Data Exploration, Visualization, and Unsupervised Learning – Assignment 2

1. Exploratory Data Analysis

1.1 Dataset Evaluation

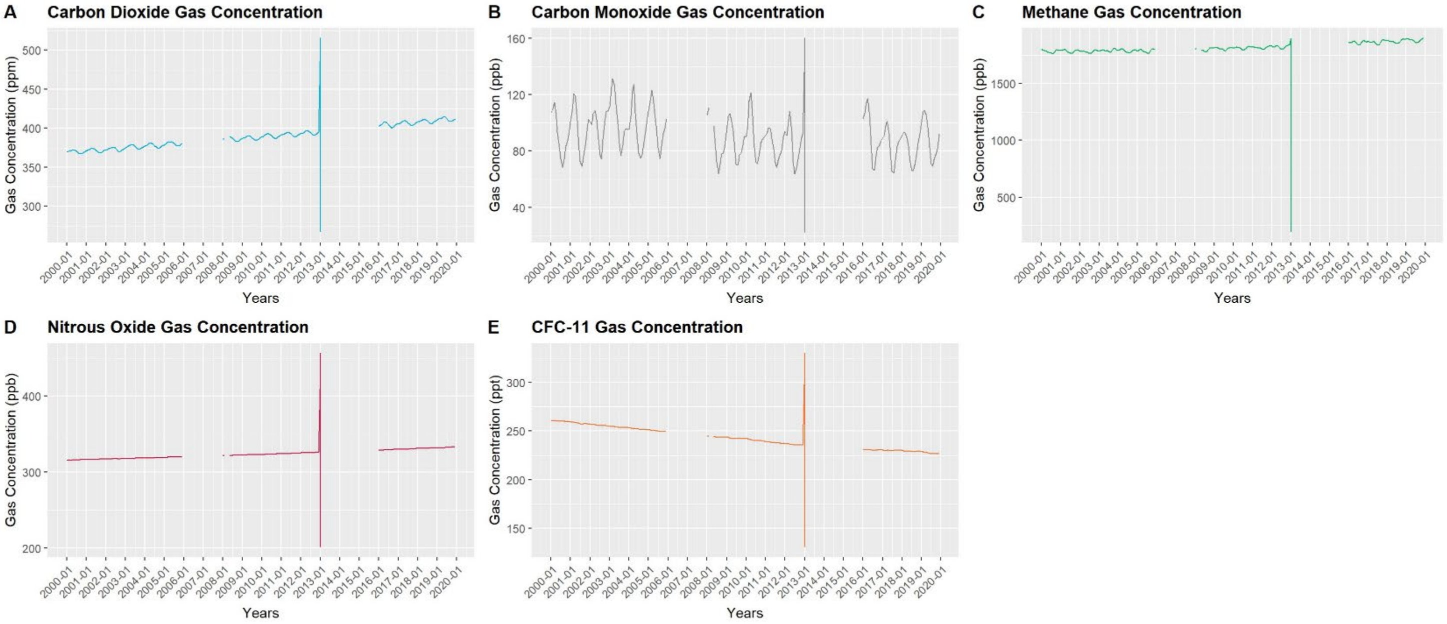
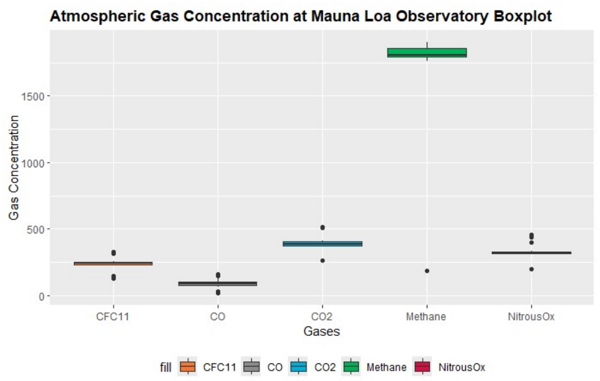


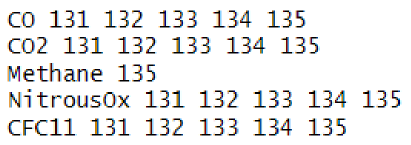
Figure 1: Time Series Plot of the Measured Atmospheric Gas Concentrations above Mauna Loa Observatory

This dataset contains the measured "selected" monthly average atmospheric gas at the Mauna Loa observatory from 2000 to 2019. The recorded date is formatted as year-month-day. Visualizing each of the gases revealed that all the gases have the same recorded period, resulting in time gaps between the data, Figure 1. The gaps are presented during 2006-2007, 2008, and 2013-2015. As the data contained "selected" months and no missing value rows (*NA* rows) are presented, we make an assumption that the time gap is intentionally made. This could be a result of measuring equipment maintenance at the Mauna Loa Observatory or other factors during data collection process. For this analysis, we will not impute the data gaps as it may introduce inaccuracies during analysis. Instead, we will acknowledge the presence of these time gaps within this analysis.

1.2 Outliers Detection and Evaluation

In addition to the gaps within the time series data, we also detect irregularity within Figure 1. During the year 2013, the plot reveals a drastic change in all gas's concentration level compared to other time periods. Through box plot, it revealed that these points are at row 131 to 135 of the original dataset, Figure 2. Looking into the dataset, we found out that from 2013-01-01 to 2013-01-05 have unusual format when compared to other observations within the dataset. It contains the first five days of the month, instead of the first day of every month. We suspected that these observations are a result of data collection misinput. As we could not investigate the data collecting methods to correct or verify these observations, we choose to remove these points (131:135) before proceeding to dimension reduction and cluster analysis. Without this step, the subsequent analysis will lack consistency as the data is not in the same Date format.





```
CO 131 132 133 134 135
CO2 131 132 133 134 135
Methane 135
NitrousOx 131 132 133 134 135
CFC11 131 132 133 134 135
```

