BUAN6337.005 – PREDICTIVE ANALYTICS FOR DATA SCIENCE

Group21 Project Report

Create data-driven strategies to help Conagra unlock future growth potential in the Tablespreads category

Group Members:

Anirudh Madhavan.	AXM220108
Shreyas Nellore	SXN210092
Priyanka Pandit	PXP210037
Hemanth Kumar Batchu	HKB220000
Akshaya Ananthakrishnan	AXA220019
Abhishek Jagannath	AXJ220033

Table of Contents

1. INTRODUCTION

"Apparently, everything is better with butter." This age-old saying gained newfound relevance during the COVID-19 pandemic when the Tiktok #butterboard trend went viral. The trend led to a surge in sales of butter and other table spreads by 49.5% as people took up baking as a hobby and shared their creations on social media platforms. This trend was not limited to Conagra but also benefited other prominent brands like Land O'Lakes'. However, as the pandemic recedes and people return to their usual routines, it remains to be seen if the trend will sustain or taper off. To mitigate the potential decline in sales, it is essential to identify data-driven strategies to unlock future growth potential in the Tablespreads category. In this report, we will analyze market research data, visualize patterns in sales, and recommend a roadmap for next steps to help Conagra stay ahead of the curve and maintain its competitive edge.

2. Data driven Strategies.

While we are trying to answer few of the questions like whether there is interactions between various attributes of table spread category, whether there is any price gap in Conagra that impacts the sales, whether there are interactions across Table spreads, Cooking Oil and Cooking Spray which pose a risk or opportunity to Conagra, we are mainly focussing on how merchandising strategies impact Conagra in this report

2.1Should Conagra have varying merchandising strategies by brand or market? Are there any segments that respond better to merchandising activity?

We are attempting to address this question through a three-fold approach:

i. **Based on geography**
ii. **Based on product within Conagra**
iii. **Based on brand**

2.1.1 Based on geography:

Our objective is to identify which geographical areas exhibit a stronger response to merchandising efforts. To achieve this, we want the target variable to be a combination of geography and the incremental dollar effect resulting from merchandising. This allows us to examine the impact of merchandising on sales across the eight different locations.

Approach 1:

Initially, we proposed developing separate regression models for each location to predict the incremental sales effect based on various features. The model can be represented as follows:

Incremental_Sales_Due_to_merchandizing_Geo_A = β0 + β1 * Sub_Category_Name_Geo_A + β2 * CAG_Count_Value_Geo_A + β3 * CAG_Ounces_Value_Geo_A + β4 * CAG_Form_Value_Geo_A + β5 * CAG_Tier_Value_Geo_A + β6 * Week_Number_Consecutive_scaled_Geo_A + β7 * Dollar_Sales_No_Merch_Geo_A + β8 * Dollar_Sales_Any_Merch_Geo_A + β9 * Price_per_Unit_No_Merch_Geo_A + β10 * Price_per_Unit_Any_Merch_Geo_A + β11 * percentage_sales_merch_Geo_A + β12 * discount_percentage_Geo_A + ε

By employing this approach, we would need to create and compare nine distinct regression models to identify any unusual findings. However, we were unable to derive any conclusive insights using this method and we felt this wasn't the best approach.

Approach 2:

In our second iteration, we considered implementing a multinomial logistic regression model, with geography as the dependent variable. We felt that this model is better suited for situations where the dependent variable has multiple categories, such as the eight IRI standard regions in our case.

The independent variables include Dollar Sales, Unit Sales, Volume Sales, Price Per Unit, Price Per Volume, Base Sales, and Incremental Sales, as well as their respective measures with and without merchandising.

Here's a brief outline of the multinomial logistic regression model:

P(Geography_i = k) = exp(β_k0 + β_k1 * Dollar_Sales_with_Merch + ... + β_k10 * Price_Per_Volume_without_Merch + β_k11 * Base_Sales + β_k12 * Incremental_Sales) / [1 + sum(exp(β_j0 + β_j1 * Dollar_Sales_with_Merch + ... + β_j10 * Price_Per_Volume_without_Merch + β_j11 * Base_Sales + β_j12 * Incremental_Sales)) for j = 1 to (K-1)]

Where:
- Geography_i: the dependent variable, representing region i
- k: index for the regions, from 1 to 8 (K-1, where K = total number of regions)
- β_k0, β_k1, ..., β_k12: the coefficients to be estimated for each region k
- P(Geography_i = k): the probability of a data point belonging to region k

By fitting this model, we can estimate the coefficients for each independent variable in each region. These coefficients will indicate the effect of the independent variables on the probability of a product belonging to a particular region.

We examined the estimated coefficients to determine the relationship between allocating more resources to product, brand or location and sales but found it difficult to interpret the coefficients in a multinomial logistic regression model. There was an increase in complexity compared to a linear regression model, as the coefficients represent the change in the log-odds of the dependent variable categories.

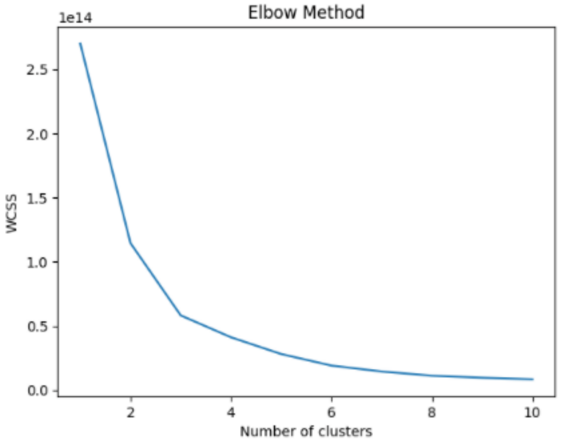
Approach 3:

In the third iteration, we placed more emphasis on interpretability.

To incorporate both geography and incremental sales as dependent variables while maintaining interpretability, we employed a two-step approach.

Step 1: Cluster analysis based on independent variables:

First, we performed k-means clustering, using the independent variables which allowed us to group data points based on similar spending and sales patterns. We then assigned each data point to a cluster.



Step 2: Multivariate regression analysis:

Subsequently, we conducted a multivariate regression analysis with two dependent variables: geography (dummy-coded for each region) and incremental sales. We used the cluster assignments from Step 1 as an independent variable, along with other relevant independent variables (e.g., product/brand dummies).

The multivariate regression model appeared as follows:
Y = X * B + E
Where:
- Y is a matrix with two columns, one for the dummy-coded geography variable and the other for incremental sales
- X is a matrix of independent variables, including the cluster assignments and any other relevant variables
- B is a matrix of coefficients to be estimated
- E is a matrix of error terms

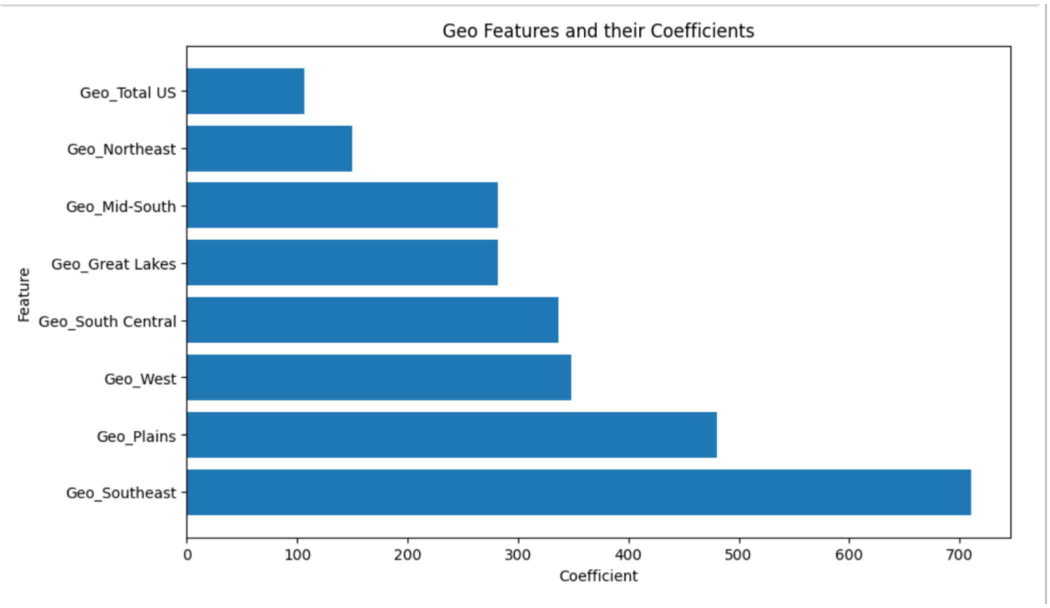
By fitting this model, we could estimate the coefficients for each independent variable, including the cluster assignments, for both geography and incremental sales.

These coefficients were interpretable in terms of their impact on geography and incremental sales, with positive coefficients indicating a positive relationship and negative coefficients indicating a negative relationship.

