ISSS602 Data Analytics Lab G3

AY2020-21T1

18 Oct 2020

Assignment 2: Be Customer Wise or Otherwise

An Insight to Philippine Household Grocery List

Prepared by Gabriella Pauline DJOJOSAPUTRO

1. Overview

Businesses strive to improve and grow. One of the strategies to achieve this is market development—selling existing products or services to a new group of customers. Going into a new market incurs cost, so market research is first conducted to identify the target market that should be pursued[1]. Effort should be concentrated on the segments which are likely to have the most profitability and opportunity to grow. The segments can also inform the promotional strategy to attract the customers based on their shared characteristics, so the efforts can be efficiently translated into business value.

This report is created for an international hypermarket retail that is going to develop their market in the Philippines. The potential customers are households in the Philippines, which are going to be segmented based on publicly available income and expenses data. The aim of this report is to identify the most profitable segment and develop profiles of the segments to inform market development strategy to attract the customers who belong to each segment.

2. Data Preparation

2.1. Data sources

Data for the analysis is retrieved from the Family Income and Expenditure Survey (FIES) conducted by the government of Philippines[2] in 2018. The survey comprises of items regarding income and expenses that was distributed to a sample of 170,917 households.

2.2. Selecting variables that are relevant for the analysis

As the analysis is performed on a public data, not all variables are relevant. Metadata and the questionnaire for the survey is examined to determine the measures to be included in the analysis. These variables are shown in Table 1. Variables that include some items which may not be sold in retail are excluded.

---

[1] nibusinessinfo.co.uk. (n.d.). *Assess your options for business growth: Market development strategy.* Retrieved October 15, 2020 from https://www.nibusinessinfo.co.uk/content/market-development-strategy.

[2] Mapa, C. D. S. (2020). *2018 Family Income and Expenditure Survey*. Philippines Statistics Authority. Retrieved October 5, 2020 from https://psa.gov.ph/sites/default/files/FIES%202018%20Final%20Report.pdf.

No	Variable	Description	Reason
1	W_REGN	Region the households belong to.	Useful to show the distribution of clusters in each region.
2	TOINC	Total income.	Likely to influence spending pattern.
3	FSIZE	Family size.	
4	BREAD	Expenses on rice, corn, flour, cereals, bread, and pasta.	Items sold in retail.
5	MEAT	Expenses on meat, whether raw or processed.	
6	FISH	Expenses on fish and seafoods, whether raw or processed.	
7	MILK	Expenses on milk, yogurt, cheese, and eggs.	
8	OIL	Expenses on butter, margarine, edible oil, and edible animal fats.	
9	FRUIT	Expenses on fruits and nuts.	
10	VEG	Expenses on vegetables, including tuber vegetables (e.g. potato, cassava) and processed vegetable products.	
11	SUGAR	Expenses on sugar, jams, honey, chocolates, and confectionery.	
12	FOOD_NEC	Expenses on salt, spices and culinary herbs, sauces, condiments, seasonings, and processed baby foods.	
13	COFFEE	Expenses on coffee, tea, and cocoa-based drinks.	
14	MINERAL	Expenses on mineral water, soft drinks, fruit and vegetable juices.	
15	ALCOHOL	Expenses on liquor, wine, and beer.	
16	TOBACCO	Expenses on cigarettes, cigars, and other tobacco products.	
17	OTHER_VEG	Expenses on betel leaves, betel nuts, mint leaf, lime, and other vegetable-based products.	
18	CLOTHING	Expenses on clothing and footwear.	

Table 1. Relevant variables from FIES for the analysis

2.3. Deriving variables that are useful for the analysis

Variables that are derived to aid the analysis are:

1. PCINC_mine

   Per capita income. It is available in the original dataset but it is recalculated for accuracy

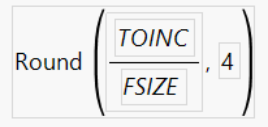


Figure 1. PCINC_mine formula

2. SHOPPING_EX

   A sum of all shopping expenses (items that are sold in retail).

Figure 2. SHOPPING_EX formula

3. EX/INC

   Ratio of shopping expenses to income.

Figure 3. EX/INC formula

2.4. Identifying data issues and clean the data accordingly

1. Numerical encoding

   W_REGN is listed as continuous variable, but it is a nominal variable because it represents different regions. The values are recoded into the region name for easier understanding. The variable type is automatically adjusted during recoding.

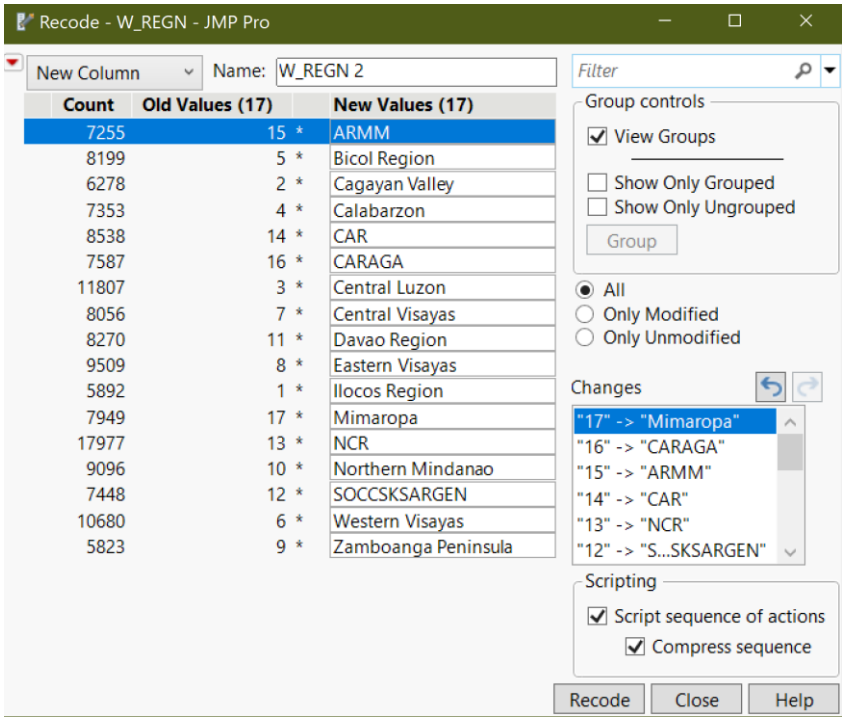


Figure 4. W_REGN recoding process

2. Incomplete data

   Records that have expenses only in very few categories are suspected to be caused by incomplete data. The respondents may not be willing to disclose the values or forgot the amount they have spent on the categories, hence leaving the items blank.

   A variable called Non-missing Value is derived to count the number of shopping expense category which value is not zero.

Figure 5. Non-missing Value formula

Records that has expenses in only 3 categories or less are excluded from the analysis. The distribution of the expenses with only 3 categories of expenses does not seem reasonable to sustain one's life.

BREAD	MEAT	FISH	MILK	OIL	FRUIT	VEG	SUGAR	FOOD_NEC	COFFEE	MINERAL	ALCOHOL	TOBACCO	OTHER_VEG	CLOTH	Non-missing Value
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	25320	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	7800	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	3820	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	3524	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	2170	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	2150	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1420	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	500	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	200	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	1
0	0	0	0	0	0	0	0	0	0	0	0	2896	0	0	1
0	0	0	0	0	0	0	0	0	0	650	0	0	0	0	1
0	0	0	0	0	240	0	0	0	0	0	0	0	0	7647	2
0	0	0	0	0	0	0	0	0	0	0	1430	0	0	6590	2
0	0	0	0	0	0	0	0	0	0	0	1248	0	0	5320	2
0	0	0	0	0	0	0	0	0	0	0	0	25520	0	4400	2
0	0	0	0	0	0	0	0	0	0	0	0	1865	0	2608	2
0	0	0	0	0	0	0	0	0	0	0	0	3620	0	1210	2
0	0	0	0	0	0	0	0	0	1086	0	0	0	0	660	2
0	0	0	0	0	200	0	0	0	0	0	0	0	0	400	2
0	0	0	2160	0	0	0	0	0	0	0	0	0	0	396	2
0	0	0	10660	0	3180	0	0	0	0	0	0	0	0	0	2
0	0	0	2715	0	0	0	0	0	2896	0	0	0	0	3750	3
0	0	0	0	0	0	0	0	0	0	715	0	6380	0	2750	3
0	0	0	0	0	0	0	0	0	1274	650	0	0	0	1800	3
0	0	0	0	0	0	0	0	0	0	0	8840	920	0	1300	3
0	0	0	2400	0	0	0	0	0	0	520	0	0	0	1060	3
0	0	0	0	0	0	0	0	0	858	0	0	12820	0	270	3
0	0	0	2118	0	450	0	0	0	0	1360	0	0	0	38439	4
0	0	0	0	0	0	0	0	0	1288	150	2340	0	0	3160	4

Figure 6. Records with values in only 3 categories or less

Records that have zero values for BREAD, MEAT, and FISH (59 records) are also excluded for the same reason.

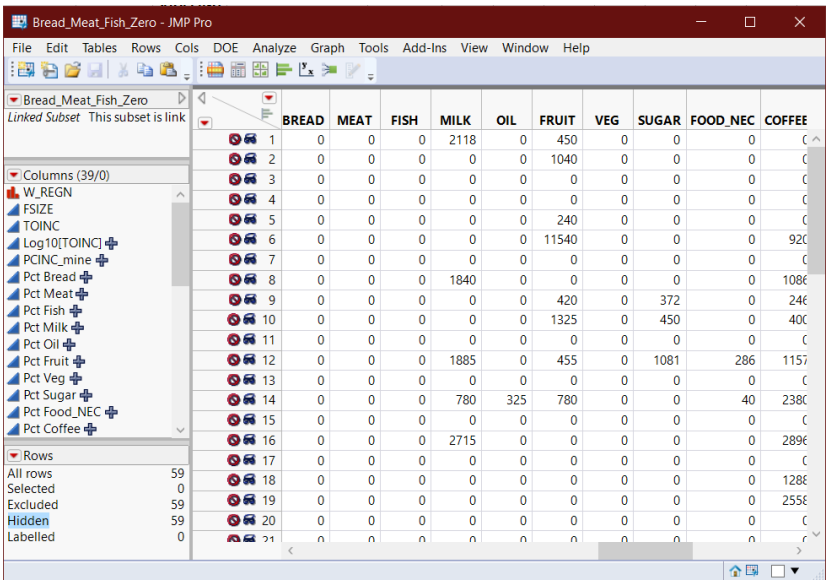


Figure 7. Records with zeroes for BREAD, MEAT, and FISH

3. Extreme outliers

   Although there are some outliers in the data, it is not a good practice to simply exclude the data point because it is an outlier[3]. Valid statistical outliers represent the variability in the population, so it should not be excluded. Distribution analysis and exploration was carried out to determine whether the records should be excluded from the analysis, as shown in Figure 9 and 10. However, because different households may exhibit different patterns of income and expenses, it is difficult to draw conclusions. To reduce the impact of these outliers, the expenses are going to be transformed as percentage of the total expenses. Aside from reducing the effect of outliers, using percentage of the expenses allows us to examine more clearly on what category of items the households spend more money. The percentage is calculated for items 4 to 18 in Table 1.

