**ST30005 Multivariate Analysis**

**Assignment 2**

**Answer 1:**

1. To create a covariance matrix in R first declare the variables which are x1,x2,x3,x4,x5,x6 then run the following command;

x1 <- c(3266.46,1343.97,731.54,1175.50,162.68,238.37)

x2 <- c(1343.97,721.91,324.25,537.35,80.17,117.73)

x3 <- c(731.54,324.25,179.28,281.17,39.15,56.80)

x4 <- c(1175.50,537.35,281.17,474.98,63.73,94.85)

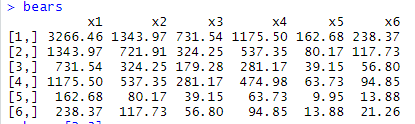
x5 <- c(162.68,80.17,39.15,63.73,9.95,13.88)

x6 <- c(238.37,117.73,56.80,94.85,13.88,21.26)

after that to create the matrix simply run the following code;

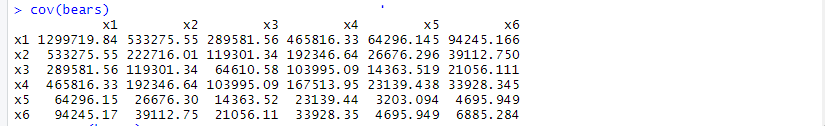
bears <- cbind(x1,x2,x3,x4,x5,x6)

and the matrix will be created which looks like below;



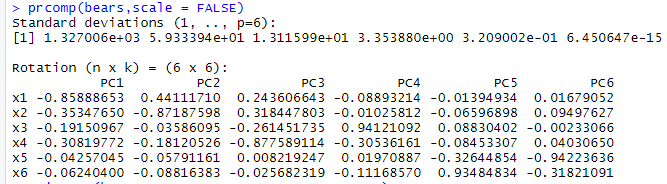
Now if we run the command bears[2,3] and bears[3,2] respectively, the first one tells the result of 2nd row and 3rd column which is 324.25 on the other hand if we run the second command the result will be the element of 3rd row and 2nd column which is 324.25, we know that values are same but they are still different because the second row i.e. x2 represents body length(cm) while 3rd row that is x3 represents neck circumference(cm) so we can say that only the values are same but they presents two very different variables.

1. To verify that “bears” is a covariance matrix we simply need to run the command cov(bears) and it will give the following matrix which is the covariance matrix of the given matrix bears.

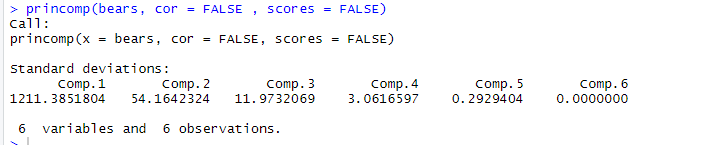


Hence we can say that “bears” is the covariance matrix.

1. To perform the principal component analysis we have to run the ‘prcomp’ or ‘princomp’ and it will give the principal component analysis of the given matrix.



Or we can run the other command which is ‘princomp’ it will give the following output.



The above command gives the components for the all the 6 variable and

6 observations. So, the data can be effectively summarized in the 6

dimensions.

1. These correlations are obtained using the correlation procedure. In the variable statement we include the first three principal components, "pc1, pc2, and pc3", in addition to all six of the original variables. We use the correlations between the principal components and the original variables to interpret these principal components.

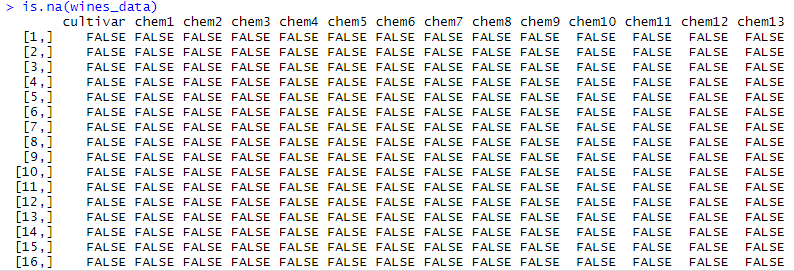
**First Principal Component Analysis:**

The first principal component is strongly correlated with five of the original variables. The first principal component increases with increasing weight, body length and neck circumference. This suggests that all six criteria vary together. If one increases, then the remaining ones tend to increase as well. In fact, we could state that based on the correlation of -0.85888653 that this principal component is primarily a measure of the weight and length.

