**CUSTOMER SEGMENTATION RECENCY FREQUENCY MONETARY ANALYSIS**

**CHIU LINH PHAT**

**OPEN UNIVERSITY MALAYSIA**

**2023**

**CUSTOMER SEGMEMENTATION RECENCY FREQUENCY MONETARY ANALYSIS**

**CHIU LINH PHAT**

A Final Year Project submitted in fulfilment of the requirements

for the degree of

Bachelor of Information Technology

OPEN UNIVERSITY MALAYSIA

2023

**DECLARATION**

Name:

Matric Number:

I hereby declare that this final year project is the result of my own work, except for quotations and summaries which have been duly acknowledged.

Signature: Date:

**CUSTOMER SEGMEMENTATION RECENCY FREQUENCY MONETARY ANALYSIS**

**ABSTRACT**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Keywords:

(Not more than 5 words/phrases)

**ACKNOWLEDGEMENT**

I would like to take this opportunity to express my gratitude and appreciation to my supervisor, Mrs. Le Thi Ngoc Tho guidance, patience and invaluable advice throughout this project. I also would like to express my appreciation to my family and friends for them endless support whenever I face problems. Without the mentioned parties, it is impossible for me to complete this project report successfully.

THANK YOU.

Chiu Linh Phat

26 December ,2023

TABLE OF CONTENTS

TITLE PAGE

DECLARATION

ABSTRACT

ABSTRAK

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CHAPTER 1 INTRODUCTION

1.1 Background to the Study

1.2 Problem Statement

1.3 Objectives of the Study

1.4 Research Questions and Hypotheses

1.5 Significance of the Study

1.6 Scope of the Study

1.7 Definitions of Terms

CHAPTER 2 REVIEW OF LITERATURE

2.1 Previous Studies

2.1.1 Project success criteria

2.2 Methodological Issues

2.3 Conceptual Framework

CHAPTER 5 DISCUSSION AND CONCLUSION

5.1 Summary of Main Findings

5.2 Discussion

5.3 Implications

5.4 Limitations of the Study

5.5 Directions for Future Research

REFERENCES

APPENDICES

A Title 100

B Title 105

Outline

CHAPTER 1: INTRODUCTION

Background to the study:

* Nguyên nhân dẫn đến việc cần thiết customer segemation
* Nêu ra một số thông tin về tại sao lại cần customer segemation

Problem Statement:

* Nêu ra về vấn đề gặp phải trong customer segmentation

Objectives of the Study:

* Nêu ra về lợi ích sau khi thực hiện quá trình customer segmentation.

Research Questions / Hypotheses:

* Đưa ra một số câu hỏi xoáy vào nhất

Scope and Limitation:

* Đưa ra các giới hạn của data set

Implementation Plan:

* outlines the step-by-step plan for executing the study.

Definition of Terms:

* This section provides clear definitions for terms specific to the study on customer segmentation.

CHAPTER 2: LITERATURE REVIEW

# CHAPTER 1: INTRODUCTION

## 1.1 Background to the Study:

Information technology and the internet have developed strongly in recent years, creating favorable conditions for e-commerce to develop. Consumers can easily access e-commerce websites to search and shop for products. Businesses can also easily build and operate e-commerce websites to reach customers. In that context, compared to traditional trading methods that are costly but less effective, Typical sales pages are growing rapidly and some stores are moving towards this market with significant changes.

Consumers increasingly tend to shop online. They can shop anytime, anywhere, as long as they have an internet connection. They can also easily compare prices, product quality and services of different stores.

Customer segmentation is the process of dividing customers into subgroups based on their common characteristics. This allows businesses to better understand their customers and provide them with more relevant products, services and experiences.

This section establishes the context for the study, emphasizing the importance of understanding and categorizing customers based on their behavior, purchasing patterns, and value to the business. It highlights the evolution from manual segmentation by marketing and sales teams to the contemporary use of data mining and machine learning for efficient and targeted customer segmentation.

