Opportunity Finder & Keyword Trend Analysis in E-Commerce

NIKHIL JHA[1], KANISHK BHADAURIA[2], APARNA JHA[3]
[1, 2, 3] *Student, Department of Computer Science & Engineering, Maharaja Agrasen Institute of Technology, Delhi, India*

Abstract- ***E-Commerce has become increasingly popular in recent years and is now an essential part of daily life for many people. It offers customers convenience, the ability to compare prices, and the option to shop without physically travelling to stores. Consequently, for businesses and sellers, the need for a reliable method to organize the customer, group the ones with similar characteristics to satisfy their demands, and form new business strategies accordingly is much needed. Using content-based filtering for shortlisting products helps in the market analysis of a particular. Based on that analysis and the trends observed, the customers can be segmented in two ways: manually using RFM analysis or using K-means clustering, a machine learning algorithm. Segmenting customers helps to understand them better and increases a company's revenue. It is a valuable tool for businesses to understand their customer base better and tailor their marketing and product development strategies. (Eg. R. Punhani, et al. 2021)***

Indexed Terms- ***E-commerce, Cluster analysis, Customer Segmentation, RFM analysis, K-means algorithm, business decisions.***

I. INTRODUCTION

In today's market, personalization is becoming increasingly important for businesses. It allows companies to stand out and get creative in their efforts to acquire and retain customers. To improve the competitiveness of enterprises, enterprises need to quickly adapt to the rapidly changing market demands and continue to take the corresponding measures to attract customers. Therefore, accurate market segmentation and differentiated marketing strategies are the difficulties that enterprise marketing must face at present [1]. One way to personalise a business's approach is through customer segmentation, which involves dividing customers into groups based on their characteristics, needs, geographic location, demographic characteristics, behaviour, and psychological traits. This can help businesses make informed decisions about new features, products, pricing, and marketing strategies. However, manual customer segmentation can be time-consuming and difficult. One solution is to use machine learning to automate the process, allowing businesses to segment their customers and better personalise their approach.

For manual customer segmentation, RFM (Recency, Frequency, and Monetary) analysis is done which generates a score by evaluating these three properties, and then customers are manually segmented into categories like the lost customer, new customer, loyal customer, active customer, etc. K-means clustering is another method that can be used for customer segmentation. We use various methods to identify the most suitable value of k for a particular product and then divide the customers into k clusters. The goal of customer segmentation is to better understand the needs and preferences of different groups of customers so that organisations can develop more targeted product development and marketing strategies. By segmenting their customers, businesses can more effectively meet the needs of different customer groups and improve their overall marketing efforts.

Content-based filtering is a method of recommending items to a user based on the features of items that the user has chosen, the seller in this case. It uses certain methods to identify items with similar features and recommend those to the user. Three main similarity metrics can be used for this purpose:-

1. Cosine Similarity: It is a measure of similarity between two vectors in an inner product space. It is

commonly used in text analysis to measure the similarity between texts. First, the text is converted into vectors using either TF (term frequency) or TF-IDF (term frequency-inverse document frequency). In this case, TF-IDF is used because it is good for search query relevance whereas TF is used for text similarity. Cosine similarity is calculated by taking the dot product of the vectors and dividing it by the product of the magnitudes of the vectors. The resulting value is the cosine of the angle between the vectors, which ranges from -1 to 1. A value of 1 indicates that the vectors are pointing in the same direction, a value of 0 indicates that the vectors are orthogonal (perpendicular) to each other, and a value of -1 indicates that the vectors are pointing in opposite directions.

2. Jaccard similarity: It is a measure of the similarity between two sets, calculated as the size of the intersection of the sets divided by the size of the union of the sets. To calculate Jaccard similarity, it is first necessary to perform lemmatization, which reduces words to their root form. This is done to ensure that words with the same meaning but different inflections are treated as the same word so that their presence or absence in the sets being compared does not affect the similarity calculation. Once lemmatization is completed, the Jaccard similarity can be calculated by dividing the size of the intersection of the sets by the size of the union of the sets. This measure can be useful for determining the similarity between two texts or for comparing the overlap between two sets of data.

3. Euclidean distance: It is a measure of similarity used in many fields, including machine learning. It is defined as the square root of the sum of the squared differences between the values of two points in n-dimensional space. The Euclidean distance between two points is a measure of the "straight-line" distance between those points and is commonly used as a measure of similarity between two items.

RFM analysis is a marketing technique used to segregate customers by analyzing three key characteristics: recency (R), frequency (F), and monetary value (M). To calculate these values, businesses look at how recently a customer made a purchase, how often they make purchases, and how much they spend in total. Customers with high R, F, and M values are considered the most valuable and targeted for marketing efforts. This technique allows businesses to focus their efforts on specific customer segments rather than trying to reach out to their entire audience. The scores of all three variables are consolidated as RFM scores ranging from 555 to 111. *(Eg. Haiying and Yu, et al. 2010)*.

It has been observed that the scores of three factors Recency, Frequency, and Monetary are directly proportional to a customer's lifetime and retention. Here, in this project, we can segment our customers based on their RFM values manually into categories like *'STARS', 'NEW', 'BIG SPENDER', 'LOYAL', 'ACTIVE', 'LOST', 'LIGHT',* and *'REGULARS',* etc.

The K-means clustering algorithm is an unsupervised machine learning algorithm used to solve data clustering problems that are effective for segmenting customers into groups based on shared characteristics. Clustering is a method for finding cluster structure in a data set that is characterised by the greatest similarity within the same cluster and the greatest dissimilarity between different clusters. *(Eg. Sinaga, Yang, et al. 2020).* In the K-means algorithm, a target number K of centroids is defined, which represents the centres of the clusters. The algorithm then assigns each data point to the nearest cluster, while keeping the centroids as small as possible, by minimising the in-cluster sum of squares. The name 'K-means' refers to the fact that the algorithm uses averaging to find the centroids. The output of the algorithm is the centroids of the clusters and the labels for each data point indicating which cluster it belongs to.

