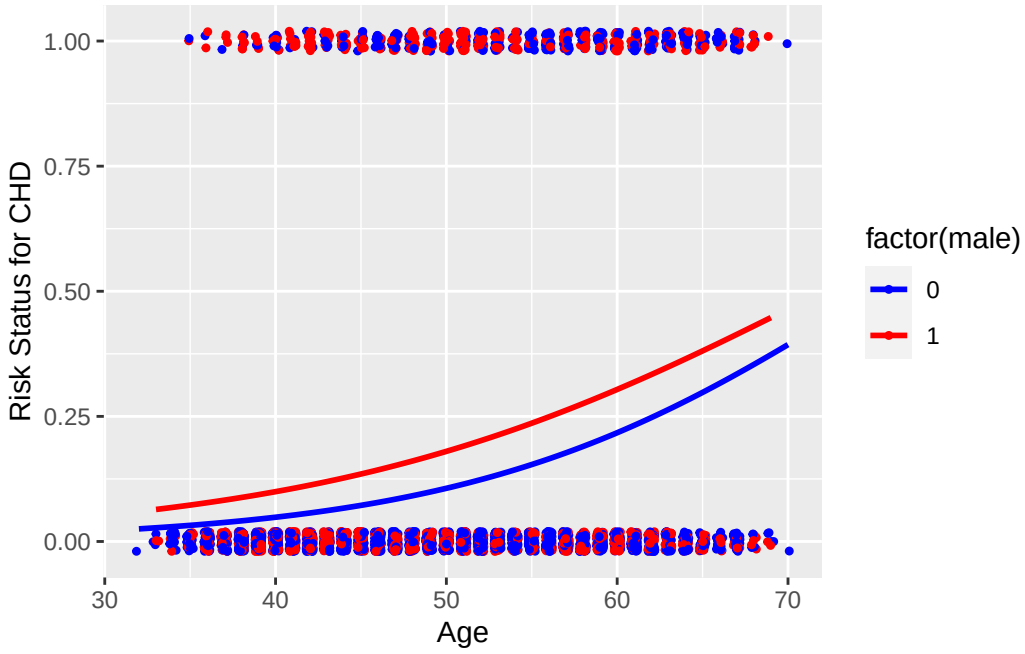
Framingham Heart Study

Josh Ye

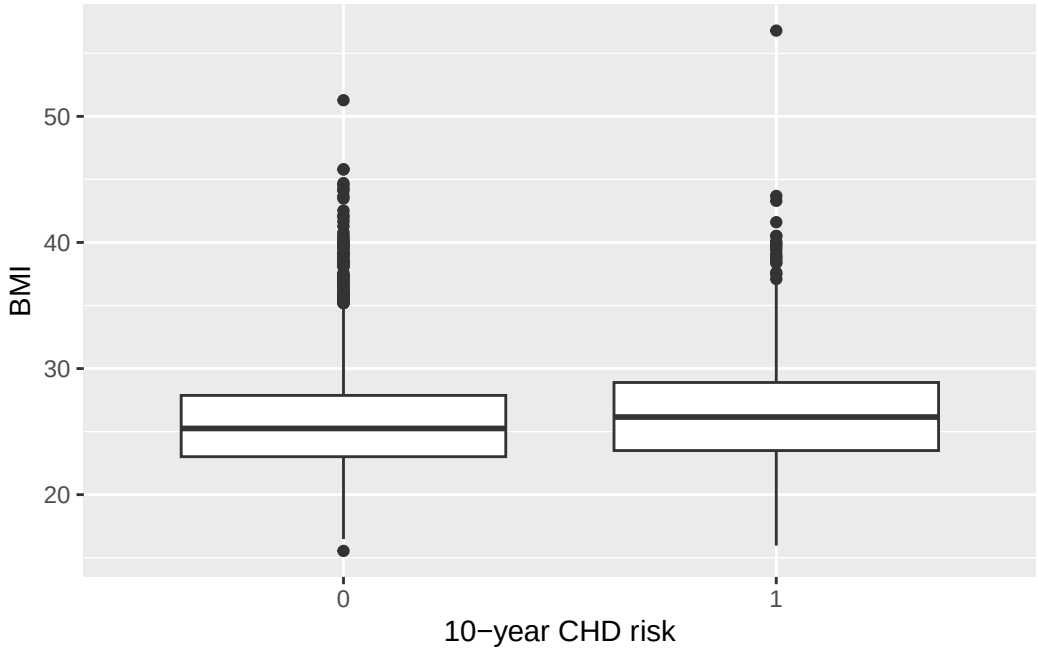
Read data and install libraries

Exploratory Data Analysis

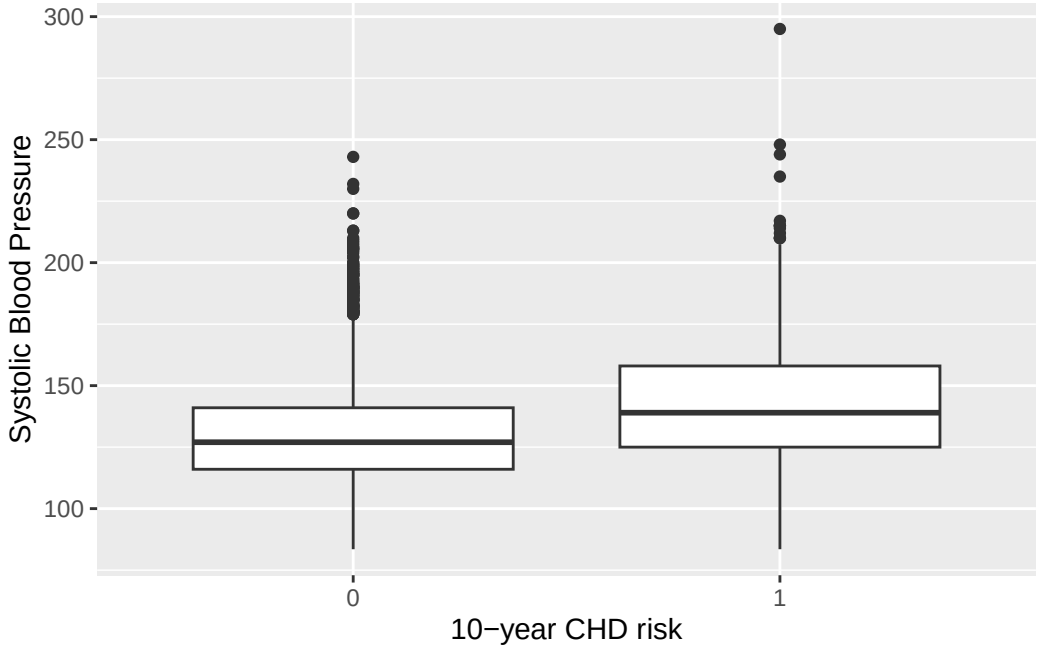
First, we take a look at the ages of individuals at risk for CHD and those not at risk. We notice that those who are at risk are older than those not at risk.



Additionally, we might assume that those who have a high BMI are generally more at risk for CHD. However, we see that there may not be such a strong statistical relationship. While the median BMI for the "at risk" category is marginally higher than the "no risk" category, there are a large large amount of outliers (on the high end) for the "no risk" group.



On the other hand, the opposite is true for Systolic Blood Pressure, where both the median and outliers (on the high end) are higher for the "at risk" group than the "no risk" group.

