DYNAMIC HEDONIC AND UTILITARIAN SEGMENTATION BASED ON INDIVIDUAL CUSTOMER PURCHASE PATTERNS

by
ZEYNEP KÜÇÜKSARI

Submitted to the Graduate School of Engineering and Natural Sciences
in partial fulfillment of
the requirements for the degree of Master of Science

Sabancı University
July 2021

DYNAMIC HEDONIC AND UTILITARIAN SEGMENTATION BASED ON INDIVIDUAL CUSTOMER PURCHASE PATTERNS

Approved by:

Date of Approval:

ABSTRACT

DYNAMIC HEDONIC AND UTILITARIAN SEGMENTATION BASED ON INDIVIDUAL CUSTOMER PURCHASE PATTERNS

ZEYNEP KÜÇÜKSARI

INDUSTRIAL ENGINEERING MASTER'S THESIS, 2021

Thesis Supervisor: Prof. Dr. Selim Balcısoy

Keywords: Hedonic and utilitarian shopping behavior, Behavior Analysis, Clustering, Unsupervised Learning, Outlier Detection

In the management and psychology literature, the consumers' motivations to make purchases have long been studied under the dichotomic perspective of hedonic vs utilitarian decisions. This perspective is proved to be relevant either to understand why people buy and to help companies to frame their operations strategies to optimize their sales efforts and to maximize customer satisfaction. In this paper, we analyze supermarket transaction data from Brazil over the course of one years to understand and identify utilitarian versus hedonic consumer behavior in a supermarket context. While current literature studies the same notion mainly for a wider set of general shopping categories, we focus on in-supermarket purchases to understand when and how consumers are inclined to make purchases that could be considered hedonic even in a supermarket setting. We develop and propose measures to quantify depth and breadth of purchases along several dimensions including brand/no brand, purchase value, purchase quantity or amount. As the definition of hedonic vs. utilitarian may change from person to person, and in the absence of ground truth, we propose an unsupervised approach to identify outlier transactions that would likely be considered hedonic purchases. A closer examination of selected customers and their transaction sets suggests that our approach produces realistic scenarios under a variety of dimensions and scenarios considered. Our approach brings new theoretical perspectives to advance the hedonic and utilitarian literature in management/operations research.

ÖZET

BİREYSEL MÜŞTERİ HARCAMA DÜZENLERİ BAZ ALINARAK DİNAMİK HEDONİK VE FAYDACI TÜKETİM SINIFLANDIRMASI

ZEYNEP KÜÇÜKSARI

ENDÜSTRİ MÜHENDİSLİĞİ YÜKSEK LİSANS TEZİ, HAZİRAN 2021

Tez Danışmanı: Prof. Dr. Selim Balcısoy

Anahtar Kelimeler: Hedonik ve faydacı tüketim davranışları, Davranış analizi, Kümeleme, Gözetimsiz öğrenme, Anormallik tespiti

Müşterilerin harcama motivasyonları hedonic ve faydacı kararlarla distomik bakış açısıyla uzun zamandır yönetim ve psikoloji alanlarında çalışılmaktadır. Bu bakış açısının, insanların neden satın aldığını anlamak ve şirketlerin satış çabalarını optimize etmek ve müşteri memnuniyetini en üst düzeye çıkarmak için operasyon stratejilerini şekillendirmelerine yardımcı olmak için alakalı olduğu kanıtlanmıştır.Bu araştırmada, süpermarket bağlamında faydacı ve hedonik tüketici davranışını anlamak ve belirlemek için Brezilya'dan bir yıllık süpermarket işlem verilerini analiz ediliyor.Mevcut literatür, aynı kavramı temel olarak daha geniş bir genel alışveriş kategorileri kümesi için incelerken, biz tüketicilerin dikkate alınabilecek satın almaları ne zaman ve nasıl yapmaya meyilli olduğunu anlamak için süpermarket içi satın alımlara odaklanıyoruz.Markalı/markasız, satın alma değeri, satın alma miktarı veya tutarı dahil olmak üzere çeşitli boyutlar boyunca satın almaların derinliğini ve genişliğini ölçmek için ölçüm değerleri geliştiriyor ve öneriyoruz. Hedonik ve faydacı tanımı kişiden kişiye değişebileceğinden ve kesin referansın yokluğunda, büyük olasılıkla hedonik satın alma olarak değerlendirilebilecek aykırı işlemleri belirlemek için denetimsiz bir yaklaşım öneriyoruz.Seçilen müşterilerin ve işlem setlerinin daha yakından incelenmesi, yaklaşımımızın dikkate alınan çeşitli boyutlar ve senaryolar altında gerçekçi senaryolar ürettiğini göstermektedir. Yaklaşımımız, yönetim / yöneylem araştırmasında hedonik ve faydacı literatürü ilerletmek için yeni teorik bakış açıları getiriyor.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my thesis supervisor Prof. Selim Balcısoy and Prof. Burçin Bozkaya for their constant support, patience, and understanding throughout the past two years. Without their encouragement and guidance, it would not be possible to complete this thesis.

In addition, I would like to thank Prof. Vinicius Brei, Carla Freitas Silveira Netto, and Prof. Nina Mazar for their invaluable comments and feedback on the developed tool. Lastly, I would like to thank Asst. Prof. Kemal Kılıç for his presence in the thesis jury.

I would like to thank other BavLab members and former members for their valuable feedbacks and suggestions on my work.

I owe special thanks to my friend Büşra Atar for her encouragement, continuing support, and patience during my thesis writing stage. Her support kept me away from stress and help me to finish my thesis successfully.

Also, I would like to thank my oldest friend Tuğba Güvercinoğlu for her support and understanding during my Master's. Her presence in my life gave me the strength to deal with any problem.

Lastly, I thank Sabancı University for the scholarship program.

"Part of the journey is the end."

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

One of the subjects studied by academics from both management and psychology is the motivations of consumers to make purchases. Several consumer behavior studies have focused on how customers buy, finding cases where consumers purchase for a necessity and other cases where consumers shop for pleasure. These cases are called as "shopping as work" theme [1] and the "shopping for fun" theme [10]. These two shopping behaviors are usually discussed in terms of "utilitarian" and "hedonic". According to the researchers, the shopping experience provides consumers with a combination of utilitarian and hedonic shopping values[10], [1], [4]. In general, when customers acquire a product out of necessity or enjoyment, they perceive utilitarian and hedonic value respectively.

Understanding customer shopping behavior is critical for developing an effective marketing strategy. Marketers have been eager to learn and detect customer goals on the purchase process to be able to get strategic advantages against their competitors. By using hedonic and utilitarian perceptions of a product, companies can allocate their digital spending and decide on marketing strategies along with personalized advertisements. Hedonic and utilitarian characteristics of purchase can help to predict customers' following purchases. Since this prediction is done based on customer's patterns, the results can capture the shopping purposes of customers. Accordingly, companies can make strategic decisions based on not only a necessity but also the social and psychological needs of their customers [3].

The hedonic and utilitarian value has been defined by many researchers from both management and psychology departments. These definitions have the same content with different adjectives. Holbrook and Hirchman (1982) have stated that utilitarian value is a task-oriented and cognitive process [10]. On the other hand, the hedonic value which is related to emotional aspects of the shopping experience is more subjective and individualistic compared to utilitarian shopping value [13]. Researchers defined hedonic shopping value as fun, freedom, fantasy, pleasure, recreation, heightened involvement, new information, and escaping from reality [7]. Utilitarian values can consider as more logical, rational, and related [18].

In the literature, there are some well-known HED/UT scales. These scales were created by conducting a survey with a small sample of students who were asked to rate product categories and brands concerning a pool of adjectives gain from published researches. However, these scales were created for several product categories and brands by generalizing students' rates. Because of these reasons, there are some limitations for these scales. One of them is that they are not individualistic. Also, these scales are not dynamic so they cannot be used in real-life problems. As a result of these limitations, hedonic and utilitarian shopping value for a customer cannot determine by the marketers. Therefore, this thesis aims to create a dynamic structure for hedonic and utilitarian segmentation based on individual consumption patterns. This study proposes a model with 2 consecutive parts.

In this study, data from a supermarket in Brazil are used. The whole data set contains 75 million transactions and 200K customers from 2017 to 2020. However, the data set were reduced to one year that is 2019 for creating the proposed model. This data set contains 37 million transactions and 40,329 customers with a unique ID. For each customer, bucket-based and product-based features were calculated. Time-series analyses were conducted for these values to detect customers' patterns overtime periods. To detect changes in buckets, total sales value, average bucket brand label, maximum bucket brand label, and brand ratio were used. These values were used to identify some pattern combinations and create product clusters. Product-based features are product score, repetition, and category brand label deviation. Individual product clusters were obtained by applying some rules based on these product-based features.

In the first part of the proposed model, product segmentation is applied by the K-means algorithm. Calculated z-values of bucket-based features are used as feature vectors to apply the K-means method. In this part, individual behavior was used to create generalized product clusters by detecting common behaviors.

In the second part, generalized product clusters are personalized by customers' product-based features. Some rules are applied to decide on the transactions of clusters. Once the individual product segmentation is done, the hedonic and utilitarian scores of buckets are calculated over the periods. Since time is an important factor, special dates in Brazil in 2019 are also used to identify which special days have more effects on a customer's hedonic behavior.

The contributions of this study are :

- Individual customer behavior can be detected without any customer interaction such as survey.

- The method is dynamic so it can be applied in different time periods as well.
- It also set a ground for prediction studies.

The remaining of this thesis is structured as follows: in Chapter 2, a comprehensive literature review is presented based on the given definitions and hypotheses. The data set explanation and preprocessing steps are presented in Chapter 3. The methodology is defined in Chapter 4. The results and discussion is stated in Chapter 5. Finally, Chapter 6 consists of the conclusion.

2. RELATED WORK

In this chapter, the literature review will be presented in three parts: an explanation of hedonic and utilitarian shopping values, the measurement methods which are mostly used in researches, and proven hypotheses according to this topic. At the end of this chapter, the gap and the limitations of the previous studies will be discussed.

2.1 Hedonic and Utilitarian Shopping Value

Customer satisfaction is the highest priority for companies that want to increase their profit and have a better reputation in the market [12]. For this reason, companies investigate the motivation behind shopping. Studies that focus on shopping experience suggested that a combination of hedonic and utilitarian shopping value provided for customers [11]. Researchers emphasized that the purchase process of a customer is affected by the customer's goals. There are two different purpose-oriented consumptions. These are task-oriented and emotional aspects [13]. Also, Hirchman (1984) mentioned that shopping experiences entail the stimulation of thoughts and senses, and hence can be considered as a process that gives cognitive (utilitarian) and affective (hedonic) advantages to the individual.

Holbrook and Hirchman (1982) have stated that utilitarian value is task-oriented and cognitive process [10]. Carpenter (2008) generalized that customers perceive utilitarian value by purchasing necessary products [4]. Childers et al. (2002) was also stated that utilitarian tasks are goal-directed and related to the need to fulfill specific tasks efficiently and effectively [5]. Babin et al. (1994) emphasized that obtaining needed products provides utilitarian shopping value which increases concerning fewer effort [1]. Therefore, utilitarian shopping behavior is more logical, rational, and related to transactions [18].

On the other hand, hedonic value is more subjective and individualistic compared to utilitarian shopping value. Holbrook and Hirchman (1982) have mentioned that hedonic value is related to emotional aspects of the shopping experience [13]. Babin, Darden, and Griffin (1994) discussed that besides from customer's experience of shopping, emotional attachment, focusing on fun, playfulness, enjoyment, excitement and the need for surprise are also define hedonic value [1]. Earlier researchers defined hedonic shopping value by using some adjectives. These are fun, freedom, fantasy, pleasure, recreation, heightened involvement, new information, and escaping from reality [7]. Hedonic shopping value refers to the level of perception in which shopping is seen as emotionally beneficial and rewarding due to a variety of happy feelings. Therefore, hedonic shopping value can be defined as the emotional benefits that a customer receives as a result of their shopping experience that is not related to the original purchase goal [14]. Similarly, hedonic shopping value was defined by Babin et al. (1994) [1] as the perceived enjoyment and emotional worth offered by shopping activities.

Figure 2.1 Adjective pools of utilitarian and hedonic values.

In terms of customer benefit, hedonic buying has both good and bad aspects. Impulse shopping, often known as compulsive shopping, is a bad aspect of hedonic shopping [11]. Shopping activity is the important value for compulsive shoppers instead of necessity [9]. The possibility for social contacts with friends, family, or even strangers, as well as sensory stimulation such as escapisms from daily life and new information about current trends and fashion, are all part of the shopping experience [22]

2.2 Hedonic and Utilitarian Characteristics Measurement

Hedonic and utilitarian values have been investigated by different disciplines such as sociology, psychology, and economics [21]. Some researchers have been stated that product/brand attitude is one-dimensional. However, after the view that attitudes are multidimensional and complex, researchers started to work on the experimental view of consumption [21].

According to Batra and Ahtola (1991) in an early attempt to quantify many aspects of product/brand attitudes, “consumers purchase products and services and engage in consuming behaviors for two primary reasons: hedonic gratification and utilitarian reasons" [2]. Voss et al. (2003) adopt this two-dimensional conceptualization by using sensations derived from product experience for hedonic dimension and functions of products for utilitarian dimension [21].