$$\text{Round}\left(\left(\frac{BREAD}{SHOPPING\_EX}\right)\cdot 100\ ,\ 2\right)$$

Figure 8. Pct Bread formula

[3] Frost, J. (2020, Jun 6). Guidelines for Removing and Handling Outliers in Data. Retrieved from *Statistics By Jim*: https://statisticsbyjim.com/basics/remove-outliers/

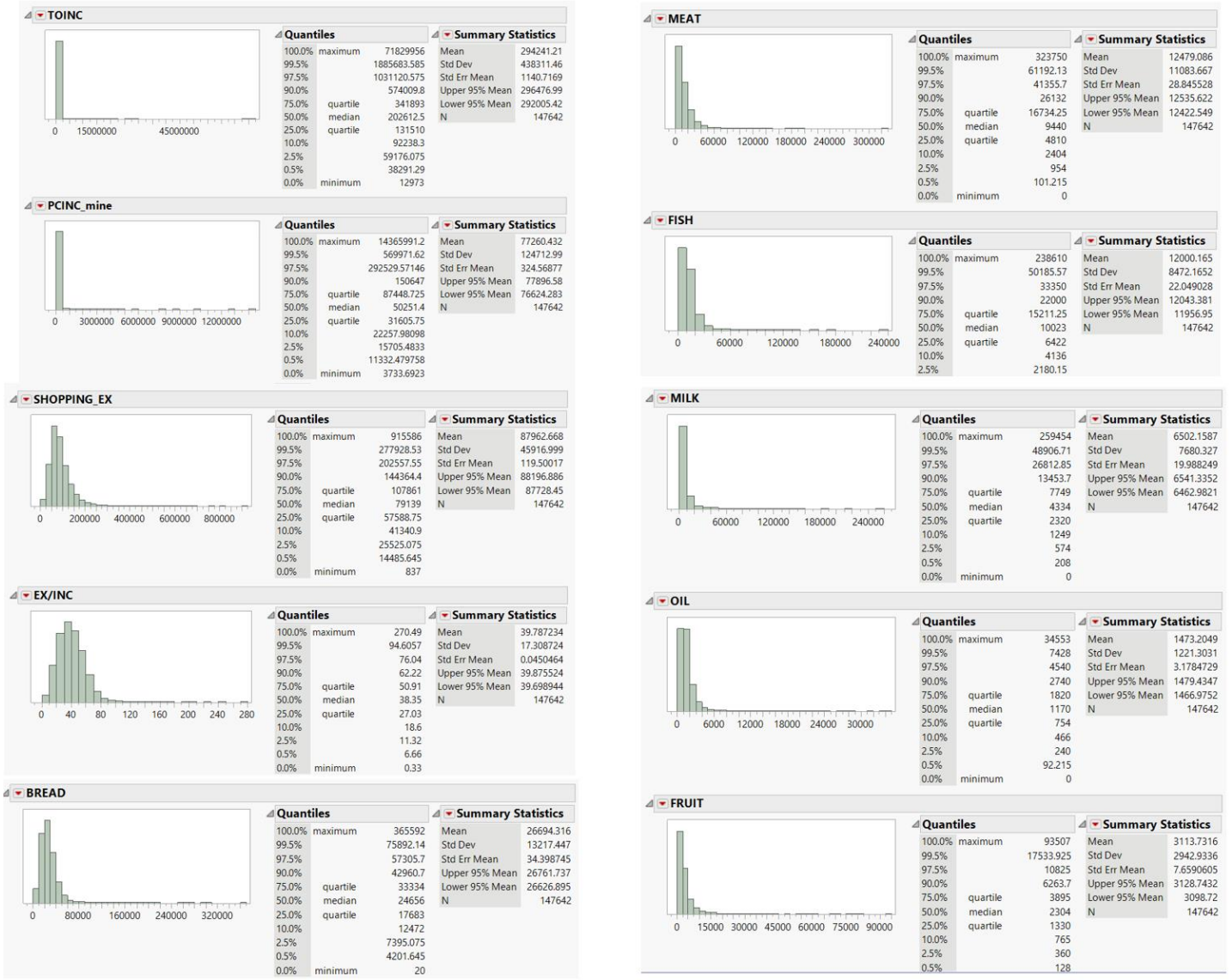
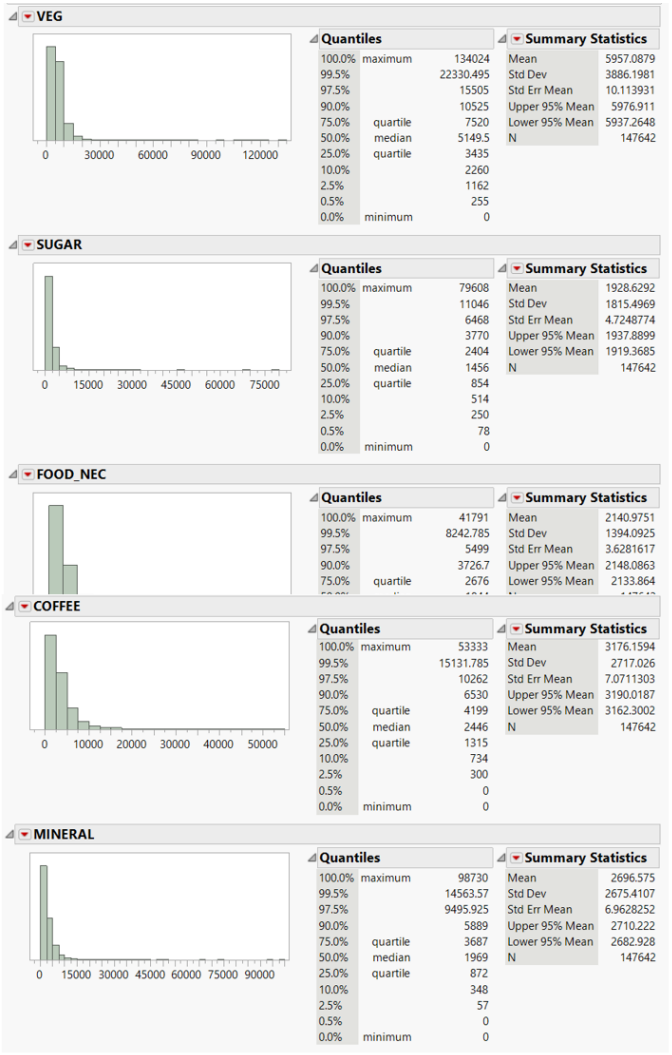


Figure 9. Distribution Analysis (pt. 1)



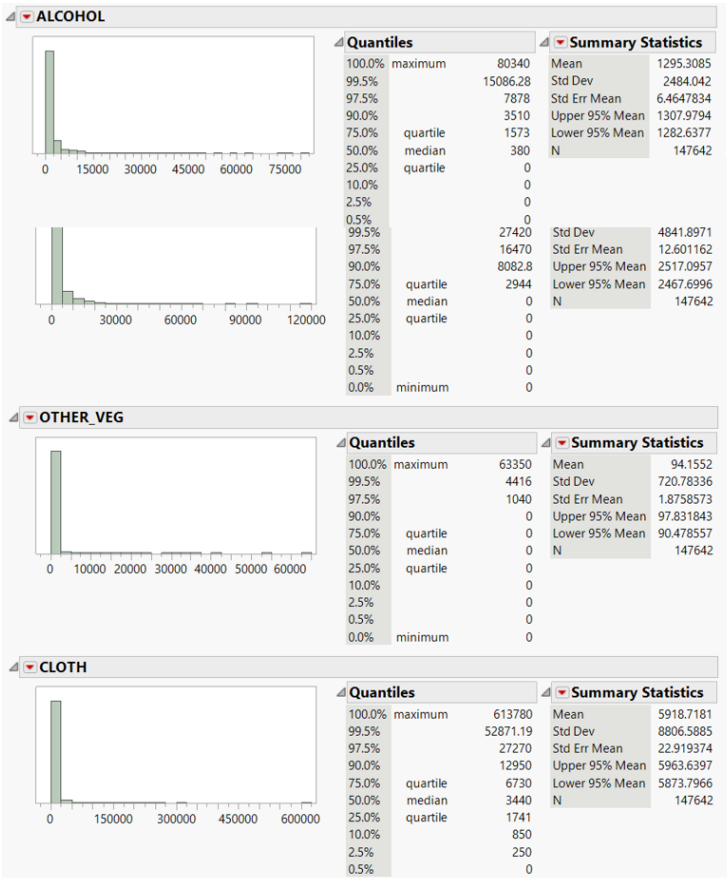


Figure 10. Distribution Analysis (pt. 2)

4. Skewed distribution and unstandardized range

   The remaining variables that are not in percentage (TOINC, SHOPPING_EX, and PCINC_mine) are shown to be highly right-skewed with a very large range. Natural log transformation is used to standardize these variables. The ranges for the three variables are reduced from millions to less than 10 with an approximately more normal distribution.