Figure 2: Boxplot of Atmospheric Gas Data (Left) and Boxplot Potential Outliers (Right)

1.3 Data Analysis

With outliers removed, we proceed to the data analysis. A plot of the relative change in the gas's concentration was plotted in Figure 3. Through Figure 1 and Figure 3, we identified that carbon dioxide carbon monoxide, and methane gas have seasonal variations as they are periodically repeated. On the other hand, nitrous oxide and CFC-11 do not have seasonal variations. Viewing the relative change, carbon dioxide, nitrous oxide, and methane gas increases over the period, whereas the CFC-11 and CO trend is decreasing. The pairs plot, Figure 3 right, reveals that carbon dioxide, nitrous oxide, methane gas, and CFC-11 are highly correlated together, with CFC-11 negatively correlated to other gases. On the other hand, carbon monoxide does not have a high correlation with other gases. This result implied that these highly correlated variables have redundant information. We could apply dimension reduction techniques such as PCA and factor analysis to reduce the data dimension.

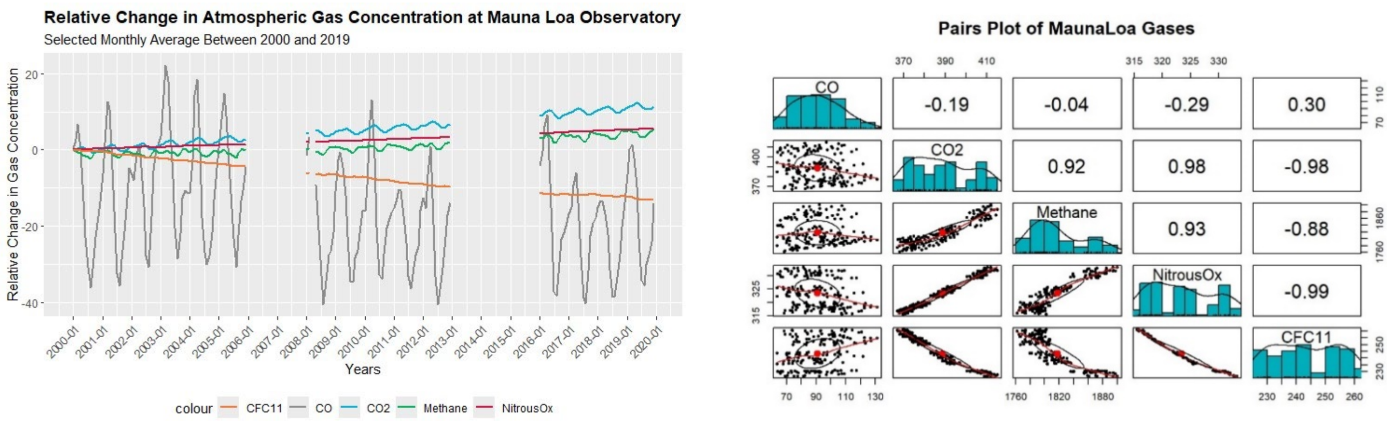


Figure 3: Relative Change in Atmospheric Gas Concentration (Left) and Gases Pairs Plot (Right) After Unusal Values Removal

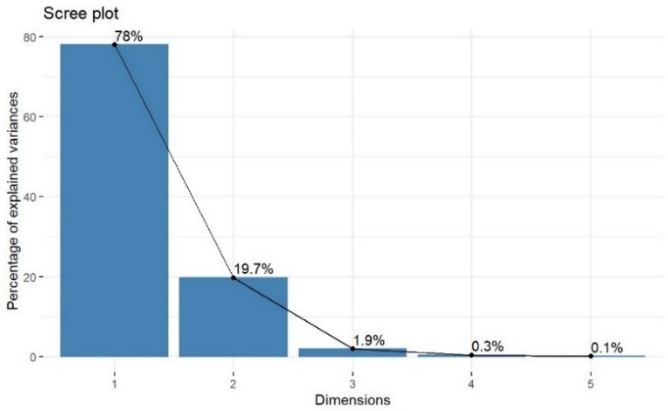
2. Dimension Reduction

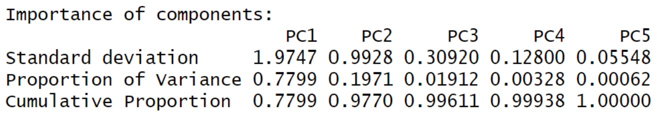
2.1 Factor Analysis vs Principal Component Analysis

With our atmospheric gases concentration data at Mauna Loa, we have the choice of Factor Analysis (FA) or Principal Component Analysis (PCA) to dimensionally reduce the five gas variables into a smaller number of variables. We select FA over the PCA method as we make an assumption that there are underlying latent variables, that influence the observed gas variables. By choosing FA, we aim to both reduce the dimensionality of the gas dataset and gain insights into the latent variables, which drive the high correlation between them as observed in the exploratory data analysis pairs plot, Figure 3. While PCA reduces variables into multiple principal components (PCs) and provides insights on which original variables contribute to the PCs in the contributions plot, it does not directly provide a relationship interpretation of those original variables. In contrast, FA provides the relationship between the factors and the original variables, which are negative, positive, or no contribution. Therefore, in this analysis, we select FA over the PCA method.

2.2 Step-by-Step Approach Towards Factor Analysis

1. We select all the columns except for the "Date" column from the Mauna Loa and apply "Scale" to make all the gases have the same scale unit as originally, they are in ppm, ppb, and ppt units.
2. We then perform PCA to determine the number of factors to be considered in the Factor Analysis. We plot the cumulative proportion of the variances and the Scree plot, Figure 4. From the Scree Plot and the Importance of components, we identify that 2 Factors (97.7 % cumulative proportion) are the optimal number as it can explain more than 80% of the variance of this gas's dataset.





```
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
```

Figure 4: PCA Scree Plot to Determine the Optimal Number of Factors (Left) and PCR Summary (Right)

3. With our goal to provide insights into the underlying relationship, we evaluate rotation patterns: No, Varimax, and Promax rotation patterns within the "factanal" function to select the method that simplifies the factors, Figure 5. From Figure 5, we select "Promax rotation" as this rotation method results in interpretable loadings. Five loadings for this rotation lie closely to either Factor 1 or Factor 2 when compared to other rotations.

