HEART STROKE PREDICTION MODEL

INTRODUCTION

This report contains the Final Project (group) done for the course ALY6015 (Intermediate Analytics). In this project, we have worked on a dataset which contains the information of the patients such as their age, gender, health issues etc. The overall goal of this project was to create predictive models for variables **age** and **stroke** by using **Multiple Linear Regression** and **Logistic Linear Regression** methods.

- **Goal1: To predict whether a person has had a heart stroke or not.**
- **Goal2: To predict the age in which people get health issues.**

We have started by cleaning the data followed by performing EDA, feature selection and fitting/predicting the models. Below are the subsections of each of the tasks performed:

ANALYSIS

2.1 Reading CSV

Reading CSV file using **read.csv()** function and saved the file in a dataframe using **data.frame()** function. With help of this CSV, we will try to understand the patterns and create our multiple linear regression model.

```
> #importing and basic checks on data
> data = read.csv(file.choose(), header = T)
> data = data.frame(data)
> head(data)      #first few records
     id gender age hypertension heart_disease ever_married     work_type Residence_type
1  9046   Male  67            0             1          Yes       Private          Urban
2 51676 Female  61            0             0          Yes Self-employed          Rural
3 31112   Male  80            0             1          Yes       Private          Rural
4 60182 Female  49            0             0          Yes       Private          Urban
5  1665 Female  79            1             0          Yes Self-employed          Rural
6 56669   Male  81            0             0          Yes       Private          Urban
  avg_glucose_level  bmi  smoking_status stroke
1            228.69 36.6 formerly smoked      1
2            202.21  N/A    never smoked      1
3            105.92 32.5    never smoked      1
4            171.23 34.4          smokes      1
5            174.12   24    never smoked      1
6            186.21   29 formerly smoked      1
> tail(data)      #last few records
        id gender age hypertension heart_disease ever_married     work_type
5105 14180 Female  13            0             0           No      children
5106 18234 Female  80            1             0          Yes       Private
5107 44873 Female  81            0             0          Yes Self-employed
5108 19723 Female  35            0             0          Yes Self-employed
5109 37544   Male  51            0             0          Yes       Private
5110 44679 Female  44            0             0          Yes      Govt_job
     Residence_type avg_glucose_level  bmi  smoking_status stroke
5105          Rural            103.08 18.6         Unknown      0
5106          Urban             83.75  N/A    never smoked      0
5107          Urban            125.20   40    never smoked      0
5108          Rural             82.99 30.6    never smoked      0
5109          Rural            166.29 25.6 formerly smoked      0
5110          Urban             85.28 26.2         Unknown      0
> 
```

As it can be observed below, there are 5118 rows and 12 columns in the dataset which includes the records of patients. It has 4 integer, 2 numeric and 6-character data types variables.

```
> nrow(data)                    #total rows
[1] 5110
> ncol(data)                  #total columns
[1] 12
> names(data)                 #variable names
 [1] "id"                "gender"            "age"               "hypertension"
 [5] "heart_disease"     "ever_married"      "work_type"         "Residence_type"
 [9] "avg_glucose_level" "bmi"               "smoking_status"    "stroke"
> data.frame(sapply(data, class))   #columns data types
                  sapply.data..class.
id                            integer
gender                      character
age                           numeric
hypertension                  integer
heart_disease                 integer
ever_married                character
work_type                   character
Residence_type              character
avg_glucose_level             numeric
bmi                         character
smoking_status              character
stroke                        integer
>
```

2.1.1 Summary Statistics

Below displaying the total number of observations in each variable, their average, SD, median, min - max range etc.

```
> describe(data)   #descriptive statistics
                  vars    n     mean       sd   median  trimmed      mad   min      max    range  skew kurtosis     se
id                   1 5110 36517.83 21161.72 36932.00 36542.26 27413.27 67.00 72940.00 72873.00 -0.02    -1.21 296.03
gender*              2 5110     1.41     0.49     1.00     1.39     0.00  1.00     3.00     2.00  0.35    -1.86   0.01
age                  3 5110    43.23    22.61    45.00    43.61    26.69  0.08    82.00    81.92 -0.14    -0.99   0.32
hypertension         4 5110     0.10     0.30     0.00     0.00     0.00  0.00     1.00     1.00  2.71     5.37   0.00
heart_disease        5 5110     0.05     0.23     0.00     0.00     0.00  0.00     1.00     1.00  3.94    13.57   0.00
ever_married*        6 5110     1.66     0.48     2.00     1.70     0.00  1.00     2.00     1.00 -0.66    -1.57   0.01
work_type*           7 5110     3.50     1.28     4.00     3.62     0.00  1.00     5.00     4.00 -0.91    -0.49   0.02
Residence_type*      8 5110     1.51     0.50     2.00     1.51     0.00  1.00     2.00     1.00 -0.03    -2.00   0.01
avg_glucose_level    9 5110   106.15    45.28    91.88    97.85    26.06 55.12   271.74   216.62  1.57     1.68   0.63
bmi*                10 5110   172.19    88.96   158.00   163.08    74.13  1.00   419.00   418.00  0.97     0.87   1.24
smoking_status*     11 5110     2.59     1.09     2.00     2.61     1.48  1.00     4.00     3.00  0.08    -1.35   0.02
stroke              12 5110     0.05     0.22     0.00     0.00     0.00  0.00     1.00     1.00  4.19    15.57   0.00
>
```

2.2 Data Cleaning

Displaying all the distinct categorical variables present in the data.

