

Group Assignment 1 Report

Group 2: Cai Anqi, Chan Yu Hang, Chin Synn Khee Joash,
Chua Cheng Ling, Chua Hua Ren, Clarence Ong

1. Introduction

In this report, we aim to use both RFM modeling and MBA, using transaction data collected over the past 3 years, to help guide the client to make strategic business decisions to improve profitability.

We hope to be able to target these 2 key problem statements and give the client clear and concise recommendations to grow their business.

1. How do we increase revenue for the client?
2. How do we increase customer brand loyalty for the client?

In the later parts, we will be further exploring and elaborating on these ideas and key problems.

We are provided with 3 datasets, namely DSA3101_Hackathon_Categories_Information.csv , DSA3101_Hackathon_Data.csv and DSA3101_Hackathon_Panelists_Demographics.csv .

2. Exploratory Data Analysis (EDA)

2.1. Preview of each dataset

DSA3101_Hackathon_Categories_Information.csv - 62 rows, 3 columns

Figure 1: First 6 rows of DSA3101_Hackathon_Categories_Information.csv

Category	Calories/100g	Price per Volume
Baby Cereal	188	29.41
Beer	43	15.12
Belacan	563	39.47
Bird Nest	46	73.45
Biscuits	416	15.57
Bouillon	16	29.09

The 62 categories are Baby Cereal, Beer, Belacan, Bird Nest, Biscuits, Bouillon, Butter, Cake, Canned Product, Cereal Beverage, Cereals, Cheese, Chicken Essence, Choc/Nut Spread, Chocolate, Coconut Milk, Coffee, Condensed/Evap Milk, Confectionery, Cooking Oils, Cooking Sauces, Cordials, Creamer, CSD, Cultured Milk, Drinking Water, Eggs, Energy Drinks, Flour, Frozen Food, Fruit/Veg Juices, Ghee, Honey, Ice Cream, Instant Noodles, Instant Soup, Isotonic Drinks, Jam, Kaya, Liquid Milk, Margarine, Milk Powder-Adult, Milk Powder-Infant, Milk Powder-Kids, MSG, Peanut Butter, Rice, RTD Coffee, RTD Tea, Salad Dressing, Savoury Spread, Seasoning Powder, Snack, Soy Milk, Spagetti, Spirits, Sugar, Tea, Tonic Food Drink, Wine, Yoghurt Drink, Yoghurts.

Figure 2: Summary Statistics

Calories/100g	Price per Volume
Min. : 0.0	Min. : 0.370
1st Qu.: 61.0	1st Qu.: 7.513
Median :210.5	Median : 15.390
Mean :243.8	Mean : 23.895
3rd Qu.:377.2	3rd Qu.: 31.957
Max. :800.0	Max. :200.890

Cross-referencing with other sources, we found that the zero calories (MSG and Drinking Water) is indeed true.

DSA3101_Hackathon_Data.csv - 1318024 rows, 6 columns

Figure 3: First 6 rows of DSA3101_Hackathon_Data.csv

Panel ID	Date	Category	Pack Size	Volume	Spend
Panel 101011101	2017-07-02	CSD	1	1.5	1.5
Panel 101011101	2017-07-02	Soy Milk	1	1.0	2.8
Panel 101011101	2017-07-02	Cooking Sauces	1	0.3	3.2
Panel 101011101	2017-07-02	Coconut Milk	2	0.4	5.3
Panel 101011101	2017-07-02	Chocolate	1	0.2	6.6
Panel 101011101	2017-07-02	Cooking Sauces	1	0.3	5.0

This dataset consists of 156 Sundays (3 full years) which spans from 2017-06-25 to 2020-06-14.

Figure 4: Summary Statistics

Pack Size	Volume	Spend
Min. : 0.000	Min. : 0.000	Min. : 0.000
1st Qu.: 1.000	1st Qu.: 0.300	1st Qu.: 3.500
Median : 1.000	Median : 0.800	Median : 5.700
Mean : 1.518	Mean : 3.209	Mean : 9.661
3rd Qu.: 2.000	3rd Qu.: 2.000	3rd Qu.: 10.800
Max. :285.000	Max. :600.000	Max. :1664.000

There are entries with value zero in the 3 numerical columns of Volume , Pack Size and Spend. It could possibly be due to rounding errors (as the values are rounded to nearest 1 decimal place) or simply a mistake in recording.

DSA3101_Hackathon_Panelists_Demographics.csv - 4026 rows, 8 columns

Figure 5: First 6 rows of DSA3101_Hackathon_Panelists_Demographics.csv

ID	BMI	Income	Ethnicity	Lifestage	Strata	#HH	location
Panel 101011101	Obese	Income 1500 - 1999	North Malay	Empty Nesters	Urban	1-3 Member HH	North
Panel 101016101	Healthy	Income 1500 - 1999	North Malay	Teens Aches	Urban	1-3 Member HH	North
Panel 101019101	Obese	Income < 1500	North Malay	Teens Aches	Urban	4 Member HH	North
Panel 101024101	Over Weight	Income 1500 - 1999	North Chinese	Nesting Families	Urban	1-3 Member HH	North
Panel 105009103	Over Weight	Income < 1500	North Malay	Empty Nesters	Rural	1-3 Member HH	North
Panel 105015101	Healthy	Income 2000 - 2999	North Malay	Teens and Toddlers	Rural	7+ Member HH	North

