IOSR Journal of Engineering (IOSRJEN)
ISSN: 2250-3021 Volume 2, Issue 8 (August 2012), PP 29-37
www.iosrjen.org

Use K-Means Cluster Analysis to Study the Classification Of Some Water Factories, According To Some Specifications On The Cover Of The Bottle

Dr. Mohammad M. Faqe Hussain[1] and Dr.Akhterkhan S.Hamad [2]

[1] (*Scholl of Administration and Economics, Statistics Science Department, University of Sulaimani, Iraqi-Kurdistan.*

[2] (*College of Administration and Economics, Statistics Science Department, University of Al-Salahaddin-Hawler, Iraqi-Kurdistan.*

ABSTRUCT: - *In this paper , the k-mean clustering analysis method is used to classify some of the factories for mineral waters in the Kurdistan Region - Iraq, according to some specifications on the cover of the bottle, such as (Ph, Cl, No3), and then compare these some specifications with international standard specifications , in other words " discuss the following question : Are these some specification on the cover of the bottle are identical with the international standard specifications or not?," and after data were collected according to the some specifications on the cover of a bottle and then organized and tabulation and analyzed using statistical package for social science (SPSS) program and apply the method of k-mean clustering analysis , we obtain two types of clusters, the first type of clusters includes the following mineral waters factories (Life, Massafi, Jin, Tiyan, Lolav, Kani-Erbil, Chiaa, Rawan, Kani-Sul, Zhian, Mira, Lava) and the second type of clusters includes the following mineral waters factories (Mazi, Rovian, Hayat, Ala, Shireen, Alhayat).The arithmetic mean for each of the specifications (Ph, Cl, No3) of the first type of clusters calculated and compare the mean of the specifications with the international standard specifications, the result show that, the first types of mineral waters factories are identical to the international standard specifications, in the same way the arithmetic mean of the second type of cluster are calculated and compare the mean of the specifications with the international standards specification, the results show that the second types of mineral waters factories are also identical to the standard classification of the world.*

Keywords: - *Some material of water drinking, K-Mean algorithm, Algorithms, Distance measures, Application of K-Mean clustering, Calculate M-Component for more than 2 attributes, Determine minimum numbers of attributes, Practical Part, Clustering Procedures.*

I. INTRODUCTION

Cluster analysis refers to a large class of techniques designed to classify a multivariate data set into some number of clusters whose members are more similar to one another than to members of other clusters. These techniques are divided into many classes, based on different notions of similarity, different emphasis placed on merging vs. splitting clusters, etc. Here, only one subclass of these techniques is discussed; the class is sufficiently large to allow for a demonstration of the abilities and restrictions of cluster analysis[7][10]. In this paper, use one of the methods of classification (k-mean cluster analysis) to study the classification of some mineral waters factories depending on some specifications such as (Ph,Cl,No3) on the cover of the bottles.

II. SOME MATERIALS OF WATER DRINKING[5][6][13]

2-1 PH:

pH is a measure of the hydrogen ion concentration in water. The pH of water indicates whether the water is acid or alkaline. The measurement of pH ranges from 1 to 14 with a pH of 7 indicating a neutral condition (neither acid nor alkaline). Numbers lower than 7 indicate acidity; numbers higher than 7 indicate alkalinity.

Drinking water with a pH between 6.5 and 8.5 is generally considered satisfactory. Acid waters tend to be corrosive to plumbing and faucets, particularly if the pH is below 6. Alkaline waters are less corrosive. Waters with a pH above 8.5 may tend to have a bitter or soda like taste.

The pH of water may have an effect on the treatment of water and also should be considered if the water is used for field application of pesticides. Water with a pH of 7.0 to 8.5 will require more chlorine for the destruction of pathogens (disease organisms) than will water that is slightly acid.

2-2 NITRATE:

Nitrate (NO3) levels should not be higher than 10 mg/l if reported as nitrogen (N) or nitrate-nitrogen (N-NO3) or higher than 45 mg/l if reported as nitrate (NO3). High nitrate may cause methemoglobinemia (infant cyanosis or "blue baby disease") in infants who drink water or formula made from water containing

nitrate levels higher than recommended. Adults can drink water with considerably higher concentrations than infants without adverse affects. Livestock water can contain up to 100 mg/l of nitrate-nitrogen, but young monogastric animals such as hogs may be affected at nitrate levels of considerably less than 100 mg/l[11].

