STA 141A Final Project

Junqing He, junehe@ucdavis.edu (mailto:junehe@ucdavis.edu)

Ziyi Zeng, jerzeng@ucdavis.edu (mailto:jerzeng@ucdavis.edu)

Yucheng Zhao, yuczhao@ucdavis.edu (mailto:yuczhao@ucdavis.edu)

Tracy Zhu, tcizhu@ucdavis.edu (mailto:tcizhu@ucdavis.edu)

Haitong Zhu, htjzhu@ucdavis.edu (mailto:htjzhu@ucdavis.edu)

11/27/2022

Contribution

Ziyi Zeng and Haitong Zhu are mainly responsible for coding and the final integration of the project report. Ziyi mainly writes code to preliminarily process the data and visually present the relationship between the data variables. Haitong mainly codes for the model establishment and analysis. Junqing He is mainly responsible for writing the introduction of the project, the description of data variables, and the analysis of data visualization. Tracy Zhu and Yucheng Zhao are mainly responsible for interpreting the results and writing the methodology, model analysis, conclusions, and answers to our research questions in our report. All team members held 3 discussion meetings together and contributed to the project efficiently and successfully.

Introduction and Background

The heart is the muscle at the center of the cardiovascular system and has the primary function of circulating blood and oxygen around the human body. Because of this re- sponsibility, small abnormalities may have significant harmful effects to the body. The association of disorders of the heart and blood vessels are defined as cardiovascular dis- ease (CVD). This includes coronary heart disease, cerebrovascular disease, peripheral arterial disease, rheumatic heart disease, continental heart disease, deep vein thrombosis and pulmonary embolism [1].

The American Heart Association (AHA) reports that about 25% of the population in the United States currently have some form of cardiovascular disease (CVD) and about 47% of all Americans have at least 1 of 3 key risk factors for heart disease – high blood pressure, high cholesterol, and smoking. One person dies every 36 seconds because of CVD and 1 in every 4 deaths is related to heart disease. Furthermore, 20% of heart attacks are silent and patients are not aware of the damage done to their bodies. Recently, the sudden onset of the COVID-19 pandemic has raised the risk for CVD patients. The most common clinical manifestations of COVID-19 are respiratory related, which can arise as a mild flu-like sickness to potentially fatal acute respiratory distress syndrome or fulminant pneumonia [2]. Acute cardiac injury, heart failure, and arrhythmia are among some of the reported cardiovascular complications in patients, along with COVID-19. This indicates people who already have pre-existing CVD are more vulnerable to COVID-19 variants than those who do not.

Many types of CVD have extensive pre-symptomatic stages during which low-cost treat- ment can help improve results. Our project aims to use statistics methods to analyze the known factors of CVD patients. The models list out the factors that correlate to the existence of CVD. Then, a patient suspecting of CVD could ask for the professional analysis such as echocardiographic screening [4], test for C-reactive protein (CPR) [3], or they can pay extra attention to their diet and daily consumption[5] to prevent the CVD.

The prevalence of CVD makes it the leading cause of death for both men and women in the United States. However, the World Health Organization estimates that over 75% of premature CVD is preventable. Therefore, detecting and preventing the factors of heart disease have a great impact on the health of the United States population.

Methodology overview

To begin with, we observed, cleaned, and organized the raw data. We attempt to measure the performance of a model for predicting heart disease. To identify and remove errors and duplicate data to create usable datasets, we first use `na.omit` to find the data with missing values and clean them out. Then we are able to remove the rows containing missing values to prevent data corruption or failure to record data.

After processing the original data, we plan to use the supervised learning technique in this project. We split data into the training set and the test set to avoid overfitting. We chose 80% of the rows in the data for the training set and the rest 20% for the testing set randomly. [1]

Then we create a Generalized linear model (GLM) model with the data in the training set, the dependent variable is heart disease, while the 17 independent variables are BMI, Smoking, Alcohol Drinking, Stroke, Physical Health, Mental Health, Diff Walking, Sex, Age Catagory, Race, Diabetic, Physical Activity, Gen Health, Sleep Time, Asthma, Kidney Disease, and Skin Cancer. It is clear that we should use binomial regression to fit the model because we want to find the relationship between the probability of success of the binomially distributed variable Heart Disease and 17 independent variables.

After creating the preliminary model, we intend to verify whether the model's predictions are valid or not. So we made a confusion matrix and create a ROC curve (receiver operating characteristic curve) based on that which uses the results from the training set to predict 20% of the data to test the accuracy rate.

Next, We calculate F1 score from the confusion matrix for the GLM model and compare the F1 scores among the different models. The model with a higher F1 score is better to classify the observations, showing more accuracy of the data performance on the data set.

We observed that the number of individuals who reported that they had heart disease (yes) was significantly less than the number who did not have heart disease (no). Since we need to balance data to make the model more accurate, we look for the total number of "yes" data in heart disease and then randomly select the same number of "no" data. Then we combine these two equal numbers of data into `data_new` .

Lastly, we did the same visualization and GLM analysis as the first model again with the new data.

Dataset Description and Exploratory Data Analysis

The chosen dataset is the 2020 annual CDC survey of the health status of 400k adults in the United States. The data is a part of the Behavioral Risk Factor Surveillance System (BRFSS), which conducts annual telephone surveys to gather data on the health status of US residents. The dataset contains 319795 rows with 18 attributes, ranging from whether the respondents have ever reported having heart disease, to their body mass index, set, race, and age, physical, and mental health status, to whether they smoke or drink alcohol, to whether they have had a stroke, difficulty walking or climbing stairs. Some attributes have boolean values of yes or no, such as smoking, drinking, and stroke, some have others have categorical values like age, race, and general health, and some have floating point values, such as physical and mental health.

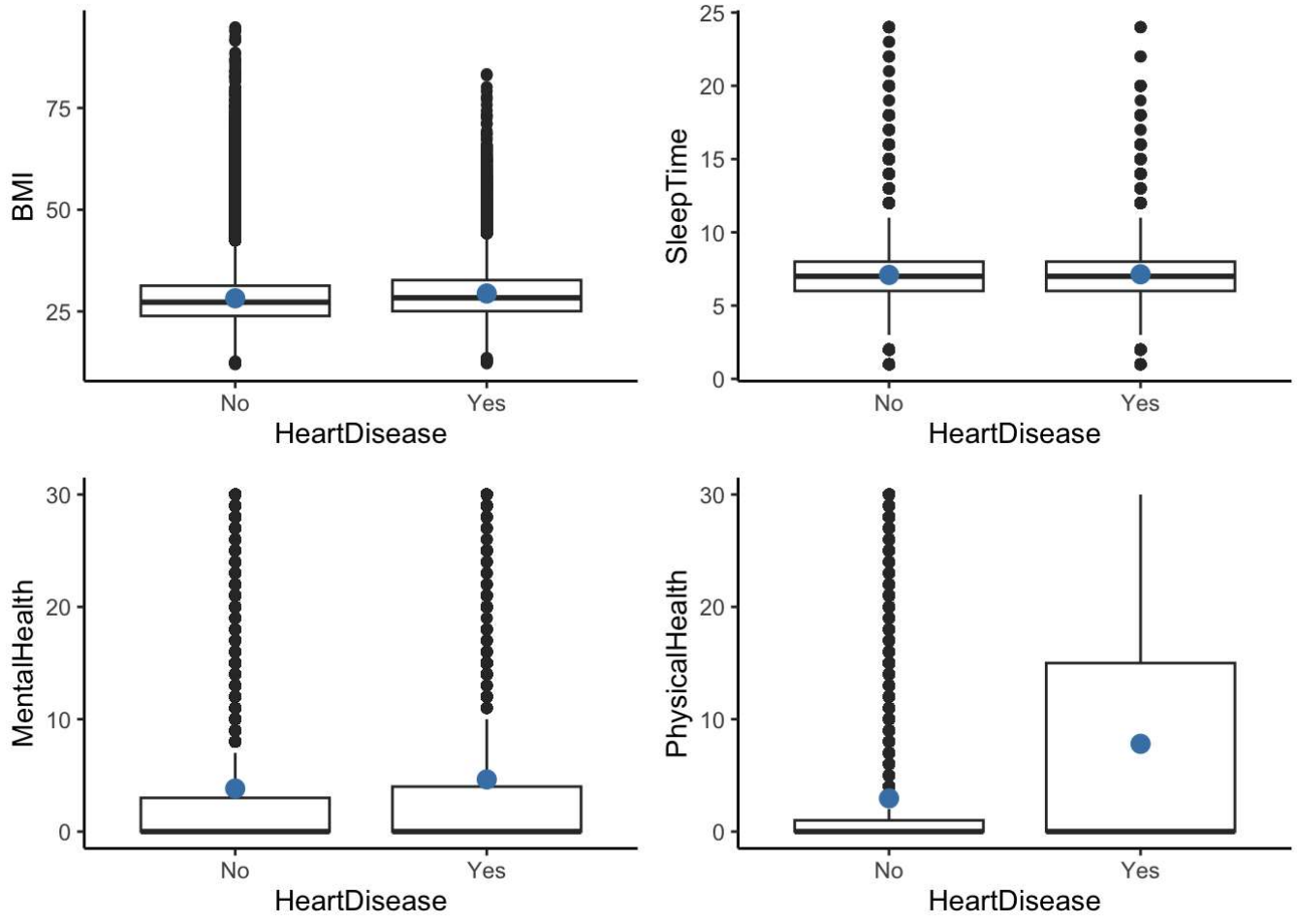
The dependent variable is under the column "HeartDisease", which represents whether the respondents have ever reported having coronary heart disease (CHD) or myocardial infarction (MI) and the independent variables will be the remaining variables of the dataset after being further reduced. The majority of the samples in the dataset report not having heart disease. A little less than 300,000 people report never having heart disease, while 25,000 people report having heart disease out of the nearly 320,000 samples collected.

