Cereal-Data Analysis

Department of Computer Science
Course: Data Warehousing & Mining
(CS-626)
Faculty: Dr. Tahseen Jillani
Batch: MCS Final 2021
4th Semester

Group Members:
Muhammad Taha (P19101040)
Asma Hassan Khan (P19101011)
Farhan Aslam (P19101036)
Rayyan Jamil (P19101054)

1 TABLE OF CONTENTS

2 DATASET INFORMATION:

The data were collected by students of Paul Velleman at Cornell University in the early 1990s. Specifically, a local Wegmans supermarket. It is not sure that if the data represent all the cereal available in the store or just a sample (potentially non-random).

Probably, the date of data collection would be between 1990 and 1993 as that is the time frame when the FDA passed a bill that mandated that all packaged foods in the US must have a nutrition label (source, see the section titled, "PASSAGE OF THE NUTRITION LABELING AND EDUCATION ACT (NLEA) OF 1990").

2.1 DETAILED INFO:

1. Title: Cereal Data
2. Number of Instances: 77
3. Number of Attributes: 13
4. For Each Attribute: (all numeric-valued)
    1. calories: calories per serving
    2. protein: grams of protein
    3. fat: grams of fat
    4. sodium: milligrams of sodium
    5. fiber: grams of dietary fiber
    6. carbo: grams of complex carbohydrates
    7. sugars: grams of sugars
    8. potass: milligrams of potassium
    9. vitamins: vitamins and minerals - 0, 25, or 100, indicating the typical percentage of FDA recommended
    10. shelf: display shelf (1, 2, or 3, counting from the floor)
    11. weight: weight in ounces of one serving
    12. cups: number of cups in one serving
    13. rating: a rating of the cereals (Possibly from Consumer Reports?)

3 ABOUT PROJECT:

3.1 SUMMARY:

The purpose of this research was to model the comparison of certain factors of the cerealsof different brands. The main objectives were whether which cereal is better or worst in quality and taste as well. The objective is to predict based on diagnostic measurements whether a cereal is good or worst. In this study, the dataset we used from the internet kaggle.

3.2 INTRODUCTION:

The project that we are working one that is we have different cereal companies products having data of their cereal products and we are applying different techniques using R. The target was to check whether which cereal is better or worst.

3.3 Understanding Important Features That Cerals Are Based On:

3.3.1 Attributes in Our Model:

1. calories: calories per serving
2. protein: grams of protein
3. fat: grams of fat
4. sodium: milligrams of sodium
5. fiber: grams of dietary fiber
6. carbo: grams of complex carbohydrates
7. sugars: grams of sugars
8. potass: milligrams of potassium
9. vitamins: vitamins and minerals - 0, 25, or 100, indicating the typical percentage of FDA recommended
10. shelf: display shelf (1, 2, or 3, counting from the floor)
11. weight: weight in ounces of one serving
12. cups: number of cups in one serving
13. rating: a rating of the cereals (Possibly from Consumer Reports?)

3.4 Methodology (Approach):

The methodology used is based on Clustering and PCA. First of all, the ggplotsbetween different variables were formed. The dendrogram represents the similarity between attibutes or variables. By using NbClust package, WSS was constructed to exactly found the number of clusters tobe created.After that, the Assumption of K-means clustering and PCA was also checked to ensure that our analysisshould be well. Simmilarity and Dissaimilaritybetween the variables in the form of Euclidean Distance were also being calculated.

3.5 Conclusion:

Our results show that the Almond Delight cereal is good in term of quality and health, whereas Smacks cereal is good in taste. Also Nutri-Grain Almond-Raisin cereal is worst in term of taste and Wheat Chex cereal is bad in term of quality.

4 R WORKING:

4.1 LIBRARIES:

The following R libraries that we used here are library(factoextra)

- library(NbClust)
- library(cluster)
- library(stats)
- library(ggplot2)
- library(ggfortify)
- library(dplyr)

4.1.1 Factoextra R Package:

Factoextra is an R package that allows you to easily extract and visualize the output of exploratory multivariate data analysis such as Principal component analysis (PCA) is used to summarize the information contained in continuous (ie, quantitative) multivariate data. Reduce the dimensions of your data without losing important information.

There are several R packages that implement principal component methods. These packages include FactoMineR, ade4, stats, ca, MASS, and Ex Position.

However, the display of the result differs depending on the package used. An R package called factoextra is used to assist in the interpretation and visualization of multivariate analyzes such as cluster analysis and dimensionality reduction analysis. The factoextra R package provides a flexible and easy-to-use method to quickly extract analysis results from the various packages mentioned above in standard human readable data formats. Reduce typing to create elegant ggplot2-based data visualizations.