## 1.2 Problem Statement:

Here, the problem statement identifies challenges inherent in traditional manual methods of customer segmentation. It suggests that these methods may struggle with large datasets and may not efficiently identify nuanced patterns among diverse customer groups. The study aims to address these limitations by proposing the use of unsupervised learning algorithms as a solution for automating and improving the efficiency of customer segmentation.

## 1.3 Objectives of the Study:

This section outlines the specific goals of the research related to customer segmentation. Objectives include exploring the effectiveness of unsupervised learning algorithms in automating segmentation, assessing their application in targeted marketing and sales strategies, and identifying common patterns and characteristics among different customer segments.

## 1.4 Research Questions / Hypotheses:

Here, the research questions or hypotheses are tailored to the context of customer segmentation. Questions may revolve around the effectiveness of unsupervised learning algorithms for segmentation, their impact on targeted marketing efforts, and the types of patterns and traits that can be identified across different customer segments.

## 1.5 Scope and Limitation:

Scope:

The scope of the study is focused on the application of unsupervised learning algorithms for customer segmentation. Customer segmentation involves grouping customers based on similar characteristics, behaviors, or preferences without predefined categories. The study aims to uncover patterns and insights within the data using unsupervised learning techniques.

Focus Area:

The primary focus is on leveraging unsupervised learning algorithms to analyze a Kaggle dataset provided by Dr. Omar Romero-Hernandez. This dataset consists of 2,240 rows of observations and 28 columns of variables. The variables include both categorical (5-character variables) and numerical (23 variables) features. The analysis aims to identify meaningful clusters within the customer data.

Limitation:

* Data Quality:

The study acknowledges that the findings may be influenced by the quality and representativeness of the Kaggle dataset. It is essential to be aware of potential biases, missing data, or inaccuracies that could impact the reliability of the results.

* External Factors:

Recognizing external factors, such as evolving market dynamics, the study understands that these factors may affect the generalizability of the results. The external environment, including economic conditions or industry trends, may impact the validity and applicability of the findings beyond the specific context of the dataset.

## 1.6 Implementation Plan:

Data Collection:

Ensure a comprehensive understanding of the dataset, including its sources, structure, and any potential biases.

Data Cleaning:

Execute thorough data cleaning procedures to address issues such as missing values, outliers, and inconsistencies in the dataset.

Standardize and preprocess the data to create a clean and reliable foundation for subsequent analysis.

Exploratory Data Analysis (EDA):

Conduct Exploratory Data Analysis to gain insights into the distribution of variables, uncover patterns, and identify key features relevant to customer segmentation.

Visualize the relationships between variables to inform the selection of appropriate segmentation criteria.

Recency, Frequency, Monetary (RFM) Analysis:

Implement Recency, Frequency, Monetary (RFM) analysis to derive meaningful customer segments based on their purchasing behavior.

Calculate and assign RFM scores to each customer, enabling the identification of high-value segments.

Visualization and Practical Application:

Utilize visualization techniques to represent the identified customer segments effectively.

Discuss the practical application of the findings in targeted marketing and sales efforts, providing actionable insights for personalized customer engagement.

Consider the integration of RFM-based segmentation into marketing strategies, tailoring promotions, and communication strategies to meet the specific needs of each customer segment.

## 1.7 Definition of Terms:

This section provides clear definitions for terms specific to the study on customer segmentation. Key terms may include "Customer Segmentation" and others relevant to the context of the research.

# CHAPTER 2: LITERATURE REVIEW

….

# CHAPTER 3: RESEARCH METHODOLOGY

## 3.1 Research Design

The research design for implementing RFM analysis involves outlining the overall plan and strategy for conducting the analysis. It encompasses decisions about the type of study, the scope of the research, and the methods employed. In the context of RFM analysis:

The process of data collection and research for RFM analysis is a critical phase that shapes the foundation of the study. To effectively implement RFM analysis, a systematic approach to gathering and understanding the data is essential. This section outlines the key steps involved in the research process, emphasizing the strategies employed for data collection and analysis.

