Assignment 2

1. Exploratory data analysis

1.1 Nature and Characteristics of Variables

The dataset comprises monthly average records spanning from 2000 to 2019. It includes measurements of the concentration of 5 gases in the atmosphere above the Mauna Loa observatory. The variables include Date, CO, CO2, Methane, Nitrous Oxide, and CFC11.

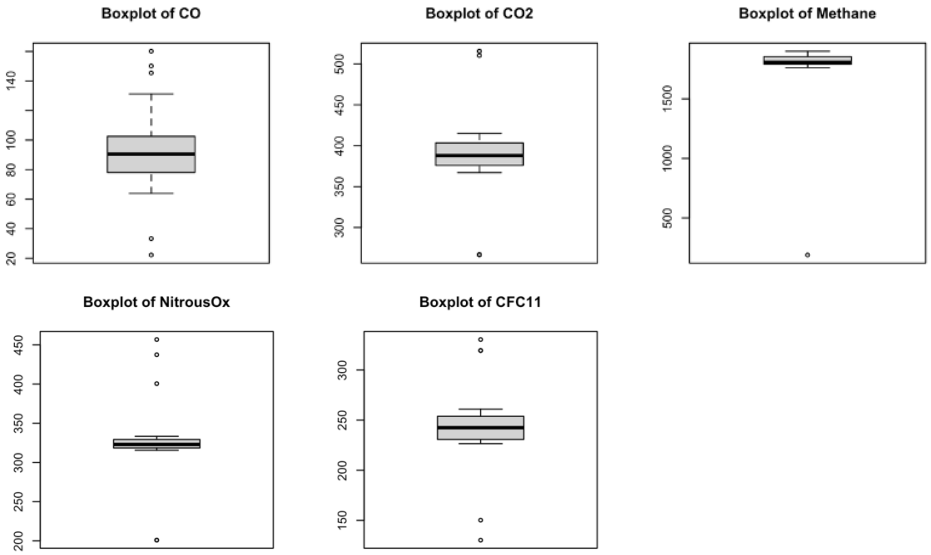

Figure 1. Boxplot of Atmospheric Gas Concentrations

The variable Date is a character variable representing the observation dates, and the dataset is organized in chronological order. As illustrated in Figure 1, CO is a numeric variable, and its boxplot shows the median lying centrally within the box, indicating a normal distribution. The width of the box suggests substantial variability in the measurements. CO2 is also a numeric variable with a normal distribution. Methane presents with a slight rightward skew and a tight box, denoting a more concentrated distribution. Both Nitrous Oxide and CFC-11 are numeric variables with normal distributions, however, Nitrous Oxide displays a tighter distribution compared to CFC-11.

1.2 Relationship Between Variables

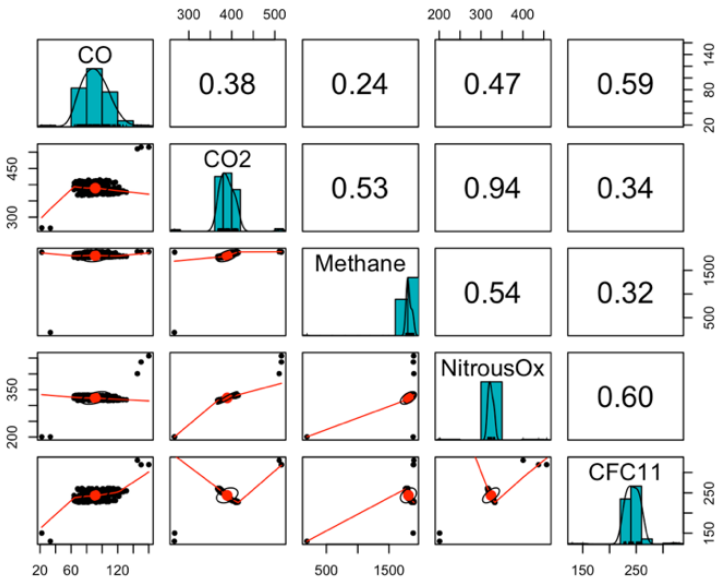

Figure 2. Atmospheric Gas Concentrations Scatterplot Matrix

According to Figure 2, CO2 and Nitrous Ox show a high positive correlation (0.94), which suggests that as the CO2 concentration increases, the Nitrous Ox concentration also tends to increase. There is also a degree of positive correlation between the other gas pairs, which may be due to the fact that these gases are all affected by similar environmental factors, such as temperature and pressure.

