Analysis of Heart Disease

Shih-Chi Chen, Kevin Xu, Selam Berekat, Wyatt Workman

7 May 2022

Abstract

Heart disease is the leading cause of death in the United States, claiming approximately 700,000 lives each year (CDC 2022). In this report, we explore a data set relating to heart disease. Our goals in this project include an investigation into which factors contribute heavily towards heart disease, establishing which predictors are statistically significant in the population, and developing a statistical model to predict the likelihood of a patient developing heart disease.

1 Data Description

The heart disease data set is collected by the health-related telephone survey that has been held annually by the CDC since 1984, known as the Behavioral Risk Factor Surveillance System (BRFSS). In this project, the data set is derived from the BRFSS survey from 2015, which can be found on Kaggle.com, and in our submitted zip folder. This data set contains 253,680 respondents and has 22 variables. Among them, 19 are category variables and 3 are continuous variables. The response variable is a binary variable indicating whether or not an individual had heart disease.

It should be noted that the variables in this data set were derived and cleaned by the individual that uploaded the data set to Kaggle. Therefore, this report will not contain any information regarding data cleaning or transformations, as they were not necessary for this data set. Also as a result, there were no missing values found in the data set.

2 Summary of Data

In this section, we will discuss our methodology for an exploratory data analysis (EDA). The main objective of this section is to explore the data by creating rich visualizations that will provide valuable insight into which features are important. Note that we will not include all of the visualizations in this report, only those that revealed valuable insights. To see all of our visualizations, please consult the notebooks section of our .zip file.

2.1 Correlation Plot

Our first step in exploring the data was to produce a simple correlation matrix of all variables in the data set. In doing so, we would be able to determine if there were any strong linear relationships among the predictors that may warrant further statistical analysis. Below, figure 1 shows the correlation matrix of all variables.

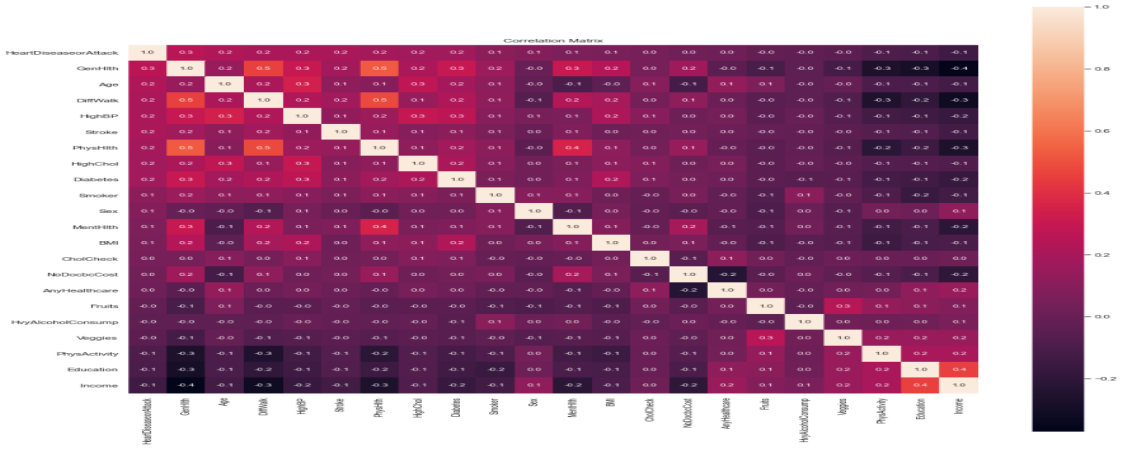


Figure 1: correlation matrix of all variables, see section 2 on continuous variables in the STA160 midproject eda Cindy.ipynb for more details.

From this plot, we were able to confirm that there were no strong correlations between any of the variables, where strong correlation is defined as an r^2 value of 0.7 or greater. However, this does not suggest that there are no associations between any of the variables. It is possible that there are strong associations between the variables and the response, but these association may be non-linear. We will examine possible associations in subsection 2.3.

2.2 Bar plots

Given that most of the variables in our data set were categorical, we decided to use bar plots to determine if there were significant differences between each covariate and the response variable. We faceted them by whether an individual experienced heart disease or not, and, for some of the bar plots, we plotted the densities rather than the raw count. We chose to do this to account for the class imbalance in the target variable. The following figures are the bar plots that our group found the most interesting, and therefore the variables that warranted further statistical analysis. All bar plots can be found in section 3 of the midterm project histograms notebook.

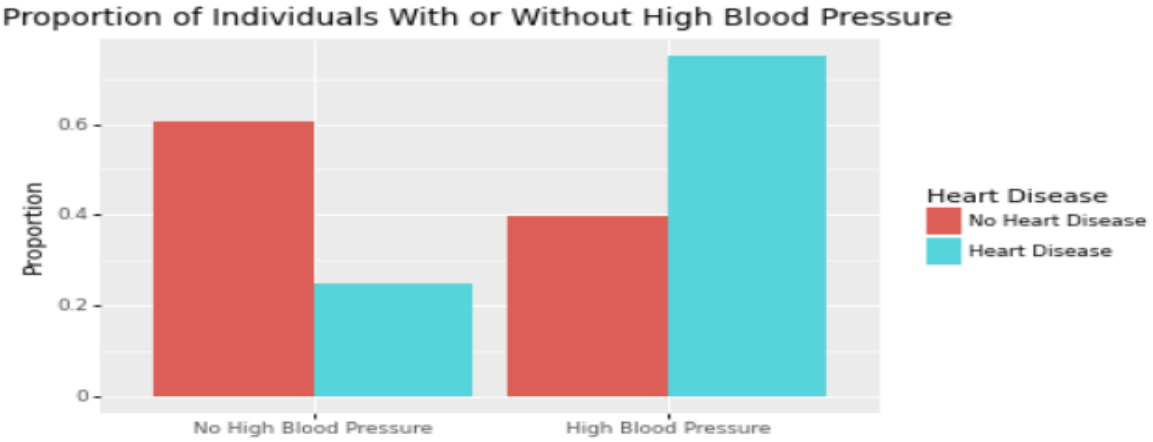


Figure 2: Bar plot of high blood pressure and heart disease.

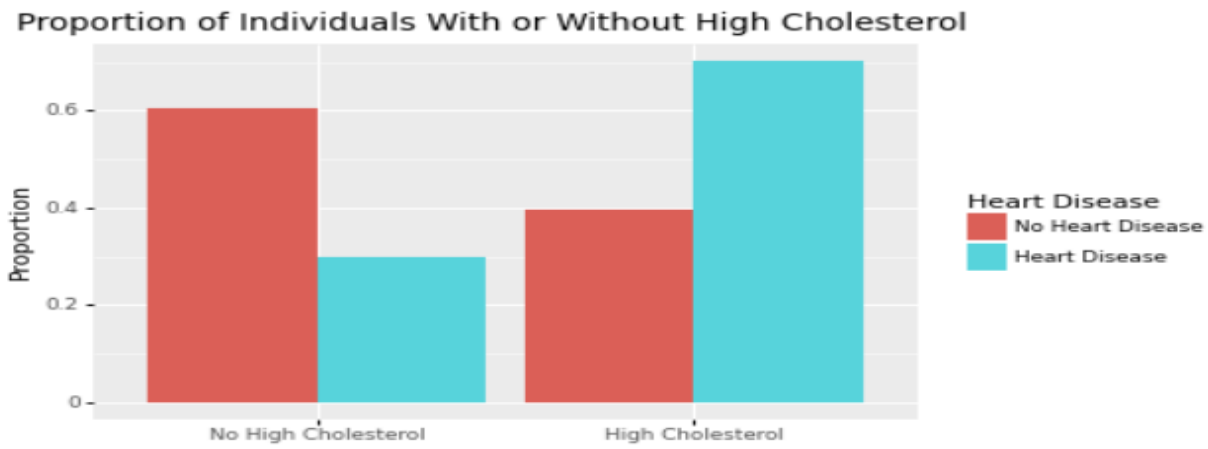


Figure 3: Bar plot of high cholesterol and heart disease.

