**Assignment 8**

**Mudasir Wazir**

1. **Predictor Table:**

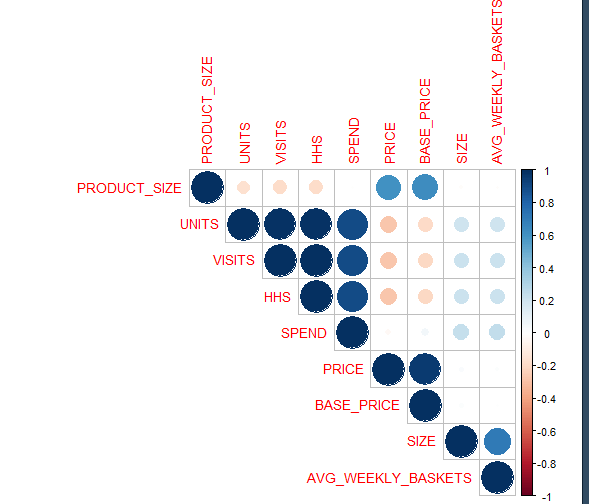
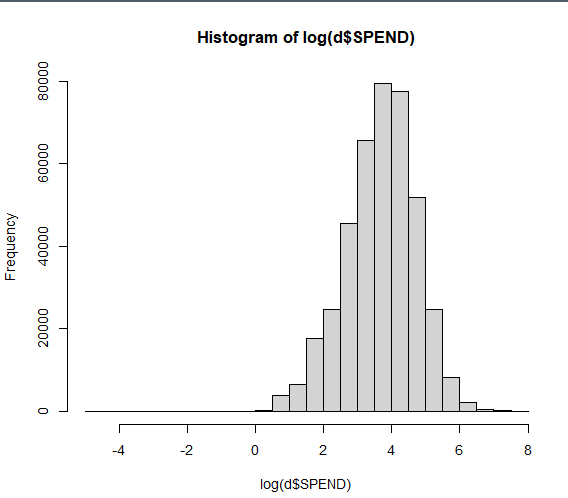
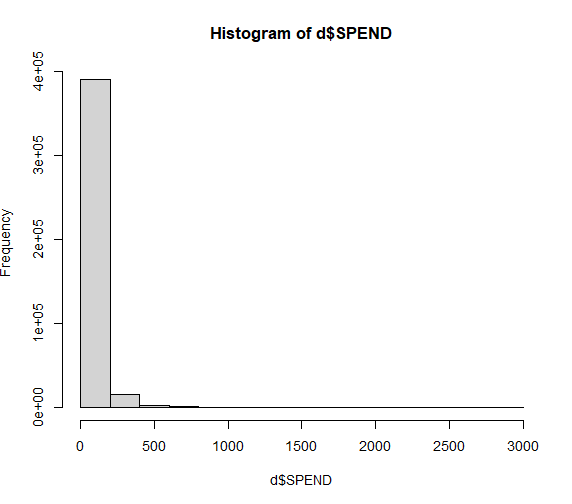
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Predictor** | **Effect of SPEND** | **Effect on UNITS** | **Effect on HHS** | **Rationale** |
| MANUFACTURER | Variable | Variable | Variable | Depends on brand strength and market presence; stronger brands may command higher spend and attract more households. **Private labels** or lesser-known brands might compete on price, leading to different purchasing patterns, potentially higher units sold due to lower prices, but lower total spend |
| CATEGORY | Variable | Variable | Variable | Different categories (e.g., cold cereals vs. frozen pizza) have different buying frequencies and price points, influencing SPEND and UNIT sales. |
| FEATURE | + | + | + | Featured products are more visible, likely increasing units sold, total spend, and the number of households purchasing. |
| DISPLAY | + | + | + | In-store displays can enhance product visibility and accessibility, likely increasing sales and household reach. |
| TPR\_ONLY | Variable | + | + | Price reductions may decrease total SPEND but increase UNIT sales and household participation. |
| SEGMENT | Variable | Variable | Variable | Store segments cater to different demographic groups, affecting spending patterns and purchase volumes. |
| PRODUCT\_SIZE | Variable | Variable | Variable | Larger sizes may increase SPEND but have mixed effects on UNITS and HHS depending on the product type and consumer preference. |
| VISITS | + | + | + | More visits imply higher exposure and potential sales, increasing all dependent measures. |
| PRICE | + | - | Variable | Higher prices typically reduce units sold but might increase SPEND; effect on HHS could vary depending on price sensitivity. |
| MSA | + | + | + | Metropolitan areas with higher population density may see higher sales volume, SPEND, and more diverse HHS. |
| SIZE | + | + | + | Larger stores might process more transactions, affecting all measures positively due to higher traffic and product availability. |
| AVG\_WEEKLY\_BASKETS | + | + | + | Indicates higher store traffic, potentially increasing sales volume, SPEND, and number of households. |
| PARKING | + | + | + | More parking may increase store accessibility and convenience, potentially increasing all dependent variables. |

