Interim Progress Report (IPR)

Project Title: Customer Segmentation for Online Retail
Module Title: Data Science and Analytics Project(7COM1039-0206-2024)
Student Name: Mihir Sanjaybhai Kalariya
Student Number: 23029707
Supervisor: Julia Goncharenko

1. Background Research and Literature Review

1.1 Customer Segmentation in Online Retail

Customer segmentation is a critical process for online retailers to optimize their marketing strategies and increase profitability. By grouping customers based on shared characteristics or behaviors, businesses can deliver more personalized experiences. There are various ways to segment customers, such as by demographics (age, gender), geography (location), psychographics (lifestyle), or behaviors (purchase patterns).

Behavioral segmentation, specifically RFM (Recency, Frequency, Monetary) analysis, is particularly valuable in the context of online retail. This method helps businesses understand customer purchase behavior based on three key metrics: how recently a customer made a purchase, how frequently they buy, and how much money they spend. These metrics provide a detailed picture of customer loyalty, potential for churn, and likelihood to respond to specific marketing strategies.

The increasing availability of transactional data has enabled retailers to implement advanced segmentation strategies, providing greater insight into customer needs. Many large online retailers, such as Amazon and eBay, use segmentation techniques to personalize product recommendations, promotions, and marketing messages to drive sales and improve customer satisfaction.

1.2 RFM Analysis: A Powerful Tool for Segmentation

RFM analysis is a proven method for identifying high-value customers by evaluating three main components:

Recency (R): How recently a customer made a purchase. Customers who have recently made a purchase are more likely to respond to new offers.

Frequency (F): How often a customer buys from the business. Frequent buyers are often more loyal and valuable.

Monetary (M): How much money a customer spends on average. High-spending customers are considered more profitable and are a key focus for retention.

This method allows businesses to identify different customer segments, such as "high-value" customers who make frequent purchases and spend a lot of money, and "at-risk" customers who haven't purchased recently. RFM analysis is advantageous for e-commerce businesses because it is relatively simple to implement and provides actionable insights into customer behavior.

1.3 Other Segmentation and Clustering Options

While RFM analysis is highly effective, there are other segmentation and clustering methods available, each with its own strengths and limitations:

Demographic Segmentation:

Groups customers based on demographic factors like age, gender, income, and education level.

Why not chosen: While useful for understanding who the customers are, it does not provide insights into their purchasing behavior. For example, knowing a customer's age does not tell us how often they shop or how much they spend.

Geographic Segmentation:

Groups customers based on their location, such as country, city, or region.

Why not chosen: This method is useful for location-based marketing but does not provide insights into customer loyalty, spending habits, or purchase frequency.

Psychographic Segmentation:

Groups customers based on lifestyle, interests, and personality traits.

Why not chosen: While insightful, this method is more subjective and harder to quantify. It often requires additional data collection (e.g., surveys), which may not be feasible for small online retailers.

Other Clustering Algorithms:

What they are: Algorithms like DBSCAN, Hierarchical Clustering, and Gaussian Mixture Models (GMM) can also be used for customer segmentation.

Why not chosen: While these algorithms are powerful, they are more complex and computationally intensive compared to K-Means. K-Means was chosen for its simplicity, efficiency, and effectiveness in handling RFM-based segmentation.

1.4 Why RFM and Clustering Were Chosen

RFM analysis was chosen because it focuses on actual customer behavior (recency, frequency, and monetary value), making it highly actionable for online retailers. Unlike demographic or geographic segmentation, RFM provides insights into how customers interact with the business, which is critical for designing targeted marketing strategies.

Clustering, specifically K-Means, was chosen because it complements RFM analysis by grouping customers with similar behaviors into distinct segments. While other clustering algorithms like DBSCAN or Hierarchical Clustering are available, K-Means is simpler to implement and works well with normalized RFM data. It also allows for easy interpretation of customer segments, which is essential for practical marketing applications.

1.5 Research Focus

The focus of this project is to explore how RFM analysis and clustering techniques can be effectively applied to small and medium-sized online retailers. While large e-commerce platforms like Amazon have extensively used these methods, there is limited research on how smaller businesses can implement them with limited data and resources. This project aims to address this gap by applying RFM and clustering to a dataset from a small online retail business and deriving actionable insights to improve customer segmentation and marketing strategies.

