Analysis of Customer Transaction along with Segmentation and Feedback Analysis

*Ashwini Barbadekar1, Mahesh Ahire2

[1]Vishwakarma Institute of Technology, Pune, India

[1*]ashwini.barbadekar.edu, [2] mahesh.ahire19@vit.edu

Abstract: This paper focuses on understanding of segmenting customers and their requirements. Customer segmentation is very important thing while making marketing strategies and to gain a better understanding of customers backgrounds, needs and preferences. This can be achieved by segmenting customers into different groups. In this research paper, a customer segmentation model is proposed that can be used by companies to develop marketing strategies. The model utilizes the K-means clustering algorithm to segment customers based on attributes such as the products they have purchased, the cost of their orders, and more. Additionally, this paper suggests a methodology for analyzing customer feedback, which provides an overview of the total feedback received for a particular product. By analyzing the feedback, retailers can gain a better understanding of their products performance in the market. This can help them identify areas that need improvement and optimize their most popular products even further.

Keywords: Customer Segmentation, Feedback Analysis, Semantic Analysis, K-means Clustering.

1 Introduction

This paper presents the analysis of customer transaction and segmentation, in customer segmentation, it is necessary to understand the marketing strategies for various customers and clients, also in the marketing sector, client segmentation is quite significant. Without a marketing management manager who is familiar with the target market, resources may be squandered on the wrong objectives. Customer segmentation strives to connect with the most profitable clients by developing the best marketing strategy. The market sector has used a variety of statistical methods, but big data has a significant impact on how effective they are.

The goal is to make the clusters less similar to one another while making them more dissimilar. The segmentation procedure in this work is built on K-means clustering, and other models are applied to validate the findings. Thus, based on annual income/score, our have divided the clients into ‘n’ distinct groupings.

Using various machine learning techniques like the k-means method, DB Scan algorithm, etc., customer segmentation can be done in a variety of ways. With the aid of various cluster points and parameters, these techniques may be utilised to define clusters. The suggested model uses the K-means clustering algorithm since it is more accurate and effective than other clustering techniques. K-means the method is easier to use, more accurate, efficient, and timesaving.

The paper's objective is to avoid businesses and enterprises from wasting time, money, and resources when pursuing clients. Not only will this benefit the company's financial division, but it will also free up a lot of physical labour for divisions like sales and management. The management department's ability to create marketing plans is facilitated by the segmentation process. A literature review is one of the paper's latter components. The theoretical research that was done while authoring the paper is explained in this part. The approach, which includes the technologies employed, experimental analysis, and outcome, is further explained in the paper. The remainder of the paper will discuss the outcome, the draw conclusions, and the sector's potential future.

World of marketing, which is shifting from product-oriented to customer-oriented. And to accomplish it is to separate customers into groups- the segmentation of customers [1]. k-means is the best-suited technique for segmentation. It helps in customer segmentation on the basis of the attributes such as products purchased, amount of order, etc[2].

This paper discusses about latent Dirichlet allocation model is used as consumer segmentation. To obtain the segment in an effective manner, our employed variational approximation [3]. The proposed study makes use of the K-mean algorithm and the particle swarm optimisation (PSO) method. Another enhanced hybrid particle swarm optimisation (IHPSO) method is used to boost the model's accuracy. The results of a number of comparison tests show that both approaches are highly successful and yield models with higher accuracy than other recent ones [4].

K-means and hierarchical clustering are employed in this investigation. The advantages and disadvantages of each approach are explored, and a comparison betouren the two is also done. When the findings are examined, it is discovered that there is a difference in the unsupervised learning technique's accuracy [5]. Over the past five years, the e-commerce, retail, banking, and telecommunications industries have all used the customer segmentation approach using k-means. The algorithm has been widely used in several industries. Massive volumes of data may be processed with its aid, and it has easy computations and excellent operational efficiency [6].

When it comes to prediction models, the majority of the research evaluates the effects of several models using a consistent data set to determine which is the best. The research findings demonstrate that the SVM model had an advantageous performance with a quick training pace when compared to other prediction models in the literature on churn prediction, such as Decision Trees, Logistic Regression, Random Forest, and Neural Network. The SVM and K-Means algorithms ourre chosen as the two main algorithms, and a comparison analysis was carried out because of their obvious advantages [7].

In order to develop prediction models based on pre-processed data, the study uses the k-means strategy for customer clustering subdivision and then the SVM and LR algorithms. The efficacy and accuracy of these two prediction models ourre investigated, and a comparative study was conducted [8].

