

# SCIENCE



# EDUCATION

**Capstone 3 Project**

**Customer Segmentation for An Online Retail**

**Zhaoqi Fan**

## Table of Content

|   |    |
|---|----|
| 1. Introduction .....                           | 1  |
| 1.1 Problem statement.....                      | 1  |
| 1.2 Scope of solution space.....                | 1  |
| 1.3 Constraints .....                           | 1  |
| 1.4 Project plan.....                           | 1  |
| 1.5 Data Source .....                           | 2  |
| 2. Data wrangling.....                          | 3  |
| 2.1 Data Integrity .....                        | 3  |
| 2.2 Feature Analysis .....                      | 4  |
| 3. Exploratory Data Analysis.....               | 7  |
| 3.1 Price vs quantity .....                     | 7  |
| 3.2 Revenue Analysis .....                      | 7  |
| 3.3 Customer Analysis .....                     | 9  |
| 4. Clustering.....                              | 11 |
| 4.1 Recency-Frequency-Monetary (RFM) value..... | 11 |
| 4.1.1 Recency calculation and clustering .....  | 11 |
| 4.1.2 Frequency calculation and clustering..... | 12 |
| 4.1.3 Monetary calculation and clustering.....  | 12 |
| 4.1.4 RFM calculation and clustering .....      | 13 |
| 4.2 RFM + Price + Quantity.....                 | 15 |
| 4.1.1 Data scaling.....                         | 15 |
| 4.1.2 Silhouette analysis.....                  | 15 |
| 5. Conclusions.....                             | 19 |
| 6. Recommendations & Future Work.....           | 19 |

## 1. Introduction

The transnational data has been collected from 2009-12-01 to 2011-12-09 for a UK-based and registered non-store online retail. The 1044848 records of customer ID, price, quantity, invoice date, description, and country are included in the data set. The manager is interested in targeting customers for developing business development strategies. This project will use various features, e.g., Recency, Frequency, and Monetary Value (RFM), for customer segmentation and explore the difference among customer groups.

### 1.1 Problem statement

How many groups we can segment the customers by using the given features?

### 1.2 Scope of solution space

- This project will focus on the price, quantity, and invoice date to explore measures for improving the RFM values
- The categories of merchandise will be a good supplement for optimizing the RFM values.

### 1.3 Constraints

- The data is mainly from the UK. The scarce data from other countries may bring noise into the analysis.

### 1.4 Project plan

This project is implemented by following the procedure below.

- a. Problem Identification
- b. Data Wrangling
  - Check data integrity
  - Clean data: missing values, duplicates, etc.
- c. Exploratory Data Analysis
  - Identify important features
  - Statistical property analysis
  - Visualize data
  - Identify trends and patterns
- d. Pre-Processing and Clustering
  - Recency (R) calculation and clustering
  - Frequency (F) calculation and clustering

- Monetary (M) calculation and clustering
- RFM value calculation and clustering
- RFM + Price + Quantity clustering
- e. Documentation  
A project report, a slide deck, and Jupyter notebooks will be generated and shared in a GitHub repo.

<https://github.com/z-fan-git/Customer-Segmentation.git>

### 1.5 Data Source

A transnational data set that 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. The data can be found on [Kaggle](#) and [UCI Machine Learning Repository](#).

There are 8 features listed below.

- Invoice No: 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: 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.

## 2. Data wrangling

The objectives of this section are to explore:

- Identify any issues that will require data cleaning
- Analyze features in the dataset

### 2.1 Data Integrity

The dataset summary listed in Table 2.1 shows features, non-null values, and data types.

Table 2.1 Dataset summary

| Column      | Non-Null | Count    | Type    |
|-------------|----------|----------|---------|
| Customer ID | 809561   | non-null | float64 |
| Invoice     | 1044848  | non-null | object  |
| InvoiceDate | 1044848  | non-null | object  |
| Price       | 1044848  | non-null | float64 |
| Quantity    | 1044848  | non-null | int64   |
| StockCode   | 1044848  | non-null | object  |
| Description | 1040573  | non-null | object  |
| Country     | 1044848  | non-null | object  |

We can find that:

- This dataset contains 1,044,848 records and 8 features.
- There are missing values in `Customer ID` and `Description`.
- 3 numerical features: Customer ID, Price, and Quantity
- 5 categorical features.

#### Missing data:

There are 235287 missing values for Customer ID and 4275 missing values for Description. The missing data in customer ID contribute nothing to the objective of this project, i.e., customer segmentation. Therefore, all these missing values are dropped. As a result, there are 809491 rows left in the dataset.

#### Duplicates:

There are also 11676 duplicated rows as shown in Table 2.2. Dropping all the duplicates yields a dataset with 797815 rows.

Table 2.2 Sample of duplicated records

|        | Customer ID | Invoice | InvoiceDate         | Price | Quantity | StockCode | Description        | Country        |
|--------|-------------|---------|---------------------|-------|----------|-----------|--------------------|----------------|
| 255587 | 12346       | C514024 | 2010-06-30 11:22:00 | 12.94 | -1       | M         | manual             | United Kingdom |
| 255589 | 12346       | C514024 | 2010-06-30 11:22:00 | 12.94 | -1       | M         | manual             | United Kingdom |
| 485600 | 12356       | 534804  | 2010-11-24 12:24:00 | 1.95  | 1        | 22629     | spaceboy lunch box | Portugal       |
| 485615 | 12356       | 534804  | 2010-11-24 12:24:00 | 1.95  | 1        | 22629     | spaceboy lunch box | Portugal       |

## 2.2 Feature Analysis

### Categorical features

- The stock code and description are the information for goods. We will ignore these two columns for now.
- The distribution of customers worldwide is shown in Figure 2.1. Most of the customers (91%) were from the United Kingdom. There were 1.8% of customers from Germany and 1.6% of customers from France.



Figure 2.1 Customer distribution in countries

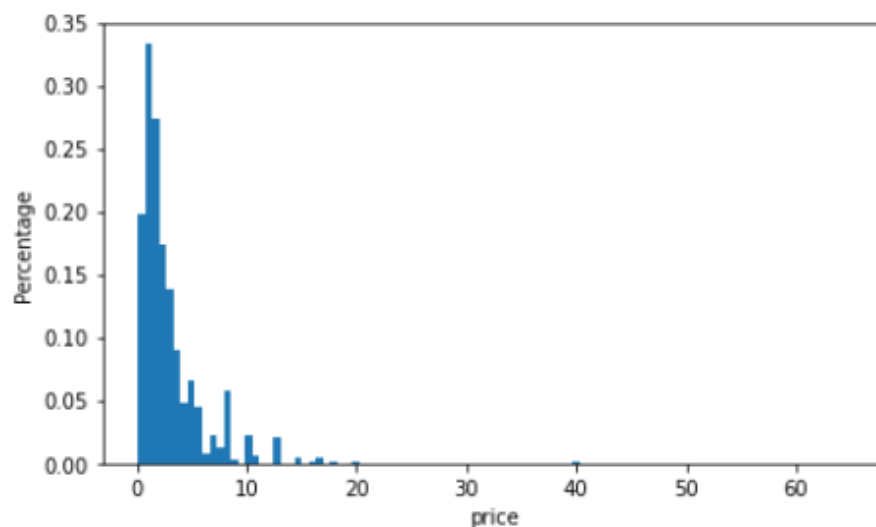
Numerical features

The statistics of numerical features is listed in Table 2.3. Both the price and quantity distribution are skewed. By removing the 0.1% and 99.9 percentile as the threshold, we removed the outliers and showed the distribution of price and quantity in Figure 2.2.

