**Heart Disease**

**Introduction:**

This dataset contains information about several risk factors, such as age, gender, smoking habits, blood pressure, diabetes, and cholesterol levels, for a group of individuals. The target column, TenYearCHD, indicates whether each individual developed coronary heart disease within ten years.

Using this dataset, it is possible to analyze the relationship between risk factors and the development of CHD. By analyzing the correlation between different factors and the development of CHD, researchers can develop preventative treatment plans.

For example, the dataset includes information about smoking habits, which is a significant risk factor for CHD. By analyzing the data related to smoking, researchers can analyze the impact of smoking on the development of CHD and encourage people to help quit smoking.

**Preprocessing the Dataset:**

**Null / Missing Values Handling:**

Many features in this dataset contains null / missing values.

Percentage of Null values in each feature:

male 0.000000%

age 0.000000%

education 2.477584%

currentSmoker 0.000000%

cigsPerDay 0.684285%

BPMeds 1.250590%

prevalentStroke 0.000000%

prevalentHyp 0.000000%

diabetes 0.000000%

totChol 1.179802%

sysBP 0.000000%

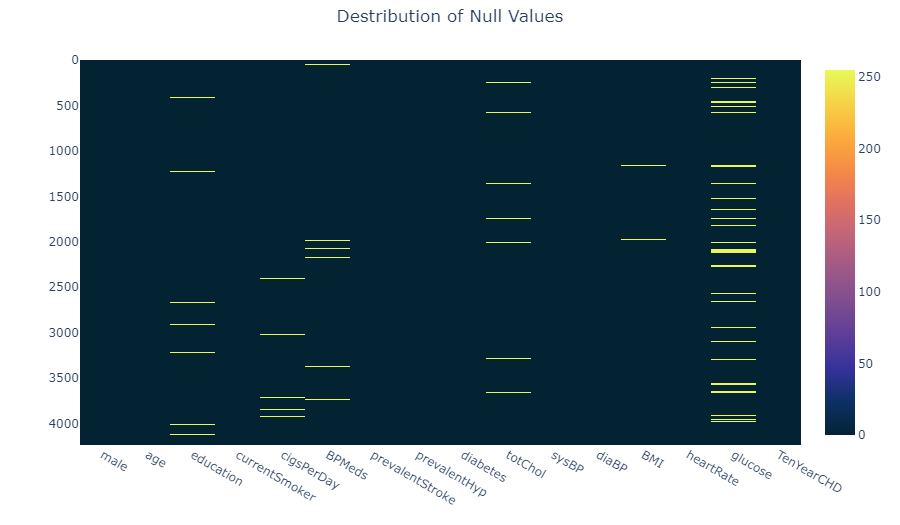
diaBP 0.000000%

BMI 0.448325%

eartrate 0.023596%

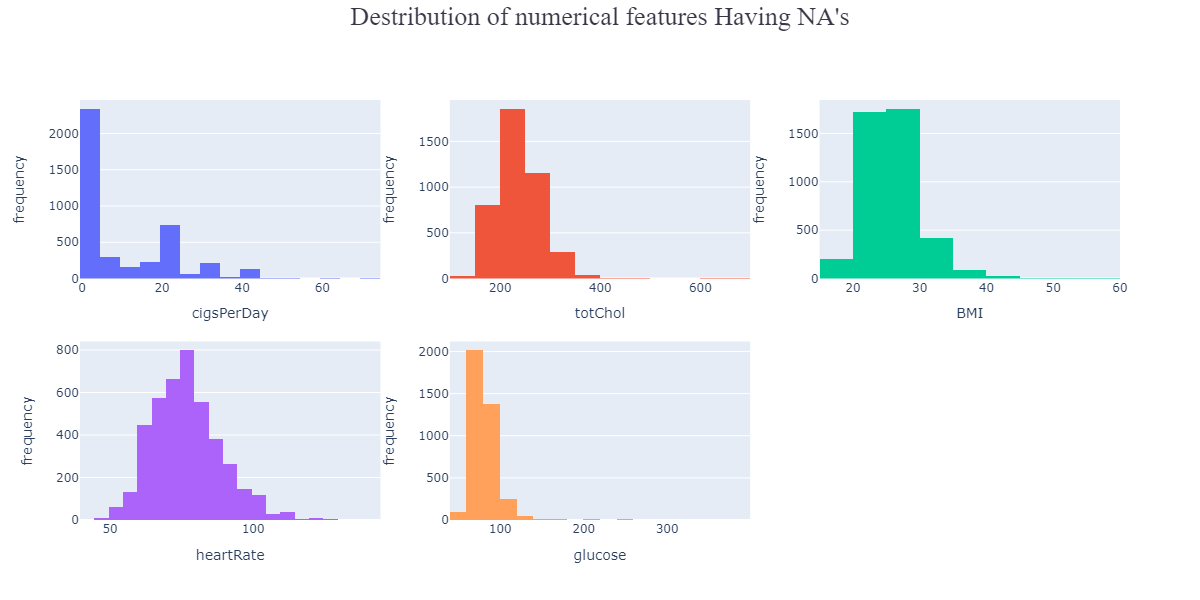
glucose 9.155262%

TenYearCHD 0.000000%

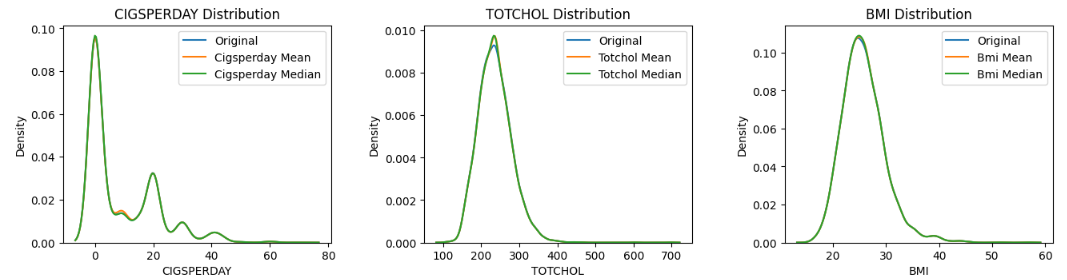


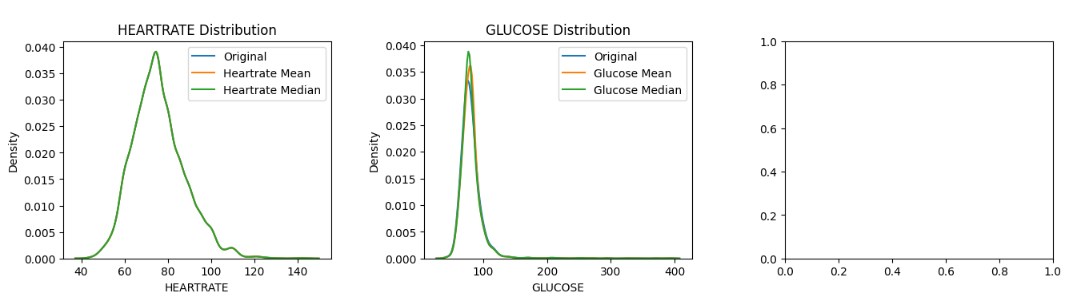
Plotting the Missing Values

Here we can see education, cigsPerDay, BPMeds, totChol, BMI, eartrate, glucose these feature contains null values.



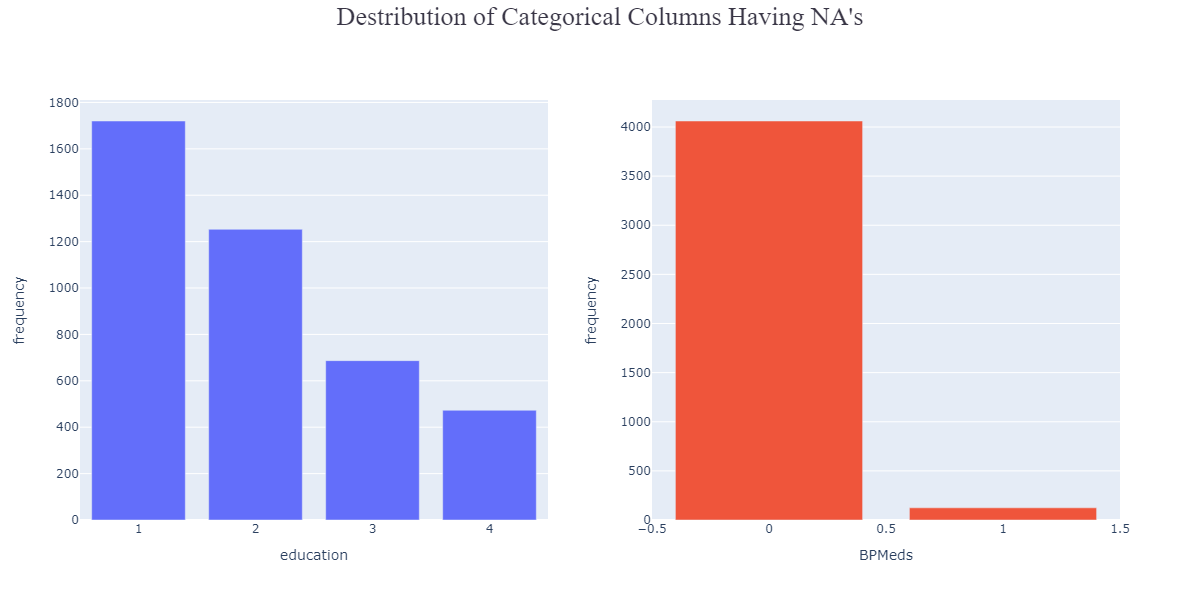
To handle the null values in numerical feature we used both mean and median, and compared them to find out the best possible outcomes.



