Analysis on E-commerce Order Cancellations Using Market Segmentation Approach

Jingyi Ye
University of California, San Diego, 9500 Gilman Dr, La Jolla 92092, CA, USA
jiy388@ucsd.edu

ABSTRACT

This study investigates the application of market segmentation on E-commerce canceled orders. It uses a transnational dataset that contains transactions of an online retail store during a year. The analysis process includes 1) an exploratory data analysis on the canceled orders which makes up a considerably amount of the dataset to show their characteristics. 2) a production segmentation that utilize the k-means clustering to create 5 product clusters. 3) a customer segmentation with k-means clustering using the production segments and customer features which results in 7 segments. In the process, the study compares silhouette scores and applies principal component analysis to optimize the number of clusters. The conclusion shows that market segmentation serves as an effective tool to distinguish products and consumers with different characteristics and help make suggestions to businesses. Also, including attitudinal features into the analysis process will result in improved customer profiles.

CCS CONCEPTS

• **Social and professional topics**; • **Professional topics**; • **Computing and business**; • **Computer supported cooperative work**;

KEYWORDS

Market segmentation, Electronic commerce, Cluster analysis, Case-based reasoning

ACM Reference Format:
Jingyi Ye. 2021. Analysis on E-commerce Order Cancellations Using Market Segmentation Approach. In *2021 The 4th International Conference on Computers in Management and Business (ICCMB 2021), January 30–February 01, 2021, Singapore, Singapore.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3450588.3450596

1 INTRODUCTION

From the rise of technology advancement and the improvement of internet environment, online shopping now plays an important role in people's daily life. The studies of E-commerce using data-driven or machine learning methods has emerged due to increasing needs to analyze this new pattern of customer behavior. Many of the related studies on E-commerce using aforementioned methods focused on logistics and supply chain management [1]. Though such studies provide reasoning on the supply side on the businesses, they did not provide analysis related to the demand side of the market. Market segmentation, however, serves as an effective way to analyze customer needs and behaviors in the business world. According to Wedel and Kamakura, "Since its introduction by Smith (1956), market segmentation has become a central concept in both marketing theory and practice" [2]. It also applies to multidiscipline including biking commuting market analysis [3], labor market [4], specialized industrial markets [5] and etc.

This study investigates the application of market segmentation on E-commerce. In particular, an exploratory data analysis and a K-means machine learning algorithm are performed on the canceled orders of a real E-commerce dataset. The study includes the following steps: (1) exploratory data analysis on the original dataset and generation of canceled orders; (2) segmentation of products using K-means clustering; (3) segmentation of customers using K-means clustering; and (4) analysis of the characteristics of the segments and a discussion of the application of such segmentation. The findings of this study help give advice to E-commerce businesses and also make suggestions to future analysis on data collecting.

2 LITERATURE REVIEW

Previous studies have shown that market segmentation has been effective to show distinct characteristics of different groups of customers [6, 7]. Such grouping helps businesses learn their customers and make decisions. For instance, a study by Beheshtian-Ardakani, Fathian, and Gholamian shows that customer loyalty helps divide customers into different segments and such segmentations promote direct marketing and help business recommend product bundles to customers [6]. Another study by Liu, Li and Peng use data from the biggest online shopping site in China, taobao.com, and perform customer segmentation using features including trust degree, interpersonal communication, favorite adding frequency and purchase duration [7]. They group the customers into six categories. Based on the sensitivity of each type of customers, they suggest different promotion strategies.

The above studies generally carry out market segmentation based on customer related features. However, they do not taken consideration the features of products themselves. Despite the various features of the customers, it is the product that they are purchasing. Therefore, it is important to include product features into the process of creating customer segments. A study by F. Daniel on E-commerce dataset integrates product segmentation into generating customer segments [8]. In particular, the author uses the product names and price range and performs k-means clustering to generate product segments. Then, author includes the segments

ICCMB 2021, January 30–February 01, 2021, Singapore, Singapore

ACM ISBN 978-1-4503-8861-0/21/01.
https://doi.org/10.1145/3450588.3450596

Table 1: Description of Variables 1

Variables	Description
Invoice No.	A six-digit number assigned to each transaction. It starts with a "C" if the transaction is canceled.
Stock code	A five-digit number assigned to each product.
Description	The name of a product.
Quantity	The amount purchased of a product during a transaction (negative if the order is canceled).
Invoice Date	The date and time of the transaction.
Unit Price (£)	The unit price of a product (sterling).
Customer ID	A five-digit number assigned to each customer.
Country	The country where the order is placed.
Total Price (£) *	The total amount of spending per order (negative if the order is canceled).