1. **Data Preparation:**

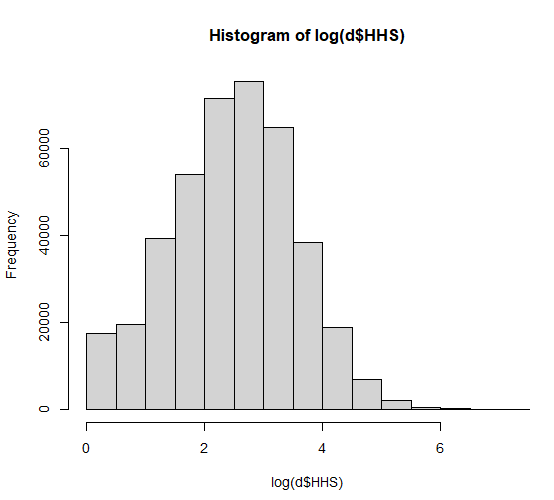
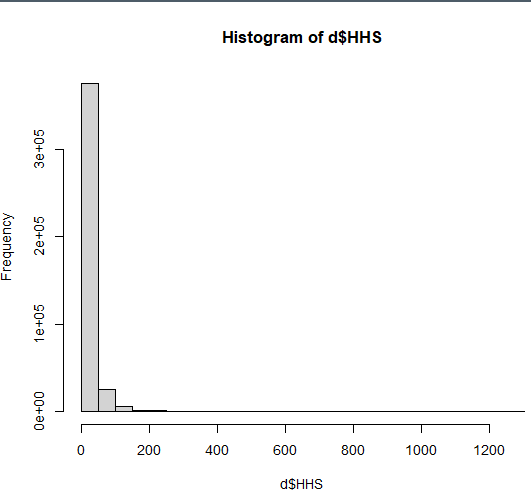
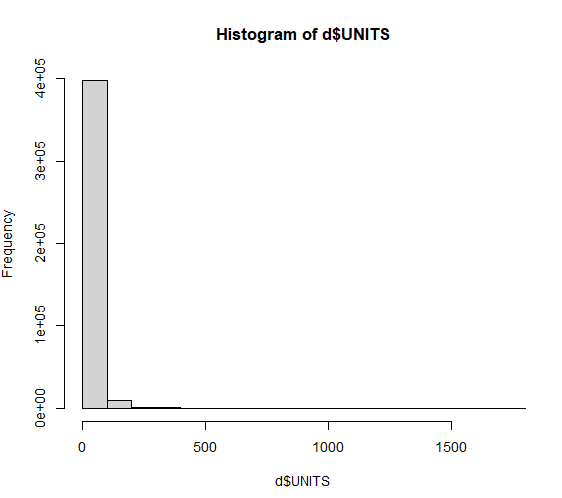
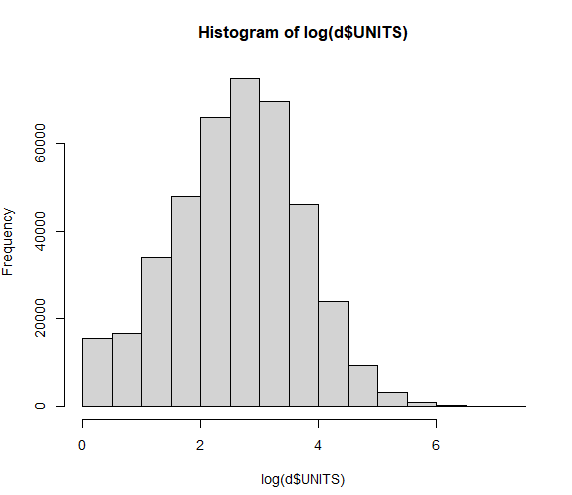
* There are 3 separate datasets – Products, Stores and Transactions on a weekly level.
* The Product and Transactions data sets are joined using “UPC” as a Foreign Key.
* This is then joined with Stores dataset with Store Num=Store ID as Foreign key.
* There are some duplicate stores (4503 & 17627) which are recorded in 2 segments; thus, we only keep only the one in the Mainstream segment since this is the most frequent segment in our data.
* Further, the “ORAL HYGIENE PRODUCTS” category is dropped.
* Next, we find that there are some NAs in our data PARKING variable has 277348 NA’s thus we drop this variable. We also remove other NA values in PRICE and BASE\_PRICE.
* We drop DESCRIPTION, SUB-CATEGORY, and STORE NAME as they might be redundant.
* We convert the categorical variables to Factors ending up with 11 factor variables, 9 numeric variables and 1 date variable.
* In the final dataset we have 408238 observations with 21 variables.

