Cardiovascular disease (CVD) encompasses a litany of medical maladies, like coronary artery disease, myocardial infraction, stroke, heart failure, etc. There are also a multitude of causes of CVD, like atherosclerosis, which are caused by things like high blood pressure, poor diet, high cholesterol, etc. Thus, if we can find a way to screen patients for CVD, we could prevent a lot of deaths before they happen. That is what we intend to do with here.

This dataset consists of 70,000 observations of patients either with or without CVD. We want to predict whether a patient has CVD or not. This is an example of supervised learning; specifically, because having CVD or not is a categorical variable, this is a classification problem. We will use a number of machine learning models for predictive purposes.

Exploratory Data Analysis:

The first thing I did was take a look at the structure of the dataset. There are thirteen variables:

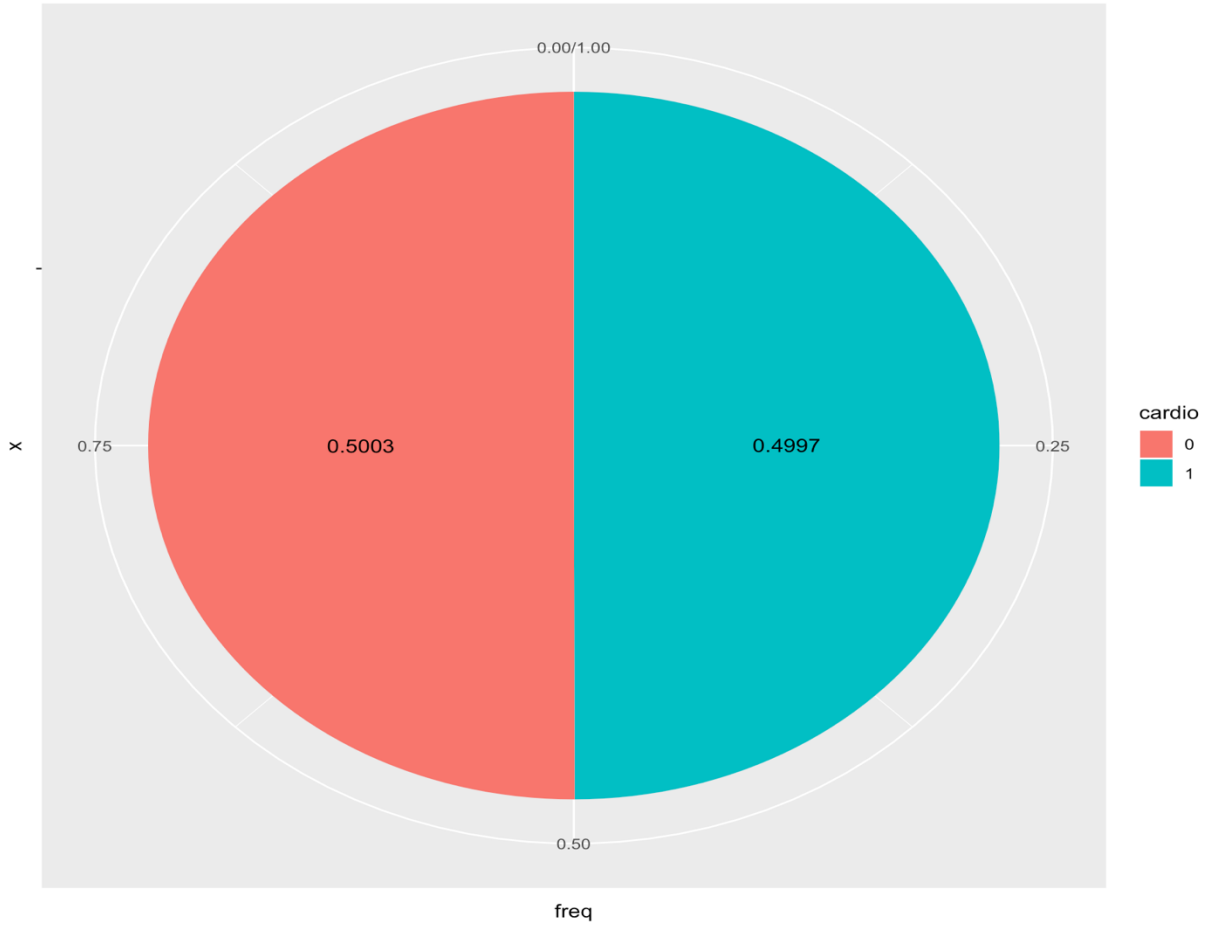
- ID number (int)
- age (int, in days)
- gender (categorical -- int,1 for women, 2 for men)
- Height (int, in centimeters)
- Weight (int, in kilograms)
- Ap_lo, or the diastolic blood pressure (int)
- Ap_hi, or the systolic blood pressure (int)
- Cholesterol (categorical -- int, 1 for normal, 2 for above normal and 3 for well above normal)
- Glucose (categorical -- int, 1 for normal, 2 for above normal and 3 for well above normal)

- Smoking (categorical – int, 0 for not smoking, 1 for smoking)
- Alcohol intake (categorical – int, 0 for no/little alcohol consumption, 1 for alcohol consumption)
- Physical activity (categorical – int, 0 for not active, 1 for active)
- Cardio (categorical – int, 0 for not having cardiovascular disease, 1 for having CVD – our target variable)

Note that alcohol intake and physical activity are subjective features here. This could be a potential cause for concern.

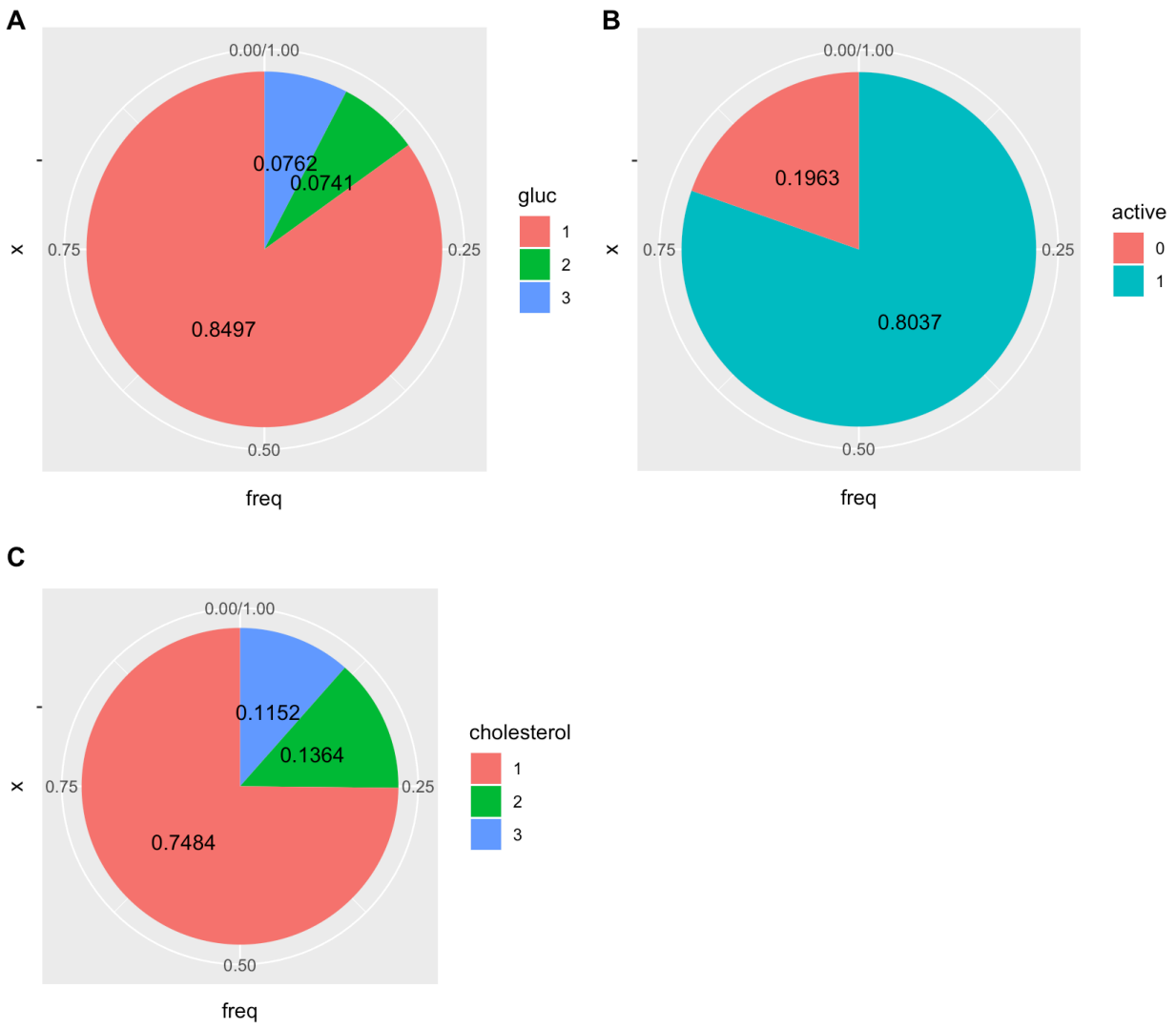
I removed the ID number from the dataset. I then checked to see how many missing values were in the dataset, and, luckily, there were none. Finally, I converted all the categorical variables from ints to factors.

