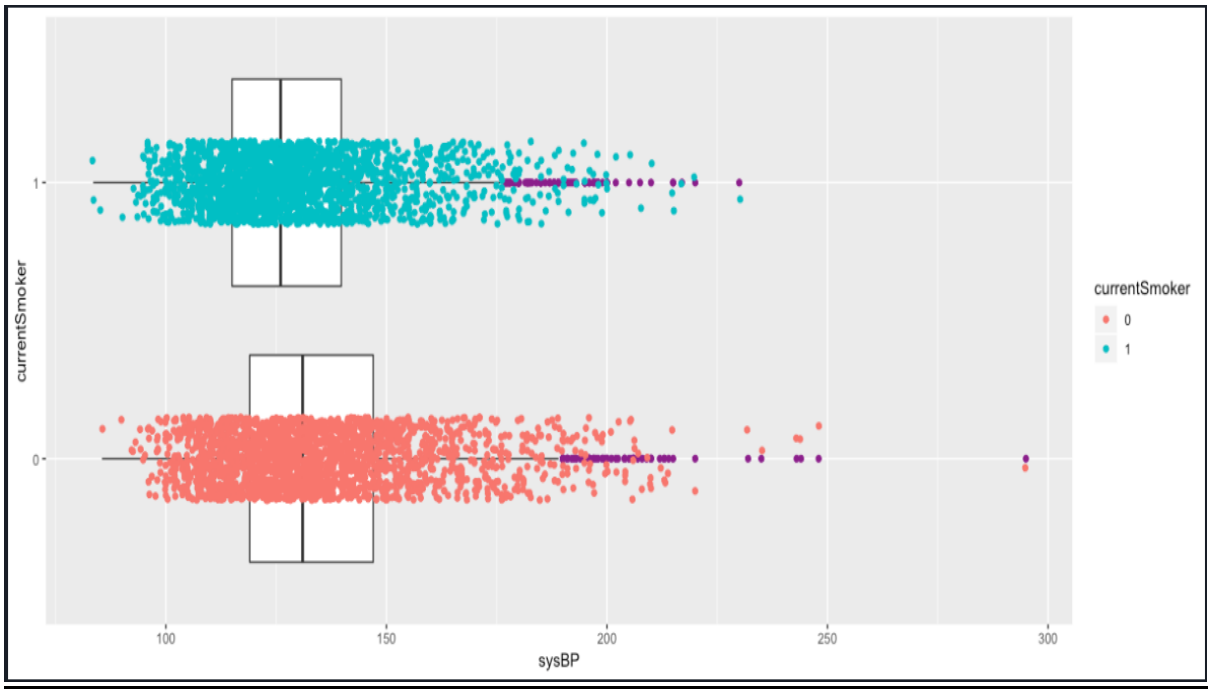
The above plot shows the correlation between the patients who are diagnosed with diabetes and the total cholesterol levels of the patients. We infer that the cholesterol levels for both categories is more or less the same and thus no evident correlation established.

Male ~ systolic blood pressure



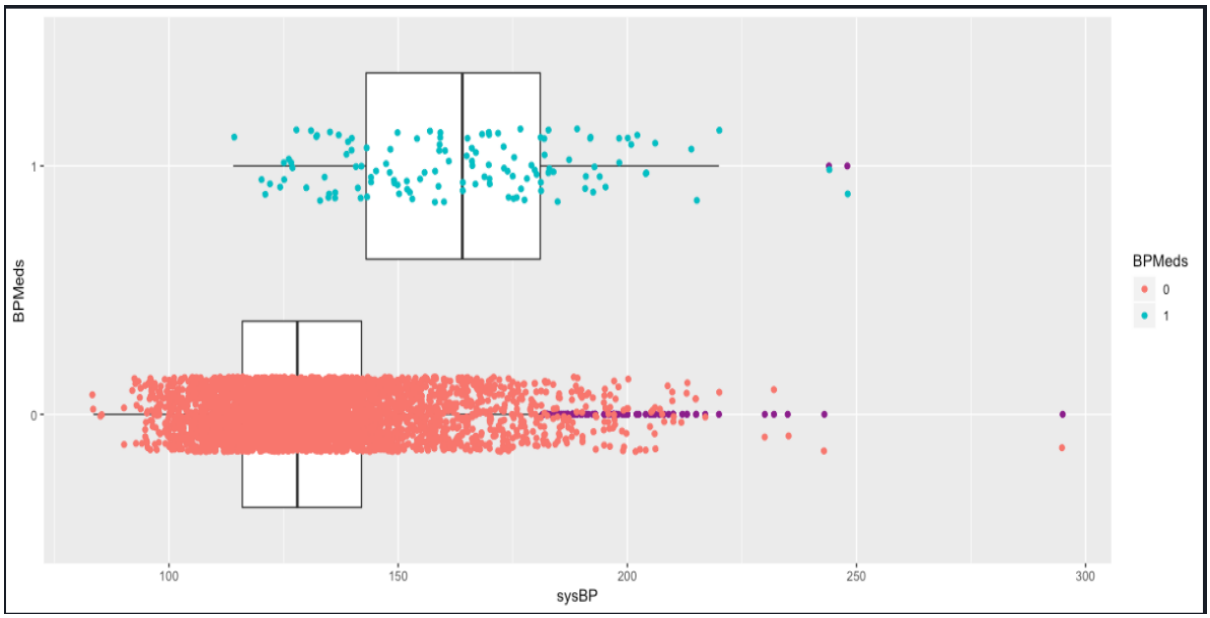
The above plot shows the correlation between the systolic blood pressure level differences grouped by gender. We infer no such difference in the blood pressure levels of males to females. The normal range of the variables is around 125 for both and the higher spreads also show a similar pattern. We will look more detailed view of how blood pressure varies with other parameters.

Current smoker ~ systolic blood pressure



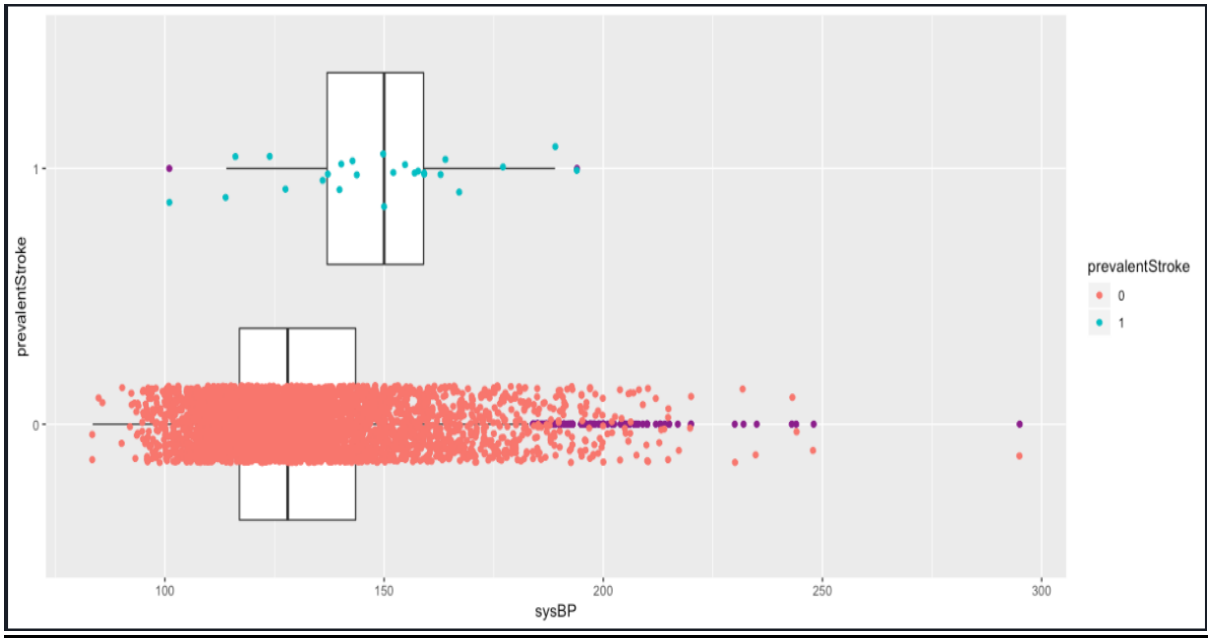
The above plot shows the correlation between patients who are current smokers and the systolic BP levels of the patients. We infer that the systolic BP of the patients who smoke is lower compared to the patients who do not smoke. It means that the smoking causes the heartrate to drop, thereby lowering the systolic BP.

BP medication ~ systolic BP



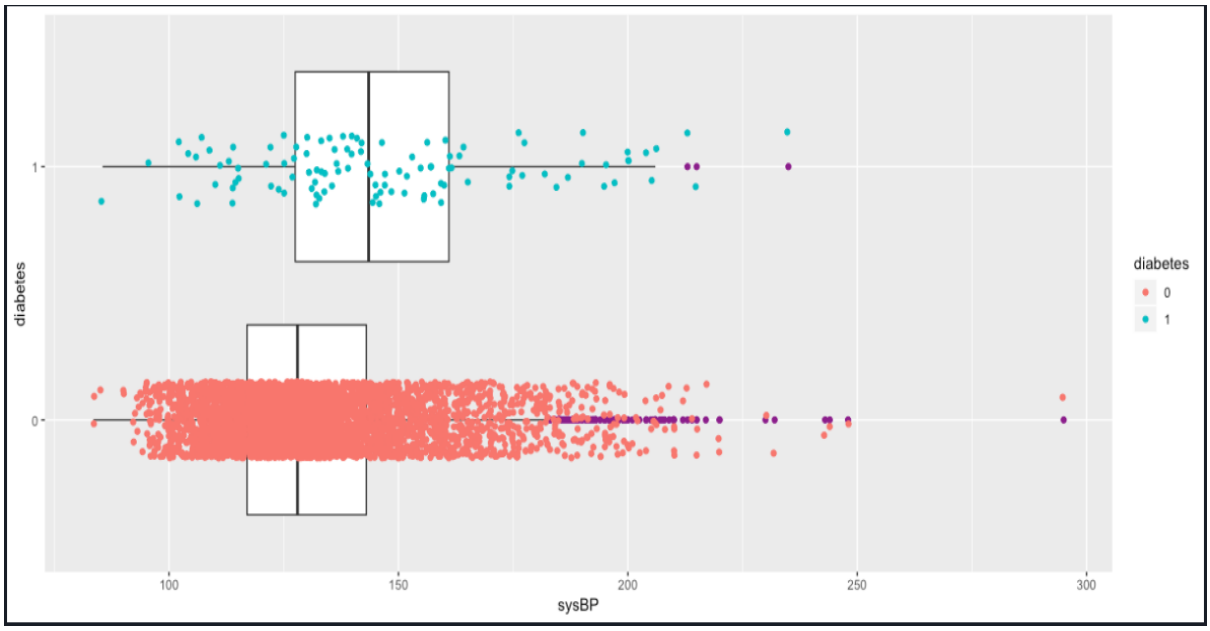
The above plot shows the correlation between whether the patients are currently prescribed on BP medication and the systolic BP level of the patients. We infer that the patients who are on BP medication have a high systolic BP compared to the rest of the patients. It may be because of the high BP levels they have been prescribed the medication or the medicines are elevating the systolic BP by inducing forced pumping to the heart.

Prevalent stroke ~ systolic BP



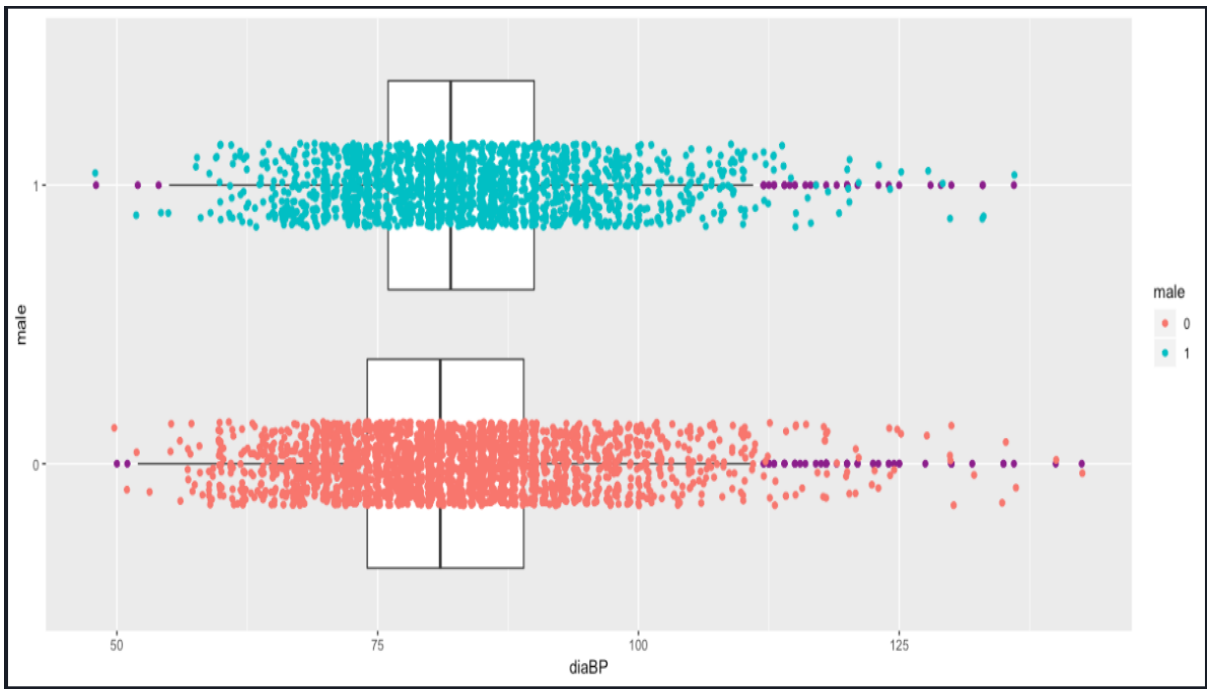
The above plot shows the correlation between patients who have had a stroke and the systolic BP measurements of the patients. We infer that the systolic BP of patients who had a stroke before is significantly higher compared to the other patients. This is because the having a stroke indeed contracts the arteries and reduces the flow to the heart, therefore making the heart pump harder to generate enough blood flow, thus increasing the systolic blood pressure.

