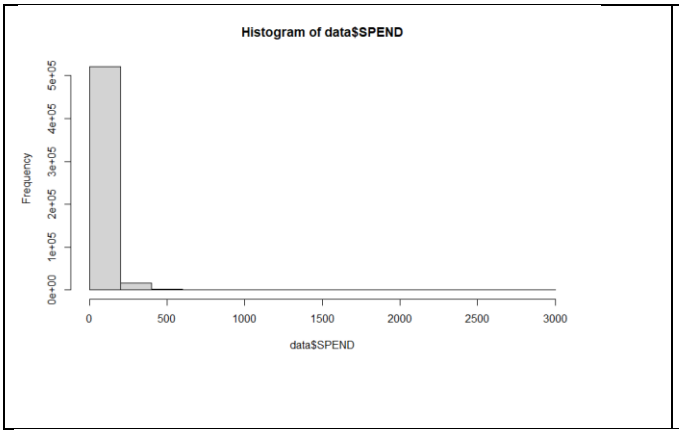
Aryan Sharma
Snack Chain

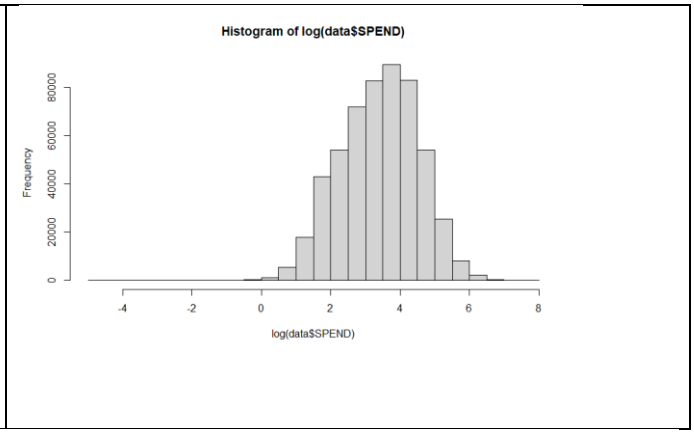
Exploratory Data Analysis & Processing

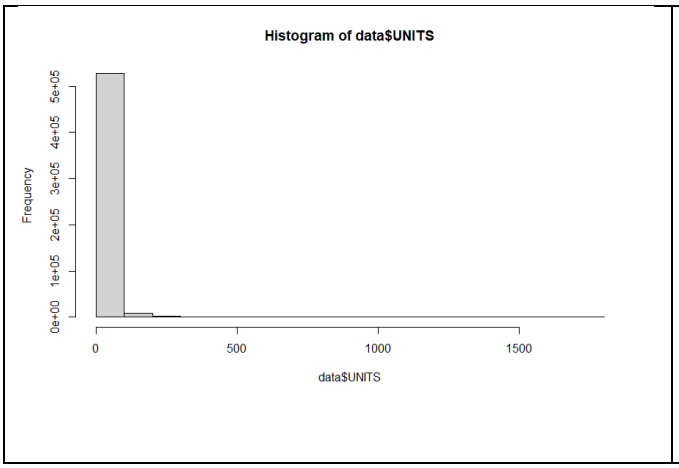
The data has 3 tables namely, stores, products, and transactions. In sheet stores there are 9 variables and 79 observations. In sheet Products, there are 6 variables and 58 observations and lastly in sheet transactions which amount for all the transactions at the store there are 12 variables and 524950 observations. The data is 37 Month data starting from Jan 2009 to Jan 2012. The joined data using R has 27 variables. A new column based on transaction weekend date is formulated highlighting the month and year in which the transactions have taken place known as Months. Out of the 28 Variables in the combined dataset 16 are numeric and 2 are in date format, the rest are character variables. Product size variables had different measurements such as OZ, Liters and Milliliters hence all the sizes are converted to OZ scale. The character variables are converted to factor variables depending on the need as mentioned below in the predictor table.

