Coronary

Heart Risk Study

Managerial Report by Chetan Suvarna

1. Introduction

1.1 Problem Statement

The dataset provides the risk factors associated with heart disease for 4240 patients and whether they have a risk of coronary heart disease in the next 10 years.

We need to predict which patients will have a coronary heart disease in the next 10 years and the factors influencing them. The second part of the project deals with providing recommendations to prevent/reduce the chances of having a heart disease.

1.2 Need of the project

The study is originally sourced from Framingham Heart Risk which was started in 1948 after the premature death of American President Franklin D Roosevelt in 1945 due to heart disease and stroke. Due to poor health care facilities at the time the doctors were not able to prognosticate as to what led to the sudden mass of deaths caused by cardiovascular failure. A 1/3rd increase in the number of deaths encountered by cardiovascular diseases from 1990-2010 makes this study even more pertinent.

This project aims to bring down the risk of having a heart disease in the coming time. This will help to spread vital information to the growing generation of youths and also the elders to lead a lifestyle which would help them in preventing from having a heart disease. It will also advice the people on what factors will lead to a heart disease and to monitor them by undergoing relevant tests on a regular basis.

1.3 Understanding business/social opportunity

This project will help the hospitals to target the right set of people. It will also help other sectors which are related to health care in the ways mentioned below:

1. **Medical insurance** – It will help the insurance companies to target the right set of audience and promote medical insurance benefits over long term
2. **Medical equipment** – Equipment's like BP check-up machine can be purchased more often by people suffering from heart disease
3. **Pharmaceutical/drugs** – Drugs being one of the main sources of revenue in the medical sector can be targeted distinctively with respect to the impact of the disease.
4. **Wearables -** Smart watches like Fitbit can be used to track the distance you walk, run, swim or cycle, as well as the number of calories you burn and take in. Some also monitor your heart rate and sleep quality. This can help you to track your fitness and also to be on top of your health
5. **Lifestyle** – Lifestyle changes like working out in the gym and adding some nutritional apps to the smartphone would help improve the lifestyle

2 Exploratory Data Analysis – Step by step approach

A Typical Data exploration activity consists of the following steps:

1. Environment Set up and Data Import
2. Variable Identification

The following steps were followed in the dataset provided.

2.1 Environment Set up and Data Import

2.1.1 Set up working Directory

Setting a working directory on starting of the R session makes importing and exporting data files and code files easier. Basically, working directory is the location/ folder on the PC where you have the data, codes etc. related to the project.

2.1.2 Import and Read the Dataset

The given dataset is in .csv format. Hence, the command 'read.csv' is used for importing the file.

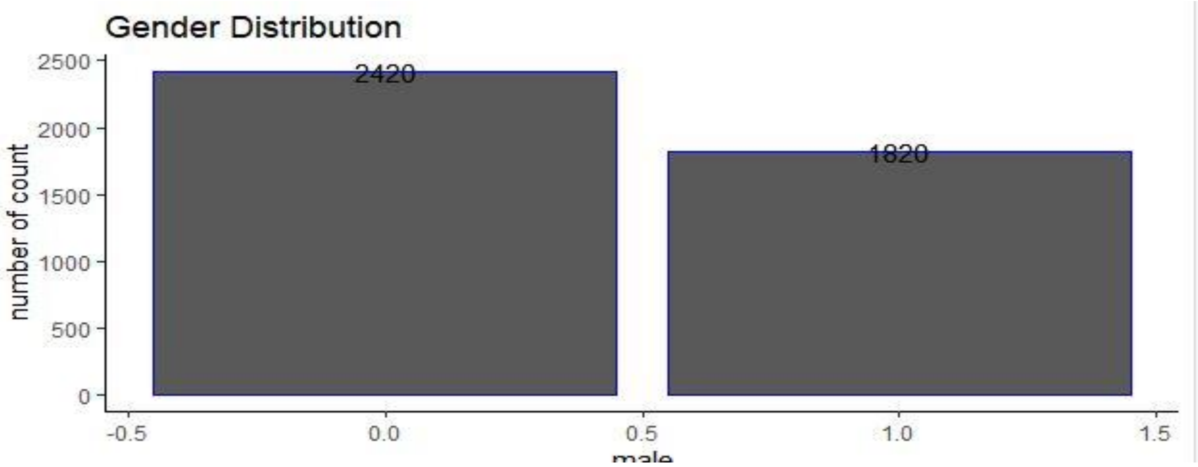
2.2 Variable Identification

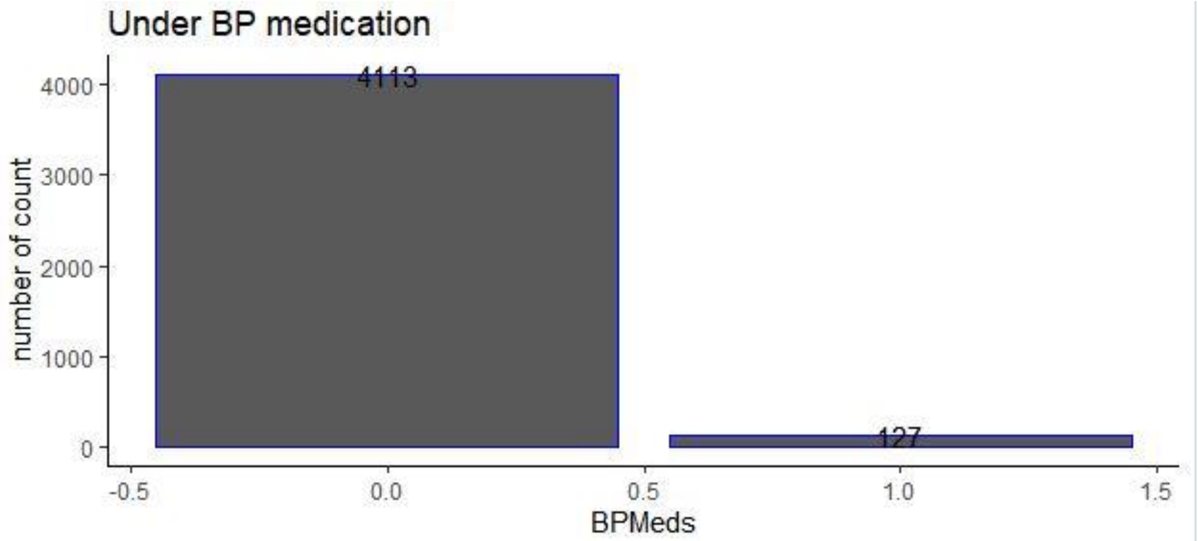
The length and breadth of the data was examined and the names of the data were pulled from the dataset. The data consisted of **4240 observations with 16 variables**. TenYearCHD is used as a categorical dependent variable to determine if the person had heart risk or not. The string type of the data was also verified by using the str() function.

