Customer Segmentation contributes to Marketing plan.

Author: Chi Nguyen

1. Introduction

2.1 Outline

The data is a sales data of a retail company which sells their product across many countries. The company sells directly to individual customers and collected information includes CustomerID, InvoiceNo, Quantity of purchase, Date of Invoice, Unit Price.

Our main goal throughout this analysis is composed of objectives:

1. What's the performance of the sales data
2. How customers can be segment and what're their characteristics
3. Suggestion for marketing plan

2.2 The data

Sales data contains information on each purchased item. Each row represents one item in an invoice.

- InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
- StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
- Description: Product (item) name. Nominal.
- Quantity: The quantities of each product (item) per transaction. Numeric.
- InvoiceDate: Invice Date and time. Numeric, the day and time when each transaction was generated.
- UnitPrice: Unit price. Numeric, Product price per unit in sterling.
- CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
- Country: Country name. Nominal, the name of the country where each customer resides.

An invoice has many item line.

2.3 Methods

2.2.1 At the first stage, a preliminary data transformation was conducted including:

- Dropping missing values in 'Description' and 'CustomerID' columns since we will proceed the customer segmentation so let us drop it instead of computing it. Besides, the percentage of missing is not major hence the dropping action will not cause information loss problem.
- Changing negative values in 'Quantity' column to zero 0 since they mean canceled items. A column 'Canceled' column was created to contain this information instead.
- Dropping 0 values in 'UnitPrice' column and computing 'TotalValue' to get the value of the order by multiplying 'UnitPrice' with ' Quantity'.
- Extracting day, time, year, month, weakday of the time data.

2.2.2 Next, customer segmentation analysis was conducted to get further insights:

- Recency, Frequency, and Monetary Value data of customer were computed from the data, and a **RFM analysis** was conducted to get an idea about customer segmentation:

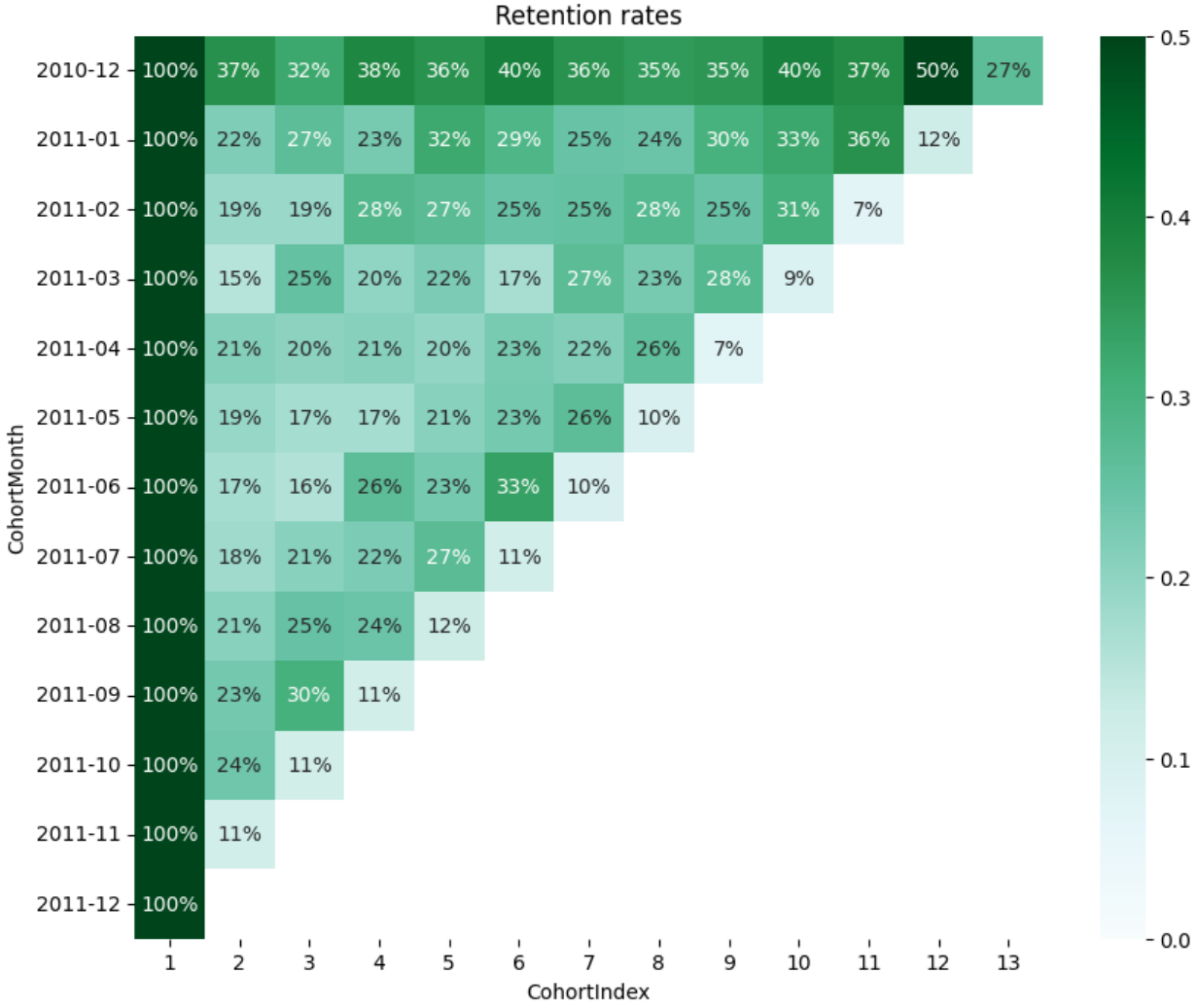
- ‘Gold’ customer: who made the latest purchase most recent, about 27 days from the latest day of data, high frequency of purchasing with 192 orders and spent a lot of money around 4406$ in average.
- ‘Silver’ customer: whose last purchase was around 100 days ago in average, made around 35 orders in average, and spent around 725$ across the year in average.
- ‘Bronze’ customer: whose last purchase was 218 days agon in average, made only 11 purchases in average, and spent only around 200$ in average.

