Capstone Project

PROJECT REPORT

MA-329

Submitted to:
Prof. Subit Kumar Jain

Submitted By:
Sahil Sood (20BMA011)
Gautam Kumar (20BMA015)
Albert Sharma (20BMA024)

ABSTRACT

This capstone project aims to conduct customer segmentation analysis for an online retail company, with the goal of identifying distinct groups of customers based on their purchasing behavior and demographic characteristics. The project will use clustering techniques, such as k-means clustering to group customers based on their buying patterns, such as frequency, recency, and monetary value. The project will also incorporate data visualization techniques to present the results of the analysis in an intuitive and easily understandable manner. The outcome of this project will provide the online retail company with valuable insights into the characteristics and needs of their customer groups, which can help them tailor their marketing and sales strategies to better meet their customers' preferences and increase their revenue.

PROBLEM STATEMENT

Ocado retail stock prices fluctuated during the given period, with a significant drop in the second half of 2011. In August 2011, the stock price dropped by more than 35% compared to the previous month, and by December 2011, it had dropped by more than 50% compared to the beginning of the year. This may have been due to a number of factors, such as increased competition in the online grocery market and concerns about the company's profitability.

- It is a critical requirement for businesses to understand the value derived from a customer. RFM is a method used for analyzing customer value.
- Customer segmentation is the practice of segregating the customer base into groups of individuals based on some common characteristics such as age, gender, interests, and spending habits
- Perform customer segmentation using RFM analysis. The resulting segments can be ordered from most valuable (highest recency, frequency, and value) to least valuable (lowest recency, frequency, and value).

INTRODUCTION

In the world of e-commerce, customer segmentation has become an essential strategy for companies looking to better understand their customers and improve their sales and marketing efforts. By dividing customers into distinct groups based on their behavior and demographics, companies can tailor their marketing messages and promotions to specific groups, leading to increased customer satisfaction and higher revenue. Ocado is a UK-based online grocery retailer that was founded in 2000. The company is headquartered in Hatfield and operates one of the largest online grocery delivery services in the UK. Ocado has continued to grow in popularity in recent years, and the company has partnered with major retailers such as Marks & Spencer to expand its offerings. The primary goal of this project is to provide online retail company with valuable insights into their customers' needs and preferences. By understanding the different segments of its customer base, the company can develop more targeted marketing and sales strategies that are tailored to the specific needs and preferences of each group. Ultimately, the project aims to help the company increase customer satisfaction and loyalty, as well as maximize its revenue potential.

OBJECTIVES

We analyzed the transactional data for an online retail company and create customer segmentation so that the company can create an effective marketing campaign. We performed the following tasks in this project:-

- Data Cleaning
- Data Transformation
- Data Modeling - RFM (Recency Frequency Monetary) model
- Data Modeling - K-means clustering algorithm

TECHNOLOGIES

- Python
- pandas
- numpy
- matplotlib
- seaborn
- sklearn

METHODOLOGY

The project will follow a mixed-methods approach, combining qualitative and quantitative research methods. Initially, the project will involve a detailed analysis of customer data and insights, with the aim of identifying common trends and patterns in customer behavior. This will be followed by a series of customer surveys, which will help to further understand customer preferences, pain points, and expectations when it comes to online shopping.

1. **Data Collection**: The first step in the methodology is to collect relevant data from the online retail company's database. This data should include customer purchase history, demographic information, and any other relevant variables that may affect customer behavior.

2. **Data Cleaning**: After collecting the data, the next step is to clean and preprocess it. This involves removing any missing or inconsistent data points, standardizing the variables, and normalizing the data.

3. **Feature Selection**: Once the data is cleaned, the next step is to select the relevant features to be used in the clustering analysis. These features may purchase history (such as purchase frequency, recency, and monetary value), and any other relevant variables.

4. **Choosing K**: The next step is to determine the optimal number of clusters (k) for the data. This can be done by using a combination of statistical techniques such as the elbow method or the silhouette score.