4.1.2 NbClust Package:

It is used for figuring out the first-rate quantity of clusters.NbClust bundle presents 30 indices for figuring out the quantity of clusters and proposes to a consumer the first-rate clustering scheme from the distinct effects acquired via way of means of various all mixtures of numerous clusters, distance measures, and clustering strategies.

The following distance measures are written for 2 vectors x and y. They are used while the facts is a d-dimensional vector bobbing up from measuring d traits on every of n items or individuals.

• **Euclidean distance**:

Usual rectangular distance among the 2 vectors (2 norm).

4.1.3 Cluster Package:

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other. The cluster package perform this task for a dataset given.

4.1.4 Stats Package:

The R package called Stats Package is developed for Statistical computing. It provides tools for statistical calculation s and the generation of random numbers. For installation of Stats package. The installation of CRAN package is must.

4.1.5 Ggplot2Package:

ggplot () is used to create the first plot object, most often followed by a + to add a component to the plot. There are three common ways to call ggplot ().ggplot2 is the most elegant and aesthetically pleasing graphic framework available in R. It has a well-planned structure.

4.1.6 Ggfortify Package:

Data visualization tools for statistical analysis results. Unified plotting tools for statistics commonly used, such as PCA families, clustering and survival analysis. This package offers a single plotting interface for these analysis result and plots in a unifies style using 'ggplot2'.

4.1.7 DplyrPackage:

Dplyr is a new package which provides a set of tools for efficiently manipulating datasets in R. dplyr is the next iteration of plyr, focusing on only data frames, dplyr is faster, has more consistent API and should be easier to use.

5 ASSUMPTIONS FOR K-MEANS:

5.1 DETERMINING NUMBER OF CLUSTERS:

Determining Optimal Number of Clusters Sum of squares of distances of every data point from its corresponding cluster centroid should be as minimum as possible. We use a method called ELBOW method to find the appropriate number of clusters. The parameter which will be taken into consideration is Sum of squares of distances of every data point from its corresponding cluster centroid which is called WSS (Within Sums of Squares). Steps involved in ELBOW method are:

- ✓ Perform K-means clustering on different values of K ranging from 1 to any upper limit.
- ✓ For each K, calculate WSS (Within Cluster Sum of Squares).
- ✓ Plot the value for WSS with the number of clusters K.

The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters. i.e the point after which WCSS doesn't decrease more rapidly is the appropriate value of K.

5.2 Comparing with K-Means clustering algorithm.

You would possibly have heard approximately the k-approach clustering algorithm; if not, test this tutorial. There are many essential variations among the 2 algorithms, despite the fact that any individual can carry out higher than the alternative in one of a kind cases. Some of the variations are:

1. Distance used: Hierarchical clustering can certainly take care of any distance metric at the same time as k-approach depend upon Euclidean distances.

2. Stability of consequences: k-approach calls for a random step at its initialization which could yield one of a kind consequences if the system is re-run. That wouldn`t be the case in hierarchical clustering.

3. Number of Clusters: While you may use elbow plots, Silhouette plot etc. to discern the proper range of clusters in k-approach, hierarchical can also use all of these however with the delivered gain of leveraging the dendrogram for the same.

4. Computation Complexity: K-approach is much less computationally pricey than hierarchical clustering and may be run on huge datasets inside an inexpensive time frame, that is the principle because k-approach is extra popular.

5.3 Implementation.

- Randomly select k points as the focus / cluster center.
- Assign data points to the closest cluster based on Euclidean distance Calculate the centroids of all points in the cluster
- Repeat until it converges. (In successive iterations, the clusters are assigned the same points) However, the only problem with this implementation is the sensitivity of initialization. Choosing different focal points during the initialization stage creates different clusters. The workaround for the problem looks like this:
- Repeating k means different initializations and choosing the best result.
- Instead of using random initialization to use a smart initialization process such as K means ++. In some cases, it is difficult to interpret centroids, for example, if you are dealing with text data, centroids are not interpretable. An approach to deal with this would be to use the K medoids algorithm. It would select the most centered member within the data as a cluster center and is generally more robust to outliers than other means.

5.4 Advantages of using K means:

- The algorithm in most cases runs in linear time.
- Simple and intuitive to understand

5.5 Limitations of using K means:

- A number of clusters need to be known beforehand.
- It is not very robust to outliers.
- Does not work very well with non convex shapes.
- Tries to generate equal-sized clusters.

