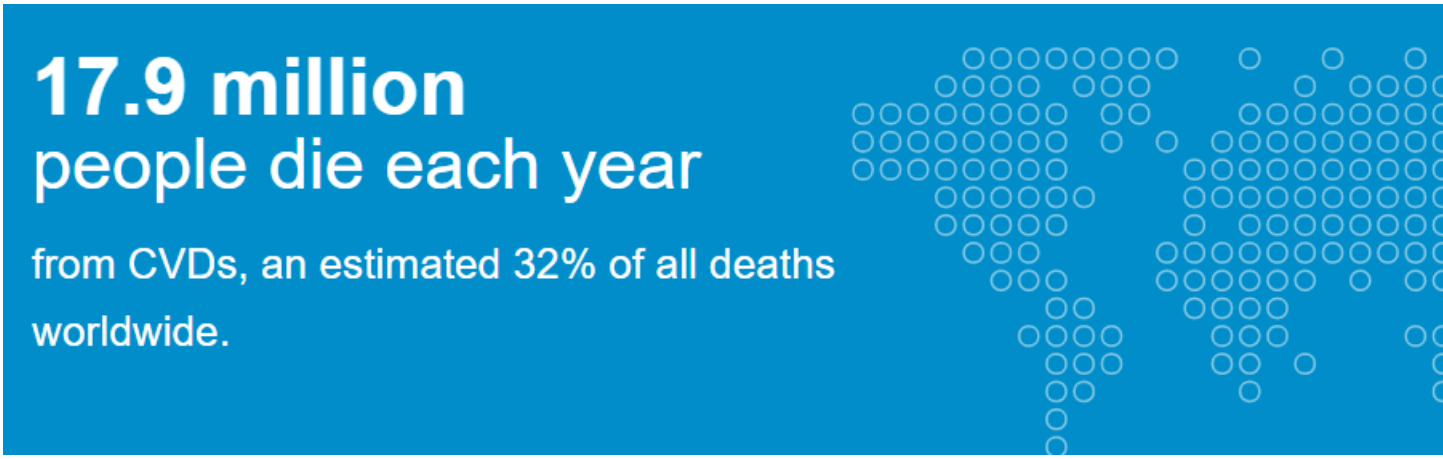
Langara.
THE COLLEGE OF HIGHER LEARNING

Heart disease prediction dataset

Why this Data Set:

According to World Health Organization statistics, cardiovascular diseases (CVDs) are the leading cause of death worldwide. Cardiovascular diseases are a group of diseases related to heart and blood vessels. It can be further classified into various subcategories of diseases where coronary heart disease is most common type of heart disease. The narrowing of the blood arteries resulting in stroke is known as coronary heart disease.



[i]In 2019, an estimated 17.9 million people died from cardiovascular diseases, accounting for 32% of all global deaths. 85 percent of these deaths were caused by a heart attack or a stroke. It is defined as a medical condition which can lead to a person's death. Since stroke has a serious impact on global health, we decided to look at the few parameters leading to it. Thus, early detection and prevention of this condition is important keeping in mind the global health conditions.

Project Overview:

During the training phase, a classification model will be constructed using several independent variables such as male, age, cigsPerDay, totChol, sysBP, and glucose, as well as a response variable (TenYearCHD class). The classification goal is to predict whether a patient has a 10-year risk of developing coronary heart disease (CHD).

Methodology:

Based on input parameters such as gender, age, education, and current health, we will attempt to predict whether a patient is likely to develop heart disease in the next ten years.

Data Overview:

The dataset is publically available on the Kaggle website, and it is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The analysis was performed on a heart dataset comprised of dimensions.

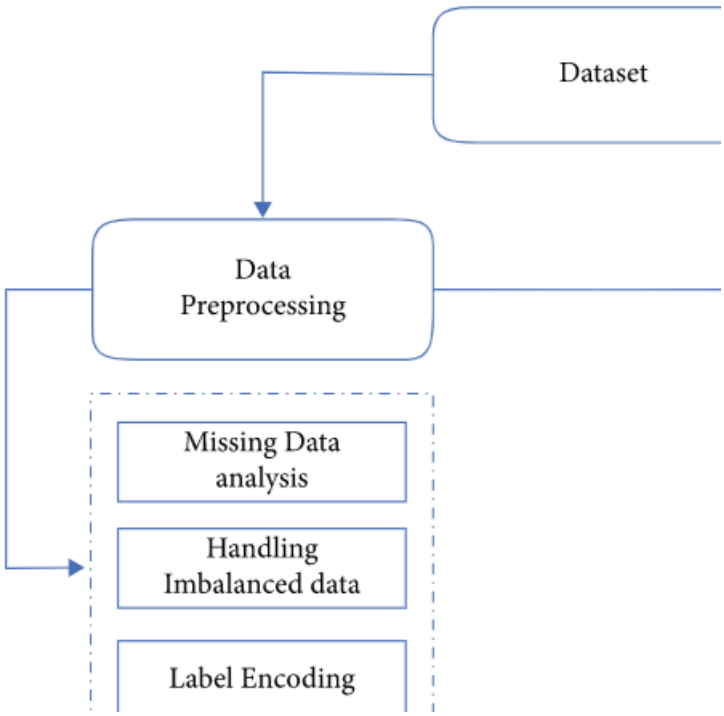
Rows	Columns
4238	16

In this section, we will predict whether a patient has a 10-year risk of developing Coronary Heart Disease (CHD). The variables with description and data type are as follows

Variable Name	Description	Kind of Variable- Type
Male	Male (1) and female (0) represent to gender	Nominal-Categorical
Age	Age of patient	Continuous-Numerical
Education	Education level of patient	Ordinal-Categorical
currentSmoker	Whether or not patient is current smoker	Nominal-Categorical
cigsPerDay	Number of cigarettes person smokes on average in one day	Continuous-Numerical
BPMeds	A patient was on blood pressure medication or no	Nominal-Categorical
prevalentStroke	A person previously had a stroke or not	Nominal-Categorical

prevalentHyp	A patient has a hypertension or not	Nominal-Categorical
diabetes	A patient has diabetes or not.	Nominal-Categorical
TenYearCHD	10 year risk of coronary Heart disease (1: yes, 0:no)	Nominal-Categorical
totchol	total cholesterol level of patient	Continuous-Numerical
sysBP	systolic blood pressure of patient	Continuous-Numerical
diaBP	diastolic blood pressure of patient	Continuous-Numerical
BMI	Body Mass Index	Continuous-Numerical
Heart Rate	heart rate of patient	Continuous-Numerical
Glucose	Glucose level of patient	Continuous-Numerical

Data Inspection and cleaning:



The figure illustrates the various steps in performed in data preprocessing.

