dunnhumby

MSIN0010 SCENARIO WEEK 3

Data Analytics Report

Giulia Zhang
Irene Liang
Tania Turdean
Jack Lim
Javier Cotoner
Alex Viessmann

Words: 1972

Content

INTRODUCTION

Stage 1:

In order to assist The Retailer to understand deeply about its consumers, the American public, we are utilizing consumer and retail data provided by Dunnhumby. Through our cooperation with Dunnhumby, we received a massive amount of data which enables us to answer two crucial retail concerns:

1. What is the price elasticity of each product sold in America?
2. What is the impact of promotion on units sold?

Stage 2:

After looking into both questions, it is necessary to explore them further, where we will study how both questions differ by geographies. Here, a comparison of Southern America (Texas, Kentucky) and Midwestern States (Ohio, Indiana) will be observed.

1.1 The Market Research Industry

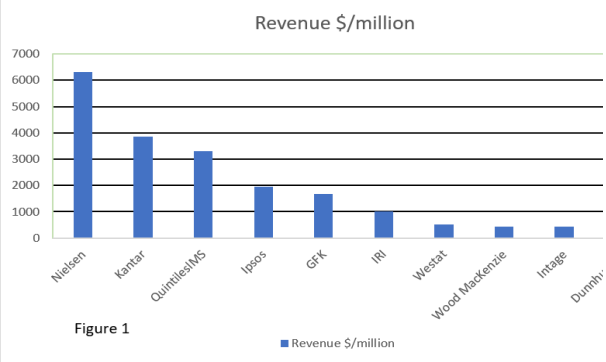


Figure 1

Figure 1 illustrates the Top 10 Market Research Firms, with their respective revenue earned in 2017. This not merely shows the significant size of the market, but also how well each firm is performing.

Figure 2

Based on a report (Figure 2) from First Research Inc, North America remains the leader in the field, followed by Europe and the Asia.

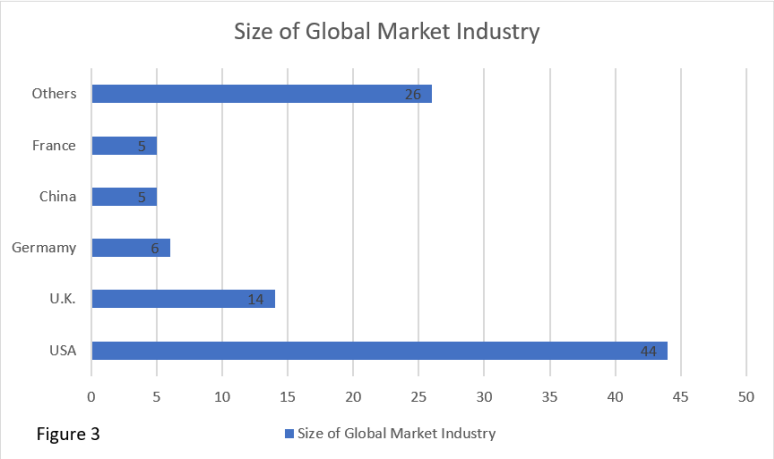


Figure 3

Figure 3 shows that, in Europe, the U.K., Germany, and France are the leading countries; China alone takes one-third of the share of Asia Pacific.

1.2 Summary

In this report, we have found the elasticity of all 55 products, ranging from -0.22 to -4.40. Besides, we found a positive relationship between promotions and units sold, where promotion is made up of three variables namely FEATURE, DISPLAY and TPR_ONLY. Here, FEATURE has the most significant impact on units sold. Lastly, we predict demand by comparing the results from different models and the number of variables. As a result, we found that the regression tree with 8 variables generates the lowest in-sample and out-of-sample RMSE.

