GENERAL SIR JOHN KOTELAWALA DEFENSE UNIVERSITY
FACULTY OF COMPUTING
DEPARTMENT OF COMPUTATIONAL MATHEMATICS

Multivariate Data Analysis - CM 3052

PROJECT REPORT

Multivariate Analysis of Well Water Quality in Maine and New Hampshire, USA

Student Name	AABB Athukorala
Student Number	D/DBA/21/0043
Lecturer in-charge	Dr. Niroshan Withanage

Intake 38

2023

Table of Contents

Introduction

Background and Context

In the regions of Maine and New Hampshire, USA, considerations about the quality and safety of well water have been increasing as the main domestic water supply for many households and communities. To protect the public's health, it is important to ensure that well water is clean. This research project "Data to Action: A Secondary School-Based Citizen Science Project to Address Arsenic Contamination of Well Water", initiated in collaboration with the NIGMS Science Education Partnership Award (SEPA), aims to comprehensively assess arsenic analysis of well water samples collected by teachers and students from local schools.They have concerns about 11 metal components in the collected well water samples:: Beryllium (Be), Chromium (Cr), Iron (Fe), Nickel (Ni), Copper (Cu), Arsenic (As), Cadmium (Cd), Barium (Ba), Thallium (Tl), Lead (Pb), and Uranium (U).This study advances importance of the quality of well water and may inform important regulatory practises and public health policies through careful data collecting and analysis.

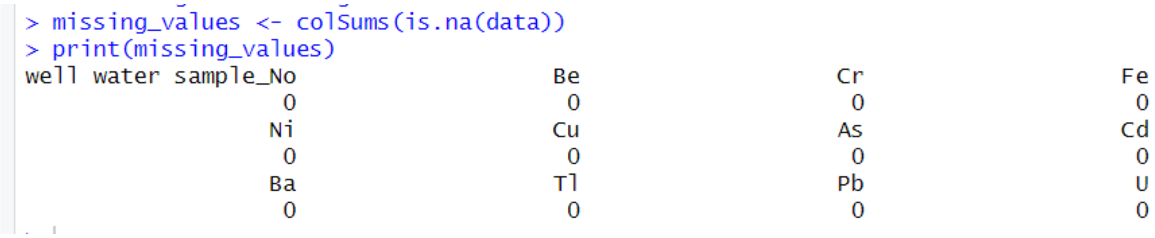
Purpose of the Research

This study's main goal is to thoroughly examine the chemical composition of the well water samples taken from the regions of Maine and New Hampshire, considering eleven important chemical elements. Three key goals of this research are as follows:(i) To identify potential subgroups within these chemical components, (ii) To cluster well water samples based on the mixture of these components, and (iii) To evaluate the compliance of chemical mixtures in well water samples with established standards within the USA.The findings of this study have a big impact on public health programs and political decisions.

Data Collection and Preprocessing

The data was sourced from 92 well water samples collected in Maine and New Hampshire, USA. Students and teachers from rural schools participated in the sample collection. Standardized protocols have been used to collect the data like, running the cold water tap for five minutes, collecting 50 mL of water, and sealing tubes with parafilm and the samples either frozen at the student's home or in the classroom for 24hrs. Data collection happened from spring 2019 to winter 2020. Key metadata include the name of the collector, the student, the well's location, its type, whether or not the water was filtered, and whether the filter was for the entire house or just at the tap. Additionally, information regarding earlier tests is gathered.

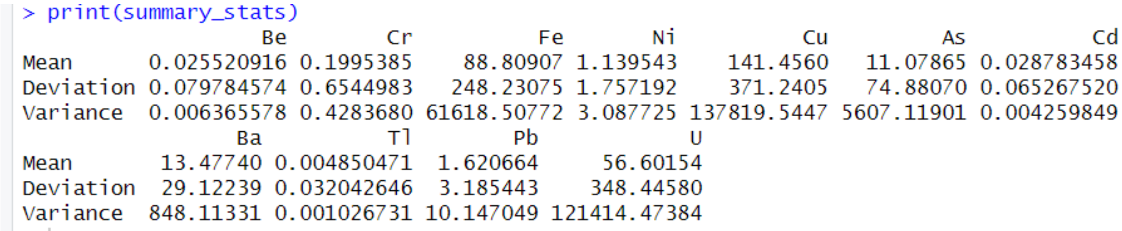
In this study, data preprocessing is an important part in getting ready the raw data for analysis, by handling the missing values and converting the dataset to a suitable format. One important thing to note is that our dataset doesn't contain any missing information. It's especially helpful in analysis like Principal Component Analysis (PCA) and cluster analysis, where having all the data reduces the chance of errors. By taking care of missing data and making sure our data is complete, we've built a strong foundation for our analysis, making our results accurate.