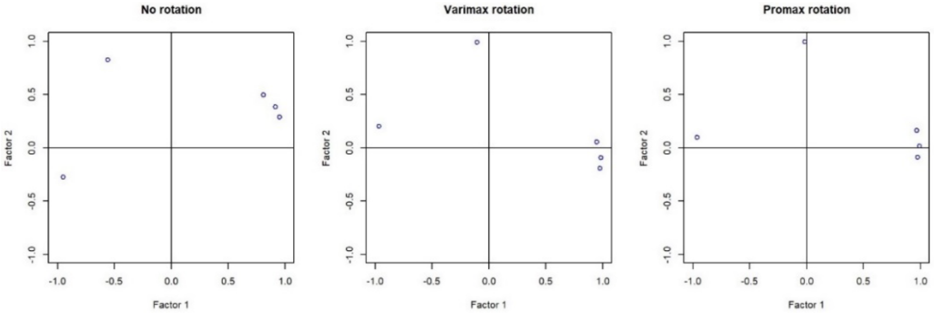


Figure 5: Evaluation of Rotation Methods for Two Factor Analysis: No (Left), Varimax (Middle), and Promax Rotation (Right)

4. We dimensionally reduced gases variables with FA using 2 factors and "Promax rotation" with "factanal" function. The FA summary and original variable projections are provided in Figure 7, left and right, accordingly.

2.3 Factor Analysis Result Interpretation

In this Mauna Loa gases dataset, 2 Factors is used to represent 5 gases. The factors are rotated with promax rotation to provides a better interpretation of the loading results, Figure 7. This choice of two factors is supported by the FA results as the cumulative proportion of the variance explained is 96.9%, which is higher than 80%. The importance of factors is visualized in Figure 6 left. The FA uniqueness results revealed that all the gas variance is mostly accounted for by the two factors as the uniqueness value is low for all gases. The test of hypothesis revels that these two factors are sufficient to explain the variance within the gas dataset. However, the low p-value of the likelihood ratio test suggests that the model is unlikely to adequately explain the data.

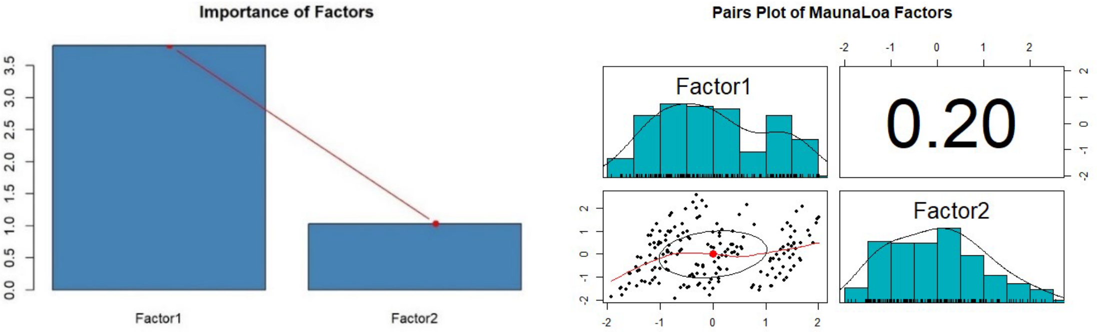
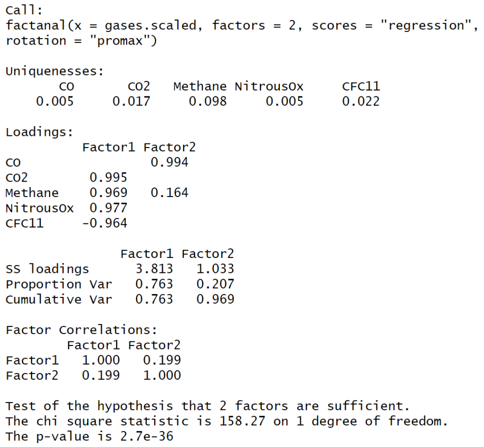


Figure 6: Importance of Factors Plot (Left) and Factors Pairs Plot (Right)

As for the loadings, the first factor reveals that carbon dioxide, methane, and nitrous oxide have a strong positive relationship and CFC-11 has a negative relationship. On the other hand, carbon monoxide has no association with the first factor. As for the second factor, only carbon monoxide and methane have an association with it with carbon monoxide have a high association. The projection of the original variables on Factor 1 and Factor 2, Figure 7, visualize these loading results. From these two factors, we interpret that "Factor 1" could be labeled as "Green House Gases" as they are gases that contribute to the global warming (United States Environmental Protection Agency, no date). As for CFC-11, which negatively associated with the first factor, it is a refrigerant for refrigerators and air conditioners that was internationally banned in 2010 with the Montreal Protocol (Montzka et al., 2021). Therefore, during the exploratory data analysis it decreases over the recorded period. On the other hand, "Factor 2" could be labeled as "Incomplete Combustion" as carbon monoxide primarily produced by incomplete combustion of fossil fuels.



```
Call:
factanal(x = gases.scaled, factors = 2, scores = "regression",
rotation = "promax")

Uniquenesses:
       CO       CO2   Methane NitrousOx     CFC11 
    0.005     0.017     0.098     0.005     0.022 

Loadings:
          Factor1 Factor2
CO                 0.994 
CO2        0.995         
Methane    0.969   0.164 
NitrousOx  0.977         
CFC11     -0.964         

               Factor1 Factor2
SS loadings      3.813   1.033
Proportion Var   0.763   0.207
Cumulative Var   0.763   0.969

Factor Correlations:
        Factor1 Factor2
Factor1   1.000   0.199
Factor2   0.199   1.000

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 158.27 on 1 degree of freedom.
The p-value is 2.7e-36
```

