Department of Mathematical Sciences

Final Year Project
Dissertation of Undergraduate of Financial Mathematics

The geographical distribution of dietary preferences in China: A study based on unsupervised learning methods

一项基于无监督学习的中国饮食偏好地域分布的研究

Name: Liangliang Zhu
ID number: 1718999

Supervisor: **Dr.Mu He**

5th May 2021

Contents

Chapter 1

Abstract

Dietary factors contribute to the risk of developing metabolic disorders, such as hypertension, diabetes, hyperglycemia, and obesity. The prevalence of these diseases has increased markedly in recent decades, which has aroused public concern. Based on the latest report Zhao et al. (2020), we obtain the dataset, which was composed of six distinctive dietary preferences, including fried food, Grilled Food, Spicy Food, Tongue-numbing Food, Instant-boiled Food, Sweet Food. Besides, there are some metabolic status data for measuring human health status, including the prevalence of diabetes and hypertension, Body Mass Index (BMI), postprandial plasma glucose (PPG), and fasting plasma glucose (FPG). This paper performs correlation analysis to investigate how dietary preference influences the prevalence of diseases and researched which provinces are the certain disease area with high incidence. Moreover, we present algorithms for K-Means, Hierarchy, Fuzzy-c Means Clustering to cluster 30 provinces in China according to the dietary preference and physical health of local citizens. In our study, we found that geographical proximity would lead to a similar dietary preference. For example, fried food and grilled food preference were mainly located in the northeast region of China. The spicy food preference and tongue-numbing food preference were centered on the southwest area. We also discovered that sweet food is the most popular food in China and spicy food is in second place. People in Liaoning, Beijing, Heilongjiang and Jilin need more care about developing diabetes, hypertension , and obesity because they intake much more fried food and grilled food than people in other provinces. However, results show that eating spicy food, tongue-numbing food can reduce the risk of diabetes and effectively decrease the value of plasma glucose. Different clustering algorithms will give different clustering outcomes. Hierarchy clustering and Fuzzy-c Means Clustering decrease the overlap area on PC1 and PC2 which perform better than K-Means Clustering in our research.

饮食因素会增加患代谢紊乱相关疾病的的风险，如高血压，糖尿病，高血糖和肥胖症。近几十年来，这些疾病的患病率显著上升，引起了公众的关注。我们从最近学者的报告中获得了数据集，其中包括6种饮食偏好，包括油炸食物，烧烤食物，辛辣食物，舌头麻木食物，煮熟食物，甜食。此外，还有一些代谢状态数据可以用来衡量人类的健康状况，包括糖尿病和高血压的患病率、体重指数(BMI)、餐后血糖(PPG)和空腹血糖(FPG)。本文通过相关分析研究了饮食偏好对疾病流行的影响，并研究了哪些省份是疾病高发区。此外，本文还使用K-Means、Hierarchy、Fuzzy-c Means聚类算法对中国30个省份居民的饮食偏好和身体健康状况进行聚类。在我们的研究中，我们发现地理位置接近的地区上有相似的饮食偏好。例如，油炸食品和烧烤食品偏好主要集中在中国东北地区。麻辣和麻舌偏好集中在西南地区。我们还发现，在中国，甜食是最受欢迎的食物，辣味食物位居第二。辽宁、北京、黑龙江、吉林等地的居民比其他省份的人民摄入更多的油炸食品和烧烤食品，因此他们需要更加小心患上糖尿病、高血压、肥胖等疾病。然而，研究发现食用辛辣食物、舌麻食物可以降低患糖尿病的风险，并能有效降低血糖值。不同的聚类算法会产生不同的聚类结果。层次聚类和Fuzzy-c Means聚类减少了在第一主成成分和第二主成成分上的重叠区域，他们在我们的研究中表现优于K-Means聚类。

Key words: Dietary Preference, Geography Distribution, Metabolic Disease, Principal component Analysis, Clustering Algorithm.

Chapter 2

Introduction

Hypertension, diabetes, obesity become more serious in recent decade in China, the prevalence rate reached an unprecedented height. Study of Wang et al. (2018) found that the burden of hypertension and cardiovascular disease in China is rising along with the urbanlization, higher income and aging of the population. Even though medical field have used antihypertensive medications to treat high blood pressure, the hypertension prevalence rate still high in China compared with high-income countries. Several epidemiological surveys reported that hypertension prevalence in China is ranging from 26.6% to 33.6%.Gao et al. (2013).

Besides this, reducing the prevalence of diabetes is other public health challenges facing in China Atlas (2015). China has the largest number of patients with diabetes, and this number continue growing. An unprecedented amount of the prevalence of diabetes in the aged population, young people and male have become a public health concern. Many scholars like Xu et al. (2013) point out that approximately 114 million Chinese suffer from diabetes, more than 10% of adults were diagnosed with diabetes.

Diet has close connection with these diseases. As World Health Organization's (WHO) reported that an unhealthy diet would lead to a series of chronic diseases, such as cardiovascular diseases, diabetes, cancer, and overweight. Therefore, it would be valuable to investigate the how dietary preference influence on human physical health.

In fact, many scholars have been made great effort in this field. For example, in Guallar-Castillón et al. (2007) and Sayon-Orea et al. (2013) study, they showed that excessive consumption of fried foods can lead to obesity and overweight. The reason was explain as Donfrancesco et al. (2008)’ s report, when the frying oil reused for many times, the fatty acid composition was changed, water content reduced and frying oil deteriorated, energy density increased. Meta-analysis of Qin et al. (2021) have shown that ingesting too much frying food also can cause the increase risk of hypertension AND type 2 diabetes mellitus.

Cai et al. (2012) found that the high temperature treated food contains too much advanced glycation end - products (AGEs) and trans - fatty acids (TFAs), which can cause metabolic diseases.Besides, Khan and Sievenpiper (2016) point out that Fructose-containing sugars can result in weight gain, diabetes and cardiovascular disease due to the excess calories they provided. As evidenced by Stanhope et al. (2009), "fructose is an unregulated metabolic substrate for fatty acid synthesis in the liver, unlike glucose, it steeres by the main rate-limiting steps of glycolysis." Further, people are likely to overconsume the fructose, due to its special incretory character. Fructose does not stimulate insulin, thus it is hard to create a feeling of satiety for human Teff et al. (2004). There are excess sugars and calories in the form of sugary food and sugar-sweetened beverages, which may cause the potential for overconsumption of sugars. Johnson et al. (2009) point out that sugary food is highly positive correlated with diabete prevalence and FPG.

Besides, in the experimental studies of Sun et al. (2016), researchers demonstrated that capsaicin could ameliorate obesity, diabetes, and hypertension and TRPV1 activation improved the competence of cardiometabolic organs. Capsaicin is beneficial for body health, fresh chili pepper, could decrease the fatality caused by diabetes Lv et al. (2015).

Hence, our study proposes to investigate how dietary preference correlated to the metabolic disease prevalence. Then providing some nutritional advice and warning to people to help them modify eating behavior towards to a healthier and more nutritious diet. Furthermore, In China, due to many external factors, including the differences of climate, economic environment, culture beliefs, altitude and many others, eating habits can vary greatly from place to place. In our research, we aim to cluster 30 provinces in China according to the dietary preference of local people to find some interesting conclusions, such as which provinces have similar dietary preferences and what their dietary preferences are.

Chapter 3

Literature review

Principal component analysis (PCA) is a dimensionality-reduction method that provides a more interpretative way to process the large datasets. PCA transforming a large set of variables into a few principal componet that still maintain the most of the information of original large dataset. PCA was first formulated in 1901 by Porter (2006), who describe it as an analogue of the principal axis theorem “lines and planes of closest fit to systems of points in space”. Yang et al. (2004) proposed a new efficient image representation technique called two-dimensional component analysis (2DPCA). 2DPCA, unlike PCA, its image covariance matrix is constructed directly by the original image matrix. It performs better than PCA in extracting of image features and recognition rate. Kernel principal component analysis is a powerful preprocessing method for classification algorithms, instead of input the pre-images into space, it uses nonlinear map. Kernel PCA as an efficient feature extractor, using the integral operator kernel functions to compute principal components in highly dimensional feature space Ge et al. (2009). Maximum likelihood PCA (MLPCA) is a generalization of principal component algorithm, which aims to infer consistent estimators in the condition of presenting the errors with known error distribution. MLPCA can be utilized to fit PCA models when missing data, and it performs better than original algorithm Folch-Fortuny et al. (2016). However, MLPCA has a major issue in the aspect of determining the number of principal components. Bayesian PCA can well tickle with this problem, it automatically select an appropriate model dimensionality Nakajima et al. (2011).

Clustering is generally used in unsupervised machine learning and act as a useful tool for statistical data analysis, it aims to partition population or data points into certain groups based on the notion of similarity. In recent years, clustering algorithm become extensively used in molecular biology for gene expression analysis. Likewise, clustering algorithm group genes according to the similarity between their expression profiles. There are many different types of clustering algorithms, like K-means clustering algorithm, Hierarchical clustering, mean-shift clustering algorithm, Spatial clustering, fuzzy-c means and so forth.

