Math 3636 Class Project Report

Title and Abstract

To: Jiatian (Justin) Xu

From: Group 9 (Christopher van Schalkwyk, Daniel Zygadlo)

Date: 04/28/2020

Title: Heart Disease and Insurance Pricing

Introduction (Problem Definition)

A critical question about our health insurance is: What are the common health traits among customers that can be considered risk factors. This question is important as it can help us price health insurance by estimating risk. There's a high proportion of people with heart diseases that share common attributes. In this project, we want to search for different trends to predict cardiovascular events. Accordingly, we want to see if there are more significant attributes that affect the presence of heart disease.

We decided to use UCI's dataset on Heart Disease since it was updated two years ago and contains three hundred and three participants. Using analytical techniques, we can help determine the relationships that can help guide the pricing department in its efforts to create fair policies for both the insurer and insuree.

Data Exploration

Our dataset from UCI on Heart Disease was updated two years ago and contains three hundred and three participants. Our data contains thirteen independent variables (Age, Sex, Chest Pain Type, Fasting Blood Suger, Resting Blood Pressure, etc.) and one dependent variable (Condition/Target). Eight of our thirteen independent variables are categorical variables and are factored accordingly.

With regard to potential risk, before implementing any marketing policies, the legal department should be consulted to determine if any of the information that we are collecting is against HIPAA

(Health Insurance Portability and Accountability Act) guidelines. There is also potential risk that customers may react negatively to us surveying, and tying that information to them in order to price their health insurance. Should any of these variables be seen as not legal, feasible, or have a high risk at potential negative response, report back to Group 9 and the model can be adjusted accordingly.

There is only one target variable, CONDITION, which is set equal to 1 if a heart disease is present in an individual, and 0 in the absence of a heart disease.

Thirteen predictor variables were extracted from the data that were deemed to have a possible relationship with our target variable: CONDITION. The list of predictor variables, as well as the descriptions of their representation is provided in Appendix A.

Statistical Summary:

AGE	SEX	CP	TRESTBPS	CHOL	FBS	THAL
Min . : 29.00	0: 96	0: 23	Min . : 94.0	Min . : 126.0	0: 254	0: 164
1st Qu.: 48.00	1: 201	1: 49	1st Qu.: 120.0	1st Qu.: 211.0	1: 43	1: 18
Median: 56.00		2: 83	Median: 130.0	Median: 243.0		2: 115
Mean : 54.54		3: 142	Mean : 131.7	Mean : 247.4		
3rd Qu. :61.00			3rd Qu. :140.0	3rd Qu.: 276.0		
Max : 77.00			Max : 200.0	Max : 564.0		

RESTECG	THALACH	EXANG	OLDPEAK	SLOPE	CA	CONDITION
0 : 147	Min . : 71.00	0: 200	Min . : 0.000	0: 139	0: 174	0: 160
1 : 4	1st Qu.: 133.00	1: 97	1st Qu.: 0.000	1: 137	1: 65	1: 137
2 : 146	Median:153.00		Median: 0.800	2: 21	2: 38	
	Mean :149.60		Mean : 1.056		3: 20	
	3rd Qu.:166.00		3rd Qu.: 1.600			
	Max :202.00		Max : 6.200			

Note: Categorical Variables are factored. CONDITION was factored for the purpose of this summary and unfactored for modeling.

The following observations stand out:

- 53.9% of the individuals in our data did not have heart disease. The sample is nicely balanced between those who did and did not have heart disease.
- 67.68% of the individuals in our data are males. This tells us that there are twice as many males as there are females. There could be a connection between gender and heart conditions.

- The age range of 29 to 77 is reasonable, but not perfect. Ideally we would want the minimum age to be at least 20, in order to use age as a more appropriate predictor of our data.
- Very few individuals had normal RESTECG. Yet 53.9% of individuals did not have heart disease. We can look out for this variable to determine its significance in predicting heart disease.
- There were a few missing values that were adjusted for by removing them from the data set.
- The max CHOL of 564.0 seems very large. This value, along with potentially others, may show to be outliers.