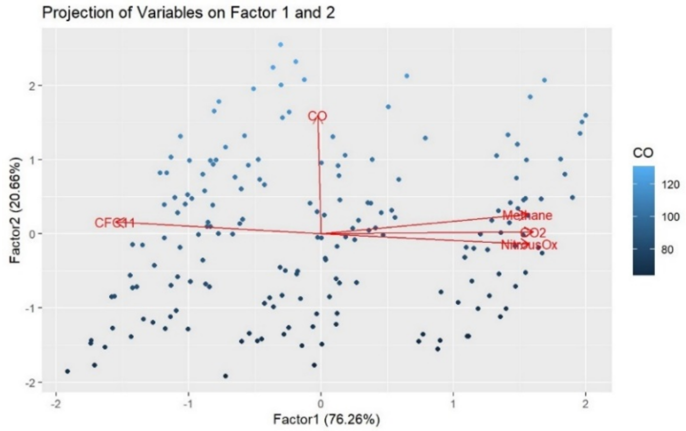


Figure 7: Factor Analysis Results (Left) and Projection for Variables on Factor 1 and 2 (Right)

3. Cluster Analysis

Through data exploration, we identified that our time series data contains three separate recorded periods, suggesting a potential groups within our dataset. Therefore, we utilize the clustering method to provide insights into the structure of this gas's concentration across this period. For this analysis, k-means and hierarchical clustering were selected. We selected the k-means clustering algorithm due to its simplicity and efficiency. While this method is known to be sensitive to outliers, we have addressed this concern through the removal of outliers during the data exploration process. Additionally, we utilized a hierarchical clustering algorithm to create a hierarchy of observations based on dissimilarity between observations. This choice is to discover structure within the data. To select an optimal number of clusters (k), we utilized both the total within-cluster sum of squares (SSE) plot and silhouette measurement plot for these algorithms.

3.1 K-means Clustering

For k-means clustering, we determine the number of clusters through the SSE within the cluster plot and the silhouette plot, Figure 8. The SSE plot reveals an elbow pattern at k = 3. As for the silhouette plot, it recommended k = 2 to maximize the silhouette width, a value closer to 1 indicates that a point is close to its assigned cluster. However, we select k = 3, as the silhouette width between k = 2 and k = 3 is comparable but at k = 3 the SSE is lower.

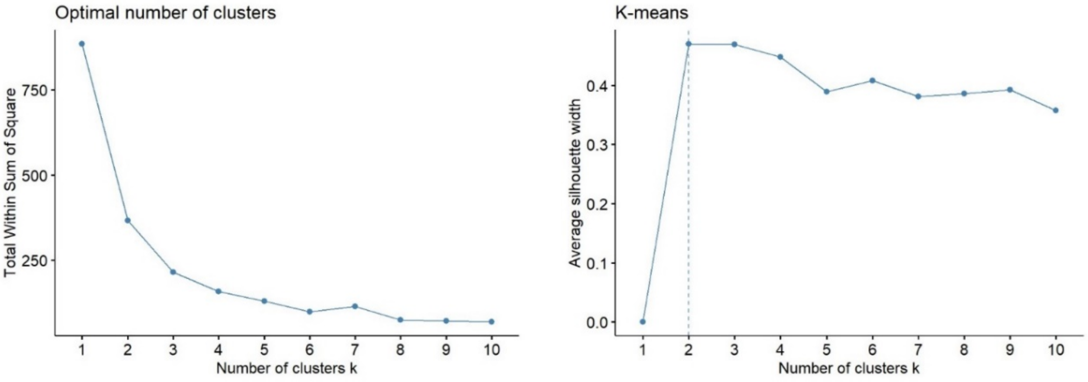


Figure 8: K-means Selection of Optimal Cluster Size: Total Sum of Square within Clusters (Left) and Average Silhouetle Width (Right)

With k = 3, we utilized "kmeans" function to create clusters, the cluster is visualized in Figure 9 (Left) and the average silhouette width for the three clusters is provided (Right).

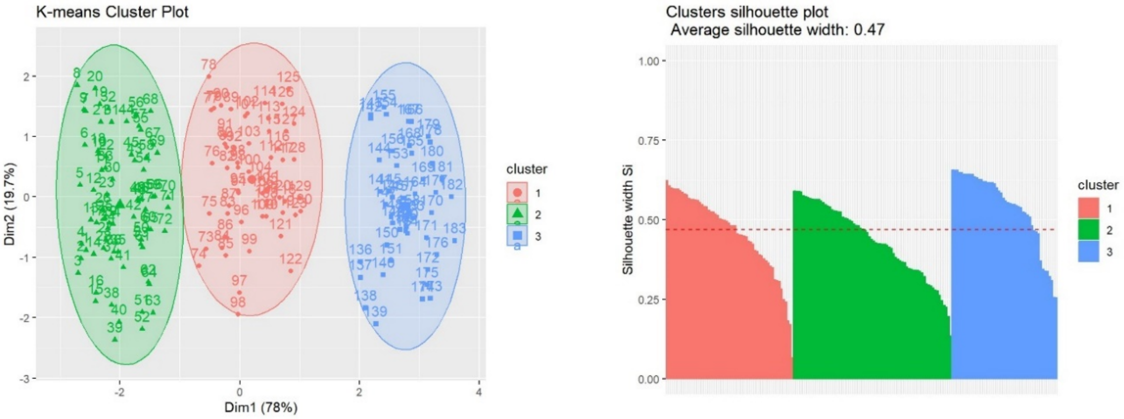
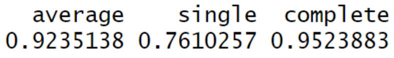


Figure 9: K-means Cluster (k = 3) Plot (Left) and Average Silhouette Width (Right)

