Xi’an Jiaotong-liverpool University

Department of Science

MTH301

Final Year Project

CLUSTERING METHODS AND ITS APPLICATION ON NUTRITION FACTS OF MCDONALD’S BREAKFAST MENU

聚类方法及其在麦当劳早餐菜单营养成分研究中的应用

Student name: Zhiyi Wang

Student ID: 1821891

Supervisor: Mu He

Date:4/May/2022

Abstract

During the period of the epidemic, due to social distancing restrictions, home fitness and exercise have prevailed, and healthy diet has also received widespread attention. In addition to requiring a balanced diet of meat and vegetables when cooking, people also pay great attention to nutritional content when purchasing packaged food and beverages. The Nutrition Facts list on packaged food and beverages is an important tool to help make healthy dietary choices. At the same time, the Nutrition Facts list of food provided by large chain restaurant enterprises such as McDonald's can also be viewed on their official website, which is very helpful for busy people who are preferring fast food to pursuit of healthy diet. This paper obtains a data set from the McDonald' s official website, which includes 42 common McDonald's breakfasts and the corresponding Nutrition Facts list including 21 nutritional content labels. This paper will analyze the data set through unsupervised machine learning algorithms, cluster 42 kinds of McDonald's breakfasts according to their nutritional components, and make inferences based on the clustering results to give recommendations for healthy diet. This paper firstly uses the principal component analysis method for dimension reduction and obtains four principal components to explain most of the information, and then uses Hierarchical Clustering, K-means Clustering, DBSCAN and spectral clustering for analysis. The clustering results are projected to these four principal components to implement visualization. In the study, it is found that the results of the four clustering methods are different, but also have some similarities. Except for Hierarchical Clustering, the other three clustering methods produce the largest cluster consisting of hamburger-like breakfasts; while Hierarchical Clustering divides hamburger-like breakfasts into two clusters based on the content of cholesterol and vitamin A. The rest of the breakfasts are clustered differently in different clustering methods. In the future, we can conduct in-depth research on the menus of large chain catering companies and learn more clustering methods.

在疫情持续的这段时间里，因为社交距离的限制，家庭健身盛行，健康饮食也受到广泛关注。人们除了在自己烹饪时要求荤素搭配，营养均衡，在购买包装食品和饮料时，也非常关注营养成分。包装食品和饮料上的营养成分表是帮助人们做出健康饮食选择的重要工具，像麦当劳这样的大型连锁餐饮企业所提供的食物的营养成分表也能在其官方网站查阅到，这对于忙碌的偏好快餐的人们追求健康饮食有很大帮助。本文从麦当劳官方网站上获取了数据集，其中包括 42 种常见的麦当劳早餐和包括 21 种营养成分的营养成分表。本文将通过无监督学习对这些数据进行分析，依据营养成分对 42 种麦当劳早餐进行聚类，并依据聚类结果进行推断并给出健康饮食的建议。本文先使用主成分分析法进行降维，获得四个主成分来解释大部分信息，然后分别使用层次聚类，K 均值聚类，DBSCAN 和谱聚类进行分析，并将聚类结果投影在前四个主成份上进行可视化。通过研究，我们发现四种聚类的结果不同，但也具有一些相似性。在除层次聚类以外的三个聚类方法中，类汉堡早餐会组成数量最多的一类；而层次聚类依据胆固醇和维他命 A 的含量将类汉堡早餐分为两类。其余的早餐类型在不同聚类方法中聚了结果不同。未来我们还可以深入大型连锁餐饮企业菜单的研究，并学习更多聚类方法。

Contents

Introduction

The foods people consume have a profound impact on their health. The scientific link between dietary choices and healthy living has been documented for decades, and healthy diet habits have a significant positive impact on healthy living.(Chrzan, J. and Brett, J. A., 2019)[1] The most basic requirement for a healthy diet is to focus on having nutrient-dense foods that provide the body with vitamins, minerals, and other health-promoting ingredients, and eating added sugar, trans fat, saturated fat, and sodium as little as possible.(Kostecka, M., 2022)[2] The FDA proposed and actively promoted the 2,000-calorie-a-day dietary pattern, suggesting that an adult's daily energy intake should be based on 2,000 calories, with small adjustments.(Silverglade, B. A., 1996)[3] This value is presented because of the Nutrition Labeling and Education Act(NLEA) which passed in 1991. This act commands the Food and Drug Administration(FDA) to make changes to food labeling regulations, which means food labels will be standardized that were previously left to the discretion of manufacturers and states.(Balasubramanian, S. K. and Cole, C., 2002)[4] The new labeling regulations require that all food labels must include a mandatory nutrition statement. In addition to vitamins and minerals, all nutrition facts must be displayed in specific units.(Jebaraj Asirvatham, Paul E. McNamara and Kathy Baylis, 2017)[5] Daily Reference Value(DRV/DV) and percent Daily Value (%DV) are also asked to be listed. Daily Reference Value is the reference amounts of nutrients consumed each day, and percent Daily Value represents the contribution of a nutrient in a serving to the maximum allowable amount of that nutrient in the total daily diet.(Silverglade, B. A., 1996)[3] The 2,000-calorie-a-day benchmark was born to help calculate %DV, which in turn helps people determine the nutrition facts needed to achieve a low-fat, low-carb, and high-fiber diet.(Petruccelli, P. J., 1996)[6]

The core elements that make up a healthy diet pattern include vegetables, legumes, fruits, non-fat or low-fat dairy products, protein-rich foods such as lean meats, eggs and seafood, and vegetable and animal oils.(Cho, S. and Kim, S., 2022)[7] And people of different ages, occupations, and physical states should adopt different dietary patterns. For example, infants and young children should actively supplement foods rich in iron and zinc within a reasonable range, and the elderly should choose appropriate foods to reduce the risk of cardiovascular disease, hypertension, type two diabetes, and even cancer.(Weschenfelder, C. et al., 2022)[8] At the same time, workers in their young and middle-aged years should also select foods that are beneficial to their health during their busy work.(Glympi, A. et al., 2020)[9]

McDonald’ s is a large multinational restaurant chain in the world. It was founded in Chicago, USA in 1955 and has about 40,000 stores in the world.(McDonald, M. and Oliver, G., 2019)[10] It mainly sells fast food such as hamburgers, French fries, fried chicken, soft drinks, ice products, salads, fruits and so on.(McDonald, M. and Oliver, G., 2019)[10] As the first and largest multinational fast-food chain, McDonald's has been the target of public debate on food-induced obesity, corporate ethics and consumer responsibility, and has been accused of affecting public health, such as high calories and lack of adequate balanced nutrition.(Tschoegl, A. E., 2007)[11] Actually, the McDonald's menu is dominated by fried foods with fewer vegetables and fruits, and the entire menu cannot be considered nutritionally balanced as a whole. However, as a convenient and clean restaurant that can be seen everywhere, McDonald's is loved and welcomed by teenagers and busy workers. Therefore, it is valuable to explore the nutritional composition of McDonald's menu, conduct cluster analytics, and give purchasing recommendations based on the clustering results.

Literature Review and Methodology

2.1 Principal Components Analysis

Principal Component Analysis (PCA) is a widely used multivariate statistical technique to visualize and explore high-dimensional data. PCA summarizes a large set of correlated features with a smaller number of representative features which mostly explain the variation of the original data. In other words, this process finds a low-dimensional representation of the data set that minimizes the information loss. Statistically, PCA is a process that tries to find lines, planes, and hyper-planes in the K-dimensional space that mostly approach the original data set in the least square sense. The new dimensions or features are called principal components(PCs) that are linear combinations of those in the original set, which ensure PCs have the largest variances and are uncorrelated with each other. Additionally, PCA also can be described by solving eigen-problems or, alternatively, obtained from the singular value decomposition (SVD) of the data matrix. In this paper, PCA is used for unsupervised machine learning because of involving a set of features and no associated responses (Abid, A. et al., 2018).[12] Actually, PCA is a flexible dimensional-reduction method without serious restrictions. It allows missing values, qualitative data, multicollinearity, and imprecise measurements. Since the main uses of PCA are descriptive and need no distributional assumptions, it can be defined as an adaptive exploratory method that can be used on numerical data of multiple types in diverse fields such as engineering, biology, and geography sciences (Jolliffe, I.T., Cadima, J., 2016).[13] Bigand, F. et al. (2021)[14] conducted PCA to decompose sign language motion into principal movements and figured out that upper-body motion can be explained by the dynamic combination of elementary movements in unconstrained continuous sign language discourses. Vinh, D.N. et al.(2021)[15] evaluated the relationship between age and seroprevalence by conducting PCA on 11 influenza antibody titers measured in 24,402 general population serum samples collected in Vietnam between 2009 to 2015. de, V. O. et al.(2021)[16] used PCA to reduce the number of dimensions of seven tested global tomography models aiming at easing the geological interpretation and making models comparison. PCA can be well combined with other analysis methods and provide a feasible solution for solving and visualizing high-dimensional problems. Its wide application in various disciplines also promotes itself to adapt to more complex problems. For example, due to the need for analysis of huge data sets in areas like bioinformatics and image processing and the sensitivity of PCA to the presence of outliers, there is an approach to define robust variants of PCA, called RPCA (Jolliffe, I.T., Cadima, J., 2016)[13]. RPCA attempts to decompose an $n \times p$ data matrix into a sum of two $n \times p$ components. The one is a low-rank component L and the other is a spares component S. This convex optimization problem was calculated by figuring out the matrix components of $X = L + S$ that minimize a linear function of two different norms of the components:

$$\min_{L,S} \; L_{*} + \lambda \; S_{1} \tag{2.1}$$

where $L_{*} = \sum_r \sigma_r(L)$, the sum of the singular values of L, is the nuclear norm of L, and $\lambda \; S_{1} = \sum_i \sum_j \mid S_{ij} \mid$ is the ℓ_1-norm of matrix S.

2.2 Clustering Methods

Clustering Methods are a collection of a broad set of techniques for finding subgroups, or clusters. In other words, clustering is a task of dividing points into distinct clusters so that data points belonging to a cluster are quite similar to each other, while data points in different clusters are as different as possible. It looks like classification, and they are both fundamental tasks in data mining. However, classification is always used as a supervised learning method and aims at creating a predictor. Clustering is a descriptive method usually for unsupervised learning. To find insightful subgroups among huge data set and draw inferences from them, it is necessary to define whether two objects are similar or dissimilar at first.

Distance measures are widely used to determine the similarity or dissimilarity between data points or one point and one cluster or clusters(Maimon, O. Rokach, L., 2005).[17] For different data types, several distance measurements are commonly known. For example, Han and Kamber(2001)[18] reported that the distance between the two data instances can be computed using Minkowski metric:

$$d(x_i, x_j) = (\mid x_{i1} - x_{j1} \mid^g + \mid x_{i2} - x_{j2} \mid^g + \cdots + \mid x_{ip} - x_{jp} \mid^g)^{1/g} \tag{2.2}$$

when there are two p-dimensional instances, $x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$ and $x_i = (x_{j1}, x_{j2}, \cdots, x_{jp})$. The widely used Euclidean distance is achieved when $g = 2$, and with $g = 1$ the sum of absolute paraxial distances (Manhattan metric) are obtained. Given $g = \infty$, the greatest of the paraxial distances (Chebychev metric) is calculated. For binary attributes, the distance between attributes can be computed based on a contingency table. However, when both of their states are equally valuable, the binary attribute is symmetric. At that time, using the simple matching coefficient can assess dissimilarity between two attributes:

$$d(x_i, x_j) = \frac{r + s}{q + r + s + t} \tag{2.3}$$

where r and s are the number of attributes that are unequal for both objects; q is the number of attributes that equal 1 for both objects; and t is the number of attributes that equal 0 for both objects. Moreover, for mixed-type attributes, the dissimilarity $d(x_{ij})$ between two instances, including p attributions of mixed type is defined as:

$$d(x_i, x_j) = \frac{\sum_{n=1}^{p} \delta_{ij}^{(n)} d_{ij}^{(n)}}{\sum_{n=1}^{p} \delta_{ij}^{(n)}} \tag{2.4}$$

where the indicator $\delta_{ij}^{(n)} = 0$ if one of the values is missing.

For the dissimilarity between one point and one cluster and two clusters, if one or both of the clusters contains multiple observations, there are four common measurements achieved by developing the notion of linkage. The following table shows the four most common types of linkage—complete, average, single, and centroid:

- Complete Linkage: Measure the farthest pair of points in two distinct clusters:

$$d_{complete}(G,H) = \max_{i \in G, j \in H} d_{ij} \tag{2.5}$$

- Single Linkage: Measure the closest pair of points in two distinct clusters:

$$d_{single}(G,H) = \min_{i \in G, j \in H} d_{ij} \tag{2.6}$$

- Average Linkage: Measure the average dissimilarity over all pairs:

$$d_{average}(G,H) = \frac{1}{n_G \cdot n_H} \sum_{i \in G, j \in H} d_{ij} \tag{2.7}$$

- Centroid Linkage: Measure the distance between the centroids of two dictinct clusters:

$$d_{centroid}(G,H) = \bar{X}_G - \bar{X}_{H\ 2} \tag{2.8}$$

Up to now, many clustering methods have been developed and different ways of distinguishing are proposed. Farley and Raftery (1998)[19] propound that clustering methods can be categorized into two main groups: hierarchical clustering and partitioning methods. Han and Kamber (2001)[18] suggest dividing clustering methods into additional three main categories: density-based methods, model-based method and grid- based methods. Estivill-Castro and Yang(2000)[20] also offer an alternative categorization based on the induction principle of the various clustering methods. In this report, K-means clustering, hierarchical clustering, Density-Based Spatial Clustering of Applications with Noise(DBSCAN) and and spectral clustering will be mainly introduced.

2.3 Hierarchical Clustering

Hierarchical clustering is provided as bottom-up or agglomerative clustering which starts from singleton sets of each point. In other words, each data point is its own cluster at first. At each following step, according to the chosen similarity measurement, the most similar cluster pairs are composed together, and this step is continued until all data points are involved in a single cluster or until some specific criteria are met(Akman, O. et al., 2019).[21] A potential advantage of hierarchical clustering is that it does not require pre-specify the number of clusters K. The number of clusters can be selected as needed after the whole process through the dendrogram which is an attractive tree-based representation of the result. The dendrogram is generally described as an upside-down tree, which is built to begin with creating leaves and then combining clusters into the trunk. The following figures are the dendrogram representation of the results of a hierarchical agglomerative clustering algorithm with different selected number of clusters K.

The left dendrogram is obtained by simulating 45 data points with hierarchical agglomerative clustering, which is observed without the cluster labels. In this diagram, each leaf represents

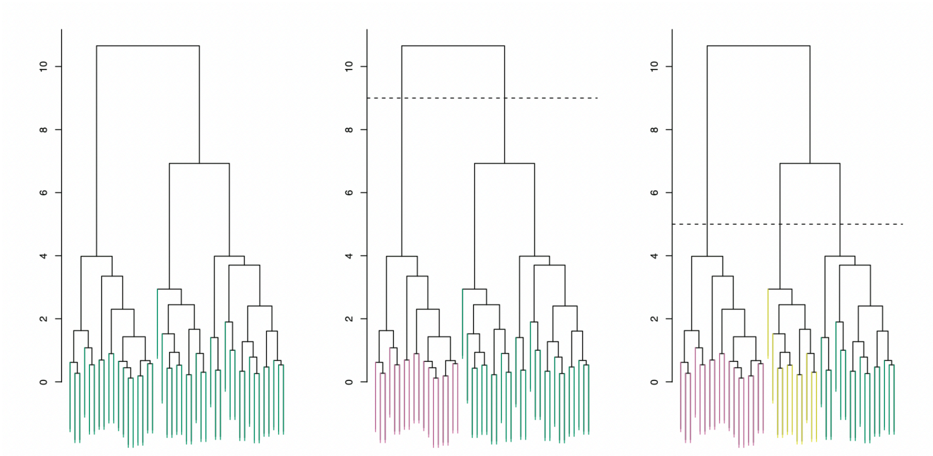


Figure 2.1: .

one of the 45 data points. With moving up the tree, some leaves are combined into branches, which refers to their similarity to each other. The earlier the combination occurs, that is, the lower the joining point is, the more similar the data points in the same clusters are. In the middle dendrogram, the dashed line serves as the stopping point of this process and produces two clusters printing different colors. The right dendrogram also comes from the left one and the dashed line cuts the results into three distinct clusters at a height of 5. The choice of dissimilarity measurement is essential, as it has a strong effect on computing the dendrogram. The four usual choices of dissimilarity measurement are mentioned before. However, for statistical questions, average, complete, and single linkage are more popular and centroid linkage is often used in genomics(James G., 2021).[22]

2.4 K-Means Clustering

K-means method is one of the most famous partitioning clustering methods. It partitions data points into K distinct, non-overlapping clusters based upon the distance metric. The value of K needs to be defined by a fixed number before the whole process, and then the K-means algorithm will assign each data points to exactly one of the clusters. Let $C_1, C_2, \cdots, C_k$ denote clusters containing data points, which satisfy two properties:

$$C_1 \cup C_2 \cup \cdots \cup C_K = 1, 2, \cdots, n \tag{2.9}$$

$$C_k \cap C_{k'} = 0 \quad \text{for all} \quad k \neq k^{'} \tag{2.10}$$

In other words, each data point involves at least one of the K clusters and the clusters are non-overlapping. The goal of K-means algorithm is that the within-clusters variation is as small as possible. The within-cluster variation of C_k is defined as $W(C_k)$. Consequently, it is tried to compute the minimum total variation.

$$\min_{C_1, \cdots . C_K} \{ \sum_{k=1}^{K} W(C_k) \} \tag{2.11}$$

For common choice, squared Euclidean distance is used

$$W(c_k) = \frac{1}{\mid C_k \mid} \sum_{i,i\prime \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \tag{2.12}$$

Where $\mid C_k \mid$ denotes the number of data points in the kth cluster. Combining above both equations gives the optimization problem.

$$\min_{C_1,\cdots .C_K} \{ \sum_{k=1}^{K} \frac{1}{\mid C_k \mid} \sum_{i,i\prime \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \} \tag{2.13}$$

The algorithm is presented by following steps:

Step1: Randomly select K points as initial clusters, and measure the distances between the initial clusters and all data points.

Step2: Assign the data points to the nearest clusters and calculate the centroid of each cluster.

Step3: Repeat the partition with clusters located by their centroids until the centroids of those clusters do not change after an iteration.

Obviously, K-means clustering method is a computationally expensive algorithm because it calculates the distance of every data point with the centroids of all the clusters at each iteration, which makes it difficult for implementing on huge data set. Additionally, there is a method to pick the best value for K. It is necessary to try different values for K and compute the total variance. It can start with $K = 1$ and let $K = K + 1$ in each round. It can be quantified its badness with the total variance when K is small. Each time a new cluster is added, the total variation within each cluster is smaller than before. And when there is only one point per cluster, the variation is equal to zero. Plotting the reduction in variance per value for K and finding the K corresponding to the largest reduction is the final step. This plot is called an ”elbow plot”, picking K is the same as seeking the ”elbow” in the plot.(James G., 2021)[22]

2.5 Density-Based Spatial Clustering of Applications with Noise(DBSCAN)

Partitioning clustering (like K-means) and hierarchical clustering work for finding spherical-shaped clusters, alternatively called convex clusters. In other words, they are only suitable for compact and well-separated clusters. Additionally, they are also seriously affected by the presence of noise in the data set. In 1996, corresponding to the requirement of discovery of clusters with arbitrary shape and good efficiency on large data sets in spatial databases, Ester et. al. (1996)[23] provided a new clustering algorithm, called density based spatial clustering of applications with noise (DBSCA), relying on a density-based notion of clusters. The key idea of DBSCAN algorithm is that for each data point of a cluster, the neighborhood of a given radius has to contain at least a minimum number of points. Compared with K-means method, DBSCAN does not specify the number of clusters as a parameter but infers the number of clusters based on the process. It has two fundamental parameters ϵ, the radius of the neighborhoods around the data point p, and minPts, the minimum number of data points in a neighborhood to define a cluster(Sander J., 2011).[24] Using these two parameters, the data points are divided into three classes.