The non-categorical variables and the target variable, CONDITION were pictured in scatter plots in an attempt to judge the effects of potential outliers. In the plots of different predictor variables we are able to see many potential outliers as pictured below:

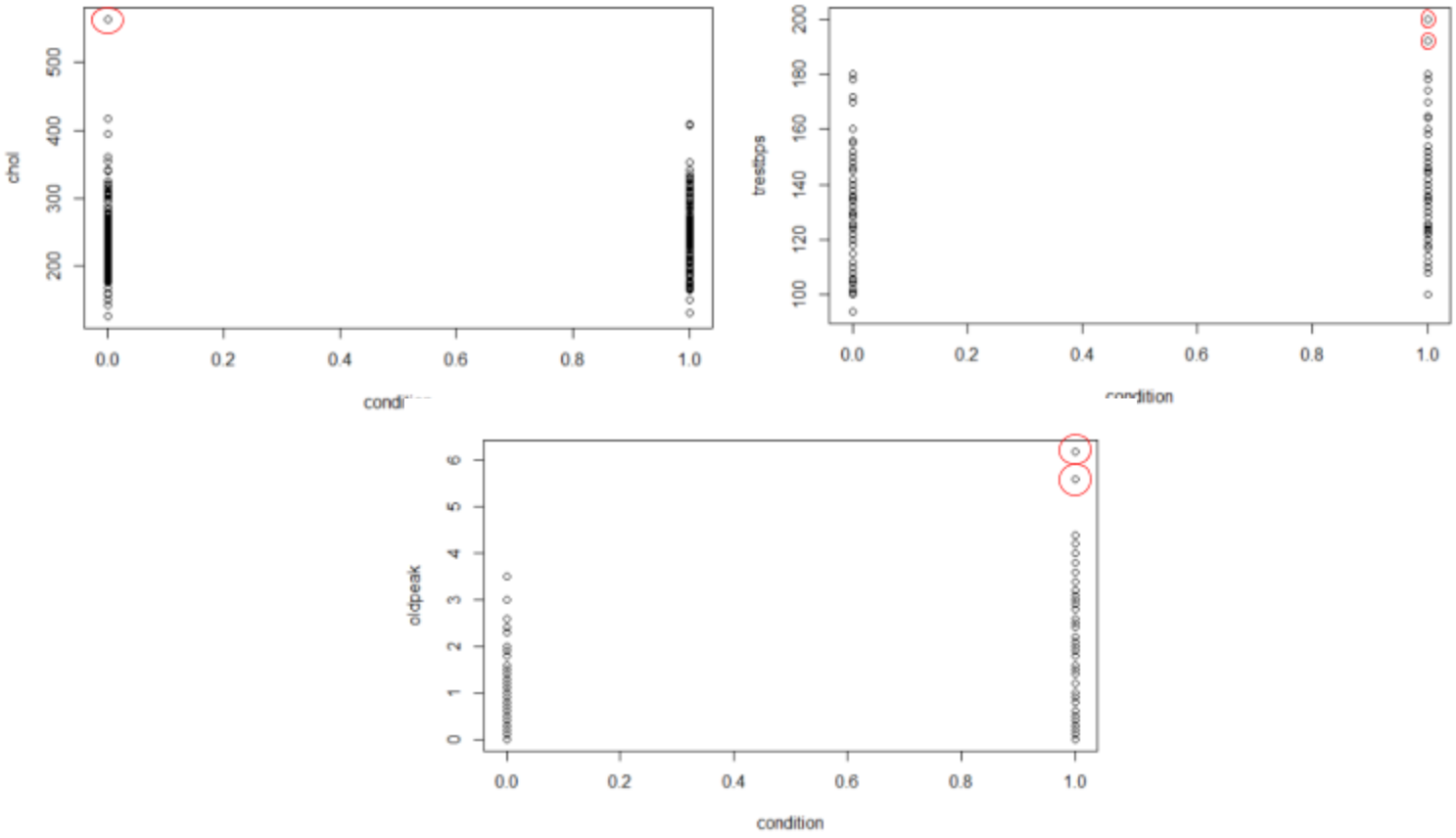

Figure 1: Example Outlier Plot For CHOL, TRESTBPS, and OLDPEAK

We then remove these outliers by eliminating any data points that are a multiple of 1.5 times greater than the IQR for each variable. Following these procedures we are able to see that the data becomes much more suitable for modeling, due to less outliers.

