A Predictive Model for Heart Attack Risk using EDA and Regression Analysis

Trang Nguyen
Mercer University

May 6, 2023

1. Introduction

A heart attack, also known as a myocardial infarction, is a major contributor to cardiovascular disease, which is one of the leading causes of death worldwide at all ages. A heart attack typically occurs when a blood clot built-up inside the walls of coronary arteries and obstructs blood flow to the heart muscle, resulting oxygen and nutrients supply to the heart muscle cells severely reduced. As a result, the affected muscle cells begin to die, and the heart can become weakened and even damaged, leading to potentially life-threatening complications[1]. According to the American Heart Association, a heart attack occurs every 40 seconds in the United States. About 805,000 people has a heart attack every year in the United States. An estimated 605,000 of these are a first-time heart attacks, and about 1 in 5 heart attacks are silent, meaning that damage to the heart muscles has occurred without any noticeable symptoms of heart attack[2]. The purpose of this study is to analyze major factors as well as to determine their relationships with the risk of heart attack to prevent and reduce the incidence of heart attack.

2. Descriptive Analysis

2.1 Data Description

Two datasets are involved in this study obtained from “Kaggle: Heart Attack Analysis & Prediction Dataset”. The first data set was collected using various medical tests and procedures. The data contains 14 attributes from 303 patients including age, gender, chest pain type, resting blood pressure, cholesterol level, fasting blood sugar level, resting electrocardiographic results, maximum heart rate achieved during exercise, exercise-induced angina (AP), ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels, Thallium stress test result, and diagnosis result of heart attack. The database is downloaded in .xlsx format with a file size of 28 KB. The second data collection contains information of oxygen saturation measurements obtained from 3586 observations. There is only one variable in this dataset, which is oxygen saturation measurements. The file is downloaded in .xlsx format with a file size of 32 KB. The target variables in these data collections are categorical variables and numerical variables, which are used to predict the risk of heart attack. To ensure the accuracy and reliability of the results, pre-processing steps are considered important steps in this analysis, which involve cleaning the data, removing any missing, erroneous values and outliers, and transforming raw databases into proper formats for analysis.

2.2 Descriptive Statistics of Study Variables

Table 1 shows the summary of descriptive statistics of the variables in the study, including Age, Resting Blood Pressure (mm Hg), Cholesterol Level (mg/dL), Maximum Heart Rate During Exercise, and Oxygen Saturation Measurements (%). The minimum, maximum, mean, standard deviation, the first quartile and the third quartile values are displayed for each variable. There were 303 patients participated in this study.

According to Table 1, the age of the participants ranged from 29 to 77 years old, with an average of 54.37 years old (SD = 9.08 years). The resting blood pressure of the participants varied between 94 mm Hg and 200 mm Hg, with an average of 131.60 mm Hg (SD = 17.54 mm Hg). The cholesterol level of the participants was widely different from 126 to 546 mg/dL, with a mean of 246.30 mg/dL and a standard deviation of 51.83 mg/dL. The average heart rate of the participants achieved during exercise was 149.60 bpm (beats per minute) with a standard deviation of 22.91 bpm. The lowest heart rate recorded during the exercise was 71.00 bpm, while the highest heart rate achieved in this exercise was 202.00 bpm. The oxygen saturation measurements of the participants ranged from 96.5% to 99.6%, with a mean of 98.24% (SD = 0.73 %).

2.3 Data Exploratory Analysis

In this study, exploratory data analysis (EDA) is applied to analyze and investigate the datasets as well as their characteristics. In the data analysis process, it is an important step to help to identify patterns, and relationships in the data. The purpose of EDA is to gain insights into the data, formulate hypothesis, and guide further analysis[4]. According to Figure 1, the distribution of continuous variables in the datasets including age, resting blood pressure, cholesterol level, heart rate, oxygen saturation measurements vary. Shapiro-Wilk normality test is also utilized in this study to access the normality of the datasets. Table 2 shows that the test statistic value of age is 0.98637, indicating that the data is slightly deviated from normality. However, the p-value of the test is 0.005798, which is larger than the critical value 0.05. It is concluded that the variable of age in this dataset is normally distributed. By contrast, although test statistic value of distribution of resting blood pressure, cholesterol level, heart rate,

