Customer Personality Analysis
for Marketing Segmentation

Junqian Zhang

University of Milan
Data Science and Economics

Abstract. This work aims to finding some customer patterns for marketing segmentation by means of unsupervised learning methods. It employs Principle Component Analysis and K-means to rebuild new feature and simultaneously reduce the dimension of data. And unsupervised learning approach Hierarchical Clustering are used for the final marketing segmentation with data of mixed type. Through cluster analysis, this work tries to interpret the characteristics of customer patterns found by algorithms and provide clues for business strategy.

Keywords: Unsupervised Learning · Market Segmentation · Principle Component Analysis.

1 Introduction

Due to the development of technology, we now are able to gain customer data with great many features. For the companies, it is crucial to find the clues for the marketing through this kind of data. This work tries to analyze the customers based on a high-dimension data and tries to find the latent patterns by implementing unsupervised learning method twice in sequence with K-means and Hierarchical method respectively. Customer personality will be explored by the cluster analysis.

2 Data Description

The dataset used in this analysis is supermarket customer information from Kaggle. I use the following 16 variables which are in three categories:

Personal:
- Year_Birth: Customer's birth year (date)
- Education: Customer's education level (character)
- Marital_status: Customer's marital status (character)
- Income: Customer's yearly household income (number)
- Kidhome: Number of children in customer's household (number)
- Teenhome: Teenhome: Number of teenagers in customer's household (number)
- Dt_Customer: Date of customer's enrollment with the company (date)

Product:
- MntWines: Amount spent on wine in last 2 years (number)
- MntFruits: Amount spent on fruits in last 2 years (number)
- MntMeatProducts: Amount spent on meat in last 2 years (number)
- MntFishProducts: Amount spent on fish in last 2 years (number)
- MntSweetProducts: Amount spent on sweets in last 2 years (number)
- MntGoldProds: Amount spent on gold in last 2 years (number)

Place:
- NumWebPurchases: Number of purchases made through the company's web site (number)
- NumCatalogPurchases: Number of purchases made using a catalogue (number)
- NumStorePurchases: Number of visits to company's web site in the last month (number)

All the variable will be transferred into numeric or categorical form and since we have so many variables here, some feature engineering will be enforced before analysis.

2.1 Personal Feature

Firstly, *Year_Birth* will be turned to *Age* calculating from this year 2021. *Education* has five categories which are "2nd Cyle", "Basic", "Graduation", "Master", and "PhD". "2nd Cyle" and "Master" are of similar level, so the two categories are combined together and assigned as "Master". *Marital_status* have 6 categories which are "Alone", "Divorced", "Married", "Single", "Together"and "Widow". To make it more simple and easy to understand, this variable can be a new variable *Living* by describing whether customer is in couple or in single. There are two variables are about children which are *Kidhome* and *Teenhome*, which are too redundant for the data, so they are summed together and therefore I get new variable *Children*. From *Dt_Customer*, we can get new variable *Seniority*.

Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Income	1730	35303	51382	52247	68522	666666
Age	25.00	44.00	51.00	52.18	62.00	128.00
Seniority	2621	2801	2976	2975	3150	3320
Children	0	0	1	0.9472	1	3
NumWebPurchases	0	2	4	4.085	6	27
NumCatalogPurchases	0	0	2	2.671	4	28
NumStorePurchases	0	3	5	5.801	8	13
MntWines	0.0	24.0	174.5	305.1	505.0	1493.0
MntFruits	0.00	2.00	8.00	26.36	33.00	199.00
MntMeatProducts	0.0	16.0	68.0	167.0	232.2	1725.0
MntFishProducts	0.00	3.00	12.00	37.64	50.00	259.00
MntSweetProducts	0.00	1.00	8.00	27.03	33.00	262.00
MntGoldProds	0.00	9.00	24.50	43.97	56.00	321.00

Table 1: Summary of Original Numeric Variables

After the restructuring the features, we can summarize the numeric variables in **Table 1**. Some features have extremely large values such as *Income* and *Age*. Because unsupervised learning methods relies on distance or similarity between observations and quite large values can be mistake and have strong impact on final clustering, I remove the observations who have values larger than 99% quantile in *Income*, *Age* and all the product features. and place features

2.2 Product Feature

In order to reduce the dimension and extract more information from the product values, 6 product variables are combined into one categorical variable which can tell people's preference on different categories. **Principle Component Analysis** (PCA) and **K-Means** are employed to cluster the customers into several groups and the new groups and their interpretation will be the new *Product* feature.

