PM566 Final- Heart disease analysis

Jayson De La O.

Which health indicator have a greater association with heart disease?

INTRODUCTION: The Centers for Disease Control and Prevention (CDC) list heart disease as the leading cause of death in the United States. The goal of this analysis is to find out which health indicators are associated heart disease and will potentially be useful in predicting whether or not someone develop heart disease. Our dataset is a subset of data collected from the CDC's 2015-Behavioral Risk Factor Surveillance System (BRFSS), a telephone survey used to collect data on the health and condition of the residents of the United States of America. The data contains the binary outcome of heart disease or heart attack status along with different health indicators that could be useful in predicting heart disease, such as BMI, smoking status, and age. Most of the health indicator variables have binary responses to a health status or condition, with BMI being the quantitative variable in the data.

METHODS: The dataset was obtained from Kaggle and it is a subset of data from the 2015 CDC's telephone survey across the 50 states. The dataset used only includes health indicators that might be useful for predicting heart disease status across the nation. The dataset was downloaded and read in as a CSV file and the dimensions, header, and footer were all checked to ensure the data was properly displaying. The str function was used to look at the structure of the data and see that all of variables are numerical values. The summary function allows us to take a closer look at the variables and see the summary statistics of each variable and any missing values that might occur in the data. No missing values are observed and most variable are binary-categorical variables. All numerical-categorical variables are transformed using ifelse statements to their corresponding category using the Kaggle legend to make the responses more intuitive. The clean dataset was then restricted to variables that would seem to have an association including the new character variables and BMI(duplicates were removed). Exploratory data analysis was run on variables of interest and it is noted that BMI has a max of 98. A BMI of 98 seems to be a suspicious value because most BMI charts only list 30 and above as the highest BMI category. Using medical journals, I was able to verify that BMI values can exceed 100, but the majority of people do not fall into that category. Additionally, I calculated the proportion of individuals that had a BMI greater than 40 in our dataset was 4% and those greater than 50 was 0.8%, so it seems to be reasonable that these BMI values are not errors and will be included in our analysis. Exploratory graphs were created for variables of interest using ggplot and summary tables were created using summarize and

kable function. Barplots and histograms were used to study the distribution of our variables and help hypothesize what we might expect for the associations between health indicators and heart disease.

RESULTS:

Our initial analysis of BMI showed that individuals who had a lower BMI were less likely to have had heart disease issues. The boxplot of individuals BMI by heart disease (Figure 1), shows that there was a difference in the mean BMI between the two groups, with higher BMI being associated with heart disease. Since the sample size for indviduals with and without heart disease is not equal, we want to confirm this association may looking at the proportions of indviduals in each category. Table 1 confirms the higher BMI in the heart disease group and displays the sample size for each group.

Since the sample size heart disease is different from each health indicator, we focused mainly on the proportions of individuals in each category in order to get results that were more intuitive.

Age and gender were both associated with higher chance of having heart disease (Table 2 and 3). Males were more likely to have developed heart disease than females, but this could also be explained by differences in habits by the two groups not necessarily as biological predisposition. Additionally, it was noted that as individuals age the odds of devloping heart disease increases.

Differences in income and education level also showed to be associated with heart disease(Table 4 and 5). A difference in education and income seemed to have a strong association with heart disease at the lowest and highest levels. In general, as income and education increased, the odds of individuals have heart disease decreased. Figure 2 shows that the indivudals who has heart diseased seemed to be relatively equal across all income categories.

Lifestyle and pre-existing conditions (high blood pressure, has had a stroke) are two of the most influential risk factors in predicting if someone will develop heart disease. Blood pressure, diabetics status, stroke status and high cholesterol status are all pre existing conditions that were studied in our analysis. The results of each of the health indicator can be found below (Tables 6,7,8,14). All pre-existing conditions showed an increase in likelihood of developing heart disease across the board.

Most of the time, lifestyle plays a major role in health outcomes. In our analysis, we saw that, for the most part, the healthier a person lives, the less likely there are to have developed heart disease. For the most part, the analysis agreed that a healthier lifestyle decreased the odds of developing heart disease. Eating habits, smoking status, exercise status and heavy alcohol use status were all studied to learn about their association with heart disease Tables(9,10,11,12,13).

All healthy lifestyle habits, with the exception of no heavy alcohol use, were associated with a decrease odds of having heart disease. However, heavy alcohol use seemed to be associated with a decreased odds of developing heart disease. Figure 3 and table 11 show the association

Table 1: Table 1:Summary Statistics

HDorAttack	n	BMI_mean
No	229787	28.26962
Yes	23893	29.46662

of heavy alcohol use with heart disease. The finding is contrary to what we would have expect because heavy alcohol use is usually a sign of an unhealthy lifestyle. The data showed a decreased in odds of having heart disease for those individual who drank alcohol heavily. Although, this is an interesting find, it should be taken with some hesitancy because of the large difference in sample size. The sample size of individuals who drank heavily is small compared to those who do not drink, which could introduce sample size bias in our interpretations.

