REPORT ON

MULTIVARIATE ANALYSIS

Individual Class Project

Source of dataset:

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney. The objective of the dataset is to diagnostically predict whether a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

Link: https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset

Data Dictionary:

Name of Variable	Description	Type of Value
Pregnancies	This field represents the number of times a person has been pregnant. It is a discrete variable with integer values ranging from 0 to a positive integer.	Discrete variable with integer values.
Glucose	This field represents the glucose level (measured in mg/dL) of the person. It is a continuous variable with numeric values representing the person's blood glucose concentration.	Continuous variable with numeric values.
Blood Pressure	This field represents the glucose level (measured in mg/dL) of the person. It is a continuous variable with numeric values representing the person's blood glucose concentration.	Continuous variable with numeric values.
Skin Thickness	This field represents the skin thickness (measured in mm) of the person. It is a continuous variable with numeric values representing the thickness of the skin at a particular location on the body.	Continuous variable with numeric values.
Insulin	This field represents the insulin level (measured in mu U/ml) of the person. It is a continuous variable with numeric values representing the person's insulin concentration.	Continuous variable with numeric values.
BMI	This field represents the Body Mass Index (BMI) of the person. It is a continuous variable with numeric values representing the person's BMI.	Continuous variable with numeric values.
Diabetes Pedigree Function	This field represents a function that scores the diabetes history of the person's ancestors. It is a continuous variable with numeric values representing the person's diabetes pedigree function score.	Continuous variable with numeric values.

Age	This field represents the age (in years) of the person. It is a discrete variable with integer values representing the person's age.	Discrete variable with integer values.
Outcome	This field represents the outcome of the person's diabetes diagnosis. It is a binary variable with values 0 or 1, where 0 indicates no diabetes and 1 indicates diabetes.	Binary variable with values 0 or 1.

Questions that I tried to answer (Hypothesis):

- Principal Component Analysis (PCA):

1. How many variables we can reduce to, by using Principal Component Analysis?
2. Are we able to identify the contribution of different features towards the overall variability or explained variance in the diabetes dataset?

- Clustering Analysis

1. Does clustering analysis help in identifying any pattern in the dataset, such as difference in glucose level, age, or insulin?

- Exploratory Factor Analysis

1. Can we reduce the number of factors using EFA, by identifying the minimum number of factors that capture the majority of the variance of the dataset?

- Multiple Regression

1. We are trying to predict the age of the person by using Multiple Regression, based on the other features like pregnancies, glucose, insulin, etc.

- Logistic Regression

1. We are trying to predict the person’s outcome (diabetic or non-diabetic) by using Logistic Regression, based on other features like glucose, insulin, age, etc.

Conclusion:

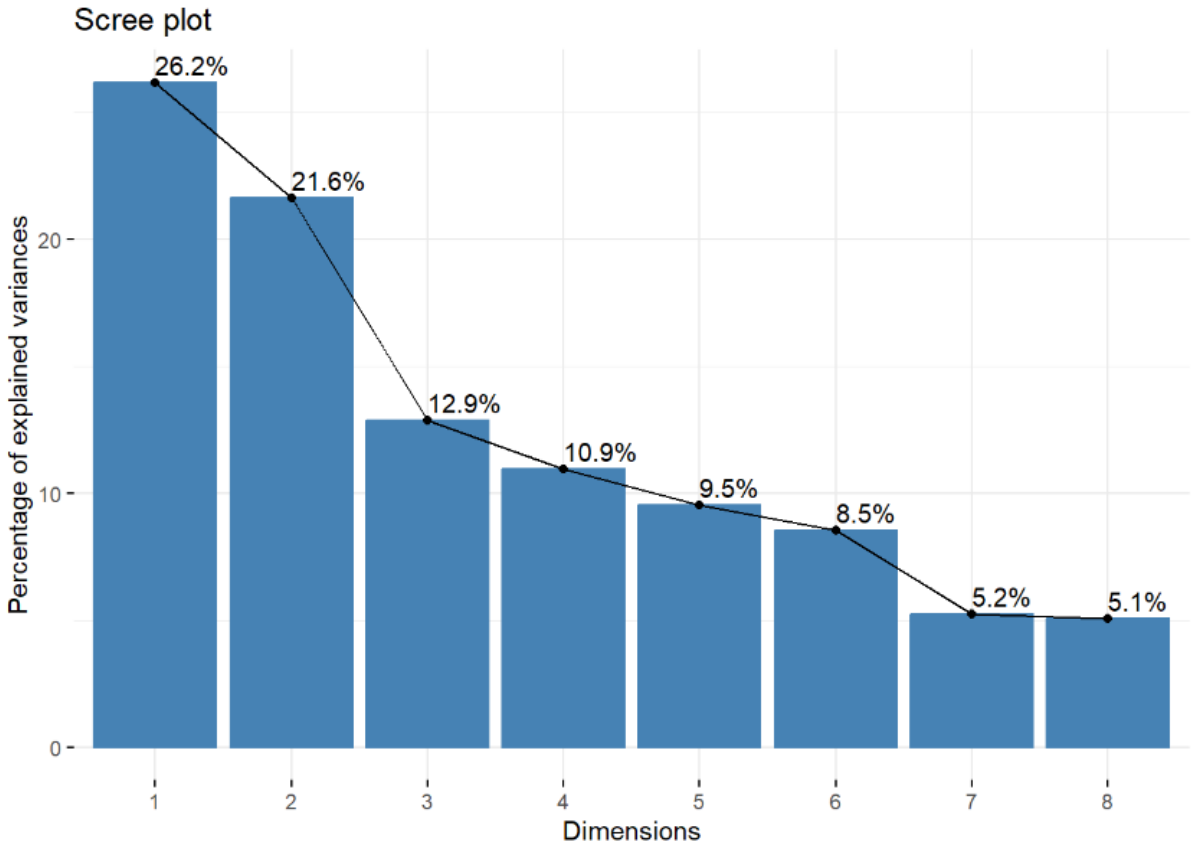
- Principal Component Analysis (PCA):

```
## Standard deviations (1, .., p=8):
## [1] 1.4471973 1.3157546 1.0147068 0.9356971 0.8731234 0.8262133 0.6479322
## [8] 0.6359733
##
## Rotation (n x k) = (8 x 8):
##                                PC1        PC2         PC3         PC4
## Pregnancies              -0.1284321  0.5937858 -0.01308692  0.08069115
## Glucose                  -0.3930826  0.1740291  0.46792282 -0.40432871
## BloodPressure            -0.3600026  0.1838921 -0.53549442  0.05598649
## SkinThickness            -0.4398243 -0.3319653 -0.23767380  0.03797608
## Insulin                  -0.4350262 -0.2507811  0.33670893 -0.34994376
## BMI                      -0.4519413 -0.1009598 -0.36186463  0.05364595
## DiabetesPedigreeFunction -0.2706114 -0.1220690  0.43318905  0.83368010
## Age                      -0.1980271  0.6205885  0.07524755  0.07120060
##                                PC5          PC6         PC7          PC8
## Pregnancies              -0.4756057  0.193598168  0.58879003  0.117840984
## Glucose                   0.4663280  0.094161756  0.06015291  0.450355256
## BloodPressure             0.3279531 -0.634115895  0.19211793 -0.011295538
## SkinThickness            -0.4878621  0.009589438 -0.28221253  0.566283799
## Insulin                  -0.3469348 -0.270650609  0.13200992 -0.548621381
## BMI                       0.2532038  0.685372179  0.03536644 -0.341517637
## DiabetesPedigreeFunction  0.1198105 -0.085784088  0.08609107 -0.008258731
## Age                      -0.1092900 -0.033357170 -0.71208542 -0.211661979
```

