# Project Paper 1

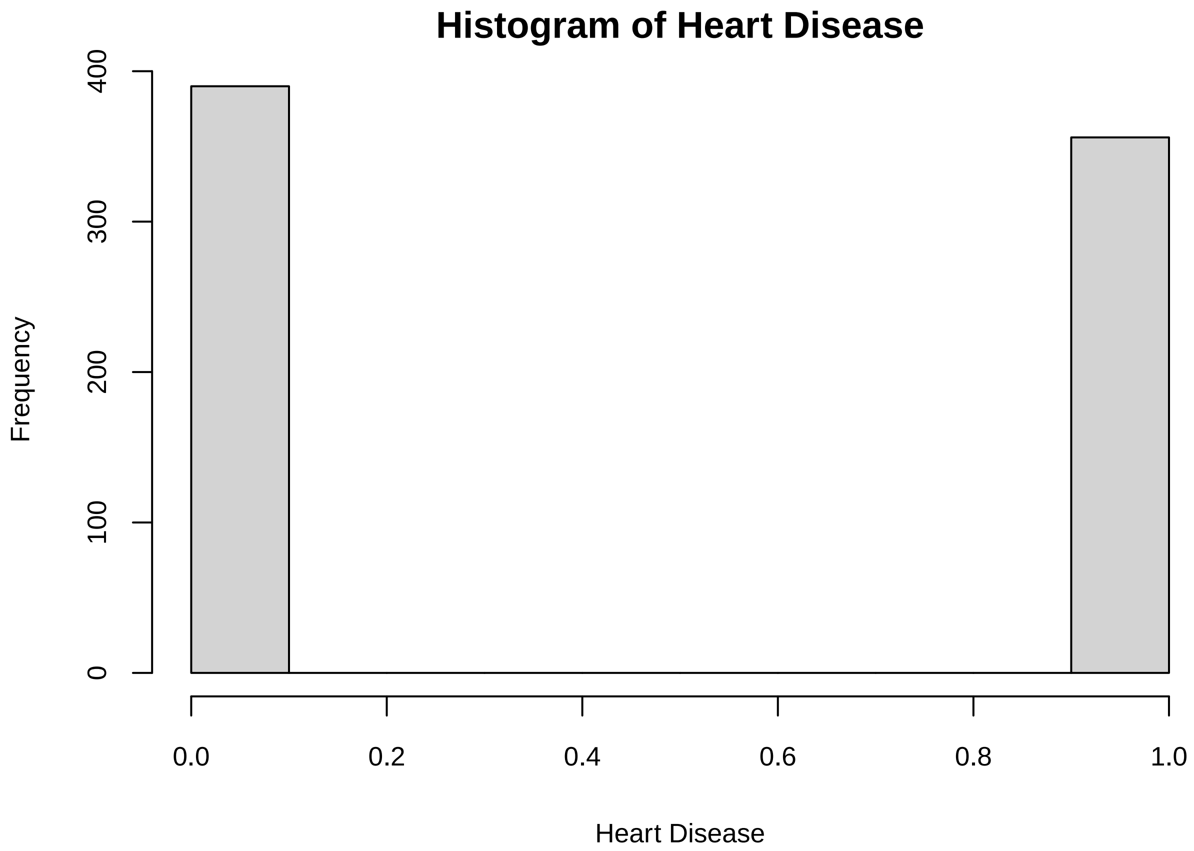
Alex Ojemann

2023-03-19

## Introduction

My project explores the prediction of heart disease based on some common bio indicators. This data is collected from five different hospitals and contains relevant indicators of cardiovascular disease. It's interesting to me because cardiovascular disease is the #1 cause of death globally and it is often said to be significantly related to factors that we can easily measure like blood pressure and cholesterol so it seems like an excellent application for regression. In this paper I will be doing some exploratory data analysis.