= *Unit Price*Quantity* generated for latter analysis.

Table 2: Comparison of Original and Canceled Dataset

	Products	Transactions	Customers
Original Dataset	3684	22190	4372
Cancellation Dataset	1920 (52.12%*)	3654 (16.47%*)	1589 (36.34%*)

* The percentage of Products, Transactions, and Customers of canceled orders comparing to the original dataset.

into the generation process of customer segments. Mainly inspired by Daniel's approach, this study applies this method. However, the aforementioned study does not analyze the canceled orders, which makes up 16.47% of the entire transactions. Therefore, this study focuses on discovering the use of market segmentation on the canceled orders with product features included.

3 METHODOLOGY

3.1 Data Source and Exploratory Data Analysis

3.1.1 Data source. Table 1 shows the description of variables using in the analysis process. This E-commerce data is originated from University of California, Irvine's Machine Learning Repository named Online Retail Data Set [9]. It is a transnational dataset that contains transactions occurs between January twelfth, 2010 and September 12, 2011from a UK-based and registered non-store online retail and the company mainly sells gifts for all occasions [9]. There are in total 541, 909 instances and 8 attributes in the dataset where each row shows information of both customer and product information of a single product from year 2010 to 2011.

3.1.2 Exploratory Data Analysis. Since the purpose of this study is to analyze the shopping cart abandonment, only the instances that indicate canceled orders is retained. There are 8872 out of 541, 909 instances containing information of such orders, which makes up 1.64 precent of the dataset. However, if measured in terms of number of products, number of transactions, or number of customers, the canceled orders make up a relative greater amount of all orders. Table 2 shows the comparison of the original dataset and the dataset containing only canceled orders.

It shows that 36.34% of the customers canceled their orders and more than 50% of the products are abandoned during the process. Also, nearly one fifth of the transactions are canceled. Therefore, it is necessary to analyze the characteristics of the customers who made the cancellations.

In addition to the 8 attributes, another feature, Total Price (Unit Price*Quantity), which measures the total amount paid for a particular transaction in sterling, is added. It is negative when Quantity is negative, meaning that the transaction is canceled. Table 3 shows the statistics of the continuous variables used in the analysis.

The pie chart shown in Figure 1 shows the distribution of the amount in each canceled order. It shows that majority of the cancellation amount is between £0 and £100, while only 10.7% of the transactions has a Total Price above £100. Another observation is that 58.6% of the cancellations have a Total Price less or equals to £20.

3.2 Product Segmentation

In order to add product features into customer segmentation in the next step, this study uses the method introduced in Daniel's notebook [8] which is to segment the products into different categories based on their product names and price range and make some adjustment.

The first step of product segmentation includes the choice of features and data encoding. The variable Description contains the product name for each product. Some examples are SET OF 3 COLOURED FLYING DUCKS, PLASTERS IN TIN CIRCUS PARADE, and VICTORIAN SEWING BOX LARGE. the keywords out of the product names are extracted using the Natural Language Toolkit, resulting in 198 unique words. Then, 13 words that are either not significant in distinguishing the products or repeated in every word category are removed. These words include pink, blue, tag, green, orange, vintage, sign, candle, pot, spot, retro, retrospot and box. Lastly, the keywords are encoded in a matrix X depicted in Table 4. Unit Price is then divided into 5 groups and added to the matrix.

Table 3: Statistics of Continuous Variables 1

Variables	Number of Observations	Mean	Standard Deviation	Minimum	Maximum
Quantity	8872.00	-30.77	1172.25	-80995.00	-1.00
Unit Price (£)	8872.00	18.90	445.19	0.01	38970.00
Total Price (£)	8872.00	-68.61	-2022.87	-0.12	-168469.60

