Online Retail Customer Segmentation and Behavioral Analytics

Springboard Data Science Career Track Capstone 3 Final Report

Yuan Yin

08/21/2021

INTRODUCTION[1]

The most successful companies today are the ones that know their customers so well that they can anticipate their needs. There are many different ways to provide business insights for companies, such as cohort analysis, CLV(Customer Lifetime Value) calculation and prediction, churn prediction, customer segmentation, and so on.

In this project, we will focus on customer segmentation and customer behavior analytics, which includes customer monthly purchasing analysis, cohort analysis, and historical CLV calculation.

Customer segmentation is the practice of dividing customers into groups based on the similarity of characteristics. Each group or segment is related to a significant customer profile so companies can design targeted marketing campaigns.

Customer segmentation is a type of clustering task in the data science field, and the best-known Machine Learning algorithm for clustering is KMeans. K-means discovers segments well only when the features are symmetrical and have the same mean values and standard deviation. Besides, KMeans requires predefining the number of clusters, i.e., k.

On top of all, whether the segments are meaningful and the customer profiles can bring business insights is crucial for customer segmentation.

- **Problem Statement**

This project is built based on a transactional dataset 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.

In addition to analyzing customer behavior, this project aims to identify a number of customer profiles, each of which is related to a significant customer segment, such that the company can market to different customers more effectively and appropriately.

DATA WRANGLING

Data Wrangling

- **Datasets**
  - The on-line retail dataset contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and non-store gift company
  - The attributes include:
    - **InvoiceNo**: Invoice number, a 6-digit integral number uniquely assigned to each transaction
    - **StockCode**: Product (item) code, a 5-digit integral number uniquely assigned to each distinct product
    - **Description**: Product (item) name
    - **Quantity**: The quantities of each product (item) per transaction.
    - **InvoiceDate**: Invoice date and time, the day and time when a transaction was generated.
    - **UnitPrice**: Unit price, product price per unit in sterling.
    - **Customer ID**: Customer number, a 5-digit integral number uniquely assigned to each customer.
    - **Country**: Country name, the name of the country where a customer resides.
- **Main Steps**
  - Removed all transactions outside the UK, having no product description, having no customer id, happening after 12/01/2011, and the transactions, of which the quantity and price are negative or zero
  - Converted date-and-time InvoiceDate attribute to date-only attribute
  - Explored the products that have the same or similar stock codes and different descriptions(product name) or have the same description but different stack codes

CUSTOMER BEHAVIOUR ANALYSIS

Customer Behaviour Analysis

- **Monthly/Seasonally Transaction Analysis**

For an on-line retail, monthly revenue can be considered as a critical metric to evaluate the company's performance. Figure 1 shows the quantity of products sold and the revenue by month, which are significantly affected by Thanksgiving and Christmas. It's reasonable that the transactions reached the highest point in November rather than December, because many customers of the company are wholesalers, who need to purchase products before not during the holidays.

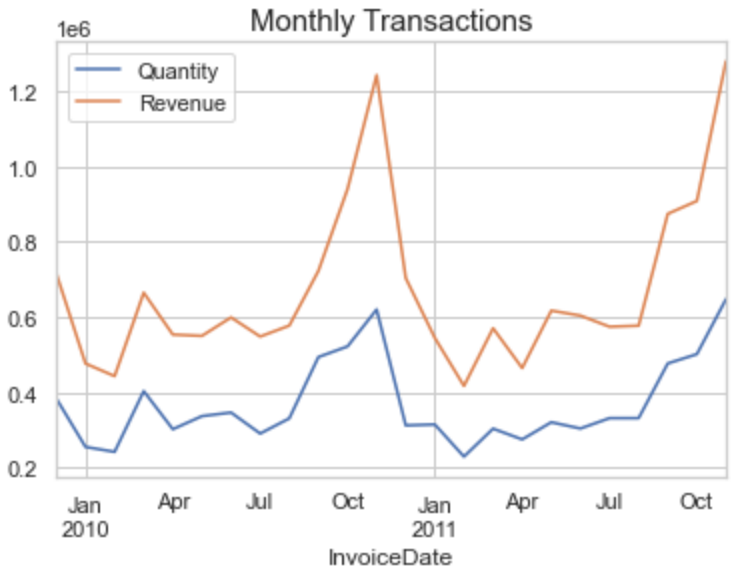

Figure 1

We can confirm the transaction pattern by calculating seasonal quantity and revenue [Figure 2]. Here, we define four seasons: December, January, and February; March, April, and May; June, July, and August; September, October, and November.

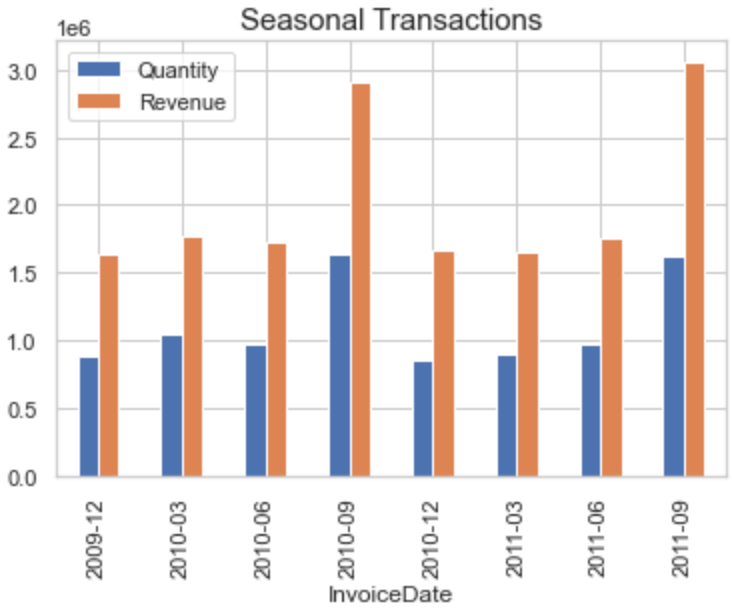

Figure 2

It's obvious that the pre-holiday season, i.e., September, October and November, has much better performance than the other three seasons.

- **Monthly/Seasonally Active Customer Analysis**

We analyse the number of customers in the same way and get similar results[Figure 3.a][Figure 3.b].

