EDA Final Report

Nitin

July 24, 2019

Exploratory Data Analysis Final Report

Research question

Hypothesis (Ho)

Is grouth in Total_Trade of India is directly relate to GDP (Real_GDP).

Describe data

Dataset:

Dataset used for this project is Government dataset QoG Standard dataset. It is largest dataset with 15403 obs. And 2199 variables for all countries in the world. This data set has over more than hundred data sources. Among available variables of 2199 on data set this project focuses on six variables which are as follows.

"gle_exp" = "Total_Export", "gle_gdp" = "Real_GDP", "gle_imp" = "Total_Import", "gle_pop" = "Population" , "gle_rgdpc" = "Real_GDP_per_Capita", "gle_trade" = "Total_Trade"

Explore data relationships

Run all required library for this project in the beginning of the project.

Read data for this project and store it in an object called Data. From this data read data from row number 5842 to 5914 data related to country of India.

Original data has 15403 obs. And 2199 variables but focus will be on only six variables gle_exp, gle_gdp, gle_imp, gle_pop , gle_rgdpc, gle_trade. Numbers 828,829,830,831,832,833 are corresponding column numbers in excel for data under consideration. Reduced data consist of seventy six observation and six variables.

Rename columns appropriately to make column names more meaningful as follows.

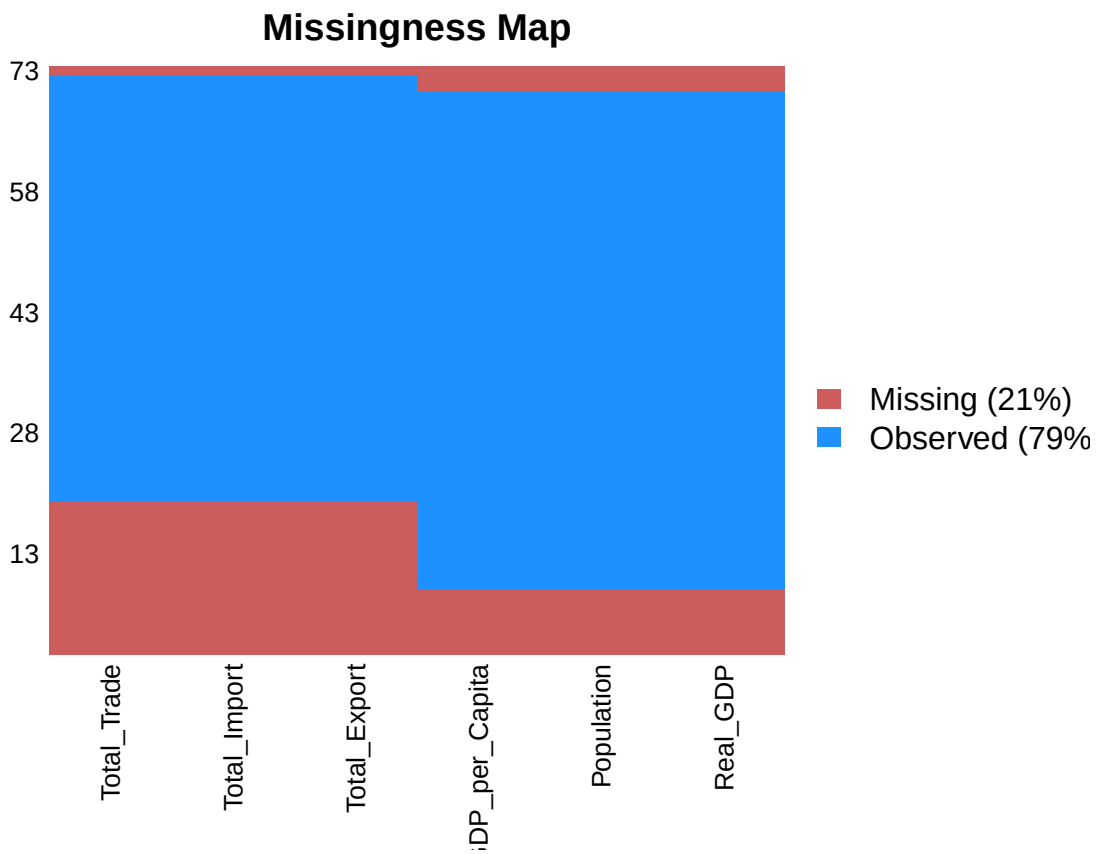
- gle_exp = Total_Export
- gle_gdp = Real_GDP
- gle_imp = Total_Import
- gle_pop = Population
- gle_rgdpc = Real_GDP_per_Capita
- gle_trade = Total_Trade

When calculated for number of rows with at list one missing observation it is observed that twenty two rows have at list one missing value.

When calculated for missing data in percentage per entire dataset it is observed that close to twenty one percent data is missing.

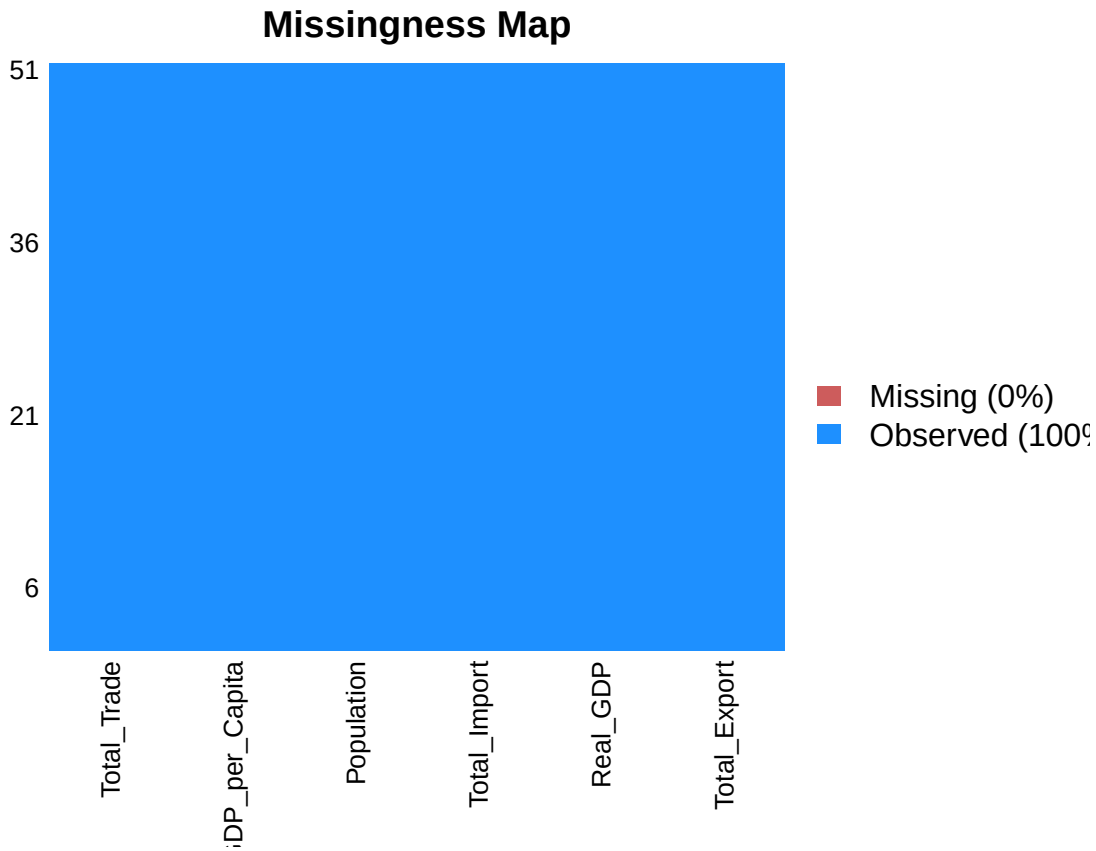
Calculate distribution of missing data by variable.

Figure bellow "Missingness Map" shows pattern of missing data in a data set used for this analysis. From figure below it could be observed that data is Missing Not at Random (MNAR).



From pattern observed for missing data in figure above it is best approach to delete entire rows with missing values.

