Customer Segmentation for Online Retailing

Jun Tian

January 2021

Table of Content

1. Introduction

This project is based on a transnational dataset which contains all the transactions occurring between December 1 2010 and December 9 2011 for a UK-based and registered non-store online retail. The dataset is downloadable from: https://archive.ics.uci.edu/ml/datasets/Online+Retail. The retailer mainly sells unique all occasion gifts and many of the customers of the retailer are wholesales, located in various countries.

There are 541,909 records in the dataset. For each of the records, there are 8 attributes, namely, InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CusomerID, and Country.

This project explores the dataset, aiming at identifying insightful findings that will help the retailer to get a better understanding of their customers, and therefore make effective and differentiated marketing decisions that are attractive to different customer groups, and therefore increase sales and profits.

2. Problem Statement

With 541,909 transaction records with 8 attributes for a period of more than 1 year, the dataset provides many perspectives that worth exploring. Due to the limitation of time and resources, this project focuses on the most important part of the business --- the customers. By looking at historical data, we would like to identify purchasing patterns of the customers. We would like to know more about customer profiles, similarities, and dissimilarities among the customers. Those information will help the retailer understand more details of the customers, and therefore can make more appropriate decisions targeting different customers. The research problem is **customer segmentation** of this online retailer.

There are many questions we would like to inquiry about customer behaviors and purchase patterns. We expect to address the following questions:

i. Who are the major customers and where are they? Any similarities among the customers with regard to geographical locations, favorable products, purchasing patterns and so on?
ii. Who are the most/least loyal customers and what are their characteristics?
iii. What are the most/least popular products?
iv. Any there sales patterns in terms of products, time, region, and so on?

3. Data

The dataset has all transactions from December 1 2010 to December 9 2011. There are 541,909 records in the dataset. For each of the records, there are 8 attributes, namely, InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CusomerID, and Country. The dataset is downloadable from: https://archive.ics.uci.edu/ml/datasets/Online+Retail. Each of the variables are described as:

- 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: 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.

4. Method

In order to address the stated problem and questions, data mining techniques will be adopted for this retailer. Data mining is a common practice and an integral part of business process in analyzing and supporting marketing. Explanatory data analysis will be conducted to explore different aspects of the dataset in terms of better understand its customers.

With regard to customer segmentation, a well-known business metrics RFM (recency, frequency, and monetary) model will be applied. R refers to recency, measuring how long is the customer's most recent purchase; F refers to frequency, how often does the customer make purchases; and M refers to monetary value, how much did the customer spend. RFM summarizes basic characteristics of customers' profitability and values therefore can be used as measurement metrics for customer segmentation.

Based on the RFM model, customers of the retailer will be segmented into various meaningful groups using the k-means clustering algorithm. Main characteristics of customers in each segment will be clearly identified. Accordingly, recommendations will be provided to the online retailer for its consideration of customer-centric marketing plans and further data analysis suggestions.

5. Data Wrangling

Good data is the very foundation of any data analysis. Data wrangling is an essential part for data preparation for further analysis. This section will follow the procedures of data wrangling to identifying and handling duplicates, missing data, data types, anomalies, and outliers, and eventually provide a clean dataset for next step research. In this process we will also standardize the timestamp since this dataset is a time series dataset. Other techniques also include proper indexing.

A snapshot of all the data looks like this:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom

A checkup on the dataset:

```
1 df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    541909 non-null  object
 1   StockCode    541909 non-null  object
 2   Description  540455 non-null  object
 3   Quantity     541909 non-null  int64
 4   InvoiceDate  541909 non-null  datetime64[ns]
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      541909 non-null  object
dtypes: datetime64[ns](1), float64(2), int64(1), object(4)
memory usage: 33.1+ MB
```

5.1 Duplicates

The very first step we would like to check if there is any duplicated record. If there are, they need to be dropped. After deleting all the duplicated records, the dataset provides the following information:

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 536641 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    536641 non-null  object
 1   StockCode    536641 non-null  object
 2   Description  535187 non-null  object
 3   Quantity     536641 non-null  int64
 4   InvoiceDate  536641 non-null  datetime64[ns]
 5   UnitPrice    536641 non-null  float64
 6   CustomerID   401604 non-null  float64
 7   Country      536641 non-null  object
dtypes: datetime64[ns](1), float64(2), int64(1), object(4)
memory usage: 36.8+ MB
```

5.2 Missing data

There are two variables with missing values ---- Description and CustomerID. There are 1,454 missing values for Description, or 0.26% of the total; and there are 135,037missing values for CustomerID, or 24.9% of the total.

For Description, since another variable StockCode is uniquely assigned to each product. StockCode and Description refer to each other and could be used as the basis for filling. We tried to look for missing Description through StockCode. However we did not find any useful information from this approach.

For CustomerID, we tried to identify the missing values through InvoiceNo. Unfortunately no useful information was obtained through this approach.

Since there is no way to fill the missing values, we would like to drop those records containing the missing values. After deleting them, we now have a dataset of 406,829 records.

5.3 Data types

Data types for all variables looks good except for CustomerID. For our analysis purpose, CustomerID should be a string that is uniquely assigned to each customer, instead of integers.

5.4 Anomalies and outliers

A check on the dataset description results in the following table:

	count	mean	std	min	25%	50%	75%	max
Quantity	401604.0	12.183273	250.283037	-80995.0	2.00	5.00	12.00	80995.0
UnitPrice	401604.0	3.474064	69.764035	0.0	1.25	1.95	3.75	38970.0

Several observations of anomalies can be made from this description table:

1. There are negative values for **Quantity**, which might be cancelled orders. They need to be cleaned from the dataset.
2. Quantity data are highly skewed, while the mean value is 12, the median is 5.00 and 75 percentile is 12.00, the maximum value is 80995.
3. There are zero values of **UnitPrice**, which is does not match with common business sense. Need further investigation.
4. UnitPrice is also highly skewed, while the mean value is 3.47, the median is 1.95 and 75 percentile is 3.75, the maximum value is 38970.0