## Response Variable

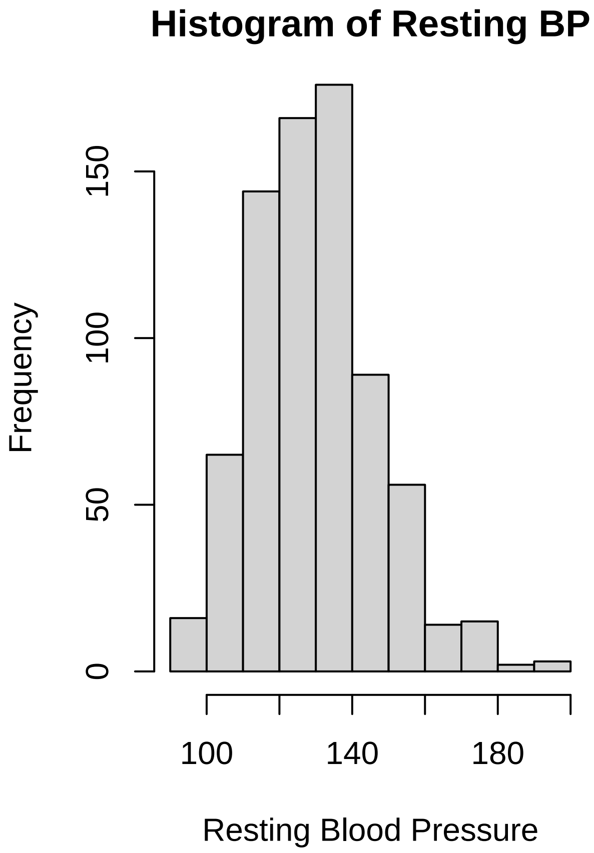


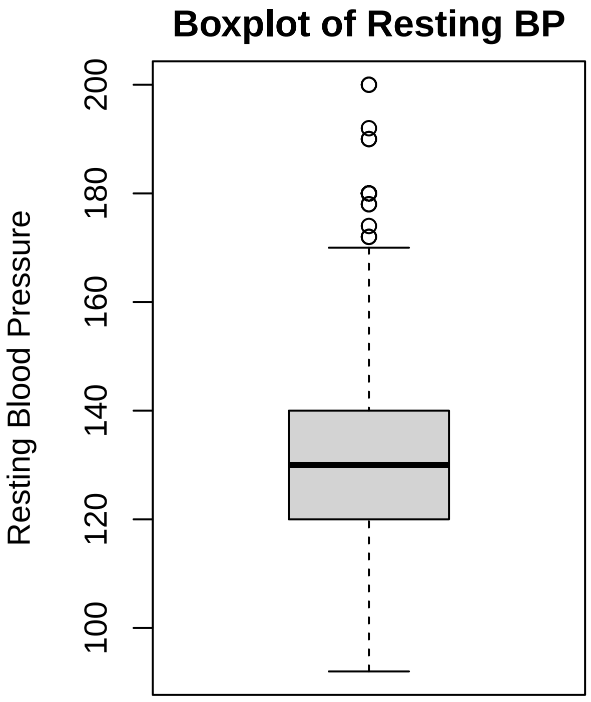
My response variable is binary, so there can't be any outliers. We can see that there are more patients that do have heart disease that don't.

Outliers are not applicable here because our response variable is binary. The distribution containing more instances of heart disease than instences of no heart disease reflects that this data is not representing the general population of people living near these hospitals, but rather the population of patients at the hospitals, specifically those who are at great enough risk of heart disease to have their biometrics in this data sets measured.

We are 95% confident that the true proportion of patients from these hospitals that have heart disease is between 0.5212175 and 0.5855363.

