WHAT CAUSES HEART DISEASE?

Arizona State University
CIS 450 | Course Number: 44897

Yiming Guo

Introduction:

The predictable diagnosis/hypothesis for heart disease is a significant key for the current/potential patients who have similar symptoms such as high cholesterol, diabetes, over-weight, smoking and family history and these factors would be possible to provide multiple patients with correct/potential treatments with fewer side-effects. It will classify if patients have heart disease or not according to factors. However, there are major changeable factors (high blood pressure, high cholesterol, smoking, over-weight, diabetes, and physical inactivity) and major non-changeable factors (increasing age, gender (biological status), and heredity (family history)). To identify factors that cause heart disease is a necessary step to achieve the analysis based on the dataset. I will focus on the issue to classify which factors affect/cause heart disease.

Background:

The main issue topic that I am going to address is "Determine factors that the patients with different health conditions cause heart disease. All patients have different or similar conditions based on the dataset such as age, gender, the chest pain experienced, the person's resting blood pressure, the person's cholesterol measurement in mg/dl, the person's fasting blood sugar, resting electrocardiographic measurement, the person's maximum heart rate achieved, exercise-induced angina, ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot), the slope of the peak exercise ST segment, the number of major vessels, a blood disorder called thalassemia, and the target heart disease. The diagnosis of heart disease is based on each patient condition and it is possible to predict which treatment with fewer side-effects or surgery or detailed information which patients want and need to know. This dataset provides various numbers of variables based on the target conditions whether the patient (or target) has or not has

the heart disease. I will focus on the 14 attributes out of 74 attributes which already filtered in Kaggle. I found this dataset in Kaggle, which is “Heart Disease UCI” to analyze the main issue topic. According to Kaggle, “This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The "goal" field refers to the presence of heart disease in the patient. It is integer-valued from 0 (no presence) to 4.”.

Dataset Content:

Attributes	**Description**
age	The person's age in years
sex	The person's sex (1 = male, 0 = female)
cp	The chest pain experienced (Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic)
trestbps	The person's resting blood pressure (mm Hg on admission to the hospital)
chol	The person's cholesterol measurement in mg/dl
fbs	The person's fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)
restecg	Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)
thalach	The person's maximum heart rate achieved
exang	Exercise induced angina (1 = yes; 0 = no)
oldpeak	ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot. See more here)
slope	the slope of the peak exercise ST segment (Value 1: upsloping, Value 2: flat, Value 3: downsloping)
ca	The number of major vessels (0-3)
thal	A blood disorder called thalassemia (3 = normal; 6 = fixed defect; 7 = reversable defect)
target	Heart disease (0 = no, 1 = yes)

I am considering two or three types of analysis. The first analysis is logistic regression on heart disease. My target variable (other binary variables such as sex, fbs, and exang are considerable) would be whether the patient has heart disease or not (0 means no and 1 means yes). My second analysis is the decision tree based on the 14 attributes that I mentioned in the data content. The decision tree would be the random forest model to check sensitivity and specificity. The last analysis is classification based on major changeable and major non-changeable factors. The reason I am considering these three analyses because it would be based on binary variables (sex, fbs, exang, and target), categorial (cp, slop, and thal) variables and continuous variables (trestbps, chol, and thalach). I will try to check in a different statement regarding the factors which cause heart disease. The example questions from the dataset can be: "Do females have a higher possibility to have heart disease compared to males?", "Which age range has the maximum heart rate achieved?", "Which age range has the most reversible defect?", "Is high resting blood pressure highly correlated to chest pain type?", "If the patient has a fixed defect, the patient's maximum heart rate achieved would be higher than normal defect patients?" I am planning to use SAS miner, Python, and Tableau to visualize/filter the dataset.

Analytical Methods:

- **Python**

  I only used Python to analyze based on 14 attributes in the dataset content to demonstrate the hypothesis of heart disease. There is a Figure 1 to Figure 11 to explain the dataset each with different visualization graphs.