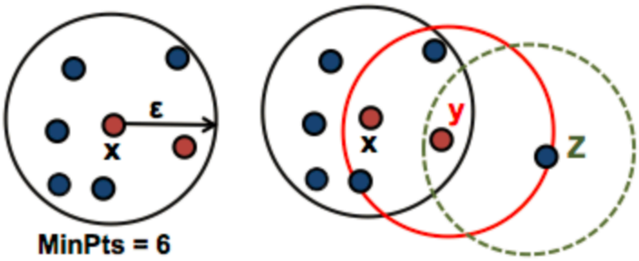


Figure 2.2

- Core point: A data point p is a core point if ϵ-neighborhood (Nbhd(p,ϵ)) of p contains at least minPts data points; $\mid Nbhd(p,\epsilon) \mid \geq minPts$
- Border point: a data point q is a border point if ϵ-neighborhood contains less than minPts data points, but q is reachable from at least one core point.
- Outlier(Noise): A data point o is an outlier if it is neither a core point nor a border point.

Based on above introduced basis, the algorithm can be abstracted by following steps:

Step1: Pick a data point randomly that has not been assigned to a cluster. Compute its neighborhood to determine whether it is a core point. If it is, start a cluster from this point. If it is not, label the point as an outlier.

Step2: Once a core point is found and becomes a cluster, expand it by adding all directly-reachable points to this cluster. Perform "neighborhood jumps" to find all density-reachable points and add them to the cluster. If an outlier is added, change the status of that point from outlier to border point.

Step3: Repeat these two steps until all points are either assigned to a cluster or designated as an outlier.

The DBSCAN algorithm finds clusters of points that are in close proximity based on a specified distance. The ordering points to identify the clustering structure algorithm (OPTICS) is also a density-based method, presented by Ankerst et al.(1999).[25] It orders the input data points according to the smallest distance to the next data point and constructs a reachability plot. The clusters are obtained based on the fewest points to be considered as a cluster, a search distance, and characteristics of the reachability plot, such as the slope and height of peaks. Moreover, Campello(2013)[26] extends DBSCAN by converting it into a hierarchical clustering, called hierarchical density-based spatial clustering of applications with noise (HDBSCAN), which tries to figure out clusters of points similar to DBSCAN but uses varying distances, allowing for clusters with varying densities based on cluster stability.

2.6 Spectral Clustering

Another widely used clustering method for a non-convex dataset is Spectral Clustering. This model is constructed by graph-based thoughts.(Higham, D. J., Kalna, G. and Kibble, M., 2007)[27] Specifically, it based on the undirected similarity graph. This graph can be expressed as $G = \{V, E\}$, where $V = \{x_i\}$ represents set of n vertices, indicating the data points and $E = \{w_{ij}\}$ means set of weighted edges indicating pair-wise similarity between points.

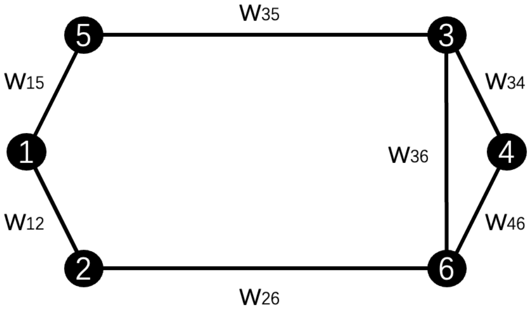


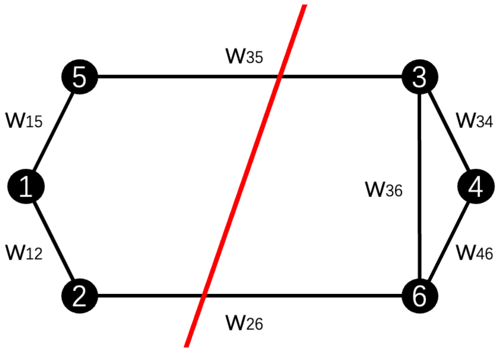
Figure 2.3: Example of Similarity Graph

From the assumption that there are N data points with P dimensions, V and E can be expressed as matrices below.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1P} \\ x_{21} & x_{22} & \cdots & x_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NP} \end{bmatrix}; \quad W = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1N} \\ w_{21} & w_{22} & \cdots & w_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ w_{N1} & w_{N2} & \cdots & w_{NN} \end{bmatrix}$$

If Gaussian Kernel is chosen, $\mathrm{w}_{ij} = \left\{ \begin{array}{ll} \exp^{\frac{||x_i - x_j||^2}{2\sigma}}; (i,j) \in E & (2.14) \\ 0; Otherwise & (2.15) \end{array} \right.$. Especially, the resulting similarities are symmetric in the sense of $w_{ij} = w_{ji}$

Turned back to the similarity graph, if these data points are divided into two categories A and B like the graph below.



If $A \subset V$, $B \subset V$ and $A \cap B = \emptyset$, the sum of the similarity between points in category A and each point in category B can be identified as

$$(A, B) = \sum_{i \in A, i \in B} w_{ij}$$

which is called ”cost” referring to the price have to pay for this split.

And then, If there are K categories, $V = \bigcup_{k=1}^{K} A_k$, $A_i \cap A_j = \emptyset$ and $\forall i,j \in \{1,2,\ldots,K\}$, the total "cost" can be repressed as

$$cut(V) = cut(A_1, A_2, \ldots, A_K) = \sum_{k=1}^{K} W(A_k, V - A_k)$$

Aiming at the points assigned to the same cluster should be highly similar and the points assigned to different clusters should be highly dissimilar, the main step is minimizing the similarities of between-categories connections.

But at first, normalization should be considered. Identifying $d_i = \sum_{j=1}^{N} w_{ij}$, called "degree", are used to normalize cuts. And then the objective becomes to

$$\min_{\{A_k\}_{k=1}^{K}} N_{cut}(V) = \sum_{k=1}^{K} \frac{W(A_k, V - A_k)}{\sum_{i \in A_k} d_i}$$

In order to find the minimization, based on Laplacian Matrix, there are efficient approximations using linear algebra. Firstly, indicator vector should be introduced, $y_i \in \{0,1\}^K$ and $\sum_{j=1}^{K} y_{ij} = 1$.

From that, $\{A_k\}_{k=1}^{K}$ can be written as Y= $\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1K} \\ y_{21} & y_{22} & \cdots & y_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ y_{N1} & y_{N2} & \cdots & y_{N_K} \end{bmatrix}$. In each column vector, there is only one element 1, and the other are all element 0. It can be figured out that what should be solved is

$$\hat{Y} = arg \min_{Y} N_{cut}(V) = \sum_{k=1}^{K} \frac{W(A_k, V - A_k)}{\sum_{i \in A_k} d_i}$$

Then $N_{cut}(V)$ can be expressed by linear algebra:

$$N_{cut}(V) = \sum_{k=1}^{K} \frac{W(A_k, V - A_k)}{\sum_{i \in A_k} d_i} = tr \begin{bmatrix} \frac{W(A_1, V - A_1)}{\sum_{i \in A_1} d_i} & 0 & \cdots & 0 \\ 0 & \frac{W(A_2, V - A_2)}{\sum_{i \in A_2} d_i} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{W(A_K, V - A_K)}{\sum_{i \in A_K} d_i} \end{bmatrix}$$

$$= tr \begin{bmatrix} W(A_1, V - A_1) & 0 & \cdots & 0 \\ 0 & W(A2_{,}V - A_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & W(A_K, V - A_K) \end{bmatrix} \cdot \begin{bmatrix} \sum_{i \in A_1} d_i & 0 & \cdots & 0 \\ 0 & \sum_{i \in A_2} d_i & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sum_{i \in A_K} d_i \end{bmatrix}^{-1}$$

$$= tr \quad O \quad \cdot \quad P^{-1}$$

W and Y can used to construct O' and P, and the degree matrix is necessary: D= $\begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_N \end{bmatrix}$

$$Y^T D Y = \sum_{i=1}^{N} y_i d_i y_i^T = P$$

$$Y^T D Y - Y^T W Y = O'$$

As the result, the final objective is to calculate the sum, which means for computing the trace.

$$tr \quad O' \quad \cdot \quad P^{-1} = tr \quad O \quad \cdot \quad P^{-1}$$

Thus,

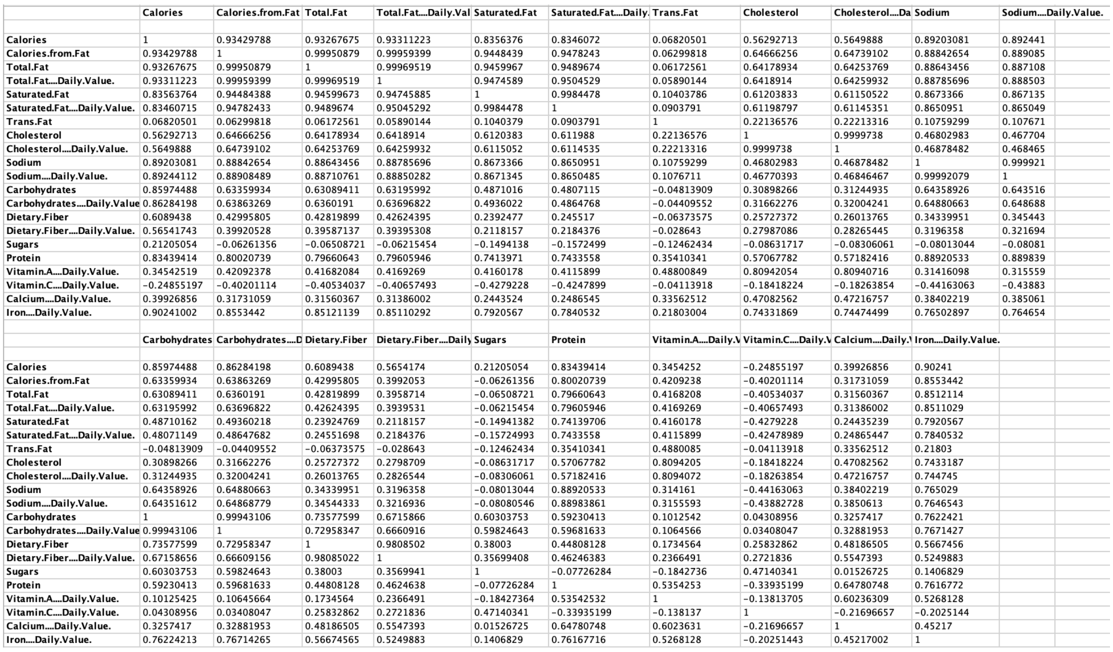
$$\hat{Y} = arg \min_{Y} tr\{Y^T(D-W)Y \cdot (Y^TDY)^{-1}\}$$

Furthermore, spectral clustering can also be interpreted as a random walk on a Markov chain, a task aimed at finding the real relaxed normalized cuts defined on the graph corresponding to the input data points.(Riaz, F. et al., 2013)[28] According to this interpretation direction, Pan, Y., Huang, C. and Wang, D. (2022)[29] proposed a multi-view spectral clustering method based on robust subspace segmentation to improve the existing multi-view clustering methods in technical There may be data corruption in all cases, resulting in a significant drop in clustering performance.In addition, spectral clustering generates spectral representations of raw data through matrix spectral analysis theory, so it is easy to implement and does not easily fall into local optima compared to other clustering methods. But cluster performance and affinity matrix learning can always be affected by redundant features and outliers, so Zhu, X. et al. (2020)[30] proposed a new spectral clustering-based clustering algorithm and half-quadratic optimization technique to solve these problems.

Results and Discussion

3.1 Data Description

This data set is constructed by the common Mcdonald' s breakfast menu and corresponding nutrition labels. These menu items and nutrition facts are scraped from the McDonald's official website. The data set contains the names of 42 common different kinds of items on the McDonald' s breakfast menu and their nutrition facts labels which are mandatory on the back of food packages by the FDA(Food and Drug Administration). The data frame with 42 observations and 21 columns, the items are set to the column name, and the 21 columns represent the nutrients contained in each serving, which are Calories, Calories from Fat, Total Fat, Total Fat (% Daily Value), Saturated Fat, Saturated Fat (% Daily Value), Trans Fat, Cholesterol, Cholesterol (% Daily Value), Sodium, Sodium (% Daily Value), Carbohydrates, Carbohydrates (% Daily Value), Dietary Fiber, Dietary Fiber (% Daily Value), Sugars, Protein, Vitamin A (% Daily Value), Vitamin C (% Daily Value), Calcium (% Daily Value), Iron (% Daily Value). All the models mentioned in the previous part will be implemented in this part. The whole process is divided into four steps, firstly read the data, and then use PCA to make dimension reduction of the data set. Next, four clustering techniques will be used to see how these breakfast items are grouped, finally, some explanations and conjectures based on the analysis results will be made. R is used as the programming language in the whole process and the package "ggplot2" is mainly used to visualize the results.



	Calories	Calories.from.Fat	Total.Fat	Total.Fat....Daily.Val	Saturated.Fat	Saturated.Fat....Daily.	Trans.Fat	Cholesterol	Cholesterol....Da	Sodium	Sodium....Daily.Value.
Calories	1	0.93429788	0.93267675	0.93311223	0.8356376	0.8346072	0.06820501	0.56292713	0.5649888	0.89203081	0.892441
Calories.from.Fat	0.93429788	1	0.99950879	0.99959399	0.9448439	0.9478243	0.06299818	0.64666256	0.64739102	0.88842654	0.889085
Total.Fat	0.93267675	0.99950879	1	0.99969519	0.9459967	0.9489674	0.06172561	0.64178934	0.64253769	0.88643456	0.887108
Total.Fat....Daily.Value.	0.93311223	0.99959399	0.99969519	1	0.9474589	0.9504529	0.05890144	0.6418914	0.64259932	0.88785696	0.888503
Saturated.Fat	0.83563764	0.94484388	0.94599673	0.94745885	1	0.9984478	0.10403786	0.61203833	0.61150522	0.8673366	0.867135
Saturated.Fat....Daily.Value.	0.83460715	0.94782433	0.9489674	0.95045292	0.9984478	1	0.0903791	0.61198797	0.61145351	0.8650951	0.865049
Trans.Fat	0.06820501	0.06299818	0.06172561	0.05890144	0.1040379	0.0903791	1	0.22136576	0.22213316	0.10759299	0.107671
Cholesterol	0.56292713	0.64666256	0.64178934	0.6418914	0.6120383	0.611988	0.22136576	1	0.9999738	0.46802983	0.467704
Cholesterol....Daily.Value.	0.5649888	0.64739102	0.64253769	0.64259932	0.6115052	0.6114535	0.22213316	0.9999738	1	0.46878482	0.468465
Sodium	0.89203081	0.88842654	0.88643456	0.88785696	0.8673366	0.8650951	0.10759299	0.46802983	0.46878482	1	0.999921
Sodium....Daily.Value.	0.89244112	0.88908489	0.88710761	0.88850282	0.8671345	0.8650485	0.1076711	0.46770393	0.46846467	0.99992079	1
Carbohydrates	0.85974488	0.63359934	0.63089411	0.63195992	0.4871016	0.4807115	−0.04813909	0.30898266	0.31244935	0.64358926	0.643516
Carbohydrates....Daily.Value	0.86284198	0.63863269	0.6360191	0.63696822	0.4936022	0.4864768	−0.04409552	0.31662276	0.32004241	0.64880663	0.648688
Dietary.Fiber	0.6089438	0.42995805	0.42819899	0.42624395	0.2392477	0.245517	−0.06373575	0.25727372	0.26013765	0.34339951	0.345443
Dietary.Fiber....Daily.Value.	0.56541743	0.39920528	0.39587137	0.39395308	0.2118157	0.2184376	−0.028643	0.27987086	0.28265445	0.3196358	0.321694
Sugars	0.21205054	−0.06261356	−0.06508721	−0.06215454	−0.1494138	−0.1572499	−0.12462434	−0.08631717	−0.08306061	−0.08013044	−0.08081
Protein	0.83439414	0.80020739	0.79660643	0.79605946	0.7413971	0.7433558	0.35410341	0.57067782	0.57182416	0.88920533	0.889839
Vitamin.A....Daily.Value.	0.34542519	0.42092378	0.41682084	0.4169269	0.4160178	0.4115899	0.48800849	0.80942054	0.80940716	0.31416098	0.315559
Vitamin.C....Daily.Value.	−0.24855197	−0.40201114	−0.40534037	−0.40657493	−0.4279228	−0.4247899	−0.04113918	−0.18418224	−0.18263854	−0.44163063	−0.43883
Calcium....Daily.Value.	0.39926856	0.31731059	0.31560367	0.31386002	0.2443524	0.2486545	0.33562512	0.47082562	0.47216757	0.38402219	0.385061
Iron....Daily.Value.	0.90241002	0.8553442	0.85121139	0.85110292	0.7920567	0.7840532	0.21803004	0.74331869	0.74474499	0.76502897	0.764654