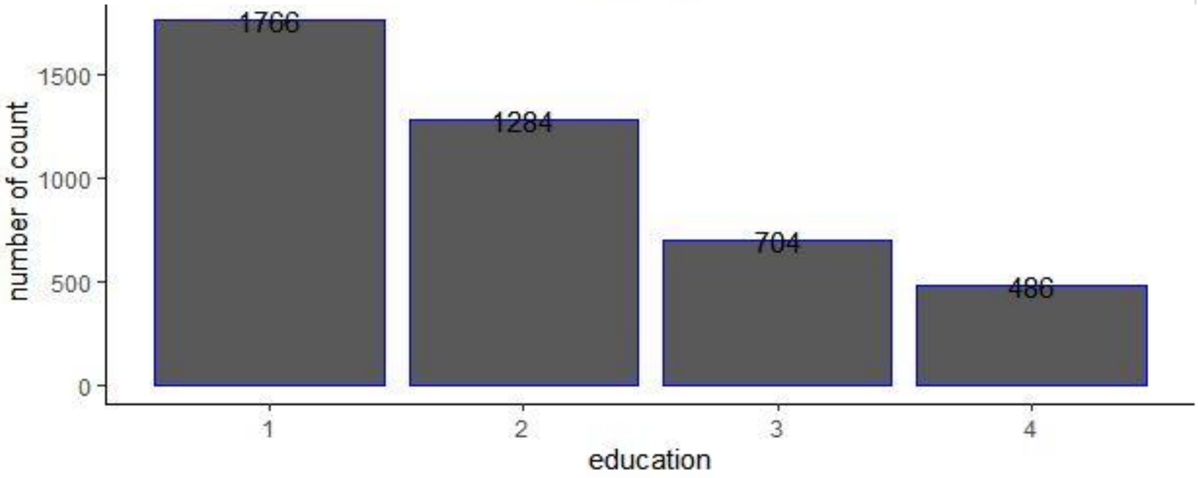
3 Visualisation Plots

3.1 Univariate analysis

3.1.1 Categorical variables



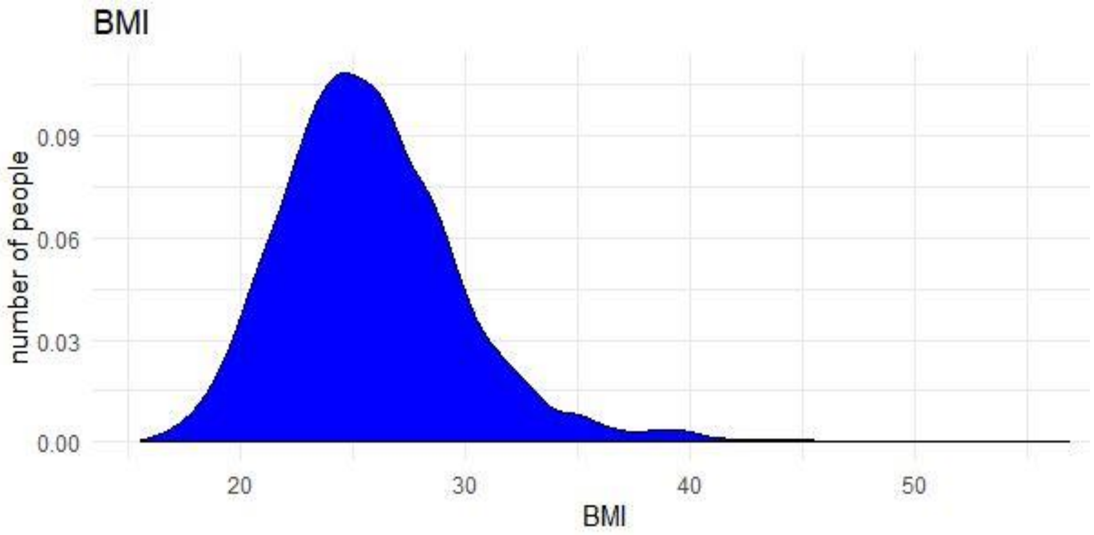




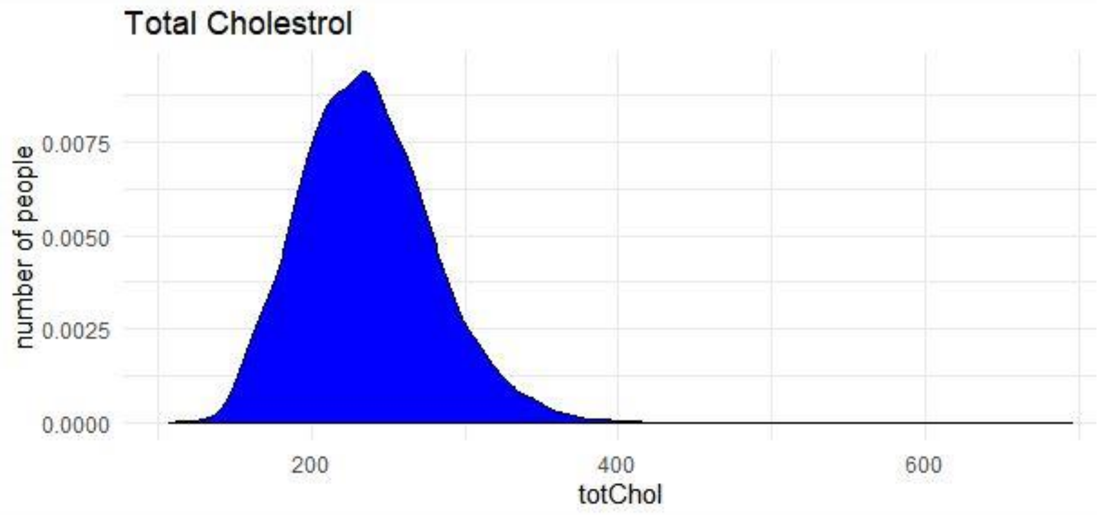
Observations

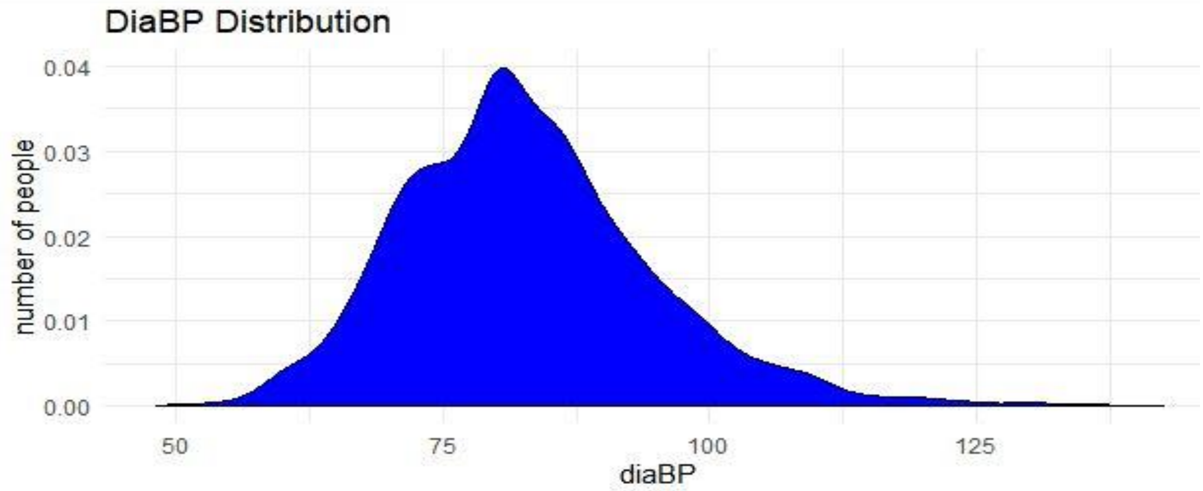
- The gender distribution is **58:42** for male and female respectively.
- Almost all the patients are under BP medication. The ones who are not, need to be evaluated for further insights which can be used to determine CHD.
- The level of education dips at a gradual rate amongst the patients as the degree goes higher.
- **18.7%** of the patients were found to be smoking with **level 1** education compared to **14.8%** for **level 4**.

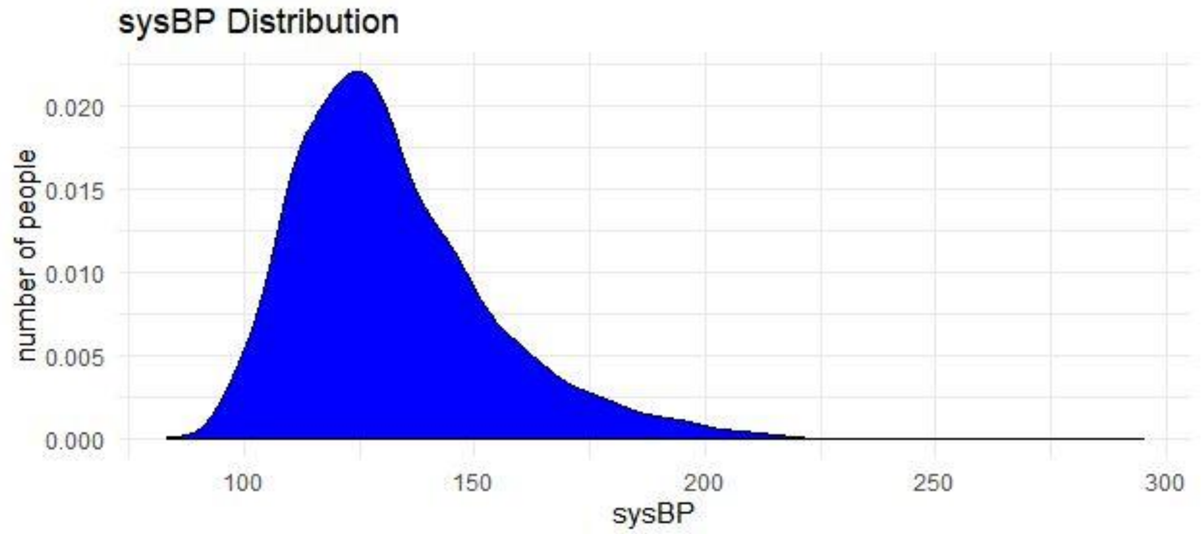
3.1.2 Continuous variables



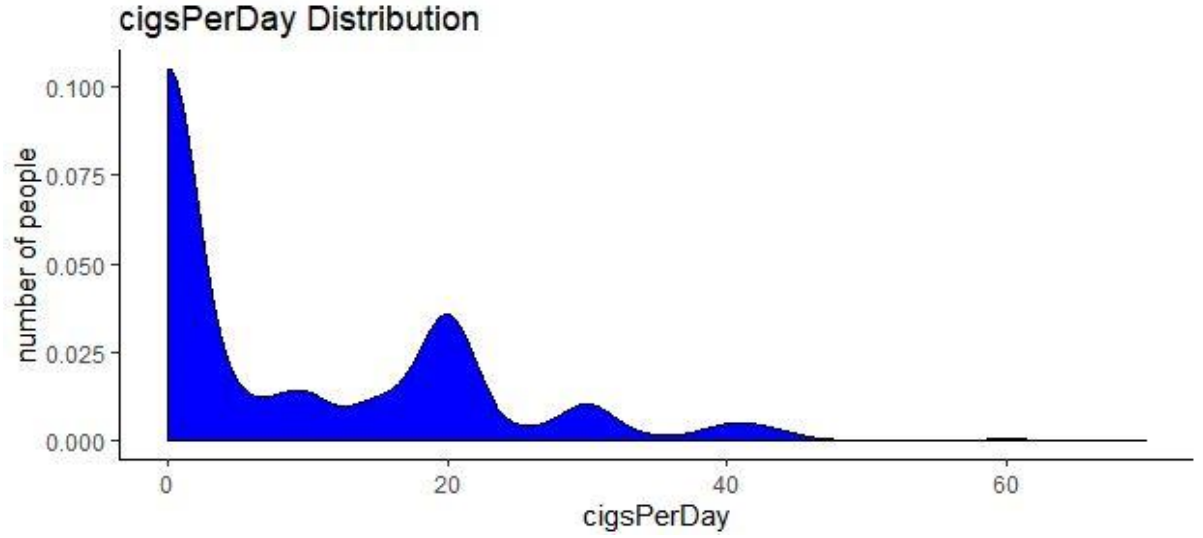
The average BMI of a healthy person lies between **18.5 to 25**. Above and under which the person can be classified as underweight and overweight. Hence, the variable needs to be categorised for detailed analysis of each segment.



