Richa Rambhia
December 18, 2021

Module 6 R Practice
Report & Outputs

Introduction

The task in this assignment is to build a regression model for the categorical values of the dataset and perform a subset analysis to determine which the best model based on the predictor variables. The dataset chosen here was the Heart Failure Prediction dataset which has 919 rows of data and 12 data fields which is nothing but the attributes of the dataset. The attributes of the dataset are age, sex, chest pain type, resting bp, cholesterol, fasting bp, heart disease, etc. The description of the attributes of the heart failure prediction dataset is as follows.

Attribute Information.

Age: Age of the patient [years]

Sex: Sex of the patient [M, F]

ChestPainType: Chest pain type

RestingBP: Resting Blood Pressure [mm Hg]

Cholesterol: Serum Cholesterol [mm/dl]

FastingBS: Fasting Blood Sugar

RestingECG: Resting ElectroCardiogram results

MaxHR: Maximum heart rate achieved

ExerciseAngina: Exercise-induced angina [Yes, No]

Oldpeak: Oldpeak = ST

ST_Slope: The slope of the peak exercise ST segment

HeartDisease: Output Class [1: heart disease, 0: Normal]

The heart failure prediction dataset along with all the attributes mentioned will thus help in predicting the possible heart disease based on the various parameters of the dataset. The phases that will be performed would be the descriptive analysis, data visualizations, linear regression model, MV regression and the subset analysis for the dataset to predict the heart disease of the person.

Source to the dataset: https://www.kaggle.com/fedesoriano/heart-failure-prediction

Data Analysis & Visualizations

Before beginning with the various phases of the data analysis processes and regression model, packages needed like broom, corrplot, gtsummary, caret and leaps were installed, and the libraries were imported. The dataset of the heart failure prediction was then read into a dataframe which would be used in further analysis. Now, to have some idea about what the dataset represents, describing the data was necessary and so we displayed the column names of the dataset which gave an overall understanding about the data fields, the starting and ending records of the dataset were displayed, the summary and structure of the dataset was displayed which gave an overview about the statistical values of the dataset and finally the class type of the attributes were displayed to know about the type of class of each attribute.

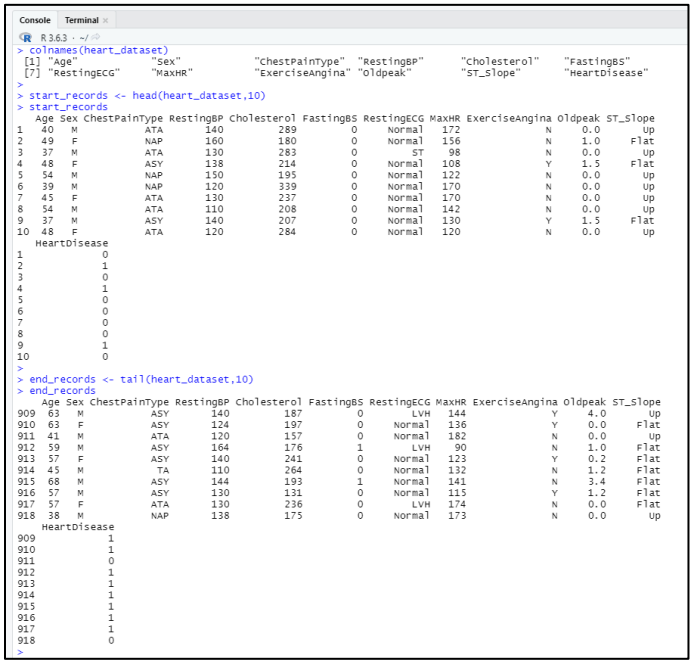
To analyze the dataset in order to be able to perform the regression analysis, descriptive analysis and data visualizations were performed to get an understanding about the values of the dataset. In the descriptive analysis, all the statistical values of the attributes were computed whereas in data visualization boxplot, bar plot and scatter plot with a regression line were created which gave an overall idea about the dataset attributes.

The output of the dataset description, descriptive analysis and data visualizations is as below.

Output:

1. <u>Dataset Description:</u>

a. Column names, start records, end records



```
Console  Terminal
R 3.6.3 · ~/
> colnames(heart_dataset)
 [1] "Age"            "Sex"            "ChestPainType"  "RestingBP"      "Cholesterol"    "FastingBS"
 [7] "RestingECG"     "MaxHR"          "ExerciseAngina" "Oldpeak"        "ST_Slope"       "HeartDisease"
>
> start_records <- head(heart_dataset,10)
> start_records
   Age Sex ChestPainType RestingBP Cholesterol FastingBS RestingECG MaxHR ExerciseAngina Oldpeak ST_Slope
1   40   M           ATA       140         289         0     Normal   172              N     0.0       Up
2   49   F           NAP       160         180         0     Normal   156              N     1.0     Flat
3   37   M           ATA       130         283         0         ST    98              N     0.0       Up
4   48   F           ASY       138         214         0     Normal   108              Y     1.5     Flat
5   54   M           NAP       150         195         0     Normal   122              N     0.0       Up
6   39   M           NAP       120         339         0     Normal   170              N     0.0       Up
7   45   F           ATA       130         237         0     Normal   170              N     0.0       Up
8   54   M           ATA       110         208         0     Normal   142              N     0.0       Up
9   37   M           ASY       140         207         0     Normal   130              Y     1.5     Flat
10  48   F           ATA       120         284         0     Normal   120              N     0.0       Up
   HeartDisease
1             0
2             1
3             0
4             1
5             0
6             0
7             0
8             0
9             1
10            0
>
> end_records <- tail(heart_dataset,10)
> end_records
    Age Sex ChestPainType RestingBP Cholesterol FastingBS RestingECG MaxHR ExerciseAngina Oldpeak ST_Slope
909  63   M           ASY       140         187         0       LVH   144              Y     4.0       Up
910  63   F           ASY       124         197         0     Normal   136              Y     0.0     Flat
911  41   M           ATA       120         157         0     Normal   182              N     0.0       Up
912  59   M           ASY       164         176         1       LVH    90              N     1.0     Flat
913  57   F           ASY       140         241         0     Normal   123              Y     0.2     Flat
914  45   M            TA       110         264         0     Normal   132              N     1.2     Flat
915  68   M           ASY       144         193         1     Normal   141              N     3.4     Flat
916  57   M           ASY       130         131         0     Normal   115              Y     1.2     Flat
917  57   F           ATA       130         236         0       LVH   174              N     0.0     Flat
918  38   M           NAP       138         175         0     Normal   173              N     0.0       Up
    HeartDisease
909            1
910            1
911            0
912            1
913            1
914            1
915            1
916            1
917            1
918            0
>
```

b. Summary, Structure and Type of dataset

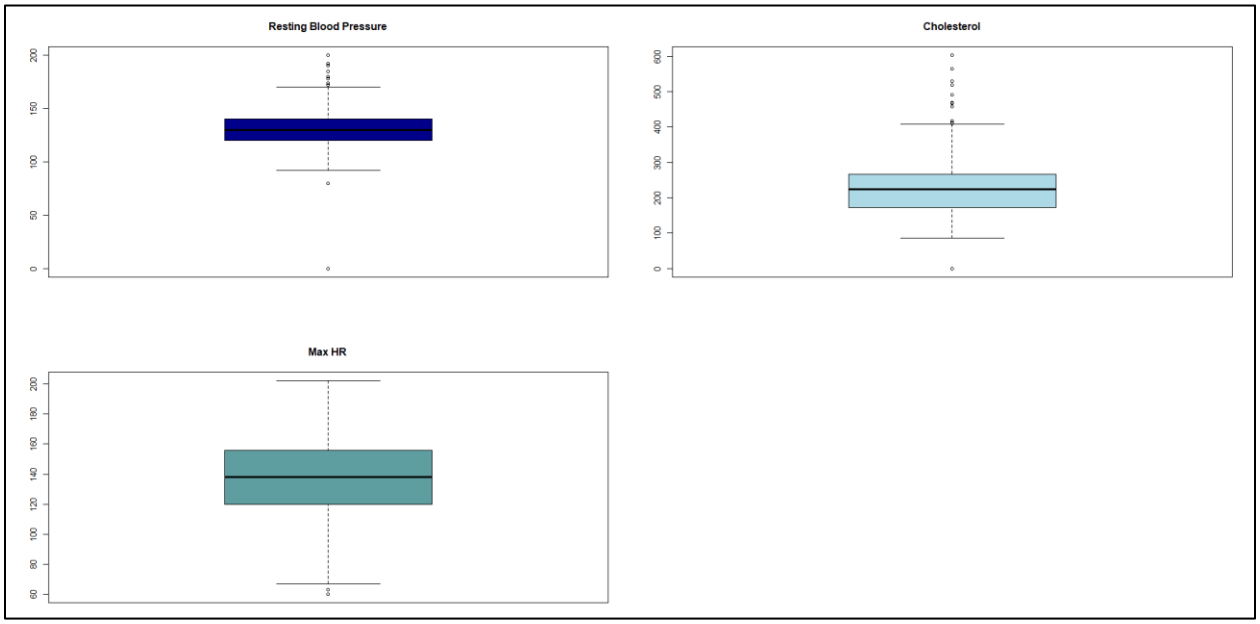
```
> dataset_summary <- summary(heart_dataset)
> dataset_summary
      Age         Sex     ChestPainType   RestingBP      Cholesterol      FastingBS       RestingECG
 Min.   :28.00   F:193   ASY:496       Min.   :  0.0   Min.   :  0.0   Min.   :0.0000   LVH   :188
 1st Qu.:47.00   M:725   ATA:173       1st Qu.:120.0   1st Qu.:173.2   1st Qu.:0.0000   Normal:552
 Median :54.00           NAP:203       Median :130.0   Median :223.0   Median :0.0000   ST    :178
 Mean   :53.51           TA : 46       Mean   :132.4   Mean   :198.8   Mean   :0.2331
 3rd Qu.:60.00                         3rd Qu.:140.0   3rd Qu.:267.0   3rd Qu.:0.0000
 Max.   :77.00                         Max.   :200.0   Max.   :603.0   Max.   :1.0000
     MaxHR       ExerciseAngina    Oldpeak        ST_Slope    HeartDisease
 Min.   : 60.0   N:547          Min.   :-2.6000   Down: 63   Min.   :0.0000
 1st Qu.:120.0   Y:371          1st Qu.: 0.0000   Flat:460   1st Qu.:0.0000
 Median :138.0                  Median : 0.6000   Up  :395   Median :1.0000
 Mean   :136.8                  Mean   : 0.8874              Mean   :0.5534
 3rd Qu.:156.0                  3rd Qu.: 1.5000              3rd Qu.:1.0000
 Max.   :202.0                  Max.   : 6.2000              Max.   :1.0000
>
> sapply(heart_dataset,class)
           Age            Sex  ChestPainType      RestingBP    Cholesterol      FastingBS     RestingECG
     "integer"       "factor"       "factor"      "integer"      "integer"      "integer"       "factor"
         MaxHR ExerciseAngina        Oldpeak       ST_Slope   HeartDisease
     "integer"       "factor"      "numeric"       "factor"      "integer"
>
> str(heart_dataset)
'data.frame':   918 obs. of  12 variables:
 $ Age           : int  40 49 37 48 54 39 45 54 37 48 ...
 $ Sex           : Factor w/ 2 levels "F","M": 2 1 2 1 2 2 1 2 2 1 ...
 $ ChestPainType : Factor w/ 4 levels "ASY","ATA","NAP",..: 2 3 2 1 3 3 2 2 1 2 ...
 $ RestingBP     : int  140 160 130 138 150 120 130 110 140 120 ...
 $ Cholesterol   : int  289 180 283 214 195 339 237 208 207 284 ...
 $ FastingBS     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ RestingECG    : Factor w/ 3 levels "LVH","Normal",..: 2 2 3 2 2 2 2 2 2 2 ...
 $ MaxHR         : int  172 156 98 108 122 170 170 142 130 120 ...
 $ ExerciseAngina: Factor w/ 2 levels "N","Y": 1 1 1 2 1 1 1 1 2 1 ...
 $ Oldpeak       : num  0 1 0 1.5 0 0 0 0 1.5 0 ...
 $ ST_Slope      : Factor w/ 3 levels "Down","Flat",..: 3 2 3 2 3 3 3 3 2 3 ...
 $ HeartDisease  : int  0 1 0 1 0 0 0 0 1 0 ...
>
```