	Carbohydrates	Carbohydrates....D	Dietary.Fiber	Dietary.Fiber....Daily	Sugars	Protein	Vitamin.A....Daily.V	Vitamin.C....Daily.V	Calcium....Daily.	Iron....Daily.Value.
Calories	0.85974488	0.86284198	0.6089438	0.5654174	0.21205054	0.83439414	0.3454252	−0.24855197	0.39926856	0.90241
Calories.from.Fat	0.63359934	0.63863269	0.42995805	0.3992053	−0.06261356	0.80020739	0.4209238	−0.40201114	0.31731059	0.8553442
Total.Fat	0.63089411	0.6360191	0.42819899	0.3958714	−0.06508721	0.79660643	0.4168208	−0.40534037	0.31560367	0.8512114
Total.Fat....Daily.Value.	0.63195992	0.63696822	0.42624395	0.3939531	−0.06215454	0.79605946	0.4169269	−0.40657493	0.31386002	0.8511029
Saturated.Fat	0.48710162	0.49360218	0.23924769	0.2118157	−0.14941382	0.74139706	0.4160178	−0.4279228	0.24435239	0.7920567
Saturated.Fat....Daily.Value.	0.48071149	0.48647682	0.24551698	0.2184376	−0.15724993	0.7433558	0.4115899	−0.42478989	0.24865447	0.7840532
Trans.Fat	−0.04813909	−0.04409552	−0.06373575	−0.028643	−0.12462434	0.35410341	0.4880085	−0.04113918	0.33562512	0.21803
Cholesterol	0.30898266	0.31662276	0.25727372	0.2798709	−0.08631717	0.57067782	0.8094205	−0.18418224	0.47082562	0.7433187
Cholesterol....Daily.Value.	0.31244935	0.32004241	0.26013765	0.2826544	−0.08306061	0.57182416	0.8094072	−0.18263854	0.47216757	0.744745
Sodium	0.64358926	0.64880663	0.34339951	0.3196358	−0.08013044	0.88920533	0.314161	−0.44163063	0.38402219	0.765029
Sodium....Daily.Value.	0.64351612	0.64868779	0.34544333	0.3216936	−0.08080546	0.88983861	0.3155593	−0.43882728	0.3850613	0.7646543
Carbohydrates	1	0.99943106	0.73577599	0.6715866	0.60303753	0.59230413	0.1012542	0.04308956	0.3257417	0.7622421
Carbohydrates....Daily.Value	0.99943106	1	0.72958347	0.6660916	0.59824643	0.59681633	0.1064566	0.03408047	0.32881953	0.7671427
Dietary.Fiber	0.73577599	0.72958347	1	0.9808502	0.38003	0.44808128	0.1734564	0.25832862	0.48186505	0.5667456
Dietary.Fiber....Daily.Value.	0.67158656	0.66609156	0.98085022	1	0.35699408	0.46246383	0.2366491	0.2721836	0.5547393	0.5249883
Sugars	0.60303753	0.59824643	0.38003	0.3569941	1	−0.07726284	−0.1842736	0.47140341	0.01526725	0.1406829
Protein	0.59230413	0.59681633	0.44808128	0.4624638	−0.07726284	1	0.5354253	−0.33935199	0.64780748	0.7616772
Vitamin.A....Daily.Value.	0.10125425	0.10645664	0.1734564	0.2366491	−0.18427364	0.53542532	1	−0.13813705	0.60236309	0.5268128
Vitamin.C....Daily.Value.	0.04308956	0.03408047	0.25832862	0.2721836	0.47140341	−0.33935199	−0.138137	1	−0.21696657	−0.2025144
Calcium....Daily.Value.	0.3257417	0.32881953	0.48186505	0.5547393	0.01526725	0.64780748	0.6023631	−0.21696657	1	0.45217
Iron....Daily.Value.	0.76224213	0.76714265	0.56674565	0.5249883	0.1406829	0.76167716	0.5268128	−0.20251443	0.45217002	1

Figure 3.1: Correlation Coefficient Table

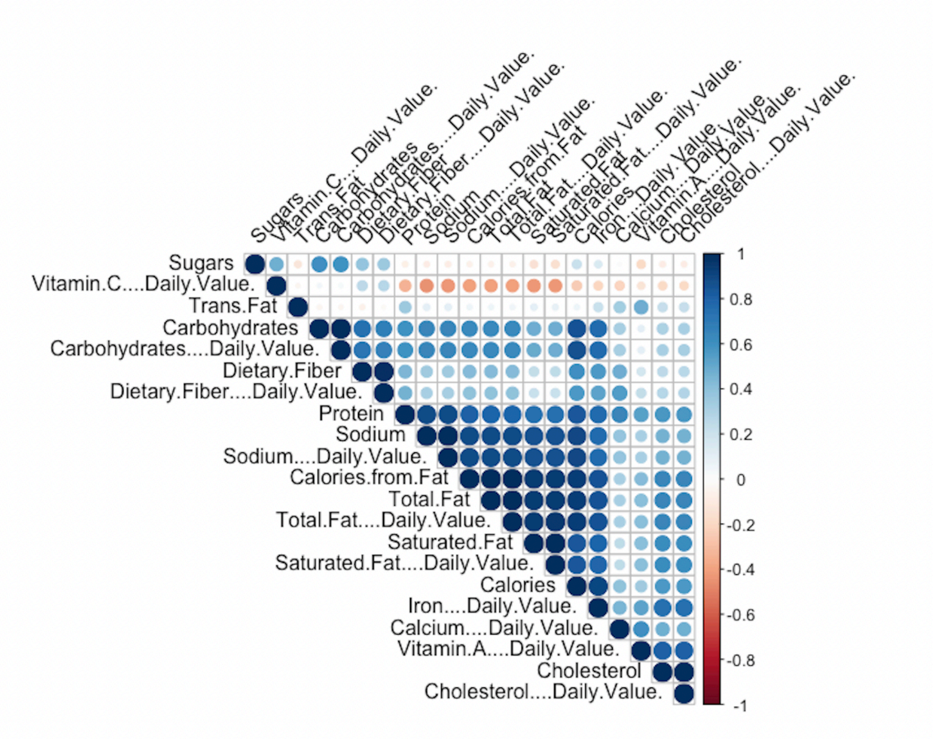


Figure 3.2: Correlation Coefficient Plot

It can be seen some details about features from the correlation coefficient table and correlation coefficient plot. Carbohydrates, Carbohydrates (% Daily Value), Protein, Calcium (% Daily Value), Sodium, Sodium (% Daily Value), Calories from Fat, Total Fat, Total Fat (% Daily Value), Saturated Fat, Saturated Fat (% Daily Value), Calories and Iron (% Daily Value) are strongly positively correlated with most variables, and the highest correlation coefficient can reach up to 0.99. However, Vitamin C (% Daily Value) is strongly negatively correlated with most features, such as Sodium, Sodium (% Daily Value), Calories from Fat, Total Fat, Total Fat (% Daily Value), Saturated Fat, Saturated Fat (% Daily Value) and Protein, and the lowest correlation coefficient can reach to -0.44. Sugars are weakly negatively correlated with Protein, Sodium, Sodium (% Daily Value), Calories from Fat, Total Fat, Total Fat (% Daily Value), Saturated Fat, Saturated Fat (% Daily Value), Cholesterol, Cholesterol (% Daily Value), Vitamin A (% Daily Value) and Trans Fat, while it is weakly positively correlated with the other variables. And Trans Fat is not drawn many correlations with the other variables.

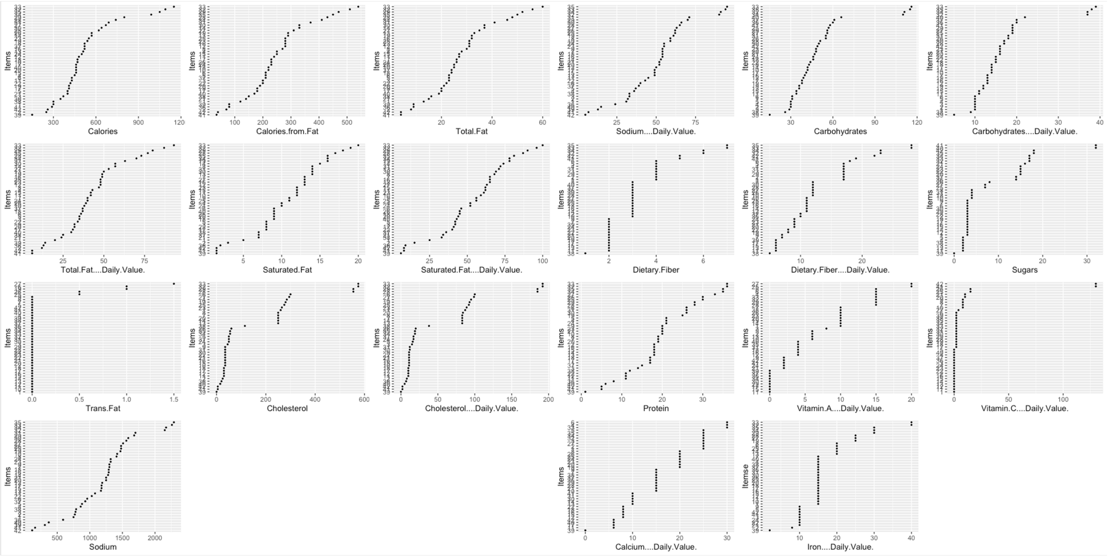


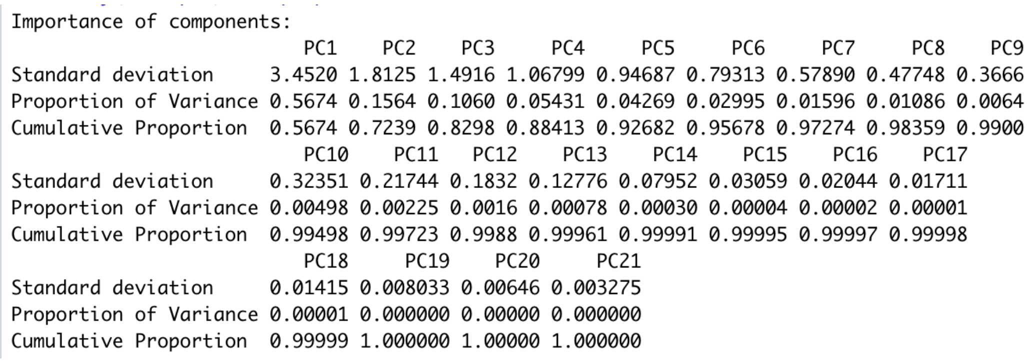
Figure 3.3: Rank Plots

Rank plot helps display that some items are outstanding in some features. For example, Big Breakfast with Hotcakes (Large Biscuit), Big Breakfast with Hotcakes (Regular Biscuit), Big Breakfast with Hotcakes and Egg Whites (Large Biscuit) and Big Breakfast (Large Biscuit) are on the top in Calories, Calories from Fat, Total Fat, Total Fat (% Daily Value), Saturated Fat, Saturated Fat (% Daily Value), Cholesterol, Cholesterol (% Daily Value), Protein and Iron (% Daily Value). Moreover, Big Breakfast with Hotcakes and Egg Whites (Large Biscuit), Big Breakfast with Hotcakes (Large Biscuit) and Big Breakfast with Hotcakes and Egg Whites (Regular Biscuit) are on the top of Sodium, Sodium (% Daily Value), Carbohydrates, Carbohydrates (% Daily Value), Dietary Fiber and Dietary Fiber (% Daily Value). However, Hash Brown, Fruit Maple Oatmeal, Fruit Maple Oatmeal without Brown Sugar and Egg White Delight rank in the bottom four of Calories, Calories from Fat, Total Fat, Total Fat (% Daily Value), Saturated Fat, Saturated Fat (% Daily Value), Cholesterol, Cholesterol (% Daily Value), Sodium, Sodium (% Daily Value) and Protein. In addition, Steak, Egg Cheese Bagel, Bacon, Egg Cheese Bagel, Fruit Maple Oatmeal without Brown Sugar and Fruit Maple Oatmeal are rich in Vitamin C (% Daily Value) and Sugars. And Egg McMuffin and Sausage Biscuit (Regular Biscuit) contain a small amount of Trans Fat, Carbohydrates (% Daily Value) and Sugars. At the same time, Hash Brown contains the lowest amount of Cholesterol, Cholesterol (% Daily Value), Carbohydrates, Carbohydrates (% Daily Value), Dietary Fiber, Dietary Fiber (% Daily Value), Sugars and Protein.

3.2 Data Processing and Results

3.2.1 Principal Components Analysis

Before implementing analytics of clustering models, PCA is performed at first to find out several principlale components to represent most of the features. "prcomp" function is chosen to make dimensional reduction, which is based on single value decomposition to obtain accurate results. This data set includes 21 nutrition facts, so the linear combination of 21 variables will be found to explain most of the variation in the data set. The following table displays the results.



```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7     PC8    PC9
Standard deviation     3.4520 1.8125 1.4916 1.06799 0.94687 0.79313 0.57890 0.47748 0.3666
Proportion of Variance 0.5674 0.1564 0.1060 0.05431 0.04269 0.02995 0.01596 0.01086 0.0064
Cumulative Proportion  0.5674 0.7239 0.8298 0.88413 0.92682 0.95678 0.97274 0.98359 0.9900
                          PC10    PC11   PC12    PC13    PC14    PC15    PC16    PC17
Standard deviation     0.32351 0.21744 0.1832 0.12776 0.07952 0.03059 0.02044 0.01711
Proportion of Variance 0.00498 0.00225 0.0016 0.00078 0.00030 0.00004 0.00002 0.00001
Cumulative Proportion  0.99498 0.99723 0.9988 0.99961 0.99991 0.99995 0.99997 0.99998
                          PC18     PC19    PC20     PC21
Standard deviation     0.01415 0.008033 0.00646 0.003275
Proportion of Variance 0.00001 0.000000 0.00000 0.000000
Cumulative Proportion  0.99999 1.000000 1.00000 1.000000
```

Figure 3.4: Importance of Components

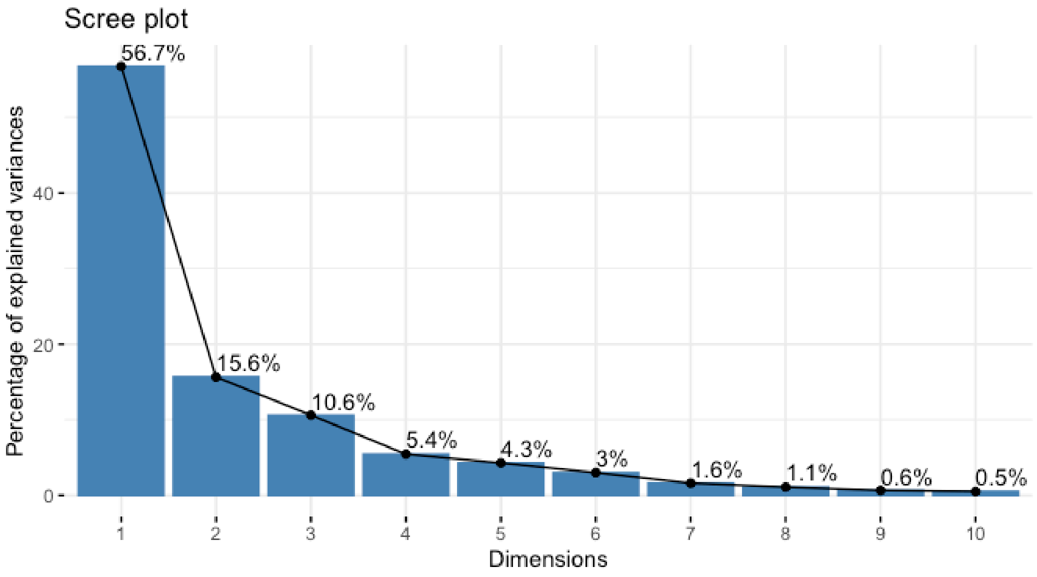
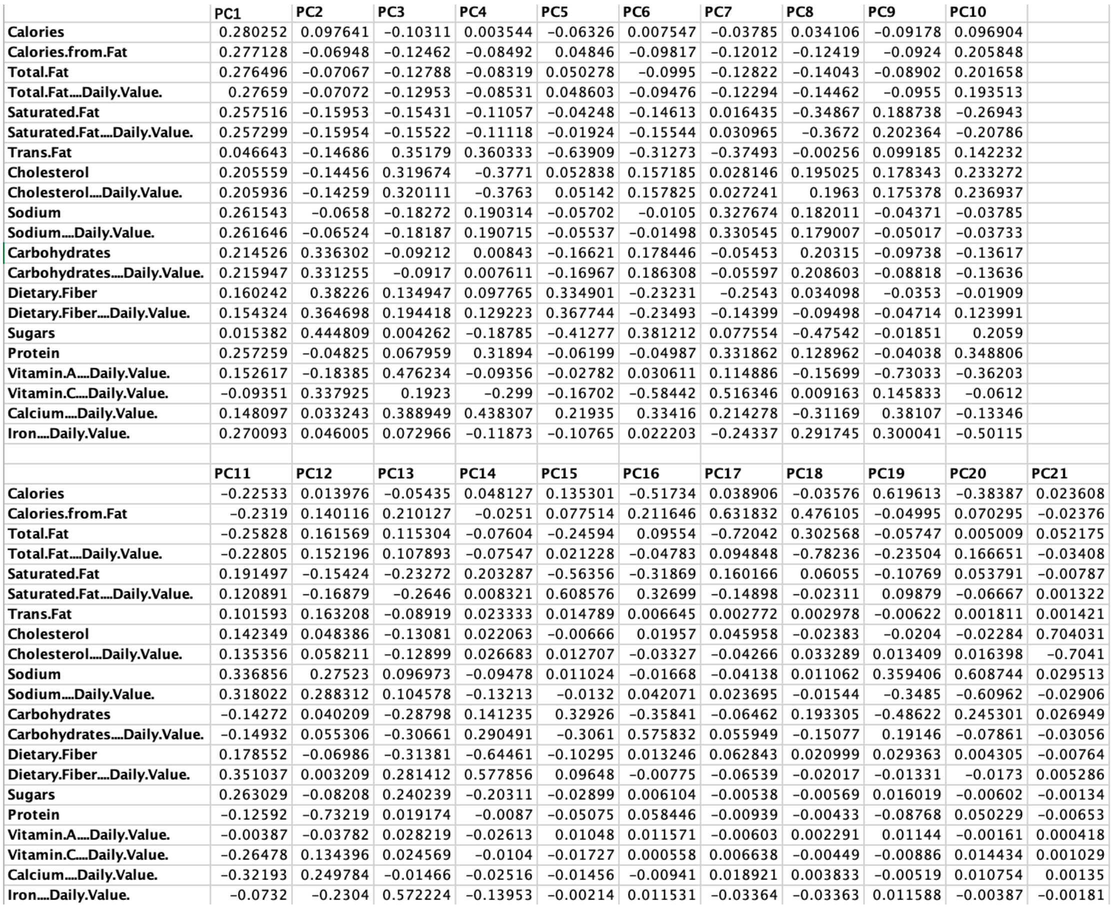


Figure 3.5: Scree Plot

As illustrated in the above table and scree plot, the top four principal components explain approximately 88% of the total variability, and each additional principal component brings more than 4.76%(1/21) increase in explanatory of overall variability, which means the top four principal components can explain most of the features and the information carried by these four principal components is not less than features before PCA transmission.



