**Customer Segmentation**

**for Online Retailing**



**Jun Tian**

**January 2021**

**Table of Content**

Contents

[1. Introduction 1](#_Toc60774780)

[2. Problem Statement 1](#_Toc60774781)

[3. Data 2](#_Toc60774782)

[4. Method 2](#_Toc60774783)

[5. Data Processing, Preparation, & Feature Engineering 4](#_Toc60774784)

[6. Explanatory Data Analysis 6](#_Toc60774785)

[7. Feature Engineering and Machine Learning 11](#_Toc60774786)

[8. 4. Recommendations 24](#_Toc60774787)

# Introduction

This project is based on a transnational dataset which contains all the transactions occurring between 1/12/2010 and 9/12/2011 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 strategic decisions to attract and keep customers, increase sales and profits.

# 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:

1. 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?
2. Who are the most/least loyal customers and what are their characteristics?
3. What are the most/least popular products?
4. Any there sales patterns in terms of products, time, region, and so on?

# 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.

# Method

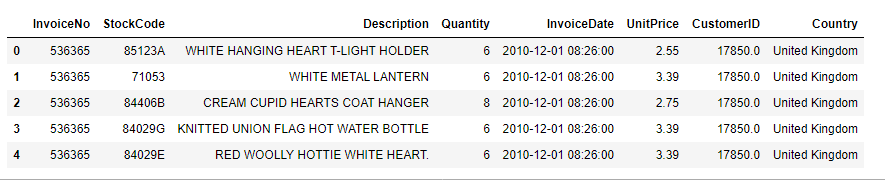
In order to address the stated problem and questions, data mining techniques will be adopted for this retailer. Data mining is a common practice and an integral part of business process in analyzing and supporting marketing. Explanatory data analysis will be conducted to explore different aspects of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

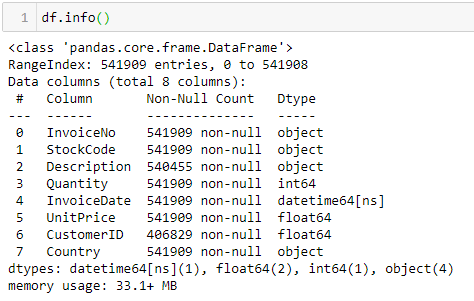
Based on the RFM model, customers of the retailer will be segmented into various meaningful groups using the k-means clustering and decision tree algorithms. Main characteristics of customers in each segment will be clearly identified. Accordingly, recommendations will be provided to the online retailer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

# 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.

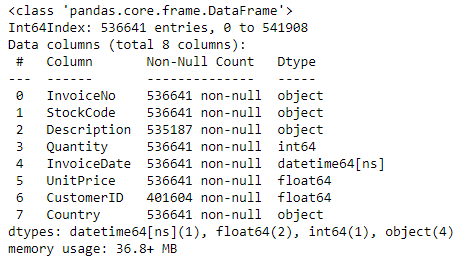
A snapshot of all the data looks like this:  


A checkup on the dataset:



## 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:



## 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.

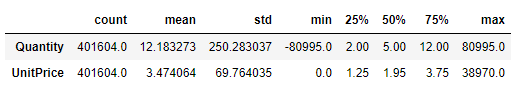
Since there is no way to fill the missing values, we would like to drop those records containing the missing values. After deleting them, we now have a dataset of 406,829 records.

## 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.

## Anomalies and outliers

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



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

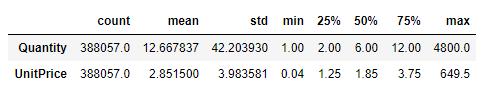
* 1. There are negative values for **Quantity**, which might be cancelled orders. They need to be cleaned from the dataset.
  2. Quantity data are highly skewed, while the mean value is 12, the median is 5.00 and 75 percentile is 12.00, the maximum value is 80995.
  3. There are zero values of **UnitPrice**, 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.