1.3 Managerial insights

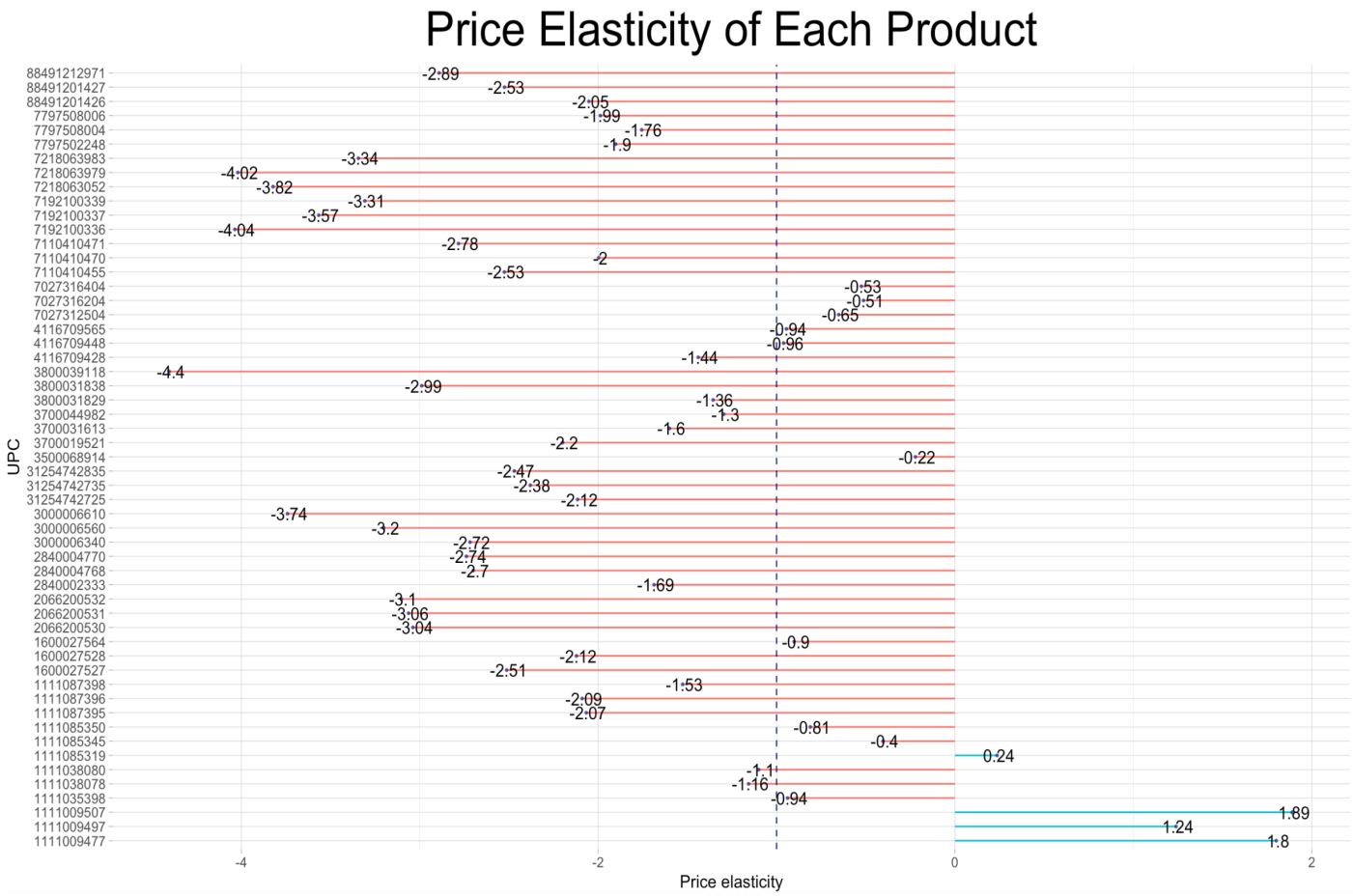
Managers from The Retailer are able to study the impact of demand changes with respect to price changes, with this information the company can generate more revenue simply by charging higher prices on inelastic products and lower prices on elastic products. Furthermore, managers can look into different regions, as we proved that different regions have different levels of elasticity. This can prevent managers from setting inaccurate prices that result in unexpected loss of revenue.

This report enables managers to distinguish the effect of different types of promotions, such as in-store circular, temporary price reductions, and in-store promotions. By choosing the one that gives the largest positive impact, The Retailer would be able to earn more revenue and save cost on unnecessary advertisements.