Research Question:
“ How can customer segmentation techniques be used to analyze purchasing behavior in an online retail setting? ”

2. Summary of Progress to Date

2.1 Literature Review Completion

The literature review has been completed with a focus on customer segmentation techniques, specifically RFM analysis and clustering. Key findings from the literature indicate that RFM is one of the most widely adopted techniques for customer segmentation in online retail, and clustering plays a crucial role in refining customer groups. The review also covered various clustering methods, including K-Means and hierarchical clustering, which were explored to determine the most suitable approach for segmenting customers in this project.

Through the literature, I have gained insights into the various ways RFM analysis can be implemented in e-commerce, including its effectiveness in identifying both loyal and at-risk customers, and how clustering can provide deeper insights into consumer behavior.

2.2 Data Collection and Cleaning

The dataset for this project was sourced from an online retail business that records transactional details such as customer ID, purchase amount, and purchase date. The data set includes over 10,000 customer records and is considered representative of the target market.

During data cleaning, several steps were taken to ensure that the data was free from errors and inconsistencies:

- **Handling missing values**: Any missing values were either imputed based on other data or removed if they were non-essential.
- **Removing duplicates**: Duplicate records were removed to avoid skewing the results.
- **Correcting inconsistencies**: Data formatting issues (e.g., inconsistent date formats) were addressed to ensure uniformity.
- **Excluding irrelevant information**: Features that were deemed unnecessary for the analysis, such as customer names or locations, were excluded.

After cleaning, the data was ready for further preprocessing to prepare for the RFM analysis.

2.3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to better understand the dataset and uncover any interesting patterns. Key tasks included:

- **Visualizing distributions**: Various visualizations, such as histograms and bar plots, were used to understand the distribution of purchase frequencies and spending amounts.
- **Identifying trends**: Insights were gathered into how often customers make purchases, their average purchase amount, and the recency of their last purchase.
- **Customer segmentation hypothesis**: Based on these visualizations, it was hypothesized that there would be several distinct customer groups based on their purchasing behavior (e.g., frequent, high-spending customers, or occasional, low-spending ones).

The findings from the EDA were essential in preparing for the next stage of the project: calculating RFM scores and applying clustering techniques.

2.4 Data Preprocessing and Feature Selection

The dataset was preprocessed to ensure that it was suitable for RFM analysis and clustering. Key preprocessing steps included:

- **Feature extraction for RFM**: Three main features were derived from the transaction data: Recency, Frequency, and Monetary.
- **Encoding categorical variables**: Any categorical variables were encoded as necessary to ensure they could be used in analysis.
- **Normalization**: The numerical data was normalized to ensure that the clustering algorithm could function correctly, as K-Means clustering is sensitive to the scale of the data.

These steps ensured that the dataset was ready for the next stages of analysis, specifically the calculation of RFM scores and the application of clustering algorithms.

2.5 Ongoing Work

The next steps in the project involve calculating RFM scores for each customer and applying clustering techniques to segment customers based on these scores:

- **RFM Calculation**: The RFM scores are being calculated for each customer based on the frequency of their purchases, the recency of their last purchase, and the total monetary value spent. These scores will be used as input features for clustering algorithms.
- **Clustering**: The clustering process is set to begin once RFM scores are calculated. Algorithms like K-Means will be applied to group customers into segments. These segments will be analyzed to determine their distinct characteristics.
- **Evaluation**: Once the clustering is complete, the effectiveness of the customer segments will be evaluated by analyzing how well they align with customer behaviors and business objectives.

---

3. Consideration of Ethical, Legal, Professional, and Social Issues

- **Privacy**: The dataset used in this project is anonymized, ensuring that no personally identifiable information (PII) is included in the analysis. Customer data privacy is a top priority, and the analysis follows all data protection regulations.
- **Legal considerations**: Data handling is compliant with GDPR and other relevant privacy laws. The dataset used does not contain sensitive information, and any potential risks related to the use of customer data have been minimized.
- **Bias in segmentation**: Bias is an important consideration in any data-driven analysis. Efforts are being made to ensure that the segmentation process is fair and does not unintentionally exclude or discriminate against any customer groups. This includes carefully reviewing the segmentation outcomes to ensure all customer groups are fairly represented.