	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Calories	0.280252	0.097641	−0.10311	0.003544	−0.06326	0.007547	−0.03785	0.034106	−0.09178	0.096904
Calories.from.Fat	0.277128	−0.06948	−0.12462	−0.08492	0.04846	−0.09817	−0.12012	−0.12419	−0.0924	0.205848
Total.Fat	0.276496	−0.07067	−0.12788	−0.08319	0.050278	−0.0995	−0.12822	−0.14043	−0.08902	0.201658
Total.Fat....Daily.Value.	0.27659	−0.07072	−0.12953	−0.08531	0.048603	−0.09476	−0.12294	−0.14462	−0.0955	0.193513
Saturated.Fat	0.257516	−0.15953	−0.15431	−0.11057	−0.04248	−0.14613	0.016435	−0.34867	0.188738	−0.26943
Saturated.Fat....Daily.Value.	0.257299	−0.15954	−0.15522	−0.11118	−0.01924	−0.15544	0.030965	−0.3672	0.202364	−0.20786
Trans.Fat	0.046643	−0.14686	0.35179	0.360333	−0.63909	−0.31273	−0.37493	−0.00256	0.099185	0.142232
Cholesterol	0.205559	−0.14456	0.319674	−0.3771	0.052838	0.157185	0.028146	0.195025	0.178343	0.233272
Cholesterol....Daily.Value.	0.205936	−0.14259	0.320111	−0.3763	0.05142	0.157825	0.027241	0.1963	0.175378	0.236937
Sodium	0.261543	−0.0658	−0.18272	0.190314	−0.05702	−0.0105	0.327674	0.182011	−0.04371	−0.03785
Sodium....Daily.Value.	0.261646	−0.06524	−0.18187	0.190715	−0.05537	−0.01498	0.330545	0.179007	−0.05017	−0.03733
Carbohydrates	0.214526	0.336302	−0.09212	0.00843	−0.16621	0.178446	−0.05453	0.20315	−0.09738	−0.13617
Carbohydrates....Daily.Value.	0.215947	0.331255	−0.0917	0.007611	−0.16967	0.186308	−0.05597	0.208603	−0.08818	−0.13636
Dietary.Fiber	0.160242	0.38226	0.134947	0.097765	0.334901	−0.23231	−0.2543	0.034098	−0.0353	−0.01909
Dietary.Fiber....Daily.Value.	0.154324	0.364698	0.194418	0.129223	0.367744	−0.23493	−0.14399	−0.09498	−0.04714	0.123991
Sugars	0.015382	0.444809	0.004262	−0.18785	−0.41277	0.381212	0.077554	−0.47542	−0.01851	0.2059
Protein	0.257259	−0.04825	0.067959	0.31894	−0.06199	−0.04987	0.331862	0.128962	−0.04038	0.348806
Vitamin.A....Daily.Value.	0.152617	−0.18385	0.476234	−0.09356	−0.02782	0.030611	0.114886	−0.15699	−0.73033	−0.36203
Vitamin.C....Daily.Value.	−0.09351	0.337925	0.1923	−0.299	−0.16702	−0.58442	0.516346	0.009163	0.145833	−0.0612
Calcium....Daily.Value.	0.148097	0.033243	0.388949	0.438307	0.21935	0.33416	0.214278	−0.31169	0.38107	−0.13346
Iron....Daily.Value.	0.270093	0.046005	0.072966	−0.11873	−0.10765	0.022203	−0.24337	0.291745	0.300041	−0.50115

	PC11	PC12	PC13	PC14	PC15	PC16	PC17	PC18	PC19	PC20	PC21
Calories	−0.22533	0.013976	−0.05435	0.048127	0.135301	−0.51734	0.038906	−0.03576	0.619613	−0.38387	0.023608
Calories.from.Fat	−0.2319	0.140116	0.210127	−0.0251	0.077514	0.211646	0.631832	0.476105	−0.04995	0.070295	−0.02376
Total.Fat	−0.25828	0.161569	0.115304	−0.07604	−0.24594	0.09554	−0.72042	0.302568	−0.05747	0.005009	0.052175
Total.Fat....Daily.Value.	−0.22805	0.152196	0.107893	−0.07547	0.021228	−0.04783	0.094848	−0.78236	−0.23504	0.166651	−0.03408
Saturated.Fat	0.191497	−0.15424	−0.23272	0.203287	−0.56356	−0.31869	0.160166	0.06055	−0.10769	0.053791	−0.00787
Saturated.Fat....Daily.Value.	0.120891	−0.16879	−0.2646	0.008321	0.608576	0.32699	−0.14898	−0.02311	0.09879	−0.06667	0.001322
Trans.Fat	0.101593	0.163208	−0.08919	0.023333	0.014789	0.006645	0.002772	0.002978	−0.00622	0.001811	0.001421
Cholesterol	0.142349	0.048386	−0.13081	0.022063	−0.00666	0.01957	0.045958	−0.02383	−0.0204	−0.02284	0.704031
Cholesterol....Daily.Value.	0.135356	0.058211	−0.12899	0.026683	0.012707	−0.03327	−0.04266	0.033289	0.013409	0.016398	−0.7041
Sodium	0.336856	0.27523	0.096973	−0.09478	0.011024	−0.01668	−0.04138	0.011062	0.359406	0.608744	0.029513
Sodium....Daily.Value.	0.318022	0.288312	0.104578	−0.13213	−0.0132	0.042071	0.023695	−0.01544	−0.3485	−0.60962	−0.02906
Carbohydrates	−0.14272	0.040209	−0.28798	0.141235	0.32926	−0.35841	−0.06462	0.193305	−0.48622	0.245301	0.026949
Carbohydrates....Daily.Value.	−0.14932	0.055306	−0.30661	0.290491	−0.3061	0.575832	0.055949	−0.15077	0.19146	−0.07861	−0.03056
Dietary.Fiber	0.178552	−0.06986	−0.31381	−0.64461	−0.10295	0.013246	0.062843	0.020999	0.029363	0.004305	−0.00764
Dietary.Fiber....Daily.Value.	0.351037	0.003209	0.281412	0.577856	0.09648	−0.00775	−0.06539	−0.02017	−0.01331	−0.0173	0.005286
Sugars	0.263029	−0.08208	0.240239	−0.20311	−0.02899	0.006104	−0.00538	−0.00569	0.016019	−0.00602	−0.00134
Protein	−0.12592	−0.73219	0.019174	−0.0087	−0.05075	0.058446	−0.00939	−0.00433	−0.08768	0.050229	−0.00653
Vitamin.A....Daily.Value.	−0.00387	−0.03782	0.028219	−0.02613	0.01048	0.011571	−0.00603	0.002291	0.01144	−0.00161	0.000418
Vitamin.C....Daily.Value.	−0.26478	0.134396	0.024569	−0.0104	−0.01727	0.000558	0.006638	−0.00449	−0.00886	0.014434	0.001029
Calcium....Daily.Value.	−0.32193	0.249784	−0.01466	−0.02516	−0.01456	−0.00941	0.018921	0.003833	−0.00519	0.010754	0.00135
Iron....Daily.Value.	−0.0732	−0.2304	0.572224	−0.13953	−0.00214	0.011531	−0.03364	−0.03363	0.011588	−0.00387	−0.00181

Figure 3.6: Coefficient Table

According to the coefficient of all principal components listed above, it can be seen that Calories, Calories from Fat, Total Fat (% Daily Value), Total Fat, Iron (% Daily Value), Sodium (% Daily Value), Sodium, Saturated Fat, Saturated Fat (% Daily Value) and Protein are selected to construct the PC1. Thus, the high value of PC1 represents a breakfast with high total number of calories, or called energy; a lot of different kinds of fat which can mostly be found in foods from both plants and animals; a large amount of salt which is commonly found in packaged and prepared foods. PC2 mainly consists of Sugars, Dietary Fiber, Dietary Fiber (% Daily Value), Vitamin C (% Daily Value), Carbohydrates, and Carbohydrates (% Daily Value). Diets with high PC2 mean getting more carbohydrates which primarily come from plant food. PC3 mainly constitutes Vitamin A (% Daily Value), Calcium (% Daily Value), Trans Fat, Cholesterol (% Daily Value) and Cholesterol. The high value of PC3 represents that a breakfast item is rich in vitamins and minerals and a part of harmful fat and cholesterol which is only in animal products. The main content of PC4 is including Calcium (% Daily Value), Trans Fat, and Protein. If a breakfast scores high in PC4 that means a high involvement of common vitamins, minerals and amino acids which are normally required for body functioning. Moreover, the contribution of each feature in each principal component is drawn above.

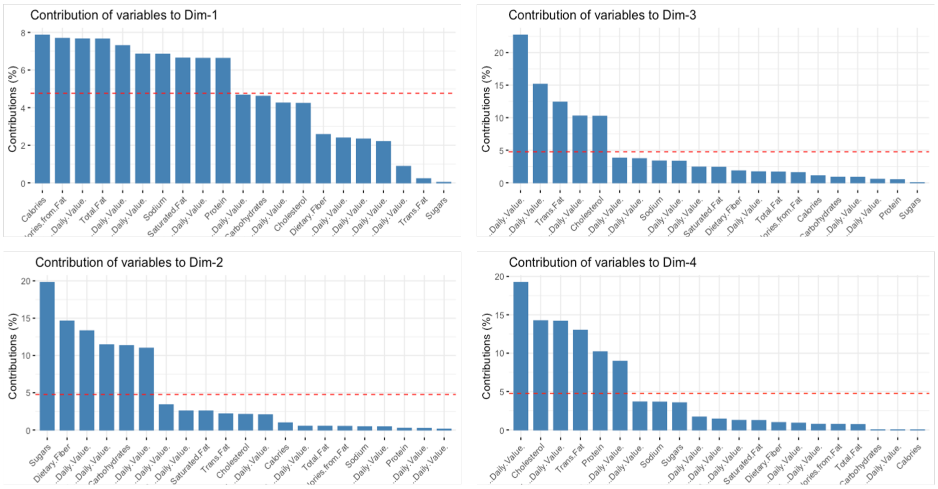


Figure 3.7: Contribution Plot

The following graph displays how much each feature contributes to each principal component. The length of the vectors shows the strength of their influence on each principal component, and the orientation of the vectors shows the direction of the influence. In this example, except for Vitamin C (% Daily Value), the other features all have a positive influence on PC1. Especially, Calories, Iron (% Daily Value) and Protein strongly positively impact PC1. Additionally, Vitamin C (% Daily Value), Sugars, Carbohydrates, Carbohydrates (% Daily Value), Dietary Fiber and Dietary Fiber (% Daily Value) have strong positive effects on PC2. From the right plot, it can be seen that Vitamin A (% Daily Value), Calcium (% Daily Value), Trans Fat, Cholesterol and Cholesterol (% Daily Value) strongly positively influence PC3. Moreover, Protein, Sodium, Sodium (% Daily Value), Dietary Fiber and Dietary Fiber (% Daily Value) have a great positive influence on PC4, but the influence of Protein is greater than any one of the others. In addition, the observation of the angles between two vectors is also worth exploring. If the two vectors form a small angle, which represents that there is a positive correlation between these two features. For example, Carbohydrates, Carbohydrates (% Daily Value) are strongly positively correlated with each other.

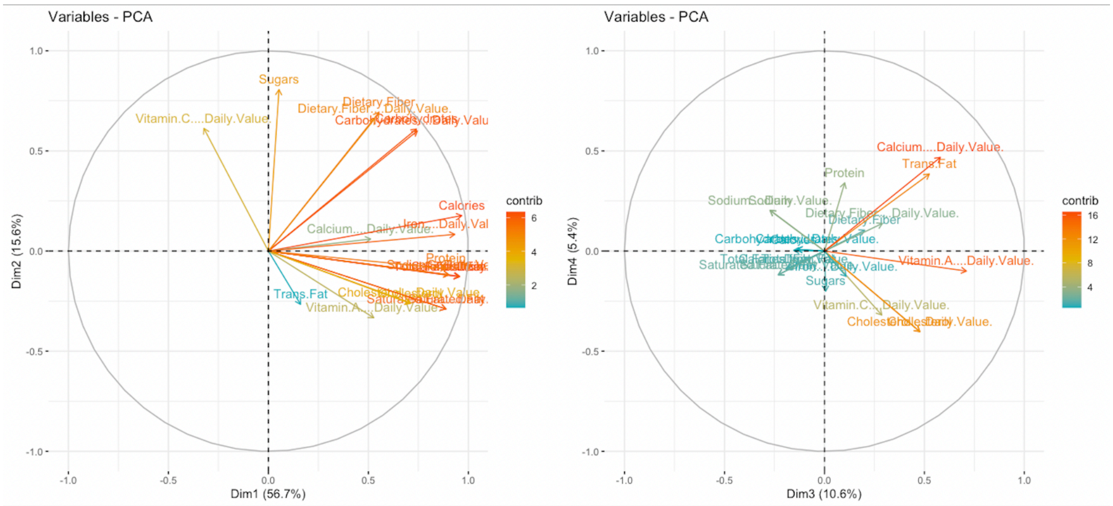
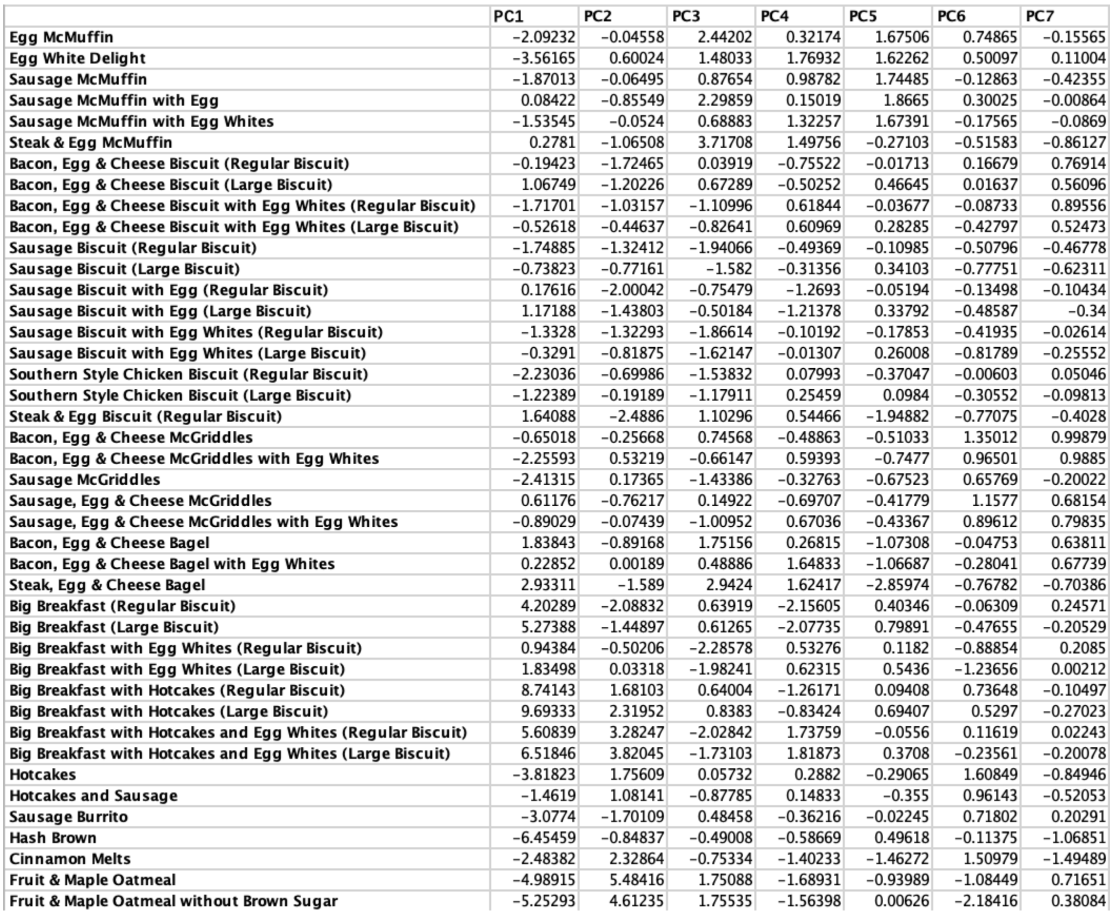


Figure 3.8: Loading Plot

According to the following graphs, it can be seen the table and the biplot of score of each breakfast item on the first four principal components. There are some inferences that can be made. Fruit Maple Oatmeal and Fruit Maple Oatmeal without Brown Sugar are relatively low in PC1 and relatively high in PC2, so they appear in the top left corner of the biplot of PC1 and PC2. In addition, they are also relatively high in PC3 and relatively low in PC4, so they appear at the bottom right corner of the biplot of PC3 and PC4. Steak Egg McMuffin is relatively high in PC3 and PC4, so it is shown at the top right corner of the biplot with PC3 and PC4. Moreover, the corresponding values of Sausage, Egg Cheese McGriddles with Egg Whites, Bacon, Egg Cheese Biscuit with Egg Whites (Large Biscuit), Bacon, Egg Cheese Biscuit with Egg Whites (Regular Biscuit) and Bacon, Egg Cheese McGriddles with Egg Whites in PC1 and PC2 are relatively close, so it can be inferred that these four items are more likely grouped in the same cluster. At the same time, the score of Egg McMuffin, Sausage McMuffin, and Sausage McMuffin with Egg Whites are relatively close in both PC1 and PC2, so these three items are more likely formed in the same cluster.



	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Egg McMuffin	−2.09232	−0.04558	2.44202	0.32174	1.67506	0.74865	−0.15565
Egg White Delight	−3.56165	0.60024	1.48033	1.76932	1.62262	0.50097	0.11004
Sausage McMuffin	−1.87013	−0.06495	0.87654	0.98782	1.74485	−0.12863	−0.42355
Sausage McMuffin with Egg	0.08422	−0.85549	2.29859	0.15019	1.8665	0.30025	−0.00864
Sausage McMuffin with Egg Whites	−1.53545	−0.0524	0.68883	1.32257	1.67391	−0.17565	−0.0869
Steak & Egg McMuffin	0.2781	−1.06508	3.71708	1.49756	−0.27103	−0.51583	−0.86127
Bacon, Egg & Cheese Biscuit (Regular Biscuit)	−0.19423	−1.72465	0.03919	−0.75522	−0.01713	0.16679	0.76914
Bacon, Egg & Cheese Biscuit (Large Biscuit)	1.06749	−1.20226	0.67289	−0.50252	0.46645	0.01637	0.56096
Bacon, Egg & Cheese Biscuit with Egg Whites (Regular Biscuit)	−1.71701	−1.03157	−1.10996	0.61844	−0.03677	−0.08733	0.89556
Bacon, Egg & Cheese Biscuit with Egg Whites (Large Biscuit)	−0.52618	−0.44637	−0.82641	0.60969	0.28285	−0.42797	0.52473
Sausage Biscuit (Regular Biscuit)	−1.74885	−1.32412	−1.94066	−0.49369	−0.10985	−0.50796	−0.46778
Sausage Biscuit (Large Biscuit)	−0.73823	−0.77161	−1.582	−0.31356	0.34103	−0.77751	−0.62311
Sausage Biscuit with Egg (Regular Biscuit)	0.17616	−2.00042	−0.75479	−1.2693	−0.05194	−0.13498	−0.10434
Sausage Biscuit with Egg (Large Biscuit)	1.17188	−1.43803	−0.50184	−1.21378	0.33792	−0.48587	−0.34
Sausage Biscuit with Egg Whites (Regular Biscuit)	−1.3328	−1.32293	−1.86614	−0.10192	−0.17853	−0.41935	−0.02614
Sausage Biscuit with Egg Whites (Large Biscuit)	−0.3291	−0.81875	−1.62147	−0.01307	0.26008	−0.81789	−0.25552
Southern Style Chicken Biscuit (Regular Biscuit)	−2.23036	−0.69986	−1.53832	0.07993	−0.37047	−0.00603	0.05046
Southern Style Chicken Biscuit (Large Biscuit)	−1.22389	−0.19189	−1.17911	0.25459	0.0984	−0.30552	−0.09813
Steak & Egg Biscuit (Regular Biscuit)	1.64088	−2.4886	1.10296	0.54466	−1.94882	−0.77075	−0.4028
Bacon, Egg & Cheese McGriddles	−0.65018	−0.25668	0.74568	−0.48863	−0.51033	1.35012	0.99879
Bacon, Egg & Cheese McGriddles with Egg Whites	−2.25593	0.53219	−0.66147	0.59393	−0.7477	0.96501	0.9885
Sausage McGriddles	−2.41315	0.17365	−1.43386	−0.32763	−0.67523	0.65769	−0.20022
Sausage, Egg & Cheese McGriddles	0.61176	−0.76217	0.14922	−0.69707	−0.41779	1.1577	0.68154
Sausage, Egg & Cheese McGriddles with Egg Whites	−0.89029	−0.07439	−1.00952	0.67036	−0.43367	0.89612	0.79835
Bacon, Egg & Cheese Bagel	1.83843	−0.89168	1.75156	0.26815	−1.07308	−0.04753	0.63811
Bacon, Egg & Cheese Bagel with Egg Whites	0.22852	0.00189	0.48886	1.64833	−1.06687	−0.28041	0.67739
Steak, Egg & Cheese Bagel	2.93311	−1.589	2.9424	1.62417	−2.85974	−0.76782	−0.70386
Big Breakfast (Regular Biscuit)	4.20289	−2.08832	0.63919	−2.15605	0.40346	−0.06309	0.24571
Big Breakfast (Large Biscuit)	5.27388	−1.44897	0.61265	−2.07735	0.79891	−0.47655	−0.20529
Big Breakfast with Egg Whites (Regular Biscuit)	0.94384	−0.50206	−2.28578	0.53276	0.1182	−0.88854	0.2085
Big Breakfast with Egg Whites (Large Biscuit)	1.83498	0.03318	−1.98241	0.62315	0.5436	−1.23656	0.00212
Big Breakfast with Hotcakes (Regular Biscuit)	8.74143	1.68103	0.64004	−1.26171	0.09408	0.73648	−0.10497
Big Breakfast with Hotcakes (Large Biscuit)	9.69333	2.31952	0.8383	−0.83424	0.69407	0.5297	−0.27023
Big Breakfast with Hotcakes and Egg Whites (Regular Biscuit)	5.60839	3.28247	−2.02842	1.73759	−0.0556	0.11619	0.02243
Big Breakfast with Hotcakes and Egg Whites (Large Biscuit)	6.51846	3.82045	−1.73103	1.81873	0.3708	−0.23561	−0.20078
Hotcakes	−3.81823	1.75609	0.05732	0.2882	−0.29065	1.60849	−0.84946
Hotcakes and Sausage	−1.4619	1.08141	−0.87785	0.14833	−0.355	0.96143	−0.52053
Sausage Burrito	−3.0774	−1.70109	0.48458	−0.36216	−0.02245	0.71802	0.20291
Hash Brown	−6.45459	−0.84837	−0.49008	−0.58669	0.49618	−0.11375	−1.06851
Cinnamon Melts	−2.48382	2.32864	−0.75334	−1.40233	−1.46272	1.50979	−1.49489
Fruit & Maple Oatmeal	−4.98915	5.48416	1.75088	−1.68931	−0.93989	−1.08449	0.71651
Fruit & Maple Oatmeal without Brown Sugar	−5.25293	4.61235	1.75535	−1.56398	0.00626	−2.18416	0.38084

