Online Retail – RFM Analysis



Customer segmentation is the technique of dividing customers into groups based on their purchase patterns to identify who are the most profitable groups. In segmenting customers, various criteria can also be used depending on the market such as geographic, demographic characteristics or behavior bases, etc. There are so many techniques out there but I will use RFM analysis for my analysis.

What is RFM?

RFM is an acronym for recency, frequency, and monetary.

**Recency**is about when was the last order of a customer. It means the number of days since a customer made the last purchase. If it’s a case for a website or an app, this could be interpreted as the last visit day or the last login time.

**Frequency**is about the number of purchases in a given period. It could be 3 months, 6 months or 1 year. So we can understand this value as to how often or how many a customer used the product of a company. The bigger the value is, the more engaged the customers are. Could we say them as our VIP? Not necessary. Cause we also have to think about how much they actually paid for each purchase, which means monetary value.

**Monetary** is the total amount of money a customer spent in that given period. Therefore big spenders will be differentiated from other customers such as MVPs or VIPs.

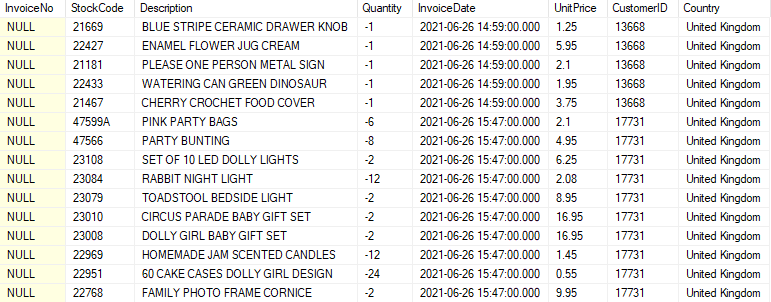
Business requirements

The Sales Manager wants me to segment customers into 4 levels: VIP, Normal, Low, and Extremely Low. The purpose of the Sales Manager when requesting me to do that because not only does he want to better understand customers but he also wants to give some attractive endowments to retain the customers based on their level.

**Data sources:** [Online Retail](https://www.kaggle.com/datasets/jihyeseo/online-retail-data-set-from-uci-ml-repo)

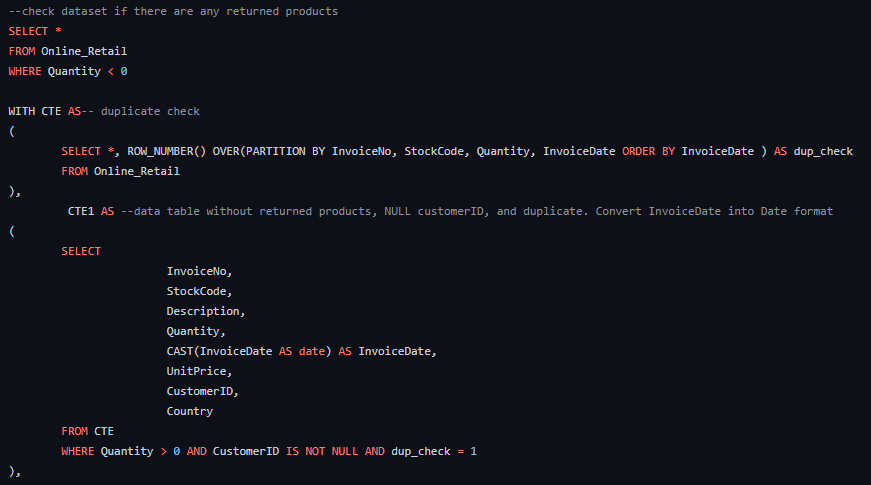
Data cleaning

First, I check data distribution in Kaggle and find that there are some missing values in the customerID column. Then, I write an easy query to check if there are any returned products. You can see that there are a lot of returned products in this dataset – the negative numbers are returned quantities.



In addition, the InvoiceDate is in Date-time format, in this analysis, I only want to focus on the date of the customer, not the time so I convert the date-time format of the InvoiceDate column into Date format. Finally, I also remove all duplicates in my dataset.

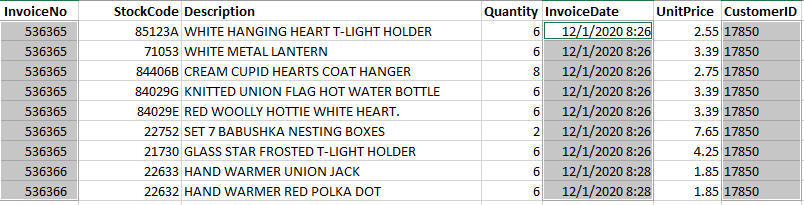
Here are all the queries that I use to clean the data



RFM analysis

**Recency**: To calculate the recency (the last time that Customers are active), I just only find the latest date that customers buy stuff by using MAX() function and subtracting it by 2021-12-10.

