# Exploratory Data Analysis

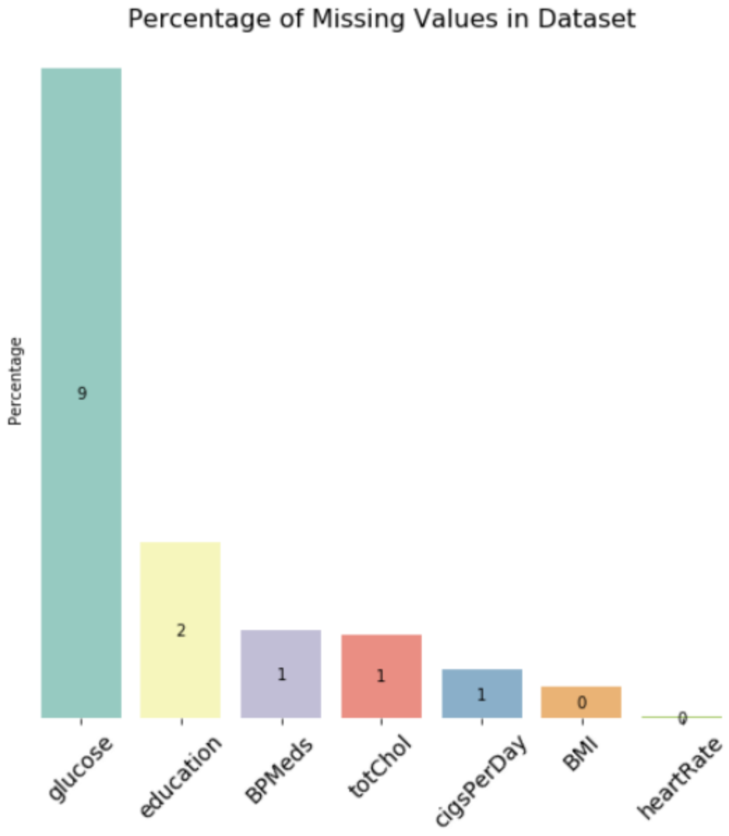
Cardiovascular disease (CVD) is the leading cause of serious illness and is the most common cause of mortality in developed and developing countries. In 1945, President Franklin Delano Roosevelt died of haemorrhagic stroke due to uncontrolled hypertension, raising awareness about the rising toll of cardiovascular disease. The Framingham Heart Study is a long-term, ongoing cardiovascular cohort study of residents of the city of Framingham, Massachusetts. The study began in 1948 under the direction of the National Heart Institute (now known as the National Heart, Lung, and Blood Institute or NHLBI) with 5,209 adult subjects from Framingham and is now on its third generation of participants. The study has not only contributed to our understanding of the natural history of cardiovascular disease and stroke, it also enabled us to identify their major risk factors. The overall impact of the Framingham Heart Study is vast, and the study continues to unveil new insights into human health until this day.

In this Section, we have performed Exploratory Data Analysis (EDA) to develop a better understanding of the Data before modelling and gaining Insights from the Data.

## About the Data

The Dataset has **4240 Rows and 16 Columns**. There are **7 features** that has Missing Values. The missing values if dropped removes **15%** of the Dataset.

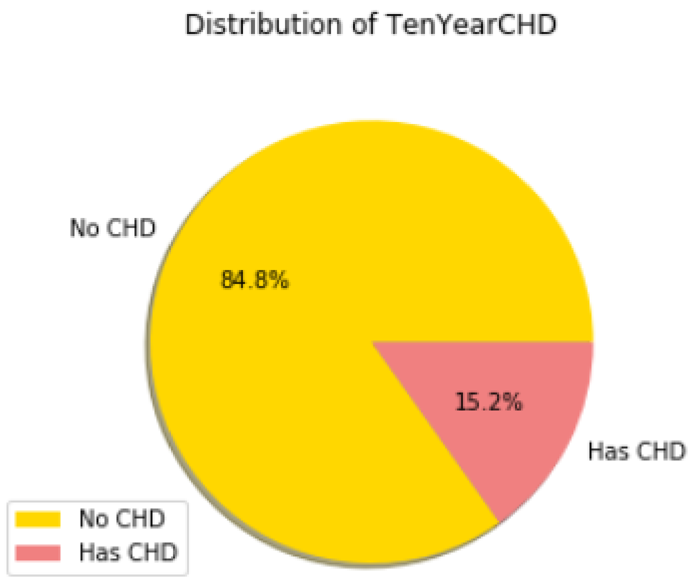
This is a Bar Chart Representation of percentage of Missing Values with respect to each feature, **“glucose”** has highest number of missing values.



## Imbalanced Classification Dataset

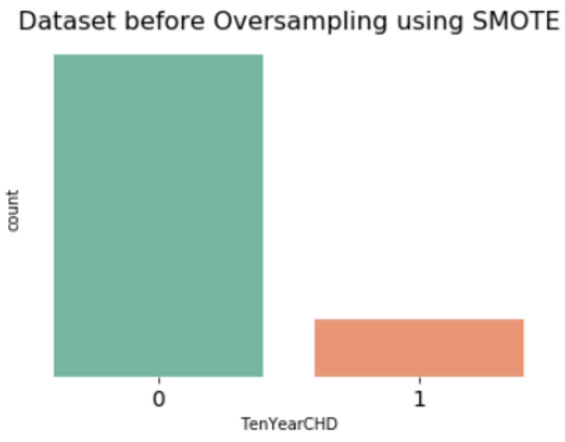
The ‘y’ variable or the Target/Prediction Label is as follows: **“TenYearCHD”** which is **10-year risk of coronary heart disease (Class 1/0)**

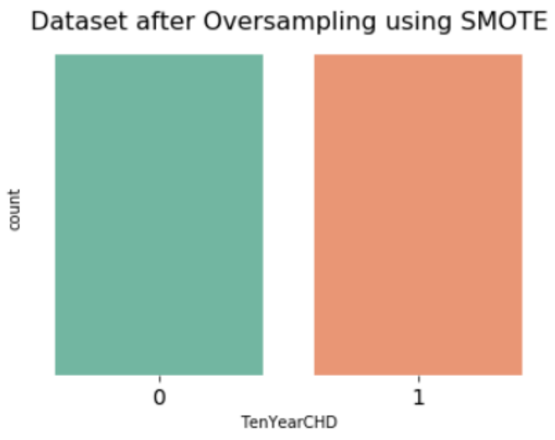
The Following Pie Chart is a representation of Distribution of variable “TenYearCHD” between class 1 and 0. We can understand from the chart below that the Dataset is Imbalanced as there are only 15.2% values in Class 1 and 84.8% in Class 0.