3.2 Hierarchy Clustering

For Hierarchy clustering, we first explore the linkage for the gas's dataset between average, single, and complete linkage types. We select the linkage type through the evaluation of the agglomerative coefficient (AC) which measures the dissimilarity between objects. A value close to 1 indicates that the cluster structure is more balanced. In this analysis, we select the "complete" linkage type as it provides the highest AC coefficient, Figure 10.



```
  average    single  complete 
0.9235138 0.7610257 0.9523883 
```

Figure 10: Hierarchy Clustering Linkage Type

With the linkage type selected, we then select the number of clusters through SSE and silhouette width. For SSE, the plot revealed a similar elbow pattern at k = 3, Figure 11. However, the average silhouette plot indicates that the optimal number of k is 2. An increasing number of k clusters result in a decrease in average silhouette width. Therefore, we will be using k = 2 for hierarchical clustering algorithm. We utilized "hcut" function to create clusters with "complete linkage" and k = 2. Hierarchical cluster and cluster silhouette plot is provided in Figure 12.

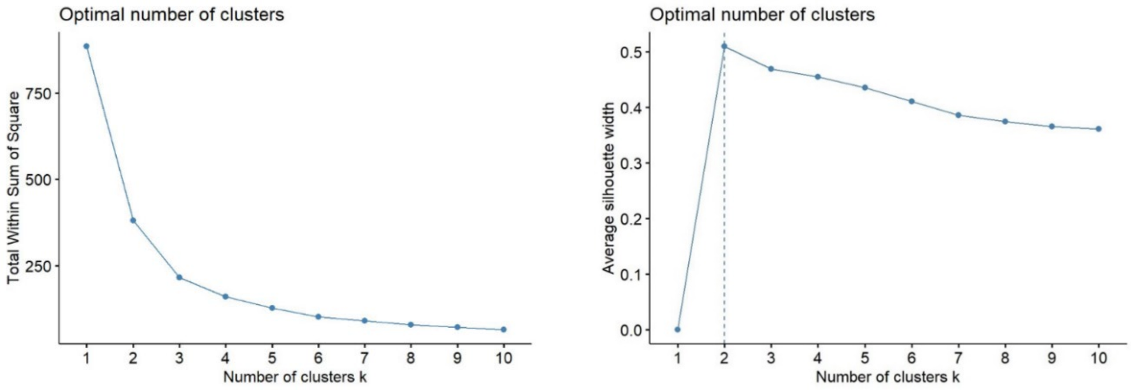


Figure 11: Hierarchy Selection of Optimal Cluster Size: Total Sum of Square within Clusters (Left) and Average Silhouetle Width (Right)

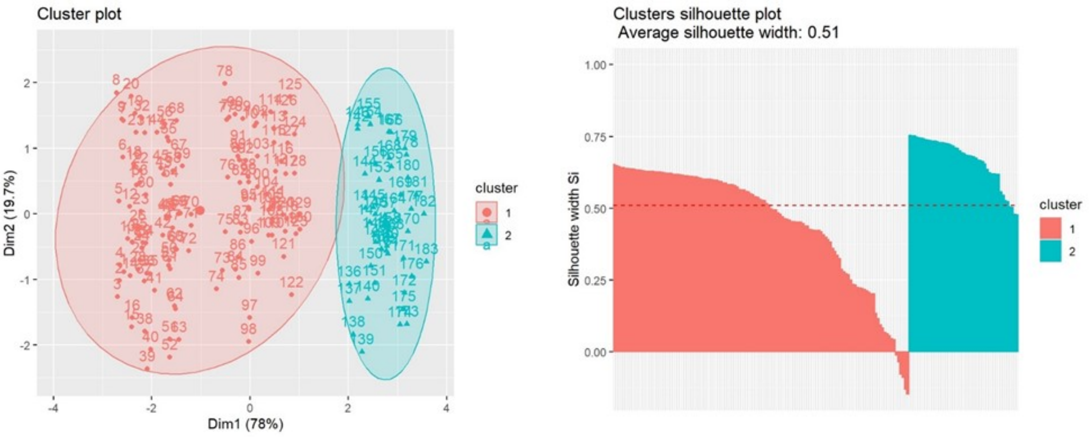
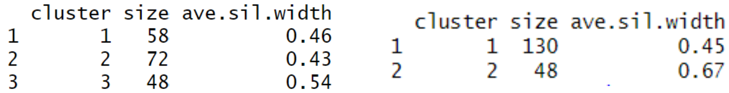


Figure 12: Hierarchical Cluster (k = 2 with complete linkage) Plot (Left) and Average Silhouette Width (Right)

3.3 Cluster Result Analysis

For cluster analysis, we utilized k-means and hierarchical cluster algorithms. Although the SSE within clusters for both algorithms reveals a similar elbow pattern at k = 3, the silhouette plot suggests a different number of clusters with k = 3 for k-means and k = 2 for hierarchical.

Visualizing the clusters, k-means identifies 3 clusters that correspond to the three time periods, which were visualized during the data exploration, 2000 to 2006, 2008 to 2012, and 2016 to 2019, Figure 13. On the other hand, hierarchical algorithm group the first two periods together, while the third period remains the same as in k-means. Using the silhouette measurement to validate these unlabeled clusters, we found that hierarchical clusters result in a higher average silhouette width at 0.51 when compared to k-means at 0.47, indicating a better separation. However, the first cluster of hierarchical clusters has some points having a negative score, indicating that they might be mis-clustered.



```
  cluster size ave.sil.width
1       1   58          0.46
2       2   72          0.43
3       3   48          0.54
```

```
  cluster size ave.sil.width
1       1  130          0.45
2       2   48          0.67
```

Figure 13: K-means (Left) and Hierarchy (Right) Silhouette Width within Cluster

4. Discussion and Conclusion

Through factor analysis, we identify two factors that represent the gases: greenhouse gases, (factor 1) and incomplete combustion (factor 2). FA revealed that carbon dioxide, methane, and nitrous oxide have a strong positive relation with the first factor, whereas CFC-11 has a negative relationship. As for incomplete combustion gases, carbon monoxide has a high positive relationship, while methane gas has a low positive relationship. These factors have a slightly positive relationship with one another, possibly due to the indirect influence of incomplete combustion gases on atmospheric gas chemistry. Alternatively, if we selected PCA instead of FA, our result would likely reveal a similar pattern with two principal components and loading variables. However, PCA would not provide information on the relationship of gas variables towards PCs and the relationship between the PCs. This is due to the aim of PCA to capture the highest variance in the data, whereas FA focuses on explaining the correlation among variables.

