# GENERAL SIR JOHN KOTELAWALA DEFENSE UNIVERSITY

## FACULTY OF COMPUTING

## DEPARTMENT OF COMPUTATIONAL MATHEMATICS

**Data Science and Business Analytics degree**

# PROJECT REPORT

Module: CM 3052 Multivariate Data Analysis

By: D/DBA/21/0007 - HAYB Hettiarachchi

Lecturer in Charge: Dr.Niroshan Withanage

Intake: Intake 38

Submitted on: 19/10/2023.

Contents

# 1. Introduction

## 1.1 Background of the study

This study conducts a comprehensive investigation into well water quality in the critical context of Maine and New Hampshire, USA, where clean and safe drinking water is a paramount concern. The core of the research involves a dataset of ninety-two meticulously collected well water samples, gathered under the NIGMS Science Education Partnership Award (SEPA) project, "Data to Action: A Secondary School-Based Citizen Science Project to Address Arsenic Contamination of Well Water." The active involvement of teachers and students from rural schools in the region is central to this scientific endeavor, emphasizing collaboration and community engagement. Through strategic partnerships and extensive training, these citizen scientists lead the data collection efforts, focusing primarily on arsenic analysis. The sampling protocol is rigorous, ensuring representativeness by capturing 50 mL of well water in plastic conical tubes following a five-minute cold-water tap run. Beyond sample collection, valuable metadata is collected, including collector identities, well characteristics, filtration status, and ethical considerations that allow families to grant consent for data sharing with authoritative entities. This commitment to ethical standards ensures the autonomy of families in sharing data with institutions such as the New Hampshire Department of Environmental Services, the Maine Centre for Disease Control, and researchers dedicated to enhancing the scientific discourse on well water quality.

## 1.2 Data and Objectives

### 1.2.1 Data Description

The dataset consists of 92 well water samples, and each sample was examined for the presence and concentration of eleven different chemical elements which are Beryllium (Be), Chromium (Cr), Iron (Fe), Nickel (Ni), Copper (Cu), Arsenic (As), Cadmium (Cd), Barium (Ba), Thallium (Ti), Lead (Pb), and Uranium (U).

### 1.2.2 Study's Objectives

The goals of this study are as follows:

(i) The eleven chemical components can be summarized into small number of subgroups.

(ii) Well water samples can be cluster into homogeneous groups according to the structure of the mixture components.

(iii) Chemical mixtures in well water samples are in line with the standard accepted values in well water samples (refer USA standards).

## 1.3 Scope of the study

It is significant to note that this study considers the overall chemical composition of well water while concentrating on specific metal chemical components like arsenic, lead, uranium, and others. The study's focus is also restricted to the parameters measured and the dataset that is currently accessible.

## 1.4 Structure of the report

The report follows a structured format with five sections. It begins with an "Introduction," which gives background information, a description of the project's scope, and its objectives. The other sections are "Methodology," which describes the research approach, "Data Exploration," which offers preliminary findings, "Analysis," which discusses statistical techniques and findings, and "Discussion and Conclusion," which offers interpretation and conclusions. This structured framework guarantees coherence and clarity while thoroughly assessing the study's objectives.

# 2. Methodology

To conduct a comprehensive multivariate data analysis for this project, several techniques can be applied. Given the nature of the study involving multiple metal chemical constituents and the need to achieve specific objectives, the following methods can be considered. We can use Principal Component Analysis (PCA) for reducing data dimensionality and summarizing relationships among chemical components (objective I), Cluster Analysis ( Hierarchical Clustering)for grouping well water samples based on chemical composition similarity to identify homogeneous groups (objective ii), and Comparative Analysis using hypothesis testing for assessing differences in chemical mixtures among well water samples and testing conformity with accepted standards (objective iii). These methods provide a comprehensive approach to address the research objectives.

## 2.1 Principle Component Analysis (PCA)

By reducing the original chemical components into a more manageable group of uncorrelated variables known as principal components, PCA helps to accomplish the first objective above mentioned. These elements describe the relationships between the chemical constituents and capture the data's most significant variance. By deconstructing the data's complicated structure, PCA makes it possible to identify underlying patterns and connections between the chemical constituents. By combining chemical parts that are comparable, this reduction in dimensionality can aid in achieving the objective of summarizing the relationships between the elements.

