DATA MINING PROJECT REPORT

RETAIL STORE CUSTOMER SEGMENTATION

NITISHA BHARATHI S

16PT25

November 18, 2020

Contents

ABSTRACT

As retail industry emerges there is increasing motivation for retailers to look for data or strategies that can help them segment or describe their customers in a succinct, but informative manner. This work focuses on *Customer Segmentation.* by integrating machine learning practices and conventional business understandings, the paths to understand customer behaviour became more intertwined and answers question like *What segments or groups of customers do we have?.*

The tasks performed in this work include customer segmentation, customer classification, RFM Analysis and Market Basket Analysis. The findings are that there are roughly five or six clusters of customers with each cluster having unique purchasing traits that define them.

CHAPTER 1
INTRODUCTION

1.1 The Business Problem

Any company in retail, no matter the industry, ends up collecting, creating, and manipulating data over the course of their lifespan. These data are produced and recorded in a variety of contexts, most notably in the form of shipments, tickets, employee logs, and digital interactions. Each of these instances of data describes a small piece of how the company operates, for better or for worse. The more access to data that one has, the better the picture that the data can delineate. With a clear picture made from data, details previously unseen begin to emerge that spur new insights and innovations.

Companies that utilize proper data science and data mining practices allow themselves to dig further into their own operating strategies, which in turn allows them to optimize their commercial practices. As a result, there are increasing motivations for investigating phenomena and data that cannot be simply answered: *Why is product B purchased more on the first Saturday of every month compared to other weekends?*, *If a customer bought product B, will they like product C?, What are the defining traits of our customers*? *Can we predict what customers will want to buy*? These questions will be the broad focus of this work.

1.2 Acquisition of Data

Finding readied, usable data for analysis in a business context is a rarity. As such, it is imperative to collect as much data as possible, but also in a format that meets a wide variety of financial, ethical, and computational considerations. Here, we use the *Online Retail Dataset*[1] for detailed analysis. The attributes of the data are described below

[1] `https://archive.ics.uci.edu/ml/datasets/online+retail`

- *InvoiceNo*: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction.
- *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*: Invoice 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.

1.3 Scope of Analysis

The scope of the paper is limited to the following four intertwined goals:

1. To cluster customers based on common purchasing behaviors for future operations/marketing projects
2. To classify the new customers using the knowledge from previous clustering
3. RFM Analysis to understand the customer behaviour
4. Market Basket Analysis to know the items buying pattern of the customers

This report is structured as follows: Chapter 2 for Exploratory Data Analysis. Chapter 3 for Customer Segmentation. Chapter 4 for Customer Classification. Chapter 5 for RFM Analysis. Chapter 6 for Market Basket Analysis.

CHAPTER 2
EXPLORATORY DATA ANALYSIS

2.1 Data Preparation

Data preparation is the act of manipulating raw data into a form that can readily and accurately be analysed. Although several attributes from the data table were listed, not all the variables could be used in the analysis as it is. Certain variables, such as the ID columns, provide necessary information to corroborate data and keep accurate calculations between instances, but are not necessarily features that merit analysis. On a similar note, features, such as the time of a specific transaction contain essential information for mining, but need to be transformed into a more usable format. Attribute such as description are on an item-based level, which require a separate transformation of their own to understand what items the customer has purchased. Nonetheless, the salient point is that it is necessary to consider the raw data, examine its format and original features, and transform them into a workable format for the task at hand.

2.1.1 Handling Missing Values

Missing data refers to those data-points whose values are not present. The concept of missing values is important to understand in order to successfully manage data. If the missing values are not handled properly then we may end up drawing an inaccurate inference about the data.

In the Online Retailer dataset, out of 541909 data points Customer ID and Description has 135080 and 1454 values missing respectively. Imputation cannot be done to both the attributes as Customer ID refers to each customer which on imputing will totally change the meaning of data. Hence the rows with missing values are dropped resulting in 406829 datapoints.

2.1.2 Feature Engineering

The process of creating or extracting features from raw data is commonly referred to as feature engineering. It is the most important step of data preprocessing because it establishes the features that the model will consider when clustering. Essentially, feature engineering involves inspecting and manipulating the raw data to somehow extract features that are worthwhile for analysis. The following are the feature engineering done

1. Day, month, year are extracted from the *InvoiceDate* attribute which would help in time analysis.
2. A new attribute *TotalCost* is created from the product of *UnitPrice* and *Quantity*.
3. All the cancelled orders are removed.

In addition to this, general preprocessing like converting *Description* to lowercase, removing punctuations etc were done.

2.2 Data Analysis

The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

2.2.1 Description Word Cloud

The word cloud of description is plotted to understand various products present. Figure 2.1 shows words like light, holder, bag, sign etc to be highly prevailing in description.

