
UTD Final Group Project

Submission 2

Group 25

| | |
|---|---|
| Pravalika Bobbala | PXB220028 |
| Claire Rosenbluth | CXR210018 |
| Yatin Kumar | YXK210022 |
| Shikhar Yadav | SXY220002 |
| Bharadwaj Narne | BXN220006 |
| Ananya Sharma | AXS220044 |

# Geographical Impacts on Unit Sales

## Approach to Data Cleaning

Data was joined to perform analysis before where the IRI POS Tablespreads data for all the years was joined with the IRI_POS_Tablespreads_Product Attribute File on the column 'UPC 13 digit code' to extract the product attributes for each product. This final dataset was called df_final.

Data Cleaning was performed before the analysis on the dataset

1. Three new columns 'Dollar Sales', 'Unit Sales', 'Volume Sales' are being added to the DataFrame 'df_final' based on calculations involving existing columns. Dollar Sales is the sum of Dollar Sales Any Merch and Dollar Sales Unit Merch and similarly Unit Sales and Volume Sales were calculated.
2. Dates were extracted from the 'Time Caller' column in the format of 'mm-dd-yyyy' and a new column called 'Date' was created.
3. Data corresponding to 'Total – US' was removed from the dataset because the data corresponding to each region was included and performed analysis on for geographical analysis.
4. Only the below columns were kept in the dataset

```
Index (['index', 'Geography', 'Product_Description', 'UPC_13_digit',
'Price_per_Unit', 'Price_per_Volume', 'Sub_Sub_Category',
       'Brand_Value','Manufacture_Value', 'CAG_Category_Value',
'CAG_Count_Value', 'CAG_Ounces_Value', 'CAG_Form_Value', 'Form',
'CAG_Tier_Value', 'Dollar_Sales', 'Unit_Sales', 'Volume_Sales', 'Date'],
dtype='object')
```

5. NA's were removed to further clean the dataset and get rid of any null values.

## Initial Insights and Observations

**Exhibit 1** contains a correlation matrix which helps us identify which variables have high correlation and can cause a multicollinearity problem in the model. The variables with high correlation were not included in our subsequent analyses.

## Geographic Analysis

### ANOVA Test to check mean unit sales across different brands

Null hypothesis: Mean Unit Sales of different brands is equal (no variation in means of groups)

Alternative hypothesis: Mean Unit Sales of at least 1 brand is different

| | sum_sq | df | F | PR(>F) |
|---|---|---|---|---|
| **Manufacture_Value** | 2.728806e+13 | 261.0 | 116.679186 | 0.0 |
| **Residual** | 5.042185e+14 | 562704.0 | NaN | NaN |

Since p-value < 0.05 we can reject NULL hypothesis and say that at least one group mean is different from other groups or Mean Unit Sales of at least 1 brand is different.

**Implication for CONAGRA**

Mean unit Sales is different for different brands and the top selling brand is Country Delight with Land O'Lakes just following. Conagra is 10$^{th}$ in the list of top brands selling Tablespreads Conagra needs to implement strategies to close the gap between the leaders in Tables spreads category and improve their position in the market.

**ANOVA Test to check mean unit sales across various CONAGRA brands**

Null hypothesis: Mean Unit Sales across different CONAGRA brands is equal

Alternative hypothesis: Mean Unit Sales of at least 1 CONAGRA brand is different

| | sum_sq | df | F | PR(>F) |
|---|---|---|---|---|
| **Brand_Value** | 1.078991e+13 | 6.0 | 1637.210911 | 0.0 |
| **Residual** | 7.612157e+13 | 69302.0 | NaN | NaN |

Since p-value < 0.05 we can reject NULL hypothesis and say that at least one group mean is different from other groups or Mean Unit Sales of at least 1 CONAGRA brand is different.

**Implication for CONAGRA**

Mean unit Sales is different for CONAGRA brands and the top selling brand amongst CONAGRA is Blue Bonnet with Parkay just following. Blue Bonnet is the most popular product amongst all the other CONAGRA brands, and it is a hugely popular brand by high margin. CONAGRA can perform a conjoint analysis to identify the features of BLUE BONNET and try to customize other products close to it so that the sales of the other CONAGRA brands can also be improved.