1. **Exploratory Data Analysis and Data Visualization:**

* SPEND, UNITS, and HHS were Log transformed to them normally distributed and eliminate the skewness.



**Correlation Plot**



**Log- transformed distribution of DVs.**

* Base price and Price are highly correlated. We will drop Base price.
* Units, Visits, HHS and SPEND are all highly correlated with each other. Thus, when using one of them as DV, others will not be used as predictors.

1. **Data Modelling:**

Based on the hierarchical nature of the data featuring multiple levels of nested data structures and the questions we need to answer, we will use Mixed Effects models for our analysis. This includes products (identified by UPCs) that are nested within stores, and both products and stores are subject to various fixed effects like pricing, promotions (FEATURE, DISPLAY, TPR\_ONLY), and store characteristics (SIZE, AVG\_WEEKLY\_BASKETS).

Mixed-effects models provide the necessary statistical framework to account for the nested data structure, control for multiple levels of variability, and handle the mixed influences of both observed and unobserved factors on the outcomes of interest. This approach ensures robust, accurate, and interpretable results that can inform strategic decisions based on the analysis.

For our analysis we will be using Store\_num and UPC for random effects. Further we will include interaction terms as they are improving the model and help us answer Question 2.

**Models:**

**Model 1: Mixed Effects Model for DV – SPEND.**

(1 | STORE\_NUM): Random intercepts for stores, allowing baseline unit sales to vary by store. This accounts for unobserved heterogeneity among stores that might affect sales.

(1 | UPC): this term allows the baseline spending to vary by product.

model\_spend <- lmer(log(SPEND) ~ PRODUCT\_SIZE+SIZE+AVG\_WEEKLY\_BASKETS +STATE+

PRICE + FEATURE + DISPLAY + TPR\_ONLY +

(FEATURE + DISPLAY + TPR\_ONLY):CATEGORY +

(FEATURE + DISPLAY + TPR\_ONLY):SEGMENT +

(1 | STORE\_NUM) + (1 | UPC), data = d, REML = FALSE)

**Model 2: Mixed Effects Model for DV – UNITS.**

For this model, we will use Log of Price as this helps us to easily interpret Price elasticity and answer the Price Elasticity questions. Further, we will use nested random effects to get the elasticities for each product.