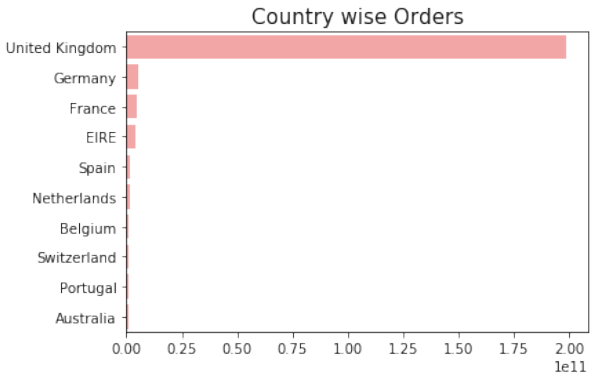
Figure 2.1: Description Attribute Word Cloud

2.2.2 Countrywise Insights

Country wise transaction insights are necessary to understand which country customers should be targeted. It answers questions like *Which country citizens make most number of purchases*, *Which country provides the most profit and least profit?*, *In which country sales should be improved?* Figure 2.2 shows that United Kingdom tops the no of orders made and from figure 2.3 (a) it is evident that revenue is also high from UK. Other than UK, the best buying (Figure 2.2 (b)) countries are Netherlands, Australia, Singapore etc. Hence, these countries can be targeted even more as chances of expanding business and customers welcoming it here is more.

2.2.3 Customer Insights

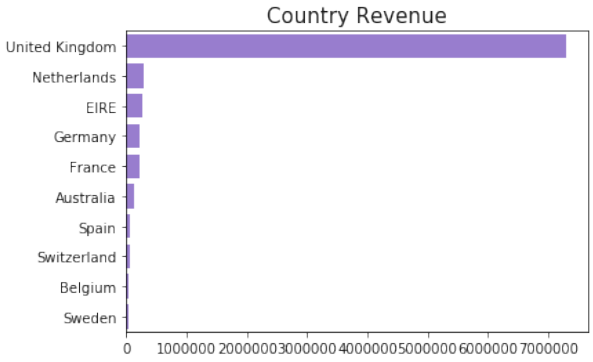
This section is to understand *Who are our best customers?*, *Which customer spends the most?* but plotting just the customer against no of orders (Figure 2.4(b)) or amount spent (Figure 2.4(a)) doesn't give a proper explanation. Customer Segmentation using clustering and RFM analysis will give a more clearer explanation on the customer behaviour.



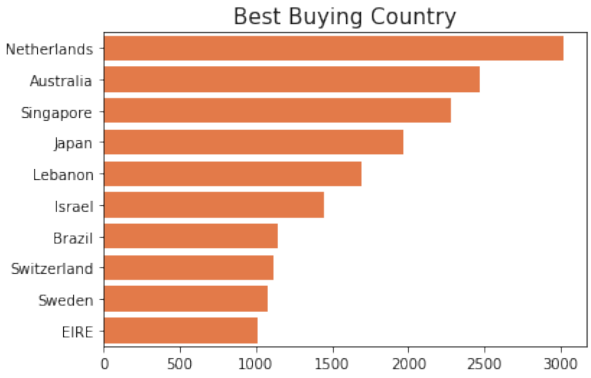
(a) Country wise Orders Barplot

(b) Country wise Orders Chromomap

Figure 2.2: Country wise Orders

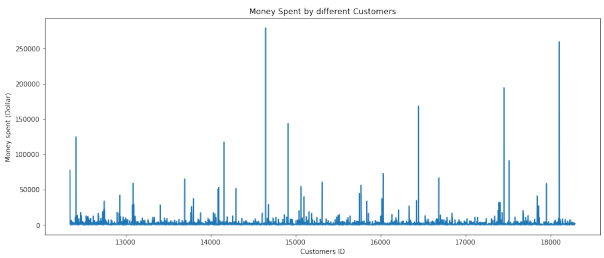


(a) Country wise Revenue

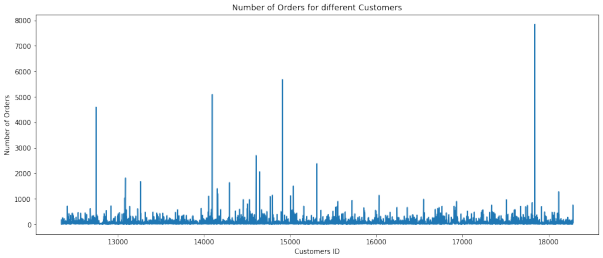


(b) Best Buying Countries

Figure 2.3: Country wise Orders



(a) Amount Spent by the Customers



(b) Orders by Customers

Figure 2.4: Amount Spent and Orders by the Customers

2.2.4 Transaction Time Insights

Time series analysis are required to understand the prime time of business. Using this, retail store can determine which period is the best for discounts, flash sales etc. From Figure 2.5 it is evident that Thursdays are the best day and store is not available on Saturdays. The prime time of sale is around 11AM to 1PM (Figure 2.6 and November is the best sale month (Figure 2.7).