**ANOVA Test to check mean unit sales across various geographies for ALL brands**

Null hypothesis: Mean Unit Sales across different geography is equal

Alternative hypothesis: Mean Unit Sales of at least 1 geography is different

| | sum_sq | df | F | PR(>F) |
|---|---|---|---|---|
| **Geography** | 3.051202e+12 | 7.0 | 464.344891 | 0.0 |
| **Residual** | 5.284554e+14 | 562958.0 | NaN | NaN |

Since p-value < 0.05 we can reject NULL hypothesis and say that at least one group mean is different from other groups or Mean Unit Sales of at least 1 geography is different.

**Implication for CONAGRA**

Mean unit Sales is different for brands across different geographies and the geography with the most people buying tablespreads is Southwest followed by Northeast. People in California and Plains don't use Tablespreads that often and thus these serve as a potential market for Tables spreads category and CONAGRA can try to get the market share from these Geographies.

**ANOVA Test to check mean unit sales across various geographies for CONAGRA brands**

Null hypothesis: Mean Unit Sales across different geographies for CONAGRA brands is equal

Alternative hypothesis: Mean Unit Sales of CONAGRA brands is different across at least 1 geography

| | sum_sq | df | F | PR(>F) |
|---|---|---|---|---|
| **Geography** | 1.820051e+12 | 7.0 | 211.757699 | 1.404766e-312 |
| **Residual** | 8.509143e+13 | 69301.0 | NaN | NaN |

Since p-value < 0.05 we can reject NULL hypothesis and say that at least one group mean is different or Mean Unit Sales of CONAGRA brands is different across at least 1 geography

**Implication for CONAGRA**

As we can see from the graph mean unit Sales for CONAGRA brands is different across geographies. CONAGRA brands is famous in Southeast and has a huge market share there. From the earlier graph we can see that Northeast and Great geographic regions have a huge market share for tablespreads category but when we narrow it down to CONAGRA the unit sales in these two regions is lower and hence CONAGRA can use these regions as potential markets for growth as the demand is there for tablespread category but the supply from CONAGRA isn't. There is a huge potential for CONAGRA to tap the customers and target Core defectors in these geographies who are willing to switch to CONAGRA.

**Capacity (Ounces) Analysis**

**Exhibit 2** contains pie charts which show the percentage of Unit Sales captured by the various levels of CAG Ounces Value both for all brands in the dataset and only CONAGRA brands.

**Implications for CONAGRA**

The first graph shows the top five ounces values for Tablespreads currently in the market irrespective of geography and the second graph shows the top five ounces for CONAGRA brand irrespective of geography. As we can see the highest selling ounce capacity for tablespreads in general is 16 OZ but for CONAGRA it is 15 OZ which means that people are inclined to buy 15 OZ packages more for CONAGRA brands because they derive more utility out of a those.

Another observation is that for CONAGRA the highest selling packages are 15 and 16 OZ which results to about 52% of the units sold.

It is also worthwhile noting that the second most popular package is the 8 OZ amongst all brands (apart from CONAGRA), but CONAGRA doesn't have the 8 OZ packaging in the top five which can serve as an opportunity for CONAGRA, if they start making more 8 OZ packaging, they can tap that segment of customers and expand their Unit Sales and hence their market share. Yes, that would impact costs but in the long run it would serve as a good option since a huge chunk of the population (nearly more than $1/4^{th}$) prefer 8 OZ packaging and it would help them break-even in the long run which would offset the cost of manufacturing.

**Capacity Analysis (ounces) for every Geography**

**Exhibit 3** contains bar charts which detail the percent of Total Unit Sales captured by each level of CAG Ounces Value across the 8 regions for all brands in the dataset.

**Implications for CONAGRA**

The above graphs show the capacities popular in each geographic region which further asserts our previous analysis that 16 OZ is the most popular one followed by 8 OZ or 15 OZ in respective geographies. Since the 8 OZ pack is more popular after 16 OZ across geographies CONAGRA should include more 8 OZ packages in their basket to tap that market segment because they already provide 15 OZ packaging.

**Capacity Analysis (ounces) for every Geography for CONAGRA brands**

**Exhibit 4** contains bar charts which detail the percent of Total Unit Sales captured by each level of CAG Ounces Value across the 8 regions for only CONAGRA brands.

**Implications for CONAGRA**

The above graphs show the capacities popular in each geographic region for CONAGRA brands which further asserts that 16 OZ is the most popular one but as seen in the graphs earlier the second most popular option is 8 OZ one which is not here for CONAGRA brands which asserts

the point that this could be a potential growth opportunity for CONAGRA. If they start producing 8 OZ packaging, then they can tap that market segment and increase their unit sales.

