# Heart Disease in Adults

Catalina Munoz

STA486C

## Introduction

Heart Disease is the top cause of death for adults and according to the CDC, 47% of Americans in the United States have at least one of the top three risk factors for heart disease - smoking, high blood pressure, or high cholesterol. In the health world, figuring out what causes this disease is very important. For this project, I will be analyzing a dataset from the CDC that highlights many different health factors and using them to see which variables could potentially influence heart disease. I chose this topic because I was interested in learning more about a condition that has affected members of my family first hand. I also wanted to do analysis on data that didn't primarily have numerical values.

## Data Description

This dataset comes from the CDC's annual health survey in the United States from 2020 to 2022 where they ask individuals 18 and older a series of questions regarding their health in order to monitor trends and develop programs with the goal of improving the well-being of adults in the United States. Coming from the CDC means this data is credible, however, it is also heavily unbalanced. About 91% of adults in the raw dataset did not have heart disease. Because of this, I knew that the results of my analysis would be biased, so I used undersampling to create a more balanced dataset. For the undersampling, I used the ROSE package to randomly exclude data from the "majority" class which were cases where individuals did not have heart disease.

There are a total of 18 variables, 16 categorical and 2 numerical that all may have a direct or indirect influence on heart disease. Nine of the factors have two levels. These are the survey questions that require a yes or no answer. The remaining categorical variables have 3 to 5 levels where survey respondents will answer questions on a scale. The last two variables, SleepTime and BMI are numerical values. The table below gives descriptive information about each variable.

| Variable | Description | Type |
|---|---|---|
| HeartDisease | Has the respondent had coronary heart disease or myocardial infarction? (Yes/No)? | Factor |
| BMI | Body Mass Index | Numeric |
| Smoking | Have they smoked 100 cigarettes in their lifetime? (Yes/No) | Factor |
| AlcoholDrinking | Are they a heavy drinker? Men: More than 14 drinks per week, Women: More than 7 (Yes/No) | Factor |
| Stroke | Have they had a stroke in their lifetime? (Yes/No) | Factor |
| PhysicalHealth | In the last 30 days, how would they classify their physical health? (Good, Fair, Bad) | Factor |
| MentalHealth | In the last 30 days, how would they classify their mental health? (Good, Fair, Bad) | Factor |
| DiffWalking | Is it difficult to walk or climb the stairs? (Yes/No) | Factor |
| Sex | Male or Female | Factor |
| AgeCategory | 13 level age category ranging from 18 to 80 years or older | Factor |
| Race | Race/Ethnicity | Factor |
| Diabetic | Have they had diabetes? (Yes, No, Borderline, During Pregnancy) | Factor |
| PhysicalActivity | Have they reported exercise in the past 30 days? (Yes/No) | Factor |
| GenHealth | In general, how good do they think their health is? (Poor, Fair, Good, Very Good, Excellent) | Factor |
| SleepTime | How many hours of sleep do they get per day? | Numeric |
| Asthma | Have they had asthma? (Yes/No) | Factor |
| KidneyDisease | Have they had kidney disease? (Yes/No) | Factor |
| SkinCancer | Have they had skin cancer? (Yes/No) | Factor |

18 row table was created for understanding each variable in this dataset.

## Data Evaluation

To see which factors affected heart disease, three visualizations were created. Figure one, shows 7 bar charts that represent the proportions of adults over 18 that may have heart disease. From this plot, graph A represents the percentage of heart disease in adults based on thirteen age categories. As the ages increase, the percentage of adults with heart disease do also. So, older people are more at risk for heart disease. Graph B shows the amount of males and females that may have heart disease. From the graph, it looks like in this data set, men tend to heart disease more. In graph C, there are four diabetes categories. For individuals that do not have diabetes, the chance of heart disease is less. People with borderline diabetes have a slight chance of having heart disease. There weren't many women who had diabetes during pregnancy, however by a slim margin, they tended to not have heart disease. Finally, over half of the individuals who do have diabetes also have heart disease.