2.3 EDA Summary

This section summarizes the insights drawn from performing Exploratory Data Analysis, in short.

1. *November* has the highest number of sale
2. Highest number of customers has arrived on *Thursday*
3. 11AM - 1PM is the prime time for sale
4. Customers from UK has made the most orders
5. Most income is from UK
6. The best set of customers are from Netherlands, Australia, Singapore.
7. Most prevailing items are lunch bag, doll, candles etc

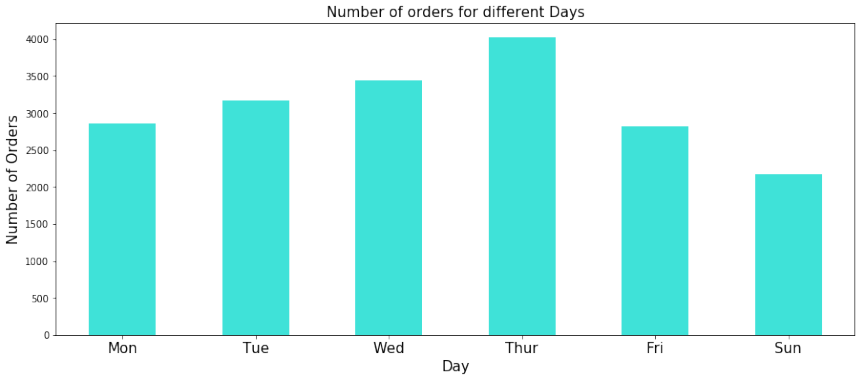


Figure 2.5: Day-wise Orders

Figure 2.6: Hour-wise Orders

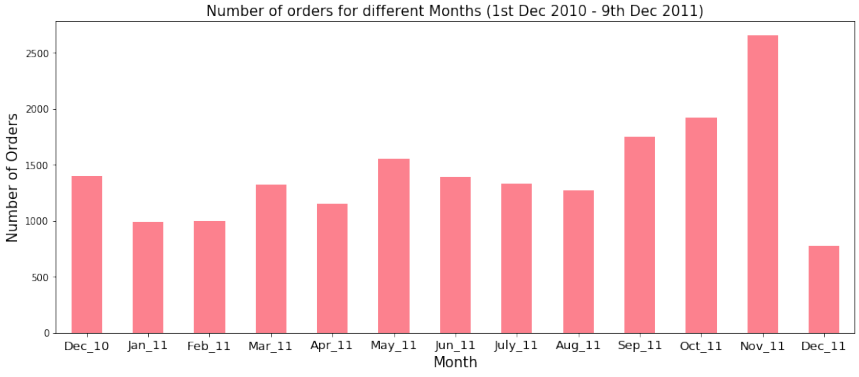


Figure 2.7: Month-wise Orders

CHAPTER 3
CUSTOMER SEGMENTATION

3.1 Brief Introduction

For a retailer, understanding the components of their consumer base is key for maximizing their potential in a market; the retailer that attracts the most customers will acquire the most market share. In fact, the high costs of gaining a new customer or getting back an old customer force retailers to seriously consider how to allocate resources to optimize not just volume of customers, but the retention of them as well. Additionally, it is a common understanding in the retail industry that the Pareto Principle—more likely than not—applies to the company: 80% of the profits come from 20% of the customers. One crucial reason why this principle holds is because retail businesses thrive on repeat purchases. As a consequence, a net change of one customer can significantly impact a business' profit in the long run. Therefore, it is generally in the best interest of the retailer to devote efforts to retaining customers by understanding them on as deep of a level as necessary.

Hence, Customer segmentation is the process of dividing customers into groups based on common characteristics so companies can market to each group effectively and appropriately. In order to perform customer segmentation analysis at a high level, retailers have begun to incorporate aspects of data mining into the analysis of their customers. More specifically, retailers are utilizing unsupervised machine learning tools such as clustering and dimensionality reduction to approach analysis in ways that cannot be matched without machine learning. Retailers across all industries are attempting to leverage clustering algorithms such as K-Means or hierarchical clustering to more accurately and quickly segment their customers. The faster and better retailers are able to cluster their customers, the quicker they can market to them and thus acquire market share.