An overview on variable **Description** found that there are some abnormal descriptions that need to be taken care of. Some of the abnormal descriptions are: “POSTAGE”, “DISCOUNT”, “CARRIAGE”, “MANUAL”, “PACKING CHARGE”, and “DOTCOM POSTAGE”. These descriptions indicate that records are directly relevant to handling and shipping.

A normal stock code is a five digit nominal. A check on variable **StockCode** found that there are some abnormal stock codes need to be taken care of. Some of the abnormal stock codes are: “POST”, “D”, “C2”, “M”, “PADS”, “DOT”, “CRUK”, “BANK CHARGES”. Again, it seems that those records are directly relevant to handling, shipping and bank fees. Those records need to be cleared.

Several steps are undertaken to clean the data, including: 1). Deleted all cancelled purchases and their counterparts in the dataset; 2) Deleted records with abnormal StockCode and Description that are obviously not relevant to products; 3) Deleted records of zero UnitPrice.

After all these steps, there are 388057 records remaining in the dataset, and the basic descriptions are as follows:

	count	mean	std	min	25%	50%	75%	max
Quantity	388057.0	12.667837	42.203930	1.00	2.00	6.00	12.00	4800.0
UnitPrice	388057.0	2.851500	3.983581	0.04	1.25	1.85	3.75	649.5

Both Quantity and UnitPrice are still highly skewed. For Quantity, the mean is 12.67 and 75 percentile is 12, the maximum is 4800. For UnitPrice, the mean is 2.85 and 75 percentile is 3.75, the maximum is 649.5. We double checked on records of high Quantity and high UnitPrice, and everything looks normal.

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
565145	22492	MINI PAINT SET VINTAGE	1152	2011-09-01 13:50:00	0.55	12415	Australia
537899	22328	ROUND SNACK BOXES SET OF 4 FRUITS	1488	2010-12-09 10:44:00	2.55	12755	Japan
579498	23084	RABBIT NIGHT LIGHT	2040	2011-11-29 15:52:00	1.79	12798	Japan
538420	17096	ASSORTED LAQUERED INCENSE HOLDERS	1728	2010-12-12 12:03:00	0.17	12875	United Kingdom
579936	21787	RAIN PONCHO RETROSPOT	1200	2011-12-01 10:07:00	0.65	12901	United Kingdom

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
44378	554836	22655	VINTAGE RED KITCHEN CABINET	1	2011-05-26 16:25:00	295.0	13015	United Kingdom
54943	536835	22655	VINTAGE RED KITCHEN CABINET	1	2010-12-02 18:06:00	295.0	13145	United Kingdom
71501	546480	22656	VINTAGE BLUE KITCHEN CABINET	1	2011-03-14 11:38:00	295.0	13452	United Kingdom
71502	547814	22656	VINTAGE BLUE KITCHEN CABINET	1	2011-03-25 14:19:00	295.0	13452	United Kingdom
179583	551393	22656	VINTAGE BLUE KITCHEN CABINET	1	2011-04-28 12:22:00	295.0	14973	United Kingdom
190602	556446	22502	PICNIC BASKET WICKER 60 PIECES	1	2011-06-10 15:33:00	649.5	15098	United Kingdom