25/1 Cols	W_REGN	FSIZE	TOINC	Log[TOINC]	PCINC_mine	Log[PCINC_mine]	SHOPPING_EX	Log[SHOPPING_EX]
	13 3 6 8 10 12 others	22 1	71.8M 13k	18.1 9.47	14.4M 3734	16.5 8.23	916k 837	13.7 6.73
1	4	6	3469740	15.059590221	578290	13.267830752	915586	13.727319576
2	3	4	6148498	15.631718382	1537124.5	14.245424021	685551	13.437978173
3	8	8.5	5879735	15.587022251	691733.5...	13.446956087	671519	13.417297589
4	7	7.5	1746096	14.372892997	232812.8	12.357989976	518107	13.157937064
5	13	8	3087700	14.942937035	385962.5	12.863495493	762259	13.544041672
6	6	4	7106670	15.776544338	1776667.5	14.390249977	394191	12.884590842
7	6	4	6268840	15.651101888	1567210	14.264807526	596001	13.297997624
8	6	3	2628260	14.781832588	876086.6...	13.6832203	411440	12.92741848

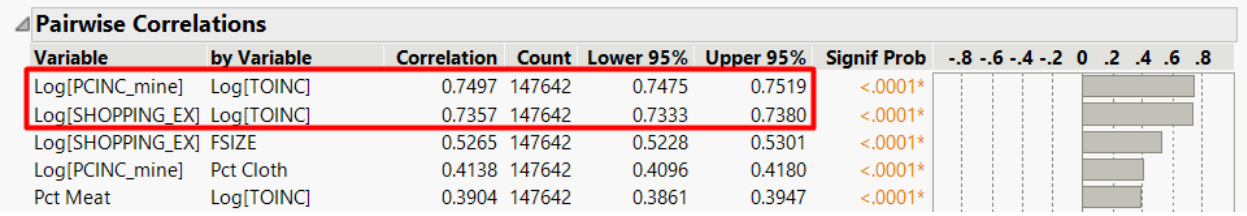
Figure 11. Log transformation for TOINC, PCINC_mine, and SHOPPING_EX

3. Data Analysis

3.1. K-means

3.1.1. Variables Selection

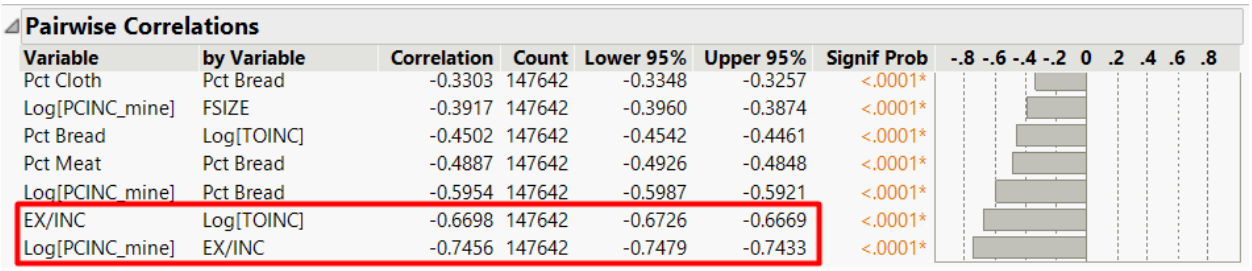
Multicollinearity can affect the clustering results by giving a higher weight for certain variables that are highly correlated. Therefore, multivariate analysis was conducted on FSIZE, transformed TOINC, SHOPPING_EX, and PCINC_mine, as well as EX/INC and the expense percentages to exclude variables that have moderately strong correlation ($|r| > 0.6$).



Pairwise Correlations

Variable	by Variable	Correlation	Count	Lower 95%	Upper 95%	Signif Prob
Log[PCINC_mine]	Log[TOINC]	0.7497	147642	0.7475	0.7519	<.0001*
Log[SHOPPING_EX]	Log[TOINC]	0.7357	147642	0.7333	0.7380	<.0001*
Log[SHOPPING_EX]	FSIZE	0.5265	147642	0.5228	0.5301	<.0001*
Log[PCINC_mine]	Pct Cloth	0.4138	147642	0.4096	0.4180	<.0001*
Pct Meat	Log[TOINC]	0.3904	147642	0.3861	0.3947	<.0001*

Figure 12. Moderately strong positive correlations



Pairwise Correlations

Variable	by Variable	Correlation	Count	Lower 95%	Upper 95%	Signif Prob
Pct Cloth	Pct Bread	-0.3303	147642	-0.3348	-0.3257	<.0001*
Log[PCINC_mine]	FSIZE	-0.3917	147642	-0.3960	-0.3874	<.0001*
Pct Bread	Log[TOINC]	-0.4502	147642	-0.4542	-0.4461	<.0001*
Pct Meat	Pct Bread	-0.4887	147642	-0.4926	-0.4848	<.0001*
Log[PCINC_mine]	Pct Bread	-0.5954	147642	-0.5987	-0.5921	<.0001*
EX/INC	Log[TOINC]	-0.6698	147642	-0.6726	-0.6669	<.0001*
Log[PCINC_mine]	EX/INC	-0.7456	147642	-0.7479	-0.7433	<.0001*

Figure 13. Moderately strong negative correlations

The four pairs of monetary variables are moderately correlated. The positive correlations are shown in Figure 12, while Figure 13 shows negative correlations. Because the expenses are shown in terms of percentage from SHOPPING_EX, this variable cannot be excluded from the analysis. The transformed value of TOINC is has correlation coefficient of 0.75 with transformed SHOPPING_EX, so it is excluded from the analysis. EX/INC is retained to show the relationship with the income instead, which means PCINC_mine also needs to be excluded because it has moderate negative correlations to EX/INC (-0.75).

3.1.2. Computing clusters

The variables highlighted in blue in Figure 14 are used to compute the clusters. Columns are scaled individually because the range of the variables greatly differs even though they have been made

into percentages and log-transformed. Using the same scale would yield negative CCC values, indicating the data is not normally distributed or skewed, and have a lot of outliers (see Appendix A). As the optimum number of clusters is unknown, a range of 3 to 12 is specified (Figure 15).

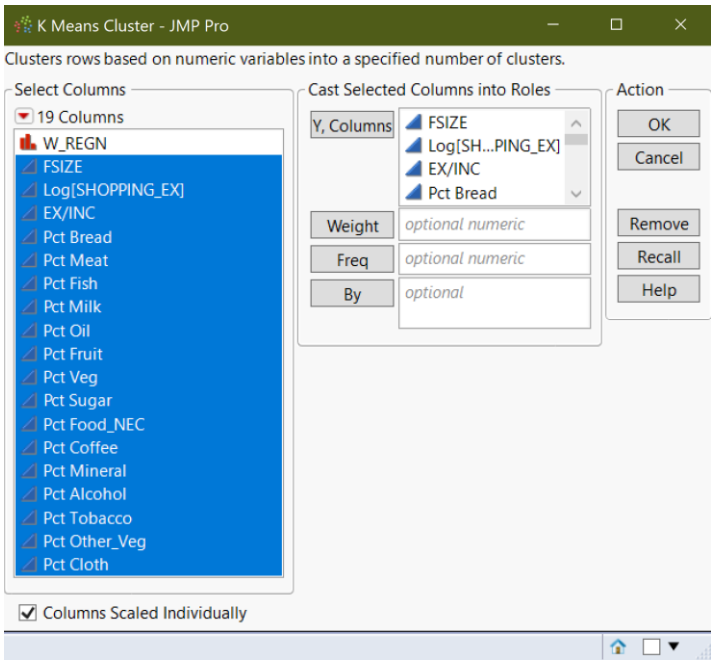


Figure 14. Clustering action window (pt. 1)

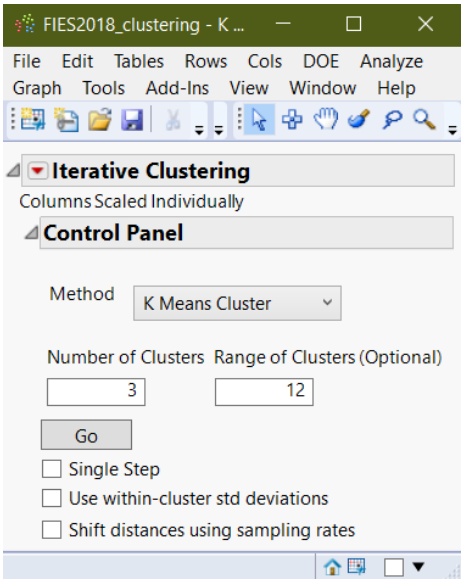


Figure 15. Clustering action window (pt. 2)

3.1.3. Selecting number of clusters

Cubic Clustering Criterion (CCC) is a measure of fit for the number of clusters for techniques that minimizes the within-cluster sum of squares[4]. The CCC values reaches a local maximum when the number of clusters is 10, so it is selected as the chosen number of clusters. Line graph is drawn to illustrate the progression of the CCC values as the clusters increase (Figure 17).

Iterative Clustering

Cluster Comparison

Method	NCluster	CCC	Best
K Means Cluster	3	-144.17	
K Means Cluster	4	-249.99	
K Means Cluster	5	-173.04	
K Means Cluster	6	-101.68	
K Means Cluster	7	-73.872	
K Means Cluster	8	-17.455	
K Means Cluster	9	4.77837	
K Means Cluster	10	53.9385	Optimal CCC
K Means Cluster	11	-7.9182	
K Means Cluster	12	75.7995	

Columns Scaled Individually

Figure 16. CCC values for 3 to 12 K-means clusters

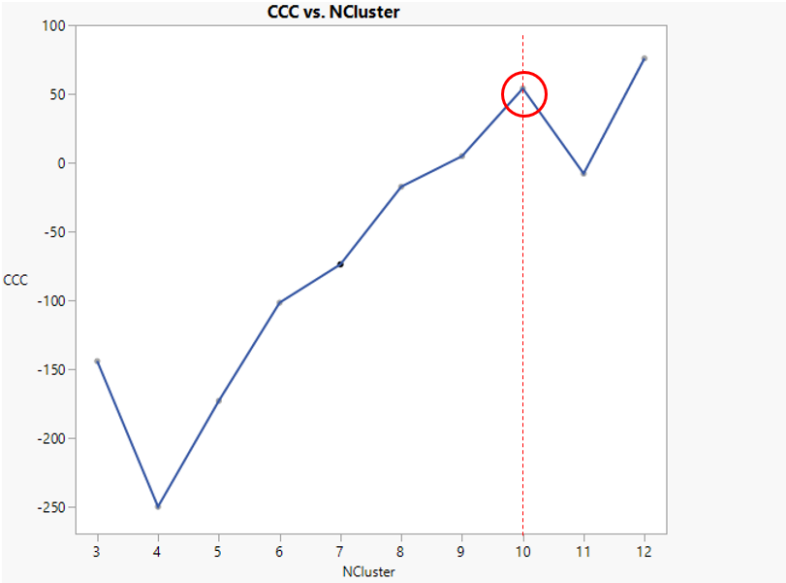
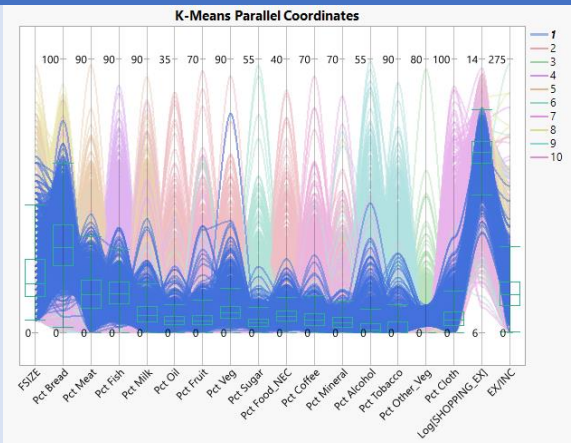
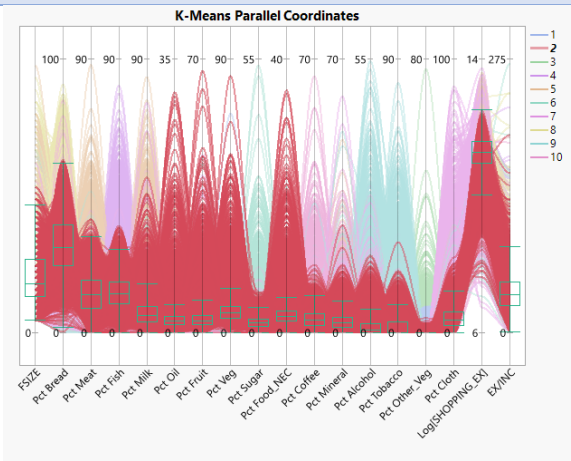
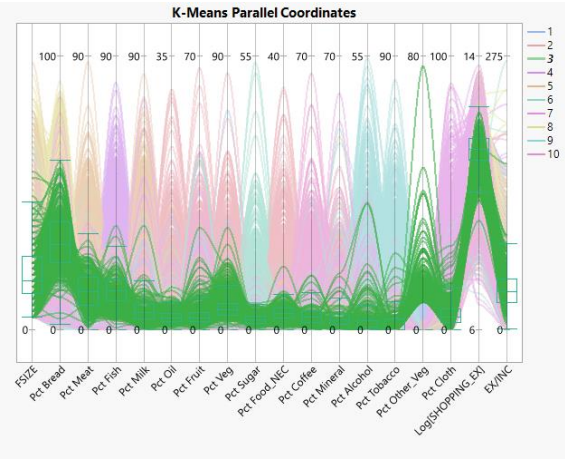


