Data Exploration, Visualisation and Unsupervised Learning - Assignment 2

1. Introduction

This report focuses on the exploration and analysis of the “MaunaLoa” dataset using unsupervised learning techniques, which consists of observations of monthly averages for selected months between 2000 and 2019 of the measurements of the concentration of 5 gases in the atmosphere above the Mauna Loa observatory.

Initially, the dataset is explored and pre-processed to facilitate subsequent dimension reduction and cluster analysis. Subsequently, Principal Component Analysis (PCA) is applied to reduce the dimensionality of the dataset. Two distinct approaches are then employed for clustering the dataset. Finally, the report summarizes the findings from the dataset exploration and discusses potential alternative methods and limitations of the current investigation.

2. Exploratory Data Analysis

2.1. Features and Relationship

Upon thorough inspection, no missing values are present in any variables. Analysis of the scatter plot matrix (Figure 1) reveals that while the distribution of Carbon monoxide (CO) appears to follow a normal distribution, the distributions of other variables exhibit skewness or compression, making them challenging to interpret due to outliers and extreme values. Similarly, the scatter plots are unreadable to interpret owing to outliers and extremes. Some relationships are observed among these variables, notably with Carbon dioxide (CO2) and Nitrous Oxide demonstrating a remarkably positive correlation of 0.94.

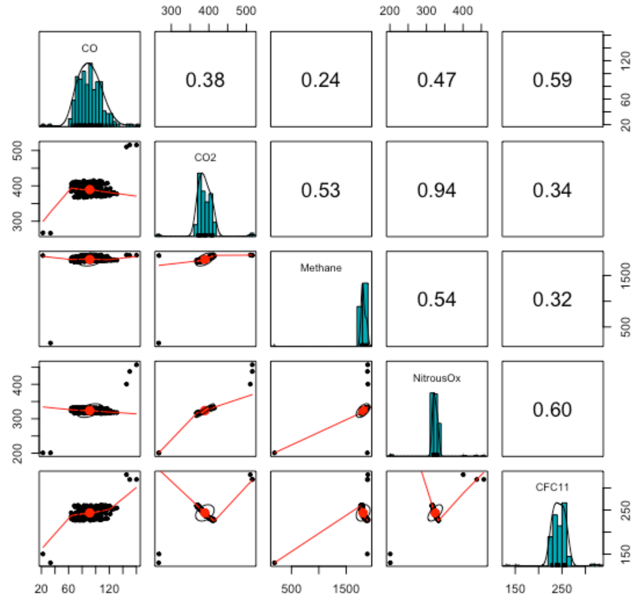


Figure 1. Scatter Plot Matrix of Monthly Averages of 5 Gases

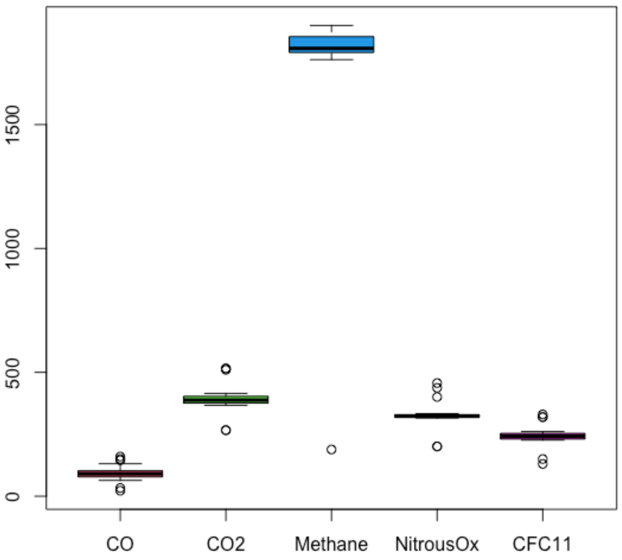


Figure 2. Boxplot of Monthly Averages of 5 Gases

2.2. Outliers and Extremes

The boxplot (Figure 2) analysis showcases that the Methane variable exhibits larger values, including an extreme, setting it apart from the other variables. However, it is notable that direct comparisons between these variables are not feasible due to differences in their units. Additionally, outliers are present in the other four variables.

2.3. Pre-processing

Scaling the dataset is necessary to facilitate further analysis and operations. And fortunately, all outliers and extremes originate from just five observations. Thus, by removing them, most of the information is still retained. The cleaned and scaled dataset is explored below (Figure 3).

Following the removal of outliers, the distributions become clearer. The variables exhibit multimodal distributions except Carbon monoxide (CO). The distributions of Carbon dioxide (CO2), Nitrous Oxide and CFC-11 demonstrate three modes, suggesting the potential presence of three subsets for each variable, which is a clue for further clustering. The scatter plot patterns also support this notion.

Furthermore, significant positive correlations are observed among Carbon dioxide (CO2), Methane, and Nitrous Oxide, while CFC-11 exhibits strong negative correlations with these three variables. In contrast, Carbon monoxide (CO) demonstrates weaker relationships with the other variables.

The boxplot (Figure 4) analysis confirms that the variables are appropriately scaled for subsequent analysis.