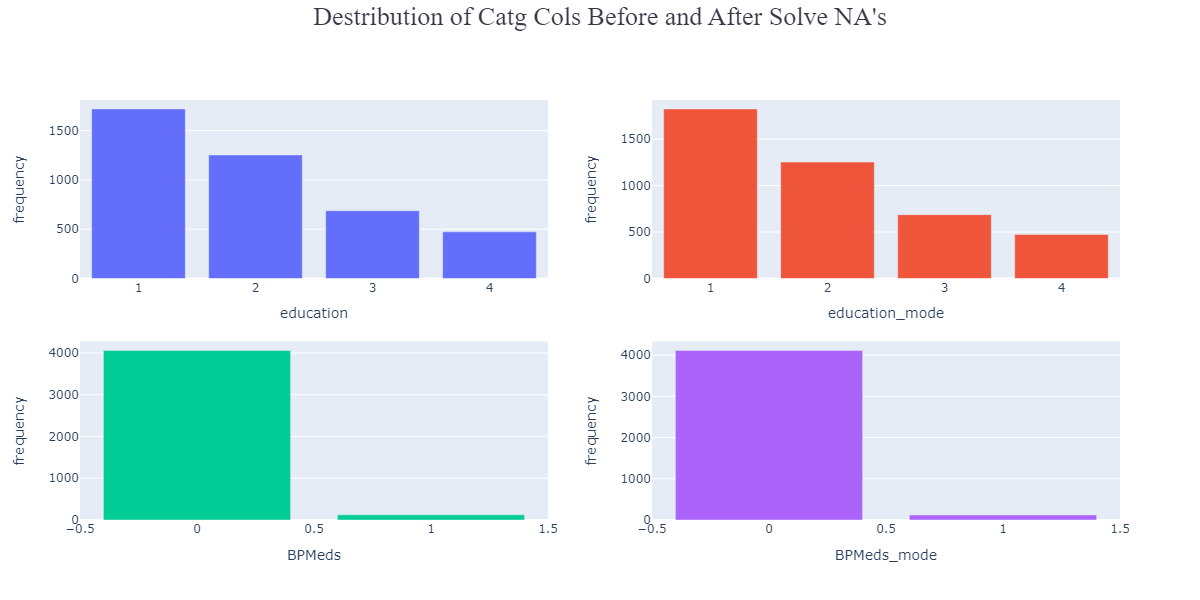


Based on the distribution analysis, it is observed that there are certain columns that demonstrate less variation in distribution after filling the missing values with the mean. These columns include totchol, bmi, heartrate, glucose.

