GENERAL SIR JHON KOTELAWALA DEFENCE UNIVERSITY

B.Sc. (Hons) Data Science and Business Analytics

INTAKE 38 SEMESTER VI

CM 3052 – Multivariate Data Analysis

Student Name	Reg no
D N R Peiris	D/DBA/21/0026
Lecturer in-charge Dr. Niroshan Withanage	

Submission Date-20.10.2023

Table of Contents

Introduction

Research Area and Background

The research topic pertains to the analysis of the chemical components of metals found in 92 drinking wells across the United States. Samples were obtained from two regions known as Maine and New Hampshire. This study was carried out within the framework of the "Data Action" project, which involved the participation of students and teachers from Maine (a rural area) and New Hampshire (an urban area). Well water was collected from neighborhoods and summer camps between spring 2019 and winter 2020. Water samples were collected using established scientific methods, and all metadata related to the water samples was gathered by the students. The samples were frozen for approximately 24 hours to minimize interference from microorganisms. The following metal components were analyzed from the 92 randomly selected water samples.

Be, Cr, Fe, Ni, Cu, As, Cd, Ba, Ti, Pb and U

Objectives of the study

The Following objectives are supposed to be covered in this analysis.

1. The eleven components are summarized into small no of subgroups.
2. Water samples are clustered into homogeneous groups according to the structure of the mixer components.
3. Check chemical mixtures in well water samples are in line with the standard accepted values in well water samples in United states.

Structure of the Report

This report will cover the methodology, Data exploration, Data Analysis and Discussion along with conclusions of the study.

Methodology

After the analysis of the 92 water samples, the results have been recorded in Excel format. These Excel files will be subjected to analysis, and conclusions will be drawn.

Firstly, exploratory data analysis will be carried out on the dataset to gain insights into it. The entire dataset will be subjected to descriptive analysis, and the necessary data will be visualized. During this process, any unnecessary data will be eliminated if deemed necessary.

Subsequently, the dataset will be subjected to various multivariate techniques to achieve the objectives outlined in the study. To accomplish the first objective, the study will employ a multivariate technique known as Principal Component Analysis. This analysis will focus on examining variances, proportions of variances, and summarizing the chemical components into smaller subgroups.

The second objective will be pursued through the implementation of "Cluster Analysis." In this phase, the number of clusters will be determined using statistical methods, and hierarchical cluster analysis will be employed to group the chemical components into homogeneous clusters.

The third and final objective will be attained through hypothesis testing on standardized values derived from the study. After calculating the test statistics and comparing them with the decision rule, the study will arrive at its conclusion.

All results obtained from the statistical tests will be discussed in the "Discussion and Conclusion" section.

Exploratory Data Analysis

This area of the study will analyze the data set according to four methodical ways.

1. Get the summary statistics of the entire data set.
2. Checking the null values of the data set.
3. Calculate and Visualize Correlation matrix for the data set.
4. Visualize the pairs of induvial data columns using scatter plots.

Summary Statistic for the Data Set

In this section, the entire dataset was subjected to analysis, and basic statistical calculations were performed on it. To carry out these calculations, the R built-in library called "skim" was utilized. The following image displays the summary statistics of the dataset.

```
>   #1.  perform skim to display summary statistics
>   skim(df)
── Data Summary ────────────────────────
                           Values
Name                       df
Number of rows             92
Number of columns          12
_____________________
Column type frequency:
  numeric                  12
_____________________
Group variables            None

── Variable type: numeric ──────────────────────────────────────────────────────────────────────────────────────────────────────────────
   skim_variable        n_missing complete_rate      mean       sd   p0     p25      p50    p75     p100 hist
 1 well.water.sample_No         0             1  46.5      26.7    1    23.8    46.5      69.2    92      ▇▇▇▇▇
 2 Be                           0             1   0.0255    0.0798 0     0       0         0       0.47 ▇▁▁▁▁
 3 Cr                           0             1   0.200     0.654  0     0       0         0.11    3.68 ▇▁▁▁▁
 4 Fe                           0             1  88.8     248.     0     2.59   11.2      43.0  1370.    ▇▁▁▁▁
 5 Ni                           0             1   1.14      1.76   0     0.0870  0.330     1.22    7.11 ▇▁▁▁▁
 6 Cu                           0             1 141.      371.     0.12  5.84   32.3     139.    3228.    ▇▁▁▁▁
 7 As                           0             1  11.1      74.9    0     0.103   0.521     3.81  718.     ▇▁▁▁▁
 8 Cd                           0             1   0.0288    0.0653 0     0       0.00665   0.02    0.412 ▇▁▁▁▁
 9 Ba                           0             1  13.5      29.1    0     0.643   3.81      9.14  198.     ▇▁▁▁▁
10 Tl                           0             1   0.00485   0.0320 0     0       0         0       0.29 ▇▁▁▁▁
11 Pb                           0             1   1.62      3.19   0.01  0.0998  0.359     1.15   20.1    ▇▁▁▁▁
12 U                            0             1  56.6     348.     0     0.116   1.00      8.18 3274.     ▇▁▁▁▁
>
```