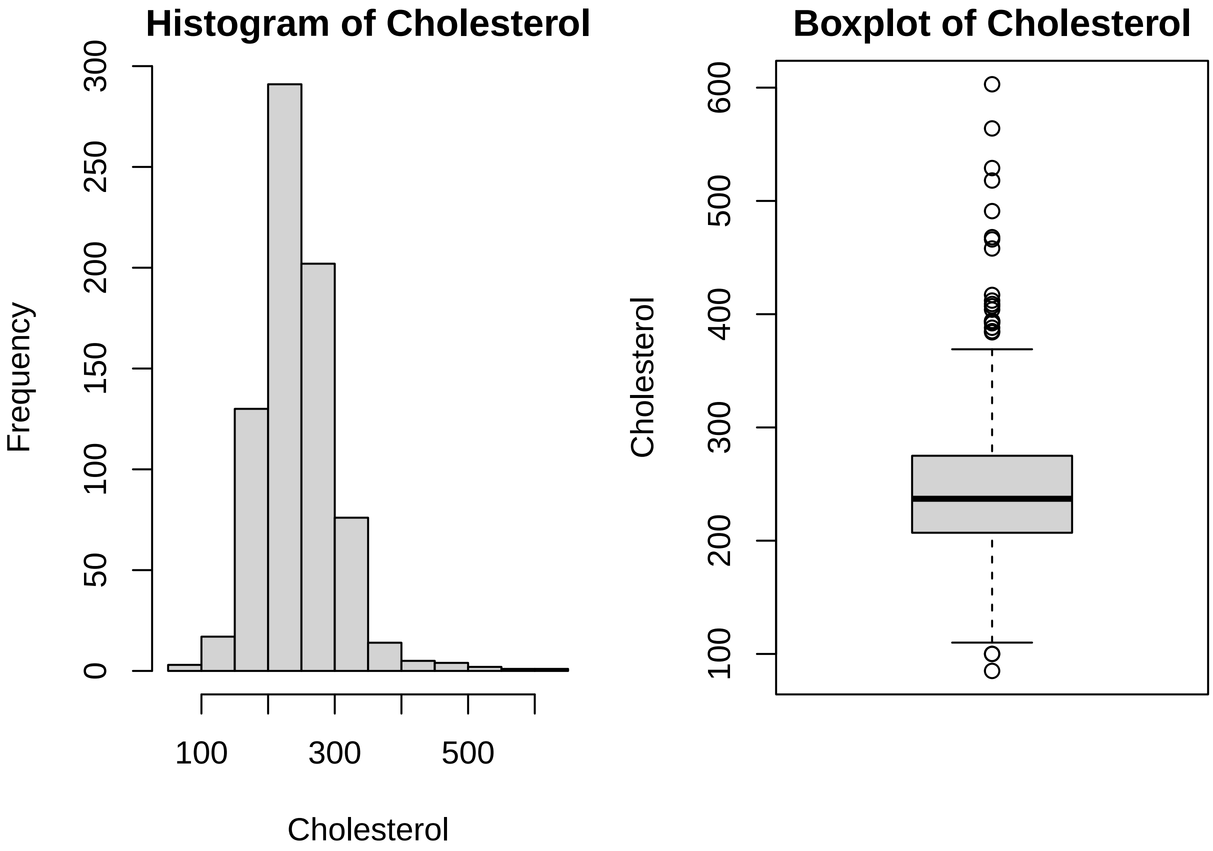
## Predictor Variables





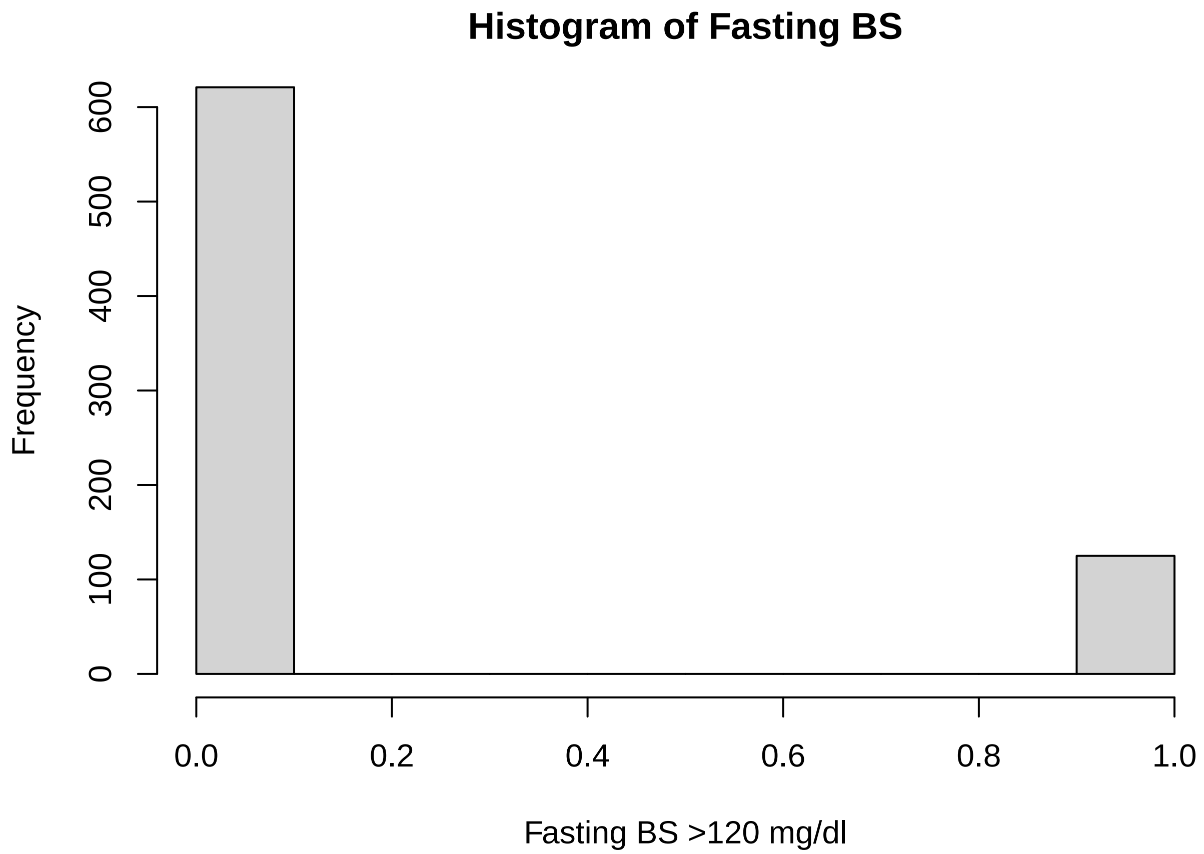
The RestingBP variable is approximately normally distributed but has a number of outliers. Most of the outliers of the RestingBP variable as shown on the boxplot are within reason, however one of them is at 0. This entry will be removed as a blood pressure of 0 mm Hg cannot exist in a living being because blood flow wouldn't be possible.