**Form analysis for every Geography**

**Exhibit 5** contains bar charts which detail the percent of Total Unit Sales captured by each level of CAG Form Value across the 8 regions for only all brands in the dataset.

**Implications for CONAGRA**
As we can see from the graphs the most popular form is the tubs form and the second main category is sticks this is because a long, thin stick of butter has a high surface area to volume ratio. This means that margarine in this form will soften relatively quickly at room temperature and hence it is more popular. This might serve as an opportunity for CONAGRA to boost its unit sales, which might be ascertained after analyzing CONAGRA data.

**Form Analysis for every Geography for CONAGRA brands**

**Exhibit 6** contains bar charts which detail the percent of Total Unit Sales captured by each level of CAG Form Value across the 8 regions for only CONAGRA brands.

**Implications for CONAGRA**
Upon analyzing CONAGRA data we see that Tubs and other forms are popular among CONAGRA customers, but sticks aren't even when it is popular for other brands. As per our previous analysis it can be asserted further that stick form can be an opportunity for CONAGRA to increase its Unit Sales and market share because of the popularity of this form.

**Packaging Count Analysis for Every Geography**

**Exhibit 7** contains bar charts which detail the percent of Total Unit Sales captured by each level of CAG Count Value across the 8 regions for all brands in the dataset.

**Implications for CONAGRA**
As we can see from the graphs the most popular packaging count is the 4 that means the packaging that has 4 pieces is most popular amongst consumers in most geographies and in the second most popular is 1 piece. Now let's see CONAGRA packaging count to see whether this can be an opportunity for CONAGRA or not.

**Packaging Count Analysis for Every Geography for CONAGRA brands**

**Exhibit 8** contains bar charts which detail the percent of Total Unit Sales captured by each level of CAG Count Value across the 8 regions for only CONAGRA brands.

## Implications for CONAGRA

Upon analyzing CONAGRA data we see that the most popular packaging count for CONAGRA brands is 1 CT, but in the last analysis we saw that 4 CT is most popular amongst consumers in general. Hence, we can suggest CONAGRA to include packaging of 4CT for its brands because of the more popularity of those. It can further be asserted by human behavior as a consumer won't go to the market to buy 1 CT because it finishes up early but would rather prefer buying a packaging that lasts and has more count of sticks.

## Initial Regression Model

```
Unit Sales = A + B * Geography_Great + C * Geography_Mid + D *
Geography_Northeast + E * Geography_Plains + F * Geography_South + G *
Geography_Southeast + H * Geography_West + Ê
```

**Base for Geography** – Geography_California

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Unit_Sales | R-squared: | 0.021 |
| Model: | OLS | Adj. R-squared: | 0.021 |
| Method: | Least Squares | F-statistic: | 190.9 |
| Date: | Sat, 22 Apr 2023 | Prob (F-statistic): | 2.49e-281 |
| Time: | 19:40:25 | Log-Likelihood: | -7.4119e+05 |
| No. Observations: | 62378 | AIC: | 1.482e+06 |
| Df Residuals: | 62370 | BIC: | 1.482e+06 |
| Df Model: | 7 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 5990.0972 | 495.380 | 12.092 | 0.000 | 5019.151 | 6961.043 |
| Geography_Great | 5736.0468 | 622.406 | 9.216 | 0.000 | 4516.130 | 6955.964 |
| Geography_Mid | 1.169e+04 | 618.464 | 18.902 | 0.000 | 1.05e+04 | 1.29e+04 |
| Geography_Northeast | 3175.0749 | 609.167 | 5.212 | 0.000 | 1981.107 | 4369.043 |
| Geography_Plains | 1593.1174 | 629.763 | 2.530 | 0.011 | 358.780 | 2827.455 |
| Geography_South | 9325.1708 | 641.025 | 14.547 | 0.000 | 8068.760 | 1.06e+04 |
| Geography_Southeast | 1.531e+04 | 633.556 | 24.166 | 0.000 | 1.41e+04 | 1.66e+04 |
| Geography_West | 875.1096 | 654.801 | 1.336 | 0.181 | -408.302 | 2158.521 |

| | | | |
|---|---|---|---|
| Omnibus: | 83427.469 | Durbin-Watson: | 1.982 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 16321215.472 |
| Skew: | 7.791 | Prob(JB): | 0.00 |
| Kurtosis: | 80.697 | Cond. No. | 11.0 |

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

## Model Interpretation

As we can see from the above regression model run on CONAGRA brands dataset that the Unit Sales is highest for Southeast followed by Mids. On average the Unit Sales in Southwest is 1.531e+4 more than Unit Sales in California.

