Capstone Project

Customer Segmentation (Unsupervised ML)

Individual Project
Tanmaya Kumar Pattanaik

Introduction

- Customer Segmentation is the process of dividing customers into groups based on common characteristics so companies can market to each group effectively and appropriately.

- A primary goal for any company and business is to understand their targeted customers.

Online Retail Data

- We will be using Online retail dataset to explore customer segmentation through the interesting task of unsupervised learning method.
- The Online Retail a transactional data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.
- In this project, our task is to identify major customer segments on a transactional data of a online retail store.

Column Information

- **InvoiceNo**: Invoice number, 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. a 5-digit integral number uniquely assigned to each distinct product.
- **Description**: Product(items) name.
- **Quantity**: The quantities of each product (item) per transaction.
- **InvoiceDate**: Invoice Date and time. The day and time when each transaction was generated
- **UnitPrice**: Product price per unit in sterlin.
- **CustomerID**: Customer number, a 5 digit integral number uniquely assigned to each customer.
- **Country**: Country name, the name of the country where each customer resides.

Required Packages

- Numpy
- Pandas
- Matplotlib
- Seaborn
- Datetime
- Sklearn Packages:

1. Cluster
2. K-Means
3. Gaussian Mixture
4. Silhouette Samples
5. Silhouette Score

Data exploration

- There are 541909 rows and 8 columns in our dataset.
- Number of transactions: 25900
- Number of products bought: 4070
- Number of customers: 4372
- Number of countries: 38

Null value treatment

- We have missing values in the CustomerID and Description columns.
- Since 25% of the customer ID's are missing, we will create and fill a new column that has a 1 when customer ID is null and a 0 when it is not.
- Since we won't be doing analysis on the descriptions of the orders, we can leave the null values as it is
- Since the customer ID's are missing, we assumed these orders were not made by the customers already in the data set because those customers already have ID's. We also don't want to assign these orders to those customers because this would alter the insights we draw from the data. Instead of dropping the null CustomerID values, we assigned those rows a unique customer ID per order. This will act as a new customer for each unique order.
- Using the values in the InvoiceNo column would be the most straightforward approach. We created a new customer ID column called NewID with the invoice numbers filling in for the missing values. Then we add the number of unique orders in df1 and to number of unique values in CustomerID and see if it equals the number of unique values in NewID. This will check if any of the new values match the existing values in the column and make sure we didn't add more orders to an existing customer.

