FDA Project Report

Northeastern University

Members – Group 8

Sneha Amin

Sai Dheeraj Malkar

Sharanaya Chikkegowda

Rakshit Dharmappa

Ankur Pandey

Contents:

Data Preprocessing

1. We loaded our CSV data into Pandas data frame “df_project”.
2. Then we checked for the columns which might contain null values irrespective of datatype.
    a. df_project.isnull().sum()
3. Filling the Null values according to the provided string for Unknown values on the Crime Data website.
    a. df_project["Description"].fillna("Unknown Description", inplace = True)
    b. df_project["CustomerID"].fillna(0, inplace = True)
4. Converting the Data types
    a. customerID column from float64 to int64 datatype
        i. df_project['CustomerID'] = df_project['CustomerID'].astype('int')
    b. InvoiceDate from string to datetime datatype
        i. df_project['InvoiceDate']= pd.to_datetime(df_project['InvoiceDate'])
5. Checking for duplicated values
    a. total_duplicates = df_project.duplicated().sum()
    we can ignore the duplicate data in this dataset as multiple customer might buy same item.

NOTE: We got more details regarding the data from the provided link for the data set https://www.kaggle.com/datasets/carrie1/ecommerce-data and also, researched further on the internet.

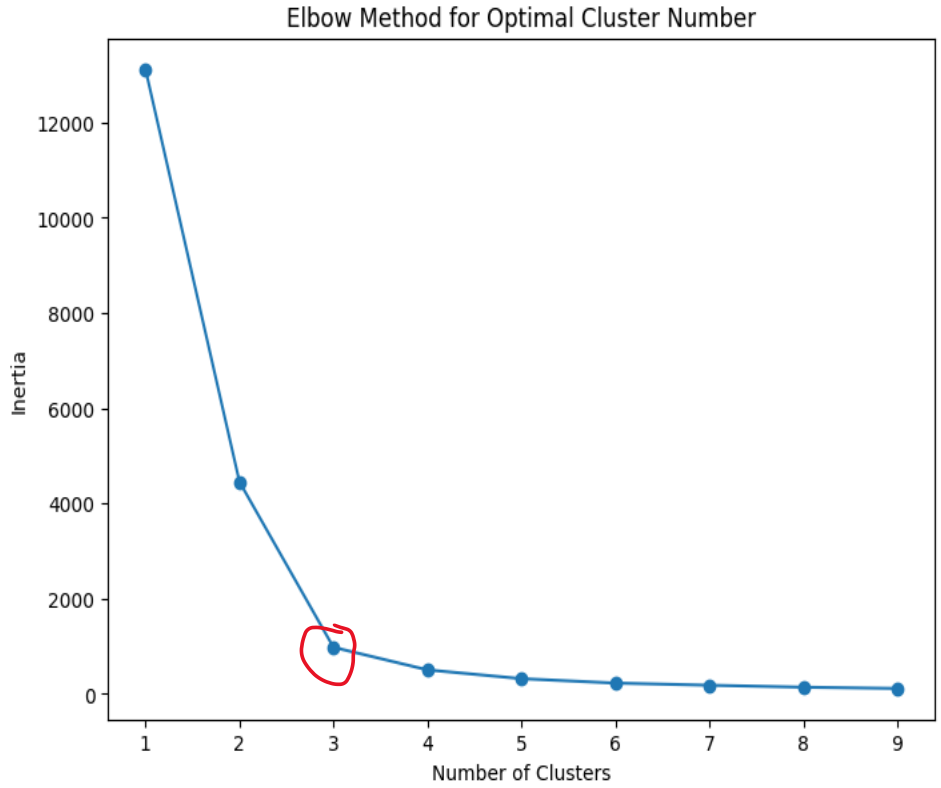
RFM Calculation

o Below are the steps to calculate RFM for the dataset.

1. Calculated the Recency by subtracting InvoiceDate from Today's datetime grouped by customerID.
    a. recency_df = df_project.groupby('CustomerID')['Recency'].min().reset_index()
2. Calculated Frequency by counting the InvoiceDates grouped by customerID.
    a. frequency_df = df_project.groupby('CustomerID')['InvoiceDate'].count().reset_index()
3. Calculated Frequency by taking sum of the UnitPrice grouped by customerID.
    a. monetary_df = df_project.groupby('CustomerID')['UnitPrice'].sum().reset_index()
4. Then we standardized the calculated columns using sklearn StandardScaler library.

