Data Fellowship 5

Machine Learning Practice Case

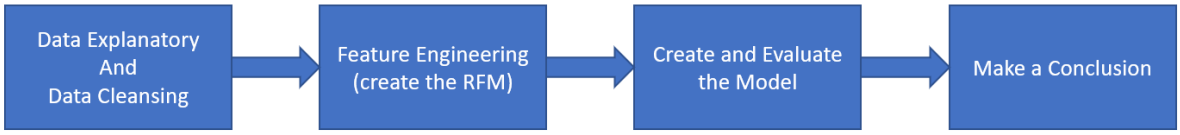
By: Ilham Wahyu Adli

Github Link: [ilhamwahyu8/Customer-Segmentation-Practice-Case: For Data Fellowship 5 Practice Case (github.com)]

The dataset given was Online Retail Data with 8 attributes:

	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	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850	United Kingdom

The attributes represent the transaction of each customer in the Online Retail Marketplace, with this information we can make customer segmentation. With customer segmentation we can manage our marketing budget, get a better knowledge of the customer's need. So our marketplace will be more efficiently handling the customer. One of method for analyse the customer value is by using RFM. RFM stands for Recency, Frequency and Monetary Value. With the attribute InvoiceDate we can see how recent the transaction (Recency), InvoiceNo we can get the total transaction made by one customer (Frequency) and Quantity and UnitPrice will help with how much the customer spend for entire transaction (Monetary Value).



Explanatory Data Analysis

1. **Null Value**
   a. CustomerID

```
1  dataset.isna().sum()
```

```
InvoiceNo          0
StockCode          0
Description     1454
Quantity           0
InvoiceDate        0
UnitPrice          0
CustomerID     135080
Country            0
dtype: int64
```

There's a lot of null value in CustomerID, because of this attribute is one of the key point of the customer so we will lookup with this problem.

Assumption:

The customer use guest account (1 time transaction) so it doesnt need an account to buy a product.

Solution:

Ask the client whether it is true or false, if its true then we can apply the NaN value with their invoice instead.

b. Description

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
622	536414	22139	NaN	56	2010-12-01 11:52:00	0.0	536414	United Kingdom
1970	536545	21134	NaN	1	2010-12-01 14:32:00	0.0	536545	United Kingdom
1971	536546	22145	NaN	1	2010-12-01 14:33:00	0.0	536546	United Kingdom
1972	536547	37509	NaN	1	2010-12-01 14:33:00	0.0	536547	United Kingdom
1987	536549	85226A	NaN	1	2010-12-01 14:34:00	0.0	536549	United Kingdom

It's so confusing with No Description and 0 price so we assume this is bad transaction, so we need to delete it.

With imputing and removing the null value we already got 0 null value in our data, so let's move on with another process