Principle Component Analysis

PCA is helpful to geometrically represent the data in a low-dimensional space but remains the key information which will improve the performance of K-means. In the result of PCA, as depicted in **Fig. 1**, the first three components take more than 80% of the total variance of data and the last eigenvalues are far smaller than 1. So I take the first three components as new variables for clustering.

Table 2 lists the loadings of the original variables and communality. We can easily find that amount spent on fruits, fish and gold have close values in all the three components. Obviously, *Component 1* is strongly negatively correlated to all the variables. It can be interpreted as the index of how much the customer like purchasing in the supermarket. The larger values in negative, the higher possibility that the customer is a shopping lover. The most correlating variables to *Component 2* are MntWines and MntGoldProds, both in negative relation. MntWines and MntGoldProds are also the top variables in *Component 3*. But MntWines has positive coefficient while MntGoldProds has negative one. Component 2 can be regarded as how much the customer dislike wines and gold while Component 3 can be taken as how much the customer like wines dislike gold.

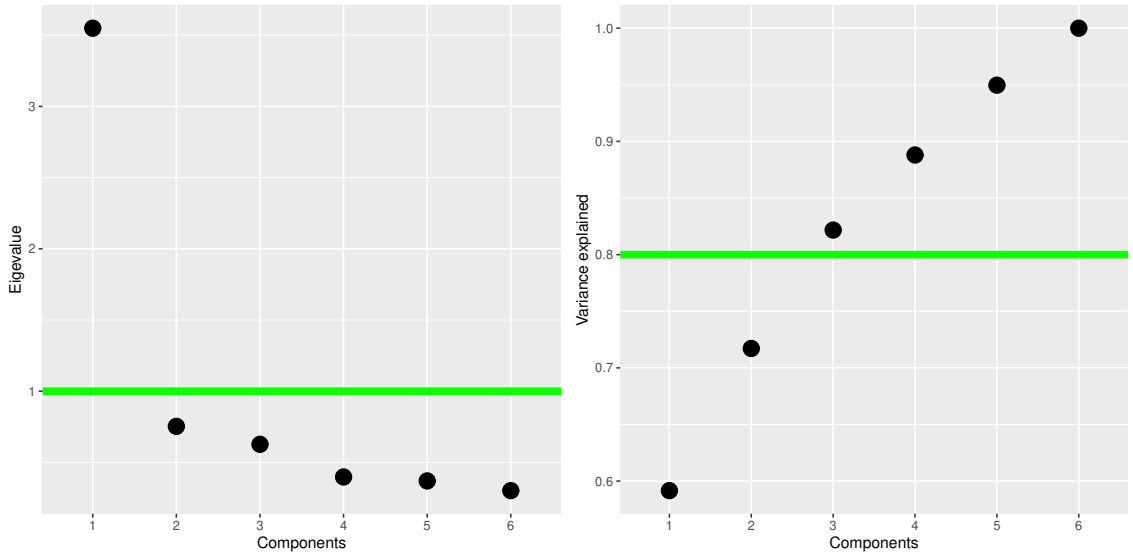

Fig. 1: Eigenvalues and Variance Explained

Table 2: Loadings and Communality

	Component 1	Component 2	Component 3	Communality
MntWines	-0.706	-0.486	0.439	0.927353
MntFruits	-0.800	0.281	-0.054	0.721877
MntMeatProducts	-0.844	-0.043	0.287	0.796554
MntFishProducts	-0.819	0.264	-0.114	0.753453
MntSweetProducts	-0.795	0.333	-0.064	0.747010
MntGoldProds	-0.638	-0.511	-0.571	0.994206