oxygen saturation measurements in the datasets is around 0.9, meaning that the data being slightly deviated from normality, these variables' p-value of the test is smaller than the critical value 0.05. Therefore, the variables of resting blood pressure, cholesterol level, heart rate, oxygen saturation measurements are not normally distributed in the datasets. To identify outliers in the data, box plot is also used as part of the exploratory data analysis (EDA) in this study. As seen in Figure 2, there are some outliers identified in the plots related to the measurements of resting blood pressure, cholesterol level, and heart rate. The outliers can affect the normal distribution of the data, the accuracy and reliability of the results.

3. Hypothesis Test

According to Centers for Disease Control and Prevention (CDC), heart attack can affect everyone at all ages. As seen in Figure 3, the diagnosis result of heart attack is displayed by different age groups ranging from 29 to 77 years old. Based on the data collection, out of 303 participants, 165 people have a history of heart attack, with 72 of them being female and 93 being males, as seen in Figure 4 and Table 2. 39 people reported that when they had a heart attack, they suffered typical angina, which is a substernal chest pain in the center of chest occurred due to physical exertion or emotional stress[5]. 41 people had atypical angina, which refers to a pressure like sensation in the chest that arises when the heart muscle does not receive an adequate oxygenated blood supply[6]. A considerable number of 69 individuals reported that they experienced non-anginal pain, a feeling of having discomfort or pain in the chest, and there are16 participants reported having a heart attack experienced asymptomatic (Figure 5). However, of 303 individuals participated in this study, Figure 6 shows that

138 individuals didn't have any past history of heart attack. Among of the remaining participants,104 people reported that they suffered a typical angina, 9 people had atypical angina, 18 people experienced non-anginal pain, and 7 individuals felt asymptomatic even though they did not experience a heart attack. It is concluded that chest pain is a common symptom of heart attack. Among chest pain types, non-angina pain is the most commonly reported type of chest pain in this group, followed by atypical angina, typical angina, and asymptomatic. However, the presence of chest pain symptom alone may not indicate the presence of a heart attack. Therefore, hypothesis testing is involved in this study for further research to confirm a diagnosis.

3.1 Hypothesis Test Performance between Two Population Means of Resting Blood Pressure

Null hypothesis: There is no difference between the average of resting blood pressure between those who had a heart attack and those who did not.

$$H_0: \mu_{\text{bp\_heart attack}} \neq \mu_{\text{bp\_w/o heart attack}}$$

Alternative hypothesis: There is a difference between the average of resting blood pressure between those who had a heart attack and those who did not.

$$H_1: \mu_{\text{bp\_heart attack}} = \mu_{\text{bp\_w/o heart attack}}$$

The t-statistic for testing the Null hypothesis is performed by following the formula:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

According to Table 3, with a degree of freedom 272.56, at alpha level 0.05, the critical value is 1.6505. The absolute value of t-statistic is 2.5083, which is larger than the critical value 1.6505. We reject the Null hypothesis at alpha level 0.05. Therefore, there is a statistically significant difference between the average of resting blood pressure between those who had a heart attack and those who did not. A 95% confidence interval for the difference between the average of resting blood pressure between those who had a heart attack and those who did not is between 129.303 mm Hg and 134.399 mm Hg. There is a significant difference between two groups.

3.2 Hypothesis Test Performance between Two Population Means of Cholesterol Level

Null hypothesis: There is no difference between the average of cholesterol level between those who had a heart attack and those who did not.

$$H_0: \mu_{\text{cl\_heart attack}} \neq \mu_{\text{cl\_w/o heart attack}}$$

Alternative hypothesis: There is a difference between the average of cholesterol level between those who had a heart attack and those who did not.

$$H_1: \mu_{\text{cl\_heart attack}} = \mu_{\text{cl\_w/o heart attack}}$$

At alpha level 0.05, with a degree of freedom 298.03, the critical value is 1.65. Table 3 shows that the absolute value of t-statistic is 1.4947, indicating the t-statistic value being smaller than the critical value 1.65. We fail to reject the Null hypothesis at alpha level 0.05. Therefore, there is no difference between the average of cholesterol level between those who had a heart attack and those who did not. A 95% confidence interval for the difference between the average of cholesterol level between those who had a heart attack and those who did not is between -20.517 mg/dL and 2.803 mg/dL.