## 2.2 Cluster Analysis

The second objective is accomplished successfully by cluster analysis techniques like Hierarchical Clustering and K-Means. They classify the samples of well water based on how similar their chemical compositions are. Cluster analysis can reveal inherent patterns and similarities in the chemical composition of well water samples by locating homogenous groupings within the dataset. In turn, this facilitates the objective of grouping well water samples into homogeneous groups based on the structure of the mixture's constituent parts, making it easier to identify any comparable characteristics or anomalies.

## 2.3 Comparative Analysis along with hypothesis testing

A comparative analysis was done using hypothesis testing to determine if the chemical combinations in the well water samples reflect the normative acknowledged values for well water in the USA. This required comparing the ninety-two well water samples reported chemical component contents (Be, Cr, Fe, Ni, Cu, As, Cd, Ba, Ti, Pb, and U) with the pertinent standards set by regulatory organizations. The study evaluated if the well water complied with these standards, which is essential for identifying potential water quality issues and guaranteeing the security of the water supply.

# 3. Data Exploration

Data exploration is a critical phase that sets the foundation for the subsequent analyses and interpretations. It provides insights into the well water samples and their chemical constituents, aligning with the study's objectives.

## 3.1 Descriptive statistics

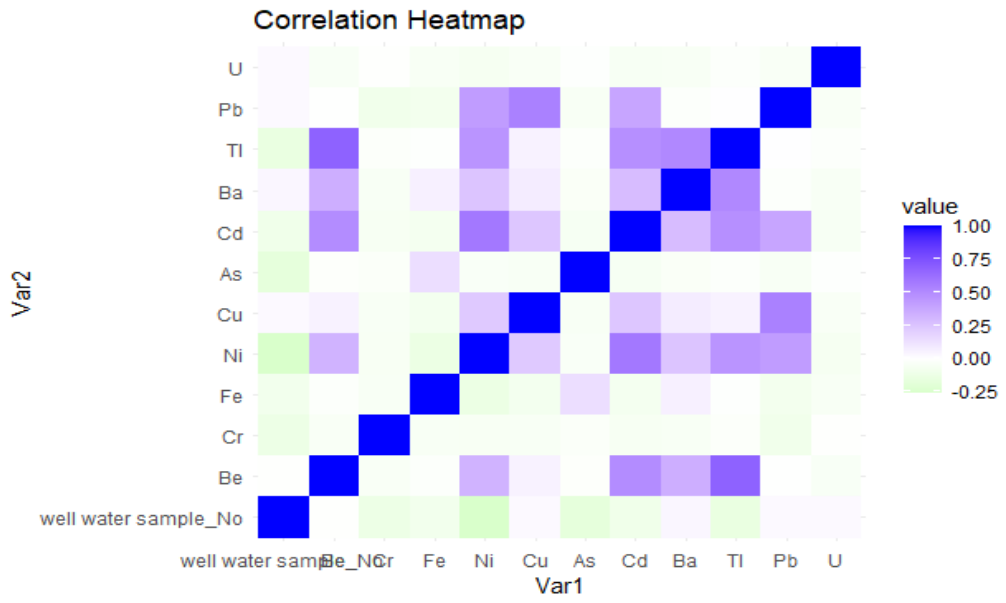
Summary Statistics provide a brief statistical overview of the dataset. For each chemical component, this contains statistics like the mean, median, minimum and maximums and quartiles. A basic knowledge of the central tendency and variability in the data is provided by descriptive statistics.

```
> # Get summary statistics for the dataset
> summary(data)
 well water sample_No       Be                Cr                Fe
 Min.   : 1.00        Min.   :0.00000   Min.   :0.0000   Min.   :   0.000
 1st Qu.:23.75        1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:   2.587
 Median :46.50        Median :0.00000   Median :0.0000   Median :  11.171
 Mean   :46.50        Mean   :0.02552   Mean   :0.1995   Mean   :  88.809
 3rd Qu.:69.25        3rd Qu.:0.00000   3rd Qu.:0.1100   3rd Qu.:  43.038
 Max.   :92.00        Max.   :0.47000   Max.   :3.6792   Max.   :1370.356
       Ni                 Cu                As                Cd
 Min.   :0.00000   Min.   :   0.120   Min.   :  0.0000   Min.   :0.000000
 1st Qu.:0.08704   1st Qu.:   5.842   1st Qu.:  0.1033   1st Qu.:0.000000
 Median :0.33013   Median :  32.280   Median :  0.5207   Median :0.006647
 Mean   :1.13954   Mean   : 141.456   Mean   : 11.0786   Mean   :0.028784
 3rd Qu.:1.21730   3rd Qu.: 139.486   3rd Qu.:  3.8104   3rd Qu.:0.020000
 Max.   :7.11000   Max.   :3228.015   Max.   :717.9056   Max.   :0.411806
       Ba                Tl                Pb                U
 Min.   :  0.0000   Min.   :0.00000   Min.   : 0.01000   Min.   :   0.000
 1st Qu.:  0.6428   1st Qu.:0.00000   1st Qu.: 0.09975   1st Qu.:   0.116
 Median :  3.8087   Median :0.00000   Median : 0.35904   Median :   1.002
 Mean   : 13.4774   Mean   :0.00485   Mean   : 1.62066   Mean   :  56.602
 3rd Qu.:  9.1386   3rd Qu.:0.00000   3rd Qu.: 1.15189   3rd Qu.:   8.181
 Max.   :197.6416   Max.   :0.29000   Max.   :20.09000   Max.   :3274.370
```

