MINI PROJECT – HANDS-ON EXPEREINCE

Principal Component Analysis of Clinical Risk Factors for Diabetes

Objective: The purpose of this project is to apply Principal Component Analysis (PCA) to uncover the major underlying patterns within key clinical health measurements and examine how these patterns differ between individuals with and without diabetes. By reducing several correlated health variables such as glucose, BMI, blood pressure, insulin, and into a smaller set of meaningful components, PCA allows us to visualize the structure of the data more clearly. This analysis also helps determine whether diabetic and non-diabetic individuals display distinct trends along the principal components, offering insights into how certain combinations of health factors are associated with diabetes risk.

Data Description: The dataset used in this project originates from the National Institute of Diabetes and Digestive and Kidney Diseases and is widely known as the Pima Indians Diabetes Dataset. The primary aim of the original study was to determine whether diabetes could be predicted based on a set of clinical and physiological measurements collected from patients.

To ensure consistency in the population being studied, several restrictions were applied during data selection. Specifically, all individuals in the dataset are females, at least 21 years old, and of Pima Indian heritage, a population with a historically high prevalence of diabetes. These constraints reduce demographic variability and allow the analysis to focus more directly on the clinical predictors of diabetes.

The dataset consists of eight medical predictor variables and one binary outcome variable.
The predictors include:

- Number of pregnancies
- Plasma glucose concentration
- Diastolic blood pressure
- Triceps skinfold thickness
- 2-hour serum insulin
- Body Mass Index (BMI)
- Diabetes Pedigree Function (an estimate of genetic influence)
- Age

The target variable, Outcome, indicates whether the individual has diabetes (1) or does not (0). These predictors represent clinically meaningful measurements commonly associated with diabetes risk. In this project, they form the basis for applying Principal Component Analysis (PCA) to uncover underlying health patterns and to explore whether PCA can visually differentiate diabetic from non-diabetic individuals.

DESCRIPTIVE STATISTICS

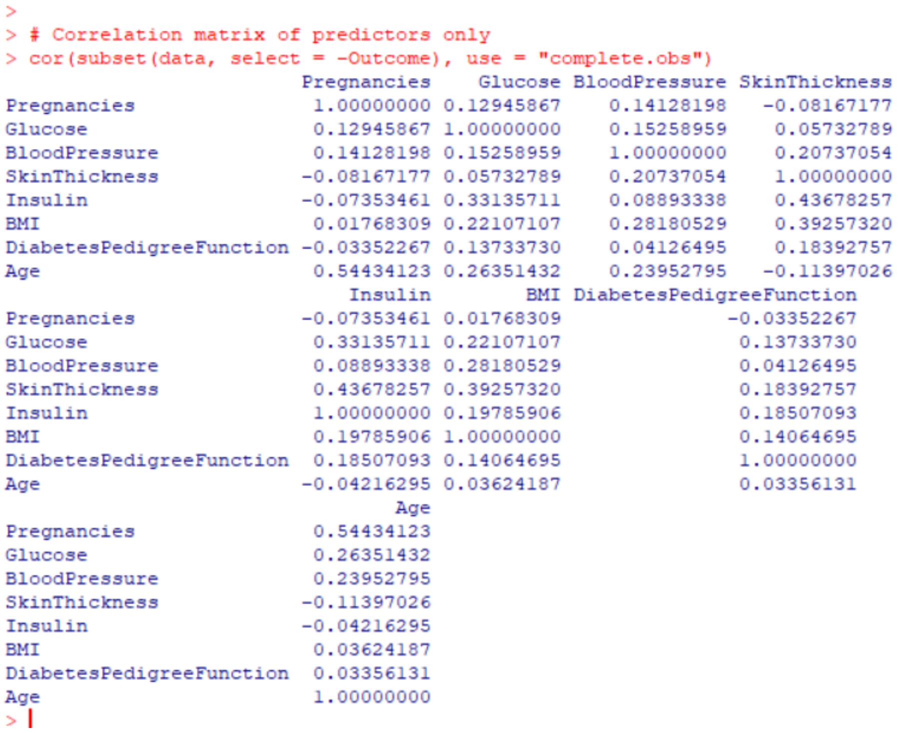
Age & Pregnancies (0.54): Older women tend to have had more pregnancies.