- Summary statistics of the dataset is as follow:

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
male	1	4238	4.292119e-01	0.49502209	0.0	0.4115566	0.000000	0.00	1.0	1.00	0.28593231	-1.9186953	0.007604035
age	2	4238	4.958495e+01	8.57215993	49.0	49.3148585	10.378200	32.00	70.0	38.00	0.22798430	-0.9908326	0.131676959
education	3	4133	1.978950e+00	1.01979069	2.0	1.8488056	1.482600	1.00	4.0	3.00	0.68953172	-0.7140909	0.015862745
currentSmoker	4	4238	4.941010e-01	0.50002420	0.0	0.4926297	0.000000	0.00	1.0	1.00	0.02358933	-1.9999153	0.007680872
cigsPerDay	5	4209	9.003089e+00	11.92009359	0.0	6.8934402	0.000000	0.00	70.0	70.00	1.24702059	1.0188050	0.183734284
BPMeds	6	4185	2.962963e-02	0.16958357	0.0	0.0000000	0.000000	0.00	1.0	1.00	5.54603234	28.7653483	0.002621417
prevalentStroke	7	4238	5.899009e-03	0.07658717	0.0	0.0000000	0.000000	0.00	1.0	1.00	12.89992563	164.4468844	0.001176456
prevalentHyp	8	4238	3.105238e-01	0.46276270	0.0	0.2632665	0.000000	0.00	1.0	1.00	0.81869805	-1.3300472	0.007108498
diabetes	9	4238	2.571968e-02	0.15831643	0.0	0.0000000	0.000000	0.00	1.0	1.00	5.99013527	33.8897174	0.002431899
totChol	10	4188	2.367216e+02	44.59033432	234.0	234.6703461	42.995400	107.00	696.0	589.00	0.87079788	4.1218162	0.689028827
sysBP	11	4238	1.323524e+02	22.03809664	128.0	130.1008255	19.273800	83.50	295.0	211.50	1.14455148	2.1486317	0.338527230
diaBP	12	4238	8.289346e+01	11.91084960	82.0	82.1982606	11.119500	48.00	142.5	94.50	0.71359676	1.2721611	0.182962575
BMI	13	4219	2.580201e+01	4.08011106	25.4	25.5295529	3.691674	15.54	56.8	41.26	0.98127617	2.6495901	0.062815558
heartRate	14	4237	7.587892e+01	12.02659635	75.0	75.1972869	10.378200	44.00	143.0	99.00	0.64402548	0.9031539	0.184762361
glucose	15	3850	8.196675e+01	23.95999819	78.0	78.9795455	11.860800	40.00	394.0	354.00	6.20856108	58.5645530	0.386150335
TenYearCHD	16	4238	1.519585e-01	0.35902299	0.0	0.0651533	0.000000	0.00	1.0	1.00	1.93836837	1.7576868	0.005514953

- Here, n denotes the number of values in the dataset.
- The mean represents the average value of each column.
- The standard deviation tells us about the spread or dispersion of our data points around mean value.

- The median is the middle value of each column, and the mean is the average of each column.
- The trimmed explains the mean after trimming minimum and maximum value.
- The mean absolute deviation tells about the variation of each point from the mean value.
- The min represents the minimum value in each column.
- The max represents the maximum value in each column.
- The range represents the difference between maximum and minimum value.
- Skew represents the shape of each column's distribution, with positive values representing right-skewed distributions and negative values representing left-skewed distributions.
- Kurtosis denotes the apex of the data distribution. If the value in the column is negative, it indicates a flat data distribution, whereas a positive value indicates a peak distribution.

- ## Missing values:

The Missing value in dataset is as follow:

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP
missing	0	0	105	0	29	53	0	0	0	50	0	0

BMI	heartRate	glucose	TenYearCHD
19	1	388	0

We discovered missing values for few columns. The numerical columns “glucose”, “CigsPerDay”, “totChol” and “BMI” had missing value less than 9.2% of the total values, thus, the values were imputed with the median of the non-missing values. All the other categorical columns that had missing values less than 5% so that values were dropped from the dataset to reduce biasness.

- ## Duplicate check:

The given dataset has no duplicate values in the dataset.

```
> any(duplicated(df))
[1] FALSE
```

- ## Rename column:

The column “male” was renamed as “gender” and the column had values “1” for “male” and “0” for “female”.

```
> names(df)
 [1] "gender"          "age"             "education"       "currentSmoker"
 [5] "cigsPerDay"      "BPMeds"          "prevalentStroke" "prevalentHyp"
 [9] "diabetes"        "totChol"         "sysBP"           "diaBP"
[13] "BMI"             "heartRate"       "glucose"         "TenYearCHD"
```

- ## Transforming data:

After checking the structure of the dataset, it was observed that there were categorical columns with datatype “integer” and some of the columns were of datatype “num”.

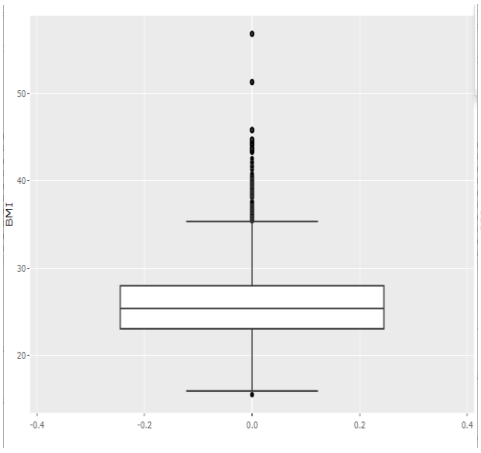
The categorical variables “gender”, “education”, “currentsmoker”, “BPMeds”, “prevalentStroke”, “prevalentHyp” and “diabetes

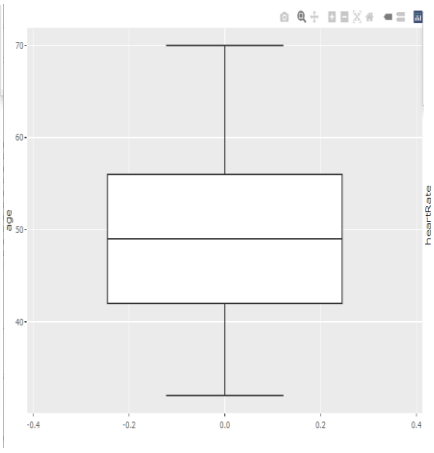
```
> str(df)
'data.frame':   4080 obs. of  16 variables:
 $ gender         : Factor w/ 2 levels "0","1": 2 1 2 1 1 1 1 1 2 2 ...
 $ age            : num  39 46 48 61 46 43 63 45 52 43 ...
 $ education      : Factor w/ 4 levels "1","2","3","4": 4 2 1 3 3 2 1 2 1 1 ...
 $ currentSmoker  : Factor w/ 2 levels "0","1": 1 1 2 2 2 1 1 2 1 2 ...
 $ cigsPerDay     : num  0 0 20 30 23 0 0 20 0 30 ...
 $ BPMeds         : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ prevalentStroke: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ prevalentHyp   : Factor w/ 2 levels "0","1": 1 1 1 2 1 2 1 1 2 2 ...
 $ diabetes       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ totChol        : num  195 250 245 225 285 228 205 313 260 225 ...
 $ sysBP          : num  106 121 128 150 130 ...
 $ diaBP          : num  70 81 80 95 84 110 71 71 89 107 ...
 $ BMI            : num  27 28.7 25.3 28.6 23.1 ...
 $ heartRate      : num  80 95 75 65 85 77 60 79 76 93 ...
 $ glucose        : num  77 76 70 103 85 99 85 78 79 88 ...
```

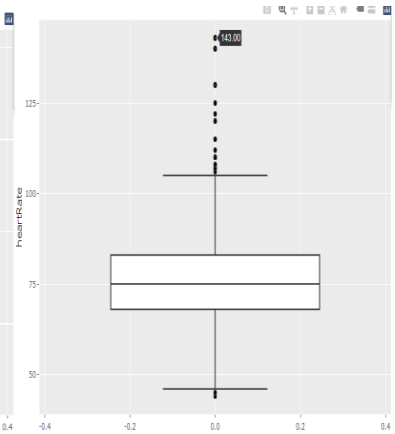
- ## Outlier Detection:

The boxplots were plotted for the numeric variables to check for any outliers in our data.