5. **Applying K-Means**: After determining the optimal value of k, the next step is to apply the k-means clustering algorithm to the data. This involves randomly initializing k centroids, assigning each data point to the nearest centroid, and then recalculating the centroids until convergence is reached.

6. **Cluster Evaluation**: This can be done by calculating metrics such as cluster cohesion, cluster separation, and silhouette scores.

7. **Visualization**: Finally, the results of the clustering analysis should be explained in an easily understandable way.

PROCESS OF DATA CLEANING

Two columns in the data had missing values.
Description - 0.27% (1454)
CustomerID - 24.93% (135080)

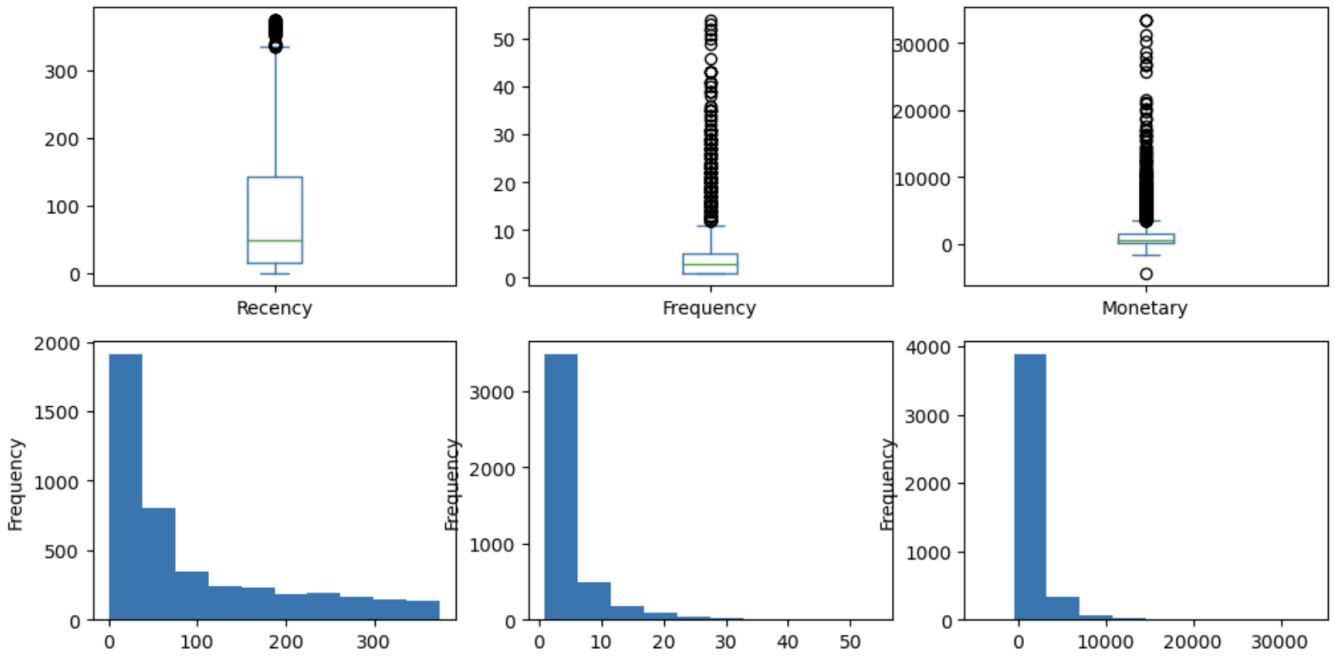
CustomerID is an important feature of our analysis since our analysis is centered around customers only so we cannot impute null values CustomerID with mean/ median/ mode in this case. We will drop the complete row having to miss CustomerID if it is not possible to track back the customerID using the Invoice Number.

We can drop the Description feature from our data since it is not going to contribute to our model.Comparing the shape of data: Shape before the data cleaning: (541909, 8) Shape after the data cleaning: (406829, 7) One column has been removed and around 140000 rows have been removed.