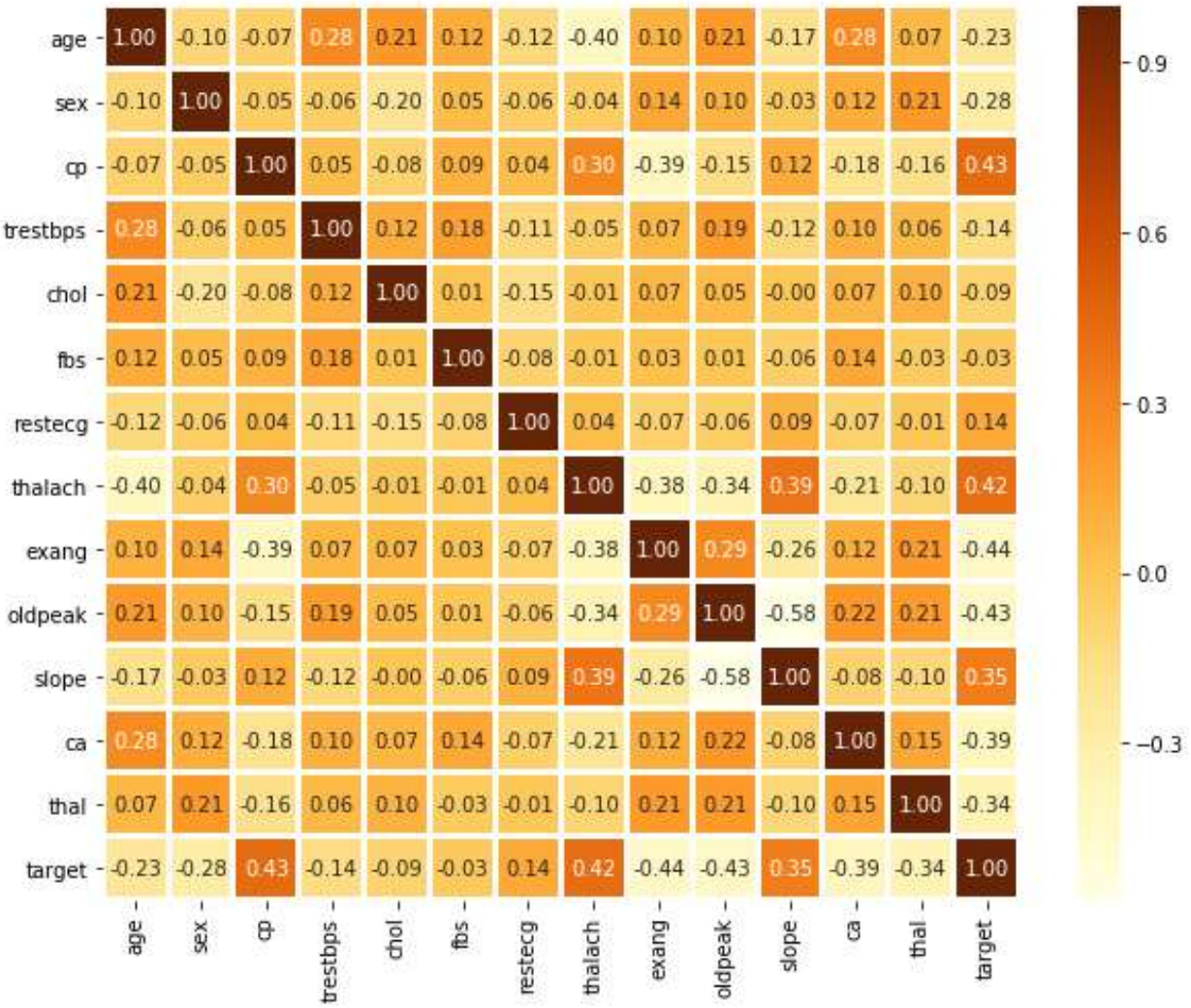


Figure 1. Heatmap (Correlation) of the entire variables:

I wanted to check there are any relationships among variables, but this correlation matrix does not correlate to each other. All numbers are below 0.5 therefore, it is hard to tell if they are correlated to each other. So, I decided to use single variables to build the model between one or two variables in further figures.

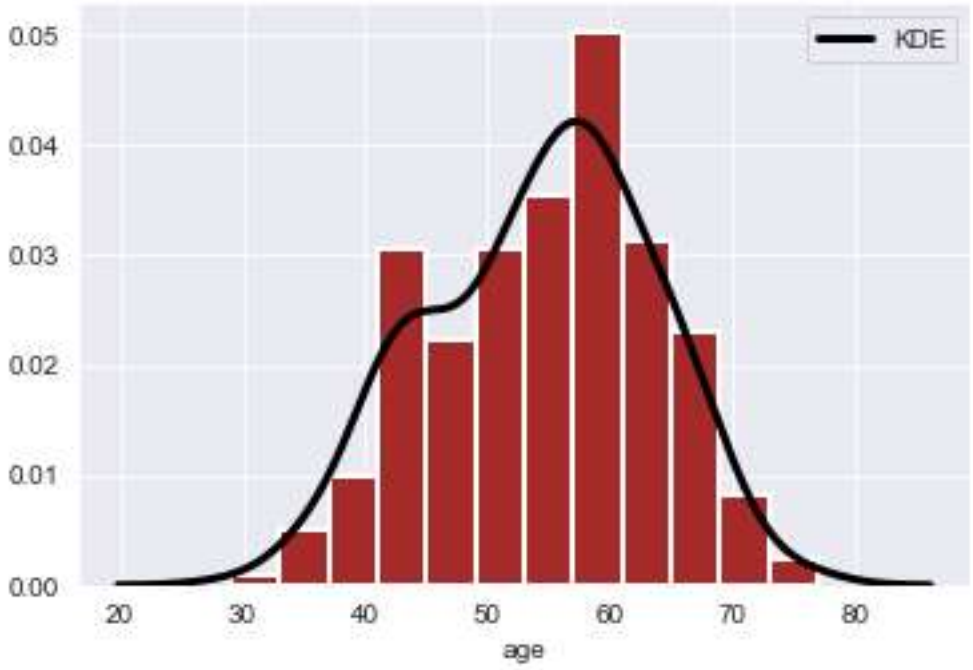
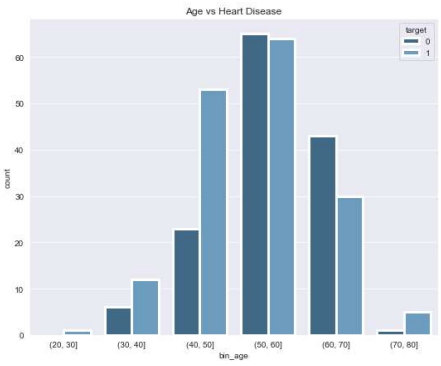
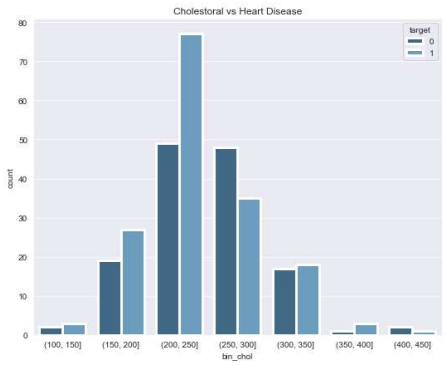