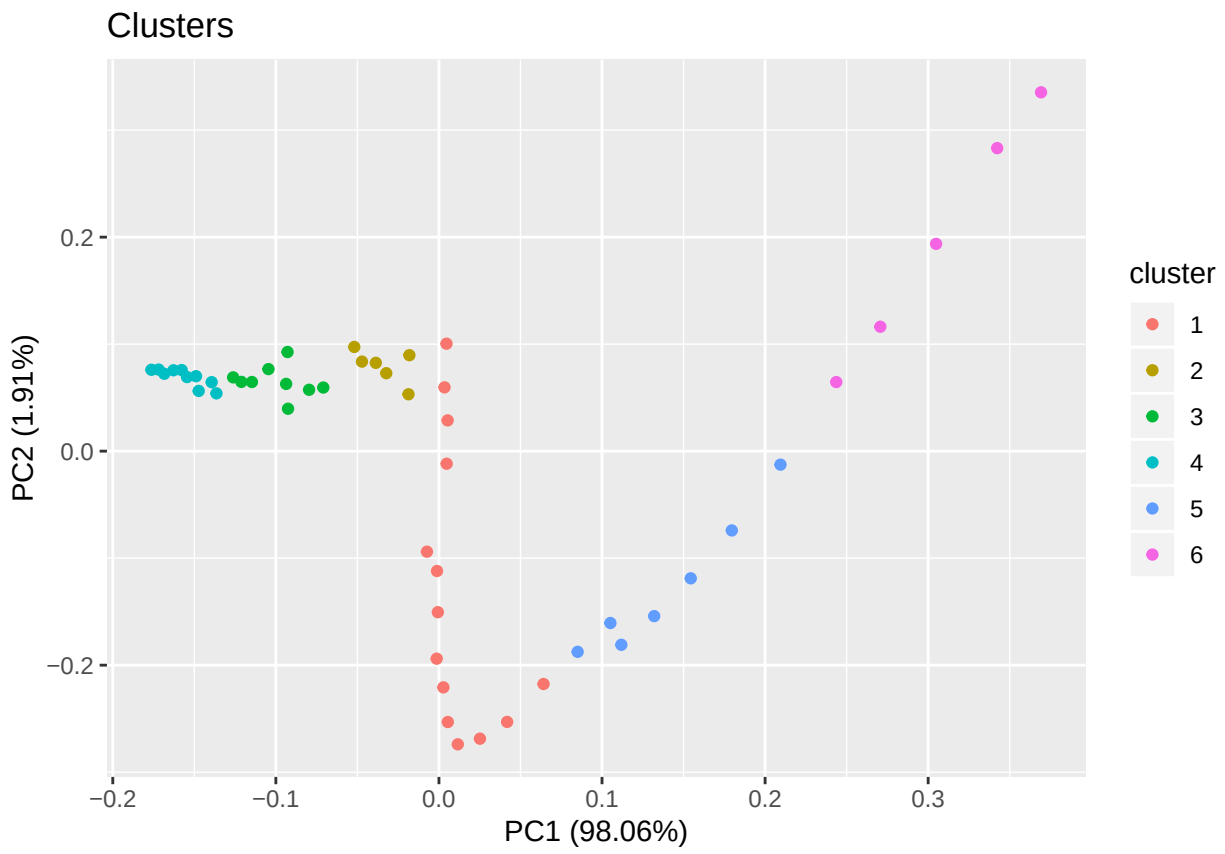
After deleting missing values it could be observed from the figure below that there no any missing data in dataset used for analysis in this project.



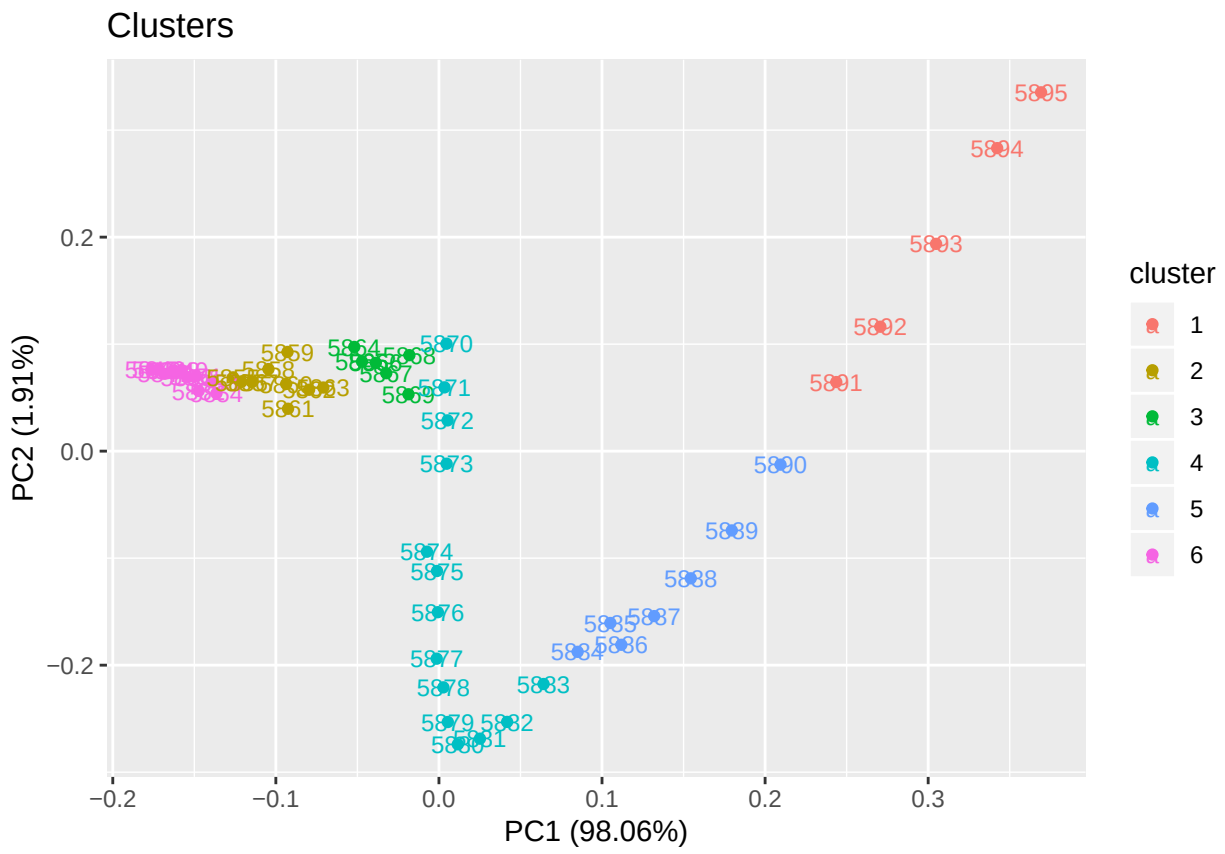
4. Cluster Analysis

###Plotting K-means

Randomize data set using function set.seed(123). Plot graph using autoplot function with number of cluster equal to six as given in the figure below.

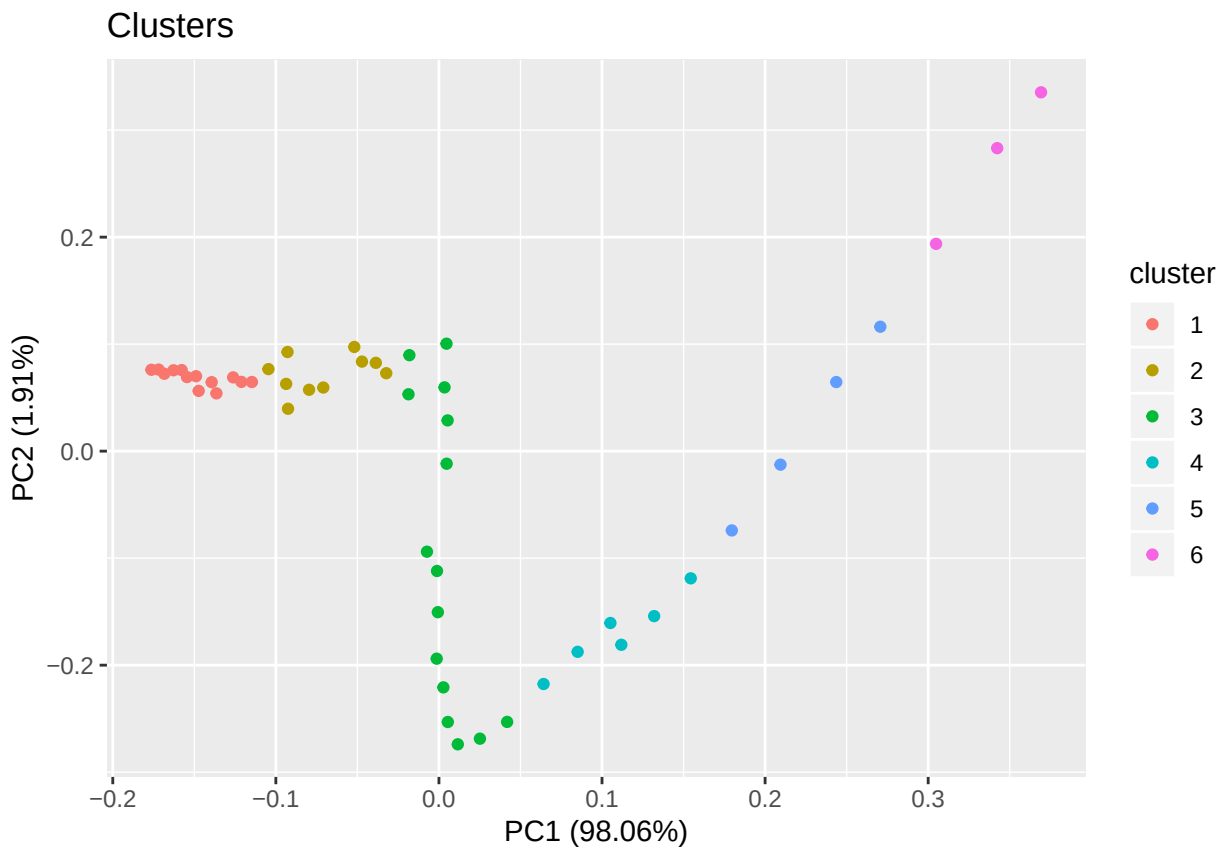


Plot graph using autoplot function with number of cluster equal to six and values for variable as given in the figure below.