We chose this dataset because it has data applicable to the residents of the United States, which allows additional users within the United States to get more accurate results. The goal of this project is to use generalized linear model to predict whether an individual is at risk of heart disease or not. With the result

information, people will be able to seek medical advice before a heart attack or something else harmful happens and lead a healthier life. The original dataset from the CDC had nearly 300 variables that were reduced to 18 variables by the creator of the dataset. And our group plans on using all 18 variables to build the model.

Firstly, we checked the numerical independent variables and found there exist some large outliers from the data because the maximum number is way much larger than the 3rd quartile in the summary table. Also, some values are not reasonable in real life. For example, a person's BMI could not be 75, and nobody would sleep around 24 hours per day. We drew the boxplots to see the distribution and outliers of the numerical data. From the boxplots, we decide to remove the rows that contain the top one percent and the bottom one percent of the `BMI` and `SleepTime` value in the dataset.

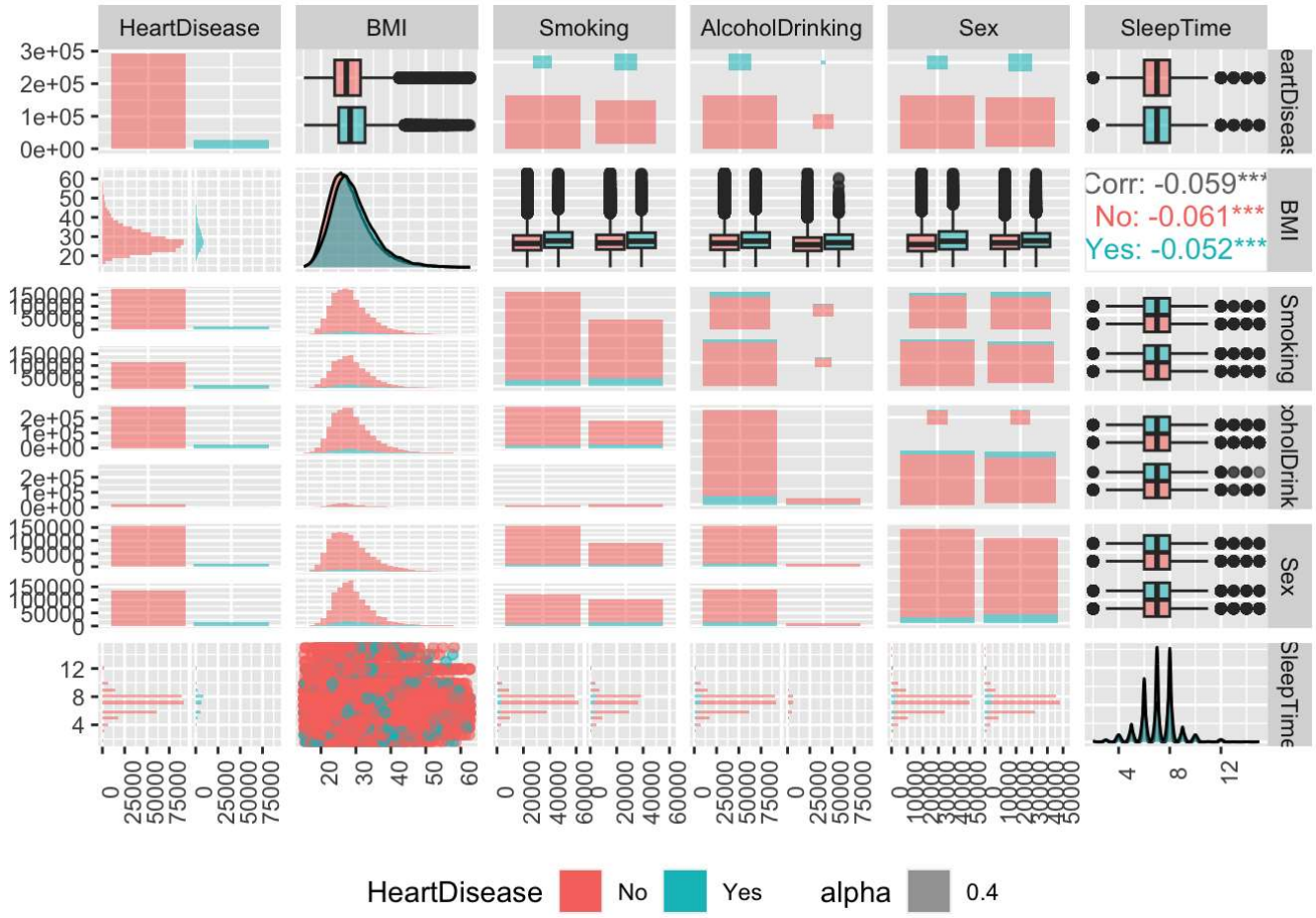
```
##       BMI         PhysicalHealth    MentalHealth      SleepTime
##  Min.   :12.02   Min.   : 0.000   Min.   : 0.000   Min.   : 1.000
##  1st Qu.:24.03   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 6.000
##  Median :27.34   Median : 0.000   Median : 0.000   Median : 7.000
##  Mean   :28.33   Mean   : 3.372   Mean   : 3.898   Mean   : 7.097
##  3rd Qu.:31.42   3rd Qu.: 2.000   3rd Qu.: 3.000   3rd Qu.: 8.000
##  Max.   :94.85   Max.   :30.000   Max.   :30.000   Max.   :24.000
```



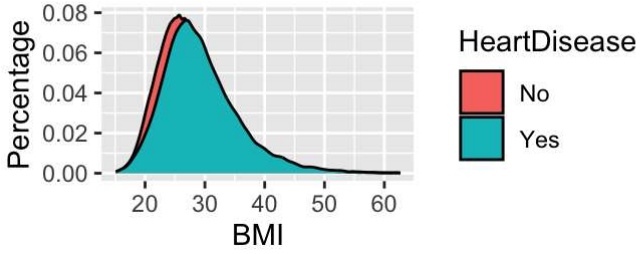
Then, we look at the relationship between categorical predictors proportions and whether the person has heart disease to further understand the data.

1. The person with higher BMI is more likely to have heart disease.
2. The proportion of males with heart disease is larger than the proportion of females with heart disease.
3. It is more likely for people who smoke to get heart disease.