Figure 1: Summary statistics of data set

Checking the Null Values of the Dataset

After executing the code for the dataset, it was observed that the sum of null values became zero.

```
>   null_val <- sum(is.null(df))
>   null_val
[1] 0
>
```

Figure 2: Checking the null values of data set.

Calculate and Visualize Correlation Matrix for the Dataset

To gain insight into the correlation among the variables, a correlation matrix was computed and subsequently visualized. It is important to note that Principal Component Analysis, as applied in this study, tends to work effectively when variables exhibit a high degree of correlation.

```
>   #3. Correlation matrix calculation
>   cor_matrix <- cor(df)
>   corrplot(cor_matrix, method = "color", tl.col = "black",tl.srt = 45, type = "full")
>
```

Figure 3: Calculating and visualizing correlation matrix

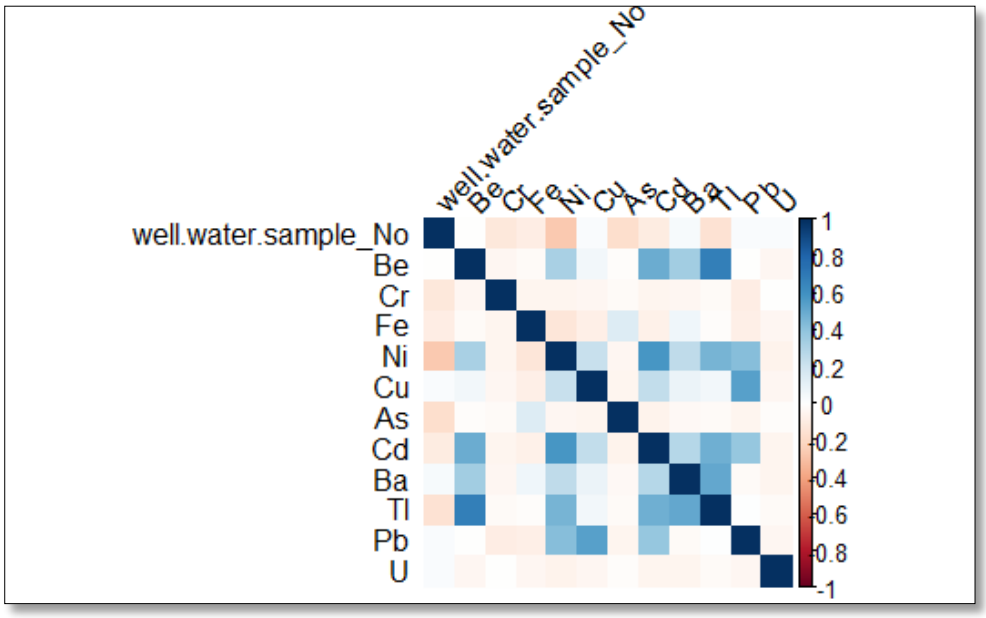


Figure 4: Correlation matrix for the dataset

Visualize the Pairs of Induvial Data Columns using Scatter Plots

Scatter plots were used to plot the pairs of induvial data behavior.

