T.C.
DOKUZ EYLÜL ÜNİVERSİTESİ
FACULTY OF SCIENCE
DEPARTMENT OF STATISTICS

LECTURE OF MULTIVARIATE STATISTICAL ANALYSIS,END OF TERM PROJECT REPORT

ANALYSIS OF LIFE INDEX OF CITIES IN TURKEY BY USING MULTIVARIATE STATISTICAL ANALYSIS METHODS

Harun ŞEN
Furkan PAŞAHAN
Furkan ERGÜNEŞ

January 2023

Thanks

Thanks to our advisor PROF. DR. Esin Firuzan for the all the help,guidance and contributions on the project and thanks to Faculty of Science,Department of Statistics for the oppurtunity.

Abstract

In real-life studies we usually deal with multivariate datasets.Multivariate datasets are usually complex and hard to deal with and we use multivariate statistical methods to simplfy complex data sets.

Data set used in this project,Lıfe Index based on Cities on Turkey was taken from TUIK. Achievements gained in the lecture of Multivariate Statistical Analysis are used to analyse life index data set in this project.

Methods used in the analysis of data sets are PCA(Principal Component Analysis), FA(Factor Analysis), Classification with Discriminant Analysis and Clustering Analysis.Mehtods are applied on R,SPSS and Microsoft Excel.

Contents

Chapter 1

Presentation of Data-Set

Observations at the datasets are the 81 cities in Turkey.There are 42 variable to measure of life satisfaction.There are total of 3402 observations at the dataset.There are no missing values. variables at the dataset are continuous numerical variables.

1.1 Variables

- X1: number of rooms per person
- X2: Toilet availability rate in the residence
- X3: rate of persons who has quality problems in the residence(%)
- X4: Employment Rate(%)
- X5: Unemployment Rate (%)
- X6: Average income per day (TL)
- 7: Job Satisfaction Rate (%)
- X8: Savings Deposits per person
- X9: Proportion of households in the middle and higher income group
- X10: Proportion of households declaring that they cannot meet their basic needs
- X11: Infant mortality rate
- X12: life expectancy at birth
- X13: Number of applications per doctor
- X14: Health satisfaction rate
- X15: Satisfaction rate of public health services
- X16: Net enrollment rate in pre-school education (%)
- X17: Average score based on TEOG system placement
- X18: Average score based on YGS
- X19: Proportion of faculty or college graduates (%)
- X20: Satisfaction rate of public education services (%)
- X21: Average of PM10 station values (Air Pollution) ($\mu g/m^3$)
- X22: Forest are per Km2 (%)
- X23: Proportion of the population provided with waste services (%)
- X24: Proportion of people experiencing noise problems from the street (%)

- X25: Satisfaction rate of the municipality's cleaning services (%)
- X26: Murder rate per Million (bir milyon kişide)
- X27: Number of fatal and injured traffic accidents per thousand
- X28: Proportion of people who feel safe walking alone at night (%)
- X29: Satisfaction rate of public security services (%)
- X30: Participation rate in local government elections (%)
- X31: Membership rate of political parties (%)
- X32: Proportion of those related to union/association activities (%)
- X33: Number of internet subscribers per hunderd
- X34: Access rate to sewage and mains water (%)
- X35: Airport access rate
- X36: Satisfaction rate of the municipality's public transportation services (%)
- X37: Number of cinema and theater audience per hundred
- X38: Mall area per thousand m2
- X39: Satisfaction rate from social relationships (%)
- X40: Social life satisfaction rate (%)
- X41: Happiness level (%)

```
print(summary(data[,1:41]))
```

```
      x1              x2              x3              x4        
 Min.   :0.75   Min.   :50.31   Min.   : 9.38   Min.   :27.80  
 1st Qu.:1.16   1st Qu.:85.71   1st Qu.:15.31   1st Qu.:43.50  
 Median :1.38   Median :89.68   Median :18.72   Median :47.20  
 Mean   :1.31   Mean   :88.06   Mean   :21.12   Mean   :46.22  
 3rd Qu.:1.49   3rd Qu.:96.21   3rd Qu.:25.64   3rd Qu.:49.90  
 Max.   :1.68   Max.   :99.92   Max.   :44.73   Max.   :59.10  
      x5               x6              x7               x8         
 Min.   : 4.200   Min.   :46.87   Min.   :63.97   Min.   :  616.2  
 1st Qu.: 6.500   1st Qu.:53.31   1st Qu.:73.60   1st Qu.: 2259.6  
 Median : 7.300   Median :56.11   Median :79.21   Median : 3899.9  
 Mean   : 8.801   Mean   :57.69   Mean   :78.77   Mean   : 4342.1  
 3rd Qu.:10.000   3rd Qu.:59.89   3rd Qu.:83.10   3rd Qu.: 5818.1  
 Max.   :23.400   Max.   :85.55   Max.   :91.60   Max.   :18131.0  
      x9              x10             x11              x12        
 Min.   :16.27   Min.   :32.78   Min.   : 5.274   Min.   :74.95  
 1st Qu.:28.78   1st Qu.:44.46   1st Qu.: 8.567   1st Qu.:77.54  
 Median :34.06   Median :47.96   Median :10.314   Median :78.00  
 Mean   :34.38   Mean   :50.95   Mean   :10.995   Mean   :78.14  
 3rd Qu.:38.92   3rd Qu.:57.17   3rd Qu.:12.869   3rd Qu.:78.70  
 Max.   :58.91   Max.   :74.95   Max.   :25.728   Max.   :80.50  
     x13             x14             x15             x16             x17        
 Min.   :2763   Min.   :59.15   Min.   :54.55   Min.   :23.53   Min.   :215.3  
 1st Qu.:4955   1st Qu.:69.32   1st Qu.:72.60   1st Qu.:30.03   1st Qu.:292.2  
 Median :5787   Median :72.02   Median :78.85   Median :35.47   Median :304.8
```

```
Mean   :5834   Mean   :72.00   Mean   :77.47   Mean   :35.27   Mean   :295.9
3rd Qu.:6774   3rd Qu.:74.29   3rd Qu.:82.59   3rd Qu.:39.45   3rd Qu.:313.8
Max.   :8067   Max.   :80.76   Max.   :89.13   Max.   :53.16   Max.   :338.0
     x18             x19              x20             x21
Min.   :178.6   Min.   : 8.561   Min.   :48.18   Min.   : 18.00
1st Qu.:195.3   1st Qu.:11.728   1st Qu.:68.78   1st Qu.: 42.00
Median :198.6   Median :12.893   Median :74.63   Median : 53.00
Mean   :197.6   Mean   :13.119   Mean   :74.09   Mean   : 55.33
3rd Qu.:201.8   3rd Qu.:14.228   3rd Qu.:81.47   3rd Qu.: 66.00
Max.   :207.9   Max.   :22.653   Max.   :88.89   Max.   :113.00
     x22               x23              x24             x25
Min.   : 0.04394   Min.   : 35.71   Min.   : 6.36   Min.   :30.96
1st Qu.:13.57082   1st Qu.: 67.85   1st Qu.:11.56   1st Qu.:55.53
Median :34.10988   Median : 76.82   Median :14.84   Median :67.67
Mean   :30.70974   Mean   : 78.73   Mean   :15.66   Mean   :63.98
3rd Qu.:44.31943   3rd Qu.: 96.77   3rd Qu.:18.69   3rd Qu.:75.46
Max.   :69.70570   Max.   :100.00   Max.   :33.75   Max.   :88.25
     x26             x27              x28             x29
Min.   : 4.472   Min.   :0.7058   Min.   :45.10   Min.   :58.88
1st Qu.:17.659   1st Qu.:1.9375   1st Qu.:62.15   1st Qu.:82.11
Median :24.081   Median :2.3919   Median :68.61   Median :86.02
Mean   :25.499   Mean   :2.4375   Mean   :67.60   Mean   :84.24
3rd Qu.:29.759   3rd Qu.:3.0697   3rd Qu.:74.67   3rd Qu.:89.42
Max.   :69.343   Max.   :4.5936   Max.   :87.23   Max.   :94.86
     x30             x31              x32             x33
Min.   :77.10   Min.   :12.44   Min.   : 3.540   Min.   : 2.163
1st Qu.:86.50   1st Qu.:18.60   1st Qu.: 5.150   1st Qu.: 6.056
Median :89.10   Median :20.30   Median : 6.440   Median : 8.797
Mean   :88.16   Mean   :21.24   Mean   : 6.735   Mean   : 8.679
3rd Qu.:90.60   3rd Qu.:23.18   3rd Qu.: 7.720   3rd Qu.:10.995
Max.   :93.10   Max.   :34.73   Max.   :22.080   Max.   :17.664
     x34             x35              x36             x37
Min.   : 31.11   Min.   :   0.00   Min.   :23.46   Min.   :  0.2867
1st Qu.: 65.99   1st Qu.:  19.22   1st Qu.:51.01   1st Qu.: 19.2736
Median : 72.93   Median :  85.64   Median :59.70   Median : 39.1120
Mean   : 74.37   Mean   : 669.86   Mean   :58.39   Mean   : 45.3530
3rd Qu.: 88.00   3rd Qu.: 468.40   3rd Qu.:66.21   3rd Qu.: 63.7632
Max.   :100.00   Max.   :9874.83   Max.   :78.81   Max.   :147.4408
     x38             x39              x40             x41
Min.   :  0.00   Min.   :78.23   Min.   :21.50   Min.   :41.98
1st Qu.:  0.00   1st Qu.:86.24   1st Qu.:46.74   1st Qu.:56.54
Median : 57.51   Median :89.71   Median :52.85   Median :60.39
Mean   : 69.54   Mean   :88.83   Mean   :54.26   Mean   :61.15
3rd Qu.:115.25   3rd Qu.:91.66   3rd Qu.:62.13   3rd Qu.:65.57
Max.   :284.01   Max.   :96.19   Max.   :80.88   Max.   :77.66
```

