Data Description –

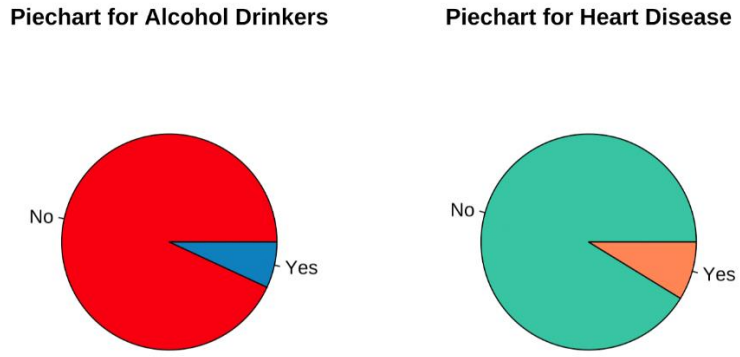
According to the Centers for Disease Control and Prevention (CDC), heart disease is one of the major causes of mortality in the United States for people of all races (African Americans, American Indians and Alaska Natives, and white people). High blood pressure, high cholesterol, and smoking are three important risk factors for heart disease that affect over half of all Americans (47 percent). Diabetic status, obesity (high BMI), lack of physical exercise, and excessive alcohol use are all important indicators. In healthcare, detecting and avoiding the variables that have the greatest influence on heart disease is critical. As a result of computational advancements, methods can be used to find "patterns" in data that can be used to forecast a patient's status.

Column Descriptions

- HeartDisease: Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI).
- BMI: Body Mass Index (BMI).
- Smoking: Have you smoked at least 100 cigarettes in your entire life?
- AlcoholDrinking: Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week
- Stroke: (Ever told) (you had) a stroke?
- PhysicalHealth: 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? (0-30 days).
- MentalHealth: Thinking about your mental health, for how many days during the past 30 days was your mental health not good? (0-30 days).
- DiffWalking: Do you have serious difficulty walking or climbing stairs?
- Sex: Are you male or female?
- AgeCategory: Fourteen-level age category. (Then calculated the mean)
- Race: Imputed race/ethnicity value.
- Diabetic: (Ever told) (you had) diabetes?
- PhysicalActivity: Adults who reported doing physical activity or exercise during the past 30 days other than their regular job.
- GenHealth: Would you say that in general your health is...
- SleepTime: On average, how many hours of sleep do you get in a 24-hour period?
- Asthma: (Ever told) (you had) asthma?
- KidneyDisease: Not including kidney stones, bladder infection or incontinence, were you ever told you had kidney disease?
- SkinCancer: (Ever told) (you had) skin cancer?

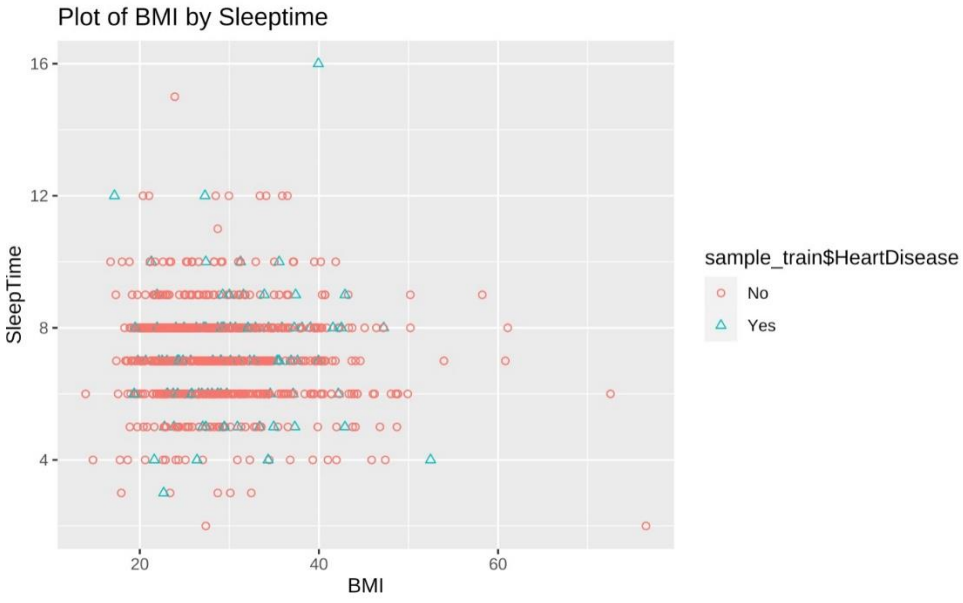
Exploratory data analysis

Research Question - Is it true that excessive alcohol use might contribute to heart disease?



