DEPARTMENT OF SCIENCE

MTH301 Final Year Project

Unsupervised Learning methods and its Applications on Customer Personality Analysis 无监督学习聚类方法及其在顾客性格分析中的应用

Student name: Xinyi Yue

Student ID: 1823376

Supervisor: Mu He

Date:2/May/2022

Abstract

In recent years, due to the prevalence of Covid-19, the offline retail of various countries has received varying degrees of negative impacts. The Internet, emerging as a new marketing channel, is rising at an unimaginable speed. So understanding the expectations and needs of online customers is regarded as the prerequisite for promoting the development of consumer-oriented electronic commerce markets. In this study, we use the method of clustering algorithm to effectively handle the online shopping market, and find the relationship between differences in customer personal information and shopping preferences. The purpose of this study is to summarize the differences of online shopping preferences among consumers with different information dimensions and carry out visual processing so as to help enterprises modify their products according to different types of customer groups. For example, instead of spending money on promoting new products to each customer in the company's database, the company can carry out targeted and efficient product promotion according to the differences of online shopping preferences. The data set involves the customers' personality (income, education, childhome, recency)and the specific shopping cost in different kinds of products.

Keywords: principal component analysis , hierarchical clustering, k-means clustering, Gaussian mixture model, online shopping

近年来，由于新冠肺炎疫情的流行，各国的线下零售都受到了不同程度的影响。互联网作为一种新的营销渠道正在以难以想象的速度崛起。因此，了解客户的期望和需求是推动以消费者为导向的电子商务市场发展的前提。在本研究中，我们使用聚类算法来有效地处理线上购物，并找到客户个人信息与购物偏好差异之间的关系。本研究的目的是总结不同个人信息维度的消费者在网络购物偏好上的差异，并进行可视化处理，以帮助企业根据不同类型的客户群体对产品进行精准投放和推广。

关键词：主成分分析, 层次聚类, k-means 聚类, 高斯混合模型, 在线购物

Contents

Introduction

Due to the rapid development of the Internet technology, e-commerce is widely used because it can eliminate the limits of in-store shopping, such as low efficiency caused by time and distance.[44] Moreover, the emergence of wireless networks and smart mobile devices provides prerequisites and convenience for online shopping[49]. In 2008, The CNNIC report has showed that there were about 30 percent people using internet for shopping, and now this proportion increased dramatically to more than 40 percent, which meant that the number of people had experience of shopping on line was more than 6500 million[52]. This result is estimated to have tripled in recent years, which means 80 percent people have had the experience shopping online. As a result, more and more traditional companies are focusing on the combination between online stores and offline retail outlets, for example Apple not only has thousands of offline stores, but has online official flagship store with million followers. However, during the critical period of integration of offline and online shopping， an unprecedented Pneumonia outbreak in Wuhan City, Hubei province in China, detected at the end of 2019, which spreads rapidly around the world through human contact and other means[47]. António Guterres, Secretary-General of the United Nations, described the COVID-19 pandemic as the most serious crisis in the last 75 years, and it is spreading human suffering, infecting the global economy and making a negative influence in people's lives[19]. In the face of this rapidly spreading epidemic, most governments restricted social life, ranging from bans on social events from offline education to economy[48]. For example, most retail stores and services has received significant constraints in these two years, which causes growing levels of economic uncertainty but supports the chance to develop the e-commerce[26].

B2C e-commerce, which means business to consumer, let consumers to search for and compare with the product features, reviews, price[24]. Through a series of studies, dividing the motivation of consumers' purchasing into the categories of hedonic and utilitarian[5]. Furthermore, Hirschman and Holbrook[22] held the idea that people seek fun and fantasies in the process of searching products and purchasing. However, fewer and fewer people think of shopping as utilitarian, that is, purposefully searching for what they want to buy and ordering it directly. In addition to purposeful direct search, more and more people choose the products they are interested in for detailed understanding and purchase when browsing the product collection pages. In order to improve the experience of browsing the shopping website and the exposure traffic conversion, more and more recommendation systems are employed[2]. Meanwhile, it is significant to determine the users' preferences when design the automatic recommendation systems[53]. Personality traits, defined as "stable, endogenous, hierarchical basic personality controlled by biological factors such as brain structures and genes", significantly influence the way people feel, think and deal with[7][42]. So in daily life, especially in the shopping situation, the products that customers are interested in are related to the personality of them to an extent, for example, the level of income and education play a role in the price and quality the products people choose; the frequency they look through in the shopping website influences if they purchase purposefully.

A clear understanding of the factors of consumers' purchase of goods not only helps us to understand the overall trend of online shopping market, but also has a deeper understanding of the operation of Internet companies and the macro grasp of specific goods on the Internet. Companies are concentrated on the products consumers purchase and the factors effecting the purchase choice. Because one such factor is the channel of information research, firms understanding the relationship between purchase choice and customer personality will benefit in the business battle[27]. For example, when a product is released, the price of the product can be adjusted in advance, and more targeted to find out the advertising promotion crowd, to maximize the use of 'push'.

To deal with these problems, machine learning can be used to improve the efficiency. The concept of machine learning can be viewed as searching through the large programs, and find the optimal performance metric by training experience[25].The field of machine learning, there are supervised and unsupervised learning can be used in the real problems, which have the differences as follows: in supervised learning, the data set is shown in advance and the correct output which have corresponding relation with the input is also known, and the questions can be divided into 'regression' and 'classification'; while, unsupervised learning is solving problems without knowing what the outcome should look like, without feedback based on the predicted outcome. Among many techniques, the cores of supervised learning and unsupervised learning are different, which is classification and clustering respectively. The original inspiration of these two thoughts is related to the habit that keep similar products together and divide different kinds of things people born with. For example, the supermarkets always put the same kind of products in one shelf, and the same brand of goods are the nearest. While, facing computational problems, we can not tag everything with a label, so 'clustering' is invented to put things without tags into groups, which means to deal with the data samples with unknown types.

Clustering analysis is a statistical method for processing data, which has been addressed in many fields such as data mining, document retrieval and pattern classification[23]. The basic thought is to divide similar individuals into the same group according to the close relationship of each item, so that the properties of individuals are homogeneous and similar. Clustering analysis studies with the variables which are not pre-divided into predictive and standard subsets, of which the purpose is to find groups of similar subjects, with the 'similarity' between each pair of subjects being some overall measure of the entire feature set. As a simple and easy-understanding method, clustering has widely use. For example, facing chaotic data, it can identify specific distributions from them, and better understand the data structure by reducing the quantity of data[43].

A Literature Review

3.1 Factors Influencing Consumer Behavior

There are a variety of factors that determine people's behaviour when they are shopping online. In the process of engaging in a purchase decision, the customers behaviors are influenced in the consumer characteristics such as age, income, education and environmental factors. The consumer characteristics include four major factors, which is culture, social factors, personality and psychology. So it means that if there exists two customers with the same age, gender and social background, but they may be show as different purchase behavior because of the psychological factors[6].

- Cultural: Culture is the fundamental determinant of person's needs and behaviors. In different country and socialization processes, people have different method to use internet to buy goods, and the target of goods are different as well. For example, fast food commodities are more popular in the US than in China because of the differences in diet habit; at the same time, there are many products born out of traditional culture that are only popular in corresponding country.

- Social: Consumer Behaviour is also influenced by family composition, social role and groups. Different occupational needs lead to different commodity needs of customers when they shop online, for example work related to live streaming of Internet celebrities always requires a great demand for cosmetics. Also, the family is also significant because with children, parents may pay more attention on the goods about toys and others.

- Personal: the decisions are majorly influenced by personal traits such as gender, income, educational level and age. There is a significant survey by French socialists shows that women have more emotional involvement than men when shopping, while men are high on efficiency and quality[14]. Women will be more likely to write shopping lists than men. Women prefer items with sentimental value, while men are more likely to choose financial and leisure products[9]. Besides gender, income is also an important effect. Compared with high-income people, low-income shoppers pay more attention on low-priced (and possibly lower quality) food products[31].
- Psychological: Expect the following fixed condition, motivation and mood also effect the shopping behavior. There was once a research showed that the same person would choose products with different complexity of patterns due to the current shopping mood, except that customers have specific preferences for patterns and colors.

3.2 Historical Development of Clustering

Clustering analysis has a rich history, which can be used in various of fields, not only a set of algorithms and tools in statistics, but in taxonomy and epidemiology as well[10]. Because data may have different size and shapes, so data clustering is regarded as a challenge in unsupervised pattern. It is a significant process in machine learning and pattern recognition. Moreover, As a process of identifying natural groupings in multiple dimensions based on data similarity, data clustering play an important role in Artificial Intelligence[20]. Clustering algorithms are used in many applications, such as machine learning, color image quantization, compression, image segmentation, data mining, etc[11]. The development can also be viewed as an interdisciplinary endeavour. According to JSTOR [12], the method of clustering first appeared in the title of an article about anthropological sources in 1954. Moreover, the article written by Saxena et al[45]. first defined clustering as a technique which groups objects according to the inherent similarity among them. This simple definition illustrate the ultimate goal of clustering: to divide an unlabeled data set into a set of natural data structures, which means using the same method grouping the objects to make them more related and similar[34]. Being viewed as mixtures of multivariate normal populations, clustering is a method of conceptualising a ’mixture’ of aggregates in a data set that can be used for sampling[8]. And this method represents a logical extension of normal distribution. While Milligan[38] used the data following truncated multivariate normal distributions and found that clusters did not overlap. In the development of clustering, Cormack[13] combined the concept of distinct groups and defined natural clusters, which should exhibit both the internal cohesion and external isolation. However, the concept of natural clusters have the limitation in some applications. If the elements in the given cluster are required to have the strong similarity to each other, then the clusters are less elongated and compact instead. Cluster connections or continuous connections can be implemented after the definition of compact clusters is modified. For example, such elongated clusters can occur if the clusters follow regression effects between variables. At the same time, in some cases, overlapping clusters make more sense. Therefore, when trying to perform cluster analysis on empirical data sets, researchers have to solve the problem of defining the concept of clustering because different clustering algorithms try to find different kinds of clustering.

As a major method in different scientific disciplines, thousands of clustering algorithms have been proposed in the published literature. All kind of algorithms can be broadly divided into two groups: partitional and hierarchical. The hierarchical clustering algorithm repeats the cycle of combining smaller groups into larger groups or dividing larger groups into smaller one to produce a ”dendogram” showing the arrangement of clusters[16]. As the most popular clustering algorithms, hierarchical methods begin with entities which can be regarded as separate clusters, and they merge at successive level in the clustering, until only one cluster, containing the entire data set. The partitions represent non-overlapping clusters and hold the property that once two elements emerge together, they are never again separated. It is possible to distinguish between different layering methods by identifying the two clusters that merge each level. The survey in 1967 showed that many agglomerative hierarchical methods obey one common recurrence formula[29].

Partitional clustering algorithms generate different partitions, which are also called non-hierarchies because there are k mutually exclusive clusters containing corresponding data. The partitional methods can be distinguished by five characteristics. This first one is about the selection of the seed points. Some K-means algorithms select the elements as starting partitions randomly, while others can allow the user to specify the start partition[3]. The second and third characteristics deal with the types of cluster assignments by assigning points to clusters through statistical criteria. Some K-means algorithms pass by assigning points in turn to the center of mass of the cluster, while others assign and update the center of mass multiple times[35].The two remaining features tell about whether the number of clusters are fixed or variable in the final stage, and the final treatment of outliers in the solution. Only a few methods, such as isodata, can form a variable number of clusters in the solution[39]. Moreover, as the simplest and most basis partitional algorithm, K-means was discovered as an independent algorithms using in different scientific field in 1995 by Steinhaus. In the following sixty years, because of the simplicity and efficiency, K-means is still the most wildly-used one.

3.3 Distance Function

Cluster analysis is essentially to classify and divide the similar points in the table. The first step is to define a distance function / metric on data. There are three commonly used distance functions, considering 2 points.

- Euclidean distance

$$d_2(x,y) = \|x-y\|_2 = \sqrt{(x_1-y_1)^2+\cdots+(x_d-y_d)^2} \tag{3.1}$$

- Manhattan distance

$$d_1(x,y) = \|x-y\|_1 = |x_1-y_1|+\cdots+|x_d-y_d| \tag{3.2}$$

- Cosine distance

$$d_p(x,y) = \sqrt{\frac{1}{2}(1-p[x,y])} \tag{3.3}$$

where p[x,y] is the Pearson correlation

Methodology

4.1 Principal Component Analysis

As an important dimension reduction method, PCA is to identify the most important aspects of the data, and use them to replace the original data [32]. Multi-indicator problems, which is necessary to use multivariate statistical analysis, have many practical significance in the real life. The ideal result is to decrease the complexity of problems, and let less indicators reflect more information[46]. Based on all the original indicators, principal component analysis is to establish as few indicators as possible, so that the correlation of the original indicator can be eliminated. To avoid information loss, the information reflected by the new indicators should keep all the original information[21]. Because of the ease of operation and understanding, principal component analysis can be used in a variety of scenarios, two of the common environment are

following: 1. to visualize. When using clustering analysis, the characters of samples can always be shown in the chart. One of the problems is how to map data from a higher dimensional space to a lower dimensional space, because it is difficult to imagine the image of four dimensions. 2. to select feature. If a sample is combined with 10 dimension characteristics, and some of them are 'noise' (just like volume, surface can both be calculated by the characteristic 'radius'), PCA can be used to reduce the noise and keep useful and basic features.
If the number of samples is m, with n features, then it can be viewed as a matrix with $m * n$ dimension. The basic theory of PCA is to find a new direction that can represent multiple original dimensions so that the dimension can reduce from n to n'. There are two thoughts to determine the 'line': The first one is that the sample points are close enough to this line, and the second is the projections of the sample points on this line are as far apart as possible. And it is similar with the derivation of PCA.