METADATA

- **Quantity:** Average quantity of each product in the transaction is 12.18. Also, the minimum value in the Quantity column is negative. This implies that some customers returned the product during our period of analysis.
- **InvoiceDate**: Our data has transactions between 01-12-2010 to 09-12-2011
- **UnitPrice:** Average price of each product in transactions is 3.47
- **InvoiceNo:** Total entries in preprocessed data are 4,01,602 but transactions are 22,190. Most entries (count of unique products) are in Invoice No. '576339' and is 542 nos.
- **StockCode:** There is a total of 3684 unique products in our data and the product with stock code '85123A' appears most frequently (2065 times) in our data.
- **CustomerID:** There are 4372 unique customers in our final preprocessed data. Customer with ID '17841' appears most frequently in data (7812 times)
- **Country:** The company has customers across 37 countries. Most entries are from the United Kingdom in our dataset (356726)

DATA TRANSFORMATION

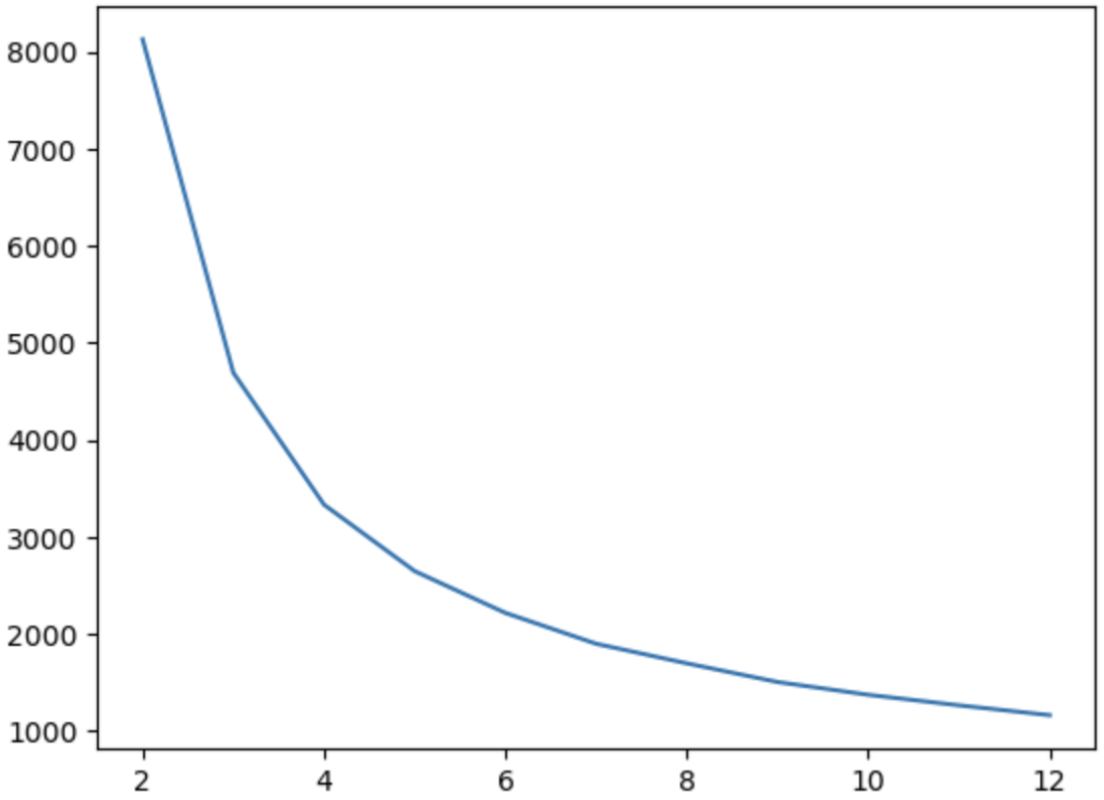


Standard Scalar Transformation: It is extremely important to rescale the features so that they have a comparable scale.