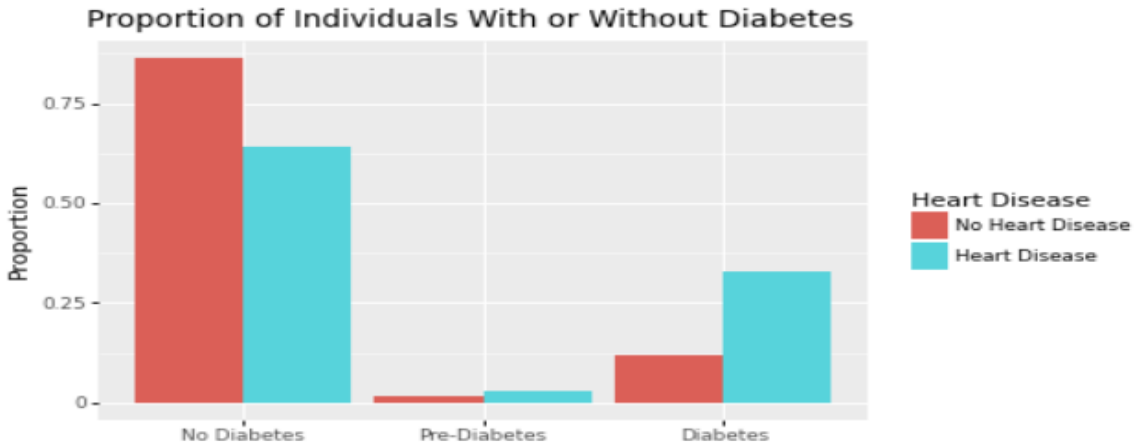


Figure 4: Bar plot of diabetes level and heart disease.

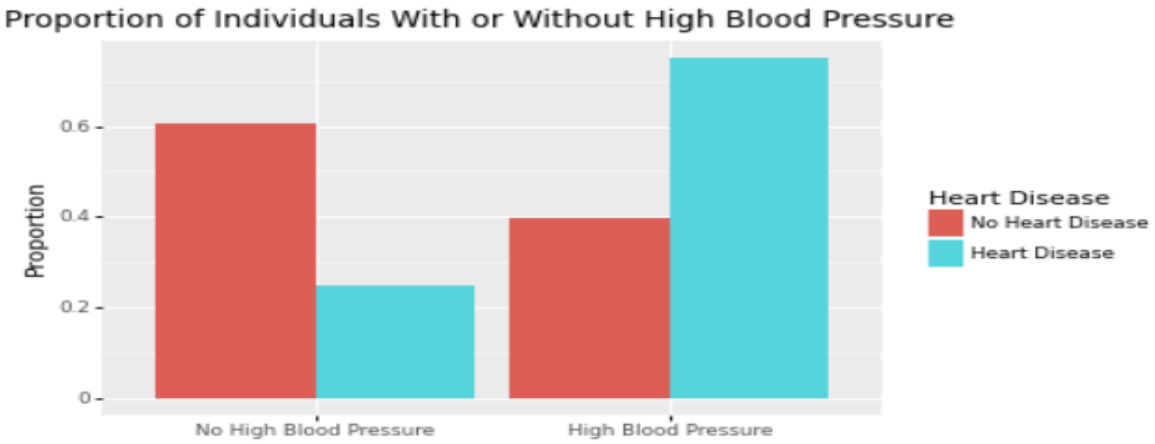


Figure 5: Bar plot of high blood pressure and heart disease.

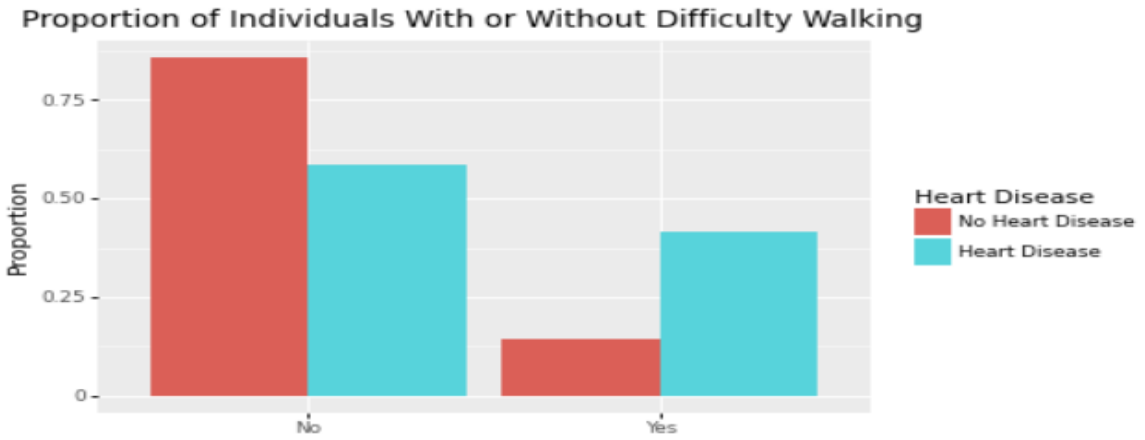


Figure 6: Bar plot of difficulty walking up stairs and heart disease.

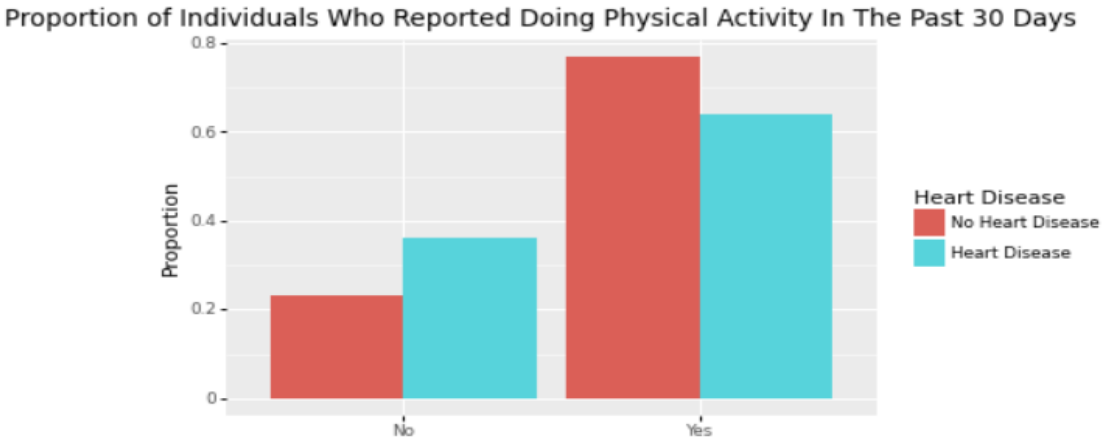
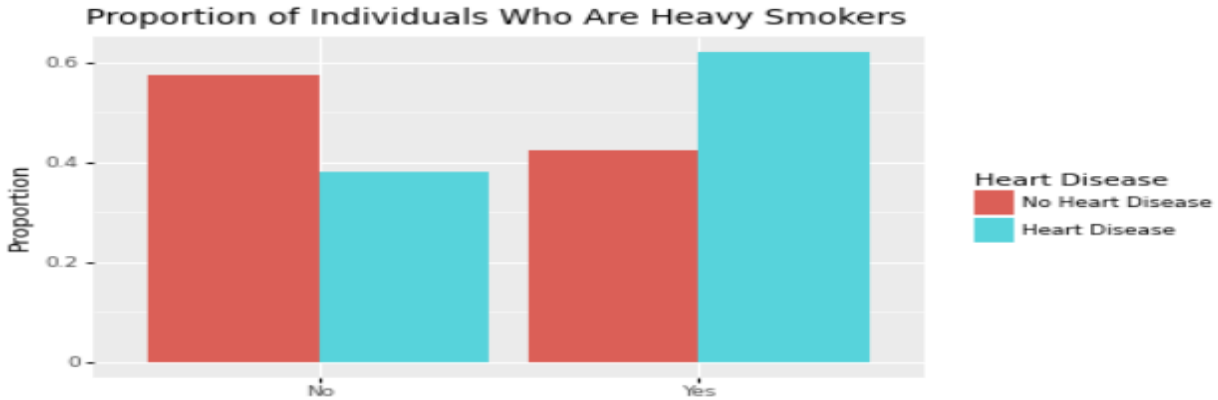


Figure 7: Bar plot of physical activity in the past 30 days and heart disease.




```
<ggplot: (137268903283)>
```

Figure 8: Bar plot of smoking and heart disease.