Now I wanted to look at the distribution of each variable. First off, let's see how our target variable, Cardio, looks like:



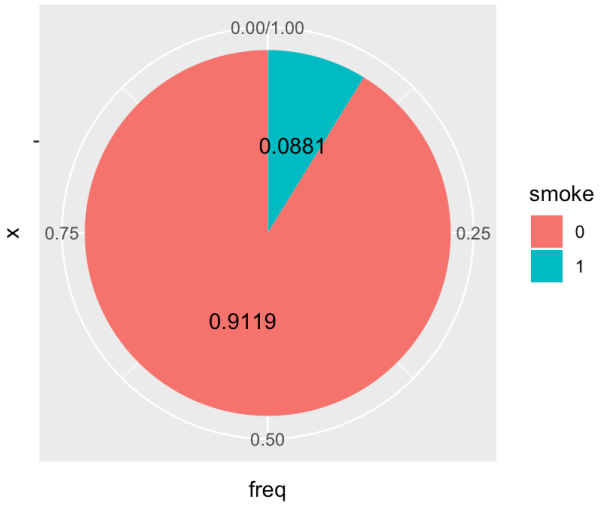
There is a pretty equal distribution of those with CVD to those without CVD. In other words, this is a balanced dataset.

Next up is the distribution of the categorical variables (CV). Note I split up the five CVs into two charts of three and two CVs apiece for easier readability:

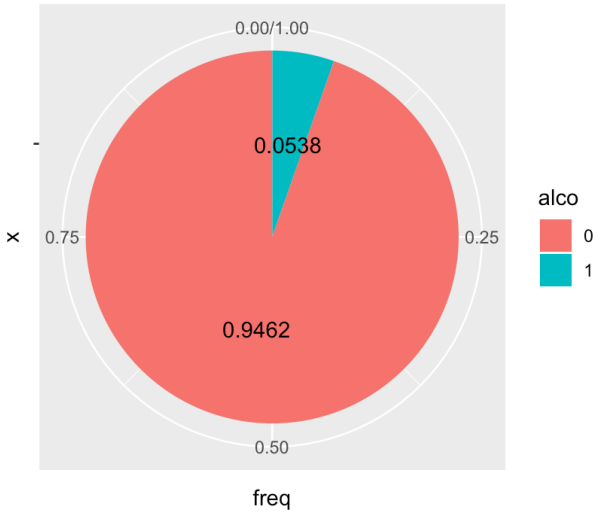


The distributions of glucose and cholesterol are fairly similar: both have a large percentage of normal levels and much smaller percentages of above normal/well above normal levels. The distribution of active-ness is similar in the sense that there is a large percentage of one level and a much smaller level of the other level (being active/not being active respectively), though in this case, this is a binary response instead of a tri-nary one.

A

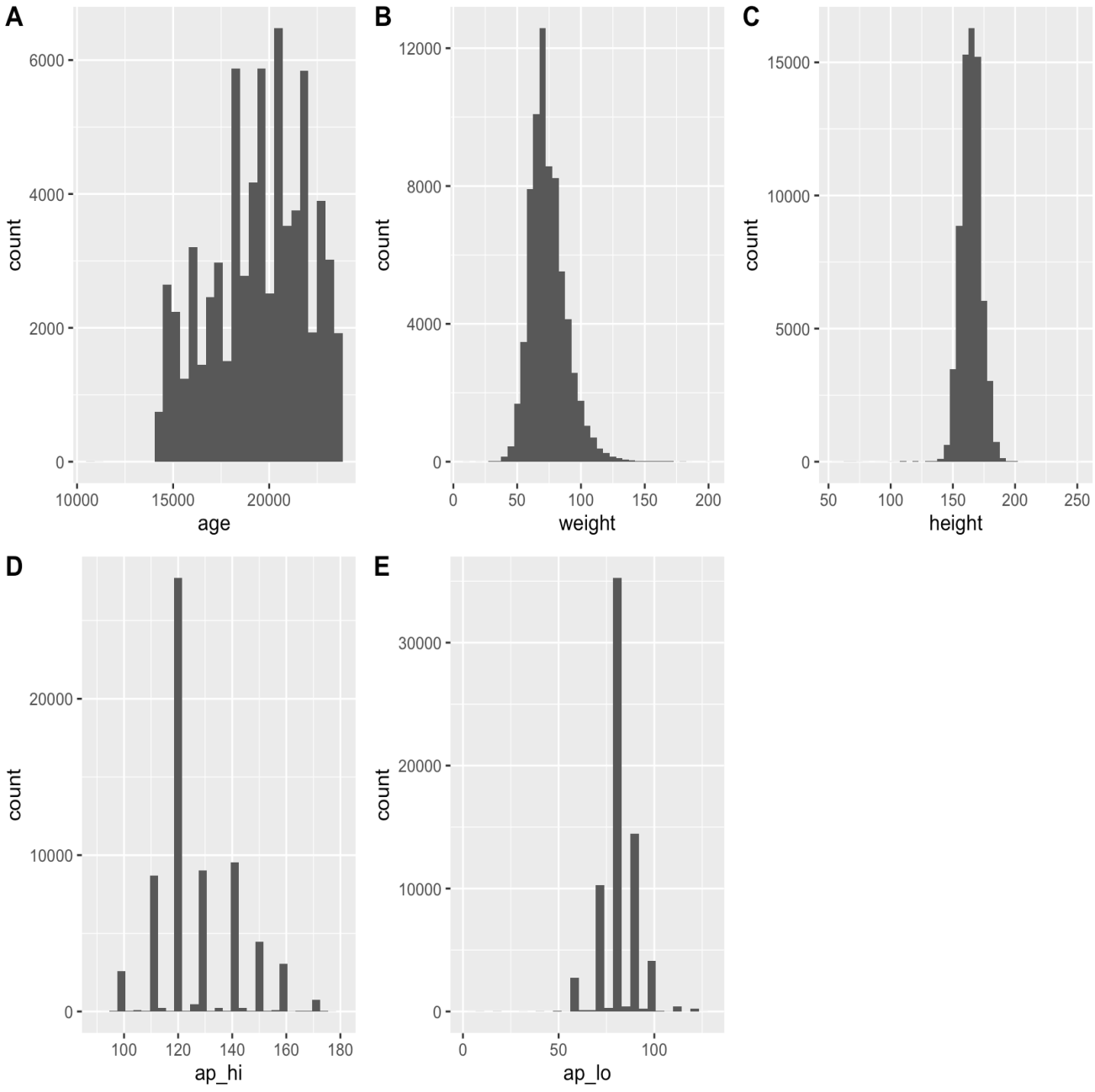


B



The distributions of smoking and alcohol intake are similar, but more extreme, than that of being active. The percentages of reported non-smokers and low/no alcohol intake are much greater than those who do smoke/do consume alcohol. Moreover, the 91.19% and 94.62% of non-smokers/alcohol consumers is far greater than the 80.37% of those who are active, which is why their distributions are more extreme than the distribution of active-ness.