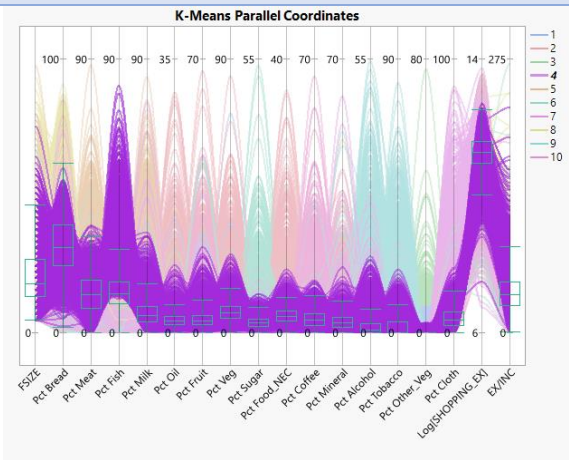
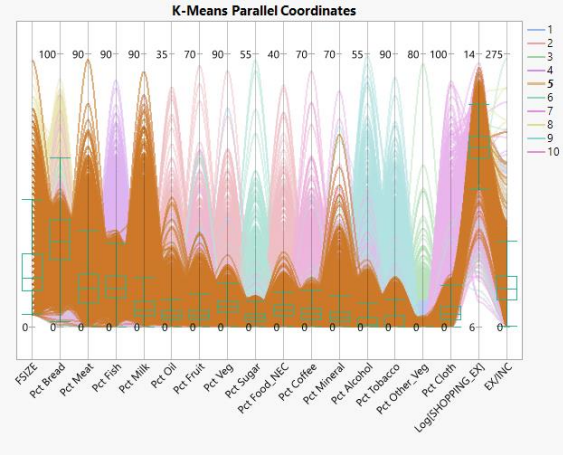
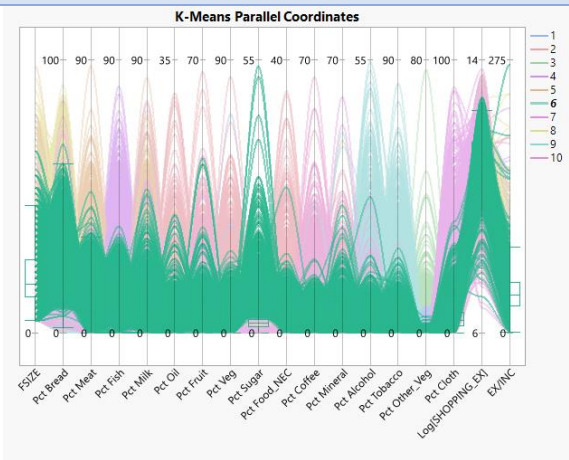
Figure 17. Line graph of CCC by number of clusters

[4] SAS Institute. (1983). *SAS Technical Report A-108, Cubic Clustering Criterion*. Cary, USA: SAS Institute Inc. Retrieved from https://support.sas.com/documentation/onlinedoc/v82/techreport_a108.pdf

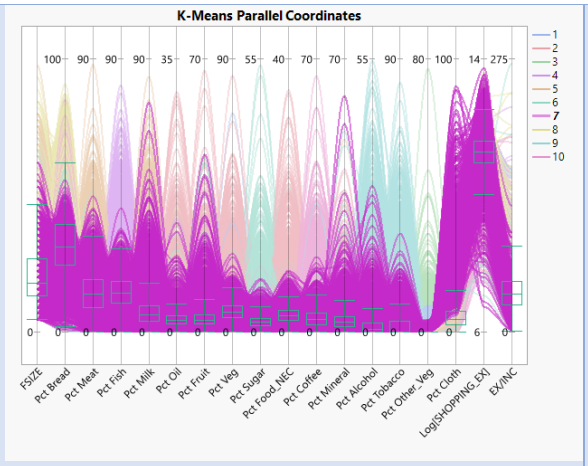
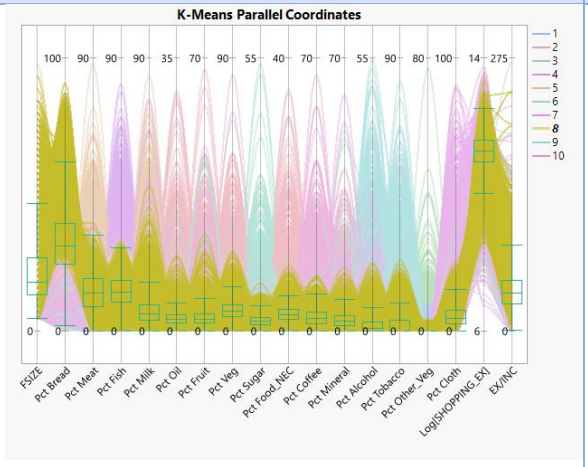
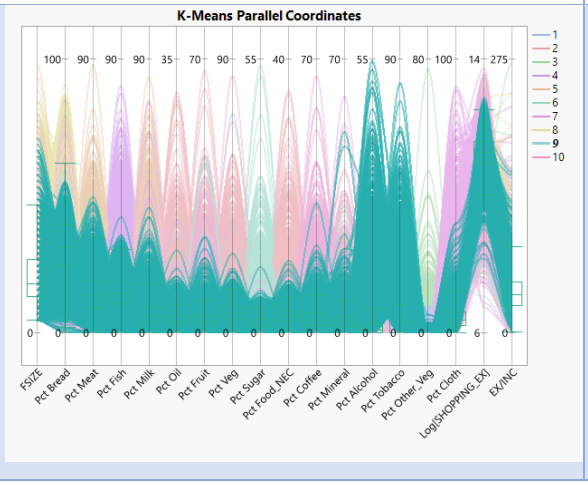
3.1.4. Clustering results

The clusters are saved and examined using parallel coordinates plot to determine the characteristic of each clusters (Table 2).

No	Parallel Coordinates	Description
1		**Average Family**: Distribution for all variables are roughly around the median. Spending pattern is just like majority of other households.
2		**Masterchef Family:** Spend more proportion of their shopping expenses on cooking ingredients (oil, fruit, vegetable, and sauces) than the rest.
3		**Traditional Household:** This group characterized by the high percentage of expense for other vegetables, which includes betel nut and leaves. Chewing betel leaves is a cultural custom in Philippines that is still

		practiced in the rural areas and by the older generations[5]
4		**Seafood Lovers**: Spends more on fish compared to other categories.
5		**Large Family with Baby:** Households with larger family size that tend to spend more on meat, milk, oil, fruit, and bottled drinks, but less on breads. High spending on milk shows likelihood of having babies in the households. Tends to have higher shopping expenses.
6		**Sweet Tooth:** Allocates more spending for sugar, jams, chocolates, and other sweets.

[5] Legarda, M. (2016, June 12). Asia's Crimson Addiction (Betel). Retrieved from illumelation: https://www.illumelation.com/blog/betel#:~:text=Betel%20Nuts%20in%20the%20Philippines,as%20bua%2C%20maman%20or%20mama.

7		**Fashionista:** Tends to have higher overall shopping expenses and very high spending on clothing. Also tend to spend more on milk, fruit, bottled drinks, alcohol and tobacco.
8		**Low Income Large Family:** Large families that has medium-low shopping expense, but a high expense to income ratio. Spends mostly on bread.
9		**Alcohol and Tobacco:** Characterized by strikingly high percentage spent on alcohol and tobacco.

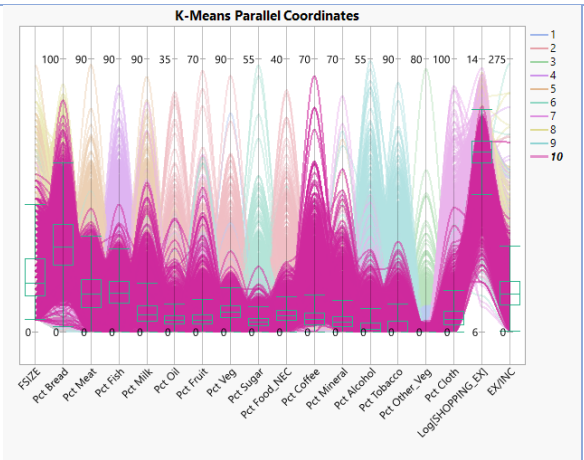
10		**Busy Workers:** Slightly lower total shopping expenses, but a high proportion of it goes to coffee and bottled drinks.

Table 2. K-means clusters parallel coordinate plots and characteristics

3.2. Latent Class Analysis

3.2.1. Binning of variables

The selected variables are binned into roughly equal quartiles (due to rounding). However, there are exceptions for FSIZE, ALCOHOL, TOBACCO, and OTHER_VEG.

The recommended cut points for equal binning is at 3, 4, and 6 for FSIZE, but spending pattern is likely to be different at the cut points 1 and 2 because they signify singles and couples. The recommended cut point at 6 is retained.