Marketers can test the impact of advertising campaigns that emphasize experiential or functional positioning techniques using measures of the hedonic and utilitarian components of attitude [16]. These measures can indicate brand differences/positions that a single dimension attitude measure may not identify [8]. Hedonic and utilitarian measurement can help managers to decide on pricing and sales promotion [21].

The hedonic/utilitarian (HED/UT) scale which was generated by Voss et al. [21] is the most used measurement model in the literature besides from the previous problematic scales such as Batra and Ahtola [2]. Voss et al. (2003) followed accepted scale development procedures by Churchill and Gilbert (1979) and created a pool of items from different domains of interest [21], [6]. This pool has 10 items with psychometric properties. This scale has been tested based on unidimensionality, reliability, discriminant, predictive, and nomological validity. Both product categories and brands were plotted with scale scores in a two-dimensional space. 608 students and 12 adjective pairs were used to create the HED/UT scale. This paper also stated that H/U characteristics are evident at the product and brand levels.

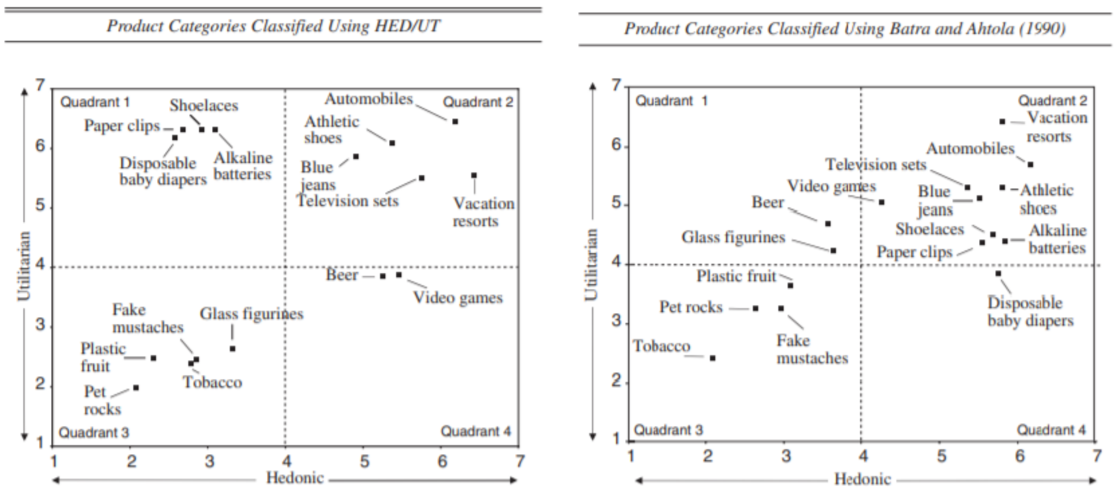


Figure 2.2 HED/UT scaling methods used in previous studies.

Comparison of two scale methods on product categories with hedonic and utilitarian scores can be seen in Figure 2.2. HED/UT scales were also applied with the brand name of the products. This classification is shown in Figure 2.3. This matrix shows that each product and brand has both hedonic and utilitarian values. In summary, Figure 2.2 highlights that the same product category can have different H/U attributes across brands and the same brand can have varying H/U characteristics for its product categories.

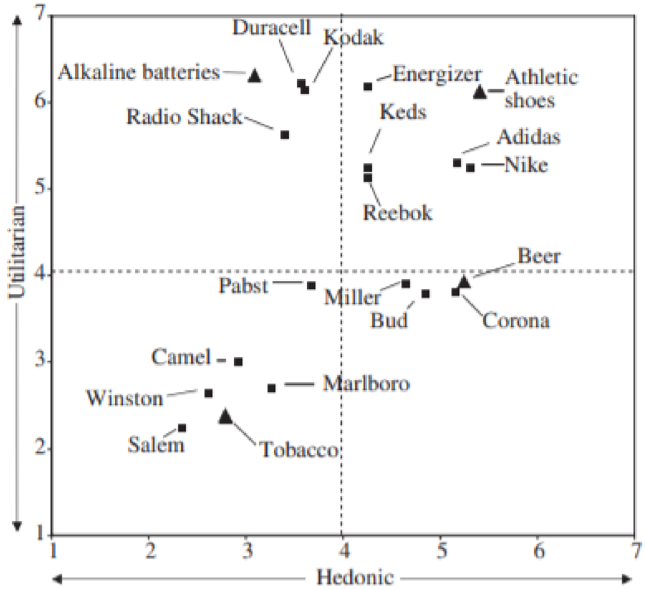


Figure 2.3 Classification of the brand names with HED/UT scale.

2.3 Hypothesis for Hedonic and Utilitarian Shopping Value

Many researchers who work on this topic, first applied a survey to the customer and then stated their hypothesis. In this part, some of these proven hypotheses and information will be stated.

Table 2.1 shows the general differences between hedonic and utilitarian shopping values. The most important statement is that hedonic value is outside of the daily routines. Many types of research investigate the connections between H/U value and price sensitivity, loyalty, customer satisfaction, and discount.

Hedonic Shopping Value	Utilitarian Shopping Value
An end itself	A means to an end
Does not necessarily include purchases	Always includes purchases
Impulsive	Planned
Efficiency not central	As efficient as possible
For pleasure	Out of necessity
Outside of daily routines	Part of daily routine
Emphasis of the experience	Emphasis of rationality

Table 2.1 Differences between hedonic and utilitarian shopping values stated by Irani et al. (2011).

According to Irani et al.(2011), both hedonic and utilitarian shopping values related to variety-seeking buying tendency [11]. They also stated that there is a negative relationship between hedonic shopping value and price sensitivity. Furthermore, both shopping values equally influenced shopping satisfaction.

2.4 Limitations and Contributions

Most of the papers on this topic design their studies with surveys and interviews. Customer interaction will be always necessary for this method. However, companies cannot always survey every purchase. This is why the data collection part is a problem and sample sizes are usually very small. Nowadays, all companies collect data from each transaction. While hedonic and utilitarian scores for products and brands are being detected, this available big data could be useful. Furthermore, even though hedonic and utilitarian scores for each product and brand can vary from one customer to another, many researchers only focus on the segmentation of products for all samples, not for the individual.

In this study, both product and customer segmentation are done. Product segmentation is based on individual pattern combinations of population. Then, by using these product clusters and some individual metrics, products are evaluated and clustered again for each customer. In the end, hedonic and utilitarian scores for customers were calculated. This process helps to create dynamic hedonic and utilitarian segmentation for individuals based on their patterns.

3. DATA AND PREPROCESSING

This chapter will be presented in 4 parts: data collection, preprocessing, feature extraction, and explanatory data analysis. The data collection part will explain the dataset in detail. In preprocessing part, brand segmentation and data reduction will be presented. Extracted features with their definitions and formulas will also be given. Finally, distributions of features will be stated in the last part.

3.1 Data Collection

In this study, secondary data are used which were obtained from one of the leading supermarket companies in Brazil. This company has over 200.000 customers and 11 stores in Brazil. Figure3.1 shows the locations of the stores. Customers have unique CPF information which is a specific client ID for each customer. Some customers used this ID in their checkout process. However, some of them do not use their ID. This is why, in our study, we only consider customers who used their client ID to identify their behavior. The time frame for the entire dataset is four years starting from July 2017 to July 2020. The dataset consists of only one table which has 13 columns and 74,520,868 rows that contain client ID information. The explanations of each column are shown by the Table 3.2.

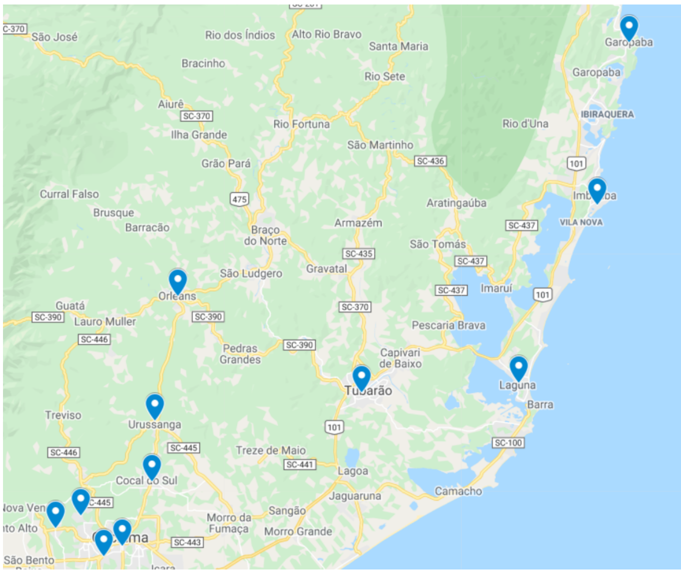


Figure 3.1 Locations of the supermarket branches.

Table 3.1 shows the detailed information for each store.

Store ID	Branch Name	# of Checkouts	Sales area/m^2
2	01-Imbituba	22	2.589,31
3	05-Michel	6	477,00
4	07-Urussanga	10	1.050,00
5	10-Garopaba	18	1.642,15
6	11-Orleans	13	1.335,08
7	12-Laguna	16	1.440,00
9	14-Tubarao	10	1.440,00
15	15-Pinheirinho	7	880,78
16	16-Cocal	5	375,56
17	17-Caravaggio	5	500,00
19	19-Metropol	5	500,00
	Total	**117**	**12.279,88**

Table 3.1 Store informations.

The top 3 sales volumes belong to store 2 (Imbituba), 7 (Laguna), and 5 (Garopaba) which are located near the seaside and have the highest number of checkouts. Seaside locations attract different customers who came to vacation. However, other stores are usually visited by people who live close to their locations.

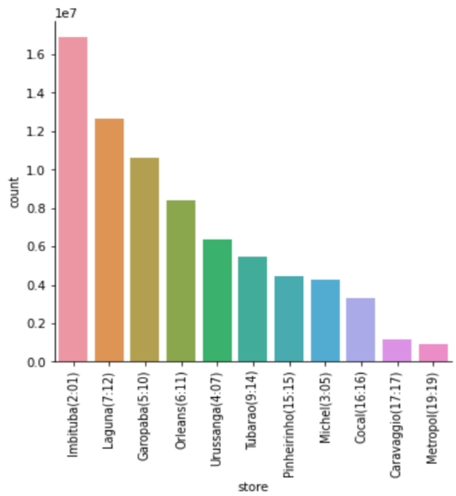


Figure 3.2 Sales distribution of stores over four years period.

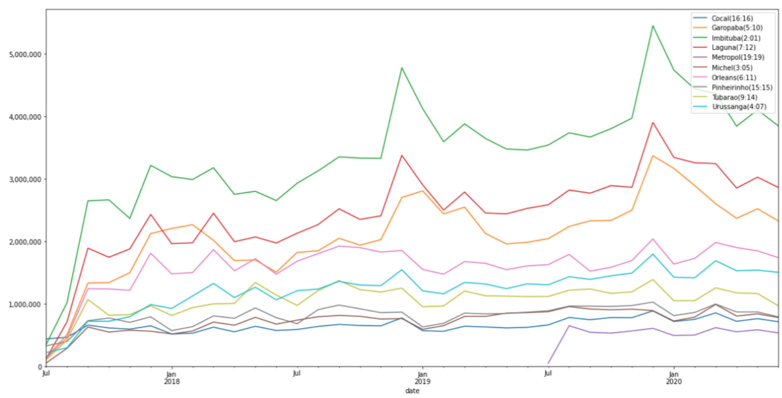


Figure 3.3 Monthly sales volume for each stores over four years period.

The dataset contains 3 types of data: categorical, numerical, and date-time. Table 3.2 shows the column information of the dataset. There are 13 columns in total. Client ID and ID tickets are unique numbers for customers and buckets respectively. Date and time are also available. Each product has its product code and product name. All products are defined with specific departments and categories. There are 14 departments, 241 categories, 18.010 products, and 1691 brands in the entire dataset.

In our study, the dataset that is obtained by the supermarket is used. There is no missing data so data cleaning was not necessary. However, the language of categorical data is in Portuguese so translations were used for category and department columns.

Column name	Data type	Explanation
client id	ID/object	unique customer ID
id ticket	ID/object	unique bucket ID
hour	Date time/datetime64	exact time in checkout
date	Date time/datetime64	date of purchase
store	Categorical/object	store ID of branches
product code	Categorical/object	unique number of each product
product name	Categorical/object	name of each product
department	Categorical/object	general department that each product belongs to
category	Categorical/object	category of each product belongs to
brand	Categorical/object	brand name of each product
sales qty	Numerical/float	Quantity of purchased good in each transaction
price	Numerical/float	unit price for each product
sales value	Numerical/float	cost of product (Quantity*price)

Table 3.2 Data Explanation

3.2 Data Preprocessing

This study focuses on sales transactions that are almost 75 million entries in a supermarket which is located in different cities of Brazil. The supermarket that is used for this research is located in various types of locations. Some of them are near to the sea so they have the highest sales volume because of visitors who came for vacation. Other stores are located inside of the country so they are mostly visited by people who live nearby. The data received from the supermarket consists of details of each sales transaction such as purchased items details, date and time, store ID, sales value, and quantity. Since the data is in Portuguese, the first step of preprocessing is translating string columns to English. For this translation process, googletrans library was applied for the category and department columns. Return items information was also included in the dataset so to avoid negative sales value, these items were removed from customer's buckets that they are registered with. The dataset contains 4 years of data. Each year the number of customers is increasing. For computational efficiency, the method that is proposed in this study is created by using only 1 year of data which is from 2019.