Figure 3.9: Score Table

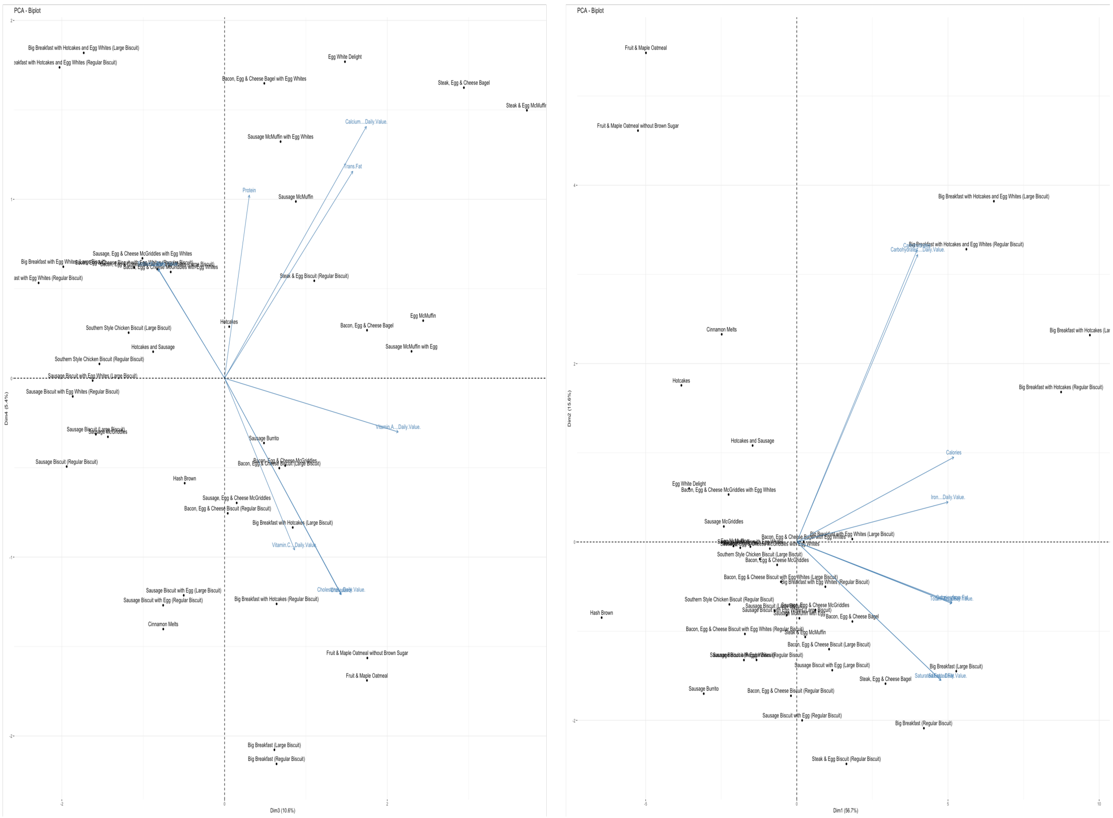


Figure 3.10: Biplot

3.2.2 Hierarchical Clustering

In this part, the key process and the results of Hierarchical Clustering method will be mainly talked about. As mentioned in the methodology part, the choice of distance linkage which means the way of calculating distance between one cluster and one point or another cluster is significant for clustering results. According to this data set, four linkages are used to make classification and implemented by the function of "hclust" with different methods. From the single linkage graph, it can be seen that this linkage measurement is unreasonable in calculating the distance between two clusters, and it is meaningless to deal with practical problems. The result calculated with "complete" method and "ward. D" is similar to each other, and next the results with "complete" measurement will be mainly discussed.

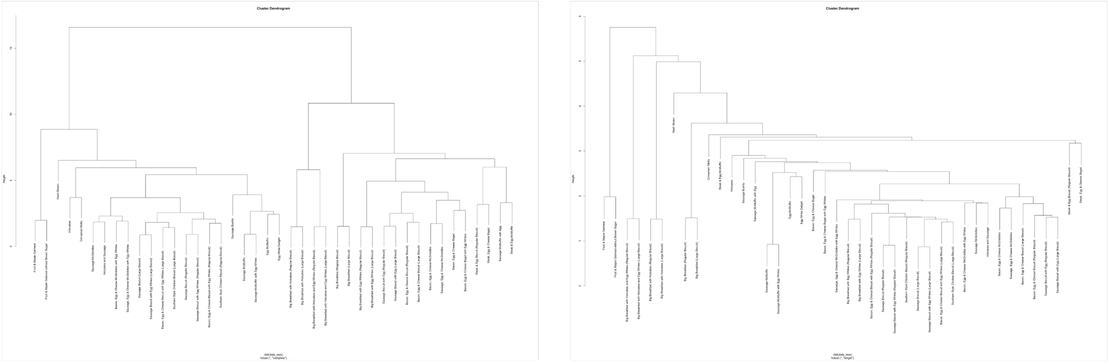


Figure 3.11

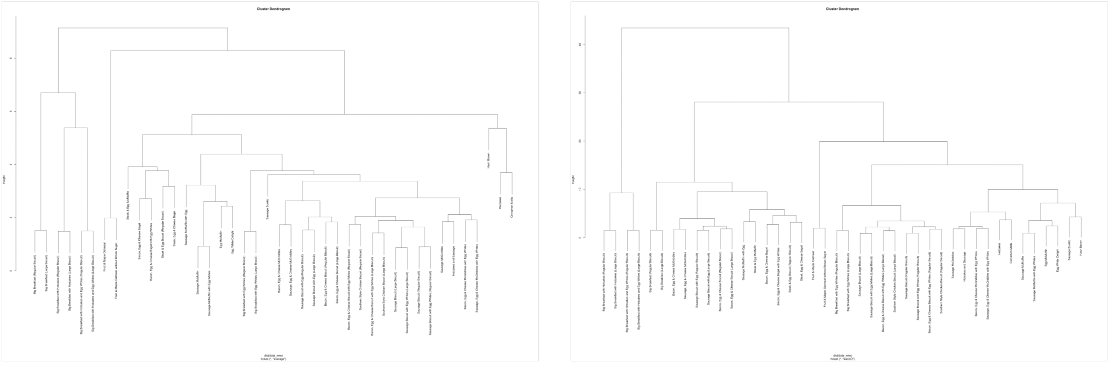


Figure 3.12: Hierarchical Clustering with Four Linkages

The tree using the "complete" method can be cut into 4 parts and the results are clearly shown in the following graph. The details of each clustering are displayed in the following table. And then in the next figure, the nutrient composition scores in each cluster are analyzed. Almost all of the nutrition facts of the first cluster are relatively low, especially for Cholesterol and Cholesterol (% Daily Value). At the same time, the breakfast items in the first large cluster are mostly simple Biscuits, English Muffin or Hotcakes with or without bacon or sausage, egg whites and cheese, except two items: Hash Brown and Cinnamon Melts. It is obvious that if a little bit more clusters are set, these two items will be attributed to another new cluster. In the second cluster, the content of Vitamin A (% Daily Value) is extremely high and the content of Cholesterol and Cholesterol (% Daily Value) is also higher than the other nutrition facts. The tree graph shows that most of the items in cluster two are of a larger size or formed with eggs but not egg whites. Moreover, the breakfast items in the cluster three mainly contain Carbohydrates, Carbohydrates (% Daily Value), Vitamin A (% Daily Value), Calories, Dietary Fiber, Iron (% Daily Value)Dietary and Fiber (% Daily Value), and the items in cluster three are mostly in a huge size with biscuit, hash browns, scrambled eggs which need more oil, sausage, even another staple food, Hotcakes. Cluster four is strongly different from breakfast items within the other clusters. Vitamin C (% Daily Value), Sugars, Dietary Fiber and Dietary Fiber (% Daily Value) are relatively high in this cluster and the items are both Fruit and Maple Oatmeals which are full of fruits and full servings of whole-grain oats.

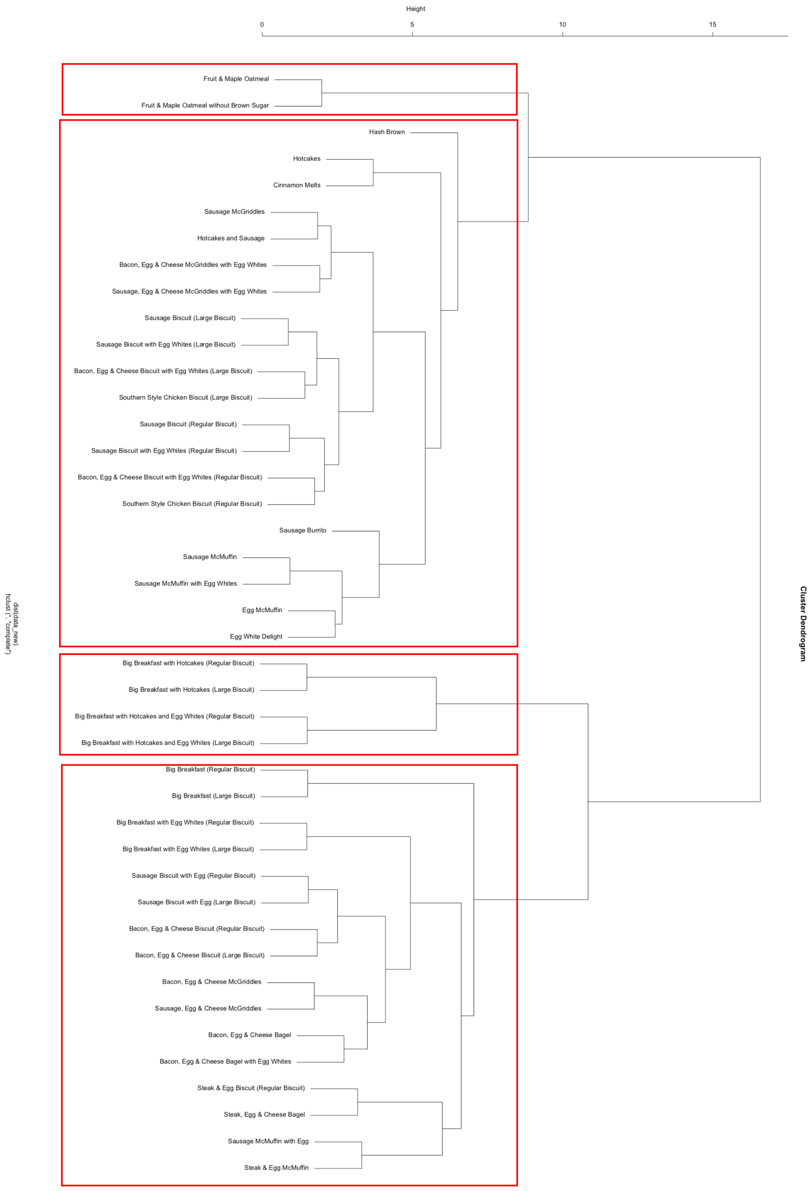
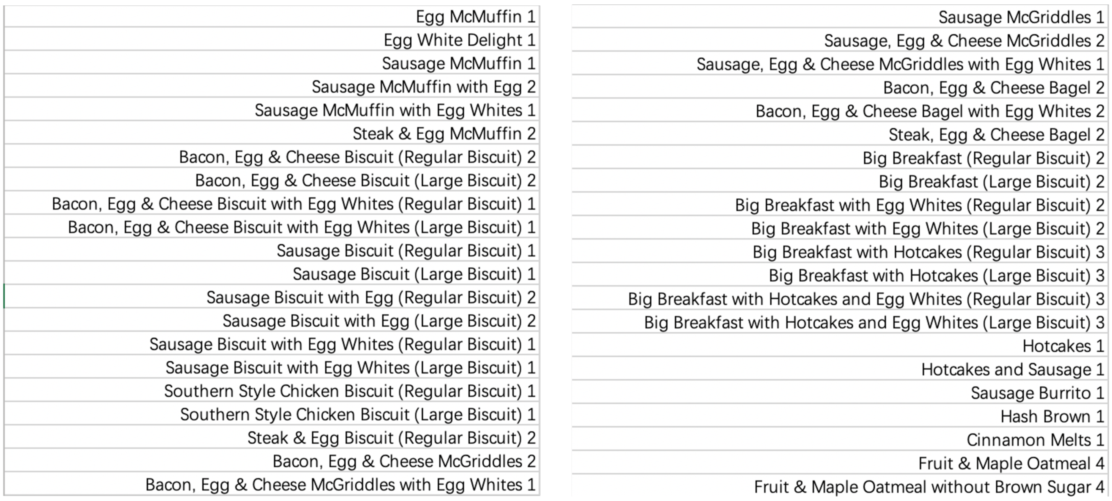
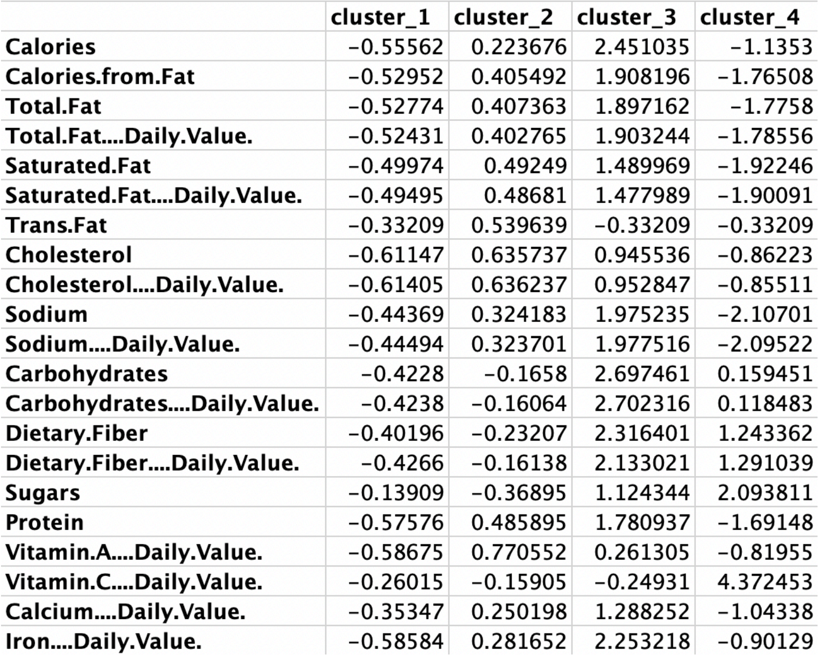


Figure 3.13: Hierarchical Clustering with Complete Linkages



Egg McMuffin 1	Sausage McGriddles 1
Egg White Delight 1	Sausage, Egg & Cheese McGriddles 2
Sausage McMuffin 1	Sausage, Egg & Cheese McGriddles with Egg Whites 1
Sausage McMuffin with Egg 2	Bacon, Egg & Cheese Bagel 2
Sausage McMuffin with Egg Whites 1	Bacon, Egg & Cheese Bagel with Egg Whites 2
Steak & Egg McMuffin 2	Steak, Egg & Cheese Bagel 2
Bacon, Egg & Cheese Biscuit (Regular Biscuit) 2	Big Breakfast (Regular Biscuit) 2
Bacon, Egg & Cheese Biscuit (Large Biscuit) 2	Big Breakfast (Large Biscuit) 2
Bacon, Egg & Cheese Biscuit with Egg Whites (Regular Biscuit) 1	Big Breakfast with Egg Whites (Regular Biscuit) 2
Bacon, Egg & Cheese Biscuit with Egg Whites (Large Biscuit) 1	Big Breakfast with Egg Whites (Large Biscuit) 2
Sausage Biscuit (Regular Biscuit) 1	Big Breakfast with Hotcakes (Regular Biscuit) 3
Sausage Biscuit (Large Biscuit) 1	Big Breakfast with Hotcakes (Large Biscuit) 3
Sausage Biscuit with Egg (Regular Biscuit) 2	Big Breakfast with Hotcakes and Egg Whites (Regular Biscuit) 3
Sausage Biscuit with Egg (Large Biscuit) 2	Big Breakfast with Hotcakes and Egg Whites (Large Biscuit) 3
Sausage Biscuit with Egg Whites (Regular Biscuit) 1	Hotcakes 1
Sausage Biscuit with Egg Whites (Large Biscuit) 1	Hotcakes and Sausage 1
Southern Style Chicken Biscuit (Regular Biscuit) 1	Sausage Burrito 1
Southern Style Chicken Biscuit (Large Biscuit) 1	Hash Brown 1
Steak & Egg Biscuit (Regular Biscuit) 2	Cinnamon Melts 1
Bacon, Egg & Cheese McGriddles 2	Fruit & Maple Oatmeal 4
Bacon, Egg & Cheese McGriddles with Egg Whites 1	Fruit & Maple Oatmeal without Brown Sugar 4

Figure 3.14: Results of Hierarchical Clustering



	cluster_1	cluster_2	cluster_3	cluster_4
Calories	−0.55562	0.223676	2.451035	−1.1353
Calories.from.Fat	−0.52952	0.405492	1.908196	−1.76508
Total.Fat	−0.52774	0.407363	1.897162	−1.7758
Total.Fat....Daily.Value.	−0.52431	0.402765	1.903244	−1.78556
Saturated.Fat	−0.49974	0.49249	1.489969	−1.92246
Saturated.Fat....Daily.Value.	−0.49495	0.48681	1.477989	−1.90091
Trans.Fat	−0.33209	0.539639	−0.33209	−0.33209
Cholesterol	−0.61147	0.635737	0.945536	−0.86223
Cholesterol....Daily.Value.	−0.61405	0.636237	0.952847	−0.85511
Sodium	−0.44369	0.324183	1.975235	−2.10701
Sodium....Daily.Value.	−0.44494	0.323701	1.977516	−2.09522
Carbohydrates	−0.4228	−0.1658	2.697461	0.159451
Carbohydrates....Daily.Value.	−0.4238	−0.16064	2.702316	0.118483
Dietary.Fiber	−0.40196	−0.23207	2.316401	1.243362
Dietary.Fiber....Daily.Value.	−0.4266	−0.16138	2.133021	1.291039
Sugars	−0.13909	−0.36895	1.124344	2.093811
Protein	−0.57576	0.485895	1.780937	−1.69148
Vitamin.A....Daily.Value.	−0.58675	0.770552	0.261305	−0.81955
Vitamin.C....Daily.Value.	−0.26015	−0.15905	−0.24931	4.372453
Calcium....Daily.Value.	−0.35347	0.250198	1.288252	−1.04338
Iron....Daily.Value.	−0.58584	0.281652	2.253218	−0.90129

Figure 3.15: Mean of Each Cluster

In order to make visualization of the hierarchical clustering, the clusters are projected on different principle components. The graph of projection on PC1 and PC2 shows that in the dimension of both PC1 and PC2, cluster three, cluster four and cluster one or two are separately located, but cluster one and cluster two are to some extent mixed together at the left bottom corner of the graph. However, in the right figure of projection on PC3 and PC4, the separation of clusters is not as clear as that in the last one. Cluster 4 is obviously separated in the dimension of PC4, but the other three clusters are not well separated in PC3 and PC4.

