MultiVariate Analysis Individual Project

Name: Sailesh Potturi

Email : sp2245@scarletmail.rutgers.edu

Github Link:

https://github.com/SaileshRBS/Multi-Variate-Analysis.git

Dataset: Heart Disease

The provided data appears to be a tabular dataset with various columns representing different attributes for each individual. The dataset consists of **4241** Rows and **15** Columns. The 15th column(Heart Disease) is the one where we need to determine whether the person has heart disease or not using the 14 variables.

Exploring the variables in the dataset:

1) Gender: binary variable indicating gender, where 1 represents male and 0 represents female.

2) age: age of the individual.

3) education: level of education, represented as a categorical variable with values 1, 2, 3, and 4.

4) currentSmoker: binary variable indicating if the individual is a current smoker, where 1 represents smoker and 0 represents non-smoker.

5) cigsPerDay: number of cigarettes smoked per day for current smokers.

6) BPMeds: binary variable indicating if the individual is on blood pressure medication, where 1 represents on medication and 0 represents not on medication.

7) prevalentHyp: binary variable indicating if the individual has prevalent hypertension, where 1 represents has hypertension and 0 represents no hypertension.

8) diabetes: binary variable indicating if the individual has diabetes, where 1 represents has diabetes and 0 represents no diabetes.

9) totChol: total cholesterol level of the individual.

10 sysBP: systolic blood pressure of the individual.

11) diaBP: diastolic blood pressure of the individual.

12) BMI: body mass index of the individual.

13) heart Rate: heart rate of the individual.

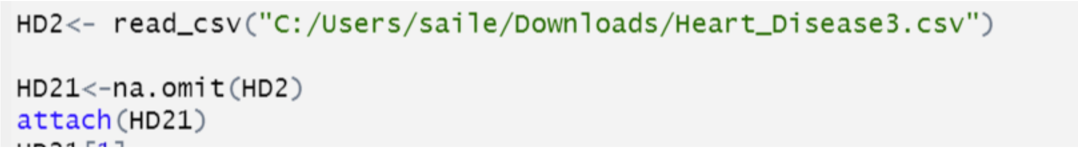
14) glucose: glucose level of the individual.

15) Heart Disease: binary variable indicating if the individual has heart disease, where 1 represents heart disease and 0 represents no heart disease.

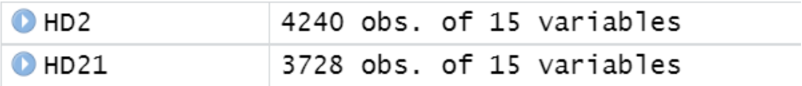
The dataset contains information about various risk factors and health attributes of individuals, and it may be used for analysis or modeling to study the relationships between these factors and heart disease.

***Data Cleaning:**

- When observed there are few values which contain NA values.
- Na.omit() function has being used and the rows have being deleted.



```
HD2<- read_csv("C:/Users/saile/Downloads/Heart_Disease3.csv")

HD21<-na.omit(HD2)
attach(HD21)
```



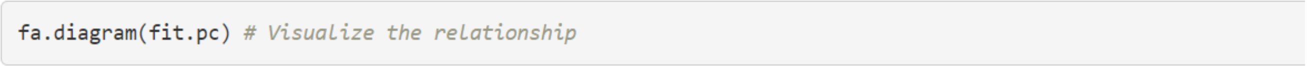
```
HD2     4240 obs. of 15 variables
HD21    3728 obs. of 15 variables
```

1^{st} Question

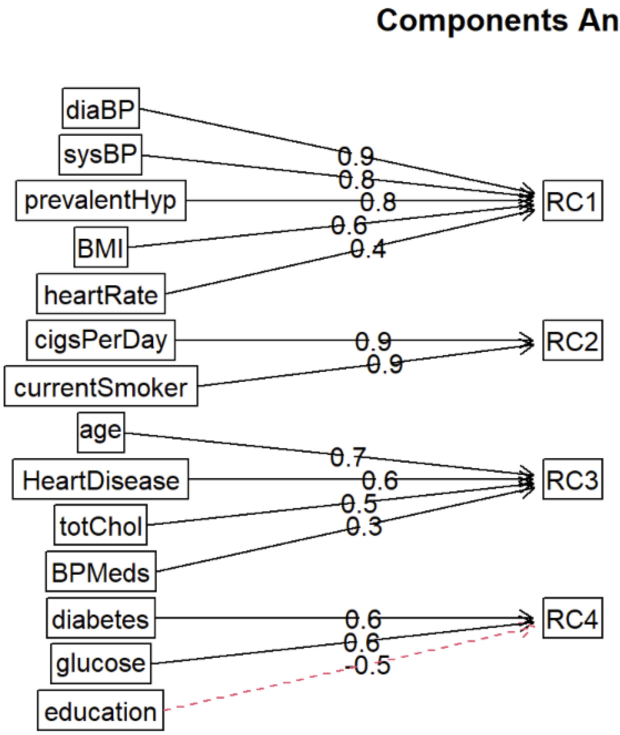
Identify the important factors underlying the observed variables and examine the relationships between the Heart Disease with respect to these factors.

