ST405-MULTIVARIATE METHOD II

(Mini Project)

S/18/846 – K.B.J.D. WIJERATHNA

Diabetes Dataset

INTRODUCTION

This report aims to analyze whether a patient has diabetes. Using Factor Analysis techniques. It explores Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) on dataset. This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. I used " diabetes.csv" dataset to find how many factors are needed and we have to perform a hypothesis test to check whether the selected factors are enough or not.

METHODOLOGY

We want to use two techniques:

- **Exploratory Factor Analysis (EFA):** This helps us find hidden patterns in the data.
- **Confirmatory Factor Analysis (CFA):** This confirms if the patterns we found are reliable.

About the dataset:

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective is to predict based on diagnostic measurements whether a patient has diabetes.

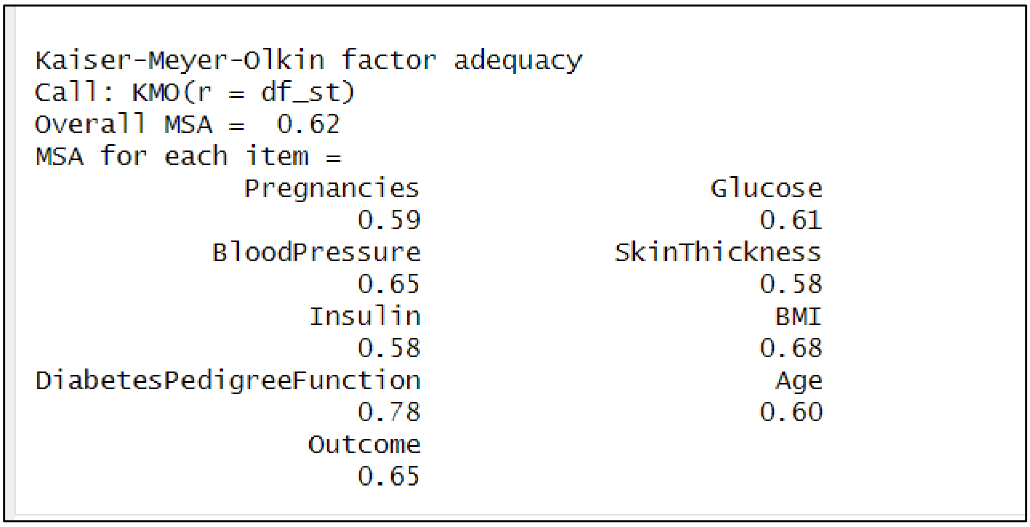
Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

Quantitative Attributes :

1. **Number of times pregnant**
2. **Plasma glucose concentration a 2 hours in an oral glucose tolerance test**
3. **Diastolic blood pressure (mm Hg)**
4. **Triceps skin fold thickness (mm)**
5. **2-Hour serum insulin (mu U/ml)**
6. **Body mass index (weight in kg/(height in m)^2)**
7. **Diabetes pedigree function**
8. **Age (years)**
9. **Class variable (0 or 1)**

Results and Discussion

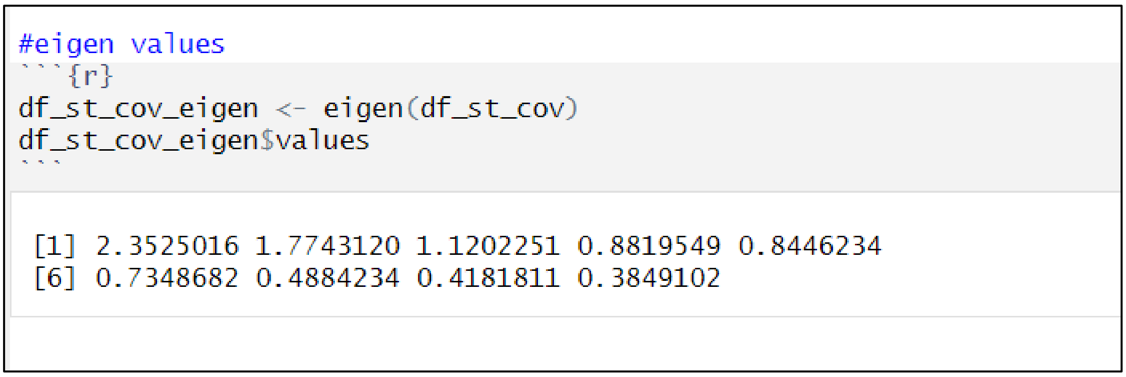
- KMO test output



```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = df_st)
Overall MSA =  0.62
MSA for each item =
                 Pregnancies                  Glucose
                        0.59                     0.61
               BloodPressure            SkinThickness
                        0.65                     0.58
                     Insulin                      BMI
                        0.58                     0.68
    DiabetesPedigreeFunction                      Age
                        0.78                     0.60
                     Outcome
                        0.65
```

