Data Science for Business Final Project

Customer Profiling – E-Commerce (Home Improvement Retailer)

Team Number – A06

Yvonne Xie / Mingyu Gu / Rachel Zhong / Nisha "Mint" Tantivess / Aritra "Auri" Shome

Introduction

Profiles describe customers categorically so they can be grouped for marketing and advertising purposes. By understanding customers in different segments, companies can make better decisions on targeting and marketing strategy to increase profits. Our team used a customer purchase history dataset from a UK retailer (spread across Europe), segmenting customers based on their purchase behavior and profiled them subsequently based on their demographics, preference of products and purchase behavior.

We aim to help the UK retailer know their customers better by answering the questions below –

a. How valuable are they to me?
b. Are they going to stop coming back?
c. How frequent are they? When was their last purchase?
d. Do they return a lot of products they buy?
e. What seasonal patterns do they exhibit?
f. Are there any high value segment that I am missing out on/engage better?
g. How can I push my customers to a higher spending bracket etc.

Data Understanding

We use a transactional dataset with all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The dataset contains detailed information on every transaction that happened in the time frame, including the time the transaction happened, description of the product, unit price, quantity, and unique customer id for identifying who make the purchases. After data cleaning, there are 4,236 unique customer ID's and 4,839,694 transactions. There are potential issues with the dataset such as duplicates and missing Customer ID's for certain transactions, but these issues are addressed in the Data Preparation section.

Business Understanding

Customers exhibit different purchasing behaviors – some make purchases regularly with relatively inexpensive items while others may come in occasionally but make place bulk orders. Therefore, it is important to understand these behaviors so that we could target different customer segments with greater precision. In particular, there are two main goals that we would like to achieve with this dataset: First, we want to increase customer spending and their frequency of visits to increase customer lifetime value in order to maximize revenue in the long term; Second, we would like to identify customers who are at risk of churning out but are highly valuable to us so that we can provide them personalized marketing content to incentivize them. This is important as the cost of acquiring a new customer is significantly higher (5x) than retaining a customer.

Data Cleaning and Preparation

- Removed Duplicates – There are rows which contains the same data for every column, so we have dropped them.
- Removed transactions with missing Customer ID – The project is aimed at profiling customers and this is only possible when a particular customer logged in while shopping (not as a guest). All "checkout as guest" customers have a blank Customer ID and have been removed from this analysis. Removed these transactions also remove items with bad description such as "??", "??missing", "sold as sets" etc.
- Stock codes having alphabets – Some of the transactions have an Alphabet as their stock (less than 1% of the transactions). They include Discounts, Bank Charges, Manual, Postage etc. We could have kept these transactions and flagged them accordingly but for less than a percent, no point going through the complications
- Converted the Date time column from "character" to "POSIXct"
- Added a column named Revenue (Quantity x Unit Price)

More data anomalies addressed

- Negative Quantity – We find out that for transactions that have negative quantity there will also be a similar transaction with a positive quantity. Therefore, the negative quantity occurs when customers return the items they bought previously.
- Unit Price extreme numbers – Price of zero or less than one euro are mostly for items that are very cheap (sharpener, pencils etc.). Zero price items can be free items (need to explore more later). Items with very high unit prices are Antique sets, Kitchen cabinets etc.

ADS for clustering – Customer level data

Apart from the original transaction-level data, we also need to transform the dataset to customer-level data since we want to classify customers based on their behaviors. We list all the possible variables for classifying customer as follows:

Calculate the following for **each customer:**

- Total number of transactions (valid/non returned) - Total Count of unique InvoiceNumber
- % returned transactions - (Total Count of InvoiceNumber with Quantity < 0)/ (Total Count of InvoiceNumber)
- Average interpurchase interval – Average number of days between a customer's transaction
- Recency – Days between the most recent purchase and the Max InvoiceData for the whole dataset
- Total Revenue – Sum of Revenue for each customer
- Units purchased – Sum of Quantity for each customer

After the transformation, we have **4236 unique customers** who satisfy the following requirements:

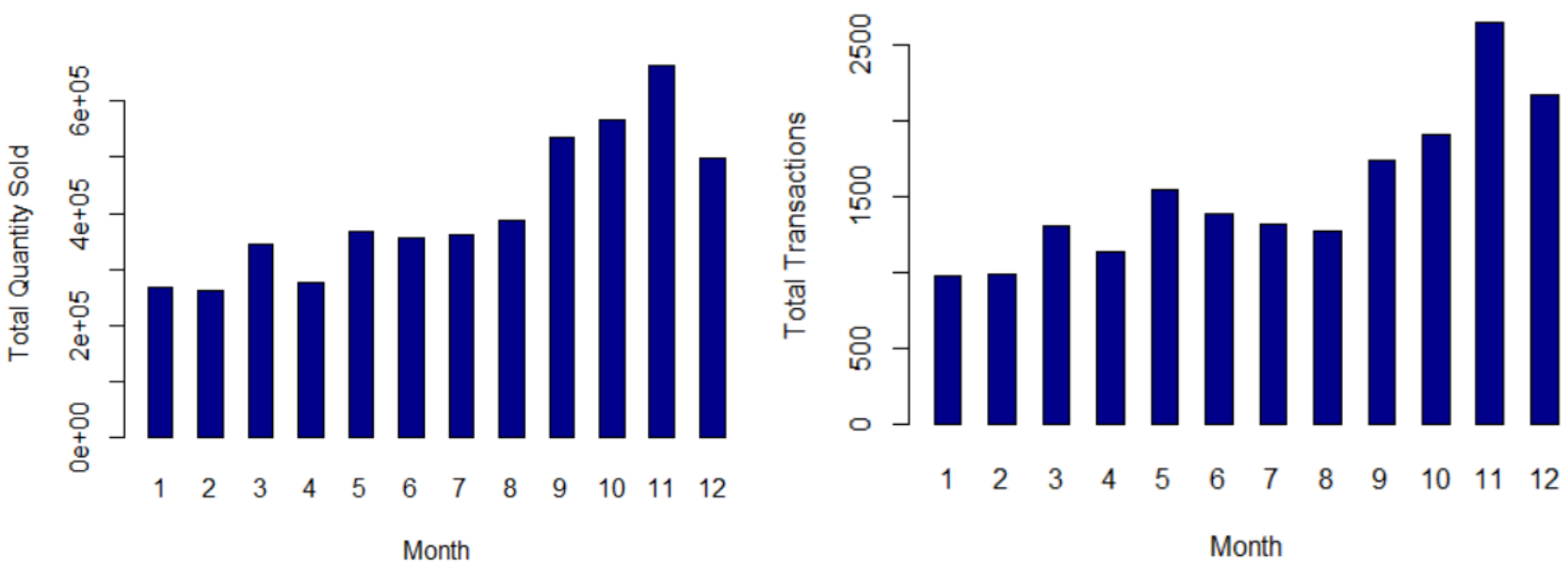
- With the unit purchase > 0: if the customer has negative unit purchase means he only have the return action in our chosen time frame which can make our prediction biased since we are missing the according purchase action,
- With return_quantity_percent < 100: For those who did not satisfied, their purchases are not complete either in our chosen time frame, they returned stuff that are purchased before it.
- With TotalTransaction > 1: we only consider customers with more than one transaction because that's when we have enough data to classify a customer and understand his/her persona

With the customer level data, we will perform clustering (for customers with more than two transactions) based on some of the above variables. After clustering, we will profile the clusters by identifying the key features of each one and train a classification model to assign each new customer a cluster (deployment).

