*sustainability*

MDPI

*Article*

# K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data

**Kayalvily Tabianan [1,*], Shubashini Velu [2] and Vinayakumar Ravi [3]**

[1] Faculty of Information Technology, Inti International University, Persiaran Perdana BBN Putra Nilai, Nilai 71800, Malaysia
[2] MIS Department, College of Business Faculty, Prince Mohammad bin Fahd University, Khobar 34754, Saudi Arabia; svelu@pmu.edu.sa
[3] Center for Artificial Intelligence, Prince Mohammad Bin Fahd University, Khobar 34754, Saudi Arabia; vravi@pmu.edu.sa
* Correspondence: kayalvily.tabianan@newinti.edu.my

**Abstract:** E-commerce system has become more popular and implemented in almost all business areas. E-commerce system is a platform for marketing and promoting the products to customer through online. Customer segmentation is known as a process of dividing the customers into groups which shares similar characteristics. The purpose of customer segmentation is to determine how to deal with customers in each category in order to increase the profit of each customer to the business. Segmenting the customers assist business to identify their profitable customer to satisfy their needs by optimizing the services and products. Therefore, customer segmentation helps E-commerce system to promote the right product to the right customer with the intention to increase profits. There are few types of customer segmentation factors which are demographic psychographic, behavioral, and geographic. In this study , customer behavioral factor has been focused. Therefore users will be analyzed using clustering algorithm in determining the purchase behavior of E-commerce system. The aim of clustering is to optimize the experimental similarity within the cluster and to maximize the dissimilarity in between clusters. In this study there are relationship between three clusters: event type, products, and categories. In this research, the proposed approach analyzed the groups that share similar criteria to help vendors to identify and focus on the high profitable segment to the least profitable segment. This type of analysis can play important role in improving the business. Grouping their customer according to their similar behavioral factor to sustain their customer for long-term and increase their business profit. It also enables high exposure of the e-offer to gain attention of potential customers. In order to process the collected data and segment the customers, an learning algorithm is used which is known as K-Means clustering. K-Means clustering is implemented to solve the clustering problems.

**Keywords:** customer segmentation; purchase behavior; K-Means clustering; profitable customer; clusters



**Citation:** Tabianan, K.; Velu, S.; Ravi, V. K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data. *Sustainability* **2022**, *14*, 7243. https://doi.org/10.3390/su14127243

Academic Editors: Abu ul Hassan Sarwar Rana and Muhammad Fazal Ijaz

Received: 3 May 2022
Accepted: 27 May 2022
Published: 13 June 2022



## 1. Introduction

In this global pandemic known as COVID-19, new terms were created such as Work from Home (WFH) and Study from Home (SFH) that requires people to stay at home and minimize outdoor activities including shopping. Large supermarkets were also opening E-commerce system to sustain their profit during this pandemic [1]. People started using online shopping website to purchase necessary items which is very convenient in this current situation.

E-commerce system has become more popular and implemented in almost all business areas. E-commerce system is a platform for marketing and promoting the products to customer through online [2]. Customer segmentation is known as dividing the customers into groups which shares similar characteristics. The purpose of customer segmentation

is to determine how to deal with customers in each category to increase the profit of each customer to the business.

When customers receive too much information or unwanted details which is not related to their regular purchase or their interest on the products, it can cause confusion on deciding their needs. This might lead their customers to give up on purchasing the items they required and effect the business to lose their potential customers. The clustering analysis will help to categorize the E-commerce customer according to their spending habit, purchase habit or specific product or brand the customers interested in. In order to process the collected data and segment the customers, an unsupervised learning algorithm is used which is known as Kmeans clustering is used [3].

Online shopping is not anymore, a new, whereas most of the business are becoming online based. There are a number of online shopping platforms keep on increasing day by day. Since most traditional business started to implement E-commerce system in their business and E-commerce system has become trending, there are more competition in the field [4]. In order for a business to sustain for a longer term and be competitive, the business should know the ways to retain their customers. For an example, if an E-commerce system continuously display a customer the products that is expensive or above their shopping budget, then the customers may decide that this E-commerce system is not suitable for them. Therefore they may look for another online shopping platforms which usually leads to high churn in E commerce platform .Customers differ in personality and have various preferences. There is evidence that inequalities in marketing exists. As a result, having the same approach and marketing for every consumer is not effective [5] and those consumers may be the most vital to the business [6]. Therefore it can be stated that inequalities in marketing and inefficiencies may cause customer churn . It is critical for a company to segment its consumers and determine the distinctions between the customer segments. For an example, the prices in the retail business in Malaysia will be increase by 50–60 percent from the 1 June 2022 onwards hence one way to keep existing customers is customer segmentation. Market segmenting according to the customer purchase behavior is important to decide the likelihood of the customer buying a specific product [7]. In this study it explains how a business can run for longer term by understanding their customer need and interest and satisfy them. The aim of this research is to conduct customer segmentation using the customer data and grouping their customers into groups that share similar criteria. Customer segmentation is carried out to find the potential and most profitable customer groups among the total customers [8]. Therefore, this helps to reduce the risk of losing the customer by selling the wrong product to the wrong customer group. Customer segmentation shows the way for E-commerce on how to make their business customer-focused and conquer a stable position in the business world.

Rachmawati et al. [9] proclaim that for the large and sophisticated data information of today's E-commerce businesses, accurate and efficient customer segmentation management should be carried out. In this competitive and developing E-commerce business, it is important to analyse the customer need and apply market segmentation mine and analyse various target customers in the system to provide different customers with distinctive marketing methods and improve their customer loyalty and satisfaction. According to Shirole et al. [10], customer segmentation is based on discovering important differentiators that split customers into target groups. A customer segmentation model allows organizations to target specific groups of customers, allowing for more effective marketing resource allocation and the maximization of cross- and up-selling capability. Customer segmentation can also help to enhance customer service and increase customer loyalty and retention. There are several aspects of online shopping behavior can be found that can influence the strategic approach in E-commerce for longer term which are the security of seller and buyer E-commerce, optimize the re-order and up-to inventory levels of e-groceries, payment method, delivery method of the item, delivery speed of the items, the user interface of E-commerce system, wide range of products choice and reasonable price for the service and products [11]. The aim of this study is to conduct an effective segmentation of E-commerce

customer. The customer segmentation proposed for this research is applying K-Means clustering algorithm. The objectives and contributions of this research are:

- To analyse the connection between customer purchase behaviour and the customer segmentation. The purpose of this study is to analyse the connection between the purchase behaviour of E-commerce platform customer and the customer segmentation.
- To comprehend how emerging technologies enable marketers to better meet the requirements and desires of consumers. This project also comprehends how the emerging technologies enable marketers to better meet the requirements and desires of consumers where technology has the ability to influence and change customer behaviour.
- To understand the consumer behaviour and focus on high profitable segment Other objective of this project is to obtain understanding on the consumer behaviour and focus on high profitable segment. Customer segmentation objective is to enhance sales by providing customized experience tailored to each segment.
- Proposal of SAPK + K-Means-based clustering approach for customer segmentation using Malaysia's E-commerce customer purchase behavior dataset.
- A detailed experimental analysis of SAPK + K-Means and other clustering approaches such as K-Means and AP + K-Means for customer segmentation.
- Data visualization and Dashboards for visualizing the results of customer segmentation to the end users.
- Various data analysis techniques were shown to understand the dataset characteristics of customer purchase behavior.