## 3.2 Correlation Analysis

Here I have Analyzed the relationships between various chemical components using correlation analysis. Potential connections between the elements can be found using scatterplot or correlation matrices. This analysis may be pertinent to Objective (I), where it is crucial to summarize relationships between the chemical constituents.

The correlations can be visualized using a heatmap as below.



## 3.3 Checking for the null values

It was identified that the sum of the null values in the dataset is zero. It means that the dataset has no missing values to be removed.

```
> null_values <- sum(is.null(data))
> null_values
[1] 0
>
```

# 4.Data Analysis

## 4.1 Principle Component Analysis (PCA)

### 4.1.1 Results of Principle Component Analysis

```
> # Performing of PCA
> #obtaining results as per the correlation matrix
> pca_result <- prcomp(data,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
 [9] 0.6381542 0.5754902 0.5039119

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

### 4.1.2 Procedure for obtaining Principle Components

This study renders use of the correlation matrix to conduct Principal Component Analysis as recommended.

Assuming,

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

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

1. $PC_1 = y_1 = 0.41z_1 – 0.06z_2 – 0.07z_3 + 0.44z_4 + 0.22z_5 – 0.05z_6 + 0.47z_7 + 0.32z_8 + 0.45z_9 + 0.25z_{10} – 0.06z_{11}$
2. $PC_2 = y_2 = - 0.32z_1 – 0.07z_2 – 0.2z_3 + 0.15z_4 + 0.5z_5 – 0.1z_6 + 0.07z_7 – 0.3z_8 – 0.34z_9 + 0.59z_{10} – 0.02z_{11}$
3. $PC_3 = y_3 = 0.04z_1 + 0.4z_2 – 0.63z_3 + 0.03z_4 – 0.13z_5 – 0.56z_6 + 0.01z_7 – 0.07z_8 + 0.05z_9 – 0.14z_{10} + 0.28z_{11}$
4. $PC_4 = y_4 = 0.02z_1 – 0.58z_2 – 0.05z_3 – 0.01z_4 – 0.03z_5 + 0.05z_6 + 0.01z_7 – 0.01z_8 + 0.03z_9 + 0.01z_{10} + 0.81z_{11}$