1.3 Outliers and Extremes

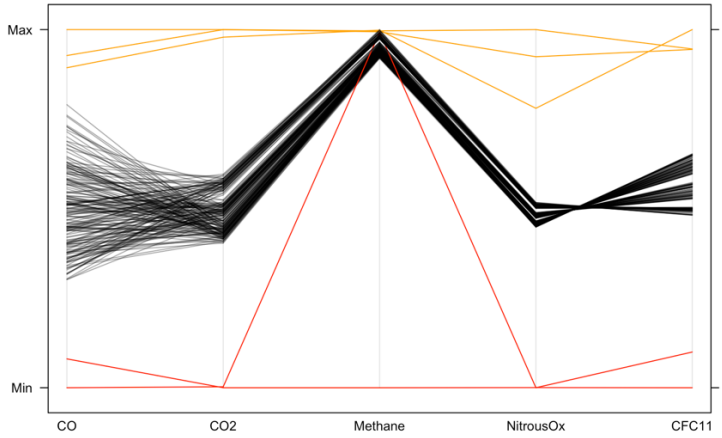

Figure 3. Outliers Visualization via Parallel Coordinate Plot

From Figure 2, CO was identified as having the most outliers. Consequently, a decision was made to visualize these outliers for CO using a Parallel coordinate plot (Figure 3), which would allow for an investigation into their relationship with outliers of other variables. In Figure 3, the orange lines represent observations with CO concentrations above 140, while the red lines represent observations with concentrations below 40. This visual analysis reveals a strong association between CO outliers and those of variables other than methane. Notably, when CO levels are exceptionally high, the concentrations of CO2, Nitrous Ox, and CFC11 also tend to be high, and conversely, these variables have correspondingly low values at very low CO concentrations. These patterns indicate that the anomalous observations are statistically significant and should be kept for further analysis.

1.4 Pre-processing

Initially, the 'Date' variable was a character variable representing the observation dates. It was then converted into a date format to extract the relevant components (year and month) for subsequent analysis. Following this, the data were standardized, a critical step in preparing for Principal Component Analysis (PCA). PCA aims to maximize variance; hence, uniform scaling of features is essential to avoid disproportionate influence on the analysis results due to varied feature variances.

2. Dimension reduction

2.1 Principal Component Analysis (PCA)

I applied Principal Component Analysis (PCA) to the dataset. The high correlation (0.94) between CO2 and Nitrous Oxide suggests they share overlapping information. PCA effectively distilled this information, capturing the essence of the data with fewer components. By reducing the number of variables, PCA simplifies the model, enhancing interpretability and streamlining computational demands.

2.2 Step-by-Step Operations

(a) Data standardization: This equalizes the influence of each variable in the analysis and was completed during data pre-processing.
(b) PCA execution: Employ the prcomp function to derive principal components.
(c) Results interpretation: Assess a summary of the principal components to determine the variance each account for within the data.
(d) Loading matrix (rotation): Ascertain the influence of each original variable on the principal components through their loadings.
(e) Scree plot: This aids in deciding how many principal components should be retained.
(f) PCA variable plot: This step visualizes the orientation and interrelationships of the original variables within the PCA framework, providing insights into their interdependencies and contributions to the principal components.