3.2 Product Clustering using K-Means

At first, the products are clustered into categories. A short description of the products are given in the *Description* attribute. Noun keywords are extracted from the Description by performing stemming and POS tags. These words cannot be directly given as input to the clustering algorithm hence a matrix $\mathcal{X}$ is created whose shape is No of Products * No of terms.

$$\mathcal{X}_{ij} = \begin{cases} 1 & \text{if product } i \text{ description contains the term } j \\ 0 & \text{Otherwise} \end{cases}$$

Before applying the algorithm one important step is to choose $\mathcal{K}$ in K-means. This can done using the *silhouette score* for each cluster. The silhouette score is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from 1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters. Figure 3.1 shows that the highest score at n=8 since, 6 to 8 there not much difference we choose the $\mathcal{K}$ value as 6. This $\mathcal{X}$ matrix is fed as

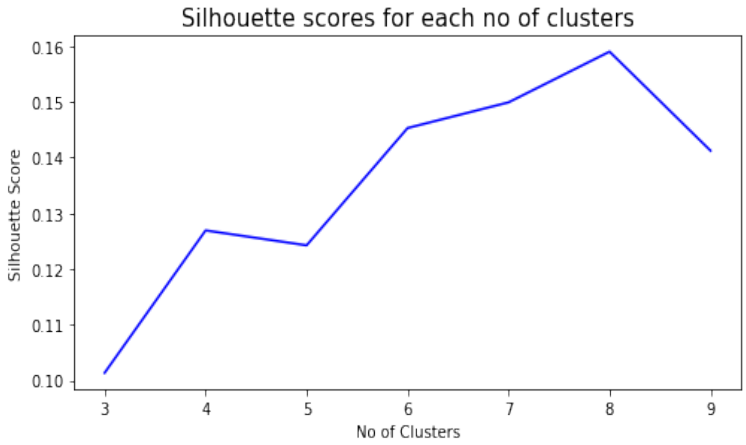


Figure 3.1: Silhouette Scores for different $\mathcal{K}$ values.

input to the K- means algorithm with $\mathcal{K}$ as 6.

3.2.1 Product Cluster Analysis and Visualization

Post Clustering, analysis and visualization is required to understand how the items are clustered. Figure 3.2 shows that clusters 2 and 3 has most number of items. The TSNE visualization from Figure 3.3 shows the clusters are segregated with few overlaps. From the clusters in Figure 3.4 we can see that one cluster is associated with christmas items i.e candles, cale,decoration etc. Another cluster is associated with luxuary items like necklace, silver, bracelet etc. Another cluster realted to vintage stuffs like retro, polkadot, retrospot. But it is observed that a few words occur in more than one cluster which make it tough to distinguish. In order to ensure if the clusters are distinct in lower dimensions Principal Component Analysis is performed.

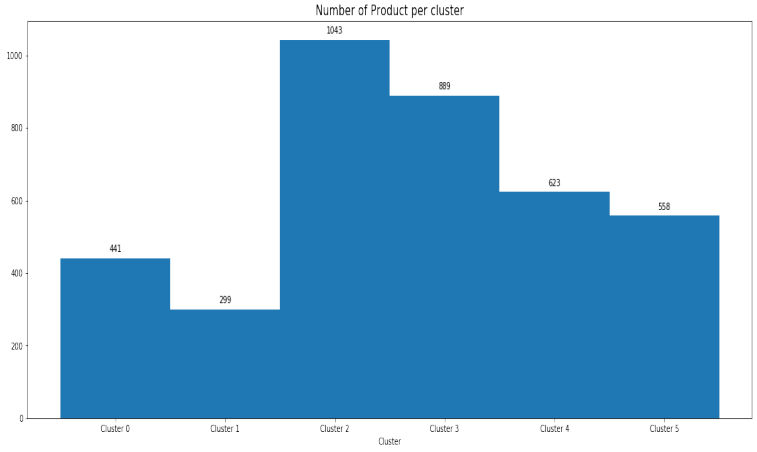


Figure 3.2: Number of Items in each Cluster

3.2.2 Principal Component Analysis

The $\mathcal{X}$ matrix defined for the item clustering has very high dimension and we see in Figure 3.4 that a few words overlaps. To check if the clustering is better in lower dimension Principal Component Analysis (PCA) is performed.

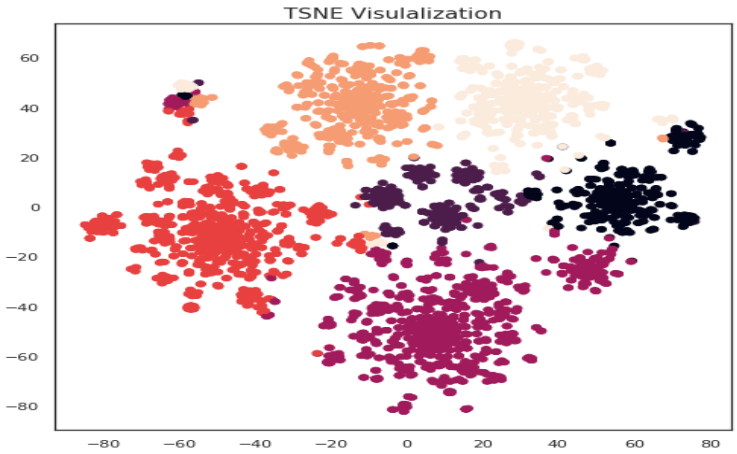


Figure 3.3: TSNE Visualization of Clusters

Figure 3.4: Item Clusters

From Figure 3.5 it is evident that even in lower dimensions there are overlaps. Hence trying other clustering algorithms are a better representation of matrix $\mathcal{X}$ may work out.

