# Visual Analytics Techniques for Customer Segmentation in an Online Retail Context

## L. Varsandan

### 1.Introduction

#### 1.1 Motivation

Customer Segmentation (CS) is a technique widely used in a range of industries as it can have multiple benefits, including informing marketing campaigns (1) and product offering strategies (2) . Additionally, if purchasing behaviours are used to segment customers in an online retail context, it can be used in collaborative filtering (3) to increase the chance of recommending new relevant products to existing customers. Since clustering algorithms are unsupervised and therefore more subjective, Naija & Sinaoui (2012) suggested using a combination of numerical validation metrics and interpretability of clusters to assess the appropriateness of the segmentation. Interpretability of clusters can also be used for getting buy-in from the key stakeholders and for business users to design meaningful strategies. Consequently, this paper will illustrate how visual analytics can play a significant role in guiding a CS analysis by suggesting which dimensions would be relevant for clustering and for validating and interpreting the clusters.

#### 1.2 Data

The data used in this paper comes from the company Instacart and is hosted by the website Kaggle (4). Instacart is an online grocery delivery service in the US (5). The dataset provides shopping details of approximatively 200,000 customers including when and how frequently they order, as well as what products they bought in each order. It also includes the product names, the aisles and the departments each product belongs to. This multivariate and text data is suitable to facilitate two types of behavioural customer segmentation (6). The first type is how frequent and when customers place an order and the second one is what types of products they buy. The data is firstly joined together using Python and then further transformed and analysed using R, both being popular open-source data science tools.

#### 1.3 Research Questions

The research questions we are going to answer with this data are:

* Is a clustering based on a text mining approach on product names going to provide more informative customer segments than a clustering based on purely continuous attributes?
* What are the main customer segments in terms of order placing behaviours (frequency and period of time) and their implication for marketing strategy?
* What are the main customer segments in terms of types of products bought and their implications for a recommender system and for a product development strategy?

### 2. Tasks and Approach

#### 2.1 Data Types

The main data types available for this analysis include multivariate continuous data, expressed as the percentage of orders for a particular user under a certain condition and text data, expressed as the names of the distinct products purchased across all orders for a particular user. Therefore most of the computational and visual techniques used are only relevant for this type of data. Examples of the order placing behaviour multivariate continuous data, product categories continuous multivariate data and product behaviour text data are given in Figure 1.

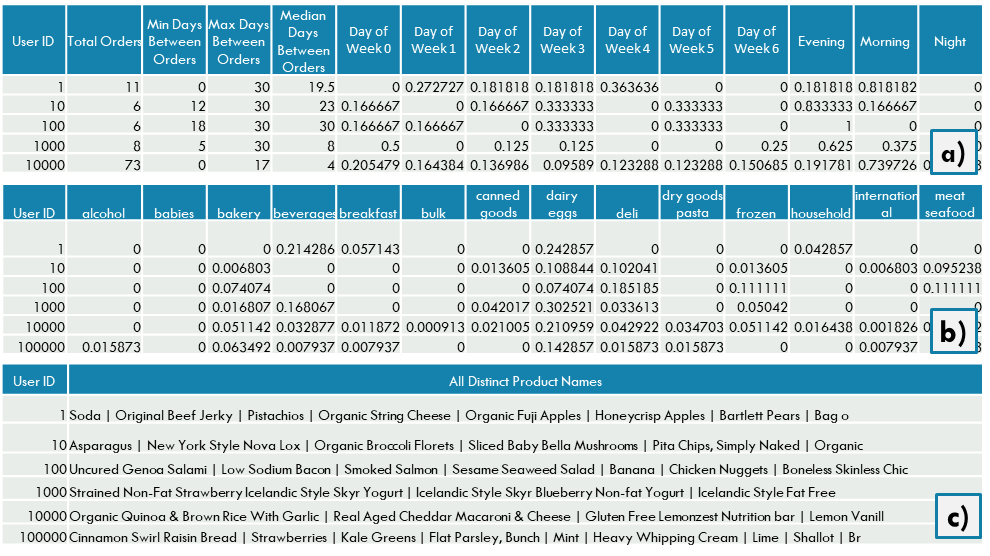


Figure 1 Data examples a) order placing behaviour multivariate continuous data, b) Departments Purchasing Habits Multivariate continuous data and c) Product Names Text Data