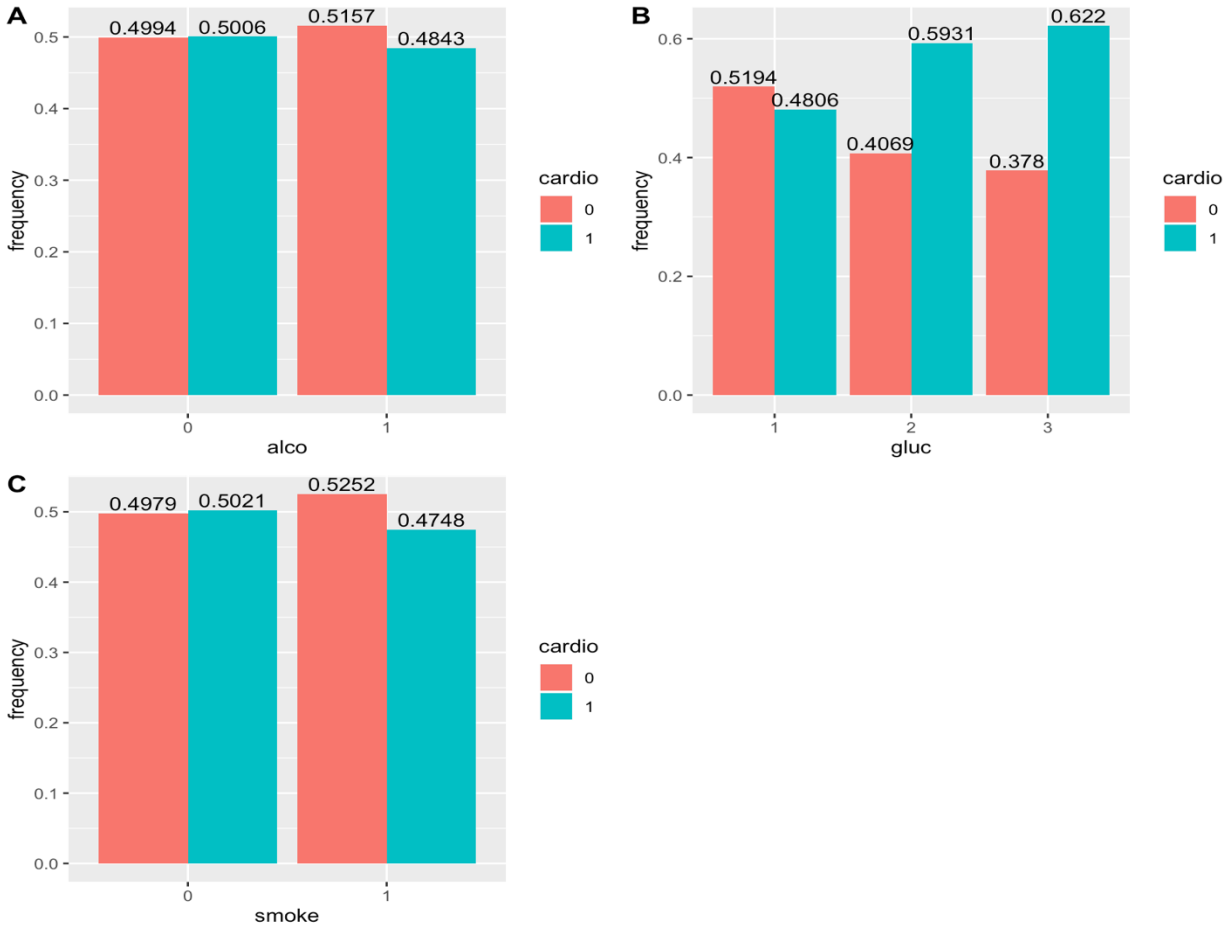
Next, let's look at the distribution of continuous variables:



Everything seems pretty standard for the age, weight height distributions. The weight and height plots are fairly normally distributed, with peaks around 75 kilograms and 160 centimeters respectively. There is some right skewness to the weight distribution; that accounts for overweight/obese patients. The age distribution seems pretty standard too: there are multiple peaks, but this does not seem abnormal. In contrast, the systolic and diastolic blood pressure

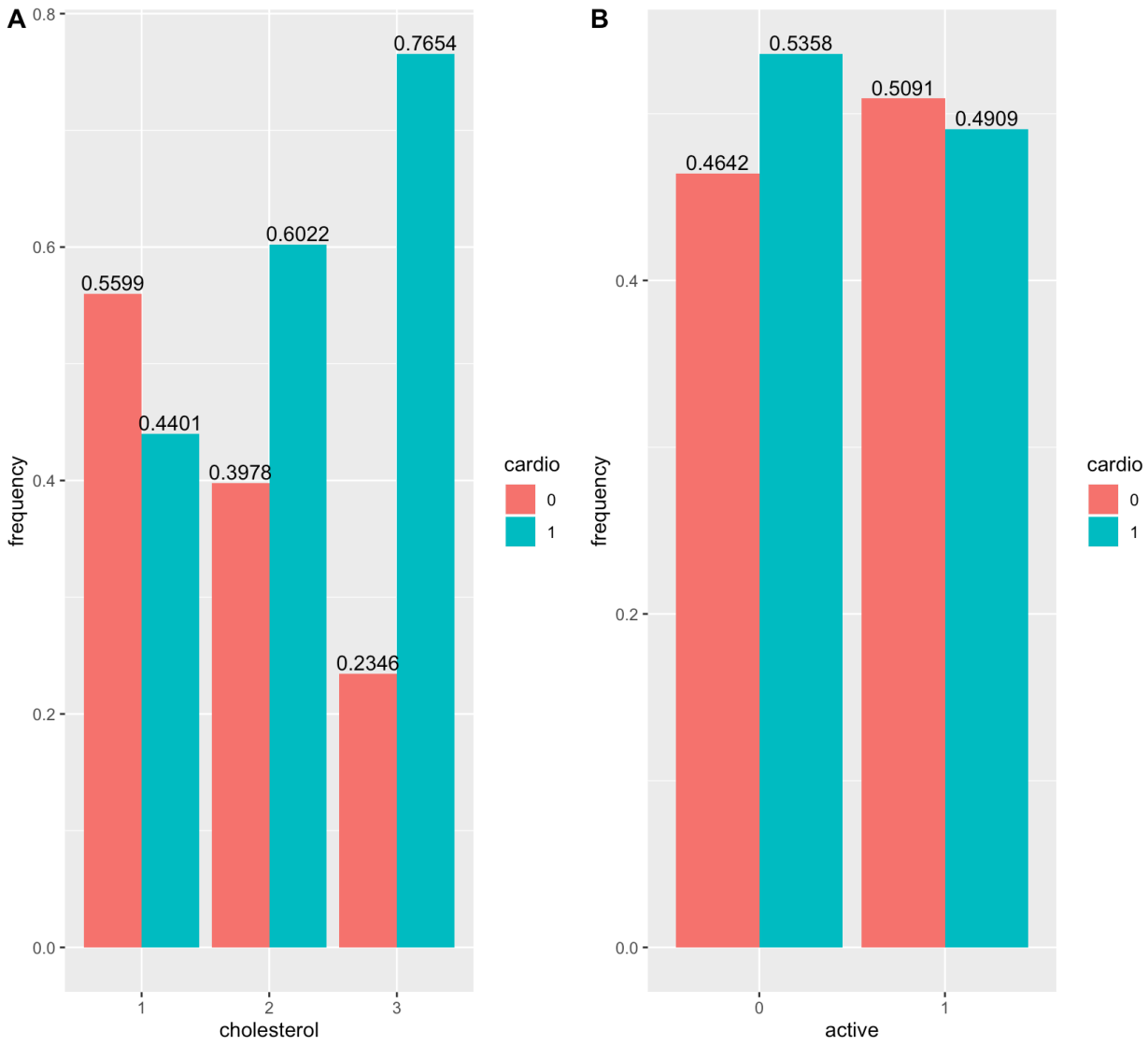
distributions look interesting. There are massive peaks and almost nothing in between the peaks. I looked more closely at the diastolic and systolic blood pressure columns and I found that every single observed value was a multiple of ten. So the systolic values were 100, 110, 120, 130, etc. and the diastolic values were 60, 70, 80, etc. There are no in-between values.

