Evaluating Multicovariate Categorical Predictors of Heart Disease via Re-Co Dynamics

STA 160 Midterm Project Final

Vincent Barletta, Sid Das, Martin Topacio, Felix Yan

2024-05-13

Introduction

Heart disease, the leading cause of death for people in the world, is a condition that carries many risk factors. With such a deadly disease causing about 1 in every 5 deaths, it would be beneficial to research the factors involved with such a condition in order to draw associations between certain variables of life and having heart disease. Related to heart disease as a risk factor and having shared risk factors is a stroke, occurring when blood supply to the brain is blocked or reduced. Another related health condition is diabetes, a disease in which the pancreas does not produce enough insulin for the body or when the body cannot use the insulin it produces. Diabetes caused 1.5 million deaths in 2019, with almost half of the deaths occurring in people under 70. With such widespread effect on the world at large, investigating how and why people have these conditions proves to be necessary. These three conditions each have ties to lifestyle and health-related factors, which we would like to explore. As such, from the BRFSS heart disease dataset from 2015, we aim to investigate lifestyle and health-related variables like smoking, heavy alcohol consumption, having diabetes, being physically active, etc. to see if they have relations to having heart disease, having a stroke, or having diabetes. In our wider investigation, there are three covariates from the dataset that we will heavily focus on in our project. We will investigate all defined variables using odds ratio tables to get an idea of how likely someone is to have heart disease, a stroke, or diabetes. We plan to choose the top three covariates with the highest likelihood of having one of the three conditions, then diving into how categories of our covariates, as well as combinations of the categories affect having a fused response variable of heart disease, a stroke, and diabetes. To achieve our investigation, the methods of the aforementioned odds ratios, entropy, and mutual information will be used on the data. By the end of this project, the objective is to determine how particular factors in health and lifestyle can contribute to and affect a person having heart disease, a stroke, or diabetes.

Data Cleaning and Summaries

```
##       Age              BMI           PhysHlth         GenHlth     
##  Min.   : 1.000   Min.   :12.00   Min.   : 0.000   Min.   :1.000  
##  1st Qu.: 6.000   1st Qu.:24.00   1st Qu.: 0.000   1st Qu.:2.000  
##  Median : 8.000   Median :27.00   Median : 0.000   Median :2.000  
##  Mean   : 8.032   Mean   :28.38   Mean   : 4.242   Mean   :2.511  
##  3rd Qu.:10.000   3rd Qu.:31.00   3rd Qu.: 3.000   3rd Qu.:3.000  
##  Max.   :13.000   Max.   :98.00   Max.   :30.000   Max.   :5.000  
##     MentHlth        Education        Income     
##  Min.   : 0.000   Min.   :1.00   Min.   :1.000  
##  1st Qu.: 0.000   1st Qu.:4.00   1st Qu.:5.000  
##  Median : 0.000   Median :5.00   Median :7.000
##  Mean   : 3.185   Mean   :5.05   Mean   :6.054
##  3rd Qu.: 2.000   3rd Qu.:6.00   3rd Qu.:8.000
##  Max.   :30.000   Max.   :6.00   Max.   :8.000
```

Our dataset has 22 different variables, with 253,680 different observations. After a simple diagnostics check for missing data, we see that no rows have missing or NA data. Therefore, no dataset cleaning will be required. As for our variables, they follow the listed criteria. The vast majority of our variables, and the primary focus for finding predictors of heart disease, will revolve around our categorical variables.

HeartDiseaseorAttack: Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI)

Gender: female: 0 and male: 1.

High Blood Pressure: Adults who have been told they have high blood pressure by a doctor, nurse, or other health professional.

High Cholesterol: Have you EVER been told by a doctor, nurse or other health professional that your blood cholesterol is high?

CholCheck: Cholesterol check within past five years.

GenHlth: an ordinal variable. 1-5 where 1 is excellent and 5 is poor. Age: convert to 14 level age buckets. 1 Age 18 to 24, 2 Age 25 to 29, 3 Age 30 to 34, 4 Age 35 to 39, 5 Age 40 to 44, 6 Age 45 to 49, 7 Age 50 to 54, 8 Age 55 to 59 9 Age 60 to 64, 10 Age 65 to 69, 11 Age 70 to 74, 12 Age 75 to 79, 13 Age 80 or older

Diabetes: 0 is for no diabetes or only during pregnancy, 1 is for pre-diabetes or borderline diabetes, 2 is for yes diabetes.

GeneralHealth: Broken up into five bins from 1 to 5, with 5 being the best general health level. Self-reported data.

MentHlth: Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?

PhysHlth: Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?

Age: 14-level age category

BMI: Patient's body mass index.

DiffWalk: Do you have serious difficulty walking or climbing stairs?

Education: ordinal variable with 1 being never attended school or kindergarten only up to 6 being college 4 years or more.

Income: Variable is already ordinal with 1 being less than $10000 all the way up to 8 being $75,000 or more.

Reference: https://www.kaggle.com/code/alexteboul/heart-disease-health-indicators-dataset-notebook

Methodology Part 1: Odds Ratio Calculations

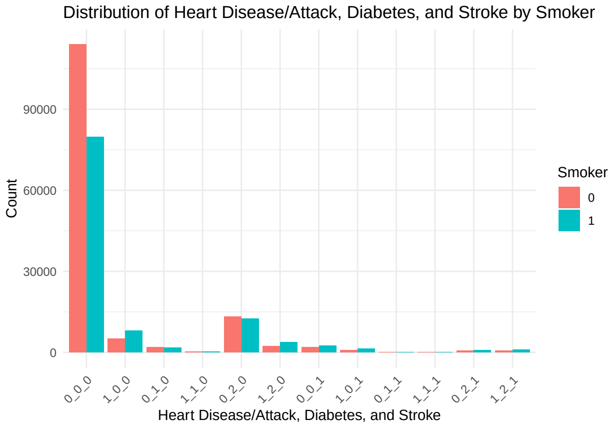
We will use two packages, epitools and InzightTools that allow us to easily calcuate the odds ratio for each of our categorical variables.

We want to use a combination of Stroke, Heart Disease, and Diabetes as our response variable to see how other variables have an affect. As stated above, when Stroke andHeart Disease =1, they are present within the subject. For Diabetes, 0 indicates no diabetes, 1 indicates pre-diabetes, and 2 indicates diabetes present. Therefore, our most unhealthy patients will have a (1,2,1) combination. We would expect these patients to also have multiple other health defects like High Blood Pressure, Difficulty Walking, High Cholesterol, etc. Let's investigate to see how these correlate. We will first begin by looking at our covariates individually. After identifying some of the highest correlated variables for our unhealthy patients, we will then look at their joint effects to find the best determinants for patients with HeartDisease, Diabetes, and Strokes.

We will first begin by looking at how smoking affects a persons odds of contracting HD, Diabetes, or a Stroke.

Smoker

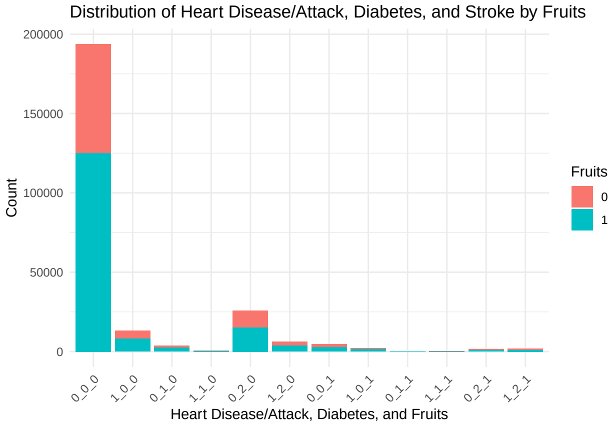
```
##                                      odds ratio with 95% C.I.
## HeartDiseaseorAttack_Diabetes_Stroke estimate    lower    upper
##                                0_0_0 1.000000       NA       NA
##                                1_0_0 2.229999 2.150833 2.312305
##                                0_1_0 1.236379 1.159261 1.318609
##                                1_1_0 2.275665 1.921726 2.700696
##                                0_2_0 1.338146 1.303756 1.373426
##                                1_2_0 2.351452 2.232630 2.477105
##                                0_0_1 1.805603 1.701996 1.915766
##                                1_0_1 2.661660 2.437986 2.907935
##                                0_1_1 1.915938 1.407631 2.619904
##                                1_1_1 4.558718 2.926920 7.369590
##                                0_2_1 1.708098 1.549064 1.883988
##                                1_2_1 2.517333 2.275939 2.786811
```



Here, we attempt to emulate the odds-ratio table calculation performed in section. However, attempting to calculate our combination variable on the x-axis yields an error during the Fisher calculation, so we will instead utilize our Response Variable on the y-axis. Our best takeaway here is that patients with HeartDiseaseorAttack, Stroke, and Diabetes have between (2.93, 7.37) odds of being a Smoker at a 95% confidence interval.

Due to the large output of including each confidence interval table with each variable, we cannot display each individual one; instead, we will report important statistics and discuss the overall averages at the end of this section.

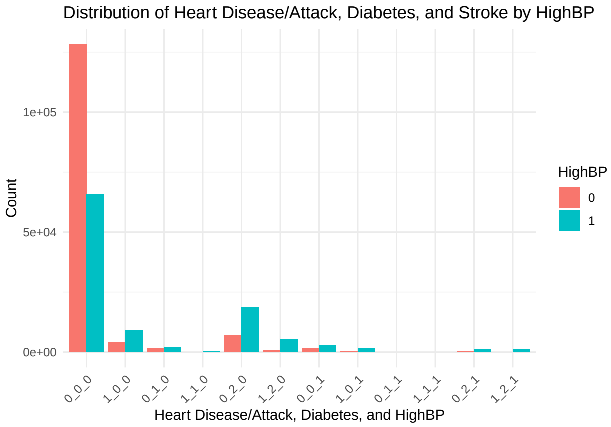
Fruits



We can see here that eating fruits or not eating fruits has very little effect on a patient contracting HeartDisease, Diabetes, or a Stroke.

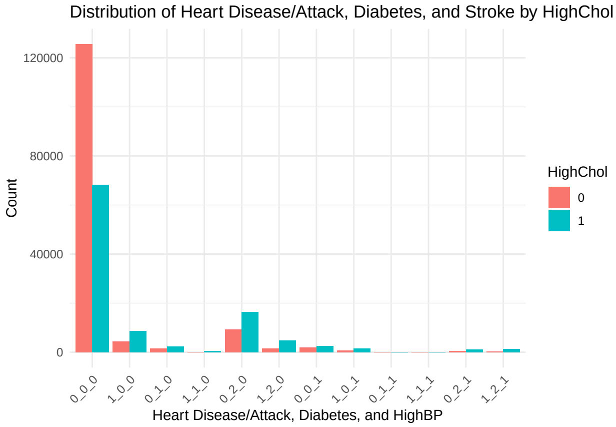
High Blood Pressure

```
##                                      odds ratio with 95% C.I.
## HeartDiseaseorAttack_Diabetes_Stroke  estimate     lower     upper
##                                0_0_0  1.000000        NA        NA
##                                1_0_0  4.308803  4.148160  4.476341
##                                0_1_0  2.878575  2.696548  3.073689
##                                1_1_0  6.977980  5.732251  8.558445
##                                0_2_0  4.994024  4.852900  5.139046
##                                1_2_0 10.197544  9.529941 10.920002
##                                0_0_1  3.583194  3.370132  3.811193
##                                1_0_1  6.243911  5.662440  6.896933
##                                0_1_1  6.239993  4.397834  9.057486
##                                1_1_1  8.371025  5.194820 14.220213
##                                0_2_1  9.343041  8.229067 10.646320
##                                1_2_1 14.418328 12.445687 16.797811
```



High Blood Pressure has a demonstrable effect for each of these conditions. The 95% confidence interval for having Stroke/HD/Diabetes and HighBP is from 12.45 to 16.80, meaning that having HighBP makes you 12.45x to 16.8x more likely to contract all three of these conditions. In addition, having Diabetes and HeartDisease makes you 9.53 to 10.92 times more likely to get HighBP.