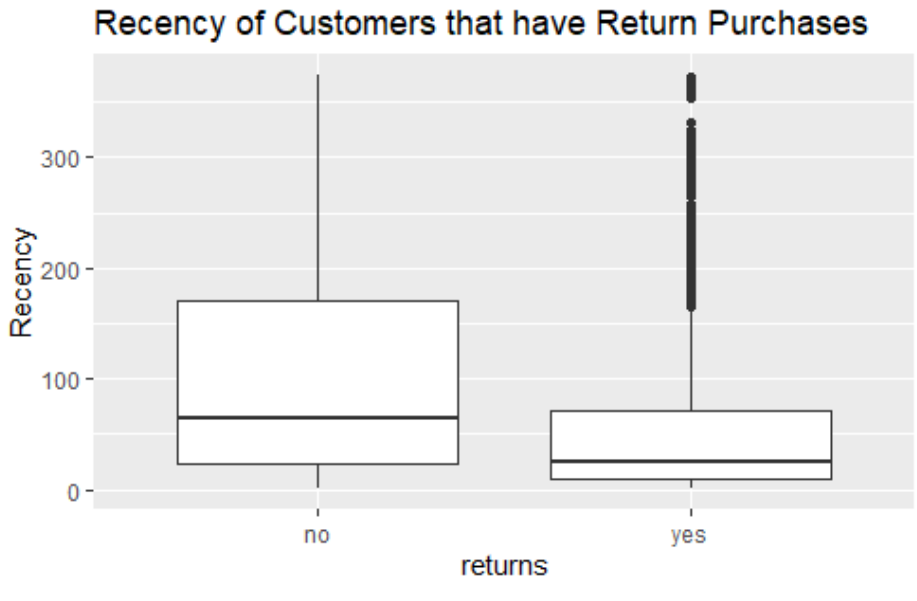
Exploratory Analysis

The results of our exploratory data analysis are based on both the cleaned-up **transaction-level** data and **customer-level** data:

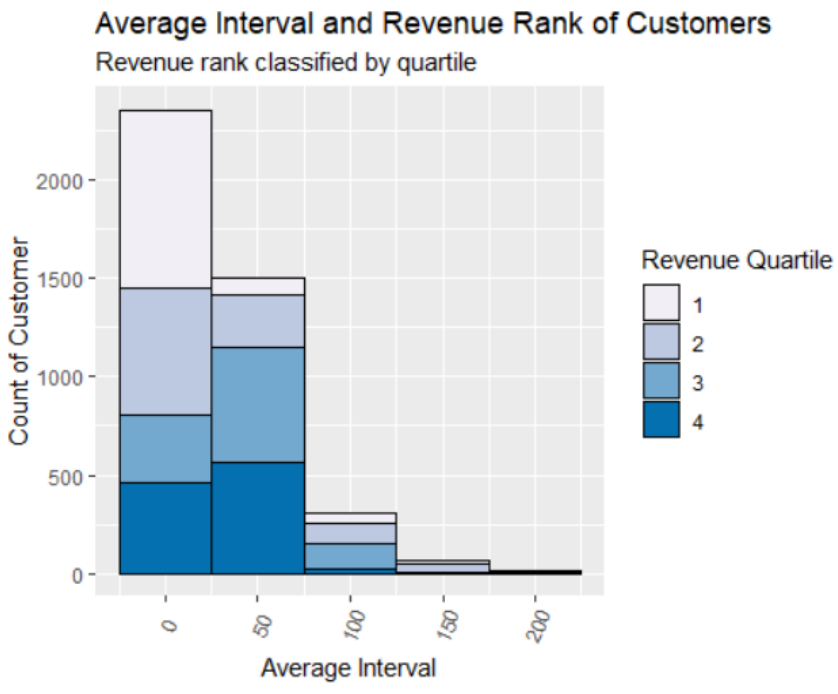
Transactions and Quantity purchased across months – Total transactions and total quantities both shown difference among different months, and the peak both are November.



The last transaction for customers that have returned a purchase are more recent.