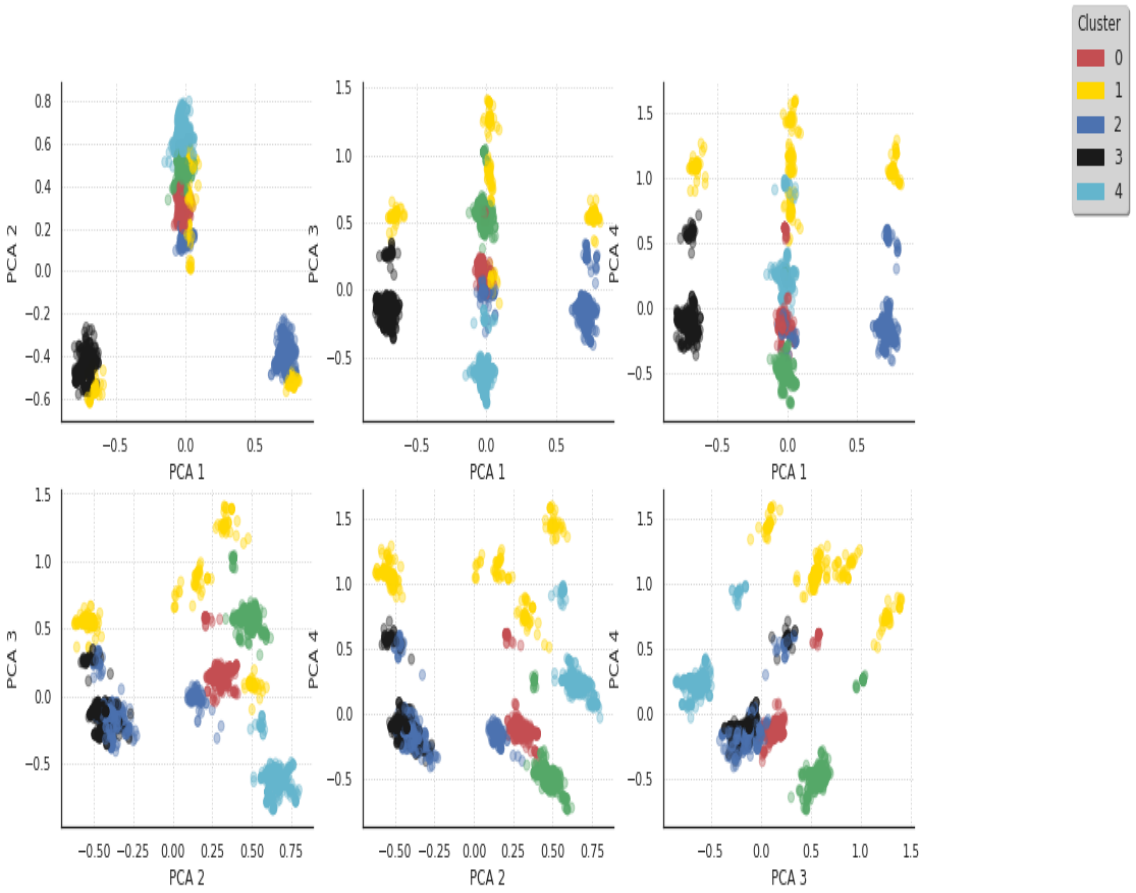


Figure 3.5: Clusters after PCA

3.3 Customer Clustering using K-Means

The item cluster values are added as another attribute to the data which would help in customer clustering. A few preparations were on the data and normalized to make the data applicable for K-means algorithm. K-means algorithm was applied to the data with $\mathcal{K}$ value as 6.

3.3.1 Customer Cluster Analysis and Visualization

Post Clustering, analysis and visualization is required to understand how the consumers are clustered. Figure 3.6 shows that clusters 1, 2, 3 have a very few customers. The TSNE visualization from Figure 3.7 shows the clusters are segregated.

Figure 3.8 is very important to understand the customer segmentation. It is a plot between the amount spent and time taken between consecutive orders.

- Customers in cluster 0 spend very less but are regular to the store.
- Customers in cluster 1 spend a lot and are regular to the store. (The best)
- Customers in cluster 2 spend very less and not regular.
- Customers in cluster 3 spend in a average way but are regular.
- Customers in cluster 4 spend below average and not much regular.
- Customers in cluster 5 are very similar to cluster 0 customers.

This can help in understanding the business requirement and which cluster should be targeted and how.

