# Heart Disease Project 1

Crystal Sawtelle

STA-486C

## Introduction

Heart disease affects hundreds of thousands of people a year and is the leading cause of death in the United States. There are several lifestyle choices and medical conditions that can put you at risk for heart disease including diabetes, unhealthy diet, lack of physical activity, overweight and obesity, smoking, etc. Another major risk of having heart disease is genetics CDC (2022). My grandfather died of a heart attack, my father had quadruple by-pass surgery, and my mother just recently had a heart attack. I chose the Heart Disease UCI to learn a little more about the different metrics doctors use to determine if you are at risk of having heart disease “Heart Disease UCI” (n.d.). Personally, I know that I am at risk of developing heart disease and my risk increases as I get older. Analyzing this data will help in understanding the tests and other symptoms to look out for.

## Data Description

As seen in Table 1 there are 13 variables and 302 observations in the Heart Disease UCI data set retrieved from Kaggle.com “Heart Disease UCI” (n.d.). There are five quantitative variables including age, resting blood pressure, cholesterol, maximum heart rate, and oldpeak. The minimum age of an individual in this study is 29 years old and the maximum is 76. The resting blood pressure numbers represents the systolic pressure (top number) when reading blood pressure. The range of the observations fall between 94 and 200 with anything above 130 to 140 indicating a cause for concern. The cholesterol levels are derived from the formula serum = LDL + HDL + 0.2 * triglycerides. Where LDL stands for low-density lipoprotein and is typically consider the “bad cholesterol.” HDL stands fro high-density lipoprotein or the “good cholesterol.” HDL absorbs cholesterol in the blood and carries it back to the liver to be flushed from the body. Triglycerides is a type of fat in you blood that your body uses for energy. Low-levels of LDL and/or low-levels HDL with high levels of triglycerides increases the risk of health problems, like a heart attack. If the serum is greater than 200 this is typically cause for concern. The maximum heart rate numbers show a range between 71 and 202. People with a maximum of over 140 are more at risk of having heart disease. Old peak measures exercise-induced ST depression versus the heart at rest, a unhealthy heart will stress more.

**Table 1:**

| Variable | Type | Description |
|---|---|---|
| Age | num | age of patient |
| Gender | factor | 0 = Female<br>1 = Male |
| Chest Pain | factor | chest pain:<br>0 = Typical Angina: chest pain related to decrease blood supply to the heart<br>1 = Atypical Angina: chest pain not related to the heart<br>2 = non-Anginal Pain: typical esophageal spasms (not heart related)<br>3 = Asymptomatic: chest pain not showing signs of heart disease |
| Resting Blood Pressure | num | Resting blood pressure (in mm Hg on admission to the hospital) anything above 130-140 is typically cause for concern |
| Cholesterol | num | Cholesterol: serum cholesterol in mg/dl (milligrams per deciliter)<br>serum = LDL + HDL + .2*triglycerides |

| Variable | Type | Description |
|---|---|---|
| Fasting Blood Sugar | factor | Fasting blood sugar: > 120 mg/dl (milligrams per deciliter)<br>1 = true<br>0 = false |
| Resting ECG | factor | Resting electrocardiographic (EKG or ECG):<br>0 = nothing to note<br>1 = ST-T Wave abnormality (can range from mild symptoms to severe problems, signals non-normal heartbeat)<br>2 = Possible or definite left ventricular hypertrophy (enlarged hearts main pumping chamber) |
| Maximum Heart Rate | num | Maximum heart rate achieved |
| Exercise Induced Angina | factor | Exercise induced angina<br>1 = true<br>0 = false |
| Oldpeak | num | ST depression induced by exercise relative to rest looks at stress of heart during exercise (unhealthy heart will stress more) |
| Slope | factor | The slope of the peak exercise ST segment<br>0 = Upsloping: better heart rate with exercise (uncommon)<br>1 = Flatsloping: minimal change (typical healthy heart)<br>2 = Downsloping: signs of unhealthy heart |
| Fluoroscopy Blood Flow | factor | Number of major vessels colored by fluoroscopy (procedure to see blood flow) |
| Heart Disease | char | 0 = do not have heart disease<br>1 = have heart disease |

