A project Report on

Customer visit segmentation in an online retail market

Submitted by

Partha Sarathi Mishra

(ROLL NO: 14MF3IM08)

Under the guidance of

Prof. J.K. Jha

DEPARTMENT OF INDUSTRIAL AND SYSTEMS ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY, KHARAGPUR

DECLARATION

I certify that

(a) The work contained in this report has been done by me under the guidance of my supervisor.

(b) The work has not been submitted to any other Institute for any degree or diploma.

(c) I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

(d) Whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references. Further, I have taken permission from the copyright owners of the sources, whenever necessary.

Date: Nov 5, 2018 Partha Sarathi Mishra

Place: Kharagpur. Roll: 14MF3IM08

DEPARTMENT OF INDUSTRIAL AND SYSTEMS ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

KHARAGPUR - 721302, INDIA

CERTIFICATE

This is to certify that the project report entitled "**Customer visit segmentation in an online retail market**" submitted by **Partha Sarathi Mishra** (Roll No:**14MF3IM08**) to Indian Institute of Technology Kharagpur towards partial fulfilment of requirements for the award of degree of Master of Technology, in Industrial and Systems Engineering, is a record of bonafide work carried out by him under my supervision and guidance during Autumn Semester, 2018-19.

Prof. Jitendra Kumar Jha

Department of Industrial and Systems Engineering,

IIT Kharagpur

Date: November 5, 2018

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to my guide, Prof. J.K. Jha, for his supervision, encouragement and mentoring. His advice in choosing the right problem for my thesis has been really crucial. Further, his valuable suggestions, when I'm stuck with a problem, has helped me make good progress.

Partha Sarathi Mishra

5th Year Dual Degree student,

Department of Industrial and Systems Engineering,

Indian Institute of Technology, Kharagpur

Table of contents

1 Abstract

Every customer has a unique intent at the back of his mind when shopping in a retail store. It could be monthly stocking, food for an outdoor party or for organizing a children's picnic. Being able to understand and predict the items a customer is likely to choose, depending on his intent, could help online retail companies to gain competitive advantage via targeted advertisements.

The extant literature in this field mainly deals with customer segmentation and basket association for products. Attempt to understand the intent of each customer visit is a relatively new area of study, which has been mainly restricted to physical retail stores. This study extends this idea for online retails, where data like customer views and add to carts are also available, allowing us to group each visit more effectively. An essential step before segmentation into groups is to identify the optimum taxonomic hierarchy or the category to which an item belongs to. Earlier attempts at choosing the right taxonomic division has been directly through managerial inputs or through development of self-designed algorithms based on that.

In our study, we intend to reiterate the segmentation based on different levels of taxonomic divisions to arrive at an optimum. We intend to identify the "shopping mission" behind every customer visit and hence we group the items a customer puts in his basket. For clustering, the data is transformed into a binary matrix of whether an item is present or not in a basket. Previous attempts at clustering have mostly focussed on using widely used clustering algorithms like the K-Means clustering. We further propose better clustering methods as opposed to the traditional methods, to exploit the special binary matrix structure of the data. Apart from the binary matrix also known as the factor table, a frequency table of the number of items of each kind contained in a basket can also be clustered. We have chosen the open source data from Retail Rocket, an online retail company, to validate our propositions. As it is an open source data with hashed private information, all our results would be both industry relevant as well as generalized for all companies in the industry. Using broadest level of item category as chosen taxonomy and K-means clustering, we report 12 clusters of customer visits which would serve as our base case for future comparisons.

[Keywords: Online retail, Data Mining, Basket Analytics, Customer Segmentation]

2 Introduction

Online retails stores are ubiquitous in today's internet era. Even physical retail stores usually tend to have at least a website displaying the myriad of their products so that they don't lag behind their competitors in appealing to the customers. Internet has not only enabled a wider reach for these online retailers but also it has given them access to all sorts of new customer data apart from what they could record in a physical store. This availability of vast amounts of data has paved the way for an intensive use of business analytics by these firms. To improve their bottom line, these firms try to understand the customer behaviours through their transaction activities and interactions with the firm's products. Using this information, they improve their marketing strategy and make improvements in their products or selling methods.