Again, we checked on Description and StockCode to make sure there are no handling, shipping, and bank charges in the dataset.

5.5 Adding new features

For better analyzing the data, a few features were added to the dataset. The first one is to translate the InvoiceDate to more time-relevant variables, such as day, month, weekday, and hour. The second one is a new variable of Spending was created by multiplying Quantity and UnitPrice.

6. Explanatory Data Analysis

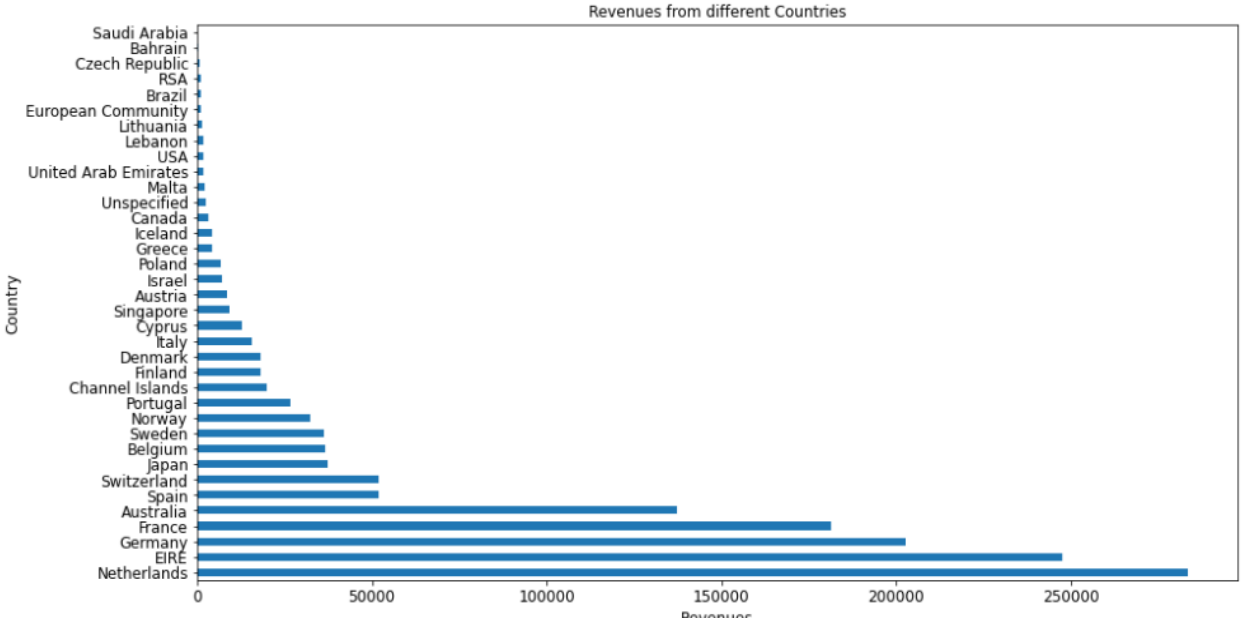
There are 388,057 transactions in the clean dataset. In total, 4,915,843 items were purchased from the retailer, and the total revenue from these purchases adds up to $8.3 million for the period 12/1/2010 to12/9/2011.

6.1 Geographical analysis

Customers from 37 countries and regions have made purchases from this online retailer. An overlook of the dataset found that UK is the predominating market of this online retailer, accounting for more than 80% of total purchases, quantity, and revenue. The following table shows top 12 countries with the highest purchases.

	# Transactions	% Transaction	Quantity	% Quantity	Revenue	% Revenue
UK	346340	89.2%	4018793	81.8%	6828573	82.3%
Netherlands	2318	0.6%	199934	4.1%	283443.5	3.4%
EIRE	6990	1.8%	136180	2.8%	247414.3	3.0%
Germany	8568	2.2%	117032	2.4%	202749.2	2.4%
France	7940	2.0%	109141	2.2%	181483.1	2.2%
Australia	1112	0.3%	83461	1.7%	137106.2	1.7%
Spain	2393	0.6%	26655	0.5%	52029.97	0.6%
Switzerland	1802	0.5%	29734	0.6%	52017.45	0.6%
Japan	320	0.1%	25976	0.5%	37314.37	0.4%
Belgium	1925	0.5%	22915	0.5%	36742.29	0.4%
Sweden	425	0.1%	35853	0.7%	36410.83	0.4%
Norway	1039	0.3%	19192	0.4%	32265.76	0.4%

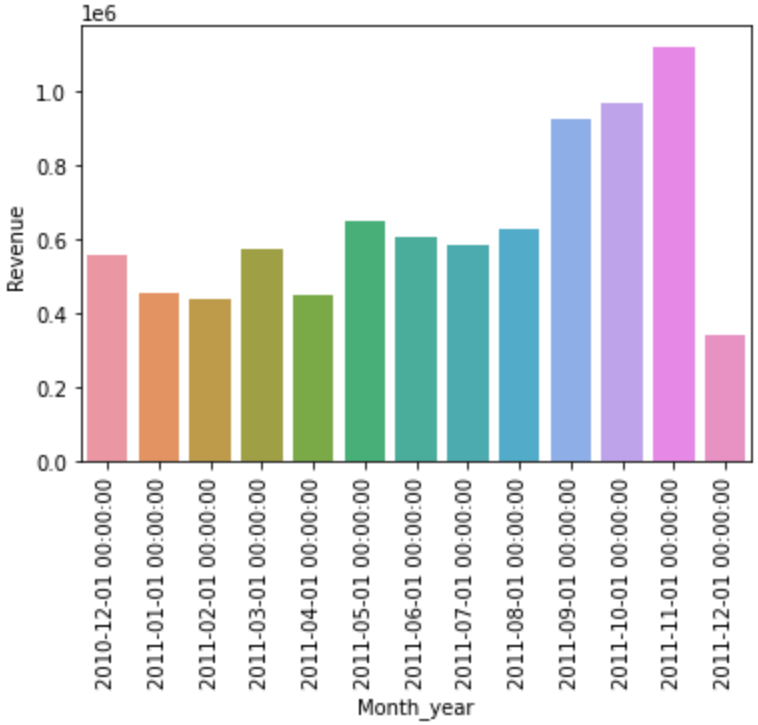
UK contributed 82.3% of the retailer's revenue. More than 25 fold of the second leading market – Netherlands, which accounts for 3.4%. Followed markets are EIRE (3.0%), Germany (2.4%),France (2.2%), and Australia (1.7%). The following graph illustrates revenues from various countries and regions, excluding UK.



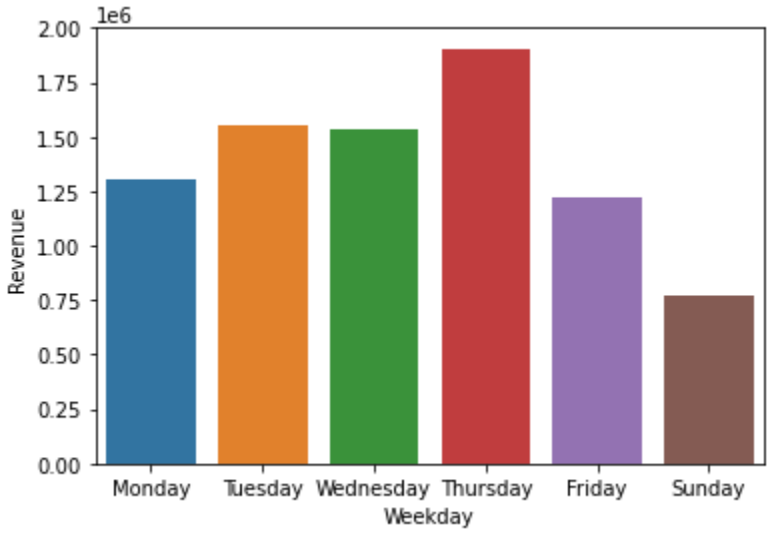
In terms of number of transactions and in terms of items purchased, there are slightly different orders of different countries and regions, but overall, the countries contribute the most to revenues are also leading countries in transactions and items purchased.

6.2 Time analysis

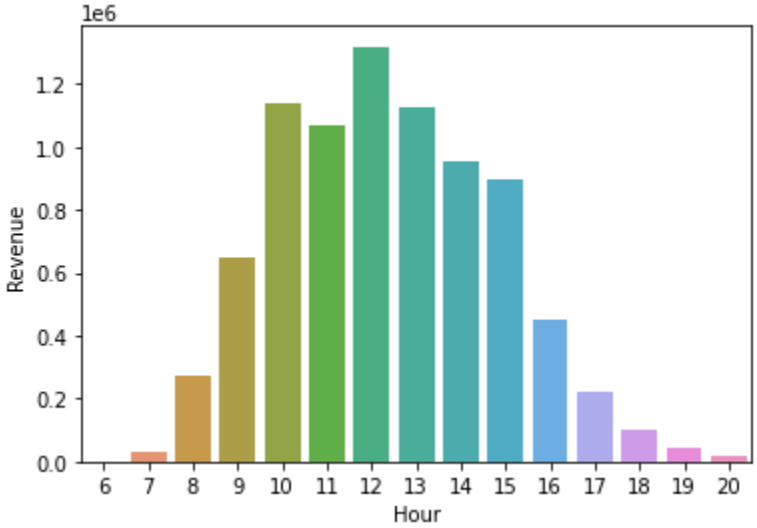
A snapshot of transactions along the timeline reveals that there are monthly variations in terms of purchases. More purchases were made close to the end of year, and therefore the revenues. November, October, and September are top three months when there are most transactions and revenues. This might due to the fact that wholesale consumers are actively preparing for the holiday season. But since the dataset covers only a period of slightly over one year, we would need more yearly data to confirm this finding.