5.6 WssPlot Technique.

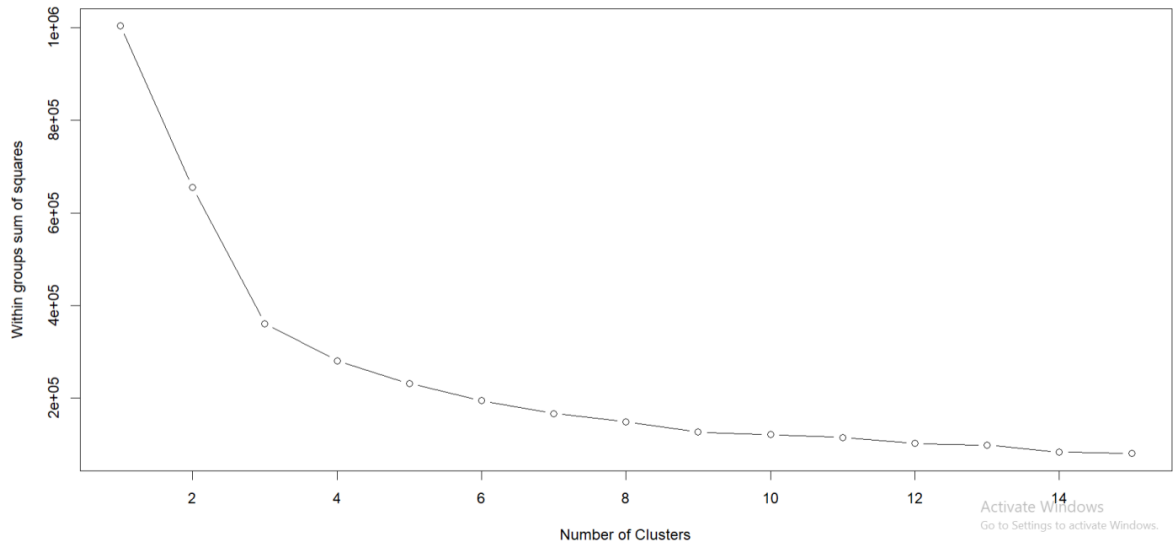

Figure 5.1

Classification and clustering are essential obligations that are there in statistics mining for long, Classification is utilized in supervised studying (Where we've got a based variable) even as clustering is utilized in un-supervised studying in which we don`t have any information approximately based variable. Clustering facilitates to organization comparable statistics factors collectively even as those

corporations are drastically special from every other.

Summary.

The k-means clustering algorithim is applied to our choosen dataset to find groups which have not been explicitly lableled in the data. The number clusters to be formed was finded out by using the WSSplot technique and the number of clusters formed are 4.

5.7 CLUSTERING :

Clustering is the process of grouping similar data. It falls into the category of unsupervised learning. NS. There is no labeled answer in the input data. Clustering algorithms are used in a variety of areas such as finance, medical, and e-commerce. In clustering, each data point belongs to a cluster, but a single data point cannot exist in more than one cluster. The performance of the clustering algorithm can be measured by metrics such as the Dunn Index (DI). The larger the inter-cluster distance (well separated) and the smaller the inter-cluster distance (compact), the higher the DI value. Based on the dataset, you can use several approaches to clustering, such as **Hierarchical Clustering**.