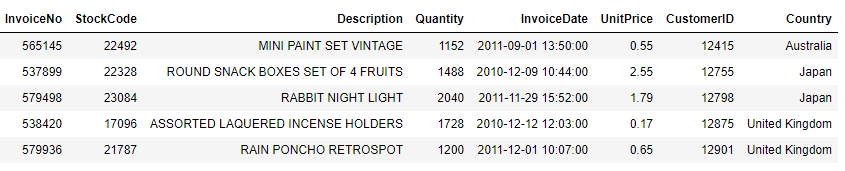
A normal stock code is a five digit nominal. A check on variable **StockCode** found that there are some abnormal stock codes need to be taken care of. Some of the abnormal stock codes are: “POST”, “D”, “C2”, “M”, “PADS”, “DOT”, “CRUK”, “BANK CHARGES”. Again, it seems that those records are directly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

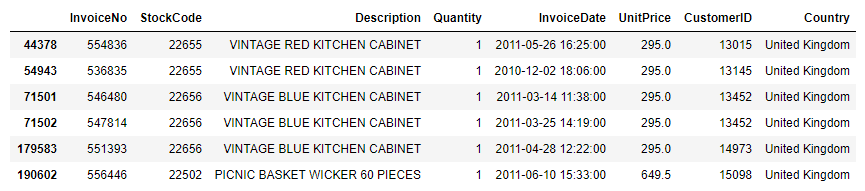
Several steps are undertaken to clean the data, including: 1). Deleted all cancelled purchases and their counterparts in the dataset; 2) Deleted records with abnormal StockCode and Description that are obviously not relevant to products; 3) Deleted records of zero UnitPrice.

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



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.





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

## 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.

# Explanatory Data Analysis

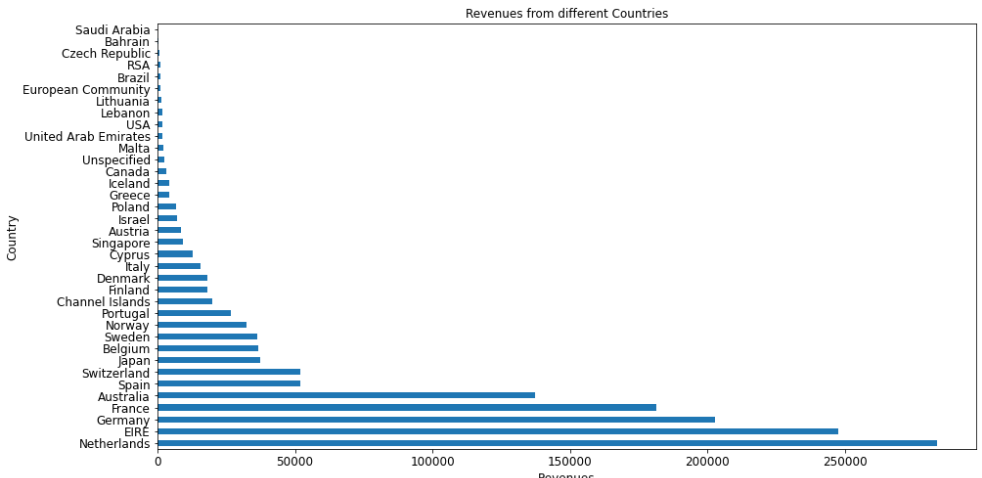
There are 388,057 transactions in the clean dataset. In total, 4,915,843 items were purchased from the retailer, and the total revenue from these purchases adds up to $8.3 million for the period 12/1/2010 to12/9/2011.

## 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% |

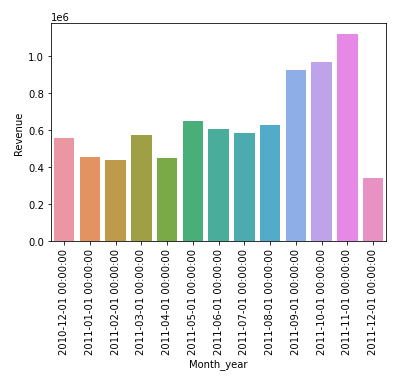
UK contributed 82.3% of the retailer’s revenue. More than 25 fold of the second leading market – Netherlands, which accounts for 3.4%. Followed markets are EIRE (3.0%), Germany (2.4%),France (2.2%), and Australia (1.7%). The following graph illustrates revenues from various countries and regions, excluding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



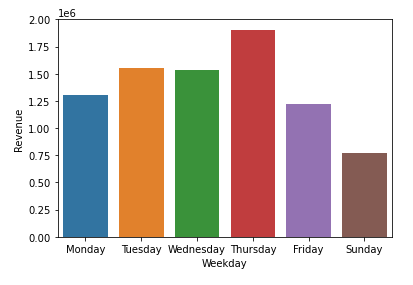
In terms of number of transactions and in terms of items purchased, there are slightly different orders of different countries and regions, but overall, the countries contribute the most to revenues are also leading countries in transactions and items purchased.

