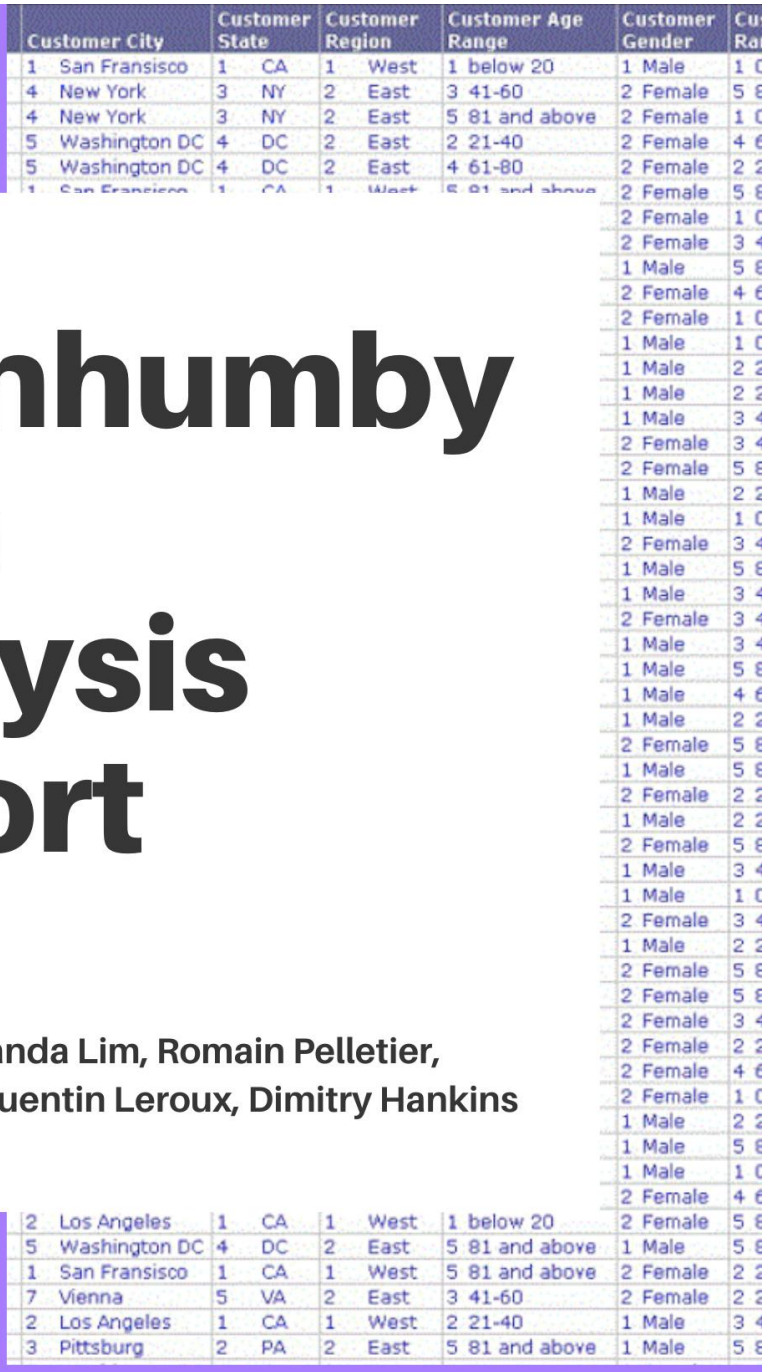


	Customer City	Customer State		Customer Region		Customer Age Range	Customer Gender	Cu Ra
1	San Fransisco	1	CA	1	West	1 below 20	1 Male	1
4	New York	3	NY	2	East	3 41-60	2 Female	5
4	New York	3	NY	2	East	5 81 and above	2 Female	1
5	Washington DC	4	DC	2	East	2 21-40	2 Female	4
5	Washington DC	4	DC	2	East	4 61-80	2 Female	2
1	San Fransisco	1	CA	1	West	5 81 and above	2 Female	5
							2 Female	1
							2 Female	3
							1 Male	5
							2 Female	4
							2 Female	1
							1 Male	1
							1 Male	2
							1 Male	2
							1 Male	3
							2 Female	3
							2 Female	5
							1 Male	2
							1 Male	1
							2 Female	3
							1 Male	5
							1 Male	3
							2 Female	3
							1 Male	3
							1 Male	5
							1 Male	4
							1 Male	2
							2 Female	5
							1 Male	5
							2 Female	2
							1 Male	2
							2 Female	5
							1 Male	3
							1 Male	1
							2 Female	3
							1 Male	2
							2 Female	5
							2 Female	5
							2 Female	3
							2 Female	2
							2 Female	4
							2 Female	1
							1 Male	2
							1 Male	5
							1 Male	1
							2 Female	4
2	Los Angeles	1	CA	1	West	1 below 20	2 Female	5
5	Washington DC	4	DC	2	East	5 81 and above	1 Male	5
1	San Fransisco	1	CA	1	West	5 81 and above	2 Female	2
7	Vienna	5	VA	2	East	3 41-60	2 Female	2
2	Los Angeles	1	CA	1	West	2 21-40	1 Male	3
3	Pittsburg	2	PA	2	East	5 81 and above	1 Male	5

Dunnhumby Data Analysis Report

FEBRUARY 2020

Colm Tobin, Amanda Lim, Romain Pelletier,
Bhavya Gupta, Quentin Leroux, Dimitry Hankins

Table of Contents

1. Introduction

1.1 Context & The Dataset

The data we have analyzed in this report comes from Dunnhumby, from one of their supermarket clients. The data provides analysis of sales of 4 product categories (cold cereal, bag snacks, frozen pizza and oral hygiene products) from locations in 4 states across the USA.

1.2 Research Objectives

This report addresses:

- What are the price elasticities of each product? How do they differ by geography?
- What is the impact of promotions on units per visits? How do they differ by geography?
- The report also addresses demand forecasting.

1.3 Summary of Managerial Insights

We have come to various key managerial insights for both the retailer and the producers of the individual goods:

1.3.1 Price Elasticity

- The product categories from most elastic to least elastic are: frozen pizza, cold cereal, oral hygiene products, bag snacks.
- Understanding price elasticities for specific products, as well as how they perform geographically, can inform managers' decisions when considering their pricing strategy, as they have a better understanding of how consumers will react.
- The states from most elastic to least elastic are Indiana, Kentucky, Ohio, Texas.

1.3.2 Impact of Promotions on Units/Visits

- Reduction ratio had by far the most significant on the number of units/visits, with an overall correlation of 0.2; compared to 0.03 for both feature and TPR Only.
- Thus, the size of the reduction is important for managers to drive additional multi-unit purchases from customers. Furthermore, display was particularly ineffective for cold cereal (correlation of -1/2), and feature was only positive with oral hygiene products, and this could form the basis of managers' promotional strategies for each product category.

1.3.3 Demand Forecasting

- According to our regression tree, the factor that influence the most the number of units sold is display, whereas the linear regression suggests that it has a minor impact in determining units sold (possible sampling error).
- Our neural network was by far the most effective method in predicting demand

2. Price Elasticity

Price Elasticity	Elastic, Inelastic, or Unit Elastic?	What does it mean?
<1	Inelastic	The quantity bought for this good is less sensitive to changes in price. A small change in price leads to a smaller change in quantity.
=1	Unit Elastic	The quantity bought for this good is exactly proportional to changes in price.
>1	Elastic	The quantity bought for this good is more sensitive to changes in price. A small change in price leads to a larger change in quantity.

Figure 1: Price Elasticity Definition

2.1 Data Visualization

2.1.1 Elasticities by Product

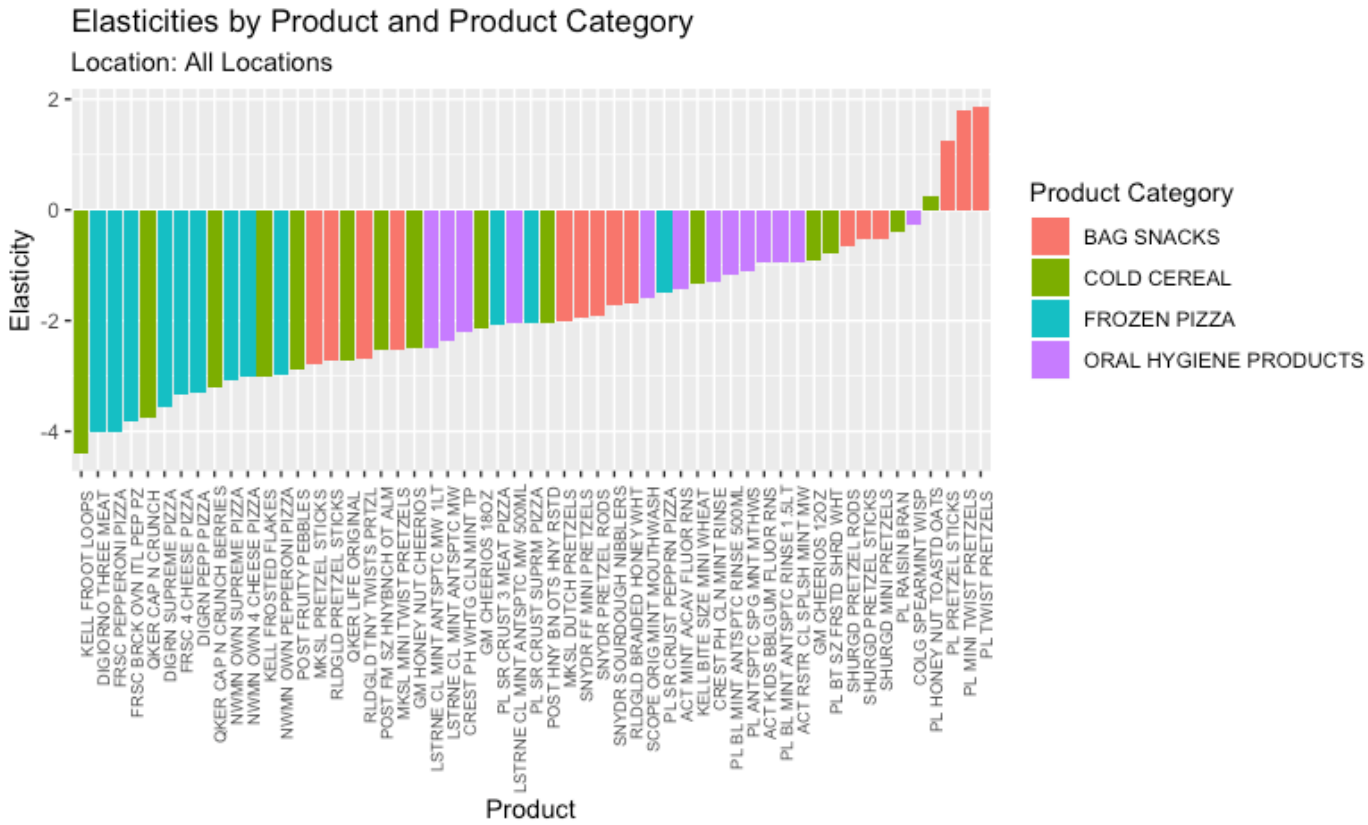


Figure 2

This illustrates the elasticities of all products in all locations, sorted from most inelastic to most elastic. Bars that are below the x axis, meaning their elasticity is less than 1, are inelastic. Bars that are above the x axis are elastic.

2.1.2 Elasticities by Product Category

We can break down this overview graph into four separate graphs – one for each product category.