to understand the structure of the dataset,we looked at the summary statistics of the variables. Our observations are the 81 cities of Turkey.There are no missing value in the entire dataset.

We can see that some variables has high variances rather than other variables that is because there are metric differences between variables.these metric differences are considered at the next stages.

1.2 *Multinormality Test*

```
HZ.test(data)
```

```
          Henze-Zirkler test for Multivariate Normality

data : data

HZ              : 1.000003
p-value         : 0.002233187

Result  : Data are not multivariate normal (sig.level = 0.05)
```

H0 : Data distributes multivariate normality
H1 : Data does not distribute multivariate normality

To check if the data distribute multivariate normality the Henze-Zirkler multivarite normality test is used. p-value is ~= 0.02 . There is statistically enough evidence to reject H0 in 95% confidence level. Data does not distribute multivarite normality

if every variable is tested one by one with shapiro wilk normality test

```
for(i in 1:ncol(data)) {
  if(shapiro.test(as.matrix(data[,i]))[2] > 0.05) {
    print(paste(i,". variable distributes normally",sep=""))
  }
}
```

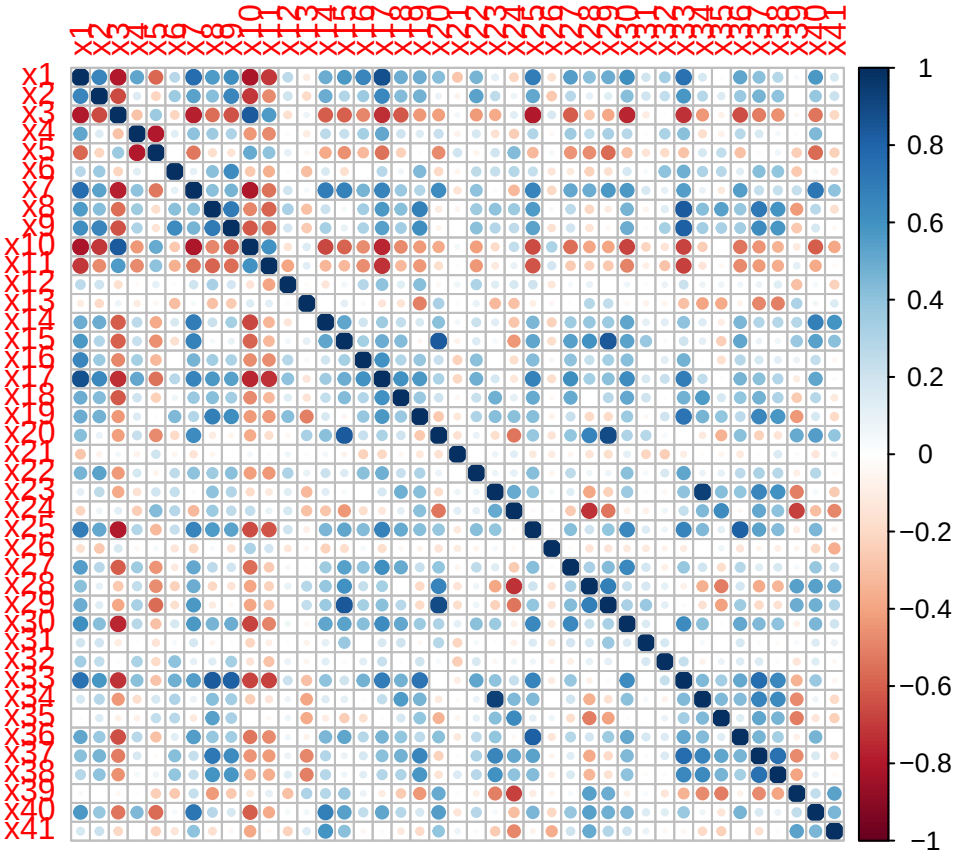
```
[1] "7. variable distributes normally"
[1] "9. variable distributes normally"
[1] "12. variable distributes normally"
[1] "13. variable distributes normally"
[1] "14. variable distributes normally"
[1] "16. variable distributes normally"
```

```
[1] "20. variable distributes normally"
[1] "27. variable distributes normally"
[1] "28. variable distributes normally"
[1] "33. variable distributes normally"
[1] "36. variable distributes normally"
[1] "40. variable distributes normally"
[1] "41. variable distributes normally"
```

the indexes of variables that ditributes normally are listed above.

1.3 Correlation Matrix

```
corrplot(cor(data))
```



at the correlation graph above, it is seen that there are highly correlated variables in the dataset.

1.3.1 List of Correlated Variables

```
# indexes of correlated variables
cor_index <- which((cor(data)>0.7) | (cor(data)< -0.7),arr.ind = T)

# correlation matrix is a symmetric matrix so repeated indexes are deleted
d_index <- c()
for(i in 1:103) {
  if (cor_index[i,1] == cor_index[i,2]) {
    d_index <- c(d_index,i)
  }
}
cor_index <- cor_index[-d_index,]

dd_index <- c()
for(i in 1:62) {
  if(i == 62) {
    break
  }
  for(x in ((i+1):62)) {
    if (all(c(cor_index[i,1],cor_index[i,2]) == c(cor_index[x,2],cor_index[x,1]))) {
      dd_index <- c(dd_index,x)
    }
  }
}


cor_index <- cor_index[-dd_index,]
rownames(cor_index) <- NULL
print(cor_index)
```

```
     row col
 [1,]   3   1
 [2,]   7   1
 [3,]  10   1
 [4,]  11   1
 [5,]  17   1
 [6,]  33   1
 [7,]  10   2
 [8,]   7   3
 [9,]  10   3
[10,]  17   3
[11,]  25   3
[12,]  30   3
```

```
[13,]  33   3
[14,]   5   4
[15,]  10   7
[16,]  40   7
[17,]   9   8
[18,]  33   8
[19,]  37   8
[20,]  33   9
[21,]  17  10
[22,]  17  11
[23,]  20  15
[24,]  29  15
[25,]  33  19
[26,]  29  20
[27,]  34  23
[28,]  28  24
[29,]  36  25
[30,]  37  33
[31,]  38  37
```

1.3.2 Determinant of Correlation Matrix

$$|cor(data)| = 6.27e^{-23}$$

determinant of correlation matrix is calculated approximately 0.There is multi-collinearity problem in the dataset.That information tells us dataset is proper for Principal Component Analyis.

Chapter 2

Application of Multivariate Statistical Analysis Mehtods

2.1 Principal Component Analysis

The intention of principal component analysis is to explain number of p variables that has multicollinearity problem between each other with m (less than p) variales that made with linear combination of p variables by minimum variance loss.

To apply PCA on data set.There has to be multicollinearity problem between variables.We can also determine the number of factors for Factor Analysis by using results of PCA.

2.1.1 KMO-Bartlett Test

H0 : p = I (correlation matrix is equal to identity matrix) There is no multicollinearity between variables
H1 : p ≠ I (correlation matrix is not equal to identity matrix) There is multicollinearity between variables

```
bartlett.test(data)
```

```
    Bartlett test of homogeneity of variances

data:  data
Bartlett's K-squared = 26030, df = 40, p-value < 2.2e-16
```

with 95% confidence There is enough evidence to reject H0. There is multicollinearity problem between variables.Dataset is proper for PCA.