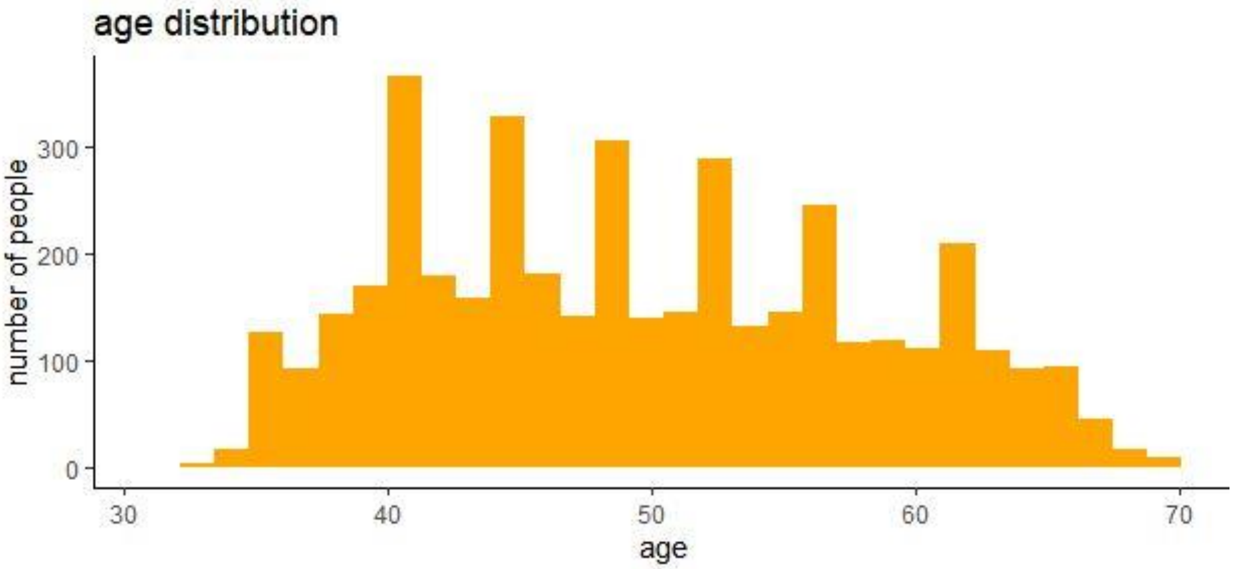




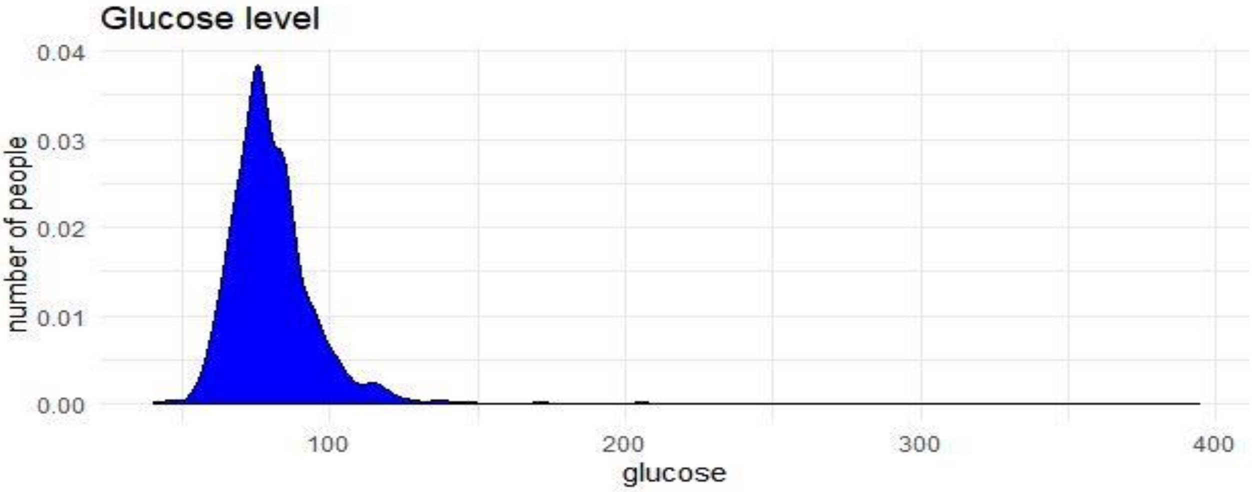
Normal Systolic and Diastolic BP is displayed by 120/80. However, as per the medical research this value holds good only for early age adults. The difference between the two BP values increases with respect to the age. Hence the dataset needs to be further categorised w.r.t age to generate better understandings for the particular age bracket.



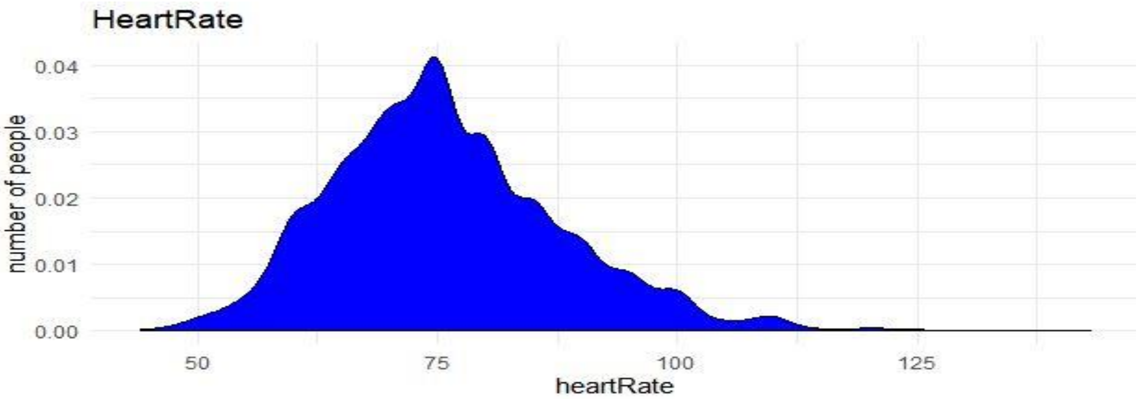
- A majority of the smoking class puffs less than 10 cigarettes per day.
- A gradual decline is observed on the count of cigarettes smoked on daily basis. There is a slight peak observed amongst the patients smoking around 20 cigarettes per day.



- Age distribution ranges from 32 to 70 with a median of around 49.



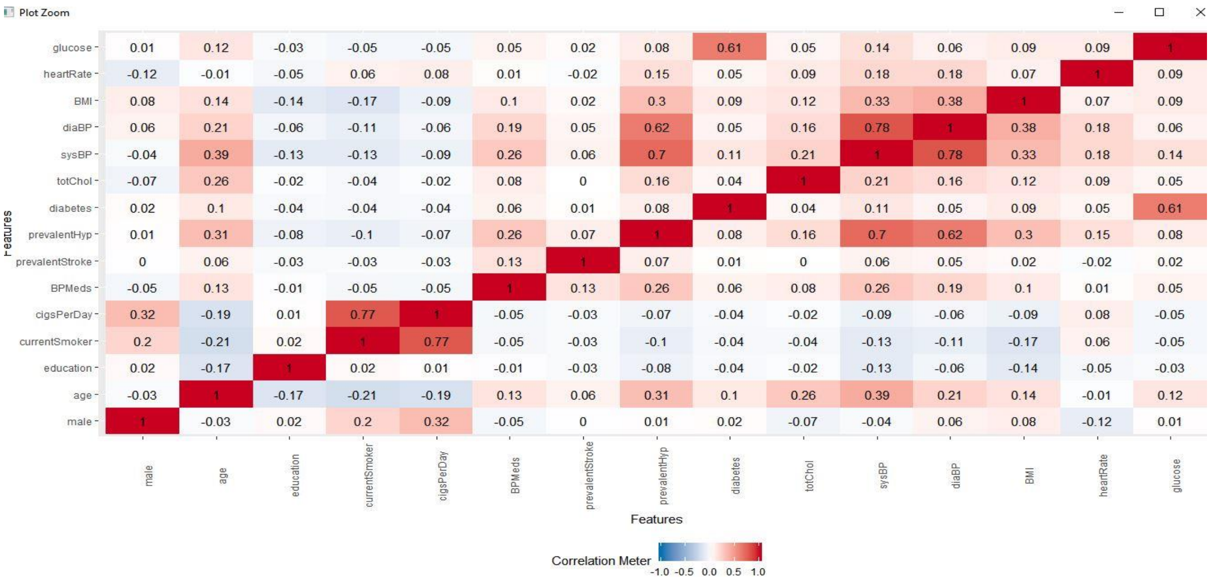
- Assuming the fact that the readings were taken during the fasting period, the glucose levels observed are more than the normal values of **70-99** for nearly **8.5%** of the patients. This also might be possible due to the fact that the patient might be suffering from diabetes.