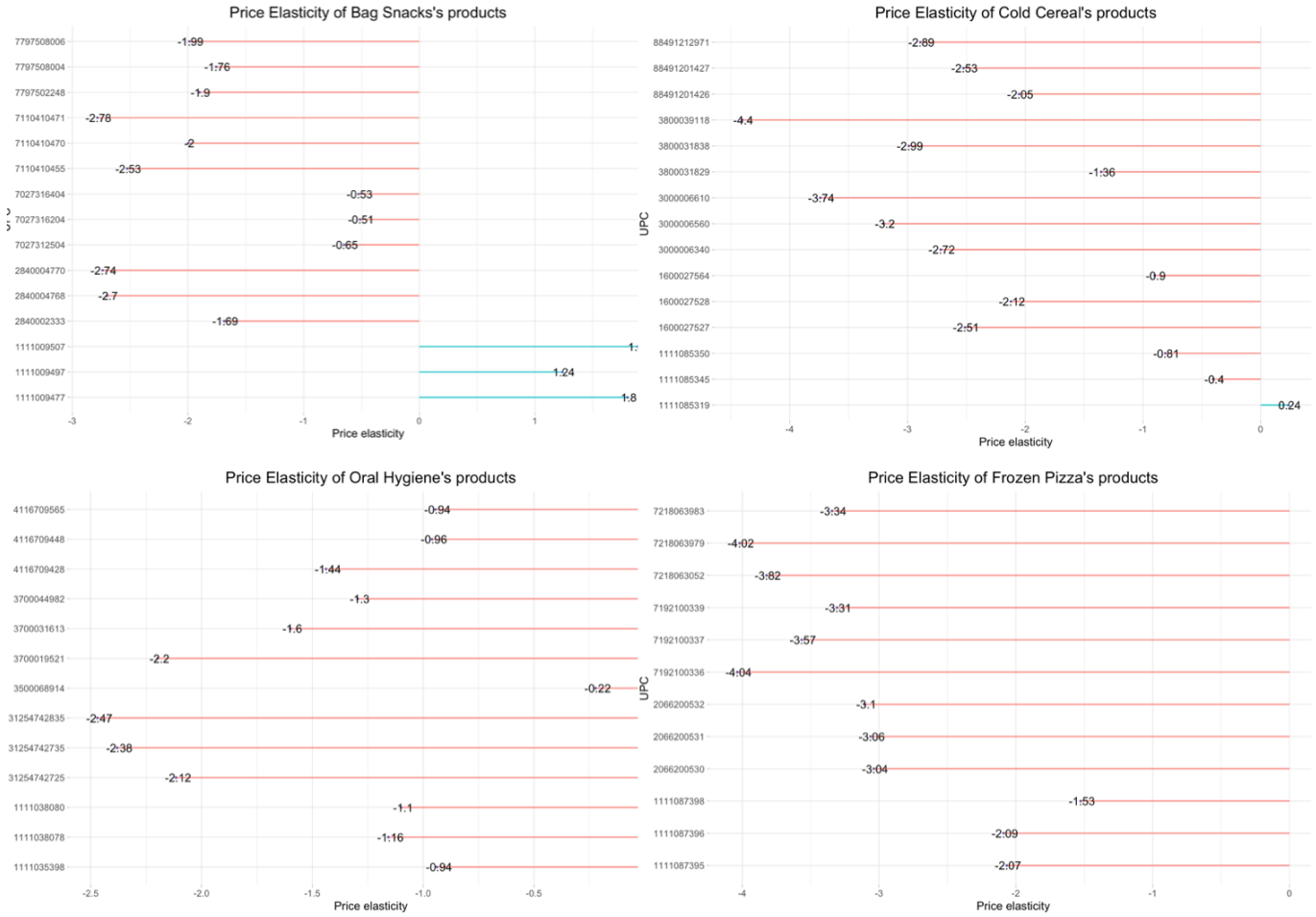
Most importantly, this report provides managers the ability to forecast demand, as we chose the most significant variables and the best regression method.