2-3 CHLORIDE:

High concentrations of chloride ions may result in an objectionable salty taste to water and the corrosion of plumbing in the hot water system. High chloride waters may also produce a laxative effect. An upper limit of 250 mg/l has been set for the chloride ions, although at this limit few people will notice the taste. Higher concentrations do not appear to cause adverse health effects. An increase in the normal chloride content of your water may indicate possible pollution from human sewage, animal manure or industrial wastes. The following chart provides a quick overview of acceptable levels for drinking water

Table 1: Show the overview of acceptable levels for drinking water [6]

A Quick Look at Safe Levels in Drinking Water	
Coliform bacteria	**No coliform bacteria is acceptable**
pH	6.5 – 8.5
Nitrates	< 10 mg/l as NO3 –N < 45 mg/l as NO3
Total dissolved solids (TDS)	< 500 mg/l
Chloride	< 250 mg/l
Fluoride	0.7 – 1.2 mg/l
Calcium and magnesium	Calcium – limits not set by EPA Magnesium > 125 mg/l may show laxative effects
Iron and manganese	Iron < 0.3 mg/l Manganese < 0.05 mg/l
Sodium	< 100 mg/l
Sulfates	< 250 mg/l
Arsenic	< 10 ppb
Conductivity	0.4-0.85 micromoles per centimeter
Total hardness	< 270 mg/l
Turbidity	1 turbidity unit (TU). Note: > 5 TUs are detectable easily in a glass of water and usually are objectionable for aesthetic reasons.
Potassium	No maximum limit has been set
Color	< 10 color units

< means less than
> means greater than
Mg/l means milligrams per liter

III. K-MEANS ALGORITHM [2][8][9]

Let us talk about the K-means algorithm. The algorithm accepts two inputs. The data itself, and "k", the number of clusters. We will talk about the implications of specifying "k" later. The output is k clusters with input data partitioned among them.

The aim of K-means (or clustering) is this : We want to group the items into k clusters such that all items in same cluster are as similar to each other as possible. And items not in same cluster are as different as possible. We use the distance measures to calculate similarity and dissimilarity. One of the important concept in K-means is that of centroid. Each cluster has a centroid. You can consider it as the point that is most representative of the cluster. Equivalently, centroid is point that is the "center" of a cluster.

Simply speaking it is an algorithm to classify or to group your objects based on attributes/features into K number of group. K is positive integer number. The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid . Thus, the purpose of K-mean clustering is to classify the data.

3-1 ALGORITHM'S[9]

The basic step of k-means clustering is simple. In the beginning, we determine number of cluster K and we assume the centroid or center of these clusters. We can take any random objects as the initial centroids or the first K objects can also serve as the initial centroids.

Then the K means algorithm will do the three steps below until convergence

Iterate until *stable* (= no object move group):

1. Determine the centroid coordinate
2. Determine the distance of each object to the centroids
3. Group the object based on minimum distance (find the closest centroid)

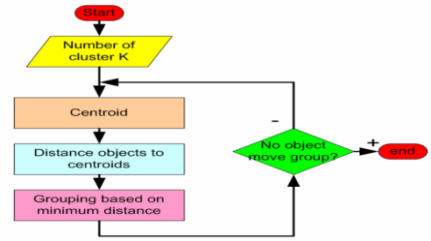

Fig 1: Show the flowchart of *k-means* Algorithm.

This procedure show the pseudo code for k-mean clustering analysis, it is shown by Alpaydin [1] to follow:

Initialize $\mathbf{m}_i$, i = 1,…,k, for example, to k random $\mathbf{x}^t$
Repeat
For all $\mathbf{x}^t$ in X
$b_i^t \leftarrow 1$ if $\| \mathbf{x}^t - \mathbf{m}_i \| = \min_j \| \mathbf{x}^t - \mathbf{m}_j \|$
$b_i^t \leftarrow 0$ otherwise
For all $\mathbf{m}_i$, i = 1,…,k
$\mathbf{m}_i \leftarrow$ sum over t ($b_i^t \mathbf{x}^t$) / sum over t (b_i^t)
Until $\mathbf{m}_i$ converge

The vector **m** contains a reference to the sample mean of each cluster. **x** refers to each of our examples, and **b** contains our "estimated [class] labels" Alpaydin[1].