5. $PC_5 = y_5 = 0.03z_1 + 0.6z_2 – 0.04z_3 + 0.12z_4 + 0.03z_5 + 0.67z_6 + 0.09z_7 – 0.16z_8 + 0.04z_9 + 0.07z_{10} + 0.38z_{11}$
6. $PC_6 = y_6 = - 0.08z_1 + 0.36z_2 + 0.61z_3 – 0.15z_4 + 0.29z_5 – 0.4z_6 – 0.06z_7 + 0.3z_8 + 0.01z_9 + 0.1z_{10} + 0.34z_{11}$
7. $PC_7 = y_7 = 0.08z_1 + 0.03z_2 + 0.41z_3 + 0.28z_4 – 0.47z_5 – 0.25z_6 + 0.32z_7 – 0.59z_8 – 0.05z_9 + 0.09z_{10} + 0.03z_{11}$
8. $PC_8 = y_8 = - 0.59z_1 + 0.02z_2 – 0.02z_3 + 0.48z_4 – 0.42z_5 + 0.03z_6 + 0z_7 + 0.48z_8 – 0.11z_9 + 0.09z_{10} + 0.03z_{11}$
9. $PC_9 = y_9 = 0.02z_1 + 0.03z_2 – 0.09z_3 – 0.51z_4 – 0.19z_5 + 0.05z_6 + 0.68z_7 + 0.25z_8 – 0.38z_9 + 0.11z_{10} + 0.01z_{11}$
10. $PC_{10} = y_{10} = - 0.27z_1 – 0.05z_2 + 0.05z_3 + 0.24z_4 + 0.41z_5 - 0z_6 + 0.39z_7 – 0.1z_8 – 0.13z_9 – 0.72z_{10} + 0.01z_{11}$
11. $PC_{11} = y_{11} = 0.54z_1 + 0.02z_2 + 0.03z_3 + 0.34z_4 + 0.02z_5 - 0z_6 – 0.2z_7 + 0.19z_8 – 0.71z_9 – 0.08z_{10} + 0.04z_{11}$

### 4.1.3 Variances of Principle Components

1. $Var(y_1) = 1.728^2 = 2.986$
2. $Var(y_2) = 1.297^2 = 1.682$
3. $Var(y_3) = 1.076^2 = 1.157$
4. $Var(y_4) = 1.005^2 = 1.010$
5. $Var(y_5) = 0.972^2 = 0.945$

6. $Var(y_6) = 0.933^2 = 0.870$
7. $Var(y_7) = 0.857^2 = 0.734$
8. $Var(y_8) = 0.788^2 = 0.620$
9. $Var(y_9) = 0.638^2 = 0.407$
10. $Var(y_{10}) = 0.575^2 = 0.331$
11. $Var(y_{11}) = 0.504^2 = 0.254$

Total of variances = 2.986 + 1.682 + 1.157 + 1.010 + 0.945 + 0.870 + 0.734 + 0.620 + 0.407 + 0.331 + 0.254

= 10.996 (nearly equals to 11)

### 4.1.4 Proportion of Total Variance Explained by Principle Components

1. 1st variance = $\frac{2.986}{11}$ x 100% = 27.16%
   1st PC explained 27.16% of the total variability in this dataset.
2. 2nd variance = $\frac{1.682}{11}$ x 100% = 15.29%
   2nd PC explained 15.29% of the total variability in this dataset.