In the second step, some filtering rules were applied to the dataset. Since some customers only visited the supermarkets a couple of times, pattern extraction for these customers cannot be applicable so after calculating each customer's frequency which represents the number of visits to the stores in one year, customers who have less frequency than 24 (quartile 1 value in the sample) were removed from the dataset.

The supermarket has a huge range of variety of products and brands. There are 1.691 brands in this range so detecting brand change was a problem. Also, the supermarket has its brand with low price products. As a solution for this problem, brands were clustered by their average price value in their product range. K-means method was used for this segmentation process. 11 clusters were created according to the average unit price of each brand. Label 0 represents noname brands such as vegetables without brand and Althoff brand which is the name of the supermarket so this brand belongs to the company. Labels from 1 to 10 are in increasing order based on the mean unit prices of brands. This value also represents how luxury brands and their brand value.

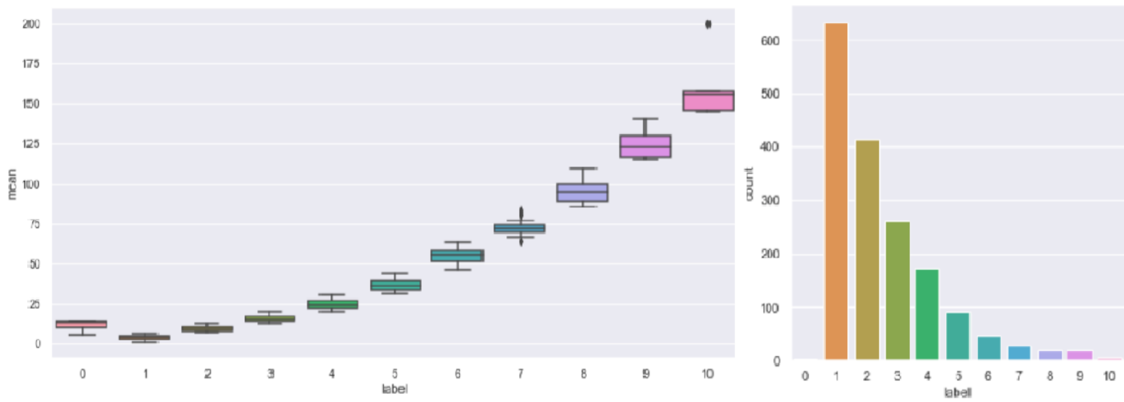


Figure 3.4 Brand labels box plot with respect to the average price and the distribution of brands into labels.

After applying these brand labels into the dataset, the average and maximum brand labels of each bucket were calculated. Figure 3.5 displays the distribution of average bucket labels. Figure 3.6 shows the distribution of the maximum brand label in each bucket. According to these distributions, customers who have not bought any product that is higher than label 4 were removed from the dataset. If a customer never bought anything above label 4, they might visit the stores only for utilitarian products so detecting hedonic and utilitarian products for these customers will not give accurate results. This is the reason behind removing customers who never get any products higher than label 4.

Figure 3.5 Average brand label of bucket distribution

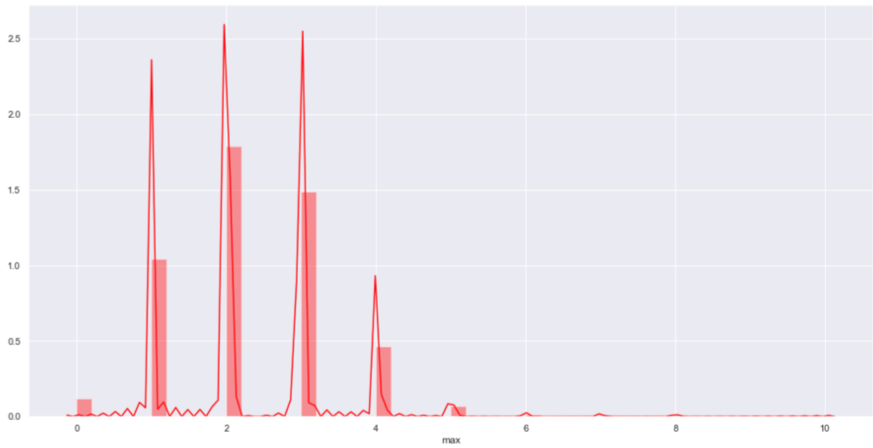


Figure 3.6 Maximum brand label of bucket distribution

After applying all these rules, the remaining number of customers is 33.487. As a result, 33.487 customers with 12 months of data from 2019 were used to create the proposed methodology for detecting hedonic and utilitarian products for each customer by focusing on outlier behavior on sales value, average brand label, maximum brand label, and the brand ratio of each bucket.

3.3 Feature Extraction

The method that is proposed in this study focuses on detecting outliers in the first step. For this purpose, some measurements were calculated based on buckets and products.

3.3.1 Bucket Based Features

Some features were calculated based on bucket information to detect outliers in the individual customer patterns. These features are sales value, average-maximum brand label, and brand ratio.

3.3.1.1 Sales Value

In the dataset, there are transactions item by item so grouping these item transactions are required to calculate general measurements for each bucket. Sales value is equal to sales quantity multiplied by unit price. To find the total sales amount of each bucket, all products' sales values which belong to the specific bucket were sum up.

3.3.1.2 Average-Maximum Brand Label

In section 3.2, the brand label was mentioned. By using these brand label that every product have based on their brand segmentation, the average label is calculated by taking the average of product label in each bucket. The maximum brand label is also calculated for buckets. These values are used to identify customer patterns and how they change over time.

3.3.1.3 Brand Ratio

There are two types of products: brand and unbrand. Brand products are labeled from 1 to 10. Unbrand products are labeled with 0. These products have the same brand as the company and they have cheap prices or brand reputation. For each bucket, the volume of the brand product was calculated as a brand ratio based on sales value.

$$B_{jk} = \frac{\text{Total sales value of brand products in the bucket k for customer j}}{\text{Total sales value of the bucket k for customer j}}$$

3.3.1.4 Z-values

Irani et al.(2011) stated that hedonic shopping values are outside of daily routine. That's why, detecting outliers by considering sales value, average bucket label, maximum bucket label, and brand ratio, is the first step in this method. To find significant differences for these 4 features in a customer's buckets, z-values were calculated. While calculating the z-scores, these steps were applied:

- Sort the buckets by their purchase date and time.
- For each bucket, select a sample from 7 previous buckets and 7 following buckets.
- Use this sample to calculate z-value for the bucket.

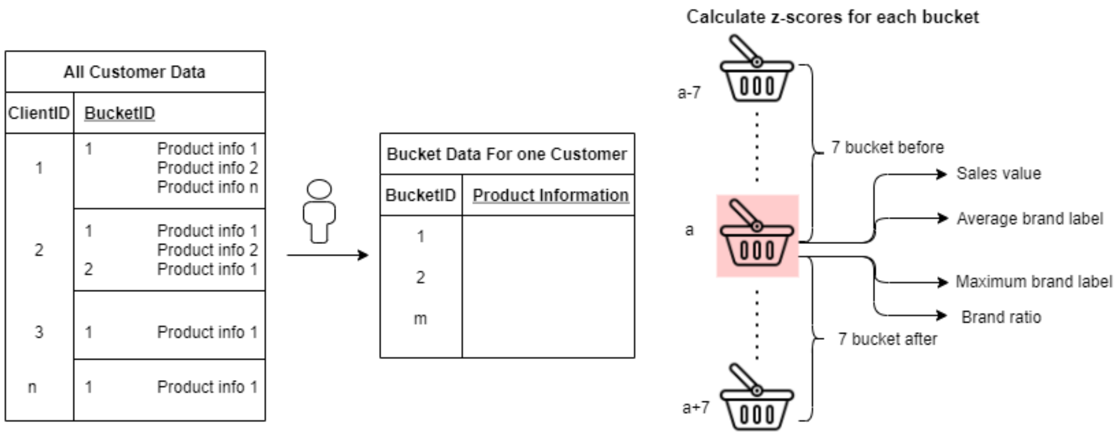


Figure 3.7 Z-values calculations process.

Figure 3.7 explains the steps for 4 z-value calculations. These values show that the selected bucket how close to other buckets. If z-values are small then the bucket is similar to the 7 previous buckets and the 7 following buckets so no outliers. On the other hand, if z-values are high, there could be some outlier in the selected bucket. These kinds of signals are helpful for the next steps.

Feature name	Data type
Client Id	object
Id Ticket	object
Hour	datetime64
Date	datetime64
Store	object
Sales Value	float
Average Label	float
Max label	integer
Brand Ratio	float
Sales Value z-value	float
Average Label z-value	float
Max Label z-value	float
Brand Ratio z-value	float

Table 3.3 Bucket table features properties.

3.3.2 Product Based Features

Since hedonic and utilitarian values are related to product/brand attributes and Voss et al. generated their HED/UT scale with product/brand pairs[21], the product is the key point in this study. However, besides from HED/UT scale which uses a sample of people and score product and brand by all their answer, this study focuses on individual scaling for each product. To be able to do this, some features were created such as product score, monthly repetitions, standard deviation from category average label, and promotion rate.

3.3.2.1 Product score

Product score is calculated as follow:

$$S_{ji} = \frac{\text{Number of bucket that contains product i for customer j}}{\text{Total number of bucket for customer j}}$$

This score is calculated for every product that is existed in a customer's buckets. S_i shows the ratio of occurrence of product i in j^{th} customer's buckets. If this score is small then the product is bought only a couple of times and this can indicate hedonic value.

3.3.2.2 Monthly Repeat and Entropy

There are some products that a customer constantly bought such as fruits, milk, and bread. These products consider as utilitarian according to Irani et al. who state that utilitarian shopping value is related to daily routines [11]. 2 features were calculated for these aspects: monthly repeat and monthly entropy.

$$R_{ji} = \text{Number of month that product i was bought by customer j}$$

Monthly entropy was calculated based on the diversity formula that is stated by Singh et al.(2015) [19].

$$D_j = \frac{-\sum_{t=1}^{N} p_{ji} \log p_{ji}}{\log M}$$

D_j is represented how customer's shopping behavior can vary over 12 months. t is a bin and represents months in this study.

$$N = \text{ number of bins (months) with the counted transactions}$$

$$p_{ji} = \frac{\text{Total sales value in month t for product i that belongs to customer j}}{\text{Total sales value of product i for customer j}}$$

$$M = \text{Number of bins (months) that customer j has transactions.}$$

Because of the normalization, the values D_j are between 0 and 1. Larger numbers mean higher diversity values for customers. In this case, when a customer buys a specific product equally almost every month, the diversity value becomes almost 1.

3.3.2.3 Category label deviation

Buying high-priced products could be an indicator of hedonic shopping value. For example, if a customer buys wine at a lower price most of the time, then in one transaction very high price wine can be bought by this customer. This product will be considered hedonic for this customer since this is impulsive and outside of routine behavior. For this reason, average category labels were calculated for every customer and product labels were subtracted from this average to find how a product's label deviates from its category average.

V_{ji}= Brand label of product i for customer j - Average category brand of product i for customer j.

3.3.2.4 Promotion rate

This company mostly give different discount rate for same products in various time intervals. By finding the maximum unit prices for every product, promotion for that product in all transactions were calculated. This value does not include the method but it is used to check the results and to see how hedonic/utilitarian values behave with various promotions.

3.4 Explanatory Data Analysis

In this section, explanatory data analysis that is conducted before proposed method will be presented.

3.4.1 Data Details and Transaction Counts

In the original data for 2019, there are 19.653.973 transactions and 40.329 customers. After the filtering conditions, the dataset that is used in this study has 18.140.521 transactions and 33.487 customers. Most of the customers have less than 100 buckets in the data set.

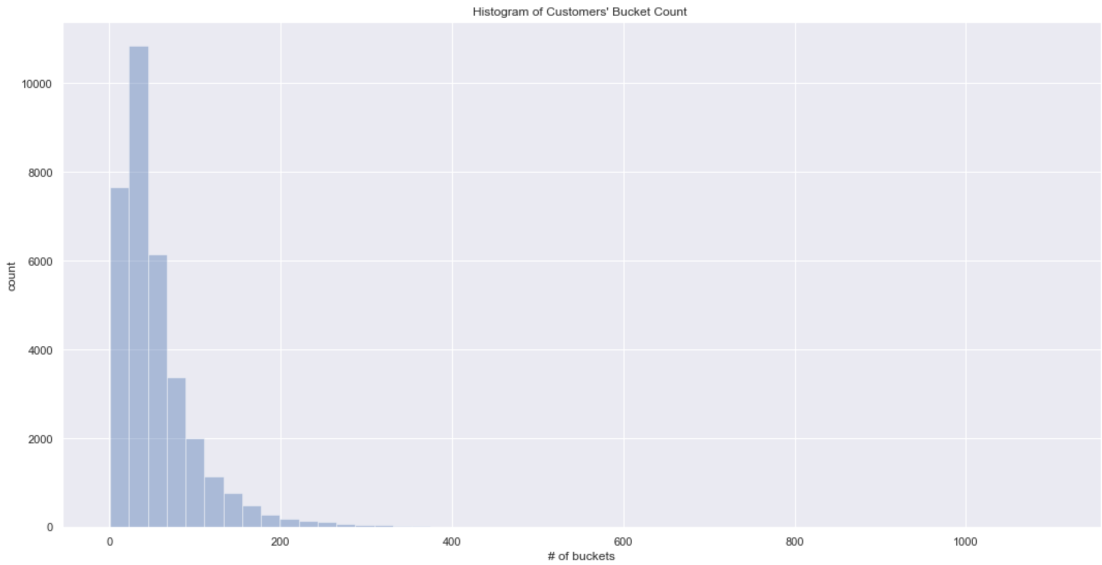


Figure 3.8 Histogram of Customers' Bucket Count in 2019.

