West Bengal State University

Bidhannagar College

Investigating the Impact of Medical Factors on Heart Failure

Registration Number - 1082011400148
Roll - 6231125
Number - 14768
College Roll Number - 20686
Session - 2020-2023

Investigating the Impact of Medical Factors on Heart Failure

Contents

Introduction 2
Executive Summary 2
Objective 2
Hypothesis 2
Data Sources 3
Attribute Information 3
Methodology 5
Analysis and Interpretation 6
Conclusion 17
Appendix 1 19
Appendix 2 24
Appendix 3 32
Gratitude 36

Introduction

Heart failure is a complex and prevalent medical condition that affects individuals of all ages, posing significant challenges for healthcare professionals worldwide. To enhance our understanding of heart failure and its underlying causes, this project aims to analyse the relationship between heart failure and various medical factors such as blood sugar level, blood pressure, cholesterol, and other relevant parameters across different age groups.

Executive Summary

This project aimed to analyze a dataset related to heart disease and develop a predictive model to identify risk factors associated with heart conditions. The dataset included various medical attributes such as age, sex, chest pain type, blood pressure, cholesterol levels, fasting blood sugar, electrocardiogram results, maximum heart rate, exercise-induced angina, old peak, and ST segment slope. After preprocessing the data and splitting it based on gender, exploratory data analysis provided insights into the distribution of variables. Barplots and histograms were utilized to visualize categorical and continuous variables, respectively.

Subsequently, a logistic regression model was built to predict the likelihood of heart disease. The significant predictors identified were sex, chest pain type, fasting blood sugar, exercise-induced angina, and old peak. The model's findings provide valuable insights into the factors contributing to heart disease and can aid in making informed decisions for prevention and early detection. However, further research with larger and diverse datasets is recommended to enhance the model's performance and accuracy, leading to more robust predictions for heart disease.

Objective

The primary objective of this project is to explore the influence of these medical factors on the occurrence and severity of heart failure in distinct age cohorts. By leveraging advanced data analytics techniques, we can uncover patterns, trends, and potential risk factors associated with heart failure, thus facilitating early detection, prevention, and personalised intervention strategies.

By conducting this project, we anticipate uncovering valuable insights regarding the relationship between heart failure and medical factors across different age groups. The project's outcomes will contribute to advancing the understanding of heart failure, facilitating early detection and prevention strategies, and optimizing patient management approaches. Furthermore, the project will provide a foundation for future research endeavors, focusing on personalized medicine and improving overall cardiovascular health outcomes.

Hypothesis

Research Hypothesis 1

H_0 : The variables *Heart Failure* and *Sex* are independent.
H_1 : The variables *Heart Failure* and *Sex* are not independent.

Research Hypothesis 2

H_0 : The variables *Heart Failure* and *Chest Pain* are independent.
H_1 : The variables *Heart Failure* and *Chest Pain* are not independent.

Research Hypothesis 3

H_0 : The variables *Heart Failure* and *Fasting Blood Suga*r are independent.
H_1 : The variables *Heart Failure* and Fasting *Blood Sugar* are not independent.

Research Hypothesis 4

H_0 : The variables *Heart Failure* and *Resting ECG* are independent.
H_1 : The variables *Heart Failure* and *Resting ECG* are not independent.

Research Hypothesis 5

H_0 : The variables *Heart Failure* and *Exercise Angina* are independent.
H_1 : The variables *Heart Failure* and *Exercise Angina* are not independent.

Research Hypothesis 6

H_0 : The variables *Heart Failure* and *ST Slope* are independent.
H_1 : The variables *Heart Failure* and *ST Slope* are not independent.

Research Hypothesis 7

H_0 : The variables *Heart Failure* is not affected by other variables like *Sex, Chest Pain, etc.*
H_1 : The variables *Heart Failure* is affected by other variables like *Sex, Chest Pain, etc.*

Data Sources

To accomplish our research objectives, we will utilize a diverse range of comprehensive datasets encompassing medical records, clinical measurements, and patient demographics. These datasets will be collected from reputable healthcare institutions, public health databases, and clinical research studies. The data will be carefully curated, anonymized, and transformed to maintain patient privacy and comply with relevant ethical guidelines.

Attribute Information

1. ***Age:*** Age of the patient in years. Age is an important risk factor for cardiovascular diseases, including heart disease. As individuals age, the risk of developing heart disease tends to increase.

2. ***Sex:*** Sex of the patient (M: Male, F: Female). Gender plays a role in the incidence and manifestation of heart disease. Males have generally been found to have a higher risk of heart disease compared to females, although the risk may vary based on other factors as well.

3. ***ChestPainType:*** Type of chest pain reported by the patient. It includes the following categories:
   - TA (Typical Angina)**:** This type of chest pain is typically described as a squeezing or pressure-like pain in the chest. It is often triggered by physical exertion or emotional stress and tends to subside with rest or medication.
   - ATA (Atypical Angina): Atypical angina refers to chest pain that does not fit the typical pattern of angina. The pain may be different in nature, location, or duration compared to typical angina.
   - NAP (Non-Anginal Pain): Non-anginal pain refers to chest discomfort or pain that is not caused by coronary artery disease. It may be due to other causes such as musculoskeletal issues, gastrointestinal problems, or anxiety.
   - ASY (Asymptomatic): Asymptomatic means the absence of any symptoms. In this case, it indicates the absence of chest pain or discomfort.

4. ***RestingBP:*** Resting blood pressure measured in millimeters of mercury (mm Hg). Blood pressure is an important indicator of cardiovascular health. Elevated resting blood pressure is a risk factor for heart disease and other cardiovascular conditions.

5. ***Cholesterol:*** Serum cholesterol levels measured in milligrams per deciliter (mm/dl). Cholesterol is a type of fat present in the bloodstream. Elevated levels of cholesterol, particularly low-density lipoprotein (LDL) cholesterol, are associated with an increased risk of developing atherosclerosis and heart disease.

6. ***FastingBS:*** Fasting blood sugar level. It is used to assess glucose metabolism and identify individuals at risk of diabetes. Elevated fasting blood sugar levels can indicate impaired glucose regulation and an increased risk of developing diabetes, which is a risk factor for heart disease.

7. ***RestingECG:*** Resting electrocardiogram results. It provides information about the electrical activity of the heart at rest. The different categories include:
   - Normal: Indicates a normal ECG reading without any significant abnormalities.
   - ST (ST-T wave abnormality): Presence of ST-T wave abnormalities, such as T wave inversions and ST segment elevation or depression, which can be indicative of ischemia or other cardiac conditions.
   - LVH (Left Ventricular Hypertrophy): Suggests an enlargement or thickening of the left ventricle of the heart, which may be a sign of underlying heart disease or hypertension.

8. ***MaxHR:*** Maximum heart rate achieved during exercise. It is a measure of cardiovascular fitness and can provide insights into the heart's functional capacity.

9. ***ExerciseAngina:*** Exercise-induced angina refers to the occurrence of chest pain or discomfort during physical exertion or exercise. It is typically a symptom of underlying coronary artery disease and can indicate reduced blood flow to the heart during exercise.

10. ***Oldpeak:*** ST segment depression measured during exercise. ST depression is a common finding in individuals with coronary artery disease and can be indicative of reduced blood supply to the heart during exercise.

11. ***ST_Slope:*** The slope of the peak exercise ST segment. It describes the shape of the ST segment on the electrocardiogram during exercise. The categories include:

- Up: Indicates an upward-sloping ST segment during exercise, which is considered a normal finding.
- Flat: Represents a flat ST segment during exercise, which may suggest ischemia or reduced blood flow to the heart.
- Down: Indicates a downward-sloping ST segment during exercise, which is also associated with ischemia or reduced blood flow to the heart.

12. ***HeartDisease***: The output class indicates whether the individual has heart disease (1) or is classified as normal (0) based on the presence or absence of relevant risk factors and diagnostic findings.

Methodology

Our project will adopt a systematic and rigorous approach to analyze the relationship between heart failure and various medical factors. The following Statistical methods will be used to do the analysis:

- ***Chi Square Test of Independence*** - The chi-square test of independence is a statistical method used to determine if there is a significant association between two categorical variables in a contingency table. It is suitable for qualitative variables because it assesses whether the observed frequencies in the contingency table differ significantly from what would be expected if the variables were independent of each other.
  Performing a chi-square test of independence on the variable "heart failure" and other categorical variables in your dataset is crucial for understanding the potential relationships and dependencies between these variables. By conducting this statistical test, you can determine if there are significant associations between heart failure and other factors, such as age, sex, chest pain type, fasting blood sugar, resting electrocardiogram results, exercise-induced angina, and the slope of the peak exercise ST segment. Identifying these associations can provide valuable insights into the risk factors and patterns that may contribute to heart failure, allowing for more informed decision-making in healthcare and potential interventions to prevent or manage heart-related conditions. Additionally, the chi-square test helps reveal whether any of the categorical variables are potentially significant predictors of heart failure, aiding in the development of predictive models and tailored treatment strategies for patients at risk.

- ***Summary of Continuous Variables -*** The summary of continuous variables is a fundamental step in analyzing and understanding quantitative data. It provides key descriptive statistics such as measures of central tendency (mean, median), dispersion (minimum, maximum, quartiles), and sometimes additional information like skewness and kurtosis. This summary is important as it allows researchers to gain insights into the distribution, range, and typical values of continuous variables within a dataset. By examining these summary statistics, researchers can identify patterns, assess the variability, and understand the overall characteristics of the data. This information is crucial for making informed decisions, identifying outliers, detecting potential issues with the data, and formulating appropriate statistical models or hypotheses. Furthermore, the summary of continuous variables enables comparisons between different groups or subgroups, providing a basis for understanding relationships and making meaningful interpretations in various research domains.

- ***KNN Method of Imputation -*** K-nearest neighbors (KNN) imputation is a technique used to fill in missing values in a dataset based on the values of its nearest neighbors. It is a non-parametric approach that estimates the missing values by taking into account the similarity between the observation with missing values and its neighboring observations.
  The KNN imputation method follows these steps:
  Identify the observations with missing values.
  1. Calculate the distance between the observation with missing values and all other observations in the dataset.
  2. Select the K nearest neighbors based on the calculated distances.
  3. For each missing value, use the values from its K nearest neighbors to impute the missing value.
  4. The imputation can be done by taking the mean, median, or mode of the values from the nearest neighbors.
  5. Repeat steps 2-4 for all observations with missing values.

The choice of K (the number of nearest neighbors) is crucial in KNN imputation. A larger K value considers a larger number of neighbors, which can provide a more accurate imputation but may also introduce more noise. A smaller K value limits the number of neighbors, which may result in a less accurate imputation but potentially reduces noise.

KNN imputation is effective when the missing values are assumed to be related to nearby observations. It can handle both numerical and categorical variables and does not assume any specific distribution of the data. However, it relies on the assumption that similar observations have similar values, which may not always hold true.

- ***Generalised Linear Model -*** The Generalized Linear Model (GLM) is well-suited for this dataset as it enables the analysis of binary outcomes, such as the presence or absence of heart failure, and accommodates both categorical (e.g., sex, chest pain type) and continuous predictors (e.g., age, cholesterol). GLM extends linear regression to handle various response distributions, making it applicable to non-normally distributed data commonly encountered in medical studies.

  By employing appropriate link functions and response distributions, GLM allows for the modeling of nonlinear relationships between predictors and the likelihood of heart failure. This flexibility enables a more accurate representation of the underlying patterns and potential risk factors associated with the outcome. Hence, GLM provides a robust and versatile approach to uncover valuable insights from this dataset, helping to identify significant predictors and understand their impact on the occurrence of heart failure.

Analysis and Interpretation

The analysis of the dataset, as represented by the R code in *Appendix 1* and *Graph 1* in *Appendix 3*, reveals that approximately 55.34% of the individuals in the dataset have experienced heart failure. This finding suggests that the dataset contains a substantial representation of both heart failure and non-heart failure data points, with heart failure cases being slightly more prevalent.

Graph 2 in *Appendix 3* reveals a notable disparity between the number of male and female observations, with a considerably higher count of males compared to females. This discrepancy indicates a potential sampling error or imbalance in the dataset, which could introduce bias in our subsequent analyses and results. By employing the R code in Appendix 1, we calculate the percentage of males and females in the dataset, yielding approximately 78.98% males and 21.02% females. This further highlights the prevailing gender imbalance and underscores the importance of acknowledging and accounting for such biases during the data analysis process.

