BUAN 6337: Predictive Analytics with SAS

Group Project (Peanut Butter)

Group 1

Members:

Ahmet Berk Gungor

Hsin Yen Lee

Shubham Bhan

Suraj Malpani

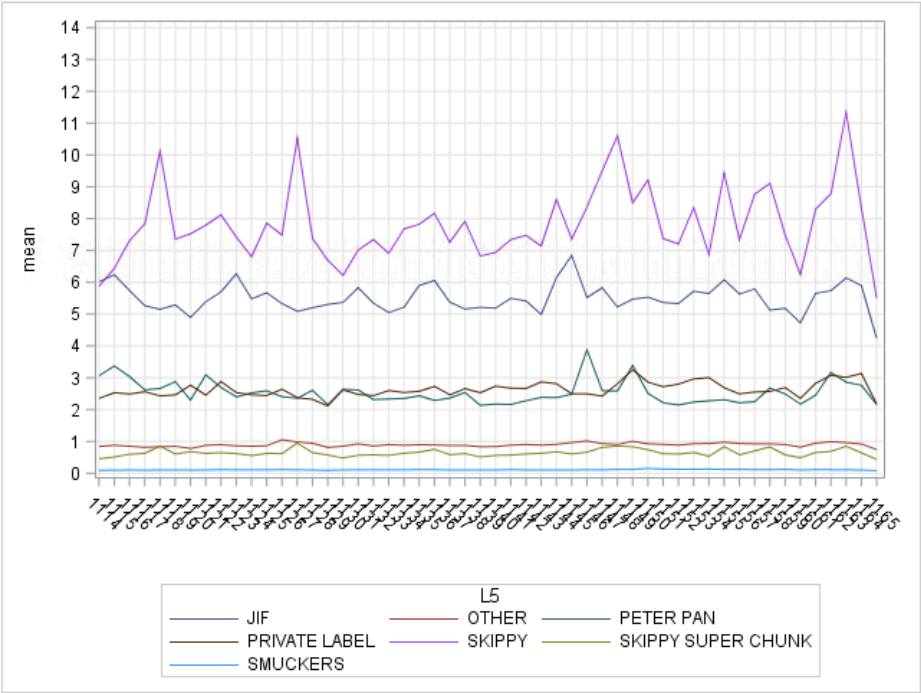
Descriptive Analysis

In the descriptive analysis section of our project, the questions we wanted to answer was how do our variables behave, what is their mean and variance, and is there any kind of patterns or insights that we can gain just by looking at the distribution of and correlation of these variables.

After initially exploring the data, it is seen that to compute the price for a brand and acquire correct results in the end, first we need to calculate weights throughout the brands. In our analysis, we chose 6 brands with the highest dollar sales and gathered all the rest of the brands under the brand name "other". To calculate price, feature and display columns correctly, we took price per ounce and multiplied it with the sales per week. After calculating this for all the observations, we calculated brand specific sums and the total market sum. After dividing brand specific sums over the total market sum, we obtained the annual market share of each of the brands we are going to analyze. After finding brand market share, then we multiplied these numbers with individual price per ounce, display and feature fields to get the weighted calculations. Of course, we first converted the character feature field into numerical. In the below proc tabulate table, you can see brand specific sums which we used to calculate the brand market share.

Obs	L5	_TYPE_	_PAGE_	_TABLE_	weights_N	weights_Sum	brandmarketshare
1	JIF	1	1	1	504338	585971490.98	0.26216
2	OTHER	1	1	1	522638	175859105.28	0.07868
3	PETER PAN	1	1	1	235555	325764522.01	0.14574
4	PRIVATE LABEL	1	1	1	496935	389779707.96	0.17438
5	SKIPPY	1	1	1	265240	652339471.58	0.29185
6	SKIPPY SUPER CHUNK	1	1	1	175863	90701088.28	0.04058
7	SMUCKERS	1	1	1	77343	14760139.57	0.00660

After generating all the necessary variables for our analysis, we built the below graph to quickly see how do brands' price change over the course of the weeks. In the graph, we can see that the brands Skippy and JIF have the highest average weighted prices overall. These two brands also have the highest market share. Having average prices higher than the competition while having the highest market shares might mean that these two brands are known to customers and are much preferable over the other brands even though their prices are a bit higher.