The overall KMO value of my dataset is 0.62. It is greater than 0.6. Therefore, to the “diabetes.csv” dataset we can apply the factor analysis techniques.

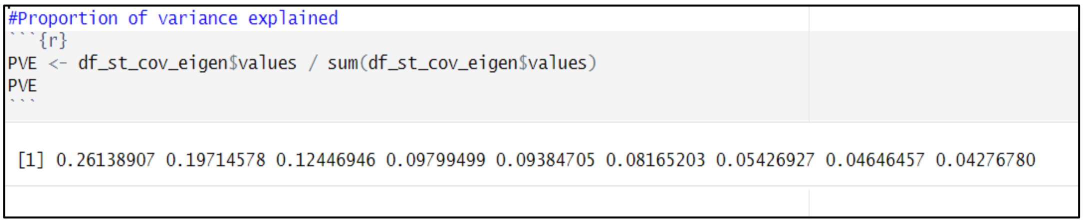
- Eigen values:



~~~
#eigen values
```{r}
df_st_cov_eigen <- eigen(df_st_cov)
df_st_cov_eigen$values
```

 [1] 2.3525016 1.7743120 1.1202251 0.8819549 0.8446234
 [6] 0.7348682 0.4884234 0.4181811 0.3849102
~~~

Using the eigen values we can say that three factors are sufficient for this dataset. Because, there are only three have the eigen values of greater than one.

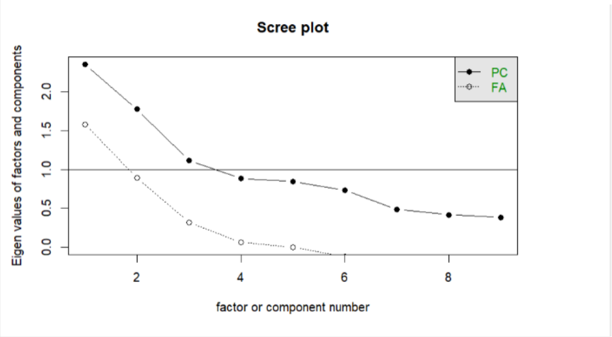
- Proportion of variance explained by each factor.

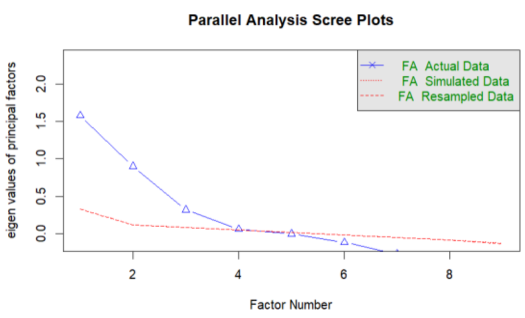


~~~
#Proportion of variance explained
```{r}
PVE <- df_st_cov_eigen$values / sum(df_st_cov_eigen$values)
PVE
```

 [1] 0.26138907 0.19714578 0.12446946 0.09799499 0.09384705 0.08165203 0.05426927 0.04646457 0.04276780
~~~

Cumulative proportion variance explained by first three factors = 0.583.
Therefore we can conclude that factor model explains 58.3% of total variance.