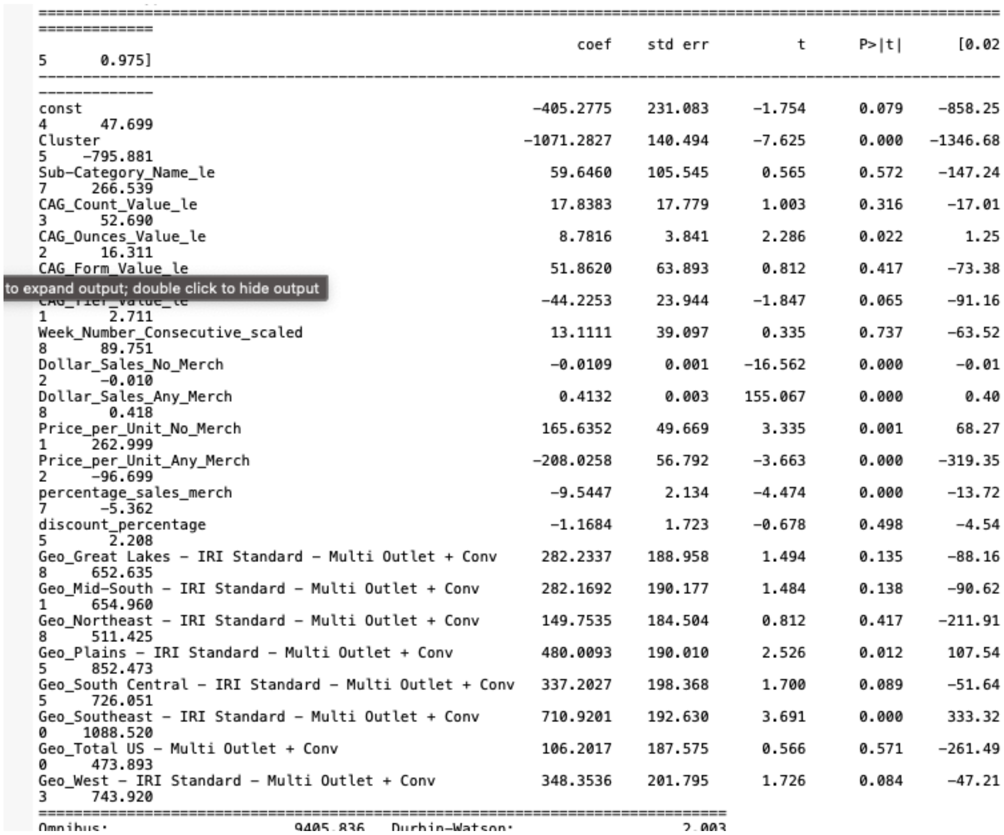
To determine whether we were spending more on a product/brand/location and receiving less increment in sales, or vice versa, we analysed the coefficients associated with the cluster assignments and identified similar spending and sales pattern. The cluster assignments represented groups of products with similar spending and sales patterns. If a cluster had a positive and significant coefficient for incremental sales, it suggested that products in that cluster tended to have higher incremental sales. Similarly, we could interpret the coefficients for geography to understand the relationships between the cluster assignments and the regional distribution of products.

This two-step approach enabled us to include both geography and incremental sales as dependent variables while maintaining the interpretability of the model results. This can further help guide us with the varying merchandising strategies that can be deployed as well as the if certain segments were more prevalent in specific regions.

Based on the results, Conagra can develop targeted merchandising strategies for different brands or markets, focusing on those segments that respond more positively to merchandising activities. This tailored approach can optimize the allocation of resources and maximize the effectiveness of merchandising efforts. We have decided to pursue this approach in our analysis.



In the results, the coefficients for the geography variables indicate the average impact of merchandising on incremental sales for each region, holding other factors constant.



```
==========================================================================================
============
                                                      coef    std err          t      P>|t|      [0.02
5      0.975]
------------------------------------------------------------------------------------------
------------
const                                            -405.2775    231.083     -1.754      0.079    -858.25
4      47.699
Cluster                                         -1071.2827    140.494     -7.625      0.000   -1346.68
5    -795.881
Sub-Category_Name_le                               59.6460    105.545      0.565      0.572    -147.24
7     266.539
CAG_Count_Value_le                                 17.8383     17.779      1.003      0.316     -17.01
3      52.690
CAG_Ounces_Value_le                                 8.7816      3.841      2.286      0.022       1.25
2      16.311
CAG_Form_Value_le                                  51.8620     63.893      0.812      0.417     -73.38
to expand output; double click to hide output
CAG_Tier_Value_le                                 -44.2253     23.944     -1.847      0.065     -91.16
1       2.711
Week_Number_Consecutive_scaled                     13.1111     39.097      0.335      0.737     -63.52
8      89.751
Dollar_Sales_No_Merch                              -0.0109      0.001    -16.562      0.000      -0.01
2      -0.010
Dollar_Sales_Any_Merch                              0.4132      0.003    155.067      0.000       0.40
8       0.418
Price_per_Unit_No_Merch                           165.6352     49.669      3.335      0.001      68.27
1     262.999
Price_per_Unit_Any_Merch                         -208.0258     56.792     -3.663      0.000    -319.35
2     -96.699
percentage_sales_merch                             -9.5447      2.134     -4.474      0.000     -13.72
7      -5.362
discount_percentage                                -1.1684      1.723     -0.678      0.498      -4.54
5       2.208
Geo_Great Lakes - IRI Standard - Multi Outlet + Conv    282.2337    188.958      1.494      0.135     -88.16
8     652.635
Geo_Mid-South - IRI Standard - Multi Outlet + Conv      282.1692    190.177      1.484      0.138     -90.62
1     654.960
Geo_Northeast - IRI Standard - Multi Outlet + Conv      149.7535    184.504      0.812      0.417    -211.91
8     511.425
Geo_Plains - IRI Standard - Multi Outlet + Conv         480.0093    190.010      2.526      0.012     107.54
5     852.473
Geo_South Central - IRI Standard - Multi Outlet + Conv  337.2027    198.368      1.700      0.089     -51.64
5     726.051
Geo_Southeast - IRI Standard - Multi Outlet + Conv      710.9201    192.630      3.691      0.000     333.32
0    1088.520
Geo_Total US - Multi Outlet + Conv                      106.2017    187.575      0.566      0.571    -261.49
0     473.893
Geo_West - IRI Standard - Multi Outlet + Conv           348.3536    201.795      1.726      0.084     -47.21
3     743.920
==============================================================================
Omnibus:                     9405.836   Durbin-Watson:                   2.003
```

Here are the coefficients for each of the 8 geographic locations:

1. Geo_Great Lakes - IRI Standard - Multi Outlet + Conv: 282.2337
2. Geo_Mid-South - IRI Standard - Multi Outlet + Conv: 282.1692
3. Geo_Northeast - IRI Standard - Multi Outlet + Conv: 149.7535
4. Geo_Plains - IRI Standard - Multi Outlet + Conv: 480.0093
5. Geo_South Central - IRI Standard - Multi Outlet + Conv: 337.2027
6. Geo_Southeast - IRI Standard - Multi Outlet + Conv: 710.9201
7. Geo_Total US - Multi Outlet + Conv: 106.2017
8. Geo_West - IRI Standard - Multi Outlet + Conv: 348.3536

These coefficients indicate the average change in incremental sales associated with a one-unit increase in the respective geography dummy variable, holding all other variables constant. For example, the coefficient for the Great Lakes region (282.2337) suggests that, on average, incremental sales in this region are expected to increase by 282.2337 units when merchandising is present, compared to when there is no merchandising, all else being equal.

2.1.2 To identify which product within Conagra responds more to merchandising, we propose the following approach:

Approach: Hierarchical Linear Modeling (HLM)

The HLM approach allows us to analyze nested data, in this case, products within Conagra. The model takes into account the variation at different levels, such as products and overall company performance. It can handle both fixed and random effects and accommodate unequal group sizes, making it suitable for this scenario.

Level 1 Model: Product-level Model

At the product level, we model the relationship between the independent variables and incremental sales.

Incremental_DollarSales_Merch_Product = β0 + β1 * Sub_Category_Name_Product + β2 * CAG_Count_Value_Product + β3 * CAG_Ounces_Value_Product + β4 * CAG_Form_Value_Product + β5 * CAG_Tier_Value_Product + β6 * Week_Number_Consecutive_scaled_Product + β7 * Dollar_Sales_No_Merch_Product + β8 * Dollar_Sales_Any_Merch_Product + β9 * Price_per_Unit_No_Merch_Product + β10 * Price_per_Unit_Any_Merch_Product + β11 * percentage_sales_merch_Product + β12 * discount_percentage_Product + e_Product

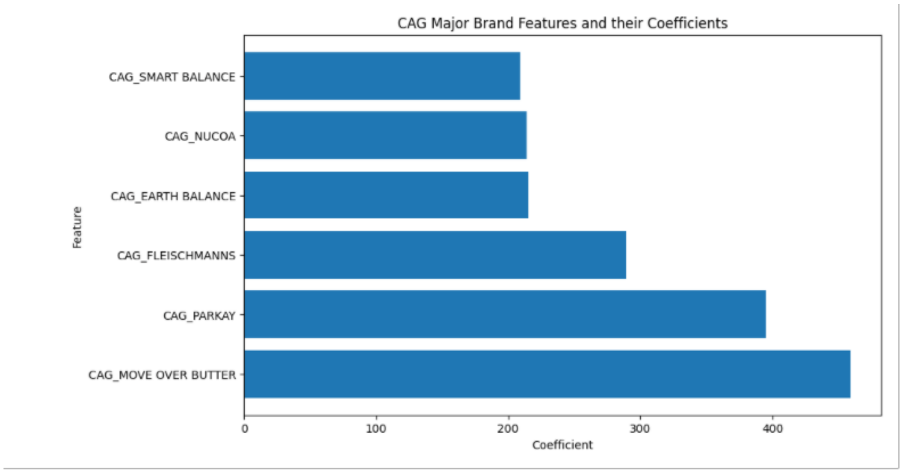
Level 2 Model: Company-level Model

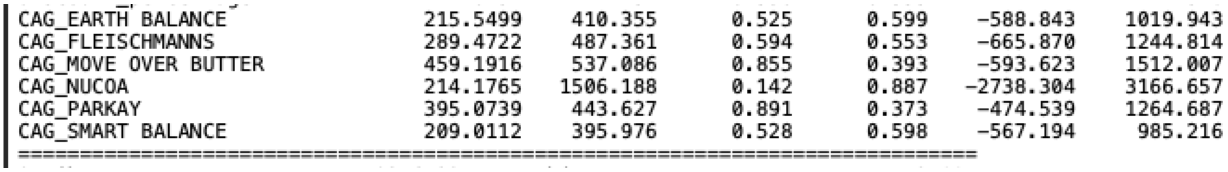
At the company level, we model the relationship between the product-level coefficients and a constant term.

β_k = γ_k0 + u_k
Where:
- β_k: the product-level coefficient from Level 1 model
- γ_k0: the company-level coefficient to be estimated
- u_k: the error term for each company-level coefficient

By fitting this HLM, we can analyse the variation in the relationship between the independent variables and incremental sales across different products within Conagra. The model will provide us with product-specific coefficients, which can be used to identify which products respond more positively to merchandising activities. Higher positive coefficients for the merchandising-related variables (e.g., Dollar_Sales_with_Merch) would indicate a stronger response to merchandising.