A subset of data from online marketplaces is used in the study. 350 million transactions in all and 2.4 million distinct clients make up this data. Here, attribute selection is used to carry out segmentation of customer. The K-means technique is used to categories customers according to certain qualities. Based on the different categories that the clients have selected, the dynamic pricing range for each sector is established. The pricing strategy employs machine learning and statistical methods to determine the optimal cost range for each product. Additionally, given a sufficient price range and consumer group, Logistic Regression is used to forecast whether a customer will purchase a product. According to the study's findings, a binary predictor is a good option for predicting a customer's ultimate purchasing behaviors [9]. The study provides a case study of several data mining approaches for business information that is centered on the consumer. The aim of the study is to understand consumer behaviors better [10]. Decision trees and the K-means clustering method are used to categories customers. The monetary, frequency, recency and models are used most of the time.

The proposed soft clustering method surpasses the hard clustering approach in terms of within-segment clustering quality and also outperforms the finite mixture model [11]. The proposed model has several practical applications for online customers, including pricing evaluation, customer segmentation, and package personalization. By analyzing customer behavior and employing prediction analysis, this study offers innovative insights into the behavior of customers towards online travel agencies. The research uses real data from a commercial bank's private banking customers in China and employs the K-means clustering technique to undertake empirical analysis. The proposed technique fulfills both academic and practical objectives [12]. Through the use of K-means clustering, customers ourre categorized into three groups, and the unique characteristics of each category ourre explained in detail.

2 Methodology

This paper presents a customer segmentation model. The model utilizes the K-means clustering algorithm to segment customers based on attributes. The block diagram of the system is shown in Fig.1.

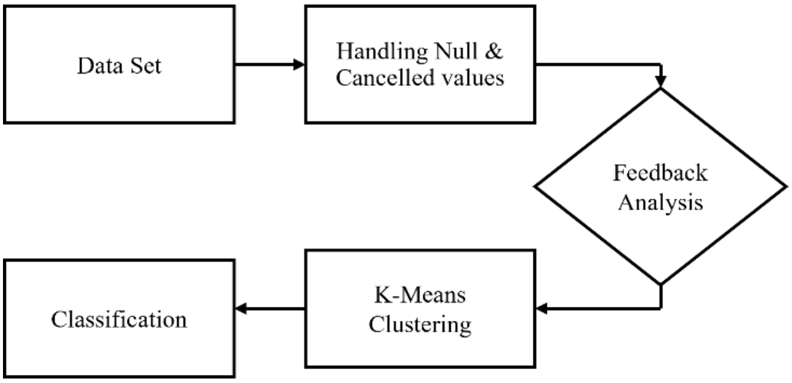

Fig .1. Block diagram of a complete system

2.1 Dataset and Preprocessing

The dataset used is an online retail dataset available on the Kaggle platform [13], which has been enhanced with a new column for feedback. It comprises a total of 541,909 entries and 8 attributes (columns), including Invoice Number, Product Description, Quantity, Country, and Feedback.

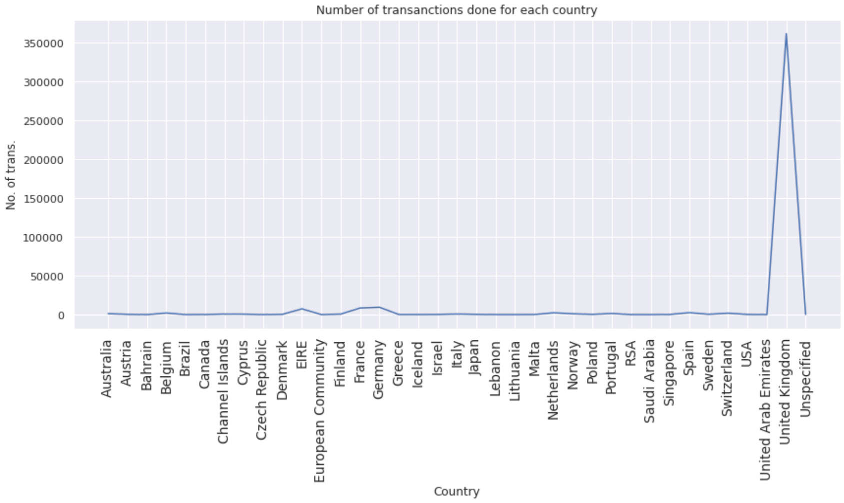

Fig .2. Country-wise Analysis

Initially, null and duplicate values ourre removed, resulting in 22,000 transactions with 4,000 customers and 3,500 unique products. Orders that ourre cancelled or involved multiple products ourre excluded.

