STA 160 Midterm Report

Nicholas Kwak, Herman Chee, Eric Liu

I. INTRODUCTION

Heart disease continues to be a leading cause of death in the United States impacting millions of Americans each year. Therefore, we will be examining survey data collected through some of the telephone responses in 2015 by the Centers for Disease Control and Prevention. We will be exploring the different indicators from this large dataset to see if survey responses can predict whether a person has or does not have heart disease. Using different analytical techniques, we can also use these findings to help with preventative screening questions for assessing risk of heart disease. Our inspiration for this report was predicting whether a person has heart disease or not based on their survey responses. Below, you can see more details regarding our dataset, methodology, and results.

II. DATA

The dataset we are using is called the Heart Disease Health Indicators Dataset. There are a total of 253,680 survey responses or rows where 229,787 people mention never having heart disease while 23,893 have had some type of heart disease. There are a total of 22 columns with the first being the binary variable of whether the respondent has heart disease or not. The 21 other columns each represent different indicators ranging from self-assessment ratings to physical readings like blood pressure and age. All the observations are either binary or numerical.

III. METHODOLOGY

One of our first techniques will be to create a heatmap so that we can observe some of the factors that are the closest related to heart disease. We will only keep the variables with the strongest correlations so that we can improve our ability to accurately predict the chances of HeartDiseaseorAttack being true. For the factors that are binary, we will create contingency tables to observe the relationship between the two variables. We will also calculate the odds ratio,

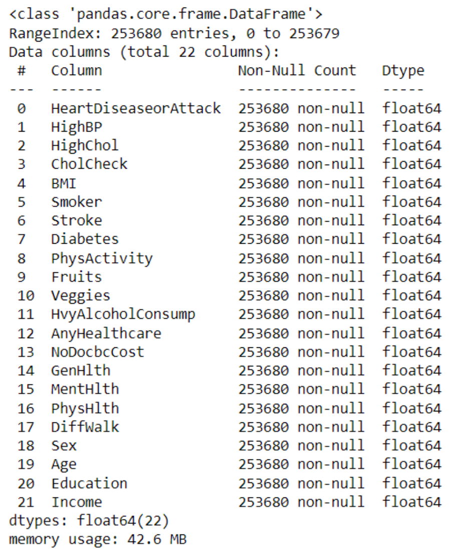
$$\frac{\frac{N(HeartDisease = 1 \,\&\, x = 1)}{N(HeartDisease = 0 \,\&\, x = 1)}}{\frac{N(HeartDisease = 1 \,\&\, x = 1)}{N(HeartDisease = 0 \,\&\, x = 0)}}$$

where x is the variable we are looking at. Furthermore, there will be several different models that we will utilize to describe which factors influence the risk of heart disease/attack. Through linear regression, odds-ratio, lasso model, and chi square testing, we will determine which model produces the best results for deciding which variable leads to heart disease.

IV. DATA PREPROCESSING

To begin the data analysis process, we first needed to check the dataset in case of any irrelevant variables or null values. To do this we first need to describe our dataset using the info function.

