Group 3

Brand Analysis of Milk Products

Submitted by:

Anubha Tripathi

Ishika Bhargava

Sanskriti Vijay

Shirin Sehgal

Table of Contentsl

1 Introduction 2

2 Descriptive Analytics 2

2.1 Exploratory Data Analysis 2

2.1.1 Total Market Share by Brands 2

2.1.2 Total sales by region 3

2.1.3 Weekly Sales Trend 4

2.1.4 Brand wise advertisement, price reduction and display 4

3 Predictive Analytics 5

3.1 Data Preparation 5

3.2 Regression Analysis 5

3.2.1 Key Insights 7

3.3 Interaction between price and feature (advertisement) 7

3.3.1 Key insights 8

3.4 Difference in means- product sub-category using ANOVA: 8

3.4.1 Key Insights 8

4 Prescriptive Analytics 9

4.1 Recommendations 9

1 Introduction

Milk is one of the most produced agricultural commodities worldwide. It is used in all the households, making it a necessity. We have conducted the analysis of the milk at the brand level. Identify top 4 brands based on highest sales. Additionally, the entire report is structured the in 3 categories:

1. Descriptive Analytics: Analyzed the data using data aggregation and data mining methods and used them to uncover trends and patterns.
2. Predictive Analytics: Used statistical modelling to identify the likelihood of future outcomes based on historical data
3. Prescriptive Analytics: Emphasized on actionable insights using the trends and output uncovered at the descriptive and predictive level

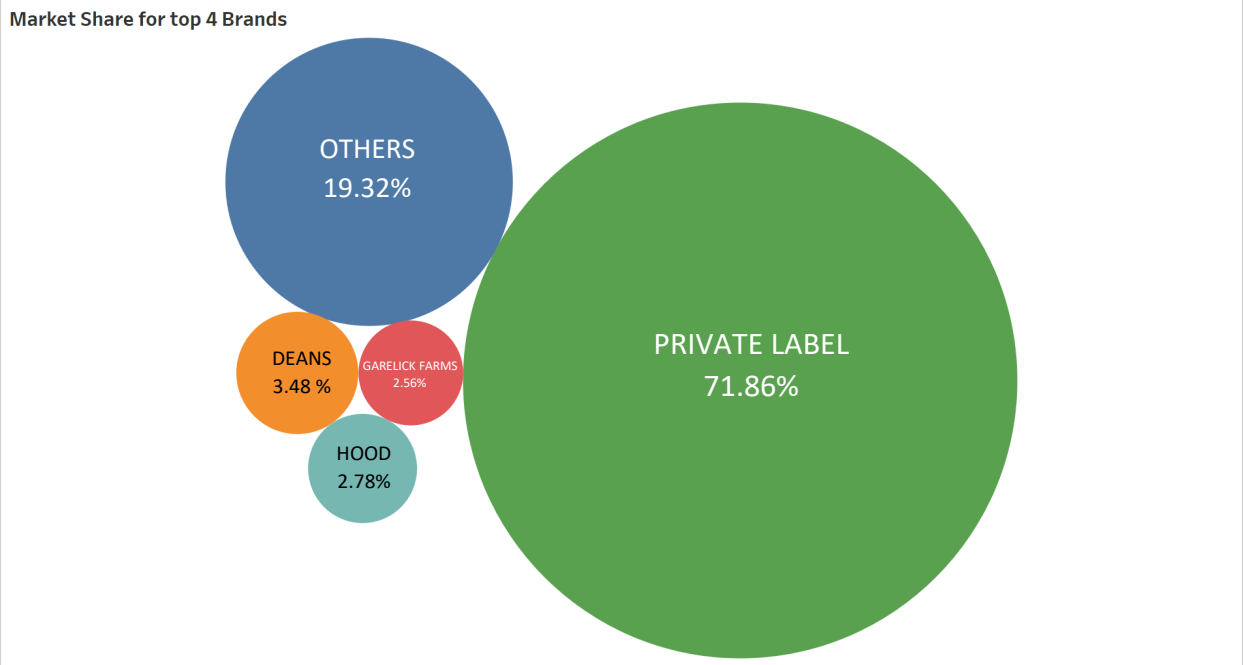
2 Descriptive Analytics

2.1 Exploratory Data Analysis

To gain a better understanding of the data and to delineate our area of focus, we have done an initial study on the number of categories each variable has. Those brands are: - Private Label, Deans, Hood and Garelick Farms. We have looked at descriptive statistics of brands like total market share, sales by region and weekly trends of top four brands.