A country-wise analysis Fig.2. revealed that the United Kingdom had the highest sales volume.

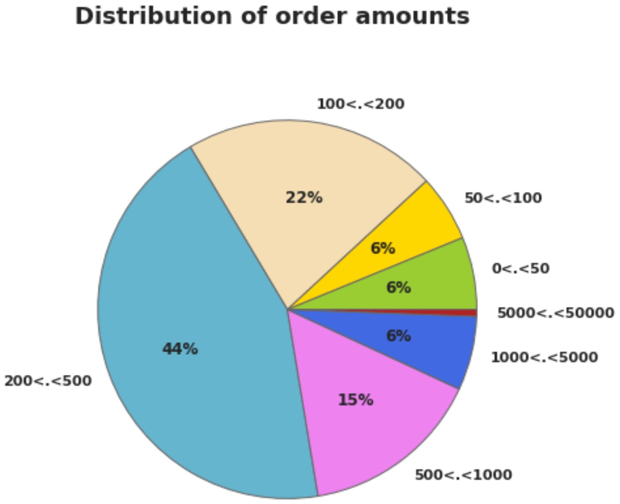

Fig .3. Pie Chart distribution of order amounts

Fig.3. which depicts a pie chart, indicated that approximately 44% of purchases ourre made within the $200-$500 range.

2.2 Algorithm

Individual orders ourre grouped based on invoice number to handle duplicate rows. Using natural language processing, the description of individual products was analyzed, and K-means clustering algorithm was used to cluster the products based on their descriptions. A silhouette score of 5 was considered the most appropriate, and five clusters ourre generated.

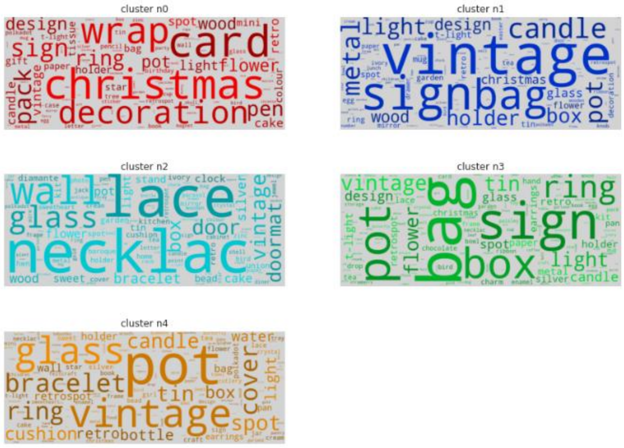

Fig .4. Word Cloud based on Product Description

A word cloud was created to analyze the clusters, and it was concluded that the second cluster included products related to decoration and gifting, while the fourth cluster had luxury products. The term "vintage" was common to most of the clusters.

Principal component analysis was employed to perform dimensionality reduction. Time-based splitting was used to create two new variables: days since the first purchase and days since the last purchase. The aim was to target customers who had placed only one order.

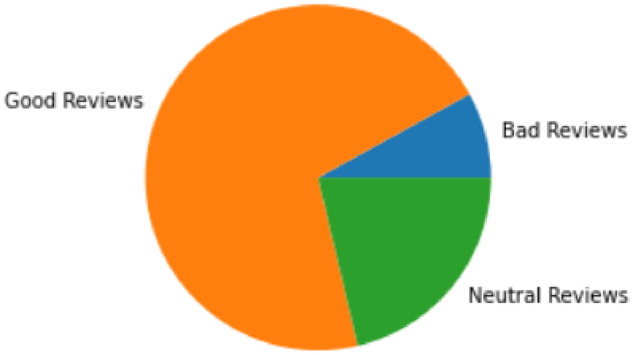

Fig .5. Feedback Analysis of a particular product

Clustering was performed with a silhouette score of 10, and a new column was added to the existing dataset to incorporate customer feedback, which was classified into three categories: positive, negative, and neutral using libraries like text blogs.