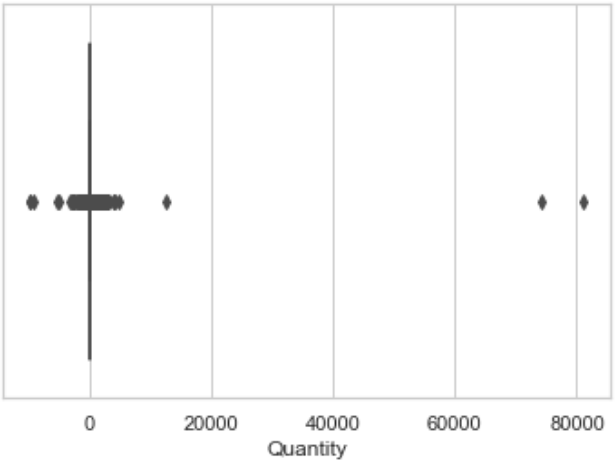
2. Explore Cancelled Invoice

```
Canceled Order:  3836
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
141	C536379	D	Discount	-1	2010-12-01 09:41:00	27.50	14527	United Kingdom
154	C536383	35004C	SET OF 3 COLOURED FLYING DUCKS	-1	2010-12-01 09:49:00	4.65	15311	United Kingdom

With 3836 data started with 'C' on the InvoiceNo, we can assume it is canceled order from the customer. But we need to check it first when we delete the 'C' is there still the history of that transaction in the StatedInvoice. And turns out there's no data got the same value if we delete the 'C'. So we just ignore all the canceled transaction.

3. Explore Quantity



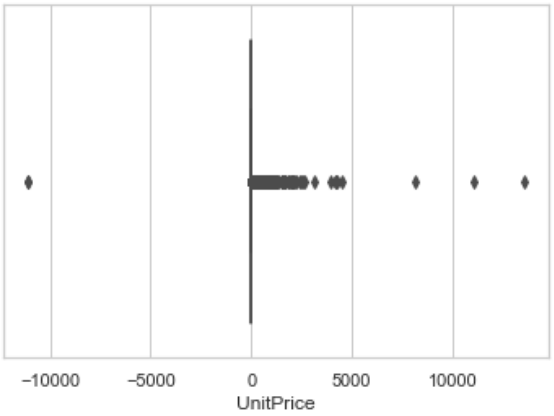
There's no way person buy value less than 0 so we will check what's happen with that transaction by the description given.

And when the quantity <= there's error in the order so we decided to delete the record. Another anomaly with the quantity columns is there's 2 transaction with more than 10000, we will look up with this problem. And the result is:

Description	Quantity
MEDIUM CERAMIC TOP STORAGE JAR	74215
ASSTD DESIGN 3D PAPER STICKERS	12540
PAPER CRAFT , LITTLE BIRDIE	80995

It still makes sense people buy a lot of that product maybe for reselling or for completing their project, so we still keep the record.

4. **Explore UnitPrice**



StockCode	Description	Quantity	InvoiceDate	UnitPrice	C
B	Adjust bad debt	1	2011-08-12 14:51:00	-11062.06	
B	Adjust bad debt	1	2011-08-12 14:52:00	-11062.06	

There's 2 data with negative UnitPrice it just doesn't make any sense and with StockCode 'B' so we need futher analysis about the StockCode 'B'. With this information we find 3 record with 'B' StockCode and the description is 'Adjust bad debt'.

StockCode	Description	Quantity	InvoiceDate	UnitPrice	C
B	Adjust bad debt	1	2011-08-12 14:50:00	11062.06	
B	Adjust bad debt	1	2011-08-12 14:51:00	-11062.06	
B	Adjust bad debt	1	2011-08-12 14:52:00	-11062.06	

Assumption:

It was company debt so we don't need the data for the customer segmentation, therefore we delete the data.

Another problem with the data is there's a lot of data with 0 price, when we look up the description, the item still makes sense.

Assumption:

It was free item on the market so people buy it, so we decided to keep the data. Until the client confirmed about this problem.

There's 6 price that above 4000, so we need to check about this feature.

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice
537632	AMAZONFEE	AMAZON FEE	1	2010-12-07 15:08:00	13541.33
551697	POST	POSTAGE	1	2011-05-03 13:46:00	8142.75
560373	M	Manual	1	2011-07-18 12:30:00	4287.63
562955	DOT	DOTCOM POSTAGE	1	2011-08-11 10:14:00	4505.17
573077	M	Manual	1	2011-10-27 14:13:00	4161.06
573080	M	Manual	1	2011-10-27 14:20:00	4161.06

After further analysis only the InvoiceNo '562955' that got different transaction, so we keep the InvoiceNo '562955' and delete the rest.

5. **Explore StockCode**

   In the description feature, there's 2 StockCode called 'POST' and 'M'.

   Assumption:

   Based on their description 'POST' stands for POSTAGE (maybe the customer need to pay some delivery) and 'M' stands for Manual (we still don't know about this, we can assume this is the transaction made by offline store) so we decided to keep this 2 StockCode.

6. **How many times each country made a transaction**

Country	InvoiceNo
United Kingdom	18015
Germany	457
France	390

Assumption:

Different country got different wages and lifestyle, so we specialize this project for the UK customer because UK got 18015 unique transaction and maybe it is their main market.

Feature Engineering

Because we want to apply the RFM method we need to get information about the latest transaction in the data, total transaction made by customer and total spending of the customer.

1. **Recency**
   How we approach the recency value is by get their latest transaction and calculate (in month) with the latest transaction in the market.

	CustomerID	LastPurchaseDate	Recency
0	12346.0	2011-01-18 10:01:00	10.833333

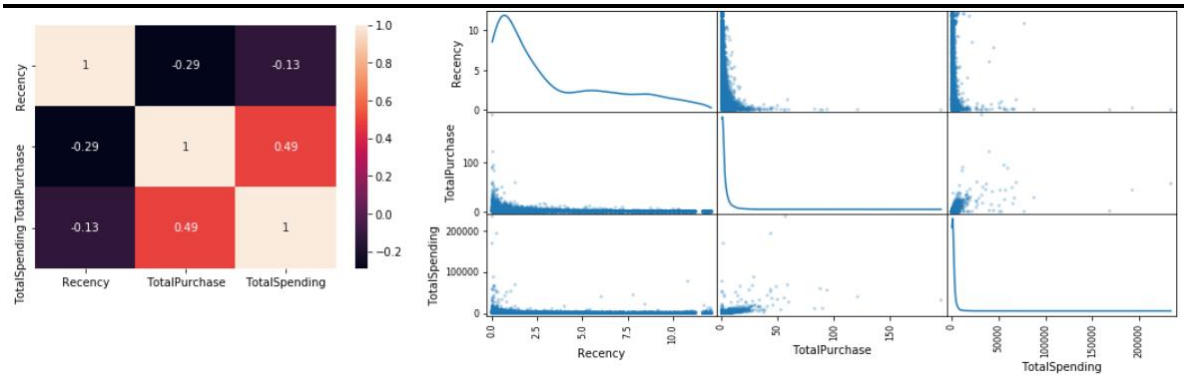
2. **Frequency**
   For the frequency we need to sum the unique InvoiceNo for each customer.

	CustomerID	LastPurchaseDate	TotalPurchase
0	12346.0	2011-01-18	1

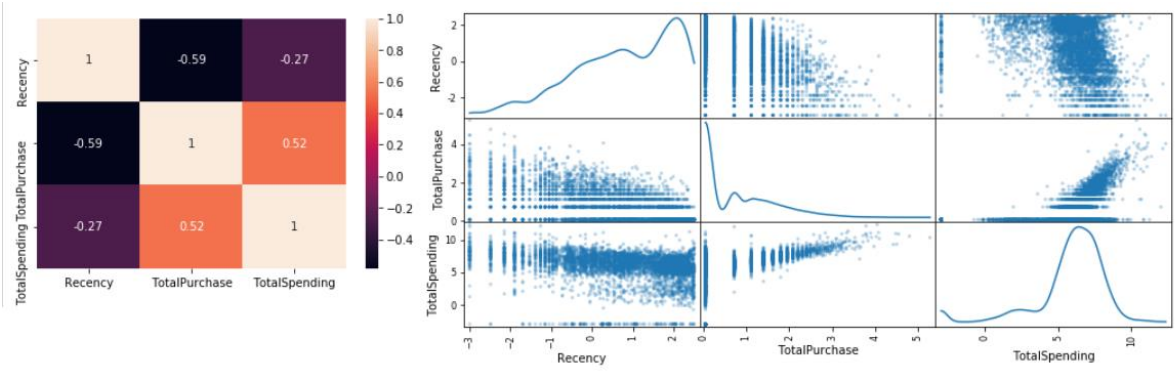
3. **Monetary Value**
   And last thing is monetary value, we need to sum all the total spending for each customer. First thing we need to do is create the column total price because the dataset given only provide quantity and price, that can be done by multiply the quantity by price. After that we can sum all the total price to gain information about monetary value.

	CustomerID	Recency	TotalPurchase	TotalSpending
0	12346.0	10.833333	1	77183.6

Correlation



Because of bad data distribution with recency, total purchase, and total spending. We apply some log transformation to make distribution better.



After applying the log transformation, the data is slightly better so we will work with this data.

Implementing the Algorithm and Evaluation