2.1.2 Anti-Image KMO values

```
KMO(data)
```

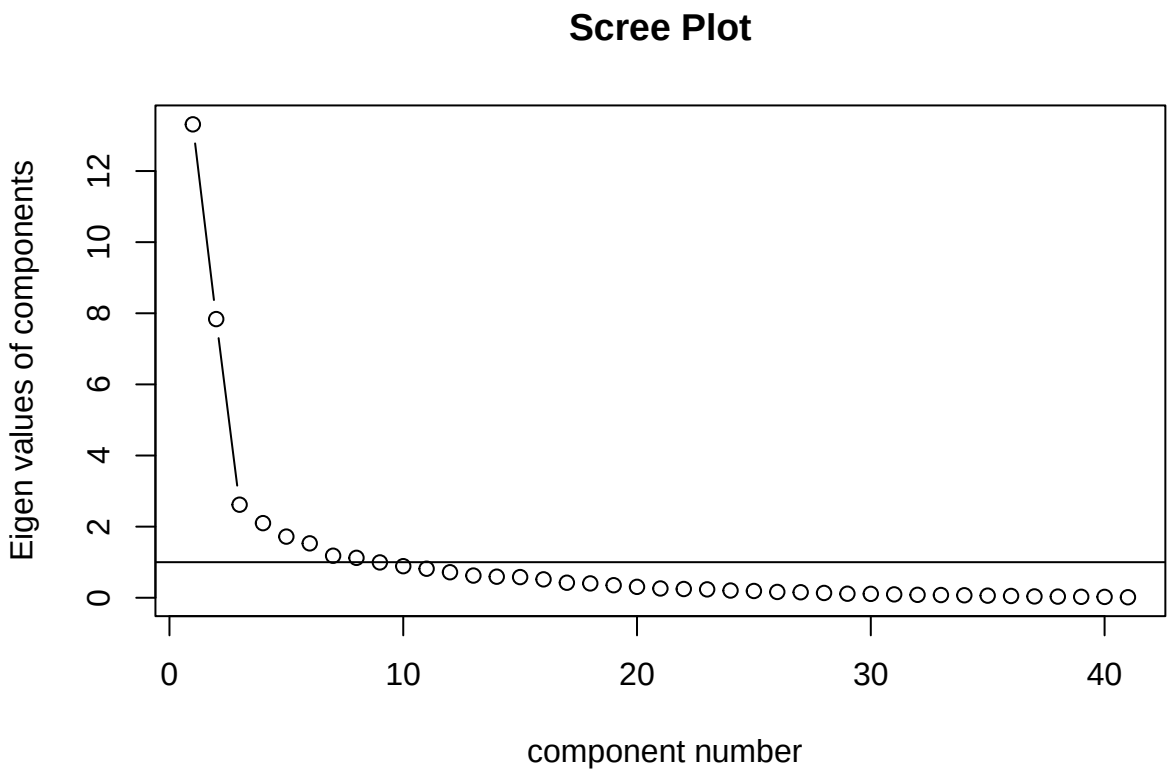
```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = data)
Overall MSA =  0.77
MSA for each item =
  x1   x2   x3   x4   x5   x6   x7   x8   x9  x10  x11  x12  x13  x14  x15  x16
0.92 0.69 0.83 0.73 0.82 0.58 0.90 0.70 0.82 0.91 0.88 0.33 0.59 0.83 0.73 0.84
 x17  x18  x19  x20  x21  x22  x23  x24  x25  x26  x27  x28  x29  x30  x31  x32
0.79 0.74 0.83 0.74 0.36 0.50 0.74 0.80 0.90 0.35 0.65 0.86 0.82 0.78 0.55 0.39
 x33  x34  x35  x36  x37  x38  x39  x40  x41
0.90 0.78 0.62 0.81 0.70 0.85 0.77 0.91 0.72
```

variables except X6,X12,X13,X21,X22,X26,X31 and X32 has KMO over 0.6. according to KMO (Measure of Sampling Adequacy) criteria. $KMO = 0.770 \geq 0.6$ since the KMO greater than 0.330. the information provided by the sample in data set is sufficient.

2.1.3 Determining Number of Components

While determining the number of components, criteria is to select the eigenvalues greater than 1 and consider the cumulative total variance explained. Scree plot is used to determine eigenvalues.

```
library(psych)
cor_mat <- cor(data)
VSS.scree(cor_mat,main="Scree Plot")
```




```
print(round(eigen(cor_mat)$values,2))
```

```
 [1] 13.31  7.84  2.62  2.10  1.72  1.53  1.18  1.12  0.99  0.89  0.82  0.72
[13]  0.62  0.59  0.58  0.52  0.42  0.40  0.35  0.31  0.26  0.25  0.23  0.20
[25]  0.19  0.16  0.15  0.13  0.11  0.11  0.10  0.08  0.08  0.07  0.06  0.05
[37]  0.04  0.03  0.03  0.02  0.01
```

```
tba.cor <- princomp(data,cor=T)
summary(tba.cor)
```

```
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4     Comp.5
Standard deviation     3.6486756 2.7994012 1.61778370 1.44839372 1.31127568
Proportion of Variance 0.3247033 0.1911377 0.06383473 0.05116694 0.04193766
Cumulative Proportion  0.3247033 0.5158410 0.57967571 0.63084265 0.67278030
                           Comp.6     Comp.7     Comp.8     Comp.9    Comp.10
Standard deviation     1.23677390 1.08693845 1.05929635 0.99653507 0.9420103
Proportion of Variance 0.03730755 0.02881549 0.02736851 0.02422152 0.0216435
Cumulative Proportion  0.71008786 0.73890335 0.76627186 0.79049337 0.8121369
                          Comp.11    Comp.12    Comp.13    Comp.14    Comp.15
Standard deviation     0.90335426 0.84682096 0.78979258 0.76943888 0.76204029
```

```
Proportion of Variance 0.01990363 0.01749038 0.01521396 0.01443991 0.01416355
Cumulative Proportion  0.83204050 0.84953089 0.86474485 0.87918475 0.89334830
                          Comp.16    Comp.17     Comp.18     Comp.19     Comp.20
Standard deviation     0.7188601 0.64998085 0.634974016 0.593487134 0.553200629
Proportion of Variance 0.0126039 0.01030427 0.009833951 0.008590902 0.007464169
Cumulative Proportion  0.9059522 0.91625647 0.926090418 0.934681320 0.942145489
                             Comp.21    Comp.22     Comp.23     Comp.24
Standard deviation     0.510287821 0.49579123 0.484708620 0.450047265
Proportion of Variance 0.006351065 0.00599534 0.005730304 0.004940062
Cumulative Proportion  0.948496554 0.95449189 0.960222198 0.965162260
                             Comp.25     Comp.26     Comp.27     Comp.28
Standard deviation     0.436102769 0.403525309 0.392226086 0.366949247
Proportion of Variance 0.004638674 0.003971529 0.003752227 0.003284189
Cumulative Proportion  0.969800934 0.973772462 0.977524689 0.980808878
                             Comp.29     Comp.30     Comp.31     Comp.32
Standard deviation     0.336877784 0.330841474 0.308264319 0.286216554
Proportion of Variance 0.002767967 0.002669661 0.002317729 0.001998047
Cumulative Proportion  0.983576845 0.986246506 0.988564235 0.990562281
                             Comp.33     Comp.34     Comp.35     Comp.36
Standard deviation     0.276259842 0.263026783 0.242610724 0.224052123
Proportion of Variance 0.001861451 0.001687392 0.001435609 0.001224374
Cumulative Proportion  0.992423733 0.994111125 0.995546734 0.996771108
                              Comp.37      Comp.38      Comp.39      Comp.40
Standard deviation     0.1925493484 0.1779551608 0.1609079486 0.1535752857
Proportion of Variance 0.0009042744 0.0007723912 0.0006314968 0.0005752529
Cumulative Proportion  0.9976753828 0.9984477740 0.9990792708 0.9996545237
                              Comp.41
Standard deviation     0.1190148266
Proportion of Variance 0.0003454763
Cumulative Proportion  1.0000000000
```

7 component are selected by considering the total variance explained and eigenvalues.73.89% of the total variance of original dataset is explained with 7 components

```
library(FactoMineR)
TBA2 <- PCA(data,scale.unit = T,ncp=7,graph=F)
```