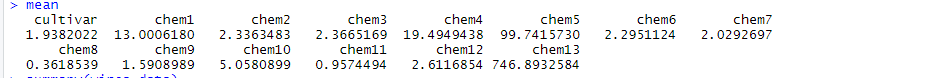
1. In the principal component analysis we can clearly see the rising trend of the figures which is from -0.85888653 to -0.06240400 so we can say it is tend to increase the values of pc1 if one value increases and that’s why the most influential figure in first principal component analysis is the value which is nearest to 0 value or tending to be positive i.e. the value of x6 -0.06240400 while the least influential value is the very first value which the value of x1 i.e. -0.85888653 as it is the value which is a least value tending to be positive or 0. As we can see the values in pc1 from x1 to x6 values vary from each other but they tend to increase so these two can be the most and the least influential values of pc1 respectively.

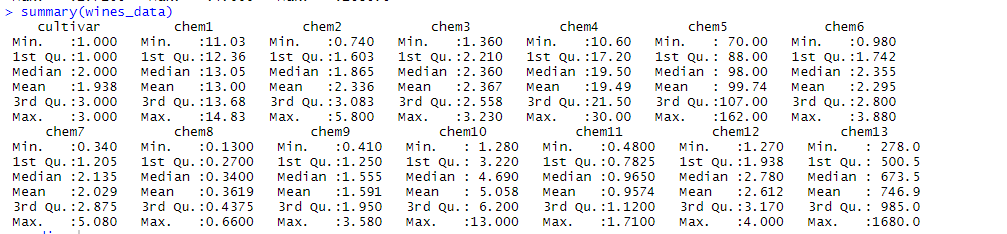
**Answer 2:**

1. The dataset is taken from the .CSV file which wines.csv this file contains the data of wines with the cultivation variables and the mixtures indicated by chem1 to chem13. To see the missing values we will use the command is.na(). The missing values will be shown as true or false.



1. To check the mean, median, mode and range of the dataset we will run the commands like colmeans, colmedian, colmode and colrange to know the values of the rows that means mean of columns, median of the columns, and finally the range of the columns. Thus we can find all the figures. The output will look like below;







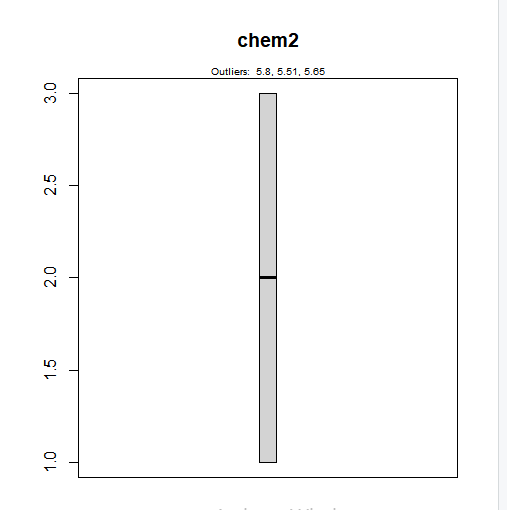
Hence from the above results we can easily find the mean, median, mode and range of the particular dataset.

1. To detect the outliers in the dataset we will just run the following command and the plot with the ratios of cultivation variable and chem 1,2,3… will be produced with the numbers of outliers.

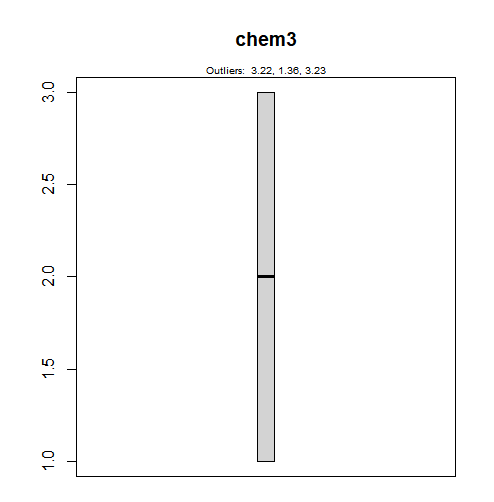
**outlier\_values <- boxplot.stats(wines\_data$chem2)$out**

**boxplot(wines\_data$cultivar, main= "chem2", boxwex=0.1)**

**mtext(paste("Outliers: ", paste(outlier\_values, collapse=", ")), cex=0.6)**

****

In the above plot we can clearly see the numbers of outliers in the ratio of cultivation variable and chem2 plot, similarly for chem3 the graph look like below;



1. To identify the potential multivariate outliers we will run the mahalnobis distance measure, this distance measure prints the number of outliers contained by the dataset which will effects the result of the dataset.