#### 2.2 Analysis Structure

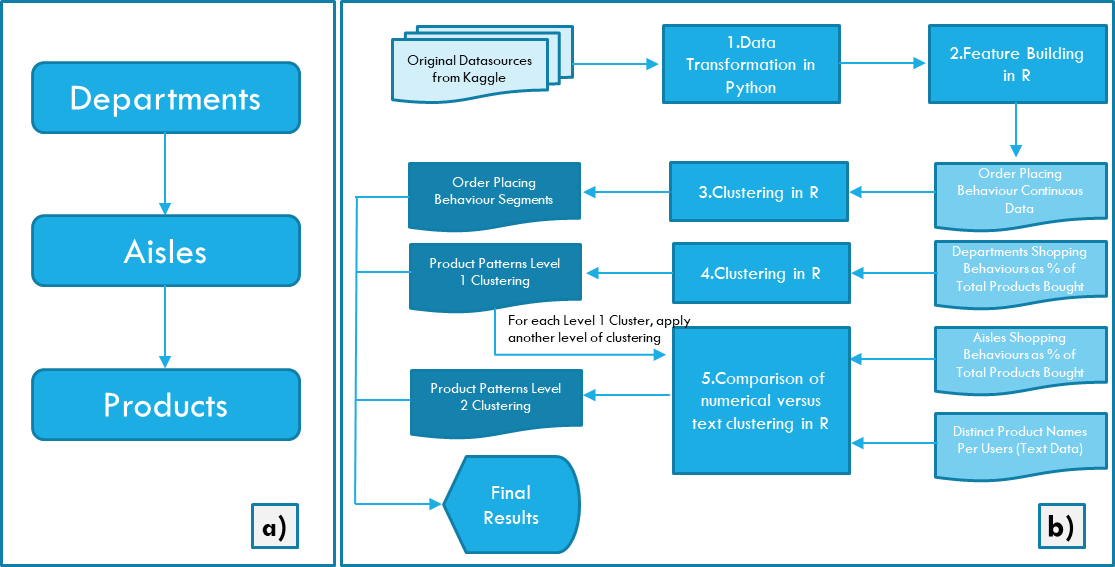


Figure 2 a) Product Category Hierarchy b) Analysis steps including key inputs and outputs after each step

This analysis consists of five main parts (Figure 2b). The first two are data manipulation steps of transforming the raw data in Python and building extra features in R. These two steps do not contain any visual analytics and therefore will not be further elaborated. The three types of clustering exercises in R include Order Placing Behaviour Patterns segmentation using continuous data only, Product Purchasing Patterns segmentation Level 1 at Department Level (the top level of product hierarchy – Figure 2a) using continuous data only and Product Purchasing segmentation Level 2 – this compares an approach using continuous data at aisles level (the second level of the product hierarchy – Figure 2a) with an approach using product text data, with the final decision being to use text data clustering in the final results.

The analytical tasks described during the rest of this section address steps taken across all three main clustering exercises and Section X will include examples across all three types of clustering for each analytical step.

#### 2.3 Exploratory Data Analysis and Feature Selection

Exploratory Data Analysis was mainly conducted on the continuous data. The text dataset was explored at the topic modelling stage (Section X).I explored which variables are of interest for the different segmentations using summary statistics like mean and variance as computational elements (7) This helped in removing the variables with low means and low variance which could have added noise and would have reduced interpretability as a dimension reduction algorithm would have been used instead to reduce to meaningful variables. These were aided by scatterplot pairs to show the relationships between two different continuous variables (8), histograms of continuous variables to show their distribution (9) and parallel coordinate plots (PCPs) to show the relationships between more than two continuous variables (8). All these have informed the decision of which features to include in the clustering algorithms.

#### 2.4 Clustering Continuous Data

The main computational methods used for segmenting the continuous data were partitioning around medoids(PAM) and hierarchical clustering (HC). I have chosen partition around medoids as it can it is less volatile to outliers (10) which in this case it is important as we try to achieve generalizable customer segments. I have used a HC as a second algorithm as it uses a dendrogram to suggest the number of clusters (11) and therefore it serves as a comparison to PAM for validating the clusters. The first decision informed by computational and visual analytics were how many clusters to use as parameter in the algorithm. For the PAM algorithm, I have used a graph of the silhouette width to determine the number of natural clusters (12)by picking the cluster number which would maximize the width. For the HC, I have used a dendrogram to pick the number of clusters by selecting the number of clusters where the height between two splits was the largest (11).

For evaluating the clusters visually, I have used a Sankey chart to see if the grouping was similar under both clustering methods (13) For the users which were grouped inconsistently, I have compared their features with the features which characterised their clusters under PAM and HC. I tried to visually determine which group they were closer to, using PCPs, in order to choose the overall clustering model. Additionally, I have interpreted the meaning of clusters by plotting the values of their centroids on a radar chart, which is often used to compare multivariate continuous data (14).

#### 2.5 Comparison between Clustering on Text Data versus Continuous Aisles Data

Once Level 1 clusters were derived from the shopping continuous data on users and departments, I proceeded to add another level of segmentation for each cluster. First, I computed the clustering on the continuous data from aisles shopping behaviour, following similar steps as described in Section 2.4.

