Group2

Advait Vaidya
Aniket Sanghvi
Foram Gohil
Manoj Thadani
Suraj Gupta
Subham Dwivedi

PROJECT REPORT – MARKETING PREDICTIVE ANALYTICS- LAUNDRY DETERGENT

Contents

Market Overview

Laundry Detergents are commonly available as powders or concentrated solutions and are commonly used for cleaning clothes and other household purposes. They come in different forms, sizes, fragrances, concentration levels, and package types.

We have data about the Panelist (Demographics and sample of their transactions at drug and grocery stores in a particular area), we have scanner data for drugs, grocery and merchandise stores and Product data (descriptive).

There 72 Companies and 137 brands. The total sales amount for the sample drug stores was $5.5M and total sales amount for the sample grocery store was $42M for the year 2001.

Data Preparation

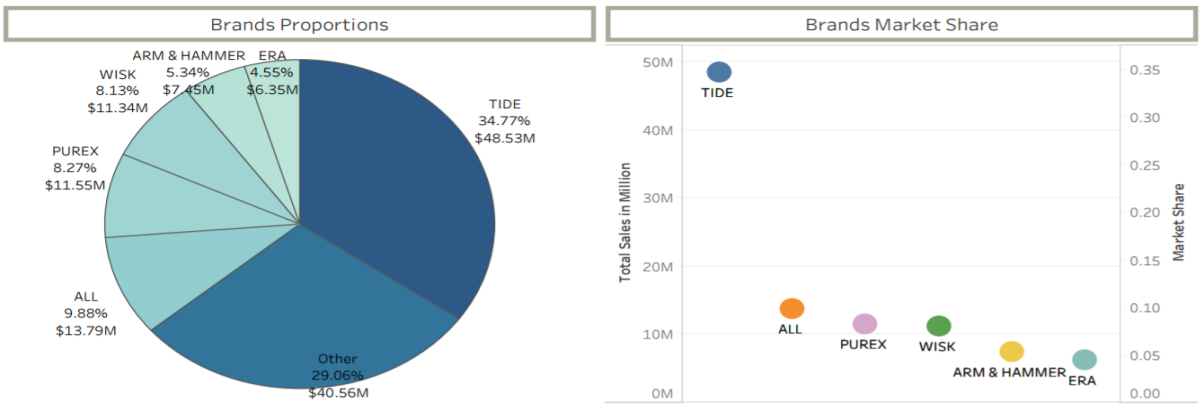
In this project, we handled four types of data:
Store level data: We had grocery stores and drug stores data and the location of the store by market.
Panel data: We had buying data of families at grocery, drug and merchandise stores. In this we had dollar sales, units bought of distinct brands (by UPC) across households and stores.
Product data: Consists of brand names and description of those brands.
Demographic data: Had data pertaining to customer demographics.

Data Merging and aggregation steps

In this we did the following steps:

- Joined Drug and Grocery store level to product data to get the description of the products sold.
- Joined Panel Data of drug and grocery store with store level data to get information of whether the product bought by the family was on discount or not or was on display/featured or not.
- We then calculated price Per Unit, price Per Volume unit, and total Volume bought.
- We then aggregated the data by grouping on Panelist ID, WEEK and individual product level.

Descriptive statistics



Sales by Form and Store Type

	Type	
Form	Grocery	Drug
Grand Total	$199.42M	$5.44M
LIQUID	$135.97M	$4.37M
POWDER	$63.36M	$0.99M
POWDER OR LI..	$0.03M	$0.09M
PLASTIC	$0.07M	$0.00M

Product Types Count by Form

L5	Liquid	Powder
ALL	46	21
ARM & HAM..	55	22
ERA	10	
PUREX	64	52
TIDE	100	131
WISK	34	25

Product Types Count by Package Type and Form

L2	Package	ALL	ARM & HAMMER	ERA	PUREX	TIDE	WISK
Liquid	MISSING	1					4
	PLASTIC BOTTLE	44	54	9	59	94	28
	PLASTIC BOTTLE IN BX					1	
	PLASTIC BOTTLE REFIL						2
	PLASTIC CONTAINER		1			1	
	PLASTIC JUG	1			7	3	2
	PLASTIC WRAPPED BSKT					1	
	REFILL PLASTC BOTTLE	3		1		2	1
Powder	BOX	21	22		48	126	25
	BOX W/PLASTIC HANDLE					1	
	PLASTIC BAG REFILL					3	
	PLASTIC BOTTLE				4	2	

- The brand Proportions indicate that TIDE (P&G) is the market leader and the second highest brand is ALL (UNILEVER). We choose ALL as our brand and decided to analyse and improve this Brand.
- Drilling down further we specifically select the liquid detergent category because the sales for liquid detergent are better compared to powder detergent.
- Other interesting fact is increasing the number of product lines does not necessarily increase the sales. Purex and Arms&Hammer have many different product lines but their sales are still less than ALL.