2. Descriptive Analysis:

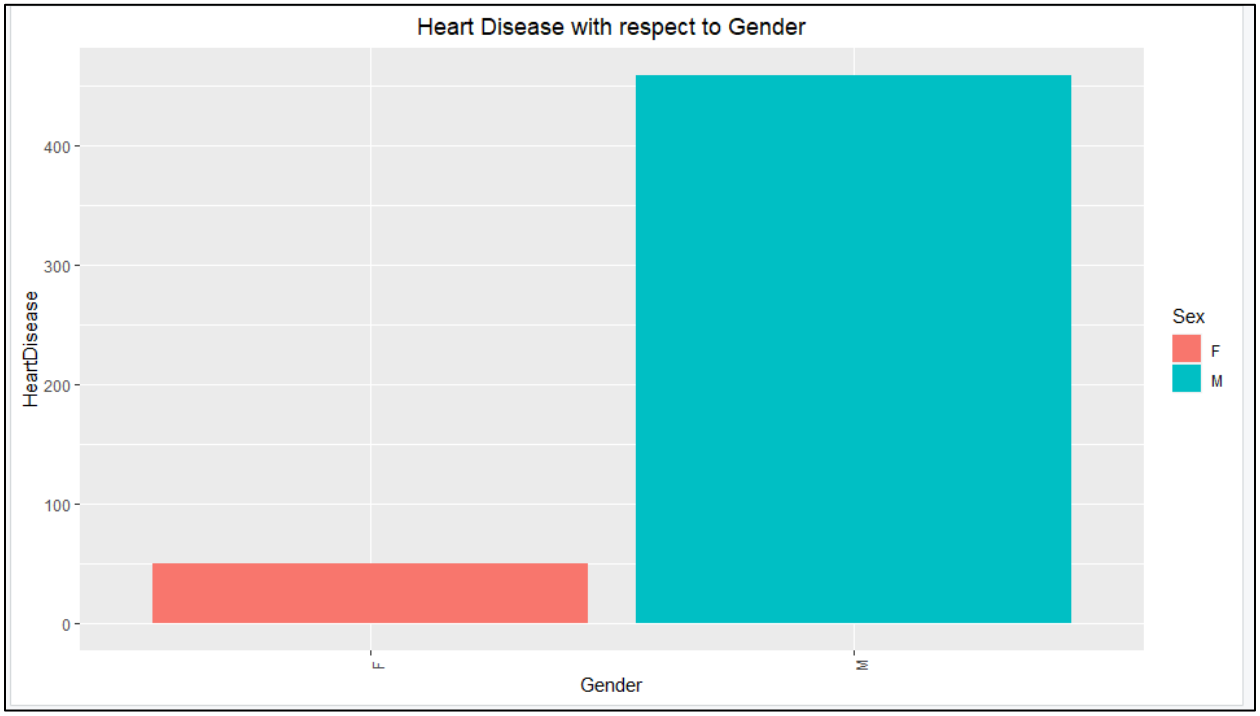
```
Console  Terminal
R 3.6.3 · ~/
> #descriptive analysis
> min(heart_dataset$RestingBP)
[1] 0
> max(heart_dataset$RestingBP)
[1] 200
> mean(heart_dataset$RestingBP)
[1] 132.3965
> median(heart_dataset$RestingBP)
[1] 130
> mode(heart_dataset$RestingBP)
[1] "numeric"
> range(heart_dataset$RestingBP)
[1]   0 200
> sd(heart_dataset$RestingBP)
[1] 18.51415
> summary(heart_dataset$RestingBP)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0   120.0   130.0   132.4   140.0   200.0
>
> min(heart_dataset$Cholesterol)
[1] 0
> max(heart_dataset$Cholesterol)
[1] 603
> mean(heart_dataset$Cholesterol)
[1] 198.7996
> median(heart_dataset$Cholesterol)
[1] 223
> mode(heart_dataset$Cholesterol)
[1] "numeric"
> range(heart_dataset$Cholesterol)
[1]   0 603
> sd(heart_dataset$Cholesterol)
[1] 109.3841
> summary(heart_dataset$Cholesterol)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0   173.2   223.0   198.8   267.0   603.0
>
> min(heart_dataset$MaxHR)
[1] 60
> max(heart_dataset$MaxHR)
[1] 202
> mean(heart_dataset$MaxHR)
[1] 136.8094
> median(heart_dataset$MaxHR)
[1] 138
> mode(heart_dataset$MaxHR)
[1] "numeric"
> range(heart_dataset$MaxHR)
[1]  60 202
> sd(heart_dataset$MaxHR)
[1] 25.46033
> summary(heart_dataset$MaxHR)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   60.0   120.0   138.0   136.8   156.0   202.0
>
```

3. Data Visualization:

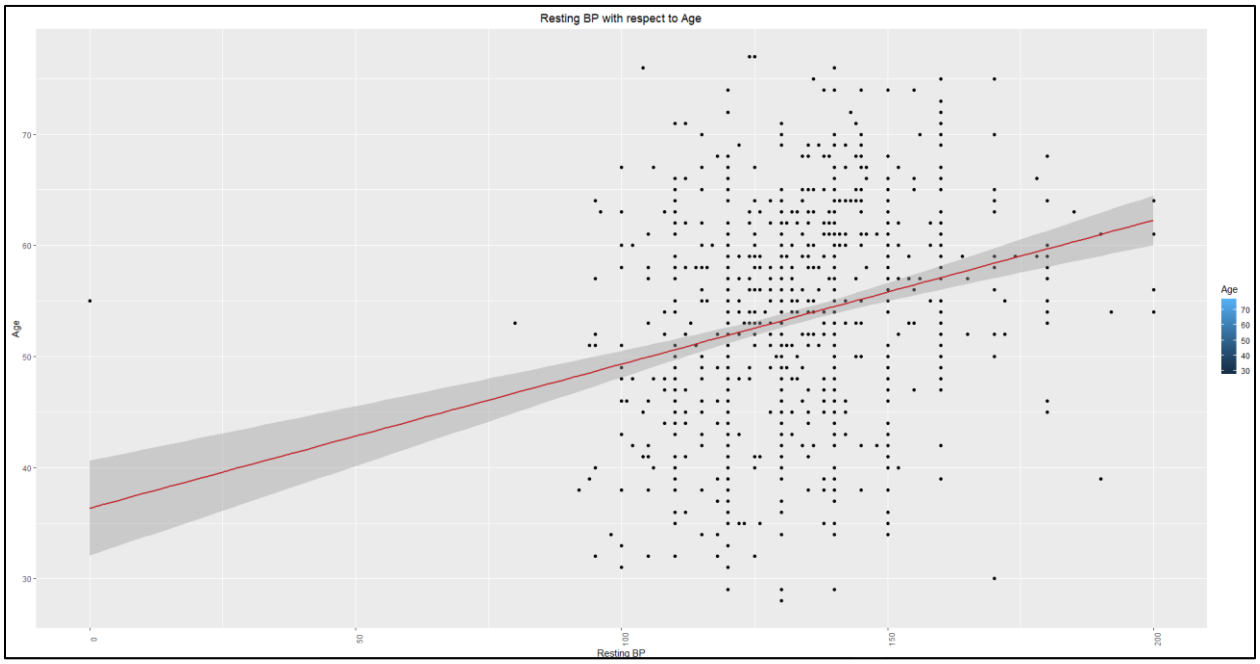
a. Graph 1 – Boxplot of the attributes



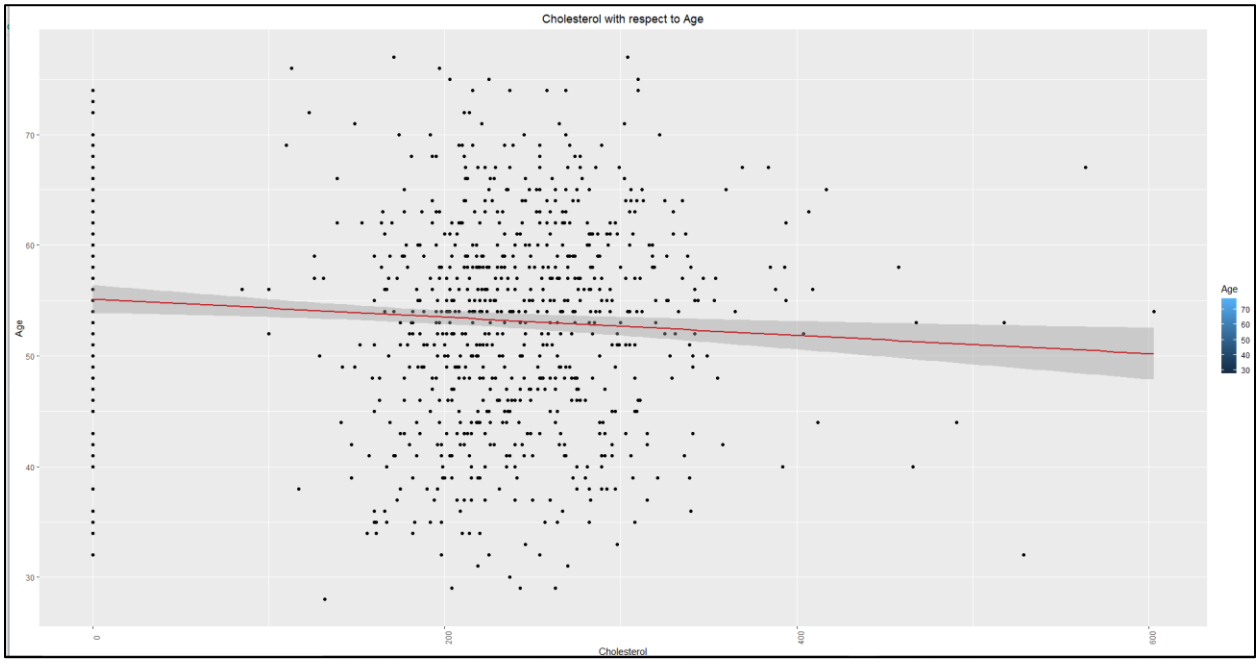
b. Graph 2 – Bar plot of heart disease with respect to gender



