Na Ni

CUSTOMER SEGMENTATION OF GOODBABY CUSTOMERS

Master’s Thesis in Information Systems
Supervisors: Prof. Shahrokh Nikou
Faculty of Social Sciences, Business and Economics
Åbo Akademi University

Åbo 2019
Abstract

When talking about China, the population has always been a hot topic. Since the two-child policy was fully opened at the end of 2015, people keep focusing on the maternal and child market that is expected to expand rapidly, and so does the author. The surge in birth population and the diversification of consumer demand have led to more demand for high quality products and services.

In recent years, this industry has been buoyed by capital, especially vertical e-commerce platform. The participation of e-commerce has made the maternal and child retail market increasingly competitive. How do traditional mother-and-baby companies compete with vertical e-commerce platforms? How do they seize the chance of transformation? Unfortunately, there is not enough literature or historical experience on the transition of the traditional mother-to-child retail company. Therefore, the author of this thesis is committed to contributing to the transformation of these traditional companies that are in dire straits. And “Goodbaby”, as one of the largest distributors and retailers selling mother and baby products in China, became the object of this study.

After six-month intern in Goodbaby, several clear problems have been found according to the basic statistics. The most fundamental cause of these problems is that they do not understand their customers, or they don't know who they are and what characteristics they have. Thus, in this thesis, the author would perform customer segmentation to assist the company make the first step in the transition.

The approach adopted to solve the problem is the combination of RFM (Recency, Frequency, Monetary) model and the famous algorithm of clustering analysis, K-means. In order to make the whole process more clear and standardized, the author regarded this experiment as a data mining work and applied the classic CRISP-DM model to implement the entire process of customer segmentation.

The results obtained in the model include seven meaningful clusters and one outlier. Combined with demographic attributes and the analysis of consumer behavior, all these seven clusters have been well defined, and corresponding marketing strategies were given according to the features of each cluster.
Table of contents

1. INTRODUCTION ........................................................................................................5
   1.1. Area of Research .................................................................................................5
   1.2. Research Question .............................................................................................7
   1.3. Structure .............................................................................................................7

2. THEORETICAL BACKGROUND ..............................................................................9
   2.1. Fertility Policy in China .................................................................................9
   2.2. Baby Boom .......................................................................................................10
   2.3. Data mining, machine learning and segmentation ...........................................12
   2.4. Customer segmentation studies ....................................................................14

3. METHOD ..................................................................................................................20
   3.1. Business Understanding ..................................................................................20
       3.1.1. Business Background ..............................................................................20
       3.1.2. New Retail ..............................................................................................20
       3.1.3. The Main difficulty faced by Goodbaby ..................................................21
   3.2. Data exploration ...............................................................................................21
   3.3. Data preparation ..............................................................................................22
       3.3.1. Business unit .........................................................................................22
       3.3.2. Channel .................................................................................................23
       3.3.3. Gender ...................................................................................................24
       3.3.4. Age ........................................................................................................25
       3.3.5. Baby’s Age ..............................................................................................26
       3.3.6. Baby’s gender .........................................................................................27
       3.3.7. Address ..................................................................................................28
       3.3.8. Shopping Guide ......................................................................................29

4. MODELING ..............................................................................................................31
   4.1. RFM model .......................................................................................................31
   4.2. Algorithm Selection .........................................................................................33
   4.3. Why RFM score formula cannot be applied? ..................................................33
   4.4. The process of K-means ..................................................................................34
       4.4.1. Check the distribution of each indicator ....................................................34
       4.4.2. Standardize the RFM indicators ...............................................................38
       4.4.3. Determine the number of clusters m .........................................................39
       4.4.4. Applying K-means for clustering. Obtain eight groups of members. .......40

5. EVALUATION ..........................................................................................................42
5.1. Check the mean of “RFM” in each cluster and compare with the total means 42

5.2. Deep analysis of eight clusters.............................................................................43

5.2.1. Cluster 1: Lost - Worst recency, frequency and monetary level. Hard to revive. 46

5.2.2. Cluster 6: Hibernating - Last purchase was long back, low spenders and low number of orders. ................................................................. 49

5.2.3. Cluster 2: Promising - Recent buyers, but haven’t spent much. .............. 50

5.2.4. Cluster 7: Potential Valuable - Spent a good amount and bought more than once. Responsive to promotions................................................. 54

5.2.5. Cluster 8: Top Customers - Bought recently, buy often and spend the most! 58

5.2.6. Cluster 4: Potential Loyalist - Bought several times but spent not much. 61

5.2.7. Cluster 3: Loyalists – Recent buyers and bought often. Not spent much each time........................................................................................................... 65

6. SUMMARY.................................................................................................................. 69

7. DISCUSSION............................................................................................................. 77

7.1. The meaning of customer segmentation for Goodbaby......................... 77

7.2. Further roles of the research......................................................................... 78

7.2.1. For other companies .................................................................................... 78

7.2.2. For the past literature .................................................................................. 78

8. LIMITATION AND FUTURE RESEARCH.................................................. 82

REFERENCE ............................................................................................................ 84

APPENDICES............................................................................................................ 88
1. INTRODUCTION

1.1. Area of Research

When people talk about China, the first word coming into their mind must be “large”, large territory and large population. According to the World Bank’s open data, China is the most populous country in the world with a population of over 1.386 billion (by the end of 2017). Another country with a population of more than 1 billion is India. However, by the year 2030, India is expected to become the most populous country in the world because the population in India keep growing while China will see a loss in population (World Population Prospects, United Nations, 2017 revision). This phenomenon is inseparable from China’s Fertility policies. In response to the issue of the aging of the population and lack of labor force, China fully opened its two-child policy at the end of 2015. This means that in the next few years, the market of maternal and child products is likely to enter a period of growth and rapid expansion.

Although this seems like a good opportunity, it is a double-edged sword for some companies. For the traditional maternal and child products companies in the face of the impact of the e-commerce industry, if the transformation succeeds, it can make a huge breakthrough, while if they do not take the opportunity to change due to digital transformation, they are likely to fall behind and will be eliminated by the competitors in the future.

With millions of competitors, attracting and retaining customers is critical for success (Ballestar et al., 2018). Therefore, more and more companies attach importance to customer relationship management (CRM). CRM is the operational model by which enterprises understand and affect consumer behavior through interaction with customers so as to attract new customers, maintain old customers and enhance customer loyalty and thereby increase profits (Chung and Chen, 2016). Ngai et al. (2009) mentioned that CRM is categorized into four dimensions which are customer identification, customer attrac-
tion, customer retention and customer development. Therefore, to maximize the performance of CRM by an enterprise, at first the company needs to have a defined, segmented group of its audience of customers (Ballestar et al., 2018).

Smith (1956) has first introduced the concept of segmentation in the marketing literature. To realize customer segmentation, there are several advanced models and algorithms like chi-square automatic interaction detection (CHAID), logistic regression and artificial neural networks (ANNs) (Cheng and Chen, 2009; Xu and Chu, 2016). However, many marketers continue to apply the traditional RFM (recency, frequency and monetary) models for their ease of use and understanding (Sarvari et al., 2016). But the traditional RFM models have some blind spots. In many cases, RFM values need to be converted to RFM scores. If R, F and M have scores from 1 to 5, there are a total of 125(5*5*5) combinations of RFM value. If the analysts were to look at how many customers they have for each RFM value, they would have to look at 125 points of data. Then they need to summarize these 125 groups into several segments to understand the customers better. All of the above processes take a long time and are not reusable. But if combined with data mining techniques like cluster analysis, things will become easier. In this thesis, the author would adopt k-means models using RFM values to perform customer segmentation. In order to formulate a comprehensive clustering model, more other attributes like consumption channel and demographic variables should not be ignored.

The aim of this thesis is to help the transformation of traditional mother-to-child retail businesses. In order to better understand the difficulties faced by this kind of companies, the author chose to investigate in depth the largest distribution and retail platform for mother and baby products in China - Goodbaby. It is one of the most famous maternal and child brands in China as well as a traditional wholesale and retail company. The current situation of this company exactly fits the description above.
The thesis can provide a reference for other companies in the similar situation, and offer a perspective for researchers interested in the Chinese maternal and infant products market through the process of customer segmentation for Goodbaby.

1.2. Research Question

In this thesis, the following questions would be answered.
- How many clusters can the consumers of Goodbaby be divided into?
- What are the characteristics of each member cluster?
- Are there corresponding marketing strategies for different cluster of members?
- Do members who prefer different consumption channels have the same loyalty to the brand?

To solve these questions, the author adopted Cross Industry Standard Process for Data Mining (CRISP-DM). First, deep understand the business and explore the existing datasets of Goodbaby. Collect the key variables and check the integrity and accuracy of them. Try to complement the missing parts through algorithms if necessary. In this thesis, the author extracted three representative variables recency, frequency and monetary from the transaction records to be used in the model. The algorithm called K-means was applied for clustering. After running the model, eight clusters were distinguished. Following the process of CRISP-DM, the results were evaluated combined with more other variables like demographic attributes.

1.3. Structure

The layout of this thesis is as follows. Chapter 2 presents the theoretical background of the thesis, including the review of fertility policies in China and the brief introduction of relevant studies such as data mining, machine learning and segmentation. Chapter 3 introduces the research method. The first three steps of CRISP-DM have been discussed in this part. Chapter 4 explains the entire modeling process and presents the RFM model and the algorithm selected in detail. Chapter 5 presents the results. Chapter 6 summaries the characteristics of each cluster and corresponding strategies in table. Chapter 7 dis-
cusses the roles of the research in the perspective of real companies and relevant literature. In the last chapter, the limitations of this thesis have been explained and the direction of the future research is recommended.
2. THEORETICAL BACKGROUND

2.1. Fertility Policy in China

As one of the countries which have the largest population size, China has experienced three drastic fertility transitions.

The first decline is the most drastic one. In 1970, the country achieved almost her highest fertility rate 5.8, which meant usually one mother had around 5.8 kids. In one decade’s time, China’s national fertility level was cut in half. The main reason for this was the launching of forceful government fertility policies on population control. For accurate, in the early 1950s after the census, the government realized the large population size and fast population growth therefore started to plan this program. The main idea in the first fertility policy could be generalized in three Chinese words: wan (late marriage), xi (longer birth intervals) and shao (fewer children). In this phase, official marriage age was raised to 25 and 22 for urban and rural males, 23 and 20 for females. The urban family was allowed to give birth to two children while the rural family could have three children. After the policy was launched, Chinese marriage age increased significantly and the fertility rate declined drastically (Feng, 2015).

The second fertility decline coincided with a much more controversial birth control policy called one-child plan which was launched in the 1980s, resulting in sharp increases in the numbers of abortions and sterilization (Hardee-Cleaveland and Banister, 1988). In this stage, each couple could only have one child. But in fact, only Chinese couples in cities (about 20% of the whole population) obeyed the One-Child policy; however, most couples in the countryside had two or even more than two children (Feeney and Yuan, 1994; Gu et al., 2007). In a word, China has completed her two fertility transitions in two decades.

However, every coin has two sides. Although the population has been successfully controlled in a debatable way, several unavoidable consequences have come. In our generation, nearly 90% urban couples end up with one child (Feng et al., 2012). As a direct result, most parents with single child will not be able to live with their married children. Then lack of functional support especially emotional support will be a big problem for
the elderly people. More importantly, for the country, a sharp decline of the scale of young labor force and expedited population aging is not an unpredictable future. Also, the unbalanced gender ratio between male and female has been brought to the forefront by the government (Feng, 2015).

To relieve this situation, the third policy has arrived in November 2011 which states that the couples who are both from one-child family can have two children. In December 2013, the policy has been corrected more flexibly that either one of the couples from one-child family can be the policyholder. The two-child policy has not been amended once more until October 2015. This time, the two-child policy is open to all the citizens to actively address the issue of the aging of the population and lack of labor force.

### 2.2. Baby Boom

Of course, Chinese parents sped up or slowed down reproduction according to the number and sex composition of existing children, and to their familial economic circumstances (Feng et al., 1995; Lavely, 2007; Lee and Campbell, 1997; Zhao, 1997). However, implementing the two-child policy could retard the negative growth trend of population more or less in the future. According to National Bureau of Statistics of the People’s Republic of China, the birth rate peaked at 12.95‰ in 2016 over the past fifteen years. During the whole year of 2016, more than 17.9 million kids have come to the world, an increase of more than 7.9% year on year. The research from iResearch shows that after the two-child policy comprehensively carried out in 2016, more than 34.2% target families expressed their willingness to have the second child. In other words, baby boom is coming. Figure 1 below shows the population change of China from 2010 to 2016 from three dimensions.
Before the “baby boom”, China has already been a large maternal and child products consumer second only to America. Till the end of 2018, the market size of maternity and baby industry in 2018 is estimated over 482.5 billion RMB (Analysys, 2018). The surge in birth population and the diversification of consumer demand have led to more demand for high quality products and services. This huge cake has become more and more conspicuous in the retail sector. Out of the optimistic view of the maternal and child market, in recent years, the mother and baby industry has been buoyed by capital, especially vertical e-commerce platform, like beibei.com, mia.com and lamabang.com. They have announced millions of dollars in financing in succession while their valuations are also rising. In a word, Chinese maternal and baby market ushers in new opportunities. But how to seize these precious chances? How do traditional mother-and-baby companies compete with vertical e-commerce platforms?

Unlike traditional mother-infant channels, the new wave of maternal and child e-commerce has emerged since 2014, and once it was born, it got onboard a fast train of the mobile Internet. Under the catalysis of both mobile e-commerce and cross-border e-commerce, maternal and infant e-commerce platforms mushroomed rapidly. After the money-
burning battle in 2015, they turn their focus to content marketing. These platforms have spent three years on making e-commerce the main channel in the mother-to-child industry.

However, literature on the transition of the traditional mother-to-child retail company is scant. In the mother-to-child field, most of the literature focuses on issues such as children’s education, environmental safety or food health.

### 2.3. Data mining, machine learning and segmentation

Customer segmentation is the practice of dividing a customer base into groups of individuals that are similar in specific ways relevant to marketing, such as age, gender, interests and spending habits. Customer segmentation, also called consumer segmentation or client segmentation. Companies employing customer segmentation operate under the fact that each customer is different but relevant groups with similar features would be found among them, and that their marketing strategies would be better achieved if they target specific, smaller clusters. Companies also hope to gain a deeper understanding of their customers' preferences.