California has the least sales for CONAGRA brands because in the above model there is no negative signed coefficient and California is the base.

This model just explains how the Unit sales is distributed across various geographies and further analysis is required to make suggestions for CONAGRA brands which would be done in next models.

## Seasonality Impacts on Total Unit Sales (No Merch + Any Merch) by Product Attribute

### Approach to Data Cleaning

1. Added column “Month” extracting the month number from Week.Ending
2. Added column “Total Sales” column by summing No Merch and Any Merch columns for Dollar, Unit and Volume Sales
3. Added column “Total Industry Sales” which is the sum of Total Dollar Sales across all brands
4. Added column “Season” to indicate the yearly season associated with Week.Ending
5. Identified top 16 brands and aggregated all other “smaller” brands using the below process, bringing dataset down from 999K to 323K records.
    a. Calculated market share for all brands as Total Dollar Sales / Total Industry Sales
    b. Aggregated all brands with market share below 0.40% into “OTHER BRANDS” category (including PRIVATE LABEL)
6. Transformed Total.Unit.Sales to logarithmic scale to account for non-normality in this variable

### Insights/observations

**Exhibit 9** contains a boxplot to visualize the Total.Unit.Sales over the various Seasons. This plot indicates that there is a higher mean value of Total.Unit.Sales occurring in the Autumn and Winter seasons.

**Exhibit 10** contains a histogram of Total.Unit.Sales to observe the distribution of the data and identify any skewness. The data is severely right-skewed, and to correct for this, we applied a log-transformation on Total.Unit.Sales. A histogram of log.unit.sales is also included.

### ANOVA Test to check mean log.unit.sales across CAG.Count.Value

Null Hypothesis: means are equal across the groups
Alternative Hypothesis: at least one mean is different

```
                    Df   Sum Sq  Mean Sq  F value   Pr(>F)
CAG.Count.Value      8    36301     4538      612  <2e-16 ***
Residuals       321020  2380334        7
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
2080 observations deleted due to missingness
```

At 5% significance level, at least one of the means of log.unit.sales is different across groups CAG Count Value, and indicates that there is evidence to investigate this product attribute further in our analysis.

**ANOVA Test to check mean log.unit.sales across CAG.Ounces.Value**

Null Hypothesis: means are equal across all groups
Alternative Hypothesis: at least one mean is different

```
                    Df  Sum Sq Mean Sq F value Pr(>F)
CAG.Ounces.Value     36  208535    5793   842.1 <2e-16 ***
Residuals        320992 2208100       7
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
2080 observations deleted due to missingness
```

At 5% significance level, at least one of the means of log.unit.sales is different across groups CAG Ounces Value, and indicates that there is evidence to investigate this product attribute further in our analysis.

**ANOVA Test to check mean log.unit.sales across Form**

Null Hypothesis: means are equal across all groups
Alternative Hypothesis: at least one mean is different

```
> form.aov = aov(log.unit.sales ~ Form); summary(form.aov)
               Df  Sum Sq Mean Sq F value Pr(>F)
Form           14   60998    4357   597.4 <2e-16 ***
Residuals   323094 2356379       7
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

At 5% significance level, at least one of the means of log.unit.sales is different across groups Form, and indicates that there is evidence to investigate this product attribute further in our analysis.

**ANOVA Test to check mean log.unit.sales across CAG.Tier.Value**

Null Hypothesis: means are equal across all groups
Alternative Hypothesis: at least one mean is different

```
                  Df  Sum Sq Mean Sq F value Pr(>F)
CAG.Tier.Value     4   74166   18541    2557 <2e-16 ***
Residuals     323104 2343212       7
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

At 5% significance level, at least one of the means of log.unit.sales is different across groups CAG.Tier.Value, and indicates that there is evidence to investigate this product attribute further in our analysis.

**Initial Regression Formula**

```
log.unit.sales = a + b * CAG.Count.Value  +  c *
CAG.Ounces.Value  +  d * Form  +  e * CAG.Tier.Value  +  f *
Season * (CAG.Count.Value  +  CAG.Ounces.Value  +  Form  +
CAG.Tier.Value)
```

We are seeking to explain the variation in log.unit.sales using the product attributes (CAG.Count.Value, CAG.Ounces.Value, Form, CAG.Tier.Value) and their individual interactions with seasonality. From the initial models we have constructed, we see that there are clear differences between the seasons in which sales are made and the product attributes. In quantifying these main and interaction effects, we will be able to identify the impacts that product attributes and seasons of the year have on Total Unit Sales.