Table 1



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 253680 entries, 0 to 253679
Data columns (total 22 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   HeartDiseaseorAttack  253680 non-null  float64
 1   HighBP                253680 non-null  float64
 2   HighChol              253680 non-null  float64
 3   CholCheck             253680 non-null  float64
 4   BMI                   253680 non-null  float64
 5   Smoker                253680 non-null  float64
 6   Stroke                253680 non-null  float64
 7   Diabetes              253680 non-null  float64
 8   PhysActivity          253680 non-null  float64
 9   Fruits                253680 non-null  float64
 10  Veggies               253680 non-null  float64
 11  HvyAlcoholConsump     253680 non-null  float64
 12  AnyHealthcare         253680 non-null  float64
 13  NoDocbcCost           253680 non-null  float64
 14  GenHlth               253680 non-null  float64
 15  MentHlth              253680 non-null  float64
 16  PhysHlth              253680 non-null  float64
 17  DiffWalk              253680 non-null  float64
 18  Sex                   253680 non-null  float64
 19  Age                   253680 non-null  float64
 20  Education             253680 non-null  float64
 21  Income                253680 non-null  float64
dtypes: float64(22)
memory usage: 42.6 MB
```

From this, we gathered that there are no missing values within the dataset and none of the variables contain strings. With 20 different indicator variables we decided to create a heatmap that displayed the correlation between those variables to see which ones are irrelevant to our analysis.

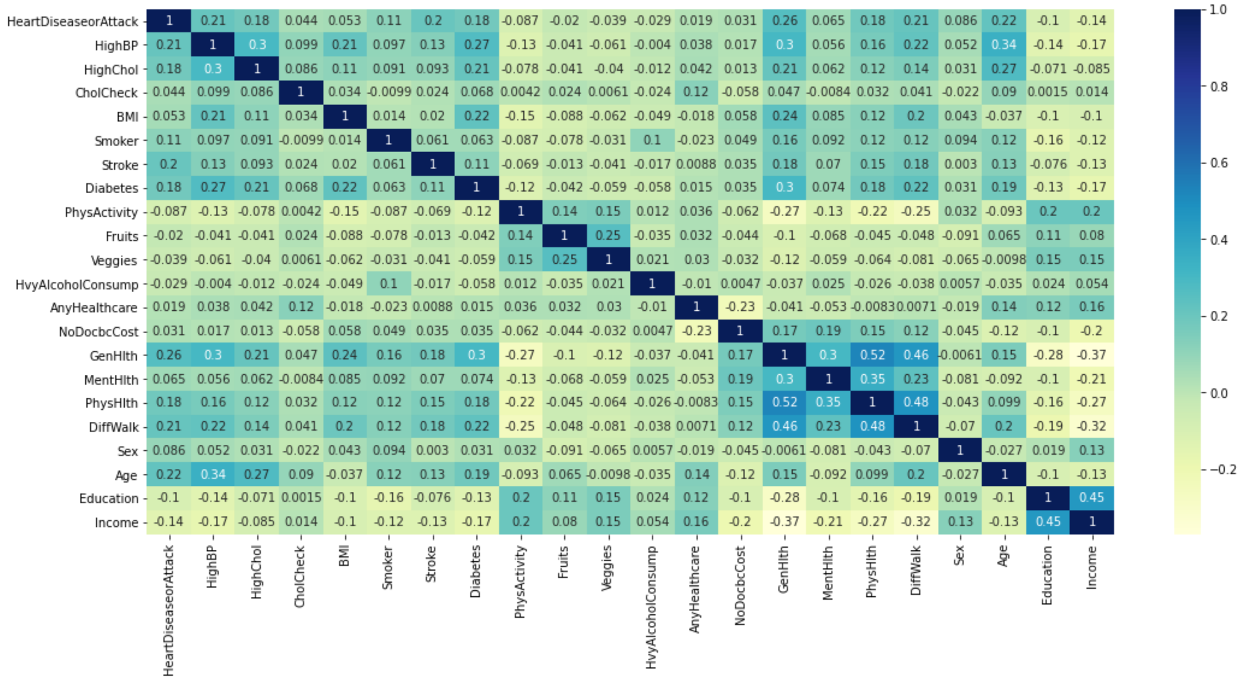


Figure 1

V. RESULTS/ANALYSIS

A. *Initial Data Analysis*