To understand customer behaviour, these retailers make use of data mining techniques to segment customers into various groups so that they can pursue a more targeted marketing approach or avail these customers of a more personalized experience. Another facet of data mining employed is to use the data on customer interaction with products to establish certain association rules among these products. Such an information on association among products can be used to suggest similar items together in case of online retail, which draws parallel with putting similar items together on shelves in a physical retail store. Retail businesses are so competitive that more and more research is being done in analyzing micro level behaviour of customers as even a small insight may cause a drastic impact. Managers are now looking into ways to understand every single customer visit and the term "shopping mission" (ECR Europe, 2011) is being explored. Being able to identify the intent behind a shopping trip of any given customer can help provide services tailored to his needs.

In India, business analytics is becoming increasingly important with the big players like Amazon and Flipkart encroaching into almost every retail domain and dominating over the traditional physical retail players. Success in the online retail space has also drawn in several new players specific to a particular domain like BigBasket and big established firms entering into this digital competition like Reliance Fresh.

This paper intends to extend the idea of using data mining in retail sector to understand each customer visit to the online retail space. As has been pointed out in Anastasia et al (2008), the

availability of vast information in the new age deems it necessary to carry an independent analysis for online retailers. For the sake of analysis, the data for a food retail sector is considered. Further, we propose the use of another algorithm that takes into account the structure of the data as opposed to a general algorithm used in previous work.

3 Literature Review

3.1 Customer Segmentation

Customer segmentation is the process of dividing heterogeneous customers into homogeneous groups on the basis of common attributes and is essential for handling a variety of customers with rich sets of diverse customer preferences more efficiently (Hong & Kim, 2012). Customer segmentation is done based on several types of data like sales data, behavioral data, demographics data, etc. Ngai, Xiu, & Chau (2009) used customer level sales data for segmentation and experimented with models like clustering, sequence discovery, forecasting and classification. Larson, Bradlow, and Fader (2005) examined common travel behaviours in a grocery store by segmenting customers based on RFID fitted in their shopping carts. Park et al. (2014) studied the customer purchase patterns in a multi category context. Customer lifestyle information was used to perform customer segmentation by Miguéis et al. (2012) on a large transactional database. Han, Ye, Fu, and Chen (2014) identified customers who purchase routine, seasonal or only in convenience categories using clustering techniques. Liao, Chen, and Hsieh (2011) use data collected via questionnaires and used it to segment customers based on their lifestyle habits and purchase behaviours.

3.2 Market basket analysis

Market basket analysis studies perform market basket analysis (also known as association rule mining), which is a data mining method that examines large transactional databases to determine which items are most frequently purchased jointly (Agrawal et al., 1993; Srikant & Agrawal, 1995). This analysis is used to change the store's layout so that it appeals more to the customer. Chen, Tang, Shen, & Hu (2005) discusses the extension of this approach to other domains like finance, telecommunications, etc. Tang et al. (2008) studies the market basket analysis in a multi-store and

multi environment.

3.3 Customer visit segmentation

Europe, E.C.R. (2011) points out that managers are increasingly looking to understand each individual customer visit. To this end, customer visit segmentation by Anastasia et al. (2018) helps to give a macro analysis of the different types of intents that customers have in their mind while they visit a physical retail store. They analyze each customer visit at look at the customer's interaction with the products at sale. With the use of K-Means clustering, they segment the customers based on what products they choose to buy in the particular visit. This is a midway between the most studied methods of customer segmentation and market basket analysis.

This paper attempts to perform customer visit segmentation in an online retail space and also proposes the use of specialized segmentation methods than the K-means clustering which is quite a general method so that the special binary structure of the data can be utilized.

4 Problem Description

As described in the previous section, most of the efforts in the past have been concentrated around either customer segmentation or basket analytics. The objective of this study is two folds:

1. It intends to extend the idea of identifying the *"shopping mission"*[1] as introduced in Anastasia et al. (2018), to online retail stores.
2. It proposes the use of a less generic algorithm, than the one used in Anastasia et al.(2018), for identifying the shopping intent in order to benefit from the unique data structure of the data.