Figure 2. Distribution of people have heart disease in different age range: This bar graph shows the rate of people who have heart disease vs. people in a different age range. Age is one of the most effective factors to analyze/predict the heart disease rate because it is easy to recognize the difference among different age groups. According to this bar graph, people who are in the 40s – 60s age groups have higher heart disease rate. The ages among 60 – 65 have the highest heart disease rate and the ages among 20 – 29 and over 80 shows 0 heart disease rate.





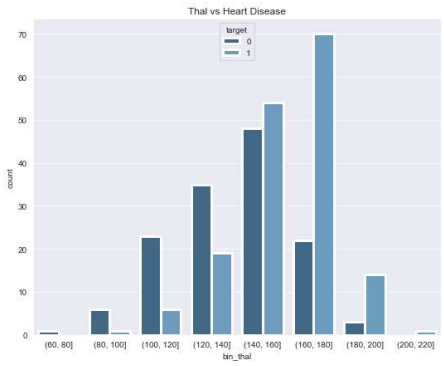


Figure 3. Heart Disease counts in different age bins, chol bins, and thal bins: Target 0 means they do not have heart disease and target 1 means they have heart disease. The first bar graph shows the count of people who have/does not have heart disease and different age bins (Age vs. Heart Disease). The people in the 40s – 60s age group have more heart disease compared to other age groups. The second bar graph shows the count of people who have/does not have heart disease and different cholesterol value bins (Cholesterol vs. Heart Disease). Most people get heart disease in 200 to 250 cholesterol range. From the second bar graph, the people who are over 250 cholesterol might appear to be at a higher risk of heart disease, but they do not have heart disease emerge at the same rate compared to the people who are in 200 - 250 cholesterol range. The last bar graph shows the count of people who have/does not have heart disease and different Thal value bins (Thal vs. Heart Disease). People who are in 160 – 180 thalach, they have a high possibility to have heart disease.

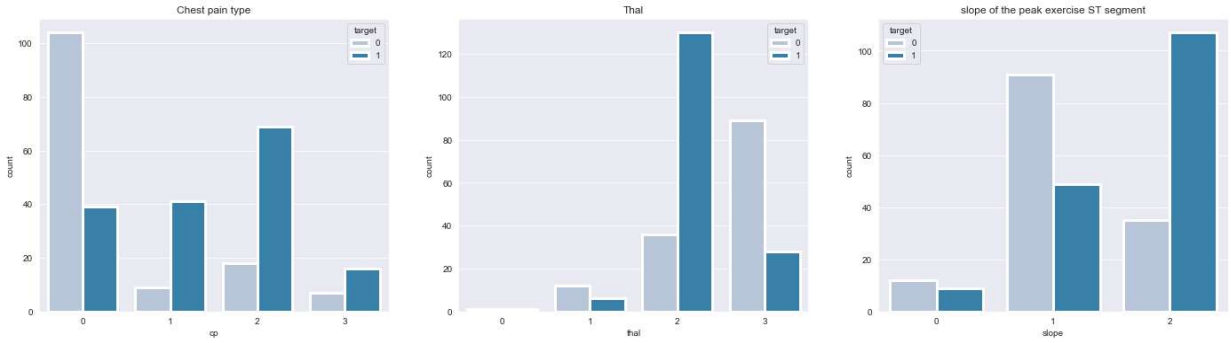


Figure 4. Heart Disease counts in different chest pain type, that, and slope of the peak exercise ST segment: Target 0 means they do not have heart disease and target 1 means they have heart disease. The first bar graph shows the count of people who have/does not have heart disease and different chest pain type (Heart Disease Count vs. Chest Pain Type). It shows the

people who are in chest pain type 2 would have the highest possibility to have heart disease. The second bar graph shows the count of people who have/does not have heart disease and different thal (Heart Disease Count vs. Thal). It shows the people who are in thal 2 would have the highest possibility to have heart disease. The last bar graph shows count of people who have/does not have heart disease and different slope of the peak exercise ST segment (Heart Disease Count vs. different slope of the peak exercise ST segment). It shows the people who are in slope 2 would have the highest possibility to have heart disease.