Answer:

- Firstly, we will use Factor Analysis to reduce the number of columns by which we can see that the columns will get in to multiple factors. We can also observe the visualization the relationship.



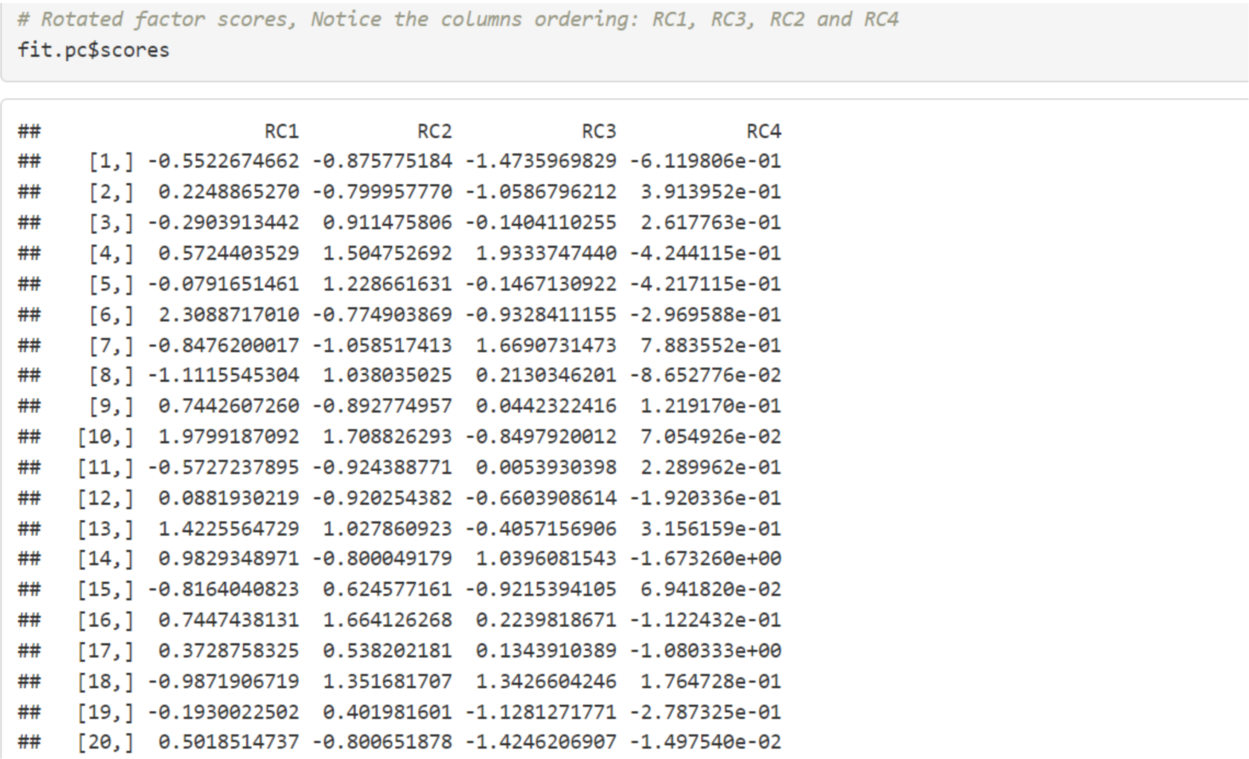
```
fa.diagram(fit.pc) # Visualize the relationship
```