Graph D shows that individuals that have a hard time walking or climbing stairs are prone to heart disease. This could tie back into age as heart disease is more prominent in older individuals, who may also have a harder time walking. Graph E indicates that many people are confident in their health. However, individuals who classify themselves with 'Good', 'Fair', or 'Poor' health are more likely to have heart disease. Further, a high proportion of people with 'Very Good' and 'Excellent' health do not have heart disease. Next, graph F shows that smoking increases a persons risk for heart disease. Lastly, graph G indicates that not many individuals in this data set have had a stroke. However, the ones that did have a high chance of having heart disease.

Figure 2 also shows heart disease in adults based on many different factors. Graph H showcases the proportion of individuals who have heart disease based on heavy alcohol consumption. This is classified by men who have at least 14 drinks per week and women who consume at least seven. From the bar chart, by a slim margin, alcohol drinkers tend to not have heart disease. Next, for people who have exercised in the past month, graph I shows that physical activity is good for your heart health as people who have heart disease tend to exercise less. For graph J, not many people in the data set have kidney disease, but the majority of individuals that do also have heart disease.

In graph K, about half of the individuals who had skin cancer also had heart disease. People without skin cancer tended to not have heart disease, however, in this data set, a bulk of survey participants have never had skin cancer. Next, in graph K, having asthma puts people more at risk for heart disease. Lastly, graph M represents adults and the quality of their mental health from the past month. In this data set, most adults classified themselves as having "Good" mental health. By a small margin, people in this category did not have heart disease. "Bad" mental health was the second largest category, and these individuals tended to have heart disease. For individuals with "Fair" mental health, the majority didn't have heart disease.

Figure 3 shows two boxplots that aim to represent the distribution of average sleep time and BMI of individuals with and without heart disease. The sleep time boxplots are very short and have similar placement. This shows that both groups have similar average sleep times at around seven hours. In the BMI boxplots, a log transformation was used to normalize the distribution. For individuals with heart disease, their BMI tends to be higher by a very small margin. Both plots have outliers which could be an indication of extreme values that may not make sense and affect the overall observation. For SleepTime, many values are very high and even reach 24 hours, which does not make sense and seems extreme. For BMI, the max was 94.66 which is very high and not possible as the average BMI of an adult is about 26.5 for men and women.

Assessing data visualizations is very important. Creating graphs to analyze is a way to find trends, insights, and patterns in order to convey information to your audience. Without visualizations, making sense of massive amounts of raw data would be very difficult. There are many ways to visualize all types of data - boxplots to see the summary of data, scatter plots to see the relationship between quantitative values,

