

Nadeem Irfan

21-September-2023

**Capstone**

**Project**

**Project-2: Online Retail Store**

Table of content

1. Problem Statement
2. Project Objective
3. Data Description
4. Data Pre-processing Steps and Inspiration
5. Choosing the Algorithm for the Project
6. Motivation and Reasons For Choosing the Algorithm
7. Assumptions
8. Model Evaluation and Techniques
9. Inferences from the Same

10) Future Possibilities of the Project

11) Conclusion

12) References

13) Appendix

Problem Statement

An online retail store (E-commerce site) has various transactions recorded by different customers, who participate in it. This project undertakes to review the purchase records from this online retail store, and provide useful insights to understand the purchase patterns of the various customers.

Project Objective

An online retail store data is provided which has various transactions recorded by different customers. The task is to come up with useful insights about the customer purchasing history that can be added advantage for the online retailers. And to segment the customers based on their purchasing behaviour.

Introduction

Companies have had to grow their profitability and company through time as a result of fierce competition in the business field to meet customer expectations and attract new clients depending on their desires. It’s tough and time-consuming to identify and respond to each customer’s needs. This is owing to the fact that, among other things, clients have a diverse set of aims, interests, and preferences.

Customer segmentation, as opposed to a "one-size-fits-all" strategy, divides customers into groups based on comparable characteristics or habits. Customer segmentation is a marketing strategy that divides a market into distinct, homogeneous groups. The data used in the customer segmentation strategy, which divides customers into categories, is based on a number of factors, including regional circumstances, economic patterns, and demographic trends, and behavioural patterns. A client segmentation technique can help a company’s marketing resources be better utilized.

Data Description

The available dataset contains 5,41,909 records (rows) and eight features (columns) as shown in the table below.

|  |  |
| --- | --- |
| **Feature Name** | **Description** |
| Invoice | Invoice number |
| StockCode | Product ID |
| Description | Product description |
| Quantity | Quantity of the product |
| InvoiceDate | Date of the invoice |
| Price | Price of the product per unit |
| CustomerID | Customer ID |
| Country | Region of purchase |

Description of the data,

From the given data of the company, it is observed that there are five lakhs, forty one thousand, nine hundred and nine (5,41,909) records and eight features captured. Further Data description is as follows:

1. IncoiceNo. : Invoice-number is a nominal, 6-digit, integral number uniquely assigned to each transaction. If this code starts with ‘c’, it indicates a cancellation.
2. StockCode : Product (Item) code is Nominal data, a 5-digit integral number uniquely assigned to each distinct product.
3. Description: Product (Item) name, (Nominal data).
4. Quantity: The quantities if each product (item) per transaction. Nominal.
5. InvoiceDate: Invoice Date and time (Numeric) describes the day and time when each transaction was generated.
6. UnitPrice: Unit Price (Numeric) describes the product price per unit.
7. CustomerID: Customer number (Nominal), is 5-digit integral number uniquely assigned to each customer.
8. Country: Country name (Nominal), is the name of the country where each customer resides.

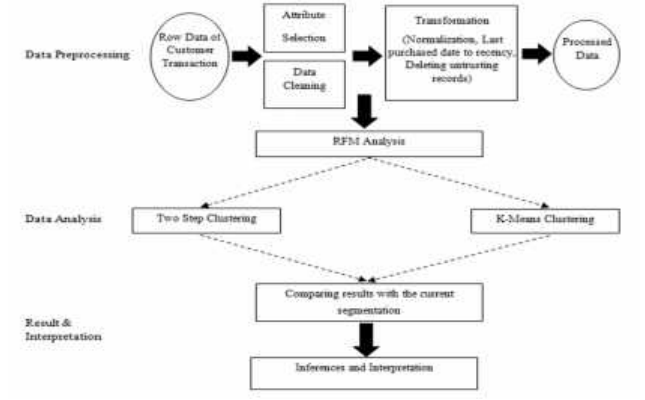
**Data Pre-processing Steps and Inspiration**

The pre-processing of the data includes the following steps:

1. Step 1: Load the data.
2. Step 2: Pre-processing of the data
   * 1. Confirming number of records in the data and how they are distributed.
     2. Check data type.
     3. Check for missing data, invalid entries, and duplicate records.
     4. Perform Feature engineering, add certain features which are derived from the given features, which will be useful in getting insights from that data
3. Step 3: Perform exploratory data analysis (EDA).