The rest of the sections are organized as follows. Section 2 includes methods, Section 3 includes K-Means clustering, Section 4 includes Data Mining, Section 5 includes Data Visualization and Dashboard, Section 6 includes Discussion, and Conclusion is included in Section 7.

## 2. Methods

The dataset used for this research is Malaysia's E-commerce dataset from MDEC repository for machine learning based on customers purchasing behavioral. There in this project the dataset contains the E-Commerce behavior data from multi category store which comprises 285,000,000 customer purchase history. Since the dataset used in this research contains millions of data, which might lead to inaccurate result or output. Therefore, selected number of datasets from the last 3 years is be used to conduct the analysis. The present paper explains the process of carrying out clustering in the Malaysian E-Commerce industry. The clustering must allow similar categories, product type and events type such as view, product in cart and purchased, to be clustered for the purpose of constructing a representative sample of the different categories. This sample must enable the generation of a set of indicators that present the inequalities that exist in the Figure 1. Furthermore, its design must revolve around the three major target groups of inequality: age, gender, working adults or students age. A sample of dataset is shown in Figure 1.

| event_date | event_time | event_type | product_id | category_id | category_code | product_detail | product_detail 1 | brand | price | user_id | user_session |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1/10/2019 | 0:00:00 | view | 44600062 | 2.10 x $10^{18}$ | | | | shiseido | 35.79 | 541312140 | 72d76fde-8bb3-4e00-8c23-a032dfed738c |
| 1/10/2019 | 0:00:00 | view | 3900821 | 2.05 x $10^{18}$ | appliances | environment | water_heater | aqua | 33.2 | 554748717 | 9333dfbd-b87a-4708-9857-6336556b0fcc |
| 1/10/2019 | 0:00:01 | view | 17200506 | 2.05 x $10^{18}$ | furniture | living_room | sofa | | 543.1 | 519107250 | 566511c2-e2e3-422b-b695-cf8e6e792ca8 |
| 1/10/2019 | 0:00:01 | view | 1307067 | 2.05 x $10^{18}$ | computers | notebook | | lenovo | 251.74 | 550050854 | 7c90fc70-0e80-4590-96f3-13c02c18c713 |
| 1/10/2019 | 0:00:04 | view | 1004237 | 2.05 x $10^{18}$ | electronics | smartphone | | apple | 1081.98 | 535871217 | c6bd7419-2748-4c56-95b4-8cec9ff8b80d |
| 1/10/2019 | 0:00:05 | view | 1480613 | 2.05 x $10^{18}$ | computers | desktop | | pulser | 908.62 | 512742880 | 0d0d91c2-c9c2-4e81-90a5-86594dec0db9 |
| 1/10/2019 | 0:00:08 | view | 17300353 | 2.05 x $10^{18}$ | | | | creed | 380.96 | 555447699 | 4fe811e9-91de-46da-90c3-bbd87ed3a65d |
| 1/10/2019 | 0:00:08 | view | 31500053 | 2.05 x $10^{18}$ | | | | luminarc | 41.16 | 550978835 | 6280d577-25c8-4147-99a7-abc6048498d6 |
| 1/10/2019 | 0:00:10 | view | 28719074 | 2.05 x $10^{18}$ | apparel | shoes | keds | baden | 102.71 | 520571932 | ac1cd4e5-a3ce-4224-a2d7-ff660a105880 |
| 1/10/2019 | 0:00:11 | view | 1004545 | 2.05 x $10^{18}$ | electronics | smartphone | | huawei | 566.01 | 537918940 | 406c46ed-90a4-4787-a43b-59a410c1a5fb |
| 1/10/2019 | 0:00:11 | view | 2900536 | 2.05 x $10^{18}$ | appliances | kitchen | microwave | elenberg | 51.46 | 555158050 | b5bdd0b3-4ca2-4c55-939e-9ce44bb50abd |
| 1/10/2019 | 0:00:11 | view | 1005011 | 2.05 x $10^{18}$ | electronics | smartphone | | samsung | 900.64 | 530282093 | 50a293fb-5940-41b2-baf3-17af0e812101 |
| 1/10/2019 | 0:00:13 | view | 3900746 | 2.05 x $10^{18}$ | appliances | environment | water_heater | haier | 102.38 | 555444559 | 98b88fa0-d8fa-4b9d-8a71-3dd403afab85 |
| 1/10/2019 | 0:00:15 | view | 44600062 | 2.10 x $10^{18}$ | | | | shiseido | 35.79 | 541312140 | 72d76fde-8bb3-4e00-8c23-a032dfed738c |
| 1/10/2019 | 0:00:16 | view | 13500240 | 2.05 x $10^{18}$ | furniture | bedroom | bed | brw | 93.18 | 555446365 | 7f0062d8-ead0-4e0a-96f6-43a0b79a2fc4 |
| 1/10/2019 | 0:00:17 | view | 23100006 | 2.05 x $10^{18}$ | | | | | 357.79 | 513642368 | 17566c27-0a8f-4506-9f30-c6a2ccbf583b |
| 1/10/2019 | 0:00:18 | view | 1801995 | 2.05 x $10^{18}$ | electronics | video | tv | haier | 193.03 | 537192226 | e3151795-c355-4efa-acf6-e1fe1bebeee5 |
| 1/10/2019 | 0:00:18 | view | 10900029 | 2.05 x $10^{18}$ | appliances | kitchen | mixer | bosch | 58.95 | 519528062 | 901b9e3c-3f8f-4147-a442-c25d5c5ed332 |

**Figure 1.** Multi category Online store dataset.

Another popular method for obtaining quantitative data is through questionnaire or survey. Therefore, for quantitative research questionnaire will be carried out in this project. The questionnaire will be sent through social media platform such as Facebook, Instagram, and WhatsApp to the targeted group of people. The response of the questionnaire will be examined, and the analysis will be created based on the collected quantitative data. The questions generated for the questionnaire are known as the foundation of the research as it will result a statistical analysis using the collected data.