Customer Segmentation

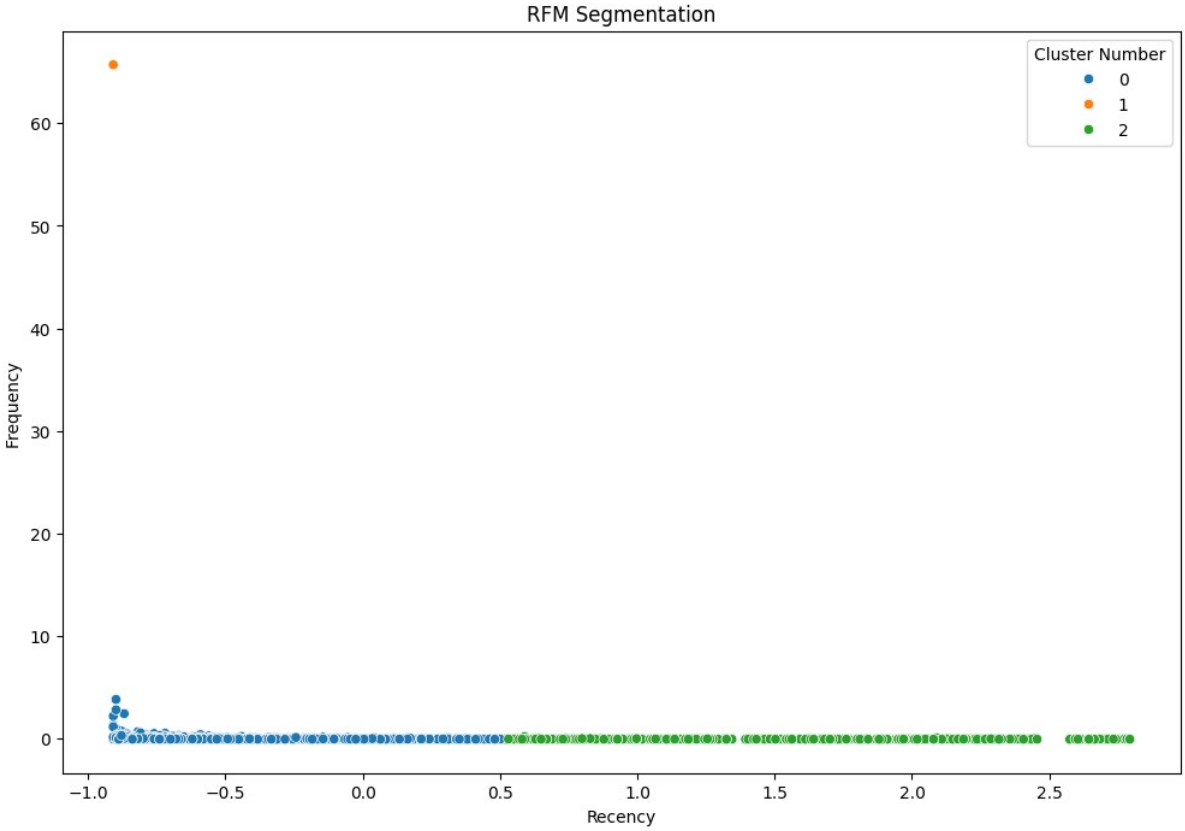
1. We calculated customer segmentation using K-means method and divided the available customer into 3 clusters by Elbow method.



2. #Choose the optimal number of clusters
    a. optimal_clusters = 3
    b. # Apply K-means clustering
    c. kmeans = KMeans(n_clusters=optimal_clusters, random_state=0)
    d. rfm_df['Cluster'] = kmeans.fit_predict(rfm_standardized_df)

RFM Segmentation

- We created RFM segments based on the recency, frequency, and Monetary scores. We noticed that the dataset is dominated by Churned and New customers.
    - #Score 1. High-Value Customers: Low recency, high frequency, and high monetary scores.
    - #Score 2. Potential Loyal Customers: Recent purchases, high frequency, moderate monetary scores.
    - #Score 3. New Customers: Very recent purchases, low frequency, and low monetary scores
    - #Score 4. Churned Customers: High recency, low frequency, and low monetary scores.