- **Gender:** There is one ’Other’ category which we will check next
- **Smoking status:** It includes ’Unknown’ which we will further analyse
- **BMI:** We can see some NA values in this column which we will handle in the next section

```
> lapply(subset(data, select = c(gender, ever_married, work_type, Residence_type, smoking_status, bmi)), unique)
$gender
[1] "Male"   "Female" "Other"

$ever_married
[1] "Yes" "No"

$work_type
[1] "Private"       "Self-employed" "Govt_job"      "children"      "Never_worked"

$Residence_type
[1] "Urban" "Rural"

$smoking_status
[1] "formerly smoked" "never smoked"    "smokes"          "Unknown"

$bmi
  [1] "36.6" "N/A"  "32.5" "34.4" "24"   "29"   "27.4" "22.8" "24.2" "29.7" "36.8" "27.3" "28.2" "30.9" "37.5" "25.8"
 [26] "22.2" "30.5" "26.5" "33.7" "23.1" "32"   "29.9" "23.9" "28.5" "26.4" "20.2" "33.6" "38.6" "39.2" "27.7" "31.4"
 [51] "28.9" "28.1" "31.1" "21.7" "27"   "24.1" "45.9" "44.1" "22.9" "29.1" "32.3" "41.1" "25.6" "29.8" "26.3" "26.2"
```

2.2.1 Gender Column

There are **2994 - females** and **2115 - males** and **1- other** category in the gender section. As female category has the majority count, have replaced ’Other’ to ’Female’ and printed the revised count again.

```
> #Gender
> table(data$gender)      # unique variables in the gender column

Female   Male  Other
  2994   2115      1
> data$gender = ifelse(data$gender == "Other", "Female", data$gender)
> table(data$gender)

Female   Male
  2995   2115
```

2.2.2 BMI

Earlier we saw, BMI has some NA values and it is of character datatype. Converted it the numeric and replaced NA with mean value. Now we can see the average BMI is 28.89.

```
> data$bmi = as.numeric(data$bmi) # Convert BMI to numeric
Warning message:
NAs introduced by coercion
> data$bmi[is.na(data$bmi)] = mean(data$bmi,na.rm=TRUE)  # Replace N/A's in BMI column with mean
> summary(data$bmi)   #New summary of the variable BMI
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.30   23.80   28.40   28.89   32.80   97.60
```

2.2.3 Smoking status

We don’t have the smoking information of 1544 people. As we have only three other categories, based on the probability of these three, replaced ’Unknown’ by the other three variables according to their weightage.

```
> table(data$smoking_status)

formerly smoked    never smoked          smokes         Unknown
            885            1892             789            1544
>
```

Now we have counts as mentioned below:

```
> prob.F = 885 / (885 + 1892 + 789)
> prob.N = 1892 / (885 + 1892 + 789)
> prob.S = 789 / (885 + 1892 + 789)
> # Replacing 'Unknown' in smoking_status by the other 3 variables according to their proportions we calculated
> data1$rand = runif(nrow(data1))
> data1 = data1%>%mutate(Probability = ifelse(rand <= prob.F, "formerly smoked",
+                                      ifelse(rand <= (prob.F+prob.N), "never smoked", ifelse(rand <= 1, "smokes", "Check"))))
> data1 = data1%>%mutate(smoking.status = ifelse(smoking_status == "Unknown", Probability, smoking_status))
> table(data1$smoking.status)   #new smoking status

formerly smoked    never smoked          smokes
           1277            2725            1108
```

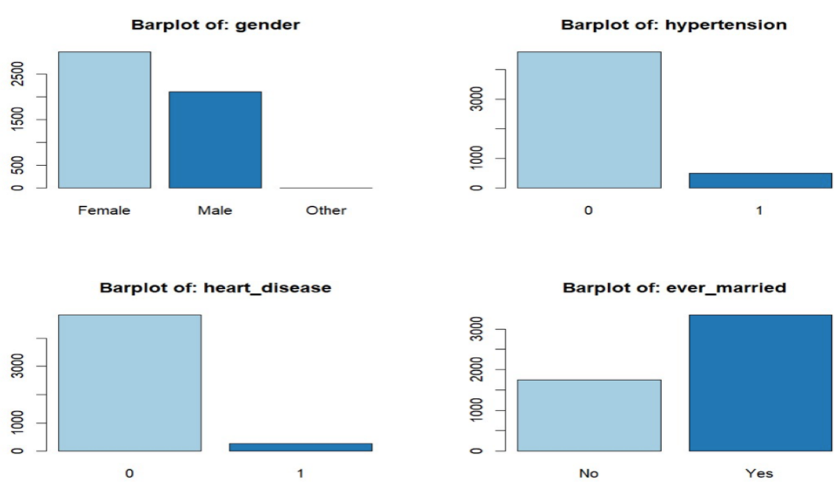
2.2.4 Removing columns

Removed columns which are not required for the analysis and below displaying the cleaned data which we will use for further analysis.

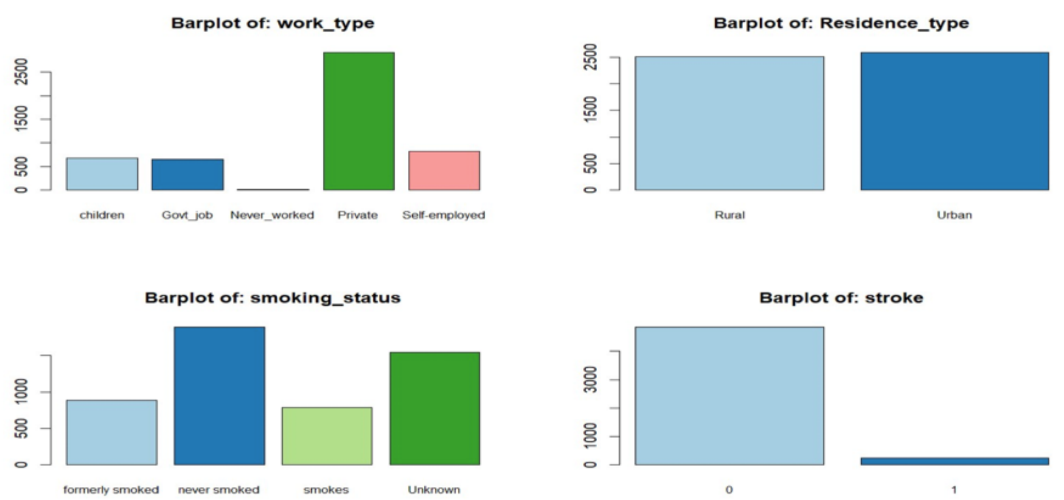
```
> data1 = subset(data1, select = -c(rand,Probability,smoking_status, id))
> view(data1)
```