Figure 6: Proportion of each columns

										Proportion								
										Central Chinese	0.1150							
										Central Malay	0.2096							
										Central Others	0.0397							
										Proportion								
										Proportion	Proportion	1-3 Member HH	0.2749627	East Coast Chinese	0.0129			
										Proportion	[0, 1500)	0.1440636	Empty Nesters	0.0675609				
										Proportion	[1500, 2000)	0.1132638	Matured Families	0.1609538	4 Member HH	0.2297566	East Coast Malay	0.1761
										Proportion	[2000, 3000)	0.2046696	Nesting Families	0.2434178	5 Member HH	0.2245405	East Coast Others	0.0044
Rural	0.1775956	Central	0.3643815	Not Available	0.1962245	[3000, 4000)	0.1671634	Teens Aches	0.2309985	6 Member HH	0.1356185	North Chinese	0.0683					
Urban	0.8224044	East Coast	0.1934923	Obese	0.1115251	[4000, 5000)	0.1048187	Teens and Toddlers	0.2635370	7+ Member HH	0.1351217	North Malay	0.1756					
										Proportion	[5000,)	0.2660209	Yankys	0.0335320	North Others	0.0268		
														South Chinese	0.0466			
														South Malay	0.1087			
														South Others	0.0158			

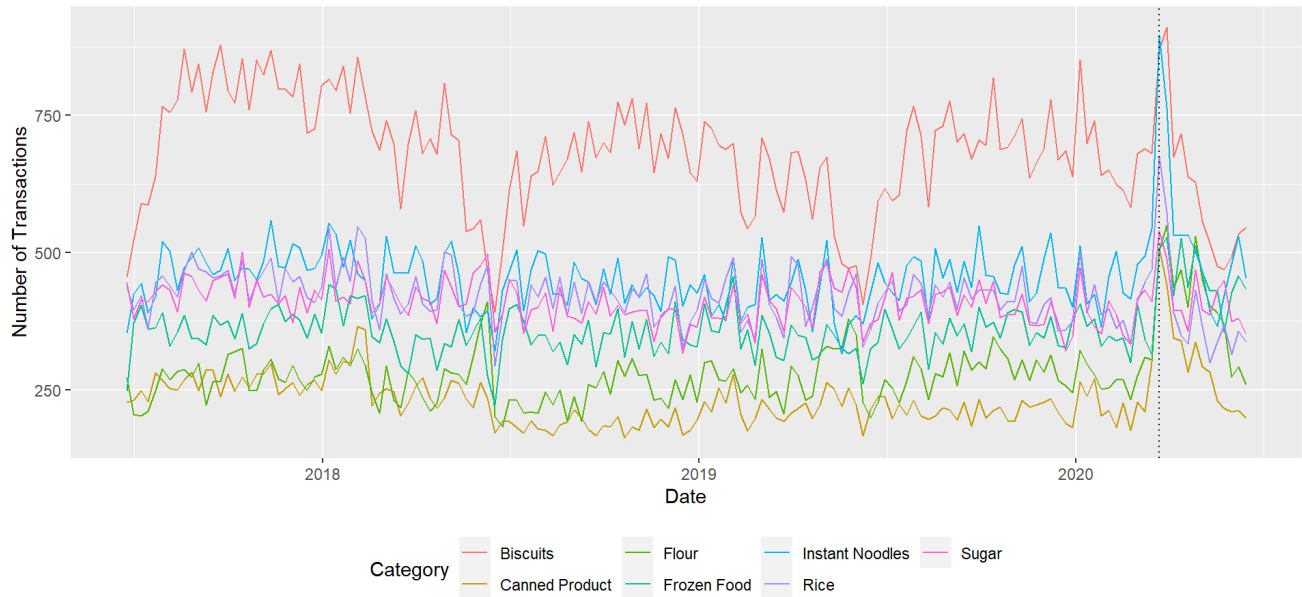
2.2. Background Information

Figure 7: Location VS Income

location	[0, 1500)	[1500, 2000)	[2000, 3000)	[3000, 4000)	[4000, 5000)	[5000,)
Central	0.0477164	0.0743013	0.1806408	0.1792774	0.1199727	0.3980913
East Coast	0.2708601	0.1450578	0.2169448	0.1270860	0.0770218	0.1630295
North	0.1981651	0.1422018	0.2201835	0.1605505	0.0871560	0.1917431
South	0.1202899	0.1144928	0.2173913	0.1971014	0.1318841	0.2188406

We observe that Central (Capital of Malaysia - Kuala Lumpur) and South (Closer to Singapore) regions are generally more wealthy than North and East Coast. Therefore one possible way of increasing profits for the client would be to increase the pricing of items in the central area.

Figure 8: Essential goods transactions over time



We observed that there was a spike in the sales transactions of essential goods such as rice, frozen food. We did further investigation and discovered that Malaysia announced their Movement Control Order (MCO) on 2020-03-18 (Wednesday). Therefore, we believe that this increase in sales may be of interest to the client, and would like to further investigate on how to increase brand loyalty due to possible increase in new customers.