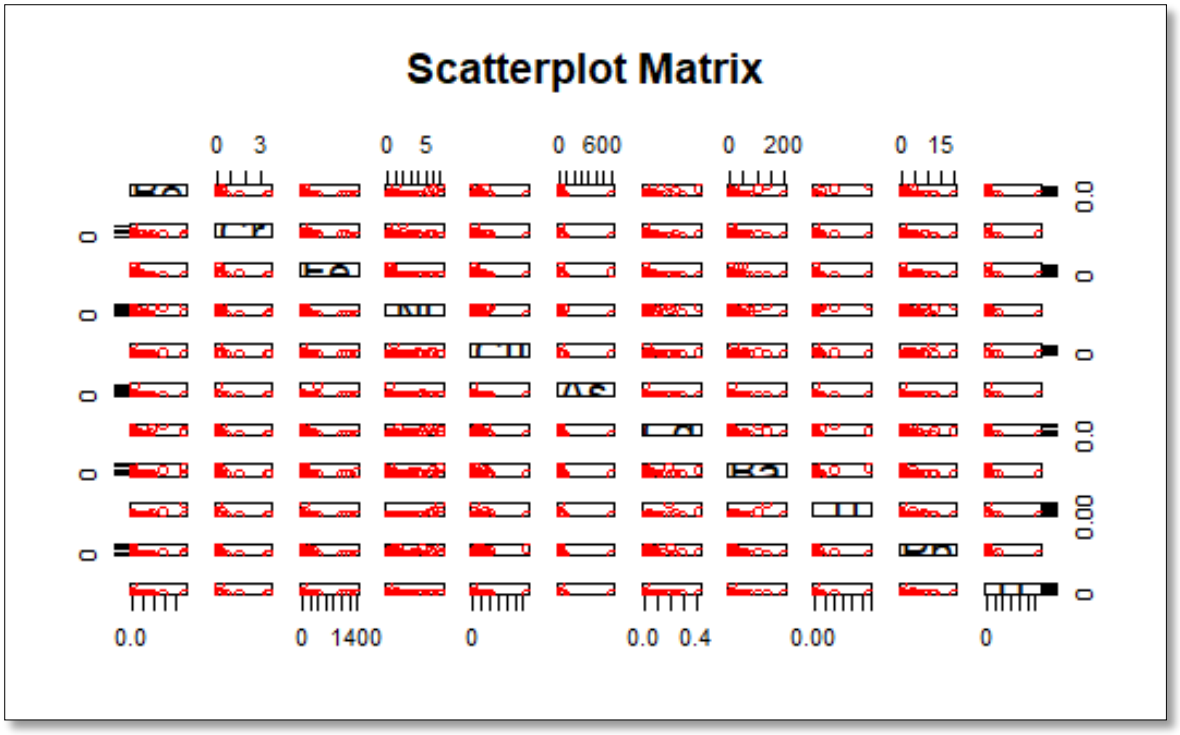


Figure 5: Scatter plot matrix for the dataset

Analysis for the Dataset

Principal Component Analysis

To accomplish the first objective, Principal Component Analysis (PCA) was performed. PCA aids in reducing the dimensionality and complexity of the dataset by capturing the variance/covariance structure through a few linear combinations of the original variables. In this study, the chemical metal components exhibited the highest variability within the dataset. PCA was employed to condense these chemical components into a smaller number of subgroups.

The following image shows the results and codes used for PCA.

```
> #Analysis
>   # Objective 1: Summarizing the chemical components into sub groups
>     # Perform Principal Component Analysis (PCA)
>     pca_result <- prcomp(df[, c("Be", "Cr", "Fe", "Ni", "Cu", "As", "Cd", "Ba", "Tl", "Pb", "U")], scale = TRUE)
>     pca_result
Standard deviations (1, .., p=11):
 [1] 1.7281203 1.2968032 1.0775815 1.0045132 0.9716089 0.9329388 0.8567597 0.7879596 0.6381542 0.5754902 0.5039119

Rotation (n x k) = (11 x 11):
           PC1         PC2          PC3          PC4         PC5         PC6         PC7          PC8
Be  0.40715651 -0.32457439  0.043925522 -0.024729664  0.02818003 -0.08397604  0.08357586 -0.5876608783
Cr -0.06322000 -0.06559069  0.400558366  0.582319131  0.59542130  0.36442814  0.02847531  0.0220228071
Fe -0.06514942 -0.19822056 -0.632414630  0.051947044 -0.04060906  0.61192611  0.40844225 -0.0165865829
Ni  0.43677200  0.15315909  0.028291761  0.014113935  0.12083120 -0.14939792  0.28051925  0.4775870067
Cu  0.21910838  0.50114675 -0.125714602  0.034863025  0.02726241  0.29389718 -0.46668846 -0.4186865707
As -0.05356205 -0.10046116 -0.563995418 -0.052455019  0.66503303 -0.39705484 -0.24977065  0.0319707226
Cd  0.46758533  0.07305314  0.005324258 -0.005273314  0.08580518 -0.05896592  0.32420485  0.0001917638
Ba  0.31776213 -0.30359436 -0.067458229  0.012742769 -0.16459469  0.29585823 -0.59047931  0.4773333773
Tl  0.44592230 -0.34436455  0.047497664 -0.029550641  0.04082053  0.01286330 -0.05447327 -0.1093594359
Pb  0.24982560  0.59153780 -0.138872461 -0.009228606  0.06715492  0.09913107  0.09250327  0.0932078477
U  -0.05888009 -0.01869519  0.277235124 -0.807638299  0.38051001  0.34400181  0.03472487  0.0314686178
            PC9          PC10          PC11
Be  0.018175472  0.265408246  0.543494087
Cr  0.028397218  0.046588067  0.023482408
Fe -0.085538372 -0.050500538  0.027339270
Ni -0.514212215 -0.243460589  0.342615407
Cu -0.188071848 -0.409840910  0.016368006
As  0.053325422  0.003728861 -0.003132092
Cd  0.683068403 -0.390273196 -0.202798166
Ba  0.254450400  0.097282044  0.191831758
Tl -0.381247532  0.132638173 -0.706920030
Pb  0.110468966  0.719813109 -0.079899047
U   0.006090362 -0.008967510  0.040074610
```