This approach allows us to account for both product-level and company-level variation while focusing on the relationship between the independent variables and the incremental sales for each product.





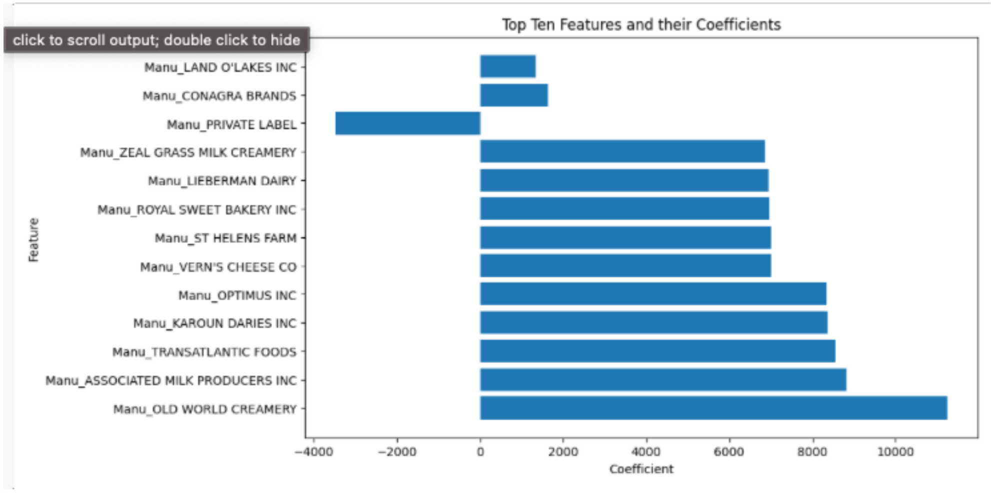
```
CAG_EARTH BALANCE               215.5499    410.355      0.525      0.599    -588.843    1019.943
CAG_FLEISCHMANNS                289.4722    487.361      0.594      0.553    -665.870    1244.814
CAG_MOVE OVER BUTTER            459.1916    537.086      0.855      0.393    -593.623    1512.007
CAG_NUCOA                       214.1765   1506.188      0.142      0.887   -2738.304    3166.657
CAG_PARKAY                      395.0739    443.627      0.891      0.373    -474.539    1264.687
CAG_SMART BALANCE               209.0112    395.976      0.528      0.598    -567.194     985.216
==============================================================================
```

CAG major brands, Earth Balance, Fleischmann's, Move Over Butter, Nucoa, and Parkay, have positive coefficients, indicating that they have a positive relationship with Incremental Dollars, meaning that these brands tend to drive higher sales.

However, the coefficients for the top three brands, Earth Balance, Fleischmann's, and Move Over Butter, are relatively higher than the other two, which suggests that they are the most effective brands in driving sales. Therefore, Conagra may want to focus on increasing merchandising efforts for these three brands in order to maximize sales.

On the other hand, the coefficients for Smart Balance and Private Label are positive, but relatively low, which indicates that they have a weaker impact on driving sales.

2.1.3 To identify which brand (Private label, Land O Lakes, Conagra) responds more to merchandising:



Here we can see the top 10 brands and their response to merchandizing. Private Label seems to have negative impact to merchandizing and is an outlier. Further analysis needs to be done on this.

To see if we are spending more on a product/brand and getting less increment in sales or vice versa, we can analyze the coefficients associated with the merchandising-related. Higher positive coefficients indicate a stronger positive relationship between merchandising spending and incremental sales, while lower or negative coefficients indicate a weaker or negative relationship. By comparing these coefficients across brands, we can identify the brands that respond better to merchandising activities and adjust spending accordingly.

Based on the three analyses conducted to understand the response to merchandising across different geographical areas, products within Conagra, and brands, we can combine the insights to develop data-driven strategies to unlock future growth potential in the Tablespreads category for Conagra.

1. Optimize merchandising strategies by region:

Using the results from the multivariate regression analysis in the first approach, identify the regions that exhibit a stronger response to merchandising efforts. Focus on increasing merchandising activities in those regions to maximize the return on investment.

2. Target high-performing products within Conagra:

Using the Hierarchical Linear Modelling approach, identify the products within Conagra that show a strong response to merchandising efforts. Allocate more resources and tailor merchandising strategies towards these high-performing products, ensuring they receive adequate visibility and promotion in both in-store and online channels.

3. Customize merchandising strategies by brand:

With the insights gained from the Multilevel Linear Regression Model, identify the brands (Private label, Land O Lakes, Conagra) that respond better to merchandising efforts. Develop targeted and customized merchandising strategies for each brand based on their responsiveness to merchandising. Focus on the unique strengths and characteristics of each brand to create compelling and effective merchandising strategies.

4. Segment-based targeting:

Based on the cluster analysis and multivariate regression results, identify segments (clusters) that respond better to merchandising activity. Develop targeted strategies for these segments, catering to their specific preferences and spending patterns.

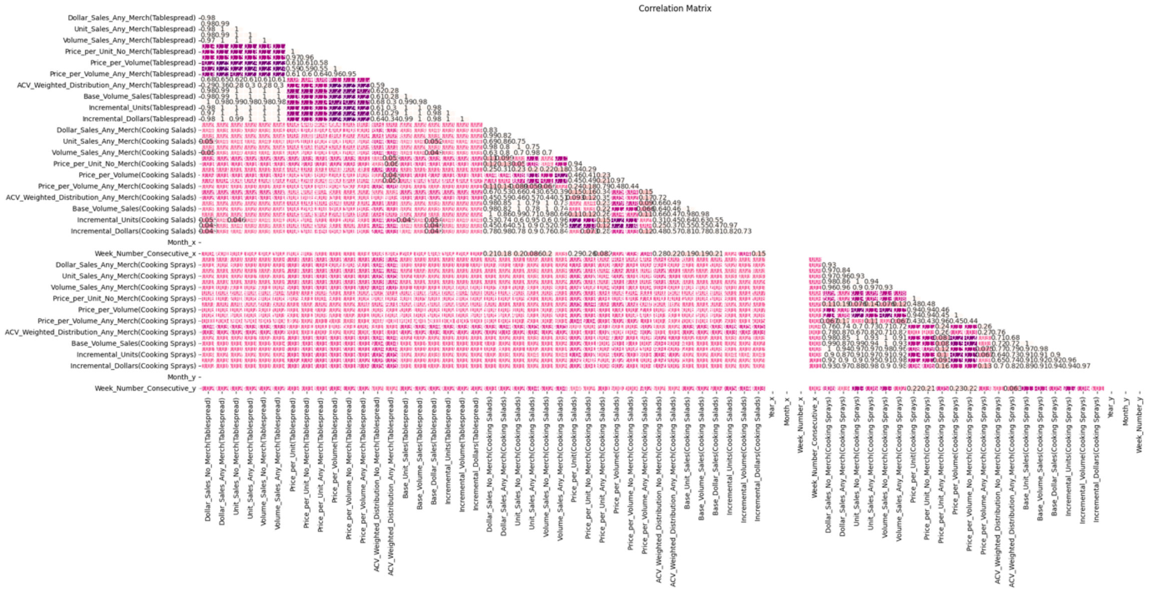
By combining the insights from these three analyses, Conagra can develop a comprehensive and data-driven merchandising strategy that targets high-performing regions, products, and brands. This tailored approach can optimize resource allocation and maximize the effectiveness of merchandising efforts, ultimately unlocking future growth potential in the Table spreads category.

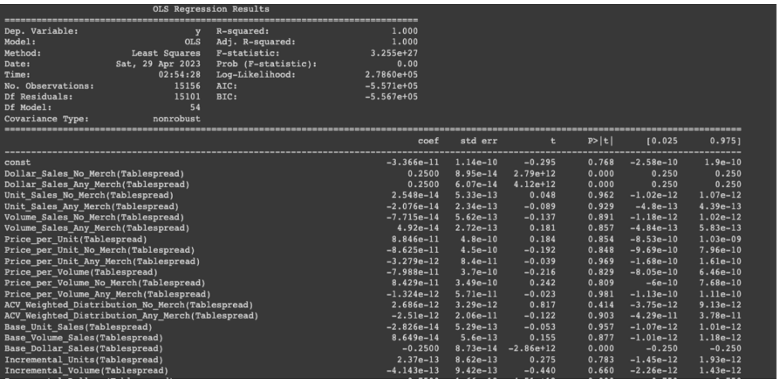
2.2 To find if there are interactions between the various attributes of table spread category and how it can impact our target variable

We are trying to observe how each attributes of Cooking Sprays and Cooking Salads are interacting with our target variable ‘Incremental Dollars’ to develop models that would help us predict the impacts. The following are the steps followed to arrive at the results:

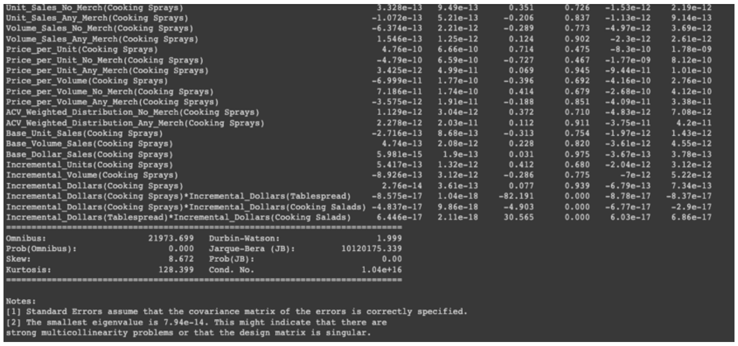
1. In order to identify the interactions across these categories we first merge all the three datasets into a single data frame.
2. We use this data frame to create a Covariance matrix and an OLS model to identify the interactions.