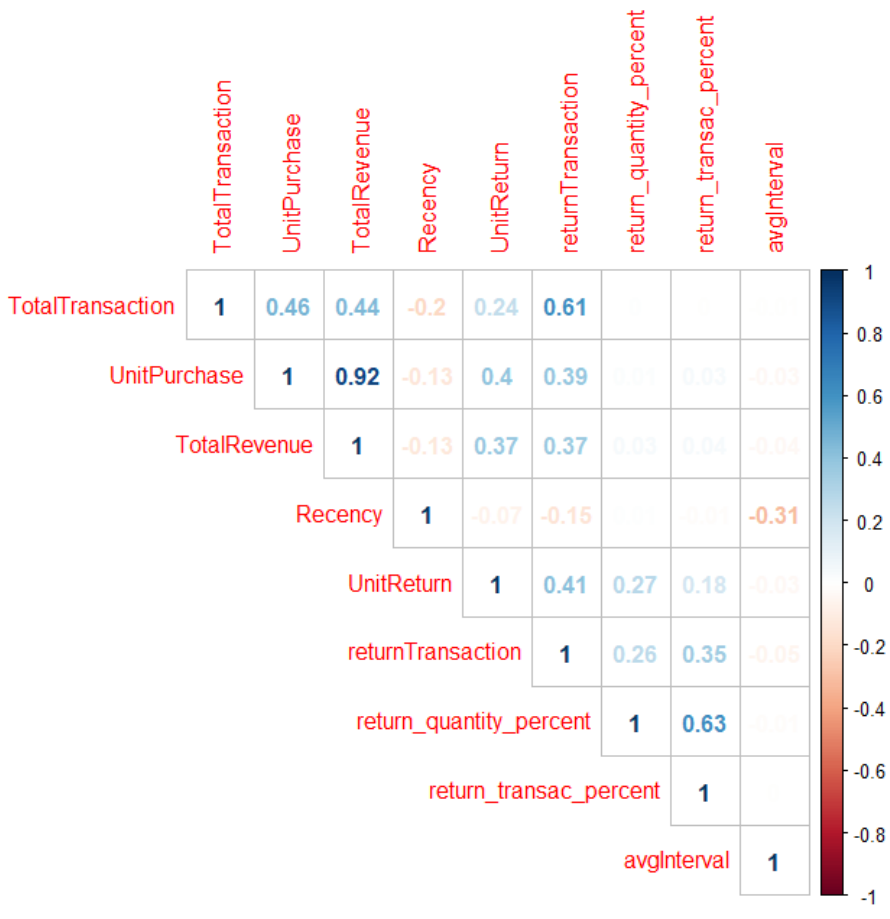


Customers are more likely to have short time gaps between transactions. To explore the different levels of spending for customers with different average intervals, the variable TotalRevenue was split into quartiles (1st quartile are customers that have spent the most, 4th quartile are those that have spent the least). After adding this variable to the analysis, it is clear that the lower the interval between transactions, the customer is more likely to be a big spender. This is most likely because many trips are made to the website.



Correlation plot across variables using customer level data – We see that Unit Purchase has a high correlation with Total Revenue, Total Transactions, and return Transactions. Therefore, we will not use pairs of variables that exhibit high correlations into clustering at the same time.

The correlation plot will help us in choosing variables while doing clustering and the following classification of new customers.



Modeling

Why are we using clustering?

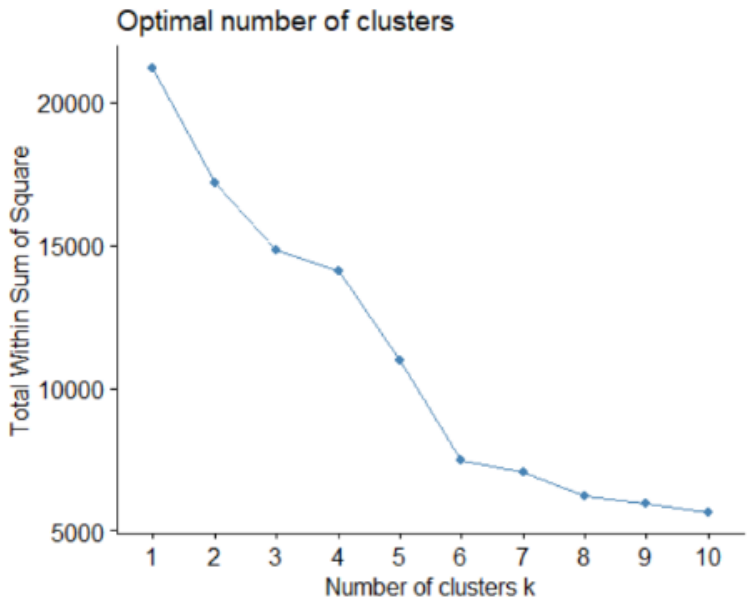
Firstly, in order to better understand the 4236 customers, we need to segment them based on certain features. We wish to profile the customers based on their past purchase behavior and leverage this information to improve CLTV and reduce attrition by offering personalized promotions and offers. Some segments we hope to identify are – customers who visit infrequently, customers who spend a lot and others.