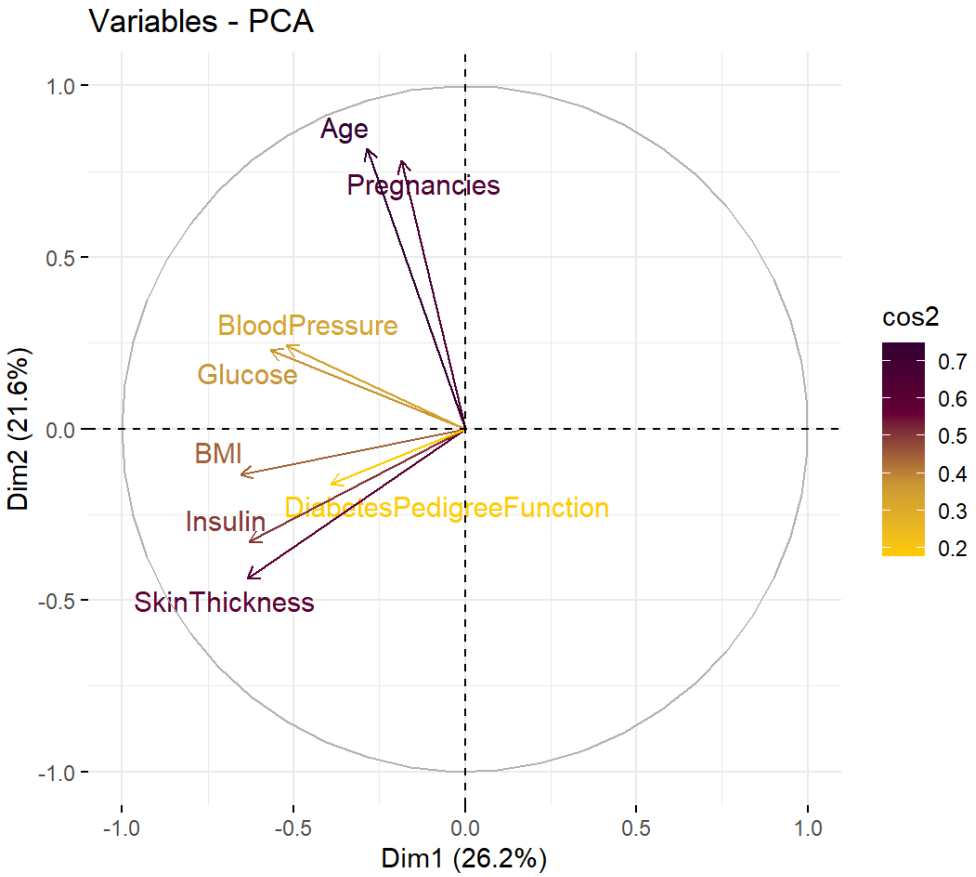
```
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     1.4472 1.3158 1.0147 0.9357 0.87312 0.82621 0.64793
## Proportion of Variance 0.2618 0.2164 0.1287 0.1094 0.09529 0.08533 0.05248
## Cumulative Proportion  0.2618 0.4782 0.6069 0.7163 0.81164 0.89697 0.94944
##                           PC8
## Standard deviation     0.63597
## Proportion of Variance 0.05056
## Cumulative Proportion  1.00000
```

- We got an 8x8 rotation matrix.
- These PCs are ordered in the order of importance, with PC1 being most important and so on.
- The numbers represented as a list shows the loadings/weights of each original variable.
- The larger the value of the loading, the more important the variable is in determining the value of that principal component.

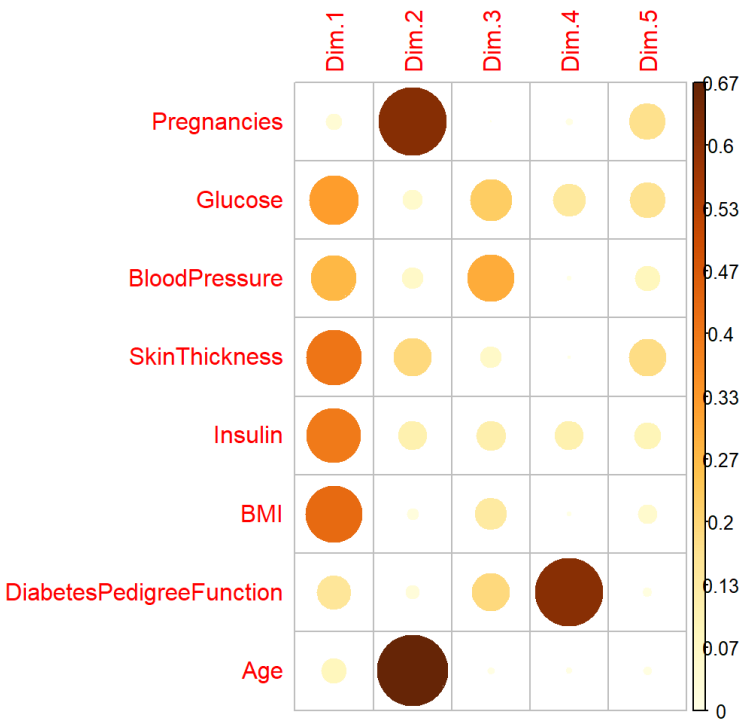
- For example, for PC2, the variables with the higher loading are Pregnancies (0.5937) and Age (0.6205).
- It indicates that these variables have large contribution to PC2.
- Note: Negative sign indicates the inverse relationship or correlation of that variable to the corresponding principal component.
- Standard Deviation indicates the amount of variability or information PC captures from the original variable.
- The proportion of Variance explains the variance explained by each PC.
- Cumulative Proportion is just the combination of the current PC and the PCs before it.



- PC1 explains around 26.2% of variance.
- PC2 explains around 21.6% of variance.
- PC3 explains around 12.9% of variance.
- PC4 explains around 10.9% of variance.
- PC5 explains around 9.5% of variance.
- PC6 explains around 8.5% of variance.
- PC7 explains around 5.2% of variance.
- PC8 explains around 5.1% of variance.