BMI & Skin Thickness (0.39): Thicker skinfold measurements are associated with higher BMI.

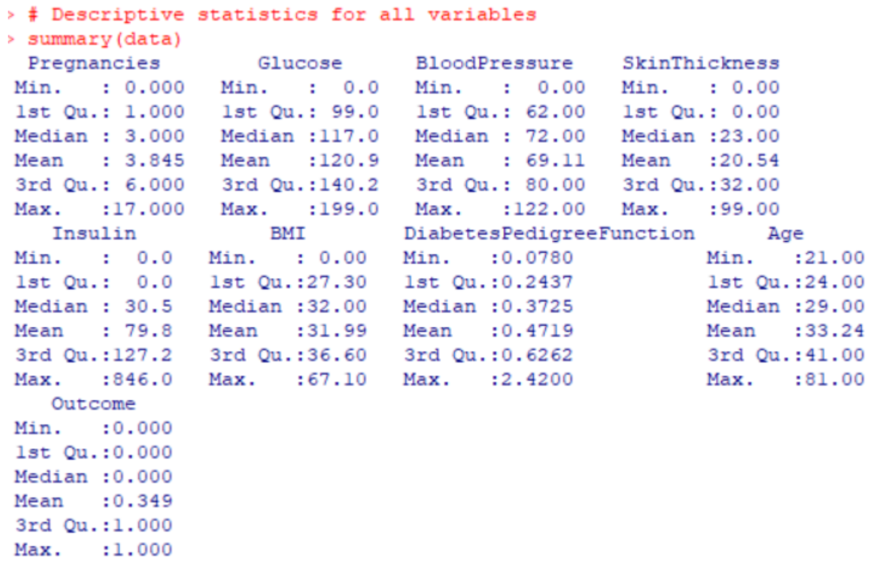
Insulin & Skin Thickness (0.44): Higher skinfold thickness correlates with higher insulin levels.

Glucose & Insulin (0.33): Higher glucose levels tend to accompany higher insulin levels.