Hypothesis testing

First Test: To check whether family size affects total amount spend. Ideally it should differ, larger family means more use of laundry detergent and which means more spending.

H0: There is no difference on total amount spent based on the family size

Ha: There is difference on total amount spent based on the family size

Family size				
Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	2617	-1.36	**0.1731**
Satterthwaite	Unequal	1195	-1.34	0.1801
Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	695	1922	1.07	**0.2702**

F test: The p value for the F critical is 0.2702 at 95% confidence. So, the variances are equal.
T test: The p value for T critical for equal variances is 0.1731 at 95% confidence. So, we fail to reject the null hypothesis
We conclude that the total amount spent does not differ based on the family size, which is opposite of what we thought.

Second Test: To check whether presence of pets affects total amount spend. Ideally it should differ, pets present in a family means more use of laundry detergent and which means more spending.

H0: There is no difference on total amount spent based on the presence of pets

Ha: There is difference on total amount spent based on the presence of pets

pet				
Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	2617	-3.15	0.0016
Satterthwaite	Unequal	2423.8	-3.13	**0.0018**
Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	1250	1367	1.48	**<.0001**

F test: The p value for the F critical is <.0001 at 95% confidence. So, the variances are unequal
T test: The p value for T critical for equal variances is 0.0018 at 95% confidence. So, we reject the null hypothesis.
We conclude that the total amount spent thus differs based on the presence of pets which is in line with what we thought.

Third Test: To check whether Family ethnicity affects total amount spend. Ideally it should not differ, pets present in a family means more use of laundry detergent and which means more spending.

H0: There is no difference on total amount spent based on ethnicity being Hispanic or non-Hispanic.

Ha: There is difference on total amount spent based on ethnicity being Hispanic or non-Hispanic.

Hispanic Non Hispanic				
Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	2617	1.04	0.2963
Satterthwaite	Unequal	16.695	1.88	**0.0781**
Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	2601	16	3.29	**0.0073**

F test: The p value for the F critical is 0.0073 at 95% confidence. So, the variances are unequal.
T test: The p value for T critical for unequal variances is 0.0781 at 95% confidence. So, we fail to reject the null hypothesis. .
We conclude that the total amount spent does not differ based on the being Hispanic or non-Hispanic which is in line with what we thought.

Regression on Grocery store data

Objective

We considering our self as Mangers for Brand 2 (ALL) with focus on research of our brand and on keeping an eye on our competitor's market. So, we thought of predicting the average amount spent (dollar value) for our brand and analyzing the effects of Feature, Display and Price Reduction on our Sales amount.

Output

We took the subset from the grocery data to get our Brand 2 (i.e. Brand 2/ALL) using Proc SQL.

We did Linear regression with step-wise selection of variables (`stb VIF Collin`) on the grocery store level for our Brands. Found the factors that affect the sales of both the brands.

Variable	Parameter Estimates	P-vaue	Standardise Estimates	VIF
Intercept	37.73158	<0.001	0	0
Average_Price per Volume	-29.347	<0.0001	-0.069	1.37
Average Feature	13.59	<0.0002	0.217	1.58
Average Display	24.65	<0.0003	0.211	1.299
Average Price Reduction	4.138	<0.0004	0.033	1.712
R Square	0.1738			
Adjusted R square	0.1738			

Analysis

- R^2: Here R^2 is 17.38% i.e. approx. 17% of variance in dependent variable can be explained by the independent variable and suggests it's a good model for the respective industry
- By observing the result (**Pr > |t|**) we can say that the Coefficient for all the independent variables **is statistically Signiant because its p-value is less than 0.05 for all variables.**
- The VIF and Condition index indicate **that there is no multi collinearity or linear dependency**.
- As expected if average price per volume of the product increases by one dollar the average amount spent decreases by approximately 29 dollars.
- If there is any feature present compared to no feature at all then the average amount spent increases by 13 dollars.
- If there is any In-store Display compared to no display then the average amount spent increases by 25 dollars.
- If price is reduced then the average amount spent increases by 4 dollars.

Post identifying them, we can predict the Average Amount Spend (Dollars) by forming linear equation using the significant variables from the above analysis.

Avg_amt_spend (Brand2) = 37.73 - 29.347 * (Avg price per VOL_EQ) + 13.59 * (Feature type) + 24.65 * (Display type) + 4.138 * (Price Reduced - Yes/No)

Regression on Panel Data of Grocery stores

Objective

We decided to run Panel data regression on only "ALL" brand to observe the buying behaviour of different families over different time periods and to check the effect of price increase, different promotion strategies.