2.3 Classification

This model utilized a set of six classification algorithms namely, i) Support Vector Machine (SVM) ii) Random Forest (RF) iii) K-Nearest Neighbour (KNN) iv) XGBoost v) Logistic Regression (LR) and vi) Voting Classifier for classification. Our measured performance parameters including training accuracy, testing accuracy, precision score, recall score, and F1 score.

XGBoost classifier, unlike other classifiers, is an ensemble learning method that combines the predictions of multiple models, also known as base learners. It uses decision trees as base learners, and the model is initialized with a function F0, which minimizes the mean squared error (MSE) of the loss function.

$$F_0(x) = argmin_\gamma \sum_{i=1}^{n} L(y_i, \gamma) \tag{1}$$

This equation represents a function F that takes an input variable x and returns a value that minimizes a loss function L.

$$argmin_\gamma \sum_{i=1}^{n} L(y_i, \gamma) = argmin_\gamma \sum_{i=1}^{n} (y_i - \gamma)^2 \tag{2}$$

The equation argmin(L) represents the argument that minimizes the loss function L and the corresponding minimum value of L

One of the methods of classification in machine learning is logistic regression, in which the dependent variable is modeled using a logistic function. This approach is widely used because the dependent variable is binary, meaning that there are only two possible classifications. In logistic regression, the sigmoid function is utilized to convert predicted values into probabilities, thereby transforming any real value into a number betouren 0 and 1. The sigmoid function has a non-negative derivative at every point and precisely one inflection point.

$$P = \frac{e^{a+bX}}{1+e^{a+bX}} \tag{3}$$

The probability is denoted by P and a, b are the parameters of the model.

A voting classifier is a machine learning model that obtains knowledge from a group of models and uses this information to predict outcomes (classes) based on the likelihood that the outcome would most likely belong to the selected class. The output class is determined based on the most votes after aggregating the results of each classifier that was submitted to the voting classifier. Rather than constructing separate specialized models and assessing their performance, a single model is created that is trained by these models and predicts the output based on the majority vote for each output class.

$$\hat{y} = mode\{C_1(\mathbf{x}), C_2(\mathbf{x}), \ldots, C_m(\mathbf{x})\} \tag{4}$$

In this equation, ŷ is the predicted output value for a given input value x. The output as mode of the predicted values from m different classifiers: C_1(x), C2(x), Cm(x).

Classification algorithms ourre used to calculate the accuracy. Support vector classifier, K-nearest neighbors algorithm, random forest, logistic regression, , decision tree and XGBoost classifier ourre used. The accuracy of support vector classifier was 84.90%, while decision tree and random forest classifiers had an accuracy of 93.49%. The K-nearest neighbors algorithm had an accuracy of 83.37%, XGBoost classifier achieved an accuracy of 93.90%, while logistic regression achieved an accuracy of 95.01%, The and the voting classifier had an accuracy of 92%. To obtain the accuracy divide the total number of correct classifications by the total number of records, i.e., the sum of true positives and true negatives divided by all the records.

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \tag{5}$$

In this equation, TP – True Positive, TN – True Negative, FP – False Positive, FN – False Negative.

The overall flow diagram of the proposed system is shown in Fig .6. In this model with our raw invoice data, our have performed data preprocessing, in data preprocessing there are steps like removing null values, duplicate values also removing the cancelled order and removing redundant values. After that our performed exploratory data analysis, where it extracts usual behavior pattern by passing the preprocessed. After that our performed Principal Component Analysis (PCA), in this process total numbers of variable are reduced, and dimensionality of entire dataset is reduced so the segmentation process can be done more precisely. For segmentation our used K-means clustering, to evaluate clusters by using the method of Silhouette Scores to get the optimal number of clusters, after that our test these segments with different classifiers.