Now that we've looked at each variable on its own, let's see how they relate to our target variable, cardio. We will first look at the relationship between the CVs and cardio. Again, we will split up the five CV-cardio graphs into three and two for readability's sake:



There is little difference in the relationship between smoking/alcohol intake and having CVD. These variables can be seen as non-essential, and will not be included in any predictive

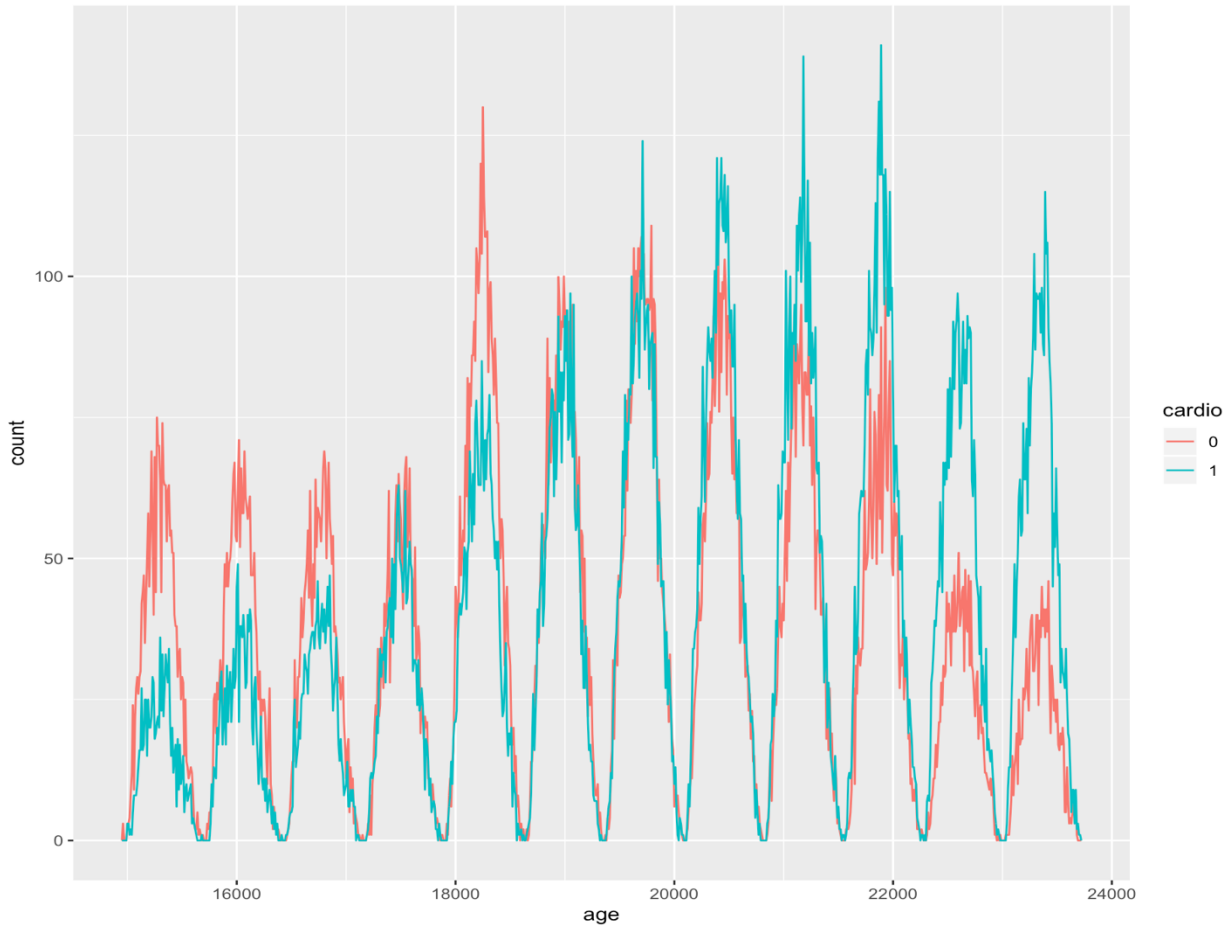
models. Glucose levels does seem significant, as the worse off your glucose levels are, the more likely you are to have CVD.



Cholesterol levels is clearly important to whether one has CVD or not, while active-ness is tricky; it definitely is more important than alcohol/smoking, but less important than cholesterol/glucose. I need to check the significance/p-value of active-ness before making a decision on whether to include it in my models or not.

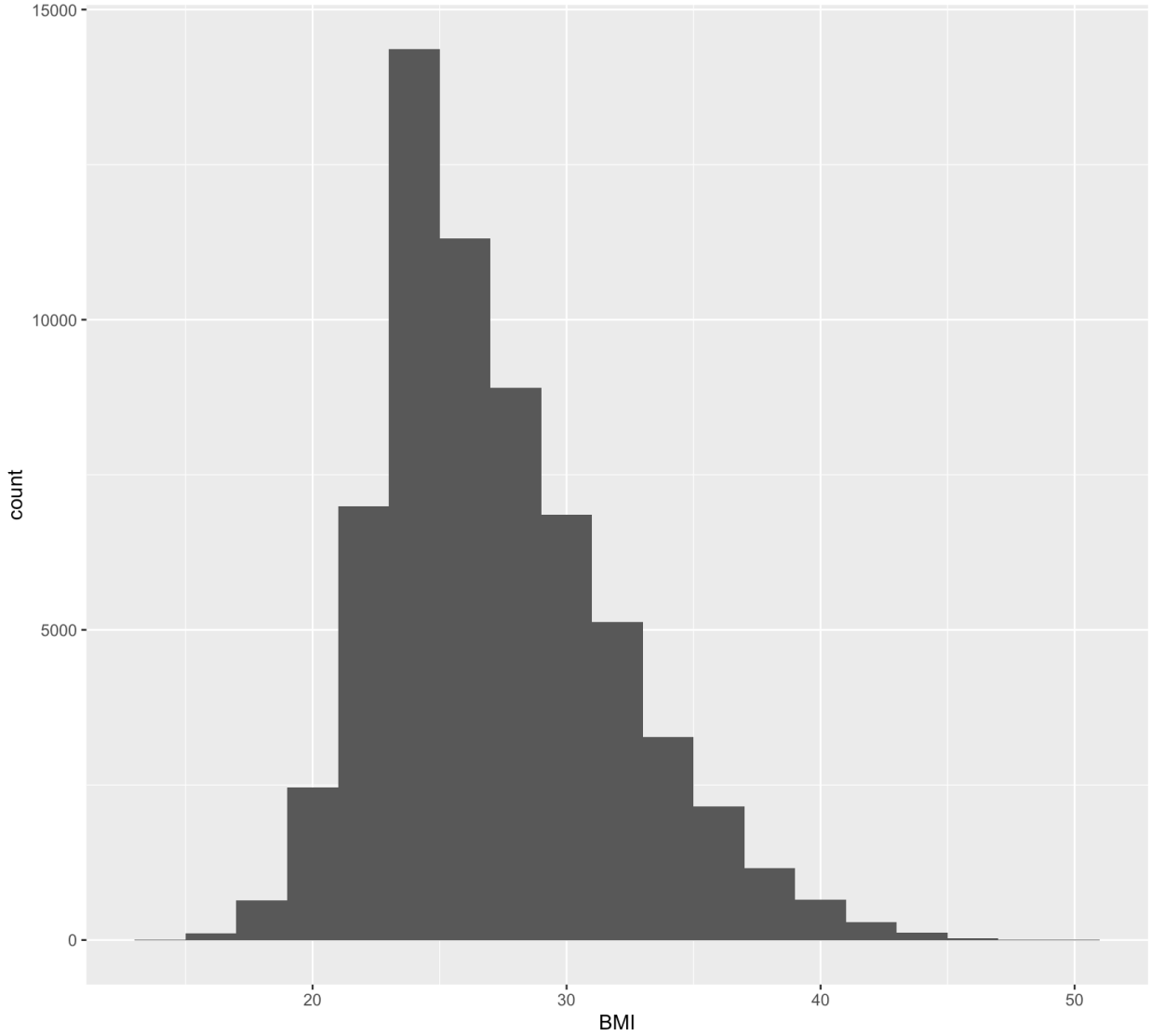
Before looking at the relationship between the continuous variables and having CVD, I wanted to get rid of outliers. I used the Q1-1.5*IQR/Q3 + 1.5*IQR method of finding outliers. There were lower and upper outliers for height, weight, systolic blood pressure and diastolic blood pressure, so I removed them from the dataset. Moreover, there were only upper outliers for age, so those were removed.