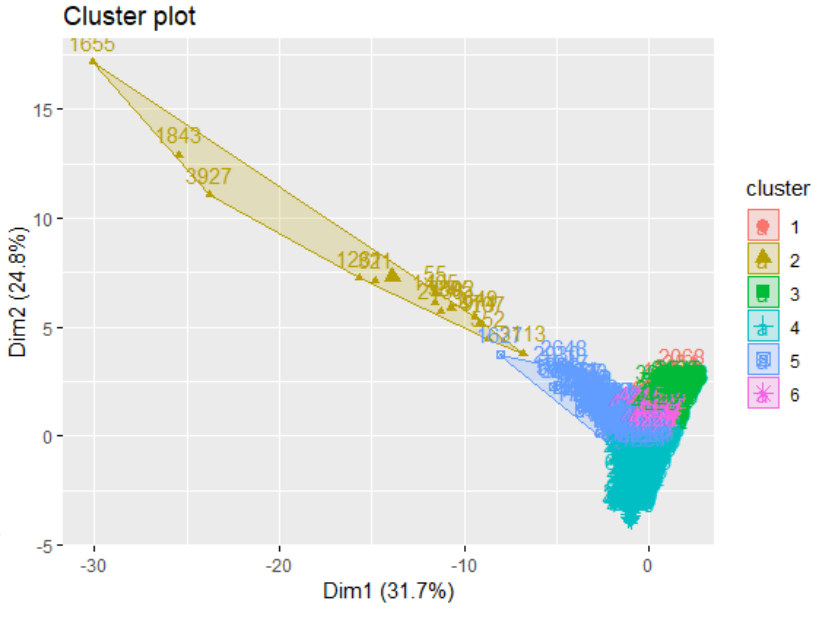
Clustering using different variables – Unsupervised learning

Based on the correlation of different variables, we choose 3 different set of variables as the input. For each set, we use **Silhouette** and **WSS** to determine the optimal *k* value for *k* means clustering. Sometimes these two methods do not yield the same optimal *k*, so we try different values of *k* based on the results:

1. Input variables: Unit purchase + TotalTransaction + avgInterval + return_quantity_percent

   Choose the optimal *k* using WSS, from the plot we choose 6 clusters to do the k mean, and the results shows like





Other input combinations with multiple *k*s that were tried out are as follows:

2. TotalRevenue + Recency + UnitReturn + avgInterval + UnitPurchase

3. TotalRevenue + UnitReturn + avgInterval + UnitPurchase

Those iterations were rejected as the differences between each cluster was not distinct enough for efficient deployment.

Customer Profiling

The output of the finalized k-means with six clusters are as follows –

Row Labels	Average of avgInterval	Average of return_transac_percent	Average of Recency	Average of TotalRevenue	Base
1	27.9027451	23.05098039	111.0980392	1761.308627	51
2	6.161333333	2.472666667	5.133333333	106855.0813	15
3	7.61245283	1.69436182	253.6770255	470.7663929	901
4	87.98922515	1.306140351	54.39473684	908.3300746	684
5	33.87321455	2.542479584	29.19896065	3454.191589	1347
6	6.398206785	1.398303716	54.64216478	645.1098223	1238

The first two clusters comprise of less than 100 customers. Diving deep into their attributes we see that cluster 1 comprises of customers who have a lot of **return transactions**. On average customers have 2% return transaction. However, this segment returns an item 50% of the times on average. We treat this as a microsegment.

Cluster 2 comprise of customers who have purchased a lot of quantities per transaction. On average customers purchase ~1200 quantities during the period. However, this segment has purchased around 60,000 items. We treat this microsegment as "**Wholesalers**".

We then decide to remove these two segments and move towards the other four segments. Also, we remove "return_txn_percent" as most of the other customer clusters have similar percentages of return transactions.