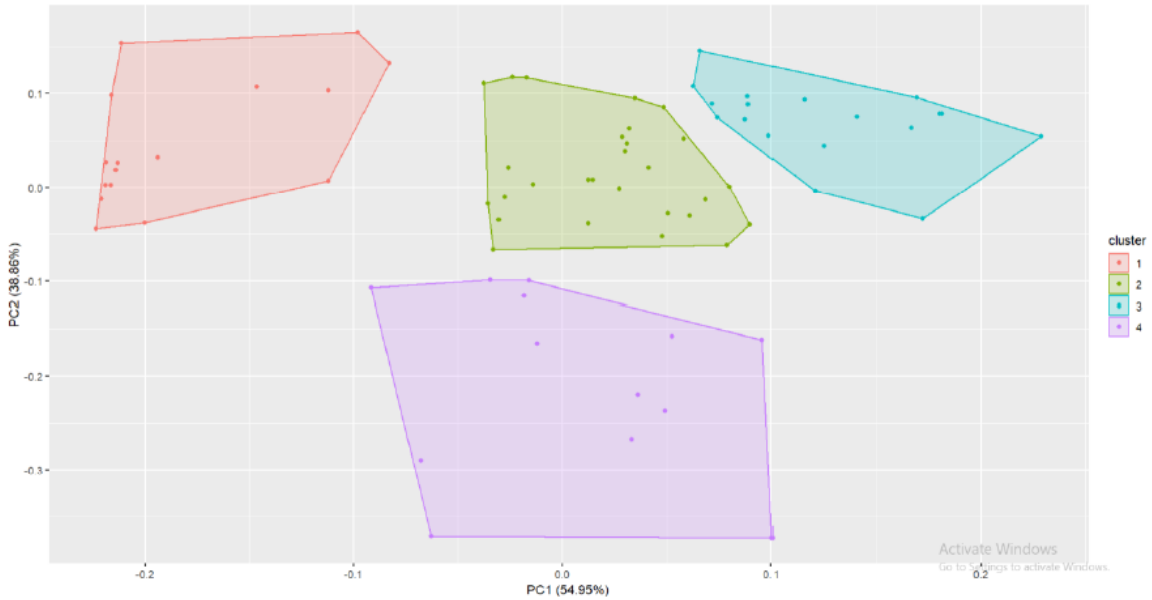

Figure 5.2

Fig 5.2 The pictorial plotting of k-means clusters.

5.8 Dendogram.

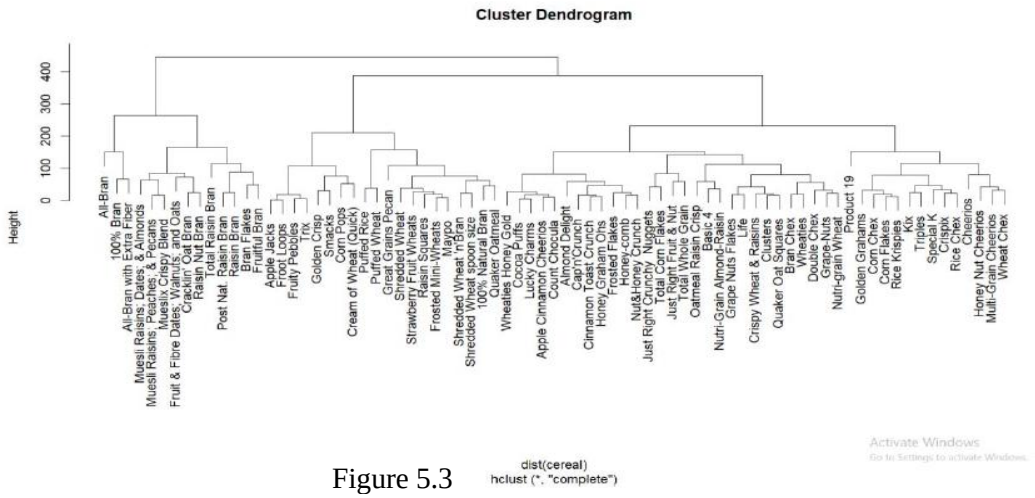

Figure 5.3

Fig. 5.3 shows the above dendogramwhich represents complete linkage between the attributes of given dataset.

5.9 HIERARCHICAL CLUSTERING DENDROGRAM:

They organize the data into nested sequences of groups that can be viewed in the form of dendrograms or tree structures. This is a set of methods for automatically classifying samples into groups using a relevance scale so that the samples in a group are similar and the samples in different groups are not.

Most aggregated hierarchical clustering algorithms are variants of single-link or fully-linked algorithms. These two basic algorithms differ only in how they characterize similarities between pairs of clusters. For the single-link method, the distance between the two clusters is the minimum distance between a pair of samples extracted from the two clusters (one element from the first cluster and the other element from the second cluster). Root-mean-square error of the entire clustering space, K. Including the cluster is the total variation within the cluster. The goal of the root-mean-squared clustering technique is to find a partition that contains a K cluster that minimizes the E2k of a particular K. The K means split cluster algorithm is the simplest and most widely used algorithm that uses the root-means-squared criterion.

Start with a random initial partition and continue to cluster the sample based on the similarity between the sample and the cluster until the convergence criteria are met. Most hierarchical clustering techniques are not based on the concept of optimization. The goal is to find a suboptimal solution of approximation by using iteration to improve the partition to convergence. Hierarchical cluster analysis algorithms can be divided into two categories: divisible algorithms and agglutination algorithms.