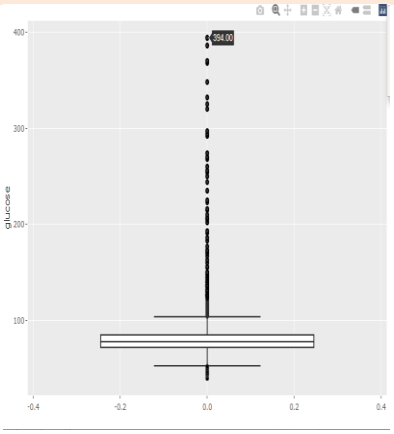
BMI	AGE	Heart Rate



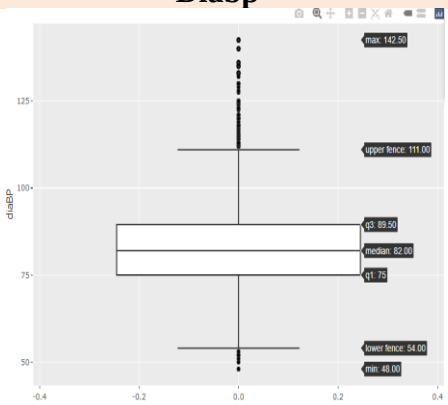




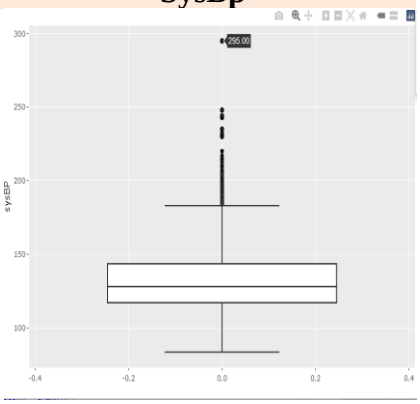
Glucose



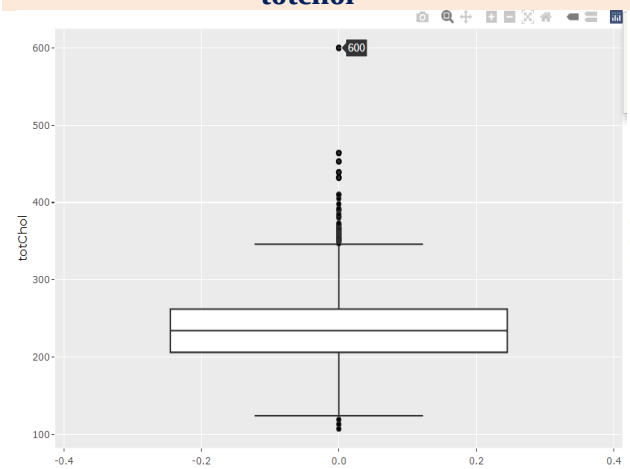
Diabp



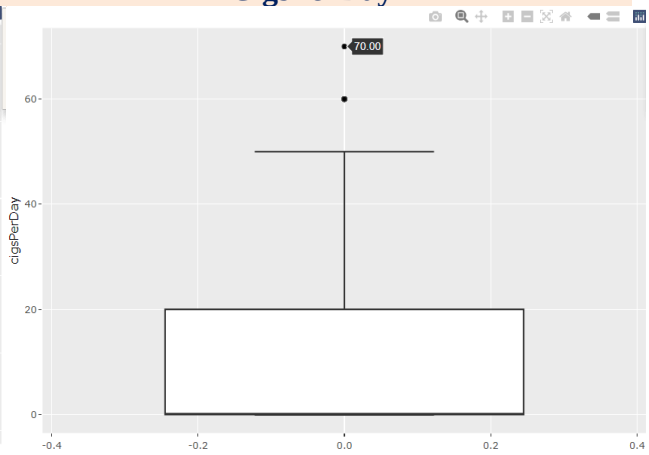
SysBp



totchol



CigsPerDay

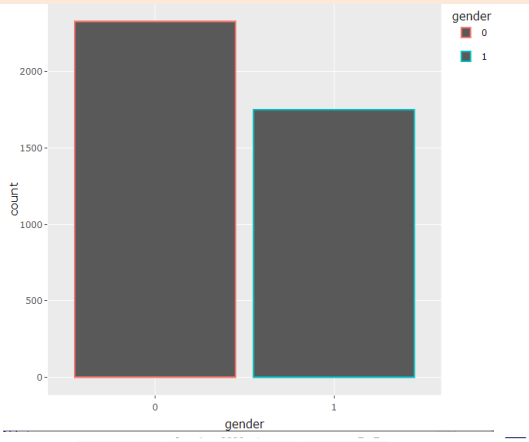
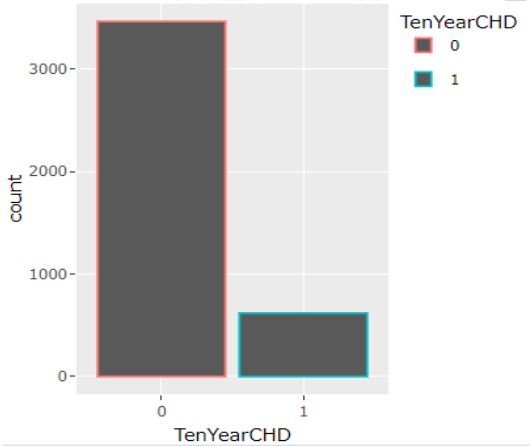


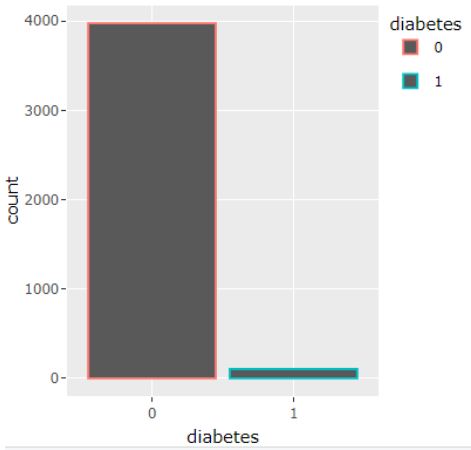
Observation:

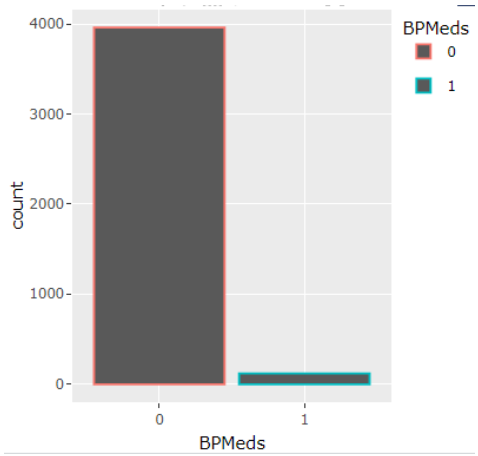
We observed that there were a few outliers for the variables “bmi”, “heartrate”, “Glucose”, “diaBP”, “sysBP”, “totchol” and “CigsPerDay”. We decided not to drop or treat the outliers as these values were significant for further analysis.

- ## Distribution of Categorical Variables:

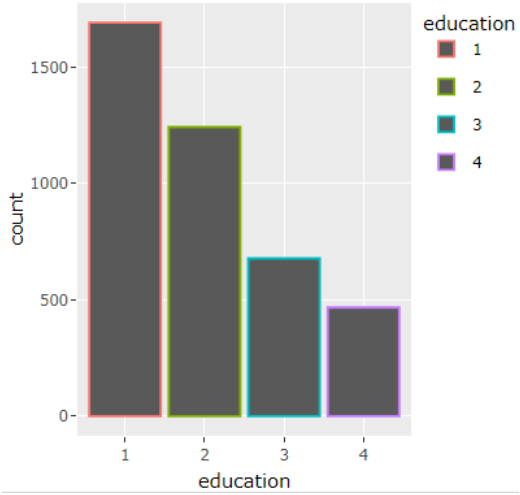
For the descriptive analysis of the categorical variables, bar graphs were plotted to check for the distribution of each categorical variable output.

Gender	TenYearCHD
	
Diabetes	**BPMeds**

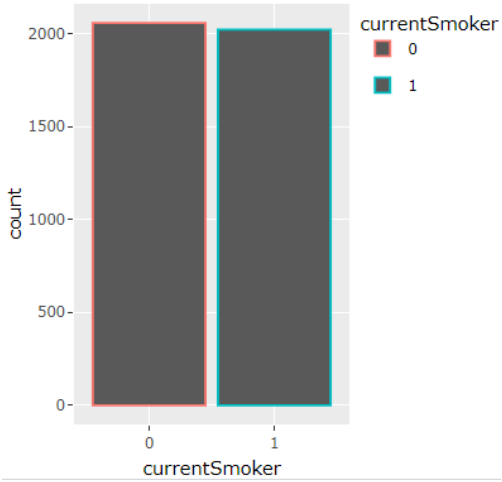




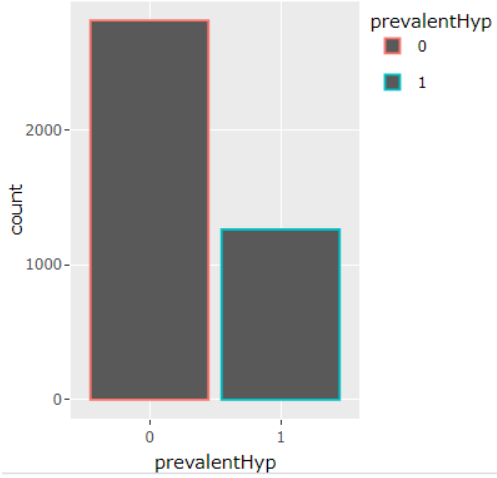
Education



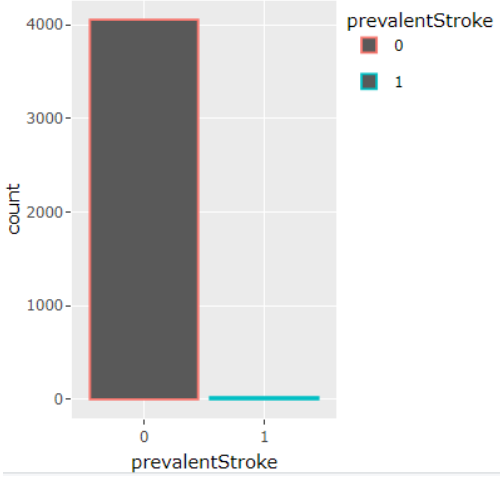
currentSmoker



prevalentHyp



prevalentStroke



Observation

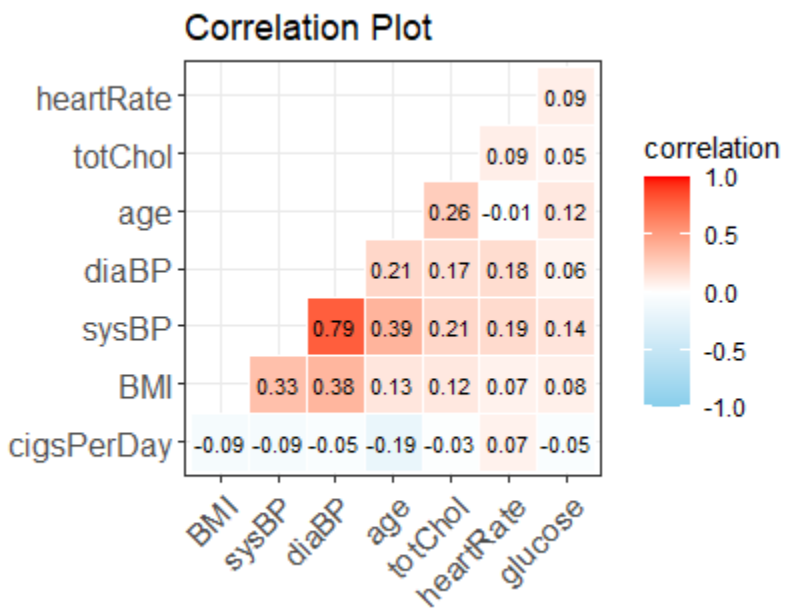
- The probability of having coronary heart disease in ten year is very less.

- Most of the participants were non-diabetic.
- Most of the participants were not taking BP medications.
- There were almost equal proportion of participants who were current smoker and were not.
- About 65% of participants do not have any prevalent hypertension history.
- Only a few participants had history of prevalent stroke.

Correlations between variables:

- ### Numeric variables

We plotted correlation plot to find if any variables were correlated with each other.



It can be observed that:

- "diaBP" is highly positively correlated with the "sysBP".
- "diaBP" and "sysBP" are positively correlated to the "BMI".
- "Age" is moderately positively correlated to "sysBP".
- "Age" is less positively correlated to the "totChol".
- "CigsPerDay" is slightly negatively correlated with all other variables except "heartrate".