Since confidence interval for the difference between two groups includes zero, meaning that there is no difference between the average of cholesterol level between those who had a heart attack and those who did not.

3.3 Hypothesis Test Performance between Two Population Means of Heart Rate

Null hypothesis: There is no difference between the mean of heart rate between those who had a heart attack and those who did not.

$$H_0: \mu_{\text{hrate\_heart attack}} \neq \mu_{\text{hrate\_w/o heart attack}}$$

Alternative hypothesis: There is a difference between the mean of heart rate between those who had a heart attack and those who did not.

$$H_1: \mu_{\text{hrate\_heart attack}} = \mu_{\text{hrate\_w/o heart attack}}$$

At alpha level 0.05, with a degree of freedom 269.9, the critical value is 1.6505. As seen in table 3, the t-statistic value is 9.953, which is larger than the critical value 1.6505. The Null hypothesis is rejected at alpha level 0.05. As a result, there is a statistically significant difference between the mean of heart rate between those who had a heart attack and those who did not. A 95% confidence interval for the difference between the mean of heart rate between those who had a heart attack and those who did not ranged from 14.571 bpm to 24.159 bpm that illustrates a statistically significant difference between the mean of two groups.

4. Regression Analysis

A heart attack occurs when the flow of blood bringing oxygen to the heart muscle severely reduced. Thallium stress test is a type of nuclear medicine imaging test evaluating how well the blood flows through the heart muscle during exercise and at rest. According to the hypothesis testing performance, it is indicated that there is a

statistically significant difference between the average of resting blood pressure and heart rate between people had a heart attack and those who did not. To further investigate heart attack risk prediction, liner regression model is utilized as a continuously statistical research methodology to determine the relationship between variables of blood pressure, heart rate and the Thallium stress test results. The dependent variables in this study are blood pressure and heart rate. The independent variable is thallium stress test result with 4 categories. The Thallium test result denotes normal blood flow as 0, mildly reduced blood flow with some restriction present as 1, moderately reduced blood flow with a significant restriction present as 2, and severely reduced blood flow with a severe restriction present as 3.

4.1 Linear Regression Model between Blood Pressure and the Thallium Stress Test Results

Regression Model: $\text{Blood Pressure} = \beta_0 + \beta_1 * D_i + u_i$

Where D_i is dummy variable, and D_i is defined as follow:

$$D_i = \begin{cases} 0 \text{ if normal blood flow} \\ 1 \text{ if mildly reduced blood flow} \\ 2 \text{ if moderately reduced blood flow} \\ 3 \text{ if severely reduced blood flow} \end{cases}$$

Table 4 shows model ummary of linear regression analysis and regression coefficients of blood pressure on 4 categories of Thallium stress test results. As seen on the table, the p-value for the coefficients of “Severely reduced blood flow”, “Moderately reduced blood flow”, and “Normal blood flow” is 0.5113, 0.0816, and 0.4917, respectively, which is greater than 0.05. It indicates that there is no association between

the categories of Thallium stress test results and blood pressure. In addition, the adjusted R-squared is 0.01228, indicating 1.228% of the variation in blood pressure is explained by the Thallium stress test results. With F-statistic value 2.251 and p-value 0.08245, it is concluded that the overall regression model is not significant. It's assumed that the exclusion of "Mildly reduced blood flow" can lead to omitted variable bias in the regression model.

4.2 Linear Regression Model between Heart Rate and the Thallium Stress Test Results

Regression Model: $\text{Heart Rate} = \beta_0 + \beta_1 * D_i + u_i$

According to Table 5, the p-value for the coefficients of "Normal blood flow" and "Severely reduced blood flow" is 0.981015, and 0.134443, respectively, which is higher than 0.05, indicating that there is no association between the categories of Thallium stress test results and blood pressure. However, there is a correlation between moderately reduced blood flow and heart rate (p-value 0.000175 < 0.05), meaning that for every unit increase in moderately reduced blood flow, the heart rate is expected to increase by nearly 21 beats per minutes on average. The adjusted R-squared in the model is 0.08522, meaning that the Thallium stress test result of moderately reduced blood flow can explain 8.522% of the variation in heart rate in this model.