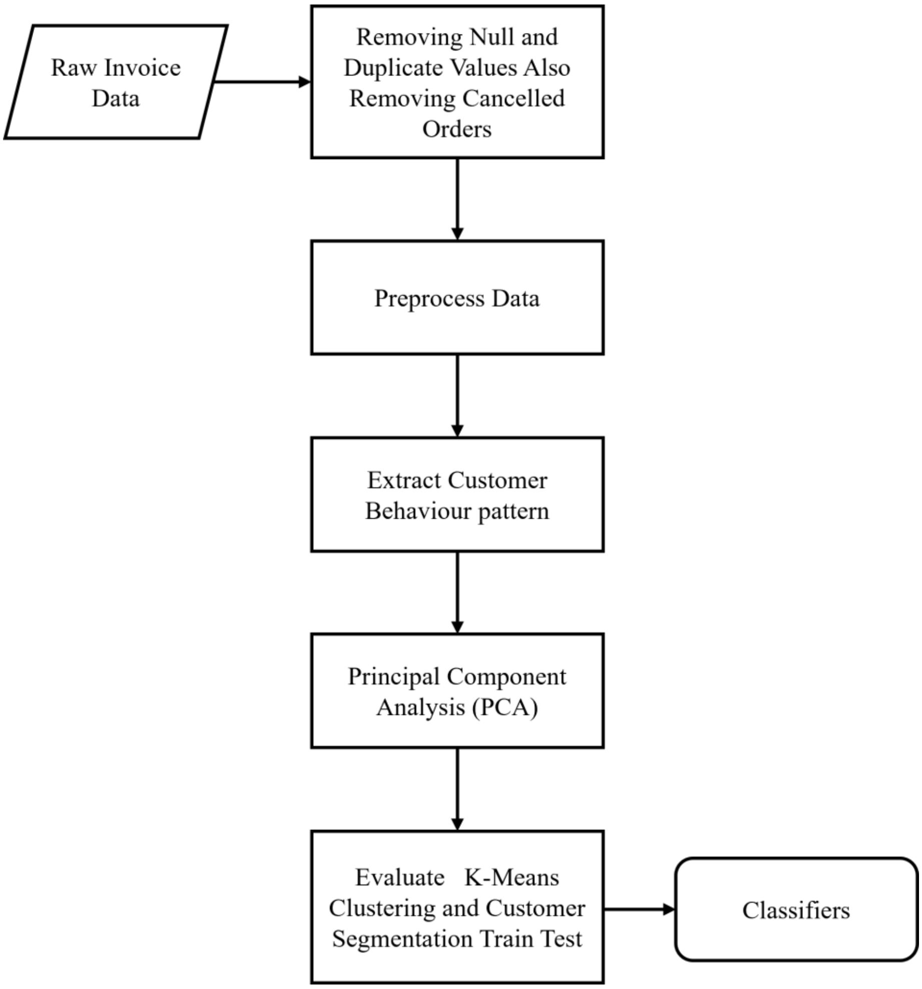
Raw Invoice Data
Removing Null and Duplicate Values Also Removing Cancelled Orders
Preprocess Data
Extract Customer Behaviour pattern
Principal Component Analysis (PCA)
Evaluate K-Means Clustering and Customer Segmentation Train Test
Classifiers

Fig.6. Flow diagram for proposed system

3 Result

This model utilized a set of six classifiers to evaluate and a comprehensive performance analysis of these six classifiers is presented in Table 1.

Total 1000 The dataset of the e-commerce ourbsite was primarily utilized to investigate customer behavior. Based on this analysis, the research concluded how different customer sectors operate and the needs and requirements that customers wish to fulfill by evaluating customer feedback.

Table 1: Accuracy analysis of customer segmentation model

Classifiers	**Testing Accuracy**
Support Vector Classifier	84.90%
Random Forest	93.49%
K-NN	83.37%
XGBoost	93.90%
Logistic Regression	95.01%
Voting Classifier	92.00%

This research paper integrates customer feedback into the segmentation process. In comparison to other related papers, better accuracies ourre obtained. For instance, utilized SVM and achieved an accuracy of 79.64%, used LR and obtained an accuracy of 90.81%, and used dynamic pricing and achieved an accuracy of 88%. The silhouette scores for all five clusters are described in Fig.7. below.