The vast availability of data, and the inefficient performance of traditional statistical techniques (or statistics-oriented segmentation tools) when handling huge amounts of data, have stimulated researchers to find more effective segmentation tools so as to mining more valuable information about their markets and customers. Thus, knowledge discovery (KD) and data mining (DM) have been seen as a solution to this problem (Sarvari et al., 2016).

Data mining is a process of discovering hidden valuable knowledge by analyzing a large amount of data. It has been defined as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data" (Frawley et al., 1992). It is widely used in various fields including healthcare, finance and retail. The common theme of these areas is that they all have huge amounts of information about their operations as well as their clients, collected in a variety of ways. In order to maximize the value of the
information, they rely on data mining technology to discover specific patterns or trend from the data (Nature Biotechnology, 2000).

There have been some efforts to define standards for the data mining process, for example the 1999 European Cross Industry Standard Process for Data Mining (CRISP-DM 1.0) and the 2004 Java Data Mining standard (JDM 1.0). Development on successors to these processes (CRISP-DM 2.0 and JDM 2.0) was active in 2006, but has stalled since JDM 2.0 was withdrawn without reaching a final draft. Here the aim is to adopt the most classical process ‘the original CRISP-DM’. Not only because it is the most widely-used analytics model but it is also industry, tool, and application neutral (Mariscal et al., 2018). CRISP-DM is an open standard process model that describes common approaches used by data mining experts. It consists of six major phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment. Also, it is a cyclical process. In this thesis, all these phases have been implemented except the last phase. The deployment phase needs to establish effective communication among relevant business units (such as marketing and customer service) about the segmentation. It will take quite a long time to deploy all the details and put theory into practice. So, this thesis will focus on the first five parts.

Disciplines such as machine learning, statistics, artificial intelligence (soft and hard computing techniques), expert systems, and data and knowledge management technologies have been incorporated within data mining, making use of their theories and algorithms (Hiziroglu, 2013). Therefore, progress made by these related fields are making significant contributions that help data mining techniques better able to handle complex tasks (Casan-dio, 2000).

Machine learning (ML) is a field of study that gives computers the ability to learn without using explicit instructions (Samuel, 1959). Three main types of machine learning problems have been distinguished, which are supervised learning, unsupervised learning and reinforcement learning. Supervised learning aims at generalizing the observations in the dataset to new inputs (Simeone, 2018). The dataset is usually divided into two parts, a training set for training the model and a testing set for testing the performance of the model. The final goal is to predict the value of the label for an input that is not in the
existing dataset. The problems of classification and regression belong to supervised learning. Unlike supervised learning, the dataset of unsupervised learning consists of an unlabeled set of training examples. Less well defined than supervised learning, unsupervised learning generally refers to the task of learning properties of the mechanism that generates this dataset (Simeone, 2018). Clustering is one of the specific tasks and applications of unsupervised learning. Furthermore, semi-supervised learning is a generalization of both supervised and unsupervised learning. Reinforcement learning refers to the problem of inferring optimal sequential decisions based on rewards or punishments received as a result of previous actions (Simeone, 2018). In this thesis, the author has adopted the method of clustering to implement the modeling part of segmentation.

2.4. Customer segmentation studies

To maximize the performance of CRM by an enterprise, it needs to have a defined, segmented group of its audience of consumers and customers. And the idea of segmentation was first introduced in the marketing literature by Smith (1956). In 1968, segmentation was mentioned as an alternative to product differentiation strategy by Claycamp and William. The main idea of segmentation or clustering is to group similar customers together. There are many clustering algorithms in the past literature. The major fundamental clustering methods can be classified as partitioning methods (like k-means), hierarchical methods (such as agglomerative and divisive hierarchical clustering algorithms), density-based methods (density-based clustering based on connected regions with high density) and grid-based methods (like STING: the statistical information grid) (Fallis, 2013). Another effective clustering algorithm is the self-organizing map (SOM) of Kohonen, which is well-known for its ability to map an input space with an neural network (NN) (Demartines and Blayo, 1992).

Without a doubt, sales record is a crucial data source to be used to help with the clustering. Besides the transaction history, Konus et al. (2008) gained segments through the way of survey and Ballestar et al. (2018) tried to utilize media contact log data to implement segmentation. Recently, Nakano and Kondo (2018) combined all of these three datasets to enhance the accuracy and comprehensibility of clustering. In this thesis, the focus is on the analysis of sales records accompanied with demographic data from Goodbaby enterprise for some unavoidable limitations of the usage of media contact log data.
To apply DM techniques into the purchase history to find the potential rules, there is a need to extract representative variables from the records to go with the modeling process such as purchase frequency, life time cycle, monetary, product category and so on. With the diversity of sales channels, customers’ consumption behavior is getting more and more complicated and diverse. In some research, the authors found that sales channels had become an important indicator to distinguish between groups. Table 1 summarizes three key past studies of customer segmentation related to this research. Table 2 provides the key findings of these studies. Konus et al. (2008) clustered the results of the questionnaire using latent-class cluster analysis to get 3 segments in 2008: Uninvolved shoppers (40%), Multichannel enthusiasts (37%) and Store focused customers (23%). They also discovered that multichannel enthusiasts tended to be more innovative and store-focused consumers generally were more loyal. The second phenomenon also happened in Goodbaby. However, it was hard to receive enough evidence to define “innovative” without the support of survey. Nakano and Kondo validated Konus's conclusions in 2018 with real consumption data, supporting some of them but also drawing some differential opinions. For example, they believed that multichannel customers tended to maintain a high level of loyalty, while single channel customers maintained a low level. But in this thesis, store-focused customers and multichannel enthusiasts are more loyal and valuable than online-focused customers. Online-focused customers have many other choices in e-commerce sites and they are easily attracted by other brands.
### Table 1: Past Studies of Customer Segmentation

<table>
<thead>
<tr>
<th>Authors</th>
<th>Title</th>
<th>Sales Channel</th>
<th>Data Collection</th>
<th>Analysis Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Konus et al. (2008)</td>
<td>Multichannel Shopper Segments and Their Covariates</td>
<td>Online &amp; offline</td>
<td>Survey</td>
<td>Latent-Class Cluster Analysis</td>
</tr>
<tr>
<td>M.T. Balleslar et al. (2018)</td>
<td>Customer segmentation in e-commerce: Applications to the cash-back business model</td>
<td>Online</td>
<td>Expenses record &amp; media contact log data</td>
<td>Two-step cluster analysis</td>
</tr>
<tr>
<td>Nakano and Kondo (2018)</td>
<td>Customer segmentation with purchase channels and media touch-points using single source panel data</td>
<td>Online &amp; offline</td>
<td>Expenses record &amp; survey &amp; media contact log data</td>
<td>Latent-Class Cluster Analysis</td>
</tr>
</tbody>
</table>
### Table 2: Key Findings of Past Studies

<table>
<thead>
<tr>
<th>Authors</th>
<th>Segmentation Result</th>
<th>Products/Stores/Websites</th>
<th>Key Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Konus et al.</td>
<td>Three segments</td>
<td>Books, Mortgage, Electronics, Holidays,</td>
<td>Multichannel enthusiasts tend to be more innovative.</td>
</tr>
<tr>
<td>(2008)</td>
<td>Uninvolved shoppers (40%)</td>
<td>Computers, Insurance</td>
<td>Store-focused consumers generally are more loyal.</td>
</tr>
<tr>
<td></td>
<td>Multichannel enthusiasts (37%)</td>
<td></td>
<td>Multichannel enthusiasts consider shopping a pleasurable experience than do the other two segments and uninvolved shoppers do not consider shopping a pleasurable experience at all.</td>
</tr>
<tr>
<td></td>
<td>Store focused customers (23%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M.T. Ballestar et al. (2018)</td>
<td>Eight segments</td>
<td>Cashback websites</td>
<td>the route via which customers enrol in the cashback website determines their role and purpose.</td>
</tr>
<tr>
<td></td>
<td>Immature referees (29.7%),</td>
<td></td>
<td>Customers who are more involved in the social network are more multi-transactional. These customers make more kinds of transactions, indicating that they are more engaged and thus more loyal to the brand.</td>
</tr>
<tr>
<td></td>
<td>Heavy user referees who invest time on the website (low profitability) (9.9%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Heavy users of all kind of transactions with low sensitivity to payouts (high profitability) (0.6%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Recommended profitability and convenience-oriented users (medium-high profitability) (6.5%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Convenience buyers with potential (11.4%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Referees in development (19.5%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Engaged convenience buyers (10.6%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Engaged referees (11.8%)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Nakano & Kondo (2018) Seven segments

<table>
<thead>
<tr>
<th>Segment</th>
<th>Description</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store-focused/anti-digital customers (21.3%)</td>
<td>Supermarkets, convenience stores, drug-stores, department stores, and specialty stores, while online stores include Internet supermarkets, e-commerce sites, and direct sales sites of brands.</td>
<td>The segmentation based on actual behavioral data in this study has provided results that support the findings from past studies. Multichannel customers tend to maintain a high level of loyalty, while single channel customers maintain a low level. Among those with considerable experience of using online channels, customers have not only high behavioral loyalty but also high psychological loyalty to the brand or store. Time-constrained customers have innovativeness and low price consciousness and use multimedia. The opinion-seeking type makes active use of social media, tends to follow the opinions of others, and has low loyalty tendencies.</td>
</tr>
<tr>
<td>Store-focused light/anti-digital customers (19.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Store-focused light/multimedia social customers (15.7%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Store-focused/multimedia social customers (15.7%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Online-favored multichannel enthusiasts/PC customers (6.5%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Store-favored multichannel enthusiasts/multimedia social customers (6.4%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uninvolved shoppers (15.4%)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Besides the sales channel, there are some important “numbers” like recency, frequency and monetary helping understand people’s consuming behavior. Customer segmentation is often based on customer lifetime value (CLV) measured by these three purchase variables: “Recency,” “Frequency” and “Monetary” (Safari et al., 2016). RFM stands for the three dimensions: Recency – How recently did the customers purchase? Frequency – How often do they purchase? Monetary – How much do they spend? Teichert (2007) segmented airline customers with demographic attributes and monetary factors and classified their customers into business and leisure categories. For lack of customers’ loyalty factors like recency and frequency, they found their conclusions doubtful and this also limited the application of their work. It was an example of a non-RFM-based segmentation.

RFM analysis originated in the practice of direct marketing by catalog sales companies in the 1960s (Blattberg et al., 2008). It is easy to use and can generally be implemented very quickly (Kahan, 1998). A study by Ambler et al. (2002) showed that RFM was the second most common method used by direct marketers, after cross-tabulation. Although
advanced data mining techniques are being consistently improved and get matured, marketers continue to employ RFM models because of their ease of use and quick implementation. Also, it is easier for decision makers to understand the outcome results of RFM analysis (Sarvari et al., 2016).

However, RFM may focus too much attention on transaction information and ignore individual difference information which is very important for assessing individual preferences (McCarty and Hastak, 2007). Later, in 2009, Chen and the others constructed a new customer segmentation method based on RFM, demographic and lifetime value data. First, they segmented the customers into different groups using RFM factors with k-means algorithm. Second, they partitioned each cluster again into new clusters using demographic data. They suggested that investigating customer demographic characteristics might suggest different marketing implications (Sarvari et al., 2016). In other words, customer clusters are better enhanced when segmentation processes are based on RFM analysis accompanied by demographic data.

In this thesis, to maintain the minimum number of variables to model and to obtain more understandable and comprehensive results, the author would use RFM to perform clustering and combine consumption channels and demographic attributes to analyze the data. The dataset includes two sales channels (online & offline).
3. METHOD

3.1. Business Understanding

3.1.1. Business Background

Goodbaby was a complete wholesaler of baby strollers at first. It was a company that only sold goods but no service. They were not predictive of the value that consumer services could bring and there was no professional staff to improve the customer experience. In recent years, the company attitude on consumer service and customer management has been influenced by the atmosphere and general direction of the entire Chinese retail market, as well as the perceived lack of competitiveness. As a result, Goodbaby realized the importance of change. In 2016, Jack Ma, Alibaba’s Founder and Chairman, proposed the concept of “new retail” in a letter to Alibaba’s shareholders. This concept highly encompasses the transformation trend of China's retail industry in the recent ten even twenty years and the growing demand of Chinese consumers for higher quality service and shopping experience.

3.1.2. New Retail

Retail is the sale of goods or commodities in small quantities directly to consumers. (American Heritage® Dictionary of the English Language). Retail activities must include three elements – the place (channel), the products (or services), and the people (or companies) (Berens,1983). The Internet has brought many changes, but the nature of retail never changes. When almost every target consumer is an internet user, the whole branding, brand promotion and marketing can be done in a digital way. Digital media not only enables consumers to receive messages but also achieves two-way interactions to promote further conversion and gain insight into these potential customers. Data can help both producers and providers to better identify and understand consumers.

On the basis of insights into consumer needs, it guides the production of goods and the optimization of services, and then provides them to consumers in all appropriate scenarios, both online and offline. The three elements of retailing have been reconstructed, from
the original “products-place-people" to “people-products-place”. This is the so-called “new retail", which is similar but different from the traditional “retail”.

In general, “new retail” is not a new business model, nor a kind of new technology. This is a shift in consciousness and decision-making. This is the realization of data-based operations which helps the company bring better service and experience to customers and provide a better resource allocation solution for the supplier.

3.1.3. The Main difficulty faced by Goodbaby

Goodbaby has several clear problems, in addition to the basic statistics. For instance, customer churn rate is high; customer’s repurchase intention is low compared to other e-platforms; distribution of online and offline inventory is uneven. But all of these above are consequences of a lack of effective member operations. The bottom line is that they do not understand their customers very well, for instance, they do not know who they are and what characteristics they have.

Nowadays, the generation after 85s and 90s were the main consumers of maternal and infant products. With the widespread popularity of the mobile Internet and smart terminals, the information is highly symmetric for both buyers and sellers. Moreover, their consumption concepts are more mature now. So, it is difficult for the retailers to “fudge” their purchases with traditional promotional methods. Goodbaby needs to utilize and make use of their existing data and resources to understand their customers in-depth and implement different marketing programs based on different customer types. Due to the lack of internet talents and lack of data management thinking, the company still does not have a sound membership system and loyalty program. The thesis extracts data from Goodbaby’s basic database then perform customer segmentation to assist the company make the first step in the transition.