5. Conclusions

A heart attack can occur to everyone at all ages even those who are relatively young. Men are more likely to experience heart attacks more than women. The symptom of heart attack is diverse; however, chest pain is a common type of heart attack. According to the findings, there is a statistically significant difference between

the average of resting blood pressure and heart rate between individuals had a heart attack and those who did not. There is also a correlation between moderately reduced blood flow and heart rate. It's concluded that people with high blood pressure and a medical history of cardiovascular disease are at a higher risk of experiencing heart attacks compared to those who don't have these risk factors.

There are some limitations as well as potential weaknesses in the analyses. The presenting of some outliers and the exclusion of a variable in the regression model can affect to the normal distribution of the data, the accuracy and reliability of the results. It's suggested that exploring different statistical techniques and modeling approaches can provide more accurate results and insights for future research. In addition, there are still additional questions related to the topic that could be analyzed for a more comprehensive understanding of the research area such as correlation between diet, exercise, and the risk of heart attack.

Appendix

Table 1. Descriptive Statistics of Continuous Variables in the Datasets

Descriptive Statistics	Age	Resting Blood Pressure (mm Hg)	Cholesterol Level (mg/dL)	Maxium Heart Rate During Exercise	Oxygen Saturation Measurements (%)
Min	29.00	94.00	126.00	71.00	96.50
1st Qu:	47.50	120.00	211.00	133.50	97.60
Median	55.00	130.00	240.00	153.00	98.60
Mean	54.37	131.60	246.30	149.60	98.24
Stdev	9.08	17.54	51.83	22.91	0.73
3rd Qu:	61.00	140.00	274.50	166.00	98.60
Max	77.00	200.00	546.00	202.00	99.60

Table 2. Shapiro-Wilk Normality Test

Variables	Statistic	Sig
Age	0.98637	0.005798
Resting Blood Pressure	0.96592	1.46E-06
Cholesterol Level	0.94688	5.37E-09
Heart Rate	0.97632	6.62E-05
Oxygen Saturation Measurements	0.86173	< 2.2e-16