c. Graph 3 – Resting BP with respect to Age



d. Graph 4 – Cholesterol with respect to Age



Regression Model

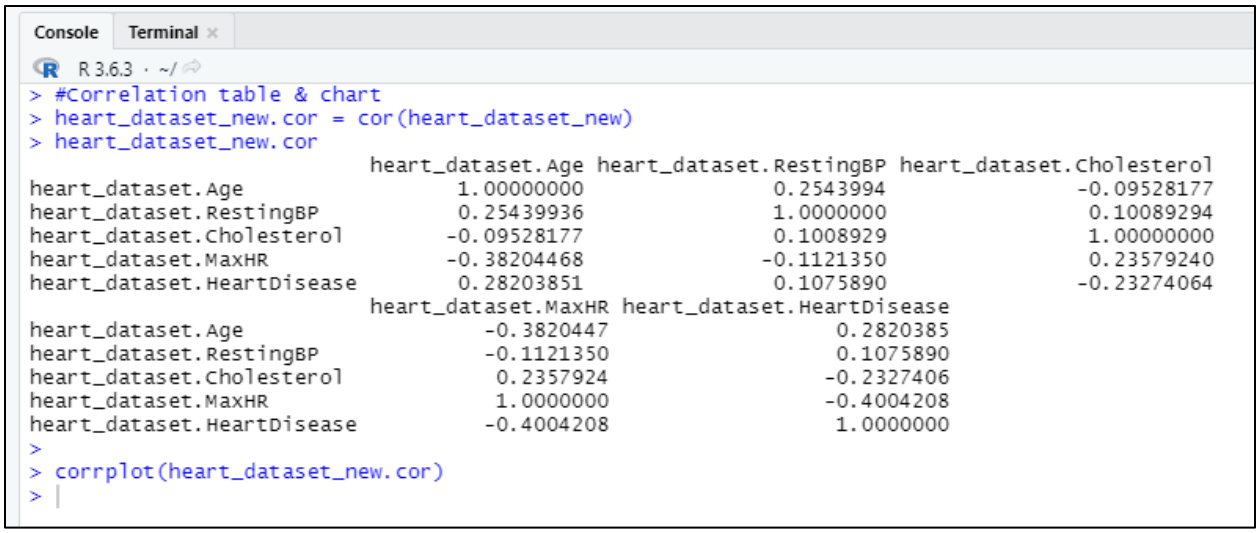
As we know, regression model is used to examine the relationship between the two attributes of the dataset and predict the outcome from the dataset. It basically estimates the relationship between a dependent variable and an independent variable of the dataset. Here in this case, we will use the regression model to predict the possible heart disease of the person with respect to the other predictor variables of the dataset. The dataset consists of numerical as well as categorical data values and thus we will be performing regression analysis on the categorical set of the data attributes as well.

Categorical variables are the variables which can take one of a limited and fixed number of possible values where the categorical data is divided into groups or categories of data based on the qualitative characteristics which does not have a number associated with it. Dummy variables on the other hand are numeric variables which represent the categorical data where their range is small and can take on only two quantitative values which are 0 or 1 representing the absence and presence of something respectively.

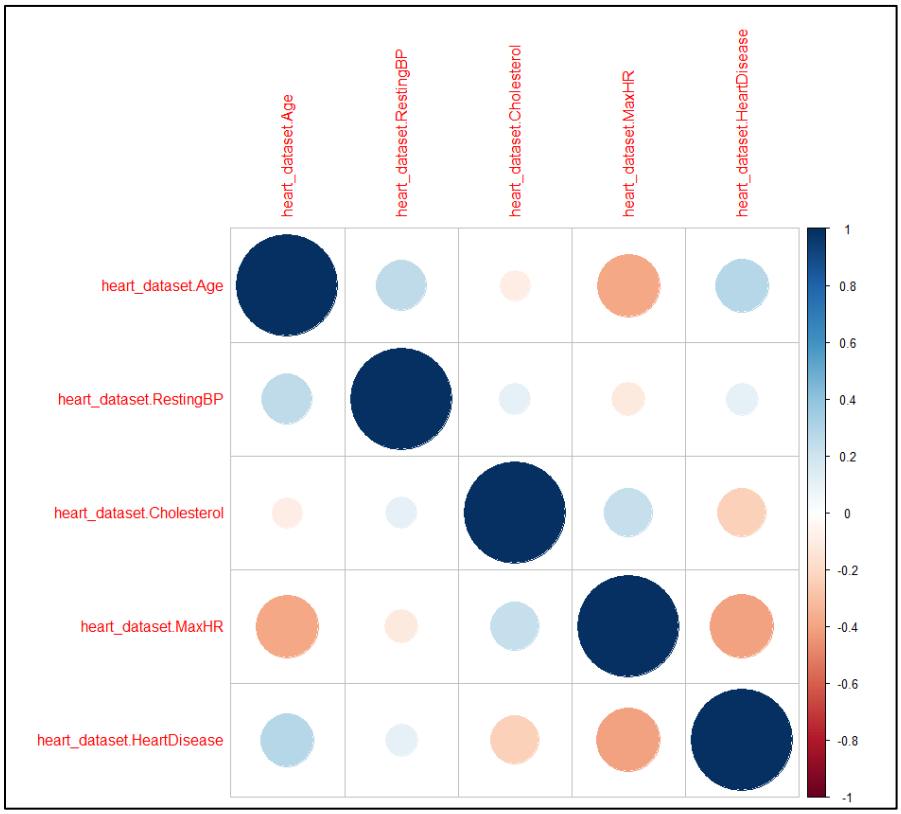
Before starting with the regression analysis and subset analysis, correlation table and chart were plotted to know about the relation between each attribute with each other. A subset of dataset was created in order to plot the correlation table and chart which had numeric attributes to determine the relation between each attribute. The output of the correlation table and the chart is as shown below.

Output:

1. Correlation Table:



1. Correlation Chart:



Now, in order to perform the regression analysis on categorical variables of the dataset, dummy variables were created on the categorical attributes of the heart failure prediction dataset. Here, the gender of the person was the categorical attribute for which the dummy variables were created representing 0 as male and 1 as female. Similarly, the dummy variable was created for the attribute of ST_Slope which was st_slope_down and st_slope_up. These variables created were then considered for the regression analysis where MV regression was performed in order to examine the relationship between the variables and predict the heart disease for the person. The regression plot for the MV regression model was also plotted in order to understand the regression analysis. The output of the regression model and the plot is as below.

Output:

1. Creating dummy variables:

```
Console  Terminal
R 3.6.3 · ~/
> #creating dummy variables
> dummy_male <- ifelse(heart_dataset$Sex == "M", 1,0)
> dummy_female <- ifelse(heart_dataset$Sex == "F", 1,0)
>
>
> st_slope_down <- ifelse(heart_dataset$ST_Slope == "Down", 1,0)
> st_slope_up <- ifelse(heart_dataset$ST_Slope == "Up", 1,0)
> st_slope_up
  [1] 1 0 1 0 1 1 1 1 0 1 1 0 1 0 1 0 0 1 0 0 1 0 1 0 1 1 0 1 1 1 0 1 0 0 1 1 0 1 1 0 1 0 1 1 0 0 1 1 0 0 0 0
 [53] 1 1 0 1 0 0 1 0 1 1 1 0 1 1 1 1 0 1 0 1 0 1 0 1 1 0 1 1 0 1 0 0 0 0 0 0 1 1 1 0 1 0 1 1 1 1 0 1 0 0
[105] 0 1 1 1 1 1 0 0 1 1 1 0 0 0 1 0 0 1 1 0 1 1 1 1 1 1 0 1 0 0 0 0 0 1 1 0 0 0 0 0 1 0 1 1 1 1 0 1 1 1 1 1 0
[157] 0 1 0 1 0 0 1 1 1 0 0 1 1 1 1 1 1 1 1 0 0 0 1 1 1 0 1 0 0 1 0 1 0 0 0 1 1 1 1 1 1 1 0 1 0 0 1 1 1 1 1 1 0
[209] 1 0 0 0 1 1 0 1 0 1 1 1 0 0 1 1 1 0 1 0 1 1 1 1 1 1 1 1 1 0 0 0 0 0 1 0 0 1 0 1 0 0 0 0 0 0 1 1 0 1 1 1 1 1
[261] 1 1 0 0 0 0 1 0 1 0 0 1 1 0 1 1 1 0 0 0 1 1 0 1 1 1 1 1 1 1 1 1 1 1 0 1 1 0 1 0 1 0 1 0 1 1 0 0 1 0 0 1 0 1 1
[313] 1 1 0 1 0 1 0 1 0 0 1 0 0 0 1 1 0 0 1 1 1 1 1 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0
[365] 0 0 0 1 0 0 1 0 1 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[417] 0 1 0 1 0 1 0 0 1 0 1 0 0 0 0 1 0 0 0 1 1 0 0 1 1 0 0 0 0 0 0 1 0 0 0 0 0 1 0 1 0 0 0 0 1 0 0 0 1 0 1 0 1
[469] 1 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 1 0 1 0 1 0 0 0 0 1
[521] 0 0 1 0 1 1 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 1 0 0 1 1 0 1 0 0 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0
[573] 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 1 1 0 0 1 0 0 1 0 0 0 0 0 0 0 1 1 0 0 0 0 0 1 0 1 0 0 0 1 0 1 1 0 0 0
[625] 0 0 1 1 0 1 0 0 0 0 1 0 0 0 1 1 1 0 1 0 1 0 1 1 0 0 1 1 1 1 1 1 1 1 1 1 0 1 1 1 0 0 0 1 1 1 0 1 0 1 0 0 1
[677] 0 1 1 0 0 1 1 0 1 0 1 1 0 1 0 0 1 0 1 0 0 1 0 0 1 1 0 1 0 1 1 0 0 1 0 0 1 1 1 0 0 1 0 0 0 0 0 1 0 0 1 1
[729] 1 1 0 1 0 0 0 0 0 0 0 1 0 1 0 0 1 1 0 1 1 0 1 0 1 0 1 1 0 1 0 1 1 0 1 0 1 1 1 0 1 0 1 1 0 1 0 0 0 0 1 0 1
[781] 1 1 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 1 0 1 1 0 1 0 0 1 0 1 1 1 0 0 0 1 1 1 1 0 1 0 0 1 0 0 0 0 0 0 0 1 1 1
[833] 1 0 1 0 0 1 1 1 0 0 0 1 0 0 0 1 1 0 0 1 0 1 1 0 1 0 0 1 1 1 1 0 0 0 1 1 1 1 1 0 0 1 1 1 0 0 1 0 1 1 0 0
[885] 0 1 1 0 0 1 0 0 0 1 0 0 1 0 1 1 0 0 0 1 0 0 1 0 1 0 1 0 0 0 0 0 0 0 1
>
```