- A small percent of people have reported heart rate above 100 who needs to be analysed further of its correlations with other predictiors

3.2 Bivariate analysis

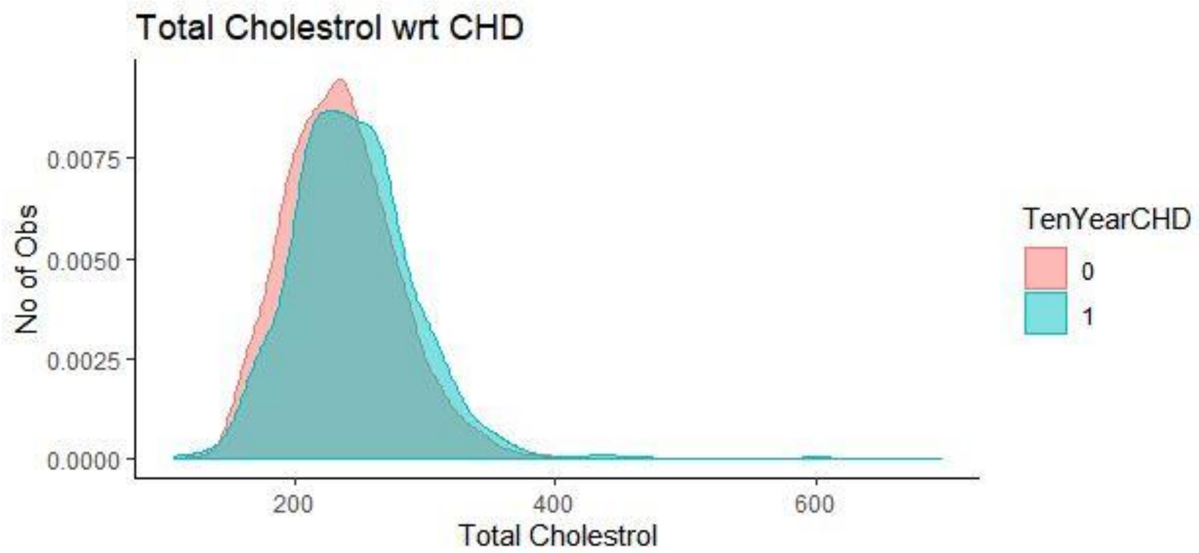
The bivariate analysis was plotted between all the variables using a corrplot. The same has been shown below.



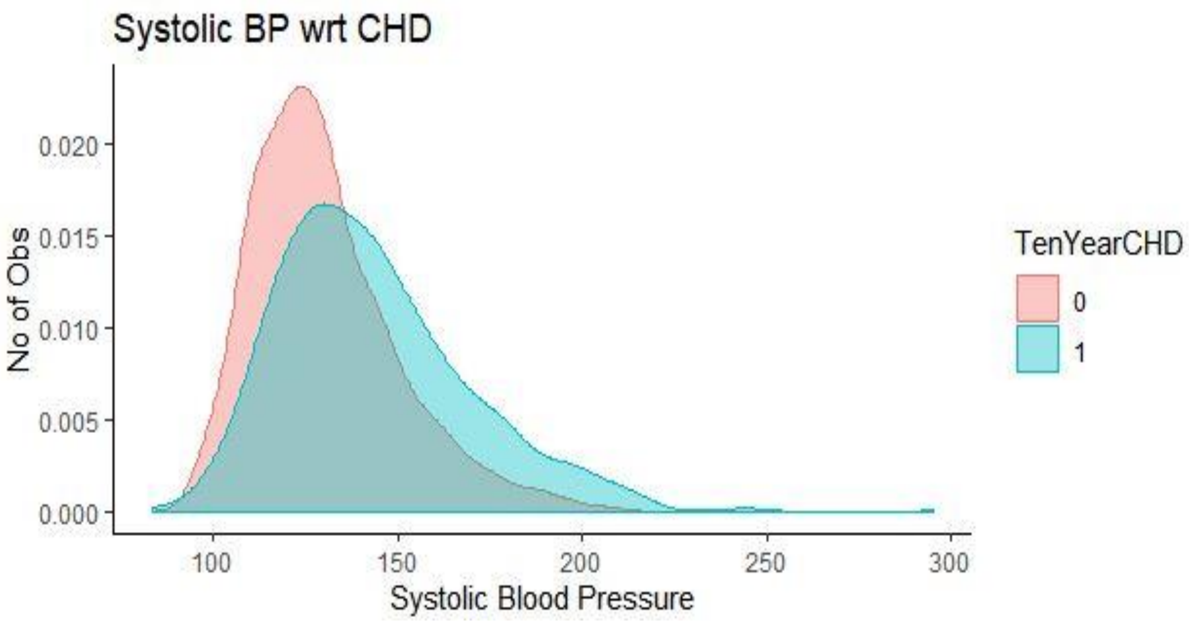
	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose
glucose	0.01	0.12	-0.03	-0.05	-0.05	0.05	0.02	0.08	0.61	0.05	0.14	0.06	0.09	0.09	1
heartRate	-0.12	-0.01	-0.05	0.06	0.08	0.01	-0.02	0.15	0.05	0.09	0.18	0.18	0.07	1	0.09
BMI	0.08	0.14	-0.14	-0.17	-0.09	0.1	0.02	0.3	0.09	0.12	0.33	0.38	1	0.07	0.09
diaBP	0.06	0.21	-0.06	-0.11	-0.06	0.19	0.05	0.62	0.05	0.16	0.78	1	0.38	0.18	0.06
sysBP	-0.04	0.39	-0.13	-0.13	-0.09	0.26	0.06	0.7	0.11	0.21	1	0.78	0.33	0.18	0.14
totChol	-0.07	0.26	-0.02	-0.04	-0.02	0.08	0	0.16	0.04	1	0.21	0.16	0.12	0.09	0.05
diabetes	0.02	0.1	-0.04	-0.04	-0.04	0.06	0.01	0.08	1	0.04	0.11	0.05	0.09	0.05	0.61
prevalentHyp	0.01	0.31	-0.08	-0.1	-0.07	0.26	0.07	1	0.08	0.16	0.7	0.62	0.3	0.15	0.08
prevalentStroke	0	0.06	-0.03	-0.03	-0.03	0.13	1	0.07	0.01	0	0.06	0.05	0.02	-0.02	0.02
BPMeds	-0.05	0.13	-0.01	-0.05	-0.05	1	0.13	0.26	0.06	0.08	0.26	0.19	0.1	0.01	0.05
cigsPerDay	0.32	-0.19	0.01	0.77	1	-0.05	-0.03	-0.07	-0.04	-0.02	-0.09	-0.06	-0.09	0.08	-0.05
currentSmoker	0.2	-0.21	0.02	1	0.77	-0.05	-0.03	-0.1	-0.04	-0.04	-0.13	-0.11	-0.17	0.06	-0.05
education	0.02	-0.17	1	0.02	0.01	-0.01	-0.03	-0.08	-0.04	-0.02	-0.13	-0.06	-0.14	-0.05	-0.03
age	-0.03	1	-0.17	-0.21	-0.19	0.13	0.06	0.31	0.1	0.26	0.39	0.21	0.14	-0.01	0.12
male	1	-0.03	0.02	0.2	0.32	-0.05	0	0.01	0.02	-0.07	-0.04	0.06	0.08	-0.12	0.01

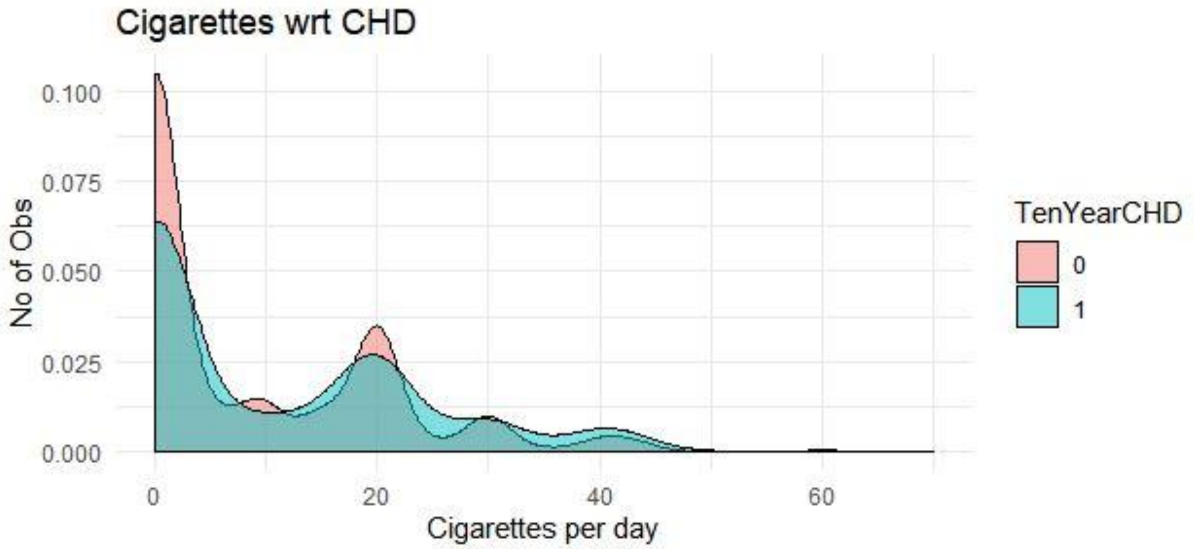
- From the above corrplot it is evident that currentSmoker and cigsPerDay are correlated. In fact, on further observation it was found that when cigsPerDay is 0 then currentSmoker is 0.
- SysBP and DiaBP were also found to be correlated. However, further investigations need to be analysed in order to perform any dimensionality reduction of any of these variables.