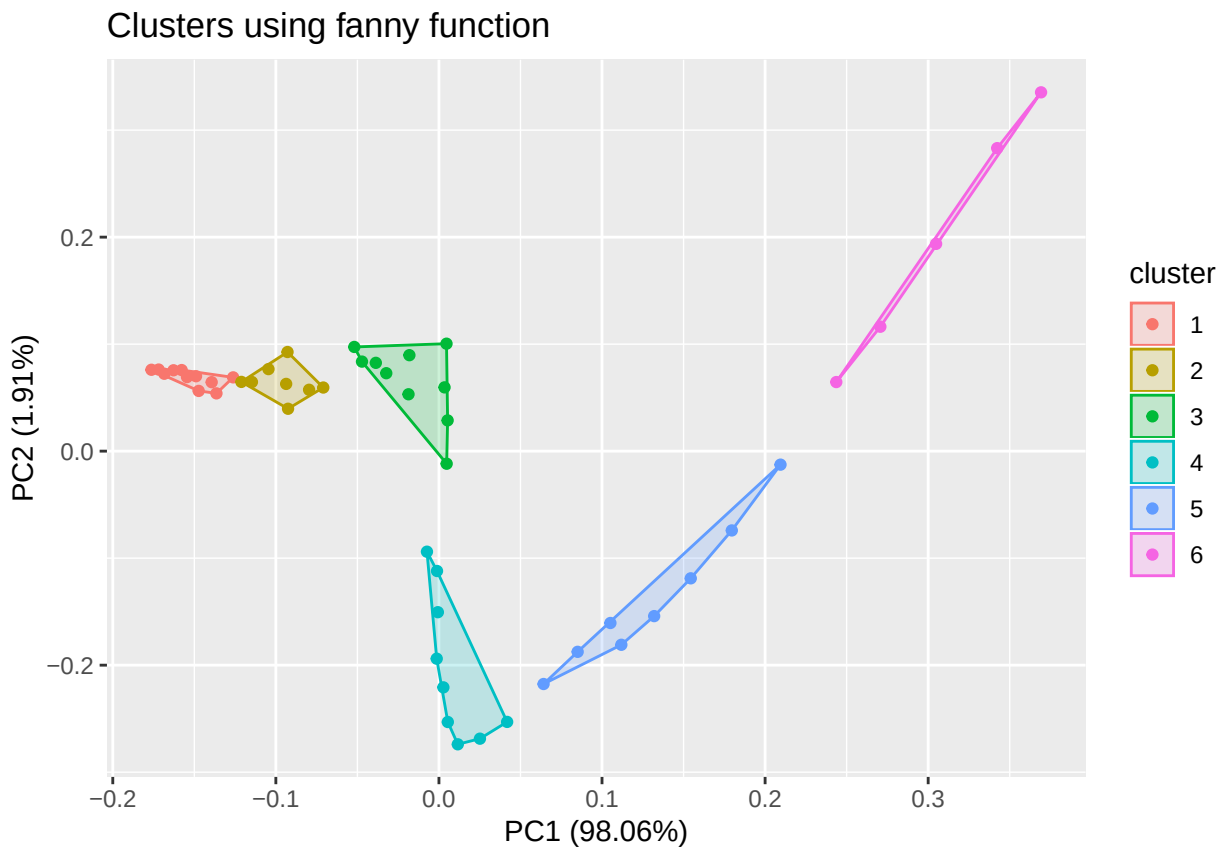


###Plotting cluster package

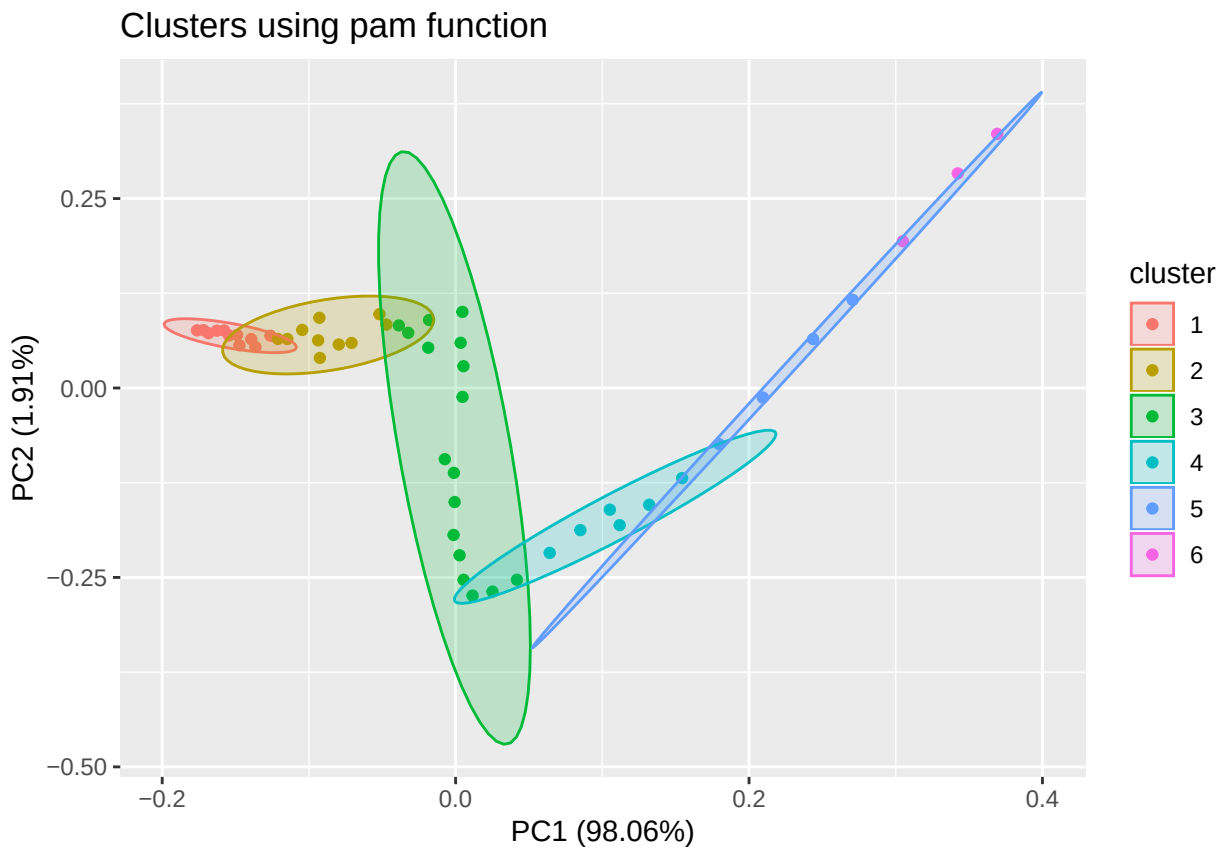
Plot cluster package using autoplot with clara function with number of cluster equal to six as given in the figure below.



Plot cluster using autoplot with fanny function and number of cluster (K) equel to six.



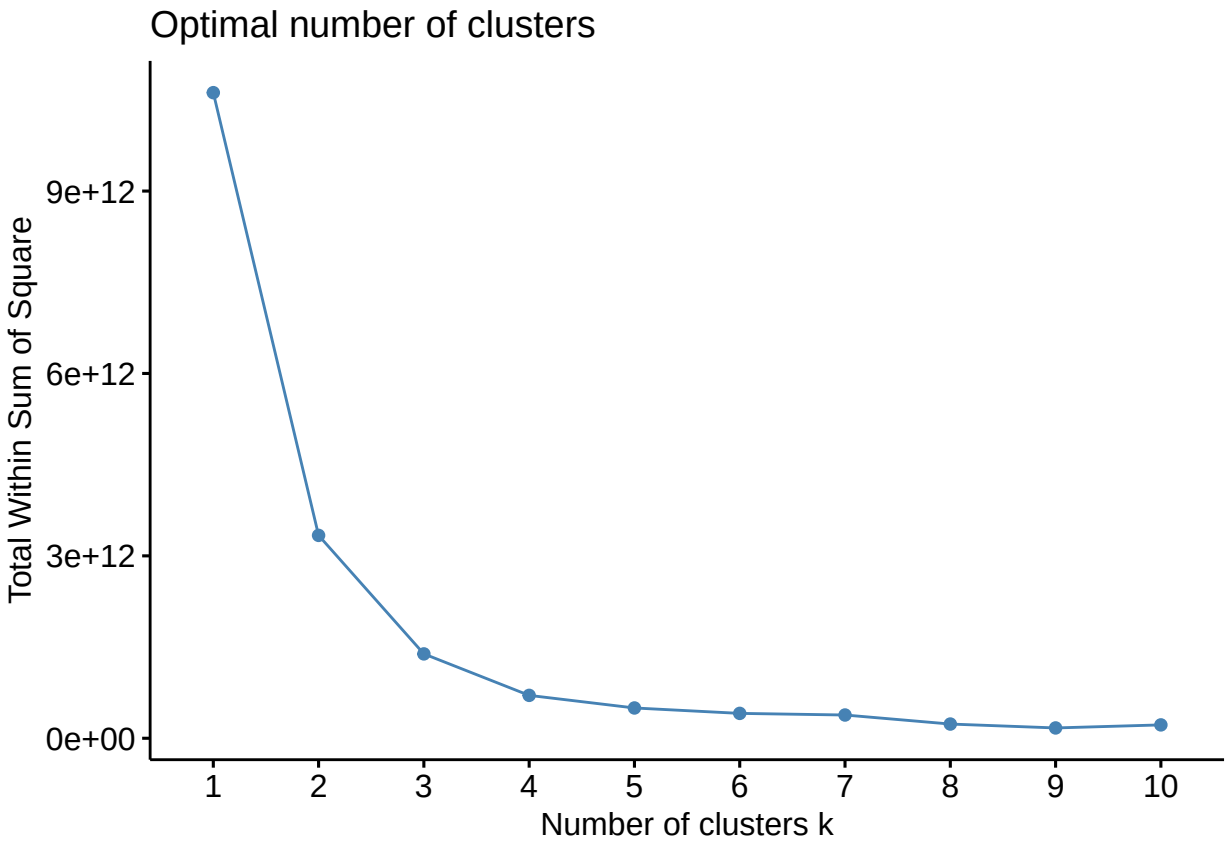
Plot cluster using autoplot with pam function and number of cluster (K) equel to six.



Calculate optimal number of clusters using various methods like Elbow method, Gap Statistic Method etc.