K-Means

Given the number of cluster K, K-means minimizes the centroid between different points of all the clusters by starting from different data points many times and picking the smallest solution for minimization. Because K-means cannot find the optimal number of the clusters by itself, I will confirm the number of clusters first. Concerning the purpose of this clustering is to build a new variable for summarizing the products features of customers, it is crucial to make sure the total intra-cluster variation which is total within-cluster sum of square (WSS), is minimized. The total WSS measures the compactness of the clustering and it should be as small as possible. **Fig. 2** depicts the total WSS as a function of the number of clusters. In this case, I use **Elbow method** to choose the optimal number k of clusters. The Elbow method looks for the k where adding another cluster does not improve much better the total WSS. In the plot, when k=4, the speed of WSS decrease slows down and there is an elbow, so 4 is set for the number of cluster in K-Means.

By means of new variables obtained in PCA, the data is categorised into 4 groups. **Fig. 3** plots how the 4 clusters vary with three components.

Cluster 1 has largest negative magnitude among all the clusters in component 1 while it has low values towards to zero in component 2 and component 3. It means this group of customers purchase a lot in all categories, so they are *active shopper*.

Compared to the other clusters, **Cluster 2** is the only group which has positive mean in component 3. Component 3 represents the preferences on wine. Low negative values in component 1 means that cluster 2 shop in this company but too much. So **Cluster 2** can be explained as *wine lover*.

Cluster 3 has significant negative values in component 2 and component 3. These two components both negatively correlates MntGoldProds. They can be viewed as *gold lover*.

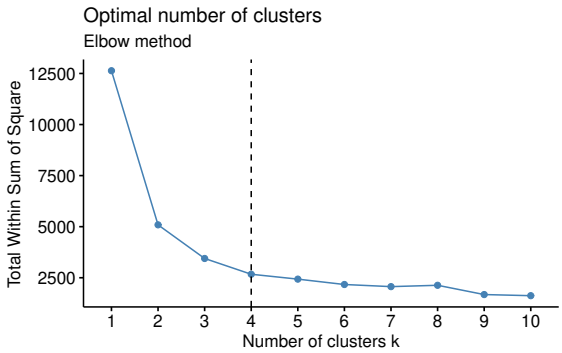

Fig. 2: Total Within Sum of Square

Cluster 4 is the only cluster which have positive mean in component1 and the means in component 2 and component 3 are nearly zero. Considering the interpretation of component 1, this cluster is the *inactive shopper* who buy less compared to others.

The new feature *Product* has the 4 category values and the original 6 variables are removed.

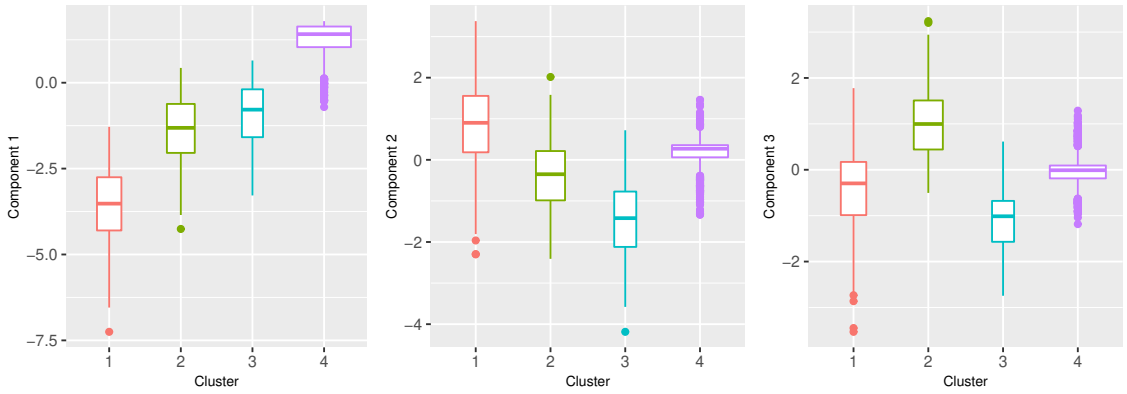

Fig. 3: Four Clusters for New Variables

2.3 Feature Summarization

After feature cleaning and rebuilding, the data now have **10** variables, including 7 numeric variables and 3 categorical variables with **2,043** observations. The summary of numeric variables and categorical variables are presented in **Table 3** and **Fig. 4**

Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Income	1730	34236	49118	49705	65526	94384
Age	25.0	44.0	51.0	51.9	61.0	76.0
Seniority	2621	2802	2977	2975	3150	3320
Children	0	0	1	1.001	1	3
NumWebPurchases	0	2	3	3.968	6	11
NumCatalogPurchases	0	0	1	2.363	4	10
NumStorePurchases	0	3	5	5.681	8	13

Table 3: Summary of Numeric Variables

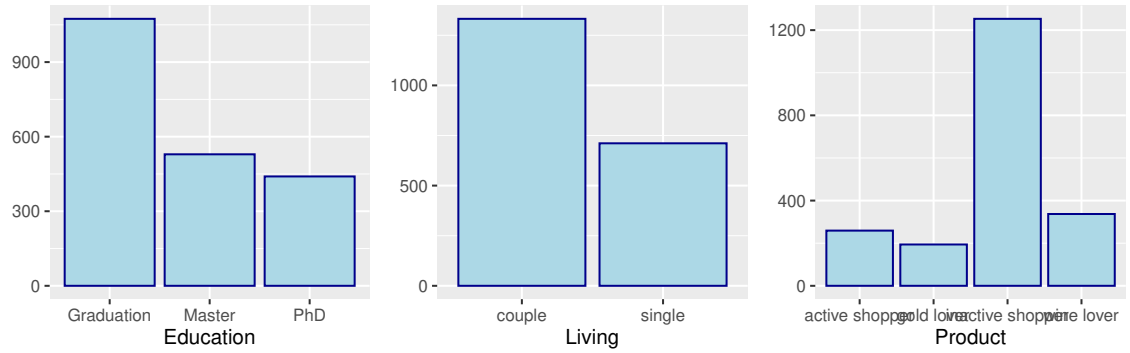

Fig. 4: Summary of Categorical Variables

3 Hierarchical Clustering

This section introduces **Hierarchical Clustering** for the final marketing segmentation strategy. Hierarchical clustering is based on the dissimilarity between each pair of the samples. It uses agglomerative approach which begins with each observation in a distinct cluster, and successively merges clusters together until a stopping criterion is satisfied.

In **Ward's**, proximity between two clusters is the magnitude by which the summed square in their joint cluster will be greater than the combined summed square in these two clusters. Concerning that the clustering will be implemented on mixed type of data and the purpose of detecting patterns, this work enforces **Ward's** method.

3.1 Dissimilarity Measure

Since the data are of mixed type, both having numeric and categorical features, **Gower Distance** are employed. It is defined as:

$$d_{G,ij} = \frac{\sum_{t=1}^{p} \delta_{ijt} s_{ijt}}{\sum_{t=1}^{p} \delta_{ijt}} \tag{1}$$

where $d_{G,ij}$ is the dissimilarity between unit i and unit j, s_{ijt} is the similarity between i and j with respect to tth variable and its value depends on the type of the variables itself, δ_{ijt} is a term allowing to control for the comparability of the observations and of the features. d_{ijt} is close to 1 if the two observations are close to each other while it is close to 0 if they are far apart along feature t. If two observations cannot be compared along feature t (because, for example, of missing values), δ_{ijt} is set to zero.