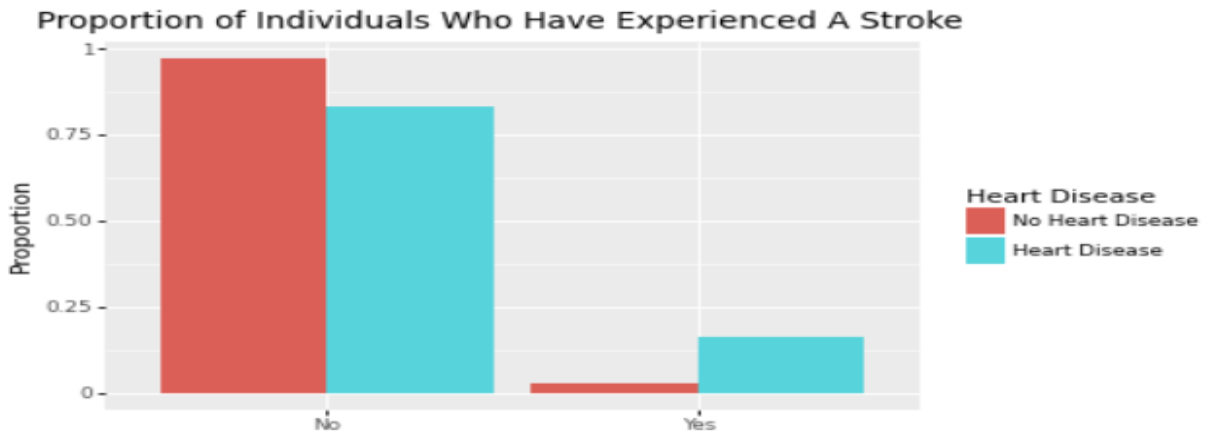


Figure 9: Bar plot of stroke and heart disease.

Although there were several other covariates, they did not have as a significant difference between groups (visually) as these variables. Again, all bar plots can be seen in section 3 of the midterm project histograms notebook.

2.2.1 Multivariate Bar Plots

Next, we considered multivariate bar plots, in order to visualize any relationships between multiple variables and the response. the proceeding figures can be found in section 3 of the STA160 midproject eda Cindy notebook.

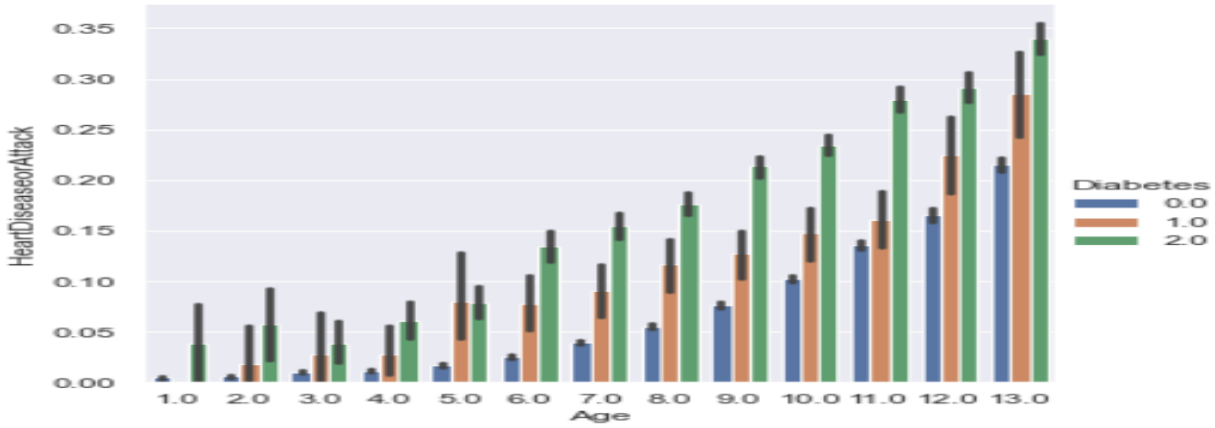


Figure 10: Bar plot of heart disease as a function of age and diabetes type.

Figure 10 shows us that not only does heart disease risk increase as age increases, heart disease also increases based on diabetes type. We can see that in almost all of the age brackets, individuals with diabetes(2) and prediabetes (1) have a much higher likelihood of developing heart disease than those who do not have any form of diabetes (0). Next, in figure 11, we consider the effects of high blood pressure and high cholesterol on heart disease.

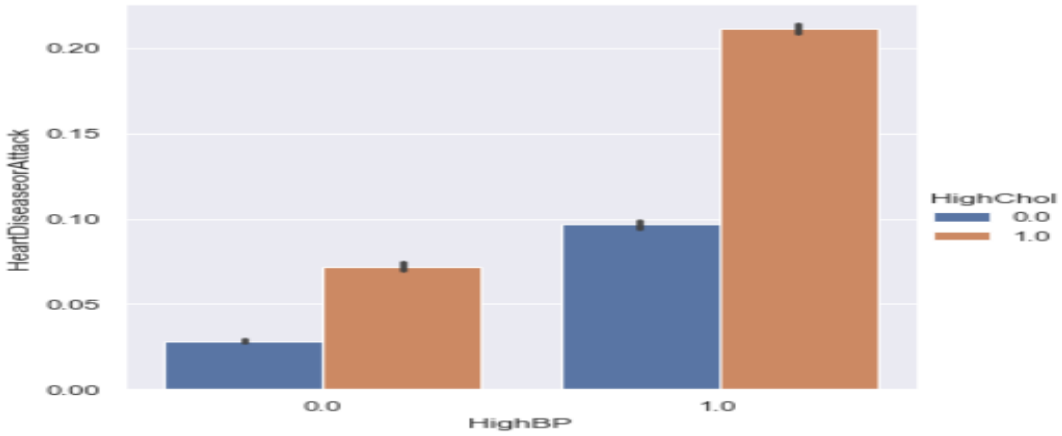


Figure 11: Bar plot of heart disease as a function of high blood pressure and high cholesterol.

Figure 11 shows that heart disease risk increases dramatically depending on if an individual has high blood pressure. Among the blood pressure groups, we see that, in both categories, individuals with high cholesterol have a much higher risk of developing heart disease. Individuals with both high blood pressure and high cholesterol are at the highest risk.

Finally, we turn our attention to figure 12, a line plot that depicts the likelihood of heart disease as a function of age and whether an individual has experienced a stroke, or was a smoker.

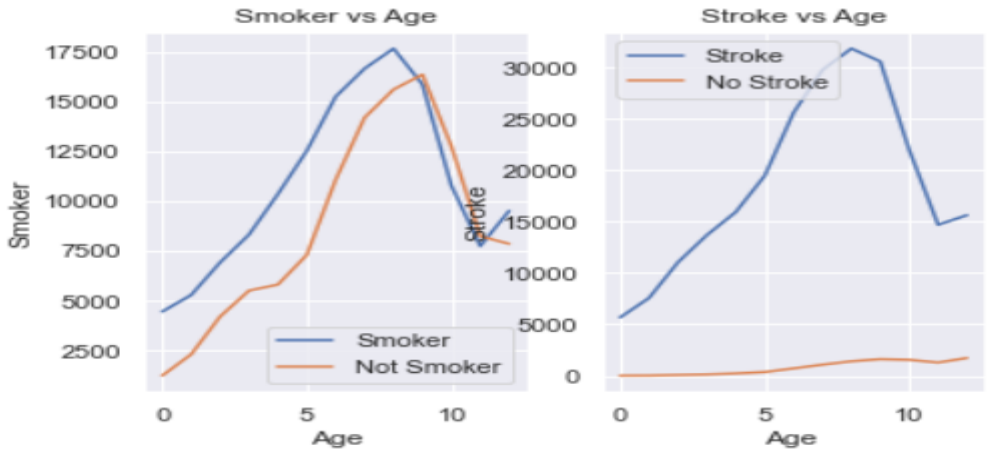


Figure 12: Side-by-side Line plots of Smoker as a function of age and Stroke as a function of Age. Plot can be found in 'EDA and Outline.ipynb',

Here, we can see that both smoking and having a stroke can elevate an individual's risk of heart disease. Also, we can see the age effect here again: regardless of smoking or stroke status, the number of individuals who experience heart disease increases as the age bracket increases.

2.3 Mutual Conditional Entropy

To determine associations among the variables, we examined multiple contingency tables and calculated their respective mutual conditional entropies. Mutual conditional entropy (MCE) is a measure of association calculated from a contingency table. Unlike correlation, association calculated from MCE does not imply linearity, but implies a general association that may be linear or non-linear in nature. The table below describes the results of the MCE's calculated from various multiple contingency tables. The code

and contingency tables for this project can be found in the mult contingency tables notebook within the zipped folder.

For the scope of this project, we chose to focus on the variables that we believed to be statistically significant, based on our findings from the plots in section 2.2. Table 1, below, describes the pair of variables that were compared to the response variable, and the associated mutual conditional entropy.

Variable Pair	**Mutual Conditional Entropy**
High Blood Pressure and High Cholesterol	0.9031
High Blood Pressure and Diabetes	0.7785
High Blood Pressure and Difficulty Walking	0.7603
High Blood Pressure and Physical Activity Within 30 days	0.8418
High Blood Pressure and Smoking	0.9340
High Blood Pressure and Stroke	0.5777

Table 1: Mutual conditional entropies of different variable pairings

In correlation analysis, a correlation coefficient between two variables of 0.7 or higher is said to be "highly Correlated". Using this same criteria for mutual conditional entropy, we can see from table 1 that all the variable pairs, except for the high blood pressure and stroke combination, are highly associated with developing heart disease.