## Category Expansion Between Tablespreads, Cooking & Salad Oils and Cooking Sprays

**Main Idea**

Capturing category expansion is important in predictive analytics because it can provide insights into how products in a particular category are related and how changes in one product can affect the sales of other related products. For example, in this case, understanding the relationship between the sales of Tablespreads and the prices of Cooking & Salad Oils and Cooking Sprays can help Conagra optimize their pricing strategies to increase overall sales and revenue. It can also help Conagra identify potential cross-selling opportunities and adjust their product offerings accordingly.

**Approach to Analysis**

- Gather the sales data for all three categories. This data should include the number of units sold, Price per Unit, and any other relevant metrics for each category like Geography, Time etc.
- Analyze the sales data to identify any patterns or trends that may indicate category expansion. For example, looking for a significant increase in sales of cooking oil and cooking spray during the same period that sales of tablespreads have decreased.
- Using visualizations to help identify any relationships between the sales data for the three categories. This include creating line graphs and bar gto show how sales of each category have changed over time, or creating heatmaps to show correlations between different metrics for each category.

**Future Scope**

- Conduct statistical analyses to test whether any observed patterns or trends are statistically significant. This may involve running regression analyses or hypothesis tests

to determine whether changes in sales of one category are associated with changes in sales of another category.

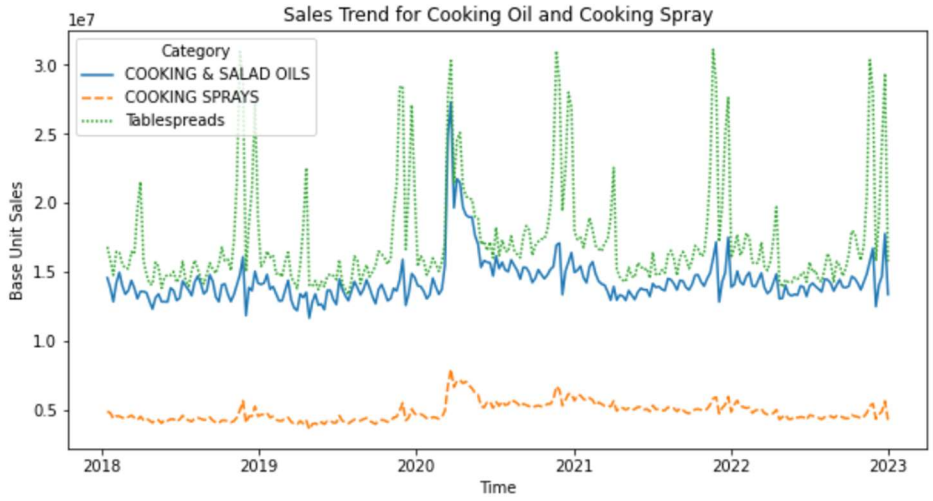
- Use the insights gained from your analysis to develop strategies for capitalizing on category expansion opportunities. For example, if you identify a significant increase in sales of cooking oil and cooking spray during a period when sales of tablespreads have decreased, you may want to consider launching a new line of cooking oils or sprays that are designed to be used in similar ways to tablespreads, or running targeted marketing campaigns to encourage customers to try cooking oil or spray as a replacement for tablespreads.

**Hypothesis**

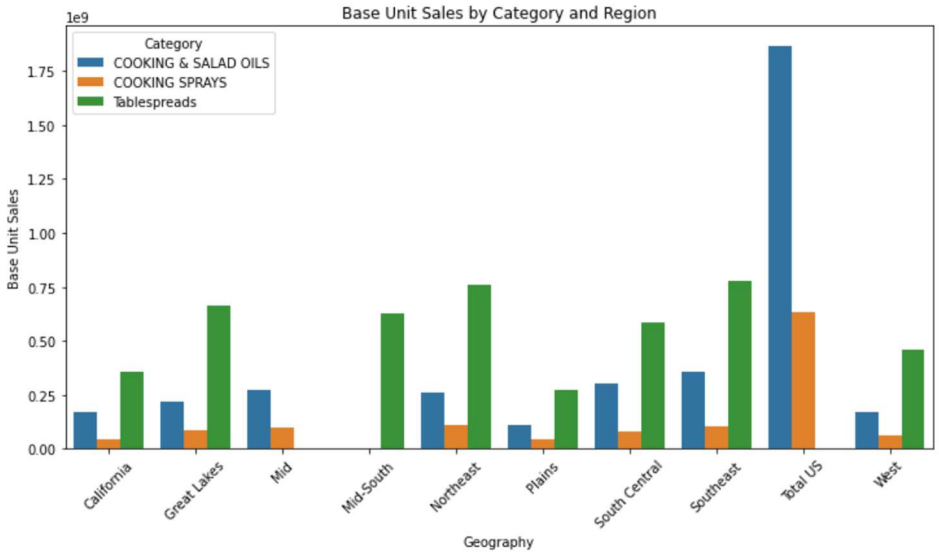
There is a change in number of sales of Tablespreads when there is an increase or decrease in the price of either Cooking Spray or Cooking & Salad Oils.