2.3 Exploratory Data Analysis

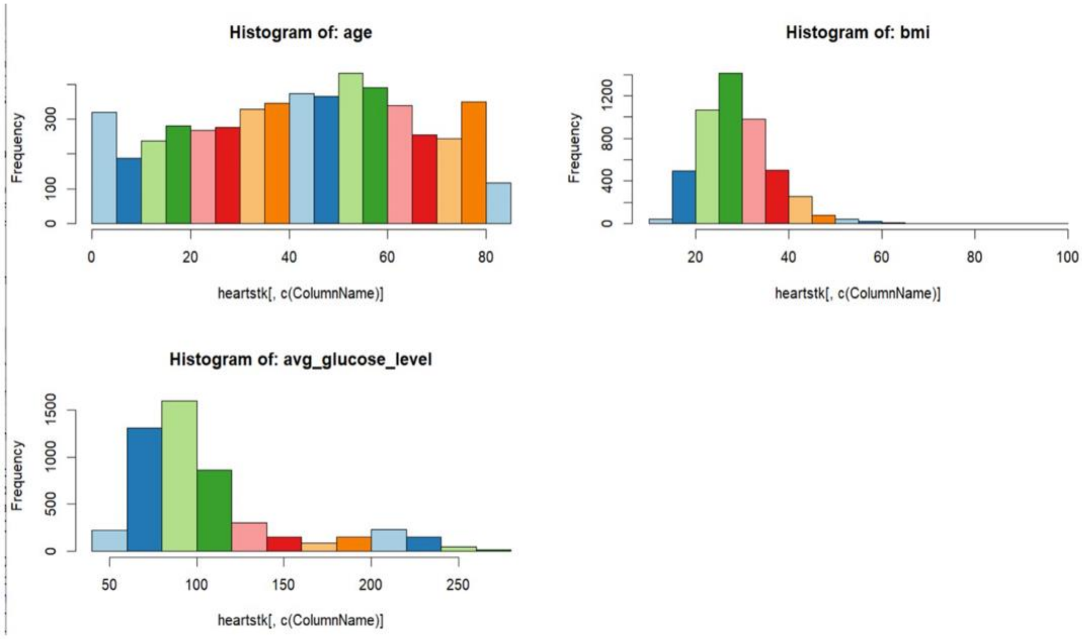
- **Gender**: Number of Female patients are more than Male patients. Initially, there was one more category named 'Other' with 1 record, we have added that to Female section since majority of the patients are females.
- **Hypertension**: Hypertension (High blood pressure) is a condition that eventually causes health problems, such as heart disease. Here, patients without hypertension are way more than the patients with hypertension.
- **Heart Disease**: The count of patients without heart disease is quite similar to the patients without hypertension. But only half of the people that are having hypertension, are also heart patients. Half of them does not have the heart problem.
- **Ever Married**: Majority of the people comes under 'ever married' category.



- **Patient work type**: A very small number of patients that have never worked. The number of patients that are either children or doing govt. job or self-employed are quite similar. But one thing to notice here, majority of the patients work at private companies. The type of work can be factor which is affecting their health.
- **Stroke**: The number of patients who have not faced any strokes is way greater than the people who faced strokes. This also indicates that more people are hypertension patients than stroke or heart problem.
- **Smoking status**: Earlier, the unknown data was randomly added to the three categories based on their weightage. Majority of the patients never smoked. A similar number of people are either current smokers or formerly smoked.
- **Residence Type**: Almost same number of patients lives Rural and urban area.



- **Age**: The distribution is close to a normal distribution with the mean of 43.22. Based on the average age information and the graph below, majority of the patients are around their 40s.
- **Average glucose:** With a mean of 106.14, the average glucose levels of the patients is right skewed.
- **BMI:** The data is right skewed with a mean of 28.89. All the NAs were replaced with the mean value in the data cleaning section.



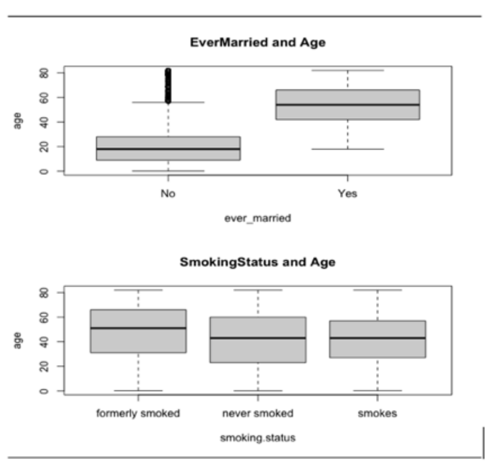
2.3.1 Boxplots & Violin-plots

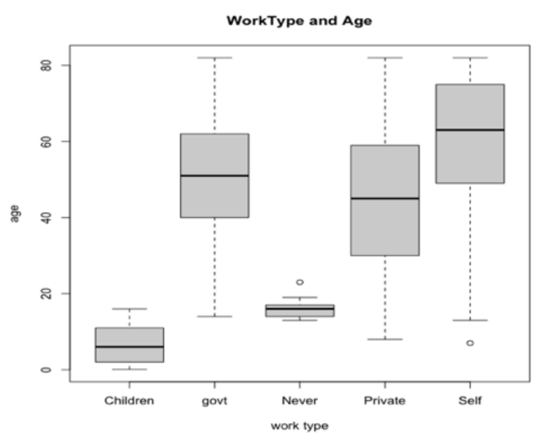
- **Ever married and Smoking status of patients by their age**

The average age of patients who are ever married is 50. There are outliers in the No category of ever married. Average age of all the categories in smoking status is similar.

- **Work type**

Most of the self-employed people are older than other categories (60+). Private companies employees' average age is around 40 and the average age of people who are govt employee is around 50.





- **Age of Patients with and Without Strokes**

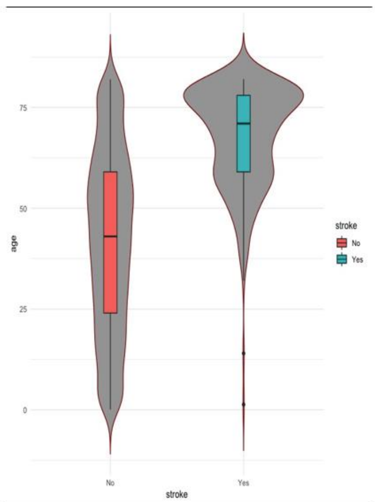
Average age of patients who have suffered strokes is around 70 or above which is much higher than the patients who are not stroke victims. This indicates that the older the patient is, the higher the chances of to be diagnosed with stroke. The plot shows a few outliers too.