Here we exlore the given sales data, for customer in-order to see the TotalAmount (target\_value) or count of transactions for different features

1. Step 3: Model for Customer Segmentation:
2. RFM (Recency, Frequency, Monetary) model for segmenting the customers.
3. K-Means cluster of customer data
4. Step 4: Inferring the results from the model



**Exploratory Data Analysis**

Here we explore the given sales data, for customer in-order to see the TotalAmount (target\_value) or count of transactions for different features

a) Top 10 products which are in sale (take a count of a given product)

b) Top 10 customers who are frequent

c) Top 5 countries which has highest Sales

d) Top 5 countries which has Least Sales

e) Top 5 customers who spend the most

f) Frequency of sales on different days

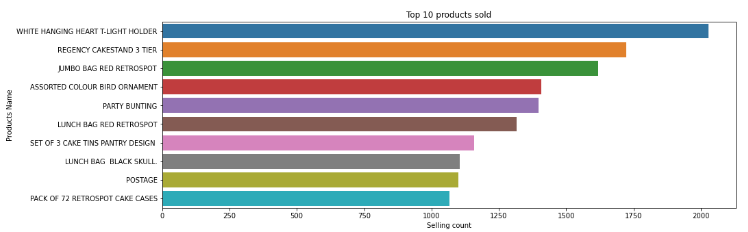
g) Sales on different Months

h) TotalAmount per month (and Year wise)

i) Why the December-2011 sales is less than nov-2011 sales

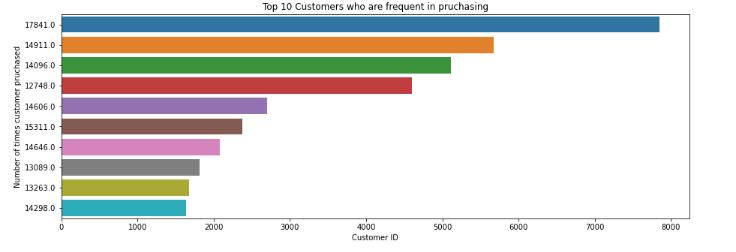
j) Sales on different part of the day

a) Top 10 products which are in sale (take a count of a given product)



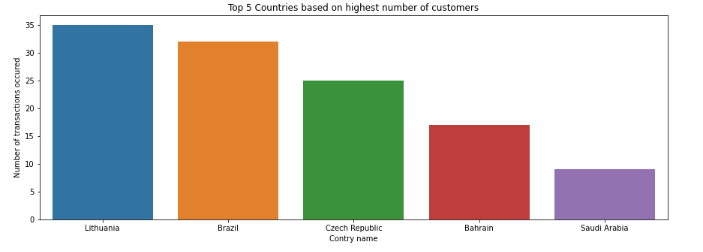
Here we see that, ‘WHITE HANGING HEART T-LIGHT HOLDER’ has the highest purchase, which is purchased 2028 times and the second highest product sold is ‘REGENCY CAKESTAND 3 TIER’, which is sold 1724 times

b) Top 10 customers who are frequent



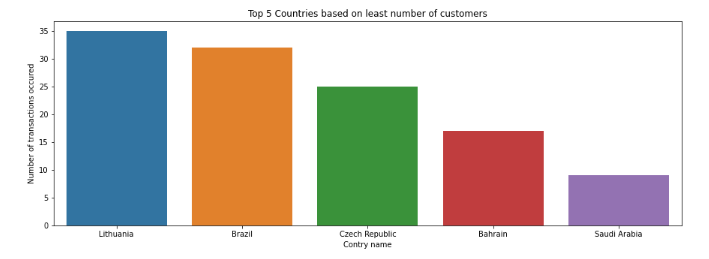
Customer ‘17841.0’ is the most frequent purchaser with frequency of 7847

c) Top 5 countries which has highest Sales



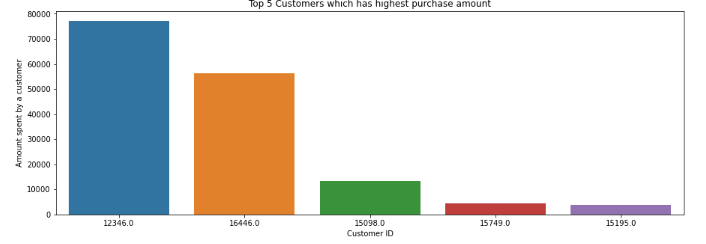
United Kingdom has the highest number of tractions with count as 354345, but the 4 other countries (Germany, France, Ireland, and Spain) almost similar sales count. United Kingdom has significantly larger amount of sales (as an outlier among all other countries)