Total daily sales value change is shown in Figure 3.9. This heat map shows that most of the sales occur on Friday and Saturday. Also, some specific days are more than the usual amount. These days correspond to the same time for each month and the reason is that people in Brazil get their salary around that time of the month.

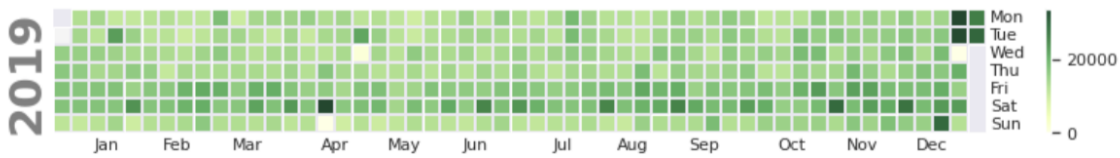


Figure 3.9 Calendar heatmap total daily sales value in 2019.

3.4.2 Z-values Distributions

There are 4 types of z-values in features. The distribution of these values for all the data set shown in Figure 3.10

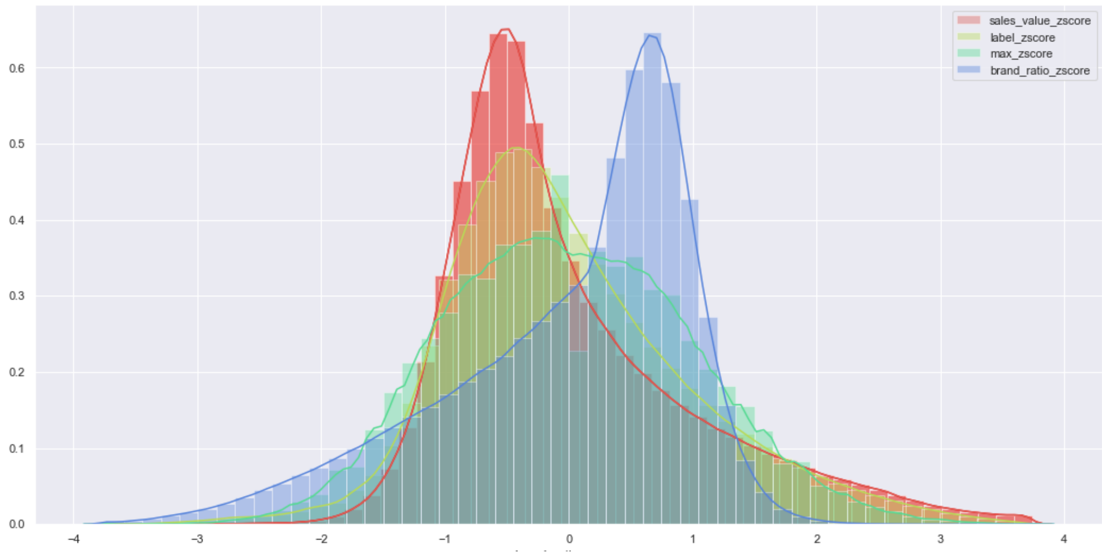


Figure 3.10 Distribution of z- values for sales value, average label, max label and brand ratio for all customers.

This figure shows that sales value and brand ratio have symmetrical distributions. While sales value has a right-skewed distribution, the brand ratio has a left-skewed distribution. This means that sales value has more positive deviated z-score and positive outliers. However, most of the brand ratios of buckets are getting smaller compare to 7 buckets before and 7 buckets after. Even though average label and max label distributions look similar, the max label has a higher variance than the average label. This means that there are some buckets that have the higher brand product in them as an outlier.

4. METHODOLOGY

In this chapter, the algorithms and methods that are used in this study will be explained. First, product segmentation is conducted with K-means clustering with the features that are created by individual patterns. Since the data does not have any label, this process is called unsupervised learning. Secondly, product segments and some individual features that were mentioned in Section 3.3.2 are used to cluster the product separately for each customer. Finally, hedonic and utilitarian scores were calculated bucket by bucket using product clusters.

4.1 Unsupervised Learning

There are two types of approaches for the data mining process: supervised learning and unsupervised learning. In supervised learning, the data has labels and outcome prediction can be performed. However, unsupervised learning is used when the data does not have the label and can perform unknown pattern detection. Because of the lack of ground truth, accuracy measurement can not be performed for the models. Cluster analysis is one of the unsupervised learning methods which also include principal component analysis and association rules. In this study, cluster analysis with unsupervised learning will be applied.

4.2 K-means Clustering

K-means clustering algorithm targets to divide the data set into k number of segments and this algorithm developed by MacQueen et al. (1967) [15]. The most efficient k number is determined before clustering by applying some methods such as the Elbow method or Silhouette score. The number of clusters is defined in the first step by setting initial cluster centers for randomly selected data. Then all points are assigned to the closes cluster center with respect to Euclidean distance. After this step, new cluster centers are calculated and cluster points are rearranged according to these centers' distance. When the centers remain the same after some iterations, the final cluster is defined.

In this study, the K-means method is used to created product segmentation with 18 features that are created by bucket-based features. In this section, detecting the number of clusters and creating product features vectors will be explained.

4.2.1 Detecting Number of Cluster

The Elbow method is used to detect the number of clusters in this study. Range from 1 to 16 is used as k and Within Clusters Sum of Squares is calculated. This process is done with the sci-kit learn library in python.

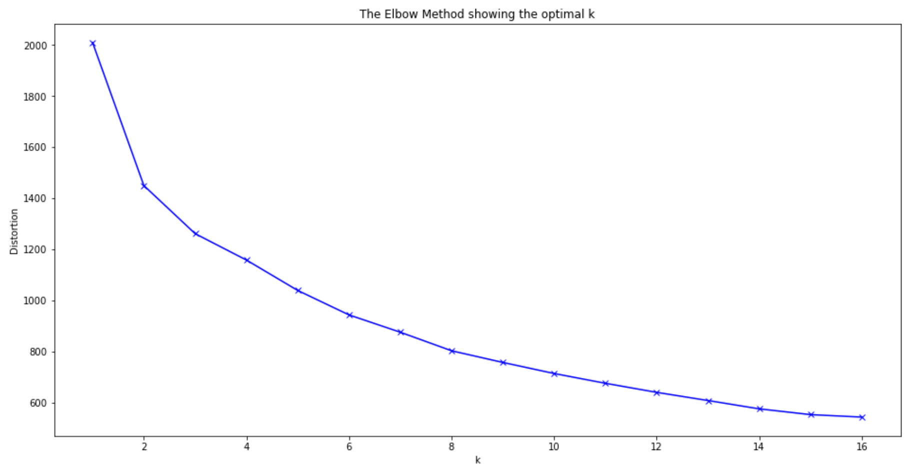


Figure 4.1 Optimal number of clusters with the Elbow method.

In addition to the Elbow method, Within Clusters Distance (WCD) and Between Cluster Distance (BCD) are calculated with the Euclidean distance formula. Then the ratio is calculated by using this formula:

$$Ratio_i = \frac{WCD_i}{WCD_i + (\frac{1}{n}\sum_{l=1}^{k} BCD_{ik})}$$

When the mean of this ratio for all product i is taken for clusters 2,3 and 4 to see the improvement, it is seen that there are no significant differences between 3 clusters and 4 clusters as seen in Table 4.1. After 3 clusters, increasing the k value will not provide better modeling of the data. Because of these reasons, the number of clusters is selected as 3.

Number of Clusters (k)	Average Ratio$_i$
2	0.306158
3	0.271085
4	0.254845

Table 4.1 Average Ratio values of 2,3, and 4 clusters.

4.2.2 Feature Vectors

In Section 3.3.1.4, z-values were presented for 4 features: sales value, average bucket label, maximum bucket label, and brand ratio for each customer. In Figure 4.3, z-values changes can be seen over time.

In general, the process of method that is used for product segmentation is described in Figure 4.2.

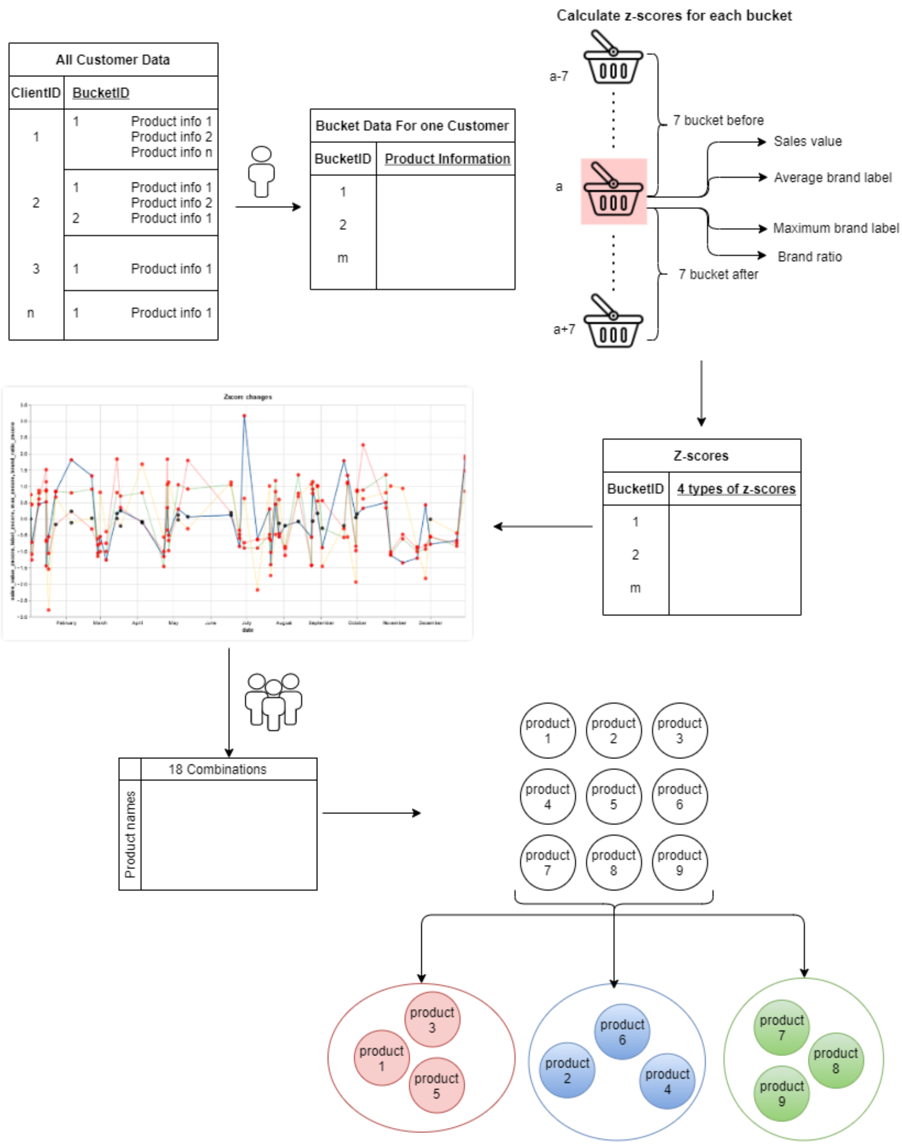


Figure 4.2 Product segmentation overview.

In the first step, for each type of z-values, %95 confidence interval was calculated. If a data point is above the upper bound of the confidence interval, it is called peak point (1). If it is below the lower bound of the confidence interval, it is called dip point (-1). If it is in the confidence interval, it is called average point (0).In Figure 4.3, red points represent peak and dip points. The same method is applied to all customers.

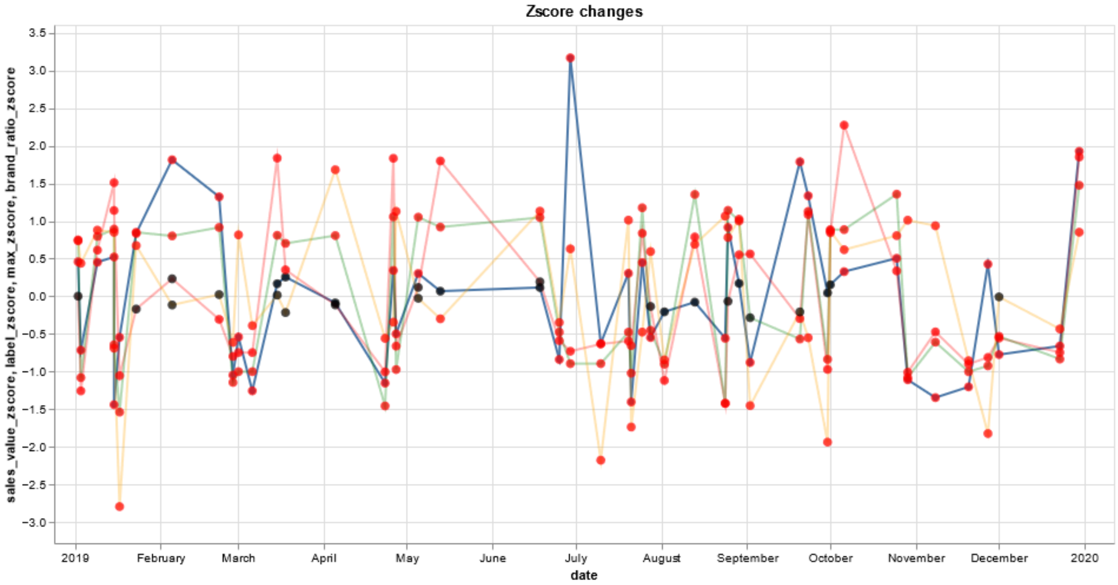


Figure 4.3 Z-values change over time for one customer.