Using our correlation heatmap, we observed the predicted relationship all the variables had with each other. Some of the strongest correlations to heart disease were GenHealth, Age, DiffWalk, and HighBP, and Stroke. Other factors did not have a strong correlation to heart disease. We ended up removing 12 of the variables in the dataset such as Sex and Healthcare due to their weak relationship to HeartDisease.

Table 2

Variable	**Odds-Ratio**
HighBP	4.592098602559366
HighChol	3.5890725604845954
Smoker	2.2039431659792883
Stroke	6.936202083608329
DiffWalk	4.2660852912766645

As you can see, having bad health tends to increase your odds of suffering from heart disease. Suffering from one of these conditions doubles your odds, and having a stroke multiplies that risk by seven times.

One of the few variables that were numerical instead of categorical were BMI so taking a closer look, we can get some interesting results from our population of American adults. BMI stands for body mass index and is a measurement used to measure someone's general health based on their height and weight. The standard for a healthy BMI range is 18.5 to 24.5 with anything below classified as underweight and anything above considered overweight. Below, you could see the healthy range only captures a minority within the black bands. Most American adults are over the healthy BMI range showing a distinct trend of unhealthy body mass approximately normally distributed around the mean BMI of 28.38.

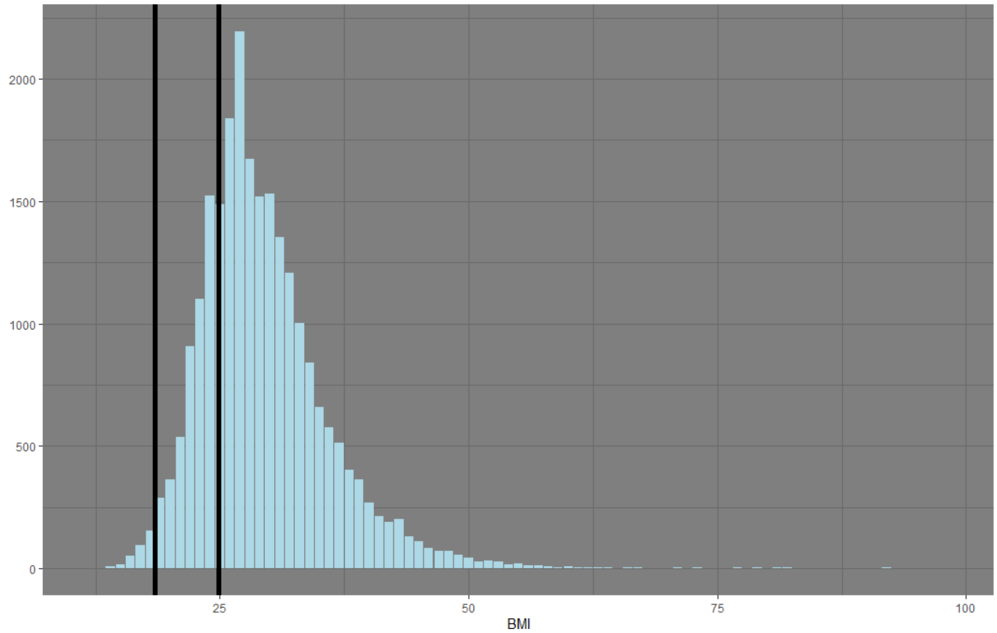


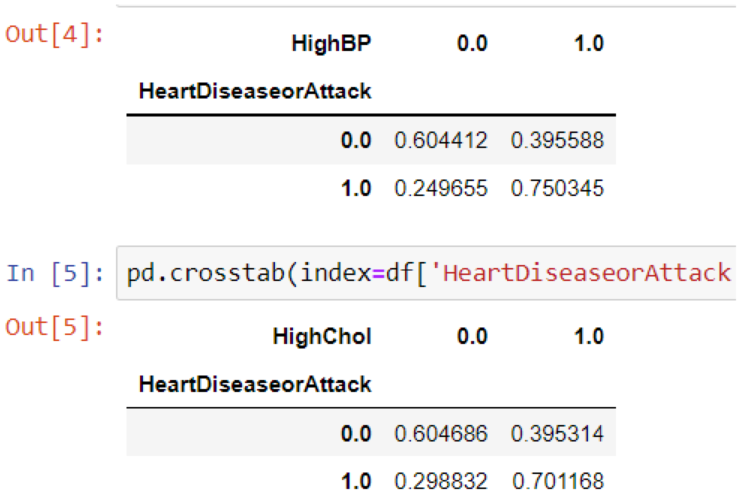
Figure 2

While BMI could be slightly inaccurate because it does not take into account bone density, it does give us a general understanding of the overall health of our population data objectively which we could compare to the subjective general health survey data. There are approximately 17% of American adults who fall within