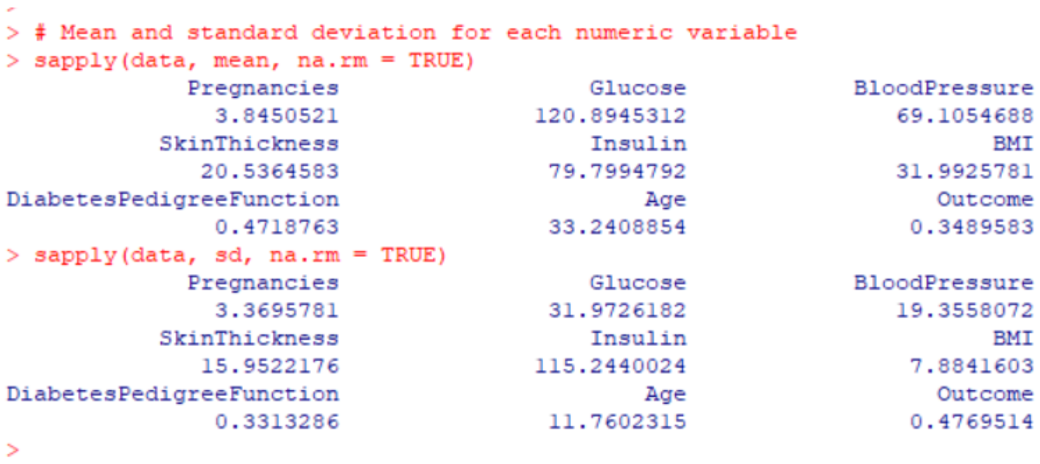
These relationships are medically reasonable and show that variables are not independent which is why PCA is useful.



```
>
> # Correlation matrix of predictors only
> cor(subset(data, select = -Outcome), use = "complete.obs")
                         Pregnancies    Glucose BloodPressure SkinThickness
Pregnancies               1.00000000 0.12945867    0.14128198   -0.08167177
Glucose                   0.12945867 1.00000000    0.15258959    0.05732789
BloodPressure             0.14128198 0.15258959    1.00000000    0.20737054
SkinThickness            -0.08167177 0.05732789    0.20737054    1.00000000
Insulin                  -0.07353461 0.33135711    0.08893338    0.43678257
BMI                       0.01768309 0.22107107    0.28180529    0.39257320
DiabetesPedigreeFunction -0.03352267 0.13733730    0.04126495    0.18392757
Age                       0.54434123 0.26351432    0.23952795   -0.11397026
                             Insulin        BMI DiabetesPedigreeFunction
Pregnancies              -0.07353461 0.01768309              -0.03352267
Glucose                   0.33135711 0.22107107               0.13733730
BloodPressure             0.08893338 0.28180529               0.04126495
SkinThickness             0.43678257 0.39257320               0.18392757
Insulin                   1.00000000 0.19785906               0.18507093
BMI                       0.19785906 1.00000000               0.14064695
DiabetesPedigreeFunction  0.18507093 0.14064695               1.00000000
Age                      -0.04216295 0.03624187               0.03356131
                                 Age
Pregnancies               0.54434123
Glucose                   0.26351432
BloodPressure             0.23952795
SkinThickness            -0.11397026
Insulin                  -0.04216295
BMI                       0.03624187
DiabetesPedigreeFunction  0.03356131
Age                       1.00000000
> 
```



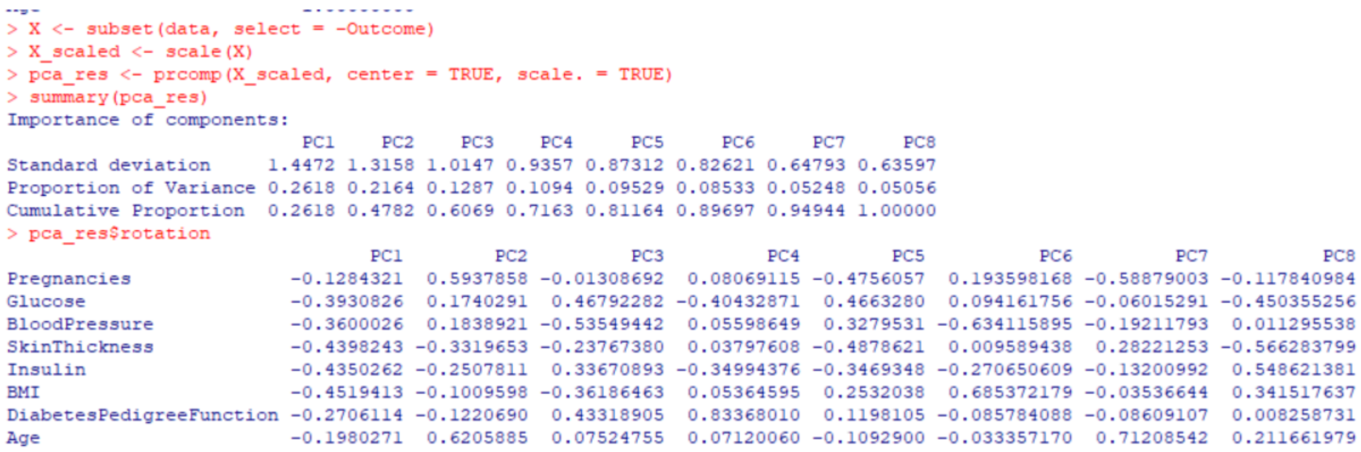
```
> # Descriptive statistics for all variables
> summary(data)