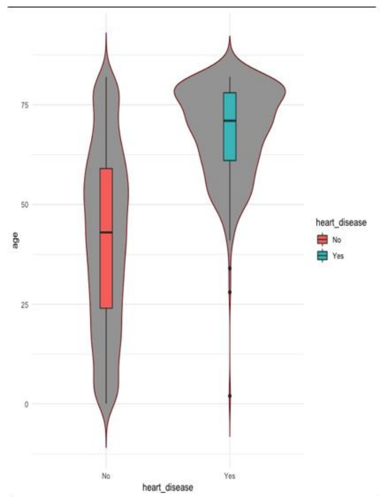
- **Age of Patients with and Without Heart Diseases**

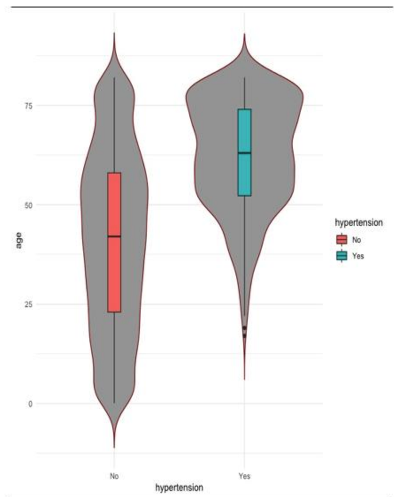
The older age patients are most likely to be diagnosed with heart diseases. Also, we can a few outliers too.

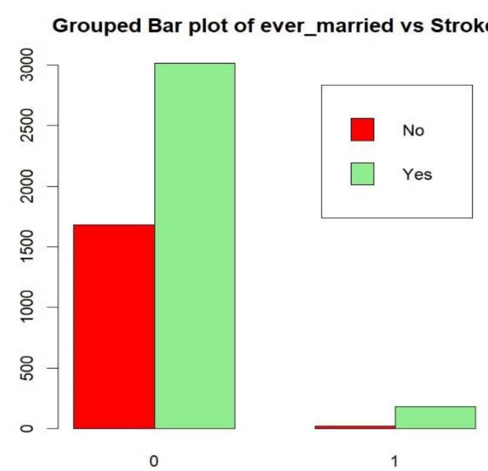
- **Age of Patients with and Without Hypertension**

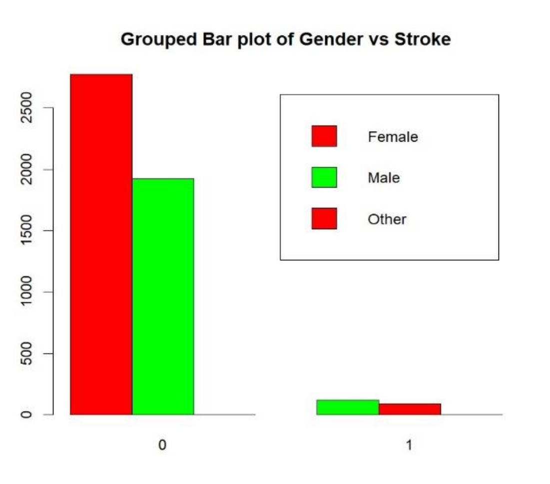
The plot shows a much higher mean age in patients who suffered hypertension than in those who have not, with a pair of low outliers among stroke victims. Same as heart disease and stroke, older patients most likely to be diagnosed with hypertension.



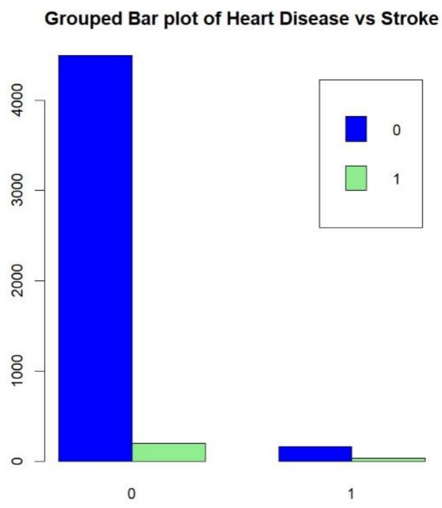


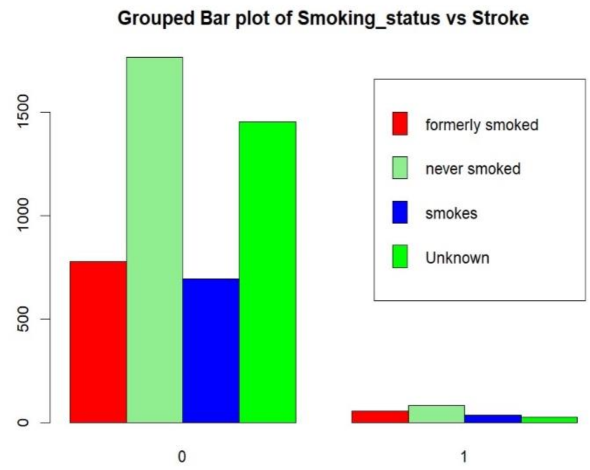






From the above bar chart of Stroke vs Ever married or not and Gender vs Stroke, it can be observed that most of the distribution is for people who have not had a stroke, People who are married have got a stroke more than the people who are not married.

