Project Delivery Two

Jessie Gong (tg2673) Yumeng Gu (yg2636) Yanghui Liu (yl4079) Jennifer Wang (yw3227)

1. Statement of the problem or question(s) being addressed

1.1 Management Dilemma

C-suites of most companies, if not all, care most about one question: how to increase the sales of the company? The company we choose for our project, a mail-order sales company in Germany, is not an exception. There are two ways to improve sales, which are expanding customers base and enlarging the basket of existing buyers. The former is considered as a better way since recruiting new customers not only increases the amounts of sales and meets the business target but also gets the company more endorsement in the total market which will help to accelerate the cycle of money for the companies. Meanwhile, even though enlarging buyer basket can push the sales in the short term as well, it may not be able to bring organic growth in the long term if the consumption of customers remains the same.

However, it is also noticeable that new customer acquisition could be pretty costly for a company, which forces the company to improve business efficiency. By identifying a target customer group, we may significantly save the money spending on acquiring new customers, and make a more competitive market campaign at the same time.

1.2 Management Question

Recently, this German company would like to launch a campaign in order to recruit new buyers. The management question is who is to recruit? This project aims to help identify segments of the population that form the target group of the company.

1.3 Research questions

1) What are the different segmentations in the population?
2) How the customer data for a mail-order sales company maps onto those population clusters.
3) Who are the overrepresented cohort, which is also the target group, in customers?
4) Is there any room for recruiting new customers?

2. Reasons behind the choice of analytical technique(s) and why the technique is suitable for the question(s) being addressed

As mentioned before, our project is to identify customer segmentation. For this objective, we choose two analytical techniques: PCA and clustering, specifically, we used k-means method to conduct analysis.

2.1 PCA

PCA is the abbreviation for principal components analysis, which is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The reasons for using PCA in our data processing is that:

1) Reduction of Dimensionality

   Since our raw data has over one hundred variables, the highly dimensional nature of the data makes it impossible to conduct direct clustering analysis. Using PCA for data dimension reduction is necessary for our data processing and analysis while keeping as much variation as possible.

2) Scaling the Original Variables

   The variable weight will significantly influence the analysis result. For distance-based clustering methods, the scale of the variable affects the weight assigned to the variable. Therefore, it is best to standardize variables.

After PCA, we transformed 110+ variables to 77 components, which were used for further analysis instead of the original set of variables.

2.2 Clustering

1) Concept

   Clustering is a classic approach for segmenting. It is the process of making a group of abstract objects into classes of similar objects since it uses mathematical models to discover groups of similar customers based on the smallest variations among customers within each group. Compared with other analytical approaches, using clustering in customer

segmentation has many obvious advantages. First, clustering has high practicality. Practically, it is impossible to use predetermined rules to accurately segment customers over many dimensions. Additionally, high homogeneity. In clustering, variances within each resulting group are very small whereas other methods typically group customers who are actually very different from one another. What's more, clustering can be very dynamic. The clustering definitions change every time the clustering algorithm runs, which ensures that the groups always accurately reflect the current state of the data.

2) K-means

During the process of clustering, we used k-means which have unique advantages as well. K-means clustering arbitrarily places centroids in the data and then iterates from that point to the final solution. The most important advantage of k-means is the relatively low computational power, when k is small, it is computationally faster than other methods with a large number of variables. Besides, k-means is easier to implement, it can produce tighter clusters, and an instance can change cluster when the centroids are recomputed.

3) Business Application

When it comes to business, using clustering methods to analyze the data can help the companies- which mainly focus on customers- gain further insights from the data and make better market campaign by locating target customers more accurately.

After conducting clustering analysis, we may compare our customer clusters with population clusters to accurately acquire customers, which means to get the specific people who have similar features with our customers.

Above all, we can see that in the process of dividing customer types into clusters based on different features, which could further help us identify target customers and deliver the market campaign to improve sales.

3. Result

3.1 Preparing Process for Getting the Result

We cluster the general population into a total of 16 clusters, each group shares similar demographic properties. The method of choosing cluster number is to see if there is an elbow appears in the ‘within sum of square’ plot.