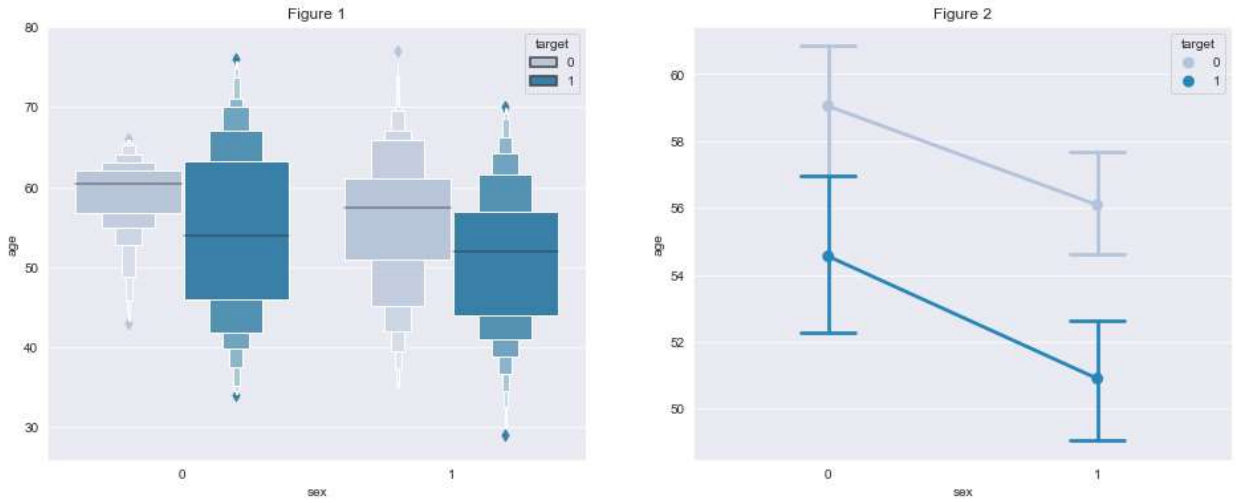


Figure 5. Heart Disease distribution in different age and sex, and mean age for female and male: Target 0 means female and Target 1 means male. The figure 1 graph shows most often the females would have heart disease in the 40s – 70s age range and most of the males would have heart disease in the 40s – 60s age range. The figure 2 graph shows the mean age of heart disease would be female: 55 years and male: 51 years.

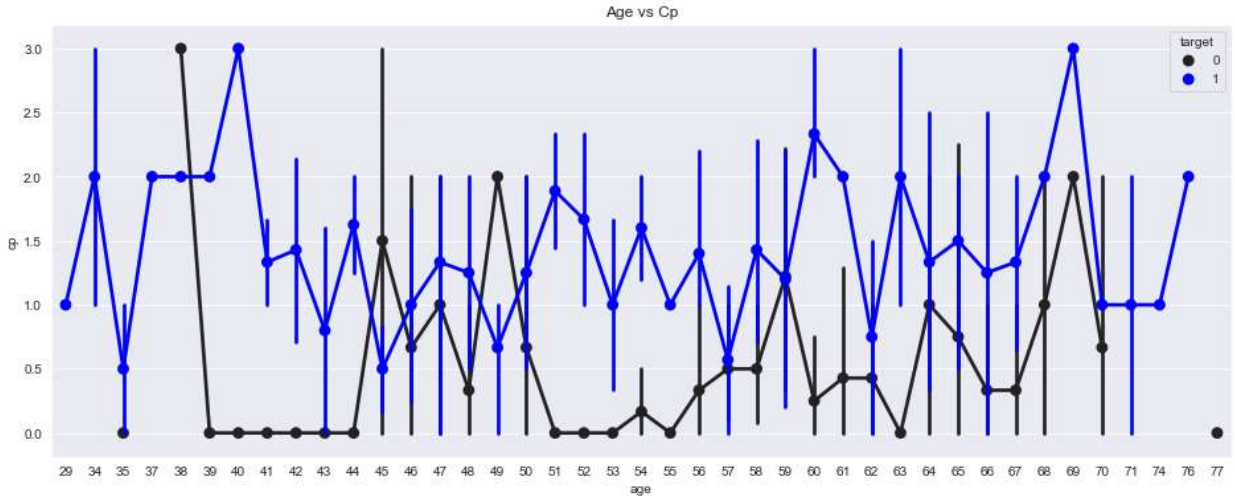


Figure 6. Distribution of people who have chest pain in a different age: Target 0 means they do not have heart disease and target 1 means they have heart disease. This subplot graph shows age vs. cp and people who have heart disease, they tend to have higher cp in entire ages except age 45 to 49.