K-means described by MacQueen et al. (1967) is one of the most popular partitional clustering algorithm, it partitions a bulk of observations into k clusters. The idea traces back to Hugo Steinhaus in 1956, and the standard algorithm was first proposed in 1957 by Stuart Lloyd which act as a technique for pulse-code modulation. K-means algorithm was modified by many times, Constrained K-means clustering (Cop-Kmeans) improve the clustering accuracy greatly. In the study of Wagstaff et al. (2001), Cop-Kmeans algorithm attains 100% accuracy overall except three data sets, which consistently outperformed the unconstrained k-means (only 58% on average). Global K-means clustering algorithm Likas et al. (2003) dynamically adds one cluster each time and employs k-means to minimize the sum of the within-cluster variances. This modification efficiently reduce the computational load and shrink the execution time.

Genetic k-means algorithm (GKA) Krishna and Murty (1999) is a hybridization of genetic algorithm (GA) and k-means clustering algorithm. GKA choose the global optimum evidenced by finite Markov chain theory, and it performs faster than other clustering algorithms. Lu et al. (2004a) proposed a developed GKA, named Fast genetic k-means algorithm (FGKA). Although GKA an FGKA have better optimum convergence than k-means algorithm, FGKA have faster speed in processing than GKA. FGKA are more flexible in the evolution process, unlike GKA, FGKA permit illegal strings exist. Thus avoiding the cost of illegal string elimination and greatly shrink the running time length. Incremental Genetic K-means Algorithm (IGKA) is an extension to FGKA, it inherits the feature of FGKA of always converge to the global optimum, but it outperforms FGKA when the mutation probability is not large Lu et al. (2004b).

Hierarchy clustering Algorithm Johnson (1967) is a clustering analysis technique that aims to build a tree structure of clusters. It got significantly improved in the early 1980s on the Lance-Williams, and related, dissimilarity update schema (de Rham, 1980; Juan, 1982). Until 1983, Murtagh (1983) introduced a noticeable improvement on algorithms based on the constructing the nearest neighbor chains and reciprocal or mutual NNs (NN-chains and RNNs). This method replicated the results found in the classical in a more exact way, but more computationally expensive. Bayesian hierarchical clustering algorithm outperforms traditional distance-based agglomerative clustering algorithms. As mentioned in the study of Heller and Ghahramani (2005), Bayesian hierarchical algorithm can work out the predictive distribution of a single point, and it merges clusters based on model-based criterion instead of ad-hoc distance metric. Further, it defines a new lower bound on the marginal likelihood of a DPM and become a novel fast bottom-up approximated inference approach for the Dirichlet process. Density-based hierarchical clustering Campello et al. (2013) can extract a simplified tree of significant clusters and it can get flat partition composed of only the most significant clusters.

Fuzzy-c means clustering (FCM) is a soft algorithm, which allows one observation to not only belong to a single one cluster but a few. FCM was developed by Dunn in 1973 and improved by Bezdek in 1981, it was widely used in many areas including computer science, engineering, mathematics and image recognition. Approximate fuzzy c-means (AFCM) use integer-valued or real-valued estimates to alternative the exact variates in the FCM equation. It directly compute Euclidean distances and exponentiation by exploiting the lookup table. Thus, the CPU time running each iteration gets decreased to about only one sixth of the time required for the original algorithm Cannon et al. (1986). Fuzzy local information C-Means (FLICM)Krinidis and Chatzis (2010) is an efficient image segmentation. It insensitive to noise and preserve the image detail, providing robustness to noisy image. Furthermore, Fuzzy-c means clustering algorithm and its variants also can be applied to many areas. For example, applying FCM into Very Large (VL) data that too big to load into the computer’ s working memory. However, numerical experiments conducted by Havens et al. (2012) shows bit-reduced FCM, approximate kernel FCM perform well in time, space and speed in approximating FCM for VL data. VL data was scale well by these algorithms. For clustering microarray data, FCM is still useful to partition memberships to genes. As Dembele and Kastner (2003) says, partitional clustering method include k–means do not provide the information of the influence of specific genes for the global shape feature of the cluster.

Chapter 4

Data illustration

The dataset in this report is made up of six common food preferences and five different metabolic data from 30 provinces in China. The six dietary preference groups were labeled according to some mainstream cooking method and dietary preference and originally processed by a dietary preference classification model in Zhao et al. (2020) 's report. The data was acquired on the internet, such as some big search engine and three best known online ordering applications in China. After getting the data, in order to visualize the data as user behaviors, Zhao et al. (2020) counted the totality of the clicks and queries that belongs to these six dietary preference then labeled a user as a fan of a certain food preference when his or her clicks or queries exceeded 25%. Metabolic data–BMI (body mass index), FPG(fasting plasma glucose), PPG(postprandial plasma glucose), diabetes prevalence rate and hypertension prevalence rate are five indicators that measure a person's physical fitness, and they are collected from 2010 China Noncommunicable Disease Surveillance.

Before detailed analysis data, we need do preprocessing of the data. Run the *na.omit()* function in R to remove the missing values, and implement the *scale()* function in R to put the values in the same range. The feature scaling function is used to transfer features to the same level of magnitudes. Because the dataset we used are variety in unit, range and magnitude, if we do not normalize it, the features with high magnitudes will have greater weight than features with low magnitudes, then introduced bias would emerge.

This study executed the correlation analysis with six different dietary preferences and some metabolic index on a provincial level. Figure 4.1 high-

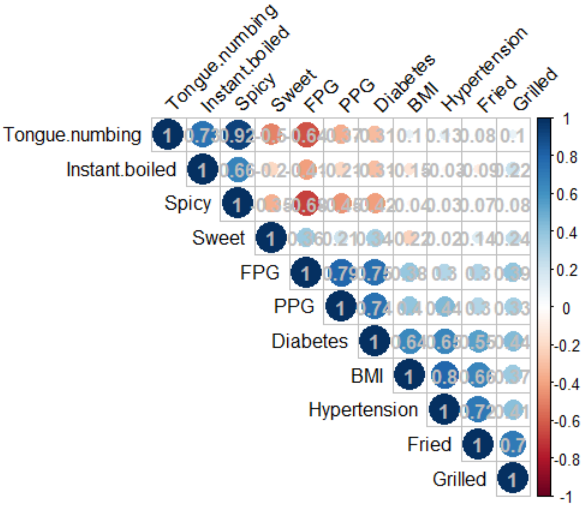
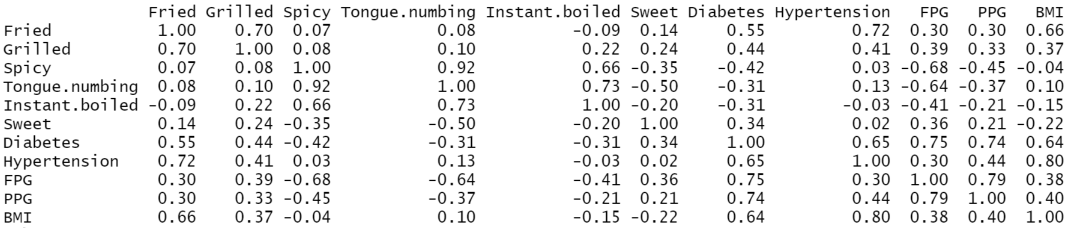


Figure 4.1: Correlation of all variables



```
                 Fried Grilled Spicy Tongue.numbing Instant.boiled Sweet Diabetes Hypertension   FPG   PPG   BMI
Fried             1.00    0.70  0.07           0.08          -0.09  0.14     0.55         0.72  0.30  0.30  0.66
Grilled           0.70    1.00  0.08           0.10           0.22  0.24     0.44         0.41  0.39  0.33  0.37
Spicy             0.07    0.08  1.00           0.92           0.66 -0.35    -0.42         0.03 -0.68 -0.45 -0.04
Tongue.numbing    0.08    0.10  0.92           1.00           0.73 -0.50    -0.31         0.13 -0.64 -0.37  0.10
Instant.boiled   -0.09    0.22  0.66           0.73           1.00 -0.20    -0.31        -0.03 -0.41 -0.21 -0.15
Sweet             0.14    0.24 -0.35          -0.50          -0.20  1.00     0.34         0.02  0.36  0.21 -0.22
Diabetes          0.55    0.44 -0.42          -0.31          -0.31  0.34     1.00         0.65  0.75  0.74  0.64
Hypertension      0.72    0.41  0.03           0.13          -0.03  0.02     0.65         1.00  0.30  0.44  0.80
FPG               0.30    0.39 -0.68          -0.64          -0.41  0.36     0.75         0.30  1.00  0.79  0.38
PPG               0.30    0.33 -0.45          -0.37          -0.21  0.21     0.74         0.44  0.79  1.00  0.40
BMI               0.66    0.37 -0.04           0.10          -0.15 -0.22     0.64         0.80  0.38  0.40  1.00
```