- Conduct a data pipeline and **K-mean clustering** method to check the previous segmentation work and involved more features to get insights about persona of different segments.
- The data pipeline includes:
  - Adding ‘Quantity of purchased item’ and ‘Cancelation history’ features for next step.
  - Preprocessing the data: remove outliers, normalize, and standardize the data.
  - K-means clustering was conducted and 2 segmentation was the best result based on Elbow method.

2. Objectives

2.1 The performance of business

- The sales were mostly witnessed in 2011 and at United Kingdom with nearly 5M items and more than 350k invoices were purchased successfully that resulted in more than 7M dollars in that year.
- The sales mostly happened in weekdays rather than weekends, and the sales season is the year-end time. That's an insightful information that marketing team will have a proper plan to invest their budget on which time of the year and which day in a week. Hence, the spending could be utilized.



- In this time-based cohort analysis, customers were grouped into cohorts based on the month of their first purchase. Customers were assigned to each month in the first column were active on that month, in other words, they purchased their first order in that month.
- The Cohort Index means the number of month that the cohort customer repurchased after the first purchase. For example, the number 26% of row '2011-06' and column '4' mean there were 26% people reactive after 6 months from 2011-06.
- According to that, we can see that customers purchased on 2010-12, 2011-01, 2011-02, 2011-03, 2011-04, 2011-08, 2011-09, 2011-10 tent to repurchase more than the other

months. And this information could raise some ideas for brand team to review and utilize marketing activities activated on these months and what trigger these groups of customers to re-purchase.

