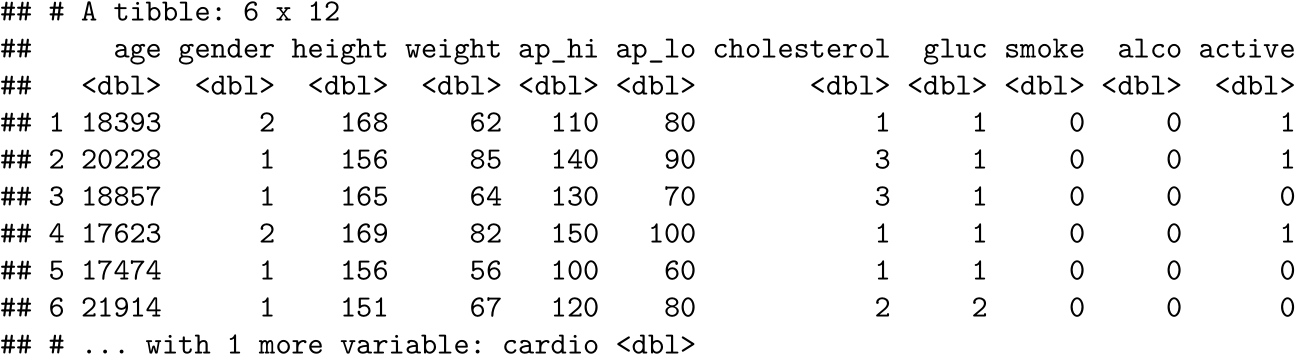
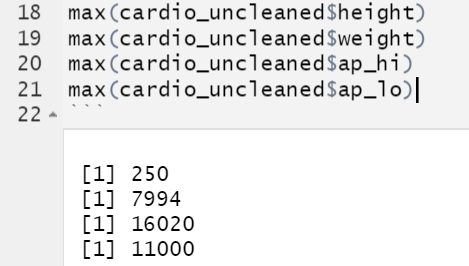
**Exploring models for Cardiovascular Disease**

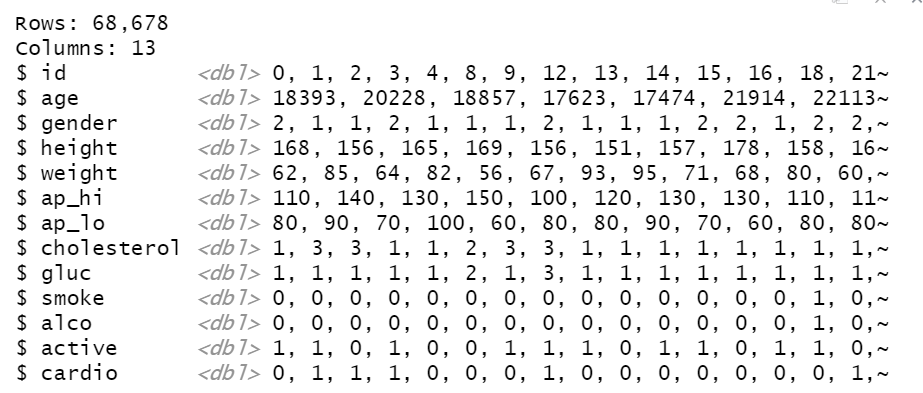
**Introduction.**

The dataset consists of 70,000 patients with data pertaining to their respective biometrics such as age (in days), height (in cm.), weight (in kg.), sex (0 for female, 1 for male), blood pressures (in mm of Hg), cholesterol levels (1 for normal, 2 for above normal, 3 for well above normal), glucose levels (same scale as cholesterol), as well as behavoral data such as smoking habits (1 if patient smokes, 0 otherwise), alcohol intake (1 if patient drinks, 0 otherwise), physical activity (1 if patient is active, 0 otherwise), and the presence of cardiovascular disease (1 if patient has Cardiovascular Disease, 0 otherwise). A sample of the data is listed in the following tibble. 

