# Heart Disease Indicators

Colin Busby

STAT 650
Fall 2022

## I. Introduction

The main research question of this data analysis is to discover which explanatory variables present in the data set used are good indicators for the risk of heart disease. According to the CDC, heart disease, which covers several types of heart conditions, is the leading cause of death in the United States. In 2020 alone approximately 697,000 Americans died from heart disease, while coronary artery disease, a specific form of heart disease, afflicts 20.1 million Americans. Ignoring the human cost, there are also financial costs: from 2017 to 2018, heart disease cost the US over $229 billion from medical costs and lost productivity. From this alone you can see the need for correctly predicting the risk of heart disease and the factors contributing to it.

## II. Data Description

The data set used for this project is Personal Key Indicators of Heart Disease, a data set containing information gathered from an annual CDC survey in 2020. The data was gathered as part of the Behavioral Risk Factor Surveillance System (BRFSS), using a telephone survey to gather US resident health data. The actual data set I used was compiled from the raw data and then cleaned by the Kaggle user Kamil Pytlak.

The dimensions of the data set are 319,795 rows (observations) by 18 columns (variables). The variables consist of the binary response variable, *HeartDisease*, and 17 explanatory variables: *BMI*, *Smoking*, *AlcoholDrinking*, *Stroke*, *PhysicalHealth*, *MentalHealth*, *DiffWalking*, *Sex*, *AgeCategory*, *Race*, *Diabetic*, *PhysicalActivity*, *GenHealth*, *SleepTime*, *Asthma*, *KidneyDisease*, and *SkinCancer*. For the purposes of this project, the variables of interest are *HeartDisease*, *Smoking*, *Stroke*, and *Sex*, all good factors in the risk of heart disease. As stated previously, *HeartDisease* is the binary categorical response variable, consisting of 'Yes' or 'No'. *Sex* is a binary categorical explanatory variable, consisting of 'Male' or 'Female'. *Stroke* is another binary categorical explanatory variable consisting of 'Yes' or 'No'. *Smoking* is a binary categorical explanatory variable consisting of 'Yes' or 'No'; it should be noted that in the context of this data set, for *Smoking* 'Yes' is defined as having smoked over 100 cigarettes, or 5 packs (20 cigarettes each), while 'No' is defined as smoking none or less than 100 cigarettes, over the course of a person's life.

## III. Methods and Results

The first visualization performed on the data was a histogram created using the response variable *HeartDisease* and the explanatory variables *BMI* and *Smoking*, as obesity and smoking are seen as good indicators of a person's risk for heart disease. As seen in Figure 1, while those who smoke account for approximately 30~40 percent of those who have no heart disease, a much larger proportion of those who have heart disease also smoke. Interestingly, *BMI* doesn't appear, just by eyeballing the graph, to have as much importance in indicating heart disease.  This would indicate that smoking is a much bigger factor in the risk of heart disease, just based on the available data.

