Capstone Project

Online Retail Customer Segmentation

Individual Project

Uthaman A

Contant

- Introduction
- Problem Statement
- Data Summary
- Approach Overview
- Exploratory Data Analysis
- Exploring UK Market
- RFM Analysis
- K-Means Implementation
- Visualize the Clusters
- Conclusion

INTRODUCTION

For the past 10 years, we have witnessed a steady and strong increase of online retail sales. According to the Interactive Media in Retail Group (IMRG), online shoppers in the United Kingdom spent an estimated £50 billion in year 2011, a more than 5000 per cent increase compared with year 2000.[1] This remarkable increase of online sales indicates that the way consumers shop for and use financial services has fundamentally changed.

Compared with traditional shopping in retail stores, online shopping has some unique characteristics: each customer's shopping process and activities can be tracked instantaneously and accurately, each customer's order is usually associated with a delivery address and a billing address, and each customer has an online store account with essential contact and payment information. These desirable, special online shopping characteristics have enabled online retailers to treat each customer as an individual with personalized understanding of each customer and to build upon customer-centric business intelligence.

Problem Statement

Many small online retailers and new entrants to the online retail sector are keen to practice data mining and consumer-centric marketing in their businesses yet technically lack the necessary knowledge and expertise to do so. In this article a case study of using data mining techniques in customer-centric business intelligence for an online retailer is presented. The main purpose of this analysis is to help the business better understand its customers and therefore conduct customer-centric marketing more effectively.

Data Summary

- 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.

Approach Overview

I ranked each customer's value to the company based on three categories, how recent was their last purchase, how often do they transact with us, and how much have the spent on our products.

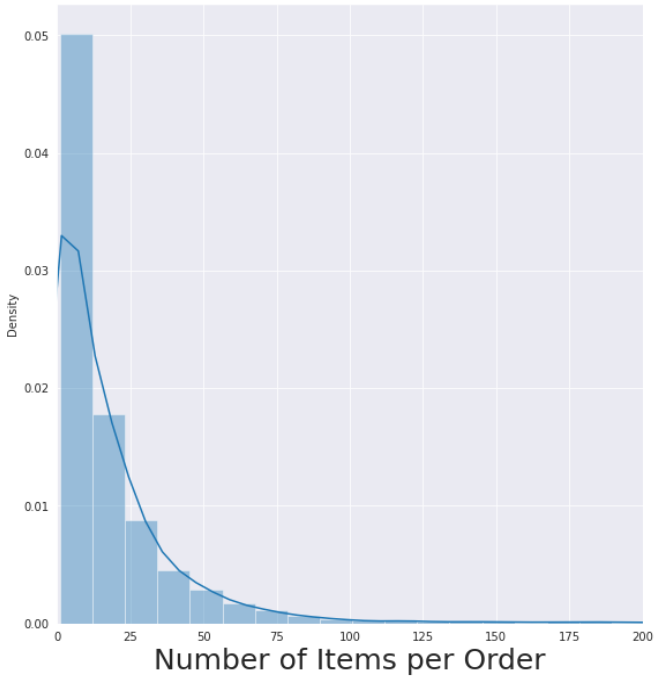
By ranking each of these categories we can group customers into 6 different segments that we can target with a different marketing and sales strategy.

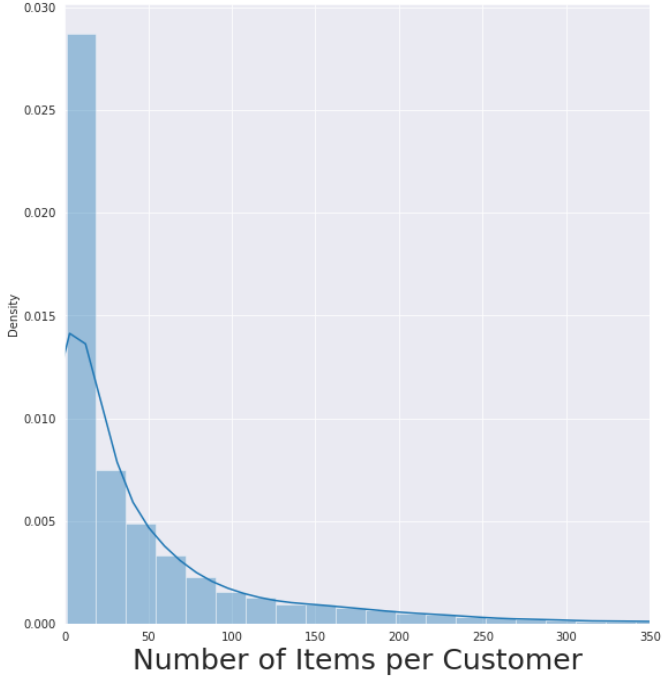
Then I built a k-means clustering model using the RFM variables to predict the classification of future customers.