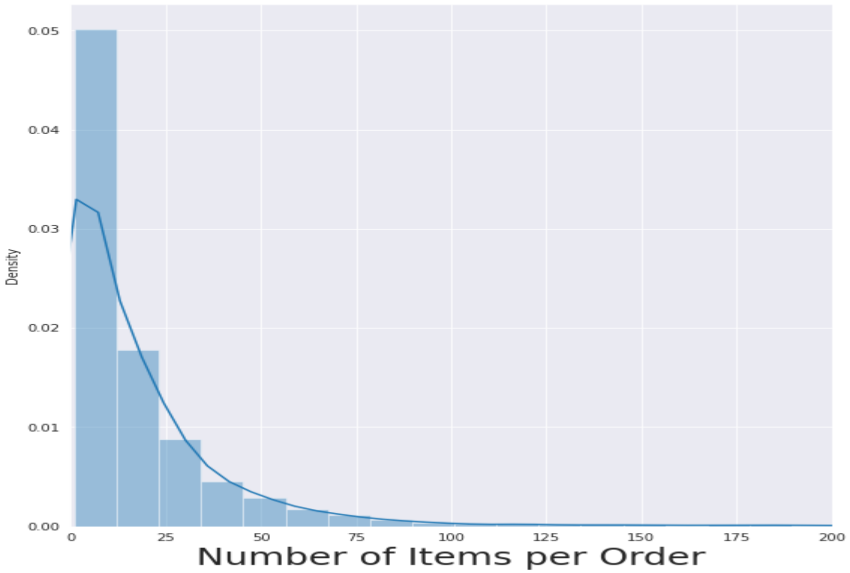
Exploratory data analysis

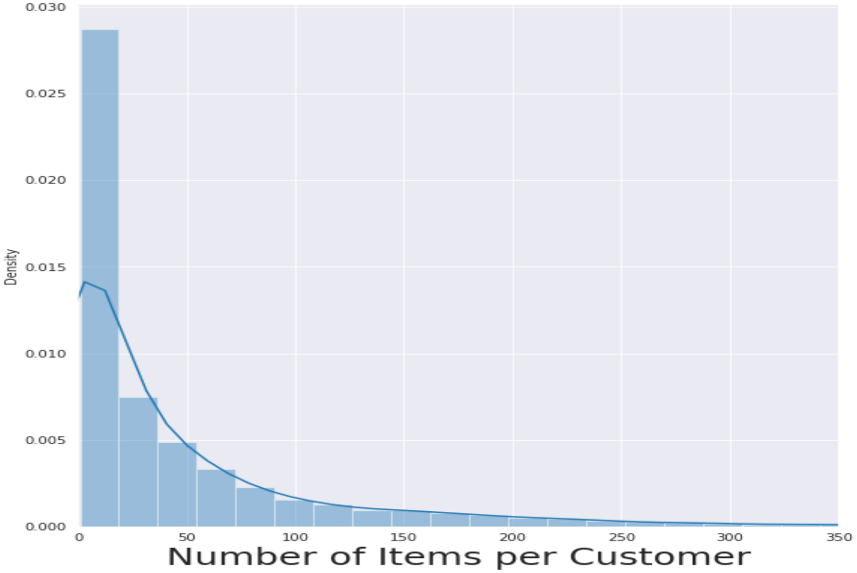
Let's deep more into the data and understand the dataset by exploring all columns one by one with the help of visualization. It will help us in understanding and building models.

Canceled orders

- When we check the summary of data, there are negative values in the Quantity and Unitprice columns. We assumed these are orders that were cancelled and items that were returned.
- Since nothing came back when we filtered the cancelled orders by Quantity > 0, this confirms that the negative values mean the order was cancelled.
- 9288 or about 36% of the orders were cancelled. Looking deeper into why these orders were cancelled may prevent future cancellations.
- The description of "Adjust bad debt" tells us that this is an adjustment for a customer with insufficient funds or an allowance for a customer who never paid for the order.