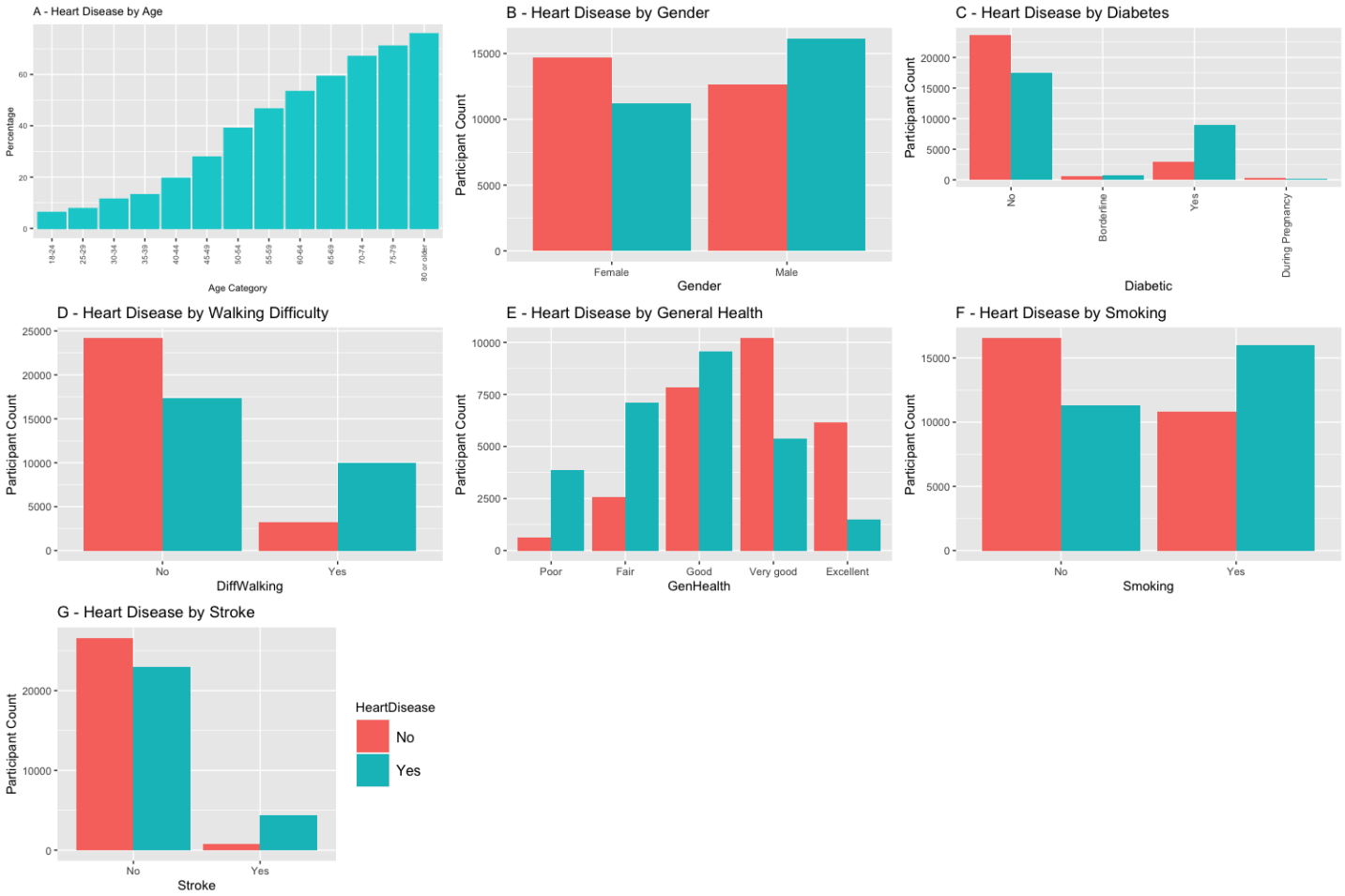
A - Heart Disease by Age
Percentage
Age Category
18-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
75-79
80 or older
B - Heart Disease by Gender
Participant Count
Gender
Female
Male
C - Heart Disease by Diabetes
Participant Count
Diabetic
No
Borderline
Yes
During Pregnancy
D - Heart Disease by Walking Difficulty
Participant Count
DiffWalking
No
Yes
E - Heart Disease by General Health
Participant Count
GenHealth
Poor
Fair
Good
Very good
Excellent
F - Heart Disease by Smoking
Participant Count
Smoking
No
Yes
G - Heart Disease by Stroke
Participant Count
Stroke
No
Yes
HeartDisease
No
Yes