We are 95% confident that the true mean resting blood pressure of patients from these hospitals is between 132.5024 and 132.5794.



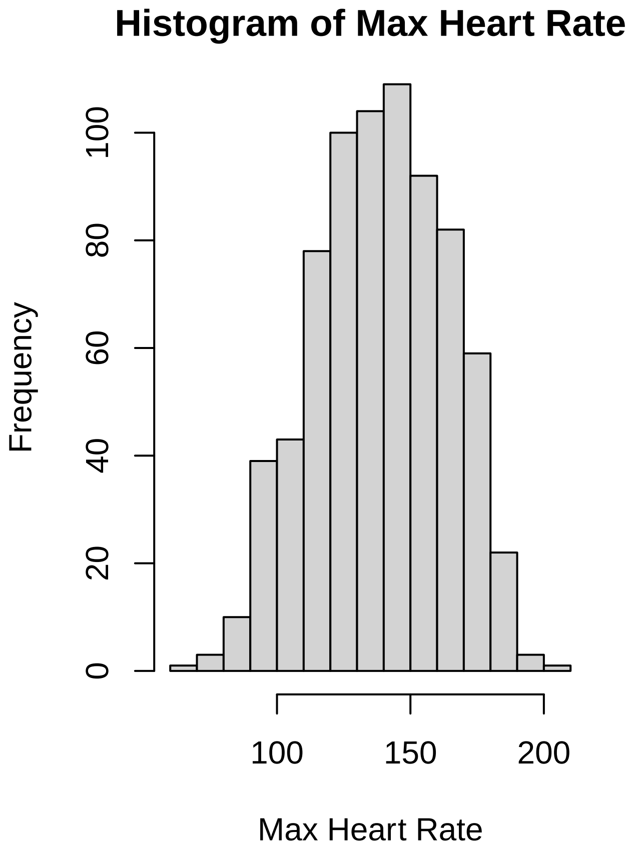
The Cholesterol variable is approximately normally distributed with a slight right skew, however it has many outliers. The outliers of the cholesterol variable above the top whisker in the boxplot aren't a concern for our purposes because given a patient is in this database they were likely deemed enough of a heart disease risk to have these measurements taken so significantly higher cholesterol than normal is possible. However it's not possible to have a cholesterol level of 0 mm/dl so those points must be removed.

We are 95% confident that the true mean cholesterol of patients from these hospitals is between 244.48 and 244.7908.