There are nine categorical variables in the data including gender, chest pain, fasting blood sugar, resting electrocardiograph, exercise induced angina, slope, fluoroscopic blood flow, and the target variable heart disease. There are roughly twice as many males in this study than females with 186 males and 92 females represented. Chest pain has four groups, the first is a typical angina, meaning the subjects have chest pain related to decrease blood supply to the heart. The second is atypical angina, which is chest pain that is not related to the heart. The non-anginal pain is typical esophageal spasms that are not related to the heart. The last group is asymptomatic chest pain with no signs of heart disease. Fasting blood sugar is represented by a 1 if the individual has a blood sugar of over 120 mg/dl and a 0 if the individual is below 120 mg/dl. Resting electrocardiographic has three groups where the first group indicates there was nothing to note from the ECG results. The second group is ST-T wave abnormality which can indicate mild to severe symptoms of a non-normal heart beat. The last group are individuals who may have a left ventricular hypertrophy, or an enlarged of the hearts main pumping chamber. This is the smallest group consisting of only three individuals, where only one was reported to have heart disease. Exercise induced angina is also represented by a 1 for individuals who suffer chest pain when exercising and a 0 for individuals who do not suffer chest pain when exercising. The slope of the peak exercise ST segment contains three groups. The first group is upsloping, which indicates a better heart rate with exercise, which is uncommon. The second group is flatsloping, this indicates minimal change in heart rate and is considered a typical healthy heart. The last group is downsloping, this is a sign of an unhealthy heart. Fluoroscopy blood flow measure the movement of blood in the body, the lower the number the better the blood flow or no indications of clots in the blood system. The last categorical variable is whether the individual has heart disease or not represented by a 1 if they have heart diseasee or 2 if they do not have heart disease.

## Data Evaluation

Figure 1 provides an exploratory analysis on the categorical variables against the target, have heart disease No/Yes. These variables where chosen to compare and find insight into what the data represents. Graph A in Figure 1 is heart disease by gender where there are 186 males represented in the data and 92 females. There are roughly twice as many males in the study than females with 75% of the females having heart disease compared to approximately 46% of the males having heart disease.

Graph B in Figure 1 is heart disease by chest pain. The largest group, 128, fall under typical angina, meaning the subjects have chest pain related to decrease blood supply to the heart. However, there is only 38 subjects, or 30% of that group, that show typical angina and have heart disease. Atypical angina is chest pain that is not related to the heart. There are 48 subjects who have atypical angina with the majority, 40 or 93% reported to have heart disease. The Non-Anginal Pain is typical esophageal spasms that are not related to the heart. There are 79 subjects that report non-anginal pain and 64, or 91% having heart disease. The last group is Asymptomatic chest pain with no signs of heart disease. There are 23 subjects in this group with approximately 70% having heart disease.