Through our study we have tried to convey what variables affect the sales of these brands. We further identified the promotional activities undertaken by each brand to increase their sales.

2.1.1 Total Market Share by Brands

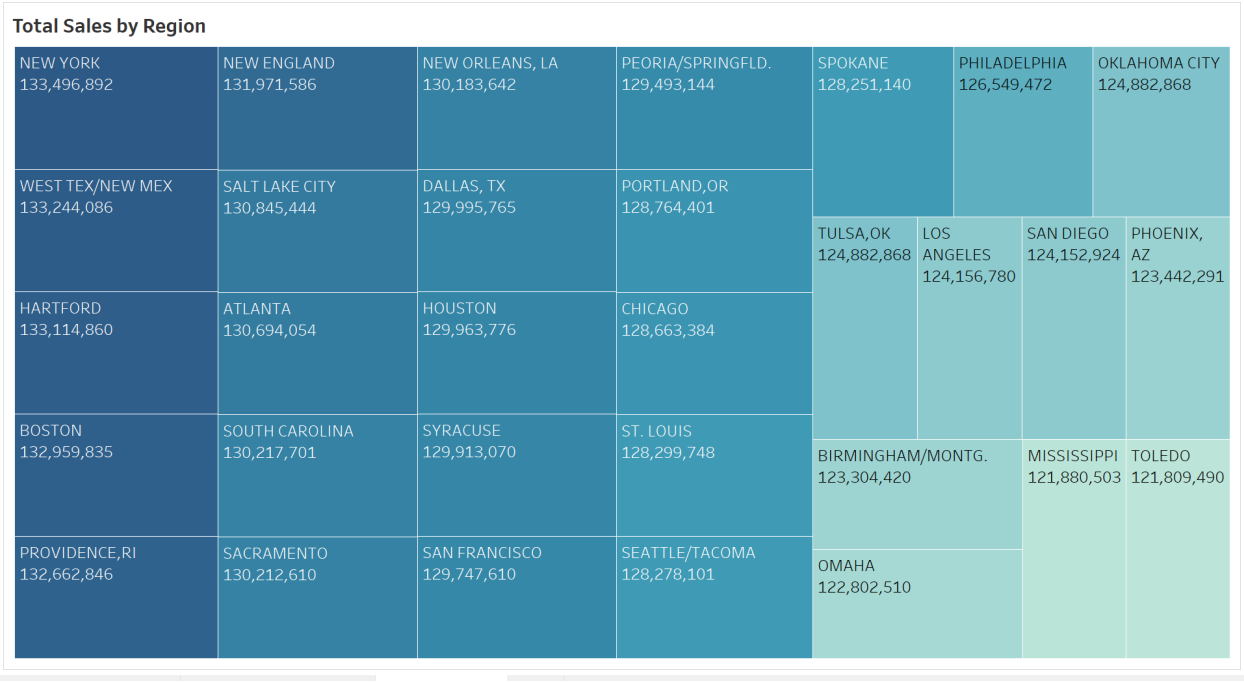


As can be seen in the pie chart above, Private Label is the Market leader in Milk brands with about 72% of market share.

The other leading brands competing with each other include Deans, Hood, and Garelick Farms with 3.48%, 2.78% and 2.56% respectively and the remaining brands contribute 19% of the market share.