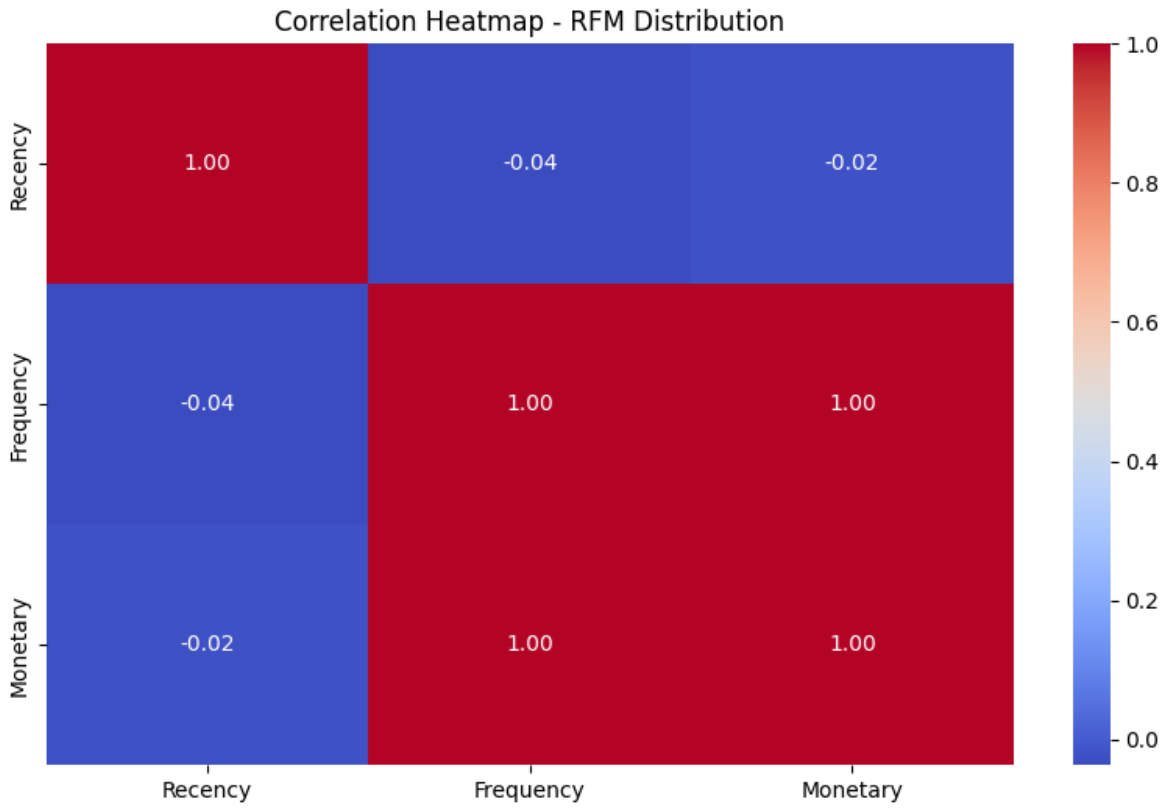
Note: We can safely ignore Cluster 1 as it is caused by the customerIDs replaced by 0 in data pre-processing steps.

Segment Profiling

	Cluster	Recency	Frequency	Monetary	CustomerID
0	0	-0.510003	-0.004860	-0.011549	3293
1	1	-0.908501	65.696532	65.921437	1
2	2	1.557321	-0.046055	-0.025850	1079

- We have two profiles of customers in our dataset.
  - Cluster 0 represents New customers
  - Cluster 3 represents Churned customers

Visualization



The above heatmap visualizes the distribution of RFM values for the dataset. As expected the recency, frequency, and monetary values are strongly related to each other for the dataset.

Marketing Recommendations

1. Decrease Processing Time:

Streamlining and optimizing order processing is imperative. Implementing efficient logistics and order fulfillment systems can significantly reduce processing time, ensuring prompt delivery and enhancing overall customer satisfaction.

2. Reduce Return and Refund Orders:

A meticulous analysis of product descriptions, images, and customer reviews can help mitigate the likelihood of returns. By enhancing the accuracy of product information, you can minimize customer dissatisfaction and subsequent return/refund requests.

3. Arrange Items for Customer Convenience:

Organizing your product catalog in a user-friendly manner enhances the overall shopping experience. Employ intuitive categorization and search features, enabling customers to find items effortlessly. Tailoring recommendations based on past purchases further adds a personalized touch.

4. Improve Customer Service for Enhanced Recency and Frequency Scores:

Elevating customer service directly impacts Recency and Frequency scores. Swift and effective resolution of customer queries, personalized communication, and proactive issue resolution contribute to a positive customer experience. Implementing customer feedback mechanisms aids continuous improvement.

By addressing these aspects, you not only optimize operational efficiency but also cultivate a customer-centric approach, fostering loyalty and boosting the likelihood of repeat business.