2.1.4 Matrix of Loadings

	Component Matrix						
	1	2	3	4	5	6	7
X1	0.906	-0.159	-0.241	0.075	-0.022	0.043	-0.059
X2	0.708	0.112	-0.076	-0.282	0.132	0.240	-0.100
X3	-0.901	0.046	-0.206	0.004	-0.136	0.056	-0.059
X4	0.480	-0.217	-0.456	0.070	-0.320	-0.433	-0.026
X5	-0.501	0.487	0.256	-0.078	0.359	0.337	0.093
X6	0.337	0.436	-0.288	-0.378	-0.221	0.152	0.138
X7	0.815	-0.365	0.016	-0.095	0.069	0.013	0.069
X8	0.677	0.498	-0.178	0.046	0.018	-0.170	0.064
X9	0.695	0.397	-0.235	-0.259	0.030	0.013	0.085
X10	-0.883	0.168	-0.009	0.209	-0.056	0.028	0.136
X11	-0.732	-0.035	0.335	-0.004	-0.121	-0.018	-0.077
X12	0.251	0.178	-0.270	0.310	0.515	0.380	-0.177
X13	-0.185	-0.491	-0.020	0.121	0.326	0.130	0.467
X14	0.622	-0.296	0.236	-0.483	-0.053	-0.104	0.096
X15	0.602	-0.569	0.368	0.132	-0.045	0.150	-0.035
X16	0.561	-0.137	-0.309	0.340	0.252	-0.077	-0.007
X17	0.872	-0.060	-0.237	0.217	0.080	0.025	-0.199
X18	0.619	0.150	0.307	0.351	-0.048	-0.001	-0.210
X19	0.583	0.561	-0.202	0.015	0.074	0.019	-0.252
X20	0.396	-0.733	0.239	0.126	-0.120	0.181	0.174
X21	-0.125	0.097	0.230	-0.281	0.483	-0.264	0.015
X22	0.516	0.098	-0.239	-0.004	0.395	0.169	0.257
X23	0.353	0.616	0.528	0.126	-0.060	0.026	0.050
X24	-0.063	0.837	0.053	0.021	0.002	0.171	0.035
X25	0.826	0.040	0.148	0.195	0.000	0.039	0.316
X26	-0.196	0.112	-0.032	0.455	-0.163	-0.367	0.273
X27	0.593	-0.261	0.186	0.275	0.086	-0.258	-0.325
X28	0.283	-0.805	-0.068	0.000	0.046	0.115	-0.057
X29	0.430	-0.729	0.104	0.215	-0.162	0.185	0.027
X30	0.766	-0.029	0.256	0.026	0.169	-0.158	0.018
X31	0.199	-0.186	0.133	0.097	-0.357	0.542	-0.281
X32	0.228	0.050	-0.513	-0.147	-0.403	0.351	0.141
X33	0.843	0.380	-0.167	-0.060	0.022	-0.132	0.065
X34	0.415	0.591	0.557	0.116	-0.122	0.055	-0.014
X35	0.119	0.694	0.073	-0.070	-0.214	0.074	0.060
X36	0.681	-0.012	0.265	0.113	-0.179	0.102	0.389
X37	0.608	0.639	0.135	-0.100	-0.143	-0.069	0.038
X38	0.516	0.579	0.221	-0.212	0.023	-0.092	-0.035
X39	-0.075	-0.768	0.057	-0.263	0.068	-0.067	0.070
X40	0.633	-0.407	0.059	-0.175	-0.111	-0.053	0.052
X41	0.170	-0.548	0.216	-0.625	0.078	-0.030	-0.189

Figure 2.1: Loadings of variables at each component

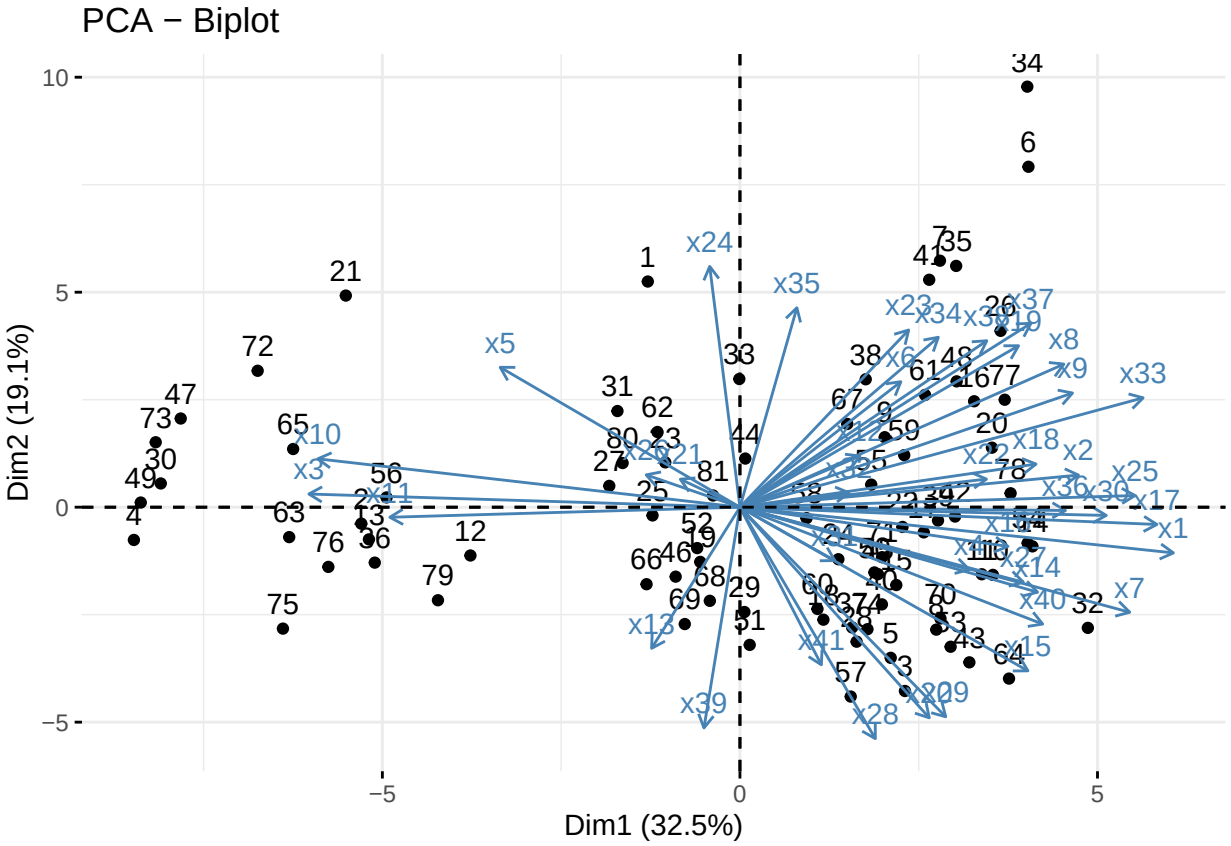
At the loadings matrix,There are effect scores of variables to components.criteria for scores is to be greater than 0.333 or lesser than -0.333.Scores out of this boundries are not considered.Scores that greater than 0.333 are painted red.Scores that lesser than -0.333 are painted green.

- 1. Component takes scores from 30 variable.
- 2. Component takes scores from 21 variable.
- 3. Component takes scores from 6 variable.
- 4. Component takes scores from 6 variable.

- 5. Component takes scores from 6 variable.
- 6. Component takes scores from 6 variable.
- 7. Component takes scores from 2 variable.

2.1.5 PCA-Biplot Graph

```
fviz_pca_biplot(TBA2)
```



At biplot graph,in 2 dimension how close variables and observations are to dimensions is seen.

2.2 Factor Analysis

Factor analysis is a method that helps to explain a structure that is explained with p number of variables that have multicollinearity between each other, with fewer new variables that are not related.

Factor analysis is an extension of Principal Component Analysis.Before the Factor Analysis PCA must be applied.Essential requirement for factor analysis is that if the dataset is consistent or not with the previously defined structure.

There are several methods to apply Factor Analysis."Maximum Likelihood" Method is used in this study.

```
FA <- factanal(x=data,factors=7,rotation="varimax")
```

2.2.1 Communalities.

```
communalities <- rowSums(FA$loadings^2)
```

Communalities

	Initial	Extraction		Initial	Extraction
X1	0.946	0.934	X22	0.822	0.429
X2	0.898	0.665	X23	0.936	0.781
X3	0.945	0.892	X24	0.865	0.700
X4	0.858	0.803	X25	0.900	0.862
X5	0.855	0.830	X26	0.634	0.192
X6	0.767	0.501	X27	0.840	0.667
X7	0.898	0.814	X28	0.831	0.725
X8	0.921	0.758	X29	0.935	0.854
X9	0.909	0.748	X30	0.849	0.680
X10	0.915	0.877	X31	0.475	0.299
X11	0.819	0.642	X32	0.734	0.463
X12	0.840	0.428	X33	0.948	0.922
X13	0.682	0.383	X34	0.948	0.883
X14	0.825	0.758	X35	0.778	0.494
X15	0.952	0.866	X36	0.851	0.701
X16	0.700	0.522	X37	0.940	0.830
X17	0.963	0.951	X38	0.820	0.658
X18	0.829	0.616	X39	0.810	0.633
X19	0.881	0.728	X40	0.809	0.579
X20	0.938	0.852	X41	0.836	0.810
X21	0.639	0.201			

Figure 2.2: Explained variance ratios of original variables at Factors

```
data <- data[,-c(21,26,31)]
FA <- factanal(x=data,factors=7,rotation="varimax",scores = "regression")
```

X21,X26 and X31 variables are removed from FA because explained variance rates of these variables are quiet low.