d) Top 5 countries which has Least Sales



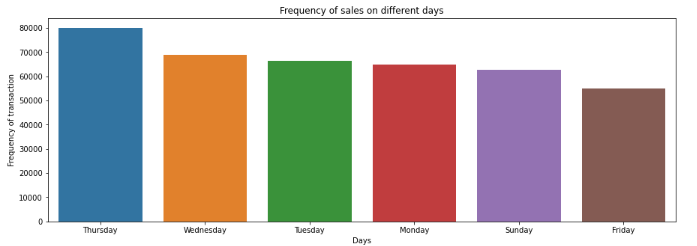
‘Lithuania’, ‘Brazil’, ‘Czech Republic’, ‘Bahrain’, ‘Saudi Arabia’ are the 5 countries with least sales, and the lowest is Saudi Arabia with number of transactions as 9

e) Top 5 customers who spend the most



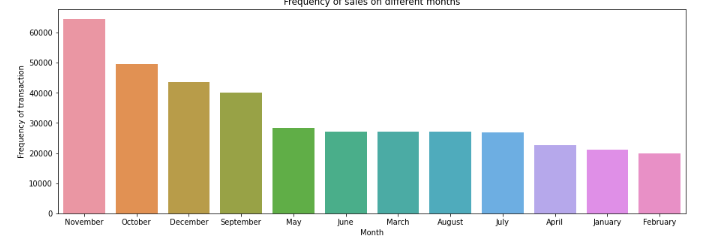
Customer ‘12346.0’ has the highest purchase, with an amount of 77183.60, However customer ‘17841.0’ is the most frequent purchaser with frequency of 7847

f) Frequency of sales on different days



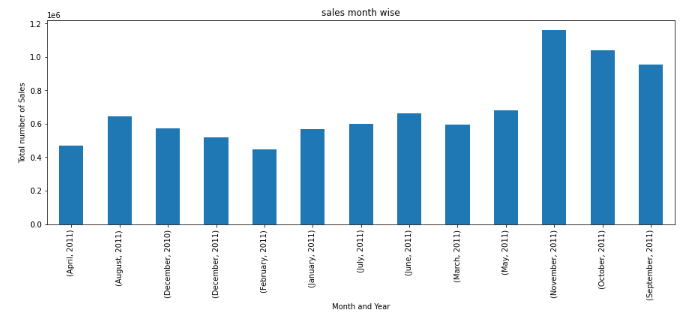
Thursday has the highest number of purchase-items of 80052 items sale. The other days have similar amount of purchase around 6000 to 7000 items sale. However there is no sale on Saturday.

g) Sales on different Months



November has the highest number of sales, and then on October, December, September has significantly higher sales than other months

h) Total Amount per month (and Year wise)

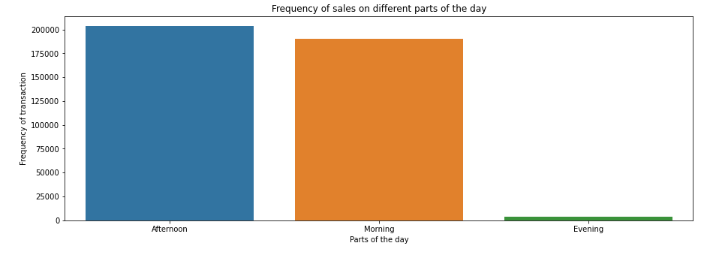


For 2010 there is only December Sales, and for 2011 entire year sales happened

i) Why the December-2011 sales is less than nov-2011 sales

Inference (Dec-2011 vs Nov-2011): Because Dec-2011 has sales for only 10 days compared to November-2011

j) Sales on different part of the day



There are high amount of sales in morning and afternoon of the day, but very less in evening

**Summary of the EDA performed**

Following is the summary for the EDA we performed:

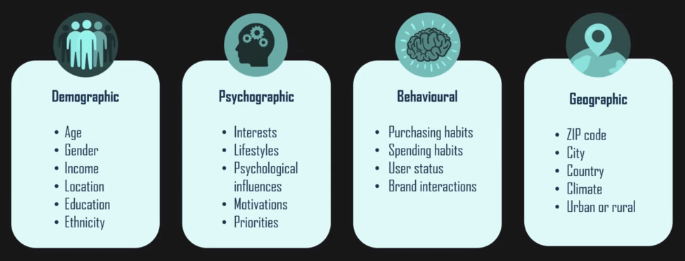
1. Here we see that, ‘WHITE HANGING HEART T-LIGHT HOLDER’ has the highest purchase, which is purchased 2028 times and the second highest product sold is ‘REGENCY CAKESTAND 3 TIER’, which is sold 1724 times.
2. Customer ‘12346.0’ has the highest purchase, with an amount of 77183.60, However customer ‘17841.0’ is the most frequent purchaser with frequency of 7847.
3. Thursday has the highest number of purchase-items of 80052 items sale. The other days have similar amount of purchase around 6000 to 7000 items sale. However there is no sale on Saturday.
4. There are high amount of sales in morning and afternoon of the day, but very less in evening.
5. November has the highest number of sales, then on October, December, September has significantly higher sales than other months
6. For 2010 there is only December Sales, and for 2011 entire year sales happened.