a. Initially we started off with using all the variables in the given data and the following are the glimpses Covariance matrix and OLS model regression table





```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
Method:                 Least Squares   F-statistic:                 3.255e+27
Date:                Sat, 29 Apr 2023   Prob (F-statistic):               0.00
Time:                        02:54:28   Log-Likelihood:             2.7860e+05
No. Observations:               15156   AIC:                        -5.571e+05
Df Residuals:                   15101   BIC:                        -5.567e+05
Df Model:                          54
Covariance Type:            nonrobust
========================================================================================================================
                                                           coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------------------------------
const                                                -3.366e-11   1.14e-10     -0.295      0.768   -2.58e-10     1.9e-10
Dollar_Sales_No_Merch(Tablespread)                       0.2500   8.95e-14   2.79e+12      0.000       0.250       0.250
Dollar_Sales_Any_Merch(Tablespread)                      0.2500   6.07e-14   4.12e+12      0.000       0.250       0.250
Unit_Sales_No_Merch(Tablespread)                      2.548e-14   5.33e-13      0.048      0.962   -1.02e-12    1.07e-12
Unit_Sales_Any_Merch(Tablespread)                    -2.076e-14   2.34e-13     -0.089      0.929    -4.8e-13    4.39e-13
Volume_Sales_No_Merch(Tablespread)                   -7.715e-14   5.62e-13     -0.137      0.891   -1.18e-12    1.02e-12
Volume_Sales_Any_Merch(Tablespread)                    4.92e-14   2.72e-13      0.181      0.857   -4.84e-13    5.83e-13
Price_per_Unit(Tablespread)                           8.846e-11    4.8e-10      0.184      0.854   -8.53e-10    1.03e-09
Price_per_Unit_No_Merch(Tablespread)                 -8.625e-11    4.5e-10     -0.192      0.848   -9.69e-10    7.96e-10
Price_per_Unit_Any_Merch(Tablespread)                -3.279e-12    8.4e-11     -0.039      0.969   -1.68e-10    1.61e-10
Price_per_Volume(Tablespread)                        -7.988e-11    3.7e-10     -0.216      0.829   -8.05e-10    6.46e-10
Price_per_Volume_No_Merch(Tablespread)                8.429e-11   3.49e-10      0.242      0.809       -6e-10    7.68e-10
Price_per_Volume_Any_Merch(Tablespread)              -1.324e-12   5.71e-11     -0.023      0.981   -1.13e-10    1.11e-10
ACV_Weighted_Distribution_No_Merch(Tablespread)       2.686e-12   3.29e-12      0.817      0.414   -3.75e-12    9.13e-12
ACV_Weighted_Distribution_Any_Merch(Tablespread)      -2.51e-12   2.06e-11     -0.122      0.903   -4.29e-11    3.78e-11
Base_Unit_Sales(Tablespread)                         -2.826e-14   5.29e-13     -0.053      0.957   -1.07e-12    1.01e-12
Base_Volume_Sales(Tablespread)                        8.649e-14    5.6e-13      0.155      0.877   -1.01e-12    1.18e-12
Base_Dollar_Sales(Tablespread)                          -0.2500   8.73e-14  -2.86e+12      0.000      -0.250      -0.250
Incremental_Units(Tablespread)                          2.37e-13   8.62e-13      0.275      0.783   -1.45e-12    1.93e-12
Incremental_Volume(Tablespread)                      -4.143e-13   9.42e-13     -0.440      0.660   -2.26e-12    1.43e-12
```