CONCLUSION: As we expected the health indicators seemed to be associated with heart disease. For the most part all the variables that seem to put you in a healthier category showed a decrease in heart disease. Although, the data did seem to have an exception were we show that an increase in heavy use of alcohol showed a decrease in heart disease, but this trend need more research because the data might have been skewed due to small sample size in the heavy use alcohol category. All variables seemed to be associated with heart disease and deserve to be furthered studied, but from the list of variables of interest heavy use of alcohol and fruit eating status seemed to be the least associated between the variables.

Figure 1. Boxplot of BMI by Heart Disease Status

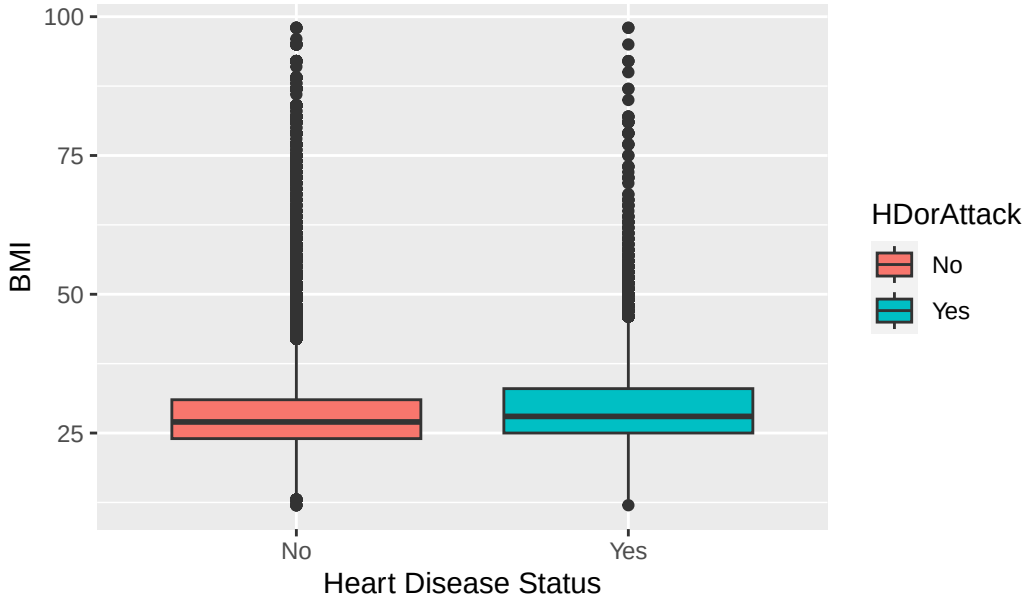

Table 2: Proportion of Heart Disease by each Age Category

	Age 18 to 24	Age 25 to 29	Age 30 to 34	Age 35 to 39	Age 40 to 44	Age 45 to 49
Heart Disease	0.9949123	0.9928929	0.9886721	0.9860378	0.9782757	0.9640749
No Heart Disease	0.0050877	0.0071071	0.0113279	0.0139622	0.0217243	0.0359251

Table 3: Proportion of Heart Disease by Gender

	Female	Male
Heart Disease	0.9281206	0.8774641
No Heart Disease	0.0718794	0.1225359