3-2 DISTANCE MEASURES[10][12][14]

The joining or tree clustering method uses the dissimilarities or distances between objects when forming the clusters. These distances can be based on a single dimension or multiple dimensions. The most straightforward way of computing distances between objects in a multi-dimensional space is to compute Euclidean distances. If we had a two- or three-dimensional space this measure is the actual geometric distance between objects in the space (i.e., as if measured with a ruler). However, the joining algorithm does not "care" whether the distances that are "fed" to it are actual real distances, or some other derived measure of distance that is more meaningful to the researcher; and it is up to the researcher to select the right method for his/her specific application.

Euclidean distance. This is probably the most commonly chosen type of distance. It simply is the geometric distance in the multidimensional space. It is computed as:

$$\text{Distance}(x,y) = \sqrt{\Sigma(x_i - y_i)^2} \qquad \ldots\ldots(1)$$

Note that Euclidean (and squared Euclidean) distances are usually computed from raw data, and not from standardized data.

Squared Euclidean distance. One may want to square the standard Euclidean distance in order to place progressively greater weight on objects that are further apart. This distance is computed as (see also the note in the previous paragraph):

$$\text{Distance}(x,y) = \Sigma(x_i - y_i)^2 \qquad \ldots\ldots(2)$$

City-block (Manhattan) distance. This distance is simply the average difference across dimensions. In most cases, this distance measure yields results similar to the simple Euclidean distance. However, note that in this measure, the effect of single large differences (outliers) is dampened (since they are not squared). The city-block distance is computed as:

$$\text{Distance}(x,y) = \Sigma|x_i - y_i|^2 \qquad \ldots\ldots(3)$$

Chebychev distance. This distance measure may be appropriate in cases when one wants to define two objects as "different" if they are different on any one of the dimensions. The Chebychev distance is computed as:

$$\text{Distance}(x, y) = Maximum|x_i - y_i| \quad \ldots\ldots(4)$$

Power distance. Sometimes one may want to increase or decrease the progressive weight that is placed on dimensions on which the respective objects are very different. This can be accomplished via the *power distance*. The power distance is computed as:

$$\text{Distance}(X, Y) = \sqrt[r]{\Sigma|x_i - y_i|^p} \quad \ldots\ldots(5)$$

Where *r* and *p* are user-defined parameters. A few example calculations may demonstrate how this measure "behaves." Parameter *p* controls the progressive weight that is placed on differences on individual dimensions, parameter *r* controls the progressive weight that is placed on larger differences between objects. If *r* and *p* are equal to 2, then this distance is equal to the Euclidean distance.

IV. APPLICATIONS OF K-MEANS CLUSTRING [2][3]

Clustering algorithms can be applied in many fields, for instance:

- *Marketing*: finding groups of customers with similar behavior given a large database of customer data containing their properties and past buying records;
- *Biology*: classification of plants and animals given their features;
- *Libraries*: book ordering;
- *Insurance*: identifying groups of motor insurance policy holders with a high average claim cost; identifying frauds;
- *City-planning*: identifying groups of houses according to their house type, value and geographical location;
- *Earthquake studies*: clustering observed earthquake epicenters to identify dangerous zones;
- *WWW*: document classification; clustering weblog data to discover groups of similar access patterns.

V. CALCULATE M-COMPONENT FOR MORE THAN 2 ATTRIBUTES[8]

To generalize the k-mean clustering into n attributes, we define the centroid as a vector where each component is the average value of that component. Each component represents one attribute. Thus each point number j has n components or denoted by $P_j(x_{j1}, x_{j2,} x_{j3,\ldots,} x_{jm,\ldots}, x_{jn})$. If we have N training points, then the m component of centroid can be calculated as:

$$\bar{x}_m = \frac{1}{N}\sum_j x_{jm} \quad \ldots\ldots(6)$$

The rest of the algorithm is just the same as above.