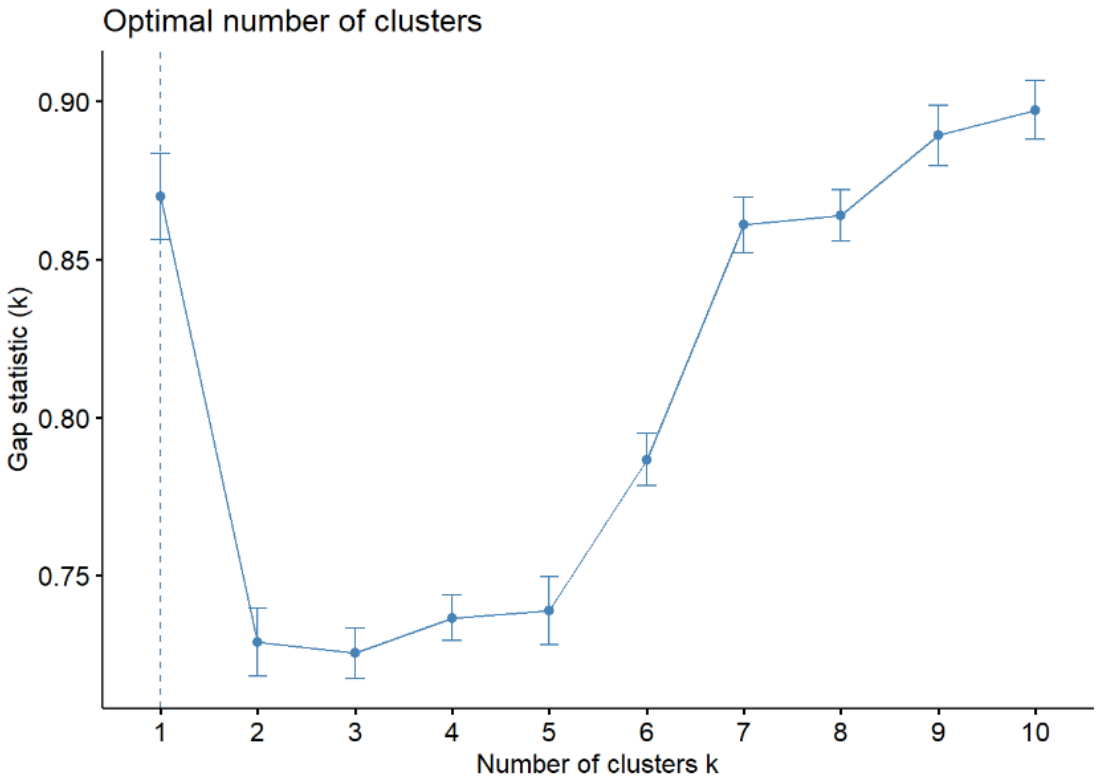
- We can see that Age and Pregnancies are close to each other and can infer that both have similar patterns and are correlated as well.
- Similarly, Blood Pressure and Glucose follow the same trend.
- Also, we can see that Age and Skin Thickness are the variables that are closer to circumference of the circle compared to other variables, which indicates that those variables have higher contributions to the explained variation in the data.
- Similarly, the variables that are closer to circumference like Blood Pressure or BMI are the ones that have lower contributions to the explained variance.



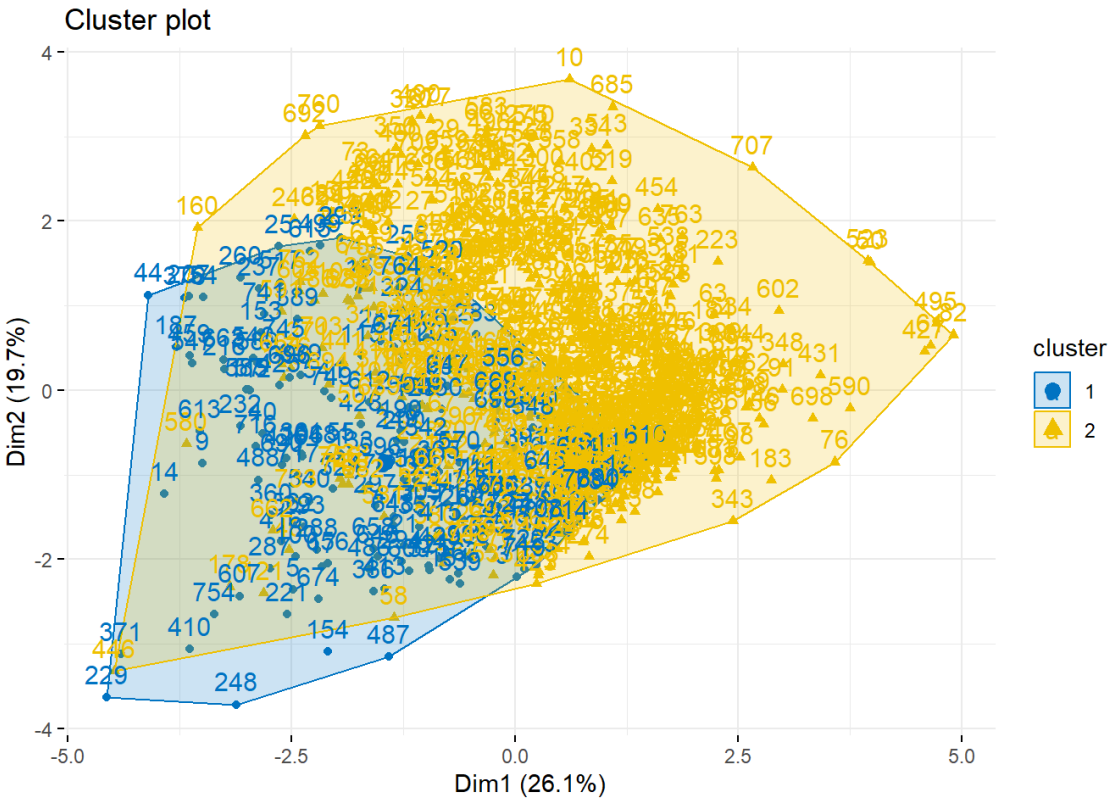
- We can see that Skin Thickness, BMI, and Insulin have much correlation with Dim. 1 (PC 1).
- Similarly, Pregnancies, and Age are more correlated with Dim. 2 (PC 2).

Based on our Principal Component Analysis (PCA) results, we are able to explain almost 61% of the variance from the first **3 principal components**. This was concluded by scree plot (from elbow on the graph), and also with T-tests, F-tests, and various other plots. We found out that how much variance is being explained by each of the components and the **first 3 components** have been selected based on it. Other interpretations are attached below the respected codes in R file/Knitted file.

➢ Clustering Analysis



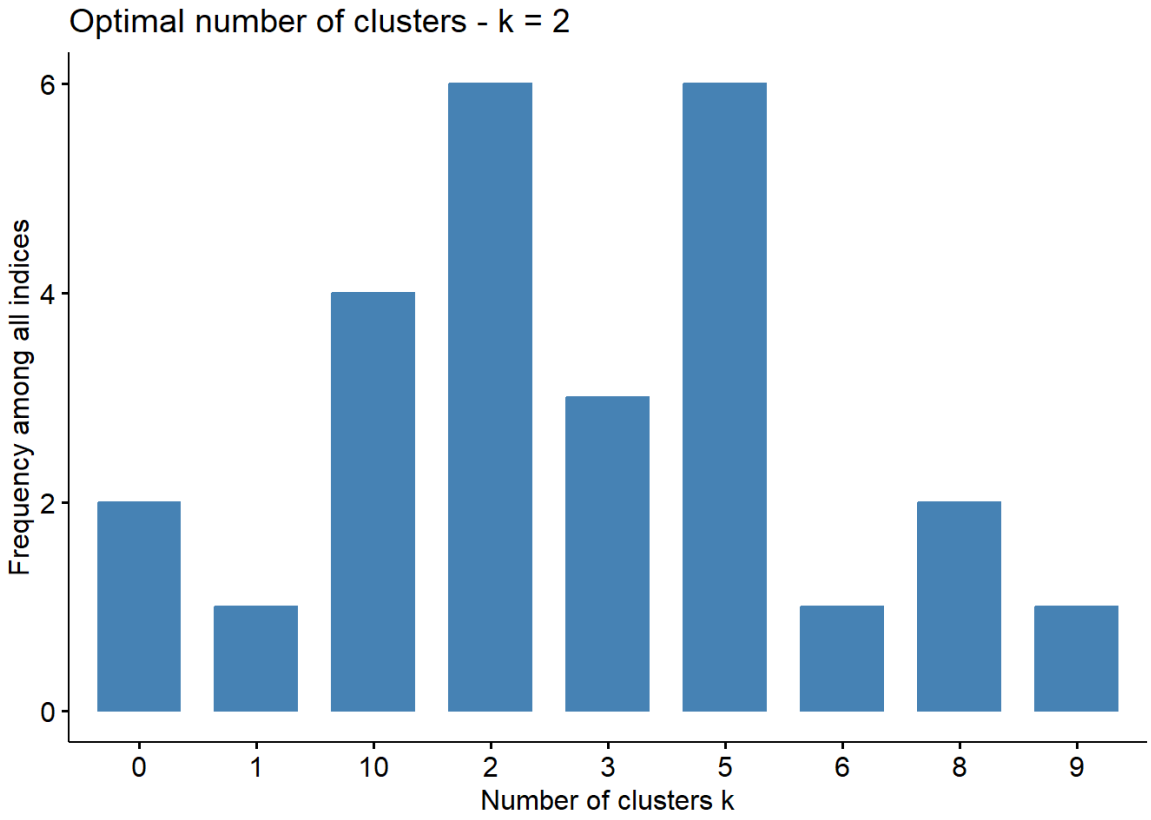
- From the graph we can see that, if number of clusters is 1, it has the gap statistics greater than 0.85.
- But for 2,3,4, and 5, the gap statistics is almost below 0.75.
- We basically prefer the number of clusters that has higher gap statistics because it indicates a better clustering structure.