Developing predictive models using Imbalanced classification datasets leads to Bias in the model. Hence, we use **SMOTE - Synthetic Minority Oversampling Technique** which is one of the most used oversampling methods to solve the imbalance problem.

The Following Bar charts represents **Distribution of “TenYearCHD”** before and after using SMOTE Oversampling Technique.





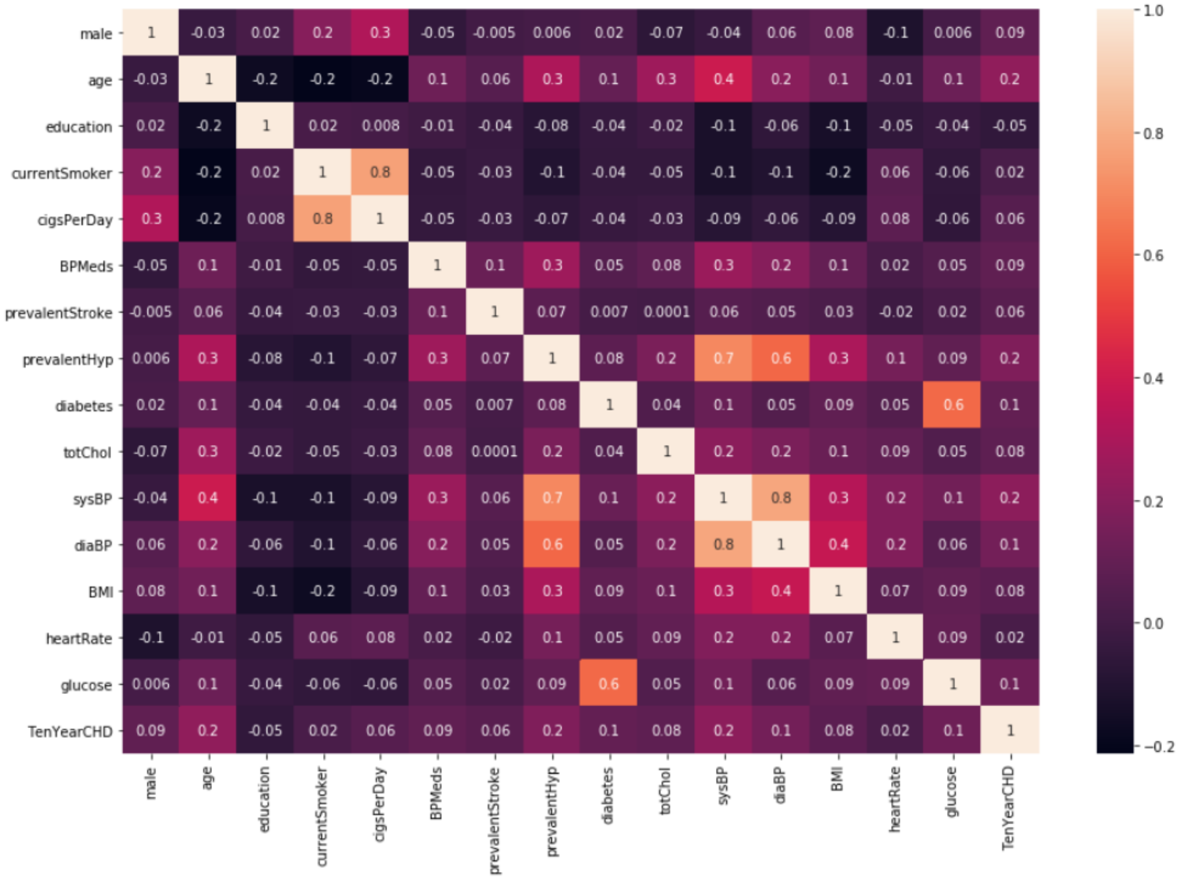
## Correlation Coefficient

Correlation Coefficient is a feature Selection method. Feature selection methods can help identify as well as remove redundant and irrelevant attributes from data that do not contribute to the predictive power of the model. The objective of feature selection are as follows:

- Improving the prediction performance.
- Providing faster and more cost-effective predictors.
- Providing a better understanding of the underlying process that generated the data.

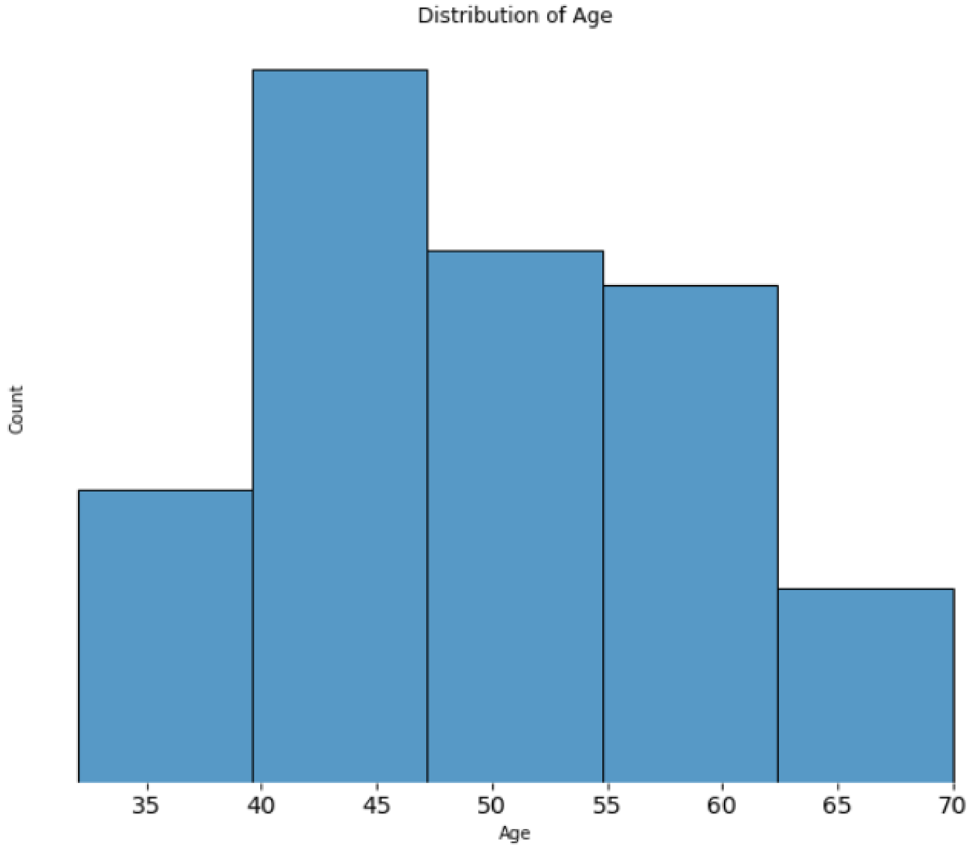
It makes intuitive sense to choose features that are highly correlated to the target variable. The more correlated the features are to the target, the easier it is for the machine to predict it. **Positive Correlation:** means that if feature A increases then feature B also increases or if feature A decreases then feature B also decreases. **Negative Correlation:** means that if feature A increases then feature B decreases and vice versa. **No Correlation:** No relationship between those two attributes.