From the analysis of *Graph 2*, it is evident that there is a significant sampling error, with a considerably higher number of males in our dataset compared to females. To investigate the impact of sex on heart failure, we split the dataset using the code in *Appendix 1* and then plotted the frequency of heart failure in each group (*Graph 3 & Graph 4* in A*ppendix 3)*. Among males, we observed a higher heart failure frequency, with 63.17% of males experiencing heart failure, while the remaining do not. However, the situation is different for females, as the heart failure rate is lower. Approximately 25.91% of females in our dataset have heart failure, and the majority do not experience heart failure. This indicates that sex may have an important role in influencing heart failure risk, with males having a higher likelihood of heart failure in our dataset compared to females.

The analysis of *Graph 5* in *Appendix 3* reveals that the majority of observations in the dataset are categorized as Asymptomatic, with a significantly higher proportion compared to other chest pain types. The second most common type is Non-anginal pain, followed by Atypical Angina, and the least frequent type is Typical angina. By quantifying each category using the code in *Appendix 1*, we find that approximately 54.03% of observations are classified as Asymptomatic, 18.85% as Atypical Angina, 22.11% as Non-anginal pain, and only 5.01% as Typical angina. This distribution highlights the prevalence of Asymptomatic cases in the dataset and provides valuable insights into the distribution of chest pain types among the patients studied.

Graph 6 in *Appendix 3* reveals that a majority of observations fall into the category 0, which corresponds to people with FastingBS levels less than or equal to 120 mg/dl. In contrast, the number of observations in category 1 (FastingBS greater than 120 mg/dl) is relatively smaller. By utilizing the code provided in *Appendix 1*, we find that approximately 76.69% of the observations are in category 0, while 23.31% of the observations belong to category 1. This distribution suggests that a substantial portion of individuals in the dataset exhibit normal fasting blood sugar levels, while a smaller subset has elevated levels, which may be of interest for further investigation or analysis.

Graph 7 in *Appendix 3* reveals that the majority of observations fall under the "Normal" category. Additionally, there are approximately equal numbers of observations in the "ST" and "LVH" categories, though both are significantly less compared to the "Normal" category. By employing the code in *Appendix 1*, we can quantify these findings and determine that 60.13% of the observations are classified as "Normal," while 20.48% belong to "LVH," and 19.39% are in the "ST" category. This information provides insights into the distribution of Resting ECG results in the dataset and can be useful for further analysis and understanding of potential heart disease correlations.

Graph 8 in *Appendix 3* reveals that a significant majority of the observations belong to the category "No," indicating that most individuals in the dataset do not experience exercise-induced angina. Using the provided R code in *Appendix 1*, we find that approximately 59.59% of the dataset falls under the "No" category, while about 40.41% are classified as "Yes" for exercise-induced angina. This information provides valuable insight into the prevalence of exercise-induced angina within the dataset and helps identify potential patterns or correlations with other variables related to heart disease.

Graph 9 in *Appendix 3* reveals that the majority of observations fall under the "Flat" category, constituting approximately 50.11% of the dataset. The "Up" category follows closely behind with around 43.03% of the observations. However, the "Down" category has the fewest observations, accounting for only 6.86% of the dataset. This indicates that a larger proportion of individuals have a flat or upward sloping ST segment during peak exercise, while a much smaller percentage exhibit a downward sloping pattern. The code in *Appendix 1* provides these quantified percentages, which shed light on the distribution and pattern of ST Slope in the dataset.

By analyzing *Graph 10* in *Appendix 3*, we see that the graph is bell-shaped and normally distributed, with most of the values lying between BP 120 to 140. As 70 to 120 is considered the normal blood pressure range, the concentration of data points above it is reassuring. However, it is essential to note a small bar at the value of 0 in the RestingBP range, which is implausible since blood pressure cannot be zero. This presence of such values indicates possible errors or missing data points in the dataset related to resting blood pressure measurements. These inconsistencies can significantly impact the accuracy and reliability of the analysis.

The analysis of *Graph 11* in *Appendix 3* reveals some notable observations. Firstly, a significant bar appears at the value of 0, indicating potential errors or missing data in the dataset related to cholesterol levels. This issue needs to be addressed to ensure data accuracy. Excluding the data point at 0, the remaining cholesterol data follows a bell-shaped distribution with a slight positive skewness. The majority of observations concentrate in the range of 200 to 250.

Graph 12 in *Appendix 3* is perfectly bell shaped with no noticeable skewness. The maximum number of observations lie in the range of 120 to 130.

Graph 13, from *Appendix 3*, is perfectly bell shaped with no noticeable skewness. The maximum number of observations lie in the range of 50 to 55.
By categorizing the dataset into different age groups ranging from 20's to 70's and focusing only on observations with heart failure, we gain better insights into the distribution of ages among individuals with heart failure. *Graph 14* in *Appendix 3* displays a bell-shaped curve with a slight negative skewness. The majority of heart failure cases are concentrated in the 50's age group, followed by the 60's, 40's, 30's, 70's, and 20's. This analysis allows us to identify the age groups most affected by heart failure, with individuals in their 50's showing the highest prevalence,

providing valuable information for further investigation and potential interventions related to heart disease in different age cohorts.

Graph 15 in *Appendix 3* of the Old Peak variable displays a slightly positively skewed pattern. The majority of observations are concentrated within the range of -1 to 0, indicating that a significant proportion of individuals in the dataset experienced a relatively small magnitude of ST segment depression during their stress tests. There are very few observations located before the value of -1, suggesting that extreme negative ST segment depressions are infrequent in this dataset. Additionally, the number of observations gradually declines as the Old Peak values increase beyond 0, indicating a decreasing occurrence of larger ST segment depressions. This distribution pattern might suggest that most individuals had mild to moderate ST segment depressions, and extreme cases were relatively rare.

The chi-square test of independence conducted on the factors "Heart Disease" and "Sex" yielded a test statistic of X-squared = 84.145 with 1 degree of freedom and an extremely small p-value (< 2.2e-16). This indicates a highly significant association between the variables "Heart Disease" and “Sex”. Based on the chi-square test of independence, we reject the null hypothesis (H_0) and accept the alternative hypothesis (H_1) in *Research Hypothesis 1.*

The chi-square test of independence conducted on the factors "Heart Failure" and "Chest Pain" yielded a test statistic of X-squared = 268.07 with 3 degrees of freedom and a p-value < 2.2e-16, indicating a highly significant association between these variables. Based on the chi-square test of independence, we reject the null hypothesis (H_0) and accept the alternative hypothesis (H_1) in *Research Hypothesis 2.*

The chi-square test of independence with Yates' continuity correction conducted on the factors "Heart Disease" and "Fasting Blood Sugar" yielded a test statistic of X-squared = 64.321 with 1 degree of freedom and a p-value = 1.057e-15, indicating a highly significant association between these variables. Based on the chi-square test of independence, we reject the null hypothesis (H_0) and accept the alternative hypothesis (H_1) in *Research Hypothesis 3.*

The chi-square test of independence conducted on the factors "Heart Disease" and "Resting ECG" yielded a test statistic of X-squared = 10.931 with 2 degrees of freedom and a p-value = 0.004229, indicating a significant association between these variables. The chi-square test of independence shows that we reject the null hypothesis (H_0) and accept the alternative hypothesis (H_1) in *Research Hypothesis 4.*

The chi-square test of independence with Yates' continuity correction conducted on the factors "Heart Disease" and "Exercise Angina" yielded a test statistic of X-squared = 222.26 with 1 degree of freedom and a p-value < 2.2e-16, indicating a highly significant association between these variables. The chi-square test of independence leads to the rejection of the null hypothesis (H_0) and acceptance of the alternative hypothesis (H_1) in *Research Hypothesis 5.*

The chi-square test of independence conducted on the factors "Heart Disease" and "ST Slope" yielded a test statistic of X-squared = 355.92 with 2 degrees of freedom and a p-value < 2.2e-16, indicating a highly significant association between these variables. The chi-square test of independence results in the rejection of the null hypothesis (H_0) and acceptance of the alternative hypothesis (H_1) in *Research Hypothesis 6.*

The summary of the continuous variable "Age" is as follows:
- *Minimum: 28.00*
- *1st Quartile: 47.00*
- *Median: 54.00*
- *Mean: 53.51*
- *3rd Quartile: 60.00*
- *Maximum: 77.00*

This summary provides key descriptive statistics of the distribution of ages in the dataset. The minimum age observed is 28 years, while the maximum age is 77 years. The median age (the middle value) is 54, indicating that half of the individuals in the dataset have an age below 54 and the other half have an age above 54.
The mean age is 53.51, representing the average age of the individuals in the dataset. The first quartile (25th percentile) is 47, indicating that 25% of the individuals in the dataset have an age below 47. The third quartile (75th percentile) is 60, indicating that 75% of the individuals have an age below 60.

The summary of the continuous variable "RestingBP" is as follows:
- *Minimum: 0.0*
- *1st Quartile: 120.0*
- *Median: 130.0*
- *Mean: 132.4*
- *3rd Quartile: 140.0*
- *Maximum: 200.0*

This summary provides key descriptive statistics for the variable "RestingBP," which represents the resting blood pressure. The minimum value of 0.0 suggests that there may be some data points with missing or invalid values. The 1st quartile (25th percentile) is 120.0, indicating that 25% of the individuals in the dataset have a resting blood pressure below 120.0.
The median resting blood pressure is 130.0, representing the middle value in the dataset. The mean resting blood pressure is 132.4, indicating the average value for the variable. The 3rd quartile (75th percentile) is 140.0, suggesting that 75% of the individuals have a resting blood pressure below 140.0.
The maximum resting blood pressure is 200.0, representing the highest value observed in the dataset. It is important to note that values of 0.0 or extreme values such as 200.0 should be further investigated to ensure data integrity and accuracy.

The summary of the continuous variable "Cholesterol" is as follows:
- *Minimum: 0.0*
- *1st Quartile: 173.2*
- *Median: 223.0*
- *Mean: 198.8*
- *3rd Quartile: 267.0*
- *Maximum: 603.0*

This summary provides key descriptive statistics for the variable "Cholesterol," which represents the cholesterol levels in the dataset. The minimum value of 0.0 suggests that there may be some data points with missing or invalid values.

The 1st quartile (25th percentile) is 173.2, indicating that 25% of the individuals in the dataset have cholesterol levels below 173.2. The median cholesterol level is 223.0, representing the middle value in the dataset. The mean cholesterol level is 198.8, indicating the average value for the variable.

The 3rd quartile (75th percentile) is 267.0, suggesting that 75% of the individuals have cholesterol levels below 267.0. The maximum cholesterol level is 603.0, representing the highest value observed in the dataset.

It is important to note that cholesterol levels are typically measured in milligrams per deciliter (mg/dL), and values of 0.0 or extreme values such as 603.0 should be further investigated to ensure data integrity and accuracy.

The summary of the continuous variable "MaxHR" (Maximum Heart Rate) is as follows:
- *Minimum: 60.0*
- *1st Quartile: 120.0*
- *Median: 138.0*
- *Mean: 136.8*
- *3rd Quartile: 156.0*
- *Maximum: 202.0*

This summary provides key descriptive statistics for the variable "MaxHR," which represents the maximum heart rate observed in the dataset. The minimum value of 60.0 indicates the lowest recorded maximum heart rate, while the maximum value of 202.0 represents the highest recorded maximum heart rate.

The 1st quartile (25th percentile) is 120.0, meaning that 25% of the individuals in the dataset have a maximum heart rate below 120.0. The median maximum heart rate is 138.0, representing the middle value in the dataset. The mean maximum heart rate is 136.8, indicating the average maximum heart rate for the variable.

The 3rd quartile (75th percentile) is 156.0, suggesting that 75% of the individuals have a maximum heart rate below 156.0.