	Initial Eigenvalues		
Factor	Total	% of Variance	Cumulative %
1	13,227	34,809	34,809
2	7,787	20,493	55,302
3	2,571	6,766	62,067
4	1,968	5,180	67,247
5	1,554	4,089	71,337
6	1,283	3,377	74,714
7	1,026	2,701	77,415

After X21,X26 and X31 removed from dataset. Total variance explained of original dataset is increased to 77.41%.

2.2.2 Residual Matrix

Residual matrix is the difference between the correlation matrix of original data and the correlation matrix of the variables in FA model.Residuals must not be higher than 0.05. Ratio of residuals greater than 0.05 to the total must be less than 20%.

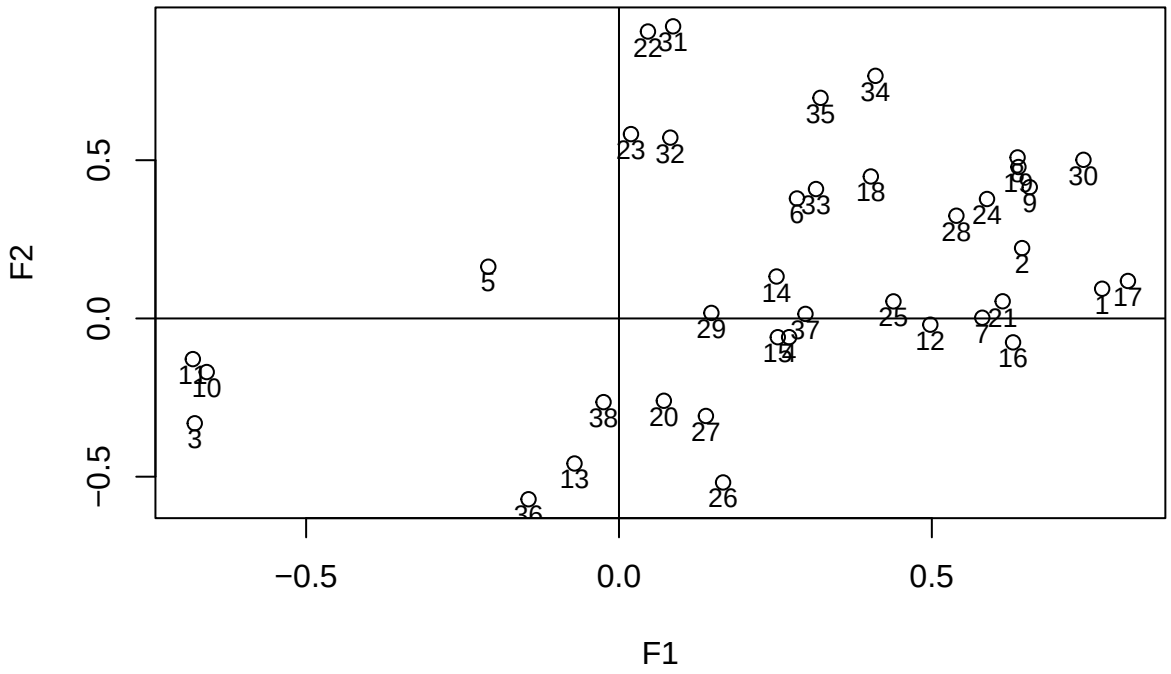
The ratio of residuals greater than 0.05 to the total is determined as 15% so our FA model is sufficient enough.

2.2.3 Naming Factors

- F1: High Life Satisfaction
- F2: Middle Income
- F3: Satisfaction of Public Services
- F4: High Sociality and Happiness
- F5: Employment Rate
- F6: Rural Life
- F7: Education Level

2.2.4 Factor Graph

```
plot(FA$loadings[,1],FA$loadings[,2],xlab="F1",ylab="F2")
abline(v=0,h=0)
text(FA$loadings[,1],FA$loadings[,2]-.05,labels=1:38,cex=0.8)
```



in graph of dimensions of PC1 and PC2 we see that x32,x23,x2,x31... variables are more close to PC2 (y-axis).X35,X34,X18... variables are far away to both PC1 and PC2 dimensions.

2.3 Discriminant Analysis

Discriminant analysis is a linear binary classification method, unlike logistic regression it requires independent variables to distribute normally.Discriminant analysis basically splits the dataset into two pieces by grouping variable.Discriminant analysis creates a rule by using the two datasets and uses this rule on a new observation to predict if it is 0 or 1.s

2.3.1 Descriptive Statistics

```
new_scores <- FA$scores
new_scores <- as.data.frame(new_scores)
colnames(new_scores) <- c("High Life Satisfaction",
"Middle Income",
"Satisfaction of Public Services",
"High Sociality and Happiness",
```

```
"Employment Rate",
"Rural Life",
"Education Level")
summary(new_scores)
```

```
High Life Satisfaction Middle Income     Satisfaction of Public Services
Min.   :-2.6353        Min.   :-1.9518   Min.   :-3.9997
1st Qu.:-0.5055        1st Qu.:-0.7298   1st Qu.:-0.4337
Median : 0.1717        Median :-0.2211   Median : 0.1936
Mean   : 0.0000        Mean   : 0.0000   Mean   : 0.0000
3rd Qu.: 0.7706        3rd Qu.: 0.8297   3rd Qu.: 0.7776
Max.   : 1.5933        Max.   : 2.6509   Max.   : 1.2269
High Sociality and Happiness Employment Rate      Rural Life
Min.   :-2.58196             Min.   :-3.19143   Min.   :-1.68536
1st Qu.:-0.49221             1st Qu.:-0.33361   1st Qu.:-0.67659
Median :-0.08242             Median : 0.09335   Median : 0.01165
Mean   : 0.00000             Mean   : 0.00000   Mean   : 0.00000
3rd Qu.: 0.55134             3rd Qu.: 0.48471   3rd Qu.: 0.53161
Max.   : 2.47560             Max.   : 2.34610   Max.   : 2.32793
Education Level
Min.   :-2.53127
1st Qu.:-0.69834
Median :-0.01643
Mean   : 0.00000
3rd Qu.: 0.61473
Max.   : 1.92532
```

2.3.2 Normallity Test

```
HZ.test(new_scores)
```

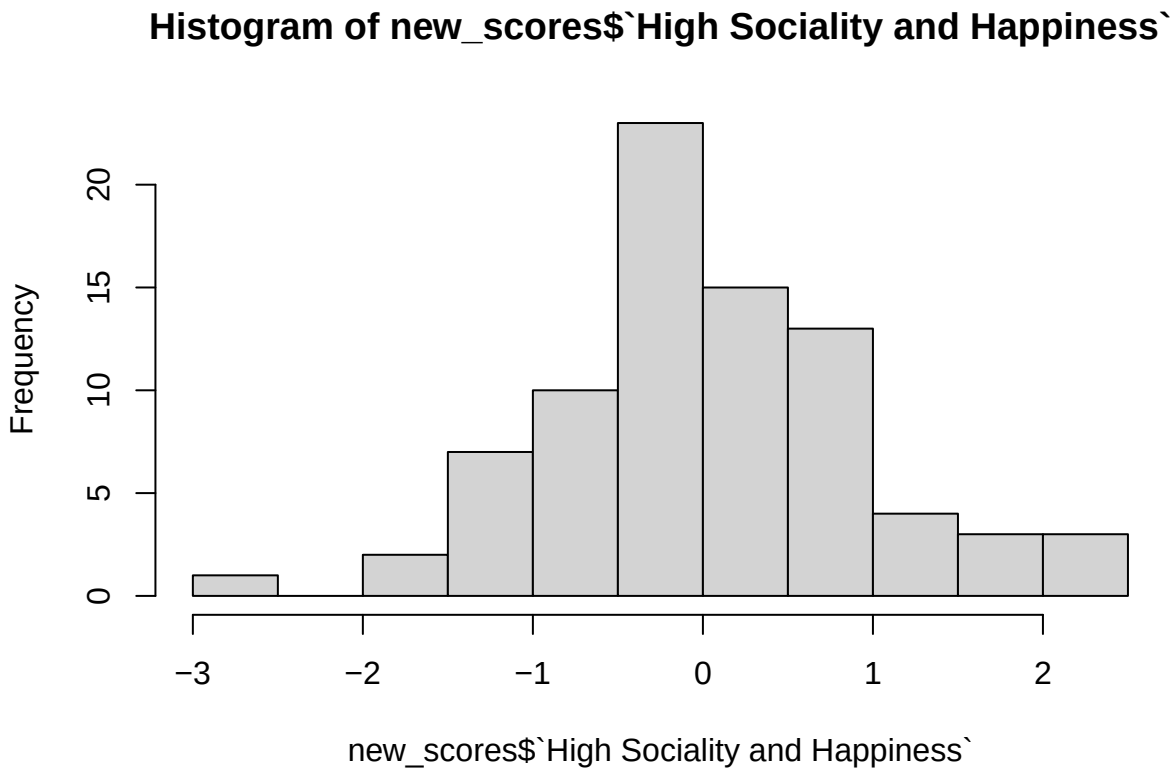
```
          Henze-Zirkler test for Multivariate Normality

data : new_scores

HZ              : 1.588391
p-value         : 0

Result  : Data are not multivariate normal (sig.level = 0.05)
```

H0: Data distributes Multivariate Normal
H1: Data does not distribute Multivariate Normal
p-value ~= 0 There is enough evidence to reject H0.Data does not distibute multivariate normal.