```
With 2 Clusters the score is: 0.0222064703936924
With 3 Clusters the score is: 0.3612310141973127
With 4 Clusters the score is: 0.3602568645839488
With 5 Clusters the score is: 0.3891383533613868
```

Based on Davies Boulding Score the best cluster for this customer segmentation problem is 2 clusters, because the minimum score of Davies Boulding Score is 0, with lower values indicate better clustering. With total customer for each cluster is 1490 for clusters 1 and 3860 for clusters 2. And information about 2 cluster provided below:

Info about cluster 1

	Recency	TotalPurchase	TotalSpending	Clusters
count	1490.000000	1490.0	1490.000000	1490.0
mean	1.422250	1.0	4.320763	0.0
std	1.070212	0.0	3.517850	0.0
min	-2.995732	1.0	-2.995732	0.0
25%	0.961901	1.0	2.079442	0.0
50%	1.783391	1.0	5.539807	0.0
75%	2.187922	1.0	7.306109	0.0
max	2.500069	1.0	10.876933	0.0

Info about cluster 2

	Recency	TotalPurchase	TotalSpending	Clusters
count	3860.000000	3860.000000	3860.000000	3860.0
mean	0.363151	4.141710	6.529806	1.0
std	1.366409	6.921233	1.248851	0.0
min	-2.995732	1.000000	-2.995732	1.0
25%	-0.538997	1.000000	5.691102	1.0
50%	0.520776	2.000000	6.465608	1.0
75%	1.500367	5.000000	7.348169	1.0
max	2.500069	195.000000	12.361952	1.0

As we can see the clusters 1 customer is the only customer that only buy 1 time only on the website and clusters 2 is the loyal customer. And the other variables feel like does not affect much about customer segmentation. Or maybe this occurs because of the assumption of guest account. Maybe if this online store give a discount for new account we can boost the market and we can track the flow of transaction easier.