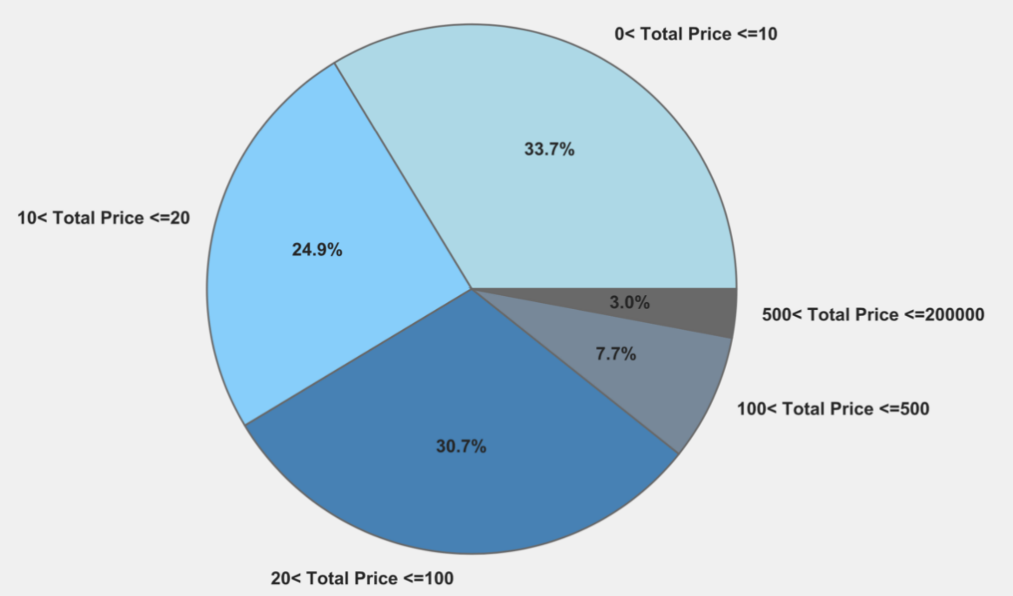

Figure 1: Distribution of Canceled Amount (£) per Order

Table 4: Matrix of Product Names

	Keyword 1	Keyword 2	. . .	Keyword m
Product 1	a1,1	a1,1	. . .	a1, m
Product 2	a2,1	a2,2	. . .	a2, m
. . .	. . .	. . .	. . .	. . .
Product n	an,1	an,2	. . .	an,m*

* an,m: m indicates which product and m indicates the keyword. an,m = 1 if product n has keyword m in its name, 0 otherwise.

The second step is to generate product segments by k-means clustering. K-means clustering is one of the most widespread algorithms, especially in marketing research [10] and it is useful when searching for a nearly optimal partition with a fixed number of clusters [11]. Since it is an unsupervised machine learning method, the optimal number of clusters are chosen based on silhouette score ($-1 \le s_i \le 1$), which measures the closeness of each point of a cluster to its center comparing to the next nearest cluster, formulated as follows:

$$s_i = \frac{b_i - a_i}{\max(b_i,\ a_i)} \tag{1}$$

a_i measures the closeness to its own cluster and b_i measures the average distance to the next nearest cluster. The number of clusters chosen is 5 since the silhouette score increases when the number of clusters is less than 5 and decreases afterwards. The average silhouette score of 5 clusters is 0.17. Figure 2 shows the scores of each datapoint in the 5 clusters. The clustering result is relatively robust since there are no negative numbers.

Note: The dash line shows the average silhouette score at 0.17.

Lastly, a word cloud shown in Figure 3 visualizes the product segments. The larger the word, the more frequent it occurs in the cluster.

Cluster 1 includes mainly Christmas gifts. The words christmas, card, wrap, and pack stand out the most.

Cluster 2 includes mainly cooking gifts. The words bottle, water, pan, tin, holder, cream and light stand out the most.

Cluster 3 includes mainly design gifts. The words design, paper, bag, glass and light stand out the most.