Determining the optimal number of clusters for a K-means clustering model is an important task, as the model's performance depends on this choice. While the k-means algorithm may converge for any value of K, not all values of K will produce the best model. Data visualization can sometimes be used to identify the optimal number of clusters, but this may not be possible for all datasets. The three methods used for the same are

1. Elbow method: It is a technique for determining the optimal number of clusters for a k-means model

by evaluating the spread of the clusters from one another. To do this, the k-means algorithm is run for multiple values of K, and the within-cluster sum of square values is calculated for each value. These values are plotted on a graph, and the optimal number of clusters is chosen as the point at which adding a cluster does not significantly change the sum of square values.

2. Average Silhouette method: It is a way to evaluate the quality of clustering by measuring how well each data point fits into its assigned cluster. A high average silhouette width indicates better clustering. This method can be used to compare different clusterings or to tune the number of clusters in a k-means model.

3. Gap statistic method: It's a technique for determining the optimal number of clusters for a k-means model by comparing the total intracluster changes for different values of K to their expected values using a null reference distribution of data points. The optimal number of clusters is the value of K that maximises the gap statistic. This method can compare different clusterings or tune the number of clusters in a k-means model.

Hyperparameter tuning is adjusting the hyperparameters of a machine-learning model to improve its performance. Hyperparameters are parameters that are not learned from data, but rather are set by the practitioner. They can have a significant impact on the accuracy and performance of a model, and so finding the optimal values for them is an important step in the development of any machine learning model. Various techniques can be used to tune hyperparameters, such as grid search, random search, and Bayesian optimization. Overall, hyperparameter tuning can be very beneficial in improving the accuracy of a machine-learning model.

II. LITERATURE REVIEW

E-commerce involves various tools and technologies such as mobile shopping and online payment encryption that enable online buying and selling. Many companies use an online store or platform to conduct e-commerce marketing and sales and to manage logistics and fulfilment. [1] E-commerce plays an essential role in advancing information technology and communication, particularly in the economy. Globalisation has made markets more international and competitive, and e-commerce can aid the economy on a local scale during pandemics that impact global trade. Improvements in the internet and logistics have allowed businesses to buy, sell, and communicate on a global scale, leading to increased interest in e-commerce.

The many advantages of e-commerce make it attractive to both companies and customers. [2] As consumers shift their shopping preferences to e-commerce, the negative effects are easily seen and reported on, but the benefits of this shift are less visible. Technological innovation has positive impacts on society, and there are some unseen benefits of e-commerce growth as well as the costs such as packaging and waste, traffic and emissions, and energy and resource consumption.

Customer segmentation allows e-commerce brands to target specific groups with personalised marketing campaigns, which is more effective than using untargeted methods. E-commerce brands can improve the effectiveness of their marketing campaigns by using customer segmentation to target specific groups with personalised messages rather than relying on non-targeted methods.
For clustering, we first needed to identify the products and find similar products. [3] A similarity measure is used to compute the similarity between users or items, but there are multiple methods, each with its own limitations. Text is different from numbers and coordinates, so it is not possible to compare the differences directly. However, a similarity score can be calculated for them.

Various techniques were used for finding the similarity but the best was Euclidean distance. [4] Segmenting customers of a business into groups based on their behaviour, using the RFM (Recency, Frequency, and Monetary) values is an efficient way to analyse customer's behaviour. It's accomplished by analyzing the transactional data of a company over a specific period. This helps in understanding customer needs and identifying potential customers. Dividing customers into segments also increases the company's revenue. It's believed that retaining customers is more

important than acquiring new ones.

RFM analysis is first performed on transactional data and then is extended to clustering which also helps a company to deploy specific marketing strategies to retain customers. [7] Innovation is a key aspect of modern business and how companies often struggle to determine which products to sell and to which customers. Machine learning can be used to better understand customer behaviour and make more informed decisions by using techniques like customer segmentation, which divides customers into groups with similar behaviours. This can help companies target specific segments of customers more effectively.

There are different factors to consider in customer segmentation such as demographic, psychographic, behavioural, and geographic. [5] The clustering algorithm is used to analyse the purchase behaviour of an E-commerce system and optimise the experimental similarity within the cluster and maximise the dissimilarity between clusters. The proposed approach helps vendors to identify and focus on the highly profitable segment and retain customers for the long term.

The K-Means clustering algorithm is used to process and segment the collected data, as it's efficient to solve clustering problems. [6] The article discusses how the growth of e-commerce creates competition among companies and highlights the importance of understanding customer behaviour and characteristics in order to identify potential customers, create effective strategies, manage customer relationships, and increase profitability. The article suggests that a clustering method with K-Means algorithm can be used to segment customers based on transaction history data. However, determining the optimal number of clusters randomly doesn't always give good results, so the Elbow, Silhouette and Gap static methods are used to improve the results.

The importance of choosing the right clustering algorithm and parameters when approaching a clustering problem and how it can be challenging due to the unsupervised nature of clustering algorithms. [8] The study proposes a framework for semi-automated hyperparameter tuning of clustering problems using a grid search to develop easy to interpret metrics that can then be used for more efficient domain-specific evaluation. The results show that internal metrics are unable to capture the semantic quality of the clusters developed and approaches driven by internal metrics would come to different conclusions than those driven by manual evaluation.