Steps Involved

1. Data Cleaning
2. Exploartory Data Analysis
3. RFM Analysis
4. K-Means Clustering

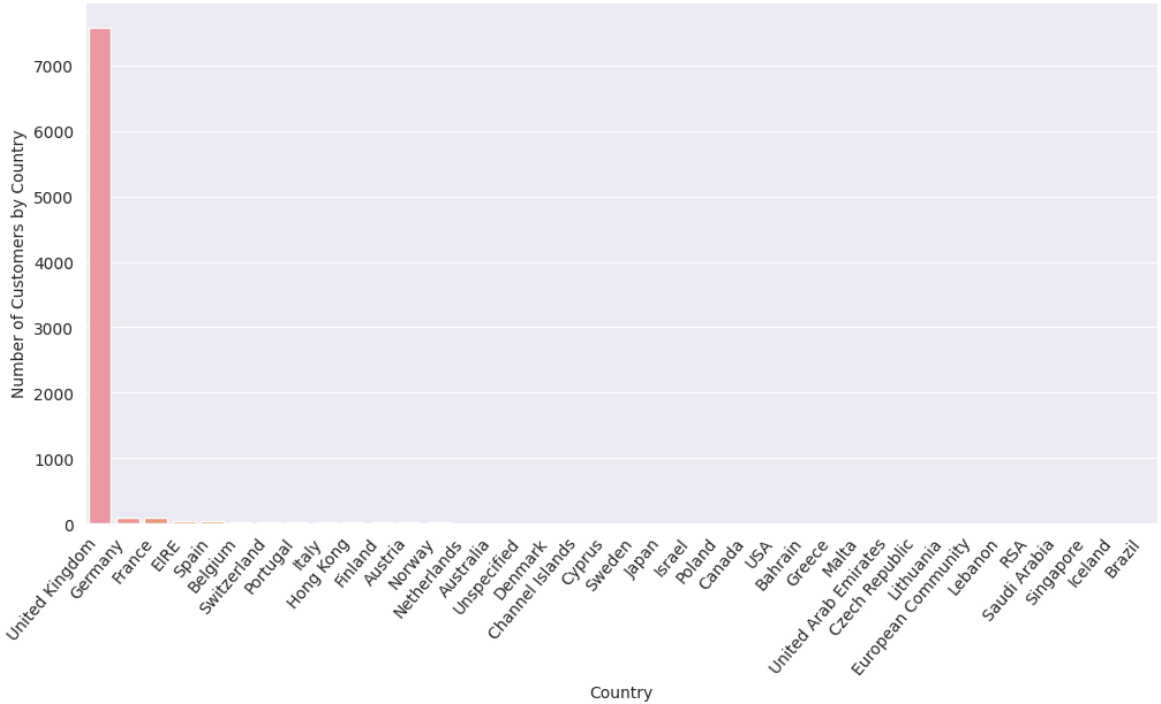
Visualize the Distribution





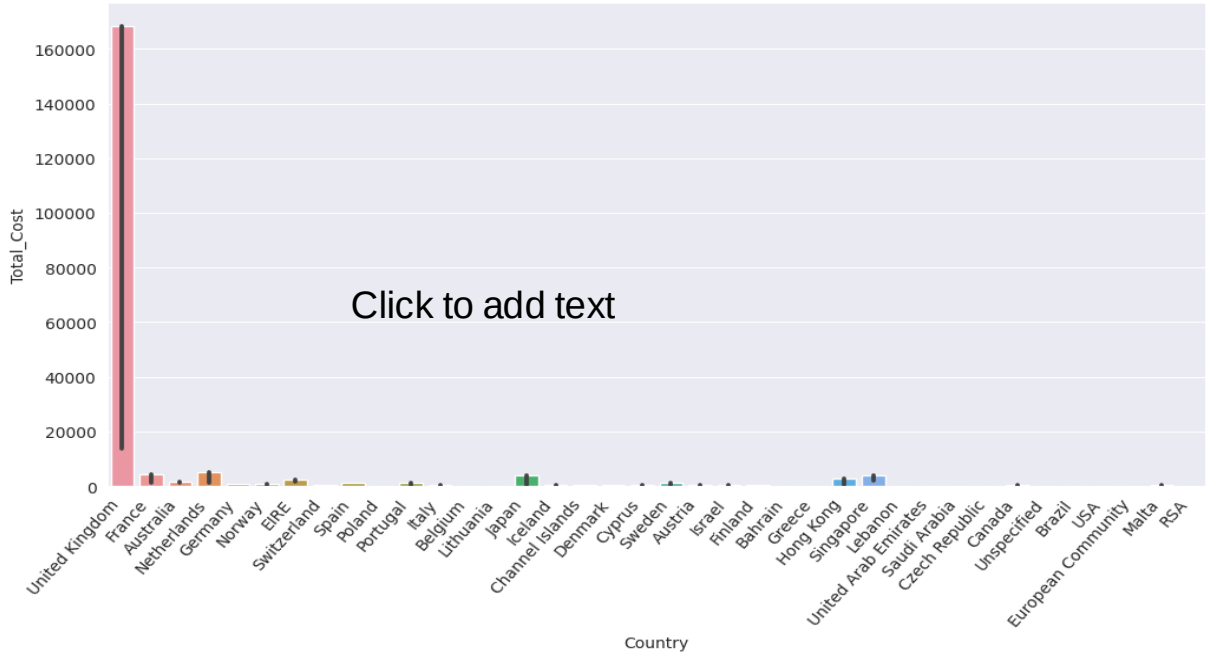
We have skewed left distributions for both plots. The average number of items per order is 20.5 and the average number of items per customer is 50.