Row Labels	Average of avgInterval	Average of Recency	Average of TotalRevenue	Base
3	7.61245283	253.6770255	470.7663929	901
4	87.98922515	54.39473684	908.3300746	684
5	33.87321455	29.19896065	3454.191589	1347
6	6.398206785	54.64216478	645.1098223	1238

Cluster 3 – *Need to win back*

We see customers in this segment have decent IPI – they are relatively frequent but have low average revenue which is driven by their high recency. On an average, customers in this cluster haven't shopped in the last eight months hence have churned out.

Recommendation – Marketing efforts need to be directed in winning them back as they used to loyal customers once. This could include coupon discounts on products that are frequently bought by this segment. In addition to this, it is also important to identify the causes for customer churn. For example a new competitor in town with lower prices could be a factor in churn.

Cluster 4 – *Sporadic visitors*

These are customers with very high IPI (Intern Purchase Interval), which means that they visit infrequently as a result of which their revenues as low as well. However, they are active because they have shopped with us in the last two months.

Recommendation – Marketing should be focussed on increasing their engagement to translate to more visits. This could come in the form of daily promotional specials that encourage customers to visit the store more often, or coupons issued post purchase that expire in a short timeframe. However, there should also be a control group that do not receive these new marketing efforts. This is to confirm whether more frequent visits will result in higher total revenue, or rather simply maintain the total revenue level with high number of transactions with lower revenue per transaction.

Cluster 5 – *High flyers*

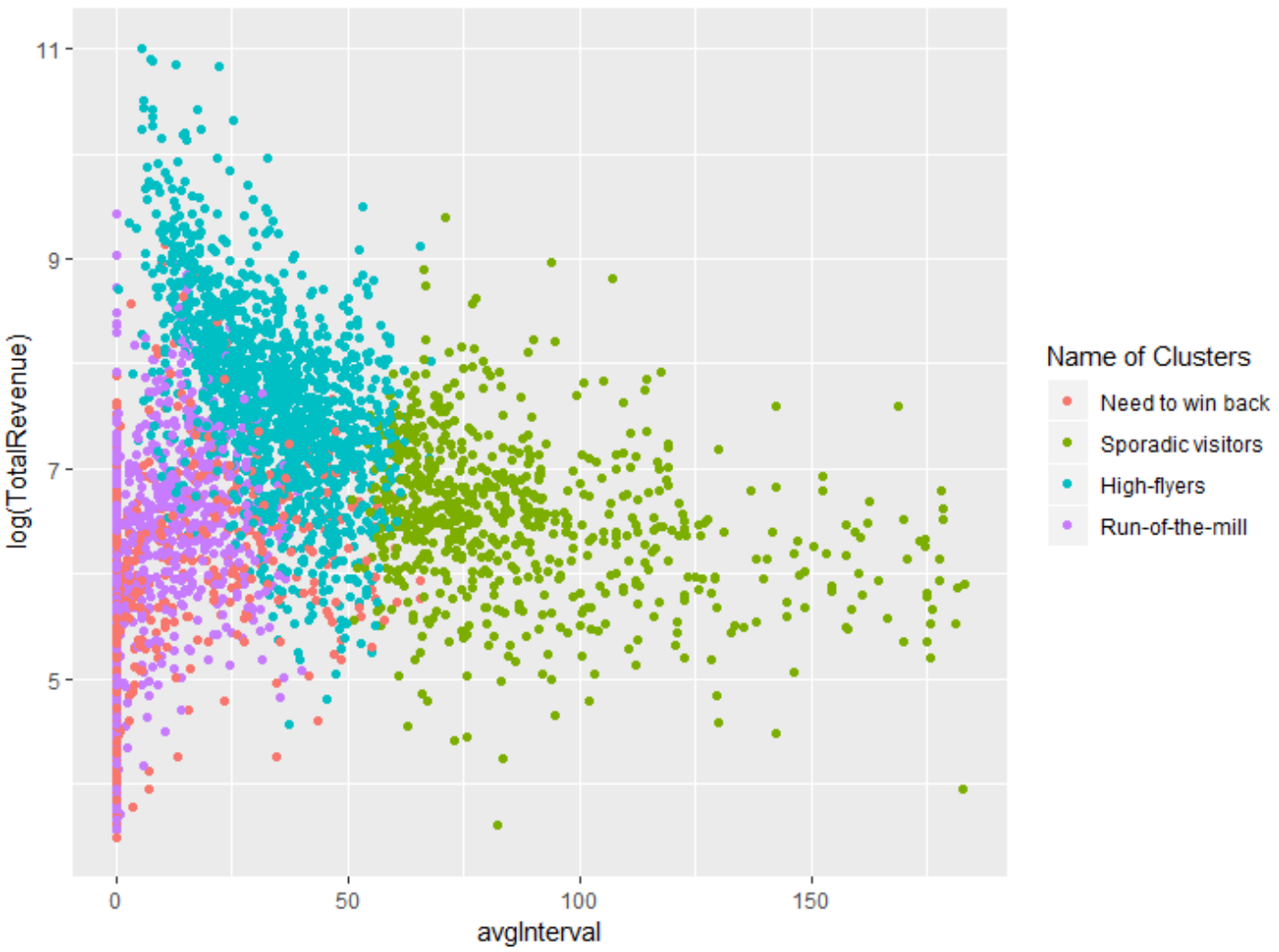
30% of the customer base contributes to more than 70% of the revenue. These customers are active, and frequently shop (at least once a month).