4. The probability of people of different races to get heart disease varies. And Asian and Hispanic people seem to have lower rate with heart disease.
5. People with fair or poor general health have higher possibility to have heart disease.
6. Adults who reported doing physical activity or exercise during the past 30 days other than their regular job has lower chance to have heart disease.
7. Surprisingly, heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) have lower probability to have heart disease. But this might because of the people with poor health condition are not allowed to drink that often.
8. If a person had a stroke before, he/she has higher chance to have heart disease.
9. The more often people with physical illness and injury during the past 30 days, the more possible for him/her to have heart disease.
10. There does not show any clear pattern between mental health and whether a person have heart disease.
11. The heart disease becomes more prevalent as an individual increases in age.
12. If a person has serious difficulty walking or climbing stairs, he/she has higher probability to have heart disease.
13. People with diabetes (not during pregnancy) or borderline diabetes are more possible to have heart disease.
14. People with asthma are more possible to have heart disease.
15. People with kidney disease have much higher chance to have heart disease.
16. Sleeping around six to eight hours per day has the lowest probability to get heart disease. Sleeping too much or too little might increase the risk of heart disease.
17. If a person had skin cancer before, he/she has higher chance to have heart disease.

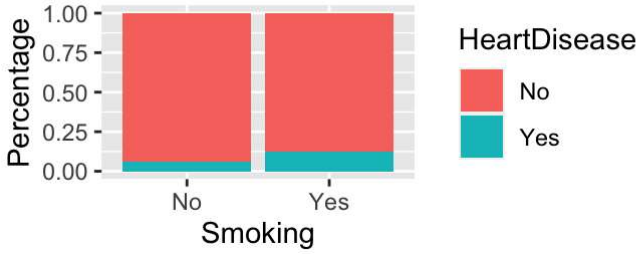


1 Percentage by BMI

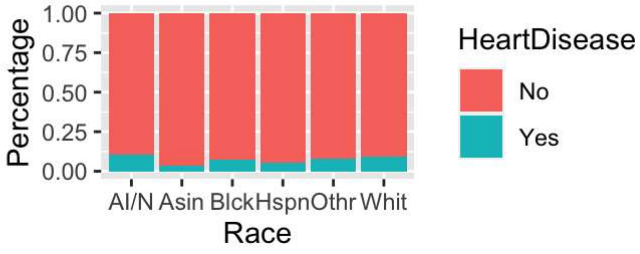


2 Percentage by Sex

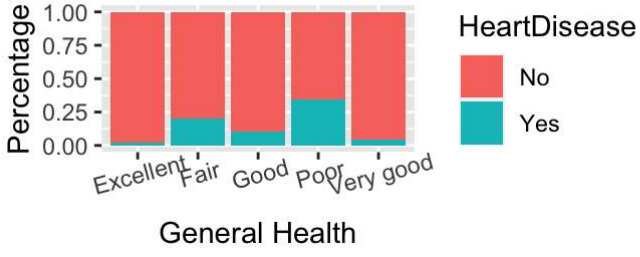
3 Percentage by Smoking status



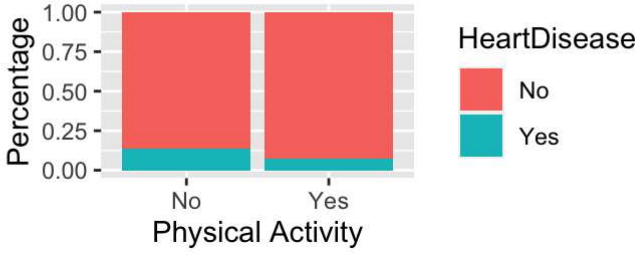
4 Percentage by Race



5 Percentage by General Health

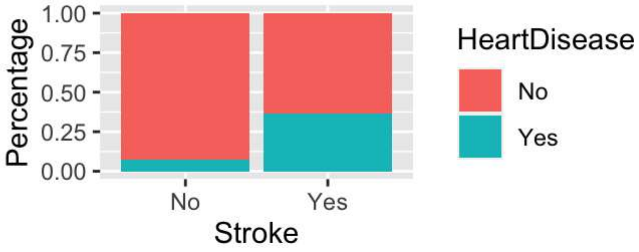


6 Percentage by Physical Activity

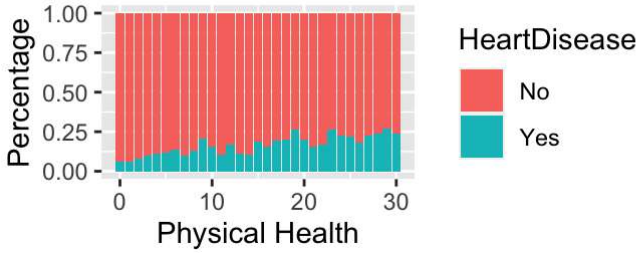


7 Percentage by AlcoholDrinking

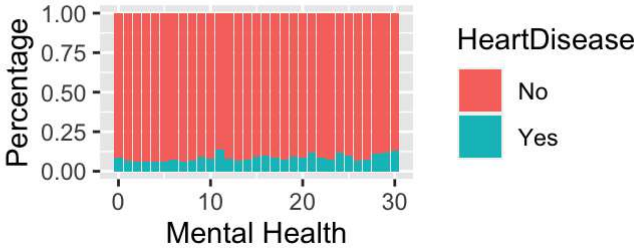
8 Percentage by Stroke



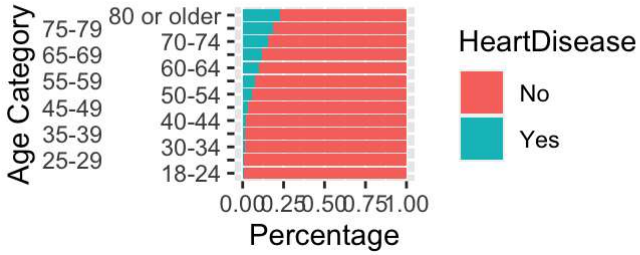
9 Percentage by Physical Health



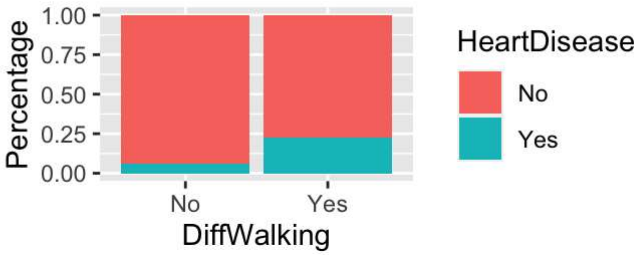
10 Percentage by Mental Health



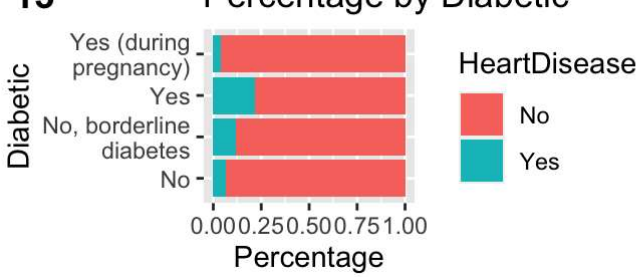
11 Percentage by Age Categor



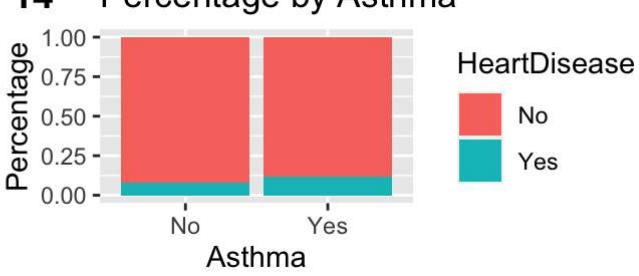
12 Percentage by DiffWalking



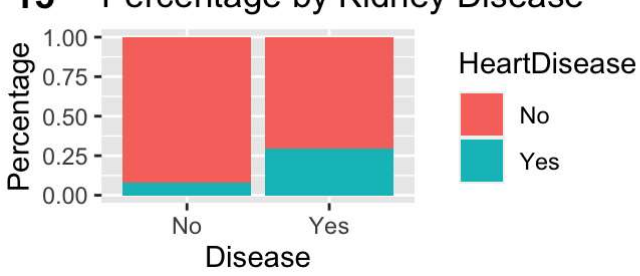
13 Percentage by Diabetic



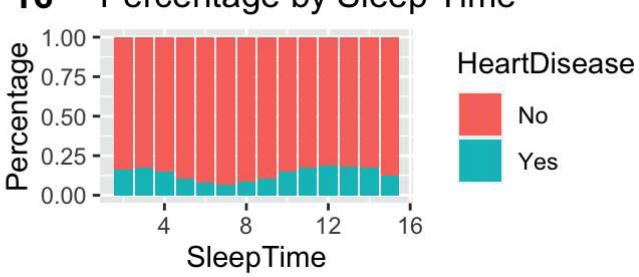
14 Percentage by Asthma



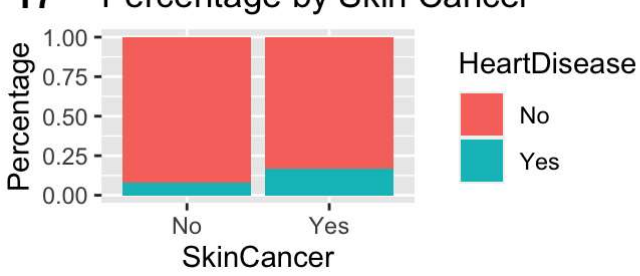
15 Percentage by Kidney Disease



16 Percentage by Sleep Time



17 Percentage by Skin Cancer



Training and Testing

Since our dependent variable HeartDisease is binary and has only yes and no categories, we converted the values under HeartDisease to 0 and 1 with the ifelse function that has 0 representing the respondent with no reported heart disease and one representing the respondent with reported heart disease. To ensure the precision and accuracy of our model, we split heart_data into a training set and a testing set, named heart_train and heart_test respectively. Here we implemented a 80 - 20 ratio. The training set consists of randomly selected 80% of the total rows (without replacement), whereas the testing set being the 20% rest. Later we will develop a Generalized Linear Model base on the training set and check the model with the testing set.

GLM model

```
##
## Call:
## glm(formula = HeartDisease ~ ., family = "binomial", data = heart_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1337  -0.4102  -0.2438  -0.1301   3.6042
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -6.2586782  0.1287039 -48.628  < 2e-16 ***