In terms of weekdays, **Thursday is the day when there are more transactions and revenues than other days.** Wednesday is the day when there are the second largest transactions and revenues. No transaction was made on Saturdays.

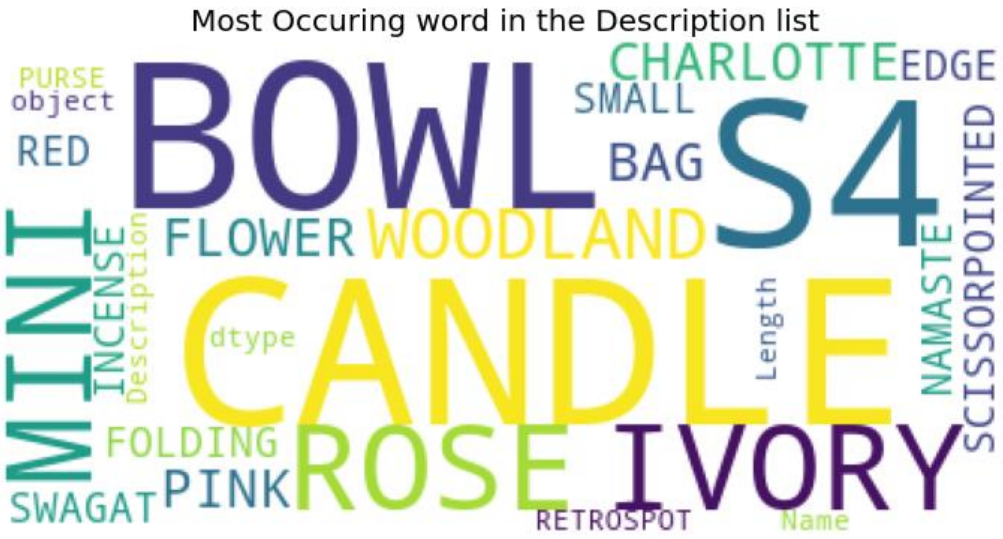


In terms of specific hours, **noon is the time when there are more transactions and revenues than any other hours.** 13PM, 14PM 11AM, 15PM, 10AM are the time when there are higher transactions. Wither regarding to revenues, 10AM, 13PM, 11AM, 14PM, 15PM and 9AM are the times when there are more revenues.



6.3 Product analysis

In total, there are 3845 different products in the dataset. A word cloud is generated based on the popularity of various products.



Most purchased products:

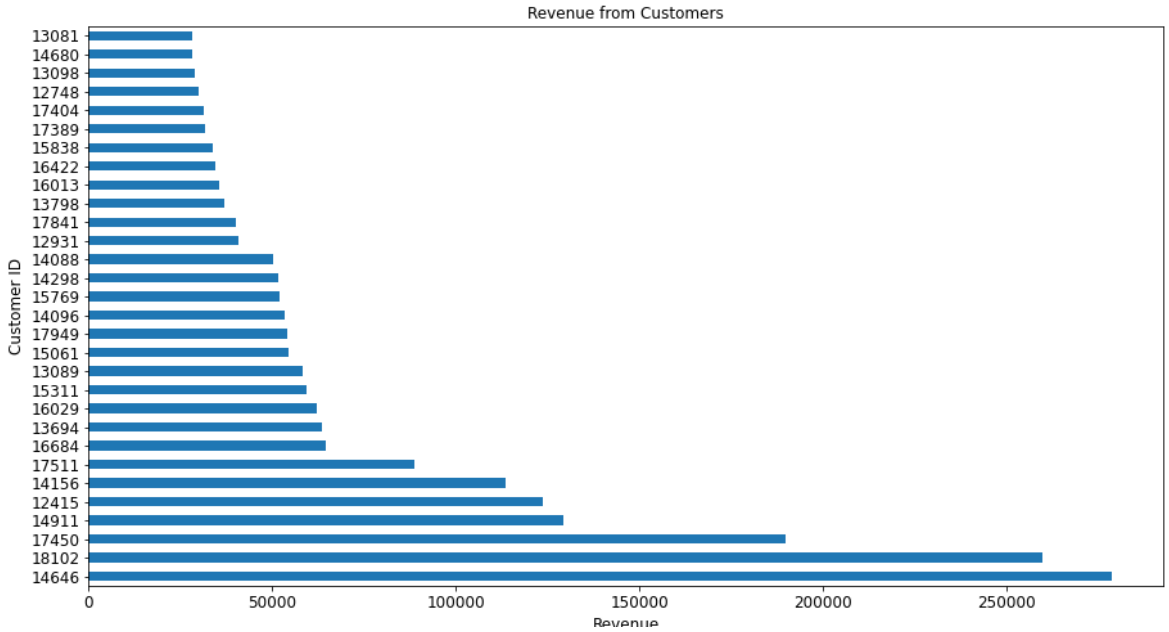
Description	
WHITE HANGING HEART TLIGHT HOLDER	1115526291
REGENCY CAKESTAND 3 TIER	930795823
JUMBO BAG RED RETROSPOT	897844732
ASSORTED COLOUR BIRD ORNAMENT	778576295
PARTY BUNTING	768658517

Most profitable products:

Description	
REGENCY CAKESTAND 3 TIER	138337.60
WHITE HANGING HEART TLIGHT HOLDER	93918.30
JUMBO BAG RED RETROSPOT	83481.96
PARTY BUNTING	67954.83
ASSORTED COLOUR BIRD ORNAMENT	56364.02

6.4 Customer analysis

In total, there are 4324 different customers and many of them are quite active. For example, 75 customers have made more than 500 purchases during the time period, and 20 of them have purchased more than 1,000 times, three of them have purchased 5000 times, and the highest purchases made by one customer is 7,566! (calculation based on InvoiceNo).



In terms of Quantity, 45 customers purchased more than10,000 items during the time period, 9 of them purchased more than 50,000 items, and customer #14646 purchased 196,556 items, more than double of the second largest customer of #14911, who purchased 77,103 items.

There are 96 customers who spent more than $10,000 during the time period, of which 6 customers spent more than $100,000. Customer #14646 spent $278,742.02 and customer #18102 spent $259,657.30.

7. Feature Engineering and Machine Learning

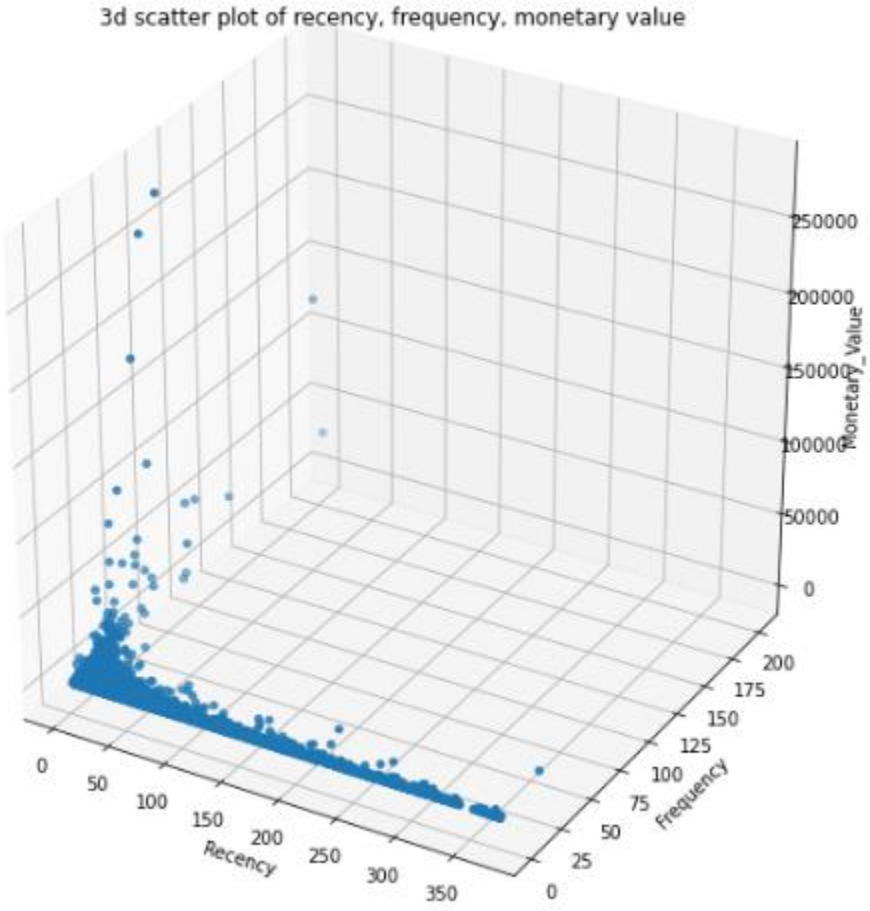
7.1 RFM model

