TheAnalyticsTeam

Sprocket Central Pty Ltd

Data analytics approach

Muhoza U Bizumuremyi
Linkedin: https://www.linkedin.com/in/muhozaursus/
Email: writetoursus@gmail.com
Phone: +14696395382
Github: https://github.com/ursus123/KPMG-Data-Analytics-Internship

Agenda

1. Introduction
2. Data Exploration
3. Model Development
4. Interpretation

Introduction

Available Data

After a careful data assessment, we are remained with a 3-month transaction (as suggested in the email) dataset where we have to analyze customer behaviors and suggest to the client the best customers (loyal customers) to focus on. There are 4871 transactions.

Method used:

- Cohort Analysis
- Customer Segmentation (RFM Metrics and K-Means Clustering)

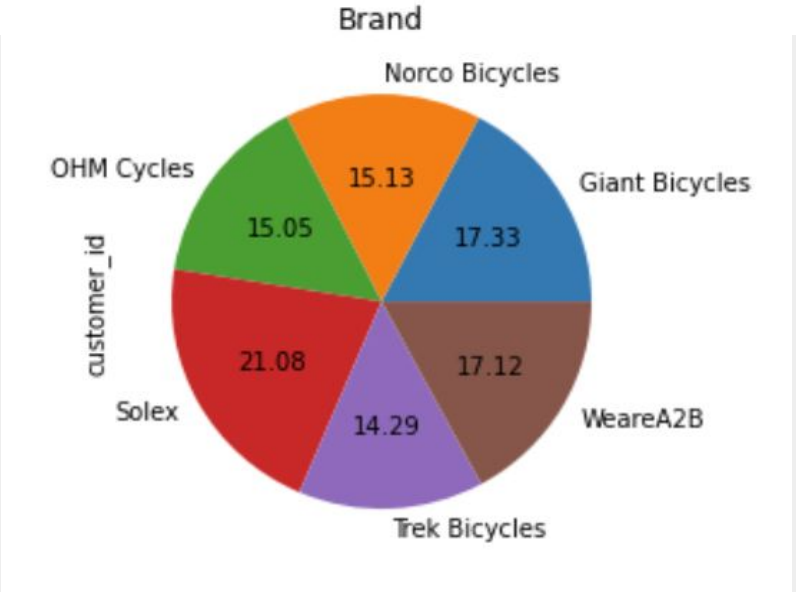
```
In [6]: t3m.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4871 entries, 0 to 4870
Data columns (total 26 columns):
 #   Column                               Non-Null Count  Dtype
---  ------                               --------------  -----
 0   customer_id                          4871 non-null   int64
 1   first_name                           4871 non-null   object
 2   last_name                            4871 non-null   object
 3   gender                               4871 non-null   object
 4   transaction_id                       4871 non-null   int64
 5   product_id                           4871 non-null   int64
 6   transaction_date                     4871 non-null   datetime64[ns]
 7   online_order                         4871 non-null   float64
 8   order_status                         4871 non-null   object
 9   brand                                4871 non-null   object
 10  product_line                         4871 non-null   object
 11  product_class                        4871 non-null   object
 12  product_size                         4871 non-null   object
 13  list_price                           4871 non-null   float64
 14  standard_cost                        4871 non-null   float64
 15  product_first_sold_date              4871 non-null   datetime64[ns]
 16  address                              4871 non-null   object
 17  postcode                             4871 non-null   float64
 18  state                                4871 non-null   object
 19  property_valuation                   4871 non-null   float64
 20  past_3_years_bike_related_purchases  4871 non-null   float64
 21  wealth_segment                       4871 non-null   object
 22  deceased_indicator                   4871 non-null   object
 23  owns_car                             4871 non-null   object
 24  DOB                                  4871 non-null   datetime64[ns]
 25  tenure                               4871 non-null   float64
dtypes: datetime64[ns](3), float64(7), int64(3), object(13)
memory usage: 989.5+ KB
```

Data Exploration

Basic Exploratory Analysis

- Almost 51% of all transactions were made by women and 49% by men
- Almost 51% of all transactions were made by customers who owns a car
- 54% of all transactions were made by customers who are situated in New South Wales, 24% are in Victoria and 21% are in Queensland
- Almost 50% of all transactions were made by Mass Customers, while high net worth customers purchased around 26% and 24% were made by Affluent customers
- Solex is the most purchased brand around 21%