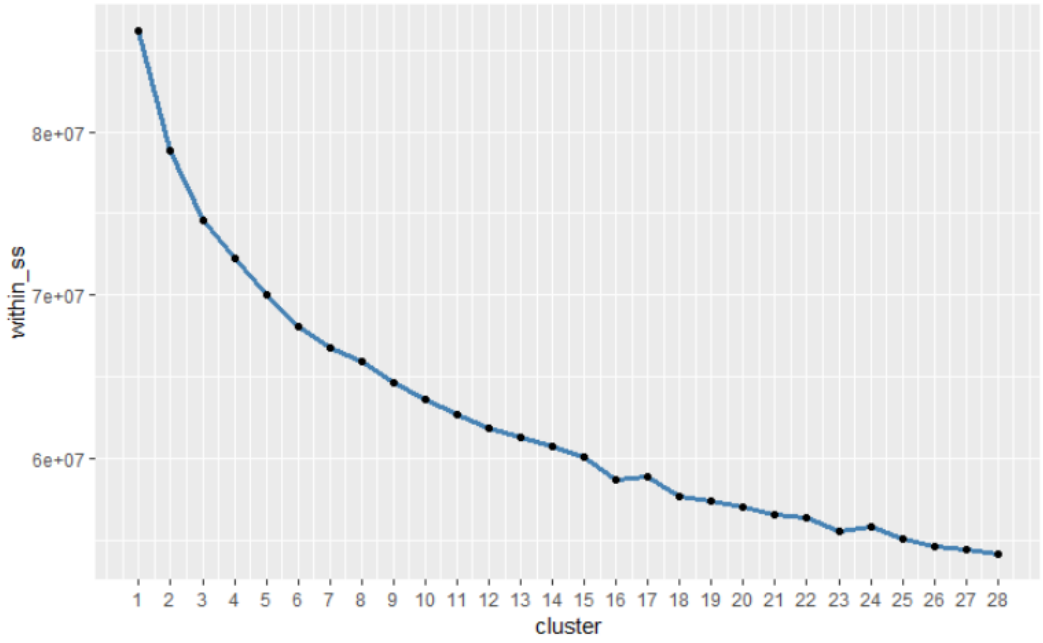

Fig. 3.1 K-means Clustering

From Fig 3.1, we can see that there is a clear elbow in 16 clusters and a small increase appears in 17 clusters. Thus, we decide to keep the cluster number as 16.

3.2 Get Target Customers by Clustering

After applying the cluster data to the general population and customer data, we then got the cluster composition (people%) for these two groups.

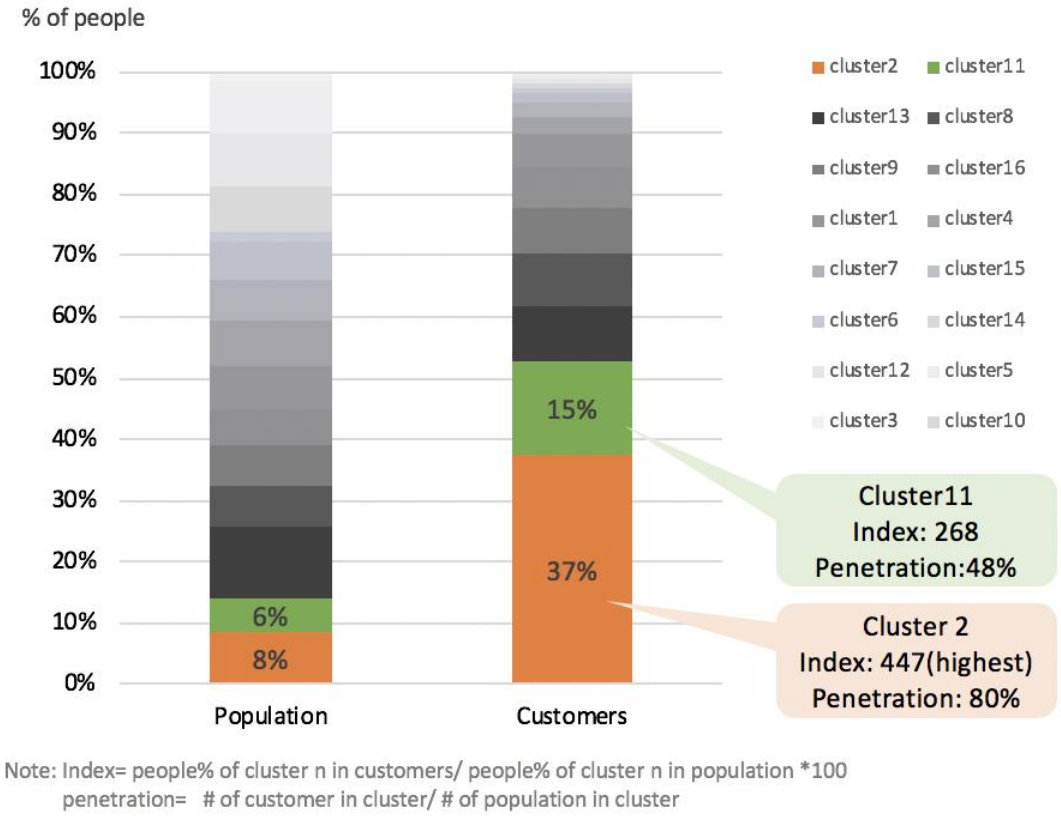

Fig 3.2 Cluster Composition Comparison

From Fig 3.2, we can identify 2 clusters in population dataset who overrepresent in the customer dataset.