Diabetes ~ systolic blood pressure



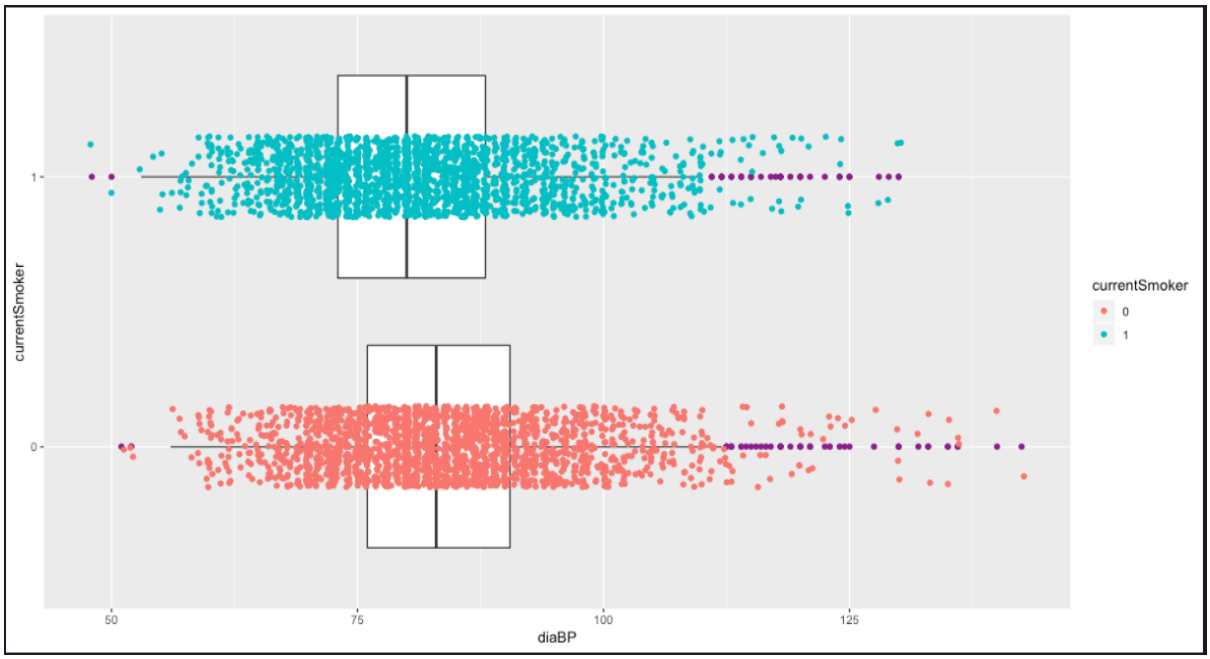
The above plot shows the correlation between the patients who have been diagnosed with diabetes and the systolic blood pressure of the patients. Similar to what we established with stroke, a patient who is diagnosed with diabetes, will have an increased blood pressure because to neutralise the insulin the heart must pump harder thereby increasing the systolic blood pressure.

Male ~ diastolic blood pressure



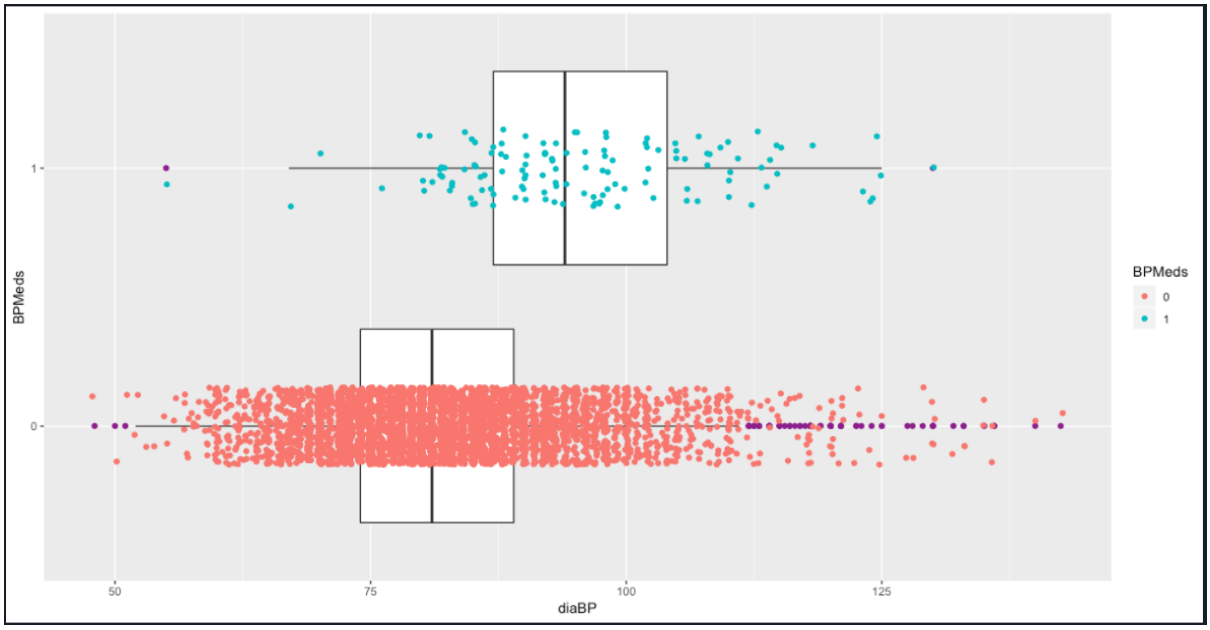
The above plot shows the diastolic blood pressure levels of the patients grouped by gender. As we observed for systolic BP, the diastolic BP levels are almost the same for both. Here the important point is to note the upper limit of the safe threshold. Unlike systolic, diastolic blood pressure levels have low tolerance for extreme values. Thus a more in-depth analysis will help us further.

Current smoker ~ diastolic BP



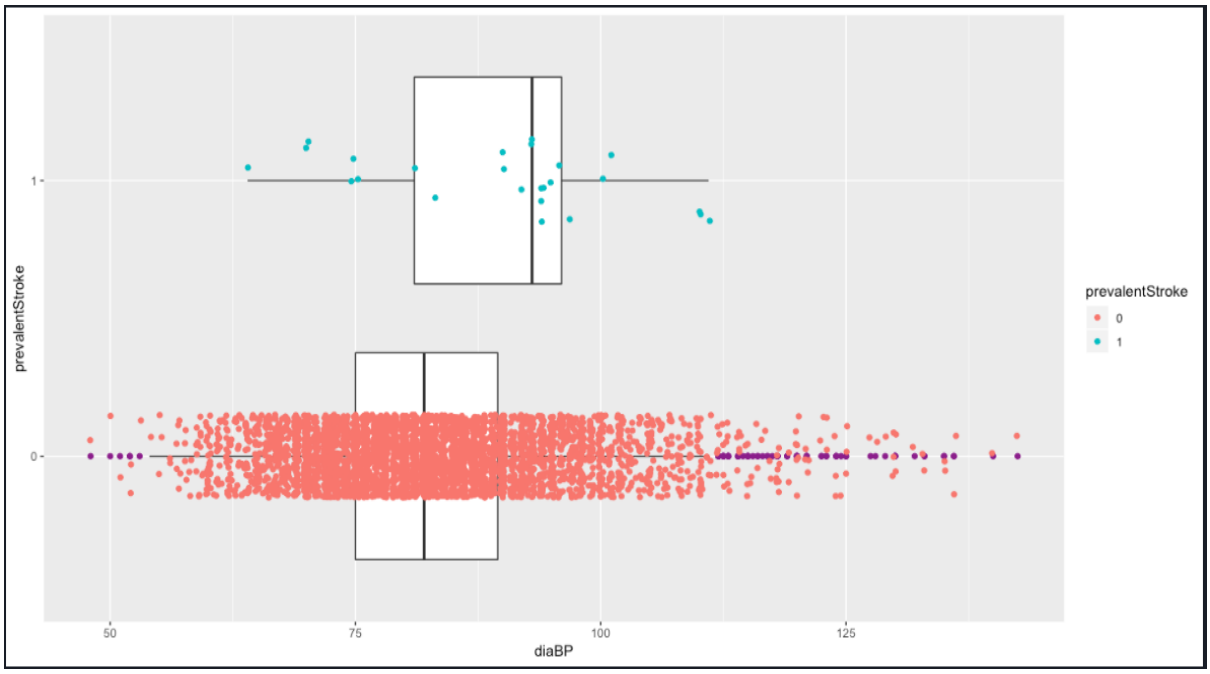
The above plot shows the correlation between the diastolic BP level of the patients and whether the patients are currently smoking or not. We infer that the diastolic BP levels for patients who smoke are lower than for rest of the patients. The reason for it is the inefficiency of the heart. Smoking long term does have a significant effect on the pumping abilities of the heart, thus reducing stamina and longevity. Reduced work levels mean higher the resting heartrate, thus lower the BP.

BP medication ~ diastolic BP



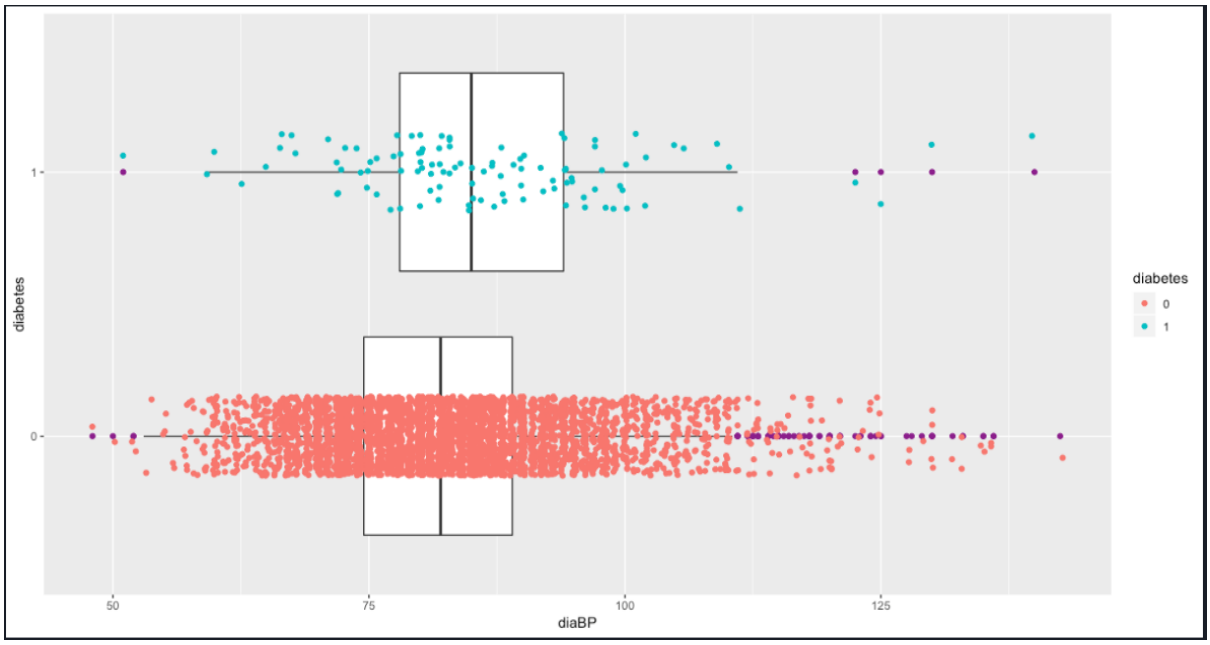
The above plot shows the correlation between the diastolic BP of the patients and whether the patient is currently prescribed on BP medication. The inference is exactly the same as we achieved with systolic BP, the BP levels will be elevated due to the medication and not so much for the rest of the patients who are not on BP medication.

Prevalent stroke ~ diastolic BP



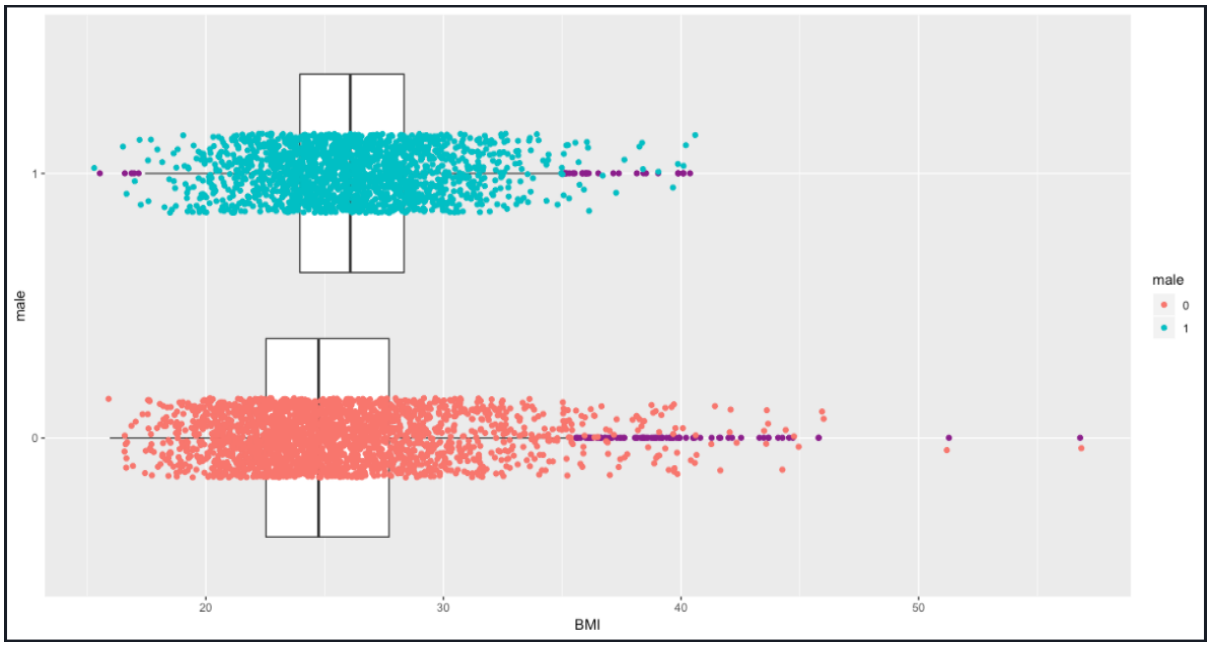
The above plot shows the correlation between the patients who had a stroke previously and the diastolic BP levels of the patients. We infer that the BP for patients who had a stroke previously is much higher than the rest of the patients. The distribution of points is similar to what we observed in systolic blood pressure. The heart which is already weak, goes through a lot of pumping activity in order to efficiently circulate blood flow, thereby increasing the diastolic BP

Diabetes ~ diastolic BP



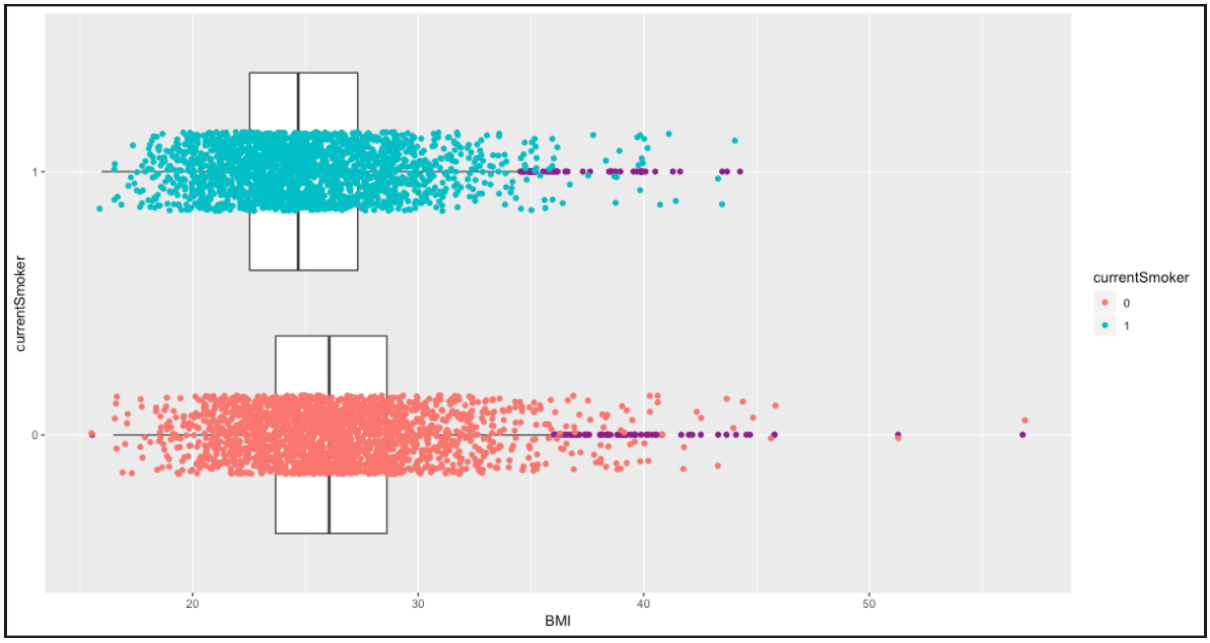
The above plot shows the correlation between the diastolic BP of the patients and whether the patients have been diagnosed with diabetes. The BP levels of diabetic patients is slightly higher than the remaining patients because of increased sugar and glucose levels.