the healthy BMI range whereas the survey reports 51% of them think they have excellent or good general health (1 and 2). This vast disparity is significant to note indicating many Americans may not be taking health as seriously as they should be thinking they are healthy when they are not. Something we could do with BMI to convert the variable to categorical is by splitting the numbers into 4 groups called underweight, healthy, overweight, and obese.

B. Contingency Tables

We also used contingency tables in our analysis giving us a general picture and understanding of the associations of our variables. We started with a 2x2 contingency table of high blood pressure compared to whether a person had heart disease or not. Contingency tables were one of the most useful tools in our initial exploratory data analysis which gave us an overview of the impact that variables had on heart disease risk.

Table 3



Out[4]:

HighBP	0.0	1.0
HeartDiseaseorAttack		
0.0	0.604412	0.395588
1.0	0.249655	0.750345

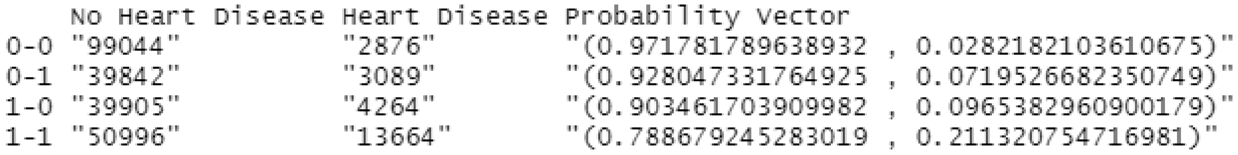
```
In [5]: pd.crosstab(index=df['HeartDiseaseorAttack
```

Out[5]:

HighChol	0.0	1.0
HeartDiseaseorAttack		
0.0	0.604686	0.395314
1.0	0.298832	0.701168

We also created contingency tables to explore the in-depth analysis of each variable when both high blood pressure and high cholesterol were considered jointly. The 4x2 contingency table below is an example of this where the first row indicates when no high blood pressure or cholesterol was present, the second row represents when high blood pressure was present but high cholesterol was not, and so on. The last row would tell us that those people in our data that had high blood pressure and high cholesterol had a 21% probability of classifying in the category of heart disease risk which we found was high.

Table 4