- Scree Plot and Parallel Analysis Scree Plots



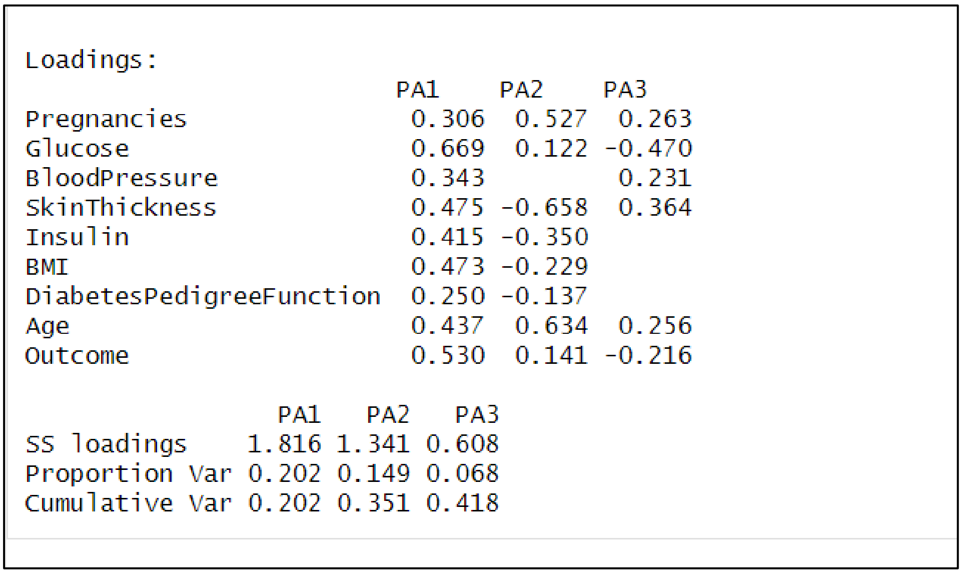


Using the Scree Plots we can see that three factors are sufficient for this dataset.

- Hypothesis testing

The harmonic n.obs is 768 with the empirical chi square 86.62 with prob < 2.2e-13
The total n.obs was 768 with Likelihood Chi Square = 106.53 with prob < 2.9e-17. Therefore we can conclude that 3 factor model is sufficient at 5% significance level.

- PCs factor loadings

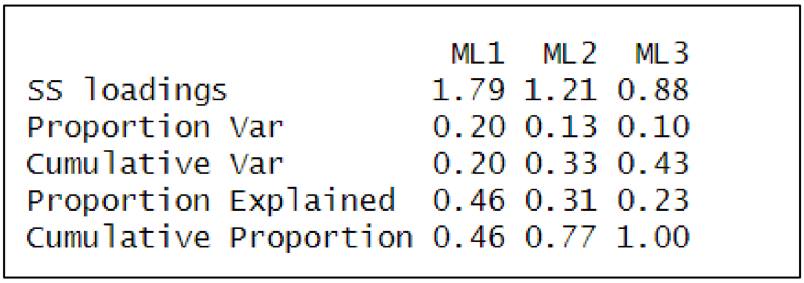


```
Loadings:
                         PA1    PA2    PA3
Pregnancies               0.306  0.527  0.263
Glucose                   0.669  0.122 -0.470
BloodPressure             0.343         0.231
SkinThickness             0.475 -0.658  0.364
Insulin                   0.415 -0.350
BMI                       0.473 -0.229
DiabetesPedigreeFunction  0.250 -0.137
Age                       0.437  0.634  0.256
Outcome                   0.530  0.141 -0.216

                 PA1   PA2   PA3
SS loadings    1.816 1.341 0.608
Proportion Var 0.202 0.149 0.068
Cumulative Var 0.202 0.351 0.418
```

- In factor 1, there is no contrast between variables.
- In factor 2, there is a contrast between ‘skinThickness’, ‘Insulin’, ‘BMI’, ‘DiabetesPedigreeFunction’ and other variables.
- In factor 3, there is a contrast between ‘Glucose’, ‘outcome’ and other variables.