3. Preprocessing

3.1 Data Imputation

From the summary statistics, we have identified that there was erroneous data for three different categories of **Pack Size**, **Volume** and **Spend**

Figure 9: Examples of erroneous data

Panel ID	Date	Category	Pack Size	Volume	Spend
Panel 101011101	2017-07-02	Seasoning Powder	1	0	0.7
Panel 101011101	2017-08-13	MSG	1	0	1.2
Panel 101011101	2017-09-24	Seasoning Powder	1	0	0.8
Panel 101011101	2017-09-24	Seasoning Powder	1	0	0.8
Panel 101011101	2017-12-24	Bouillon	1	0	1.3
Panel 101011101	2017-12-24	Bouillon	2	0	2.6

We observe that there are 49539 such instances of misrecorded/missing data therefore, we propose a few ways to handle the missing data.

1. Since the missing data takes up 3.8% of the entire dataset, we can choose to disregard the missing data as missing data that is less than 5% of the dataset is usually inconsequential. (Schafer, 1999)
2. We can choose to impute the data using kNN imputation. However, we will need to decide which data we want to use to impute.

We will try to impute the data using the median, as we observe that the dataset has some great outliers. (Anil et al, 2019)

We first encode these different categories into labels.

Next, we replace all the zero values in these categories to NaN for imputation

Next, we do impute with the median value

Some final processing bits for the data

Testing to see if there are any errors with the new imputed data

Figure 10: Summary Statistics after imputation

Pack Size	Volume	Spend
Min. : 1.00	Min. : 0.100	Min. : 0.100
1st Qu.: 1.00	1st Qu.: 0.400	1st Qu.: 3.500
Median : 1.00	Median : 1.000	Median : 5.700
Mean : 1.52	Mean : 3.304	Mean : 9.669
3rd Qu.: 2.00	3rd Qu.: 2.600	3rd Qu.: 10.800
Max. :285.00	Max. :600.000	Max. :1664.000

Examples from the previous dataset against the new imputed.

Figure 11: Examples of preimputed erroneous data

Panel ID	Date	Category	Pack Size	Volume	Spend
Panel 108052110	2017-06-25	Biscuits	1	0	1.5
Panel 108052110	2017-06-25	Seasoning Powder	2	0	2.6

Figure 12: Examples of postimputed erroneous data

Panel ID	Date	Category	Pack Size	Volume	Spend
Panel 108052110	2017-06-25	Biscuits	1	2.6	1.5
Panel 108052110	2017-06-25	Seasoning Powder	2	2.7	2.6

However we understand the pros and cons of these various methods on how to handle the missing data.

1. Deleting Missing Data Pros: It is a simple method that we can use easily.

Cons: If the data removed is significant, it may affect our results. However, it may not significantly impact the insights we obtain, as the data is less than 5% of the entire dataset.

2. Imputation using regression imputer Pros: Regression imputer may reduce bias of the dataset. We know that it will not introduce any additional assumptions.

Cons: It may be difficult to accomplish as we do not have much experience in it.

Therefore, our team ultimately chose to proceed to just remove the data.

4. Analysis

4.1. RFM Modelling over Time

Assumptions:

1. Weekly data represents the purchases a customer has made in that particular week. Since there is no data on receipts, we shall treat this as a single visit for computing frequency score.

Example: Customer A who has made 10 transactions 4 weeks ago is not as frequent as Customer B who has made 1 transaction each consistently for 4 weeks.