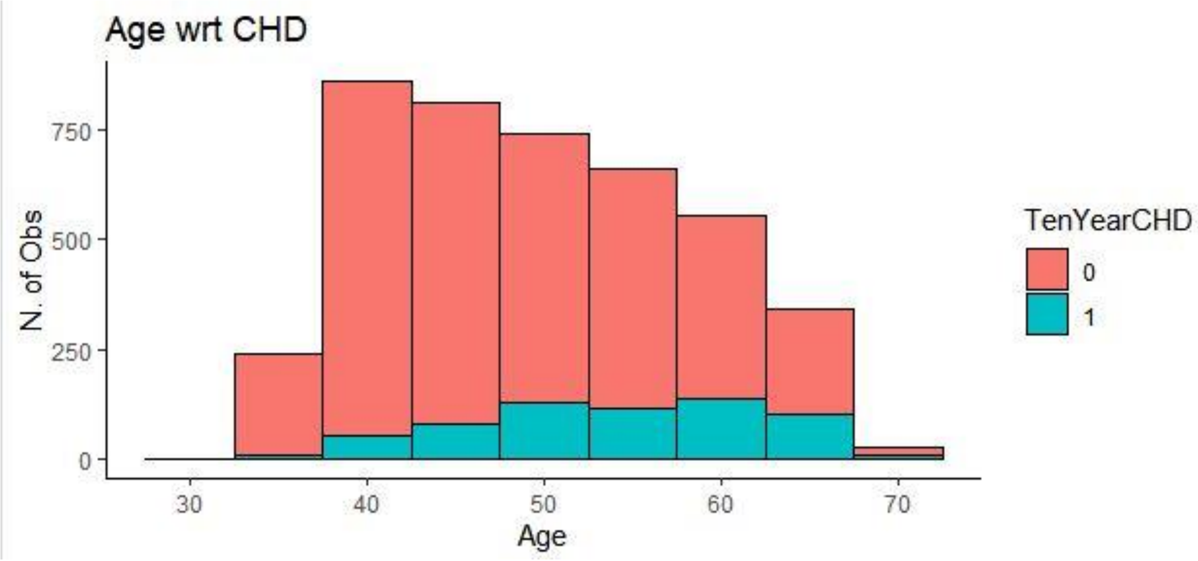
Bivariate analysis of certain categorical and continuous variables was performed w.r.t TenYearCHD.



- BP and cholesterol are found to be slightly right skewed for patients having a heart disease.
- Also, it must be noted that BP increases w.r.t age, which might mean that people who have a good control over their BP live a healthy lifestyle and are not prone to coronary heart disease.







- Here we observe that people aged above 50 tend to be more prone to having a heart disease.
- Hence, proper medications need to be ensured at later stage in life. Health check-ups for elderly people should be a must at a regular interval in order to keep up at the best.
- Here, the insurance covers might come handy to provide claims.

4. Data Pre-processing

Missing values

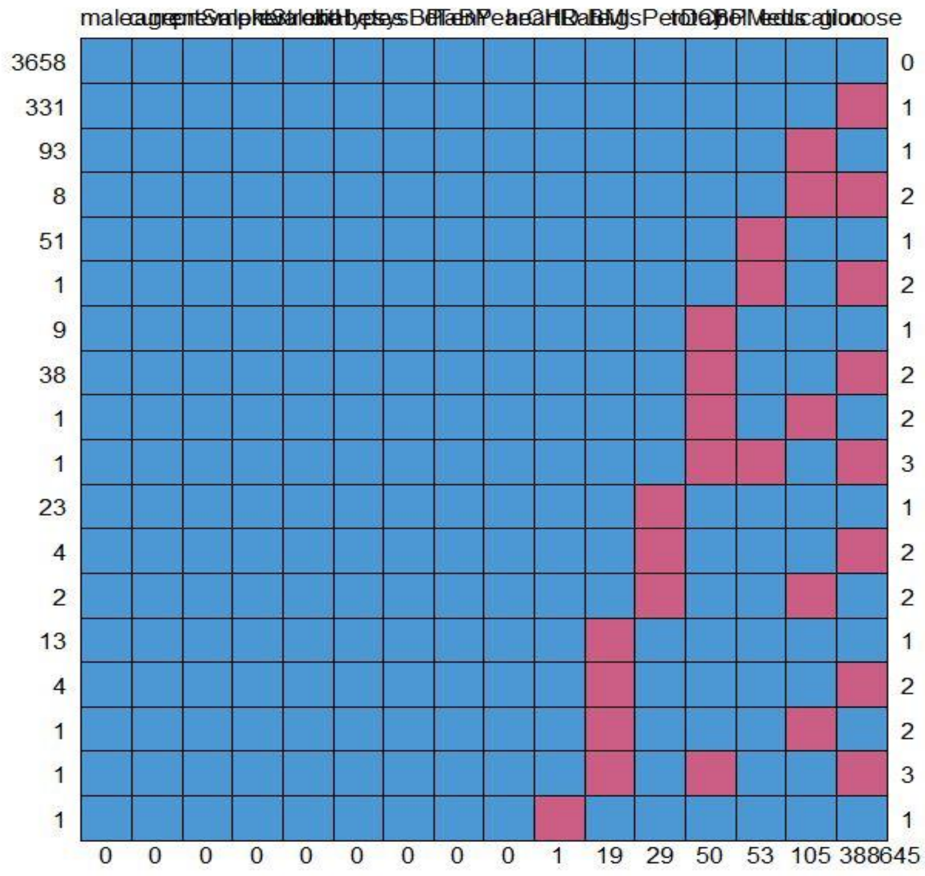
645 observations recorded missing observation of 1 or more variables influencing the heart disease which amounted to nearly **15%** of the dataset. Hence imputation technique was performed to fill all the missing observations.

- The missing values were treated using the mice package in R studio.
- The imputation was performed on all the missing values of the entire dataset and imputation was performed thrice.
- Null values were cross checked using is.na() and no further missing value was obtained.

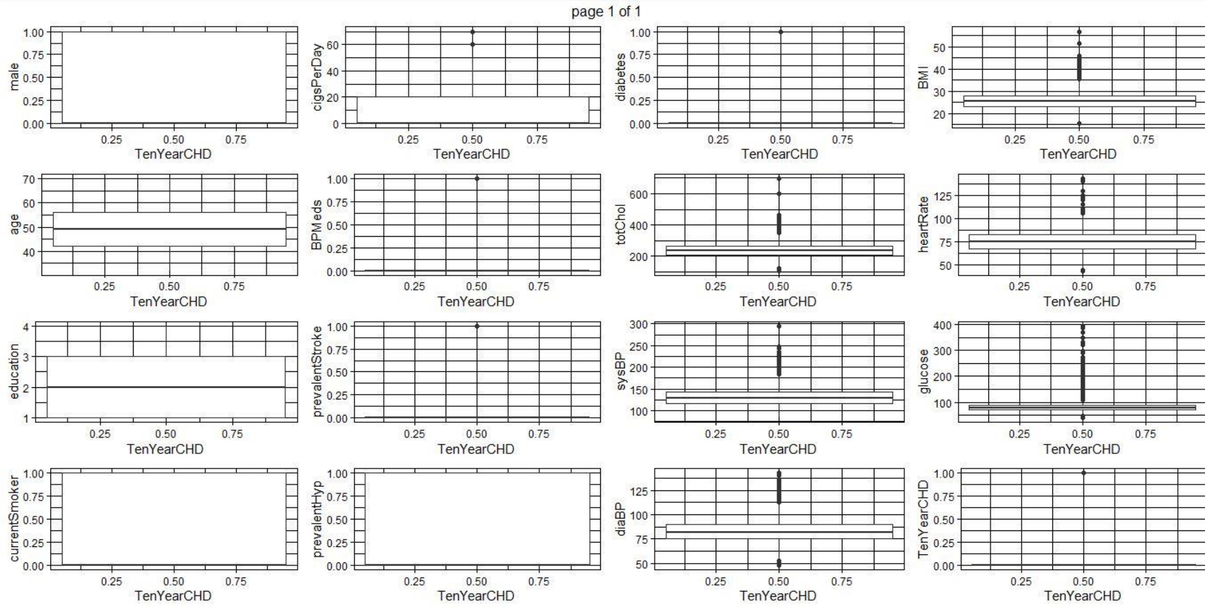
The percentage of missing variables was found to be below

male	**Age**	**education**	**currentSmoker**	**cigsPerDay**
0	0	2.47641509	0	0.68396226
BPMeds	**prevalentStroke**	**prevalentHyp**	**diabetes**	**totChol**
1.25	0	0	0	1.17924528
sysBP	**diaBP**	**BMI**	**heartRate**	**glucose**
0	0	0.44811321	0.02358491	9.1509434