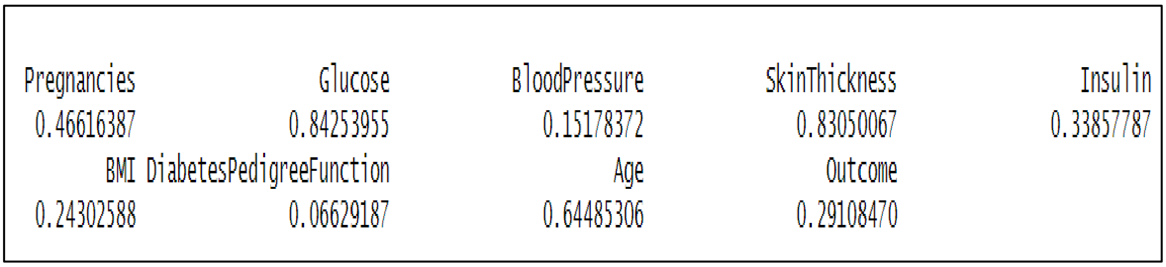
Factor loadings are not giving clear conclusion about the model. Therefore, we have to rotate them.

- Factor analyze using Maximum Likelihood method.



```
                       ML1  ML2  ML3
SS loadings           1.79 1.21 0.88
Proportion Var        0.20 0.13 0.10
Cumulative Var        0.20 0.33 0.43
Proportion Explained  0.46 0.31 0.23
Cumulative Proportion 0.46 0.77 1.00
```

- Communalities



```
Pregnancies              Glucose    BloodPressure    SkinThickness          Insulin
 0.46616387           0.84253955       0.15178372       0.83050067       0.33857787
        BMI DiabetesPedigreeFunction          Age          Outcome
 0.24302588           0.06629187       0.64485306       0.29108470
```

Conclusions and recommendation

- According to the analysis we can get three factors.

- The 58.3% of the total proportion is explained by the first three factors. 41.7% of the total proportion is explained by the other factors respectively.

- The total communality is 3.874821.

- The Proportion of the total variation explained by the three factors is 32.29%.

- The model explains Glucose, SkinThickness the best.

References

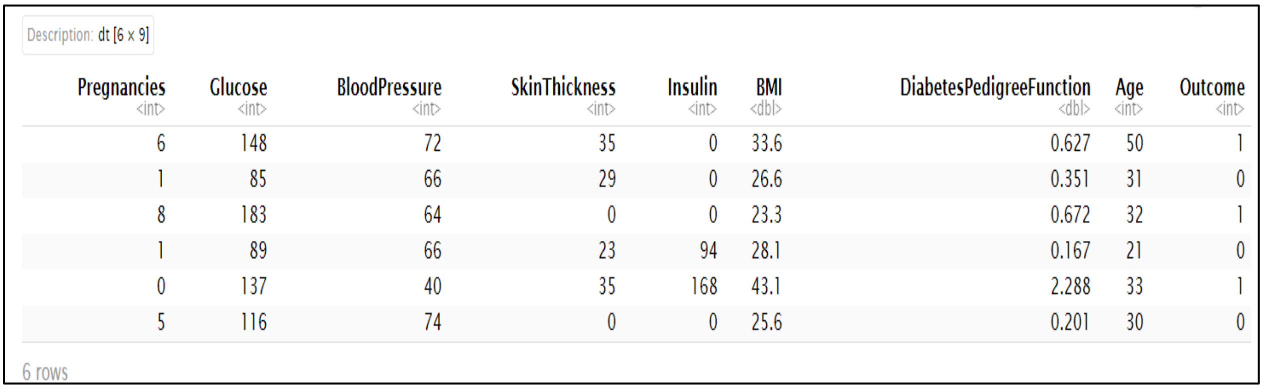
- National Institute of Diabetes and Digestive and Kidney Diseases. (n.d.). Original owners

- Sigillito, V. (1990). Donor of database.