Male ~ BMI



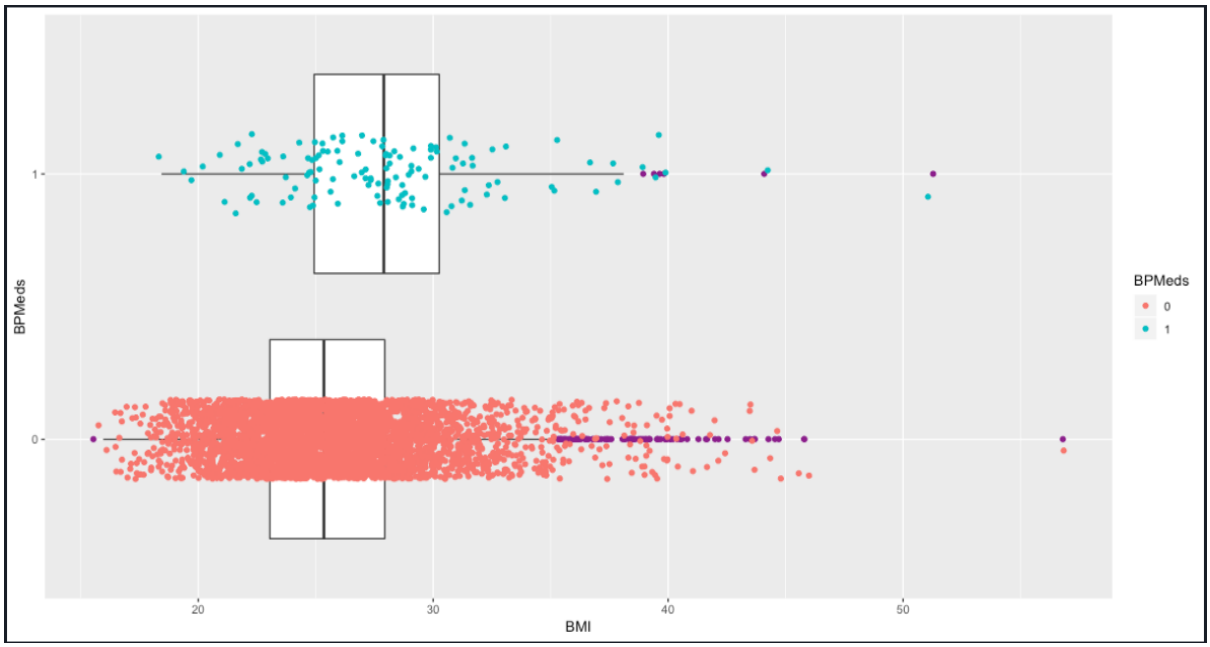
The above plot shows the relation between the BMI of the patients grouped by gender. We infer that the BMI levels for males are higher than of females, which is a well-known fact. The ideal range for males is 25-28 and for females is 19-23 and we infer that both of the ideal ranges are perfectly met. The distribution is slightly spread for females mostly because of ranging heights. We also observe a few extreme outlier values, which probably have the highest risk of a heart disease.

Current smoker ~ BMI



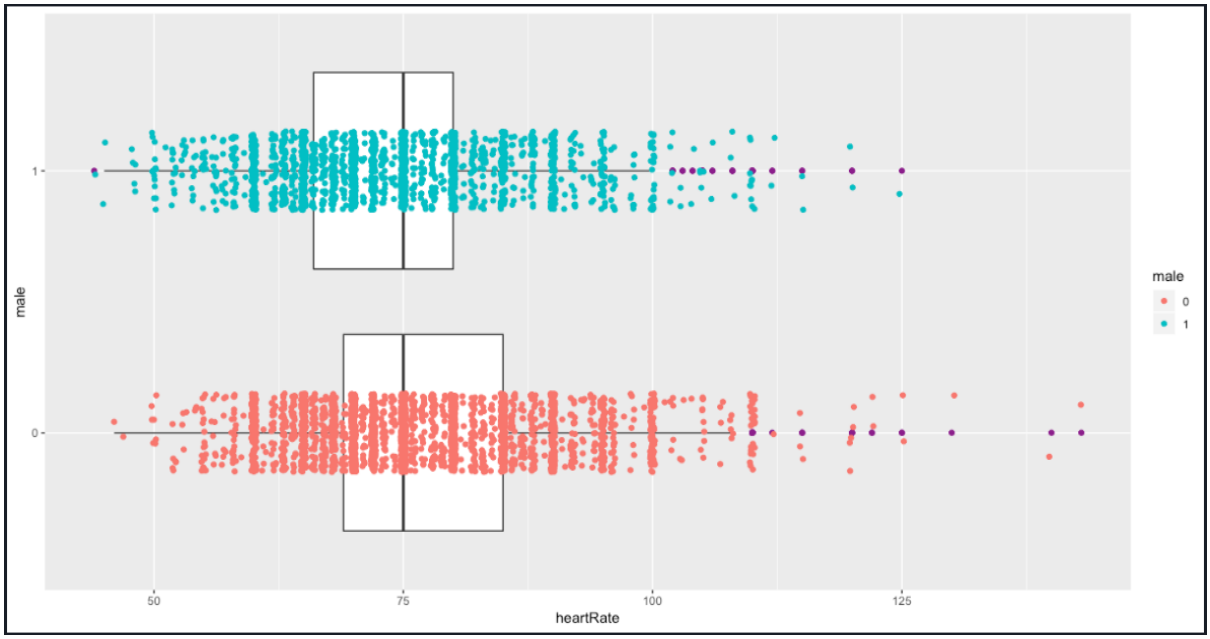
The above plot shows the correlation between the BMI of the patients and whether the patient is a current smoker or not. We infer that the BMI for patients who smoke is less compared to those who do not smoke, which is likely because smoking causes loss of hunger and makes a person more thirsty, thereby increasing the water intake and reducing the food intake.

BP medication ~ BMI



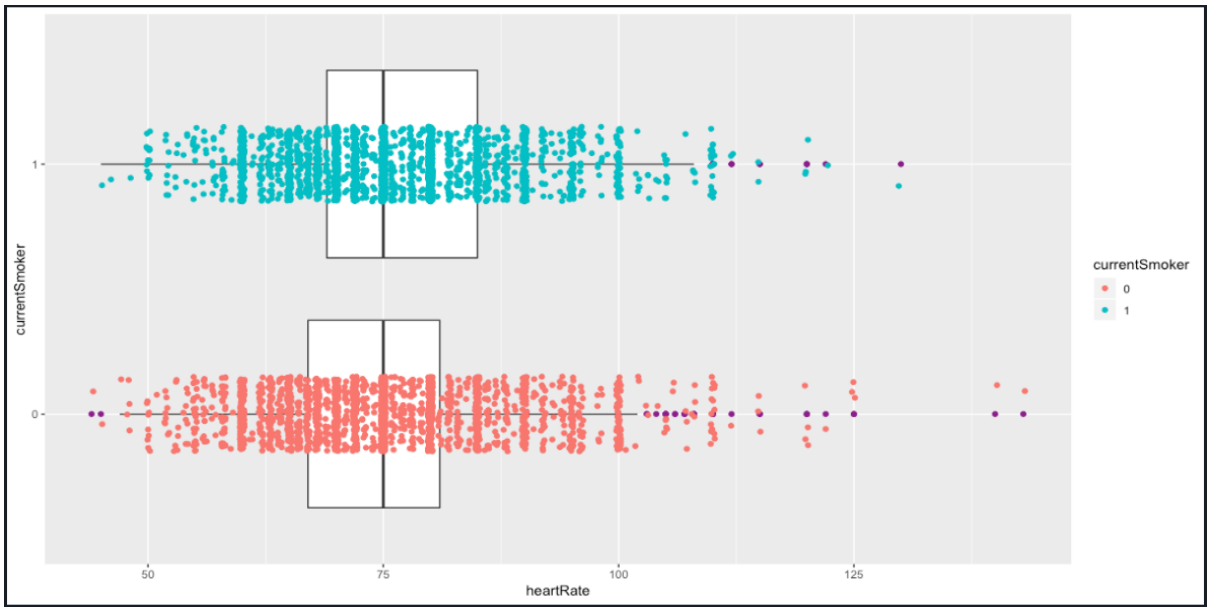
The above plot shows the correlation between the BMI levels of the patients and whether the patients are prescribed on BP medication. We infer that the patients who are on BP medication have higher BMI levels compared to those who are not on BP medication. This is because, the increased BP must be a symptom of obesity or being overweight. Thus having a BMI value which is larger than the generally accepted safe levels for males and females.

Male ~ heartrate



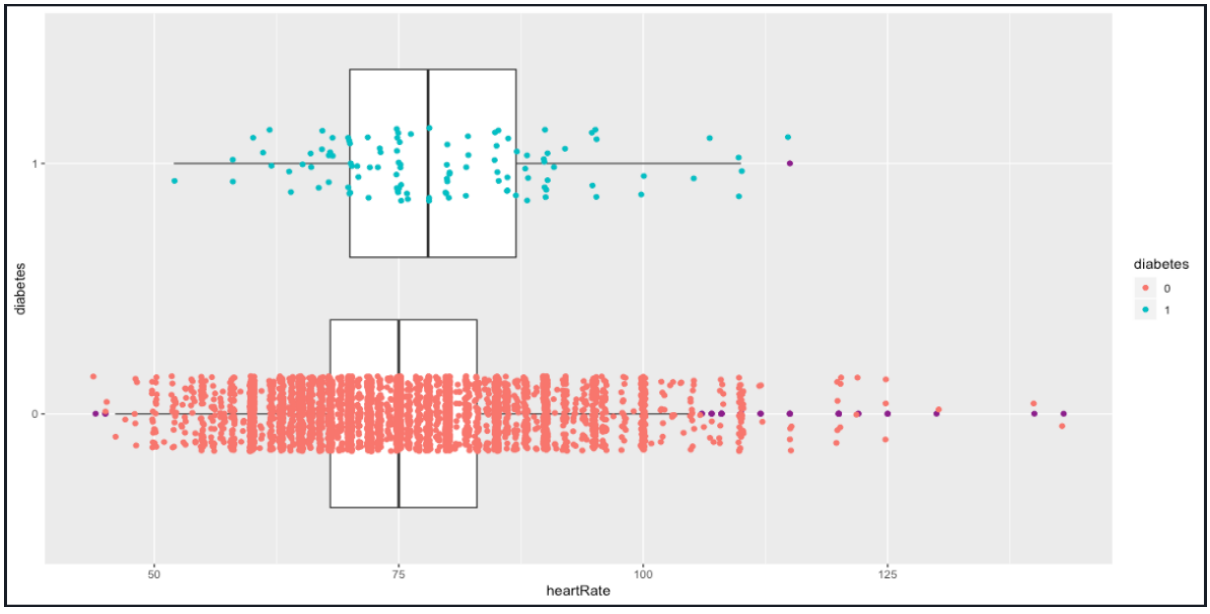
The above plot shows the relation between the heartrate of the patients grouped by gender. If closely observed there is no difference in the heart rates of male and female patients. The only difference being the spread of female patients is on the larger side after the median and exactly opposite when compared of males.

Current smoker ~ heartrate



The above plot shows the correlation between the heartrate of the patients and whether a patient is or not a current smoker. We infer that even though there is no difference in the heart rates of the patients who smoke and do not smoke, one slight difference is the spread of observations tends to be upward for patients who smoke. It might be because smoking does not cause a gradual rise or decline in heartrate. The heartrate measurement is likely to change only if the patient has smoked just prior to the heartrate measurement and not otherwise.

Diabetes ~ heartrate



The above plot shows the correlation between the heartrate of the patients and whether or not a patient is diagnosed with diabetes. We infer that the patients who are diagnosed with diabetes tend to exhibit a higher heartrate compared to the non-diabetic patients, as diabetic patients will have the BP levels elevated and in order to tackle the insulin from the body, the heartrate or the pulse of diabetic patients will be higher compared to the other patients.

d) Categorical variable bivariate analysis:

In bivariate analysis of categorical variables, I'll be exploring the correlation between variables with the help of "Pearson's chi-squared test of independence". The chi-squared test assumes a null hypothesis and an alternate hypothesis. The general practice is, if the p-value that comes out in the result is less than a pre-determined significance level, which is 0.05 usually, then we reject the null hypothesis.

H0: The two variables are independent
H1: The two variables are dependent

The null hypothesis of the chi-squared test is that the two variables are independent and the alternate hypothesis is that they are related.