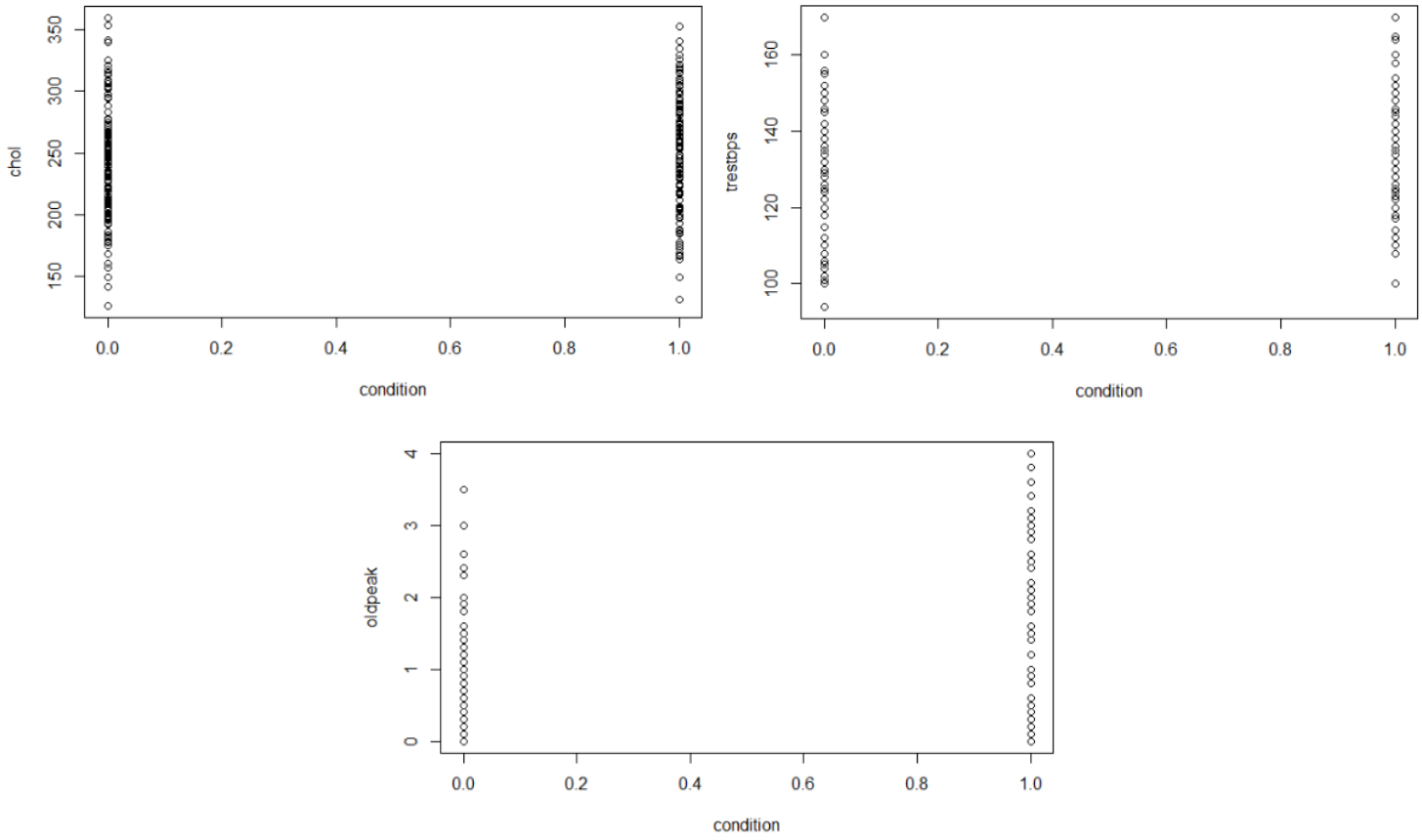

Figure 2: Example Plots for CHOL, TRESTBPS, and OLDPEAK following removal of outliers.

As you can see, the ranges of the plots become much more compact, and there is less variance in the plots following the removal of the outliers. For more outlier plots and a greater insight into the procedure, refer to the RMD file that accompanies this report.

The plot of residuals as well as cook's distance indicate that we must do some form of transformation, to assure that we are not violating any testing and modeling assumptions.