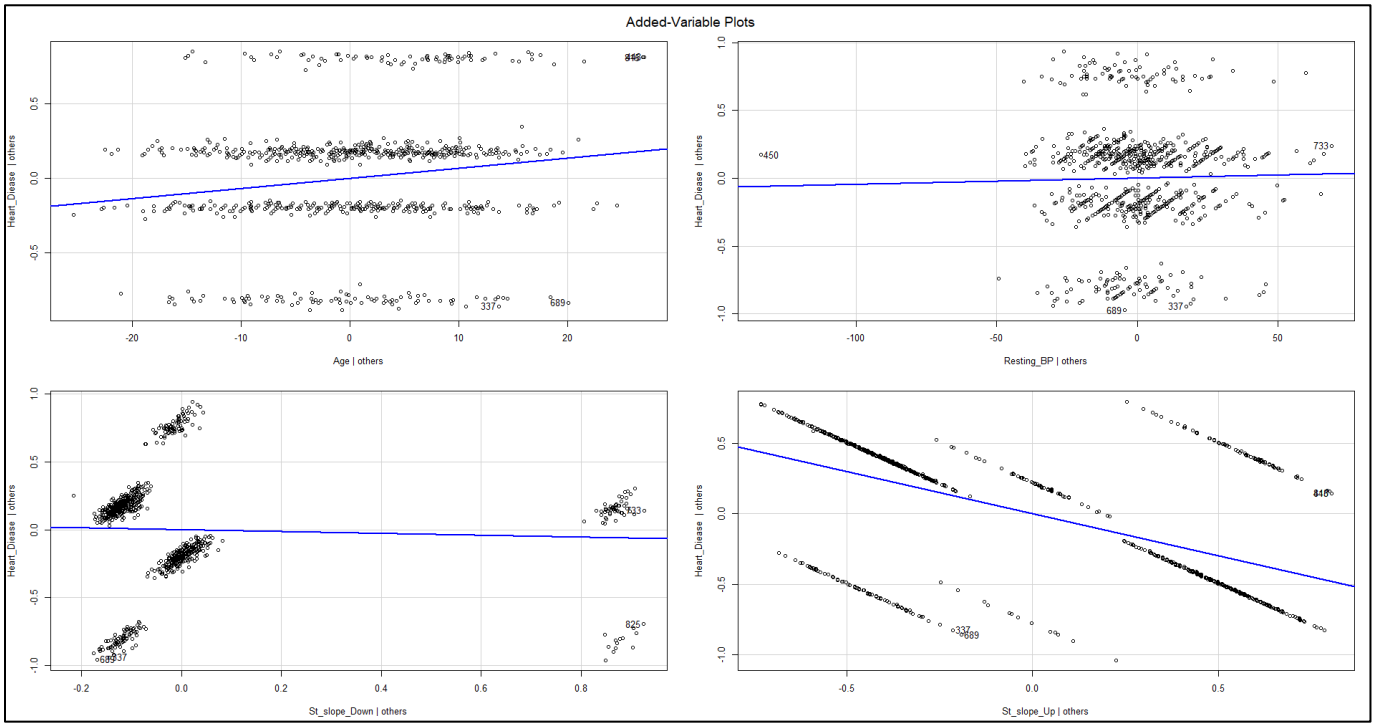
2. Creating data frame:

```
Console  Terminal
R 3.6.3 · ~/
> #create data frame
> dataframe_regression <- data.frame(Age = heart_dataset$Age,
+                                    Male = dummy_male,
+                                    Female = dummy_female,
+                                    Heart_Diease = heart_dataset$HeartDisease,
+                                    Resting_BP = heart_dataset$RestingBP,
+                                    Cholesterol = heart_dataset$Cholesterol,
+                                    St_slope_Down = st_slope_down,
+                                    St_slope_Up = st_slope_up)
> dataframe_regression
    Age Male Female Heart_Diease Resting_BP Cholesterol St_slope_Down St_slope_Up
1    40    1      0            0        140         289             0           1
2    49    0      1            1        160         180             0           0
3    37    1      0            0        130         283             0           1
4    48    0      1            1        138         214             0           0
5    54    1      0            0        150         195             0           1
6    39    1      0            0        120         339             0           1
7    45    0      1            0        130         237             0           1
8    54    1      0            0        110         208             0           1
9    37    1      0            1        140         207             0           0
10   48    0      1            0        120         284             0           1
11   37    0      1            0        130         211             0           1
12   58    1      0            1        136         164             0           0
13   39    1      0            0        120         204             0           1
14   49    1      0            1        140         234             0           0
15   42    0      1            0        115         211             0           1
16   54    0      1            0        120         273             0           0
17   38    1      0            1        110         196             0           0
18   43    0      1            0        120         201             0           1
19   60    1      0            1        100         248             0           0
20   36    1      0            1        120         267             0           0
21   43    0      1            0        100         223             0           1
22   44    1      0            0        120         184             0           0
23   49    0      1            0        124         201             0           1
24   44    1      0            1        150         288             0           0
25   40    1      0            0        130         215             0           1
26   36    1      0            0        130         209             0           1
27   53    1      0            0        124         260             0           0
28   52    1      0            0        120         284             0           1
29   53    0      1            0        113         468             0           1
30   51    1      0            0        125         188             0           1
31   53    1      0            1        145         518             0           0
32   56    1      0            0        130         167             0           1
33   54    1      0            1        125         224             0           0
34   41    1      0            1        130         172             0           0
35   43    0      1            0        150         186             0           1
36   32    1      0            0        125         254             0           1
37   65    1      0            1        140         306             0           0
38   41    0      1            0        110         250             0           1
39   48    0      1            0        120         177             0           1
40   48    0      1            0        150         227             0           0
41   54    0      1            0        150         230             0           1
42   54    0      1            1        130         294             0           0
43   35    1      0            0        150         264             0           1
44   52    1      0            0        140         259             0           1
45   43    1      0            1        120         175             0           0
46   59    1      0            0        130         318             0           0
```

3. MV Regression Model:

```
Console  Terminal
R 3.6.3 · ~/
> #MV regression model
> reg_model <- lm(Heart_Diease ~ Age + Resting_BP + St_slope_Down + St_slope_Up, data = dataframe_regression)
> reg_model

Call:
lm(formula = Heart_Diease ~ Age + Resting_BP + St_slope_Down +
    St_slope_Up, data = dataframe_regression)

Coefficients:
  (Intercept)            Age     Resting_BP  St_slope_Down    St_slope_Up
    0.3905459      0.0067956      0.0004635     -0.0701849     -0.5981079

> summary(reg_model)

Call:
lm(formula = Heart_Diease ~ Age + Resting_BP + St_slope_Down +
    St_slope_Up, data = dataframe_regression)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9719 -0.2047  0.1025  0.1937  0.9461

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.3905459  0.1099572   3.552 0.000402 ***
Age            0.0067956  0.0014418   4.713 2.82e-06 ***
Resting_BP     0.0004635  0.0007109   0.652 0.514576
St_slope_Down -0.0701849  0.0519316  -1.351 0.176874
St_slope_Up   -0.5981079  0.0271702 -22.013  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3847 on 913 degrees of freedom
Multiple R-squared:  0.4044,    Adjusted R-squared:  0.4018
F-statistic:   155 on 4 and 913 DF,  p-value: < 2.2e-16

>
```

4. <u>Regression Plot:</u>



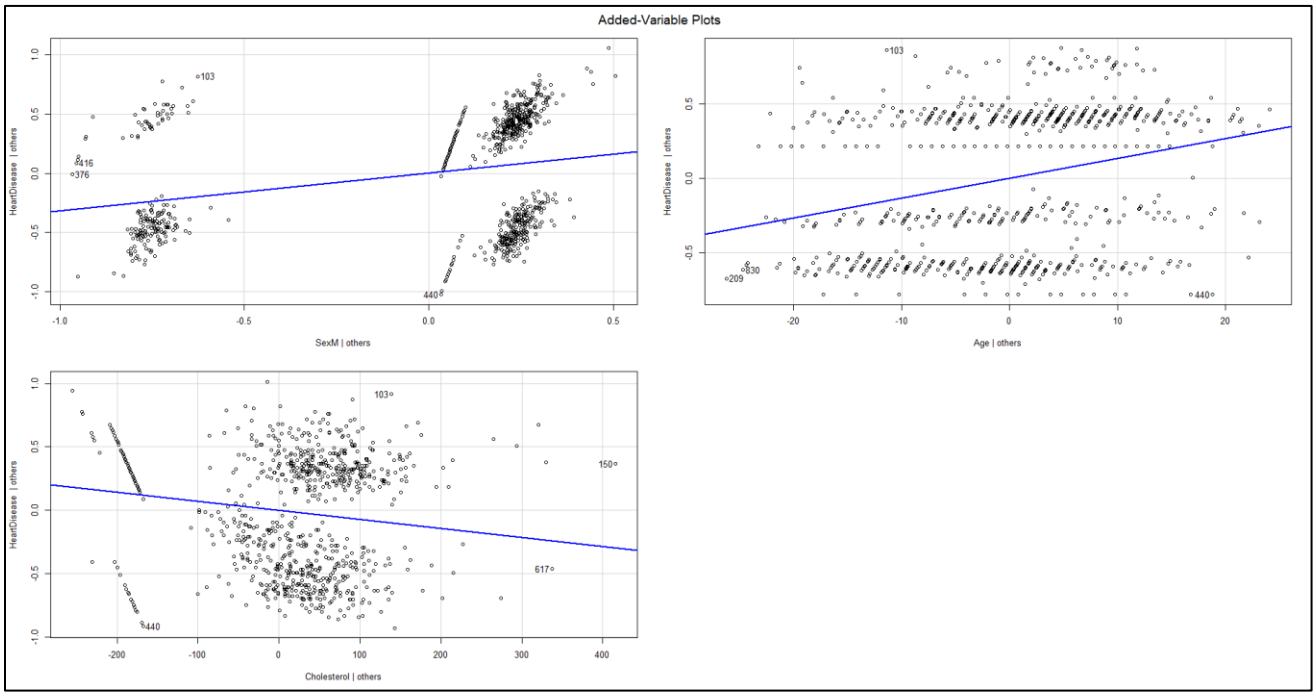
In the second part, we created a MV regression model again on numeric as well as categorical data but this time the dummy variables were automatically created by R. The output of the same is as show below.