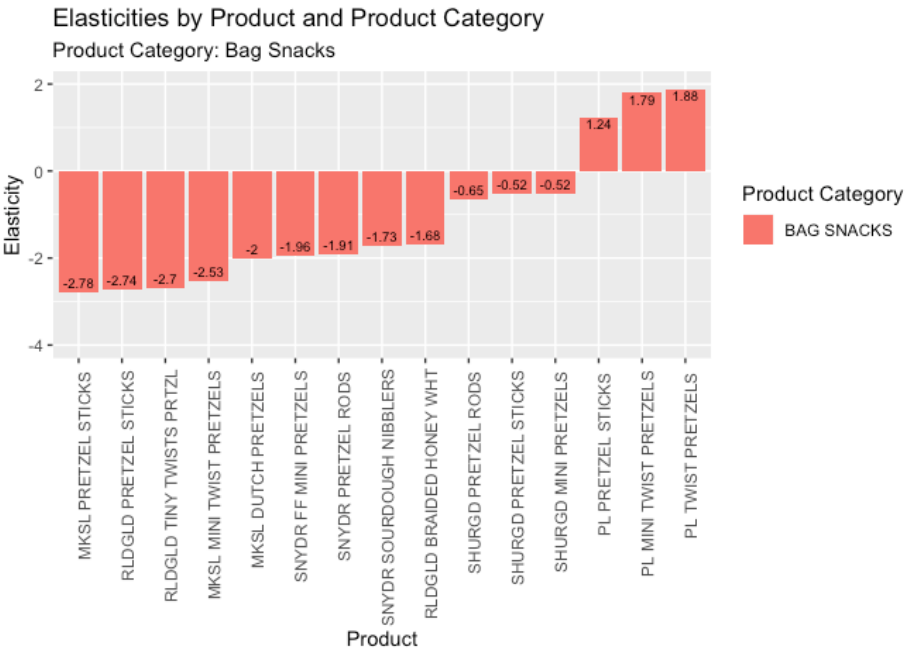


Figure 3.1

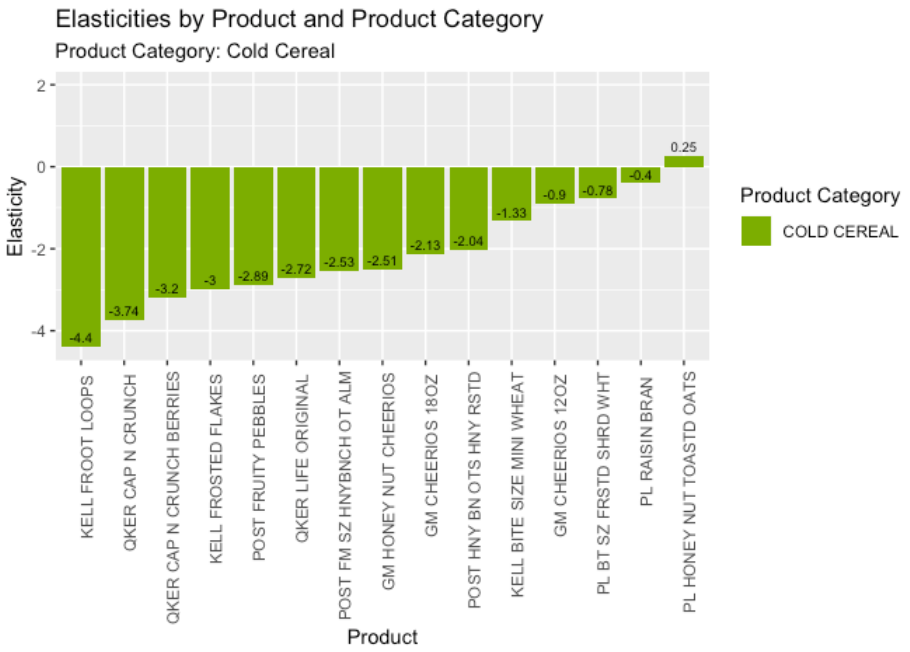


Figure 3.2

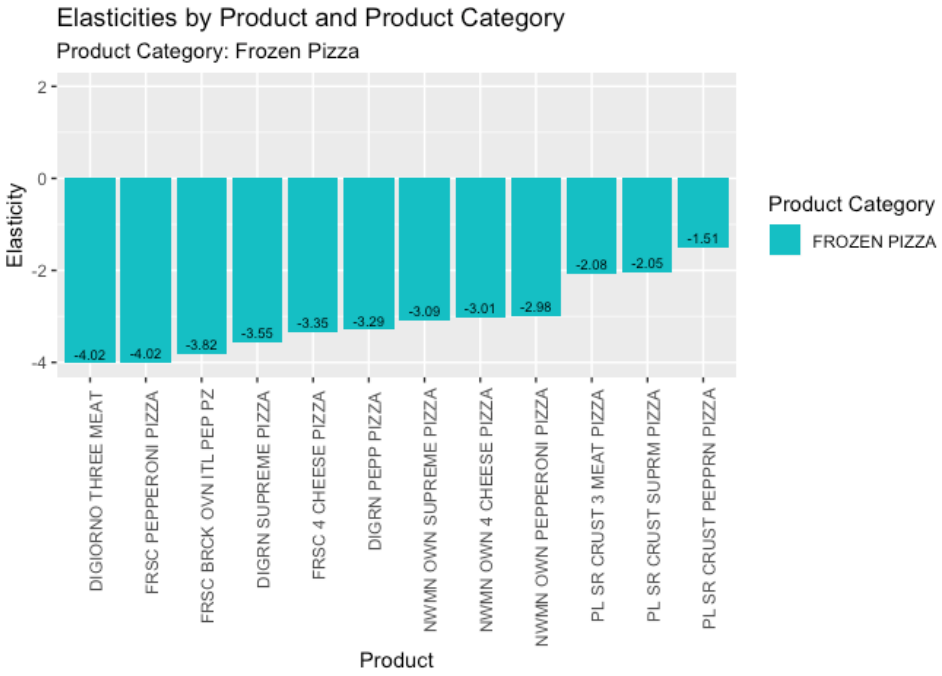


Figure 3.3

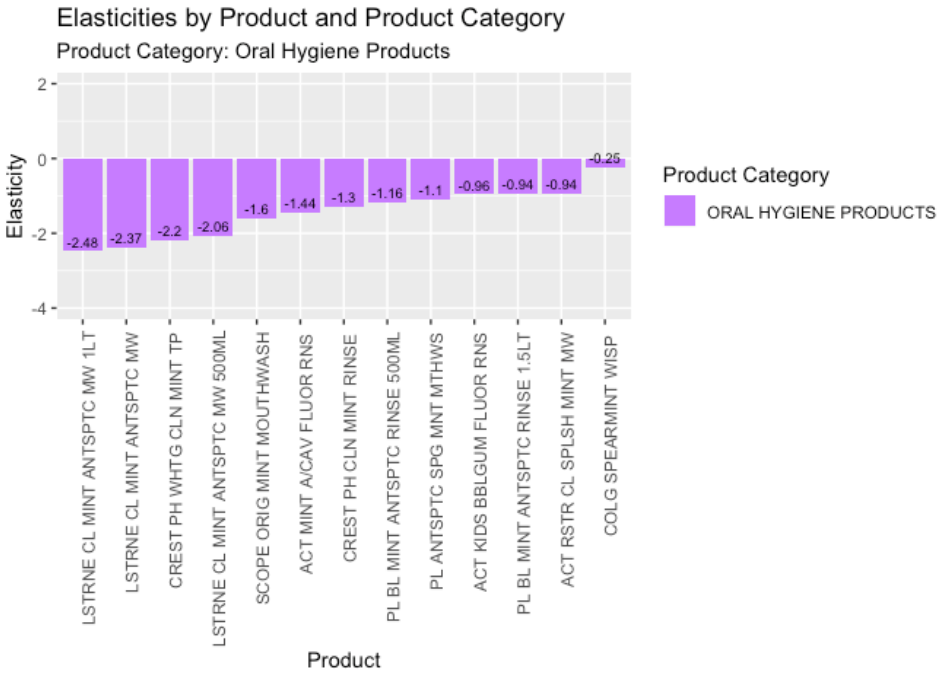


Figure 3.4

The mean elasticities of each product category are as follows:

Product Category	Mean Elasticity
Bag Snacks	-1.120785
Cold Cereal	-2.154896
Frozen Pizza	-3.064944
Oral Hygiene Products	-1.446919

Figure 4

2.1.2 Elasticities by Location

Below are the elasticities of each product filtered by state. Note that not all products are sold in all states (i.e. Texas and Kentucky), therefore not all products are shown some graphs.