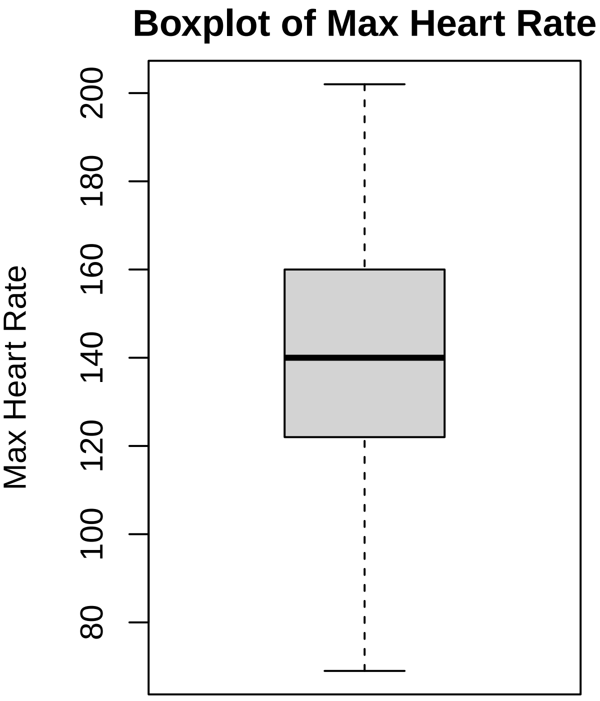


The FastingBS variable represents whether a patient's fasting blood sugar is above 120 mg/dl which is binary so outliers are not possible. The distribution is heavily slanted towards not having a fasting blood sugar above 120 mg/dl. This means a lot of nuance is likely being left out in the group of patients below 120 mg/dl. For example, a patient with 50 mg/dl of fasting blood sugar is treated the same as a patient with 110 mg/dl, which isn't ideal for prediction.

We are 95% confident that the true proportion of patients from these hospitals that have fasting blood sugar levels above 120 mg/dl is between 0.14076 and 0.1943607.