To run the mahalnobis distance run the following command;

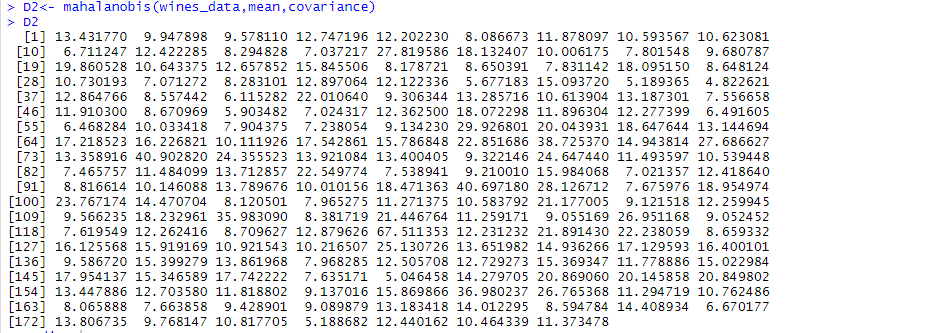
**covariance <- cov(wines\_data)**

**covariance**

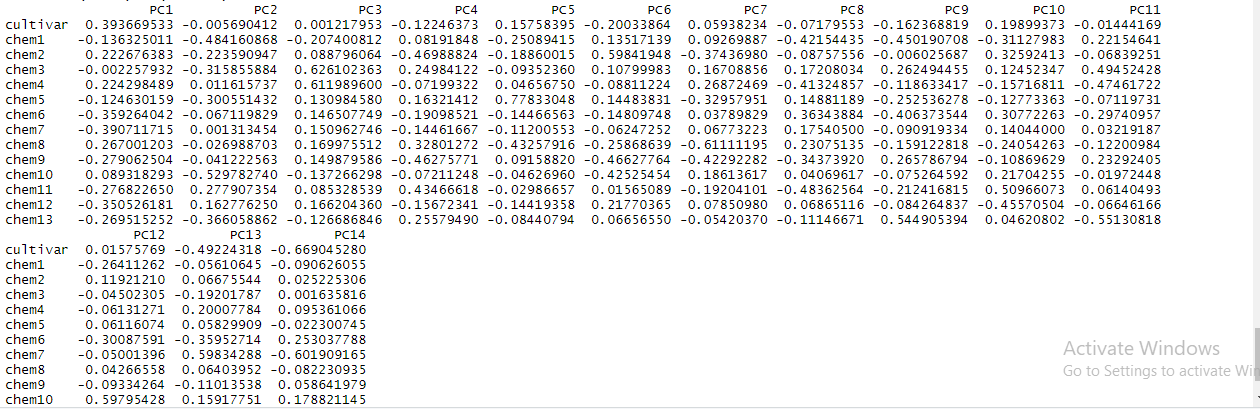
**D2<- mahalanobis(wines\_data,mean,covariance)**