## BMI                             0.0093400  0.0013273   7.037 1.97e-12 ***
## SmokingYes                      0.3502064  0.0161481  21.687  < 2e-16 ***
## AlcoholDrinkingYes             -0.2475536  0.0378644  -6.538 6.24e-11 ***
## StrokeYes                       1.0427726  0.0255659  40.788  < 2e-16 ***
## PhysicalHealth                  0.0030600  0.0009726   3.146  0.00165 **
## MentalHealth                    0.0040570  0.0009987   4.062 4.86e-05 ***
## DiffWalkingYes                  0.2195929  0.0203817  10.774  < 2e-16 ***
## SexMale                         0.7235178  0.0163846  44.159  < 2e-16 ***
## AgeCategory25-29                0.1562770  0.1364672   1.145  0.25214
## AgeCategory30-34                0.4769533  0.1233308   3.867  0.00011 ***
## AgeCategory35-39                0.4950251  0.1195516   4.141 3.46e-05 ***
## AgeCategory40-44                0.9969561  0.1109714   8.984  < 2e-16 ***
## AgeCategory45-49                1.2859392  0.1071669  11.999  < 2e-16 ***
## AgeCategory50-54                1.7295618  0.1031807  16.762  < 2e-16 ***
## AgeCategory55-59                1.9472188  0.1016371  19.159  < 2e-16 ***
## AgeCategory60-64                2.1862578  0.1007421  21.702  < 2e-16 ***
## AgeCategory65-69                2.4364800  0.1004046  24.267  < 2e-16 ***
## AgeCategory70-74                2.7296061  0.1003260  27.207  < 2e-16 ***
## AgeCategory75-79                2.9199224  0.1009528  28.924  < 2e-16 ***
## AgeCategory80 or older          3.1872224  0.1006503  31.666  < 2e-16 ***
## RaceAsian                      -0.5288123  0.0942047  -5.613 1.98e-08 ***
## RaceBlack                      -0.3724892  0.0646351  -5.763 8.27e-09 ***
## RaceHispanic                   -0.2634873  0.0656744  -4.012 6.02e-05 ***
## RaceOther                      -0.0475989  0.0714102  -0.667  0.50506
## RaceWhite                      -0.1001993  0.0575665  -1.741  0.08176 .
## DiabeticNo, borderline diabetes  0.1284200  0.0472027   2.721  0.00652 **
## DiabeticYes                     0.4783814  0.0187861  25.465  < 2e-16 ***
## DiabeticYes (during pregnancy)  0.1927652  0.1150962   1.675  0.09397 .
## PhysicalActivityYes             0.0101599  0.0180276   0.564  0.57304
## GenHealthFair                   1.5056013  0.0367938  40.920  < 2e-16 ***
## GenHealthGood                   1.0449874  0.0330613  31.608  < 2e-16 ***
## GenHealthPoor                   1.9035193  0.0459788  41.400  < 2e-16 ***
## GenHealthVery good              0.4580481  0.0339671  13.485  < 2e-16 ***
## SleepTime                      -0.0256094  0.0052462  -4.882 1.05e-06 ***
## AsthmaYes                       0.2781430  0.0215594  12.901  < 2e-16 ***
## KidneyDiseaseYes                0.5720177  0.0274177  20.863  < 2e-16 ***
## SkinCancerYes                   0.1099311  0.0219693   5.004 5.62e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 148361  on 254509  degrees of freedom
## Residual deviance: 115234  on 254472  degrees of freedom
```

```
## AIC: 115310
##
## Number of Fisher Scoring iterations: 7
```

We then used function glm() to fit the generalized linear model with the training set. Since the goal is to determine what are the significant variables leading to cardiovascular disease, we chose HeartDisease(Whether or not a respondent having heart disease) as the respondent variable and the other 17 variables, including BMI, Smoking, AlcoholDrinking, Stroke, PhysicalHealth, MentalHealth, DiffWalking, Sex, AgeCategory,Race, Diabetic, PhysicalActivity, GenHealth, SleepTime, Asthma, KidneyDisease, and SkinCancer, as the independent variables.

From the general linear model above, we looked at the p-values associated with each predictor and determined that the following predictors are statistically significant at a significance level of 0.05:

The intercept is significant indicates that the value of the response variable(HeartDisease) is -6.2113 when the unit of all the predictors is 0.