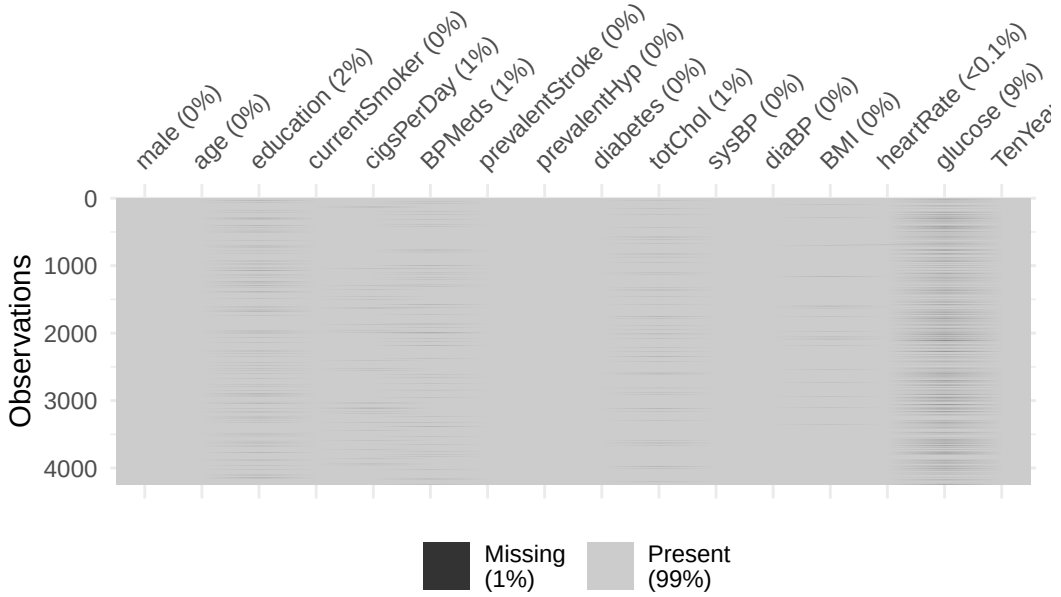


Research Question

Our main research question will be to determine which factors are most important for predicting the Ten-Year risk level for Coronary Heart Disease based on the data provided in this dataset. We will investigate model selection and variable selection techniques, then we will thoroughly investigate the model that was produced using our chosen techniques.

Missing Data

Firstly, we had to determine what to do with the missing data in the model.



Based on the graph above, we can see that the `glucose` predictor had the highest missing data rate. While this could possibly imply that there is some issue specific to measuring glucose levels which causes the data to be missing, this is hard to determine conclusively. Therefore, although it is possible that the data is MNAR, we will *not* impute data and simply continue our analysis by removing the missing data.

Variable Selection and Testing

Here, it is important to consider what potential interactions could occur before we proceed any further. [DO SOME EXPLORATORY DATA ON THIS]

Since our dependent variable, `TenYearCHD` is binary, this is a binary classification problem and a *logistic regression* model is best suited for this type of analysis. Furthermore, we have a few highly correlated variables.

```
# A tibble: 6 x 3
# Groups:   value [6]
  var1       var2          value
  <chr>      <chr>         <dbl>
1 diaBP      sysBP         0.787
2 cigsPerDay currentSmoker 0.774
3 sysBP      prevalentHyp  0.698
4 diaBP      prevalentHyp  0.618
5 glucose    diabetes      0.615
6 sysBP      age           0.389
```

As we can see, `diaBP` and `sysBP` , as well as `cigsPerday` and `currentSmoker` are the two most highly correlated pairs of predictors. The method that we utilized for variable selection was stepwise backward AIC. There were two reasons why this method was chosen. Firstly, the LASSO variable selection process, with a lambda produced by k-fold cross validation, did not produce meaningfully better results, especially because the cross-validation process produced an extremely low λ value. Secondly, by inspecting the results of backward AIC satisfactorily removing predictors that had little statistical significance in the model.

The results of our backward AIC selection process is the following model:

```
Call:  glm(formula = TenYearCHD ~ male + age + cigsPerDay + prevalentStroke +
    prevalentHyp + totChol + sysBP + glucose, family = "binomial",
    data = data)

Coefficients:
    (Intercept)             male              age       cigsPerDay
      -8.739521         0.553152         0.065337         0.019574
prevalentStroke     prevalentHyp          totChol            sysBP
       0.751412         0.226231         0.002248         0.014219
        glucose
       0.007314

Degrees of Freedom: 3655 Total (i.e. Null);  3647 Residual
Null Deviance:      3121
Residual Deviance: 2757     AIC: 2775
```

As we can see, while this refined model contains only eight predictors, most of them are statistically significant, while most of the statistically *insignificant* predictors have been dropped.

Key Results and Findings