Output:

1. Regression Model:

```
Console  Terminal
R 3.6.3 · ~/
> #SUBSET 2
>
> #dummy variables created by R for gender attribute and MV regression model
> model1 <- lm(HeartDisease ~ Sex + Age + Cholesterol, data = heart_dataset)
> model1

Call:
lm(formula = HeartDisease ~ Sex + Age + Cholesterol, data = heart_dataset)

Coefficients:
(Intercept)         SexM          Age  Cholesterol
 -0.2683913    0.3172343    0.0133214   -0.0007123

> summary(model1)

Call:
lm(formula = HeartDisease ~ Sex + Age + Cholesterol, data = heart_dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0346 -0.4223  0.1386  0.3700  1.0148

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.2683913  0.0976389  -2.749   0.0061 **
SexM         0.3172343  0.0371616   8.537  < 2e-16 ***
Age          0.0133214  0.0015809   8.426  < 2e-16 ***
Cholesterol -0.0007123  0.0001389  -5.127 3.59e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4492 on 914 degrees of freedom
Multiple R-squared:  0.1871,    Adjusted R-squared:  0.1845
F-statistic: 70.14 on 3 and 914 DF,  p-value: < 2.2e-16

>
```

2. Regression Plot:



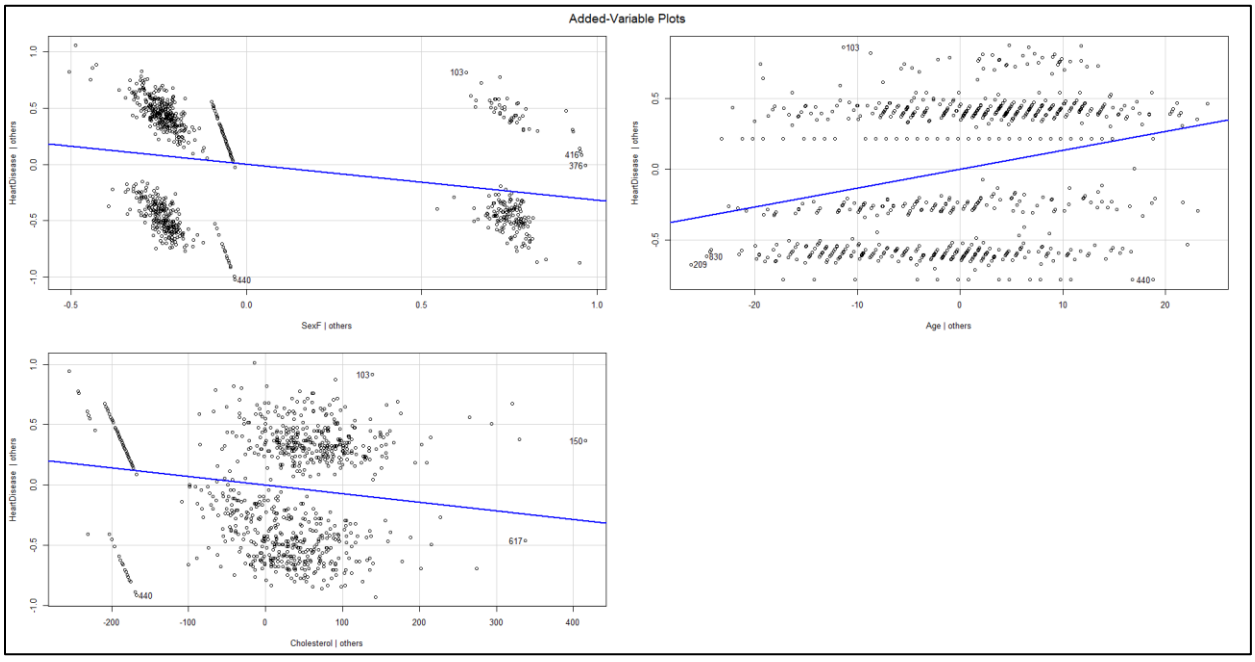
Now, using the relevel function for regression model we get the following output.

Output:

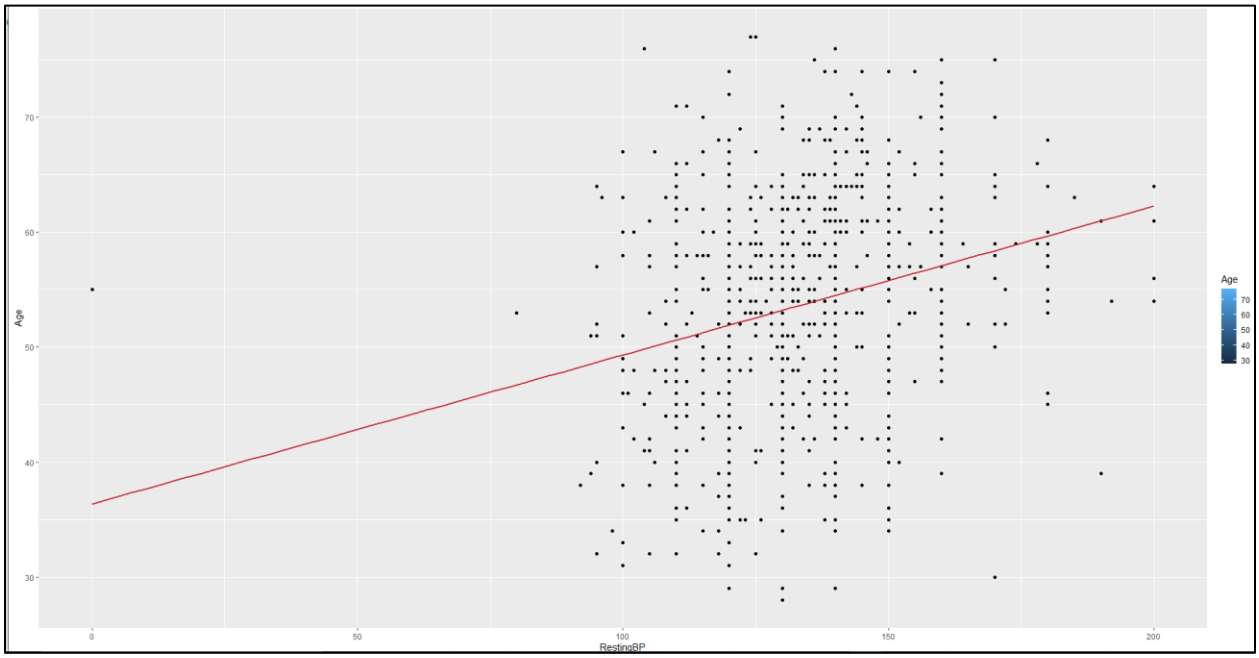
1. Regression Model:

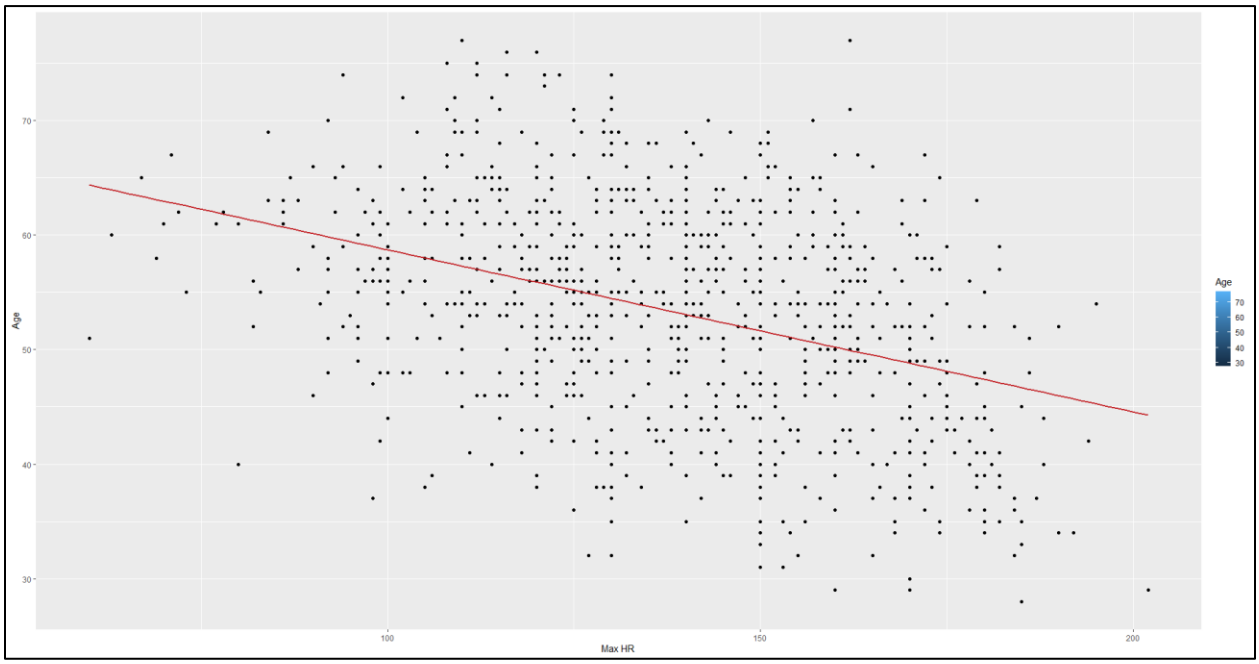
```
Console  Terminal
R 3.6.3 · ~/
> #using the re level function for regression model
> heart_dataset <- heart_dataset %>%
+   mutate(Sex = relevel(Sex, ref = "M"))
> #using the re level function for regression model
> heart_dataset <- heart_dataset %>%
+   mutate(Sex = relevel(Sex, ref = "M"))
>
>
> #regression model
> model2 <- lm(HeartDisease ~ Sex + Age + Cholesterol , data = heart_dataset)
> model2

Call:
lm(formula = HeartDisease ~ Sex + Age + Cholesterol, data = heart_dataset)

Coefficients:
(Intercept)         SexF          Age  Cholesterol
  0.0488431   -0.3172343    0.0133214   -0.0007123

> summary(model2)

Call:
lm(formula = HeartDisease ~ Sex + Age + Cholesterol, data = heart_dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0346 -0.4223  0.1386  0.3700  1.0148

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0488431  0.0925595   0.528    0.598
SexF        -0.3172343  0.0371616  -8.537  < 2e-16 ***
Age          0.0133214  0.0015809   8.426  < 2e-16 ***
Cholesterol -0.0007123  0.0001389  -5.127 3.59e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4492 on 914 degrees of freedom
Multiple R-squared:  0.1871,    Adjusted R-squared:  0.1845
F-statistic: 70.14 on 3 and 914 DF,  p-value: < 2.2e-16

>
>
> #regression plot
> avPlots(model2)
>
```