- ### Categorical Variables

The association between the categorical variables was checked using Chi- Square test. We tried looping the variables to check the dependency of one variable with all other categorical variables.

```
[1] "1 .  education  and currentSmoker  are  Dependent"
[1] "2 .  education  and prevalentHyp  are  Dependent"
[1] "3 .  education  and diabetes  are  Dependent"
[1] "4 .  currentSmoker  and BPMeds  are  Dependent"
[1] "5 .  currentSmoker  and prevalentHyp  are  Dependent"
[1] "6 .  currentSmoker  and diabetes  are  Dependent"
[1] "7 .  BPMeds  and prevalentStroke  are  Dependent"
[1] "8 .  BPMeds  and prevalentHyp  are  Dependent"
[1] "9 .  BPMeds  and diabetes  are  Dependent"
[1] "10 .  prevalentStroke  and prevalentHyp  are  Dependent"
[1] "11 .  prevalentHyp  and diabetes  are  Dependent"
```

Initial Variable Selection

The numerical variables and categorical variables were tested to check if the variables are significant to be including in the regression model to detect the possibility of coronary heart disease in ten years.

- **Numeric variable**

For the numeric variables, we used t-test to check if the variable is statistically significant or not.

Variable	**t-value**	**p-value**
Age	-14.88	< 2.2e-16
CigsPerDay	-3.3382	0.0008819
totChol	-4.5989	4.932e-06
sysBP	-11.838	< 2.2e-16
DiaBP	-8.1974	1.037e-15
BMI	-4.3731	1.387e-05
heartrate	-1.3526	0.1825
glucose	-4.6366	4.268e-06

The p-value is less than all other variables except “heartrate”.

The hypothesis for t-test is:

Null hypothesis: The mean group difference is zero.

Alternate hypothesis: The mean group difference is not zero.

We can conclude that “heartrate” is not significant for any further analysis.

- **Categorical variables**

For the categorical variables, we used **Chi-Square test** to check if the variable is statistically significant or not. We ran a loop for testing each variable association with all other variables.

```
[1] "1 .  gender  and education  are  Dependent"
[1] "2 .  gender  and currentSmoker  are  Dependent"
[1] "3 .  gender  and BPMeds  are  Dependent"
[1] "4 .  education  and currentSmoker  are  Dependent"
[1] "5 .  education  and prevalentHyp  are  Dependent"
[1] "6 .  education  and diabetes  are  Dependent"
[1] "7 .  currentSmoker  and BPMeds  are  Dependent"
[1] "8 .  currentSmoker  and prevalentHyp  are  Dependent"
[1] "9 .  currentSmoker  and diabetes  are  Dependent"
[1] "10 .  BPMeds  and prevalentStroke  are  Dependent"
[1] "11 .  BPMeds  and prevalentHyp  are  Dependent"
[1] "12 .  BPMeds  and diabetes  are  Dependent"
[1] "13 .  prevalentStroke  and prevalentHyp  are  Dependent"
[1] "14 .  prevalentHyp  and diabetes  are  Dependent"
```

Building Logistic Regression

We ran a logistic regression model considering all the variables as explanatory variables except "heartrate" and "current smoker".

```
Call:
glm(formula = TenYearCHD ~ . - heartRate - BPMeds - currentSmoker,
    family = "binomial", data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9371  -0.5913  -0.4303  -0.2938   2.8382

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)      -7.9757073  0.6156979 -12.954  < 2e-16 ***
gender1           0.4691000  0.1024681   4.578 4.69e-06 ***
age               0.0609376  0.0063837   9.546  < 2e-16 ***
education2       -0.1996700  0.1169345  -1.708  0.08772 .
education3       -0.0986015  0.1404071  -0.702  0.48252
education4        0.0249472  0.1549579   0.161  0.87210
cigsPerDay        0.0209869  0.0039577   5.303 1.14e-07 ***
prevalentStroke1  0.9388131  0.4524137   2.075  0.03798 *
prevalentHyp1     0.2467163  0.1305626   1.890  0.05881 .
diabetes1         0.2124032  0.3005806   0.707  0.47979
totChol           0.0014881  0.0010762   1.383  0.16672
sysBP             0.0148746  0.0035818   4.153 3.28e-05 ***
diaBP            -0.0032919  0.0061032  -0.539  0.58963
BMI               0.0008848  0.0119840   0.074  0.94115
glucose           0.0061690  0.0021540   2.864  0.00418 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3466.6  on 4079  degrees of freedom
Residual deviance: 3080.4  on 4065  degrees of freedom
AIC: 3110.4
```

We can conclude from the output of the model that

- "gender", "age" and "cigsPerDay" are highly significant in predicting the probability of having coronary heart disease in ten years.
- "sysBp" is moderately significant in the model.
- "prevalentHyp", "totChol" and "glucose" are less significantly associated in predicting CHD.

Checking Multicollinearity

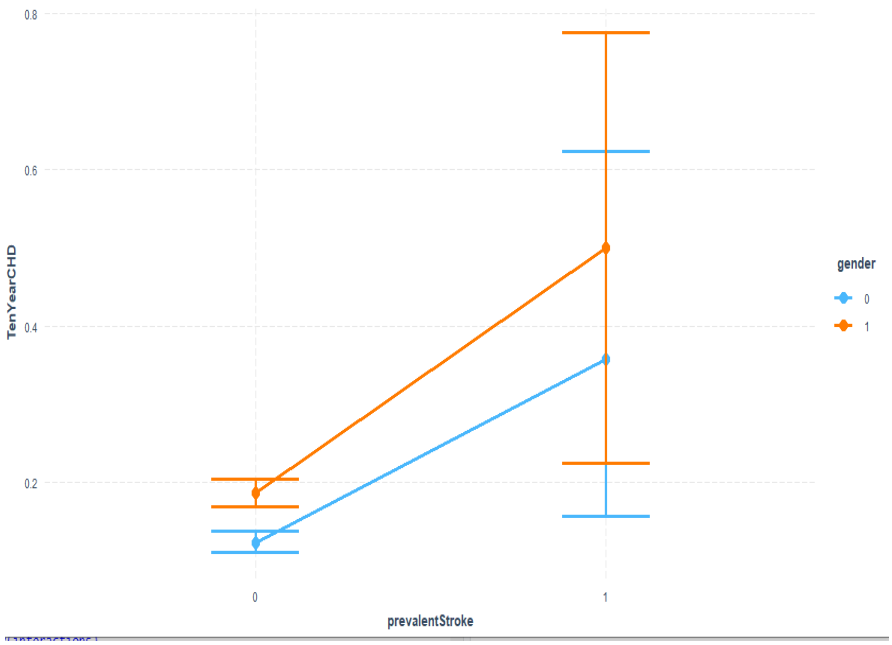
The model was then tested for any multicollinearity issue and Variance Inflation Factor (VIF) was used.

```
> vif(model_a)
                    GVIF Df GVIF^(1/(2*Df))
gender          1.228970  1        1.108589
age             1.313037  1        1.145878
education       1.111945  3        1.017842
cigsPerDay      1.237813  1        1.112570
prevalentStroke 1.010566  1        1.005269
prevalentHyp    1.979979  1        1.407117
diabetes        1.762691  1        1.327664
totChol         1.065455  1        1.032209
sysBP           3.539315  1        1.881307
diaBP           2.828180  1        1.681719
BMI             1.186770  1        1.089390
glucose         1.761301  1        1.327140
```