The summary of the continuous variable "Oldpeak" is as follows:
- *Minimum: -2.6000*
- *1st Quartile: 0.0000*
- *Median: 0.6000*
- *Mean: 0.8874*
- *3rd Quartile: 1.5000*

- Maximum: 6.2000

This summary provides key descriptive statistics for the variable "Oldpeak," which represents the ST depression induced by exercise relative to rest. The negative minimum value of -2.6000 suggests that the ST depression can go below the resting value.
The 1st quartile (25th percentile) is 0.0000, indicating that 25% of the individuals in the dataset have an Oldpeak value of 0.0000 or below. The median Oldpeak value is 0.6000, representing the middle value in the dataset. The mean Oldpeak is 0.8874, indicating the average value for the variable.
The 3rd quartile (75th percentile) is 1.5000, suggesting that 75% of the individuals have an Oldpeak value below 1.5000. The maximum Oldpeak value is 6.2000, representing the highest value observed in the dataset.

Model Fitting to Original Dataset

- Intercept: The intercept term represents the log-odds of heart failure when all other predictor variables are held constant. In this case, the estimated intercept is -1.163656. However, the p-value (0.411197) suggests that the intercept is not significantly different from zero.

- Age: The coefficient for Age is 0.016550, indicating that for every one-unit increase in age, the log-odds of heart failure increase by 0.016550. However, the p-value (0.209803) suggests that the effect of Age may not be statistically significant.

- Sex: The coefficient for SexM is 1.466477, suggesting that being male (SexM = 1) is associated with an increase in the log-odds of heart failure by 1.466477 compared to being female (SexM = 0). The p-value (< 2.2e-16) indicates that the effect of Sex is statistically significant.

- ChestPainType: The coefficients for ChestPainTypeATA, ChestPainTypeNAP, and ChestPainTypeTA represent the differences in log-odds of heart failure compared to the reference category (ChestPainTypeASY). For example, for ChestPainTypeATA, the coefficient is -1.830289, suggesting that individuals with ATA chest pain type have a lower log-odds of heart failure compared to those with ASY chest pain type. The p-values indicate that all three categories of ChestPainType are significantly associated with heart failure.

- RestingBP: The coefficient for RestingBP is 0.004194, indicating that a one-unit increase in resting blood pressure is associated with a 0.004194 increase in the log-odds of heart failure. However, the p-value (0.485296) suggests that the effect of RestingBP is not statistically significant.

- Cholesterol: The coefficient for Cholesterol is -0.004115, implying that a one-unit increase in cholesterol level is associated with a decrease of 0.004115 in the log-odds of heart failure. The p-value (0.000154) indicates that Cholesterol has a statistically significant effect on heart failure.

- FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak, and ST_Slope: The coefficients for these variables represent the log-odds ratios associated with the respective predictors. The p-values indicate that all of these variables have statistically significant effects on heart failure.

The deviance measures (Null deviance and Residual deviance) provide information about the goodness of fit of the model. In this case, the Residual deviance (594.19) is substantially lower than the Null deviance (1262.14), indicating that the model explains a significant amount of the variability in the data.

Replacing 0’s from the missing places with “NA” then fitting model

- The intercept term represents the estimated log-odds of having heart disease when all predictor variables are zero. In this case, the intercept is -5.4373.

- The coefficients estimate the change in log-odds of having heart disease associated with a one-unit increase in each predictor variable, holding all other variables constant.

- Age: For every one-unit increase in age, the log-odds of having heart disease increase by 0.0314, with a p-value of 0.0341 (indicating statistical significance).

- Sex: Being male (SexM) is associated with a higher risk of heart disease. The estimated increase in log-odds is 1.8655, with a very low p-value (2.64e-09), indicating statistical significance.

- ChestPainType: Having certain types of chest pain (ATA, NAP, TA) compared to the reference category is associated with a lower risk of heart disease. The estimated coefficients are negative, indicating a decrease in log-odds.

- RestingBP and Cholesterol: The coefficients for these variables represent the estimated change in log-odds of having heart disease associated with a one-unit increase in these variables. However, they are not statistically significant based on their p-values.

- FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak, ST_Slope: These variables have statistically significant coefficients, indicating they are associated with a change in the log-odds of having heart disease.

- The model's null deviance is the measure of the total variability explained by the model with only an intercept term. The residual deviance is the variability left unexplained after fitting the model with predictor variables. Smaller residual deviance indicates a better fit to the data.

- The AIC (Akaike Information Criterion) value is a measure of model quality, balancing goodness of fit and complexity. Lower AIC values suggest a better model fit.

Overall, this model suggests that age, sex, chest pain type, and several other variables are significant predictors of heart disease. However, it's important to note that the interpretation of coefficients in logistic regression models assumes linearity and other assumptions, so further analysis and validation are recommended.

Using KNN Method of Imputation then fitting the model

1. **Intercept -** The intercept represents the log-odds of HeartDisease when all predictor variables in the model are zero. In this case, the estimated intercept is -2.272021. However, the p-value (Pr(>|z|)) is 0.125519, which means that the intercept is not statistically significant at the usual significance level of 0.05. This suggests that the log-odds of HeartDisease when all predictor variables are zero is not significantly different from zero.
2. **Age -** For every one-unit increase in Age, the log-odds of HeartDisease increases by 0.019284. However, the p-value (Pr(>|z|)) is 0.141979, which indicates that the relationship between Age and HeartDisease is not statistically significant at the usual significance level of 0.05. In other words, Age does not have a significant effect on the likelihood of having heart disease in this model.
3. **Sex -** Being male (SexM) is associated with a significant increase in the log-odds of HeartDisease. Specifically, males have approximately exp(1.655037) = 5.22 times higher odds of having heart disease compared to females. The p-value (Pr(>|z|)) is extremely small (close to zero), indicating that this effect is highly statistically significant.
4. **ChestPainTypeATA -** Having chest pain type "ATA" (asymptomatic) is associated with a significant decrease in the log-odds of HeartDisease. Specifically, individuals with chest pain type ATA have approximately exp(-1.911300) = 0.148 times the odds of having heart disease compared to other chest pain types. The p-value (Pr(>|z|)) is extremely small, indicating that this effect is highly statistically significant.
5. **ChestPAinTypeNAP -** Having chest pain type "NAP" (non-anginal pain) is associated with a significant decrease in the log-odds of HeartDisease. Specifically, individuals with chest pain type NAP have approximately exp(-1.613028) = 0.199 times the odds of having heart disease compared to other chest pain types. The p-value (Pr(>|z|)) is extremely small, indicating that this effect is highly statistically significant.
6. **ChestPainTypeTA -** Having chest pain type "TA" (typical angina) is associated with a significant decrease in the log-odds of HeartDisease. Specifically, individuals with chest pain type TA have approximately exp(-1.471749) = 0.229 times the odds of having heart disease compared to other chest pain types. The p-value (Pr(>|z|)) is small, indicating that this effect is statistically significant.
7. **RestingBP -** The variable RestingBP represents the resting blood pressure. The estimate of 0.002163 means that for every one-unit increase in RestingBP, the log-odds of HeartDisease increases by 0.002163. However, the p-value (Pr(>|z|)) is 0.725193, indicating that RestingBP is not statistically significant in predicting the likelihood of heart disease. Therefore, RestingBP does not seem to have a significant effect on the odds of having heart disease in this model.
8. **Cholesterol -** The variable Cholesterol represents the cholesterol level. The estimate of 0.002875 means that for every one-unit increase in Cholesterol, the log-odds of HeartDisease increases by 0.002875. However, the p-value (Pr(>|z|)) is 0.149811, indicating that Cholesterol is not statistically significant in predicting the likelihood of heart disease. Therefore, Cholesterol does not seem to have a significant effect on the odds of having heart disease in this model.
9. **FastingBS -** FastingBS represents the fasting blood sugar level, where 1 indicates high fasting blood sugar and 0 indicates normal fasting blood sugar. The estimate of 1.323663 means that having high fasting blood sugar is associated with a significant increase in the log-odds of HeartDisease. Specifically, individuals with high fasting blood sugar have approximately exp(1.323663) = 3.760 times higher odds of having heart disease compared to those with normal fasting blood sugar. The p-value (Pr(>|z|)) is extremely small, indicating that this effect is highly statistically significant.

10. **RestingECGNormal -** RestingECGNormal is a binary variable indicating whether the resting electrocardiographic results are normal (1) or not (0). The estimate of 0.028619 indicates that individuals with normal resting ECG results have a slightly higher log-odds of HeartDisease, but the effect is not statistically significant. The p-value (Pr(>|z|)) is 0.914827, suggesting that RestingECGNormal is not a significant predictor of heart disease in this model.
11. **RestingECGST -** RestingECGST is a binary variable indicating whether the resting electrocardiographic results show probable or definite left ventricular hypertrophy by Estes' criteria (1) or not (0). The estimate of 0.023513 suggests that individuals with this type of electrocardiographic result have a slightly higher log-odds of HeartDisease, but again, the effect is not statistically significant. The p-value (Pr(>|z|)) is 0.945560, indicating that RestingECGST is not a significant predictor of heart disease in this model.
12. **MaxHR -** MaxHR represents the maximum heart rate achieved during exercise. The estimate of -0.007709 means that for every one-unit increase in MaxHR, the log-odds of HeartDisease decreases by 0.007709. However, the p-value (Pr(>|z|)) is 0.115619, indicating that MaxHR is not statistically significant in predicting the likelihood of heart disease. Therefore, MaxHR does not seem to have a significant effect on the odds of having heart disease in this model.
13. **ExerciseAnginaY -** ExerciseAnginaY is a binary variable indicating whether exercise induces angina (chest pain) (1) or not (0). The estimate of 0.829775 means that individuals with exercise-induced angina have a significantly higher log-odds of HeartDisease. Specifically, individuals with exercise-induced angina have approximately exp(0.829775) = 2.292 times higher odds of having heart disease compared to those without angina during exercise. The p-value (Pr(>|z|)) is extremely small, indicating that this effect is highly statistically significant.
14. **Oldpeak -** Oldpeak is the ST depression induced by exercise relative to rest. The estimate of 0.364049 means that for every one-unit increase in Oldpeak, the log-odds of HeartDisease increases by 0.364049. The p-value (Pr(>|z|)) is extremely small, indicating that Oldpeak is a highly statistically significant predictor of heart disease. Individuals with higher ST depression during exercise are more likely to have heart disease.
15. **ST_SlopeFlat -** ST_SlopeFlat is a binary variable indicating a flat ST segment during exercise (1) or not (0). The estimate of 1.240525 means that individuals with a flat ST segment during exercise have a significantly higher log-odds of HeartDisease. Specifically, individuals with a flat ST segment have approximately exp(1.240525) = 3.461 times higher odds of having heart disease compared to those without a flat ST segment. The p-value (Pr(>|z|)) is extremely small, indicating that this effect is highly statistically significant.
16. **ST_SlopeUp -** ST_SlopeUp is a binary variable indicating an upsloping ST segment during exercise (1) or not (0). The estimate of -1.099357 suggests that individuals with an upsloping ST segment during exercise have a significantly lower log-odds of HeartDisease. Specifically, individuals with an upsloping ST segment have approximately exp(-1.099357) = 0.333 times lower odds of having heart disease compared to those without an upsloping ST segment. The p-value (Pr(>|z|)) is 0.013745, indicating that this effect is statistically significant at the 0.05 significance level.

Therefore, we can conclude that the variables Heart Failure is affected by other variables like Sex, Chest Pain, ExerciseAngina, Oldpeak, and ST_Slope, and we reject the null hypothesis (H0) in favor of the alternative hypothesis (H1) for Research Hypothesis 7.

Fitting GLM model separately to dataset containing only Males

1. The intercept coefficient is 0.737, which represents the estimated log-odds of having heart disease when all other predictors are zero.
2. `Age` has a coefficient of 0.018, indicating that for every one-unit increase in age, the log-odds of having heart disease increase by 0.018. However, it is not statistically significant (p-value > 0.05).
3. The variables `ChestPainTypeATA`, `ChestPainTypeNAP`, and `ChestPainTypeTA` represent the effect of different types of chest pain. They are all statistically significant (p-values < 0.001) and have negative coefficients, which suggests that individuals with specific types of chest pain have lower odds of having heart disease compared to the reference category.
4. `FastingBS` (Fasting Blood Sugar) has a coefficient of 1.086, indicating that individuals with fasting blood sugar greater than 120 mg/dl have higher odds of having heart disease compared to those with lower fasting blood sugar levels.
5. `MaxHR` (Maximum Heart Rate achieved) has a coefficient of -0.011, suggesting that for every one-unit increase in maximum heart rate, the log-odds of having heart disease decrease by 0.011.
6. `ExerciseAnginaY` (Exercise-induced angina) has a coefficient of 0.814, indicating that individuals with exercise-induced angina have higher odds of having heart disease compared to those without it.
7. `Oldpeak` has a coefficient of 0.393, meaning that for every one-unit increase in oldpeak (ST segment depression), the log-odds of having heart disease increase by 0.393.
The residual deviance is 485.44 on 710 degrees of freedom, and the AIC (Akaike Information Criterion) is 515.44. These values help assess the goodness of fit of the model. The null deviance, which represents the residual deviance when only the intercept is included in the model, is 954.15 on 724 degrees of freedom.

