**Name: Himanshu Sharma Topic: SNACK CHAIN**

***Handling Missing Values:***

*PRICE* has missing some values. We will handle it by considering the value of BASE\_PRICE and TPR\_ONLY. If TRP\_ONLY is 0 ie there is no promotion discount and therefore PRICE = BASE\_PRICE.

*BASE\_PRICE* has missing some values. We will handle it by considering the value of PRICE and TPR\_ONLY. If TRP\_ONLY is 0 ie there is no promotion discount and therefore PRICE = BASE\_PRICE.

*PARKING* has missing values. It has over 0.2 million missing values. It is not possible to impute such a large number and hence we will drop the column.

***Handling Duplicate Values:***

*STORE\_ID has* values. We will handle it by considering the value of AVG\_BASKET\_SIZE to find out which segment the store belongs to. We will see the mean price for store SEGMENT to relate to the closest entry.

***Merging the Tables:***

We will join the table products and transactions by UPC. This table will be merged with the store’s table using the primary key STORE\_ID. To join the table we will rename the column name from STORE\_NUM to STORE\_ID in the transaction table.

***Dropping Not Relevant Features:***

We will drop features that we do not require for analysis PARKING, WEEK\_END\_DATE.

Also dropping the value for category ORAL HYGIENE PRODUCTS as we are not including it in our analysis

***Feature Engineering:***

*PRODUCT\_SIZE:* It has the size and the units in a column. We will separate out the units and the values to consider the value in the analysis as numeric.

*DISPLAY, FEATURE, TPR\_ONLY, HHS, UNITS, VISITS, UPC, STORE\_NUM,* *MSA:* Consider as factors.

***Correlation Matrix:***

Table

Description automatically generated

The corrplot shows a high correlation between UNITS, VISITS, HHS, and SPEND.

The co-relation is expected as visits will be converted into units purchased, number of household purchased and spend amount.

Also, more number of units purchased and more number of house hold purchases will contribute to spend.

Also, we can see a high correlation between BASE\_PPRICE and PRICE but this is expected as a PRICE consists of temporary price reduction.

***Exploratory Data Analysis of Dependent Variable :***

The histogram of SPEND,UNITS and HHS shows the distribution is right skewed. Considering the distribution of the SPEND, UNITS and HHS, the log transformation shows a normal distribution.

The box plot above shows that the each of the distributions is weighted towards lower spend for SEGMENTS .

Similary, the box plot above shows that the distribution is weighthed towards lower spend for CATEEGORY too.

|  |  |
| --- | --- |
| SPEND: Without log transformation | SPEND: With log transformation |
| Graphical user interface, chart, text, application  Description automatically generated | Chart, histogram  Description automatically generated |
| UNITS: Without log transformation | UNITS: With log transformation |
| Chart  Description automatically generated | Chart, histogram  Description automatically generated |
| HHS: Without log transformation | HHS: Withlog transformation |
| Chart  Description automatically generated | Chart, histogram  Description automatically generated |
| Plot for SPEND distribution for SEGEMNT | Plot for SPEND distribution for CATEGORY |
| A picture containing chart  Description automatically generated | A picture containing diagram  Description automatically generated |