  Pregnancies        Glucose       BloodPressure     SkinThickness
 Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00
 1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00
 Median : 3.000   Median :117.0   Median : 72.00   Median :23.00
 Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54
 3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00
 Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
    Insulin          BMI        DiabetesPedigreeFunction      Age
 Min.   :  0.0   Min.   : 0.00   Min.   :0.0780           Min.   :21.00
 1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437           1st Qu.:24.00
 Median : 30.5   Median :32.00   Median :0.3725           Median :29.00
 Mean   : 79.8   Mean   :31.99   Mean   :0.4719           Mean   :33.24
 3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262           3rd Qu.:41.00
 Max.   :846.0   Max.   :67.10   Max.   :2.4200           Max.   :81.00
    Outcome
 Min.   :0.000
 1st Qu.:0.000
 Median :0.000
 Mean   :0.349
 3rd Qu.:1.000
 Max.   :1.000
```



```
> # Mean and standard deviation for each numeric variable
> sapply(data, mean, na.rm = TRUE)
             Pregnancies                  Glucose            BloodPressure
               3.8450521              120.8945312               69.1054688
           SkinThickness                  Insulin                      BMI
              20.5364583               79.7994792               31.9925781
DiabetesPedigreeFunction                      Age                  Outcome
               0.4718763               33.2408854                0.3489583
> sapply(data, sd, na.rm = TRUE)
             Pregnancies                  Glucose            BloodPressure
               3.3695781               31.9726182               19.3558072
           SkinThickness                  Insulin                      BMI
              15.9522176              115.2440024                7.8841603
DiabetesPedigreeFunction                      Age                  Outcome
               0.3313286               11.7602315                0.4769514
>
```