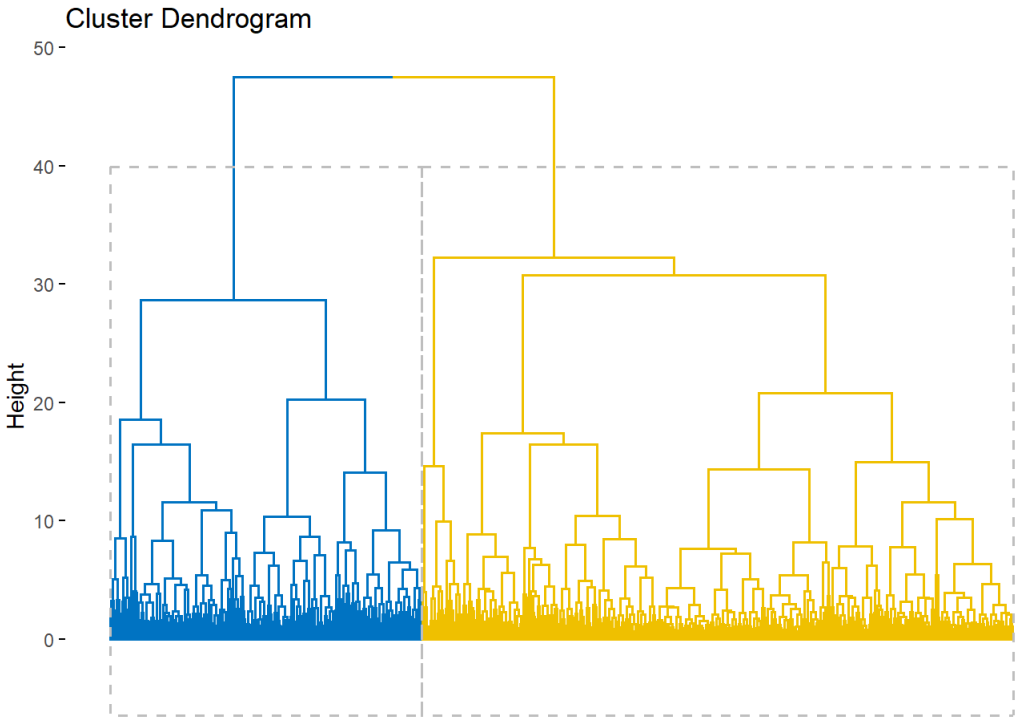
- We can observe that we are not able to distinguish the clusters easily when number of clusters are 2.
- We can see that yellow cluster has overlapped blue cluster, but noticeably, we can see that yellow cluster has half of the formed cluster which is not overlapped.

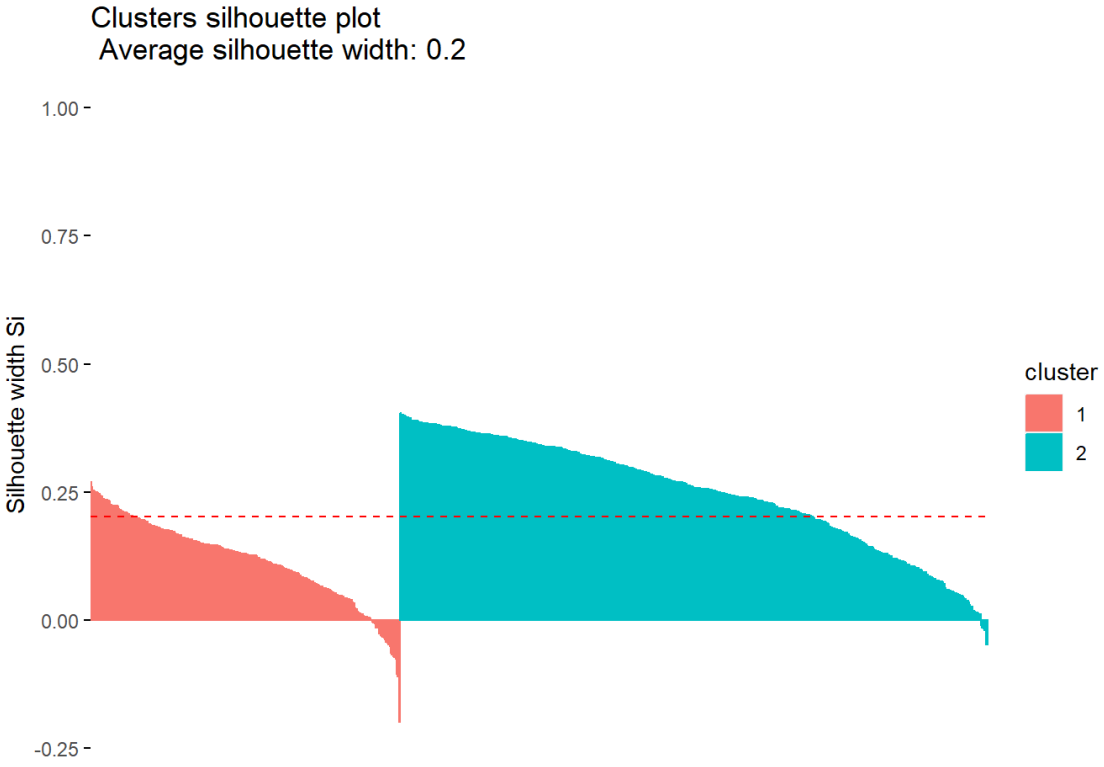
```
## *** : The D index is a graphical method of determining the number of clusters.
##                 In the plot of D index, we seek a significant knee (the significant peak in Dindex
##                 second differences plot) that corresponds to a significant increase of the value of
##                 the measure.
##
## *******************************************************************
## * Among all indices:
## * 6 proposed 2 as the best number of clusters
## * 3 proposed 3 as the best number of clusters
## * 6 proposed 5 as the best number of clusters
## * 1 proposed 6 as the best number of clusters
## * 2 proposed 8 as the best number of clusters
## * 1 proposed 9 as the best number of clusters
## * 4 proposed 10 as the best number of clusters
##
##                   ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is  2
##
##
## *******************************************************************
```

- We are using Hubert Index and D-Index method, which are graphical methods to determine the number of clusters.
- From the majority rule, we can say that the best number of clusters is 2 for our dataset.



- We can see that, number of clusters = 2 has the highest frequency among all indices.
- Therefore, we can say that 2 is the optimal number of clusters for our dataset.