Components Analysis

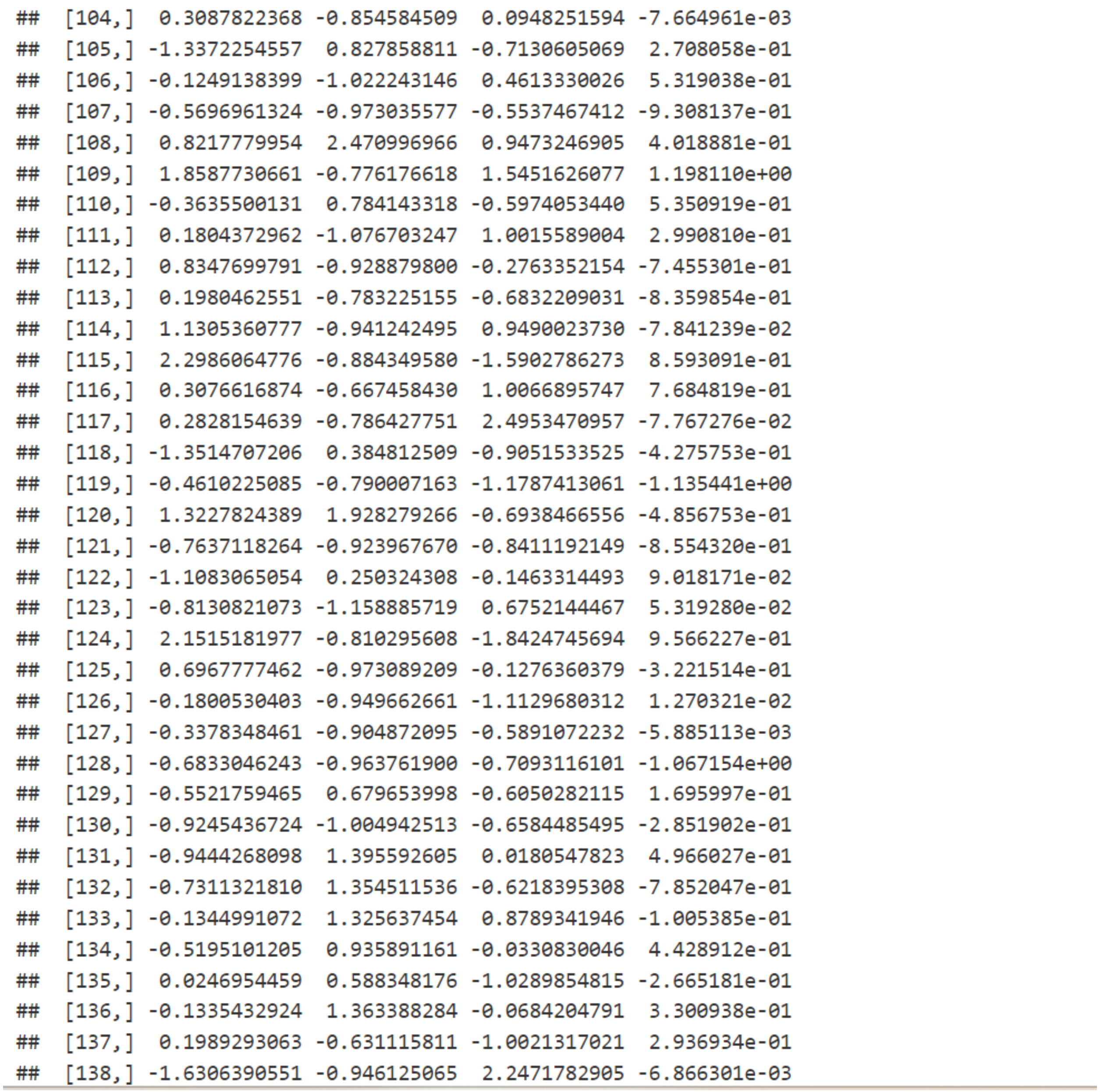
- **RC1**: This is trying to tell us that these columns diaBP, sysBP, prevalentHyp, BMI and heartRate are giving me the data related to heart.
- **RC2**: The columns cigsPerDay,currentSmoker is showing data related to consumptions of cigarettes per day of an individual.
- **RC3**: HeartDisease , totChol,BPmeds column are giving data related to the factors which can affect heart issues or health problems related to heart.
- **RC4**: diabetes, glucose these columns are related to problem sugar in an individual which can affect the heart and cause HeartDisease.

Now, we see the fit.pc$score values by the columns RC1, RC2, RC3 and RC4. Since the dataset has many columns in it. The below code shows random columns of the RC1, RC2, RC3 and RC4