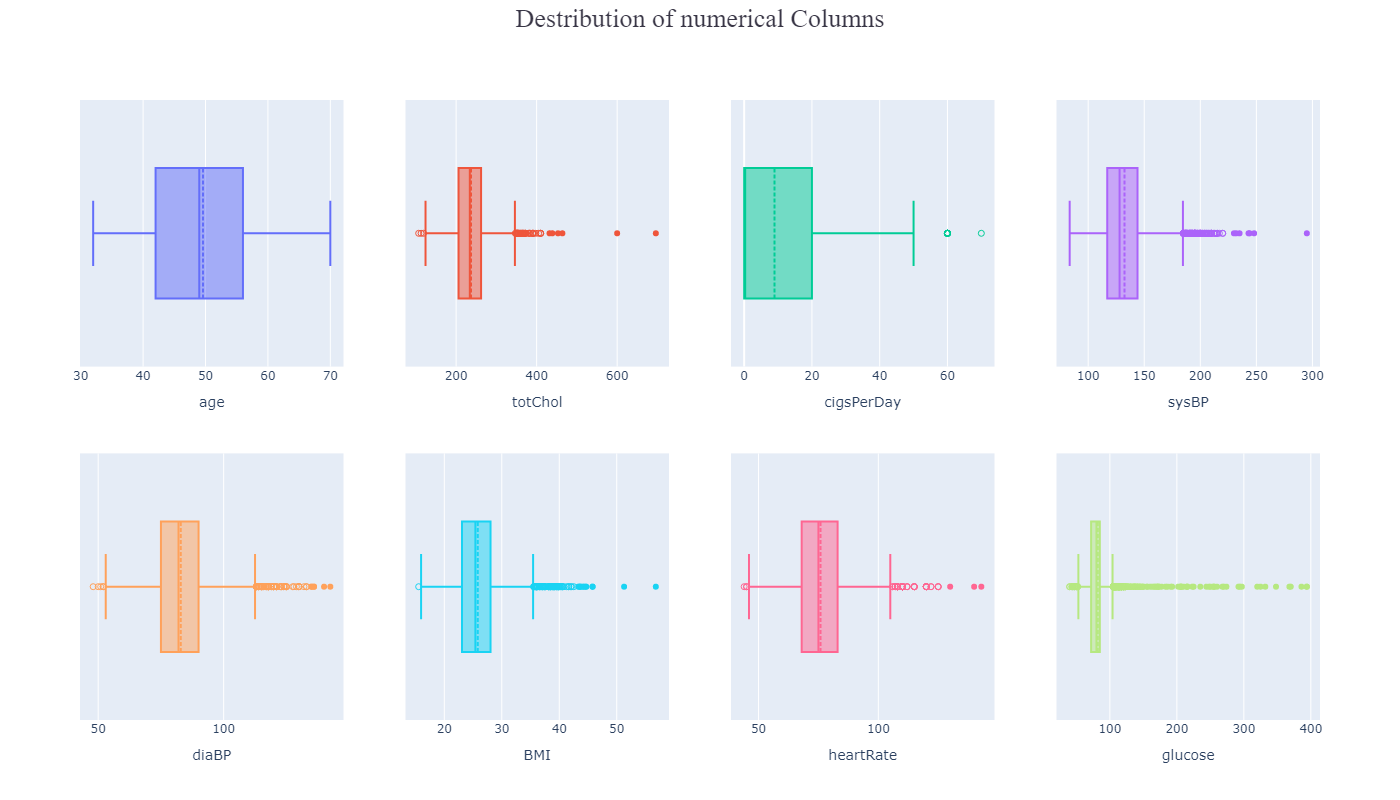
On the other hand, in cigsperday feature the distribution shows significantly less variation after filling the missing values with the median. Therefore, we have to use the median to fill the missing values in these columns.



To handle the null values in categorical feature we used mode to fill those cells up.



**Handling Outliers:**



**Summery Statistics of outliers in each feature:**

Age Percentage: 0.00%

TotChol Percentage: 1.34%

TotChol Mean: 371.23

TotChol Median: 366.00

TotChol Std: 83.46

CigsPerDay Percentage: 0.28%

CigsPerDay Mean: 60.83

CigsPerDay Median: 60.00

CigsPerDay Std: 2.76

SysBP Percentage: 2.97%

SysBP Mean: 199.27

SysBP Median: 195.50

SysBP Std: 15.21

DiaBP Percentage: 1.91%

DiaBP Mean: 116.90

DiaBP Median: 119.00

DiaBP Std: 18.42

BMI Percentage: 2.29%

BMI Mean: 38.97

BMI Median: 38.54

BMI Std: 4.10

HeartRate Percentage: 1.79%

HeartRate Mean: 110.28

HeartRate Median: 110.00

HeartRate Std: 15.00

Glucose Percentage: 6.18%

Glucose Mean: 135.73

Glucose Median: 116.00

Glucose Std: 62.25

From the above analysis we can see feature that have outliers are 'totChol', 'cigsPerDay', 'totChol', 'sysBP', 'diaBP', 'BMI', 'heartRate', 'glucose'.

To handle the outlier we used capping method.



**Hypothesis Test:**

On this dataset hypothesis testing can be performed to determine the statistical significance of the relationship between the target feature TenYearCHD(risk of developing Colonary Heart Disease within the Next Ten Years). By performing Anova(for numerical columns) and Chi-square (for categorical columns) tests it is possible to test hypothesis for this dataset.

If there is no significant difference in the mean ‘TenYearCHD’ across different levels of categorical features in the data then it is Null Hypothesis, and if there is a significant difference then we call it alternative hypothesis.