2.3.3 Grouping Variable

Between factor variables,High Sociallity and Happines is selected for classification.Values lesser than or equal to 0 are coded as 0 and Values Greater than 0 are coded as 1.The reason threshold is selected 0 because data is distributed around mean 0 symmetrical.

```
hist(new_scores$`High Sociality and Happiness`)
```




```
cat_var <- c()
for(x in new_scores$`High Sociality and Happiness`) {
  if (x < 0) {
    cat_var <- c(cat_var,0)
  } else {
    cat_var <- c(cat_var,1)
  }
}

new_scores$cat_var <- cat_var
new_scores$cat_var <- as.factor(new_scores$cat_var)
new_scores <- new_scores[,-4]
```

```
table(new_scores$cat_var)
```

```
 0  1
43 38
```

frequency table of grouping variable is listed above.

```
library(MASS)
disc_lda <- lda(new_scores$cat_var~. , data=new_scores)


predicted_val <- predict(disc_lda,newdata=new_scores[,-7])
```

2.3.4 Wilk's Lambda Test

Wilk's lambda test used to test if there is difference between mean by grouping variable.

```
library(rrcov)
Wilks.test(new_scores[,-7],grouping=new_scores$cat_var,method="c")
```

```
    One-way MANOVA (Bartlett Chi2)

data:  x
Wilks' Lambda = 0.97039, Chi2-Value = 2.2844, DF = 6.0000, p-value =
0.8918
sample estimates:
  High Life Satisfaction Middle Income Satisfaction of Public Services
0            0.008110115   -0.04635419                      -0.1200339
1           -0.009177235    0.05245342                       0.1358279
  Employment Rate  Rural Life Education Level
0      0.04304380  0.06370616      0.03199056
1     -0.04870746 -0.07208855     -0.03619985
```

H0 : There is no difference between means by grouping variable
H1 : There is difference between mean by grouping variable
p-value ~= 0.89. There is not enough evidence to reject H0 so There is no difference between means by grouping variable.

2.3.5 BoxM Test.

BoxM test used to test if the variance-covariance matrixes are equal by grouping variable.

```
boxM(new_scores[,-7],group=new_scores$cat_var)
```

```
    Box's M-test for Homogeneity of Covariance Matrices

data:  new_scores[, -7]
Chi-Sq (approx.) = 49.793, df = 21, p-value = 0.0003898
```

H0 : Variance-Covariance matrices are equal by grouping variable.
H1 : Variance-Covariance matrices are not equal by grouping variable
p-value ~= 0.0004, There is enough evidence to reject H0 so Variance-Covariance matrices are not equal by grouping variable.

2.3.6 Canonic Discriminant Function

Canonic Discriminant Function gives the explained variance rate on the grouping variable.Its better as close to 1.

Summary of Canonical Discriminant Functions

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	,021[a]	100,0	100,0	,145

a. First 1 canonical discriminant functions were used in the analysis.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	,979	1,612	6	,952

Figure 2.3: Summary of Canonical Discriminant Function

H0: Canonical Discriminant function is not statically significant
H1: Canonical Discriminant function is statistically significant
p-value ~=0.95. There is not enough evidence to reject H0.Canonical Discriminant Function is not statistically significant.

2.3.7 Fisher's Linear Discriminant Function

To classify the observations, values on this function used to determine whether the observation belongs to 0 or 1.

Classification Function Coefficients

	MutlulukBinary	
	0	1
YuksekYasamKonforu	-,029	,034
OrtaSinif	-,097	,116
KamuHizmetiMemnuniyeti	,025	-,029
IstihdamDuzeyi	,068	-,081
KirsalYasam	-,015	,018
EgitimDuzeyi	-,046	,055
(Constant)	-,702	-,706

Fisher's linear discriminant functions

Discriminant Equations

Unhappy(0) = -0.702 - 0.029(High Life Satisfaction) - 0.097(Middle Income) + 0.025(Satisfaction of Public Services) + 0.068(Employment Rate) - 0.015(Rural Life) - 0.046(Education Level)

Happy(1) = -0.706 + 0.034(High Life Satisfaction) + 0.116(Middle Income) - 0.029 (Satisfaction of Public Services) - 0.081(Employment Rate) + 0.018(Rural Life) + 0.055(Education Level)

2.3.8 Classification Results

```
table(new_scores$cat_var, predicted_val$class, dnn = c('Actual Group', 'Predicted Group'
```

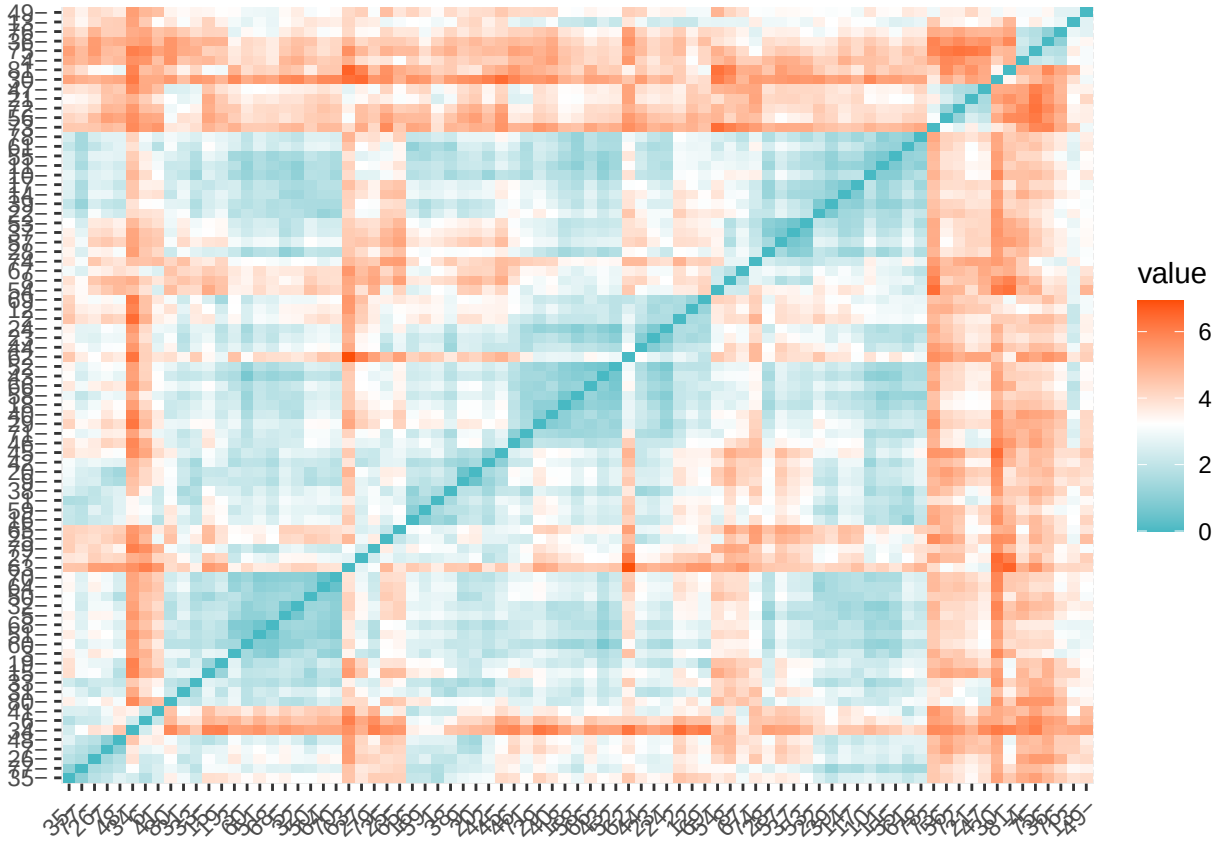
```
            Predicted Group
Actual Group  0  1
           0 32 11
           1 19 19
```