2. Regression Plot:



3. Regression Graph:

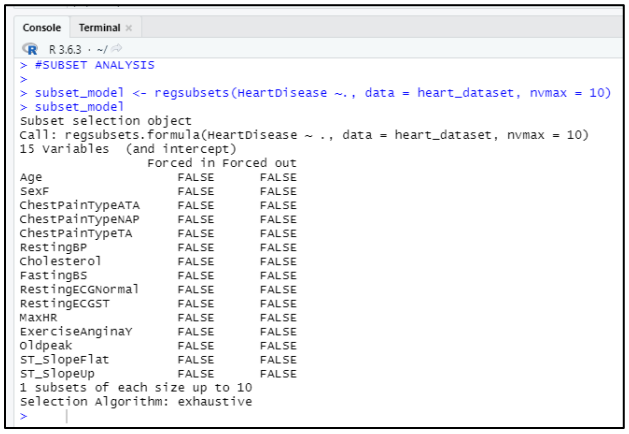




Subset analysis or subset regression is a model selection approach that consists of testing all possible combination of the predictor variables and then selecting the best model according to some of the statistical criteria. This analysis basically helps us to figure out which would be the best model for prediction based on the attributes we selected for prediction. Here, we performed a subset regression analysis to identify the different best models of different sizes. The R function regsubsets() was used for the same where the function returns up to the best 5-variables model. The output of the subset regression analysis performed is as shown below.

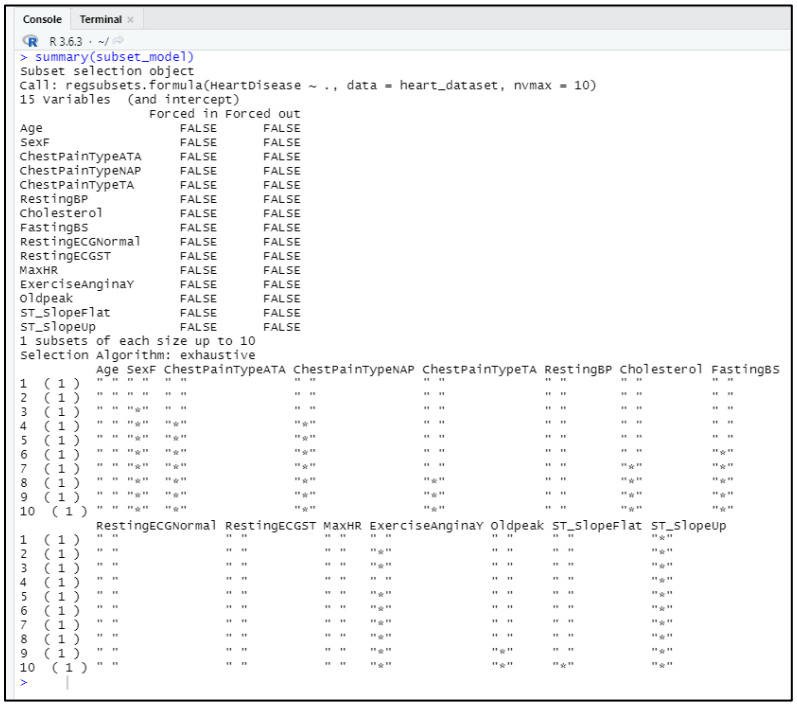
Output:

1. Subset model:



```
Console  Terminal
R 3.6.3 · ~/
> #SUBSET ANALYSIS
>
> subset_model <- regsubsets(HeartDisease ~., data = heart_dataset, nvmax = 10)
> subset_model
Subset selection object
Call: regsubsets.formula(HeartDisease ~ ., data = heart_dataset, nvmax = 10)
15 Variables  (and intercept)
                  Forced in Forced out
Age                   FALSE      FALSE
SexF                  FALSE      FALSE
ChestPainTypeATA      FALSE      FALSE
ChestPainTypeNAP      FALSE      FALSE
ChestPainTypeTA       FALSE      FALSE
RestingBP             FALSE      FALSE
Cholesterol           FALSE      FALSE
FastingBS             FALSE      FALSE
RestingECGNormal      FALSE      FALSE
RestingECGST          FALSE      FALSE
MaxHR                 FALSE      FALSE
ExerciseAnginaY       FALSE      FALSE
Oldpeak               FALSE      FALSE
ST_SlopeFlat          FALSE      FALSE
ST_SlopeUp            FALSE      FALSE
1 subsets of each size up to 10
Selection Algorithm: exhaustive
>
```

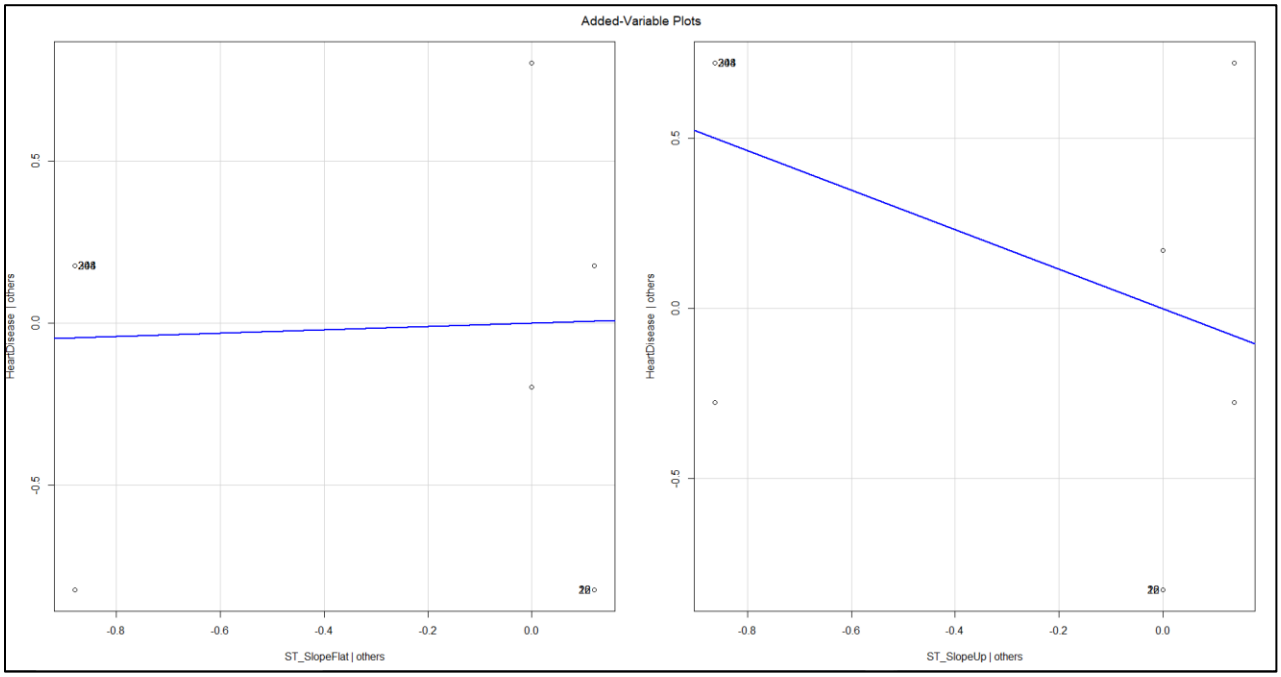
2. Summary of subset model:



```
Console  Terminal
R 3.6.3 · ~/
> summary(subset_model)
Subset selection object
Call: regsubsets.formula(HeartDisease ~ ., data = heart_dataset, nvmax = 10)
15 Variables  (and intercept)
                  Forced in Forced out
Age                   FALSE      FALSE
SexF                  FALSE      FALSE
ChestPainTypeATA      FALSE      FALSE
ChestPainTypeNAP      FALSE      FALSE
ChestPainTypeTA       FALSE      FALSE
RestingBP             FALSE      FALSE
Cholesterol           FALSE      FALSE
FastingBS             FALSE      FALSE
RestingECGNormal      FALSE      FALSE
RestingECGST          FALSE      FALSE
MaxHR                 FALSE      FALSE
ExerciseAnginaY       FALSE      FALSE
Oldpeak               FALSE      FALSE
ST_SlopeFlat          FALSE      FALSE
ST_SlopeUp            FALSE      FALSE
1 subsets of each size up to 10
Selection Algorithm: exhaustive
          Age SexF ChestPainTypeATA ChestPainTypeNAP ChestPainTypeTA RestingBP Cholesterol FastingBS
1  ( 1 )  " "  " "  " "            " "             " "             " "       " "         " "
2  ( 1 )  " "  " "  " "            " "             "*"             " "       " "         " "
3  ( 1 )  " "  " "  "*"            " "             " "             " "       " "         " "
4  ( 1 )  " "  " "  "*"            "*"             " "             " "       " "         " "
5  ( 1 )  " "  " "  "*"            "*"             "*"             " "       " "         " "
6  ( 1 )  " "  " "  "*"            "*"             "*"             " "       " "         "*"
7  ( 1 )  " "  " "  "*"            "*"             "*"             " "       " "         "*"
8  ( 1 )  " "  " "  "*"            "*"             "*"             " "       "*"         "*"
9  ( 1 )  " "  " "  "*"            "*"             "*"             " "       "*"         "*"
10  ( 1 ) " "  " "  "*"            "*"             "*"             " "       "*"         "*"
          RestingECGNormal RestingECGST MaxHR ExerciseAnginaY Oldpeak ST_SlopeFlat ST_SlopeUp
1  ( 1 )  " "              " "          " "   " "            " "     " "          "*"
2  ( 1 )  " "              " "          " "   "*"            " "     " "          "*"
3  ( 1 )  " "              " "          " "   "*"            " "     " "          "*"
4  ( 1 )  " "              " "          " "   " "            " "     " "          "*"
5  ( 1 )  " "              " "          " "   "*"            " "     " "          "*"
6  ( 1 )  " "              " "          " "   "*"            " "     " "          "*"
7  ( 1 )  " "              " "          " "   "*"            " "     " "          "*"
8  ( 1 )  " "              " "          " "   "*"            " "     " "          "*"
9  ( 1 )  " "              " "          " "   "*"            "*"     " "          "*"
10  ( 1 ) " "              " "          " "   "*"            "*"     "*"          "*"
>
```