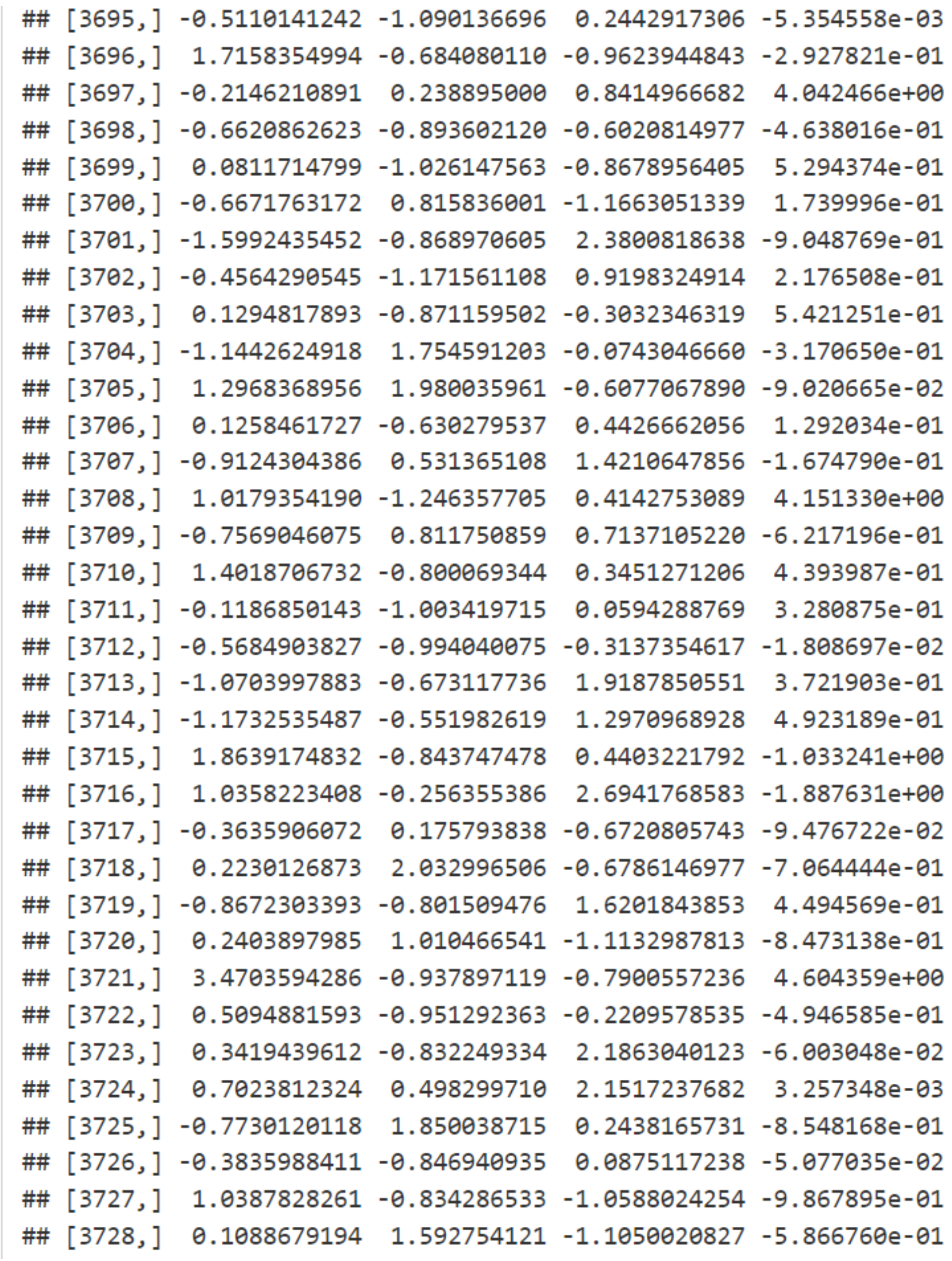


```
# Rotated factor scores, Notice the columns ordering: RC1, RC3, RC2 and RC4
fit.pc$scores
```

```
##                RC1          RC2          RC3           RC4
##    [1,] -0.5522674662 -0.875775184 -1.4735969829 -6.119806e-01
##    [2,]  0.2248865270 -0.799957770 -1.0586796212  3.913952e-01
##    [3,] -0.2903913442  0.911475806 -0.1404110255  2.617763e-01
##    [4,]  0.5724403529  1.504752692  1.9333747440 -4.244115e-01
##    [5,] -0.0791651461  1.228661631 -0.1467130922 -4.217115e-01
##    [6,]  2.3088717010 -0.774903869 -0.9328411155 -2.969588e-01
##    [7,] -0.8476200017 -1.058517413  1.6690731473  7.883552e-01
##    [8,] -1.1115545304  1.038035025  0.2130346201 -8.652776e-02
##    [9,]  0.7442607260 -0.892774957  0.0442322416  1.219170e-01
##   [10,]  1.9799187092  1.708826293 -0.8497920012  7.054926e-02
##   [11,] -0.5727237895 -0.924388771  0.0053930398  2.289962e-01
##   [12,]  0.0881930219 -0.920254382 -0.6603908614 -1.920336e-01
##   [13,]  1.4225564729  1.027860923 -0.4057156906  3.156159e-01
##   [14,]  0.9829348971 -0.800049179  1.0396081543 -1.673260e+00
##   [15,] -0.8164040823  0.624577161 -0.9215394105  6.941820e-02
##   [16,]  0.7447438131  1.664126268  0.2239818671 -1.122432e-01
##   [17,]  0.3728758325  0.538202181  0.1343910389 -1.080333e+00
##   [18,] -0.9871906719  1.351681707  1.3426604246  1.764728e-01
##   [19,] -0.1930022502  0.401981601 -1.1281271771 -2.787325e-01
##   [20,]  0.5018514737 -0.800651878 -1.4246206907 -1.497540e-02
```