Figure 4.2: Correlation coefficient of all variables

lighted a high and positive correlation between the tongue numbing food and spicy food, which reach up to r equals to 0.92. Correlation analysis reported the proportion of tongue numbing food preference was significantly negatively correlated with FPG (r=-0.64) and also negatively correlated with PPG (r=-0.37) and diabetes prevalence (r=0.31). The same result was found in instant - boiled food preference and spicy food, consuming instant-boiled food can decrease FPG, PPG and diabetes with correlation coefficient r=-0.41, r=-0.21, r= -0.31 respectively, taking in spicy food can significantly decrease the value of FPG with correlation coefficient r=-0.63 and r=-0.45 for PPG, r=-0.42 for diabetes prevalence. There was a negative correlation between sweet food preference and BMI (r = 0.22). On the contrary, fried food preference has strong positive correlation with hypertension (r=0.72), BMI (r=0.66), diabetes (r=0.55). In addition, it also positively correlated with FPG and PPG (r=0.3 both). Grilled food shows the similar pattern

Table 4.1: Full ranges for all variables

observation.	Range	units
Fried	0.168-0.261	in percentage
Grilled	0.156-0.27	in percentage
Spicy	0.27-0.443	in percentage
Tongue.numbing	0.123-0.25	in percentage
Instant.boiled	0.106-0.203	in percentage
Sweet	0.363-0.522	in percentage
diabetes prevalence	0.058-0.227	in percentage
hypertension prevalence	0.24-0.511	in percentage
FPG	4.67-6.06	mmol/L
PPG	5.37-6.91	mmol/L
BMI	22.3-25.7	kg/m2

as fried food, it was positively associate with diabetes prevalence, hypertension prevalence, BMI, FPG, and PPG with correlation coefficient r=0.44, r=0.41, r=0.37, r=0.39, r=0.33 respectively.

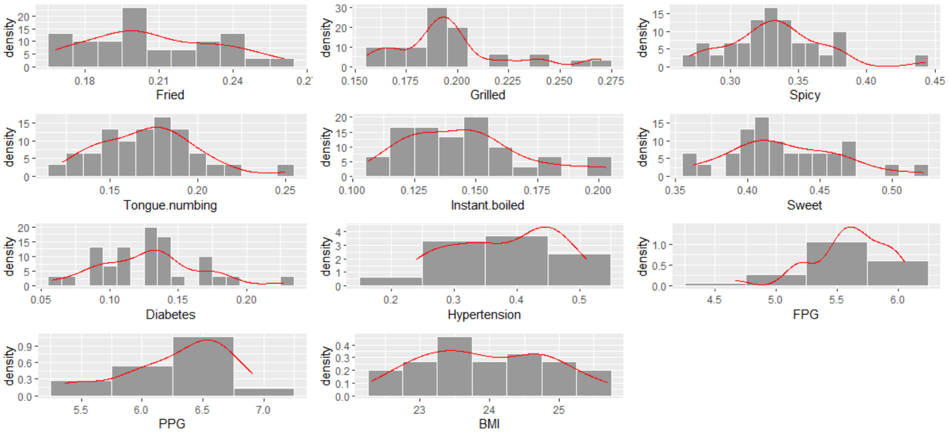


Figure 4.3: Density plot

Density Plot helps display where values are concentrated over the interval. As shown from the first six subplots in figure 4.3, each column is probability density that demonstrates the percentage of the population in each province that prefers a particular diet. For example, the peak of density plot of fried food means there are 20 percent of the people like to eat fried food in the most provinces. Further, shown in spicy food preference subfigure, there is an isolated bin on the right, this is because normally the data range on the x axis goes from 30% to approximately 37%. Therefore, we expect the mean height of the density curve to be (30%+37%)/2 = 33.5%, that is generally about 33.5% people in each province like spicy taste food, however, there is a special province where 44% citizens have a spicy food preference. This phenomenon also can be seen in the tongue numbing food preference and sweet food preference, namely, there is one or two provinces where people give a particularly preference on this certain food compared with the most provinces. Density plot also shows which foods are more popular at the national level. sweet food is the most popular food (most bars concentrate on 0.41) and the spicy food is in second place (around 0.33). As for the rest, people show more or less similar preference on grilled food, tongue numbing and instant boiled food. The plasma glucose before dinner (FPG) usually distributed around 5.8 mg/dL and the plasma glucose after dinner (PPG) mostly distributed in 6.5 mg/dL. Most Chinese have the BMI index between 22 and 26, people with 23.5 BMI index take up the highest proportion.

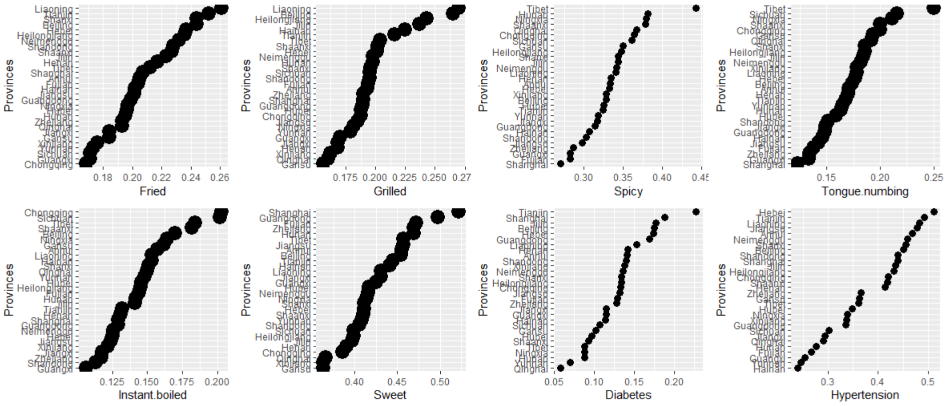


Figure 4.4: Plot the ordered countries for the variables

Figure 4.4 shows the people's dietary preference in 30 province in China. Predictably, tongue - numbing food preference was mainly located in the region where centered on Sichuan province, such as NingXia, Shaanxi, Chongqing. The highest proportion region of spicy food was also centered on Sichuan province and the neighborhood western cities. However, contrary to expectation, Statistics show that Xizang is the province with the highest preference in spicy food and tongue numbing food. Shanghai is the region most prefer the sweet food, Gansu shows the least preference on sweet food. Most people who preferred fried food were distributed in northeast of China such as Liaoning, Tianjin. Liaoning also was the region with the highest proportion of grilled food preferences. The province with the largest proportion of diabetes patients is Tianjin, followed by Shanghai and Beijing. Over half of people in Hebei suffer from hypertension, Hainan is the province with the least proportion of hypertension prevalence.

Chapter 5

Methodology

5.1 Association rule

Association rule is a technique to find correlations and co-occurrences between data seemingly independent relational databases or other data repositories, and it states a pattern that when an event occurs, another event occurs with certain probability.

An association rule has two parts: antecedent and consequent, both of which are a list of items. An antecedent is something that found in data, and a consequent is an item that is found in combination with the antecedent. Here we represent antecedent as A, and consequent as B (A , B are mutually exclusive events). And then define an association rule, called A⇒ B.There are three basic criteria, support confidence and lift, which can identify the relationships and rules through analyzing data. These criteria can be written as a form of joint probability:

❶ Support $(A \Rightarrow B) = P(AB)$

❷Confidence $(A \Rightarrow B) = P(B/A)=P(AB)/P(A)$

❸Lift $(A \Rightarrow B) = P(B/A)/P(B)=\text{confidence}((A \Rightarrow B)/P(B)$

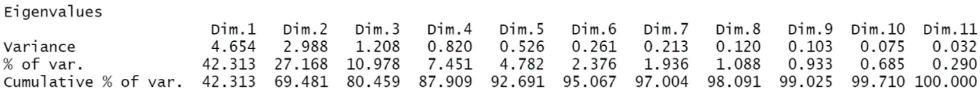
Furthermore, in a given dataset, association rules with support and confidence always be required to greater than, or equal to a user-specified minimum support threshold MinSup, and a confidence threshold MinConf, respectively.

There are several algorithms about the association rules, Apriori algorithm, Eclat algorithm and FP-growth algorithm. We mainly use the first one, Apriori algorithm is the first frequent itemset minin algorithm. It was later improved by R Agarwal and R Srikant and came to be known as Apriori. Apriori algorithm is an iterative process to find the most frequent itemset in the given dataset, frequent here are defined as when the probability of a certain item greater than or equal to the minimum support threshold, otherwise, the itemsets would be consider as unfrequent.

Overall, Apriori algorithm scans the datasets just once that verified its high effectiveness. It greatly cut down the size of the item sets in the database, thus providing good performance. Association Rule is also play a significant role in many application areas, including market basket analysis, recommendation analysis, bioinformatics, Web usage mining and so forth. It has helped data scientists find out patterns they never knew existed.