Continuing to our analysis, we checked the means of average price and weighted display and feature among brands. Although Skippy brand has the highest average price, its standard deviation was also the highest, meaning the average price of the brand fluctuates a lot. This problem is also seen with JIF but not so much with the remaining brands. When it came to weighted display and feature, Skippy also had the highest mean display and feature, followed by JIF. This correlation might be a causation, and it is worth looking into. From these charts, we can also see that total observations for Skippy is a lot less than that of JIFs even after accounting for Skippy and Skippy Super Chunk are the same brands. This might have led to favorable situation for Skippy in the analysis, since the means shown here for Skippy's best competitor, JIF, are closer to its true mean since JIF has more related observations to its brand. But it should not be a problem in our case.

Analysis Variable : averageweightedprice						
L5	N Obs	N	Mean	Std Dev	Minimum	Maximum
JIF	504373	504338	5.5115397	9.7712819	0.0675790	652.2540800
OTHER	522796	522638	0.8970186	2.3432307	0.0224800	1160.35
PETER PAN	235566	235555	2.5406755	5.6409952	0.0971600	516.7849717
PRIVATE LABEL	496954	496935	2.6300391	4.9162715	0.0613818	572.8080741
SKIPPY	265255	265240	7.8356842	14.8830673	0.1444658	1005.03
SKIPPY SUPER CHUNK	175865	175863	0.6434166	1.0763703	0.0162320	49.0660896
SMUCKERS	77352	77343	0.1111249	0.0990472	0.0032340	1.2654180

L5	N Obs	Variable	N	Mean	Std Dev	Minimum	Maximum
JIF	504373	weightedD	504338	0.0065117	0.0536147	0	0.5243200
		weightedF	504373	0.0284233	0.1337841	0	1.0486400
OTHER	522796	weightedD	522638	0.000594197	0.0088358	0	0.1573600
		weightedF	522796	0.0061446	0.0354918	0	0.3147200
PETER PAN	235566	weightedD	235555	0.0044535	0.0329467	0	0.2914800
		weightedF	235566	0.0175990	0.0798474	0	0.5829600
PRIVATE LABEL	496954	weightedD	496935	0.0092619	0.0520796	0	0.3487600
		weightedF	496954	0.0196369	0.0918942	0	0.6975200
SKIPPY	265255	weightedD	265240	0.0083317	0.0649437	0	0.5837000
		weightedF	265255	0.0642465	0.2162103	0	1.1674000
SKIPPY SUPER CHUNK	175865	weightedD	175863	0.0013035	0.0096133	0	0.0811600
		weightedF	175865	0.0046853	0.0229894	0	0.1623200
SMUCKERS	77352	weightedD	77343	0.000015787	0.000401737	0	0.0132000
		weightedF	77352	0.000082253	0.0011562	0	0.0264000

Below is the frequency table for the features. From the table, we learn that all brands mostly focused for featuring through medium sized ads. After these came the large size ads and then retailer coupons. Skippy gave around the same amount of medium and large ads, but far more coupons to its customers, which is a valuable insight from this table.

Frequency
Percent
Row Pct
Col Pct

Table of F by L5

F	L5(L5) JIF	OTHER	PETER PAN	PRIVATE LABEL	SKIPPY	SKIPPY SUPER CHUNK	SMUCKERS	Total
A	7815	6143	5044	9177	7222	2631	76	38108
	0.34	0.27	0.22	0.40	0.32	0.12	0.00	1.67
	20.51	16.12	13.24	24.08	18.95	6.90	0.20	
	1.55	1.18	2.14	1.85	2.72	1.50	0.10	
A+	774	1238	393	845	2612	1317	32	7211
	0.03	0.05	0.02	0.04	0.11	0.06	0.00	0.32
	10.73	17.17	5.45	11.72	36.22	18.26	0.44	
	0.15	0.24	0.17	0.17	0.98	0.75	0.04	
B	12998	7565	5293	11458	12592	3434	237	53577
	0.57	0.33	0.23	0.50	0.55	0.15	0.01	2.35
	24.26	14.12	9.88	21.39	23.50	6.41	0.44	
	2.58	1.45	2.25	2.31	4.75	1.95	0.31	
C	2007	1685	1112	2059	1034	268	98	8263
	0.09	0.07	0.05	0.09	0.05	0.01	0.00	0.36
	24.29	20.39	13.46	24.92	12.51	3.24	1.19	
	0.40	0.32	0.47	0.41	0.39	0.15	0.13	
NONE	480744	506007	223713	473396	241780	168213	76900	2170753
	21.10	22.21	9.82	20.78	10.61	7.38	3.38	95.30
	22.15	23.31	10.31	21.81	11.14	7.75	3.54	
	95.32	96.82	94.97	95.26	91.16	95.65	99.43	
Total	504338	522638	235555	496935	265240	175863	77343	2277912
	22.14	22.94	10.34	21.82	11.64	7.72	3.40	100.00