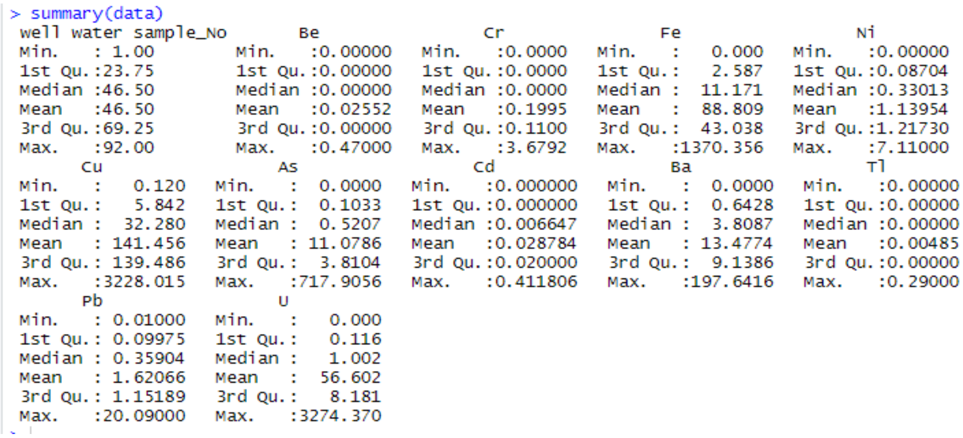
```
> missing_values <- colSums(is.na(data))
> print(missing_values)
well water sample_No                   Be                   Cr                   Fe
                   0                    0                    0                    0
                  Ni                   Cu                   As                   Cd
                   0                    0                    0                    0
                  Ba                   Tl                   Pb                    U
                   0                    0                    0                    0
```

Data Exploration

Data exploration is the phase where we take a closer look at the dataset to understand its ins and outs. This phase explores the characteristics of each chemical component, descriptive statistics like mean,deviation ets and summarize and visualize the information it holds.By using different visualizations,we can identify interesting patterns, relationships of these chemical components.

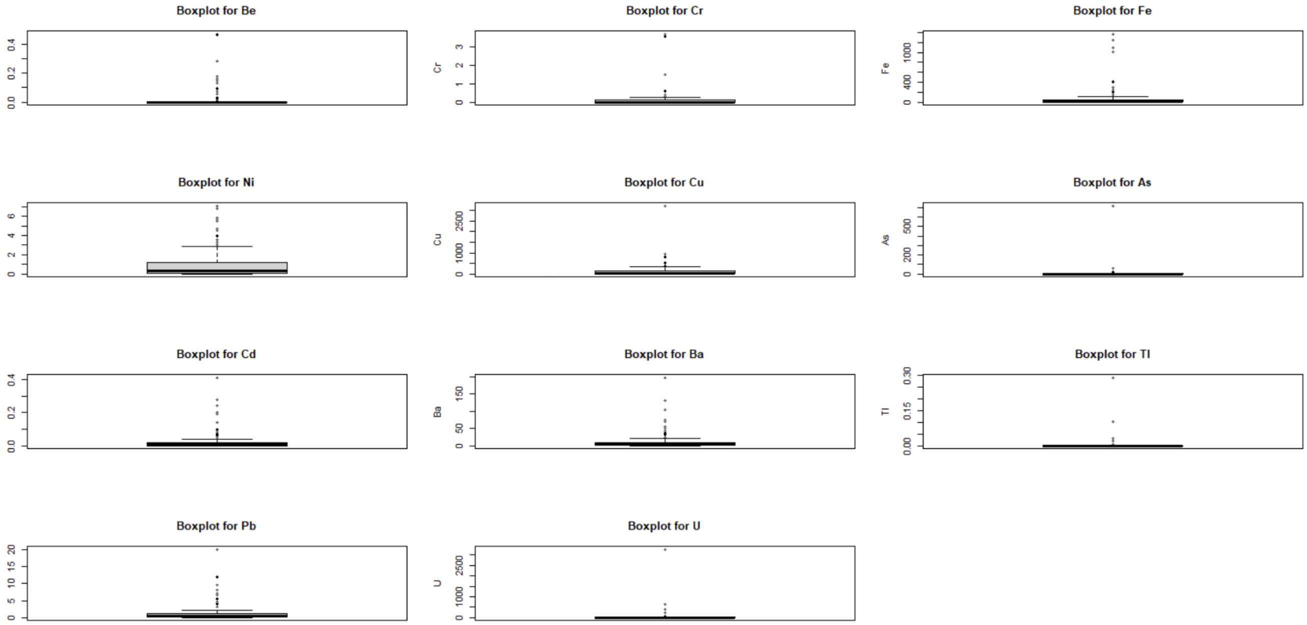


```
> print(summary_stats)
                   Be        Cr          Fe       Ni          Cu         As          Cd
Mean      0.025520916 0.1995385    88.80907 1.139543    141.4560   11.07865 0.028783458
Deviation 0.079784574 0.6544983   248.23075 1.757192    371.2405   74.88070 0.065267520
Variance  0.006365578 0.4283680 61618.50772 3.087725 137819.5447 5607.11901 0.004259849
                 Ba          Tl        Pb            U
Mean       13.47740 0.004850471  1.620664     56.60154
Deviation  29.12239 0.032042646  3.185443    348.44580
Variance  848.11331 0.001026731 10.147049 121414.47384
```