The clustering of customers based on the product names in their baskets, consisted of finding topics for each Level 1 Clusters and then applying PAM on how many words fell under each topic for each user. In order to derive the basket topics, I have used the Latent Dirichlet Allocation (LDA) algorithm, which is typically used in topic modelling (15).I then used the LDAvis package in R (16) in order to interactively visualise the results of the LDA and choose which was a suitable number of topics for each cluster. This was done by spatializing the topics using Multidimensional Scaling (16) and by visualizing which terms were relevant for each topic (16).Once the number of topic were chosen, the topics for each user were incorporated as features in a PAM algorithm for the same reasons, using the silhouette width to choose the number of topics. Since the silhouette width was a lot closer to 1 than in the case of Level 1 Clustering, I no longer wasted resources to validate through other clustering methods. Finally, in order to interpret the clusters based on the basket topics, I computed the term frequency(tf) and the term frequency – inverse document frequency(tf-idf) for each cluster and word. The term frequency illustrated what are the predominant words in each cluster, which can overlap between clusters. They are important in terms of the overall strategy of the supermarket. However, the tf-idf indicates better what are the words are more important in the context of each cluster (17). The tf-idf can inform more on the characteristics of each cluster.

Once the topic modelling was done, I compared the two types of segmentation – the one based on continuous data from aisles shopping behaviour and the one based on text – by interpreting the meaning of the clusters using a radar chart for the continuous clustering and a table heat map of tf and idf for the top terms for each cluster. I also used a Sankey chart to see if the users are grouped similarly under the two methods.

### 3. Analytical Steps

#### 3.1 Exploratory Data Analysis and Feature Selection

Figure 3 shows examples of the visual analytics used during data exploration and feature selection. Figure 3a and 3b we used in the clustering for order placing behaviour. This involved checking whether there is a meaningful difference in the amount of orders placed by users and how frequently they placed those orders. From Figure 3b we can draw the conclusion, as suspected that the more orders customers place, the more frequently they place them i.e the median number of days between orders is lower. A limitation is that we do not know the time context under which the data was provided to the Kaggle website, so we cannot be certain that users placing only two orders have not placed more orders outside the data made available. This graph shows that there could be 3 potential groups of users: a highly loyal group of customers that have placed a high number of orders frequently, a group of customers who have place few orders but quite frequently and a group of customers who have placed few orders and less frequently. The next component of order placing is when the order is made. The feature available are the time of day and the day of week. In order to explore the relationships between these, I have used PCPs as illustrated in Figure 3d, 3e, 3f. Whilst the day of week added noise, but the time of day showed some potentially meaningful patterns. For example, there are some customers who lean towards either evening or morning, or prefer both of them equally, whereas there are only very few customers who order during the evening and night. Following these explorations, I have the total orders, median days between orders, and time of day features in the clustering algorithm.

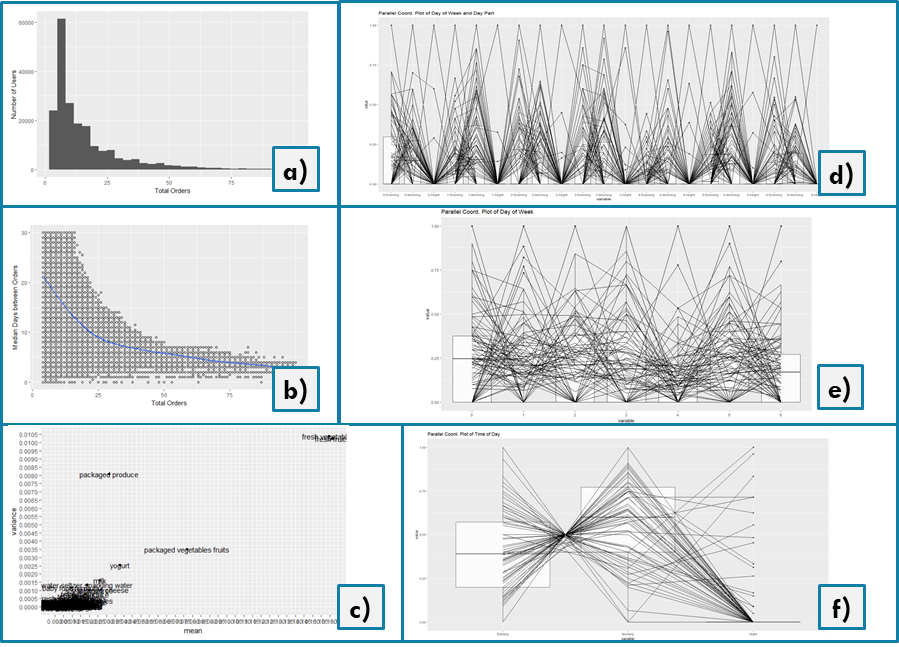


Figure 3 a) Histogram of total number of orders used for order placing behaviour b) Scatterplot of total orders and the median number of days between orders to show potentially loyal groups c) Scatterplot of mean and variance for Level 2 Segmentation of Cluster 2 of Product Patterns Segmentation on Department d) PCP of time of day and day of week features combined e) PCP of day of week features f) PCP of time of day feature