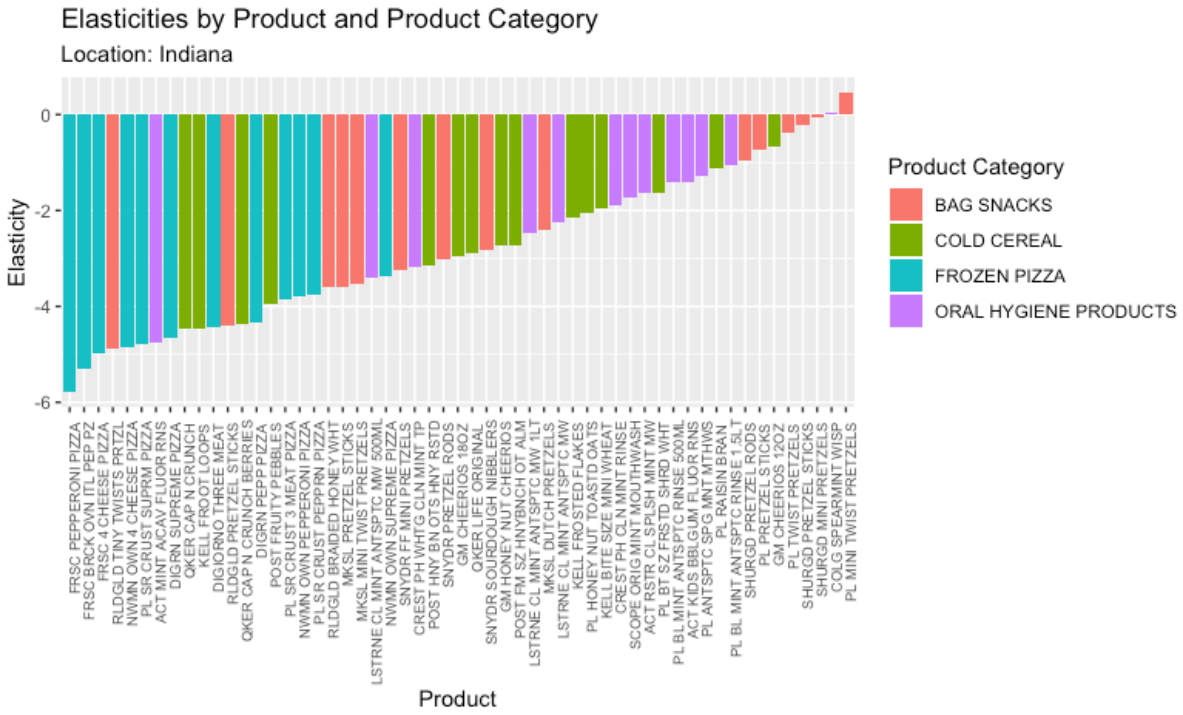


Figure 4.3

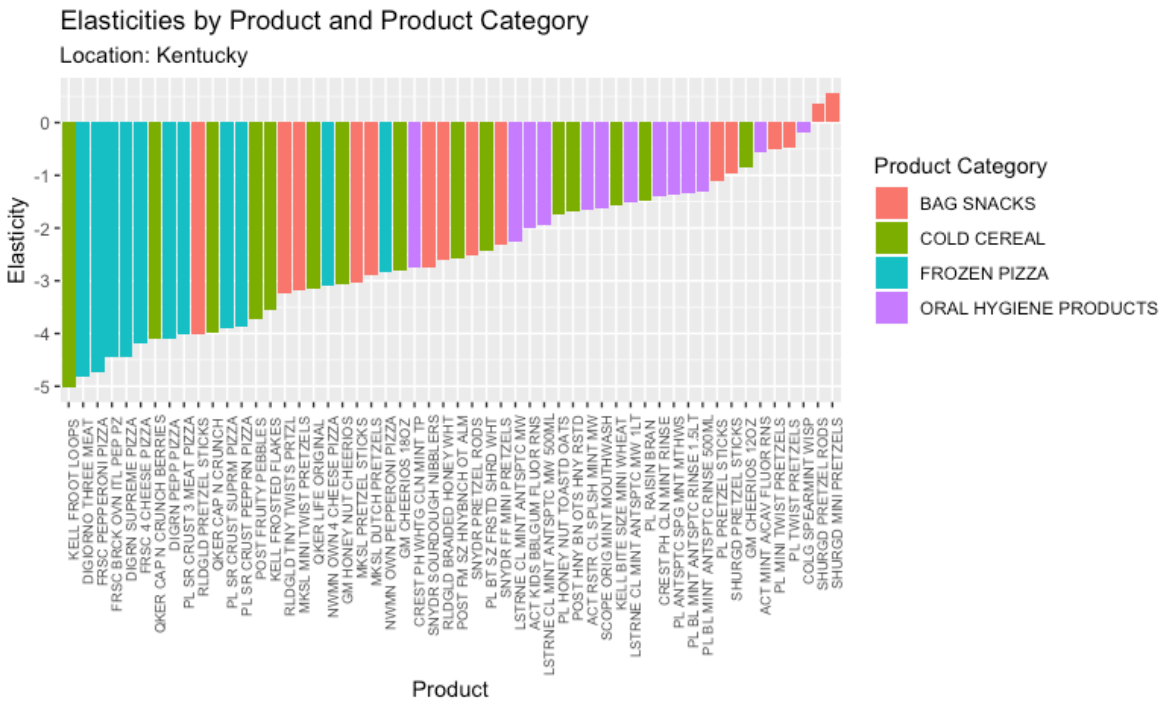


Figure 4.4

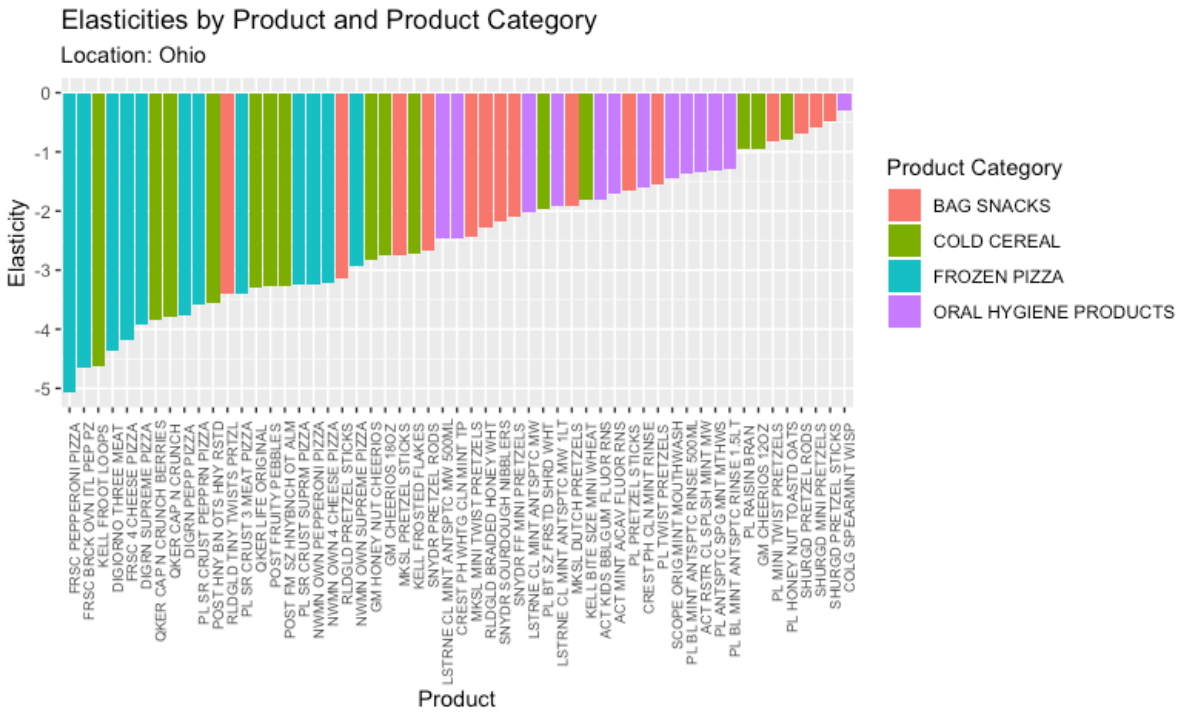


Figure 4.5

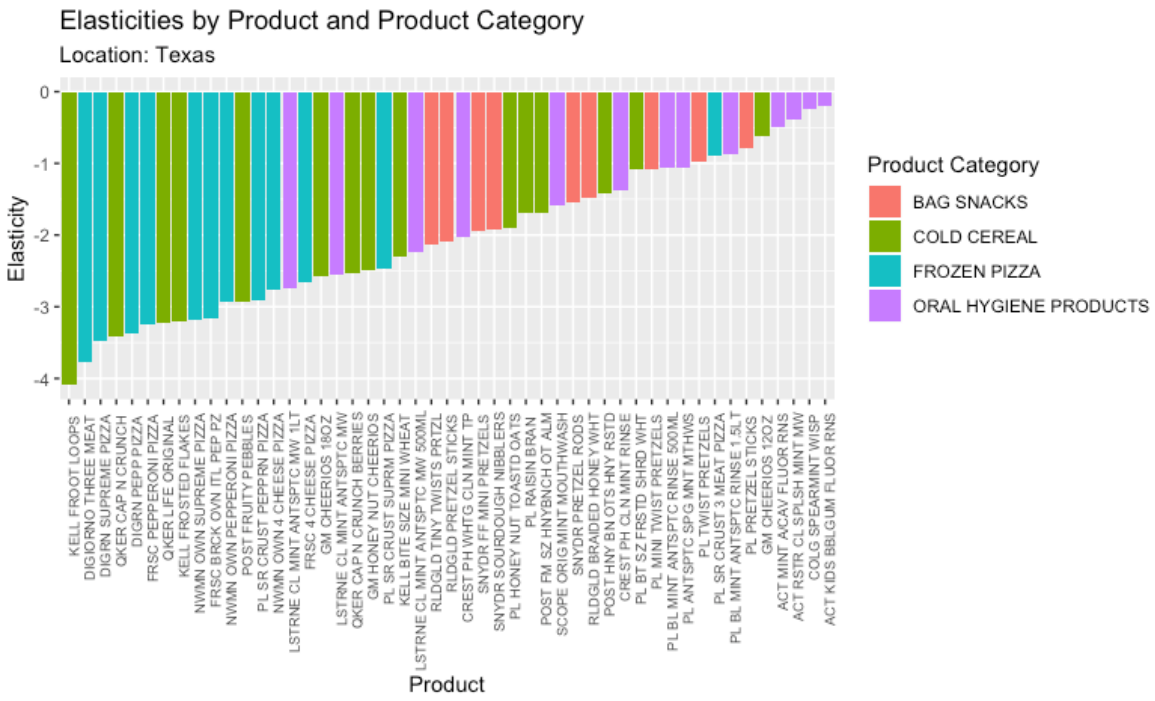


Figure 4.6

2.2.3 Elasticities by Product Category by Location

We grouped the products into their product categories and found their mean elasticities per state.

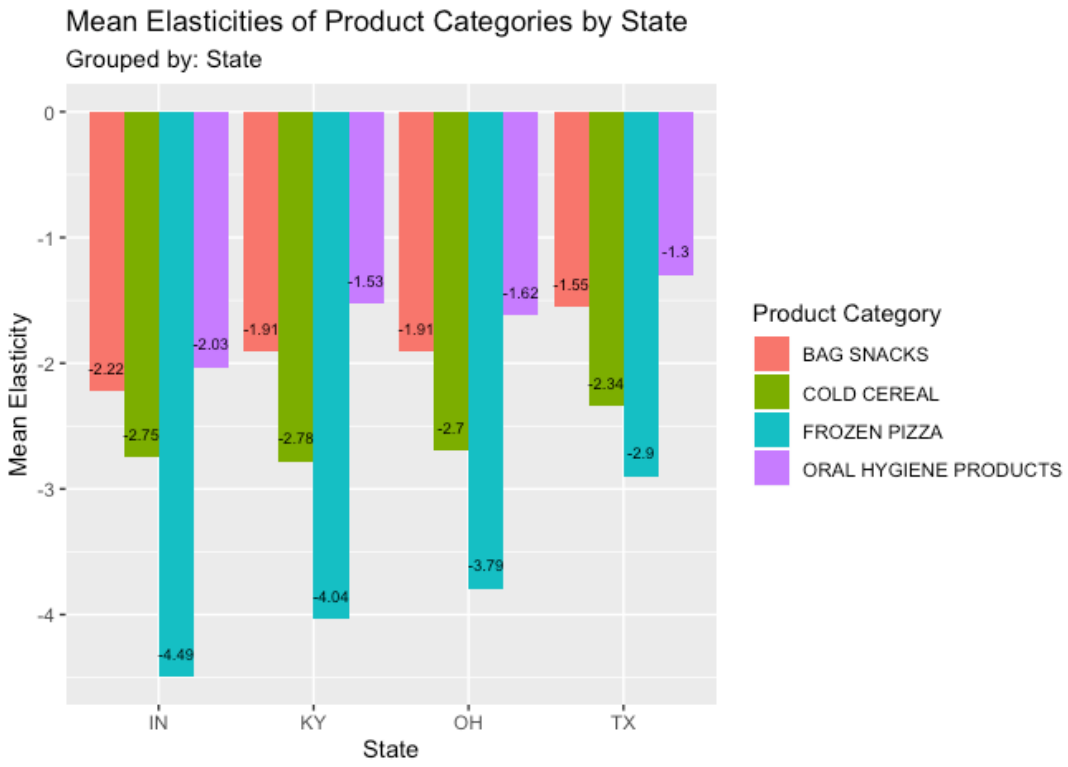


Figure 5

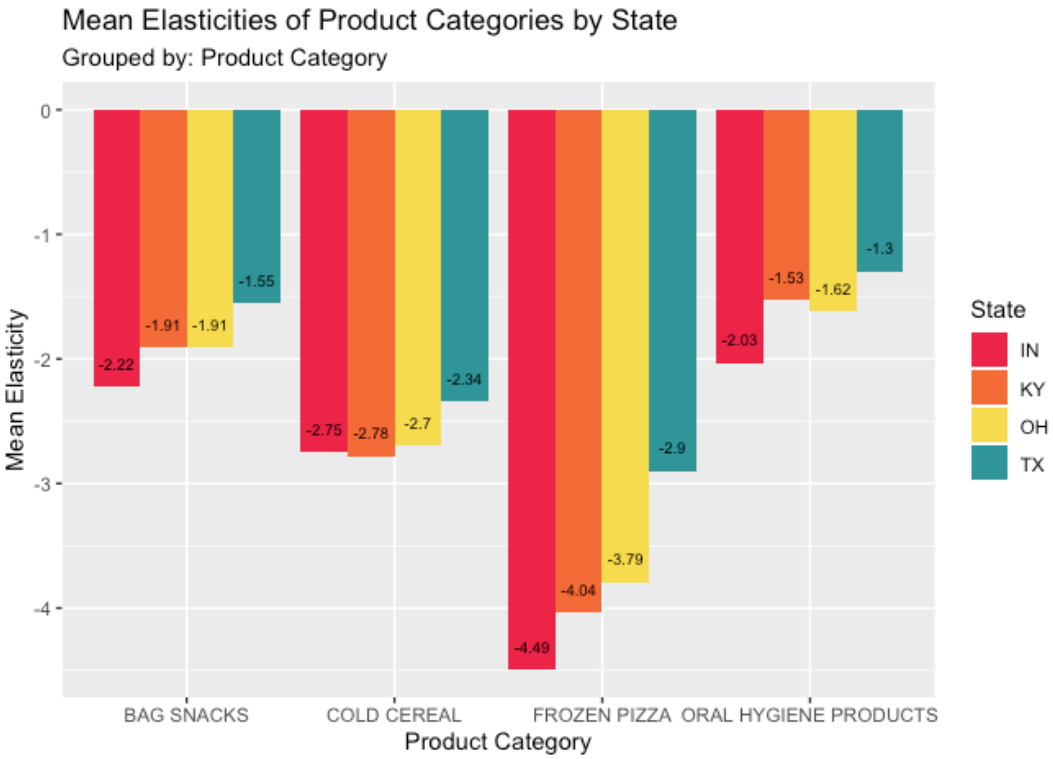


Figure 6

The mean elasticities of each product category by state is as follows:

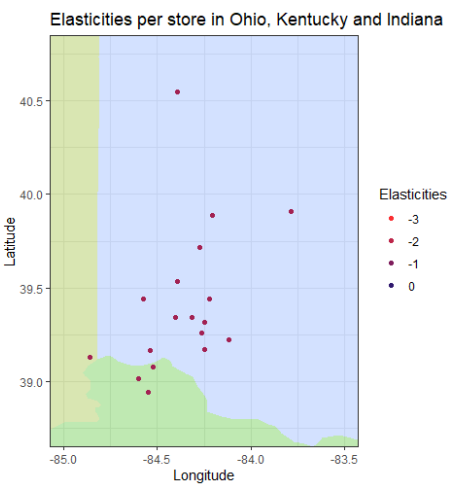
	Mean Elasticity			
Product Category	**IN**	**KY**	**OH**	**TX**
Bag Snacks	-2.224401	-1.909096	-1.909061	-1.553162
Cold Cereal	-2.747900	-2.783242	-2.695273	-2.342732
Frozen Pizza	-4.491630	-4.036041	-3.794761	-2.900059
Oral Hygiene Products	-2.034227	-1.526801	-1.619154	-1.297211
Mean**	**-2.816891**	**-2.493147**	**-2.466384**	**-2.056814**

***Calculated by taking the mean of all elasticities in that state (rather than the mean of the mean of the product categories)*

Figure 7

2.1.3 Elasticities by Store by Location

We visualized elasticities by store on maps:



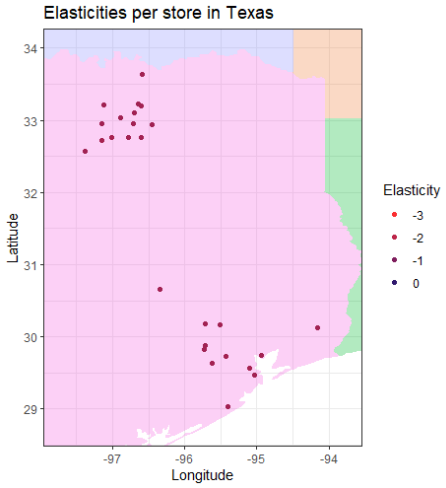


Figure 8.1 & 8.2

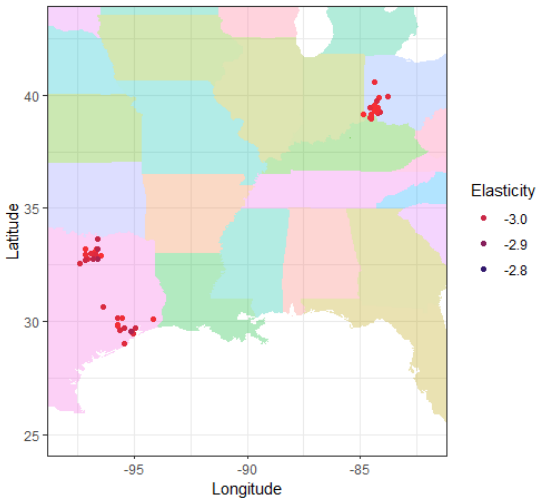


Figure 8.3 All states overview

2.2 Modeling Results

2.2.1 Description of Models

We performed a linear regression model regressing the log of units by the log of price. This is equivalent to taking the percent change of units divided by the percent change in price, which is the formula for elasticity. We calculated this for each of the 55 UPCs (unique product codes) getting the elasticity coefficient for each. We used the results of this model to create the visualisations of elasticities by product, product category, and location. Below shows one iteration of the multiple linear regression for **PL MINI TWIST PRETZELS** showing that its elasticity is 1.7947.

```
Call:
lm(formula = log(UNITS) ~ log(PRICE), data = plpretz)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1744 -0.4338  0.0895  0.5696  3.4718

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.2327     0.0144  224.46   <2e-16 ***
log(PRICE)    1.7947     0.0484   37.08   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8539 on 11859 degrees of freedom
Multiple R-squared:  0.1039,    Adjusted R-squared:  0.1038
F-statistic:  1375 on 1 and 11859 DF,  p-value: < 2.2e-16
```

Figure 9

2.3 Insights & Managerial Implications

2.3.1 Elasticities by Product

In *Figure 2*, four products are elastic. While elastic goods rarely exist in the real world, we can infer this positive elasticity coefficient (reduction in price leads to a reduction in units bought) is due to sampling error.

2.3.2 Elasticities by Product Category

Figure 4 shows that the product categories from most elastic to least elastic are frozen pizza, cold cereal, oral hygiene products, bag snacks. The specific elasticities are shown in *Figure 3.1 to 3.4.* In other words, for frozen pizza, the same percent change in price leads to a greater change in units sold than for the other categories, and vice versa. This finding suggests that The Retailer could expect a greater increase in the units bought of products in the bag snacks and oral hygiene categories if they chose to reduce the price than cold cereal and frozen pizza.

2.3.3 Elasticities by Location

In *Figures 4.1 to 4.4,* the order of products changes, however, frozen pizza products remain consistently towards the left (most elastic).

2.3.3 Elasticities by Product Category by Location

In *Figure 7,* elasticity differs per state but the order of product categories from most elastic to least elastic remains the same: frozen pizza, cold cereal, bag snacks, oral hygiene products. There is some sampling error (see appendix for further explanation).

We can also see that the states from most inelastic to least inelastic are Indiana, Kentucky, Ohio, Texas. This means that consumers in Indiana are the most sensitive to changes in price of the products sold at The Retailer. When deciding on promotions, The Retailer could

anticipate a greater change in units sold in Indiana given a constant reduction in price across states.

2.3.3 Elasticities by Store Location

Visualization by individuals store location further supports our findings in 2.2.3.

Figure 8.3 shows that the stores are concentrated between 3 areas. When we take stores individually, most of their elasticities are between -2 and -3. Thus, managers of individual stores have a better understanding of how price changes could impact their sales for specific products.

3. Impact of Promotion on Units/Visits

3.1 Data Visualization

We visualized the impact on units/visits based on each type of promotion: display, reduction ratio, temporary price reduction, and featuring.