```
     No Heart Disease Heart Disease Probability Vector
0-0 "99044"          "2876"         "(0.971781789638932 , 0.0282182103610675)"
0-1 "39842"          "3089"         "(0.928047331764925 , 0.0719526682350749)"
1-0 "39905"          "4264"         "(0.903461703909982 , 0.0965382960900179)"
1-1 "50996"          "13664"        "(0.788679245283019 , 0.211320754716981)"
```

Using the tables with conditional and marginal probabilities, we were able to compute how likely a person with other variables present could end with heart disease. For example, our data suggests that given the fact you have diabetes, you have a 22% likelihood of heart disease at some point in your life. That is a 8% increase of risk compared to those with pre-diabetes and 15% increase compared to those who do not have diabetes which is shown below. Thus, we can say those with diabetes have a slightly higher chance of heart disease compared to those with a stroke and high blood pressure according to our data.

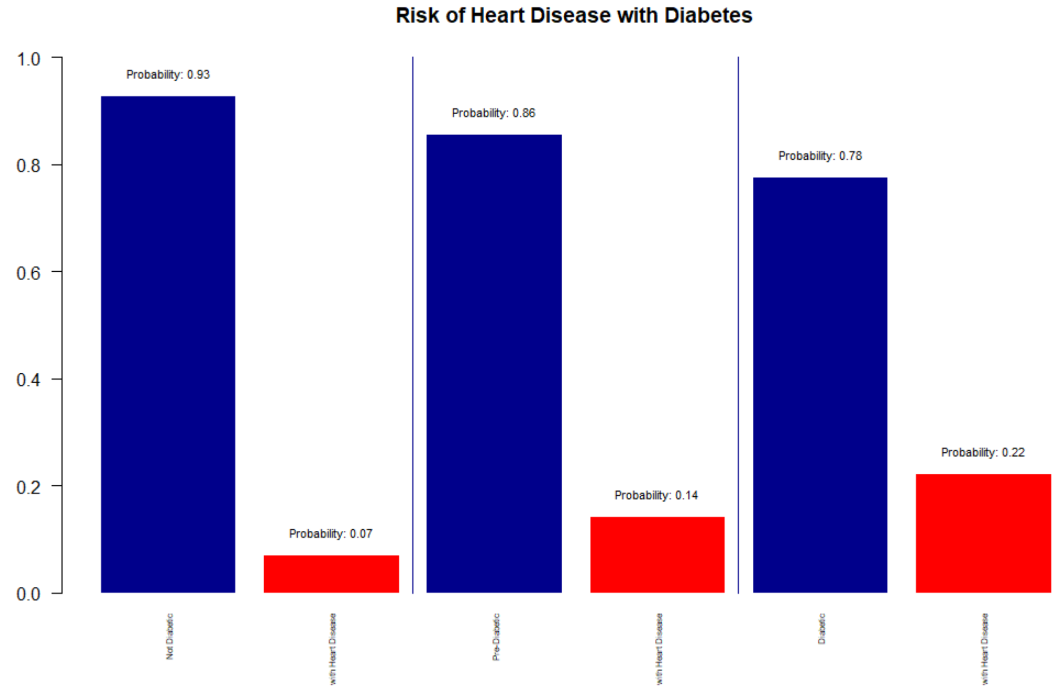


Figure 3

We also noted how stroke has the highest log-odds ratio so we computed contingency tables comparing this to other important variables to see the effect when both features are present. We computed joint contingency tables with stroke and the 4 following variables: Blood pressure, Cholesterol, Smoking, and Difficulty Walking. Using a bar plot, you can see how the risks increase dramatically when stroke is present along with these 4 different risk factors. The biggest increase in probability of heart disease occurs when a person simultaneously has reported difficulty walking and a stroke.