Approach

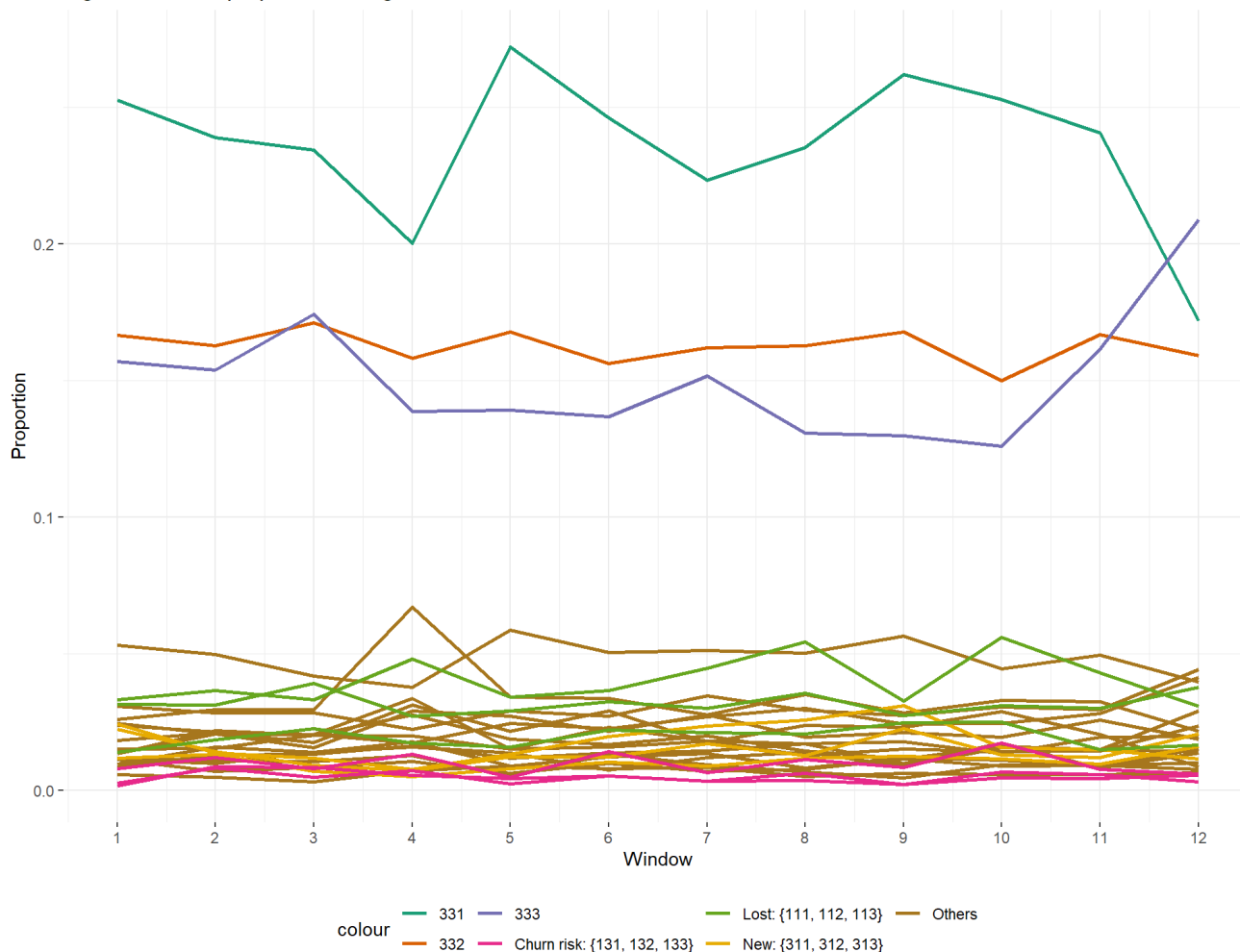
We group by the Panel ID and Date first to compile the transactions in 1 single receipt for that week. Then we group by Panel ID to apply RFM. The monetary is computed based on mean expenditure rather than the total sum to as the total sum is easily influenced by the frequency of the visits (collinearity issues)

Due to the largely skewed dataset, we will **not** be performing quantile cuts. This is so Instead, we will be doing manual segmentation as it is more robust and the results are more interpretable.

This is how we will be representing the different R, F and M values.

- Recency
 - 3: Last visit was ≤ 1 week ago
 - 2: Last visit was ≤ 2 weeks ago
 - 1: Last visit was > 2 weeks ago
- Frequency
 - 3: ≥ 9 visits in 12 weeks
 - 2: < 9 visits in 12 weeks (average twice every three weeks)
 - 1: < 6 visits in 12 weeks (average fortnightly visits)
- Monetary
 - 3: \geq RM60 mean spending in a week
 - 2: $<$ RM60 mean spending in a week
 - 1: $<$ RM30 mean spending in a week

Figure 13: RFM proportion changes over time



In general, there are no big changes (with the exception of RFM = 331) in terms of the proportion of the RFM for in each window (13 weeks period).

Figure 17: Proportion of each columns for Income > 5000 with RFM = 331 after MCO

				Proportion							
		Proportion			1-3 Member HH	0.4000000					
Proportion		Empty Nesters	0.1333333					Proportion			
Healthy	0.6166667	Matured Families	0.2833333	Proportion		4 Member HH	0.3166667	Central	0.4333333	Proportion	
Obese	0.0666667			Rural	0.1333333	5 Member HH	0.1666667	East Coast	0.0500000	Chinese	0.6833333
Over Weight	0.1500000	Nesting Families	0.1833333	Urban	0.8666667			North	0.3500000	Malay	0.3166667
Under Weight	0.1666667	Teens Aches	0.2833333			6 Member HH	0.0666667	South	0.1666667		
		Teens and Toddlers	0.0833333			7+ Member HH	0.0500000				
		Yankys	0.0333333								

From the above analyses on the subgroups of customers in the high income bracket, they are largely coming from smaller households, which could partly explain the lower consumption (hence lower expenditure).

Figure 18: Proportion of each columns for customers with R = 3, F = 3 and experienced an increase in M after MCO

Proportion				Proportion			
Healthy	0.4716981	[0, 1500)	0.1283019	Empty Nesters	0.0754717	1-3 Member HH	0.2754717
		[1500, 2000)	0.0754717	Matured Families	0.1811321	4 Member HH	0.2943396
Obese	0.1358491	[2000, 3000)	0.2150943	Nesting Families	0.1433962	5 Member HH	0.2490566
Over Weight	0.2603774	[3000, 4000)	0.1698113	Teens Aches	0.3547170	6 Member HH	0.1094340
Under Weight	0.1320755	[4000, 5000)	0.1207547	Teens and Toddlers	0.2188679	7+ Member HH	0.0716981
Yankys				0.0264151			
				Central			
				East Coast			
				North			
				South			
				Chinese			
				Malay			
				Others			

Therefore, we recommend to push more advertisements for smaller but more luxurious towards these customers that were identified as RFM = 331 and high income level. This is because we identify these customers as most likely to spend, but due to their demographic may not buy as much items.