Figure 1. Histograms of Distribution 5 Continuous Variables of Datasets

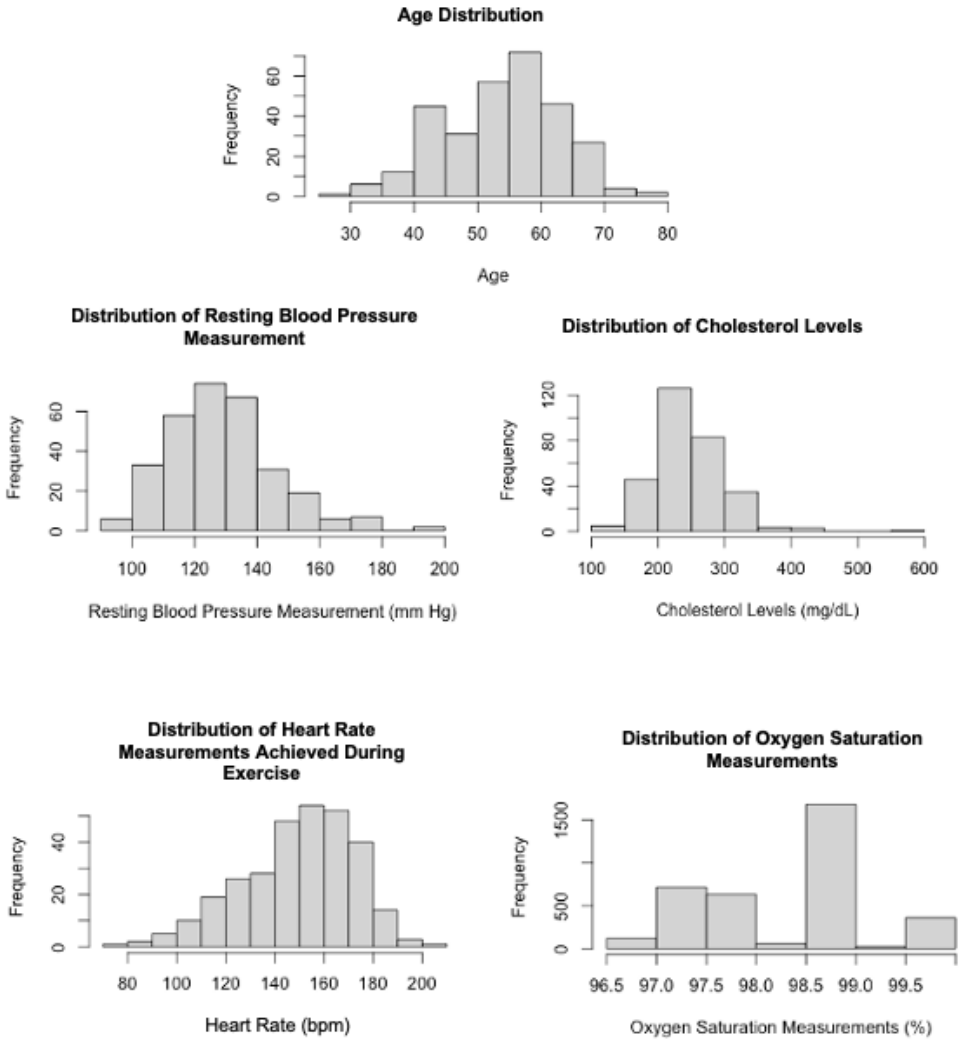


Figure 2. Box Plots of Distribution 5 Continuous Variables of Datasets

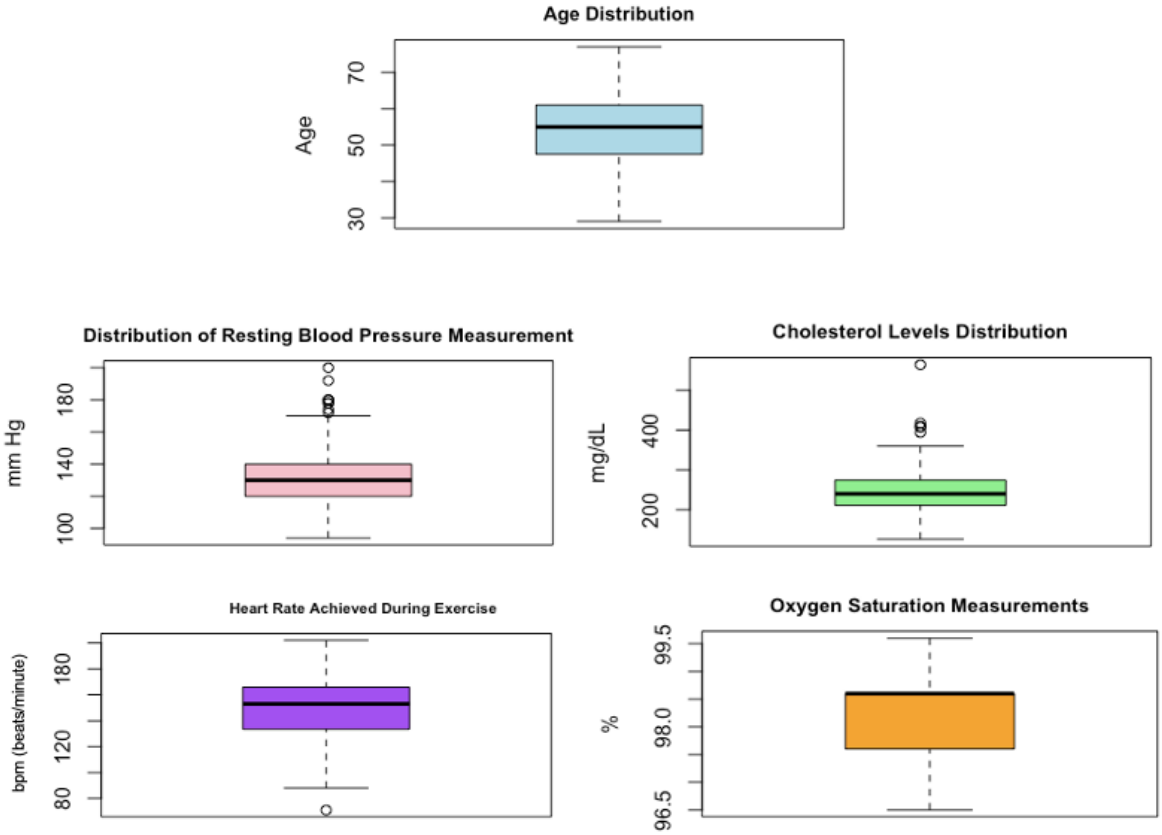


Figure 3. Diagnosis Result of Heart Attack by Age Group

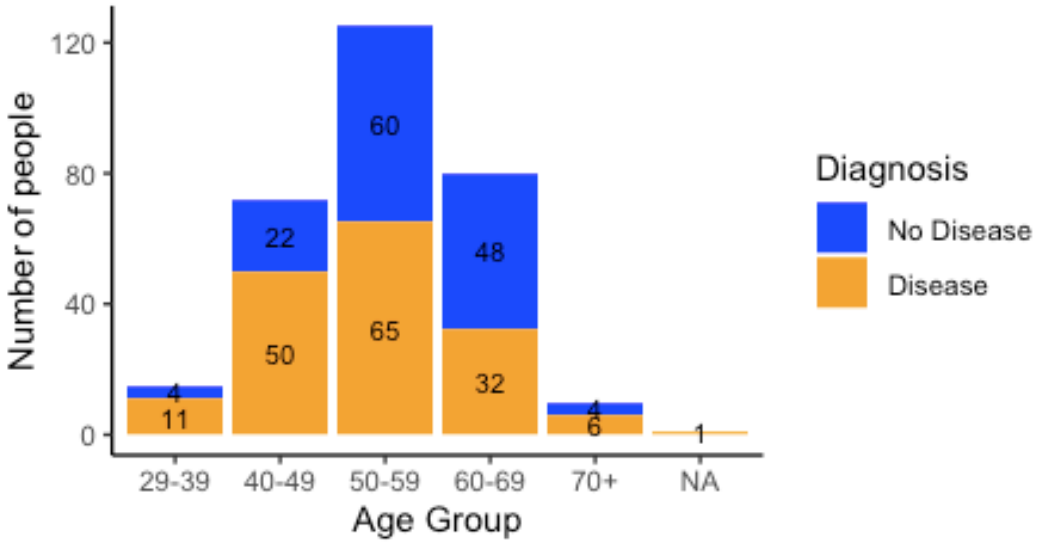


Figure 4. Gender Distribution of Heart Attack Patients

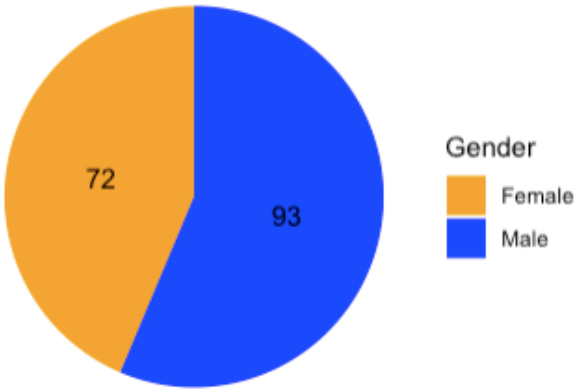


Table 2. Distribution of Heart Attack Cases by Gender

Gender	Number of Heart Attack Cases
Female	72
Male	93

Figure 5. Distribution of Chest Pain Types in Heart Attack Cases

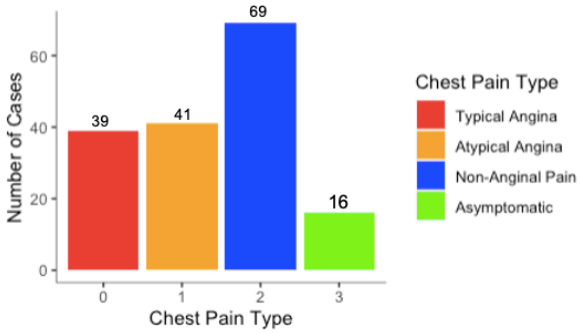
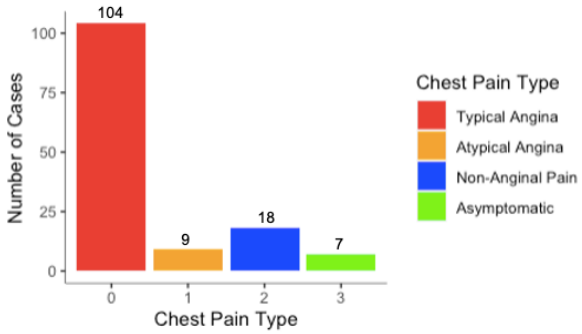


Figure 6. Distribution of Chest Pain Types in No Heart Attack Cases



Hypothesis Test Performance between Two Population Means in R

Table 3. Welch Two Sample t-test between Two Population Means

Variables	$\overline{X}$ *heart* attack	$\overline{X}$ *no heart* attack	df	t-statistic	p-value	[95% Conf. Interval]	
Resting Blood Pressure	129.303	134.399	272.56	-2.5083	0.01271	129.303	134.399
Cholesterol Level	242.230	251.087	298.03	-1.4948	0.136	-20.517	2.803