2. DATA VISUALISATION

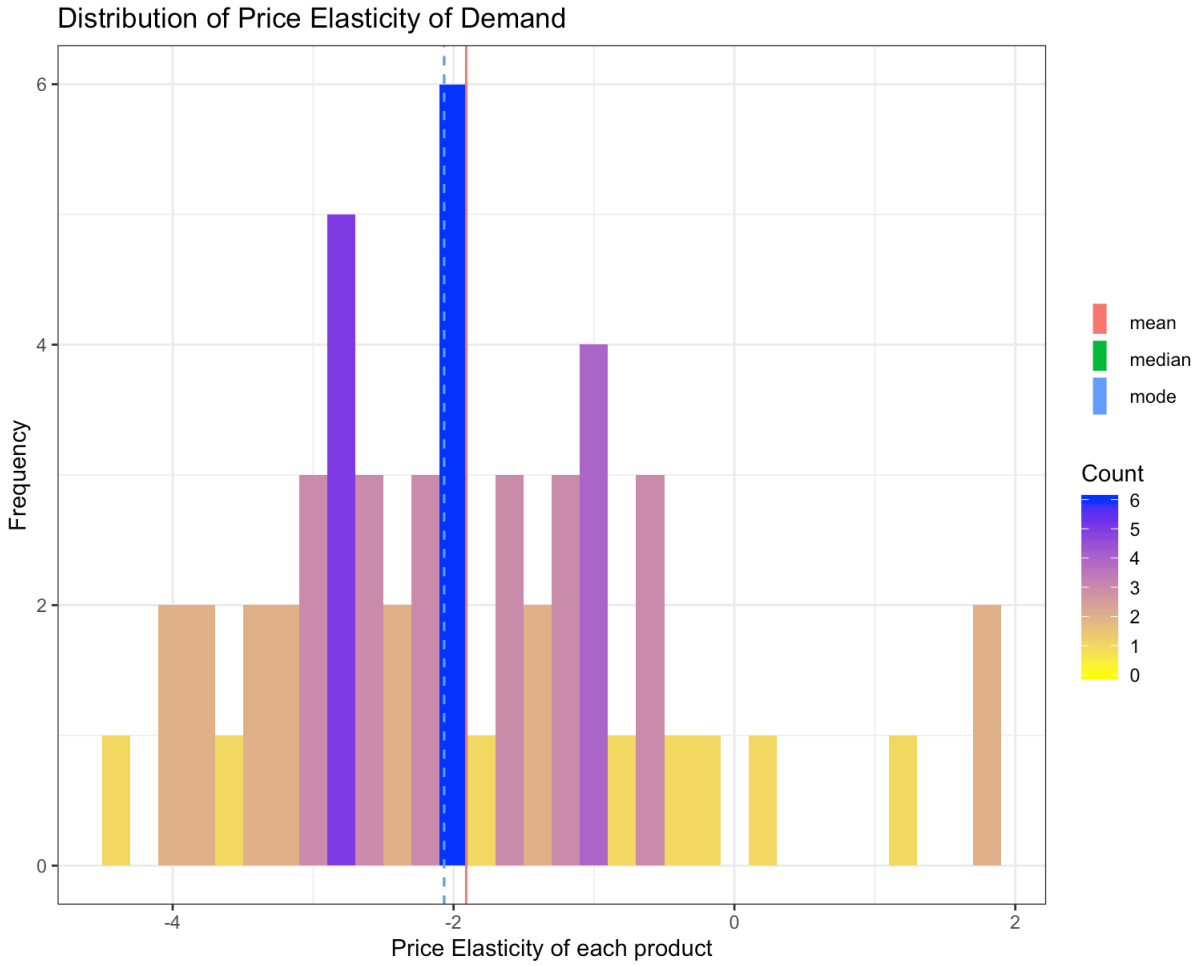
2.1 What is the price elasticity for each product?



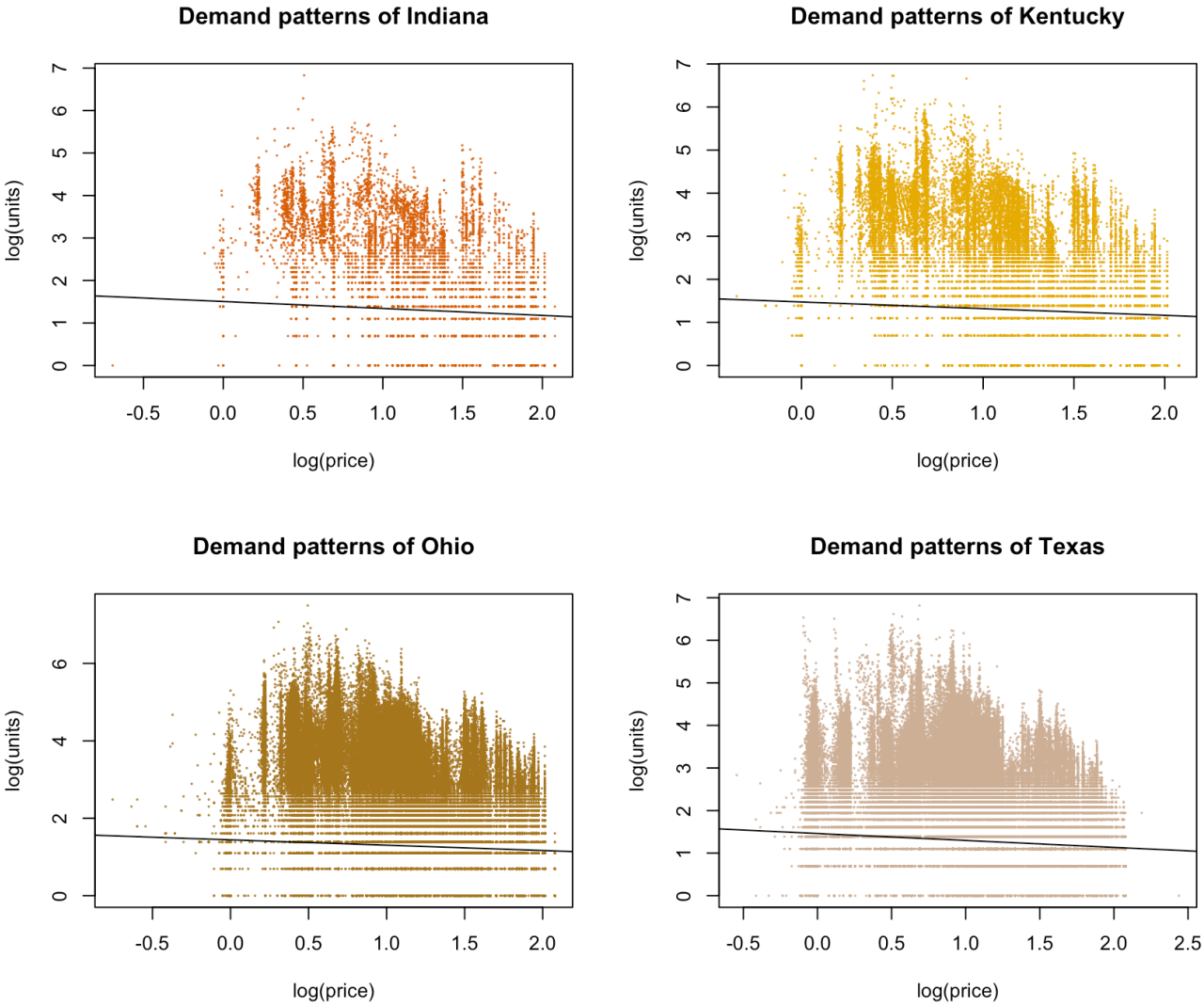
The lollipop chart shows the price elasticity for each product, which is round to two decimals. A vertical line is added to demonstrate how many products are above or below -1. It can be observed that most of the products are relatively elastic as their price elasticity is smaller than -1.