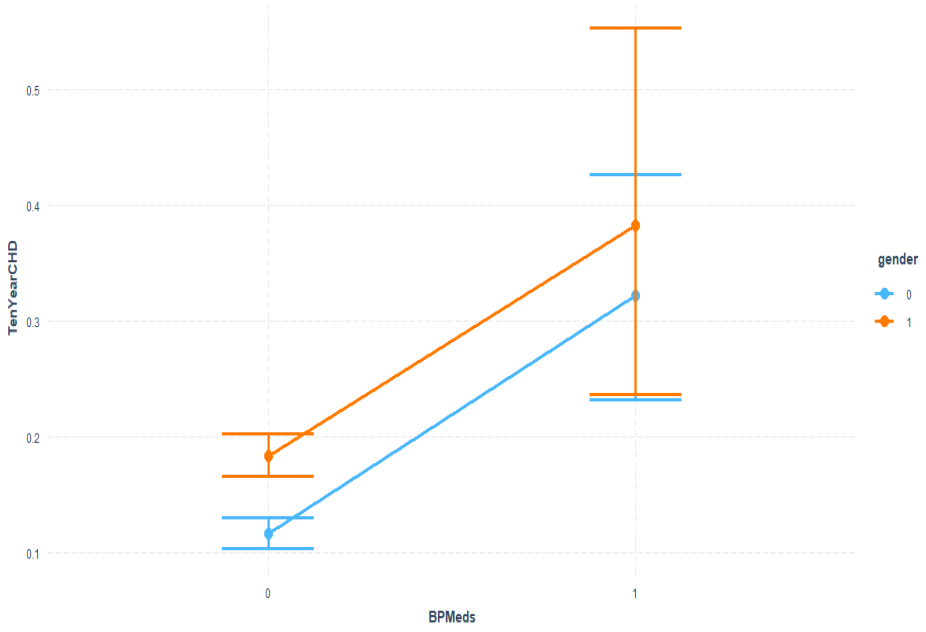
No vif value is greater than 5, so there is no multicollinearity in our model.

Interaction Analysis

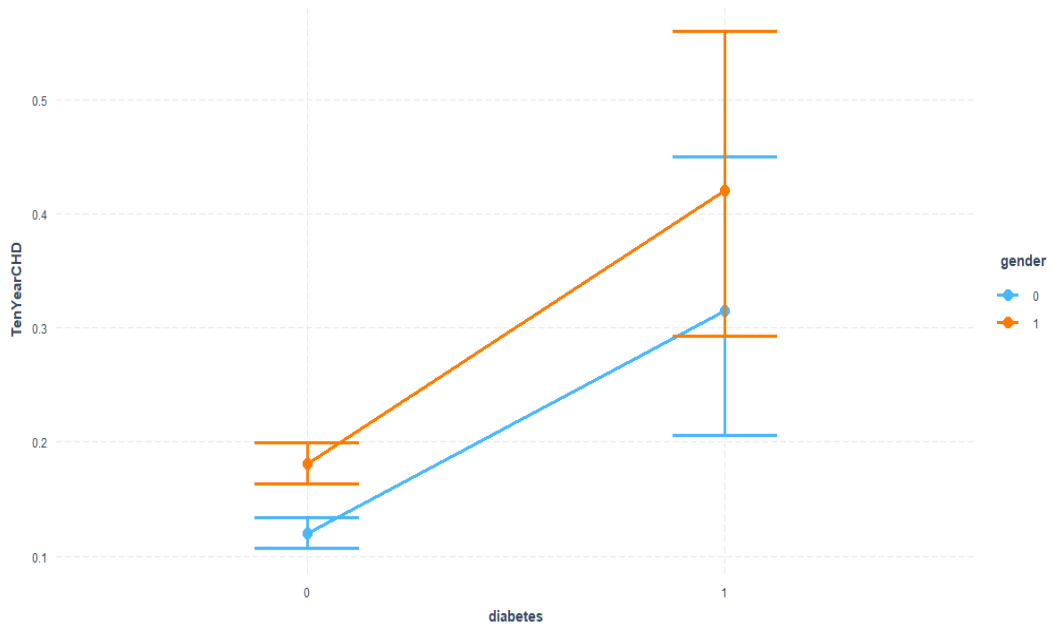
We analyzed various interaction plots, Here are some



As, we have observed from the interaction plot between Prevalent Stroke , TenYearCHD and Gender that there is interaction between these variables as the lines are not parallel and will intersect at one point to the left of the plot



In the interaction plot of TenYearCHD, BPMeds and gender we see parallel lines which says that there is no interaction between these variable.



Observation:

In the interaction plot between TenYearCHD, diabetes and gender. We see that there is interaction between these variables as the lines are not parallel and will intersect at one point to the left of the plot

Data Splitting

For Data Splitting used 80-20% splitting technique.

	Obsevation	**Variables**
Training Data	3264	16
Testing Data	816	16

Remark: Used splitting technique to check the Accuracy of the model.

Variable Selection Technique

We used backward selection technique

- Backward selection technique starts with all variables and removes those which are not significant to the model.
- At the end the leave only those variables that contribute substantially to the model.