Heart Rate	158.467	139.101	269.9	7.953	5.02E-14	14.571	24.159

Linear Regression Model

Table 4. Model Summary of Linear Regression Analysis and Regression Coefficients – Blood Pressure

Explanatory Variables	Estimate Regression Coefficients	Std. Error	t value	Pr(>|t|)
(Interncept)	136.944	4.108	33.333	<2e-16
Moderately reduced blood flow	-7.559	4.325	-1.748	0.0816
Normal blood flow	-8.944	12.992	-0.688	0.4917
Severely reduced blood flow	-2.902	4.413	-0.658	0.5113
Residual standard error: 17.43 on 299 degrees of freedom				
Multiple R-squared: 0.02209, Adjusted R-squared: 0.01228				
F-statistic: 2.251 on 3 and 299 DF, p-value: 0.08245				

Table 5. Model Summary of Linear Regression Analysis and Regression Coefficients – Heart Rate

Explanatory Variables	Estimate Regression Coefficients	Std. Error	t value	Pr(>|t|)
(Interncept)	135.1111	5.1636	26.166	< 2e-16
Moderately reduced blood flow	20.6600	5.4364	3.800	0.000175
Normal blood flow	0.3889	16.3288	0.024	0.981015
Severely reduced blood flow	8.3248	5.5466	1.501	0.134443
Residual standard error: 21.91 on 299 degrees of freedom				