(1 + log (PRICE) | UPC): Random intercepts and random slopes for the effect of log (PRICE) by product. This allows both the baseline level of sales and the sensitivity to price changes to vary across products. It's crucial for capturing product-specific elasticity that might differ due to unique product characteristics or consumer preferences.

model\_units <- lmer(log(UNITS) ~ log(PRICE)+ FEATURE + DISPLAY + TPR\_ONLY +

PRODUCT\_SIZE+SIZE+AVG\_WEEKLY\_BASKETS +STATE+

(FEATURE + DISPLAY + TPR\_ONLY): CATEGORY +

(FEATURE + DISPLAY + TPR\_ONLY): SEGMENT + (1 | STORE\_NUM) +

(1 +log(PRICE) | UPC) , data = d, REML = FALSE)

**Model 3: Mixed Effects Model for DV – HHS.**

model\_HHS <- lmer(log(HHS) ~ PRICE + FEATURE + DISPLAY + TPR\_ONLY +

PRODUCT\_SIZE +SIZE+AVG\_WEEKLY\_BASKETS +STATE+

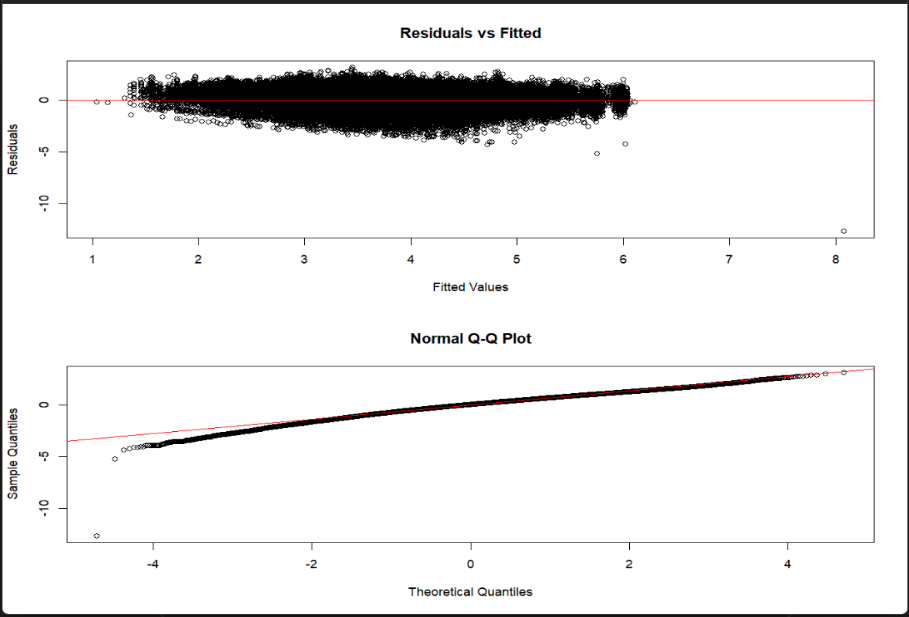
(FEATURE + DISPLAY + TPR\_ONLY): CATEGORY +

(FEATURE + DISPLAY + TPR\_ONLY): SEGMENT +

(1 | UPC) + (1 | STORE\_NUM), data = d, REML = FALSE)