PRINCIPAL COMPONENT ANALYSIS RESULTS

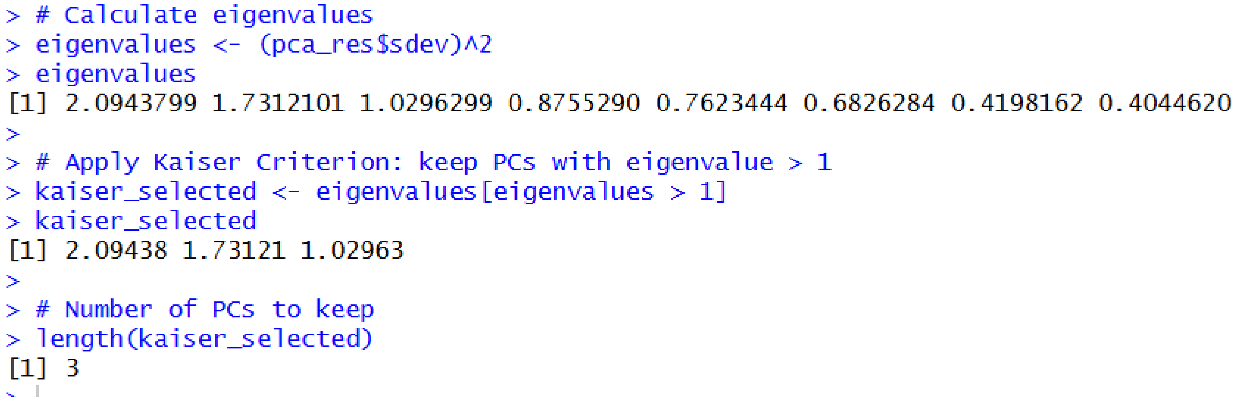
PC1 (Metabolic Health Component):
The first principal component had strong negative loadings from BMI (–0.4519), skin thickness (–0.4398), insulin (–0.4350), glucose (–0.3931), and blood pressure (–0.3600). These variables are all well-established indicators of metabolic status and diabetes risk. Because PC1 is dominated by metabolic and obesity-related measurements, this component represents overall metabolic health. Individuals with higher PC1 scores tend to have higher glucose levels, greater insulin resistance, elevated BMI, thicker skinfold measurements, and higher blood pressure, patterns associated with increased likelihood of diabetes.