Figure 1

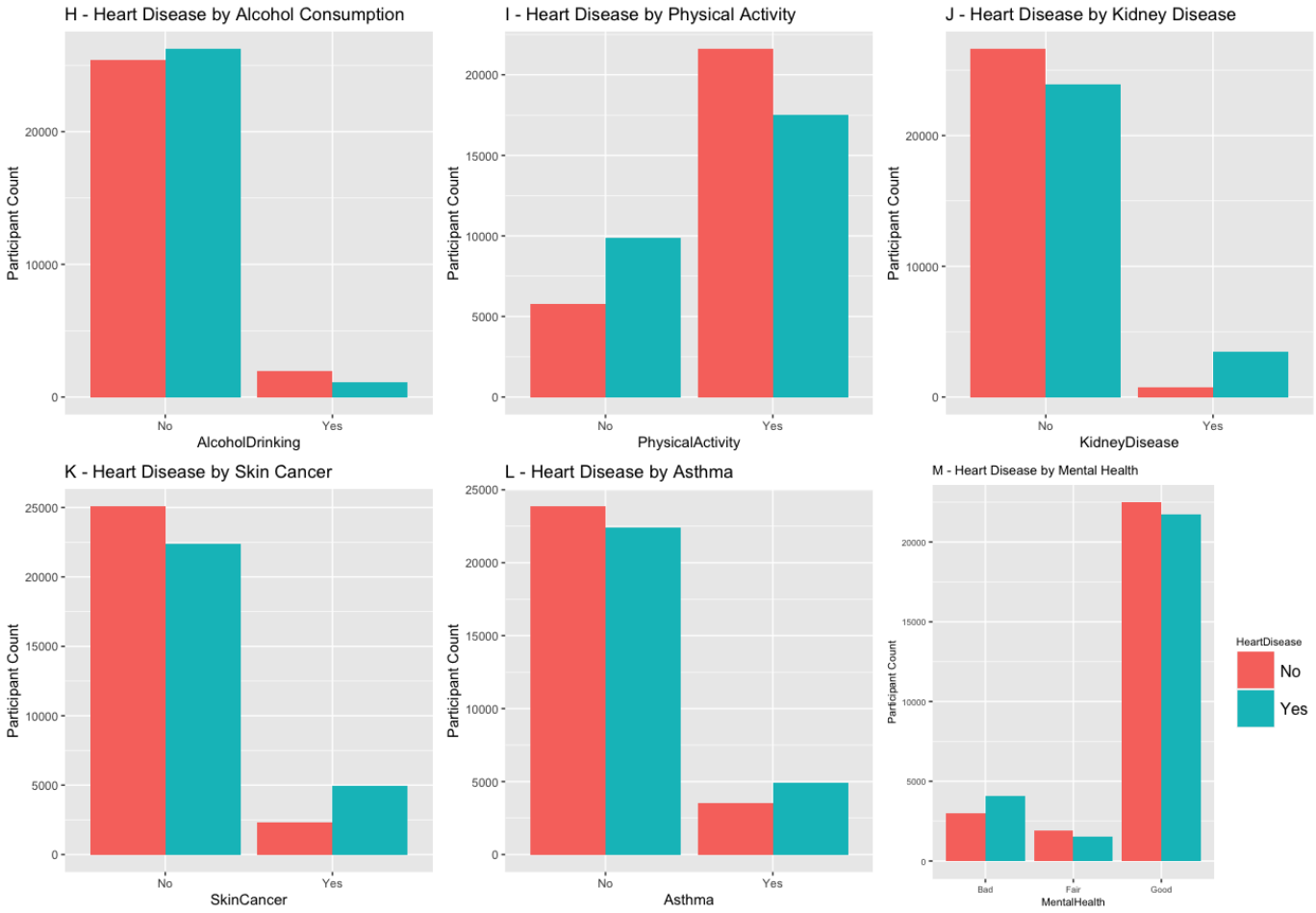
H - Heart Disease by Alcohol Consumption
Participant Count
AlcoholDrinking
No
Yes
I - Heart Disease by Physical Activity
PhysicalActivity
J - Heart Disease by Kidney Disease
KidneyDisease
K - Heart Disease by Skin Cancer
SkinCancer
L - Heart Disease by Asthma
Asthma
M - Heart Disease by Mental Health
MentalHealth
Bad
Fair
Good
HeartDisease
No
Yes