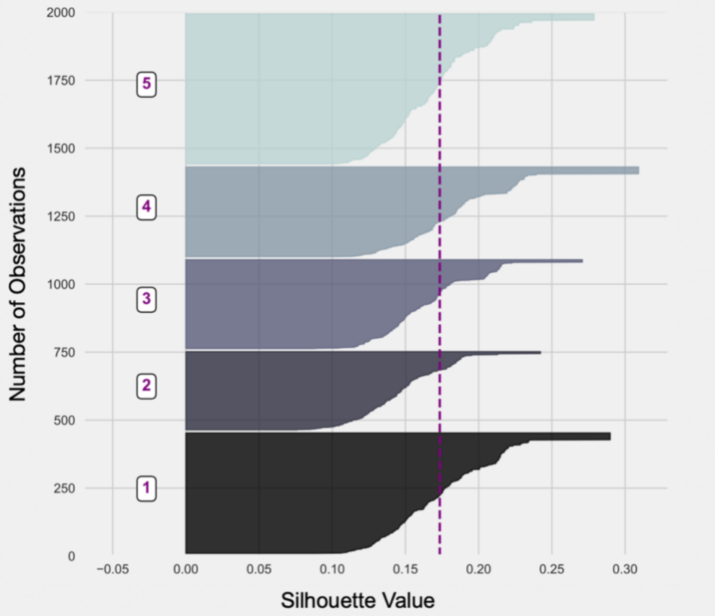

Figure 2: Silhouette Scores of 5 Product Segments

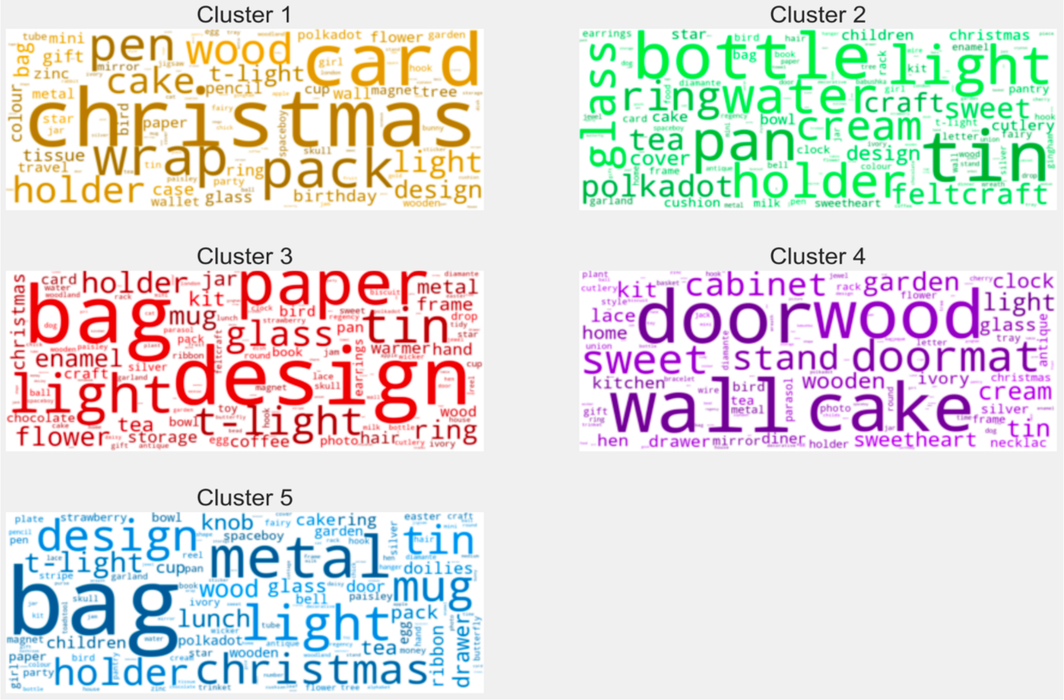

Figure 3: Word Cloud of Product Clusters

Cluster 4 includes mainly home decoration gifts. The words door, wood, doormat, wall, cake and sweet stand out the most.

Cluster 5 includes mainly random gifts. The words bag, metal, light, design and christmas stand out the most.

From the word cloud above, it can conclude that k-means clustering does a relatively good job of making product segments based on product names and price range. But one problem with the algorithm is that it does not do well with overlapping clusters [11]. The

Table 5: Description of Variables 2

Variables	Description
count	The number of cancellations each customer makes
min (£)	The minimum amount of the cancellations each customer spends in sterling.
max (£)	The maximum amount of the cancellations each customer spends in sterling.
median (£)	The median amount of the cancellations each customer spends in sterling.
sum (£)	The total amount of the cancellations each customer spends in sterling.
c1 (%)	The percent of products that belongs to product segment 1 for each customer.
c2 (%)	The percent of products that belongs to product segment 2 for each customer.
c3 (%)	The percent of products that belongs to product segment 3 for each customer.
c4 (%)	The percent of products that belongs to product segment 4 for each customer.
c5 (%)	The percent of products that belongs to product segment 5 for each customer.

Table 6: Statistics of Continuous Variables 2

Variables	Number of Observations	Mean	Standard Deviation	Minimum	Maximum
count	1589	2.30	3.03	1.00	47.00
min (£)	1589	215.08	4685.21	0.39	168469.60
max (£)	1589	312.31	4802.93	0.42	168469.60
median (£)	1589	246.33	4712.56	0.42	168469.60
sum (£)	1589	383.06	4880.13	0.42	168469.60
c1 (%)	1589	8.35	22.56	0.00	100
c2 (%)	1589	39.23	40.92	0.00	100
c3 (%)	1589	18.63	31.24	0.00	100
c4 (%)	1589	14.21	27.23	0.00	100
c5 (%)	1589	19.58	31.70	0.00	100

word light appears frequently in every cluster. Also, Cluster 5 do not have a strong inclination to a specific category of gifts.