- We are using this method to measure the quality of clustering in dataset.

- Its value ranges from -1 to 1, where a higher value indicates better clustering.
- Of the 768 values, 265 belong to cluster 1 and 503 belongs to cluster 2.
- Since we observed an overlap in our cluster, we are observing few negative silhouette values for cluster 1 and 2.

Based on our Clustering Analysis result, we were not exactly able to distinguish between the two group of people, that is diabetic and non-diabetic, because the clusters were getting overlapped. We also tried to measure the quality of clusters using Silhouette value, and the results are attached in the R files/Knitted file. Overall, we found out that **glucose and insulin** were quite correlated with the person being diabetic. We also found that the optimal number of clusters for our dataset would be **2**.

➢ Exploratory Factor Analysis

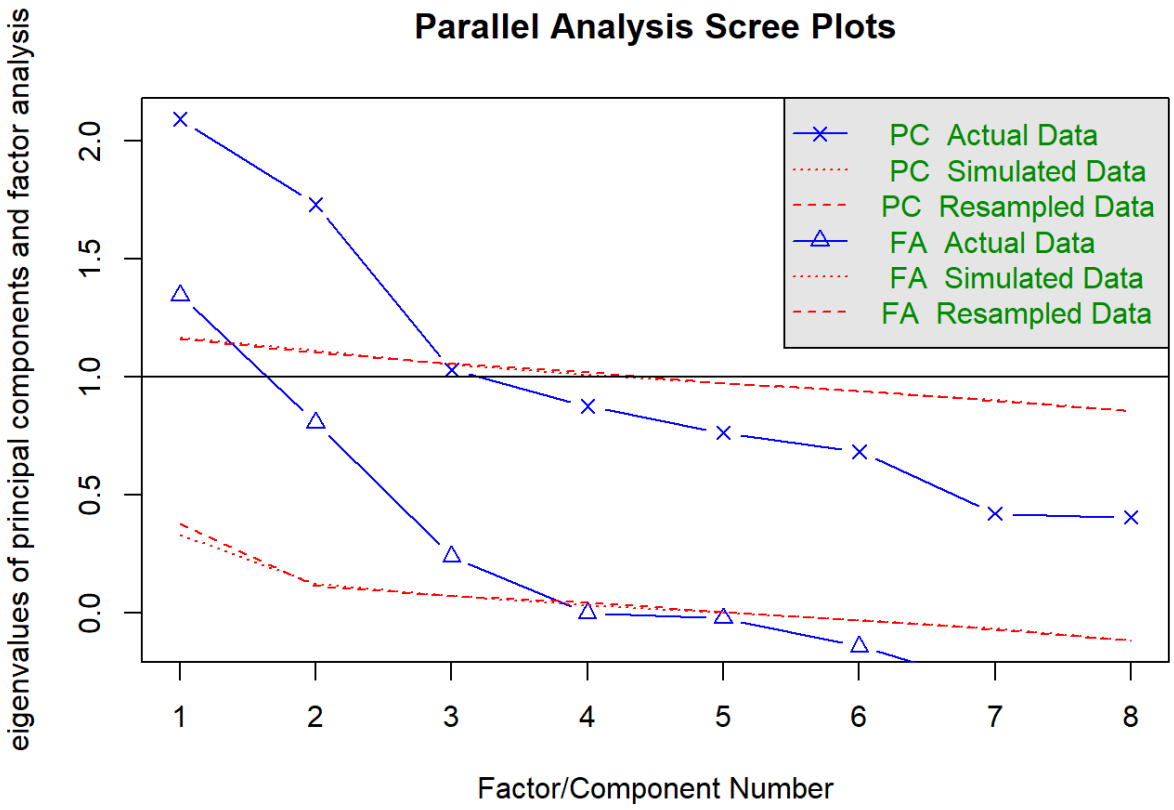
```
## Principal Components Analysis
## Call: principal(r = diabetes[-9], nfactors = 4, rotate = "varimax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                            RC2  RC1   RC3   RC4   h2    u2 com
## Pregnancies               0.81 0.02 -0.01 -0.02 0.65 0.349 1.0
## Glucose                   0.32 0.03  0.80  0.02 0.74 0.255 1.3
## BloodPressure             0.35 0.70 -0.07 -0.10 0.63 0.372 1.5
## SkinThickness            -0.28 0.70  0.25  0.17 0.66 0.345 1.8
## Insulin                  -0.20 0.25  0.78  0.11 0.73 0.271 1.4
## BMI                       0.02 0.74  0.15  0.08 0.58 0.417 1.1
## DiabetesPedigreeFunction  0.02 0.08  0.09  0.98 0.98 0.019 1.0
## Age                       0.86 0.03  0.10  0.04 0.76 0.241 1.0
##
##                        RC2  RC1  RC3  RC4
## SS loadings           1.74 1.60 1.36 1.02
## Proportion Var        0.22 0.20 0.17 0.13
## Cumulative Var        0.22 0.42 0.59 0.72
## Proportion Explained  0.30 0.28 0.24 0.18
## Cumulative Proportion 0.30 0.58 0.82 1.00
##
## Mean item complexity =  1.3
## Test of the hypothesis that 4 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.11
##  with the empirical chi square  489.1  with prob <  6.2e-107
##
## Fit based upon off diagonal values = 0.75
```

- We passed nfactors = 4 which refers to the number of factors or PCs.
- RC2, RC1, RC3, and RC4 have some values which represents the loadings of each variables.
- Positive loading indicates that there is a positive correlation between the original variable and extracted Principal Components.
- While negative loading indicates a negative relationship.
- h2 shows the proportion of variance in each variable explained by the principal components.
- u2 shows the proportion of variance in each variable unexplained by the principal components.
- com shows the complexity to explain the variance.
- SS Loadings shows the squared loading of each principal components.
- Proportion Variance shows the proportion of total variance explained by each principal component.

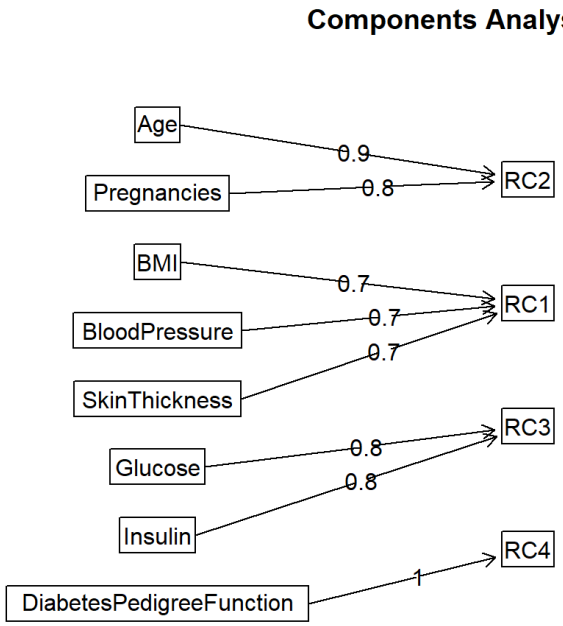
- Cumulative Variance shows the cumulative proportion of total variance explained by each principal component.
- Proportion Explained shows the proportion of total variance explained by each principal component, expressed as a percentage.
- Cumulative Proportion shows the cumulative proportion of total variance explained by each principal component, expressed as a percentage.

```
##
## Loadings:
##                           RC2    RC1    RC3    RC4
## Pregnancies                0.806
## Glucose                    0.323         0.800
## BloodPressure              0.345  0.704
## SkinThickness             -0.281  0.695  0.254  0.169
## Insulin                   -0.203  0.254  0.782  0.110
## BMI                               0.744  0.153
## DiabetesPedigreeFunction                        0.983
## Age                        0.864         0.101
##
##                  RC2   RC1   RC3   RC4
## SS loadings    1.741 1.605 1.362 1.024
## Proportion Var 0.218 0.201 0.170 0.128
## Cumulative Var 0.218 0.418 0.588 0.716
```