4.1.1 Derivation of PCA

- **Based on the minimum projection distance**

  It can simplify as the minimum distance between samples and the obtained line segment or plane. After being centralized, the new coordinates translate into $\{w_1, w_2, ..., w_n\}$, in which 'w' is standard orthogonal basis. If reduce the dimension from n to n', then the new coordinate is $\{w_1, w_2, ..., w'_n\}$, and the projection in n prime coordinates is $z^i = (z^i_1, z^i_2, ..., z'^i_n)^T$, where $z^i_j = w^T_j * x^i$. Then the recovery data is

  $$\overline{x}^i = \sum_{j=1}^{n'} z^i_j * w_j = Wz^i$$

  where W is the matrix of orthonormal bases. So in order to let all samples are close enough to the hyperplane, which means to minimize

  $$\sum_{i=1}^{m} \|\overline{x}^i - x^i\|_2^2$$

  Then this equation can reduce as

  $$-tr(W^T XX^T W) + \sum_{i=1}^{m} x^{iT} x^i$$

  where $\sum_{i=1}^{m} x^{iT} x^i$ is the covariance matrix of the original data set, so we can minimize the equation $-tr(W^T XX^T W)$, which can use lagrangian function. As a result,the dimensions of the original data set can reduce to n' using 'the smallest projected distance' if $z^i = W^T x^i$

- **Based on the maximum projection variance**

  The definitions of x, zW are all same as the above one, and any sample x^i has the projection $W^T x^i$ in the new coordinate, and the projection variance is

  $$x^{iT} W W^T x^i$$

  In order to maximize projection variance, to find

  $$argmaxtr(W^T X X^T W) s.t. W^T W = 1$$

  and use the lagrangian function, $J(w) = TR(W^T X X^T W + \lambda(W^T W - I))$, where λ is combined with the eigenvalues of XX^T. so we need to find all the eigenvectors corresponding to the largest n' eigenvalues and combine them into the matrix W (the same as $z^i = W^T x^i$)

4.1.2 Code Representation

It is easy to have PCA by R, because there are $prcomp()$ and $princomp()$ that can return results directly.

- **Input data**

  firstly, to simulate the data, and can use $plot()$ to determine the distribution of data set. And in this process, the code is similar to the code in K-means.

```
set.seed(1995)
#Random number is the number taken from the seed of random number.
    A seed is a serial number
data=matrix(abs(round(rnorm(a, mean=b, sd=c))), 10,
    10)
#Random positive integers generate matrices
#abs() means the absolute value
#round() means round numbers to the nearest possibility
#rnorm() means generate numbers randomly based on normal
    distribution
#the matrix is 10*10
colnames(data)=paste()
#determine the column
rownames(data)=paste()
#determine the row
```

- **Data standardization**

  In order to unify the dimension and centralize the data, it is necessary to standardize the raw data before principal component analysis, and sometimes z-score standardization can be used. This method minus the mean divided by the standard deviation of the data, so that let the data fits the standard normal distribution.

```
    data2=scale(data, center=T, scale=T)
    #center means minus the means
    #scale(normalize) means dividing the standard deviation on the basic of centralized
```

- **PCA**

  1. using prcomp()

```
    data.pca <- prcomp(data, center=F, scale=F)
    #on the step 2, the data is nomalized, so 'scale = F' means no need to repeat data standardization
    summary(data.pca)
    #return the result of PCA
```

  2. using princomp()
  'princomp()' only works with matrices with more rows than columns, and the method can divide into Standard deviation、Proportion of Variance、Cumulative Proportion, all of which have the same results

```
    data2.pca <- princomp(data,cor = T, scale=F)
    #If WE want to use correlation coefficient matrix as the source of processing, set cor to TRUE
    #cor=False means calculating using covariance
    summary(data2.pca)
```

- **the difference between prcomp() princomp()**

  1. prcomp() is applicable to both R mode and Q mode, princomp() is applicable to R mode
  R mode means investigating the relationship of variables in all observations, such as which one is typical and basic; Q mode is the analysis about observation, which can be viewed as investigating the relationship of observations in the variable, such as Which two records are similar.
  2. prcomp() is based one the eigens from covariance and correlation ('cor = T & F' are used in the algorithm), and princomp() is based on SVD methods.
  SVD is the short for Singular Value Decomposition, and the definition of the SVD of matrix A is $A = U^T$ and U is a $m*m$ matrix that is combined with all eigenvectors of AA^T, and V $(n*n)$is from the eigenvectors of A^TA, $\sum$ is a $m*n$ matrix with only singular values on diagonal.And SVD is a significant method

that widely used for for feature decomposition in dimensionality reduction algorithms.

4.2 K-means Clustering

K-means clustering was first published in 1955 by Macqueen [33] and has been regarded as the most popular clustering algorithm since then. This algorithm uses the distance of each cluster centroid to divide the cluster similarity [55], so that each individual in the dataset belongs to only one cluster with similar attributes.

Assuming $D = \{X_1, X_2, \ldots, X_n\}$, each object has attributes of m dimensions, in order to aggregate n objects into k specified clusters according to the similarity. Each object belongs to and only belongs to a class cluster whose distance from the center of the class cluster is the smallest.

Algorithm execution steps:

1. Select K points as cluster centers of initial aggregation

2. Using Euclidean distance/equation (1) to describe the distance between the objectives and divide each point into different clusters according to distance

3. Then repeat 1&2, to minimize the sum of the squared error of all the K clusters

$$\sum_{k=1}^{k} \sum_{x_i \in c_k} \|x_i - \mu_i\|^2 \tag{4.1}$$

4. As the addition of new points causes a change in the cluster centroid, different centres of mass need to be recalculated. When k new centre of masses are found, a new binding should be created between the same data point and the nearest new centroid, generating a loop. As a result of this cycle, the K centroid may gradually change its position until it remains unchanged, which is the end of the algorithm[40] .

Code implementation:

```
input D={X_1:X_n};
for t = (1:T)
for every x_i;
calculate dist(x_i, center k);
#center k is the cluster centers of initial aggregation
divide x_i into the least distance clusters;
end for
Update the cluster center
calculate the distance
calculate the difference between two iterations a
```

```
if a < error
#which is the given minimum acceptable error
output
break;
end if
end for
```

4.2.1 Code Representation

If the process is divided into different steps for code representation:

- **Read Data**

  Firstly, save the data file as : comma separated value(.csv) or tab delimited text file(.txt), which contains numerical values split by commas, we should read the data from this EXCEL, and save the data into a list. every element in the new list can be seen as another list containing the value of a feature item.

```
    data1 <- read.csv(file.choose(),header=T)
    #header=T means that the first row is used as the column name for the Excel data substituted in, and the concrete data starts at the second row
```

```
    data2 <- read.table(file.choose(),header=T, sep=",")
    #In the input, the original content is separated by ",", so as to read the data more correct.
```

- **Euclidean Distance**

  Because of the definition of the k-means method, we should calculate the distance so that to divide the points into different classes. And euclidean distance can be used as a metric of similarity.

```
     mydata <-rbind(a,b,c)
     #After converting the categories of each individual into vector form, a data matrix can be formed.
     mydatas <-scale(mydata)
     #Standardize the data
     mydist <-dist(mydatas,method="euclidean")
```

- **Initialize Means**

  Because the aim of k-means clustering is to converge on an optimal class of cluster centers so that to minimize the distance, it is clear that the iteration of the k-means clustering algorithms and the strength of convergence both depend on the positioning of the initial centroids[15], so that I use the method to find the min and max of each feature, and then to initialize each mean’ s feature values randomly.

```
        set.seed()
        #Set seed of random number to ensure repeatable experiment
        #the number in () is the number of k
        km_result <- kmeans(df, a, nstart = 25)
        #K-mean is used for clustering
        #”a” means divide into k groups
        #”nstart” means that randomly starts k-means algorithm n times and
            return the best. usually take nstart=20 or 25 unless you have very
            big dataset.
```

- **Visualization of clustering results**

  After confirming the initial point and calculating the Euclidean distance, we can use plot to visualize the result and output it.

```
        plot(x, col = cl$cluster)
        #”clcluster”meanstoshowtheclassesintowhichalldataisgrouped
        points()
```

4.2.2 K - means Optimization

When use K-means algorithm, K should be determined, which is very difficult to estimate and calculate. So it is hard to set the number of appropriate categories in advance; Also, in k-means algorithm, it is necessary to know an initial partition based on the initial clustering center, and then change the partition by calculating the distance, so the whole calculation depends on the selection of the initial clustering center[51]. Finally, the whole progress needs to adjust the cluster classification and the new clustering center constantly, so facing large number of data, the algorithm’s time cost is very large.

- **K-means++**:

  K-means++ method defines a specific way to determine the new initial clustering point when merging. After chose a sample as the initial clustering center randomly fro the data set (C_1), then calculate the distance between every points and corresponding C_1, and the distance is D_x, then the next clustering center is chosen with the probability $D(x)^2/\overset{x}{\sum} D(x)^2$ [4].

- **ISODATA**:
  ISODATA is short for 'iterative self-organizing data analysis'. In ISODATA method, means should be recalculated and the clusters should be reclassified after every iteration. Comparing with K-means method in which the number of clusters K remains the same throughout the whole process, ISODATA saves the time significantly [1].

- **Kernel K-means**:
  In k-means, Euclidean distance function is appropriate to describe the similarity. The drawback is that it is not suitable in all situation. Kernel K-means combine it with the vector, and map all samples to another feature space for clustering.

4.3 Hierarchical Clustering

Hierarchical clustering is a kind of cluster analysis creating clusters that there is a primary and secondary order from top to bottom. Different from K-means that require us to pre-specify the number of clusters at first. Hierarchical clustering initially combines individuals based on the similarity to form n clusters, and then continuously further classifies the clusters formed by the combination until they become a whole. According to the difference of initial objects and clustering order, hierarchical can divide into agglomerative clustering and divisive clustering. As a bottom-up method, the idea of agglomerative clustering is to merge clusters with the smallest intergroup difference into one cluster, and the process ends until all clusters become a large cluster after multiple mergers. contrary to the agglomerative clustering, divisive clustering is to split the known cluster with the largest intergroup dissimilarity. Furthermore, either method can finally show an attractive tree-based representation, which clearly expresses all kinds of associations. This "tree" is called dendrogram. The process of this algorithm is similar with the storage method of files and folders in computers, and the information with different similarity is classified and processed through continuous integration and segmentation.

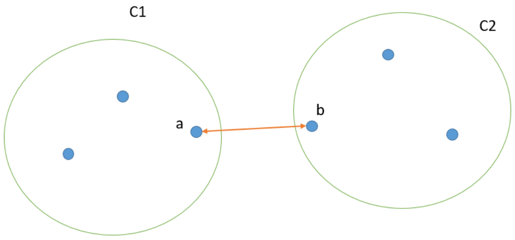
When judging the similarity of clusters, we need to use the distance function, and in hierarchical clustering analysis, there are different methods to define the similarity in clusters depending on the distance between between individuals in each group. The definitions of three common clustering methods are given below:

4.3.1 Hierarchical Clustering Methods

- **Single linkage clustering:**
  One of the simplest and most common agglomerative hierarchical clustering methods. The distance between different groups is equal to the minimum distance between the closest group of samples.

$$D(C_1, C_2) = Min\{d(a,b)\} \tag{4.2}$$

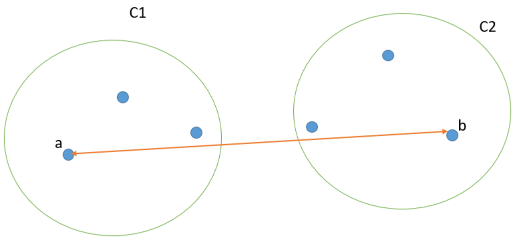
  At each stage, the clusters can be merged due to the minimum distance da, b, and the new formed cluster will have minimum pairwise distances. Repeating, all samples can merge into a final cluster according to the similarity.



- **Complete linkage clustering:**
  Different from the single linkage clustering, Complete linkage clustering defines the distance between every clusters as the most distant pair of samples.

$$D(C_1, C_2) = Max\{d(a,b)\} \tag{4.3}$$

  After using the distance function to compute the distance between every possible object pair, the distance between two clusters can be obtained. And at each stage, find the minimum distance between every clusters, and merge them together.



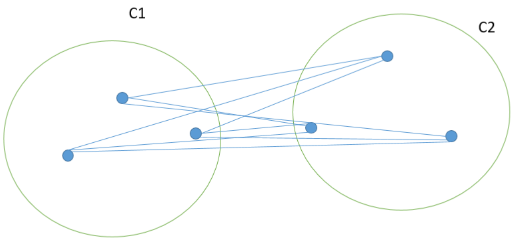
- **Average linkage clustering:**
  The disadvantage of the following two methods is that there is a large error in

comparison of similarity, so the distances between two clusters can be defined as the average of distances between all samples in these corresponding class.

And when compare the regional global similarity, the error is smaller.

$$D(C_1, C_2) = T_c/(N_1 * N_2) \tag{4.4}$$

T_c means the sum of the distances corresponding to points in regions C_1 and C_2; N_1 &N_2 means the size of C_1 and C_2.



4.3.2 Divisive Approach & Agglomerative Method

Because the similarity of each cluster is different, the whole process needs to be repeated continuously. Unlike the divisive clustering, the agglomerative method can be repeated all the time in the ideal state. Therefore, generally, there will be restrictions on the number of clusters or errors to terminate the cycle.

- **Divisive approach**: Firstly, calculate all distance between every samples in the known clusters, and find out a and b with the farthest distance
  Secondly, allocate a and b to different class clusters C1 and C2, and calculate dis(other samples, a) and dis(other samples, b) , if dis(a)<dis(b), then assign the sample points to C1, otherwise, assign the sample points to C2

- **Agglomerative method**:

  Firstly, calculate the distance between two types of clusters, and find c1 and c2 with the minimum distance(which means the largest similarity).
  Then merge c1 and c2 into a new cluster, and repeat the operation until reach the number of clusters.

4.3.3 Code Representation

If the process is divided into different steps for code representation:

- **Read data & Data normalized & Distance fuction**
  Same to the K-means methods, firstly, we should read data and calculate the distance(usually use ”Euclidean Distance” method Also, in the process of data mining, in order to eliminate the influence of dimension, data need to be normalized.

- **Process data set**
  Usually, agglomerative hierarchical clustering could be used, and the idea is to constantly merge clusters by comparing their similarity (using distance), in order to control the error, average linkage clustering is appropriate.

```
hc <- hclust(dist(customer,method = "euclidean"),
    method = "average")
#'hclust()' is the fuction of hierarchical clustering
#'euclidean' is the method to describe the distance between clusters
    (similarity)
#'average' is cluster method, including:'ward.D2','ward','centroid'....
plot(hc, hang=-1, cex=.8, main="Average Linkage
    Clustering")
#Generate the dendogram
```

4.4 Model-Based Clustering

Model-based clustering is a significant method in the application of cluster analysis to fit the data set by formulating the probabilistic model. Compared with heuristic clustering methods, it is more explicit [37]. Model-based clustering means using the mixture models to perform clustering and is widely used in different fields[18].
Let **X** be the N * J data matrix, where each row $x_i = (x_{i1},...,x_{ij})$ is the realization of a J-dimensional vector of random variables $\mathbf{X} = (X_1, X_2,....,X_J)$. Also assuming that each observation comes from a mixture of finite G-probability distributions, each of them represents a different cluster or group. The form of the mixture distribution is as follows:

$$p(x_i;\theta) = \sum_{g=1}^{G} \pi_g p(x_i, \theta_g) \tag{4.5}$$

where the π_g is the mixing probability, which means that $\sum_{g=1}^{G} \pi_g = 1$ and θ_g is the parameter set of g.
Then assume z_i =(z_{i1},...,z_{iG}) as the component membership of observation i, so $z_{ig} = 1$ if i belongs to component g and $z_{ig} = 0$ otherwise. After parameter estimation, each

observation can be assigned to the different clusters by using the maximum a posteriori rule in statics[37]. So as a result, the observing cluster z can be shown as follows:

$$\hat{z}_{ig} = \frac{\hat{\pi}_g p(x_i|\hat{\theta}_g}{\sum_{g=1}^{G} \hat{\pi}_h p(x_i|\hat{\theta}_h)} \tag{4.6}$$

The most popular distributions are Multinomial and Gaussian. However, more flexible distribution hypothesis can be appointed, allowing for heavy tails, skewness and different data types[17]. For example, non-Gaussian components[37], multivariate t-Student distributions[36], skew-t and skew normal distributions[30].