III. RESEARCH & APPROACH

A. Content Based Filtering

Content-based filtering is a method of recommending items to users in an e-commerce setting based on the features or content of the items themselves. It uses a set of features or "content" to describe each item, such as text, image, or video, and then compares those features to a user's past behaviour or preferences in order to make recommendations. The main idea behind content-based filtering is that similar items should be recommended to users who have shown an interest in similar items in the past.

One of the key advantages of content-based filtering is that it does not require any information about other users or their behaviour. This makes it relatively simple to implement, and it can also be effective for recommending items to new users who do not have any past behaviour history. In an E-commerce setting, content-based filtering can be used to recommend products to users based on their past browsing and purchase history. It can be used to suggest products similar to those that a user has previously looked at or purchased, or to suggest products that are similar to items that the user has shown an interest in. Content-based filtering can also be used to recommend products to users based on specific features such as colour, material, or brand. This type of filtering can be particularly useful for recommending clothing and fashion items, or for recommending products in a specific category or niche.

Another use case is recommending items to users in a news or media setting, such as a news website or social media platform, where articles or videos can be recommended to users based on the topics or keywords they have shown an interest in in the past.

In summary, the importance of content-based filtering is that it allows for personalised recommendations for each user, increasing the chances of users finding what they are looking for, which in turn could increase the chances of them converting to a purchase. As a result, it can increase the customer satisfaction and revenue for an e-commerce website.

1. Cosine Similarity

Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them. It is commonly used in natural language processing and information retrieval to compare the similarity of two texts or documents.

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Eqn 1: Cosine Similarity

In the context of clustering, cosine similarity can be used as a similarity metric to group similar documents together. Clustering is the process of grouping similar objects together in order to discover underlying patterns and relationships in the data.

Cosine similarity can be used as a measure of similarity between documents to group them into clusters of similar documents. The basic idea behind using cosine similarity for clustering is that documents that have similar content will have similar vectors in vector space representation, thus the angle between them will be small, resulting in a cosine similarity close to 1. Conversely, documents that are dissimilar will have vectors pointing in different directions, resulting in a cosine similarity close to 0. For instance, for k-means clustering algorithm, which is one of the most popular clustering algorithms, Cosine similarity can be used as a distance metric instead of euclidean distance; this approach is known as k-means clustering with cosine similarity.

It's important to note that, for high-dimensional sparse data, Euclidean distance-based methods can be sensitive to the curse of dimensionality. Cosine similarity based approaches are less sensitive to the curse of dimensionality and have been found to work well in practice in document clustering and recommendation systems.

Another thing to notice is that Cosine similarity is not a proper distance measure, it doesn't obey all the distance metric properties like non-negativity, triangle inequality and identity of indiscernibles.

2. Euclidean Distance

Euclidean distance, also known as L2 distance, is a measure of the straight-line distance between two points in Euclidean space. It is defined as the square root of the sum of the squared differences between the coordinates of the two points. Euclidean distance is commonly used as a distance metric in many machine learning algorithms, including clustering algorithms.

$$Euclidean\,(A, B) = \sqrt{\sum_{i=1}^{N} (fa_i - fb_i)^2}$$

Eqn 2: Euclidean Distance

In the context of clustering, Euclidean distance is often used to measure the similarity or dissimilarity between data points. Clustering algorithms such as k-means and hierarchical clustering use Euclidean distance as a measure of similarity to group similar data points together in the same cluster. K-means algorithm, for example, is an iterative algorithm that partitions a dataset into k clusters by iteratively reassigning each data point to the cluster with the nearest mean. In the K-means algorithm, Euclidean distance is used to compute the distance between a data point and the cluster mean. Euclidean distance can also be used as a similarity metric in other machine learning algorithms such as classification, anomaly detection, and dimensionality reduction.

In a classification problem, for instance, Euclidean distance can be used to measure the distance between a new data point and the decision boundary or the distance between a new data point and the nearest training example of each class. It's worth noting that Euclidean distance is sensitive to the curse of dimensionality which occurs when the number of

dimensions increases, and the data becomes increasingly sparse. For high-dimensional data, alternative distance measures like Cosine Similarity or Manhattan Distance are preferred over Euclidean distance as they are less sensitive to dimensionality. Additionally, it's important to keep in mind that Euclidean distance is not the only distance metric that can be used in clustering and machine learning algorithms, it depends on the problem and the type of data you have, different distance metric may work better depending on the situation.

3. Jaccard Similarity

Jaccard similarity, also known as Jaccard coefficient, is a measure of similarity between two sets of data.

It is defined as the size of the intersection of two sets divided by the size of their union. Jaccard similarity is commonly used in information retrieval, natural language processing, and data clustering to measure the similarity between two sets of data.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

Eqn 3: Jaccard Similarity

In the context of data clustering, Jaccard similarity is often used as a measure of similarity between sets of data points, with each data point being represented as a set of features. Jaccard similarity can be used to group similar data points into clusters by identifying the data points with the highest similarity values and assigning them to the same cluster. Jaccard similarity is particularly useful when working with binary data, where each feature of a data point is either present or absent. It's important to notice that, Jaccard similarity is only defined for binary attributes, for non-binary attributes other similarity measures such as Cosine similarity or Euclidean distance may be more appropriate.

Cosine Similarity	Euclidean Distance	Jaccard Similarity
The similarity measure between two non-zero vectors, measuring the cosine of the angle b/w them.	The measure of the straight-line distance between two points in an Euclidean space.	The measure of similarity between two sets of data.
The range varies from 0 to 1. If the score is 1, it means that they are the same in orientation (not magnitude).	The Score means the distance between two objects. If it is 0, it means that both objects are identical.	The range lies from 0 to 1. If the score is 1, it means that they are identical.