Firstly, the selection of data sources is a pivotal decision in the research design. In the context of RFM analysis, the data often originates from customer transactions, encompassing information on recency, frequency, and monetary value. The identification of relevant databases or datasets is crucial to ensure the availability of comprehensive and accurate information. This may involve collaboration with relevant departments, such as marketing or sales, to access transactional records and customer data.

The next step involves defining the parameters for recency, frequency, and monetary value, which are fundamental to RFM analysis. Careful consideration is given to the time frame for recency, the definition of a 'frequent' transaction, and the criteria for categorizing customers based on their monetary contributions. Clear and well-defined parameters contribute to the accuracy and reliability of the analysis.

Once the data sources and parameters are established, the data collection process commences. This may involve the extraction of transactional data, customer profiles, and other relevant information from the identified sources. The data is then organized and formatted for further analysis. Quality assurance measures, such as data cleaning and validation, are implemented to address any inconsistencies or inaccuracies in the dataset.

The subsequent phase entails the application of RFM analysis techniques to the prepared dataset. This involves the calculation of recency, frequency, and monetary values for each customer, leading to the segmentation of the customer base. Various statistical and analytical tools may be employed to derive meaningful insights from the data, such as clustering algorithms or regression analysis.

Throughout the entire process, it is essential to maintain a rigorous documentation of the methodology and decisions made during data collection and analysis. This documentation not only ensures transparency and reproducibility but also facilitates the evaluation of the study's validity and reliability.

### 3.1.1 Type of Study

Determine whether the research is exploratory, descriptive, or explanatory in nature. Consider if the focus is on understanding customer behavior, describing patterns, or explaining the relationships between RFM metrics and business outcomes.

The nature of the research design can be characterized based on its overarching objectives and the level of depth in understanding customer behavior, describing patterns, or explaining the relationships between RFM metrics and business outcomes. In this context, the research design for RFM analysis exhibits elements of both descriptive and explanatory research, with a primary focus on unraveling the intricacies of customer behavior and its impact on business outcomes.

Primarily, the research leans towards being descriptive as it seeks to depict and characterize the patterns inherent in customer transactions through the lens of recency, frequency, and monetary value. Descriptive research aims to provide a comprehensive overview of a phenomenon, and in this case, the objective is to delineate customer segments based on their RFM metrics. The analysis involves classifying customers into distinct groups, shedding light on their purchasing habits, transaction frequency, and overall monetary contributions. By doing so, the research endeavors to offer a clear and detailed snapshot of customer behavior within the given dataset.

Concurrently, the research design incorporates elements of explanatory research, as it delves deeper into understanding the relationships between RFM metrics and business outcomes. While descriptive research identifies patterns, explanatory research seeks to elucidate the underlying reasons and causal relationships. In the context of RFM analysis, the investigation may extend beyond the mere categorization of customers to explore the impact of these segments on business performance. For example, the study may aim to explain why certain customer segments exhibit higher monetary value and whether this correlates with increased profitability or customer loyalty.

Moreover, the research design may involve probing into the effectiveness of marketing strategies or personalized approaches tailored to different RFM segments. This explanatory dimension is crucial for organizations seeking actionable insights to enhance their business strategies based on a nuanced understanding of customer behavior.