3.3 Customer Segmentation

The final phrase of this study is to use K-means clustering to generate customer segmentation. There are 2 steps in total and the resulting clusters will be discussed in Results.

The first step includes the choice of features and data encoding. Rows in the dataset are transformed so that each row contains information of one customer. The resulting data frame contains 1598 rows. After reorganizing the data, the 5 categories of products are encoded into the data frame by assigning a percentage to each category based on the total amount of cancelation of each customer. For example, customer A canceled an order containing 20 dollars on each category of products, then the percentage for each category will be 20%. Table 5 shows the description of variables and Table 6 shows the statistics.

The variable median (£), which represents the median amount of spending of each customer, is used instead of the mean. This is because it represents the middle number of a set of data and large outliers do not affect median as much as they affect mean. For this dataset, the distribution of spending is skewed to the left (Figure 4). Therefore, a median is a better representative of the distribution of datapoint comparing to the mean.

The second step is to generate product segments by k-means clustering. the number of features to include in the algorithm is selected based on principle component analysis. It allows me to do dimensionality reduction if less than 9 features are enough to explain the variation of the dataset. Figure 5 and Figure 6 shows the result: 7 components are enough to explain the variation of customers cancellations. The features retained are count, median, and c1 to c5.

After finalizing the features, k-means clustering is performed, and the number of clusters are chosen based on the silhouette scores. The optimal number of clusters is 7 in this case since the score increases when the number of clusters is less than 7 and decreases afterwards. Figure 7 shows the distribution of silhouette scores for the points in each cluster. The clustering result is relatively robust since there is only few negative numbers in cluster 3, 5, 6 and 7. The average silhouette score is 0.56, which is much higher than that of the product clusters.

Note: The dash line shows the average silhouette score at 0.56.

4 RESULTS

Figure 8 displays the distribution of the number of canceled orders. It shows that the majority of cancellations are between 1and 2 which indicates that 75.4% of the customers only make cancellation(s) less or equal to 2 times.