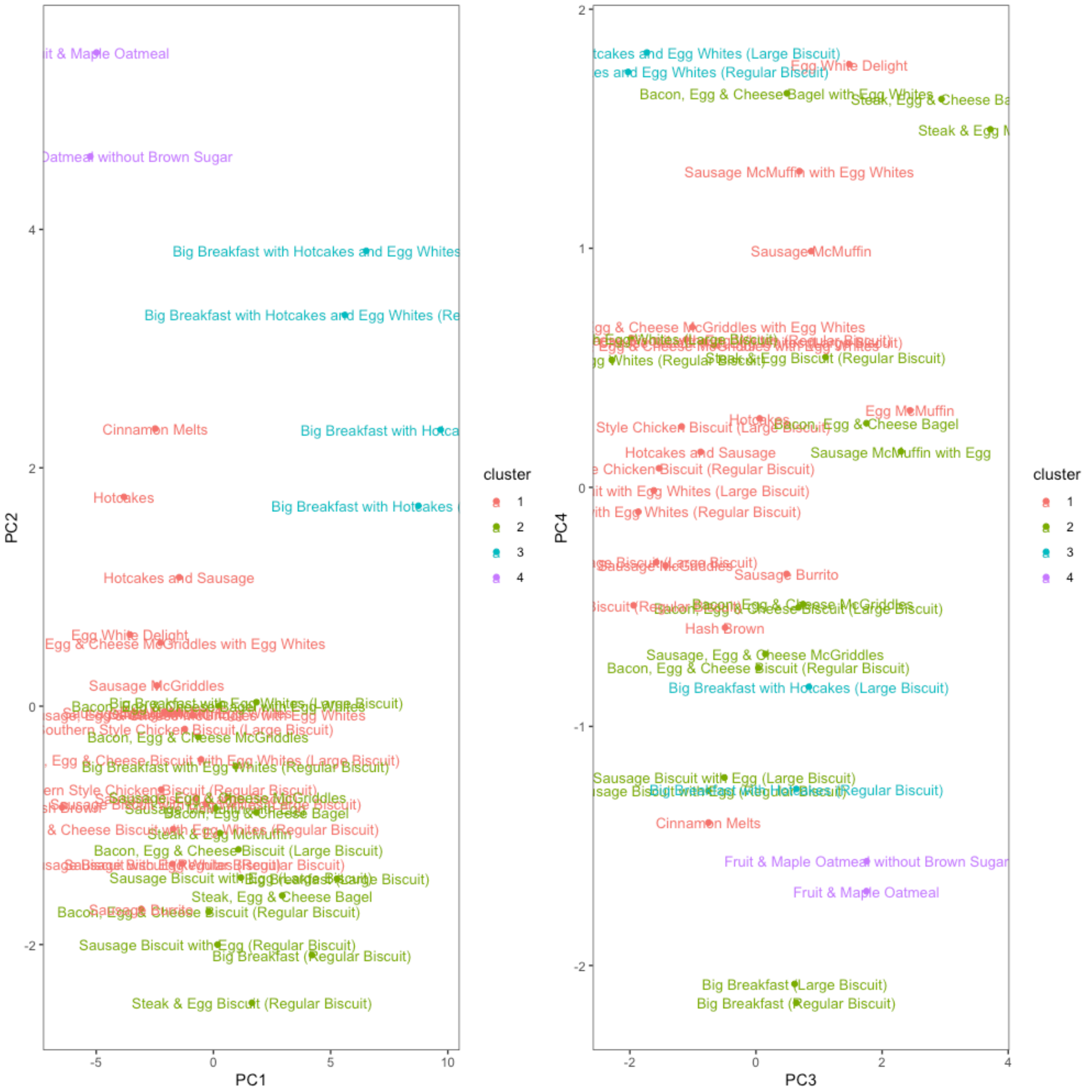


Figure 3.16: Projection on Principle Components

3.2.3 K-Means Clustering

In the method of K-means, the most important prior process is to choose the value of "K" which is the exact number of clusters, which directly affects the quality of clustering. There are two widely used methods to find the optimal "K" , the one is Elbow method, and the other is Sihouette coefficient method. The corresponding graphs are demonstrated below. It can be seen that there is little change in slope after K=3, so it can be inferred that the elbow point of the line in the left graph is K=3, which is an optimal value. In the right graph, it shows the average silhouette width from 1 to 10, and the maximum value of average silhouette width appears when K=2. Thus, K=3 is chosen as the number of the clusters, and then the implementation of K-means method is practicable.

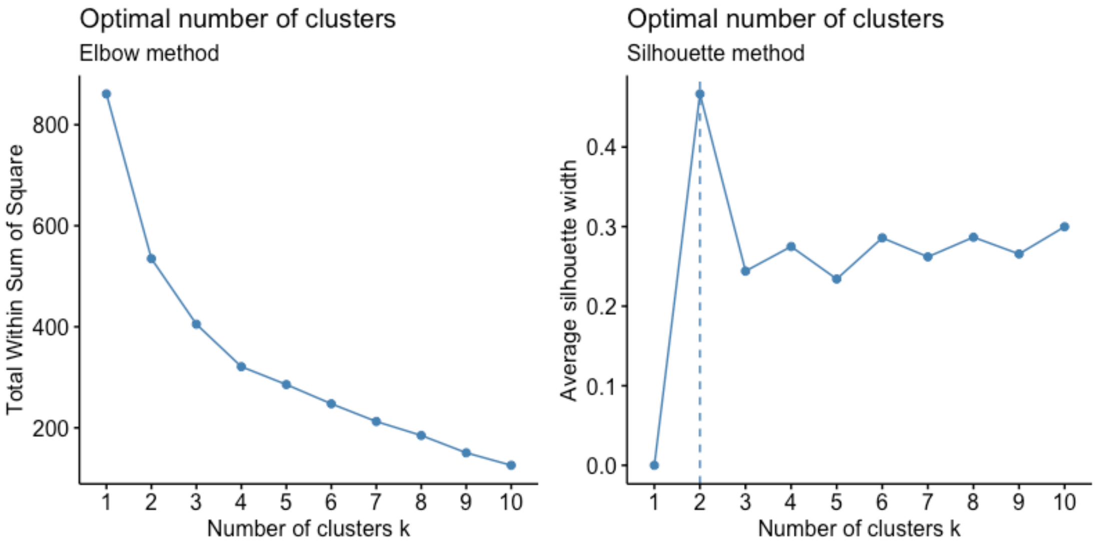
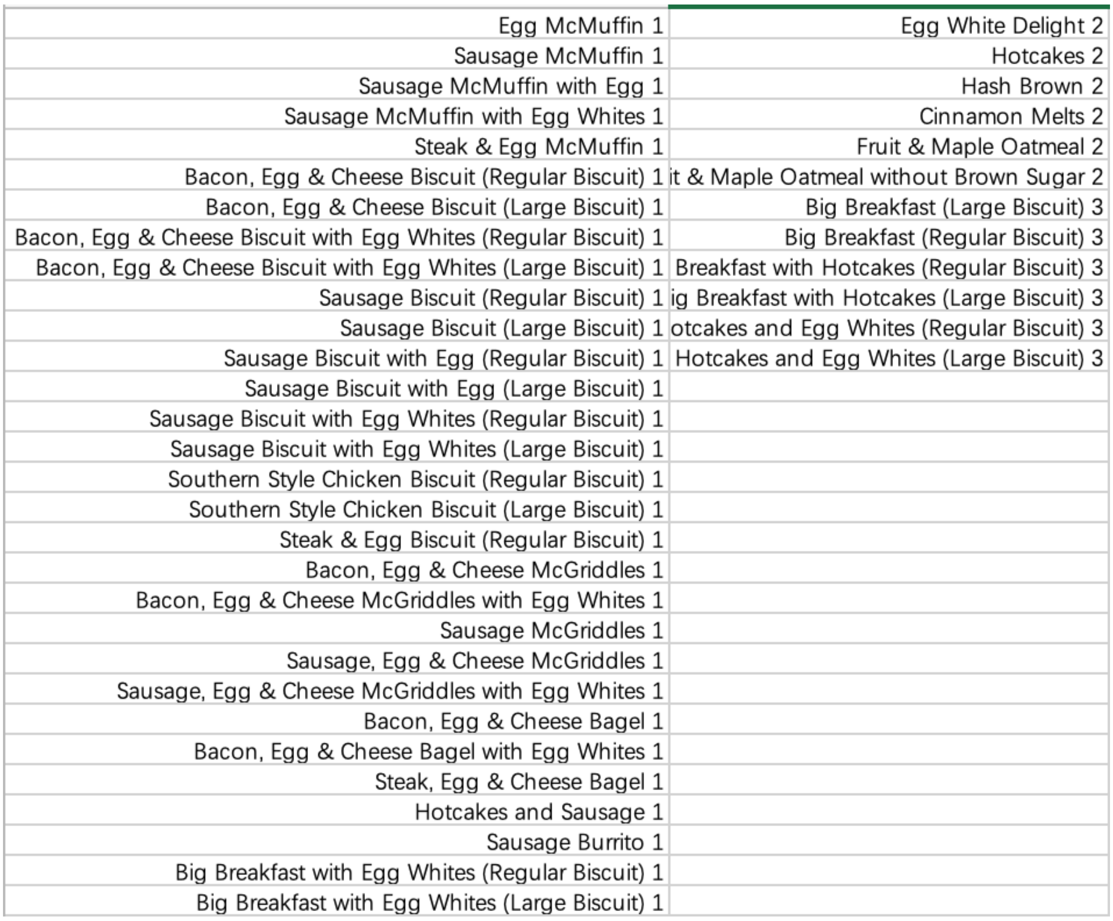


Figure 3.17: Optimal Number of Clusters in K-means Method

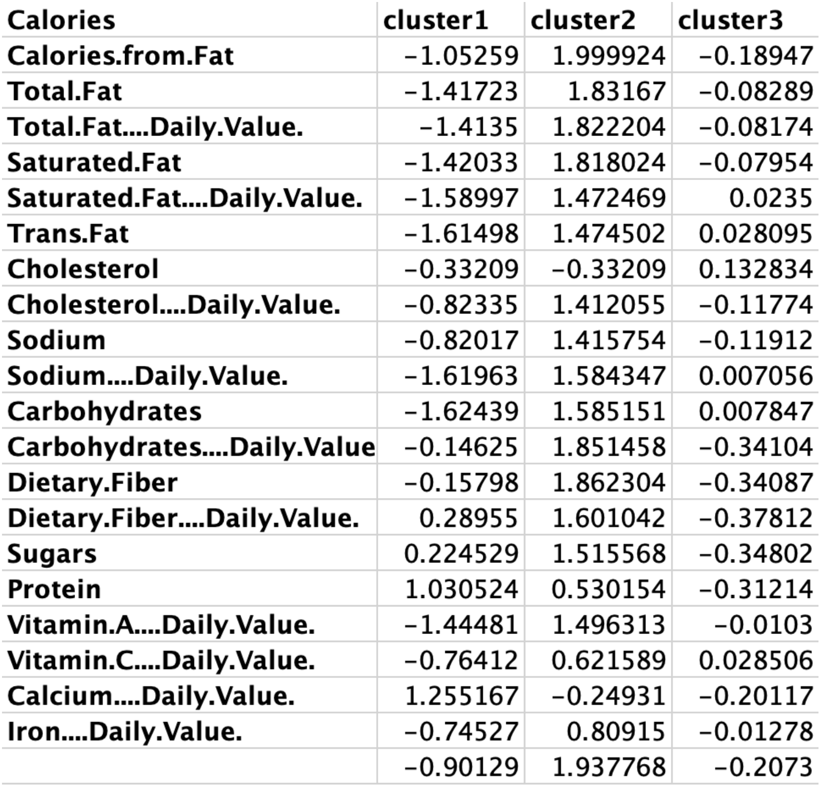
The following graph lists each breakfast item in each cluster, and it is clear that the components of the largest cluster are mostly similar to the results of hierarchical clustering. The items mostly include Biscuits, McMuffin, McGriddles and Bagels with sausage or bacon, egg or only egg whites and cheese of standard size. However, compared with the hierarchical clustering, this method collects more items together and obtains a more unbalanced result. There are respectively 6 items grouped in cluster two and three, but 30 items are contained in the first cluster. Moreover, the cluster two is composed of most "big breakfasts" which are including more ingredients and energy, which is similar to the third cluster in hierarchical method. In addition, fruits and small snacks, like Cinnamon Melts and Hotcakes, form the cluster three, which mostly lack meat.



Egg McMuffin 1	Egg White Delight 2
Sausage McMuffin 1	Hotcakes 2
Sausage McMuffin with Egg 1	Hash Brown 2
Sausage McMuffin with Egg Whites 1	Cinnamon Melts 2
Steak & Egg McMuffin 1	Fruit & Maple Oatmeal 2
Bacon, Egg & Cheese Biscuit (Regular Biscuit) 1	it & Maple Oatmeal without Brown Sugar 2
Bacon, Egg & Cheese Biscuit (Large Biscuit) 1	Big Breakfast (Large Biscuit) 3
Bacon, Egg & Cheese Biscuit with Egg Whites (Regular Biscuit) 1	Big Breakfast (Regular Biscuit) 3
Bacon, Egg & Cheese Biscuit with Egg Whites (Large Biscuit) 1	Breakfast with Hotcakes (Regular Biscuit) 3
Sausage Biscuit (Regular Biscuit) 1	ig Breakfast with Hotcakes (Large Biscuit) 3
Sausage Biscuit (Large Biscuit) 1	otcakes and Egg Whites (Regular Biscuit) 3
Sausage Biscuit with Egg (Regular Biscuit) 1	Hotcakes and Egg Whites (Large Biscuit) 3
Sausage Biscuit with Egg (Large Biscuit) 1	
Sausage Biscuit with Egg Whites (Regular Biscuit) 1	
Sausage Biscuit with Egg Whites (Large Biscuit) 1	
Southern Style Chicken Biscuit (Regular Biscuit) 1	
Southern Style Chicken Biscuit (Large Biscuit) 1	
Steak & Egg Biscuit (Regular Biscuit) 1	
Bacon, Egg & Cheese McGriddles 1	
Bacon, Egg & Cheese McGriddles with Egg Whites 1	
Sausage McGriddles 1	
Sausage, Egg & Cheese McGriddles 1	
Sausage, Egg & Cheese McGriddles with Egg Whites 1	
Bacon, Egg & Cheese Bagel 1	
Bacon, Egg & Cheese Bagel with Egg Whites 1	
Steak, Egg & Cheese Bagel 1	
Hotcakes and Sausage 1	
Sausage Burrito 1	
Big Breakfast with Egg Whites (Regular Biscuit) 1	
Big Breakfast with Egg Whites (Large Biscuit) 1	

Figure 3.18: Results of K-means Clustering

From the following table, the difference in nutritional composition between clusters can be found. The items in cluster one are rich in Vitamin C (% Daily Value), Sugars, Dietary Fiber and Dietary Fiber (% Daily Value), which almost cannot be made by the body and must be obtained from food. Cluster two contains breakfast items with high values of Calories, Calories from Fat, Total Fat, Total Fat (% Daily Value), Carbohydrates, Carbohydrates (% Daily Value) and Iron (% Daily Value). In order to keep fit, it is necessary for everybody to balance the number of calories balance calories in and calories out. The breakfast items in the third cluster are rich in Saturated Fat, Saturated Fat (% Daily Value), Trans Fat, Sodium, Sodium (% Daily Value), and Vitamin A (% Daily Value), which are mostly associated with increased risk of suffering cardiovascular disease, high blood pressure and heart attacks.



Calories	cluster1	cluster2	cluster3
Calories.from.Fat	−1.05259	1.999924	−0.18947
Total.Fat	−1.41723	1.83167	−0.08289
Total.Fat....Daily.Value.	−1.4135	1.822204	−0.08174
Saturated.Fat	−1.42033	1.818024	−0.07954
Saturated.Fat....Daily.Value.	−1.58997	1.472469	0.0235
Trans.Fat	−1.61498	1.474502	0.028095
Cholesterol	−0.33209	−0.33209	0.132834
Cholesterol....Daily.Value.	−0.82335	1.412055	−0.11774
Sodium	−0.82017	1.415754	−0.11912
Sodium....Daily.Value.	−1.61963	1.584347	0.007056
Carbohydrates	−1.62439	1.585151	0.007847
Carbohydrates....Daily.Value	−0.14625	1.851458	−0.34104
Dietary.Fiber	−0.15798	1.862304	−0.34087
Dietary.Fiber....Daily.Value.	0.28955	1.601042	−0.37812
Sugars	0.224529	1.515568	−0.34802
Protein	1.030524	0.530154	−0.31214
Vitamin.A....Daily.Value.	−1.44481	1.496313	−0.0103
Vitamin.C....Daily.Value.	−0.76412	0.621589	0.028506
Calcium....Daily.Value.	1.255167	−0.24931	−0.20117
Iron....Daily.Value.	−0.74527	0.80915	−0.01278
	−0.90129	1.937768	−0.2073

Figure 3.19: Centers of each cluster

The following left figure shows that there is a good separation between cluster one and two in PC1 dimension. However, in the projection of PC3 and PC4, clusters are mixed together and show almost little separation.