Data Preparation Steps

We have merged the panel data file which is demographics about the panellist and grocery transaction data for the panel data analysis. We filtered our data for only our brand "ALL". We aggregated feature, display, price reduction across a single week using appropriate weights (units sold).

Output

After running the proc panel model with fixed effects and random effects, the results of the fixed effects are as follows:

- F Test for no fixed effects is significant i.e P(r) > F

So, we reject the null hypothesis that there is fixed effects present and conclude that there is Random Effects in our data and decide to run Panel Data regression with Random effects.

Variable	**Estimate**	**Pr > |t|**	**Variable**	**Estimate**	**Pr > |t|**
Intercept	7.714075	<.0001	FAM_SIZE2	-0.01952	0.9355
avg_pricePerVOLEQ	**-2.54196**	**0.0003**	ISHISPANIC1	-0.81176	0.3933
AVG_F4	**2.286522**	**<.0001**	CPIHH	0.023038	0.5288
AVG_F3	**1.983878**	**<.0001**	CGC	0.001631	0.9727
AVG_F5	**-1.35256**	**0.0269**	TRP1	-0.31379	0.2172
AVG_F2	1.545266	0.3332	MS0	-0.75116	0.545
AVG_D2	-0.0825	0.6606	MS1	0.033162	0.9338
AVG_D1	**2.190476**	**<.0001**	MS3	-0.25244	0.4046
AVG_PR1	**-0.87627**	**0.0023**	**MS4**	**-0.65535**	**0.0326**
PETPRESENT1	**0.359453**	**0.0454**	MS5	0.450193	0.4307

Analysis

- R^2: In the random effects model, the R^2 is 8.36% that means the explanatory variables explain about 8% of the variation in the dependent variable.
- The significant variables are Average Price per Volume, Feature 4(Large size ad), Feature 3(medium size ad), Feature 5(A+ ad – also known as "Q" or "R" – retailer coupon or rebate), Display 1 (Minor), Price Reduction, Pet present, Marital Status 4(Widowed).
- As expected as average price per volume of a product increases by $1 then the total dollar amount spent decreases by $2.54.
- Features –
    - If there is a retailer coupon or rebate compared to no feature in that particular week, the total amount spent **decreases by $1.35. This is surprising**.
    - If there is a large size Ad compared to no feature (or no Ad) in that particular week, the total amount spent increases by $2.28.
    - If there is a medium size ad compared to no feature in that particular week, the total amount spent increases by $1.98.
    - The effect of large size ad and medium size ad is not significantly different.

    - But C - small ad, usually 1 line of text is not significantly different then not having any feature i.e. the effect of a small ad as to that of no feature is the same.
    - To summarize: **Large size Ad = Medium Size Ad > Small Ad = No feature > Retailer Coupon or rebate**.
- Display –
    - If the display is minor as compared to no display, then the total amount spent increases by $2.19.
    - A Major display is not significantly different then not having a display at all.
    - To Summarize: **Minor Display > Major Display = No Display**
- **Contrast to what we thought**, if there is Price Reduction compared to no price reduction, then the total amount spent decreases by $0.88.
- As expected if there is a pet present compared to no pet present, then the total amount spent increases by $0.36.
- If the marital status is widowed compared to married, then the total amount spent decreases by $0.65. The effect of single, divorced, separated is the same as that of the married individual (our base group).

Cutomer Segmentation and RFM analysis

Objective

The purpose of segmentation and RFM analysis is to get significant insights into the characteristics of consumers in the laundry detergent market. So, we are not restricting this segmentation to just our brand, as we are not the market leader and there is a huge difference in market share when compared to top brand (TIDE). We would like to get the overall idea of the different types of consumers in this product segment.

Steps followed

- Step1, did RFM analysis on Panel grocery data. We decided to use 3 bins for performing RFM. So, expected 27 segments.
- Step2, checked correlation between R, F, and M values. Found out F and M are highly correlated, so dropped M and our segments reduced to 9.
- Step3, we then joined demographic data with the RFM scores and then performed cluster analysis on it.
- Step4, based on the dendogram decided to have 4 clusters.