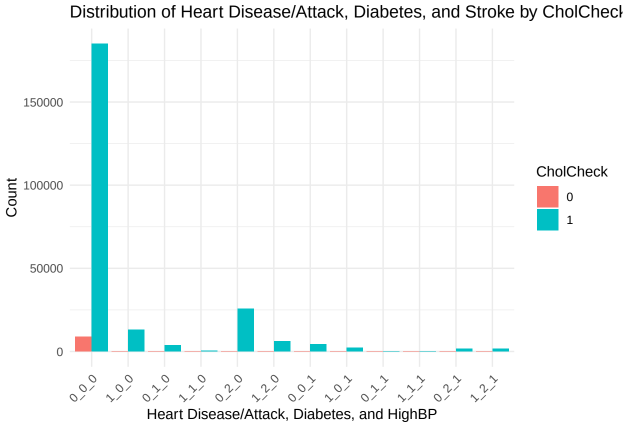
High Cholesterol



HighCholesterol also has a large effect on increasing likelihood of HD, Diabetes, and Stroke. Those that have HD, Pre-Diabetes, and Stroke have a 4.16x to 10.79x increased chance of having high cholesterol. Also, having HighChol makes you 5.92 to 7.50 more likely to have HD, Diabetes, and Stroke.

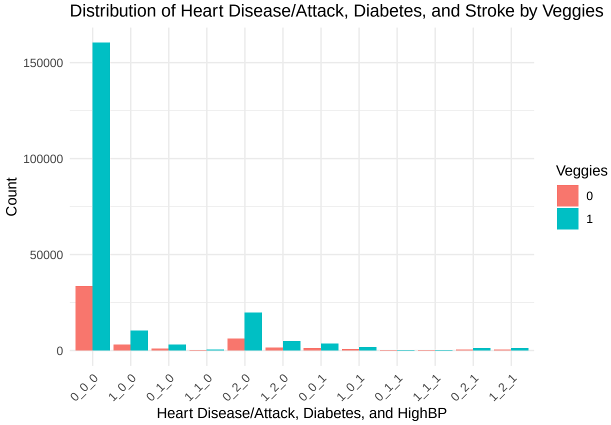
Cholesterol Check

```
##                                      odds ratio with 95% C.I.
## HeartDiseaseorAttack_Diabetes_Stroke estimate     lower     upper
##                              0_0_0 1.000000        NA        NA
##                              1_0_0 3.530403 3.0482252  4.118039
##                              0_1_0 3.586037 2.7432757  4.808212
##                              1_1_0 4.333329 2.1268590 11.019855
##                              0_2_0 6.787786 5.8751065  7.896256
##                              1_2_0 7.791784 5.7499774 10.918147
##                              0_0_1 2.297776 1.8811857  2.844345
##                              1_0_1 3.038987 2.2022098  4.352579
##                              0_1_1 2.445072 0.9307044 10.211963
##                              1_1_1 1.489278 0.5610262  6.251923
##                              0_2_1 9.535933 5.1111531 20.987502
##                              1_2_1 5.489172 3.3795551  9.780465
```



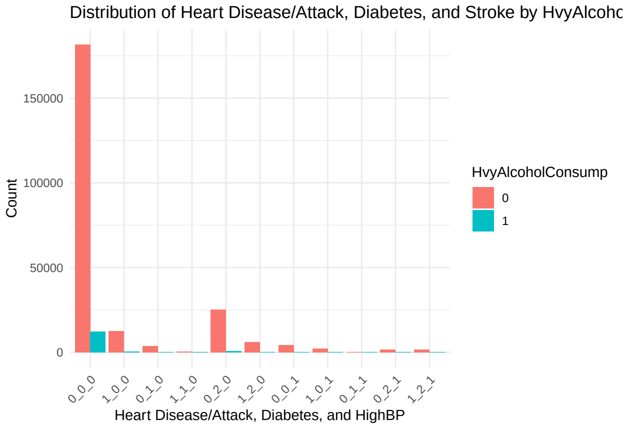
If you have not had your Cholesterol Checked in the past year, you have an estimated 9.54x, with a (5.11, 20.987) 95% confidence interval, increased chance of having Diabetes and a Stroke. Not having cholesterol checked in the past year also yields an estimated 5.48x, with a (3.38, 9.78) 95% confidence interval, increased chance of having heart disease, diabetes, and a stroke.

Veggies



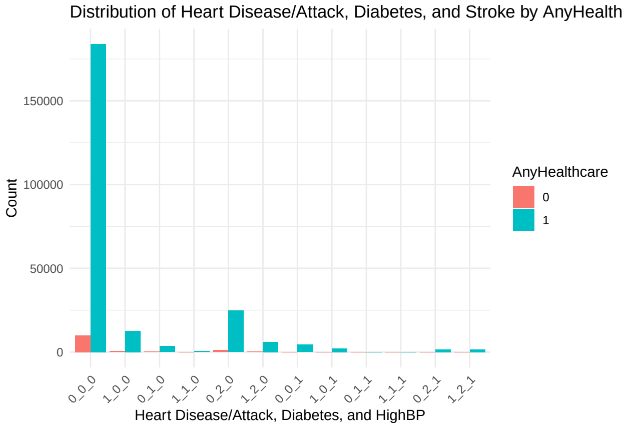
As we can see from the odds ratio table, every estimate where at least one disability is present yields a value less than 1, with every lower bound and every upper bound except for (1,1,1) also having values less than 1. This indicates that there is a weak correlation and whether or not someone eats at least 1 vegetable a day or not has very little effect on someone having heart disease, diabetes, or a stroke.

Heavy Alcohol Consumption



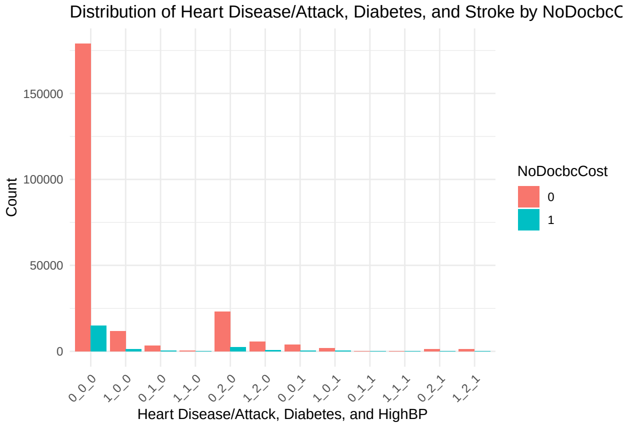
Similarly to the fruits odds ratio table for our fused response variable, there is a weak correlation between heavy alcohol consumption and someone having heart disease, diabetes, or a stroke because almost all values in the odds ratio table for heavy alcohol consumption are less than 1.

Any Healthcare



Having access to any kind of healthcare yields a 1.42x increased chance of having diabetes and a stroke, with a 95% confidence interval of (1.11, 1.86). Further, this covariate yields a 1.21x increased chance of having all three conditions, with a 95% confidence interval of (0.96, 1.56).

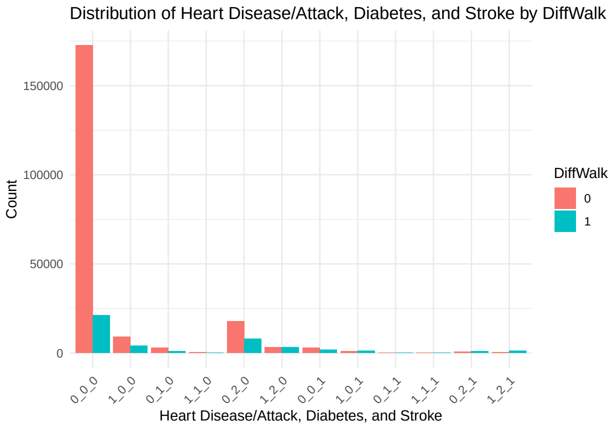
No Doctor Because of Cost



Being financially able to visit the doctor has a 1.6x increased likelihood of having diabetes and a stroke, with a 95% confidence interval of (1.37, 1.86). Further, being financially able to visit the doctor has a 2.31x increased chance of having all three conditions, with a 95% CI of (2.01, 2.63).

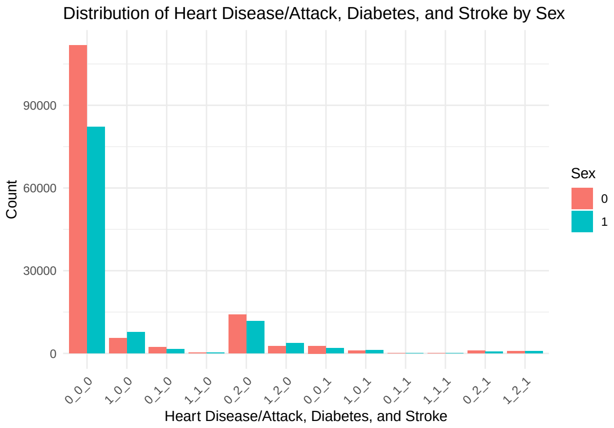
Difficulty Walking

```
##                                      odds ratio with 95% C.I.
## HeartDiseaseorAttack_Diabetes_Stroke  estimate     lower     upper
##                                0_0_0  1.000000        NA        NA
##                                1_0_0  3.758225  3.612700  3.908594
##                                0_1_0  2.561737  2.373625  2.762227
##                                1_1_0  6.211540  5.250503  7.339416
##                                0_2_0  3.599215  3.492653  3.708798
##                                1_2_0  8.197931  7.785472  8.634075
##                                0_0_1  4.864513  4.573036  5.172861
##                                1_0_1  8.892662  8.169368  9.680663
##                                0_1_1  5.335671  3.883233  7.283048
##                                1_1_1 14.018044  9.395788 21.230158
##                                0_2_1  9.738638  8.826337 10.747831
##                                1_2_1 19.782798 17.777192 22.055374
```



Having difficulty walking or climbing stairs yields a 19.78x increased chance of having heart disease, diabetes, and a stroke, with a 95% confidence interval of (17.78, 22.06). This same covariate yields a 9.74x increased likelihood of having diabetes and a stroke, with a 95% CI of (8.83, 10.75).