***Predictor Table:***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **SPEND** | **UNITS** | **HHS** | **Effect** |
| BASE\_PRICE | Yes(+ve) | Yes(+ve) | Yes(+ve ) | Higher base pricing can make the product less appealing, resulting in lower sales and spending. |
| DISPLAY | Yes(+ve) | Yes(+ve) | Yes(+ve) | Products on an in-store promotional display have more visibility and can attract more people, resulting in increased sales and spending. |
| FEATURE | Yes(+ve) | Yes(+ve) | Yes(+ve) | Products included in an in-store circular have more visibility and can attract more customers, resulting in higher sales and spending. |
| HHS | No | No | \_ | An increase in spend and an increase in the number of units is the effect of the number of increased households purchase and it is not a cause |
| PRICE | No | No | No | The price consists of promotional effects and hence we will not include price as we are considering the promotion and base\_price for granularity |
| SPEND | \_ | No | No | An increase in spend is due to a greater number of units purchased and hence would lead to high co-relation.  A greater number of household purchasing would lead to high spend and hence would lead to high co-relation. |
| TPR\_ONLY | Yes(+ve) | Yes(+ve) | Yes(+ve) | Temporary price cuts may attract price-sensitive clients who are not loyal to the brand, resulting in lower spending. |
| UNITS | No | \_ | No | An increase in the number of units purchased would lead to an increase in spend and would lead to high co-relation |
| VISITS | No | No | No | A person entering a store is likely to make a purchase and hence including it will lead to a correlation |
| WEEK\_END\_DATE | No | No | No | The date variable may not provide any relevant information related to seasonality or promotions. |
| UPC | \_ | \_ | \_ | Considering product elasticity analysis. |
| STORE\_ID | Yes(?) | Yes(?) | Yes(?) | Considered for joining transactions and store tables. Also, the purchasing pattern vary for different stores |
| CATEGORY | Yes(?) | Yes(?) | Yes(?) | Categories may have different price sensitivities or client preferences, which can influence the effect. |
| DESCRIPTION | No | No | No | Considering UPC Universal Product Code instead. UPC corresponds to the Description. Using only for product name extraction |
| MANUFACTURER | No | No | No | The effect of different manufacturers will be captured in the product category and therefore we will not consider manufacturers to avoid correlation |
| PRODUCT\_SIZE | Yes(+ve) | Yes(+ve) | Yes(+ve) | Consumers may be more likely to purchase larger sizes if they perceive them to be better value for money. |
| SUB\_CATEGORY | No | No | No | We would be considering the category in the analysis and it will show a correlation with the category |
| CITY | No | No | No | This effect is captured in MSA therefore not including to avoid correlation |
| STATE | No | No | No | This effect is captured in MSA therefore not including to avoid correlation |
| AVG\_WEEKLY\_BASKETS | Yes(?) | Yes(?) | Yes(?) | A product is typically purchased together with other items in the store, then the number of baskets sold can be a good indicator of the demand for the product. |
| MSA | Yes(?) | Yes(?) | Yes(?) | People in densely populated areas may have different buying habits than people in rural areas and may affect consumer spending patterns. |
| SEGMENT | Yes(?) | Yes(?) | Yes(?) | The store's appeal is an important factor that can influence the purchasing behavior of customers. |
| SIZE | No | No | No | We would be including the outlet segment and average baskets to capture this effect. So we will not be including the size |
| STORE\_NAME | No | No | No | We are including STORE\_ID in our analysis to capture the effect. Therefore, not including to avoid co-relation |
| PARKING | No | No | No | Dropped due to a large number of missing values |

***Models:***

Because it contains information at various levels of analysis, this data is multi-level. The data, for example, contains information on Level1: Individual purchases (e.g., spend, units, households)

Level2: Products (e.g., category, size, manufacture)

Level3: Stores where they were sold (e.g., store size, location, segment).

To fully account for the interdependence and variability at each level of analysis, the data's nested structure necessitates the employment of multilevel modeling tools. We will therefore be considering random effects models.

As the distribution for the target is not normal we will try for log transformation of target variables.

By introducing interaction terms, we can capture these nonlinear interactions and gain a deeper understanding of how different factors interact to influence spending. This can lead to more accurate and intelligent spending projections.

Also, as we want to study the effect of display, feature, and TPR on SPEND vary by product categories and store segments we will only include interaction terms for spend.

Model for SPEND:

m1\_spend = lmer( log(SPEND\_I) ~ DISPLAY\*CATEGORY + FEATURE\*CATEGORY + TPR\_ONLY\*CATEGORY +

DISPLAY\*SEGMENT + FEATURE\*SEGMENT + TPR\_ONLY\*SEGMENT +BASE\_PRICE + PRODUCT\_SIZE +

AVG\_BSKT + ( 1 | STORE\_ID ) + (1 | MSA), data = df)

Model for UNITS:

m1\_unit = lmer(log(UNITS) ~ DISPLAY + FEATURE + TPR\_ONLY + BASE\_PRICE+ CATEGORY +

PRODUCT\_SIZE + SEGMENT + AVG\_BSKT + ( 1 | STORE\_ID ) + (1 | MSA), data = df)

Model for HHS:

m1\_HHS = lmer(log(HHS)~ DISPLAY + FEATURE + TPR\_ONLY + BASE\_PRICE+ CATEGORY +

PRODUCT\_SIZE + SEGMENT + AVG\_BSKT + ( 1 | STORE\_ID ) + (1 | MSA), data = df)