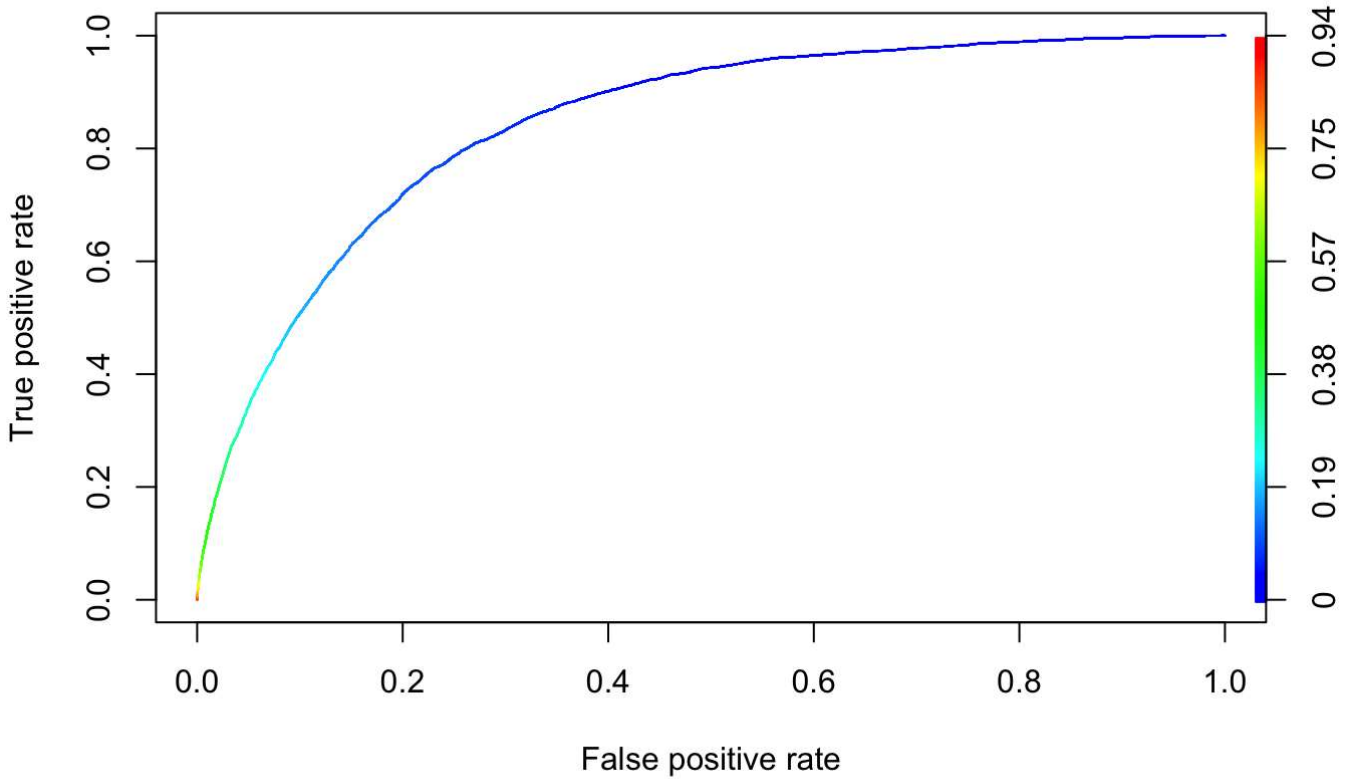
1. BMI: one unit increase in BMI is associated with an average change of 0.0092 of the response variable HeartDisease.
2. SMoking: one unit increase in Smoking(whether the respondent smokes) is associated with an average change of 0.3490 of the response variable HeartDisease.
3. AlcoholDrinking: one unit increase in AlcoholDrinking(whether the respondent drink alcohol) is associated with an average change of -0.2225 of the response variable HeartDisease.
4. Stroke: one unit increase in Stroke(whether the respondent had stroke before) is associated with an average change of 1.0353 of the response variable HeartDisease.
5. PhysicalHealth: one unit increase in PhysicalHealth is associated with an average change of 0.0026 of the response variable HeartDisease.
6. MentalHealth: one unit increase in MentalHealth is associated with an average change of 0.0048 of the response variable HeartDisease.
7. DiffWalking: one unit increase in DiffWalking(whether the respondent has difficulty in walking or climbing stairs) is associated with an average change of 0.2162 of the response variable HeartDisease.
8. Sex: one unit increase in Sex(Whether the respondent is male) is associated with an average change of 0.7040 of the response variable HeartDisease.
9. AgeCategory(except AgeCategory25-29): 30-34: 0.4939 35-39: 0.5225 40-44: 0.9756 45-49: 1.2917 50-54: 1.7095 55-59: 1.9234 60-64: 2.1799 65_69: 2.4352 70-74: 2.7276 75-79: 2.9173
10. Race(except RaceOther): Asian:average change of -0.5639 on HeartDisease. Black: average change of -0.3725 on HeartDisease.Hispanic: average change of -0.3032 on HeartDisease.White:average change of -0.1128 on HeartDisease.
11. Diabetic: one unit increase in Diabetic(whether the respondent has diabetes) is associated with an average change of 0.4718(have diabetes) and 0.1182(not have diabetes but on the borderline) of the response variable HeartDisease.
12. GenHealth: Poor: average change of 1.9154 on HeartDisease. Fair: average change of 1.5226 on HeartDisease. Good: average change of 1.0412 on HeartDisease. Very good: average change of 0.4472 on HeartDisease.
13. SleepTime: one unit increase in SleepTime is associated with an average change of -0.0280 of the response variable HeartDisease.
14. Asthma: one unit increase in Asthma(whether the respondent have asthma) is associated with an average change of 0.2781 of the response variable HeartDisease.
15. KidneyDisease: one unit increase in KidneyDisease(Whether the respondent have kidney disease) is associated with an average change of 0.5720 of the response variable HeartDisease.
16. SkinCancer: one unit increase in SkinCancer(Whether the respondent have skin cncer) is associated with an average change of 0.1307 of the response variable HeartDisease. [6]

Confusion Matrix

```
##
##      FALSE  TRUE
##   0 57695   474
##   1  4887   572
```

We created a confusion matrix that provided a contrast of predicted values against true values. Each of its columns represents the true result(FALSE = don't have heart disease, TRUE = have heart disease), while each row represents the predicted result(0 = don't have heart disease, 1 = have heart disease). Precision looks at the accuracy of the positive prediction, while Recall is the ratio of positive instances that are correctly detected by the classifier. We calculated that precision = 0.5266 and recall = 0.1018. With these values, we then calculated the F1 score for the comparison of different models that we have done later. Since a higher F1 score represents a higher precision and recall, the model with the highest F1 score will the best one in our consideration. [7]

ROC Curve

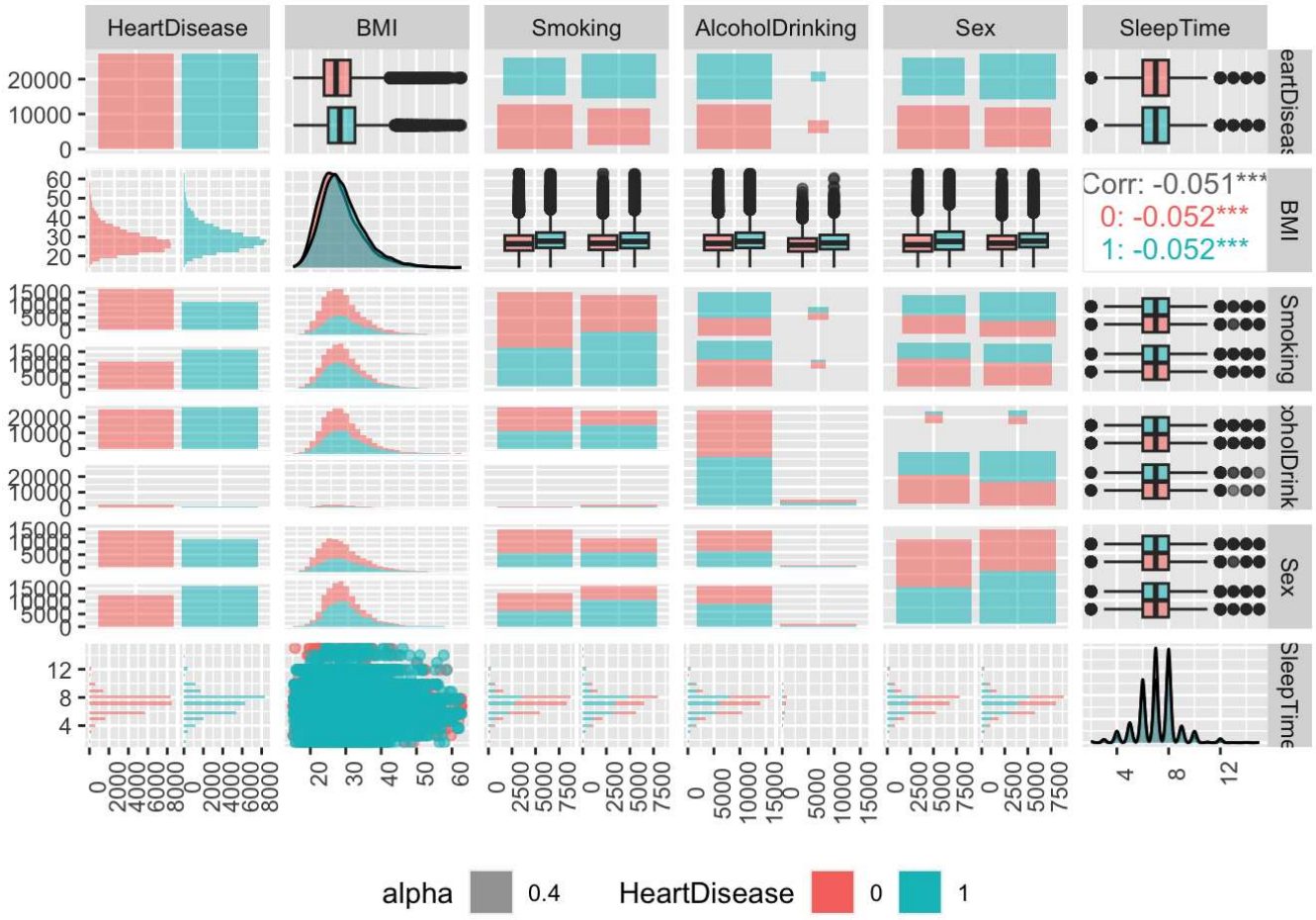


The ROC Curve of this model indicates a trade off between true positive rate and false positive rate. Thus, the model that has a curve closer to the top left corner and more area under curve is considered better. Later, We will compare the ROC curve of each model as well. [8]

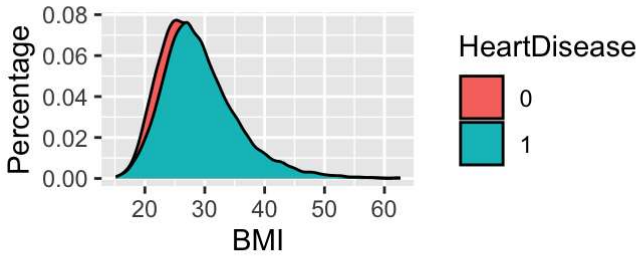
Because the dataset contains significantly more people who have never reported having heart disease (292,422) compared to the people who have reported having heart disease (27,373), we must balance the data. Otherwise, it will lower the accuracy of our model and lead to biased conclusions. In order to solve this problem, we use the following method to balance the original dataset. We first select the rows according to "Yes"(1) or "No"(0) in HeartDisease. Then, we randomly select equal amount of data of "No" in `HeartDisease` as the amount of "Yes" in `HeartDisease` and recombine them with the original data that have heart disease, which provides us a new balanced dataset. Then, we try to use it to train the model again.