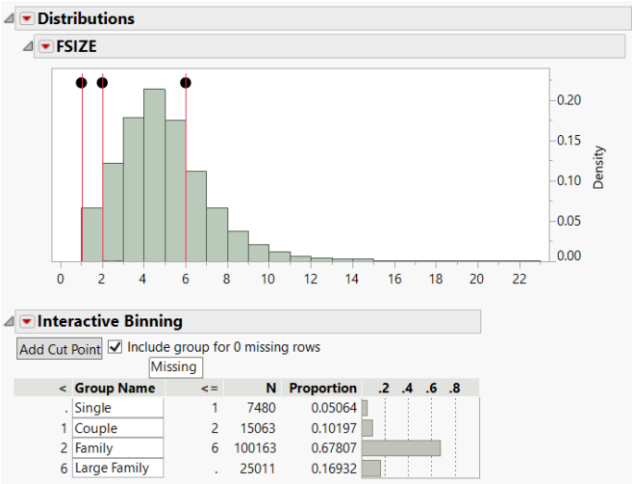


Figure 18. Interactive binning for FSIZE

There are 38% records which has 0 for Pct Alcohol, so the remaining records are split into two roughly equal bins (Figure 19). The same logic follows for Pct Tobacco which has 50.8% records with the value 0 (Figure 20).

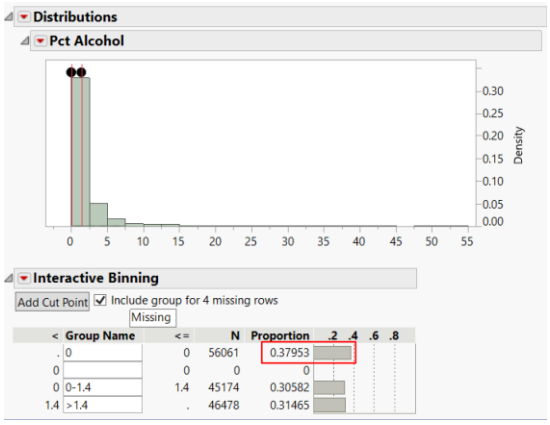


Figure 19. Interactive binning for Pct Alcohol

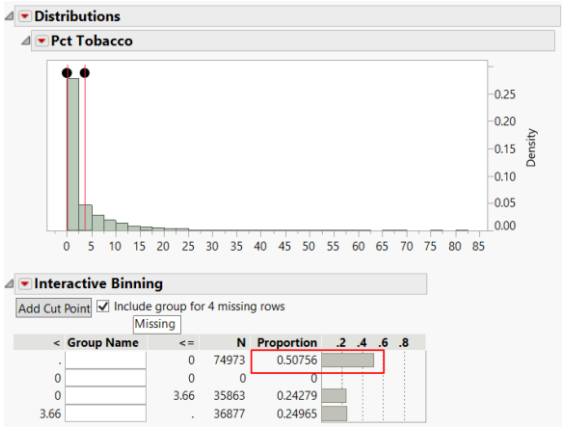


Figure 20. Interactive binning for Pct Tobacco

The percentage of zeroes is much higher in Pct Other_Veg at 94.7% (Figure 21), so this variable is not going to be included in the latent class analysis (LCA).

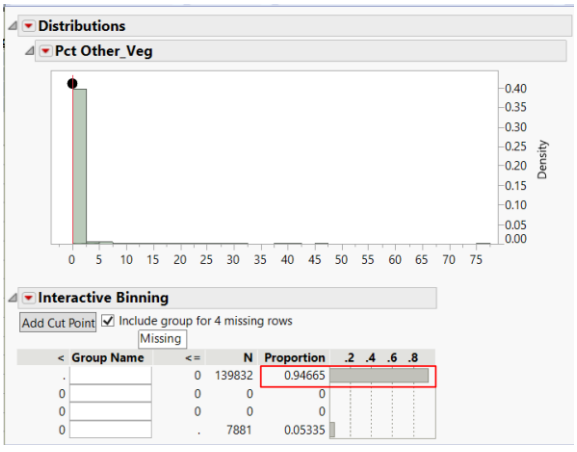


Figure 21. Interactive binning for Pct Other_Veg

3.2.2. Selecting number of clusters

Scree plot is used to determine the optimum number of clusters because the number keeps on decreasing (Figure 22). The values of AIC and BIC seem to have a slight shift of direction in 6 clusters and 10 clusters (Figure 23). Ten is chosen as the optimum number of clusters to maintain consistency with the results observed from K-means clustering techniques, with percentage of expenditure on bread, shopping expenses, and family size as the top three factors that affect the clustering for LCA (Figure 24).

Latent Class Analysis

Cluster Comparison

NCluster	-LogLikelihood	BIC	AIC	Best
8	3178592	6361934	6357982	
9	3173595	6352535	6348088	
10	3168374	6342688	6337747	
11	3164735	6336005	6330569	
12	3161464	6330058	6324127	
13	3158671	6325066	6318639	
14	3156130	6320580	6313658	Smallest BIC Smallest AIC

Figure 22. AIC and BIC for LCA

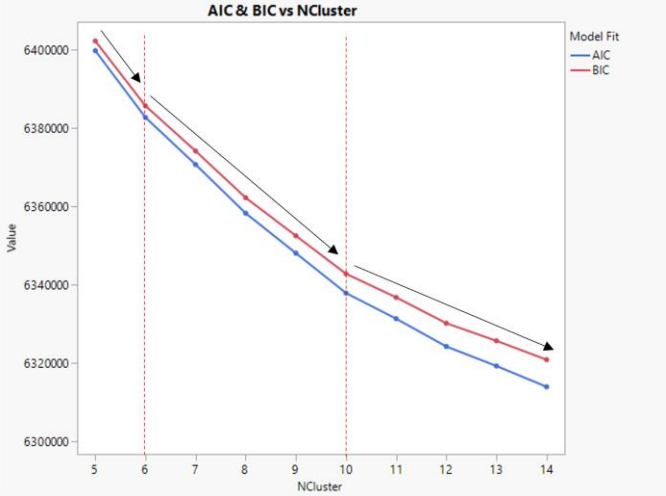


Figure 23. AIC and BIC Scree plot

Effect Sizes

Column	Effect Size	LR Logworth
Pct Bread Groups	1.0309	36951
SHOPPING_EX Groups	0.8852	27427
FSIZE Groups	0.8068	17793
Pct Meat Groups	0.6553	14599
EX/INC Groups	0.5516	9811.2
Pct Veg Groups	0.5065	8435.1
Pct Food_NEC Groups	0.4865	7862.6
Pct Cloth Groups	0.4809	6849.9
Pct Tobacco Groups	0.4763	7365.8
Pct Alcohol Groups	0.4241	5907.3
Pct Mineral Groups	0.4109	5535.9
Pct Fruit Groups	0.3987	5290.7
Pct Milk Groups	0.3702	4381.7
Pct Oil Groups	0.3662	4442.6
Pct Fish Groups	0.3005	2663.2
Pct Sugar Groups	0.2996	2879.8
Pct Coffee Groups	0.2067	1345.5

Figure 24. LCA Variable Effect Size

3.2.3. Clustering results

The class membership probabilities and item response probabilities are shown in Figure 25 and further illustrated in Figure 26.

Parameter Estimates

		Pct Fruit Groups				Pct Veg Groups				Pct Sugar Groups				Pct Food_NEC Groups				Pct Coffee Groups				Pct Mineral Groups				Pct Alcohol Groups			Pct Tobacco Groups			Pct Cloth Groups			
Cluster	Overall	<=2	2-3	3-4.5	>4.5	<=5	5-6.5	6.5-8.5	>8.5	<=1.2	1.2-1.9	1.9-2.8	>2.8	<=1.7	1.7-2.3	2.3-3.1	>3.1	<=1.8	1.8-3	3-4.9	>4.9	<=1.3	1.3-2.5	2.5-4	>4	0	0-1.4	>1.4	0	0-3.7	>3.7	<=2.6	2.6-4.5	4.5-7.7	>7.7
Cluster 1	0.15037	0.2691	0.2882	0.2731	0.1696	0.2074	0.2614	0.2919	0.2393	0.1908	0.2885	0.2845	0.2362	0.1294	0.2704	0.3111	0.2892	0.2322	0.2391	0.2650	0.2637	0.1955	0.2587	0.2599	0.2859	0.3805	0.3503	0.2692	0.6424	0.2169	0.1408	0.2155	0.2776	0.3026	0.2043
Cluster 2	0.13179	0.3773	0.3002	0.2127	0.1098	0.4917	0.2715	0.1637	0.0730	0.3421	0.2933	0.2176	0.1470	0.4127	0.3054	0.2000	0.0818	0.1981	0.2132	0.2824	0.3064	0.1273	0.2589	0.3187	0.2951	0.1331	0.3549	0.5120	0.2394	0.2717	0.4889	0.2468	0.2727	0.2704	0.2101
Cluster 3	0.12320	0.1408	0.2228	0.3108	0.3256	0.0634	0.2073	0.3434	0.3860	0.1820	0.2482	0.2701	0.2997	0.0405	0.1907	0.3409	0.4280	0.2043	0.2471	0.3020	0.2466	0.0817	0.2077	0.3284	0.3822	0.2794	0.3174	0.4032	0.5139	0.2202	0.2659	0.1880	0.2774	0.3247	0.2098
Cluster 4	0.11988	0.4697	0.2670	0.1722	0.0911	0.4721	0.2385	0.1715	0.1179	0.2862	0.3116	0.2344	0.1678	0.4104	0.3128	0.1950	0.0819	0.2691	0.2708	0.2427	0.2175	0.4087	0.2882	0.1791	0.1240	0.3246	0.4163	0.2591	0.3735	0.3672	0.2592	0.3319	0.2984	0.2469	0.1229
Cluster 5	0.11622	0.0962	0.1848	0.3065	0.4125	0.3689	0.2614	0.2142	0.1555	0.2269	0.2440	0.2382	0.2909	0.3919	0.2843	0.2141	0.1097	0.3080	0.2864	0.2660	0.1397	0.0837	0.2197	0.3388	0.3578	0.4261	0.3338	0.2401	0.7396	0.1426	0.1179	0.0347	0.0568	0.1662	0.7424
Cluster 6	0.10768	0.3617	0.2508	0.2206	0.1670	0.2656	0.2277	0.2475	0.2592	0.2003	0.3135	0.2710	0.2152	0.1912	0.2707	0.2944	0.2438	0.2835	0.2661	0.2425	0.2079	0.5343	0.2510	0.1352	0.0795	0.4886	0.3106	0.2008	0.5663	0.2933	0.1404	0.3724	0.2905	0.2385	0.0986
Cluster 7	0.08655	0.1198	0.1593	0.2529	0.4680	0.0656	0.1174	0.2119	0.6052	0.2430	0.2566	0.2512	0.2492	0.1106	0.1679	0.2492	0.4723	0.2968	0.2125	0.2252	0.2654	0.3051	0.2250	0.2195	0.2504	0.6994	0.1527	0.1479	0.8537	0.0934	0.0530	0.3014	0.2337	0.2342	0.2307
Cluster 8	0.08071	0.3815	0.2361	0.1933	0.1891	0.2396	0.2017	0.2339	0.3248	0.2486	0.3148	0.2515	0.1851	0.1891	0.2548	0.2937	0.2624	0.2382	0.2344	0.2421	0.2853	0.4210	0.2856	0.1804	0.1129	0.1348	0.2851	0.5801	0.1617	0.2792	0.5591	0.4006	0.2765	0.2125	0.1104
Cluster 9	0.04794	0.0474	0.1716	0.3655	0.4155	0.0271	0.0995	0.2425	0.6309	0.0047	0.0799	0.2704	0.6450	0.0203	0.0924	0.2456	0.6418	0.0900	0.3096	0.3648	0.2356	0.2506	0.3269	0.2296	0.1928	0.8240	0.1056	0.0704	0.3125	0.5193	0.1683	0.2800	0.3586	0.2680	0.0934
Cluster 10	0.03567	0.3306	0.1732	0.1741	0.3221	0.7510	0.1252	0.0753	0.0485	0.4824	0.1758	0.1227	0.2191	0.5980	0.1871	0.1279	0.0870	0.2005	0.1026	0.1641	0.5329	0.1666	0.1505	0.2131	0.4698	0.5104	0.1112	0.3784	0.6275	0.0472	0.3253	0.1827	0.1389	0.1840	0.4944