3.2 Cluster Evaluation

Since this is mainly for business strategy, the company do not need too many clusters limited to the budget. To get a better interpretation for final market segmentation, I experiment the clustering on the two linkage methods with respect to 3 and 4 clusters. To evaluate the performance in the four experiment pairs, the unsupervised learning result is evaluated by **Internal clustering validation** which use the internal information of the clustering process to evaluate the goodness of a clustering structure without reference to external information. The main reason for the choice is that I want to make sure that the observations in the same cluster are similar as much as possible, while the observations in different clusters are highly distinct from each other. **Dunn's Index** is selected as measures. It is the ratio of the smallest distance between observations not in the same cluster to the largest intra-cluster distance. The Dunn's Index has a value between zero and infinity, and the one with larger Dunn's Index will be selected.

From **Table 3**, apparently, clustering with 4 clusters takes advantages over the one with 3 clusters so the following analysis will focus on 4 clusters.

Table 4: Dunn's Index Comparison

	Ward's	
	3 clusters	**4 clusters**
Dunn's Index	0.1388741	0.147637

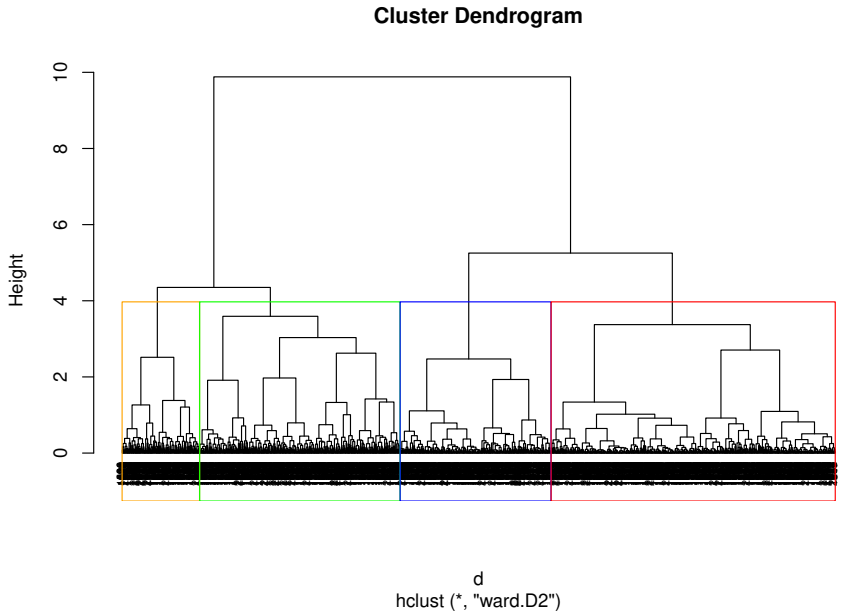

Fig. 5: Cluster Dendrogram

3.3 Cluster Interpretation

Fig. 5 dicpts the dengrogram generated by Ward's method. The data are grouped into 4 clusters, each containing 421, 821, 572, and 219 customers respectively. According to the cluster result, **Fig. 6** and **Fig.7** present the differences among clusters on all the variables.

Firstly, in the left plot of **Fig. 6**, the bar chart shows the percentage of each education level in different clusters. And Apparently, they cannot distinguish from each other, meaning that the partition strategy does not concern about education.

For living state, all the customers are in cluster 1 and cluster 4 are in single, while cluster 2 and cluster 3 are mostly in couple.

In the right plot in **Fig. 6** for Product preference, cluster 1 and cluster 2 are full of inactive shoppers. Cluster 4 is combined by active shopper without gold lover or inactive shopper. In cluster 3, there is nearly no inactive shopper.

In the numeric variables, we can find that, cluster 1 and cluster 2 always have similar behaviors and simultaneously, cluster 3 and cluster 4 have similaer behaviours. Cluster 1 and cluster 2 have slightly lower magnitude in the age than cluster 3 and cluster 4. Manifestly, cluster 3 and cluster 4 have much higher income than cluster 1 and cluster2, and there is nearly no overlapping between the plots. And cluster 3 and cluster 4 have longer membership. But cluster 3 and cluster 4 have less children.