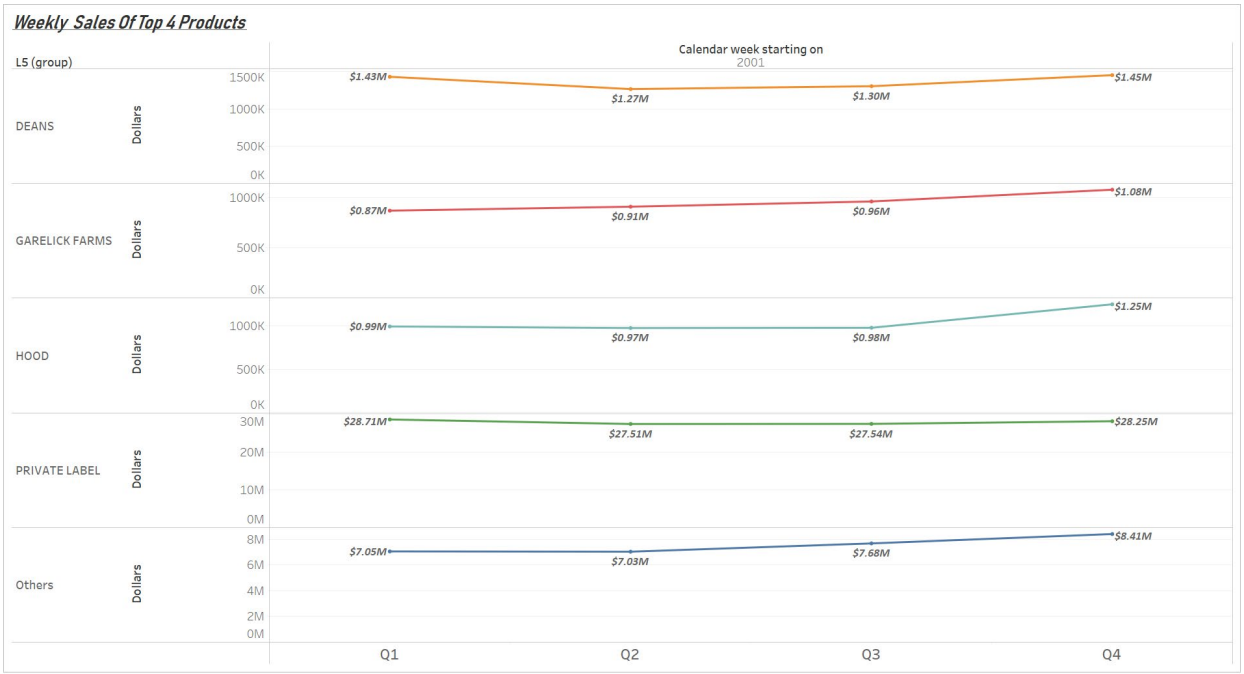
2.1.2 Total sales by region

We will further be checking the distribution of the market share by region to better understand which brand is preferred in which region.



We can infer from the tree map that the most sales are in the cities New York, West Texas, Hartford, and Boston. Three of the four cities with the highest sales being in the east region implies that the east region comprises of the most sales of milk products and dominate the entire market.

2.1.3 Weekly Sales Trend



Based on sales, we have identified the top 4 brands which cover around 80% of the market. Based on the quarterly trends across each milk brand, we can infer that the highest player**, Private Label** is performing consistently good with slight increase in sales sporadically while other brands with highest sales show an increasing trend across the quarters.

2.1.4 Brand wise advertisement, price reduction and display

L5_new	Advertisement	price_reduction	display
Private Label	19381	37660	3081
Deans	4301	5646	177
Hood	1357	1778	128
Garelick Farms	1293	2029	152
Others	34130	103004	1769

We did an analysis of brands with respect to their promotional efforts- advertising, price reduction and display.

We found out that the frequency of advertising, price reduction and display is highest in Private Label brand followed by Deans, Hood, and Garelick Farms.

Investing in promotional activities might be an underlying reason of the highest market share of Private Label indicating that promotional efforts accounts for a significant customer base.

3 Predictive Analytics

3.1 Data Preparation

In order to draw conclusive results, we carried out an extensive data cleaning process. To summarize:

1. Joined the data from all the three data tables i.e., Milk Groc, Delivery Stores and Product Milk using UPC code and IRI Key
2. Calculated price per ounce using Dollars/ (units*Size (in OZ))
3. Checked the total dollar sales by brands, to check which brands are contributing highest sales thereby leading to capture largest market share (as concluded at the descriptive stage) and identified top 4 brands
4. Created dummy variables for top 4 brands based on sales
5. Created dummy variables for based on highest frequency by identifying top 4-5 categories for the variables: feature (f), Name of group (L3), Product sub-category(L2), Name of company within a group (L4), Product type, Package, Process, Fat content, Flavor scent and Display (d)
6. For accurate analysis, checked correlation between all the variables as well as checked multicollinearity using VIF in the regression equation. As per the output, some variables had VIF greater than 5, so eventually dropped the variables like FLAVOR_REGULAR, L4_DEAN, FLAVOR_WHITE, PRODUCT_MILK because they had high correlation with other variables
7. Based on highest frequencies, we have classified our data into categories: -
    a. Product Type – Milk, Milk substitute, Buttermilk, eggnog
    b. Package- Carton, plastic jug, plastic bottle
    c. Process- pasteurized hmgnzd, ultra-pasteurized, ultra pstrz hmgnzd, organic, pasteurized, homogenized, cultured
    d. Fat Content- 2% Milk Fat, 1% Milk Fat, whole, reduced fat, NT STT on PKG FR NND, Low fat
    e. Flavor Scent- white, chocolate, regular, vanilla, strawberry
    f. L2 - RFG FLAVORED MILK/EGGNOG/BUTTERM, RFG KEFIR/MILK SUBSTITUTE2S/SOYMI, RFG WHOLE MILK, RFG MILKSHAKES/NON-DAIRY DRINKS
    g. L4- Private label, white wave inc, mcneil consumer products, nestle USA inc, Deans Food co, Garelick farms inc