In the second step, It can be seen that these points form some combinations. In general, there are 81 combinations with 4 features and 3 points. However, 18 features were created by aggregating the combinations that include average points in them.

A matrix is created for all products with combinations. These 18 features are used for product segmentation via the K-means algorithm. C_{if} is between 0 and 1. Figure 4.4 shows the features and products with numerical values.

$$C_{if} = \frac{\text{Total sales value of product i that occurs in combination f}}{\text{Total sales value of product i}}$$

product_name	0	1	2	3	4	5	6	7	8	...	10	11	12	13	14	15	16	17	
##LANTERNA RAYOVAC ILUMINA C/ 4 PIL.PEQ.	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	
##LANTERNA RAYOVAC SPORT P/USO C/BATERIA	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
##SEMENTE ISLA TODAS	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
ABACATE KG	0.008301	0.052777	0.008321	0.044464	0.024994	0.114894	0.051362	0.014084	0.028085	...	0.006800	0.001332	0.013253	0.003593	0.001164	0.000831	0.003532	0.596583	1
ABACAXI HAVAI UN	0.029884	0.044826	0.014366	0.014942	0.013924	0.127349	0.044029	0.014942	0.029087	...	0.000000	0.013924	0.029308	0.000000	0.000000	0.000000	0.000000	0.593536	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
YAKISSOBA SEARA CX 500G	0.006235	0.030514	0.109566	0.064014	0.075190	0.014994	0.027727	0.058702	0.021887	...	0.021756	0.009330	0.000000	0.000000	0.000000	0.003183	0.000000	0.553696	
YAKISSOBA SEARA M GOURMET CX 350G	0.006305	0.010862	0.130683	0.071151	0.049519	0.035755	0.026866	0.048298	0.024969	...	0.017175	0.006663	0.000000	0.003203	0.002189	0.008400	0.000000	0.553132	
YAKISSOBA VERD FACIL 500G	0.011471	0.039778	0.033579	0.051514	0.040709	0.041374	0.033060	0.007886	0.018899	...	0.019531	0.023097	0.020797	0.000000	0.008008	0.000000	0.004802	0.630749	
ZZZFORMINHA ALEGRIA N4 AZUL C/200 FL	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
ZZZGARFO CHURRASCO SOL 55CM STOLF FL	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	

Figure 4.4 Product-Category Matrix with $C_{if} values$.

Feature number	sales value	average label	max label	brand ratio	name of combination
0	0	0	0	0	avg
1	1	1	1	-1	slm-b
2	-1	-1	-1	-1	-s-l-mb
3	1	1	1	1	slmb
4	1	-1	1	1	s-lmb
5	-1	-1	-1	-1	-s-l-m-b
6	1	-1	-1	-1	s-l-m-b
7	1	-1	-1	1	s-l-mb
8	1	-1	1	-1	s-lm-b
9	-1	1	1	-1	-slm-b
10	-1	1	1	1	-slmb
11	-1	1	-1	1	-sl-mb
12	-1	1	-1	-1	-sl-m-b
13	-1	-1	1	-1	-s-lm-b
14	1	1	-1	1	sl-mb
15	-1	-1	1	1	-s-lmb
16	1	1	-1	-1	sl-m-b
17	-1/1/0	-1/1/0	-1/1/0	-1/1/0	include 0

Table 4.2 Combinations and feature numbers.

Table 4.2 shows the peak, dip, and average points that create combinations. Name of combinations are abbreviations that stand for:

- s: sales value z-score has a peak.
- -s:sales value z-score has a dip.
- l:average bucket label z-score has peak.
- -l:average bucket label z-score has dip.
- m:maximum label z-score has peak.
- -m:maximum label z-score has dip.
- b:brand ratio z-score has peak.
- -b:brand ratio z-score has dip.
- all zero: average points in all 4 z-scores.

4.3 Individual Segmentation

In the previous section, product segmentation is done based on individual patterns and combinations in an aggregate way. However, hedonic and utilitarian values vary from customer to customer. This is why one more step is required in the method. After getting clusters to form product segmentation, 3 features are also involved in the system for the final cluster decision about product segment for a customer. Product score, monthly repeat, and category label deviation features are used to create a series of rules. In general, Figure 4.5 represents the process.

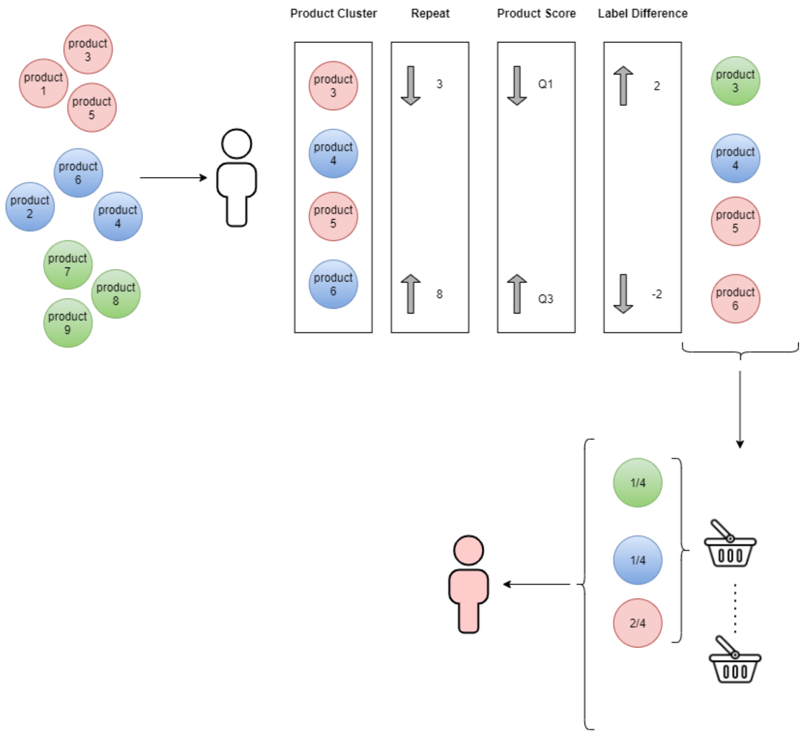


Figure 4.5 Individual segmentation process flow.

First, for each customer, product cluster and product-based features that are mentioned in Section 3.3.2 are consider. These rules will be applied for all products:

- Monthly repeat is less than 3 AND
- Product score is less than quartile 1 value of the given customer's all product score AND
- Category label deviation is greater than 2.

This product will be consider as hedonic for this customer.

Utilitarian conditions are:

- Monthly repeat is greater than 8 AND

- Product score is less than quartile 3 value of the given customer's all product score AND
- Category label deviation is less than -2,

Upper and lower bounds will be explained in Chapter 5.

5. RESULTS AND DISCUSSION

In this chapter, a description of the product segmentation and dimension reduction methods' results will be given. Furthermore, randomly selected customer's product segmentation and hedonic/utilitarian scores will be represented with various graphs. Lastly, the histograms that show the change of product segment after individual segmentation applied for 11.000 customers.

5.1 Product Segments

The K-means clustering algorithm is used based on 18 features regarding the customers' shopping pattern combinations that are determined by sales value, average bucket label, max bucket label, and brand ratio z-scores.

5.1.1 Dimension Reduction

Since the visualization is hard for 18 features, dimension reduction methods are applied to feature vectors. Figure 5.1 shows that the k-means clustering method is applied with high success and divides the products into 3 clusters with minor mistakes.

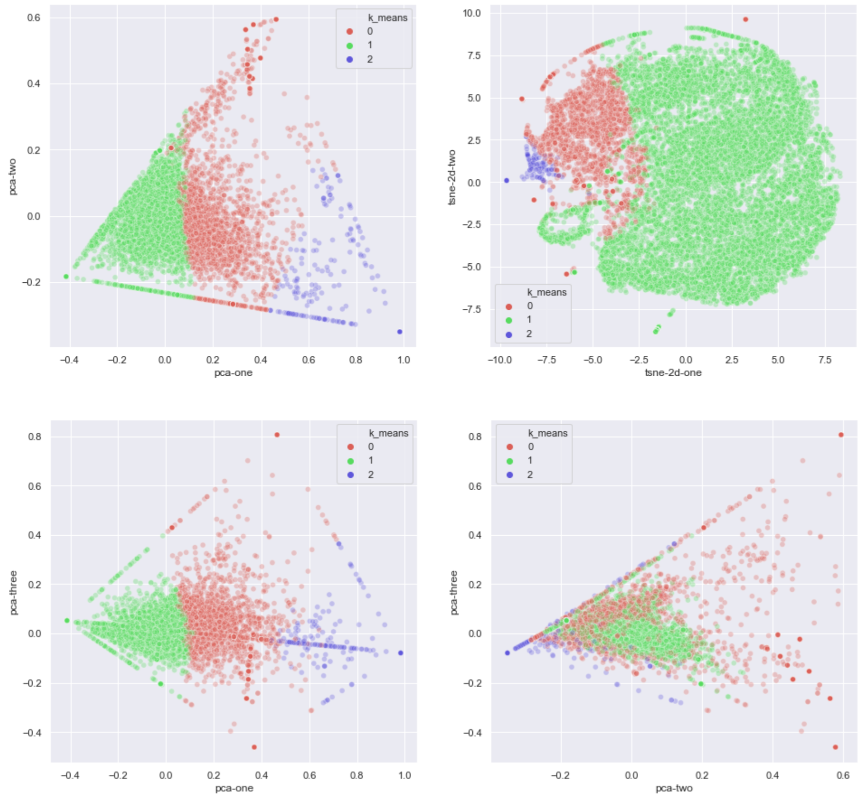


Figure 5.1 Dimension reduction with PCA and t-SNE methods and cluster divisions.

PCA stands for Principal Component Analysis and it is used for dimension reduction in massive distributed data sets. This method was developed by Qu et al. (2002) [17]. The algorithm does not require the raw data. Only feature vectors are used for the process. The basis of this method is based on within clusters covariances and between clusters covariance. To sum up, PCA is an unsupervised linear dimensionality reduction and data visualization method for high dimensional data.

t-SNE stands for t-distributed stochastic neighborhood embedding. t-SNE is presented by Maaten et.al (2008) [20]. It is also an unsupervised non-linear dimensionality reduction and data visualization method. In contrast to PCA, it seeks to retain the data's local structure by minimizing the Kullback–Leibler divergence (KL divergence) between the two distributions concerning the map's point locations.

As a result of dimension reduction and visualization, it can be said that the segmentation is working well.

5.1.2 Cluster Centers and Personas

K-means cluster algorithm creates 3 clusters but which one of these clusters has a hedonic or utilitarian shopping value indicator is not defined yet. Center points of clusters can give a better understanding of the behavior value of each cluster. Figure 5.2 shows that the data frame that contains the center values for 18 features in an order which starts with a high number of peaks to a high number of dips.

	slmb	slm-b	sl-mb	s-lmb	-slmb	-s-lmb	-slm-b	-sl-mb	all zero	s-lm-b	sl-m-b	s-l-mb	-s-l-mb	-s-lm-b	-sl-m-b	s-l-m-b	-s-l-m-b	other
0	0.244463	0.055317	0.004641	0.063756	0.118040	0.004793	0.021143	0.009519	0.006600	0.019428	0.001507	0.018048	0.022709	0.000880	0.001969	0.010485	0.015362	0.381341
1	0.836132	0.020257	0.001119	0.017215	0.026293	0.001070	0.002134	0.000009	0.000044	0.002611	0.000000	0.005628	0.001377	0.000000	0.000000	0.001822	0.004686	0.079602
2	0.086028	0.025653	0.002958	0.056022	0.026008	0.004605	0.009086	0.007753	0.006379	0.019524	0.001471	0.030115	0.043978	0.001615	0.003747	0.021770	0.028229	0.625059

Figure 5.2 Cluster centers for 18 features.

According to center information, cluster labels will be changed as follows.

Cluster Number	Cluster Label
0	both hedonic and utilitarian
1	utilitarian
2	hedonic

Table 5.1 Cluster labels.

Combination and category with clusters are given in Figure 5.3. The x-axis is a combination name by decreasing peak and increasing dip order. The y-axis is the category with an increased order of average brand price.