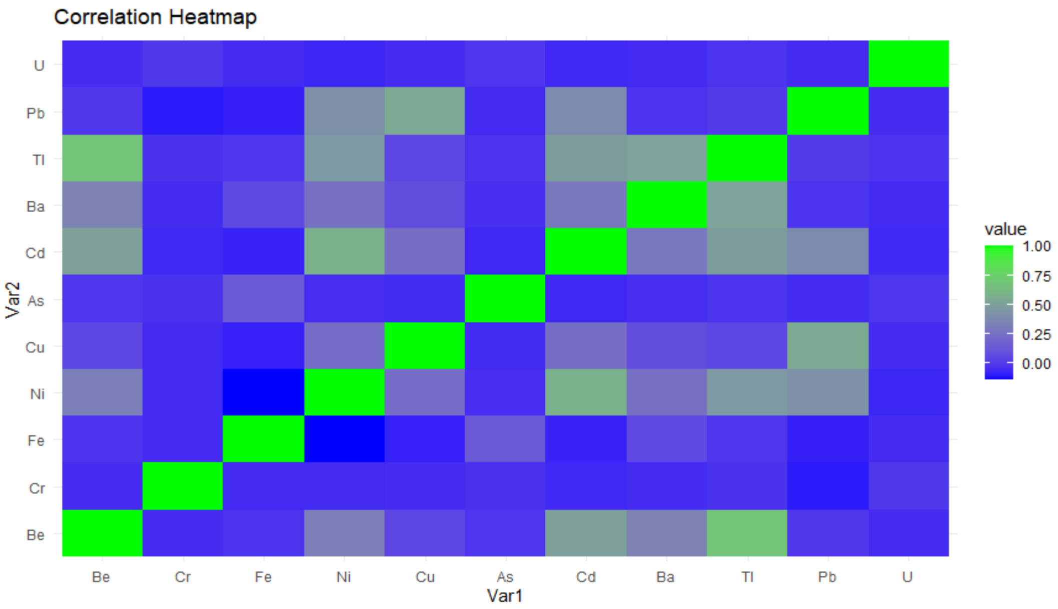


```
> summary(data)
 well water sample_No       Be                Cr               Fe                Ni
 Min.   : 1.00        Min.   :0.00000   Min.   :0.0000   Min.   :   0.000   Min.   :0.00000
 1st Qu.:23.75        1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:   2.587   1st Qu.:0.08704
 Median :46.50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

Below is a set of boxplots for each of the 11 chemical components (Be, Cr, Fe, Ni, Cu, As, Cd, Ba, Tl, Pb, U) in the dataset. Boxplots are a useful way to visualize the distribution of data, including characteristics like the median, quartiles, and any potential outliers for each component.



Correlation analysis is used to identify the relationships between variables, in this study, the 11 chemical components in the dataset. It helps to understand the relationships between the chemical components and whether there are any patterns or dependencies among them.



Methodology

In this section, we figure out the methods, models, and techniques that are used to extract valuable insights from the dataset. As the most suitable methods for the analysis, Principal Component Analysis (PCA) and cluster analysis have chosen to deliberate these objectives of this study.

PCA helps to simplify complex data by reducing its dimensionality while keeping important information. In PCA, we have used scree plot and PCA results to identify the principal components that explain the total variability. It helps to summarize the chemical components into small number of subgroups.

Cluster analysis helps to identify natural groupings within the dataset. This method is well-suited to achieve the second objective of clustering the well water samples into homogeneous groups according the structure of the mixture components. In cluster analysis, K-Means Clustering, Hierarchical Clustering and Dendrogram Analysis methods have been used to cluster well water samples to homogeneous groups.

Lastly to check whether the standard of the well water samples, we have used multivariate statistical test referred as Hotelling's T².

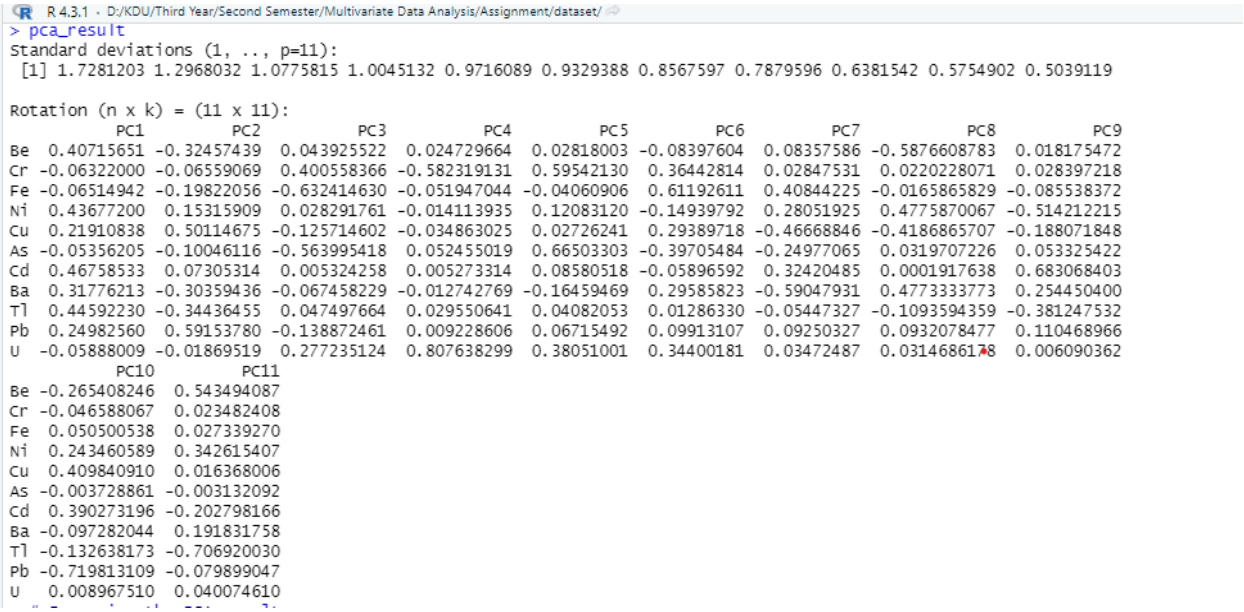