**Data Processing**

****

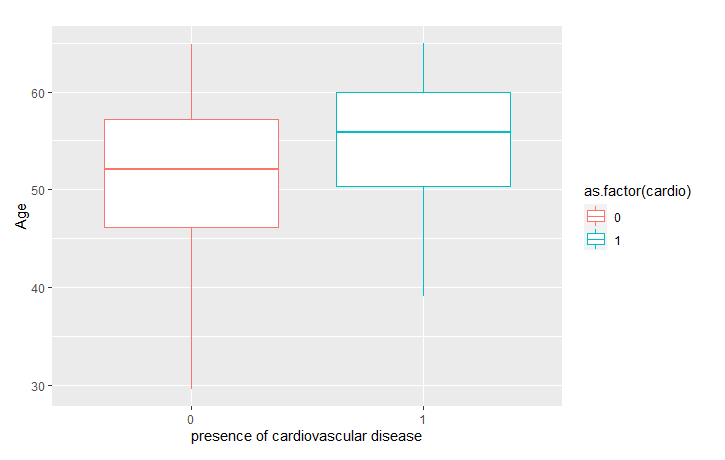
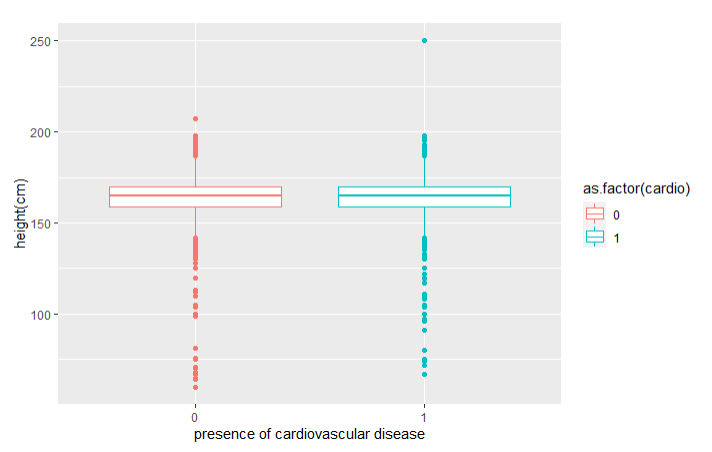
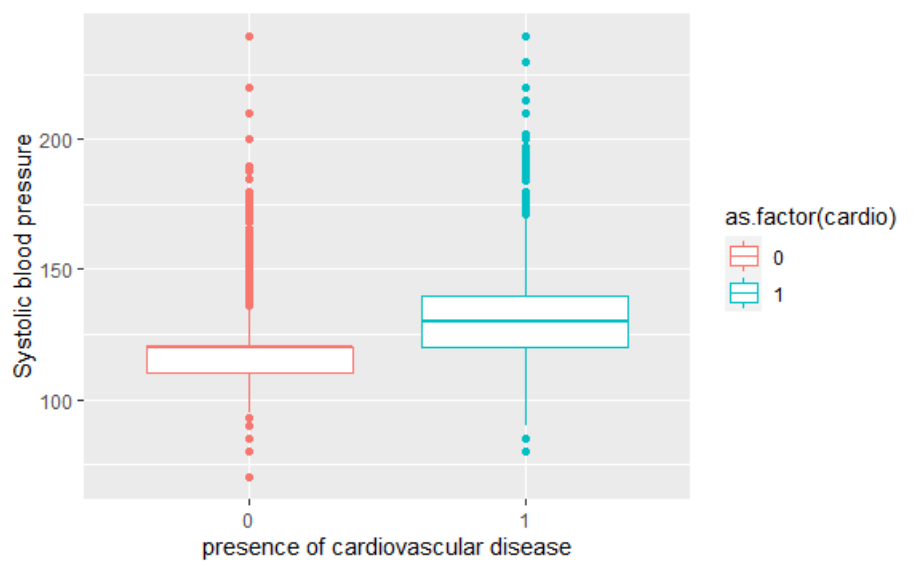
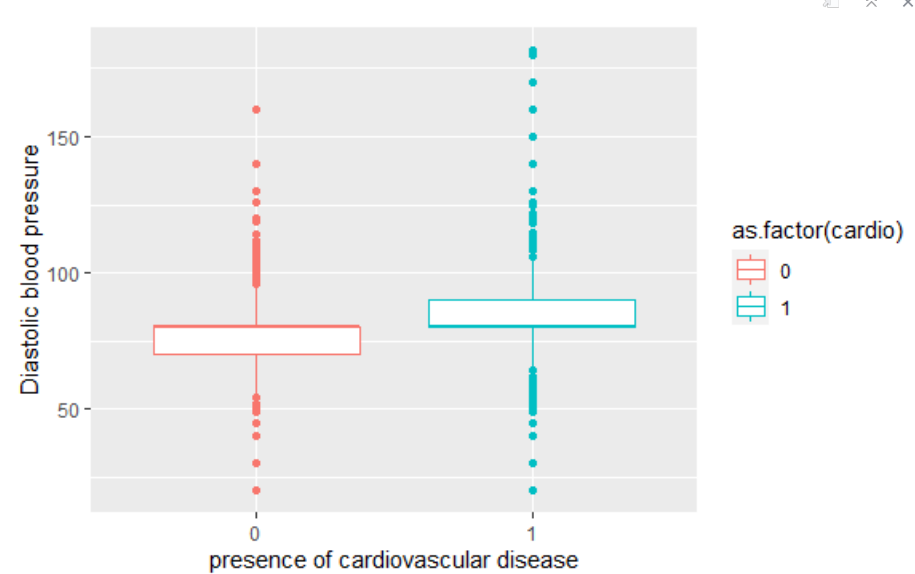
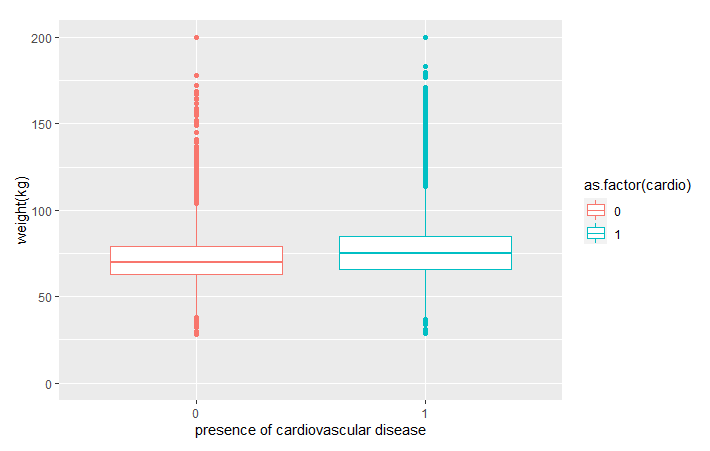
While researching datasets, we found a lot of unreasonable data. So we cleaned the data first: we removed unreasonable data (systolic blood pressure lower than diastolic blood pressure), and extreme data (extreme weight , height, diastolic and systolic blood pressure values).