Sex

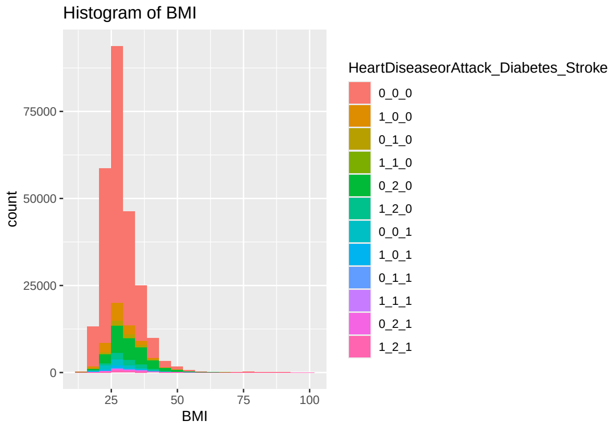


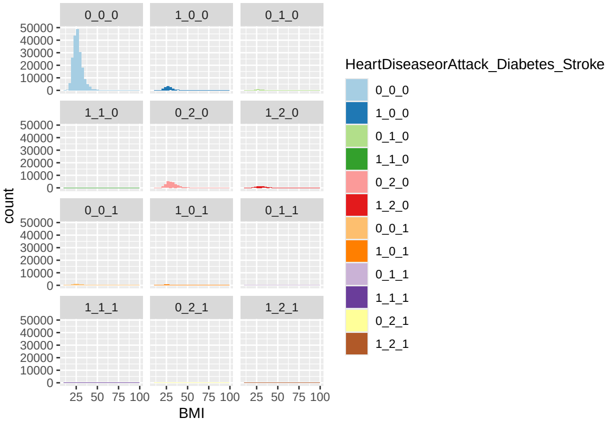
Males have a 1.01x increased likelihood of having diabetes and a stroke compared to females, with a (0.91, 1.11) 95% CI. Further, males have a 1.44x increased likelihood of having all three conditions, with a 95% CI of (1.31, 1.59)

Non-categorical Variable EDA

For the other variables in our dataset that are not categorical, we will simply look at some basic EDA; primarily in the form of histograms to see how their distributions differ.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



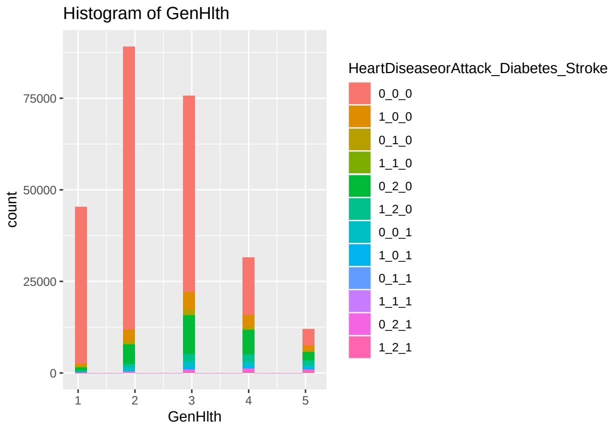


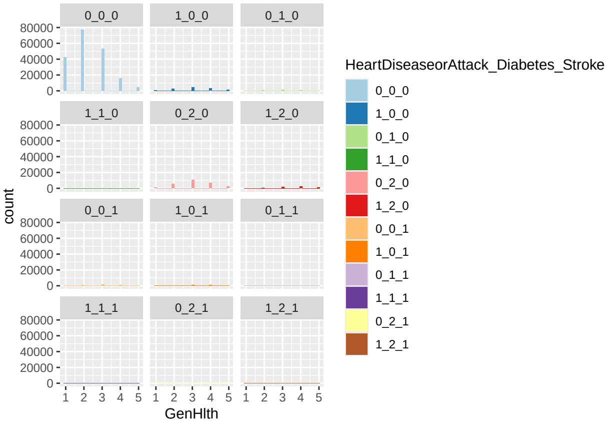
BMI values less than 18.5 are considered underweight. BMI values between 18.5 and 24.9 are classified as healthy. BMI values between 25 and 29.9 are considered overweight. Lastly, BMI values greater than or equal to 30 are classified as obese.

The histograms illustrate that a large majority of people with underweight, healthy, or overweight BMI scores did not experience heart disease, diabetes, or stroke. However, we can see that a small portion of people with overweight to obese BMI scores have experienced heart disease, diabetes, or a combination of heart disease and diabetes. This indicates that there is relationship between overweight to obese BMI scores, and people that have experienced 1 or more of heart disease, diabetes, and stroke.

General Health

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

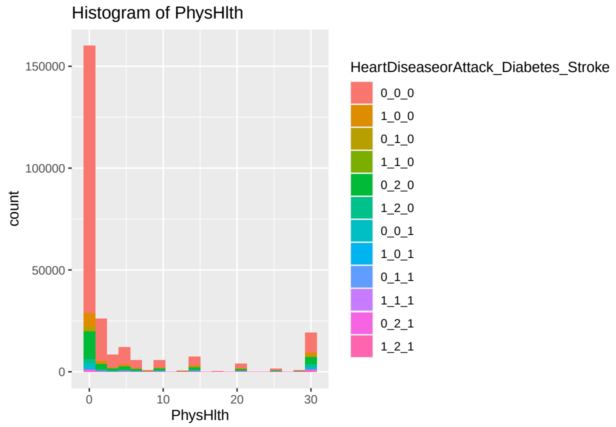


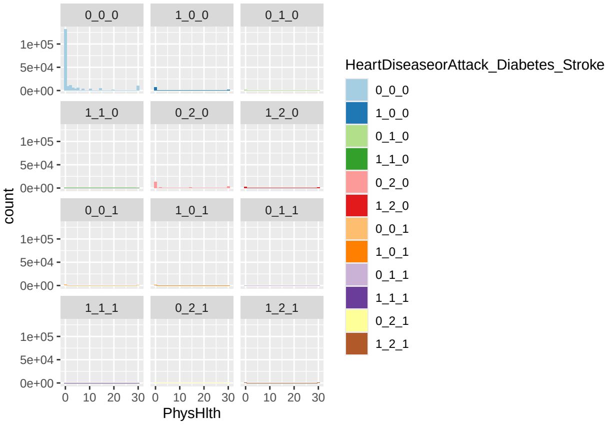


The general health histograms indicate that most people with excellent general health scores have not experienced heart disease, diabetes, or stroke. However, people with mediocre to unhealthy general health scores tend to experience heart disease, diabetes, or a combination of heart disease and diabetes.

Physical Health

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```

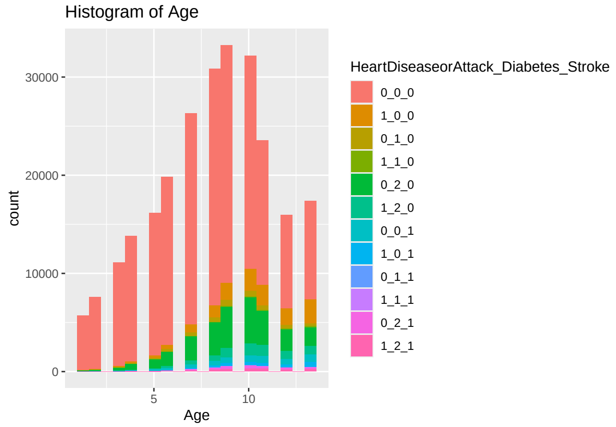


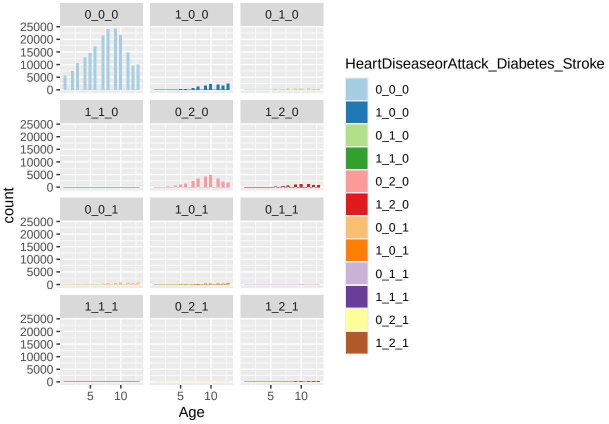


Similar to general health, the histograms for physical health indicate that most people with excellent physical health have not experienced heart disease, diabetes, or stroke. However, a portion of people with excellent physical health have experienced heart disease, diabetes, or a combination of heart disease and diabetes.

Age

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```

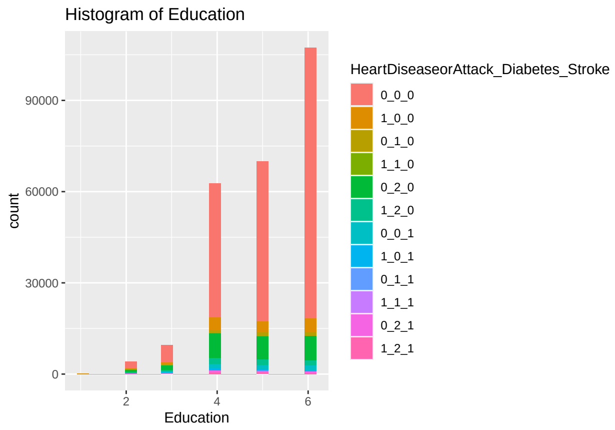


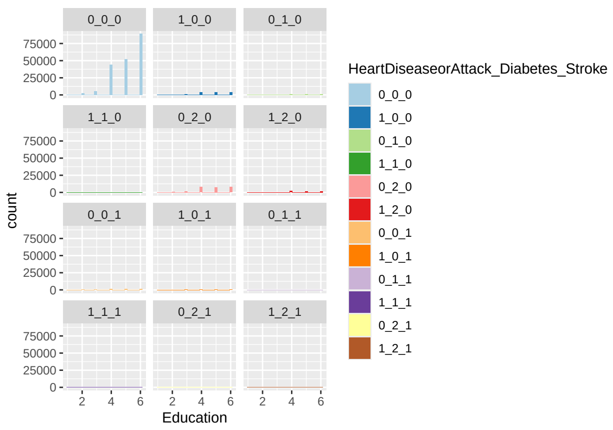


The age histograms indicate that younger people usually have not experienced heart disease, diabetes, or stroke. However, older people have a higher rate of experiencing heart disease, diabetes, or a combination of heart disease and diabetes.