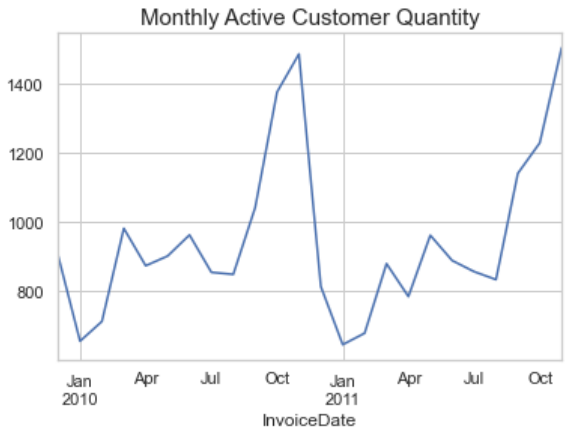

Figure 3.a

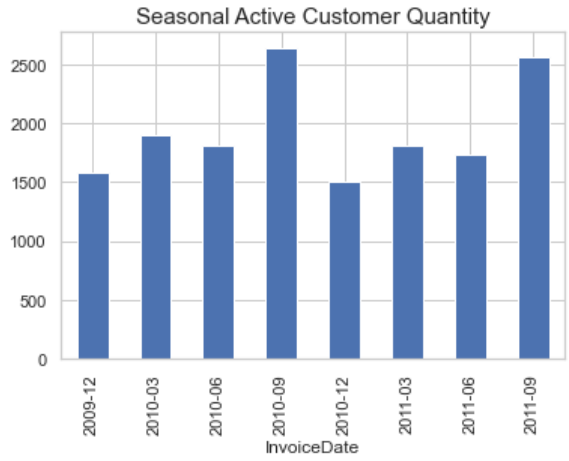

Figure 3.b

As we can see, the active customers in the pre-holiday season are much more than the other season. This pattern brings a signal of whether a customer is active or not: if a customer did not buy anything in this peak season, we may consider them as churn.

- **Time Cohort Analysis**

Time cohorts are customers who signed up for a product or service during a particular time frame. Analyzing these cohorts shows the customers' behavior depending on the time they started using the company's products or services. There are other types of cohort , such as behavior cohort and size cohort. For this object, we will focus on time cohort analysis by grouping customers into cohorts based on the month of their first purchase.

We can calculate monthly active customers from each cohort to check customer's retention rate, i.e., the ratio of how many of these customers came back in the subsequent months [Figure 4]. Here, Cohort Index 1 represents the month when the customers in the cohort purchase for the first time; 2 represents the following month, and so on. So as we can expect, the retention rate of the first month is 100%.

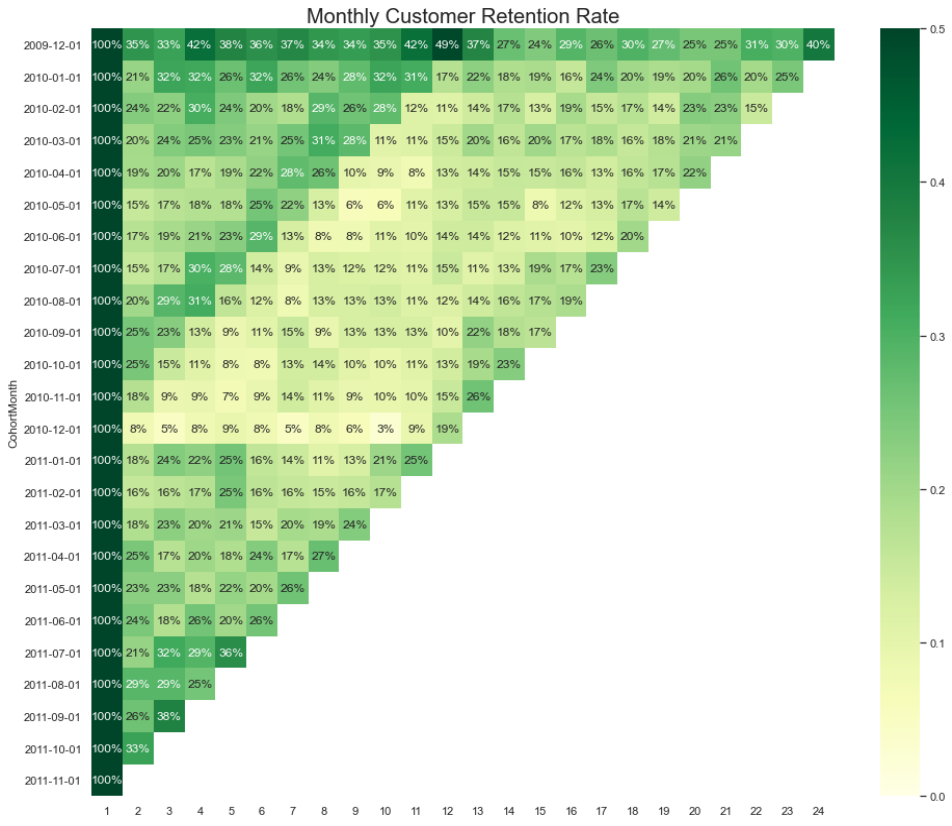

Figure 4

Note: The statistics of the first several months of the dataset, especially December 2009, can not present the real world because of the absence of the previous months. According to the heatmap, the retention rates are relatively higher in the pre-holiday season (the darker squares) than the other periods (the lighter squares).

We can also use heatmap to display the average monthly revenue for each cohort [Figure 5].

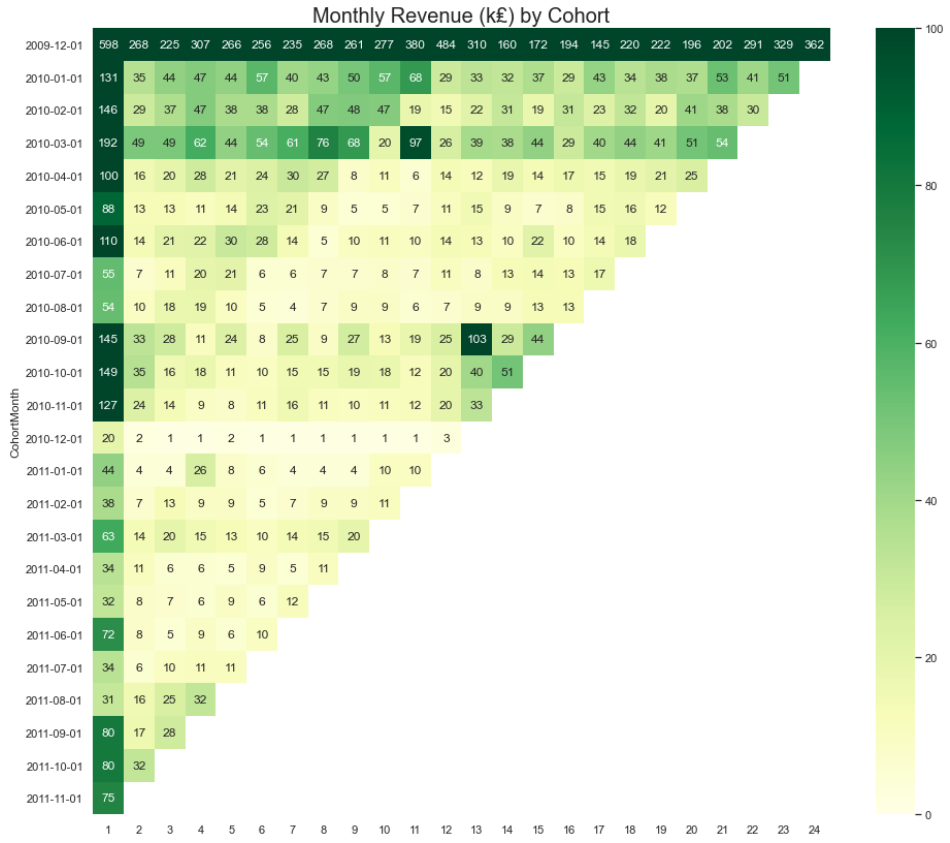

Figure 5

Revenue heatmap has a similar pattern with retention rate heatmap. The revenue from each cohort is higher in pre-holidays season and, of course, the acquisition month. In the same way, we need to ignore the first few months since the revenue of these months were generated by all the previous customers which are not included in our dataset.