### 3.1.2 Scope of Research

The specific objectives of the RFM (Recency, Frequency, Monetary) analysis are multifaceted, encompassing a comprehensive exploration of customer behavior to inform targeted strategies, enhance marketing initiatives, and improve overall customer engagement. Firstly, a primary objective of RFM analysis is to gain a nuanced understanding of customer segments. By categorizing customers based on recency, frequency, and monetary value, the analysis seeks to identify distinct groups within the customer base. This segmentation enables businesses to move beyond a one-size-fits-all approach and tailor their strategies to meet the unique needs and behaviors of each segment. Understanding these segments can illuminate patterns and trends in customer behavior, facilitating more personalized and effective engagement strategies. Identifying opportunities for targeted marketing is another key objective of RFM analysis. Once customer segments are defined, businesses can leverage this knowledge to design and implement precise marketing campaigns. For example, customers who exhibit high recency and frequency but lower monetary value may benefit from targeted promotions to increase their average transaction value. On the other hand, customers with high monetary value but low frequency might be targeted with loyalty programs or incentives to increase their engagement. Moreover, the analysis aims to uncover insights into customer preferences and behaviors that contribute to improved overall customer engagement. By understanding the factors that drive customer satisfaction and loyalty, businesses can refine their products, services, and customer interactions. This, in turn, enhances the overall customer experience, fostering stronger relationships and increasing the likelihood of repeat business.

RFM analysis can also play a pivotal role in customer retention efforts. By identifying at-risk customer segments or those showing signs of decreased engagement, businesses can proactively implement retention strategies. For instance, targeted communication, exclusive offers, or personalized recommendations can be employed to re-engage customers and prevent attrition.

Furthermore, the analysis contributes to strategic decision-making by providing actionable insights into the performance of marketing initiatives, the effectiveness of promotions, and the overall health of customer relationships. This information empowers businesses to allocate resources effectively, optimize marketing budgets, and refine their approach based on measurable outcomes.

### 3.1.3 Sampling Strategy

Decide on the sampling method to select a representative subset of data for analysis. Consider factors such as data availability, time constraints, and the desired level of precision in the results. Data availability is a key determinant in selecting the most appropriate sampling method. If the entire dataset is readily accessible and manageable, a census approach may be feasible. This involves analyzing the entire population of interest, providing a complete and accurate representation of customer behavior. However, in scenarios where the dataset is vast and time-consuming to process, a more pragmatic approach is to employ a sampling method. Considering time constraints is essential in determining the sampling method. If there is a need for swift results due to time limitations, a random sampling technique may be suitable. Random sampling involves selecting a subset of data points purely by chance, ensuring that each element in the population has an equal probability of being included. While this method can be quick, it may not account for specific characteristics or patterns within the data, potentially leading to less precision. Alternatively, if time allows for a more comprehensive approach, stratified sampling might be advantageous. This method involves dividing the population into distinct strata or subgroups based on relevant criteria (e.g., customer segments, geographic regions) and then randomly sampling from each stratum. This ensures representation from various segments, providing a more nuanced understanding of the population. The desired level of precision in the results is a critical factor influencing the choice of a sampling method. If a high level of precision is required, a stratified or cluster sampling method may be more suitable.

These methods consider the heterogeneity within the population, allowing for more accurate insights into specific subgroups. On the other hand, if the research aims to provide a broad overview and general trends, a simple random sampling approach might be sufficient.

### 3.1.4 Time Horizon

Defining the time period for RFM analysis is a crucial decision that significantly influences the insights derived from the study. The time horizon determines whether the focus is on recent data, historical trends, or a balanced combination of both. Each approach carries its own implications for understanding customer behavior through the Recency, Frequency, and Monetary metrics.