Multiple R-squared: 0.09431, Adjusted R-squared: 0.08522				
F-statistic: 10.38 on 3 and 299 DF, p-value: 1.621e-06				

References

1. Heart Health and Aging https://www.nia.nih.gov/health/heart-health-and-aging

2. Tsao CW, Aday AW, Almarzooq ZI, Beaton AZ, Bittencourt MS, Boehme AK, et al. Heart Disease and Stroke Statistics—2022 Update: A Report From the American Heart Association. *Circulation.* 2022;145(8):e153–e639.

3. What is exploratory data analysis? https://www.ibm.com/topics/exploratory-data-analysis

4. AlBadri A, Leong D, Bairey Merz CN, Wei J, Handberg EM, Shufelt CL, Mehta PK, Nelson MD, Thomson LE, Berman DS, Shaw LJ, Cook-Wiens G, Pepine CJ. Typical angina is associated with greater coronary endothelial dysfunction but not abnormal vasodilatory reserve. Clin Cardiol. 2017 Oct;40(10):886-891. doi: 10.1002/clc.22740. Epub 2017 Jun 12. PMID: 28605043; PMCID: PMC5680106.

5. Angina Pectoris (Stable Angina) https://www.heart.org/en/health-topics/heart-attack/angina-chest-pain/angina-pectoris-stable-angina#:~:text=Angina%20usually%20causes%20uncomfortable%20pressure,'t%20related%20to%20angina