Compared to cluster 3 and cluster 4, cluster 1 and cluster 2 have significant low values on all of NumWebPurchases, NumCatalogPurchases and NumStorePurchases. This corresponds to the fact that customers in cluster 1 and cluster 2 are inactive shoppers.

According to the observations above, the characteristics of each cluster can be summarized as following:

Cluster 1 are customers in single who are inactive shoppers, and they have lower income but more children, and their frequency to the supermarket is relatively low. Compared to catalog purchase and web purchase, store purchase is their preference.

Cluster 2 are customers in couple who are inactive shoppers. They have children but lower income. Same as customers in cluster 1, they do not have frequent purchase and prefer store purchase.

Cluster 3 are customers in couple who buy in this company and go to the company frequently. They have longer membership and no children. And they are richer.

Cluster 4 are customers in single who like shopping and shop frequently. Same as cluster 3, they have longer membership and no children. Also they have high income.

Cluster 3 and **Cluster 4** can viewed as the loyal clients and target customers. They have ability to pay and are willing to pay. The business strategy for them should mind the living status. **Cluster 1** and **Cluster 2** are the customers who need more sales incentives in stores.

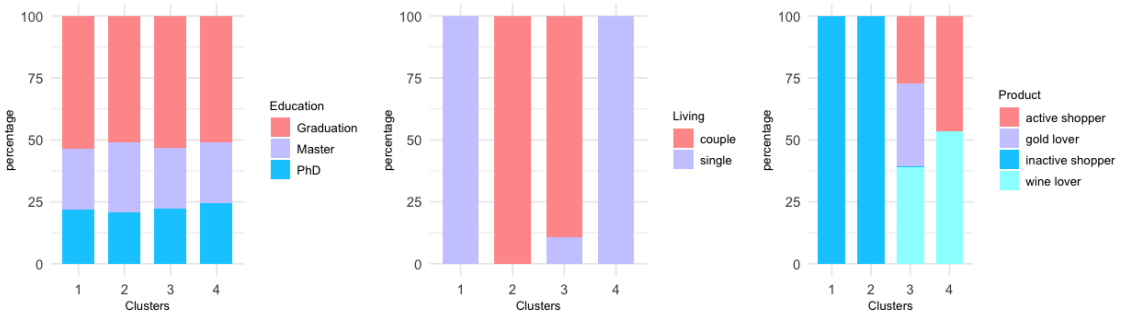

Fig. 6: Clusters and Categorical Variables

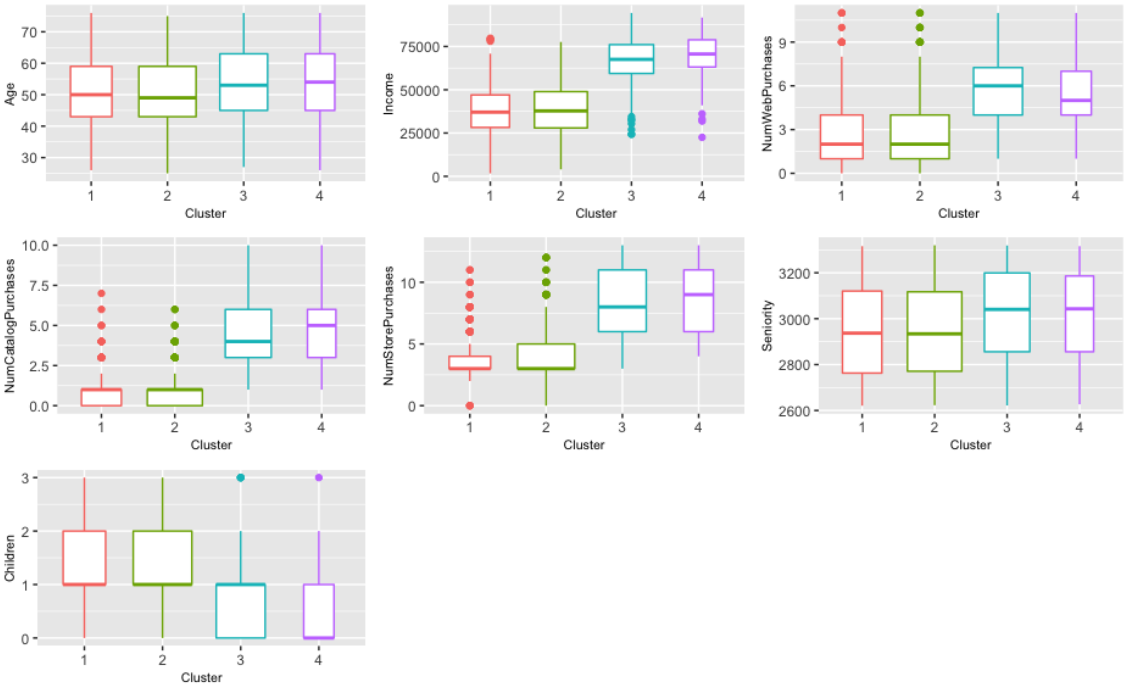

Fig. 7: Clusters and Numeric Variables

4 Conclusion