For the data set chosen in this research pre-processing techniques will be done to to check missing values, noisy data, and other inconsistencies before executing it to the algorithm. RStudio and Microsoft Office Excel are the tools that will be used to perform the data pre-processing and using K-Means algorithm . KMeans clustering is known as an unsupervised learning which is used to solve problems related to clustering [12]. K-Means clustering is a process of classifying the dataset into certain number of clusters where each cluster will be defined with k centers. The k centres should be strategically placed since various locations produce various results. The result will be better if each cluster is as far as possible. The ideal number of clusters k that leads to the maximum distance which can be calculated from the dataset. One of the ways to choose the optimum number of clusters is elbow method. A practical technique would be to compare the results of numerous runs with multiple k and select the best one based on a predetermined criterion. In general, a high k reduces error but raises the likelihood of overfitting.

The target audience of this analysis are users from 18 to 60 years old. Users from 18 to 23 years old are most likely to be students who uses E-commerce system to purchase electronic devices, accessories, and clothing. User from 24 years and above mostly likely to be working person who gets pay and their purchase budget might be higher than student groups. It is because working people holds credit card or debit cards to make their payments and prefer that the items to be delivered to their home rather going shopping after their work schedule. This research could benefit the target user by promoting the product categories according to the right target group. For example, for people who are working, the items they will prefer are kitchen appliance, furniture, and electronic appliances. Meanwhile, students would not be interested in those items.

## 3. K-Means Clustering

Once the variables were selected, a clustering process was carried out to detect the product popularity based on target customer group. In this study SAPK + Kmeans clustering algorithm is used because improvised algorithm, the square error result will be smaller compared to Affinity Propagation Algorithm (AP). AP Algorithm will not provide the number of cluster or the cluster center, but then, it will consider all the samples as the exemplar which known as potential cluster center. This improved algorithm is always searching for the optimal cluster center value and maximizing the objective function value during the implementation process [13]. Explained by Deng et al. [14] is SAPK + K-Means which is the combination AP algorithm and K-Means algorithm. Since SAPK + Kmeans is an The SAPK + K-Means algorithm only requires one iteration to complete the operation, resulting in better clustering effect for the number of elements and satisfactory outcomes [14]. Figure 2 explains the process of SAPK + K-Means algorithm.

In addition, there are a few steps in applying SAPK + K-Means algorithm in customer segmentation, as shown in Figure 3. The first step is obtaining data regarding E-commerce purchase and check if the data obtained has clustering trend. If there is a trend, then conduct clustering, else end the clustering. The next step is to apply the SAPK + K-Means algorithm to the dataset and divide the dataset into several clusters such as c1, c2, c3...cn. Then, summary each class into one or more rules according to the characteristics of each class based on the data object characteristics in the class and analyze the clustering outcomes. If the clustering result is extremely dependable, it is confirmed for the actual application. If not, the clustering analysis is repeated using different clustering techniques. As conclusion,

SAPK + K-Means algorithm helps to increase the quality of customer data clustering and increase the effectiveness of E-commerce activities enterprises [15].

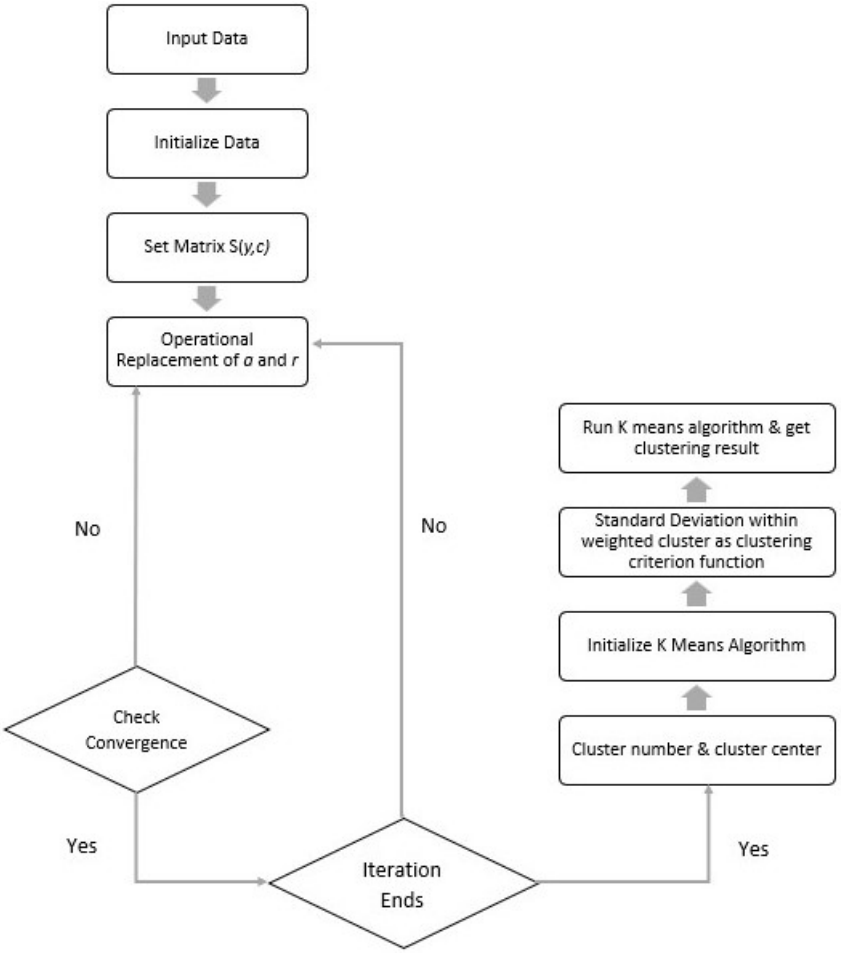



**Figure 2.** SAPK + K-Means Process.

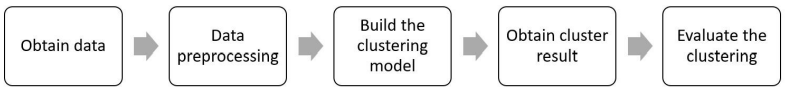



**Figure 3.** Process of applying SAPK+K-Means algorithm in customer segmentation.

*Comparison between Clustering Algorithms*

The clustering time and the error rate of K-Means algorithm, AP + K-Means algorithm and SAPK + K-Means algorithm are described in Figure 4 below. According to the analysis, the SAPK + K-Means algorithm enhanced in this work has a low error rate in the clusters derived from the two data sets, like the simulated data set, followed by AP + K-Means method, and the K-Means algorithm has worst effect. However, the SAPK + K-Means algorithm in this study consumes more time than the AP + K-Means algorithm and the K-Means algorithm, which may be related to clustering structure of the dataset. Since, the SAPK + K-Means algorithm has a high accurate rate, the longer time required is reasonable.