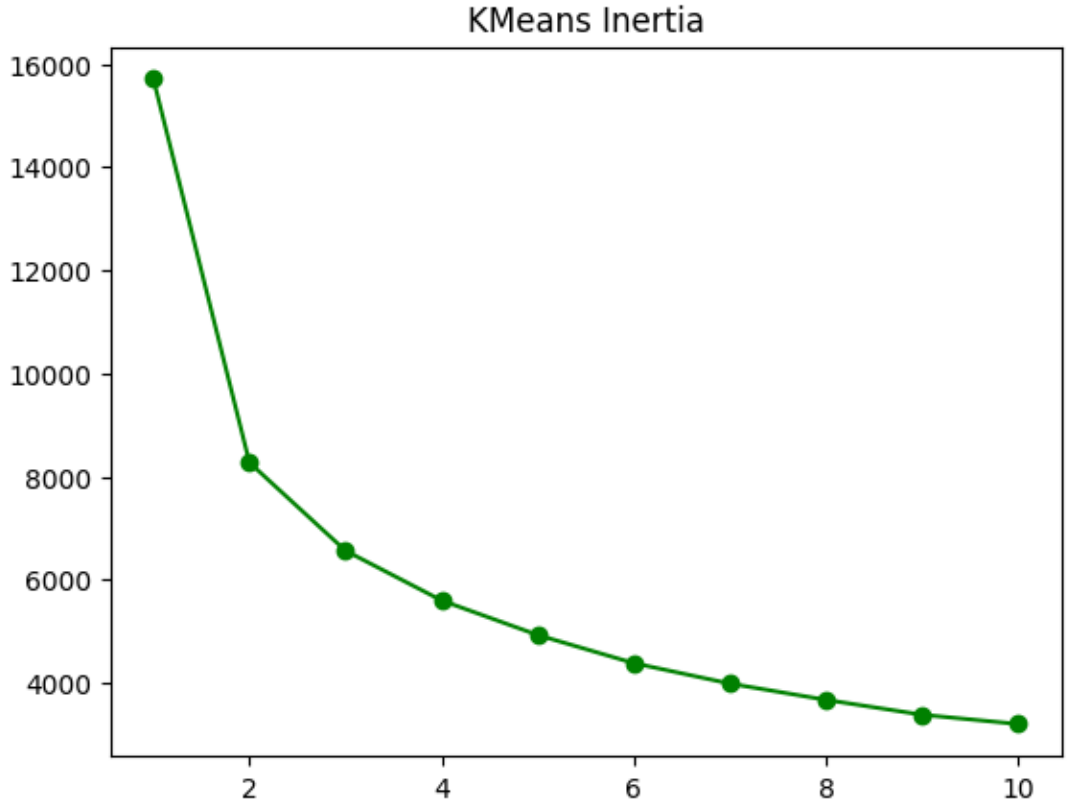
2.2 Customer segmentation and their characteristics

With RFM analysis, customers were recognized as 3 different groups based on Recency, Frequency, Monetary. However, this approach could cause some bias since the scale range is different between different attributes. Therefore, we proceed K-means clustering method with adding 2 more attributes to get more insights.