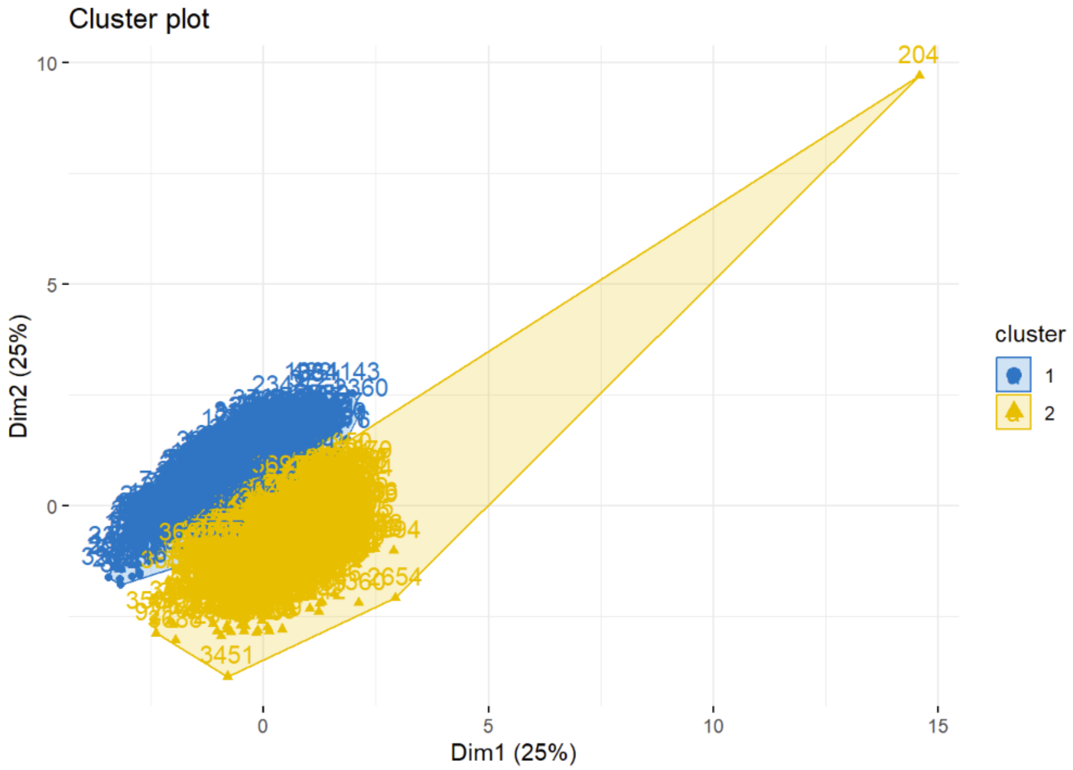


```
## [104,]  0.3087822368 -0.854584509  0.0948251594 -7.664961e-03
## [105,] -1.3372254557  0.827858811 -0.7130605069  2.708058e-01
## [106,] -0.1249138399 -1.022243146  0.4613330026  5.319038e-01
## [107,] -0.5696961324 -0.973035577 -0.5537467412 -9.308137e-01
## [108,]  0.8217779954  2.470996966  0.9473246905  4.018881e-01
## [109,]  1.8587730661 -0.776176618  1.5451626077  1.198110e+00
## [110,] -0.3635500131  0.784143318 -0.5974053440  5.350919e-01
## [111,]  0.1804372962 -1.076703247  1.0015589004  2.990810e-01
## [112,]  0.8347699791 -0.928879800 -0.2763352154 -7.455301e-01
## [113,]  0.1980462551 -0.783225155 -0.6832209031 -8.359854e-01
## [114,]  1.1305360777 -0.941242495  0.9490023730 -7.841239e-02
## [115,]  2.2986064776 -0.884349580 -1.5902786273  8.593091e-01
## [116,]  0.3076616874 -0.667458430  1.0066895747  7.684819e-01
## [117,]  0.2828154639 -0.786427751  2.4953470957 -7.767276e-02
## [118,] -1.3514707206  0.384812509 -0.9051533525 -4.275753e-01
## [119,] -0.4610225085 -0.790007163 -1.1787413061 -1.135441e+00
## [120,]  1.3227824389  1.928279266 -0.6938466556 -4.856753e-01
## [121,] -0.7637118264 -0.923967670 -0.8411192149 -8.554320e-01
## [122,] -1.1083065054  0.250324308 -0.1463314493  9.018171e-02
## [123,] -0.8130821073 -1.158885719  0.6752144467  5.319280e-02
## [124,]  2.1515181977 -0.810295608 -1.8424745694  9.566227e-01
## [125,]  0.6967777462 -0.973089209 -0.1276360379 -3.221514e-01
## [126,] -0.1800530403 -0.949662661 -1.1129680312  1.270321e-02
## [127,] -0.3378348461 -0.904872095 -0.5891072232 -5.885113e-03
## [128,] -0.6833046243 -0.963761900 -0.7093116101 -1.067154e+00
## [129,] -0.5521759465  0.679653998 -0.6050282115  1.695997e-01
## [130,] -0.9245436724 -1.004942513 -0.6584485495 -2.851902e-01
## [131,] -0.9444268098  1.395592605  0.0180547823  4.966027e-01
## [132,] -0.7311321810  1.354511536 -0.6218395308 -7.852047e-01
## [133,] -0.1344991072  1.325637454  0.8789341946 -1.005385e-01
## [134,] -0.5195101205  0.935891161 -0.0330830046  4.428912e-01
## [135,]  0.0246954459  0.588348176 -1.0289854815 -2.665181e-01
## [136,] -0.1335432924  1.363388284 -0.0684204791  3.300938e-01
## [137,]  0.1989293063 -0.631115811 -1.0021317021  2.936934e-01
## [138,] -1.6306390551 -0.946125065  2.2471782905 -6.866301e-03
```