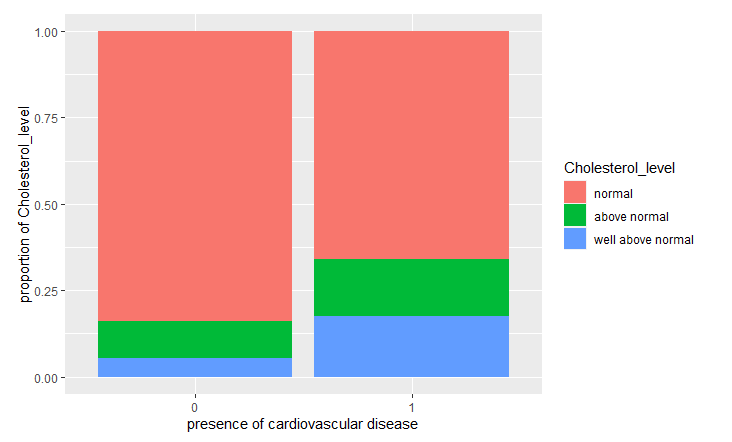
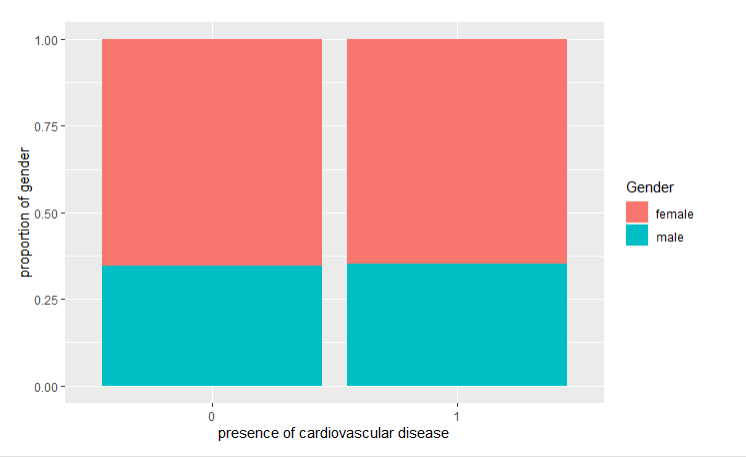