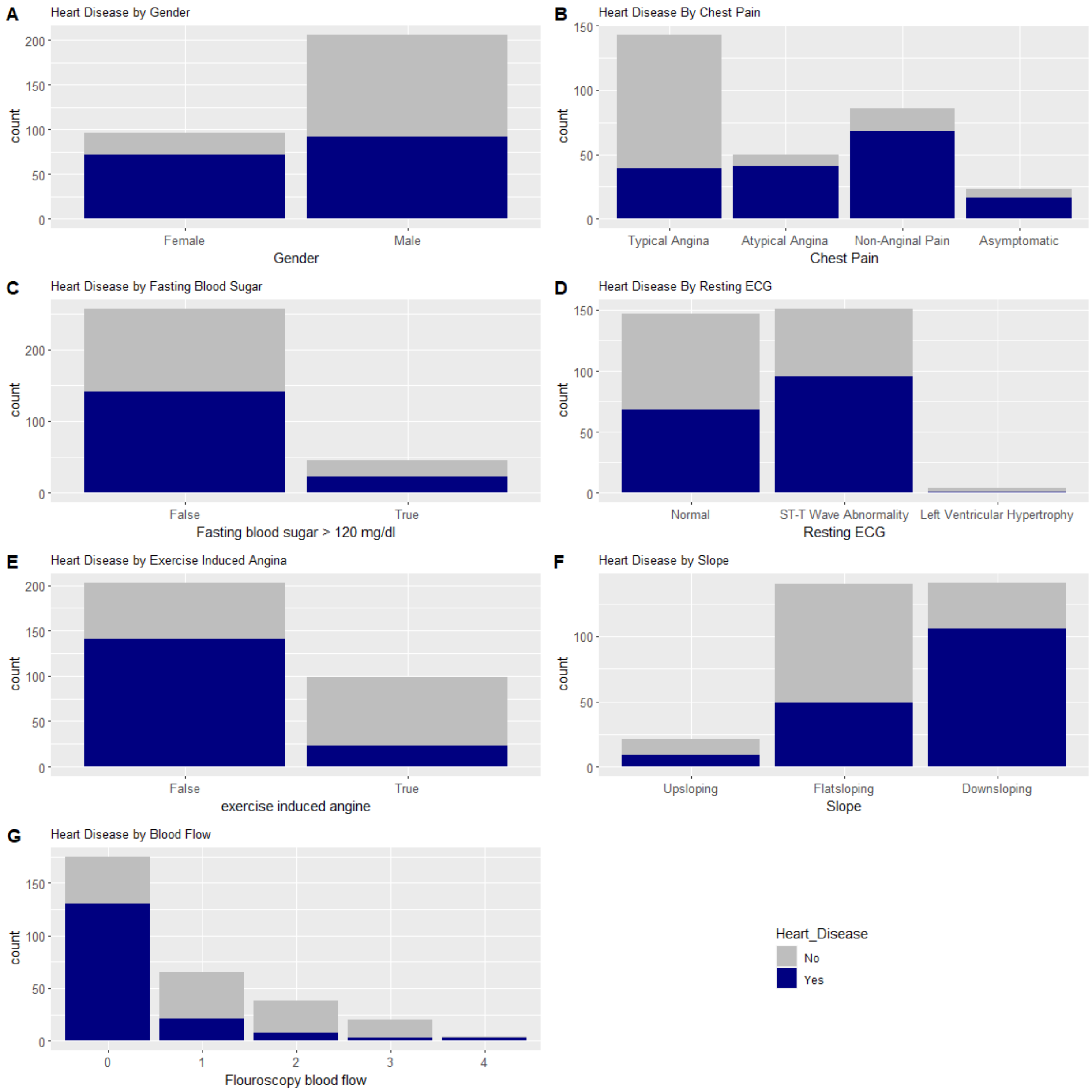
Fasting blood sugar displayed in Graph C Figure 1 has 239 individuals that have a fasting blood sugar below 120 ml/dl and 137 of them, or 57% report having heart disease. There are 39 individuals who have fasting blood sugar above 120 mg/dl and 21, or about 54% report having heart disease. Since these figures are roughly close to 50% for both True and False, fasting blood sugar may have little to no effect on the prediction of whether an individual has heart disease or not.

Graph D in Figure 1 is heart disease by resting electrocardiographic (ECG). There were 67 individuals out of the 133 that reported having heart disease or approximately 50% where the ECG indicated there was nothing to note from the results. The second group, representing the ST-T wave abnormality which can indicate mild to severe symptoms of a non-normal heart beat, is the largest group consisting of 142 individuals where 90, or approximately 63% reported having heart disease. The last group are individuals who may have a left ventricular hypertrophy, or an enlarged of the hearts main pumping chamber. This is the smallest group consisting of only three individuals, where only one was reported to have heart disease.

Heart disease by exercise induced angina in Figure 1 Graph E has 187 individuals that did not have exercise induced angina, out of those individuals 135 reported having heart disease or roughly 72%. There were 91 individuals who reported exercise induced angina where 23, or 25% reported having heart disease. These results appear to be incorrect, but the graph is reporting correctly verified by the raw data.

Figure 1 Graph F is heart disease by the slope of the peak exercise ST segment. The first group is upsloping, which indicates a better heart rate with exercise, which is uncommon. There were 19 individuals that had a better heart rate and 9 of them have reported having heart disease. The second group is flatsloping, this indicates minimal change in heart rate and is considered a typical healthy heart. There were 125 individuals that had a flatsloping heart rate and 47, or approximately 38% that reported having heart disease. The last and largest group is downsloping which is a sign of an unhealthy heart consisting of 134 individuals where 102 reported having heart disease or approximately 76%.