2.3 Scree Plot Visualization

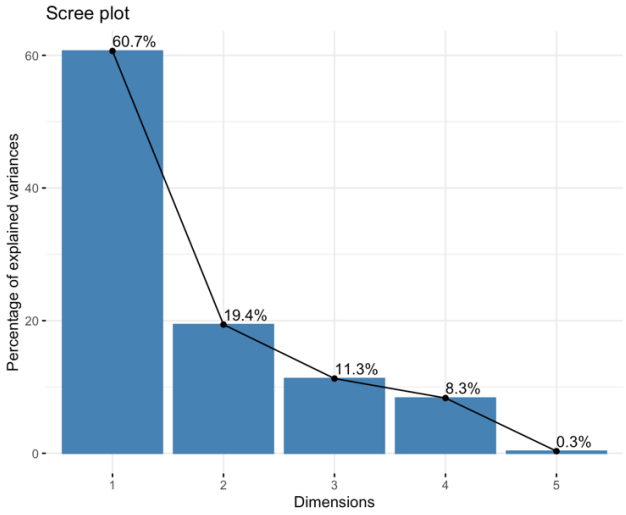

Figure 4. Scree Plot Visualization for PCA

According to Figure 4, it can be determined that retaining two new variables is necessary. The scree plot reveals that PC1 accounts for 60.7% of the variance, and PC2 accounts for 19.4%. Together, PC1 and PC2 explain a significant proportion (> 80%) of the variance, providing an informative overview of the dataset.

2.4 Relationship Between Original and New Variables

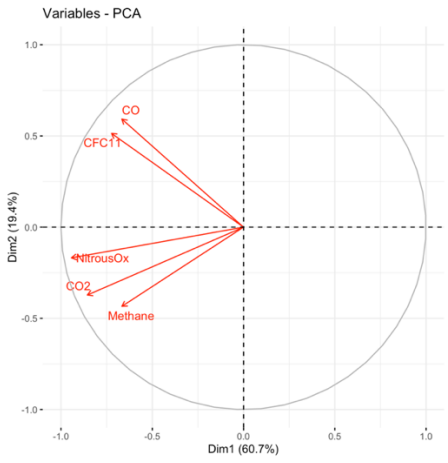

Figure 5. PCA variable plot

As depicted in Figure 5, the first principal component (PC1) is responsible for a large proportion of the variance, indicating it captures a significant trend or pattern within the data. CO2 and Nitrous Oxide significantly contribute to the first principal component (PC1). In contrast, the second principal component (PC2), which accounts for a smaller portion of the variance, sees its most substantial contributions from the CO and CFC11 variables. Methane, while not the primary contributor to either the first or second principal components, has a slightly greater influence on PC2 than PC1.

2.5 Interpretation of New Variables

PC1: The first principal component primarily comprises CO2 and Nitrous Oxide, both recognized as greenhouse gases. This suggests that PC1 likely reflects the level of greenhouse gases that may be directly related to climate change.

PC2: The second principal component is influenced mainly by CO, CFC11, and Methane, gases typically related to combustion processes and industrial emissions. Hence, PC2 likely signifies the influence of industrial activities on atmospheric gas concentrations.

3. Cluster analysis

3.1 Motivation for Cluster Analysis

Firstly, PCA indicates that the first two principal components account for a significant majority of the variance (80%). This suggests that cluster analysis can be conducted in a more simplified and information-dense space, likely yielding clearer and more insightful outcomes. Secondly, cluster analysis proves particularly beneficial for the Mauna Loa dataset as it facilitates the grouping of similar atmospheric gas concentration observations, unveiling natural temporal patterns. These clusters may align with specific environmental conditions, seasonal variations, or times of heightened industrial activities.

3.2 Choice of Clustering Methods

3.2.1 K-medoids：

Justification: K-medoids is less sensitive to outliers compared to k-means, a feature especially relevant given the dataset's outliers. Furthermore, the representative objects (medoids) are actual data points from the dataset, enhancing the interpretability of the results.