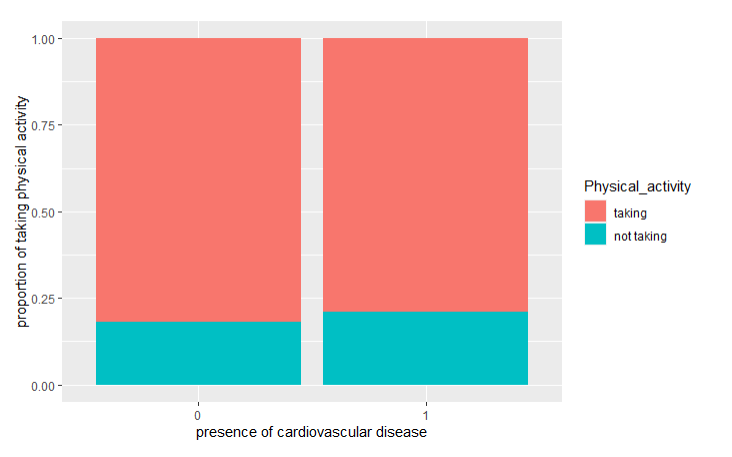
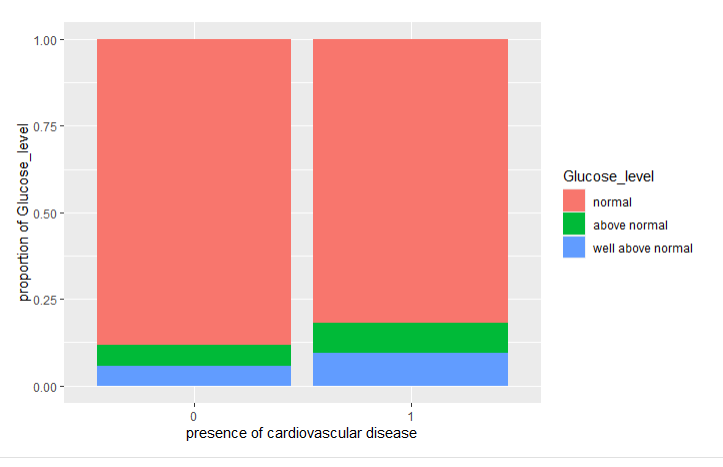
After we removed some of the data, we were left with 68678 data. This is still a large sample and supports us in drawing solid conclusions.

**Method.**

Before formulating the model, the independent variables were judged to find whether or not each independent variable had a relationship with the response variable, the presence of cardiovascular disease. To do this, box and whisker plots were created and used to compare the presence of cardiovascular disease with the value of each numerically distinct independent variable to show visually if there were any difference between the range and concentration of the independent variable when cardiovascular disease was present vs when the disease was not. The following graphs depict this process.