The Following Chart is a Heat map that shows the Correlation between features. We can see feature like ‘**DiaBP’** and ‘**SysBP’** are highly correlated with ‘**prevalent Hypertension’** and ‘**glucose’** is highly positively correlated with feature **‘diabetes’**.

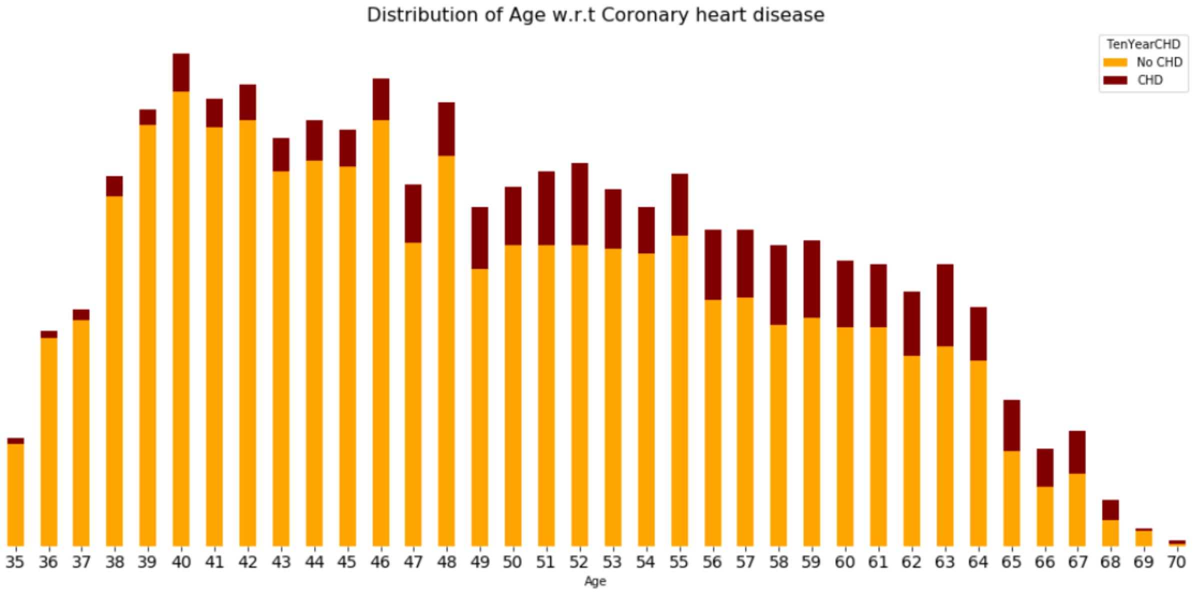


## Insights from Data

The Histogram Chart below gives us a distribution of all the **4240** Patients Age during the time of test. The Age Group approximately falls between **32** to **70**.

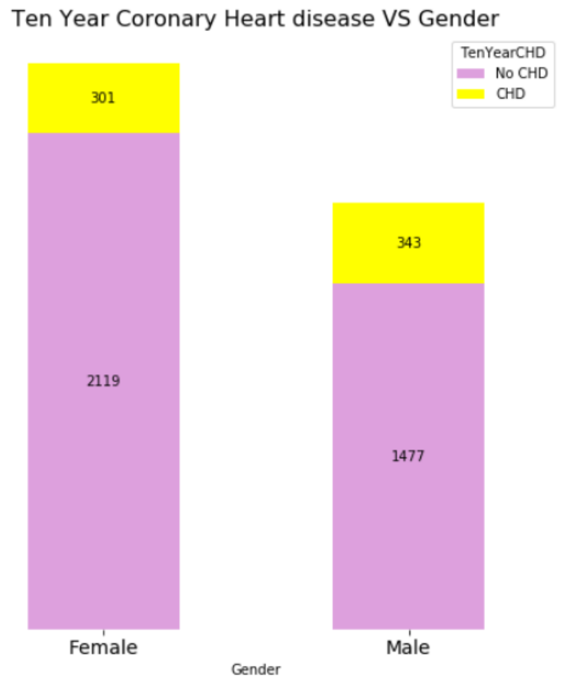


The Bar Chart below is a distribution of ‘Age’ of Patients with respect to 10 year risk of Coronary heart disease (CHD) and we learn that people between age group **47 to 65 have highest risk of developing Coronary Heart Disease.**