## 6.2 Time analysis

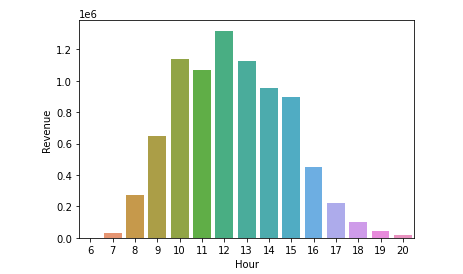
A snapshot of transactions along the timeline reveals that there are monthly variations in terms of purchases. More purchases were made close to the end of year, and therefore the revenues. November, October, and September are top three months when there are most transactions and revenues. This might 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.

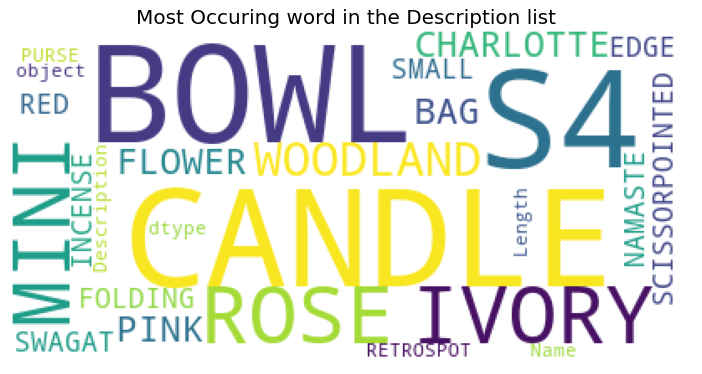


In terms of specific hours, **noon is the time when there are more transactions and revenues than any other hours.** 13PM, 14PM 11AM, 15PM, 10AM are the time when there are higher transactions. Wither regarding to revenues, 10AM, 13PM, 11AM, 14PM, 15PM and 9AM are the times when there are more revenues.

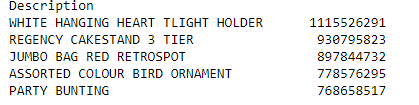


## 6.3 Product analysis

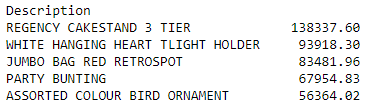
In total, there are 3845 different products in the dataset. A word cloud is generated based on the popularity of various products.



Most purchased products:

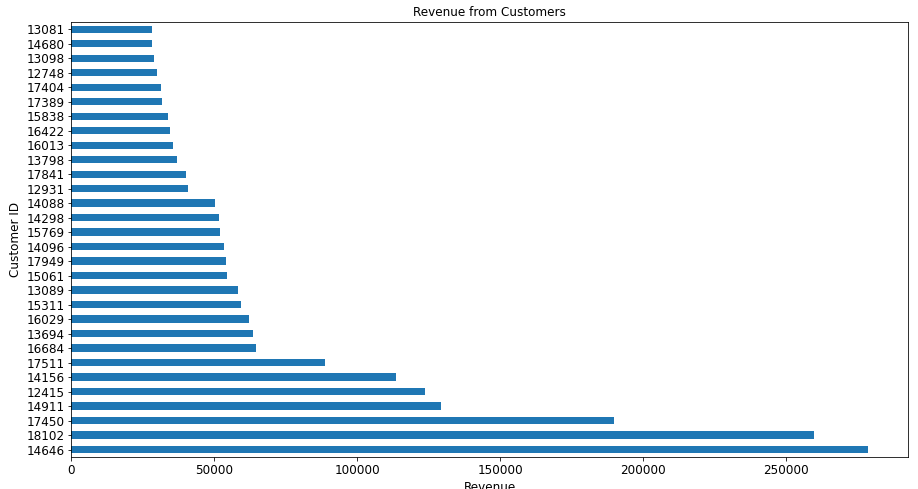


Most profitable products:



## 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.

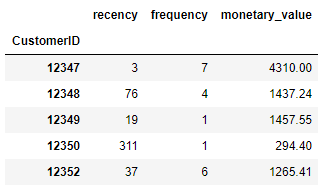
There are 96 customers who spent more than $10,000 during the time period, of which 6 customers spent more than $100,000. Customer #14646 spent $278,742.02 and customer #18102 spent $259,657.30.