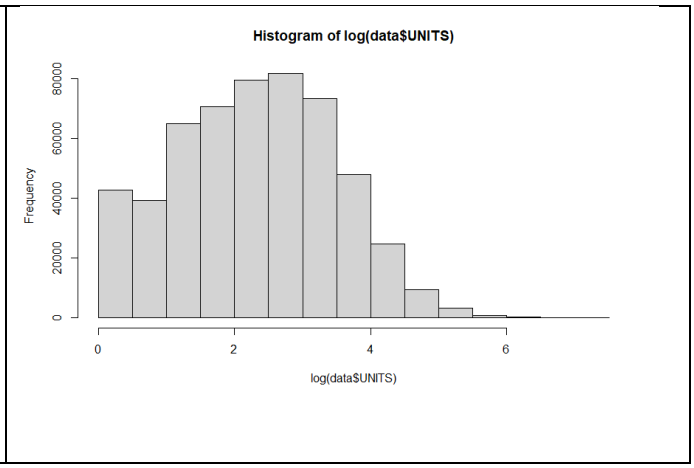
Histograms of DV:

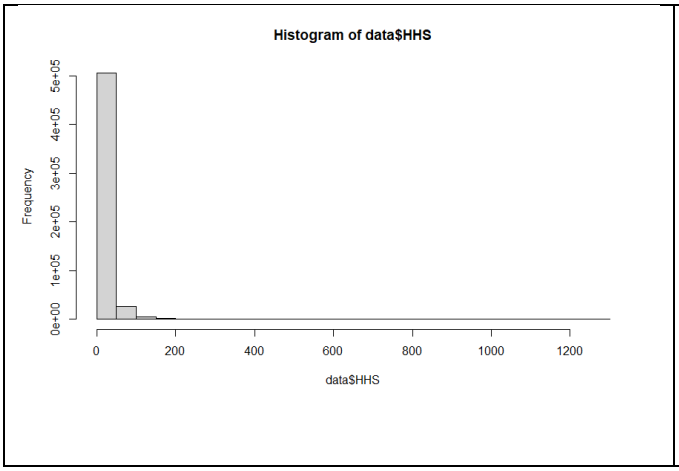
The distributions of **Spend, Units & HHS** are right-skewed. The distributions of **Spend, Units & HSS** are close to normal, and therefore, more suited for MLS regression. The variable is Poison distribution hence we'll have to use MLS models. Then we can determine which model is the best fit and has least violation of assumptions.