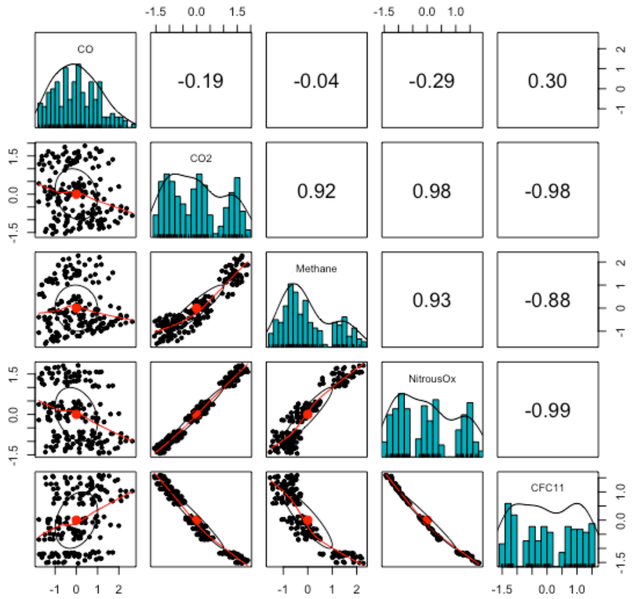


Figure 3. Scatter Plot Matrix after Omitting Outliers and Scaling

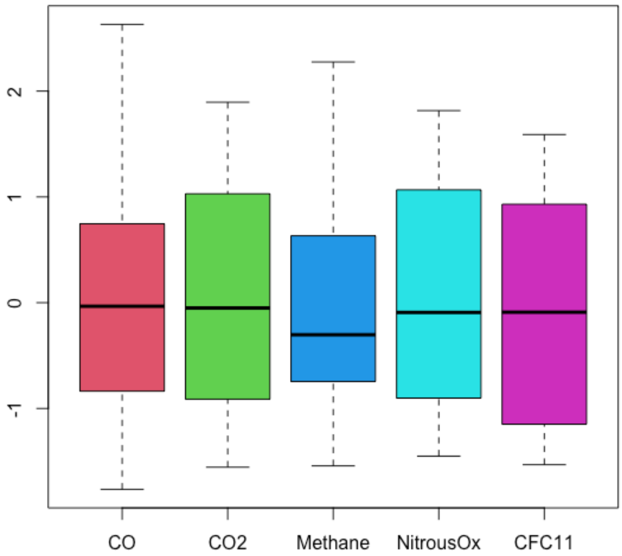


Figure 4. Boxplot after Omitting Outliers and Scaling

3. Dimension Reduction

3.1 Technique Choosing

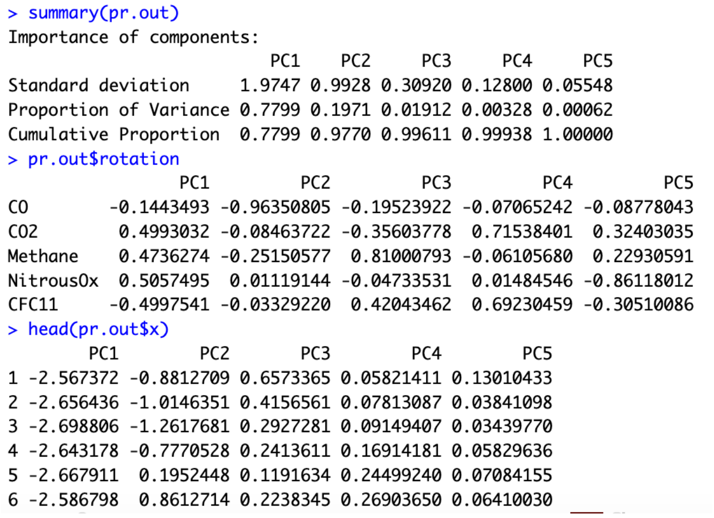
Carbon dioxide (CO2), Methane, Nitrous Oxide, and CFC-11 exhibit significant correlations, as indicated by the correlation matrix. This implies redundant information among these variables, making Principal Component Analysis (PCA) a suitable method for removing this redundancy, thereby reducing the dimension of the data. The higher the correlation, the higher the redundancy in the data in which case PCA becomes more effective and facilitates clearer interpretation.

3.2 Scaling

Data is scaled to account for variations in units across variables. Carbon dioxide (CO2), measured in parts per million, displays a notably higher variance among the original dataset, which aligns with the fact that Carbon dioxide (CO2) is more prevalent in the atmosphere than other gases.

3.3 Dimension Reducing

The summary (Figure 5) provides insights into standard deviation, proportion of explained variance, and cumulative proportion for each principal component. Additionally, the loading vectors are the columns of the rotation matrix**.** Each row of scores, denoting the coefficients of the observations in the basis (ranging from PC1 to PC5), is computed and will be further visualized.



```
> summary(pr.out)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5
Standard deviation     1.9747 0.9928 0.30920 0.12800 0.05548
Proportion of Variance 0.7799 0.1971 0.01912 0.00328 0.00062
Cumulative Proportion  0.7799 0.9770 0.99611 0.99938 1.00000
> pr.out$rotation
                 PC1         PC2         PC3         PC4         PC5
CO        -0.1443493 -0.96350805 -0.19523922 -0.07065242 -0.08778043
CO2        0.4993032 -0.08463722 -0.35603778  0.71538401  0.32403035
Methane    0.4736274 -0.25150577  0.81000793 -0.06105680  0.22930591
NitrousOx  0.5057495  0.01119144 -0.04733531  0.01484546 -0.86118012
CFC11     -0.4997541 -0.03329220  0.42043462  0.69230459 -0.30510086
> head(pr.out$x)
        PC1        PC2       PC3        PC4        PC5
1 -2.567372 -0.8812709 0.6573365 0.05821411 0.13010433
2 -2.656436 -1.0146351 0.4156561 0.07813087 0.03841098
3 -2.698806 -1.2617681 0.2927281 0.09149407 0.03439770
4 -2.643178 -0.7770528 0.2413611 0.16914181 0.05829636
5 -2.667911  0.1952448 0.1191634 0.24499240 0.07084155
6 -2.586798  0.8612714 0.2238345 0.26903650 0.06410030
```

Figure 5. PCA Summary, Loadings and Scores

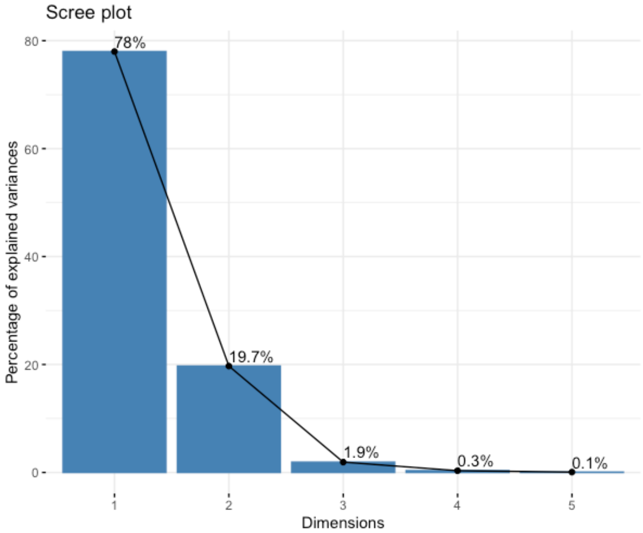


Figure 6. Scree Plot of the Proportion of Variance Explained

3.4 Number of New Variables Choosing (Scree Plot):

The scree plot (Figure 6) demonstrates that PC1 encodes 78% of the variance, retaining an amount of information. PC1 and PC2 together capture 97.7% of the variance in the dataset, exceeding the 80% threshold, significant enough to provide an informative representation.