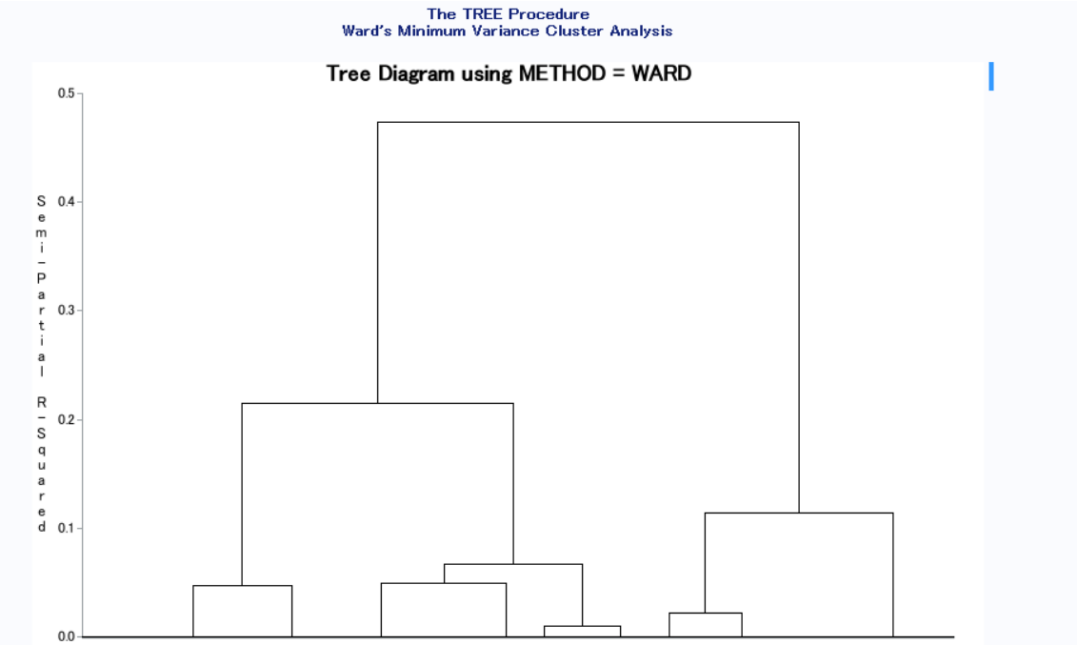
Output

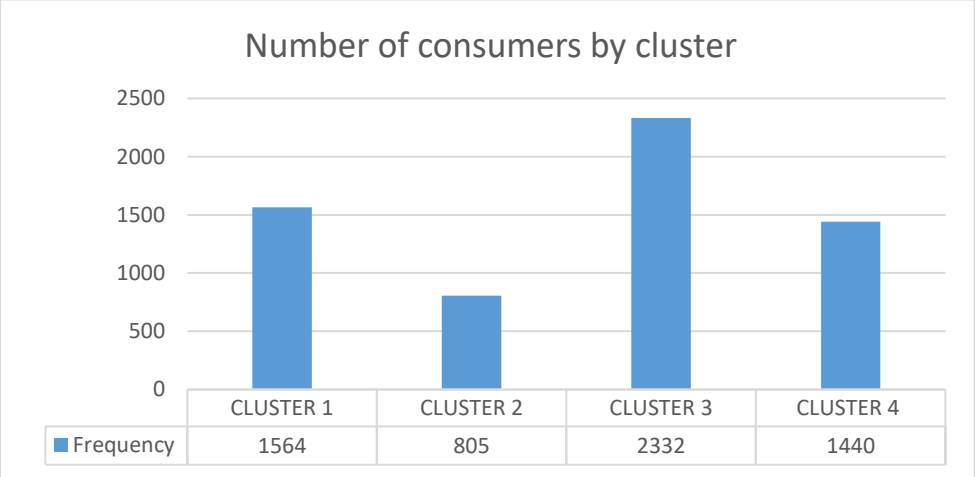
Correlation between R, F and M values:

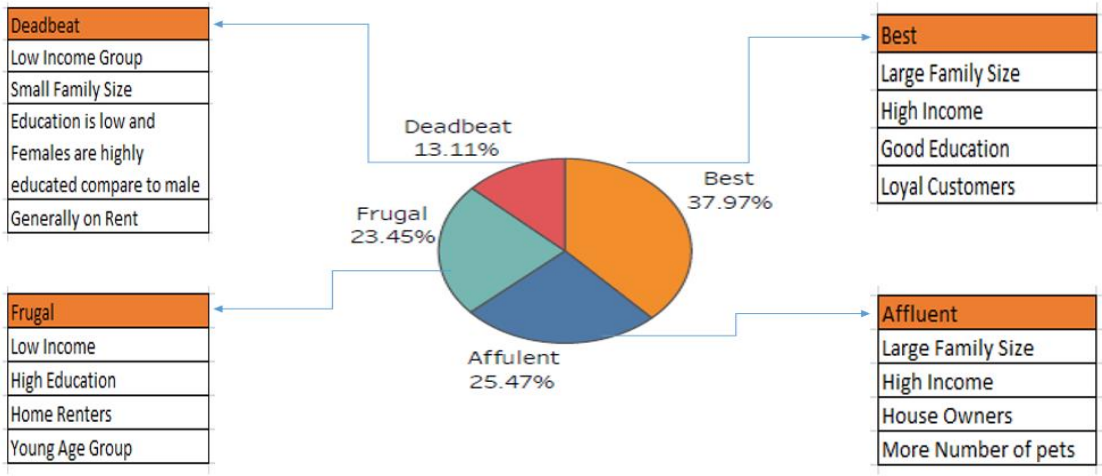
Pearson Correlation Coefficients, N = 6141
Prob > |r| under H0: Rho=0