I have compared a few combinations of categorical variables, which may contribute significantly towards CHD. I'll compare the p-values of each of the following against our threshold and accordingly accept or fail to accept our null hypothesis.

```
Correlation between categorical variables
```{r}
chisq.test(data.cat$currentSmoker, data.cat$prevalentStroke, correct = FALSE)

chisq.test(data.cat$currentSmoker, data.cat$prevalentHyp, correct = FALSE)

chisq.test(data.cat$BPMeds, data.cat$prevalentStroke, correct = FALSE)

chisq.test(data.cat$BPMeds, data.cat$diabetes, correct = FALSE)
```
```

```
        Pearson's Chi-squared test

data:  data.cat$currentSmoker and data.cat$prevalentStroke
X-squared = 4.6119, df = 1, p-value = 0.03175


        Pearson's Chi-squared test

data:  data.cat$currentSmoker and data.cat$prevalentHyp
X-squared = 45.605, df = 1, p-value = 1.447e-11

Chi-squared approximation may be incorrect
        Pearson's Chi-squared test

data:  data.cat$BPMeds and data.cat$prevalentStroke
X-squared = 57.679, df = 1, p-value = 3.086e-14

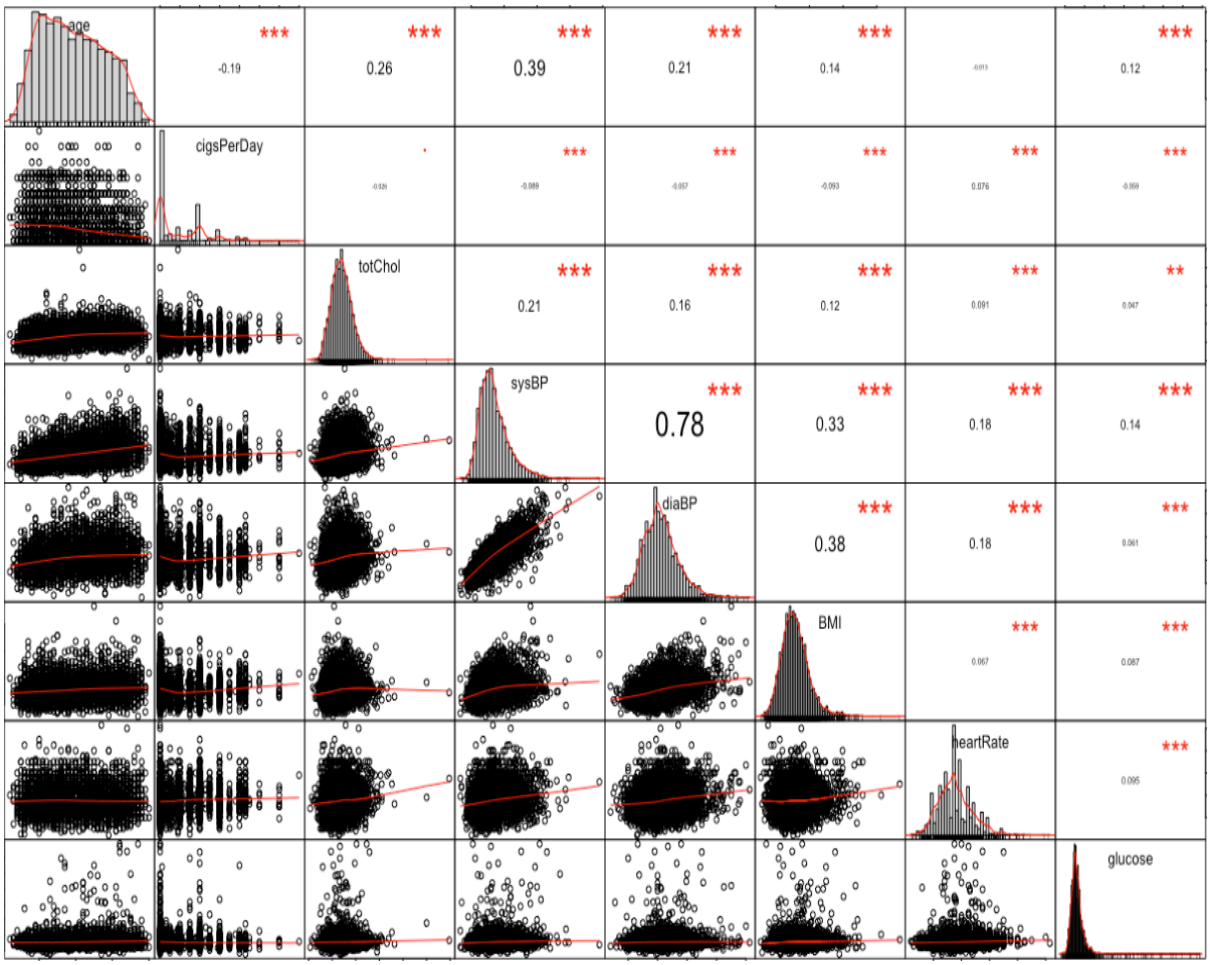
Chi-squared approximation may be incorrect
        Pearson's Chi-squared test

data:  data.cat$BPMeds and data.cat$diabetes
X-squared = 11.348, df = 1, p-value = 0.0007555
```

We can infer from the above statistic, that each of the 4 tests conclude that the p-value is less than the threshold value of 0.05, we can safely reject the null hypothesis and thus affirming that alternate hypothesis is true, i.e. the variables current smoker and prevalent stroke are dependent, similarly current smoker and prevalent hypertension are dependent.

e) Multivariate analysis

To conclude our correlation between variables, I'll plot a correlation chart, which includes a few additional details such as the significance of variables w.r.t p-values, the scatter plot distribution between variables with a regression (ab-line) line and also the histogram frequency distribution between variables with a density plot.



As explained above, the scatter plot distribution between variables also show correlation order. The regression line in the scatter plots, show how high or low the correlation is between variables. Consider, the scatter plot between systolic and diastolic blood pressure, it is the only plot where a significant slope is evident on the regression line, meaning high correlation between them. The other regression lines are pretty much flat with very little slope, meaning correlation values closer to zero i.e. low correlation between variables.

The histograms explain the frequency distribution and density plots explain the spread of the variables. As we observe, most of the variables have a density plot, which shows a relative right or positive skew. It means that the mean values are greater compared to the median values. Visual inspection shows that the variables are moderately skewed i.e. have skew values between the range of [-1, -0.5 to 0.5, 1]. Glucose however shows a high skew value.

The density plots also helps us identify the kurtosis. Except for age, all other variables show a right tail i.e. positive excess kurtosis or "leptokurtic". In general a base value of 3 is considered as normal distribution or "mesokurtic". Glucose again shows extreme right tail, meaning that the value is way greater than 3. This also confirms our initial observation that glucose level is the variable which is exhibiting high skew and kurtosis values, which in turn can produce extreme outlier values.

Data Pre-Processing:

f) Skewness & kurtosis:

Skewness refers to a measure of lack of symmetry. In statistics, we assume every data to be normally distributed. However this scenario is ideal and most of the times we have data which is skewed. If the data is not normally distributed, we have two possible outcomes, the data is either skewed positively or negatively.

Kurtosis refers to a measure the weight of the lack of symmetry i.e. how exactly is the skew, if the data is skewed, then we know that mean and the median are different, which has two possible scenarios,

1. Mean is less than the median
2. Mean is more than the median

In either scenario, the data will leave a tail and the measure of whether the tail is light or heavy is nothing but kurtosis. Like skewness, kurtosis also can be positive or negative. If the situation 1 is true then the data is left skewed or negative kurtosis, whereas if the situation 2 is true then the data is right skewed or positive kurtosis.

- If the kurtosis value is equal to 0, then the data is said to be mesokurtic
- If the kurtosis value is greater than 3, then the data is said to be leptokurtic
- If the kurtosis value is less than 3, then the data is said to be platykurtic

Similarly,

- If the skewness is less than -1 or greater than 1, then the data is highly skewed
- If the skewness is between -1 to -0.5 or 0.5 to 1, then the data is moderately skewed
- If the skewness is between -0.5 to 0.5, then the data is in symmetry

Now let us analyse the dataset for skewness and kurtosis:

```
      age cigsPerDay    totChol      sysBP      diaBP        BMI  heartRate    glucose
0.2287861  1.2466081  0.8715684  1.1448798  0.7129979  0.9818342  0.6441438  6.2125279
      age cigsPerDay    totChol      sysBP      diaBP        BMI  heartRate    glucose
 2.009857   4.016784   7.123531   5.152666   4.272396   5.652742   3.904911  61.626013
```

Skewness interpretation : The distribution of the data is for all variables slightly skewed towards right, except for glucose which has a significant right skew. So data is positively skewed.

Kurtosis interpretation : The distribution is a mixture of negative and positive excess kurtosis. Only Age shows an excess kurtosis value of 2.0098 which is < 3, hence is a negative excess kurtosis or simply put - platykurtic. Meaning that Age variable shows a thinner tail and therefore has a less chance of producing outliers compared to a normal distribution.

However all other variables show an excess kurtosis value > 3, which implies a positive excess kurtosis or - leptokurtic. Meaning that these variables show a comparatively fatter tail and therefore can produce outliers compared to a normal distribution.

Important factor however is the extent of excess kurtosis, all variables except glucose, are in close proximity to 3 (Normal distribution), hence will produce outliers which might not affect the dataset, glucose on the other hand has a positive excess kurtosis value of 61.62 which is way greater than 3, thus can produce extreme outliers to significantly affect the data, unless treated accordingly. This also concludes that the variable glucose is red-flagged by both skewness and kurtosis values.

g) Missing value identification and treatment:

Missing values are common and can be induced in a dataset by human error, unrecorded data etc. However there are types of missing values: Let us look into a bit more detail:

MCAR – Missing completely at random
MAR – Missing at random
NMAR – Not missing at random

I have first identified missing values and then treated them using MICE package. MICE is an acronym for multivariate imputation of chained equations and one of the most popular missing data imputation packages in R. As most of our missing data was continuous as well as categorical, I used the PMM – predictive mean modelling for continuous data and logistic regression for binary classification, and polynomial regression for multi-level classification.

I also categorized the missing values in the types mentioned above:

I identified 3 variables which were MAR (Missing at random):

- Education
- BP medication
- Total cholesterol

I identified 3 variables which were MCAR (Missing completely at random):

- BMI
- Heartrate
- Glucose

Only 1 variable was NMAR (Not missing at random)

- Cigarettes per day

I have imputed the missing data and checked if any data is still missing.

```
 age cigsPerDay    totChol      sysBP      diaBP        BMI  heartRate    glucose
   0          0          0          0          0          0          0          0
```

```
male   currentSmoker         BPMeds prevalentStroke    prevalentHyp       diabetes      education
   0              0              0                0              0              0              0
```

h) Outlier identification and treatment:

Similar to missing values, outliers were present in the data. As we observed in our previous analysis, I have identified the outliers in multiple ways using QQ-plots and box-plots. I also marked the outlier values in box-plots so that they can be capped from the data before modelling.

In terms of treatment of outliers, I treated the outliers with capping off the extremes found from the QQ-normal and QQ-line plots. There could have been an alternate way to handle, is to replace outliers with missing values and impute them using MICE.

i) Data balancing:

We inferred from our categorical variable analysis that, few variables did exhibit high data imbalance. It is referred to as a mis proportion of classes in our variables. Prevalent stroke and BP medication are the 2 variables with highest level of data imbalance.

There are several ways to treat the imbalanced classes, a few of them are:

Random oversampling : In random oversampling, we oversample the minority class. This is a resampling technique which will alter the dimensions of the data. However as we resample the data with random observations, there is a high chance that model will overfit the data.

Random under-sampling : In under-sampling, we under-sample the majority class. This is also a resampling technique which will alter the dimensions of the data. Here the model might underfit.

Synthetic oversampling : SMOTE acronym for synthetic minority oversampling technique is an advanced technique which uses a KNN algorithm to generate samples and also retain the dataset size. I have used SMOTE to balance the classes to prepare the data for modelling.

Here, I have used under sampling and oversampling techniques, to test which sample performs better, it turns out that under sampling is much efficient than over sampling.

j) Variable transformation

The final step in our exploration is the variable transformation. It refers to as transforming the variables or scaling the data in order to retain the shape. The scaled data will always have mean 0 and standard deviation 1.

Scaling is important to retain the scale of each variable. In this case, the total cholesterol levels and glucose levels were ranging from 100-600 whereas all other variables were 25-150. Now this imbalance in scaling reflects to add noise to our model and thus making the model to overfit the data. Scaling enables us to reflect each variable as it contributes and delivers perfect fit.

While scaling is possible for continuous data, we can apply log transformation for logistic regression and classification problems. It is equivalent of scaling in terms of classification levels. However this only helps for multi-level classification problems.

Now that we have concluded our exploratory analysis, we will discuss the analytical approach we will be using to achieve the outcome of our problem. Just to re-iterate, we have a classification problem wherein, we have to predict whether or not a patient will be diagnosed with coronary heart disease in the upcoming 10 years.

1.4. Analytical approach:

As re-iterated previously, we are dealing with a binary classification problem. Possible approaches to solving this problem are:

- Logistic Regression
- Naïve Bayes classifier
- CART model for classification
- Random forest for classification
- KNN for classification

Logistic Regression is one of the most used ML algorithms for binary classification. Thus it would be my first choice. After the model building and evaluation, I would further develop Naïve Bayes classifier, second of the linear classifiers.

Linear models explain yield both “direction of influence” & “global interpretability”, meaning that the models would not only provide us with an outcome, but also justify. Linear models are easy to interpret and thus explain the relationship between the X’s and the Y. Finally, linear models are less likely to be overfitted, as the linearity assists in capturing less noise compared to polynomials.

Non-linear models such as classification trees and random forests are also ideal choices. Both CART and RF help explain the non-linearity extremely well and also require little training. In addition the user defined parameter tuning extracts precise output. The only limitation is the criteria and model transparency are defined by the model with no insight on model functionality decisions.

Last but not the least, KNN, one of the easiest to implement ML algorithm also is a possible choice as our dataset is relatively small and with so many independent variables closely dependent, KNN would be an ideal choice.

As we have prepared the data nicely till this point, I’ll build and run all the above models.

Coming to model evaluation, we are dealing with a classification problem, thus primary metrics would be confusion matrix and it’s sub metrics. Rather than focusing on accuracy, I would apply a problem outcome based approach.

In our problem, **the most important metric will be correctly predicting the patients who will not be diagnosed with CHD** i.e. True Negatives.

Most of the models mentioned earlier will result in probabilistic outputs i.e. models will give us a probability for each observation, stating its chances of correctly or incorrectly predicting whether or not will the patient be diagnosed with CHD. One important factor to consider is setting the decision threshold value.

We would have to set an extremely high decision threshold as we are dealing with a healthcare domain problem and proceeding with usual decision threshold of 0.5 will result is disastrous outcomes.

Along with the confusion matrix – I will focus on ROC and AUC curves and plots, log loss/cross entropy and f1 score. In order to check if the models were actually overfitted, I would also implement cross validations techniques such as K-fold or leave one out.

Model Building:

Pre-modelling checklist

- Dataset is split into training and testing data in ratio 75-25%.
- The ratio of training to testing data is equal.
- The data is assumed to be normally distributed.
- Missing values have been imputed.
- Outliers have been removed or treated.
- The variables have been transformed and scaled.
- Insignificant variables have been removed.
- Classes with level imbalance, have been balanced in equal ratios.
- Variables have appropriate data types, numeric and factor.

All of the steps have been completed, thus making the data ready to be fed to our various models. I will first model our linear algorithms namely Logistic Regression and Naïve Bayes.

Model building & Model validation

A) Logistic Regression

Logistic Regression (LR) is one of the most widely used binary classification algorithms in the industry. Since the outcome of our problem is binary classification, LR is our first choice of variety of algorithms. LR is a probability based linear modelling algorithm, which provides us with a probability of accurately predicting the output of each of the observations. The probability ranges between 0 and 1, with 1 being absolute certainty of accurate prediction and 0 being no certainty of accurate prediction. Thus it is we who decide on a decision threshold value to determine how accurate we want our predictions to be, which is obviously dependent on the business context.

B) Naïve Bayes

Naïve Bayes (NB) is a probability based supervised classification algorithm and it relies on the fact that the predictor variables are independent. It is called Naïve because it assumes predictor variables to be independent, however in real life, some correlation does exist and it is impossible to have perfectly independent predictor variables. NB is actually the extension of “Bayes Theorem”, which calculates conditional probability i.e. calculating probability of an event, given another event in the past. It functions with building a joint probability distribution table which calculates the individual probabilities and then the conditions are applied, finally we can calculate the conditional probabilities by grouping them into categories.

C) CT – Classification Trees

CART, acronym for classification and regression trees is a supervised learning algorithm and is based on segmenting or branching the possible outcomes based on a algorithmically derived condition. We will use the modified CT as we are dealing with a classification problem. It follows an inverted tree structure where each node represents a predictor variable and each leaf node represents an outcome variable. The segmentation or decision criterion are the branches which connect the node and the leaf nodes. Unlike LR and NB, CART works well with non-linear data. Like all ML algorithms, this also is based on the percentage of variance explained by the independent variables without deteriorating the accuracy of the model. In CART, we can decide this threshold and prune the tree to a point where minimum number of predictor variables, explain maximum amount of data. The only disadvantage being, that non-linear algorithms are more likely to be overfitted, thus pruning the tree is extremely important to avoid overfitting.