4.4.1 Gaussian Mixture Model

With the basic of maximum likelihood estimation and single-Gaussian mixture model, we can find that the expression of Gaussian probability density function is

$$N(x; u, \sum) = \frac{1}{\sqrt{2\pi|\sum|}} exp[-\frac{1}{2}(x-u)^T \sum^{-1}(x-u)] \tag{4.7}$$

Also the Gaussion Mixture Model (GMM) can be presented as

$$p(x) = \sum_{k=1}^{K} \pi_k N(x; u_k, \sum_k) \tag{4.8}$$

in which π_k is the weight factor, and $\sum_{k=1}^{K} \pi_k = 1$. The Gaussian mixture model (GMM) can be seen as an optimisation of the k-means model, which attempts to find a mixture representation of the probability distribution of a multidimensional Gaussian model and thus fit an arbitrarily shaped data distribution. GMMs are widely used in the daily life and industry. For example, in biometrics systems, most significantly in the systems of voice recognition, because it is able to represent large sample distributions. GMM is a mixture of two models using discrete sets of Gaussian functions, and the mean and the covariance matrix are allowed to a more efficient modeling capability.

In biometrics, the use of GMM to represent feature distributions may be based on the direct idea that individual component density may have underlying relationships too. For example, in the field of applications to speaker recognition, one can assume that the acoustic space of features associated with the spectrum corresponds to a wide range of events, such as the vowel, nasal or fricative sounds of a speaker. These acoustic categories reflect the vocal tract configuration that speakers generally rely on, which can greatly improve the accuracy of identifying a speaker. For example, the spectral shape of the i^th acoustic class can be represented by the average u_i of the density of the i^th component, and the covariance matrix $\sum_i$ can show the change of the the mean spectral shape. Because all the features used to train GMM are unlabeled, acoustic classification in an observed category is unknown[41].

4.4.2 EM Algorithms of Gaussian Mixture Model

EM clustering algorithm is a general method of finding the Maximum Likelihood Estimation. Consider the data set $X_1, X_2, ..., X_n$ as a random sample of size n form the d-variate mixture model.

$$p(x;\alpha,\theta) = \sum_{k=1}^{c} \alpha_k f(x,\theta_k) \tag{4.9}$$

and $\alpha_k > 0$ means the mixing proportions with the constraint $\sum_{k=1}^{c} \alpha_k = 1$. Also $f(x,\theta_k)$ is the density of x with parametersθ_k. Then we can assume the value of matrix Z_{ki} which is a binomial distribution. The log likelihood function is obtained as follows:

$$L(\alpha,\theta;x_1,...x_n;Z_1,...Z_n) = \sum_{i=1}^{n}\sum_{k=1}^{c} z_{ki} ln[\alpha_k f(x_i;\theta_k] \tag{4.10}$$

To use EM algorithm solving GMM, it can be devided into two steps: E-step and M-step. E-step is the process of calculating the rough values of the estimated parameters, and M-step is the process of maximizing the likelihood function based on the results before[54].

$$P(x;\pi,u,\sum) = \sum_{k=1}^{K} \pi_k N(x;u_k,\sum_k) \tag{4.11}$$

- E-step:

  In the previous section, z$_kisunknownandZistheimplicitvariable.Inthispart,Assumingthatthem$

  $$\gamma(i,k) = \alpha_k P(z_k;x_i;\pi,u,\sum) \tag{4.12}$$

  The conditional expected value can be substituted, and based on the Baye's Theorem, we have

  $$\gamma(i,k) = \frac{\pi_k N(x_i;u_k,\sum_k)}{\sum_{j=1}^{K} \pi_j N(x_j;u_j,\sum_j)} \tag{4.13}$$

- M-step:

  Under the constraint $\sum_{k=1}^{c} \alpha_k = 1$, to maximize

  $$L(\alpha,\theta;x_1,...x_n;Z_1,...Z_n) = \sum_{i=1}^{n}\sum_{k=1}^{c} z_{ki} ln[\alpha_k f(x_i;\theta_k] \tag{4.14}$$

  Then the updated equation for mixing proportions can be calculated

  $$\pi_k = \frac{N_k}{N} \tag{4.15}$$

  Then use the parameter θ_k with the mean vector u_k and the covariance matrix $\sum_k$, the updated equations of parameters are follows:

  $$N_k = \sum_{i=1}^{n} \gamma(i,k) \tag{4.16}$$

$$u_k = \frac{1}{N_K}\sum_{i=1}^{n}\gamma(i,k)x_i \tag{4.17}$$

$$\sum_k = \frac{1}{N_K}\sum_{i=1}^{n}\gamma(i,k)(x_i - u_k)(x_i - u_k)^T \tag{4.18}$$

Thus, make sure that both the parameters and the log-likelihood function are convergent, if not, return and calculate again.

Results and Discussions

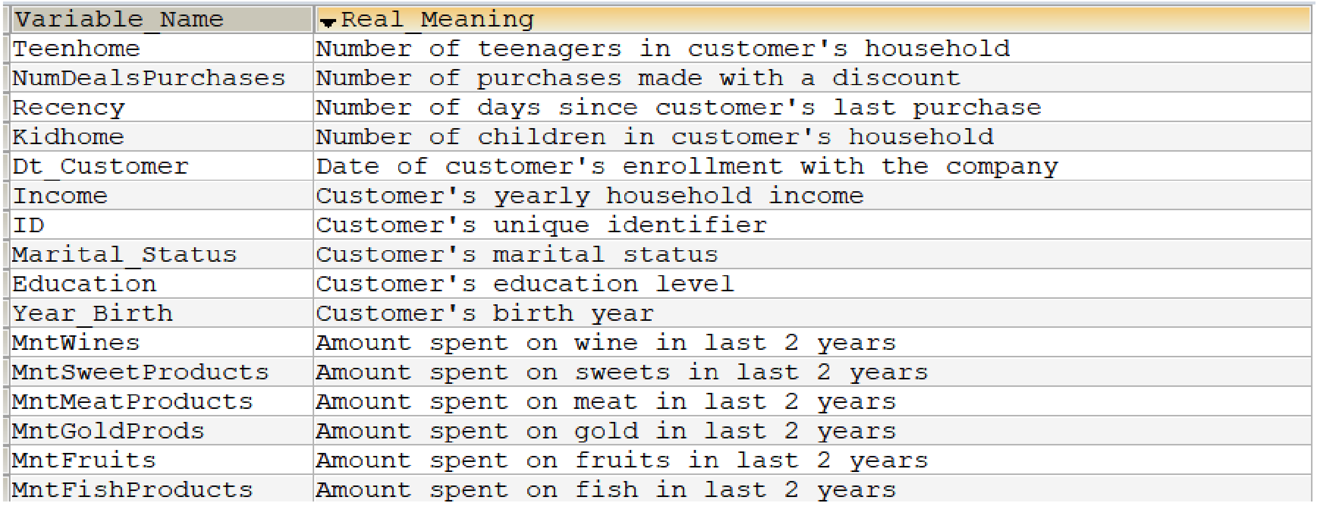
Shopping helps to drive the economy to some extent in 21st century. Some reports use principal component analysis and cluster analysis of customer personality to describe shopping behaviour [50]. In this part, all the techniques would be used to solve a real problem. By using PCA to reduce the dimension of data set, then use the method of clustering and the result of relationship between shopping action and customer personality will be shown as diagrams. Also, the corresponding description will be made based on the results. R studio is dominant used to generate and visualize results.

5.1 Description of Dataset

In order to find the general laws of the relationship, I found the data set which includes more than two thousand people's manners in online shopping (2240 in total) and their personal information. The 'Education' and 'Marital Status' are described as text, others are all numbers. The variables can be divided into 4 parts.

- personal information,including income, education level, the number of kids in the family...
- the choice of products in 2 years, including wine, fruits, meat...
- promotion using 0 and 1 to describe if customers accept the offer in the campaign
- purchase way, including through the website catalogue or directly in the official website...

In this article, we simplify the information and only explore the relationship between customer personality and the amount of money customers spend on different types of goods. To make it easier to understand, the variable names and actual names have been organized in the following table.



Variable_Name	Real_Meaning
Teenhome	Number of teenagers in customer's household
NumDealsPurchases	Number of purchases made with a discount
Recency	Number of days since customer's last purchase
Kidhome	Number of children in customer's household
Dt_Customer	Date of customer's enrollment with the company
Income	Customer's yearly household income
ID	Customer's unique identifier
Marital_Status	Customer's marital status
Education	Customer's education level
Year_Birth	Customer's birth year
MntWines	Amount spent on wine in last 2 years
MntSweetProducts	Amount spent on sweets in last 2 years
MntMeatProducts	Amount spent on meat in last 2 years
MntGoldProds	Amount spent on gold in last 2 years
MntFruits	Amount spent on fruits in last 2 years
MntFishProducts	Amount spent on fish in last 2 years

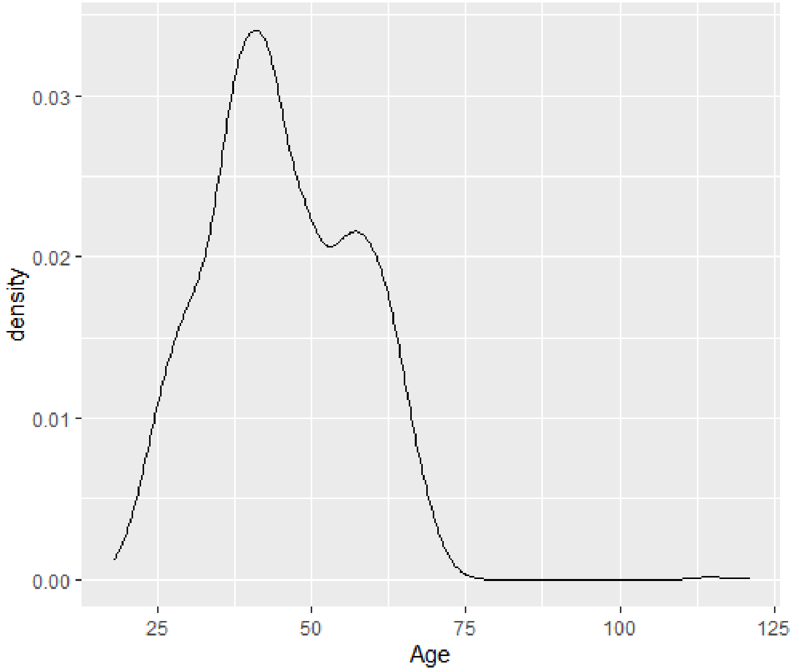
5.2 Description Of Variables

Because the large sample size reduces the relative error, we can view this data set as a microcosm of the entire Internet shopping crowd. In this part, we can find the average situation of people shopping on the Internet through individual analysis, and at the same time, we can use simple cluster analysis to find the relationship between these personal conditions and customers' purchasing situation.

In this part, I will analyze five pieces of information separately, which are education, income, age, and the situation about marriage and children and kids in family.

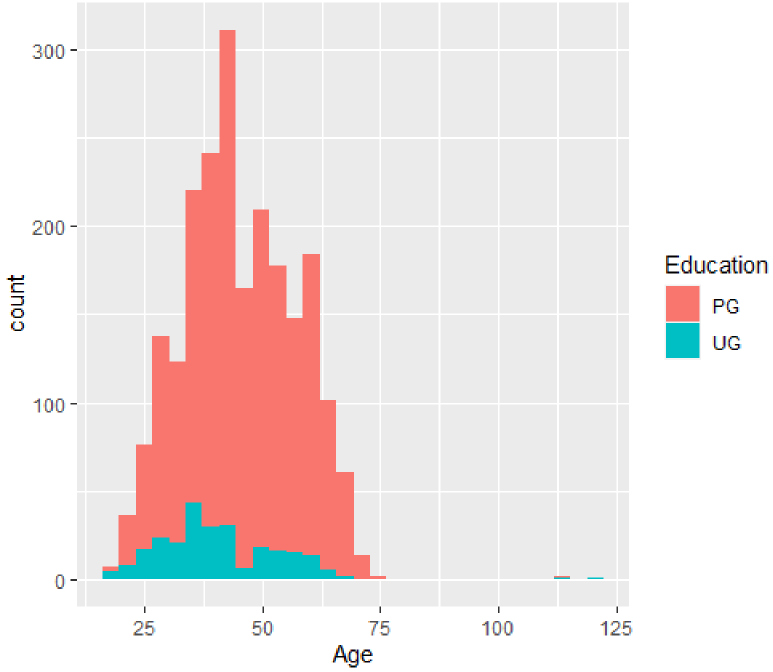
5.2.1 Age

Because the data set only record the birth, so it is necessary to calculate the age in order to find if it is an essential factor in surveying shopping behaviour. Although, the density of age in this study which has more than 2000 customers can also be regarded as the age distribution of shoppers in real life. The chart shows that most people purchasing online is between 25 and 40. The need and habit of shopping online has increased dramatically as people get older, and the density of people in forties is more than other ages.



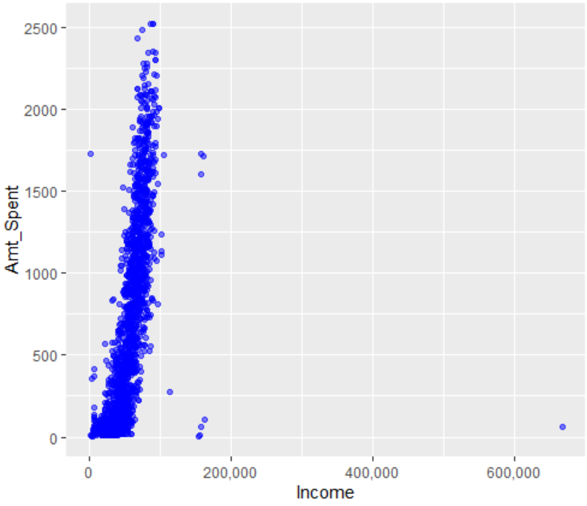
5.2.2 Education

In the data set, there are five categories of education for all people: 2n cycle, basic, graduation, master, PhD. And I group them into two broad categories, namely PG, which includes ”graduation”, ”master”, ”PhD”, and ”UG” with ”2n cycle” and ”basic”. Because the situation of education is fixed, which leads the histogram having less reference value. So I combine the age with education situation, so that in the following table, not only the relationship between education and count, but the relationship between age and count can be shown.
According to 5.3, people with high education cost much more than people who have 2n cycle and basic degree. Also, similar with the density of customers' age, people between 40 and 50 cost the most in online shopping.