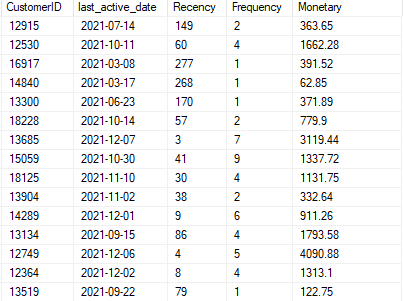
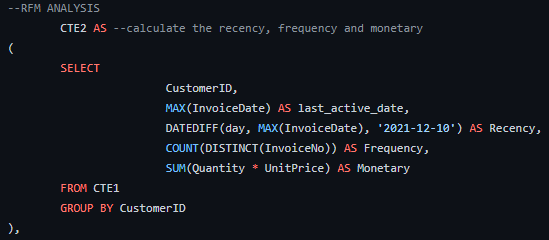
**Frequency**: When calculating frequency, at first I don’t realize that there are differences between the number of orders and the number of products that customers buy. After encountering errors, I gradually understand that to calculate the frequency we have to count the number of Orders not the number of Products because one order may have lots of products. We want to find out how many times customers buy stuff not how many products they have bought.



Look at the picture above, the customer who has CustomerID 17850 bought 2 times. One is in 8:26 and the last one is in 8:28 with the Invoice being respectively 536365 and 536366. One order may have a lot of products like there are lots of products in InvoiceNo 536365 but actually, they just buy stuff only 2 times.

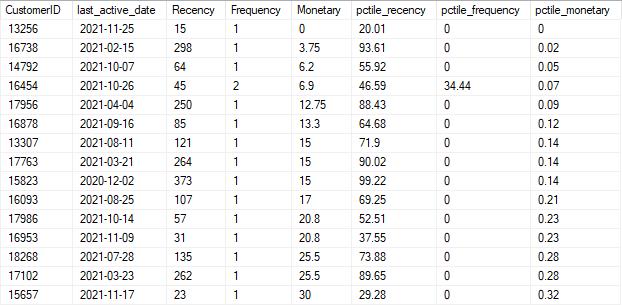
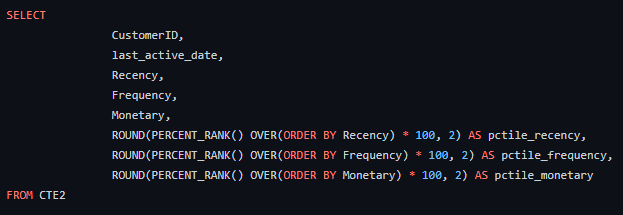
Therefore when I count the orders I have to combine the COUNT and DISTINCT functions.

**Monetary**: It is so easy to find out how much customers have spent by multiplying the quantity of each product by its price then we sum it up by each customer.



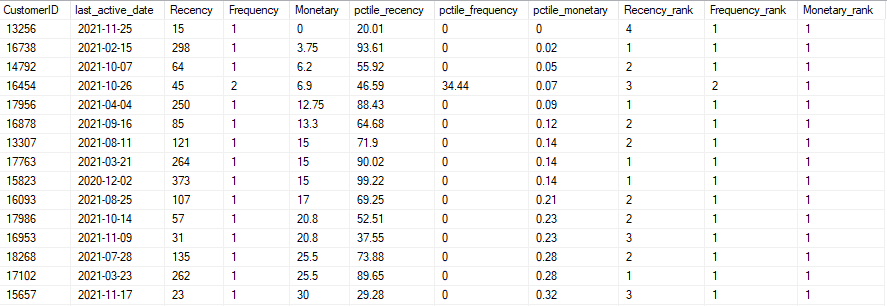
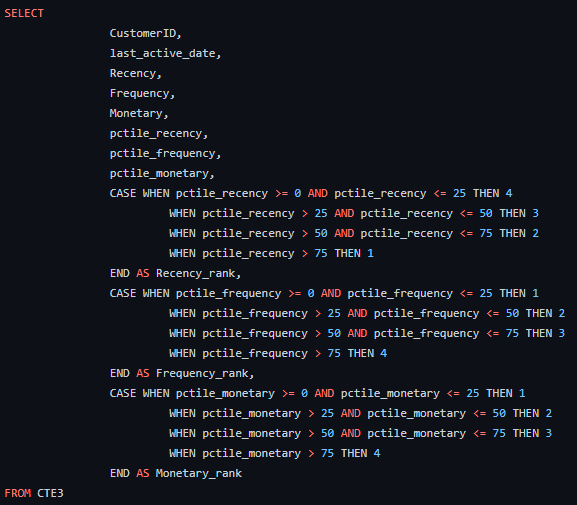
After having recency, frequency, and monetary, I apply the Quantile concept in statistics and respectively divide the values of each of them into 4 levels of Quantile ( 0-25th, 25th-50th, 50th-75th, and 75th-100th)

Before doing that, I have to calculate the percentile of each value by using the PERCENT\_RANK function