Figure 3c shows each aisle variable and its mean and variance plotted as a scatterplot. I have only kept the ones that had variability more than 0.0005 and a mean higher than 0.010, reducing the number of aisles used in the clustering algorithm which initially were 134. I used this instead of more advanced dimensionality reduction algorithms in order to reduce noise (7) and maintain interpretability of clusters. As highlighted in Section 1.1 maintaining interpretability was key to assess the meaningfulness of a cluster.

#### 3.2 Clustering Continuous Data

Figure 4a illustrate the silhouette width graph used to choose the number of clusters for the order placing behaviour segmentation. Since 1 or 4 clusters had the highest silhouette width, I have chosen 4 as the final number for the parameter k in the PAM algorithm, which has led to the 4 main clusters of order placing behaviour, illustrated in the final results. Figure 4b shows the dendrogram for the Level 1 product pattern clustering at Department level, which again, has the largest height between 4 clusters. These clusters were then compared with the PAM clusters in the Sankey Chart in Figure 4c. The Clusters on the left were the PAM clusters and the clusters on the right were the HC clusters. It can be seen that the clustering was largely consistent between the two methods. The points that were inconsistently grouped, for example from Cluster 2a to Cluster 3b, were explored in a PCP and compared with the centroid points of the two Clusters in Figure 4d. Their patterns did fall somewhere between the two clusters, for example on the ‘pets’ department, they were closer to the PAM Cluster 2a, but on ‘frozen’ more points were closed to the HC Cluster 3b. As a result, I decided to manually group the inconsistently clustered points in a fifth cluster that would be called ‘Other’ and allow the Level 2 of the segmentation to bring more clarifications on these segments.

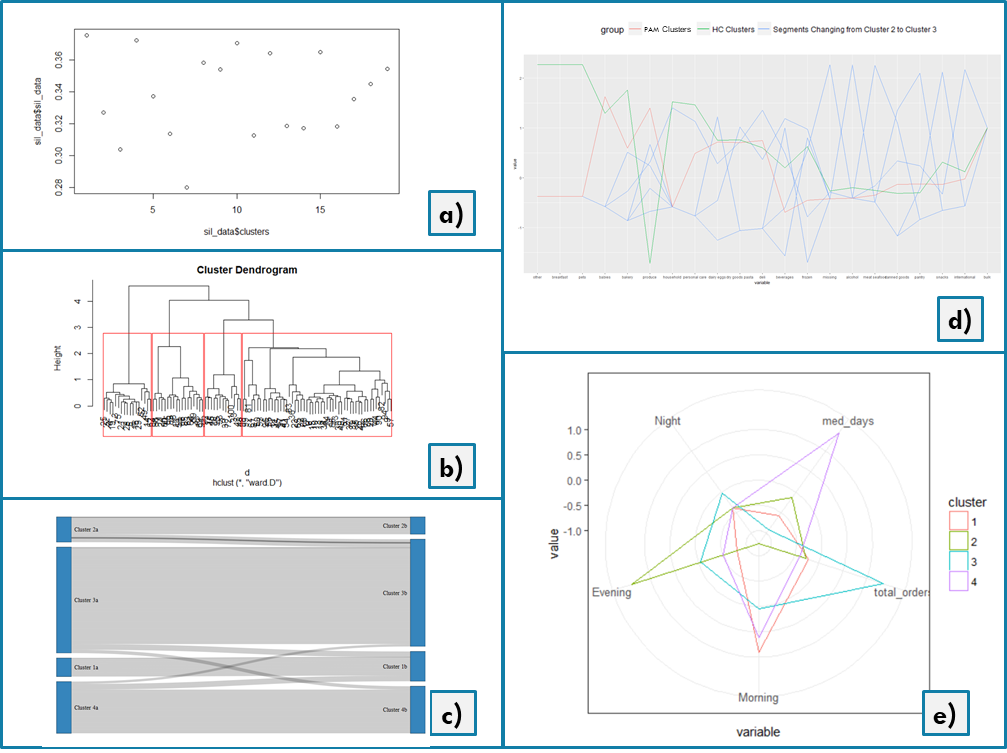


Figure 4 a) Silhouette width graph used in PAM clustering of order placing behaviour b) Dendrogram used in HC of Departments Product Behaviours (Level 1 Segmentation) c) Sankey chart comparing PAM clusters with HC clusters for the Department Level 1 Segmentation d) PCP showing the segmented points versus the PAM and HC clusters characteristics e) Radar chart illustrating the traits of the final segments of order placing behaviour.