4.2. Market Basket Analysis

We thus now turn to analysing the market baskets of the 2 groups to explain the discrepancy and hopefully obtain some insight as to how we can promote greater spending. After generating the MBA results for the two groups before and after MCO started with minimum support = 0.01 and minimum lift = 2, we then ranked them by lift.

Figure 19: preMCOHIHM

antecedents	consequents	antecedent.support	consequent.support	support	confidence	lift
CSD	Isotonic Drinks	0.0628210	0.0637302	0.0155007	0.2467438	3.871696
Isotonic Drinks	CSD	0.0637302	0.0628210	0.0155007	0.2432240	3.871696
Tea	Seasoning Powder	0.0641393	0.0717305	0.0170917	0.2664777	3.714983
Seasoning Powder	Tea	0.0717305	0.0641393	0.0170917	0.2382763	3.714983
Coconut Milk	Seasoning Powder	0.0752307	0.0717305	0.0196373	0.2610272	3.638997
Seasoning Powder	Coconut Milk	0.0717305	0.0752307	0.0196373	0.2737643	3.638997
Seasoning Powder	Margarine	0.0717305	0.0468203	0.0115460	0.1609632	3.437894
Margarine	Seasoning Powder	0.0468203	0.0717305	0.0115460	0.2466019	3.437894
Coconut Milk	Bouillon	0.0752307	0.0471385	0.0101368	0.1347432	2.858453
Bouillon	Coconut Milk	0.0471385	0.0752307	0.0101368	0.2150434	2.858453

Figure 20: preMCOHILM

antecedents	consequents	antecedent.support	consequent.support	support	confidence	lift
Flour	Sugar	0.0837964	0.1057290	0.0271945	0.3245305	3.069455
Sugar	Flour	0.1057290	0.0837964	0.0271945	0.2572093	3.069455
Tea	Condensed/Evap Milk	0.0389968	0.1069093	0.0119498	0.3064313	2.866274
Condensed/Evap Milk	Tea	0.1069093	0.0389968	0.0119498	0.1117755	2.866274
Tea	Sugar	0.0389968	0.1057290	0.0111630	0.2862547	2.707437
Sugar	Tea	0.1057290	0.0389968	0.0111630	0.1055814	2.707437
Cereals	Liquid Milk	0.0336366	0.1311040	0.0111630	0.3318713	2.531359
Liquid Milk	Cereals	0.1311040	0.0336366	0.0111630	0.0851463	2.531359
Sugar	Coconut Milk	0.1057290	0.0451930	0.0117040	0.1106977	2.449442
Coconut Milk	Sugar	0.0451930	0.1057290	0.0117040	0.2589771	2.449442

Figure 21: postMCOHIHM

antecedents	consequents	antecedent.support	consequent.support	support	confidence	lift
CSD	Isotonic Drinks	0.0560267	0.0621090	0.0137634	0.2456576	3.955266
Isotonic Drinks	CSD	0.0621090	0.0560267	0.0137634	0.2216004	3.955266
Seasoning Powder	Coconut Milk	0.0774364	0.0849089	0.0213054	0.2751346	3.240350
Coconut Milk	Seasoning Powder	0.0849089	0.0774364	0.0213054	0.2509210	3.240350
Seasoning Powder	Tea	0.0774364	0.0662797	0.0157445	0.2033214	3.067626
Tea	Seasoning Powder	0.0662797	0.0774364	0.0157445	0.2375459	3.067626
Coconut Milk	Bouillon	0.0849089	0.0501182	0.0112262	0.1322145	2.638055
Bouillon	Coconut Milk	0.0501182	0.0849089	0.0112262	0.2239945	2.638055
Flour	Seasoning Powder	0.1458015	0.0774364	0.0288127	0.1976162	2.551981
Seasoning Powder	Flour	0.0774364	0.1458015	0.0288127	0.3720826	2.551981

Figure 22: postMCOHILM

antecedents	consequents	antecedent.support	consequent.support	support	confidence	lift
Flour	Sugar	0.0808218	0.0974616	0.0283555	0.3508403	3.599781
Sugar	Flour	0.0974616	0.0808218	0.0283555	0.2909408	3.599781
Coconut Milk	Sugar	0.0377791	0.0974616	0.0101876	0.2696629	2.766864
Sugar	Coconut Milk	0.0974616	0.0377791	0.0101876	0.1045296	2.766864
Chocolate	Biscuits	0.0278462	0.2169115	0.0129892	0.4664634	2.150479
Biscuits	Chocolate	0.2169115	0.0278462	0.0129892	0.0598826	2.150479

It is observed that for high spenders, there is in general a slight increase in lift, which means that they may be sticking tighter to their routine of buying things that they always buy together. The number of rules with lift > 2 also increased after MCO started.

For the low spenders, there is a decrease in lift. This could be attributed to them purchasing randomly, i.e. hoarding, since they seldom had a fix regime of getting things previously when compared to the high spenders. Hence, this implies that they may be a little less prepared in terms of getting things. The number of rules also decreased largely after MCO started, suggesting that their market baskets are shrinking.