- **Customer Lifetime Value(CLV)**

The customer lifetime value is a measurement of how much a company expects to earn from an average customer in a lifetime. It can be historical, and we also can predict the customer lifetime value. In this project, we calculate historical CLV by summing up each customer's revenue instead of profit since we have no cost data.

There are different ways to calculate historical CLV, which is based on the company type, customer status, and the goal of calculating CLV. Here, we focus on traditional CLV calculation.

$$Traditional\ CLV\ = Average\ Monthly\ Revenue\ \times \frac{Average\ Monthly\ Retention\ Rate}{1 -\ Average\ Monthly\ Retention\ Rate}$$

The traditional CLV is one of the most popular descriptive customer lifetime value techniques. It incorporates retention and churn rates, which is defined as 1 minus retention. The retention to churn ratio can act as a proxy to expected length of the customer lifespan with the company. So here, we don't need to define customer lifespan by ourselves.

It is worth our attention that the traditional CLV assumes the churn is final. In other words, when we calculate traditional CLV based on average monthly revenue, we assume that the customers who don't come back the next month, are not coming back in the later periods. But it's not always the case in the real world, especially in non-contract business like our object. As a result, the CLV of this online retail company is only about £153, which is much underestimated.

DATA PRE-PROCESSING

Data Pre-processing

As we mentioned above, for this project, we will use RFM (Refrency, Frequency, and Monetary Value) and RRMT (RFM plus Tenure) as metrics respectively. All these features are numeric, which are accepted by KMeans algorithm. But KMeans works well only when the features are symmetrical and have the same mean values and standard deviation.

As we can see, all the features are skewed and not on the comparable scales [Figure 6]. Here, we unskew data with box-cox method and scale data with StandardScaler.

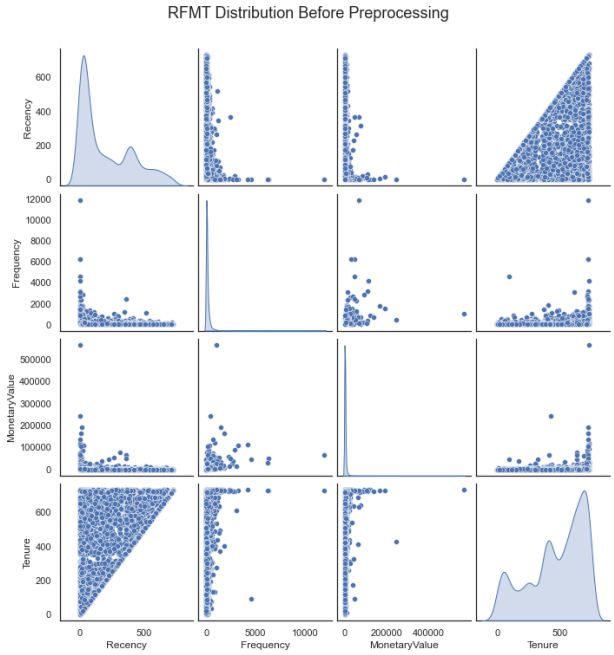

Figure 6

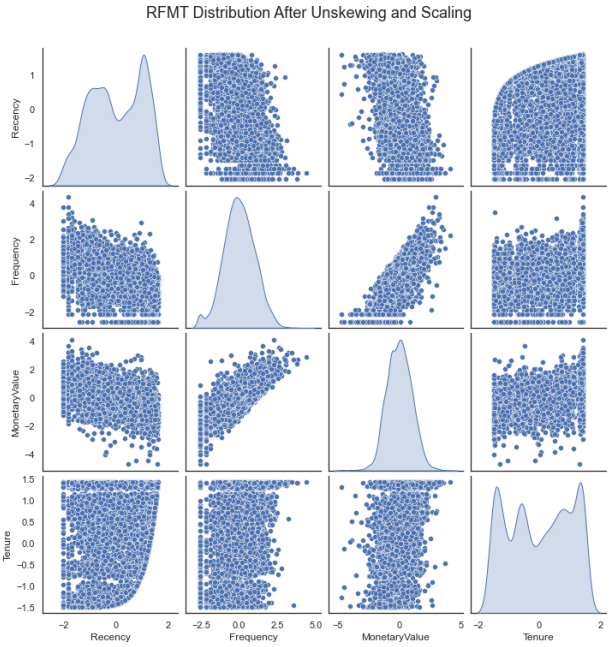

Figure 7

After pre-processing, Recency and Tenure look not perfect but much better than before [Figure 7].

CUSTOMER SEGMENTATION AND PROFILING

Customer Segmentation and Profiling

- **Metrics for Segmentation**

For customer segmentation, we can select a number of attributes to describe customers. Some typical examples include demographic information, location information, transactional information such as the category and quantity of the products that the customers have purchased and how much revenue generated from the transactions, the membership status, if they are active, how long have they been customers, etc.

In this project, we try to identify customer clusters based on two slightly different sets of metrics, i.e., Recency, Frequency, and Monetary Value (RFM) and Recency, Monetary

Value, and Tenure(RMT).

- ➔ **Recency**: measures how recent was each customer's last purchase
- ➔ **Frequency**: measures how many purchases the customer has done in the past period of time
- ➔ **Monetary Value**: measures how much has the customer spent in the past period of time
- ➔ **Tenure**: measures how long the customer has been with the company since their first transaction.

RFM are very popular metrics to define a customer's profile, but for this online retail dataset, Frequency and Monetary Value are strongly correlated after un-skewing. In other words, Frequency can not provide more additional information to identify customer segments if we do clustering based on customer's Monetary Value using KMeans algorithm. So after trying 3-segments, 4-segments, and 5-segments models with RFM metrics, we update the metrics by dropping Frequency and importing Tenure to figure out the most meaningful and insightful customer segmentation model.

- **Selecting Optimal Number of Clusters**

KMeans algorithm requires predefining the number of clusters, and there are different ways to optimize the number of clusters. Here, we select the optimal number of clusters with the elbow criterion method, which plots the sum of squared errors for each clustering. The elbow-point, 4, represents the optimal number, and we build models around this number [Figure 8].