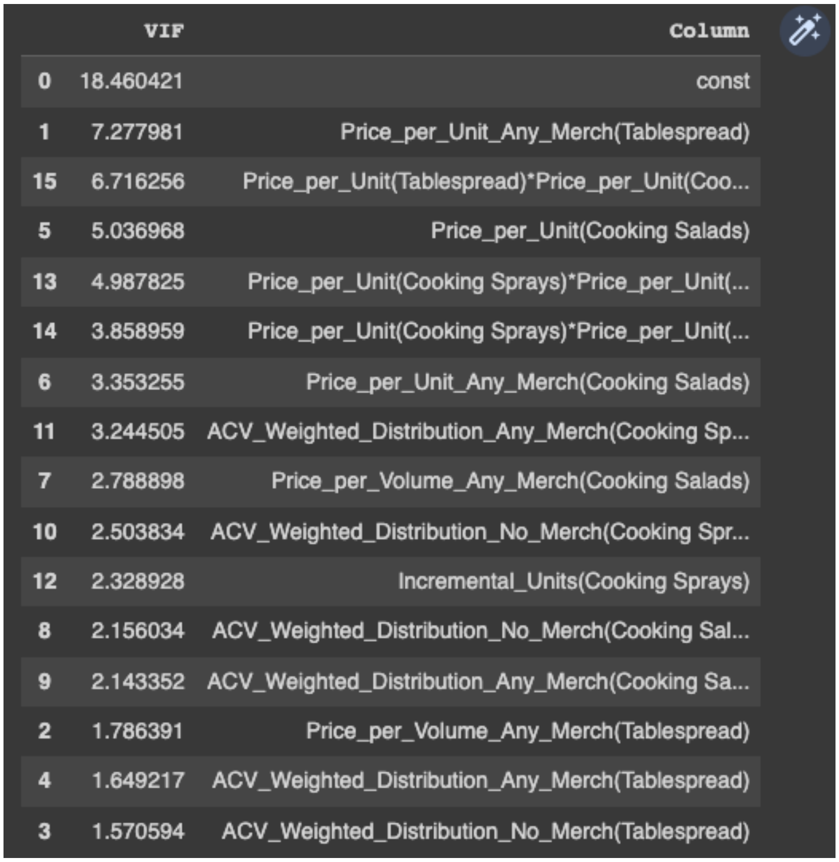
```
Unit_Sales_No_Merch(Cooking Sprays)                                       3.328e-13   9.49e-13      0.351      0.726   -1.53e-12    2.19e-12
Unit_Sales_Any_Merch(Cooking Sprays)                                     -1.072e-13   5.21e-13     -0.206      0.837   -1.13e-12    9.14e-13
Volume_Sales_No_Merch(Cooking Sprays)                                    -6.374e-13   2.21e-12     -0.289      0.773   -4.97e-12    3.69e-12
Volume_Sales_Any_Merch(Cooking Sprays)                                    1.546e-13   1.25e-12      0.124      0.902    -2.3e-12    2.61e-12
Price_per_Unit(Cooking Sprays)                                             4.76e-10   6.66e-10      0.714      0.475    -8.3e-10    1.78e-09
Price_per_Unit_No_Merch(Cooking Sprays)                                   -4.79e-10   6.59e-10     -0.727      0.467   -1.77e-09    8.12e-10
Price_per_Unit_Any_Merch(Cooking Sprays)                                  3.425e-12   4.99e-11      0.069      0.945   -9.44e-11    1.01e-10
Price_per_Volume(Cooking Sprays)                                         -6.999e-11   1.77e-10     -0.396      0.692   -4.16e-10    2.76e-10
Price_per_Volume_No_Merch(Cooking Sprays)                                 7.186e-11   1.74e-10      0.414      0.679   -2.68e-10    4.12e-10
Price_per_Volume_Any_Merch(Cooking Sprays)                               -3.575e-12   1.91e-11     -0.188      0.851   -4.09e-11    3.38e-11
ACV_Weighted_Distribution_No_Merch(Cooking Sprays)                        1.129e-12   3.04e-12      0.372      0.710   -4.83e-12    7.08e-12
ACV_Weighted_Distribution_Any_Merch(Cooking Sprays)                       2.278e-12   2.03e-11      0.112      0.911   -3.75e-11     4.2e-11
Base_Unit_Sales(Cooking Sprays)                                          -2.716e-13   8.68e-13     -0.313      0.754   -1.97e-12    1.43e-12
Base_Volume_Sales(Cooking Sprays)                                          4.74e-13   2.08e-12      0.228      0.820   -3.61e-12    4.55e-12
Base_Dollar_Sales(Cooking Sprays)                                         5.981e-15    1.9e-13      0.031      0.975   -3.67e-13    3.78e-13
Incremental_Units(Cooking Sprays)                                         5.417e-13   1.32e-12      0.412      0.680   -2.04e-12    3.12e-12
Incremental_Volume(Cooking Sprays)                                       -8.926e-13   3.12e-12     -0.286      0.775      -7e-12    5.22e-12
Incremental_Dollars(Cooking Sprays)                                        2.76e-14   3.61e-13      0.077      0.939   -6.79e-13    7.34e-13
Incremental_Dollars(Cooking Sprays)*Incremental_Dollars(Tablespread)     -8.575e-17   1.04e-18    -82.191      0.000   -8.78e-17   -8.37e-17
Incremental_Dollars(Cooking Sprays)*Incremental_Dollars(Cooking Salads) -4.837e-17   9.86e-18     -4.903      0.000   -6.77e-17    -2.9e-17
Incremental_Dollars(Tablespread)*Incremental_Dollars(Cooking Salads)      6.446e-17   2.11e-18     30.565      0.000    6.03e-17    6.86e-17
==============================================================================
Omnibus:                    21973.699   Durbin-Watson:                   1.999
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         10120175.339
Skew:                           8.672   Prob(JB):                         0.00
Kurtosis:                     128.399   Cond. No.                     1.04e+16
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 7.94e-14. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

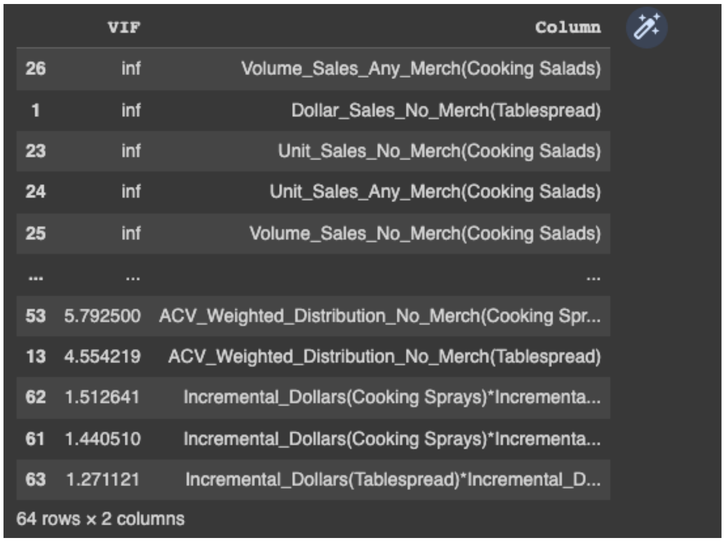
Looking at the above results it is clearly visible that there is multicollinearity in the model and therefore these results cannot be accurate.

To get rid of the multicollinearity problem we calculated the VIF (Variance Inflation Factor) for each independent variables. We decided to not include the columns with VIF value greater than 5.

- The one on the right is the initial VIF table
- And the one on the left is the new VIF table after including only required columns



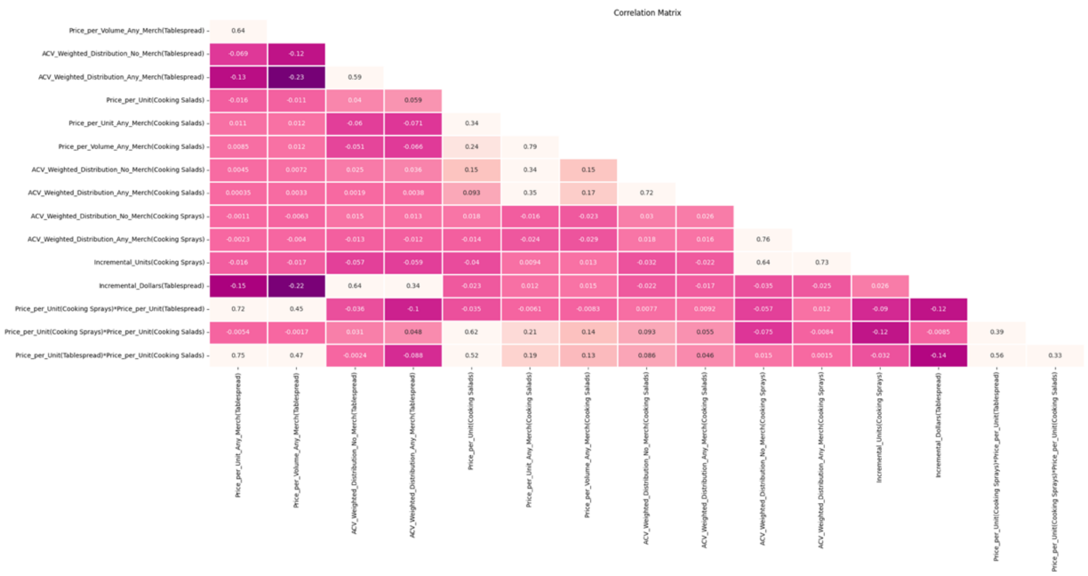
	VIF	Column
0	18.460421	const
1	7.277981	Price_per_Unit_Any_Merch(Tablespread)
15	6.716256	Price_per_Unit(Tablespread)*Price_per_Unit(Coo...
5	5.036968	Price_per_Unit(Cooking Salads)
13	4.987825	Price_per_Unit(Cooking Sprays)*Price_per_Unit(...
14	3.858959	Price_per_Unit(Cooking Sprays)*Price_per_Unit(...
6	3.353255	Price_per_Unit_Any_Merch(Cooking Salads)
11	3.244505	ACV_Weighted_Distribution_Any_Merch(Cooking Sp...
7	2.788898	Price_per_Volume_Any_Merch(Cooking Salads)
10	2.503834	ACV_Weighted_Distribution_No_Merch(Cooking Spr...
12	2.328928	Incremental_Units(Cooking Sprays)
8	2.156034	ACV_Weighted_Distribution_No_Merch(Cooking Sal...
9	2.143352	ACV_Weighted_Distribution_Any_Merch(Cooking Sa...
2	1.786391	Price_per_Volume_Any_Merch(Tablespread)
4	1.649217	ACV_Weighted_Distribution_Any_Merch(Tablespread)
3	1.570594	ACV_Weighted_Distribution_No_Merch(Tablespread)



	VIF	Column
26	inf	Volume_Sales_Any_Merch(Cooking Salads)
1	inf	Dollar_Sales_No_Merch(Tablespread)
23	inf	Unit_Sales_No_Merch(Cooking Salads)
24	inf	Unit_Sales_Any_Merch(Cooking Salads)
25	inf	Volume_Sales_No_Merch(Cooking Salads)
...	...	...
53	5.792500	ACV_Weighted_Distribution_No_Merch(Cooking Spr...
13	4.554219	ACV_Weighted_Distribution_No_Merch(Tablespread)
62	1.512641	Incremental_Dollars(Cooking Sprays)*Incrementa...
61	1.440510	Incremental_Dollars(Cooking Sprays)*Incrementa...
63	1.271121	Incremental_Dollars(Tablespread)*Incremental_D...

64 rows × 2 columns

2. After solving for multicollinearity issue we arrive at the following Covariance matrix and an OLS model.

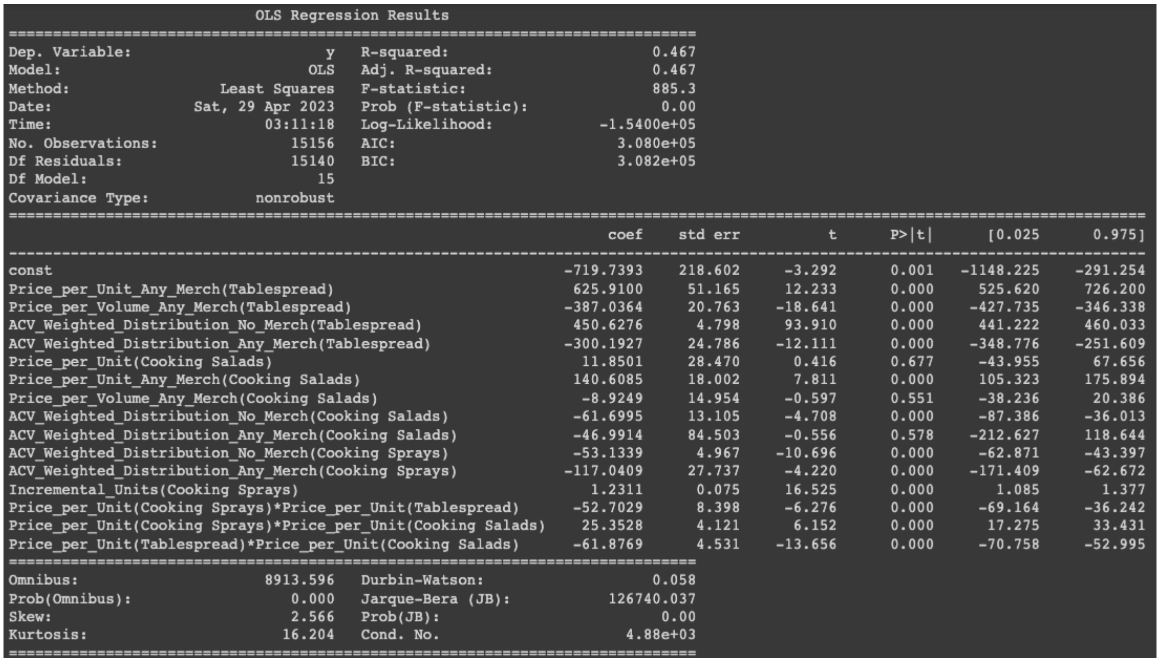


Correlation Matrix

	Price_per_Unit_Any_Merch(Tablespread)	Price_per_Volume_Any_Merch(Tablespread)	ACV_Weighted_Distribution_No_Merch(Tablespread)	ACV_Weighted_Distribution_Any_Merch(Tablespread)	Price_per_Unit(Cooking Salads)	Price_per_Unit_Any_Merch(Cooking Salads)	Price_per_Volume_Any_Merch(Cooking Salads)	ACV_Weighted_Distribution_No_Merch(Cooking Salads)	ACV_Weighted_Distribution_Any_Merch(Cooking Salads)	ACV_Weighted_Distribution_No_Merch(Cooking Sprays)	ACV_Weighted_Distribution_Any_Merch(Cooking Sprays)	Incremental_Units(Cooking Sprays)	Incremental_Dollars(Tablespread)	Price_per_Unit(Cooking Sprays)*Price_per_Unit(Tablespread)	Price_per_Unit(Cooking Sprays)*Price_per_Unit(Cooking Salads)
Price_per_Volume_Any_Merch(Tablespread)	0.64														
ACV_Weighted_Distribution_No_Merch(Tablespread)	-0.069	-0.12													
ACV_Weighted_Distribution_Any_Merch(Tablespread)	-0.13	-0.23	0.59												
Price_per_Unit(Cooking Salads)	-0.016	-0.011	0.04	0.059											
Price_per_Unit_Any_Merch(Cooking Salads)	0.011	0.012	-0.06	-0.071	0.34										
Price_per_Volume_Any_Merch(Cooking Salads)	0.0085	0.012	-0.051	-0.066	0.24	0.79									
ACV_Weighted_Distribution_No_Merch(Cooking Salads)	0.0045	0.0072	0.025	0.036	0.15	0.34	0.15								
ACV_Weighted_Distribution_Any_Merch(Cooking Salads)	0.00035	0.0033	0.0019	0.0038	0.093	0.35	0.17	0.72							
ACV_Weighted_Distribution_No_Merch(Cooking Sprays)	-0.0011	-0.0063	0.015	0.013	0.018	-0.016	-0.023	0.03	0.026						
ACV_Weighted_Distribution_Any_Merch(Cooking Sprays)	-0.0023	-0.004	-0.013	-0.012	-0.014	-0.024	-0.029	0.018	0.016	0.76					
Incremental_Units(Cooking Sprays)	-0.016	-0.017	-0.057	-0.059	-0.04	0.0094	0.013	-0.032	-0.022	0.64	0.73				
Incremental_Dollars(Tablespread)	-0.15	-0.22	0.64	0.34	-0.023	0.012	0.015	-0.022	-0.017	-0.035	-0.025	0.026			
Price_per_Unit(Cooking Sprays)*Price_per_Unit(Tablespread)	0.72	0.45	-0.036	-0.1	-0.035	-0.0061	-0.0083	0.0077	0.0092	-0.057	0.012	-0.09	-0.12		
Price_per_Unit(Cooking Sprays)*Price_per_Unit(Cooking Salads)	-0.0054	-0.0017	0.031	0.048	0.62	0.21	0.14	0.093	0.055	-0.075	-0.0084	-0.12	-0.0085	0.39	
Price_per_Unit(Tablespread)*Price_per_Unit(Cooking Salads)	0.75	0.47	-0.0024	-0.088	0.52	0.19	0.13	0.086	0.046	0.015	0.0015	-0.032	-0.14	0.56	0.33

Looking at the above results the following can be concluded:

1. The covariance value between the variables Incremental_Dollars(Tablespread) and Price_per_Unit(Cooking Salads) is -0.023 which is almost close to 0. This means that there is not much interaction effect of Price_Per_Unit(Cooking Salads) on the Increment Dollars of Tablespreads
2. From the regression table the coefficient of Incremental_Units(Cooking Sprays) is 1.2311. Therefore it can be said that there is an impact of Incremental_Units of Cooking Sprays on Incremental Dollars of Tablespreads and if Incremental_Units of Cooking Sprays increase by 1 unit, the Incremental Dollars value increases by 1.2311 units.

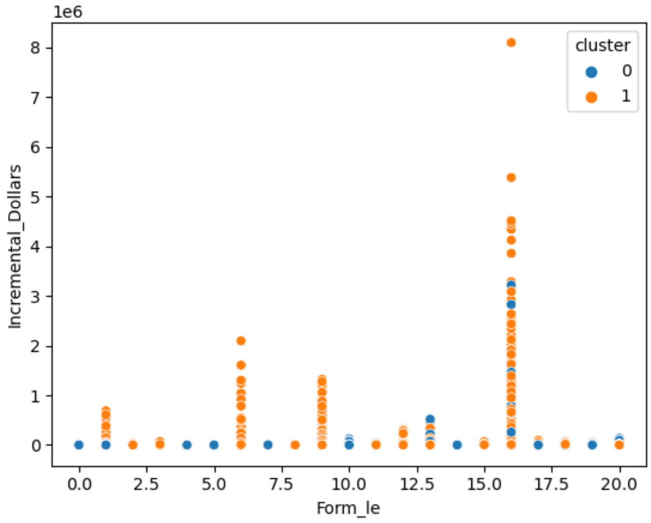


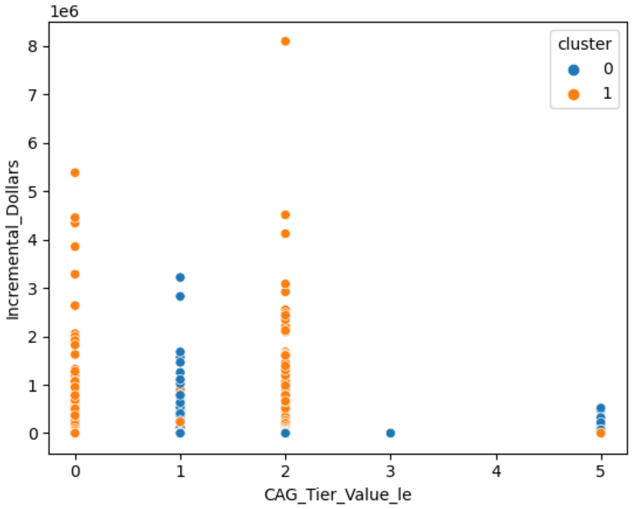
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.467
Model:                            OLS   Adj. R-squared:                  0.467
Method:                 Least Squares   F-statistic:                     885.3
Date:                Sat, 29 Apr 2023   Prob (F-statistic):               0.00
Time:                        03:11:18   Log-Likelihood:            -1.5400e+05
No. Observations:               15156   AIC:                         3.080e+05
Df Residuals:                   15140   BIC:                         3.082e+05
Df Model:                          15
Covariance Type:            nonrobust
=================================================================================================================================
                                                                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------------------------------
const                                                          -719.7393    218.602     -3.292      0.001   -1148.225    -291.254
Price_per_Unit_Any_Merch(Tablespread)                           625.9100     51.165     12.233      0.000     525.620     726.200
Price_per_Volume_Any_Merch(Tablespread)                        -387.0364     20.763    -18.641      0.000    -427.735    -346.338
ACV_Weighted_Distribution_No_Merch(Tablespread)                 450.6276      4.798     93.910      0.000     441.222     460.033
ACV_Weighted_Distribution_Any_Merch(Tablespread)               -300.1927     24.786    -12.111      0.000    -348.776    -251.609
Price_per_Unit(Cooking Salads)                                   11.8501     28.470      0.416      0.677     -43.955      67.656
Price_per_Unit_Any_Merch(Cooking Salads)                        140.6085     18.002      7.811      0.000     105.323     175.894
Price_per_Volume_Any_Merch(Cooking Salads)                       -8.9249     14.954     -0.597      0.551     -38.236      20.386
ACV_Weighted_Distribution_No_Merch(Cooking Salads)              -61.6995     13.105     -4.708      0.000     -87.386     -36.013
ACV_Weighted_Distribution_Any_Merch(Cooking Salads)             -46.9914     84.503     -0.556      0.578    -212.627     118.644
ACV_Weighted_Distribution_No_Merch(Cooking Sprays)              -53.1339      4.967    -10.696      0.000     -62.871     -43.397
ACV_Weighted_Distribution_Any_Merch(Cooking Sprays)            -117.0409     27.737     -4.220      0.000    -171.409     -62.672
Incremental_Units(Cooking Sprays)                                 1.2311      0.075     16.525      0.000       1.085       1.377
Price_per_Unit(Cooking Sprays)*Price_per_Unit(Tablespread)      -52.7029      8.398     -6.276      0.000     -69.164     -36.242
Price_per_Unit(Cooking Sprays)*Price_per_Unit(Cooking Salads)    25.3528      4.121      6.152      0.000      17.275      33.431
Price_per_Unit(Tablespread)*Price_per_Unit(Cooking Salads)      -61.8769      4.531    -13.656      0.000     -70.758     -52.995
==============================================================================
Omnibus:                     8913.596   Durbin-Watson:                   0.058
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           126740.037
Skew:                           2.566   Prob(JB):                         0.00
Kurtosis:                      16.204   Cond. No.                     4.88e+03
==============================================================================
```