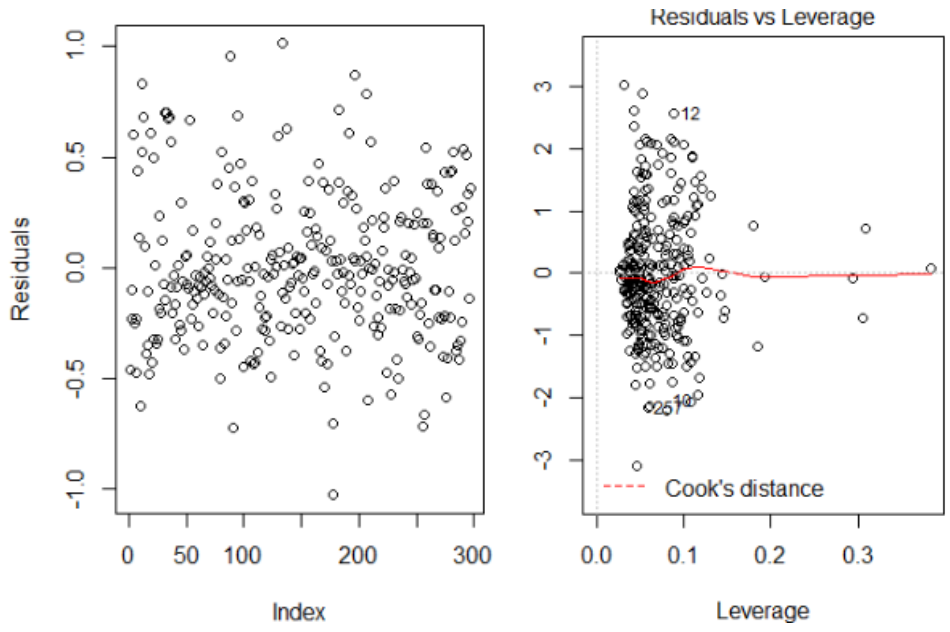

Figure 3: Residuals and Cook's Distance Of Pre-Transformation Data

We apply the transformation technique of the method of maximum likelihood. We do this as the logistic and probit regression models are nonlinearly on their own. Thus we estimate these models using the method of maximum likelihood in order to fit these logistic and probit models using the glm function.

After our transformations using link functions "probit" and "logit", we can see that the plot of residuals become uniformly distributed, indicating a more favorable distribution for modeling. An example of this is the constant variance. Cook's distance becomes much more clustered around 0, indicating that there is a great reduction in potential outliers.

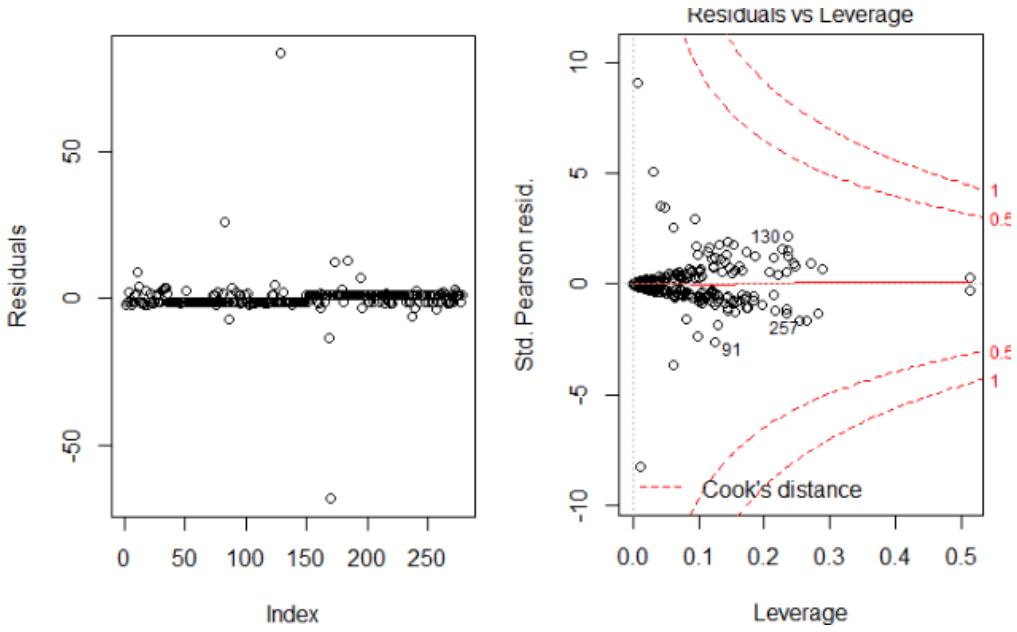

Figure 4: Logit-Full Post-Transformation Plot of Residuals and Cook's Distance for others refer to RMD file

These plots and figures were analyzed for each of the 6 models. Through these figures we can determine that the linear models with no link function transformations would be inferior to their probit and logit counterparts as the data becomes much more favorable for modeling with these transformations by violating less assumptions. An example of this is that these link functions transform the data from a binomial distribution to a normal distribution.