Results of these methods will be discussed in the discussion phase.

Analysis

This phase focuses on further analysis of the dataset to address the objectives of the study.

Objective 1: <u>The eleven chemical components can be summarized into small number of subgroups.</u>

As the first objective, to summarize the eleven chemical components into a smaller number of subgroups, Principal Component Analysis techniques have been used to identify the patterns and reduce the dimensionality of data. By using PCA results we can identify the chemical components that explains the high variability in the dataset.


R 4.3.1 · D:/KDU/Third Year/Second Semester/Multivariate Data Analysis/Assignment/dataset/

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

Obtaining the principal components

In PCA, the original variables, represented by matrix X (data matrix), are transformed into a new set of variables, represented by matrix Z (the principal components.)

$Z=X*V$

V = Matrix of loadings, representing the coefficients for the linear combinations of the original variables.

Z1	= Be
Z2	= Cr
Z3	= Fe
Z4	= Ni
Z5	= Cu
Z6	= As
Z7	= Cd
Z8	= Ba
Z9	= Tl
Z10	= Pb
Z11	= U

$PC = y = \underline{e_j}^{T\,(}\underline{X)}$

$PC_1 = y_1 = 0.41z_1 – 0.06z_2 – 0.07z_3 + 0.44z_4 + 0.22z_5 – 0.05z_6 + 0.47z_7 + 0.32z_8 + 0.45z_9 + 0.25z_{10} – 0.06z_{11}$

$PC_2 = y_2 = - 0.32z_1 – 0.07z_2 – 0.2z_3 + 0.15z_4 + 0.5z_5 – 0.1z_6 + 0.07z_7 – 0.3z_8 – 0.34z_9 + 0.59z_{10} – 0.02z_{11}$

$PC_3 = y_3 = 0.04z_1 + 0.4z_2 – 0.63z_3 + 0.03z_4 – 0.13z_5 – 0.56z_6 + 0.01z_7 – 0.07z_8 + 0.05z_9 – 0.14z_{10} + 0.28z_{11}$

$PC_4 = y_4 = 0.02z_1 – 0.58z_2 – 0.05z_3 – 0.01z_4 – 0.03z_5 + 0.05z_6 + 0.01z_7 – 0.01z_8 + 0.03z_9 + 0.01z_{10} + 0.81z_{11}$

$PC_5 = y_5 = 0.03z_1 + 0.6z_2 – 0.04z_3 + 0.12z_4 + 0.03z_5 + 0.67z_6 + 0.09z_7 – 0.16z_8 + 0.04z_9 + 0.07z_{10} + 0.38z_{11}$

$PC_6 = y_6 = - 0.08z_1 + 0.36z_2 + 0.61z_3 – 0.15z_4 + 0.29z_5 – 0.4z_6 – 0.06z_7 + 0.3z_8 + 0.01z_9 + 0.1z_{10} + 0.34z_{11}$

$PC_7 = y_7 = 0.08z_1 + 0.03z_2 + 0.41z_3 + 0.28z_4 – 0.47z_5 – 0.25z_6 + 0.32z_7 – 0.59z_8 – 0.05z_9 + 0.09z_{10} + 0.03z_{11}$