4. Project Plan

Tasks Completed:

- Literature review on customer segmentation techniques, including RFM analysis and clustering methods.
- Data collection, cleaning, and preprocessing.
- Exploratory Data Analysis (EDA) to identify trends and patterns in customer behavior.

Upcoming Tasks:

- RFM score calculation and implementation.
- Application of clustering algorithms to segment customers.
- Evaluation of clustering results and final report writing.

Timeline:

27/1 to 3/2	**Finalize Project Idea**	Choose the project topic, define the scope, and finalize research objectives.
3/2 to 10/2	**Detailed Project Proposal (DPP)**	Prepare and submit DPP, outlining the methodology, objectives, and expected deliverables.
10/2 to 17/2	**Literature Review**	Conduct a literature review on customer segmentation and RFM analysis.
17/2 to 24/2	**Data Cleaning & Preprocessing**	Handle missing values, duplicates, and prepare the data for analysis.
24/2 to 3/3	**Exploratory Data Analysis (EDA)**	Perform EDA, visualizing customer purchase behaviors and identifying key trends.
3/3 to 17/3	**RFM Analysis**	Calculate RFM scores for customers based on Recency, Frequency, and Monetary value.

17/3 to 31/3	**Clustering Analysis**	Apply clustering algorithms (like K-Means) on RFM scores to segment customers.
1/4 to 21/4	**Final Report & Viva Demo**	Finalize the report with analysis results, visualizations, and conclusions. Prepare for the Viva presentation.

5. Referencing

Khajvand, M., Zolfaghar, K., Ashoori, S., & Alizadeh, S. (2011). Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study. Procedia Computer Science, 3, 57-63.

Why it’s relevant: This paper discusses the application of RFM analysis in estimating customer lifetime value (CLV), which is directly related to this project.

available on: ScienceDirect.

Liu, D. R., & Shih, Y. Y. (2005). Hybrid approaches to product recommendation based on customer lifetime value and purchase preferences. Journal of Systems and Software, 77(2), 181-191.

Why it’s relevant: This paper discusses hybrid approaches to customer segmentation, including RFM analysis, and how it can be used to improve product recommendations.

Fully available for free on: ScienceDirect.

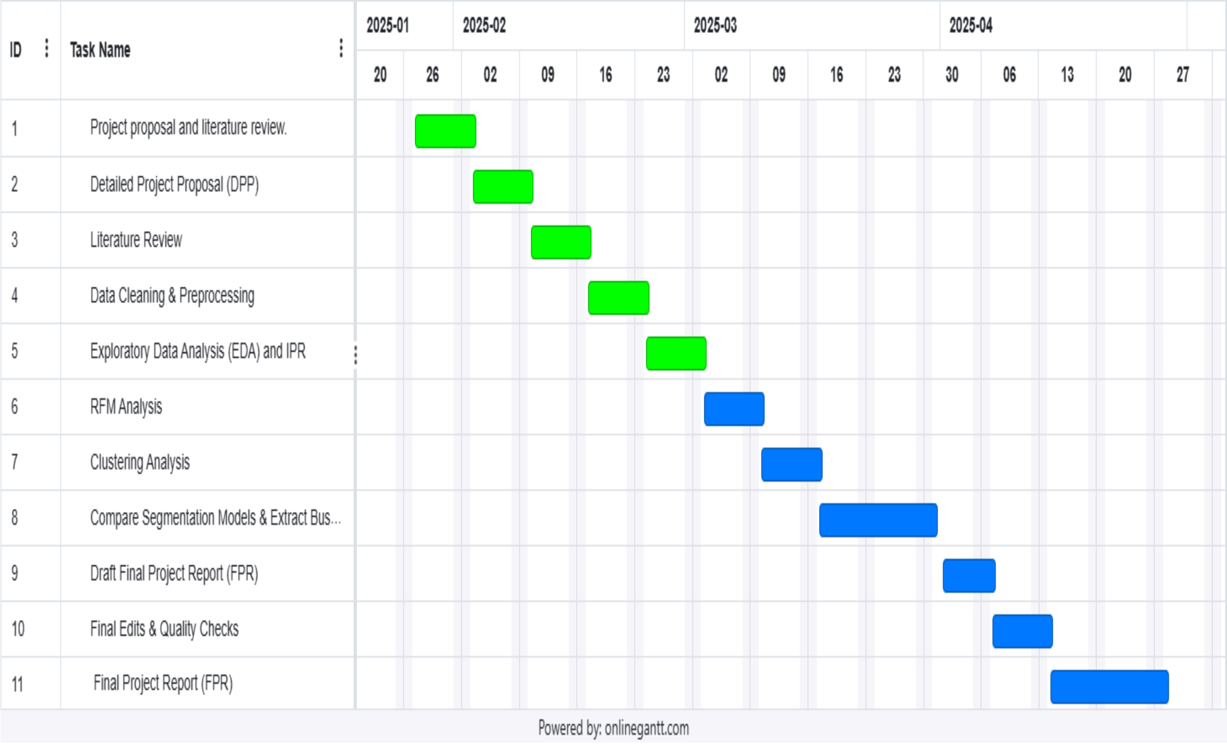
Chen, D., Sain, S. L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using K-Means clustering. International Journal of Electronic Commerce, 16(3), 81-106.