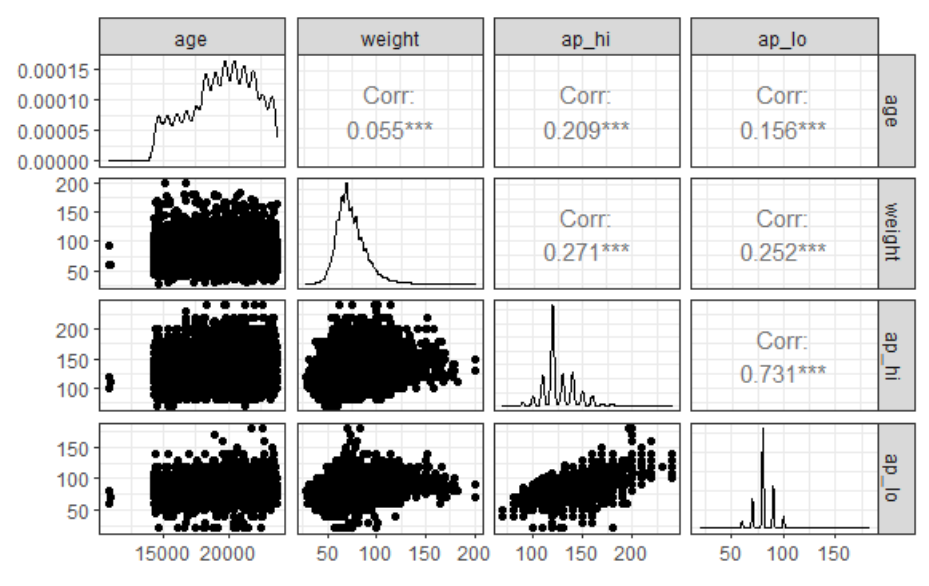
The findings from these graphs are that the variables that seem significant are age, diastolic blood pressure, and systolic blood pressure for determining if a patient has cardiovascular disease, while height seems insignificant, and weight might play a role. Next, the same process was applied for the ordinal and binary data, but instead with stacked bar graphs to show the proportion of those at each level.



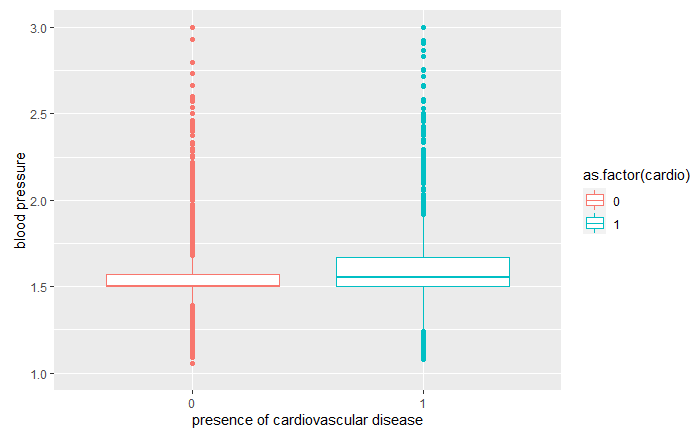
As shown, the sex of the patient and their alcohol intake or lack thereof seem to have no relationship with whether or not the patient has cardiovascular disease, while the cholesterol level of the patient and the patient’s glucose levels seem to have a relationship. Intuitively,

one might assume that the smoking habits of the patient would influence their risk of cardiovascular disease, yet its graph did not depict a strong relationship between them.

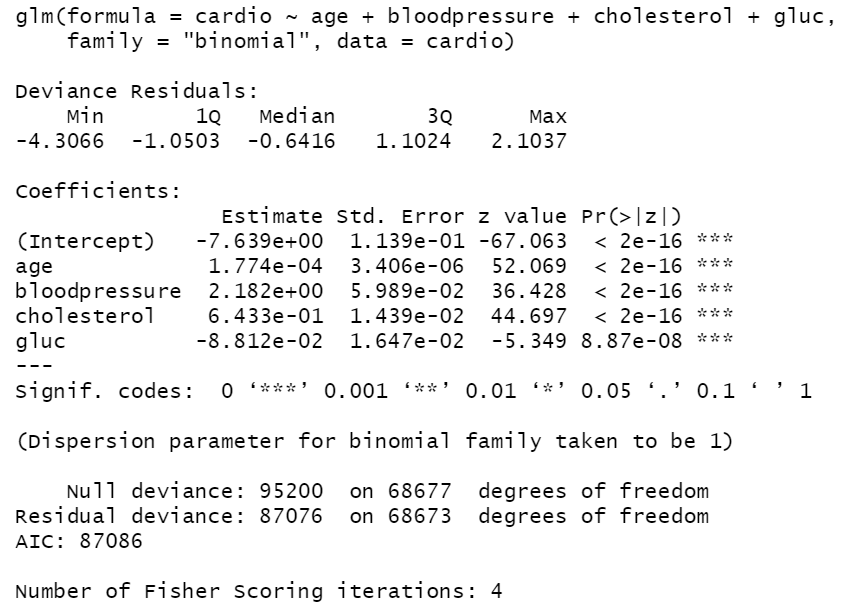
Before testing the model, we need to test the correlations between continuous independent variables. We would not include height since it will not influence the dependent variable.