$PC_8 = y_8 = - 0.59z_1 + 0.02z_2 – 0.02z_3 + 0.48z_4 – 0.42z_5 + 0.03z_6 + 0z_7 + 0.48z_8 – 0.11z_9 + 0.09z_{10} + 0.03z_{11}$

$PC_9 = y_9 = 0.02z_1 + 0.03z_2 – 0.09z_3 – 0.51z_4 – 0.19z_5 + 0.05z_6 + 0.68z_7 + 0.25z_8 – 0.38z_9 + 0.11z_{10} + 0.01z_{11}$

$PC_{10} = y_{10} = - 0.27z_1 – 0.05z_2 + 0.05z_3 + 0.24z_4 + 0.41z_5 - 0z_6 + 0.39z_7 – 0.1z_8 – 0.13z_9 – 0.72z_{10} + 0.01z_{11}$

$PC_{11} = y_{11} = 0.54z_1 + 0.02z_2 + 0.03z_3 + 0.34z_4 + 0.02z_5 - 0z_6 – 0.2z_7 + 0.19z_8 – 0.71z_9 – 0.08z_{10} + 0.04z_{11}$

Find the variances of PC’s.

$Var(y_1) = 1.728^2 = 2.99$
$Var(y_2) = 1.297^2 = 1.68$
$Var(y_3) = 1.078^2 = 1.16$

Var(y_4) = 1.005^2 = 1.01

Var(y_5) = 0.972^2 = 0.94

Var(y_6) = 0.933^2 = 0.87

Var(y_7) = 0.857^2 = 0.73

Var(y_8) = 0.788^2 = 0.62

Var(y_9) = 0.638^2 = 0.41

Var(y_{10}) = 0.575^2 = 0.33

Var(y_{11}) = 0.504^2 = 0.25

Total variances of PC's = 10.99 ~ 11

Finding the proportions explained by each PC.

PC1 = $\frac{2.99}{11} \times 100 = 27.18\%$ First PC explains 27.18% of the total variability in this dataset

PC2 = $\frac{1.68}{11} \times 100 = 15.27\%$ Second PC explains 15.27% of the total variability in this dataset

PC3 = $\frac{1.16}{11} \times 100 = 10.55\%$ Third PC explains 10.55% of the total variability in this dataset

PC4= $\frac{1.01}{11} \times 100 = 9.18\%$ Fourth PC explains 9.18% of the total variability in this dataset

PC5= $\frac{0.94}{11} \times 100 = 8.55\%$ Fifth PC explains 8.55% of the total variability in this dataset

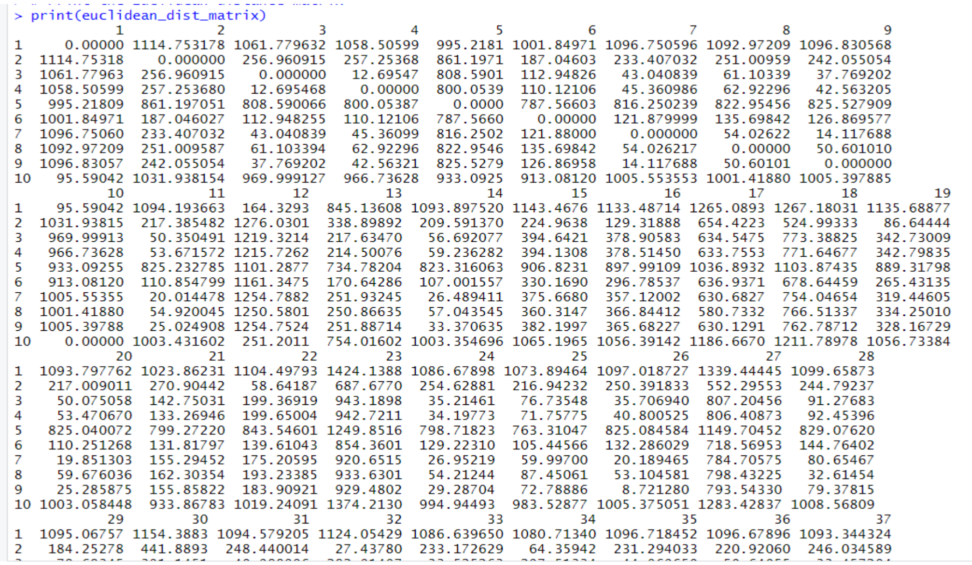
PC6= $\frac{0.87}{11} \times 100 = 7.91\%$ Sixth PC explains 7.91% of the total variability in this dataset

PC7= $\frac{0.73}{11} \times 100 = 6.64\%$ Seventh PC explains 6.64% of the total variability in this dataset

According to above findings, 7 PC's together explains around 85.28% of the total variability in this dataset. So, 7 PCs are sufficient.

Objective 2: <u>Well water samples can be cluster into homogeneous groups according to the structure of the mixture components.</u>

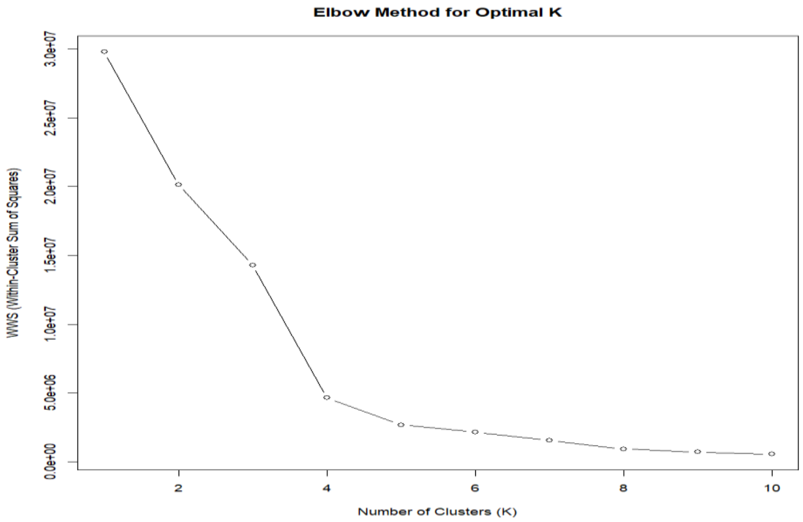
To fulfil this objective, firstly we need to measure Euclidean distance of the well water samples to measure the similarity and grouping similar once into clusters.