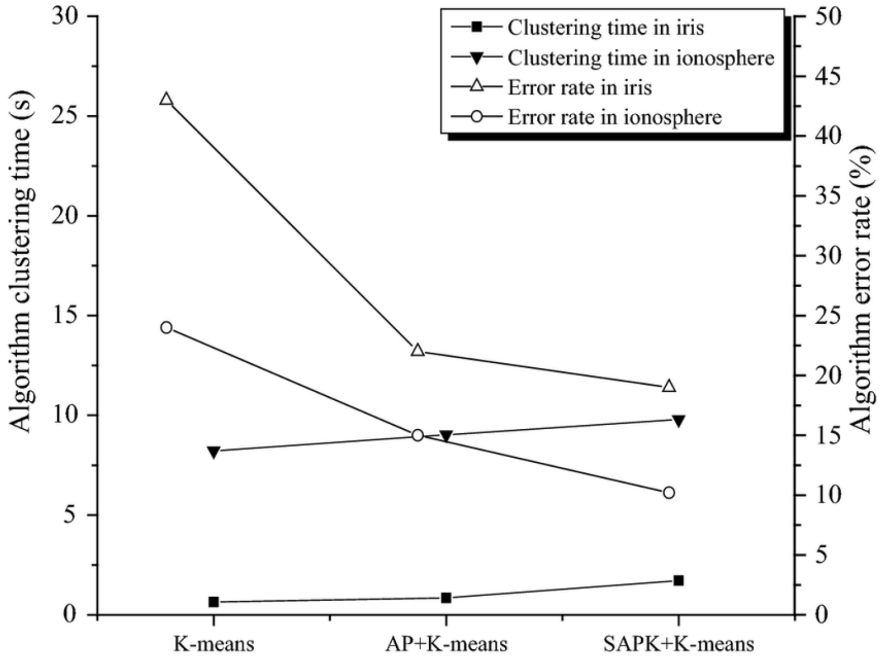



**Figure 4.** Comparison Analysis of Clustering algorithm.

## 4. Data Mining

The histogram in Figure 5 represents the frequency of each brand. As shown in Figure 5, the brand is converted to numeric from character. Each brand is represented as a number and the histogram shows the frequency of each brand viewed or purchased or added to cart. According to the histogram Figure 6 below, view for event_type has the highest frequency compared to cart and purchase events. This shows that the highest number of users view the products than purchasing and adding the products to cart. This could be because users compare the product by checking multiple items from online retail store before purchase an item. view for event type has the highest frequency compared to cart and purchase events. This shows that the highest number of users view the products than purchasing and adding the products to cart. This could be because users compare the product by checking multiple items from online retail store before purchase an item. The category histogram in Figure 7 shows that the highest frequency is category 2 which is appliances and least frequency is category 6 which is Country yard. In a nutshell, appliance is the most viewed category by the users. shows the code to represent histogram for Product category by the view event type. The output of the figure below is shown in Figure 8. The output shown in Figure 8 is almost similar to the previous histogram because the number of views occurred is 431,539 and higher than the event type cart and purchase.

As shown in the Figure 8 above, it shows the frequency of the 10 product categories. The categories with high frequency are category 2 and 7 which are Appliances and electronic and least frequency is categories are 6 which is country yard. In short, appliances and electronic are the category users most purchased.

The Figure 9 above shows the K-Means algorithm applied to for the three groups across the two variables which are event_type and product_id. The Figure 10 below shows the cluster centers means.

```
hist(training$brand,
     xlab = "Brand",
     col="aliceblue",
     main = "Histogram of Product Brand")
```

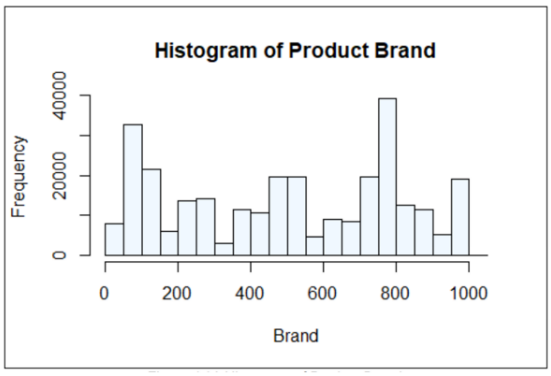



**Figure 5.** Histogram of Product Brand.

```
hist(training$event_type,
     xlab = "Event type",
     col="lightpink",
     main = "Histogram of Event Type")
```

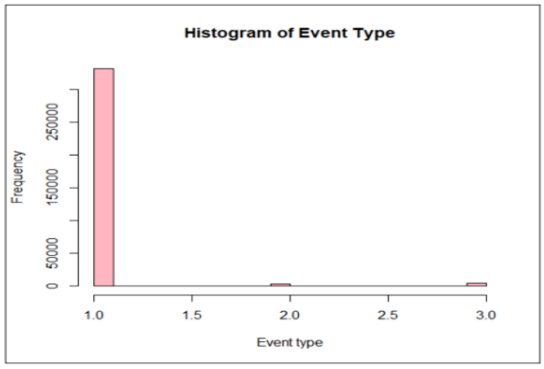



**Figure 6.** Histogram of Event Type.

```
hist(training$category_code,
     xlab = "Category code",
     col="lightblue",
     main = "Histogram of Category Code")
```

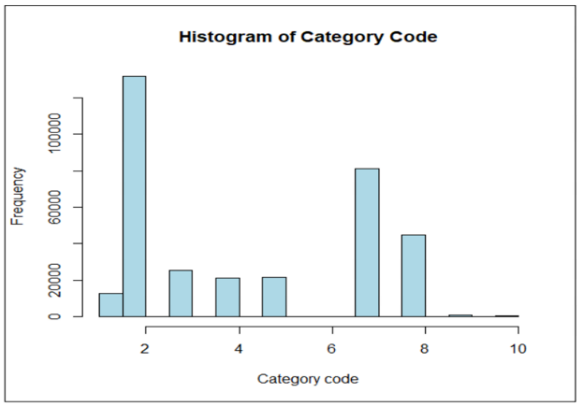
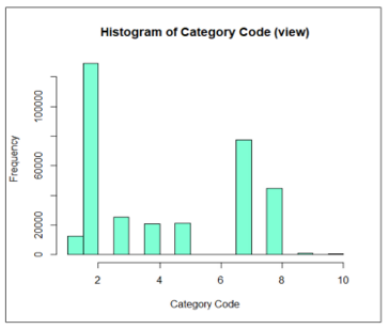