Heart disease by blood flow displayed in Graph G of Figure 1 signifies the higher the number the better the blood flow or no indications of clots in the blood system. The first and largest group with the lowest blood flow contains 175 individuals, where 130, or 72% report having heart disease. The second group contains 65 individuals where 21 of them report having heart disease or roughly 32%. The third group which contains 38 individuals where 7, or 18% report having heart disease. The fourth group has 20 individuals, three of which have heart disease or 15%. The last group which has the highest blood flow contains four individuals where three out of the four have heart disease. This shows a clear trend that the higher the blood flow the less individuals with heart disease.



## Figure 1:

*Exploratory analysis of categorical data, gray is no heart disease and navy is yes heart disease.* **Graph A:** *Heart Disease by Gender.* **Graph B:** *Heart Disease by Chest Pain; Typical angina, atypical angina, Non-Angina Pain, and Asymptomatic.* **Graph C:** *Heart Disease by Fasting Blood Sugar; fasting blood sugar > 120 mg/dl equals true, otherwise false.* **Graph D:** *Heart Disease by Resting ECG; Normal, ST-T Wave Abnormality, Left Ventricular Hypertrophy.* **Graph E:** *Heart Disease by Exercise Induced Angina.* **Graph F:** *Heart Disease by Slope; Upsloping, Flatsloping, and Downsloping.* **Graph G:** *Heart Disease by Blood Flow; higher blood flow (0-4) less likely to have heart disease.*

The box plots in Figure 2 are an exploratory analysis of the quantitative variables in the Heart Disease data set. The box plots where chosen to display the statistical summaries of the quantitative variable and to compare between having heart disease and not having heart disease. Figure 2 Graph A shows heart disease by age. For individuals who have not reported having heart disease, the boxplot show that the median age to be 58 years old, with the majority of the observations falling between the ages of 52 and 61. The overall range is between age 39 and 70 with a few outliers below the lower whiskers. For individuals who have reported having heart disease, the median age is 52, with a overall range between the age of 29 and 76. The majority of the observations fall between the age of 44 and 59. It is interesting that the individuals with heart disease appear to be younger in age than the individuals without heart disease.

Graph B in Figure 2 is heart disease by resting blood pressure. This number represents the systolic pressure (top number) when reading blood pressure. There is not much difference in resting blood pressure between having heart disease and not having heart disease. Both have a median value of 130 and a quantile 1 value of 120. The quantile 3 value for individuals without heart disease is 144 and with heart disease is 140. The majority of resting blood pressure, regardless of heart disease is between 120 and 140. The overall range for individuals without heart disease is between 100 and 180 with a few outliers above the upper whiskers with a max at 200. The overall range for individuals with heart disease is between 94 and 170 which also have a few outliers above the upper whiskers with a max at 180. It appears that the systolic blood pressure value from this data set does not indicate whether an individual has heart disease or not.

Figure 2 Graph C is heart disease by cholesterol level. This number was derived by serum = LDL + HDL + .2 * triglycerides. If serum calculates to be greater than 200, there is cause for concern. For individuals reported to not have heart disease, the median value is 248.5 with the majority of observations between 215.8 and 281.2. With an overall range between 131 and 353, with an outlier at 409. For individuals who have reported to have heart disease, the median value is 235.5 with the majority of observations between 208.2 and 268.8. The overall range for individuals with heart disease is between 126 and 360, with a few outliers and a max of 564. The large majority of individuals in the is study regardless of if they have heart disease or not have a cholesterol level above 200. So it also appears that the cholesterol level from this dataset does not indicate whether an individual has heart disease or not.

Heart disease by maximum heart rate is displayed in Figure 2 Graph D. For individuals who are reported not having heart disease the median maximum heart rate is 142.5 with the majority of observations falling between 125 and 158. The total overall range is from 88 to 195 with an outlier at 71. For individuals who reported to have heart disease the median at 161 with the majority falling between 149.2 and 172. The overall range is from 115 to 202 with a few outliers under the maximum heart rate of 115. The minimum being 96. These box plots indicate that people with heart disease tend to have a higher maximum heart rate.