**D2**

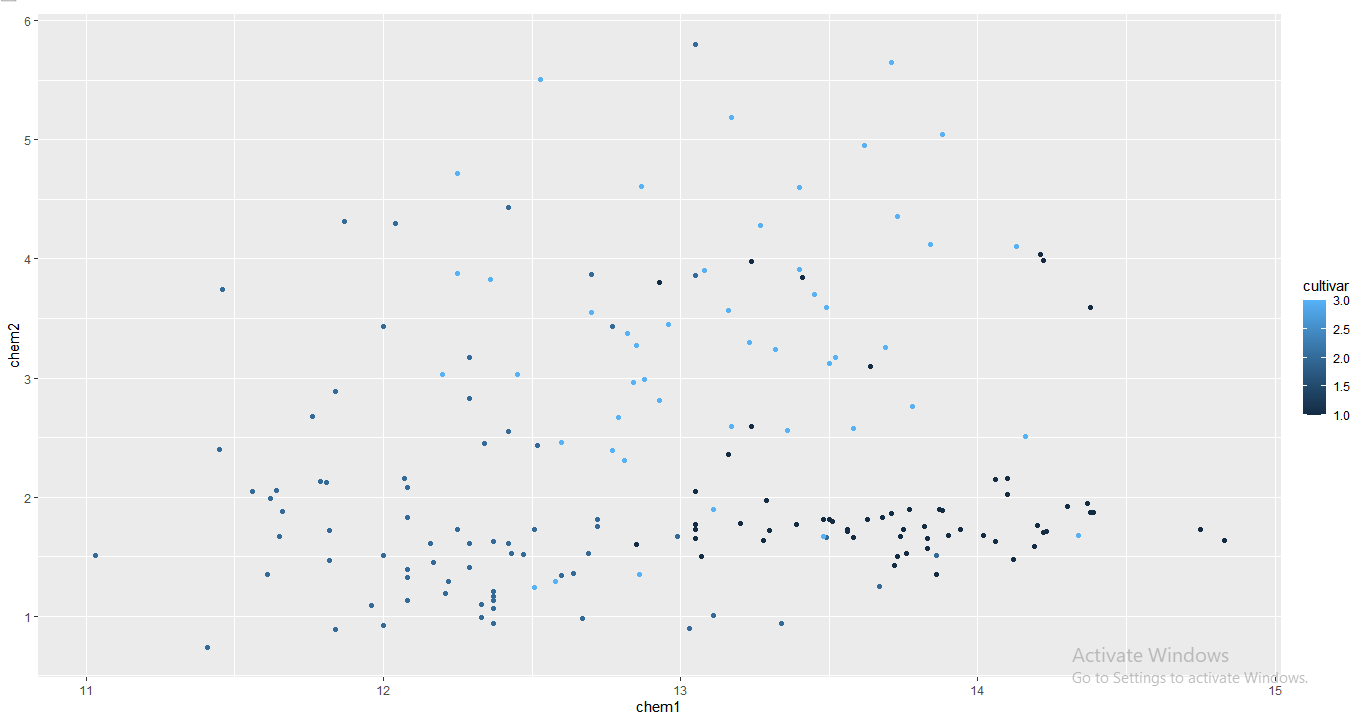
The output will look like below;

****

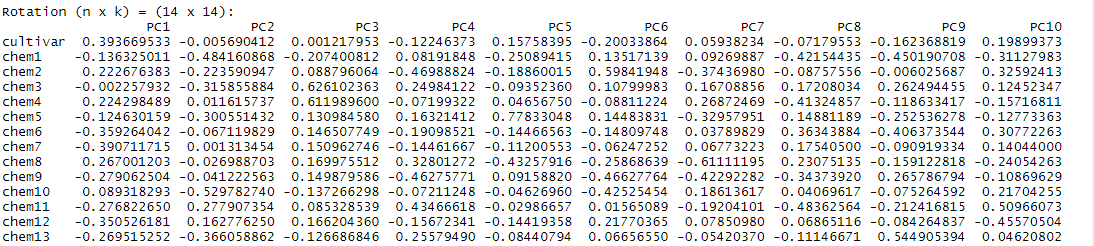
1. The scatterplot of matrix of principal component analysis shows the outliers and the principal components which will affect the dataset and results, principal components are denoted as PC1, PC2 etc The output for the pca look like below;

****

The scatterplot for the matrix will look like below;



1. The outliers removal process applies the PCA method to check whether there is any further outliers remained in the dataset and if there is any we should remove it immediately. The PC’s which are remained after removal of outliers is PC1 to PC10 with their eigen values in the form of matrix.



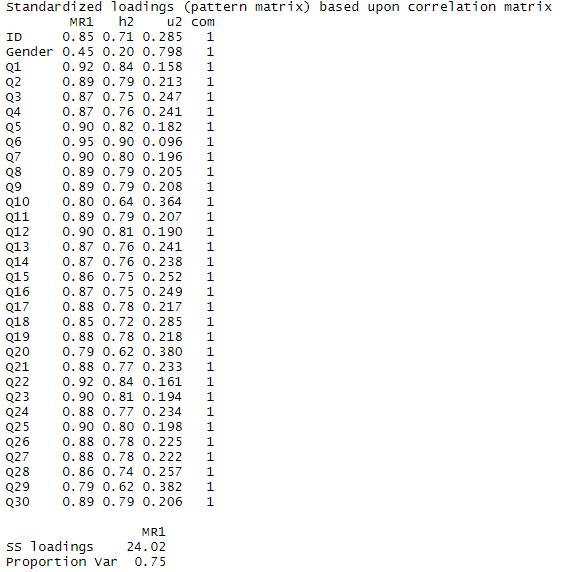
1. The second largest eigen value in the principal component analysis is 0.77833048 which is the part of PC5, it indicates that most number of outliers and largest outliers were found in PC5’s chem4 analysis.
2. The retained principal component looks like this in the form of equation;

H u(x) = E u(x) where x=cultivar i.e. cultivation variable.