2.4 Hierarchical Clustering Diagrams

Hierarchical clustering is a agglomerative (bottom-up) clustering method in which the general structure of data points can be observed from a distribution-free perspective. It observes the closeness between data points based on the L2 norm of k continuous variables in $\mathbb{R}^k$ space.

Procedure of Hierarchical Clustering:

- Take random subsets containing 10 percent of total observations (25,368 obs per subset) for computational purposes.
- Assign data points to clusters using using L2 Minkowski distance under an 'average-linkage', based on the 3 continuous variables in our data: BMI, Mental Health, and Physical Health. (average linkage is referred to as the smallest distance between any centroid to the centroid of a different cluster.)
- Clusters with the smallest L2 distance from each other are combined to form bigger clusters. This process is repeated until all clusters are connected.

Below shows our results from two random subsets.

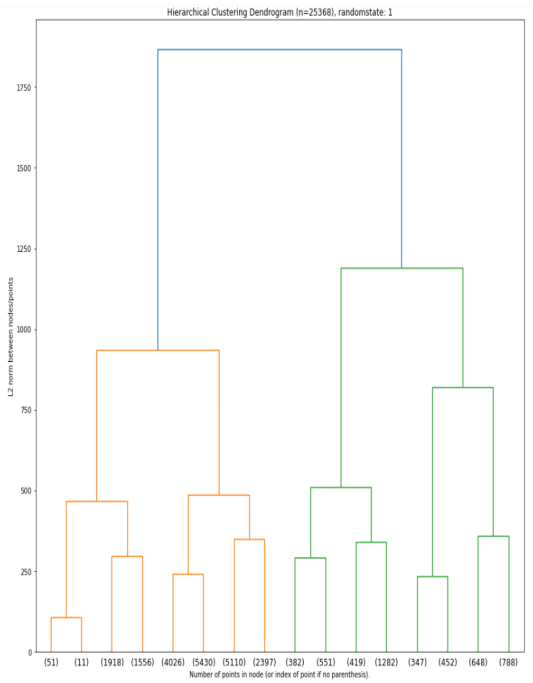


Figure 13: Dendrogram based on a 10 percent random subset. Plot can be found in Section 3 of 'EDA and Outline.ipynb'

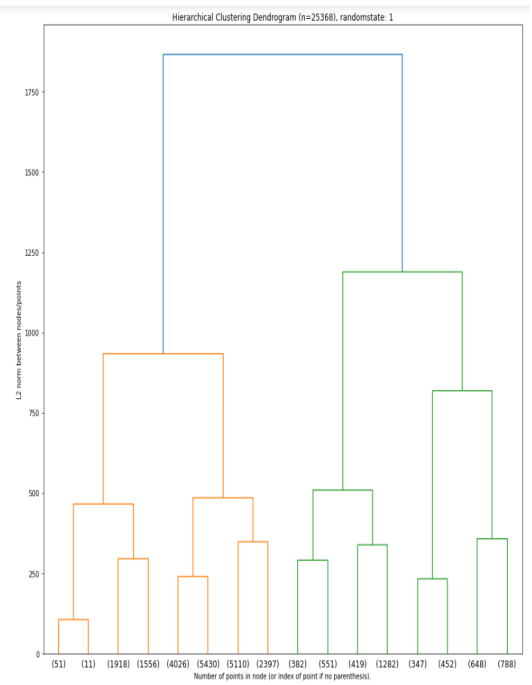


Figure 14: Dendrogram based on a 10 percent random subset. Plot can be found in Section 3 of 'EDA and Outline.ipynb'

Interpretation of Hierarchical Clusters:

- We observed that the highest level of distances are strictly between clusters with a significant proportion of datapoints: there appears to be no outlying observations.
- There exists a large L2-norm distance between these last two cluster's centroids: more so than between each mini-cluster and individual observations. This observation shows that there may exist aggregate hierarchical structure between heart disease and no heart disease. I.e. The two largest structures are highly different since the distance between the highest clave and all other claves is large.
- Note that the final two clusters are simply two emergent structures from our data- they may not be associated with our response classes: heart disease and no heart disease.

3 Statistical Analysis

Next, we perform appropriate statistical tests to formally determine if our findings from section 2 hold true in the population. This will allow us to confirm if, based on the sample data, the variables we believe to be highly associated with heart disease indeed are. In doing so, we can determine if using these variables to predict heart disease is meaningful. This would make our model much more applicable to the population, as well as more interpretable by healthcare professionals.

3.1 K-sample test

Our decision to conduct a K-sample test is motivated the curiosity of whether levels of age has an effect on the affliction of heart-disease. To this end, we selected the large sample approximation to the Kruskal-Wallis test to avoid the normality and linearity assumptions that are needed in ANOVA; it is a non-parametric method to compare whether k-groups are significantly different. We determined that

a large-sample approximation was applicable because there are at least 5000 observations per age group. Our pre-hoc analysis investigates whether there is a significant difference between all 13 age groups and our post-hoc analysis focuses on which age groups are significantly different, given that there is a significant difference between at least one pair of age groups.
For our pre-hoc analysis, we had the following hypotheses:

$$H_0 : F_1(x) = F_2(x) = ... = F_{12}(x) = F_{13}(x)$$

$$H_A : F_i(x) \geq F_j(x) or F_i(x) \geq F_j(x), i \neq j$$

where $F_i(x)$ represents the cumulative distribution function of the ith age group with respect to the heart-disease response variable.
Under a large-sample approximation, our Kruskal-Wallis test statistic is distributed under a chi-squared distribution with 12 degrees of freedom (since there are 13 age groups), the appropriate test statistic is as follows:

$$KW_{obs} = \frac{1}{(S_R)^2} \sum_{i=1}^{k} n_i (\bar{R}_i - \frac{N+1}{2})^2 \sim \chi^2_{k-1}$$

where $(S_R)^2$: variance of ranks regardless of group, $\bar{R}_i$: mean rank of group i, $i = 1, ..., k$, n_i : sample size of group i, $i = 1, ..., k$, N: overall sample size
Based on our highly significant p-value ≈ 0, we reject H_0. Age appears to have a significant affect on whether a given person has heart disease. Based on these results we conduct a post-hoc test to identify which of the k groups are significantly different.
For our post-hoc analysis, we make pairwise comparisons between all pairs of age-groups to test if they have a significantly different effect on heart disease. We found the absolute value of pairwise differences between the mean rank of each age group, and compared them with the Bonferroni and Tukey's Honest Significant Difference Correction for the pairwise difference of those associated age groups. In other words, if $|\bar{Rank}_i - \bar{Rank}_j|$ >Bonferroni cutoff, then age group i and age group j are significantly different; and similarly for the Tukey's HSD cutoff value. These corrections are used to account for the large number of comparisons to be made.
Bonferroni correction is calculated as follows:

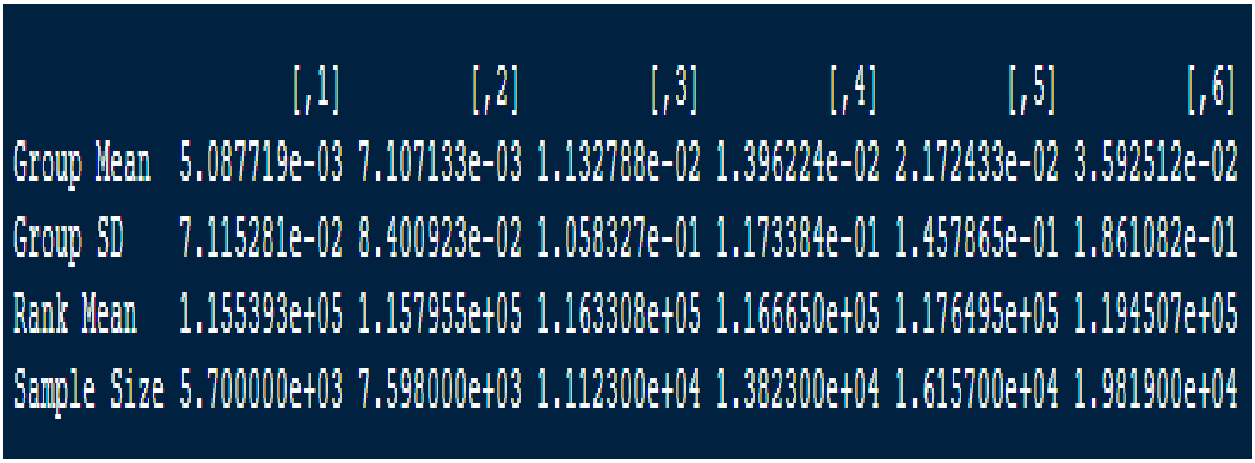
$$Z_{1-\frac{\alpha}{2g}} * \sqrt{(SR)^2 * (\frac{1}{n_i} + \frac{1}{n_j})}$$