# Feature Engineering and Machine Learning

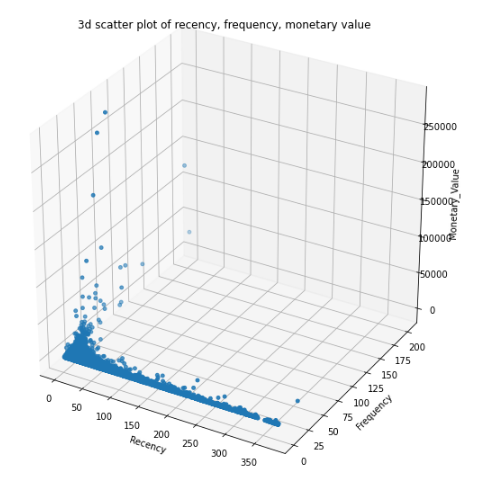
## 7.1 RFM model

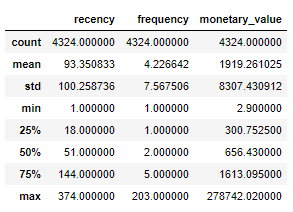
There are many approaches to analyze customers. For this research we will adopt a well-known model called RFM. R represents recency, or days since last purchase. It answers the question of how many days ago was the customer’s last purchase. Recency was calculated by deducting most recent purchase date 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:

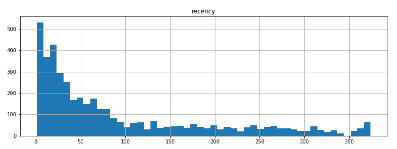


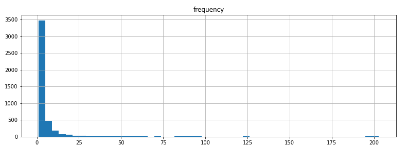
The 3D scatter plot of recency, frequency, and monetary values illustrate that the data are highly skewed. A brief statistical description of RFM confirms this. Recency ranges from 1 to 374, with a mean of 93 and median of 51. Frequency has a wider range, varies from 1 to 203, with the mean of 4 and 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.

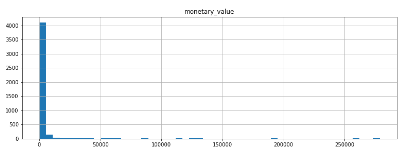




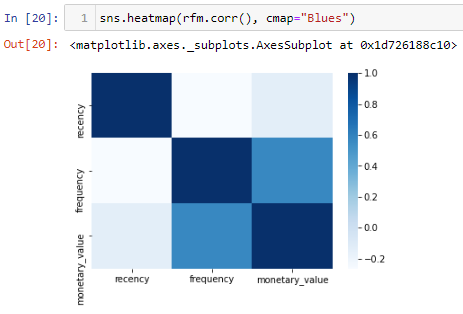
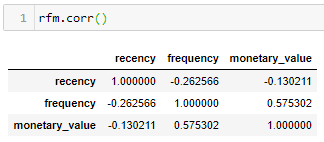
The following three histograms show the distribution of recency, frequency, and monetary values:



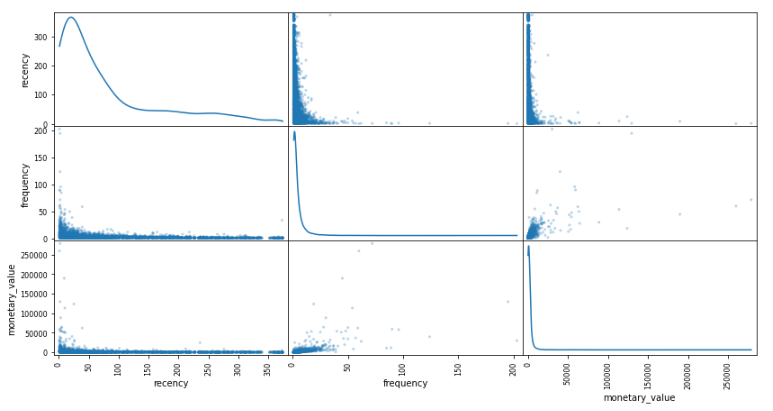




A further look into the correlation among R, F, and M found that frequency and monetary value has a positive correlation at 0.58. Recency has a negative correlation with frequency and monetary value, at a coefficient of -0.26 and -0.13 respectively.