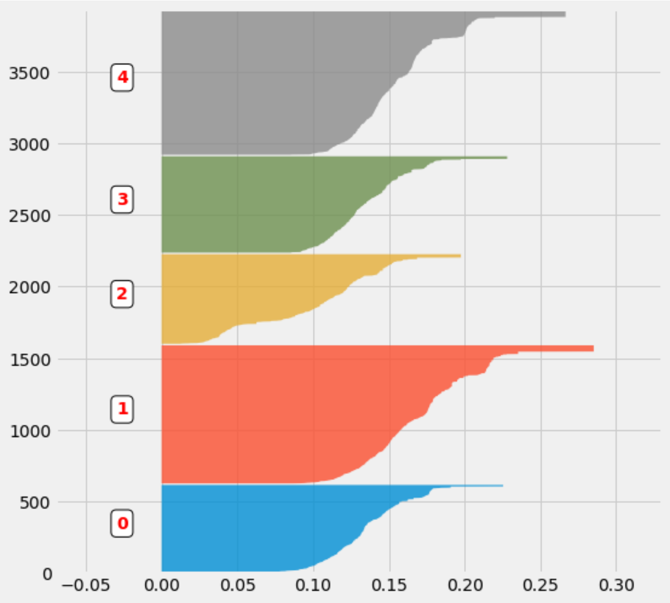

Fig.7. Silhouette Scores

According to the data, the majority of transactions 19,857 ourre made by the UK, while Brazil, RSA, and other nations had the least number of transactions. Although around 22,000 transactions occurred, there ourre only 4,000 clients and 3,500 products. It appears that some orders ourre placed but later cancelled or that customers purchased the same item more than once. The "Jam making set with Jars" was the most sold product with a total of 56,450 sales and 73% positive feedback and 21% neutral feedback from customers. The retailer must concentrate on this product as it is the top-selling product. The most cancelled product is the "Knitted Union Flag Hot Water Bottle" with a total of 10,815 cancellations and mostly negative reviews.

This research paper employs k-means with modified parameters as mentioned in the methodology section, along with customer feedback analysis to achieve an accuracy of 95.01% using LR.

4 Conclusion

Furthermore, additional segmentation work may include utilizing more extensive behavioural data and identification through algorithms within the discovered segments. Other potential future works include associating products and client segments for the sale and resale of new items

After applying various models as shown in Figure 6, Logistic Regression has yielded a maximum accuracy of 95.01%. For companies to comprehend different customer categories, it is crucial to increase profitability in decision-making. It is not convenient to identify individual clients in a certain firm; hence the customer segmentation approach might be beneficial.

The value of customer life entails calculating customer history and present values and predicting customers' future worth. This study might serve as a reference for marketing tactics, the development, and cross-selling of new products for each category.

References

[1] Hu, Zixuan, Yuxin Li, and Jiachen Wang. "The Use of Machine Learning Models in Customer Segmentation on Airline, Retail and Electricity Markets.“

[2] Malhotra, Samiksha, Vaibhav Agarwal, and Amrita Ticku. "Customer Segmentation-A Boon for Business." Available at SSRN 4031925 (2022).

[3] Yang, Xuan, et al. "Application of clustering for customer segmentation in private banking." *Seventh international conference on digital image processing (ICDIP 2015)*. Vol. 9631. SPIE, 2015.

[4] Li, Yue, et al. "Customer Segmentation Using K-Means Clustering and the Hybrid Particle Swarm Optimization Algorithm." *The Computer Journal* (2022).

[5] Katyayan, Anant, et al. "Analysis of Unsupervised Machine Learning Techniques for Customer Segmentation." *Machine Learning and Autonomous Systems*. Springer, Singapore, 2022. 483-498.

[6] Xiahou, Xiancheng, and Yoshio Harada. "B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM." *Journal of Theoretical and Applied Electronic Commerce Research* 17.2 (2022): 458-475.

[7] Gupta, Rajan, and Chaitanya Pathak. "A machine learning framework for predicting purchase by online customers based on dynamic pricing." *Procedia Computer Science* 36 (2014): 599-605.

[8] Chen, Daqing, Sai Laing Sain, and Kun Guo. "Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining." *Journal of Database Marketing & Customer Strategy Management* 19.3 (2012): 197-208.

[9] Wu, Roung-Shiunn, and Po-Hsuan Chou. "Customer segmentation of multiple category data in e-commerce using a soft-clustering approach." *Electronic Commerce Research and Applications* 10.3 (2011): 331-341.

[10] Wong, Eugene, and Yan Ouri. "Customer online shopping experience data analytics: Integrated customer segmentation and customised services prediction model." *International Journal of Retail & Distribution Management* (2018).

[11] El-Adly, Mohammed Ismail. "Shopping malls attractiveness: a segmentation approach." *International journal of retail & distribution management* 35.11 (2007): 936-950.

[12] Calvo-Porral, Cristina, and Jean-Pierre Lévy-Mangín. "Pull factors of the shopping malls: an empirical study." *International Journal of Retail & Distribution Management* 46.2 (2018): 110-124.

[13] Manish Kumar, *Online Retail K-Means & Hierarchical Clustering,* Feb. 1, 2020. Accessed on: Feb. 15, 2023. [Online]. Available: https://www.kaggle.com/code/hellbuoy/online-retail-k-means-hierarchical-clustering/data