Frequency Missing = 249

For a better look into the brands and price, we used proc univariate. We see that the mean weighted price among all brands was $3.23, while the variance of the price was 62.7, meaning that we have huge variation in the dataset. Then we looked at the correlation between brand average price, weighted Display and weighted features. Price is 27% correlated with display and 28% with feature while the correlation between display and feature is around 18%. These levels are all acceptable for our analysis.

The UNIVARIATE Procedure
Variable: averageweightedprice

Moments			
N	2277912	Sum Weights	2277912
Mean	3.22839817	Sum Observations	7354006.93
Std Deviation	7.91768008	Variance	62.6896578
Skewness	21.2929959	Kurtosis	1099.35594
Uncorrected SS	166543124	Corrected SS	142801461
Coeff Variation	245.25104	Std Error Mean	0.00524601

Basic Statistical Measures			
Location		Variability	
Mean	3.228398	Std Deviation	7.91768
Median	1.274948	Variance	62.68966
Mode	0.235253	Range	1160
		Interquartile Range	2.79031

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	615.4002	Pr > |t|	<.0001
Sign	M	1138956	Pr >= |M|	<.0001
Signed Rank	S	1.297E12	Pr >= |S|	<.0001

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
averageweightedprice	2277912	3.22840	7.91768	7354007	0.00323	1160
weightedD	2277912	0.00513	0.04323	11687	0	0.58370
weightedF	2278161	0.02165	0.11204	49325	0	1.16740

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	averageweightedprice	weightedD	weightedF
averageweightedprice	1.00000 2277912	0.27235 <.0001 2277912	0.28966 <.0001 2277912
weightedD	0.27235 <.0001 2277912	1.00000 2277912	0.18922 <.0001 2277912
weightedF	0.28966 <.0001 2277912	0.18922 <.0001 2277912	1.00000 2278161

Although not being considered a part of a descriptive analysis naturally, we also wanted to include a metric conjoint analysis using proc transreg in this part of our project to get further insights. When we look at the results, we see that Skippy is the most preferred brand with a utility of 478.2, followed by JIF with a utility of 158.2. Our reference group here is Smuckers, which is the second least preferred brand followed by Skippy Super Chunk, but that brand is considered in Skippy so it really is the Smuckers that is in trouble here brand sales wise. When it comes to display, major display provides the highest utility to the brand sales, which is expected. It is worth mentioning that no display provides far less of a utility than minor display when it comes to brand sales. For features, our reference group is coupons. It seems that large size ads provide the most utility, followed by medium-sized ads and then coupons. The difference between small ads and no featuring at all seems very small.

Utilities Table Based on the Usual Degrees of Freedom				
Label	Utility	Standard Error	Importance (% Utility Range)	Variable
Intercept	5596.8	91.864		Intercept
L5 JIF	158.2	25.461	6.706	Class.L5JIF
L5 OTHER	-45.9	25.391		Class.L5OTHER
L5 PETER PAN	26.4	27.326		Class.L5PETER_PAN
L5 PRIVATE LABEL	-84.1	25.501		Class.L5PRIVATE_LABEL
L5 SKIPPY	478.2	26.985		Class.L5SKIPPY
L5 SKIPPY SUPER CHUNK	-134.8	28.446		Class.L5SKIPPY_SUPER_CHUNK
D 0	-4977.9	44.153	54.462	Class.D0
D 1	-2719.1	66.341		Class.D1
Fnew 0	-632.2	77.845	38.832	Class.Fnew0
Fnew 1	-688.0	106.281		Class.Fnew1
Fnew 2	1103.3	82.715		Class.Fnew2
Fnew 3	2861.3	84.731		Class.Fnew3

Recommendations: From what we have seen so far, the higher the brand average annual price, the higher its total sales are. However, this correlation is there for the top and known brands like JIF and Skippy, which means that people are preferring these brands even when their prices are higher than competitors'. We can say that people are used to the taste of these brands. The higher the display option, the higher brand total sales. On the other hand, the best performing feature option is not the best option that gives customers the most utility, which are coupons. Rather it is the large and medium-sized ads that increase the brand total sales the most. Since we configured our mapping as coupons having the highest-ranking value, this may result in negative coefficients for feature on the potential regressions.