Pie chart was used to present our problem statement, which states that people who consume excessive amounts of alcohol are not at risk for heart disease, based on the sample used, which is in fact dubious, as the data set contains more values for people who do not have heart disease than for those who do.

Research Question - Is it possible that being overweight and sleeping poorly induce heart disease?



The scatterplot above depicts the relationship between poor sleep and body mass index (BMI) on heart diseases in people over a long period of time. We may infer that there are a variety of reactions to a high body mass index and poor sleep, indicating that while these variables contribute considerably to heart disease, the results for this group are not exact due to a lack of sufficient data.

Regression Analysis Method

We used Regression Analysis Method here, when a continuous quantity needs to be predicted, regression can be useful. Numerical values are the numbers that regression analysis can predict. Like our use case necessitates the prediction of continuous numerical values we went with regression analysis to focus on the Numerical data present in our dataset for heart diseases, that included the BMI value which is nothing but the combination of the factorial values of weight and Height that is distinguishable as often interpreted that Obese people have a higher chance of developing heart diseases.

Similarly, we focused on three more such features including Mental Health, Physical Health which is quite similar to the BMI index and finally the Sleep Time of any given individual.

We then remove the categorical values to focus on the numerical values to find out and present a correlation matrix

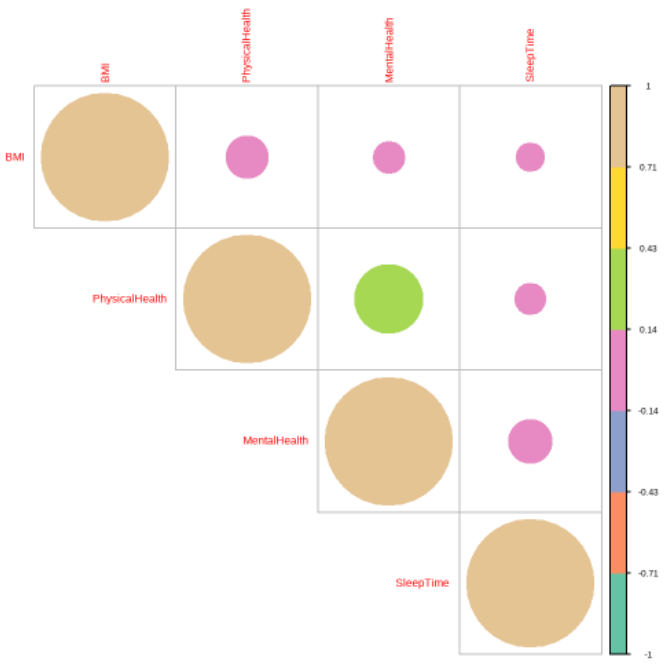
Correlation coefficients value for the numerical values

	BMI	PhysicalHealth	MentalHealth	SleepTime
BMI	1	0.11	0.064	-0.052
PhysicalHealth	0.11	1	0.288	-0.061
MentalHealth	0.064	0.288	1	-0.12
SleepTime	-0.052	-0.061	-0.12	1

Showing 1 to 4 of 4 entries Previous 1 Next

A correlation matrix is essentially a table which displays the correlation coefficients for distinct variables. The correlation between all possible pairings of values in a table is represented by the matrix.

The unlist function was used to remove the category values from the whole data set, and the cor function on the unlisted numeric data was used to construct a correlation matrix using the remaining numerical data set. We can see that the correlation values for the same columns, such as BMI, Physical Health, are both 1, indicating perfect correlation wherein the maximum correlation observed apart from that is 0.288 which is for Mental Health and Physical Health indicating not so strong correlation, also we can observe certain negative values indicating no correlation whatsoever.



In our case, per say if a person has higher BMI then the chances of him having a heart disease should be more, so keeping BMI as a dependent variable and changing the values of independent variables over, Physical Health, Mental Health and Sleep Time.

While we calculate the regression values using the lm function, we find that the mentioned variables depict very less correlation, probably close to none suggesting for a fact that according to our dataset the BMI value is not dependent on the other factors as generally anticipated.

```
##
## Call:
## lm(formula = HeartIs$BMI ~ HeartIs$PhysicalHealth + HeartIs$MentalHealth +
##     HeartIs$SleepTime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.245  -4.269  -0.983   3.137  66.867
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            29.287044   0.057755  507.09   <2e-16 ***
## HeartIs$PhysicalHealth  0.078617   0.001466   53.64   <2e-16 ***
## HeartIs$MentalHealth    0.024584   0.001473   16.69   <2e-16 ***
## HeartIs$SleepTime      -0.186352   0.007828  -23.81   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.308 on 319791 degrees of freedom
## Multiple R-squared:  0.01495,    Adjusted R-squared:  0.01494
## F-statistic:  1618 on 3 and 319791 DF,  p-value: < 2.2e-16
```

After finding out that there were too many outliers while using the regression function we used the mean square values to reduce the non-linearity of the data but it didn't make that much difference as the R2 and adjusted R2 square values were very low.