Tasks

1. **Data Overview**
    - What is the size of the dataset in terms of the number of rows and columns?
        - df_project.shape
    - Can you provide a brief description of each column in the dataset?
        - df_project.describe
    - What is the time period covered by this dataset?
        - We found that Time Period Covered in the dataset is From 2010-12-01 08:26:00 to 2011-12-09 12:50:00

2. **Customer Analysis**
    - How many unique customers are there in the dataset?
        - Number of Unique Customers: 4372
    - What is the distribution of the number of orders per customer?

    - Can you identify the top 5 customers who have made the most purchases by order count?

- Top 5 Customers with Most Purchases by Order Count:

Customer ID	Total Purchases
14911	248
12748	224
17841	169
14606	128
13089	118

3. Product Analysis

- What are the top 10 most frequently purchased products?
  - Top 10 Most Frequently Purchased Products:

Description	**Quantity**
WHITE HANGING HEART T-LIGHT HOLDER	2070
REGENCY CAKESTAND 3 TIER	1905
JUMBO BAG RED RETROSPOT	1662
ASSORTED COLOUR BIRD ORNAMENT	1418
PARTY BUNTING	1416
LUNCH BAG RED RETROSPOT	1358
SET OF 3 CAKE TINS PANTRY DESIGN	1232
POSTAGE	1196
LUNCH BAG BLACK SKULL.	1126
PACK OF 72 RETROSPOT CAKE CASES	1080

- What is the average price of products in the dataset?
  - Average Price of Products: $ 3.46\

- Can you find out which product category generates the highest revenue?
  - Product Category Generating the Highest Revenue: REGENCY CAKESTAND 3 TIER
  - Total Revenue of the Product Category: 132870.4

4. Time Analysis

- Is there a specific day of the week or time of day when most orders are placed?

- Day of the week with most orders: Thursday
- Hour of the day with most orders: 12:00 - 13:00

- What is the average order processing time?
  - The average processing time for a customer is approx: 24.2 minutes.

- Are there any seasonal trends in the dataset?

  - We can observe that sales of products increase during the November month, which is an expected phenomenon as it is considered as holiday and festival season. Hence, customers shop more than usual during this season.

5. Geographical Analysis

- Can you determine the top 5 countries with the highest number of orders?
    - Top 5 Countries with the Highest Number of Orders:

Country	Number of Orders
United Kingdom	361878
Germany	9495
France	8491
EIRE	7485
Spain	2533

- Is there a correlation between the country of the customer and the average order value?
    - Ans: We did not found any correlation between country of the customer and the average order value in the dataset.

6. Payment Analysis

- What are the most common payment methods used by customers?
- Is there a relationship between the payment method and the order amount?

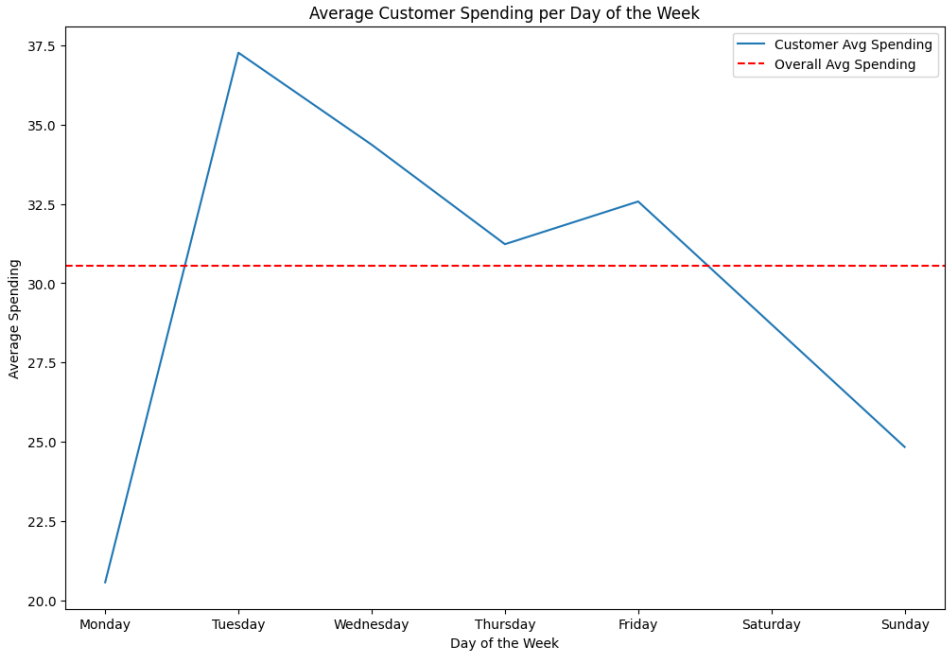
Ans: We do not have sufficient data present in the dataset to derive the payment methods being used by the customers.