3.2. Data exploration
To embrace the new retail era, companies must first digitize their retail ports. Whether it is online or offline, as long as consumer behavior is real, existing technology can turn it into a digital transaction.

Although Goodbaby has been building a membership system to collect customer information so as to provide better-customized services in the past few years, after viewing the database, it seems that besides the most basic transaction data (from POS systems and EC channels), the basic attributes like consumers' age and gender, baby's age and gender were mostly missing or error records. Moreover, the conversion rate of members is quite low. This is a big challenge for the company, which means that to perform the customer segmentation and get a deeper understanding of consumer behavior, the researcher needs to extract other necessary content from the consumer's trading behavior or existing data (such as judging adults’ gender by name, judging babies’ gender and age by purchased products, etc.). What is more, the stores have a large number of shopping guides. When they check out the customers, they are likely to use their own membership card number. This leads to numerous shopping guide records mixed into members’ transaction records. One shopping guide’s records usually represent the consumption behavior of multiple consumers, which will be a non-negligible impact on our analysis results. Therefore, this part of the data must be removed before modeling and analysis.

3.3. Data preparation

After understanding the basic situation of the database, the author of this thesis decided to try to extract some other basic information on members to increase the variable richness, which will help the understanding of model results and increase the practical guiding significance of analysis.

3.3.1. Business unit

Goodbaby is a large enterprise with many business units. It has not only its own brands like Goodbaby and Happydino, but also obtained the proxy permission of famous sports brands (NIKE KIDS, ADIDAS KIDS and etc.) and Mothercare in China. Therefore, among 9 million members, some of them are members of Goodbaby, some are members
of Mothercare, and others have only bought a pair of shoes in the sports store, leaving personal information for discounts. Due to a large number of business divisions, if we convert them into 0-1 variables, the number of fields would be too much, which would definitely reduce the running effect and interpretability of the model. Also, the target groups and brand positioning of different brands are also discrepant. What is more, a sophisticated computer with high computing power is needed to run too large datasets. Based on all of these issues, in this thesis, the focus is only on the customer segmentation of all the 2.93 million members who once purchased the products of the brand ‘Goodbaby’ in 2017 or 2018.

3.3.2. Channel

Based on past literature, the consumption channel is likely to be a key variable that can distinguish different customer groups. The consumption behavior of store-focused customers and online-focused customers are always different. The consumption channel here was divided into two parts: online e-commerce platform (including the comprehensive e-commerce platform like Taobao and Jingdong, Goodbaby's official websites, self-operated APP “Mamahao”, etc.) and offline stores.

To facilitate the process of analysis, each channel is converted into one binary variable. For example, when the variable “is_e” equals one, it means that the member has once spent at the e-commerce platform. Around 74 thousand members spent at stores in the past two years, while about 2.69 million members purchased online through Tmall platform (one of Alibaba’s e-commerce platforms) accounting for 92% of total members. The reasons for this include that if you purchase at an offline store, there is no mandatory requirement for registration, while once online consumption will turn you into a member of this e-shop by default. Although the number of offline members was far less than the number of online members, bricks and mortar stores contributed 54% (including sales created by unregistered offline customers) of the total sales in 2017 and 2018, which indicated huge potential in the offline market.
3.3.3. Gender

In the member profile, “name” was a relevant complete field. The value of the name was either a full name or the family name with a title like Mr. or Miss. The latter could be easily judged by gender but for the former one an algorithm was needed to determine whether this name was more likely to be a male or a female. After reviewing some academic papers on how to judge a person’s gender by his name, it was found that many authors used the similar way to first model the training set (names and gender are both known) to calculate the probability that a name with the given name was a lady or a gentleman when the given name was in a different position, then predict the gender of each name in the test set based on the result of the training model. In China, a full name usually consists of two words or three words including a one-word surname. For example, for a name like “Xiaoting Pan”, Xiao and Ting are given names and Pan is a surname.

This was certainly a good method but it was quite tough and challenging to gain a large enough training set so that the model could perform well on the test set. As a result, the author took another simple and straight way to judge the gender of members. Plenty of names commonly used by boys and girls from name-recommended websites have been downloaded, and a name library with boys’ names as well as a name library with girls’ names has been established separately. When a member's first name existed in the boy's name library, it would be classified as a boy, and when a member's name existed in the girl's name library, it would be classified as a girl. And if the member's name has three words, according to the Chinese custom of naming, the gender bias of the third word should be given priority. If the name library did not contain any words in the name, it would be marked as "unknown".

By continuously expanding the name library, the author has finally gained nearly 80% of the members’ gender. Figure 2 below shows the detailed distribution of each gender. After randomly verifying 100 members, the accuracy rate was confirmed to be over 82%.
3.3.4. Age

Members’ age was also a very tricky variable to get. Because of the input mistakes made by the shopping guides, the adult’s birthday and the baby’s birthday were mixed together, which needed to be divided into two variables. Based on the existing values of birthday, the age at which the members registered was calculated. If the age was more than 14 years old, the birthday was deemed to be the adult's birthday; if the age was less than 14 years and was greater than minus ten months (expected date of childbirth could be filled in), this birthday was considered to be the baby's birthday. After running the algorithm, the birthday dates of 2.9 million out of the 2.93 million members were missing values, so the integrity of the field “age” was far from enough. It had to be abandoned. Figure 3 below shows the detailed distribution of members’ age.
3.3.5. Baby’s Age

With the method mentioned above, the author managed to fix a small part of babies’ birthdays, but by comparison, they were found the same as the members’ registration dates. This is obviously unreasonable. Of course, mothers cannot go shopping in the shopping mall just after production. After observing the data, it was confirmed that this was still the default value issue. The default value of the birthday was “today”. This means that information on the baby’s age should be obtained in other ways.

If a member has made a purchase and the applicable age of the product was clear, then the baby’s birthday could be calculated by estimating the age range of the baby when this deal happened. Through this way, whatever the baby’s current age or the age when purchasing, I could easily reckon it.

It was worth noting that with the release of the second child policy, the consumers’ purchase history might contain items purchased for two or more babies. The information on the number of babies is also very valuable to the company. Considering the simplicity of the algorithm, the author decided to ignore the situation of more than two births. If the minimum birthday extracted from the purchase history differs from the maximum birthday by two years, or the purchase records contain the products applicable for two genders,
the family is considered to have at least two babies. Through this algorithm, 1.13 million records of baby’s age out of 2.93 million members was extracted. Figure 4 and figure 5 below respectively show the elder baby’s age distribution and the younger baby’s age distribution.

![Figure 4: The Elder Baby’s Age Distribution (0–Years Old)](image1)

*Source: Goodbaby*

![Figure 5: The Younger Baby’s Age Distribution (0–Years Old)](image2)

*Source: Goodbaby*

### 3.3.6. Baby’s gender

Baby’s gender also needs to be collected from the purchase history, but only if the goods purchased have gender identification, the baby’s gender can be discerned. When only one
gender was distinguished from history while the number of babies has been judged to be two, it was assumed that both babies were of the same gender (For example, if a family having two babies have only bought the products applicable for boys, then both of the children are considered to be boys). Of course, this algorithm would cause unavoidable errors, which is a limitation in this research. Figure 6 below shows the distribution of baby’s gender.

**FIGURE 6: THE DISTRIBUTION OF BABY’S GENDER**

*Source: Goodbaby*

### 3.3.7. Address

“Address” is a relatively easy-to-obtain variable. As long as the member has made a purchase, the delivery address or the store address would be used as the member's address. The data preparation process to be done in this stage was to subdivide the address into three variables which were province, city and district so that we can analyze the overall geographical distribution of members, as shown in Figure 7. The data showed that about 70% of the addresses were accurate to the city. After checking the entire distribution, it was found that the provinces with the largest number of members were Jiangsu Province, followed by Guangdong Province.
3.3.8. Shopping Guide

In order to eliminate the interference of the data of shopping guides on the model, all the members whose mobile phones have been registered as Goodbaby employees were excluded. Of course, there were a lot of mothers who were both consumers and employees, but in order to alleviate the impact of outliers on the model, I still removed the transaction records of these 20 thousand accounts. However, after the rejection, there was still a lot of abnormal data (consumption behavior was too frequent). After investigation, most of the data was still proved from shopping guides. The reason for not being removed with the steps above was that these shopping guides all used another phone number for registration. When the consumers were not willing to become members in the store, the shopping guides logged in with their own member accounts to perform the checkout operation. So, these consumption records include the purchase behavior of multiple consumers.

Considering that it was impossible to collect all the registered phones of these shopping guides, an algorithm was written to solve the problem. Once a member spent more than three orders a day, it would be marked as a shopping guide and his or her purchase history would be excluded. The algorithm could eliminate most of the abnormal data, but there were still some accounts slipping through the cracks. I could do nothing but see if the
result could identify this part of the data automatically after modeling. After deleting all suspected anomalous data, there were 2.83 million members left available for analysis.
4. MODELING

For the topic of customer segmentation, one frequently-used method is Cluster Analysis. In China, the process of clustering has an interesting metaphor that means "Birds of a feather flock together". With some specific business indicators, observed objects can be divided into different groups according to their similarity and dissimilarity. After division, the similarity among objects within each group will be high and the objects in different groups will have a high degree of dissimilarity with each other. In this thesis, members have been clustered according to their consumption behavior. Then I tried to understand the different characteristics of customers in each group.

After taking the time to extract these key columns, it was found that except for gender and address, the situation of missing values was so serious that these variables could not be directly applied into the modeling process. Moreover, there was no other way to make predictions of the missing values. Anyhow it was of important meaning for the company to understand the customer segments combined with their demographic properties. In practice, there has been a mature and effective method to solve this problem by combining non-clustering variables with clustering results for analysis, refinement and mining. This complementary approach has been widely accepted and adopted in the application of Cluster Analysis (Lu, 2013). Therefore, for the variables with a large number of missing values, I did not have to put them into the model for clustering, but to combine the clustering results with them to analyze, which not only had no negative impact on the clustering results, but also made these indicators play a role in my analysis results. This thesis would use the more complete variables extracted from consumption records for modeling so as to classify the members, then combine the clustering results with the attributes mentioned above for the thorough analysis to obtain conclusions that would be instructive to practical operation.

4.1. RFM model

How to measure the value of a consumer according to his or her consumption behavior? How to transform the transaction records into several variables for modeling? RFM model could help me solve these problems.
RFM is a method used for analyzing customer value. It is commonly used in database marketing and direct marketing. It has received particular attention in retail and professional services industries. RFM helps divide customers into various categories or clusters to identify customers who are more likely to respond to promotions and also for future personalization services.

RFM stands for the three dimensions:

- **Recency** – How recently did the customer purchase?
  Last purchase time is an important indicator for maintaining customer relationship. Customers who have recently bought goods or services are the ones who are most likely to buy something from the same retailer again. Moreover, it is much easier to attract a customer who gave his or her patronage to the shop a few months ago than to attract one who went there more than a year ago. Marketers who embrace the strong marketing philosophy of building long-term relationships with customers rather than just selling things will keep customers engaged and win their loyalty.

This thesis would calculate the time interval “Recency” between the last consumption time of each member and September 2, 2018 on a weekly basis. The smaller the R value was, the more attention the member has paid to the brand recently.

- **Frequency** – How often do they purchase?
  We can say that the customers who buy the most frequently are also the customers with the highest satisfaction. If we believe in the concept of brand loyalty, the customers who purchase the most frequently are also the customers with the highest loyalty.

- **Monetary Value** – How much do they spend?
  RFM analysis is based on the marketing axiom that "80% of your business comes from 20% of your customers." If you have a small budget and can only provide the product information to 2,000 to 3,000 customers, would you mail the information to customers who contributed 80% of your income, or to those who contributed 20%? The cost saved
by such marketing strategy would be considerable. Of course, this is just an extreme example. In real life, marketing is often not so simple. This thesis has counted the total consumption amount of each consumer in the past two years.

### 4.2. Algorithm Selection

The algorithms of clustering analysis can be divided into partitioning methods, hierarchical methods, density-based methods, grid-based methods and model-based methods. Among them, the first two methods are the most commonly used (Lu, 2017). To be specific, the two most famous algorithms are K-means and agglomerative hierarchical clustering.

K-means clustering is a method of vector quantization, originally from signal processing. K-means clustering aims to partition $n$ observations into $k$ clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

Hierarchical clustering (also called hierarchical cluster analysis or HCA) is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types:

- **Agglomerative**: This is a "bottom-up" approach. Each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
- **Divisive**: This is a "top-down" approach. All observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

However, the standard algorithm for hierarchical agglomerative clustering (HAC) has a time complexity of $O(n^3)$ and requires $O(n^2)$ memory, which makes it too slow for even medium data sets. Since the dataset has almost three million records, K-means is the best choice for this moment.

### 4.3. Why RFM score formula cannot be applied?
If the standard RFM analysis process is followed, once one has RFM values from the purchase history, it should assign a score from one to five to recency, frequency and monetary values individually for each customer. However, after exploring the data, it was found that about 70% of the members only consumed once, and monetary had a large range and was not evenly distributed. It was difficult to calculate RFM scores on scale of 1-5, which meant that the population distribution of clusters would be uneven. If traditional RFM analysis will be used, it may lead to inaccurate or incomprehensible results. So, in this thesis, the author has used the RFM analysis indicators and the general analysis idea, but the modeling process was slightly different. It would be possible to carry on the modeling with the initial data of RFM after simple standardization through. After the results of clustering analysis have been obtained, the analysis thinking of RFM model would be continued.

### 4.4. The process of K-means

#### 4.4.1. Check the distribution of each indicator

First, the overall distributions of the three indicators were assessed in Table 3.