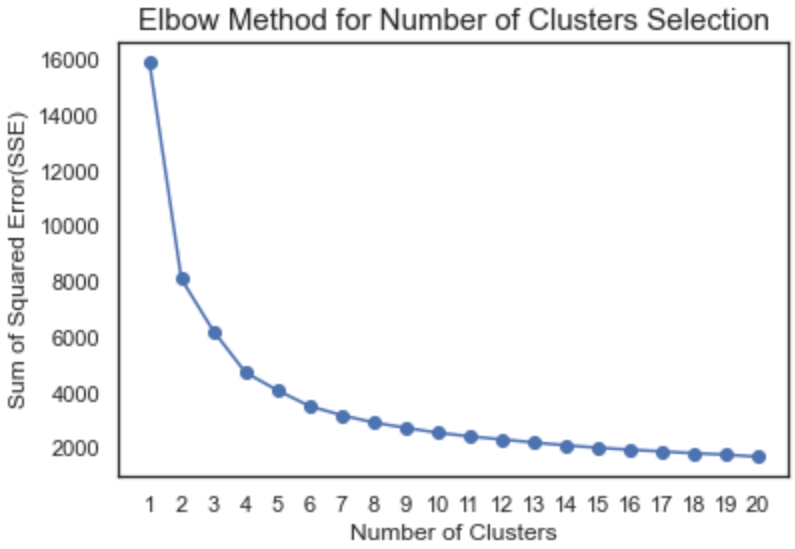

Figure 8

- **Approaches to Explore and Identify Customer Profile**

There are multiple ways to explore and identify customer persona, with statistics or through visualization. The approaches applied in this project include:

- ➔ **Summary Statistics**: mean, median, max, min, standard deviation
- ➔ **Relative Importance**: median values of metrics(RFM or RMT) for each cluster divided by the median values for total population. **Note**: From a company's standpoint, the more revenue a customer generates (higher Monetary Value), the more frequently a customer purchases(higher Frequency), and more recently a customer is active(Lower Recency), the better. We visualize the result of relative importance with heatmap, so we take the reciprocal of Recency rather than Recency for a more intuitive display.
- ➔ **Snake Plot**: plots segments and their RFM/RMT values on a line chart.
- ➔ **2D kde Plot**: displays the distribution of each segments and the relationship between the segments on each metric in the kde (Kernel Density Estimators) style
- ➔ **3D Scatter Plot**: displays each customer's position based on his/her RFM or RMT values

Note: The scales of each metric are significantly different, so we scale all the metrics to make them comparable within the same charts, including snake plot, 2D kde plot and 3D scatter plot.

- **Basic Segmentation: 3-Segments RFM/RMT Model**
    - **Summary Statistics** [Table 1] [Table 2] and **Segment Size Ratio** [Figure 9] [Figure 10]

	Recency					Frequency				
	mean	median	min	max	std	mean	median	min	max	std
Segment										
0	144.0	77	1	721	149.0	69.0	57.0	2	397	50.0
1	397.0	402	2	730	189.0	19.0	15.0	1	110	15.0
2	39.0	20	1	521	59.0	365.0	242.0	24	11952	526.0

	MonetaryValue					Size
	mean	median	min	max	std	
Segment						
0	1186.0	920.0	96.0	44534.0	1583.0	2173
1	304.0	242.0	3.0	3424.0	255.0	1768
2	7766.0	3896.0	479.0	564101.0	21022.0	1372

Table 1. Statistics of RFM 3-Segment Model

	Recency					MonetaryValue			
	mean	median	min	max	std	mean	median	min	max
Segment									
0	425.0	409	28	730	151.0	584.0	385.0	3.0	6117.0
1	68.0	42	1	308	69.0	878.0	606.0	6.0	8594.0
2	61.0	31	1	683	83.0	6823.0	3316.0	230.0	564101.0

	Tenure					Size
	std	mean	median	min	max	std	
Segment							
0	580.0	546.0	564.0	197	730	127.0	2014
1	923.0	211.0	194.0	1	616	152.0	1668
2	19462.0	649.0	672.0	93	730	86.0	1631

Table 2. Statistics of RMT 3-Segment Model

The statistics of metrics provides us with important but not intuitive information about the segmentation.

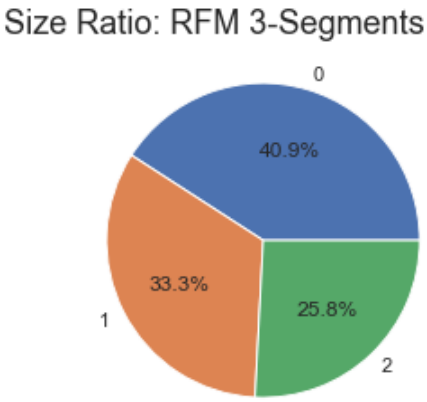

Figure 9

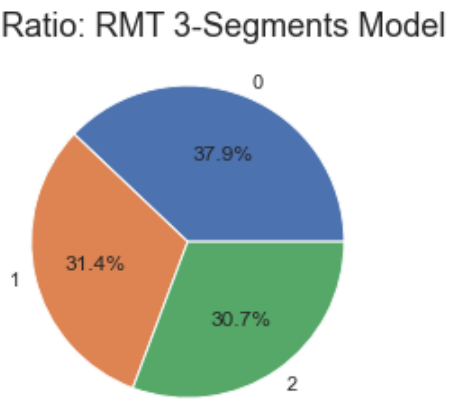

Figure 10

Generally, we want our segments to differ from the overall population, and have distinctive properties of their own, so we can use Relative Importance and Snake Plot to identify each attribute.

- **Relative Importance**[Figure 11] [Figure 12] and **Snake Plot**[Figure 13] [Figure 14]