Multiple Regression Formula

```
## [1] "Multiple regression formula, y = 29.287 + 0.079 x1 + 0.025 x2 + -0.186 x3"
```

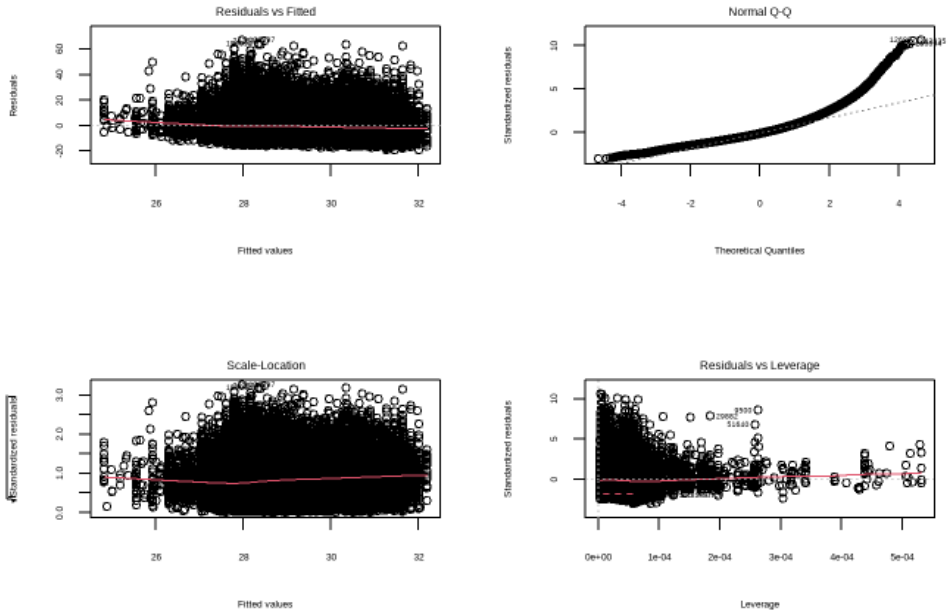
Correlation & Determination Coefficient values

	Correlation Coefficient	Determination Coefficient
BMI & Physical Health	0.11	0.01
BMI & Mental Health	0.06	0.00
BMI & Sleep Time	-0.05	0.00

Showing 1 to 3 of 3 entries

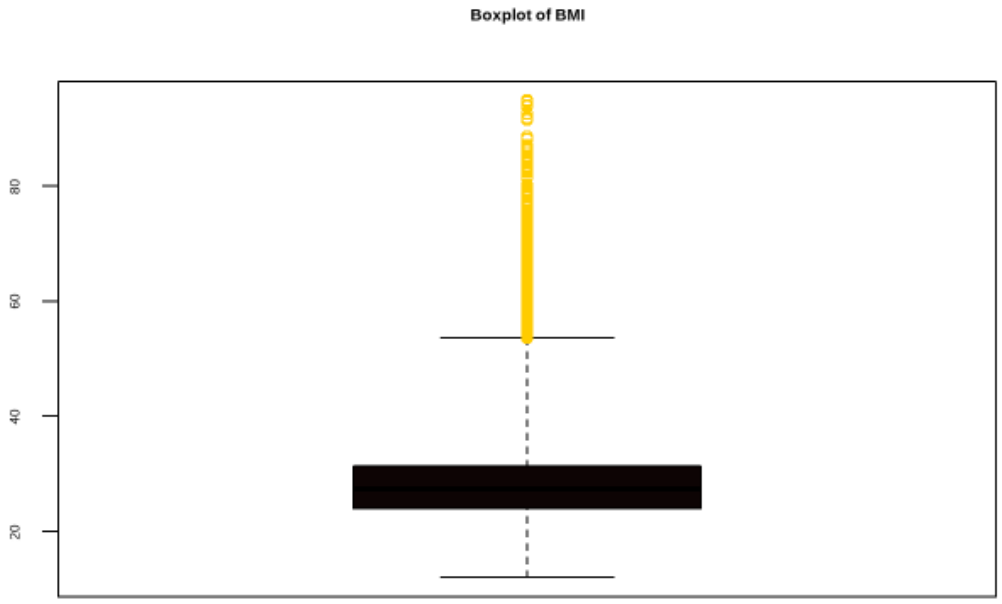
Previous 1 Next

The derived values were taken through the coefficients and the Multiple regression formula was presented, then further the values of the coefficient of the determination and determination with respect to them were put together in a tabular format, where we can observe that for BMI and the other factors the values are almost close to zero indicating that there is no correlation at all.




```
## HeartIs$PhysicalHealth   HeartIs$MentalHealth      HeartIs$SleepTime
##               1.091311               1.103007               1.015354
```