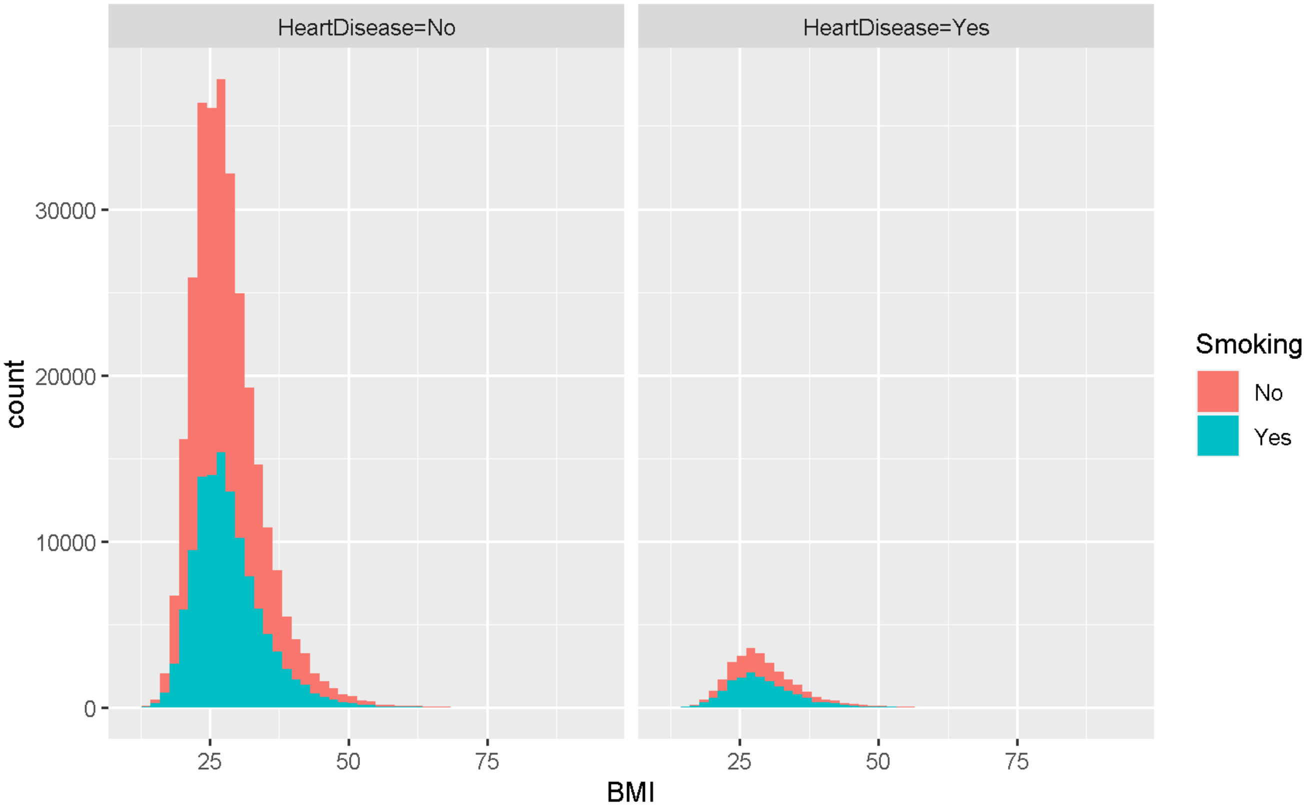

*Figure 1. Histogram of BMI according to heart disease status, showing proportion of smoking status by color.*

The second visualization was a bar chart using the response variable *HeartDisease* and the explanatory variables *Sex*, *Smoking* and *Stroke*.  *Sex*, *Smoking* and *Stroke* were chosen as I wanted to explore how strokes and smoking would affect the gender differences present in heart disease,as until the onset of old age, women suffer less from cardiovascular disease.