3.1.1 Impact of Promotional Display on Units/Visits

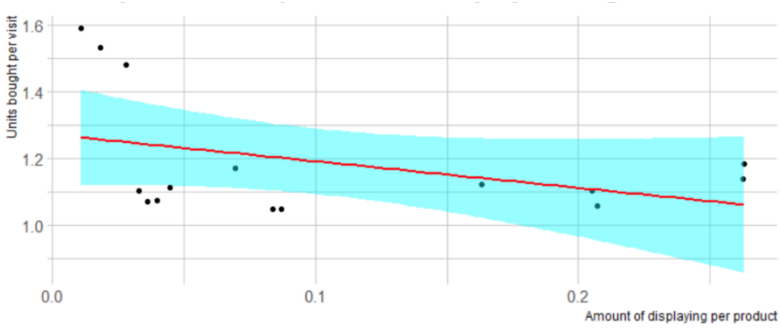


Figure 10.1: Bag Snacks

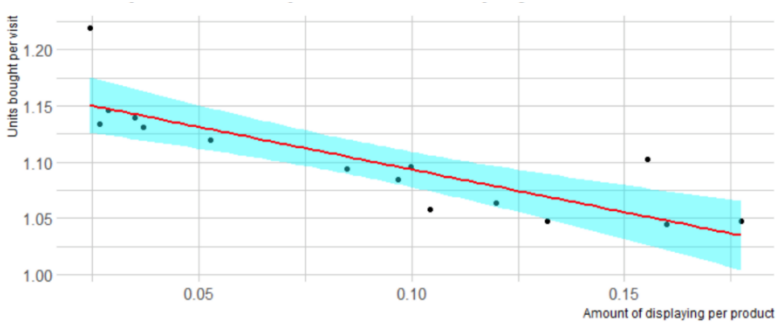


Figure 10.2: Cold Cereal

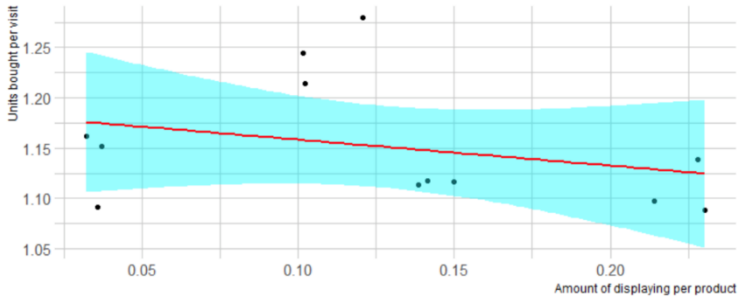


Figure 10.3: Frozen Pizza

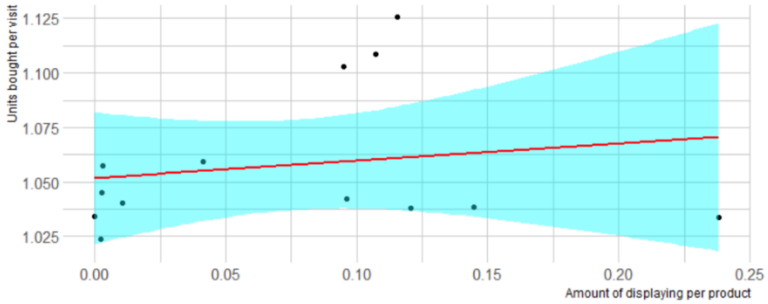


Figure 10.4: Oral Hygiene Products

3.1.2 Impact of Reduction Ratio on Units/Visits

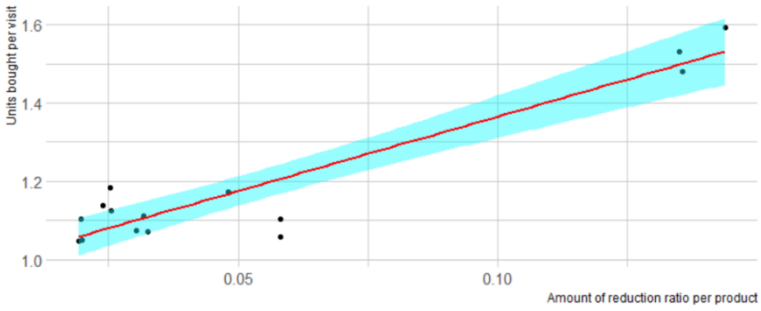


Figure 11.1: Bag Snacks

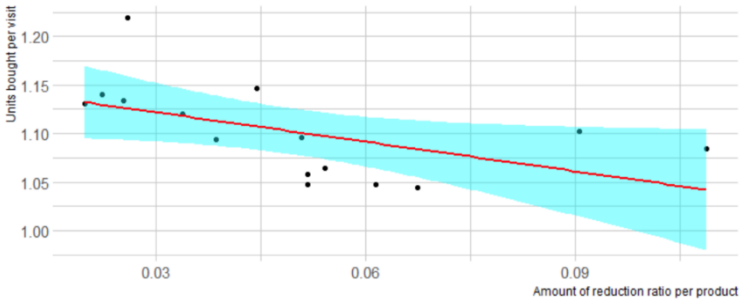


Figure 11.2: Cold Cereal

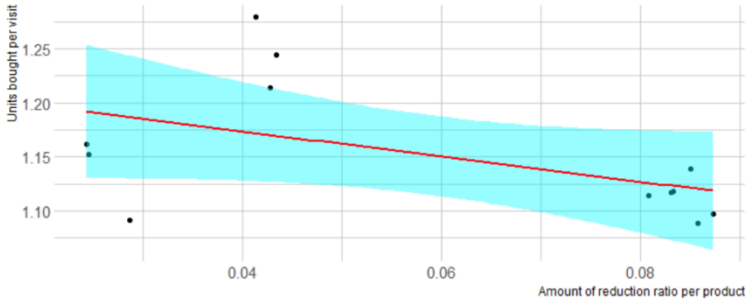


Figure 11.3: Frozen Pizza

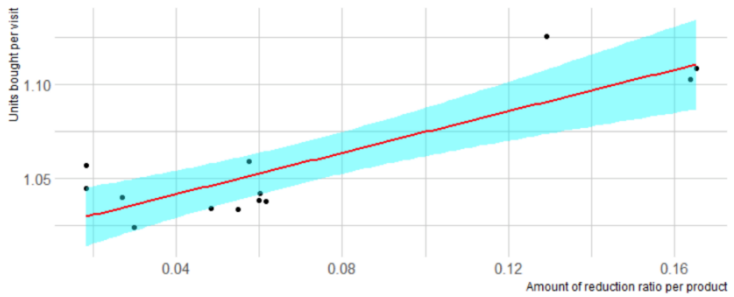


Figure 11.4: Oral Hygiene Products

3.1.3 Impact of Temporary Price Reduction on Units/Visits

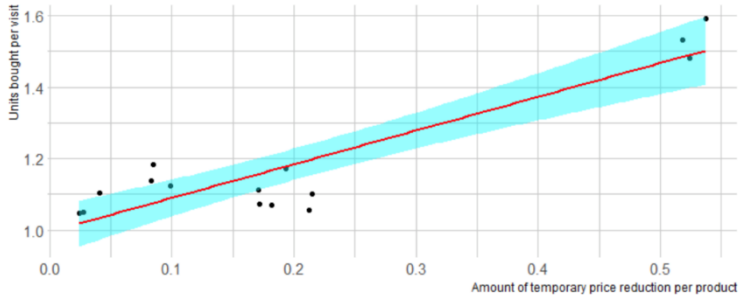


Figure 12.1: Bag Snacks

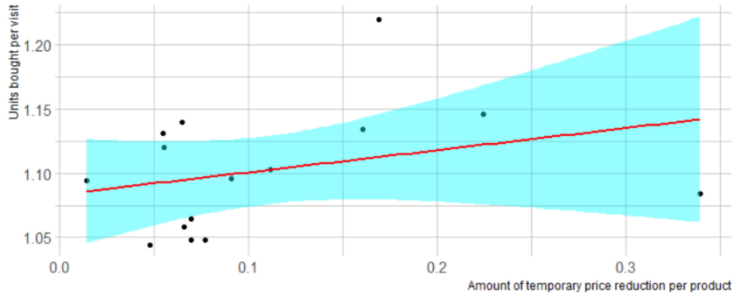


Figure 12.2: Cold Cereal

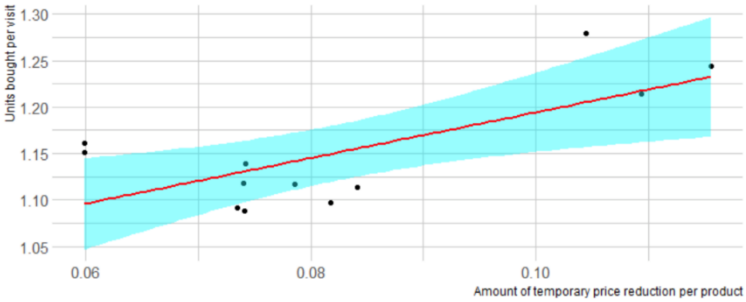


Figure 12.3: Frozen Pizza

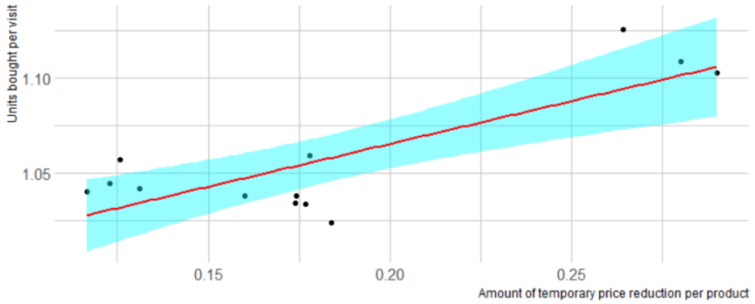


Figure 12.4: Oral Hygiene Products

3.1.4 Impact of Featuring on Units/Visits

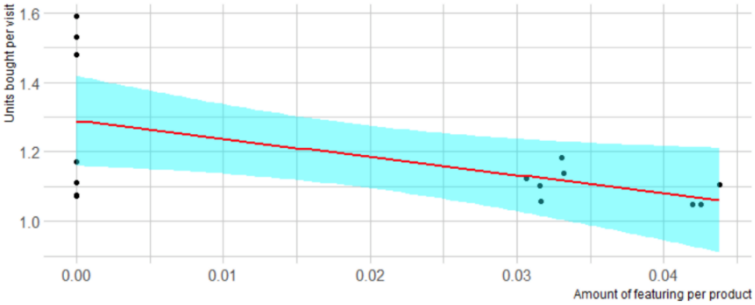


Figure 13.1: Bag Snacks

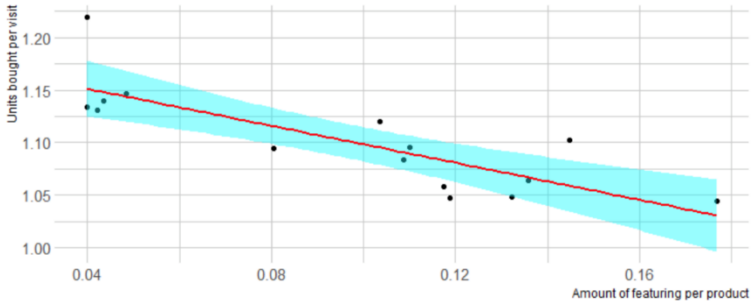


Figure 13.2: Cold Cereal

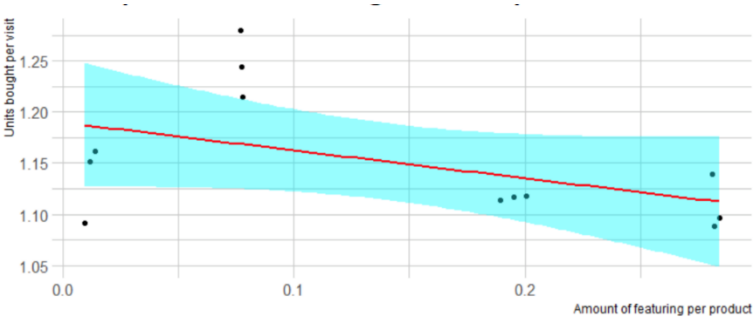


Figure 13.3: Frozen Pizza

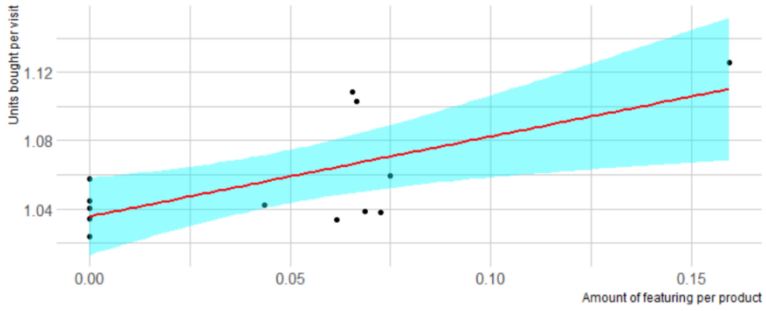


Figure 13.4: Oral Hygiene Products

3.1.5 Units/Visits and individual store Promotions by individual store Location

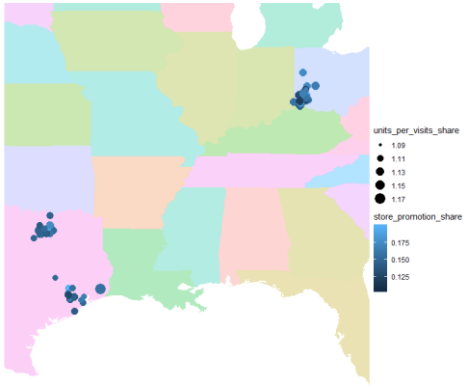


Figure 15.1 All states overview

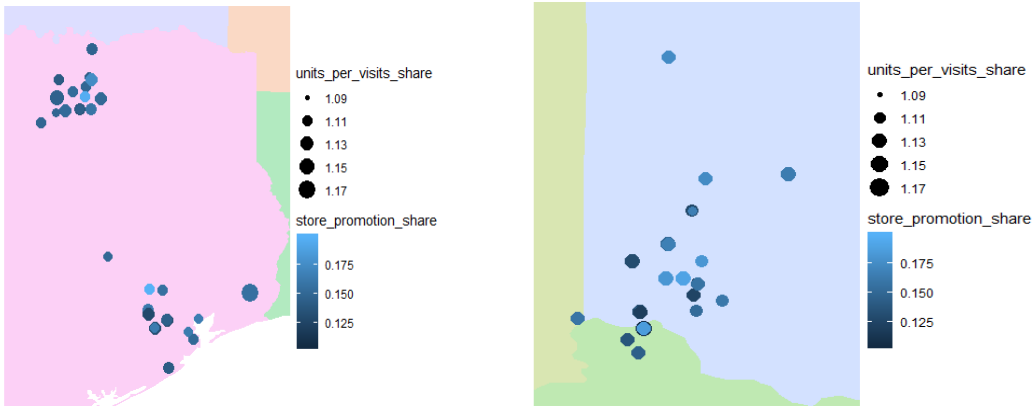


Figure 15.2 (Units/visits and store promotion in Texas)

& Figure 15.3 (Units/visits and store promotion in Ohio, Indiana and Kentucky)

Units_per_visits_share = Average number of units/visit at specific stores

store_promotion_share = The amount of products at that store featuring 1 or more promotion

3.2 Modeling Results

3.2.1 Description of Models

We performed a linear regression model regressing units/visits against the promotional variables seen above.

```
Coefficients:
                  Estimate Std. Error  t value Pr(>|t|)
(Intercept)      1.0992092  0.0003423 3211.280  < 2e-16 ***
FEATURE          0.0327160  0.0014349   22.801  < 2e-16 ***
DISPLAY         -0.0046323  0.0011066   -4.186 2.84e-05 ***
TPR_ONLY         0.0341357  0.0012975   26.310  < 2e-16 ***
REDUCTION_ratio  0.2050786  0.0044627   45.954  < 2e-16 ***
```

3.3 Insights & Managerial Implications

3.3.1 Impact of Promotional Display on Units/Visits

Promotional display only had a positive effect for oral hygiene products, and negative for the other 3 product categories. This could be due to sampling error, or the type of items that were displayed (e.g. if more 'premium' items are displayed, fewer consumers are likely to be able to afford multi-unit purchases).

3.3.2 Impact of Reduction Ratio on Units/Visits

The reduction ratio has a positive impact on the number of units/visits for oral hygiene products and bag snacks (from 1 to 1.5 bag snacks per visits), but negative for frozen pizza and cold cereal. This could be due to more customer loyalty in those categories, meaning customers are less likely to be swayed by price reductions.

3.3.3 Impact of Temporary Reduction on Units/Visits

Temporary Reduction seems to have a significant positive impact on all product types. Thus, managers may want to focus on this promotional strategy further.

3.3.4 Impact of Featuring on Units/Visits

Feature only has a positive impact on oral hygiene products. It can be noted that oral hygiene products as a category had a positive correlation with every promotional method. Thus, promotions in general were much more effective for oral hygiene products than the other categories.

4. Demand Forecasting

4.1 Visualization & Modelling

4.1.1 Linear Regression

We produced a linear regression model predicting “UNITS”, with input variables “PRICE”, “FEATURE” , “DISPLAY”, “TPR_ONLY”, “HHS” and “REDUCTION”.

We then used the model to predict demand from both the test and train data sets, to estimate the accuracy of our modelling.

```
Coefficients:
                Estimate Std. Error  t value Pr(>|t|)
(Intercept)    -1.3868734  0.0216316  -64.113   <2e-16 ***
PRICE           0.1101221  0.0051843   21.241   <2e-16 ***
FEATURE         1.5783079  0.0361849   43.618   <2e-16 ***
DISPLAY        -0.4975532  0.0275047  -18.090   <2e-16 ***
Reduction      -1.6304570  0.0349422  -46.662   <2e-16 ***
TPR_ONLY        0.0233815  0.0316622    0.738     0.46
HHS             1.2134894  0.0003297 3680.738   <2e-16 ***
ReductionRatio  9.4418544  0.1695058   55.702   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.139 on 520792 degrees of freedom
Multiple R-squared:  0.9706,    Adjusted R-squared:  0.9706
F-statistic: 2.46e+06 on 7 and 520792 DF,  p-value: < 2.2e-16
```

‘General Linear Regression Model’

The above general linear regression model suggests that the amount of reduction offered (reduction ratio) has a much larger and positive impact on unit sales than any of the other variables.

We then filtered each data set to only include entries of one particular product category, and repeated this for all categories. This provides added insight such as how effective promotional strategies, such as display, are for specific categories.