3. 3rd variance = $\frac{1.157}{11}$ x 100% = 10.52%
   3rd PC explained 10.52% of the total variability in this dataset.

4. 4th variance = $\frac{1.01}{11}$ x 100% = 9.18%
   4th PC explained 9.18% of the total variability in this dataset.

5. 5th variance = $\frac{0.945}{11}$ x 100% = 8.59%
   5th PC explained 8.59% of the total variability in this dataset.

6. 6th variance = $\frac{0.87}{11}$ x 100% = 7.91%
   6th PC explained 7.91% of the total variability in this dataset.

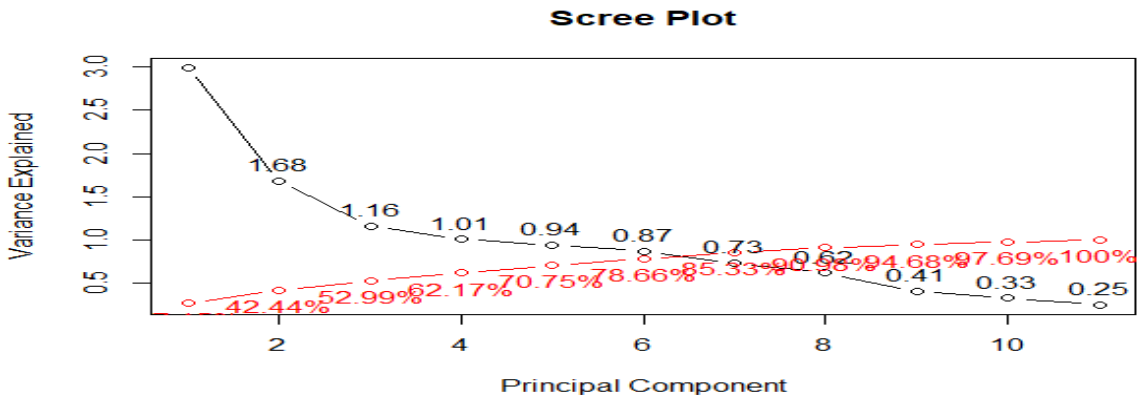
7. $7^{th}$ variance = $\frac{0.734}{11}$ x 100% = 6.67%

   $7^{th}$ variance explained 6.67% of the total variability in this dataset.

27.16% + 15.29% + 10.52% + 9.18% + 8.59% + 7.91% + 6.67% = 85.32%

The first 7 PCs together explain 85.32% of the total variability in this dataset. Therefore, the first 7 PCs are retained to define the significance of the dataset.

Using the scree plot too, the elbow occurs when PC = 7 as shown below.




## 4.2 Cluster Analysis of chemical components

Cluster analysis is a statistical technique used to group similar data points or observations into clusters or groups based on their similarity or dissimilarity. In this study, the primary objective of cluster analysis is to categorize the well water samples based on the chemical components present in the water.

### 4.2.1 Distance Matrix Calculation

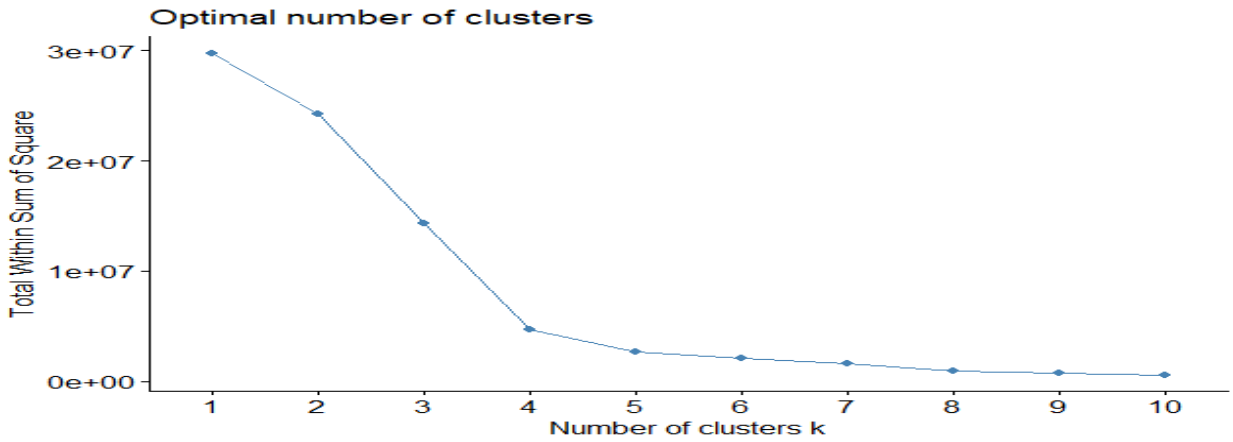
It is necessary to determine a measure of dissimilarity or distance between each pair of well water samples to group comparable samples together. The total of squared differences in the concentrations of all chemical components would be used to calculate the Euclidean distance between two well water samples.

Depending on the data and the study's objectives, decided to use Euclidean distance as the distance metric.