VI. DETRMINE MINIMUM NUMBER OF ATTRIBUTE[8]

As you may guess, the minimum number of attribute is one. If the number of attribute is one, each example point represents a point in a distribution. The k-mean algorithm becomes the way to calculate the mean value of k distributions. Figure below is an example of k = 2 distributions.

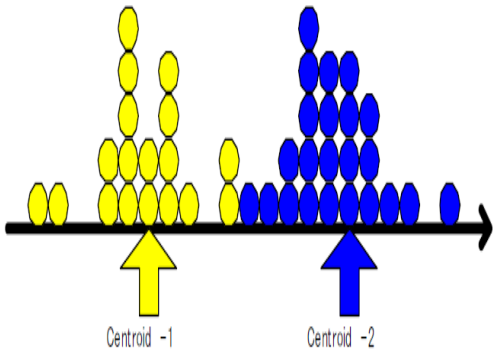

Fig. 2 : Show an example of k=2 distribution

VII. PRACTICAL PARTS

Through the practical part, the data was collected from (19) mineral waters factories in the Kurdistan Region - Iraq, the data was about some types of water in some mineral waters factories according to some specifications on the surface of the bottle such as (Ph,Cl,No3) , a k-mean cluster analysis based on Euclidean distances used to determine a suitable number of segments. The Figure below explain the data of the research and use the program SPSS (Statistical Package for Social Science 19) for the analysis.

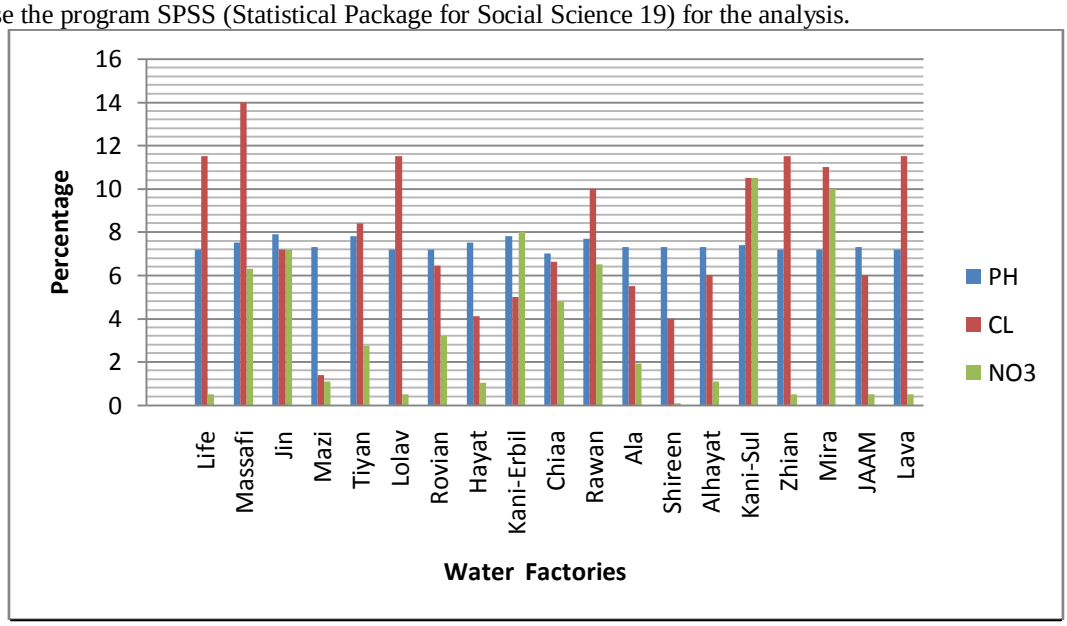

Fig. 3: Show the details of some specifications of water drinking of some water factories

VIII. CUSTERING PROCEDURES

We want to use the k-means method on the data. We have previously seen that we need to specify the number of segments when conducting k-means clustering. SPSS then initiates cluster centers and assigns objects to the clusters based on their minimum distance to these centers. We run a K-Means analysis by clicking → Analyze → Classify → K-Means cluster .The results are shown as follows:-

The Initial Cluster Centers table shows the first step in the k-means clustering in finding the k centers.

Table 2: Show the Initial Cluster Centers

	Cluster	
	1	2
Ph	7.20	7.30
Cl	11.00	1.40
No3	10.00	1.10

The Iteration History table shows the number of iterations that were enough until cluster centers did not change substantially.