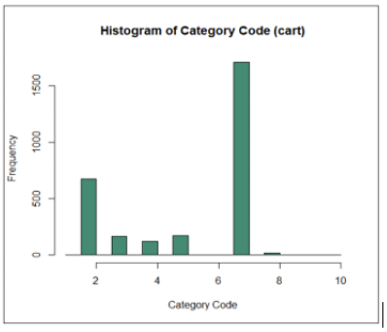

**Figure 7.** Histogram of Category Code.

```
hist(training$category_code[training$event_type==1],
     xlab = "Category Code",
     col="aquamarine1",
     main = "Histogram of Category Code (view)")
```

```
hist(training$category_code[training$event_type==2],
     xlab = "Category Code",
     col="aquamarine4",
     main = "Histogram of Category Code (cart)")
```

```
hist(training$category_code[training$event_type==3],
     xlab = "Category Code",
     col="cyan3",
     main = "Histogram of Category Code (purchase)")
```



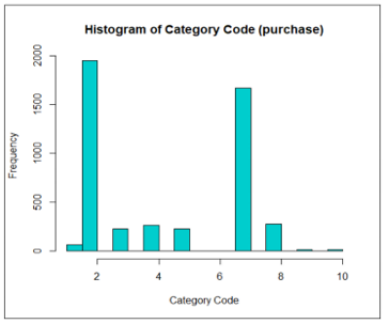




**Figure 8.** Histogram of Category_code by event_type—view, cart and purchase.

```
> cluster <- kmeans(training[,2:3],3) #event_type & product_id
> str(cluster)
List of 9
 $ cluster     : Named int [1:340367] 3 2 1 3 2 3 2 3 3 3 ...
  ..- attr(*, "names")= chr [1:340367] "2" "3" "9" "11" ...
 $ centers     : num [1:3, 1:2] 1.01 1.02 1.04 3.28e+07 1.50e+07 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:3] "1" "2" "3"
  .. ..$ : chr [1:2] "event_type" "product_id"
 $ totss       : num 2.6e+19
 $ withinss    : num [1:3] 1.44e+18 5.83e+17 7.00e+17
 $ tot.withinss: num 2.72e+18
 $ betweenss   : num 2.33e+19
 $ size        : int [1:3] 26324 66632 247411
 $ iter        : int 3
 $ ifault      : int 0
 - attr(*, "class")= chr "kmeans"
```

**Figure 9.** K-Means code.

```
> cluster
K-means clustering with 3 clusters of sizes 26324, 66632, 247411

Cluster means:
  event_type product_id
1   1.013600   32821175
2   1.019570   15015233
3   1.043216    4034613

Clustering vector:
   2    3    9   11   15   17   18   22   25   28   31   32   41   44   46
   3    2    1    3    2    3    2    3    3    3    1    2    3    3    3
  48   49   51   60   72   76   89   93   94  102  104  120  125  129  134
   3    3    3    2    3    2    3    1    2    3    3    1    3    3    1
 138  148  166  169  171  193  209  220  226  230  239  247  249  252  260
   3    1    3    1    3    3    3    3    1    1    3    1    3    3    2
 267  272  279  280  283  297  298  306  344  352  367  377  378  385  386
   3    2    2    2    1    1    1    3    3    3    3    3    2    3    3
 387  390  392  400  401  409  421  425  433  437  443  451  468  469  475
   2    3    3    3    2    3    3    3    2    3    2    3    3    3    3
 483  488  492  494  495  496  499  513  517  525  527  531  535  536  541
   1    3    2    3    2    1    3    3    3    2    3    3    1    3    3
 542  558  562  569  588  589  593  594  602  604  605  608  612  616  625
   3    3    3    3    3    2    3    3    3    3    3    3    3    3    3
 631  647  655  657  666  668  670  671  672  673  679  688  689  700  704
   3    3    3    2    3    3    3    3    2    3    3    3    2    3    3
 709  724  727  732  736  738  739  746  750  753  758  761  765  766  769
   3    3    3    3    3    3    3    3    3    3    3    3    3    2    3
 777  779  789  797  800  802  827  829  840  848  849  856  866  884  886
   1    3    1    3    3    3    2    3    3    3    3    3    3    3    3
 893  899  905  907  916  918  923  928  930  932  949  951  952  960  961
   3    3    3    1    3    3    3    3    3    3    3    2    3    3    3


 966  971  972  981 1002 1006 1011 1030 1049 1054 1061 1065 1095 1096 1099
   3    3    2    3    3    3    3    3    3    3    3    3    2    1    3
1109 1129 1133 1136 1137 1149 1158 1160 1182 1188 1200 1205 1216 1224 1227
   3    3    2    2    3    3    3    3    3    3    3    3    3    3    3
1228 1235 1251 1272 1274 1290 1304 1306 1310 1332 1335 1337 1355 1374 1379
   3    3    2    3    3    1    3    3    3    1    1    3    3    1    3
1382 1386 1389 1391 1401 1437 1440 1442 1449 1452 1456 1460 1464 1465 1471
   3    3    2    3    3    3    3    2    1    3    1    1    3    2    3
1475 1482 1484 1485 1489 1503 1505 1520 1521 1522 1525 1545 1551 1557 1562
   2    3    3    3    1    2    3    3    1    3    3    2    3    3    3
1563 1568 1570 1571 1572 1591 1597 1598 1605 1611 1624 1627 1630 1642 1643
   3    1    3    3    3    3    3    3    3    2    3    3    1    3    3
1649 1659 1674 1680 1682 1688 1689 1710 1734 1747 1762 1770 1778 1793 1810
   3    3    3    3    3    3    3    3    2    3    2    3    2    3    1
1825 1830 1838 1839 1840 1842 1854 1857 1872 1882 1891 1898 1909 1911 1916
   3    3    3    3    3    3    3    3    3    3    2    3    3    3    1
1927 1937 1942 1970 1978 1994 1995 1998 2018 2019 2030 2033 2034 2042 2043
   3    3    2    3    3    3    3    3    2    3    1    3    3    3    2
2046 2050 2053 2063 2066 2078 2097 2103 2121 2124 2127 2133 2134 2146 2161
   3    3    3    3    3    3    3    3    2    3    3    2    3    3    3
2162 2172 2180 2181 2192 2195 2200 2209 2210 2214 2219 2221 2224 2238 2241
   2    3    1    3    2    1    3    2    2    3    3    3    3    3    3
2242 2248 2250 2254 2257 2263 2270 2284 2285 2287 2298 2301 2305 2308 2320
   3    3    3    3    3    2    3    1    3    2    3    3    3    3    3
2328 2339 2344 2349 2363 2394 2402 2403 2404 2415 2425 2436 2446 2450 2468
   3    3    1    3    3    3    3    3    3    3    3    3    3    3    2
2475 2476 2483 2490 2495 2496 2498 2499 2509 2516 2517 2522 2536 2539 2541
   1    2    3    1    3    1    1    2    3    1    3    3    3    2    3
2545 2552 2558 2559 2571 2573 2579 2583 2584 2589 2627 2631 2633 2638 2666
   2    3    1    2    3    3    2    3    3    2    2    3    3    3    3
```