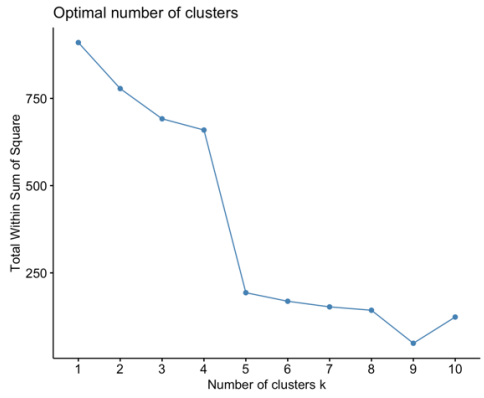

Figure 6. Within-cluster Sum of Squares (WSS)

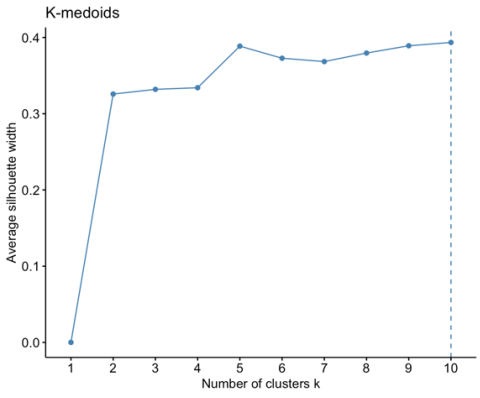

Figure 7. Silhouette

Model Parameters: The Within-cluster Sum of Squares (WSS) and the Silhouette method were utilized to

ascertain the optimal cluster number. Figures 6 and 7 suggest that k=5 is the optimal choice.

3.2.2 Model-based clustering：

Justification: Model-based clustering assumes that the data originates from a mixture of underlying probability distributions. It takes into account the intrinsic properties of the data distribution, enabling it to more accurately identify and adapt to the true shape of the data, including elliptical or other non-spherical clusters. This consideration of the distribution helps achieve more precise clustering in the presence of complexly distributed data.

Model parameters：By comparing models with varying numbers of clusters and covariance structures using the Bayesian Information Criterion (BIC), the highest BIC value obtained was 949.7278, achieved with the VEV model for 4 clusters. Consequently, G=4 and modelNames='VEV' were selected for the model.

3.3 Comparison of Results

	CO	CO2	Methane	NitrousOx	CFC11
[1,]	0.9124821	-0.4299486	-0.17477321	-0.22120659	0.41838137
[2,]	-0.7750671	-0.1909296	-0.06992669	-0.03960613	-0.03330324
[3,]	-0.2351608	0.6298372	0.42918393	0.30850335	-0.70703904
[4,]	2.9654391	4.8993643	0.68848687	3.85801273	4.69801184
[5,]	-3.1666280	-4.9838677	-12.85414829	-6.18082076	-6.13754457

Table 1. K-medoids Cluster Centers Table

	CO	CO2	Methane	NitrousOx	CFC11
[1,]	0.3250745	-0.58714224	-0.18255335	-0.295196688	0.6235859
[2,]	-0.2556009	0.02843446	0.01582594	0.005276574	-0.1799082
[3,]	-0.2523216	0.74961897	0.47248144	0.360910785	-0.7684317
[4,]	0.6034133	1.03697582	-2.05591164	0.777743460	0.3425547

Table 2. Model-based Clustering Means Table

An analysis of Tables 1 and 2 allows for a comparison between the two clustering methods regarding the number of clusters and their characteristics.

Number of Clusters： K-medoids method clusters observations into five groups, whereas Model-based clustering organizes them into four groups.

Cluster features:

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**
K-medoids	High CO and CFC11	Low emission	High CO2 ，Methane and Nitrous OX	extremely high emissions / outliers	extremely low emissions/ outliers
Model-based Clustering	High CO and CFC11	Low emission	High CO2 ，Methane and Nitrous OX	Extreme emissions/ outliers	

Table3. Comparison of Clustering Features between Two Methods

(a) Cluster 1: High concentrations of CO and CFC11 may suggest frequent industrial activity during this period.

(b) Cluster 2: Low atmospheric concentrations of all five gases could indicate the successful implementation of emission control policies at this time.

(c) Cluster 3: Elevated levels of Methane and Nitrous Oxide might reflect an increase in greenhouse gas emissions due to factors like increased fossil fuel consumption, deforestation, and agricultural practices.

(d) Clusters 4 and 5: The unusual fluctuations in atmospheric concentrations of the five gases could result

from specific events or changes, such as natural disasters or industrial accidents.