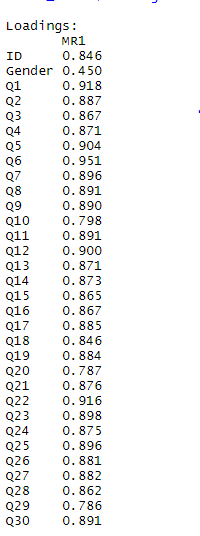
1. The first two principal components shows that the components vary on the large scale of positive and negative differences which shows the variability of the cultivar to chem1…14.

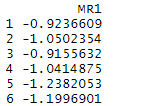
**Answer 3:**

1. The dataset named as household2.csv contains the series of questionnaire asked to different group of people for their living conditions and whether they’re satisfied with the situations or not. The gender is not included in the analysis because in the dataset the gender and ids are almost same and the system can not differentiate whether it is male or female that’s why we exclude the gender.
2. The principal component analysis (PCA) is used to analyze the outliers using univariate and multivariate outliers analysis method while the exploratory factors analysis used to analyze the factors to see what kind of factors effects the dataset and in which way. The EFA analysis of the dataset looks like below;

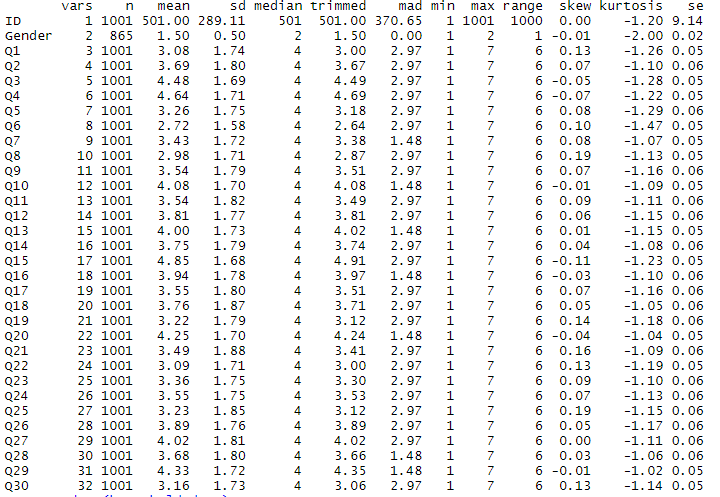
****

1. Variables that are included in the EFA are valid according to the analysis, the analysis below shows that the description of the data is valid for the give dataset.