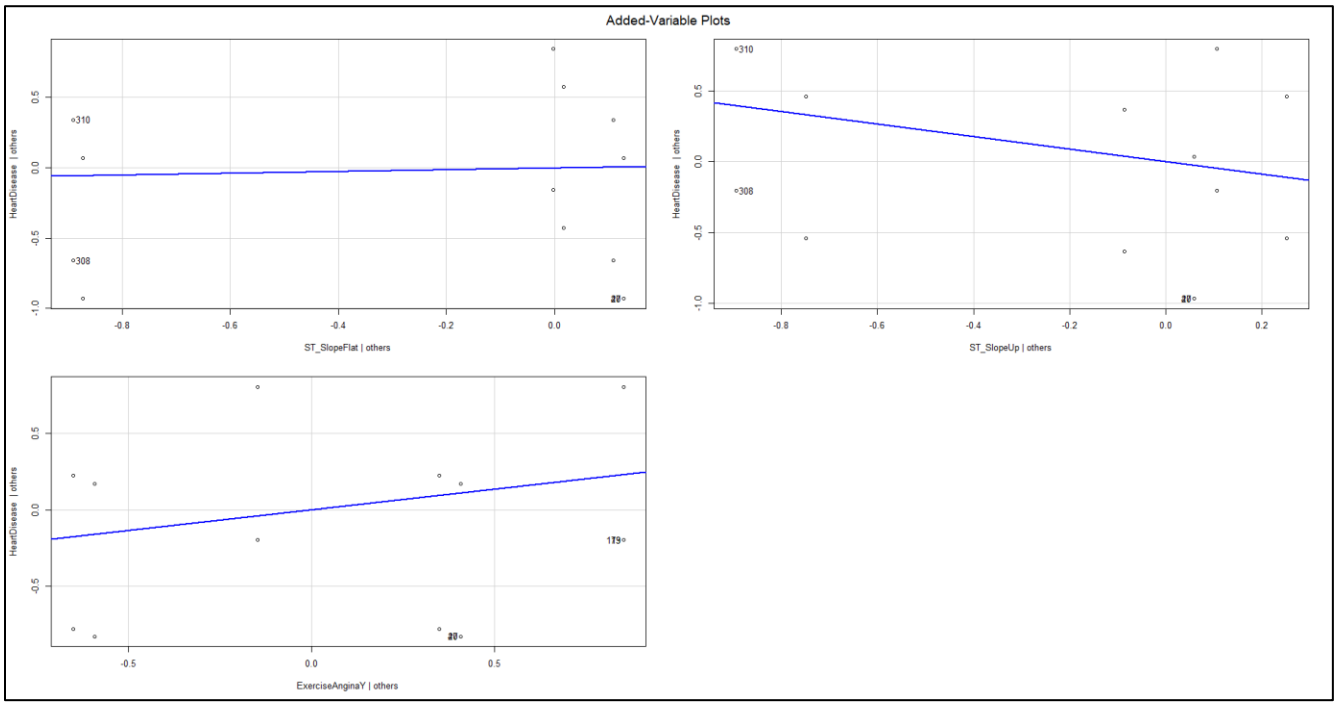
3. Best 1 variable model:

```
Console  Terminal
R 3.6.3 · ~/
> #best 1 variable model
> best1_model <- lm(HeartDisease ~ ST_Slope, data = heart_dataset)
> best1_model

Call:
lm(formula = HeartDisease ~ ST_Slope, data = heart_dataset)

Coefficients:
 (Intercept)  ST_SlopeFlat    ST_SlopeUp
     0.77778       0.05048      -0.58031

>
> summary(best1_model)

Call:
lm(formula = HeartDisease ~ ST_Slope, data = heart_dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8283 -0.1975  0.1717  0.1717  0.8025

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.77778    0.04909  15.844   <2e-16 ***
ST_SlopeFlat  0.05048    0.05234   0.964    0.335
ST_SlopeUp   -0.58031    0.05286 -10.978   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3896 on 915 degrees of freedom
Multiple R-squared:  0.3877,    Adjusted R-squared:  0.3864
F-statistic: 289.7 on 2 and 915 DF,  p-value: < 2.2e-16

>
```



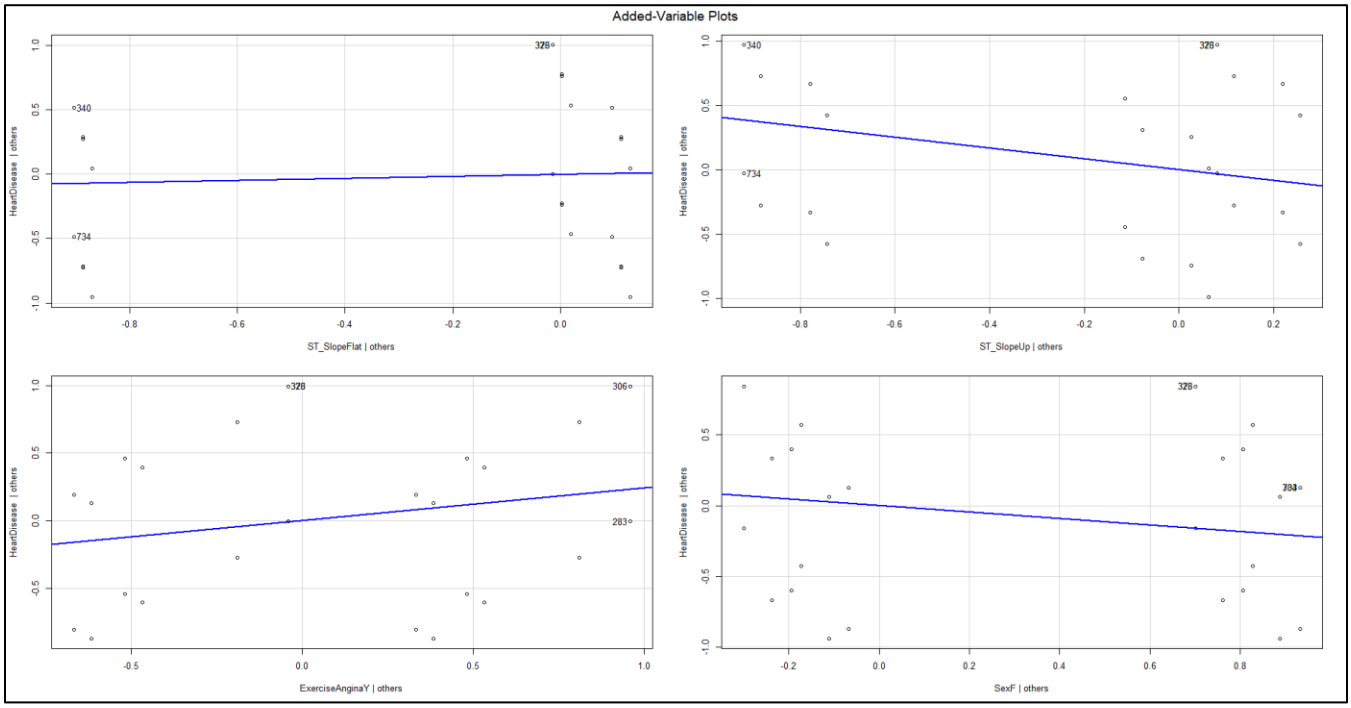
4. Best 2 variable model:

```
Console  Terminal
R 3.6.3 · ~/
> #best 2 variable model
> best2_model <- lm(HeartDisease  ~ ST_Slope + ExerciseAngina, data = heart_dataset)
> best2_model

Call:
lm(formula = HeartDisease ~ ST_Slope + ExerciseAngina, data = heart_dataset)

Coefficients:
    (Intercept)     ST_SlopeFlat       ST_SlopeUp  ExerciseAnginaY
        0.60158          0.06659         -0.44387          0.27074

>
> summary(best2_model)

Call:
lm(formula = HeartDisease ~ ST_Slope + ExerciseAngina, data = heart_dataset)

Residuals:
     Min       1Q   Median       3Q      Max
-0.93891 -0.15771  0.06109  0.12768  0.84229

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.60158    0.05024  11.975   <2e-16 ***
ST_SlopeFlat   0.06659    0.04993   1.334    0.183
ST_SlopeUp    -0.44387    0.05234  -8.481   <2e-16 ***
ExerciseAnginaY  0.27074    0.02808   9.642   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3714 on 914 degrees of freedom
Multiple R-squared:  0.4442,    Adjusted R-squared:  0.4424
F-statistic: 243.5 on 3 and 914 DF,  p-value: < 2.2e-16

>
```