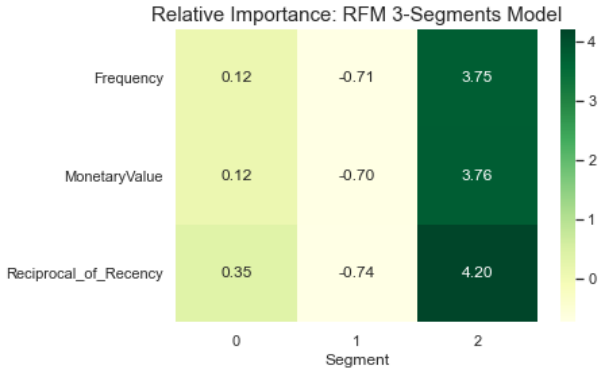

Figure 11

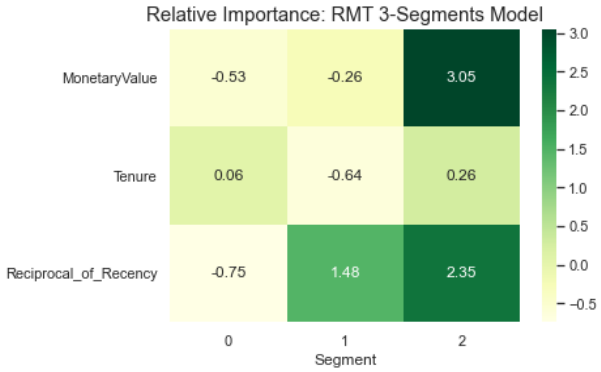

Figure 12

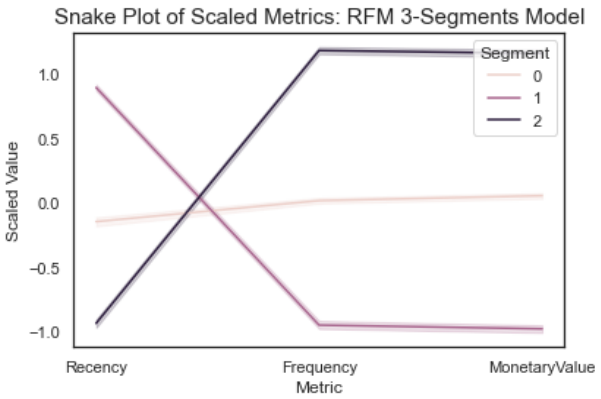

Figure 13

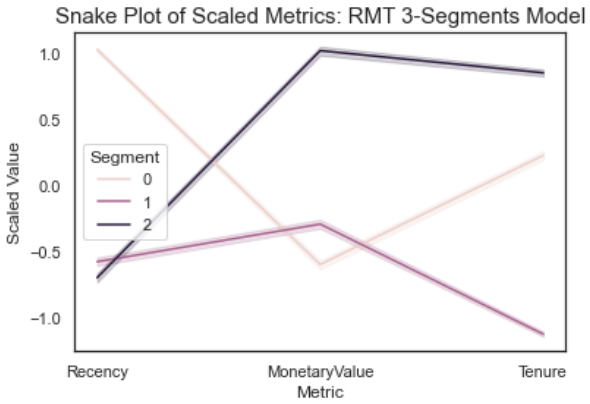

Figure 14

There are big differences between the two 3-Segment models. RFM model looks simple and easy to distinguish different clusters based on the median value of RFM. For the clusters of RMT model, there are only two parts on Recency and Monetary Value: High -value and Low-value, Active and Inactive.

The shortcoming of the Relative Importance and Snake Plot is they only reflect the overall performance of a segment. We can get to know more about the distribution of each metric within a group and the relationship between the groups by 2D kde plot and 3D scatter plot.

- **2D KDE Plot** [Figure 15] [Figure 16]

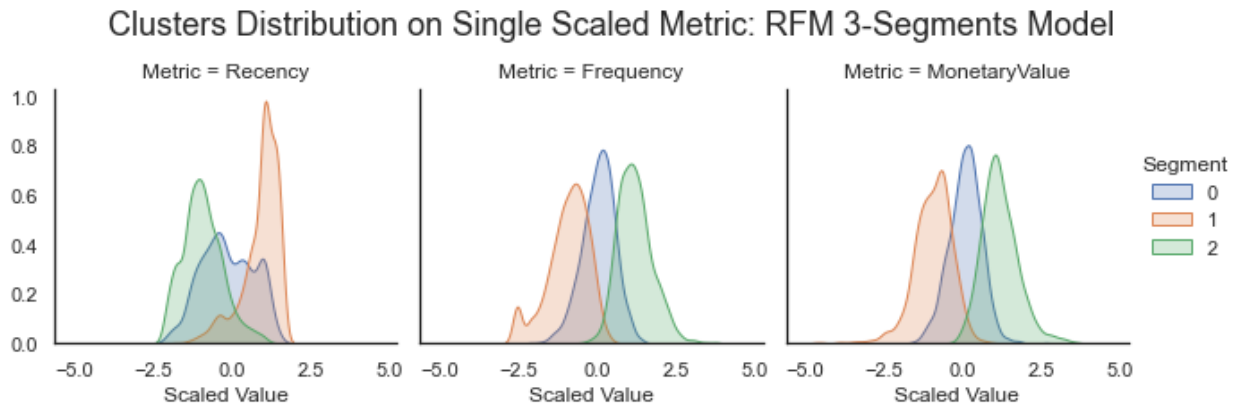

Figure 15

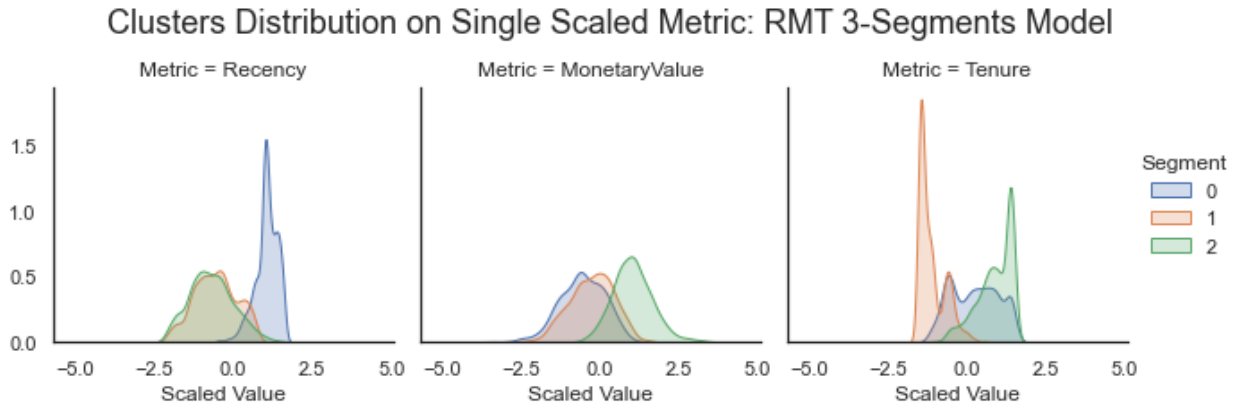

Figure 16

As we can see, for the RFM model, Segment 0 has a big range over Recency, and the Segment 0 in the RMT spreads evenly over the Tenure scope, which means we can not distinguish such segments from the others well in the corresponding dimension.

- **3D Scatter Plot** [Figure 17] [Figure 18]

According to the performance of the segmentation, we label each cluster to identify the customer persona.

For the RFM model [Figure 17]:

- ➔ Segment 0: not very high or low in Frequency and Monetary Value, having a big range on Recency, labeled as **Medium**.
- ➔ Segment 1: low in Frequency and Monetary Value, high in Recency, labeled as **Inactive_Rare_Low**
- ➔ Segment 2: high in Frequency and Monetary Value, low in Recency, labeled as **Active_Frequent_High**.

For the RMT model [Figure 18]:

- ➔ Segment 0: low in Recency and Monetary Value, having a big range on Tenure, labeled as **Inactive_Low**
- ➔ Segment 1: low in Recency, Monetary Value and Tenure, labeled as **Active_Low_New**
- ➔ Segment 2: low in Recency and Monetary Value, high in Tenure, labeled as **Active_High_Old**

We can display the single customer in each segment with the labels by scatter plot.