**Figure 10.** Cluster.

Figure 11 shows that k means algorithm is applied to column 4 to 7 with the k value 3. According to the figure above the integer vector from 1 to 340,367 specifying the cluster to which each point belongs. It shows the cluster center, sum of squares in the cluster and between clusters and number of points in each cluster.

```
> cluster1 <- kmeans(training[,4:7],3)
> str(cluster1)
List of 9
 $ cluster     : Named int [1:340367] 2 3 2 2 2 2 2 2 2 2 ...
  ..- attr(*, "names")= chr [1:340367] "2" "3" "9" "11" ...
 $ centers     : num [1:3, 1:4] 5.1 4.26 4.64 11.7 8.06 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:3] "1" "2" "3"
  .. ..$ : chr [1:4] "category_code" "product_detail" "product_detail.1" "price"
 $ totss       : num 2.37e+10
 $ withinss    : num [1:3] 1.45e+09 1.55e+09 1.94e+09
 $ tot.withinss: num 4.94e+09
 $ betweenss   : num 1.88e+10
 $ size        : int [1:3] 6806 255324 78237
 $ iter        : int 2
 $ ifault      : int 0
 - attr(*, "class")= chr "kmeans"
```

**Figure 11.** K-Means Algorithm 1.

This code in Figure 12 display the illustration of plot as shown in Figure 13. The code training [,2:3],3 is indicating that column two and three from the training dataset which are event type and product id, is chosen and k is set as 3. R will randomly choose 3 point and compute the Euclidean distance and create the clusters. R will randomly choose 3 points and compute the Euclidean distance and create the clusters. The illustration with plot shown above displays the clustering analysis in two dimensions between the event type and product id variables. Cluster 1 in black, Cluster 2 in red, and Cluster 3 in green are the three clusters with centroids identified. There is one large black cluster at the top, one red cluster below the black color cluster and one green cluster smaller than the red cluster at the bottom. Cluster 1 indicates the product types that customers are most interested in. Cluster 2 represents the types of products in which customers have a moderate interest. Cluster 3 indicates the products that customers have the least interest in.

```
> kmeans.ani(training [,2:3],3)
```

**Figure 12.** K-Means animation.

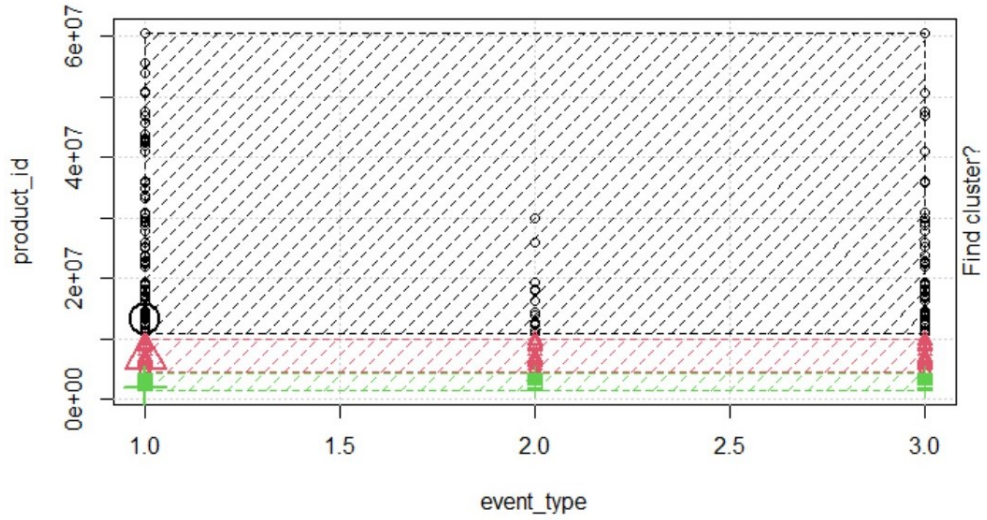



**Figure 13.** Animation 1, ○ High interest, △ Moderate interest, ▮ Low interest.

This code will display the illustration of plot as shown in Figure 14. The code training [,4:7],3 is indicating that column four to seven from the training dataset which are category code, product detail, product detail 1 and price, are chosen and k is set as 3. R will randomly choose 3 points and compute the Euclidean distance and create the clusters. Then it will compute the centroids and repeat the process until there is no changes in the illustration. The clusters can be labelled as Cluster 1 in black, Cluster 2 in red, and Cluster 3 in green. The size of cluster 1 is 255,324, cluster 2 is 78,237 and cluster 3 is 6806. Figure 15 below shows the cluster in black color is considered as the high profitable segment, cluster in red represents moderate profitable segment and green cluster represents the least profitable segment. Therefore, the conclusion from the Figures 14 and 15, shows that products in which customers with high interest yield high profit, products in which customers with moderate interest yield moderate profit, and products that has low interest yield low profit.

```
Kmeans.ani(training [4:7],3)
```