Our approaches therefore include two parts:

1. Compare the MBA results of pre- and post-MCO for high income high spenders and look for difference/change after MCO. The target is to retain this group of customers by giving certain promotions to items that can improve sales.
2. Compare the MBA results of pre- and post-MCO for high income low spenders and look for difference/change after MCO. The target is to shift their purchasing habits back to 'normal period', which is before MCO started.

We first look at the results of pre- and post-MCO for both groups:

```
## Warning in kable_styling(., full_width = F): Please specify format in
## kable. kableExtra can customize either HTML or LaTeX outputs. See https://
## haozhu233.github.io/kableExtra/ for details.
```

Figure 23: Pre-MCO HIHM

antecedents	counts
Seasoning Powder	9
Flour	8
Coconut Milk	7
Condensed/Evap Milk	7
Margarine	6
Tea	6

```
## Warning in kable_styling(., full_width = F): Please specify format in
## kable. kableExtra can customize either HTML or LaTeX outputs. See https://
## haozhu233.github.io/kableExtra/ for details.
```

Figure 24: Post-MCO HIHM

antecedents	counts
Coconut Milk	9
Flour	8
Sugar	7
Condensed/Evap Milk	6
Seasoning Powder	6
Bouillon	5

```
## Warning in kable_styling(., full_width = F): Please specify format in
## kable. kableExtra can customize either HTML or LaTeX outputs. See https://
## haozhu233.github.io/kableExtra/ for details.
```

Figure 25: Pre-MCO HILM

antecedents	counts
Sugar	5
Condensed/Evap Milk	2
Cooking Sauces	2
Liquid Milk	2
Tea	2
Bouillon	1
Cereals	1
Coconut Milk	1
Flour	1
Frozen Food	1
Ice Cream	1
Rice	1
Seasoning Powder	1
Yoghurts	1

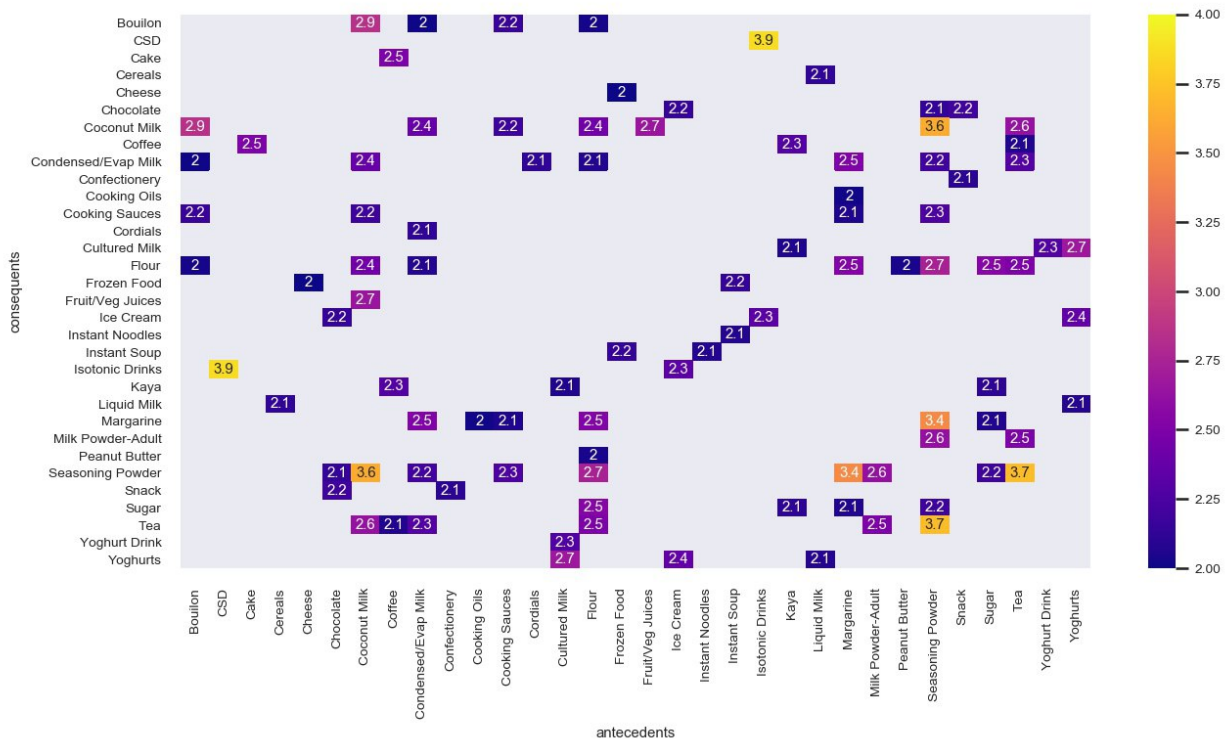
```
## Warning in kable_styling(., full_width = F): Please specify format in
## kable. kableExtra can customize either HTML or LaTeX outputs. See https://
## haozhu233.github.io/kableExtra/ for details.
```