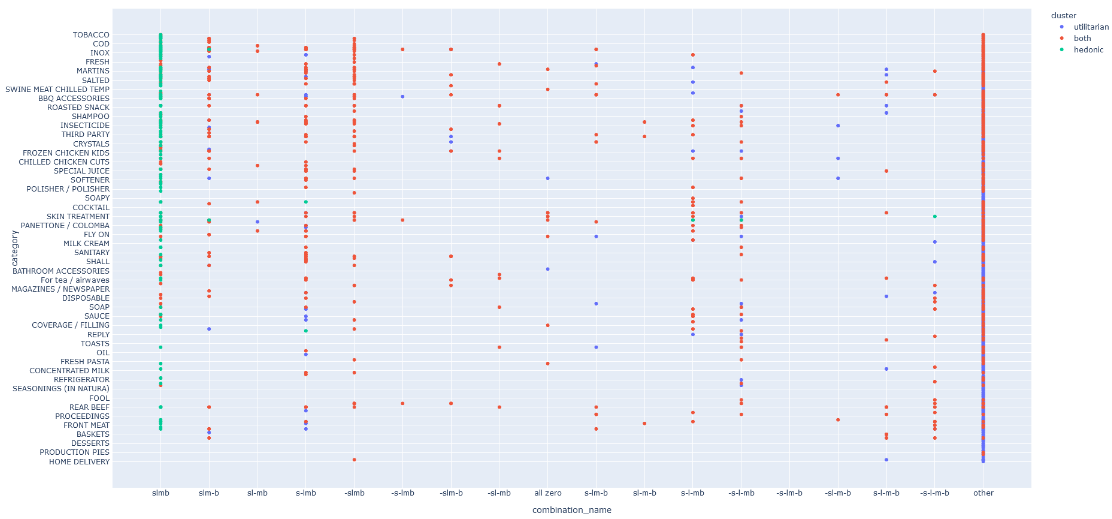


Figure 5.3 Combination and category with increasing price based on clusters.

Both cluster centers and Figure 5.3 are considered to decide segments' labels. The properties of clusters are as follows.

Hedonic products' properties:

The products that are labeled as hedonic mostly occur in a bucket with the following properties:

- All 4 z-values for the bucket are significantly different from other purchases in a given time (14 visits).
- If the sales value does not have a big difference average bucket label, max label, and brand ratio should have a higher variance from the samples.
- It can not occur in any bucket within confidence interval which means the bucket is in daily routine.

Utilitarian products' properties:

The products that are labeled as utilitarian mostly occur in a bucket with the following properties:

- Any one of the 4 z-values is in the confidence interval. This means that the bucket is in daily routine so no big differences can be caused by impulsive shopping behavior.
- Even if sales value has a huge increase, it still is considered as utilitarian if all label-related z-values making dip points. This is an example of a bulk purchase.

Both hedonic and utilitarian products' properties:

The products that are labeled as both hedonic and utilitarian mostly occur in a bucket with the following properties:

- In the case of buckets with average points in any z-values, these products occur with high labeled categories. This can be seen in Figure 5.3 and also in the centers the combinations that contain peak points in maximum bucket label z-value have a higher ratio for this label.
- Since this label is a mix of two behavior, it mostly occurs all of the combinations. However, there is a noticeable difference between utilitarian and hedonic products based on their brand label. Utilitarian products usually have lower brand labels.

Even though there is no feature related to the brand label of a product, Figure 5.4 shows that the K-means divide in a reasonable way according to brand label. This means that using pattern combinations as features is a good way to segment the products.

Figure 5.4 Brand label and category comparison for clustered products.

5.1.3 Hedonic/Utilitarian Scores for Categories

After the segmentation is done for products, categorical hedonic and utilitarian weights can be calculated with respect to the sales value ratio of products' clusters in each category and department. Scores are calculated by the following formula. m is the number of clusters (m=3).

$$W_{mc} = \frac{\text{Total sales value of m labeled products in category c}}{\text{Total sales value of category c}}$$

These scores are between 0 and 1. If it is closed to 1 for a cluster, then the category mostly contains that cluster.

In Figures 5.5, 5.6 and 5.7, color represent departments, size of the points are the average label of the category, and the points are for categories. In Figure 5.5, there is a noticeable diagonal line which means that these categories have only products with utilitarian and both labels. It can also be seen that categories with higher cluster product scores have a higher brand price as well.

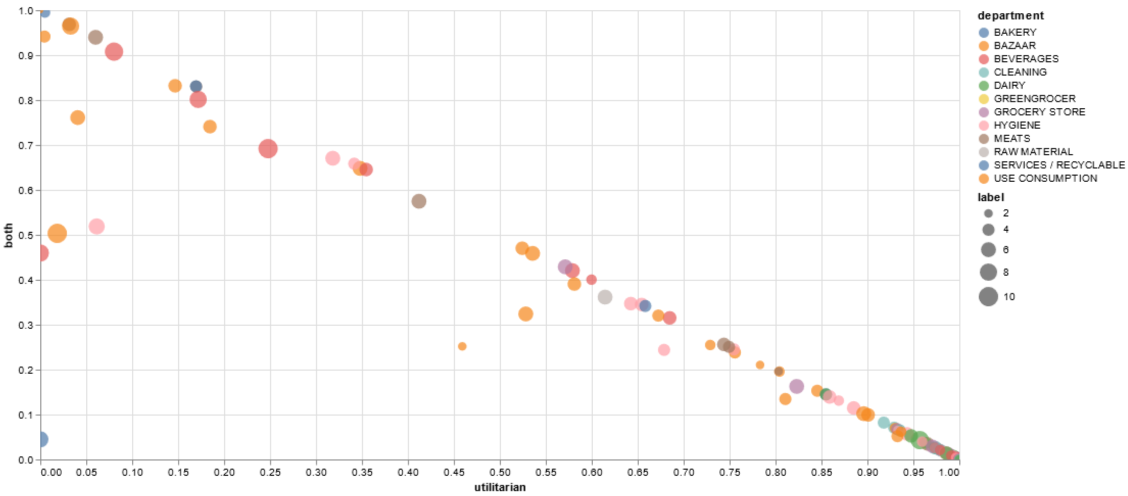


Figure 5.5 Utilitarian and both label category scores.

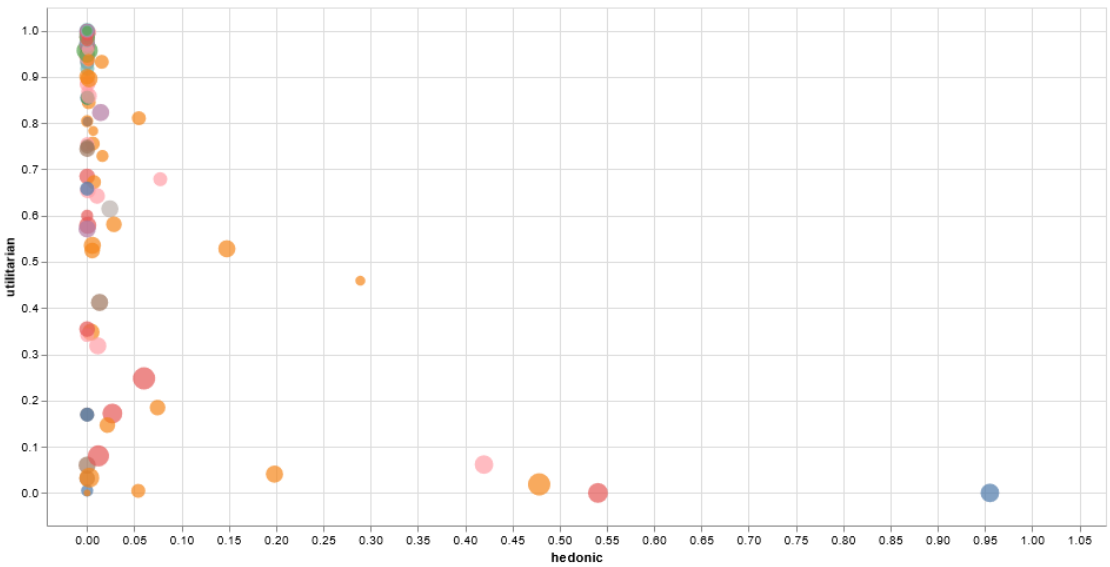


Figure 5.6 Utilitarian and Hedonic label Category scores.

In Figure 5.7, the categories with higher hedonic scores are very small and have higher brand labels. These graphs are making sense when the previous studies' definitions is considered.

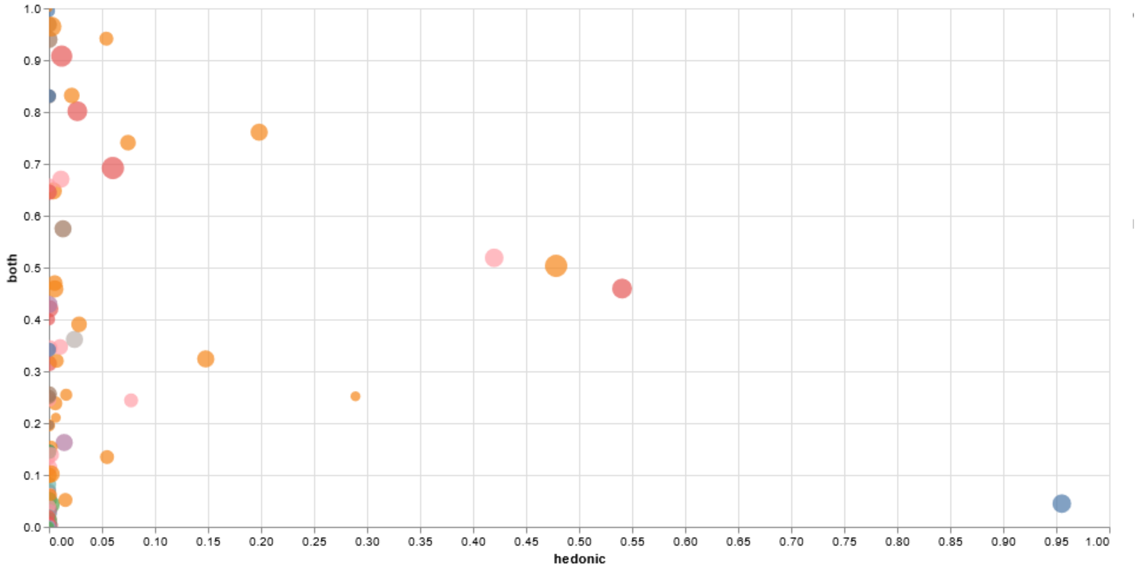


Figure 5.7 Hedonic and both label category scores.

According to Batra et al. (1990) some product categories, brands, and behaviors might be more positively closed to one of the dimensions (hedonic and utilitarian) than other [2]. These scores are shown above.

5.2 Individual Customer Product Segments

In the previous section, product segmentation is done in aggregate settings. However, each product changes its hedonic/utilitarian values from one customer to another. Because of this issue, some rules are applied for each customer's products. These rules were mentioned in Section 4.3. In this section, there will be a case study with a selected customer. For 11.000 customers, the number of products that change their segments after this method is represented in Figure 5.9.

cluster_names	cluster_ind	product_name
both	both	2084
both	hedonic	366
hedonic	hedonic	261
utilitarian	hedonic	179
utilitarian	utilitarian	11744

Figure 5.8 Overall product segment changes.

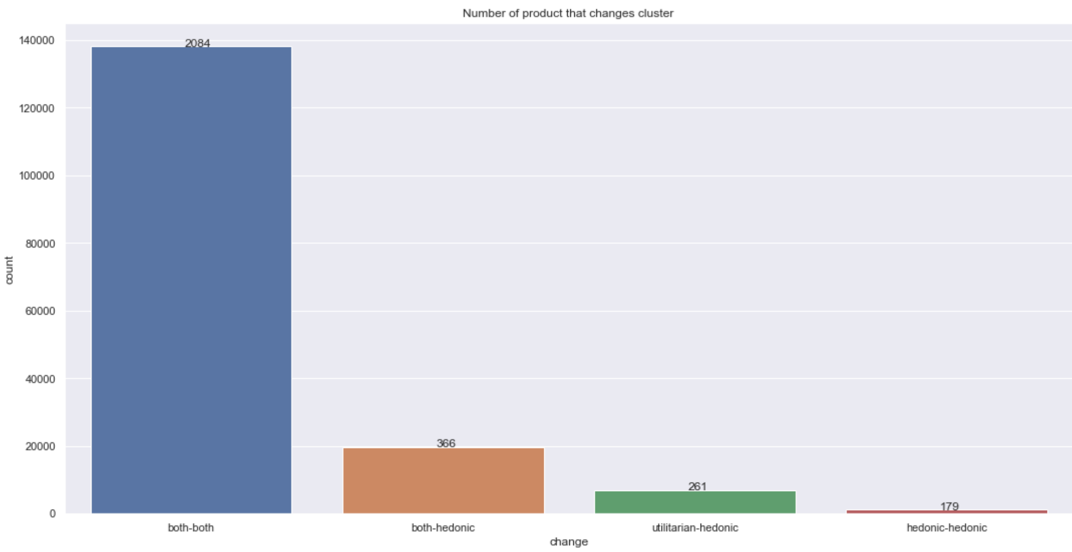


Figure 5.9 Overall product segment changes histogram.

5.2.1 Case Study : Customer 1

This customer has 414 buckets in one year. Sales value changes over one year period is given in Figure 5.10.