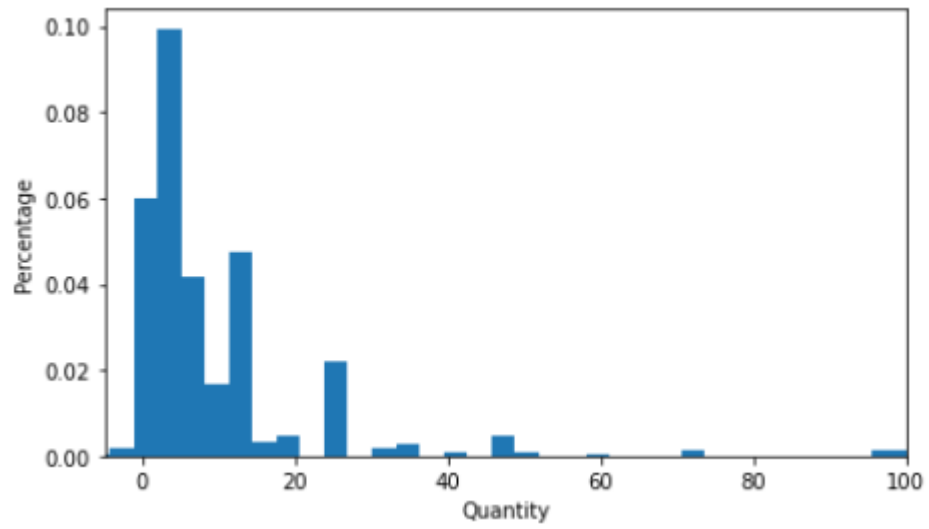
- The distribution is extremely skewed.
- It suggests a transformation in modeling.

Table 2.2 Statistics of numerical features

|              | Price         | Quantity      |
|--------------|---------------|---------------|
| <b>count</b> | 1.044848e+06  | 1.044848e+06  |
| <b>mean</b>  | 4.590546e+00  | 9.993649e+00  |
| <b>std</b>   | 1.217042e+02  | 1.742185e+02  |
| <b>min</b>   | -5.359436e+04 | -8.099500e+04 |
| <b>0.1%</b>  | 0.000000e+00  | -1.141530e+02 |
| <b>25%</b>   | 1.250000e+00  | 1.000000e+00  |
| <b>50%</b>   | 2.100000e+00  | 3.000000e+00  |
| <b>75%</b>   | 4.130000e+00  | 1.000000e+01  |
| <b>99.9%</b> | 2.145879e+02  | 5.000000e+02  |
| <b>max</b>   | 3.897000e+04  | 8.099500e+04  |



(a)



(b)

Figure 2.1 Price (a) and quantity (b) distribution



### 3. Exploratory Data Analysis

The objective of EDA is to explore the characteristics of features and the use of features to derive valuable insights like RFM value.

#### 3.1 Price vs quantity

There is no clear correlation between price and quantity as shown in Figure 3.1.

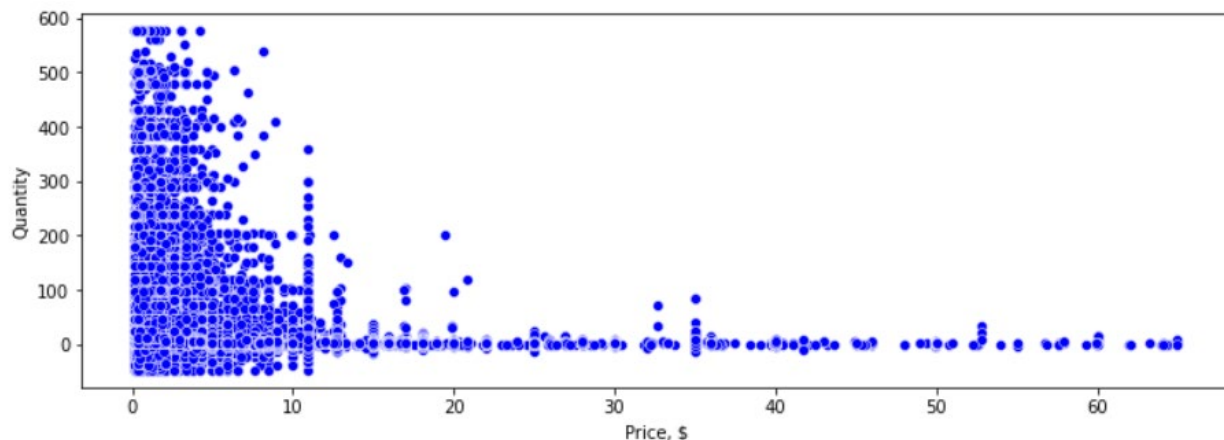


Figure 3.1 Price vs quantity

#### 3.2 Revenue Analysis

The revenue is calculated by using the equation below.

$$\text{Revenue} = \text{Active Customer Count} * \text{Order Count} * \text{Average Revenue per Order}$$

As there is no cost reported, the revenue can also be treated as the lifetime value for each customer as indicated by the equation below. Note that the time window is determined by the business. It can be 1 month, 3 months, 6 months, 1 year, etc.

$$\text{Lifetime Value} = (\text{Total Gross Revenue} - \text{Total Cost}) \text{ within a time window}$$

Based on the invoice date, we can estimate:

- Monthly revenue (Figure 3.2). A strong seasonality exhibits in the monthly revenue, which shows two peaks around November. In addition, the monthly revenue is quite stable in the first half year and progressively increases after August.
- Monthly revenue growth rate (Figure 3.3). The growth rate varies significantly in quarters. The trend is in line with the monthly revenue.
- Monthly revenue per customer (Figure 3.4). The customers spend more money in September - November and much less in April - June.