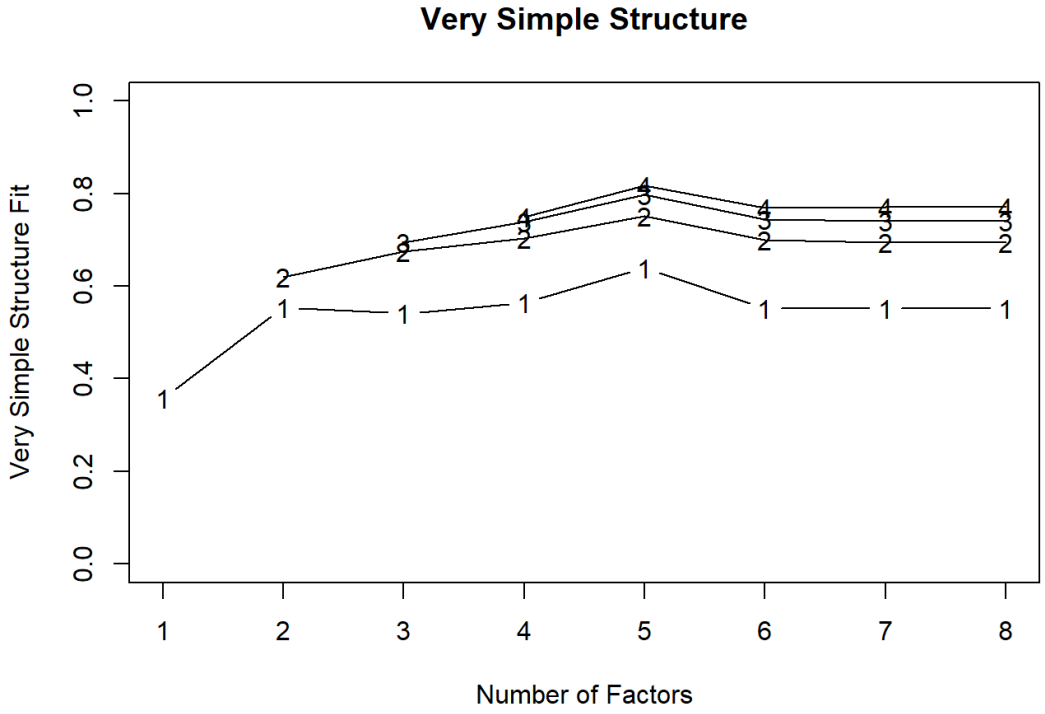
- From the result we can make following conclusions:
- RC1 shows the relation with Blood Pressure, Skin Thickness, and BMI.
- RC2 shows the relation with Pregnancies, and Age.
- RC3 shows the relation with Glucose, and Insulin.
- RC4 shows the relation with only Diabetes Pedigree Function.
- All four RCs explain around 72% (71.6% precise) of the variance.



- Parallel analysis suggests that the number of factors = 3 and the number of components = 2.



- We can say that we reduced the factors from 8 to 4.
- It can be seen from the Components Analysis Plot.



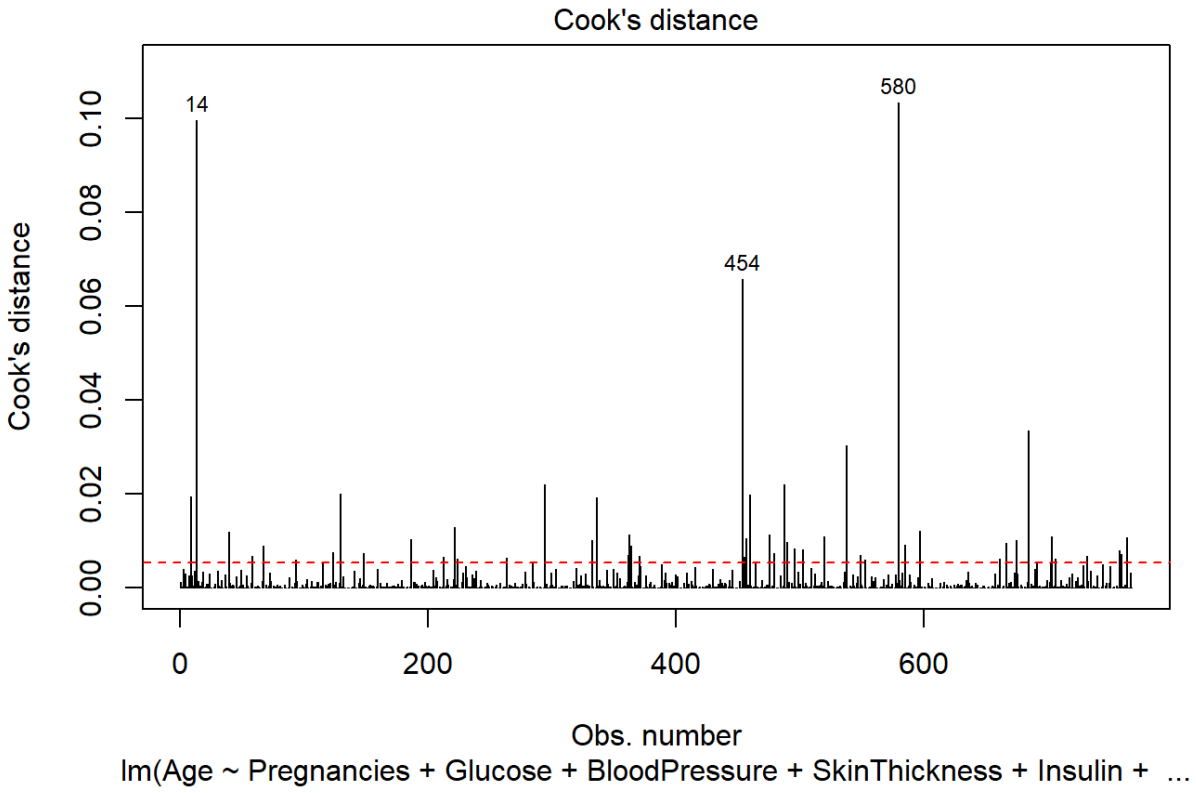
- We can also see from Very Simple Structure that after factor number 4, we can't see any increase in the explained variance.
- Therefore, we can conclude that number of factors = 4 is a good fit.