<table>
<thead>
<tr>
<th></th>
<th>recency</th>
<th>frequency</th>
<th>monetary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.</td>
<td>0.0</td>
<td>1.000</td>
<td>0.0</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>15.0</td>
<td>1.000</td>
<td>68.0</td>
</tr>
<tr>
<td>Median</td>
<td>38.0</td>
<td>1.000</td>
<td>133.7</td>
</tr>
<tr>
<td>Mean</td>
<td>36.7</td>
<td>1.601</td>
<td>335.3</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>56.0</td>
<td>2.000</td>
<td>313.8</td>
</tr>
<tr>
<td>Max.</td>
<td>87.0</td>
<td>281.000</td>
<td>223198.7</td>
</tr>
</tbody>
</table>

Secondly, the specific distribution of every indicator respectively was explored.
As shown in Figure 8 above, the horizontal axis represents the time interval between the last consumption and September 2 which is calculated in weeks. Four weeks represents approximately one month. The vertical axis represents the number of members. The distribution of the time interval was relatively even. 6.4% of members have consumed the last month, 17.8% of members have consumed in the past three months, and 62.4% of members have not purchased anything for half a year. It is worth noticed that 11.4% of members have paid the last order in November last year. In China, November 11th is regarded as “Singles’ Day” standing for a holiday for single people. Taobao, Tmall (both are Alibaba’s e-commerce platforms) and other platforms would carry out all-category promotions on this day, claiming to be the biggest promotion day of the year, which is similar to the Black Friday in America. Over time, people regarded November 11th as a shopping festival by default. This also explained why so many members would purchase products in bulk in November last year.
In figure 9, the horizontal axis of the figure shows the frequency of consumption, and the vertical axis represents the number of members. Since most members' frequency was less than 10 times, the frequency was divided into different lengths. As can be seen from Figure 9, 70% of the consumers have only spent once, which means that they have never consumed again after the first time in the past two years. For this condition in Goodbaby Company, there are several possible reasons. First, they have bought all they needed on discount day like “Singles’ Day”. Second, they once participated in a promotional event for the giveaways and never spent again because of poor shopping experience or lack of relevant demand. Third, online consumers have more choices and tend to try multiple brands. To raise the repurchase rate, the company needs to start with these single-time consumers to find out the reasons.

For members who have spent more than 10 times and less than 50 times, they might be our “golden” members with high values, or loyal members with high consumption frequency but low average consumption. Members who have spent more than 50 times were most likely to be unrecognized shopping guides or individual dealers.
Figure 10 shows the distribution of monetary of all the members in the past two years. The horizontal axis represents the consumption amount and the vertical axis shows the number of members. Similarly, since more than 80% of members spent less than 500 yuan (the unit of money used in the People's Republic of China), the monetary was divided into different lengths. The figure informs us that most of the members spent only a small amount of money, which is also consistent with the phenomenon we saw in the histogram of frequency that there were a large number of single-time consumers. This implies a certain correlation between the frequency of consumption and the amount of consumption. Figure 11 below demonstrates the correlation among the three variables via R. Frequency and monetary do have a positive correlation of 0.46. However, the research’s aim is to classify members instead of calculating their values. Therefore, the issue of the weight of the indicators that may be caused by the correlation between the two indicators would not be considered for the time being.
4.4.2. Standardize the RFM indicators.

Since the dimensions of three indicators are different, the values vary widely. In order to eliminate the influence of large differences in distribution and dimensions, it was necessary to consider centralizing and standardizing the data before modeling. The so-called centralization is the result of subtracting the mean from each row of records. For example, there is a set of data (1, 2, 3, 4, 5, 6, 7). The mean value is 4. The centralized data is (-3, -2, -1, 0, 1, 2, 3). Standardization is based on the centralized data and divided by the standard deviation of the data. In the R language, the data can be centralized and standardized directly with the “scale” function. The following table shows the result in R including several standardized records and the mean value of each indicator within each step.

Table 4: The result of standardization

<table>
<thead>
<tr>
<th></th>
<th>recency</th>
<th>frequency</th>
<th>monetary</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1]</td>
<td>1.38223876</td>
<td>-0.3985771</td>
<td>4.098999e+00</td>
</tr>
<tr>
<td>[2]</td>
<td>-0.31939598</td>
<td>1.5920769</td>
<td>2.965593e+00</td>
</tr>
<tr>
<td>[3]</td>
<td>0.46916646</td>
<td>0.9285256</td>
<td>5.085971e-02</td>
</tr>
<tr>
<td>[4]</td>
<td>0.22014675</td>
<td>4.2462822</td>
<td>1.768406e+00</td>
</tr>
</tbody>
</table>
4.4.3. Determine the number of clusters m.

The most difficult point of clustering with K-MEANS is to pre-set the number of clusters. Unlike hierarchical clustering, K-means clustering requires specifying the number of clusters to extract in advance. Actually, the “NbClust” package in R could be used as a guide. However, due to the large data set and equipment constraints, it was possible to use “NbClust” function to get the best number of clusters. Therefore, another method was adopted to solve this problem. A plot of the total within-groups sums of squares (WSS) against the number of clusters in a K-means solution can be helpful. A bend in the graph can suggest the appropriate number of clusters.
Figure 12 indicates that there is a distinct drop in within-groups sum of squares when moving from 1 to 8 clusters. After eight clusters, this decrease drops off, suggesting that an 8-cluster solution may be a good fit to the data.

4.4.4. Applying K-means for clustering. Obtain eight groups of members.

Due to the different center points of the initial selection, the output of K-means may be different, which means that not every clustering is optimal. With the “nstart” parameter in R, the model was set to run 25 times, and different initial points were selected to return the optimal solution. After running the models, eight clusters of members were obtained. Table 5 below shows the number of members in each cluster.

<table>
<thead>
<tr>
<th>cluster</th>
<th>member_num</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>673094</td>
</tr>
<tr>
<td>2</td>
<td>994983</td>
</tr>
<tr>
<td>3</td>
<td>32273</td>
</tr>
<tr>
<td>4</td>
<td>254362</td>
</tr>
<tr>
<td>5</td>
<td>94</td>
</tr>
<tr>
<td>6</td>
<td>777111</td>
</tr>
<tr>
<td>7</td>
<td>96260</td>
</tr>
<tr>
<td>8</td>
<td>1363</td>
</tr>
</tbody>
</table>
The result gave a BSS/TSS ratio (between_SS / total_SS) of 75.7% indicating a not bad fit. Next, I would define and analyze each group based on the clustering results and other attributes.
5. EVALUATION

5.1. Check the mean of “RFM” in each cluster and compare with the total means

First, the average values of “RFM” in each cluster with the overall means of “RFM” were compared. There were two possible outcomes for each comparison which were either greater than (or equal to) the total means or less than the total means. After several rounds of comparison, the overall situation of the eight groups of members was emerged. Based on the comparison results, Table 6 represents the eight groups.

<table>
<thead>
<tr>
<th>cluster</th>
<th>member_num</th>
<th>recency_mean</th>
<th>frequency_mean</th>
<th>monetary_mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>673094(23.79%)</td>
<td>↑70.3</td>
<td>↓1.16</td>
<td>↓200</td>
</tr>
<tr>
<td>6</td>
<td>777111(27.46%)</td>
<td>↑42.8</td>
<td>↓1.26</td>
<td>↓207</td>
</tr>
<tr>
<td>2</td>
<td>994983(35.16%)</td>
<td>↓13.7</td>
<td>↓1.2</td>
<td>↓189</td>
</tr>
<tr>
<td>7</td>
<td>96260(3.40%)</td>
<td>↑37.4</td>
<td>↑2.25</td>
<td>↑2329</td>
</tr>
<tr>
<td>8</td>
<td>1363(0.05%)</td>
<td>↓16.5</td>
<td>↑15.43</td>
<td>↑15659</td>
</tr>
<tr>
<td>3</td>
<td>32273(1.14%)</td>
<td>↓16.2</td>
<td>↑10.11</td>
<td>↑2037</td>
</tr>
<tr>
<td>4</td>
<td>254362(8.99%)</td>
<td>↓21.7</td>
<td>↑3.94</td>
<td>↑578</td>
</tr>
<tr>
<td>5</td>
<td>94(0.00%)</td>
<td>↓7.7</td>
<td>↑41.89</td>
<td>↑73825</td>
</tr>
<tr>
<td>total_mean</td>
<td>36.7</td>
<td>1.601</td>
<td>335.3</td>
<td></td>
</tr>
</tbody>
</table>

As can be seen from the above table, the frequency of members in clusters 1, 2 and 6 is relatively low at around one time, and the average monetary is also below 200 yuan. The main difference among these three clusters is the value of recency. The members of cluster 1 have not consumed for nearly one year and a half. The last time members of cluster 6 spent with GB was ten months ago, which was just around the date of the Singles’ Day which has mentioned above. Members of cluster 2 have consumed once the last three
months. These three types of members account for up to 86% of the total members. They should be focused on and analyzed in depth.

The members of cluster 7 account for 3.4% of total members. Their general frequency is not very high. On the average, each person in cluster 2 has purchased two times, but the monetary is much higher than the average level which is more than one thousand yuan. However, their last purchase time is more than half a year ago. How to maintain these valuable customers is a key question worthy of further investigation.

Clusters 3, 8, 4 and 5 are all members who have consumed in the past six months and whose consumption frequency and total amount are much higher than the average. Cluster 8 has only about 1,000 members, but the frequency and the monetary of members in this group are quite high. Their consumption capacity per order is similar to that of the members in cluster 7. They are likely to be the most valuable members of GB which could be called “golden members”. The members of clusters 3 and 4 account for 10% of total members which is much higher than other clusters except for clusters 1, 2 and 6. The average frequency and the average monetary of cluster 3 are much higher than cluster 4, but the consumption capacity per order of the two clusters is almost the same which is below 200 yuan. Cluster 5 is suspected to be the group of unremoved shopping guides or personal dealers because of its excessively high frequency and monetary (an average of two shopping times per month and around 1800 yuan per order).

After viewing the means of RFM of each cluster, the eight groups were reclassified according to the frequency and the consumption capacity per order. Clusters 1, 6 and 2, clusters 7 and 8, clusters 3 and 4, cluster 5 have been classified into four big classes separately, which provided a new path of thinking for the following analysis. Next, the detailed distribution of the three indicators in the eight clusters were checked to define the consumption behavior of each group.

**5.2. Deep analysis of eight clusters**

The above analysis gave an idea of the overall level of the three key indicators in the eight clusters, but the mean value method might mask certain characteristics of every cluster.
Next, the detailed distribution of each indicator in every cluster was investigated respectively. Each indicator was presented in the most accessible way. Figure 13 shows the detailed distribution of recency. Each row in different color represents one of the eight clusters and the size of the square represents the proportion of the number of members in the current cluster. The larger the square is, the more members there are in this range (0 represents 0 to 4 weeks). The following Figure 14 and Figure 15 represent the frequency distribution and the monetary distribution of each cluster. The shaded color represents the proportion of the number of members in the current cluster. The darker the color is, the more members there are in this range.

![Figure 13: Recency Distribution of Each Cluster](image)

FIGURE 13: REGENCY DISTRIBUTION OF EACH CLUSTER
Na Ni: Customer Segmentation of Goodbaby Customers

**FIGURE 14: FREQUENCY DISTRIBUTION OF EACH CLUSTER**

**FIGURE 15: MONETARY DISTRIBUTION OF EACH CLUSTER**
Clusters 2, 6, 1

By investigating the distribution maps of clusters 1, 6 and 2 in the three figures above, it was known that all the members in these three clusters consumed less frequently. More than 80% of them only spent once. Moreover, the consumption amount was also lower than other members. 60% of them spent less than 200 yuan in total. The only significant difference between the three clusters is the last purchase time. Members of cluster 1 have not consumed for at least thirteen months, and the last time when members of cluster 6 consumed was between the past six to thirteen months, most of which was in November. While members of Cluster 2 have consumed within half a year which means they are still active. Because members of the three clusters mentioned above account for more than 85% of the total members and the small initial purchase amount does not mean that the consumption capacity is definitely weak, it is crucial to conduct in-depth analysis and establish a sound marketing plan for this part of members, especially for the members of Cluster 2, who are still unfamiliar with the brand.

5.2.1. Cluster 1: Lost - Worst recency, frequency and monetary level. Hard to revive.

Since the members of cluster 1 have not consumed for at least 13 months, they have almost been defined as “lost” members. Therefore, it is difficult to retrieve them. However, this observation could not simply generalize the reasons for the loss because the quantity of the lost members is not the minority. Before summarizing the reasons for loss, it is important to find out their status before they lost. It is not known how long they have been GB’s members in view of the transaction data merely covering the year of 2017 and 2018. Fortunately, there was another variable, the date of the first order. Then the time interval between the first order and the last order was redefined as the consumer's life cycle. If a member's life cycle is within seven days, obviously he or she is still a freshman who can be called “early buyers”, meaning that the member has little knowledge about the brand and is in the early stage of the life cycle with the first attempt.

Among these lost members, 79% of them were early buyers which means that they have never looked back after consuming that week. About 11% of the lost members had a life cycle of six months, approximately 8% had a life cycle of six to thirteen months, and only
2% had one of more than thirteen months, showing that the majority of the lost members were fresh members instead of regular customers.

It is difficult to retrieve these 79% early buyers due to the fact that they had little comprehension of the brand and so they lacked the so-called loyalty. In addition, the time they lost was quite long, which proved that all the promotional activities failed to retrieve them. For this part of members, it is of much higher value trying to find the possible reasons for the loss from their purchase history than regaining them.

For instance, if a customer purchased a gift box, we can speculate that the gifts were prepared for their relatives or friends, indicating that they do not have a child or have no demand for baby products anymore. If there were a large number of members asking for changing or refunding, it means that the product did not meet their expectations in some respects. If there were numerous complaints about after-service, it indicates that there were problems in after-sales service. There might be various causes for the loss, it will be a considerable project to find the sticking points and improve them. In the light of the current records, most of the lost members have purchased FMCG (fast moving consumer goods) such as wet wipes, toiletries, nursing products and etc. 18% of the lost members buying wet wipes just paid no more than five yuan or even close to zero, indicating that this part of members registered and consumed in one promotion of which the gift or the promotional items was a bag of wipes for they do not have real requirements. I examined all the members who have participated in similar activities and found that 40% were in the lost (cluster 1) or dormant status (cluster 6) while about 30% were active (cluster 2, new members). Customers were not growing as fast as they were leaving. This phenomenon also illustrates that using items such as wet wipes, mats or laundry soaps as gifts could indeed attract enough customers, but the strategy for activating them after the first conversion was imperfect which caused that these members had no further desire for consumption. On the other hand, this sort of goods is not only available for babies, but also for adults, resulting in that the attracted members were not necessarily GB’s target group.

