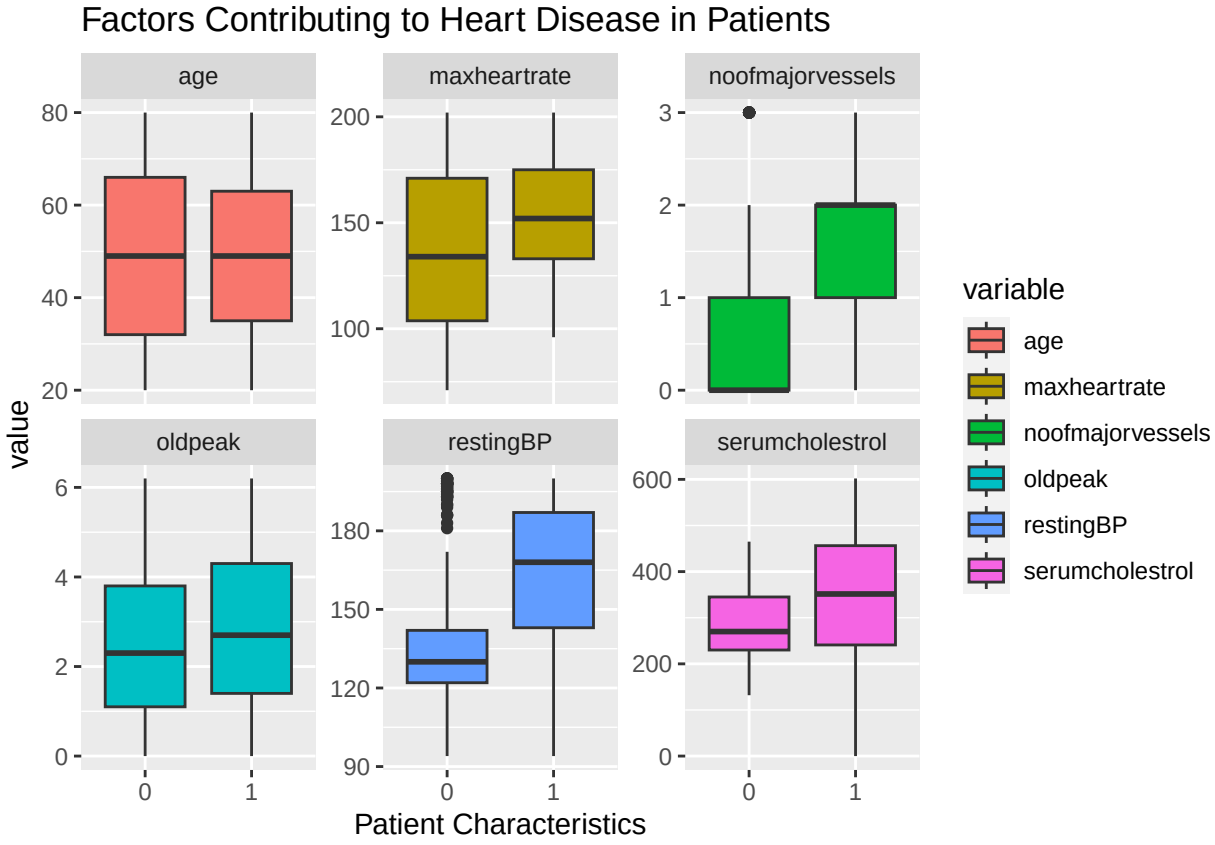
Not Project 1

Ethan May

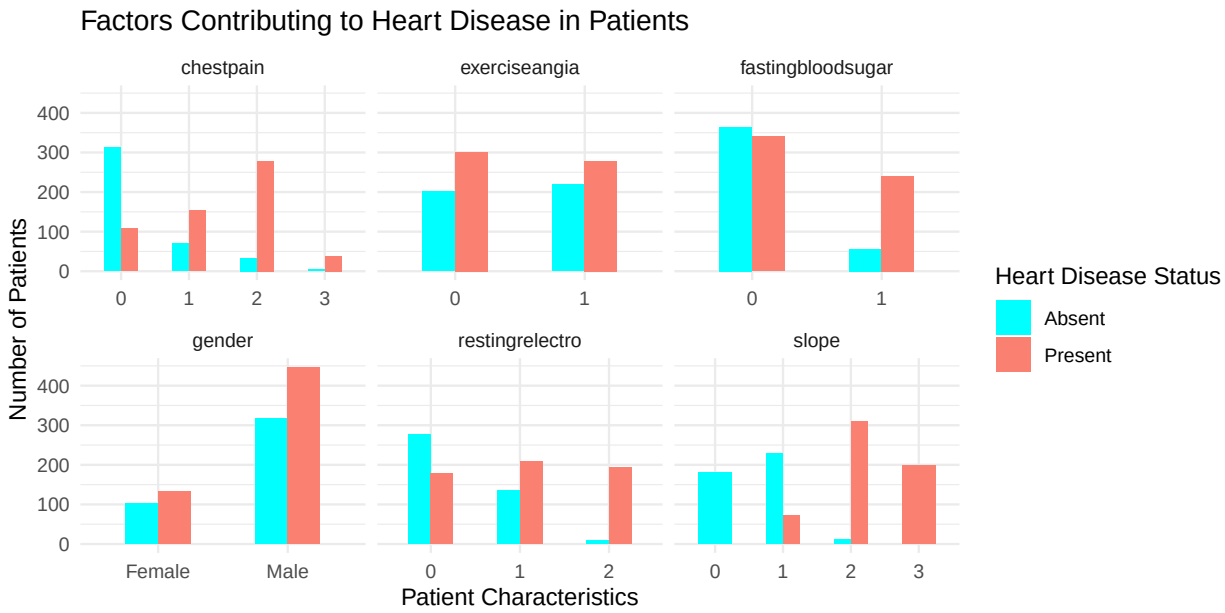
2024-02-14

This sample of data was chosen as it could help health professionals detect or even predict potential heart disease in their patients. From this project my goal was to learn how to find from both numerical and categorical variables what best can predict a target variable.

Cleaning this data was a straightforward process that involved omitting all NA values, ensuring R interpreted all the Categorical variables as factors rather than numeric values, and then, for the sake of readability making the gender variable more descriptive for the sake of readability. The most important variable in this dataset is "target" which shows whether or not the patient has heart disease or not. There are then 6 numerical variables and 6 categorical variables that are predictors to this target.



This Box plot includes all the numerical predictor variables put up against the target variable that shows whether or not the patient has heart disease. From this graph we can see which numerical variables have the most impact on the target variable. For instance we see that age and oldpeak are not very effective predictors of heart disease, while number of major blood vessles, and resting blood pressure are much better.



This bar graph includes all of the categorical predictor variables paired against the target variable. Here we can see the Characteristics that are most likely associated with heart disease. For example patients with a ST segment slope factor of 2 or 3 are very likely to have heart disease, and patients with a restingelectro value of 2 are also very likely. We also see from this visualization that Gender is not a very effective predictor, while slope and restingelectro are.

In a previous project i was able to use random forests and knn models to help predict the target variable. From this I was able to learn how to create predictive models that are effective and avoid over fitting. The data set however does suffer from data bias, as every observation is a patient currently at a hospital. There is no observations of healthy patients. This limits the usefulness of these predictive models as it would only help people who already need to be hospitalized. If a data set included data from health patients who only came in for a physical, the models may be more useful.