**Insights/observations**

**Exhibit 11** contains a boxplot of Total Sales across the Cooking & Salad Oils, Cooking Sprays and Tablespreads categories. Tablespreads have the highest number of sales followed by Cooking & Salad Oils. Percentage of total sales from Cooking Oil and Cooking Spray: 52.63% which is a major chunk of total sales that provides an opportunity for Conagra to take away market share from these two categories through competitive pricing.




The sales of all 3 categories follows similar trend over the years




Tablespreads has the highest sales in the Northeast region and Southeast region. The highest sales for each category occur in different regions, indicating that the performance of each category may be influenced by regional preferences or factors.

**Initial Regression Equation**

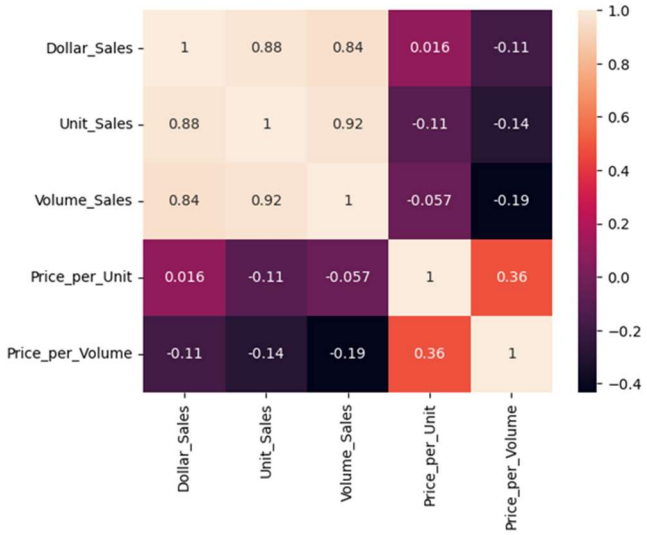
Base Unit Sales (Tablespreads) = Intercept + a*(Price per Unit_Cooking Spray) + b*(Price per Unit_Cooking & Salad Oils)

What story we are trying to tell with this regression
Whether there is category switching happening between the three different categories based on the change in price across the weeks. This regression will provide insights about whether these categories are substitute for each other when the price of one becomes undesirable for the consumer.