Figure 26: Post-MCO HILM

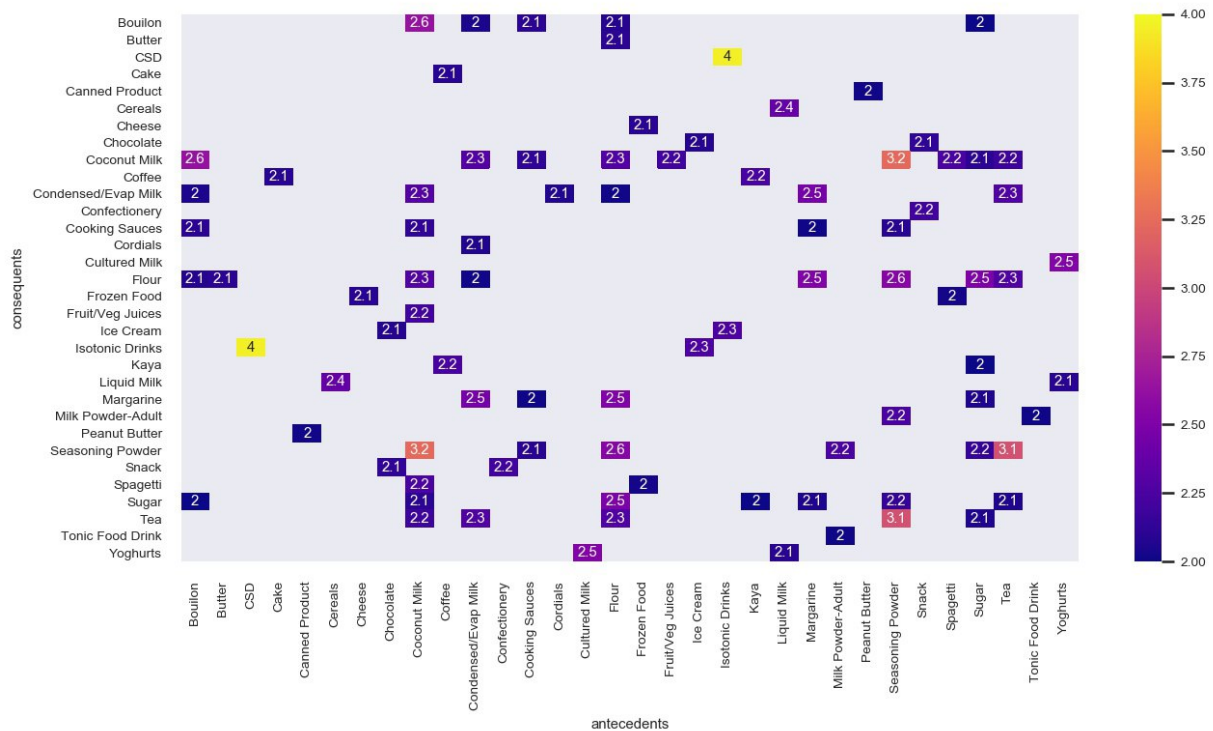
antecedents	counts
Sugar	2
Biscuits	1
Chocolate	1

antecedents	counts
Coconut Milk	1
Flour	1

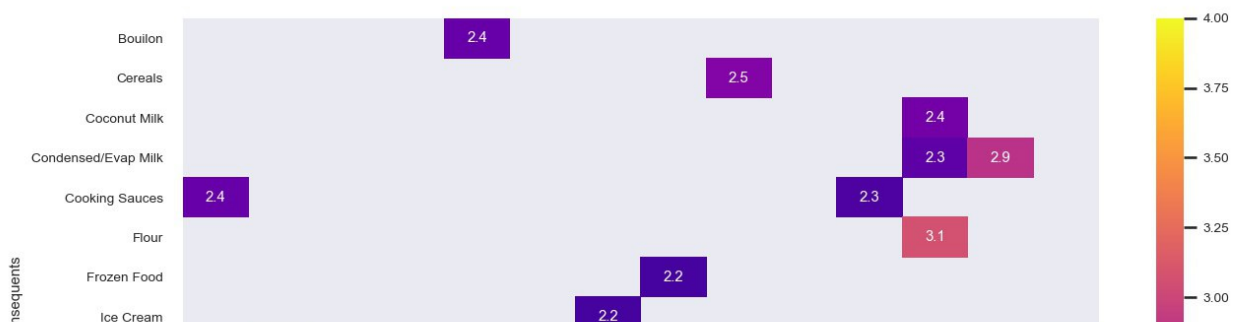
Lift of High Income and High Monetary Spending Consumers Pre-MCO