Table

Description automatically generated

|  |  |
| --- | --- |
| **Test** | **Observation: SPEND** |
| *Multicollinearity: Passed*  We can see from the test all variables below 5  We can see a high VIF between interaction terms, but the high VIF is expected. | > vif(m1\_spend)  GVIF Df GVIF^(1/(2\*Df))  DISPLAY 3.857266 1 1.963992  CATEGORY 9.217688 2 1.742431  FEATURE 12.317177 1 3.509584  TPR\_ONLY 3.309048 1 1.819079  SEGMENT 1.078530 2 1.019080  BASE\_PRICE 3.405966 1 1.845526  PRODUCT\_SIZE 3.584099 1 1.893172  AVG\_BSKT 1.075611 1 1.037117  DISPLAY:CATEGORY 6.132528 2 1.573656  CATEGORY:FEATURE 20.400089 2 2.125240  CATEGORY:TPR\_ONLY 2.964756 2 1.312192  DISPLAY:SEGMENT 2.729858 2 1.285390  FEATURE:SEGMENT 2.859347 2 1.300369  TPR\_ONLY:SEGMENT 1.832228 2 1.163442 |
| **Test** | **Observation: UNITS** |
| *Multicollinearity: Passed*  We can see from the test all variables below 5 | > vif(m1\_unit)  GVIF Df GVIF^(1/(2\*Df))  DISPLAY 1.286191 1 1.134104  FEATURE 1.328100 1 1.152432  TPR\_ONLY 1.028508 1 1.014154  BASE\_PRICE 3.348060 1 1.829770  CATEGORY 6.891653 2 1.620246  PRODUCT\_SIZE 3.516390 1 1.875204  SEGMENT 1.075627 2 1.018393  AVG\_BSKT 1.075600 1 1.037111 |
| **Test** | **Observation: HHS** |
| *Multicollinearity: Passed*  We can see from the test all variables below 5 | > vif(m1\_HHS)  GVIF Df GVIF^(1/(2\*Df))  DISPLAY 1.286192 1 1.134104  FEATURE 1.328102 1 1.152433  TPR\_ONLY 1.028508 1 1.014154  BASE\_PRICE 3.348059 1 1.829770  CATEGORY 6.891645 2 1.620245  PRODUCT\_SIZE 3.516396 1 1.875206  SEGMENT 1.075624 2 1.018392  AVG\_BSKT 1.075600 1 1.037112 |

***Assumptions:***

***Interpretations:***

Equation for SPEND:

log(SPEND) = 1.723 + 0.809(DISPLAY1) + 1.024(CATEGORYCOLD CEREAL) - 0.083(CATEGORYFROZEN PIZZA) + 0.302(FEATURE1) - 0.075(TPR\_ONLY1) + 0.168(SEGMENTUPSCALE) - 0.337(SEGMENTVALUE) + 0.020(BASE\_PRICE) + 0.036(PRODUCT\_SIZE) + 0.00003(AVG\_BSKT) - 0.175(DISPLAY1:CATEGORYCOLD CEREAL) - 0.168(DISPLAY1:CATEGORYFROZEN PIZZA) + 0.313(CATEGORYCOLD CEREAL:FEATURE1) + 0.386(CATEGORYFROZEN PIZZA:FEATURE1) + 0.151(CATEGORYCOLD CEREAL:TPR\_ONLY1) + 0.265(CATEGORYFROZEN PIZZA:TPR\_ONLY1) - 0.059(DISPLAY1:SEGMENTUPSCALE) + 0.030(DISPLAY1:SEGMENTVALUE) - 0.207(FEATURE1:SEGMENTUPSCALE) + 0.093(FEATURE1:SEGMENTVALUE) - 0.057(TPR\_ONLY1:SEGMENTUPSCALE) - 0.140(TPR\_ONLY1:SEGMENTVALUE)

Equation for UNITS:

log(UNITS) = 1.674 + 0.812(DISPLAY1) + 0.918(CATEGORYCOLD CEREAL) - 0.022(CATEGORYFROZEN PIZZA) + 0.780(FEATURE1) + 0.214(TPR\_ONLY1) + 0.131(SEGMENTUPSCALE) - 0.319(SEGMENTVALUE) - 0.246(BASE\_PRICE) + 0.032(PRODUCT\_SIZE) + 0.00002(AVG\_BSKT)

Equation for HHS:

log(HHS) = 1.533 + 0.808(DISPLAY1) + 0.912(CATEGORYCOLD CEREAL) - 0.082(CATEGORYFROZEN PIZZA) + 0.725(FEATURE1) + 0.155(TPR\_ONLY1) + 0.131(SEGMENTUPSCALE) - 0.350(SEGMENTVALUE) - 0.218(BASE\_PRICE) + 0.029(PRODUCT\_SIZE) + 0.00002(AVG\_BSKT)