Figure 6: PCA for the dataset

To perform Principal Component Analysis, this study uses correlation matrix because it is the recommended method to perform PCA and the measurements of the dataset doesn't appear in different ranges.

Consider,

Be	Cr	Fe	Ni	Cu	As	Cd	Ba	Tl	Pb	U
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}

Because this study considers about correlation matrix, $PC_1=\underline{e_j^t X}$

Therefore, considering normalize values for X variables, $Z_1=\frac{X1-Avg(X1)}{SD(X1)}$

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9	Z_{10}	Z_{11}

Principal Components appear as follows.

1. $PC_1= 0.41Z_1 – 0.06Z_2 – 0.07Z_3 + 0.44Z_4 + 0.22Z_5 – 0.05Z_6 + 0.47Z_7 + 0.32Z_8 + 0.45Z_9 + 0.25Z_{10} – 0.06Z_{11}$

2. $PC_2 = - 0.32Z_1 – 0.07Z_2 – 0.2Z_3 + 0.15Z_4 + 0.5Z_5 – 0.1Z_6 + 0.07Z_7 – 0.3Z_8 – 0.34Z_9 + 0.59Z_{10} – 0.02Z_{11}$

3. $PC_{3=}\ 0.04Z_1 + 0.4Z_2 - 0.63Z_3 + 0.03Z_4 - 0.13Z_5 - 0.56Z_6 + 0.01Z_7 - 0.07Z_8 + 0.05Z_9 - 0.14Z_{10} + 0.28Z_{11}$

4. $PC_4 = 0.02Z_1 - 0.58Z_2 - 0.05Z_3 - 0.01Z_4 - 0.03Z_5 + 0.05Z_6 + 0.01Z_7 - 0.01Z_8 + 0.03Z_9 + 0.01Z_{10} + 0.81Z_{11}$

5. $PC_{5=}\ 0.03Z_1 + 0.6Z_2 - 0.04Z_3 + 0.12Z_4 + 0.03Z_5 + 0.67Z_6 + 0.09Z_7 - 0.16Z_8 + 0.04Z_9 + 0.07Z_{10} + 0.38Z_{11}$

6. $PC_{6=} = -\ 0.08Z_1 + 0.36Z_2 + 0.61Z_3 - 0.15Z_4 + 0.29Z_5 - 0.4Z_6 - 0.06Z_7 + 0.3Z_8 + 0.01Z_9 + 0.1Z_{10} + 0.34Z_{11}$

7. $PC_7 = 0.08Z_1 + 0.03Z_2 + 0.41Z_3 + 0.28Z_4 - 0.47Z_5 - 0.25Z_6 + 0.32Z_7 - 0.59Z_8 - 0.05Z_9 + 0.09Z_{10} + 0.03Z_{11}$

8. $PC_8 = -\ 0.59Z_1 + 0.02Z_2 - 0.02Z_3 + 0.48Z_4 - 0.42Z_5 + 0.03Z_6 + 0Z_7 + 0.48Z_8 - 0.11Z_9 + 0.09Z_{10} + 0.03Z_{11}$