Fitting GLM model separately to dataset containing only Females

1. The intercept coefficient is -7.702, which represents the estimated log-odds of having heart disease for females when all other predictors are zero.
2. `Age` has a coefficient of 0.022, indicating that for every one-unit increase in age, the log-odds of having heart disease slightly increase by 0.022. However, it is not statistically significant (p-value > 0.05), suggesting that age may not be a strong predictor of heart disease among females in this dataset.
3. The variables `ChestPainTypeATA`, `ChestPainTypeNAP`, and `ChestPainTypeTA` represent the effect of different types of chest pain. Among females, only `ChestPainTypeATA` and `ChestPainTypeNAP` are statistically significant (p-values < 0.05), indicating that individuals with these specific types of chest pain have lower odds of having heart disease compared to the reference category.
4. `FastingBS` (Fasting Blood Sugar) has a coefficient of 3.225, suggesting that females with fasting blood sugar greater than 120 mg/dl have significantly higher odds of having heart disease compared to those with lower fasting blood sugar levels.
5. None of the other variables (such as `Cholesterol`, `RestingBP`, `RestingECG`, `MaxHR`, `ExerciseAngina`, `Oldpeak`, and `ST_Slope`) are statistically significant predictors of heart disease among females in this dataset (p-values > 0.05).

The residual deviance is 105.49 on 178 degrees of freedom, and the AIC (Akaike Information Criterion) is 135.49. These values help assess the goodness of fit of the model. The null deviance, which represents the residual deviance when only the intercept is included in the model, is 220.82 on 192 degrees of freedom.

Conclusion

This project utilized Generalized Linear Models (GLM) to explore and analyze a dataset related to heart disease, focusing on understanding the significance of various demographic and clinical variables in predicting heart failure. The dataset comprised factors such as age, chest pain type, blood pressure, cholesterol levels, fasting blood sugar, resting ECG results, maximum heart rate, exercise-induced angina, and ST segment slope.

In the GLM models, we observed interesting insights regarding the significance of the predictor variables. In the model fitted to the entire dataset, the variables "Sex," "ChestPainType," "FastingBS," "ExerciseAngina," "Oldpeak," and "ST_SlopeFlat" were found to be statistically significant, as indicated by their low p-values ($p < 0.05$). These variables are considered important predictors in determining heart failure risk.

However, certain variables were found to be statistically insignificant in the model, with p-values greater than 0.05. These variables include "Age," "RestingBP," "Cholesterol," "RestingECGNormal," "RestingECGST," "MaxHR," and "ST_SlopeUp." While these variables may not be significant in predicting heart failure in the overall dataset, it is essential to consider that significance might change when analyzing subsets of the data.

Upon splitting the dataset based on gender, we fitted separate GLM models to each group. In the model for males, "ChestPainType," "FastingBS," and "ST_SlopeFlat" remained significant predictors, while "Age," "RestingBP," "Cholesterol," "RestingECGNormal," "RestingECGST," "MaxHR," "ExerciseAngina," "Oldpeak," and "ST_SlopeUp" were not statistically significant. For females, "ChestPainType" and "FastingBS" remained significant, but "Age," "RestingBP," "Cholesterol," "RestingECGNormal," "RestingECGST," "MaxHR," "ExerciseAngina," "Oldpeak," "ST_SlopeFlat," and "ST_SlopeUp" were not significant.

In summary, the GLM analysis revealed that certain variables, such as "ChestPainType" and "FastingBS," were consistently significant predictors of heart failure across the entire dataset and within gender-specific subsets. Other variables showed varying significance levels, suggesting that their impact on heart failure prediction may differ based on gender. These findings highlight the importance of considering gender-specific factors and may guide further investigations into personalized risk assessment and treatment strategies for heart disease. However, it is essential to conduct further research to validate and expand upon these results for a comprehensive understanding of heart failure risk factors.

Based on the results of the chi-square test of independence, which was performed to assess the relationship between categorical variables and heart failure, certain variables were found to be significant predictors of heart failure. The variables "Sex," "ChestPainType," "FastingBS," "ExerciseAngina," "Oldpeak," and "ST_SlopeFlat" were found to be statistically significant in predicting heart failure. These results align with the findings from the GLM models, indicating that these variables have a strong association with heart failure risk.

On the other hand, some variables were found to be statistically insignificant in both the chi-square test and GLM models. These variables include "Age," "RestingBP," "Cholesterol," "RestingECGNormal," "RestingECGST," "MaxHR," and "ST_SlopeUp." The lack of significance suggests that these variables may not be strong predictors of heart failure in the analyzed dataset.

The consistency of significant and insignificant variables between the chi-square test and GLM models adds confidence to the findings. Variables that exhibit significance in both analyses can be considered robust and informative in predicting heart failure risk. Conversely, variables that are insignificant in both analyses may be less influential and may not warrant significant attention when assessing heart failure risk.

Overall, combining the results of the chi-square test and GLM models provides a comprehensive view of the factors that contribute to heart failure in the dataset. This information can guide further research and medical interventions, allowing for a more personalized and targeted approach to prevent, diagnose, and manage heart disease effectively.

Appendix 1

R codes -

```
#Loading dataset
data <- read.csv(file.choose()) #file name - "heart.csv"
head(data)


###Graphs

##Plotting Categorical Variables
# Define colors for each bar plot
barplot_colors <- c("salmon", "lightgreen", "skyblue", "orchid", "turquoise",
"lightpink", "lightskyblue", "lightcoral")

# Barplot for Heart Failure (Graph 1)
table_heart_disease <- table(data$HeartDisease)
par(mar = c(5, 4, 4, 2) + 0.1) # Adjust margins
barplot(table_heart_disease, main = "Barplot on Heart Failure", col =
barplot_colors[1:2], ylim = c(0, max(table_heart_disease)*1.2))
legend("topright", legend = c("0: Normal", "1: Heart Disease"), fill =
barplot_colors[1:2])

# Barplot for Number of Male and Female (Graph 2)
table_sex <- table(data$Sex)
par(mar = c(5, 4, 4, 2) + 0.1) # Adjust margins
barplot(table_sex, main = "Barplot for Number of Male and Female", col =
barplot_colors[3:4], ylim = c(0, max(table_sex)*1.2))
legend("topright", legend = c("M: Male", "F: Female"), fill = barplot_colors[3:4])

# Barplot for types of Chest Pain (Graph 5)
table_chest_pain <- table(data$ChestPainType)
par(mar = c(5, 4, 4, 4)) # Adjust margins
barplot(table_chest_pain, main = "Barplot for types of Chest Pain", col =
barplot_colors[5:8], ylim = c(0, max(table_chest_pain)*1.2))
legend("topright", legend = c("TA: Typical Angina", "ATA: Atypical Angina", "NAP: Non-
Anginal Pain", "ASY: Asymptomatic"), fill = barplot_colors[5:8])

# Barplot for Blood Sugar (Graph 6)
table_fasting_bs <- table(data$FastingBS)
par(mar = c(5, 4, 4, 2) + 0.1) # Adjust margins
barplot(table_fasting_bs, main = "Barplot for Blood Sugar", col = barplot_colors[3:4],
ylim = c(0, max(table_fasting_bs)*1.2))
legend("topright", legend = c("0: otherwise", "1: if FastingBS > 120 mg/dl"), fill =
barplot_colors[3:4])

# Barplot for Resting ECG (Graph 7)
table_resting_ecg <- table(data$RestingECG)
par(mar = c(5, 4, 4, 5)) # Adjust margins
barplot(table_resting_ecg, main = "Barplot for Resting ECG", col = barplot_colors[5:7],
ylim = c(0, max(table_resting_ecg)*1.2))
legend("topright", legend = c("Normal: Normal", "ST: having ST-T wave abnormality", "LVH:
showing probable or definite left ventricular hypertrophy"), fill = barplot_colors[5:7])

# Barplot for Exercise Angina (Graph 8)
table_exercise_angina <- table(data$ExerciseAngina)
par(mar = c(5, 4, 4, 2) + 0.1) # Adjust margins
barplot(table_exercise_angina, main = "Barplot on Exercise Angina", col =
barplot_colors[3:4], ylim = c(0, max(table_exercise_angina)*1.2))
legend("topright", legend = c("Y: Yes", "N: No"), fill = barplot_colors[3:4])

# Barplot for ST Slope (Graph 9)
table_st_slope <- table(data$ST_Slope)
par(mar = c(5, 4, 4, 3)) # Adjust margins
barplot(table_st_slope, main = "Barplot on ST Slope", col = barplot_colors[5:7], ylim =
c(0, max(table_st_slope)*1.2))
legend("topright", legend = c("Down", "Flat", "Up"), fill = barplot_colors[5:7])

##Plotting Continuous Variables
```

```
# Set colors for each histogram
hist_colors <- c("skyblue", "salmon", "lightgreen", "lightpink", "turquoise")

# Histogram of Resting Blood Pressure (Graph 10)
hist(data$RestingBP, main = "Histogram of Resting Blood Pressure", xlab = "Resting BP",
col = hist_colors[1])

# Histogram of Cholesterol (Graph 11)
hist(data$Cholesterol, main = "Histogram of Cholesterol", xlab = "Cholesterol", col =
hist_colors[2])

# Histogram of Maximum Heart Rate (Graph 12)
hist(data$MaxHR, main = "Histogram of Maximum Heart Rate", xlab = "Maximum Heart Rate",
col = hist_colors[3])

# Histogram of Age (Graph 13)
hist(data$Age, main = "Histogram of Age", xlab = "Age", col = hist_colors[4])

# Histogram of Oldpeak (Graph 15)
hist(data$Oldpeak, main = "Histogram of Oldpeak", xlab = "Oldpeak", col = hist_colors[5])


#Chi Square Test of Independence
chisq.test(data$HeartDisease,data$Sex)
chisq.test(data$HeartDisease,data$ChestPainType)
chisq.test(data$HeartDisease,data$FastingBS)
chisq.test(data$HeartDisease,data$RestingECG)
chisq.test(data$HeartDisease,data$ExerciseAngina)
chisq.test(data$HeartDisease,data$ST_Slope)

#Summary of Continuous Variables
summary(data$Age)
summary(data$RestingBP)
summary(data$Cholesterol)
summary(data$MaxHR)
summary(data$Oldpeak)


##Fitting

model<-
glm(data$HeartDisease~data$Age+data$Sex+data$ChestPainType+data$RestingBP+data$Cholestero
l+data$FastingBS+data$RestingECG+data$MaxHR+data$ExerciseAngina+data$Oldpeak+data$ST_Slop
e,family = binomial)
summary(model)
#Replacing the missing values with NA

summary(data)

data1 <- data

data1$RestingBP[data$RestingBP == 0] <- NA
data1$Cholesterol[data$Cholesterol == 0] <- NA

summary(data1)

model1<-
glm(data1$HeartDisease~data1$Age+data1$Sex+data1$ChestPainType+data1$RestingBP+data1$Chol
esterol+data1$FastingBS+data1$RestingECG+data1$MaxHR+data1$ExerciseAngina+data1$Oldpeak+d
ata1$ST_Slope,family = binomial)
summary(model1)


#Replacing missing values with KNN Method of imputation
#Pagkage VIM
data2 <- kNN(data1,variable = c("Cholesterol","RestingBP"))
#write.csv(data2,"/Volumes/STRANGER/6th Sem Project/Swapnil//Imputed
Data.csv",row.names=FALSE)

data2 <- subset(data2,select = -c(Cholesterol_imp,RestingBP_imp))
data2
```

```
model2<-
glm(data2$HeartDisease~data2$Age+data2$Sex+data2$ChestPainType+data2$RestingBP+data2$Chol
esterol+data2$FastingBS+data2$RestingECG+data2$MaxHR+data2$ExerciseAngina+data2$Oldpeak+d
ata2$ST_Slope,family = binomial)

summary(model2)

#Splitting the Dataset for Males and Females
dataM <- subset(data2, Sex == "M", select = -c(Sex))
dataF <- subset(data2, Sex == "F", select = -c(Sex))

#Fitting model Separately for males and Females
modelM <-
glm(dataM$HeartDisease~dataM$Age+dataM$ChestPainType+dataM$RestingBP+dataM$Cholesterol+da
taM$FastingBS+dataM$RestingECG+dataM$MaxHR+dataM$ExerciseAngina+dataM$Oldpeak+dataM$ST_Sl
ope,family = binomial)
summary(modelM)

modelF <-
glm(dataF$HeartDisease~dataF$Age+dataF$ChestPainType+dataF$RestingBP+dataF$Cholesterol+da
taF$FastingBS+dataF$RestingECG+dataF$MaxHR+dataF$ExerciseAngina+dataF$Oldpeak+dataF$ST_Sl
ope,family = binomial)
summary(modelF)

#finding percentage of heart failure in each sex

#Males
# Count the number of heart failure cases (1) in the "dataM" dataset
heart_failure_count <- sum(dataM$HeartDisease == 1)

# Calculate the total number of observations in the "dataM" dataset
total_observations <- nrow(dataM)

# Calculate the percentage of heart failure in "dataM"
heart_failure_percentage_M <- (heart_failure_count / total_observations) * 100
round(heart_failure_percentage_M, 2)

#Females
# Count the number of heart failure cases (1) in the "dataF" dataset
heart_failure_count <- sum(dataF$HeartDisease == 1)

# Calculate the total number of observations in the "dataF" dataset
total_observations <- nrow(dataF)

# Calculate the percentage of heart failure in "dataM"
heart_failure_percentage_F <- (heart_failure_count / total_observations) * 100
round(heart_failure_percentage_F, 2)

#Plotting frequency graph for heart failure among males and females separately
# Define colors for the bar plots
barplot_colors <- c("skyblue", "salmon")

# Barplot on Heart Failure among Males (Graph 3)
table_heart_disease_m <- table(dataM$HeartDisease)
max_freq_m <- max(table_heart_disease_m)
barplot(table_heart_disease_m, main = "Barplot on Heart Failure among Males", col =
barplot_colors, ylim = c(0, max_freq_m * 1.2))
legend("topright", legend = c("0: Normal", "1: Heart Disease"), fill = barplot_colors)

# Barplot on Heart Failure among Females (Graph 4)
table_heart_disease_f <- table(dataF$HeartDisease)
max_freq_f <- max(table_heart_disease_f)
barplot(table_heart_disease_f, main = "Barplot on Heart Failure among Females", col =
barplot_colors, ylim = c(0, max_freq_f * 1.2))
legend("topright", legend = c("0: Normal", "1: Heart Disease"), fill = barplot_colors)


##Age Grouping

data3 <- data2
data3$AgeGroup = cut(data2$Age, breaks = c(20,30,40,50,60,70,80),labels =
c("20's","30's","40's","50's","60's","70's"),right = FALSE)
```

```r
data3

# Create a new dataset data3_heart_failure containing only the observations with heart
      failure
data3_heart_failure <- data3[data3$HeartDisease == 1, ]


# Plot the bar plot to see the frequency distribution of AgeGroup in observations with
      heart failure (Graph 14)
# Define colors for the bar plot
barplot_colors <- c("lightblue", "lightgreen", "lightpink", "lightcoral", "lightsalmon",
"lightseagreen")

# Create a table for the frequency distribution of AgeGroup in observations with heart
      failure
table_age_group <- table(data3_heart_failure$AgeGroup)

# Barplot for Frequency Distribution of AgeGroup in Observations with Heart Failure
barplot(table_age_group, main = "Frequency Distribution of AgeGroup in Observations with
Heart Failure",
        xlab = "AgeGroup", ylab = "Frequency", col = barplot_colors)


## Percentage Calculation

#Heart Failure %
# Count the number of people with heart failure (HeartDisease = 1)
num_heart_failure <- sum(data2$HeartDisease == 1)

# Calculate the total number of people in the dataset
total_people <- nrow(data2)

# Calculate the percentage of people with heart failure
percentage_heart_failure <- (num_heart_failure / total_people) * 100

round(percentage_heart_failure, 2)


#Sex %
# Count the number of males and females
num_males <- sum(data2$Sex == "M")
num_females <- sum(data2$Sex == "F")

# Calculate the total number of observations
total_observations <- nrow(data2)

# Calculate the percentage of males and females
percentage_male <- (num_males / total_observations) * 100
percentage_female <- (num_females / total_observations) * 100

round(percentage_male, 2)
round(percentage_female, 2)


#Chest Pain Type
# Calculate the frequency of each type of Chest Pain Type
chest_pain_freq <- table(data2$ChestPainType)

# Calculate the total number of observations in the dataset
total_observations <- nrow(data2)

# Calculate the percentage of each type of Chest Pain Type
chest_pain_percentage <- chest_pain_freq / total_observations * 100

# Print the results
chest_pain_percentage


#fastingBS
# Calculate percentage observations in each category of Fasting Blood Sugar
percentage_fasting_bs <- table(data2$FastingBS) / length(data2$FastingBS) * 100
```

```
# Print the percentage observations for each category
percentage_fasting_bs

#Resting ECG
# Calculate the percentage of observations in each category of Resting ECG
resting_ecg_percentages <- round(prop.table(table(data2$RestingECG)) * 100, 2)

# Print the percentages
resting_ecg_percentages

#Exercise Angina
# Calculate the percentage of observations in each category of Exercise Angina
percentage_exercise_angina <- prop.table(table(data2$ExerciseAngina)) * 100

# Print the results
percentage_exercise_angina

#ST Slope
# Calculate the percentage of observations in each category of ST Slope
percentage_st_slope <- prop.table(table(data2$ST_Slope)) * 100

# Display the percentages
percentage_st_slope
```