1. **What are the effects of product display, being featured on in-store circular, and temporary price reduction on product sales (spend), unit sales, and number of household purchasers?**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Effects of products display | Effects of in-store circular | Effects of the temporary reduction |
| SPEND | 80% more than the effect of not displaying the products | 30% more than the effect of not being in-store circular | 7.5% less than the effect of not having a temporary price reduction |
| UNITS | 81.2% more than the effect of not displaying the products | 78% more than the effect of not being in-store circular | 21.4% more than the effect of not having a temporary price reduction |
| HHS | 80.8% more than the effect of not displaying the products | 72.5% more than the effect of not being in-store circular | 15.5% more than the effect of not having a temporary price reduction |

1. **How do the effects of display, feature, and TPR on SPEND vary by product categories (cold cereals, frozen pizza, bag snacks) and store segments (mainstream, upscale, value)?**

**Product Category for Display (BAG SNACKS – Base Case)**

The effect on spend of the product being on display for category cold cereals is 63.4% more than the category BAG SNACKS

The effect on spend of the product being on display for category frozen pizza is 64.1% more than the category BAG SNACKS

**Product Category for Feature (BAG SNACKS – Base Case)**

The effect on spend of the product being on feature for category cold cereals is 61.5% more than the category BAG SNACKS

The effect on spend of the product being on feature for category frozen pizza is 68.8% more than the category BAG SNACKS

**Product Category for TPR (BAG SNACKS – Base Case)**

The effect on spend of the product being on price reduction for category cold cereals is 7.6 % more than the category BAG SNACKS

The effect on spend of the product being on price reduction for category frozen pizza is 19 % more than the category BAG SNACKS

**Segment for Display (MAINSTRAEM - Base Case)**

The effect of display for store being UPSCALE on the spend is 75% more than the segment MAINSTREAM.

The effect of display for store being VALUE on the spend is 83.9% more than the segment MAINSTREAM.

**Segment for Feature (MAINSTRAEM - Base Case)**

The effect of feature for store being UPSCALE on the spend is 9.5% more than the segment MAINSTREAM.

The effect of feature for store being VALUE on the spend is 90.2% more than the segment MAINSTREAM.

**Segment for TPR (MAINSTRAEM - Base Case)**

The effect of price reduction for store being UPSCALE on the spend is 13.2% less than the segment MAINSTREAM.

The effect of price reduction for store being VALUE on the spend is 21.5% less than the segment MAINSTREAM.

1. **What are the five most price elastic and five least price elastic products? Price elasticity is the change in units sold for change in product price.**

Product elasticity, also known as price elasticity of demand, is a measure of the responsiveness of the quantity requested of a product to price changes. It is a measure of the percentage change in a product's demand caused by a one percent change in its price.

glm(UNITS ~ PRICE, data = data,family = poisson(link = log). We will collect the beta coefficient.

Most Elastic 5 Products:

UPC Elasticity Description Avg Price

2066200532 0.7401054 NWMN OWN SUPREME PIZZA 6.209281

7218063979 0.6490624 FRSC PEPPERONI PIZZA 6.87836

7218063983 0.6454552 FRSC 4 CHEESE PIZZA 6.880683

7218063052 0.5991506 FRSC BRCK OVN ITL PEP PZ 6.880339

2066200530 0.5268903 NWMN OWN PEPPERONI PIZZA 6.310389

Least Elastic Products :

UPC Elasticity Description

1111085319 0.01334842 PL HONEY NUT TOASTD OATS

1111085345 0.01469244 PL RAISIN BRAN

7027316204 0.01603153 SHURGD MINI PRETZELS

7027316404 0.01652280 SHURGD PRETZEL STICKS

1111085350 0.02295120 PL BT SZ FRSTD SHRD WHT

1. **As the retailer, which products would you lower the price to maximize (a) product sales and (b) unit sales, and why?**

Items with higher elasticity are often more sensitive to price changes, which means that lowering the price might result in a considerable rise in demand, whereas items with lower elasticity are less responsive to price changes.

Maximize product sales:

As the sales depend on both price and number of units, we need to find products with a comparatively higher price and higher elasticity. The overall average price for FRSC BRCK OVN ITL PEP PZ , FRSC 4 CHEESE PIZZA, and FRSC PEPPERONI PIZZA i.e. Manufacturer TONYS products are comparatively high out of the high elastic products. So we should consider lowering the price by a smaller fraction to result in higher demand and hence higher sales. We will consider the cost-benefit tradeoff here.

Maximize unit sales:

The retailer should concentrate on the products with the highest elasticities because they are the most sensitive to price adjustments and a lower price is likely to result in an increase in unit sales. The most elastic product in this example is NWMN OWN SUPREME PIZZA, which has an elasticity of 0.74. To enhance unit sales, the store should consider decreasing the price of this product.