Brand Choice Analysis

In this section of our project, we wanted to conduct a brand choice analysis since each brand in our case has different prices and promotions and utility varies only by the other brand's characteristics of price, feature, and display.

We decided to use proc mdc with conditional logit to conduct our brand choice analysis since the dependent variable "decision" will be a multinomial discrete variable. However, to use this type of model, we first needed to decide on which variables to use in the model and reshape the model accordingly. In this model, we chose to use price, display and feature variables to see how these 3 major variables affect brand choice. These three variables are keys to understand most of the variation in decision making.

We proceeded to assign a brand number to each of the 6 brands we have (this time we considered Skippy Chunky in the brand Skippy). Then, we reshaped the price, display and feature variables from long format to wide format, which is the format that choice models need to work. We then ran a do loop to construct a multinomial decision variable so that one individual "id" would have 6 different brand choice records but only the chosen brand has the decision variable of 1 with the rest being decision of 0. On this step, we had a problem dealing with missing data, which was

making it so that for some ids in the model no choice was given. So, to exclude these subjects from the model, we included the ids where price, display and feature columns were all non-missing. Then, since the sum of the decision variable should be non-zero across all of the ids in the analysis, we ran the proc mdc model with only those observations that have a non-zero sum for the decision column. With that, our model was ready to run.

Discrete Response Profile

Index	CHOICE	Frequency	Percent
0	1	504338	22.14
1	2	522638	22.94
2	3	235555	10.34
3	4	496935	21.82
4	5	441103	19.36
5	6	77343	3.40

Goodness-of-Fit Measures

Measure	Value	Formula
Likelihood Ratio (R)	8.16E6	2 * (LogL - LogL0)
Upper Bound of R (U)	8.16E6	- 2 * LogL0
Aldrich-Nelson	0.7818	R / (R+N)
Cragg-Uhler 1	0.9722	1 - exp(-R/N)
Cragg-Uhler 2	1	(1-exp(-R/N)) / (1-exp(-U/N))
Estrella	1	1 - (1-R/U)^(U/N)
Adjusted Estrella	1	1 - ((LogL-K)/LogL0)^(-2/N*LogL0)
McFadden's LRI	1	R / U
Veall-Zimmermann	1	(R * (U+N)) / (U * (R+N))

N = # of observations, K = # of regressors

The SAS System

The MDC Procedure

Conditional Logit Estimates

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
price	1	2878	5997	0.48	0.6313
feature	0	-285.3647	.	.	.
display	0	-312.2025	.	.	.

In the above discrete response profile, we see that brand 2 was the most frequently chosen, which corresponds to the brand Other, followed by JIF, Private Label and then Skippy. Then, when we look at the actual Conditional Logit Estimates, we see that the higher the price of a brand, the higher the likelihood of the brand being selected, although the price is not significant at a 5% significance level in this model. A higher feature meaning mostly coupons, in this case, leads to a lower likelihood of the brand being selected. Similarly, the higher display leads to a lower likelihood of the brand being selected. As you can see, there are some issues with these results. Previously, we have seen that leading brands have higher prices and major displays, but they got the most utility out of large-sized ads opposed to coupons. These points might explain some

portion of the coefficients of price and feature, but then the coefficient on display does not make sense.

These coefficients tell us that the price of the brand does not matter on the choice of the brand and the lower the display and feature, the higher the choice of that brand. Therefore, we made a mistake in our analysis.

This mistake could be that our variables are not enough to explain most of the variance of the decision or that when we are setting up the model, we miscalculated a step although we rechecked all of the calculations. The code is attached at the end of the project. It could also be that we used a proc mdc model rather than a bchoice model. However, we could not get a bchoice model to work after 6 hours of trying.