Table 1: Comparing- Cosine Similarity, Euclidean Distance & Jaccard Similarity

B. Customer Segmentation

Customer segmentation is the process of dividing a customer base into smaller groups of individuals that have similar needs or characteristics. In the context of e-commerce, this can help businesses to better understand and target their desired audience, and tailor There are a variety of ways to segment customers in e-commerce, such as demographics, behaviour, and purchase history. Demographic segmentation can include factors such as age, gender, income, and location. Behavioural segmentation can include factors such as past purchases, browsing history, and search queries. Purchase history segmentation can include factors such as purchase frequency, purchase history, and lifetime value. Once customers have been segmented, businesses can develop targeted marketing campaigns, personalised product recommendations, and specialised customer service to better serve each segment. This can help to increase customer retention, as well as attract new customers who belong to the targeted segment.

Personalised product recommendations can be made through purchase history segmentation, showing products that customers have shown interest in before. Additionally, businesses can use specialised

customer service to address specific concerns of each segment, such as offering a loyalty program for frequent customers or providing extra support to first-time customers. Overall, customer segmentation in e-commerce can help businesses to increase customer retention and attract new customers who belong to the targeted segment. Additionally, it can help to improve the customer experience and drive sales by providing tailored marketing, product offerings, and customer service.

1. Manual Segmentation using RFM Analysis

Manual segmentation is a method of dividing customers into groups or segments based on specific characteristics or attributes. This is typically done by manually analyzing data, such as customer demographics, purchase history, or behaviour, and then grouping customers together based on similarities

The method involves dividing customers into groups based on three key factors:

Recency: This refers to the last time a customer made a purchase. Customers who have made a purchase recently are considered more valuable than those who have not made a purchase in a long time.

Frequency: This refers to how often a customer makes a purchase. Customers who make frequent purchases are considered more valuable than those who make infrequent purchases.

Monetary: This refers to the amount of money a customer has spent. Customers who have spent more money are considered more valuable than those who have spent less.

To perform RFM analysis, a business will typically collect data on the purchase history of its customers, including the date of each purchase, the number of purchases, and the total dollar value of each purchase. The data is then used to create a score for each customer based on the recency, frequency, and monetary values. These scores are then used to segment customers into groups. Once customers have been segmented using RFM analysis, businesses can develop targeted marketing campaigns, personalised product recommendations, and specialised customer service to better serve each group. For example, businesses can use RFM analysis to identify customers who have recently made a purchase and offer them a special promotion or discount to encourage them to make another purchase.

On the other hand, businesses can use the same analysis to target customers who have not made a purchase in a while and send them re-engagement campaigns to bring them back. RFM analysis is considered a manual method of segmentation because it requires manual data collection and calculations to create the scores for each customer. It is a relatively simple and easy-to-implement method of segmentation that can be a good starting point for businesses that are new to customer segmentation. However, keep in mind that RFM is a simple method which doesn't take in account many factors that other modern and complex methods do.

2. K-Means Clustering

K-means is a popular clustering algorithm that groups similar data points together. The "k" in k-means refers to the number of clusters to be created. The algorithm works by first randomly initialising k "centroids", which are the centre points of each cluster. Then, each data point is assigned to the cluster whose centroid it is closest to. Next, the centroid for each cluster is re-calculated as the mean of all the data points.

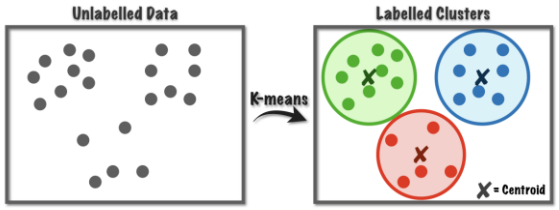

Fig 1: K-means Clustering representation

This process is repeated until the assignments of data points to clusters no longer change or a maximum number of iterations is reached. One of the advantages of k-means is that it is computationally efficient and easy to implement. However, the algorithm can be sensitive to the initialization of centroids and may not always converge to a global optimum. Also, it assumes that clusters are spherical

in shape, so it may not be suitable for data with non-uniformly sized clusters. It's commonly used for a variety of use cases such as market segmentation, image compression, and anomaly detection. The K-means algorithm is a method of vector quantization, which is a technique for approximating a multidimensional probability density function by a mixture of simpler distributions. In the case of k-means, the "simpler distribution" is a multivariate normal distribution with a diagonal covariance matrix.

The K-means algorithm begins by randomly initialising k centroids, which are the centre points of each cluster. The centroids can be initialised in various ways, such as randomly selecting k data points from the dataset, or using some heuristic method. Once the centroids are initialised, the algorithm proceeds through the following steps:

1. Assign each data point to the cluster whose centroid it is closest to, according to some distance metric (such as Euclidean distance).
2. Recalculate the centroid of each cluster as the mean of all the data points in the cluster.
3. Repeat steps 1 and 2 until the assignments of data points to clusters no longer change, or a maximum number of iterations is reached.

The goal of the k-means algorithm is to minimise the sum of squared distances between each data point and its closest centroid. This criterion is called the within-cluster sum of squares (WCSS), and it is used as a measure of cluster compactness.

There are some variants of k-means, such as the k-medoids algorithm, which uses the medoid (i.e., the most centrally located data point) instead of the mean as the representative of a cluster.

One of the main advantages of k-means is that it is computationally efficient, especially for large datasets, because it requires only O(tkn) computations to classify n points among k clusters in t iterations, where t is the number of iterations. However, k-means also has some drawbacks, such as being sensitive to the initialization of centroids, and assuming that clusters are spherical in shape and have roughly the same size, which may not be the case for many datasets. Additionally, k-means does not work well with categorical variables, so these must be transformed to numerical values before using this method. In the case of k-means, the time complexity is often given as O(tkn), where t is the number of iterations and n is the number of data points, and k is the number of clusters. This means that the running time of k-means is at most proportional to the product of t, n, and k. In other words, as the number of iterations, data points or clusters increase, the running time of the algorithm increases as well.