New Dataset

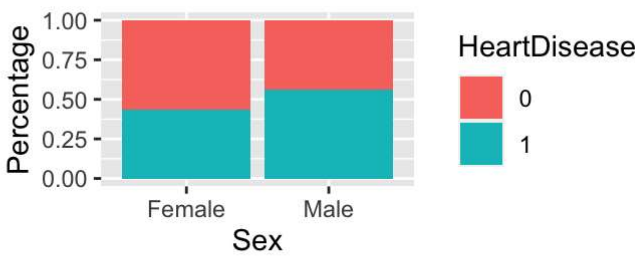
Data Visualization



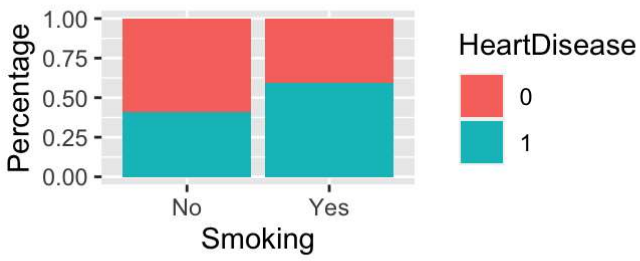
1 Percentage by BMI



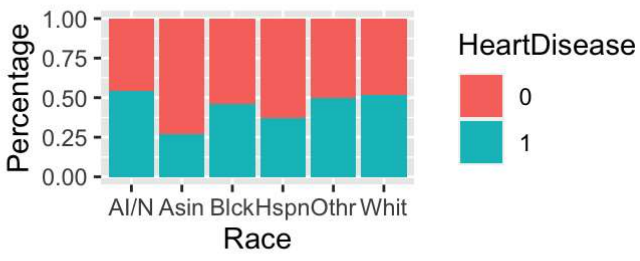
2 Percentage by Sex



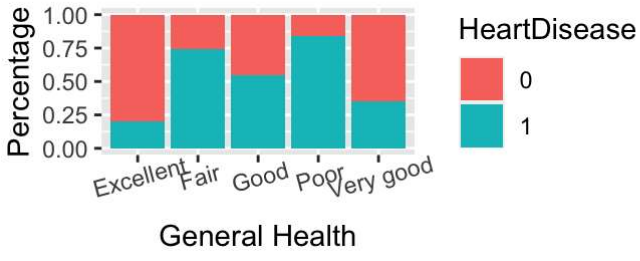
3 Percentage by Smoking status



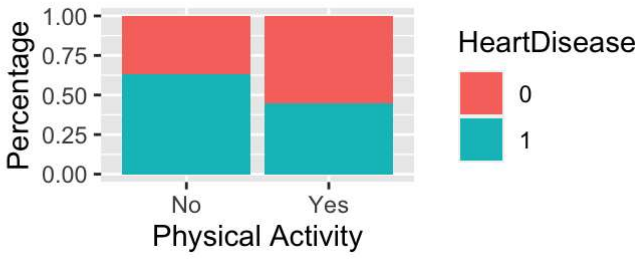
4 Percentage by Race



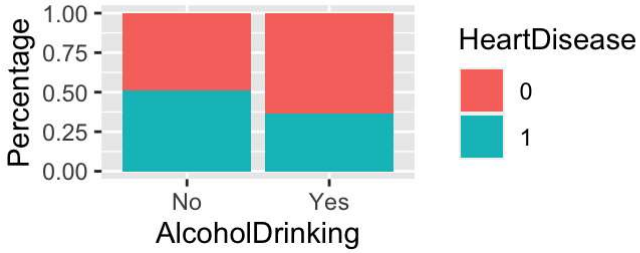
5 Percentage by General Health



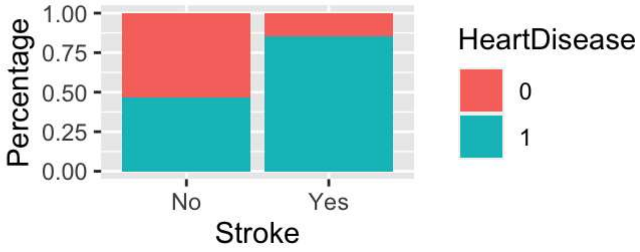
6 Percentage by Physical Activity



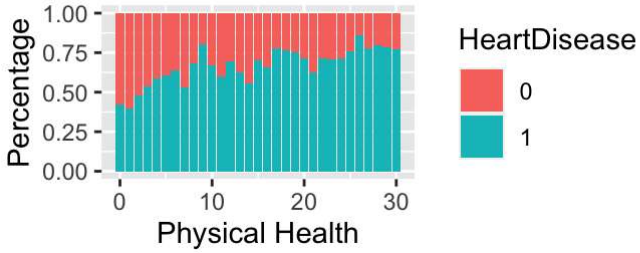
7 Percentage by AlcoholDrinking



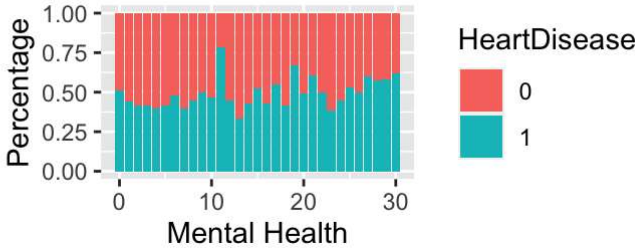
8 Percentage by Stroke



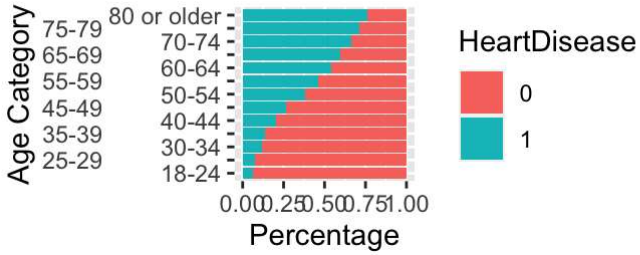
9 Percentage by Physical Health



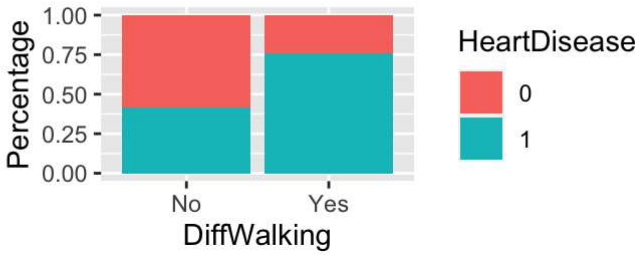
10 Percentage by Mental Health



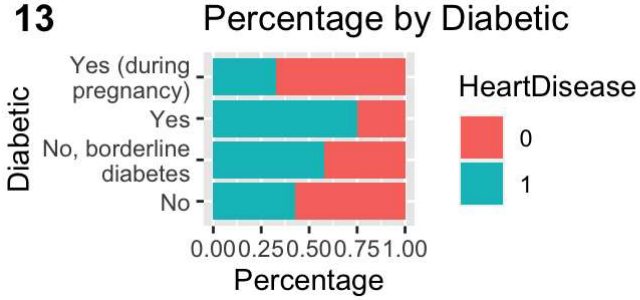
11 Percentage by Age Categor



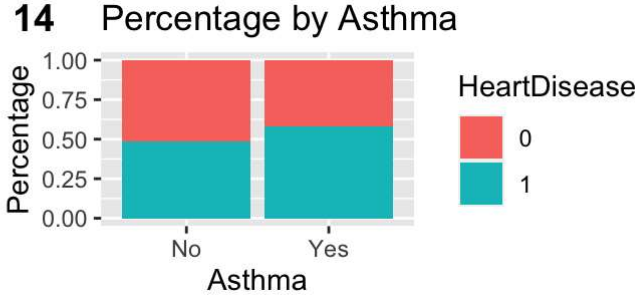
12 Percentage by DiffWalking



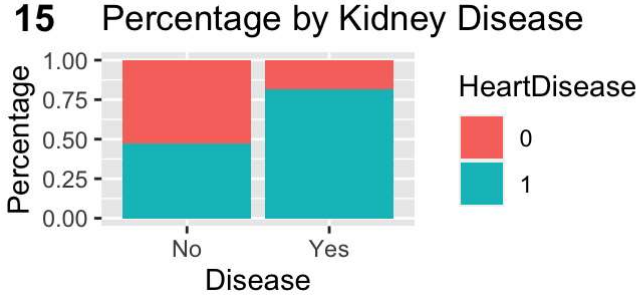
13 Percentage by Diabetic



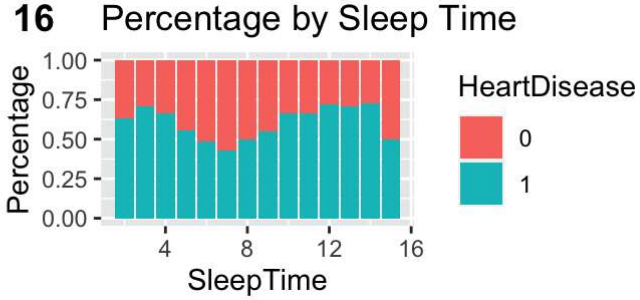
14 Percentage by Asthma



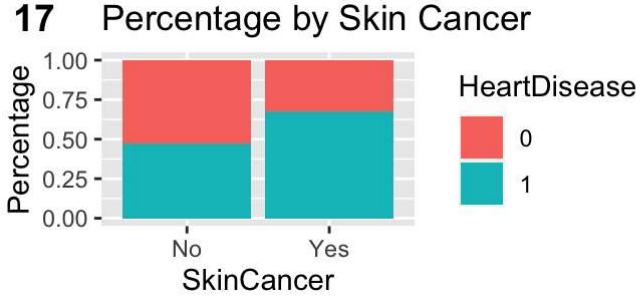
15 Percentage by Kidney Disease



16 Percentage by Sleep Time



17 Percentage by Skin Cancer



Training and Testing of Balanced Data

GLM Model

```
##
## Call:
## glm(formula = HeartDisease ~ ., family = "binomial", data = train_new)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1146  -0.7837   0.1257   0.8154   2.9070
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -1.586265   0.174324  -9.100  < 2e-16 ***
## BMI                            0.008858   0.002079   4.260 2.04e-05 ***
## SmokingYes                     0.374273   0.024454  15.305  < 2e-16 ***
## AlcoholDrinkingYes            -0.224357   0.052547  -4.270 1.96e-05 ***
## StrokeYes                      1.149378   0.050858  22.600  < 2e-16 ***
## PhysicalHealth                 0.004988   0.001572   3.174 0.001503 **
## MentalHealth                   0.007374   0.001609   4.583 4.58e-06 ***
## DiffWalkingYes                 0.250717   0.033836   7.410 1.26e-13 ***
## SexMale                        0.715922   0.024833  28.830  < 2e-16 ***
## AgeCategory25-29               0.231160   0.149116   1.550 0.121093
## AgeCategory30-34               0.515675   0.137120   3.761 0.000169 ***
## AgeCategory35-39               