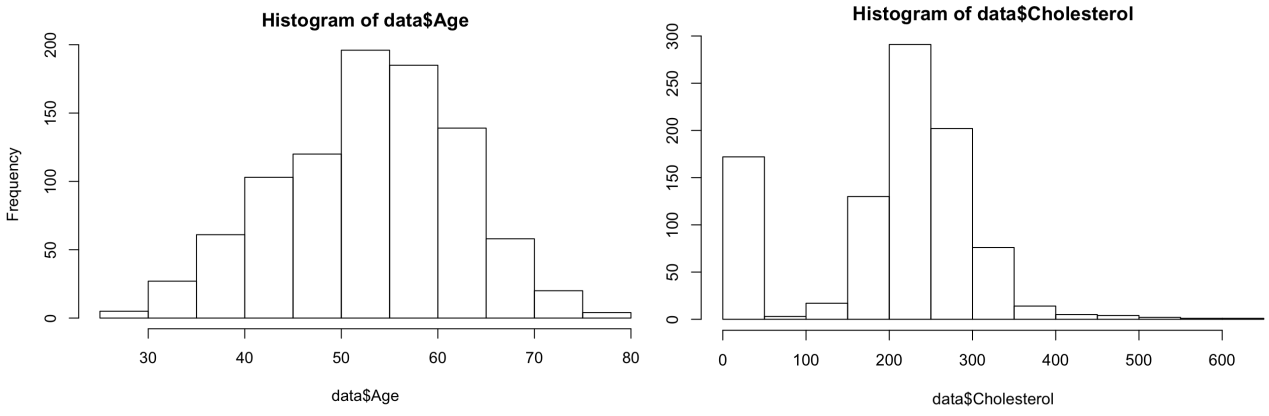
Jacob Hurley, Sofia Gray

Data 467

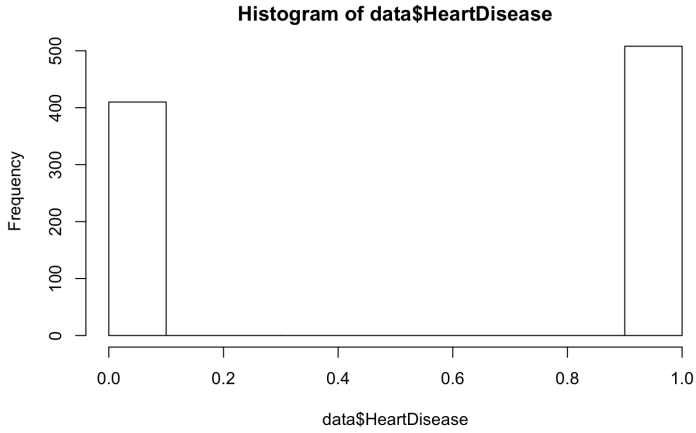
9/30/2021

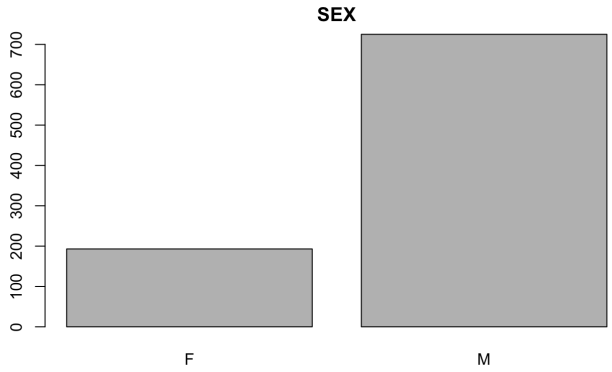
Results

We started by converting our data set into a data frame to make it easier to read and analyse. We found the summary statistics for each variable and found the range of possible values for each. From there, we found the Pearson correlation between each variable. By performing correlation tests on the numerical data, we observed that the range of correlations of any two variables were between -0.5 and 0.5. This shows that there is no strong correlation between any two variables. Because there are no strong correlations between any two variables, we plan on building a predictive model to single out which variables determine whether someone has heart disease or not. When plotting the numerical data, we observed that some of the data is not applicable. For example, some of the individuals reported their cholesterol level to be 0 which is not possible, as well as some of the ST_slope data reported at less than 0 which is also not possible. In our analyses, we plan on omitting these data sets in order to build a more accurate model. We also observed that the distribution of some variables seemed to follow the normal distribution. The age of individuals, the cholesterol levels (excluding a few outliers), and maximum heart rate all seem to follow the normal distribution. This can be seen in the histograms below.



Some variables, such as heart disease and sex, are binary. The distribution for whether a person has heart disease or not seems to be equally distributed, half of the sample seem to have it and half seem to not. This is good as it will allow for an analysis of the response variable without having unwanted bias. If the response variable had a large majority of either, then the model may predict heart disease happening more often than is the case for the population. We predict that the older you are with a cholesterol level of 200 or higher, the greater your chances are of having heart disease.





The barplots above show the number of people who have or don't have heart disease and how many of them are male or female. As we can see, the share of those who have heart disease is in the same amount as those who don't but there seems to be about three times as many men as there are women. So we can already tell from these barplots, that heart disease seems to be more common in men.

Code:

https://github.com/JacobHurley/Data467