Data has to be scaled and outliers to be removed before applying the clustering algorithm. If outliers appear in the data, they significantly deteriorate the quality of clusters. It is worth emphasizing at this point that the outlier can be both a given error or information noise and real outlier data. Clustering algorithms such as K-means do need feature scaling before they are fed to the algorithm. Since clustering techniques use Euclidean Distance to form the cohorts, it will be wise e.g to scale the variables having heights in meters and weights in KGs before calculating the distance.

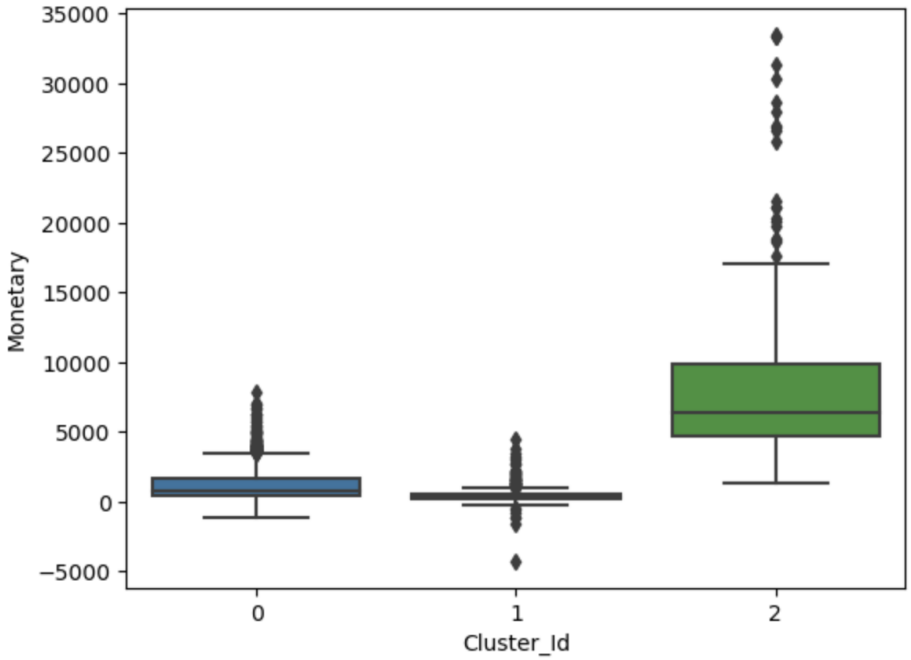
In our analysis we have used standard scalar transformation StandardScaler removes the mean and scales each feature/variable to unit variance. This operation is performed feature-wise in an independent way.