Gift-giving can assist the company in raising member’s conversion rate and collecting the information of potential customers, but if the customer relationship management in
the later period is barely satisfactory, these "guest sources" would only become "lost" members.

Among the rest lost members (not early buyers), 16% of them had a total consumption of 500 to 1,000 yuan (about twenty thousand people) in their “life cycle” in GB, and 12% had a total consumption of more than 1,000 yuan (about seventeen thousand people). Their life cycle all exceeded one week and they all purchased GB’s products more than once. They had a certain understanding and loyalty to the brand and they also possessed considerable spending power. There might be several reasons that could result in losing such as no longer demand as the baby growing up, the terrible ultimate shopping experience, shifting to other brands or other causes that were difficult to acquire from the existing data. The data shows that the babies of these valuable lost members were mostly aged one to five years old while the products of GB are mainly available for babies from newborn to three years old, which means part of them were naturally lost.

Although it was mentioned before that most of our members came from the e-commerce platform, it is still worth noting that the consuming channel of 98.5% of the lost members was e-commerce. The former phenomenon is easy to understand because as long as you have an account of the EC platform (such as Tmall account), you can purchase GB’s products on the platform without register again. It is obvious that re-filling information for registration could result in a low conversion rate of offline consumers. However, the data suggested that 24% of all the online members belonged to cluster 1 (Lost members) and 27% belonged to cluster 6 (Hibernating members), while only 14% of offline members belonged to cluster1 and 18% belonged to cluster6. The situation demonstrates that members of e-commerce are more likely to lose. Their loyalty is lower than that of offline members. Consequently, how to attract more offline members, as well as how to improve the loyalty of online members, is the first challenge for GB.

Generally speaking, there was no lack of valuable customers among the lost members. However, it is better to summarize the experience and check the changes of the members’ status regularly, then intervene in time when they are dormant or have a tendency to lose instead of consuming a lot of energy to regain the members who have already been lost. It may bring about great effects to adopt methods such as new product trials, promotional
events, customized product recommendation, customer concern, etc. Without a doubt, the specific type of intervention depends on what style of members they were before.

5.2.2. Cluster 6: Hibernating - Last purchase was long back, low spenders and low number of orders.

The last shopping time of members in cluster 6 was between six months and thirteen months ago, which I called the hibernate period. The next state of hibernating members was most likely to be lost if without intervention. Once in a state of loss, waking-up would turn into a much tougher thing. 80% of members in Cluster 6 were early buyers as same as the situation in Cluster 1 (Lost members), meaning that they have not consumed anymore after spending once. Customer retention is becoming increasingly essential. According to Bain & Co, increasing customer retention rates by 5%, can increase profits by 25% to 95%. The probability of selling to an existing customer is 60-70% compared to 5-20% to a new customer (Marketing Metrics). For this part of members, the company can premeditate the following methods to attempt to wake up them.

a. Promotional information

It is worth mentioning that 30% of the hibernating members had their last shopping time concentrated in the Single’s Day last year. This phenomenon indicates that these members are sales-sensitive. Large-scale promotions may evoke their enthusiasm for shopping. Therefore, besides the Single’s Day, the information about big sales on Children's Day, outlets and other promotion activities can be sent to them in advance by SMS, official accounts or other means. It is certainly that messages should not be sent too frequently, which can lead to resentment of customers.

b. Convert online members to offline members by offering the store's new information and coupons when they arrive nearby.

From the analysis above, we can conclude that online members have lower loyalty than offline members; therefore, resulting in higher churn rate. Especially for maternal and child products, the product exhibition and introduction by professional shopping guides, as well as the sharing of parenting experiences can leave a profound
impression on customers, which brings about more opportunities to establish the value of the brand in their minds. In cooperation with related companies, it is technically possible to locate members in time through the APP on the mobile phone. When they are close to the relevant store or shopping mall, APP could immediately transmit the product information and preferential activities prepared by the store. In the point of consumers' view, they can grasp which activities are going on and decide whether to enter the shop when they pass by, which can yet be regarded as a convenient and modern shopping experience. However, the acquisition of such information is passive. In China, more and more consumers are aware of the importance of personal privacy, so this kind of “modern” shopping pattern may cause panic in “information disclosure”. Although there is no technical obstacle, it still needs time and improvement to make consumers informed and accept this new shopping mode.

c. Provide online and offline parenting trial lessons to strengthen consumer’s dependence on the brand.

Among the hibernating members, most of them have purchased skin-friendly supplies like wet wipes, pacifiers or toiletries, indicating that they have a certain degree of recognition for the safety of the brand. Most of these people are parents for the first time and lack a wealth of knowledge about baby care. It can be easier to establish connections with consumers and increase brand exposure by offering some babysitting courses instead of product recommendations.

5.2.3. Cluster 2: Promising - Recent buyers, but haven't spent much.

As the first two groups, more than 80% of members in Cluster 2 were early buyers, meaning that they did not know much about the brand and were in the initial phase of the member's life cycle. For GB, they are still fresh and active members. In terms of the company, “Fresh” means new. The company does not know so much about them that their ability and preferences of consumption cannot be accurately assessed. After all, the low spending of the first purchase does not mean that the consuming ability is weak. Perhaps it was just an attempt on the brand. “Active” represents that the members have consumed in the last six months, so it would be easier for the company to get in touch
with them again and establish the value of the brand. In a word, they were promising customers. Therefore, enough attention should be paid to this cluster.

Firstly, it is necessary to find out what characteristics these new members possess. In the perspective of geographical location, about 30% of them come from Jiangsu Province, Zhejiang Province and Guangdong Province due to the affection of GB’s business strategy and store distribution. Among the thirteen prefecture-level cities in Jiangsu Province, Suzhou has the biggest membership ratio of 29% because GB is headquartered in Kunshan, Suzhou. It declares that local citizens are more responsive to local brands, so it is a good choice to invite local members to participate in user surveys or product trials regularly.

It is difficult to determine whether the baby is currently a second child or the first child because the “fresh” members have just produced a few records of consumption. But the gender and the age of the baby can be surmised by the commodities they bought. In cluster 2, 55% of the members are girls and 45% are boys without considering the second pregnancy. This phenomenon must be unimaginable for China decades ago for the idea of patriarchal tradition was much widespread in the old society, especially in rural areas. Nowadays in China, the dilution of gender discrimination and the favor for daughters have led more and more families to no longer pay attention to their children's gender, and even more incline to have daughters. Although this is not something that a person is able to decide, the change of concept has made the birth rate of girls gradually higher. Another interesting finding is that Shanghai is the only place with a higher proportion of boys than girls. Baby boys account for 53% and girls 47%, while the proportion of baby girls is as high as 58% in the less developed province compared with Shanghai such as Hebei, Shanxi and Henan. This is a phenomenon that cannot be explained by existing data, but more relevant information is expected to interpret this phenomenon in the future.

In the view of the baby’s age, 54% of the babies are newborns, and 30% are one to three years old (the babies of unknown age are excluded). Among them, the proportion of newborns in Beijing and Shanghai is slightly lower, 41% and 44% respectively, which is also related to the low birth rate of the newborns in these two first-tier cities in the past few years. The high proportion of newborns is a piece of good news for the enterprise because
it means that GB has a wide reputation in the maternal and infant industry. Many families have a preliminary impression of GB and basic trust in the brand, and that is why they have chosen GB as their first choice after the baby was born. In these families with newborns, 37% of them purchased baby strollers and 28% purchased jumpsuits. The price of the baby stroller usually reflects the spending power of a family.

Among the promising members who have purchased baby strollers, 31% of them chose the price range of 600 to 800 yuan, 25% chose 200 to 400 yuan, 19% chose 400 to 600 yuan, and 14% chose 1000 to 1500 yuan. Considering 800 yuan as a cut-off, if the consumption of a stroller of 800 yuan or above represents high spending power, 24% of them have a strong spending power. In other words, if a new member purchases a product with a large price range, we can estimate the member’s spending power through the price, then we can recommend the commodities in accordance with the member’s spending power in the later stage to increase the look-to-buy rate. For example, we can try to recommend a 1200 yuan car seat to a member who has bought a 1000 yuan baby stroller to further explore her demand and spending power. Excellent after-sales services and a return visit via phone are also good for retaining these valuable promising members.

In addition to buying big-ticket items like baby strollers and car seats, there were still a large number of members who have purchased feeding supplies (such as bottles, pacifiers, etc.), toiletries (such as baby shampoo, shower gel, etc.) and wet wipes. If the member has only bought this kind of low-priced supplies, it is quite difficult to measure the buyer’s spending power. In detail, 20% of members in cluster 2 purchased feeding supplies, 16% purchased toiletries, 16% purchased wet wipes, and 7% purchased safety products (e.g. tub, safety scissors for babies, medicine feeder, etc.). The general characteristics of these products are that the requirements for quality and safety level are relatively high while there is no high requirement for the design and style of the products and the unit price is generally low with no more than 100 yuan. Except for toiletries and wipes which are FMCG (fast moving consumer goods), others are durable goods that do not need to be repurchased in a short time. If the company wants to stimulate the repurchase rate of these consumers, some other related products should be recommended to them. For example, if one buys a bathtub, we can recommend toiletries. If one buys a feeding-bottle, we can recommend a medicine feeder, a plate, etc. The rule of recommendation can be speculated
from the receipts of other buyers. If one buys FMCG like wet wipes, we can offer discounts for bulk purchase and other product recommendations which can better reflect the brand value. Offering free trials can assist them in knowing about other new merchandise.

In summary, there are several measures which can be taken to carry out follow-up relationship maintenance with promising members.

a. For local members with high recognition and responsiveness, product research and new product trial activities cannot only increase their brand loyalty, but also assist the company in receiving feedback on product use then making timely adjustments and optimizations. Local members account for a large proportion of all the promising members, but most of them only made one-time consumption through two big outlets each year and there is not much attention paid to new products. If they can participate in the research and development of new products increasing their involvement, thus promoting their willingness to purchase new products, I believe they will bring higher value to the company based on their huge population.

b. According to the content of the first orders of these new members, the corresponding strategy should be made in time. For the members of whom the first-time purchase contained durable products with a wide price range (such as baby strollers), they have shown the potential of becoming loyal or high-value customers. Perfect after-sales services and timely return visits via phone are required to manage the relationship with them.

Based on the price they chose; the consumer capacity can be estimated. Then we can recommend other related products that match their spending power. For example, if one buys a baby stroller with a high price, we can recommend an infant walker with the same quality standard. Of course, recommendations should be made at the appropriate time and with persuasive price offering. For high-value members with high consumption amount, regular tracking and return visits can prevent their loss to some extent. However, if the first order only includes low-priced durable goods like feeding bottle, bathtub, etc., it is necessary to promptly recommend other related products to increase the repurchase rate and cultivate their purchasing habits. For instance, a baby-specific sunscreen cream can be recommended when someone has purchased a
baby hat. If the first order only contained FMCG like wet wipes, hand sanitizers, etc., discounts in bulk purchases can facilitate their secondary consumption. The recommendation of other products with some brand characteristics and the provision of free trials can increase consumers' understanding of product dimensions, not just of current products.

In general, for new customers with a single purchase, quickly improving their understanding of the brand as well as promoting the secondary purchase is the top priority before the state transition.

5.2.4. Cluster 7: Potential Valuable - Spent a good amount and bought more than once. Responsive to promotions.

Cluster 7 accounts for 3.4% of the total members at around 96,000. Although their purchase frequency was not as high as the company expected, they have consumed more than once in GB. Their spending time had a certain periodicity and they generally liked to shop on big promotional days. The average purchase amount was usually more than one thousand yuan per time. There is no doubt that they are potential valuable consumers who are worthy of focus. Therefore, the company needs to get familiar with their consumption behavior, shopping preferences and interested products, provide them with visible membership benefits, increase the frequency of purchase and thus enhance their brand loyalty.

In order to understand their behavior, it is essential to find some potential characteristics on the category of merchandise, the price level or the purchasing channel during their first purchase. Since the dataset only includes the transaction records in 2017 and 2018, only 84 thousand members who registered after 2017 were extracted for the analysis so as to ensure the integrity of their life cycle, which means that one’s first record in the dataset was indeed her "first order." After analysis, it was found that there were 44 thousand (in Cluster 7, a total of 56 members bought the baby stroller) out of the 84 thousand members who have purchased baby strollers during their first purchase time, accounting for over 50%. 38% of the members who bought baby strollers chose to buy baby stroller accessories together, such as mats, blankets, etc. The second place in the first shopping
list is bedding (32%) including bed and bedding, followed by car seat (22.3%), feeding supplies (13.2%), wet wipes (10.5%) and toiletries (10.5%).

In terms of price, 71% of the members who bought the baby stroller chose the price range of 1,000 to 2,000. 74.5% of the members who bought the bed chose the price range of 800 to 1,500 and 85% of the members who bought the car seat chose the price range of 1,000 to 3,000. It is worth mentioning that the price range of these durable goods is extremely wide, as low as two hundred yuan or less, up to six thousand yuan or more. In general, most of the potential valuable members have bought the higher-priced durable goods in their first order. Or in other words, the members buying high-priced durable goods are most likely to be the potential valuable members, who will bring more value to the company in the future.

In addition, the features of such durable goods are that the first use requires some guidance and subsequent use may require after-sales service or replacement of accessories. This brings the company an entry point for customer relationship maintenance.

In order to explore the member's next consumption intention, the second order of the repurchased members can be examined. For members who have purchased a stroller in the first order, 30.4% would purchase baby stroller accessories in the second order, and 15.5% would purchase feeding supplies. Unexpectedly, 15.3% would buy the baby stroller again. Excluding the return and exchange situation, 10.6% of the members would buy the second new baby stroller immediately after purchasing the first one. After analyzing the time interval between two purchases, which can be seen in figure 16, it was found that more than 50% of the members purchased the second stroller within one month, so the conjecture for the reason of the second child was ruled out.
After looking over the original data, it was discovered that two main reasons for this surprising phenomenon can be speculated. First, the second stroller has different usage scenarios and functions. For instance, the member has first bought a medium-sized stroller then bought a pocket stroller. The former one has great effects of insulation and anti-seismic so it can be used for ordinary egress. The latter one can be used for aircraft travel for its light weight and dimensions. Second, they bought the same or similar strollers for their relatives or friends a few days after the first purchase. This is indeed a remarkable discovery. It not only provides another way of thinking for joint sales strategy, but also shows that mothers are born with the gift of publicizing products through word of mouth (WoM). Each of them has the potential to spread the value of the brand.