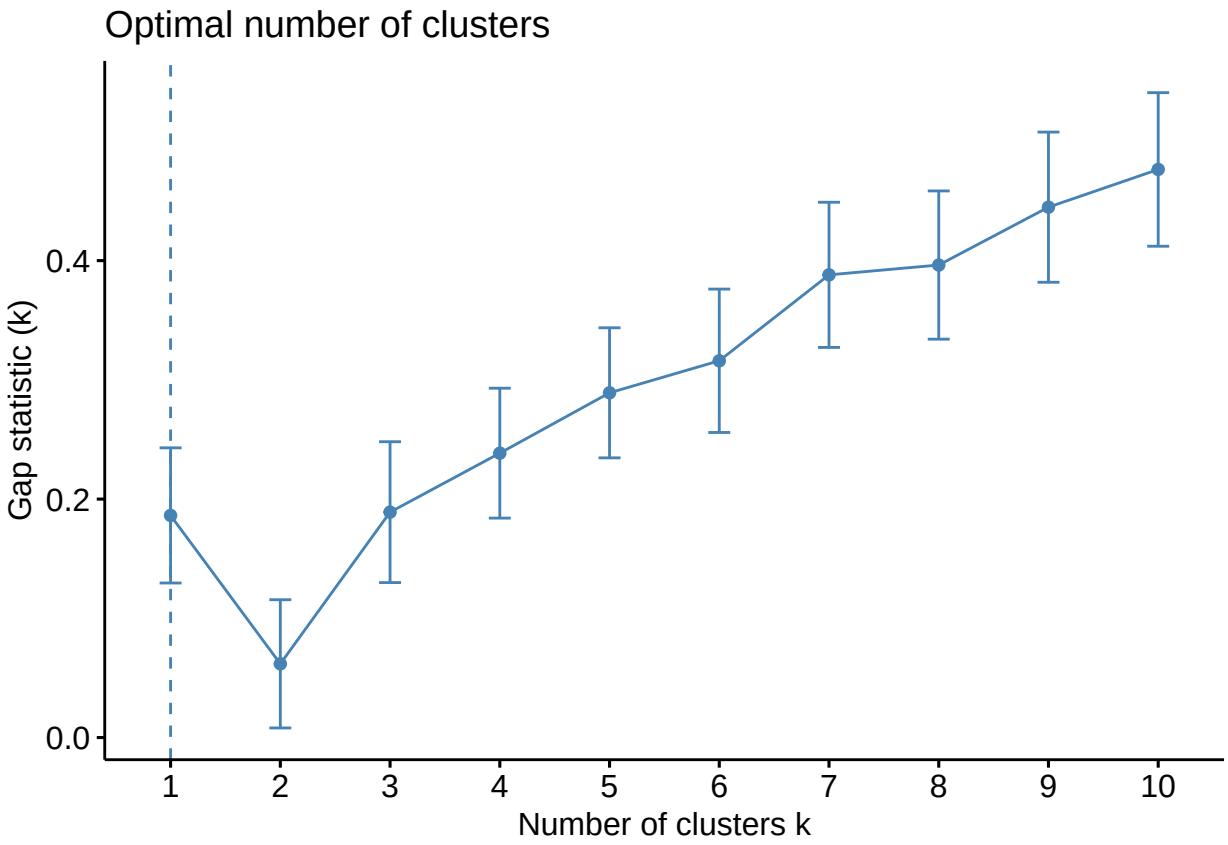
Elbow Method

Figure below showes optimal number of clusters using Elbow method. Elbow method gives obtimal number of cluster as two.



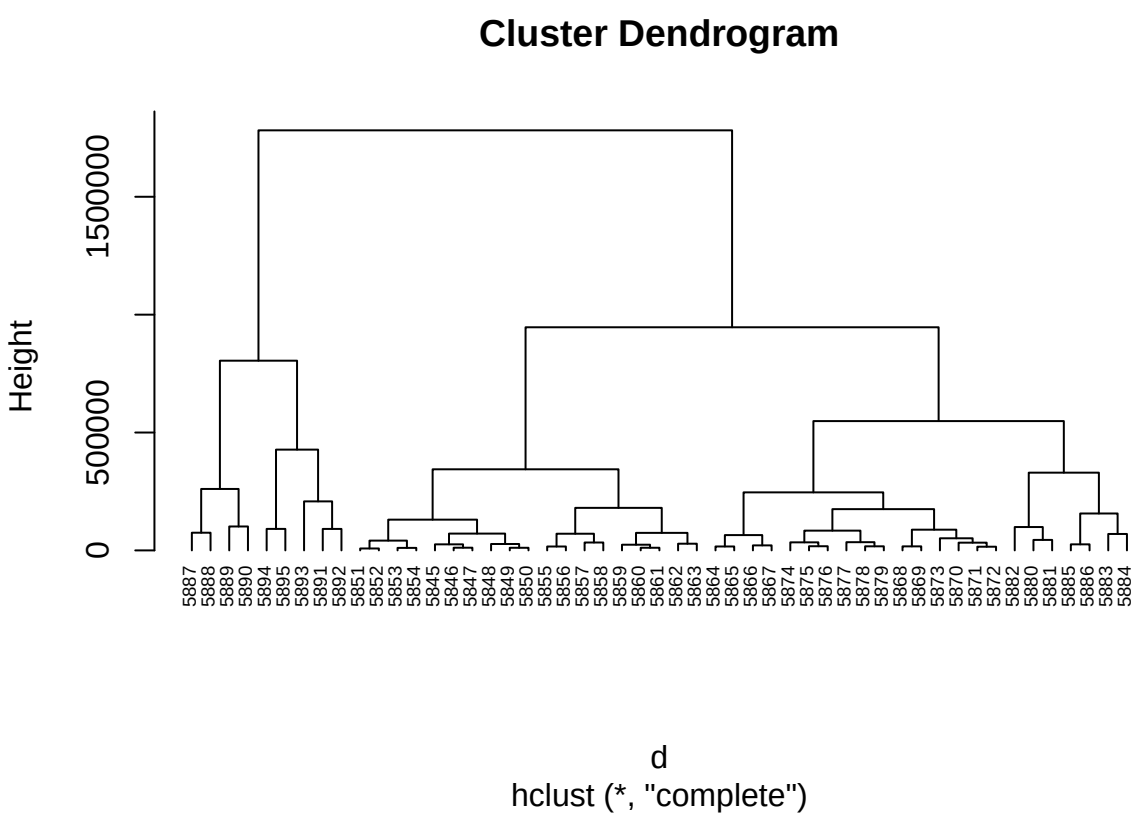
Gap Statistic Method

Figure below showes optimal number of clusters using Gap Statistic Method. Gap Statistic Method gives obtimal number of cluster as one.



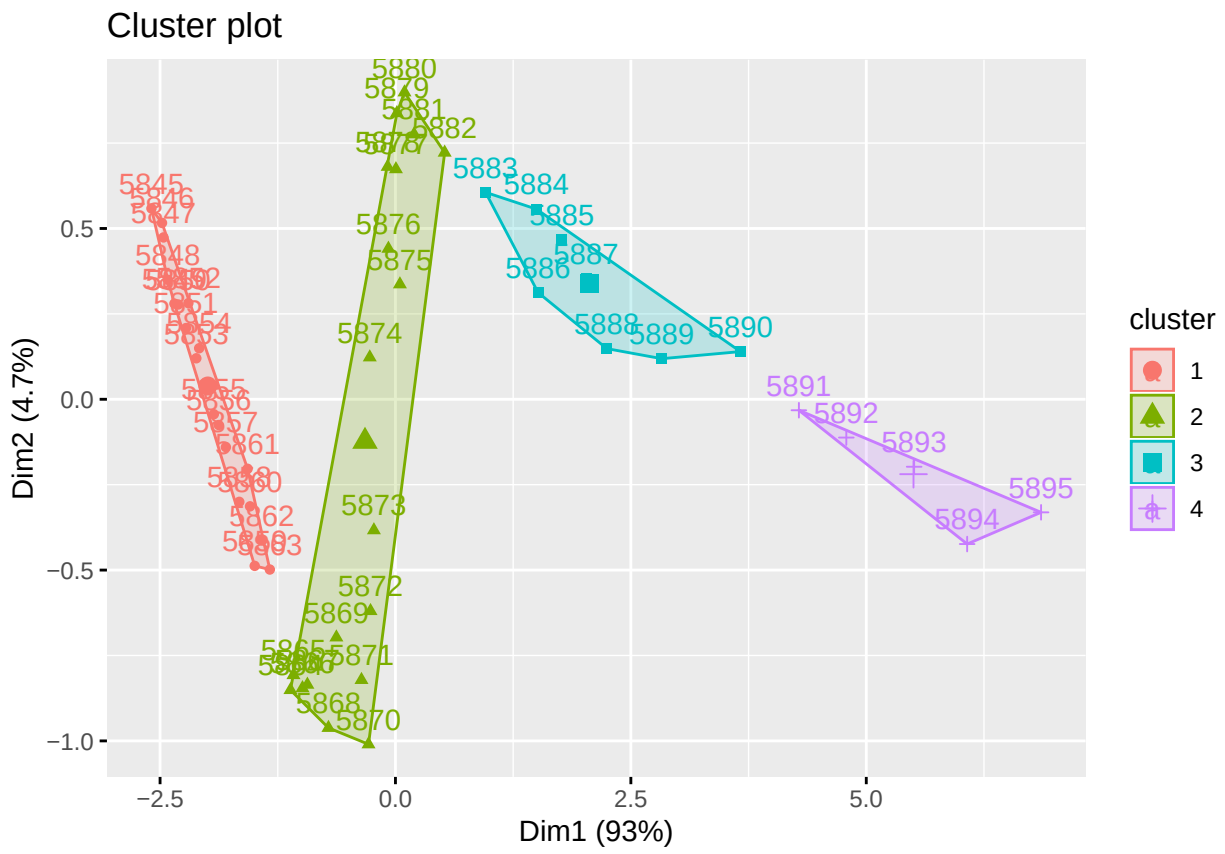
Hierarchical clustering

Figure below showes cluster dendrogram.