Without Interaction	**With Interaction**

```
Start:  AIC=2516.02
TenYearCHD ~ (gender + age + education + currentSmoker + cigsPerDay +
    BPMeds + prevalentStroke + prevalentHyp + diabetes + totChol +
    sysBP + diaBP + BMI + heartRate + glucose) - heartRate -
    BPMeds - currentSmoker

                Df Deviance    AIC
- BMI            1   2486.1 2514.1
- education      3   2490.2 2514.2
- diabetes       1   2486.4 2514.4
- diaBP          1   2487.2 2515.2
- prevalentHyp   1   2487.3 2515.3
- totChol        1   2487.7 2515.7
<none>               2486.0 2516.0
- glucose        1   2489.5 2517.5
- prevalentStroke 1  2491.1 2519.1
- cigsPerDay     1   2502.1 2530.1
- sysBP          1   2506.0 2534.0
- gender         1   2508.4 2536.4
- age            1   2552.1 2580.1
```

```
Start:  AIC=2527.35
TenYearCHD ~ (gender + age + education + currentSmoker + cigsPerDay +
    BPMeds + prevalentStroke + prevalentHyp + diabetes + totChol +
    sysBP + diaBP + BMI + heartRate + glucose) - heartRate -
    BPMeds - currentSmoker + age * BMI + gender * prevalentStroke +
    cigsPerDay * diaBP + prevalentHyp * totChol + totChol * sysBP +
    sysBP * diaBP + sysBP * BMI

                     Df Deviance    AIC
- gender:prevalentStroke 1   2483.4 2525.4
- totChol:sysBP          1   2483.4 2525.4
- age:BMI                1   2483.4 2525.4
- prevalentHyp:totChol   1   2483.4 2525.4
- cigsPerDay:diaBP       1   2483.5 2525.5
- education              3   2487.6 2525.6
- sysBP:BMI              1   2483.6 2525.6
- diabetes               1   2483.7 2525.7
<none>                       2483.3 2527.3
- sysBP:diaBP            1   2485.4 2527.4
- glucose                1   2486.7 2528.7
```

```
Step:  AIC=2509.04
TenYearCHD ~ gender + age + cigsPerDay + prevalentStroke + totChol +
    sysBP + glucose

                Df Deviance    AIC
- totChol        1   2494.8 2508.8
<none>               2493.0 2509.0
- prevalentStroke 1  2498.2 2512.2
- glucose        1   2501.7 2515.7
- cigsPerDay     1   2509.1 2523.1
- gender         1   2517.9 2531.9
- sysBP          1   2546.8 2560.8
- age            1   2575.7 2589.7

Step:  AIC=2508.76
TenYearCHD ~ gender + age + cigsPerDay + prevalentStroke + sysBP +
    glucose

                Df Deviance    AIC
<none>               2494.8 2508.8
- prevalentStroke 1  2499.7 2511.7
- glucose        1   2503.3 2515.3
- cigsPerDay     1   2511.2 2523.2
- gender         1   2518.2 2530.2
- sysBP          1   2551.5 2563.5
- age            1   2582.1 2594.1
```

```
Step:  AIC=2512.2
TenYearCHD ~ gender + age + cigsPerDay + prevalentStroke + prevalentHyp +
    diabetes + totChol + sysBP + diaBP + glucose + sysBP:diaBP

                Df Deviance    AIC
- diabetes       1   2488.7 2510.7
- totChol        1   2490.2 2512.2
<none>               2488.2 2512.2
- sysBP:diaBP    1   2490.2 2512.2
- prevalentHyp   1   2490.4 2512.4
- glucose        1   2491.4 2513.4
- prevalentStroke 1  2493.2 2515.2
- cigsPerDay     1   2503.9 2525.9
- gender         1   2514.2 2536.2
- age            1   2562.3 2584.3

Step:  AIC=2510.66
TenYearCHD ~ gender + age + cigsPerDay + prevalentStroke + prevalentHyp +
    totChol + sysBP + diaBP + glucose + sysBP:diaBP

                Df Deviance    AIC
<none>               2488.7 2510.7
- totChol        1   2490.7 2510.7
- sysBP:diaBP    1   2490.7 2510.7
- prevalentHyp   1   2490.8 2510.8
- prevalentStroke 1  2493.7 2513.7
- glucose        1   2497.0 2517.0
- cigsPerDay     1   2504.2 2524.2
- gender         1   2515.0 2535.0
- age            1   2563.0 2583.0
```

It has been observed that the interaction plot has some extra variables. Moreover, we can see that the final model has AIC 2508.76 in without interaction term and with interaction terms has AIC 2510.66. which shows that Without interaction performs better which shows that interaction terms are not significant our further analysis

Model without Interaction Terms:

TenYearCHD ~ gender + age + cigsPerDay + prevalentStroke +

sysBP + glucose

Model with Interaction Terms:

TenYearCHD ~ gender + age + cigsPerDay + prevalentStroke + prevalentHyp +

totChol + sysBP + diaBP + glucose + sysBP:diaBP

Model Comparison

For model Comparison we can Check AIC and test using Log Likelihood Ratio Test. The result is as given bellow:

Without Interaction	With Interaction

```
> summary(model2)

Call:
glm(formula = TenYearCHD ~ gender + age + cigsPerDay + prevalentStroke +
    sysBP + glucose, family = "binomial", data = training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8207  -0.5950  -0.4421  -0.3046   2.7768

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -8.134396   0.437541 -18.591  < 2e-16 ***
gender1          0.537887   0.111162   4.839 1.31e-06 ***
age              0.061653   0.006712   9.185  < 2e-16 ***
cigsPerDay       0.018105   0.004410   4.105 4.04e-05 ***
prevalentStroke1 1.045174   0.455915   2.292  0.02188 *
sysBP            0.017049   0.002264   7.529 5.12e-14 ***
glucose          0.005665   0.001908   2.969  0.00299 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2778.1  on 3263  degrees of freedom
Residual deviance: 2494.8  on 3257  degrees of freedom
AIC: 2508.8

Number of Fisher Scoring iterations: 5
```

```
Call:
glm(formula = TenYearCHD ~ gender + age + cigsPerDay + prevalentStroke +
    prevalentHyp + totChol + sysBP + diaBP + glucose + sysBP:diaBP,
    family = "binomial", data = training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8084  -0.5936  -0.4392  -0.3059   2.7291

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -5.5345782  1.7399871  -3.181  0.00147 **
gender1          0.5823445  0.1138979   5.113 3.17e-07 ***
age              0.0588125  0.0069245   8.493  < 2e-16 ***
cigsPerDay       0.0176709  0.0044281   3.991 6.59e-05 ***
prevalentStroke1 1.0575654  0.4577264   2.310  0.02086 *
prevalentHyp1    0.2215401  0.1499973   1.477  0.13969
totChol          0.0017113  0.0012033   1.422  0.15499
sysBP            0.0013201  0.0120512   0.110  0.91277
diaBP           -0.0339495  0.0197077  -1.723  0.08495 .
glucose          0.0055954  0.0019092   2.931  0.00338 **
sysBP:diaBP      0.0001746  0.0001226   1.424  0.15434
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2778.1  on 3263  degrees of freedom
Residual deviance: 2488.7  on 3253  degrees of freedom
AIC: 2510.7

Number of Fisher Scoring iterations: 5
```

AIC

	Without Interaction Term	**With Interaction Term**
AIC Value	**2508.87**	**2510.7**

AIC is Akaike information criterion. Lower the AIC the better the model is. Hence it is cleared that model without interaction term has lower AIC value than With Interaction Term.

Log Likelihood Ratio Test

Log Likelihood is used to check which model is better in terms of significance. In this we have taken Model with interaction and Model without interaction terms.

H0: Reduced model is appropriate

Ha: Full Model is appropriate

```
Analysis of Deviance Table

Model 1: TenYearCHD ~ gender + age + cigsPerDay + prevalentStroke + sysBP +
    glucose
Model 2: TenYearCHD ~ gender + age + cigsPerDay + prevalentStroke + prevalentHyp +
    totChol + sysBP + diaBP + glucose + sysBP:diaBP
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      3257     2494.8
2      3253     2488.7  4   6.0954   0.1921
```

It shows that p-value is greater than 0.05 which means we accept null hypothesis says model without interaction terms is better.

Prediction Accuracy

To check the Accuracy of the data we checked classification report and ROC curve for both with and without interaction model. For this model is trained on training dataset and to check accuracy we used testing dataset.