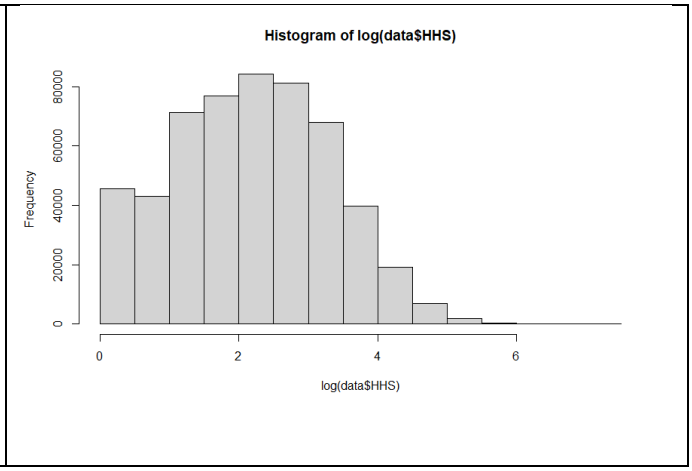






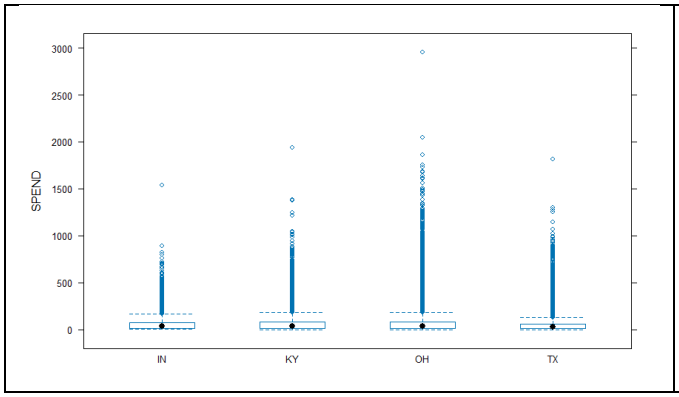


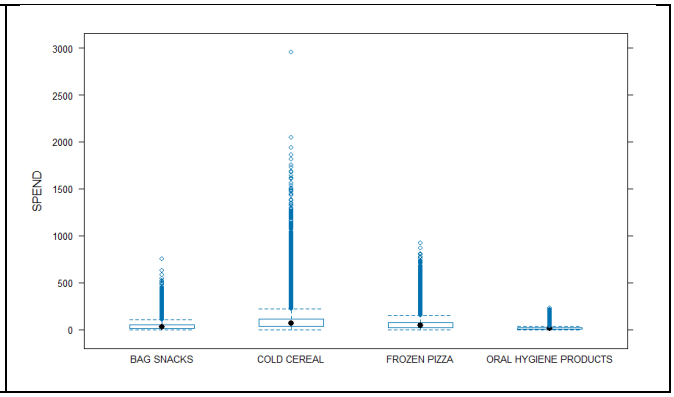


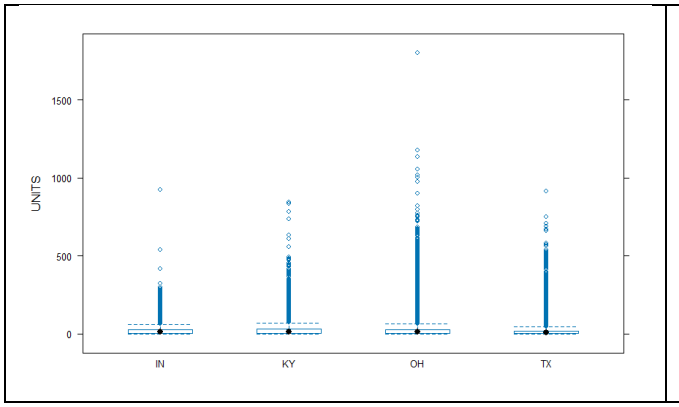


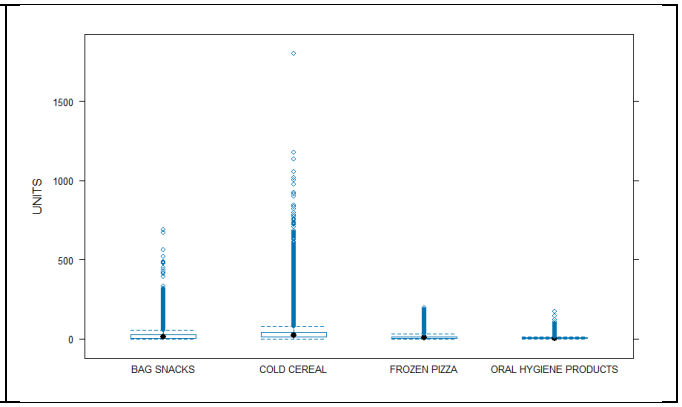
Boxplots:

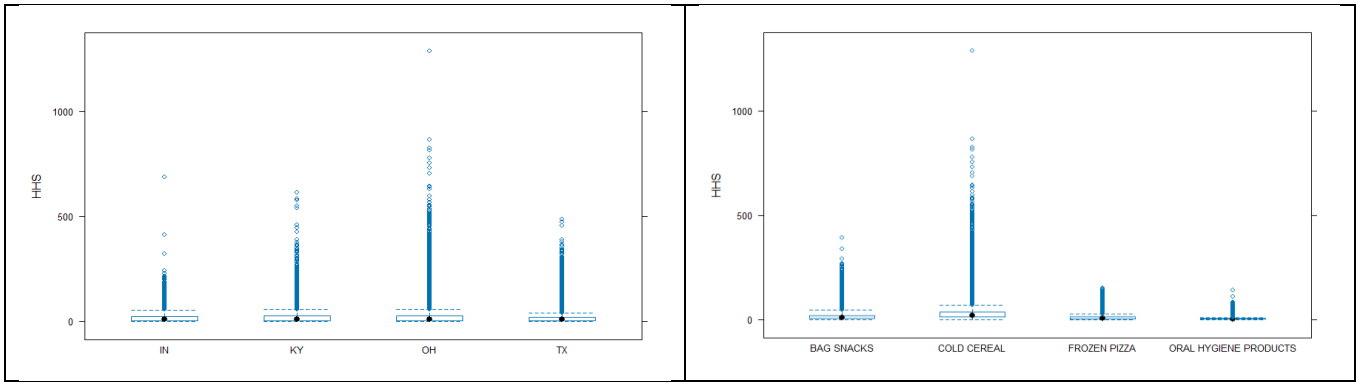
Boxplot of Spend, Units & HSS *with State and Category.*