The aggregation algorithm initially considers each object as an initial cluster. The cluster is merged into the coarser partition and the merging process continues until it reaches the trivial partition. That is, all objects are in a large cluster. This clustering process is a bottom-up process where partitions are divided from fine to coarse. Aggregation algorithms are generally used more often than divisible methods in real-world applications. The availability of a large collection of clustering algorithms in the literature and in various software environments can easily confuse users who are trying to choose

the right approach to the problem at hand. It is important to note that there is no one-size-fits-all clustering technique for revealing the various structures that exist in multidimensional datasets. Understanding the user in question and the corresponding data type is the best criterion for choosing the right method. Since similarity is the basis of the definition of a cluster, the quality of the clustering process depends on this decision, so this measure should be chosen very carefully. The results of both the divisible and agglutinating clustering techniques are displayed in the form of a two-dimensional diagram.

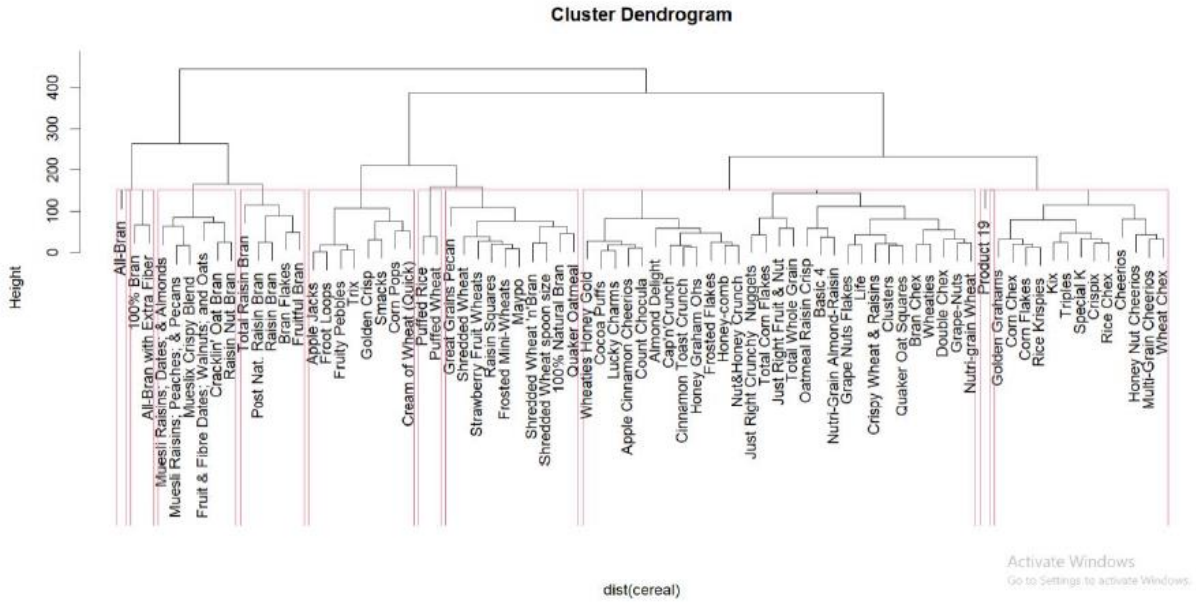

Figure 5.4

Fig. 5.4 shows the rectangles in the hierarchical cluster dendogram represents the number of clusters i.e 9 clusters are formed and shows what attribute belongs to which cluster.

Summary.

We applied the hierarchical clustering to our data and in hierarchical clustering we applied the agglomerative clustering since this algorithm is more efficient and used for the industrial uses. The dendogram formed out of this data is of complete linkage.which shows the sequence of cluster fusion and the distance at which each fusion took place.

6 PRINCIPAL COMPONENT ANALYSIS(PCA):

Principal component analysis (PCA) is an unsupervised machine learning technique. Perhaps the most specialized application of principal component analysis is dimensionality reduction. Principal component analysis (PCA) is becoming a popular tool for identifying well-constructed patterns from complex biological datasets. That is, the PCA captures the core of the data with several key components that convey most of the variations in the dataset.

6.1 PLOTTING OF PRINCIPLE COMPONENT ANALYSIS:

Not only can PCA be used as a data preparation technique, but it can also be used for data visualization. Visualized data makes it easy to gain insights and determine the next steps in a machine learning model. The PCA plot shows a cluster of samples based on similarity. PCA does not discard attributes or characteristics. Instead, reduce a huge number of dimensions by creating a key component (PC). PCs describe variation and account for the varied influences of the original characteristics. Such influences, or loadings, can be traced back from the PCA plot to find out what produces the differences among clusters.
We used Principal component analysis (PCA) technique for reducing the dimensionality of such datasets, increasing interpretability but at the same time minimizing information loss. It does so by creating new uncorrelated variables that successively maximize variance.