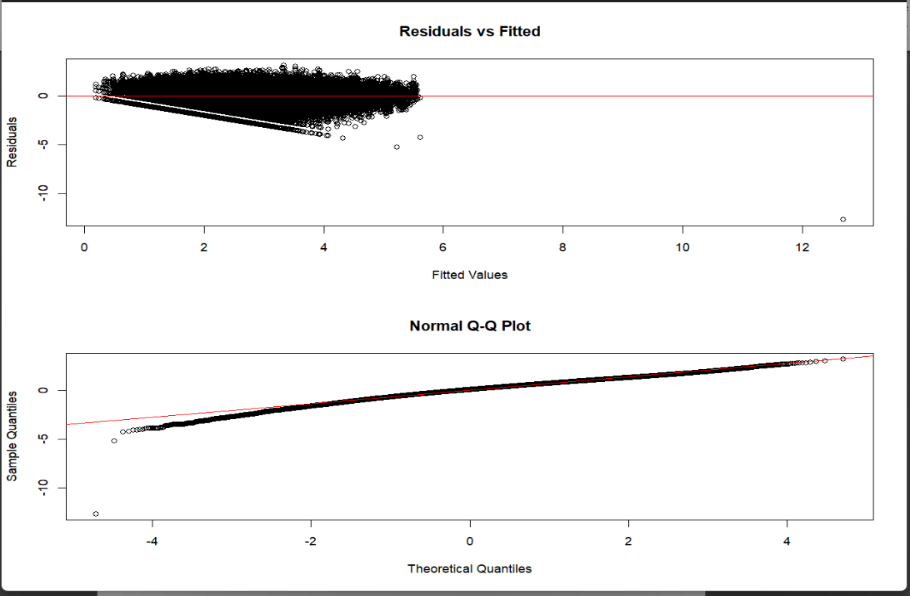
**Assumptions Check:**

* **Linearity**: The Residuals vs. Fitted plot shows a relatively even spread around the zero line without any systematic pattern, suggesting that the linearity assumption is largely met.



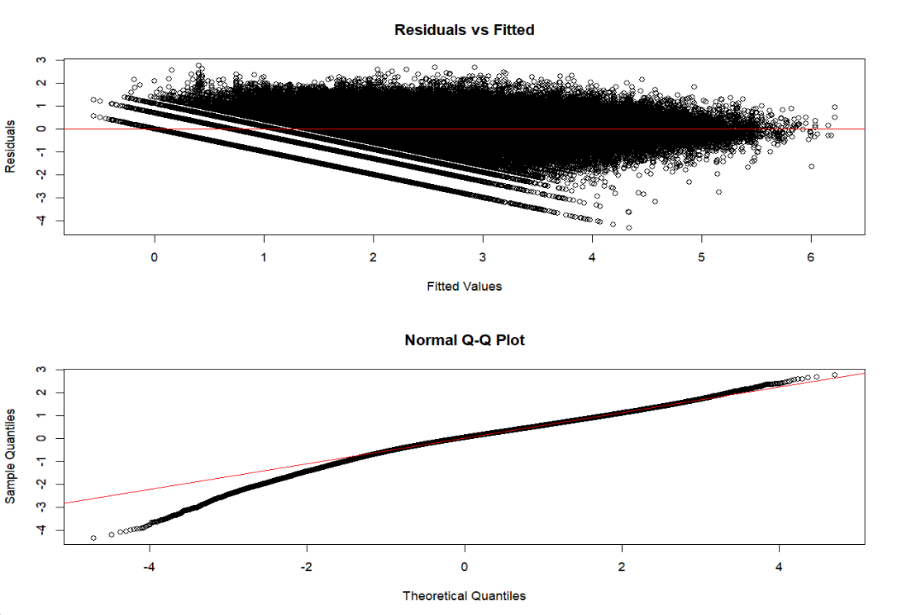
**Model 1: Residual plot and QQ plot**

* **Normality**: The Q-Q plot reveals slight deviations from the theoretical line at one tail, indicating some departure from normality but its not extreme.
* **Constant Variance**: The plot does not display any obvious signs of increasing or decreasing spread along the range of fitted values, indicating that the assumption of constant variance is reasonably satisfied.
* **Linearity**: The residuals show an apparent pattern, suggesting violation of the linearity assumption.