# Appendix

## Geographical Impacts on Unit Sales

### Exhibit 1




### Exhibit 2

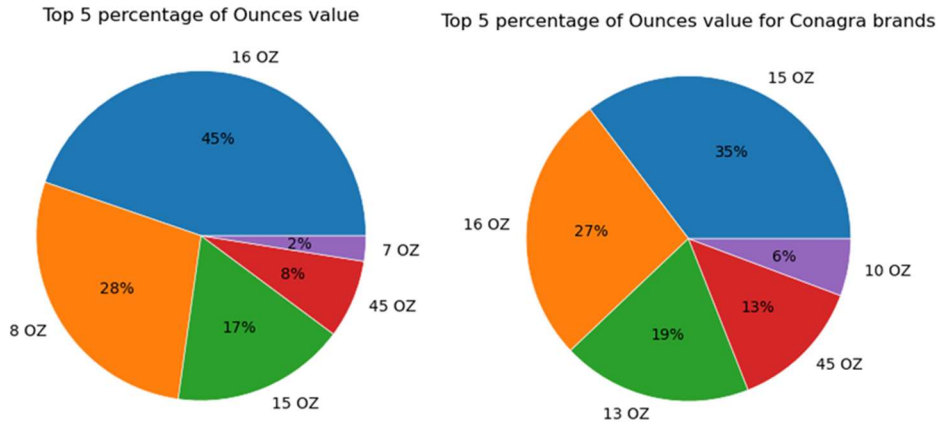
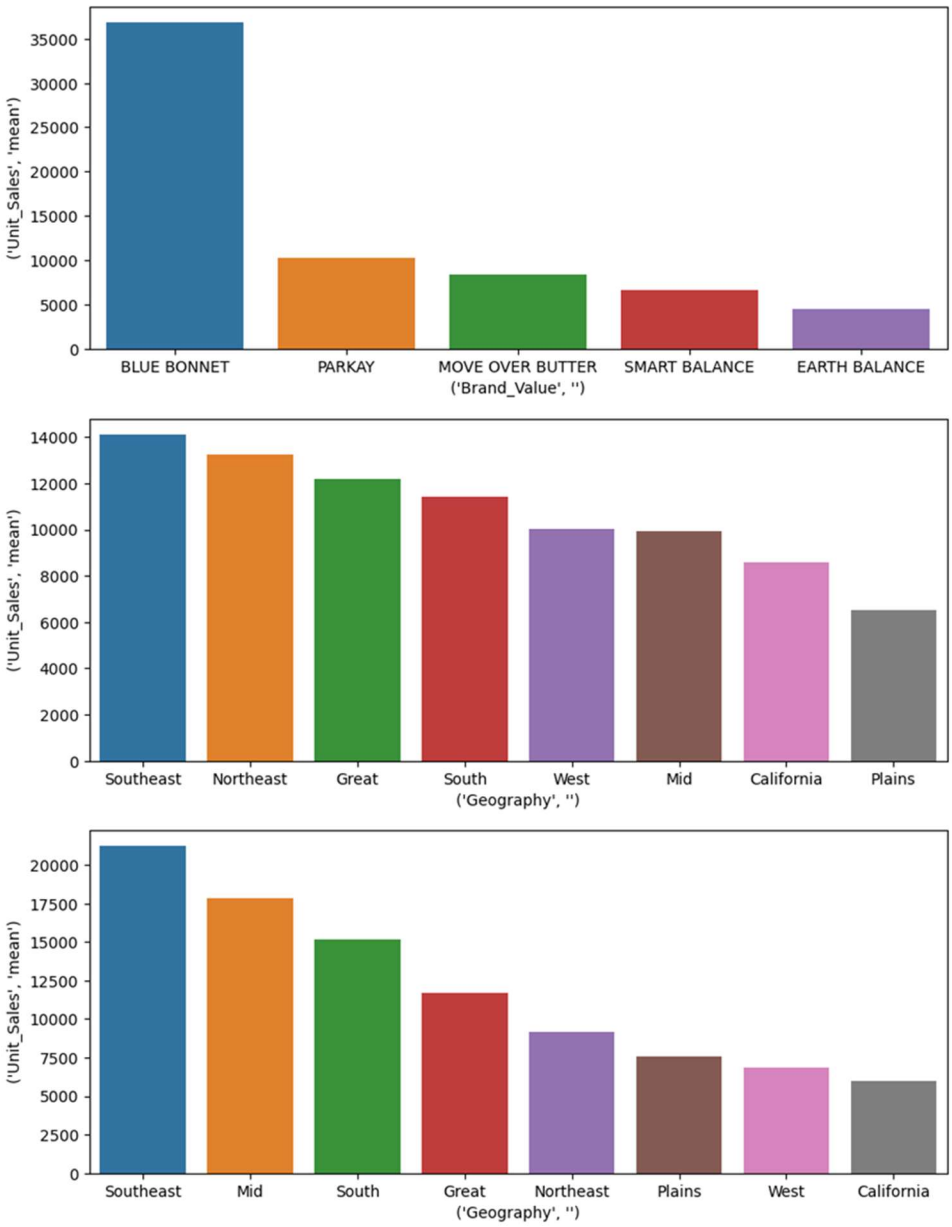