Form sub groups for further analysis.

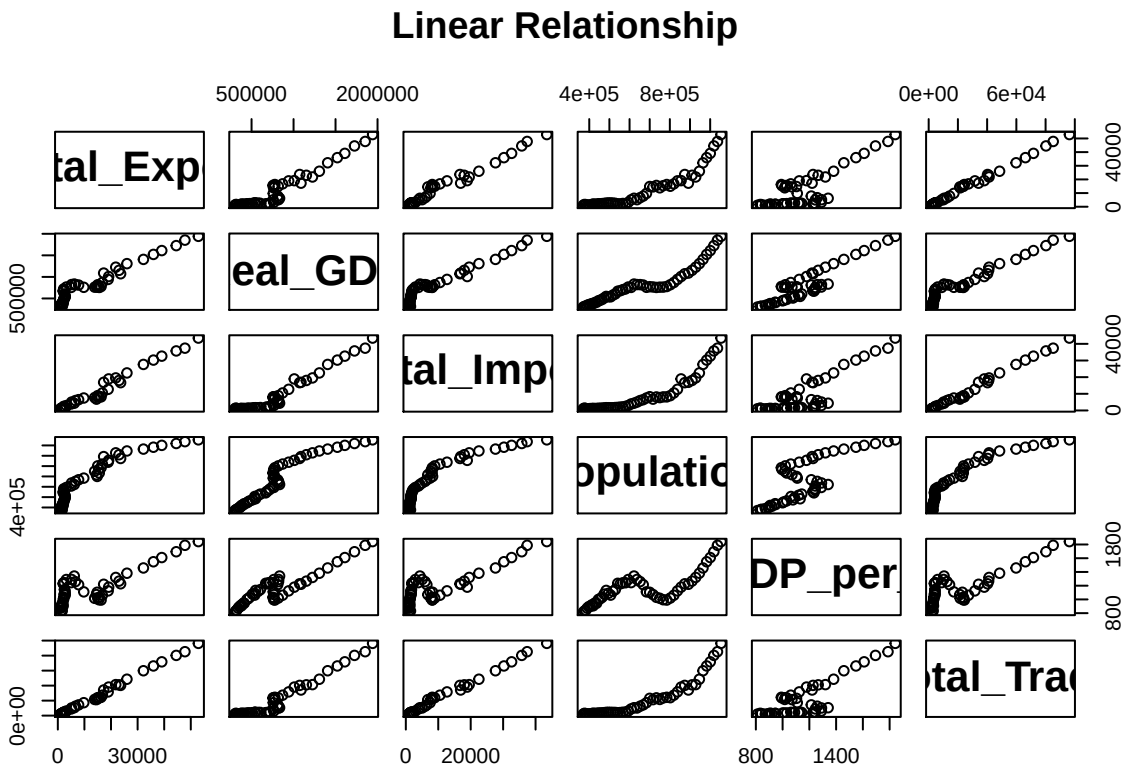
Figure below is ploted with fviz_cluster and list function.



Principal Component Analysis (PCA)

Data Correlation

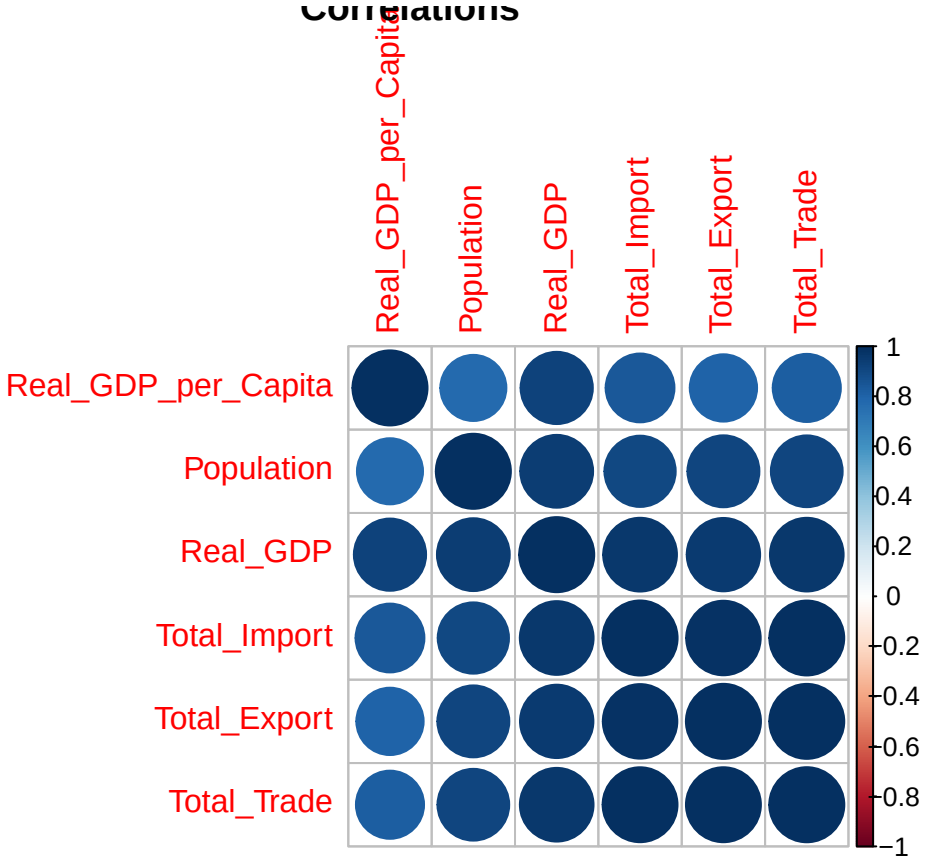
STEP 1: Use pair to create scatterplots and inspect the existence of a linear relationship



Step 2: Pearson correlation

For missing data:

Figure below shows both positive and negative correlation ship of all variables.



STEP 3: Run a principal component analysis

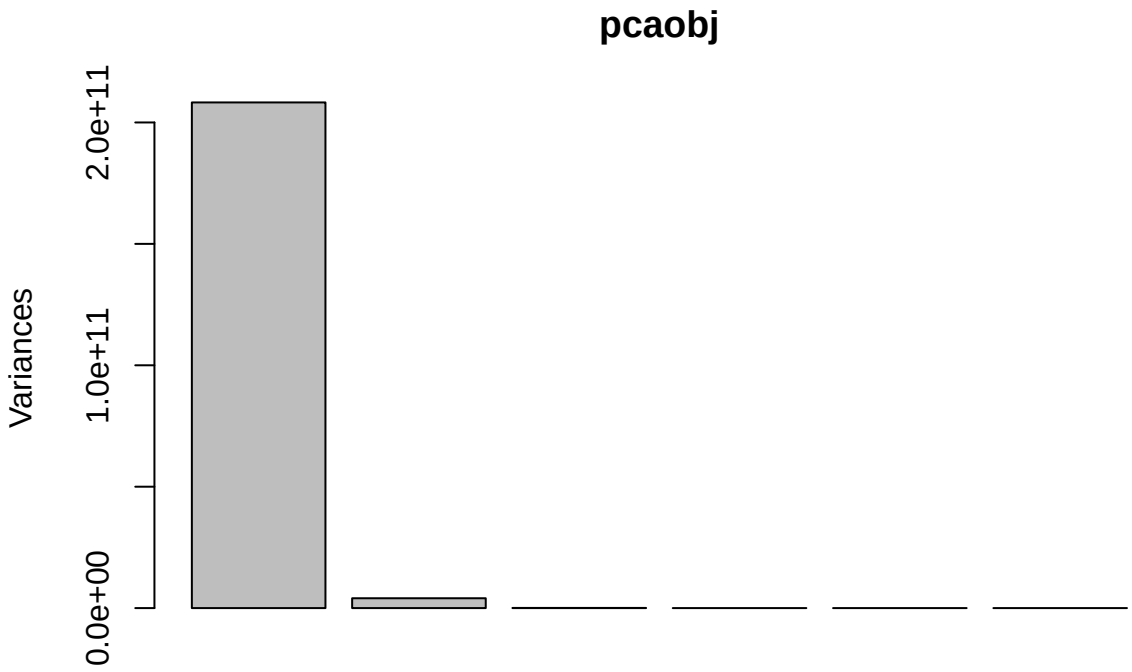
The directions of the new rotated axes are called the eigenvectors of the covariance matrix.

From result displed above equation for PC1 could be as followes.