I first compared age with having CVD:

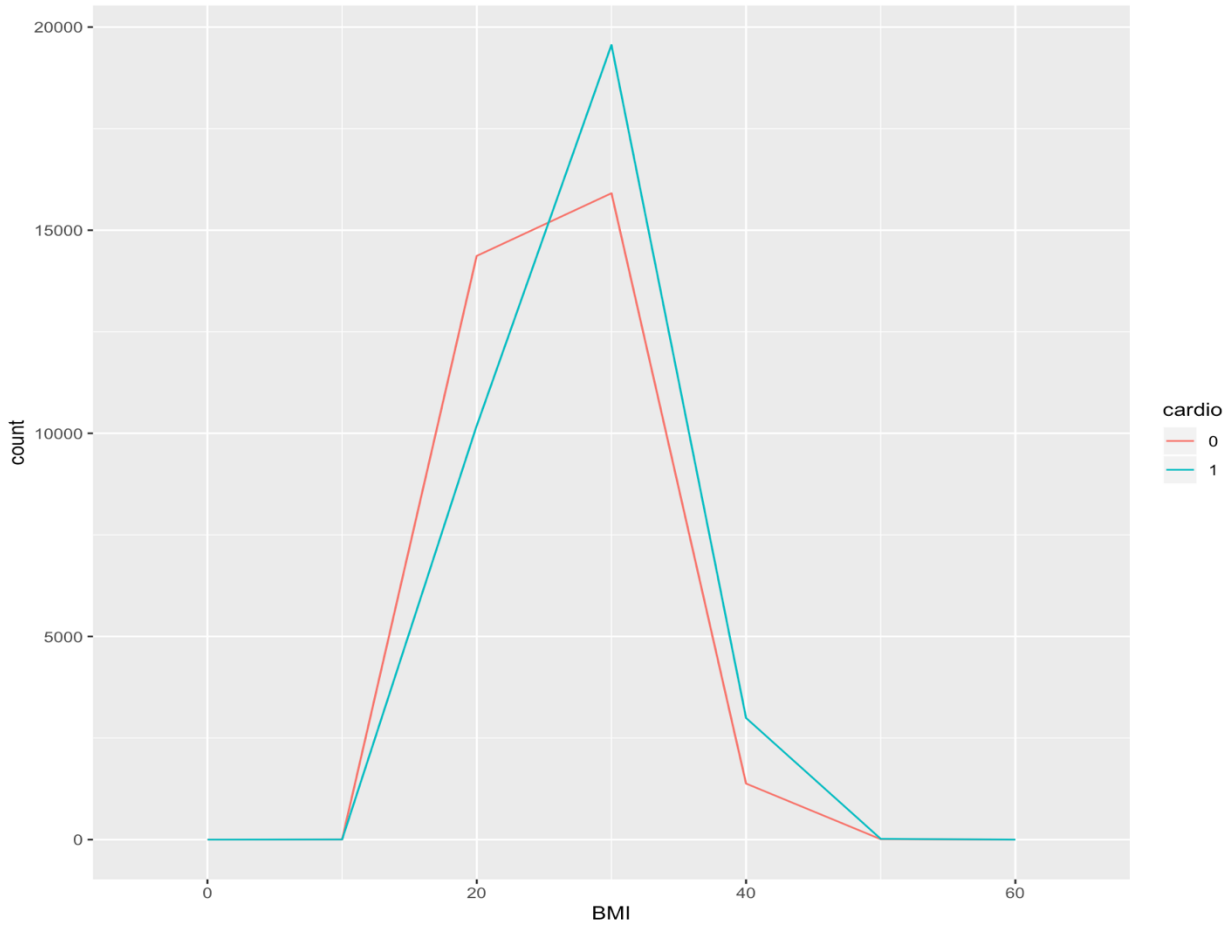


There is a split in the data: for those younger than roughly 18,000 days (50 years), there are more people without CVD than with it, but this flips for those who are older than 18,000 days. For those around 19,000 days (55 years), the proportion of those with and without CVD is pretty similar, and for everyone older, more people have CVD than those who do not.

I could have used height and weight by themselves, but these two variables give me a perfect excuse to perform some feature engineering. Height and weight can be used to calculate BMI. The formula for BMI is weight (in kg) / [height (in m)]^2 . Height is currently in centimeters, so I converted it to meters before calculating BMI. Here is how BMI looks by itself:

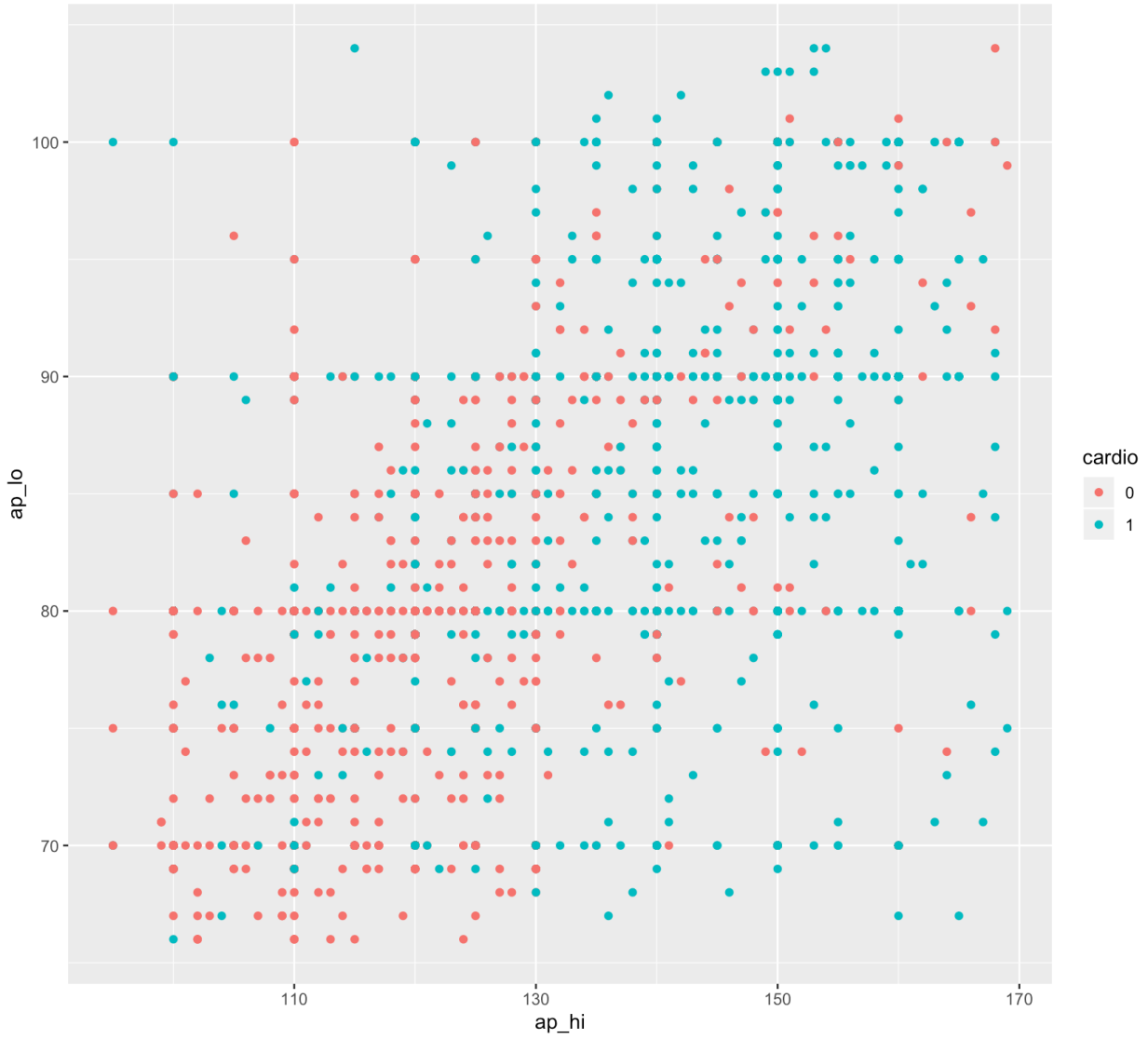


And in relation with cardio:



Its distribution has a right skew to it, and there is a clear split in the data centered around a BMI of around 25.