Appendix 2

R Outputs -

```
> #Loading dataset
> data <- read.csv(file.choose()) #file name - "heart.csv"
> head(data)
  Age Sex ChestPainType RestingBP Cholesterol FastingBS RestingECG MaxHR
1  40   M           ATA       140         289         0     Normal   172
2  49   F           NAP       160         180         0     Normal   156
3  37   M           ATA       130         283         0         ST    98
4  48   F           ASY       138         214         0     Normal   108
5  54   M           NAP       150         195         0     Normal   122
6  39   M           NAP       120         339         0     Normal   170
  ExerciseAngina Oldpeak ST_Slope HeartDisease
1              N     0.0       Up            0
2              N     1.0     Flat            1
3              N     0.0       Up            0
4              Y     1.5     Flat            1
5              N     0.0       Up            0
6              N     0.0       Up            0


> #Chi Square Test of Independence
> chisq.test(data$HeartDisease,data$Sex)

	Pearson's Chi-squared test with Yates' continuity correction

data:  data$HeartDisease and data$Sex
X-squared = 84.145, df = 1, p-value < 2.2e-16

> chisq.test(data$HeartDisease,data$ChestPainType)

	Pearson's Chi-squared test

data:  data$HeartDisease and data$ChestPainType
X-squared = 268.07, df = 3, p-value < 2.2e-16

> chisq.test(data$HeartDisease,data$FastingBS)

	Pearson's Chi-squared test with Yates' continuity correction

data:  data$HeartDisease and data$FastingBS
X-squared = 64.321, df = 1, p-value = 1.057e-15

> chisq.test(data$HeartDisease,data$RestingECG)

	Pearson's Chi-squared test

data:  data$HeartDisease and data$RestingECG
X-squared = 10.931, df = 2, p-value = 0.004229

> chisq.test(data$HeartDisease,data$ExerciseAngina)

	Pearson's Chi-squared test with Yates' continuity correction

data:  data$HeartDisease and data$ExerciseAngina
X-squared = 222.26, df = 1, p-value < 2.2e-16

> chisq.test(data$HeartDisease,data$ST_Slope)

	Pearson's Chi-squared test

data:  data$HeartDisease and data$ST_Slope
X-squared = 355.92, df = 2, p-value < 2.2e-16


> #Summary of Continuous Variables
> summary(data$Age)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  28.00   47.00   54.00   53.51   60.00   77.00 
> summary(data$RestingBP)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    0.0   120.0   130.0   132.4   140.0   200.0 
> summary(data$Cholesterol)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    0.0   173.2   223.0   198.8   267.0   603.0 
> summary(data$MaxHR)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   60.0   120.0   138.0   136.8   156.0   202.0 
> summary(data$Oldpeak)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-2.6000  0.0000  0.6000  0.8874  1.5000  6.2000 


> ##Fitting
> 
> model<-
glm(data$HeartDisease~data$Age+data$Sex+data$ChestPainType+data$RestingBP+data$Cholestero
l+data$FastingBS+data$RestingECG+data$MaxHR+data$ExerciseAngina+data$Oldpeak+data$ST_Slop
e,family = binomial)
> summary(model)

Call:
glm(formula = data$HeartDisease ~ data$Age + data$Sex + data$ChestPainType + 
    data$RestingBP + data$Cholesterol + data$FastingBS + data$RestingECG + 
    data$MaxHR + data$ExerciseAngina + data$Oldpeak + data$ST_Slope, 
    family = binomial)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)    
(Intercept)          -1.163656   1.416003  -0.822 0.411197    
data$Age              0.016550   0.013197   1.254 0.209803    
data$SexM             1.466477   0.279834   5.241 1.60e-07 ***
data$ChestPainTypeATA -1.830289   0.326293  -5.609 2.03e-08 ***
data$ChestPainTypeNAP -1.685682   0.266001  -6.337 2.34e-10 ***
data$ChestPainTypeTA  -1.488392   0.432572  -3.441 0.000580 ***
data$RestingBP        0.004194   0.006010   0.698 0.485296    
data$Cholesterol     -0.004115   0.001087  -3.785 0.000154 ***
data$FastingBS        1.136482   0.274999   4.133 3.59e-05 ***
data$RestingECGNormal -0.177033   0.271925  -0.651 0.515022    
data$RestingECGST    -0.268546   0.350020  -0.767 0.442945    
data$MaxHR           -0.004288   0.005023  -0.854 0.393249    
data$ExerciseAnginaY  0.900292   0.244513   3.682 0.000231 ***
data$Oldpeak          0.380643   0.118466   3.213 0.001313 ** 
data$ST_SlopeFlat     1.453902   0.429086   3.388 0.000703 ***
data$ST_SlopeUp      -0.994101   0.450196  -2.208 0.027234 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1262.14  on 917  degrees of freedom
Residual deviance:  594.19  on 902  degrees of freedom
AIC: 626.19

Number of Fisher Scoring iterations: 6


> #Replacing the missing values with NA
> 
> summary(data)
      Age            Sex            ChestPainType        RestingBP    
 Min.   :28.00   Length:918         Length:918         Min.   :  0.0  
 1st Qu.:47.00   Class :character   Class :character   1st Qu.:120.0  
 Median :54.00   Mode  :character   Mode  :character   Median :130.0  
 Mean   :53.51                                         Mean   :132.4  
 3rd Qu.:60.00                                         3rd Qu.:140.0  
 Max.   :77.00                                         Max.   :200.0  
  Cholesterol      FastingBS       RestingECG            MaxHR      
 Min.   :  0.0   Min.   :0.0000   Length:918         Min.   : 60.0  
 1st Qu.:173.2   1st Qu.:0.0000   Class :character   1st Qu.:120.0  
 Median :223.0   Median :0.0000   Mode  :character   Median :138.0  
 Mean   :198.8   Mean   :0.2331                      Mean   :136.8  
 3rd Qu.:267.0   3rd Qu.:0.0000                      3rd Qu.:156.0  
 Max.   :603.0   Max.   :1.0000                      Max.   :202.0  
 ExerciseAngina        Oldpeak          ST_Slope          HeartDisease   
 Length:918         Min.   :-2.6000   Length:918         Min.   :0.0000  
 Class :character   1st Qu.: 0.0000   Class :character   1st Qu.:0.0000  
 Mode  :character   Median : 0.6000   Mode  :character   Median :1.0000  
                    Mean   : 0.8874                      Mean   :0.5534  
                    3rd Qu.: 1.5000                      3rd Qu.:1.0000  
                    Max.   : 6.2000                      Max.   :1.0000  
> 
> data1 <- data
> 
> data1$RestingBP[data$RestingBP == 0] <- NA
> data1$Cholesterol[data$Cholesterol == 0] <- NA
> 
> summary(data1)
      Age             Sex            ChestPainType        RestingBP    
 Min.   :28.00   Length:918         Length:918         Min.   : 80.0  
 1st Qu.:47.00   Class :character   Class :character   1st Qu.:120.0  
 Median :54.00   Mode  :character   Mode  :character   Median :130.0  
 Mean   :53.51                                         Mean   :132.5  
 3rd Qu.:60.00                                         3rd Qu.:140.0  
 Max.   :77.00                                         Max.   :200.0  
                                                       NA's   :1      
  Cholesterol      FastingBS       RestingECG            MaxHR      
 Min.   : 85.0   Min.   :0.0000   Length:918         Min.   : 60.0  
 1st Qu.:207.2   1st Qu.:0.0000   Class :character   1st Qu.:120.0  
 Median :237.0   Median :0.0000   Mode  :character   Median :138.0  
 Mean   :244.6   Mean   :0.2331                      Mean   :136.8  
 3rd Qu.:275.0   3rd Qu.:0.0000                      3rd Qu.:156.0  
 Max.   :603.0   Max.   :1.0000                      Max.   :202.0  
 NA's   :172                                                        
 ExerciseAngina        Oldpeak          ST_Slope          HeartDisease   
 Length:918         Min.   :-2.6000   Length:918         Min.   :0.0000  
 Class :character   1st Qu.: 0.0000   Class :character   1st Qu.:0.0000  
 Mode  :character   Median : 0.6000   Mode  :character   Median :1.0000  
                    Mean   : 0.8874                      Mean   :0.5534  
                    3rd Qu.: 1.5000                      3rd Qu.:1.0000  
                    Max.   : 6.2000                      Max.   :1.0000  

> 
> model1<-
glm(data1$HeartDisease~data1$Age+data1$Sex+data1$ChestPainType+data1$RestingBP+data1$Chol
esterol+data1$FastingBS+data1$RestingECG+data1$MaxHR+data1$ExerciseAngina+data1$Oldpeak+d
ata1$ST_Slope,family = binomial)
> summary(model1)

Call:
glm(formula = data1$HeartDisease ~ data1$Age + data1$Sex + data1$ChestPainType + 
    data1$RestingBP + data1$Cholesterol + data1$FastingBS + data1$RestingECG + 
    data1$MaxHR + data1$ExerciseAngina + data1$Oldpeak + data1$ST_Slope, 
    family = binomial)

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)    
(Intercept)            -5.4373046  1.7625169  -3.085 0.002036 ** 
data1$Age               0.0313784  0.0148105   2.119 0.034119 *  
data1$SexM              1.8655490  0.3134065   5.952 2.64e-09 ***
data1$ChestPainTypeATA -1.6731804  0.3544226  -4.721 2.35e-06 ***
data1$ChestPainTypeNAP -1.5730121  0.3029404  -5.192 2.08e-07 ***
data1$ChestPainTypeTA  -1.6332529  0.4838117  -3.376 0.000736 ***
data1$RestingBP         0.0117792  0.0072988   1.614 0.106557    
data1$Cholesterol       0.0024955  0.0019773   1.262 0.206928    
data1$FastingBS         0.2923999  0.3311265   0.883 0.377212    
data1$RestingECGNormal -0.2297888  0.2842091  -0.809 0.418791    
data1$RestingECGST     -0.1746017  0.3941671  -0.443 0.657792    
data1$MaxHR             0.0005807  0.0057810   0.100 0.919991    
data1$ExerciseAnginaY   0.9073515  0.2671360   3.397 0.000682 ***
data1$Oldpeak           0.4108355  0.1406671   2.921 0.003493 ** 
```

```
data1$ST_SlopeFlat       1.3038217  0.5197574   2.509 0.012124 *
data1$ST_SlopeUp        -1.2100372  0.5655279  -2.140 0.032382 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1032.63  on 745  degrees of freedom
Residual deviance:  483.58  on 730  degrees of freedom
  (172 observations deleted due to missingness)
AIC: 515.58

Number of Fisher Scoring iterations: 6


> #Replacing missing values with KNN Method of imputation
> #Pagkage VIM
> data2 <- kNN(data1,variable = c("Cholesterol","RestingBP"))
> #write.csv(data2,"/Volumes/STRANGER/6th Sem Project/Swapnil//Imputed
Data.csv",row.names=FALSE)
>
> data2 <- subset(data2,select = -c(Cholesterol_imp,RestingBP_imp))
> head(data2)
  Age Sex ChestPainType RestingBP Cholesterol FastingBS RestingECG MaxHR
1  40   M           ATA       140         289         0     Normal   172
2  49   F           NAP       160         180         0     Normal   156
3  37   M           ATA       130         283         0         ST    98
4  48   F           ASY       138         214         0     Normal   108
5  54   M           NAP       150         195         0     Normal   122
6  39   M           NAP       120         339         0     Normal   170
  ExerciseAngina Oldpeak ST_Slope HeartDisease
1              N     0.0       Up            0
2              N     1.0     Flat            1
3              N     0.0       Up            0
4              Y     1.5     Flat            1
5              N     0.0       Up            0
6              N     0.0       Up            0
> model2<-
glm(data2$HeartDisease~data2$Age+data2$Sex+data2$ChestPainType+data2$RestingBP+data2$Chol
esterol+data2$FastingBS+data2$RestingECG+data2$MaxHR+data2$ExerciseAngina+data2$Oldpeak+d
ata2$ST_Slope,family = binomial)
>
> summary(model2)

Call:
glm(formula = data2$HeartDisease ~ data2$Age + data2$Sex + data2$ChestPainType +
    data2$RestingBP + data2$Cholesterol + data2$FastingBS + data2$RestingECG +
    data2$MaxHR + data2$ExerciseAngina + data2$Oldpeak + data2$ST_Slope,
    family = binomial)

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)            -2.272021   1.483028  -1.532 0.125519
data2$Age               0.019284   0.013132   1.468 0.141979
data2$SexM              1.655037   0.280454   5.901 3.61e-09 ***
data2$ChestPainTypeATA -1.911300   0.324334  -5.893 3.79e-09 ***
data2$ChestPainTypeNAP -1.613028   0.259083  -6.226 4.79e-10 ***
data2$ChestPainTypeTA  -1.471749   0.429887  -3.424 0.000618 ***
data2$RestingBP         0.002163   0.006154   0.352 0.725193
data2$Cholesterol       0.002875   0.001996   1.440 0.149811
data2$FastingBS         1.323663   0.267099   4.956 7.21e-07 ***
data2$RestingECGNormal  0.028619   0.267590   0.107 0.914827
data2$RestingECGST      0.023513   0.344347   0.068 0.945560
data2$MaxHR            -0.007709   0.004899  -1.573 0.115619
data2$ExerciseAnginaY   0.829775   0.241383   3.438 0.000587 ***
data2$Oldpeak           0.364049   0.115733   3.146 0.001658 **
data2$ST_SlopeFlat      1.240525   0.427128   2.904 0.003680 **
data2$ST_SlopeUp       -1.099357   0.446195  -2.464 0.013745 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 1262.14  on 917  degrees of freedom
Residual deviance:  606.82  on 902  degrees of freedom
AIC: 638.82

Number of Fisher Scoring iterations: 5


> #Splitting the Dataset for Males and Females
> dataM <- subset(data2, Sex == "M", select = -c(Sex))
> dataF <- subset(data2, Sex == "F", select = -c(Sex))
>
> #Fitting model Separately for males and Females
> modelM <-
glm(dataM$HeartDisease~dataM$Age+dataM$ChestPainType+dataM$RestingBP+dataM$Cholesterol+da
taM$FastingBS+dataM$RestingECG+dataM$MaxHR+dataM$ExerciseAngina+dataM$Oldpeak+dataM$ST_Sl
ope,family = binomial)
> summary(modelM)

Call:
glm(formula = dataM$HeartDisease ~ dataM$Age + dataM$ChestPainType +
    dataM$RestingBP + dataM$Cholesterol + dataM$FastingBS + dataM$RestingECG +
    dataM$MaxHR + dataM$ExerciseAngina + dataM$Oldpeak + dataM$ST_Slope,
    family = binomial)

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)             0.737028   1.634009   0.451  0.65195
dataM$Age               0.017688   0.014850   1.191  0.23361
dataM$ChestPainTypeATA -2.000647   0.369975  -5.408 6.39e-08 ***
dataM$ChestPainTypeNAP -1.659005   0.289863  -5.723 1.04e-08 ***
dataM$ChestPainTypeTA  -1.394707   0.466213  -2.992  0.00278 **
dataM$RestingBP        -0.004641   0.007043  -0.659  0.50993
dataM$Cholesterol       0.002672   0.002425   1.102  0.27065
dataM$FastingBS         1.086024   0.287715   3.775  0.00016 ***
dataM$RestingECGNormal -0.081983   0.307360  -0.267  0.78967
dataM$RestingECGST     -0.022215   0.377044  -0.059  0.95302
dataM$MaxHR            -0.010733   0.005447  -1.970  0.04879 *
dataM$ExerciseAnginaY   0.814001   0.273943   2.971  0.00296 **
dataM$Oldpeak           0.393343   0.131323   2.995  0.00274 **
dataM$ST_SlopeFlat      1.496766   0.456971   3.275  0.00106 **
dataM$ST_SlopeUp       -0.885424   0.471832  -1.877  0.06058 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 954.15  on 724  degrees of freedom
Residual deviance: 485.44  on 710  degrees of freedom
AIC: 515.44

Number of Fisher Scoring iterations: 5

>
> modelF <-
glm(dataF$HeartDisease~dataF$Age+dataF$ChestPainType+dataF$RestingBP+dataF$Cholesterol+da
taF$FastingBS+dataF$RestingECG+dataF$MaxHR+dataF$ExerciseAngina+dataF$Oldpeak+dataF$ST_Sl
ope,family = binomial)
> summary(modelF)

Call:
glm(formula = dataF$HeartDisease ~ dataF$Age + dataF$ChestPainType +
    dataF$RestingBP + dataF$Cholesterol + dataF$FastingBS + dataF$RestingECG +
    dataF$MaxHR + dataF$ExerciseAngina + dataF$Oldpeak + dataF$ST_Slope,
    family = binomial)

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)            -7.702274   3.731602  -2.064 0.039011 *
dataF$Age               0.021907   0.032087   0.683 0.494768
dataF$ChestPainTypeATA -1.795796   0.740998  -2.423 0.015372 *
dataF$ChestPainTypeNAP -1.589162   0.692359  -2.295 0.021717 *
```