From our Exploratory Factor Analysis (EFA) result, we concluded that the number of factors we **can reduce to is 4**. The evidence has been provided by using Parallel Analysis Plot, and Very Simple Structure (VSS) Plot, which has been attached in the R file along with the interpretations. We were able to explain around **72% (approx.)** of the variance from our 4 generated factors. Therefore, we reduced the number of features from **8 to 4.**

➢ Multiple Regression

```
##
## Call:
## lm(formula = Age ~ Pregnancies + Glucose + BloodPressure + SkinThickness +
##     Insulin + BMI + DiabetesPedigreeFunction + Outcome, data = diabetes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.565  -5.622  -2.146   3.348  46.781
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              14.250793   2.063405   6.906 1.05e-11 ***
## Pregnancies               1.658286   0.105288  15.750  < 2e-16 ***
## Glucose                   0.062631   0.012858   4.871 1.35e-06 ***
## BloodPressure             0.106469   0.018695   5.695 1.76e-08 ***
## SkinThickness            -0.074257   0.025880  -2.869  0.00423 **
## Insulin                  -0.004448   0.003505  -1.269  0.20479
## BMI                      -0.050250   0.050087  -1.003  0.31606
## DiabetesPedigreeFunction  1.422907   1.060137   1.342  0.17994
## Outcome                   1.434674   0.847562   1.693  0.09092 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.363 on 759 degrees of freedom
## Multiple R-squared:  0.3728, Adjusted R-squared:  0.3662
## F-statistic: 56.39 on 8 and 759 DF,  p-value: < 2.2e-16
```

- The estimated coefficient for "Pregnancies" is 1.66541, which means that for every one-unit increase in the number of pregnancies, the predicted age increases by approximately 1.67 units.
- The multiple R-squared is 0.3695, which means that the model explains about 37% of the variability in the age of individuals.
- P-value (p-value: < 2.2e-16) suggests that the model is statistically significant, indicating that at least one of the predictor variables has a significant effect on the predicted age.
- The estimated residual standard error is 9.369, which indicates the average difference between the observed and predicted age values is approximately 9.37 units.



- We are basically looking for influential data points.
- The horizontal red line shows the cutoff.
- The points above the line indicate that those are influential data points and have negative impact on the regression model.

```r
#Predicting the Age based on model_1 for some new data
new_data <- data.frame(
  Pregnancies = 3,
  Glucose = 155,
  BloodPressure = 72,
  SkinThickness = 35,
  Insulin = 10,
  BMI = 33,
  DiabetesPedigreeFunction = 0.55,
  Outcome = 1
)

# Replacing the values with the values of our test data
predicted_age <- predict(model_1, newdata = new_data)
print(predicted_age)
```

```
##        1
## 34.51478
```

- By passing some random values (test data) on our model, the model predicted the age to be 34.5 years.

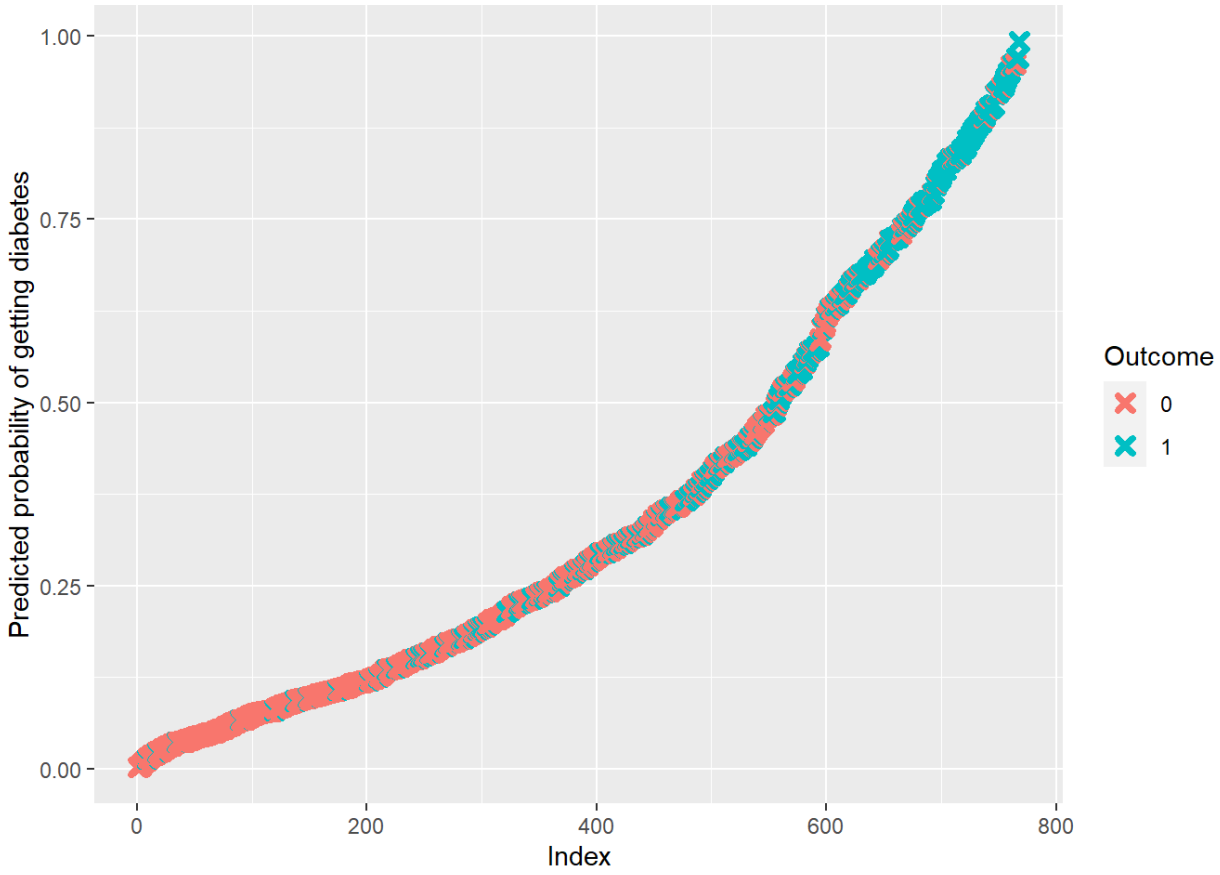
From our Multiple Regression result, we made a model that was able to **predict the age** of the person based on other features like pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, Diabetes Pedigree Function, and Outcome. The interpretation of the results is attached along the codes in the R file/Knitted file. We also predicted a person's age by passing new random data, and model was able to predict the age, and the results of it are also attached with the code.

- Logistic Regression

```
##
## Call:
## glm(formula = Outcome ~ ., family = "binomial", data = diabetes)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5566  -0.7274  -0.4159   0.7267   2.9297
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -8.4046964  0.7166359 -11.728  < 2e-16 ***
## Pregnancies               0.1231823  0.0320776   3.840 0.000123 ***
## Glucose                   0.0351637  0.0037087   9.481  < 2e-16 ***
## BloodPressure            -0.0132955  0.0052336  -2.540 0.011072 *
## SkinThickness             0.0006190  0.0068994   0.090 0.928515
## Insulin                  -0.0011917  0.0009012  -1.322 0.186065
## BMI                       0.0897010  0.0150876   5.945 2.76e-09 ***
## DiabetesPedigreeFunction  0.9451797  0.2991475   3.160 0.001580 **
## Age                       0.0148690  0.0093348   1.593 0.111192
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 993.48  on 767  degrees of freedom
## Residual deviance: 723.45  on 759  degrees of freedom
## AIC: 741.45
##
## Number of Fisher Scoring iterations: 5
```

- The intercept is the log(odds) a person will be diabetic.
- Each one unit change in the number of pregnancies will increase the log odds of getting diabetes by 0.123.
- Each one unit change in the glucose level will increase the log odds of getting diabetes by 0.0351.
- Even the p-value states that, both these attributes are quite significant in determining the diabetes.
- The interpretation of Blood Pressure and Insulin is different.
- Getting Blood Pressure and Insulin changed by one unit will decrease the log odds of getting diabetes by -0.0132 and -0.00119 respectively.