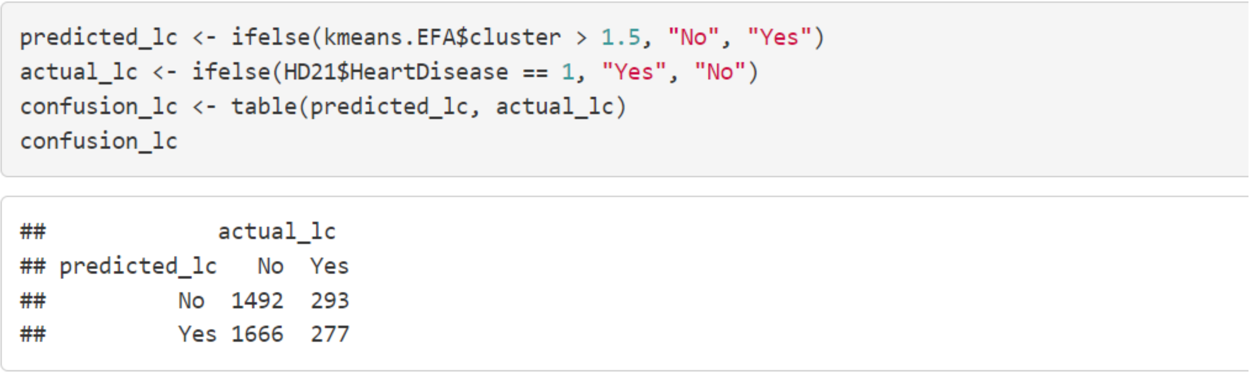


```
## [3695,] -0.5110141242 -1.090136696  0.2442917306 -5.354558e-03
## [3696,]  1.7158354994 -0.684080110 -0.9623944843 -2.927821e-01
## [3697,] -0.2146210891  0.238895000  0.8414966682  4.042466e+00
## [3698,] -0.6620862623 -0.893602120 -0.6020814977 -4.638016e-01
## [3699,]  0.0811714799 -1.026147563 -0.8678956405  5.294374e-01
## [3700,] -0.6671763172  0.815836001 -1.1663051339  1.739996e-01
## [3701,] -1.5992435452 -0.868970605  2.3800818638 -9.048769e-01
## [3702,] -0.4564290545 -1.171561108  0.9198324914  2.176508e-01
## [3703,]  0.1294817893 -0.871159502 -0.3032346319  5.421251e-01
## [3704,] -1.1442624918  1.754591203 -0.0743046660 -3.170650e-01
## [3705,]  1.2968368956  1.980035961 -0.6077067890 -9.020665e-02
## [3706,]  0.1258461727 -0.630279537  0.4426662056  1.292034e-01
## [3707,] -0.9124304386  0.531365108  1.4210647856 -1.674790e-01
## [3708,]  1.0179354190 -1.246357705  0.4142753089  4.151330e+00
## [3709,] -0.7569046075  0.811750859  0.7137105220 -6.217196e-01
## [3710,]  1.4018706732 -0.800069344  0.3451271206  4.393987e-01
## [3711,] -0.1186850143 -1.003419715  0.0594288769  3.280875e-01
## [3712,] -0.5684903827 -0.994040075 -0.3137354617 -1.808697e-02
## [3713,] -1.0703997883 -0.673117736  1.9187850551  3.721903e-01
## [3714,] -1.1732535487 -0.551982619  1.2970968928  4.923189e-01
## [3715,]  1.8639174832 -0.843747478  0.4403221792 -1.033241e+00
## [3716,]  1.0358223408 -0.256355386  2.6941768583 -1.887631e+00
## [3717,] -0.3635906072  0.175793838 -0.6720805743 -9.476722e-02
## [3718,]  0.2230126873  2.032996506 -0.6786146977 -7.064444e-01
## [3719,] -0.8672303393 -0.801509476  1.6201843853  4.494569e-01
## [3720,]  0.2403897985  1.010466541 -1.1132987813 -8.473138e-01
## [3721,]  3.4703594286 -0.937897119 -0.7900557236  4.604359e+00
## [3722,]  0.5094881593 -0.951292363 -0.2209578535 -4.946585e-01
## [3723,]  0.3419439612 -0.832249334  2.1863040123 -6.003048e-02
## [3724,]  0.7023812324  0.498299710  2.1517237682  3.257348e-03
## [3725,] -0.7730120118  1.850038715  0.2438165731 -8.548168e-01
## [3726,] -0.3835988411 -0.846940935  0.0875117238 -5.077035e-02
## [3727,]  1.0387828261 -0.834286533 -1.0588024254 -9.867895e-01
## [3728,]  0.1088679194  1.592754121 -1.1050020827 -5.866760e-01
```

- The 2nd part is that we implement Cluster Analysis using the fit.pc$score values of the RC's columns.
- Here , we scale the data derived from EFA with respect to the RC'S columns and clustering is performed.



- From the above plot we can see that there are many points formed in to 2 clusters 1 and cluster 2.
- But we can’t which cluster and its points determine Heart Disease.
- For a better understanding we perform confusion matrix for the above plot.



```
predicted_lc <- ifelse(kmeans.EFA$cluster > 1.5, "No", "Yes")
actual_lc <- ifelse(HD21$HeartDisease == 1, "Yes", "No")
confusion_lc <- table(predicted_lc, actual_lc)
confusion_lc
```

```
##             actual_lc
## predicted_lc   No  Yes
##          No  1492  293
##          Yes 1666  277
```

Based on the confusion matrix and also considering the RC values of the columns mentioned earlier there are 277 individuals who are affected with Heart Disease and 1492 who are not affected with heart disease. But the Predicted model says 1959 was wrongly predicted. In the earlier analysis I’ve applied PCA to the dataset but, the PC1 and PC2 components together are showing less than 50% which means PCA doesn’t work good for the HeartDisease dataset.