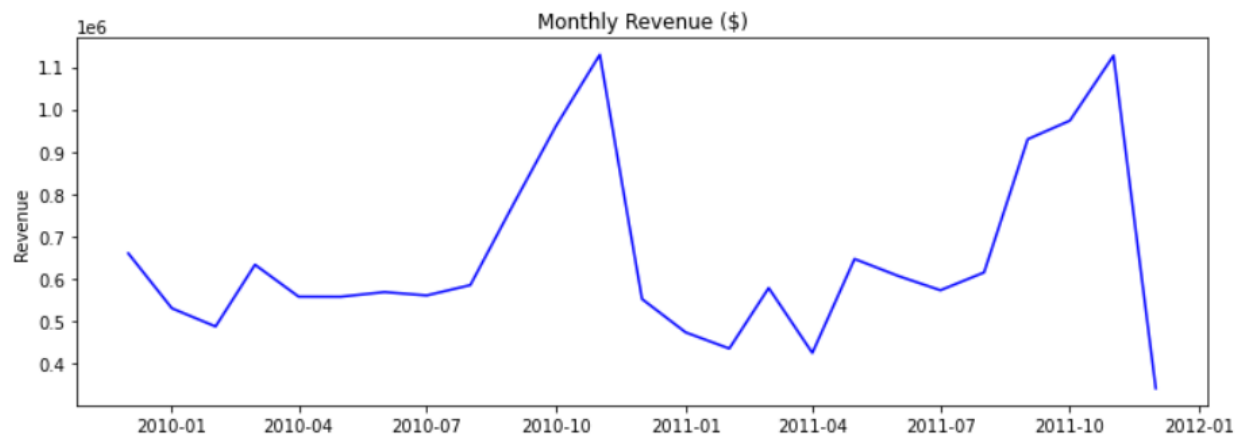


Figure 3.2 Monthly revenue variation

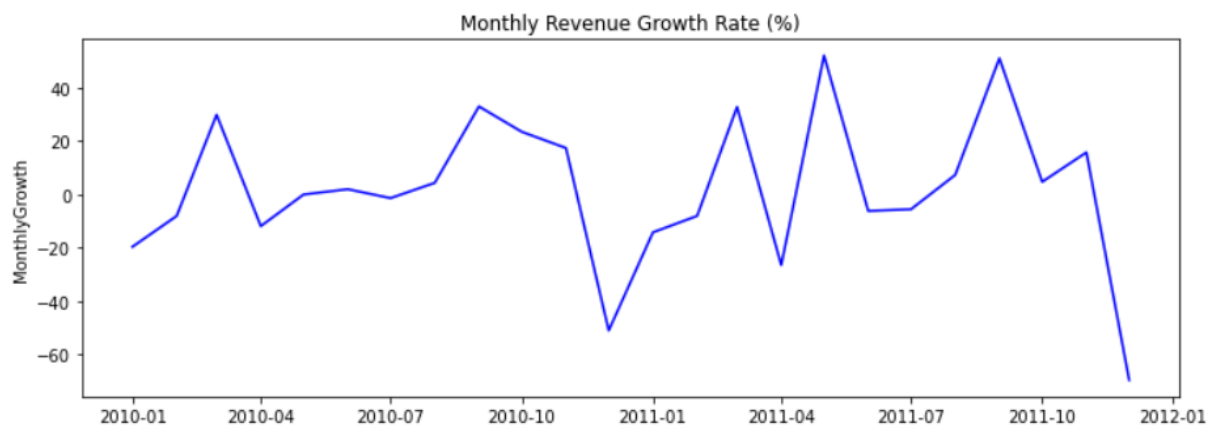


Figure 3.3 Monthly revenue growth rate variation

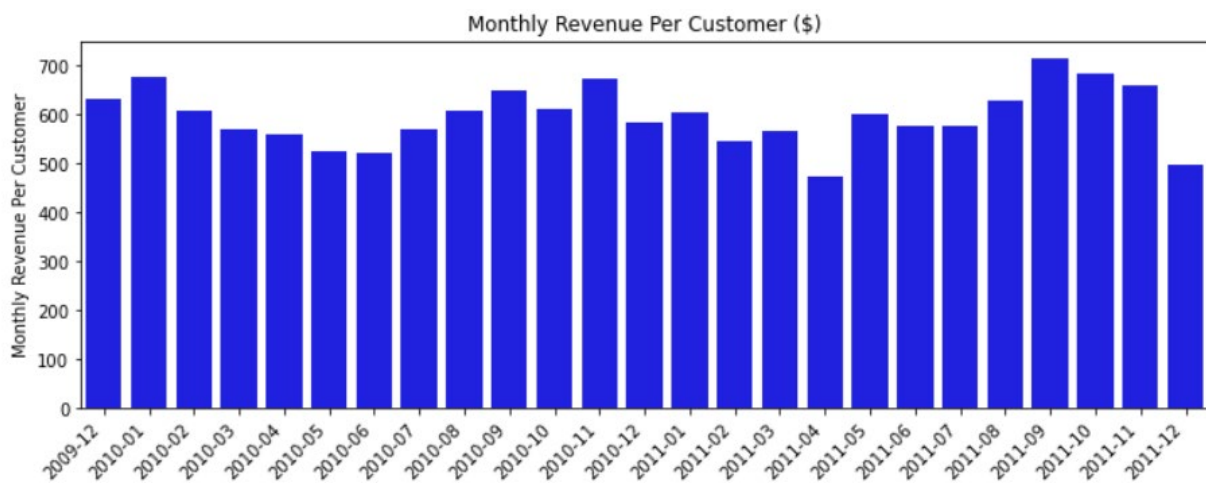


Figure 3.4 Monthly revenue per customer variation

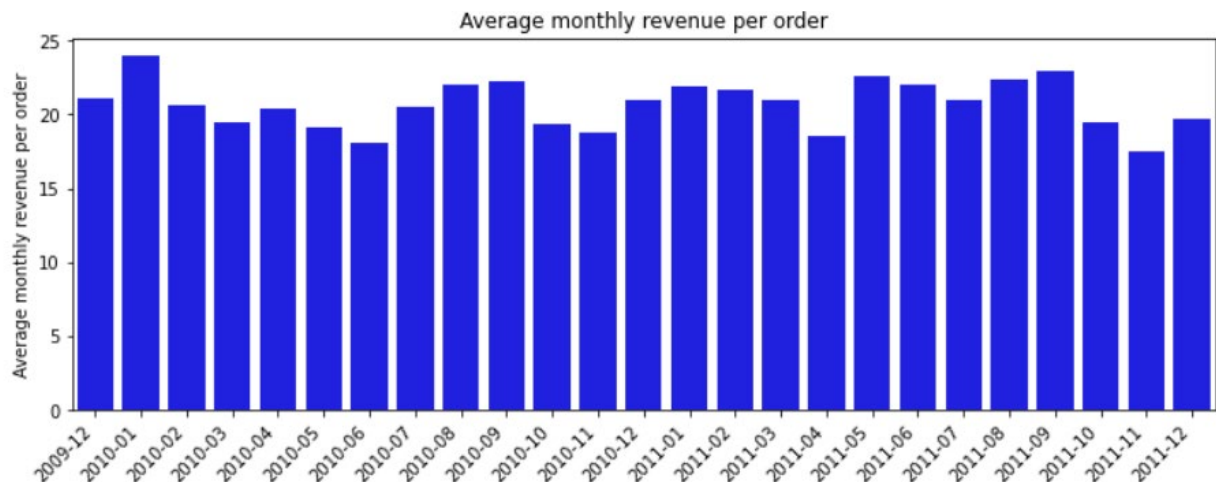


Figure 3.5 Monthly revenue per order variation

- Monthly revenue per order (Figure 3.5). The trend is slightly different. The revenue per order doesn't increase significantly in October and November, which implies customers buy more goods instead of buying more expensive ones.