Instead of segmenting the customers based on their entire history of visits, to determine their behaviour, we look at each individual customer visit. In basket analytics, only an association is derived between the products that are being bought together. But in our study, we use this data on the products being bought in each basket and link it to the customer profile. Then we use this data, to segment various customers on the basis of their intent of visit, physical or digital, to the retail store. One advantage that retail stores wield over the physical store is the availability of more data. Data is

of prime importance in business analytics and the more the data, the better the chances at achieving accurate segmentation of the customers. Online retails have several extra data measures like add to cart, number of item views and so on. Almost anything done in digital space can be recorded and economically exploited. Hence, it makes sense to perform such an analysis again in the online retail domain to gain critical insights on customer behaviours not available otherwise to the physical stores.

To identify the customer segmentation by visit, the approach followed in this paper is clustering. Clustering, an unsupervised learning method, makes use of several basket features of each customer to segment them based on their similarity. Then by inspecting the items in each basket, we can determine the intent of visit of the customer. Here, the data matrix containing basket data for each customer visit, called factor table, is used for clustering. This table has a binary structure. The approach taken earlier was to use the most popular K-Means algorithm in clustering. However, K-Means clustering uses the idea of mean which is absent in binary data. Even if, Hamming's distance is used instead of the Euclidean distance, this approach would not be accurate. Hence, we propose the use of an algorithm suitable for this structure. Borrowing the idea of clustering in machine cells, which clusters machines and parts and has a binary structure, we use rank order clustering.

5 Methodology

As discussed before, to achieve the first objective of determining the shopping intent of a customer, we employ the method of clustering on each individual customer visit. In each visit, a customer interacts with the products on sale and this interaction data is used to form different clusters, unlike traditional customer segmentation where the customer's previous history of visits is used in segmentation. This allows us to understand the intent of a group of customers by looking at which products they interact with during their visit.

Clustering is employed for data mining purpose and to standardize the steps in this process, the CRISP-DM[2] or cross industry standardized process for data mining is employed. This constitutes of 6 steps originally:

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

The contribution of this study is mainly in the data understanding and modeling phases.

5.1 Business Understanding

This step deals with pinpointing the objective behind performing the data mining task. In our problem, the objective from a business point of view is to utilize the customer interaction data to improve the bottomline of a firm. The approach of this paper, is to segment each individual customer visit based on the customer's interaction with the products on sale so as to understand the intent of each customer's visit.

5.2 Data understanding

The data required for this purpose should contain the customer's interact with products on sale in a retail firm. Examples of such data can be, in a particular visit, which products is a customer adding to the cart, which among these is he finally buying, which are the ones he discards and finally what does he store in his wishlist. As has been observed before, some of this information like add to cart and wishlist would only be available in the online retail case and hence this study is more likely to identify customer visit segments more accurately as compared to offline physical store case. This phase consists of several substeps:

5.2.1 Data Collection

The data required for such an analysis of baskets for each visit is confidential and of high value to any firm. Hence, it is difficult to obtain a complete dataset for this purpose without confidentiality agreements. For our purpose, however, we have selected an open source data repository containing information relevant to us. This dataset has data on online food retail sector and has been contributed by Retail Rocket, an online retail firm, with some confidential data hashed for privacy. While we may not be able to determine the exact item details for which value is provided, such a dataset also

provides a salient advantage. Using this dataset is better than applying theory on randomly generated data as it would have the characteristics particular to the industry and also the hashing of specific data restricts us to derive insights that are generic rather than specific to the firm. The data sets are arranged in the following files:

1. *category_tree.csv:* This is a 2d matrix that maps each item category to its parent category.
2. *events.csv:* This contains customer firm interaction data like the time and nature of action performed by a customer. For example, If a customer viewed or added an item to his cart or made a transaction.
3. *item_properties1.csv, item_properties2.csv:* The item property files contain information about an item like its price, category id, availability, etc in different time snapshots. Of these data, only category id and availability are not hashed.