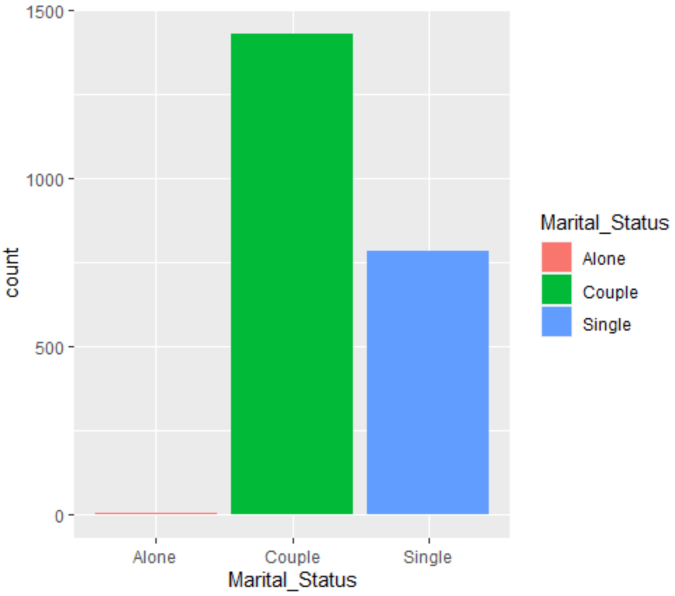
5.2.3 Income

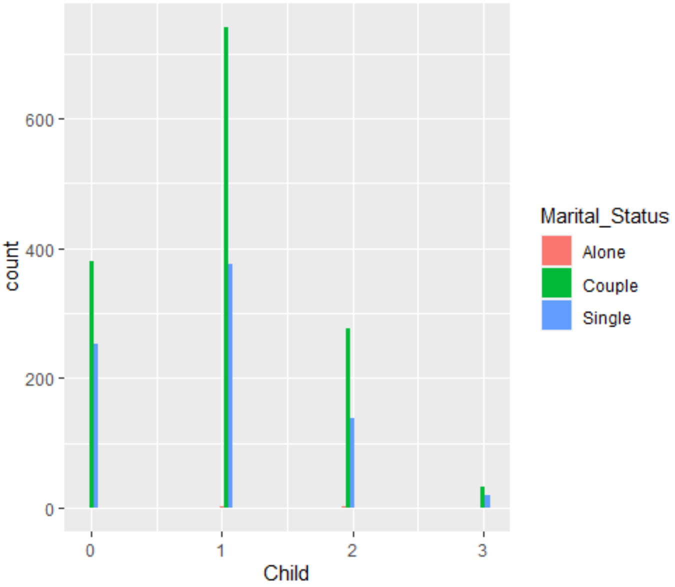
The relation between income and cost is shown as dot figure. And it is clear that more and more money will be spent on online shopping as income increase. This results also confirm the research on economic situation and shopping put forward by sociologists in 2016 Economic constraints were paramount for low-income householders, so that they always find ways to purchase food at the lowest possible cost [28]. Conversely, people on higher incomes will focus more on using online shopping to improve their quality of life, so they will spend much more.



5.2.4 Family Situation

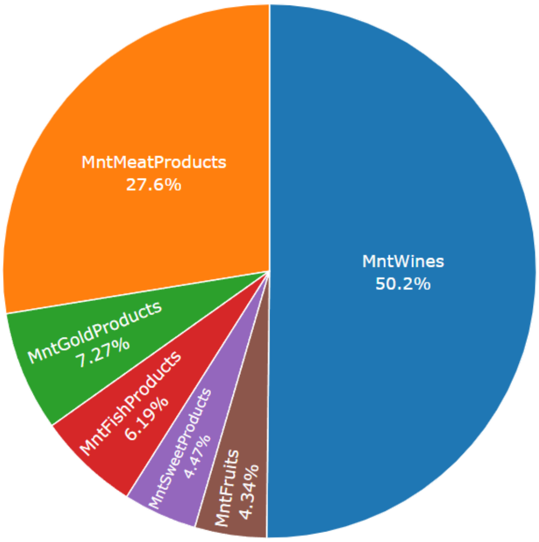
Family situation includes the marital status and kids number in family. The marital status includes alone, single and couple. People's status will affect the demand for online shopping, for example, having a child will increase the demand for baby products than single people. The number of each marital status in these 2000 people will be shown. Families with 1, 2 and 3 kids are also counted. In particular, families with children were more finely divided into marital status, which can break the data set down into nine groups. The marital status and fertility in the data set are presented in the form of bar charts, after which it is easy to find the influence of family factors on the propensity to go shopping online.

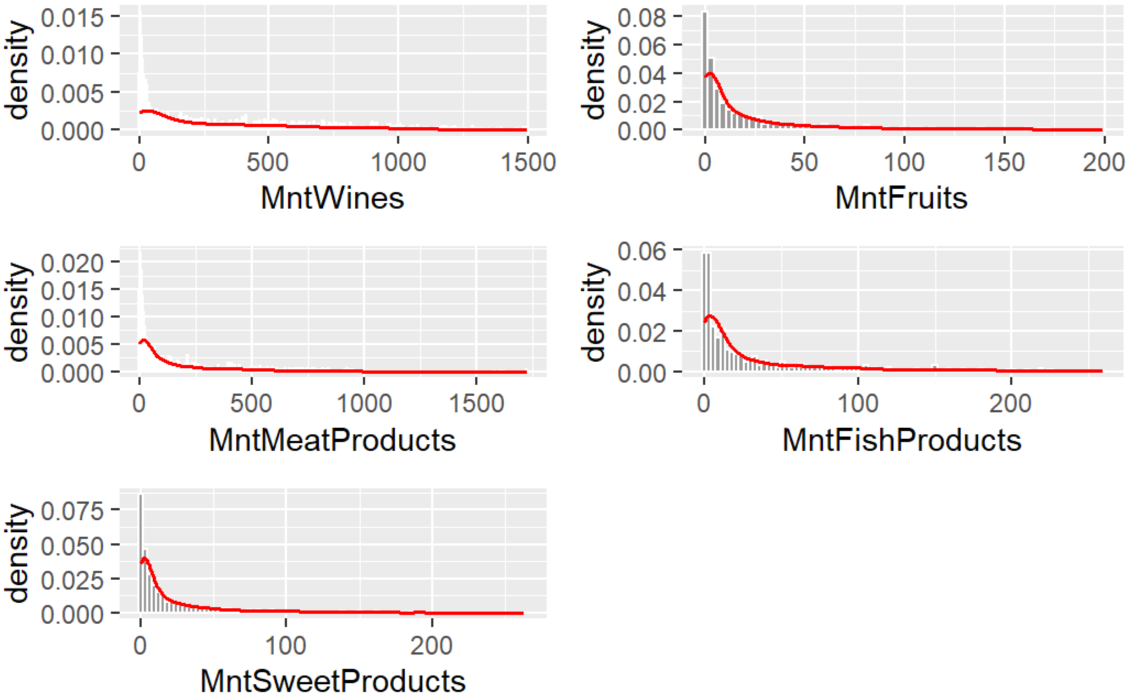




5.2.5 Commodity Category Preference

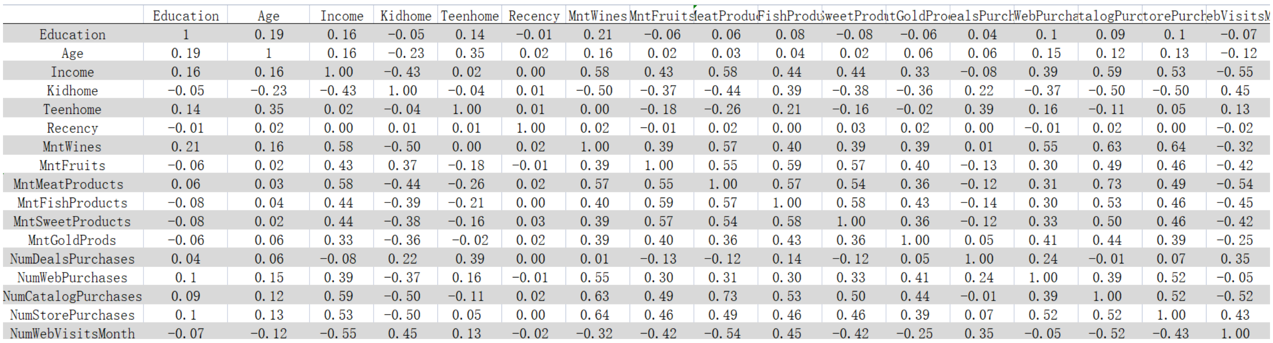
In the data set, the cost of five kinds of food by every interviewees is counted which are wine, fruits, fish, meat, sweet. And according to the maximum cost in each kind of food, each category is divided into different sections with different measure, for example the cost of fish products only be divided into trisection with difference in 100, while some customers bought more than 1500 of meat. We find that people spent the most money on meat and wine in average, and as the total cost increases, the density is decrease gently.



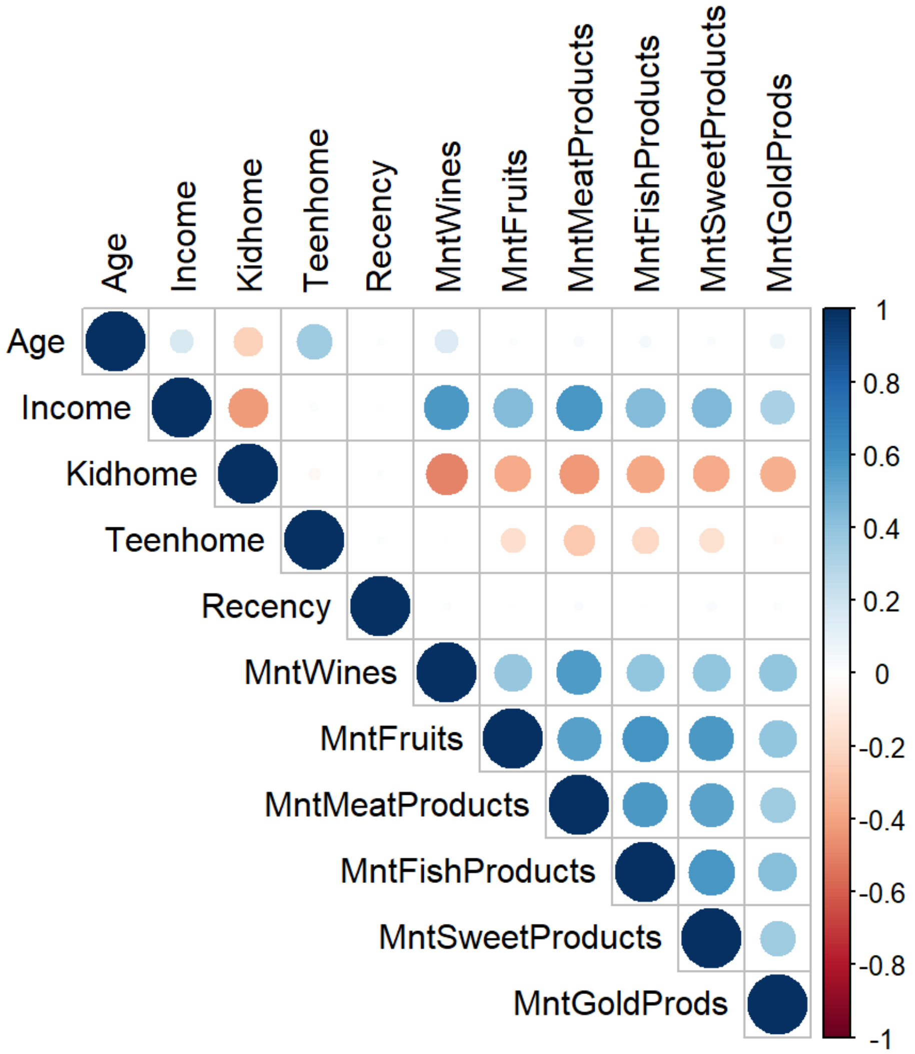


5.2.6 Correlation

We use R to visualize the correlation between the data and represent it with a table and an image.



	Education	Age	Income	Kidhome	Teenhome	Recency	MntWines	MntFruits	leatProdu	FishProdu	weetProdu	tGoldPro	ealsPurch	WebPurcha	talogPurc	torePurch	ebVisitsM
Education	1	0.19	0.16	-0.05	0.14	-0.01	0.21	-0.06	0.06	0.08	-0.08	-0.06	0.04	0.1	0.09	0.1	-0.07
Age	0.19	1	0.16	-0.23	0.35	0.02	0.16	0.02	0.03	0.04	0.02	0.06	0.06	0.15	0.12	0.13	-0.12
Income	0.16	0.16	1.00	-0.43	0.02	0.00	0.58	0.43	0.58	0.44	0.44	0.33	-0.08	0.39	0.59	0.53	-0.55
Kidhome	-0.05	-0.23	-0.43	1.00	-0.04	0.01	-0.50	-0.37	-0.44	0.39	-0.38	-0.36	0.22	-0.37	-0.50	-0.50	0.45
Teenhome	0.14	0.35	0.02	-0.04	1.00	0.01	0.00	-0.18	-0.26	0.21	-0.16	-0.02	0.39	0.16	-0.11	0.05	0.13
Recency	-0.01	0.02	0.00	0.01	0.01	1.00	0.02	-0.01	0.02	0.00	0.03	0.02	0.00	-0.01	0.02	0.00	-0.02
MntWines	0.21	0.16	0.58	-0.50	0.00	0.02	1.00	0.39	0.57	0.40	0.39	0.39	0.01	0.55	0.63	0.64	-0.32
MntFruits	-0.06	0.02	0.43	0.37	-0.18	-0.01	0.39	1.00	0.55	0.59	0.57	0.40	-0.13	0.30	0.49	0.46	-0.42
MntMeatProducts	0.06	0.03	0.58	-0.44	-0.26	0.02	0.57	0.55	1.00	0.57	0.54	0.36	-0.12	0.31	0.73	0.49	-0.54
MntFishProducts	-0.08	0.04	0.44	-0.39	-0.21	0.00	0.40	0.59	0.57	1.00	0.58	0.43	-0.14	0.30	0.53	0.46	-0.45
MntSweetProducts	-0.08	0.02	0.44	-0.38	-0.16	0.03	0.39	0.57	0.54	0.58	1.00	0.36	-0.12	0.33	0.50	0.46	-0.42
MntGoldProds	-0.06	0.06	0.33	-0.36	-0.02	0.02	0.39	0.40	0.36	0.43	0.36	1.00	0.05	0.41	0.44	0.39	-0.25
NumDealsPurchases	0.04	0.06	-0.08	0.22	0.39	0.00	0.01	-0.13	-0.12	0.14	-0.12	0.05	1.00	0.24	-0.01	0.07	0.35
NumWebPurchases	0.1	0.15	0.39	-0.37	0.16	-0.01	0.55	0.30	0.31	0.30	0.33	0.41	0.24	1.00	0.39	0.52	-0.05
NumCatalogPurchases	0.09	0.12	0.59	-0.50	-0.11	0.02	0.63	0.49	0.73	0.53	0.50	0.44	-0.01	0.39	1.00	0.52	-0.52
NumStorePurchases	0.1	0.13	0.53	-0.50	0.05	0.00	0.64	0.46	0.49	0.46	0.46	0.39	0.07	0.52	0.52	1.00	0.43
NumWebVisitsMonth	-0.07	-0.12	-0.55	0.45	0.13	-0.02	-0.32	-0.42	-0.54	0.45	-0.42	-0.25	0.35	-0.05	-0.52	-0.43	1.00



Because the education is described in text, I transformed them into number to simple the correlation calculation. 1 is Basic, 2 is 2nd Cycle, 3 is Graduation, 4 is Master, 5 is PhD. The higher the number, the higher the degree.