Finally, plots like the radar chart shown in Figure 4e helped with the interpretation and meaning of each cluster. For example, in terms of the order placing behaviour, the four clusters could be described in the following way:

* Cluster 1: Potentially highly loyal customers since the days between orders is the second lowest out of all clusters and the second highest in terms of total orders, with a high preference for ordering in the morning
* Cluster 2: Medium loyalty customers, with a high preference of ordering in the evening
* Cluster 3: The most loyal customers, without any strong preference for any time of day
* Cluster 4: The least loyal customers, with a high preference of ordering in the morning

#### 3.3 Comparison between Clustering on Text Data versus Continuous Aisles Data

Figure 5 illustrates the output from the LDAvis package for topic modelling done on users grouped in Cluster 2 of the Level 2 Clustering. Having tried the model with the parameter k of 3 (Figure 5a) and 4(Figure 5b) topics, it can be seen that the distance between 3 topics is the clearest in terms of the PCA. However, a closer look at the terms for topic 2 under the 4 clusters LDA shows finer terms such as ‘gluten’ or ‘coconut’ which are specific for topic 2, whereas topic 3 does not contain those in the top terms, even though it is close to topic two in terms of the principal components. Since gluten free products and coconut milk could be traits of a group of users that have dietary requirements, I have decided to keep the 4 topics LDA model instead of the 3 and apply PCA on it.

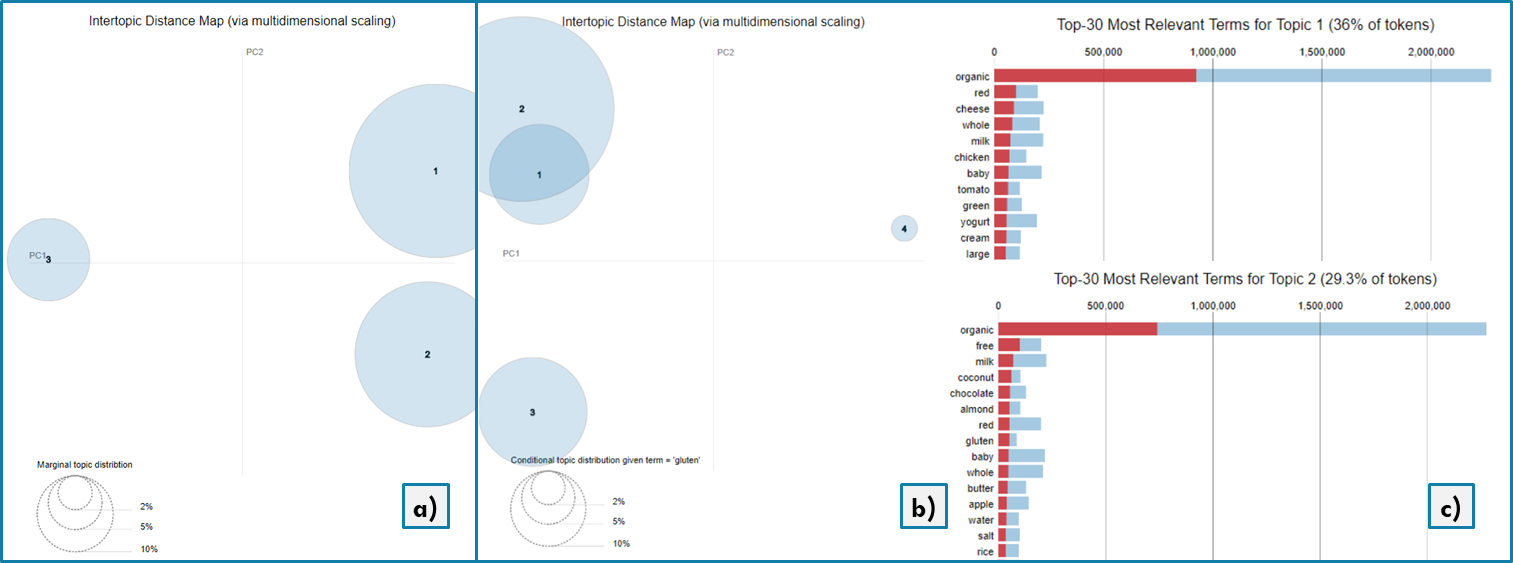


Figure 5 a) Intertopic Distance Map for 3 topics LDA on Cluster 2 of the Level 1 Product Behaviour Segmentation b) Intertopic Distance Map for 4 topics LDA on Cluster 2 of the Level 1 Product Behaviour Segmentation , showing the conditional topic distribution for the term ‘gluten’ c) The top 30 most relevant terms for Topic 2 and Topic 3 of the 4 Topic LDA on Cluster 2 of the Level 1 Product Behaviour Segmentation