Table 3 shows that, for the first three cluster features, the two clustering methods yield generally consistent results. However, in handling outliers, K-medoids achieves a more nuanced clustering, whereas Model-based Clustering consolidates all outliers into one cluster.

3.4 Visualization of Cluster Results

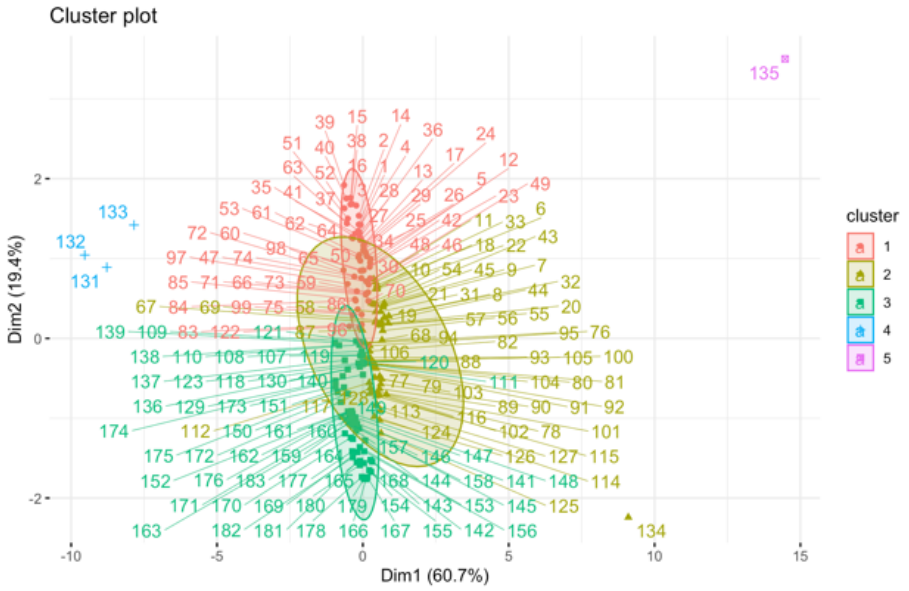

Figure 8. Cluster Plot using K-medoids

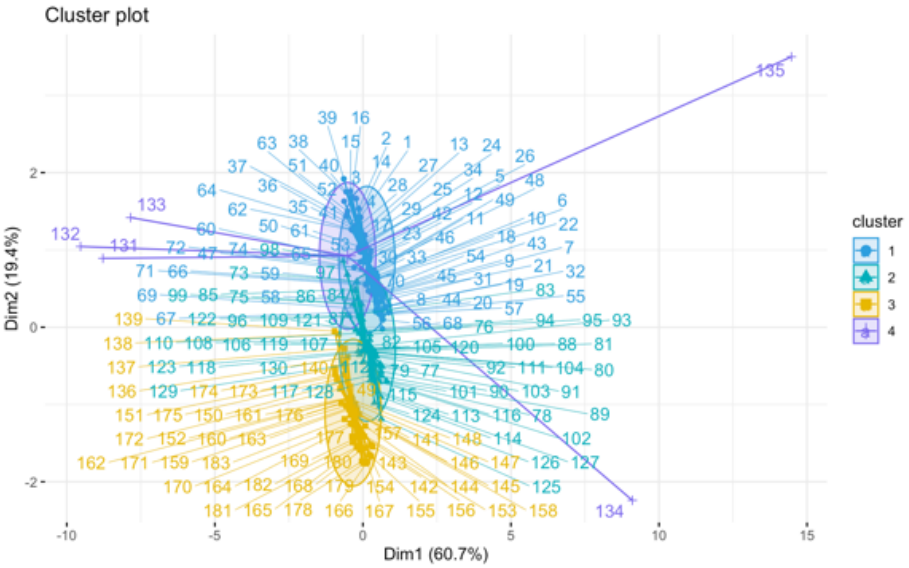

Figure 9. Cluster Plot using Model-based Clustering

Figure 8 reveals that there is some overlap among the five clusters in the K-medoids clustering, with outliers distributed across different categories. This indicates the possibility of complex relationships or transitional patterns among various atmospheric conditions. In contrast, the four clusters identified by Model-based clustering (Figure 9) show little to no overlap, with clearly defined clusters, and all outliers grouped into the same cluster. This suggests the presence of distinct potential emission sources or environmental conditions.