**Model 2: Residual plot and QQ plot**

* **Normality**: The Q-Q plot for this model shows a closer adherence to the line compared to Model 1, although slight deviations at the tails are still present, suggesting minor issues with normality.
* **Constant Variance**: The variance of the residuals also appears to change across the range of fitted values, hinting at potential heteroscedasticity.
* **Linearity**: There is some pattern in the Residuals vs. Fitted plot, suggesting potential non-linearity in the relationship between predictors and the response variable.



**Model 3: Residual plot and QQ plot**

* **Normality**: The Q-Q plot exhibits noticeable deviations from normality, particularly with heavy tail on one side, indicating that the residuals do not follow a normal distribution.
* **Constant Variance**: The residuals display a pattern of increasing spread with higher fitted values, suggesting heteroscedasticity, which is a violation of the constant variance assumption.

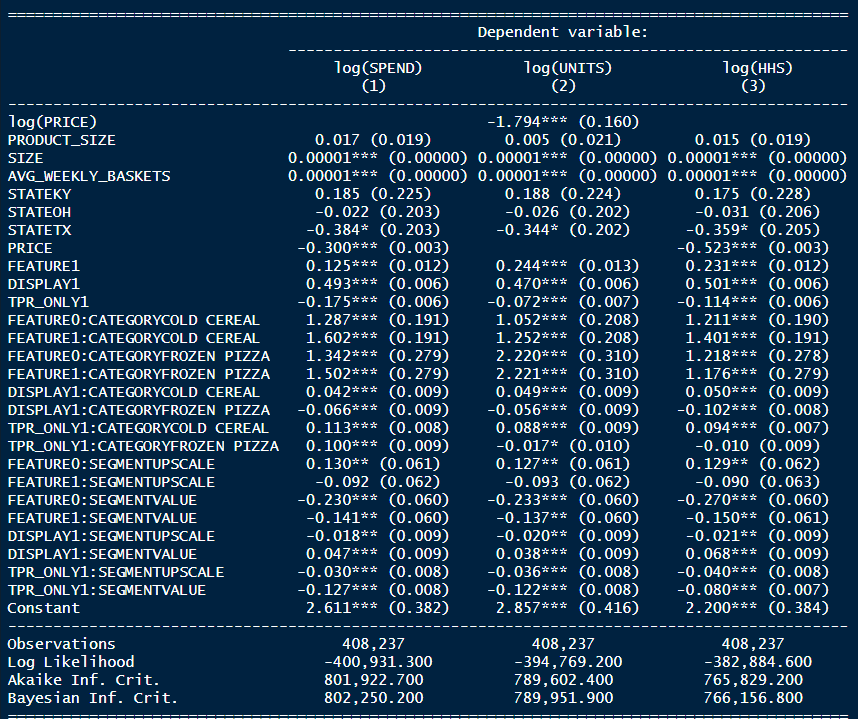
***While Model 1 shows better adherence to the assumptions, Models 2 and 3 demonstrate issues.***

***For future analyses a GLMER model might be a better approach which captures non-linear relations and non-normal distribution.***

Further, the assumption of **Independence** is generally addressed by the random effects structure of mixed models.

**Multicollinearity**: Only interaction terms introduced little multicollinearity.

1. **Outputs**



**Stargazer output for all three models**

**Q1: 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?**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variables** | **Effects on Product Sales**  **(SPEND)** | **Effects on Unit Sales (UNITS)** | **Effects on Number of Household Purchasers (HHS)** |
| **Display** | The presence of a product display leads to a multiplicative increase in SPEND by about 63.7% (exp (0.493)) compared to not having display, suggesting that displays promote product sales. | Displays increase unit sales by about 60.0% (exp (0.470)), compared to not having display, demonstrating their effectiveness in boosting product quantity purchases. | Product displays result in a 65 % increase in the number of purchasing households (exp (0.501)) compared to not having display. |
| **Feature** | Being featured in an in-store circular increases SPEND by about 13.2% (exp (0.125)) compared to not featured, indicating enhanced revenue generation. | Featured products experience a 27.6% increase in unit sales (exp (0.244)) compared to not featured, showcasing the impact of advertising promotions in increasing volume. | Being featured boosts the number of household purchasers by about 26% (exp (0.231)) compared to not featured. |
| **Temporary Price Reduction** | Temporary price reductions decrease SPEND by approximately 16 % (exp (-0.175)) compared to not doing price reductions, reflecting the direct impact of lower prices on sales. | A temporary price reduction leads to a 7 % decrease in unit sales (exp (- 0.072)) compared to not doing price reductions. | Temporary price reductions decrease the number of household purchasers by 11% (exp (-0.114)) compared to not doing price reductions. |