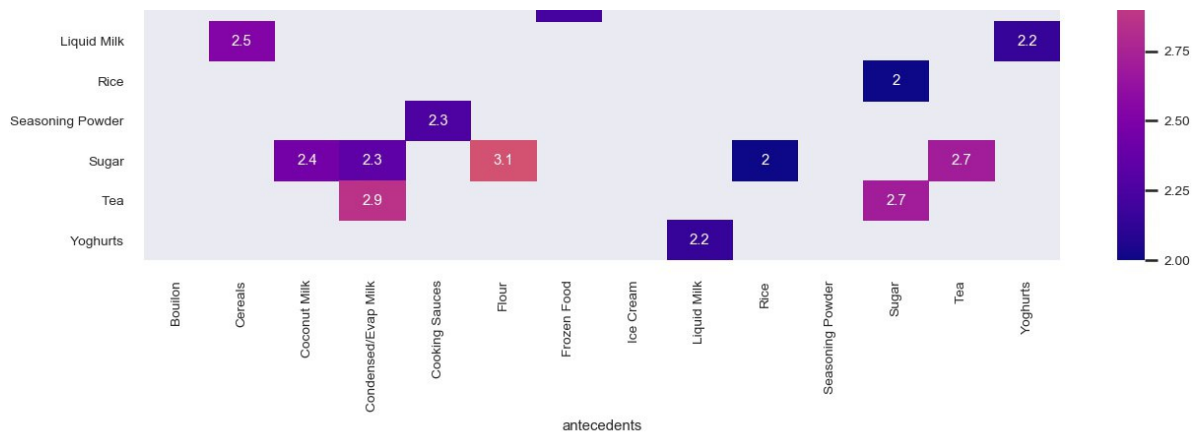


Lift of High Income and High Monetary Spending Consumers Post-MCO

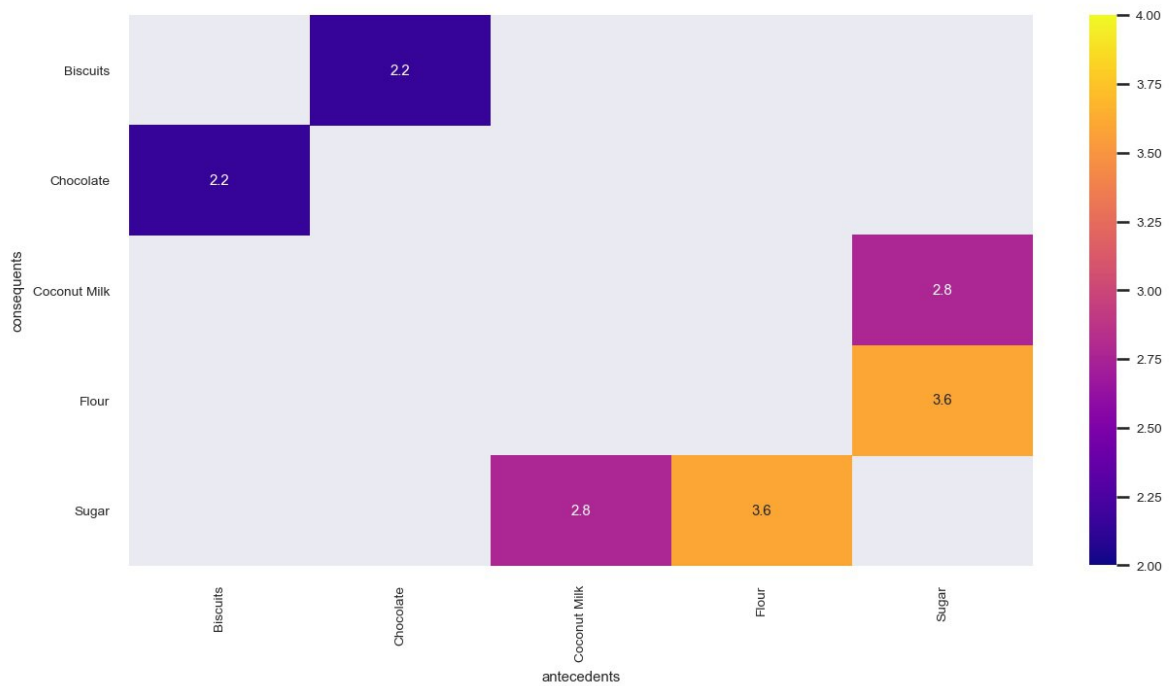


Lift of High Income and Low Monetary Spending Consumers Pre-MCO





Lift of High Income and Low Monetary Spending Consumers Post-MCO



From the figures and the tables, we could see that:

1. For high spenders, items like Coconut Milk and Sugar have experienced an increase in the number of antecedent counts, indicating that customers are purchasing a larger variety of goods in tandem with these items.
2. For low spenders, it is observed that while these customers often buy sugar together with things such as tea and milk before MCO started. However, they are not buying them together so often now.

5. Conclusion

1. We observe that most of the customers with higher purchasing power are in the central and southern location. Therefore, we believe that if we have increased pricing options in the central and southern regions as compared to the rest of the regions these customers tend to be more affluent.
2. We observe that there are a large group of customers that have high affluence and have been recent spenders but are not spending as much. So we propose to target customers with RFM score of 331 and promote advertisements of more luxurious items to increase the profit/customer for those that are more affluent. This can be done by giving them thank you notes or more discount vouchers on the higher priced items vs lower priced items to reward this customer behaviour.
3. We observe that high spenders have been purchasing more after MCO. To encourage them to continue with this behaviour, we recommend to give promotions to items such as coconut milk and sugar, to further increase spending from high-income-high-spending customers. We also observe that low spenders are purchasing less after MCO. It is thus recommended to place sugar and items such as tea and milk together to encourage high-income-low-spending customers to shift back to their normal purchasing styles, or hold campaigns for this kind of item sets, such as bundle sales.

6. Bibliography

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8(1), 3–15.
<https://doi.org/10.1177/096228029900800102> (<https://doi.org/10.1177/096228029900800102>)

Anil Jadhav, Dhanya Pramod & Krishnan Ramanathan (2019) Comparison of Performance of Data Imputation Methods for Numeric Dataset, *Applied Artificial Intelligence*, 33:10, 913-933, DOI: 10.1080/08839514.2019.1637138