Parameter Estimates

		FSIZE Groups				SHOPPING_EX Groups				EX/INC Groups				Pct Bread Groups				Pct Meat Groups				Pct Fish Groups				Pct Milk Groups				Pct Oil Groups			
Cluster	Overall	Single	Couple	Family	Large Family	<=57.5K	57.5K-79K	79K-108K	>108K	<=27	27-38	38-51	>51	<=24	24-31	31-39	>39	<=8	8-12	12-17	>17	<=9.5	9.5-12.5	12.5-16.5	>16.5	<=3.5	3.5-5.5	5.5-8.5	>8.5	<=1	1-1.5	1.5-2	>2
Cluster 1	0.15037	0.0000	0.0011	0.9258	0.0731	0.1634	0.4337	0.3491	0.0537	0.2022	0.2873	0.3153	0.1953	0.0000	0.2494	0.6232	0.1274	0.0931	0.2987	0.3753	0.2329	0.2493	0.2721	0.2763	0.2022	0.1483	0.2832	0.3212	0.2473	0.1181	0.2899	0.2772	0.3148
Cluster 2	0.13179	0.0000	0.0052	0.6083	0.3865	0.0000	0.0512	0.2810	0.6678	0.1837	0.2924	0.3031	0.2207	0.3077	0.3938	0.2982	0.0002	0.0784	0.1630	0.3117	0.4469	0.2525	0.2372	0.2611	0.2492	0.1663	0.2265	0.2495	0.3577	0.3072	0.3503	0.1905	0.1521
Cluster 3	0.12320	0.0014	0.0774	0.8591	0.0622	0.0336	0.2325	0.3813	0.3526	0.2871	0.3120	0.2566	0.1443	0.5669	0.4330	0.0001	0.0000	0.0107	0.1014	0.3238	0.5641	0.1557	0.2589	0.3532	0.2322	0.1210	0.2634	0.3378	0.2778	0.0604	0.2296	0.2768	0.4331
Cluster 4	0.11988	0.0000	0.0013	0.4343	0.5644	0.0121	0.3097	0.4416	0.2366	0.0804	0.1752	0.2922	0.4523	0.0001	0.0003	0.1359	0.8637	0.4623	0.3004	0.1796	0.0577	0.3184	0.2457	0.2212	0.2147	0.3647	0.2704	0.2189	0.1461	0.3796	0.3537	0.1639	0.1028
Cluster 5	0.11622	0.0079	0.0901	0.7938	0.1082	0.0040	0.0894	0.2466	0.6599	0.6757	0.2345	0.0733	0.0166	0.6761	0.2577	0.0643	0.0019	0.0668	0.1666	0.2876	0.4790	0.3027	0.2393	0.2371	0.2209	0.0996	0.1965	0.2727	0.4311	0.1727	0.2748	0.2308	0.3217
Cluster 6	0.10768	0.0439	0.1063	0.7931	0.0567	0.6918	0.2989	0.0093	0.0000	0.0892	0.2193	0.3295	0.3620	0.0000	0.0000	0.0064	0.9935	0.5801	0.2790	0.1173	0.0236	0.3159	0.2550	0.2507	0.1784	0.4461	0.2558	0.1881	0.1100	0.2877	0.2987	0.2136	0.2000
Cluster 7	0.08655	0.2587	0.4630	0.2769	0.0013	0.8180	0.1552	0.0261	0.0007	0.3638	0.2729	0.2231	0.1402	0.2392	0.3683	0.3394	0.0531	0.2483	0.2624	0.2725	0.2168	0.1763	0.2035	0.2604	0.3597	0.2233	0.2308	0.2465	0.2993	0.1208	0.2152	0.2172	0.4468
Cluster 8	0.08071	0.0937	0.2167	0.6838	0.0058	0.4054	0.3703	0.2030	0.0213	0.0475	0.1736	0.3181	0.4608	0.1786	0.3627	0.4586	0.0001	0.2898	0.3051	0.2493	0.1558	0.1207	0.1561	0.2421	0.4812	0.4282	0.2340	0.1940	0.1438	0.3001	0.3286	0.2075	0.1638
Cluster 9	0.04794	0.0004	0.0264	0.7068	0.2663	0.2941	0.3962	0.2577	0.0520	0.0502	0.1273	0.2701	0.5524	0.0679	0.2299	0.3850	0.3172	0.9018	0.0833	0.0148	0.0001	0.1185	0.1910	0.2346	0.4559	0.5085	0.2501	0.1591	0.0824	0.0673	0.2259	0.2692	0.4376
Cluster 10	0.03567	0.4051	0.2993	0.2920	0.0035	0.7355	0.1919	0.0638	0.0088	0.6680	0.2051	0.0963	0.0306	0.3366	0.2150	0.2142	0.2342	0.4733	0.1972	0.1616	0.1679	0.6284	0.1454	0.0970	0.1292	0.2972	0.1736	0.1779	0.3513	0.4418	0.2212	0.1351	0.2018

Figure 25. Class membership probabilities and item response probabilities

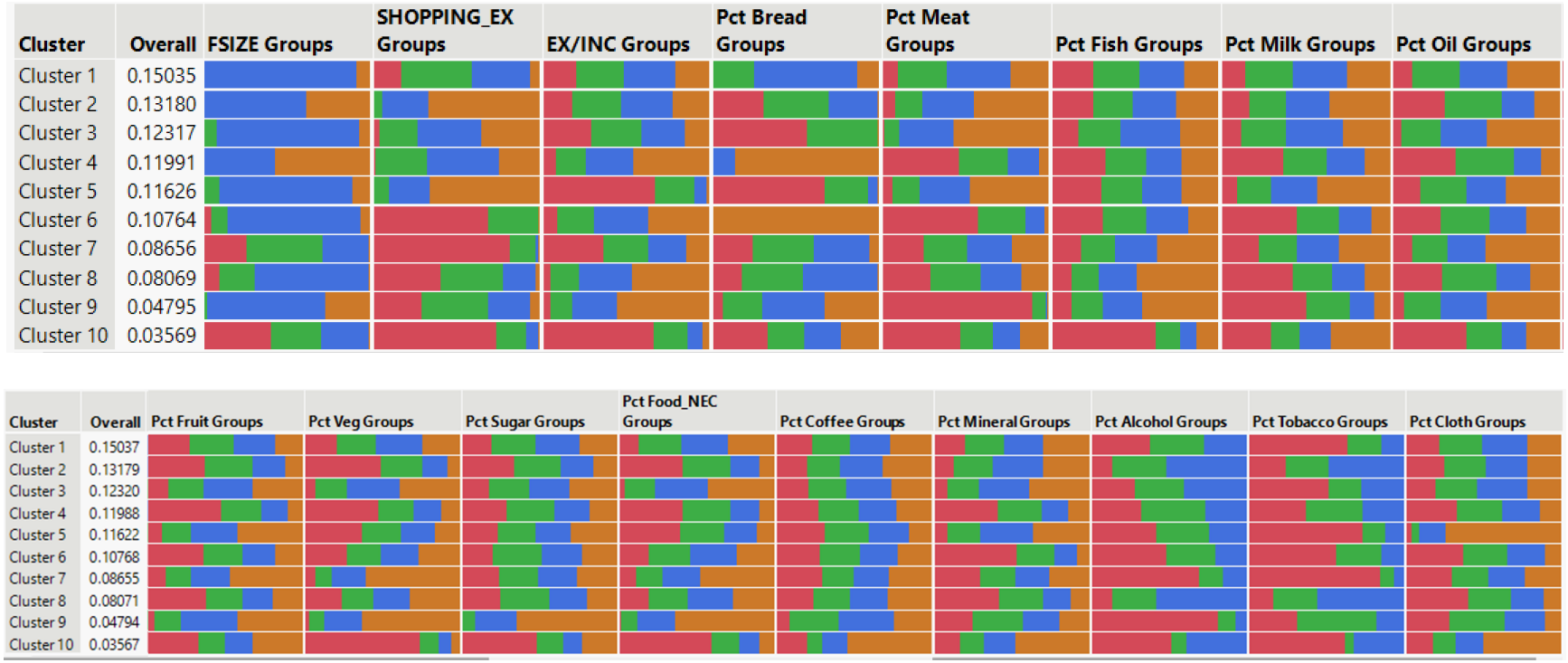


Figure 26. Class membership probabilities and item response probabilities plot

The characteristics of each clusters are described in Table 3.

No	Title	Description
1	Average family	Families that has up to 6 members. Spending pattern for all categories are mostly within the middle range (25^{th} to 75^{th} percentile), but mostly are non-smokers. Do not excessively allocate their expenses to buy a particular item.
2	High spending families	64% chance to be in the top 25% of total expenses spent for items that are sold in retail. Spend less on bread, oil, fruit, vegetable, sugar and jams, and sauces, but allocate more proportion of their expenses to buy meat, milk, alcohol, and tobacco.
3	Keto diet	Have moderate to high shopping expenses, but have 57% chance of being in the first quartile and 43% chance of being in the second quartile for percentage spent on bread. 88% are in the top half for percentage spent on meat. Also spending more on oil, fruit, vegetable, sauces and bottled drinks.
4	Large family	Mostly large families with middle to high shopping expenses, but high expense to income ratio. 86% of them have the top 25% highest proportion of expenses on bread compared to other customer segments, while spending a lot less for everything else.
5	High spending high income families	Even though most of them are in the top 25% for shopping expenses, their expenses to income ratio is in the lowest 25% which means their income is very high. Spend less on bread and higher for meat, milk, fruit. 74% are the top 25% on percentage spent on clothing.
6	Low expenditure low income	69% in the lowest quartile of shopping expenses and with medium to high expense and income ratio. 99% of them are in the top quartile for proportion of expenses spent on bread, while 58% in the lowest quartile for meat and 44% in the lowest quartile for milk.
7	Health-conscious singles and couples	Mostly are singles and couples with the 25% lowest shopping expenses. Buying more fish, oil, fruit, vegetable, sauce, and mostly do not smoke or drink alcohol.
8	Low expenditure alcoholic and smokers	Spending less proportion of expenses on staple food items, spending more proportion on alcohol and tobacco.