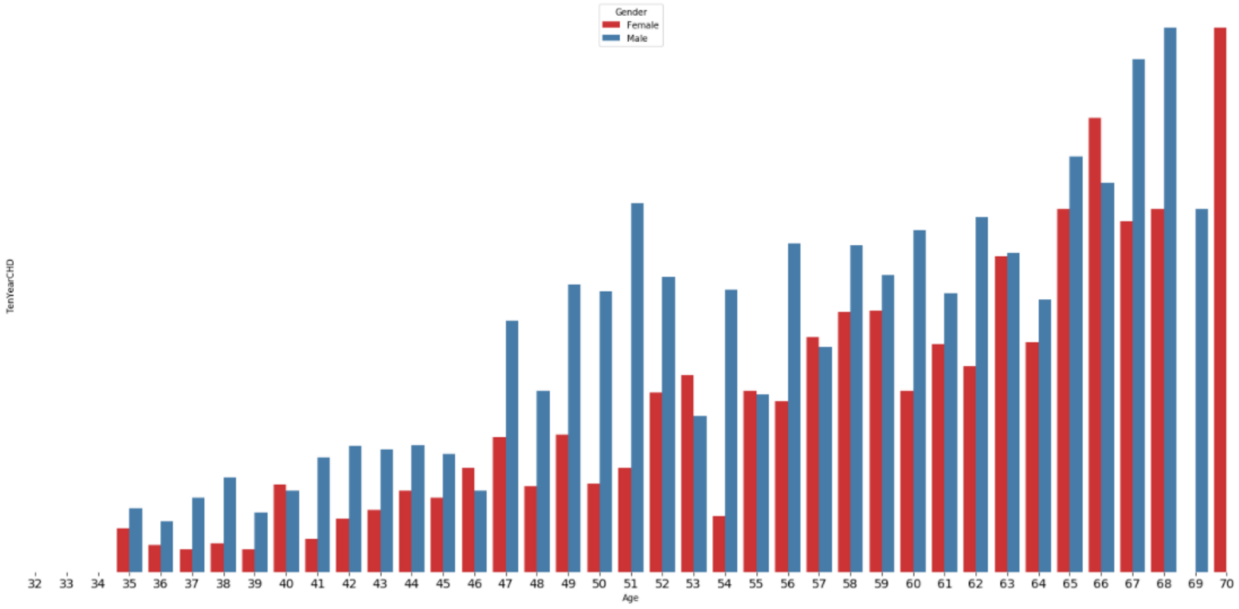


## Gender and Risk Factor

The Stacked Bar chart below is a representation of Gender distribution with respect to Coronary Heart disease. We can see that out of total patients in the dataset **59.3% are Females** and **41.0% are Males**. Out of 59% percent Females, 14.2% have a risk of developing Coronary Heart disease (CHD) and out of 41% of Males 23.2% Males have a risk of developing CHD. Hence, we can say **Males have a higher risk of developing Coronary Heart Disease when compared to Females.**

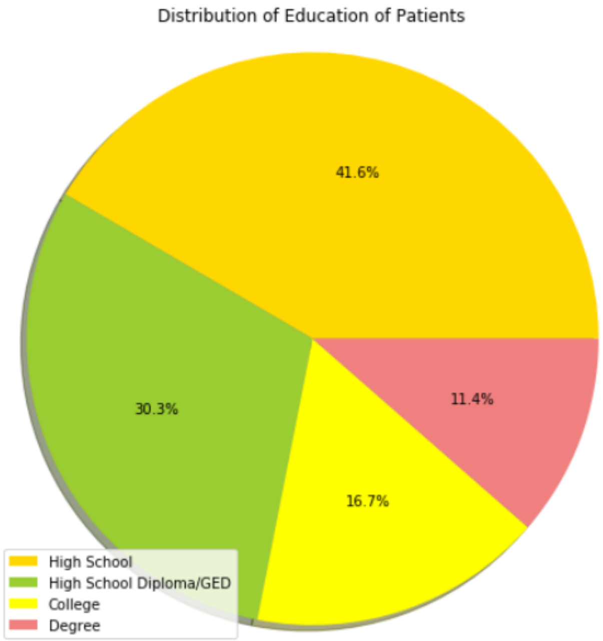


However, in **1976** using Framingham Heart Study Dataset it was observed that **Heart disease risk is found to increase in women after Menopause.** We can observe from chart below that women **above age of 51** which is the average age for Menopause are prone to develop Coronary Heart Disease.

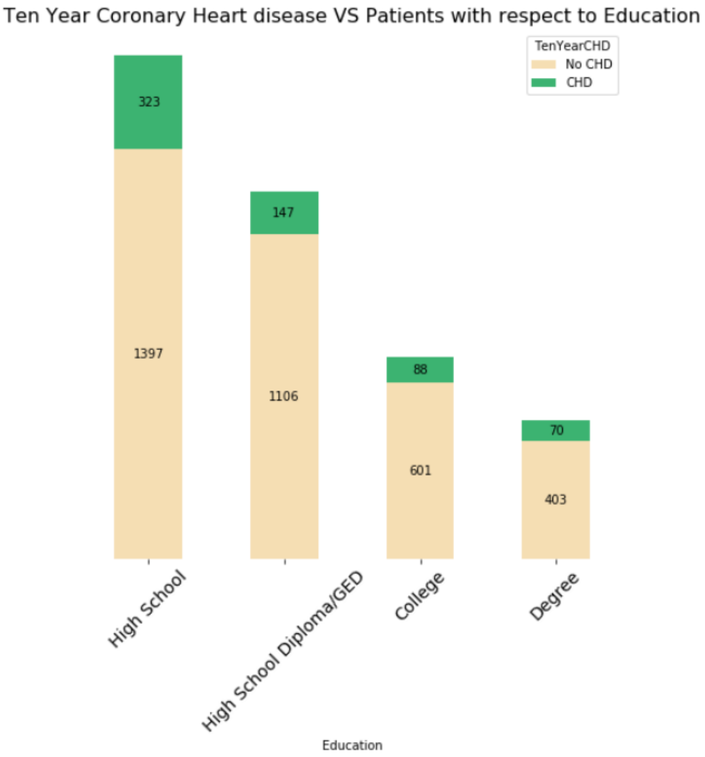


## Education and Risk Factor

The Following Chart below is a Pie-chart which shows distribution of all the Patients Education Level.



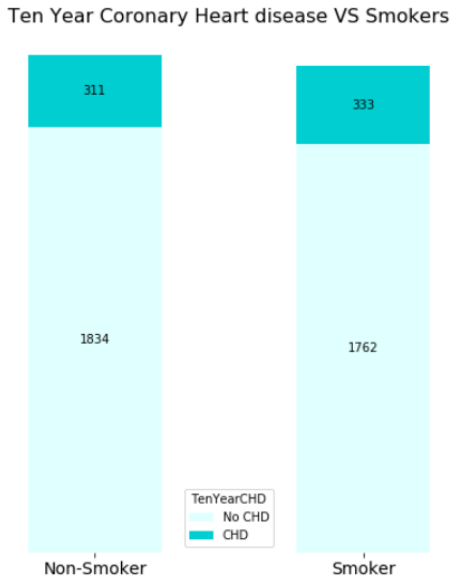
In the given Stacked Bar-Chart below, we are unable to infer any valuable insight with respect to, whether a person's Education affects Risk Factor implicitly or explicitly. Also, in the Heat map above we had observed the **Correlation coefficient** of features **'education'** and **'TenYearCHD'** is **-0.05.** Hence, it is safe to say that there might not be a link between a person's education and Coronary Heart disease.