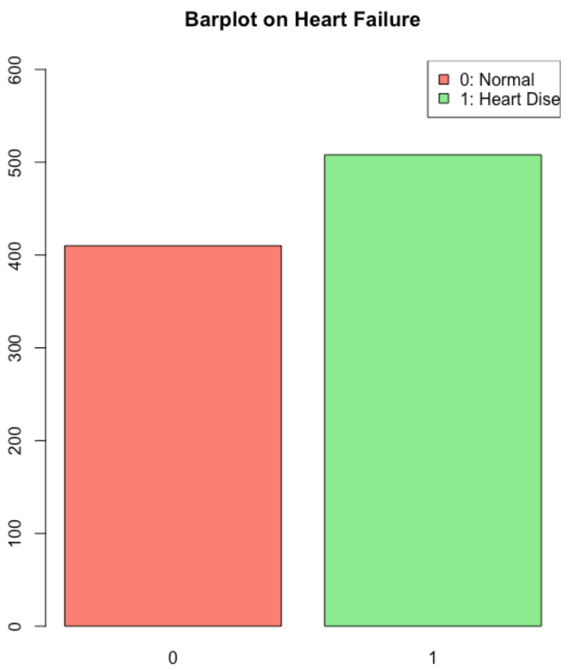
```
dataF$ChestPainTypeTA  -3.234036   1.642962  -1.968 0.049020 *
dataF$RestingBP         0.032814   0.015983   2.053 0.040071 *
dataF$Cholesterol       0.003116   0.004045   0.770 0.441146
dataF$FastingBS         3.225450   0.878498   3.672 0.000241 ***
dataF$RestingECGNormal  0.684054   0.623873   1.096 0.272876
dataF$RestingECGST      0.537035   1.072001   0.501 0.616396
dataF$MaxHR             0.007272   0.013401   0.543 0.587383
dataF$ExerciseAnginaY   0.452371   0.617702   0.732 0.463957
dataF$Oldpeak           0.264202   0.318308   0.830 0.406527
dataF$ST_SlopeFlat     -0.128953   1.508131  -0.086 0.931860
dataF$ST_SlopeUp       -3.031566   1.711327  -1.771 0.076482 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 220.82  on 192  degrees of freedom
Residual deviance: 105.49  on 178  degrees of freedom
AIC: 135.49

Number of Fisher Scoring iterations: 6


> #finding percentage of heart failure in each sex
>
> #Males
> # Count the number of heart failure cases (1) in the "dataM" dataset
> heart_failure_count <- sum(dataM$HeartDisease == 1)
>
> # Calculate the total number of observations in the "dataM" dataset
> total_observations <- nrow(dataM)
>
> # Calculate the percentage of heart failure in "dataM"
> heart_failure_percentage_M <- (heart_failure_count / total_observations) * 100
>
> round(heart_failure_percentage_M, 2)
[1] 63.17


>
> #Females
> # Count the number of heart failure cases (1) in the "dataF" dataset
> heart_failure_count <- sum(dataF$HeartDisease == 1)
>
> # Calculate the total number of observations in the "dataF" dataset
> total_observations <- nrow(dataF)
>
> # Calculate the percentage of heart failure in "dataM"
> heart_failure_percentage_F <- (heart_failure_count / total_observations) * 100
>
> round(heart_failure_percentage_F, 2)
[1] 25.91



> ## Percentage Calculation
>
> #Heart Failure %
> # Count the number of people with heart failure (HeartDisease = 1)
> num_heart_failure <- sum(data2$HeartDisease == 1)
>
> # Calculate the total number of people in the dataset
> total_people <- nrow(data2)
>
> # Calculate the percentage of people with heart failure
> percentage_heart_failure <- (num_heart_failure / total_people) * 100
>
> round(percentage_heart_failure, 2)
[1] 55.34
```

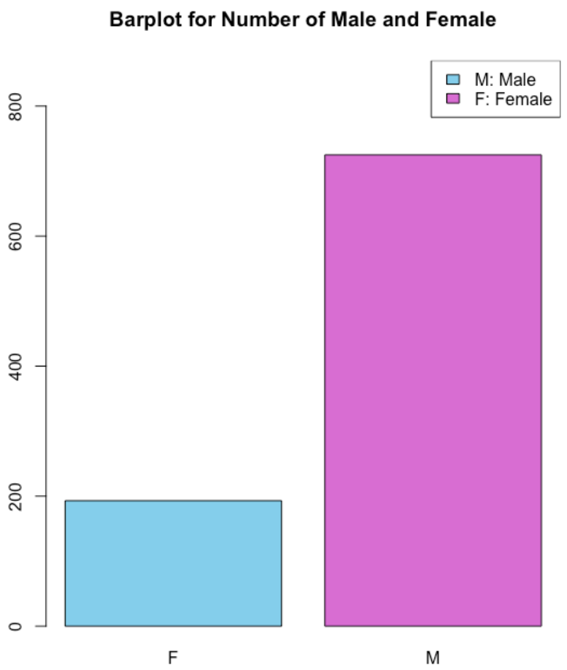
```
> #Sex %
> # Count the number of males and females
> num_males <- sum(data2$Sex == "M")
> num_females <- sum(data2$Sex == "F")
>
> # Calculate the total number of observations
> total_observations <- nrow(data2)
>
> # Calculate the percentage of males and females
> percentage_male <- (num_males / total_observations) * 100
> percentage_female <- (num_females / total_observations) * 100
>
> round(percentage_male, 2)
[1] 78.98


> round(percentage_female, 2)
[1] 21.02



>
>
> #Chest Pain Type
> # Calculate the frequency of each type of Chest Pain Type
> chest_pain_freq <- table(data2$ChestPainType)
>
> # Calculate the total number of observations in the dataset
> total_observations <- nrow(data2)
>
> # Calculate the percentage of each type of Chest Pain Type
> chest_pain_percentage <- chest_pain_freq / total_observations * 100
>
> # Print the results
> chest_pain_percentage

      ASY       ATA       NAP        TA 
54.030501 18.845316 22.113290  5.010893 


>
>
> #fastingBS
> # Calculate percentage observations in each category of Fasting Blood Sugar
> percentage_fasting_bs <- table(data2$FastingBS) / length(data2$FastingBS) * 100
>
> # Print the percentage observations for each category
> percentage_fasting_bs

       0        1 
76.68845 23.31155 


>
> #Resting ECG
> # Calculate the percentage of observations in each category of Resting ECG
> resting_ecg_percentages <- round(prop.table(table(data2$RestingECG)) * 100, 2)
>
> # Print the percentages
> resting_ecg_percentages

   LVH Normal     ST 
 20.48  60.13  19.39 


>
> #Exercise Angina
> # Calculate the percentage of observations in each category of Exercise Angina
> percentage_exercise_angina <- prop.table(table(data2$ExerciseAngina)) * 100
>
> # Print the results
> percentage_exercise_angina
```

```
       N        Y 
59.58606 40.41394 


> 
> #ST Slope
> # Calculate the percentage of observations in each category of ST Slope
> percentage_st_slope <- prop.table(table(data2$ST_Slope)) * 100
> 
> # Display the percentages
> percentage_st_slope

     Down      Flat        Up 
 6.862745 50.108932 43.028322 
```