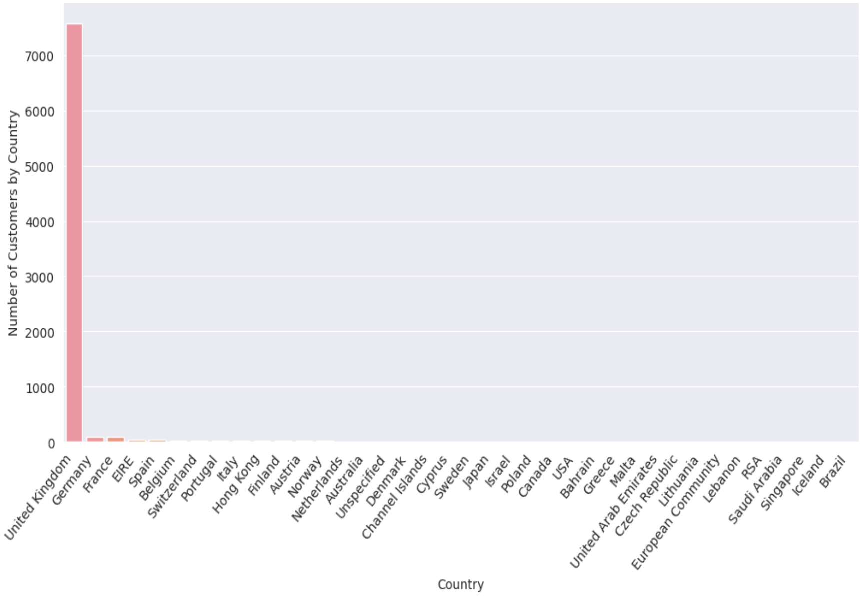
Exploring the Oders

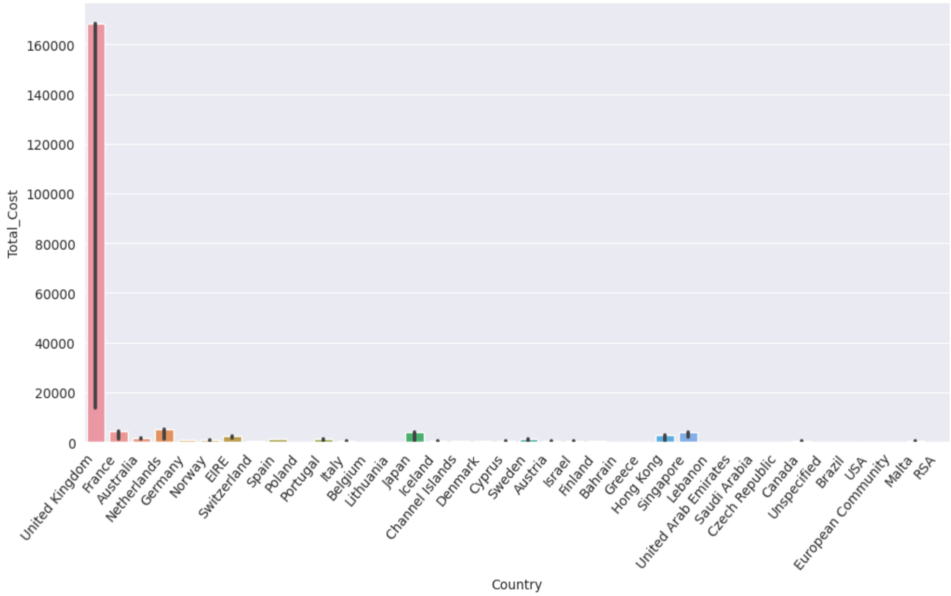




- We have skewed right distributions for both plots.
- The average number of items per order is 20.5
- The average number of items per customer is 50.

Customers by Country





- The United Kingdom has significantly more customers than the other countries in our data set.
- The UK not only has the most sales revenue, but also the most customers.