Among the remaining members, 12% of the 84,000 members bought toiletries or bedding, 9.6% bought wet wipes, and only 8.8% bought car seats, which proved that the parents usually do not buy car seats at birth, or baby strollers and car seats are often not purchased at the same time. Therefore, the current joint sales strategy of strollers and car seats is not effective.

Similarly, the members who have purchased bedding for the first time would still prefer to buy bedclothes most for the second time, then followed by baby strollers, feeding supplies, safety products (safety scissors, nail clippers, diapers, etc.). Also, most of the members who bought the car seat for the first time purchased matching accessories for the second time, followed by baby strollers and car seats. If the first order of the members only included relatively low-priced and fast-moving products, the content of the second order could be anything, which means I found no rules from it.

Combined with the above analysis, the following conclusions can be made. The average expense per order of the most members in cluster 7 is higher than 1,000 yuan, because the majority of members have purchased higher-priced durable products such as baby stroller, car seat or bed. The common features of this kind of products are that the acces-
ories can be replaced and it needs instructions for the first installation. Such characteristics have motivated the company to have more natural contact with these customers and to leave a deeper brand impression on customers.

For the customers who have purchased durable products are most likely to buy the relevant products such as the matching accessories or the goods of the same category at the next time. The discovery lets us realize a new way of joint sales as well as the mother’s natural ability of product propaganda. How to serve these KOLs (Key Opinion Leader) and arrange with them to play a bigger role in brand promotion will be another topic worth exploring.

Furthermore, the data shows that the baby stroller and the car seat were often not purchased at the same time or purchased continuously. The parents often prepare the stroller before or just after the baby’s birth, while the car seat will not be bought until the baby is a little bit older. However, there is no rule to find from the content of the second orders if the content of the first orders only contains FMCG.

Even so, the repurchase rate of them was far higher than that of members who have purchased the durable goods. It must be related to the life cycle of the product itself but it also shows that the later type of customers was not familiar with or trusted in other categories of GB’s products especially FMCG to some extent. It is also related to GB’s position as “the king of baby strollers” in this industry. Then, in order to increase the purchase frequency of the potential valuable members, the company has to increase the promotion cost and develop the strategy of joint sales for FMCG. The methods of sending gifts with the purchase of a stroller and offering free trials are both considerable. The display of the stores is also a key factor to concern about.

Fortunately, the members’ sensitivity to big sales makes us gain a clear idea of their purchase frequency so that the company can provide coupons to stimulate consumption on time and intervene in time before they convert into a hibernating state. Of course, attractive membership benefits such as giving points spent as money will also promote the healthy development of the potential valuable members.
5.2.5. Cluster 8: Top Customers - Bought recently, buy often and spend the most!

The members in cluster 8 meet all the highest standards that the company want their members to have. They possess strong spending power, short consumption cycles and price insensitivity, which respectively means that they spend a lot on shopping almost every time, they consume frequently and their consumption time is not limited to the date of big sales. Those who satisfy all the qualities above form the top members of the company. The good news is that there are top members for the company which has just developed its membership system, but the bad news is that top members only account for 0.05% of all the members of GB with accurately 1,363 people. This ratio is far from adequate for a large enterprise as well as a famous brand with more than ten thousand employees. It is precisely because of the small number of the top members that the records of each member are precious, which also implies that outliers can affect the judgment of the whole group. Thus, before the process of deep analysis, the “individual retailers” that may exist in the group of top customers again had to be removed.

The main performance of their consumer behavior is to purchase a large number of goods of the same type at one time, but for durable goods, FMCG and clothes, the number of pieces used to determine whether the order is “abnormal” is not the same. After reviewing the overall distribution of the number of pieces, it was decided to mark the member who has purchased more than five pieces of the same durable goods or more than ten pieces of the same clothes or more than one hundred pieces of the same FMCG one time as an individual retailer, then excluded them from the scope of analysis finally obtained 1,291 top customers. For these valuable members, the company must not only concentrate on maintenance, but also understand their characteristics so as to tap or foster more top members.

Let's first pay attention to consumer behavior of top members. In terms of life cycle, 40% of these members have a life cycle of more than thirteen months and 24% between six months and thirteen months. From the category of goods purchased, their shopping lists
are more comprehensive compared with other clusters which have mentioned above, generally including baby strollers, feeding bottles, toiletries, wet wipes, outerwear, bedding, underwear, accessories, safety products, diapers and so on.

On average, each top member has purchased at least 5 categories of products during their life cycle and usually purchased 2.3 categories per purchase. In terms of the number of pieces, each member purchased an average of 6.7 items per order. This kind of behavior has clearly distinguished this group of members from other groups. They have purchased almost everything needed for their children during the babyhood in GB, from diapers to wet wipes, from underwear to outerwear, from bedding to strollers. In particular, the purchase enthusiasm of outwear has almost surpassed that of FMCG (except wet wipes), which was less common in other groups of members.

It is worth mentioning that the trust has been shown from the first purchase. From 778 members whose transaction records are whole (they registered as members after 2017), the analysis shows that each person has purchased an average of 3.6 categories and 10.4 pieces of goods during the first purchase. This behavior implies that these customers had a comprehensive understanding and trust in the brand before the baby was born. Therefore, if the company wants to activate or cultivate more top members, the spread of brand value and product information needs to be fully laid out before the members arrive at the store.

It is also worth noting that 56.2% of the orders paid by top members were completed in the stores, 9.2% were completed on the company's own APP, and only 33.5% were completed on Tmall (third-party e-commerce platform). It represents that top members were more inclined to shop offline, which also confirms previous inference that offline consumers have higher loyalty to the brand. In addition, 23% of top members have used more than one way to make purchases.

After learning about the spending habits of top members, let's take a look at their own demographic attributes. Surprisingly, 44% of top members come from second-tier cities, 21.8% are from third-tier cities, and only 14.6% are from first-tier cities. To be specific, as shown in figure 17 below, nearly 50% of top members are from Henan, Shandong and
Jiangsu these three provinces. In Jiangsu Province, although there are so many members in Kunshan City where GB’s headquarter is located, the number of top members in Kunshan is less than Nanjing, ranking second. Facts prove that people in second- and third-tier cities have greater recognition of GB than first-tier cities, and their brand loyalty is also higher. Of course, this may be related to fewer brand choices in second- and third-tier cities. In addition, another notable feature is that 74% of top members have a family with two babies, most of whom consist of one girl and one boy. 87% of the elder babies are over three years old now, and most of the smaller babies are under three years old, which means that their demands cover a very wide age range.

![FIGURE 17: GEOGRAPHICAL DISTRIBUTION OF TOP CUSTOMERS](image)

After fully understanding the basic characteristics and consumption habits of top members, the company can try to improve the operational strategy from the following aspects.

a. Increase the publicity of the membership system in second- and third-tier cities, not only in first-tier cities. The benefits like reward points, exclusive discounts for different levels of members, new product trials cannot only help the company maintain these top members, but also enable shopping guides to quickly identify these high-
value members and provide exclusive one-to-one services. It is also more accurate and efficient in the recommendation of goods.

b. The emergence of top members has made us aware of the importance of “pre-branding”. In order to tap more valuable members, how to make the potential customers know about the key strengths and product features of the brand then develop trust before they arrive at the store is a difficult point that the marketing department needs to consider about. For instance, the choice of the spokesperson, the laying of physical advertisements, title sponsors and live web broadcast which is quite popular in China in recent years may be important marketing tools that can do more with less. In addition, consulting these members about the way to know GB and the reasons for choosing GB will also provide the company with a good reference.

c. Cooperate with the top members to promote the brand. Such frequent consumption and high spending power make them the best experiencers and promoters of the brand. The appreciation of the products as well as the comprehension of the baby makes them naturally have the ability to promote the products and summarize the selling points. Collaborating with them so as to drive the potential customers around them to explore the brand is an efficient marketing method.

5.2.6. Cluster 4: Potential Loyalist - Bought several times but spent not much.

Members of Cluster 4 account for 9% of the total members with more than 25 thousand people. Compared with other clusters except for cluster 1, cluster 6 and cluster 2, cluster 4 is a relatively large group. Compared with the potential valuable members of Cluster 7, they have the common features that the consumption frequency of the most members was concentrated in two to five times and they were responsive to the sales promotions, but there are also some significant differences between these two clusters. First, the average frequency in cluster 4 was slightly higher than cluster 7 with around four times. Second, on the whole, the last purchase time of the members in cluster 4 was within half a year which indicates that they are still active. However, the average consumption amount per
order of the members in cluster 4 was much lower than that in cluster 7 with an average of 100 yuan to 150 yuan.

The members of both cluster 4 and cluster 7 are the types of members that promising customers (cluster 2) may evolve into in the later stage. Since this type of members have purchased more than once and they have a certain understanding of the brand, but the frequency of purchase was not particularly high, I define them as potential loyalists.

In order to further increase the purchase frequency of potential loyalists and to cultivate them to become our true loyalists, it is necessary to explore their consumption cycle and the categories of products purchased so as to make, e.g., product recommendations and deliver coupons in a timely manner. Under the premise of maintaining high frequency, the company should consider about how to stimulate their consumption and increase their expense per order by increasing the probability of success in joint sales. So, it is essential to know their purchase preferences and whether they still have potential demand for other category of goods (such as baby stroller, car seat, etc.).

From the analysis of the consumption cycle, the average time interval between two consecutive purchases of 14% of members was within one month, and 58% was between one and four months, see Figure 18.

![FIGURE 18: THE CONSUMPTION CYCLE OF THE MEMBERS IN CLUSTER 4](image)

In terms of the product categories, most of them have bought FMCG and apparel. The specific distribution is shown in Table 7.
The orders including baby strollers only account for 2.3% of the total orders paid by members of cluster 4 (only 21 thousand members out of the 250 thousand members have ever purchased the stroller) and other durable goods have fared even worse among the members of cluster 4, which explains why the average expense of this group was at such a low level. What’s more, 80% of these orders contained only one category of merchandise, and 14.5% contained two categories, standing in stark contrast to the consumption behavior of top customers. In order to explore the reasons behind this phenomenon, the consumption channels for generating these orders have been assessed.

Data shows that 94.5% of the orders were completed on the Tmall platform, which indicates that most people are accustomed to searching for merchandise categories rather than brands when shopping online with aims. When they need other category of products, they are used to searching again. But at this moment, search results will show different brands of goods again, so customers are likely to no longer choose the last selected brand unless there is money off. Finally, consumers’ orders will be assigned to the corresponding merchants, and each merchant's order will often contain only one category of merchandise. If people do not have a preferred brand, this phenomenon often happens. Therefore, the brand effect is weakened a lot in this shopping model. For this reason, the company should consider selling packages with different relevant products on each product page, not just on the store homepage. Besides, in order to improve the effectiveness of product recommendation and coupon issuance, it is indispensable to understand the natural usage cycle of each category of FMCG, and predict the next replenishment date of the member based on the number of pieces the member purchased. For example, on average, a mother uses about three diapers a day, and purchases two bags (22 pieces per bag) of diapers at

<table>
<thead>
<tr>
<th>Product Category</th>
<th>Wet Tissue</th>
<th>Feeding</th>
<th>Bathing</th>
<th>Underwear</th>
<th>Baby Safety</th>
<th>Accessories</th>
<th>Bedding (Furniture Excluded)</th>
<th>Outerwear</th>
<th>Strollers</th>
</tr>
</thead>
<tbody>
<tr>
<td>% of Total Orders</td>
<td>30.6%</td>
<td>21.7%</td>
<td>17.7%</td>
<td>11.9%</td>
<td>8.9%</td>
<td>8.9%</td>
<td>8.2%</td>
<td>5.5%</td>
<td>2.3%</td>
</tr>
</tbody>
</table>

Table 7: Commodity Type Distribution of Cluster 4 (Potential Loyalist)
a time. Then, on the second week after the purchase, the coupon for diapers can be delivered to the mother to promote the repurchase, effectively preventing the loss of the member and fostering her long-term loyalty to the brand.

Maintaining a member's repurchase can stabilize its brand loyalty and transform it from a potential loyalist to a real loyalist, and the increase in per customer transaction (the average expense per transaction per customer) is one of the ways to convert a normal member into a high-value member. In other words, high frequency means loyalty while high expense means high value. Among the members of cluster 4, 76% are single-child families and 24% are two-child families. Among the first-child families, 20.8% of the children are newborns, 58.8% are 1 to 3 years old, and 15% are 3 to 5 years old. In the two-child families, 22.2% of the elder children are 1 to 3 years old, 31.6% are 3 to 5 years old and 29.4% are 5 to 8 years old while 29.4% of the younger children are newborn and 48.3% are 1 to 3 years old. Therefore, in terms of durable goods, there may still be potential demand for baby strollers for a total of about 23% of newborns in Cluster 4. However, according to the previous analysis, many parents had already purchased the stroller before the baby was born, thus the recommendation effectiveness at this time may be worse than before the baby’s birth. However, a car seat available for children up to 12 years old can be used for other children other than newborns (the previous data analysis shows that parents rarely buy a car seat when the baby is just born).

Of course, tapping the consumers’ demand for other product categories as well as recommending the products which are related to the categories purchased by them can be a good way to lift their expense, but the actual conversion effect is inseparable from the recommendation timing and method. The estimation of the baby's age needs to be further refined to narrow the age range in order to make more accurate recommendations. In addition, mastering the purchasing behaviors of parents (how many to buy, how much they consume in a day, when to buy the next time, etc.) is also the key to successful conversion. It would be helpful for the company to understand the consumers’ purchasing habits by combining external market research with customer interviews.
5.2.7. Cluster 3: Loyalists – Recent buyers and bought often. Not spent much each time.

The members of cluster 3 account for 1.14% of the total members with about 32 thousand people. Although each of them only spent an average of two hundred yuan per transaction, they have contributed a lot to the company's income because of their high-frequency consumption. Compared to Cluster 4, they have similar per customer transaction and recency, but the difference is that the frequency of cluster 3 is concentrated in 6 to 15 times while the frequency of Cluster 4 is concentrated in 2 to 5 times. Not like members of cluster 4, the members of cluster 3 are more stable in consumption and more loyal to the brand, so they have been defined as loyalists. The company needs to dig deeper into the consumption patterns of this group so as to provide a reference for the conversion of potential loyalists (cluster 4). Besides, in order to increase per customer transaction of loyalists then convert them into more valuable members, it is still imperative to understand their needs in depth.