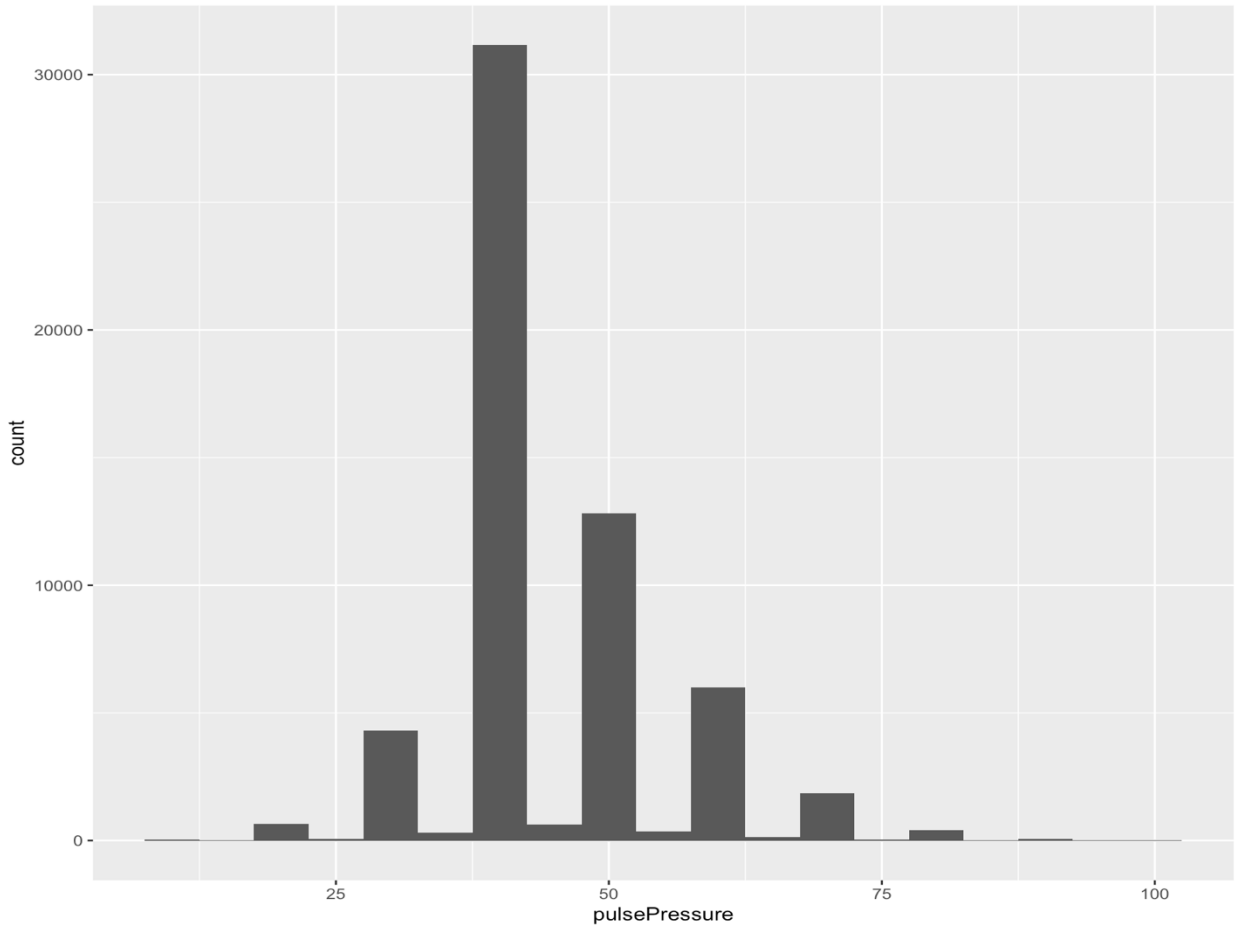
Next, let's look at the distribution of systolic and diastolic blood pressures:

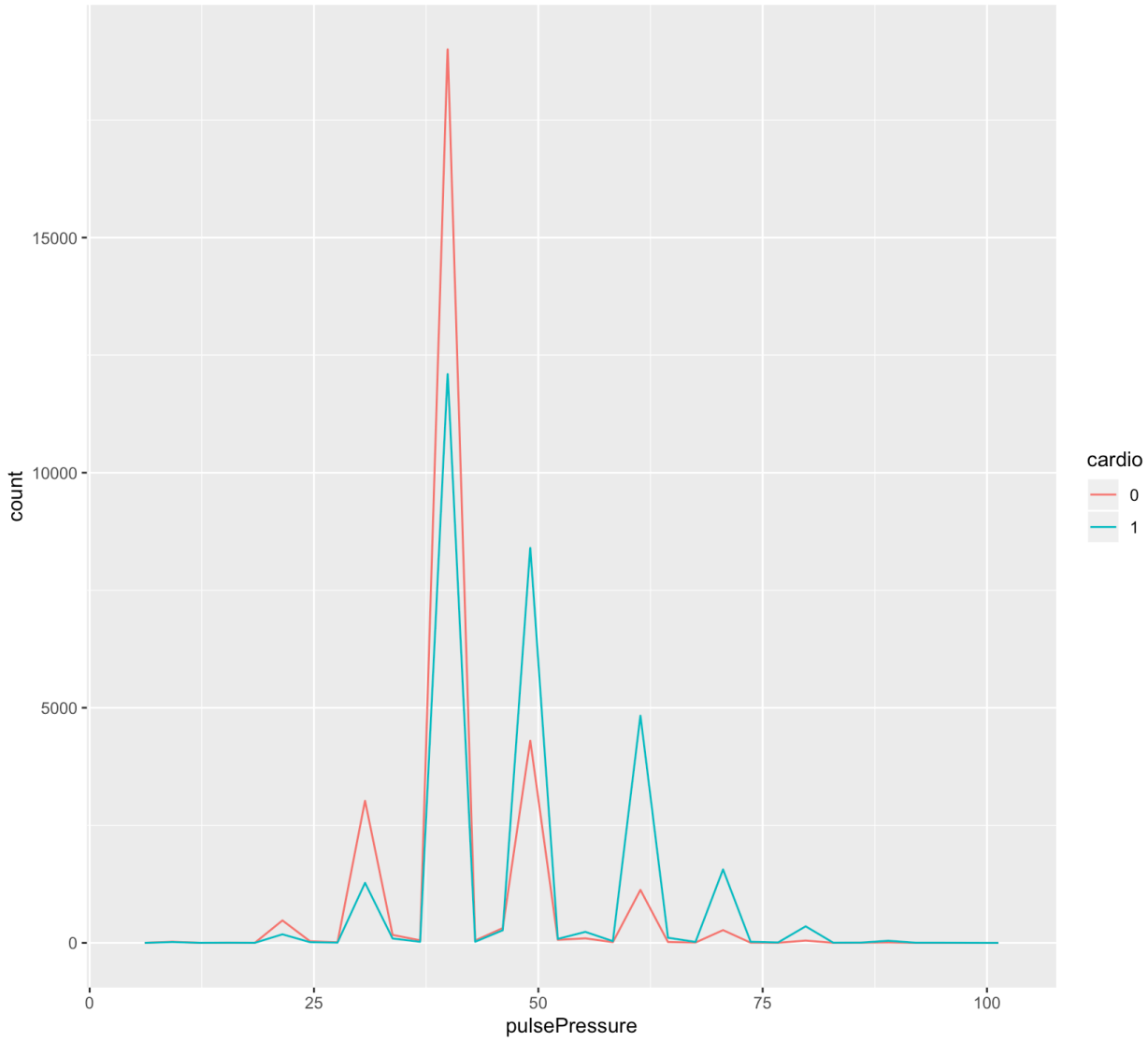


It looks like most of the patients on the right side of the graph, with high systolic and diastolic blood pressures, have CVD, while most of those with low systolic and diastolic blood pressures do not have CVD. This makes me think these two variables can be combined into one variable.

I created two variables out of the combination of systolic and diastolic blood pressure: pulse pressure, which is systolic blood pressure minus diastolic blood pressure, and systolic-to-diastolic ratio (S-to-D), which is systolic blood pressure divided by diastolic blood pressure. There