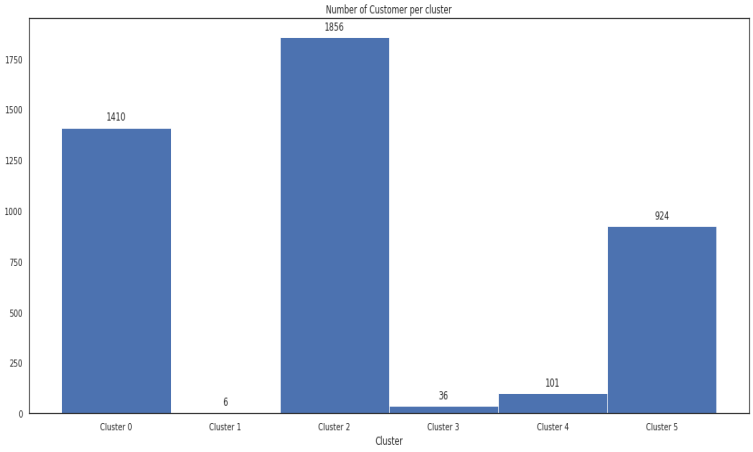


Figure 3.6: Number of Customers in each Cluster

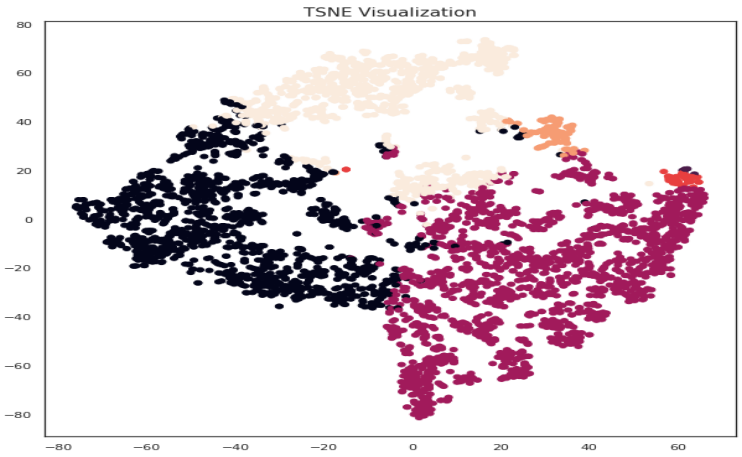


Figure 3.7: TSNE Visualization of Clusters

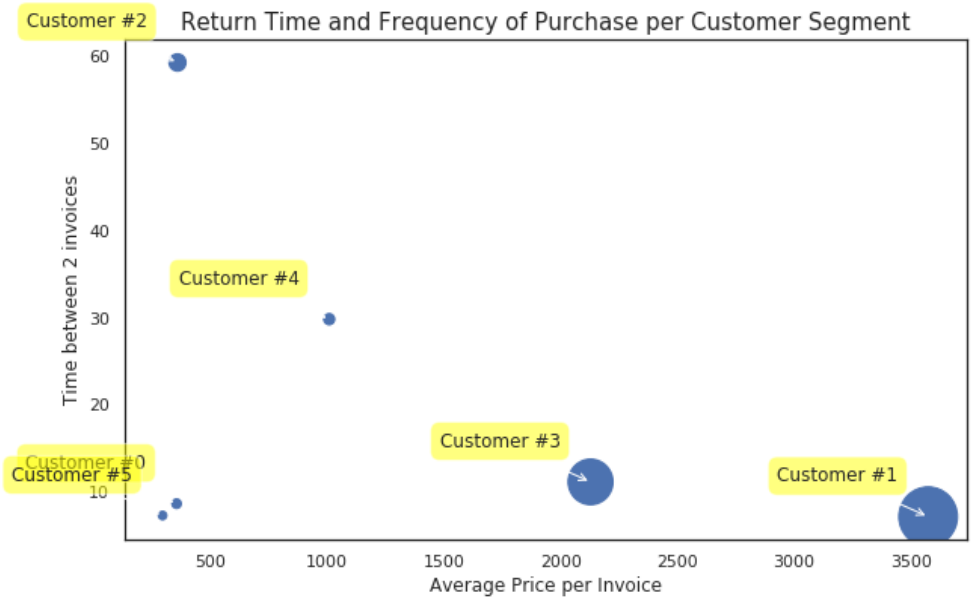


Figure 3.8: Customer Groups

CHAPTER 4
CUSTOMER CLASSIFICATION

4.1 Brief Introduction

We used clustering to segment the customers in chapter 3 which is unsupervised because no prior information about the customer behaviour was available. Now that we have segmented the customers into clusters we have the cluster information. This can be used to classify new set of customers. When a different set of customers are incoming, the knowledge learned from the clusters can be used to segregate them and understand how these customers should be targeted.

4.2 Supervised Customer Segmentation

With the cluster number as target variable the dataset is separated as train and test set. Test set has 25% data points. Various machine learning algorithms were applied and their results are displayed in Table 4.1.