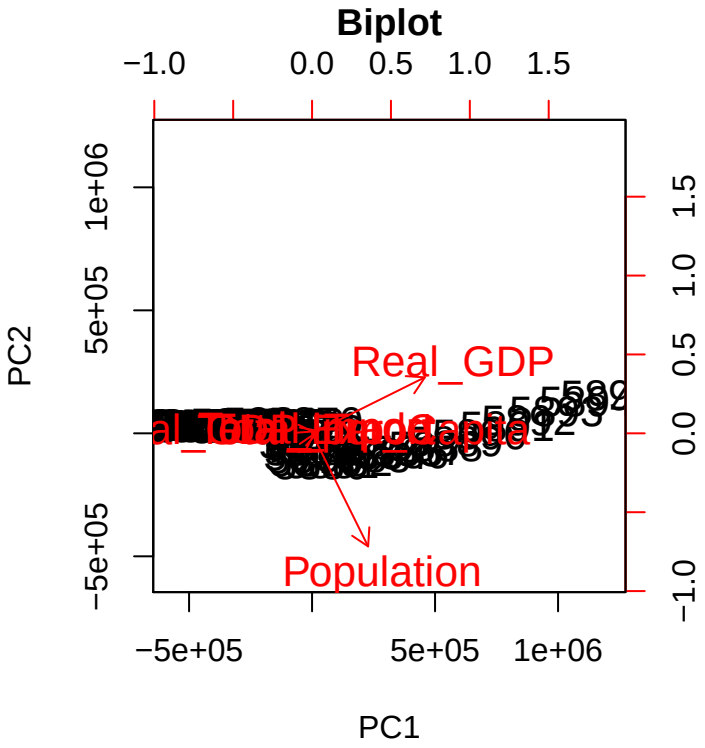
PC1 = 0.028 (Total_Export) + 0.89 (Real_GDP) + 0.023 (Total_Import) + 0.44 (Population) + 0.00 (Real_GDP_per_Capita) + 0.05 (Total_Trade)

From this equation it can be conclude that Real_GDP and Population contribute more than 90 % for PC1 so this variables should be used for analysis. On the other had other variables like Total_Export, Total_Import, Real_GDP_per_Capita, Total_Trade do not contribute much therefore should be eliminated.

STEP 4. We can use a Scree plot to assess Principal Components



From figure below it could be observed that Real_GDP and Population are significant.



From figure below it could be observed that comp.1 is significent.

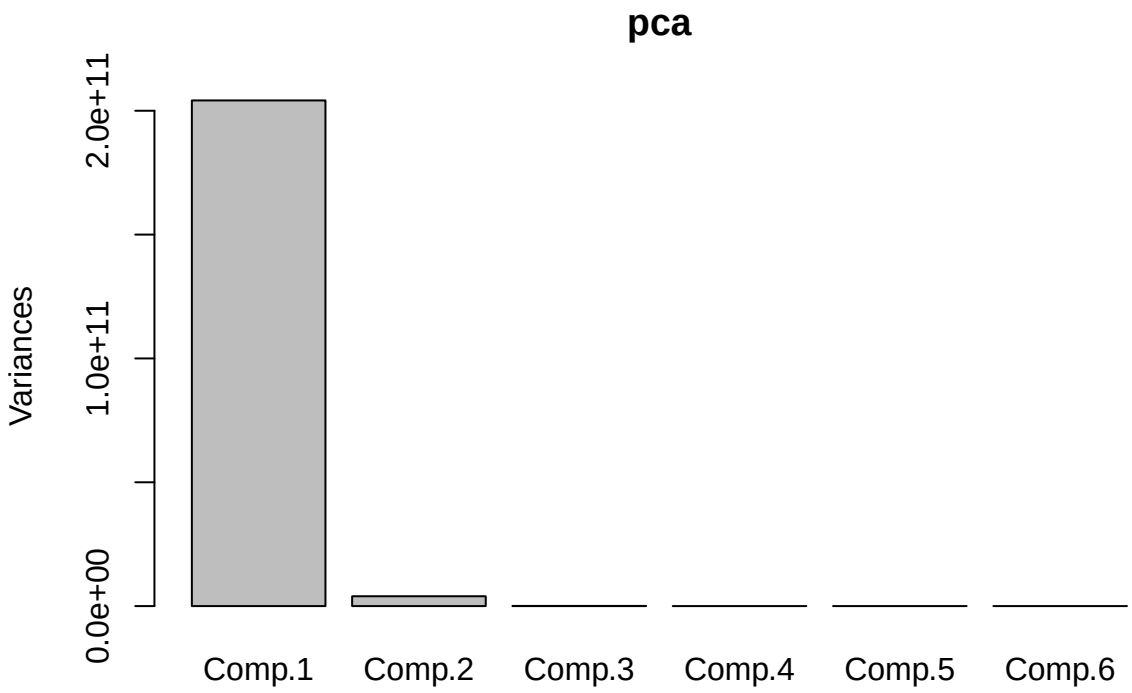


Figure below using autoplot and prcomp function.

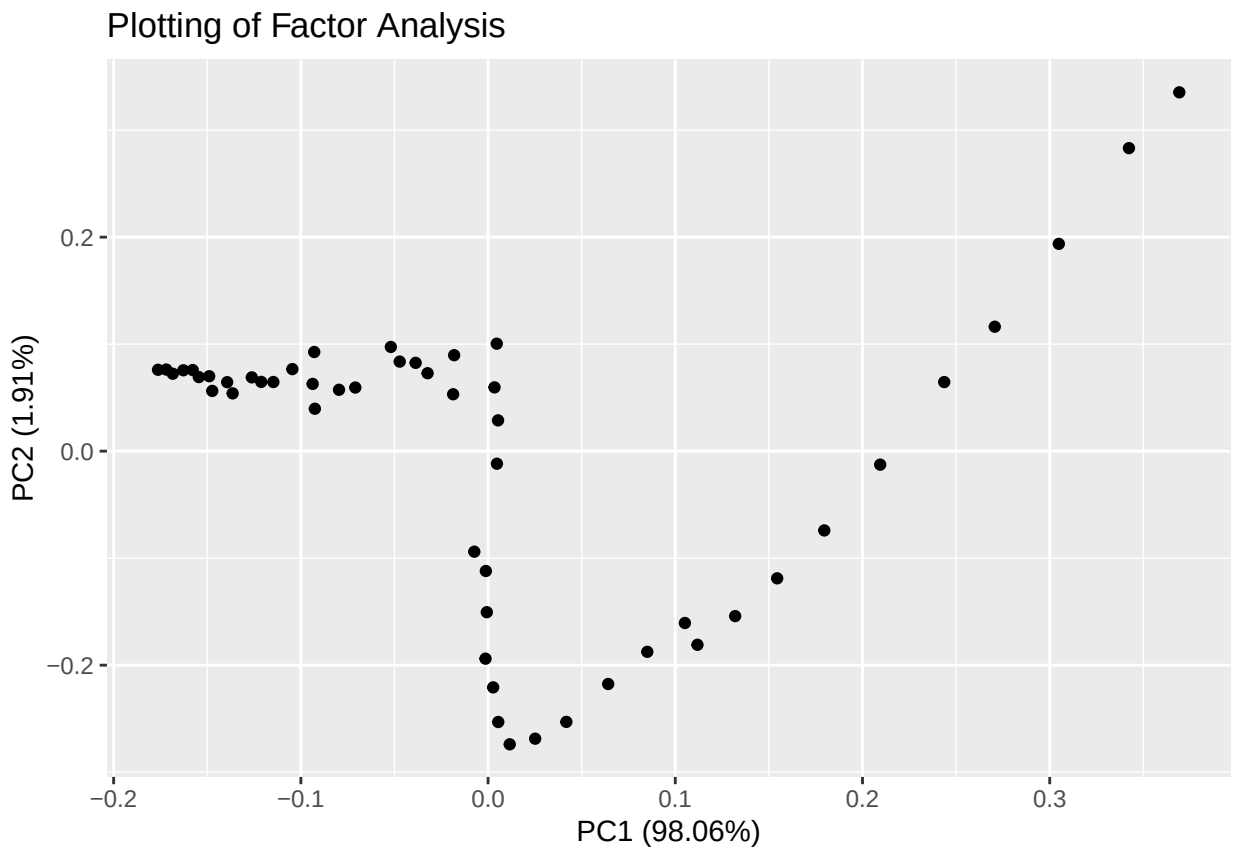


Figure below using autoplot and prcomp function and colour as Total_Trade.