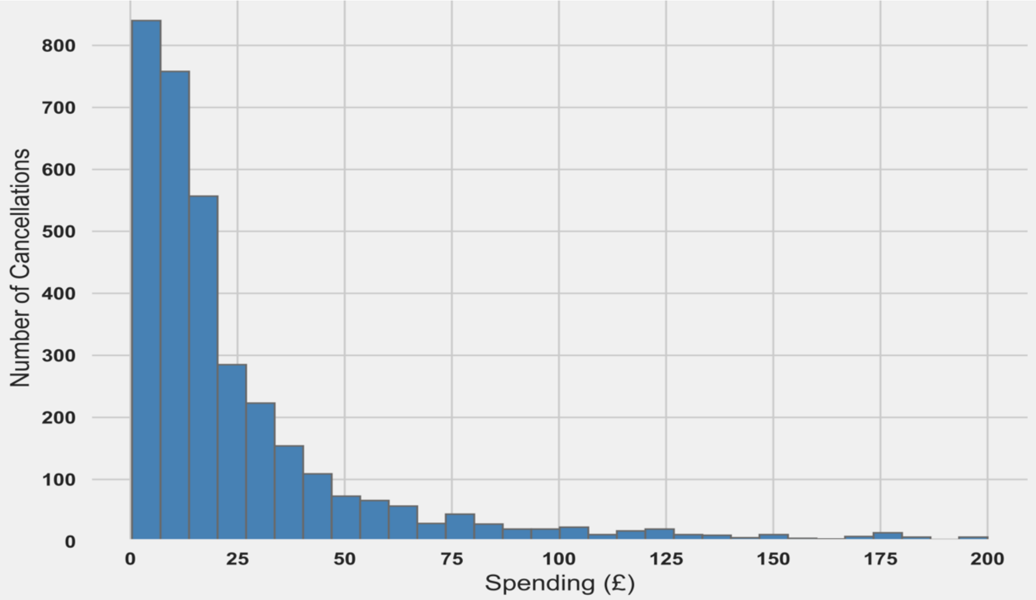

Figure 4: Distribution of Basket Price below £200

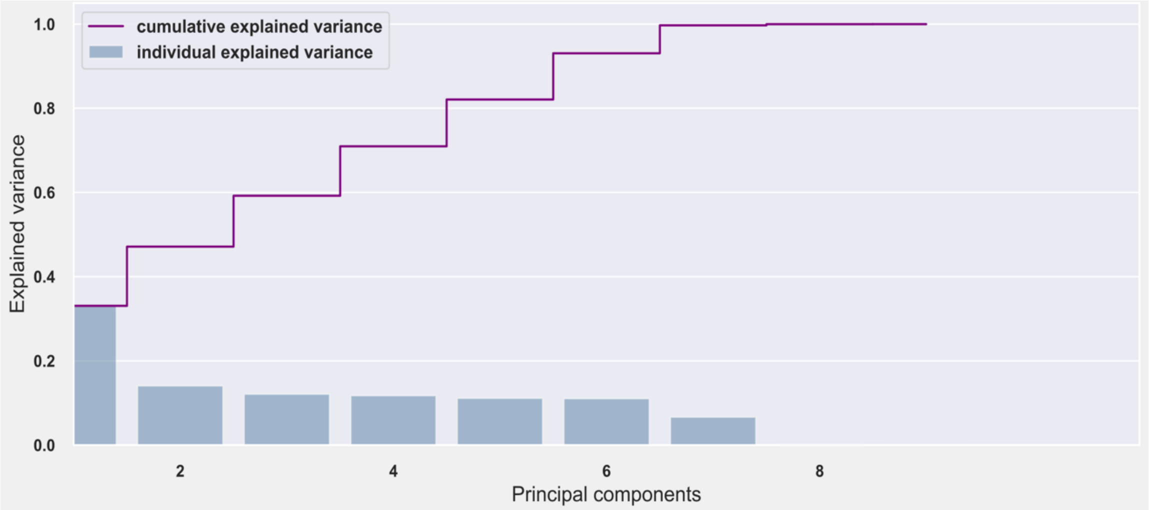

Figure 5: PCA of Customer Clustering 1

Table 7 shows the result of customer segmentation. The features within each segment are summarized as follows:

Segment1 (S1) is a group of customers who cancel the most home decoration related orders and the number of orders they cancel is close to 2.

Segment2 (S2) is a group of customers who cancel the most design related orders and the number of orders they cancel is above to 2.

Segment3 (S3) is a group of customers who cancel the most random products and the number of orders they cancel is close to 2.

Segment4 (S4) is a group of customers cancel the most Christmas related orders and the number of orders they cancel is close to 2.

Segment5 (S5) is a group of outlier customers who cancel one cooking related order and one random product order.

Segment6 (S6) is a group of customers who cancel the most cooking related products and the number of orders they cancel is close to 2.

Segment7 (S7) is a group of customers where nearly half of them canceled design related products and the number of orders they cancel is the most on average, which is above 17.

