Data

Sara Shao

12/5/2021

Data

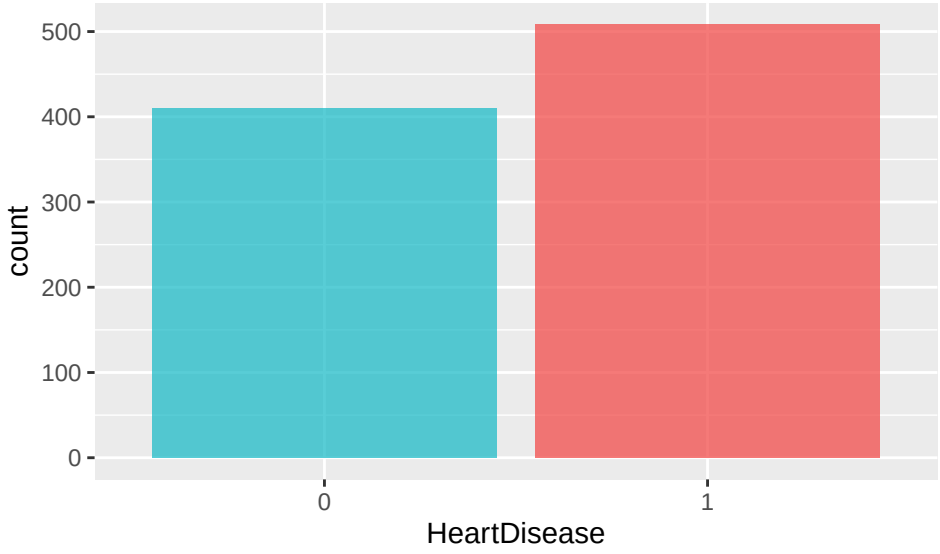
Description

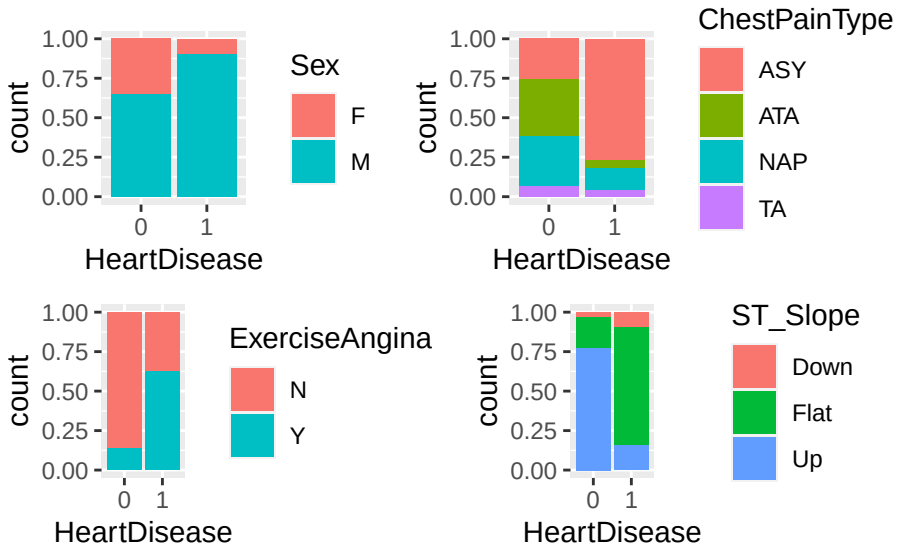
Link to data: https://www.kaggle.com/fedesoriano/heart-failure-prediction

We obtained our data from Kaggle. This data is the result of 5 heart disease datasets combined into one dataset. These five datasets are from Cleveland, Hungaria, Switzerland, Long Beach, and Stalog, with similar amounts of observations from each. It is apparently one of the largest heart disease datasets available. The dataset has 11 predictor variables, which are age, sex, chest pain types, resting BP, cholesterol, fasting blood sugar, resting ECG results, maximum heart rate, exercise-induced angina, oldpeak, and slope of the peak exercise ST segment. Sex, chest pain types, resting ECG, exercise-induced angina, and ST slope are categorical variables, while the rest are numeric. It has one predictor variable, which is whether or not the patient has heart disease. It has roughly the same amount of positive and negative heart disease diagnoses. Because in the dataset the incidence of heart failure is roughly equal to the incidence of non-heart failure, the data is likely not based on the general population, where we would expect the incidence of heart disease to be a lot lower. This means that probabilistic prediction models may be skewed in their results, but we can still make inferences about important predictors.

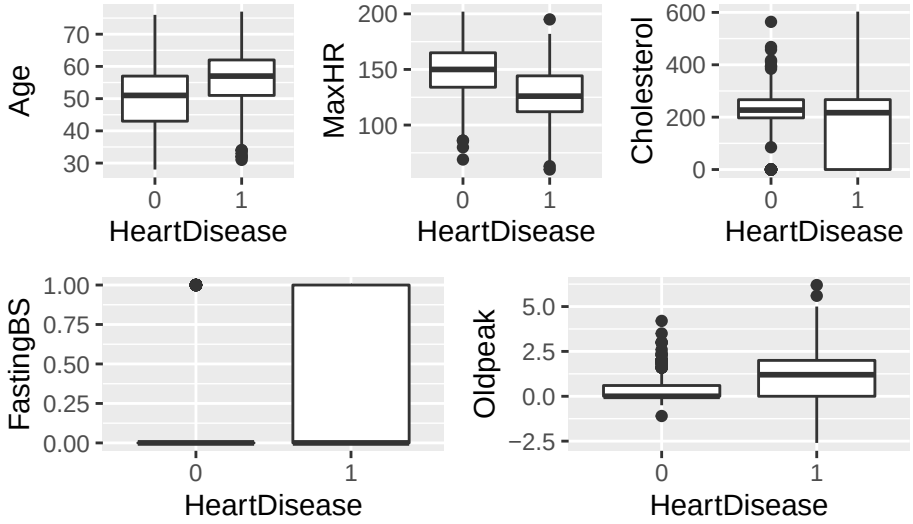
EDA

The plot below shows the spread of our response variable `HeartDisease`.





The above 4 categorical variables, sex, chest pain types, exercise-induced angina, and ST slope seem to have some relationship with heart disease incidence, so these variables will be good to look out for in our final model.



It seems that these above 5 numeric variables, age, max heart rate, cholesterol, fasting blood sugar, and oldpeak may have a relationship with heart heart disease and would be worth exploring in our models.