Sensivity (True Positive Rate) = 32/51 = 0.6274
Specifity (True Negative Rate) = 19/30 = 0.634

2.4 Clustering

2.4.1 Euclidean Distances Between Observations

```
k1 <- new_scores[,-7]
sk1 <- k1 %>% mutate_if(is.numeric, scale)

library(factoextra)
distance <- get_dist(sk1,method = "euclidean")

fviz_dist(distance, gradient = list(low = "#00AFBB", mid = "white", high = "#FC4E0
```



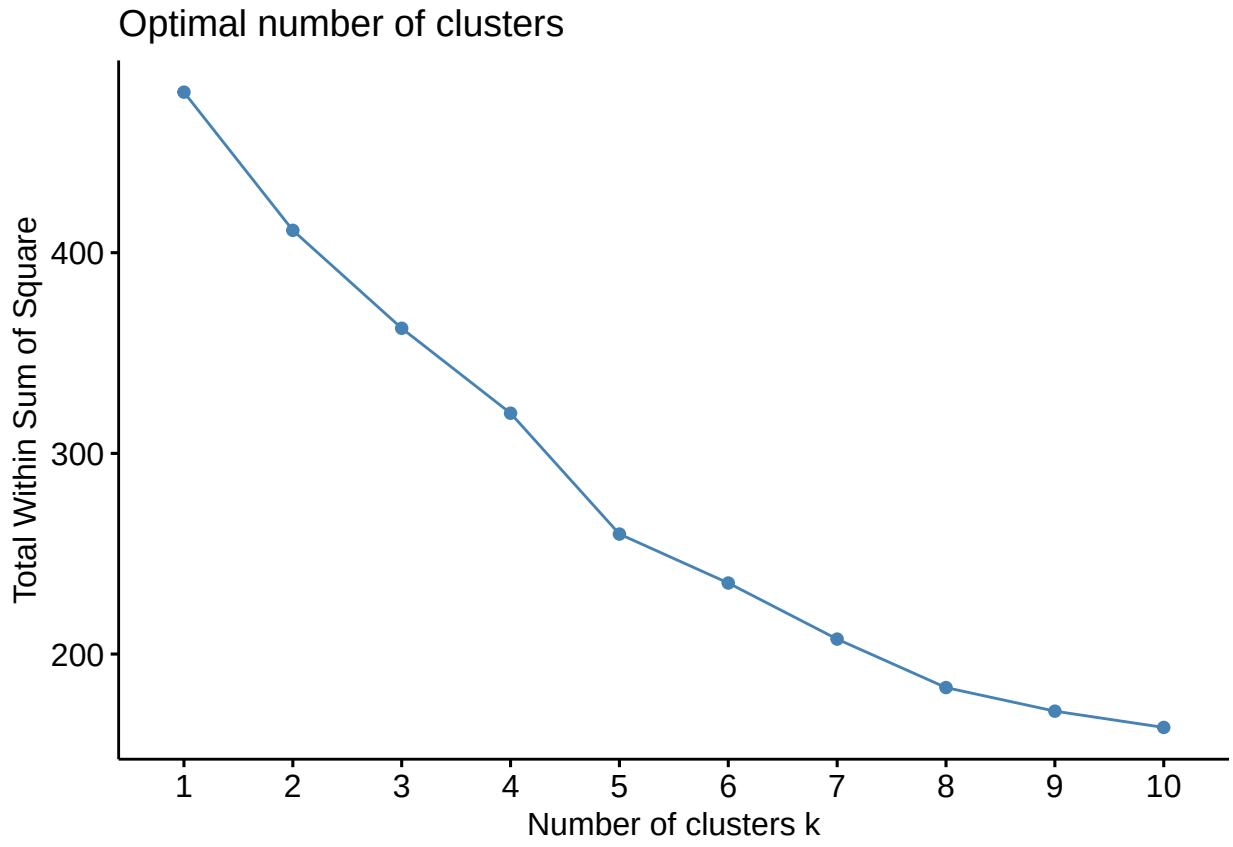
Heatmap graph is drawn with euclidean distances.

2.4.2 Determining Number of Clusters

To apply clustering methods,We have to determine the number of clusters first. To determine optimal number of clusters there are a few methods: - Within Sum of Square (WSS) - Average Sillhouette Indeks - Gap Statistic In this project.We used WSS and Silhoutte methods to determine number of clusters.

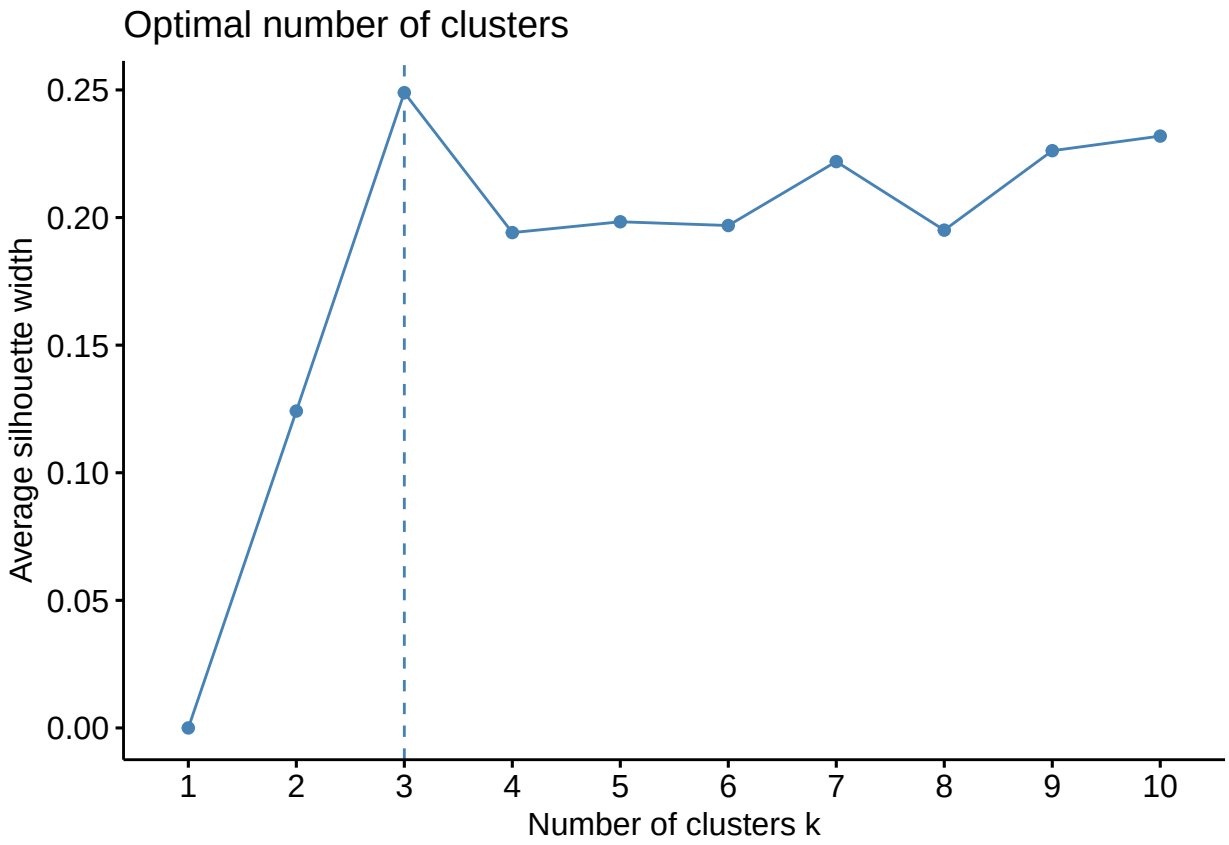
WSS Method

```
library(factoextra)
fviz_nbclust(sk1, kmeans, method = "wss")
```