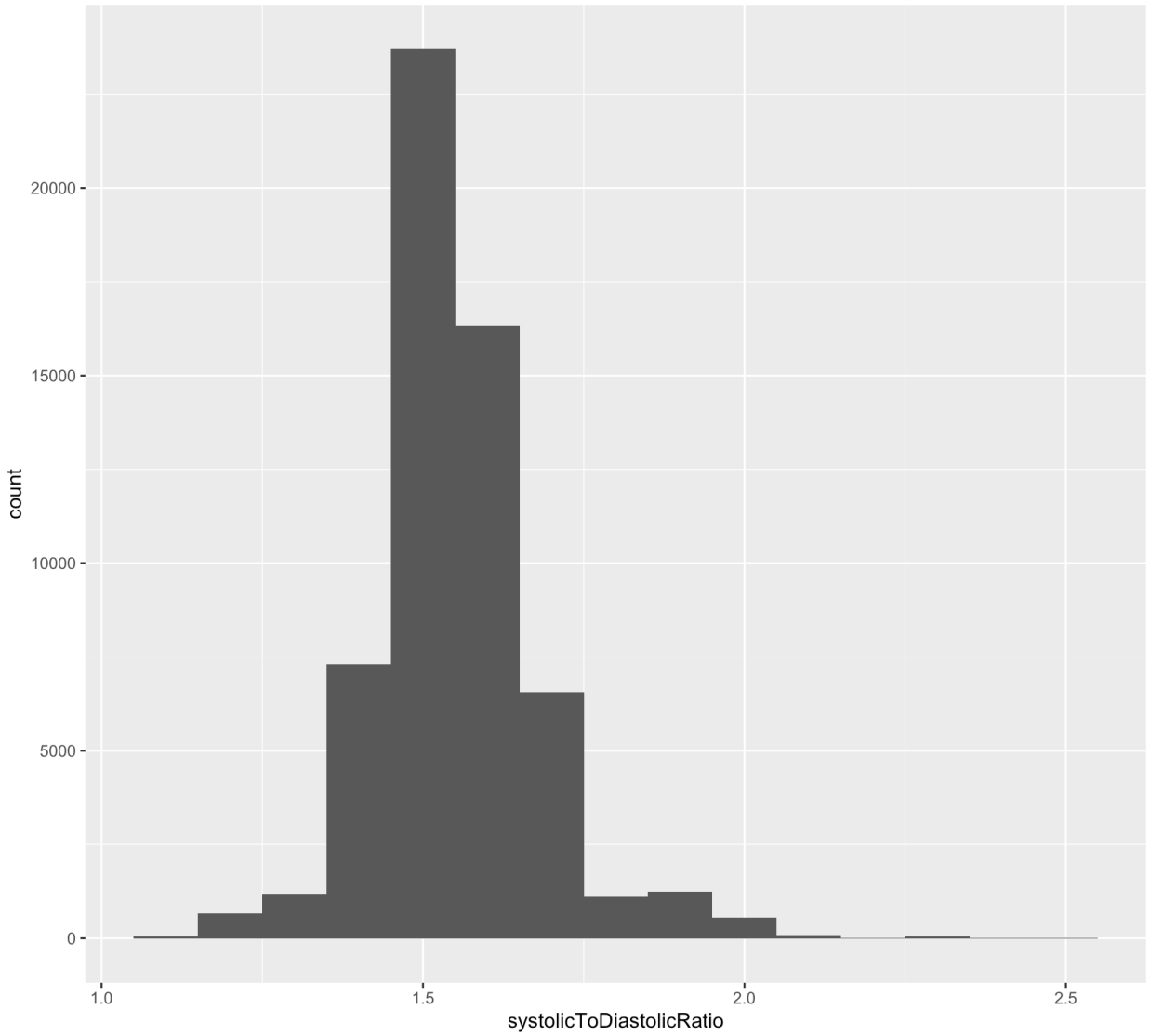
was a negative pulse pressure, meaning that there was a diastolic blood pressure greater than a systolic blood pressure, which is impossible. I kept only pulse pressures greater than 0. I checked the distribution of pulse pressure and its relationship to cardio:



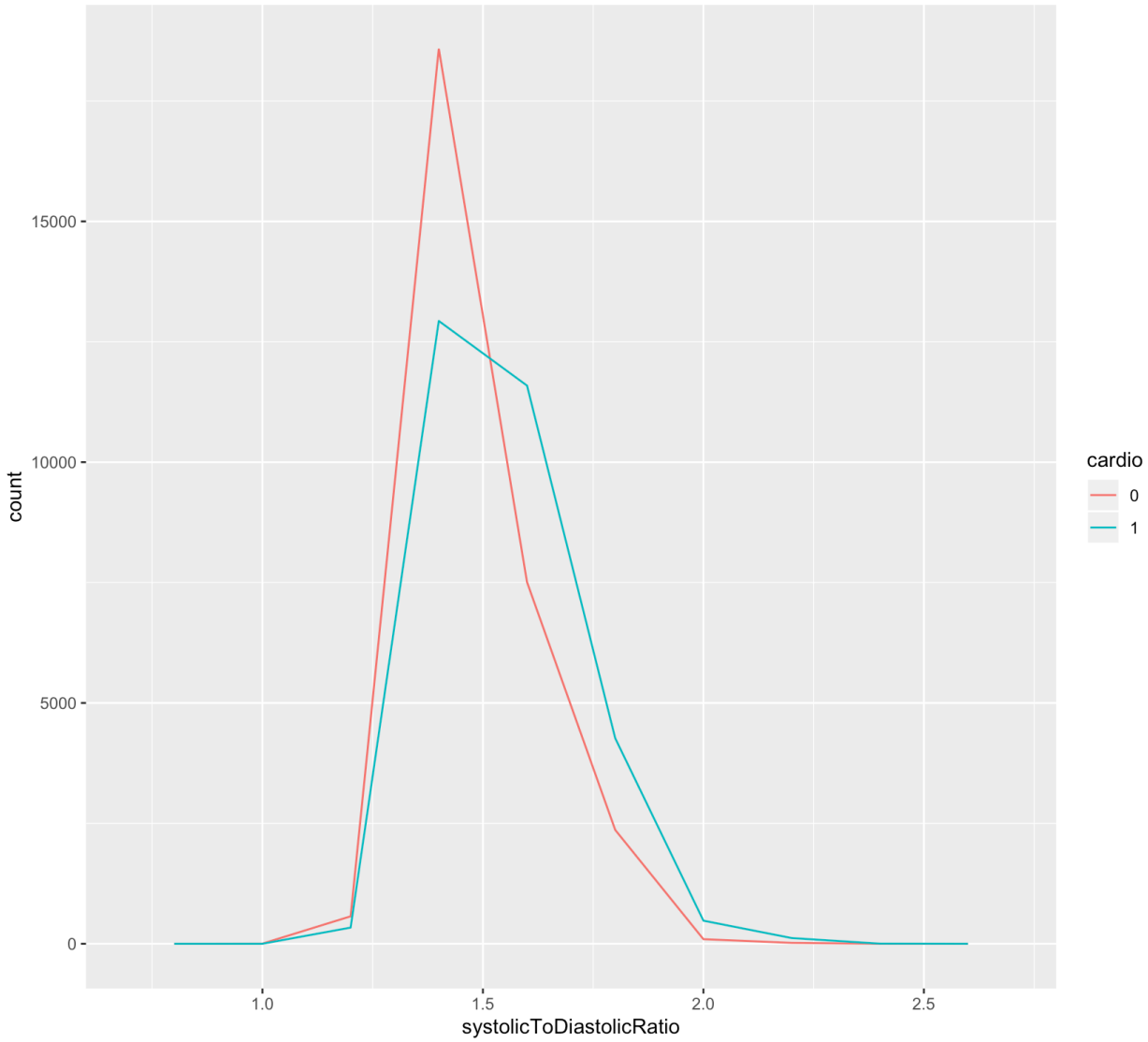


Its distribution is pretty similar to ap_lo and ap_hi in the sense that there are peaks around multiples of ten and nothing else. In this case though, the peaks range from 20 to 80 instead of the higher numbers of ap_lo and ap_hi. The pulse pressure graph has a pretty clear split: those with <=40 pulse pressure are more likely not to have CVD, while those with >= 50 pulse pressure are more likely to have CVD.

Next, let's look at the S-to-D ratio's distribution.



Most S-to-D ratios seem to be around 1.5, with a slight (and I mean slight) rightward skew.



Again, there looks to be a split in the data – those with a S-to-D ratio of less than around 1.55/1.6 are more likely not to have CVD than those with a S-to-D ratio of greater than 1.55/1.6.

So, in conclusion, the features that I selected are BMI, pulse pressure, systolic to diastolic ratio, age, cholesterol, glucose, and active.