CHOOSING K FOR K MEANS CLUSTERING

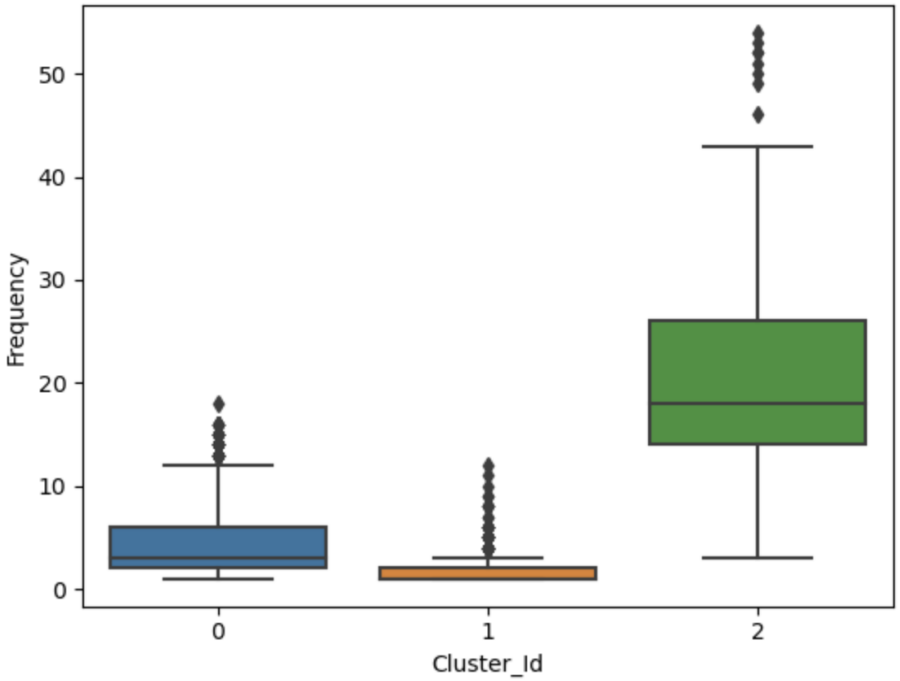


The elbow method is a graphical representation of finding the optimal 'K' in a K-means clustering. It works by finding the Within-Cluster Sum of Square i.e. the sum of the square distance between points in a cluster and the cluster centroid.
Here the elbow is formed on two points **3 and 4** but for ease of understanding, we have considered the number of clusters equal to 3.

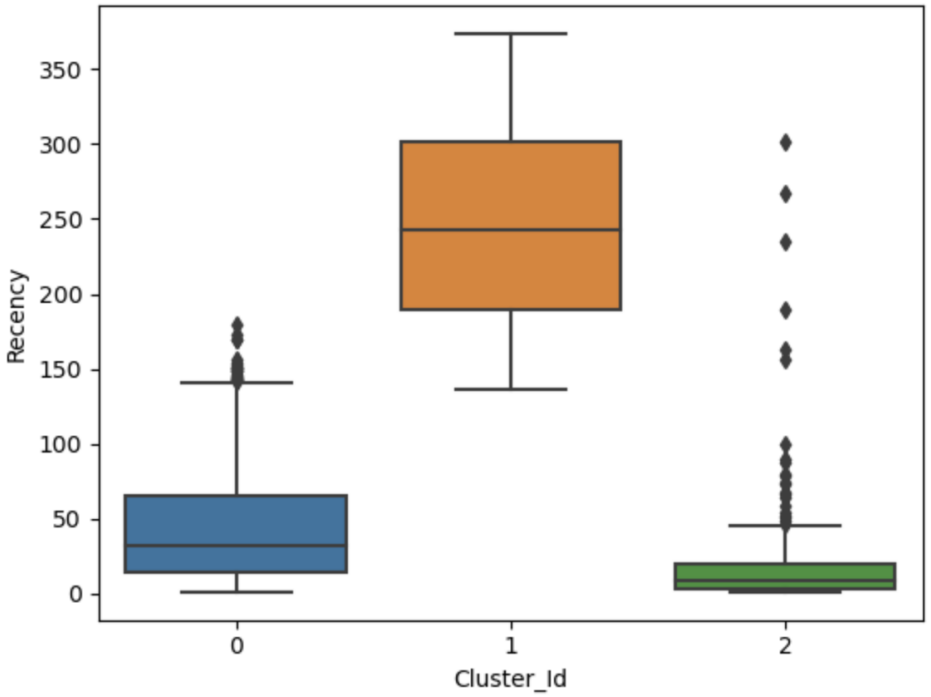
PERFORMING K MEANS CLUSTERING



Clusters based on monetary spending



Clusters based on Frequency



Clusters based on Recency

CONCLUSION

As we can observe from the above boxplots that our model has nicely created 3 segments of customers with the interpretation as below:

- **Customers with Cluster Id 0** are less frequent buyers with low monetary expenditure and also they have not purchased anything in recent times and hence are least important for business.

- **Customers with Cluster Id 1** are the customers having Frequency and Monetary scores in the medium range. But the Recency score is higher so they can be *potentially long-term customers* of the firm.

- **Customers with Cluster Id 2** are the most frequent buyers and spend a high amount on their orders so they are the *most important customers from the business point of view*. But the *recency of this consumer segment has been declining*, ***that means that these high-value customers might have shifted to another retailer***. So special emphasis should be given to these customers by giving better customer service to improve sales as they are the highest source of revenue for the company.

The company can now form relevant strategies to enhance customer satisfaction and customer retention.