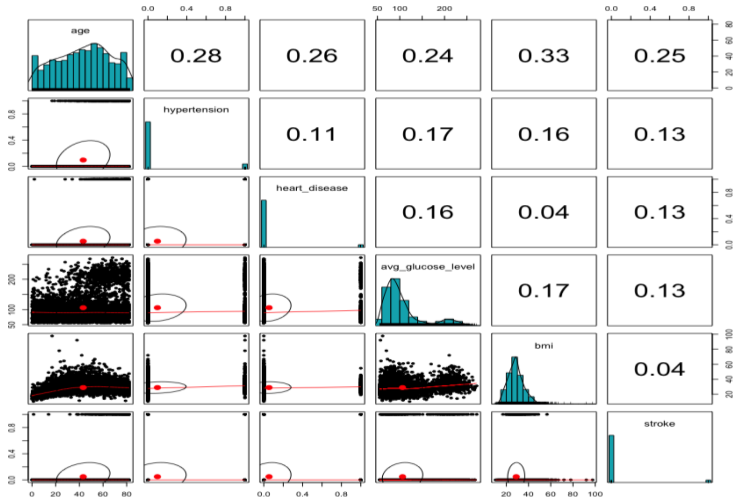




From the above bar chart of Stroke vs heart disease and Smoking status vs Stroke, it can be observed that most of the distribution is for people who have not had a stroke, People who have not had a heart disease and have had a stroke is higher than the people who have got a heart disease. Also, People who do not smoke have never smoked have had strokes more than the other categories. Overall people who have not smoked has higher chances of not getting a stroke.

2.4 Correlation between numerical variables

From below scatter plot matrix, we can see that all numeric variables are positively correlated to the dependent variable [age] (not strong correlation though). BMI has the highest correlation with age. Based on the data, stroke, hypertension, heart disease and average glucose level are the main driving factors to predict the age of a patient.



2.5 Feature Selection

2.5.1 Forward Selection Method

It is an iterative method in which we start with having no feature in the model and in each iteration, we keep adding the feature which is the best for our model. It gives us the AIC value in each step. Below displaying the selected features for our model.

```
Step:  AIC=26335.64
age ~ work_type + ever_married + heart_disease + stroke + hypertension +
    avg_glucose_level + smoking.status + bmi

                 Df Sum of Sq    RSS   AIC
<none>                        879943 26336
+ Residence_type  1   168.900 879774 26337
+ gender          1     0.004 879943 26338

Call:
lm(formula = age ~ work_type + ever_married + heart_disease +
    stroke + hypertension + avg_glucose_level + smoking.status +
    bmi, data = data1)
```

2.5.2 Backward Elimination

Opposite to the forward selection, we will start with all the features and removes the least significant

```
Step:  AIC=26335.64
age ~ hypertension + heart_disease + ever_married + work_type +
    avg_glucose_level + bmi + stroke + smoking.status

                    Df Sum of Sq     RSS   AIC
<none>                            879943 26336
- bmi                1       620  880562 26337
- smoking.status     2     10112  890055 26390
- avg_glucose_level  1     13368  893311 26411
- hypertension       1     26996  906939 26488
- stroke             1     35837  915779 26538
- heart_disease      1     42285  922227 26574
- ever_married       1    261723 1141666 27664
- work_type          4    289812 1169754 27782
```

feature at each iteration which improves the performance of our model. As we can see, it gave us the same features.

2.5.3 Stepwise Selection (Bi-directional)

In this method, we add predictors to the model sequentially just like we did in forward selection and after adding each predictor we also remove the predictors that no longer provided an improvement in model fit. However, all the feature selection methods gave us the same result.

```
Step:  AIC=26335.64
age ~ hypertension + heart_disease + ever_married + work_type + 
    avg_glucose_level + bmi + stroke + smoking.status

                    Df Sum of Sq     RSS   AIC
<none>                            879943 26336
+ Residence_type     1       169  879774 26337
- bmi                1       620  880562 26337
+ gender             1         0  879943 26338
- smoking.status     2     10112  890055 26390
- avg_glucose_level  1     13368  893311 26411
- hypertension       1     26996  906939 26488
- stroke             1     35837  915779 26538
- heart_disease      1     42285  922227 26574
- ever_married       1    261723 1141666 27664
- work_type          4    289812 1169754 27782

Call:
lm(formula = age ~ hypertension + heart_disease + ever_married + 
    work_type + avg_glucose_level + bmi + stroke + smoking.status, 
    data = data1)
```

2.6 Splitting the data into test and train sets (70-30)

Now, let’s split features into training and testing sets (70-30) for training and testing our model. We will train the 70% data and test on 30% of the data.

```
> # Create Train and Test set - maintain % of event rate (70/30 split)
> library(caret)
> trainIndex = sort(sample(x = nrow(data1), size = nrow(data1) * 0.7))
> sample_train = data1[trainIndex,]
> sample_test = data1[-trainIndex,]
> dim(sample_train)
[1] 3577   11
> dim(sample_test)
[1] 1533   11
> 
```

2.7 Model Building

<u>MULTIPLE REGRESSION MODEL</u>

Model 1

It can be seen from below displaying figure that p-value of the F-statistic is < 2.2e-16, which is highly significant. This means that, at least, one of the predictor variables is significantly related to our outcome variable. However, by looking at the p-values, we can see that BMI is not statistically significant with our outcome variable.

- Age = b0 + b1*hypertension + b2*heart_disease + b3*ever_married + b4*work_type + b5*avg_glucose_level+ b6*bmi + b7*smoking.status + b8*stroke

```
> model1 = lm(age ~ hypertension + heart_disease + ever_married +
+                work_type + avg_glucose_level + bmi + smoking.status + stroke, data = sample_train)
> summary(model1)

Call:
lm(formula = age ~ hypertension + heart_disease + ever_married +
    work_type + avg_glucose_level + bmi + smoking.status + stroke,
    data = sample_train)

Residuals:
    Min      1Q  Median      3Q     Max
-35.507  -8.542  -0.991   7.670  53.228

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 6.459489   1.052075   6.140 9.17e-10 ***
hypertension                8.285735   0.775994  10.678  < 2e-16 ***
heart_disease              13.508714   1.018619  13.262  < 2e-16 ***
ever_marriedYes            19.119938   0.574480  33.282  < 2e-16 ***
work_typeGovt_job          26.420305   1.012682  26.089  < 2e-16 ***
work_typeNever_worked       9.957347   3.569889   2.789  0.00531 **
work_typePrivate           23.055662   0.825079  27.944  < 2e-16 ***
work_typeSelf-employed     33.477562   0.988320  33.873  < 2e-16 ***
avg_glucose_level           0.034376   0.005021   6.847 8.87e-12 ***
bmi                        -0.044787   0.032870  -1.363  0.17310
smoking.statusnever smoked -2.655193   0.536058  -4.953 7.64e-07 ***
smoking.statussmokes       -3.425858   0.645293  -5.309 1.17e-07 ***
stroke                     11.608301   1.037827  11.185  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.14 on 3564 degrees of freedom
Multiple R-squared:  0.6678,    Adjusted R-squared:  0.6667
F-statistic:   597 on 12 and 3564 DF,  p-value: < 2.2e-16
```

Model 2

Let's remove the least significant predictor variable and fit a new model. After removing BMI, the **R2 = 0.66** which is not different from our Model 1, meaning that both the models have 66% of the variance in the measure of age can be predicted by our predictor variables.