9	I don't like meat	90% in the lowest 25% for percentage of expenses spent on meat, 46% in the top 25% for fish. Spending more on oil, fruit, vegetable, sugar, sauces as well. In the top 25% for sugar and sauces.
10	Working singles and couples	Spend less on shopping, but with low expense to income ratio, indicating a higher income. Mostly in the lowest 25% for staples (meat, fish, oil, vegetable, sugar) but in the highest 25% for coffee and bottled drinks. Probably do not have time to cook at home so only shop for fruit and caffeine.

Table 3. LCA Clusters characteristics

4. Discussion and Interpretation

To examine the clustering results of K-means and LCA, descriptive analysis is performed. A new variable Max Category is added to see which expense category has the highest value.

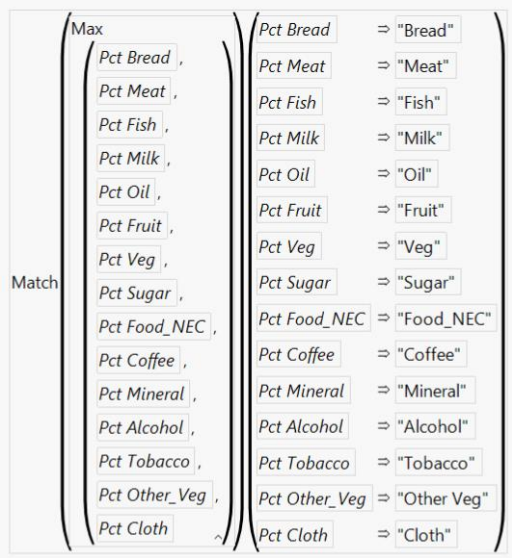


Figure 27. Formula for Max Category

By comparing the distribution of clusters within each maximum category, it seems that the k-means (Figure 28) is better at explaining the spending pattern of households in the Philippines than LCA (Figure 29). While the distribution of LCA clusters does not seem to show a discernible pattern, the distribution of k-means clusters shows a pattern that confirms the analysis of each cluster's characteristics.

Alcohol and Tobacco cluster segment accounts for 98% and 100% of records that has the highest expenditure in Alcohol and Tobacco category respectively. Masterchef Family accounts for 100%

of records with the highest expenditure in Food_NEC (sauces) and oil, as well as 68% of fruit and 91% of vegetables. Fashionista also has the highest proportion that makes up Cloth Max Category (95%), Busy Workers for Coffee (99%), Seafood Lovers for Fish (61%), Traditional Household for Other Vegetables (100%), and Sweet Tooth for Sugar (100%). Large Family with Baby makes up 69%, 76%, and 46% of Meat, Milk, and Mineral (bottled drinks).

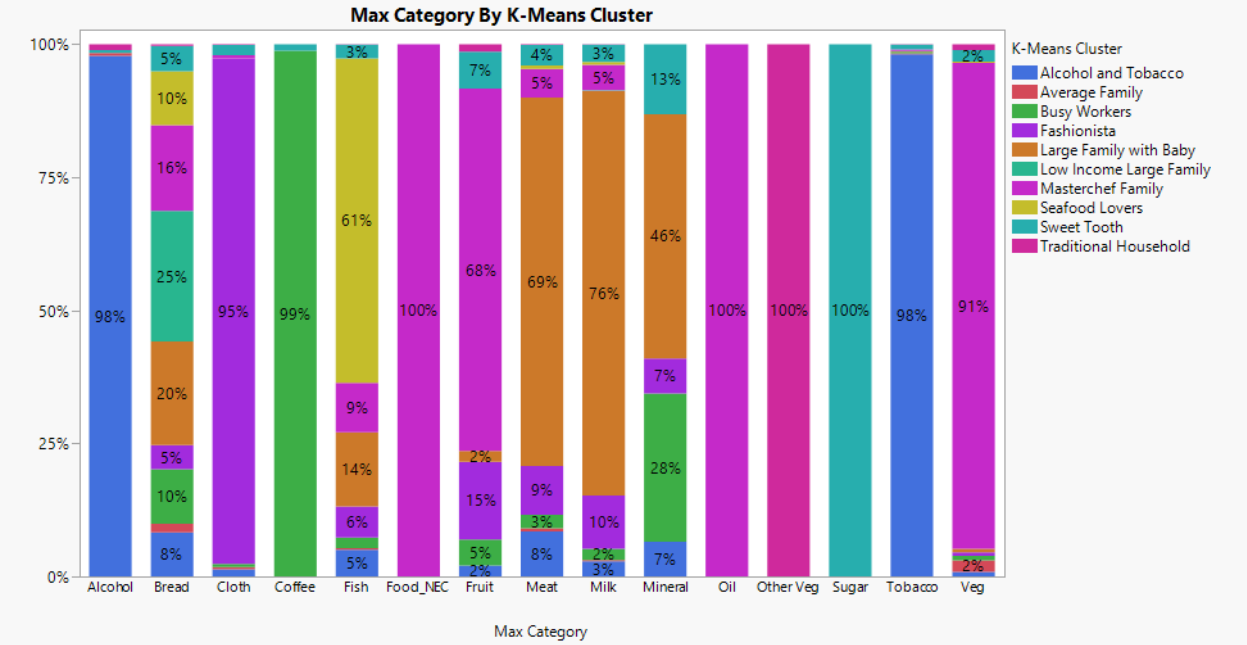


Figure 28. Distribution of K-means cluster in each Max Category

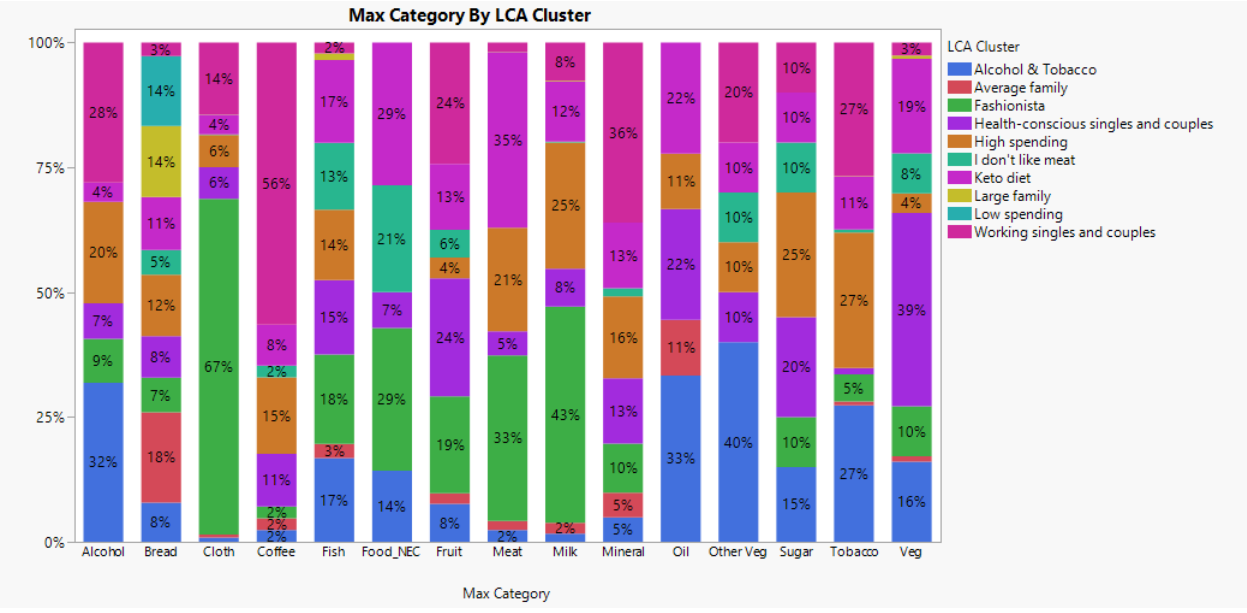


Figure 29. Distribution of LCA Cluster in each Max Category

To determine the most profitable customer segment, the total shopping expenses contributed by each cluster were examined (Figure 30). Large Family with Baby has the largest contribution to the total shopping expenses (31.95%), so they are likely to be the most profitable for the retail. While Fashionista also have a lot of expenses, it largely goes into clothing items, which they are likely to buy in a more expensive brand rather than hypermarket. Busy Workers group is more likely to purchase their items from a minimarket, while Traditional Household is more likely to shop in traditional market. Average Family is also discarded because it only contributes less than 2% of the total shopping expenses. Key customer segments that are ranked by median shopping expenses are identified in Table 4.

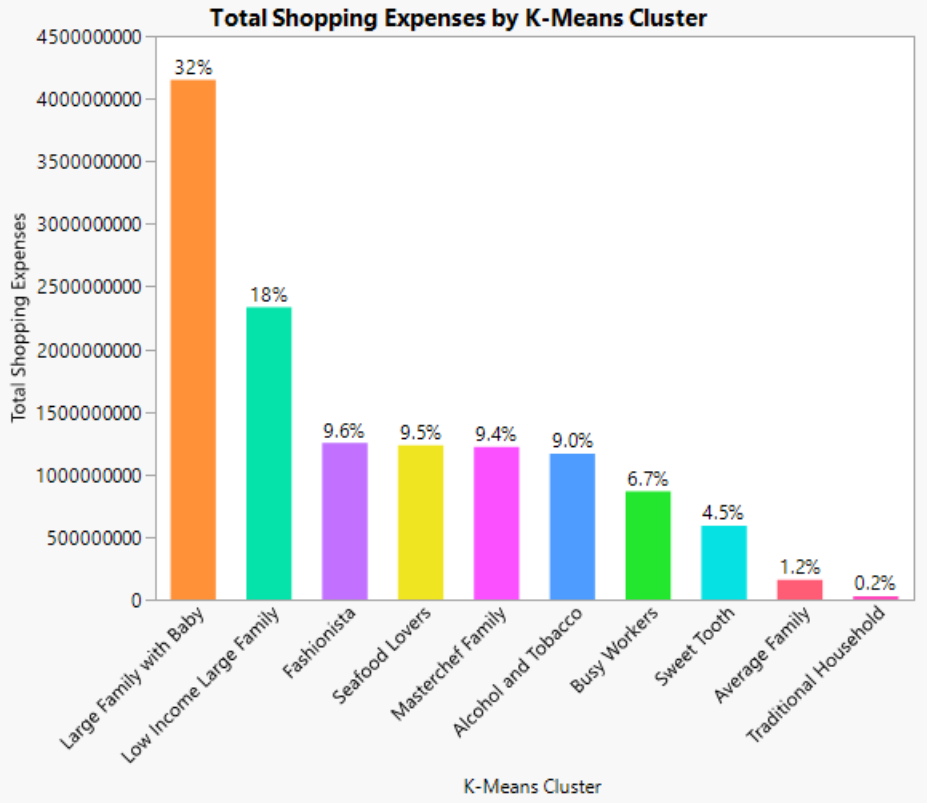


Figure 30. Total shopping expenses by k-means cluster treemap plot

Rank	Customer Segment	Number of Members	Preferred Expense Category	Median Family Size	Median Total Income (₱)	Median Shopping Expenses (₱)
1	Large Family with Baby	34,905	Meat, Milk, Mineral	5	350,440	109,028
2	Low Income Large Family	29,644	Bread	6	153,319	74,061
3	Seafood Lovers	16,310	Fish	4	143,852	71,262.5
4	Masterchef Family	21,439	Oil, Fruit, Vegetable, Food_NEC	3	141,364	53,487
5	Alcohol and Tobacco	12,833	Alcohol, Tobacco	4	210,435	84,388
6	Sweet Tooth	6,603	Sugar	4	220,009	81,669

Table 4. Key customer segments

The distribution of clusters in each region in the Philippines is shown in Figure 33. ARMM has the highest total percentage of the key customer segments, followed by Zamboanga Peninsula and SOCCSKSSARGEN.