7. Customer Behavior

- How long, on average, do customers remain active (between their first and last purchase)?
    - Observation: On average, customers' activity between 1st and last purchase is: 133.39 minutes.

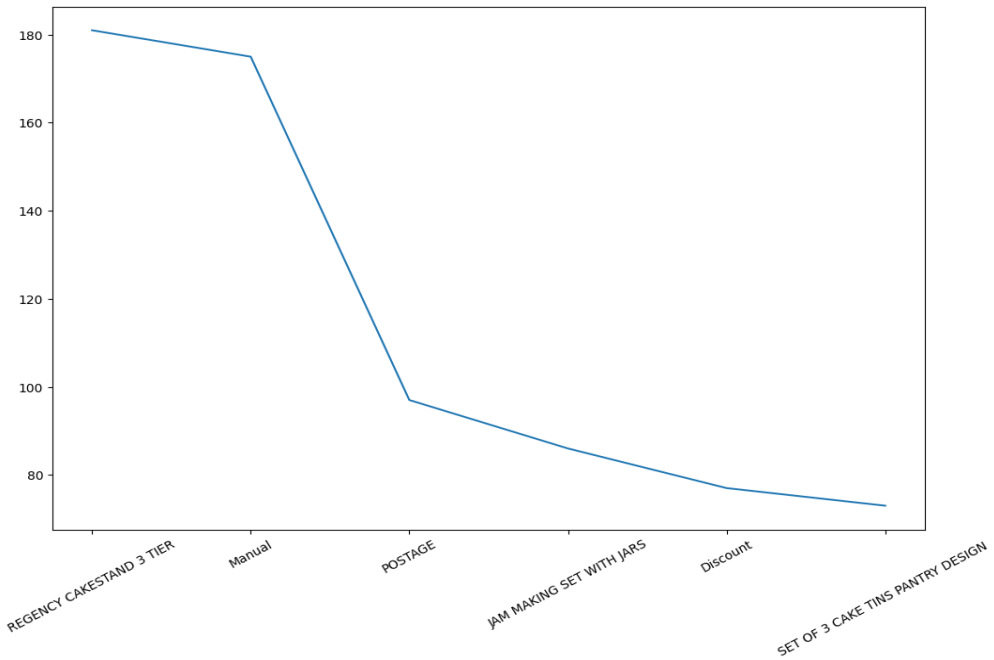
- Are there any customer segments based on their purchase behavior?

- Observation: We observed that customers tend to make more purchase mostly on Tuesdays.



8. Returns and Refunds

- What is the percentage of orders that have experienced returns or refunds?
  - There is total -274808 orders that have experienced returns or refunds. Negative sign represents returned or refunded orders.
- Is there a correlation between the product category and the likelihood of returns?
  - We found that the items that were mostly returned by the customers generally belong to utility categories, this suggested that customer wanted to try out the product and return them according to the convenience of use.



9. Profitability Analysis

- Can you calculate the total profit generated by the company during the dataset's time period?
- What are the top 5 products with the highest profit margins?

Ans: We cannot calculate the profit generated by the company as we do not have the required details to derive this variable i.e., cost price or profit margin on products.

10. Customer Satisfaction

- Is there any data available on customer feedback or ratings for products or services?

Ans: We do not have any recorded variable to provide a measure for the customer satisfaction in our dataset. Hence, we will be using our RFM analysis to get an overall idea about the customer satisfaction level. If customers are satisfied with the store's services, then the recency and frequency would be greater.

1. High-Value Customers: Low recency, high frequency, and high monetary scores.
2. Potential Loyal Customers: Recent purchases, high frequency, moderate monetary scores.
3. New Customers: Very recent purchases, low frequency, and low monetary score.
4. Churned Customers: High recency, low frequency, and low monetary scores.

Now, we can observe that we have we have greater number of Either new customers or Churned customers from our RFM analysis. This means that customer satisfaction of the stores is not great, and they need to work on certain things, as suggested under market recommendations.

- Can you analyze the sentiment or feedback trends, if available?
    - Ans: We do not have any data to understand the sentiment or feedback trend in our dataset.

References:

- E-Commerce Data (kaggle.com)
- What does negative value represent for Column 'Quantity'?
- Quiz 18 FDA class.