**Model evaluation and techniques**

**Choosing the Algorithm for the Project**

Studying the customers purchase behaviour can be evaluated and segmented using many robust Machine-Learning algorithms.

Marketing strategy for better sales is done by understanding the customer’s behaviour; they can be segmented by studying their Demographic, Psychological, Behavioural or Geographic insights. [[1]](#Reference_1)



One of the ways to study the customer’s purchasing behaviour is by using RFM model, which is rank / categorize the customer based on their purchasing behaviour.

**Model Selection:**

In order to understand the behaviour of customers purchase, we need to derive their behaviour of purchase using parameters like Recency of the customer (R), Frequency of the customer (F) and Monetary of the customer (M). Then using their scores derived from RFM, we can segment them using a very famous unsupervised learning model named, K-Means Clustering.

1. **RFM:** [**[2]**](#Reference_2)

Segmenting the customers according to their data became vital in this context. RFM (recency, frequency and monetary) values have been used for many years to identify which customers valuable for the company, which customers need promotional activities, etc. Data-mining tools and techniques widely have been used by organizations and individuals to analysis their stored data. Clustering, which one of the tasks of data mining has been used to group people, objects, etc.

RFM is a method to analyse customer value. RFM stands for Recency, Frequency, and Monetary.

* Recency: How recently did the customer visit our website or how recently did a customer purchase.
* Frequency: How often do they visit or how often do they purchase.
* Monetary: How much revenue we get from their visit or how much do they spend when they purchase.

Why we need RFM?

RFM analysis is a marketing framework that is used to understand and analyse customer behaviour based on the 3 factors Recency, Frequency and Monetary.

RFM analysis will help Business to segment their customer base into different homogenous groups so that they can engage with each other with each group with different targeted marketing strategy.

1. **Customer Segmentation Using K-Means Clustering:**

The most well-known unsupervised partitioning clustering approach is K-Means Clustering. This clustering approach, commonly known as the centroid-based technique, divides data into non-hierarchical categories.

**Model Evaluation and Technique**

[[3]](#Reference_3)[[4]](#Reference_4) The dataset is separated into a collection of k groups in this sort of partitioning, where K is the number of pre-defined groups or clusters. When compared to another cluster centroid, the cluster centre is built so that the distance between data points in one cluster is as low as possible.

**Model Evaluation**

The following techniques and steps were involved in the evaluation of the model

1. Load necessary libraries
2. Load the dataset
3. Deriving scores of each customer purchase, and categorising them.
4. With respect to each customer, extracting Recency, Frequency and Monetary.
5. Examine their distribution
6. Ranking and partitioning the customers into categories, according to the score produced from Recency, Frequency and Monetary.
7. Examine the distribution of categories which are derived for different customers.
8. Clustering the customers as per their categories.
9. Extracting the results

**Model Evaluation and Technique**

**Model Approach:**

1. RFM model:

* To get the recency, frequency and monetary of a customer, where
* Recency = Latest Date of Invoice Date,
* Frequency = Count of invoice No. of transactions,
* Monetary = Sum of Total Amount for each customer
* Study the distribution of Recency, Frequency and Monetary data
* Split the RFM score into 4 parts and rank the customers accordingly.
* Also Plot the graph for Recency vs Frequency, Frequency vs Monetary, Recency vs Monetary.

1. K-Means clustering model:

* Normalize and standardize the sales data, because K-Means accepts normally distributed data.
* Get the optimal K-value using elbow method
* Train the model with optimal k-value.
* Labelling the customers with corresponding cluster of their RFM\_Score.

**Inference from the Project**

**Model result**

First we have categorised the customers behaviour using R.F.M model, and then segmented then using K-Means clustering.