9. $PC_9 = 0.02Z_1 + 0.03Z_2 - 0.09Z_3 - 0.51Z_4 - 0.19Z_5 + 0.05Z_6 + 0.68Z_7 + 0.25Z_8 - 0.38Z_9 + 0.11Z_{10} + 0.01Z_{11}$

10. $PC_{10} = -\ 0.27Z_1 - 0.05Z_2 + 0.05Z_3 + 0.24Z_4 + 0.41Z_5 - 0Z_6 + 0.39Z_7 - 0.1Z_8 - 0.13Z_9 - 0.72Z_{10} + 0.01Z_{11}$

11. $PC_{11} = 0.54Z_1 + 0.02Z_2 + 0.03Z_3 + 0.34Z_4 + 0.02Z_5 - 0Z_6 - 0.2Z_7 + 0.19Z_8 - 0.71Z_9 - 0.08Z_{10} + 0.04Z_{11}$

The following table consists of variances represented by 11 principal components.

Principal Component	Variance
PC_1	$1.73^2 = 2.9929$
PC_2	$1.30^2 = 1.69$

PC_3	$1.08^2 = 1.1664$
PC_4	$1.00^2 = 1$
PC_5	$0.97^2 = 0.9409$
PC_6	$0.93^2 = 0.8649$
PC_7	$0.86^2 = 0.7396$
PC_8	$0.79^2 = 0.6241$
PC_9	$0.64^2 = 0.4096$
PC_{10}	$0.58^2 = 0.3364$
PC_{11}	$0.50^2 = 0.25$
Sum	11.0148 ~ 11

Then calculated the proportion of total variance explained from each principal component as a percentage.

1. From $PC_1 = \frac{2.9929}{11}$ X 100% = 27.21%: Explained 27.21% of variability in the dataset.
2. From $PC_2 = \frac{1.69}{11}$ X 100% = 15.36%: Explained 15.36% of variability in the dataset.
3. From $PC_3 = \frac{1.1664}{11}$ X 100% = 10.60%: Explained 10.60% of variability in the dataset.
4. From $PC_4 = \frac{1}{11}$ X 100% = 9.09%: Explained 9.09% of variability in the dataset.
5. From $PC_5 = \frac{0.9409}{11}$X 100% = 8.55%: Explained 8.55% of variability in the dataset.
6. From $PC_6 = \frac{0.8649}{11}$X 100% = 7.86%: Explained 7.86% of variability in the dataset.
7. From $PC_7 = \frac{0.7396}{11}$X 100% = 6.72%: Explained 6.72% of variability in the dataset.
8. From $PC_8 = \frac{0.6241}{11}$X 100% = 5.67%: Explained 5.67% of variability in the dataset.
9. From $PC_9 = \frac{0.4096}{11}$X 100% = 3.72%: Explained 3.72% of variability in the dataset.
10. From $PC_{10} = \frac{0.3364}{11}$X 100% = 3.06%: Explained 3.06% of variability in the dataset.
11. From $PC_{11} = \frac{0.25}{11}$X 100% = 2.27%: Explained 2.27% of variability in the dataset.

Out of 11 principal components, first **7** principal components explain,

(27.21% + 15.36% + 10.60% + 9.09% + 8.55% + 7.86%+6.72% = 85.39%)

of total variability of this dataset.

Cluster Analysis

The second objective was achieved by conducting Cluster Analysis on the dataset. To achieve that Hierarchical Cluster Analysis was performed on the dataset. It was performed according to the following steps.

1. Identified the optimal number of clusters to perform cluster analysis.
2. Calculate the distance matrix.
3. Applied hierarchical clustering and plot the dendrogram for clustered dataset.

Identify the Optimal Number of Clusters

To Identify the optimal number of clusters for dataset, 2 methods were used.

The first one is plotting the scree plot and identifying the number of clusters until elbow point.

```
>     #determine the no of clusters
>     fviz_nbclust(df[-1],kmeans,method = "wss")
>
```