## Smoking as a Risk Factor for CHD

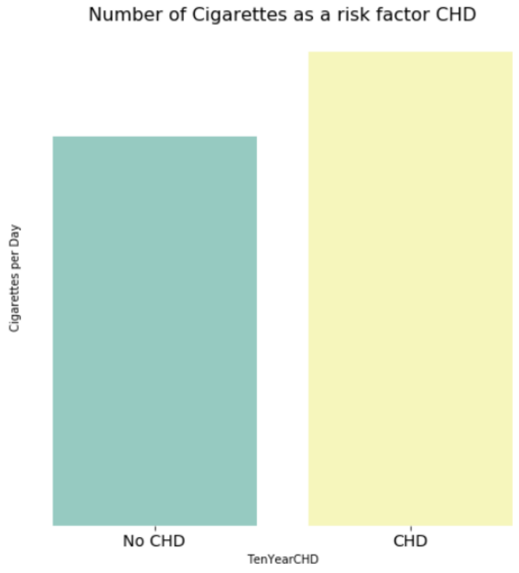
The Stacked Bar Graph below is a representation on how smoking is linked to Coronary Heart Disease (CHD). **Approximately 49.41% of total patients are Smokers. Cigarette smoking is linked to CHD was determined in Framingham in 1959** and a first report was released in **1964** on Smoking and Health.

Out of Non-Smokers **16.9%** Patients have a risk of developing Coronary Heart Disease and out of Smokers **19%** Patients have a Risk of developing CHD.

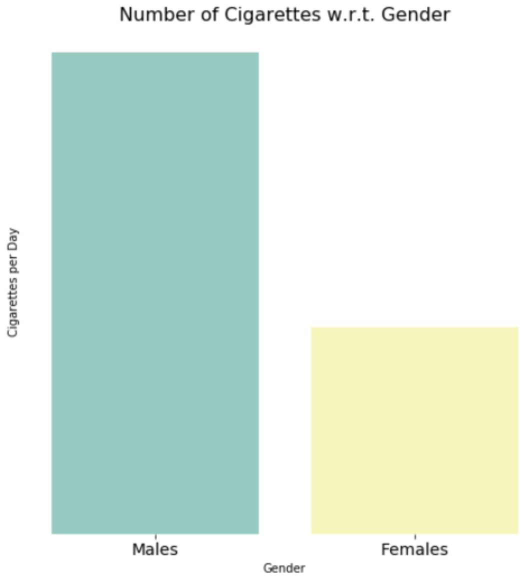


We can also confirm from the below **graph (A)** that **Patients who consume a greater number of Cigarettes have a Higher risk of developing Coronary Heart Disease.**

From **Graph (B)** we have observed **Males consume a greater number of Cigarettes in Comparison to Females** and Females have lower risk of CHD. Hence, Number of Cigarettes and Smoking both are linked to Coronary Heart Disease.



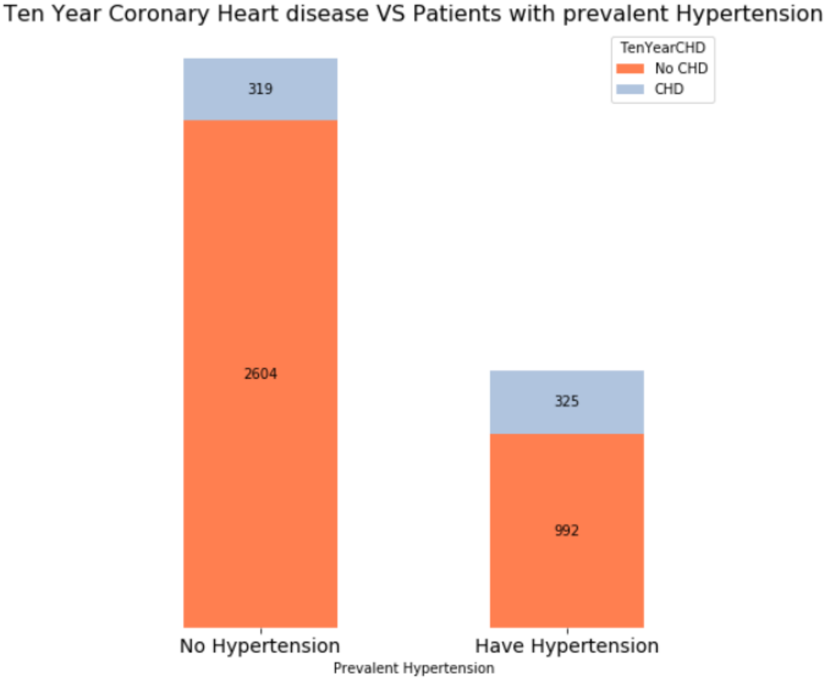
**(A)**



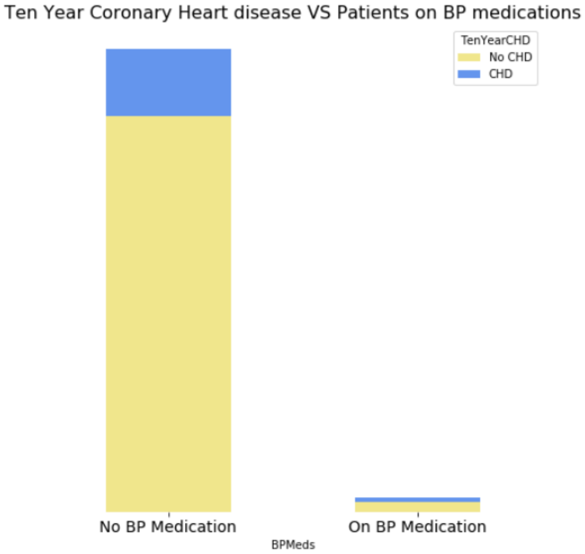
**(B)**

## Hypertension and Blood pressure Medications a Risk Factor for CHD

In 1961, in Framingham Heart study, High Blood pressure is a risk factor for Coronary Heart Disease. **31% of Total Patients have Hypertension**. Out of **31% approximately 24% of Patients have CHD** and **out of remaining 69% who do not have Hypertension only 12% people have a risk of developing CHD**. Hence, we can say that **people with prevalent condition of Hypertension are prone to Coronary Heart diseases.** In 1970, it was determined High Blood Pressure is linked to increased risk of Stroke.