This data has been most widely used developing recommender systems. Hence, mostly predictive analysis has been performed using it. This study however is more interested in data mining through customer visit segmentation.

	categoryid	parentid
0	1016	213.0
1	809	169.0
2	570	9.0
3	1691	885.0
4	536	1691.0

Fig 1: category_tree.csv

	timestamp	visitorid	event	itemid	transactionid
0	1433222531378	57036	view	334662	NaN
1	1433223239808	1377281	view	251467	NaN
2	1433223236124	287857	addtocart	5206	NaN
3	1433224244282	1370216	view	176721	NaN
4	1433221078505	158090	addtocart	10572	NaN
5	1433222276276	599528	transaction	356475	4000.0
6	1433193500981	121688	transaction	15335	11117.0

Fig 2: events.csv

	timestamp	itemid	property	value
0	1435460400000	460429	categoryid	1338
1	1441508400000	206783	888	1116713 960601 n277.200
2	1439089200000	395014	400	n552.000 639502 n720.000 424566
3	1431226800000	59481	790	n15360.000
4	1431831600000	156781	917	828513

Fig 3: item_properties.csv

5.2.2 Exploratory Data Analysis (EDA)

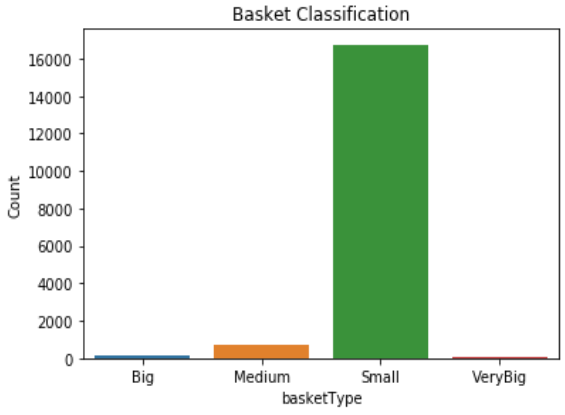

Fig 4: Basket Classification

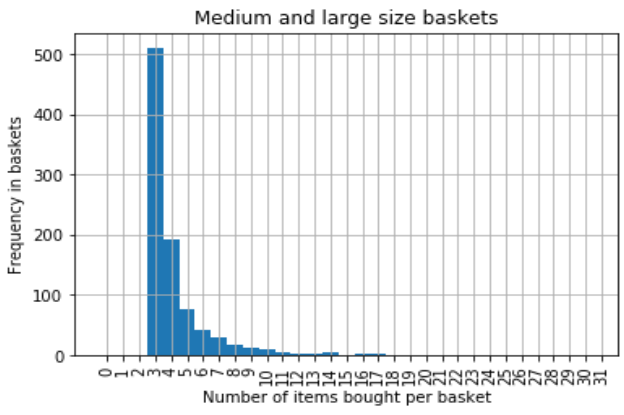

Fig 5: Medium to large size basket histogram

Exploratory analysis for our purpose of clustering can help in outlier detection. A very big basket consisting of too many items may not reveal any purpose of visit and may just be a general store visit. On the other hand, very small baskets could be useful because they may represent visits for a specific product of immediate need.