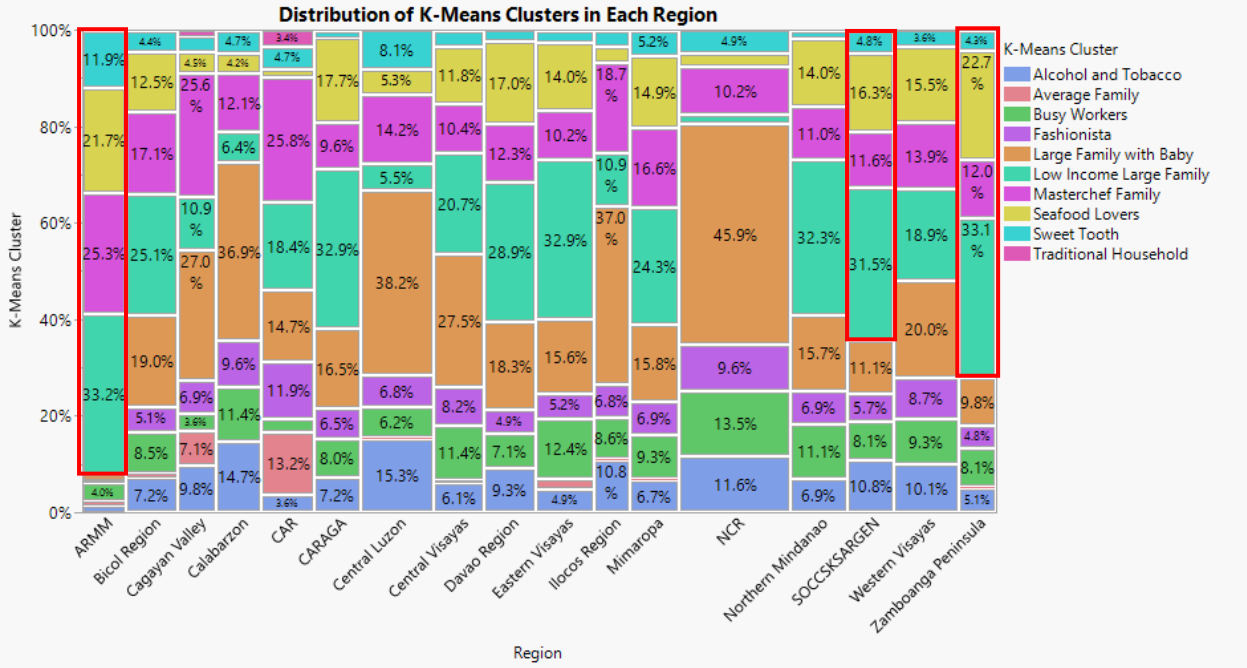


Figure 31. Distribution of K-means clusters in each region

Appendix

Appendix A. Unused Analysis Results

Iterative Clustering

Cluster Comparison

Method	NCluster	CCC	Best
K Means Cluster	3	-120.56	
K Means Cluster	4	-95.195	
K Means Cluster	5	-114.72	
K Means Cluster	6	-110.7	
K Means Cluster	7	-108.88	
K Means Cluster	8	-112.1	
K Means Cluster	9	-100.53	
K Means Cluster	10	-100.58	
K Means Cluster	11	-90.164	
K Means Cluster	12	-81.288	Optimal CCC

Figure 32. K-means clustering without scaling columns individually

Appendix B. Data Change Log

Item	Field Name	Issue	Comments
1	W_REGN	Values are numerically encoded.	Recoded into a new column "Region" with the actual region name to make it easier to understand.
2		Wrong modelling type.	Converted into nominal modelling type.
3	Non-missing Value	New formula column.	Count the number of expense categories which values are not zero.
4		Some records have only 3 non-zero values or less.	Hidden and excluded from analysis
5	BREAD, MEAT, FISH	Some records have the value 0 for all three variables.	Hidden and excluded from analysis

6	PCINC_mine	New formula column to calculate the income for every person in the household.	TOINC divided by FSIZE.
7	SHOPPING_EX	New formula column to calculate the total expenses for items that might be sold in retail.	Sum of all expense categories.
8	EX/INC	New formula column to calculate the ratio of shopping expenses to total income.	SHOPPING_EX divided by TOINC.
9	EX/INC	Some expenses to income ratio is greater than 100%.	No action taken. Probably due to bad financial management so they overspend.
10	Log[TOINC]	New formula column to standardize TOINC.	Log transformed TOINC using natural base.
11	PCINC_mine	New formula column to standardize PCINC_mine.	Log transformed PCINC_mine using natural base.
12	SHOPPING_EX	New formula column to standardize SHOPPING_EX.	Log transformed SHOPPING_EX using natural base.
13	Pct Bread	New formula column to calculate the percentage allocated for rice, bread, and cereals out of the total shopping expenses.	BREAD divided by SHOPPING_EX.
14	Pct Meat	New formula column to calculate the percentage allocated for meat out of the total shopping expenses.	MEAT divided by SHOPPING_EX.
15	Pct Fish	New formula column to calculate the percentage	FISH divided by SHOPPING_EX.

		allocated for fish and seafood out of the total shopping expenses.	
16	Pct Milk	New formula column to calculate the percentage allocated for milk out of the total shopping expenses.	MILK divided by SHOPPING_EX.
17	Pct Oil	New formula column to calculate the percentage allocated for butter and oil out of the total shopping expenses.	OIL divided by SHOPPING_EX.
18	Pct Fruit	New formula column to calculate the percentage allocated for fruits out of the total shopping expenses.	FRUIT divided by SHOPPING_EX.
19	Pct Veg	New formula column to calculate the percentage allocated for vegetables out of the total shopping expenses.	VEG divided by SHOPPING_EX.
20	Pct Sugar	New formula column to calculate the percentage allocated for sugar, sweets, jams, and chocolates out of the total shopping expenses.	SUGAR divided by SHOPPING_EX.
21	Pct Food_NEC	New formula column to calculate the percentage allocated for sauce and condiments out of the total shopping expenses.	FOOD_NEC divided by SHOPPING_EX.
22	Pct Coffee	New formula column to calculate the percentage	COFFEE divided by SHOPPING_EX.

		allocated for coffee, tea, and cocoa out of the total shopping expenses.	
23	Pct Mineral	New formula column to calculate the percentage allocated for mineral water, juice, and other bottled drinks out of the total shopping expenses.	MINERAL divided by SHOPPING_EX.
24	Pct Alcohol	New formula column to calculate the percentage allocated for alcohol out of the total shopping expenses.	ALCOHOL divided by SHOPPING_EX.
25	Pct Tobacco	New formula column to calculate the percentage allocated for tobacco out of the total shopping expenses.	TOBACCO divided by SHOPPING_EX.
26	Pct Other_Veg	New formula column to calculate the percentage allocated for betel nut, betel leaves, mint leaf and lime out of the total shopping expenses.	OTHER_VEG divided by SHOPPING_EX.
27	Pct Cloth	New formula column to calculate the percentage allocated for clothing and footwear out of the total shopping expenses.	CLOTH divided by SHOPPING_EX.
32	FSIZE Groups	Binned variable.	FSIZE binned into equal quartiles.
33	SHOPPING_EX Groups	Binned variable.	SHOPPING_EX binned into equal quartiles.

34	EX/INC Groups	Binned variable.	EX/INC binned into equal quartiles.
35	Pct Bread Groups	Binned variable.	Pct Bread binned into equal quartiles.
36	Pct Meat Groups	Binned variable.	Pct Meat binned into equal quartiles.
37	Pct Fish Groups	Binned variable.	Pct Fish binned into equal quartiles.
38	Pct Milk Groups	Binned variable.	Pct Milk binned into equal quartiles.
39	Pct Oil Groups	Binned variable.	Pct Oil binned into equal quartiles.
40	Pct Fruit Groups	Binned variable.	Pct Fruit binned into equal quartiles.
41	Pct Veg Groups	Binned variable.	Pct Veg binned into equal quartiles.
42	Pct Sugar Groups	Binned variable.	Pct Sugar binned into equal quartiles.
43	Pct Food_NEC Groups	Binned variable.	Pct Food_NEC binned into equal quartiles.
44	Pct Coffee Groups	Binned variable.	Pct Coffee binned into equal quartiles.
45	Pct Mineral Groups	Binned variable.	Pct Mineral binned into equal quartiles.
46	Pct Alcohol Groups	Binned variable.	Pct Alcohol binned into equal quartiles.
47	Pct Tobacco Groups	Binned variable.	Pct Tobacco binned into equal quartiles.
48	Pct Other_Veg Groups	Binned variable.	Pct Other_Veg binned into equal quartiles.

49	Pct Cloth Groups	Binned variable.	Pct Cloth binned into equal quartiles.
50	Cluster	New column created by K-means clustering. Number of the cluster.	-
51		-	Recoded to the customer segment title instead of number (automatically changed the data type into character).
52		Similar name with LCA clustering results.	Renamed into “K-means Cluster”.
53	Distance	New column created by K-means clustering. Distance from the centroid.	-
54	Most Likely Cluster	New column created by LCA clustering.	-
55		-	Recoded to the customer segment title instead of number (automatically changed the data type into character).
56		-	Renamed into “LCA Cluster”.
57	Max Category	New formula column.	Return the expense category with the highest value.