### 3.3 Customer Analysis

In this section, the new and existing customers are identified by using the invoice date. Figure 3.6 shows that existing customers contribute way more than new customers.

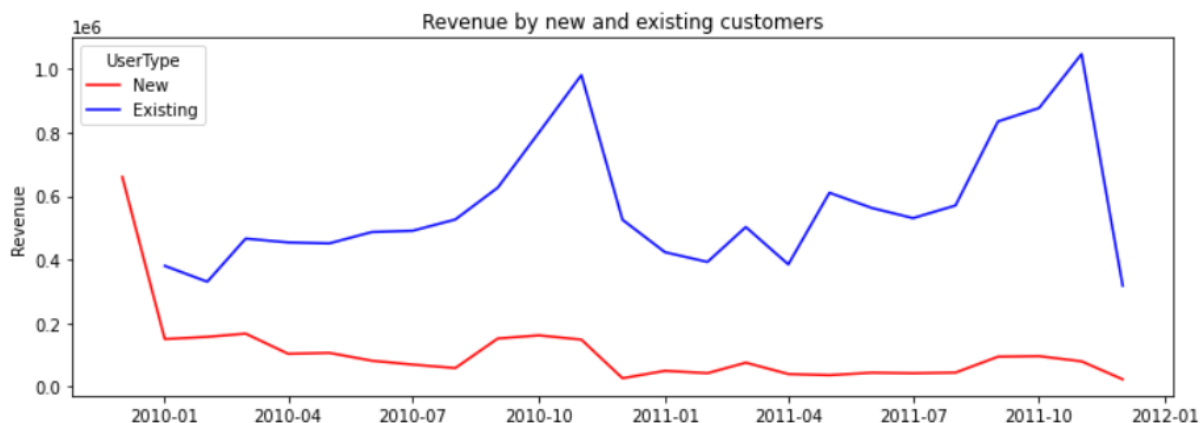


Figure 3.6 Revenue by customers

The ratio of new and existing customers each month is a good indicator of the growth of a business. Figure 3.7 shows that more new customers are coming to the store in February – March and September – November each year. We can convert this ratio into the retention rate shown in Figure 3.8. The retention rate is 0.4 on average and it also

shows seasonality to a certain extent. The retention rate is higher in October-November, which is in line with the revenue analysis.

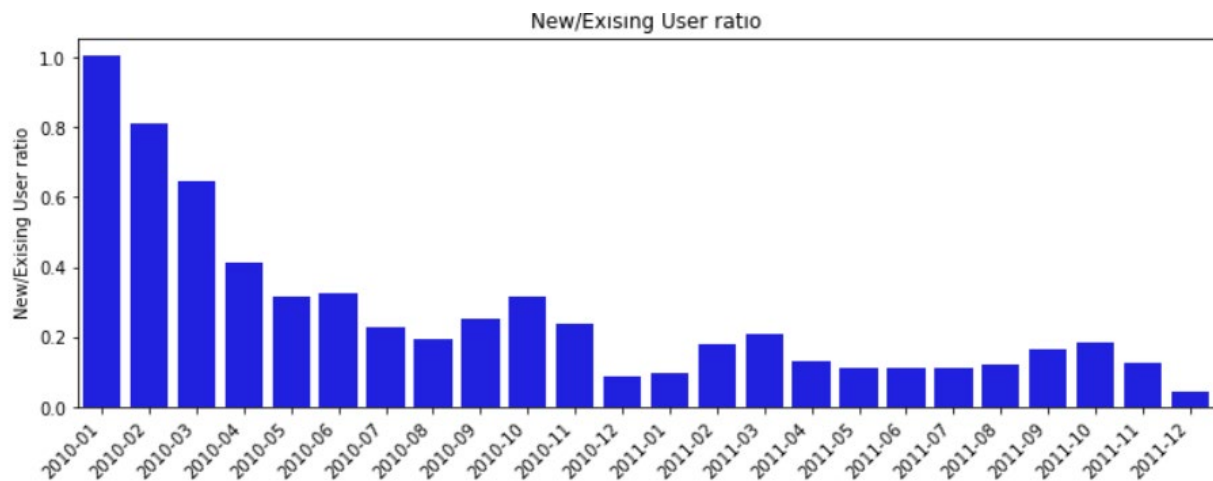


Figure 3.7 New – existing customer ratio

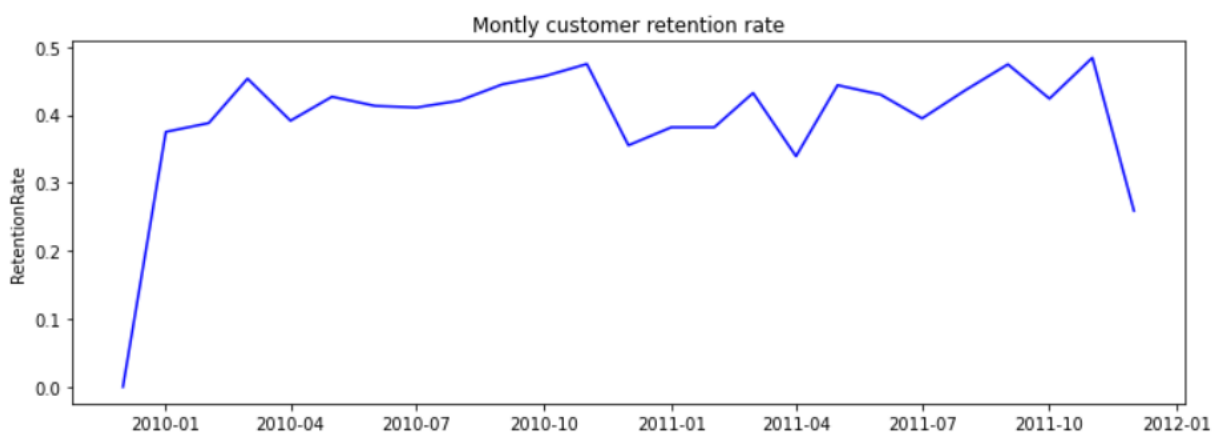


Figure 3.8 Retention rate

## 4. Clustering

The objectives of this step are

- Pre-process the features as necessary
- Segment customers based on various features.

### 4.1 Recency-Frequency-Monetary (RFM) value

Recency-Frequency-Monetary (RFM) value is a commonly used feature for customer segmentation.