5.2.3 Verification of data quality

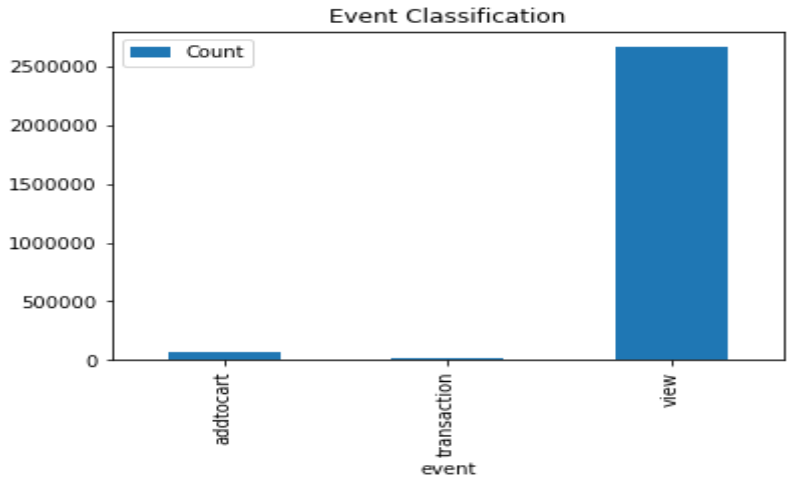

Fig 6: Frequency of different events in the data set

Usually to validate the data, ad-hoc measures proposed by management is used. For our purposes, a part of EDA depicting the frequency of each of the events, add to cart, transaction and views can be compared with our common knowledge of their relative frequencies to check that the data is indeed applicable to the industry. As is common knowledge, most of the products are viewed, some of which are added to cart and a smaller of fraction of which is finally bought. Our data confirms with this observation.

5.3 Data preparation

In this step, the data is preprocessed before it is used. To segment the customers per visit first the data needs to be transformed. The customer data from the events file had to be merged with the item properties file to create the the customer product interaction matrix. This matrix could either contain if a customer's basket contained a product or not, factor table, or the number of products of a

particular type present in the basket, frequency table. Further, the category tree file had to be used to identify the category id of a product at our chosen taxonomy level.

```
In [71]: factor_table.loc[[11117]]
```

Out[71]:

transactionid	140.0	1600.0	653.0	1224.0	395.0	859.0	1698.0	1482.0	1532.0	250.0	679.0	791.0	1579.0	1490.0	803.0	378.0	1452.0	431.0
11117	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Fig 7: Factor table

```
In [68]: frequency_table[frequency_table.transactionid ==11117]
```

Out[68]:

	transactionid	140.0	1600.0	653.0	1224.0	395.0	859.0	1698.0	1482.0	1532.0	250.0	679.0	791.0	1579.0	1490.0	803.0	378.0	1452.0	431.0
10897	11117.0	5.0	0.0	6.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Fig 8: Frequency table

5.4 Modeling

This phase deals with the choice of taxonomic level and clustering for our segmentation purpose.

In our initial study, we have chosen the topmost level of classification of category id i.e. the category ids considered don't have any parent categories. This gave a reasonable number of items in each category to begin our study.

5.4.1 K-Means

For clustering, we begin with the K-Means clustering method as used by Anastasia et al (2018). This algorithm works as follows:

Input: *k* (the number of clusters),
D (a set of lift ratios)
Output: a set of k clusters
Method:
Arbitrarily choose *k* objects from *D* as the initial cluster centers;
Repeat:
1. (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
2. Update the cluster means, i.e., calculate the mean value of the objects for each cluster
Until no change;

Fig 9: K-Means clustering pseudo code[13]

To obtain the optimum number of clusters, we make use of the elbow curve. This curve is a plot of the sum squared error vs the number of clusters. The optimum is chosen by visual inspection at the point where the elbow curve flattens. This suggest that there is no further significant decrease in the error with increase in the number of clusters.

5.4.2 Rank order clustering (ROC)

This is a specialized clustering technique seen to be more effective for binary data structures. Since the original ROC is computationally very expensive, there are research on many several modified approaches. The modified approach used here is:

1) Obtain the factor table

2) Compute a set of the top-k nearest neighbors for each customer visit

3) Compute pairwise distances between each customer visit and its top-k nearest neighbor.

4) Transitively merge all pairs of faces with distances below a threshold

5.5 Evaluation

As the product details data is encrypted, we cannot comment on the nature of clusters by taking managerial input or comparing it with common sense. We have initially applied the K-Means clustering algorithm in which finally we obtain several clusters along with a cluster mean value at

the centre of the cluster. So, if the clustering is effective, then all the data points belonging to a particular cluster would be closer to the cluster mean. If we define, sum squared error (SSE) as the sum total of squared deviation of each data point from its respective cluster mean for a given number of clusters using apatticular method, then we can expect a better clustering method to give smaller SSE. We use this metric for our evaluation.