We can do Anova test for following features:

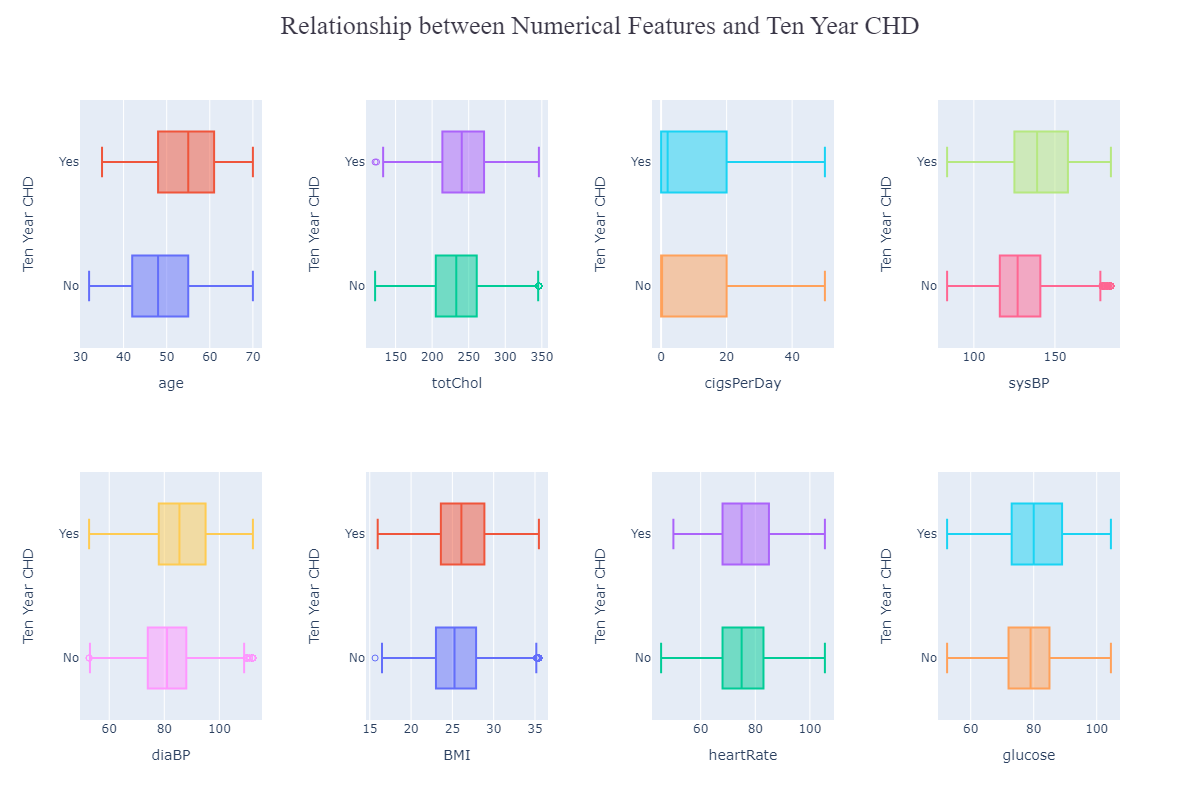
age, totChol, cigsPerDay, sysBP, diaBP, BMI, heartRate