#### 4.1.1 Recency calculation and clustering

To calculate recency, we need to find out the most recent purchase date of each customer and see how many days they are inactive. Figure 4.1 shows the recency in days for all customs. After having the number of inactive days for each customer, we apply the K-means clustering algorithm to assign customers a recency cluster/score as shown in Table 4.1. A high score means the customer is more active while a low score means the customer is less active.

Table 4.1 Recency scores/clusters

|                | count  | mean       | std       | min   | 25%   | 50%   | 75%    | max   |
|----------------|--------|------------|-----------|-------|-------|-------|--------|-------|
| RecencyCluster |        |            |           |       |       |       |        |       |
| 0              | 673.0  | 621.983655 | 65.012623 | 516.0 | 568.0 | 618.0 | 674.00 | 738.0 |
| 1              | 1133.0 | 409.353045 | 49.093665 | 310.0 | 378.0 | 407.0 | 441.00 | 515.0 |
| 2              | 968.0  | 209.136364 | 52.880052 | 123.0 | 164.0 | 205.0 | 254.25 | 309.0 |
| 3              | 3165.0 | 35.881201  | 31.345800 | 0.0   | 9.0   | 27.0  | 57.00  | 122.0 |

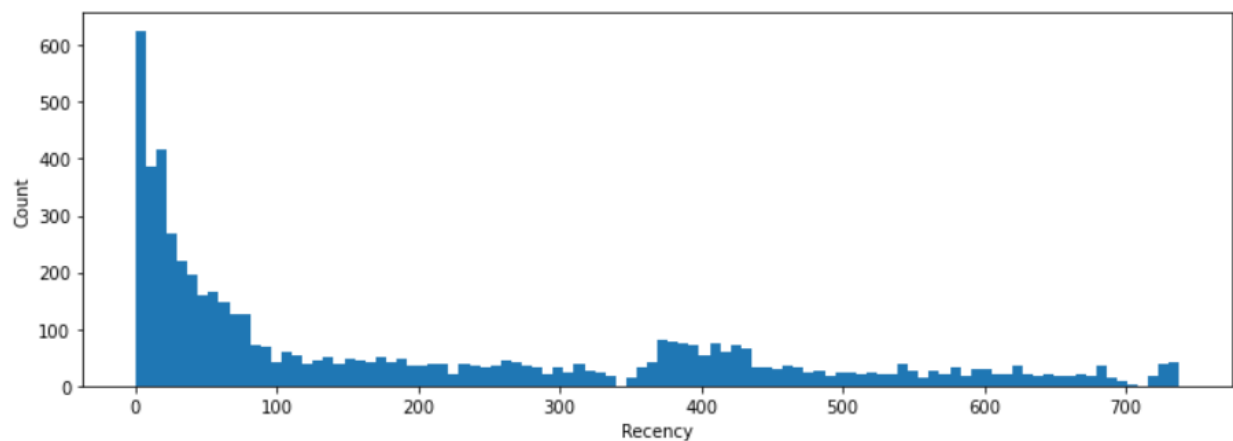


Figure 4.1 Recency distribution

### 4.1.2 Frequency calculation and clustering

To create frequency clusters, we need to find the total number of orders for each customer. Figure 4.2 shows the frequency in the number of orders for all customers. Then, we apply the KMeans clustering algorithm to assign customers a frequency cluster/score as shown in Table 4.2. A high score means the customer placed more orders while a low score means the customer placed fewer orders.

Table 4.2 Frequency scores/clusters

|                  | count  | mean         | std         | min     | 25%      | 50%     | 75%      | max     |
|------------------|--------|--------------|-------------|---------|----------|---------|----------|---------|
| FrequencyCluster |        |              |             |         |          |         |          |         |
| 0                | 5515.0 | 79.655485    | 85.588707   | 1.0     | 19.00    | 46.0    | 110.00   | 385.0   |
| 1                | 408.0  | 689.620098   | 306.629750  | 387.0   | 465.75   | 589.0   | 821.25   | 2077.0  |
| 2                | 14.0   | 3790.714286  | 1445.820844 | 2352.0  | 2728.00  | 3244.0  | 4463.75  | 6660.0  |
| 3                | 2.0    | 12040.000000 | 845.699710  | 11442.0 | 11741.00 | 12040.0 | 12339.00 | 12638.0 |

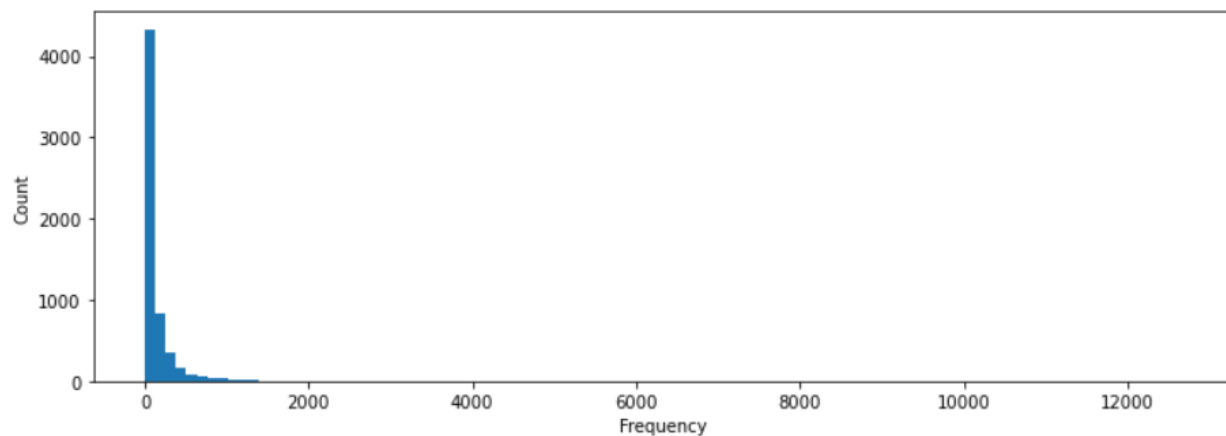


Figure 4.2 Frequency distribution

### 4.1.3 Monetary calculation and clustering

To create monetary clusters, we need to use the revenue. Figure 4.3 shows the monetary revenue for all customers. Then, we apply KMeans clustering algorithm to assign customers a monetary cluster/score as shown in Table 4.3. A high score means the customers have more monetary value while a low score means have a less monetary value.