Feature Engineering

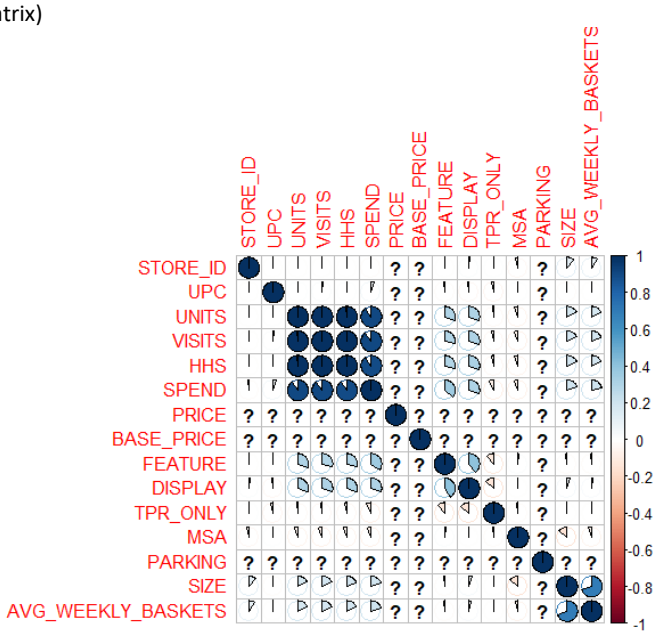
All the three tables have a unique key which can combine them. The table Products and transaction has UPC as a foreign key which connects those two tables. Once this is merged, we can use the store ID present in table transaction and stores to connect the data. Then from table transaction using variable WEEK_END_DATE I have created a new column month which has month and year of transaction.

stores_data = read_excel("C:/Users/91884/Desktop/BAIS/Statistical Data Mining/8/SnackChain (2).xlsx", sheet = "stores")
product_data = read_excel("C:/Users/91884/Desktop/BAIS/Statistical Data Mining/8/SnackChain (2).xlsx", sheet = "products")
transaction_data = read_excel("C:/Users/91884/Desktop/BAIS/Statistical Data Mining/8/SnackChain (2).xlsx", sheet = "transactions")
merged_data1 <- left_join(transaction_data, product_data, by = "UPC")
data <- left_join(merged_data1, stores_data, by = "STORE_ID")

Correlation:

There are only 15 variables in the data which are numeric. Hence, we are calculating whether they have any correlation between them or not.

```
correlation_matrix <- Cor(data[c("STORE_ID",
"UPC","UNITS","VISITS","HHS","SPEND","PRICE","BASE_PRICE","FEATURE","DISPLAY","TPR_ONLY","MSA","PARKING","SIZE","AVG_WEEKLY_BASKETS")])
corrplot(correlation_matrix, method = "pie")
corrplot(correlation_matrix, method = "pie")
print(correlation_matrix)
```



There is a correlation between Units -Visits and HHS, spend – Units, Visits and HHS, hence we are only taking Spend variable.