Here we used the vif function to check the multicollinearity of the given model, which is a great factor to compare the predictor values from one to another, which if we compare our cases are all below 5 that indicates that there is no multicollinearity for this particular case.



The Boxplot depicts a lot of outliers for our given variable, indicating that could be the reason for the attained values.

```
## Subset selection object
## Call: regsubsets.formula(HeartI$BMI ~ HeartI$SleepTime + HeartI$PhysicalHealth +
##     HeartI$MentalHealth, data = HeartI, nbest = 4)
## 3 Variables  (and intercept)
##                       Forced in Forced out
## HeartI$SleepTime          FALSE      FALSE
## HeartI$PhysicalHealth     FALSE      FALSE
## HeartI$MentalHealth       FALSE      FALSE
## 4 subsets of each size up to 3
## Selection Algorithm: exhaustive
##          HeartI$SleepTime HeartI$PhysicalHealth HeartI$MentalHealth
## 1  ( 1 ) " "              "*"                   " "
## 1  ( 2 ) " "              " "                   "*"
## 1  ( 3 ) "*"              " "                   " "
## 2  ( 1 ) "*"              "*"                   " "
## 2  ( 2 ) " "              "*"                   "*"
## 2  ( 3 ) "*"              " "                   "*"
## 3  ( 1 ) "*"              "*"                   "*"
```

To obtain the optimal predictor model values, the leaps library's regsubsets function was used, followed by the summary function. Physical Health is the best one predictor model, and Mental Health and Sleep time are followed by it.

Interpretation

So, looking at the conclusions, we could say that the considered numerical variable of BMI may not be the best fit to predict whether or not a person is suffering from Heart Disease as when we calculated the values, we were able to depict very low correlation values between the features, suggesting the same that the dependency is not clear.

This could also be due to the Dataset containing varied outliers, and not being a properly collected data set as it suggests that there are more cases for people not having a heart disease and they overpower the dataset into reaching a conclusion that probably isn't justified but Regression analysis on this dataset lead us to the conclusion that the given features are not correlated and that a given individuals chances of having a heart disease do not depend on the BMI index value.

Logistic Regression

So, we used and choose Logistic Regression Method here, which is similar to linear regression but it predicts True and False instead of prediction something continuous, it's an S Shape curve that lies between the values 0 and 1. For instance, we choose Heart Disease which gets predicted by BMI and Sleep Time to see if someone will develop Heart Disease based on these factors.

```
##
## Call:
## glm(formula = test_DS$HeartDisease ~ test_DS$BMI + test_DS$SleepTime,
##     family = binomial(link = "logit"), data = test_DS)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8756  -0.4319  -0.4094  -0.3912   2.4117
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -3.195440   0.076918 -41.543   <2e-16 ***
## test_DS$BMI        0.025986   0.001656  15.690   <2e-16 ***
## test_DS$SleepTime  0.009651   0.007924   1.218    0.223
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 55706  on 95937  degrees of freedom
## Residual deviance: 55474  on 95935  degrees of freedom
## AIC: 55480
##
## Number of Fisher Scoring iterations: 5
```

We divide our data into training and testing sets in a 70/30 split for Logistic Regression and make a confusion Matrix for the testing data set.

In our case for the confusion matrix, if the threshold probability of a person having Heart Disease is greater than 20% then it will classify it as "person with heart disease" else "person with not heart disease"