3.5 Relationship between the Original and New Variables:

The biplot (Figure 7) reveals that the similar lengths of the five vectors suggest that the first two principal components represent fairly the original variables. As anticipated, Carbon dioxide (CO2), Methane, Nitrous Oxide, and CFC-11 contribute significantly to the first principal component, while Carbon monoxide (CO) predominantly contributes to the second principal component since Carbon monoxide (CO) exhibits less correlation with other variables. Observations positioned on the left indicate higher proportions of Carbon dioxide (CO2), Methane, and Nitrous Oxide, whereas observations at the bottom denote a higher proportion of Carbon monoxide (CO).

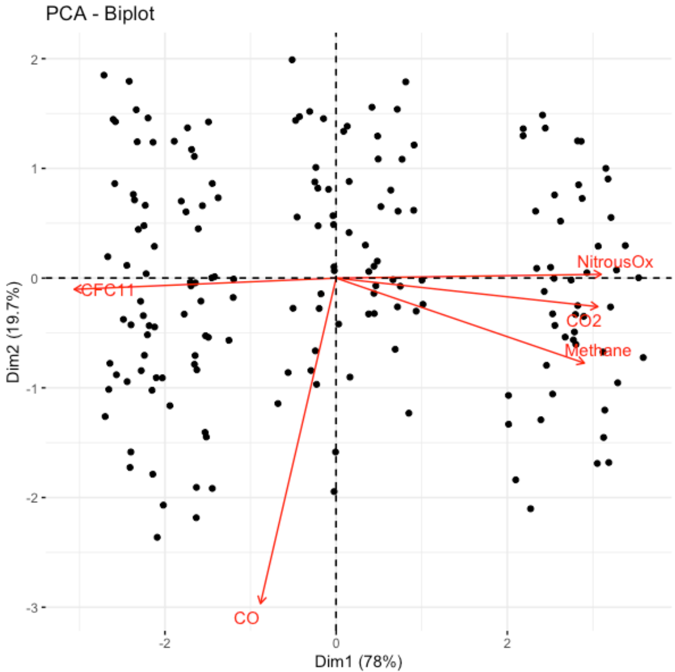


Figure 7. Biplot of PCA

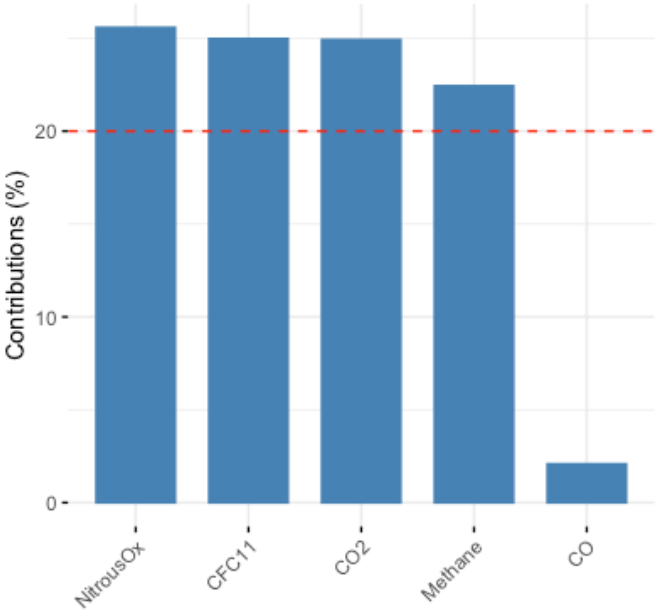


Figure 8. Contribution of Variables to Dim-1

The contribution plots (Figures 8&9) and matrix (Figure 10) provide detailed and quantitative insights into the contributions of variables to dimensions. The first principal component is contributed by CO2 (24.93%), Methane (22.43%), Nitrous Oxide (25.58%), and CFC11 (24.98%), while Carbon monoxide (CO) contributes 92.83% to the second principal component.

The correlation between variables and principal components (Figure 11) shows that Carbon dioxide (CO2), Methane, and Nitrous Oxide exhibit high positive correlations with the first principal component, whereas CFC-11 demonstrates a significant negative correlation. Comparably, there is a significant negative correlation between the second principal component and Carbon monoxide (CO).

The quality of representation (Figure 12) suggests that Carbon dioxide (CO2), Methane, Nitrous Oxide, and CFC-11 variables are adequately explained by the first principal component, with respective explanations of 97.21%, 87.47%, 99.74%, and 97.39%. Additionally, the second principal component largely accounts for 91.50% variability of Carbon monoxide (CO).

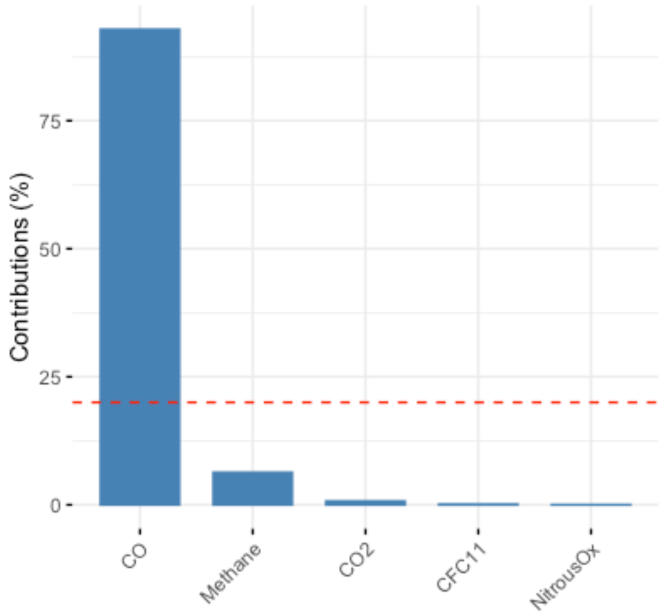
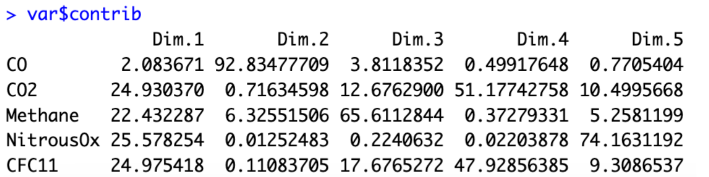
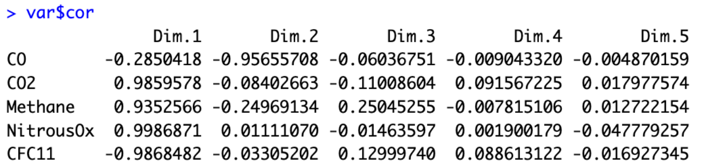