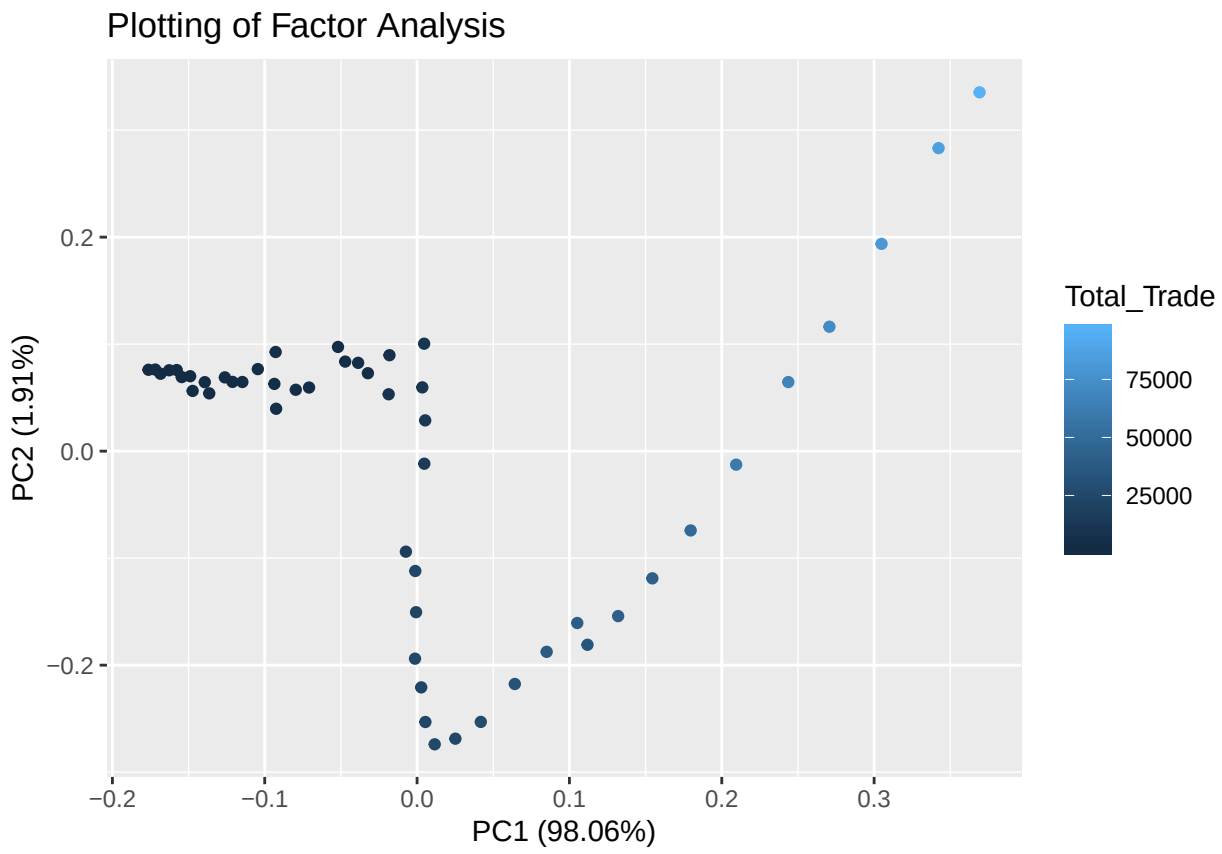


Figure below using autoplot and prcomp function and colour as Total_Trade and label of variable.

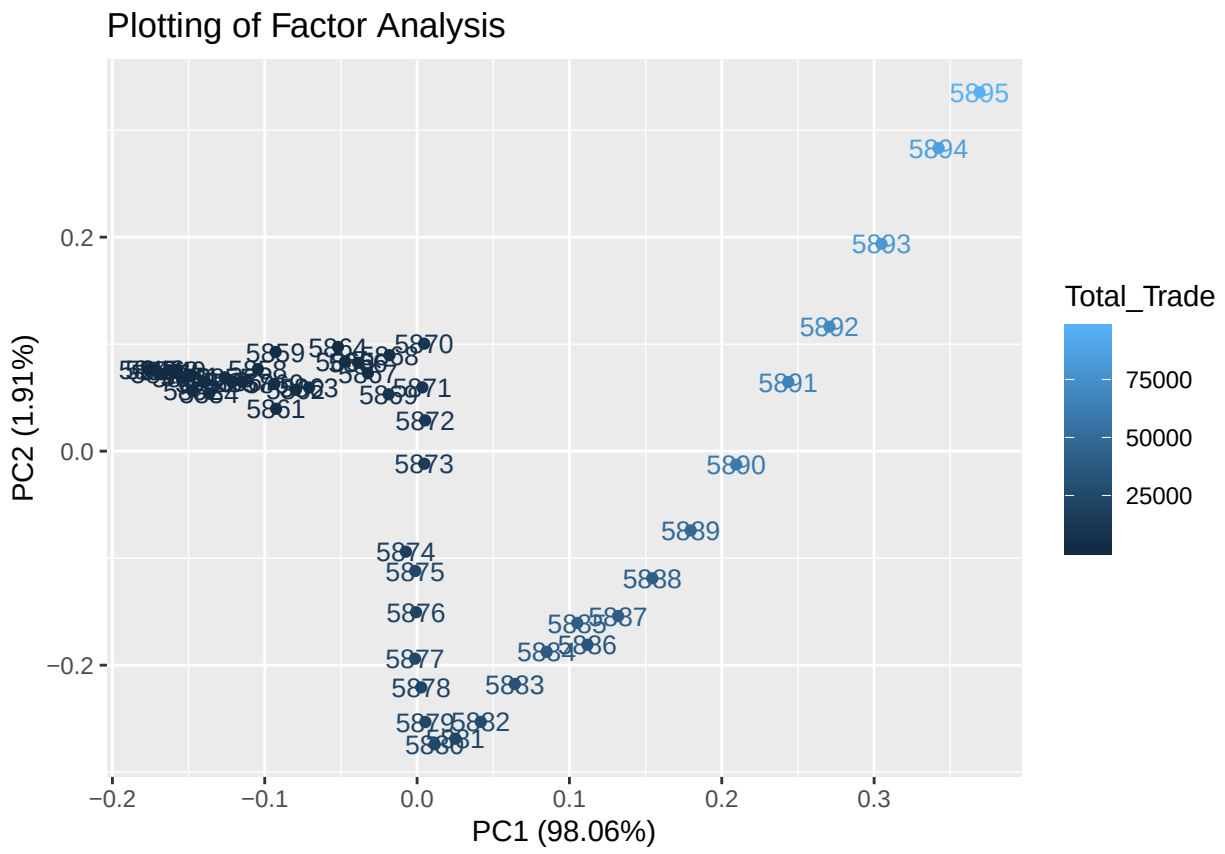


Figure below using autoplot and prcomp function and colour as Total_Trade and without shape of the variable.

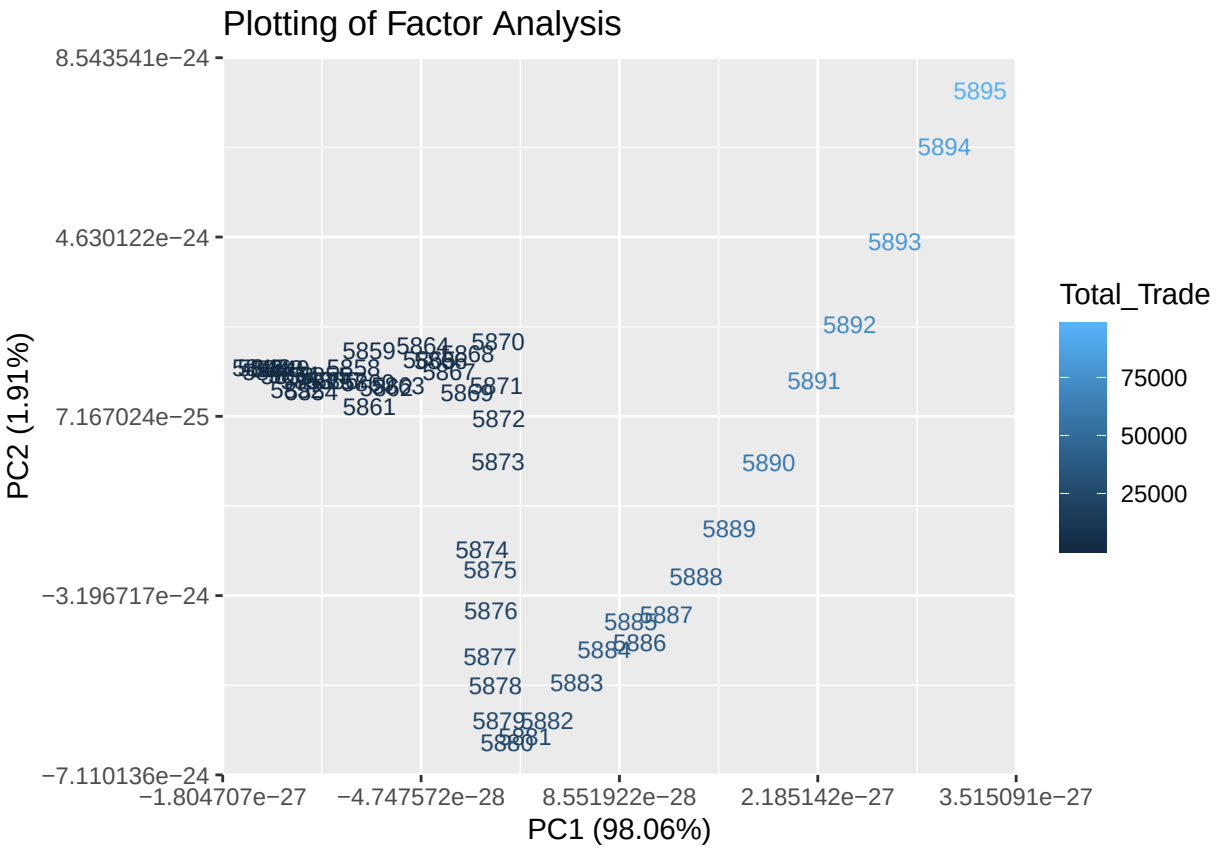


Figure below using autoplot and prcomp function and colour as Total_Trade and loadings as TRUE.