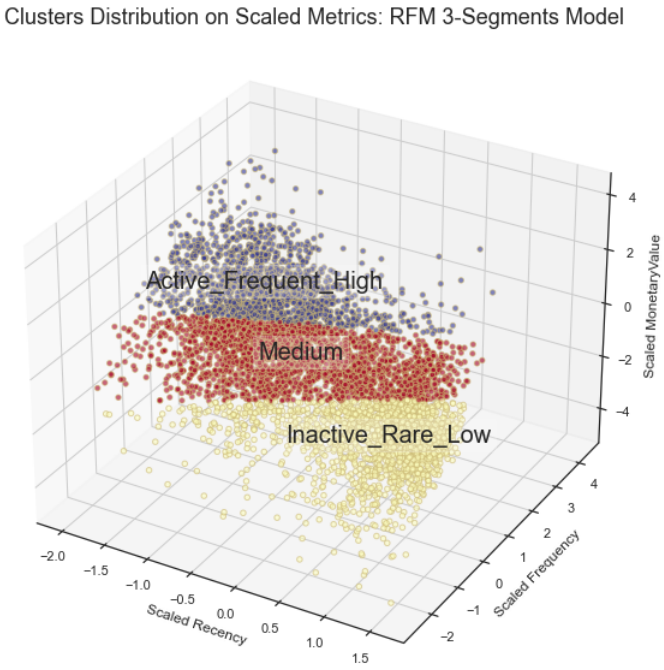

Figure 17

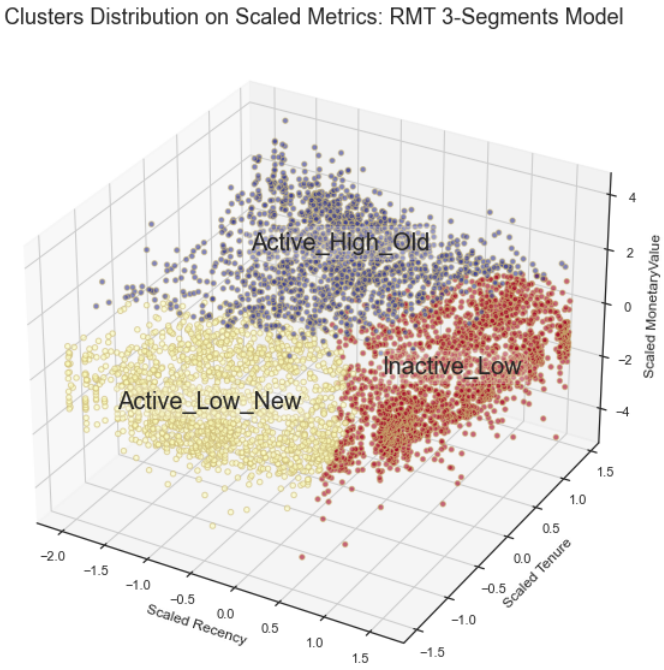

Figure 18

- **5-Segments RFM Model v.s. 6-Segments RMT Model**

The segments in both 3-segment models can be distinguished from each other based on the overall performance. But they are too simple to bring more detailed business insights. In addition to the 3-segments models, we also build 4-Segments RFM/RMT model, 5-segment RFM/RMT model and 6-segment RMT model, and compare each segment profile. As a result, 5-segments RFM model and 6-segments RMT model look more meaningful and the customer profiles can bring practical business insights.

- **Improved Segmentation: 5-Segments RFM Model/6-Segments RFM Model**
    - **Summary Statistics** [Table 3] [Table 4] and **Segment Size Ratio** [Figure 19] [Figure 20]

	Recency					Frequency				
	mean	median	min	max	std	mean	median	min	max	std
Segment										
0	39.0	29	1	157	32.0	103.0	92.0	6	376	58.0
1	439.0	424	48	730	167.0	14.0	12.0	1	57	10.0
2	47.0	42	1	165	35.0	22.0	20.0	1	91	14.0
3	36.0	17	1	521	55.0	478.0	334.0	33	11952	622.0
4	334.0	354	90	730	145.0	75.0	55.0	3	440	62.0

	MonetaryValue					Size
	mean	median	min	max	std	
Segment						
0	1705.0	1403.0	239.0	16246.0	1164.0	1216
1	232.0	197.0	3.0	2803.0	178.0	1164
2	404.0	341.0	21.0	3166.0	297.0	710
3	10545.0	5572.0	1323.0	564101.0	25503.0	892
4	1303.0	849.0	173.0	77347.0	2876.0	1331

Table 3 Statistics of RFM 5-Segment Model

	Recency					MonetaryValue			
	mean	median	min	max	std	mean	median	min	max
Segment									
0	315.0	358	37	576	111.0	528.0	432.0	96.0	4512.0
1	31.0	18	1	521	41.0	9286.0	4600.0	462.0	564101.0
2	33.0	23	1	227	32.0	1604.0	1126.0	211.0	44534.0
3	90.0	58	1	326	83.0	262.0	212.0	3.0	1862.0
4	259.0	231	6	730	172.0	1787.0	1276.0	96.0	77347.0
5	538.0	557	42	730	135.0	236.0	195.0	3.0	1437.0

	Tenure					Size	
	std	mean	median	min	max	std	
Segment							
0	387.0	388.0	401	105	587	97.0	857
1	24047.0	663.0	686	93	730	79.0	1020
2	2354.0	285.0	297	1	659	170.0	998
3	196.0	133.0	84	1	639	122.0	628
4	2778.0	651.0	658	403	730	64.0	1066
5	169.0	591.0	603	350	730	98.0	744

Table 4 Statistics of RMT 6-Segments Model

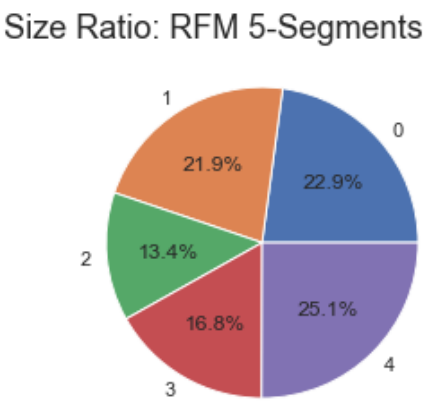

Figure 19

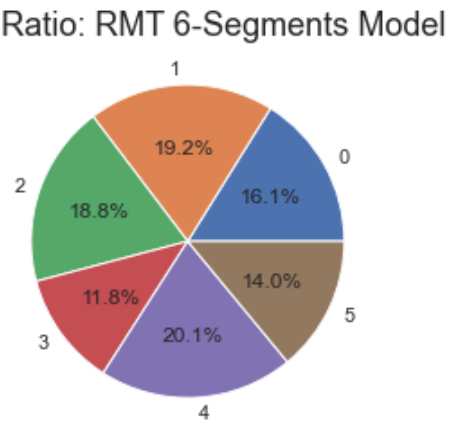

Figure 20

- **Relative Importance**[Figure 21] [Figure 22] and **Snake Plot**[Figure 23] [Figure 24]