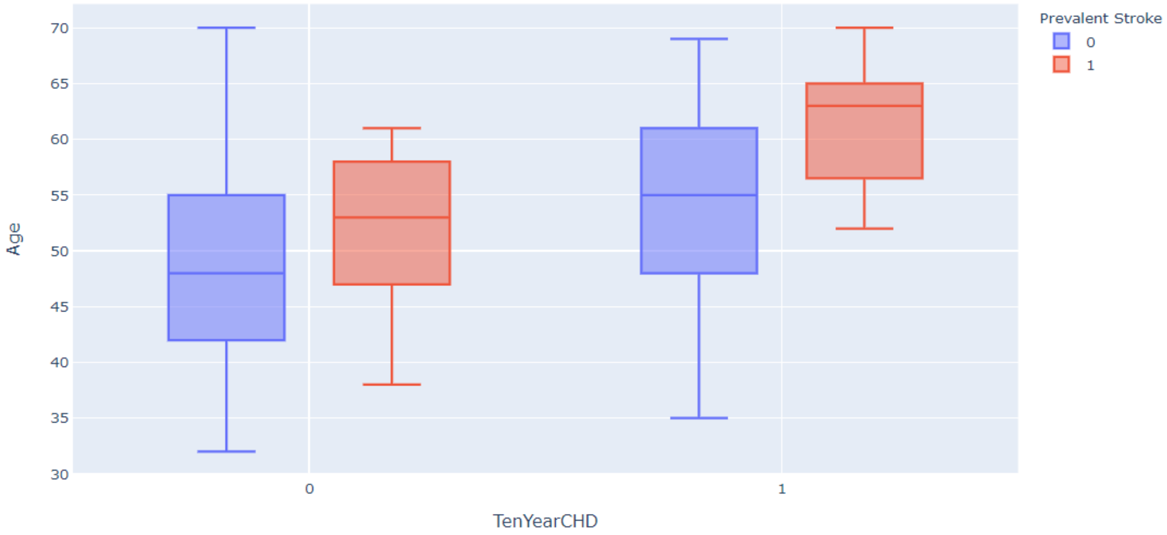


From the below chart, we understand that feature “BPMeds” is imbalanced. However, from available data we learn that people who are on medication for High Blood pressure have a lesser risk of developing Coronary Heart Disease. However, there is no enough data to conclude this information.

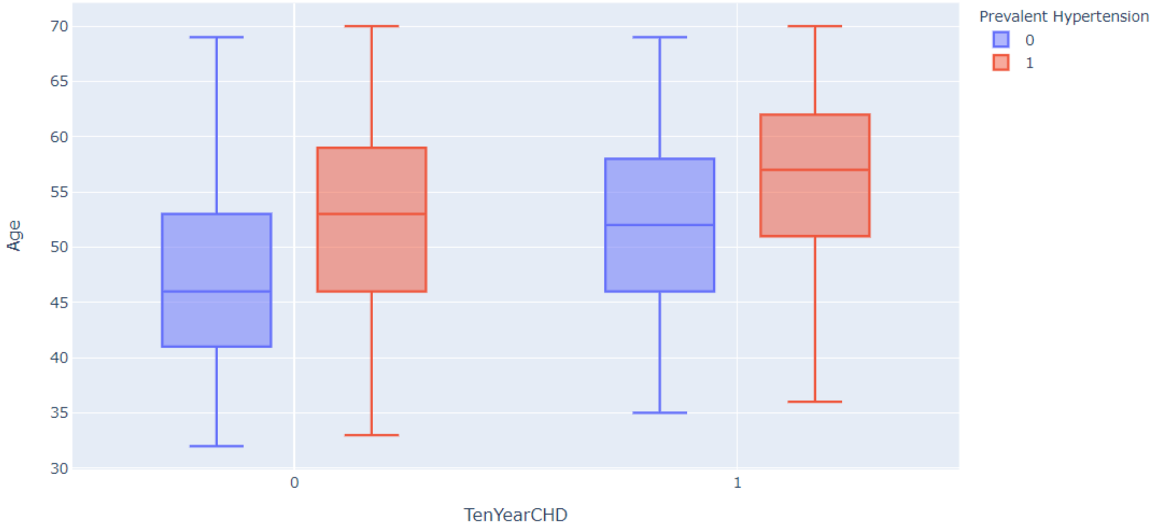


## Age affects Prevalent Hypertension and Stroke

In the Box plots given below, Image **(C)** shows “Prevalent stroke w.r.t. Age and CHD” and Image **(D)** shows “Prevalent Hypertension w.r.t. Age and CHD” respectively. From both the images we can say that Age is an important factor as older people tend to develop Hypertension, Stroke and thus a high risk for Coronary Heart disease.