2.3 To find if there are interactions between the various attributes of tablespread category and how it can impact our target variable

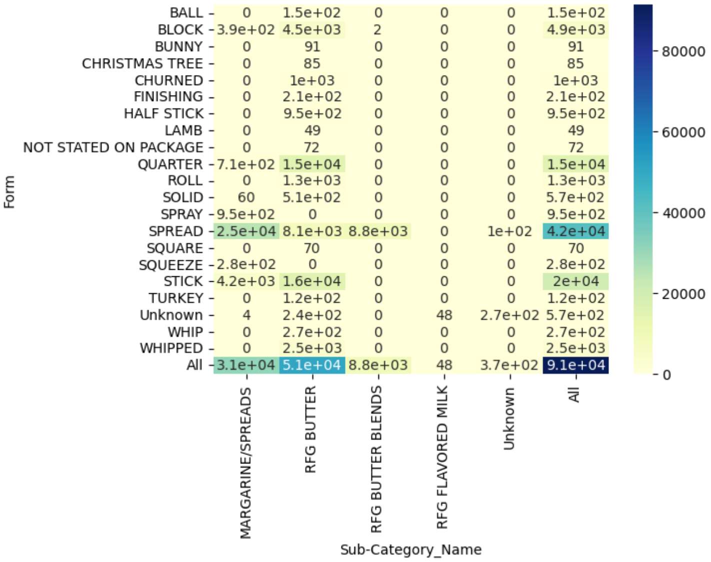
We try to observe how each tablespread attributes are interacting within each other and with target variable ‘Incremental Dollars’ to develop models that would help us predict the impacts.

Firstly, we perform k-means clustering using 2 nodes (which was found to be optimal with the elbow method) of the attributes along with ‘Incremental Dollars’. We were able to observe that attributes like Form and Tier value are clustered more distinctively as compared to Sub_Category or Ounces Value.