Once the topics for Cluster 2 have been defined, I applied a PAM model to group the users depending on what topics they had in their basket. Next, Figure 6a) illustrates the radar chart with the clusters from the continuous data. There are some clusters that have some interesting traits, such as a high percentage of products in ‘fresh fruits’ or in ‘yogurt’ and ‘baby food formula’. However, compared to the topic modelling clustering seen in Figure 6c and Figure 6d, it contains less detail. For example, aspects such as the fact that the majority of the customers prefer ‘Organic’ products is not obvious in the clustering done on continuous data. Moreover, if we look at the tf-idf metric, we can see that the Cluster 6 has a strong preference for the 2nd Foods and Graduates baby food, as well as trimmed and husked vegetables and 160ct fabric softener. In order to understand the meaning of the terms, I have searched for them in the list of products. The text data provides a lot more information about the size of the products, the brands and the types of products in a particular aisle, rather than just knowing that they shop a lot in a particular aisle.

Additionally, Figure 6b illustrates that even the groupings are different when it comes to the topic modelling clustering. It has produced 7 more even customers rather than just 6 cluster under continuous aisle data. This suggest that customers might be more alike in terms of the types of products bought, even though they might shop more in a particular aisle.

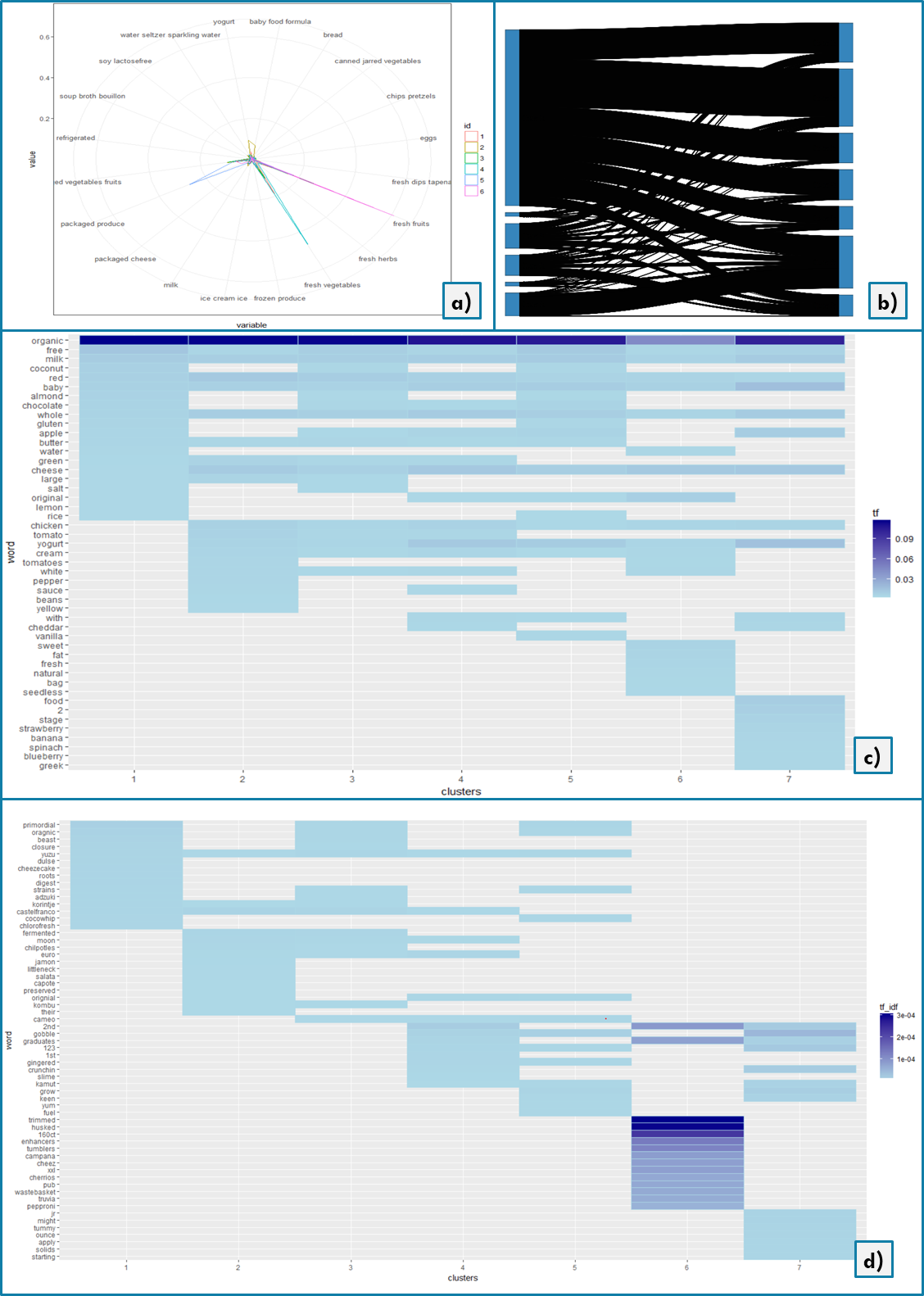


Figure 6 a) Radar Chart for Interpreting Cluster 2 Level 2 Segments using continuous aisle data b) Comparing continuous aisle data clusters (left) with topic modelling clusters on the right c) Cluster 2 Level 2 Segmentation on topic modelling, showing terms by term frequency d) Cluster 2 Level 2 Segmentation on topic modelling, showing terms by term frequency – inverse document frequency

### 4. Findings

#### 4.1 Text clustering as a more informative segmentation tool

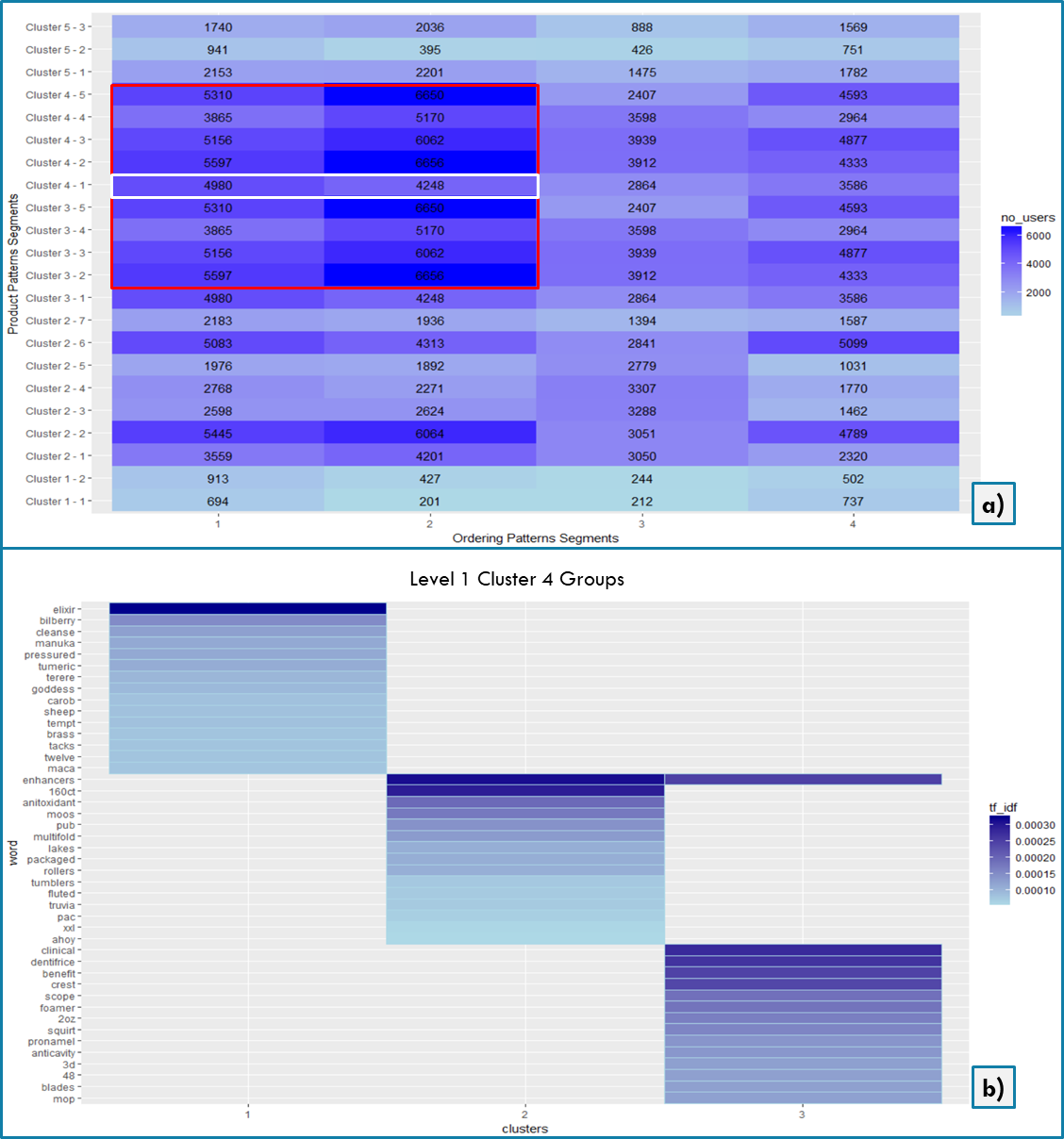
As highlighted in section 3.3, one of the main findings to answer our first research question is that customer segmentation done on text mining provides a richer interpretation (Figure 6c and d) than traditional clustering on continuous multivariate data based on what percentage of products each aisle occupies in from the total products bought (Figure 6a). The types of actions that can be taken based on text segmentations are more specific and could therefore be more efficient. For example, a practical implication of the example given in section 3.3 is that customers generally seem to prefer organic products. This means that the supermarket could increase the variety of organic products at an overall level of they want to attract repeated purchases from those customers. Additionally, they could send offers on the brands that came top for particular groups, such as 2nd Foods for Cluster 6 in Figure 6d. As a result, I have decided to use the text clustering as the final Level 2 segmentation.