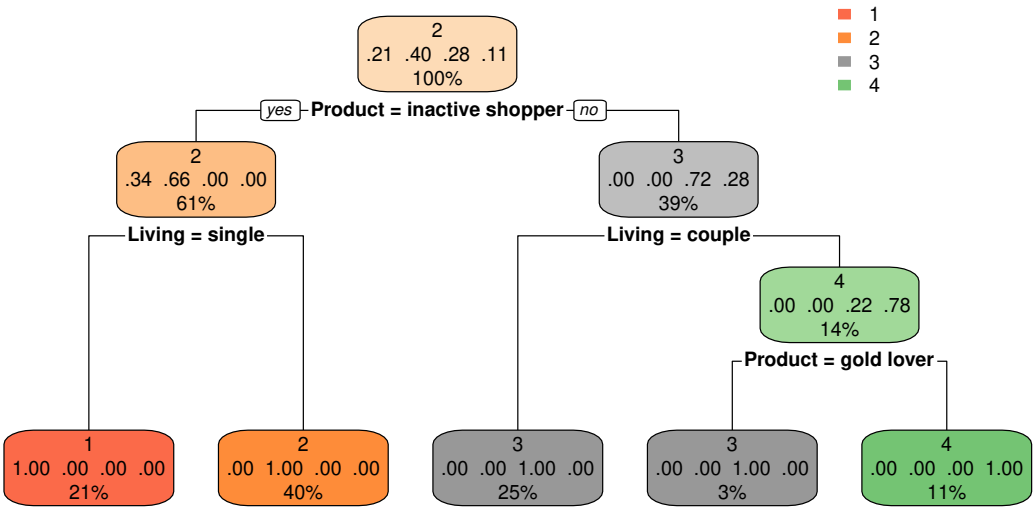

Fig. 8: Clusters and Numeric Variables

By using unsupervised learning methods twice, this work intends to reduce the dimension and aggregate the information in features. It finally divides the data into 4 groups and describes the pattern behind. From the final interpretation, we can find that the work so far relies more on categorical variables. Without categorical variables, present clustering fails to distinguish cluster 1 from cluster 2, and also fails to tell cluster 3 and cluster 4 apart. To be more clear, decision tree is used to show the mechanism behind. As the result presented in **Fig. 8**, the main influencing factors are *Product* and *Living Status* while the other numeric variables have no impacts. What's more, we can see that frequency of purchase in different place and amount of products customers buy have some positive relation. This kind of relation also appears between income and amount of products customers buy. So the future work may try Categorical Principle Component Analysis to find the latent variables behind which will help improve the performance of clustering.

Besides the issues in Hierarchical Clustering, some issues come from K-means the first time the unsupervised learning is executed. Through the result of K-means, the category value distribution is quite unbalanced. This unbalancing may be the main reason causing that cluster 1 and cluster 2 are full of inactive shoppers and attracting too much attention from the agglomerative algorithm.

5 Appendix

This project is performed through R, and code book is shared on Github.