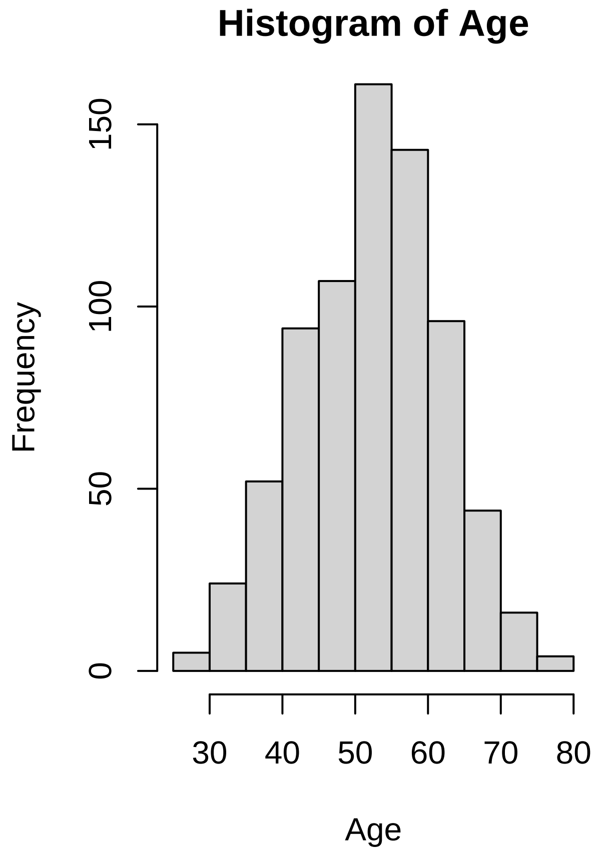
Histogram of Max Heart Rate

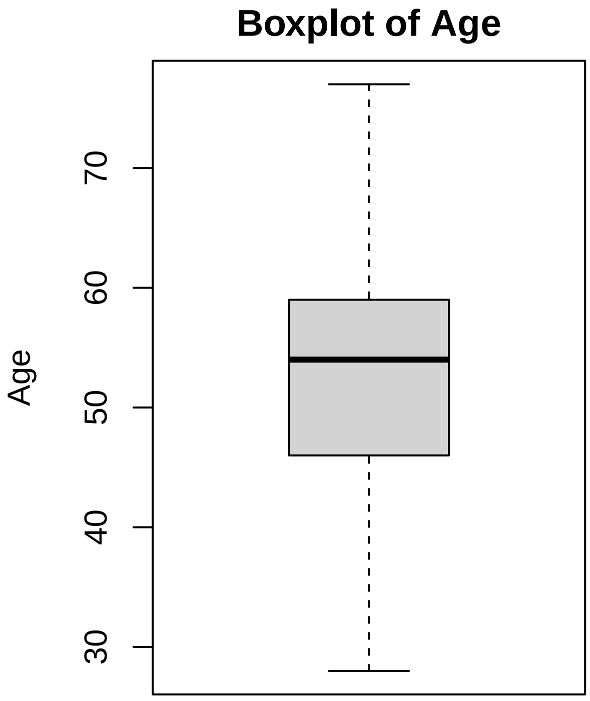


Boxplot of Max Heart Rate

The MaxHR variable is very close to normally distribued with no outliers, which is ideal for prediction.

We are 95% confident that the true mean maximum heart rate of patients from these hospitals is between 140.1621 and 140.291.



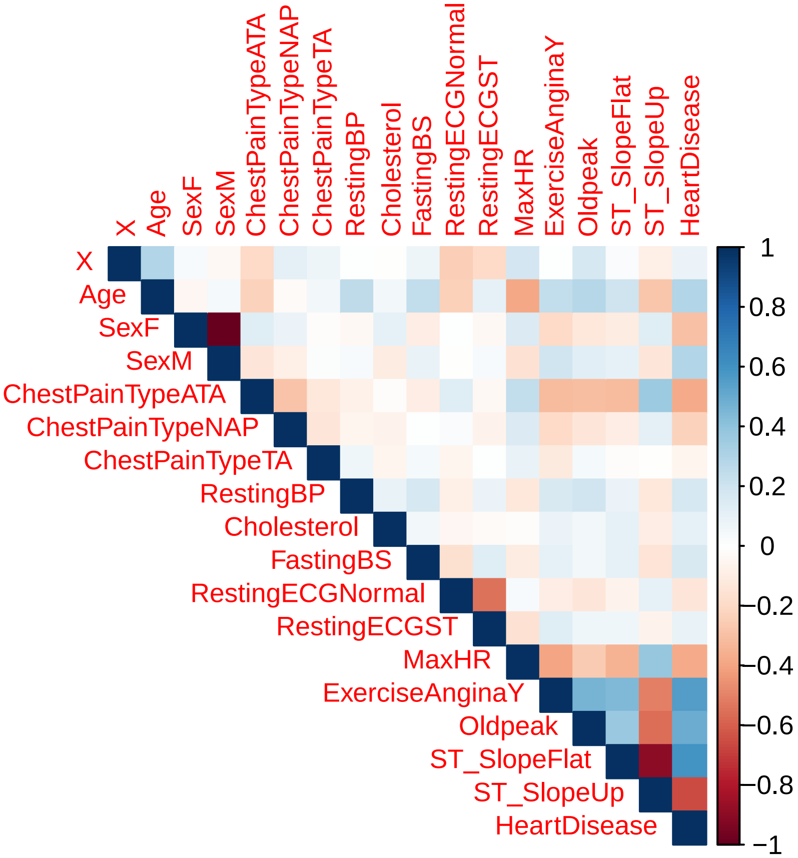


The Age variable is very close to normally distributed with no outliers. This is ideal for prediction. However, the age variable having a normal distribution centered around 55 doesn't align with the ages of the entire human population. It shows that the patients in this data set are generally older than the average person.

We are 95% confident that the true mean age of patients from these hospitals is between 52.85706 and 52.90701.

## Multicollinearity

```
## corrplot 0.92 loaded
```



None of the predictor variables have a correlation > 0.5 (moderate) other than one hot encoded variables within the same category (ST_SlopeFlat and ST_SlopeUp for example).

Most of the relatively high correlation values are between ExerciseAngina, OldPeak, and ST_Slope. These variables may be more advanced and associated with one another, however their correlations are all less than 0.5 thus it's not worth holding them out of feature selection. One of the few pairs of features outside these three that have a correlation magnitude above 0.35 is age and maximum heart rate, which corroborates domain knowledge, but once again the correlation isn't strong enough to assume they're redundant.