Figure 7: Scree plot

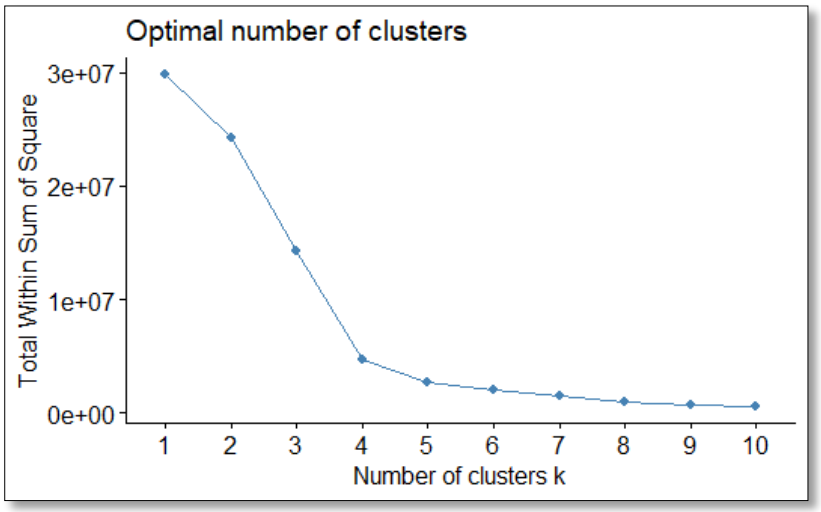


Figure 8: Scree plot for the data set

The second method was using "Nbclust" R inbuilt library.

```
#2nd method to obtain no of optimal clusters
NbClust(data = df[-1], diss = NULL, distance = "euclidean", min.nc = 2, max.nc = 15,
        method = "ward.D2", index = "all", alphaBeale = 0.1)
```

Figure 9: Code for Nbclust function

```
* 2 proposed 4 as the best number of clusters
* 2 proposed 5 as the best number of clusters
* 3 proposed 7 as the best number of clusters
* 1 proposed 8 as the best number of clusters
* 3 proposed 15 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  3
```

Figure 10: Conclusion of NBclust function

According to both methods optimal number of clusters = 3

Calculate the Distance Matrix

After obtaining the optimal number of clusters, distance matrix was calculated according to "Euclidean Distance."

```
>     d_df <- dist(df[-1])
>     d_df
           1          2          3          4          5          6          7          8          9
2  1114.753178
3  1061.779632  256.960915
4  1058.505987  257.253680   12.695468
5   995.218090  861.197051  808.590066  800.053869
6  1001.849710  187.046027  112.948255  110.121058  787.566029
7  1096.750596  233.407032   43.040839   45.360986  816.250239  121.879999
8  1092.972095  251.009587   61.103394   62.922955  822.954560  135.698420   54.026217
9  1096.830568  242.055054   37.769202   42.563205  825.527909  126.869577   14.117688   50.601010
10   95.590419 1031.938154  969.999127  966.736275  933.092548  913.081196 1005.553553 1001.418801 1005.397885
11 1094.193663  217.385482   50.350491   53.671572  825.232785  110.854799   20.014478   54.920045   25.024908
          10         11         12         13         14         15         16         17         18
2
3
4
```

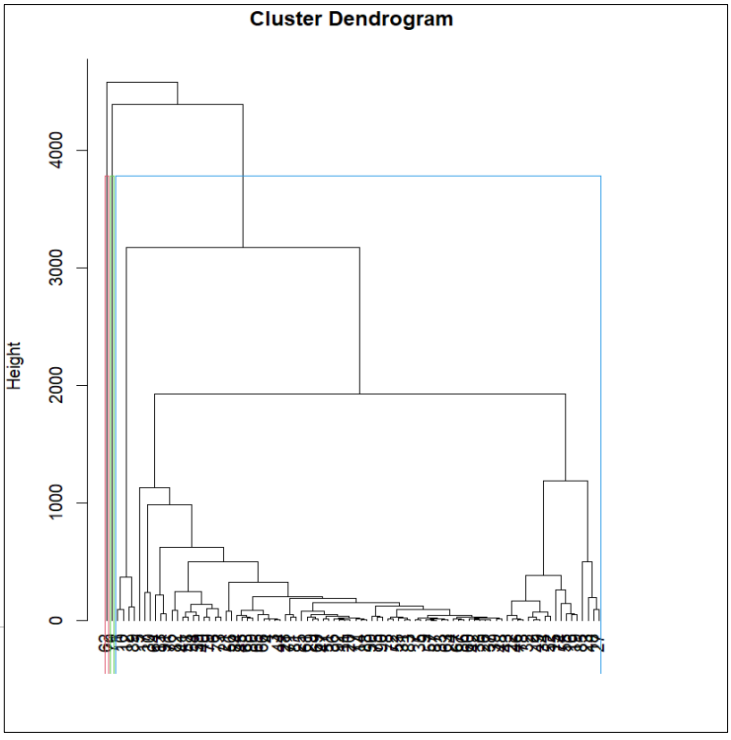
Figure 11: Distance Matrix Calculation and Code