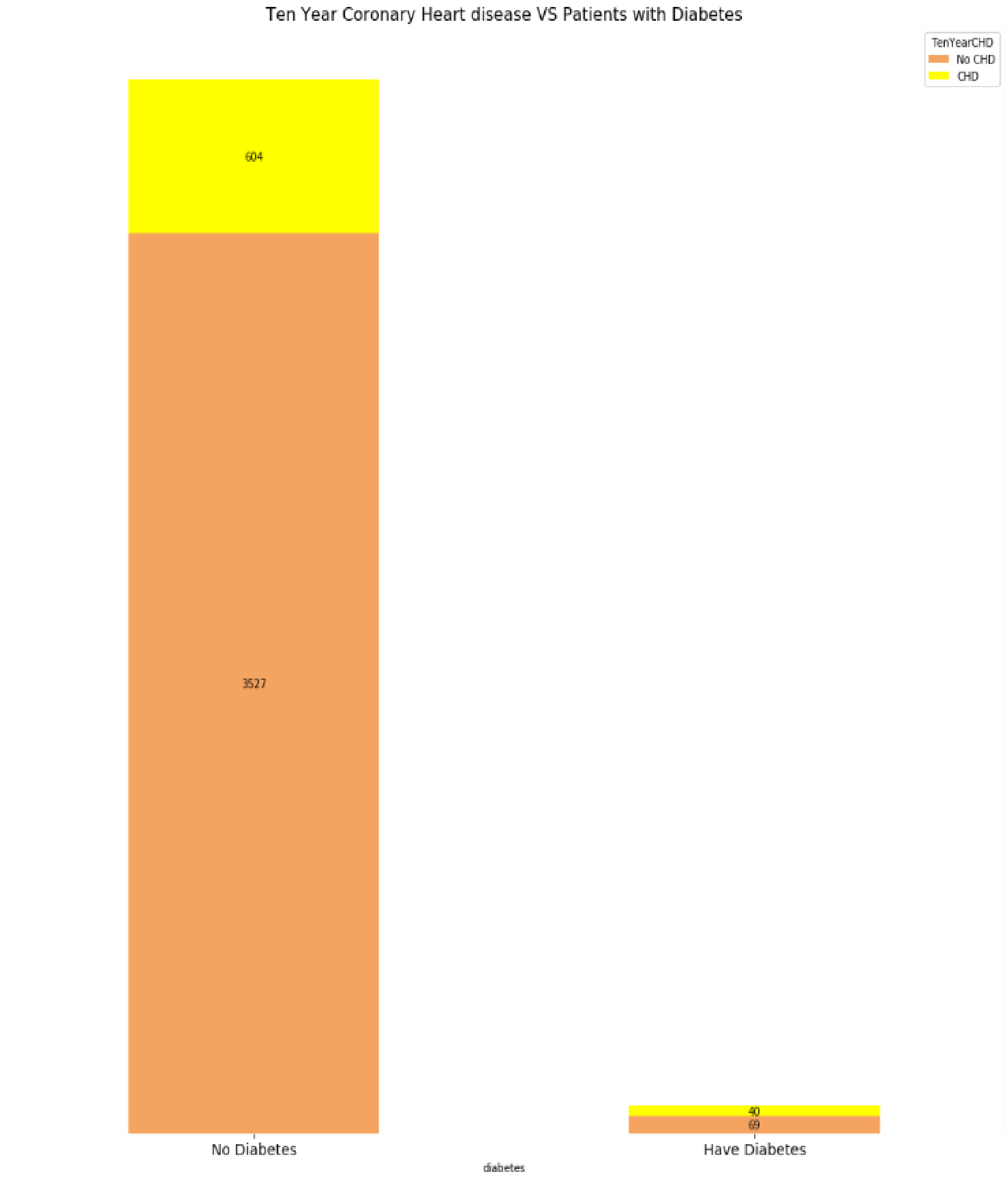
**(C)**



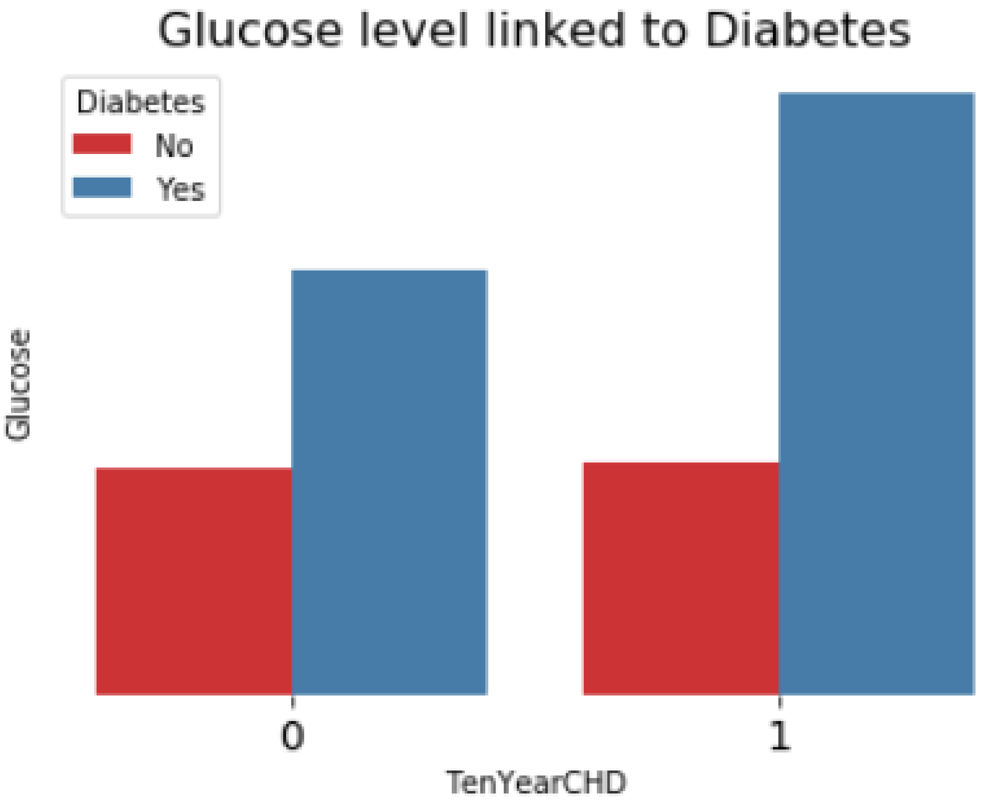
**(D)**

## Diabetes as a risk Factor for CHD

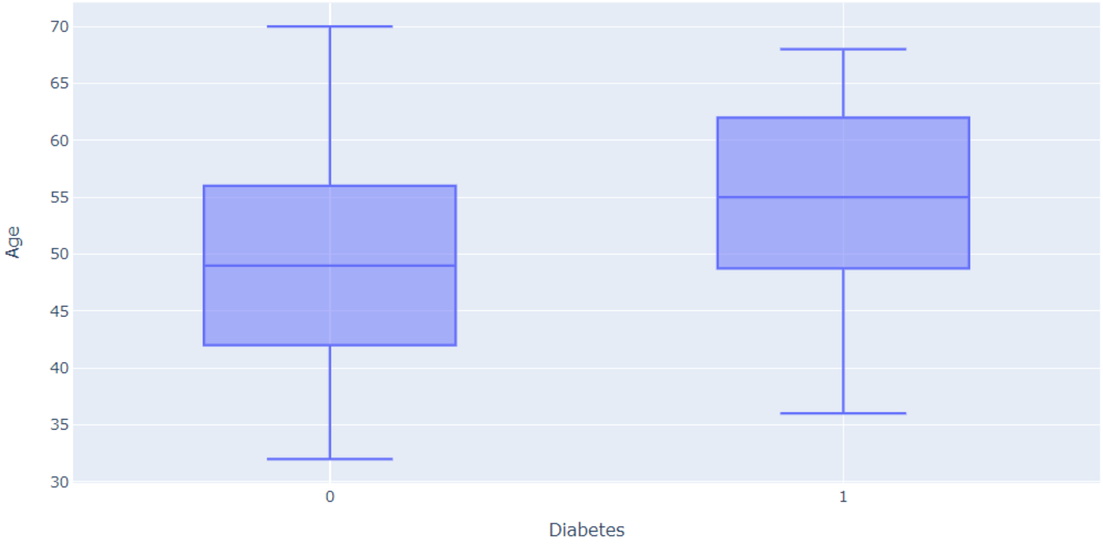
In the given Chart below, from the available data we can observe that only **2.5% of the total people have Diabetes** and within that **2.5% almost 36.4% people have Coronary Heart Disease.** Hence**, Diabetes is linked to Coronary Heart Disease.**



In the given Chart below, we can see that Glucose level is linked Diabetes. As higher the glucose level in blood higher the chances of suffering from Diabetes and thus higher risk of Coronary Heart Disease.