2.2.1 Adding candidate features, transforming, and scaling them

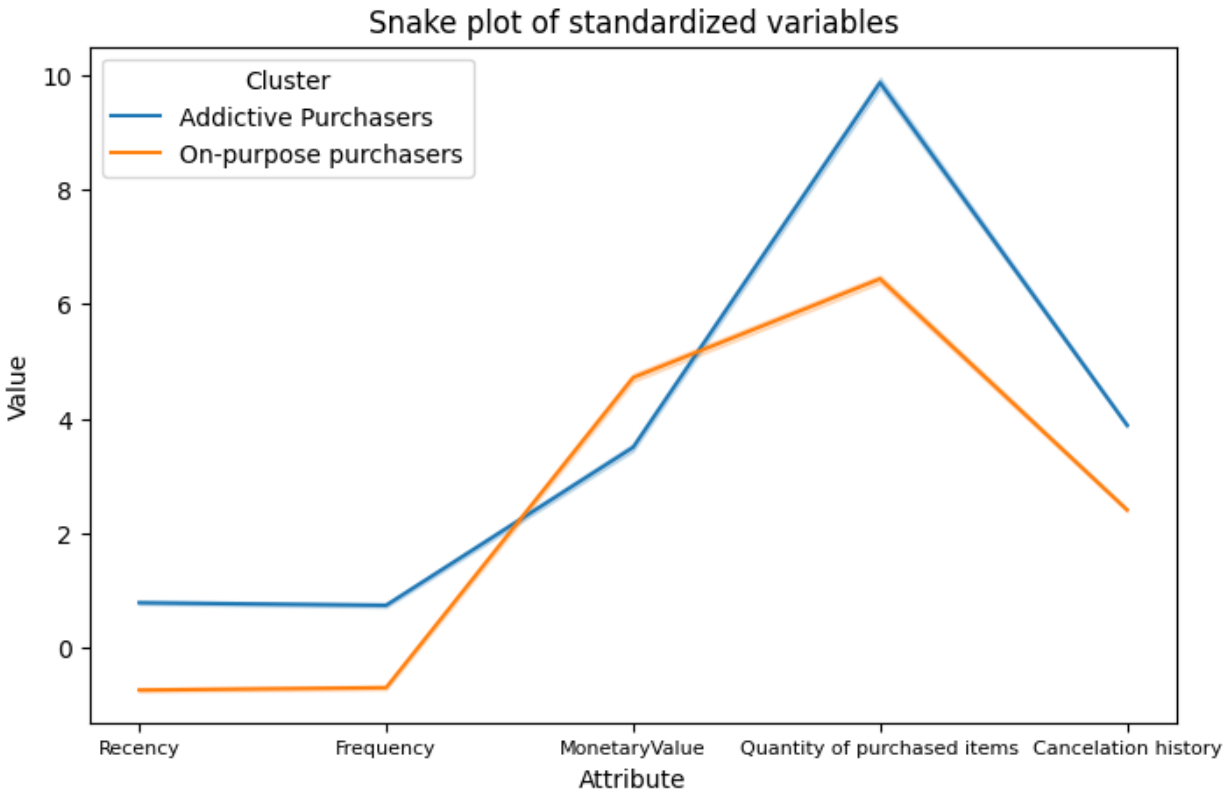
With my business acumen understanding, once we look at the Monetary Value attribute, we should take Quantity of purchased items into account. Besides, 'cancelation history' is also a potential attribute that can provide insight about behavior of customers. The goal of final segments was interpretability so instead of going with RFM analysis, I looked for segments with more attributes which could distinguish the final groups with informative insights.

After normalizing and scaling the data, we applied K-means clustering method to figure out the 2 best groups of customers. The whole data pipeline process is provided in the attached code ipynb file.



2.2.2 Distinguishing each segment characteristics

The snake plot can help us to imply the general characteristics of each groups:



Why these groups named “Addictive Purchasers” and “On-purpose Purchasers”?

- **Addictive Purchasers**:
  - They have the highest number of purchased items during the year with low quantity of cancellation history. However, they haven’t spent the as much money as the other group, but their frequency and recency are higher. It can be implied that they are eager for purchase everything very that they are interested in, with a high amount of quantity but low price.
- **On-purpose Purchasers**:
  - They have much lower number of purchased items with lower frequency and recency compare with the “Addictive” group. However, they spend more money. It can be assumed that they only purchase what they need, and their behavior is target on premium product with high price.