Education

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```

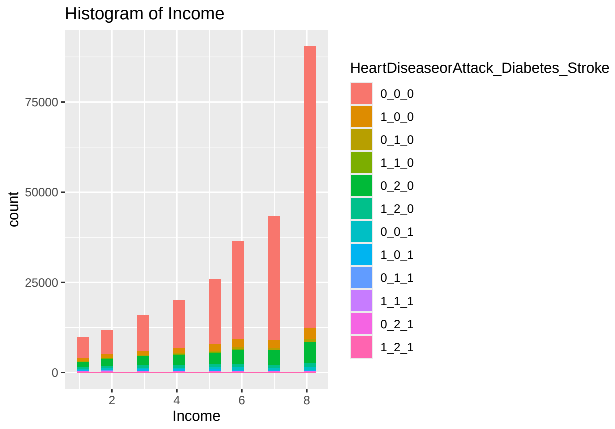


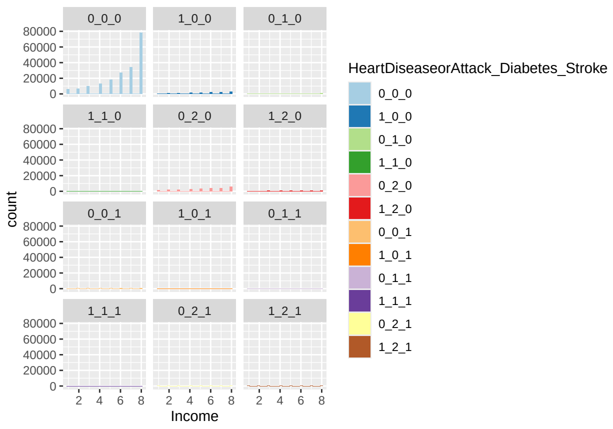


The histograms indicate that in terms of education, the largest group has completed their college education. In addition, we can see that people with more education have a higher rate of experiencing heart disease, diabetes, or a combination of heart disease and diabetes. I am curious whether this relationship is due to the fact that education can cause higher levels of stress, unhealthy eating habits, and less sleep, which can lead to heart disease and diabetes.

Income

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```





The histograms illustrate that there is a relationship between people with all levels of income and experiencing heart disease, diabetes, or a combination of heart disease and diabetes. In addition, we can see that people with higher levels of income have a higher rate of experiencing heart disease or diabetes. Similar to education, I wonder if higher levels of income are related to stress, sleep, etc., which can lead to heart disease and diabetes.

Methodologies Part 1 Results

Now, we want to see on average how each covariate increases the likelihood of one of our primary disabilities (Diabetes, Stroke, HeartDiseaseorAttack). We will begin by removing our first row from each of our Odds tables (as it represents having no disabilities).

```
##  [1] 2.2362718 0.8078346 7.0505834 4.5790065 4.5750506 0.6161100 0.5612826
##  [8] 1.2353811 1.9320723 7.9055432 1.3309869
```

Now, we can look back at our measures and see what our highest correlating variables are. In descending order, Difficulty Walking, High Blood Pressure, Cholesterol Check, High Cholesterol, and Smoking are our five highest correlating factors.

For those with difficulty walking, their odds of having Stroke/Diabetes/HD are 19.782x higher than someone who does not have difficulty walking, with the 95% CI spanning from 17.777 to 22.1.

For people with high blood pressure, their odds of having all three conditions are 14.42x higher than those with low blood pressure, with the 95% CI having values between 12.45 and 16.8.

Those that have gotten a cholesterol check in the past year are 5.49x more likely to have heart disease, diabetes, and a stroke than those that have not checked their cholesterol in the past year, with the 95% CI yielding values between 3.38 and 9.78.

In order to avoid further complicating our already large variable matrix, we will focus exclusively on the top three variables (DiffWalk,HighBP,CholCheck) and how they may jointly affect our three response variables (Stroke/Diabetes/HD).

Methodologies Part 2: Entropy, Conditional Entropy, and Mutual Information

Entropy measures the amount of ways that a system, or random variable, can be arranged or distributed. In the context of this project, entropy quantifies the expected value of the information contained in a vector. So higher entropy indicates more disorder, or randomness, in that particular subset of data and thus, its distribution is much more random and harder to use in predictive or discriminatory settings. Whereas lower entropies suggest that the data is more predictable, as they are much more structured and have more predictable distributions. Also, when selecting features, it is important to analyze entropy values as features with lower entropy provide much more discriminatory, or predictive, power.

Conditional Entropy is the measure of uncertainty in a given variable Y when some variable X is known. In the context of this project, we will use Conditional Entropy to quantify interaction effects and to measure the effectiveness of our chosen predictor variables. That is, a feature (or predictor) with lower conditional entropies may indicate that it provides more information in terms of predicting the response variable.

Mutual Information is similar to Conditional Entropy as it quantifies the "amount of information" obtained about one variable, Y, by observing another variable, X. In the context of this project, Mutual Information will be used as a quantity that measures the mutual dependence of the two random variables. Predictor variables with higher mutual information with the response variable are often considered more important for understanding and potentially predicting. Thus, features with higher mutual information with the response variable are often regarded as more important or relevant.

Using our odds ratio analyses, we have ascertained that the three most important predictor variables that we need to further analyze at a granular level are: Difficulty Walking (DiffWalk), Cholesterol Check (Chol Check), and High Blood Pressure (HighBP). Thus, we will further analyze how these predictor variables affect the randomness of our response variables, and thusly how important they are to consider, through an analysis of their entropies, conditional entropies, and mutual information.

Note that we are using log base (1) in our calculations of entropy because we are more interested in relative changes in our data rather than absolute values, which means that we wish to relatively simplify our calculations. However, this distinction may lead to some entropy values to be greater than 1, which is normally not possible with log base (2), but since we are using log base (1), in some cases with extremely disorganized data, it can occur.

Difficulty Walking

Total Count Tables for HD x DW, Diabetes x DW, Stroke x DW, Combined Response x DW

```
##                      DiffWalk
## HeartDiseaseorAttack      0      1    Sum
##                  0   197027  32760 229787
##                  1    13978   9915  23893
##                  Sum 211005  42675 253680
```

```
##         DiffWalk
## Diabetes      0      1    Sum
##      0   185434  28269 213703
##      1     3346   1285   4631
##      2    22225  13121  35346
##      Sum 211005  42675 253680
```

```
##       DiffWalk
## Stroke      0      1    Sum
##    0   205750  37638 243388
##    1     5255   5037  10292
##    Sum 211005  42675 253680
```

```
##                                      DiffWalk
## HeartDiseaseorAttack_Diabetes_Stroke      0      1    Sum
##                                 0_0_0 172543  21250 193793
##                                 1_0_0   8990   4161  13151
##                                 0_1_0   2891    912   3803
##                                 1_1_0    319    244    563
##                                 0_2_0  17901   7935  25836
##                                 1_2_0   3106   3136   6242
##                                 0_0_1   2851   1708   4559
##                                 1_0_1   1050   1150   2200
##                                 0_1_1     99     65    164
##                                 1_1_1     37     64    101
##                                 0_2_1    742    890   1632
##                                 1_2_1    476   1160   1636
##                                 Sum   211005  42675 253680
```

Using the tables above, we can find the total counts for each of our response variable in relation to the experimental unit's difficulty walking. Then with these counts, we can compute the overall entropy values for our response variables and the Difficulty Walking variable.

Overall Entropy Values for Response Variables and Difficulty Walking Variable

```
## The entropy value for Difficulty Walking Values: 0.4530585
```

```
## The entropy value for Heart Disease or Attack Values: 0.3121163
```

```
## The entropy value for Diabetes Values: 0.4921534
```

```
## The entropy value for Stroke Values: 0.169754
```

```
## The entropy value for Combined Response Values: 0.9457289
```

Here we can see the overall entropy values for each of our four response variables and our Difficulty Walking variable. The entropy values of our response variables tells us that the Stroke variable has the lowest uncertainty and thusly the most the most homogenous as it has the lowest entropy value at 0.169754. The Combined Response variable, conversely, has the highest entropy value at 0.9457289, which indicates that it has the most randomness and most heterogenous of all of our response variables. The other two response variables, Diabetes and Heart Disease, both have moderate entropies, which suggest that the data is somewhat random but not entirely so, and somewhat homoegenous, but also not entirely so. Furthermore, the entropy value of our Difficulty Walking Value is 0.4530585, which suggests that it probably has moderate predictive power as it contains valuable information that can definitely contribute to making any sort of prediction or classification, but it cannot tell the whole story. However, since it does still contain valuable information, it is worth analyzing further.

Entropy Values for Heart Disease Dependent on Difficulty Walking Values

```
## Entropy values for Heart Disease with no Difficulty Walking: 0.2438155
```

```
## Entropy values for Heart Disease with Difficulty Walking: 0.5420853
```

Here we have our entropy values for heart disease that are dependent on our Difficulty Walking variable. The entropy value for heart disease without difficulty walking is 0.244, indicating not much randomness and a good level of homogeneity, and the entropy value with difficulty walking is 0.542, which indicates a level of randomness similar to that of a coin flip, since we can round that value to 0.5. This means that there is a medium level of predictive power, but it cannot predict very accurately.

Entropy Values for Diabetes Dependent on Difficulty Walking Values

```
## Entropy values for Diabetes with no Difficulty Walking: 0.4163034
```

```
## Entropy values for Diabetes with Difficulty Walking: 0.7409158
```

Next, we analyze our entropy values for the diabetes response dependent on the difficulty walking covariate. Without difficulty walking, the entropy value for diabetes is 0.416, which indicates a moderate level of predictive power. The entropy value for diabetes with difficulty walking is a high 0.741, indicating a relatively high level of randomness and disorder for that subset of data. By this logic, this subset of data would be difficult to use in predictive settings.

Entropy Values for Stroke Dependent on Difficulty Walking Values

```
## Entropy values for Stroke with no Difficulty Walking: 0.1165572
```

```
## Entropy values for Stroke with Difficulty Walking: 0.3629847
```

Here, we examine the entropy values for the stroke response variable dependent on the difficulty walking covariate. With difficulty walking, the entropy value is 0.363, and without, it is 0.117. Both of these entropy values are relatively low, indicating a low level of disorder in the subset of data, making it a good fit for prediction.

Entropy Values for Combined Response Dependent on Difficulty Walking Values

```
## Entropy values for Combined Response with no Difficulty Walking: 0.7622747
```

```
## Entropy values for Combined Response with Difficulty Walking: 1.615083
```

In the above section, we can find the relative entropy of our response variables dependent on if the patients had no difficulty walking or they had difficulty walking. Although each response variable's entropy was different from one another, they all shared a similar pattern, which is that all of our response variables entropy values were lower for patients who had no difficulty walking rather than those who did have difficulty walking. This might suggest that our response variables become more homogenous, and thus less random, when the patients did not have difficulty walking. Which in context of the data set means that subjects who had no difficulty walking often reported more similar results in terms of heart disease, diabetes, stroke, and a permutation of all three. Whereas subjects who reported difficulty walking experienced a much more uncertain distribution of results in terms of heart disease, diabetes, stroke, and a permuation of all three.

Proportion of Difficulty Walking Values

```
##         0         1
## 0.8317763 0.1682237
```

Here we can see that about 83.2% of patients in our data set did not report difficulty walking, and about 16.8% of patients did report difficulty walking. These proportions are essential in calculating and understanding the conditional entropies in the next section.