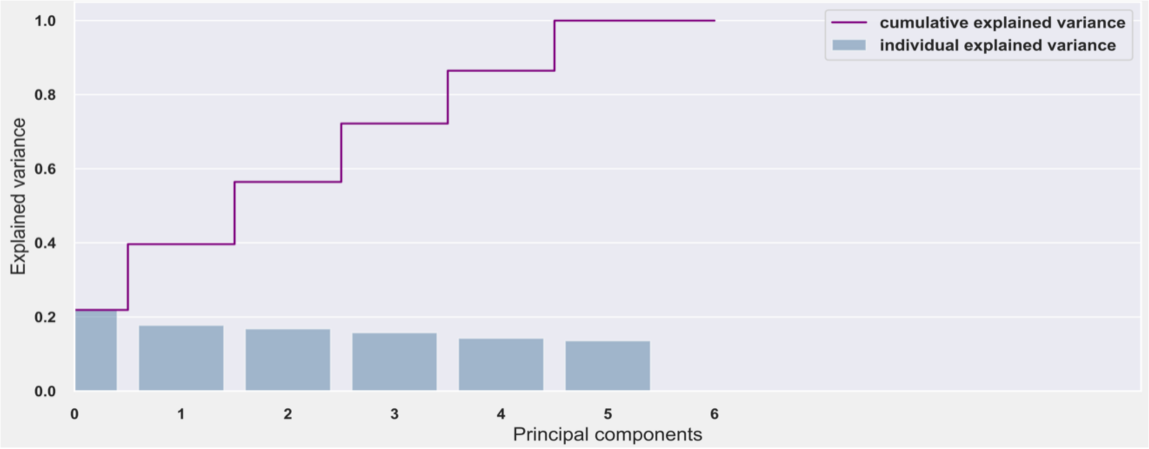

Figure 6: PCA of Customer Clustering 2

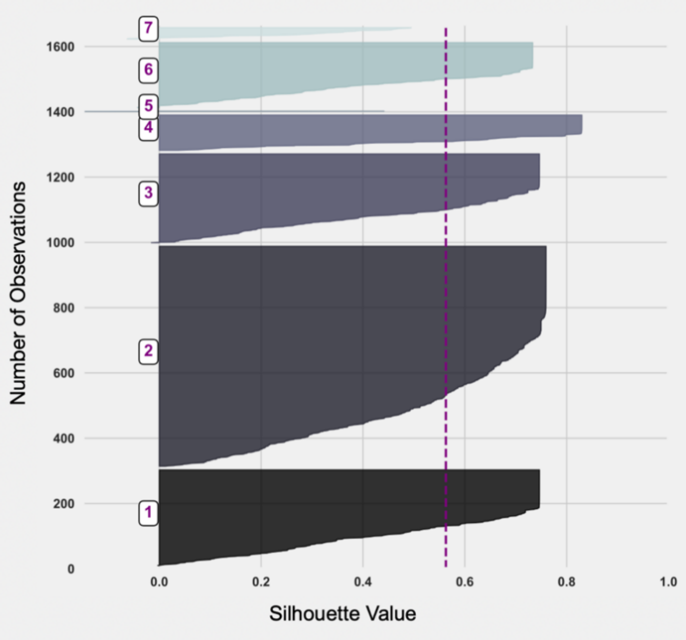

Figure 7: Silhouette Scores of 7 Customer Segments

The above segments show distinct characteristics of customer groups. Segment 1, 2, 3, 4 and 6 each has cancellation majored in one segment while segment 5 and 7 has more. The reason from this difference can be attributed to cluster size and outliers. Segment 2 has only 2 customers but results in the highest median spending. Segment 5, on the other hand, has a relatively even distribution of cancellations. Both segments have the lowest number of samples comparing to the others. In addition, design related projects (C3) is cancelled the most in two segments, 2 and 7. Both segments have the number of cancellations above 2, which makes only 25% of the entire customers (Figure 8). It indicates that people are more likely to cancel orders containing design related goods. Lastly, segment 5 and 6 result contain the highest two cancellation amount. One order in segment 5 is cooking related product and in segment6, the most cancellation are also cooking goods. It is therefore necessary for the business to track these particular orders and understand the reasons behind them.

5 CONCLUSION

This study demonstrates the application of market segmentation on cancelled orders of a real E-commerce dataset. Firstly, an exploratory data analysis is applied to analyze the canceled order. Then, k-means clustering is used to create production segments. Lastly, the process results in 7 customer segments by k-means clustering with customer features and the product segments. The

Table 7: Customer Segmentation results

Features	Thresholds	Segments						
		S1 (Size = 294)	S2 (Size = 675)	S3 (Size = 273)	S4 (Size = 109)	S5 (Size = 2)	S6 (Size = 200)	S7 (Size = 36)
Count	1 ≤. ≤ 2	1.83		1.97	1.41	1.00	1.82	
	2 <. ≤ 30		2.14					17.03
Median (£)	0 <. ≤ 100	44.90		59.80	50.08			31.83
	100 <.		115.22			122865.19	159.25	
C1 (%)	0 <. ≤ 50	2.81	1.84	4.58		0.00	2.68	3.86
	50 <. ≤ 100				85.01			
C2 (%)	0 <. ≤ 50	4.87	4.97	4.61	3.14	50.00		15.89
	50 <. ≤ 100						77.69	
C3 (%)	0 <. ≤ 50	7.23		5.64	3.65	0.00	5.47	47.47
	50 <. ≤ 100		82.30					
C4 (%)	0 <. ≤ 50		6.30	4.79	2.40	0.00	6.37	17.14
	50 <. ≤ 100	79.22						
C5 (%)	0 <. ≤ 50	5.87	4.59		5.80	50.00	7.79	15.64
	50 <. ≤ 100			80.38				

Note: grey cell indicates the threshold each cluster center belongs to. The numbers within are numeric value of the centers.