Model Selection and Training

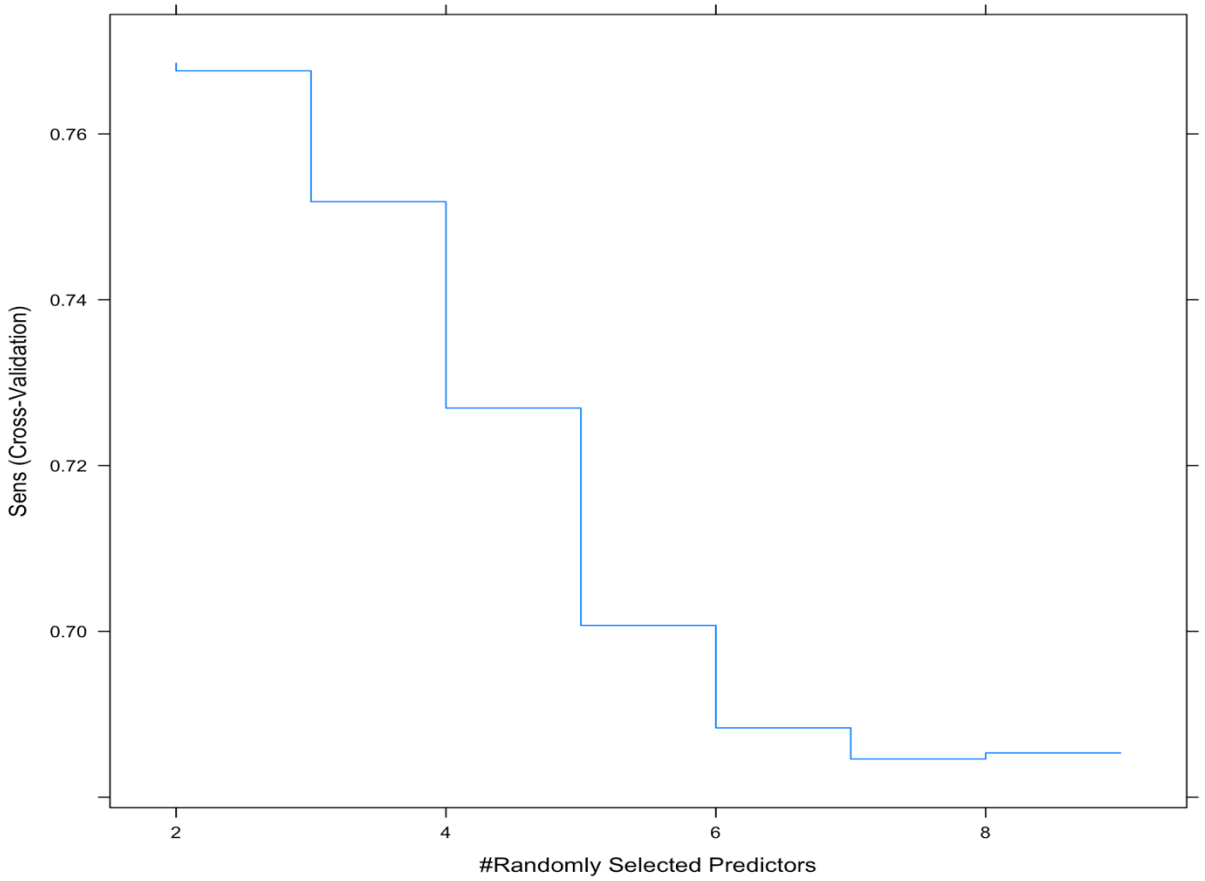
This is a supervised classification problem, so I used three models: logistic regression, random forest, and k nearest neighbors (KNN). Before applying any models, I split up the data into a test set and a training set (80% and 20% respectively). I used a 10-fold cross validation (CV) for the logistic regression and the knn-neighbors, but only a 5-fold CV for the random forest. The reason is the training dataset contained 44,146 observations, and running a 10-fold CV for the random forest proved to be too taxing on my personal computer. I also created a hard voting classifier, which aggregates the predictions of each model and predicts the CVD option that had the most votes. Note that I hyperparameter tuned the number of features used for the random forest model (mtry) and the number of neighbors (k) for the KNN model.

Before running my logistic regression model, I wanted to see whether active-ness is a statistically significant variable or not. Turns out it is statistically significant; it has a p-value of less than 2.2e -16. Furthermore, this is the same p-value as each of the other variables included, so active-ness is as important as each other input variable.

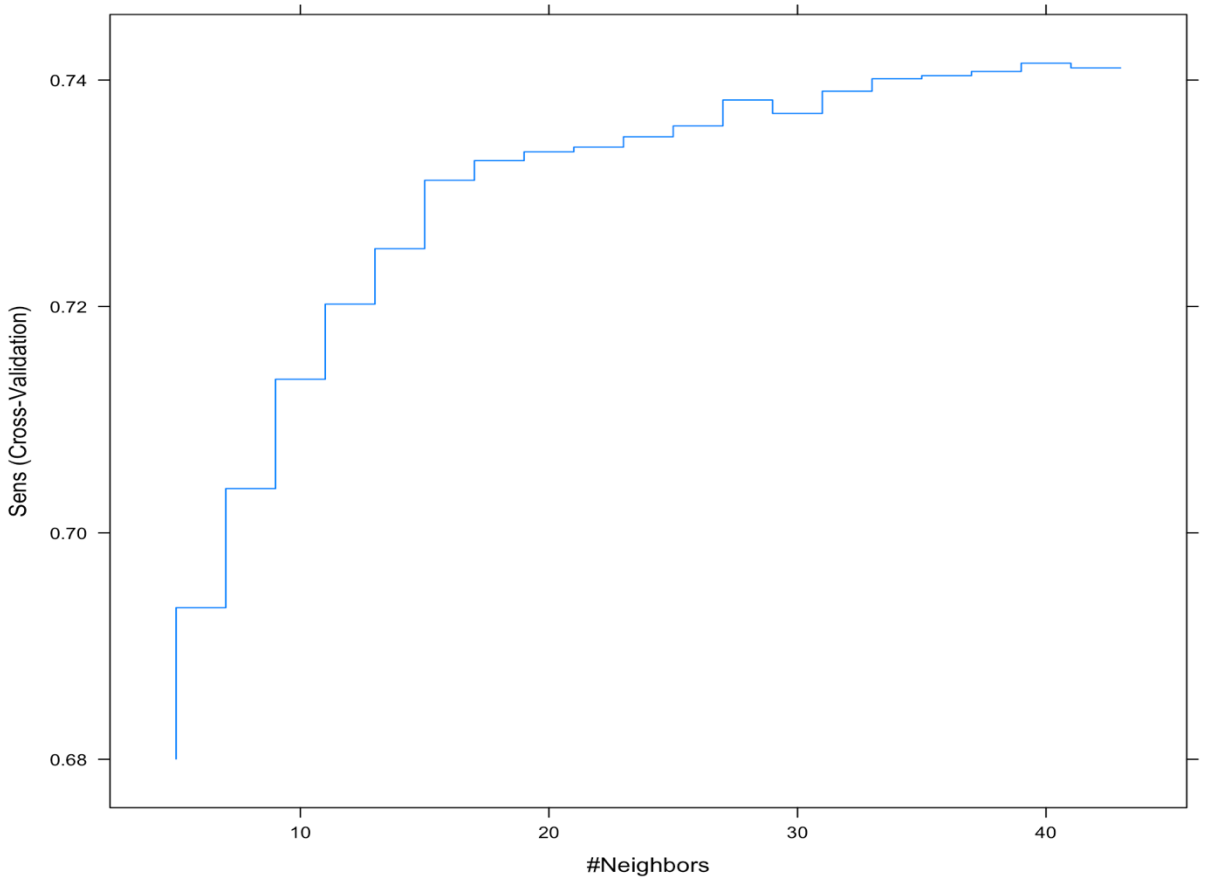
There are a ton of evaluation metrics for classification problems, but not every metric works for every problem. This problem is determining whether a patient has a disease or not, so the most critical thing to minimize false negatives. In other words, the worst thing for us to do is to predict that a patient does not have a disease, when in reality, they have it. That would be a disaster. The evaluation metric that takes false negatives into account is recall, or sensitivity. The formula for recall is true positives / (true positives + false negatives), or, in our case, patients with CVD correctly identified / (patients with CVD correctly identified + patients who are predicted to not have CVD when they actually have it). A high recall value (which goes from 0 to 1) has a small false negative value, which is what we want.

	Recall/Sensitivity for training set	Recall/Sensitivity for test set		
Logistic Regression (10-CV)	0.7771824	0.7780		
Random Forest (5-CV) – the best model had an mtry = 2	0.7687762	0.7666		
KNN (10-CV) – the best model had a k = 41	0.7414884	0.7438		
Hard voting classifier		.7649		

This is what the recall values were for each mtry value (it goes up to 8 because we only used 8 variables):



Same for the k value:



First and foremost, the model that produced the best recall value is the Logistic Regression model. There was very little overfitting, as the recall values for the training and test sets for each model were very close to each other. Interestingly, recall drops off significantly as the number of features used increases, a rather unexpected outcome. Another interesting tidbit is that the ensemble method did not produce the best recall value – I thought that ensemble methods always produce better results than each of its constituents, but both the Random Forest and Logistic Regression models produced a better recall value. Maybe ensemble methods work better with traditional evaluation metrics, like RMSE and accuracy, than with other evaluation metrics.

Conclusions and Sources of Improvement

Overall, I was pleased with the outcome, but there are things that could have been improved. Namely:

1) This dataset was highly flawed. I have no idea what normal, above normal and well above normal means for cholesterol/glucose levels. Moreover, I have no idea what 0/1 means for alcohol intake– does a 0 mean a patient doesn't drink often? Or that they don't drink at all? The same confusion applies to exercise – does 0 mean you don't exercise at all, or that you exercise a little bit, but don't consider yourself physically active? None of this was made clear, and a solution would have been to measure cholesterol/glucose levels in the blood (turning this into a continuous variable) and having clear alcohol and exercise bins (drink/exercise < 2 times a week, drink/exercise 2-4 times a week, drink/exercise 4+ times a week, etc.) to place patients in.

Overall, this project helped me learn a lot about classification problems and served as a great stepping stone for future growth in this field.