4. Discussion and conclusion

4.1 Conclusions

Principal Component Analysis (PCA) indicates that the first principal component has a strong correlation with CO2 and N2O, hinting that it might represent the overall trend of greenhouse gases in the atmosphere. The second principal component is significantly influenced by CO, CFC11, and Methane, potentially reflecting specific industrial emissions or other human activities.

Cluster Analysis suggests that atmospheric concentration changes of the five gases at Mauna Loa from 2000 to 2019 can be categorized into four to five distinct groups. This further suggests a potential correlation between the variations in gas concentrations and specific time periods.

4.2 Reflection on Different Choices

4.2.1 Dimension Reduction

Given that the dataset primarily consists of numerical variables, Correspondence Analysis is not applicable. Factor Analysis, with only 2 factors considered, produced a p-value of 0.00283, significantly below 0.05, leading to the rejection of the null hypothesis (H0). This suggests a need for more factors. However, given the dataset comprises 5 variables, limiting to two factors restricts its applicability for comprehensive analysis.

4.2.2Clustering

Utilizing hierarchical clustering could provide a layered structure of clusters, potentially revealing various levels of aggregation within the data. This might allow the observation of multi-scale patterns in gas concentration changes from 2000 to 2019, such as identifying sub-clusters of minor changes and their integration into larger clusters over time.

Using k-means, which presupposes spherical clusters, may not suit all data distributions. Given the presence of extreme values and outliers within this dataset, k-means' sensitivity to outliers could disproportionately influence the calculation of cluster centers, complicating the accurate interpretation of clusters.

4.3 Open Questions

The cause of the outliers in the observed data has not yet been determined. These anomalies could result from natural disasters, industrial incidents, or recording errors. Determining the precise cause of an outlier may necessitate collaboration with domain experts and analysis of multiple data sources.

The clusters have not been examined alongside time series analysis. The relationship between these clusters and years needs to be further explored and could reveal whether these clusters correspond to specific time periods or events.

Appendix

#1. Exploratory data analysis

```r
install.packages('psych')
install.packages('naniar')

MaunaLoa<- read.csv("/Users/sutingting/Documents/data science/data exproration/MaunaLoa.csv")
# Display the structure of the data
str(MaunaLoa)

# View the first few rows of data
head(MaunaLoa)

#1. Exploratory data analysis

summary(MaunaLoa)
#the distribution of the numerical variables
par(mfrow=c(2, 3))
boxplot(MaunaLoa$CO, main="Boxplot of CO")
boxplot(MaunaLoa$CO2, main="Boxplot of CO2")
boxplot(MaunaLoa$Methane, main="Boxplot of Methane")
boxplot(MaunaLoa$NitrousOx, main="Boxplot of NitrousOx")
boxplot(MaunaLoa$CFC11, main="Boxplot of CFC11")

#check outliers and extremes
library(scales)
colours <- rep('black',length=nrow(MaunaLoa[,2:6])) ## create a vector of same length as the data
colours <- alpha(colours,0.4) ## use some transparency to fade most of the cases

which(MaunaLoa$CO<40)     ##  find out which row in the data set has a VECT mark below 10
colours[which(MaunaLoa$CO<40)] <- 'red' ## replace the colour for that row with "red"

which(MaunaLoa$CO>140)
colours[which(MaunaLoa$CO>140)] <- 'orange' ## replace the colour for that row with "red"

library(lattice)
parallelplot(MaunaLoa[,2:6], col=colours,  horizontal=FALSE)

#detect missing values
library(naniar)
gg_miss_var(MaunaLoa,show_pct = TRUE)
#no missing values