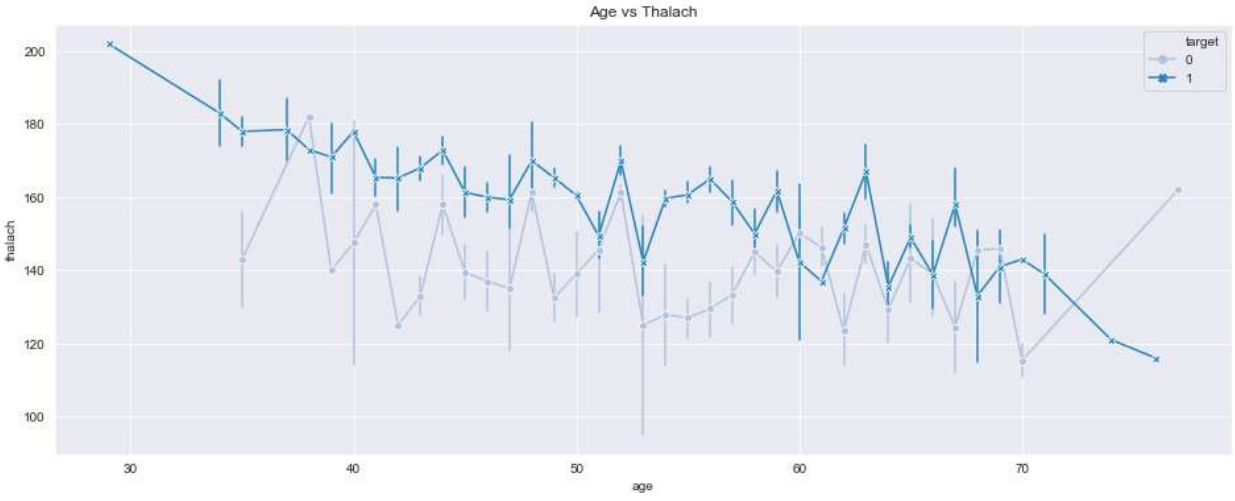


Figure 7. Distribution of people who have different thalach value in a different age: Target 0 means they do not have heart disease and target 1 means they have heart disease. This

subplot shows people who have heart disease always have high thalach and when their ages are increasing, the thalach seems to trend down. Other factors may play a similar/different role in heart disease.

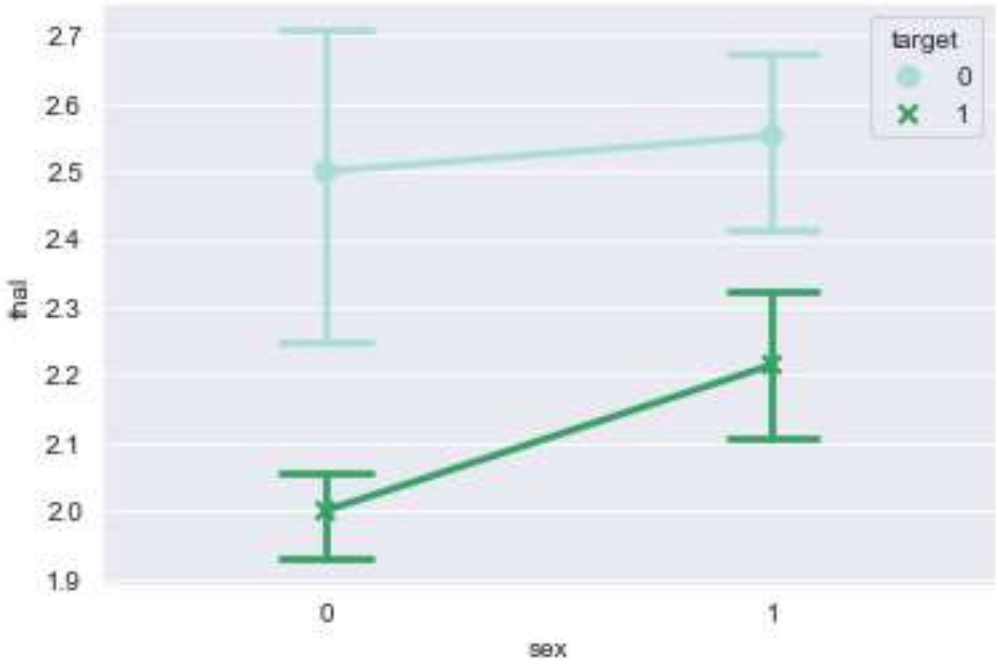


Figure 8. Mean of different genders in different thal value: 0 means female and 1 means male. This point plot graph shows both females and males who do not have heart disease have higher thalach and males who have heart disease have higher thalach than females.