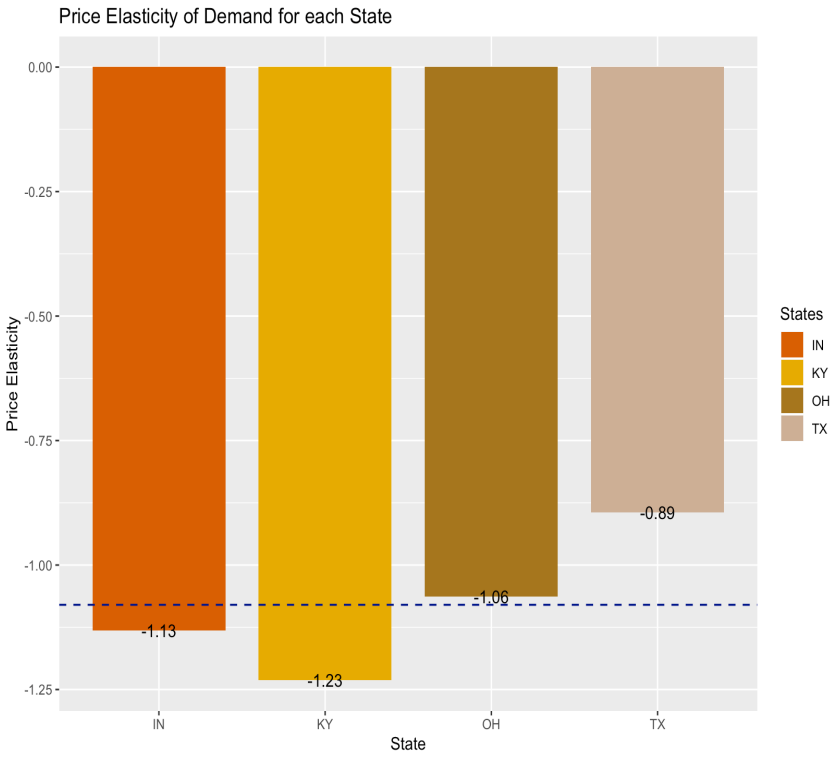
The diagrams above show the price elasticity of the 4 categories. We can see that within the same category, the price elasticity varies a lot. This means that when making a managerial decision, each product has to be taken into consideration for a more accurate analysis.



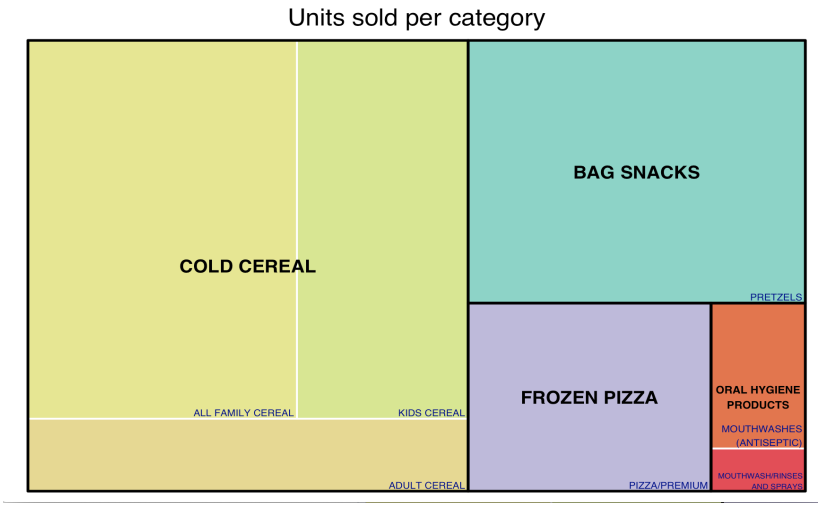
It can be observed from the histogram that the most frequent price elasticity is approximately -2.0 with a frequency of 6 (products). The mean, median and mode, are all allocated in the interval between -2.1 and -1.9. Therefore, a small decrease in the price might result in a big increase in demand since most of the products are relatively elastic.