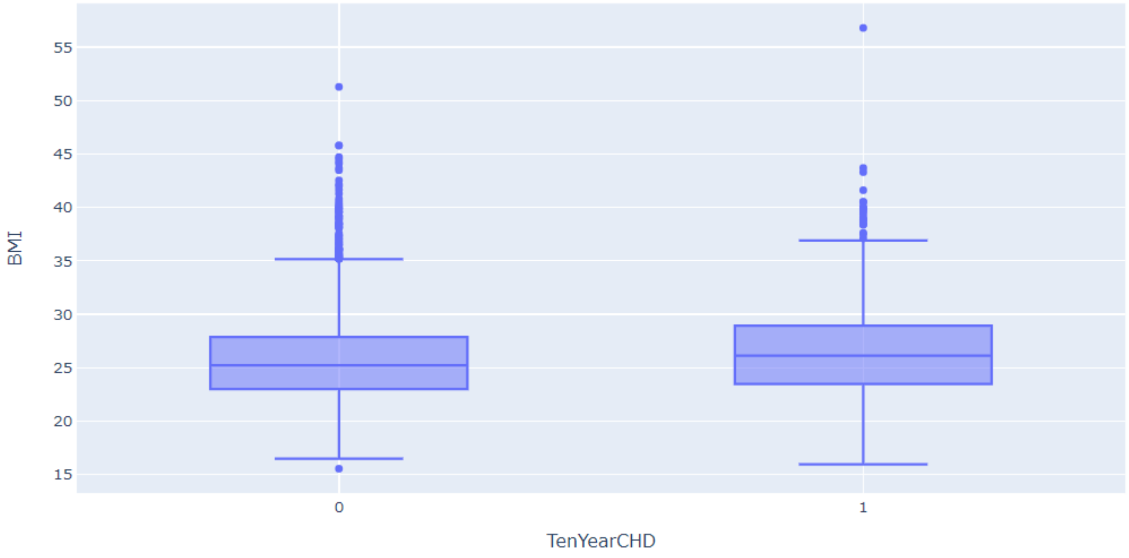


From the Below chart, we can see that Age Factor Contributes to Diabetes as well, as older people have a higher risk of developing Diabetes and in turn a high risk for Coronary Heart Disease.

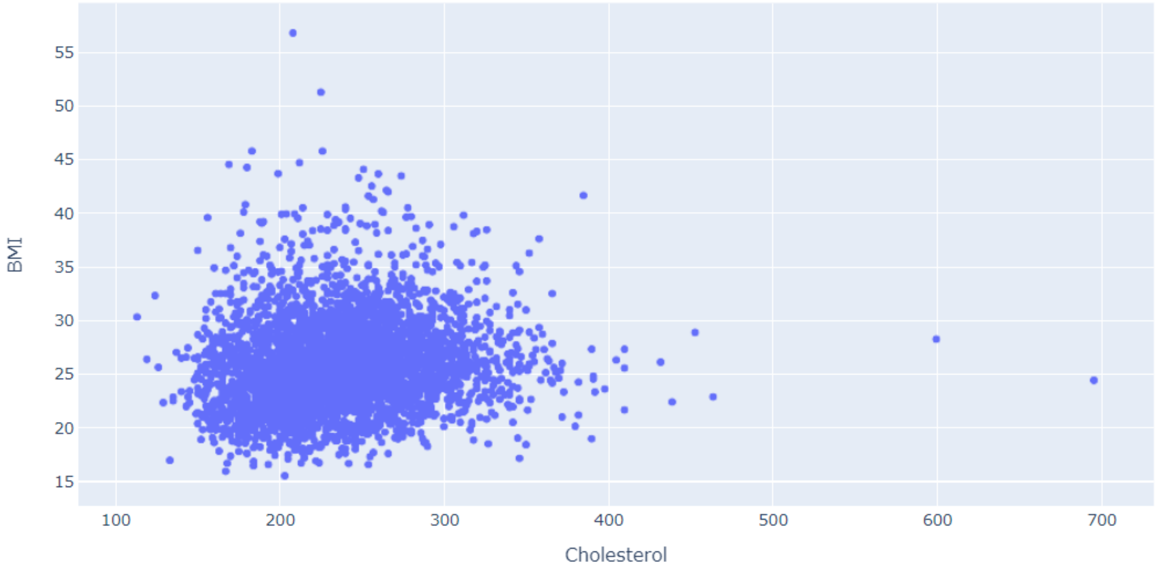


## Body Mass Index(BMI) and Obesity a Risk Factor for Coronary Heart Disease

In the image **(E)** we see a box plot of "Body Mass Index vs Cholesterol" and in image **(F)** we see a scatter plot of "Body Mass Index vs CHD". In **1961,** it was found **High Cholesterol is a risk factor for Coronary Heart disease** also in **1967** it was determined **Obesity and Physical Inactivity increase risk for developing Coronary Heart Disease.**



**(E)**



**(F)**

## Summary

1. Age of a person is a high contributing factor in Coronary heart disease.
2. Men have a higher risk of developing Coronary Heart Disease when compared to Women.
3. Heart disease risk is found to increase in women after Menopause.
4. Cigarette smoking is linked to CHD.
5. Males consume a greater number of Cigarettes in Comparison to Females and have higher risk for developing Coronary Heart Disease.
6. High Blood pressure is a risk for Stroke and Heart Disease.
7. Diabetes is linked to Coronary Heart Disease.
8. Risk of Hypertension, Stroke and Diabetes increases with Age.
9. Obesity and high Cholesterol levels increase risk of Coronary Heart Disease.

## References

[1] https://www.nih.gov/sites/default/files/about-nih/impact/framingham-heart-study.pdf

[2]https://en.wikipedia.org/wiki/Framingham_Heart_Study#:~:text=The%20Framingham%20Heart%20Study%20is,its%20third%20generation%20of%20participants.

[3] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4159698/

[4] https://framinghamheartstudy.org/fhs-for-researchers/data-available-overview/