Figure 2

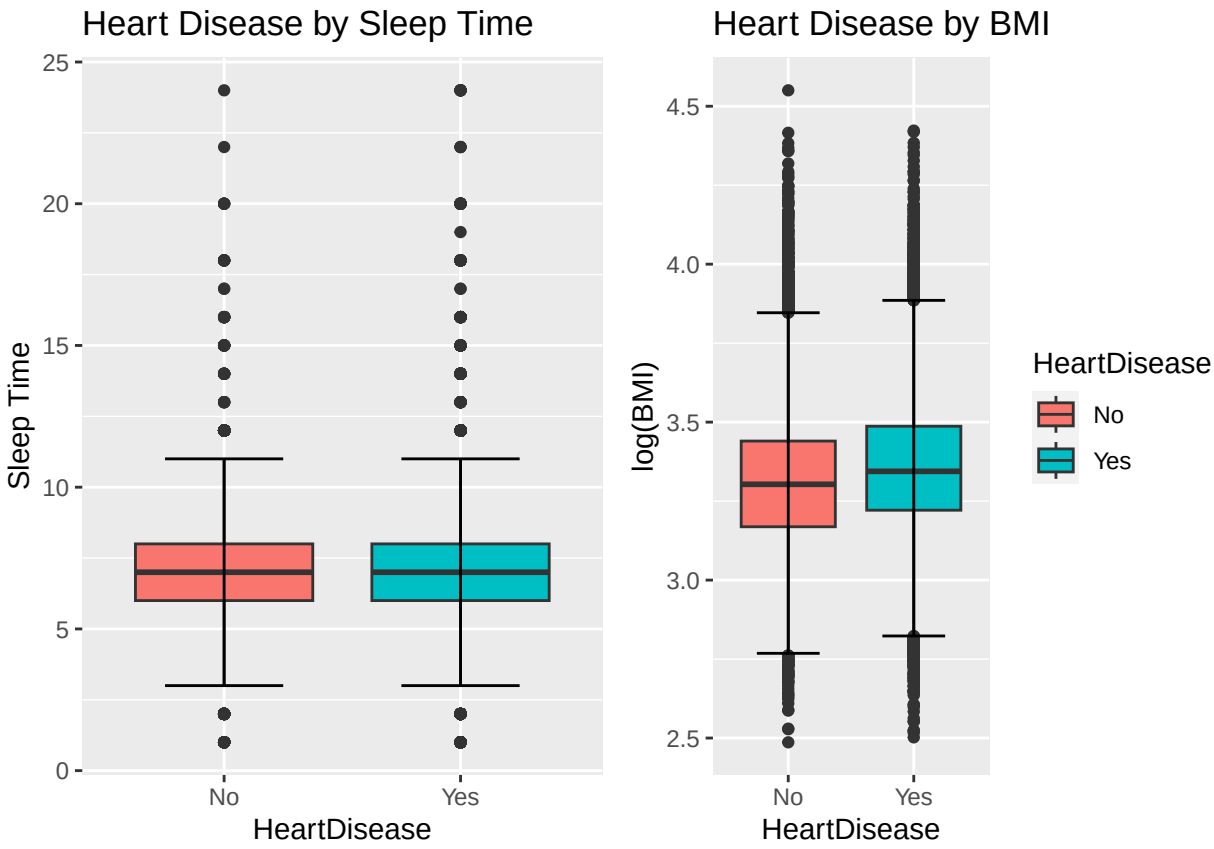
Heart Disease by Sleep Time
Sleep Time
25
20
15
10
5
0
No
Yes
HeartDisease
Heart Disease by BMI
log(BMI)
4.5
4.0
3.5
3.0
2.5
No
Yes
HeartDisease
HeartDisease
No
Yes

Figure 3

bar charts to see comparisons of categorical groups of data, maps to see geographical data, and many more to visualize every data type.