Appendices

- Dataset - https://www.kaggle.com/datasets/mathchi/diabetes-data-set

- Head of the Dataset –

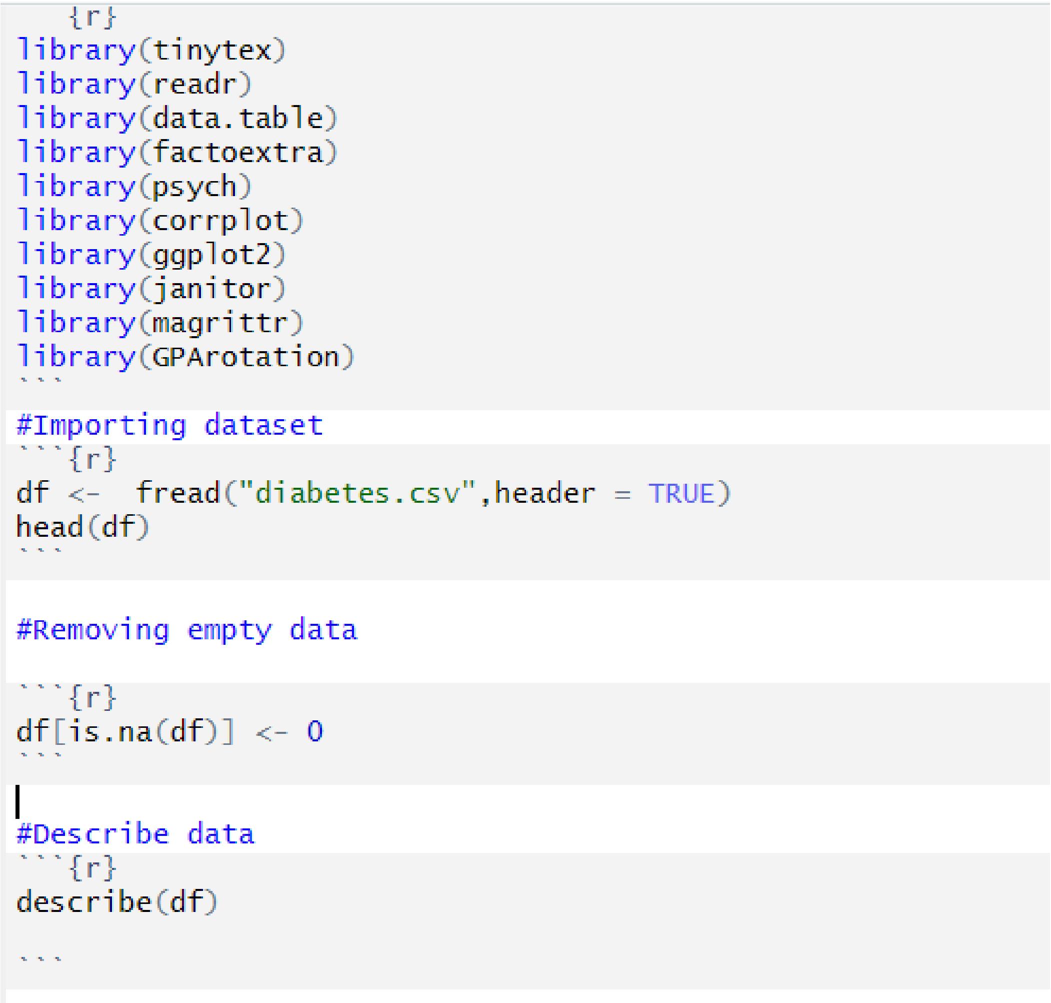


Description: dt [6 x 9]