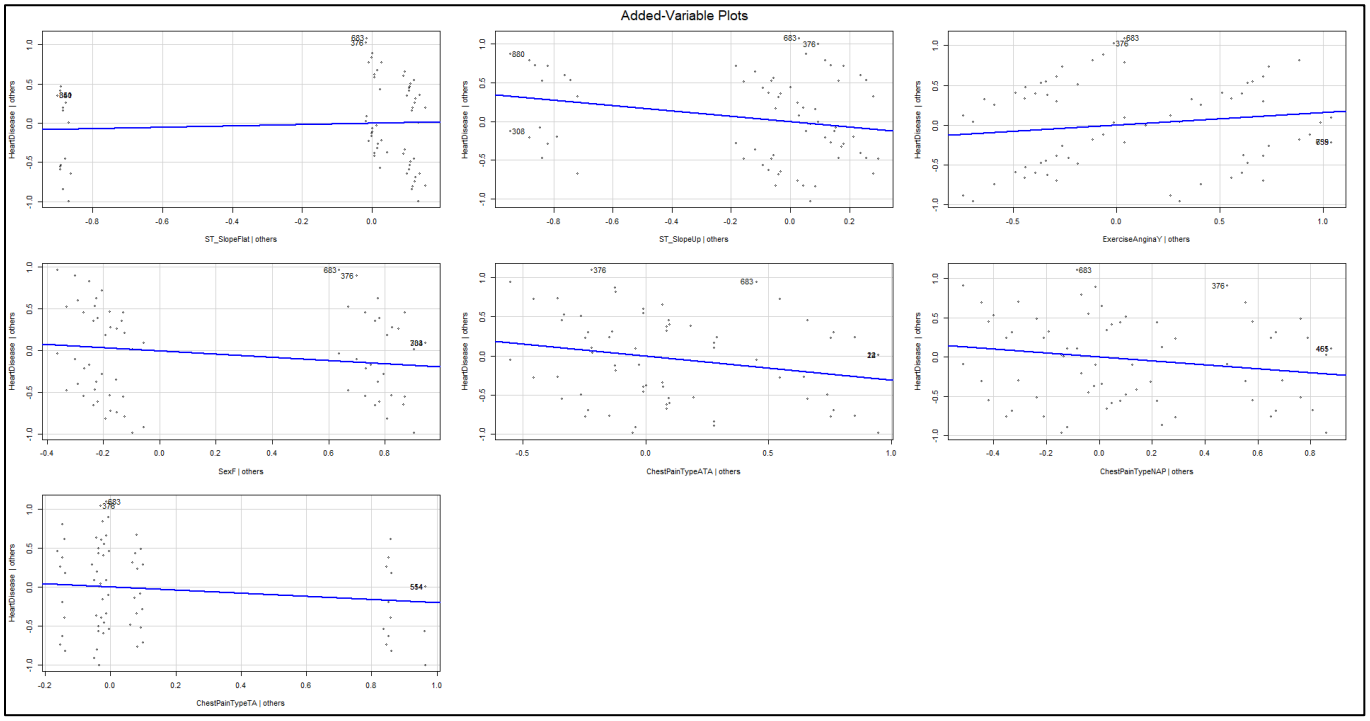
5. Best 3 variable model:

```
Console  Terminal
R 3.6.3 · ~/
> #best 3 variable model
> best3_model <- lm(HeartDisease  ~ ST_Slope + ExerciseAngina + Sex, data = heart_dataset)
> best3_model

Call:
lm(formula = HeartDisease ~ ST_Slope + ExerciseAngina + Sex,
    data = heart_dataset)

Coefficients:
    (Intercept)     ST_SlopeFlat       ST_SlopeUp  ExerciseAnginaY             SexF
        0.64599          0.07677         -0.41957          0.24165         -0.22928

>
> summary(best3_model)

Call:
lm(formula = HeartDisease ~ ST_Slope + ExerciseAngina + Sex,
    data = heart_dataset)

Residuals:
     Min       1Q   Median       3Q      Max
-0.96441 -0.22642  0.03559  0.26488  1.00287

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.64599    0.04905  13.170  < 2e-16 ***
ST_SlopeFlat     0.07677    0.04843   1.585    0.113
ST_SlopeUp      -0.41957    0.05085  -8.252 5.43e-16 ***
ExerciseAnginaY  0.24165    0.02749   8.791  < 2e-16 ***
SexF            -0.22928    0.02981  -7.692 3.74e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3601 on 913 degrees of freedom
Multiple R-squared:  0.4781,    Adjusted R-squared:  0.4758
F-statistic: 209.1 on 4 and 913 DF,  p-value: < 2.2e-16

>
```



6. Best 4 variable model:

```
Console  Terminal
R 3.6.3 · ~/
> #best 4 variable model
> best4_model <- lm(HeartDisease  ~ ST_Slope + ExerciseAngina + Sex + ChestPainType, data = heart_dataset)
> best4_model

Call:
lm(formula = HeartDisease ~ ST_Slope + ExerciseAngina + Sex +
    ChestPainType, data = heart_dataset)

Coefficients:
      (Intercept)       ST_SlopeFlat         ST_SlopeUp   ExerciseAnginaY             SexF  ChestPainTypeATA
          0.76521            0.08189           -0.35046            0.15605           -0.19533           -0.30967
ChestPainTypeNAP   ChestPainTypeTA
         -0.25297           -0.19511

>
> summary(best4_model)

Call:
lm(formula = HeartDisease ~ ST_Slope + ExerciseAngina + Sex +
    ChestPainType, data = heart_dataset)

Residuals:
     Min       1Q   Median       3Q      Max
-1.00315 -0.16179 -0.00315  0.15290  1.09024

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.76521    0.04799  15.944  < 2e-16 ***
ST_SlopeFlat     0.08189    0.04578   1.789 0.073984 .
ST_SlopeUp      -0.35046    0.04870  -7.196 1.30e-12 ***
ExerciseAnginaY  0.15605    0.02738   5.700 1.62e-08 ***
SexF            -0.19533    0.02837  -6.884 1.08e-11 ***
ChestPainTypeATA -0.30967    0.03420  -9.055  < 2e-16 ***
ChestPainTypeNAP -0.25297    0.02999  -8.436  < 2e-16 ***
ChestPainTypeTA  -0.19511    0.05381  -3.626 0.000304 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3403 on 910 degrees of freedom
Multiple R-squared:  0.5354,    Adjusted R-squared:  0.5318
F-statistic: 149.8 on 7 and 910 DF,  p-value: < 2.2e-16

>
```



Analysis

The correlation plot gave an overview of the relationship between the attributes of the dataset. Once the relationship was determined between the variables, we understood the positive correlation and negative correlation between them which further helped in our analysis to create subset of the dataset for the regression model. With respect to our dataset, there are categorical variables for which dummy variables needed to be created. Subset analysis was performed to find the best fit model in order to predict the heart disease of the person based on the predictor variable.

A subset_model variable was created which stored the best fit model of the dataset and we displayed the summary of the same. The summary of the subset regression model returned the best set of variables for each model size. From the output above, an asterisk specifies that a given variable is included in the corresponding model. Thus, depending on the asterisk we would get the best fit variables for our regression model which would help to choose the optimal model. After the subset regression model was performed and we got an overview of the best fit variables, regression model was built for these best fit variables in order to check the performance of these models to better predict the heart disease of the person.

We performed the best 1 variable model which had the st_slope attribute for predicting the heart disease, the best 2 variable model which had the st_slope and exercise angina attribute, the best 3 variable model having the st_slope, exercise angina and gender attribute and the best 4 variable model which had the st_slope, exercise angina, gender, and the chest pain type attribute for the prediction of the heart disease. The regression model was built for all these best fit variables along with the avplots which is the plot for MV regression.

The output of the regression model along with the plots helped in analyzing and predicting the heart disease based on the various predictor variables. The summary of the regression model returned statistical values of the model helping to analyze further on the model. The r-squared value came out to be 0.5354 for the best 4 variable model which is the highest compared to the other best variable models which could be considered for predicting the heart disease. Similarly, the p-values were also returned where it was less than 0.05 in the best 4 variable model and thus, we rejected the null hypothesis.

Summary

The heart failure prediction dataset helps in predicting the heart disease of the person via the regression model. This dataset consisted of both numerical and categorical values and in order to have categorical values in the regression model, dummy variables were created for gender and st_slope based on the dataset to build a regression model. As we know, regression model is used for predicting and examining the relationship between the attributes of the dataset. But before performing the regression analysis in order to predict the heart disease of the person with respect to the predictor variables, subset analysis was performed. Subset analysis allows the model selection approach consisting of testing all possible combination of the predictor variables and then selecting the best model according to the statistical criteria.

Now, in order to perform the regression analysis, it was important to have an understanding about the dataset for which descriptive analysis and data visualizations were performed giving us some insights about the dataset. The statistical values were computed in the descriptive analysis for the various attributes of the heart prediction dataset and visualizations like the boxplot, bar plot and scatter plot with regression lines were plotted. These visualizations gave an overall analysis about the dataset and which factor would lead to the heart disease for a person.

After the descriptive analysis and data visualizations were performed, correlation plot and regression model were built for the subset of dataset created for the categorical variables. The correlation chart was plotted for the attributes of the dataset to determine the correlation between them for analysis. MV regression model was built for the categorical attributes of the dataset and avplots were plotted for the same. The summary of the regression models returned the r-squared value and other statistical values such as the p-value which helped in further analysis of the dataset. Subset analysis helps in determining which would be the best model based on the predictor variables. The subset regression model was built for the dataset of the heart failure prediction and the summary of this model returned the best 5 variable model which could be used to perform the regression model and predict the heart disease of a person.

This analysis is much more ideal as compared to the normal regression model built as it allows us to choose the predictor variable for the regression model. The regression model built gave a r-squared value lesser as compared to the best 4 variable model built, which implied that the regression line was not best fitted when performing the analysis for the attributes as compared to when done for the subset analysis.

References

Zach, S. (2021b, February 2). How to Create Dummy Variables in R (Step-by-Step). Statology.Org. Retrieved December 15, 2021, from https://www.statology.org/dummy-variables-in-r/

Kassambara. (2018, March 11). Regression with Categorical Variables: Dummy Coding Essentials in R. STHDA.Com. Retrieved December 15, 2021, from http://sthda.com/english/articles/40-regression-analysis/163-regression-with-categorical-variables-dummy-coding-essentials-in-r/

Zach, S. (2020b, December 23). How to Plot Multiple Linear Regression Results in R. Statology.Org. Retrieved December 15, 2021, from https://www.statology.org/plot-multiple-linear-regression-in-r/

Kassambara. (2018a, March 11). Best Subsets Regression Essentials in R. STHDA.Com. Retrieved December 16, 2021, from http://sthda.com/english/articles/37-model-selection-essentials-in-r/155-best-subsets-regression-essentials-in-r

Regression with Categorical Variables. (n.d.). Faculty.Nps.Edu. Retrieved December 16, 2021, from https://faculty.nps.edu/rbassett/_book/regression-with-categorical-variables.html

Marsja, E. (2020, May 24). How to Create Dummy Variables in R (with Examples). Marsja.Se. Retrieved December 16, 2021, from https://www.marsja.se/create-dummy-variables-in-r/