Figure 9. Contribution of Variables to Dim-2



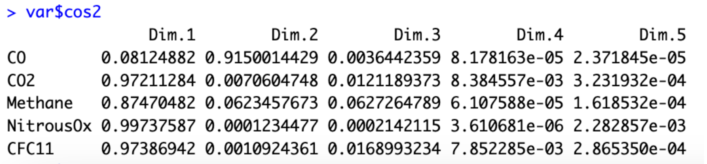
```
> var$contrib
              Dim.1       Dim.2      Dim.3       Dim.4      Dim.5
CO         2.083671 92.83477709  3.8118352  0.49917648  0.7705404
CO2       24.930370  0.71634598 12.6762900 51.17742758 10.4995668
Methane   22.432287  6.32551506 65.6112844  0.37279331  5.2581199
NitrousOx 25.578254  0.01252483  0.2240632  0.02203878 74.1631192
CFC11     24.975418  0.11083705 17.6765272 47.92856385  9.3086537
```

Figure 10. Contribution Matrix of Variables



```
> var$cor
               Dim.1       Dim.2       Dim.3        Dim.4        Dim.5
CO        -0.2850418 -0.95655708 -0.06036751 -0.009043320 -0.004870159
CO2        0.9859578 -0.08402663 -0.11008604  0.091567225  0.017977574
Methane    0.9352566 -0.24969134  0.25045255 -0.007815106  0.012722154
NitrousOx  0.9986871  0.01111070 -0.01463597  0.001900179 -0.047779257
CFC11     -0.9868482 -0.03305202  0.12999740  0.088613122 -0.016927345
```

Figure 11. Correlation between variables and principal components



```
> var$cos2
               Dim.1        Dim.2        Dim.3        Dim.4        Dim.5
CO        0.08124882 0.9150014429 0.0036442359 8.178163e-05 2.371845e-05
CO2       0.97211284 0.0070604748 0.0121189373 8.384557e-03 3.231932e-04
Methane   0.87470482 0.0623457673 0.0627264789 6.107588e-05 1.618532e-04
NitrousOx 0.99737587 0.0001234477 0.0002142115 3.610681e-06 2.282857e-03
CFC11     0.97386942 0.0010924361 0.0168993234 7.852285e-03 2.865350e-04
```

Figure 12. Quality of Representation

3.6 Interpretation

Through a series of analyses, the first principal component explains better the variance of the Carbon dioxide (CO2), Methane, Nitrous Oxide and CFC11 variables, which are all greenhouse gases. While not as precise as factor analysis, the first principal component could be interpreted as the new feature representing greenhouse gas. However, the second component still predominantly reflects Carbon monoxide (CO) due to its substantial weight within this component.

4. Cluster Analysis

4.1. Motivation

Cluster analysis can effectively summarize data with a reduced representation, gain fresh insights into data structure, and generate a new categorical variable for input in further analysis or supervised learning modelling. The MaunaLoa dataset comprises the date variable and monthly averages of 5 gas variables, which are complex numerical values impacting comprehension and interpretation. Through cluster analysis, inherent patterns identified could solve the lack of intuitive and categorical variables. The newly generated variable can be used for subsequent modelling and prediction tasks.

4.2. Method

K-means and model-based clustering are employed here. K-means clustering is a basic and relatively straightforward method, whereas model-based clustering uses advanced techniques. Although complex and advanced approaches do not necessarily guarantee better performance, comparing these two methods could yield interesting insights and effective results.

K-means Clustering

Starting with choosing the number of clusters, while there is no prior personal and scientific judgement to have a sensible number of clusters, a pattern of 3 groups is showcased according to the previous PCA analysis. Both the scree-like plot and silhouette measure provide scientific evidence. The elbow of the scree-like plot (Figure 13) suggests that 3 clusters are optimal. The silhouette plot (Figure 14) indicates that 2 clusters are the best option, while 3 clusters show the same silhouette width. Considering all analyses, the optimal number of clusters is 3.

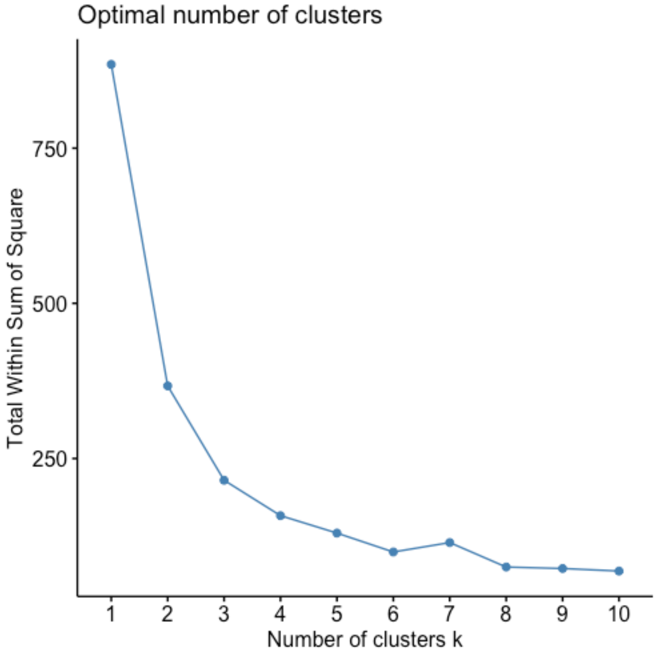


Figure 13. Scree-like Plot of Optional number of Clusters

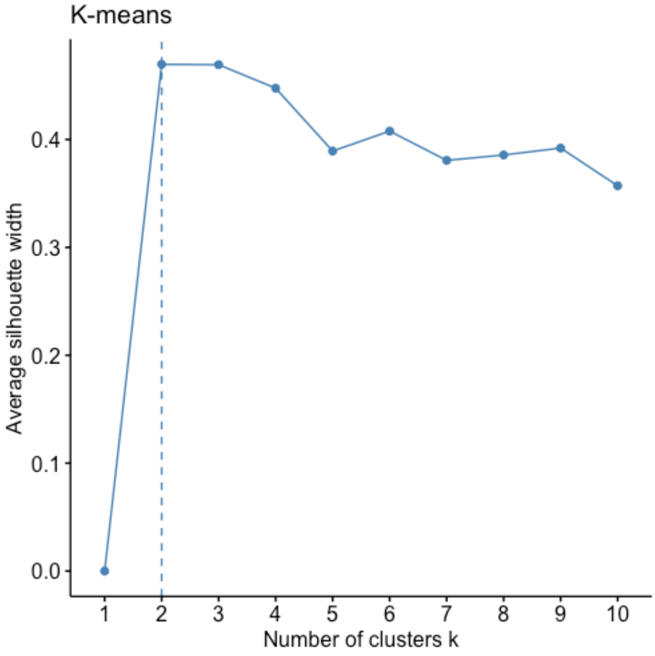
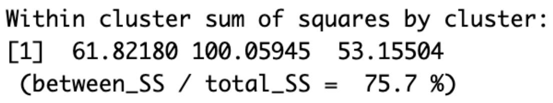