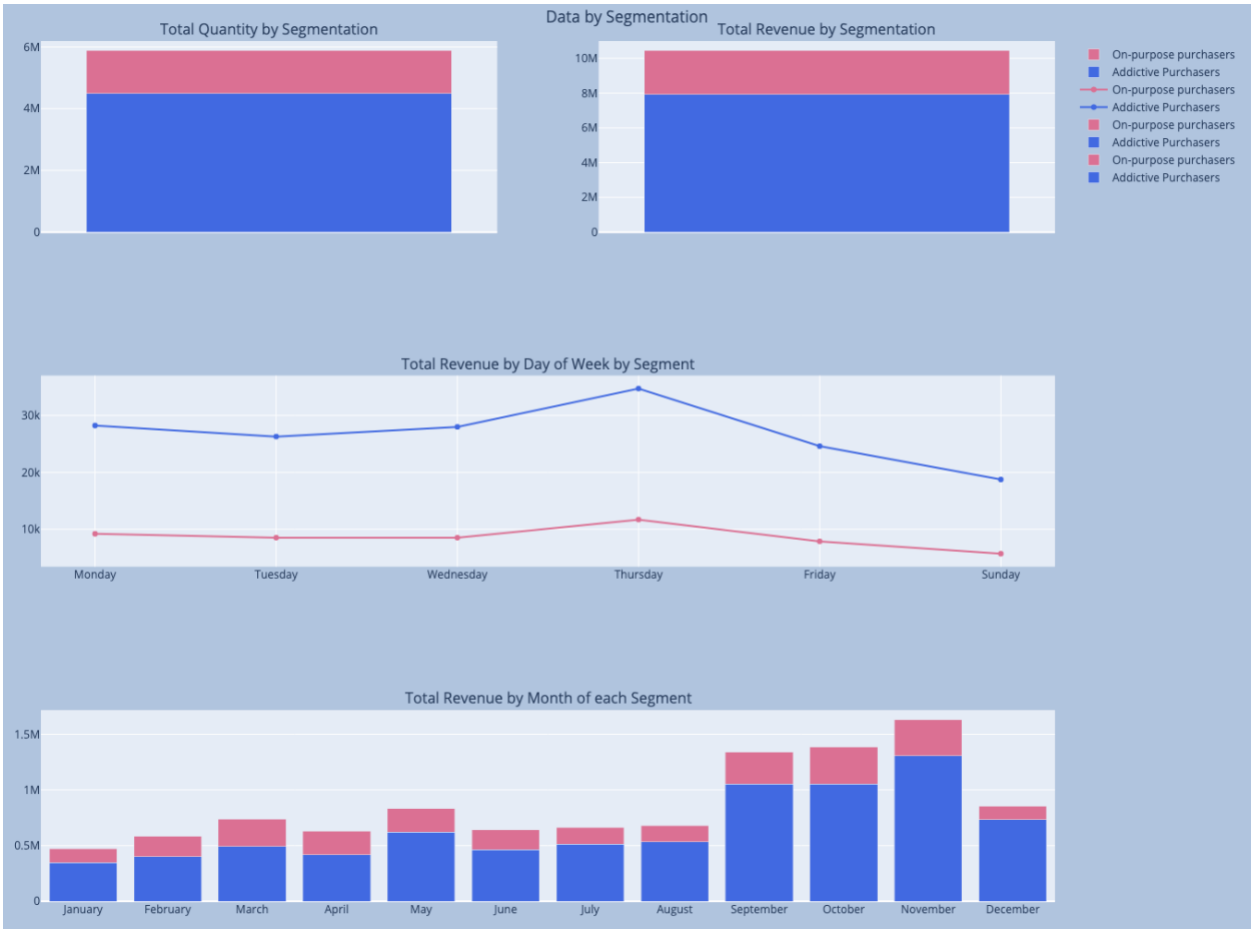
2.3 Suggestion for marketing plan

With K-means clustering method, we now have two different customer group with different behavioral pattern from the other, each has their own unique sets of characteristics, and we can have a directional strategic plan for our marketing and customer service team. The following are the suggestions to contribute for marketing plan, based on their shopping behavior.

- **Addictive Purchasers:**
  - Discover their favorite category, product to prioritize and utilize investment in recruiting new customers like them.

- Figure out their favorite promotional program and point of trigger to encourage them purchase more, and utilize the program to apply for another product (slow moving stocks, for example)

- **On-purpose Purchasers:**
  - Focus on their most purchased categories, products.
  - Deep dive into why they canceled their orders in the history to reduce the cancelation rate and increase the purchase.
  - Focus on promote product features regularly to make them remember when they have demand as well as to persuade them better to make purchase decision.



Marketing budget allocation suggestion:

- 70% budget should focus on bigger customer segment which is the ‘Addictive Purchasers’ since this is the main source of revenue.
- Promotion programs can be considered to be pushed strongly on weekdays rather than weekends to encourage more purchases.
- September, October, November are the top priority months for marketing plan.