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

## L. Varsandan

### 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 (Gil-Saura & Ruiz-Molina, 2009) and product offering strategies (Noori, 2015) . Additionally, if purchasing behaviours are used to segment customers in an online retail context, it can be used in collaborative filtering (Liu, Lai, & Lee, 2009) 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 (Kaggle, 2017). Instacart is an online grocery delivery service in the US (Instacart, 2017). 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 (Kotler & Keller, 2006). 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:

* What are the main customer segments in terms of order placing behaviours (frequency and period of time) and their implication for marketing strategy?
* 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 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.

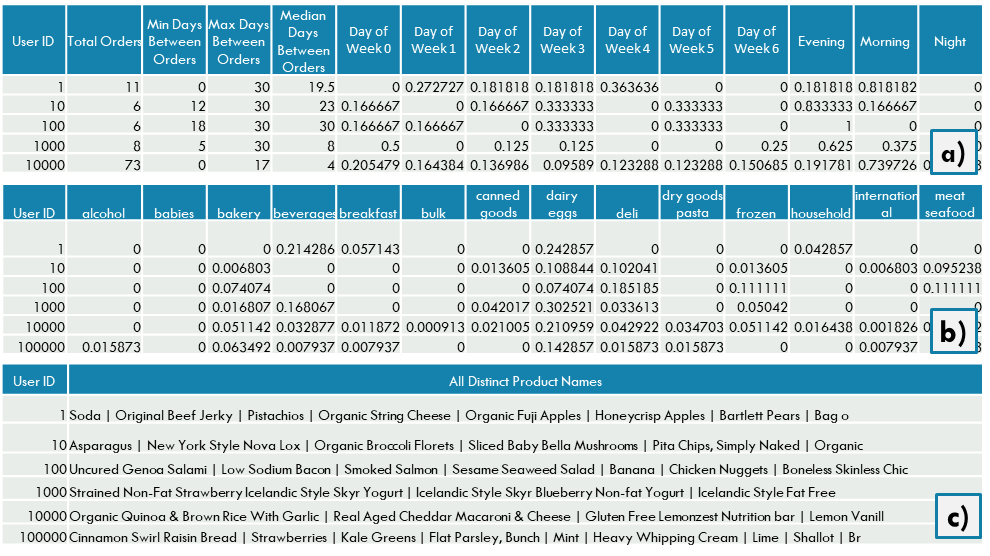


Figure 1 Data examples a) order placing behaviour multivariate continuous data, b) Departments Purchasing Habits Multivariate continuous data and c) Product Names Text Data

#### 2.2 Analysis Structure

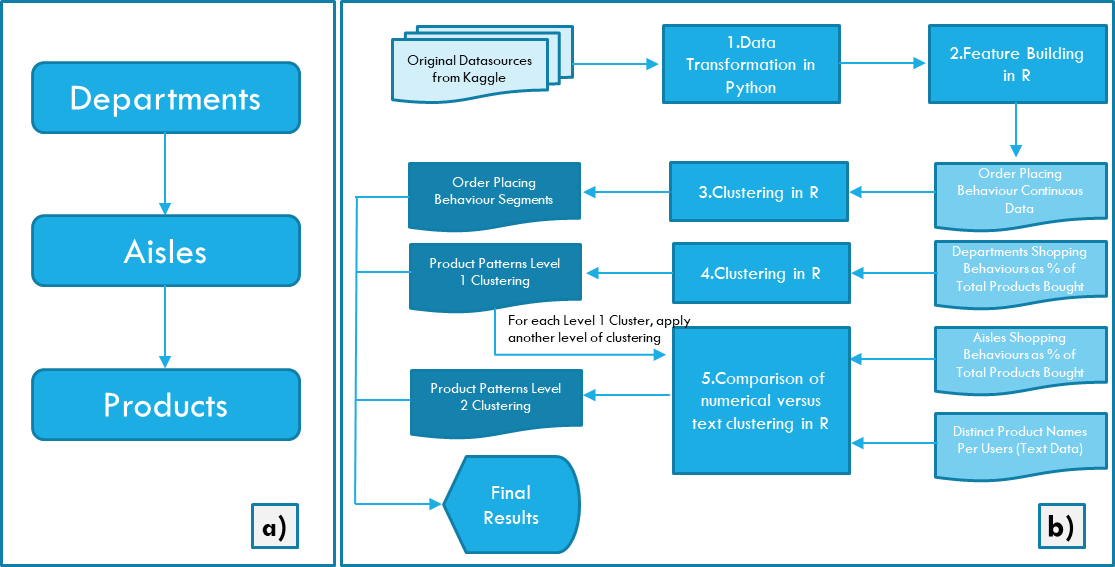


Figure 2 a) Product Category Hierarchy b) Analysis steps including key inputs and outputs after each step

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 (Yang, Zhang, Zou, Hu, & Qiu, 2013) 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 (Im, McGuffin, & Leung, 2013), histograms of continuous variables to show their distribution (Huzurbazar, 2005) and parallel coordinate plots (PCPs) to show the relationships between more than two continuous variables (Im, McGuffin, & Leung, 2013). 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. I have chosen partition around medoids as it can it is less volatile to outliers[ref] which in this case it is important as we try to achieve generalizable customer segments. I have used a hierarchical clustering as a second algorithm as it uses a dendrogram to suggest the number of clusters[ref] and therefore it serves as a validation that the clusters are robust [ref]. 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 silhouette width [ref] to determine the number of clusters that would maximize the width. For the hierarchical clustering, 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 [ref].

For evaluating the clusters visually, I have used a Sankey chart to see if the groups were similar under both clustering methods [ref]. For the users which were grouped inconsistently, I have compared their features with the features which characterised their clusters under PAM and hierarchical clustering. I tried to visually determine which group they were closer to, 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 spider diagram, which is often used to compare multivariate continuous data.

#### 2.5 Clustering Text Data

### 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 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 xx and a mean higher than xx, 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 (Yang, Zhang, Zou, Hu, & Qiu, 2013) and maintain interpretability of clusters. As highlighted in Section 1.1 maintaining interpretability was key to assess the meaningfulness of a cluster.