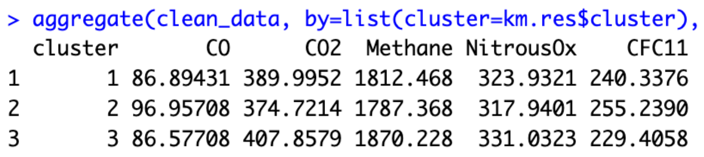
Figure 14. Silhouette Measure Plot of K-means Clustering

The portion of variability explained (Figure 15) is 75.7%, which is fairly good. The means of five variables based on the newly created clusters (Figure 16) are shown below. Carbon dioxide (CO2), Methane, Nitrous Oxide and CFC11 exhibit clear patterns, with their means being highest in cluster 3, followed by cluster 1, and lowest in cluster 2. CFC11 instead shows an opposite pattern. However, the mean of Carbon monoxide (CO) seems to have no clear trend.



```
Within cluster sum of squares by cluster:
[1]  61.82180 100.05945  53.15504
 (between_SS / total_SS =  75.7 %)
```

Figure 15. the Portion of Variability Explained

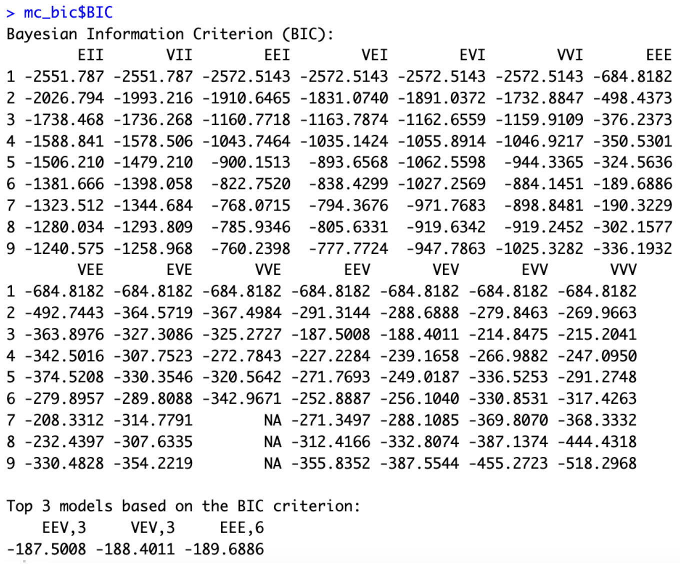


```
> aggregate(clean_data, by=list(cluster=km.res$cluster),
  cluster       CO      CO2  Methane NitrousOx    CFC11
1       1 86.89431 389.9952 1812.468  323.9321 240.3376
2       2 96.95708 374.7214 1787.368  317.9401 255.2390
3       3 86.57708 407.8579 1870.228  331.0323 229.4058
```

Figure 16. The Means based on the Newly Created Clusters

Model-based Clustering

The maximum BIC value here is -187.5 using EEV (Figure 17&18). Therefore, the optimal number of clusters is determined to be 3.



```
> mc_bic$BIC
Bayesian Information Criterion (BIC):
        EII       VII        EEI        VEI        EVI        VVI       EEE
1 -2551.787 -2551.787 -2572.5143 -2572.5143 -2572.5143 -2572.5143 -684.8182
2 -2026.794 -1993.216 -1910.6465 -1831.0740 -1891.0372 -1732.8847 -498.4373
3 -1738.468 -1736.268 -1160.7718 -1163.7874 -1162.6559 -1159.9109 -376.2373
4 -1588.841 -1578.506 -1043.7464 -1035.1424 -1055.8914 -1046.9217 -350.5301
5 -1506.210 -1479.210  -900.1513  -893.6568 -1062.5598  -944.3365 -324.5636
6 -1381.666 -1398.058  -822.7520  -838.4299 -1027.2569  -884.1451 -189.6886
7 -1323.512 -1344.684  -768.0715  -794.3676  -971.7683  -898.8481 -190.3229
8 -1280.034 -1293.809  -785.9346  -805.6331  -919.6342  -919.2452 -302.1577
9 -1240.575 -1258.968  -760.2398  -777.7724  -947.7863 -1025.3282 -336.1932
        VEE       EVE       VVE       EEV       VEV       EVV       VVV
1 -684.8182 -684.8182 -684.8182 -684.8182 -684.8182 -684.8182 -684.8182
2 -492.7443 -364.5719 -367.4984 -291.3144 -288.6888 -279.8463 -269.9663
3 -363.8976 -327.3086 -325.2727 -187.5008 -188.4011 -214.8475 -215.2041
4 -342.5016 -307.7523 -272.7843 -227.2284 -239.1658 -266.9882 -247.0950
5 -374.5208 -330.3546 -320.5642 -271.7693 -249.0187 -336.5253 -291.2748
6 -279.8957 -289.8088 -342.9671 -252.8887 -256.1040 -330.8531 -317.4263
7 -208.3312 -314.7791        NA -271.3497 -288.1085 -369.8070 -368.3332
8 -232.4397 -307.6335        NA -312.4166 -332.8074 -387.1374 -444.4318
9 -330.4828 -354.2219        NA -355.8352 -387.5544 -455.2723 -518.2968

Top 3 models based on the BIC criterion:
    EEV,3     VEV,3     EEE,6 
-187.5008 -188.4011 -189.6886
```