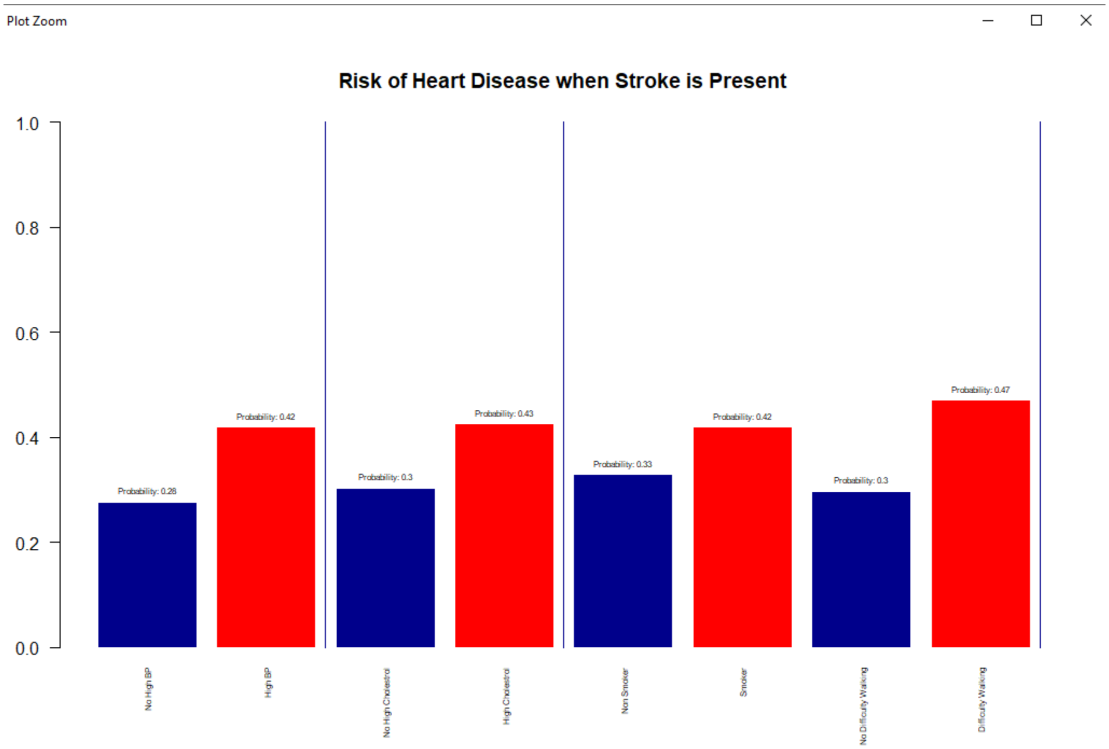


Figure 4

From the barplot below of people who never had a stroke before, you can see how the 4 different variables each a high probability of heart disease has when looking at the data alone. When people do not have a stroke, difficulty walking seems to be the highest risk factor for heart disease since it has a probability of 0.2 whereas the others have much lower probability. For example, people who smoke and didn’t have a stroke have the lowest risk of heart disease which is a surprising result since smoking with a stroke has a dramatically high increase in heart disease risk. This is also true when stroke was present which indicates difficulty walking may be the 2nd most important variable when considering risk of heart disease after looking at stroke.

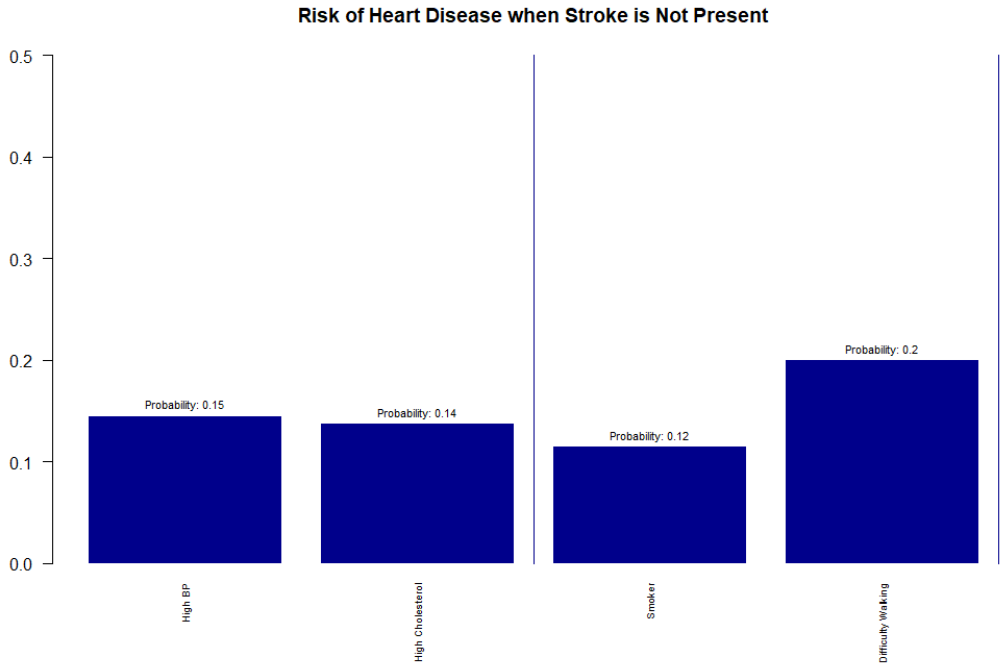
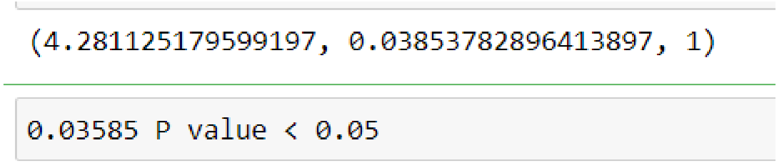


Figure 5

C. Chi Squared Testing

Another method we decided to use in finding the most impactful factors for heart disease was chi square testing. This test allows us to determine whether different categorical variables are likely to be related to each other. Using the chi square test statistic, we can see which variables affect the dependent variable of heart disease most by checking for dependency. After performing the test on all the highly correlated variables, we found that the most dependent variable of heart disease is Stroke. By performing the test on our previous contingency tables for stroke, we have the following results.

Table 5