Conditional Entropy of Response Variables on Difficulty Walking

```
## Conditional Entropy of Heart Disease on Difficulty Walking: 0.2939915

## Conditional Entropy of Diabetes on Difficulty Walking: 0.4709109

## Conditional Entropy of Stroke on Difficulty Walking: 0.1580122

## Conditional Entropy of Combined Response on Difficulty Walking: 0.9057372
```

We can find that Stroke and Heart Disease have the lowest conditional entropies, which suggests that knowing if a patient has difficulty walking might help to predict if the patient has experienced a Stroke or Heart Disease. Thus, we know that there is a relatively strong relationship between Difficulty Walking and Stroke and Difficulty Walking and Heart Disease. Conversely, we also ascertained that there is a relatively weak relationship between Difficutly Walking and Diabetes and Difficulty Walking and our Combined Response. This suggests that knowing if a patient has difficulty walking does not have as big of an impact on the randomness of the data distribution for Diabetes and our Combined Response as it does for Stroke and Heart Disease.

Cholesterol Check

Overall Entropy Values for Cholesterol Check

```
## The entropy value for Cholesterol Check: 0.1593655

## The entropy value for Heart Disease or Attack Values: 0.3121163

## The entropy value for Stroke Values: 0.169754

## The entropy value for Diabetes: 0.4921534

## The entropy value for Combined Response: 0.9457289
```

Displayed are the overall entropy values for all four of our response variables and our Cholesterol Check variable. From the response variables, we can see that the entropy value for stroke is the lowest at 0.170, indicating the least uncertainty and the most homogeneity. Our combine response variable, in contrast, has the highest entropy value at 0.946, meaning it has the most randomness and disorder. Heart disease and diabetes both have mid-level entropy values, indicating that these subsets of data are only somewhat random, with diabetes being moreso, almost to the level comparable to that of a coin flip.

Entropy Values for Heart Disease Dependent on Cholesterol Check

```
## Entropy values for Heart Disease with no Cholesterol Check: 0.1299002
## Entropy values for Heart Disease with Cholesterol Check: 0.3178341
```

As we can glean from these entropy values for heart disease dependent on cholesterol check, both values are relatively low, with the entropy value without cholesterol check being very low at 0.130, indicating a very low level of randomness, and in turn, higher predictive power. The entropy value with cholesterol check is a higher, but still relatively low 0.318.

Entropy Values for Diabetes Dependent on Cholesterol Check

```
## Entropy values for Diabetes with no Cholesterol Check: 0.1578263
```

```
## Entropy values for Diabetes with Cholesterol Check: 0.5017507
```

The entropy values for diabetes now, reveal more of a contrast compared to heart disease when it is dependent on cholesterol check. Without the covariate, the entropy value is a low 0.158, meaning not so much randomness and an easy use in predictive settings. With the covariate, the entropy value is roughly akin to the randomness of a coin flip at 0.502. With this information, we can see that the data for diabetes with cholesterol check is not very predictable.

Entropy Values for Stroke Dependent on Cholesterol Check

```
## Entropy values for Stroke with no Cholesterol Check: 0.08354337
```

```
## Entropy values for Stroke with Cholesterol Check: 0.1727117
```

The entropy values for stroke with or without the covariate are both very low, at 0.173 and 0.084, respectively. This indicates that the data for stroke, whether or not there has been a cholesterol check in the past year, is very predictable.

Entropy Values for Combined Response Dependent on Cholesterol Check

```
## Entropy values for Combined Response with no Cholesterol Check: 0.3524323
```

```
## Entropy values for Combined Response with Cholesterol Check: 0.9642584
```

Proportion of Cholesterol Check Values

The entropy value for the combined response without the covariate is a moderate 0.352, indicating some level of randomness, while the entropy value with the covariate is an extremely high 0.964, rendering it very difficult to predict.

```
##         0         1
## 0.0373305 0.9626695
```

Here we can see that about 96.3% of patients in our data set reported getting a cholesterol check in the past year, and about 3.7% of patients did not report getting a cholesterol check in the past year.

Conditional Entropy of Response Variables Dependent on Cholesterol Check:

```
## Conditional Entropy of Heart Disease on Cholesterol Check: 0.3108184
```

```
## Conditional Entropy of Diabetes on Cholesterol Check: 0.4889118
## Conditional Entropy of Stroke on Cholesterol Check: 0.169383
```

```
## Conditional Entropy of Combined Response on Cholesterol Check: 0.9414186
```

From these conditional entropy values, we find that stroke, then heart disease had the lowest values, indicating that knowing if the person has had a check on their cholesterol in the past year may help predict if the person has had a stroke or heart disease. Harder to predict is diabetes, since the conditional entropy value is almost 0.5, rendering the predictability comparable to a coin flip. Worst of all is the combined response, with has a value of 0.941, making prediction extremely difficult and suggests that knowing the covariate in this case does not have a major impact on the combined response.

Blood Pressure

Overall Entropy Values for High Blood Pressure and Response Variables

```
## The entropy value for High Blood Pressure Values: 0.6830313
```

```
## The entropy value for Heart Disease or Attack Values: 0.3121163
```

```
## The entropy value for Diabetes Values: 0.4921534
```

```
## The entropy value for Stroke values: 0.169754
```

```
## The entropy value for Combined Response Value: 0.9457289
```

From these overall entropy values, we find that stroke has the lowest entropy value at 0.170, while the combined response variable has the highest entropy value at 0.946. Heart disease has a moderate entropy value at 0.312, and diabetes has a higher, medium entropy value at 0.492. The least random subset of all is the stroke data, but the highest level of randomness resides in the combined reponse.

Entropy Values for Heart Disease Dependent on High Blood Pressure

```
## Entropy values for Heart Disease with no High Blood Pressure: 0.1716771
```

```
## Entropy values for Heart Disease with High Blood Pressure: 0.4474397
```

Without high blood pressure, the entropy value for heart disease is a low 0.172, meaning this subset of data is more structured and has a more predictable distribution. With high blood pressure, the entropy value is a moderate 0.447, meaning there is a moderate randomness to this subset of data.

Entropy Values for Diabetes Dependent on HBP

```
## Entropy values for Diabetes with no High Blood Pressure: 0.2915768
```

```
## Entropy values for Diabetes with High Blood Pressure: 0.6718584
```

For diabetes, no high blood pressure resulted in a moderate entropy value of 0.292, which indicates some level of randomness to the data. With high blood pressure, the entropy value is a somewhat high 0.672, indicating a somewhat high level of randomness, rendering this data moderately difficult to predict.

Entropy Values for Stroke Dependent on HBP

```
## Entropy values for Stroke with no High Blood Pressure: 0.09179293
```

```
## Entropy values for Stroke with High Blood Pressure: 0.2538046
```

For stroke, both entropy values with or without high BP are very low. Without resulted in a value of 0.092, and with resulted in a value of 0.254. Since without high BP has a lower value, this subset of data is much easier to predict, but the high BP subset does not have much randomness either.

Entropy Values for Combined Response Dpeendent on HBP

```
## Entropy values for Combined Response with no High Blood Pressure: 0.542712
```

```
## Entropy values for Combined Response with High Blood Pressure: 1.349472
```

For the combined response, the entropy values show levels of randomness from both ends. Without high BP, the randomness is comparable to a coin flip at 0.543, and with high BP, there is an extreme level of randomness since the entropy value is well over 1 at 1.349, rendering this subset of data extremely difficult to predict.

Proportion of Blood Pressure Distribution

```
##         0         1
## 0.5709989 0.4290011
```

About 57.1% of people did not report high blood pressure, while about 42.9% of people reported high blood pressure.

Conditional Entropy of Response Variables on Blood Pressure

```
## Conditional Entropy of Heart Disease on High Blood Pressure: 0.2899795
```

```
## Conditional Entropy of Diabetes on High Blood Pressure: 0.454718
```

```
## Conditional Entropy of Stroke on High Blood Pressure: 0.1612961
```

```
## Conditional Entropy of Combined Response on High Blood Pressure: 0.8888127
```

Looking at the conditional entropies for the response variables on high blood pressure, we can see that the conditional entropy for stroke is the lowest at 0.161, suggesting that knowing if someone has high blood pressure can help predict a stroke. The conditional entropies for heart disease and diabetes suggest moderate links to knowing if someone with high BP has had either condition. The combined response has a high conditional entropy at 0.889, indicating randomness and a weak association with having high blood pressure.

Conditional Entropy of Response Variables Based on Combined Predictor Covariate:

```
## Conditional Entropy of Heart Disease with respect to Combined Predictor Covariate
## 0.2781586
```

```
## Conditional Entropy of Diabetes with respect to Combined Predictor Covariate
## 0.4411987
## Conditional Entropy of Stroke with respect to Combined Predictor Covariate
## 0.1529631
```

```
## Conditional Entropy of Combined Response Variable with respect to Combined Predictor Covariate
## 0.8609792
```

Looking at the conditional entropies for the response variables with respect to the combined predictor covariate, we find that the lowest value is the conditional entropy of stroke at 0.153, indicating a strong association between stroke and the combined predictor covariate. We can assume that the combined predictor covariate can be effective in predicting the stroke response. Heart disease is a little less easy to predict, as it has a conditional entropy of 0.278, but there is an association with the combined predictor covariate. The diabetes response variable has a weaker link with the combined predictor covariate, with a conditional entropy of 0.441, but the weakest association is with the combined predictor covariate is the combined response variable, with a conditional entropy of 0.861.

Interaction Effects

```
## Difference in Heart Disease Entropy from using Combined Covariate vs Individual Predictor Variables:
## -0.007601597
```

```
## Difference in Diabetes Entropy from using Combined Covariate vs Individual Predictor Variables:
## -0.01096469
```

```
## Difference in Stroke Entropy from using Combined Covariate vs Individual Predictor Variables:
## -0.003779785
```

```
## Difference in Combined Response Entropy from using Combined Covariate vs Individual Predictor Variabl
## -0.01646844
```

Since the difference in the entropy for all 4 potential response variables when considering a covariate that combines all three of our predictor variables is less than the difference in entropy when considering each predictor variable individually, we can conclude that there is no extra clarity gained from combining all three predictor variables into a single covariate variable. That is, there is no interaction effect between the three predictor variables that might make a predictive model potentially more accurate if they had combined all three predictors into one.

Mutual Information

Mutual Information Table of Heart Disease

```
##                            Variables Mutual_Information
## 1 Heart Disease x Difficulty Walking         0.01812472
## 2         Heart Disease x Chol Check         0.00129782
## 3     Heart Disease x Blood Pressure         0.02213673
## 4 Heart Disease x Combined Covariate         0.03395767
```

Using the mutual information table above, we can see that our Combined Predictor Covariate and Blood Pressure variables have the strongest dependency, or relationship, to the Heart Disease variable. This suggest that if we were only focusing on Heart Disease as a response variable, the most important features to consider would be the Combined Covariate and Blood Pressure, as these two variables provide the most information about Heart Disease.

Mutual Information Table of Heart Disease

```
##                         Variables Mutual_Information
## 1 Diabetes x Difficulty Walking        0.021242447
## 2         Diabetes x Chol Check        0.003241549
## 3     Diabetes x Blood Pressure        0.037435367
## 4 Diabetes x Combined Covariate        0.050954672
```

From the mutual information table for diabetes, we find that the two strongest dependencies are with the combined covariate at 0.051 and with blood pressure at 0.037. This indicates that if we were to solely focus on diabetes as a response variable, the combined covariate and blood pressure would give us the most information on diabetes.

