Heart disease prediction
By Jacob Hurley and Sofia Gray

Introduction:

According to the CDC, heart disease is the leading cause of death for people of most racial and ethnic groups in the United States. About 655,000 Americans die from heart disease every year and between the years of 2014 and 2015, it cost the United States about $219 billion each year. Heart attacks can happen with no warning and also be silent (meaning that the person is not aware of it happening). Additionally, about 805,000 Americans every year have heart attacks and about 605,000 of these are their first heart attack. Some known factors that can put people at risk for heart disease are high blood pressure, high blood cholesterol, overweight and obesity, an unhealthy diet, and physical inactivity.

Specifics:

One of the questions we are thinking of asking in our project is if we are able to predict whether someone has heart disease or not and possibly estimate when they are going to develop it based on several different variables that are easily observed and measurable. The data set that we will be using contains both qualitative and quantitative variables and the link for it can be found at the top of the page. Some of the qualitative variables include age, resting blood pressure, serum cholesterol, maximum heart rate, and oldpeak (ST). The qualitative or categorical explanatory variables we can explore include sex, chest pain type, fasting blood sugar, resting electrocardiogram results, excersice angina, and the peak exercise ST slope. The response variable is qualitative or categorical and will determine whether or not the person has heart disease. The data was obtained from kaggle, which combined 5 heart disease datasets(Cleveland, Hungarian, Switzerland, Long Beach, and Stalog) with 11 common features. The final dataset contains 918 observations.