However, the actual time complexity of k-means can depend on the specific implementation and the data characteristics. Also, the O(tkn) is just an estimate of the time complexity of k-means. The actual running time of k-means can depend on the specific implementation, the data characteristics, and the hardware used. Some variations of the algorithm can have different complexity estimates.

RFM	K Means
RFM (Recency, Frequency, Monetary) analysis is a customer segmentation technique that is used to identify the value of a customer to a business.	K means clustering is a machine learning algorithm that is used to group data points into clusters based on their similarity.
It is based on the idea that the more recently a customer has made a purchase, the more likely they are to make another purchase soon.	Divides the data into a specified number of clusters and iteratively assigns each data point to the cluster with the nearest mean.
RFM analysis involves calculating these three values for each customer and then using them to segment customers into different groups based on their value to the business.	K means clustering is often used for customer segmentation, but it is a more general technique that can be applied to any type of data.

Table 2: Comparing- RFM Analysis & K-Means Clustering

IV. PROPOSED METHODOLOGY

In this research, we present a novel approach for generating keyword analysis for sellers using content based filtering techniques and customer segmentation. The dataset under observation was acquired from a UK based retail company which consisted of purchase history of various customers.

First, the entire dataset was narrowed down to the products that matched a keyword (here, a product description). For this purpose, we used content-based filtering also known as text similarity. It uses the text of an item, such as a document, article, or a product description to recommend similar items. This is done by comparing the text of the item to other items in a database, and finding those with the most similar text. This process can be done using various techniques such as the vector space model, cosine similarity, Jaccard similarity, and others. But here we are using majorly two techniques:

Cosine Similarity and Euclidean Distance. The Text similarity model generates a list of items that are most similar to the input item, which can then be utilised to narrow down the dataset to the purchase history of only those customers who have bought similar items. This allows us to target them as a market segment for our further marketing campaigns.

We then carried out customer segmentation on this filtered dataset. There are several techniques available for customer segmentation, including Demographic Segmentation, Behavioural segmentation, etc. In order to analyse purchasing behaviour, we employed two methods: RFM Analysis and K-means Clustering. Manual segmentation using RFM analysis is a simple yet effective method to understand customer behaviour, which can be useful for businesses to target their marketing efforts, improve customer retention and increase revenue. Using a clustering algorithm to group customers based on their similarities in their behaviour and attributes is another effective way which we implemented for our research.

We evaluated the segments generated by our proposed approach using various cluster sizes. For cluster size, we used the Elbow method, Gap Statistic method and the Silhouette method to determine the optimal value of size k and then used the average of these three values. The results of our experiments showed that clustering customers around a product description will allow the sellers to draw some very interesting conclusions about the effectiveness of their marketing strategy for that particular product.

One challenge that we faced during our research was the implementation of Jaccard Similarity because different sized sets with the same number of common members also will result in the same Jaccard similarity and the overall computational cost of it was very high in our method. Our proposed solution showed the potential of using RFM Analysis and K-Means Clustering in combination with a content-based filtering system to generate segments based on a product chosen by a seller. We believe that this approach can be applied to various real-world applications and hope that it will inspire further research in this field.

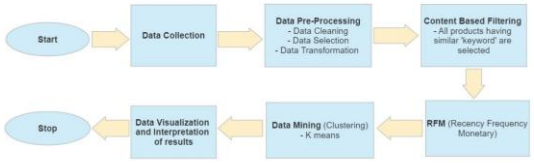

Fig 2: Architecture of the proposed methodology

V. APPROACH

A. Dataset

For this study, we have used a UK based registered non-store online retail company's transactional dataset.

1. The transnational data set contains all the transactions occurring between the time period, starting from 01/12/2010 until 09/12/2011.
2. It's based on a UK-based and registered non-store online retail store, which mainly sells unique all-occasion gifts. Many customers of the company include wholesalers.
3. Analyses for this dataset could include time series, clustering, classification and more.
4. The dataset contains 8 columns, namely: Invoice Number, Stock Code, Description, Quantity, Invoice Date, Unit Price, Customer ID and the

Country.

5. As per the UCI Machine Learning Repository, the data has been made available by Dr Daqing Chen, Director: Public Analytics group of School of Engineering, London South Bank University, London SE1 0AA, UK.

B. Initial EDA

We obtained the training dataset for our two segmentation techniques from the UCI Machine Learning Repository. In the initial EDA on the dataset, we checked for the missing values and removed the duplicates. This involved identifying any null or empty values in the dataset and either dropping or imputing those values as necessary. We also dropped any duplicate rows in order to ensure that the data was clean and accurate for further analysis. It was discovered that the main market for the retail company was based in the United Kingdom. However, there were also a small number of buyers from France, Germany, Belgium, and Ireland. This information can help to inform future marketing and sales strategies for the company, as well as provide insight into the demographics and purchasing habits of the company's customers.

C. Training

In order to match similar products in the filtering process, we used TF-IDF scores to rank the importance of words in a document and converted each unique product into a vector of scores. Cosine Similarity and Euclidean Distance were then used to predict the 100 most similar products. We then calculated Recency, Frequency, and Monetary means for each customer who bought a similar product and manually segmented them into some groups.