Mutual Information Table for Stroke

```
##                       Variables Mutual_Information
## 1 Stroke x Difficulty Walking       0.0117417530
## 2         Stroke x Chol Check       0.0003709951
## 3     Stroke x Blood Pressure       0.0084578449
## 4 Stroke x Combined Covariate       0.0167908082
```

For the stroke mutual information table, we find that the combined covariate at 0.017 and difficulty walking at 0.012 have the highest values in the table, suggesting that if we were to only focus on stroke as the response variable, then the combined covariate and difficulty walking would be the covariates that give us the most information on the stroke response.

Mutual Information Table for Combined Response

```
##                                  Variables Mutual_Information
## 1 Combined Response x Difficulty Walking        0.039991658
## 2         Combined Response x Chol Check        0.004310261
## 3     Combined Response x Blood Pressure        0.056916175
## 4 Combined Response x Combined Covariate        0.084749652
```

The mutual information table for our combined response shows that the highest values are the combined covariate at 0.085, and blood pressure at 0.057. This indicates that these two covariates would give us the most information if we were to solely focus on the combined response.

From these results of the mutual information tables, we find that the combined covariate had the highest mutual information value for all four response variables, which is to be expected as the combined covariate is, by definition, a combination of all three potential predictor variables. However, the most interesting phenomenon to note here is that the Blood Pressure Variable produces the second highest mutual information for all but the Stroke response variable. This might suggest that Blood Pressure has a much closer relationship with Diabetes, Heart Disease, and a combination of Heart Disease, Diabetes, and Stroke, than any of the other individual predictor variables.

Conclusion

In our analysis of our three response variables, Diabetes, Stroke, and HeartDiseaseandAttack, we considered how all of the other categorical and factored variables may act as a significant predictor for these health defects. We began by calculating the odds ratios of Smoker, HighBP, Veggies, Fruits, DiffWalk, NoDocBcCost, HeavyAlcoholConsumption,HighChol, CholCheck, AnyHealthcare, and Sex against all 12 possible combinations of our three response variables. Based upon both the highest odd ratios for contracting all three defects, a singular defect, and the average across all combinations, we decided that Difficulty Walking, High Blood Pressure, and Cholesterol Check are the three most impactful categorical predictors. Although we did not delve deeper into the other factored variables, it is of important consideration to that people with low-incomes, lower education levels, worsened mental, general, and physical health levels are much more predisposed to developing one of these major health ailments.

To analyze these variables at a granular level, we looked at their entropy values, where we found that our Cholesterol Check variable had the lowest overall entropy of all of our predictor variables and that our Stroke variable had the lowest overall entropy of all our response variables. Then, using the total counts of the joint distributions of these variables, we calculated each response variable's conditional entropy with respect to each of the chosen predictor variables, which we then used to construct a way to see if there are potentially any interaction effects between Difficutly Walking, High Blood Pressure, and Cholesterol Check. Since we had found that there is no added reduction in our response variable's entropy when combining the three into a singular covariate rather than considering them alone, we concluded that there is no special interaction effect between the three predictor variables. Then we used our calculations of conditional entropy to find the mutual information that each predictor variable gives us. In these tables, we found that the combined covariate gave the most amount of information about each of our four chosen response variables, which is to be expected. However, we also found that the High Blood Pressure variable gave much more information for almost all of our response variables (except the Stroke variable) in comparison to the other predictor variables. Thus, we have effectively analyzed the potential effects that each one of our chosen predictor variables have on our chosen response variables in terms of reducing the randomness in their data and making it easier to predict.