3.2 Regression Analysis

To know the effect on sales of milk across the brands, PROC REG was run with sales as a dependent variable and L2, L3, L4, L5, product type, process, package, fat content, Flavor scent, Display, Feature, price deduction, volume, and price per unit as independent variables.

Model Fit Stats: - We ran a model with 1,055,061 observations. Based on R square and adjusted R square which is equal to 0.33 implies that all the explanatory variables explain 33% of the variance in the dependent variable ‘Sales' indicating a good model fit.

Parameter Estimates

Variable	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	-282.96674	2.64428	-107.01	<.0001	0
VOL_EQ	78.2385	0.25564	306.05	<.0001	0.37709
price_unit	1427.29988	24.37792	58.55	<.0001	0.06933
F_flag	129.01669	1.79095	72.04	<.0001	0.06765
PR	27.8334	1.20567	23.09	<.0001	0.02193
DISPLAY	247.34234	5.02047	49.27	<.0001	0.03947
L3_Dean	-7.03897	1.40642	-5	<.0001	-0.00735
L3_Johnson	17.22659	1.46477	11.76	<.0001	0.01147
L3_NESTLE	120.32846	1.8191	66.15	<.0001	0.07135
L3_PRIVATE	0	.	.	.	.
L4_PRIVATE	347.69502	1.09353	317.96	<.0001	0.32928
L4_WHITEWAVE	6.78008	1.85719	3.65	0.0003	0.00524
L4_MCNEIL	0	.	.	.	.
L4_NESTLE	0	.	.	.	.
L4_GARELICK	-151.49106	4.51482	-33.55	<.0001	-0.06006
L5_PRIVATE	0	.	.	.	.
L5_DEANS	109.47902	2.1035	52.05	<.0001	0.05116
L5_HOOD	48.28984	2.32289	20.79	<.0001	0.01737
L5_GARELICK	197.75216	4.86938	40.61	<.0001	0.06984
PRODUCT_MILKSUB	101.0426	1.90219	53.12	<.0001	0.07139
PRODUCT_BM	-63.30834	1.75113	-36.15	<.0001	-0.03137
PRODUCT_EGGNOG	-36.33675	2.81723	-12.9	<.0001	-0.01112
FAT_FREE	6.10579	1.35516	4.51	<.0001	0.00543
FAT_MILKFAT	119.25667	1.35606	87.94	<.0001	0.10384
FAT_1Milkfat	-27.5732	1.32336	-20.84	<.0001	-0.02363
FAT_WHOLE	78.11091	1.35653	57.58	<.0001	0.06503
FAT_REDUCE	-18.89422	2.1233	-8.9	<.0001	-0.0082
FAT_NSOPFN	0	.	.	.	.
FAT_LOW	0	.	.	.	.
PACKAGE_CARTON	-82.54782	1.46742	-56.25	<.0001	-0.09311
PACKAGE_JUG	18.27444	1.71378	10.66	<.0001	0.01752
PACKAGE_BOTTLE	-17.29968	1.68887	-10.24	<.0001	-0.01414
FLAVOR_CHOCOLATE	-65.9332	1.07528	-61.32	<.0001	-0.06001
FLAVOR_VANILLA	-33.66394	2.11536	-15.91	<.0001	-0.01536
FLAVOR_STRAWBERRY	6.24642	2.13016	2.93	0.0034	0.00275

We are looking at standardized betas to draw our analysis. Through Standardized Betas, we found out that Product subcategories(L4), Volume and Milk Fat are the most impactful variables on sales.