```
> print(euclidean_dist_matrix)
            1           2           3          4        5          6           7          8           9
1     0.00000 1114.753178 1061.779632 1058.50599 995.2181 1001.84971 1096.750596 1092.97209 1096.830568
2  1114.75318    0.000000  256.960915  257.25368 861.1971  187.04603  233.407032  251.00959  242.055054
3  1061.77963  256.960915    0.000000   12.69547 808.5901  112.94826   43.040839   61.10339   37.769202
4  1058.50599  257.253680   12.695468    0.00000 800.0539  110.12106   45.360986   62.92296   42.563205
5   995.21809  861.197051  808.590066  800.05387   0.0000  787.56603  816.250239  822.95456  825.527909
6  1001.84971  187.046027  112.948255  110.12106 787.5660    0.00000  121.879999  135.69842  126.869577
7  1096.75060  233.407032   43.040839   45.36099 816.2502  121.88000    0.000000   54.02622   14.117688
8  1092.97209  251.009587   61.103394   62.92296 822.9546  135.69842   54.026217    0.00000   50.601010
9  1096.83057  242.055054   37.769202   42.56321 825.5279  126.86958   14.117688   50.60101    0.000000
10   95.59042 1031.938154  969.999127  966.73628 933.0925  913.08120 1005.553553 1001.41880 1005.397885
           10          11        12        13          14        15         16        17         18         19
1    95.59042 1094.193663  164.3293 845.13608 1093.897520 1143.4676 1133.48714 1265.0893 1267.18031 1135.68877
2  1031.93815  217.385482 1276.0301 338.89892  209.591370  224.9638  129.31888  654.4223  524.99333   86.64444
3   969.99913   50.350491 1219.3214 217.63470   56.692077  394.6421  378.90583  634.5475  773.38825  342.73009
4   966.73628   53.671572 1215.7262 214.50076   59.236282  394.1308  378.51450  633.7553  771.64677  342.79835
5   933.09255  825.232785 1101.2877 734.78204  823.316063  906.8231  897.99109 1036.8932 1103.87435  889.31798
6   913.08120  110.854799 1161.3475 170.64286  107.001557  330.1690  296.78537  636.9371  678.64459  265.43135
7  1005.55355   20.014478 1254.7882 251.93245   26.489411  375.6680  357.12002  630.6827  754.04654  319.44605
8  1001.41880   54.920045 1250.5801 250.86635   57.043545  360.3147  366.84412  580.7332  766.51337  334.25010
9  1005.39788   25.024908 1254.7524 251.88714   33.370635  382.1997  365.68227  630.1291  762.78712  328.16729
10    0.00000 1003.431602  251.2011 754.01602 1003.354696 1065.1965 1056.39142 1186.6670 1211.78978 1056.73384
            20         21         22        23         24         25          26         27         28