If the emphasis is on recent data, the analysis zeroes in on a specific, short-term timeframe. This approach is particularly valuable for businesses seeking to understand the most current customer behaviors and preferences. Recent data provides insights into the immediate impact of marketing campaigns, changes in consumer behavior, and the effectiveness of current strategies. Analyzing short-term trends allows businesses to adapt quickly to dynamic market conditions and make timely adjustments to their customer engagement strategies. Conversely, opting for a focus on historical trends involves analyzing a more extended timeframe. This approach is beneficial for uncovering patterns and behaviors that may have evolved over time. By considering historical data, businesses can identify long-term trends, seasonal variations, and patterns that might not be apparent in a snapshot of recent data. This perspective is particularly useful for strategic planning and understanding the evolution of customer segments over different seasons or economic cycles. A balanced approach that combines both recent and historical data offers a comprehensive view of customer behavior. This hybrid model provides insights into immediate trends while also considering the broader context of long-term customer engagement. It allows businesses to capture the nuances of changing customer preferences while recognizing the enduring patterns that contribute to overall customer loyalty and satisfaction. The time horizon selected significantly impacts the recency aspect of RFM metrics. A shorter time period emphasizes the most recent transactions, offering a snapshot of current customer behavior. In contrast, a longer time period encompasses a broader range of transactions, providing a more historical perspective. Striking the right balance depends on the specific objectives of the analysis and the nature of the business. Additionally, when deciding on the time period, it is crucial to align with the industry's characteristics, product lifecycle, and the frequency of customer interactions. Industries with rapid changes or seasonal fluctuations may benefit from shorter timeframes, while those with more stable patterns may find value in a more extended historical analysis.

The determination of the time period for RFM analysis involves a thoughtful consideration of the balance between recent and historical data. Whether focusing on immediate trends, long-term patterns, or a combination of both, businesses can tailor their approach to gain actionable insights into customer behavior and inform strategic decision-making in the dynamic landscape of today's markets.

## 3.2 Data Collection Method

The success of RFM (Recency, Frequency, Monetary) analysis hinges on the meticulous and thorough collection of high-quality and comprehensive data. The data collection process is a critical phase that involves obtaining transactional data, customer information, and other relevant data points to ensure the accuracy and relevance of the insights derived from the analysis. One of the primary considerations in data collection for RFM analysis is the source of the data. Transactional data, which includes details of customer purchases, dates of transactions, and corresponding monetary values, forms the backbone of RFM metrics. This information is typically sourced from sales databases, e-commerce platforms, or point-of-sale systems. It is imperative to collaborate closely with relevant departments, such as sales, marketing, or IT, to access and extract this transactional data accurately. In addition to transactional data, collecting comprehensive customer information is vital. This includes demographic details, contact information, and any relevant attributes that can enhance the segmentation of customers. Customer profiles contribute to a more nuanced understanding of the different segments identified through RFM analysis, enabling businesses to tailor their strategies to specific customer characteristics. The completeness and accuracy of the data are paramount. To ensure data integrity, it is essential to implement robust data cleaning and validation processes. This involves identifying and rectifying any discrepancies, inaccuracies, or missing values within the dataset. Inaccurate or incomplete data can skew the results of RFM analysis, leading to unreliable insights and potentially misguided strategic decisions. Furthermore, the data collection method should align with the ethical considerations and legal requirements governing the use of customer data. Adhering to data privacy regulations and obtaining necessary permissions for data usage is crucial to maintain trust and compliance with applicable laws. Clear communication with customers regarding data collection practices can also foster transparency and build positive relationships. Timing is another critical aspect of data collection. The frequency at which data is collected impacts the recency aspect of RFM metrics. Real-time data collection provides the most up-to-date insights, especially beneficial for businesses with rapidly changing customer behaviors. However, the frequency should be balanced with practical considerations, such as the availability of resources and the need for timely analysis. Moreover, businesses may consider incorporating external data sources, such as market trends or economic indicators, to enrich their understanding of the context in which customer behaviors unfold. Integrating diverse data sets can contribute to a more comprehensive analysis, offering a holistic view of the factors influencing customer engagement.