When compared to other linear models tree models performs very well. Light GBM gives best results followed by XG boost. SVM performs the least. The confusion matrix for the predictions are shown in Figure 4.1.

Model	Accuracy	F1 Measure
Decision Tree	97.23%	0.972
Random Forest	97.88%	0.981
SVM	42.62%	0.426
Logistic Regression	93.82%	0.938
Naive Bayes	93.63%	0.936
XG Boost	98.52%	0.985
Ada Boost	73.52%	0.735
Light GBM	98.43%	0.984
KNN	88.75%	0.887
Perceptron	75.28%	0.752

Table 4.1: Customer Classification Results

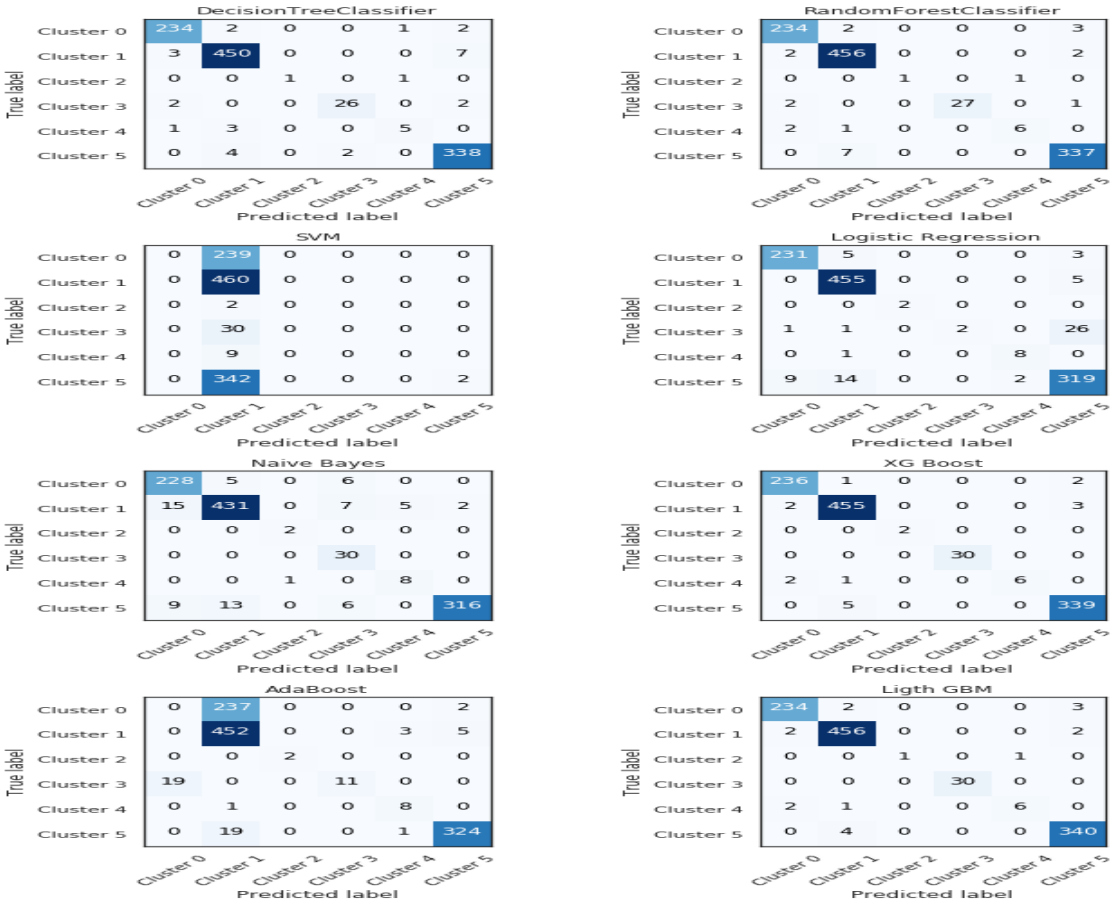


Figure 4.1: Item Clusters

CHAPTER 5
RFM ANALYSIS

5.1 Brief Introduction

RFM (Recency Frequency Monetary) analysis is a marketing technique used for analyzing customer behavior such as how recently a customer has purchased (*Recency*), how often the customer purchases (*Frequency*), and how much the customer spends (*Monetary*). It is a useful method to improve customer segmentation by dividing customers into various groups for future personalization services and to identify customers who are more likely to respond to promotions.

5.2 Steps of RFM

1. Calculate the Recency, Frequency, Monetary values for each customer.
2. Add segment bin values to RFM table using quartile.
3. Sort the customer RFM score in ascending order.

The recency, frequency, and monetary value are calculated by aggregating the InvoiceDate, TotalCost and CustomerID. These values are divided into bins and the values to which bin they belong is added as another attribute. *RFM score* is a 3 digit number representing each bin. Using the RFM score the customers are segmented.