0.708676   0.131354   5.395 6.85e-08 ***
## AgeCategory40-44               1.040473   0.125394   8.298  < 2e-16 ***
## AgeCategory45-49               1.298693   0.121597  10.680  < 2e-16 ***
## AgeCategory50-54               1.713607   0.117215  14.619  < 2e-16 ***
## AgeCategory55-59               2.000883   0.115114  17.382  < 2e-16 ***
## AgeCategory60-64               2.290993   0.113909  20.112  < 2e-16 ***
## AgeCategory65-69               2.598987   0.113530  22.892  < 2e-16 ***
## AgeCategory70-74               2.887615   0.113775  25.380  < 2e-16 ***
## AgeCategory75-79               3.108909   0.115630  26.887  < 2e-16 ***
## AgeCategory80 or older         3.422842   0.115327  29.679  < 2e-16 ***
## RaceAsian                     -0.410917   0.137455  -2.989 0.002795 **
## RaceBlack                     -0.244620   0.102510  -2.386 0.017019 *
## RaceHispanic                  -0.124420   0.103103  -1.207 0.227527
## RaceOther                      0.028838   0.114067   0.253 0.800408
## RaceWhite                     -0.055861   0.092645  -0.603 0.546532
## DiabeticNo, borderline diabetes  0.126101   0.073058   1.726 0.084339 .
## DiabeticYes                    0.498183   0.031447  15.842  < 2e-16 ***
## DiabeticYes (during pregnancy)   0.123903   0.162222   0.764 0.444995
## PhysicalActivityYes            0.012614   0.028578   0.441 0.658942
## GenHealth                     -0.503296   0.014211 -35.415  < 2e-16 ***
## SleepTime                     -0.028799   0.008377  -3.438 0.000586 ***
## AsthmaYes                      0.290365   0.034557   8.403  < 2e-16 ***
## KidneyDiseaseYes               0.634617   0.052130  12.174  < 2e-16 ***
## SkinCancerYes                  0.101838   0.035821   2.843 0.004469 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 60247  on 43458  degrees of freedom
## Residual deviance: 42992  on 43424  degrees of freedom
## AIC: 43062
```

```
##
## Number of Fisher Scoring iterations: 5
```

Similar to our first model, we looked at the p-values associated with each predictor and determined that the following predictors are statistically significant at a significance level of 0.05:

The intercept is significant and indicates that the value of the response variable(HeartDisease) is -1.5849 when the unit of all the predictors is 0.