### 3.2.1 Data Sources

Identify the sources of data, which may include transaction records, customer databases, and any additional relevant datasets. Ensure that the data sources are reliable, up-to-date, and representative of the target customer population. Identifying and leveraging reliable data sources is paramount to the success of RFM (Recency, Frequency, Monetary) analysis. The accuracy and representativeness of the data play a pivotal role in generating meaningful insights into customer behavior. The primary sources of data for RFM analysis include transaction records, customer databases, and additional relevant datasets. Transaction records serve as the foundational data source for RFM analysis. These records contain detailed information about customer purchases, including the date of each transaction and the corresponding monetary value. Organizations typically extract this data from sales databases, point-of-sale systems, or e-commerce platforms. Ensuring the accuracy and completeness of transactional data is essential, as it directly influences the precision of recency, frequency, and monetary metrics. Real-time access to transaction records provides the most current insights into customer behavior, especially for businesses operating in dynamic markets. Customer databases are another critical source of data for RFM analysis. These databases house comprehensive customer information, including demographic details, contact information, and historical interactions. Accessing and integrating this data enables a more nuanced segmentation of customers based on factors beyond transactional behavior. The reliability of customer databases is crucial, and organizations should implement measures to update and maintain accuracy, particularly as customer information evolves over time. In addition to transaction records and customer databases, businesses may incorporate additional relevant datasets to enrich their understanding of customer behavior. These datasets could include external factors such as market trends, economic indicators, or even social media interactions. Integrating diverse datasets enhances the context of RFM analysis, allowing organizations to identify external influences that may impact customer engagement. It is imperative to evaluate the reliability and relevance of these additional datasets to ensure their contribution to the overall analysis. Ensuring the representativeness of the target customer population is equally crucial. The selected data sources should encompass a diverse and inclusive sample that accurately reflects the demographic and behavioral diversity of the customer base. This prevents biases in the analysis and ensures that insights derived from RFM metrics are applicable across different customer segments.

Regular audits and data quality assessments are essential practices to maintain the reliability of data sources. Organizations should establish protocols for data validation, cleaning, and updating to address any inaccuracies or inconsistencies. This proactive approach ensures that the data used for RFM analysis remains trustworthy and continues to provide valuable insights over time.

### 3.2.2 Data Collection Techniques

The techniques employed to collect the necessary data for RFM (Recency, Frequency, Monetary) analysis are critical components of the overall research process. Depending on the data sources and the nature of the business operations, various methods can be employed to extract transactional and customer-related information. The choice of techniques often revolves around data extraction from databases, API (Application Programming Interface) integration, and other methods suitable for gathering comprehensive and accurate datasets. One primary method for data collection in RFM analysis involves extracting transactional data from internal databases. This entails accessing and querying the organization's sales databases, point-of-sale systems, or e-commerce platforms. Through SQL (Structured Query Language) queries or database management tools, businesses can retrieve specific fields such as transaction dates, monetary values, and customer IDs. This method is particularly effective for businesses with well-structured internal databases, providing a direct route to the detailed transactional information needed for RFM analysis. API integration is another powerful technique for collecting data, especially in scenarios where information is sourced from external platforms or services. Many e-commerce platforms, payment gateways, and customer relationship management (CRM) systems offer APIs that allow seamless integration with other applications. By leveraging APIs, businesses can programmatically access transactional and customer-related data, ensuring real-time synchronization and reducing manual data entry errors. API integration is particularly valuable for organizations operating across multiple platforms, as it facilitates a centralized and automated approach to data collection.

## 3.3 Data Analysis Method

The data analysis method involves the application of RFM scoring and subsequent segmentation. It includes techniques for interpreting results and deriving actionable insights.

### 3.3.1 RFM Scoring Algorithm

Define the algorithm used to calculate RFM scores based on recency, frequency, and monetary value. This may involve assigning numerical values, establishing score ranges, or applying weighted scores based on business priorities.

### 3.3.2 Segmentation Criteria

Specify the criteria used to segment customers based on their RFM scores. Determine the number and characteristics of segments (e.g., Champions, Loyal Customers, Loyal Customers, Recent Customers, Promising, Need Attention, About to Sleep, At Risk), and outline how these segments align with business objectives.

### 3.3.3 Statistical Methods

If applicable, describe any statistical methods employed for validating the significance of the identified segments or assessing the reliability of the RFM analysis.

Detail how the results of the RFM analysis will be interpreted and translated into actionable insights. This may involve recommendations for targeted marketing strategies, personalized customer experiences, or operational improvements.