For the approximate ROC that we use, however, we need to specify a certain threshold that it uses to determine if data points being compared belong to a cluster or not. This is a very computationally expensive process and seems to persistently end in consuming all of the free RAM available. So, while a similar plot can be used by gradually increasing the threshold, hence reducing the number of clusters, there is a need to first manage the data storage more efficiently. This can possibly be done by using big data techniques.

6 Results and Discussions

Anastasia et al, in their segmentation of physical retails have employed K-Means clustering for the purpose of segmentation. They also use an elbow curve to identify the optimum number of clusters. This is essential because, if we use too many number of clusters, we may miss the big picture, the customer intent, which is our core objective. Whereas very few clusters also pose the risk of not supplying any valuable insight. Hence, to choose the optimum number of clusters, they plot the sum squared errors of all the data points in the clusters from the cluster mean obtained using K-Means clustering vs number of final number of clusters. Hence, once the change in sum squared error becomes insignificant, we can identify that further clustering isn't extracting much valuable information.

Our first objective is to obtain clusters from the customer-product interaction data. To check if our approach by combining additional data available to online retail firms results in a more accurate cluster than the case of physical store, we can compare the sum squared errors obtained in clustering the data. The following plot shows the result from performing K-means clustering on our data set using all the data also available in physical stores.

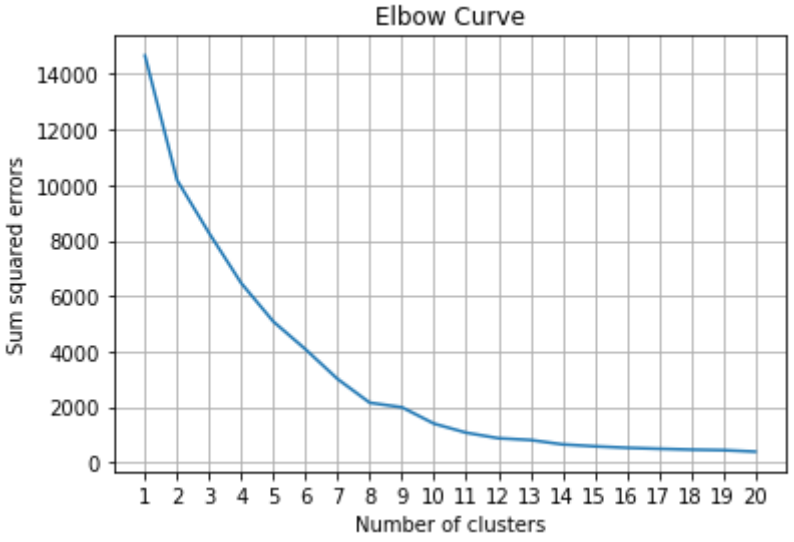

Fig10: Elbow curve

Through visual inspection, we can observe that after 12 clusters, the decrease in SSE is less than 1000. Hence, this can be chosen as the optimal number of clusters.

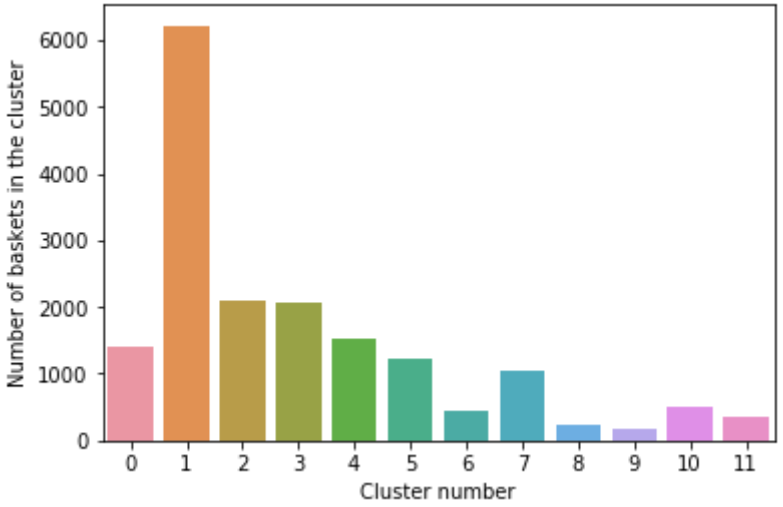

Fig 11: Number of baskets in each of the clusters

- After clustering, an inspection of the items in each cluster can reveal the intent or the shopping mission behind each visit. However, with our present data, since we are privy to the