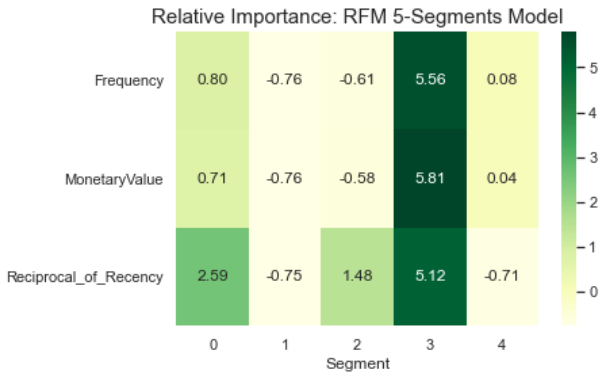

Figure 21

Figure 22

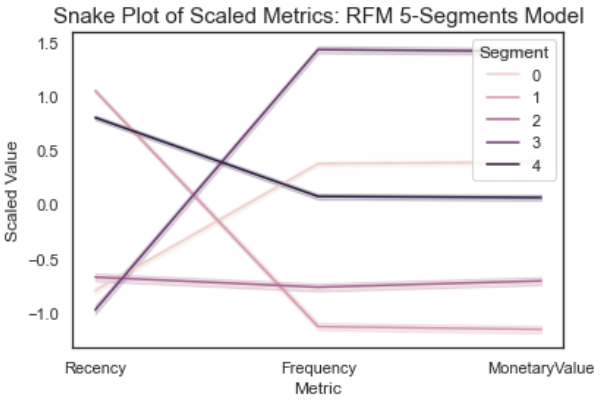

Figure 23

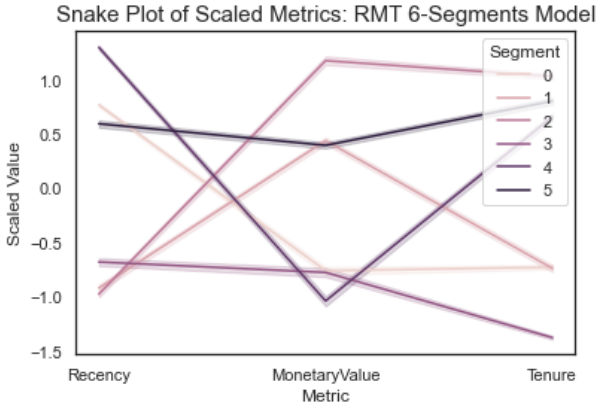

Figure 24

The RFM 5-segments model can significantly separate the customers into active and inactive parts based on Recency, which is much better than the RFM 3-segments model. However, from the Relative Importance heatmap and the snake plot of the RFM model, it's easy to find that the preprocessed Frequency and Monetary Value change simultaneously. High Monetary Value means high frequency, Low Monetary Value also means low Frequency, and vice versa. The correlation coefficient of the two features is 0.81, which means unskewed Frequency and Monetary Value are strongly correlated. So compared to the RFM model, the RMT model provides more information about the customers.

The RMT 6-segments mode divides the customer into active/inactive parts based on Recency and high/medium/low value parts based on Monetary Value. Besides, the RMT 6-segment model also separates new customers from old customers.

- 2D KDE Plot [Figure 25] [Figure 26]

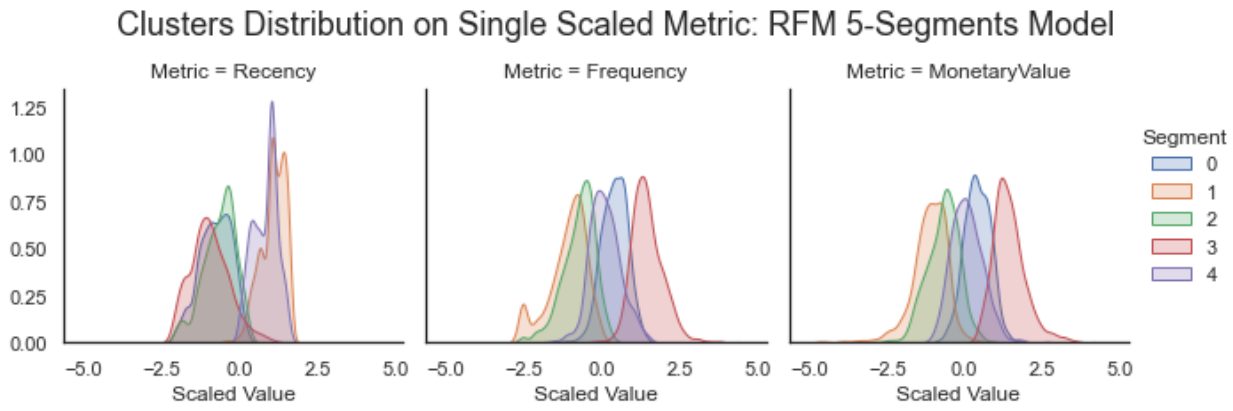

Figure 25

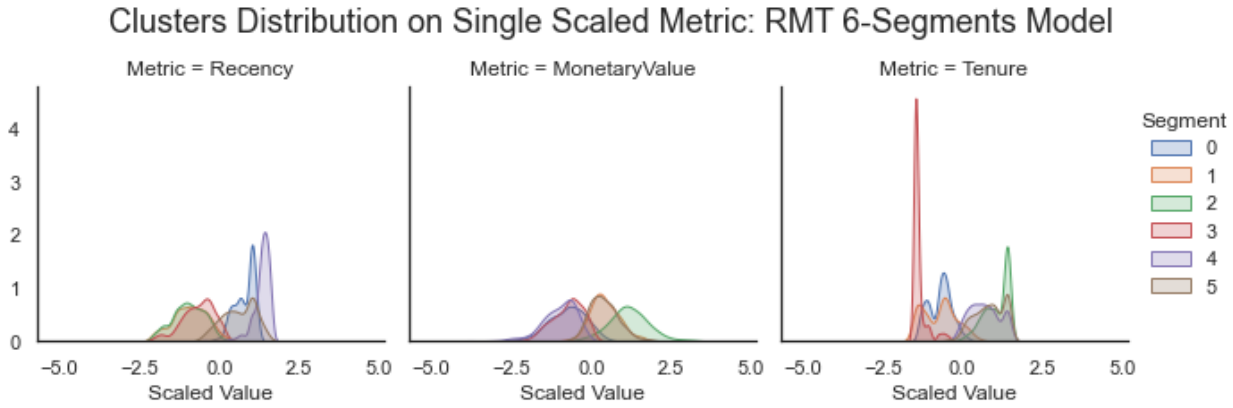

Figure 26

For the RFM 5-segment model, the distributions of Frequency and Monetary Value appear to be staggered based on KDE plot[Figure 25], but if we take a look at the scatter plot[Figure 27] we will find that the Monetary Value boundaries between the segments within active and inactive are pretty significant.

RFM 6-segment model has relatively clear boundary between active and inactive clusters. According to the summary statistics, the median values of Recency of all the active clusters are less than 40 days and the values of standard deviation are about 30 days, which means most of these active customers purchase in the last pre-holiday season (last 92 days). The median values of recency of all the inactive clusters are more than 8 months.

- 3D Scatter Plot [Figure 27][Figure 28]

Based on the statistics and the plots, we can label the RFM 5-segments model and RMT 6-segments model as follows.