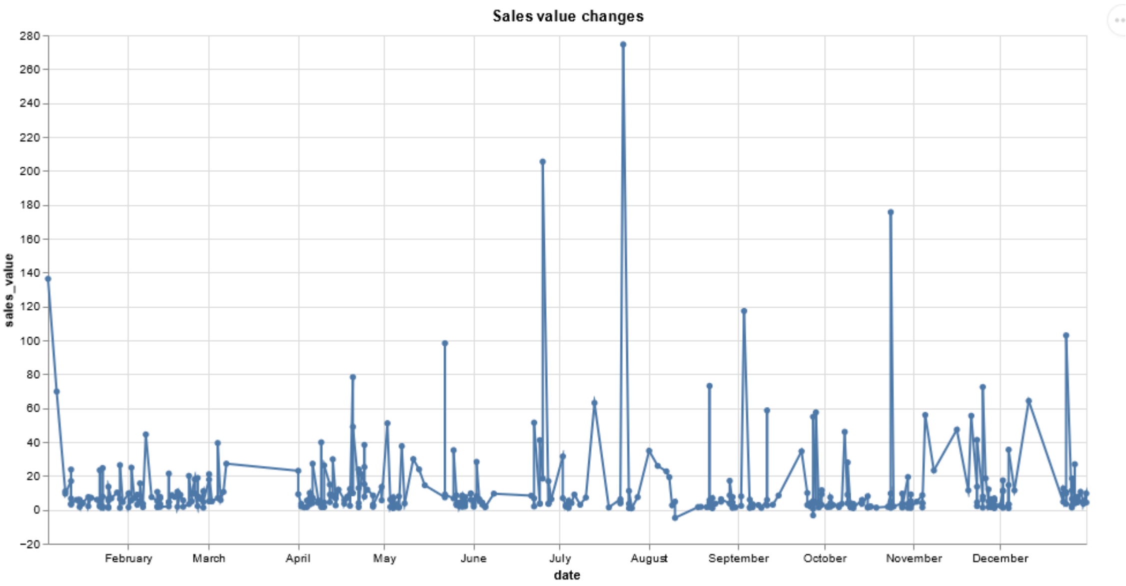


Figure 5.10 Sales value change for Customer 1.

According to product segmentation, this customer does not have any hedonic product. In Figure 5.11, some products are only bought at one time with a very small product score so some of these products should be hedonic for this customer.

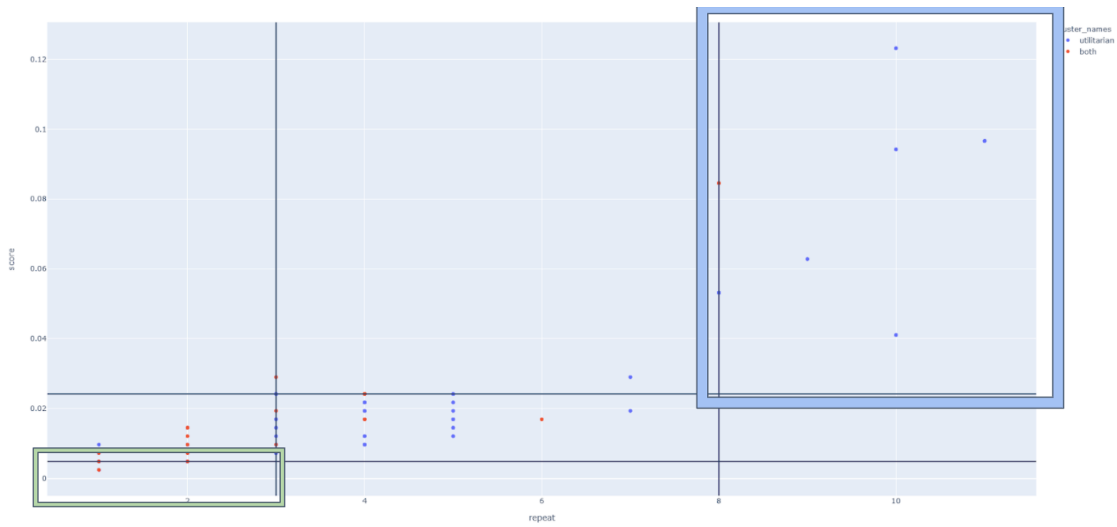


Figure 5.11 Monthly repeat and product scores with initial clusters.

If a product's brand label is higher than 2, this shows that the customer buys a product that is higher than its average category label. For example, this customer usually buys chocolate with label 1. However, one day he/she bought chocolate with label 4.

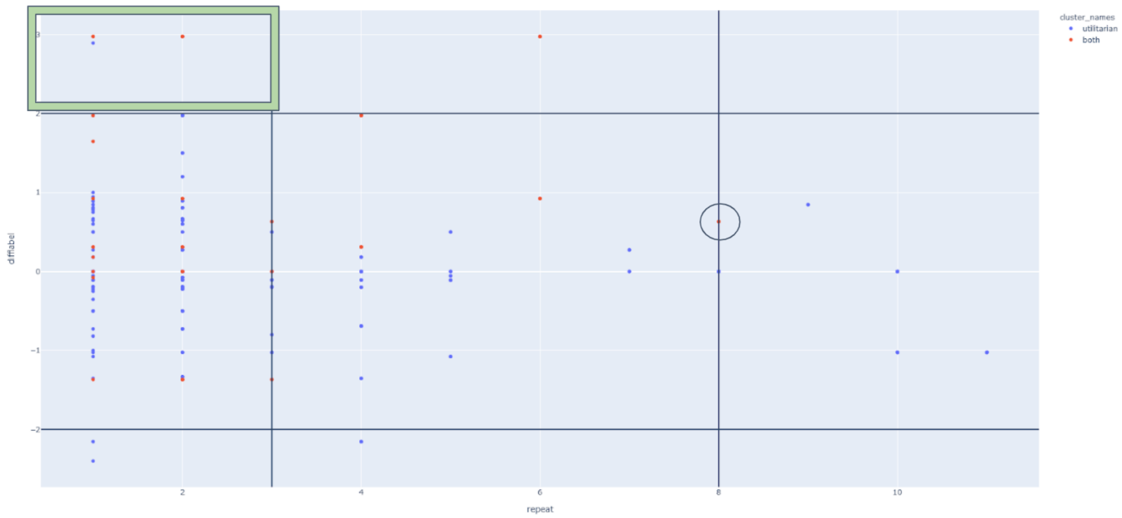


Figure 5.12 Monthly repeat and brand label deviations with initial clusters.

After applying the conditions, final results are:

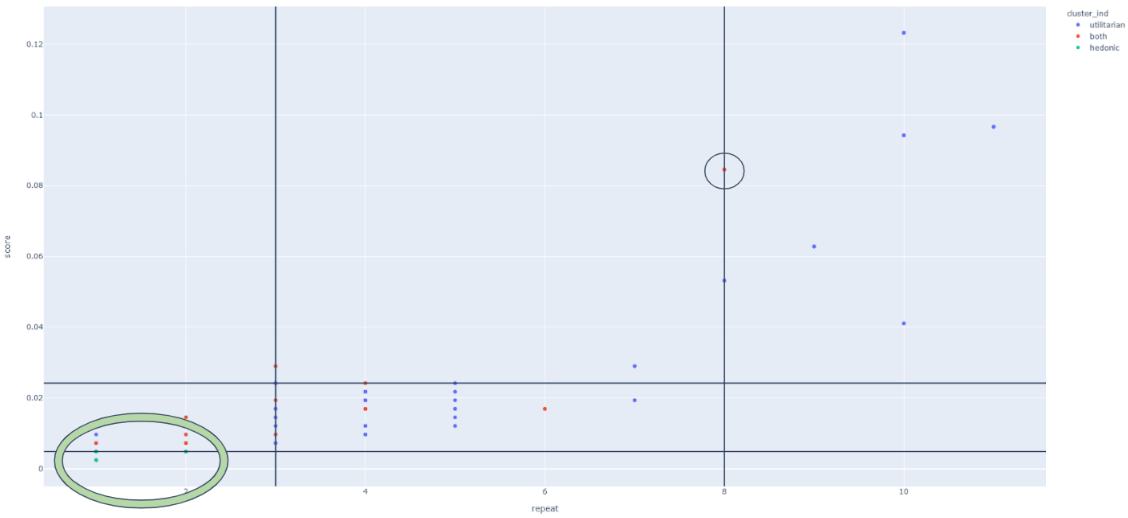


Figure 5.13 Monthly repeat and product scores after individual clusters.

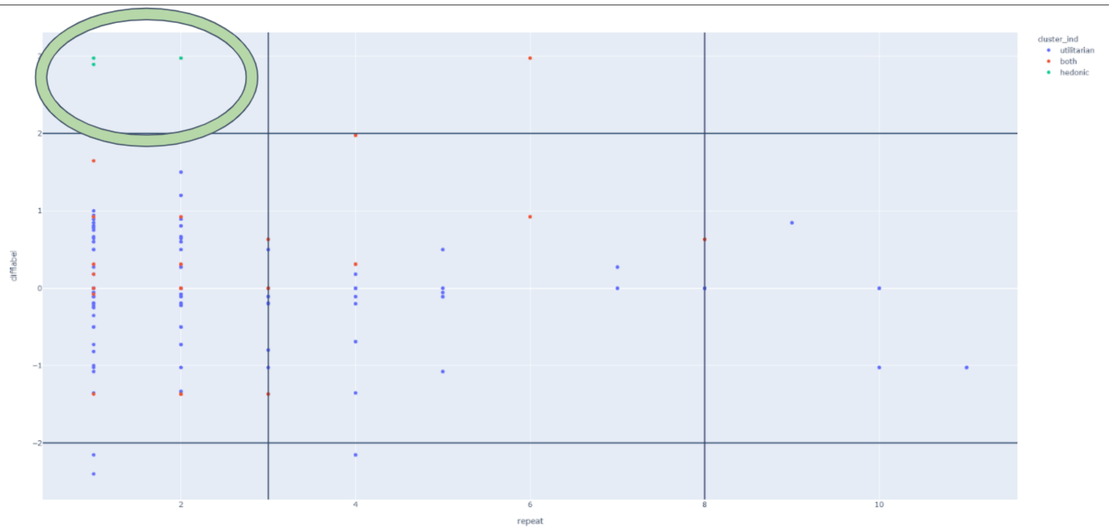


Figure 5.14 Monthly repeat and brand label deviations after individual clusters.

7 products clustered as hedonic after this method. These products are:

store	hour	product_code	product_name	department	category	brand	sales_qty	sales_value	price	date	client_id	label_x	cluster_names	cluster_ind
2	15:15:00	4533	SALG ASSADO FOLH GULA COQ DIVERSOS KG	BAKERY	ROASTED SNACK	GULA	0.09	2.72	29.89	2019-04-05	216438	4	both	hedonic
2	09:35:00	4945	FOLHADO GULA FRANGO/REQ UN	BAKERY	ROASTED SNACK	GULA	1.00	3.75	3.75	2019-07-22	216438	4	both	hedonic
2	13:27:00	4950	MINI CROISSANT GULA CHOCOLATE KG	BAKERY	ROASTED SNACK	GULA	0.05	1.82	37.92	2019-08-30	216438	4	both	hedonic
2	13:35:00	4950	MINI CROISSANT GULA CHOCOLATE KG	BAKERY	ROASTED SNACK	GULA	0.03	1.14	38.00	2019-08-30	216438	4	both	hedonic
2	13:20:00	4522	MINI PASTEL NATA GULA MACA KG	BAKERY	ROASTED SNACK	GULA	0.09	3.42	38.00	2019-09-27	216438	4	both	hedonic
2	13:11:00	4945	FOLHADO GULA FRANGO/REQ UN	BAKERY	ROASTED SNACK	GULA	1.00	3.75	3.75	2019-10-31	216438	4	both	hedonic
2	16:07:00	53202	OVO PASCOA ARCOR TORT ESBUG BCO 150G	GROCERY STORE	CHOCOLATES	ARCOR	1.00	29.90	29.90	2019-04-20	216438	4	utilitarian	hedonic

Figure 5.15 Products that change cluster both to hedonic for Customer 1.

These 7 products are considered hedonic since the price and brand label of the product is higher than this customer's regular pattern.

After deciding individual product segments, Bucket scores will be calculated based on the total sales value of each cluster.

$$Bucketscore_{mk} = \frac{\text{Total sales value of m labeled products in bucket k}}{\text{Total sales value of bucket k}}$$

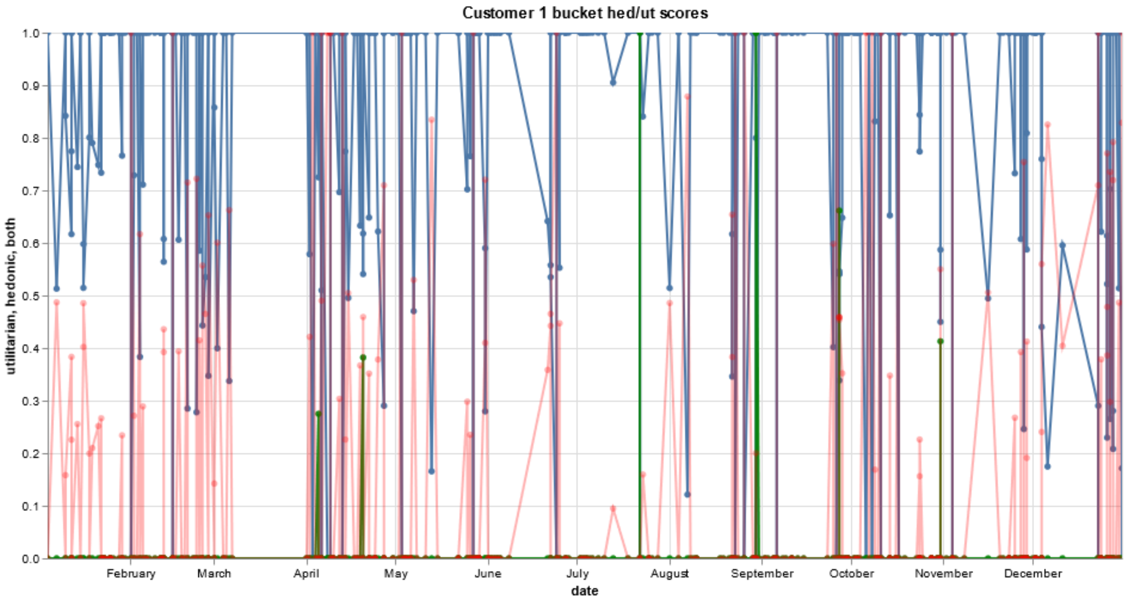


Figure 5.16 Bucketscore change over a year for customer1.