Feature Selection

Based on the glm fit models we are able to determine the significant variables by looking at the summary portion of each respective fit. Using this summary we can see that the variables: SEX, CP, CA, THAL, TRESTBPS, SLOPE, CHOL, and OLDPEAK all are significant to the $\boldsymbol{\alpha} = 10\%$. We then take these eight variables and separate them into their own reduced model and fit new generalized linear models to see if we can improve the efficiency of the model by lowering the amount of variables.

Model Selection and Interpretation

CONDITION Target Variable

The variable that indicates whether a heart disease is present within an individual is CONDITION, which is 0 if a heart disease is not present and 1 if a heart disease is present. There are thirteen variables, of which eight are categorical variables, that we could use to predict the target variable, CONDITION.

We fit two types of models: General Linear Model with link function "logit", and General Linear Model with link function "probit". Each of these models has a full version in which all thirteen predictor variables were used, and a reduced version in which only variables deemed significant to a significance value of $\alpha = 10\%$ were used.

Comparing the AIC of each of the four general linear models, we noticed that the AIC's are all relatively similar in the 205- 215 threshold. As all of these models are so close in AIC, we decided to continue working with both the full and reduced models for the logit and probit link function models in order to later compare each model's accuracy to determine the best.

We use the predict function in R in order to obtain the fitted values of our general linear model. A problem that arises is that our fitted values are probabilities, ranging from 0 to 1. This does not reflect our target variable, CONDITION, as this variable can only take binary values of 0 or 1. We remedy this by editing our fitted values by rounding our fitted values to the closest value of 0 or 1. This means any values less than 0.5 rounds to 0, and any values greater than 0.54 rounds to 1. We decided 0.54 as our probability cutoff, as our data is fairly split as 53.9% of individuals did not have heart disease.

We then create a confusion matrix by aligning the predicted responses, and the observed responses. We use this matrix to determine the accuracy, as well as the possibility of error in our model by following this format:

		Predicted Response	
		Positive	Negative
Observed Response	Positive	True Positive	False Negative (Type II Error)
	Negative	False Positive (Type I Error)	True Negative

Figure 5: Table Legend for Confusion Matrix

After comparing each of the four confusion matrices, their statistical summaries, and procedures, we chose to focus on the Reduced General Linear Model with the link function “logit” for four reasons:

- The General Linear Models with link functions “probit” and “logit” do not violate modeling assumptions whereas the Linear Model does.
- The Reduced General Linear Model with link function “logit” has a higher accuracy than the Reduced General Linear Model with link function “probit”.
- By using the Reduced General Linear Model with link function “logit”, we got rid of five insignificant predictor variables and gained 0.7194% accuaracy. This tradeoff is favorable as we are able to use a model that is more accurate, for nearly half the predictors. This will save costs in data collection, as well as over fitting of the data, for a slight increase in accuracy.
- The reduced logit model has a lower AIC of 209.06, in comparison to the reduced probit model of 210.78.

To assess the performance of our models, we decided to use the Rand Accuracy to evaluate the out-of-sample performance of all of our models. For example, the Confusion Matrix of the model that we chose, the Reduced General Linear Model “logit”, shows us the True Positive (142/278 = 51.08%), False Negative (12/278 = 4.32%), False Positive (22/278 = 7.91%) and True Negative (102/278 = 36.69%). From this observation we can calculate the accuracy of our model by summing the True Positive and True Negative which results as 87.78%.

		Predicted Response	
		0	1
Observed	0	142	12
Response	1	22	102

Figure 6: Confusion Matrix for Reduced Generalized Linear Model with link function logit

Summary and Concluding Remarks

Multiple methods were conducted in order to predict whether a target individual has heart disease. As the data is collected in a medical setting, we are able to accurately create a model that can use the predictive variables in our model to determine the likelihood of an individual having heart disease, forecasted by our 87.78% accuracy.