1  1093.797762 1023.86231 1104.49793 1424.1388 1086.67898 1073.89464 1097.018727 1339.44445 1099.65873
2   217.009011  270.90442   58.64187  687.6770  254.62881  216.94232  250.391833  552.29553  244.79237
3    50.075058  142.75031  199.36919  943.1898   35.21461   76.73548   35.706940  807.20456   91.27683
4    53.470670  133.26946  199.65004  942.7211   34.19773   71.75775   40.800525  806.40873   92.45396
5   825.040072  799.27220  843.54601 1249.8516  798.71823  763.31047  825.084584 1149.70452  829.07620
6   110.251268  131.81797  139.61043  854.3601  129.22310  105.44566  132.286029  718.56953  144.76402
7    19.851303  155.29452  175.20595  920.6515   26.95219   59.99700   20.189465  784.70575   80.65467
8    59.676036  162.30354  193.23385  933.6301   54.21244   87.45061   53.104581  798.43225   32.61454
9    25.285875  155.85822  183.90921  929.4802   29.28704   72.78886    8.721280  793.54330   79.37815
10 1003.058448  933.86783 1019.24091 1374.2130  994.94493  983.52877 1005.375051 1283.42837 1008.56809
           29        30          31         32          33         34          35         36          37
1  1095.06757 1154.3883 1094.579205 1124.05429 1086.639650 1080.71340 1096.718452 1096.67896 1093.344324
2   184.25278  441.8893  248.440014   27.43780  233.172629   64.35942  231.294033  220.92060  246.034589
```

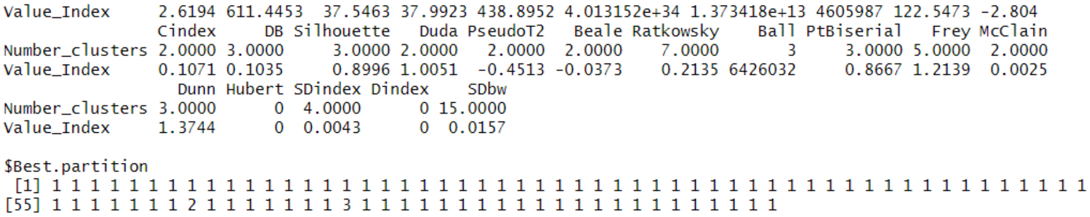
Elbow method

The elbow method has been used to determine the optimal number of clusters in the dataset. When we increase the number of clusters (K), the variance within each cluster tends to decrease. The elbow point, which resembles the bend in the shape of an elbow, is the value of K when this reduction in variance starts to slow down.



According to the above plot we can divide the dataset into 03 clusters.

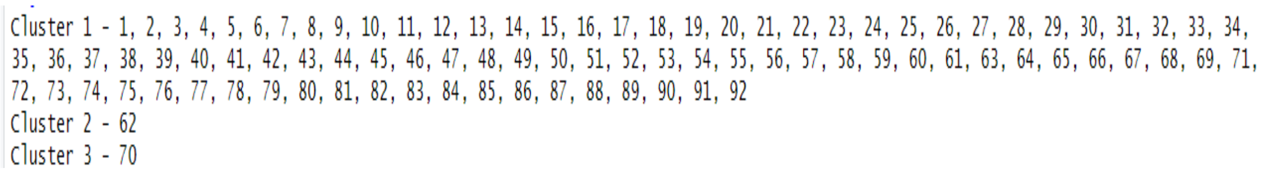
Method2: NbClust method



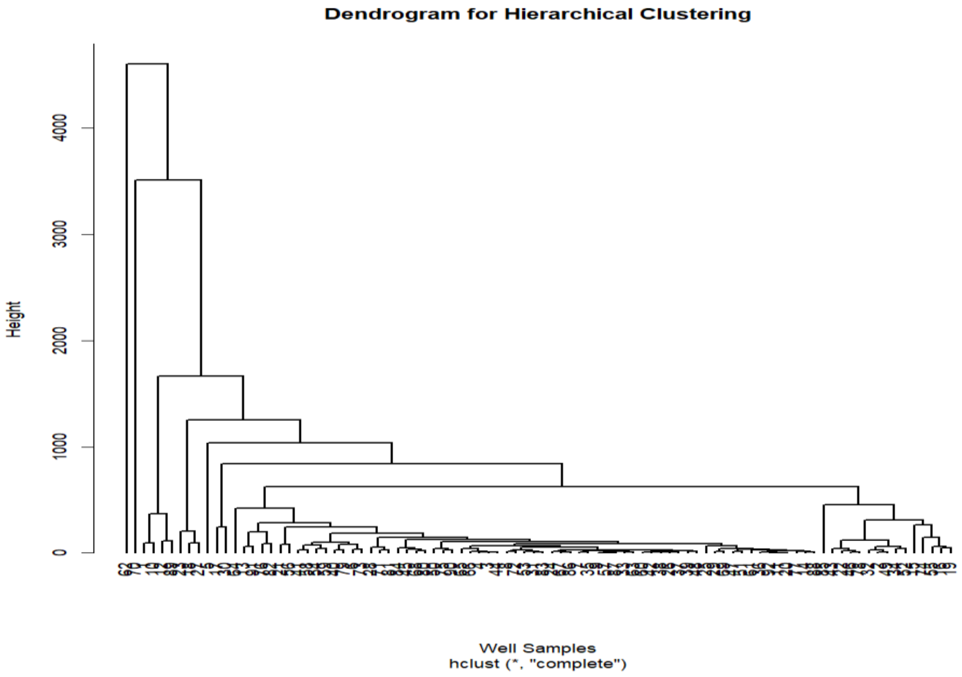
```
Value_Index     2.6194 611.4453  37.5463 37.9923 438.8952 4.013152e+34 1.373418e+13 4605987 122.5473 -2.804
                Cindex       DB Silhouette    Duda PseudoT2    Beale Ratkowsky     Ball PtBiserial    Frey McClain
Number_clusters 2.0000 3.0000     3.0000 2.0000   2.0000  2.0000    7.0000        3     3.0000 5.0000  2.0000
Value_Index     0.1071 0.1035     0.8996 1.0051  -0.4513 -0.0373    0.2135 6426032     0.8667 1.2139  0.0025
                  Dunn Hubert SDindex Dindex    SDbw
Number_clusters 3.0000      0  4.0000      0 15.0000
Value_Index     1.3744      0  0.0043      0  0.0157