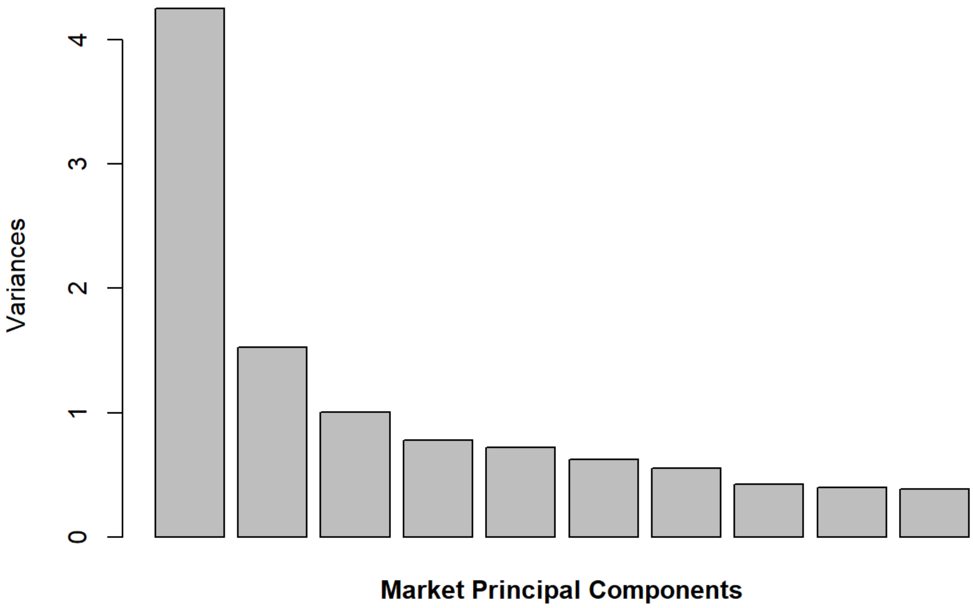
Shopping behaviour can be divided into a distinction between buying channels and the types of goods, and in this section we use this to analyse the influence of personality on both of these aspects. According to the correlation table and the correlation plot, There is no great correlation between academic qualifications and the purchase of most types of goods, except for alcohol. There is a strong positive correlation between education levels and money spent on the purchase of alcoholic beverages. At the same time, people with higher education levels spend less money on sweet. The data proves that the older you are, the more money you spend on each category. Compared to education and age, income is a better determinant of the type of goods purchased, as can be seen from the icons, the amount of all goods purchased is positively correlated with income, while the amount of money spent on alcohol and meat is strongly correlated with income. The number of children in the household is negatively correlated with the purchase of most goods, with the purchase of meat having the greatest impact; however, the amount spent on fish increases significantly with the number of children in the household. In conclusion, of all the personalities, income most influences the amount spent on each commodity type, while education has the least influence on them.

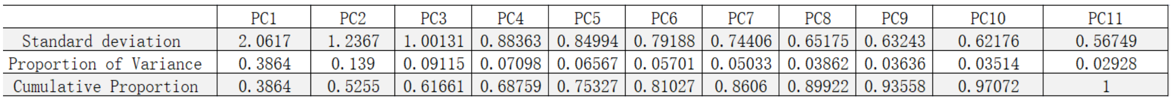
These inferences from tabular information are the same as real life experiences.

5.3 Data Processing and Results

5.3.1 PCA

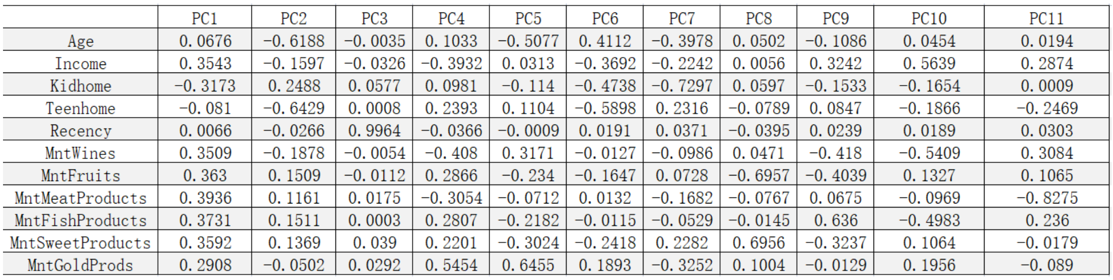
Because the high dimensions of the data set, so it is necessary to have dimensionality reduction through PCA before the clustering analysis. 11 principal components are obtained and shown in the following table.





	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
Standard deviation	2.0617	1.2367	1.00131	0.88363	0.84994	0.79188	0.74406	0.65175	0.63243	0.62176	0.56749
Proportion of Variance	0.3864	0.139	0.09115	0.07098	0.06567	0.05701	0.05033	0.03862	0.03636	0.03514	0.02928
Cumulative Proportion	0.3864	0.5255	0.61661	0.68759	0.75327	0.81027	0.8606	0.89922	0.93558	0.97072	1

According to the bar chart above, we can find that most of the variance is included in the first six principal components, and the table shows that the first six principal components explain 81 percent of the overall rate of change, so, in conducting this part of the analysis, we mainly choose the first six principal components for specific analysis.

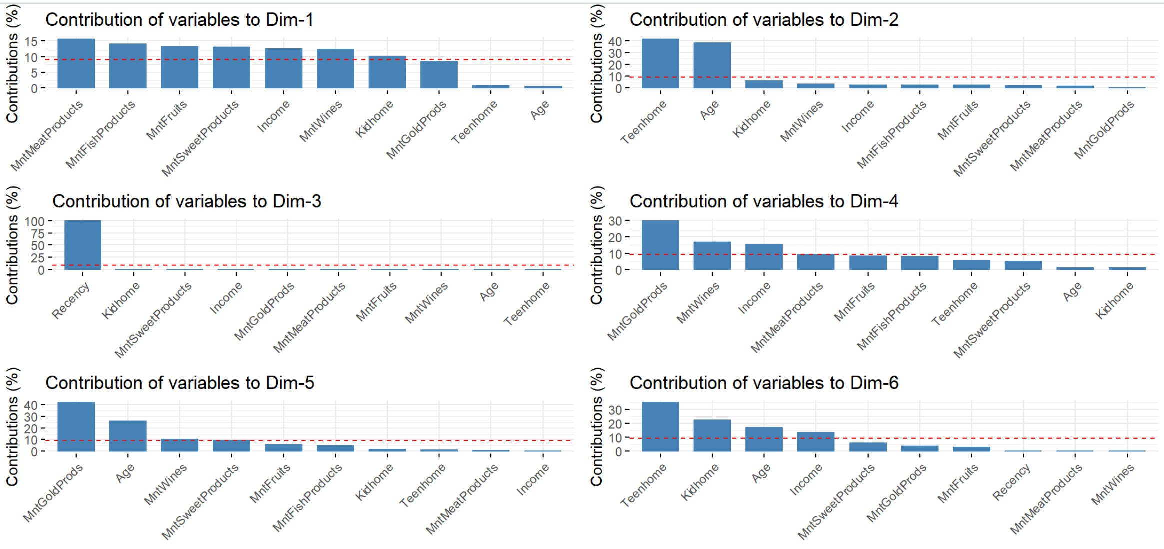


	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
Age	0.0676	-0.6188	-0.0035	0.1033	-0.5077	0.4112	-0.3978	0.0502	-0.1086	0.0454	0.0194
Income	0.3543	-0.1597	-0.0326	-0.3932	0.0313	-0.3692	-0.2242	0.0056	0.3242	0.5639	0.2874
Kidhome	-0.3173	0.2488	0.0577	0.0981	-0.114	-0.4738	-0.7297	0.0597	-0.1533	-0.1654	0.0009
Teenhome	-0.081	-0.6429	0.0008	0.2393	0.1104	-0.5898	0.2316	-0.0789	0.0847	-0.1866	-0.2469
Recency	0.0066	-0.0266	0.9964	-0.0366	-0.0009	0.0191	0.0371	-0.0395	0.0239	0.0189	0.0303
MntWines	0.3509	-0.1878	-0.0054	-0.408	0.3171	-0.0127	-0.0986	0.0471	-0.418	-0.5409	0.3084
MntFruits	0.363	0.1509	-0.0112	0.2866	-0.234	-0.1647	0.0728	-0.6957	-0.4039	0.1327	0.1065
MntMeatProducts	0.3936	0.1161	0.0175	-0.3054	-0.0712	0.0132	-0.1682	-0.0767	0.0675	-0.0969	-0.8275
MntFishProducts	0.3731	0.1511	0.0003	0.2807	-0.2182	-0.0115	-0.0529	-0.0145	0.636	-0.4983	0.236
MntSweetProducts	0.3592	0.1369	0.039	0.2201	-0.3024	-0.2418	0.2282	0.6956	-0.3237	0.1064	-0.0179
MntGoldProds	0.2908	-0.0502	0.0292	0.5454	0.6455	0.1893	-0.3252	0.1004	-0.0129	0.1956	-0.089

Figure 5.1: correlation plot

From the above table, we can conclude that PC1 is mainly related to income and purchase of each category, and is positively correlated, which means that if the customer's PC1 is relatively large, it means that he has more income and spends a lot of money on the shopping site; PC2 shows a clear negative correlation between age and the number of children in the family, so it can be understood that the customer is younger and has a simpler family composition. In PC3, the frequency of visits to the shopping site is positive and close to 1; in PC4, income, the amount of money spent on alcohol and meat products is large and negative; in PC5, age and the amount of money spent on fish and sugar products are correlated and both are negative; in PC6, the correlation coefficients for the number of children in the household at each age are all negative and correlated.

The relationships between principal components and variables represent the coordinates of the variable in the PC coordinate system. To plot variables, we can use the function $fvizpcavar$.



In the graph above, the different contributions of each variable are shown. The red dashed line represents the average expected contribution. The value of each variable is a measure of the usefulness of a variable. A higher value indicates that the variable is more important in the principal component analysis. Higher values indicate that the variable contributes more to the principal component, i.e. near the edge of the circle in the correlation plot. Lower values indicate that the variable is not well represented by principal components and that the variable is near the centre of the circle in the correlation plot.

Using a graph to show the variable correlation, which shows the relationship between the variables and the principal components within the variable group. The positively correlated variables are close to each other, while the negatively correlated variables are opposite. The length from the centroid to the variable indicates the proportion of the variable in this dimension (also known as quality). "Contrib" shows how well variables and PC components are represented. And we can find the variables that are most relevant to the composition.

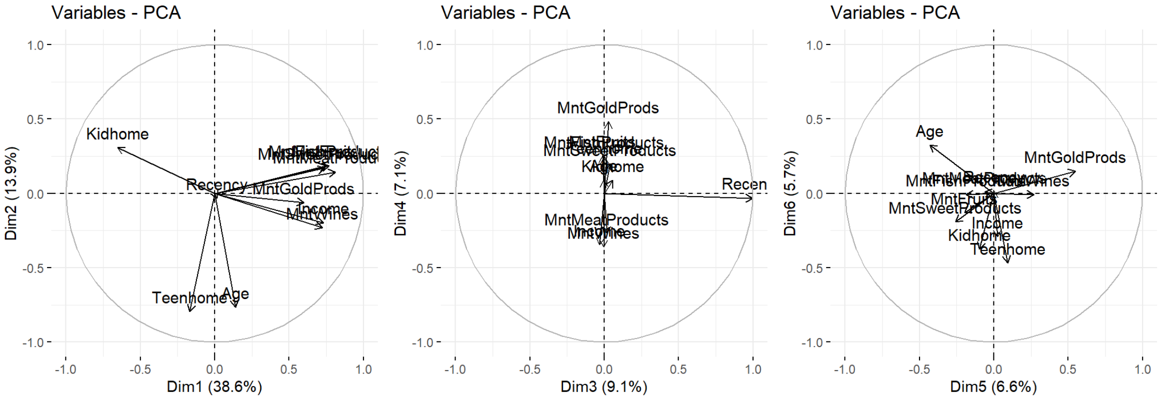


Figure 5.2: Factor negative load diagram

Above, we show the Factor negative load diagram in three different dimensions, where the image is more obvious when the dimension is large, i.e. the image below