From the consumption cycle point of view, the average time interval between two consecutive purchases of 21.5% of the loyalists was within one month, 48.5% was between one and two months (20% of potential loyalists), 24.5% was between two and three months and only 4.2% was between three and four months. Broadly speaking, the average time interval of almost all the loyal members was within four months, with nearly 50% was between one and two months. Compared with loyalists, the time interval of consumption of potential loyalists was more dispersed. In terms of the category of goods, what the loyalists have bought and the order share of each category were quite similar to that of potential loyalists, except for outerwear and diapers. The orders of loyalists including outerwear account for 8.31% and diapers account for 4.82%, which was about 3% more than that of potential loyalists respectively. Outerwear is generally not suitable for newborns, and it is also a weak product line of GB. As the baby grows up, the choice of outerwear is more diverse so the company faces greater competitiveness. The high proportion of outerwear here highlights the loyalty of the members in cluster 3 to the brand. Similarly, for diapers, compared to other famous diaper brands such as Pampers, Merries and Huggies, GB's diapers have fewer styles, less popularity because they are not the
main products of the company. The members who have always bought diapers in GB represented their overall trust in the brand.

In retrospect, since the goods demanded by loyalists and potential loyalists were almost identical and contained a large number of FMCG, why is the consumption frequency of the former type of members so much higher than the latter? After exploring the data, it was found that 19.4% of the loyal members came from stores, official websites or self-operated APP, while there was only 11.3% for potential loyal members. Members from these three channels represent a certain understanding of the brand and the product variety. It also determines the difference in the purchasing channels of the two clusters. In cluster 4 (potential loyalists), orders from Tmall account for 94.5% of the total orders while in cluster 3 (loyalists), the proportion of orders from Tmall was 86.7% and orders from stores, APP and official websites account for about 12%. Generally, e-commerce platforms sell FMCG products in bulk. If a member has purchased a box of wet wipes in one time online, of course, the usage cycle will be longer. But for a member who usually shops offline, the basket generally contains no more than 5 packs of wet wipes. When the wipes run out, the customer will go to the store to buy some again, so the purchase cycle is shorter than that of the members purchased in bulk on the e-commerce platform. This explains to a certain extent why loyalists have to shop more frequently than potential loyalists.

Then another question emerged. Since the demand pool was stable, why was the per customer transaction of loyalists higher than that of potential loyalists? If it was only a difference in frequency, it was not enough to prove the loyalty of the members in cluster 3. Let’s explain this question from another aspect. The data shows that 93.4% of the members in cluster 4 have purchased no more than five categories of products during their whole life cycle, while more than half of the members in cluster 3 have purchased over five categories. This explains why the average expense of cluster 3 was a little higher than that of cluster 4 because their needs were more diversified. If the company wants to realize the conversion of potential loyalists to loyalists, it needs to get them familiar with more product dimensions of GB and enhance brand awareness. Offering free trials with purchase, bundle sales and offline purchase discounts are all common ways to solve this kind of problem.
If GB wants to stimulate consumption of loyalists and improve the associated purchase rate, they cannot start with FMCG for which the demand is stable and limited. Instead, the company should tap their demand for durable goods or clothing such as car seats, shoes, toys because the demand for these kinds of products is unclear and flexible. Regular customer care and return visits can give the company a main idea of the potential needs of these members. Besides, they can also try to get reference answers from the transaction records of members with more purchased categories.

Then, the demographic attributes of loyal members were examined. Similar to the top members, the proportion of two-child households in loyalists was as high as 45%. These families purchased products of GB when both of their babies were born, which reflected their recognition and loyalty to the brand from the side. In the one-child families, 71.4% of the babies are between 1 and 3 years old. In the two-child families, 30% of the elder babies are between 1 and 3 years old, 33% are between 3 and 5 years old, and 24.8% are between 5 and 8 years old, while 36% of the younger babies are newborns, 54.2% are between 1 and 3 years old. As a result, the families with babies between 1 and 3 years old have become the major group among loyalists, and so the products applicable to this age group have also turned into the largest demand of this cluster. The second similarity with top members is that about 67% of loyalists come from second- or third-tier cities. The remaining 33% are from first-tier cities and fourth-tier cities in equal proportions. Specifically, 17.9% of the loyalists locate in Jiangsu Province, followed by 9.3% in Guangdong Province and 9% in Zhejiang Province. Suzhou City, where the company’s headquarter is located, contains 45% of the loyal members of Jiangsu Province. Such similarities may remind us that if the consumption capacity is sufficient, loyalists have the potential to develop into top members.

In order to find out whether there are potential top members with high spending power among loyalists, in a separate assessment, the distribution of the maximum order amount of the two clusters of members has been examined. Among the top members, more than 95% of them had a maximum order amount of more than 1,000 yuan, of which about 53% were between 1,000 yuan and 5,000 yuan, and 42% were more than 5,000 yuan. Among the loyal members, one-quarter of them have a maximum order amount of more than
1,000 yuan, but the orders of more than 5,000 yuan almost never occurred. Anyhow, such data distribution proves that 25% of loyal members have certain spending power and are expected to be converted into GB’s top members.

As for the members of cluster 5, it was judged as individual retailers after the sample survey. They are not the target group for this analysis, so they are not considered at this moment. However, this part of data can be used to extract features so that we can distinguish between real members and individual retailers more accurately and the results are expected to be more in line with the actual situation.
6. SUMMARY

After the customer segmentation, the implementation of the specific marketing methods in the later stage requires a long period of discussion and cooperation among all the relevant departments, so here in this thesis, the last step of deployment in CRISP-DM has been ignored.

To summarize the process of analysis, Table 8 shows the definition and key features of the seven clusters, then summaries several suggested tips distilled from the analysis based on the operation targets. To help the readers review how these suggestions were derived from, Table 9 displays the key particular phenomenon found in each cluster.

### Table 8: Customer Segmentation of GB members

<table>
<thead>
<tr>
<th>CI</th>
<th>Group</th>
<th>Key Feature</th>
<th>Operation Targets</th>
<th>Suggested Tips</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>Lost</td>
<td>Worst recency, frequency and</td>
<td>Summarize the reasons for loss according to the</td>
<td>Possible reasons:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>monetary level. Hard to revive.</td>
<td>purchase history</td>
<td>- Buy gifts for relatives or friends, no follow-up requirements</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- Product problems</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- Problems in after-sales service</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- Bad CRM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- The members attracted by some activities were not necessarily GB’s target</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>group</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Online members account for the majority but they are more likely to lose</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>compared with offline members.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Attract more offline members and improve the loyalty of online members</td>
</tr>
<tr>
<td>6</td>
<td>Hibernating</td>
<td>Last purchase was long back,</td>
<td>Timely intervention to prevent loss</td>
<td>Sending information on promotion</td>
</tr>
<tr>
<td></td>
<td></td>
<td>low spenders and low number of</td>
<td></td>
<td>Convert online members to offline by offering the store's new information and</td>
</tr>
<tr>
<td></td>
<td></td>
<td>orders.</td>
<td></td>
<td>coupons when they arrive nearby</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Provide online and offline parenting trial lessons to strengthen consumer’s</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>dependence on the brand</td>
</tr>
</tbody>
</table>
Table 8: Customer Segmentation of GB members

<table>
<thead>
<tr>
<th></th>
<th>Customer Segment</th>
<th>Characteristics</th>
<th>Strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Promising</td>
<td>Recent buyers, but have not spent much.</td>
<td>Find out their characteristics.</td>
</tr>
<tr>
<td>7</td>
<td>Potential Valuable</td>
<td>Spent a good amount and bought more than once.</td>
<td>Understand their consumption behavior and interested products. Offer visible benefits of membership. Increase their consumption frequency so as to enhance their brand loyalty.</td>
</tr>
<tr>
<td>8</td>
<td>Top Customers</td>
<td>Bought recently, buy often and spend the most!</td>
<td>Maintain the relationship with them carefully. Know about their features so as to tap or foster more top members. Increase the publicity of the membership system in second- and third-tier cities, not only in first-tier cities. Offer the membership benefits like reward points, exclusive discounts for different level of members, new product trials.</td>
</tr>
</tbody>
</table>
Table 8: Customer Segmentation of GB members

<table>
<thead>
<tr>
<th>Cluster ID</th>
<th>Group Definition</th>
<th>Relevant Variable</th>
<th>Variable Explanation</th>
<th>Phenomenon</th>
</tr>
</thead>
</table>
| 4          | Potential Loyalist | Bought several times but spent not much. | Further increase the purchase frequency to cultivate them to become true loyalists. Stimulate their consumption and increase their expense per order. | - For online sales channels, sell packages with different relevant products in each product page, not just in the store homepage  
- Understand the natural usage cycle of each category of FMCG, and predict the next replenishment date of the member based on the number of pieces the member purchased |
| 3          | Loyalist          | Recent buyers and bought often. Not spent much each time. | Dig deeper into the consumption patterns of loyalists so as to provide reference for the conversion of potential loyalists. Understand the differences between loyalists and top members and tap more valuable members from loyalists. | - The average expense of loyalists was a little higher than that of potential loyalists because their needs were more diversified. Get potential loyalists familiar with more product dimensions of GB and enhance the brand awareness  
- Tap loyalists’ demand for durable goods or clothing such as car seats, shoes, toys  
- 25% of loyal members have certain spending power and are expected to be converted into GB’s top members |

Table 9: Particular Phenomenon of Each Cluster

<table>
<thead>
<tr>
<th>Cluster ID</th>
<th>Group Definition</th>
<th>Relevant Variable</th>
<th>Variable Explanation</th>
<th>Phenomenon</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Lost</td>
<td>Recency</td>
<td>The time interval between the last consumption time of each member and today</td>
<td>Have not consumed for at least 13 months</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Life cycle</td>
<td>The time interval between the first order and the last order of a member</td>
<td>The life cycle of 79% of them was within seven days, defined as “early buyers”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Product Type</td>
<td>The type of goods, including FMCG (fast-moving consumer goods) and durable goods.</td>
<td>Most of them have purchased FMCG such as wet wipes, toiletries, nursing products and etc.</td>
</tr>
</tbody>
</table>
### Table 9: Particular Phenomenon of Each Cluster

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Recency</th>
<th>Location</th>
<th>Baby’s gender</th>
<th>Baby’s age</th>
<th>Category of product</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>Hibernating</td>
<td>The last shopping time of members in cluster 6 was between six months and thirteen months ago. 30% of them had their last shopping time concentrated in the Single’s Day last year</td>
<td>About 30% of them come from Jiangsu Province, Zhejiang Province and Guangdong Province. Suzhou City has the biggest membership ratio of Jiangsu with 29%</td>
<td>55% of the members are girls and 45% are boys. Shanghai was the only place with a higher proportion of boys than girls, while the proportion of baby girls is as high as 58% in the less developed province such as Hebei, Shanxi and Henan.</td>
<td>54% of the babies are newborns, and 30% are one to three years old. The proportion of newborns in Beijing and Shanghai is slightly lower, 41% and 44% respectively.</td>
</tr>
<tr>
<td>2</td>
<td>Promising</td>
<td>The time interval between the last consumption time of each member and today</td>
<td></td>
<td></td>
<td>Such as baby strollers, underwear, wet wipes, etc.</td>
</tr>
</tbody>
</table>
### Table 9: Particular Phenomenon of Each Cluster

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Phenomenon</th>
</tr>
</thead>
</table>
| 7       | **Potential Valuable**  
Spending powder | Evaluate their spending power based on the price range of the strollers they purchased  
Cluster 2 purchased feeding supplies, 16% purchased toiletries, 16% purchased wet wipes, and 7% purchased safety products.  
Considering 800 yuan as a cut-off, if the consumption of a stroller of 800 yuan or above represents a high spending power, 24% of them have a strong spending power. |
|         | **First purchase**  
Price range | The price segment of the product  
71% of the members who bought the baby stroller chose the price range of 1,000 to 2,000. 74.5% of the members who bought the bed chose the price range of 800 to 1,500 and 85% of the members who bought the car seat chose the price range of 1,000 to 3,000. |
|         | **Second purchase**  
Second purchase | The goods of the second order of each member  
For members who have purchased a stroller in the first order, 30.4% would purchase baby stroller accessories in the second order, and 15.5% would purchase feeding supplies. Unexpectedly, 10.6% would buy the baby stroller again. Two main reasons have been found that the second stroller has different usage scenarios and functions, or they bought the same or similar strollers for their relatives or friends a few days after the first purchase. Only 8.8% bought car seats during the first purchase. |
| 8       | **Top Customers**  
Life cycle | The time interval between the first order and the last order of a member  
40% of these members have a life cycle of more than thirteen months and 24% between six months and thirteen months. |
<table>
<thead>
<tr>
<th>Category of product</th>
<th>Such as baby strollers, underwear, wet wipes, etc.</th>
<th>Their shopping lists were more comprehensive compared with other clusters, especially including a large amount of outerwear.</th>
</tr>
</thead>
<tbody>
<tr>
<td>The number of categories</td>
<td>The number of different categories of goods ever purchased</td>
<td>On average, each top member has purchased at least 5 categories of products and usually purchased 2.3 categories per purchase.</td>
</tr>
<tr>
<td>Associated purchase rate</td>
<td>The number of pieces purchased per order</td>
<td>Each member purchased an average of 6.7 items per order.</td>
</tr>
<tr>
<td>The number of categories in the first purchase</td>
<td>How many categories the member have bought during the first purchase</td>
<td>Each person has purchased an average of 3.6 categories of goods during the first purchase.</td>
</tr>
<tr>
<td>Associated purchase rate of the first purchase</td>
<td>The number of pieces of the first purchase</td>
<td>Each person has purchased an average of 10.4 pieces of goods during the first purchase.</td>
</tr>
<tr>
<td>Consumption channel</td>
<td>Channel for purchasing items</td>
<td>56.2% of the orders were completed in the stores, 9.2% were completed on the company's own APP, and only 33.5% were completed on Tmall.</td>
</tr>
<tr>
<td>The number of consumption channels</td>
<td>How many different channels the member has ever used to purchase</td>
<td>23% of top members have used more than one way to make purchases.</td>
</tr>
<tr>
<td>The degree of development of the city</td>
<td>It was divided into first-tier to sixth-tier, and the first-tier cities are the most developed.</td>
<td>44% of top members come from second-tier cities, 21.8% are from third-tier cities, and only 14.6% are from first-tier cities.</td>
</tr>
<tr>
<td>The number of babies</td>
<td>How many babies a family (member) have</td>
<td>74% of top members have a family with two babies. 87% of the elder babies are over three years old now, and most of the smaller babies are under three years old.</td>
</tr>
</tbody>
</table>

| 4 | Potential Loyalist Frequency | How many times the member have purchased | Concentrated in two to five times, with an average of four times |
Table 9: Particular Phenomenon of Each Cluster