For the RFM 5-segments model [Figure 27]:

- ➔ Segment 0: low in Recency, medium in Frequency and Monetary Value, labeled as **Active_Medium**.
- ➔ Segment 1: high in Recency, low in Frequency and Monetary Value, labeled as **Inactive_Rare_Low**
- ➔ Segment 2: low in Recency, Frequency and Monetary Value, labeled as **Active_Rare_Low**
- ➔ Segment 3: low in Recency, high in Frequency and Monetary Value, labeled as **Active_Frequent_High**
- ➔ Segment 4: high in Recency, medium in Frequency and Monetary Value, labeled as **Inactive_Medium**.

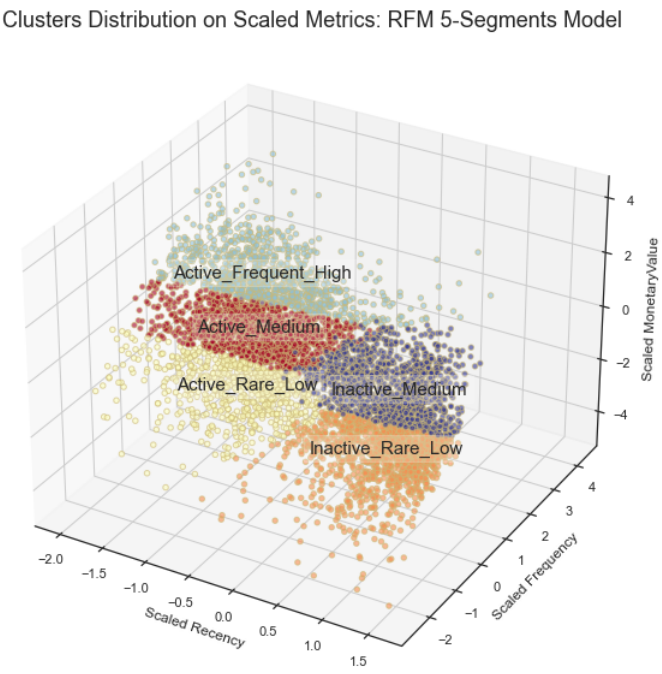

Figure 27

For the RMT 6-segments model [Figure 28]:

- ➔ Segment 0: **One_Time**. The median values of both Recency and Tenure are around one year, which means most of the customers within this segment purchased products about one year ago, for the first and last time.
- ➔ Segment 1: **Potential_Valuable**. This segment is worthy to highlight since the customers within this segment are relatively new, compared to the high-value customers, but the revenue generated by these customers are also considerable. Besides, this segment is pretty active.
- ➔ Segment 2: **High_End**. The customers within this segment are most important for the company, who are active, loyal, and have generated much revenue.
- ➔ Segment 3: **New**. The customers within this segment are new to the company. The median value of Tenure is only 64 days, which is much lower than the other segments
- ➔ Segment 4: **Low_and_Lost**. It seems not worthy to pay close attention to this segment. The customers in this segment generated very low revenue since most of them have not purchased anything for more than one year and a half.
- ➔ Segment 5: **Once_Valuable_Potential_Lost**. Different from the Low-and-Lost segment, although most of the customers within this segment did not purchase recently (the median value of Recency is 249 days) but they have generated pretty high revenue. So the company may take special actions to prevent these customers from being really lost.

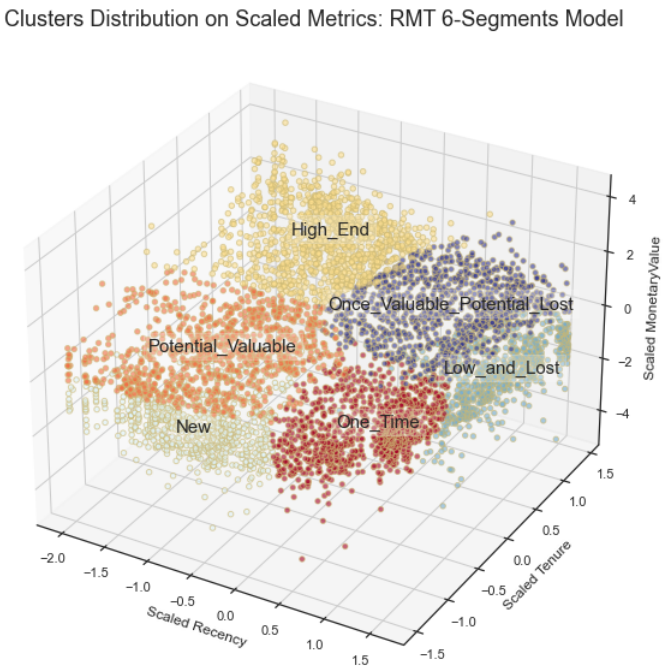

Figure 28

CONCLUSION

- This project includes two parts: customer behaviour analysis and customer segmentation, both of which are based on the transaction dataset of a online retail from 12/01/2019 to 11/30/2011

- In customer behaviour analysis part, we completed:
    - Monthly/seasonal transaction and customer analysis, including sales volume, revenue and the number of active customers. **There is a significant pattern of pre-holiday season, i.e., September, October and November, during which the quantity of products sold, the revenue and the number of active customers are much higher than the other months.**

    - Time cohort analysis, including monthly revenue by cohort and monthly retention rate by cohort. The result shows that **the monthly revenue and retention rates by cohort are both relatively higher in the pre-holiday season than the other periods**.

- Traditional Customer Lifetime Value(CLV) calculation, which is based on the average monthly retention rate. The traditional CLV calculation assumes the churn, i.e., one minus retention rate, is final. but it's not always the case in the real world, especially in non-contract business like retail. As a result, **the CLV of this online retail company is only about £153, which is much underestimated**.

- We clustered customer with KMeans algorithm and defined significant customer profiles
  - **Metrics** selection is very important for customer segmentation. In this project, we use **Recency, Frequency, and Monetary Value(RFM)**, a well-known set metrics to measure customer's performance firstly, and then update to **Recency, Monetary Value, and Tenure(RMT)** to build more meaningful customer segmentation
  - We unskewed and scaled data to meet KMeans algorithm's requirement and predefined the number of clusters with elbow method.
  - We compared the customer profiles, which are identified by different models, in different ways, including summary statistics, relative importance, snake plot, 2D kde plot, and 3D scatter plot.
  - **The best model is the 6-segments model based on Recency, Monetary Value and Tenure**, which clusters more than 5k customers into 6 groups：High-end customer, Potential-valuable customer, New customer, One-time customer, Once-valuable-potential-lost customer, low-and-lost customer. Compared with other solutions, the 6-segments RMT is more targeted and can bring practical business insights. Besides, **the 5-segments RFM model also does well in distinguishing active/inactive customers and high/medium/low value customers**.
  -

FUTURE WORK

- In this project, we segment customers using KMeans. We can try to identify meaningful segments using other methods such as Non-negative Matrix

Factorization (NMF) or Hierarchical Clustering

- Except for customer segmentation, we also can try market basket analysis, product segmentation, Customer Lifetime Value prediction, next transaction prediction, customer churn prediction, and so on.

REFERENCES

1. Data Science (The MIT Press Essential Knowledge series), By John D. Kelleher and Brendan Tierney, 2018, The MIT Press

2. Customer Segmentation with K Means Clustering