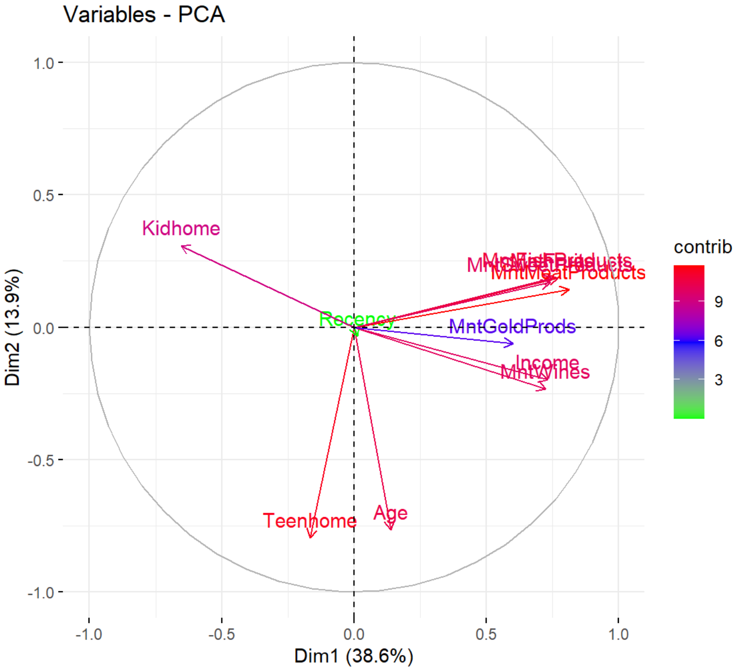
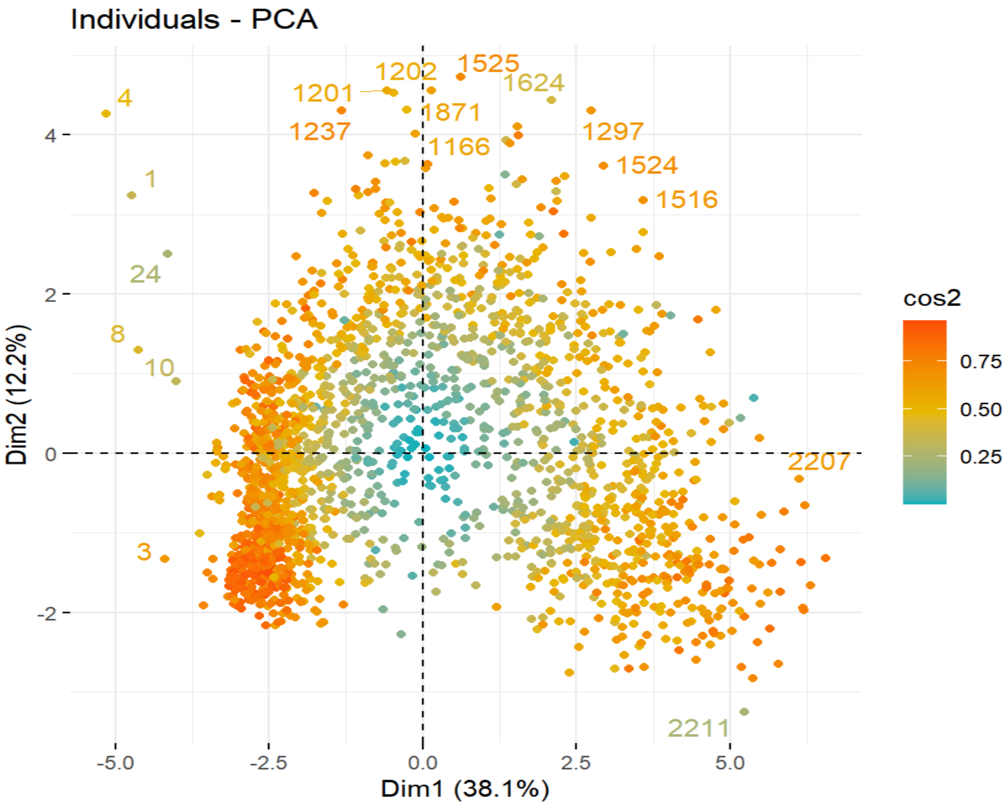
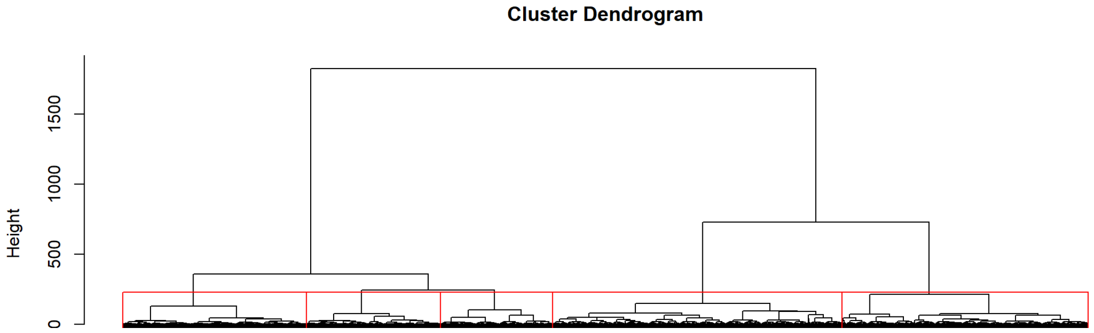


Figure 5.3: Factor negative load diagram

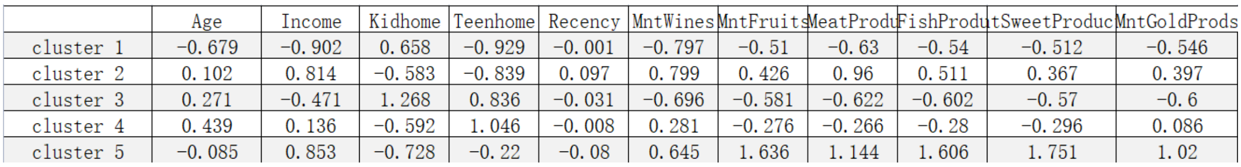


5.3.2 Hierarchical Clustering

Dealing with this data set we will mainly discuss agglomerative procedures for hierarchical clustering.
However, because there are too many data, the image of hierarchical clustering cannot analyze the data. Thus, one of the limitations of hierarchical clustering is that it is suitable for data sets with relatively small data volume. For example, because there are too many individual in this data set, so the image of hierarchical clustering is very complex. So I've only taken a small part of the image, so you can see that all 2000 data can be roughly divided into five categories. But obviously it is not very scientific to analyse such a large amount of data on a case-by-case basis.

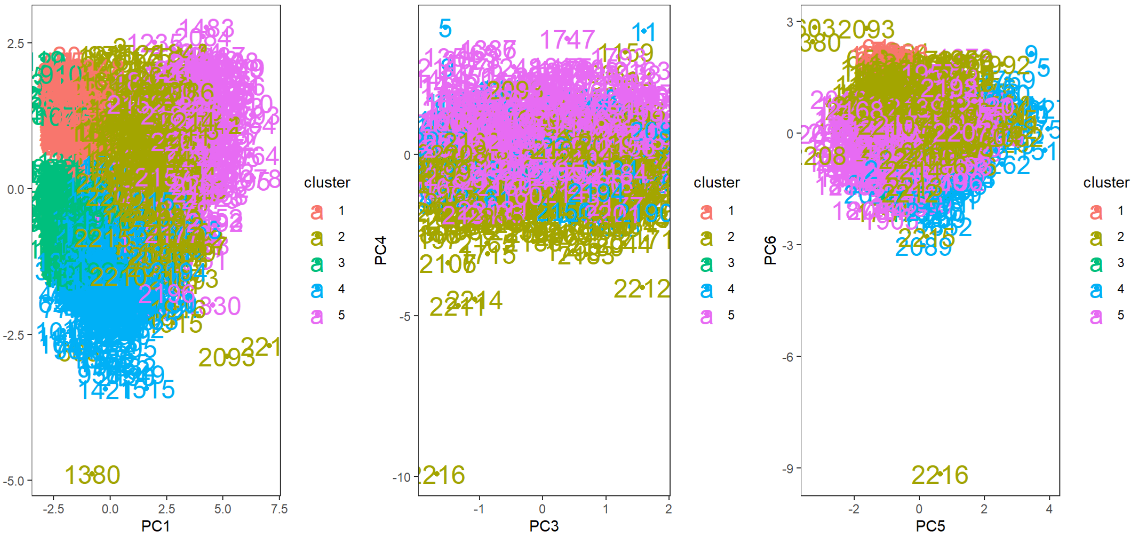


Although the tree diagram does not show the data clearly because there are too many individuals, we can use R to find the contribution and relevance of each dimension to the different clusters, which is shown in the following figure.



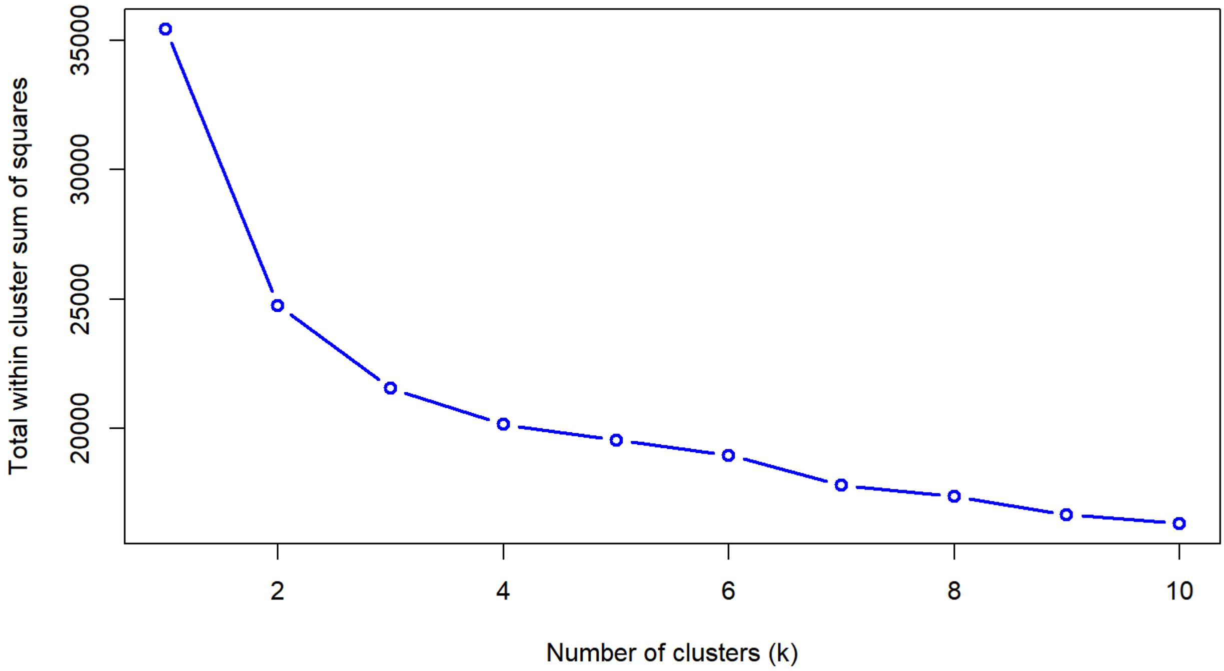
	Age	Income	Kidhome	Teenhome	Recency	MntWines	MntFruits	MeatProdu	FishProdu	tSweetProduc	MntGoldProds
cluster 1	-0.679	-0.902	0.658	-0.929	-0.001	-0.797	-0.51	-0.63	-0.54	-0.512	-0.546
cluster 2	0.102	0.814	-0.583	-0.839	0.097	0.799	0.426	0.96	0.511	0.367	0.397
cluster 3	0.271	-0.471	1.268	0.836	-0.031	-0.696	-0.581	-0.622	-0.602	-0.57	-0.6
cluster 4	0.439	0.136	-0.592	1.046	-0.008	0.281	-0.276	-0.266	-0.28	-0.296	0.086
cluster 5	-0.085	0.853	-0.728	-0.22	-0.08	0.645	1.636	1.144	1.606	1.751	1.02

The projections of the clusters on the different principal components are shown below. Because of the large amount of data, we can see if the grouping is reasonable by looking at the colours of the different populations. As can be seen from the graphs, all the data in PC3,PC4,PC5 and PC6 do not show clear groupings, while the first graph can clearly show the situation of each different grouping.

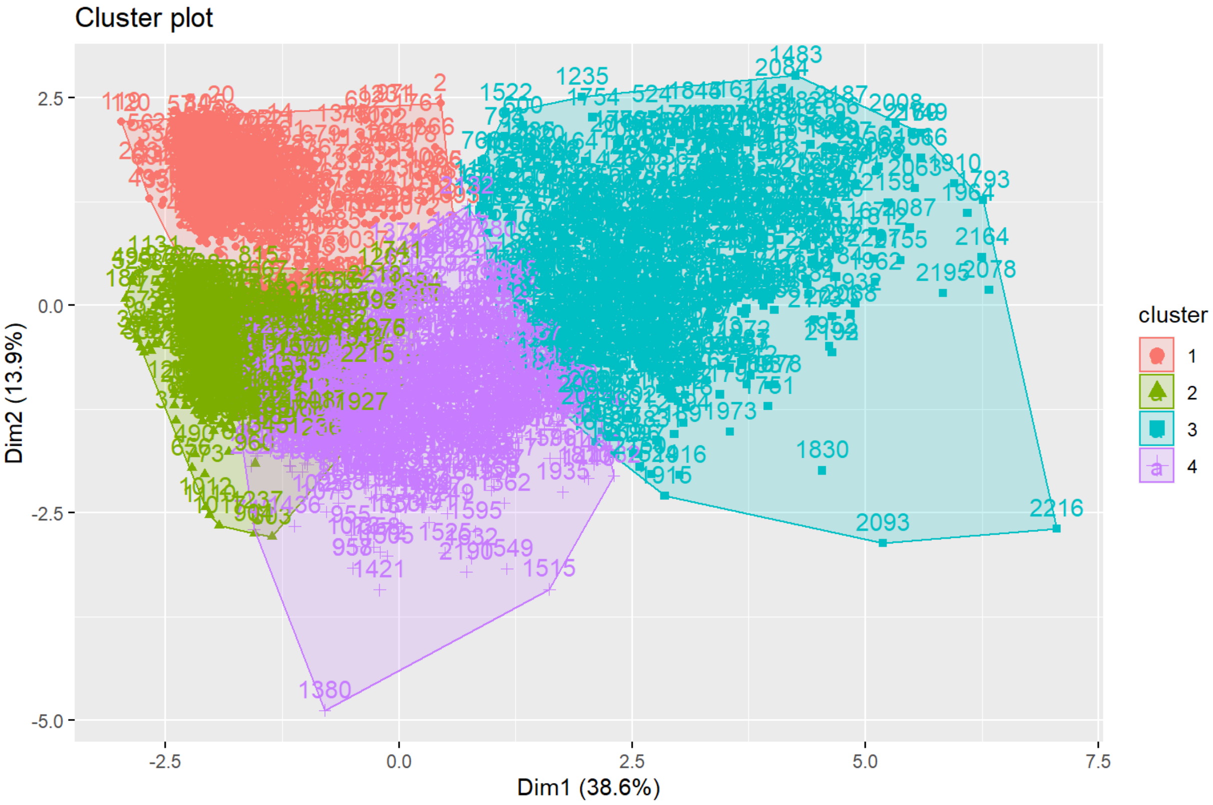


5.3.3 K-means

The first step of k-means algorithm for clustering analysis is to find the appropriate value of K. So the answer of k-means algorithm is more convincible. In cluster analysis, the elbow method is a heuristic method for determining the number of clusters in a data set. The method involves plotting the variance being explained as a function of the number of clusters and selecting the elbow of the curve as the number of clusters to be used.