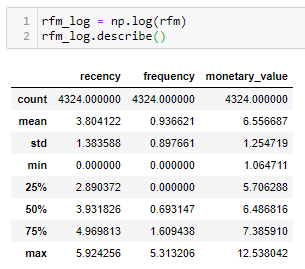
 

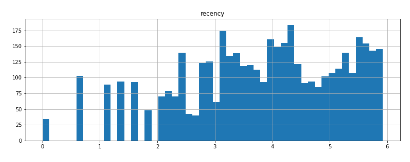
The scatter plots below provide a better illustration:

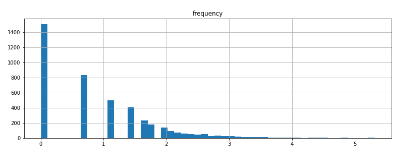


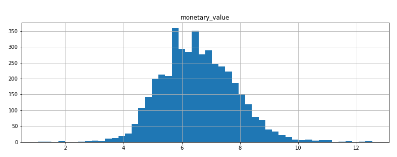
## Feature scaling

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:



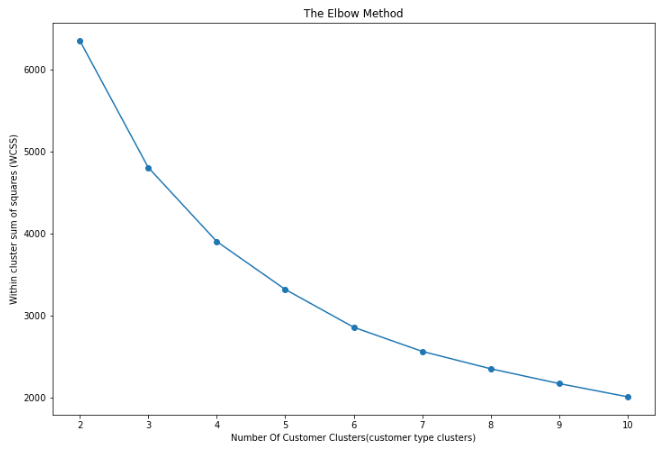




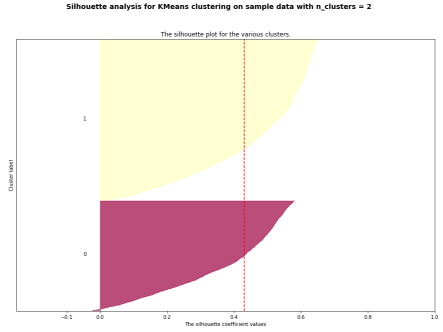
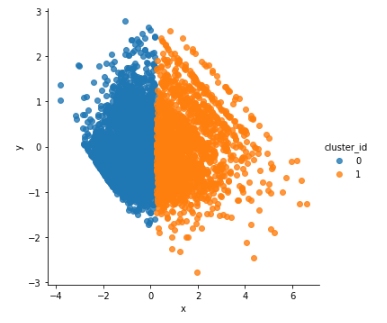


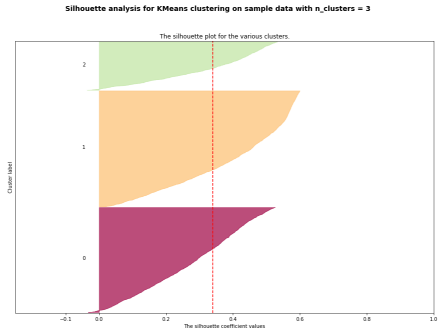
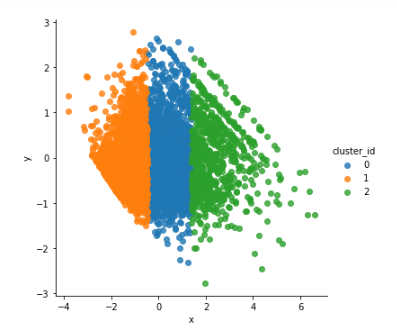
## Machine learning applications