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    No   Yes
##        No  87746  8125
##        Yes    57    10
##
##               Accuracy : 0.9147
##                 95% CI : (0.9129, 0.9165)
##     No Information Rate : 0.9152
##     P-Value [Acc > NIR] : 0.7094
##
##                  Kappa : 0.0011
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.0012293
##             Specificity : 0.9993508
##          Pos Pred Value : 0.1492537
##          Neg Pred Value : 0.9152507
##              Prevalence : 0.0847943
##          Detection Rate : 0.0001042
##    Detection Prevalence : 0.0006984
##       Balanced Accuracy : 0.5002900
##
##        'Positive' Class : Yes
##
```

Rows are what the machine learning algo predicted and the columns are known truths or actual data.

So, we have “Does not have heart disease” and “Have heart disease” for both the rows and columns.

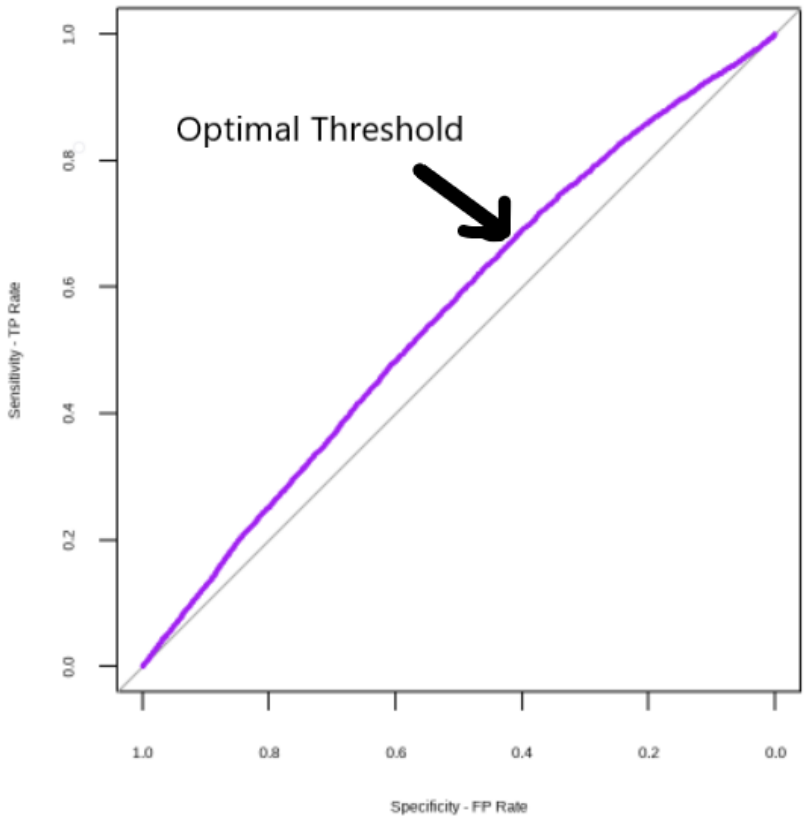
The bottom right corner is True Positives (10 in our case). These are patients that had heart disease that were correctly identified by the algorithm.

The top left corner is True Negatives (87746 in our case). These are patients that did not have heart disease that were correctly identified by the algorithm.

The top right corner is False Negatives (8125 in our case). These are patients that have heart disease, but the algorithm said they didn’t.

The bottom left corner is False Positives (57 in our case). These are patients that do not have heart disease, but the algorithm said they do.

So, the diagonal (True Positive and True Negative) tells how many times the samples are correctly classified. The numbers not on the diagonal (False Positive and False Negative) are samples the algorithm messed up.



Ideally, we want the purple line to go straight to the top and then right horizontally. That is a model that is capable of perfect predictions but it never happens. But what we do not want is this purple line hugging the gray diagonal line. Because that will be the model that will make no correct predictions. The diagonal line shows where the TP Rate is the same as the FP Rate. So, the ROC curve is a pretty good way to understand how your model is performing over specified thresholds.

We can use different confusion matrix based on their threshold to get the Sensitivity and Specificity values to plot this ROC curve for our testing data. Here, our ROC graph summaries all the of the confusion matrices that each threshold produce and we can notice the optimal threshold around the 0.7 Sensitivity and 0.4 Specificity.

Interpretation

So, looking at our confusion Matrix we can notice that our model for 20% probability threshold had 8175 False Negatives which we want to avoid here because these are patients that have heart disease, but the algorithm said they didn't. We can use different threshold range to predict a better model but looking at the ROC curve we can notice that it's not that far from the gray line

So, our False Negative, False Positive Values will still be there because these are the factors which determine Specificity and Sensitivity because when the data set was collected it was mostly of the people who didn't have heart disease. So rather than, trying to fix the model as we can already see the ROC curve optimal threshold is not that far from the gray line focusing more on the data collection can fix the issue.

RIDGE REGRESSION

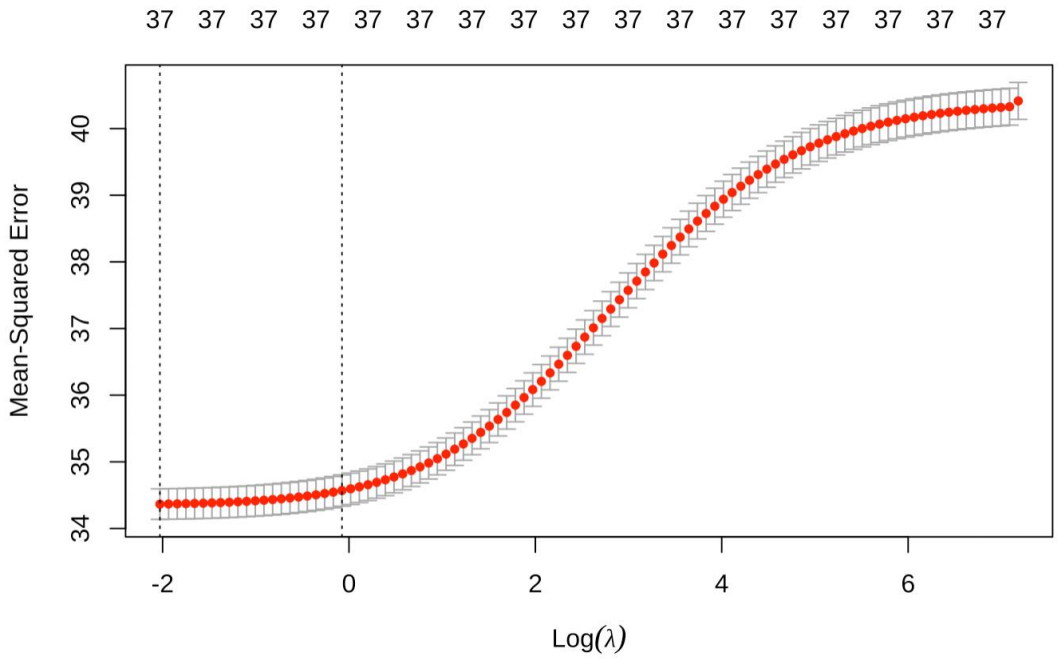
Ridge regression is a variation of linear regression in which the loss function is changed to reduce model complexity. This is accomplished by adding a penalty parameter that is equal to the square of the coefficients' magnitude.

Using GLMNET() function: Since glmnet function only takes matrix as input, we split the dataset into training and testing sets to convert the data into a matrix. Now, using glmnet function we obtained the plot of ridge regression.