Exhibit 3




Product Attribute Impacts on Total Unit Sales

**Exhibit 9**

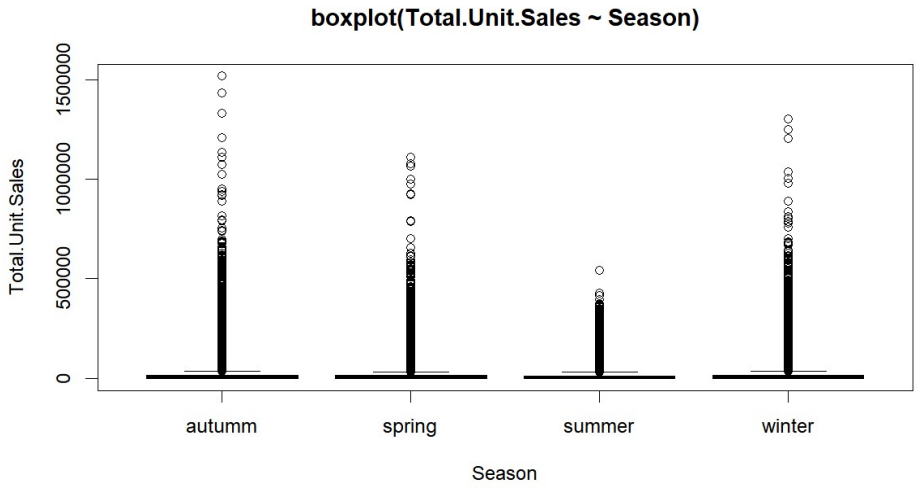
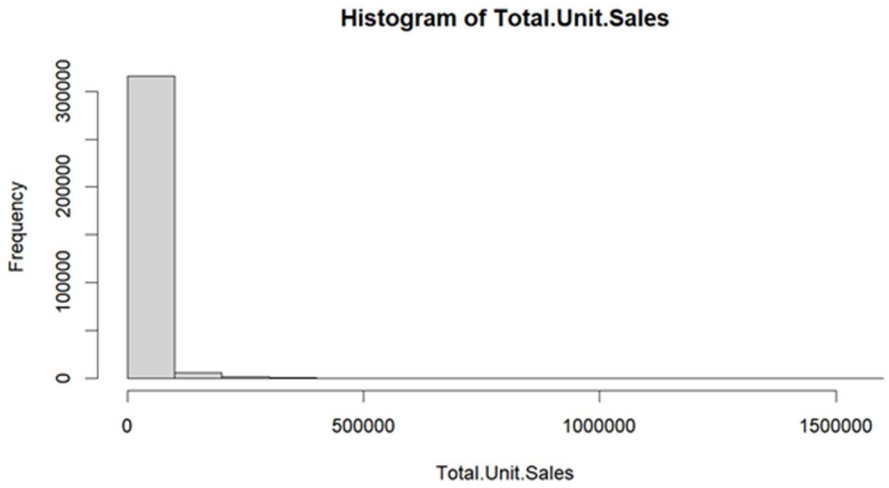
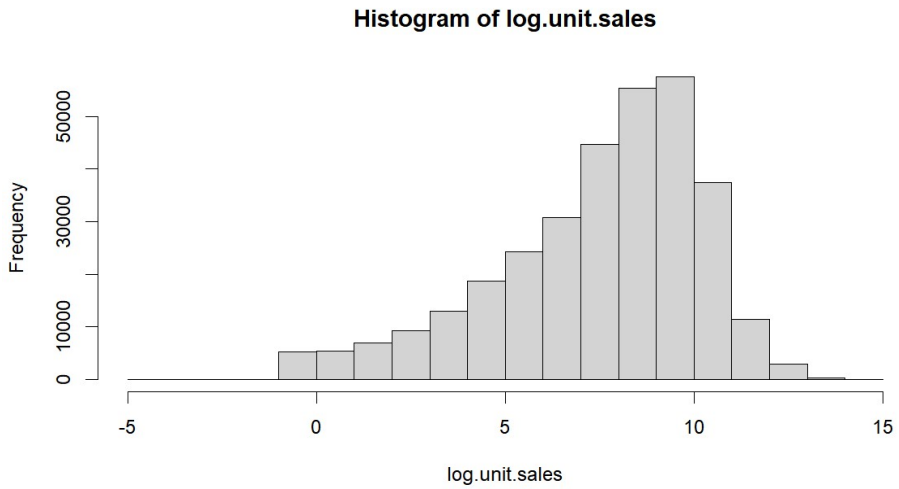

boxplot(Total.Unit.Sales ~ Season)
Total.Unit.Sales
0
500000
1000000
1500000
autumm
spring
summer
winter
Season

**Exhibit 10**


Histogram of Total.Unit.Sales
Frequency
0
100000
200000
300000
0
500000
1000000
1500000
Total.Unit.Sales


Histogram of log.unit.sales
Frequency
0
10000
30000
50000
-5
0
5
10
15
log.unit.sales

Category Expansion

**Exhibit 11**