#### 4.2 Order Placing Segmentation and Product Behaviour Segmentation.

The final results of the order placing segmentation have been already touched upon in Section 3.2 (Figure 4e). Based on those groups, the online retailer might decide that in order to encourage repeated purchases from more of the potentially highly loyal customers, they could lower delivery prices during the morning or they could send promotional emails to them during the morning.

The results for product behaviour in terms of text mining have also been partially discussed in the previous sections, with examples being given of actions which can be taken from the more detailed interpretation of clusters.

More value could perhaps be added by overlapping the order placing segmentation with the product behaviour segmentation (Level 1 based on departments preference and Level 2 based on text mining). Figure 7a shows how this overlap looks. Since the segments cannot be easily described in a few words, I have chosen to leave the cluster numbers as labels, with further detail being available in the Appendix. A key finding from this overlap is that Clusters 3 and 4 from Level 1 seem to contain the highest number of customers in the highly loyal groups 1 and 2 from order placing clustering. So the retailer might therefore seek ways to increase the number of customers in this group or to improve their customer experience to retain them. For example, Cluster 4-1 is one of the lowest in that group. Figure 7b shows that this group of customers are very health focused, with products such as Manuka Honey and Turmeric as one of the favourite products. Therefore, the retailer could perhaps increase the range of products with healthy benefits and then let them those customers know that they have introduced new products. A deeper dive into the relevant terms for each cluster could uncover more of such actions.



### 5. Critical Reflection

#### 5.1 Implications of the findings for the online retailer Instacart

The customers segments in terms of order placing behaviour enables Instacart to easily identify which groups are likely to be more loyal and therefore enables them to focus their marketing efforts on increasing and retaining those groups. Additionally, the text segmentation provides a detailed description of each cluster which could be used in product range decisions, new product recommendations to customers through collaborative filtering and informing the overall strategy of the company, such as perhaps focusing on organic healthy products. This analysis could be refined by incorporating additional information such as profitability of each customer, the average price range of each order etc. This would add another layer of focus on the highly profitable segments or segments that generate most revenue.

#### 5.2 The quality of methodology used in answering the proposed research questions

The visual analytics used in this analysis were built in R, which is an open source tool. Perhaps more advanced computational and visual methods could be used in other software tools. The results generally answer the research questions well, with clear implications, however, more work could be done to increase confidence in the results. For example, the silhouette width for the order placing behaviour is only 0.37 (Figure 4a). This means that the clusters are not clearly separable and therefore by segmenting customers in clear groups we might be artificially imposing some limits on there. Perhaps other variables such as profitability could have increased the quality of clusters. Another limitation of the data is that we do not know the context in which the data was collected. For example, Cluster 1 of Order Placing Segmentation might have fewer orders than Cluster 2 because they have churned quite quickly. In that case, it might not be in Instacart’s interest to continue growing that segment. Additionally, clustering is a subjective exercise and perhaps there might be other valid reasons why a clustering approach on continuous data would have been preferred instead of the text mining clustering.

#### 5.3 Applicability to other domains

This exercise could be applied to other retail industries where the customers are likely to do repeat purchases. However, for the industries such as clothing where customers might not buy the same thing twice and might be quite hard to predict what they would need next, and therefore might be difficult to cluster according to product names. Additionally, the computational and visual analytics methods used in this analysis address numerical continuous data and text data. In the cases where categorical data exists, this analysis would have to be modified and might held different results.

### 6. Conclusion

This paper has provided examples of how visual analytics can guide decisions in a customer segmentation exercise for an online grocery retailer. The segmentation has been done at 3 levels: order placing behaviours such as order frequency and time of ordering, Level 1 segmentation of product preference based on which depaertments each customer is likely to focus their shopping on and Level 2 segmentation based on text mining of the product names they have purchased. The text mining approach was deemed more suitable than a clustering based on numerical data as it provided more detailed insights into the customer characteristics. By overlapping these segments, Instacart can find what are the core customer segments they should focus on, what should be their competing strategy, when they should communicate to customers, what new products they could recommend to existing customers and what new products they should add to their range.