**P-value, Test of Association:**

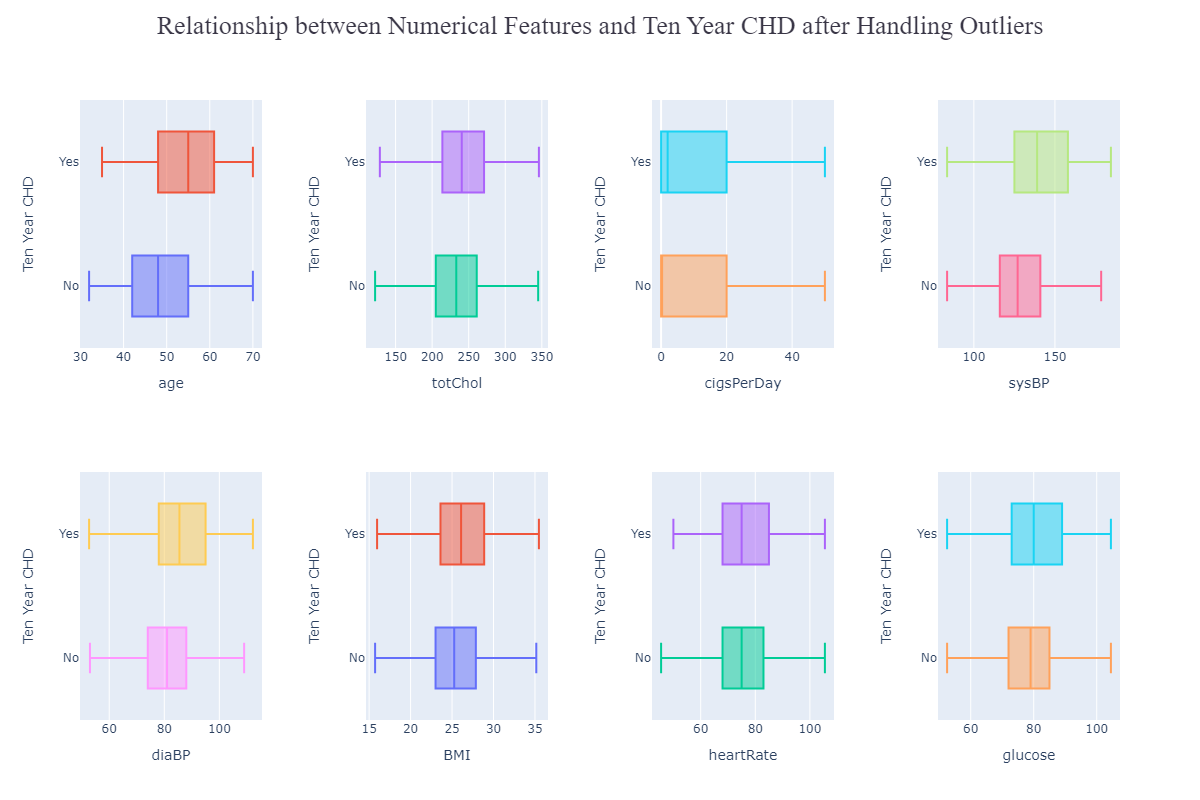
Test of association simply means Chi-square test, which is a statistical test to determine if two categorical features are associated with each other or not. It calculates a p-value to determine the association between variables if the null hypothesis is true or false. Lower p-value means null hypothesis in rejected and alternative hypothesis is accepted.

For Chi-square test we have following features:

male, education, currentSmoker, BPMeds, prevalentStroke, prevalentHyp, diabetes.



From the above distribution we see that feature having outlier are 'totChol’,'sysBP', 'diaBP', 'BMI'. As a result, we used capping method to fill up the outlier with respective upper and lower limit.