There are many approaches to analyze customers. For this research we will adopt a well-known model called RFM. R represents recency, or days since last purchase. It answers the question of how many days ago was the customer's last purchase. Recency was calculated by deducting most recent purchase date from the current date. F represents frequency, or total number of transactions. It answers the question of how many times has the customer purchased from the retailer. Frequency was calculated by counting how many times the customer purchased during the time period. M represents total money spent. It answers the question o how much has the customer spent in the time period. M was calculated by simply adding up the money from all transactions.

Overall, RFM summaries some of the most important characteristics of the customers and can be used as the foundation for customer segmentation. A sample of RFM for this online retailer looks as follows:

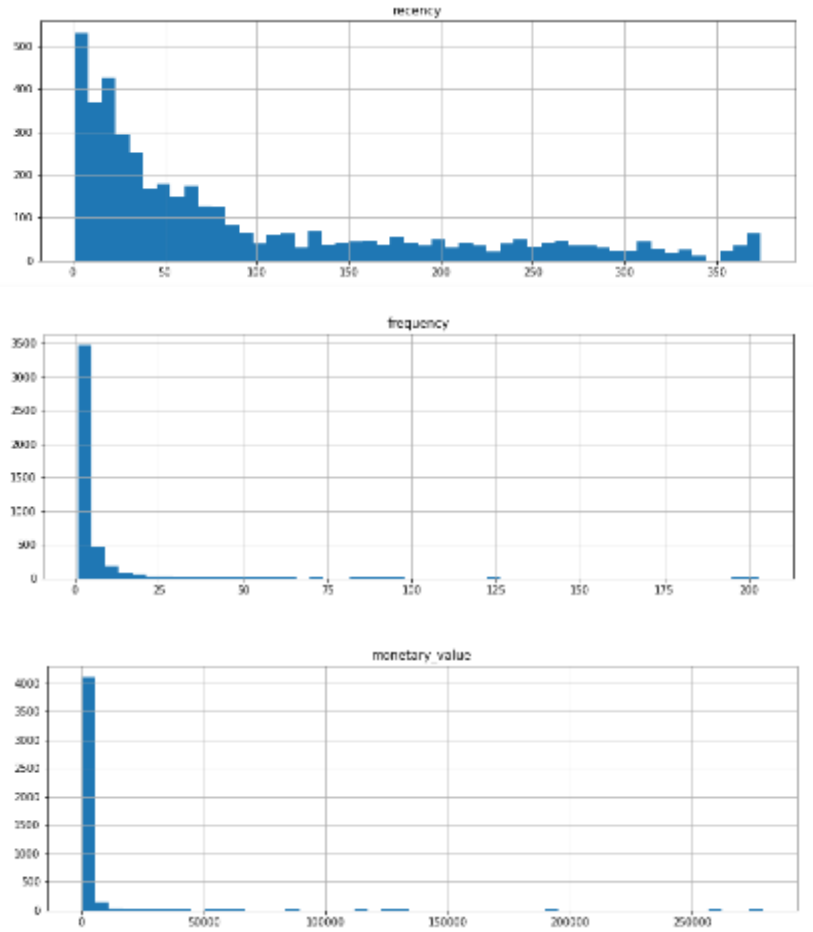
CustomerID	recency	frequency	monetary_value
12347	3	7	4310.00
12348	76	4	1437.24
12349	19	1	1457.55
12350	311	1	294.40
12352	37	6	1265.41

The 3D scatter plot of recency, frequency, and monetary values illustrate that the data are highly skewed. A brief statistical description of RFM confirms this. Recency ranges from 1 to 374, with a mean of 93 and median of 51. Frequency has a wider range, varies from 1 to 203, with the mean of 4 and 75 percentile of 5. Monetary value varies the most, the minimum is 2.9 and maximum is 278,742, the mean is 1,919 and 75 percentile is 1,613.

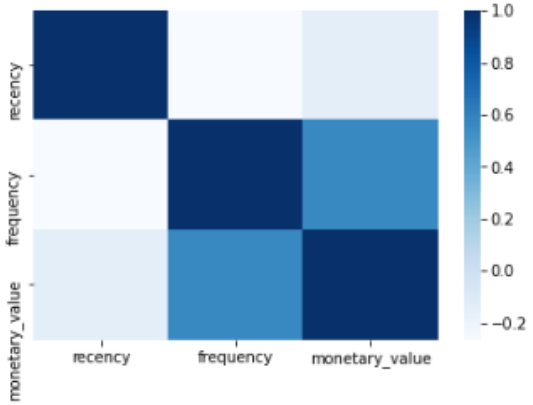


	recency	frequency	monetary_value
count	4324.000000	4324.000000	4324.000000
mean	93.350833	4.226642	1919.261025
std	100.258736	7.567506	8307.430912
min	1.000000	1.000000	2.900000
25%	18.000000	1.000000	300.752500
50%	51.000000	2.000000	656.430000
75%	144.000000	5.000000	1613.095000
max	374.000000	203.000000	278742.020000

The following three histograms show the distribution of recency, frequency, and monetary values:

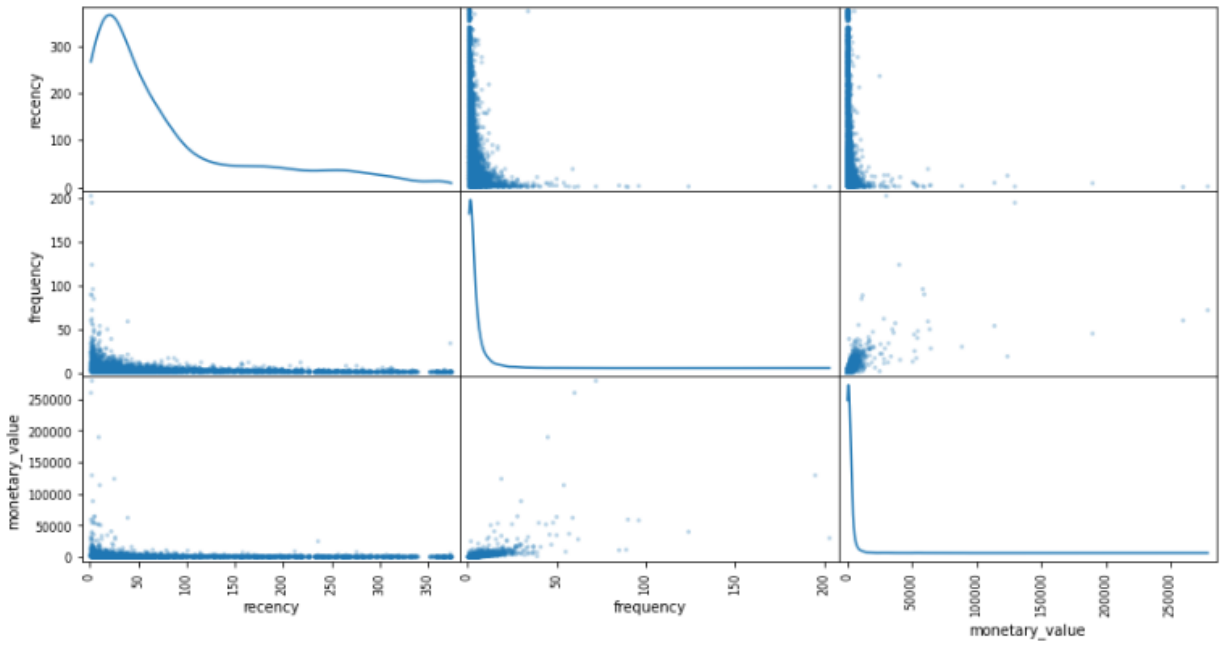


A further look into the correlation among R, F, and M found that frequency and monetary value has a positive correlation at 0.58. Recency has a negative correlation with frequency and monetary value, at a coefficient of -0.26 and -0.13 respectively.



	recency	frequency	monetary_value
recency	1.000000	-0.262566	-0.130211
frequency	-0.262566	1.000000	0.575302
monetary_value	-0.130211	0.575302	1.000000

The scatter plots below provide a better illustration:

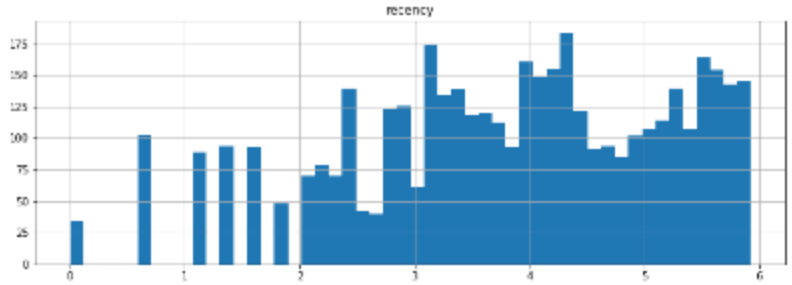


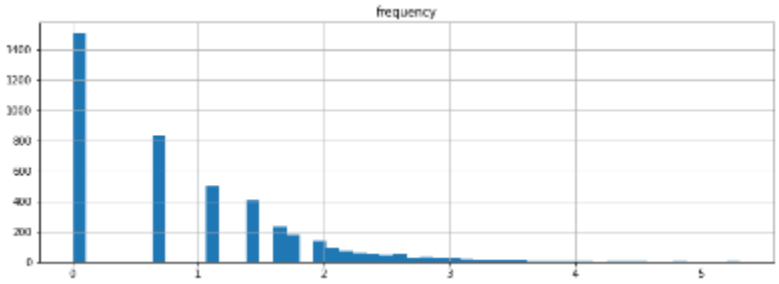
7.2 Feature scaling

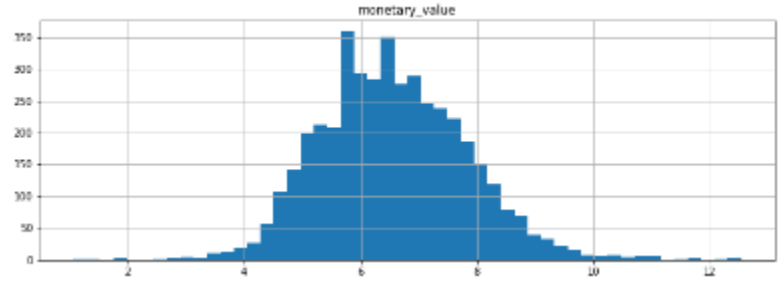
RFM values are all highly skewed. To make the data ready for machine learning, log transformation is made. After the transformation, here is the statistics of the data:

```
1 rfm_log = np.log(rfm)
2 rfm_log.describe()
```

	recency	frequency	monetary_value
count	4324.000000	4324.000000	4324.000000
mean	3.804122	0.936621	6.556687
std	1.383588	0.897661	1.254719
min	0.000000	0.000000	1.064711
25%	2.890372	0.000000	5.706288
50%	3.931826	0.693147	6.486816
75%	4.969813	1.609438	7.385910
max	5.924256	5.313206	12.538042

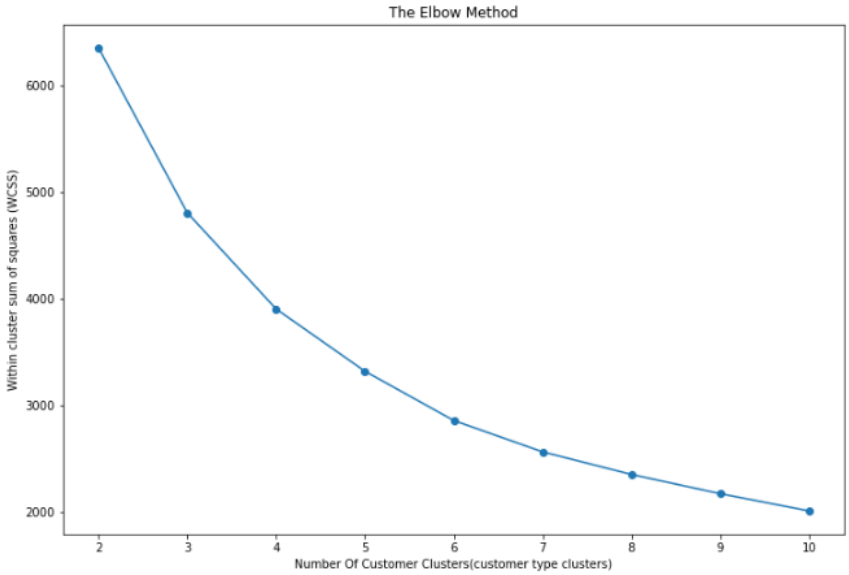




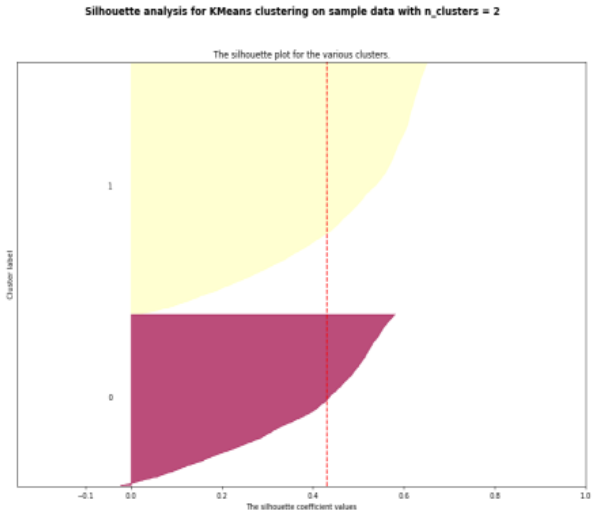


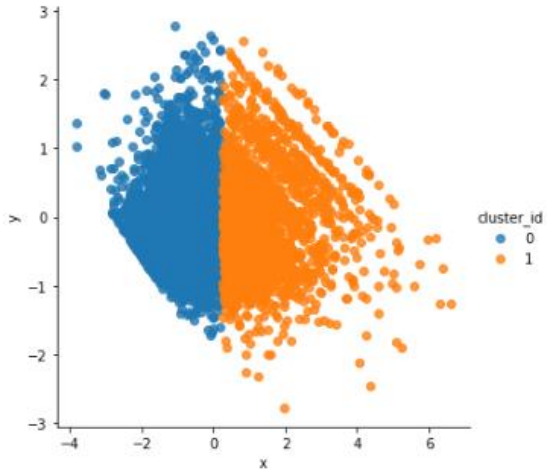
7.3 Machine learning applications

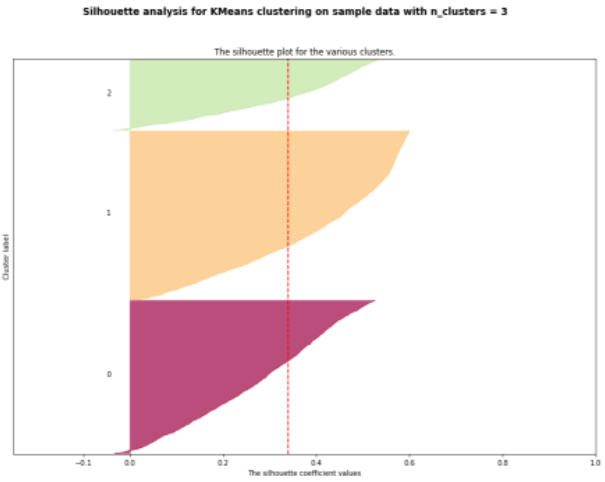
The log transformed RFM data was feed in a k-means algorithm for clustering. With a range of k from 2 to 11, a graph was plotted to help the selection of a suitable k value for machine learning.

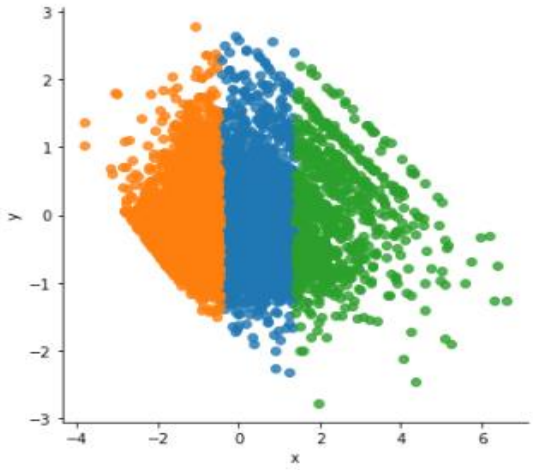


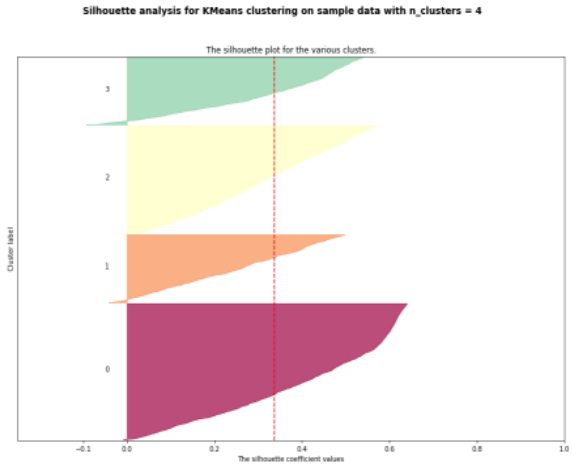
With the elbow method, k = 4 seems a good choice. We will confirm this with silhouette analysis and graph illustration.

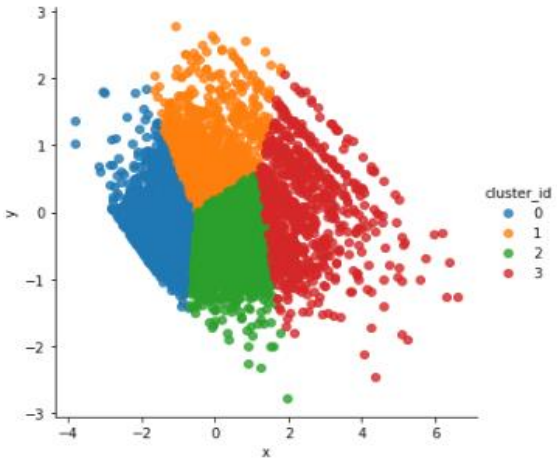






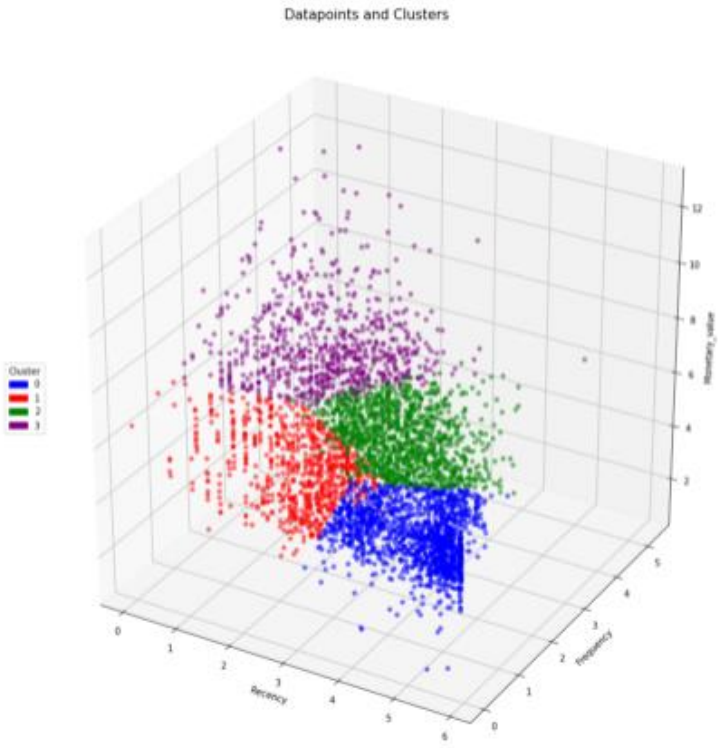






From the graphs, k = 4 is a plausible selection for the model. Thus, the customers are clustered into 4 groups. Each group contains customers that present similar characteristics in terms of recency, frequency and monetary values. A breakdown of clusters and customers show that the first cluster (cluster 0) has 1,551 customers, the second one (cluster 1) has 775 customers, the third one (cluster 2) has 1,233, and the fourth one (cluster 3) has 765 customers.

The following 3D scatter plot shows the four clusters of customers with relevant to log transformed recency, frequency, and monetary values.

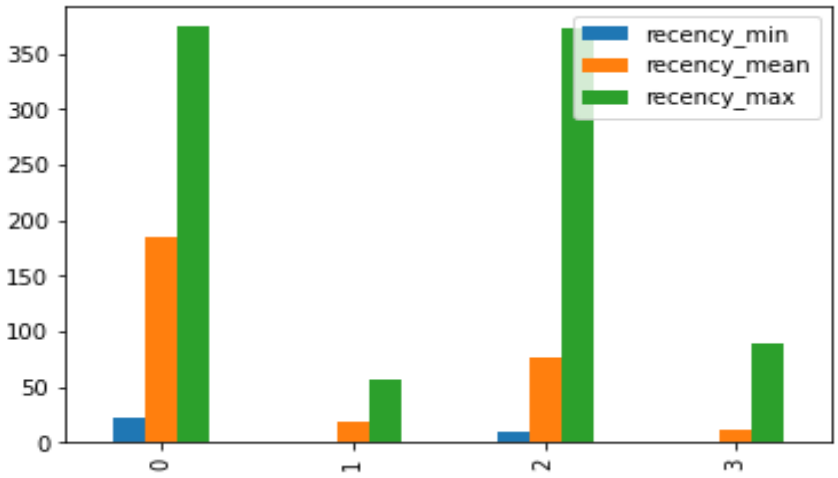


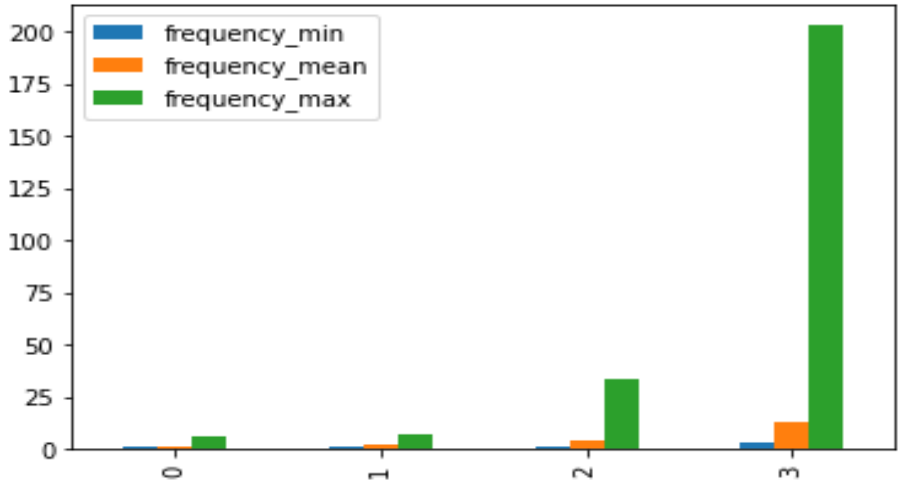
Then we go back to the RFM characteristics of different clusters. The first cluster of 1,551 customers has the longest recency, indicating that they are more likely old customers. They are the least frequency buyers, meaning that their purchasing activities are quite inactive. As a result, they contribute the least to the monetary values that the retailer made.

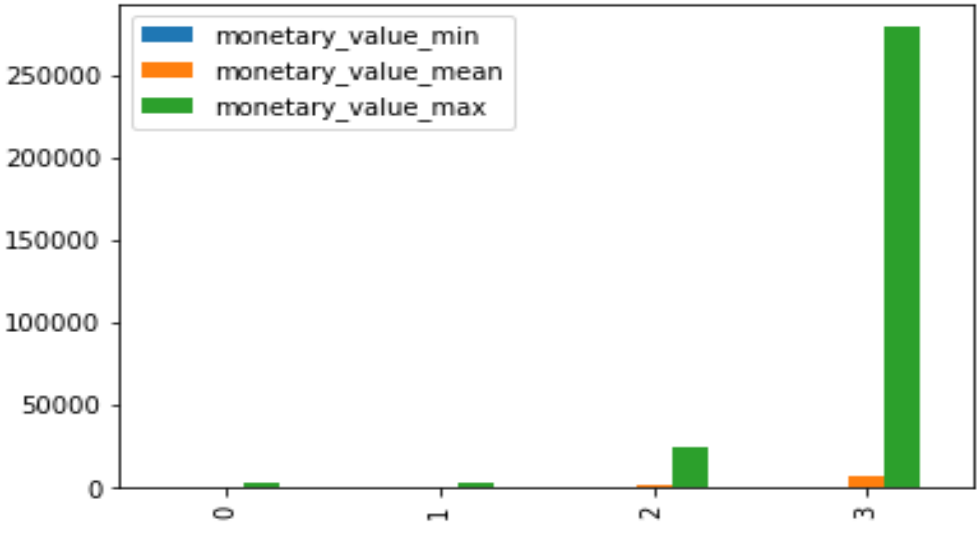
The second cluster of 775 customers has the shortest recency, indicating that their last purchases were recent. Their frequency value is low. As a result, they did not contribute much to the monetary values.