5.2 Principal component analysis

Principal component analysis (PCA) is a widely used statistical technique applied to a large dataset of multi-dimensional variables where there are much common and difficult to interpret. PCA has considerably utility in increasing the interpretability and minimize the information loss at the same time, it reducing a large number of observed variables down to a few number of factors, which called principal components. Algebraically, Principal components are particular linear combinations of several observed variables, where each linear combination is a factor. These factors are uncorrelated and successively maximize variance. Therefore, some small variances of the indices will be so low as to be negligible, only important factors will be retained as the interest of measuring the underlying dimensions in the data. Run the *function PCA* in the R package FactoMineR to print the principal components.



```
Eigenvalues
                       Dim.1  Dim.2  Dim.3  Dim.4  Dim.5  Dim.6  Dim.7  Dim.8  Dim.9 Dim.10  Dim.11
Variance               4.654  2.988  1.208  0.820  0.526  0.261  0.213  0.120  0.103  0.075   0.032
% of var.             42.313 27.168 10.978  7.451  4.782  2.376  1.936  1.088  0.933  0.685   0.290
Cumulative % of var.  42.313 69.481 80.459 87.909 92.691 95.067 97.004 98.091 99.025 99.710 100.000
```

Figure 5.1: PCtable

As shown in PCtable, PC1, PC2, PC3 are first principal component, second principal component, third principal component, which are statistically independent representation of datasets and ranked by their size of proportion of variance. The first principal component has the maximum variance, which account for 42.31% of total variance. The proportion of total variance explained by the second principal component is 27.17%. The cumulative proportion of variance is 69.48%, which carry large enough information. Therefore, we retain first two principal components that explain a significant portion of the data variance. Another straightforward way to determine the number of the principal components is scree plot.

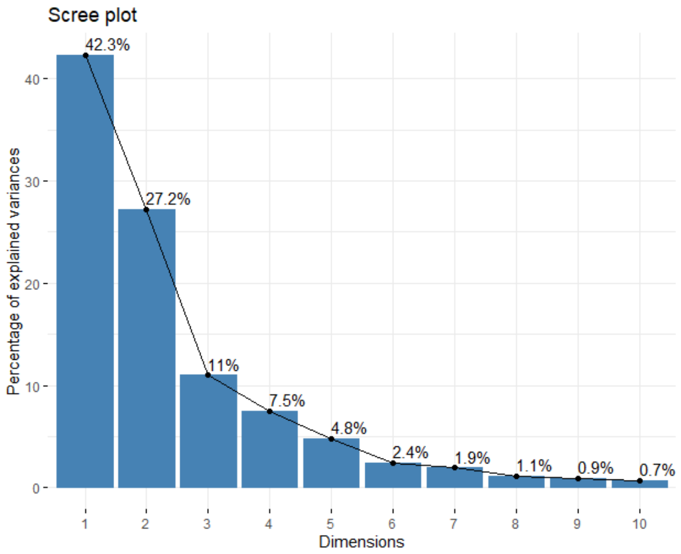
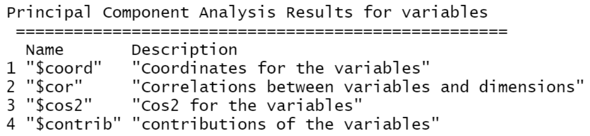


Figure 5.2: Proportion of variance graphs

Figure 5.2 is a scree plot, this is method was initially proposed by Cattell (1966). It visualizes the representation of the eigenvalues and sort the eigenvalues in a descending order then join the midpoints to one line. Sarmento and Costa (2017) It purposes to find a large cut in the size of eigenvalues, with the rest of the smaller eigenvalues aggregating rubbles. Therefore, in our model we consider two components, because the elbow in this plot starts a less steep decline in the second place, which account for 69.5% of the explained variance. Geometrically, these factors represent a new coordinate system obtained by rotating the factors, which increase the interpretability. The factor summarizes the patters of correlations in the observed correlation matrix. The function *get_pca_var()* can be used to extract the results for variables, the function provide a list of matrices involving all the results from the active variables, such as coordinate, correlation between variables and axes, square of correlation coefficient and contributions of a certain variable. As shown below:

Run function *fviz_pca_var()* to visualize the coord and cor between



```
Principal Component Analysis Results for variables
 ===================================================
  Name       Description                                    
1 "$coord"   "Coordinates for the variables"                
2 "$cor"     "Correlations between variables and dimensions"
3 "$cos2"    "Cos2 for the variables"                       
4 "$contrib" "contributions of the variables"               
```

Figure 5.3: extract the results for variables

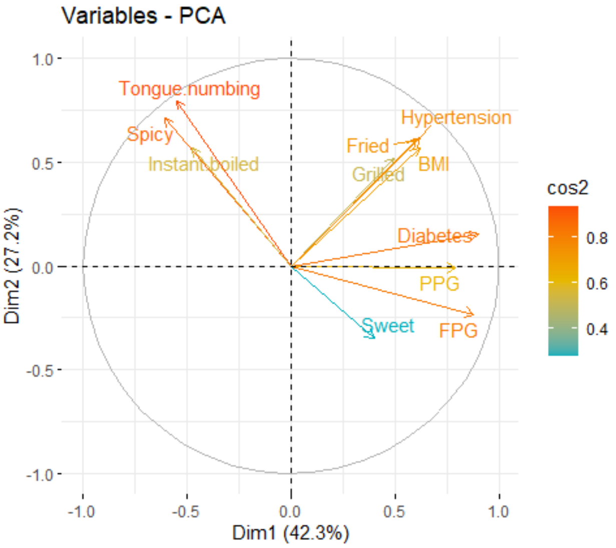


Figure 5.4: loading plot

variables. Coord means coordination that also represnet the true loading. Hence, the figure 5.4 is a loading plot, which shows how strongly each characteristic influences a principal component. See how these vectors far away from the coordinate origin. Their project values on each PC illustrate how much weight they have on that PC. In our dietary preference example, hypertension, BMI, diabetes, FPG, fried and grilled food strongly positive influence PC1,whereas the arrows of Tongue numbing food and spicy food go in the opposite direction. It means they have adverse impact on PC1. Similarly, observing the PC2, FPG and sweet food preference both have negative influence on PC2, but FPG has a stronger effect than sweet food. All the remaining objects have the positive influence on PC2 except the PPG whose projection on PC2 is about 0.0. In addition, the observation on angles between two vectors are also valuable. When two vectors are close, forming a small angle, this represent that these two variable are positively correlated. For example, hypertension and fried food are highly correlated with each other. Besides, when two variables meet each other at 90°, they are likely to be uncorrelated, such as spicy food and BMI. Furthermore, when there is a 180° angle, we say two variables are negative correlated, such as sweet food and instant boiled food.

These results can be seen clearly by executing function *corrplot* and *fviz_cos2*. The shades and sizes of the circles in figure 5.5 demonstrate the correlation between variables and axis (PC). The darker the color, the more relevant it is.

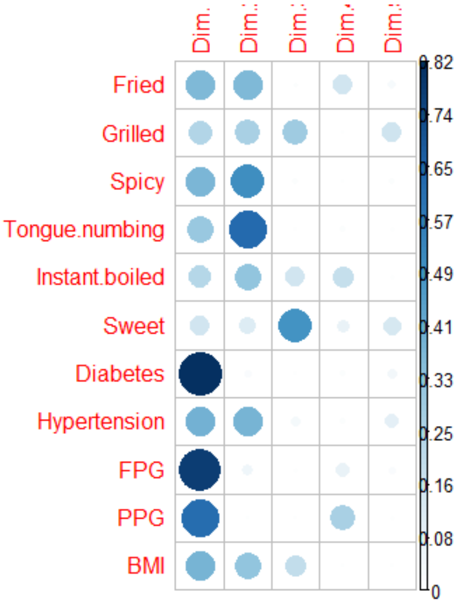


Figure 5.5: correlation between variables and PCs

Figure 5.6 compared the absolute value of each variable which is much easier to see who has the most influence on the principal component. Tongue numbing food preference, spicy food preference and diabetes top the list, then FPG and hypertension rank the fourth place and fifth place respectively.

Figure 5.7 is score plot and biplot. Clearly, we find that PCA biplot simply merge a PCA score plot with a loading plot. The score plots project the variable vectors onto the span of the principal components. The first two PCs is shown in the left subfigure, and it uses different scales for the axes. There are no clearly separated clusters, but notice that sample 25 and 20 look like outliers. The loading plot indicates the correlation between variables and principal components. Biplot has the functionality of these

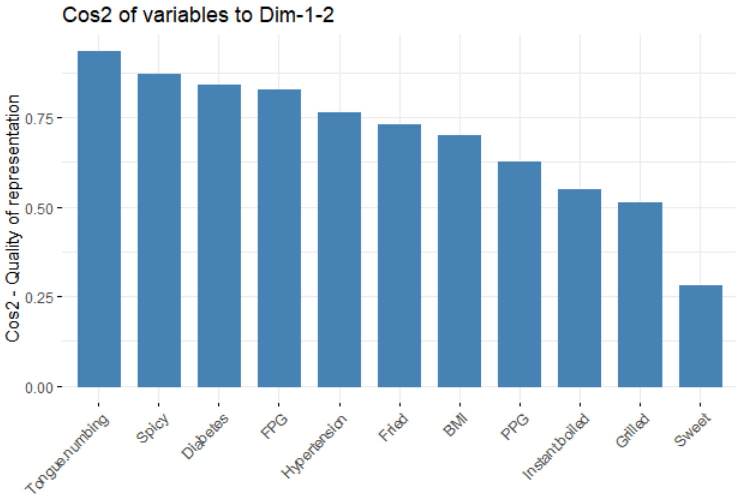


Figure 5.6: proportion of explained variance

two graphs, further it also illustrates how 30 province correlated with six food preferences and 5 metabolic indexes. Notice that Liaoning, Beijing, Heilongjiang, Shanxi, Hebei, Tianjin, they have similar food preference and similar physical condition. They eat much fried food and grilled food in daily life, and need to be more care of their BMI and the risk of hypertension prevalence. Moreover, we see people in Shanghai and Guangdong prefer the sweet food and people in Shaanxi, Chongqing, Sichuan, Ningxia have a dietary preference of instant boiled food and numb spicy food. And people in Yunnan provience are not biased in dietary preference.