Recommendation – Marketing should be targeted to maintain their loyalty. Communications on new products can be rolled out to keep the retailer brand top of mind. Additional offers need not be rolled out for them as they already contribute a lot to revenue.

Cluster 6 – *Run-of-the-mill*

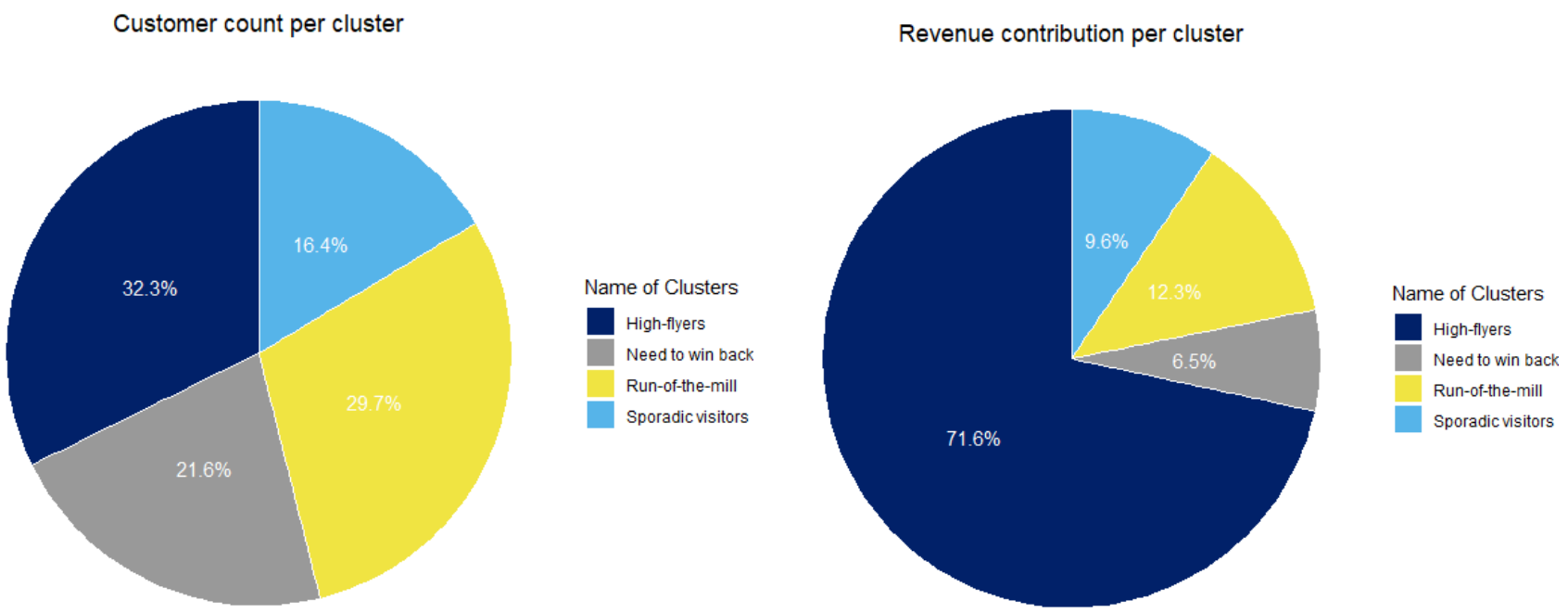
The basic customer base with high frequency and low IPI meaning they frequently shop with me but for low transaction value as a result of which their contribution to revenue is low as well. Close to 30% of the base only contributing to just above 10% of the revenue.