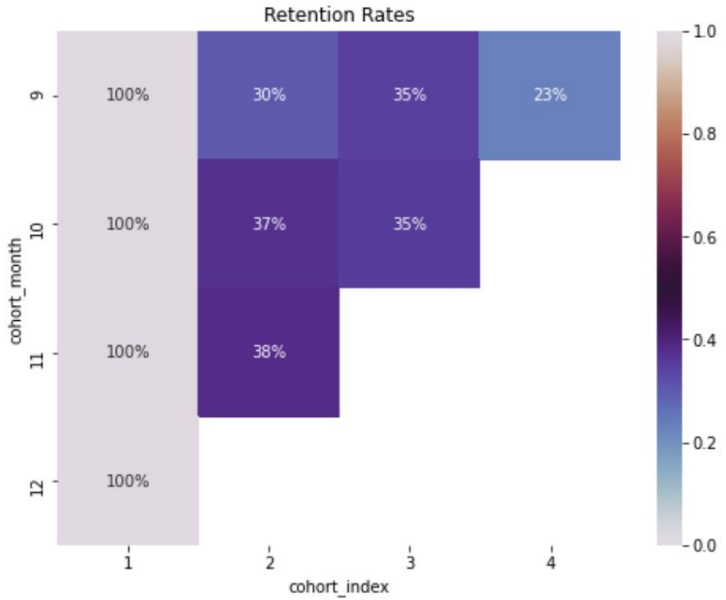
Data Exploration

Cohort Analysis

- 23% of all customers who made their first purchases in september were still active in December
- 35% of all customers who made their first purchases in October were still active in December
- 38% of all customers who made their first purchases in November were still active in December

This shows that there is some kind of retention

- Almost 54% of all customers have made more than 1 purchases in the last 3 month



Data Exploration

Customer Segmentation

The method used here: RFM Metrics
R: Recency of a customer
F: Frequency of a customer
M: Monetary Value of a customer
Then we gave RFM Score and we grouped customers into Gold(Good Customers), Silver and Bronze categories

As you can see in the table: we have 584 customers a recency (the lower the better) of 21, mean frequency of 3 (meaning they have at least 3 purchases for the last 3 month) and 3465 monetary value considering each brand costs $1000

(The list of those customers can be found in github repository provide at the end of this presentation)

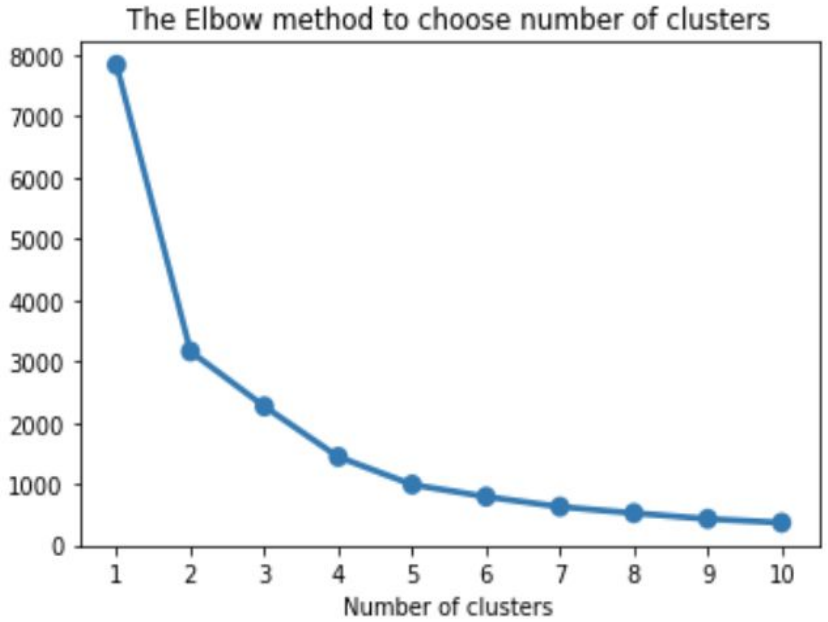
	Recency	Frequency	MonetaryValue	
	mean	mean	mean	count
Segments				
Bronze	63.210859	1.000000	1000.000000	792
Gold	21.354452	3.465753	3465.753425	584
Silver	27.090688	1.663968	1663.967611	1235

Model Pre_Processing and Processing

The methodology used is K-Means clustering (an unsupervised machine learning algorithm). This algorithm will help us group our customers into groups/segments or clusters and then we will see what kind of characteristics those groups share.