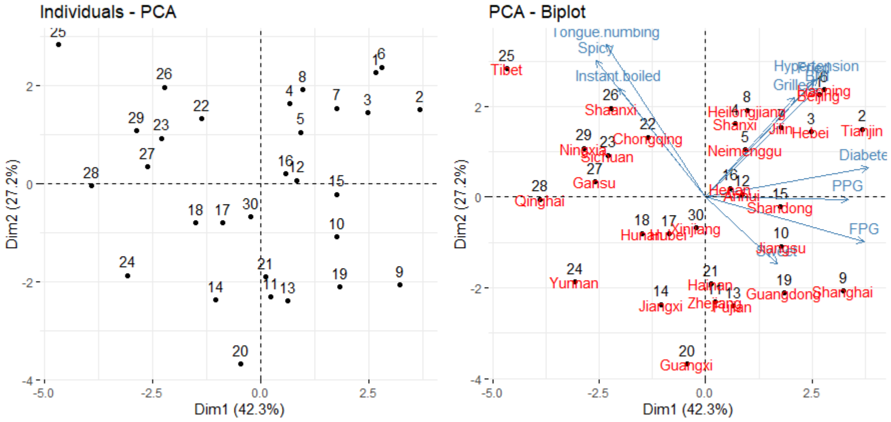


Figure 5.7: score plot and biplot

However, as Tabachnick et al. (2007) metioned in his book "Using Multivariate Statistics" , there are some limitations and issues caused by the sensitivity of PCA to the size of correlations, such as problems arise from missing data, sensitivity to outlying cases and degradation of correlations between poorly distributed variables.

❶ sample size and missing data
When data is extract from small sample, correlation coefficients tend to be less reliable. Hence, in order to obtain reliable estimated correlations, it is critical to choose a large enough sample size. Furthermore, the case of having missing data need to be taken into consideration, consider what the missing values distribution is and decision between estimation of missing variable and deletion of missing variable.

❷ linearity
The relationship between pairs of variables should be linear, otherwise the analysis is degraded due to the disability of coefficient in reflecting the nonlinear relationship. The method of screening for linearity is Scatterplot in R, when it inspect nonlinearity exist, transformation of variables is required.

❸ normality
The assumption of multivariate normality is tenable. If normal distribution hypothesis of variables is hold, the solutions is enhanced. If not, the solution is degraded.

❹ absence of outliers among cases
When there are outliers in our study, they will have more influence on the factor solutions than other variables. Therefore, detecting and reducing the effect of univariate and multivariate outliers is necessary.

5.3 Clustering Algorithm

Cluster analysis is a class of techniques that are used to group or segment a collection of objects or cases into subsets such that the similarity between those within the same cluster are greater than objects assigned to different clusters.

5.3.1 Hierarchy clustering

Partitioning algorithms based on specifying an initial number of groups and iteratively reassignment observations among groups to convergence. On the contrary, hierarchical clustering combines or divides existing groups, grouping data points into a tree of cluster. It creates the hierarchical representations of clusters at each level of the hierarchy and merge clusters to the next lower level.

Hierarchical clustering strategies can classified into agglomerative (bottom-up) and divisive (top-down) hierarchical clustering. Agglomerative approach start at the bottom, each node represents a single cluster, and at each step two clusters which have the smallest intergroup dissimilarity would merge into the same cluster. Divisive methods start at the top and at each step split one cluster into two new clusters according to choosing the largest between-group dissimilarity.

Hierarchy clustering requires distance measures (i.e Eucliden distance, Manhattan distance, etc.) to measure the dissimilarity between each pair of observations. The Euclidean distance between two vectors $x_i = (x_{i1}, x_{i2}, \ldots, x_{iM})$ and $x_j = (x_{j1}, x_{j2}, \ldots, x_{jM})$ is defined as:

$$d(x_i, x_j) = \sum_{m=1}^{M} (x_{im} - x_{jm}^2) \tag{5.1}$$

The Manhattan distance measures distance in the number of horizontal and vertical units, represented by :

$$d(x_i, x_j) = \sum_{m=1}^{M} |x_{im} - x_{jm}| \tag{5.2}$$

Define the dissimilarity d(G,H) between two clusters G and H, the dissimilarities between pairwise observation denoted as d_{ii}, where one member of the pair i is in G and the other $i^{'}$ is in H. There are three methods of agglomerative hierarchical clustering, including Single Linkage (SL) agglomerative clustering, Complete linkage (CL) agglomerative clustering and Group average (GA) clustering.

❶ Single linkage (SL) agglomerative clustering measures the intergroup dissimilarity between the closest pair.

$$d_{SL}(G, H) = \min_{\substack{i \in G, \\ i' \in H}} d_{ii'} \tag{5.3}$$

❷ Complete linkage (CL) agglomerative clustering takes the intergroup distance as that of the farthest pair.

$$d_{CL}(G, H) = \max_{\substack{i \in G, \\ i' \in H}} d_{ii'} \tag{5.4}$$

❸ Group average (GA) clustering denote the average dissimilarity between the groups.

$$d_{CL}(G,H) = \frac{1}{N_G N_H} \sum_{i \in G} \sum_{i' \in H} d_{ii'} \tag{5.5}$$

Dendrogram provides a way of visualizing the Hierarchical clustering result, it constructs a tree-like diagram that records the series of merges and splits. The cophenetic correlation coefficient can be used to represent what the extent the hierarchical structure in dendreogram can represent the data itself. This is the correlation defined between the N(N1)/2 pair-wise observation dissimilarities $d_{ii'}$, and the corresponding cophenetic dissimilarities $C_{ii'}$, . The cophenetic dissimilarity obeys the ultrametric inequality: $C_{ii'} \leq \max_{C_{ik},C_{i'k}}$, In this data set, we tried four different linkage methods built in hclust function.“Single” measurement method will cause the chaining problem, which will make the cluster result invalid. “Complete”, “Ward.D”, “Average” produced the similar results. Figure 5.8 shows the cluster pattern of “Ward.D” linkage measure.

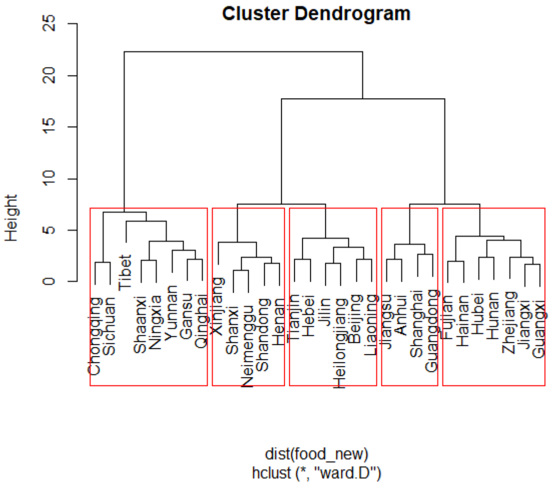


Figure 5.8: Hierarchical clustering with “wald” linkage measure

The function *cutree* divided the clusters into 5 groups, we find that the province in each cluster have the similar geographical region, which illustrate that people from neighborhood area tend to have the same food preference. Observing the figure 5.8 combined with the Figure 5.3 “Plot the ordered countries for the variables”, from left to right, cluster 1 province tend to consume more spicy and tongue numbing food. Cluster 2 provinces show medium preference in fried food and grilled food, however among this 5 province, Xinjiang show relatively lower interest in this two type of food. Cluster 3 contains provinces with higher preference of grilled food but lower interest in sweet food. Province classified into Cluster 4 are prefer sweet food but show less favor of the tongue numbing, province in cluster 5 also show higher preference in sweet food but lower preference in spicy food.