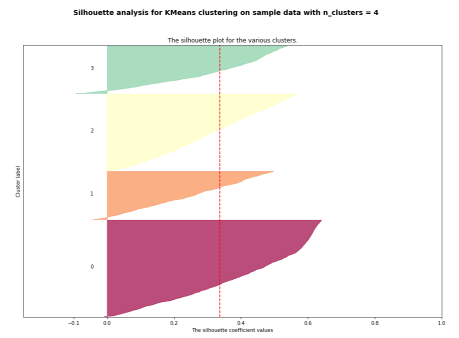
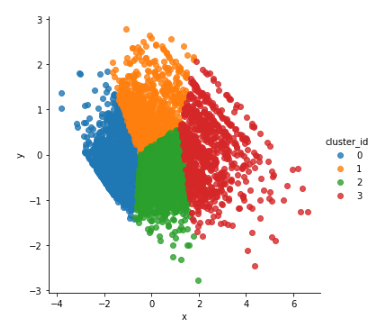
The log transformed RFM data was feed in a k-means algorithm for clustering. With a range of k from 2 to 11, a graph was plotted to help the selection of a suitable k value for machine learning.



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

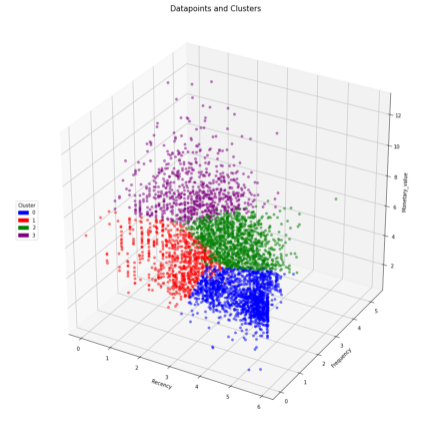
 

From the graphs, k = 4 is a plausible selection for the model. Thus, the customers are clustered into 4 groups. Each group contains customers that present similar characteristics in terms of recency, frequency and monetary values. A breakdown of clusters and customers show that the first cluster (cluster 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.

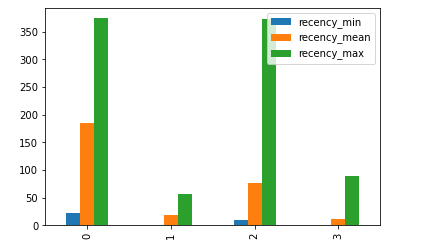


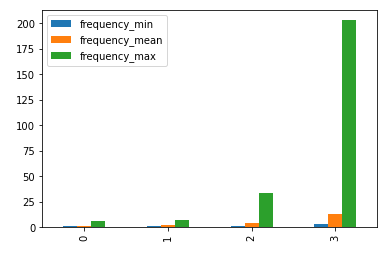
Then we go back to the RFM characteristics of different clusters. The first cluster of 1,551 customers has the longest recency, indicating that they are more likely old customers. They are the least frequency buyers, meaning that their purchasing activities are quite inactive. As a result, they contribute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

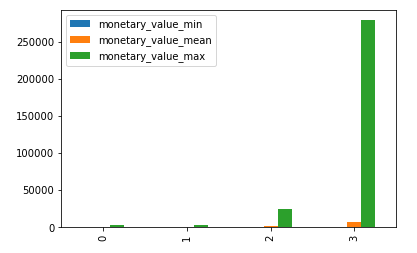
The second cluster of 775 customers has the shortest recency, indicating that their last purchases were recent. Their frequency value is low. As a result, they did not contribute much to the monetary values.

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.







# 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 strategy is 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. Main strategy is the reactivate them.
* The 775 customers in the second cluster are inactive, but their last purchase was recent. Main strategy is to activate them
* The 1,233 customers in the third cluster are old customers who used to be active but not anymore. Main strategy is to reactivate them.

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