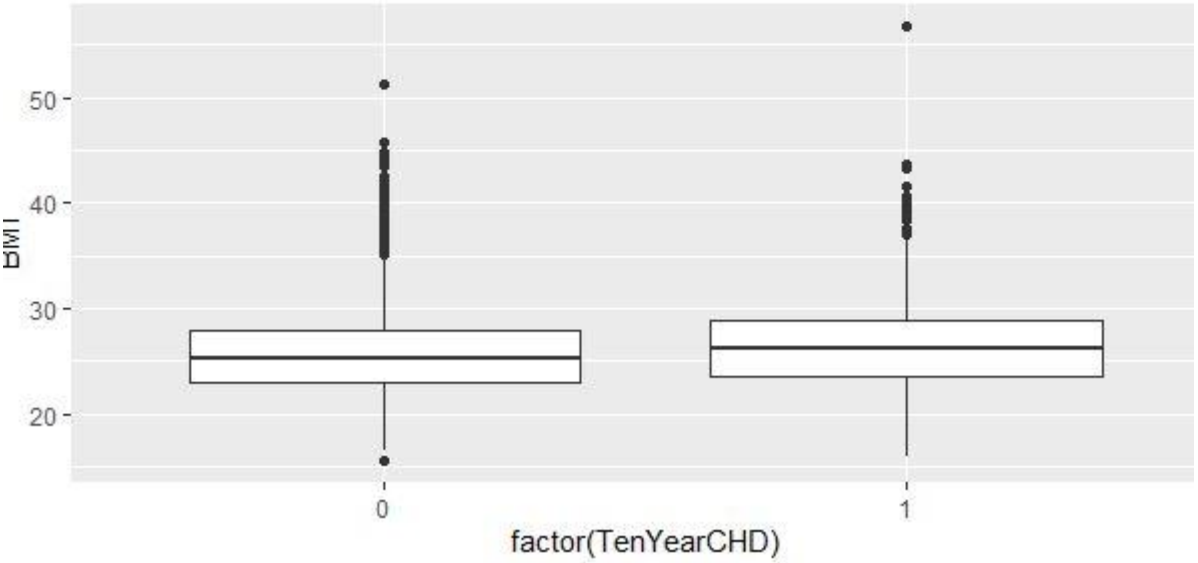
Missing value pattern was found to be below



Outlier treatment



- Certain variables like Total cholesterol, sysBP, DiaBP, BMI, heart rate and glucose tend to have outliers. However, they cannot be ignored as these values contain a major section of people having a heart disease. In a healthcare sector, the outliers are the actual cases which needs to be monitored and provided with necessary aid.



Patients who have CHD tend to have a slight increase in BMI when compared to the latter.

Variable Transformation/Creation of new variables

- BMI was a continuous variable which had been transformed into 4 categorical segments as follows

Condition	Name
<18.5	Underweight
18.6-24.5	Healthy
24.6-29.9	Overweight
>30	Obese

These were then studied individually for their significance.

- Education levels were also categorised individually as education 1 to 4 to study and identify their impacts on CHD.

Removal of unwanted variables

Below is the list of variables which were removed as they had either impacted from multi-collinearity, high VIF value or poor significance in the model building.

- currentSmoker
- BPMeds
- diabetes
- diaBP
- BMI
- heartrate

Splitting the data

Smote was applied to treat the unbalanced dataset of 85:15 for TenYearCHD. The data was then split based on the dependent variable into train and test respectively in 70:30 ratio.

5. Model building

5.1 Logistic Regression

Logistic regression was used for building the model and the final model was built on the below variables based on the significance levels and low VIF values.

- Male
- Age
- Cigs per day
- PrevStroke
- PrevHYP
- Totchol
- SysBP
- Glucose

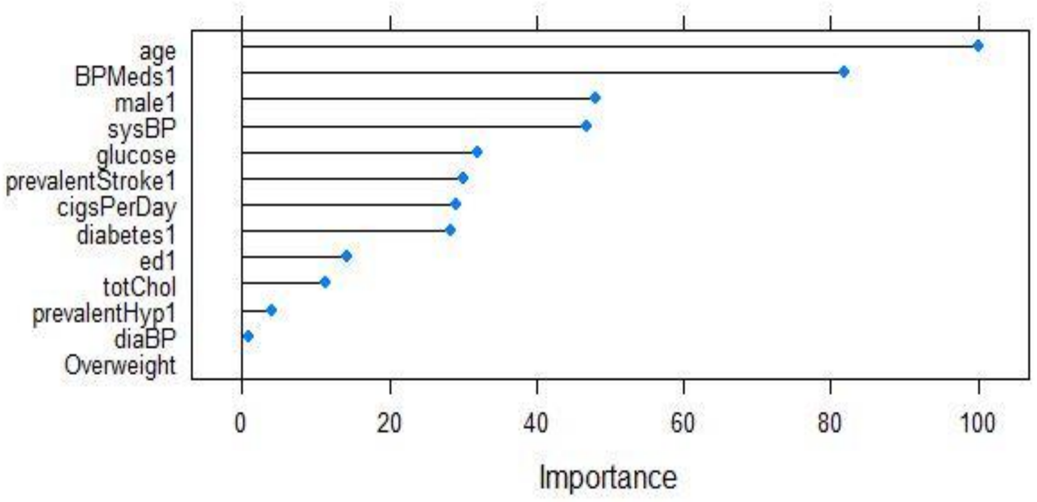
- With log.pred =0.3 we were able to obtain the best logistic regression accuracy at **70.41%** using a confusion matrix.
- Below is the confusion matrix obtained for logistic regression

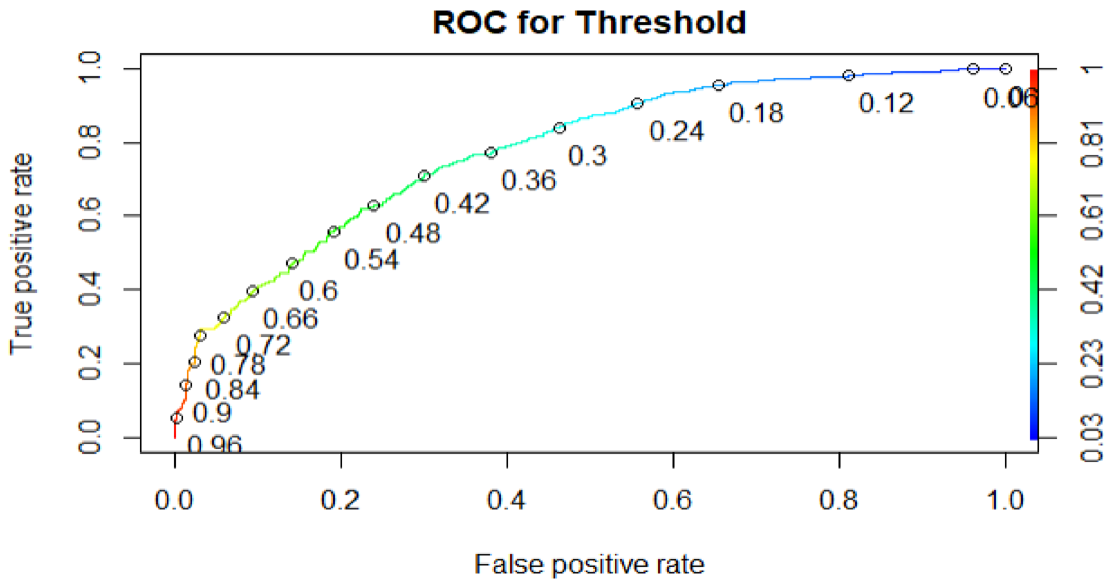
	Reference	
Prediction	0	1
0	708	152
1	293	351

Variable Importance

- Age, BPMeds, Male and SysBP turned out to be the most important predictors in the model.

ROC curve was plotted and a threshold of 0.3 was selected.