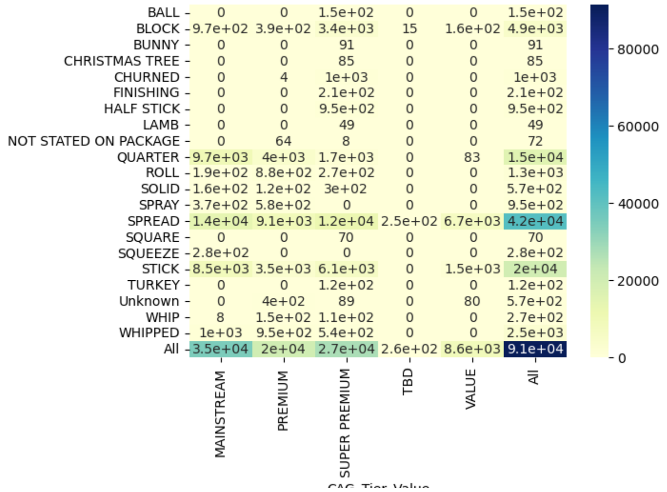


2. We also use cross tabulation tables to understand how much of items do two attributes have in common.

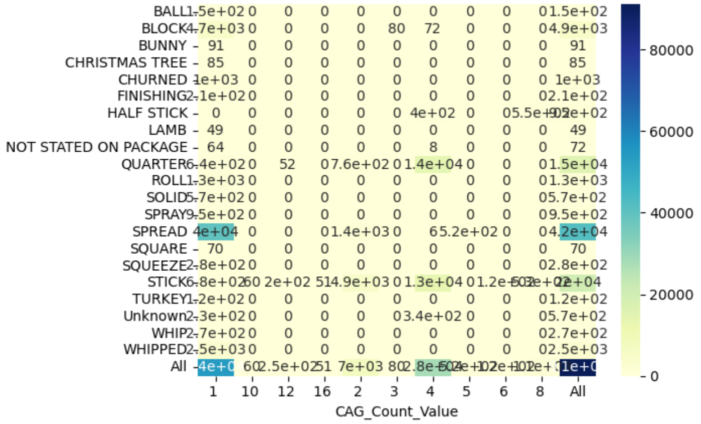


Form	MARGARINE/SPREADS	RFG BUTTER	RFG BUTTER BLENDS	RFG FLAVORED MILK	Unknown	All
BALL	0	1.5e+02	0	0	0	1.5e+02
BLOCK	3.9e+02	4.5e+03	2	0	0	4.9e+03
BUNNY	0	91	0	0	0	91
CHRISTMAS TREE	0	85	0	0	0	85
CHURNED	0	1e+03	0	0	0	1e+03
FINISHING	0	2.1e+02	0	0	0	2.1e+02
HALF STICK	0	9.5e+02	0	0	0	9.5e+02
LAMB	0	49	0	0	0	49
NOT STATED ON PACKAGE	0	72	0	0	0	72
QUARTER	7.1e+02	1.5e+04	0	0	0	1.5e+04
ROLL	0	1.3e+03	0	0	0	1.3e+03
SOLID	60	5.1e+02	0	0	0	5.7e+02
SPRAY	9.5e+02	0	0	0	0	9.5e+02
SPREAD	2.5e+04	8.1e+03	8.8e+03	0	1e+02	4.2e+04
SQUARE	0	70	0	0	0	70
SQUEEZE	2.8e+02	0	0	0	0	2.8e+02
STICK	4.2e+03	1.6e+04	0	0	0	2e+04
TURKEY	0	1.2e+02	0	0	0	1.2e+02
Unknown	4	2.4e+02	0	48	2.7e+02	5.7e+02
WHIP	0	2.7e+02	0	0	0	2.7e+02
WHIPPED	0	2.5e+03	0	0	0	2.5e+03
All	3.1e+04	5.1e+04	8.8e+03	48	3.7e+02	9.1e+04

We see a similar trend as our previous image with Form and sub-category in the cross tabulation as well. We can see that a majority of the items are from RFG butter and that is why it was all concentrated in one category while clustering as well. This means the focus must be on the sales of butter while we can remove the contribution from milk or add it with unknown for ease while working on models reducing



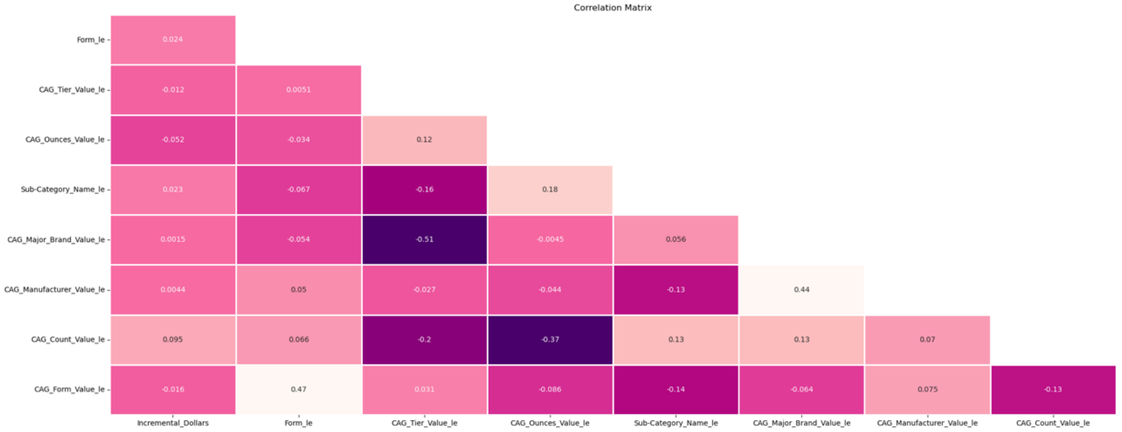
In form and tier value we can just focus on the 3 main groups Mainstream, Premium and Super Premium with extra focus on Super Premium as it constitutes to a major percentage in the total.



This heatmap gives an interesting finding in-terms of form with count value. Over here we can notice that while all forms have a majority in 1 apart from that each item has its own specific count value for packaging. This helps us further ideate to see how some interaction between the form and a specific count of each form can add value that can impact the target.

Also, what can happen if Conagra changes the packaging of a particular form to increase or decrease the count value? Can this impact the Incremental dollars is a question to further investigate.

3. We further proceed to create a correlation matrix to derive the interrelation of each of our independent variables and how are they correlated with the target variable.



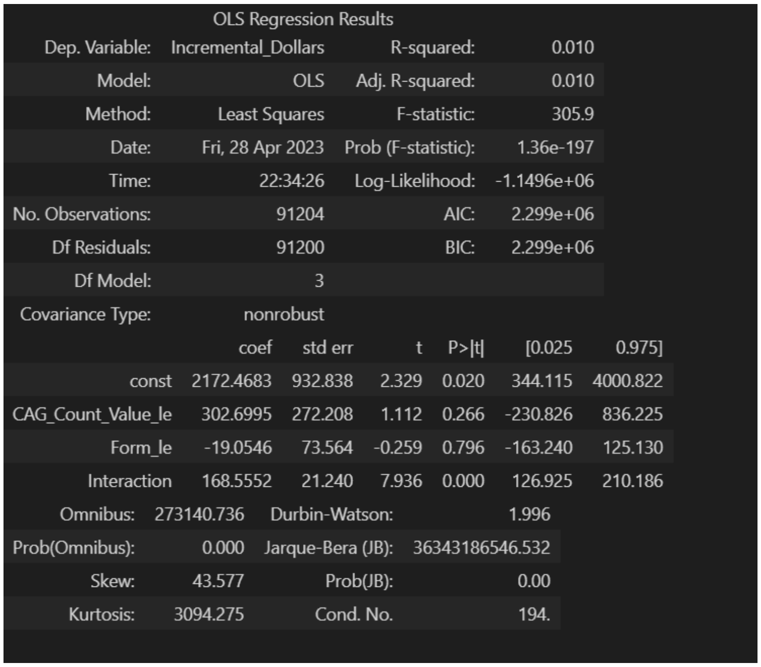
From the correlation matrix we can see that of the attribute variables Form_le, SAG_Count value has some amount of positive correlation with the incremental dollars, CAG_Ounces_Value has a high negative correlation.

It is important we try to develop models with all these variables individually to understand their significant effects before checking the interactions.

4. With all the initial analysis we then further proceed to draft some basic ols models to understand the significance of each of these attributes and their interactions with Incremental Dollars.

Initially drawing ols models with y=b0+b1*Form_le and y=b0+b1*count_value_le does give some significant coefficients but with very low goodness of fit. So when we add the interaction effect such that y=b0+b1*Form_le+b2*count_value_le+b3*Form_le*count_value_le

We get

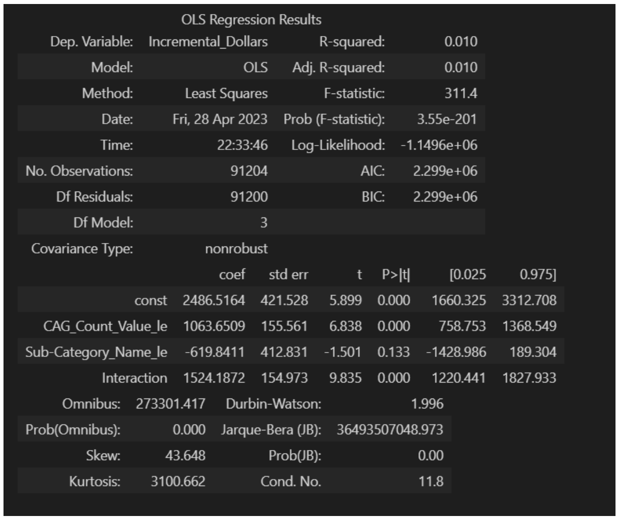


```
                       OLS Regression Results
    Dep. Variable:  Incremental_Dollars        R-squared:        0.010
            Model:                  OLS   Adj. R-squared:        0.010
           Method:        Least Squares      F-statistic:        305.9
             Date:     Fri, 28 Apr 2023 Prob (F-statistic):   1.36e-197
             Time:             22:34:26   Log-Likelihood:  -1.1496e+06
No. Observations:                 91204              AIC:     2.299e+06
     Df Residuals:                91200              BIC:     2.299e+06
         Df Model:                    3
  Covariance Type:            nonrobust
                        coef   std err       t   P>|t|    [0.025     0.975]
             const  2172.4683  932.838   2.329   0.020   344.115   4000.822
CAG_Count_Value_le   302.6995  272.208   1.112   0.266  -230.826    836.225
           Form_le   -19.0546   73.564  -0.259   0.796  -163.240    125.130
       Interaction   168.5552   21.240   7.936   0.000   126.925    210.186
       Omnibus: 273140.736   Durbin-Watson:            1.996
Prob(Omnibus):       0.000 Jarque-Bera (JB): 36343186546.532
          Skew:     43.577        Prob(JB):             0.00
      Kurtosis:   3094.275        Cond. No.             194.
```

While we can see that the individual variables are not that significant, their interaction effect seems to be highly significant with a value of 168.5 and the r^2 also comes to 0.010.

This means it is important to consider the interaction effect of Form and Count value in our final model.