## Results and Analysis

Through data visualization, I can see that a majority of the variables in this dataset influence heart disease, some more than others. In particular, age, walking difficulty, smoking, and stroke have the greatest affect on heart disease in adults based on the visual information from this dataset. Another way to evaluate variables that affect heart disease is a Chi-square test of association to see the relationship each of one of them has with the HeartDisease variable. For this test, the null hypothesis is $H_0$: There is no association between the variables, and the alternative is $H_A$: There is an association between the variables.

For the test on each variable to heart disease, a p-value less than $\alpha = 0.05$ would mean that we accept the alternative hypothesis and there is an association between the variables. Further, a greater chi-squared/X-squared value means a greater association between HeartDisease and the target variable. From the chi-squared test on the categorical variables, each one of them had a p-value $<$ 2.2e-16. Using a p-value adjustment, the p-value of each variable is zero. Since this value is less than $\alpha = 0.05$, I can say that we reject the null hypothesis in favor of the alternative, and there is an association between each predictor variable and heart disease. Looking a the chi-squared values, in order of most associated to least associated, the variables are AgeCategory, GenHealth, DiffWalking, Diabetic, Stroke, Smoking, KidneyDisease, PhysicalActivity, SkinCancer, Sex, Asthma, MentalHealth, and AlcoholDrinking.

Additionally, in figure three, there are two groups visualized in each graph - SleepTime and BMI of individuals with and without heart disease. From looking at each graph, the difference between each group looks small and it's difficult to tell if there is a significant influence on heart disease based on an individuals BMI or average nightly sleep time. To determine if there is a difference between groups, I performed a t-test on each variable. The null hypothesis for these tests is $H_0$: The two population means are equal, and the alternative, $H_A$: The two population means are not equal.

Starting with BMI, from the two sample t-test, the p-value is $<$ 2.2e-16 which is a very small number less than 0.05. We reject the null in favor of the alternative and can say that there is a significant difference in means between the BMI value of people with and without heart disease. The mean BMI of individuals with heart disease is 29.40, and the mean without heart disease is 28.18.

Next is sleep time. After running a t-test, the p-value was .00221 which is also less than .05. The mean sleep time of individuals with heart disease is 7.136 hours and without heart disease is 7.094 hours. So, adults without heart disease tend to sleep less by a very small margin.

## Conclusion

Overall, through visualization, I found that of the 13 categorical variables, 12 of them had an affect on heart disease in adults. From the chi-squared association test, the top three variables that influence heart disease are age - individuals who are older have a higher risk. General health - people who classify themselves as having poor health have a higher probability of having heart disease. And walking difficulty - adults who struggle walking or climbing stairs are more likely to have heart disease. The predictors least associated with heart disease are mental health - a small proportion of people with bad mental health have heart disease and the majority of people with heart disease in this data set also had good mental health. Alcohol drinking also didn't seem to influence heart disease. For people who drank more, they tended to not have heart disease.

Further, from the t-tests of the numerical variables, the difference in mean values of sleep time and BMI for individuals with and without heart disease is very slim. I would say that BMI has a small affect on heart

disease in adults. The higher your BMI, the more influence it has on your risk for heart disease. On the other hand, the hours of sleep that survey participants get each night does not seem to have an affect on heart disease as the mean sleep time of each group was almost identical.

Lastly, if I was given more time on this project, I would use more than visualizations to determine which factors cause heart disease. Using classification models such as decision trees or logistic regression with cross validation would be a way to find the most important predictors that affect heart disease.

## Appendix

- Citations

"Know Your Risk for Heart Disease." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 9 Dec. 2019, https://www.cdc.gov/heartdisease/risk_factors.htm.

Pytlak, Kamil. "Personal Key Indicators of Heart Disease." Kaggle, 16 Feb. 2022, https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease/discussion/316999.

- GitHub

https://github.com/catalinamunoz/HeartProject