```
> data_matrix <- as.matrix(data)
> # Calculate the Euclidean distance between samples
> dist_matrix <- dist(data_matrix)
> dist_matrix
             1           2           3           4           5           6           7           8
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
11 1094.193663  217.385482   50.350491   53.671572  825.232785  110.854799   20.014478   54.920045
             9          10          11          12          13          14          15          16
2
3
4
```

### 4.2.2 Identifying the optimal number of clusters.

After plotting the scree plot, the points till the elbow point, which is 3, could be taken as the number of clusters.




Using "NbClust" R inbuilt function too, the optimal number of clusters received was 3.

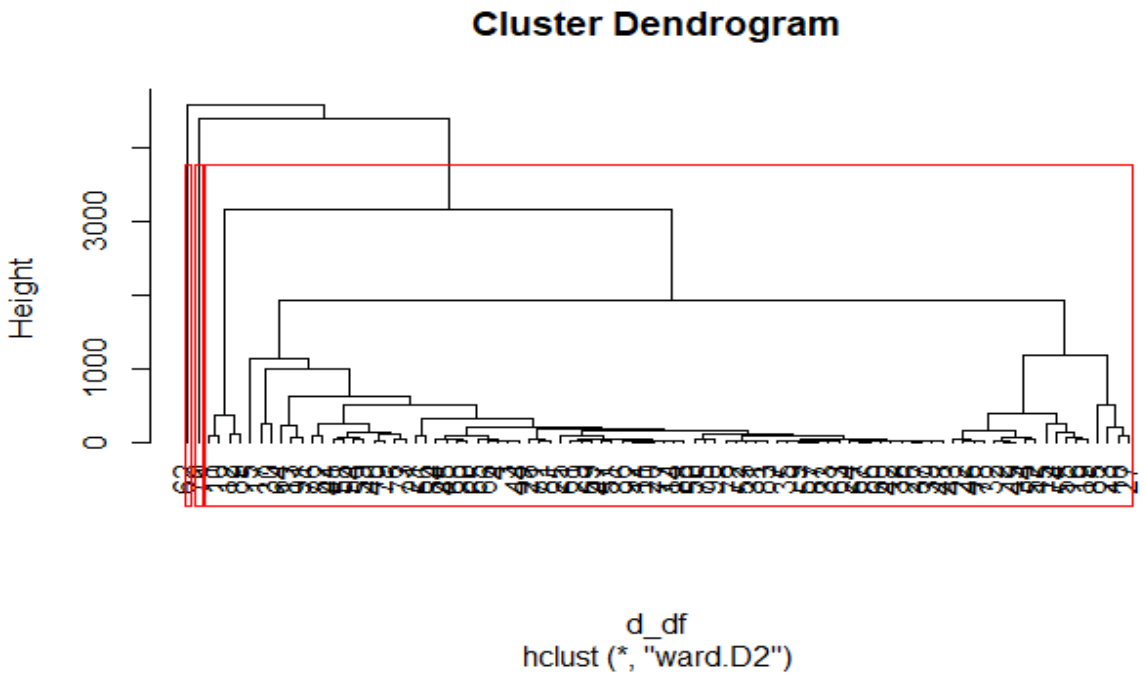
```
*******************************************************************
* Among all indices:
* 5 proposed 2 as the best number of clusters
* 8 proposed 3 as the best number of clusters
* 2 proposed 4 as the best number of clusters
* 2 proposed 5 as the best number of clusters
* 3 proposed 7 as the best 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
```

So, it is concluded that the optimal number of clusters for a ward minimum variance method usin g Euclidian distance is `3.`

### 4.2.3 Dendogram Interpretation after performing Hierarchical Clustering

After calculating distance matrix, Hierarchical clustering was done and then the dendogram was plotted using 3 clusters.

The dendogram will show how the samples are grouped at various levels of similarity.




Finally, the three clusters were classified as follows.