Why it’s relevant: This paper provides a comprehensive case study on applying RFM analysis and K-Means clustering in the online retail industry.

Fully available for free on: ResearchGate.

Appendices

- **Appendix 1:** Gantt Chart for Project Timeline



ID	Task Name
1	Project proposal and literature review.
2	Detailed Project Proposal (DPP)
3	Literature Review
4	Data Cleaning & Preprocessing
5	Exploratory Data Analysis (EDA) and IPR
6	RFM Analysis
7	Clustering Analysis
8	Compare Segmentation Models & Extract Bus...
9	Draft Final Project Report (FPR)
10	Final Edits & Quality Checks
11	Final Project Report (FPR)

- **Appendix 2:** Code Snippets for Data Preprocessing and EDA

In [6]:
```
retail.isnull().sum()
```
Out[6]:
```
InvoiceNo           0
StockCode           0
Description      1454
Quantity            0
InvoiceDate         0
UnitPrice           0
CustomerID     135080
Country             0
dtype: int64
```

In [7]:
```
(retail.select_dtypes(include=['number']) < 0).sum()

```
Out[7]:
```
Quantity      10624
UnitPrice         2
CustomerID        0
dtype: int64
```

In [8]:
```
retail.describe()
```
Out[8]:

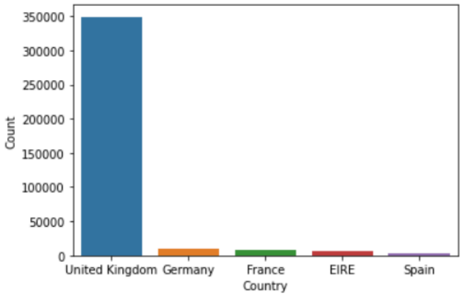
	Quantity	UnitPrice	CustomerID
count	541909.000000	541909.000000	406829.000000
mean	9.552250	4.611114	15287.690570
std	218.081158	96.759853	1713.600303
min	-80995.000000	-11062.060000	12346.000000
25%	1.000000	1.250000	13953.000000
50%	3.000000	2.080000	15152.000000
75%	10.000000	4.130000	16791.000000
max	80995.000000	38970.000000	18287.000000

Which countries has the most number of customers?

In [35]:
```
top_countries = retail['Country'].value_counts().reset_index().rename(columns={'index': 'Country', 'Country': 'Count'})
top_countries.head(5)
```
Out[35]:

	Country	Count
0	United Kingdom	349203
1	Germany	9025
2	France	8326
3	EIRE	7226
4	Spain	2479

In [36]:
```
sns.barplot(x = 'Country', y = 'Count', data = top_countries.head(5))
```
Out[36]:
```
<AxesSubplot:xlabel='Country', ylabel='Count'>
```


Count
United Kingdom
Germany
France
EIRE
Spain
Country