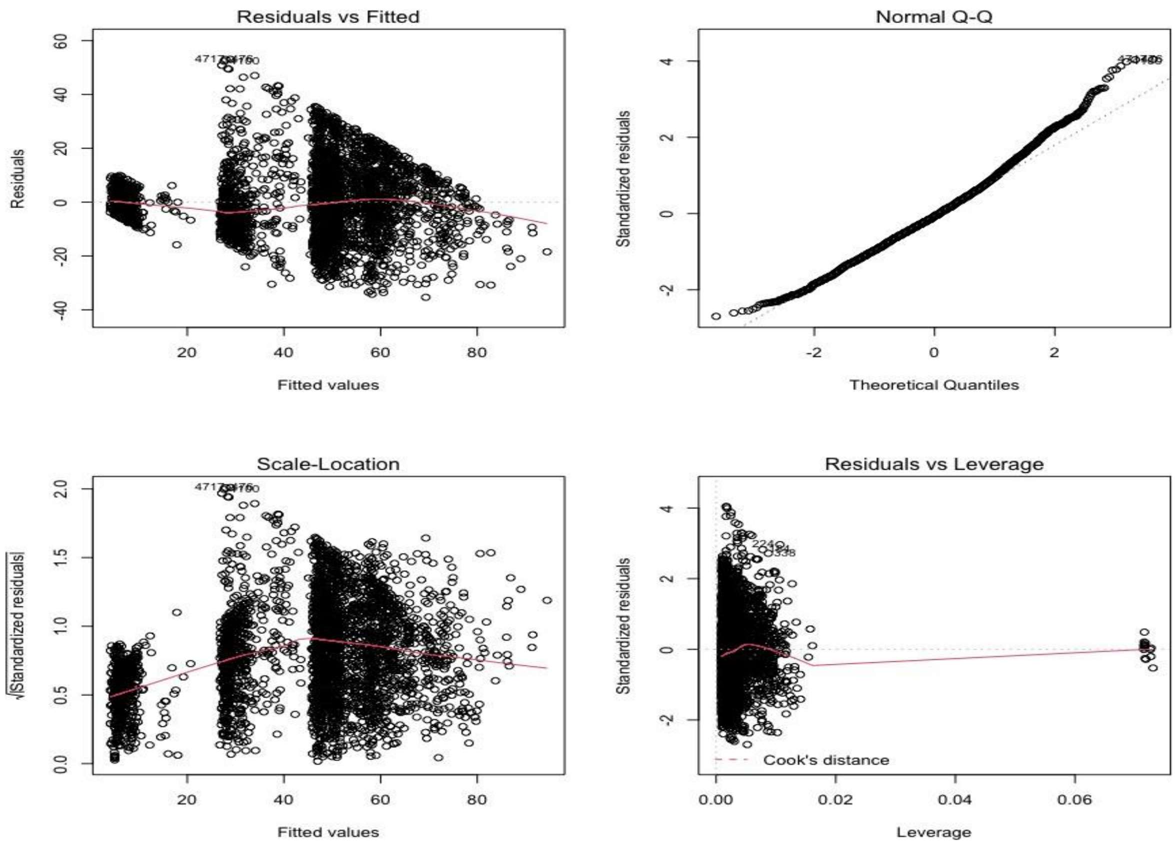
Age = b0 + b1*hypertension + b2*heart_disease + b3*ever_married + b4*work_type + b5*avg_glucose_level
+ b6 *smoking.status + b7*stroke

Residual Standard Error (RSE): The RSE estimate gives us a measure of error of prediction. The lower the RSE, the more accurate the model is. The error rate can be estimated by dividing the RSE by the mean outcome variable. For our model, the RSE is 13.14 corresponding to 30% error rate.

```
> sigma(model2)/mean(sample_train$age)
[1] 0.3041276
>
```

2.8 Diagnostic Plots

- **Residual vs Fitted:** This plot is used to determine linearity. As the red line across the center of the plot is roughly horizontal then we can assume that the residuals follow a linear pattern.
- **Normal QQ:** This plot is used to determine the Normality of the residuals. As we can see in the plot, the points fall along the straight diagonal line, hence we can assume the residuals are normally distributed.
- **Scale-Location:** This plot is used to check the assumption of equal variance (homoscedasticity). As the red line is roughly horizontal across the plot, hence we can say that the assumption of equal variance is likely met.
- **Residuals vs Leverage:** This plot is used to identify unusual (influential) observations. If any points in this plot fall outside of Cook's distance (the dashed lines) then it is an influential observation. The horizontal line is not deviating much and none of the points were influencing the model.



2.9 Multiple Regression Model: Interpretation

- The Residuals are the difference between the actual values and the predicted values of Age.
- Interpretations of the coefficients of the model:
- The intercept value 5.616746 gives us the estimated Y value (Age) when all the independent variables values are zero.
- The slope of hypertension is 8.169943, the slope of heart disease is 13.5399 and so on. Here the slope of each independent variable effects the age of the patients and simultaneously adjusts and controls the rest of the independent variables in the model.
- The t-statistic helps to find the p-value. The P-value of the Predictor variables in the summary indicates how significant is the variable to the model, any p-value below 0.05 indicates that the variable is significant for the model. Since here all the variables have p-values below 0.05, they are significant variable for the model except BMI which we have removed in Model 2.
- The Residual Standard Error tells us if the model is fitting data well or not. The lower it is the better the model. Here, the Residual Standard Error 13.14 which means the model predicts the age with an average error rate corresponding to 30%.
- Multiple R-Squared shows how well the data points are fitting along the curve or the line. Here, in the above model the R-Squared is 66% which means that approximately 66% of variability in the age can be explained by this regression model.
- F-statistic and p-value of the model: To test the Validity and significance of the regression model, hypothesis test is conducted on the global model. Here, it can be observed that the F-statistic is large, and the p-value is less than 0.05, therefore the model is significant. There is strong evidence that relationship does exist between the age and at least one of the independent variables.