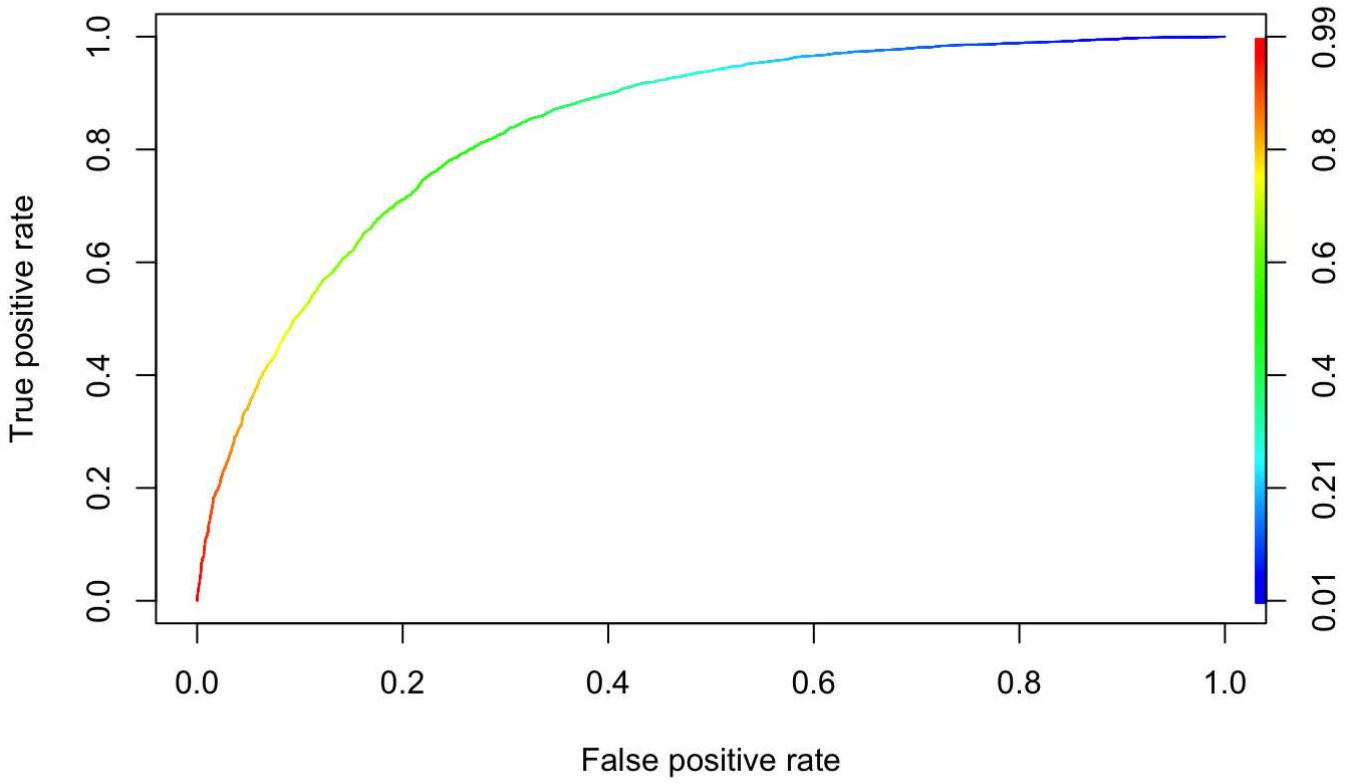
1. BMI: one unit increase in BMI is associated with an average change of 0.0083 of the response variable HeartDisease.
2. SMoking: one unit increase in Smoking(whether the respondent smokes) is associated with an average change of 0.3523 of the response variable HeartDisease.
3. AlcoholDrinking: one unit increase in AlcoholDrinking(whether the respondent drink alcohol) is associated with an average change of -0.2217 of the response variable HeartDisease.
4. Stroke: one unit increase in Stroke(whether the respondent had stroke before) is associated with an average change of 1.1986 of the response variable HeartDisease.
5. PhysicalHealth: one unit increase in PhysicalHealth is associated with an average change of 0.0048 of the response variable HeartDisease.
6. MentalHealth: one unit increase in MentalHealth is associated with an average change of 0.0056 of the response variable HeartDisease.
7. DiffWalking: one unit increase in DiffWalking(whether the respondent has difficulty in walking or climbing stairs) is associated with an average change of 0.2581 of the response variable HeartDisease.
8. Sex: one unit increase in Sex(Whether the respondent is male) is associated with an average change of 0.7040 of the response variable HeartDisease.
9. AgeCategory: 25-29: 0.3633 30-34: 0.7025 35-39: 0.7309 40-44: 1.0565 45-49: 1.3888 50-54: 1.7762 55-59: 2.0781 60-64: 2.4097 65_69: 2.6517 70-74: 2.9804 75-79: 3.1388 80 or older: 3.4104
10. Race(except RaceOther, RaceWhite, RaceHispanic): Asian:average change of -0.4698 on HeartDisease. Black: average change of -0.2218 on HeartDisease.
11. Diabetic(except DiabeticNo, borderline diabetes and DiabetesYes(during pregnancy)): one unit increase in Diabetic(whether the respondent has diabetes) is associated with an average change of 0.4811(have diabetes) of the response variable HeartDisease.
12. GenHealth: one unit increase in GenHealth is associated with an average change of -0.5149 of the response variable HeartDisease.
13. SleepTime: one unit increase in SleepTime is associated with an average change of -0.0333 of the response variable HeartDisease.
14. Asthma: one unit increase in Asthma(whether the respondent have asthma) is associated with an average change of 0.3184 of the response variable HeartDisease.
15. KidneyDisease: one unit increase in KidneyDisease(Whether the respondent have kidney disease) is associated with an average change of 0.4478 of the response variable HeartDisease.
16. SkinCancer: one unit increase in SkinCancer(Whether the respondent have skin cncer) is associated with an average change of 0.1358 of the response variable HeartDisease.

The AIC of the balanced model is significantly lower than the AIC of our first unbalanced model, 43,270 < 115,609. Thus, we considered the balanced model providing a better fit. [8]

Confusion Matrix

Following the same steps as we did for the unbalanced model, we developed the confusion matrix and calculated the f1 score of our balanced model, which is 0.7681. The f1 score of the balanced model is larger than that of the unbalanced model(0.7681 > 0.5266), indicating a higher precision and recall. Thus, we considered that the balanced model is better than the unbalanced model. [7]

```
##
##      FALSE TRUE
##   0  4067 1385
##   1  1144 4269
```



From visualizing the ROC curves of both models, we are able to determine that the curve of the balanced model has more area under curve and is closer to the upper left corner. Therefore, together with the results we obtained through the comparisons of f1 score and AIC, we considered that the balanced model is better than the unbalanced model.

Conclusion

Questions

To answer the question we asked:

1. What are the significant variables leading to heart disease? We can see there are multiple variables leading to heart disease according to the Balanced GLM model: BMI, Smoking, Alcohol Drinking, Stroke, Mental Health, Difficulty Walking, Male, Age, Asian, Diabetics, Gen Health, Sleep Time, Asthma, Kidney Disease, and Skin Cancer. The most significant variable according to the coefficient estimates is 80 years and older people are most likely to have heart disease. Also, we can see through the trend of coefficient estimate that as people get older, they are more likely to have heart disease.
2. How precise is our prediction? The precision of an estimate for the unbalanced model is 0.526616, and the precision of an estimate for the balanced model is 0.7605198 which is considered in the good range and much higher than the unbalanced one.

Analysis

Among all the variables, we can conclude BMI, stroke, and older age affect heart disease the most by comparing the absolute value of coefficient estimates.

Although through visualization, we cannot determine which curve is closer to the upper left corner; however, by comparing the values of AIC (115609 > 30852) and F1 scores (0.1705928 < 0.7681), we are still able to determine that the balanced model provides a better fit than the unbalanced model.

References

Dataset:

https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease (https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease)

Literature Reivew:

[1] “Cardiovascular Diseases (Cvds).” World Health Organization, World Health Organization, 11 June 2021, www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases- (cvds).

[2] Bansal, Manish. “Cardiovascular Disease and COVID-19.” Diabetes Metabolic Syndrome: Clinical Research Reviews, vol. 14, no. 3, May 2020, pp. 247–250., doi:https://doi.org/10.1016/ (doi:https://doi.org/10.1016/)

[3] Ridker, Paul M. “Clinical Application of C-Reactive Protein for Cardiovascular Disease Detection and Prevention.” Circulation, vol. 107, no. 3, 2003, pp. 363–369., doi:10.1161/01.cir.0000053730.47739.3c (doi:10.1161/01.cir.0000053730.47739.3c).

[4] Celermajer, David S., et al. “Cardiovascular Disease in the Developing World.” Journal of the American College of Cardiology, vol. 60, no. 14, Oct. 2012, pp. 1207–1216., doi:10.1016/j.jacc.2012.03.074 (doi:10.1016/j.jacc.2012.03.074).

[5] Getz, Godfrey S., and Catherine A. Reardon. “Nutrition and Cardiovascular Disease.” Arteriosclerosis, Thrombosis, and Vascular Biology, vol. 27, no. 12, 22 Oct. 2007, pp. 2499–2506., doi:10.1161/atvbaha.107.155853 (doi:10.1161/atvbaha.107.155853).

[6] “Training and Test Sets: Splitting Data; Machine Learning; Google Developers.” Google, 18 July 2022, https://developers.google.com/machine-learning/crash-course/training-and-test-sets/splitting-data?hl=en (https://developers.google.com/machine-learning/crash-course/training-and-test-sets/splitting-data?hl=en).

[7]“How to Interpret Glm Output in R (With Example).” Statology, 23 Feb. 2022, www.statology.org/interpret-glm-output-in-r. Accessed 7 Dec. 2022.

[8]“How to Calculate F1 Score in R (Including Example).” Statology, 8 Sept. 2021, www.statology.org/f1-score-in-r. Accessed 6 Dec. 2022.

[9]Chan, Carmen. “What Is a ROC Curve - How to Interpret ROC Curves.” Displayr, 23 Aug. 2022, www.displayr.com/what-is-a-roc-curve-how-to-interpret-it.