For clustering analysis, k-mean results suggest three clusters, whereas the hierarchical clustering identified two clusters. We attribute this difference in the number of clusters to both the gaps within our data and the distinct mechanisms between these algorithms. K-mean aims to minimize within-cluster SSE. The presence of gaps between observations would lead to clusters group around these gaps, resulting in three clusters. On the other hand, hierarchical clustering is based on dissimilarity between observations. The size of the gap potentially influences the cluster as the smaller gap between 2006 to 2008 has resulted in one large cluster. Alternatively, if we have selected k-medoids instead of k-means, the clusters would be different. K-medoids center the cluster around the data points, potentially resulting in one cluster for 2016-2019 and multiple clusters between 2000-2012.

In conclusion, this analysis provides insights into the underlying relationship of selected atmospheric gases at the Mauna Loa observatory and its inherited structure. Through data exploration, we have identified and removed unusual values in 2013. With FA, we identify greenhouse incomplete combustion gases as the factors for this dataset. Clustering analysis reveals the influence of the gaps and the algorithm differences in cluster structures. One question that remains open is the underlying causes of the gap within the time series data and the unusual values. Future studies should investigate these issues, especially the presence of the time gaps, as they directly influence our analysis. Moreover, these analyses are limited to data exploration it could not be used to model future trends for these gases' concentrations.

References:

Montzka, S.A. *et al.* (2021) ‘A decline in global CFC-11 emissions during 2018−2019’, *Nature*, 590(7846), pp. 428–432. doi:10.1038/s41586-021-03260-5.

United States Environmental Protection Agency (no date) *Overview of Greenhouse Gases*, *EPA*. Available at: https://www.epa.gov/ghgemissions/overview-greenhouse-gases (Accessed: 21 March 2024).