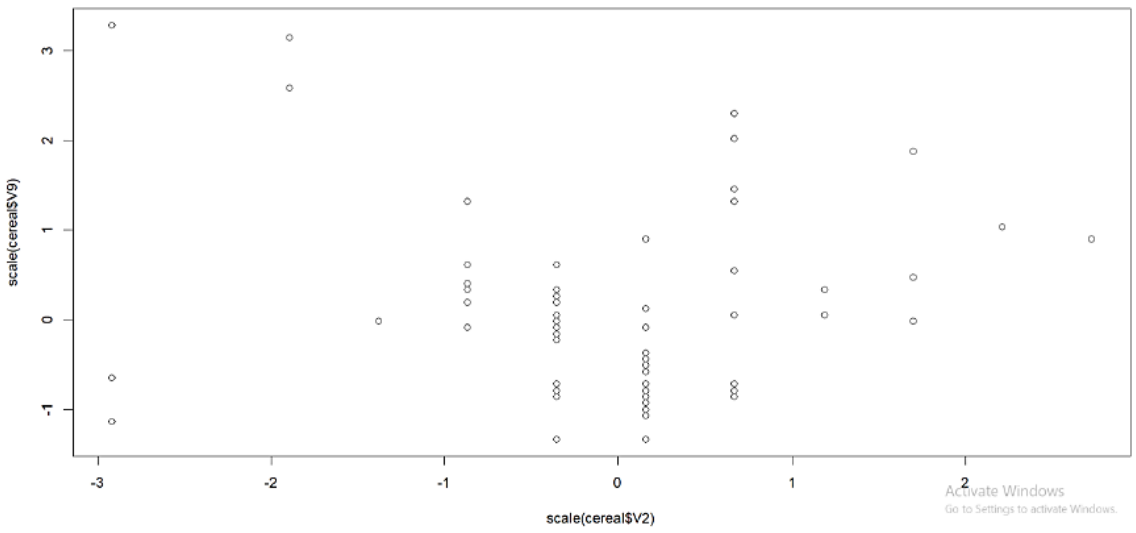

Figure 6.1

Fig 6.1 represents the plot of principle component analysis between cereals.

6.2 Mathematics Behind PCA.

- Take the whole dataset consisting of d+1 dimensions and ignore the labels such that our new dataset becomes d dimensional.
- Compute the mean for every dimension of the whole dataset.
- Compute the covariance matrix of the whole dataset.

$$cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n} (Xi - \bar{x})(Yi - \bar{y})$$

- Compute eigenvectors and the corresponding eigenvalues.
- Sort the eigenvectors by decreasing eigenvalues and choose k eigenvectors with the largest eigenvalues to form a d × k dimensional matrix **W.**
- Use this d × k eigenvector matrix to transform the samples onto the new subspace.

Summary.

We used Principal component analysis (PCA) technique for reducing the dimensionality of such datasets, increasing interpretability but at the same time minimizing information loss. It does so by creating new uncorrelated variables that successively maximize variance.

7 SIMILARITY MEASURE /DISTANCE MEASURE:

Clustering includes grouping positive items which can be just like every other, it is able to be used to determine if gadgets are comparable or varied of their properties. In a Data Mining sense, the similarity degree is a distance with dimensions describing item features. That approach if the gap amongst records factors is small then there may be a excessive diploma of similarity some of the items and vice versa. The similarity is subjective and relies upon closely at the context and application. Most clustering procedures use distance measures to evaluate the similarities or variations among a couple of items, the maximum famous distance measures used are:

7.1 Euclidean Distance:

The class of observations into organizations calls for a few techniques for computing the space or the dissimilarity among every pair of observations. The end result of this computation is called a dissimilarity or distance matrix. There are many techniques to calculate this distance information; the selection of distance measures is a vital step in clustering. It defines how the similarity of elements (x, y) is calculated and it'll affect the form of the clusters. The preference of distance measures is a vital step in clustering. It defines how the similarity of elements (x, y) is calculated and it'll affect the form of the clusters.

- ✓ **Mathematical formula for Euclidean distance.**

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$$

The classical strategies for distance measures are Euclidean and Manhattan distances, which might be described as follow:

Euclidean distance is taken into consideration the conventional metric for issues with geometry. It may be genuinely defined because the regular distance among points. It is one of the maximum used algorithms within side the cluster analysis. One of the algorithms that use this system could be K-mean. Mathematically it computes the foundation of squared variations among the coordinates among objects.