<u>LOGISTIC REGRESSION MODEL</u>

2.9.1 Splitting the data into Testing and Training sets

```
> ##########################################################################################
> # Sampling | Splitting data into 70% for training 30% for testing
> TrainingSampleIndex=sample(1:nrow(DataForML), size=0.7 * nrow(DataForML) )
> DataForMLTrain=DataForML[TrainingSampleIndex, ]
> DataForMLTest=DataForML[-TrainingSampleIndex, ]
> DataForMLTrain = na.omit(DataForMLTrain)
> DataForMLTest = na.omit(DataForMLTest)
> dim(DataForMLTrain)
[1] 3577   12
> dim(DataForMLTest)
[1] 1533   12
```

In the above figure, the heart disease dataset has been split into training and testing data in the ratio of 70:30 for Training: Testing data.

2.9.2 Full Model: Target variable with all the predictor variables

```
Coefficients:
                            Estimate Std. Error z value Pr(>|z|)    
(Intercept)               -7.194e+00  1.104e+00  -6.518 7.12e-11 ***
id                        -4.378e-06  4.313e-06  -1.015 0.310085    
genderMale                -2.913e-02  1.914e-01  -0.152 0.879064    
genderOther               -1.021e+01  1.455e+03  -0.007 0.994405    
age                        7.276e-02  7.864e-03   9.251  < 2e-16 ***
hypertension1              4.451e-01  2.182e-01   2.040 0.041352 *  
heart_disease1             4.719e-01  2.675e-01   1.764 0.077675 .  
ever_marriedYes           -1.155e-01  3.102e-01  -0.372 0.709523    
work_typeGovt_job         -1.214e+00  1.167e+00  -1.040 0.298199    
work_typeNever_worked     -1.018e+01  4.570e+02  -0.022 0.982229    
work_typePrivate          -8.935e-01  1.145e+00  -0.780 0.435275    
work_typeSelf-employed    -1.154e+00  1.170e+00  -0.986 0.324200    
Residence_typeUrban       -4.017e-02  1.823e-01  -0.220 0.825549    
avg_glucose_level          5.885e-03  1.633e-03   3.605 0.000312 ***
bmi                        1.089e-02  1.425e-02   0.764 0.444589    
smoking_statusnever smoked -5.658e-03  2.321e-01  -0.024 0.980552    
smoking_statussmokes       3.396e-01  2.824e-01   1.202 0.229181    
smoking_statusUnknown     -1.667e-01  3.032e-01  -0.550 0.582409    
Prediction1                1.583e-03  4.195e-01   0.004 0.996989    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1182.94  on 3435  degrees of freedom
Residual deviance:  927.45  on 3417  degrees of freedom
AIC: 965.45

Number of Fisher Scoring iterations: 14
```

2.9.3 Interpretations of the full model:

- From the above table in figure 2, it can be observed that **age and avg_glucose_level is most significantly** associated with the **target variable, hypertension1** is also associated with the target variable but is slightly less significant than age.
- The **co-efficient estimate** of age is **positive.** This means with an **increase in Number of Enrollment of new students** will be associated **with an increased probability of getting a Heart stroke.**
- The **goodness of** fit is measured by the **difference between the Null deviance and the Residual deviance**. The greater the difference between them, the better is the model. Null deviance is the value when all the predictor variables are 0 and there is only intercept term. Residual deviance is the value considering all the predictor variables into account. In the above model the difference is moderate in between the Null and the Residual deviance, hence it is a good model.
- The **Akaike Information Criterion (AIC) is a method to evaluate how well does the model fit the data it has been generated from.** The lower the AIC compared to other models the better it is because it means with lower number of predictor variables It will have similar accuracy. Here, the AIC value is 965.45.

2.9.4 Model output using stepwise feature selection criteria.

```
> step.model <- LR_Model %>% stepAIC(trace = FALSE)
> coef(step.model)
      (Intercept)               age     hypertension1    heart_disease1 avg_glucose_level
     -7.794485423       0.067026948       0.462368760       0.505103726       0.006089103
> summary(step.model)

Call:
glm(formula = TargetVariable ~ age + hypertension + heart_disease +
    avg_glucose_level, family = "binomial", data = DataForMLTrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1451  -0.2868  -0.1585  -0.0779   3.6085

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -7.794485   0.468297 -16.644  < 2e-16 ***
age                0.067027   0.006796   9.863  < 2e-16 ***
hypertension1      0.462369   0.211521   2.186   0.0288 *
heart_disease1     0.505104   0.243707   2.073   0.0382 *
avg_glucose_level  0.006089   0.001495   4.072 4.66e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1182.94  on 3435  degrees of freedom
Residual deviance:  934.33  on 3431  degrees of freedom
AIC: 944.33

Number of Fisher Scoring iterations: 7
```

2.9.5 Interpretations of the stepwise regression model:

- From the above table in figure 2, it can be observed that **age is most significantly** associated with the **target variable, hypertension1** and **avg_glucose_level** is also associated with the target variable but is slightly less significant than age. Apart from that **heart_disease1** is also considered important as per stepwise feature selection criteria.
- The **co-efficient estimate** of all the predictor variables is **positive.** This means with an **increase in predictor variables** will be associated **with an increased probability of getting a Heart stroke.**
- The **goodness of** fit is measured by the **difference between the Null deviance and the Residual deviance**. Compared to the full model the difference is moderate in between the Null and the Residual deviance, hence it is a better model.
- The **Akaike Information Criterion (AIC) is a method to evaluate how well does the model fit the data it has been generated from.** The lower the AIC compared to other models the better it is because it means with lower number of predictor variables It will have similar accuracy. Here, in the model as per stepwise feature selection criteria the **AIC is 944.33** compared to the previous model which had an **AIC of 965.45**.