Figure 17. Bayesian Information Criterion of Model-based Clustering

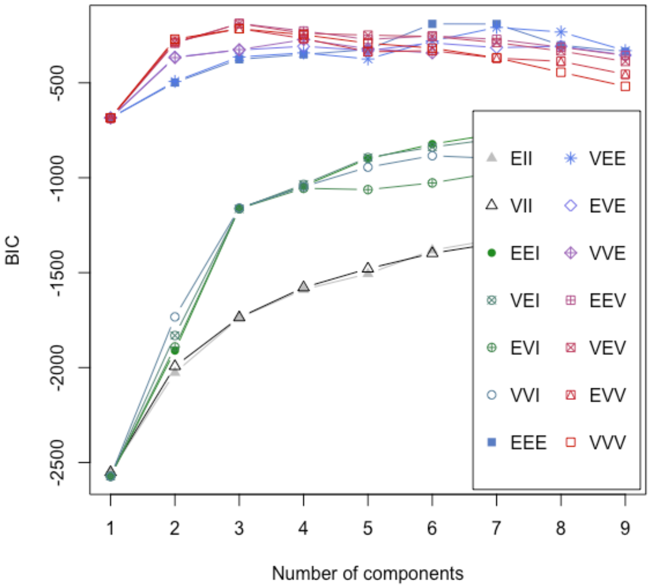


Figure 18. BIC Plot of Model-based Clustering

The uncertainty plot (Figure 19) demonstrates that there is an observation that could potentially be misclustered.

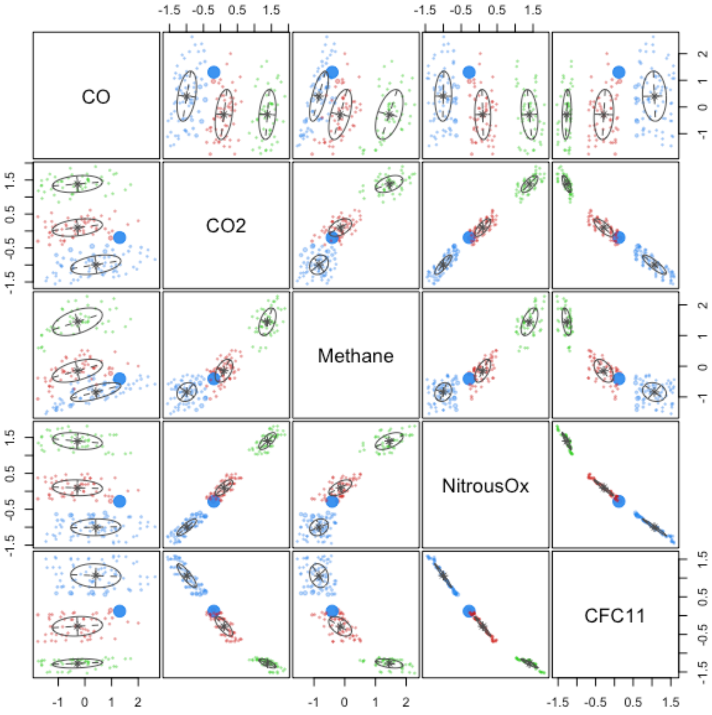
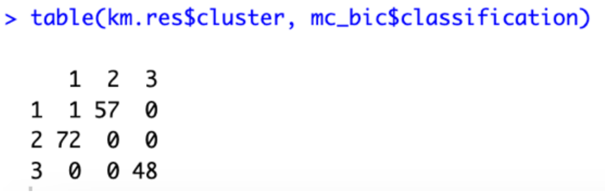


Figure 19. Uncertainty Plot of Model-based Clustering

4.3. Result Comparison

A classification table (Figure 20) compares the outcomes of two different methods. Impressively, the k-means clustering shows equivalent performance compared to model-based clustering, with only one observation clustered differently.

The cluster silhouette plots (Figures 21&22) reveal that both methods exhibit the same average silhouette width of 0.47. However, a negative silhouette value is observed in cluster 1 using model-based clustering, indicating a potential misclassification of this observation (consistent with the uncertainty plot findings). Therefore, k-means clustering outperforms model-based clustering, suggesting that simple methods yield better results in this scenario.



```
> table(km.res$cluster, mc_bic$classification)
   
     1  2  3
  1  1 57  0
  2 72  0  0
  3  0  0 48
```