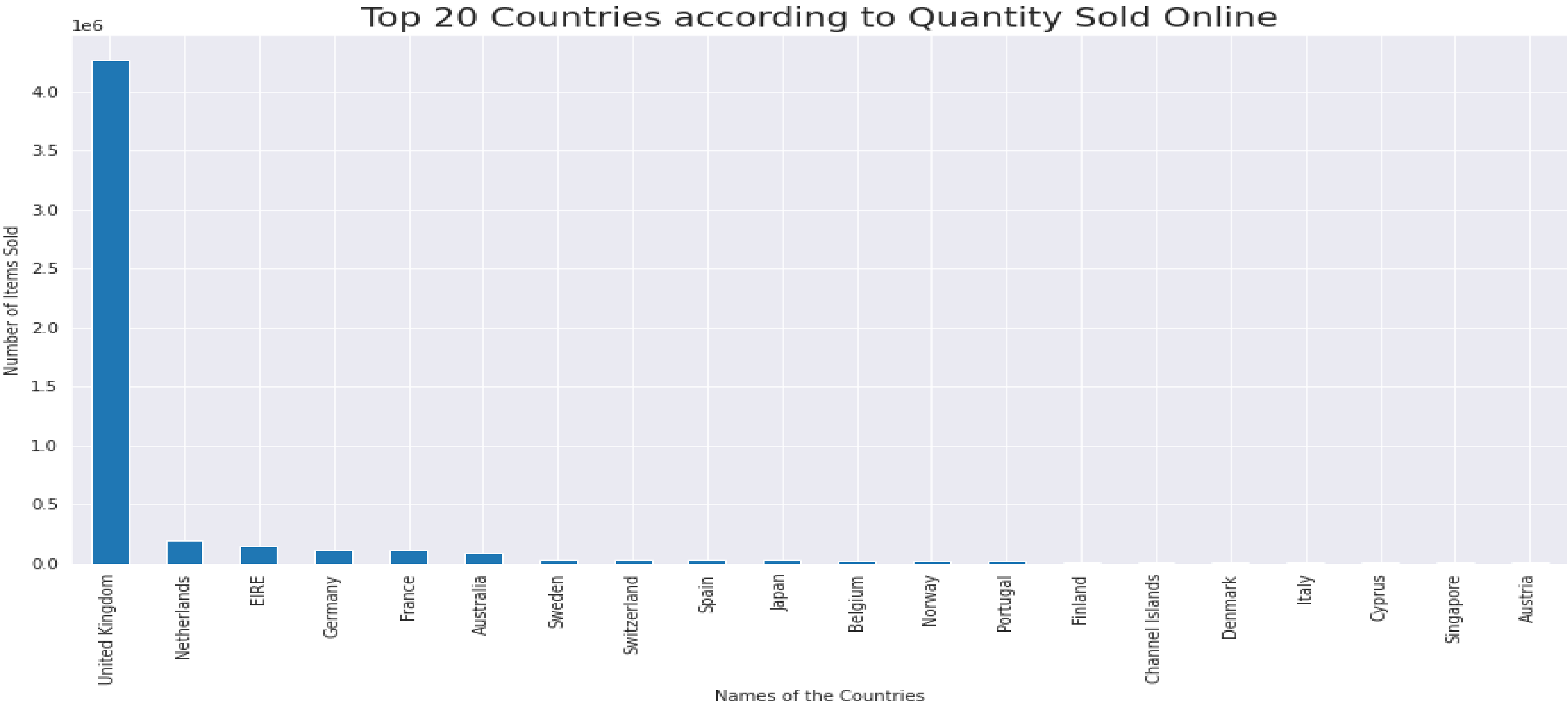
Exploring the UK Market

- 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

Description	Quantity
WORLD WAR 2 GLIDERS ASSTD DESIGNS	48326
JUMBO BAG RED RETROSPOT	43167
POPCORN HOLDER	34365
ASSORTED COLOUR BIRD ORNAMENT	33679
WHITE HANGING HEART T-LIGHT HOLDER	32901
PACK OF 12 LONDON TISSUES	25307
PACK OF 72 RETROSPOT CAKE CASES	24702
VICTORIAN GLASS HANGING T-LIGHT	23242
BROCADE RING PURSE	22801
ASSORTED COLOURS SILK FAN	20322

Top 20 countries according to Quantity sold online



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 presence.

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 behaviors and preferences.

RFM

- Recency (Days since last purchase): To calculate recency, we need to choose a date as a point of reference to evaluate how many days ago was the customer's last purchase.
- Frequency (Number of purchases): To calculate how many times a customer purchased something; we need to count how many invoices each customer has. Customer ID 12748 has 210 frequency.
- Monetary (Total amount of money spent): The monetary value is calculated by adding together the cost of the customers' purchases. The customer ID 12346 has spent 77183.60 and 12748 has spent 3841.31.

Customer Segmentation with RFM Model

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.

RFM Model

CustomerID	Recency	Frequency	Monetary	R_Quartile	F_Quartile	M_Quartile	RFM_Score
12346	325	1	77183.6	1	1	4	114
12747	2	11	689.49	4	4	4	444
12748	0	210	3841.31	4	4	4	444
12749	3	5	98.35	4	3	3	433
12820	3	4	58.2	4	3	3	433

- High recency (more days since last purchase) is bad, while high frequency and monetary value is good.
- A score of 4 represents the customer being in the 75th percentile for that category.