(each individual linear regression model is explained in the appendix)

```
                Estimate Std. Error  t value Pr(>|t|)
(Intercept)     2.408182   0.076860   31.332  < 2e-16 ***
PRICE          -1.027673   0.029031  -35.399  < 2e-16 ***
FEATURE        -1.001545   0.126802   -7.898 2.85e-15 ***
DISPLAY        -0.790856   0.060104  -13.158  < 2e-16 ***
TPR_ONLY        1.011402   0.081670   12.384  < 2e-16 ***
HHS             1.168273   0.000951 1228.434  < 2e-16 ***
Reduction      -2.869795   0.268334  -10.695  < 2e-16 ***
REDUCTION_ratio  7.585791   0.787812    9.629  < 2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.189 on 127369 degrees of freedom
Multiple R-squared:  0.9442,    Adjusted R-squared:  0.9442
F-statistic: 3.078e+05 on 7 and 127369 DF,  p-value: < 2.2e-16
```

(Bag snacks linear regression)

```
Coefficients:
                Estimate Std. Error  t value Pr(>|t|)
(Intercept)      2.1372401  0.0259620   82.322  < 2e-16 ***
PRICE           -0.3926991  0.0043194  -90.916  < 2e-16 ***
FEATURE         -0.2155376  0.0237108   -9.090  < 2e-16 ***
DISPLAY         -0.1547273  0.0206115   -7.507  6.1e-14 ***
TPR_ONLY        -0.5684405  0.0274431  -20.713  < 2e-16 ***
HHS              1.1947870  0.0006368 1876.270  < 2e-16 ***
Reduction       -2.1169382  0.0467379  -45.294  < 2e-16 ***
REDUCTION_ratio 13.0789535  0.3246036   40.292  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.848 on 111152 degrees of freedom
Multiple R-squared:  0.9809,    Adjusted R-squared:  0.9809
F-statistic: 8.15e+05 on 7 and 111152 DF,  p-value: < 2.2e-16
```

(Frozen pizza linear regression) linear regression)

```
Coefficients:
                Estimate Std. Error  t value Pr(>|t|)
(Intercept)     -2.4202483  0.0790262  -30.626  < 2e-16 ***
PRICE            0.0528371  0.0260560    2.028 0.042579 *
FEATURE          3.6979940  0.0848564   43.579  < 2e-16 ***
DISPLAY         -0.9108308  0.0807319  -11.282  < 2e-16 ***
TPR_ONLY        -0.2939324  0.0816219   -3.601 0.000317 ***
HHS              1.2285518  0.0005997 2048.610  < 2e-16 ***
Reduction        3.6862316  0.2691045   13.698  < 2e-16 ***
REDUCTION_ratio -6.3948852  0.9823712   -6.510 7.55e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.793 on 169669 degrees of freedom
Multiple R-squared:  0.973,     Adjusted R-squared:  0.973
F-statistic: 8.748e+05 on 7 and 169669 DF,  p-value: < 2.2e-16
```

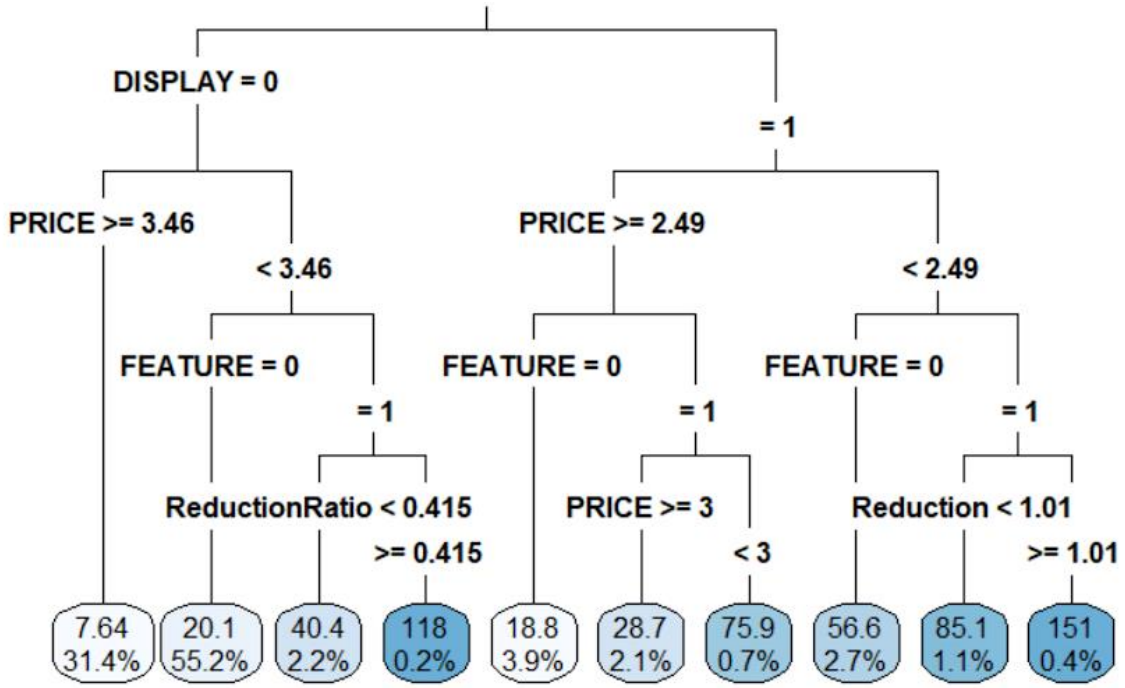
(Cold cereal linear regression)

```
Coefficients:
                Estimate Std. Error  t value Pr(>|t|)
(Intercept)     -0.3607604  0.0123030  -29.323  < 2e-16 ***
PRICE            0.0158777  0.0027181    5.842 5.18e-09 ***
FEATURE          0.5065808  0.0170751   29.668  < 2e-16 ***
DISPLAY         -0.1660412  0.0126498  -13.126  < 2e-16 ***
TPR_ONLY        -0.3113896  0.0134174  -23.208  < 2e-16 ***
HHS              1.1263410  0.0008291 1358.555  < 2e-16 ***
Reduction       -0.6582530  0.0156480  -42.066  < 2e-16 ***
REDUCTION_ratio  4.8104559  0.0596480   80.647  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.03 on 116463 degrees of freedom
Multiple R-squared:  0.9535,    Adjusted R-squared:  0.9535
F-statistic: 3.412e+05 on 7 and 116463 DF,  p-value: < 2.2e-16
```

(Oral hygiene linear regression)

4.1.2 Regression Tree

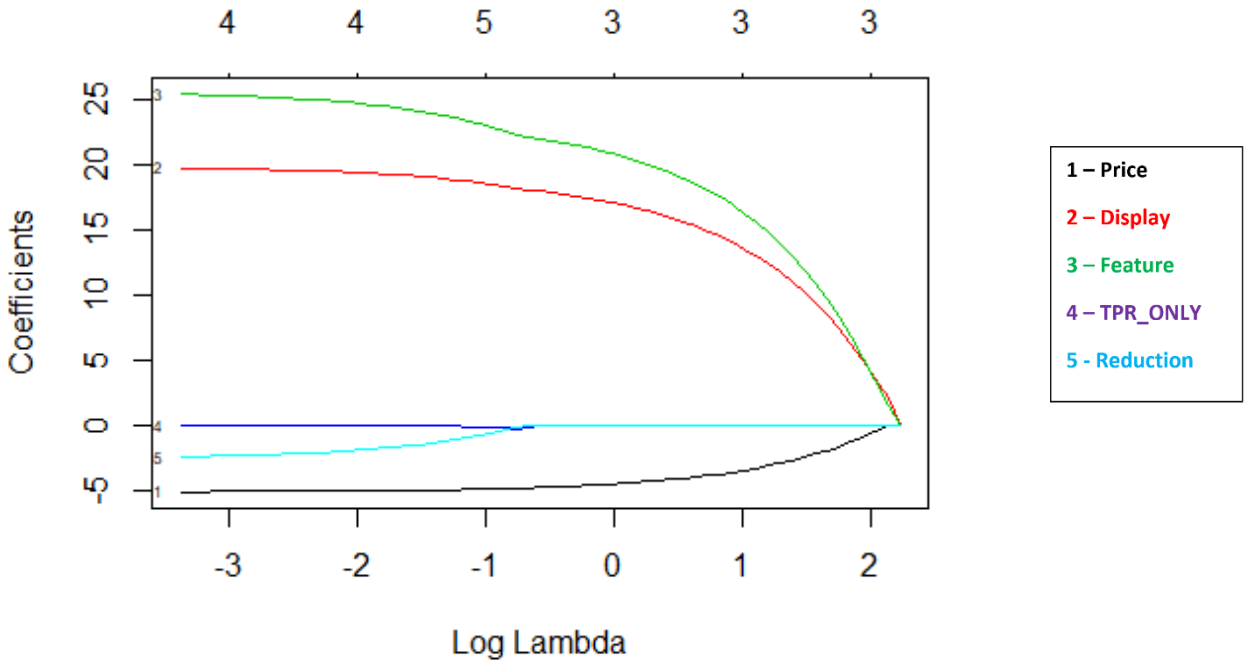


From *the regression tree,* display had the most important impact in determining the number of units sold. This is because the value that appears as the root node is the most important when using regression trees. It can be seen from the data that the majority of observations (89.2%) appear underneath the no display category, but a far greater proportion of the large unit sales appear under the display node. For example, the result of 118 units with a 0.2% chance appears under the no display node, but under the display node there is a 75.9 total with 0.7% chance, an 85.1 with a 1.1% chance and a 151 total with a 0.4% chance. Therefore, more of the very-high performing sales figures (2.2% vs 0.2%) appeared under the display node, and as such display likely had a positive impact on selling greater numbers of units.

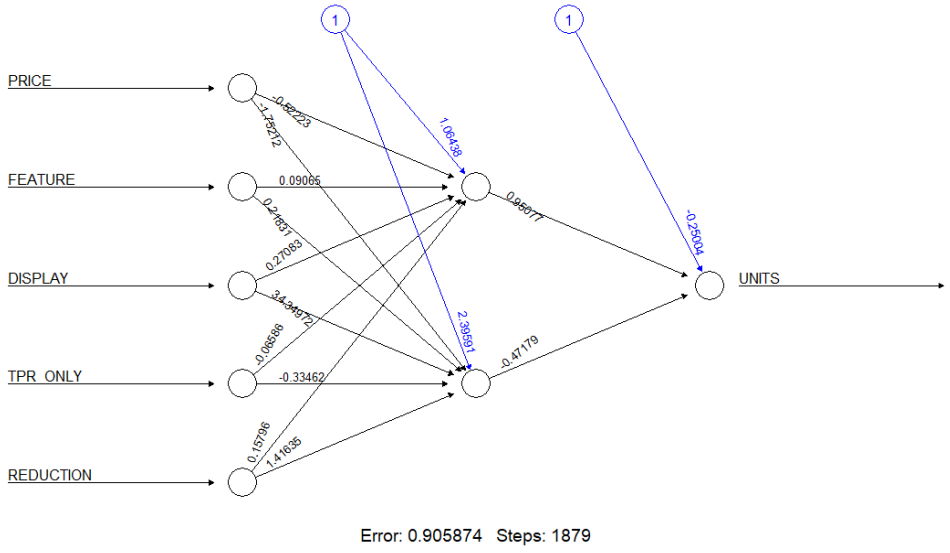
Store planners could take this information, and try to design stores that will have a higher percentage of total products on display, such as including more pop-up displays.

4.1.3 LASSO

The LASSO is a method of automatic variable selection which helps us select predictors (variables) of a target variable from a larger set of potential variables. In our case, it eliminated TPR_ONLY and REDUCTION to build a model to forecast demand. (refer Appendix)



4.1.4 Neural Network



(^ a neural network, explained in 5.6)

4.2 Accuracy of the models and RMSE

To calculate RMSE's for our models, we built the models using the train database, and then the in-sample RMSE (Root mean squared error) was calculated by applying this model on the training dataset. Alternatively, the predictive RMSE came from applying the model onto the test dataset.

Our errors recorded are shown below:

	In-sample RMSE	**Predictive RMSE**
Linear regression	5.139459	3.483139
Regression tree	32.1076	19.14385
LASSO	26.78779	18.54838
Neural network*	0.01447424	0.148043

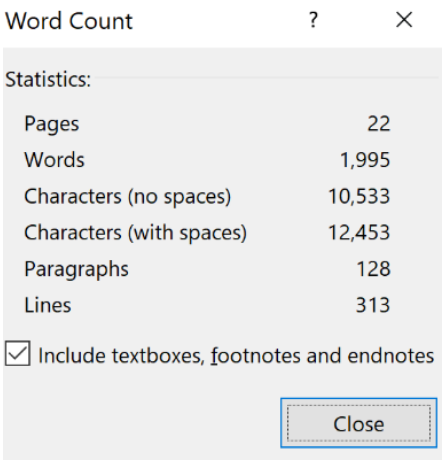
*From the first 10,000 rows

Why are some models better than others?