```
(4.281125179599197, 0.03853782896413897, 1)
```

```
0.03585 P value < 0.05
```

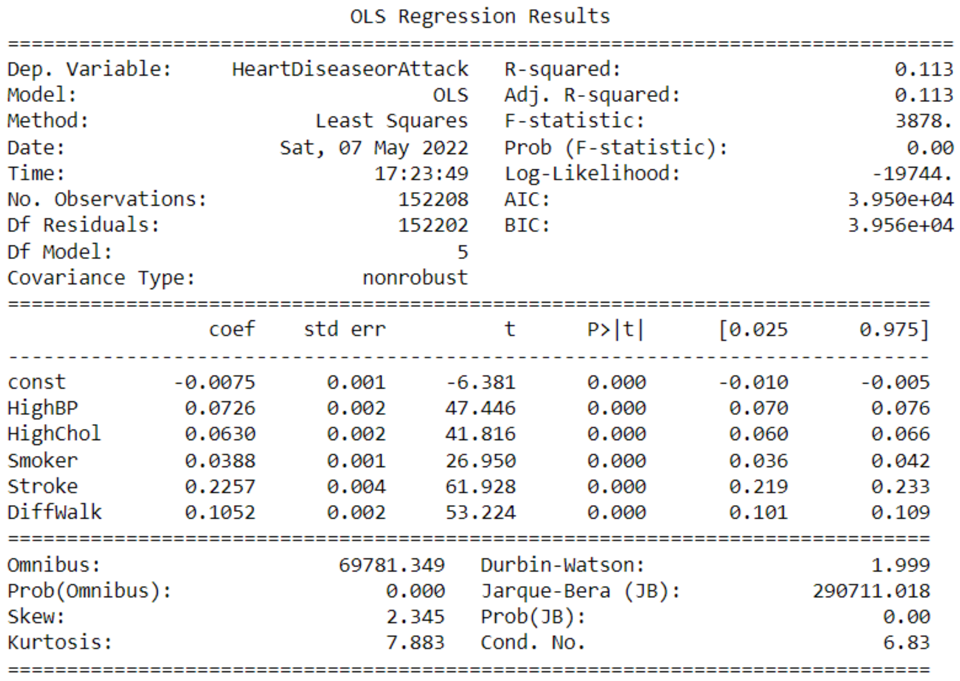
We find that our variable for stroke has a p value of 0.0365 for independence. This means that with a significance value of 0.05, we can conclude that there is significant evidence that this variable has dependency with heart disease/attack. This result also lines up with the findings from our odd-ratio test where the stroke variable had the largest ratio. Therefore, this is further evidence that stroke is most likely the most important variable when determining the risk of heart disease of a person.

D. Linear Regression

Now that we have a pretty good idea of what the most defining variable for heart disease is, we can perform linear regression on our most correlated variables to see just how much each variable would affect our heart disease dependent variable.

When creating our linear model, we used the 5 variables we kept for our explanatory variables, while having HeartDiseaseorAttack as our response variable. After creating the regression model, we have the following summary statistics.

Table 6



```
                            OLS Regression Results
==============================================================================
Dep. Variable:     HeartDiseaseorAttack   R-squared:                       0.113
Model:                              OLS   Adj. R-squared:                  0.113
Method:                   Least Squares   F-statistic:                     3878.
Date:                  Sat, 07 May 2022   Prob (F-statistic):               0.00
Time:                          17:23:49   Log-Likelihood:                -19744.
No. Observations:                152208   AIC:                         3.950e+04
Df Residuals:                    152202   BIC:                         3.956e+04
Df Model:                             5
Covariance Type:              nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.0075      0.001     -6.381      0.000      -0.010      -0.005
HighBP         0.0726      0.002     47.446      0.000       0.070       0.076
HighChol       0.0630      0.002     41.816      0.000       0.060       0.066
Smoker         0.0388      0.001     26.950      0.000       0.036       0.042
Stroke         0.2257      0.004     61.928      0.000       0.219       0.233
DiffWalk       0.1052      0.002     53.224      0.000       0.101       0.109
==============================================================================
Omnibus:                    69781.349   Durbin-Watson:                   1.999
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           290711.018
Skew:                           2.345   Prob(JB):                         0.00
Kurtosis:                       7.883   Cond. No.                         6.83
==============================================================================
```

Since we are most interested in how one of our factors affects the heart disease variable, we look to see which variable has the highest coefficient. From this result, the stroke variable again has the highest influence on whether a person gets heart disease. The other variables are much lower with the closest one being the difficulty of walking variable.

VI. CONCLUSION

Based on our in-depth analysis of the main five risk factors on heart disease, we were able to conclude the individual and joint effect they each have on a person's risk of developing heart disease sometime in their lifetime. Along with that, we were able to derive some interesting results on how each of these factors play a role together to magnify the risk and develop conclusions about the increased probability risk for specifically adults in America. With the techniques of log-odds ratio, contingency tables, chi-squared tests, and linear regression, we were able to figure out the order of heart disease risk factors of the five we focused on. For example, stroke was the highest risk factor followed by difficulty walking. While the correlation heatmap and contingency gave us a general idea of the data, the analytical techniques helped to confirm these initial exploratory findings. We also found interesting results about how diabetes was a risk factor that did not have a high correlation in the heatmap. Another surprising result was the BMI indicator which showed many adults lying in unhealthy ranges when over 50% believed their general health was good.

APPENDIX/SUPPLEMENTARY MATERIALS

1. Kaggle Dataset: https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset
2.