- Cluster 2 is overrepresented the most in the customer data since it enjoys the highest index (447) against the population among all clusters. Meanwhile, cluster 2 also enjoys the largest customer base, occupying 37% of the total customers of the company. However, the penetration of cluster 2 is as high as 80%, which indicates that the majority of this cluster in the population has already been our customers, and customer acquisition may soon reach its ceiling.
- Cluster 11 also enjoys index higher than 200, occupying 15% of the total customer base. The penetration of this cluster is just 48%, indicating a large room to improve.

Thus, cluster 2 and cluster 11 are the target groups of this company, and both of them should be discussed when finding the features of our target audience of the campaign to be launched.

3.3 Features of Target Groups (Comparing with Population Average)

After clustering, we can get features which are significant from each component. However, considering the business reality, not all these features can be directly used to get business insights. Thus, features are chosen for further investigated based on their importance and feasibility to navigate and conduct market strategies.

By looking at the previous analysis, the two most important features from PCA analysis are income and degree of Minimalist, which will be examined first in the feature comparison part. Some other features such as gender and age will also be considered.

1) Income

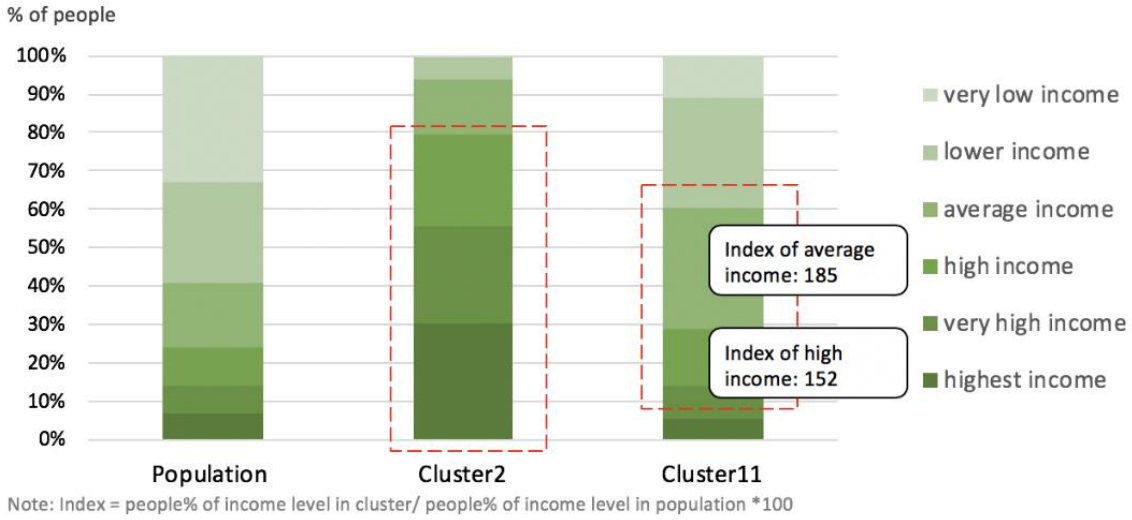

Fig 3.3 Feature Analysis - Income

As shown in Fig 3.3, the company targets on people with middle and high income.

- Approximately 80% of cluster 2 falls into the category of high-income people; while in population only 24% are classified as high income.
- Cluster 11 is not as much skewed to high income as cluster 2, but it has a higher proportion of middle and high income than population.

2) Degree of Minimalist

Fig 3.4 Feature Analysis - Degree of Minimalist

As shown in Fig 3.4, the company targets on people who are not a minimalist.

- Most people of both cluster 2(73%) and cluster 11(65%) shows little similarity in this type comparing to the population average. Minimalist are those who are not willing to spend money on their daily life; while people we are targeting won't save too much money, instead, they have plenty of cash in hand and are willing to purchase. This also aligns with the income data, since people with higher income normally have higher personal expenditure.