```
#Spliting the data into a train and test set.
set.seed(123)
train_index <- createDataPartition(AVM4GrpData$HeartDisease, p = 0.6, list = FALSE)
sample_train <- AVM4GrpData[ train_index,]
sample_test <- AVM4GrpData[-train_index,]

sample_train_x <- model.matrix(sample_train$BMI~.,sample_train )[,-1]
sample_test_x <- model.matrix(sample_test$BMI~.,sample_test )[,-1]

sample_train_y <- sample_train$BMI
sample_test_y <- sample_test$BMI
```

```
#RIDGE REGRESSION
#finding best lambda using cross validation process
set.seed(123)
cv.ridge<- cv.glmnet(sample_train_x, sample_train_y,alpha=0, nfolds=10 )

#plot
plot(cv.ridge)
```

This plots the cross-validation curve (red dotted line) along with upper and lower standard deviation curves along the λ sequence (error bars). Two special values along the λ sequence are indicated by the vertical dotted lines.



Ridge keeps all variables and shrinks the coefficients towards zero. In the plot, as lambda decreases, the mean squared error decreases. Ridge includes all the variables in the model and the value of lambda selected is indicated by the vertical lines.

```
#Regression coefficients
coef(ridge_model.1se)
```

```
## 38 x 1 sparse Matrix of class "dgCMatrix"
##                                          s0
## (Intercept)                      27.500149471
## HeartDiseaseYes                   0.065094673
## SmokingYes                       -0.267594193
## AlcoholDrinkingYes               -0.605864694
## StrokeYes                        -0.531270439
## PhysicalHealth                    0.008001261
## MentalHealth                      0.002556761
## DiffWalkingYes                    1.840869479
## SexMale                           0.424091403
## AgeCategory25-29                  0.355269930
## AgeCategory30-34                  0.961927505
## AgeCategory35-39                  1.187109598
## AgeCategory40-44                  1.329243399
## AgeCategory45-49                  1.406983143
## AgeCategory50-54                  1.257740666
## AgeCategory55-59                  0.857053807
## AgeCategory60-64                  0.524541684
## AgeCategory65-69                  0.337948722
## AgeCategory70-74                 -0.198373669
## AgeCategory75-79                 -0.720770609
## AgeCategory80 or older           -2.023836224
## RaceAsian                        -2.490420530
## RaceBlack                         0.929888057
## RaceHispanic                      0.106725437
## RaceOther                         0.114164591
## RaceWhite                        -0.121846626
## DiabeticNo, borderline diabetes   1.390209805
## DiabeticYes                       2.487312275
## DiabeticYes (during pregnancy)    0.582142036
## PhysicalActivityYes              -1.143537572
## GenHealthFair                     1.671524393
## GenHealthGood                     1.645998893
## GenHealthPoor                     0.889928195
## GenHealthVery good                0.664080428
## SleepTime                        -0.052500582
## AsthmaYes                         0.916712698
## KidneyDiseaseYes                  0.186320614
## SkinCancerYes                    -0.246690950
```

RMSE INTERPRETATION