	R	F	M
R Rank for Variable day	1.00000	0.51578 <.0001	0.45319 <.0001
F Rank for Variable freq	0.51578 <.0001	1.00000	0.81768 <.0001
M Rank for Variable monetary	0.45319 <.0001	0.81768 <.0001	1.00000

Dendogram:


The TREE Procedure
Ward's Minimum Variance Cluster Analysis
Tree Diagram using METHOD = WARD
Semi-Partial R-Squared
0.5
0.4
0.3
0.2
0.1
0.0


Number of consumers by cluster
2500
2000
1500
1000
500
0
CLUSTER 1
CLUSTER 2
CLUSTER 3
CLUSTER 4
Frequency
1564
805
2332
1440


Deadbeat
Low Income Group
Small Family Size
Education is low and Females are highly educated compare to male
Generally on Rent
Best
Large Family Size
High Income
Good Education
Loyal Customers
Deadbeat 13.11%
Best 37.97%
Frugal 23.45%
Affulent 25.47%
Frugal
Low Income
High Education
Home Renters
Young Age Group
Affluent
Large Family Size
High Income
House Owners
More Number of pets

Analysis

Based on the above findings we can segment the market into 4 clusters.

CLUSTER 1: **Affluent Customers: High Income, Large Families**
Market Share: 25.5%.
Brings the highest monetary value, they are also frequent buyers of laundry detergents. here are a few key characteristics.

1. Largest Family size.
2. Cluster with the highest combined pretax income.
3. Have their own house and generally do not rent.
4. Tend to have the highest number of pets.

CLUSTER 2: **Deadbeat Customers: Young professionals**
Market Share: 13%.
Brings the highest monetary value after cluster 1.

1. Don't have a high income, however due the less dependents (smallest family size), they have a higher purchasing power.
2. Not very highly educated, but the females of the household in this cluster have better education than the males.
3. Almost always are home renters and do not own property.

CLUSTER 3: **Best Customers: Educated Middle class families**
Market Share: 38%.
Contrary to what we thought, these consumers don't bring much value to the market.

1. Consumers in this cluster have large families and come from a good education with a higher income.
2. They usually home owners and own pets.
3. Both males and females have high education levels.
4. This cluster also comprises of older consumers hence we see a lower demand from them.

CLUSTER 4: **Frugal Customers: Low income, highly educated**
Market Share: 23.5%.
Brings the lowest value to the market.

1. Having a low income yet being highly educated could mean this cluster comprises of students.
2. Usually are home renters.
3. Small family size.
4. Young age.

Recommendations

1. Target the cluster “Best Customers” and “Affluent Customers”. The overall Market share of these would be around 53.5%. Also, the family size of both the groups is more, the consumption of laundry detergent will be more. Since they have high income and they are home owners with most of them owning pets, the consumption of laundry detergent is high. Thus, we conclude that targeting these segments with appropriate business strategies would generate more sales for our brand “ALL”.
2. Based on Standardized Beta Values the most important variable affecting sales is Features and next is Display. So instead of focusing on reducing the price, if we improve our strategy in Feature and Display our sales would increase.
3. Focus more on having Minor Display of our products and not focus on Major Display. Also, this will help reduce cost as to have a Major Display in stores would be expensive than having minor displays.
4. Focus on Ads of Medium Size in stores as compared to other type of Ads.
5. Price reduction is not helping. So, no need to reduce price for our products that are sold in grocery stores.

Challenges Faced while preparing data for Multinomial Logistic regression

The Data that we were using to model MNL was Panel Drug transaction data.

So, when we were grouping the data by week and store ID there were a lot of missing values for Price, Display, Feature and price reduction - because it was not necessary that in that particular week in that particular store all of the top four brands were brought. For example: there were a lot of missing values for WISK which is the 3rd most top brand by sales.