Such as *'ACTIVE'*, *'BIG SPENDER'*, *'LIGHT'*, *'LOST'*, *'LOYAL'*, *'POTENTIAL'*, and *'STARS'*. Additionally, a K-Means clustering model was trained, ran for 10 iterations with different centroid seeds, in order to cluster the data further. For the optimal value of cluster size k, we took the average of the values obtained from three different methods - Elbow method, Gap Statistic method and the Silhouette method. For this dataset the value of k obtained was 3. To avoid outliers, we removed the customers whose z-score of RFM means were outside bounds i.e, >3 or <-3.

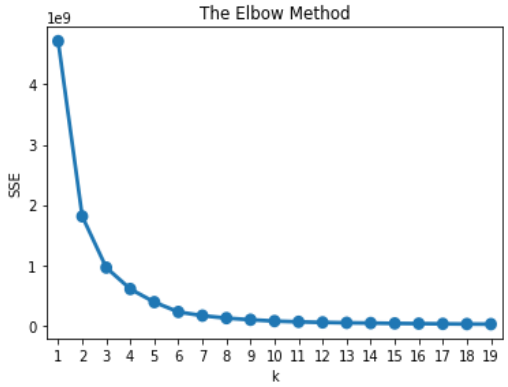

Fig 3: Graphical representation of the implemented Elbow method

VI. RESULTS

A. Recorded Metrics

We visualized the performance of our model using a streamlit application. We used a Treemap plot, a 2-D scatter plot and a 3-D scatter plot to display the RFM mean values and better understand the outcome of the two segmentation methods used - Manual Segmentation using RFM Analysis and clustering using K-Means clustering.

Apart from making sure that our dataset has a *clustering tendency* and choosing the optimal value of cluster size *k*, there are two types of measures used to evaluate the quality of a clustering algorithm: extrinsic measures that require ground truth labels, such as Adjusted Rand index and Fowlkes-Mallows scores, and intrinsic measures that do not require ground truth labels, such as Silhouette Coefficient and Calinski-Harabasz Index.

These measures can be used to evaluate the clustering performance by assessing the minimal intra-cluster distance and maximal inter-cluster distance. Since we do not know the true segment class of each customer we have used intrinsic measures to evaluate our model’s performance.

Intrinsic Measures	Extrinsic Measures
Evaluate the quality of clustering based on the data within the clusters.	Evaluate the quality of clustering based on some external criterion or external data.
Examples include silhouette coefficient, Davies-Bouldin index, and Calinski-Harabasz index.	Examples include adjusted Rand index, Normalised Mutual Information (NMI), and Fowlkes-Mallows index.
Assess the similarity of data points within a cluster and the dissimilarity of data points between clusters.	Compare the cluster labels obtained by the clustering algorithm to some external criterion, such as true class labels in a supervised learning setting.
Evaluates the internal structure of the cluster.	Evaluates the external relationship between the cluster and some external criterion.

1. Silhouette Coefficient

Silhouette score is a measure used to evaluate the quality of a clustering algorithm. It calculates the similarity of an observation to its own cluster compared to other clusters. The score ranges from -1 to 1, where a high score indicates that the observation is well-matched to its own cluster, and a low score indicates that the observation is mis-matched to its own cluster. A score of 0 means that the observation is on the decision boundary between two clusters.

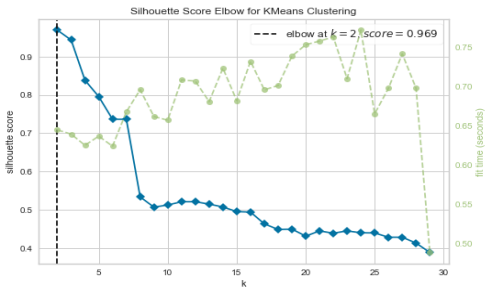

Fig 4: Graphical representation of the represented Silhouette Coefficient

The Silhouette score is calculated for each observation in the dataset and is based on two main components: the average distance between an observation and all other observations in the same cluster (a), and the average distance between an observation and all other observations in the next nearest cluster (b).

The silhouette score for each observation is then calculated as [(b-a) / Max (a,b)]. The final silhouette score for the entire dataset is the average of the silhouette scores for all observations. The measure of distance used in the calculation of a(i) and b(i) was Euclidean Distance.

We achieved a Silhouette score of 0.695 on the K-Means clustering model. A Silhouette score close to 1 suggests that the clustering algorithm is performing well and that the observations are well-matched to their own clusters. This is considered as a good score, and it is a positive indication of the clustering performance.

2. Calinski-Harabasz Index

The Calinski-Harabasz index (also known as the Variance Ratio Criterion) is a measure used to evaluate the quality of a clustering algorithm. It compares the ratio of the between-cluster variance to the within-cluster variance. A higher Calinski-Harabasz index value indicates better clustering performance. The Calinski-Harabasz index is calculated as:

$$s = \frac{\mathrm{tr}(B_k)}{\mathrm{tr}(W_k)} \times \frac{n_E - k}{k - 1}$$

Eqn 4: Calinski-Harabasz Index Formula

where,

s - is the Calinski-Harabasz index value
tr(Bk) - trace of the between-cluster scatter matrix
tr(Wk) - trace of the within-cluster scatter matrix
k - is the number of clusters
n - is the total number of observations

We were able to achieve a Calinski-Harabasz index

value of 4709.345, which suggests that the clustering algorithm is performing well. The higher the value of the index, the better is the clustering. This value indicates that the between-cluster variance is relatively high in comparison to the within-cluster variance. This suggests that the observations are well separated into different clusters and that the clusters are relatively distinct.

3. Davies-Bouldin Index

The Davies-Bouldin index (DBI) is a measure used to evaluate the quality of a clustering algorithm, it is often used to evaluate the performance of K-Means clustering algorithm.