Blue lines represent the utilitarian scores of buckets. Red lines represent both labeled buckets. Finally, Green ones are for hedonic scores of buckets. For this customer, most of the buckets consist of utilitarian and both labeled scores. However, some buckets contain hedonic scores. These scores can be helpful to predict the next purchases that this customer makes.

In this study, a dashboard was created to see the differences after the proposed model was applied. In this dashboard, monthly repetition can be changed by the user to get a closer look and see the list of products in a given range of repetition. There is also a button to select before and after the model graphs. Furthermore, utilitarian, hedonic, and both bucket scores can be seen separately. Figure 5.17 is the screenshot of the dashboard.

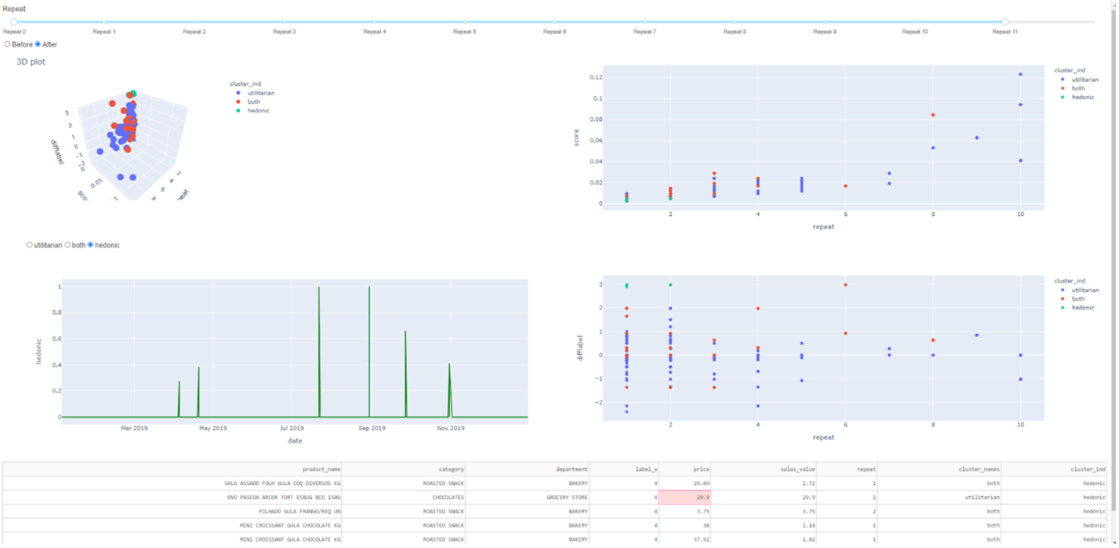
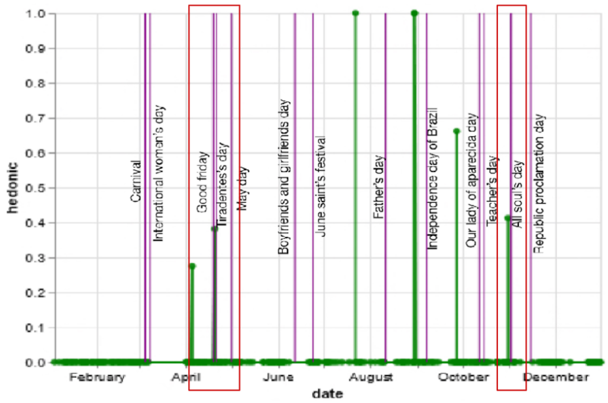


Figure 5.17 Dashboard screen for customer 1.

Table 5.2 shows the special day in Brazil. People can be affected by holidays to make a hedonic purchase. By using these dates and applying them to the hedonic bucket score graph, it can be easy to identify which days affect this customer. Figure 5.18 shows that 3 of these special days are important for this customer and he/she made hedonic purchases during these days. These days are: Good Friday, Tiradentes' Day and All Soul's Day.

Date	Name of the Day
3/5/2019	Carnival
3/8/2019	International Women's Day
4/19/2019	Good Friday
4/21/2019	Tiradentes' Day
5/1/2019	May Day
6/12/2019	Boyfriends and Girlfriends Day
6/24/2019	June Saint's Festivals
8/11/2019	Father's Day
9/7/2019	Independence Day of Brazil
10/12/2019	Our Lady of Aparecida Day
10/15/2019	Teacher's Day
11/2/2019	All Soul's Day
11/15/2019	Republic Proclamation Day

Table 5.2 Special days in Brazil



product_name	category	department	label_x	price	sales_value	repeat	cluster_names	cluster_ind
SALG ASSADO FOLH GULA COQ DIVERSOS KG	ROASTED SNACK	BAKERY	4	29.89	2.72	1	both	hedonic
OVO PASCOA ARCOR TORT ESBUG BCO 150G	CHOCOLATES	GROCERY STORE	4	29.9	29.9	1	utilitarian	hedonic
FOLHADO GULA FRANGO/REQ UN	ROASTED SNACK	BAKERY	4	3.75	3.75	2	both	hedonic
MINI CROISSANT GULA CHOCOLATE KG	ROASTED SNACK	BAKERY	4	38	1.14	1	both	hedonic
MINI CROISSANT GULA CHOCOLATE KG	ROASTED SNACK	BAKERY	4	37.92	1.82	1	both	hedonic
MINI PASTEL NATA GULA MACA KG	ROASTED SNACK	BAKERY	4	38	3.42	1	both	hedonic
FOLHADO GULA FRANGO/REQ UN	ROASTED SNACK	BAKERY	4	3.75	3.75	2	both	hedonic

Figure 5.18 Hedonic bucketscore changes over time for customer 1.

5.3 Future Works

In this thesis, the proposed model helps to identify the hedonic and utilitarian purchase patterns. This segmentation model can help marketing managers in many ways such as individual advertisement design and optimization, designing a marketing strategy for online and offline channels, allocating products based on the customer, and product segmentation in the stores.

In the next step, some ideas can be considered regarding the proposed model.

- The frequency of a customer's hedonic purchases can be compared with all population's frequency to detect common behaviors.
- The seasonality factor can apply to the model. For example, when a customer buys an ice cream in summer, it may not be considered hedonic. There is an obvious time-dependence in hedonic and utilitarian shopping value. Seasonal product scores can be compared and used to make an accurate decision.
- Statistical hypothesis testing can be used to examine the effects of discounts on hedonic and utilitarian shopping value.
- The proposed model can train with different parameters. Instead of buckets,

days can be used as sample intervals. The number of buckets can be changed during the sampling process.

- Lifestyle changes such as getting a new job or raising salary can be detected from time-series analysis.
- The distribution of hedonic and utilitarian segmented customers can be compared with the literature to justify the results.
- Accumulated time series values of utilitarian and hedonic bucket scores can be compared with Brazil's macroeconomic changes. During an economic crisis, people may tend to buy more utilitarian products for themselves. If the population's time series values overlap with economic situations, this can also be used to test the proposed model.
- A Retraining cycle can be created for companies to apply for a different time in a short interval.
- A visual analytic tool can be created for management's point of view to make it easy to make decisions.

6. CONCLUSION

The consumers' motivations to make the purchases is one of the topics that researchers from both management and psychology have been studied. Two shopping values are provided for a customer during shopping experiences: hedonic and utilitarian. These two shopping values are relevant and required to understand the reason why customers buy and to help companies to build their management strategies to increase profit and customer satisfaction.

In this thesis, supermarket transaction data from Brazil were analyzed for one year. As opposed to the literature, individual product, category, brand segmentation were applied. Instead of using a survey method, data transactions were used to identify customers' intentions. In the first step, brands were clustered into 10 segments. Some brands belong to the supermarket company. These brands were considered unbranded products. Secondly, for each customer, bucket and product-based features were calculated. By looking at the combinations in bucket features, products were clustered with the K-means algorithm. Since the dimension reduction methods: PCA and t-SNE visualizations proved that the clusters were well defined, labels of the product were decided and their personas were created. After the product segmentation, individual product features were used such as monthly repeat, product score, and label differences to cluster the product in an individual aspect. Finally, bucket scores were calculated and time-series or patterns of customers were represented in a line graph.

In this method, unsupervised learning with the K-means algorithm was applied. Since the is no ground truth to check if the resulting labels are correct or not is not possible. However, in future studies, a survey can be used to check the predicted scores and labels.

All the measures were created according to the literature hypothesis. Furthermore, the resulting outcomes support some papers. As Batra et al.(1991) stated that product categories, brands, and products can have both perspectives at the same time. Because one item can be hedonic for one person but utilitarian for another.

This study is conducted based on this information and create hedonic/utilitarian scores for product categories, department, and buckets.

In conclusion, this research might be helpful for the companies to develop better strategical methods to approach their target customer.

Bibliography

[1] Barry J. Babin, William R. Darden, and Mitch Griffin. Work and/or fun: Measuring hedonic and utilitarian shopping value. *Journal of Consumer Research*, 20(4):644, March 1994.

[2] Rajeev Batra and Olli Ahtola. Measuring the hedonic and utilitarian sources of consumer attitudes. *Marketing Letters*, 2, 04 1991.

[3] Rajeev Batra and Kevin Keller. Integrating marketing communications: New findings, new lessons and new ideas. *Journal of Marketing*, 80, 07 2016.

[4] Jason Carpenter. Consumer shopping value, satisfaction and loyalty in discount retailing. *Journal of Retailing and Consumer Services*, 15:358–363, 09 2008.

[5] Terry Childers, Christopher Carr, Joann Peck, and Stephen Carson. Hedonic and utilitarian motivations for online retail shopping behavior. *Journal of Retailing*, 77:511–535, 12 2001.

[6] Gilbert A. Churchill. A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research*, 16(1):64–73, 1979.

[7] William Darden and Fred Reynolds. Shopping orientations and product usage rates. *Journal of Marketing Research*, 8, 11 1971.

[8] William Dillon, Thomas Madden, Amna Kirmani, and Soumen Mukherjee. Understanding what?s in a brand rating: A model for assessing brand and attribute effects and their relationship to brand equity. *Journal of Marketing Research - J MARKET RES-CHICAGO*, 38:415–429, 11 2001.

[9] Ronald Faber and Thomas O'Guinn. A clinical screener for compulsive buying. *J Consum Res.*, 19:459–469, 01 1992.

[10] Morris B. Holbrook and Elizabeth C. Hirschman. The experiential aspects of consumption: Consumer fantasies, feelings, and fun. *Journal of Consumer Research*, 9(2):132, September 1982.

[11] Neda Irani and Kambiz Hanzaee. The effects of variety-seeking buying tendency and price sensitivity on utilitarian and hedonic value in apparel shopping satisfaction. *International Journal of Marketing Studies*, 3, 08 2011.

[12] Kai Kristensen, Anne Martensen, and Lars Grønholdt. Measuring the impact of buying behaviour on customer satisfaction. *Total Quality Management*, 10:602–614, 11 2009.

[13] Jingjing Li, Ahmed Abbasi, Amar Cheema, and Linda Abraham. Path to purpose? how online customer journeys differ for hedonic versus utilitarian purchases. *Journal of Marketing*, 84:002224292091162, 03 2020.

[14] Deborah Macinnis and Linda Price. The role of imagery in information processing: Review and extensions. *Journal of Consumer Research*, 13:473–491, 01 1987.

[15] J. MacQueen. Some methods for classification and analysis of multivariate observations. 1967.

[16] C. Park, Bernard Jaworski, and Deborah Macinnis. Strategic brand concept-image management. *Journal of Marketing*, 50:135, 10 1986.

[17] Yongming Qu, George Ostrouchov, Nagiza Samatova, and Al Geist. Principal component analysis for dimension reduction in massive distributed data sets. 04 2002.

[18] John Sherry. A sociocultural analysis of a midwestern american flea market. *Journal of Consumer Research*, 17:149–164, 01 1990.

[19] Vivek Singh, Burcin Bozkaya, and Alex Pentland. Money walks: Implicit mobility behavior and financial well-being. *PLOS ONE*, 10:e0136628, 08 2015.

[20] Laurens van der Maaten and Geoffrey Hinton. Vialuzing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 11 2008.

[21] Kevin Voss, Eric Spangenberg, and Bianca Grohmann. Measuring the hedonic and utilitarian dimensions of consumer attitude. *Journal of Marketing Research - J MARKET RES-CHICAGO*, 40:310–320, 08 2003.

[22] Robert Westbrook and William Black. A motivation-based shopper typology. *Journal of Retailing*, 61:78–103, 01 1985.