Tukey's HSD correction is calculated as follows:

$$q(\alpha, k, df = N - k) \sqrt{\frac{(SR)^2}{2} * (\frac{1}{n_i} + \frac{1}{n_j})}$$

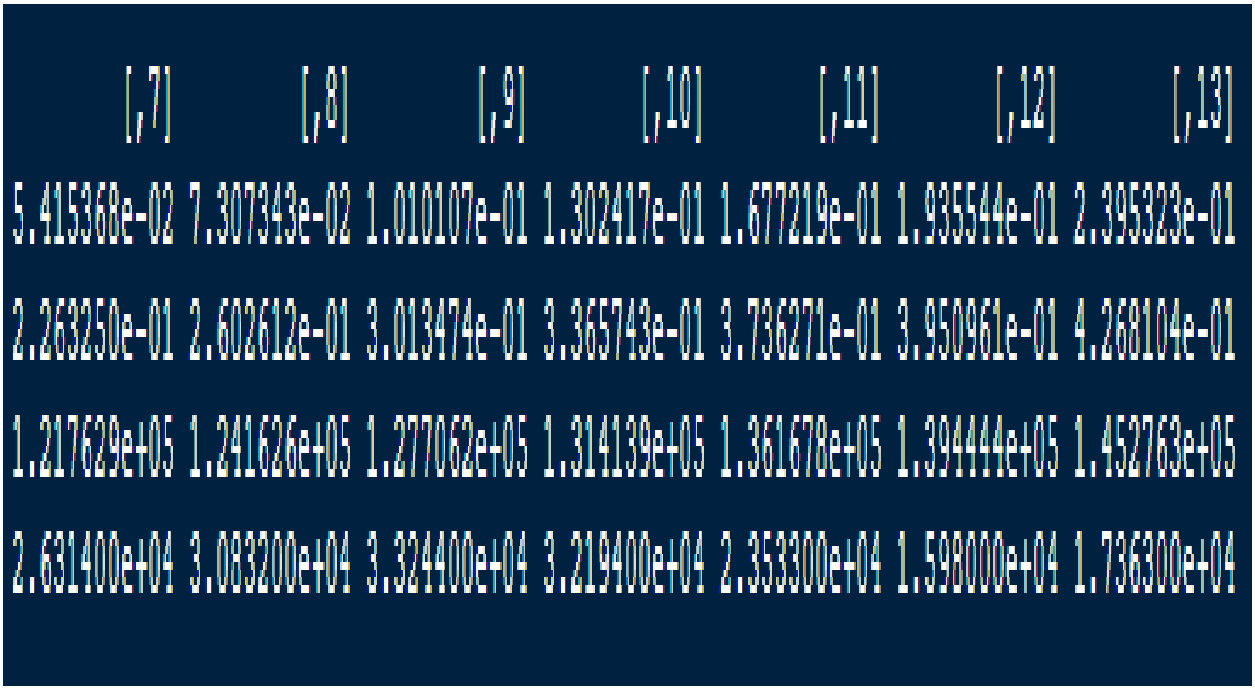
Note: Z() is the standard normal distribution and q() is the studentized range statistic.

Some summary statistics of each of our age groups is shown below:



	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
Group Mean	5.087719e-03	7.107133e-03	1.132788e-02	1.396224e-02	2.172433e-02	3.592512e-02
Group SD	7.115281e-02	8.400923e-02	1.058327e-01	1.173384e-01	1.457865e-01	1.861082e-01
Rank Mean	1.155393e+05	1.157955e+05	1.163308e+05	1.166650e+05	1.176495e+05	1.194507e+05
Sample Size	5.700000e+03	7.598000e+03	1.112300e+04	1.382300e+04	1.615700e+04	1.981900e+04

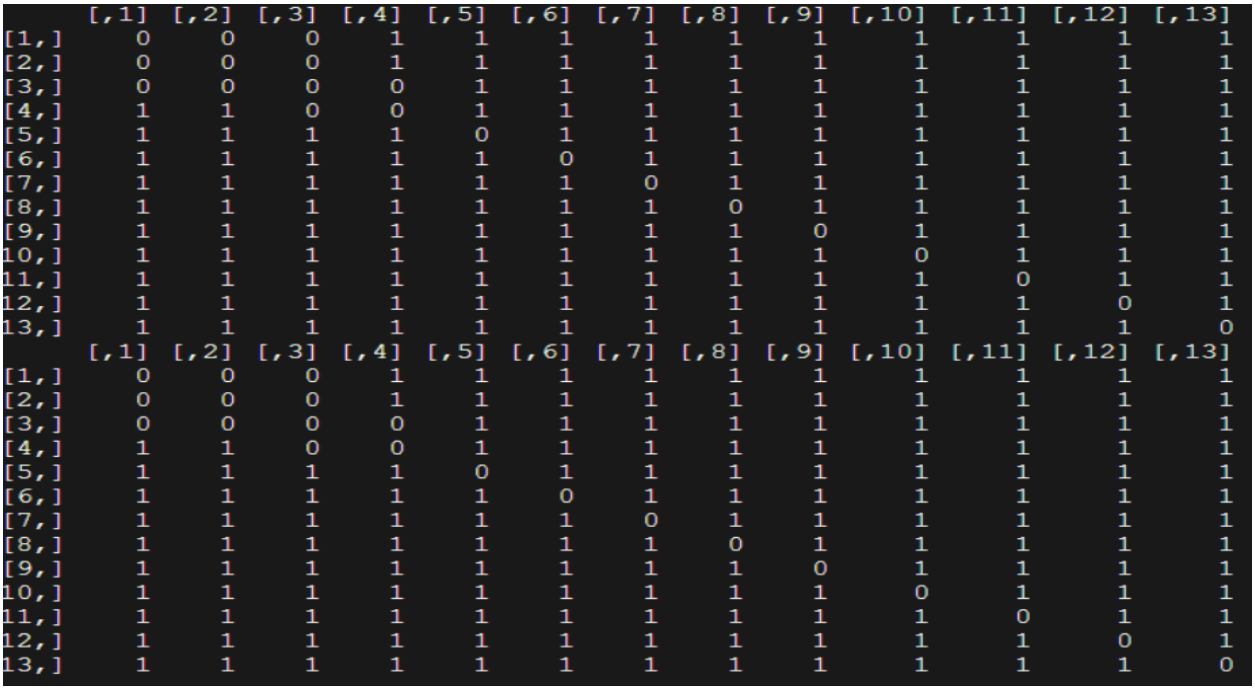
Figure 15: Summary statistics of age levels 1-6, see section 1 in K-groups Significant Difference.ipynb for more details.



[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]
5.415368e-02	7.307343e-02	1.010107e-01	1.302417e-01	1.677219e-01	1.935544e-01	2.395323e-01
2.263250e-01	2.602612e-01	3.013474e-01	3.365743e-01	3.736271e-01	3.950961e-01	4.268104e-01
1.217629e+05	1.241626e+05	1.277062e+05	1.314139e+05	1.361678e+05	1.394444e+05	1.452763e+05
2.631400e+04	3.083200e+04	3.324400e+04	3.219400e+04	2.353300e+04	1.598000e+04	1.736300e+04

Figure 16: Summary statistics of age levels 7-13, see section 1 in K-groups Significant Difference.ipynb for more details.