Table 3:Show the Iteration History

Iteration	Change in Cluster Centers	
	1	2
1	5.078100884872100	2.661412349463640
2	0.362721491776578	0.380201764209090
3	0.025908677984040	0.054314537744156
4	0.001850619856004	0.007759219677737
5	0.000132187132572	0.001108459953962
6	0.000009441938040	0.000158351421995

7	0.000000674424146	0.000022621631714
8	0.000000048173152	0.000003231661673
9	0.000000003440941	0.000000461665953
10	0.000000000245781	0.000000065952279
11	0.000000000017556	0.000000009421754
12	0.000000000001254	0.000000001345965
13	0.000000000000090	0.000000000192280
14	0.000000000000005	0.000000000027469
15	0.000000000000000	0.000000000003925
16	0.000000000000000	0.000000000000561
17	0.000000000000000	0.000000000000080
18	0.000000000000000	0.000000000000011
19	0.000000000000000	0.000000000000002
20	0.000000000000000	0.000000000000000

- The iteration history shows the progress of the clustering process at each step.
- In early iterations, the cluster centers shift quite a lot.
- By the 6th iteration, they have settled down to the general area of their final location, and the last five iterations are minor adjustments.
- If the algorithm stops because the maximum number of iterations is reached, you may want to increase the maximum because the solution may otherwise be unstable.
- For example, if you had left the maximum number of iterations at 10, the reported solution would still be in a state of flux.

The **Cluster Membership** table gives you the case cluster each case belongs to and the Euclidean distance of each case to the cluster center. Below is a print out of the all cases. Visual inspection of distances is necessary to check for outliers that may not adequately reflect the population

Table 4 :Show the distance of the cluster membership

Case Number	Company	Ph	Cl	No3	Distance			
					No. of Iterations= 5	No. of Iterations= 10	No. of Iterations= 15	No. of Iterations =20
1	Life	7.2	11.5	0.5	4.613	4.613	4.613	4.613
2	Massafi	7.5	14	6.3	4.654	4.654	4.654	4.654
3	Jin	7.9	7.194	7.2	3.515	3.515	3.515	3.515
4	Mazi	7.3	1.4	1.1	3.105	3.105	3.105	3.105
5	Tiyan	7.8	8.397	2.753	2.345	2.345	2.345	2.345
6	Lolav	7.2	11.5	0.5	4.613	4.613	4.613	4.613
7	Rovian	7.2	6.437	3.19	2.955	2.955	2.955	2.955
8	Hayat	7.5	4.1	1.039	0.442	0.442	0.442	0.442
9	Kani-Erbil	7.8	5	8	5.690	5.690	5.690	5.69
10	Chiaa	7	6.614	4.796	3.041	3.041	3.041	3.041
11	Rawan	7.7	10	6.5	1.851	1.851	1.851	1.851
12	Ala	7.3	5.51	1.924	1.395	1.395	1.395	1.395
13	Shireen	7.3	4	0.1	0.997	0.997	0.997	0.997
14	Alhayat	7.3	6	1.1	1.505	1.505	1.505	1.505
15	Kani-Sul	7.4	10.5	10.5	5.855	5.855	5.855	5.855
16	Zhian	7.2	11.5	0.5	4.613	4.613	4.613	4.613
17	Mira	7.2	11	10	5.469	5.469	5.469	5.469
18	JAAM	7.3	6	0.5	1.568	1.568	1.568	1.568
19	Lava	7.2	11.5	0.5	4.613	4.613	4.613	4.613