**Comparing Our Anova Test function wth Built-in Anova Fuction:**

age

stat = 226.42482952300466, p\_value = 6.8450085874350565e-50

my\_stat = 226.42482952300432, my\_p\_value = 6.8450085874350565e-50

totChol

stat = 26.896520053510475, p\_value = 2.246671581802369e-07

my\_stat = 26.89652005351028, my\_p\_value = 2.246671581802369e-07

cigsPerDay

stat = 15.025334270599298, p\_value = 0.0001076833974413811

my\_stat = 15.025334270599334, my\_p\_value = 0.0001076833974413811

sysBP

stat = 206.41830554811094, p\_value = 9.716367662190807e-46

my\_stat = 206.41830554811006, my\_p\_value = 9.716367662195639e-46

diaBP

stat = 88.71014175059128, p\_value = 7.301881020945315e-21

my\_stat = 88.71014175059173, my\_p\_value = 7.301881020941683e-21

BMI

stat = 22.486967523452538, p\_value = 2.185287570160557e-06

my\_stat = 22.48696752345222, my\_p\_value = 2.185287570160557e-06

heartRate

stat = 2.216695482358314, p\_value = 0.13659944260506415

my\_stat = 2.2166954823592198, my\_p\_value = 0.13659944260509999

**Comparing Our Chi-square Function with Built-in Chi-square Function:**

male

stat = 33.13876412178594, p\_value = 8.581080129462392e-09

my\_stat = 33.13876412178594, my\_p\_value = 8.581080179759226e-09

education

stat = 30.93552881168504, p\_value = 8.770368701361081e-07

my\_stat = 30.935528811685035, my\_p\_value = 8.770368701283004e-07

currentSmoker

stat = 1.6042792098599596, p\_value = 0.20529783701529147

my\_stat = 1.6042792098599596, my\_p\_value = 0.20529783701529503

BPMeds

stat = 31.649053344312318, p\_value = 1.8470461189725918e-08

my\_stat = 31.649053344312318, my\_p\_value = 1.8470461160546847e-08

prevalentStroke

stat = 16.191149461946694, p\_value = 5.726102177501766e-05

my\_stat = 16.191149461946694, my\_p\_value = 5.7261021775056875e-05

prevalentHyp

stat = 133.6780899180762, p\_value = 6.425269735903649e-31

my\_stat = 133.6780899180762, my\_p\_value = 0.0

diabetes

stat = 40.135995000258504, p\_value = 2.368836764784068e-10

my\_stat = 40.135995000258504, my\_p\_value = 2.368836238275662e-10

**Feature Selection Based on P-value:**

prevalentHyp: 0.0

age: 6.8450085874350565e-50

sysBP: 9.716367662190807e-46

diaBP: 7.301881020945315e-21

diabetes: 2.368836238275662e-10

male: 8.581080179759226e-09

BPMeds: 1.8470461160546847e-08

totChol: 2.246671581802369e-07

education: 8.770368701283004e-07

BMI: 2.185287570160557e-06

prevalentStroke: 5.7261021775056875e-05

cigsPerDay: 0.0001076833974413811

heartRate: 0.13659944260506415

currentSmoker: 0.20529783701529503

The null hypothesis for each p-value is that there is no relationship between the feature and TenYearCHD. Therefore, if the p-value is less than the significance level (usually 0.05), we can reject the null hypothesis and conclude that there is a significant relationship between the feature and TenYearCHD.

From the above p-values, we can see that all the features except 'heartRate' and 'currentSmoker' have p-values less than 0.05. This means that these features are statistically significant and have a significant relationship with TenYearCHD.

Therefore, the significant features based on the given p-values are:

prevalentHyp, age, sysBP, diaBP, diabetes, male, BPMeds, totChol, education, BMI, prevalentStroke, cigsPerDay.

We can formally state that there is a statistically significant relationship between TenYearCHD and the above mentioned features at a significance level of 0.05.

**Relevant Exploratory Data Analysis:**

1. What is the relationship between total cholesterol levels and systolic blood pressure, and how does this impact the risk of developing heart disease?

A picture containing text, screenshot, colorfulness, majorelle blue

Description automatically generated

The correlation value of 0.2132 indicates a positive correlation between total cholesterol levels and systolic blood pressure. This means that as the levels of total cholesterol increase, the systolic blood pressure tends to increase as well. However, the strength of this correlation is only moderate, suggesting that other factors may also contribute to variations in blood pressure and cholesterol levels. Furthermore, it is important to note that correlation does not imply causation, and further investigation would be needed to establish a causal relationship between these variables and the risk of developing heart disease.

2. Is there a correlation between BMI and heart rate, and does this vary between smokers and non-smokers?

A picture containing text, screenshot, colorfulness, line

Description automatically generated

The correlation coefficient of 0.06 suggests a weak positive correlation between BMI and heart rate. However, this correlation is not statistically significant or practically meaningful.

3. What is the distribution of glucose levels among patients with and without diabetes, and does this have any impact on the risk of developing heart disease?

A screenshot of a graph

Description automatically generated with low confidence

The distribution of glucose levels among patients with and without diabetes reveals distinct patterns. Patients with diabetes exhibit higher glucose levels, with the majority exceeding 100. In contrast, patients without diabetes show a relatively normal distribution of glucose levels, centered around 80 or above. This shows that those who have diabetes are more likely to have increased glucose levels, which may also be a sign of the disease. Additionally, the finding that people without diabetes also have a considerable percentage of blood sugar levels that are above 80 suggests that there may be a link between blood sugar levels and an increased risk of developing heart disease.

4. How does the average number of cigarettes smoked per day vary across different age groups in the dataset?

A picture containing text, diagram, screenshot, plot

Description automatically generated

The distribution of the number of cigarettes smoked per day among different age groups reveals interesting patterns. Among individuals aged up to 50 years, there is a notable presence of heavy smokers, with some individuals reporting up to 50 cigarettes per day. However, as the age increases to 60 years and above, the number of heavy smokers decreases significantly. This suggests that older individuals tend to smoke fewer cigarettes per day on average. Furthermore, the distribution of cigarette consumption among individuals below 40 years appears to follow a relatively normal pattern.

5. Is there a significant difference in the average BMI between patients who do and do not take blood pressure medication?

A screenshot of a graph

Description automatically generated with low confidence

The t-statistic is 5.59, which indicates that there is a significant difference in the mean BMI between patients who take blood pressure medication and those who do not. The p-value is very small (1.29e-07), which indicates strong evidence against the null hypothesis that there is no difference in mean BMI between the two groups. Therefore, we can reject the null hypothesis and conclude that there is a significant difference in the average BMI between patients who take blood pressure medication and those who do not. Specifically, it appears that patients who take blood pressure medication have a higher mean BMI compared to those who do not.

6. What is the relationship between age and the total cholesterol levels of patients?

A picture containing text, screenshot, line, diagram

Description automatically generated

Since the Correlation coefficient between age and total cholesterol levels is 0.27 so we can say that there is no relation between age and total cholesterol levels.