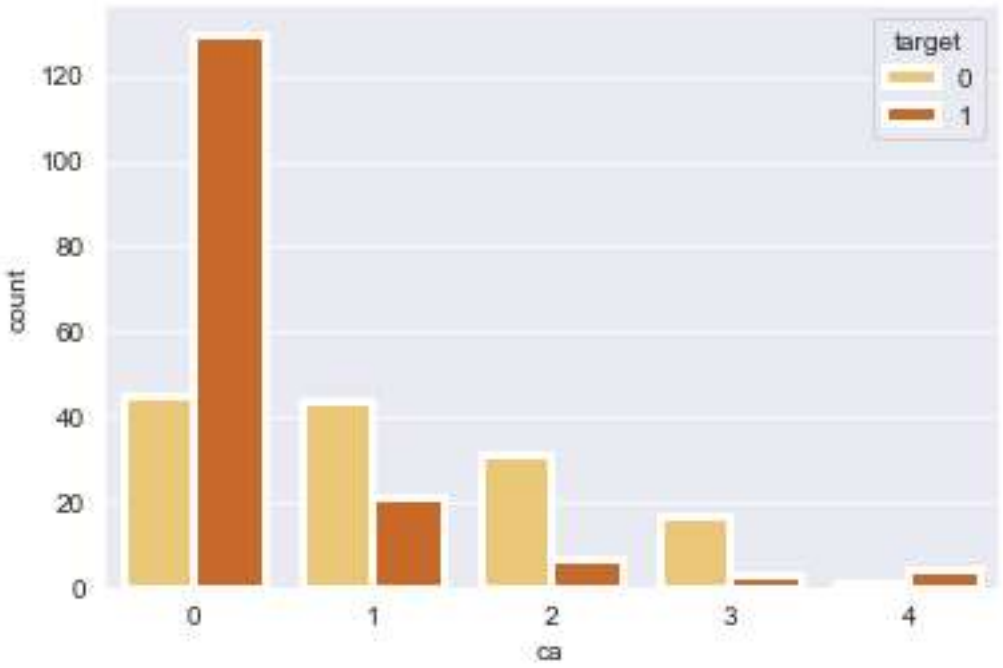


Figure 9. Count of people who have heart disease in different ca type: Target 0 means they do not have heart disease and target 1 means they have heart disease. This count plot shows people who have ca 0 have the highest possibility to have heart disease.

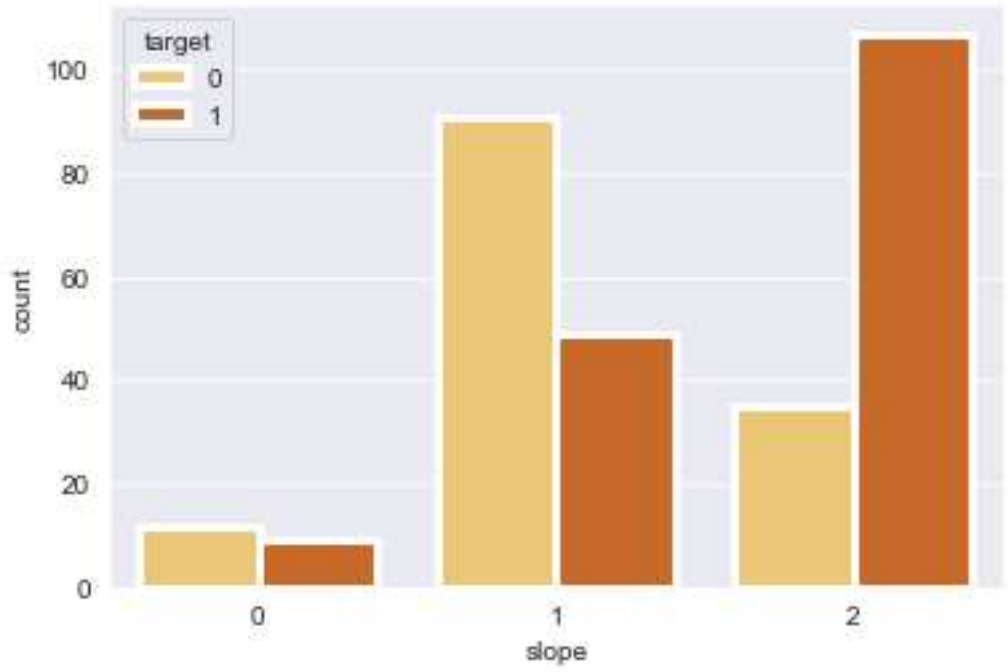


Figure 10. Count of people who have heart disease in different ca type: Target 0 means they do not have heart disease and target 1 means they have heart disease. This count plot shows slope 2 has the highest people who have heart disease.

	precision	recall	f1-score	support
0	0.95	0.84	0.89	44
1	0.87	0.96	0.91	47
micro avg	0.90	0.90	0.90	91
macro avg	0.91	0.90	0.90	91
weighted avg	0.91	0.90	0.90	91

0.9010989010989011

Figure 11. Logistic Regression result: Target 0 means they do not have heart disease and target 1 means they have heart disease. This logistic regression proves Hyperparameter that this model with logistic regression is 90.1 % accurate.

- **SAS Enterprise Miner**

I used SAS Enterprise Miner to execute logistic regression model which I used in Python and it gave me an output which shows the analysis of Maximum Likelihood Estimates. I set up target as a target variable because target shows whether the person has heart disease or not. The target attribute is the main key to show the result of my analysis/hypothesis about the questions that I mentioned in the background.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)	95% Confidence Limits	
Intercept		1	-2.4346	0.6728	13.10	0.0003		0.088	-3.7533	-1.1160
ca		1	-0.8830	0.2478	12.70	0.0004	-0.4994	0.414	-1.3687	-0.3973
cp		1	1.0332	0.2187	22.31	<.0001	0.6385	2.810	0.6045	1.4620
sex	0	1	1.0292	0.2711	14.41	0.0001		2.799	0.4979	1.5606
slope		1	1.8076	0.4226	18.30	<.0001	0.6144	6.096	0.9794	2.6358

Figure 12. Logistic Regression result of Analysis of Maximum likelihood Estimates:

P values of variable ca, cp, sex0, and slope are much smaller than 0.05 which means, the corresponding variables have the significant effect on the target variable(target) at 95% confidence level.