3) Gender

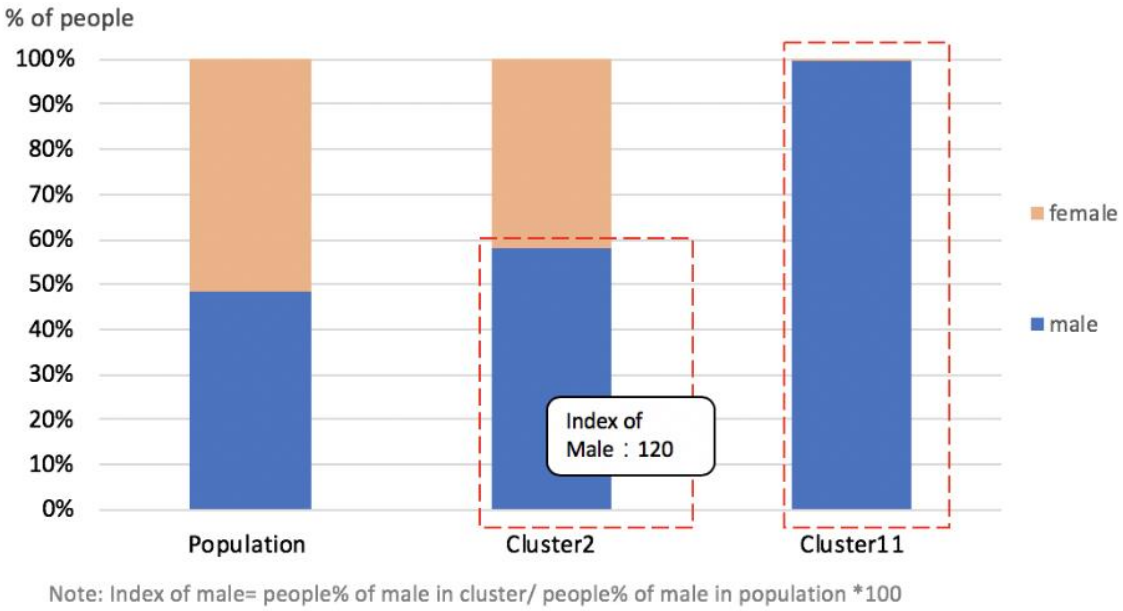

Fig 3.5 Feature Analysis - Gender

As shown in Fig 3.5, the company mainly targeted on male.

- Cluster 2 slightly skew to male when compared with the population average, with an index of 120.
- Nearly all of cluster 11 are male. This might infer that males are more likely to purchase through the mail than females since males are more goal-driven and not willing to waste time and energy on wandering in a physical store.

4) Age

Fig 3.6 Feature Analysis - Age

As shown in Fig 3.6, the company mainly targets on elder customers.

- Almost all target groups (both cluster 2 and cluster 11) consist of people aging from 46 and plus. People over 60 years old are highly skewed in target group, especially in cluster 11.

5) Other character features

From the analysis, we also find our target groups tend to be rational minded, material minded, dominant minded, not dreamily, and less social minded in both cluster 2 and cluster 11. They are more critical and have fightful attitudes as well. These characteristics align with the gender distribution-- since there are more male in target groups comparing with population average. Meanwhile, the target groups are also more religious and traditional minded. This part can further be used for designing the forms and contents of campaigns.

4. Conclusion and recommendation

4.1 Conclusion

The target group of the company is cluster 2 and cluster 11, which share similar features, with just slight differences in income and gender. In all, the mail-order company prefer customers

who have higher income and willing to spend money on daily spending. Meanwhile, demographic wise, the company targets relatively more on male customers and elder customers. There exist some differences between cluster 2 and cluster 11, though.

- Cluster 2 is more skewed to extremely high income; while cluster 11 more skewed to middle and moderately high income.
- Cluster 2 is slightly skewed to male; while 99% of cluster 11 are male.

4.2 Recommendations

Recommendation 1:

Since the company targets customers who have higher income and prefer to spend money in improving life quality, the market campaign should be conducted on higher educated people with honorable jobs and stress the concept of 'high life quality' when conveying the message to the audience.

Recommendation 2:

Since the company targets more on senior customers, the campaign would better be conducted in a traditional way, such as offline activities or TV ad, to ensure the messages will be successfully delivered to them. Meanwhile, characters in the ads can be set as elder males for more emotional attachment for potential customers.

Recommendation 3:

Since the company targets more on male, the company should cater to this group of people. For example, the content of the ad should be neat and focus more on functionality.

4.3 Future Study

For future study, it would be better to combine the customer segmentation with their sales data. Leveraging on the sales data, we will be able to further investigate the purchase habit of different clusters (frequency, basket size, channel, preference on product specification, or sensitivity to price).