7.2 Summary.

The Euclidean distance was used to find out the similarity between the different brands producing cereals that shows how the calories or sodium or etc are similar to each of the different brands

8 GGPLOT:

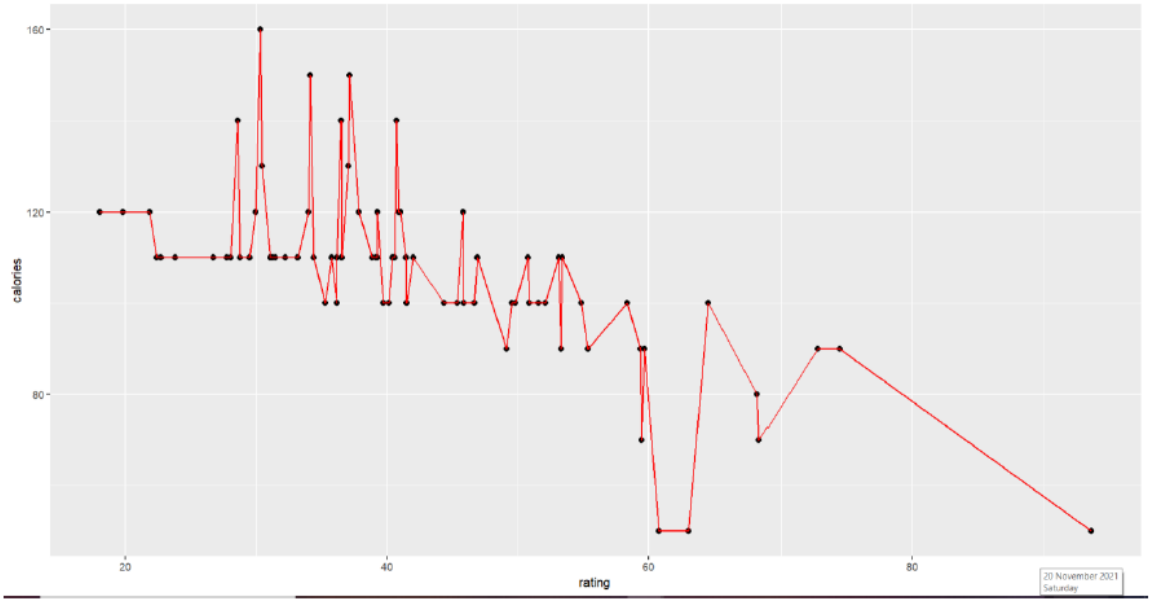

Figure 7.1

fig 7.1 represents the ggplot of calories along with the ratings of different cereal manufacturing companies.