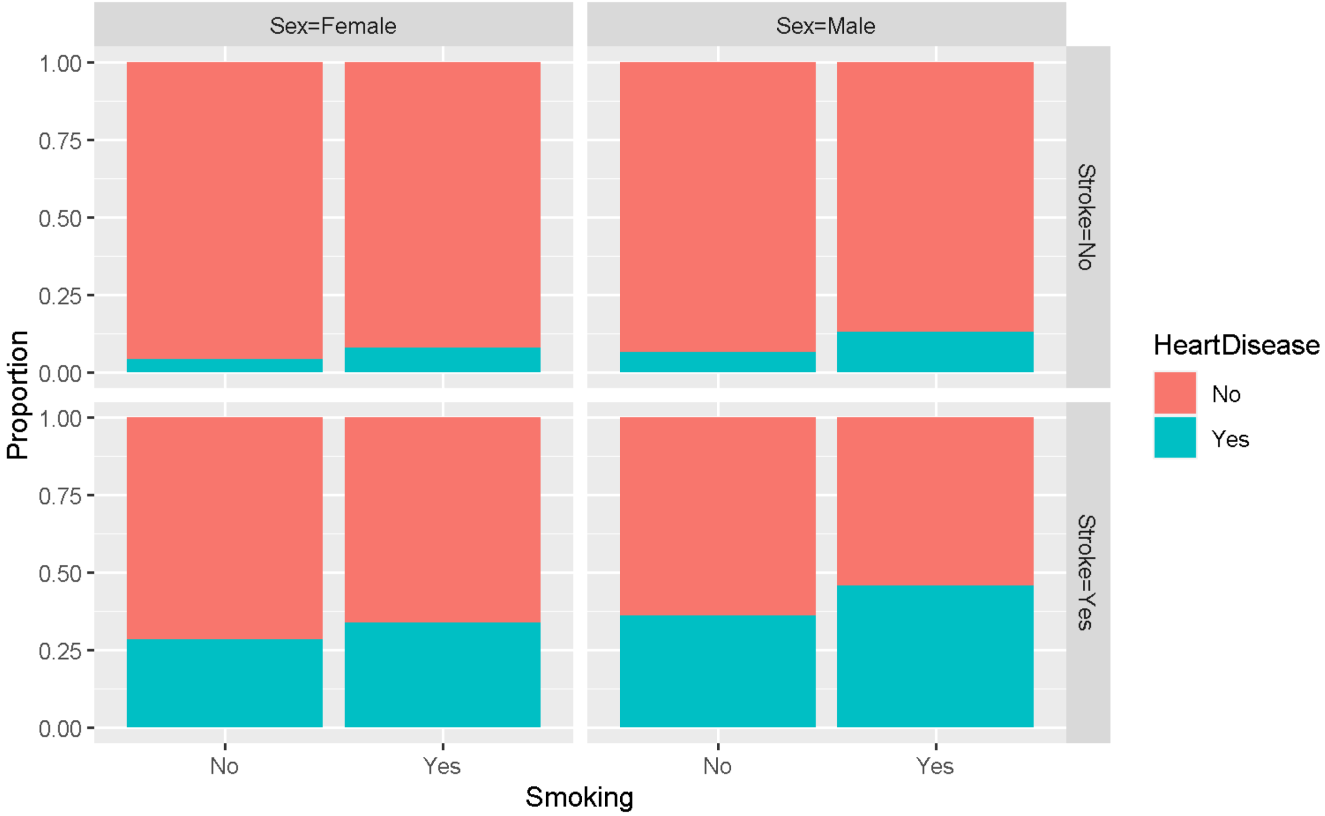

*Figure 2. Bar charts of the proportion of heart disease among participants according to sex and stroke status*

As seen in Figure 2, *Sex* plays a definite role in a person's chances of heart disease, regardless of their *Stroke* and *Smoking* status, with women suffering less from heart disease. Having had a stroke leads to a large increase in having heart disease, approximately 3 to 4 times as great. Finally, *Smoking* does lead to an increase in heart disease, though a much smaller one than that of a stroke.

## IV. Conclusion

Based upon the two figures, the average person at the greatest risk of heart disease would be a male, potentially overweight, who either smokes or has smoked over a 100 cigarettes in the past, and has had at least one stroke in their life. Interestingly, the chances for a woman with the same parameters to have heart disease are about slightly less than those of a man who doesn't smoke but has had a stroke before. Taken all together, the data shows that risk factors influenced by lifestyle choices, such as smoking and the risk factors of strokes (high blood pressure, diabetes, high cholesterol, etc.), are of greater importance than gender, as they can be controlled and changed. In conclusion, if you don't smoke and live a relatively healthy lifestyle, your chances of heart disease should be fairly minimal.

## Code Appendix

Github Repository: BusbyCl11/STAT-650-Advanced-R/STAT 650 Project.Rmd