Our post-hoc analysis results are shown below. The first matrix is associated with the Bonferroni correction, and the second is associated with Tukey's HSD correction. Both matrices contains 13 rows and columns corresponding to the 13 age groups. Entries containing 1 describe age groups that have a absolute mean rank difference that is greater than the associated cutoff value type. i.e. if entry (1,2) of the first matrix is 1, then age group 1 and 2 are significantly different (as determined by the Bonferroni correction).



```
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13]
[1,]     0    0    0    1    1    1    1    1    1     1     1     1     1
[2,]     0    0    0    1    1    1    1    1    1     1     1     1     1
[3,]     0    0    0    0    1    1    1    1    1     1     1     1     1
[4,]     1    1    0    0    1    1    1    1    1     1     1     1     1
[5,]     1    1    1    1    0    1    1    1    1     1     1     1     1
[6,]     1    1    1    1    1    0    1    1    1     1     1     1     1
[7,]     1    1    1    1    1    1    0    1    1     1     1     1     1
[8,]     1    1    1    1    1    1    1    0    1     1     1     1     1
[9,]     1    1    1    1    1    1    1    1    0     1     1     1     1
10,]     1    1    1    1    1    1    1    1    1     0     1     1     1
11,]     1    1    1    1    1    1    1    1    1     1     0     1     1
12,]     1    1    1    1    1    1    1    1    1     1     1     0     1
13,]     1    1    1    1    1    1    1    1    1     1     1     1     0
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13]
[1,]     0    0    0    1    1    1    1    1    1     1     1     1     1
[2,]     0    0    0    1    1    1    1    1    1     1     1     1     1
[3,]     0    0    0    0    1    1    1    1    1     1     1     1     1
[4,]     1    1    0    0    1    1    1    1    1     1     1     1     1
[5,]     1    1    1    1    0    1    1    1    1     1     1     1     1
[6,]     1    1    1    1    1    0    1    1    1     1     1     1     1
[7,]     1    1    1    1    1    1    0    1    1     1     1     1     1
[8,]     1    1    1    1    1    1    1    0    1     1     1     1     1
[9,]     1    1    1    1    1    1    1    1    0     1     1     1     1
10,]     1    1    1    1    1    1    1    1    1     0     1     1     1
11,]     1    1    1    1    1    1    1    1    1     1     0     1     1
12,]     1    1    1    1    1    1    1    1    1     1     1     0     1
13,]     1    1    1    1    1    1    1    1    1     1     1     1     0
```

Figure 17: Significance of pairwise difference between age-groups associated with Bonferroni correction and Tukey's HSD correct respectfully. See section 2 in K-groups Significant Difference.ipynb for more details.

We find that the same groups are significantly different for both the Bonferroni cutoff and the Tukey's HSD cutoff value. Additionally, all pairwise differences including age groups 5-13 are significantly different. Age groups i = 1,2,3 are found to be mutually statistically insignificant. Age group 4 is not significantly different from 3, but is significantly different from 5. We suspect that these results imply that the number of people afflicted by heart-disease changes dramatically as one's age increases past age-group 4.

3.2 χ^2 Independent Test and Odds Ratio

Pearson χ^2 test can be used if there are more observations (e.g. lager than 10) for individual cells.

Hypotheses:
H_0: There is no association between the two variables.
H_a: There is an association between the two variables.

High blood pressure
The resulting p-value obtained from the Independent χ^2 test is very small, which means that there is stronger evidence in favor of the alternative hypothesis. In addition, the odd ratio obtained from the Table 2, it shows that people who have high blood pressure are more likely (4.592099 times) to have heart disease than people who do not have high blood pressure. Additionally, since the odd ratio is not equal to 1, it means that two variables are not independent, which is same as the results obtained by the Independence Test.

Table 2: HighBP vs Heart Disease contingency table

	No Heart Disease	Heart Disease
No High blood pressure	138886	5965
High blood pressure	90901	17928

High cholesterol

The resulting p-value obtained from the Independent χ^2 test is very small, which means that there is stronger evidence in favor of the alternative hypothesis. In addition, the odd ratio obtained from the Table 3, it shows that people who have high cholesterol are more likely (3.589073 times) to have heart disease than people who do not have high cholesterol. Additionally, since the odd ratio is not equal to 1, it means that two variables are not independent, which is same as the results obtained by Independent Test.

Table 3: HighChol vs Heart Disease contingency table

	No Heart Disease	Heart Disease
No High cholesterol	138949	7140
High cholesterol	90838	16753

Diabetes

The resulting p-value obtained from the Independent χ^2 test is very small, which means that there is stronger evidence in favor of the alternative hypothesis. In addition, the odd ratio obtained from the Table 4, it shows that people are less likely(0.04516371 times) to have heart disease when they do not have diabetes. From the odd ratio for Pre-Diabetes, it shows that people are more likely(1.627189 times) to have heart disease when they have pre-diabetes. From the odd ratio for Diabetes, it shows that people are more likely(3.623253 times) to have heart disease when they have diabetes.

Table 4: Diabetes vs Heart Disease contingency table

	No Heart Disease	Heart Disease
No Diabetes	198352	15351
Pre-Diabetes	3967	664
Diabetes	27468	7878

Difficulty walking

The resulting p-value obtained from the Independent χ^2 test is very small, which means that there is stronger evidence in favor of the alternative hypothesis. In addition, the odd ratio obtained from the Table 5, it shows that people who have serious difficulty walking are more likely (4.266085 times) to have heart disease than people who do not have serious difficulty walking. Additionally, since the odd ratio is not equal to 1, it means that two variables are not independent, which is same as the results obtained by Independent Test.

Table 5: Diffwalk vs Heart Disease contingency table

	No Heart Disease	Heart Disease
No Difficulty walking	197027	13978
Difficulty walking	32760	9915

Phys activity

The resulting p-value obtained from the Independent χ^2 test is very small, which means that there is stronger evidence in favor of the alternative hypothesis. In addition, the odd ratio obtained from the Table 6, it shows that people who doing physical activity or exercise during the past 30 days other than their regular job are less likely (0.5359804 times) to have heart disease than people who do not do the physical activity. Additionally, since the odd ratio is not equal to 1, it means that two variables are not independent, which is same as the results obtained by Independent Test.

Table 6: PhysAct vs Heart Disease contingency table

	No Heart Disease	Heart Disease
No physical activity	53167	8593
Physical activity	176620	15300

Smoking

The resulting p-value obtained from the Independent χ^2 test is very small, which means that there is stronger evidence in favor of the alternative hypothesis. In addition, the odd ratio obtained from the Table 7, it shows that people who have smoke are more likely (2.203943 times) to have heart disease than people who do not smoke. Additionally, since the odd ratio is not equal to 1, it means that two variables are not independent, which is same as the results obtained by Independent Test.

Table 7: Smoking vs Heart Disease contingency table

	No Heart Disease	Heart Disease
No smoking	132165	9092
Smoking	97622	14801

Stroke

The resulting p-value obtained from the Independent χ^2 test is very small, which means that there is stronger evidence in favor of the alternative hypothesis. In addition, the odd ratio obtained from the Table 8, it shows that people who have stroke are more likely (6.936202 times) to have heart disease than people who do not have stroke. Additionally, since the odd ratio is not equal to 1, it means that two variables are not independent, which is same as the results obtained by Independent Test.

Table 8: Stroke vs Heart Disease contingency table

	No Heart Disease	Heart Disease
No stroke	223432	19956
Stroke	6355	3937

Age

The resulting p-value obtained from the Independent χ^2 test is very small, which means that there is stronger evidence in favor of the alternative hypothesis. In addition, from Table 9 different age groups odd ratio, it is found that the odd ratio increases as age increases. Interestingly, it is also found that the odd ratio becomes larger than 1 after the age group 8. Therefore, people with the age over 60 will have higher chances to get the heart disease, especially those people who are over 80. On the other hands, the odd ratio tends to close to zero as age decreases. Therefore, the young people have less chances to get the heart disease.

Table 9: Age vs Heart Disease contingency table

	No Heart Disease	Heart Disease
Age group 1 (18-24)	5671	29
Age group 2 (25-30)	7544	54
Age group 3 (31-35)	10997	126
Age group 4 (36-40)	13630	193
Age group 5 (41-45)	15806	351
Age group 6 (46-50)	19107	712
Age group 7 (51-55)	24889	1425
Age group 8 (56-60)	28579	2253
Age group 9 (61-65)	29886	3358
Age group 10 (66-70)	28001	4193
Age group 11 (71-75)	19586	3947
Age group 12 (76-80)	12887	3093
Age group 13 (over 80)	13204	4159

Table 10: Different age groups odd ratio table

	Odd ratio
Age group 1 (18-24)	0.04802508
Age group 2 (25-30)	0.06673169
Age group 3 (31-35)	0.105475
Age group 4 (36-40)	0.1291464
Age group 5 (41-45)	0.2018445
Age group 6 (46-50)	0.3386715
Age group 7 (51-55)	0.5221325
Age group 8 (56-60)	0.7329972
Age group 9 (61-65)	1.093788
Age group 10 (66-70)	1.533826
Age group 11 (71-75)	2.123735
Age group 12 (76-80)	2.502789
Age group 13 (over 80)	3.456946

Overall Comparison
From the table 10, it shows that Stroke has the highest odd ratio, which is 6.936. This means that people who have stroke have the highest chance to get heart disease than people with other factors since the highest odd ratio is obtained in this group. In addition, the second high odd ratio occurs on the HighBP variable, which is 4.5921. Therefore, the people who have high blood pressure are more likely to get heart disease than others.
Since p-values are too small to show in the R, test statistics will be used to compare the test result. The test result shows that Age and Diff walk variables have the largest test statistics. Therefore, there is a strong evidence to conclude that these two variables are not independent from heart disease. They are strong associated with heart disease.