**Figure 14.** Kmeans Animation.

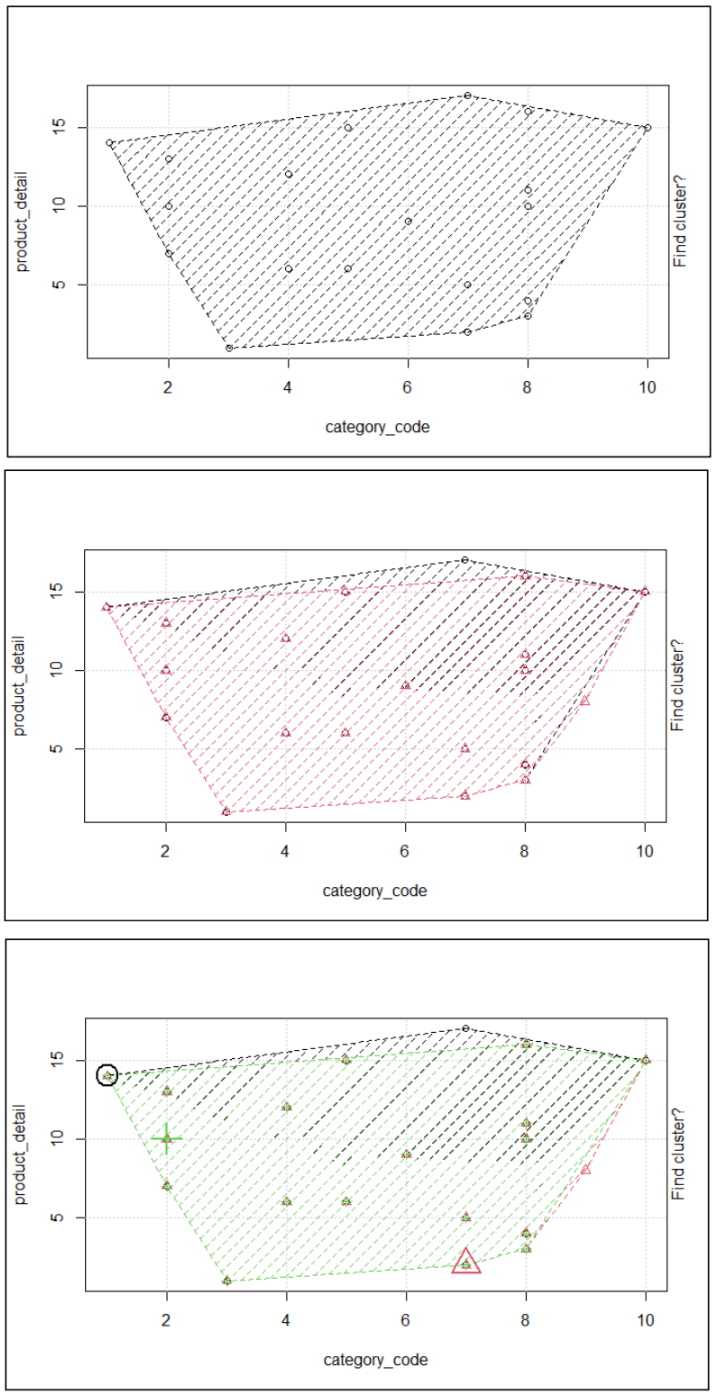



**Figure 15.** Animation 2.

## 5. Data Visualization and Dashboard

Figure 16 shows the overview of the dashboard. The diagram below illustrates the overview of all the e category code, brand, price, product details and event type. This page shows the overall distribution and analysis on the data for all the attributes or variables analyzed.


**Figure 16.** Overview of Dashboard.

Figure 17 below shows the product brand page of the customer segmentation dashboard. This page explains on the brand preferred by the E-commerce users. The visualizations show the most popular product brand to the least popular product. The dashboard also shows the table that display the number of product brands each user viewed, purchased, or added to cart. The data analysis is shown in table and clustered column chart to visualize it.

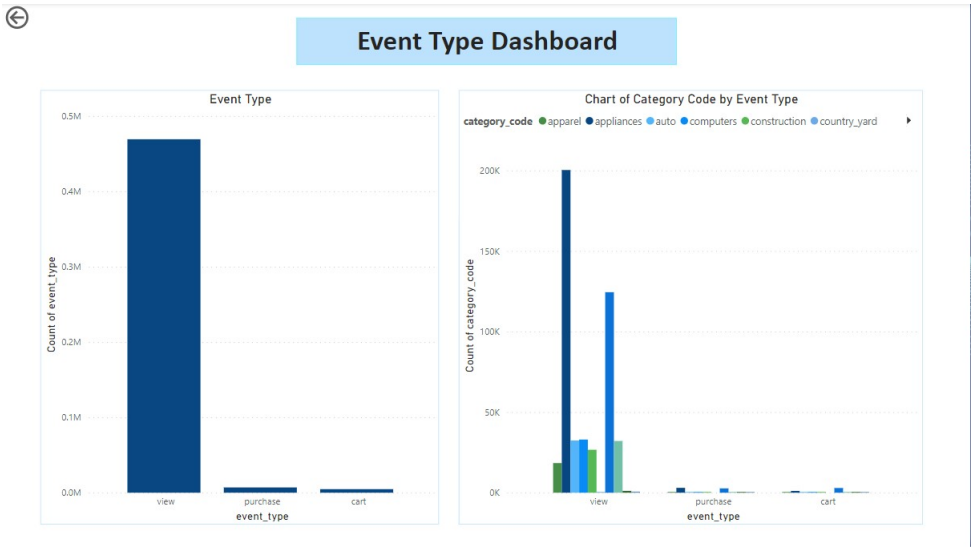



**Figure 17.** Event Type Dashboard.

Figure 18 above shows the analysis on time of customer segmentation dashboard. This analysis explains the time of user login to the online store. The dashboard can help to understand the most popular time of the online store. The visualization of the data analytics for time attribute is shown in as clustered column chart. Example, according to the dashboard, the time with the highest frequency is 3.59 pm. From this chart, it can be concluded that the popular or active time is from around 2 pm to 4 pm.

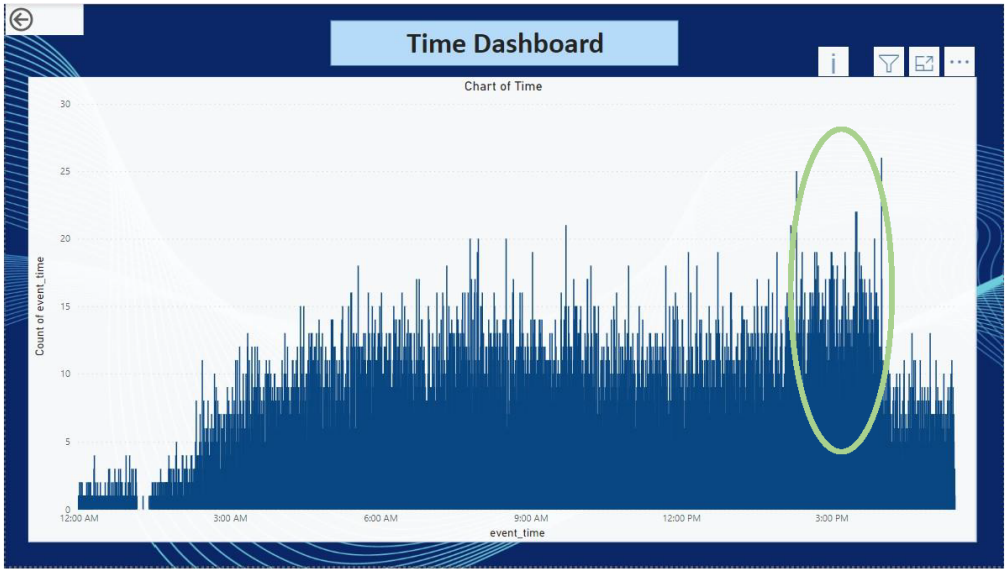



**Figure 18.** Time Dashboard.

## 6. Discussion

The research was performed on understanding the E-commerce platform. Research on the algorithms such as hierarchical clustering, K-Means clustering, and mean shift clustering were evaluated for customer segmentation. From the research, K-Means clustering was chosen as most suitable for conducting the customer segmentation. The chosen algorithm performance is shown on customer purchase history data for customer segmentation.