Predictor table:

Predictor	Spend	Units	HHS	Rationale
FEATURE	+	+	+	Whether the product was in store circular can be helpful in sales as it drives the publicity up
DISPLAY	+	+	+	If a product is visible to the customer the chances of sales may increase
PRICE	-	-	-	If the price of the same commodity is more since difference in brand, the shopper may tend to purchase the cheaper one
STATE	+/-	+/-	+/-	Population and socio-economic factors of the state can be a factor in determining the spend
PRODUCT SIZE	-	-	-	The quantity of the product will increase price and high price can lead to low sales
CATEGORY	+/-	+/-	+/-	Product belonging to which category can be a factor in determining the sales
SEGMENT	+/-	+/-	+/-	Segments like essential goods always will have more sales
TPR_ONLY	+	+	+	Rollback and high discount can lead to more sales
STORE ID	+/-	+/-	+/-	To identify which outlet is doing the sale

Regression Analysis

```
model_1 <- lmer(log(SPEND)~ FEATURE + DISPLAY + log(PRICE) + STATE + PRODUCT_SIZE + CATEGORY + SEGMENT
+(1|STORE_ID),data=data,REML=FALSE)
model_2 = lmer(log(UNITS)~ FEATURE + DISPLAY + log(PRICE) + STATE + PRODUCT_SIZE + CATEGORY + SEGMENT
+(1|STORE_ID),data=data,REML=FALSE)
model_3 = lmer(log(HHS)~ FEATURE + DISPLAY + log(PRICE) + STATE + PRODUCT_SIZE + CATEGORY + SEGMENT
+(1|STORE_ID),data=data,REML=FALSE)
model_4 = lmer(log(SPEND) ~ FEATURE * DISPLAY * TPR_ONLY * CATEGORY * SEGMENT + log(PRICE) + STATE +
PRODUCT_SIZE + (1|STORE_ID), data = data, REML = FALSE)
```

```
================================================================================
                                          Dependent variable:
                                ---------------------------------------------
                                   log(SPEND)        log(UNITS)         log(HHS)
                                       (1)               (2)               (3)
--------------------------------------------------------------------------------
FEATURE1                        0.609*** (0.004)  0.609*** (0.004)  0.565*** (0.004)
DISPLAY1                        0.664*** (0.004)  0.664*** (0.004)  0.666*** (0.004)
log(PRICE)                      0.149*** (0.003) -0.851*** (0.003) -0.769*** (0.003)
STATEKY                           0.043 (0.324)     0.043 (0.324)     0.027 (0.327)
STATEOH                           0.092 (0.295)     0.092 (0.295)     0.085 (0.298)
STATETX                          -0.221 (0.294)    -0.221 (0.294)    -0.209 (0.296)
PRODUCT_SIZE                    0.018*** (0.0002) 0.018*** (0.0002) 0.017*** (0.0002)
CATEGORYCOLD CEREAL             0.971*** (0.003)  0.971*** (0.003)  0.962*** (0.003)
CATEGORYFROZEN PIZZA            0.085*** (0.005)  0.085*** (0.005)  0.018*** (0.004)
CATEGORYORAL HYGIENE PRODUCTS -1.026*** (0.004) -1.026*** (0.004) -0.970*** (0.004)
SEGMENTUPSCALE                    0.003 (0.009)     0.003 (0.009)     0.003 (0.009)
SEGMENTVALUE                   -0.415*** (0.077) -0.415*** (0.077) -0.438*** (0.077)
TPR_ONLY1                       0.052*** (0.003)  0.052*** (0.003)  0.016*** (0.003)
Constant                        2.777*** (0.290)  2.777*** (0.290)  2.599*** (0.292)
--------------------------------------------------------------------------------
Observations                         538,619           538,619           538,619
Log Likelihood                    -622,321.200      -622,321.200      -605,946.700
Akaike Inf. Crit.                 1,244,674.000     1,244,674.000     1,211,925.000
Bayesian Inf. Crit.               1,244,854.000     1,244,854.000     1,212,105.000
================================================================================
Note:                                                *p<0.1; **p<0.05; ***p<0.01
> #AIC and BIC
> AIC(model_1,model_2,model_3)
         df      AIC
model_1 16 1244674
model_2 16 1244674
model_3 16 1211925
> BIC(model_1,model_2,model_3)
         df      BIC
model_1 16 1244854
model_2 16 1244854
model_3 16 1212105
```