Table 11: Comparison table

	HighBP	HighChol	Diabetes	Diff Walk	Phys Act	Smoking	Stroke	Age
Test Statistics	11119.3	8289.27	8244.889	11477.75	1933.324 (smallest)	3322.39	10454.1	13731.04 (largest)
Odd Ratio	4.5921	3.58907	0: 0.04516 1: 1.62719 2: 3.62325	4.266085	0.53598	2.20394	6.936	1: 0.0480251 13:3.45695
Disease Chance	High BP more	High Chol more	Diabetes more No diabetes less	Diff walk more	Physical activity less	Smoking more	Stroke more	Age group 1 less Age group 13 more

4 Heart Disease Prediction Model

Data Pre-processing: First, the data set is arranged into feature (X) variables and the target (Y) variable by indexing from the data frame. Then, we split the data using the `train_test_split` function into two. Training contains 80 percent of randomly split data, and the remaining 20 percent is stored as test data. Every machine learning algorithm assumes the data is normalized/scaled at equal measurement since the variables can have different data types that need to be scaled equally. Thus, we standardize the features in both the training and testing data by making the mean equal to zero and the variance equal to one. This is achieved using the function `StandardScaler` in the sklearn Python library. This approach of

standardizing the data improves the accuracy score which we will see in the later results for the applied classification models. Following, we use the `fit_transform` function to find the parameters scaled and transform them into the required shape.

Classification Models: This is a technique applied to both structured and unstructured data that identifies and sorts the data points according to their given category which is either heart disease or not. This method first initializes a classifier to map input data or patient X to its relevant category Y. There are different classifiers that we will use to compare, and they are covered in the following sections. Using this classifier then we fit the model by training the given X and Y from the training dataset. Given the unlabeled observation X, the model predicts the target by returning the predicted Y. Finally, the model is evaluated using the new testing data to ensure the performance of the model.

Cross-Validation: In this method, the goal is to prevent the over-fitting of our predictive models. This over-fitting means if the optimal model is tested on a different dataset, we want to ensure the computation works well. This method works by first setting a fixed fold, for instance, we select ten folds to our data. This splits the data into ten folds where nine are training and one is testing. We run the analysis in the procedure iteratively until each fold forms a one testing set. Then, we return the average of the overall error estimate from this testing data.

Linear Discriminant Analysis: This method uses Bayes theorem to estimate the probability by finding the linear combination of all input features that are separated according to which class of Y they belong. If the output is either a patient with heart disease or a patient without heart disease and the input X is one of the features, The output will then return the highest probability of each class label which helps to predict. The method ensures all assumptions of normality, independence, and collinearity are met to estimate the mean and variance of the data.

Logistic Regression Analysis: This method uses a logistic function to distinguish the binary outcome, whether patients have heart disease or not, by analyzing the relationship between the independent features. The probability of each outcome is computed by calculating the weighted sum of input features and then estimating the probability of every patient that belongs to a specific class label in Y. It assumes the data is normally distributed, observations are independent of each other, features should not have multicollinearity, and absence of extreme outliers.

Decision Tree: This method is like a flowchart with a tree structure that displays possible solutions given certain conditions. So that it classifies the input features into their relevant class labels, it produces a sequence of conditions starting at the root of the tree continuing to the branch of the tree ending in the leaf of the tree. The root contains the important features, the branch contains several decisions, and the leaf node holds the outcome decision. produces a sequence of conditions to classify the input features into their relevant class labels. For this model, we set the parameter of the classifier to entropy because we want to measure the level of uncertainty of the random variables.

Support Vector Machine: This method creates a decision boundary using a hyperplane, a plane in 3-dimension space, that segregates the input features into their relevant classes by putting the observation of each patient into the right category. The support vector classifier is when an observation of a patient that can be from either of the classes is closer to the hyperplane, then we would choose the hyperplane that maximizes the margin between classes.

Naïve Bayes: This method uses Bayes theorem to get the estimated probability of the class labels from the data. Given the observation of patients from the input features, the probability of the class label is multiplied by the joint conditional distribution of the features given by the label. Then, the product term is divided by the predictor prior to distribution. The method assumes the data are independent.

Results: For the given data set, we compare the different algorithms using a cross-validation technique to find the optimal algorithm. Given the results of each algorithm score of the unseen data, we selected both Logistic Regression and Linear Discriminant Analysis to train the model.

Table 12: Model Evaluation Results		
No	Classification Type	Accuracy Score
1	Decision Tree	90%
2	Logistic Regression	90%
3	Support Vector Machine	90%
4	Linear Discriminant Analysis	89%
5	kNN Neighbor	88%
6	Naive Bayes	87%

We would further compare LDA and logistic classifiers to access their performance using the metrics confusion matrices, classification reports, and ROC curves.