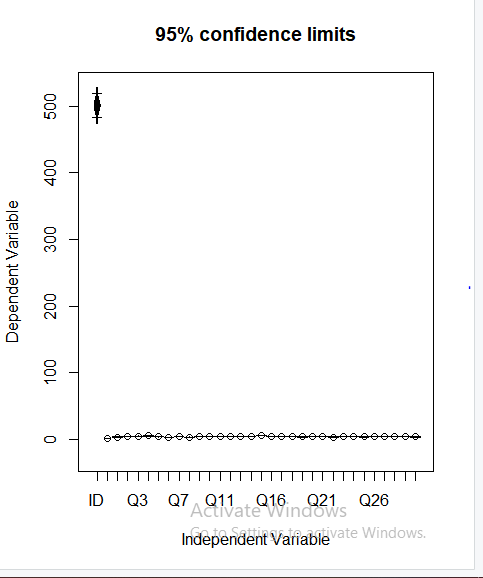
****

****

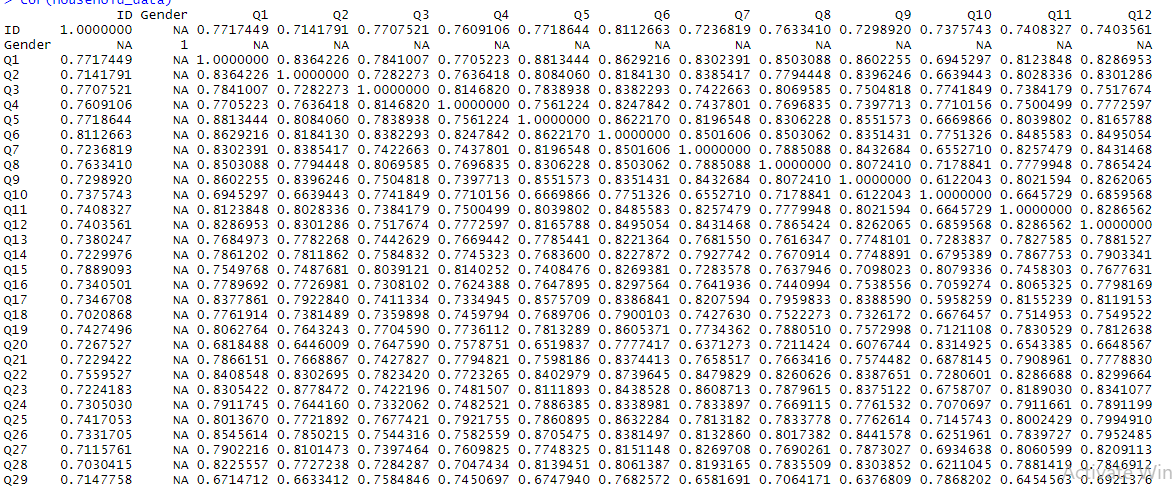
1. As we know that the outliers are treated in the dataset and now only figures which can show the precise result are the part of dataset and the factors which need to work on are standard deviation, variance, kurtosis of the analysis of the graph and hence the following analysis will show the precise results of the dataset.

****

The following plot shows the error in the dataset and needed to resolve.

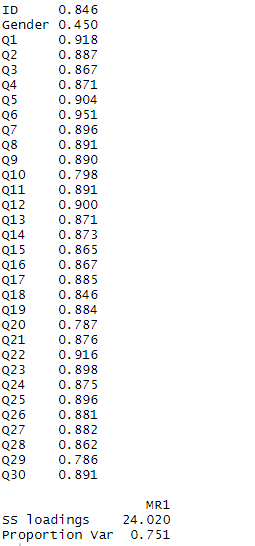


1. The correlation matrix created from EFA analysis looks like below;

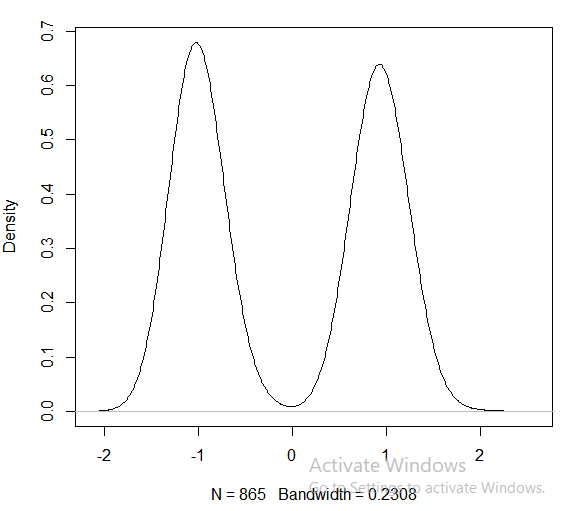


According to the above analysis we can see that factors which are needed to be investigated are the missing values in the matrix, the figures below the diagonal elements which looks unbalanced or which can distort the result of the whole analysis needs to be investigated.

1. The eigenvalue of the matrix are related to the rank of the matrix and in the above matrix the second largest eigen value is 0.8813444. As we can see the largest value is 1.00000 which is the diagonal element of the dataset so the second largest value above the diagonal is 0.8813444.
2. The varimax rotation isn’t applicable on the missing values of the dataset hence it didn’t work on the correlation matrix which is mentioned above but it worked on factor analysis model and the ratio of EFA and load factor. The analysis look like below;



1. If we compare the un-rotated solution of part (e) with the rotated solution which have used the varimax rotation we can see that the solution which came from the varimax rotation is easy for the analysis while the solution in part (e) is actually untreated and contains many missing values which makes it difficult to analyze.
2. According to the answer in the previous question the next part of the analysis is to visualize the result which means to perform the visualization for the particular analysis and by visualization, it will make clear about the ups and downs of the analysis the smooth result will be shown. The visualization of the above analysis looks like below.



1. While using PCA we have to find the covariance of the data apart from that while using PCA method we have to call several libraries in the R packages and it much complex, on the other hand EFA is much simpler than PCA, for EFA one single library includes all its dependencies and its plotting isn’t much complicated like PCA, also EFA is time-efficient than the PCA method, so I would prefer EFA over PCA for analyzing the large dataset.