The third cluster of 1,233 customers has the second longest recency, indicating that they are old customers. They have the second largest frequency, meaning that they used to be active buyer, but somehow they stopped doing that anymore. Their contribution to monetary value was second to the fourth cluster.

The fourth cluster of 765 customers has relative short recency, indicating that their last purchases were recent. They have the highest frequency, meaning that they are quite active buyers. As a result, they contribute the most to the monetary values the retailer made. Those customers are very important to the retailer.


recency_min
recency_mean
recency_max
350
300
250
200
150
100
50
0
0
1
2
3


frequency_min
frequency_mean
frequency_max
200
175
150
125
100
75
50
25
0
0
1
2
3


monetary_value_min
monetary_value_mean
monetary_value_max
250000
200000
150000
100000
50000
0
0
1
2
3

8. Recommendations

This research focused on customer segmentation with k-means algorithm. The customers were group into 4 clusters and basic characteristics of each cluster are summarized.

A few recommendations to the online retailer:

- The fourth cluster of 765 customers is very important to the business. They have been very active in purchasing and the main marketing decision should focus on how to retain them to the business.
- The 1,551 customers in the first cluster are old customers who are not active for a long time.
- The 775 customers in the second cluster are inactive, but their last purchase was recent.
- The 1,233 customers in the third cluster are old customers who used to be active but not anymore.

Four clusters of customers presented quite different purchasing characteristics. The retailer should take these characteristics into account when making marketing decisions targeting each cluster of customers.

To better understand each cluster of customers, further research is required to explore their characteristics in details in addition to RFM.

As customer behaviors change over time, more recent data is needed. And data for a longer period of time would be better than one year of data.

In the data cleaning process, we deleted transactions with missing CustomerIDs. Those account for 24.9% of the total transaction, which might affect the model result. Therefore the retailer is recommended to have a complete record of transactions.