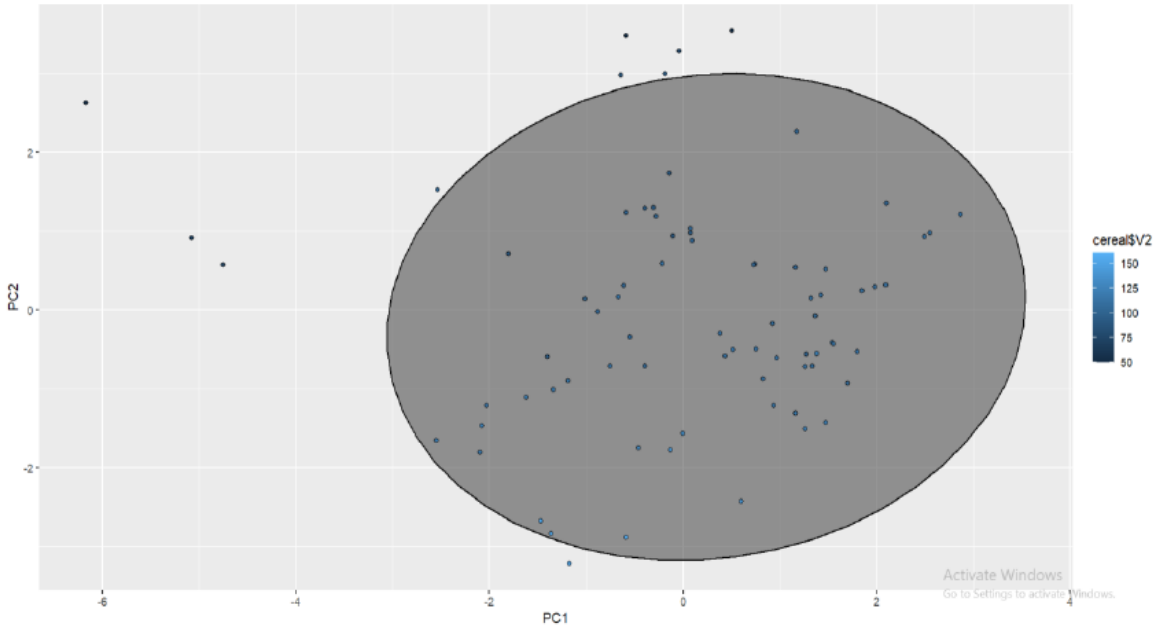

Figure 7.2

fig 7.2 show the ggploting of principle component analysis considering PC1 and PC2, this shows the similarity between PC1 and PC2.

9 CALCULATIONS:

```
> # Calculating the euclidean distance
> dd <- dist(mydata[71:77, -1], method = "euclidean")
> dd
          71        72        73        74        75        76
72 120.92560
73 195.69875 103.24243
74 223.99777 128.64680 116.06464
75 143.45731  80.96295  58.77074 127.69886
76 142.32006  75.03333  70.91544 104.55620  30.41381
77 186.21224  90.32165  50.54701  69.65630  62.88879  50.29911
```

the distance matrix by using Euclidean method.

```
Importance of components:
                          PC1    PC2    PC3    PC4    PC5     PC6     PC7     PC8     PC9
Standard deviation     1.6334 1.4324 1.2756 0.9984 0.8016 0.72328 0.61431 0.26177 0.21150
Proportion of Variance 0.2964 0.2280 0.1808 0.1108 0.0714 0.05813 0.04193 0.00761 0.00497
Cumulative Proportion  0.2964 0.5244 0.7052 0.8160 0.8874 0.94549 0.98742 0.99503 1.00000
```

Calculation of standard dev, variance and cumulative proportion of PCA's formed.

10 REFERENCES:

1. Analytics, Perceptive. “How to Perform Hierarchical Clustering Using R | R-Bloggers.”
2. R-BLOGERS, 18 Dec. 2017, www.r-bloggers.com/2017/12/how-to-perform-hierarchical-clustering-using-r/#:~:text=In%20Divisive%20method%20we%20assume. Accessed 13 Dec. 2021.
3. DataCamp. “NbClust Function | R Documentation.”Rdocumentation.org, 2013, www.rdocumentation.org/packages/NbClust/versions/3.0/topics/NbClust.
4. Davidson, Ian, and S. S. Ravi. “Agglomerative Hierarchical Clustering with Constraints: Theoretical and Empirical Results.”
5. Knowledge Discovery in Databases: PKDD 2005, 2005, pp. 59–70, 10.1007/11564126_11. Accessed 13 Sept. 2020.
6. Gulzar, Maria. “K-Means Clustering: Concepts and Implementation in R for Data Science.”Medium, 10 June 2021, towardsdatascience.com/k-means-clustering-concepts-and-implementation-in-r-for-data-science-32cae6a3ceba. Accessed 13 Dec. 2021
7. Hayden, Luke. “PCA Analysis in R.”DataCamp Community, 2018, www.datacamp.com/community/tutorials/pca-analysis-r.
8. “Hierarchical Clustering in R Programming.”GeeksforGeeks, 18 June 2020, www.geeksforgeeks.org/hierarchical-clustering-in-r-programming/. Accessed 13 Dec. 2021.
9. “Hierarchical Clustering in R: Dendrograms with Hclust.”DataCamp Community, 24 July 2018, www.datacamp.com/community/tutorials/hierarchical-clustering-R. Accessed 13 Dec. 2021.
10. Kassambara, Alboukadel, and Fabian Mundt. “Factoextra: Extract and Visualize the Results of Multivariate Data Analyses.”
11. R-Packages, 1 Apr. 2020, cran.r-project.org/web/packages/factoextra/index.html.
12. Lin, Zhenhua, et al. “Interpretable Functional Principal Component Analysis.”
13. Biometrics, vol. 72, no. 3, 18 Dec. 2015, pp. 846–854, 10.1111/biom.12457.
14. “Principal Components Analysis in R.”YouTube, 10 July 2017,www.youtube.com/watch?v=xKl4LJAXnEA. Accessed 3 Dec. 2019.
15. “RPubs - K-Means Clustering.”Rpubs.com, rpubs.com/violetgirl/201598.
16. Zach. “How to Calculate Euclidean Distance in R (with Examples).”Statology, 16 Oct. 2020, www.statology.org/euclidean-distance-in-r/. Accessed 13 Dec. 2021.