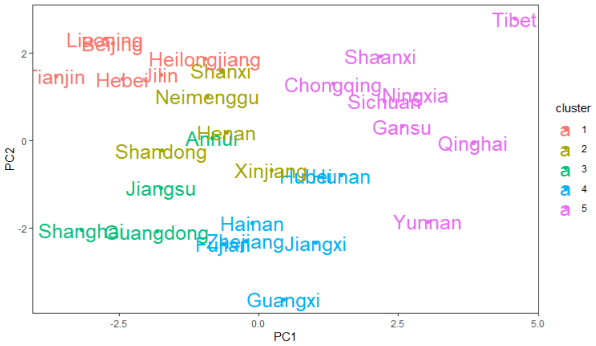


Figure 5.9: Projection the clusters onto the first two principal components

To visualize the hierarchical clustering, we project the data onto two principal components. Figure 5.9 shows the data in PC1 and PC2. We can see the cluster 5 are entirely separated with any other clusters in the dimension of PC1. Cluster 1, 2, 3 and 4 are mixed in the center of the dimension. Cluster 2 and 1 are mixed around PC1 eqauls -0.5, PC2 equals 1.5, it means Heilongjiang province and Shanxi Province are similar in dietary preference in some extend in PC1. Furthermore, cluster 2 and cluster 3 have overlap in the position of Henan and Anhui. And cluster 2 and 4 are intersect in the position of PC1= 0.5, PC2= -1. These results also can be seen from the pairwise plot of Figure 5.10. The upper triangular part numerically interpret the result, it shows the correlation coefficient between five clusters. The scatter plot on the lower triangle graphically demonstrate the relationship between two variables. The density plot on the diagonal allows us to see the distribution of a single principal component.

5.3.2 K-means clustering

K-means clustering is the most widely used unsupervised machine learning algorithm for partitioning a set of n objects into k $\geq$ 2 clusters, such

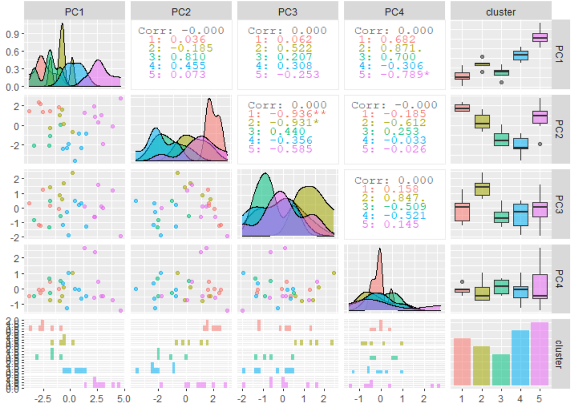


Figure 5.10: Pairwise Plot

that the objects in the same cluster are as similar as possible and distinct to others in the different clusters. It is an iterative process which aims to minimize the within-cluster variances, ie. minimize the sum of distance between the points and respective cluster centroid. K-means group the data points by alternating between

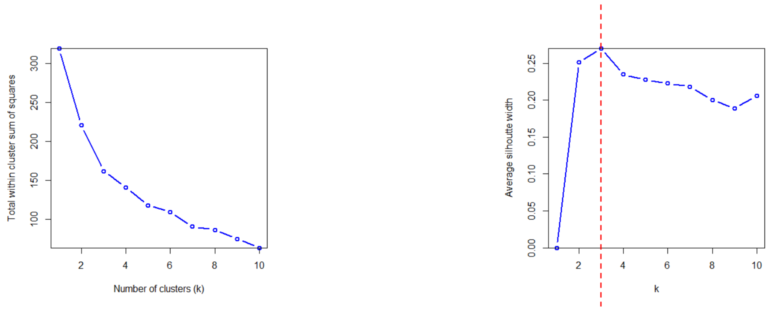
❶ Assigning each object to the the nearest cluster that with the least squared Euclidean distance. (Mathematically, For a given cluster assignment C, express the initial set of K centroids as $\{m_1, m_2, \cdots, m_k\}$, then the process of finding the smallest in-cluster distance can be displayed as:

$$C(i) = \underset{\mathbf{1 \leq k \leq K}}{\arg\min} \quad \|x_i - m_k\| \tag{5.6}$$

❷ Update the centroid based on the recalculation of the mean of all data points in the cluster. Iterating and repeating the assignment of points to the cluster until points stop changing clusters. The algorithm obtain the optimization $C*$ is:

$$C* = \min_{C,\{m_k\}} \sum_{k=1}^{K} N_k \sum_{C(i)=k} \|x_i - m_k\|^2 \tag{5.7}$$

There are two methods to optimal value of k (i.e. the number of cluster) in a data set: Elbow method and Sihouette coefficient method. Elbow method plot the explained variation as a function of the number of clusters and pick the turning point of the curve as the value of k. Another method to check the optimal number of clusters is Silhouette coefficient, it shows how the data well matched to its own cluster and it ranges from -1 to 1. The value near to 1 indicates the observation is well clustered, and a small value indicates the observation lies between two clusters, and a negative value of silhouette means the observation probably in the wrong cluster. Furthermore, the silhouette score also can be used to validate the working of clustering algorithm when there are high dimensions. Mola (2015)



(a) Total with cluster sum of squares (b) Average silhouette for k=1 to 10

Figure 5.11: Choosing the number of cluster

The function *kmeans* obtain the total within cluster sum of squares pattern in figure 5.11(a), which choose the optimal number of cluster at the turning point k=2, because at this point the sum of error square decrease at the largest degree.

In the Silhouette order determination method, the average silhouette is maximized at k=3. Hence, the silhouette coefficient method obtains the number of cluster at 3. The clusters under the Silhouette method contain the observation of 8,12,10 respectively. And the average silhouette width is 0.27. All the results are shown in figure 5.12.

The clustering result of K-means is shown in Figure 5.13. A total of 30 datapoints has been divided into three clusters, but there is mixture in the center of the plot which means k-means clustering not seperated this dataset well.

5.3.3 Fuzzy C-means Clustering

Fuzzy c-means clustering (FCM) is a soft algorithm used to cluster multidimensional fuzzy data in which an object is not only a membership of a cluster centered from 0 to 100 percent but member of many clusters

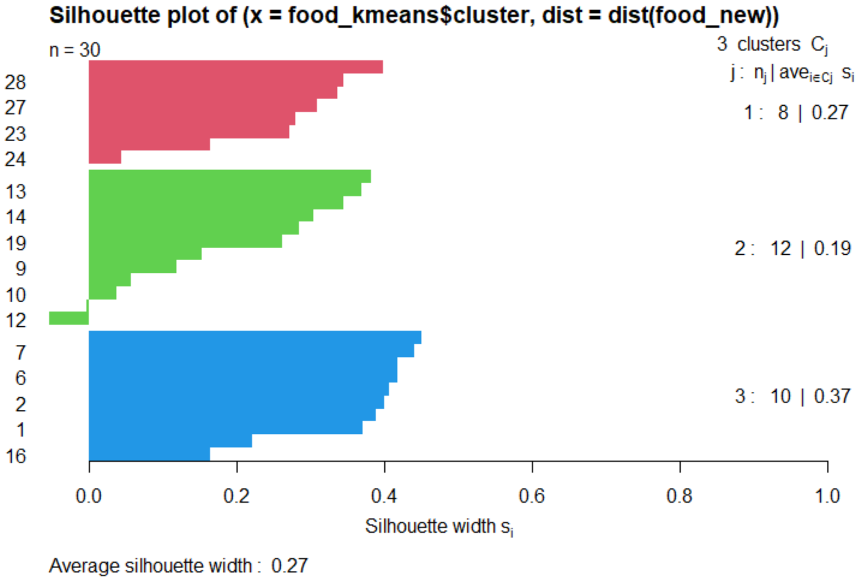


Figure 5.12: Silhouette plot

in varying degree of memberships as well. Compared with the traditional hard-threshold clustering where every data point assigned to a crisp, exact label. FCM increase the expressiveness of the clustering analysis and using a membership matrix to successively provide a more comprehensive view of relationship present in the dataset. Membership matrix computes membership degrees at each iteration, which gives each membership degree a proportional manifestation to its proximity to the representitives. This method mitigates the problem that presented in K-means cluster that inadequate hard assign the data points randomly to cluster or another when the points are equally distance between two cluster centers.Döring et al. (2006)

In R environment, Fuzzy c-means clustering is accomplished via *skfuzzy.cmeans*, and the output of this function can be repurpose to do predication-classifying new data according to computed clusters via *skfuzzy.cmeans_predict*.