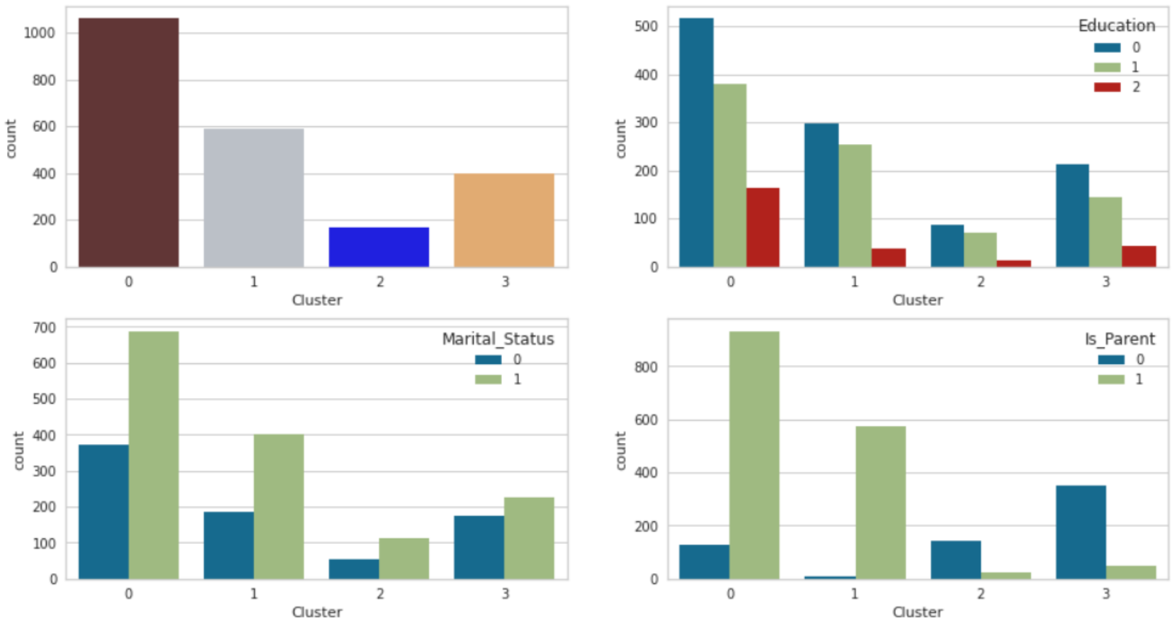


The search for the optimal k can be analysed using the above line graph. From the graph it can be seen that the slope is larger when k is less than 4, but when k is greater than 4 the slope is on the smaller side, so it can be concluded that dividing all the data into 4 populations is the most appropriate, which can not only show the type and relationship of the data well, but also save time.



From this figure, we can see that after ignoring a few outliers, all the data sets can be divided into four broad categories with clear boundaries between them, which can indicate that it is desirable when k is taken as 4. Afterwards, we can analyse the four broad categories to classify the data sets.

At the same time, the amount of data is too large to visualise the relationships between the groups and to define them through graphs. I will therefore present the data for each variable in the four groups in the form of a table, hoping to find the definitions of the four different groups through this operation.



Using the bar chart, I can observe that these individuals can be classified into different clusters by the following properties:

- cluster 0:

  1. Low Spent and Low Income
  2. Majority of these are parent
  3. At the max 5 members in the family

- cluster 1:

  1. Average Spent and Average Income
  2. Definitely a parent
  3. At the max 4 members in the family

- cluster 2:

  1. High Spent and High income
  2. Definitely not a parent
  3. At the max 2 members in the family

- cluster 3:

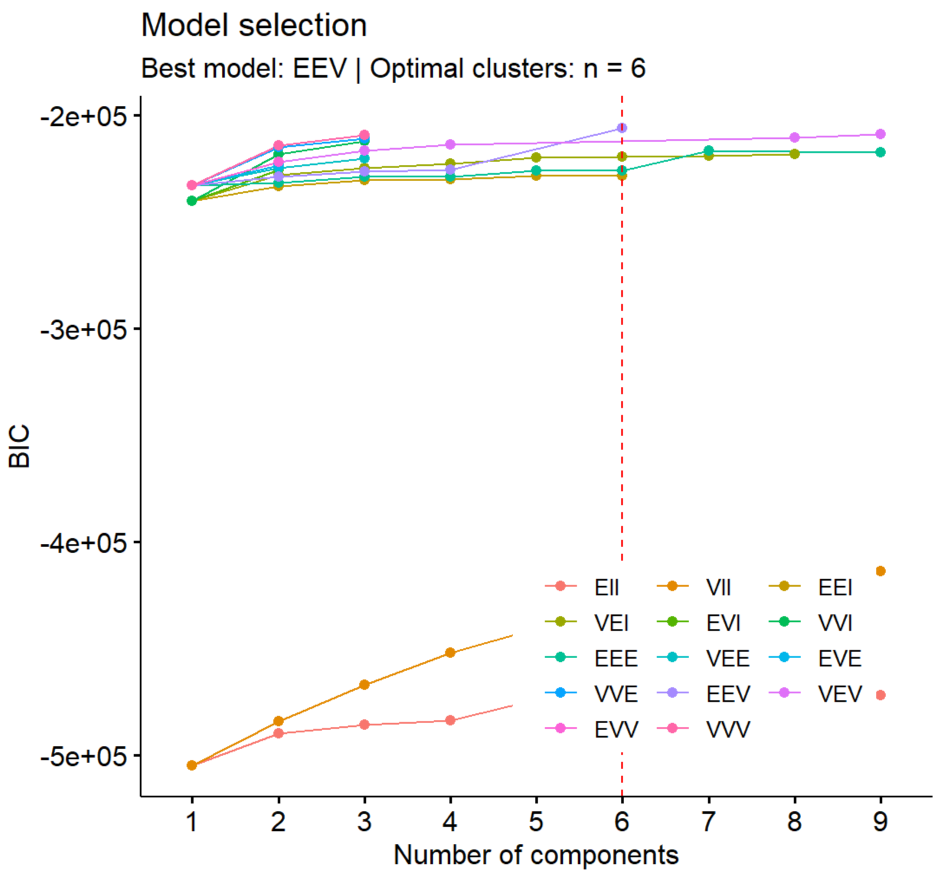
  1. High Spent and Average Income
  2. Definitely not a parent
  3. At the max 2 members in the family

5.3.4 Gaussian Mixture Model

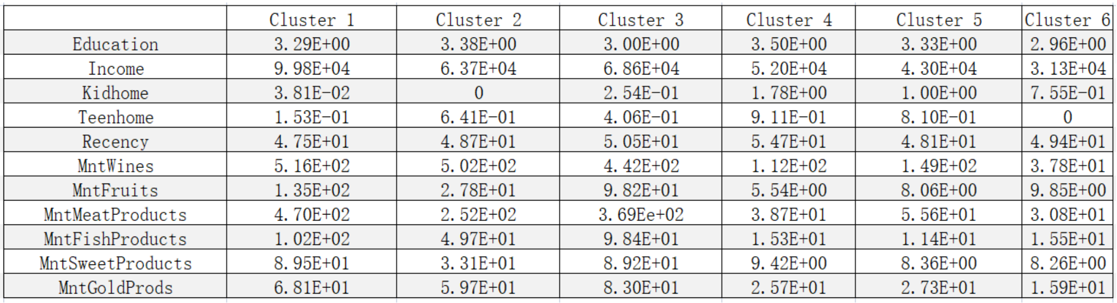
Model-based clustering enables clustering, classification and density estimation based on Gaussian finite mixture models. For Gaussian mixture models with various covariance structures, it provides parameter prediction functions according to the EM algorithm. It is also possible to visualise the fitted models in clustering, classification and density estimation results.

In the scenario of this question, there is no normal distribution across the data, and Mclust assumes that the observations are the result of sampling from one or more mixed Gaussian distributions, and Mclust needs to infer the best possible model parameters based on the available data, and how many sets of distributions are sampled from q.

Mclust provides a total of 14 models, and each model has a different two-dimensional form for the data. First, we need to identify the model and determine the number of groupings. In this question, we need to first identify the model to be fitted and the number of groupings; that is, Mclust obtains the results of the analysis for groups 1 to 9 of the 14 models by default, and then selects the final model and the number of groupings based on certain criteria. Also, Mclust provides two methods for evaluating the likelihood of different models under different groupings: BIC (Bayesian Information Criterion) and ICL (integrated complete-data likelihood). Next, we plot the BIC change curve.



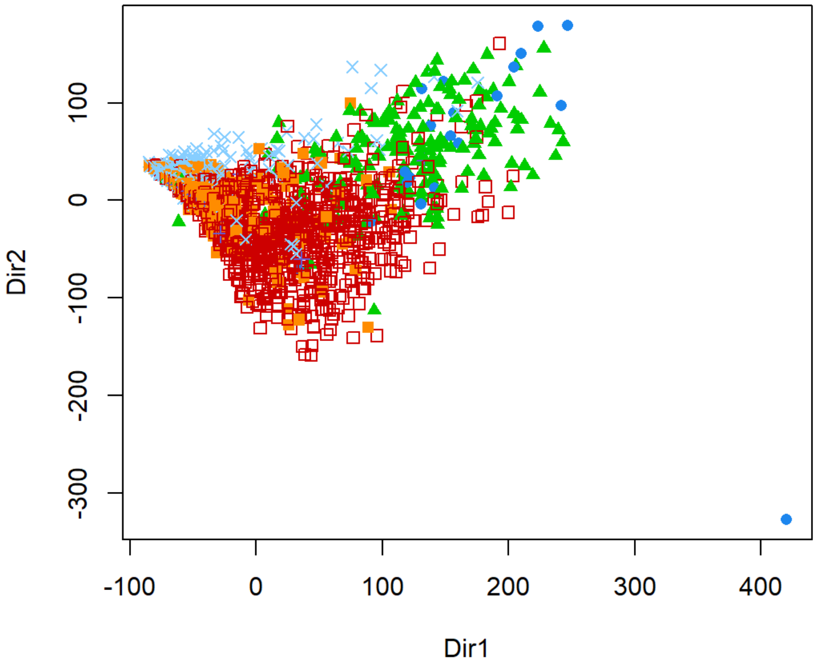
Using a Gaussian model, we partitioned the data set and divided it into six groups, the first cluster has 25 variables, the second with 957 variables, the third with 186 variables, the fourth with 22 variables, and the fifth and sixth clusters have 449 and 577 variables respectively.And the mixing probabilities are 0.0118, 0.432, 0.08677698 0.00991, 0.2244, 0.2346 in that order.

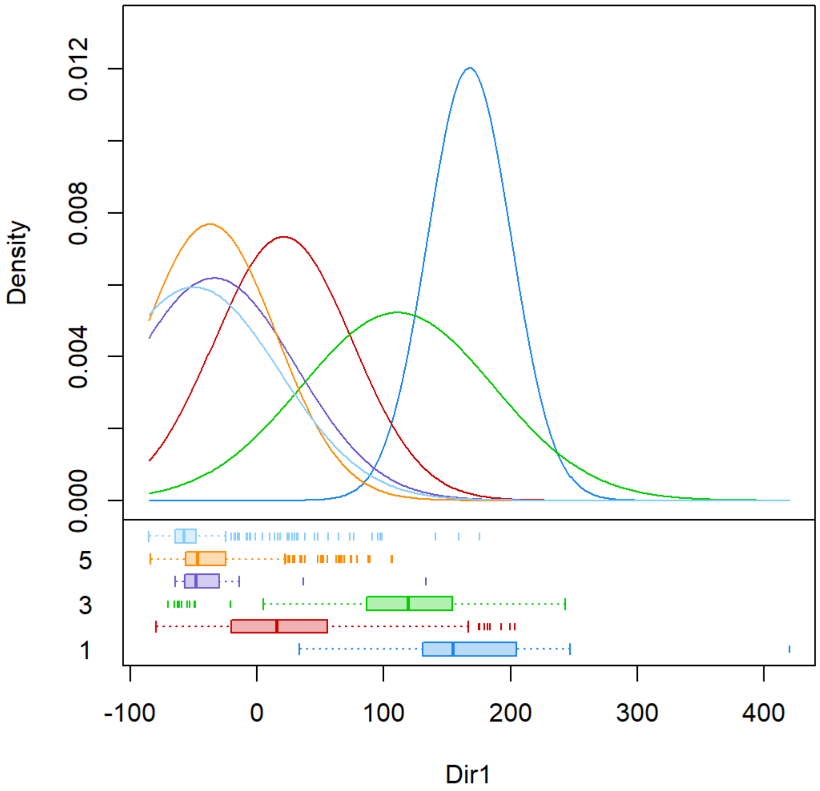


	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Education	3. 29E+00	3. 38E+00	3. 00E+00	3. 50E+00	3. 33E+00	2. 96E+00
Income	9. 98E+04	6. 37E+04	6. 86E+04	5. 20E+04	4. 30E+04	3. 13E+04
Kidhome	3. 81E-02	0	2. 54E-01	1. 78E+00	1. 00E+00	7. 55E-01
Teenhome	1. 53E-01	6. 41E-01	4. 06E-01	9. 11E-01	8. 10E-01	0
Recency	4. 75E+01	4. 87E+01	5. 05E+01	5. 47E+01	4. 81E+01	4. 94E+01
MntWines	5. 16E+02	5. 02E+02	4. 42E+02	1. 12E+02	1. 49E+02	3. 78E+01
MntFruits	1. 35E+02	2. 78E+01	9. 82E+01	5. 54E+00	8. 06E+00	9. 85E+00
MntMeatProducts	4. 70E+02	2. 52E+02	3. 69Ee+02	3. 87E+01	5. 56E+01	3. 08E+01
MntFishProducts	1. 02E+02	4. 97E+01	9. 84E+01	1. 53E+01	1. 14E+01	1. 55E+01
MntSweetProducts	8. 95E+01	3. 31E+01	8. 92E+01	9. 42E+00	8. 36E+00	8. 26E+00
MntGoldProds	6. 81E+01	5. 97E+01	8. 30E+01	2. 57E+01	2. 73E+01	1. 59E+01

In all six clusters, there is a significant share of income. In addition to this, there is a more significant share of spending on meat in cluster one and more on beer in cluster two, while the other four groups do not have a pronounced shopping tendency, as evidenced by the fact that they do not differ significantly in their purchases of different types of goods.This could indicate that income plays a decisive role in all customers characteristics; and in terms of shopping expenses, people can be divided according

to the money spent on alcoholic and meat goods. The other dimensions may not play decisive roles in differentiating between different groups of customers.
After that, the model and grouping with the largest BIC among them is selected as the final result. So in the context of this data set, which can be divided into a total of n populations, we can visualise the corresponding information in the form of scatter and density.