Appendix:

```
## DEVUL - Assignment 2
## library Imports
library(ggplot2)
library(ggpubr)
library(visdat)
library(scales)
library(psych)
library(factoextra)
library(cluster)
library(tidyverse)

## 0. Data Loading and Pre-processing
MaunaLoa <- read.table("File_Directory",
              sep = ",", header = 1)

MaunaLoa$Date <- as.Date(MaunaLoa$Date, format = "%Y-%m-%d") # Format as a data format

## Fill in the skipped months
every_month_first_day <- data.frame(Date = seq(as.Date("2000-01-01"), # Min
                                    as.Date("2019-12-01"), # Max
                                    by = "month")) # Create dataframe that contains only  day of the month
MaunaLoa_NA_Gaps <- merge(every_month_first_day,
                    MaunaLoa,
                    by = "Date",
                    all.x = TRUE)
MaunaLoa_NA_Gaps$Date <- as.Date(MaunaLoa_NA_Gaps$Date, format = "%Y-%m-%d")

## Add 132:135 from original dataset
strange_records <- MaunaLoa[132:135,] # In the original dataset
strange_records$Date <- as.Date(strange_records$Date, format = "%Y-%m-%d")
strange_records # Strange records
MaunaLoa_NA_Gaps <- rbind(MaunaLoa_NA_Gaps, strange_records) # bind them
MaunaLoa_NA_Gaps <- MaunaLoa_NA_Gaps[order(MaunaLoa_NA_Gaps$Date), ] # Sort by values
rownames(MaunaLoa_NA_Gaps) <- NULL # Reset index start from 0
```

```
## 1. Exploratory Data Analysis (15 %)

## Visualize the Missing Values
missing_val_plot <- vis_miss(MaunaLoa_NA_Gaps)
missing_val_plot

## Standardize each Gases Showing relative Change
MaunaLoa_NA_Gaps$CO_rel <- 100 * (MaunaLoa_NA_Gaps$CO/MaunaLoa_NA_Gaps$CO[1]-1)
MaunaLoa_NA_Gaps$CO2_rel <- 100 * (MaunaLoa_NA_Gaps$CO2/MaunaLoa_NA_Gaps$CO2[1]-1)
MaunaLoa_NA_Gaps$Methane_rel <- 100 * (MaunaLoa_NA_Gaps$Methane/MaunaLoa_NA_Gaps$Methane[1]-1)
MaunaLoa_NA_Gaps$NitrousOx_rel <- 100 * (MaunaLoa_NA_Gaps$NitrousOx/MaunaLoa_NA_Gaps$NitrousOx[1]-1)
MaunaLoa_NA_Gaps$CFC11_rel <- 100 * (MaunaLoa_NA_Gaps$CFC11/MaunaLoa_NA_Gaps$CFC11[1]-1)

## Plot Standardize Series - Showing Relative Change
ggplot(data = MaunaLoa_NA_Gaps, aes(x = Date)) +
  geom_line(aes(y = CO_rel, color = "CO"), linewidth = 1) +
  geom_line(aes(y = CO2_rel, color = "CO2"), linewidth= 1) +
  geom_line(aes(y = Methane_rel, color = "Methane"), linewidth = 1) +
  geom_line(aes(y = NitrousOx_rel, color = "NitrousOx"), linewidth = 1) +
  geom_line(aes(y = CFC11_rel, color = "CFC11"), linewidth = 1) +
  scale_color_manual(values = c("CO" = "#8c8c8c", "CO2" = "#00aedb", "Methane" = "#00b159",
                                "NitrousOx" = "#d11141", "CFC11" = "#f37735")) +
  labs(title = "Relative Change in Atmospheric Gas Concentration at Mauna Loa Observatory",
       subtitle = "Selected Monthly Average Between 2000 and 2019",
       x = "Years",
       y = "Relative Change in Gas Concentration") +
  theme(plot.title = element_text(face = "bold"),
        legend.position = "bottom",
        axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_x_date(labels = date_format("%Y-%m"), date_breaks = "1 year")

## Visualize the time series as a year - After Removal of Strange Values
MaunaLoa_Remove_Strange <- MaunaLoa_NA_Gaps[-c(157:161),]
ggplot(data = MaunaLoa_Remove_Strange, aes(x = Date)) +
  geom_line(aes(y = CO_rel, color = "CO"), linewidth = 1) +
  geom_line(aes(y = CO2_rel, color = "CO2"), linewidth= 1) +
  geom_line(aes(y = Methane_rel, color = "Methane"), linewidth = 1) +
  geom_line(aes(y = NitrousOx_rel, color = "NitrousOx"), linewidth = 1) +
  geom_line(aes(y = CFC11_rel, color = "CFC11"), linewidth = 1) +
  scale_color_manual(values = c("CO" = "#8c8c8c", "CO2" = "#00aedb", "Methane" = "#00b159",
                                "NitrousOx" = "#d11141", "CFC11" = "#f37735")) +
  labs(title = "Relative Change in Atmospheric Gas Concentration at Mauna Loa Observatory",
       subtitle = "Selected Monthly Average Between 2000 and 2019",
       x = "Years",
       y = "Relative Change in Gas Concentration") +
  theme(plot.title = element_text(face = "bold"),
        legend.position = "bottom",
        axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_x_date(labels = date_format("%Y-%m"), date_breaks = "1 year")
```

```
## Time Series One by One - Adding NA Gaps between skipped months
## CO
mis_CO <- ggplot(data = MaunaLoa_NA_Gaps, aes(x = Date, y = CO)) +
  geom_line(color = "#8c8c8c")+
  labs(title = "Carbon Monoxide Gas Concentration",
       x = "Years",
       y = "Gas Concentration (ppb)") +
  theme(plot.title = element_text(face = "bold"),
        axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_x_date(labels = date_format("%Y-%m"), date_breaks = "1 year")

## CO2
mis_CO2 <- ggplot(data = MaunaLoa_NA_Gaps, aes(x = Date, y = CO2)) +
  geom_line(color = "#00aedb") +
  labs(title = "Carbon Dioxide Gas Concentration",
       x = "Years",
       y = "Gas Concentration (ppm)") +
  theme(plot.title = element_text(face = "bold"),
        axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_x_date(labels = date_format("%Y-%m"), date_breaks = "1 year")

## Methane
mis_Met <- ggplot(data = MaunaLoa_NA_Gaps, aes(x = Date, y = Methane)) +
  geom_line(color = "#00b159") +
  labs(title = "Methane Gas Concentration",
       x = "Years",
       y = "Gas Concentration (ppb)") +
  theme(plot.title = element_text(face = "bold"),
        axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_x_date(labels = date_format("%Y-%m"), date_breaks = "1 year")

## NitrousOx
mis_NOx <- ggplot(data = MaunaLoa_NA_Gaps, aes(x = Date, y = NitrousOx)) +
  geom_line(color = "#d11141") +
  labs(title = "Nitrous Oxide Gas Concentration",
       x = "Years",
       y = "Gas Concentration (ppb)") +
  theme(plot.title = element_text(face = "bold"),
        axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_x_date(labels = date_format("%Y-%m"), date_breaks = "1 year")

## CFC11
mis_cfc <- ggplot(data = MaunaLoa_NA_Gaps, aes(x = Date, y = CFC11)) +
  geom_line(color = "#f37735") +
  labs(title = "CFC-11 Gas Concentration",
       x = "Years",
       y = "Gas Concentration (ppt)") +
  theme(plot.title = element_text(face = "bold"),
        axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_x_date(labels = date_format("%Y-%m"), date_breaks = "1 year")

## Plot them all
missing_gases <- ggarrange(mis_CO2, mis_CO, mis_Met,