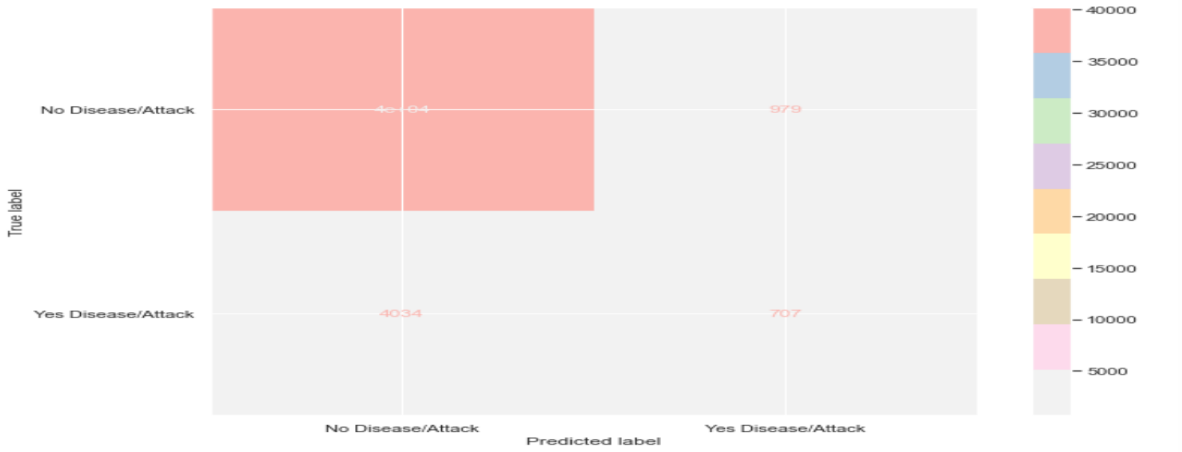


Figure 18: Confusion Matrix for Linear Discriminant

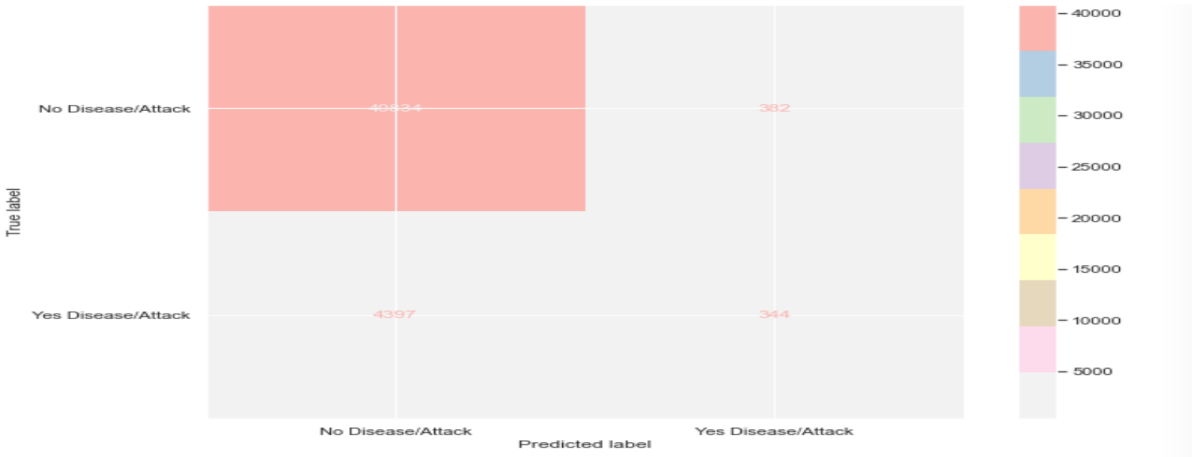


Figure 19: Confusion Matrix for Logistic Regression

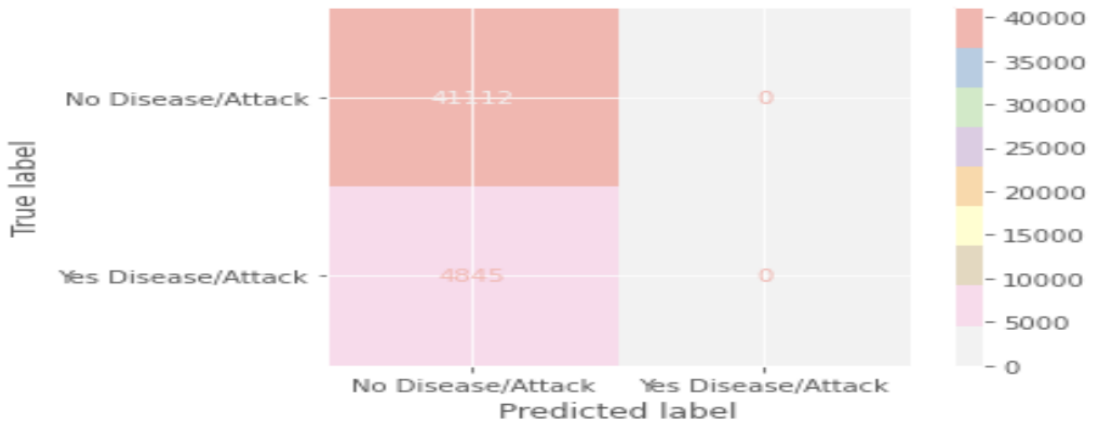


Figure 20: Confusion Matrix for Decision Tree

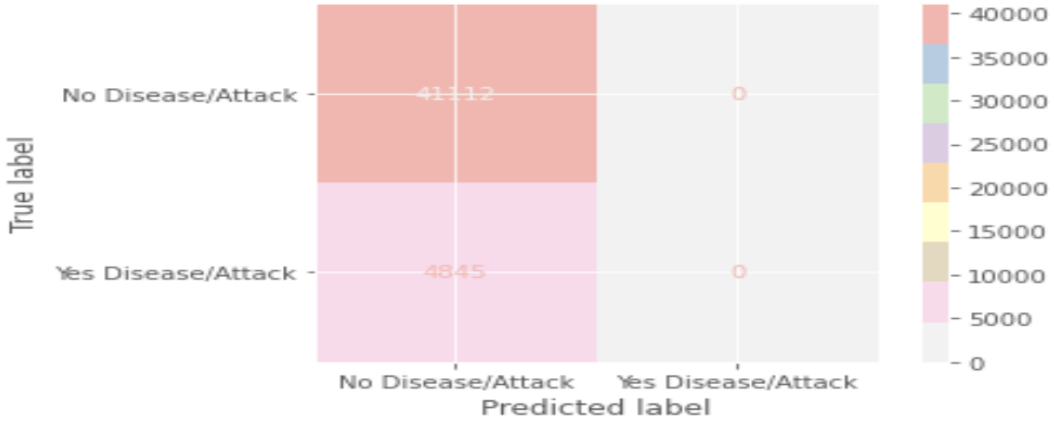


Figure 21: Confusion Matrix for Support Vector

The contingency tables illustrate the proportion of how the classifiers were able to correctly and incorrectly classify the data according to the labels. This contingency table is presented in a confusion matrix where it includes both the information of observed and unseen or predicted values of our data. This is achieved by the `confusion_matrix''` function in SKLEARN library in Python.
The results indicate that both classifiers Linear Discriminant Analysis and Logistic Regression have higher true negative values that represent our model and can correctly classify negative observed and predicted values as actually negative values. Whereas, both Decision Tree and Support Vector Machine have 0's for both False Negative and True Negative. This indicates class in balance problem, when the number of sample in a specific class is significantly higher than the sample in the other class. This can be resolved by either training multiple models or by performing an under sampling or oversampling techniques.

	precision	recall	f1-score	support
No Disease/Attack	0.91	0.98	0.94	41216
Yes Disease/Attack	0.42	0.15	0.22	4741
accuracy			0.89	45957
macro avg	0.66	0.56	0.58	45957
weighted avg	0.86	0.89	0.87	45957

Computation running time in Linear Discriminant Classifier: 0.02 secs.

Figure 22: Classification Report for Linear Discriminant Analysis

	precision	recall	f1-score	support
No Disease/Attack	0.90	0.99	0.94	41216
Yes Disease/Attack	0.47	0.07	0.13	4741
accuracy			0.90	45957
macro avg	0.69	0.53	0.54	45957
weighted avg	0.86	0.90	0.86	45957

Computation running time in Logistic Regression: 0.02 secs.

Figure 23: Classification Report for Logistic Regression

	precision	recall	f1-score	support
No Disease/Attack	0.89	1.00	0.94	41112
Yes Disease/Attack	0.00	0.00	0.00	4845
accuracy			0.89	45957
macro avg	0.45	0.50	0.47	45957
weighted avg	0.80	0.89	0.84	45957

Figure 24: Classification Report for Decision Tree

	precision	recall	f1-score	support
No Disease/Attack	0.89	1.00	0.94	41112
Yes Disease/Attack	0.00	0.00	0.00	4845
accuracy			0.89	45957
macro avg	0.45	0.50	0.47	45957
weighted avg	0.80	0.89	0.84	45957

Figure 25: Classification Report for Support Vector

We use the function `classification_report` from Scikit-Learn's library in Python. In this classification metric, we focus on the recall that measures how well the models are able to classify the positive observations as actually positive. It is computed by taking the sum of true positive divided by the sum of both true positive and false negative values. Precision measures how well the models are able to predict positive values that are actually positive.

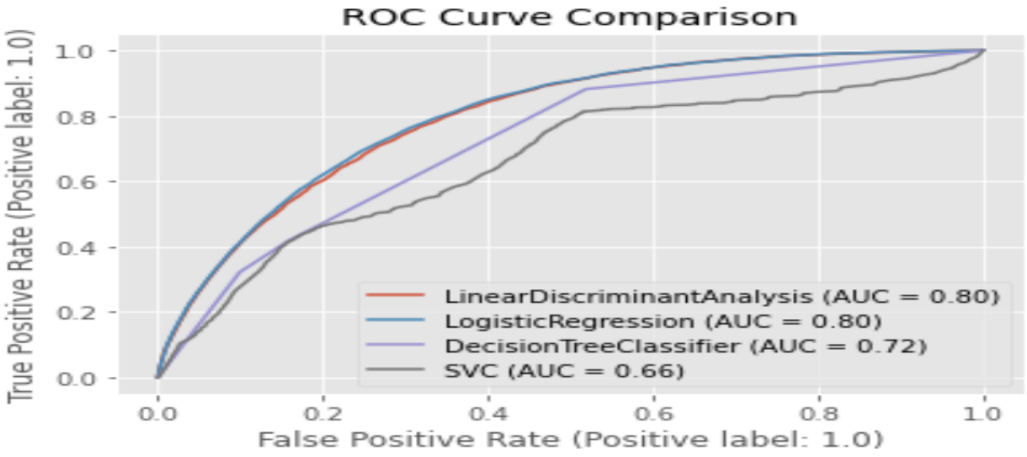


Figure 26: ROC Curve of Selected Classifiers

To get a thorough comparison between the model, we will calculate the area under the curve (AUC) and plot the ROC curve after we train the models. The model in logistic regression has been correctly classified at the rate of 0.90. And, the model in the decision tree classifier has correctly classified at the rate of 0.90 as well. To distinguish which is a better predictive model, we adjust the threshold for classifying an observation to the positive class label.

5 Conclusion

This report detailed our methodology into exploring some of the underlying predictors of heart disease. We began by examining correlation and association between each predictor and the response variable. Through our exploratory data analysis, we were able to identify several potentially significant predictor variables. We then proceeded to establish the statistical significance of these variables by utilizing a non-parametric k-samples test to establish significant differences among groups within each variable, and a

chi-square test for independence to establish whether each predictor is related to the response.

Having established the statistically significant variables, we used these in a variety of machine learning models to make predictions on whether an individual would be susceptible to heart disease. On the basis of validation accuracy, we can conclude that Logistic Regression Classifier is the optimal model to be used for this data set, after comparing the accuracy score and AUC score of the three classification models.

References

[1] Centers for Disease Control and Prevention. (2022, January 13). FASTSTATS - leading causes of death. Centers for Disease Control and Prevention. Retrieved from https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm

[2] Garg,R. (2021, October 7). 7 types of classification algorithms. Analytics India Magazine. Retrieved May 6, 2022, from https://analyticsindiamag.com/7-types-classification-algorithms/

[3] Mishra, R. (2021, May 22). Heart failure prediction in python! Medium. Retrieved May 6, 2022, from https://medium.com/analytics-vidhya/heart-failure-prediction-in-python-70ce2a033a18

[4] Wicker, E. (2021, February 7). Linear discriminant analysis 2. Ethan Wicker. Retrieved May 6, 2022, from https://ethanwicker.com/2021-02-07-linear-discriminant-analysis-002/