Figure 20. Classification Table of Two Methods

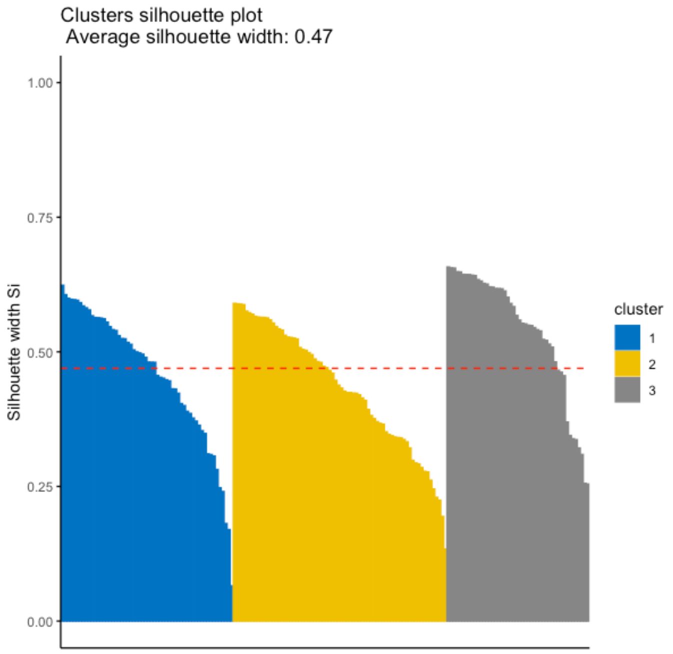


Figure 21. Clusters Silhouette Plot of K-means Clustering

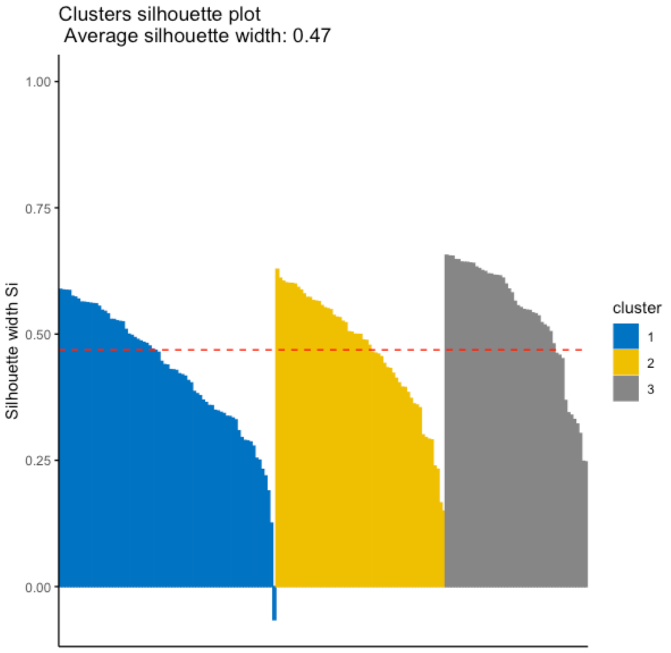


Figure 22. Clusters Silhouette Plot of Model-based Clustering

4.4. Visualization and Interpretation

The plots (Figures 23&24) show that the dataset is divided into three groups, with only one differently clustered observation (observation 74). Recalling the directions of loadings of PCA, the classification of clusters is determined by the positive proportionality of Carbon dioxide (CO2), Methane, and Nitrous Oxide, and the negative proportionality of CFC11. Carbon monoxide (CO) does not contribute to the clustering.

In addition, based on the number of observations, the dataset can be divided into three distinct time stages with clear boundaries. The three clusters identified by K-means clustering are respectively 1-72, 73-130, 136-183, while those determined by model-based clustering are 1-72 and 74, 73 and 75-130, 136-183 (observations 131-135 are outliers). The underlying reasons might be that the levels of greenhouse gases are increasing over time due to factors such as the rising consumption of mineral fuels, the expansion of animal husbandry, and rapid industrial development, while the decrease in CFC11 levels could be attributed to the prohibition and substitution for its use.

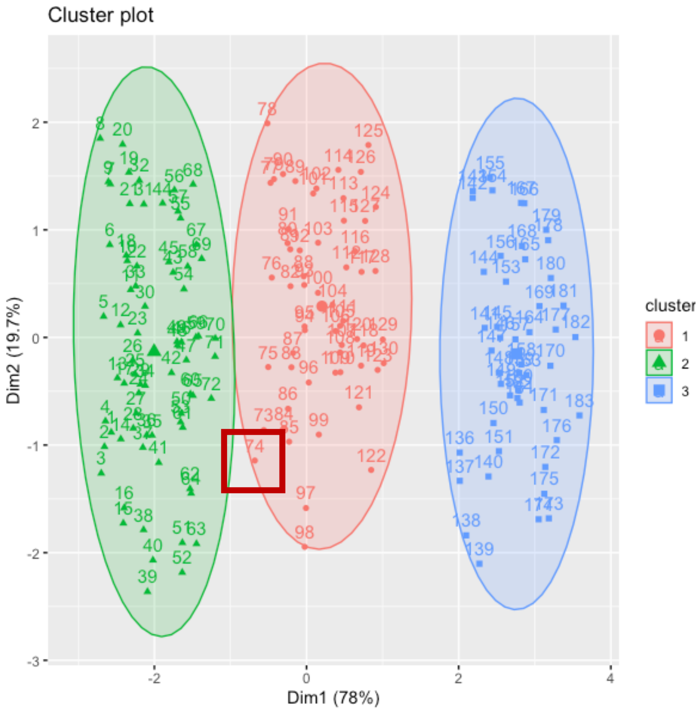


Figure 23. Cluster Plot of K-means Clustering

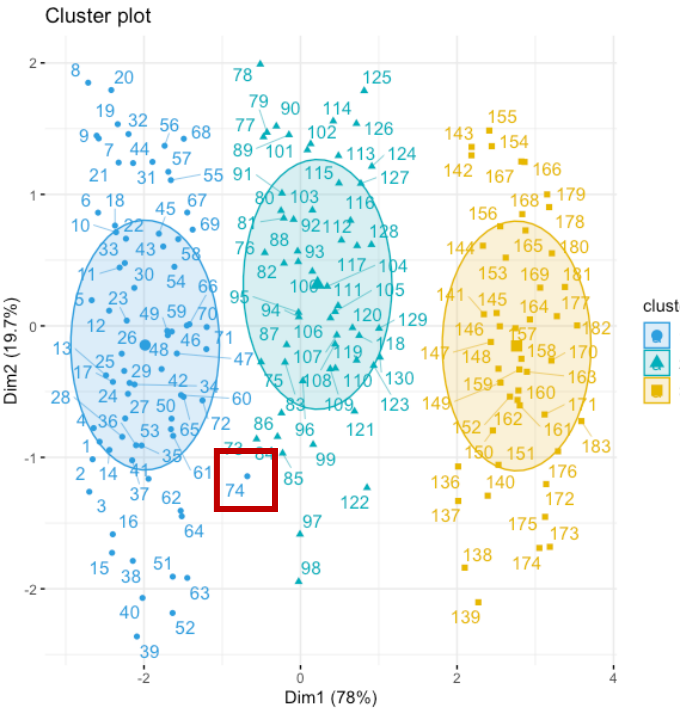


Figure 24. Cluster Plot of Model-based Clustering

5. Discussion and Conclusion

5.1. Conclusion

To sum up, PCA effectively reduces dataset redundancy and explains its variance by dividing variables into Carbon monoxide (CO) and greenhouse gases. The clustering analysis provides insights into the dataset pattern, revealing a trend of increasing greenhouse gases over time, alongside a decrease in CFC11 levels, with three distinct time periods.

5.2. Hypothesis

Theoretically, factor analysis is considered a more reliable method for explaining latent variables and providing direct and meaningful interpretations of new variables. However, in this data, significant strong relationships (approximately 0.9) exist. As a result, factor analysis may yield similar results, identifying Carbon monoxide (CO) and greenhouse gases as new dataset variables.

K-medoids clustering could be explored for clustering analysis, as it is more robust against the influence of outliers. Outliers were initially removed from the dataset. Therefore, k-means and k-medoids clustering may yield similar results. K-medoids clustering may demonstrate better performance working with the original dataset containing outliers.

5.3. Limitation and Further Plan

Interpretation limitations arise from a lack of expertise in atmospheric science and cannot be addressed solely through unsupervised learning methods.

The capability of PCA to explain the meaning of new variables is limited. Interpretations of principal components rely on personal knowledge. Future research could involve collaboration with atmospheric experts to elucidate the meaning and underlying information of new variables using factor analysis.