D) Random Forest

Random forest (RF) is again a supervised learning algorithm which extends the decision tree logic by building a forest comprising of multiple trees. The idea of CART is to maximise the information gain at each split when the leaf node is branched under a node and ensemble models such as RF and GBM (gradient boosting) use the same concept to maximise the variance explained by setting high performance thresholds. RF works well with both classification and regression data. In either case, RF interprets the prediction factor with help of 2 main parameters: Entropy and information gain. *"Entropy refers to as the measure of impurity or uncertainty in the data, it decides how and based on which parameter the model splits the data"*. Whereas, *"Information gain is the most significant measure of the prediction, it indicates how much information does the feature, gives us about the prediction or the outcome."*

Model validation:

Now that we have completed the model building phase, we will cross validate our models to check if either of the models have underfitted or overfitted the data. The cross validation techniques are useful to identify the fit of the model. I have included cross validation techniques within the modelling and have applied k-fold cross validation with 10 folds in each of the models. I will now jump straight onto modelling our linear and ensemble models.

Linear Models:

A) Logistic Regression

I built 4 LR models, 2 each with oversampling and under sampling data. It was important to know how the observations were synthetically oversampled and which were under sampled. Initially, the models were built as default i.e. response variable against all predictor variables. By doing this, we will identify our significant variables. The model summaries are as follows:

```
Call:
glm(formula = Class ~ ., family = "binomial", data = up_train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.11870  -1.02859  -0.02958   1.01801   2.04153

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -5.0636088  0.4917906 -10.296  < 2e-16 ***
male1             0.4547927  0.0679856   6.690 2.24e-11 ***
age               0.0628748  0.0042121  14.927  < 2e-16 ***
Education2       -0.0501399  0.0752645  -0.666  0.50529
Education3       -0.1786961  0.0912209  -1.959  0.05012 .
Education4       -0.2262268  0.1045413  -2.164  0.03046 *
currentSmoker1   -0.0960492  0.1018705  -0.943  0.34575
cigsPerDay        0.0212015  0.0043492   4.875 1.09e-06 ***
BPMeds1           0.1298202  0.1718721   0.755  0.45005
prevalentStroke1  0.3536011  0.3843249   0.920  0.35754
prevalentHyp1     0.4733629  0.0865849   5.467 4.58e-08 ***
diabetes1         0.7836049  0.1767395   4.434 9.26e-06 ***
totChol           0.0019910  0.0007971   2.498  0.01250 *
sysBP             0.0085530  0.0027160   3.149  0.00164 **
diaBP            -0.0069793  0.0043900  -1.590  0.11188
BMI               0.0068359  0.0093844   0.728  0.46635
heartRate         0.0003069  0.0028750   0.107  0.91498
glucose           0.0011740  0.0028058   0.418  0.67565
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6981.4  on 5035  degrees of freedom
Residual deviance: 6230.2  on 5018  degrees of freedom
AIC: 6266.2
```

The important model interpretation parameters are AIC (Akaike Information Criteria), the null and the residual deviance, confusion matrix and ROC (Receiver Operating Characteristics) curve. Let us understand each one of them in detail.

AIC, is the analogous metric of adjusted R^2 in logistic regression. It is the measure of fit which penalizes model for the number of model coefficients. Therefore, we always prefer model with minimum AIC value. In our case, the AIC value is 6266.2, which is by no means close to minimum. Thus, we infer that the oversampling data has not been fitted properly.

However we observe from the significant levels that variables like education, BP medication, current smoker, stroke, BMI, heartrate, glucose and both BP levels are insignificant across all confidence intervals, meaning that the amount of explained variance from these variables is negligible compared to the AIC value of the LR model. In our next model, we will remove the insignificant variables and see if the metrics improve.

```
Call:
glm(formula = Class ~ . - Education - currentSmoker - diaBP -
    BMI - heartRate - glucose - BPMeds - prevalentStroke, family = "binomial",
    data = up_train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.14130  -1.03366  -0.02107   1.02532   2.04343

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -5.2933772  0.3364290 -15.734  < 2e-16 ***
male1          0.4450541  0.0644228   6.908 4.90e-12 ***
age            0.0657582  0.0039776  16.532  < 2e-16 ***
cigsPerDay     0.0180923  0.0027738   6.523 6.91e-11 ***
prevalentHyp1  0.4695280  0.0835214   5.622 1.89e-08 ***
diabetes1      0.8182396  0.1759759   4.650 3.32e-06 ***
totChol        0.0018685  0.0007869   2.374   0.0176 *
sysBP          0.0066691  0.0021254   3.138   0.0017 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6981.4  on 5035  degrees of freedom
Residual deviance: 6243.7  on 5028  degrees of freedom
AIC: 6259.7

Number of Fisher Scoring iterations: 4
```

There seems to be no evident improvement in the model. The value of AIC has reduced, but again is nowhere close to being minimum.

Here, all the variables are significant, with most of them significant at an extremely high confidence interval of 99.99%. Ideally the model performance should have been much better. So it means that the oversampling algorithm introduced a lot of noise and possibly redundant observations, which is yielding such a poor AIC metric.

In our subsequent models, we have used the under sampling data as here, no additional observations were introduced or synthetically imputed, here only the oversampled class has been reduced to match with the under sampled class.

```
Call:
glm(formula = Class ~ . - Education - currentSmoker - diaBP -
    BMI - heartRate - glucose - BPMeds - prevalentStroke, family = "binomial",
    data = down_train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.20099  -1.02616  -0.08881   1.02407   2.02739

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.190648   0.807104  -6.431 1.27e-10 ***
male1         0.248555   0.153967   1.614 0.106454
age           0.057468   0.009222   6.232 4.62e-10 ***
cigsPerDay    0.025628   0.006678   3.838 0.000124 ***
prevalentHyp1 0.576353   0.198705   2.901 0.003725 **
diabetes1     1.182061   0.448190   2.637 0.008354 **
totChol       0.001664   0.001875   0.888 0.374732
sysBP         0.009303   0.005088   1.829 0.067474 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1250.4  on 901  degrees of freedom
Residual deviance: 1112.4  on 894  degrees of freedom
AIC: 1128.4

Number of Fisher Scoring iterations: 4
```

In the under sampled model, we see a drastic improvement. The AIC value is 1128.4 way better than that of oversampled model. Here again we have only added significant parameters. It is not possible to find the order of significant variables in LR, thus it is not possible to derive further metrics.

Null and Residual Deviance - Null Deviance indicates the response predicted by a model with nothing but an intercept. Lower the value, better the model. Residual deviance indicates the response predicted by a model on adding independent variables. Lower the value, better the model.

The difference between the null and residual deviances is slightly on the lower side, meaning that the predicted intercept in the model is of negating the effect of the independent variables. The value drops after we add the independent variables.

B) Naïve Bayes

Here I built only a single model with under sampled data. Let us look at the model and the measures of evaluation.

```
Naive Bayes

902 samples
 15 predictor
  2 classes: '0', '1'

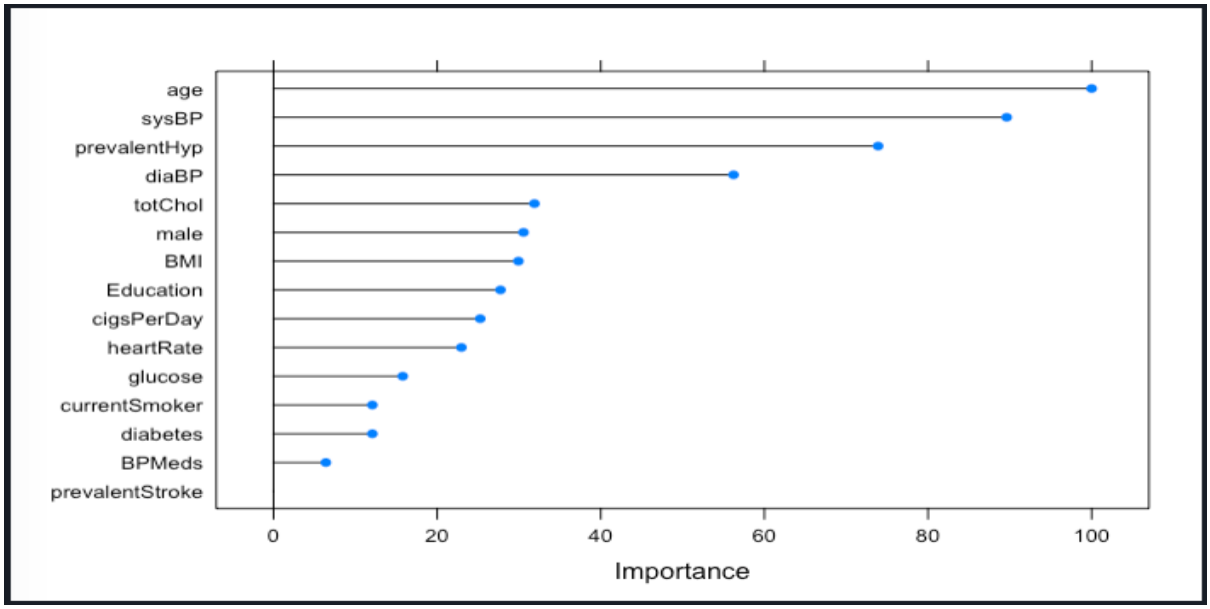
No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 812, 811, 812, 812, 811, 812, ...
Resampling results across tuning parameters:

  usekernel  Accuracy   Kappa
  FALSE      0.6474603  0.2947887
   TRUE      0.6208791  0.2416141

Tuning parameter 'fL' was held constant at a value of 0
Tuning parameter 'adjust' was held constant at a value of 1
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were fL = 0, usekernel = FALSE and adjust = 1.
```

Here I have applied a cross validation of 10 folds to ensure that the data is not overfitted. We observe that the model was fitted with an overall accuracy of ~65%.

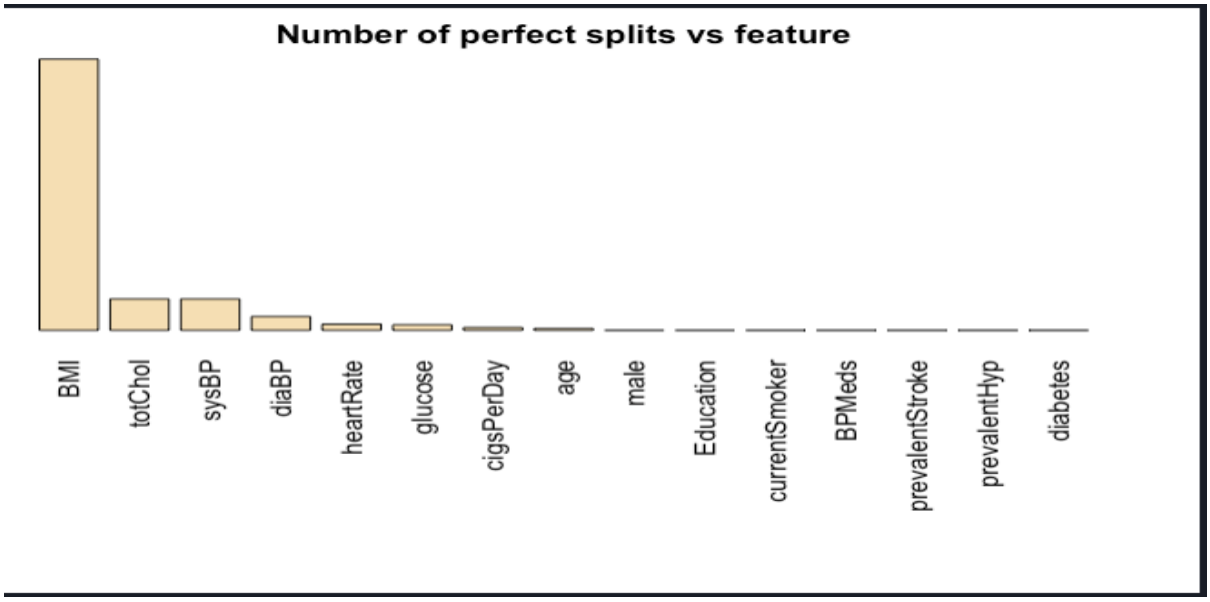
Note that NB model does not provide any intercepts or confidence intervals for variables like LR, however we can see the significant variables in NB model with the variable importance plot.



We infer that age is the most significant variable in our model. It is pretty obvious when dealing with heart diseases, the higher the age, higher the risk. The model also considers systolic BP, hypertension, diastolic BP, cholesterol and gender as other significant parameters.

C) CART

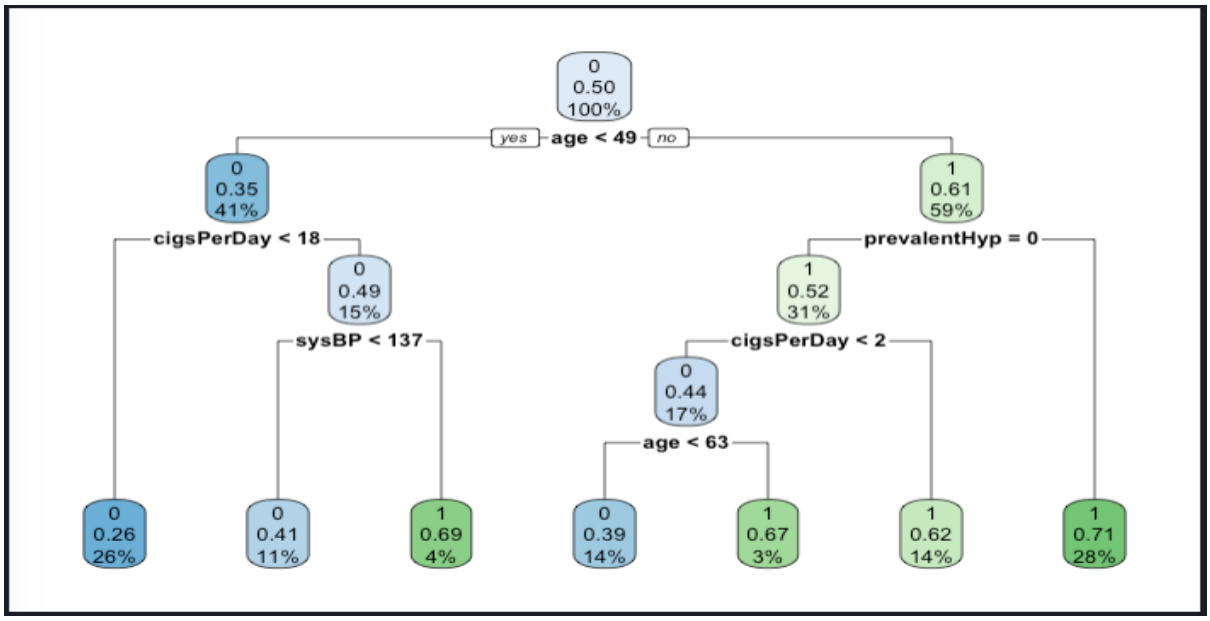
Moving to ensemble modelling techniques, I have modelled our data with CART as well as Random Forest. These are the most widely used classification algorithms which handle non-linear data. Now that we have modelled our data with linear models, I will also model the data with ensemble methods. Let us first model our data with CART.



We observe that, this plot is similar to a variable importance plot as in Naïve Bayes and explains exactly the same concept. This plot shows the best possible split root nodes with which information gain within the CART decision splits is maximum.