Table 4.3 Revenue scores/clusters

|                | count  | mean          | std          | min       | 25%         | 50%        | 75%        | max       |
|----------------|--------|---------------|--------------|-----------|-------------|------------|------------|-----------|
| RevenueCluster |        |               |              |           |             |            |            |           |
| 0              | 5858.0 | 1821.652251   | 2830.594995  | -25111.09 | 317.0125    | 800.135    | 2052.270   | 21535.90  |
| 1              | 71.0   | 41740.448493  | 22088.363188 | 21893.53  | 25539.2050  | 33480.820  | 52250.470  | 111739.36 |
| 2              | 8.0    | 195182.235000 | 62923.301710 | 124961.98 | 142827.5300 | 179256.230 | 239982.390 | 296063.44 |
| 3              | 2.0    | 546861.340000 | 33261.270611 | 523342.07 | 535101.7050 | 546861.340 | 558620.975 | 570380.61 |

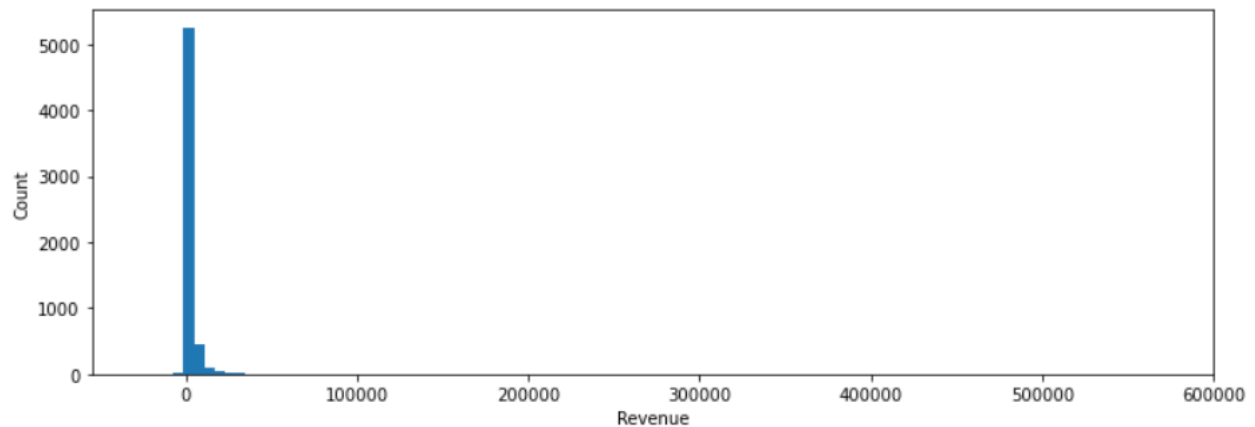


Figure 4.3 Frequency distribution

#### 4.1.4 RFM calculation and clustering

RMF values are calculated by adding up recency, frequency, and monetary scores, which have been shown in Table 4.4.

Table 4.4 RFM values

|           | Recency    | Frequency   | Revenue       |
|-----------|------------|-------------|---------------|
| RFM value |            |             |               |
| 0         | 621.494815 | 23.662222   | 275.126803    |
| 1         | 408.922872 | 49.085106   | 748.954301    |
| 2         | 209.935857 | 64.652997   | 1168.003012   |
| 3         | 39.865217  | 114.338768  | 2105.548216   |
| 4         | 20.097765  | 633.480447  | 9776.187405   |
| 5         | 14.553191  | 974.808511  | 32971.483468  |
| 6         | 4.384615   | 2747.000000 | 103344.460769 |
| 7         | 2.400000   | 5155.000000 | 231052.704000 |
| 8         | 0.500000   | 7663.500000 | 394549.990000 |

The scoring above clearly shows us that customers with a score of 8 are our best customers whereas 0 is the worst. We can name these scores:

- 0 to 2: Low Value: good candidates for improving retention
- 3 to 4: Mid Value: good candidates for improving retention and increasing frequency
- 5+: High Value: good candidates for increasing frequency

The clustering of customers can be visualized in Figures 4.4 and 4.5. In general, the RFM value segment the customers into three categories satisfactorily. But we can still see some customers have been mistakenly grouped.

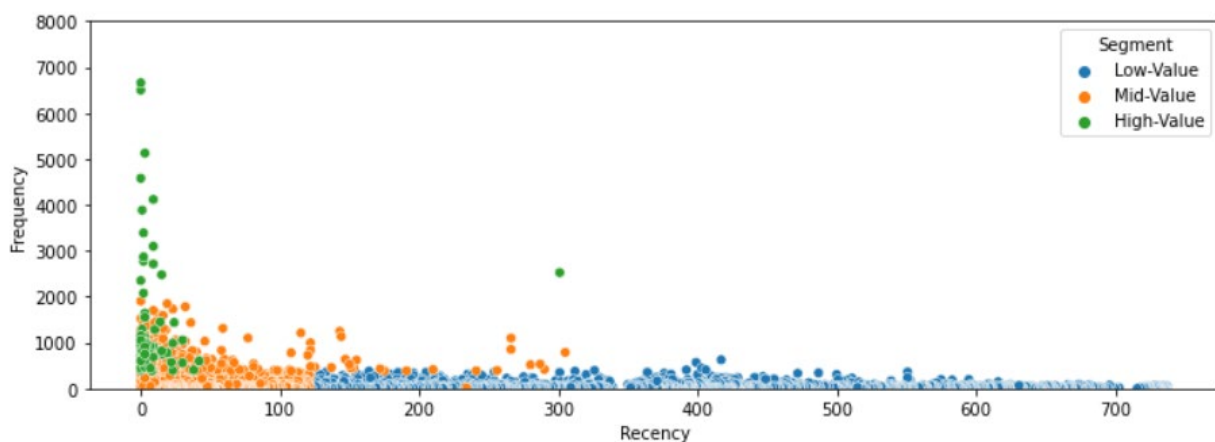


Figure 4.4 Customer Segmentation (Frequency vs. Recency)

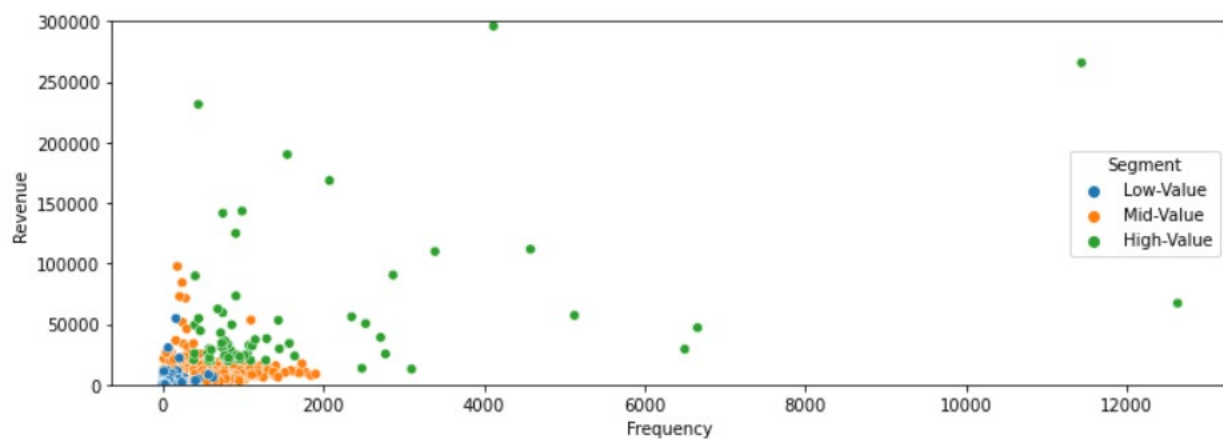


Figure 4.5 Customer Segmentation (Revenue vs. Frequency)



## 4.2 RFM + Price + Quantity

### 4.1.1 Data scaling

In this section, we will use the recency score, frequency score, and monetary score, together with price and quantity to build the customer clusters. Recency-Frequency-Monetary (RFM) value is a commonly used feature for customer segmentation. Table 4.5 shows the data for clustering. As demonstrated in previous sections, all the features are skewed. Therefore, scaling is applied to improve the quality of data. KMeans cluster relies on the distance calculation, which is sensitive to values of features. Therefore, it is necessary to process the multiple-feature data for clustering. Figure 4.6 illustrates the distribution of each feature after the following transformation.