$Best.partition
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[55] 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

Dendrogram Interpretation – Used to identify the optimal number of clusters

Samples in each cluster:



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



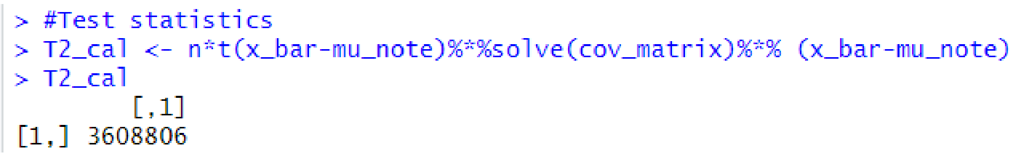
Objective 3: <u>Chemical mixtures in well water samples are in line with the standard accepted values in well water samples.</u>

To fulfill this objective, Hotelling's T-squared test, which is a multivariate statistical test used to Compare the mean values of multiple variables in a dataset with specified standard mean values. By using this test we can identify whether the well water samples in line with accepted standards.

Standard mean values of chemical components are:

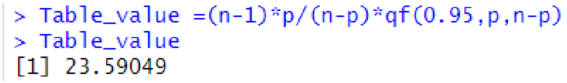
Be = 4, Cr = 100, Fe = 300, Ni = 20, Cu = 1300, As = 10, Cd =5, Ba = 2000, Tl = 0.5, Pb = 15, U = 30

By calculating the Hotelling's T-squared test using observed mean values, the standard mean values, and the covariance matrix. This statistic quantifies the difference between the observed and standard mean values.



```
> #Test statistics
> T2_cal <- n*t(x_bar-mu_note)%*%solve(cov_matrix)%*% (x_bar-mu_note)
> T2_cal
        [,1]
[1,] 3608806
```

Calculate the table value (critical value) based on the F distribution and it depends on the degrees of freedom and a specified significance level (0.05 in this case). This critical value is used to check whether the test statistic is significant.



```
> Table_value =(n-1)*p/(n-p)*qf(0.95,p,n-p)
> Table_value
[1] 23.59049
```

Null Hypothesis (H0): There is no significant multivariate difference between the sample means and the standard means.
Alternative Hypothesis (H1): There is a significant multivariate difference between the sample means and the standard means.

According to the above results, the test statistic value is greater than table value, So it rejects the null hypothesis. Which determines that there is a significant difference between the sample mean values and the standard values. Therefore, well water samples deviate from the standards, which means collected well water samples not in line with accepted standards.

Discussion and Conclusion

In this study, we performed a comprehensive analysis of well water samples, focusing on their chemical composition to achieve three main objectives.

As part of the data preprocessing phase for our initial data exploration, we handled outliers, checked for missing values, and scaled the data to verify that all variables were on a similar scale. For the ensuing analyses to be reliable and valid, the preprocessing stage is essential.

we performed Principal Component Analysis (PCA) to reduce the dimensionality of the eleven chemical components into a smaller set of principal components. Using the PCA results, the coefficients of the PC's are rounded up to two decimal places. Deviation rounded up to three decimal places and the variance get from that rounded up to two decimal places. Then finding the proportion of total variance explained by each component until it gives over 80% of the total variability in the dataset from the total proportions of each component. In this study we got 85.28% for 7PC's.So in this study 07 PCs are sufficient.

We then moved on to cluster analysis, to group the well water samples into homogeneous clusters based on their chemical compositions. We have used two methods to identify the optimal number of clusters. As the first method we used elbow method, and it shows the elbow point at 4. By using NbClust method, it showed 3 clusters. According to these results, we have identified that the number of clusters as 3. Next a dendrogram plot visually illustrated the hierarchical clustering of samples, and we successfully segmented the samples into three distinct clusters. 90 well water samples were allocated to one cluster while the other two samples allocated to two clusters.

In the final stage of our analysis, we assessed the well water samples in comparison to predefined standard mean values to determine their conformity with accepted standards. We have used Hotelling's T-squared test, a multivariate statistical method, to evaluate the multivariate differences between the means of our dataset and the established standard means. Hypotheses were formulated and tested to identify the results. According to the results, the well water samples are not in line with accepted standards.

In conclusion, our thorough analysis of the well water samples provided important information regarding the chemical composition and quality of the water. We were successful in achieving our goals and provided a comprehensive idea of the dataset through data preparation, PCA, cluster analysis, and hypothesis testing. The findings of this study can help stakeholders and decision-makers make correct decisions on the management and quality evaluation of well water sources.

References

[1] "National Primary Drinking Water Regulations," EPA, [Online]. Available: https://www.epa.gov/ground-water-and-drinking-water/national-primary-drinking-water-regulations.

Appendix

R code : code