3.2.1 Key Insights

1. Through our observation, we found out that if a person chooses Fat_Milkfat, sales will increase by 119 dollars, which is quite impactful.
2. Since Volume is the most important variable in our model as per standardized beta, we can say that one ounce increase in volume can result in 78 dollars of increase in the sales.
3. If there is 5% Deduction in the price, sales will go up by 28 dollars. This reflects the fact that the promotional activities like price reduction have a positive impact on sales.
4. Packaging has some of the important effects on the sales as well; if the packaging is carton or bottle, then the sales would decrease by 83 and 17 dollars respectively whereas if the packaging is jug, sales will increase by 18 dollars indicating jug is the most important packaging feature amongst the customers. This could mean that customers prefer a product if the type of packaging is jug, thereby increasing the sales.
5. Feature being a significant attribute, we found out that if there is the presence of brand in a store flyer then sales will increase by 129 dollars.
6. If there is an in-store display highlighting the fact that the brand or UPC is on a special deal, then the sales would increase by 247 dollars.
7. Amongst the brands, Garelick has the most significant impact on the sales. If the customer chooses Garelick, the sales will increase by 198 dollars.

3.3 Interaction between price and feature (advertisement)

Different stores have carried out various promotional activities like:

1. Feature(F_flag): It means that a brand is being featured in a store flyer. This is an advertisement that brands have carried out
2. Price Reduction (PR): If there was a price reduction for a product
3. Display (D): It means the in-store display highlighting if a brand is on a special deal

We would like to check the effect of these promotional activities on dollar sales. We are also interested in looking at the combined effect of price and advertisement (i.e., feature), therefore we have included an interaction between price and advertisement.

Parameter Estimates

Variable	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	400.61397	1.04481	383.43	<.0001
price_unit	-5466.44144	19.23398	-284.21	<.0001
F_flag	868.3133	4.33855	200.14	<.0001
PR	16.17204	1.35654	-11.92	<.0001

DISPLAY	390.09894	5.68586	68.61	<.0001
Price_feature	-18418	102.97592	-178.85	<.0001

3.3.1 Key insights

1. If the price is increased by a dollar, the sales will decrease by a significant amount which is 5466 dollars indicating a price sensitive market.
2. If the brand is featured on the in-store flyer, the sales are likely to increase by 868 dollars which tells us advertising has an advantageous effect on sales.
3. If there is a price reduction of 5% for the product, the sales will increase by 16 dollars which can be used as a promotional tool by the brand to increase the sales.
4. If there is an in-store display for the product, the sales will increase by 390 dollars implying one more beneficial way of promoting sales.
5. When we witness the interaction between price and featuring the product, we notice that there is a synergistic effect i.e., advertising makes the price effect stronger (as more people will know about the lower price); there is a significant increase of 18418 dollars.

3.4 Difference in means- product sub-category using ANOVA:

Here we are looking at the difference in means of sales of product sub-categories. In this case the following would be our null and alternative hypothesis:

Ho: Means of sales for different product sub-category is equal.

Ha: At least two means of sales for different product sub-category is equal.is different.

Source	DF	Anova SS	Mean Square	F Value	Pr > F
L2_new	4	12798269959	3199567490	17354	<.0001

3.4.1 Key Insights

In the above table, the p-value is <0.0001 so we could reject the null hypothesis. This means we have sufficient evidence to say that there is a statistically significant difference between at least two means of sales for different product sub-category. Therefore, we can conclude that different sub-categories would contribute different towards the dollar sales.

4 Prescriptive Analytics

4.1 Recommendations

1. Through our analysis of the market share, we saw that Private label is the market leader followed by the market followers Deans, Hood, and Garelick Farms. As we can see that some of the potential market is untapped by the market leader namely Omaha, Mississippi, Toledo, Birmingham, Phoenix, San Diego, Los Angeles, and Tulsa. Since Private Label is the market leader, it can try to occupy these regions as well and followers can choose to follow the leader's strategy or can target to acquire these markets and increase their market shares significantly to become the leader.

2. The followers Deans, Hood, and Garelick Farms can take a brand differential advantage through intangibles, perceived qualities, packaging, and distribution innovations, like using Plastic Jug as a form of packaging as it had the highest effect on the sales as compared to other packaging

3. As seen from our regression analysis, we found out that the market is price sensitive i.e., a small amount of change in the price results in a significant amount of change in sales. Since there is availability of substitutes, the price elasticity is high. The leader and followers should make sure that there are economies of scale which gives them the benefit of low prices in the market.

4. After the analysis, we cannot ignore the advertising impact on the sales. A potential segment of the customers who are coupon prone can be targeted**.** The brands can follow the pull strategy to create more demand in the form of Advertisings, rebates, premiums, and coupons. This will not only increase brand loyalty but also attract new customers.