The plot has the highest significance for BMI. BMI is directly related to body weight, fat and glucose levels. Obesity is an already serious condition which exhibits similar artery contractions as experienced in a heart disease. As we move further, we observe the similar parameters such as cholesterol levels, blood pressure, heartrate and glucose which primarily focus on obesity. It infers that body fat, weight and obesity play a significant role in prediction of heart diseases.

Now that we have established variable importance, I will now plot the RPART plot.



Analysing the tree plot, we infer that age is considered as the root node, the primary decision split node because, dealing with heart diseases will have a direct relation with patient's age. We already are aware that risk of heart disease will only increase with increase in age.

Patients which are older than 49 years old are primary suspects of CHD. These patients will further be tested with prevalent hypertension. As hypertension is directly related to stress levels, with increasing age, this will become a crucial identifier.

- If patients have hypertension, the probability is significantly increased to be 71% of being diagnosed with CHD.
- If patients do not have hypertension, the next check is for the number of cigarettes patients smoke in a day.
- If the count of cigarettes is more than 2, the probability is 62%.
- If the count of cigarettes is less than 2, the patient's age is reconsidered and if it is more than 63, probability is 67%.

Patients which are younger than 49 years are less susceptible to being diagnosed with CHD. Here the number of cigarettes per day count is increased from 2 to 18, primarily due to age group.

- If the count of cigarettes per day is less than 18, the chances are low.
- However if the number of cigarettes per day is more than 18, the next check is of systolic BP. Systolic BP measures the pumping heart rates and is ideally in range of 120.
- Setting the threshold at 137, if BP is less than the threshold, the chances are still low at 41%.
- However if the BP increases more than 137, it induces the risk of hypertension, thereby increasing the probability to 69%.

D) Random Forest

Last of our models, Random Forest is an extension of CART. Here I have modelled the RF with cross validation with k-folds at 10.

```
Random Forest

902 samples
 15 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 812, 812, 812, 811, 812, 812, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
   2    0.6596947  0.3194865
   9    0.6431136  0.2861762
  17    0.6242369  0.2484093

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 2.
```

Here we observe that, the accuracy is decreasing as the value of m-try is increased. This parameter tells us the number of available features which can be split further. The lower the number means the criteria are strict and most of the significant decisions are split on categorical variables as opposed to continuous ones.

Model evaluation:

We discussed the relative evaluation metrics in brief in the above data modelling and validation section, here we will explore in depth the confusion matrix, AUC and ROC plots of each of our models and interpret them.

A) Logistic Regression

LR is a probability based algorithm and as already mentioned earlier, setting the decision threshold value is a key parameter in model evaluation. I have considered our under sampled data model and ran 5 decision threshold splits starting from 0.5 to 0.9.

Threshold	True Positive	False Positive	True Negative	False Negative
0.5	141	392	52	686
0.6	110	229	83	849
0.7	65	114	128	964
0.8	21	23	172	1055
0.9	4	1	189	1077

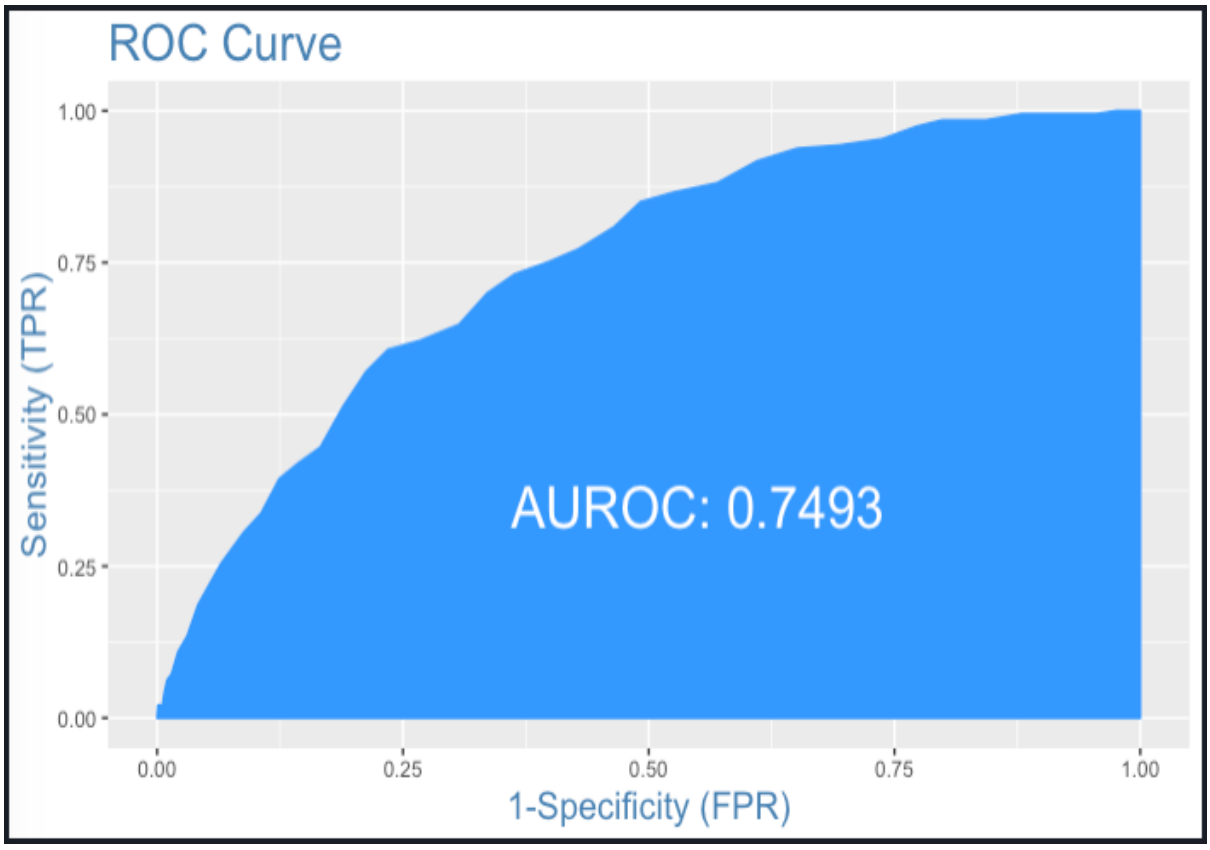
As we are dealing with a likelihood of prediction of heart disease problem, the primary objective is to predict negative for patients who did not have heart diseases. The alternative objective is to not predict positive for patients who did not have heart diseases. So basically negatives have a higher preference over positives.

We infer from the table, as the decision threshold is increased, the value of **false positives** is **dropping** and value of **false negatives** is **rising**. This is an extremely encouraging sign for our model. It means that the model at 0.5 threshold was incorrectly predicting positive for patients who did not show signs of CHD and as we increase the threshold, that value is consistently dropping.

Similarly the false negatives are increasing. It means that the patients who did not exhibit signs of CHD, are diagnosed as negatives i.e. will not be predicted as possible CHD patients. So the high threshold value is recommended for our healthcare domain.

Model	Accuracy	TPR	TNR	FNR	FPR	Precision
0.5	65.07%	73.06%	63.64%	26.94%	36.36%	26.45%
0.6	75.45%	56.99%	78.76%	43.01%	21.24%	32.45%
0.7	80.96%	33.68%	89.42%	66.32%	10.58%	36.31%
0.8	84.66%	10.88%	97.87%	89.12%	2.13%	47.73%
0.9	85.05%	2.07%	99.91%	97.93%	0.09%	80.00%

Now finally the ROC and AUC curve plots:



The area under the curve for the selected threshold is 74.9%, which is a good model performance.

B) Naïve Bayes

The confusion matrix for NB is as follows:

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 708  66
         1 370 127

               Accuracy : 0.657
                 95% CI : (0.6301, 0.6831)
    No Information Rate : 0.8482
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1912

 Mcnemar's Test P-Value : <2e-16

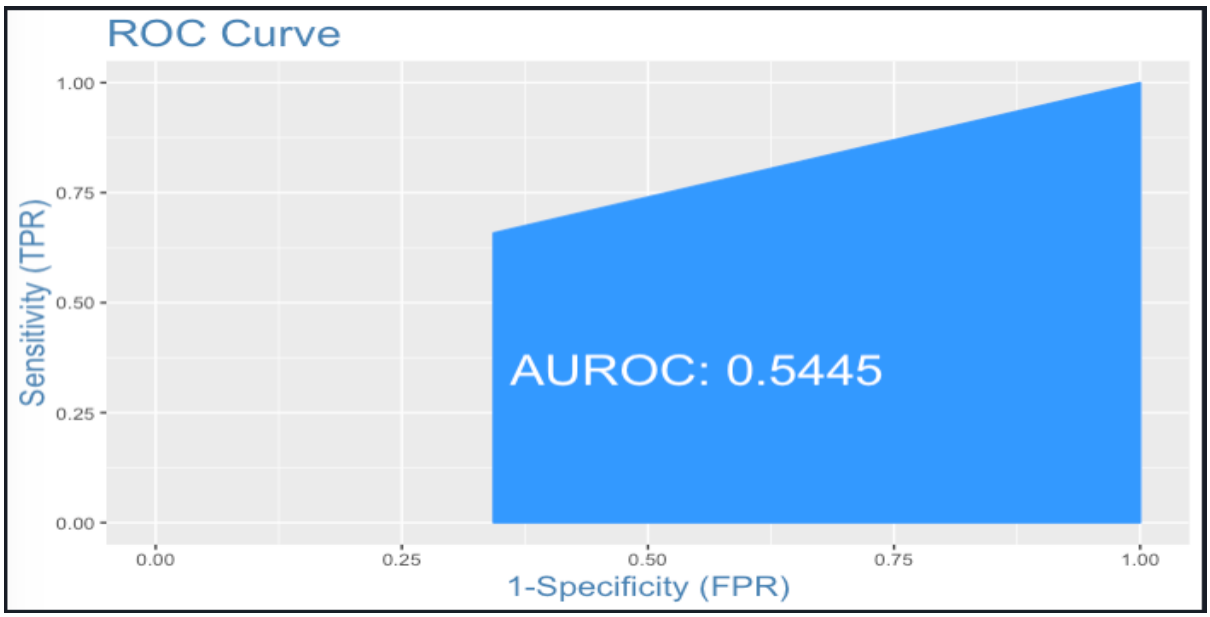
            Sensitivity : 0.6568
            Specificity : 0.6580
         Pos Pred Value : 0.9147
         Neg Pred Value : 0.2555
             Prevalence : 0.8482
         Detection Rate : 0.5570
   Detection Prevalence : 0.6090
      Balanced Accuracy : 0.6574

       'Positive' Class : 0
```

The overall accuracy of the model is 65.7%. As previously mentioned our focus is on negatives. Therefore, negative prediction value is at 25%, which is good model again. The detection rate also is at 55%, which is average but still fine.

The positive class however here is important. We're focused on negatives, thus the positive class or the prediction class is 0.

Let's look at the AUC and ROC plot.



The area under the curve for the selected threshold is 54.5%, which is average model performance.

C) CART

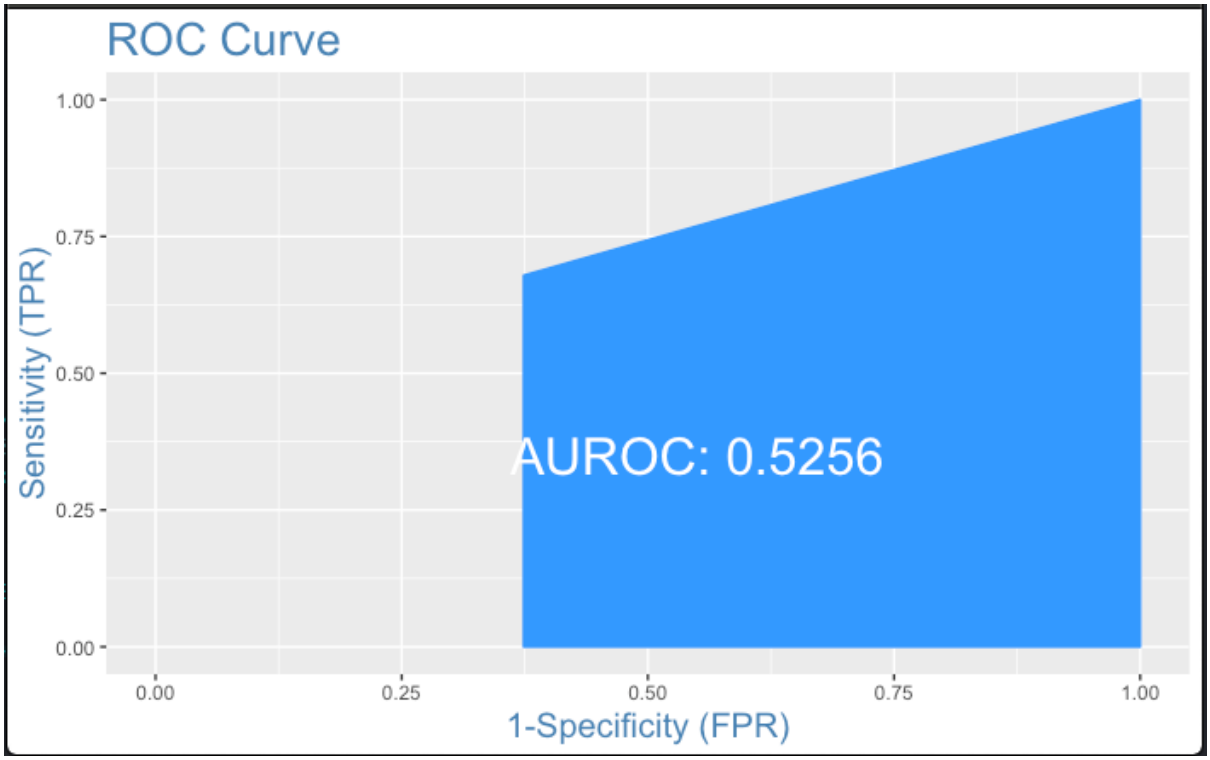
The confusion matrix for CART is as follows:

```
   pred_CART
       0   1
0 675 403
1  62 131
```

Accuracy	TPR	TNR	FNR	FPR	Precision
63.41%	67.88%	62.62%	32.12%	37.38%	24.53%

The overall accuracy of the model is 63.41%. The negative rates are higher than Naïve Bayes, but still in a descent performance range. It is safe to say that the linear models performed better than the non-linear models.

The ROC plot is as follows:



The area under the curve for the selected threshold is 52.5%, which is average model performance.

D) Random Forest

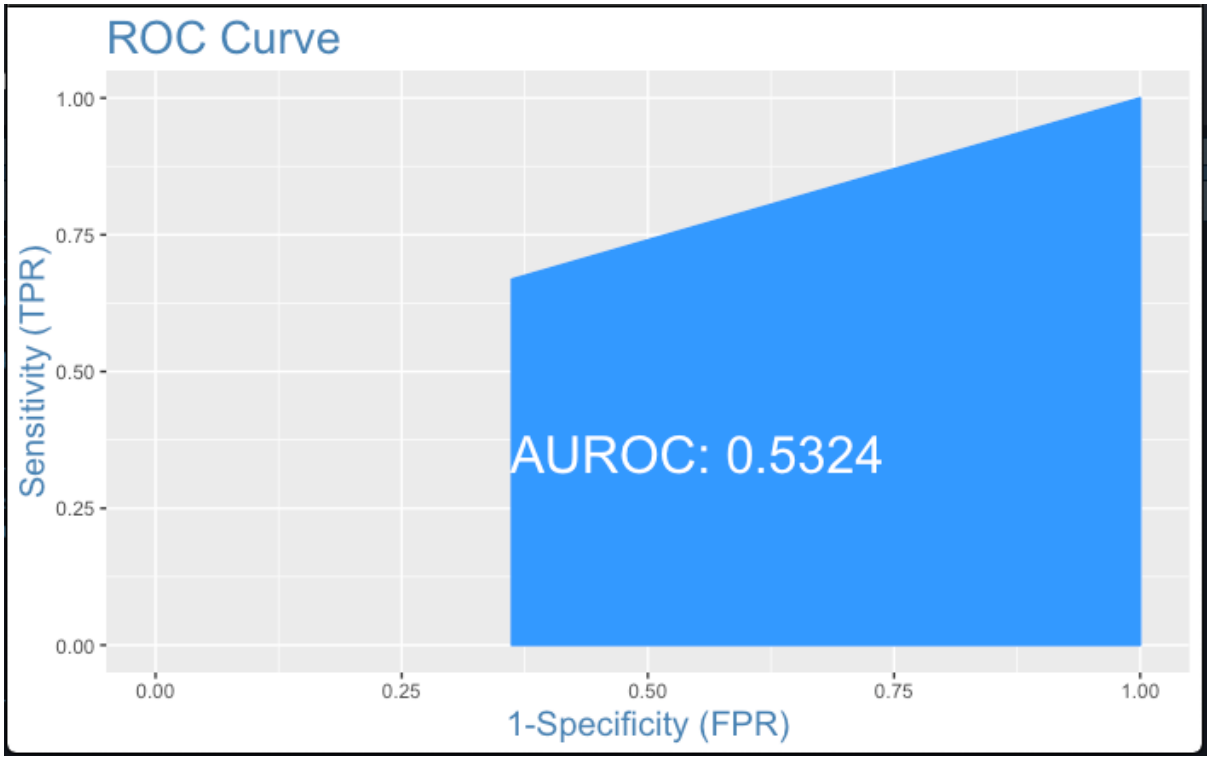
The confusion matrix for RF is as follows:

```
      pred_RF
         0   1
    0  688 390
    1   64 129
```

Accuracy	TPR	TNR	FNR	FPR	Precision
64.28%	66.84%	63.82%	33.16%	36.18%	24.86%

The overall accuracy of the model is 64.28%. The negative rates are higher than Naïve Bayes, but still in a descent performance range. It is safe to say that the linear models performed better than the non-linear models.

The ROC plot is as follows:



The area under the curve for the selected threshold is 52.5%, which is average model performance.

Model comparison:

Model	Accuracy
Logistic Regression	85.05%
Naïve Bayes	65.70%
CART	63.41%
Random Forest	64.28%

Model Interpretation:

- The Logistic Regression model is the best performer among our models.
- The LR model has significant advantages in case of a binary classification problem as ours, as it considers multiple parameters such as the intercepts and the deviances from the significant as well as the insignificant parameters.
- The null and residual deviances also incorporate the variance of only the intercept as well as the independent variables.
- As our problem was from healthcare domain, setting the specified decision threshold comes in as a very handy tuning, as it increases the accuracy as well as false negative and true negative rates.
- Here, I have focused on the negatives as we would want as low a number of patients being diagnosed with CHD. However, multiple perspectives could be applied and addressed for the same problem.

Significant predictor variables:

- **Age** is the one of the most important factors in determining the patient diagnosis of CHD. Our analysis states, that males with age > 51 and post-menopausal females (i.e. age > 49) are at maximum risk of being diagnosed with CHD, with additional contributors.

- **BMI** is in my opinion the most critical factor to be considered. BMI is a composition of height and weight, and the effect of many contributing factors exemplifies with increasing weight. Thus our analysis suggests, a BMI with 38 or more in both males & females is a high risk.

- **Gender** only signifies that pre-menopausal females are less likely to develop artery contractions. Because of the recurring menstrual cycles, the blood is well circulated throughout, thus once a menopause is approached, the risk slightly intensifies.

- **Current smoker** is another obvious factor contributing to diagnosis of CHD. The effect is extremely gradual and even though our analysis suggests that number of cigarettes smoked in a day should not exceed 20, in practicality, the number needs to be much lower.

- **BP levels** (diastolic) is yet another decider. The BP levels justify the rate of blood flow and a higher BP suggests, artery contractions. Our focus is primarily on diastolic, as it is the resting rate and if it exceeds 95 and patient exhibits hypertension, then it is a high risk.

- One of the most important factor is however was missing in the data i.e. Serum cholesterol which includes HDL and LDL. Acronyms for high and low density lipoproteins, they are extremely crucial in CHD diagnosis. HDL is healthy cholesterol while LDL is a risk. Inclusion of these parameters would have made the analysis completely sound.

Business Impact & Recommendations:

The direct impact of this project is primarily on the following industries:

- **Insurance –** It is a need if not a mandatory requirement to have a medical insurance for health situations. Problems related to heart are among most expensive and thus a first point of contact is insurance.

- **Pharmaceuticals –** Again, treatment is most of the times of limited time-frame. In order to avoid the symptoms over a period of time, a prescribed medication is essential, where the pharma industry is involved.

- **Telecommunications –** It is relatively new, multiple cross technology application developers have started rolling out health related apps, which track the general symptoms from your smartphone. A boon to routine life goers.

- **Technology –** The implications may not be restricted to medication, at times specially designed preventive equipment is prescribed by doctors such as *pacemakers* and *implantable cardioverter defibrillators (ICDs).*

- **Social Media –** In a digitally connected world, social media must have a contribution. As soon as a research paper is published, a worldwide circulation is possible via social media. Gyms, hospitals start taking measures to ensure safety standards etc.

Recommendations:

The outcome of every project is incomplete without the recommendations. Here, we would like to spread awareness in following areas to best avoid being diagnosed with CHD. The gender and the age are factors beyond our control, so let's review the following:

- Try to be physically active

- Try workouts or stretching

- Quit smoking

- Control body fat & weight

- Develop healthy eating habits

- Meditate to release anxiety or stress

- Take routine health check-ups

- Avoid foods with LDL

Project References:

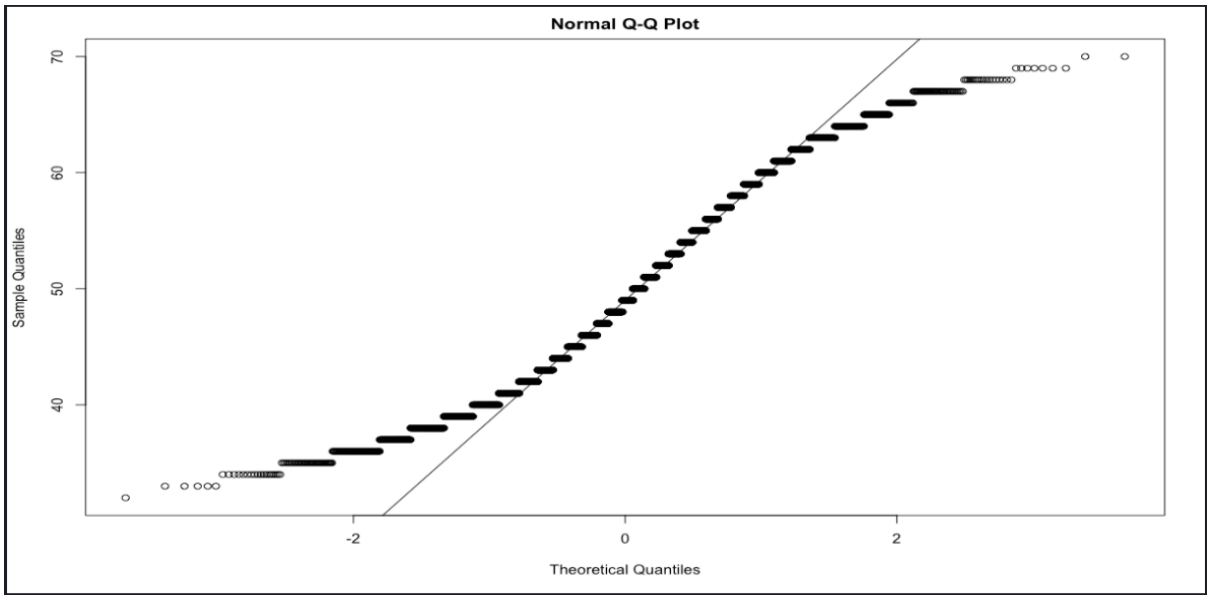
- https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/
- https://www.kdnuggets.com/2017/09/missing-data-imputation-using-r.html
- https://stackoverflow.com/questions/39916565/plot-a-clear-graph-to-show-the-skewness-and-kurtosis
- https://www.evolven.com/blog/data-and-problem-definition.html
- https://towardsdatascience.com/heart-disease-prediction-73468d630cfc
- https://developers.google.com/machine-learning/crash-course/classification/check-your-understanding-accuracy-precision-recall
- https://analyticsindiamag.com/7-types-classification-algorithms/
- https://www.r-bloggers.com/to-eat-or-not-to-eat-thats-the-question-measuring-the-association-between-categorical-variables/
- http://www.sthda.com/english/wiki/correlation-matrix-a-quick-start-guide-to-analyze-format-and-visualize-a-correlation-matrix-using-r-software
- http://www.sthda.com/english/wiki/scatter-plot-matrices-r-base-graphs
- https://towardsdatascience.com/simple-fast-exploratory-data-analysis-in-r-with-dataexplorer-package-e055348d9619
- https://machinelearningmastery.com/data-visualization-in-r/
- https://towardsdatascience.com/a-guide-to-data-visualisation-in-r-for-beginners-ef6d41a34174
- https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/
- https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/
- https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem/
- https://datasciencebeginners.com/2018/11/18/10-how-to-detect-outliers/
- https://rcompanion.org/handbook/I_12.html
- https://www.datanovia.com/en/blog/ggplot-colors-best-tricks-you-will-love/
- https://towardsdatascience.com/practical-guide-to-outlier-detection-methods-6b9f947a161e
- https://www.kdnuggets.com/2017/01/3-methods-deal-outliers.html
- https://machinelearningmastery.com/how-to-identify-outliers-in-your-data/
- https://medium.com/@george.seif94/15-python-tips-and-tricks-so-you-dont-have-to-look-them-up-on-stack-overflow-90cec02705ae

Please refer the appendices next page onwards.

Appendix: Reference plots

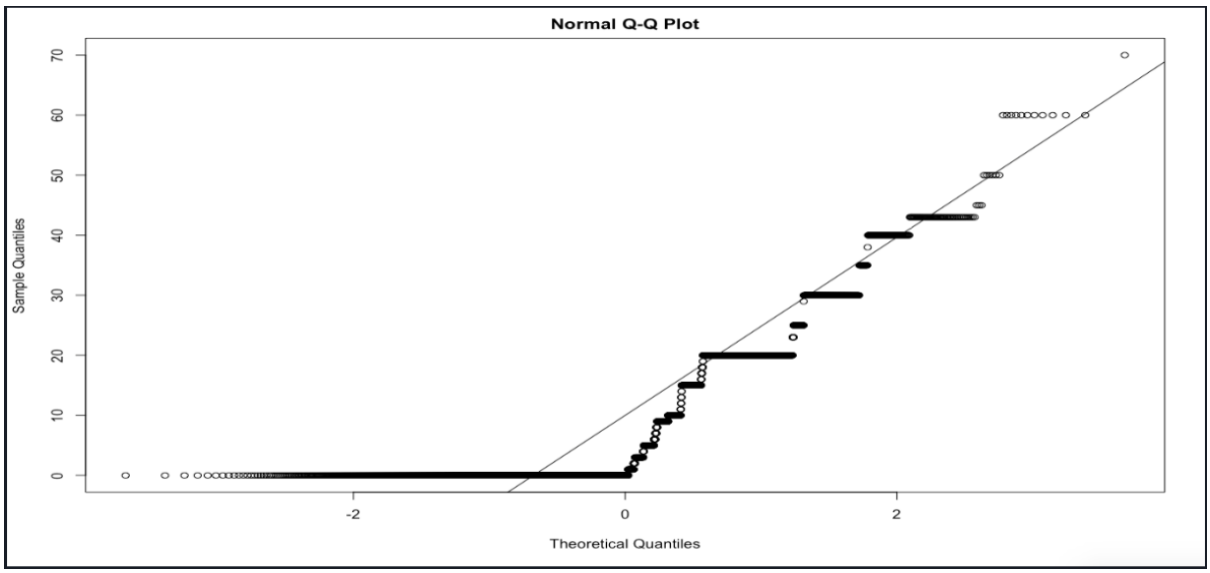
Univariate reference plots refer to plots which provide an insight on the distribution of data. In simple words, check the data for **skewness** and **kurtosis**. These plots also enable us to detect the presence of outliers. I will be covering two types of quantile plots, the normal QQ plots and line QQ plots. The normal QQ plot helps us understand whether our data is normally distributed or skewed.

Age



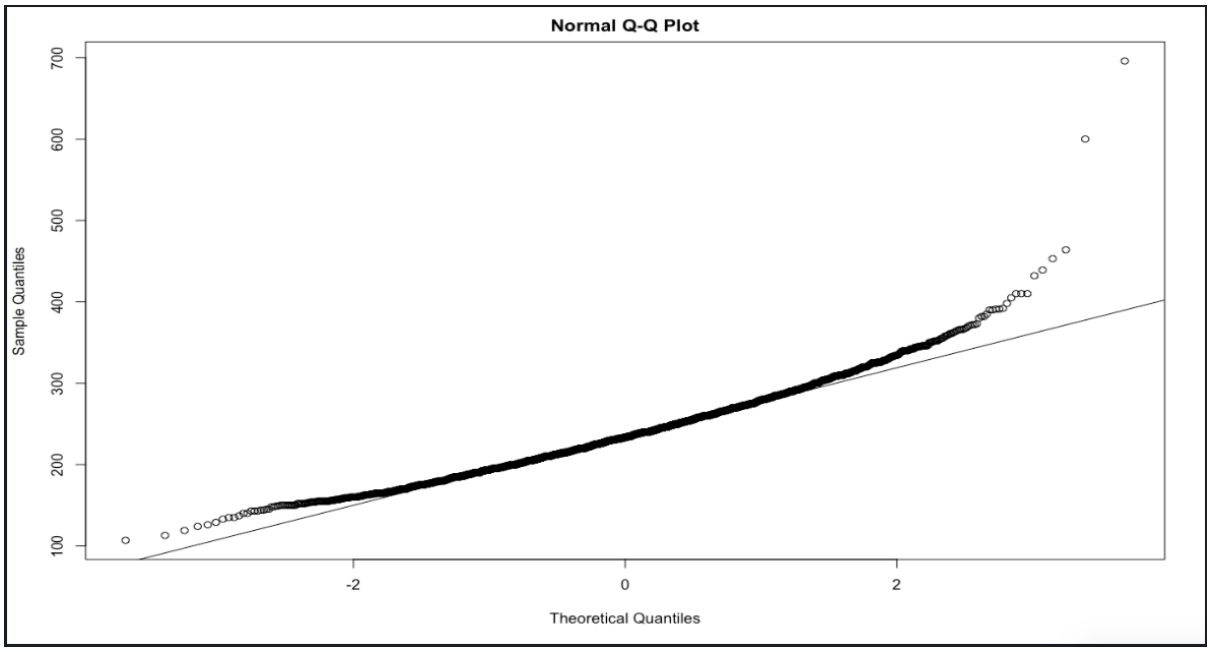
The QQ plot shows that the observations follow along the path in a relatively straight line across the centre of the plot, stating that the data points in age variable are fairly normally distributed.

Cigarettes per day



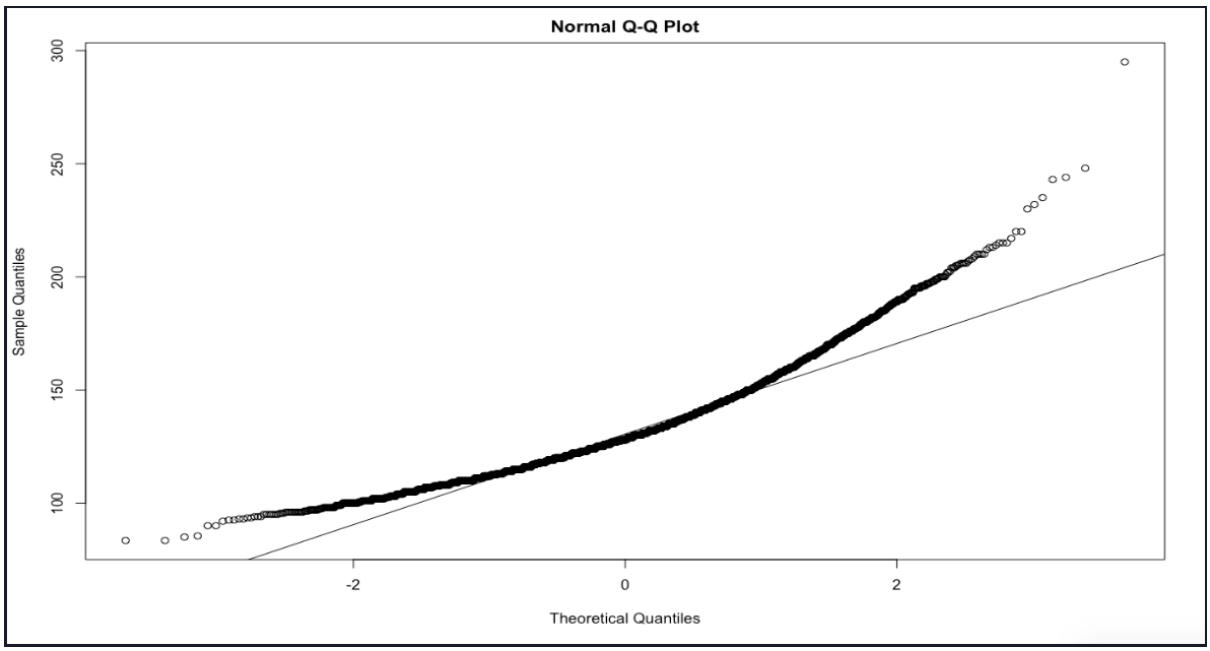
The above QQ plot infers that the data points in given variable have breaks and the data exhibits an extremely long tail at 0, meaning the data is highly skewed towards the left or negative skewed.

Total cholesterol



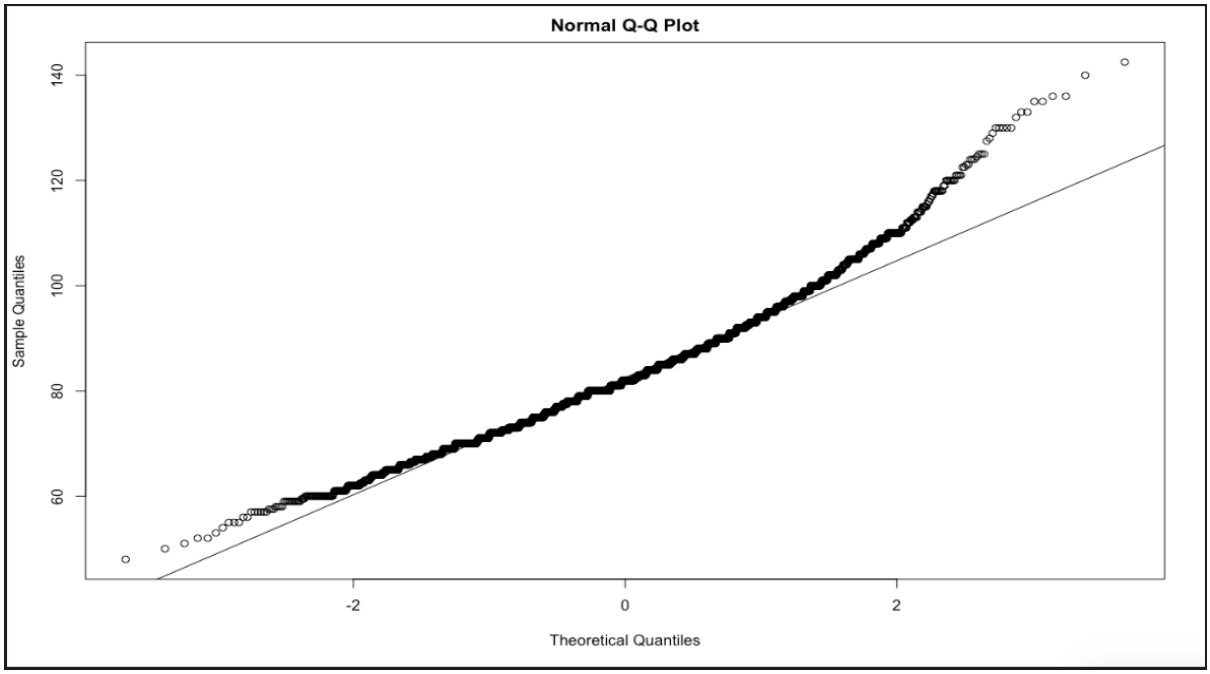
The above QQ plot shows that the observations are closely aligned with the QQ-line plot, which means that the observations do not exhibit extreme fluctuations until they start to spread out. In this case observations greater than 400, can be capped as outliers.

Systolic blood pressure



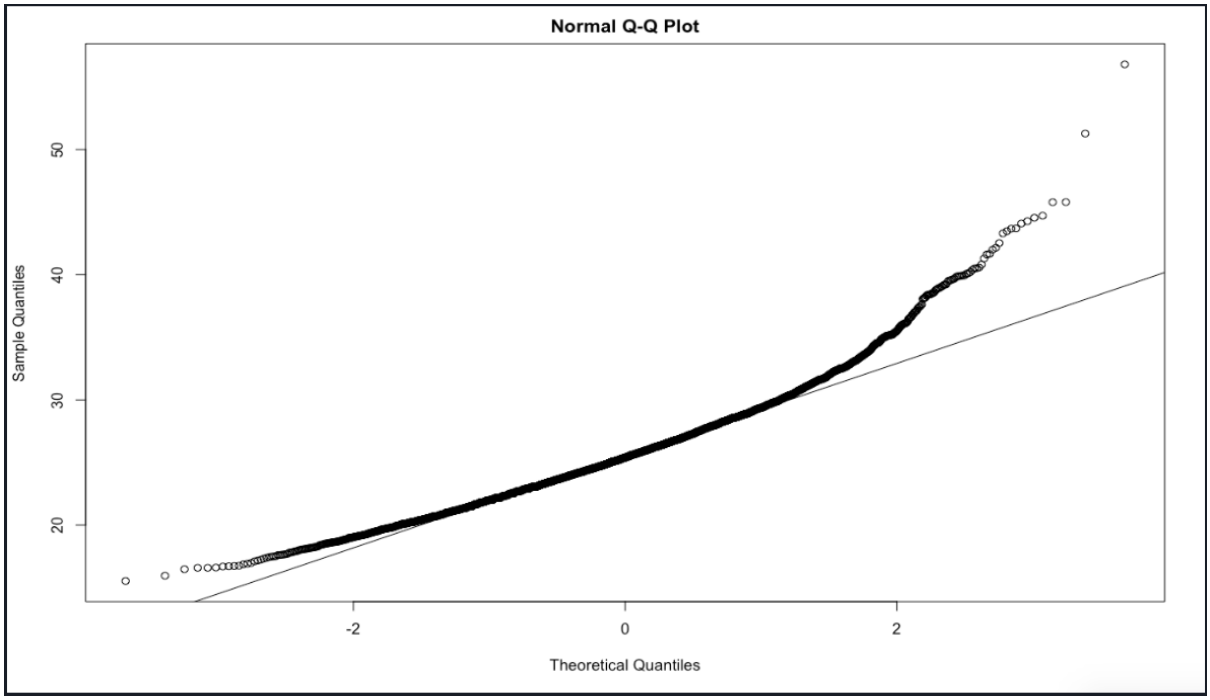
The slight curve in the plot, suggests that the data is slightly skewed towards right, positively skewed. The points start spreading away from QQ-line suggesting the kurtosis. As systolic BP increases, the effect of high BP also proportionately increases.

Diastolic blood pressure



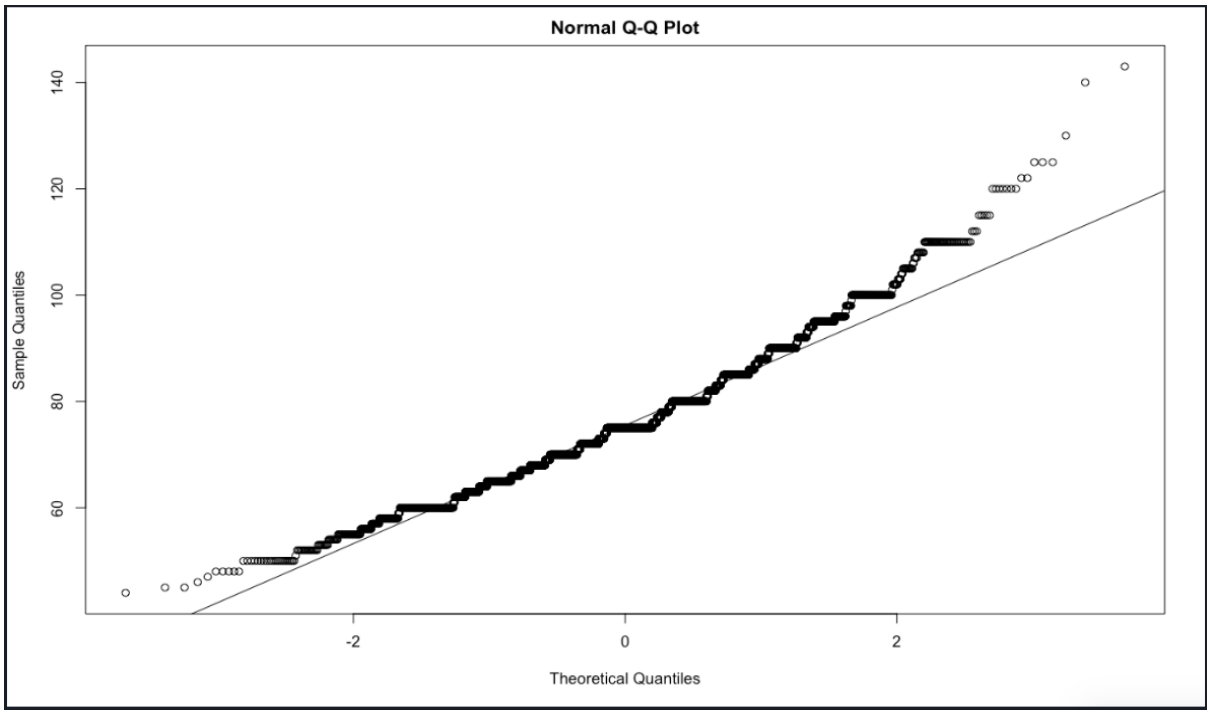
The above QQ-plot exhibits the most natural form of normal distribution. The data points are very closely aligned with the QQ-line and only the few observations greater than 130 exhibit outlier properties. All other observations are normally distributed and exhibit no skew.

Body mass index



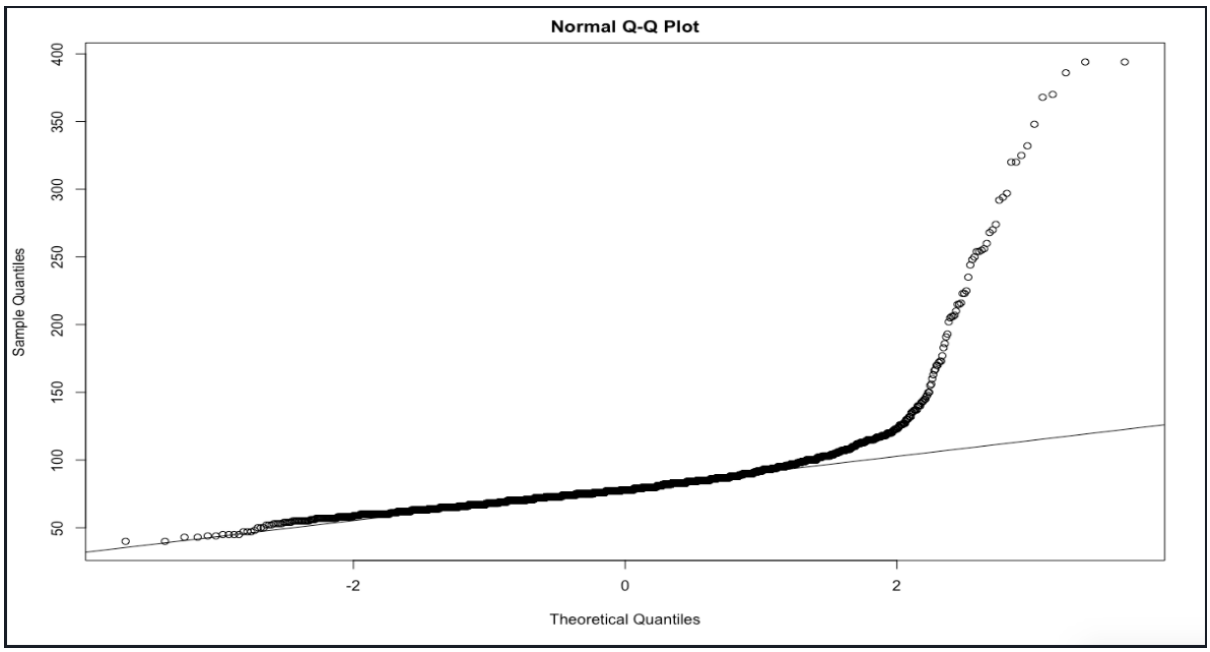
The BMI QQ-plot shows that the tail of the data exhibit a large spread indicating the presence of skewness and kurtosis. Also in addition, the data points greater than 40 are away from the QQ-line indicating that presence of outliers has a strong impact.

Heartrate



The heartrate QQ-plot is discrete. The observations are repetitive as the ideal range of pulse is only a limited buffer between 60-75. Overall the data does not exhibit sings of skewness and high impact outlier values except for the ones greater than 120. Anyways 120 pulse is extremely unlikely and at a very high risk, so the theory of it being an outlier is supported.

Glucose



The above QQ-plot suggests a steep increase in the glucose levels after 125. However we have assumed that values till 140 are normal, but the remainder values with values greater than 150 are again rare and unlikely, thus supports our theory as they being as outliers.

Project analysis conducted and report prepared by Chinmay Govilkar.

Thank you for reading!!