- logarithmic transformation is applied to price and frequency
- cubic root transformation is applied to quantity and revenue
- We leave the recency unchanged.

Table 4.5 Statistics of data for clustering

|              | Price        | Quantity    | Recency     | Frequency    | Revenue       |
|--------------|--------------|-------------|-------------|--------------|---------------|
| <b>count</b> | 5939.000000  | 5939.000000 | 5939.000000 | 5939.000000  | 5939.000000   |
| <b>mean</b>  | 19.835088    | 18.512167   | 201.784812  | 134.334905   | 2742.884541   |
| <b>std</b>   | 405.031023   | 83.275333   | 211.727459  | 348.092835   | 13679.955199  |
| <b>min</b>   | 0.151333     | -16.000000  | 0.000000    | 1.000000     | -25111.090000 |
| <b>25%</b>   | 2.323423     | 5.155944    | 24.000000   | 20.000000    | 321.365000    |
| <b>50%</b>   | 3.010753     | 9.075472    | 95.000000   | 52.000000    | 823.530000    |
| <b>75%</b>   | 3.962383     | 13.409903   | 380.000000  | 140.000000   | 2143.280000   |
| <b>max</b>   | 25111.090000 | 3255.074627 | 738.000000  | 12638.000000 | 570380.610000 |

### 4.1.2 Silhouette analysis

The silhouette analysis indicates that 2 clusters yield the best result and 3 clusters may also be a good option if we know more information about the customers.

Table 4.5 Statistics of data for clustering

| Clusters | Silhouette score |
|----------|------------------|
| 2        | 0.485            |
| 3        | 0.392            |
| 4        | 0.350            |
| 5        | 0.374            |