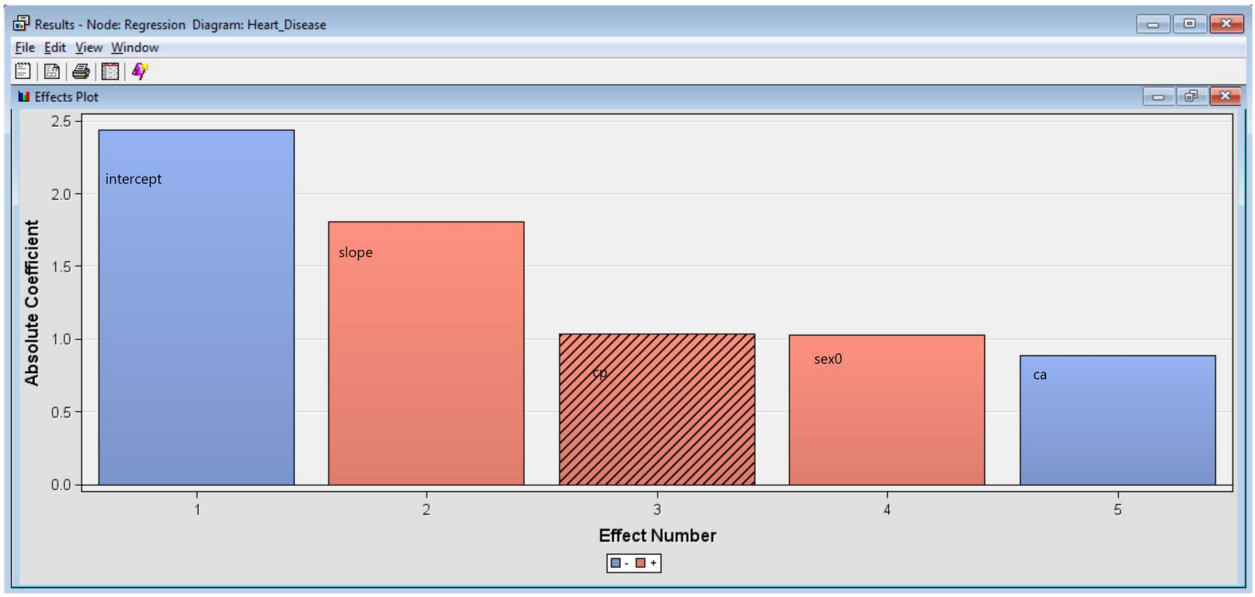


Figure 13. Bar chart of Absolute Coefficient of new model factors:

All of the variables have the absolute coefficient larger than 0.5.

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
cp		1	1.0000	1.0000	1.0000
thal		1	0.6533	0.4563	0.6985
slope		1	0.4451	0.0000	0.0000

Figure 14. Importance statistics: It shows the cp, thal, and slope variable importance.