Linear regression works by fixing a linear model to the data set. Therefore, when the data can best be modelled as a straight line, linear regression is effective, and it struggles when the data is harder to interpret as a straight line.

The regression tree model, while demonstrating the degree of impact of each independent variable more clearly, demonstrated a very high error rate in both the predictive and in sample methods. This could be due to it having a small number of trees (9), and a greater number of trees can allow for it to be a far more reliable predictor. The fact that linear regression performs much better than the regression tree, suggests that the data is more suited to being modelled as a straight line.

The lasso curve includes a penalty term, and applies this to variables which it believes do not impact the final results, this can lead to inaccuracies. Neural networks produce a complex link of layers, and interprets these more insightful results.



5. Appendix

5.1 Elasticities Table by UPC

Mean elasticities by product category and location are insightful for managers, however, below are the elasticities of all products:

UPC	DESCRIPTION	ELASTICITY
3800039118	KELL FROOT LOOPS	-4.3991347
7192100336	DIGIORNO THREE MEAT	-4.0218905
7218063979	FRSC PEPPERONI PIZZA	-4.0189729
7218063052	FRSC BRCK OVN ITL PEP PZ	-3.8210038
3000006610	QKER CAP N CRUNCH	-3.7412646
7192100337	DIGRN SUPREME PIZZA	-3.5542073
7218063983	FRSC 4 CHEESE PIZZA	-3.3461439
7192100339	DIGRN PEPP PIZZA	-3.2929238
3000006560	QKER CAP N CRUNCH BERRIES	-3.1986774
2066200532	NWMN OWN SUPREME PIZZA	-3.0885715
2066200531	NWMN OWN 4 CHEESE PIZZA	-3.0135903
3800031838	KELL FROSTED FLAKES	-2.9998889
2066200530	NWMN OWN PEPPERONI PIZZA	-2.9838624
88491212971	POST FRUITY PEBBLES	-2.8938987
7110410471	MKSL PRETZEL STICKS	-2.7847370
2840004770	RLDGLD PRETZEL STICKS	-2.7356525
3000006340	QKER LIFE ORIGINAL	-2.7187835
2840004768	RLDGLD TINY TWISTS PRTZL	-2.6985039
88491201427	POST FM SZ HNYBNCH OT ALM	-2.5336247
7110410455	MKSL MINI TWIST PRETZELS	-2.5252143
1600027527	GM HONEY NUT CHEERIOS	-2.5107959
31254742835	LSTRNE CL MINT ANTSPTC MW 1LT	-2.4811898
31254742735	LSTRNE CL MINT ANTSPTC MW	-2.3698517
3700019521	CREST PH WHTG CLN MINT TP	-2.1970363
1600027528	GM CHEERIOS 18OZ	-2.1257837
1111087396	PL SR CRUST 3 MEAT PIZZA	-2.0786419
31254742725	LSTRNE CL MINT ANTSPTC MW 500ML	-2.0571354
1111087395	PL SR CRUST SUPRM PIZZA	-2.0521808
88491201426	POST HNY BN OTS HNY RSTD	-2.0431962
7110410470	MKSL DUTCH PRETZELS	-2.0033888
7797508006	SNYDR FF MINI PRETZELS	-1.9562406
7797502248	SNYDR PRETZEL RODS	-1.9097607
7797508004	SNYDR SOURDOUGH NIBBLERS	-1.7316079
2840002333	RLDGLD BRAIDED HONEY WHT	-1.6808449
3700031613	SCOPE ORIG MINT MOUTHWASH	-1.6026277
1111087398	PL SR CRUST PEPPRN PIZZA	-1.5073371
4116709428	ACT MINT A/CAV FLUOR RNS	-1.4428462
3800031829	KELL BITE SIZE MINI WHEAT	-1.3282616
3700044982	CREST PH CLN MINT RINSE	-1.3043067
1111038078	PL BL MINT ANTSPTC RINSE 500ML	-1.1571935
1111038080	PL ANTSPTC SPG MNT MTHWS	-1.1030957
4116709448	ACT KIDS BBLGUM FLUOR RNS	-0.9614665
1111035398	PL BL MINT ANTSPTC RINSE 1.5LT	-0.9405101
4116709565	ACT RSTR CL SPLSH MINT MW	-0.9390749
1600027564	GM CHEERIOS 12OZ	-0.9007926
1111085350	PL BT SZ FRSTD SHRD WHT	-0.7801285
7027312504	SHURGD PRETZEL RODS	-0.6522803
7027316404	SHURGD PRETZEL STICKS	-0.5248605
7027316204	SHURGD MINI PRETZELS	-0.5187390
1111085345	PL RAISIN BRAN	-0.4031625
3500068914	COLG SPEARMINT WISP	-0.2536169
1111085319	PL HONEY NUT TOASTD OATS	0.2539534
1111009497	PL PRETZEL STICKS	1.2367869
1111009477	PL MINI TWIST PRETZELS	1.7946512
1111009507	PL TWIST PRETZELS	1.8786217

5.2 Sampling Error for Elasticities by Product Category by State

The elasticity results summarized in *Figure 6* have sampling error as not all goods are sold in every state. This is supported by computing the unique UPCs in each state; we get 55 UPCs for Indiana, 54 for Kentucky, 55 for Ohio, and 49 for Texas. This could mean that a very inelastic good sold in one state but not the other can skew results.

Specifically, the following products are not sold in Texas:

- **SHURGD PRETZEL RODS**

- **SHURGD MINI PRETZELS**
- **SHURGD PRETZEL STICKS**
- **MKSL MINI TWIST PRETZELS**
- **MKSL DUTCH PRETZELS**
- **MKSL PRETZEL STICKS**

Coincidentally, all of these products fall under the bag snacks category, the least inelastic product category.

And, **NWMN OWN SUPREME PIZZA** is not sold in Kentucky, a product that falls under the frozen pizza category, the most inelastic product category.

The absence of these goods sold in Texas and Kentucky could explain the difference between the mean elasticity of one product category calculated by across all states vs. the results of *Figure 4,* as well as the order of most inelastic to least inelastic product categories in *Figure 7.*

5.3 Distribution of Prices of Products by Sub Category

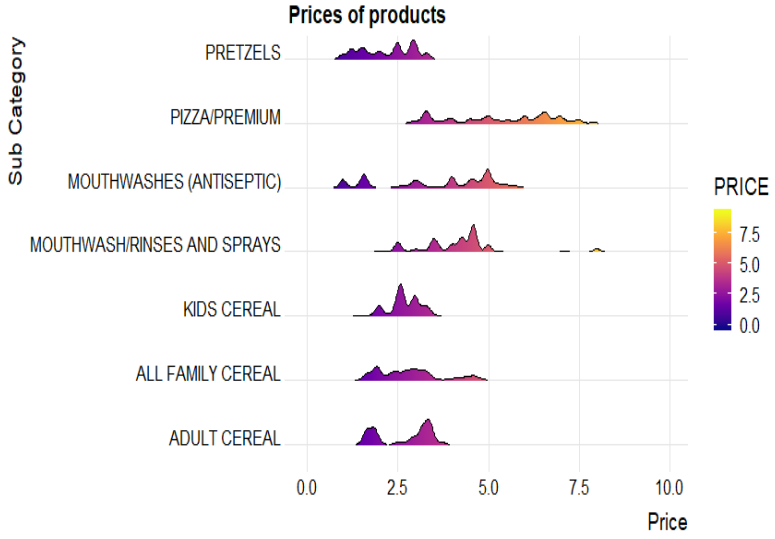


Figure XXX

5.4 Distribution of Units per Visits

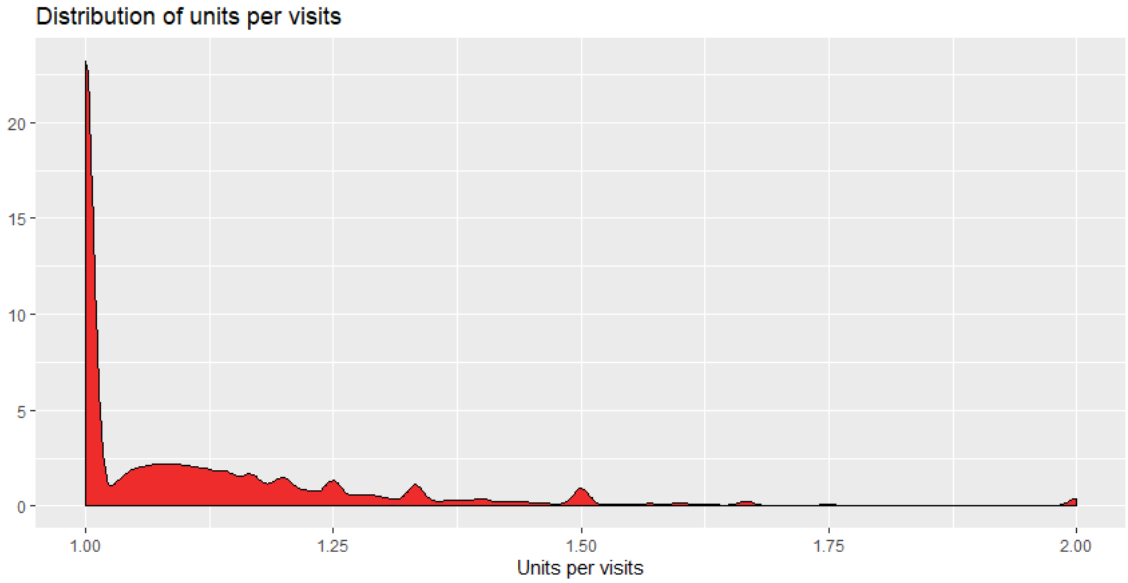


Figure XXX

Figure XXX shows that the data is highly skewed around 1 unit per visit.

5.5. Demand Forecasting: Linear Regression analysis by product category

Price had a negative correlation for bag snacks and frozen pizza, suggesting that as the price charged increased, the unit sales fell. -1 and -0.4 were the size of these correlations – a much larger effect on bag snacks. For oral hygiene products and cold cereal this was not the case, and both correlations were close to zero. This could be due to customers being willing to pay more to upgrade to higher quality products in these categories.

Feature was split between product categories, but did vary far more than the price. Bag snacks and frozen pizza both had negative correlations (-1 and -0.2), whilst oral hygiene products had a slight positive correlation (~0.5). Cold cereal however displayed a very strong correlation of 3.7. Therefore, store managers could look to have a greater number of cold cereal products in their feature, as it had a more positive impact on units sold than the likes of on bag snacks.

Display has a negative correlation for all values, suggesting that featuring a product in a display had a negative effect between 0 and 1 for all values, suggesting that it had a minor impact on determining units sold. This could be due to sampling error or it may have had varying effects on specific types of products, such as 'premium' products which may have performed better with a display.

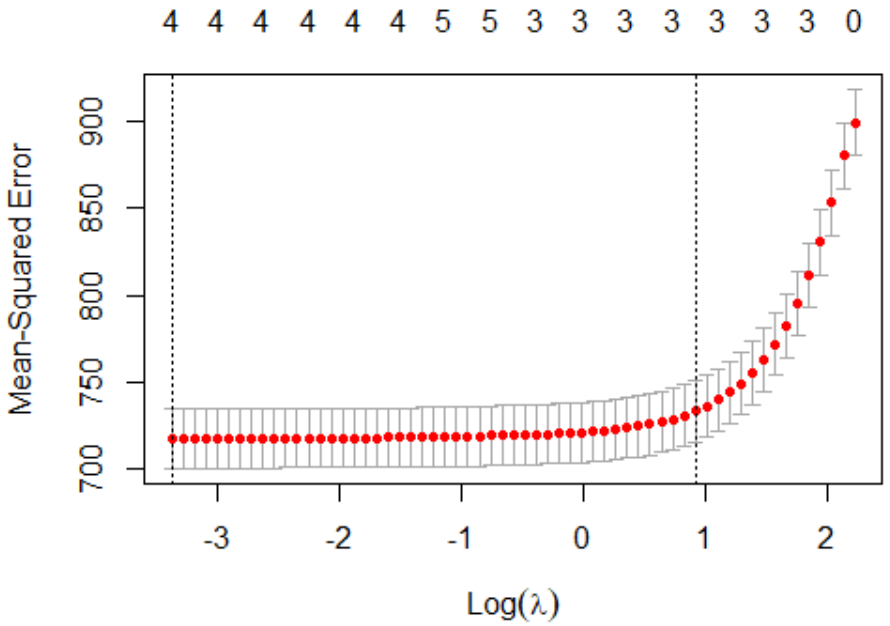
Reduction had a positive correlation of 3.69 for cold cereal sales, but negative for the other values. Store managers could consider having a greater proportion of their stock featuring a reduction be cereal

Everything except cold cereal resulted in a strongly positive correlation with reduction ratio, therefore the other 3 categories appeared more sensitive to the size of the reduction.