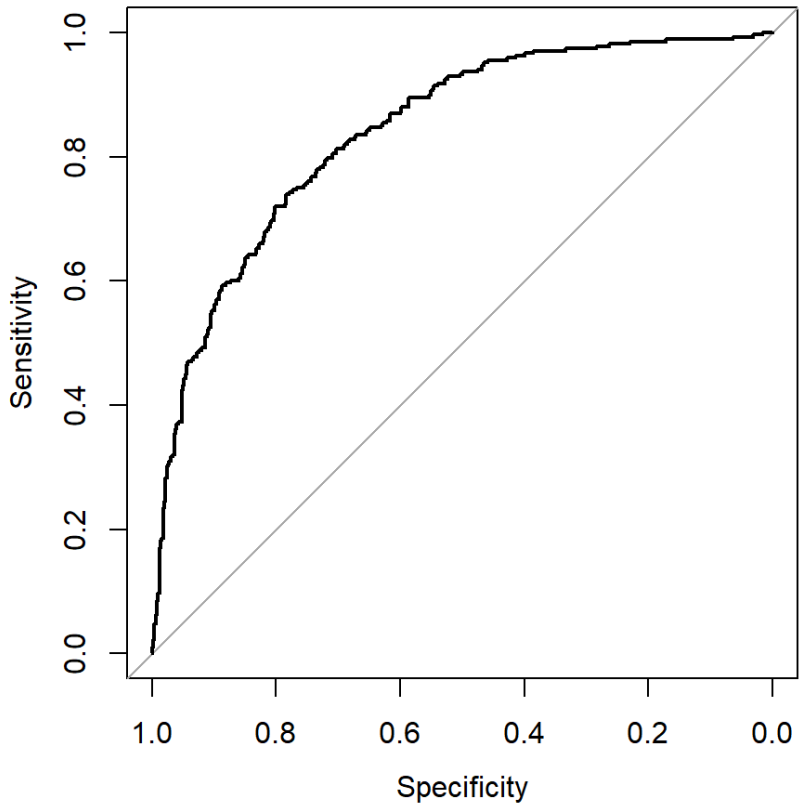
- Difference between Null deviance and Residual deviance tells us that the model is a good fit.
- Greater the difference means better the model.



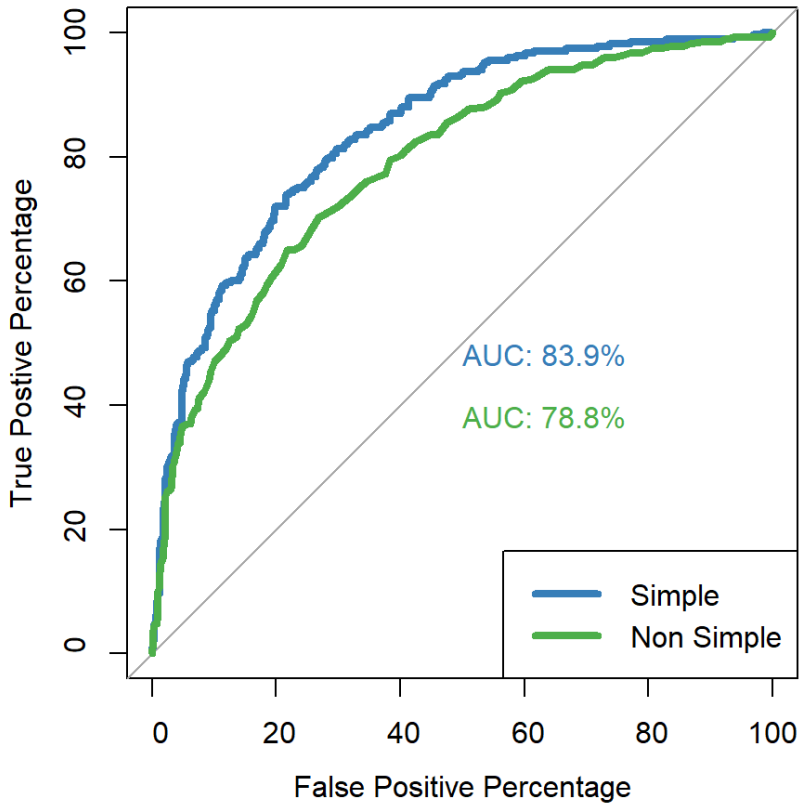
- We know that as probability of diabetes increases, a person has greater chance of having diabetes and it is showed with blue color.
- Similarly, a person having less probability of diabetes has less chance of having it.
- In our case, 0.50 (approximately from the graph), can be identified as the cut-off or threshold of having diabetes and not having diabetes.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 445 112
##          1  55 156
##
##                Accuracy : 0.7826
##                  95% CI : (0.7517, 0.8112)
##     No Information Rate : 0.651
##     P-Value [Acc > NIR] : 1.373e-15
##
##                   Kappa : 0.4966
##
##  Mcnemar's Test P-Value : 1.468e-05
##
##             Sensitivity : 0.8900
##             Specificity : 0.5821
##          Pos Pred Value : 0.7989
##          Neg Pred Value : 0.7393
##              Prevalence : 0.6510
##          Detection Rate : 0.5794
##    Detection Prevalence : 0.7253
##       Balanced Accuracy : 0.7360
##
##        'Positive' Class : 0
##
```

- 445 + 156 = 601 were predicted as 0 or 1 correctly (diagonal)
- 55 + 112 = 167 were predicted incorrectly (off-diagonal).
- The accuracy is reported as 0.7826, which means that the model correctly predicted around 78.26% of the instances.
- The NIR is reported as 0.651, which is the accuracy that would be achieved by always predicting class 0.
- A higher kappa value indicates a better agreement between the model's predictions and the actual values.
- In our case, the kappa is reported as 0.4966, which is not as much good as expected.
- Sensitivity is reported as 0.8900, indicating that the model correctly predicted around 89.00% of the instances of class 1.
- Specificity is reported as 0.5821, indicating that the model correctly predicted around 58.21% of the instances of class 0.
- Balanced Accuracy is the average of Sensitivity and Specificity.
- We got balanced accuracy as 0.7360.



- ROC curve is a graphical plot which shows the trade-off between the true positive rate (Sensitivity) and the false positive rate (Specificity).
- Higher value indicates better overall model performance.
- The AUC is reported as 0.8394, indicating that the model has good discriminatory power with an AUC value close to 1, suggesting that it can effectively distinguish between the two classes (0 and 1) based on the predicted probabilities from the logistic regression model.



- The AUC in terms of percentage for logistic (with Glucose) is 78.8% (indicated by green color).

```r
#Predicting the Outcome based on our logistic model for some new data
new_data <- data.frame(
  Pregnancies = 0,
  Glucose = 135,
  BloodPressure = 92,
  SkinThickness = 25,
  Insulin = 0,
  BMI = 45,
  DiabetesPedigreeFunction = 0.67,
  Age = 40
)

# Replacing the values with the values of our test data
predicted_outcome <- predict(logistic_simple, newdata = new_data)
print(predicted_outcome)
```

```
##         1
## 0.3992632
```

- By passing some random values (test data) on our model, the model predicted the outcome probability to be 0.399.
- This means that the person is less likely or does not have diabetes based on our model's output.

From our Logistic Regression result, we were able to **predict the person's outcome**, whether he/she is diabetic or not by using other features from the dataset. We got an accuracy of **78.26%**, that means the model correctly predicted **0.7826** of the instances correctly. We also plotted the AUC curve to evident our results and predicted the outcome of a new random data of the person, and the results are attached on the R file.

Supporting Evidence:

All the supporting evidence and interpretation are attached to the code. I had attached the GitHub link where the all the methodology, code, rmd file, and data dictionary is attached as well.

GitHub Link:

https://github.com/Prince0511/Multivariate-Analysis/tree/main/Diabetes%20Analysis%20Project