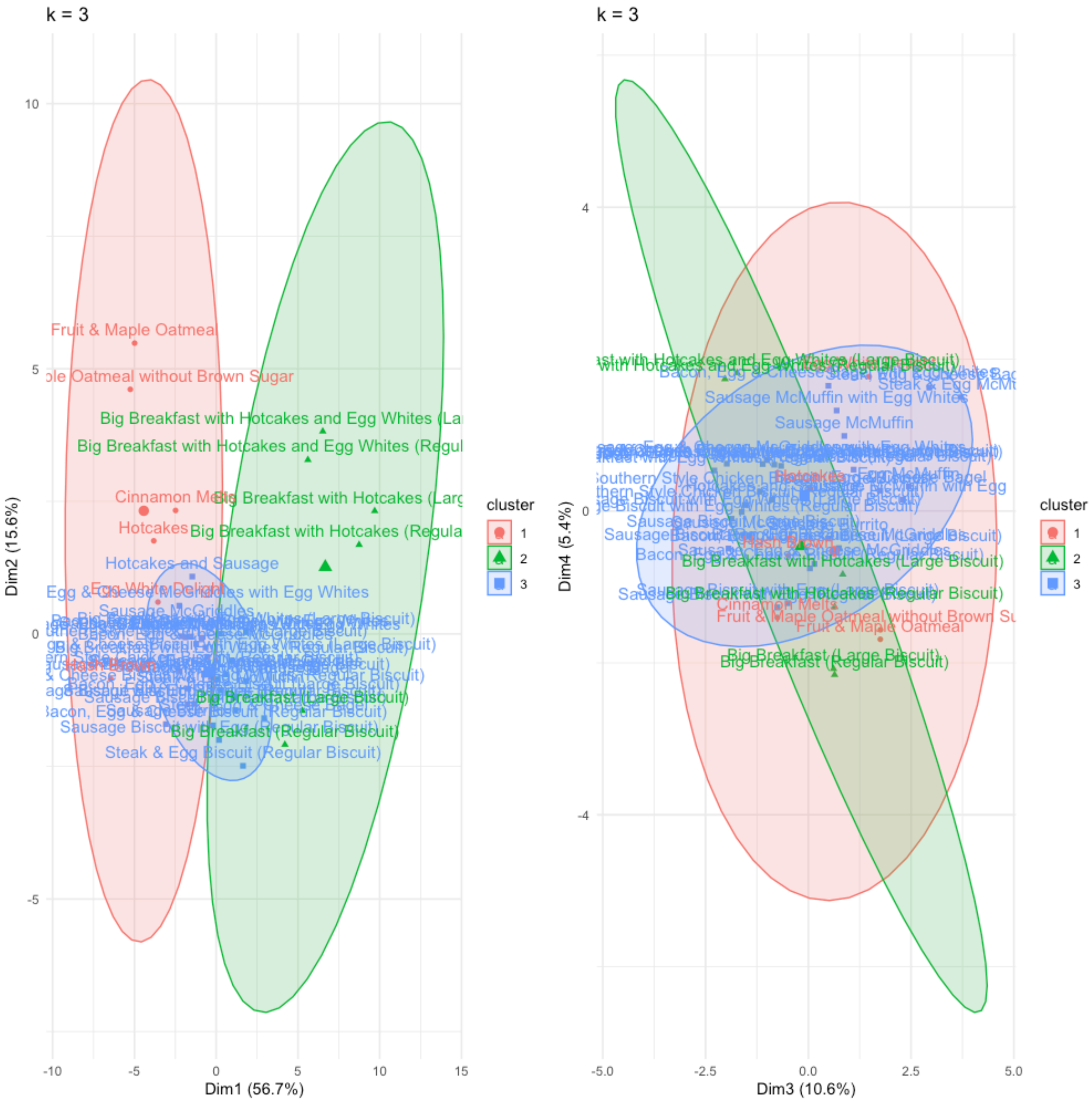


Figure 3.20: Projection on Principle Components

3.2.4 Data-Based Spatial Clustering of Application with Noise(DBSCAN)

In this part, the method of DBSCAN will be mainly discussed. It is a density-based clustering algorithm based on high-density connected regions, which divides regions with high enough density into clusters and finds clusters of arbitrary shapes with noises. At the same time, because of the characteristics of function "dbscan" , it is more appropriate to use the data set after implementing dimensional reduction. In this function, two parameters are necessarily set at first, "eps" (the radius of the neighborhoods around the chosen data point p) and "minPts" (the minimum number of data points in a neighborhood to define a cluster). The value of "eps" can be obtained from the inflection point of the K-distance graph. The choice of epsilon is essential, if the parameter is set too small, most of the data points cannot be clustered; if the parameter is set too large, multiple clusters and most objects will be merged into the same cluster. From the following graph, "eps" with the value of 5 can be selected. The value of "minPts" is always determined by dimensions. If the value is too small, the result in the sparse cluster will be considered as the boundary point and will not be used for further expansion, because the density is less than ”minPts”; if the value is too large, the two adjacent clusters with higher density may be merged into the same cluster. Therefore, it is usually set the value as a number greater than the number of features plus one, so in this data set, 22 is chosen.

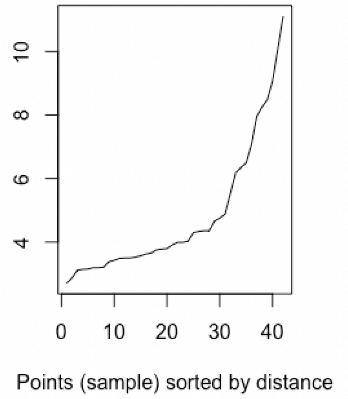


Figure 3.21: Optimal Value of Epsilon

As the following graph and table display, the data set is separated into two clusters, the first cluster is formed by six “Big Breakfasts” , two Fruit Maple Oatmeal, two kinds of burgers with steak and Hash Brown. These items are mostly with high PC2 and low PC4, which means getting more carbohydrates, common vitamins, minerals and less protein. The rest items are collected in cluster two, which are consist of different kinds of burgers with different kinds of meat, eggs and cheese. However, compared with the above two methods, this model cannot separate items in a clear standard level, which is obviously shown in the following table. Hence, this model is not suitable for the data set.

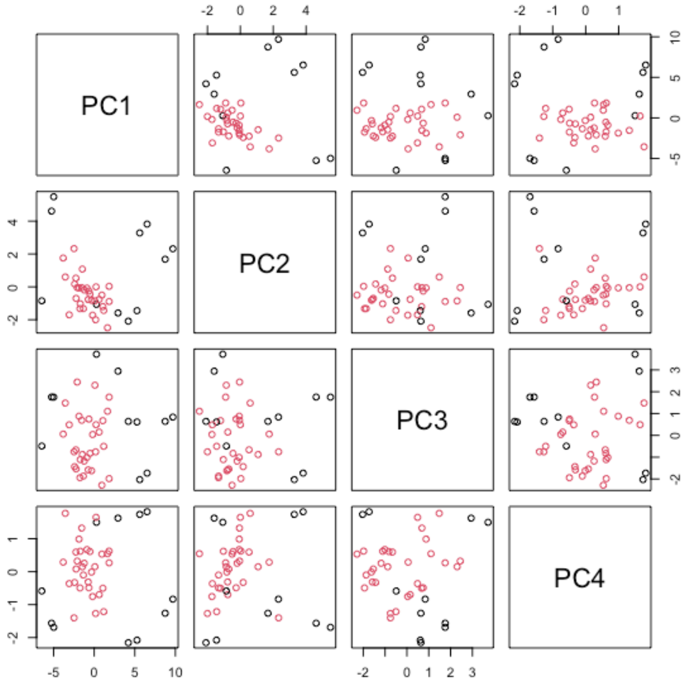
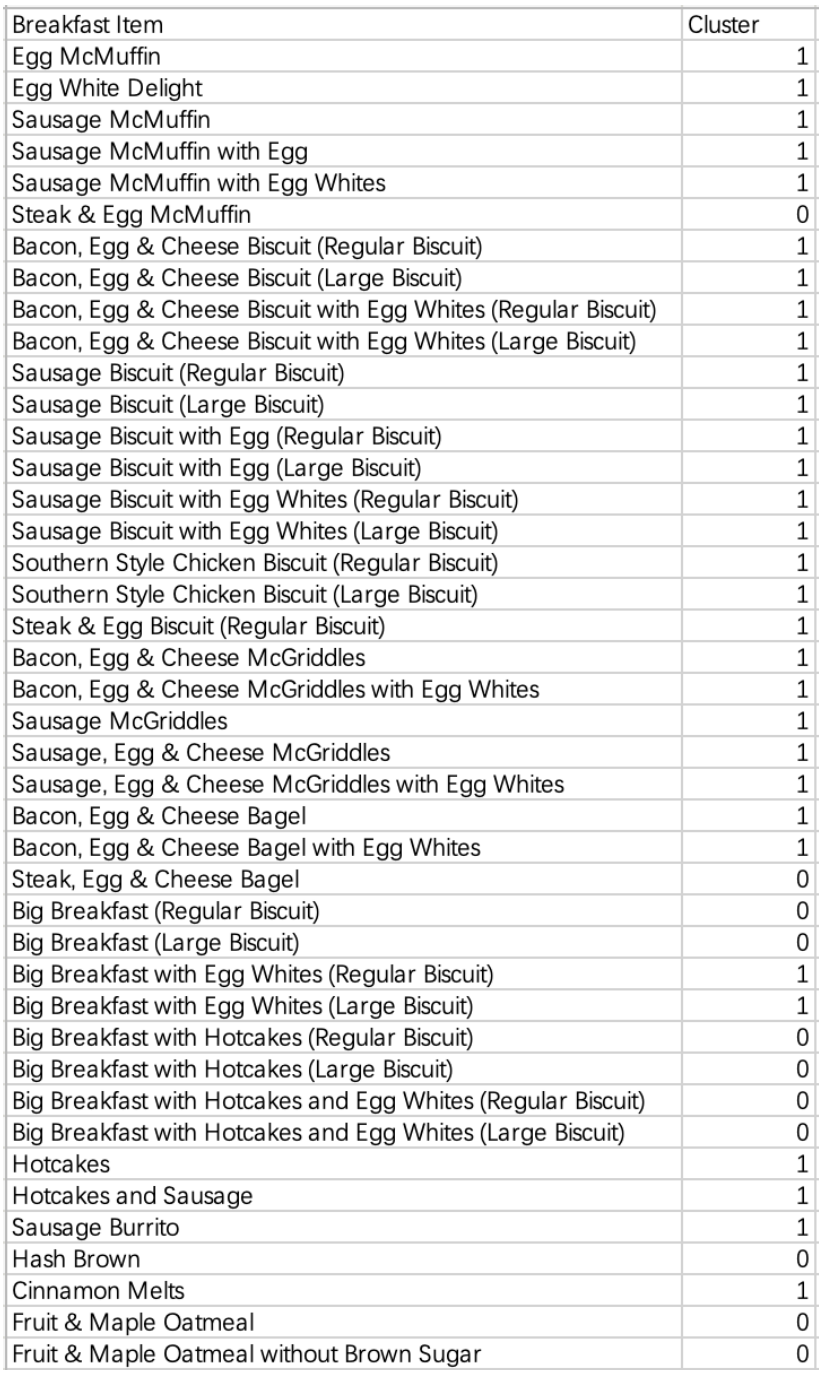


Figure 3.22: Results in Each Dimension



Breakfast Item	Cluster
Egg McMuffin	1
Egg White Delight	1
Sausage McMuffin	1
Sausage McMuffin with Egg	1
Sausage McMuffin with Egg Whites	1
Steak & Egg McMuffin	0
Bacon, Egg & Cheese Biscuit (Regular Biscuit)	1
Bacon, Egg & Cheese Biscuit (Large Biscuit)	1
Bacon, Egg & Cheese Biscuit with Egg Whites (Regular Biscuit)	1
Bacon, Egg & Cheese Biscuit with Egg Whites (Large Biscuit)	1
Sausage Biscuit (Regular Biscuit)	1
Sausage Biscuit (Large Biscuit)	1
Sausage Biscuit with Egg (Regular Biscuit)	1
Sausage Biscuit with Egg (Large Biscuit)	1
Sausage Biscuit with Egg Whites (Regular Biscuit)	1
Sausage Biscuit with Egg Whites (Large Biscuit)	1
Southern Style Chicken Biscuit (Regular Biscuit)	1
Southern Style Chicken Biscuit (Large Biscuit)	1
Steak & Egg Biscuit (Regular Biscuit)	1
Bacon, Egg & Cheese McGriddles	1
Bacon, Egg & Cheese McGriddles with Egg Whites	1
Sausage McGriddles	1
Sausage, Egg & Cheese McGriddles	1
Sausage, Egg & Cheese McGriddles with Egg Whites	1
Bacon, Egg & Cheese Bagel	1
Bacon, Egg & Cheese Bagel with Egg Whites	1
Steak, Egg & Cheese Bagel	0
Big Breakfast (Regular Biscuit)	0
Big Breakfast (Large Biscuit)	0
Big Breakfast with Egg Whites (Regular Biscuit)	1
Big Breakfast with Egg Whites (Large Biscuit)	1
Big Breakfast with Hotcakes (Regular Biscuit)	0
Big Breakfast with Hotcakes (Large Biscuit)	0
Big Breakfast with Hotcakes and Egg Whites (Regular Biscuit)	0
Big Breakfast with Hotcakes and Egg Whites (Large Biscuit)	0
Hotcakes	1
Hotcakes and Sausage	1
Sausage Burrito	1
Hash Brown	0
Cinnamon Melts	1
Fruit & Maple Oatmeal	0
Fruit & Maple Oatmeal without Brown Sugar	0

Figure 3.23: Results of DBSCAN

3.2.5 Spectral Clustering

Spectral clustering is an algorithm coming from graph theory, which is more adaptable to uncommon data distribution, and the clustering effect is always valuable. At the same time, the calculation amount of this clustering is also much smaller. The basic idea is to use the eigenvectors obtained after the eigendecomposition of the Laplace matrix of the sample data for clustering. It can identify the sample space of any shape and converge to the optimal solution for the overall situation.

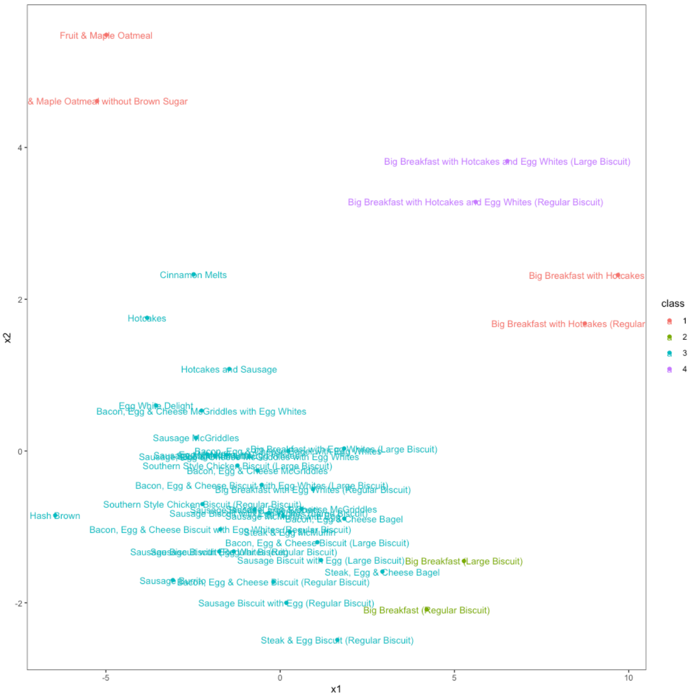


Figure 3.24: Results in Each Dimension

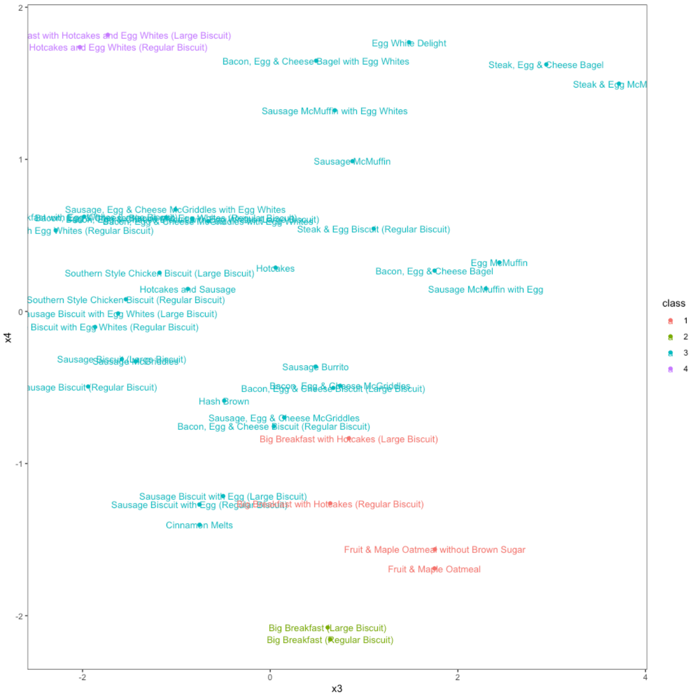
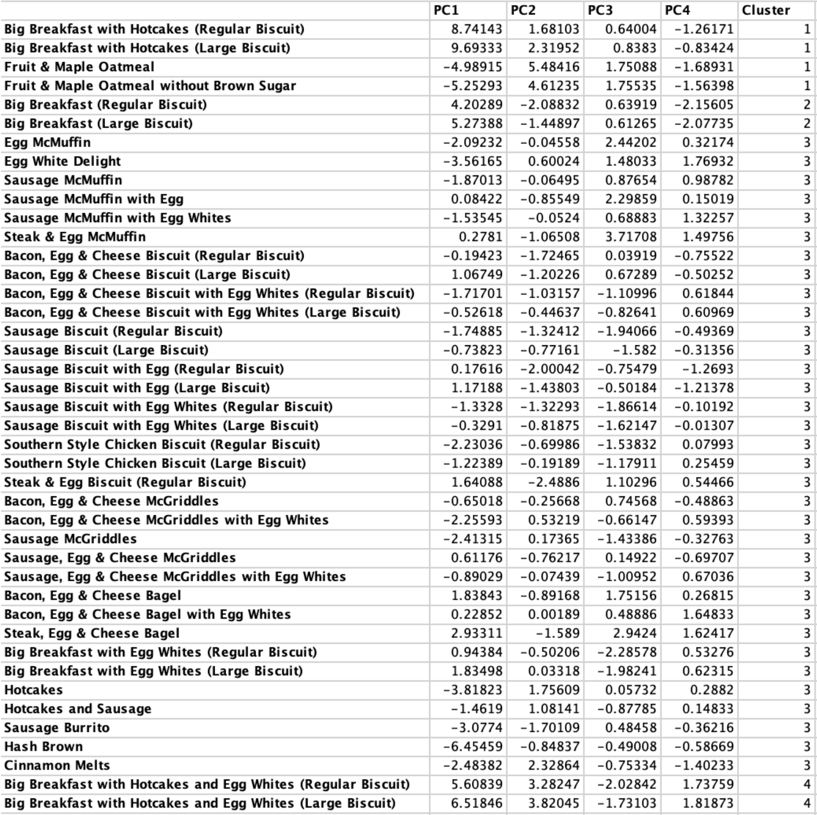


Figure 3.25: Results in Each Dimension

From the above graphs, it can be seen that the data set are grouped into four clusters, The first cluster consists of two Big Breakfast with hotcake and two Fruit Maple Oatmeals, which have relatively high PC2 and relatively low PC4. It means that items in this cluster are rich in Vitamin C, sugars, dietary fiber and other carbohydrates, while they lack common vitamins, minerals and amino acids. Cluster two and cluster four are both with high PC1, which means containing high total energy and calories. But the cluster four is higher in PC2 and PC4, which means the items are rich in common vitamins, minerals, carbohydrates and amino acids. These benefits may come from hotcakes and egg whites, but not scrambled eggs. The other items are getting together in cluster three, which are mostly burgers.