The DBI is calculated as the average similarity between each cluster and its most similar cluster. The lower the DBI value, the better the clustering performance, as it indicates that the observations are well separated into different clusters and the clusters are relatively distinct.

The Davies-Bouldin index is an internal evaluation method, which uses properties and characteristics within the dataset to assess the quality of the clustering. This allows for the validation of the clustering performance without the need for external information.

Our model gave us a DBI of 0.557. A lower DBI value is considered to be good and is an indication that the clustering algorithm has performed well.

B. Evaluation

An analysis of the dataset using RFM segmentation revealed that a significant portion of customers for the product "4 BLUE DINNER CANDLES SILVER FLOCK" fall into the "Big Spender" category. People tend to spend more on antique dinner candles.

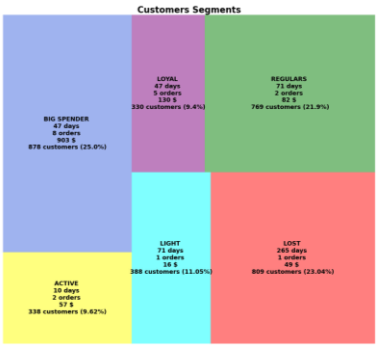

Fig 5: Treemap Plot using RFM Analysis (Manual Segmentation)

Additionally, the presence of “Loyal” and “Regular” customers suggests that the current market strategy for these products is successful and does not require adjustments for this segment.

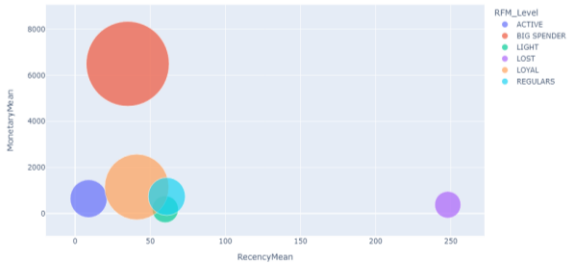

Fig 6: 2D Scatter Plot using RFM Analysis (Manual Segmentation)

An examination of the K-Means clustering model applied to the dataset for the product "4 BLUE DINNER CANDLES SILVER FLOCK" revealed that most customers fall into the "Gold Customers" category. This high concentration in a single category highlights the popularity of dinner candles and similar products in 2011. The quality of the model is considered to be good as it effectively separates customers into their respective segments. It's worth noting that no single measure can be considered the "best" for evaluating clustering quality, as it depends on the specific characteristics of the data and the goals of the clustering task. It's often a good idea to use multiple measures to evaluate the performance of a clustering algorithm.

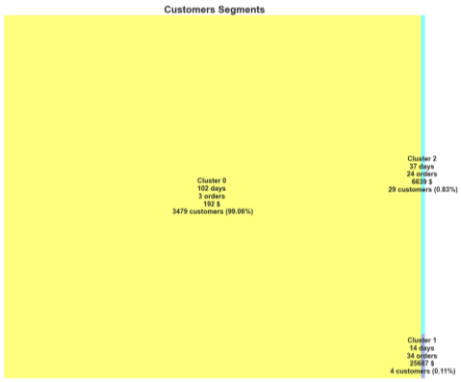

Fig 7: Treemap Plot using K-Means Clustering

CONCLUSION

Our clustering model and RFM models have been well-trained, providing valuable insights on customer segmentation. This approach allows sellers to filter products based on keywords and convert them into a market. By using this method of customer segmentation, businesses can gain valuable insights on their customers and make informed decisions on how to target them effectively in the future. This strategy can be beneficial for sellers as it will help them to target their customers with more precision and increase their sales and revenue. Additionally, the rich statistics provided by our models can assist sellers in identifying trends and patterns in customer behaviour, which can inform future marketing strategies and product development. Overall, using our clustering and RFM models for customer segmentation can give a competitive edge to the businesses in the market.

In conclusion, we found out that the K-means clustering performed on this dataset, fetches us with the customers segmenting into 3 different clusters. Cluster 0 has a high recency rate, while Cluster 1 and Cluster 2 have low recency, thus putting the latter in the race for Platinum and Gold customers. Cluster 0 has a low frequency rate, while Cluster 1 and Cluster 2 have high frequency, thus putting them in the race for Platinum and Gold customers. Cluster 0 has a low Monetary rate, while Cluster 1 has the highest Monetary rate (Platinum) whereas Cluster 2 has a medium level (Gold). Hence-

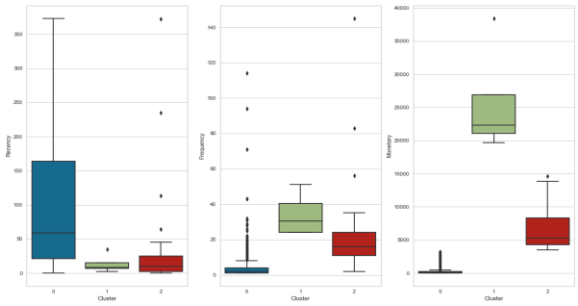

Fig 8: Boxplots for RFM Values of different Clusters

Following results were concluded after the clustering of customers:

Cluster 0	Silver Customers
Cluster 1	Platinum Customers
Cluster 2	Gold Customers

Thus, the proposed research concludes the following: Cluster 0 being the Silver customers. Cluster 1, the Platinum & Cluster 2, the Gold customers.

VII. SUMMARY

This research presents a novel approach for generating keyword analysis for sellers using content based filtering techniques and customer segmentation. The dataset under observation was acquired from a UK based retail company which consisted of purchase history of various customers. The dataset was narrowed down to the products that matched a keyword using content-based filtering, then customer segmentation was carried out on this filtered dataset using RFM Analysis and K-means Clustering. The results showed that clustering customers around a product description will allow the sellers to draw some very interesting conclusions.