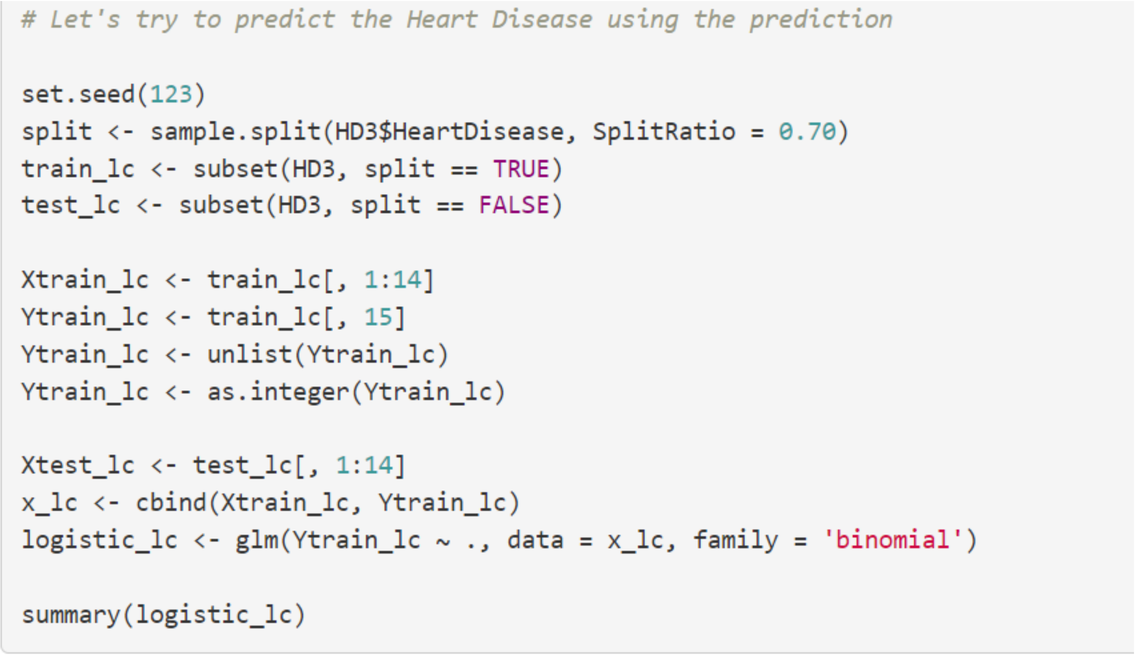
2ND Question:

Using the heart disease dataset predict the number of individuals who are affected with Heart Disease and the accuracy of the derived result

Answer:

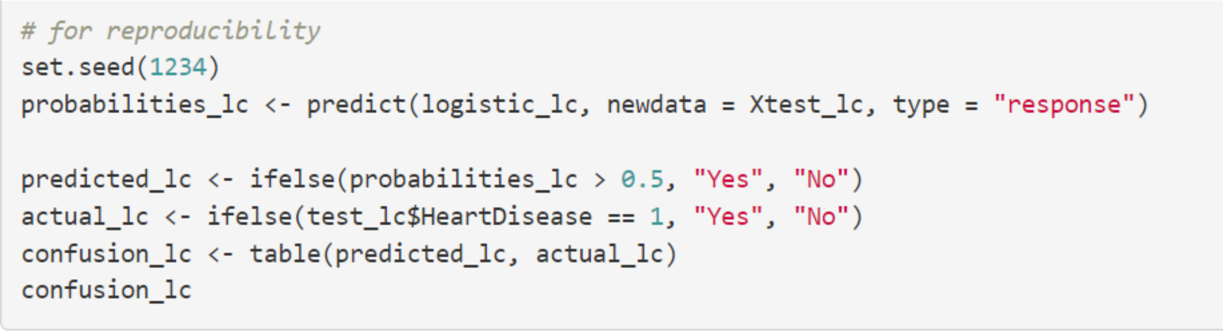
- The heart disease dataset is a type of dataset where we need to predict ‘1’ or ‘0’ or “Yes” or” No”. We use Logistic Regression for this.

- Firstly, We Split the dataset into Train and Test data with a Split ratio of 0.70.

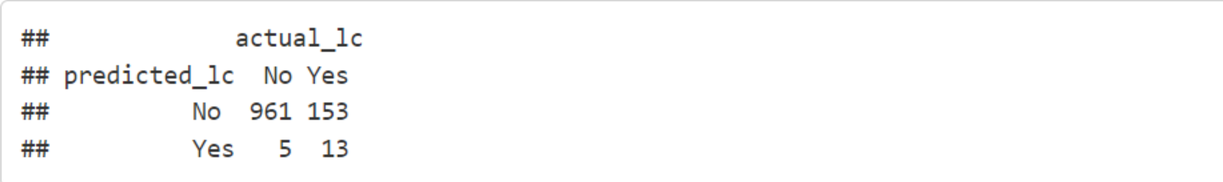


```
# Let's try to predict the Heart Disease using the prediction

set.seed(123)
split <- sample.split(HD3$HeartDisease, SplitRatio = 0.70)
train_lc <- subset(HD3, split == TRUE)
test_lc <- subset(HD3, split == FALSE)

Xtrain_lc <- train_lc[, 1:14]
Ytrain_lc <- train_lc[, 15]
Ytrain_lc <- unlist(Ytrain_lc)
Ytrain_lc <- as.integer(Ytrain_lc)

Xtest_lc <- test_lc[, 1:14]
x_lc <- cbind(Xtrain_lc, Ytrain_lc)
logistic_lc <- glm(Ytrain_lc ~ ., data = x_lc, family = 'binomial')

summary(logistic_lc)
```