Code Appendix

```r
hd <- read.csv("heart_disease_health_indicators_BRFSS2015.csv")
nrow(hd)
print("The number of rows with ")
nrow(!is.na(hd)) == nrow(hd)
length(hd)
hd_noncat <- hd %>% select(Age, BMI, PhysHlth, GenHlth, MentHlth, Education, Income)
summary(hd_noncat)
rv_data = combine_vars(hd, vars = c('HeartDiseaseorAttack','Diabetes','Stroke'), sep = "_")
tab1=xtabs(~HeartDiseaseorAttack_Diabetes_Stroke+Smoker,data=rv_data)
#print(tab1)

measure1 = oddsratio(tab1,conf.level = 0.95)$measure
print(measure1)

tab1df = as.data.frame(tab1)
ggplot(tab1df, aes(x = HeartDiseaseorAttack_Diabetes_Stroke, y = Freq, fill = Smoker)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Distribution of Heart Disease/Attack, Diabetes, and Stroke by Smoker",
       x = "Heart Disease/Attack, Diabetes, and Stroke",
       y = "Count",
       fill = "Smoker") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

tab2=xtabs(~HeartDiseaseorAttack_Diabetes_Stroke+Fruits,data=rv_data)
measure2 = oddsratio(tab2,conf.level = 0.95)$measure
#print(measure2)
tab2df = as.data.frame(tab2)
ggplot(tab2df, aes(x = HeartDiseaseorAttack_Diabetes_Stroke, y = Freq, fill = Fruits)) +
  geom_bar(stat = "identity") +
  labs(title = "Distribution of Heart Disease/Attack, Diabetes, and Stroke by Fruits",
       x = "Heart Disease/Attack, Diabetes, and Fruits",
       y = "Count",
       fill = "Fruits") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
tab3=xtabs(~HeartDiseaseorAttack_Diabetes_Stroke+HighBP,data=rv_data)
```

```
#tab3
measure3 = oddsratio(tab3,conf.level = 0.95)$measure
print(measure3)
tab3df = as.data.frame(tab3)
ggplot(tab3df, aes(x = HeartDiseaseorAttack_Diabetes_Stroke, y = Freq, fill = HighBP)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Distribution of Heart Disease/Attack, Diabetes, and Stroke by HighBP",
       x = "Heart Disease/Attack, Diabetes, and HighBP",
       y = "Count",
       fill = "HighBP") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

tab4=xtabs(~HeartDiseaseorAttack_Diabetes_Stroke+HighChol,data=rv_data)
#tab4
measure4 = oddsratio(tab4,conf.level = 0.95)$measure
#print(measure4)
tab4df = as.data.frame(tab4)
ggplot(tab4df, aes(x = HeartDiseaseorAttack_Diabetes_Stroke, y = Freq, fill = HighChol)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Distribution of Heart Disease/Attack, Diabetes, and Stroke by HighChol",
       x = "Heart Disease/Attack, Diabetes, and HighBP",
       y = "Count",
       fill = "HighChol") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

tab5=xtabs(~HeartDiseaseorAttack_Diabetes_Stroke+CholCheck,data=rv_data)
measure5 = suppressWarnings(oddsratio(tab5,conf.level = 0.95)$measure)
print(measure5)
tab5df = as.data.frame(tab5)
ggplot(tab5df, aes(x = HeartDiseaseorAttack_Diabetes_Stroke, y = Freq, fill = CholCheck)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Distribution of Heart Disease/Attack, Diabetes, and Stroke by CholCheck",
       x = "Heart Disease/Attack, Diabetes, and HighBP",
       y = "Count",
       fill = "CholCheck") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
tab6=xtabs(~HeartDiseaseorAttack_Diabetes_Stroke+Veggies,data=rv_data)
#tab6
measure6 = oddsratio(tab6,conf.level = 0.95)$measure
#measure6
tab6df = as.data.frame(tab6)
ggplot(tab6df, aes(x = HeartDiseaseorAttack_Diabetes_Stroke, y = Freq, fill = Veggies)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Distribution of Heart Disease/Attack, Diabetes, and Stroke by Veggies",
       x = "Heart Disease/Attack, Diabetes, and HighBP",
       y = "Count",
       fill = "Veggies") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
tab7=xtabs(~HeartDiseaseorAttack_Diabetes_Stroke+HvyAlcoholConsump,data=rv_data)
tab7 = tab7[-10,]
measure7 = oddsratio(tab7,conf.level = 0.95)$measure
#measure7
tab7df = as.data.frame(tab7)
ggplot(tab7df, aes(x = HeartDiseaseorAttack_Diabetes_Stroke, y = Freq, fill = HvyAlcoholConsump)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Distribution of Heart Disease/Attack, Diabetes, and Stroke by HvyAlcoholConsump",
       x = "Heart Disease/Attack, Diabetes, and HighBP",
       y = "Count",
       fill = "HvyAlcoholConsump") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

tab8=xtabs(~HeartDiseaseorAttack_Diabetes_Stroke+AnyHealthcare,data=rv_data)
```

```r
#tab8
measure8 = oddsratio(tab8,conf.level = 0.95)$measure
#measure8
tab8df = as.data.frame(tab8)
ggplot(tab8df, aes(x = HeartDiseaseorAttack_Diabetes_Stroke, y = Freq, fill = AnyHealthcare)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Distribution of Heart Disease/Attack, Diabetes, and Stroke by AnyHealthcare",
       x = "Heart Disease/Attack, Diabetes, and HighBP",
       y = "Count",
       fill = "AnyHealthcare") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
tab9=xtabs(~HeartDiseaseorAttack_Diabetes_Stroke+NoDocbcCost,data=rv_data)
#tab9
measure9 = oddsratio(tab9,conf.level = 0.95)$measure
#measure9
tab9df = as.data.frame(tab9)
ggplot(tab9df, aes(x = HeartDiseaseorAttack_Diabetes_Stroke, y = Freq, fill = NoDocbcCost)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Distribution of Heart Disease/Attack, Diabetes, and Stroke by NoDocbcCost",
       x = "Heart Disease/Attack, Diabetes, and HighBP",
       y = "Count",
       fill = "NoDocbcCost") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
tab10=xtabs(~HeartDiseaseorAttack_Diabetes_Stroke+DiffWalk,data=rv_data)
#tab10
measure10 = oddsratio(tab10,conf.level = 0.95)$measure
measure10
tab10df = as.data.frame(tab10)
ggplot(tab10df, aes(x = HeartDiseaseorAttack_Diabetes_Stroke, y = Freq, fill = DiffWalk)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Distribution of Heart Disease/Attack, Diabetes, and Stroke by DiffWalk",
       x = "Heart Disease/Attack, Diabetes, and Stroke",
       y = "Count",
       fill = "DiffWalk") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
tab11=xtabs(~HeartDiseaseorAttack_Diabetes_Stroke+Sex,data=rv_data)
#tab11
measure11 = oddsratio(tab11,conf.level = 0.95)$measure
#measure11
tab11df = as.data.frame(tab11)
ggplot(tab11df, aes(x = HeartDiseaseorAttack_Diabetes_Stroke, y = Freq, fill = Sex)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Distribution of Heart Disease/Attack, Diabetes, and Stroke by Sex",
       x = "Heart Disease/Attack, Diabetes, and Stroke",
       y = "Count",
       fill = "Sex") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
ggplot(data=rv_data, aes(x=BMI,fill = HeartDiseaseorAttack_Diabetes_Stroke)) +
  geom_histogram(bins=20,position='stack')+
  labs(x ='BMI', title ='Histogram of BMI')
pal <- brewer.pal(12, "Paired")
ggplot(rv_data, aes(x = BMI, fill = HeartDiseaseorAttack_Diabetes_Stroke)) +
  geom_histogram() +
  scale_fill_brewer(palette = "Paired") +
  facet_wrap(~ HeartDiseaseorAttack_Diabetes_Stroke, ncol = 3)

ggplot(data=rv_data, aes(x=GenHlth,fill = HeartDiseaseorAttack_Diabetes_Stroke)) +
  geom_histogram(bins=20,position='stack')+
  labs(x ='GenHlth', title ='Histogram of GenHlth')
ggplot(rv_data, aes(x = GenHlth, fill = HeartDiseaseorAttack_Diabetes_Stroke)) +
  geom_histogram() +
  scale_fill_brewer(palette = "Paired") +
  facet_wrap(~ HeartDiseaseorAttack_Diabetes_Stroke, ncol = 3)
```

```
ggplot(data=rv_data, aes(x=PhysHlth,fill = HeartDiseaseorAttack_Diabetes_Stroke)) +
  geom_histogram(bins=20,position='stack')+
  labs(x ='PhysHlth', title ='Histogram of PhysHlth')
ggplot(rv_data, aes(x = PhysHlth, fill = HeartDiseaseorAttack_Diabetes_Stroke)) +
  geom_histogram() +
  scale_fill_brewer(palette = "Paired") +
  facet_wrap(~ HeartDiseaseorAttack_Diabetes_Stroke, ncol = 3)
ggplot(data=rv_data, aes(x=Age,fill = HeartDiseaseorAttack_Diabetes_Stroke)) +
  geom_histogram(bins=20,position='stack')+
  labs(x ='Age', title ='Histogram of Age')
ggplot(rv_data, aes(x = Age, fill = HeartDiseaseorAttack_Diabetes_Stroke)) +
  geom_histogram() +
  scale_fill_brewer(palette = "Paired") +
  facet_wrap(~ HeartDiseaseorAttack_Diabetes_Stroke, ncol = 3)
ggplot(data=rv_data, aes(x=Education,fill = HeartDiseaseorAttack_Diabetes_Stroke)) +
  geom_histogram(bins=20,position='stack')+
  labs(x ='Education', title ='Histogram of Education')
ggplot(rv_data, aes(x = Education, fill = HeartDiseaseorAttack_Diabetes_Stroke)) +
  geom_histogram() +
  scale_fill_brewer(palette = "Paired") +
  facet_wrap(~ HeartDiseaseorAttack_Diabetes_Stroke, ncol = 3)
ggplot(data=rv_data, aes(x=Income,fill = HeartDiseaseorAttack_Diabetes_Stroke)) +
  geom_histogram(bins=20,position='stack')+
  labs(x ='Income', title ='Histogram of Income')
ggplot(rv_data, aes(x = Income, fill = HeartDiseaseorAttack_Diabetes_Stroke)) +
  geom_histogram() +
  scale_fill_brewer(palette = "Paired") +
  facet_wrap(~ HeartDiseaseorAttack_Diabetes_Stroke, ncol = 3)
measure_list <- list(measure1, measure2, measure3, measure4, measure5, measure6, measure7, measure8, measure9, measure10, measur
for (i in 1:11) {
  measure_list[[i]] <- measure_list[[i]][-1, ]
}
measure_avgs <- numeric(11)  # Create an empty numeric vector to store averages
for (i in 1:11) {
  measure_avgs[i] <- mean(measure_list[[i]][, 1])
}
measure_avgs
df = rv_data
tab1=xtabs(~HeartDiseaseorAttack+DiffWalk,data=df)
tab1=addmargins(tab1)
print(tab1)
tab1_Diabetes=xtabs(~Diabetes+DiffWalk,data=df)
tab1_Diabetes=addmargins(tab1_Diabetes)
print(tab1_Diabetes)
tab1_Stroke=xtabs(~Stroke+DiffWalk,data=df)
tab1_Stroke=addmargins(tab1_Stroke)
print(tab1_Stroke)
tab1_Combined=xtabs(~HeartDiseaseorAttack_Diabetes_Stroke+DiffWalk,data=df)
tab1_Combined=addmargins(tab1_Combined)
print(tab1_Combined)
DiffWalk_entropy = Entropy(tab1[3,1:2],base=exp(1))
HeartDieseaseorAttack_entropy = Entropy(tab1[1:2,3],base=exp(1))
Diabetes_entropy = Entropy(tab1_Diabetes[1:3,3],base=exp(1))
Stroke_entropy = Entropy(tab1_Stroke[1:2,3],base=exp(1))
Combined_entropy = Entropy(tab1_Combined[1:12,3],base=exp(1))

cat("The entropy value for Difficulty Walking Values:",entropy(df$DiffWalk), fill = T)
cat("The entropy value for Heart Disease or Attack Values:", entropy(df$HeartDiseaseorAttack), fill = T)
cat("The entropy value for Diabetes Values:", Diabetes_entropy, fill = T)
cat("The entropy value for Stroke Values:", Stroke_entropy, fill = T)
cat("The entropy value for Combined Response Values:", Combined_entropy, fill = T)
entropy_no_diffwalking = Entropy(tab1[1:2,1],base=exp(1))
entropy_yes_diffwalking = Entropy(tab1[1:2,2],base=exp(1))

cat("Entropy values for Heart Disease with no Difficulty Walking:", entropy_no_diffwalking, fill = T)
cat("Entropy values for Heart Disease with Difficulty Walking:", entropy_yes_diffwalking, fill = T)
entropy_no_diffwalking_diabetes = Entropy(tab1_Diabetes[1:3,1],base=exp(1))
```

```
entropy_yes_diffwalking_diabetes = Entropy(tab1_Diabetes[1:3,2],base=exp(1))

cat("Entropy values for Diabetes with no Difficulty Walking:", entropy_no_diffwalking_diabetes, fill = T)
cat("Entropy values for Diabetes with Difficulty Walking:", entropy_yes_diffwalking_diabetes, fill = T)
entropy_no_diffwalking_stroke = Entropy(tab1_Stroke[1:2,1],base=exp(1))
entropy_yes_diffwalking_stroke = Entropy(tab1_Stroke[1:2,2],base=exp(1))

cat("Entropy values for Stroke with no Difficulty Walking:", entropy_no_diffwalking_stroke, fill = T)
cat("Entropy values for Stroke with Difficulty Walking:", entropy_yes_diffwalking_stroke, fill = T)
entropy_no_diffwalking_combined = Entropy(tab1_Combined[1:12,1],base=exp(1))
entropy_yes_diffwalking_combined = Entropy(tab1_Combined[1:12,2],base=exp(1))

cat("Entropy values for Combined Response with no Difficulty Walking:", entropy_no_diffwalking_combined, fill = T)
cat("Entropy values for Combined Response with Difficulty Walking:", entropy_yes_diffwalking_combined, fill = T)
prop_w=tab1[3,1:2]/sum(tab1[3,1:2])
print(prop_w)
diffwalk_entropy=apply(tab1[c(1,2),c(1,2)],2,function(o) Entropy(o,base=exp(1)))