Model 4:

	Estimate	Std. Error	t value
(Intercept)	*2.8320642*	*0.2874573*	*9.852*
FEATURE1	*0.4242979*	*0.0249972*	*16.974*
DISPLAY1	*0.8203795*	*0.0084502*	*97.084*
TPR_ONLY1	*-0.0508115*	*0.0077062*	*-6.594*
CATEGORYCOLD CEREAL	*0.8950238*	*0.0044201*	*202.492*
CATEGORYFROZEN PIZZA	*-0.0132293*	*0.0058533*	*-2.260*
CATEGORYORAL HYGIENE PRODUCTS	*-1.1113824*	*0.0053423*	*-208.035*
SEGMENTUPSCALE	*0.2722969*	*0.0108229*	*25.159*
SEGMENTVALUE	*-0.9127170*	*0.0764334*	*-11.941*
log(PRICE)	*0.1512678*	*0.0031077*	*48.675*
STATEKY	*0.0432271*	*0.3219170*	*0.134*
STATEOH	*0.0910191*	*0.2926972*	*0.311*
STATETX	*-0.2223107*	*0.2915137*	*-0.763*
PRODUCT_SIZE	*0.0176901*	*0.0001596*	*110.871*
FEATURE1:DISPLAY1	*-0.2284370*	*0.0359864*	*-6.348*
FEATURE1:CATEGORYCOLD CEREAL	*0.2244392*	*0.0280990*	*7.987*
FEATURE1:CATEGORYFROZEN PIZZA	*0.4619526*	*0.0276748*	*16.692*
FEATURE1:CATEGORYORAL HYGIENE PRODUCTS	*0.3099981*	*0.0293497*	*10.562*
DISPLAY1:CATEGORYCOLD CEREAL	*-0.2013732*	*0.0164387*	*-12.250*
DISPLAY1:CATEGORYFROZEN PIZZA	*-0.0126924*	*0.0152824*	*-0.831*
DISPLAY1:CATEGORYORAL HYGIENE PRODUCTS	*-0.3710475*	*0.0140813*	*-26.350*
TPR_ONLY1:CATEGORYCOLD CEREAL	*0.1670582*	*0.0109278*	*15.287*
TPR_ONLY1:CATEGORYFROZEN PIZZA	*0.2553410*	*0.0132957*	*19.205*
TPR_ONLY1:CATEGORYORAL HYGIENE PRODUCTS	*0.1607554*	*0.0108085*	*14.873*
FEATURE1:SEGMENTUPSCALE	*-0.1508489*	*0.0465504*	*-3.241*
FEATURE1:SEGMENTVALUE	*-0.0572373*	*0.0448067*	*-1.277*
DISPLAY1:SEGMENTUPSCALE	*-0.2505182*	*0.0159594*	*-15.697*
DISPLAY1:SEGMENTVALUE	*0.1646993*	*0.0192408*	*8.560*
TPR_ONLY1:SEGMENTUPSCALE	*-0.1310856*	*0.0150132*	*-8.731*
TPR_ONLY1:SEGMENTVALUE	*0.0368579*	*0.0142435*	*2.588*
CATEGORYCOLD CEREAL:SEGMENTUPSCALE	*-0.3197020*	*0.0081391*	*-39.280*
CATEGORYFROZEN PIZZA:SEGMENTUPSCALE	*-0.4083789*	*0.0090116*	*-45.317*
CATEGORYORAL HYGIENE PRODUCTS:SEGMENTUPSCALE	*-0.1852934*	*0.0088859*	*-20.852*
CATEGORYCOLD CEREAL:SEGMENTVALUE	*0.6587827*	*0.0082415*	*79.935*
CATEGORYFROZEN PIZZA:SEGMENTVALUE	*0.7106940*	*0.0095012*	*74.800*
CATEGORYORAL HYGIENE PRODUCTS:SEGMENTVALUE	*0.6019315*	*0.0094930*	*63.408*
FEATURE1:DISPLAY1:CATEGORYCOLD CEREAL	*0.2594207*	*0.0418814*	*6.194*