To choose the number of groups (segments, clusters), we used a common method called the elbow method, in our case both 2 and 3 groups were used.

However, to use this algorithm we had to unskew and standardize the data.



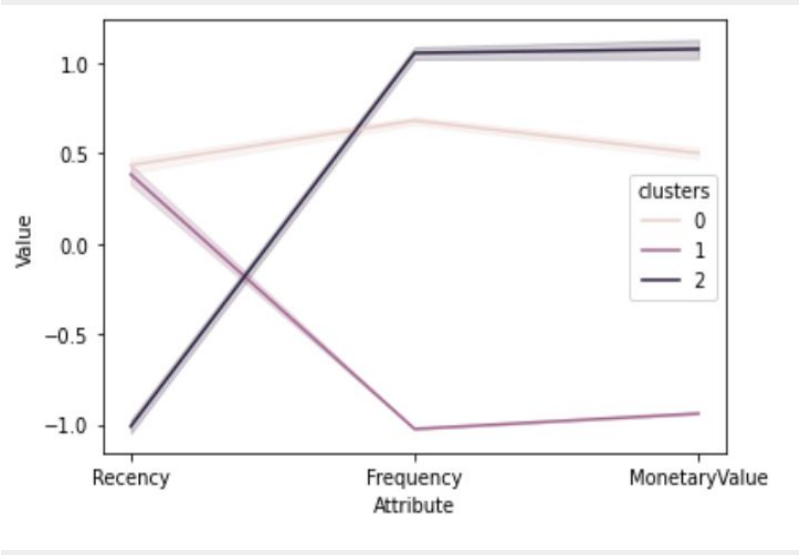
Interpretation

Final Thoughts

Using the 3 groups/segments (number of clusters) we can see that group 2 buys more often (more frequency), they recently purchased products (low recency) and they spent more.

They are about 744 customers. I have provided the list in my github repository, check the link at the last page)

These are the customers who are very important.



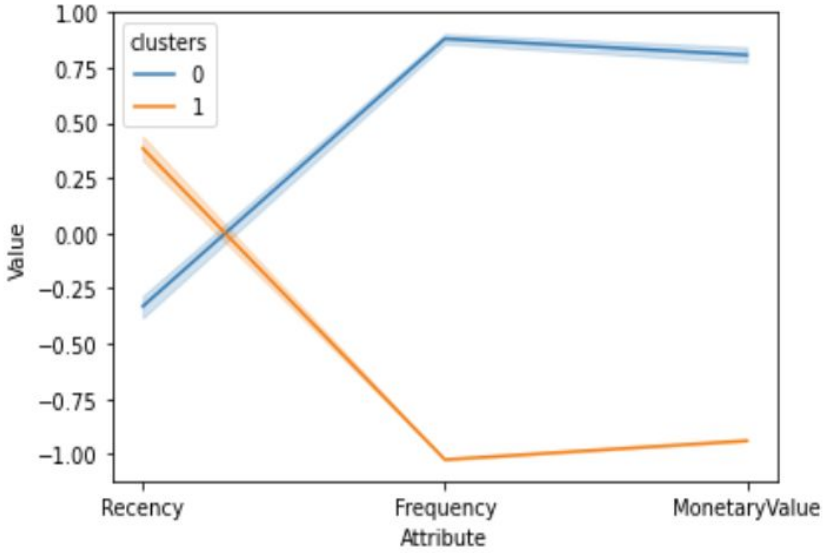
Interpretation

Final Thoughts

Using the 2 groups/segments (number of clusters) we can see that group 0 buys more often (more frequency), they recently purchased products (low recency) and they spent more.

They are about 1404 loyal customers. I have provided the list in my github repository, check the link at the last page)

These are the customers who are very important.



THANK YOU

Appendix

For more information visit my github page :
(https://github.com/ursus123/KPMG-Data-Analytics-Internship/blob/main/KPMG%20Data%20Analytics%20Internship%20Part%202%20EDA%20and%20Modeling.ipynb)