How many customers do we have in each segment?

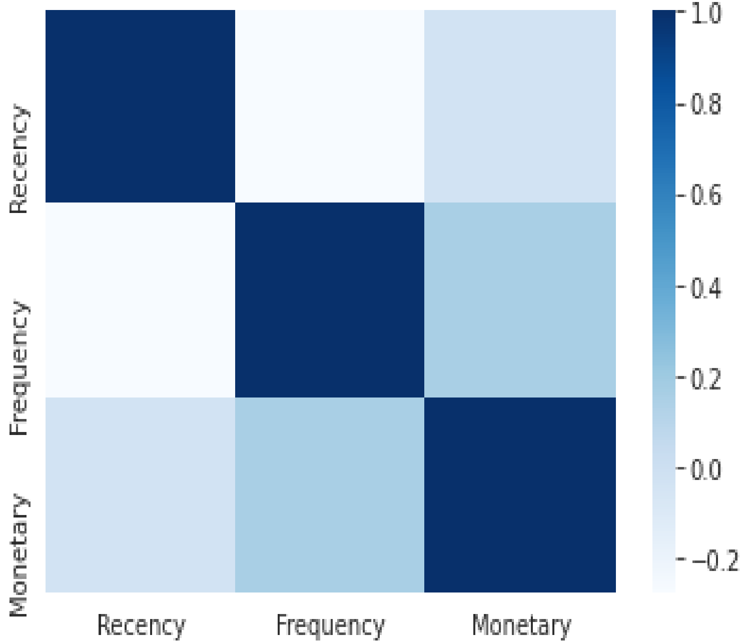
- Best customers: 370
- Loyal customers: 791
- Big spenders: 980
- Almost lost: 65
- Lost customers: 11
- Lost cheap customers: 377

➢ We could reward our Best Customers and Loyal Customers or create a "Refer a Friend" promotional offer targeted for them.

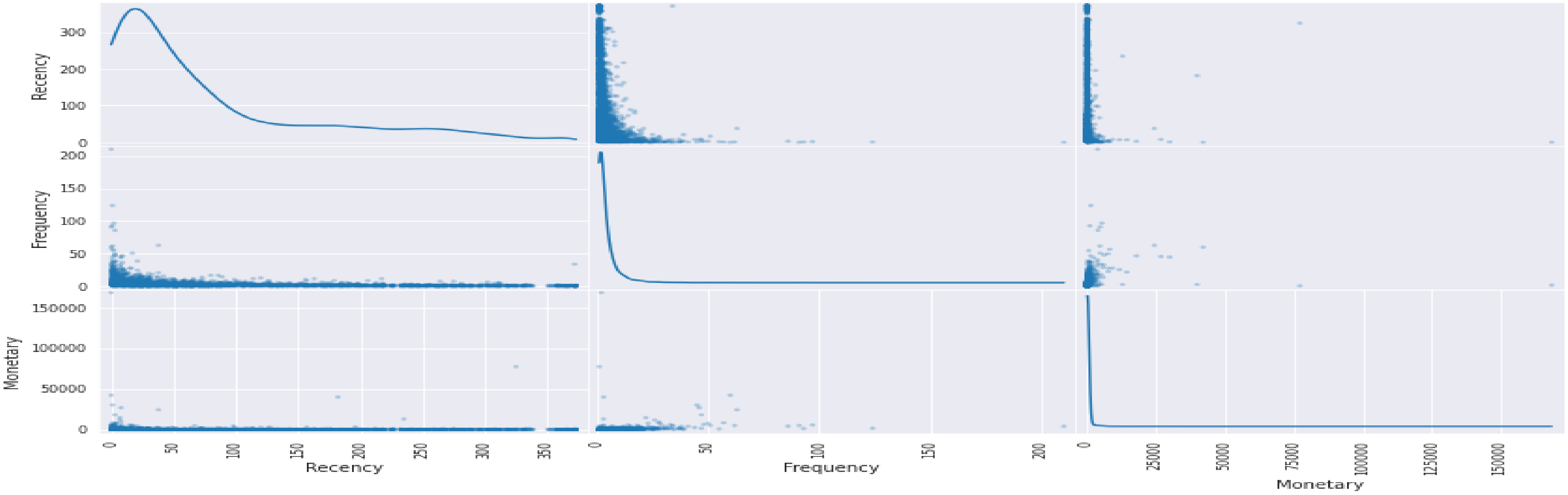
➢ For the Almost Lost customers, we could aggressively market towards them with great deals so we don't lose them forever.