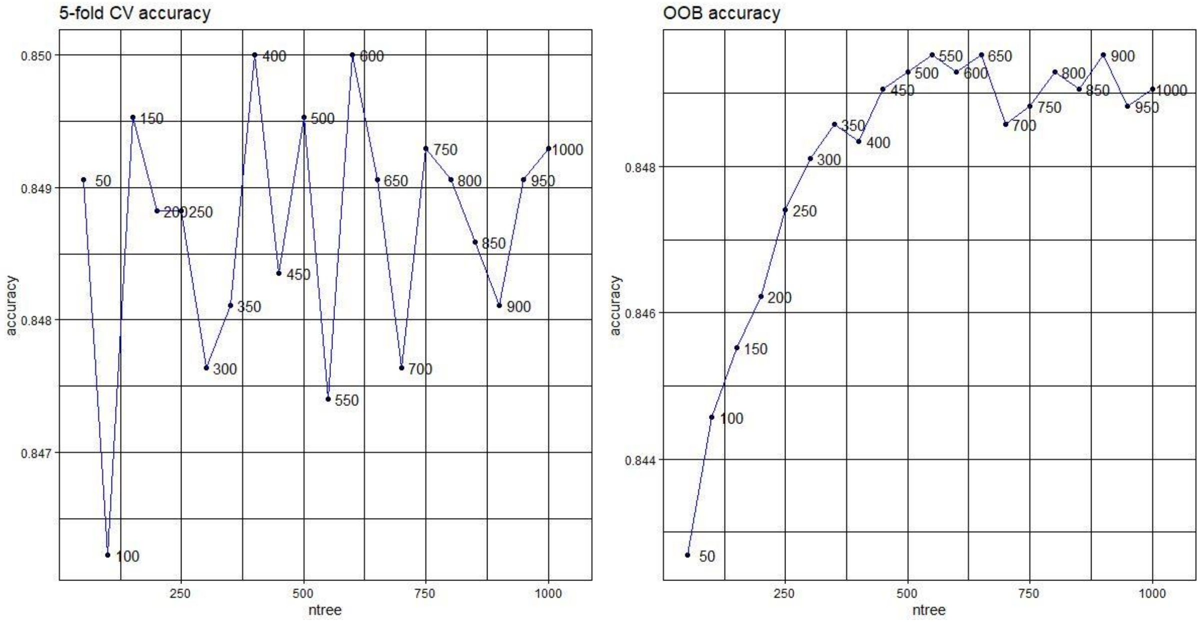




5.2 Random Forest

- K fold Cross Validation and Out of Bag technique was used in Random Forest.
- At ntree=400 best accuracy for the model was achieved

```
|   | ntree|  accuracy|
|:--|-----:|---------:|
|8  |   400| 0.8500000|
|12 |   600| 0.8500000|
|10 |   500| 0.8495283|
|3  |   150| 0.8495283|
|15 |   750| 0.8492925|
|20 |  1000| 0.8492925|
|13 |   650| 0.8490566|
|16 |   800| 0.8490566|
|19 |   950| 0.8490566|
|1  |    50| 0.8490566|
|4  |   200| 0.8488208|
|5  |   250| 0.8488208|
|17 |   850| 0.8485849|
|9  |   450| 0.8483491|
|7  |   350| 0.8481132|
|18 |   900| 0.8481132|
|6  |   300| 0.8476415|
|14 |   700| 0.8476415|
|11 |   550| 0.8474057|
|2  |   100| 0.8462264|
```

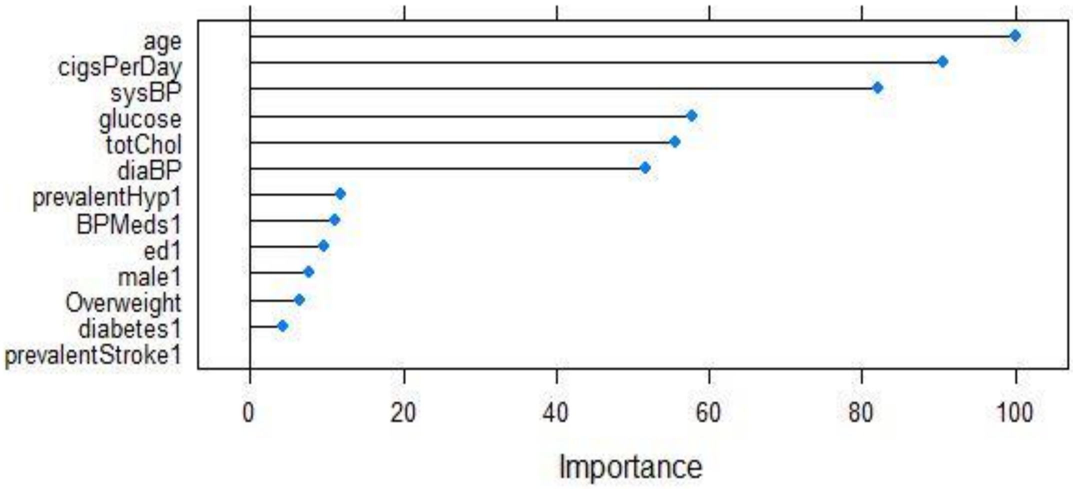


- Here, CV and out-of-bag (OOB) have been used for evaluation of RF performance.
- At ntree=400 we are able to reach the maximum accuracy with 81.45**%**.

- The confusion matrix for the model has been shown below

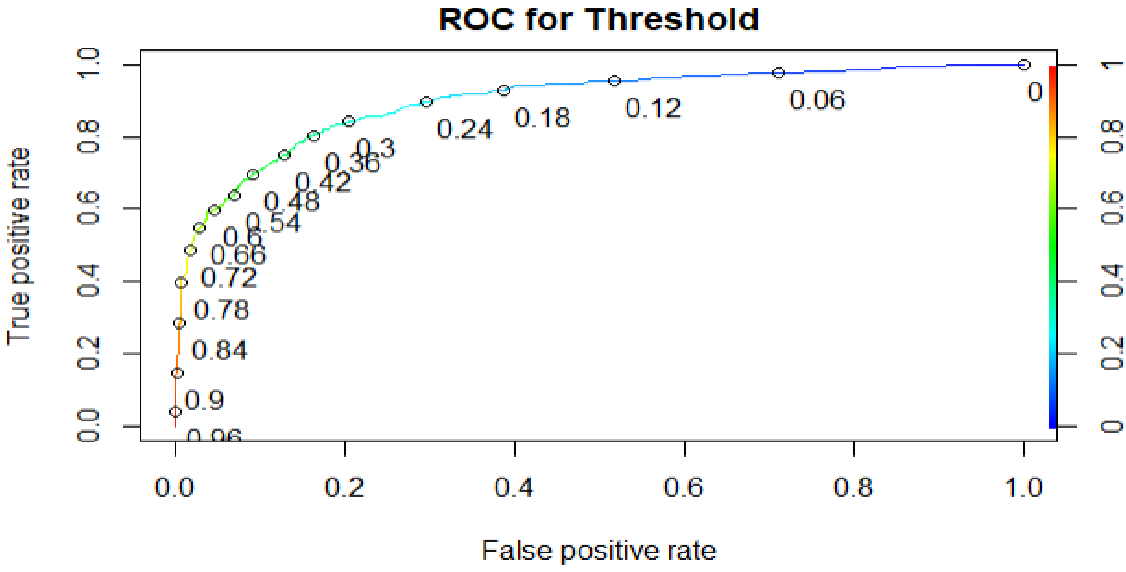
	Reference	
Prediction	0	1
0	652	86
1	208	558

- Sensitivity and specificity were observed at 84.01% and 79.53% respectively
- Variable importance was also performed on the dataset and the same has been plotted below



- We observe that Age, cigs per day and systolic BP, glucose and total cholesterol turns out to be the top most important variables.