5.3 RFM Visualization

Figure 5.1 shows how the customers are clustered using RFM technique. The various segments of customers are described below

1. *Core* - Best Customers

   Highly engaged customers who have bought the most recent, the most often, and generated the most revenue.

2. *Loyal* - Most Loyal Customers

   Customers who buy the most often from the store.

3. *Whales* - Highest Paying Customers

   Customers who have generated the most revenue for the store.

4. *Rookies* - New Customers

   First time buyers.

5. *Slipping* - Once Loyal, Now Gone

   Great past customers who haven't bought in awhile.

6. *Regular* - Average Customers

   Customer who have average metrics across each RFM scores.

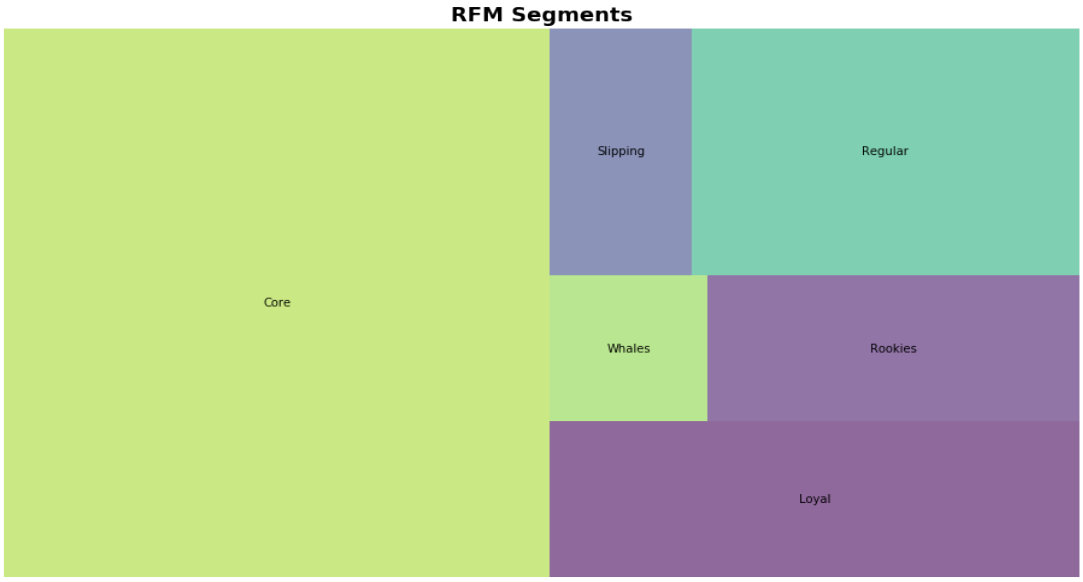


Figure 5.1: RFM Segments

CHAPTER 6
MARKET BASKET ANALYSIS

6.1 Brief Introduction

Market Basket analysis is a data mining method focusing on discovering purchase patterns of the customers by extracting association or co-occurrences from a store's transactional data. For example, when the person checkout items in a supermarket all the details about their purchase goes into the transaction database. Later, this huge data of many customers are analyzed to determine the purchasing pattern of customers. Also decisions like which item to stock more, cross selling, up selling, store shelf arrangement are determined. A retailer must know the needs of customers and adapt to them. Market basket analysis is one possible way to find out which items can be put together.

6.2 Analysis using Apriori Algorithm

The data is converted to a binary format as Apriori accepts only categorical values. Then the rules are fetched by applying the Apriori Algorithm. But most of the rules were quite obvious ones. Better formatting of the data could provide interesting patterns. Sample rules are listed in Table 6.1.

Antecedent	Consequent
Candle, light	christmas light
Tea	saucer
necklace, bracelet	silver, box
bag	vintage pot, box

Table 6.1: Sample Apriori Rules

CHAPTER 7
CONCLUSION

The tasks performed in this work include customer segmentation, customer classification, RFM Analysis and Market Basket Analysis. The findings are that there are roughly five or six clusters of customers with each cluster having unique purchasing traits that define them. Customer classification performed well using Light GBM. RFM analysis of the customers showed potential customers who can be targeted. Market Basket Analysis showed obvious rules, no interesting patterns could be drawn

7.0.1 Future Work

The finding in these work can be extended by trying out other clustering algorithms. Recommendations using Item based Collaborative Filtering and User Based Collaborative Filter can be performed.

Bibliography

[1] UCI Dataset
`https://archive.ics.uci.edu/ml/datasets/online+retail`

[2] Ryan Henry Papetti. *Customer Segmentation Analysis of Cannabis Retail Data : A Machine Learning Approach*. 2019.

[3] RFM Analysis in Python
`https://www.datacamp.com/community/tutorials/introduction-customer-segmentation-python`