In the predictive variables that determine the CONDITION of an individual, are we able to make out some overwhelming correlations. More specifically:

- The sex of the individual is a significant predictor of heart disease.
    - As we can see in this matrix, only 21.2% of females have heart disease. Compared to, 54.92% of males have heart disease. Further research indicates that males are twice as likely to experience heart attacks and thus heart disease, than females (Read more in Appendix C). Which is supported by the dataset and the significance of sex in the predictive model.

		CONDITION	
		NO DISEASE	DISEASE
SEX	FEMALE	67	18
	MALE	87	106

Figure 7: Confusion Matrix for the presence of heart disease and the corresponding gender

- Other trends include a high significance of CA, which is the number of major vessels colored by the fluoroscopy. This makes sense as if a vessel is colored, there is something wrong with the vessel, blockage, for example. Thus a high number of blockages will translate to a high chance of heart disease which is useful for predicting the heart disease.

To depict the effect of other variables, we describe the reduced model that was used.

Factor	Adjustment	Significance (Lower is better)
Intercept	-10.293091	**0.0000178**
Sex1	1.351937	0.010075
CP1	1.098036	0.168779
CP2	0.023333	0.973881
CP3	2.369551	**0.000635**
CA1	2.296227	**0.00000294**
CA2	2.624812	**0.000306**
CA3	2.128688	0.025344
THAL1	0.179538	0.813508
THAL2	1.689438	0.**000127**
TRESTBPS	0.022387	0.075380
CHOL	0.008447	0.076377
SLOPE1	1.247702	0.007467
SLOPE2	0.854630	0.328748
OLPEAK	0.601963	0.013415

Figure 8: The model that was used to predict heart disease with the highest accuracy relative to variables required

This model can be adjusted for the data that was collected. It is presented here to provide an idea of the significance of each predictor variable, and the intensity of its effects on our predictions. We add each of these values accordingly to the intercept depending on the appearance of each of the factors per individual. We can then analyze the probability of the "response" by rounding the values to 0 or 1. A 0 indicates no heart disease, and a 1 indicates heart disease.

The pricing department can use this information to determine the risk of any individual who wishes to purchase health insurance. These variables can be appropriate to determine whether an individual is at risk of heart disease to a α = 10%. If a lower alpha is desired, the model can be adjusted

to only include variables whose significance is less than the desired $\boldsymbol{\alpha}$. Although this may not be necessary as our reduced model has higher accuracy than the full model, possibly demonstrating that we have the most significant predictors available. The significance of the predictors, as well as the 87.78% accuracy of the model, leads to the conclusion that this is a viable model to determine potential heart disease in individuals.

As noted earlier, some of these variables may not be practical, appropriate, or allowed via HIPAA. Let Group 9 know if such issues arise.

References and Appendix

Appendix A - Data Dictionary

The following table describes the thirteen variables that were used in this analysis.

Variable Name	Description
AGE	Age in years
SEX	Sex (1 - Male; 0 - Female)
CP	Chest Pain type: 0 - Asymptomatic 1 - Atypical Angina 2 - Non-anginal Pain 3 - Typical Angina
TRESTBPS	Resting Blood Pressure
CHOL	Serum cholesterol in mg/dl
FBS	Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
RESTECG	Resting Electrocardiographic Results 0 - Showing probable or definite left ventricular hypertrophy by Estes' criteria 1- Normal 2 - Having ST-T Wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
THALACH	Maximum heart rate achieved
EXANG	Exercise including angina (1 = yes; 0 = no)
OLDPEAK	ST depression induced by exercise relative to rest
SLOPE	The slope of the peak exercise ST segment
CA	Number of major vessels (0-3) colored by fluoroscopy
THAL	1 - Fixed Defect 2 - Normal 3 - Reversible Defect
CONDITION (target)	0 - No heart disease 1- Indicative heart disease

Appendix B - Source of Data

Below is the cited source for our data:

Ronit. “Heart Disease UCI.” *Kaggle*, 25 June 2018, www.kaggle.com/ronitf/heart-disease-uci.

Appendix C - Heart Disease Between Males and Females

Below is the source for this article:

Harvard Health Publishing. “Throughout Life, Heart Attacks Are Twice as Common in Men than Women.” *Harvard Health*, Harvard Heart Letter, Nov. 2016, https://www.health.harvard.edu/heart-health/throughout-life-heart-attacks-are-twice-as-common-in-men-than-women