```

```r
#the relationship of the numerical variables
library(psych)
pairs.panels(MaunaLoa[,2:6],
             method = "pearson", # correlation method
             hist.col = "#00AFBB",
             density = TRUE,  # show density plots
             ellipses = TRUE # show correlation ellipses
)


#pre-processing

# Convert the Date variable to date type,
#retaining only the last two digits of the year and the month
#for easier visualization in subsequent analyses

# Create a new dataframe
new_MaunaLoa <- MaunaLoa
# Replace the Date column
new_MaunaLoa$Date <- format(as.Date(MaunaLoa$Date), "%y-%m")
# View the first few rows to confirm the changes
head(new_MaunaLoa)

#scale
datascaled <- scale(new_MaunaLoa[, 2:6])

```

#2. Dimension reduction

```r
#2.Dimension reduction

#2.1 PCA

pr_out <- prcomp(datascaled)
#names(pr_out)
summary(pr_out)
pr_out$rotation

#scree plot visualization
library(factoextra)
fviz_pca_var(pr_out, axes = c(1, 2), repel = TRUE,
             col.var = "red")

fviz_screeplot(pr_out, addlabels = TRUE)




#plot the contribution of the variables to PC1
fviz_contrib(pr_out, choice = "var", axes = 1, top = 10)
#plot the contribution of the variables to PC2
fviz_contrib(pr_out, choice = "var", axes = 2, top = 10)



#2.2 FA
fa2 <- factanal(datascaled, factors = 2)
fa2
# p-value = 0.00283 < 0.05, reject the hypothesis that two factors are sufficient.
# require more factors to better explain the correlations among variables.

fa3 <- factanal(datascaled, factors = 3)
# An error message was received indicating that 3 factors are too many for 5 variables.
# Therefore, Factor Analysis is not suitable for this dataset.

```

#3. Cluster analysis

```r
#3. Cluster analysis

#3.1 K-medoids
library(cluster)
#choose a suitable number of clusters
fviz_nbclust(datascaled, kmeans, method = "wss")
fviz_nbclust(datascaled, clara, method = "silhouette") + labs(title = "K-medoids")

# Compute K-medoids with k = 5
pam_res <- clara(datascaled, k=5)
pam_res$medoids
pam_res$i.med


#The cluster for the observations
pam_res$clustering

#Visualise K-medoids clustering
library(factoextra)
fviz_cluster(pam_res, datascaled,
             repel = TRUE,
             ellipse.type = "norm",
             ggtheme = theme_minimal()
             )



#3.2 Model-based clustering
install.packages("mclust")
library(mclust)

#find the model combination which gives the largest BIC
mc_bic2 <- Mclust(datascaled, G = 2)
mc_bic2$BIC
max(mc_bic2$BIC, na.rm = TRUE) #ignore NA


mc_bic3 <- Mclust(datascaled, G = 3)
mc_bic3$BIC
max(mc_bic3$BIC, na.rm = TRUE) #ignore NA


mc_bic4 <- Mclust(datascaled, G = 4)
mc_bic4$BIC
max(mc_bic4$BIC, na.rm = TRUE) #ignore NA

mc_bic5 <- Mclust(datascaled, G = 5)
mc_bic5$BIC
max(mc_bic5$BIC, na.rm = TRUE) #ignore NA


mc_bic6 <- Mclust(datascaled, G = 6)
mc_bic6$BIC
max(mc_bic6$BIC, na.rm = TRUE) #ignore NA


#The maximum bic value here is 949.7278 using VEV
mcMau <- Mclust(datascaled, G = 4, modelNames = "VEV") # Model-based-clustering
summary(mcMau)

cluster_means <- mcMau$parameters$mean
t(cluster_means)

mcMau$classification

#Visualise Model-based clustering
library(factoextra)
fviz_cluster(mcMau, data = datascaled,
             palette = c("#2E9FDF", "#00AFBB", "#E7B800","#7B68EE"),
             ellipse.type = "euclid", # Concentration ellipse
             star.plot = TRUE, # Add segments from centroids to items
             repel = TRUE, # Avoid label overplotting (slow)
             ggtheme = theme_minimal()
)
```