Recommendation – Upsell strategies to such customers to encourage higher spends for each visit. Additional product association analysis can also be run to create an algorithm to recommend additional products. Spend stretch discounts can also be utilized, for example “Buy for amount X to get discount Y”.



Now let’s look at the customers with cluster label 3,4,5 and 6 – From the graph above we see that *Need to win back* and *Run-of-the-mill* customers are very similar in terms of Avg_IPI and Revenue contribution. Only thing that differentiates these two clusters is the attribute **Days since last purchase** which is very large for the *Need to win back*. We also see that the *High Flyers* are concentrated on the upper left corner of

the graph – Less Average IPI (frequent customers) and high revenue. Sporadic are the ones with average contribution to revenue but high IPI (infrequent)



Deployment (Classification Model)

In order to scale and operationalize this solution, we have to come up with a deployment plan. A model that automatically flags each new customer (with 2 or more transactions) to a pre-defined profile.

For customers in the first two clusters or microsegments (Customers who return a lot of the stuff they buy and Wholesalers) we take a rule-based segmentation approach instead of an ML classification model. Reason behind this is their base is very small and for a classifier to tag a customer to a persona with low representation is going to be difficult.

- If the customer has more than 40% of his/her transactions as return – flag them as *Customers who return a lot*
- If the customer has more purchased more than 40,000 items in a year – flag them as *Wholesaler*

To classify and assign personas (*Need to win back, Sporadic, High flyers and Run-of-the-mill*) to the rest of the customers we train a classifier.

We experiment with different classification models (namely **Logistic** and **Random Forest**) to train a classifier. The variables that were chosen for this model were based on our profiling of each of the clusters. We could define each cluster (3,4,5 and 6) by looking at their Average revenue, Average IPI and their recency (days since last purchase).

Below is the OOS prediction output of the **multinomial logistic classifier**. We see that it has a very good accuracy with some misclassification rates in assigning clusters 5 and 6 (high flyers and run-of-the-mill)

	predicted.classes			
	3	4	5	6
3	224	0	0	0
4	1	175	6	0
5	0	7	301	17
6	0	0	20	292

For the Random Forest classifier, we get the below OOS prediction output.

	3	4	5	6
3	224	0	0	0
4	4	169	9	0
5	0	7	301	17
6	0	0	25	287

Like the logistic classifier, we see that the RF model also has trouble classifying clusters 5 & 6 correct. We stick to the Logistic model as it gives better accuracy.

Conclusion

For this Home improvement retailer, we leverage a raw transaction level dataset to develop customer level insights and strategies. We also segmented the entire customer base into cohorts that the Sales and Marketing team can use to personalize their efforts to improve revenue and conversions. And in order to scale this solution for future customers we developed a model that can continue running in the future at a fixed cadence (every three months for instance) to create customer cohorts.

APPENDIX

Team Member Contribution

Yvonne Xie : Data cleaning, Metrics creation for customer data, clustering

Mingyu Gu : Data visualization, clustering iterations

Rachel Zhong : Data visualization, Formatting, Business Understanding

Nisha "Mint" Tantivess : Data visualization, Formatting, Business Understanding

Aritra "Auri" Shome: Model building and Testing, customer profiling