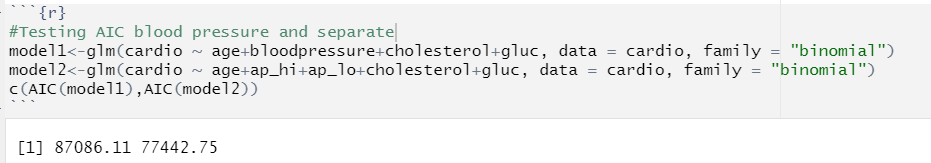


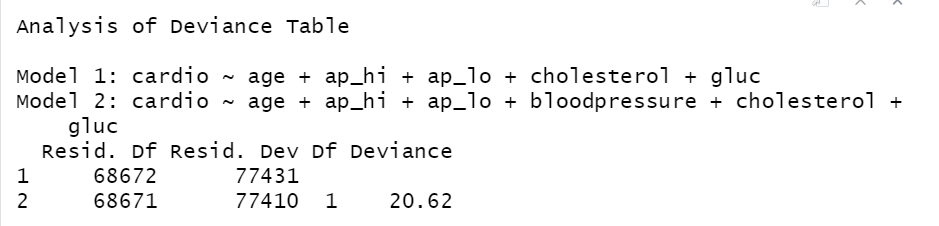
From the correlation scatterplot matrix, we find out that the systolic blood pressure and diastolic blood pressure are highly correlated. As we searched on the Internet, we learned that systolic blood pressure and diastolic blood pressure should be correlated and blood pressure is calculated by systolic blood pressure and diastolic blood pressure. So we use the formula to calculate the blood pressure and use blood pressure as another independent variable.



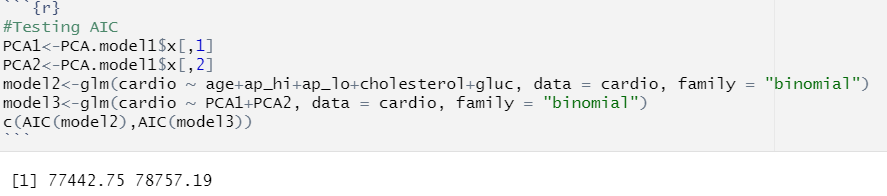
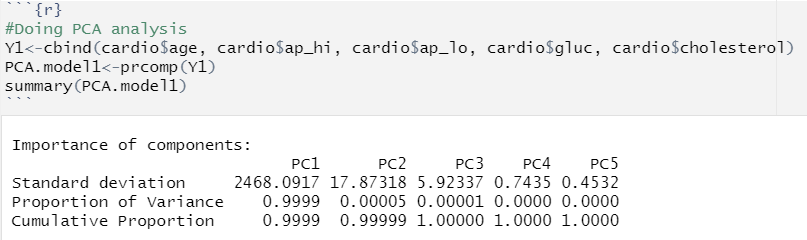
With knowledge of which variables are most significant, the model was formulated next. A binomial regression model was chosen to predict the odds of the patient possessing cardiovascular disease vs not possessing the disease.



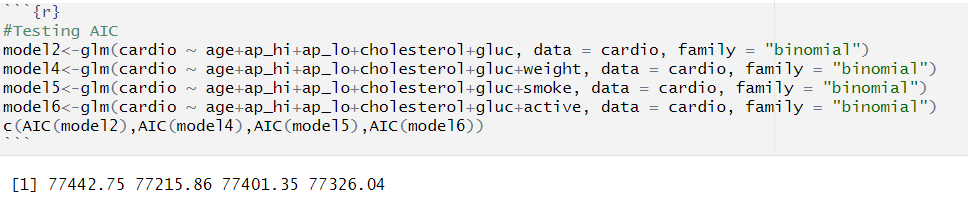
Because the blood pressure in the original dataset was divided among two variables: systolic and diastolic, a comparison was made between the AIC values of two different models, one where blood pressure is a single variable and one with the two divided variables for blood pressure, the decision was made to keep blood pressure as two variables instead.