FEATURE1:DISPLAY1:CATEGORYFROZEN PIZZA	-0.1612573	0.0409514	-3.938
FEATURE1:DISPLAY1:CATEGORYORAL HYGIENE PRODUCTS	0.0659829	0.0463597	1.423
FEATURE1:DISPLAY1:SEGMENTUPSCALE	0.3001740	0.0671364	4.471
FEATURE1:DISPLAY1:SEGMENTVALUE	0.0270228	0.0710737	0.380
FEATURE1:CATEGORYCOLD CEREAL:SEGMENTUPSCALE	-0.0310480	0.0522853	-0.594
FEATURE1:CATEGORYFROZEN PIZZA:SEGMENTUPSCALE	0.0109125	0.0517877	0.211
FEATURE1:CATEGORYORAL HYGIENE PRODUCTS:SEGMENTUPSCALE	0.0543988	0.0548406	0.992
FEATURE1:CATEGORYCOLD CEREAL:SEGMENTVALUE	-0.0586642	0.0501416	-1.170
FEATURE1:CATEGORYFROZEN PIZZA:SEGMENTVALUE	-0.1162296	0.0493532	-2.355
FEATURE1:CATEGORYORAL HYGIENE PRODUCTS:SEGMENTVALUE	-0.0263169	0.0531626	-0.495
DISPLAY1:CATEGORYCOLD CEREAL:SEGMENTUPSCALE	0.1307704	0.0321627	4.066
DISPLAY1:CATEGORYFROZEN PIZZA:SEGMENTUPSCALE	0.2242855	0.0305646	7.338
DISPLAY1:CATEGORYORAL HYGIENE PRODUCTS:SEGMENTUPSCALE	0.3656792	0.0261872	13.964
DISPLAY1:CATEGORYCOLD CEREAL:SEGMENTVALUE	-0.2130118	0.0311223	-6.844
DISPLAY1:CATEGORYFROZEN PIZZA:SEGMENTVALUE	0.0139705	0.0329862	0.424
DISPLAY1:CATEGORYORAL HYGIENE PRODUCTS:SEGMENTVALUE	-0.1976299	0.0306420	-6.450
TPR_ONLY1:CATEGORYCOLD CEREAL:SEGMENTUPSCALE	0.0993863	0.0213163	4.662
TPR_ONLY1:CATEGORYFROZEN PIZZA:SEGMENTUPSCALE	0.0559472	0.0251935	2.221
TPR_ONLY1:CATEGORYORAL HYGIENE PRODUCTS:SEGMENTUPSCALE	0.1335876	0.0206637	6.465
TPR_ONLY1:CATEGORYCOLD CEREAL:SEGMENTVALUE	-0.2242148	0.0202029	-11.098
TPR_ONLY1:CATEGORYFROZEN PIZZA:SEGMENTVALUE	0.0102794	0.0246879	0.416
TPR_ONLY1:CATEGORYORAL HYGIENE PRODUCTS:SEGMENTVALUE	-0.0468379	0.0202731	-2.310
FEATURE1:DISPLAY1:CATEGORYCOLD CEREAL:SEGMENTUPSCALE	-0.1944257	0.0787308	-2.470
FEATURE1:DISPLAY1:CATEGORYFROZEN PIZZA:SEGMENTUPSCALE	-0.1883731	0.0772902	-2.437
FEATURE1:DISPLAY1:CATEGORYORAL HYGIENE PRODUCTS:SEGMENTUPSCALE	-0.3187632	0.0856897	-3.720
FEATURE1:DISPLAY1:CATEGORYCOLD CEREAL:SEGMENTVALUE	0.1423157	0.0805755	1.766
FEATURE1:DISPLAY1:CATEGORYFROZEN PIZZA:SEGMENTVALUE	-0.0929021	0.0807460	-1.151
FEATURE1:DISPLAY1:CATEGORYORAL HYGIENE PRODUCTS:SEGMENTVALUE	0.0509744	0.0936429	0.544