```
#making predictions on training data set
ridge_pred_train <- predict(ridge_model.1se, newx=sample_train_x)
ridge_train_rmse <- rmse(sample_train_y, ridge_pred_train)
ridge_train_rmse
```

```
## [1] 5.890161
```

```
#making predictions on testing data set
ridge_pred_test <- predict(ridge_model.1se, newx=sample_test_x)
ridge_test_rmse <- rmse(sample_test_y, ridge_pred_test)
ridge_test_rmse
```

```
## [1] 5.970784
```

The RMSE is the square root of the variance of the residuals. It indicates the absolute fit of the model to the data–how close the observed data points are to the model's predicted values.

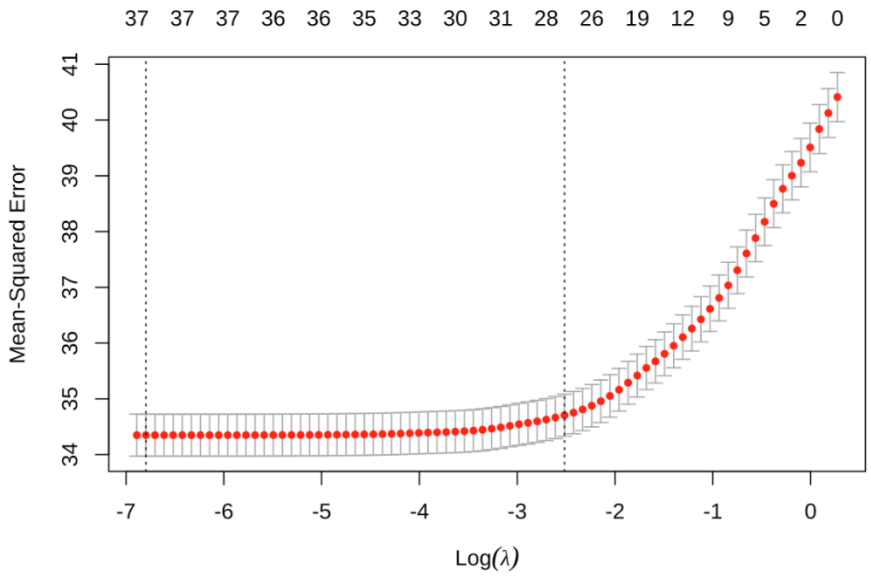
In Ridge model, for the training set the RMSE value is 5.890161 and for the testing set the RMSE value is 5.970784 which means there is less than 0.5 gap between the training and testing set which suggests that there is no overfitting in the model.

A low RMSE value indicates that the simulated and observed data are close to each other showing a better accuracy. Thus lower the RMSE better is model performance. The RMSE is a good measure for evaluating the performance of a model because RMSE is proportional to the observed mean.

LASSO REGRESSION

When there is multicollinearity in the data, we can use Lasso regression to create a regression model. Least squares regression seeks to produce coefficient estimates that minimize the sum of squared residuals (RSS).

```
#LASSO REGRESSION
#finding best lambda using cross validation process
set.seed(123)
cv.lasso<- cv.glmnet(sample_train_x, sample_train_y, nfolds=10 )

#plot
plot(cv.lasso)
```



To summarize, when lambda is zero, then the lasso model simply gives the least squares fit. Lasso can produce a model involving any number of variables. In contrast, ridge regression will always include all the variables in the model.

Using Lasso Regression, we only focused only on 1se values for lambda. Note that Lasso shrinks the coefficient estimates towards zero and it has the effect of setting variables exactly equal to

zero when lambda is large enough. In our model, 27 coefficients took the non-zero values whereas 10 variables were set equal to zero.