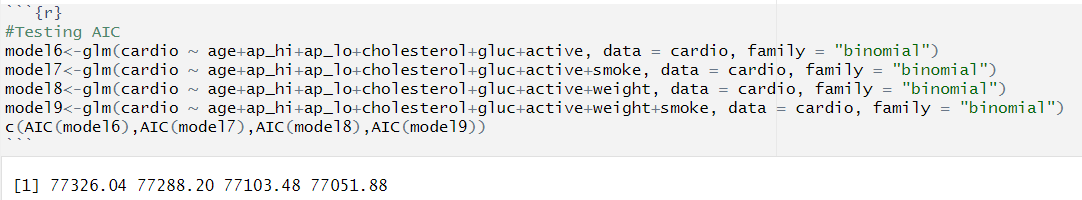
Afterwards, a deviance goodness of fit test was done between the chosen model representing blood pressure with two variables and a model where an interaction between the two blood pressure variables are used in addition to the two variables representing blood pressure. Due to the negligible change in residual deviance, the non-interaction model was chosen over the interaction model. 

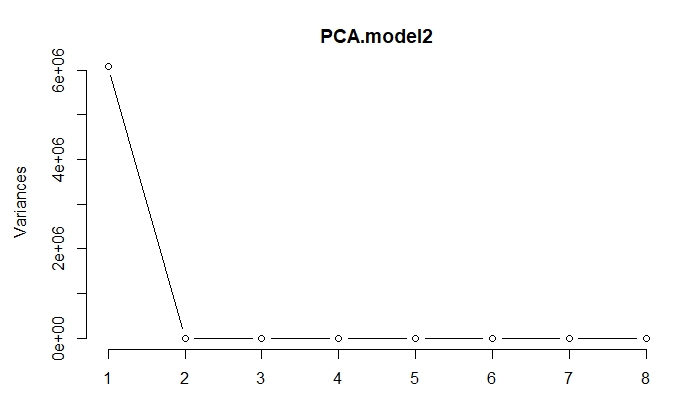
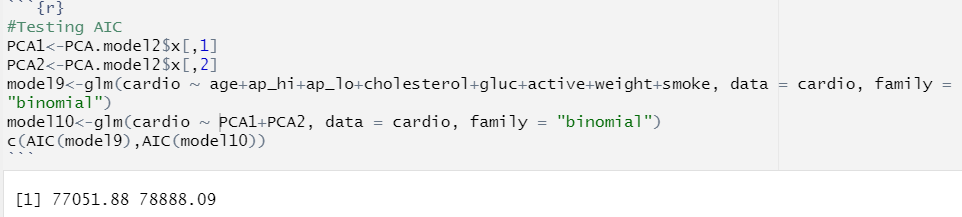
A principal component analysis was made to see if a reduction of the dimensionality of the model could reveal relationships within the model in order to reduce the dimensions if necessary. The first component, however, captured 99.99% of the variance, and an AIC comparison revealed that when using the first two principal components to predict the odds of a patient possessing cardiovascular disease, the AIC would increase past the regular model.



Next, an AIC test was administered to all possible combinations of the original model containing blood pressure, cholesterol levels, and glucose levels along with the newly added variables for weight, smoking, and physical activity to see if more of the variance of the response variable could be explained with these additional variables.





The model found to have the lowest AIC value was the full model with the variables pertaining to weight, smoking, and physical activity added. To see if principal components could be used for this full model, a principal component analysis was once again conducted.

The scree plot shows that once again the proportion of variance of the first principal component was significantly larger than that of all other principal components and in addition, the AIC of the full model without using principal components to determine the odds of a patient possessing cardiovascular disease was less than that of the model using the principal components.

**Conclusion.**

The variables of age, glucose level, cholesterol level, diastolic blood pressure, and systolic blood pressure seem to be helpful with the prediction of the odds of a patient possessing cardiovascular disease. Weight, physical activity, and smoking was shown to play a lessened role in predicting the odds of a patient possessing cardiovascular disease. While, lastly, gender, alcohol intake, and height of the patient seem to have no relationship with the odds of a patient possessing cardiovascular disease. As a reminder, since we only have binary data for alcohol intake, smoking, and physical activity, the conclusion is made on whether do or not do influence the odds of a patient possessing cardiovascular disease. It is possible that drinking too much alcohol will increase the odds of a patient possessing cardiovascular disease. But we have at least shown that moderate alcohol consumption is not harmful to getting cardiovascular disease.