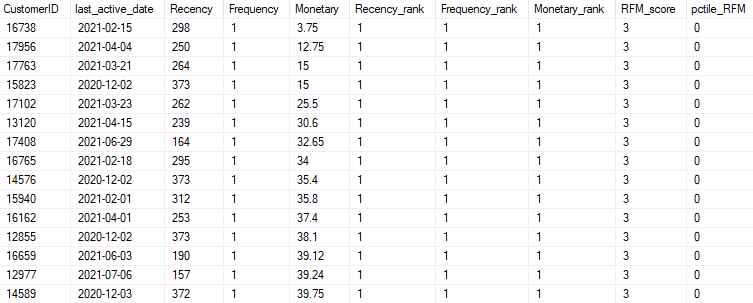
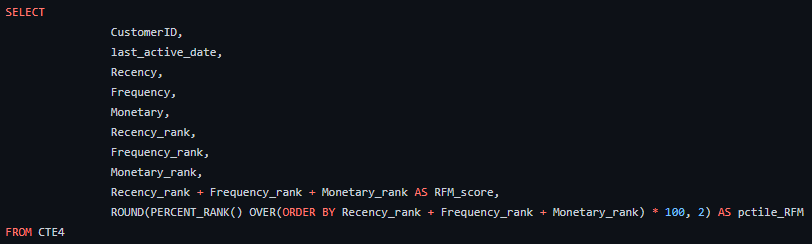
Now I will respectively divide the values of each recency, frequency, and monetary into 4 levels of Quantile ( 0-25th, 25th-50th, 50th-75th, and 75th-100th)

**Note**: I label 1, 2, 3, and 4 respectively for 0-25th, 25th-50th, 50th-75th, and 75th-100th for frequency and monetary. However, I only label 4, 3, 2, and 1 respectively for 0-25th, 25th-50th, 50th-75th, and 75th-100th for recency, because recency means how much time has elapsed since a customer’s last order. Therefore the smaller the value is, the more engaged a customer is with that brand.

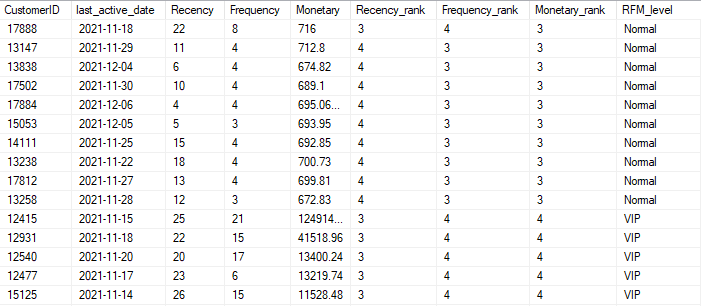
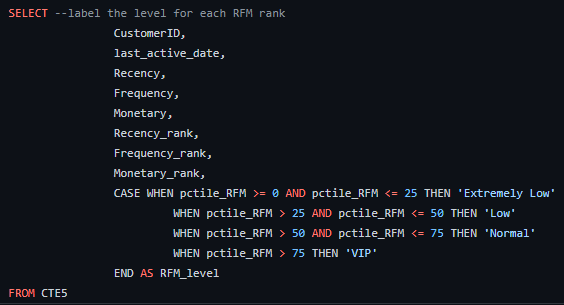


When I have done all of the steps above, it is time to actually segment the customer into 4 levels like the request of the Sales Manager: VIP, Normal, Low, and Extremely Low.

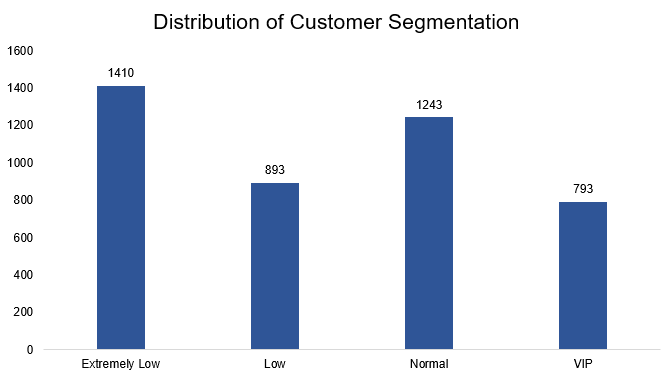
First, I will sum up the values of Recency\_rank, Frequency\_rank, and Monetary\_rank altogether and I attach those values to the RFM\_score column. Then, I continue applying the Quantile concepts.



Now, instead of labeling 1, 2, 3, and 4 respectively for 0-25th, 25th-50th, 50th-75th, and 75th-100th, I label VIP, Normal, Low, and Extremely Low respectively for 0-25th, 25th-50th, 50th-75th, and 75th-100th.



This is the final result of the RFM analysis, customers are finally divided into 4 levels. However, I found that the result is so tedious for my Sales Manager to follow, so I decided to create a visualization of the customer segmentation. Here is the visualization that I will show to my Manager.



As you can see, the Extremely Low group accounts for the majority of the Customer Segmentation levels. Therefore, we need to create more attractive discounts or events to increase the conversion rate of those customers from the Extremely Low group to higher levels like Normal or VIP. In addition, the VIP group has the least members so the Sales Manager has to construct some strategies to retain those customers because those customers in the VIP group can contribute huge money to our company.