```
#Regression coefficients
coef(lasso_model.1se)
```

```
## 38 x 1 sparse Matrix of class "dgCMatrix"
##                                        s0
## (Intercept)                    27.289611209
## HeartDiseaseYes                 .
## SmokingYes                     -0.195912431
## AlcoholDrinkingYes             -0.428140278
## StrokeYes                      -0.269745426
## PhysicalHealth                  .
## MentalHealth                    .
## DiffWalkingYes                  2.055986798
## SexMale                         0.329600528
## AgeCategory25-29                .
## AgeCategory30-34                0.605930678
## AgeCategory35-39                0.871408845
## AgeCategory40-44                1.031361103
## AgeCategory45-49                1.113708383
## AgeCategory50-54                0.928228310
## AgeCategory55-59                0.424725675
## AgeCategory60-64                0.010018078
## AgeCategory65-69                .
## AgeCategory70-74               -0.452518153
## AgeCategory75-79               -1.086148833
## AgeCategory80 or older         -2.696485692
## RaceAsian                      -2.501438273
## RaceBlack                       0.899243492
## RaceHispanic                    .
## RaceOther                       .
## RaceWhite                      -0.007685893
## DiabeticNo, borderline diabetes  1.246494958
## DiabeticYes                     2.792081089
## DiabeticYes (during pregnancy)  .
## PhysicalActivityYes            -1.167997805
## GenHealthFair                   1.979092592
## GenHealthGood                   1.985682778
## GenHealthPoor                   0.894295141
## GenHealthVery good              0.855776312
## SleepTime                      -0.012123887
## AsthmaYes                       0.849183603
## KidneyDiseaseYes                .
## SkinCancerYes                  -0.043127973
```

RMSE INTERPRETATION

```
#making predictions on training data set
lasso_pred_train <- predict(lasso_model.1se, newx=sample_train_x)
lasso_train_rmse <- rmse(sample_train_y, lasso_pred_train)
lasso_train_rmse
```

```
## [1] 5.888511
```

```
#making predictions on testing data set
lasso_pred_test <- predict(lasso_model.1se, newx=sample_test_x)
lasso_test_rmse <- rmse(sample_test_y, lasso_pred_test)
lasso_test_rmse
```

```
## [1] 5.969952
```

In Lasso model, for the training set the RMSE value is 5.888511 and for the testing set the RMSE value is 5.969952 which means there is less than 0.5 gap between the training and testing set which suggests that there is no overfitting in the model.

Here both the values of RMSE in Ridge and Lasso are lower which indicates that the simulated and observed data are close to each other showing a better accuracy as lower the RMSE, better is model performance. In this case, since the RMSE value of Lasso model is less than the Ridge model I would say Lasso model performed better.

CONCLUSION AND LEARNINGS

By using three methods on our dataset, we were able to develop a good model by using Ridge and Lasso regression method as the RMSE value is extremely low which shows a better accuracy and good model performance.

Learnings from each method in this project:

- Learnt about multiple linear regression formula: Y=A+B1X1+B2X2+B3X3+…. BnXn and was able to obtain these values individually to present the multiple linear regression formula using R. Finally, we learnt how to use scatterplots to make conclusions on whether two variables have positive or negative correlation and understand the behavior of the dataset.
- By using the confusion matrix, understood the difference between true positive, true negatives, false positive and false negatives. Also, gained more knowledge on how it affects the accuracy, recall, precision, and Specificity. Understood the significance of Area under curve and ROC curve.
- Learnt that Ridge regression imposes a penalty on the coefficients to shrink them towards zero, but it doesn't set any coefficients to zero.
  In Lasso regression, some coefficients are allowed to be absolutely zero. As a result, lasso selects features and returns a final model with fewer parameters. The predictor variables that are less relevant for explaining the variation in the response variable are penalized in Lasso and Ridge regression. They allow us to concentrate on the strongest predictors to better understand how the response variable changes.

FUTURE CONSIDERATIONS & RECOMMENDATIONS

- While working on the dataset, we identified that the dataset we chose has patients who did not have a heart disease. We would ensure to pick a dataset which primarily focus on patients who have heart disease.
- Another consideration is that we can pick a dataset that is not biased. As the heart disease dataset we analyzed, had patients who did not have heart disease which was supported by false negatives being around 8000. Suggestion made by one our classmates was to pick a dataset that is unbiased.

Bibliography

1) Data source: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease?select=heart_2020_cleaned.csv

2) Kida, Y. (2020, December 17). Generalized linear models. Medium. Retrieved May 4, 2022, from https://towardsdatascience.com/generalized-linear-models-9cbf848bb8ab

3) Multiple Regression: Definition, Uses and 5 Examples. (2020). Indeed Career Guide. https://www.indeed.com/career-advice/career-development/multiple-regression