2.9.6 Confusion matrix of the testing dataset:

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1008   13
         1  398   54

               Accuracy : 0.721
                 95% CI : (0.6973, 0.7438)
    No Information Rate : 0.9545
    P-Value [Acc > NIR] : 1

                  Kappa : 0.14

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

              Precision : 0.9873
                 Recall : 0.7169
                     F1 : 0.8307
             Prevalence : 0.9545
         Detection Rate : 0.6843
   Detection Prevalence : 0.6931
      Balanced Accuracy : 0.7614

       'Positive' Class : 0
```

2.9.7 Interpretations:

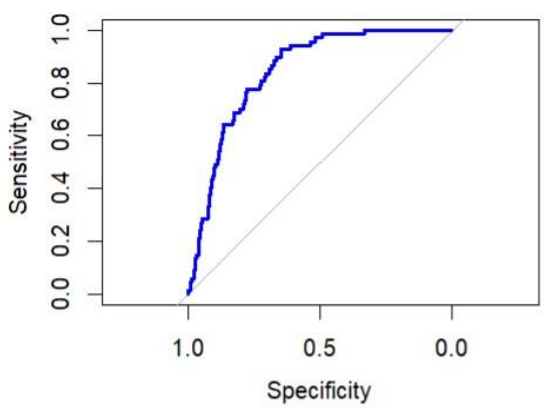
- The **confusion matrix** provides us with the details of the model accuracy and the errors in the prediction.
- **True Negative:** When the model predicted 0 and it is originally 0, here it is **1008.**
- **True Positive:** When the model predicted 1 and it is originally 1, here it is **54.**
- **False Negative:** When the model predicted 0 but it is originally 1, here it is **13**.
- **False Positive:** When the model predicted 1 but it is originally 0, here it is **398.**
- **Accuracy**: Accuracy is **72.1%** means model predicts **72.1%** correctly whether a person have had a heart stroke or not.
- **Precision:** Precision checks how correctly does the model predicts true positives. The precision value of **98.73%** means that the **model predicts correctly 98.73 of times if the person has had a heart stroke or not.**
- **Recall/True Positive Rate/Sensitivity:** Recall checks how often does the model predict yes when it is originally yes. The Recall value of **71.69%** shows **that the model predicts the Heart stroke correctly 71.69% of times when originally a heart stroke has occurred.**
- **F1-Score:** The F1-score of the model for the testing dataset is **83.07%.**

```
> Accuracy1[['table']]
          Reference
Prediction    0    1
         0 1008   13
         1  398   54
> Accuracy1[['byClass']]
         Sensitivity          Specificity       Pos Pred Value       Neg Pred Value
           0.7169275            0.8059701            0.9872674            0.1194690
           Precision               Recall                   F1           Prevalence
           0.9872674            0.7169275            0.8306551            0.9545146
      Detection Rate Detection Prevalence    Balanced Accuracy
           0.6843177            0.6931432            0.7614488
> |
```

- **In above figure,** it can be observed that the **Specificity is 80.05%.** It is also referred to as True Negative Rate, shows the proportion of negative class correctly.

2.9.8 The ROC (Receiver Operating Characteristic Curve):

The ROC curve shows the **performance of the classification model at all the classification thresholds**. It is produced by **plotting the True Positive Rate (TPR) against the False Positive Rate (FPR).**



The ROC curve for this model fits moderately. Improvements can be made if more data were to be collected. They also might be made if additional factors were collected, specifically ones which are risk factors by those with domain knowledge.

2.9.9 AUC (Area under the (ROC) curve)

The Area under the curve tells us how better the model will perform. The **higher the AUC curve the better** will the classifier of the model perform on a given task.

```
> ## Area under the curve
> auc(auc_gbm)
Area under the curve: 0.8437
```

The AUC value of 0.8437, which is near to 1 indicates that the **model is good at predicting whether the Person have had a Heart Stroke or not.**

CONCLUSION

The Heart stroke dataset is chosen from Kaggle to build models with these 2 Goals:

Goal2: To predict the Age in which people get health issues

To predict Age, we have cleaned the data followed by doing Exploratory data analysis. Visualized the relationship between our target (**Age**) and other variables including numerical & categorical. Next, by plotting a scatterplot matrix, checked the numerical variables which are correlated with the target variable. Also, performed the **feature selection methods** to choose the variable which can improve our model. Built the **Multiple Linear Regression** model with all the selected predictors with **R Squared 0.66** and **RSE is 13.14**.

Goal1: To predict whether a person has had a heart Stroke or not

For the above Goal2, the dataset has been cleaned and from the dataset the missing values of the variable BMI have been removed, after conducting Exploratory Data Analysis, the variables have been checked for correlation with the stroke variable. The **Logistic Regression model** has been built using all the predictor variables which gave an **AIC value of 965.45.** Further, **stepwise feature selection criteria have been used** to eliminate insignificant variables and a model has been constructed with all the significant predictor variables which has **a lower AIC of 944.33 and the accuracy of the model is 72.2%**. Finally, **ROC** (Receiver Operating Characteristics Curve) has been plotted and **the area under the curve is 84.37%**.

REFERENCES

- https://www.analyticsvidhya.com/blog/2021/05/how-to-create-a-stroke-prediction-model/
- https://www.stat.cmu.edu/capstoneresearch/spring2021/315files/team16.html
- https://www.kaggle.com/adrynh/exploratory-data-analysis-on-stroke-dataset
- https://www.analyticsvidhya.com/blog/2021/06/25-questions-to-test-your-skills-on-linear-regression-algorithm/
- https://degreesofbelief.roryquinn.com/common-evaluation-measures-for-classification-models#:~:text=Detection%20Rate%20%2Dshows%20the%20number,a%20proportion%20of%20all%20predictions.
- https://stackoverflow.com/questions/38829646/confusion-matrix-of-bsttree-predictions-error-the-data-must-contain-some-leve
- https://stats.stackexchange.com/questions/175767/why-the-probability-in-my-logistic-regression-is-far-from-0-5-in-r