Appendix 3

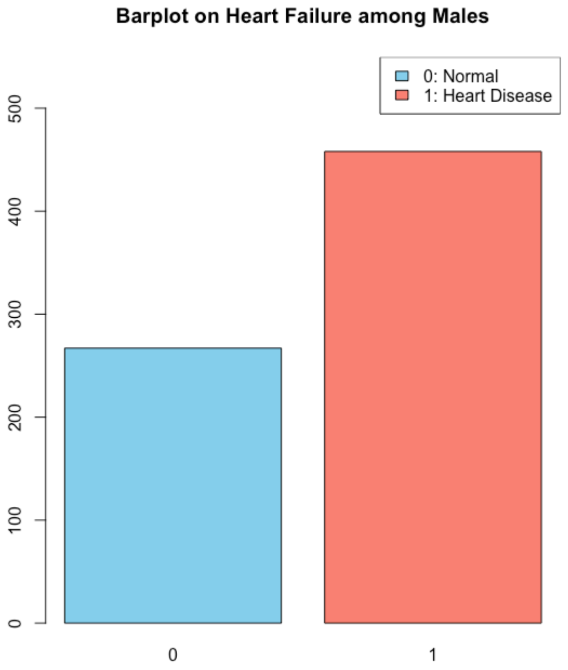
Graphical Representations -

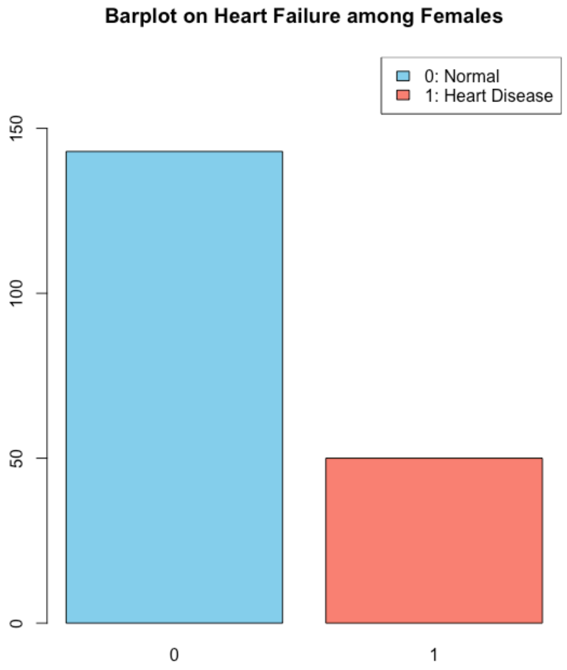




Graph 1

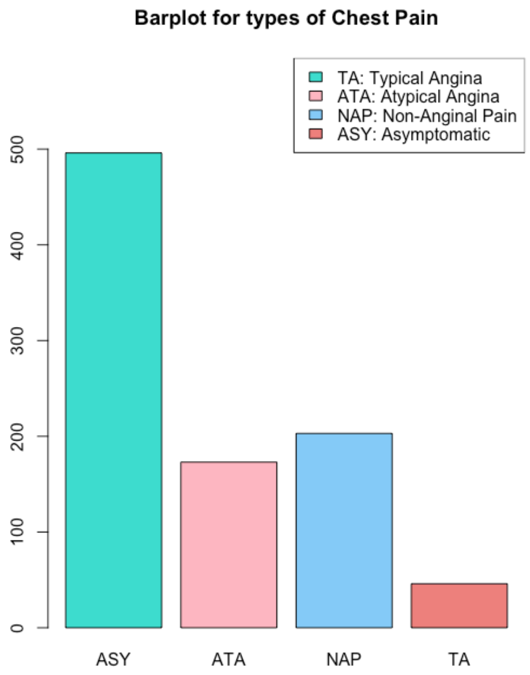
Graph 2



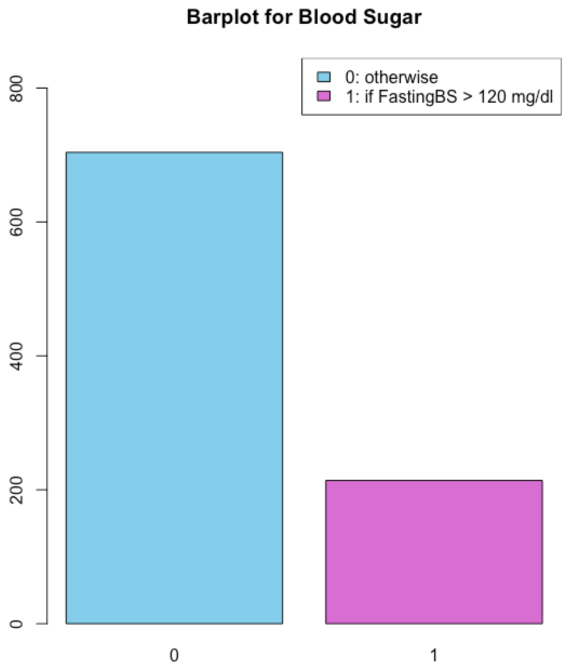


Graph 3

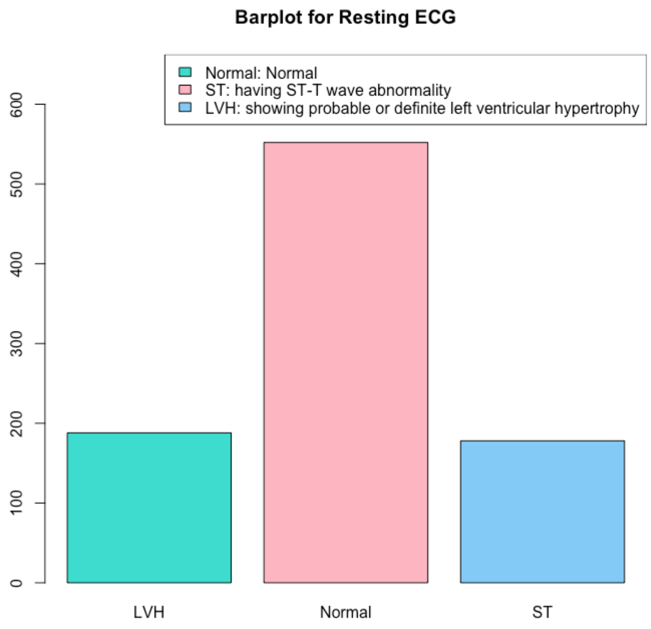
Graph 4



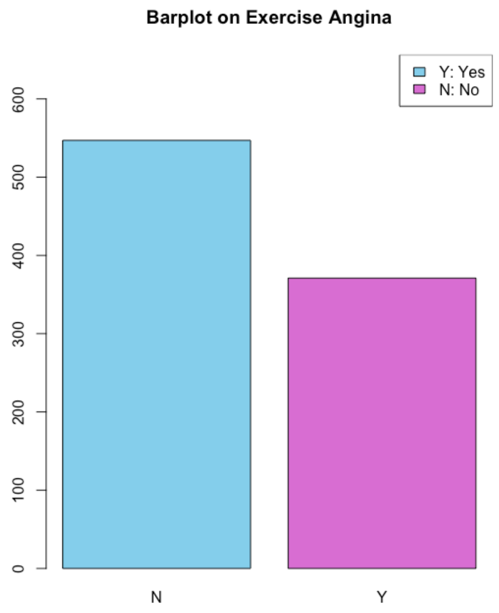
Graph 5



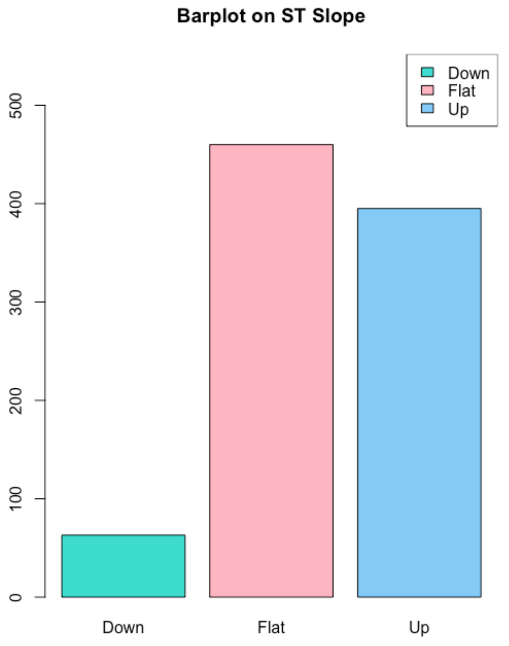
Graph 6



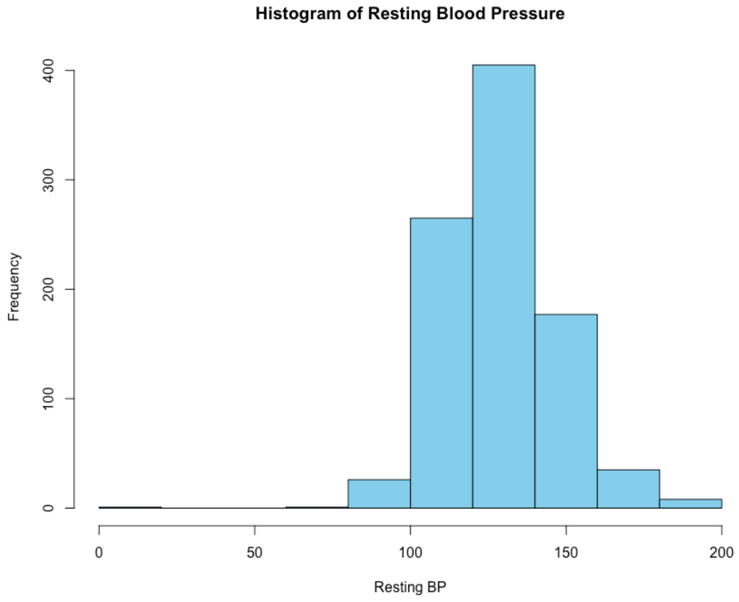
Graph 7



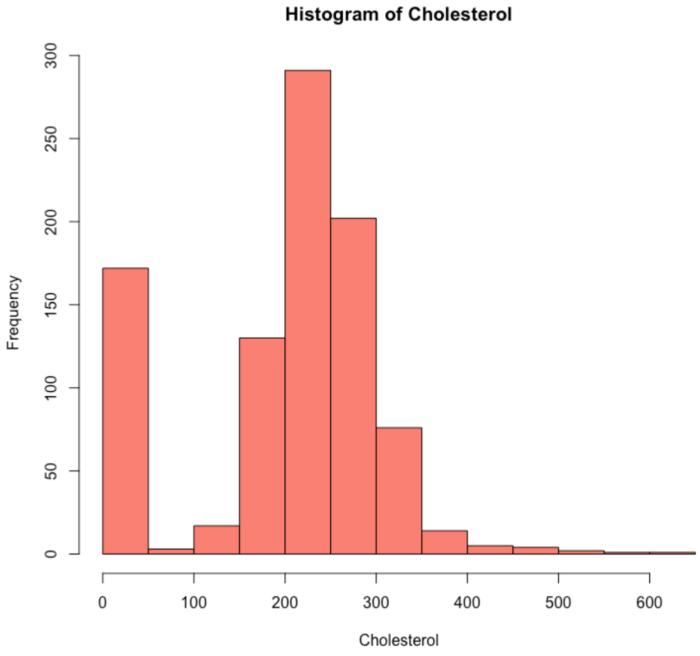
Graph 8



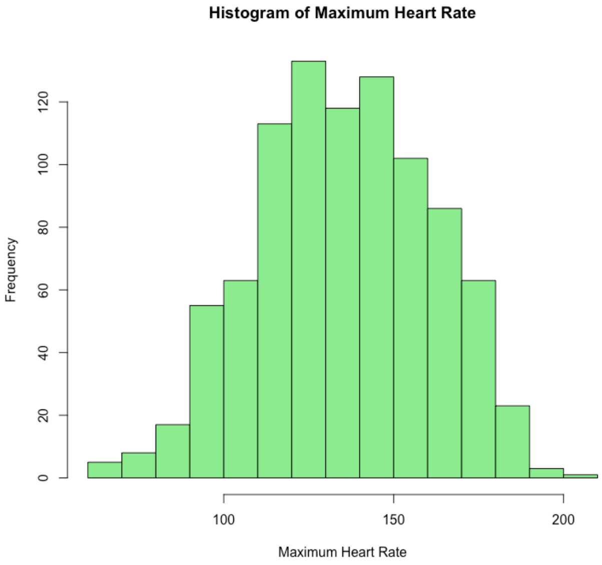
Graph 9



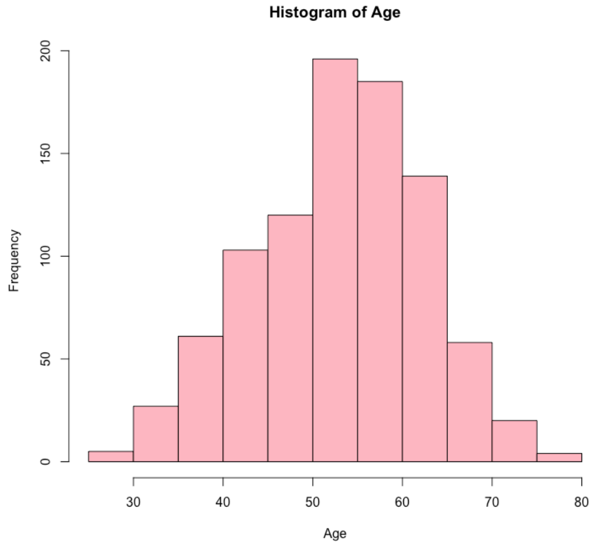
Graph 10



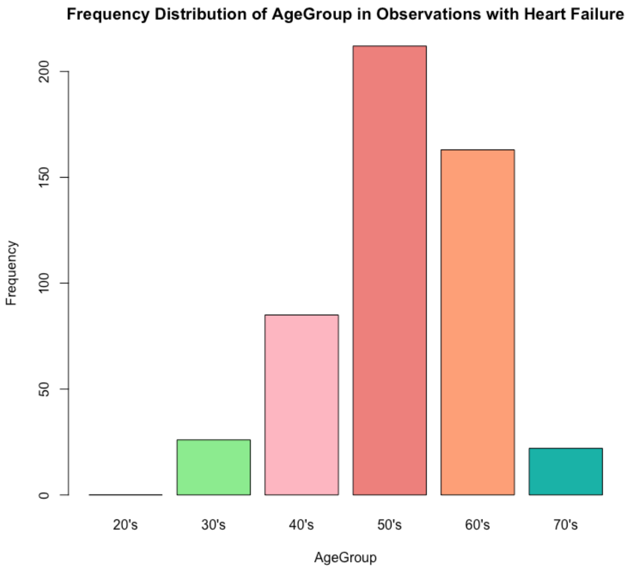
Graph 11



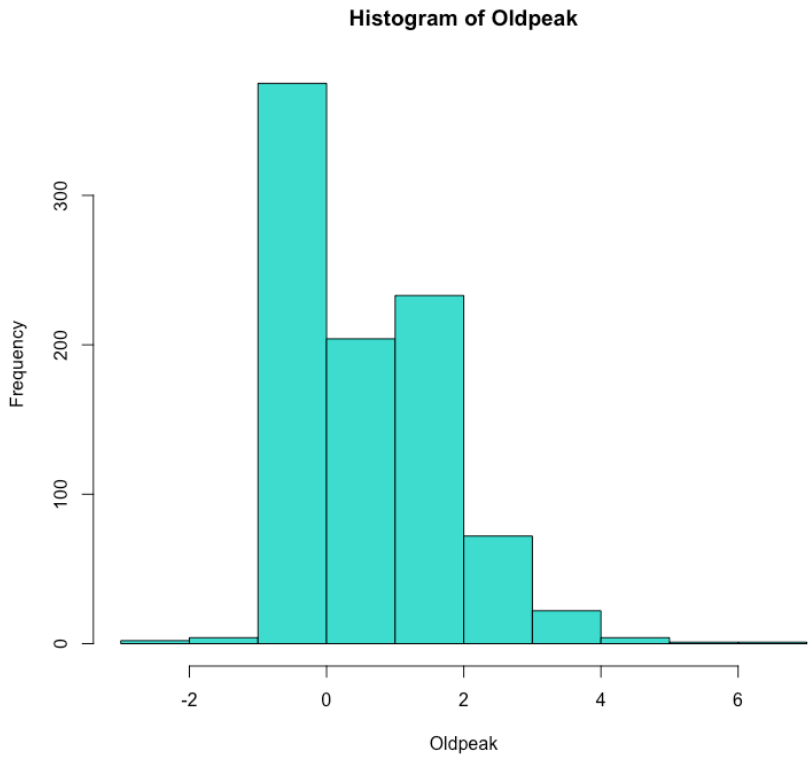
Graph 12



Graph 13



Graph 14



Graph 15

Gratitude

I would like to express my heartfelt gratitude to all those who have supported and guided me throughout this project. Without their valuable assistance and encouragement, completing this endeavour would not have been possible.

First and foremost, I extend my sincere thanks to **Professor Soumyadeep Das** for his unwavering guidance, expertise, and continuous motivation. His insightful feedback and constructive criticism have been instrumental in shaping the direction of this project.

I am also deeply indebted to Professor **Dr. Kiranmoy Chatterjee** for his invaluable insights and scholarly advice. His profound knowledge and expertise in the subject have enriched my understanding and significantly contributed to the quality of this work.

My heartfelt appreciation goes to **Professor Suryasish Chatterjee** for his constant support and encouragement. His dedication to nurturing students' intellectual growth has been a great source of inspiration for me.

I would like to acknowledge the support and friendship of my fellow classmates and friends. Their camaraderie, discussions, and exchange of ideas have been a source of encouragement and positivity throughout this project.

Lastly, I am grateful to all the professors and faculty members of my department at Bidhannagar College, Kolkata. Their commitment to imparting knowledge and fostering a conducive learning environment has been pivotal in shaping my academic journey.

In conclusion, I extend my deepest gratitude to all the individuals mentioned above and everyone else who has played a role in supporting me during this project. Your contributions have been invaluable, and I am honoured to have had the opportunity to work with such remarkable mentors and friends.