5.4 Discussion

As this data is collectively large and of great commercial use, in this study we focus on going over how these users are grouped and what the characteristics of the different groupings are. At the same time, we find that the conclusions using the three clustering methods above were similar.

5.4.1 Clusters based on Hierarchical Clustering

Because of there are a great number of individuals, we don't have a way to visualize the number of clusters corresponding to all individuals. So, hierarchical clustering is not as useful as the other two methods when dealing with large data sets. Cluster1 and Cluster2 are both highly influenced by income and the number of children, and people spend more money in wines. While Cluster1 was also influenced to some extent by age, but Cluster 2 was not related to age. Cluster3 and Cluster4 describe people have kids and teens at home, while people in Cluster3 pay more attention on wines meat and fish. There is no clear shopping preference, i.e. the amount spent on each type of product is not very different. And Cluster5 means people with high income, but pay more attention on daily goods expect wines.

In all, recency hardly affects customers' shopping choices, and customers' personal characteristics are not relevant to how often they use shopping sites. Income levels and family situation are significant in shopping tendency, especially the money spent on meat, wine and fish.

5.4.2 Clustering based on K-means Clustering

In this study, k-means divides the data set into four groups, each of which is differentiated by income, expenditure and household size. For example, the behaviour of low-income, low-spending households with two children on online shopping will differ significantly from that of high-income, high-spending users with no children. However, there are some drawbacks to this methods. For example, wine purchases are not the same as fruit and vegetable purchases when it comes to online shopping behaviour, using only a single k-means will only reveal the general shopping behaviour of users, but if we want to go deeper into the differences between each category and the purchase channels, more optimisation will be needed.

However, in general, k-means has made a great contribution to our analysis of the relationship between online shopping behaviour and personality. There are four clusters in this method. The first group is users with high income and high expenditure who also have up to three children at home, the second group is users with high income and high expenditure and the presence of single-parent families, the third group is users with high income and high expenditure and mostly childless families with two members, and the fourth group is users with average income and high expenditure and mostly childless families. There are significant shopping differences between these four groups of users and a small number of outliers.

5.4.3 Clustering based on Gaussian Mixture Model

The essence of the 2-dimensional k-means model is that it draws a circle with the centre of each cluster as the centre and the maximum Euclidean distance from the cluster midpoint to the cluster centroid as the radius. This circle rigidly truncates the training set. Furthermore, k-means requires that the shape of these clusters must be circular. As a result, the clusters fitted by the k-means model differ significantly from the actual data distribution, so the Gaussian mixture model is used as an optimisation of k-means to cluster the data set more clearly.

In this data set, the Gaussian mixture model divides the users into six groups. And most people are in Cluster2, Cluster3, Cluster5 and Cluster6, so it is similar to the result in k-means. And income is still the most significant factor, the number of kids and teens at home also plays an important role. As with the other two methods, these conditions largely determine the amount of money they spend on wine and meat and fish. It is worth noting that the Gaussian mixture model takes into account the variance of each group when clustering compared to k-means, which makes the results more realistic.

Conclusion

Based on the distance function learned before, this thesis investigates the commonly used unsupervised machine learning method - cluster analysis, including the principal component analysis (PCA) which is used for data pre-processing, and three common clustering methods including hierarchical clustering, k-means, and Gaussian Mixture Models (GMM). The underlying mathematical logic of these three methods and the fundamental codes are understood in order to be used for practical applications to specific sets of numbers. Meanwhile, the data set selected for this paper is real data from Amazon customers, containing user personality, such as user age, education, marital status, fertility status, as well as user shopping behaviour, including the purchase amount for each item category (wine, sweet, meat, gold, fruits and fish) , in order to find the influence of user personality on user behaviour and to group more than 2000 users to find commonalities.

The aim of the paper is to apply unsupervised learning techniques to the online shopping industry. Due to the epidemic, online shopping platforms have grown by leaps and bounds. However, in the face of such a bright market, many merchants do not know how to find their target customers and third-party shopping platforms are facing an excessive workload. Therefore, our findings can also add some business value, for example, e-merchants can stratify their users accurately according to the products they sell, thus reducing the cost of advertising distribution; e-merchant platforms can also manage different kinds of products on the platform by stratifying them through user portraits, increasing the turnover of the platform.

Research analysis shows that most important determinants of purchase intentions are the income of the user and the childbearing status of the family, for example, the purchase of alcohol products is more likely to be made by people with high incomes and no children, and the purchase of sweet products is also made by people without children. At the same time, higher income groups tend to spend more money on alcohol and gold products. Although there is no strong correlation between education and purchasing preferences, there is a close relationship with income. Therefore, in reality, wine distributors can focus more on the unmarried and childless high-education and high-income group, and meat and fruit distributors can pay more attention to the information of adults who have already had children.

However, there are certain limitations to this study. The first limitation is the methodology, in this paper I only used the three most commonly used methods and came up with two answers, and the number of clusters in the K-means is by myself according to the relationship between number of clusters and sum of square, which is somewhat subjective. In subsequent applications, the staff involved could have used more clustering techniques and a more objective approach to determine the number of clusters, trying to find more detailed patterns and relationships. The second limitation is that the data set is too small. Although this data set contains information on more than two thousand users, according to information from the end of 2021, only Amazon had reached 200 million users, so the two thousand data set does not show all the patterns. The third limitation is that compared to the user's personality, the user's path to purchase and the strength of discounts on each country's trend platform can largely determine the user's purchase behaviour, and it is somewhat one-sided to study only the user's own label. So in the real life, e-commerce companies need to refine their clustering analysis according to their own product categories, which can refer to the user's personality, purchase channels, platform activities, social trends and other dimensions of analysis.

References

[1] Arbab Waseem Abbas et al. “K-Means and ISODATA clustering algorithms for landcover classification using remote sensing”. In: *Sindh University Research Journal-SURJ (Science Series)* 48.2 (2016).

[2] Gediminas Adomavicius and Alexander Tuzhilin. “Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions”. In: *IEEE transactions on knowledge and data engineering* 17.6 (2005), pp. 734–749.

[3] Michael R Anderberg. “The broad view of cluster analysis”. In: *Cluster analysis for applications* 1.1 (1973), pp. 1–9.

[4] David Arthur and Sergei Vassilvitskii. *k-means++: The advantages of careful seeding*. Tech. rep. Stanford, 2006.

[5] Barry J Babin, William R Darden, and Mitch Griffin. “Work and/or fun: measuring hedonic and utilitarian shopping value”. In: *Journal of consumer research* 20.4 (1994), pp. 644–656.

[6] Swarna Bakshi. “Impact of gender on consumer purchase behaviour”. In: *Journal of Research in Commerce and Management* 1.9 (2012), pp. 1–8.

[7] Murray R Barrick and Michael K Mount. “The big five personality dimensions and job performance: a meta-analysis”. In: *Personnel psychology* 44.1 (1991), pp. 1–26.

[8] Roger K Blashfield. “Mixture model tests of cluster analysis: accuracy of four agglomerative hierarchical methods.” In: *Psychological Bulletin* 83.3 (1976), p. 377.

[9] Lauren G Block and Vicki G Morwitz. “Shopping lists as an external memory aid for grocery shopping: Influences on list writing and list fulfillment”. In: *Journal of Consumer Psychology* 8.4 (1999), pp. 343–375.

[10] Hans-Hermann Bock. “Clustering methods: a history of k-means algorithms”. In: *Selected contributions in data analysis and classification* (2007), pp. 161–172.

[11] Max Bramer. *Clustering*. Springer, 2007.

[12] Forrest E Clements. “Use of cluster analysis with anthropological data”. In: *American Anthropologist* 56.2 (1954), pp. 180–199.

[13] Richard M Cormack. “A review of classification”. In: *Journal of the Royal Statistical Society: Series A (General)* 134.3 (1971), pp. 321–353.

[14] Helga Dittmar, Karen Long, and Rosie Meek. “Buying on the Internet: Gender differences in on-line and conventional buying motivations”. In: *Sex roles* 50.5 (2004), pp. 423–444.

[15] Mohammed El Agha and Wesam M Ashour. “Efficient and fast initialization algorithm for k-means clustering”. In: *International Journal of Intelligent Systems and Applications* 4.1 (2012).

[16] Brian S Everitt and Anders Skrondal. “The Cambridge dictionary of statistics”. In: (2010).

[17] Michael Fop and Thomas Brendan Murphy. “Variable selection methods for model-based clustering”. In: *Statistics Surveys* 12 (2018), pp. 18–65.

[18] Chris Fraley and Adrian E Raftery. “Model-based clustering, discriminant analysis, and density estimation”. In: *Journal of the American statistical Association* 97.458 (2002), pp. 611–631.

[19] António Guterres. “Mental health services are an essential part of all government responses to COVID-19”. In: *United Nations, COVID-19 Response* 13 (2020).

[20] Gregory James Hamerly. *Learning structure and concepts in data through data clustering*. University of California, San Diego, 2003.

[21] Peter A Henderson and Richard MH Seaby. *A practical handbook for multivariate methods*. Pisces Conservation Lymington, England, 2008.

[22] Elizabeth C Hirschman and Morris B Holbrook. “Hedonic consumption: emerging concepts, methods and propositions”. In: *Journal of marketing* 46.3 (1982), pp. 92–101.

[23] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. “Data clustering: a review”. In: *ACM computing surveys (CSUR)* 31.3 (1999), pp. 264–323.

[24] Mohammad Hossein Moshref Javadi et al. “An analysis of factors affecting on online shopping behavior of consumers”. In: *International journal of marketing studies* 4.5 (2012), p. 81.

[25] Michael I Jordan and Tom M Mitchell. “Machine learning: Trends, perspectives, and prospects”. In: *Science* 349.6245 (2015), pp. 255–260.

[26] Julia Koch, Britta Frommeyer, and Gerhard Schewe. “Online shopping motives during the COVID-19 pandemic—lessons from the crisis”. In: *Sustainability* 12.24 (2020), p. 10247.

[27] Gauri Kulkarni, Brian T Ratchford, and PK Kannan. “The impact of online and offline information sources on automobile choice behavior”. In: *Journal of Interactive Marketing* 26.3 (2012), pp. 167–175.

[28] Brenda L Beagan, Gwen E Chapman, and Elaine M Power. “Cultural and symbolic capital with and without economic constraint: Food shopping in low-income and high-income Canadian families”. In: *Food, Culture & Society* 19.1 (2016), pp. 45–70.

[29] Godfrey N Lance and William Thomas Williams. “A general theory of classificatory sorting strategies: 1. Hierarchical systems”. In: *The computer journal* 9.4 (1967), pp. 373–380.

[30] Sharon X Lee and Geoffrey J McLachlan. “Finite mixtures of canonical fundamental skew t-distributions”. In: *Statistics and computing* 26.3 (2016), pp. 573–589.

[31] Ephraim S Leibtag and Phillip R Kaufman. *Exploring food purchase behavior of low-income households: how do they economize?* Tech. rep. 2003.

[32] Andrzej Maćkiewicz and Waldemar Ratajczak. “Principal components analysis (PCA)”. In: *Computers & Geosciences* 19.3 (1993), pp. 303–342.

[33] James MacQueen et al. “Some methods for classification and analysis of multivariate observations”. In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1. 14. Oakland, CA, USA. 1967, pp. 281–297.

[34] Amandeep Kaur Mann and Navneet Kaur. “Review paper on clustering techniques”. In: *Global Journal of Computer Science and Technology* (2013).

[35] Francis Henry Charles Marriott. “Practical problems in a method of cluster analysis”. In: *Biometrics* (1971), pp. 501–514.

[36] Geoffrey J McLachlan and David Peel. “Robust cluster analysis via mixtures of multivariate t-distributions”. In: *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*. Springer. 1998, pp. 658–666.

[37] Paul D McNicholas. “Model-based clustering”. In: *Journal of Classification* 33.3 (2016), pp. 331–373.

[38] Glenn W Milligan. “An examination of the effect of six types of error perturbation on fifteen clustering algorithms”. In: *psychometrika* 45.3 (1980), pp. 325–342.

[39] Glenn W Milligan and Martha C Cooper. "Methodology review: Clustering methods". In: *Applied psychological measurement* 11.4 (1987), pp. 329–354.

[40] KA Abdul Nazeer and MP Sebastian. "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm". In: *Proceedings of the world congress on engineering*. Vol. 1. Citeseer. 2009, pp. 1–3.

[41] Douglas A Reynolds. "Gaussian mixture models." In: ().

[42] Estrella Romero et al. "Traits, personal strivings and well-being". In: *Journal of Research in Personality* 43.4 (2009), pp. 535–546.

[43] ZHANG Rong et al. "Overviewing of visual analysis approaches for clustering high-dimensional data". In: *Journal of Graphics* 41.1 (2020), p. 44.

[44] Mustafa Savci et al. "The development of the Turkish craving for online shopping scale: a validation study". In: *International Journal of Mental Health and Addiction* (2021), pp. 1–17.

[45] Amit Saxena et al. "A review of clustering techniques and developments". In: *Neurocomputing* 267 (2017), pp. 664–681.

[46] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. "Kernel principal component analysis". In: *International conference on artificial neural networks*. Springer. 1997, pp. 583–588.

[47] Catrin Sohrabi et al. "World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19)". In: *International journal of surgery* 76 (2020), pp. 71–76.

[48] Erwin Stolz, Hannes Mayerl, and Wolfgang Freidl. "The impact of COVID-19 restriction measures on loneliness among older adults in Austria". In: *European journal of public health* 31.1 (2021), pp. 44–49.

[49] Dingna Tang et al. "What determines online consumers to migrate from PCs to mobile devices?-An empirical approach on consumers' internet cross-channel behaviours". In: *International Journal of Services Technology and Management* 22.1-2 (2016), pp. 46–62.

[50] Zofija Tupikovskaja-Omovie and David Tyler. "Clustering consumers' shopping journeys: eye tracking fashion m-retail". In: *Journal of Fashion Marketing and Management: An International Journal* (2020).

[51] Kiri Wagstaff et al. "Constrained k-means clustering with background knowledge". In: *Icml*. Vol. 1. 2001, pp. 577–584.

[52] Na Wang, Dongchang Liu, and Jun Cheng. “Study on the influencing factors of online shopping”. In: *Proceedings of the 11th Joint Conference on Information Sciences, Published by Atlantis Press*. 2008, pp. 1–4.

[53] Bo Xiao and Izak Benbasat. “E-commerce product recommendation agents: Use, characteristics, and impact”. In: *MIS quarterly* (2007), pp. 137–209.

[54] Miin-Shen Yang, Chien-Yo Lai, and Chih-Ying Lin. “A robust EM clustering algorithm for Gaussian mixture models”. In: *Pattern Recognition* 45.11 (2012), pp. 3950–3961.

[55] Sobia Zahra et al. “Novel centroid selection approaches for KMeans-clustering based recommender systems”. In: *Information sciences* 320 (2015), pp. 156–189.