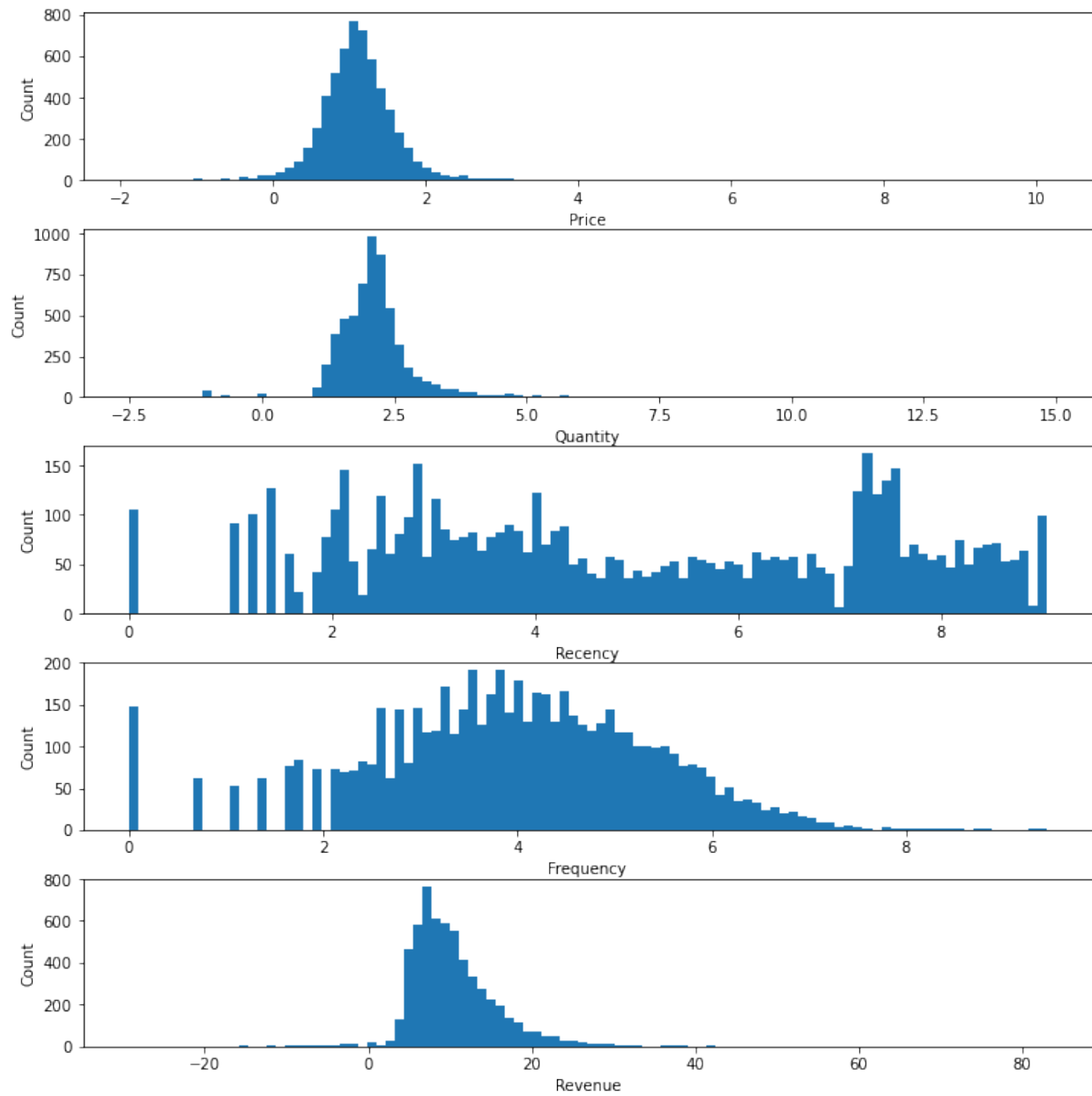


Figure 4.6 Distribution of features after transformation

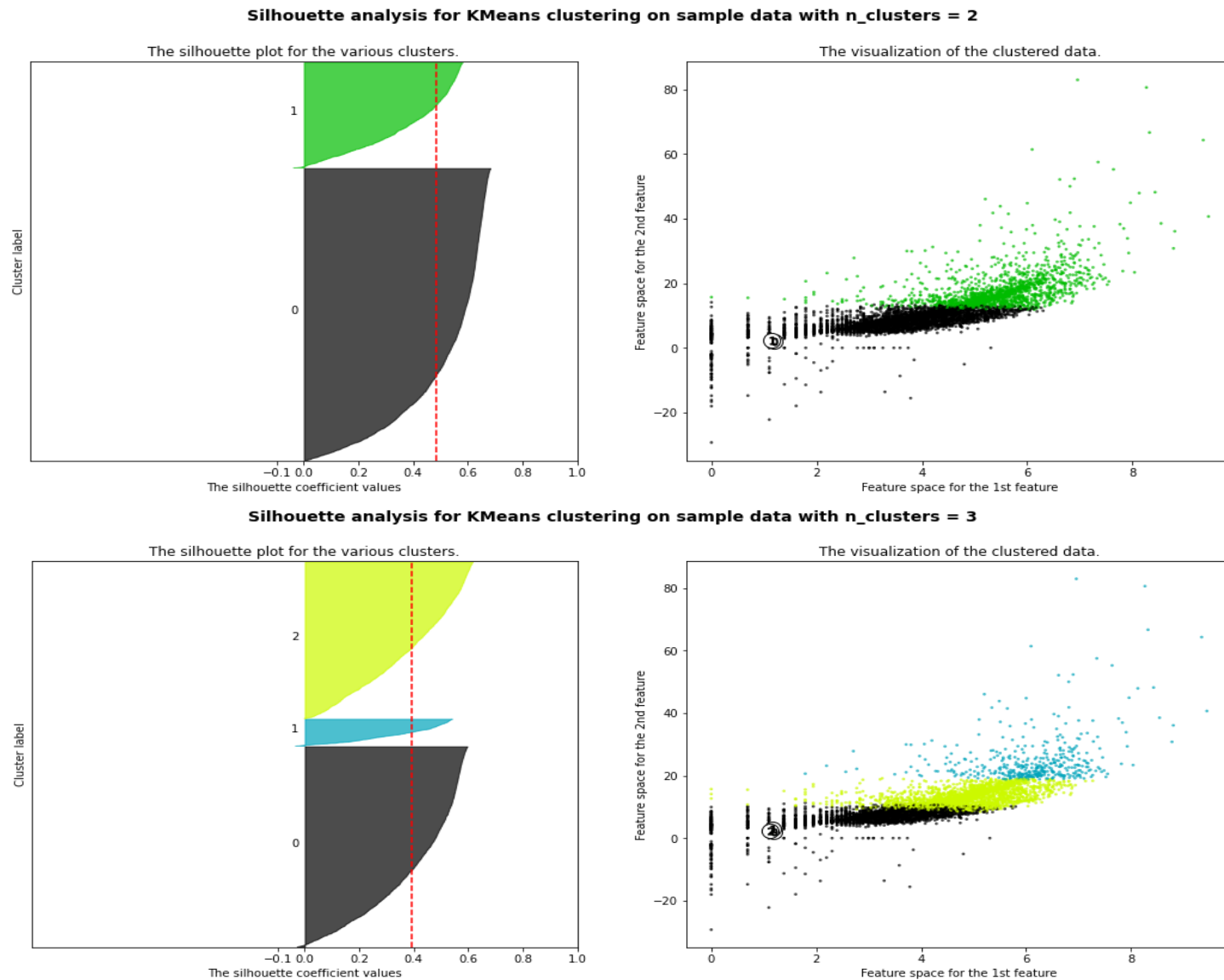


Figure 4.7 Silhouette analysis (clusters = 2 or 3)

Figure 4.7 shows more details for the 2 and 3 clusters silhouette analysis. It can be seen 2 clusters segment customers pretty well and the size of the clusters is reasonable.

Figure 4.8 shows two clusters on the revenue and frequency plot. Cluster 0 represents customers with high frequency and high revenue while Cluster 1 represents customers with relatively low frequency and low revenue. But there are still overlaps between the two clusters.

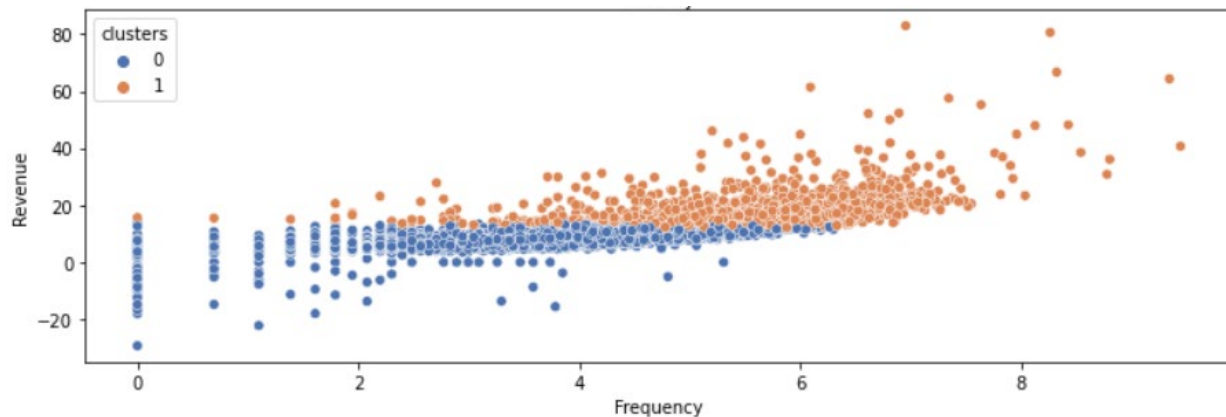


Figure 4.8 Two cluster customer segmentation

For three clusters, there is one relatively small size of the cluster as shown in Figure 4.7. Meanwhile, Figure 4.9 demonstrates three clusters on revenue and frequency plots. The small size cluster represents the highest frequency and revenue; therefore, it may be a good choice to make it an individual segment.

## 5. Conclusions

The customers of an online store have been segmented based on the numerical features of the dataset by using KMeans clustering. Two clusters work quite well with a silhouette score of 0.485. However, three clusters may also be a good option due to a more specific subset can be defined. The following observations are summarized from this project.

- All missing and duplicated records are dropped;
- EDA is conducted on all available features.
- The revenue analysis and customer analysis reveal the seasonality associated with sales and customer retention rates.
- Only numerical features are used in the modeling process.
- The Elbow method and silhouette analysis are used to determine the best clusters of KMeans algorithm.
- Two or three clusters of customers are recommended. With various business plans or more information about customers, corresponding strategies may be developed for either two or three clusters.

## 6. Recommendations & Future Work

To build a more robust clustering model for stakeholders, the following actions are recommended.

- Using natural language processing tools to interpret the stock code and description features, which may reveal customers' preference for goods and further split subsets for customer segmentation.
- Collecting and importing more data into the modeling. For example, the cost of goods can be used to estimate profit. High revenue doesn't mean high profit. It is critical for a business to identify which goods have the highest profit and which group of customers contribute the highest profits.