Correlation Analysis

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



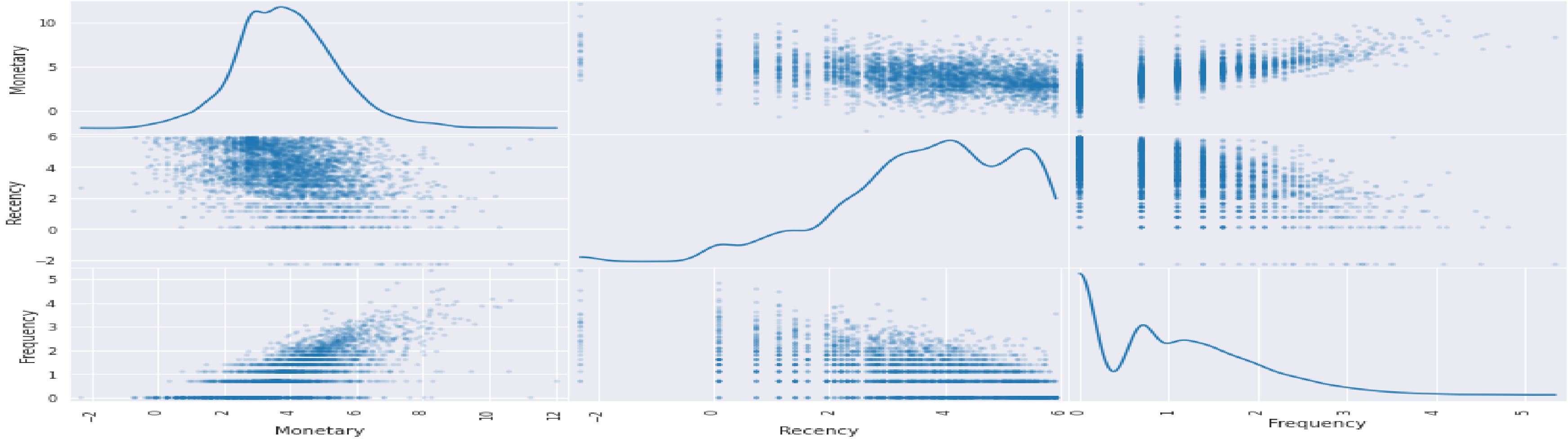
Visualizing Feature Distributions



There is a skewed distribution for the 3 variables and there are outliers.

Since clustering algorithms require a normal distribution, normalization of the data is required.