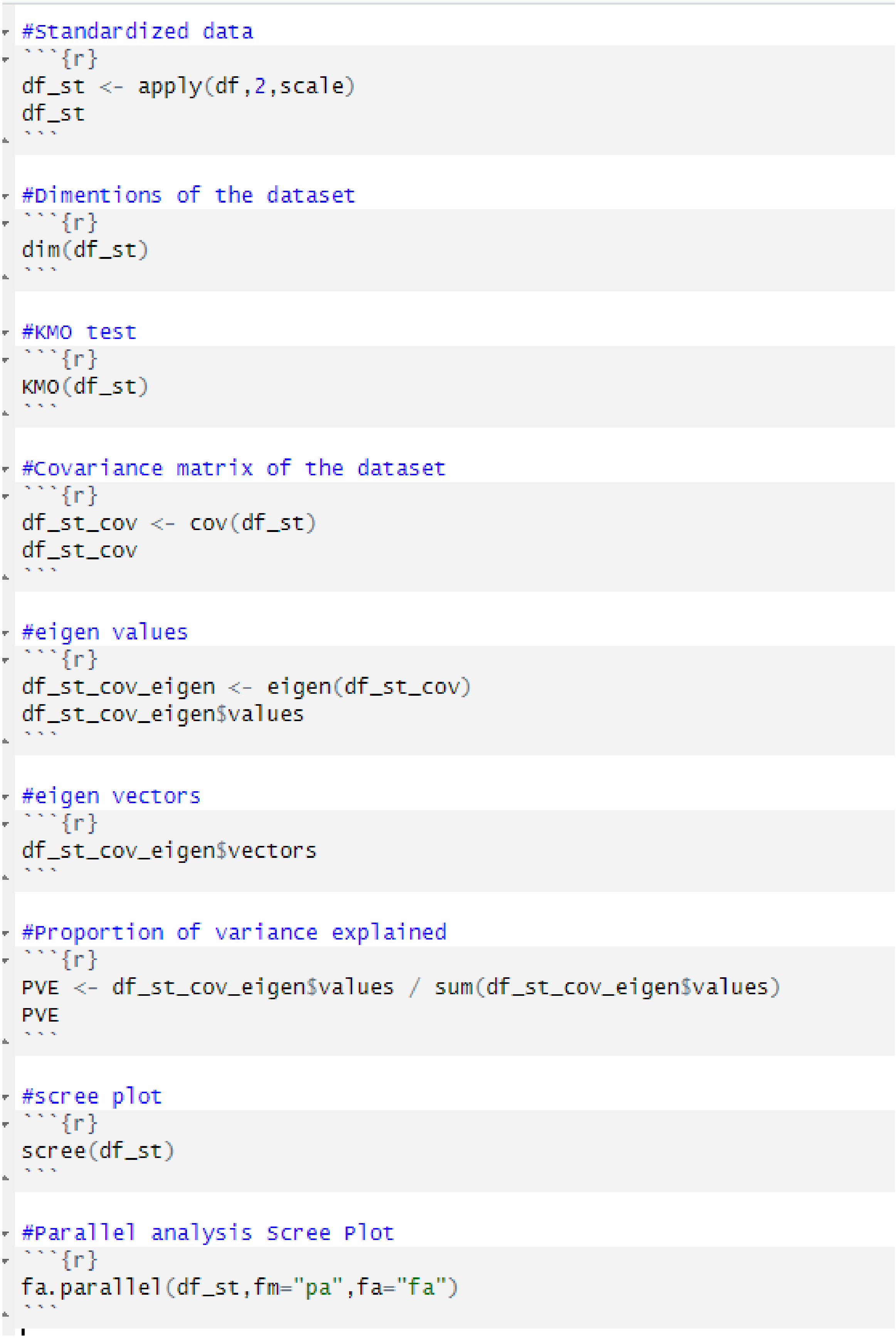
Pregnancies <int>	Glucose <int>	BloodPressure <int>	SkinThickness <int>	Insulin <int>	BMI <dbl>	DiabetesPedigreeFunction <dbl>	Age <int>	Outcome <int>
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0

