**Coronary Artery Disease Dataset – Analysis and Implications**

**Link to GitHub Repository:** <https://github.com/bmay51213/DSC-530-Final-Project>

**Background:**

For this course’s final project, I sought to analyze and explore a dataset regarding characteristics for heart disease. The dataset can be found at the following URL: [https://www.kaggle.com/ronitf/heart-disease-uci#](https://www.kaggle.com/ronitf/heart-disease-uci). The data here were obtained from 4 different databases from Cleveland, Hungary, Switzerland, and a VA Facility in Long Beach, California.

Coronary artery disease or Heart Disease is the presence of atherosclerotic plaques in the coronary arteries around the heart. It is more common as someone ages. Heart disease is a clinical disease entity that can be associated with many different other pathologies including high cholesterol (hyperlipidemia), high blood pressure (hypertension), diabetes (or impaired glucose tolerance). If someone is diagnosed with this based on angiography (a specific technique to image the arteries of the heart), the typical value would be a stenosis >50% to have heart disease. At this level of stenosis, there may be considerations for stenting and/or aggressive lifestyle modifications.

However, even in our current medical setting, coronary artery disease can be difficult to diagnose, symptoms can be elusive and not typical of coronary artery disease. Therefore, analyzing datasets like this could be useful to see what combination of risk factors could be predictive of the development of coronary artery disease. Obviously, those who smoke, have high blood pressure, and high cholesterol are highly likely to have coronary artery disease. I sought to learn more information about these factors to see if there are otherwise unknown variables or combination of factors that could be used to predict coronary artery disease in this dataset.

**Hypotheses and Key Questions:**

I hypothesize that the presence of coronary artery disease can be predicted using a combination of age, cholesterol level, systolic blood pressure, fasting blood sugar, findings on an EKG, and the type of chest pain. Further hypotheses include:

* Atypical angina will not significantly correlate with the presence of coronary artery disease.
* The greater the age, cholesterol level, and resting BP will have a positive correlation with the presence of coronary artery disease and will be a stronger association than individually.
* Those with a fasting blood sugar >120 (pre-diabetic range) and those with typical angina (angina score of 1) will have a significant positive correlation predictive of significant coronary artery disease.
* For those that are asymptomatic (angina score of 4), the higher the cholesterol, resting blood pressure, and fasting blood sugar will have a positive correlation with the chance of predicting coronary artery disease.
* For EKG findings, those who have EKG changes consistent with possible ischemia, will have a stronger association with CAD than those who have LVH criteria (those with enlarged hearts) on their EKG.

**Conclusions:**

After initial analysis, my dataset does not necessarily appear to be normally distributed. There are outliers in several variables including age, cholesterol, and blood pressure. However, these were not excluded on the basis that these are all within possible ranges when it comes to human physiology. Removing them would not necessarily be beneficial. There were approximate equal numbers of CAD presence and absence leading me to believe that our dataset was balanced. The vast majority of people were asymptomatic, had fasting blood sugars <120, did not have exercise induced angina (chest pain), and had 0 vessels imaged through fluoroscopy. Spearman correlation was used due to possible non-normal distributions and being more robust in the presence of outliers. Regression analysis was done using logistic regression due to the target variable being categorical.

**Key Findings:**

* Small positive correlations were noted between age and BP and age and cholesterol.
* Strong negative correlation (-0.43) between CAD presence and Exercise Induced Angina.
* Moderate negative correlation (-0.23) between CAD presence and Age. Graphs indicated higher amount of individuals at younger ages had CAD.
* Max HR on Stress Testing seemed to have a positive correlation with CAD presence.
* Logistic Regression:
  + Age, Cholesterol, and BP when correlated with CAD presence all had odd ratios less than 1 indicating that the higher these values, the less likely one is to have CAD. The only statistically significant variable was age (when controlling for cholesterol and resting BP) with P of 0.001.
  + Resting EKG changes and CAD presence had a statistically significant (P=0.017) odds ratio of 1.70 indicating that if one was to have abnormal EKG changes, you were 1.7x more likely to have CAD.
  + Blood sugar and presence of exercise induced angina had low odds ratios of 0.88 and 0.13 respectively when correlated with CAD presence. This indicates that those with exercised induced angina were far less likely to have CAD.
  + Number of vessels imaged had a statistically significant (P=0.00) odds ratio of 0.40 indicating as the number of vessels imaged increased, it was less likely for CAD to be diagnosed.

**Limitations:**

Since much of this data was categorical, I had to create dummy values in order to get a logistic regression model to run. Looking at the dataset source site, it appears that this is primarily being used in the machine learning space so more specialized techniques may be more valid. Furthermore, many of these findings directly contradict my hypotheses generated from my basic knowledge of cardiac risk factors for heart disease.

My current skills in Python do not allow me to do further machine learning algorithms to evaluate this and unsure how to specifically isolate which categorical variable I’m looking for in my regression analysis. It could be that logistic regression is simply not a good tool based on this dataset. I think an easier way to do an analysis here is to have specifically isolated variables, such as ST changes on EKG and LVH changes on EKG as two separate columns. This would make it easier to code and evaluate. Furthermore, the degree of stenosis, if quantified better would possibly lead itself to further numerical analysis.

Further directions would be to look at numerical data regarding the individuals regarding CAD presence vs. absence to better quantify relationships. Being able to see the actual main dataset and determine the same variables using numerical data may be easier to work with and use. This significantly limited my analysis as I am unfamiliar with using categorical variables for analysis in Python though I learned new methods to analyze the data. I would also be interested in knowing the baseline demographics of the studies population as I do not think that it was sampled equally and that there appears to be a higher likelihood of CAD presence at younger ages which is opposite from what would believe.