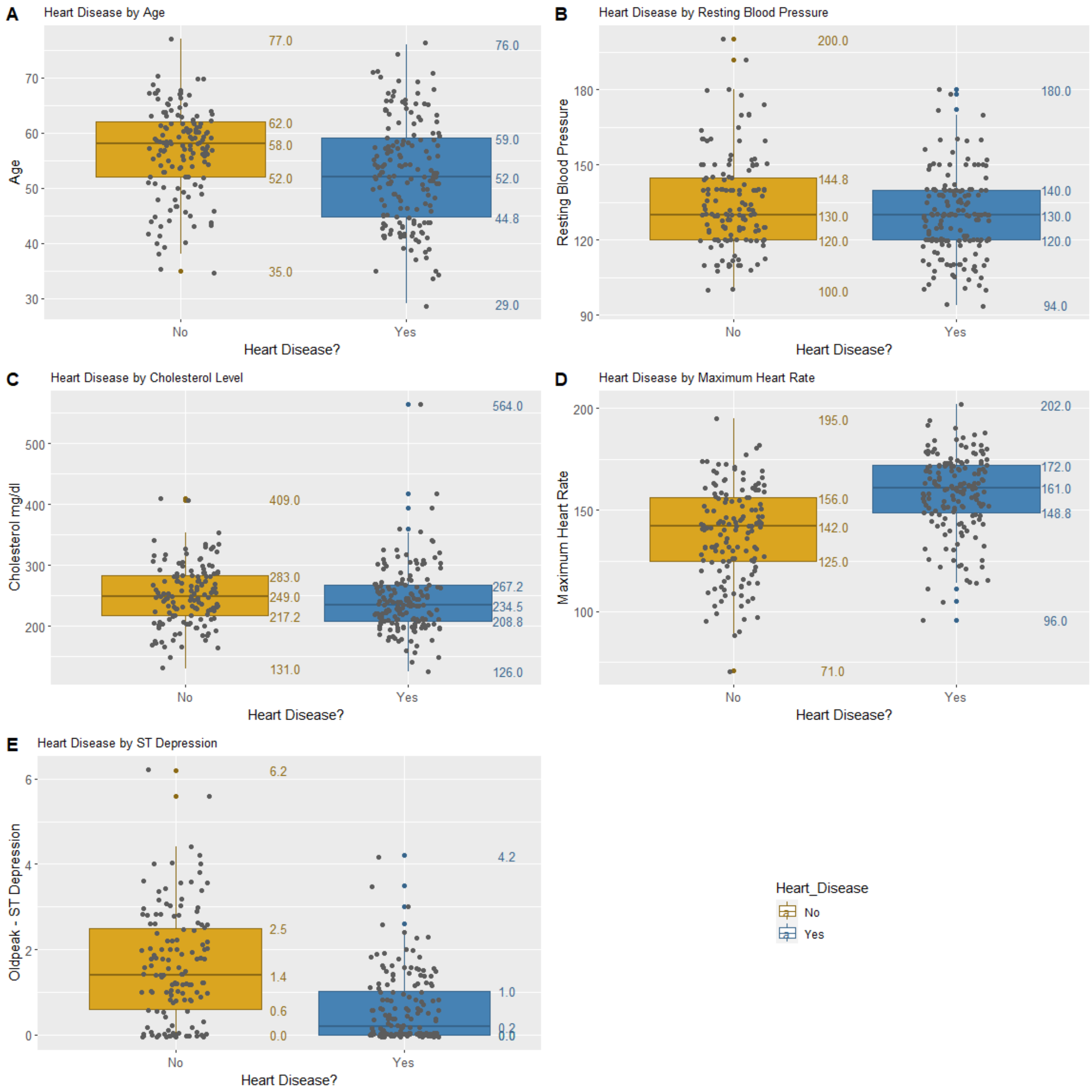
Graph E in Figure 2 is heart disease by old peak ST depression. The box plot for individuals who do not have heart disease has a median value of 1.4 and an overall range between 0 and 4. The majority of the observations fall between 0.6 and 2.5, with an outlier at 5.6. For individuals with heart disease the median is at 0.2 with an overall range between 0 and 2.6. The majority of the observations fall between 0 and 1.1 with a few outliers and a max of 4.2. The majority of all observations are at zero and the further away you get from zero the less likely you will have heart disease.