Main objective of fuzzy c-means algorithm is to minimize:

$$J(U,V) = \sum_{i=1}^{n}\sum_{j=1}^{c}(\mu_{ij})^m\|x_i - v_j\|^2) \tag{5.8}$$

Figure 5.14 shows the membership grade and visualise it in corrplot. The higher degree of the membership, the darker the dot is, the closer it gets

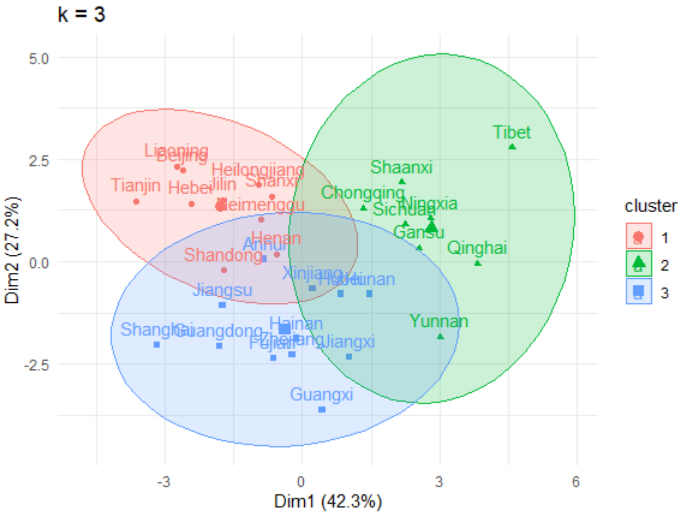


Figure 5.13: Visualization of the K-means clustering in the projection of principal components

Table 5.1: important variable

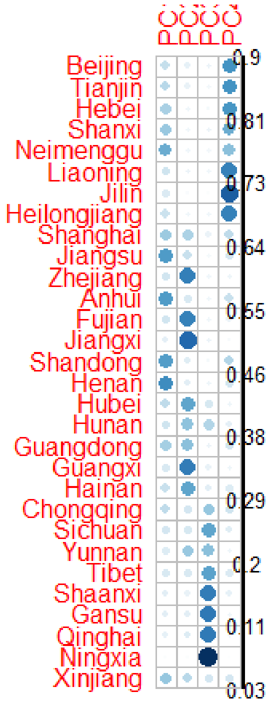
1	n	the number of data points
2	v_j	the jth cluster center
3	m	the fuzziness index
4	c	the number of cluster center
5	μ_{ij}	the membership of ith data to jth cluster
6	$\|x_i - v_j\ \|$	the Euclidean distance between ith data and jth cluster center

to the center of the corresponding cluster. For example, in PC2 Zhejiang, Fujian and Jiangxi are closer to the cluster center than others. In PC3 Ningxia and Gansu are in the center of the cluster.

Graphically, cluster plot in Figure 5.15 seperate 30 provinces' data in China into 4 clusters. Observe that Fuzzy C-means cluster Algorithm perform better in seprating the clusters than K-means clustering, it significantly decrease the size of overlap communities in K-means clustering.

```
                    PC1        PC2        PC3        PC4
Beijing      0.23016538 0.11935944 0.10068444 0.54979075
Tianjin      0.26748927 0.11183494 0.07634624 0.54432955
Hebei        0.30697322 0.08702251 0.06475880 0.54124547
Shanxi       0.35251461 0.08676278 0.09992894 0.46079366
Neimenggu    0.45477593 0.08491624 0.07115086 0.38915697
Liaoning     0.20566871 0.09959757 0.08142324 0.61331048
Jilin        0.16006380 0.05269095 0.04130297 0.74594227
Heilongjiang 0.21675896 0.08114723 0.08316905 0.61892475
Shanghai     0.34622558 0.30063114 0.10329593 0.24984735
Jiangsu      0.52261727 0.22387997 0.06775494 0.18574782
Zhejiang     0.19917439 0.62247453 0.08574705 0.09260404
Anhui        0.52203536 0.17130291 0.08604594 0.22061579
Fujian       0.19991291 0.61453081 0.08046095 0.10509532
Jiangxi      0.13066912 0.71169067 0.09956086 0.05807936
Shandong     0.50936663 0.14467422 0.06464664 0.28131252
Henan        0.57433244 0.13505563 0.09282657 0.19778536
Hubei        0.25271285 0.48181946 0.17589883 0.08956886
Hunan        0.20767399 0.38114828 0.28729366 0.12388407
Guangdong    0.31762729 0.39057539 0.09888136 0.19291596
Guangxi      0.16992089 0.63598033 0.10584179 0.08825699
Hainan       0.22473443 0.48752229 0.13290844 0.15483483
Chongqing    0.25285585 0.18563443 0.35895911 0.20255061
Sichuan      0.18731232 0.18244660 0.48257489 0.14766619
Yunnan       0.15736981 0.35623186 0.39008103 0.09631730
Tibet        0.17086690 0.17106747 0.49799303 0.16007261
Shaanxi      0.13790462 0.11039162 0.63057545 0.12112831
Gansu        0.13518775 0.13187853 0.65007325 0.08286046
Qinghai      0.12074118 0.15145729 0.64403958 0.08376196
Ningxia      0.03695882 0.03621653 0.90000308 0.02682157
Xinjiang     0.36014783 0.26677023 0.18871670 0.18436524
```

(a) Membership degrees matrix



(b) corrplot of membership degree

Figure 5.14: membership grade

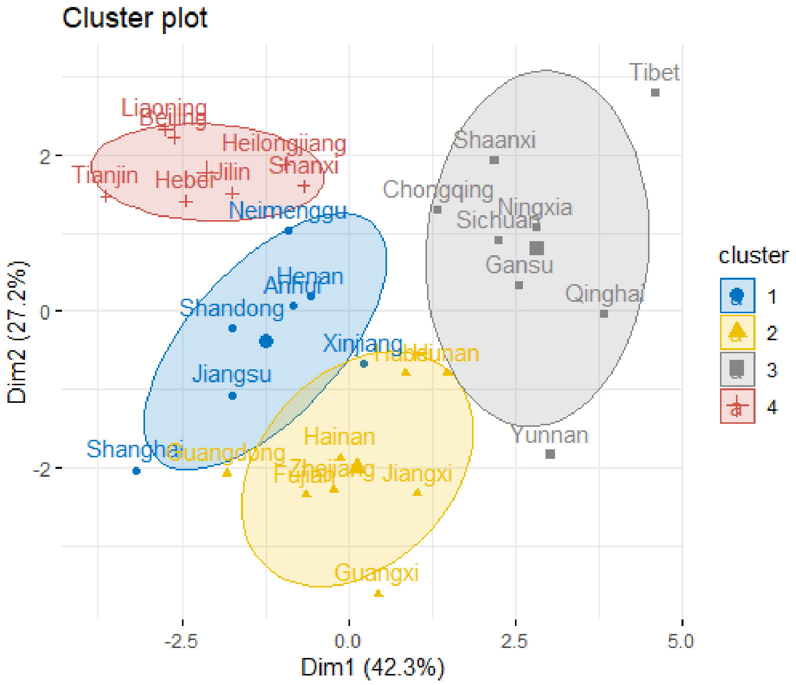

Cluster plot
Liaoning
Beijing
Heilongjiang
Tianjin
Hebei
Jilin
Shanxi
Neimenggu
Henan
Anhui
Shandong
Xinjiang
Hubei
Hunan
Jiangsu
Shanghai
Guangdong
Hainan
Fujian
Zhejiang
Jiangxi
Guangxi
Tibet
Shaanxi
Chongqing
Ningxia
Sichuan
Gansu
Qinghai
Yunnan
Dim2 (27.2%)
2
0
-2
-2.5
0.0
2.5
5.0
Dim1 (42.3%)
cluster
1
2
3
4

Figure 5.15

Chapter 6

Conclusion

In our project, we aim to find some correlations between metabolic diseases and certain kind of food, which province has the highest incidence of the metabolic disease, and how unsupervised learning cluster these 30 province based on the dietary preference data and disease prevalence data.

From the correlation analysis in our study, we find fried food, grilled food and sweet food are main culprit of developing diabetes, whereas diabetes prevalence is negatively correlated to the preference of spicy food, tongue numbing food and instant boiled food. Hence we conclude that increasingly intake the spicy taste food in a diet could decrease the diabetes-caused mortality. Further, people with diabetes are especially vulnerable to high blood pressure. Because greater frequency of fried food and grilled food consumption also contribute a lot to hypertension prevalence. Spicy food preference is so similar to tongue - numbing food preference in dietary taste and they also have a strong negative relationship with FPG and PPG. This result is consistent with Zhao et al. (2020)'s opinion: “spicy food may reducing oxidative stress and thus increase the metabolism and decrease the prevalence of chronic disease” . Besides, the group of people who have hypertension and diabetes often accompanied by high BMI value. Therefore, people should choose food carefully, avoid eating too much fried and grilled food could protect you from many metabolic diseases.

In our research, we find that sweet food has the highest proportion of fans in any province and spicy food is the second popular food. Most of Chinese people have 23.5 BMI value. People live in the near geographical area have similar dietary preference. Our research shows that the people in the northeast of China prefer fried food and grilled food than all others, especially Liaoning and Tianjin. So people who live there need more careful about developing hypertension, diabetes and obesity. The region with highest proportion of spicy food is centered in Sichuan province and some neighborhood like Chongqing and Ningxia. Tongue numbing food preference was also mainly distributed in these midwest cities. Beyond the imagination, Xizang is the highest proportion province of spicy food preference and tongue numbing food preference. People in southeast coastal cities are favor sweet food, especially the Shanghai where have a highest proportion of sweet food. People there should beware of diabetes. However, our statistic shows that the province with the largest proportion of diabetes patients is Tianjin, followed by Shanghai and Beijing. People least likely to develop diabetes are those in Qinghai. Citizens in Hebei should more care about the risk of hypertension, over half of people in Hebei have high blood pressure.