Applied Hierarchical Clustering and Plot the Dendrogram for Clustered Dataset

After calculating the distance matrix hierarchical clustering applied and dendrogram plotted using 3 clusters. Ward.D2 method was used to

```
>     wardse <- hclust(d_df,"ward.D2")
>     plot(wardse, hang = -1)
>     rect.hclust(wardse, k=3 , border="blue")
>     rect.hclust(wardse, k=3 , border= 2:5)
```

Figure 12: hierarchical clustering and dendrogram



Finally created a data frame including well water sample number along with their cluster number.

```
>     clusters <- cutree(wardse,k=3)
>     df2 <- data.frame(df$well.water.sample_No,clusters)
>     head(df2)
  df.well.water.sample_No clusters
1                       1        1
2                       2        1
3                       3        1
4                       4        1
5                       5        1
6                       6        1
>
```

Figure 13: sample numbers along with their cluster number

Hypothesis Testing

To achieve the third objective, hypothesis testing was applied by comparing the mean values of the study dataset with standardized values for well drinking water. The standardized values for well drinking water were obtained from the National Primary Drinking Water Regulations website (*National Primary Drinking Water Regulations*, n.d.).

The mean values of the dataset and the standardized values were placed into two separate vectors. Subsequently, Hotelling's T-squared test was employed to make a comparison between the mean values of the dataset and the standardized values of the chemical metal components in water. A matrix named "ex1" was created to incorporate the mean values of the dataset. Variables known as "p" and "n" were established to define the number of variables and observations in the dataset, which amounted to 11 variables and 92 observations. Following this, the T-squared statistics significance, and the test statistics value was compared with the critical value from the F-Distribution table at a significance level of 0.05.

1. Null Hypothesis H_0: Chemical mixtures in well water samples are in line with the standard accepted values in well water.
2. Alternative Hypothesis H_1: Chemical mixtures in well water samples are **not** in line with the standard accepted values in well water.

```
>     #Test statistics
>     T2_cal <- n*t(x_bar-mu_note)%*%solve(S)%*% (x_bar-mu_note)
>     T2_cal
          [,1]
[1,] 3608806
>     Table_value =(n-1)*p/(n-p)*qf(0.95,p,n-p)
>     Table_value
[1] 23.59049
```

Figure 14: Test statistics and Table value

TheT squared value is greater than table value. Therefore, **we have enough** evidence to reject null hypotheses, which is Chemical mixtures in well water samples are in line with the standard accepted values in well water.

Discussion

To achieve the mentioned objectives three multivariate analyses were conducted.

In the Principal Component Analysis, the scores of Principal components were rounded off to two decimal places, and the standard deviation values were similarly rounded off to two decimal places. The determination of the number of Principal components was based on the total variance explained by these components.

In this study, it was found that 7 Principal components explained the total variability, accounting for 85.39% of the variance. Consequently, these 7 Principal Components were selected to effectively summarize the dataset.

In the Cluster Analysis, the optimal number of clusters was achieved through two methods. The first method involved the examination of a Scree plot, which identified the elbow point as 4. To further validate the optimal number of clusters, the "NbClust" library was utilized.

Based on the results of both methods, the optimal number of clusters was determined to be **three**. The dataset was then subjected to hierarchical clustering, employing Euclidean Distance and the "Ward.D2" method. After the dataset had been clustered, a new data frame, "df2," was created, which included the relevant well water sample numbers and their corresponding cluster assignments. Most samples (90) were clustered into one group, while the remaining two samples were allocated to the other clusters.

In the Hypothesis testing, it was assumed that the means of the two populations were both equal to zero. All the standardized values were then converted into Micrograms per liter (µg/l). Ultimately, the study reached the conclusion that there was a significant difference between the sample well water and the standardized well water metal parameters.

References

[1]J. Smith et al., “Placeholder Text: A Study,” Citation Styles, vol. 3, Jul. 2021, doi: 10.10/X.

[2] (Smith et al., 2021)Smith, J., Petrovic, P., Rose, M., De Souz, C., Muller, L., Nowak, B., & Martinez, J. (2021). Placeholder Text: A Study. The Journal of Citation Styles, 3. https://doi.org/10.10/X

Appendix

Code: R file