## Analysis and Results

The chi-squared test of independence is used to determine if there is an association between two categorical variables. The first analysis was to find if there is any association between each categorical variable and whether they have heart disease, the predictor variable. My hypothesis test is as follows:

$H_0$: There is no association between the chosen variable and having heart disease

$H_a$: There is an association between the chosen variable and having heart disease



## Figure 2:

*Exploratory Analysis of quantitative variables. Gold represents individuals without heart disease and blue represents individuals with heart disease.* **Graph A:** *Heart Disease by age.* **Graph B:** *Heart Disease by Resting Blood Pressure.* **Graph C:** *Heart Disease by Cholesterol Level.* **Graph D:** *Heart Disease by Maximum Heart Rate.* **Graph E:** *Heart Disease by ST Depression.*

**Table 2:**

| Variable | X_squared | P_Value |
|---|---:|---:|
| Gender vs Heart_Disease | 23.08387946 | 0.00000155 |
| Chest Pain vs Heart_Disease | 80.97876151 | 0.00000000 |
| Fasting Blood Sugar vs Heart_Disease | 0.09240836 | 0.76113747 |
| Resting ECG vs Heart_Disease | 9.72968231 | 0.00771305 |
| Exercise Induced Angina vs Heart_Disease | 55.45620298 | 0.00000000 |
| Slope vs Heart_Disease | 46.88947660 | 0.00000000 |
| Blood Flow vs Heart_Disease | 73.68984583 | 0.00000000 |

As seen in Table 2 the X_squared test statistic for gender, chest pain, resting ECG, exercise induced angina, slope and blood flow are mostly higher numbers and using $\alpha \leq 0.05$ the p-values are below 0.05. For these variables we would reject the null hypothesis, concluding there is evidence that there is an association between these specific variable and having heart disease. For fasting blood sugar we would fail to reject the null hypothesis because the X_squared value is low and the p-value is approximately 0.76, which is greater than 0.05. There is no evidence there is an association between fasting blood sugar and having heart disease. This chi-squared test of independence verified what the visualizations in Figure 1 displayed, that fasting blood sugar is not a significant variable when determining if an individual has heart disease.

The second analysis was to find if there is any association between each categorical variable and the slope, the slope of the peak exercise ST segment. I chose the slope to see if there was any association because the data showed a clear distinction that people with downsloping ST segment typically have heart disease. My hypothesis test is as follows:

$H_0$: There is no association between the chosen variable and the slope

$H_a$: There is an association between the chosen variable and the slope

**Table 3:**

| Variable | X_squared | P_Value |
|---|---:|---:|
| Gender vs Slope | 0.6700921 | 0.71530514 |
| Chest Pain vs Slope | 27.3926971 | 0.00012222 |
| Fasting Blood Sugar vs Slope | 3.3472109 | 0.18756957 |
| Resting ECG vs Slope | 10.6435610 | 0.03087589 |
| Exercise Induced Angina vs Slope | 24.7556872 | 0.00000421 |
| Blood Flow vs Slope | 11.2115315 | 0.18999920 |

What can be seen in Table 3 is that gender has a p-value of $> 0.71$ which is greater than $\alpha = 0.05$, we would fail to reject the null hypothesis, there is no evidence there is an association between gender and the slope. The same can be said for both fasting blood sugar and blood flow, both have a p-value equal to $\approx .19$. Again, showing not evidence there is an association between the two variables and the slope. The three remaining variables are chest pain, resting ECG and exercise induced angina. All of which have p-values less than 0.05. In this case we would reject the null hypothesis, there is evidence there is an association between these three variables and the slope.

## Discussion and Conclusion

In reviewing this data, I have learned there are a number of warning signs that I personally can look out for that could be a sign of heart disease. Although most of these metrics can only be determined by a doctors visit and necessary tests, chest pain and exercise induced angina can be clear signs without the need of a doctor. If I had more time to analyze this data further, I would like to pull data directly from the source and not through the kaggle website. The There are a lot of other things that this data left out that could help predict heart disease, like weight, exercise and other lifestyle parameters. Additionally, I would have done further analysis on the quantitative variables and produce a logistic regression model with test and training sets to better predict heart disease.

## Appendix

1. CDC. 2022. "Heart Disease Facts  Cdc.gov." *Centers for Disease Control and Prevention*. https://www.cdc.gov/heartdisease/facts.htm.

2. "Heart Disease UCI." n.d. Accessed February 23, 2023. https://www.kaggle.com/datasets/hartman/heart-disease-uci.

3. Github code