Silhoutte Graph

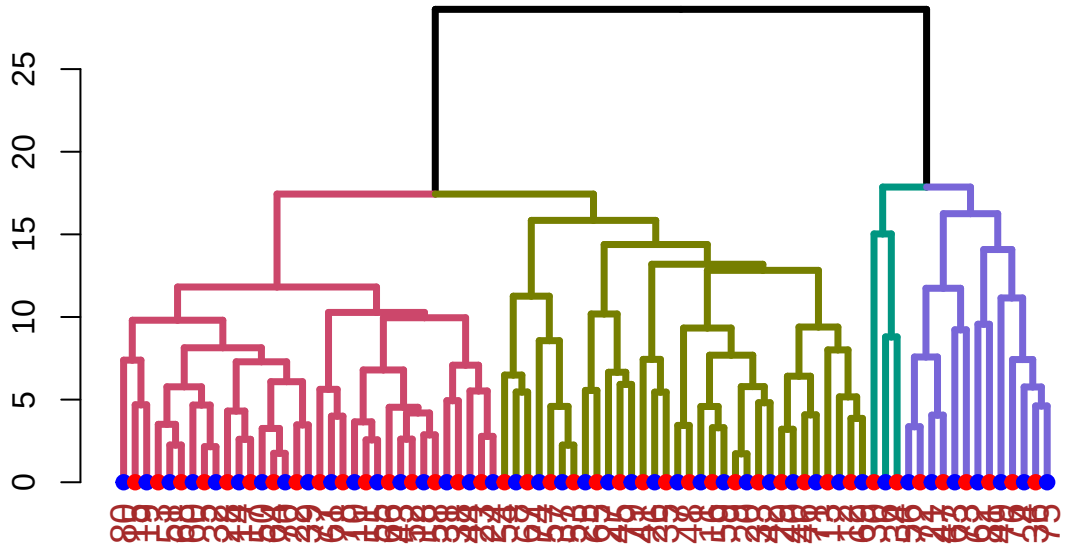
```
fviz_nbclust(sk1, kmeans, method = "silhouette")
```



number of clusters is determined 3 based Silhoutte graph. 3 or 4 looks like a good choice for number of clusters based on wss and silhoutte graph.

2.4.3 Hierarchical Clustering

```
library(magrittr)
library(ggdendro)
library(dendextend)

dend <- distance %>% dist %>% hclust %>% as.dendrogram %>% set("branches_k_color",
plot(dend)
```



its seen number of 4 clusters is sufficient for hierarchical clustering.

Number of members in each cluster.

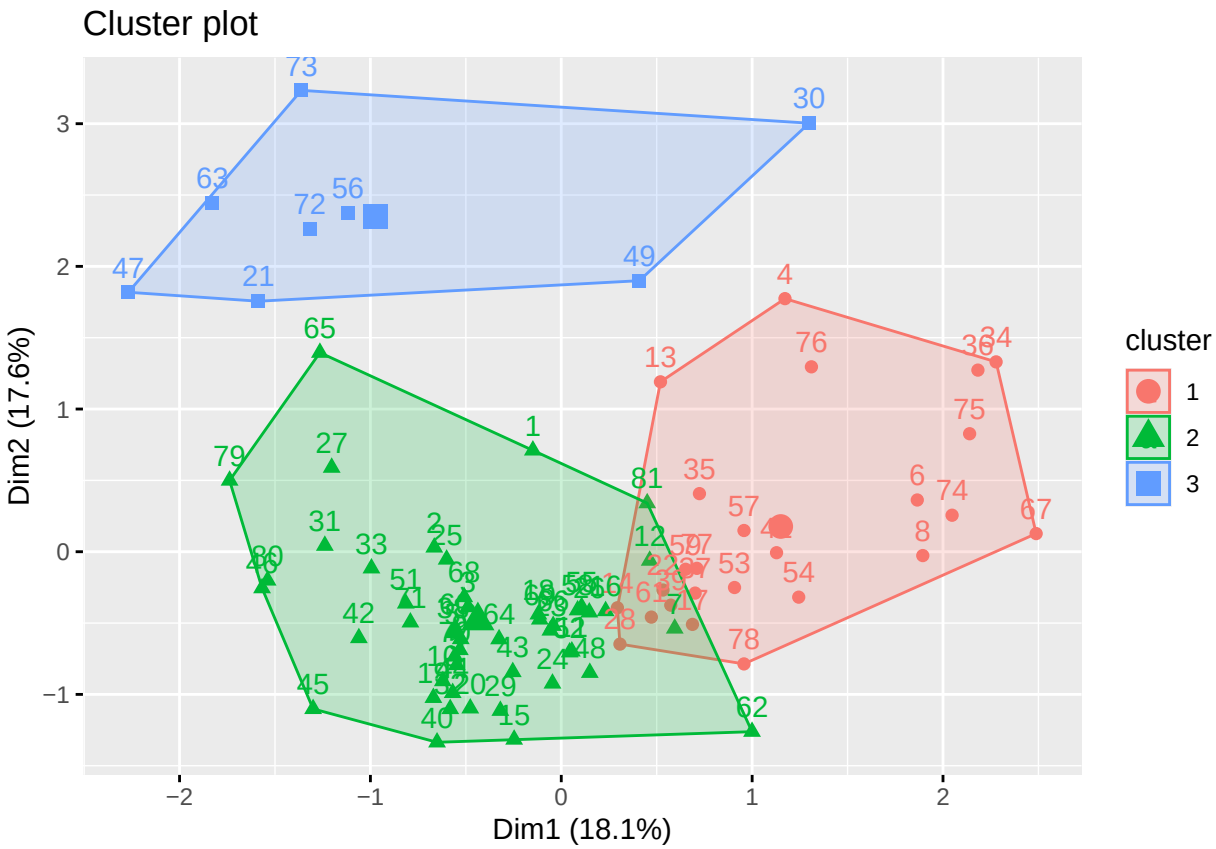
```
member_hc_c = cutree(dend,4)
table(member_hc_c)
```

```
member_hc_c
 1  2  3  4
32 33 13  3
```

2.4.4 K-Means Clustering

K-Means is an unsupervised learning clustering algorithm.K represent the number of clusters and its a hyperparameter that should be determined before the application of method. There are K clusters and center of every cluster is the means of members of that cluster and that is why it's called K-means

```
kmeans.re <- kmeans(sk1, centers = 3, nstart=25)
library(factoextra)
fviz_cluster(kmeans.re, data = sk1,ellipse=T,ellipse.type="convex",shwo.clust.cent=T,poi
```



it is seen that with number of 3 clusters KMeans gives a good result.

Frequencies of member on each cluster

```
members = kmeans.re$cluster
table(members)
```

```
members
 1  2  3
25 48  8
```

Cluster Centers

```
centers = kmeans.re$centers
centers
```

```
  High Life Satisfaction Middle Income Satisfaction of Public Services
1            0.324840222   -0.09738261                      -0.005055839
2            0.006549512    0.11443326                       0.184889644
3           -1.054422769   -0.38227892                      -1.093538368
  Employment Rate Rural Life Education Level
```

```
1       0.2335200  0.8541586      -0.8580362
2       0.2225889 -0.3610304       0.5424959
3      -2.0652834 -0.5030634      -0.5736120
```

Conclusion

A preliminary study was carried out because there are too many variables and metric differences in the data set we have.It has been determined that there is a multicollinearity problem between the variables in the data set.For this reason, it was deemed appropriate to use the Principal Component Analysis method first.In this analysis, it has been investigated how many variables can be reduced due to the large number of variables.By applying principal component analysis, 41 variables and outputs that can be reduced to 7 variables were found. With these 7 variables, 77% of the variance in the data set can be explained.

Factor analysis was carried out in line with the determined 7 variables. Meaningful naming was done by looking at the loads of the factors. Discriminant analysis was performed with the newly created variables.In this analysis, a threshold value was obtained by determining the target variable. According to the values below and above this threshold value, binary classifications were made as high sociability and happiness 1, and low sociability and unhappiness 0.First of all, the assumptions of the discriminant analysis were checked. In the assumptions, the variables do not come from multivariate normal distributions. It was determined that the mean between groups was not different from each other by Wilk's Lambda Test. With Wilk's Lambda test statistic, it was concluded that the canonical discriminant function is not significant. There was no homogeneity of variance between the groups, which is the basic assumption of the discriminant analysis.The explanation rate of the total variance occurring in the group variable of the canonical correlation was determined as 14%. This value was found to be very low in terms of classification. As a result, the positive success rate (TRUE POSITIVE RATE or SENSIVITY) of the discriminant analysis was determined as %62.74.The negative success rate (TRUE NEGATIVE RATE or SPECIFITY) of the discriminant analysis was determined as %63.4.
The clustering technique, which is based on the experimental unit, was also applied in the project. Firstly, it started with the determination of the number of clusters. WSS and Silhouette methods were used. It was agreed to create 4 clusters for Hierarchical clustering and 3 clusters for KMeans using these methods. Then Hierarchical clustering and K-means clustering Cluster analysis was carried out using methods. In line with these clustering methods, our data set showed good performance.

In line with this study, gains were obtained for the multivariate statistics lecture by

using principal component analysis, factor analysis, discriminany analysis and cluster analysis methods. Since the assumptions and prerequisites suitable for the analysis cannot be provided and good outputs cannot be obtained,In line with this study, quadratic discriminant or (non-linear methods) logistic regression should be tried instead of linear discriminant.

References

10 Zelterman, D. (2015). *Applied multivariate statistics with r*. Springer.