exact items in each cluster we can infer about the characteristics of the cluster. For example, in Fig2, we can see that not all clusters are equitably distributed. Hence, the managers could prioritize their actions on the customer visit segmentation analysis by looking at the concentration of visits in each cluster. If there are more number of baskets or individual customer visits corresponding to a cluster, that means that those type of customers are more frequent visitors to the retail store and would contribute more to their bottom line.

7 Future work:

1. A comparison can be made in the accuracy of clustering using K-Means clustering technique between the online retail vs physical retail stores using the sum squared error metric.
2. Another comparison in clustering accuracy can be made between the use of factor table and the frequency table in segmenting customer visits.
3. The current use of taxonomic classification can also be improved by choosing an optimum category level so that the product category tree is balanced or the height of all the branches are nearly the same.
4. It is also observed that the size of data is large and in the data transformation phase, it multiplies causing a problem in storage and easy manipulation. Hence, a big data solution for data handling has to be explored.

8 References:

[1] Europe, E. C. R. (2011). The Consumer and Shopper Journey Framework. https://www.ecrireland.ie/uploadedfiles/shopper/projects- ecr- eu/the- consumer- and-shopper- journey- framework- .pdf

[2] Hong, T. , & Kim, E. (2012). Segmenting customers in online stores based on factors that affect the customer's intention to purchase. Expert Systems with Applications, 39 (2),

2127–2131

[3] Ngai, E. W. T. , Xiu, L. , & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. Expert Systems with Applications, 36 (2), 2592–2602 .

[4] Larson, J. S. , Bradlow, E. T. , & Fader, P. S. (2005). An exploratory look at supermarket shopping paths. International Journal of Research in Marketing, 22 (4), 395–414

[5] Park, C. H. , Park, Y.-H. , & Schweidel, D. A. (2014). A multi-category customer base analysis. International Journal of Research in Marketing, 31 (3), 266–279 .

[6] Miguéis, V. L. , Camanho, A. S. , Falcão, e. , & Cunha, J. (2012). Customer data mining for lifestyle segmentation. Expert Systems with Applications, 39 (10), 9359–9366 .

[7] Han, S. , Ye, Y. , Fu, X. , & Chen, Z. (2014). Category role aided market segmentation approach to convenience store chain category management. Decision Support Systems, 57 (1), 296–308 .

[8] Liao, S. , Chen, Y. , & Hsieh, H. (2011). Mining customer knowledge for direct selling and marketing. Expert Systems with Applications, 38 (5), 6059–6069 .

[9] Agrawal, R. , Imieli ´nski, T. , & Swami, A. (1993). Mining association rules between sets of items in large databases. In ACM SIGMOD International conference on management of data (pp. 207–216) .

[10] Chen, Y.-L. , Tang, K. , Shen, R.-J. , & Hu, Y.-H. (2005). Market basket analysis in a mul- tiple store environment. Decision Support Systems, 40 (2), 339–354 .

[11] Tang, K. , Chen, Y.-L. , & Hu, H.-W. (2008). Context-based market basket analysis in a multiple-store environment. Decision Support Systems, 45 (1), 150–163 .

[12] A. Griva, C. Bardaki, K. Pramatari, D. Papakiriakopoulos. Retail business analytics: Customer visit segmentation using market basket data, Expert Systems with Applications, 100 (2018), pp. 1-16

[13]https://www.researchgate.net/figure/The-pseudo-code-for-K-means-clustering-algorithm_fig2_273063437

[14] C Otto, D Wang, AK Jain (2016), arXiv preprint arXiv:1604.00989