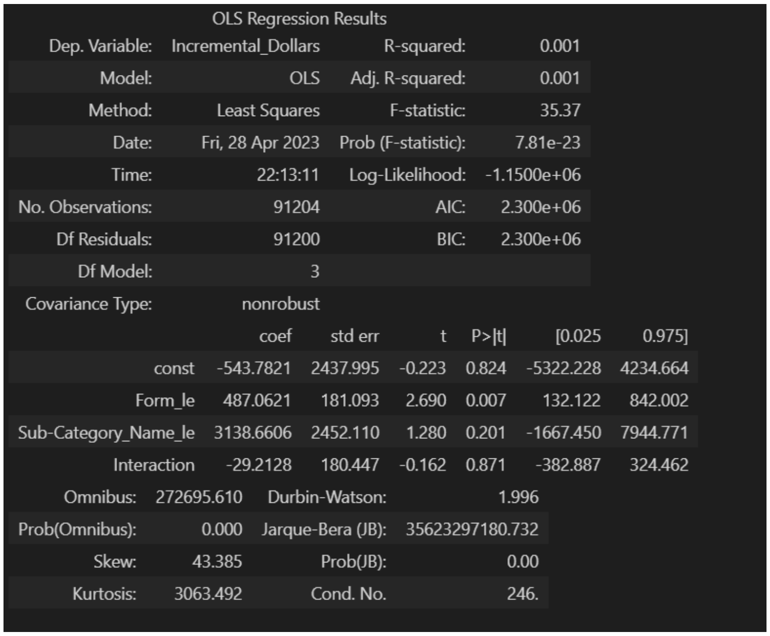
Another observation would be that of the interaction between CAG_Count Value and Sub_Category_Name



```
                    OLS Regression Results
  Dep. Variable:   Incremental_Dollars        R-squared:       0.010
          Model:                   OLS   Adj. R-squared:       0.010
         Method:         Least Squares        F-statistic:     311.4
           Date:      Fri, 28 Apr 2023  Prob (F-statistic):  3.55e-201
           Time:              22:33:46    Log-Likelihood:  -1.1496e+06
No. Observations:                91204               AIC:    2.299e+06
    Df Residuals:                91200               BIC:    2.299e+06
        Df Model:                    3
 Covariance Type:            nonrobust
                        coef    std err       t    P>|t|     [0.025     0.975]
             const   2486.5164  421.528   5.899   0.000   1660.325   3312.708
  CAG_Count_Value_le 1063.6509  155.561   6.838   0.000    758.753   1368.549
Sub-Category_Name_le -619.8411  412.831  -1.501   0.133  -1428.986    189.304
         Interaction 1524.1872  154.973   9.835   0.000   1220.441   1827.933
      Omnibus:  273301.417    Durbin-Watson:            1.996
Prob(Omnibus):       0.000  Jarque-Bera (JB):  36493507048.973
         Skew:      43.648          Prob(JB):             0.00
     Kurtosis:    3100.662          Cond. No.             11.8
```

Over this as well the interaction effect has a high impact on our target variable that we can use to improve the efficiency of our final model.

On the very similar grounds if we check for interactions between Form and Sub Category



```
                    OLS Regression Results
  Dep. Variable:   Incremental_Dollars        R-squared:       0.001
          Model:                   OLS   Adj. R-squared:       0.001
         Method:         Least Squares        F-statistic:     35.37
           Date:      Fri, 28 Apr 2023  Prob (F-statistic):  7.81e-23
           Time:              22:13:11    Log-Likelihood:  -1.1500e+06
No. Observations:                91204               AIC:    2.300e+06
    Df Residuals:                91200               BIC:    2.300e+06
        Df Model:                    3
 Covariance Type:            nonrobust
                        coef    std err       t    P>|t|     [0.025     0.975]
               const  -543.7821 2437.995  -0.223   0.824  -5322.228   4234.664
             Form_le   487.0621  181.093   2.690   0.007    132.122    842.002
Sub-Category_Name_le  3138.6606 2452.110   1.280   0.201  -1667.450   7944.771
         Interaction   -29.2128  180.447  -0.162   0.871   -382.887    324.462
      Omnibus:  272695.610    Durbin-Watson:            1.996
Prob(Omnibus):       0.000  Jarque-Bera (JB):  35623297180.732
         Skew:      43.385          Prob(JB):             0.00
     Kurtosis:    3063.492          Cond. No.             246.
```

We can observe that none of the variables are significant including the interaction so we can eliminate this combination of the interaction effect.

The above stated are only a few distinct observations of the many combinations of interactions among the attribute variables. We can follow the similar pattern to decide on which interactions would create an impact(both positive and negative) on the target variable to use it them as our independent variables in our final models to further predict the Profits for Conagra.

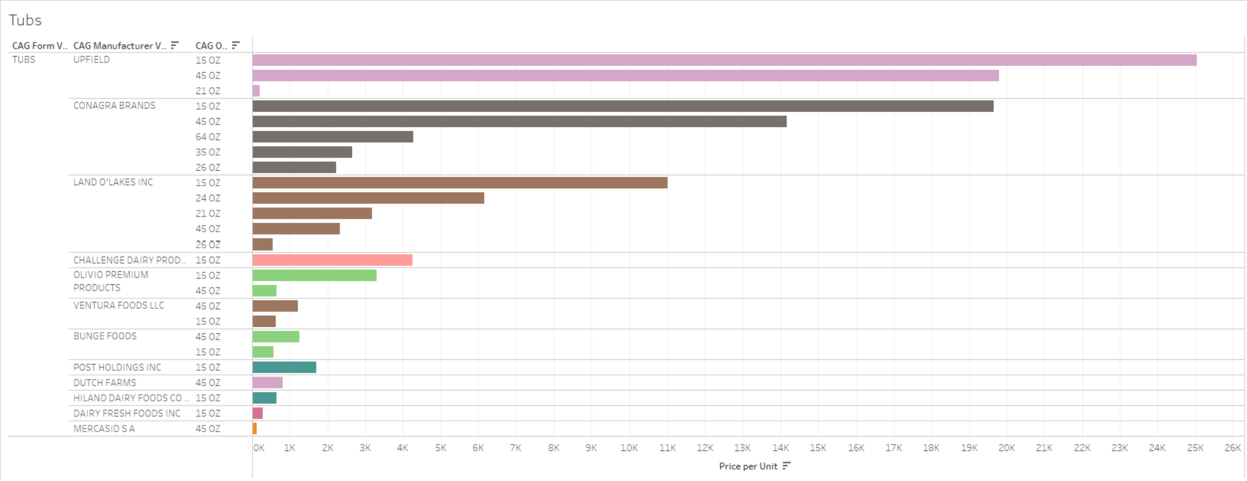
3. Future Roadmap

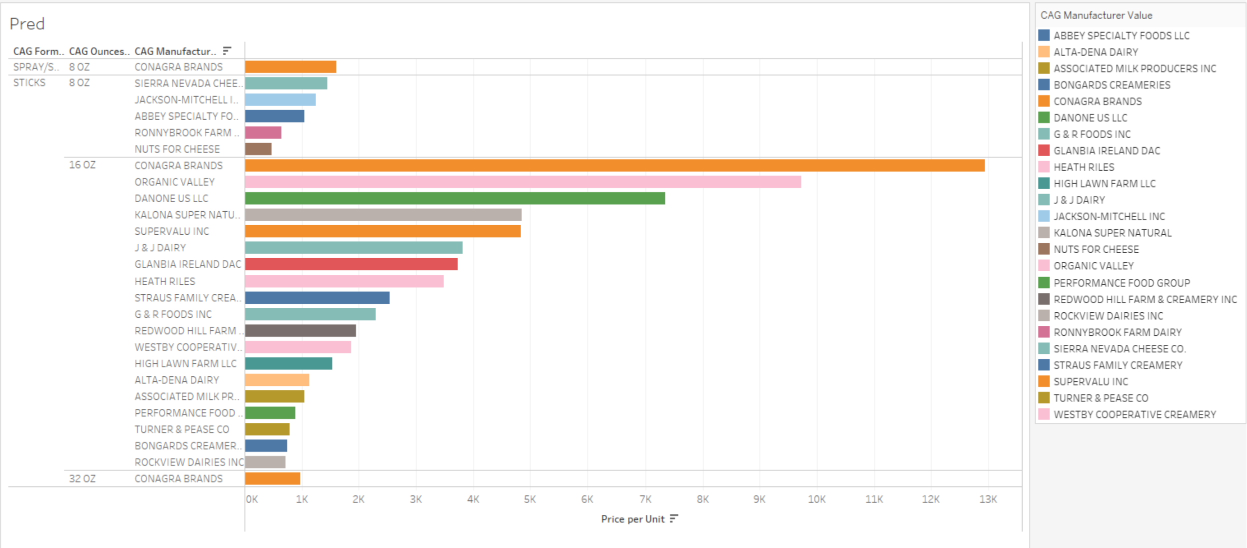
- We have explored several data-driven strategies that Conagra could use to optimize their merchandising strategies
- These strategies include tailoring approaches based on geography, product types, and specific brands
- We have also identified important interactions between different categories such as table spreads, cooking oil, and cooking spray
- Additionally, we found interactions between various attributes of the table spread category and potential price gaps
- Concepts such as varying confidence intervals, interaction effects, collinearity, principal component analysis, and model selection were already explored
- We will delve even deeper into these areas in our final report.

4. Appendix

4.1 Are there price gaps and/or price thresholds that cause an unexpected impact on sales and/or velocities?

Here we are checking the price gap across brands





We compared the prices of sticks across three different sizes: 8 oz, 16 oz, and 32 oz of the top 20 brands, and then compared them against the Conagra Brands. Based on our analysis, Conagra Brands consistently has the highest price per unit in all three segments. This significant price gap is especially noteworthy when compared to other top brands such as Organic Valley and Danone US LLC, whose price per unit was under 38% of Conagra's. In fact, Conagra's price per unit was vastly different from the majority of other brands in the market.

In addition, it may be worth considering whether Conagra's high prices could potentially limit its market share or lead to decreased customer loyalty over time. While Conagra's pricing power is undoubtedly a strength, there may be opportunities for the company to explore more competitive pricing strategies that could appeal to price-sensitive consumers and drive higher sales volumes. Overall, our analysis suggests that Conagra's pricing strategy has positioned the company as a leader in the market, but there may be opportunities to refine this strategy in order to maintain its competitive edge and drive long-term growth.

.