The scatterplots show the demand pattern of each state with a line whose slope indicates the price elasticity. As the regression line of KY is the steepest, it is possible to conclude that it is the most elastic one.

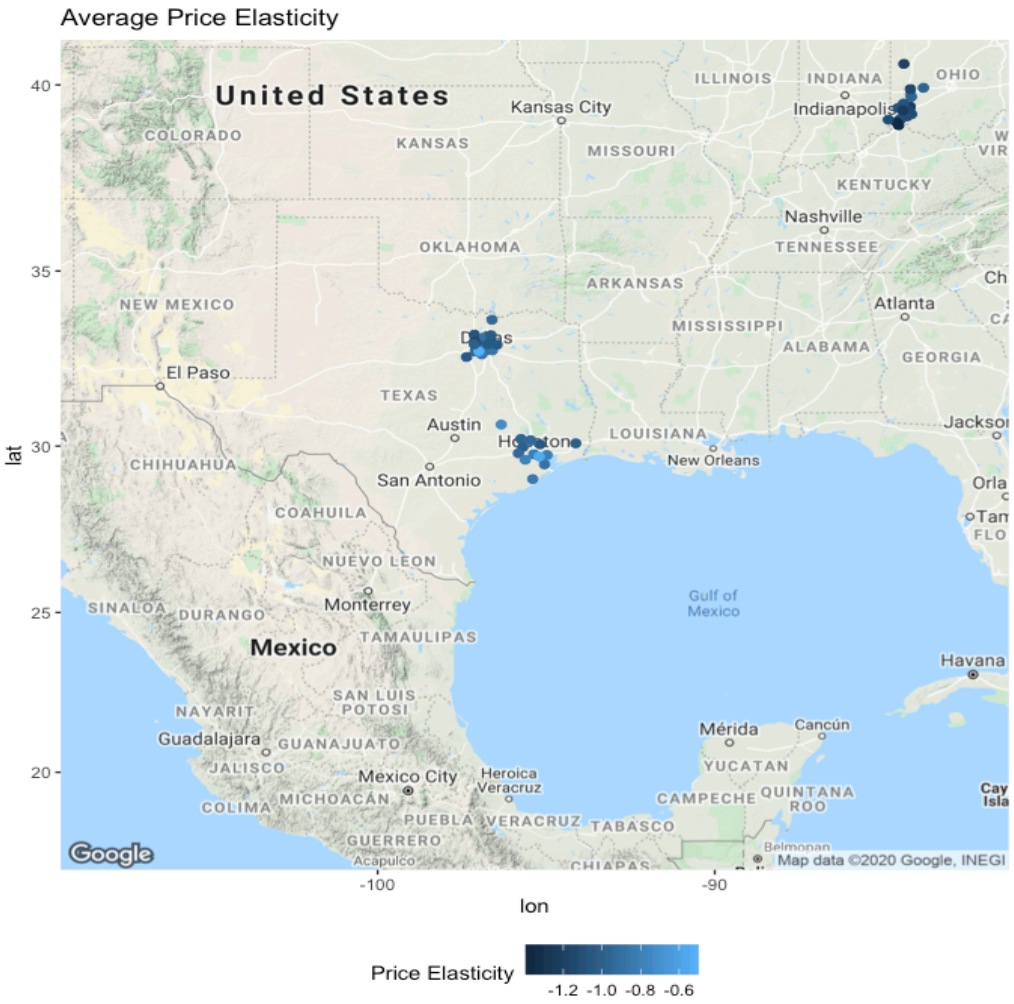


The bar graph demonstrates the price elasticity of demand of each respective state with a dashed line showing the mean of the price elasticity. The price elasticity of TX is between -1 and 0, which indicates a relatively inelastic relationship, whereas for IN, KY and OH, it is lower than -1, making their products relatively elastic.