- We use the confusion matrix to predict the values of the individuals who are affected with heart diseases.



```
# for reproducibility
set.seed(1234)
probabilities_lc <- predict(logistic_lc, newdata = Xtest_lc, type = "response")

predicted_lc <- ifelse(probabilities_lc > 0.5, "Yes", "No")
actual_lc <- ifelse(test_lc$HeartDisease == 1, "Yes", "No")
confusion_lc <- table(predicted_lc, actual_lc)
confusion_lc
```

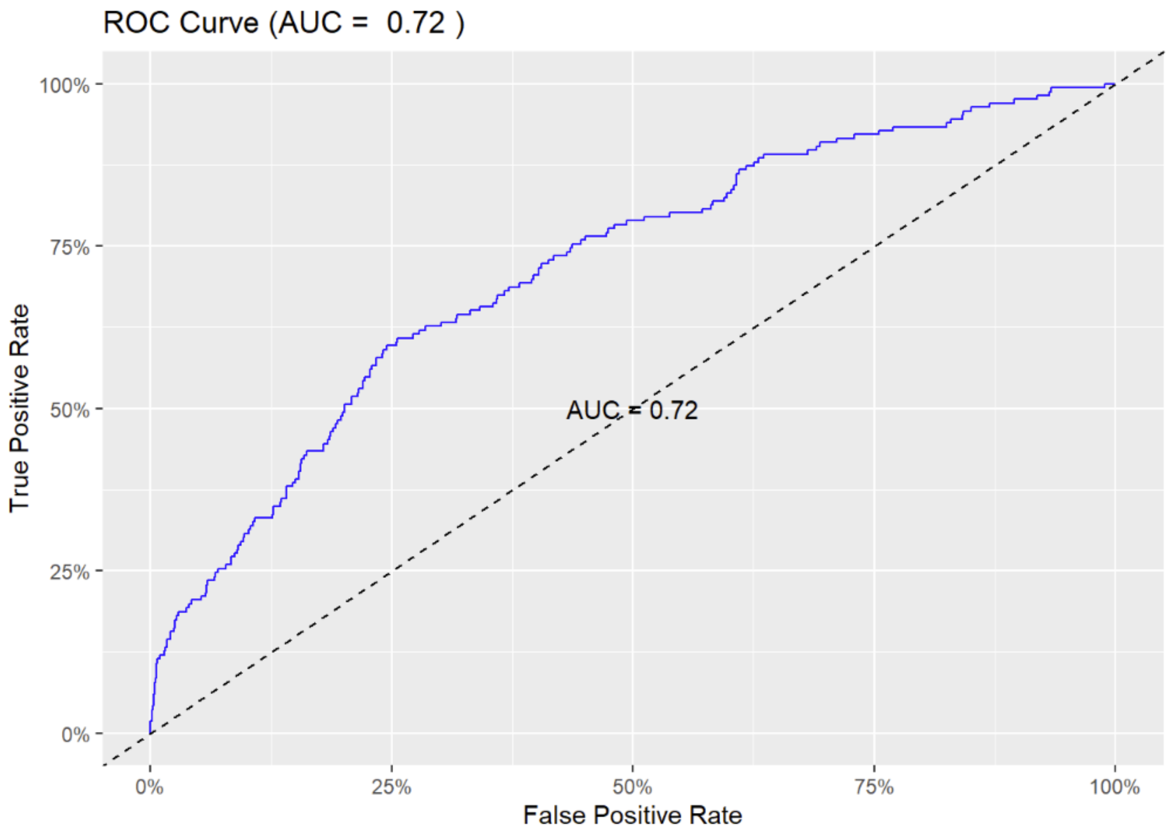


```
##             actual_lc
## predicted_lc  No Yes
##          No  961 153
##          Yes   5  13
```

From the above code we can conclude that there are 961 individuals who are not affected with heart disease and 13 are affected with heart Disease.

- **974 records were correctly predicted.**
- **158 records were wrongly predicted.**
- **153 records having heart disease were predicted as not having heart disease.**
- **5 records not having heart disease were predicted as having heart disease.**

The accuracy of the dataset can be shown using the ROC Curve



- **From the above plot we see the dataset accuracy is 0.72(72%). We can say that the predicted was not accurate to the mark basing on the confusion matrix mentioned above and needs more improvement in prediction.**

Github Link:

https://github.com/SaileshRBS/Multi-Variate-Analysis.git