Classification Report

It is the detailed table which we got from R console which tells us about the Accuracy, sensitivity and specificity of the model

Model without Interaction

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 692 115
         1   2   7

               Accuracy : 0.8566
                 95% CI : (0.8307, 0.88)
    No Information Rate : 0.8505
    P-Value [Acc > NIR] : 0.3327

                  Kappa : 0.0881

Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.99712
            Specificity : 0.05738
         Pos Pred Value : 0.85750
         Neg Pred Value : 0.77778
             Prevalence : 0.85049
         Detection Rate : 0.84804
   Detection Prevalence : 0.98897
      Balanced Accuracy : 0.52725

       'Positive' Class : 0
```

- The accuracy rate of the model is 85.66%. Whereas Misclassification rate is 14%
- The Sensitivity of detecting positive case is 99.7% that is people likely to have heart disease in 10 years and is predicted by the model.
- The specificity of detecting the negative case is 0.05

Model With Interaction

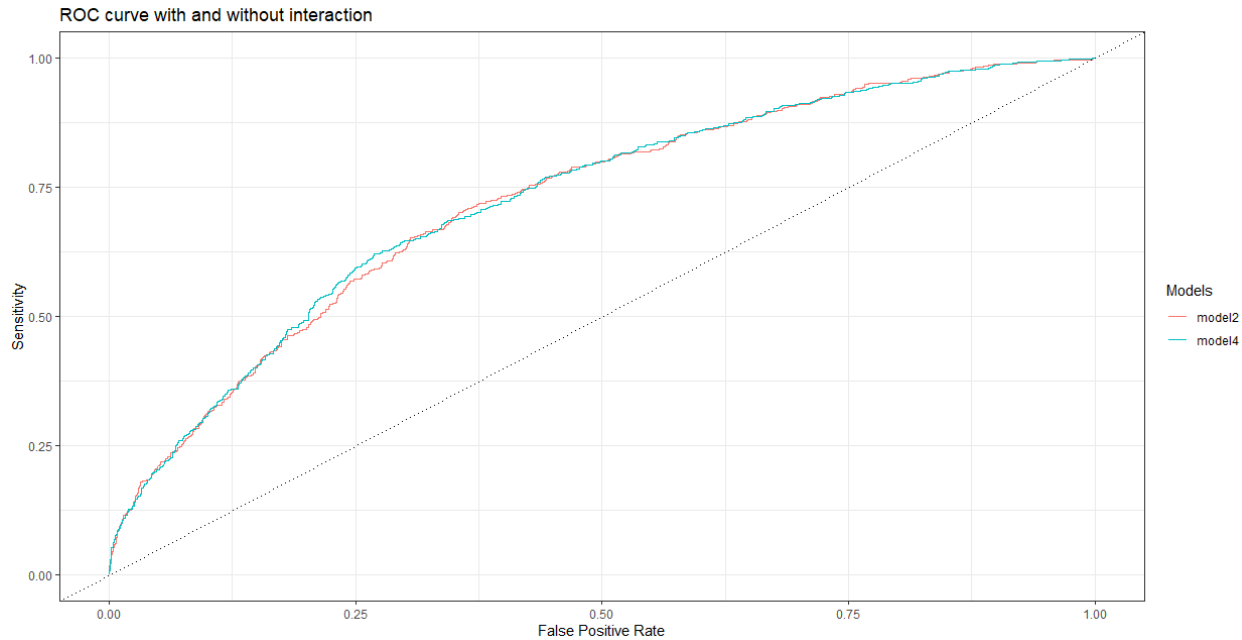
```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 691 115
         1   3   7

               Accuracy : 0.8554
                 95% CI : (0.8294, 0.8788)
    No Information Rate : 0.8505
    P-Value [Acc > NIR] : 0.3694

                  Kappa : 0.0853

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.99568
            Specificity : 0.05738
         Pos Pred Value : 0.85732
         Neg Pred Value : 0.70000
             Prevalence : 0.85049
         Detection Rate : 0.84681
   Detection Prevalence : 0.98775
      Balanced Accuracy : 0.52653

       'Positive' Class : 0
```

- The accuracy rate of the model is 85.54%.
- The Sensitivity of detecting positive case is 99.5% that is people likely to have heart disease in 10 years and is predicted by the model.
- The specificity of detecting the negative case is 0.05.

ROC Curve

ROC curve is the probability curve. It indicates how well the model can distinguish between classes. Where AUC measures the separability. Hence, Higher the AUC the better the model is for prediction

The ROC curve with and without interaction in one plot. As shown below



Observation:

Model2 – AUC value = 0.722

Model4- AUC Value = 0.724

From, plot it has been observed that AUC value is approximately same for model which is 72%. So, we consider our model without interaction plot is better as it will be less complex in analysis.

Hosmer-Lemeshow Test

Hosmer Lemeshow test is a goodness of fit test used in logistic regression. It tells us how the data fits the model.

H0: Reduced model is appropriate

Ha: Full model is appropriate

```
> h1 = hoslem.test(model2$y, fitted(model2), g=10)
> h1

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  model2$y, fitted(model2)
X-squared = 9.0688, df = 8, p-value = 0.3365
```

```
> h2 = hoslem.test(model4$y, fitted(model4), g=10)
> h2

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  model4$y, fitted(model4)
X-squared = 6.9392, df = 8, p-value = 0.5432
```

The p-value of both model is greater than 0.05, so we fail to reject null hypothesis which shows that our model is good fit to predict whether a patient has a 10-year risk of developing coronary heart disease (CHD).

Conclusion:

Coronary heart disease is a life-threatening disease which should be treated at the earliest. So, in our project report, we analyzed some of the factors that might cause it. As they are impacting global health therefore doing early detection and appropriate management can prevent any further loss. We have analyzed the various risk factors responsible for a person to develop CHD (coronary heart disease) in ten years thereby prevention of these can help in reducing the risk of developing the heart disease. The factors that influence if a person will have a coronary disease in 10 years are gender", "age" and "cigsPerDay. We got three models one with backward selection, one with interaction terms and one without interaction terms but after comparison we have chosen the model without any interaction terms as it is less complex and there is data imbalance in our model so the accuracy of both models is almost identical with a slight variation. The high sensitivity indicates that our model will correctly predict if a person will have coronary disease in 10 years and less specificity indicating that the model will not correctly predict if the person will not develop CHD in 10 years.

The insights found in the dataset are:

- "gender", "age" and "cigsPerDay" are highly significant in predicting the probability of having coronary heart disease in ten years.
- "sysBp" is moderately significant in the model.
- "prevalentHyp", "totChol" and "glucose" are less significantly associated in predicting CHD.
- The model without the interaction terms is better and less complex than the model with the interaction terms.
- Our model has accuracy rate of 86%.
- Our model is more sensitive than specific.
- There is a presence of data imbalance in our model.

Recommendation:

- The methodology used for the data collection could have been improvised by selecting more significant factors contributing if a person will develop coronary heart disease such as being obese, body fat percent, stress and having a family history of CHD.

References:

WHO. (2021). Cardiovascular diseases. *Health topics*.