Implement different clustering Algorithm will lead to distinct clustering result. There are five clusters in hierarchical clustering. The clustering result looks well overall, provinces in close geographic proximity tend to have same food preference and physical condition. Liaoning, Beijing, Heilongjiang, Hebei, Tian jin, Jilin are clustered in the same group, But in K-means clustering method, Shanxi, Neimenggu, Henan, Shandong also belong to this cluster, and altogether there are only three clusters. These three clusters are mixed in the center, which shows clustering outcome of k-means is not very desirable. Compared with the traditional hard threshold clustering algorithm—k-means clustering, fuzzy-c clustering algorithm has better clustering result of multidimensional data. It clustered 30 provinces into 4 groups and mitigate the problem that presented in k-means clustering that has large size of overlapped area.

Overall, these three clustering algorithm obtain some common results in grouping. Hierachy clustering and k-means clustering perform well in the first principal component and the second principal component , they split the members neat with not a lot of overlap. However, K-means clustering shows a considerate degree of overlap of three clusters.

Bibliography

Atlas, D. (2015). International diabetes federation. *IDF Diabetes Atlas, 7th edn. Brussels, Belgium: International Diabetes Federation.*

Cai, W., Ramdas, M., Zhu, L., Chen, X., Striker, G. E., and Vlassara, H. (2012). Oral advanced glycation endproducts (ages) promote insulin resistance and diabetes by depleting the antioxidant defenses age receptor-1 and sirtuin 1. *Proceedings of the National Academy of Sciences*, 109(39):15888–15893.

Campello, R. J., Moulavi, D., and Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer.

Cannon, R. L., Dave, J. V., and Bezdek, J. C. (1986). Efficient implementation of the fuzzy c-means clustering algorithms. *IEEE transactions on pattern analysis and machine intelligence*, (2):248–255.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate behavioral research*, 1(2):245–276.

Dembele, D. and Kastner, P. (2003). Fuzzy c-means method for clustering microarray data. *bioinformatics*, 19(8):973–980.

Donfrancesco, C., Noce, C. L., Brignoli, O., Riccardi, G., Ciccarelli, P., Dima, F., Palmieri, L., and Giampaoli, S. (2008). Italian network for obesity and cardiovascular disease surveillance: a pilot project. *BMC family practice*, 9(1):53.

Döring, C., Lesot, M.-J., and Kruse, R. (2006). Data analysis with fuzzy clustering methods. *Computational Statistics & Data Analysis*, 51(1):192–214.

Folch-Fortuny, A., Arteaga, F., and Ferrer, A. (2016). Assessment of maximum likelihood pca missing data imputation. *Journal of Chemometrics*, 30(7):386–393.

Gao, Y., Chen, G., Tian, H., Lin, L., Lu, J., Weng, J., Jia, W., Ji, L., Xiao, J., Zhou, Z., et al. (2013). Prevalence of hypertension in china: a cross-sectional study. *PloS one*, 8(6):e65938.

Ge, Z., Yang, C., and Song, Z. (2009). Improved kernel pca-based monitoring approach for nonlinear processes. *Chemical Engineering Science*, 64(9):2245–2255.

Guallar-Castillón, P., Rodríguez-Artalejo, F., Fornés, N. S., Banegas, J. R., Etxezarreta, P. A., Ardanaz, E., Barricarte, A., Chirlaque, M.-D., Iraeta, M. D., Larrañaga, N. L., et al. (2007). Intake of fried foods is associated with obesity in the cohort of spanish adults from the european prospective investigation into cancer and nutrition. *The American journal of clinical nutrition*, 86(1):198–205.

Havens, T. C., Bezdek, J. C., Leckie, C., Hall, L. O., and Palaniswami, M. (2012). Fuzzy c-means algorithms for very large data. *IEEE Transactions on Fuzzy Systems*, 20(6):1130–1146.

Heller, K. A. and Ghahramani, Z. (2005). Bayesian hierarchical clustering. In *Proceedings of the 22nd international conference on Machine learning*, pages 297–304.

Johnson, R. J., Perez-Pozo, S. E., Sautin, Y. Y., Manitius, J., Sanchez-Lozada, L. G., Feig, D. I., Shafiu, M., Segal, M., Glassock, R. J., Shimada, M., et al. (2009). Hypothesis: could excessive fructose intake and uric acid cause type 2 diabetes? *Endocrine reviews*, 30(1):96–116.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254.

Khan, T. A. and Sievenpiper, J. L. (2016). Controversies about sugars: results from systematic reviews and meta-analyses on obesity, cardiometabolic disease and diabetes. *European journal of nutrition*, 55(2):25–43.

Krinidis, S. and Chatzis, V. (2010). A robust fuzzy local information c-means clustering algorithm. *IEEE transactions on image processing*, 19(5):1328–1337.

Krishna, K. and Murty, M. N. (1999). Genetic k-means algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 29(3):433–439.

Likas, A., Vlassis, N., and Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern recognition*, 36(2):451–461.

Lu, Y., Lu, S., Fotouhi, F., Deng, Y., and Brown, S. J. (2004a). Fgka: A fast genetic k-means clustering algorithm. In *Proceedings of the 2004 ACM symposium on Applied computing*, pages 622–623.

Lu, Y., Lu, S., Fotouhi, F., Deng, Y., and Brown, S. J. (2004b). Incremental genetic k-means algorithm and its application in gene expression data analysis. *BMC bioinformatics*, 5(1):1–10.

Lv, J., Qi, L., Yu, C., Yang, L., Guo, Y., Chen, Y., Bian, Z., Sun, D., Du, J., Ge, P., et al. (2015). Consumption of spicy foods and total and cause specific mortality: population based cohort study. *Bmj*, 351:h3942.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Mola, A. (2015). Development and validation of a bangladeshi pediatric silhouette scale (bpss).

Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms. *The computer journal*, 26(4):354–359.

Nakajima, S., Sugiyama, M., and Babacan, S. D. (2011). On bayesian pca: Automatic dimensionality selection and analytic solution. In *ICML*.

Porter, T. M. (2006). *Karl Pearson: The scientific life in a statistical age*. Greenwood Publishing Group.

Qin, P., Liu, D., Wu, X., Zeng, Y., Sun, X., Zhang, Y., Li, Y., Wu, Y., Han, M., Qie, R., et al. (2021). Fried-food consumption and risk of overweight/obesity, type 2 diabetes mellitus, and hypertension in adults: a meta-analysis of observational studies. *Critical Reviews in Food Science and Nutrition*, pages 1–12.

Sarmento, R. and Costa, V. (2017). *Comparative approaches to using R and python for statistical data analysis*. IGI Global.

Sayon-Orea, C., Bes-Rastrollo, M., Basterra-Gortari, F., Beunza, J., Guallar-Castillon, P., De la Fuente-Arrillaga, C., and Martinez-Gonzalez, M. (2013). Consumption of fried foods and weight gain in a mediterranean cohort: the sun project. *Nutrition, Metabolism and Cardiovascular Diseases*, 23(2):144–150.

Stanhope, K. L., Schwarz, J. M., Keim, N. L., Griffen, S. C., Bremer, A. A., Graham, J. L., Hatcher, B., Cox, C. L., Dyachenko, A., Zhang, W., et al. (2009). Consuming fructose-sweetened, not glucose-sweetened, beverages increases visceral adiposity and lipids and decreases insulin sensitivity in overweight/obese humans. *The Journal of clinical investigation*, 119(5):1322–1334.

Sun, F., Xiong, S., and Zhu, Z. (2016). Dietary capsaicin protects cardiometabolic organs from dysfunction. *Nutrients*, 8(5):174.

Tabachnick, B. G., Fidell, L. S., and Ullman, J. B. (2007). *Using multivariate statistics*, volume 5. Pearson Boston, MA.

Teff, K. L., Elliott, S. S., Tscho??p, M., Kieffer, T. J., Rader, D., Heiman, M., Townsend, R. R., Keim, N. L., DAlessio, D., and Havel, P. J. (2004). Dietary fructose reduces circulating insulin and leptin, attenuates postprandial suppression of ghrelin, and increases triglycerides in women. *The Journal of Clinical Endocrinology & Metabolism*, 89(6):2963–2972.

Wagstaff, K., Cardie, C., Rogers, S., Schroedl, S., et al. (2001). Constrained k-means clustering with background knowledge. In *Icml*, volume 1, pages 577–584.

Wang, Z., Chen, Z., Zhang, L., Wang, X., Hao, G., Zhang, Z., Shao, L., Tian, Y., Dong, Y., Zheng, C., et al. (2018). Status of hypertension in china: results from the china hypertension survey, 2012–2015. *Circulation*, 137(22):2344–2356.

Xu, Y., Wang, L., He, J., Bi, Y., Li, M., Wang, T., Wang, L., Jiang, Y., Dai, M., Lu, J., et al. (2013). Prevalence and control of diabetes in chinese adults. *Jama*, 310(9):948–959.

Yang, J., Zhang, D., Frangi, A. F., and Yang, J.-y. (2004). Two-dimensional pca: a new approach to appearance-based face representation and recognition. *IEEE transactions on pattern analysis and machine intelligence*, 26(1):131–137.

Zhao, Z., Li, M., Li, C., Wang, T., Xu, Y., Zhan, Z., Dong, W., Shen, Z., Xu, M., Lu, J., et al. (2020). Dietary preferences and diabetic risk in china: A large-scale nationwide internet data-based study.