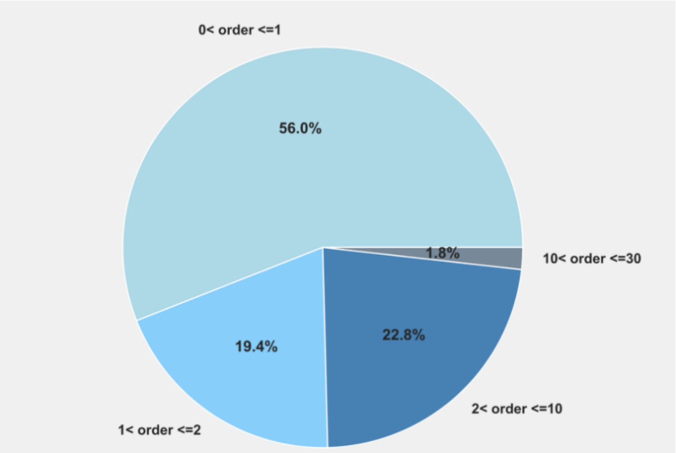

Figure 8: Distribution of Numbers of Canceled Order(s)

resulting segments demonstrate different characteristics and help make suggestions to businesses.

There are some limitations to this study. Though the final segmentation result shows the variation among different customer groups, it is not sound enough for businesses to draw complete consumer profiles. Other descriptive features, such as gender and age should be collected before the analysis process. Also, attitudinal factors are recommended. The business could send out surveys asking attitudinal questions such as how often you cancel an order, or do you cancel an order due to the price or dissatisfaction with the service. Therefore, a combination of the attitudinal, product-related and customer-related factors in the analysis should result in a more complete customer profile. The author recommends future studies using this approach.

ACKNOWLEDGMENTS

The author would like to thank Professor S. Coggeshall from the University of Southern California for his guidance through the process of this research and F. Daniel, the author of Customer Segmentation notebook on Kaggle for the inspiration.

REFERENCES

[1] Xu, G., Qiu, X., Fang, M., Kou, X., & Yu, Y. (2019), Data-driven operational risk analysis in E-Commerce Logistics, Advanced Engineering Informatics, Volume 40, Pages 29-35, ISSN 1474-0346, https://doi.org/10.1016/j.aei.2019.03.001.

[2] Wedel M. & Kamakura W. (2000). Market Segmentation: Conceptual and Methodological Foundations. New York: Springer Science+Business Median, LLC. https://books.google.com.hk/books?id=XxLaBwAAQBAJ&lpg=PA4&ots=LZwVC3BSnT&dq=market%20segmentation%20scholarly%20articles&lr&hl=zh-CN&pg=PP5#v=onepage&q&f=false.

[3] Li, Z., Wang, W., Yang, C., & Ragland, D. R. (2012), Bicycle commuting market analysis using attitudinal market segmentation approach, Transportation Research Part A: Policy and Practice, Volume 47, Pages 56-68, ISSN 0965-8564, https://doi.org/10.1016/j.tra.2012.10.017.

[4] Reich, M., Gordon, D., & Edwards, R. (1973). A Theory of Labor Market Segmentation. The American Economic Review, 63(2), 359-365. Retrieved October 6, 2020, from http://www.jstor.org/stable/1817097

[5] Doyle, P., & Saunders, J. (1985). Market Segmentation and Positioning in Specialized Industrial Markets. Journal of Marketing, 49(2), 24–32. https://doi.org/10.1177/002224298504900202

[6] Beheshtian-Ardakani, A., Fathian, M & Gholamian, M. (2018). A novel model for product bundling and direct marketing in e-commerce based on market segmentation. Decision Science Letters, 7(1), 39-54.http://doi.org/10.5267/j.dsl.2017.4.005

[7] Liu, Y., Li, H., Peng, G. et al. (2015). Online purchaser segmentation and promotion strategy selection: evidence from Chinese E-commerce market. Ann Oper Res 233, 263–279. https://doi.org/10.1007/s10479-013-1443-z

[8] Daniel, F. (2017). Customer Segmentation. Kaggle Notebooks. https://www.kaggle.com/fabiendaniel/customer-segmentation

[9] UCI Machine Learning Repository. (2015). Online Retail Data Set. https://archive.ics.uci.edu/ml/datasets/Online+Retail

[10] Green, P.E., Krieger, A.M. (1995). Alternative approaches to cluster-based market segmentation. Journal of the Market Research Society, 3 (1995), 221-239. https://doi.org/10.1177/147078539503700302

[11] Kim, K., Ahn, H. (2008). A recommender system using GA K-means clustering in an online shopping market. Expert Systems with Applications. Volume 34, Issue 2. Pages 1200-1209. ISSN 0957-4174. https://doi.org/10.1016/j.eswa.2006.12.025