Total Revenue per country



We have skewed left distributions for both plots. The average number of items per order is 20.5 and the average number of items per customer is 50.

Most Customers per country



The UK not only has the most sales revenue, but also the most customers. Since the majority of this data set contains orders from the UK, we can explore the UK market further by finding out what products the customers buy together and any other buying behaviors to improve our sales and targeting strategy.

Exploring UK Market

- Percentage of customers from the UK: 93.88 %
- Number of transactions: 23494
- Number of products Bought: 4065
- Number of customers: 7587

Most popular products that are bought in the UK

	StockCode	Description	Quantity
3154	84077	WORLD WAR 2 GLIDERS ASSTD DESIGNS	48326
4340	85099B	JUMBO BAG RED RETROSPOT	43167
1237	22197	POPCORN HOLDER	34365
3274	84879	ASSORTED COLOUR BIRD ORNAMENT	33679
4353	85123A	WHITE HANGING HEART T-LIGHT HOLDER	32901
1677	22616	PACK OF 12 LONDON TISSUES	25307
437	21212	PACK OF 72 RETROSPOT CAKE CASES	24702
1216	22178	VICTORIAN GLASS HANGING T-LIGHT	23242
41	17003	BROCADE RING PURSE	22801
11	15036	ASSORTED COLOURS SILK FAN	20322

RFM Analysis

- In the age of the internet and e-commerce, companies that do not expand their businesses online or utilize digital tools to reach their customers will run into issues like scalability and a lack of digital precsence.

- An important marketing strategy e-commerce businesses use for analyzing and predicting customer value is customer segmentation.

- Customer data is used to sort customers into group based on their behaviours and preferences:
    1. Recognize who are our most valuable customers
    2. Increase revenue
    3. Increase customer retention
    4. Learn more about the trends and behaviours of our customers
    5. Define customers that are at risk

Customer Segmentation with RFM Model

RFM Scoring

- The simplest way to create customer segments from an RFM model is by using Quartiles.
- We will assign a score from 1 to 4 to each category (Recency, Frequency, and Monetary) with 4 being the highest/best value.
- The final RFM score is calculated by combining the individual RFM values.
- **Note:** Data can be assigned into more groups for better granularity, but we will use 4 in this case

Choosing a Predictive Model

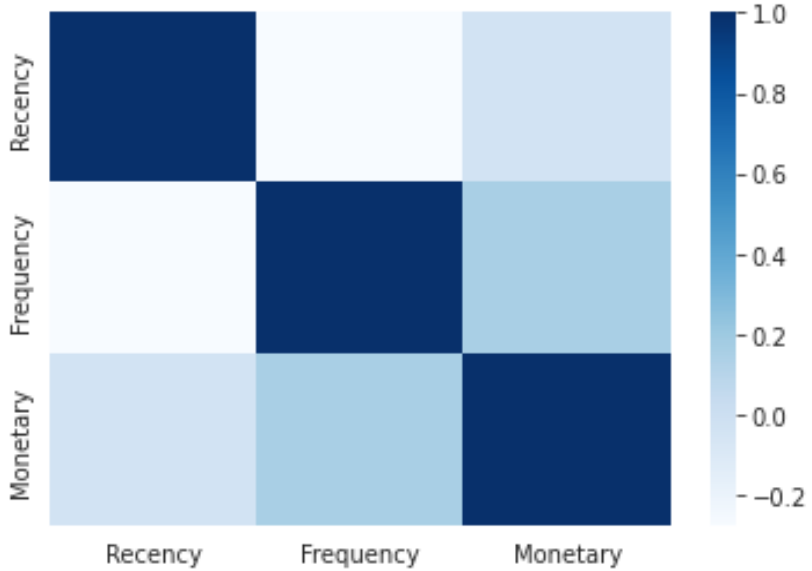
- Now that we have our customers segmented into 6 different categories, we can gain further insight into customer behaviour by using predictive models in conjunction with out RFM model.