```
cluster 1 ;
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55
56 57 58 59 60 61 63 64 65 66 67 68 69 71 72 73 74 75 76 77 78 79 80 81 82 83
84 85 86 87 88 89 90 91 92

cluster 2; 62

cluster 3; 70
```

## 4.3 Comparative Analysis using hypothesis testing

Here, hypothesis testing was applied comparing means of well water samples and the standard values. The standardized values for well drinking water were obtained from the National Primary Drinking Water Regulations website (*National Primary Drinking Water Regulations*, n.d.).

Two distinct vectors were created using the dataset's mean values and its standardized values. The mean values of the dataset and the standardized values of the chemical metal components in water were then compared using Hotelling's T-squared test. To include the dataset's mean values, a matrix called "sample_means" was created. The dataset consisted of 11 variables(p) and 92 observations, and after that, at a significance level of 0.05, the T-squared statistics significance and test statistics value were compared with the crucial value from the F-Distribution table.

**Null Hypothesis H0: The chemical combinations in well water samples are consistent with the standardized levels.**

**Alternative Hypothesis H1: Chemical compositions in well water samples do not match accepted levels.**

```
>     #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
```

The calculated T2 value (3608806) was greater than the F-table value (23.59049). So. we can say there are enough evidence to say that the chemical compositions in well water samples do not match accepted levels, meaning that the null hypothesis was rejected.

# 5. Discussion and Conclusion

## 5.1 Objective 1: Summarizing Chemical Components

The findings of the Principal Component Analysis (PCA) offer helpful insights into the fundamental structure of your well water sample data. Through principal component analysis (PCA), it was discovered that the top seven principle components (PCs) collectively account for 85.32% of the total variance, indicating a considerable reduction in dimensionality while keeping most of the data's variability. This result is consistent with Objective 1, which sought to condense the eleven chemical constituents into more manageable categories.

These seven important PCs can be interpreted as follows:

1. PC1 appears to be strongly influenced by Beryllium (Be), Nickel (Ni), and Cadmium (Cd).
2. PC2 is primarily associated with Chromium (Cr), Copper (Cu), and Thallium (Tl).
3. The elements iron (Fe) and lead (Pb) define PC3.
4. Uranium (U) concentrations are mostly represented by PC4.
5. Arsenic (As), Cadmium (Cd), and Barium (Ba) are represented by PC5.
6. PC6 displays negative loadings for most of the components and may indicate a compositional contrast.
7. PC7 displays a more intricate pattern with several contributing factors.

Each of these PCs' relevance in capturing the data's structure is shown by the percentage of the total variance explained by each of them, which helps in achieving Objective 1.

In conclusion, the Principal Component Analysis effectively minimized the dataset's dimensionality by condensing the chemical constituents into smaller, more easily understood subgroups. For further analysis, the first seven PCs are kept since they jointly account for 85.32% of the overall variability. The investigation of homogeneous clusters among well water samples and the evaluation of their adherence to accepted values are made easier thanks to this dimension reduction, which also helps to meet the larger study objectives.

## 5.2 Objective 2: Clustering the Chemical Components

We applied the Euclidean distance in the cluster analysis to divide well water samples into groups that made sense based on the presence and concentrations of their chemical constituents. Calculating the Euclidean distance between samples demonstrated how different they were from one another and demonstrated how samples with comparable chemical constituents are more similar. We discovered that the ideal number of clusters for the ward minimum variance technique with Euclidean distance is three, which is in line with the study's second objective, using the NbClust function in R and the scree plot elbow method.

The dendrogram, a tree-like graphic representation of hierarchical clustering, and its interpretation helped to further illuminate the dataset's structure and the way samples are categorized according to degrees of similarity. Most of the samples are included in Cluster 1, showing a large cluster of well water components. One sample from Cluster 2 stands out as different from the others. Another separate group is represented by Cluster 3, highlighting the dataset's variability.

In conclusion, the three clusters formed by the successful classification of well water samples by cluster analysis serve as a helpful framework for comprehending the variety in chemical

compositions. This analysis is an important stage in achieving the study's goals and helps determine whether these samples conform to the accepted values for well water, which is crucial for the study's overall objectives.

### 5.3 Objective 3: Checking for USA standards.

In this study, we conducted a rigorous hypothesis test to compare the means of well water samples with established standard values, which were sourced from the National Primary Drinking Water Regulations. Employing Hotelling's T-squared test, we juxtaposed the mean values of the dataset against the standardized values of chemical metal components in water. The resulting T-squared statistics were evaluated at a significance level of 0.05 by comparing them with critical values from the F-Distribution table. Our null hypothesis (H0) posited that the chemical combinations in well water samples conform to standardized levels, while the alternative hypothesis (H1) suggested the opposite. Significantly, the calculated T-squared value exceeded the F-table value, providing compelling evidence for rejecting the null hypothesis. Consequently, we conclude that the chemical compositions in well water samples do not align with accepted standards. This underscores the need for continued monitoring and proactive measures to address any potential water quality concerns, prioritizing the health and well-being of the community.

## 6. References


[1] *National Primary Drinking Water Regulations | US EPA*. Available at: https://www.epa.gov/ground-water-and-drinking-water/national-primary-drinking-water-regulations (Accessed: 19 October 2023).

[2] *Clustering data in R - Amazon Web Services*. Available at: https://rstudio-pubs-static.s3.amazonaws.com/599072_93cf94954aa64fc7a4b99ca524e5371c.html (Accessed: 19 October 2023).


## Appendix

Rcode: Multivariate project(D-DBA-21-0007)