- ROC curve for the model has been shown below



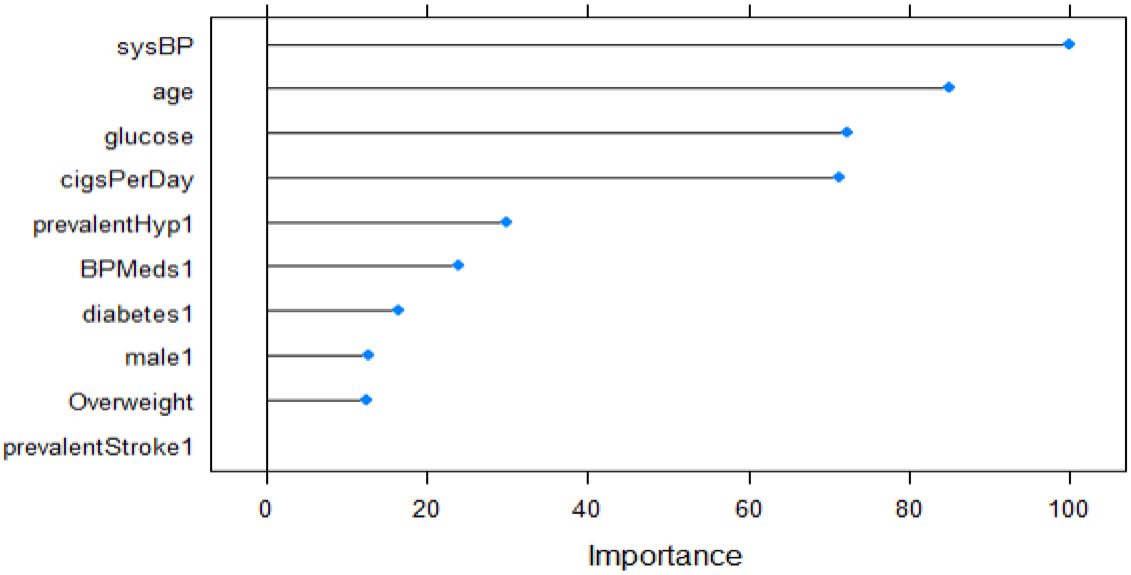
6. Ensembling techniques

6.1 Bagging

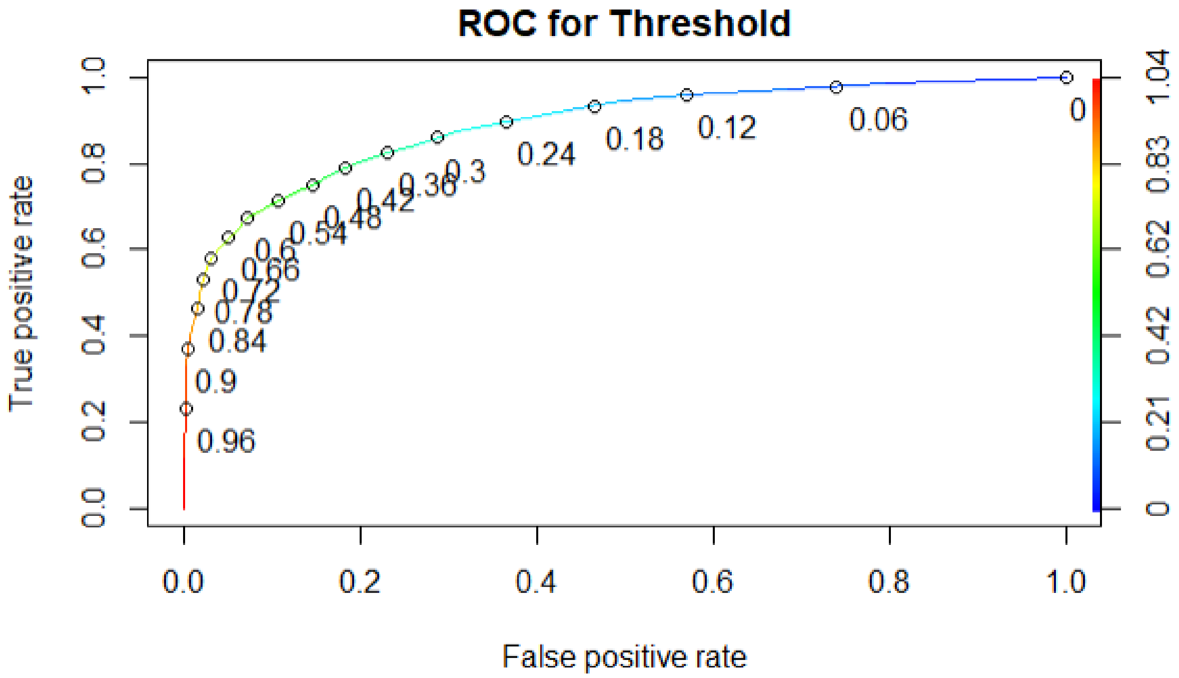
- Bagging was performed on the dataset and different minsplit and max depth was used to predict the model
- With minsplit=20 and max depth = 6 below confusion matrix was obtained
- Confusion matrix for the model has been shown below

	Reference	
Prediction	0	1
0	654	90
1	206	554

- Accuracy of 80.32% was achieved by the model on the test dataset.
- Sensitivity of 84.32% and Specificity of 73.60% was recorded which shows an improvement in determining the True positive rate and True negative rate in a person having a disease.
- Variable Importance for the model has been shown below



ROC for the model has been shown below



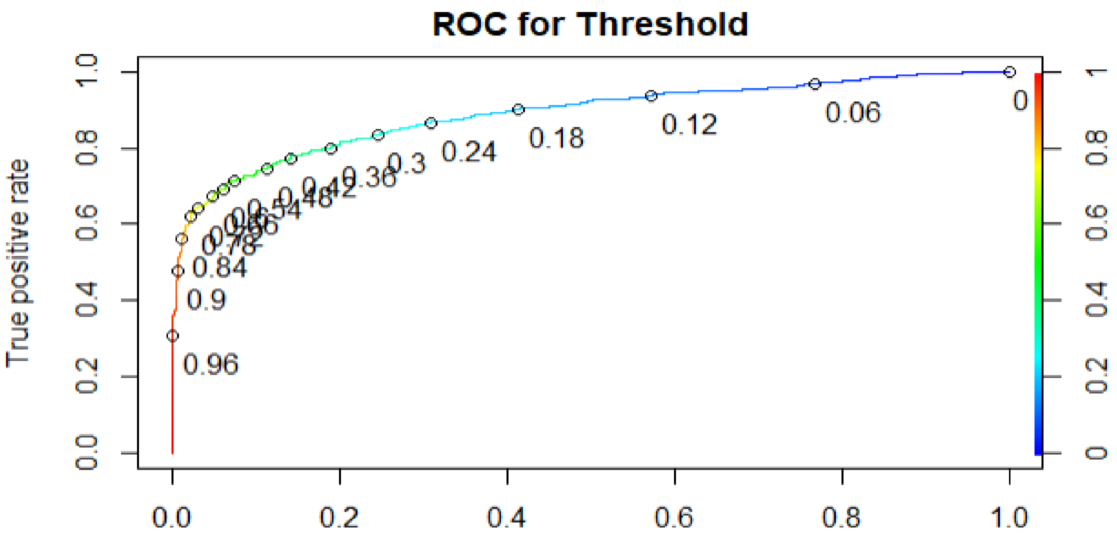
6.2 Boosting

- Boosting was performed on the dataset by using the default values of eta, max_depth and nrounds.

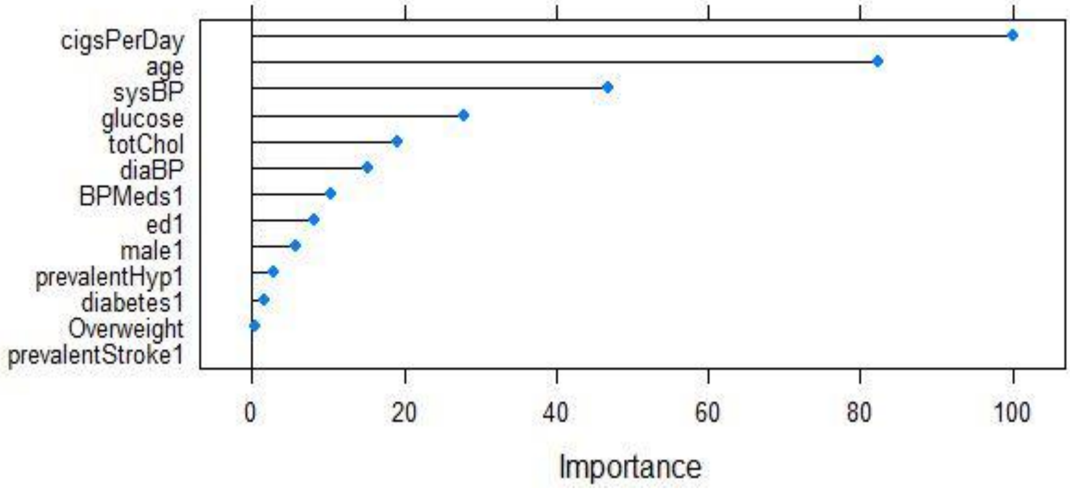
Below are the outputs recorded for XGBoost

	Reference	
Prediction	0	1
0	787	73
1	185	459

- XGBoost recorded 82.85% accuracy when applied on SMOTE.
- With selecting the best value for eta, max depth and n rounds at 0.9, 15 and 50 for **XGBoost** respectively, an improvement was observed in predicting the target variable to 459 accurate cases.
- Sensitivity: 84.49% and Specificity: 81.97% was recorded which gave the best results amongst all the models.
- Boosting when applied with SMOTE reduced the percentage of False Negative cases.



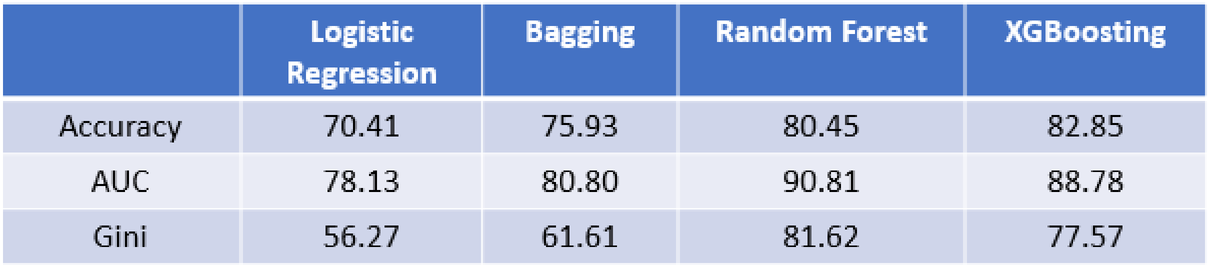
ROC curve for the model is shown above



- Variable Importance was plotted for XGBoost with Cigs Per Day , Age and sysBP accounted for the most important predictors as shown below

Interpretation of the model

- Of all the models used for predicting the dataset SMOTE when applied on XGBoost performs best when compared to others in predicting more accurate cases of having a heart disease.
- XGBoost is an ideal choice in terms of accuracy and also in determining True Positive rate and True Negative rate of the model.
- Although SMOTE helped in improving the accuracy of the model, it is more recommended if the dataset a greater number of actual cases to improvise on the prediction accuracy.



	Logistic Regression	Bagging	Random Forest	XGBoosting
Accuracy	70.41	75.93	80.45	82.85
AUC	78.13	80.80	90.81	88.78
Gini	56.27	61.61	81.62	77.57

- Smoking doubles the risk of having a heart disease
- BP and cholesterol were found to be slightly higher for patients having a heart disease.
- Also, it must be noted that BP increases w.r.t age, which might mean that people who have a good control over their BP live a healthy lifestyle and are not prone to CHD.
- People aged above 50 tend to be more prone to having a heart disease.

Business Insight

- Health insurance companies can target the right age group and also sell group insurance as family cover if all the members are adults
- **Wearables** - Smart watches like Fitbit can be used to track the distance travelled as well as the calories you burn. Some also monitor your heart rate and sleep quality. This can help you to track your fitness and also to be on top of your health
- Smart watches can undergo a tie up with insurance companies to improve their sales by targeting the needy ones and use them as a word of mouth to enhance the business exponentially

Recommendations

- It is recommended for everyone to indulge in physical exercises to overcome the problem of obesity and sedentary lifestyle
- People aged over 50 are advised to undergo routine check-up and take necessary medications if required
- Stress being one of the reasons of smoke should be replaced with activities like yoga
- Maintain a reasonable body weight
- Adults to undergo a routine check-up, especially if your family has a history of heart disease