	PC1	PC2	PC3	PC4	Cluster
Big Breakfast with Hotcakes (Regular Biscuit)	8.74143	1.68103	0.64004	-1.26171	1
Big Breakfast with Hotcakes (Large Biscuit)	9.69333	2.31952	0.8383	-0.83424	1
Fruit & Maple Oatmeal	-4.98915	5.48416	1.75088	-1.68931	1
Fruit & Maple Oatmeal without Brown Sugar	-5.25293	4.61235	1.75535	-1.56398	1
Big Breakfast (Regular Biscuit)	4.20289	-2.08832	0.63919	-2.15605	2
Big Breakfast (Large Biscuit)	5.27388	-1.44897	0.61265	-2.07735	2
Egg McMuffin	-2.09232	-0.04558	2.44202	0.32174	3
Egg White Delight	-3.56165	0.60024	1.48033	1.76932	3
Sausage McMuffin	-1.87013	-0.06495	0.87654	0.98782	3
Sausage McMuffin with Egg	0.08422	-0.85549	2.29859	0.15019	3
Sausage McMuffin with Egg Whites	-1.53545	-0.0524	0.68883	1.32257	3
Steak & Egg McMuffin	0.2781	-1.06508	3.71708	1.49756	3
Bacon, Egg & Cheese Biscuit (Regular Biscuit)	-0.19423	-1.72465	0.03919	-0.75522	3
Bacon, Egg & Cheese Biscuit (Large Biscuit)	1.06749	-1.20226	0.67289	-0.50252	3
Bacon, Egg & Cheese Biscuit with Egg Whites (Regular Biscuit)	-1.71701	-1.03157	-1.10996	0.61844	3
Bacon, Egg & Cheese Biscuit with Egg Whites (Large Biscuit)	-0.52618	-0.44637	-0.82641	0.60969	3
Sausage Biscuit (Regular Biscuit)	-1.74885	-1.32412	-1.94066	-0.49369	3
Sausage Biscuit (Large Biscuit)	-0.73823	-0.77161	-1.582	-0.31356	3
Sausage Biscuit with Egg (Regular Biscuit)	0.17616	-2.00042	-0.75479	-1.2693	3
Sausage Biscuit with Egg (Large Biscuit)	1.17188	-1.43803	-0.50184	-1.21378	3
Sausage Biscuit with Egg Whites (Regular Biscuit)	-1.3328	-1.32293	-1.86614	-0.10192	3
Sausage Biscuit with Egg Whites (Large Biscuit)	-0.3291	-0.81875	-1.62147	-0.01307	3
Southern Style Chicken Biscuit (Regular Biscuit)	-2.23036	-0.69986	-1.53832	0.07993	3
Southern Style Chicken Biscuit (Large Biscuit)	-1.22389	-0.19189	-1.17911	0.25459	3
Steak & Egg Biscuit (Regular Biscuit)	1.64088	-2.4886	1.10296	0.54466	3
Bacon, Egg & Cheese McGriddles	-0.65018	-0.25668	0.74568	-0.48863	3
Bacon, Egg & Cheese McGriddles with Egg Whites	-2.25593	0.53219	-0.66147	0.59393	3
Sausage McGriddles	-2.41315	0.17365	-1.43386	-0.32763	3
Sausage, Egg & Cheese McGriddles	0.61176	-0.76217	0.14922	-0.69707	3
Sausage, Egg & Cheese McGriddles with Egg Whites	-0.89029	-0.07439	-1.00952	0.67036	3
Bacon, Egg & Cheese Bagel	1.83843	-0.89168	1.75156	0.26815	3
Bacon, Egg & Cheese Bagel with Egg Whites	0.22852	0.00189	0.48886	1.64833	3
Steak, Egg & Cheese Bagel	2.93311	-1.589	2.9424	1.62417	3
Big Breakfast with Egg Whites (Regular Biscuit)	0.94384	-0.50206	-2.28578	0.53276	3
Big Breakfast with Egg Whites (Large Biscuit)	1.83498	0.03318	-1.98241	0.62315	3
Hotcakes	-3.81823	1.75609	0.05732	0.2882	3
Hotcakes and Sausage	-1.4619	1.08141	-0.87785	0.14833	3
Sausage Burrito	-3.0774	-1.70109	0.48458	-0.36216	3
Hash Brown	-6.45459	-0.84837	-0.49008	-0.58669	3
Cinnamon Melts	-2.48382	2.32864	-0.75334	-1.40233	3
Big Breakfast with Hotcakes and Egg Whites (Regular Biscuit)	5.60839	3.28247	-2.02842	1.73759	4
Big Breakfast with Hotcakes and Egg Whites (Large Biscuit)	6.51846	3.82045	-1.73103	1.81873	4

Figure 3.26: Results of Spectral Clustering

Conclusion

In the study, the whole process is focusing on four popular unsupervised machine learning algorithms, Hierarchical Clustering, K-means Clustering, DBSCAN(Density-Based Spatial Clustering of Applications with Noise) and Spectral Clustering. After understanding the development process and common practical applications of these algorithms, the basic principles are introduced. Afterwards, these algorithms were applied to the McDonald's breakfast data set. Based on the nutrition facts list, attempts were made to cluster 42 common McDonald's breakfasts and to explore what commonalities and similarities were found in the nutrient content of different categories of breakfasts. After dimensionality reduction by principal component analysis, four principal components are selected. These four principle components can explain more than 88% variation of the data set. The importance of the components shows that PC1, PC2, PC3 and PC4 respectively represent 56.7%, 15.6%, 10.6% and 5.4% variation of the whole data set.

Although the number of clusters is different because of different algorithms, there are several similarities in obtained clusters. In addition to hierarchical clustering divides the data set into four relatively balance groups, the remaining three methods divide the data set into two or three unbalanced categories, in which hamburgers with different kinds of bread and meat, eggs or egg whites and cheese are mostly clustered into the same category. This largest cluster contains usually more than thirty items. The nutrition facts of these breakfasts did not differ much, with similar values for calories, various fats, sodium, carbohydrates, dietary fiber, protein and iron. However, the values for cholesterol, vitamin A, and vitamin C are quite different. And this is the reason why data set are clustered into four groups within Hierarchical Clustering method. This largest cluster containing different kinds of hamburgers and obtained from K-means Clustering, DBSCAN and Spectral Clustering methods are mostly separated into two groups by Hierarchical clustering. The breakfasts in one separated group are extremely rich in cholesterol and with relatively high values of vitamin A. And the nutrition facts of those items in another separated group are similar to that largest cluster. Moreover, from the data set it can be seen that these hamburger items with high values of cholesterol and vitamin A are mostly including eggs but not egg whites. Cholesterol is necessary for our body and it can only be found in animal products, especially egg yolks that are rich in it.(Daeun Kim et al., 2021)[31] However, all the cholesterol that the human body needs can be made by the human body self, and diets with high cholesterol are always correlated with an increased risk of suffering cardiovascular disease.(Poli, A. et al., 2021)[32] According to the 2000 calories a day benchmark, cholesterol intake should be less than 300 mg per day, and each meal intake should be less than 20% of the daily value.(Lim, C. G. Y., Tai, E. S. and van, D. R. M., 2022)[33] However, Cholesterols(% Daily Value) of these items are almost greater than 80, which means if possible, keeping a diet with cholesterol as low as possible and skipping these breakfasts.

Breakfasts which are different from burgers, such as Hash Brown, Hotcakes, Cinnamon melts, Fruit Maple Oatmeals with and without brown sugar and "Big Breakfasts" are separated in different clusters by different algorithms. In Hierarchical Clustering, K-means Clustering and DBSCAN, four "Big Breakfasts with Hotcakes" are grouped in a separate cluster with the highest values about calories, different kinds of fat, sodium, carbohydrates, dietary fiber, sugars, protein, minerals calcium and iron. It can be referred that the reason is the large serving sizes of these breakfasts. Especially, Hotcakes and Hash Brown are always in the same cluster because their nutrition facts are quite similar, and they are both rich in Carbohydrates. Carbohydrates provide calories for the body, and the body can break down carbohydrates into glucose, the glucose in the blood can be used directly to provide energy for the brain and muscles, and the glucose can also be stored in the liver or muscles to provide energy for future activities.(Romeo B.Batacan Jr et al., 2018)[34] If someone needs a little extra energy urgently, Hotcakes and Hash Brown are good options. Meanwhile, Fruit Maple Oatmeals with and without brown sugar are always clustered away from different kinds of burgers. In Hierarchical Clustering, they are in a separate group, and the reason is explicable. They consist of oatmeal, diced apples, cranberry raisin blend and light cream, which are low in calories, fat, cholesterol, sodium, protein, and rich in carbohydrates, dietary fiber and vitamins. If the consumers pursue a healthy diet, even want to lose weight and prefer McDonald' s breakfasts, Fruit Maple Oatmeals are the best choice. From the results computed from these four clustering algorithms, it can be concluded that the

Hierarchical Clustering method is more suitable for this data set. Within this method, breakfast items are classified into four groups and the degree of similarity within the same cluster is high, but the differences between different clusters are large.

There are also some suggestions for choosing McDonald’ s breakfasts that can be inferred. As a multinational chain restaurant famous for its fast food, McDonald’s has attracted many people with a fast-paced lifestyle. They need healthy food in a short amount of time, and especially want a quick breakfast of various nutrients. Nutrient-dense foods provide vitamins, minerals, and other energy-enhancing and health-promoting ingredients with little or no saturated fat, trans fat, added sugar, and sodium. Therefore, based on the nutrition facts label and the cluster analytics done, Fruit Maple Oatmeals will be preferentially recommended to workers who do not need a lot of physical labor. They will get the right amount of nutrients, such as Vitamin C, dietary fiber, carbohydrates and sugars, by consuming oatmeal, apples, raisins, dried cranberries and cream. That will help keep full, limit calories intake, and increase satisfaction, while also providing adequate energy and vitamins. If consumers are manual workers, they can choose Big Breakfast with Egg Whites, which contains more protein, carbohydrates, vitamins and minerals within the limit of 2000 calories per day in the largest cluster obtained from K-means Clustering, DBSCAN and Spectral Clustering. In addition, people with chronic diseases and those at risk for disease should be more cautious in their breakfast choices. For example, people with cardiovascular disease should try to reduce sodium, cholesterol and fat intake, so the items in cluster two obtained from Hierarchical Clustering are not suggested, such as Steak Egg Biscuit and Steak Egg McMuffin; and people with type two diabetes should reduce their intake of sugar, saturated fat and trans fat, so the items in cluster two obtained from hierarchical clustering are not suggested, such as Bacon, Egg Cheese McGriddles with Egg Whites.(Hajime Haimoto et al., 2021)[35]

References

[1] Janet Chrzan and John A. Brett. Food health : nutrition, technology, and public health. Research methods for anthropological studies of food and nutrition: Volume III. Berghahn Books, 2019.

[2] Malgorzata Kostecka. The effect of the ”colorful eating is healthy eating” long-term nutrition education program for 3- to 6-year-olds on eating habits in the family and parental nutrition knowledge. International journal of environmental research and public health, 19(4), 2022.

[3] Bruce A. Silverglade. The nutrition labeling and education act: Progress to date and challenges for the future. Journal of Public Policy Marketing, 15(1):148 – 150, 1996.

[4] Siva K. Balasubramanian and Catherine Cole. Consumers' search and use of nutrition information: The challenge and promise of the nutrition labeling and education act. Journal of Marketing, 66(3):112 – 127, 2002.

[5] Asirvatham Jebaraj, McNamara Paul E., and Baylis Kathy. Informational campaign effects of the nutrition labeling and education act (nlea) of 1990 on diet. Cogent Social Sciences, 3(1), 2017.

[6] Paul J. Petruccelli. Consumer and marketing implications of information provision: The case of the nutrition labeling and education act of 1990. Journal of Public Policy Marketing, 15(1):150 – 153, 199.

[7] Seonghee Cho and Sooyeol Kim. Does a healthy lifestyle matter? a daily diary study of unhealthy eating at home and behavioral outcomes at work. Journal of Applied Psychology, 107(1):23 – 39, 2022.

[8] Camila Weschenfelder, Philip Sapp, Terrence Riley, Kristina Petersen, Jacqueline Tereza da Silva, Angela Cristine Bersch-Ferreira, Rachel Helena Vieira Machado, Erlon Oliveira de Abreu-Silva, Lucas Ribeiro Silva, Bernardete Weber, Alexandre Schaan de Quadros, Penny Kris-Etherton, and Aline Marcadenti. Absolute and relative agreement between the current and modified brazilian cardioprotective nutritional program dietary index (balance di) and the american heart association healthy diet score (aha-ds) in post myocardial infarction patients. Nutrients, 14(7), 2022.

[9] Glympi Alkyoni, Chasioti Amalia, and Bälter Katarina. Dietary interventions to promote healthy eating among office workers: A literature review. Nutrients, 12(3754):3754, 2020.

[10] Malcolm McDonald and Grant Oliver. Malcolm McDonald on value propositions : how to develop them, how to quantify them. Kogan Page Limited, 2019.

[11] Adrian E. Tschoegl. Mcdonald's – much maligned, but an engine of economic development. Global Economy Journal, 7(4):1 – 16, 2007.

[12] Abubakar Abid, Martin J Zhang, Vivek K Bagaria, and James Zou. Exploring patterns enriched in a dataset with contrastive principal component analysis. Nature communications, 9(1):2134, May 2018.

[13] Ian Jolliffe and Jorge Cadima. Principal component analysis: A review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374:20150202, 04 2016.

[14] Félix Bigand, Elise Prigent, Bastien Berret, and Annelies Braffort. Decomposing spontaneous sign language into elementary movements: A principal component analysis-based approach. PloS one, 16(10):e0259464, 2021.

[15] Vinh Dao Nguyen, Nhat Nguyen Thi Duy, Bruin Erwin de, Vy Nguyen Ha Thao, Thao Tran Thi Nhu, Phuong Huynh Thi, Anh Pham Hong, Todd Stacy, Quan Tran Minh, Thanh Nguyen Thi Le, Lien Nguyen Thi Nam, Ha Nguyen Thi Hong, Hong Tran Thi Kim, Thai Pham Quang, Choisy Marc, Nguyen Tran Dang, Simmons Cameron P., Thwaites Guy E., Clapham Hannah E., Chau Nguyen Van Vinh, Koopmans Marion, and Boni Maciej F. Age-seroprevalence curves for the multi-strain structure of influenza a virus. Nature Communications, 12(1):1 – 9, 2021.

[16] Olivier de Viron, Michel Van Camp, Alexia Grabkowiak, and Ana M. G. Ferreira. Comparing global seismic tomography models using varimax principal component analysis. Solid Earth, 12(7):1601 – 1634, 2021.

[17] Oded Maimon and Lior Rokach. Data Mining and Knowledge Discovery Handbook. [electronic resource]. Springer US, 2005.

[18] Jiawei Han, Micheline Kamber, and Jian Pei. 10 - cluster analysis: Basic concepts and methods. In Jiawei Han, Micheline Kamber, and Jian Pei, editors, Data Mining (Third Edition), The Morgan Kaufmann Series in Data Management Systems, pages 443–495. Morgan Kaufmann, Boston, third edition edition, 2012.

[19] C. Fraley and A. E. Raftery. How many clusters? which clustering method? answers via model-based cluster analysis. The Computer Journal, 41(8):578–588, 1998.

[20] Vladimir Estivill-Castro and Jianhua Yang. Fast and robust general purpose clustering algorithms. In Riichiro Mizoguchi and John Slaney, editors, PRICAI 2000 Topics in Artificial Intelligence, pages 208–218, Berlin, Heidelberg, 2000. Springer Berlin Heidelberg.

[21] Olcay Akman, Timothy Comar, Daniel Hrozencik, and Josselyn Gonzales. Chapter 11 - data clustering and self-organizing maps in biology. In Raina Robeva and Matthew Macauley, editors, Algebraic and Combinatorial Computational Biology, MSE/Mathematics in Science and Engineering, pages 351–374. Academic Press, 2019.

[22] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. Unsupervised Learning, pages 497–552. Springer US, New York, NY, 2021.

[23] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96, page 226–231. AAAI Press, 1996.

[24] Joerg Sander. Density-Based Clustering, pages 270–273. Springer US, Boston, MA, 2010.

[25] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: Ordering points to identify the clustering structure. SIGMOD Rec., 28(2):49–60, jun 1999.

[26] Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. Density-based clustering based on hierarchical density estimates. In Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu, editors, Advances in Knowledge Discovery and Data Mining, pages 160–172, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[27] Desmond J. Higham, Gabriela Kalna, and Milla Kibble. Spectral clustering and its use in bioinformatics. Journal of Computational and Applied Mathematics, 204(1):25 – 37, 2007.

[28] F. Riaz, F. B. Silva, M. D. Ribeiro, and M. T. Coimbra. Impact of visual features on the segmentation of gastroenterology images using normalized cuts. IEEE Transactions on Biomedical Engineering, Biomedical Engineering, IEEE Transactions on, IEEE Trans. Biomed. Eng, 60(5):1191 – 1201, 2013.

[29] Y. Pan, C. Huang, and D. Wang. Multiview spectral clustering via robust subspace segmentation. IEEE Transactions on Cybernetics, Cybernetics, IEEE Transactions on, IEEE Trans. Cybern, 52(4):2467 – 2476, 2022.

[30] Xiaofeng Zhu, Jiangzhang Gan, Guangquan Lu, Jiaye Li, and Shichao Zhang. Spectral clustering via half-quadratic optimization. World Wide Web: Internet and Web Information Systems, 23(3):1969 – 1988, 2020.

[31] Kim Daeun, Hanzawa Fumiaki, Sun Shumin, Laurent Thomas, Ikeda Saiko, Umeki Miki, Mochizuki Satoshi, and Oda Hiroaki. Delayed meal timing, a breakfast skipping model, increased hepatic lipid accumulation and adipose tissue weight by disintegrating circadian oscillation in rats fed a high-cholesterol diet. Frontiers in Nutrition, 8, 2021.

[32] Andrea Poli, Franca Marangoni, Alberto Corsini, Enzo Manzato, Walter Marrocco, Daniela Martini, Gerardo Medea, and Francesco Visioli. Phytosterols, cholesterol control, and cardiovascular disease. Nutrients, 13(8), 2021.

[33] Charlie G Y Lim, E Shyong Tai, and Rob M van Dam. Replacing dietary carbohydrates and refined grains with different alternatives and risk of cardiovascular diseases in a multi-ethnic asian population. American Journal of Clinical Nutrition, 115(3):854 – 863, 2022.

[34] Jr Romeo B.Batacan, J.Duncan Mitch, J.Dalbo Vincent, L.Buitrago Geraldine, and S.Fenning Andrew. Effect of different intensities of physical activity on cardiometabolic markers and vascular and cardiac function in adult rats fed with a high-fat high-carbohydrate diet. 运动与健康科学（英文版）/ JOURNAL OF SPORT AND HEALTH SCIENCE(JSHS), 7(1):109 – 119, 2018.

[35] Haimoto Hajime, Watanabe Shiho, Maeda Keiko, Murase Takashi, and Wakai Kenji. Reducing carbohydrate from individual sources has differential effects on glycosylated hemoglobin in type 2 diabetes mellitus patients on moderate low- carbohydrate diets. Diabetes and Metabolism Journal (DMJ), 45(3):390, 2021.