The challenge faced was the implementation of Jaccard Similarity due to high computational cost. The proposed solution showed the potential of using RFM Analysis and K-Means Clustering in combination with a content-based filtering system to generate segments based on a product chosen by a seller, which can be applied to various real-world applications.

VIII. FUTURE SCOPE

The future scope of customer segmentation is vast and promising. Advancements in technology, such as machine learning and big data analytics, are making it easier for businesses to segment their customer base in new and more accurate ways. Alternative criteria for customer segmentation can be used, such as geographical location or demographic characteristics, to better understand customers' habits and preferences. Additionally, as more companies adopt a customer-centric approach to business, customer segmentation will become increasingly important for understanding and catering to the unique needs of different groups of customers. Furthermore, the use of customer segmentation in personalization and targeting of marketing efforts, customer service, and product development will continue to be a key area for future growth.

REFERENCES

[1] Taher, G., 2021. E-commerce: advantages and limitations. International Journal of Academic Research in Accounting Finance and Management Sciences, 11(1), pp.153-165.

[2] Tokar, T., Jensen, R. and Williams, B.D., 2021. A guide to the seen costs and unseen benefits of e-commerce. Business Horizons, 64(3), pp.323-332.

[3] Jain, G., Mahara, T. and Tripathi, K.N., 2020. A survey of similarity measures for collaborative filtering-based recommender systems. In Soft computing: theories and applications (pp. 343-352). Springer, Singapore.

[4] Christy, A.J., Umamakeswari, A., Priyatharsini, L. and Neyaa, A., 2021. RFM ranking–An effective approach to customer segmentation. Journal of King Saud University-Computer and Information Sciences, 33(10), pp.1251-1257.

[5] Tabianan, K., Velu, S. and Ravi, V., 2022. K-means clustering approach for intelligent customer segmentation using customer purchase behaviour data. Sustainability, 14(12), p.7243.

[6] Siagian, R., Sirait, P.S.P. and Halima, A., 2021. E-Commerce Customer Segmentation Using K-Means Algorithm and Length, Recency, Frequency, Monetary Model. JOURNAL OF INFORMATICS AND TELECOMMUNICATION ENGINEERING, 5(1), pp.21-30.

[7] Kansal, T., Bahuguna, S., Singh, V. and Choudhury, T., 2018, December. Customer segmentation using K-means clustering. In 2018 international conference on computational techniques, electronics and mechanical systems (CTEMS) (pp. 135-139). IEEE.

[8] Ditton, E., Swinbourne, A., Myers, T. and Scovell, M., 2021. Applying Semi-Automated Hyperparameter Tuning for Clustering Algorithms. arXiv preprint arXiv:2108.11053.

[9] Liu, Y., 2021. Research on E-commerce Enterprise Customer Segmentation Based on Cluster Analysis-Taking Jingdong Century Trading Co., Ltd as an Example.

[10] Anwar, T. and Uma, V., 2021. Comparative study of recommender system approaches and movie recommendation using collaborative filtering. International Journal of System Assurance Engineering and Management, 12(3), pp.426-436.

[11] Thongtan, T. and Phienthrakul, T., 2019, July. Sentiment classification using document embeddings trained with cosine similarity. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop (pp. 407-414).

[12] Gunawan, D., Sembiring, C.A. and Budiman, M.A., 2018, March. The implementation of cosine similarity to calculate text relevance between two documents. In Journal of physics: conference series (Vol. 978, No. 1, p. 012120). IOP Publishing.

[13] Xia, S., Xiong, Z., Luo, Y. and Zhang, G., 2015. Effectiveness of the Euclidean distance in high dimensional spaces. Optik, 126(24), pp.5614-5619.

[14] Vijaymeena, M.K. and Kavitha, K., 2016. A survey on similarity measures in text mining. Machine Learning and Applications: An International Journal, 3(2), pp.19-28.

[15] Bank, J. and Cole, B., 2008. Calculating the jaccard similarity coefficient with map reduce for entity pairs in wikipedia. Wikipedia Similarity Team, 1, p.94.

[16] Adolfsson, A., Ackerman, M. and Brownstein, N.C., 2019. To cluster, or not to cluster: An analysis of clusterability methods. Pattern Recognition, 88, pp.13-26.

[17] Dogan, O., Ayçin, E. and Bulut, Z.A., 2018. Customer segmentation by using RFM model and clustering methods: a case study in the retail industry. International Journal of Contemporary Economics and Administrative Sciences, 8(1), pp.1-19.

[18] Cardenas, C.E., Yang, J., Anderson, B.M., Court, L.E. and Brock, K.B., 2019, July. Advances in auto-segmentation. In Seminars in radiation oncology (Vol. 29, No. 3, pp. 185-197). WB Saunders.

[19] Sheshasaayee, A. and Logeshwari, L., 2018, May. Implementation of clustering technique based RFM analysis for customer behaviour in online transactions. In 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 1166-1170). IEEE.

[20] Deng, Y. and Gao, Q., 2020. A study on e-commerce customer segmentation management based on an improved K-means algorithm. Information Systems and e-Business Management, 18(4), pp.497-510.

[21] Punhani, R., Arora, V.S., Sabitha, S. and Shukla, V.K., 2021, March. Application of clustering algorithm for effective customer segmentation in E-commerce. In 2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE) (pp. 149-154). IEEE.

[22] Christy, A.J., Umamakeswari, A., Priyatharsini, L. and Neyaa, A., 2021. RFM ranking–An effective approach to customer segmentation. Journal of King Saud University-Computer and Information Sciences, 33(10), pp.1251-1257.