6 rows

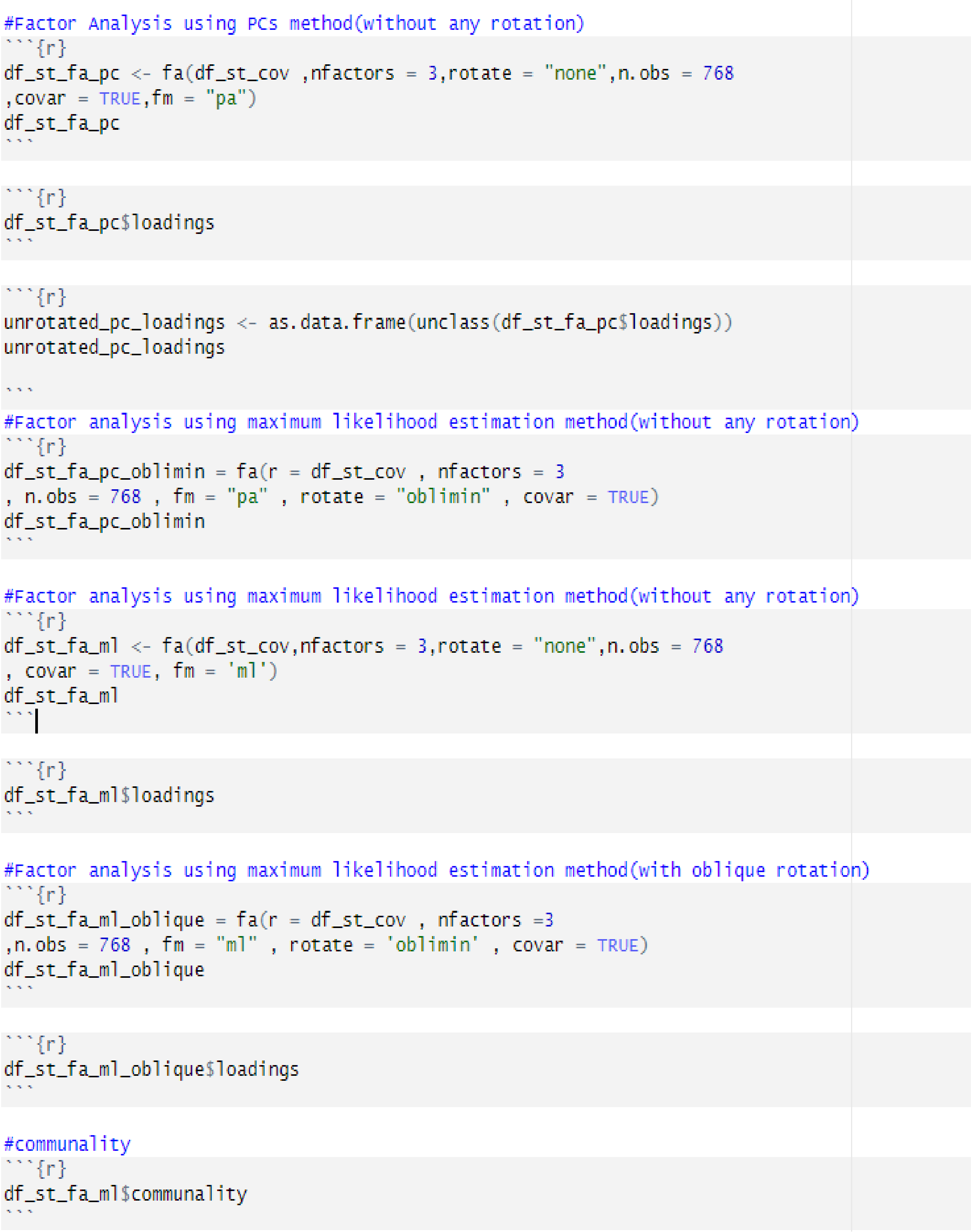
- **R Code**



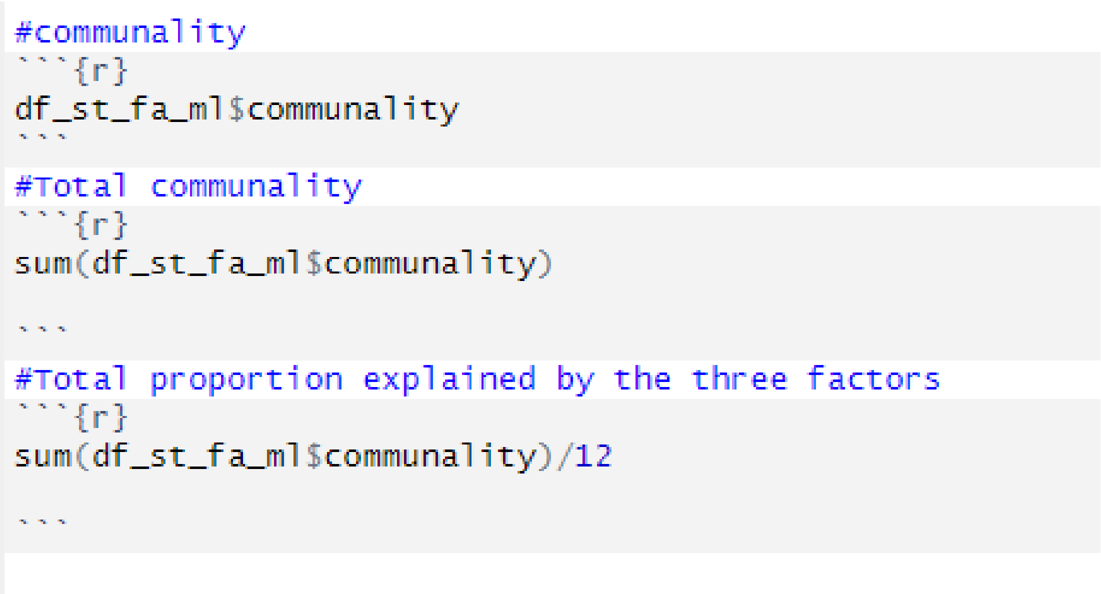
~~~
   {r}
library(tinytex)
library(readr)
library(data.table)
library(factoextra)
library(psych)
library(corrplot)
library(ggplot2)
library(janitor)
library(magrittr)
library(GPArotation)
```

#Importing dataset
```{r}
df <-  fread("diabetes.csv",header = TRUE)
head(df)
```

#Removing empty data

```{r}
df[is.na(df)] <- 0
```

#Describe data
```{r}
describe(df)

```
~~~



~~~
#Standardized data
```{r}
df_st <- apply(df,2,scale)
df_st
```

#Dimentions of the dataset
```{r}
dim(df_st)
```

#KMO test
```{r}
KMO(df_st)
```

#Covariance matrix of the dataset
```{r}
df_st_cov <- cov(df_st)
df_st_cov
```

#eigen values
```{r}
df_st_cov_eigen <- eigen(df_st_cov)
df_st_cov_eigen$values
```

#eigen vectors
```{r}
df_st_cov_eigen$vectors
```

#Proportion of variance explained
```{r}
PVE <- df_st_cov_eigen$values / sum(df_st_cov_eigen$values)
PVE
```

#scree plot
```{r}
scree(df_st)
```

#Parallel analysis Scree Plot
```{r}
fa.parallel(df_st,fm="pa",fa="fa")
```
~~~



~~~
#Factor Analysis using PCs method(without any rotation)
```{r}
df_st_fa_pc <- fa(df_st_cov ,nfactors = 3,rotate = "none",n.obs = 768