                           mis_NOx, mis_cfc,
                             labels = c("A", "B", "C", "D", "E"),
                             ncol = 3, nrow = 2)
missing_gases
```

```
## Plot Boxplot
ggplot(data = MaunaLoa) +
  geom_boxplot(mapping = aes(x = "CO", y = CO, fill = "CO")) +
  geom_boxplot(mapping = aes(x = "CO2", y = CO2, fill = "CO2")) +
  geom_boxplot(mapping = aes(x = "Methane", y = Methane, fill = "Methane")) +
  geom_boxplot(mapping = aes(x = "NitrousOx", y = NitrousOx, fill = "NitrousOx")) +
  geom_boxplot(mapping = aes(x = "CFC11", y = CFC11, fill = "CFC11"))+
  scale_fill_manual(values = c("CO" = "#8c8c8c", "CO2" = "#00aedb", "Methane" = "#00b159",
                               "NitrousOx" = "#d11141", "CFC11" = "#f37735"))+
  labs(title = "Atmospheric Gas Concentration at Mauna Loa Observatory Boxplot",
       x = "Gases",
       y = "Gas Concentration") +
  theme(plot.title = element_text(face = "bold"),
        legend.position = "bottom")

## Get the outlines function
get_q_outliers <- function(data, variable){
  ## Outliers = Q1 - 1.5*IQR or Q3 + 1.5*IQR
  outlier_quantile <- boxplot.stats(data[[variable]])$out
  outlier_idx <- which(data[[variable]] %in% outlier_quantile) # Get indexes
  return(outlier_idx)
}

gases_names <- names(MaunaLoa)[-1] # get the gases names
print("Potential Outliers")
for (gas in gases_names){ # loop and get outliers
  gas_outliers <- c(get_q_outliers(MaunaLoa, gas))
  cat(gas, gas_outliers, "\n") # Print out outliers
}

## Create new dataset by removing these outliers
MaunaLoa_remove_outliers <- MaunaLoa[-c(131:135),]

## Plotting Pairs after the removal of outliers
pairs.panels(MaunaLoa_remove_outliers[,-1],
             method = "pearson",
             hist.col = "#00AFBB",
             density = TRUE,
             ellipses = TRUE,
             main = "Pairs Plot of MaunaLoa Gases"
)
```

```
## 2. Dimension Reduction (30 %) - Factor Analysis
## 2.1 Using the data without the "Date" columns and Scale the Data
gases_no_outliners <- MaunaLoa_remove_outliers[,-1] # Only select the gases columns
gases.scaled <- scale(gases_no_outliners) # Scale the gases

## 2.2 Using PCA to determine the numbers of Factors
gases.pr <- prcomp(gases.scaled) # data already scaled
summary(gases.pr) # we select 2 Factors
gases.pr

fviz_screeplot(gases.pr,
            addlabels = TRUE)

plot(summary(gases.pr)$importance[2,],
    type = "b",
    xlab ="PCs",
    ylab = "Variability explained",
    main = "Gases Scree Plot") # Visualization
## We select 2 factors

## 2.3 Select Rotation Type
gases.fa.none <- factanal(gases.scaled, factors = 2, rotation = "none", scores="regression")
gases.fa.varimax <- factanal(gases.scaled, factors = 2, rotation = "varimax", scores="regression")
gases.fa.promax <- factanal(gases.scaled, factors = 2, rotation = "promax", scores="regression")

## Visualize different Rotation Types
par(mfrow = c(1, 3))
plot(gases.fa.none$loadings[, 1], gases.fa.none$loadings[, 2],
    xlab = "Factor 1", ylab = "Factor 2", ylim = c(-1, 1), xlim = c(-1, 1),
    col= "blue", main = "No rotation")
abline(h = 0, v = 0)

plot(gases.fa.varimax$loadings[, 1], gases.fa.varimax$loadings[, 2],
    xlab = "Factor 1", ylab = "Factor 2", ylim = c(-1, 1), xlim = c(-1, 1),
    col= "blue", main = "Varimax rotation")
abline(h = 0, v = 0)

plot(gases.fa.promax$loadings[, 1], gases.fa.promax$loadings[, 2],
    xlab = "Factor 1", ylab = "Factor 2", ylim = c(-1, 1), xlim = c(-1, 1),
    col= "blue", main = "Promax rotation")
abline(h = 0, v = 0)
par(mfrow = c(1, 1))

## 2.4 Evaluation
gases.fa.promax  # We select "Promax" rotation
gases.fa.promax$uniquenesses # Uniqueness are unexplained variability proportion by the factors.
gases.fa.promax$loadings # Showing contributions of variables into these two factors

## Variance Explained
L <- gases.fa.promax$loadings
SS <- apply(L^2, 2, sum)
barplot(SS, col='steelblue' , main="Importance of Factors")
lines(x = c(.75,2), SS, type="b", pch=19, col = "red")

## Visualize the loadings
autoplot(gases.fa.promax, data = MaunaLoa_remove_outliers,
       loadings = TRUE, loadings.label = TRUE,
       loadings.label.size  = 3.5, color = "CO",
       main = "Projection of Variables on Factor 1 and 2")

## Pairs of the two factors
gases.fa <- as.data.frame(gases.fa.promax$scores)
pairs.panels(gases.fa,
          method = "pearson",
          hist.col = "#00AFBB",
          density = TRUE,
          main = "Pairs Plot of MaunaLoa Factors"
)
```

```
## 3. Cluster Analysis (30 %) - kmeans and cluster
## 3.1 Scaled the data
MaunaLoa_Scaled <- scale(MaunaLoa_remove_outliers[-1]) # Select only variables

## 3.2 k-Mean Clustering
## 3.2.1 Select the best K
fviz_nbclust(MaunaLoa_Scaled,
             kmeans,
             method = "wss") # Elbow pattern at 3 clusters
fviz_nbclust(MaunaLoa_Scaled,
             kmeans,
             method = "silhouette")+ # Avg Silhouette width recommend at 2 clusters
  labs(title = "K-means")

## 3.2.2 Best k k-means
set.seed(123) # reproducibility
kmean_output <- kmeans(MaunaLoa_Scaled,
                       centers=3, # we select 3 clusters as Avg Silhouette width is similar for 2 and 3 clusters
                       nstart=25)

fviz_cluster(kmean_output, MaunaLoa_Scaled, ellipse.type = "norm")+
  labs(title = "K-means Cluster Plot") # Plot the clusters

## 3.2.3 Best k-Mean Silhouette
kmean_sil <- silhouette(kmean_output$cluster, dist(MaunaLoa_Scaled))
fviz_silhouette(kmean_sil)

## 3.3 Hierarchical Clustering
## 3.3.1 Determine the linkage type
link_methods <- c("average","single","complete")
names(link_methods) <- c("average","single","complete")
ac <- function(x){
  agnes(MaunaLoa_Scaled, method = x)$ac
}
map_dbl(link_methods,ac) # call the function
## Complete linkage is the best

## 3.3.2 Determine the number of clusters
fviz_nbclust(MaunaLoa_Scaled, hcut, method = "wss")
fviz_nbclust(MaunaLoa_Scaled, hcut, method = "silhouette")

## 3.3.3 Create hcut
hcut_output <- hcut(MaunaLoa_Scaled, k = 2, hc_method = "complete")
fviz_cluster(hcut_output, ellipse.type = "norm")
fviz_dend(hcut_output, cex = 0.5, color_labels_by_k = TRUE, ggtheme = theme_gray(),
          main = "Hierarchical Cluster Dendrogram"
)

## 3.3.4 Best k-cluster Silhouette
fviz_silhouette(hcut_output)
```