7. Is there a significant difference in systolic blood pressure between male and female patients?

After using student t-test we found the following answer,

t-statistic: -1.81, p-value: 0.07089

The t-test was conducted to determine if there is a significant difference in systolic blood pressure between male and female patients. The negative t-statistic value (-1.81) suggests that the mean systolic blood pressure is lower in female patients than in male patients. The p-value of 0.07089 is greater then the significance level of 0.05, indicating no statistically significant difference between male and female patients.

8. What is the average heart rate of patients with diabetes compared to those without diabetes? Is there a significant difference between them?

Average Heart rate patients having diabetes: 79.28

Average Heart rate patients not having diabetes: 75.66

After using student t-test we found the following answer,

t-statistic: 3.21, p-value: 0.00131

The p-value of 0.00131 is less than the common threshold for statistical significance of 0.05, indicating that we can reject the null hypothesis that there is no difference in the average heart rate between the two groups. Therefore, we can conclude that patients with diabetes have a significantly different average heart rate compared to those without diabetes.

9. Is there a relationship between the number of cigarettes smoked per day and systolic blood pressure?

A picture containing text, screenshot, line, diagram

Description automatically generated

The correlation coefficient between the number of cigarettes smoked per day and systolic blood pressure was found to be -0.09. This suggests a weak negative correlation between the two variables, indicating that as the number of cigarettes smoked per day increases, systolic blood pressure tends to slightly decrease. However, the correlation coefficient value is relatively small, indicating that the relationship between the two variables is not very strong. Therefore, while there is some association between cigarette smoking and systolic blood pressure, the effect size is weak.

10.Destribution of age in the dataset in terms of Gender

A picture containing screenshot, text, diagram, plot

Description automatically generated

From the above plot we can see that the destribution is rightly skewed and the majority of the people are between the age 38 - 43. Also based on the group we can analyze that majority number of people are female.

11. Distribution of education levels Inerms of Gender

A picture containing text, screenshot, rectangle, diagram

Description automatically generated

The distribution of education levels indicates that the majority of individuals in the dataset have an education rank of 1. Additionally, among all the education categories, category 4 stands out as having a higher proportion of males compared to females.

This observation suggests that there is an imbalance in the distribution of education levels, with a significant number of individuals falling into the education rank 1 category.

12. Destribution of smokers interms of diabetes

A picture containing text, screenshot, rectangle, diagram

Description automatically generated

This count distribution indicates a gender disparity in smoking behavior within the dataset, with a higher proportion of females abstaining from smoking compared to males. However, among smokers, the difference in the count between males and females is less pronounced, suggesting a relatively comparable proportion of male and female smokers.

13. Is there a significant difference in the proportion of smokers between males and females?

Built in : 165.47002460557053 7.221626137546635e-38

Ours : 165.47002460557053 0.0

The chi-square test for association yielded a test statistic of 165.47 and an extremely low p-value of 7.22e-38. Based on these results, we can conclude that there is a significant association between gender and smoking status in the dataset.

The low p-value indicates strong evidence against the null hypothesis, suggesting that the proportion of smokers differs significantly between males and females. Therefore, we reject the null hypothesis and accept the alternative hypothesis, which suggests a notable difference in smoking prevalence based on gender.

14. distribution of the target variable (TenYearCHD)

A picture containing diagram, circle, screenshot, font

Description automatically generated

The distribution of gender among individuals with TenYearCHD (Ten-Year Risk of Coronary Heart Disease) shows that approximately 46.7% are female and 53.3% are male. On the other hand, the distribution of gender among individuals without TenYearCHD indicates that approximately 58.9% are female and 41.1% are male.

This suggests that there may be a difference in the gender distribution between those with and without TenYearCHD. Females seem to be more prevalent in both groups, but the difference is more pronounced among individuals without TenYearCHD.

15. Relationship between age and cigsperday.

A picture containing text, screenshot, line, font

Description automatically generated

From the above distribution we can conclude that as the age increases, there tends to be a slight decrease in the number of cigarettes smoked per day. However, the correlation is weak, suggesting that the relationship between age and cigarette consumption is not strongly linear.

16. Variation Heart Rate across AGE groups

A picture containing text, screenshot, line, colorfulness

Description automatically generated

The correlation coefficient between age and heart rate in the dataset is -0.01. This indicates a very weak or negligible linear relationship between these variables. The correlation coefficient being close to zero suggests that there is no substantial linear association or dependence between age and heart rate.

**Discussion:**

**Conclusion:**

**Reference:**