5.6 Demand Forecasting: LASSO

The LASSO is a method of automatic variable selection which helps us select predictors (variables) of a target variable from a larger set of potential variables. Each curve corresponds to a variable (1,2,3,4,5 in the order of the variables written in the function). λ is the penalizing factor. It shows the path of its coefficient against the log of lambda as λ varies. The axis above indicates the number of nonzero coefficients at the current λ, which is the effective degrees of freedom for the lasso. It shows from left to right the number of nonzero coefficients (more coefficients reach the value 0 as lambda increases thereby being eliminated), the percent (of null) deviance explained (%dev) and the value of λ. In this case TPR_ONLY and REDUCTION were eliminated.

Cross-validation was used to select the best lambda at which predictions are made. The first one is the value at which the minimal mean squared error is achieved and the second is for the most regularized model whose mean squared error is within one standard error of the minimal lambda at which predictions are made.



We then predict the units sold.

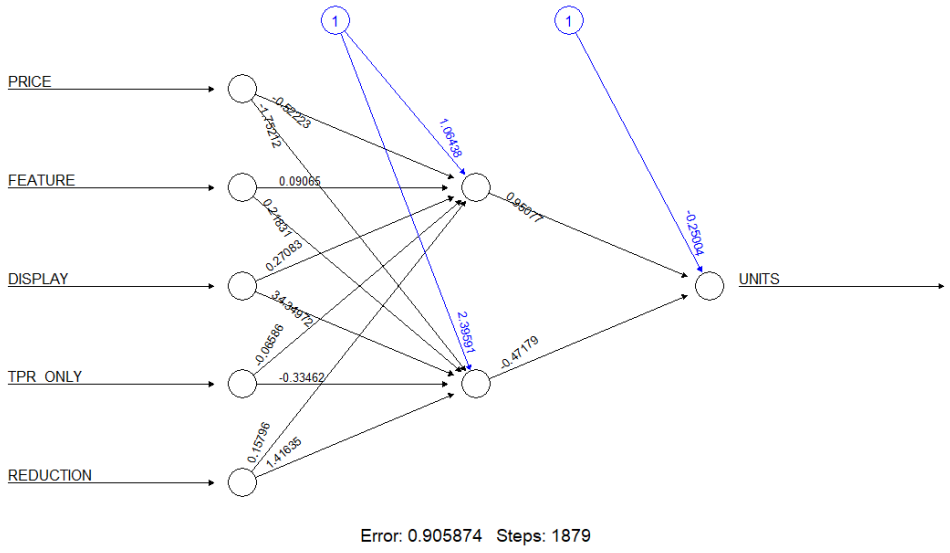
5.7 Demand Forecasting: Neural Network

We used a neural network to provide a more inclusive and accurate model to forecast demand.

We scaled the data to improve our predictions, after which we plotted the neural network. Note that running the neural network for the entire train set takes a long time, so we randomly sampled 10,000 rows for our predictions.

The circles in the first layer represent each input variable. Each circle in the second layer is shown below is called a perceptron. A perceptron takes several inputs and provides a single output. It makes a decision based on weights, which are numbers expressing the importance of the respective inputs to the output. The input variables are linked with the black lines which show the connections between each layer and their weights while the blue lines show the bias term (equivalent to the intercept in a linear model) added in each step. There are two hidden layers. The more layers of perceptrons, the more complex data the network can handle, and the network becomes more and more insightful.

We then predicted on the test data and calculated the error, which is quite low.



5.8 Model Definitions

The **random forest** is an algorithm capable of randomly combining multiple decision trees, training each one based on a set of observations and splitting nodes in each tree. It usually relies on a collection of decision models in order to improve its accuracy.

The **regression tree** is a type of decision tree, used as a predictive model in order to go from observations about an item (shown in the branches) to conclusions about the item's target value.

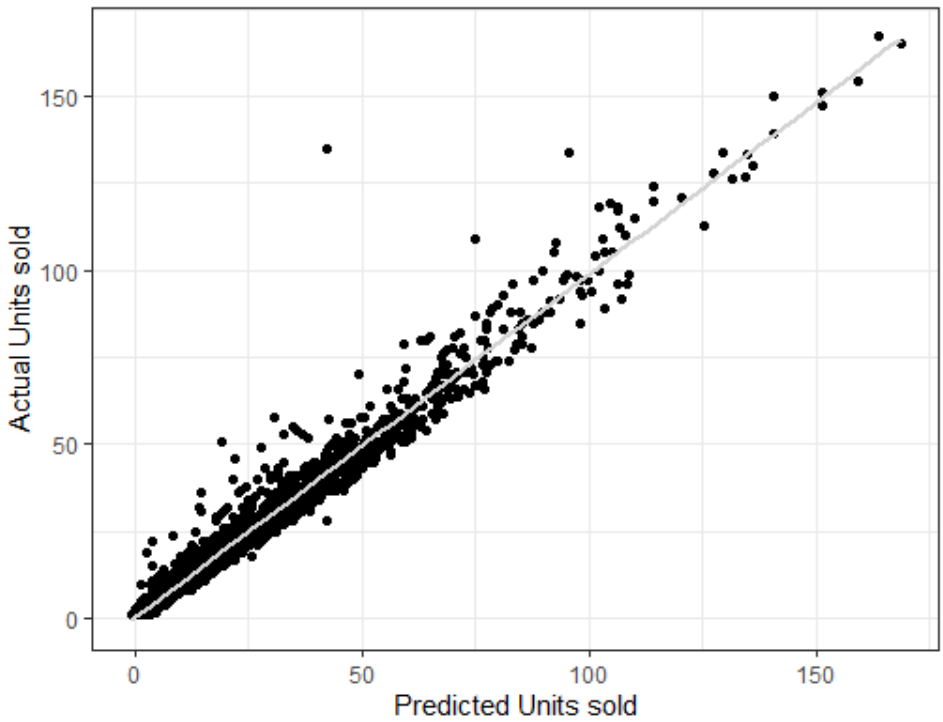
Linear regression is a type of predictive analysis. Linear regression is used to verify if a set of predictor variables are useful in predicting an outcome variable, but also to determine which specific variables are significant predictors of the outcome variable. It also attempts to model the relationship between two variables by fitting a linear equation to observed data.

Lasso (least absolute shrinkage and selection operator) is a regression analysis method that performs variable selection, choosing the best “subset” of predictors, and regularization (process of adding information in order to solve a problem) in order to enhance the prediction accuracy and interpretability of the statistical model it produces.

The **neural network** is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates.

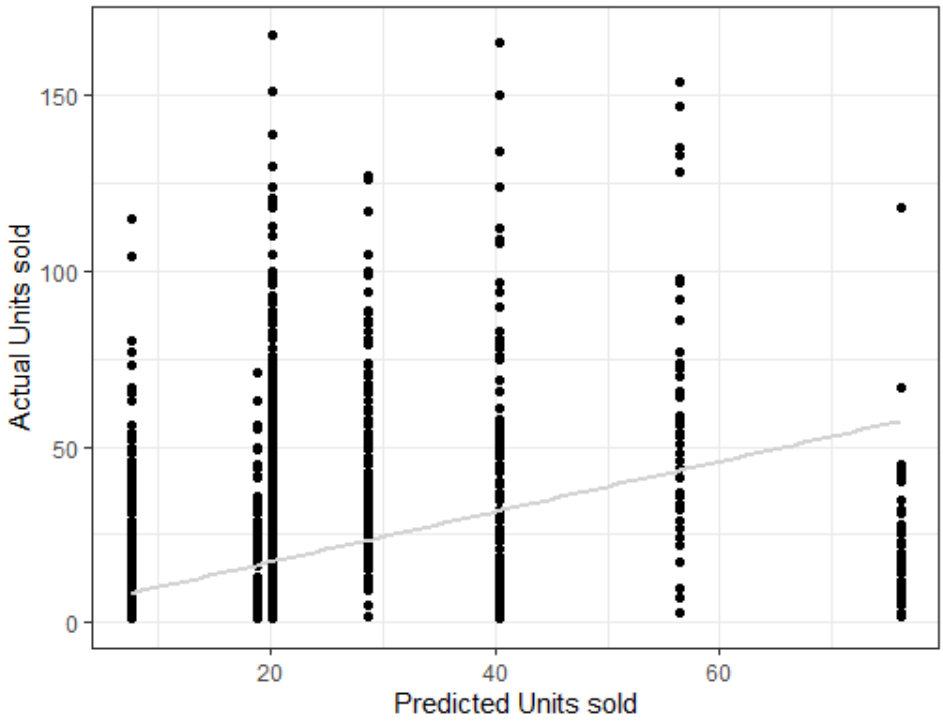
5.9 Demand Forecasting: Graphs to Assess Accuracy of Models

Linear regression:



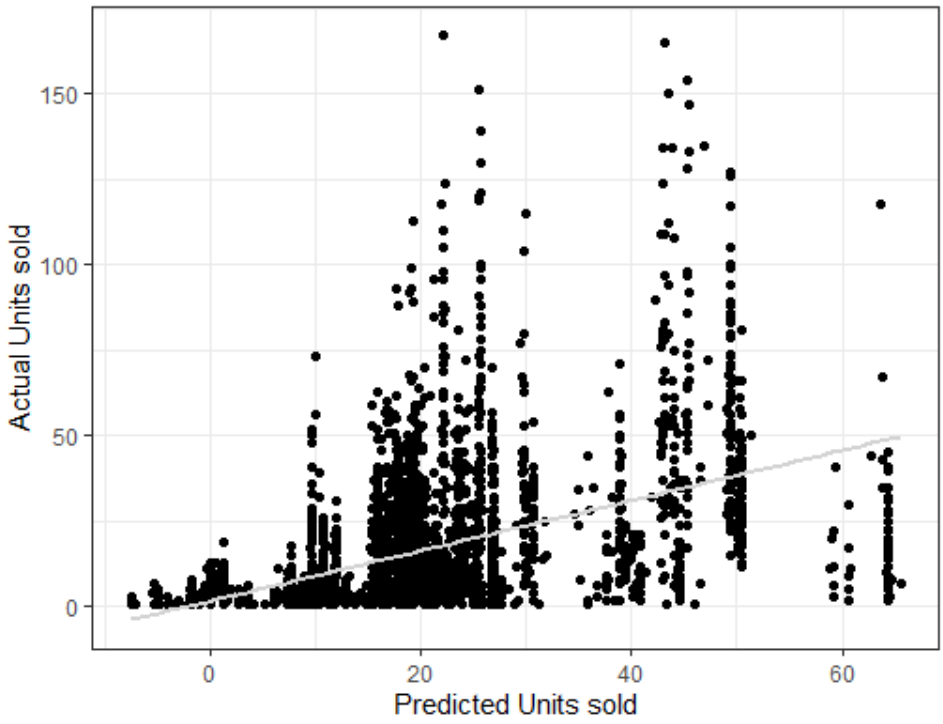
Since this line almost resembles a y=x line, it means that this model is quite accurate as it predicts the actual units correctly.

Regression Tree:



This graph reveals that the regression tree is not really efficient in predicting sales. The regression tree determines which independent variables have the most important impact on the dependent variable. This can be shown by the node positions of all the dependent variables (display is the most important variable according to our model). These produce clear demonstrations of the impact specific variables have on the output data.

LASSO:



LASSO seems to be less accurate in predicting sales, possibly because variable selection leads to deterioration in predictions. Some variables may need to be in the model regardless of any measure of significance and they might be necessary control variables. Moreover, standardization of predictors and classifying them as big or small might not be the best way to standardize categorical predictors. In this case, it decided to drop the variables TPR_ONLY and REDUCTION, which is very counter-intuitive as we can assume they play an important role, as proven by the promotions section.