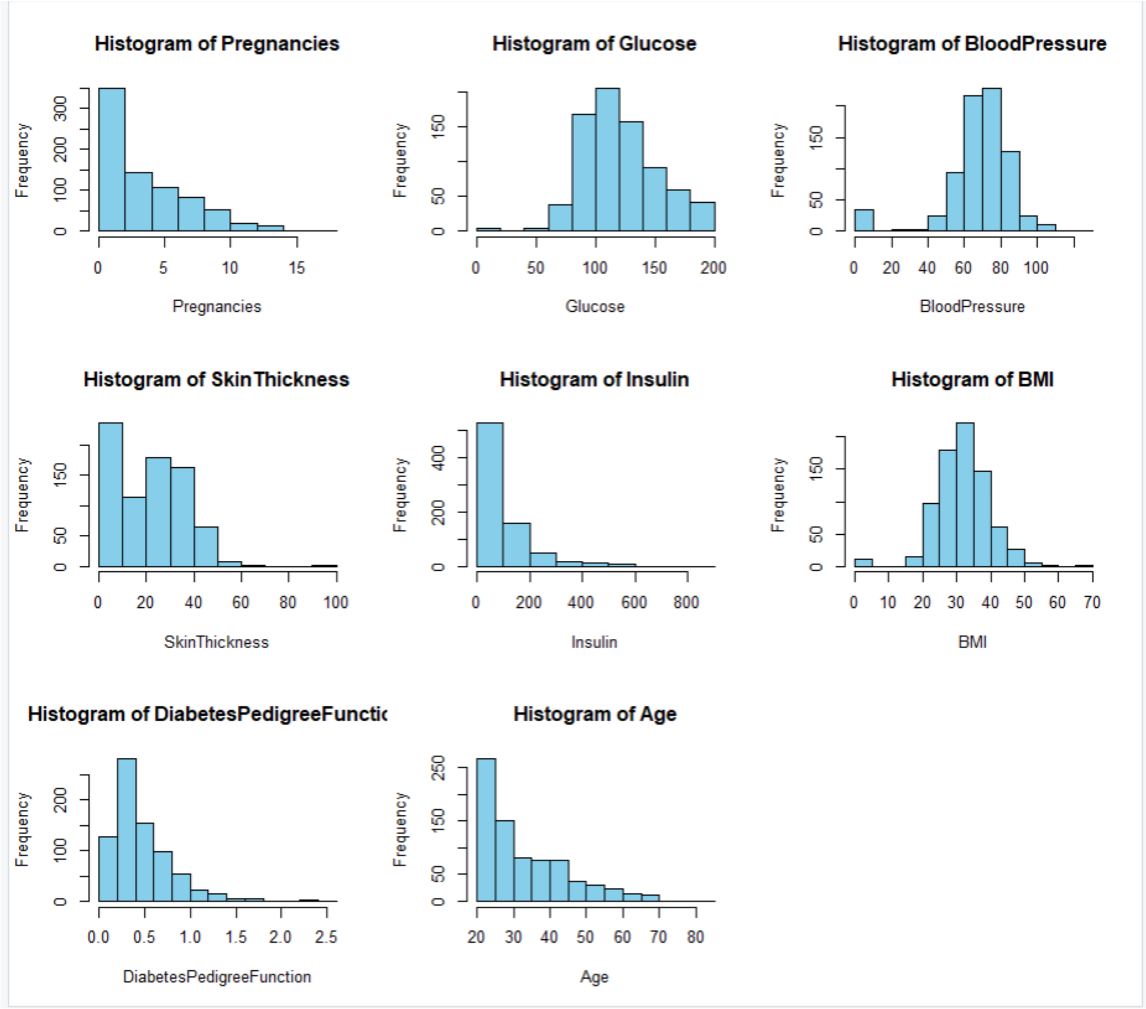
```
> X <- subset(data, select = -Outcome)
> X_scaled <- scale(X)
> pca_res <- prcomp(X_scaled, center = TRUE, scale. = TRUE)
> summary(pca_res)
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6     PC7     PC8
Standard deviation     1.4472 1.3158 1.0147 0.9357 0.87312 0.82621 0.64793 0.63597
Proportion of Variance 0.2618 0.2164 0.1287 0.1094 0.09529 0.08533 0.05248 0.05056
Cumulative Proportion  0.2618 0.4782 0.6069 0.7163 0.81164 0.89697 0.94944 1.00000
> pca_res$rotation
                                PC1        PC2         PC3         PC4        PC5          PC6         PC7          PC8
Pregnancies              -0.1284321  0.5937858 -0.01308692  0.08069115 -0.4756057  0.193598168 -0.58879003 -0.117840984
Glucose                  -0.3930826  0.1740291  0.46792282 -0.40432871  0.4663280  0.094161756 -0.06015291 -0.450355256
BloodPressure            -0.3600026  0.1838921 -0.53549442  0.05598649  0.3279531 -0.634115895 -0.19211793  0.011295538
SkinThickness            -0.4398243 -0.3319653 -0.23767380  0.03797608 -0.4878621  0.009589438  0.28221253 -0.566283799
Insulin                  -0.4350262 -0.2507811  0.33670893 -0.34994376 -0.3469348 -0.270650609 -0.13200992  0.548621381
BMI                      -0.4519413 -0.1009598 -0.36186463  0.05364595  0.2532038  0.685372179 -0.03536644  0.341517637
DiabetesPedigreeFunction -0.2706114 -0.1220690  0.43318905  0.83368010  0.1198105 -0.085784088 -0.08609107  0.008258731
Age                      -0.1980271  0.6205885  0.07524755  0.07120060 -0.1092900 -0.033357170  0.71208542  0.211661979
```