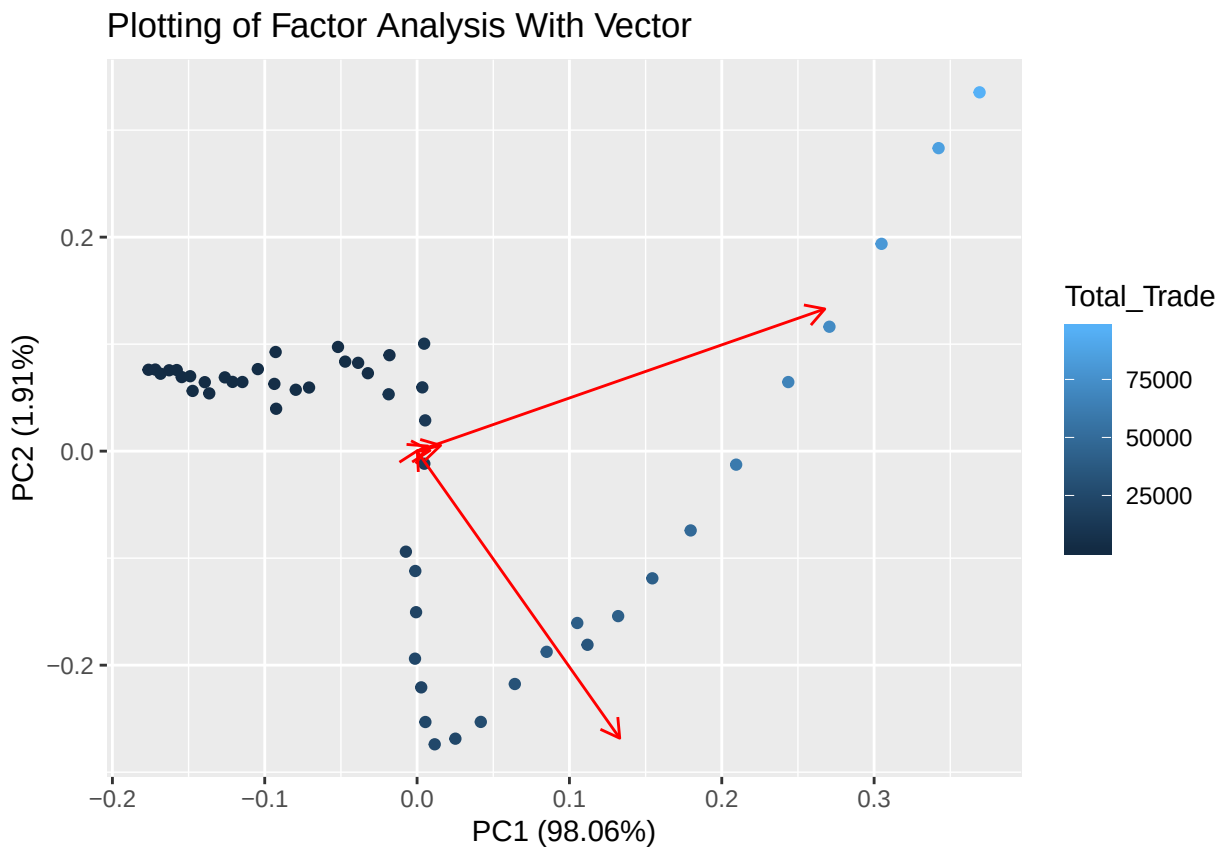


Figure below with autoplot function and loading.colour equal to blue.

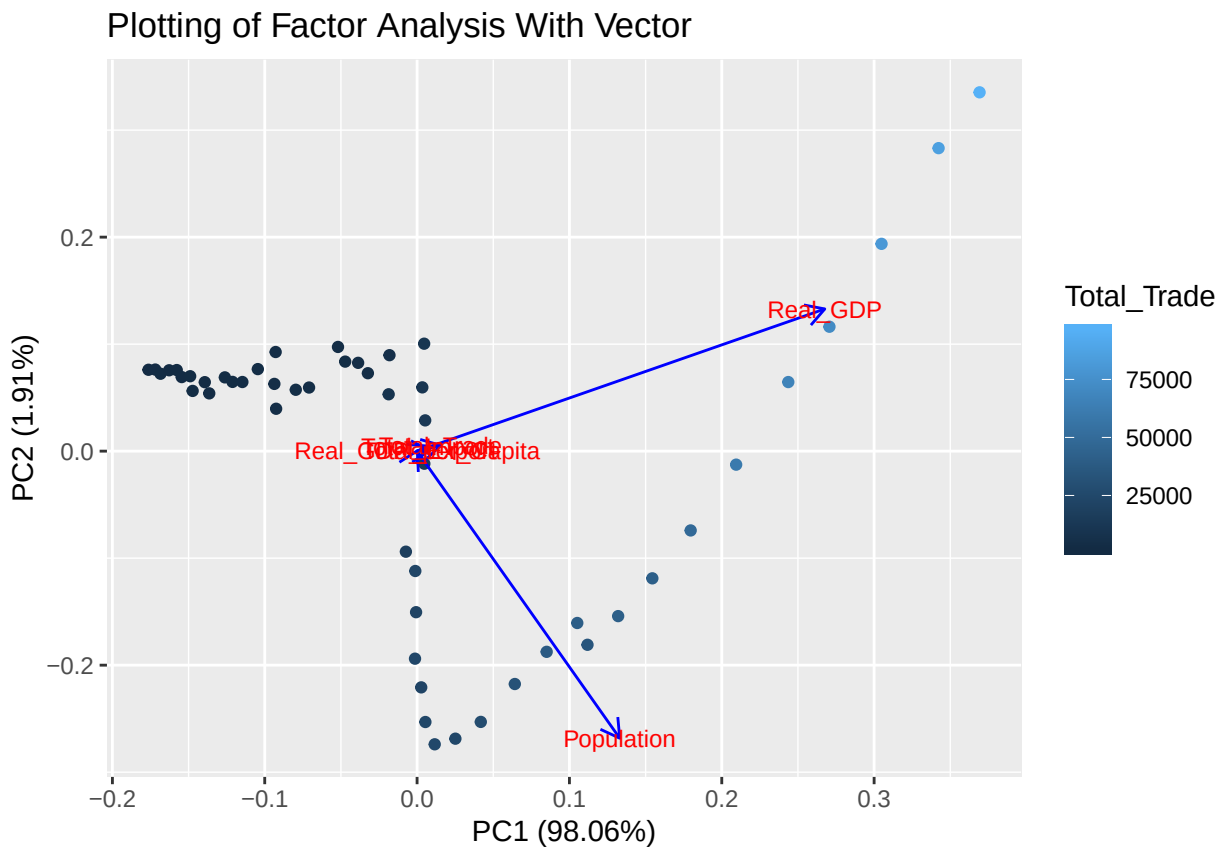
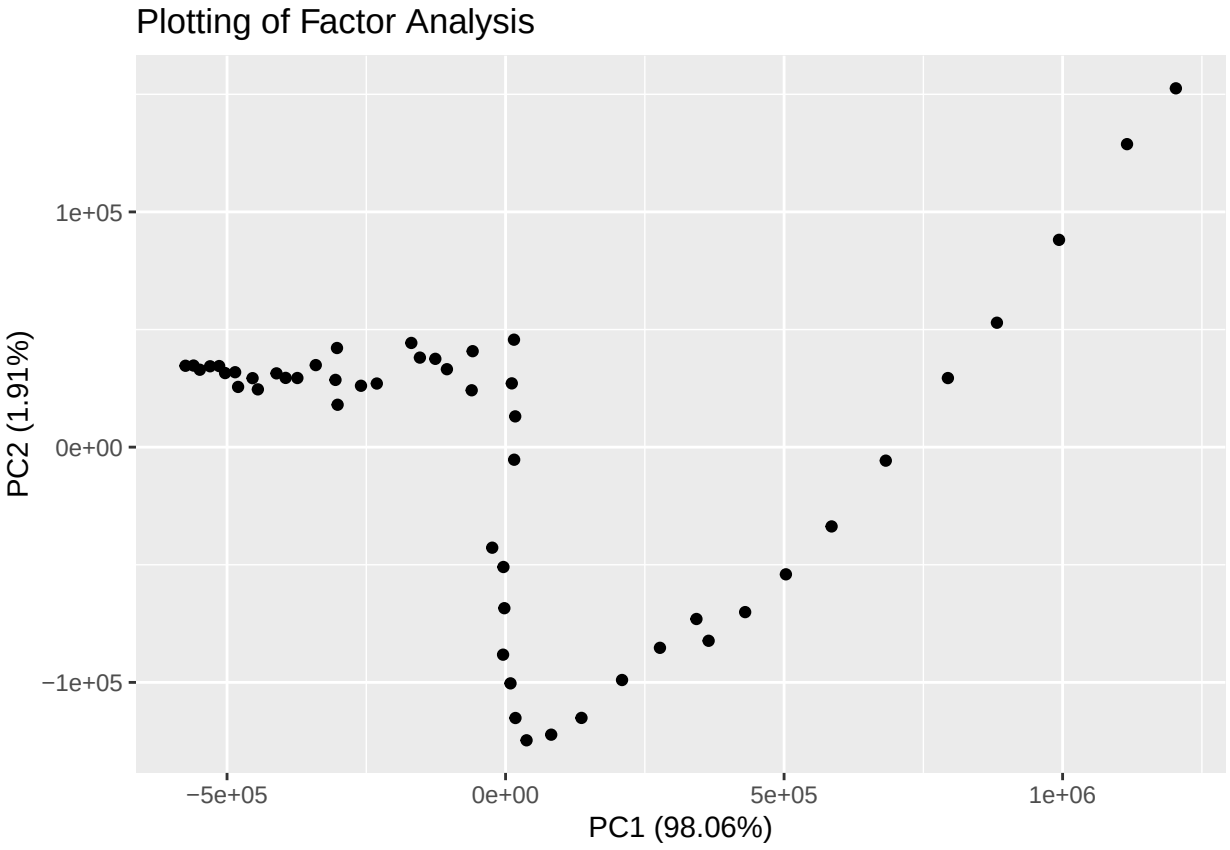


Figure below with autoplot and prcomp function with scale equal to zero.



Conclusion

From Cluster Analysis and PCA Analysis it can be concluded that Total_Trade of India is directly relate to GDP (Real_GDP) along with population growth.

Reference

https://cran.r-project.org/web/packages/ggfortify/vignettes/plot_pca.html

https://qog.pol.gu.se/data/datadownloads/qogstandarddata

http://daviddalpiaz.github.io/appliedstats/model-diagnostics.html

https://www.datacamp.com/community/tutorials/pca-analysis-r

https://bookdown.org/rdpeng/exdata/k-means-clustering.html

https://uc-r.github.io/kmeans_clustering

https://uc-r.github.io/hc_clustering

Class PPT