The explanations of clusters are concluded on intuition for pattern visualization. Further exploration of the reasons behind time-based clustering could benefit from the help of atmospheric domain specialists.

Appendix

```
MaunaLoa <- read.csv('/Users/huangdengwei/Downloads/MDS(DH)/DEVUL/Assignment 2/MaunaLoa.csv')
library(psych)
library(factoextra)
library(mclust)
library(cluster)
#------------------------------------------------------------------------------------------------------------------------------------
#------------------------------------------------------------------------------------------------------------------------------------
## 2. Exploratory Data Analysis
#------------------------------------------------------------------------------------------------------------------------------------
## 2.1. Features and Relationship
summary(MaunaLoa)
data <- MaunaLoa[,-1]
# Missing Values Checking
any(is.na(data))
# Figure 1. Scatter Plot Matrix of Monthly Averages of 5 Gases
pairs.panels(data, breaks = 20, method = "pearson", hist.col = "#00AFBB", density = TRUE, cex.main = 1,
        cex.cor = 1.5, ellipses = TRUE, main = "Scatter Plot Matrix of Monthly Averages of 5 Gases")
# Figure 2. Boxplot of Monthly Averages of 5 Gases
boxplot(MaunaLoa[,-1], col=2:6, cex.main = 1, main = "Boxplot of Monthly Averages of 5 Gases")
#------------------------------------------------------------------------------------------------------------------------------------
## 2.2. Outliers and Extremes
threshold <- 1.5
Q1 <- apply(data, 2, quantile, probs = 0.25)
Q3 <- apply(data, 2, quantile, probs = 0.75)
IQR_vals <- Q3 - Q1
outliers <- apply(data, 1, function(x) any(x < Q1 - threshold * IQR_vals | x > Q3 + threshold * IQR_vals))
clean_data <- data[!outliers, ]
#------------------------------------------------------------------------------------------------------------------------------------
## 2.3. Pre-processing
sc.data <- scale(clean_data)
# Figure 3. Scatter Plot Matrix after Omitting Outliers and Scaling
pairs.panels(sc.data, breaks = 20, method = "pearson", hist.col = "#00AFBB", density = TRUE, cex.main = 0.8,
        cex.cor = 1.5, ellipses = TRUE, main = "Scatter Plot Matrix of Monthly Averages of 5 Gases after Omitting Outliers and Scaling")
# Figure 4. Boxplot after Omitting Outliers and Scaling
boxplot(sc.data, col=2:6, cex.main = 0.9, main = "Boxplot of Monthly Averages of 5 Gases after Omitting Outliers and Scaling")
#------------------------------------------------------------------------------------------------------------------------------------
#------------------------------------------------------------------------------------------------------------------------------------
## 3. Dimension Reduction
#------------------------------------------------------------------------------------------------------------------------------------
# 3.3. Dimension Reducing
pr.out <- prcomp(sc.data)
# Figure 5. PCA Summary, Loadings and Scores
summary(pr.out)
pr.out$rotation
head(pr.out$x)
#------------------------------------------------------------------------------------------------------------------------------------
## 3.4. Number of New Variables Choosing
# Figure 6. Scree Plot of the Proportion of Variance Explained
fviz_screeplot(pr.out, addlabels = TRUE)
#------------------------------------------------------------------------------------------------------------------------------------
## 3.5. Relationship between the Original and New Variables
# Figure 7. Biplot of PCA
fviz_pca_biplot(pr.out, axes = c(1, 2), repel = TRUE, geom = "point", col.var = "red")
```

```
# Figure 8. Contribution of Variables to Dim-1
fviz_contrib(pr.out, choice = "var", axes = 1, top = 10)
# Figure 9. Contribution of Variables to Dim-2
fviz_contrib(pr.out, choice = "var", axes = 2, top = 10)
# Figure 10. Contribution Matrix; Figure 11. Correlation; Figure 12. Quality of Representation
var <- get_pca_var(pr.out)
var$contrib
var$cor
var$cos2
#---------------------------------------------------------------------------------------------------------------------------
#---------------------------------------------------------------------------------------------------------------------------
## 4. Cluster Analysis
#---------------------------------------------------------------------------------------------------------------------------
## 4.2. Method
# K-means Clustering
# Figure 13. Scree-like Plot of Optional number of Clusters
fviz_nbclust(sc.data, kmeans, method = "wss")
# Figure 14. Silhouette Measure Plot of K-means Clustering
fviz_nbclust(sc.data, kmeans, method = "silhouette") + labs(title = "K-means")
# Compute k-means with k = 3
set.seed(123)
km.res <- kmeans(sc.data, 3, nstart = 25)
# Figure 15. The Portion of Variability Explained
km.res
# Figure 16. The Means based on the Newly Created Clusters
aggregate(clean_data, by=list(cluster=km.res$cluster), mean)
#---------------------------------------------------------------------------------------------------------------------------
## Model-based Clustering
mc_bic <- Mclust(sc.data)
# Figure 17. Bayesian Information Criterion of Model-based Clustering
mc_bic$BIC
# Figure 18. BIC Plot of Model-based Clustering
plot(mc_bic, what = "BIC")
# Figure 19. Uncertainty Plot of Model-based Clustering
plot(mc_bic, what = "uncertainty")
#---------------------------------------------------------------------------------------------------------------------------
## 4.3. Result Comparison
# Figure 20. Classification Table of Two Methods
table(km.res$cluster, mc_bic$classification)
# Clusters Silhouette Plot
# Figure 21. Clusters Silhouette Plot of K-means Clustering
km.sil <- silhouette(km.res$cluster, dist(sc.data))
fviz_silhouette(km.sil, palette = "jco", ggtheme = theme_classic())
# Figure 22. Clusters Silhouette Plot of Model-based Clustering
mc.sil <- silhouette(mc_bic$classification, dist(sc.data))
fviz_silhouette(mc.sil, palette = "jco", ggtheme = theme_classic())
#---------------------------------------------------------------------------------------------------------------------------
## 4.4. Visualization and Interpretation
# Figure 23. Cluster Plot of K-means Clustering
fviz_cluster(km.res, clean_data, ellipse.type = "norm")
# Figure 24. Cluster Plot of Model-based Clustering
fviz_cluster(mc_bic, data = sc.data, palette = c("#2E9FDF", "#00AFBB", "#E7B800"),
        ellipse.type = "euclid", star.plot = FALSE, repel = TRUE, ggtheme = theme_minimal())
```