Data Normalization

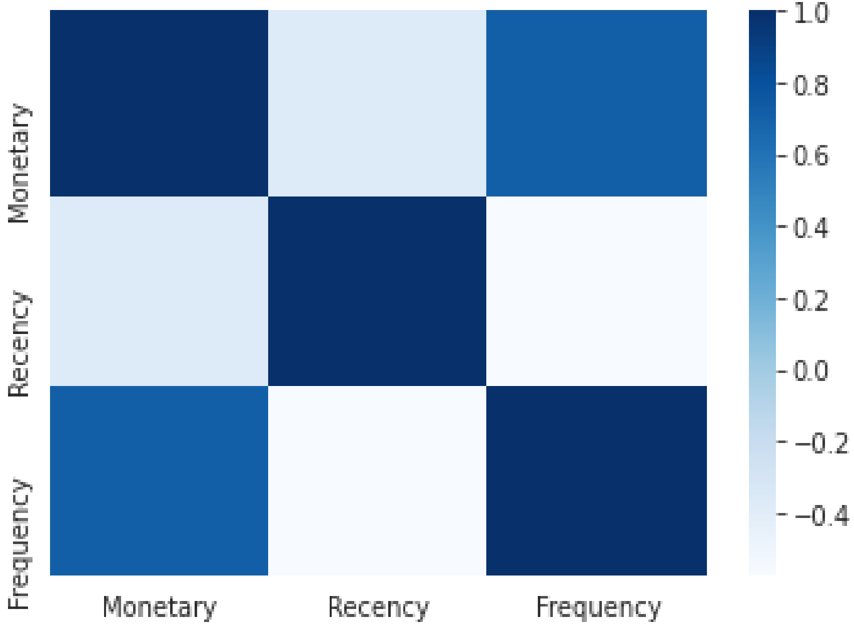


The distributions of Monetary and Recency are more normailized, but recency is skewed to the left.

Frequency was also skewed right because of a lot of customers only buying from us once.

Correlation

Now, Monetary and Frequency are more strongly correlated, so we will use those two variables in our K-Means model.



K-Means Implementation

For k-means, we 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).

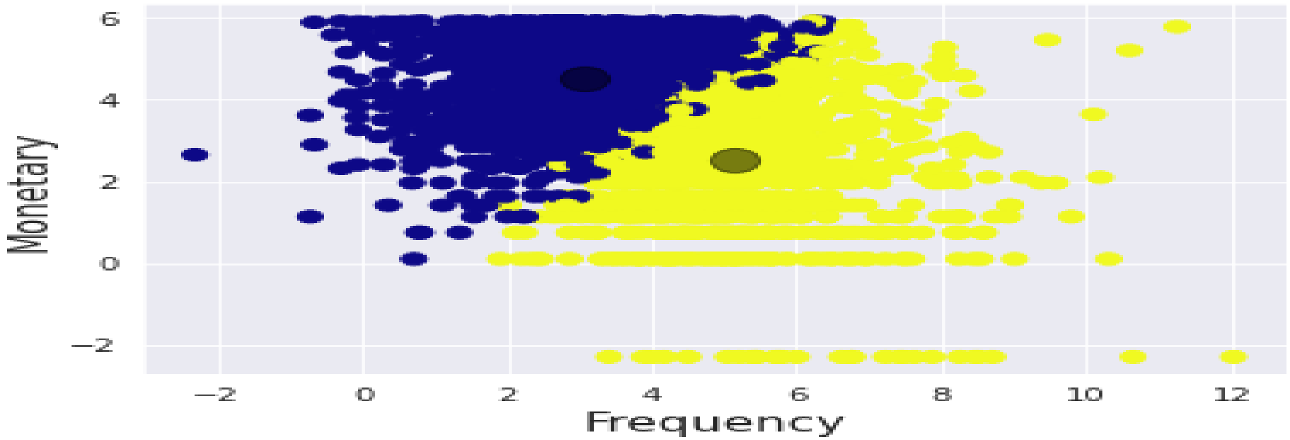
K-means is sensitive to initializations because they are critical to qualifty of optima found. Thus, we will use smart initialization called k-means++

Silhouette score

n_clusters	Average Silhouette score
2	0.38
3	0.30
4	0.31
5	0.29
6	0.29
7	0.29
8	0.28
9	0.29

The best silhouette score obtained is when there are **2 clusters.**

Visualize the Clusters



The yellow cluster has a centroid at around (5, 2.5) and represents the "low value customers".

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

Conclusion

- In this project we covered various aspects of the Machine learning development cycle. We observed that the data exploration and variable analysis is a very important aspect of the whole cycle and should be done for thorough understanding of the data.
- 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“.
- We got cluster 2 gives the best silhouette score.

Thank You