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>The consumption cycle</td>
<td>the average time interval between two consecutive purchases</td>
<td>14% was within one month, and 58% was between one and four months.</td>
</tr>
<tr>
<td>Category of product</td>
<td>Such as baby strollers, underwear, wet wipes, etc.</td>
<td>Most of them have bought FMCG and apparel. The orders including baby strollers only account for 2.3%</td>
</tr>
<tr>
<td>The number of categories</td>
<td>The number of different categories of goods ever purchased</td>
<td>80% of the orders contained only one category, and 14.5% contained two categories.</td>
</tr>
<tr>
<td>Consumption channel</td>
<td>Channel for purchasing items</td>
<td>94.5% of the orders were completed on the Tmall platform.</td>
</tr>
<tr>
<td>Baby’s age</td>
<td>Estimated current age range of the baby.</td>
<td>76% are single-child families and 24% are two-child families. Among the first-child families, 20.8% of the children are newborns, 58.8% are 1 to 3 years old, and 15% are 3 to 5 years old. In the two-child families, 22.2% of the elder children are 1 to 3 years old, 31.6% are 3 to 5 years old and 29.4% are 5 to 8 years old while 29.4% of the younger children are newborn and 48.3% are 1 to 3 years old</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3 Loyalist Frequency</th>
<th>How many times the member have purchased</th>
<th>Concentrated in six to fifteen times</th>
</tr>
</thead>
<tbody>
<tr>
<td>The consumption cycle</td>
<td>the average time interval between two consecutive purchases</td>
<td>Almost 100% was within four months, with nearly 50% was between one and two months</td>
</tr>
<tr>
<td>Category of product</td>
<td>Such as baby strollers, underwear, wet wipes, etc.</td>
<td>Quite similar to potential loyalists, except for outerwear and diapers. The orders including outerwear account for 8.31% and diapers account for 4.82%, which was about 3% more than that of potential loyalists respectively.</td>
</tr>
<tr>
<td>Registration channel</td>
<td>The channel for a customer registering as a member</td>
<td>19.4% of the loyalists came from stores, official websites or self-operated APP, while there was only 11.3% for potential loyalists.</td>
</tr>
</tbody>
</table>
Table 9: Particular Phenomenon of Each Cluster

<table>
<thead>
<tr>
<th>Consumption channel</th>
<th>Channel for purchasing items</th>
<th>Orders from Tmall accounted for 86.7% and orders from stores, APP and official web-sites account for about 12%</th>
</tr>
</thead>
<tbody>
<tr>
<td>The number of categories</td>
<td>The number of different categories of goods ever purchased</td>
<td>93.4% of the potential loyalist have purchased no more than five categories of products while more than half of loyalists have purchased over five categories</td>
</tr>
<tr>
<td>The number of babies</td>
<td>How many babies a family (member) have</td>
<td>The proportion of two-child households was as high as 45%</td>
</tr>
<tr>
<td>Baby’s age</td>
<td>Estimated current age range of the baby</td>
<td>The families with babies between 1 and 3 years old have become the major group among loyalists</td>
</tr>
<tr>
<td>The degree of development of the city</td>
<td>It was divided into first-tier to sixth-tier, and the first-tier cities are the most developed.</td>
<td>About 67% of loyalists come from second-or third-tier cities</td>
</tr>
<tr>
<td>Spending power</td>
<td>Estimated by the maximum order amount of the member</td>
<td>One-quarter of them have a maximum order amount of more than 1,000 yuan, proving that 25% of loyalists have certain spending power</td>
</tr>
</tbody>
</table>
7. DISCUSSION

7.1. The meaning of customer segmentation for Goodbaby

Followed the process of CRISP-DM, this thesis applied the customer segmentation of GB’s members and obtained seven different segments. In order to make the results more relevant to the company's reality and practical for real operations, the author of this thesis has utilized transaction records and demographic attributes for a more detailed analysis. Although some statistical processes were barely visible in the analysis, it required a long time to constantly explore and collide in the world of data. So here the question is, are these analyses and suggestions really meaningful to the company? According to the author’s personal experience gained after completing the internship at the company, the answer should be "yes".

For Goodbaby, with the help of online sales channels, it got a large number of members, but the company lacks a profound understanding about the consumption characteristics of these members and the difference between them and offline members. Obviously, blindly recruiting new members without knowing how to maintain the customer relationships fails to improve the repurchase rate and retain valuable members. This is evident in the analysis which has been illustrated and discussed above. Consequently, recur to customer segmentation, the company can clearly understand the composition of its membership, and thus develop corresponding operational goals and strategies.

After conducting segmentation and deep analysis based on the results, we know that Goodbaby's members can be divided into seven clusters. Most of them were from online platforms with only one purchase. The number of new members is growing fast, but the number of old members is shrinking faster. Online members are less loyal than offline members and are more likely to lose. Despite this, there are still a small number of loyal members who consumed frequently and valuable members who spent much. Another good news is that there are also many people who have the potential to develop into loyal or high-value members. After mastering their spending habits, the company has enough confidence to tap their needs and convert them.
7.2. Further roles of the research

7.2.1. For other companies

Goodbaby is not only a representative enterprise in China's maternal and child industry, but also a typical traditional enterprise in transition. In China, there are countless traditional companies looking for transformation to resist the impact of e-commerce. Creating a superior offline shopping experience and a comprehensive membership system to retain more high-quality customers is the common goal and development plan of these enterprises. Opening online sales channels was a choice that had to be made in response to market competition, but how to make targeted operational strategies for online and offline members is an unknown area for them.

This study gives these companies in transition a way to understand their members better and informs them which indicators can help them analyze the results. The detailed analysis can provide them with a reference.

Furthermore, the results of this thesis may give the researchers interested in the maternal and child industry a clearer understanding of the development status of China's major maternal and child groups and help them think more deeply in light of China's fertility policy.

7.2.2. For the past literature

As shown in the literature review, the clusters have been labeled based on customers’ consumption channels because members using different channels for purchase have shown distinct behavioral characteristics. However, for GB, the number of online and offline members varies considerably. Simply distinguishing between online members and offline members does not make much sense for the company, so, the author of this thesis labeled the clusters based on the current status and the value of the members. Despite this, in the process of analysis, some conclusions about the difference in loyalty between online members and offline members have been drawn, which can be discussed in more detail here.
In keeping with the definitions of “online” and “offline” mentioned in other literature, the channel of self-operated APP as well as official websites have been excluded. “Offline” means store-focused and “online” means e-commerce platform-favored. “Multichannel” means that the customer has ever purchased through two channels and “multichannel with preference” means that the customer has once used two channels for purchase but preferred to use one of them.

a. Store-focused customers show more loyalty than online-favored customers.

The consumption cycle of store-focused customers was shorter than online-favored customers. In other words, offline members purchased more frequently. On average, store-focused customers went shopping twice in 2.4 months while for online-favored customers, the average consumption cycle was 3.2 months. In terms of the spending amount, offline consumers spent an average of 665.7 yuan per order while online consumers spent only 186.6 yuan per order. Furthermore, online members have purchased an average of 2.67 categories of products throughout their life cycle in GB while offline members have purchased an average of 5.44 categories, which also shows more loyalty and understanding to the brand.

Although the online membership base was large, more than half of them were in a dormant or lost state, and (potential) valuable and (potential) loyal members only accounted for a low percentage with around 14%, while the percentage was 30% among offline members. In summary, store-focused customers showed more loyalty than online-favored customers.

b. Multichannel customers show more loyalty than online-favored customers.

Although it was hard to distinguish which group was more loyal according to their consumption cycle because it was quite similar between multichannel and online-favored customers, the average consumption amount of these two groups showed a huge gap. Compared with the bill of 186.6 yuan per order of online-favored customers, multichan-
nel customers with no preference paid an average of 488.1 yuan per order, and the multi-channel group with a preference for online paid 357.2 yuan while the people with a preference for store paid 463.3 yuan each time.

Furthermore, on average, multichannel customers usually bought 0.8 to 1.2 pieces of items more during each purchase than online-favored customers. And as mentioned earlier, online members have purchased an average of 2.67 categories of products throughout their life cycle in GB, while most multichannel customers have purchased more than 5 categories. More details are showed in Table 10.

<table>
<thead>
<tr>
<th>The consumption channel</th>
<th>The average number of pieces purchased per order</th>
<th>The average number of categories purchased</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store-focused</td>
<td>3.74</td>
<td>5.44</td>
</tr>
<tr>
<td>Online-favored</td>
<td>2.27</td>
<td>2.67</td>
</tr>
<tr>
<td>Multichannel with a preference for store</td>
<td>3.51</td>
<td>7.19</td>
</tr>
<tr>
<td>Multichannel with a preference for online</td>
<td>3.10</td>
<td>6.24</td>
</tr>
<tr>
<td>Multichannel with no preference</td>
<td>3.08</td>
<td>4.95</td>
</tr>
</tbody>
</table>

The greater number of pieces and categories of products included in the shopping cart implied that multichannel consumers had more trust and higher loyalty than online-favored members.

It is worth mentioning that connected with the result of customer segments, potential loyalists and loyalists accounted for over 60% of multichannel consumers with a preference whatever for store or online channels. 38% of the rest members consisted of potential valuable customers and top members. In other words, multichannel consumers with a preference were less likely to lose. Their willing to purchase was more stable and stayed active.

In this research, there is not enough evidence to prove that multichannel customers are more loyal than store-focused members. And it cannot be firmly proved that the difference in preference among multichannel customers means the difference in the value and
loyalty to the brand. Because on several key indicators, each group in its own way has made an important contribution.
8. LIMITATION AND FUTURE RESEARCH

As a representative brand in the maternal and child market in China, the research on the segmentation of GB members cannot only bring new marketing ideas to this industry, but also enable researchers in related fields to better understand the current situation and difficulties faced by the traditional maternal and child companies.

In the context of booming e-commerce, more and more entity economies are seeking transformation to comply with the digital transformation phenomenon. Optimizing offline shopping experience and developing a perfect membership system to seek more quality and loyal members are the common goals of these companies. As one of the traditional companies in transition, I believe that the problems and breakthroughs encountered by GB in the process of developing the membership system will also be confronted by other traditional companies. Therefore, the suggestions given to GB for different phenomena in this thesis may also provide a reference to traditional companies in other industries.

However, the limitation of the research is still obvious. First, in terms of the scope of application of the thesis, most marketing strategies were made based on the fundamental realities of China and the company's specific conditions, thus they may not have a good guiding role for foreign companies. Second, regarding the accuracy of the results, the huge difference in the number of online members and offline members may have a certain impact on the clustering results and bias the analysis. Because promising, hibernating and lost members accounted for a large proportion, there is probably more potential information hidden in these types of members. Combined with more other indicators, we can continue to perform secondary segmentation. Last but not the least, after the model was completed and the segments were obtained, the interpretation of the clusters and the corresponding operational strategies were generated based entirely on the author's understanding of the business and a small part of references, so a more experienced operator may be able to give more mature opinions.

Regarding the future development direction of this research, the first one is to apply “theory” into practice based on the analysis results to verify the practical significance of the
research, which is the process of “deployment”. Besides, GB also has a good reputation and sales performance in foreign countries, especially in the European market. The process of segmentation and analysis of the customers in other countries can let us understand the differences and characteristics of membership development in different countries, as well as the consumption habits of consumers in different areas.
REFERENCE


Data mining. (2000). *Nature Biotechnology, 18*(S10), IT35-IT36. doi: 10.1038/80073


APPENDICES

R codes:

# combine several datasets
data12 <- rbind(data1, data2)
data <- rbind(data12, data3)
rm(data1)
data2 <- data
data[, 2:4] <- lapply(data[, 2:4], as.numeric)
data_rfm <- data[, 2:4]
dim(data_rfm)
str(data_rfm)
summary(data_rfm)

# explore the correlation between three variables
data_cor <- cor(data_rfm)
library(car)
scatterplotMatrix(data_rfm)

library(corrplot)
corrplot(corr = data_cor, method = 'color', addCoef.col = "grey")

# make the histogram of each variable
hist(data_rfm$recency)
hist(data_rfm$frequency)
hist(data_rfm$monetary)
data_rfm2 <- data_rfm
data_rfm2 <- scale(data_rfm2)
summary(data_rfm2)

library(factoextra)

# decide the best number of clusters (failed, dataset is too big)
fviz_nbclust(data_rfm2, kmeans, method = "wss") + geom_vline(xintercept = 4, linetype = 2)

# another way to decide the best number of clusters (success)
wssplot <- function(data, nc=15, seed=1234) {
    wss <- (nrow(data)-1)*sum(apply(data,2,var))
    for (i in 2:nc) {
        set.seed(seed)
        wss[i] <- sum(kmeans(data, centers=i)$withinss)
    }
    plot(1:nc, wss, type="b", xlab="Number of Clusters", ylab="Within groups sum of squares")
}
wssplot(data_rfm2)
# Set random number seeds to ensure repeatable experimentation
set.seed(123)

# Run k-means
fit.km<-kmeans(data_rfm2,8,nstart=25)

# Check the result of the model
print(fit.km)
fit.km$size
fit.km$centers

# Extract the cluster tag and merge with raw data
aggregate(data_rfm, by=list(cluster=fit.km$cluster), mean)
data2 <- cbind(data2, cluster = fit.km$cluster)
data_cluster<-cbind(data$usr_num_id,cluster=fit.km$cluster)

# Check the number of members of each cluster
table(fit.km$cluster)