**1. R.F.M values:**

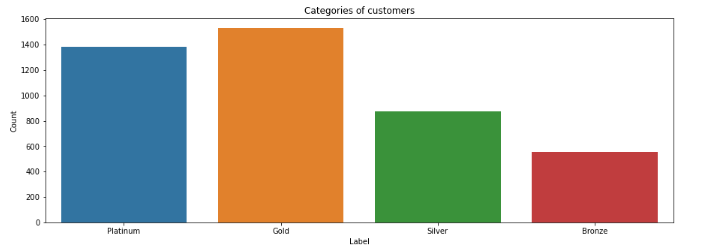
We have calculated the RFM values of different customers and we have categorized them into 4 categories as Platinum, Gold, Silver, and Bronze. Their respective counts and descriptions are

Platinum customers = 1380 (less recency but high frequency and heavy spending)

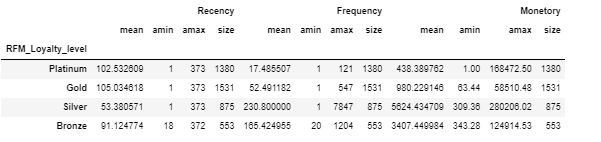
Gold customers = 1531 (good recency, frequency and monetary)

Silver customers = 875 (high frequency, low frequency and low spending)

Bronze customers = 553 (very high recency but very low frequency and spending)

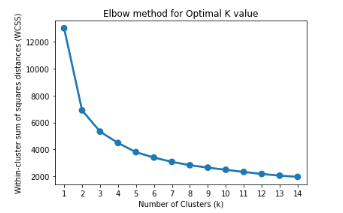


A detailed summary of the 4 categories of customers are as follows:



**2. K-Means clustering:**

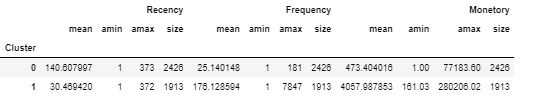
After categorizing the customers as per their Recency, Frequency and Monetary scores, optimal K-value is calculated as 2 using Elbow method. Following is the result of k-value elbow method.



The data used to train the K-Means cluster was RFM-Scores of different customers, and then customers are segmented into 2 clusters.

Cluster 0 has high recency rate but very low frequency and monetary. Cluster 0 contains 2426 customers.

Cluster 1 has low recency rate but they are frequent buyers and spends very high money than other customers as mean monetary value is very high. Cluster 1 contains 1913 customers.



**Future Possibilities:**

We can use more robust analysis for the clustering, using not only RFM but other metrics such as demographics or product features.

By enabling enterprises to better understanding both the customers’ and business functionality behaviour, Machine Learning has enabled companies to offer better/ targeted customer services leading to more loyal customers and ultimately improved sales revenue.

**Conclusion:**

The project undertook the study of an online retail company to study different customer’s behaviour. Some of the important findings from the report include the followings:

* Customers are segmented into 4 categories based on their purchasing behaviour and their attendance in the stores.
* They are also being clustered based on their evaluation of scores.

Following are the useful recommendations for improving the sales of the store, after studying the customer purchasing history.

1. Customers, who visit the store, are mostly in Gold and Platinum category, and less in Silver and less in Bronze category. This implies the sales are very good, for the existing customers.
2. However in order to improve more sales, studying the demand in market and the sales of the competitors, advertising needs to be changed in order to attract new customers.
3. Customers who have visited we need recommend similar products, which they purchase by studying their type of products likes. This will help those customers who do not purchase expensive product, can visit and explore more products.
4. At end yearend sales are very high, so it’s better to attract more Bronze level customer and new customers, by providing offers of discount.
5. Explore other services options that have worked well for similar companies, such as same-day or next-day home delivery.

It may just be that some customers with very low RFM score may just leave/ unsubscribe the online store, if they don’t like the products.

**References:**

1) Different marketing strategy to get customers.

(<https://www.yext.com/blog/2019/06/how-demographic-psychographic-and-behavioral-marketing-inform-your-intent-marketing-strategy> )

2) Customer Segmentation by using RFM model and clustering methods: A case study for retail industry, International Journal of Contemporary Economics and Administrative Sciences, Volume: 8, Issue: 1, Year 2018, pp. 1-19.

3) “Customer Segmentation Using K Means Clustering,” Towards Data Science, Apr. 2019.

4) Customer Segmentation using K-Means Clustering

(<https://www.ijert.org/> ) International Journal of Engineering Research & Technology (IJERT), Vol. 11 Issue 03, March-2022.

**Appendix**

There is Code file named ‘Online Retail Sales Analysis.ipynb’ is attached to the zip file, as a reference, to see the functions used to evaluate the result in this report.