**Q2: 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)?**

**Display1**

***Effect on Product Categories:***

* **Cold Cereals:** Increases SPEND by approximately 4.2 % compared to Bag snacks.
* **Frozen Pizza:** Decreases SPEND by approximately 6.4%. compared to Bag snacks.

***Effect on Store Segments:***

* **Value Segment:** Decreases SPEND by approximately 2%, compared to Mainstream.
* **Upscale Segment:** Increases SPEND by approximately 4.7 %, compared to Mainstream.

**Feature1**

***Effect on Product Categories****:*

* **Cold Cereals**: Increases SPEND by 396.2% compared to Bag snacks.
* **Frozen Pizza**: Increases SPEND by 350 % compared to Bag snacks.

***Effect on Store Segments****:*

* **Value Segment**: Decreases SPEND by 13% compared to Mainstream.
* **Upscale Segment**: Decreases SPEND by 9 % compared to Mainstream.

**Temporary Price Reduction (TPR)1**

***Effect on Product Categories****:*

* **Cold Cereals**: Increases SPEND by approximately 12 % compared to Bag snacks.
* **Frozen Pizza**: Increases SPEND by approximately 10 % compared to Bag snacks.

***Effect on Store Segments****:*

* **Value Segment**: Decreases SPEND by 12% compared to Mainstream.
* **Upscale Segment**: Decreases SPEND by 3 % compared to Mainstream.

**Q3: 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.**

A screenshot of a computer screen

Description automatically generatedUsing Model 2, we compute the coefficients of log(Price) for each product (UPC) which is the sum of fixed effect and random effect shown below.

The UPCs with the highest absolute values have the highest price elasticity and those with lowest **absolute** values have lowest price elasticity. Log of price indicates the percentage change in Units for a percentage change in price. (elasticity)

The 5 **most** price elastic products are:

* A screen shot of a computer screen

  Description automatically generatedFRSC PEPPERONI PIZZA
* DIGIORNO THREE MEAT
  + - FRSC BRCK OVN ITL PEP PZ
    - KELL FROOT LOOPS
    - MKSL PRETZEL STICKS

The 5 **least** price elastic products are:

* A screenshot of a computer screen

  Description automatically generatedPL PRETZEL STICKS
* PL HONEY NUT TOASTD OATS
* GM CHEERIOS
* SHURGD PRETZEL STICKS
* SHURGD MINI PRETZELS

**Q4: As the retailer, which products would you lower the price to maximize (a) product sales and (b) unit sales, and why?**

Lowering prices on the most elastic products is likely to result in a substantial increase in units sold and thus also product sales. Thus, the prices of most elastic products calculated for Q3 would be lowered.

* FRSC PEPPERONI PIZZA
* DIGIORNO THREE MEAT
* FRSC BRCK OVN ITL PEP PZ
* KELL FROOT LOOPS
* MKSL PRETZEL STICKS

However, for product sales, we must ensure that the increase in volume compensates for the loss in revenue per unit due to lowering the price. Products with a high negative elasticity (where a small decrease in price leads to a large increase in volume) are still candidates, but the calculation needs to consider the revenue impact directly. Ensure that the projected increase in volume from a price reduction translates into a net increase in revenue.