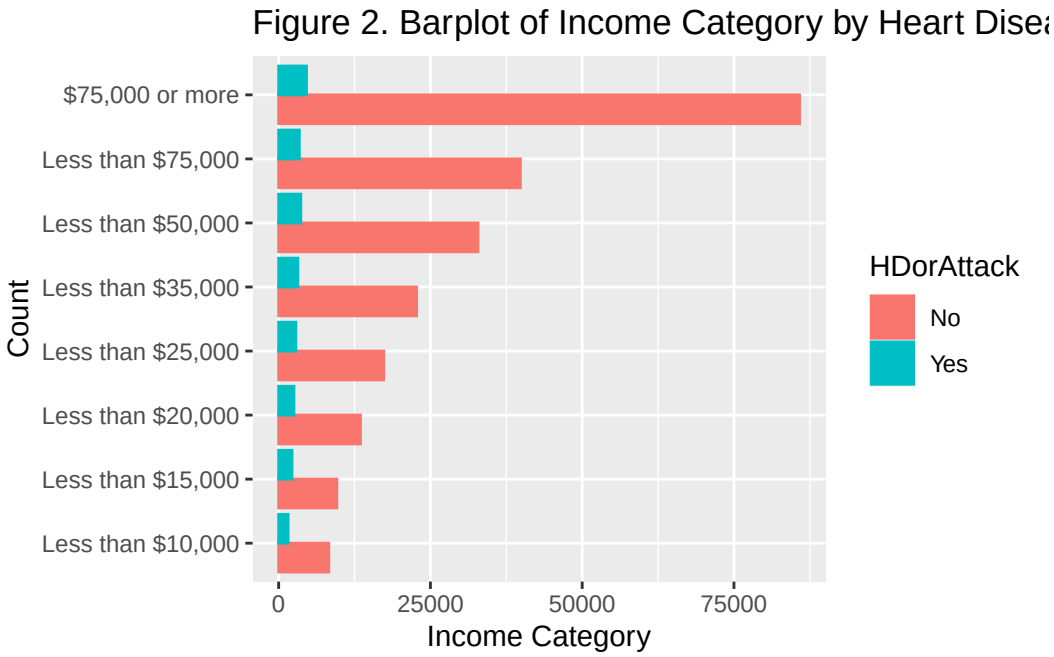

Table 4: Proportion of Heart Disease by Education Level

	College graduate	Elementary	High School	No school	Some college	Some High Sc
Heart Disease	0.9340042	0.8075686	0.881004	0.8333333	0.9010442	0.829
No Heart Disease	0.0659958	0.1924314	0.118996	0.1666667	0.0989558	0.170

Table 5: Proportion of Heart Disease by Income

	$75,000 or more	Less than $10,000	Less than $15,000	Less than $20,000	Less than
Heart Disease	0.9492726	0.8417083	0.8135449	0.8425034	0.
No Heart Disease	0.0507274	0.1582917	0.1864551	0.1574966	0.

Table 6: Proportion of Heart Disease by Blood Pressure

	High BP	No High BP
Heart Disease	0.9588198	0.8352645
No Heart Disease	0.0411802	0.1647355

Table 7: Proportion of Heart Disease by High Cholestrol

	High Cholestrol	No High Cholestrol
Heart Disease	0.9511257	0.8442899
No Heart Disease	0.0488743	0.1557101

Table 8: Proportion of Heart Disease by Stroke Status

	stroke	No stroke
Heart Disease	0.9180075	0.6174699
No Heart Disease	0.0819925	0.3825301

Table 9: Proportion of Heart Disease by Smoking status

	Smoke	No smoke
Heart Disease	0.935635	0.8683454
No Heart Disease	0.064365	0.1316546

Table 10: Proportion of Heart Disease by Exercise status

	excercise	No exercise
Heart Disease	0.8608646	0.9202793
No Heart Disease	0.1391354	0.0797207

Table 11: Proportion of Heart Disease by Heavy Alocohol Use

	Alcohol Use	No Alcohol
Heart Disease	0.9037482	0.9405163
No Heart Disease	0.0962518	0.0594837

Table 12: Proportion of Heart Disease by if they eat fruit

	Eats fruits	No fruits
Heart Disease	0.8982022	0.910204
No Heart Disease	0.1017978	0.089796

Figure 3.Barplot of High Alcohol use Status by Heart Diseas

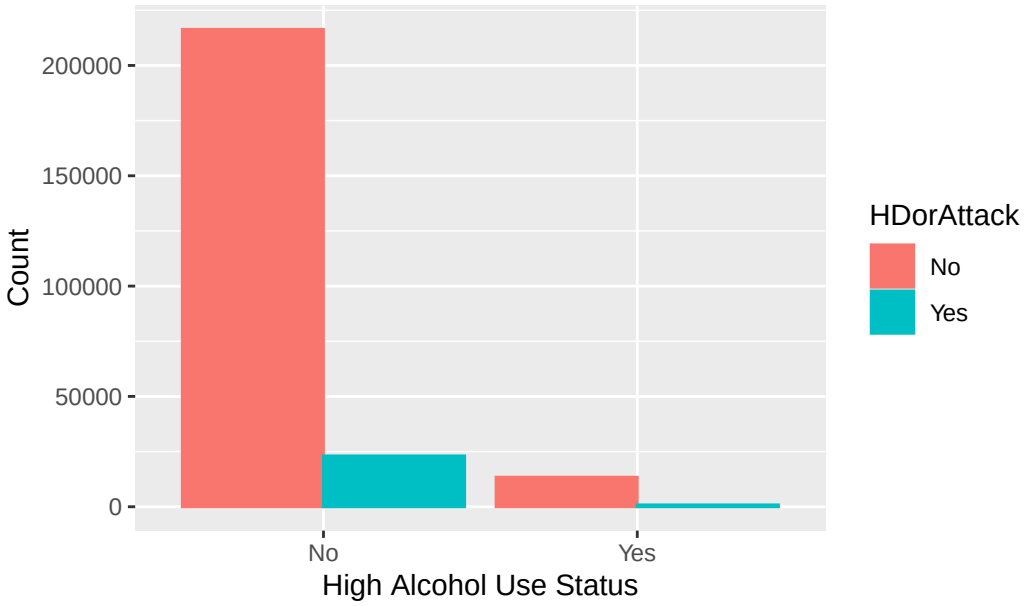

Table 13: Proportion of Heart Disease by if they eat vegetables

	Eats Vegetables	No Vegetables
Heart Disease	0.8820837	0.9113296
No Heart Disease	0.1179163	0.0886704

Table 14: Proportion of Heart Disease by Diabetes status

	Diabetes	No Diabetes	Pre-diabetes or borderline diabetes
Heart Disease	0.7771176	0.9281667	0.8566184
No Heart Disease	0.2228824	0.0718333	0.1433816