Table 5 : Show the classification of the mineral waters factories

Case Number	Company	Ph	Cl	No3	Clusters			
					No. of Iterations= 5	No. of Iterations= 10	No. of Iterations= 15	No. of Iterations =20
1	Life	7.2	11.5	0.5	1	1	1	1
2	Massafi	7.5	14	6.3	1	1	1	1
3	Jin	7.9	7.194	7.2	1	1	1	1
4	Mazi	7.3	1.4	1.1	2	2	2	2
5	Tiyan	7.8	8.397	2.753	1	1	1	1
6	Lolav	7.2	11.5	0.5	1	1	1	1
7	Rovian	7.2	6.437	3.19	2	2	2	2
8	Hayat	7.5	4.1	1.039	2	2	2	2
9	Kani-Erbil	7.8	5	8	1	1	1	1
10	Chiaa	7	6.614	4.796	1	1	1	1
11	Rawan	7.7	10	6.5	1	1	1	1
12	Ala	7.3	5.51	1.924	2	2	2	2
13	Shireen	7.3	4	0.1	2	2	2	2
14	Alhayat	7.3	6	1.1	2	2	2	2
15	Kani-Sul	7.4	10.5	10.5	1	1	1	1
16	Zhian	7.2	11.5	0.5	1	1	1	1
17	Mira	7.2	11	10	1	1	1	1
18	JAAM	7.3	6	0.5	2	2	2	2
19	Lava	7.2	11.5	0.5	1	1	1	1

The **Final Cluster Centers** table below allows you to describe the clusters by the variables.

Table 6 : Show the Final Cluster Centers

	Cluster	
	1	2
Ph	7.43	7.31
Cl	9.89	4.78
No3	4.84	1.28

The Differences between Final Cluster Centers in the table below shows the Euclidean distances between the final cluster centers. Greater distances between clusters mean there are greater similarities.

Table 7: Show the Distances between Final Cluster Centers

Cluster	1	2
1		6.231
2	6.231	

Table 8: Show the ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Ph	.054	1	.063	17	.860	.367
Cl	115.621	1	5.398	17	21.419	.000
No3	55.981	1	9.705	17	5.768	.028

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.
According to the data presented in the ANOVA table, Material Cl have maximum influence in forming the clusters and Ph the least.

Table 9 : Show the number of Cases in each Cluster

Cluster	1	12.000
	2	7.000

Table 9 presents data for the number of units in each cluster as well as their total number and missing units (if there are any).

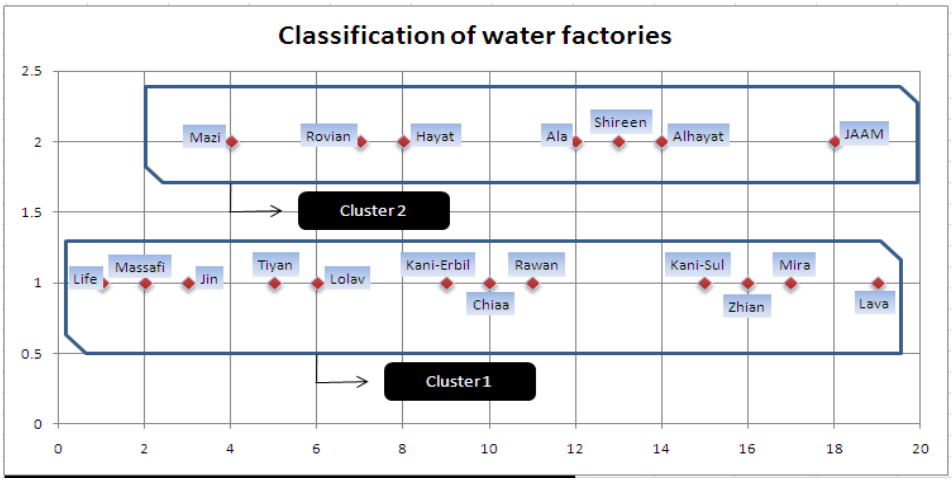

Fig. 4: Show the distribution of mineral waters factories by k-means method

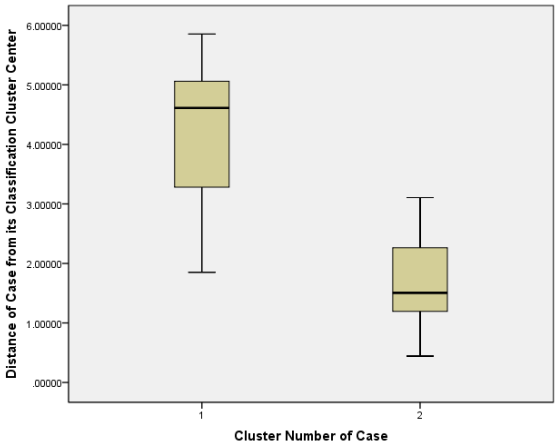

Fig. 5: Show the box-plot for detecting outliers

This is a diagnostic plot that helps you to find outliers within clusters. There is a lot variability in cluster1, but all the distances are not within reason.