However, going by our estimates, this model taught us that price point is not a good determinant of the peanut butter brand choice. As we have seen from our descriptive analysis and proc transreg procedure before, the top brands are chosen despite of their higher prices. The insignificant price coefficient is associated with this fact. When it comes to the negative coefficient on feature, on our analysis we gave A+ feature, which are the coupons a 4 rating, large size ads a 3 and so on. Since large size ads had the most utility for brand sales, and coupons were the third most utility, this negative coefficient can also be explained partially.

Time Series Analysis

For time series analysis, our goal was to forecast sales for the brand. For this reason, we created separate datasets for each brand's weekly sales. After separating them we rand PROC ARIMA to check ACF, PACF, IACF plots to check for any trend or autocorrelations. We could also get an initial idea about which model we are going to fit using these plots based on when ACF, PACF plots trail to 0. Later, we checked each data for stationarity using ADF, PP tests in PROC ARIMA. If we could reject the null hypothesis at 5% significance level, which indicated the series are stationary and no further treatment is needed. These tests also confirmed that the series is not a 'Random walk'.

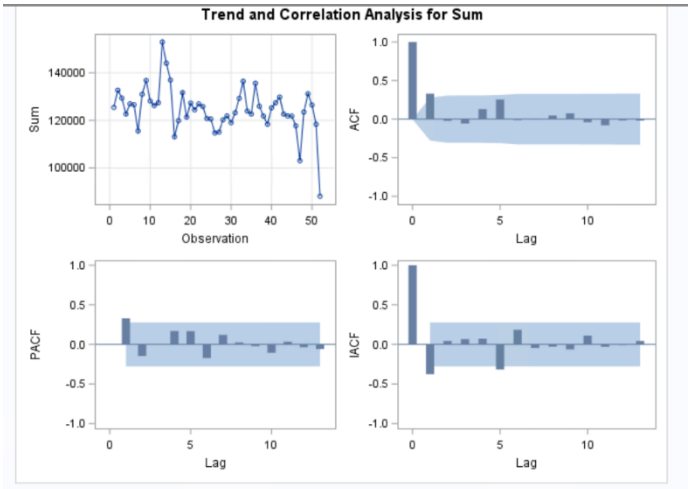
Afterward, we used 'MINIC' in conjunction with PROC ARIMA to figure out p and q values of the ARMA models. The lowest Minimum Information Criterion value gave us the most significant model for that Series, and we used the same model to Estimate and Forecast the future weekly sales for all the Data series/Brands. We forecasted the next 8 weeks of Dollar sales for all the brands. One sample of tests, ACF, PACF plots and a forecast with the confidence interval looks like below:

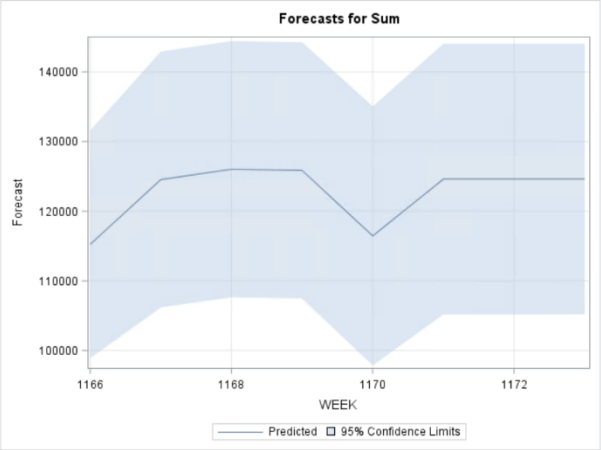
Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	11.07	6	0.0861	0.330	-0.022	-0.056	0.129	0.254	-0.011
12	12.13	12	0.4350	-0.006	0.046	0.075	-0.040	-0.080	-0.016

Minimum Information Criterion						
Lags	MA 0	MA 1	MA 2	MA 3	MA 4	MA 5
AR 0	18.23881	18.05986	18.13211	18.17608	18.24765	17.94662
AR 1	18.09156	18.13391	18.2078	18.25187	18.32297	18.02245
AR 2	18.11909	18.19229	18.24092	18.31219	18.37822	18.09377
AR 3	18.19285	18.26298	18.3071	18.3757	18.44367	18.16904
AR 4	18.19952	18.26304	18.33386	18.40506	18.36759	18.22222
AR 5	18.00382	18.0459	18.10582	18.18051	18.22411	18.29816

Error series model: AR(8)