,covar = TRUE,fm = "pa")
df_st_fa_pc
```

```{r}
df_st_fa_pc$loadings
```

```{r}
unrotated_pc_loadings <- as.data.frame(unclass(df_st_fa_pc$loadings))
unrotated_pc_loadings

```
#Factor analysis using maximum likelihood estimation method(without any rotation)
```{r}
df_st_fa_pc_oblimin = fa(r = df_st_cov , nfactors = 3
, n.obs = 768 , fm = "pa" , rotate = "oblimin" , covar = TRUE)
df_st_fa_pc_oblimin
```

#Factor analysis using maximum likelihood estimation method(without any rotation)
```{r}
df_st_fa_ml <- fa(df_st_cov,nfactors = 3,rotate = "none",n.obs = 768
, covar = TRUE, fm = 'ml')
df_st_fa_ml
```

```{r}
df_st_fa_ml$loadings
```

#Factor analysis using maximum likelihood estimation method(with oblique rotation)
```{r}
df_st_fa_ml_oblique = fa(r = df_st_cov , nfactors =3
,n.obs = 768 , fm = "ml" , rotate = 'oblimin' , covar = TRUE)
df_st_fa_ml_oblique
```

```{r}
df_st_fa_ml_oblique$loadings
```

#communality
```{r}
df_st_fa_ml$communality
```
~~~



~~~
#communality
```{r}
df_st_fa_ml$communality
```
#Total communality
```{r}
sum(df_st_fa_ml$communality)

```
#Total proportion explained by the three factors
```{r}
sum(df_st_fa_ml$communality)/12

```
~~~