Table 10: Show the mean of the Clusters

Specifications	Cluster1	Cluster2
	Arithmetic Mean	Arithmetic Mean
PH	7.425	7.314
Cl	9.892	4.779
No3	4.838	1.279

IX. CONCLUSTION

1) When using the K-Mean Clustering analysis for the classification of the mineral waters factories depending on the some specifications of the waters such as (Ph,Cl,NO3) on the cover of the bottle we can distribute the mineral waters factories to the following clusters, the first cluster includes the following mineral waters factories (Life, Massafi,Jin, Tiyan, Lolav, Kani-Erbil, Chiaa, Rawan, Kani-Sul, Zhian, Mira, Lava) and the second type of cluster fitted with the following mineral water factories(Mazi, Rovian, Hayat, Ala, Shireen, Alhayat,JAAM).
2) The international standards specifications for (Ph) in drinking water between (6.5-8.5) mg/l and the arithmetic mean for the first type of cluster is equal (7.425)mg/l but the arithmetic mean for the second type of cluster is equal (7.314)mg/l when we comparing both clusters with the international standards specifications the (Ph) for the both clusters within the limits of international standard specifications s, and this is a good sign for the mineral waters factories for drinking water in the Kurdistan Region – Iraq
3) The arithmetic mean for the specifications or materials (Cl ,NO3) for the first and second cluster are equal to (9.892,4.779) (4.838,1.272) respectively, and compare the arithmetic mean for both clusters with the international standards specification the result show that the both clusters are identical with the international standards specifications.
4) In the K-Mean Clustering analysis whenever the numbers of iterations are increasing the value of the distance are fixed .In other words does not change the value of the distance.
5) In the K-Mean Clustering analysis whenever the numbers of iterations are increasing the Classification of the mineral waters factories does not change.
6) In the K-Mean Clustering analysis whenever the numbers of iterations are increasing the number of case in each cluster does not change. In the table (9)the number of case in the first cluster = 12 and the number of case in the second cluster = 7.

X. RECOMANDATIONS

1. They need to conduct more studies and research in this area because the water is the main source of life.
2. Use other methods of cluster analysis in the analysis of this type of data .
3. Use more precise specifications other than the specifications in the study of the water for the classification of factories in the Kurdistan Region.

REFERENCES

[1] Alpaydin, Ethem. *Introduction To Machine Learning*. Cambridge, Massachusetts: MIT Press. 2004.
[2] Anderberg, M.R., (1973), Cluster Analysis for Applications, Academic Press, New York, pp.162-163.
[3] Denderfer, M. S., and R. K. Blashfield, (1984), Cluster Analysis, Newbury Park: Sage Publications.
[4] P. Arabie, L. J. Hubert, and G. De Soete.,(1996),Clustering and Classification. World Scientific,
[5] Robert L.Mahler ,Alex Colter,and Ronda Hirnyck, Quality Water for Idaho-Nitrate and Groundwater, (2007),university of Idaho, College of Agricultural and life science, CIS 872.
[6] Roxanne ,J. and Tom,S.(2012),Drinking water quality: Testing and interpreting your results,WQ-1341
[7] Steven M.Holland (2006),Cluster analysis ",Department of Geology,University of Georgia,Atbens,GA 30602-2501
[8] Kaufman,L. and Rousseeuw,P.J. ,(1990),Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley and Sons.
[9] Teknomo,Kardi,K-Means Clustering Tutorials.http:\\people.revoledu.com\kardi\ tutorial\kMean\ Last Update: July 2007
[10] Thomas B. Fomby, April 2008 and 2010, Cluster Analysis, Southern Methodist University , Department of Economics, Dallas, TX 75275
[11] Winton, E.E., Tardiff, R.G., and McCabe L.J., January 1971, Nitrate in Drinking Water, Jour. AWWA, 63:95
[12] See website http://WWW.statsoft.com/textbook/cluster-analysis
[13] U.S. EPA, Secondary Drinking Water Standards: http://www.epa.gov/safewater/consumer/2ndstandards.html
[14] See website http://mathworld.wolfram.com