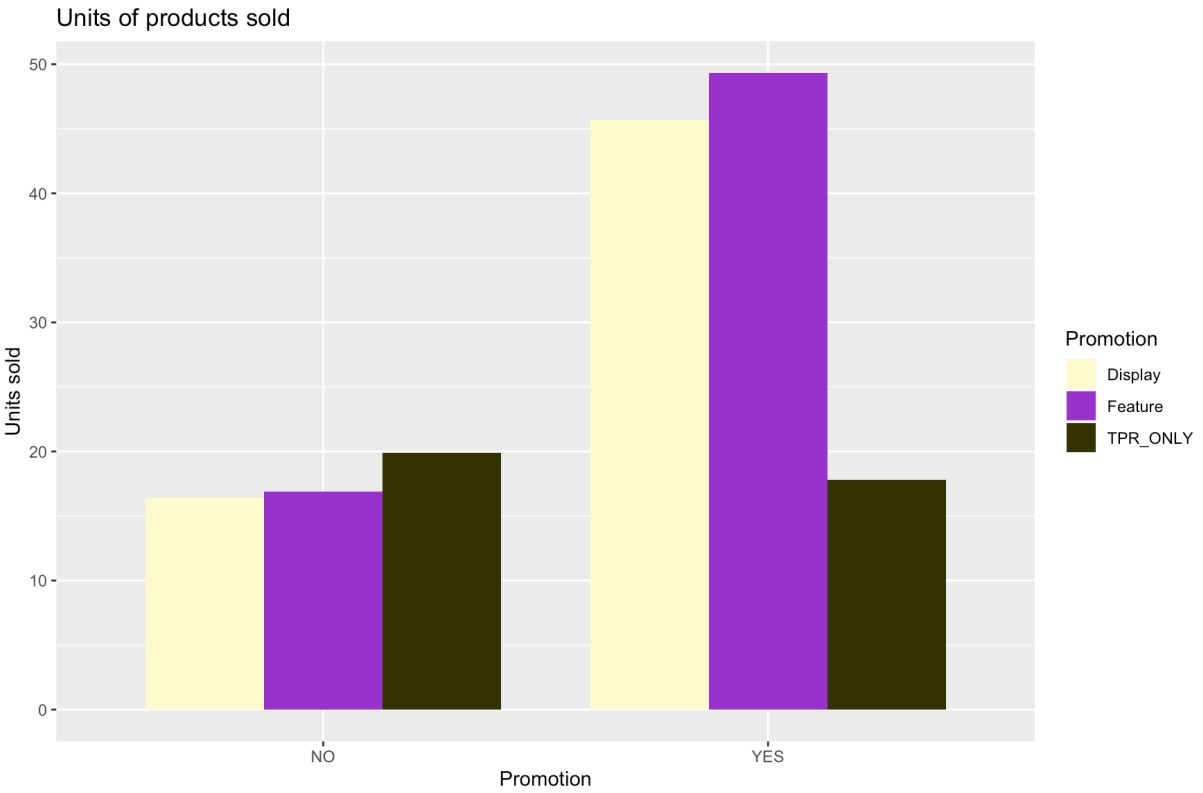
The category which has the highest average units sold is Cold Cereal, then followed by Bag Snacks.
Moreover, Bag Snacks has the lowest Price elasticity of demand, which means that it is the most elastic one.
Therefore, a small decrease in price of Bag Snacks will result in a huge increasement of demand.

	CATEGORY	term	estimate
1	BAG SNACKS	log(PRICE)	-1.5120919
2	COLD CEREAL	log(PRICE)	-0.5210160
3	FROZEN PIZZA	log(PRICE)	-1.3980548
4	ORAL HYGIENE PRODUCTS	log(PRICE)	-0.5778718