Minimum Table Value: BIC(0,5) = 17.94662



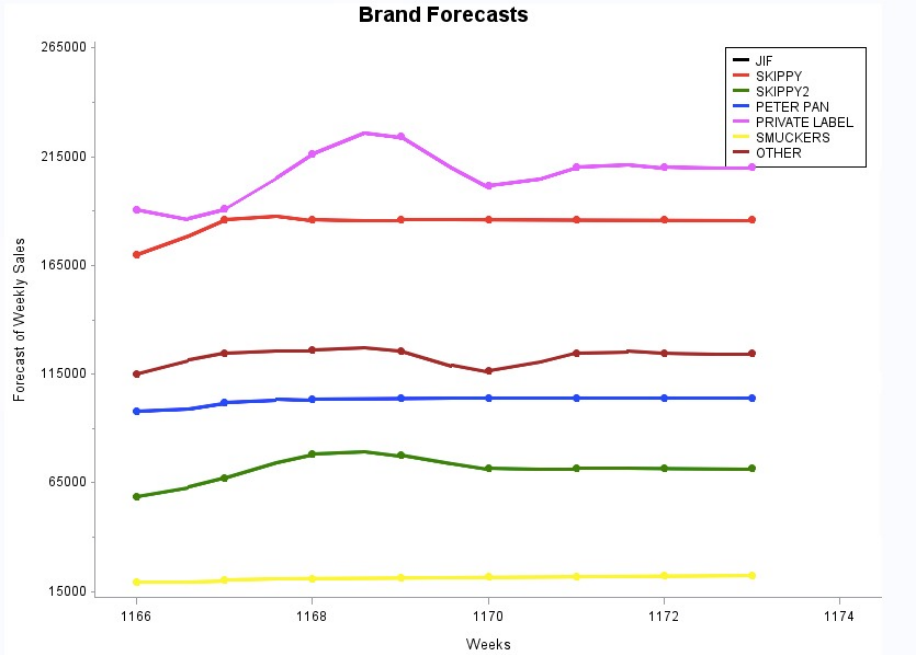


After forecasting using the decided models for each brand separately, we get the following results for the next 8 weeks. We forecasted only 8 weeks since ARIMA models are optimum for short term forecasting and the forecasts become steady after a point because of moving averages.

As we can see in the following graph, Private Label's total Dollar sales see a peak in the next 3-4 weeks and then become steady. Private Label was already dominating the market in terms of the dollar sales and continues to do so. For Skippy, Sales seem to grow in the next 2 weeks and then become steady. All the rest of the brands follow a similar pattern, as their sales become steady after 3-4 weeks.

For Skippy, they have an opportunity to take over Private label's sales in the next a couple of weeks when their sales come very close. Also, they need to improve their marketing strategies as

their sales don't seem to grow much after the next 3-4 weeks.



Conclusion

In this project, we provide recommendations to the peanut butter brand, Skippy. From the analysis above, we can find that Skippy and its top competitor JIF, counts for more than 50% of the all, are leading brands in the peanut butter market. Although JIF and Skippy have the highest weighted average price among the all, price is not the main determinant when customers are choosing which brand to buy. One possible reason is that peanut butter is not a luxury product. Compared to luxury goods, peanut butter is not so expensive and the risk of buying wrong peanut butter is acceptable for most customers, so customers do not conduct complex and hard buying decision processes. We can say that when choosing peanut butter, customers tend to do habitual buying or simply routine problem-solving. From the above analysis, we found that coupons, display, and feature will influence customers' decision making. Hence, if Skippy wants to boost its sales in the next 3 to 4 weeks, the company can take the following suggestions based on our analysis. First, show or increase the large size ads at the point-of-sales, since the large size ad presents the highest utility from the brand choice analysis. The large size ad could be a reminder or a trigger when customers are standing in front of the shelves, so customers would be attracted by the ads and are likely to give Skippy a try. It is comparatively inexpensive and effective than other forms of promotion. Second, checking and ensuring the product display is the best level in the store, since the higher the display score can also lead to higher sales. Last, re-evaluate the performance or conversion rates of sending coupons. From the data, we can know that Skippy tends to give coupons to customers as part of its promotional strategies. However, after our analysis, we found that coupons do not give the customers most utilities. Therefore, Skippy should consider that coupons are not the most effective promotion to boost their peanut butter sales.