The execution of the algorithms and data sets show how the validation improves when working with more stable data, such as the nominal values smoothed by the z score. This stability is translated into less variability in the construction of the clusters in the three periods. The variability improves when working with the smoothed data set. This is a relevant point when considering the design of a longitudinal study to find the individuals that are representative of the same type of municipality. Of the clustering presented, three of the maps can be identified as the most representative of profitable segments. The maps generated with the nominal and smoothed data sets, using the hierarchical K-Means algorithm, which showed three clusters: event_type and products and Categories. This research we analyzed the groups that share the similar criteria helps vendors to identify and focus on the high profitable segment to the least profitable segment that can play important role in improving the business. In addition, the proposed approach can be employed on any business related datasets for example sustainability and health related data to identify the customers behaviours.

The current research findings and interpretation inform marketing decision with e-commerce purchasing patterns to guide vendors toward effective engagement when confronted with volatility on data analysis. The market segmentation research, regardless the method, has informative value for marketing decision only if a new value proposition will target customers and prospects who respond similarly to marketing programs.

## 7. Conclusions

This research aims to help researchers and other E commerce stakeholder for a comparison of the structuring and grouping of the E-commerce purchasing pattern small areas. It also endeavors to show an optimal way of transforming and working on datasets to facilitate the resulting groupings. Therefore, this research allows us to segment a cohort from E-commerce behavior data from multiple category store and purchasing histories. In addition, it would capture the inequality that can be observed between the high profitable segments and low profitable segment in category of products better.

In conclusion there are some limitation such as working with microdata both making comparisons and modeling and clustering, especially if they are business management and sales financial data. The difficulties of working with indicators, indexes, and rates complicate the data mining process and, later, the reading of the results. A smoothing or standardization process is necessary to work effectively [16,17]. It must be considered that using percentages with such small data sets mean that these can drastically change from year to year. These possible irregularities accentuate the variations and generate an elevated volatility. This volatility affects the clustering and models, making their classification difficult. Therefore this is another point to further studies in microdata and subgroups. This was not possible in this study due to the lack of data. Recent literature survey shows that deep learning-based methods performing better compared to machine learning. Thus, applying deep learning-based approaches for customer segmentation and comparing the performance with the SAPK + K-means clustering algorithm will be considered as one of the significant directions towards future work.

**Author Contributions:** Conceptualization, K.T. and S.V.; methodology, K.T. and S.V.; software, K.T. and S.V.; validation, K.T., S.V. and V.R.; formal analysis, K.T., S.V. and V.R.; investigation, K.T., S.V. and V.R.; resources, K.T. and S.V.; data curation, K.T. and S.V.; writing—original draft preparation, K.T. and S.V.; writing—review and editing, K.T., S.V. and V.R.; visualization, K.T. and S.V.; supervision, K.T. and S.V.; project administration, K.T. and S.V.; funding acquisition, K.T.; All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data presented in this study are available on request from the corresponding author.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Sayyida, S.; Hartini, S.; Gunawan, S.; Husin, S.N. The Impact of the Covid-19 Pandemic on Retail Consumer Behavior. *Aptisi Trans. Manag. (ATM)* **2021**, *5*, 79–88. [CrossRef]
2. Bhaskara, G.I.; Filimonau, V. The COVID-19 pandemic and organisational learning for disaster planning and management: A perspective of tourism businesses from a destination prone to consecutive disasters. *J. Hosp. Tour. Manag.* **2021**, *46*, 364–375. [CrossRef]
3. Nie, F.; Li, Z.; Wang, R.; Li, X. An Effective and Efficient Algorithm for K-means Clustering with New Formulation. *IEEE Trans. Knowl. Data Eng.* **2022**. Available online: https://ieeexplore.ieee.org/abstract/document/9723527/ (accessed on 2 May 2022).
4. Brandtner, P.; Darbanian, F.; Falatouri, T.; Udokwu, C. Impact of COVID-19 on the customer end of retail supply chains: A big data analysis of consumer satisfaction. *Sustainability* **2021**, *13*, 1464. [CrossRef]
5. Khong, D.W.K. Rents: How Marketing Causes Inequality by Gerrit De Geest. *Asian J. Law Policy* **2021**, *1*, 83–86. [CrossRef]
6. Manero, K.M.; Rimiru, R.; Otieno, C. Customer Behaviour Segmentation among Mobile Service Providers in Kenya using K-Means Algorithm. *Int. J. Comput. Sci. Issues* **2018**, *15*, 67–76.
7. Janardhanan, S.; Muthalagu, R. Market segmentation for profit maximization using machine learning algorithms. *J. Phys. Conf. Ser.* **2020**, *1706*, 012160. [CrossRef]
8. Dawane, V.; Waghodekar, P.; Pagare, J. RFM Analysis Using K-Means Clustering to Improve Revenue and Customer Retention. In Proceedings of the International Conference on Smart Data Intelligence (ICSMDI 2021), Online, 29–30 April 2021. [CrossRef]
9. Rachmawati, I.K. Collaboration Technology Acceptance Model, Subjective Norms and Personal Innovations on Buying Interest Online. *Int. J. Innov. Sci. Res. Technol.* **2020**, *5*, 115–122.
10. Shirole, R.; Salokhe, L.; Jadhav, S. Customer Segmentation using RFM Model and K-Means Clustering. *Int. J. Sci. Res. Sci. Technol.* **2021**, *8*, 591–597. [CrossRef]
11. Ekren, B.Y.; Mangla, S.K.; Turhanlar, E.E.; Kazancoglu, Y.; Li, G. Lateral inventory share-based models for IoT-enabled E-commerce sustainable food supply networks. *Comput. Oper. Res.* **2021**, *130*, 105237. [CrossRef]
12. Sinaga, K.P.; Yang, M.S. Unsupervised K-means clustering algorithm. *IEEE Access* **2020**, *8*, 80716–80727. [CrossRef]
13. Kuruba Manjunath, Y.S.; Kashef, R.F. Distributed clustering using multi-tier hierarchical overlay super-peer peer-to-peer network architecture for efficient customer segmentation. *Electron. Commer. Res. Appl.* **2021**, *47*, 101040. [CrossRef]

14. Deng, Y.; Gao, Q. A study on e-commerce customer segmentation management based on improved K-means algorithm. *Inf. Syst. Bus. Manag.* **2020**, *18*, 497–510. [CrossRef]
15. Suryadi, D.; Kim, H.M. A data-driven methodology to construct customer choice sets using online data and customer reviews. *J. Mech. Des. Trans. ASME* **2019**, *141*, 111103. [CrossRef]
16. Miloudi, S.; Wang, Y.; Ding, W. A Gradient-Based Clustering for Multi-Database Mining. *IEEE Access* **2021**, *9*, 11144–11172. [CrossRef]
17. Mishra, S.K.; Dwivedi, V.; Sarvanan, C.K.; Pathak, K. Pattern Discovery in Hydrological Time Series Data Mining during the Monsoon Period of the High Flood Years in Brahmaputra River Basin. *Int. J. Comput. Appl.* **2013**, *67*, 7–14. [CrossRef]