Interpretation

1. Product display, being featured in the in-store circular, and temporary price reduction.

Predictor	Product Display	In store Circular	Temporary Price Reduction
Spend	If the product was a part of in-store promotional display, then the log spend would be 66.4% more	If the product was in store circular, then the log spend would be 60.9% more	If the prices of product are reduced temporary, then 5% more log spending would happen
Unit Sales	If the product was a part of in-store promotional display, then log(unit) would be 66.4% more	If the product was in store circular log unit sold would be 60.9% more	If the prices of product are reduced temporary, then 5% more log unit would happen
HHS	If in store promotion happen then there is a 5.6% chance of increase in log of no. of purchasing households	If the product was in store circular, then there is a 5.6% chance of increase in log of no. of purchasing households	If the product was price reduction on products, then there is a 1.6% chance of increase in log of no. of purchasing households

2. the effects of display, feature, and TPR on SPEND vary by product categories (cold cereals, frozen pizza, bag snacks) and store segments

Product Category	Product Display	In store Circular (Feature)	Temporary Price Reduction
Cold Cereals	With bag snacks as the base, if the product is displayed the chances of sales decline by 20.13%	With bag snacks as the base, if there is an in-store feature then sales grow by 22.44%	With bag snacks as the base, if TPA happens then 16.70% growth would be there
Frozen pizza	With bag snacks as the base, if the product is displayed the chances of sales decline by 1.2%	With bag snacks as the base, if there is an in-store feature then sales grow by 46.19%	With bag snacks as the base, if TPA happens then 25.53% growth would be there
Hygiene Products	With bag snacks as the base, if the product is displayed the chances of sales decline by 37.1%	With bag snacks as the base, if there is an in-store feature then sales grow by 30.99%	With bag snacks as the base, if TPA happens then 16.70% growth would be there

Store Segment	Product Display	In store Circular (Feature)	Temporary Price Reduction
Upscale	With mainstream as base, there would be a 25.05% less sales if the product is displayed	With mainstream as base, there would be a 15.08% less sales if the product is in circular	With mainstream as base, there would be a 13.13% less sales if the product price is reduced
Value	With mainstream as base, there would be 16.46% more sales if the product is displayed	With mainstream as base, there would be a 5.7% less sales if the product is in circular	With mainstream as base, there would be 3.6% more sales if the product price is reduced

3. the five most price elastic and five least price elastic products

```
> print(top_five)
          UPC Price_Elasticity
55 2066200532        -3.785198
32 7218063979        -3.345891
33 7218063983        -3.245785
31 7218063052        -3.014127
25 4116709428        -2.895474
> print("Bottom Five Least Price Elastic Products:")
[1] "Bottom Five Least Price Elastic Products:"
> print(bottom_five)
          UPC Price_Elasticity
27 7027316404       -0.03942242
26 7027316204       -0.03824922
2  1111009497        0.03424793
8  1111085345       -0.02483807
7  1111085319        0.01555133
```

Top 5	Product Name & Volatility	Bottom 5	Product Name & Volatility
2066200532	Own Supreme Pizza	7027316404	Shurgd pretzel sticks
7218063979	Pepperoni Pizza	7027316204	Shurgd mini pretzels
7218063983	4 Cheese Pizza	1111009497	Pl pretzel sticks
7218063052	Brck OVN ITL Pep pz	1111085345	PL raisin bran
4116709428	Mint Fluor RNS	1111085319	PL honey nut toasted oats

4. Products would you lower the price to maximize (a) Spend and (b) unit sales

A) To maximize spend the price of pretzels i.e. 1111009497 should be dropped so that the spend is more.

B) To maximize unit sales UPC 3700019521 i.e CREST PH WHTG toothpaste should be sold.

Recommendations:

Exploring additional features or transform existing ones to capture more complex relationships in the data. Consider interaction terms, polynomial features, or domain-specific transformations to better represent the underlying patterns. Regularization helps control model complexity and prevents extreme parameter estimates, leading to better performance on unseen data.