diffwalk_entropy_diabetes=apply(tab1_Diabetes[c(1:3),c(1,2)],2,function(o) Entropy(o,base=exp(1)))
diffwalk_entropy_stroke=apply(tab1_Stroke[c(1,2),c(1,2)],2,function(o) Entropy(o,base=exp(1)))
diffwalk_entropy_combined=apply(tab1_Combined[c(1:12),c(1,2)],2,function(o) Entropy(o,base=exp(1)))
entropy_w=sum(prop_w*diffwalk_entropy)
cat("Conditional Entropy of Heart Disease on Difficulty Walking:", entropy_w, fill = T)
entropy_w_diabetes=sum(prop_w*diffwalk_entropy_diabetes)
cat("Conditional Entropy of Diabetes on Difficulty Walking:", entropy_w_diabetes, fill = T)
entropy_w_stroke=sum(prop_w*diffwalk_entropy_stroke)
cat("Conditional Entropy of Stroke on Difficulty Walking:", entropy_w_stroke, fill = T)
entropy_w_combined=sum(prop_w*diffwalk_entropy_combined)
cat("Conditional Entropy of Combined Response on Difficulty Walking:", entropy_w_combined, fill = T)
tab2=xtabs(~HeartDiseaseorAttack+CholCheck,data=df)
tab2=addmargins(tab2)
tab2_diabetes=xtabs(~Diabetes+CholCheck,data=df)
tab2_diabetes=addmargins(tab2_diabetes)
tab2_stroke=xtabs(~Stroke+CholCheck,data=df)
tab2_stroke=addmargins(tab2_stroke)
tab2_combined=xtabs(~HeartDiseaseorAttack_Diabetes_Stroke+CholCheck,data=df)
tab2_combined=addmargins(tab2_combined)
Stroke_entropy = Entropy(tab2_stroke[1:2,3],base=exp(1))
HeartDieseaseorAttack_entropy = Entropy(tab2[1:2,3],base=exp(1))
diabetes_entropy = Entropy(tab2_diabetes[1:3,3], base = exp(1))
chol_entropy = Entropy(tab2_combined[13,1:2], base = exp(1))
cat("The entropy value for Cholesterol Check:", chol_entropy, fill = T)
cat("The entropy value for Heart Disease or Attack Values:", HeartDieseaseorAttack_entropy, fill = T)
cat("The entropy value for Stroke Values:",Stroke_entropy, fill = T)
cat("The entropy value for Diabetes:", diabetes_entropy, fill = T)
cat("The entropy value for Combined Response:", entropy(df$HeartDiseaseorAttack_Diabetes_Stroke), fill = T)
entropy_no_Check = Entropy(tab2[1:2,1],base=exp(1))
entropy_yes_Check = Entropy(tab2[1:2,2],base=exp(1))

cat("Entropy values for Heart Disease with no Cholesterol Check:", entropy_no_Check, fill = T)
cat("Entropy values for Heart Disease with Cholesterol Check:", entropy_yes_Check, fill = T)
entropy_no_Check_diabetes = Entropy(tab2_diabetes[1:3,1],base=exp(1))
entropy_yes_Check_diabetes = Entropy(tab2_diabetes[1:3,2],base=exp(1))

cat("Entropy values for Diabetes with no Cholesterol Check:", entropy_no_Check_diabetes, fill = T)
cat("Entropy values for Diabetes with Cholesterol Check:", entropy_yes_Check_diabetes, fill = T)
entropy_no_Check_stroke = Entropy(tab2_stroke[1:2,1],base=exp(1))
entropy_yes_Check_stroke = Entropy(tab2_stroke[1:2,2],base=exp(1))

cat("Entropy values for Stroke with no Cholesterol Check:", entropy_no_Check_stroke, fill = T)
cat("Entropy values for Stroke with Cholesterol Check:", entropy_yes_Check_stroke, fill = T)
entropy_no_Check_combined = Entropy(tab2_combined[1:12,1],base=exp(1))
entropy_yes_Check_combined = Entropy(tab2_combined[1:12,2],base=exp(1))

cat("Entropy values for Combined Response with no Cholesterol Check:", entropy_no_Check_combined, fill = T)
cat("Entropy values for Combined Response with Cholesterol Check:", entropy_yes_Check_combined, fill = T)
prop_s=tab2[3,1:2]/sum(tab2[3,1:2])
print(prop_s)
```

```r
heart_entropy_chol=apply(tab2[c(1,2),c(1,2)],2,function(o) Entropy(o,base=exp(1)))
diabetes_entropy_chol = apply(tab2_diabetes[c(1:3),c(1,2)],2,function(o) Entropy(o,base=exp(1)))
stroke_entropy_chol = apply(tab2_stroke[c(1:2),c(1,2)],2,function(o) Entropy(o,base=exp(1)))
combined_entropy_chol = apply(tab2_combined[c(1:12),c(1,2)],2,function(o) Entropy(o,base=exp(1)))
entropy_s=sum(prop_s*heart_entropy_chol)
cat("Conditional Entropy of Heart Disease on Cholesterol Check:", entropy_s, fill =T)
entropy_diabetes_chol = sum(prop_s*diabetes_entropy_chol)
cat("Conditional Entropy of Diabetes on Cholesterol Check:", entropy_diabetes_chol, fill= T)
entropy_stroke_chol=sum(prop_s*stroke_entropy_chol)
cat("Conditional Entropy of Stroke on Cholesterol Check:", entropy_stroke_chol, fill =T)
combined_chol=sum(prop_s*combined_entropy_chol)
cat("Conditional Entropy of Combined Response on Cholesterol Check:", combined_chol, fill =T)
tab3=xtabs(~HeartDiseaseorAttack+HighBP,data=df)
tab3=addmargins(tab3)
tab3_diabetes = xtabs(~Diabetes+HighBP, data = df)
tab3_diabetes = addmargins(tab3_diabetes)
tab3_stroke = xtabs(~Stroke+HighBP, data = df)
tab3_stroke = addmargins(tab3_stroke)
tab3_combined = xtabs(~HeartDiseaseorAttack_Diabetes_Stroke+HighBP, data = df)
tab3_combined = addmargins(tab3_combined)
Blood_entropy = Entropy(tab3[3,1:2],base=exp(1))
HeartDieseaseorAttack_entropy = Entropy(tab3[1:2,3],base=exp(1))

cat("The entropy value for High Blood Pressure Values:",Blood_entropy, fill = T)
cat("The entropy value for Heart Disease or Attack Values:", HeartDieseaseorAttack_entropy, fill = T)
cat("The entropy value for Diabetes Values:", Diabetes_entropy, fill = T)
cat("The entropy value for Stroke values:", Stroke_entropy, fill = T)
cat("The entropy value for Combined Response Value:", Combined_entropy, fill = T)
entropy_no_high_bp = Entropy(tab3[1:2,1],base=exp(1))
entropy_yes_high_bp = Entropy(tab3[1:2,2],base=exp(1))

cat("Entropy values for Heart Disease with no High Blood Pressure:", entropy_no_high_bp, fill = T)
cat("Entropy values for Heart Disease with High Blood Pressure:", entropy_yes_high_bp, fill = T)
entropy_no_high_bp_diabetes = Entropy(tab3_diabetes[1:3,1],base=exp(1))
entropy_yes_high_bp_diabetes = Entropy(tab3_diabetes[1:3,2],base=exp(1))

cat("Entropy values for Diabetes with no High Blood Pressure:", entropy_no_high_bp_diabetes, fill = T)
cat("Entropy values for Diabetes with High Blood Pressure:", entropy_yes_high_bp_diabetes, fill = T)
entropy_no_high_bp_stroke = Entropy(tab3_stroke[1:2,1],base=exp(1))
entropy_yes_high_bp_stroke = Entropy(tab3_stroke[1:2,2],base=exp(1))

cat("Entropy values for Stroke with no High Blood Pressure:", entropy_no_high_bp_stroke, fill = T)
cat("Entropy values for Stroke with High Blood Pressure:", entropy_yes_high_bp_stroke, fill = T)
entropy_no_high_bp_combined = Entropy(tab3_combined[1:12,1],base=exp(1))
entropy_yes_high_bp_combined = Entropy(tab3_combined[1:12,2],base=exp(1))

cat("Entropy values for Combined Response with no High Blood Pressure:", entropy_no_high_bp_combined, fill = T)
cat("Entropy values for Combined Response with High Blood Pressure:", entropy_yes_high_bp_combined, fill = T)

prop_b=tab3[3,1:2]/sum(tab3[3,1:2])
print(prop_b)
blood_entropy=apply(tab3[c(1,2),c(1,2)],2,function(o) Entropy(o,base=exp(1)))
blood_entropy_diabetes = apply(tab3_diabetes[c(1:3),c(1,2)], 2, function(o) Entropy(o, base=exp(1)))
blood_entropy_stroke = apply(tab3_stroke[c(1,2),c(1,2)],2,function(o) Entropy(o,base=exp(1)))
blood_entropy_combined = apply(tab3_combined[c(1:12),c(1,2)],2,function(o) Entropy(o,base=exp(1)))
entropy_b=sum(prop_b*blood_entropy)
cat("Conditional Entropy of Heart Disease on High Blood Pressure:", entropy_b, fill =T)
entropy_b_diabetes = sum(prop_b*blood_entropy_diabetes)
cat("Conditional Entropy of Diabetes on High Blood Pressure:", entropy_b_diabetes, fill =T)
entropy_b_stroke = sum(prop_b*blood_entropy_stroke)
cat("Conditional Entropy of Stroke on High Blood Pressure:", entropy_b_stroke, fill =T)
entropy_b_combined = sum(prop_b*blood_entropy_combined)
cat("Conditional Entropy of Combined Response on High Blood Pressure:", entropy_b_combined, fill =T)

df$Walk_CholCheck_BP=paste(df$DiffWalk,df$CholCheck, df$HighBP,sep=',')
tab4=xtabs(formula = ~HeartDiseaseorAttack+Walk_CholCheck_BP,data=df)
tab4=addmargins(tab4)
```

```r
tab4_diabetes=xtabs(formula = ~Diabetes+Walk_CholCheck_BP,data=df)
tab4_diabetes=addmargins(tab4_diabetes)
tab4_stroke=xtabs(formula = ~Stroke+Walk_CholCheck_BP,data=df)
tab4_stroke=addmargins(tab4_stroke)
tab4_combined=xtabs(formula = ~HeartDiseaseorAttack_Diabetes_Stroke+Walk_CholCheck_BP,data=df)
tab4_combined=addmargins(tab4_combined)

probs=tab4[3,1:8]/sum(tab4[3,1:8])
# use apply function to obtain CE[Y|X=x]
H_yxx=apply(tab4[1:2,1:8],2,function(o) Entropy(o,base=exp(1)))

CE_yx=sum(probs*H_yxx)
CE_yx_diabetes = condentropy(df$Diabetes, df$Walk_CholCheck_BP)
CE_yx_stroke = condentropy(df$Stroke, df$Walk_CholCheck_BP)
CE_yx_combined = condentropy(df$HeartDiseaseorAttack_Diabetes_Stroke, df$Walk_CholCheck_BP)
cat('Conditional Entropy of Heart Disease with respect to Combined Predictor Covariate',CE_yx, fill = T)
cat('Conditional Entropy of Diabetes with respect to Combined Predictor Covariate',CE_yx_diabetes, fill = T)
cat('Conditional Entropy of Stroke with respect to Combined Predictor Covariate',CE_yx_stroke, fill = T)
cat('Conditional Entropy of Combined Response Variable with respect to Combined Predictor Covariate',CE_yx_combined, fill = T)

CE_y= Entropy(tab1[1:2,3],base=exp(1))
CE_yx1=entropy_w
CE_yx2=entropy_s
CE_yx3=entropy_b

Difference_Covariate = (CE_y - CE_yx)
Difference_No_Covariate = ((CE_y - CE_yx1) + (CE_y - CE_yx2) + (CE_y - CE_yx3))

Interaction_effect_hd = Difference_Covariate - Difference_No_Covariate
cat('Difference in Heart Disease Entropy from using Combined Covariate vs Individual Predictor Variables:', Interaction_effect_h


Difference_covariate_diabetes = (entropy(df$Diabetes) - CE_yx_diabetes)
Difference_No_covariate_diabetes = ((entropy(df$Diabetes) - entropy_w_diabetes)+(entropy(df$Diabetes) - entropy_diabetes_chol)+

interaction_effect_diabetes = Difference_covariate_diabetes - Difference_No_covariate_diabetes
cat('Difference in Diabetes Entropy from using Combined Covariate vs Individual Predictor Variables:', interaction_effect_diabet

Difference_covariate_stroke = (entropy(df$Stroke) - CE_yx_stroke)
Difference_No_covariate_stroke = ((entropy(df$Stroke) - entropy_w_stroke)+(entropy(df$Stroke) - entropy_stroke_chol)+ (entropy(d

interaction_effect_stroke = Difference_covariate_stroke - Difference_No_covariate_stroke
cat('Difference in Stroke Entropy from using Combined Covariate vs Individual Predictor Variables:', interaction_effect_stroke,


Difference_covariate_combined = (entropy(df$HeartDiseaseorAttack_Diabetes_Stroke) - CE_yx_combined)
Difference_No_covariate_combined = ((entropy(df$HeartDiseaseorAttack_Diabetes_Stroke) - entropy_w_combined)+(entropy(df$HeartDis

interaction_effect_combined = Difference_covariate_combined - Difference_No_covariate_combined
cat('Difference in Combined Response Entropy from using Combined Covariate vs Individual Predictor Variables:', interaction_effe

MI_W = CE_y - CE_yx1
MI_S = CE_y - CE_yx2
MI_B = CE_y - CE_yx3
MI_covariate = CE_y - CE_yx

mutual_info_table = data.frame(
  Variables = c("Heart Disease x Difficulty Walking", "Heart Disease x Chol Check", "Heart Disease x Blood Pressure", "Heart Dis
Mutual_Information = c(MI_W,MI_S,MI_B,MI_covariate)
)

mutual_info_table
mutual_info_table_diabetes = data.frame(
  Variables = c("Diabetes x Difficulty Walking", "Diabetes x Chol Check", "Diabetes x Blood Pressure", "Diabetes x Combined Cova
Mutual_Information = c(mutinformation(df$Diabetes, df$DiffWalk),mutinformation(df$Diabetes, df$CholCheck),mutinformation(df$Diab
)
```

```
mutual_info_table_diabetes
mutual_info_table_stroke = data.frame(
  Variables = c("Stroke x Difficulty Walking", "Stroke x Chol Check", "Stroke x Blood Pressure", "Stroke x Combined Covariate"),
Mutual_Information = c(mutinformation(df$Stroke, df$DiffWalk),mutinformation(df$Stroke, df$CholCheck),mutinformation(df$Stroke, 
)

mutual_info_table_stroke
mutual_info_table_combined = data.frame(
  Variables = c("Combined Response x Difficulty Walking", "Combined Response x Chol Check", "Combined Response x Blood Pressure"
Mutual_Information = c(mutinformation(df$HeartDiseaseorAttack_Diabetes_Stroke, df$DiffWalk),mutinformation(df$HeartDiseaseorAtta
)

mutual_info_table_combined
```