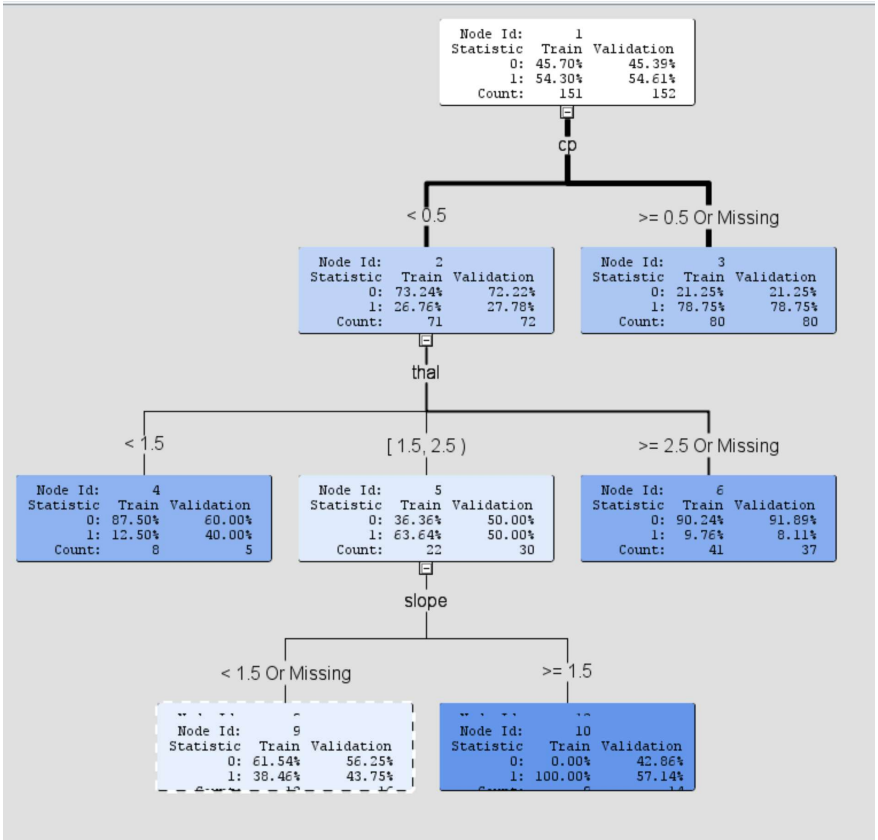


Figure 15. Decision Tree of the model: Statistic is based on Train and Validation. Target 0 means they do not have heart disease and target 1 means they have heart disease.

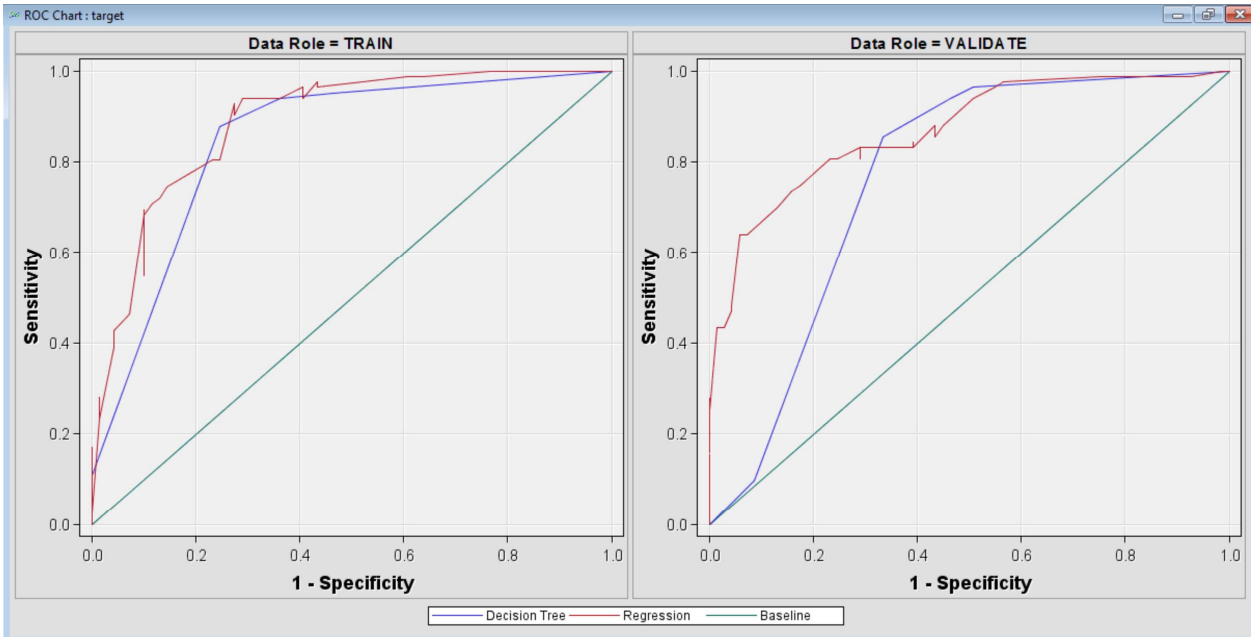


Figure 16. ROC Curves: Decision Tree and Regression show significantly difference between 0.2 to 0.4 in specificity and 0.6 to 1.0 in sensitivity.

Conclusions:

I wanted to determine the factors which could cause heart disease and the questions I addressed based on these factors. After analyzed the dataset with python, I found that age, chest pain, thal value, the slope of the peak exercise ST segment, and ca are the most possible factors to indicate heart disease. I found people who are in 40 to 70 years old, men in the 40 to 60 year range, and women from 40 to 70 years old are the most at risk for possible heart disease. For the chest pain, people with type 2 chest pain are significantly more likely to have heart disease. People not at age 40 to 49 with type 2 chest pain see an increase in their heart disease rate. In thal values, people who are with thal from 160 to 180 will have a higher chance to have heart disease. Also, slope 2 people have a higher chance to have heart disease and people with ca value 0 have the

highest chance of heart disease. Therefore, these results demonstrate the hypothesis based on solving my example questions which cause the heart disease is the attributes of age, chest pain, thal value, the slope of the peak exercise ST segment, and ca are the most possible factors to cause heart disease.

According to the Figure 11 to Figure 16, the accuracy of logistic regression is 90.1% which is high and SAS figures demonstrate the reliability of the model I created. Therefore, base on the result of visualizations, regressions, and ROC curve, the logistic regression model is reliable and the significant factors which affect the heart disease rate are age, gender, chest pain, thal(blood disorder) value, the slope of the peak exercise ST segment, and ca(The number of major vessels).

References

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

UCI Machine Learning Repository: Heart Disease Data Set, archive.ics.uci.edu/ml/datasets/Heart+Disease.