Proportion of Variance from the output above

- PC1 → 26.18%
- PC2 → 21.64%
- PC3 → 12.87%
- PC4 → 10.94%
- PC5–PC8 → small values



```
> # Calculate eigenvalues
> eigenvalues <- (pca_res$sdev)^2
> eigenvalues
[1] 2.0943799 1.7312101 1.0296299 0.8755290 0.7623444 0.6826284 0.4198162 0.4044620
>
> # Apply Kaiser Criterion: keep PCs with eigenvalue > 1
> kaiser_selected <- eigenvalues[eigenvalues > 1]
> kaiser_selected
[1] 2.09438 1.73121 1.02963
>
> # Number of PCs to keep
> length(kaiser_selected)
[1] 3
```



The histograms indicate substantial variation across predictors, with many variables showing right-skewed distributions and several containing zeros that represent missing measurements. This variation provides a strong basis for PCA, as it reflects meaningful differences in metabolic, clinical, and demographic characteristics among individuals.

PCA Results

Variance Explained

The PCA summary showed:

- **PC1:** 26.18% variance
- **PC2:** 21.64% variance
- **PC3:** 12.87% variance

Together, the first three components account for 60.69% of the total variation.

Interpretation of Principal Components

PC1 – Metabolic Health Component

PC1 had strong loadings from:

- BMI (–0.4519)
- SkinThickness (–0.4398)
- Insulin (–0.4350)
- Glucose (–0.3930)
- BloodPressure (–0.3600)

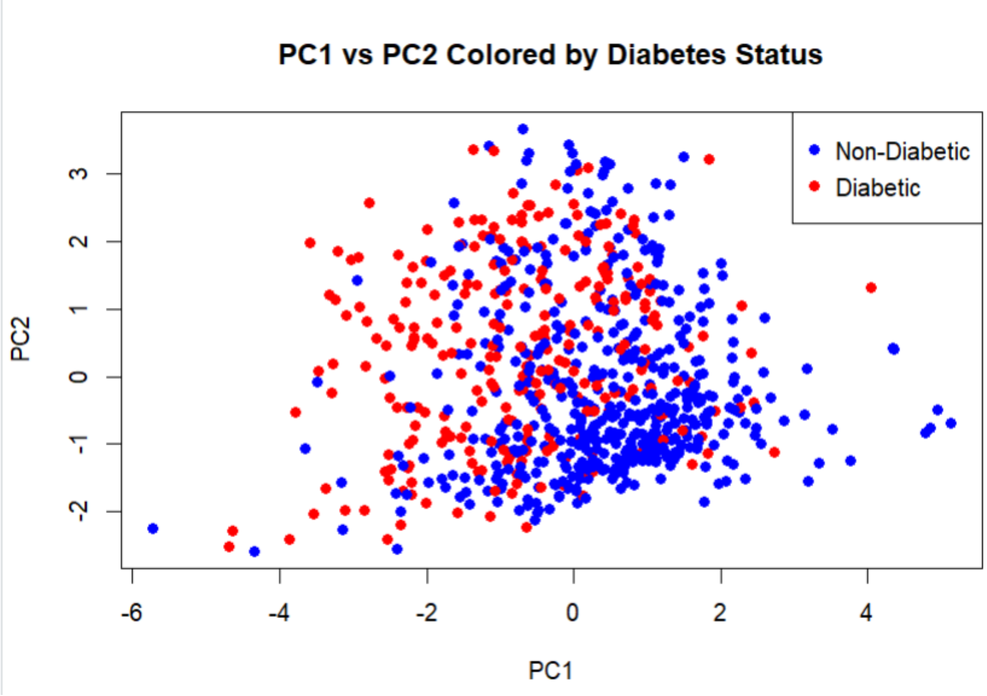
These measurements reflect obesity, insulin resistance, glucose regulation, and cardiometabolic health. Therefore: PC1 represents overall metabolic health.
Individuals with more negative PC1 scores tend to have poorer metabolic profiles associated with diabetes risk. PC2 – Age and Reproductive Component

PC2 was dominated by:

- Age (0.6209)
- Pregnancies (0.5938)

These variables describe demographic differences rather than metabolic ones. Thus: PC2 represents age and reproductive history, distinguishing older women with more pregnancies from younger individuals. PC3 – Blood Pressure & Genetic Predisposition Component PC3 had notable contributions from: Blood Pressure (–0.535) Diabetes Pedigree Function (0.434) BMI (–0.361) This suggests: PC3 reflects a secondary pattern involving blood pressure variation and family history of diabetes.

Visualization of PC1 and PC2



A scatter plot of PC1 vs PC2, colored by diabetes status, revealed: Non-diabetic individuals (blue) cluster more toward positive PC1 scores, reflecting healthier metabolic status. Diabetic individuals (red) tend to have more negative PC1 scores, indicating poorer metabolic health. PC2 did not meaningfully separate the two groups, consistent with its demographic interpretation. This shows that PCA successfully captured meaningful health-related differences, especially along PC1.

Conclusion

This PCA project successfully reduced eight correlated clinical variables into three principal components that collectively explain 60.69% of the total variance.

PC1 captured metabolic health status, PC2 reflected age and reproductive history, and PC3 represented blood pressure and family history factors. The PC1 vs PC2 plot revealed partial separation between diabetic and non-diabetic individuals, indicating that metabolic factors are strongly associated with diabetes outcomes in this dataset. Although PCA was not used for prediction in this project, it provided valuable insights into the underlying structure of the data and highlighted how health measurements cluster in relation to diabetes status.