- Possible algorithms include Logistic Regression, K-means Clustering, and K-nearest Neighbor.

- We will go with K-means since we already have our distinct groups determined.

- K-means has also been widely used for market segmentation and has the advantage of being simple to implement.

Feature Correlations

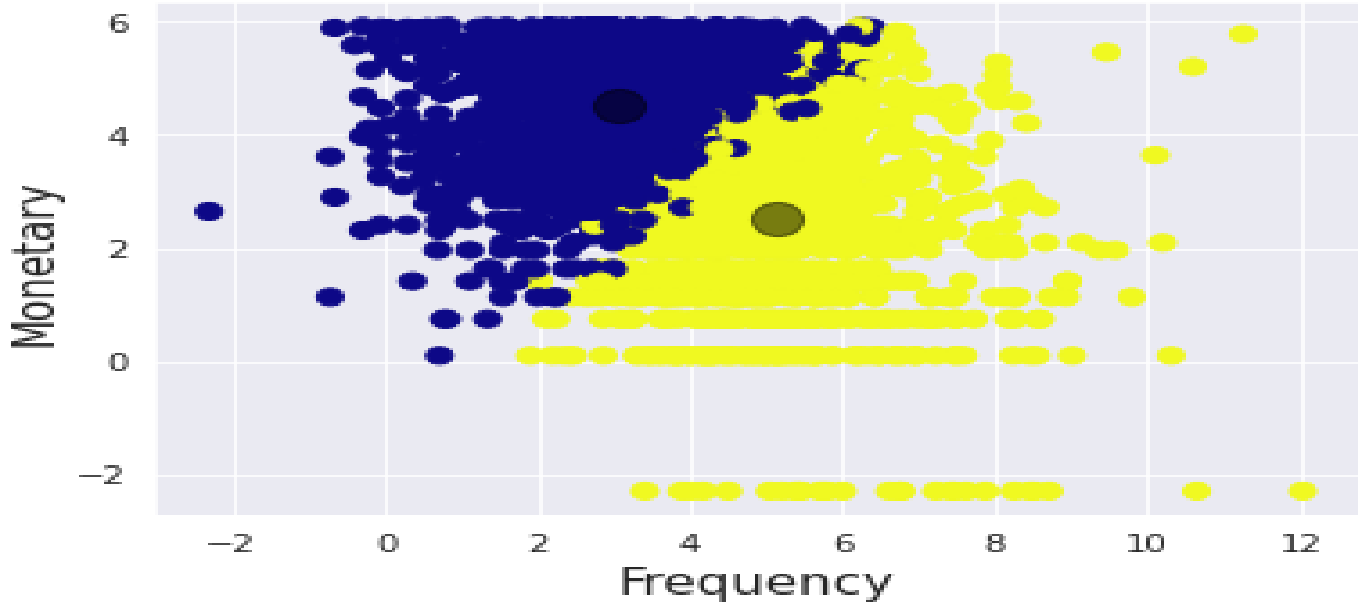


Looking at this heatmap, we see that there is a negative correlation between Recency : Frequency and Recency : Monetary, but there is a positive correlation between Frequency : Monetary

K-Means Implementation

- For k-means, you have to set k to the number of clusters you want, but figuring out how many clusters is not obvious from the beginning.
- We will try different cluster numbers and check their silhouette coefficient.
- The silhouette coefficient for a data point measures how similar it is to its assigned cluster from -1 (dissimilar) to 1 (similar).
- **Note:** K-means is sensitive to initializations because they are critical to qualify of optima found. Thus, we will use smart initialization called k-means++

Visualize the Clusters



The yellow cluster has a centroid at around (0.5, 3) and represents the "low value customers". The dark blue cluster has a centroid at around (1.8, 5) and represents the "high value customers".

Conclusion

Although we didn't obtain two clearly separated clusters, we were able to build a model that can classify new customers into "low value" and "high value" groups. Generally, if a customer only transacted with us a few times, they needed to be at least in the top 50th percentile in monetary spending to be considered a "high value customer". The clusters assignments are muddled, which may be due to outliers that weren't removed.

Limitations of k-means clustering:

- There is no assurance that it will lead to the global best solution.
- Can't deal with different shapes(not circular) and consider one point's probability of belonging to more than one cluster.

These disadvantages of k-means show that for many datasets (especially low-dimensional datasets), it may not perform as well as you might hope.

Thank You