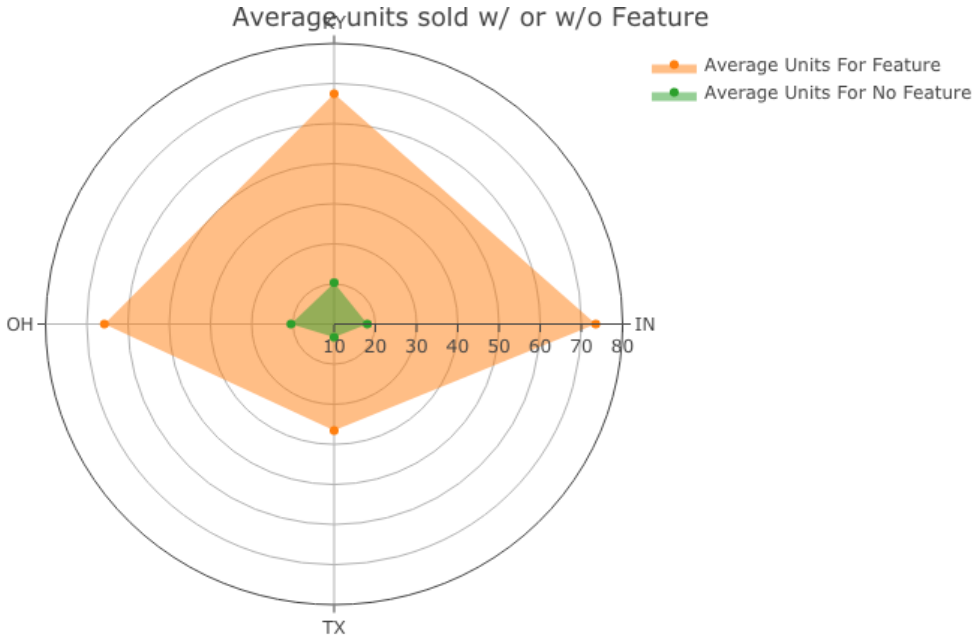
The map shows the price elasticity of each city across states. Each point represents the average units in a city. The darker the point is, the lower the price elasticity is. It can be observed that the midwestern part tends to have lower price elasticity than the southwestern part.

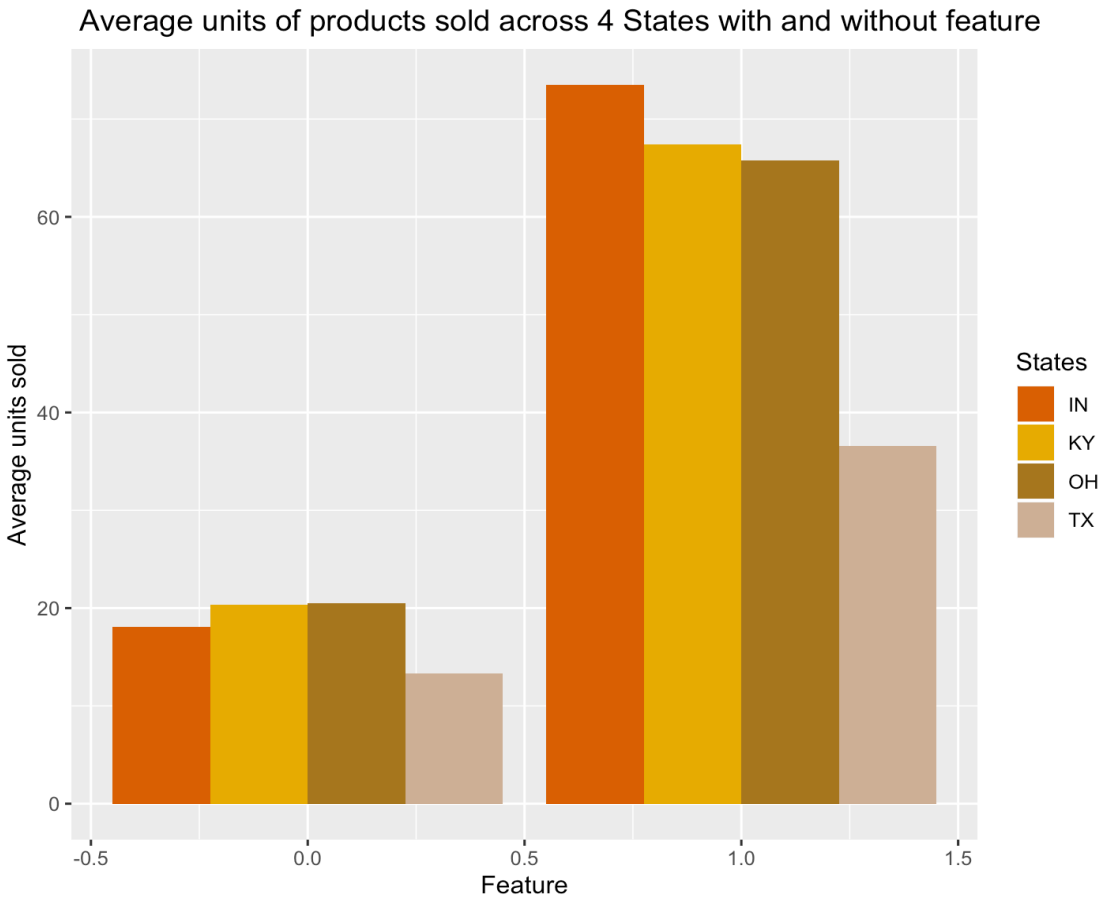
2.2 What is the impact of promotions on units?



The diagram above shows the impact of promotion. Here, we can clearly observe that when there are no promotions, the average units sold is less than when promotions occur. For example, the number of units sold increased by 188% when in-store circular is used, compared to the absence of itself.

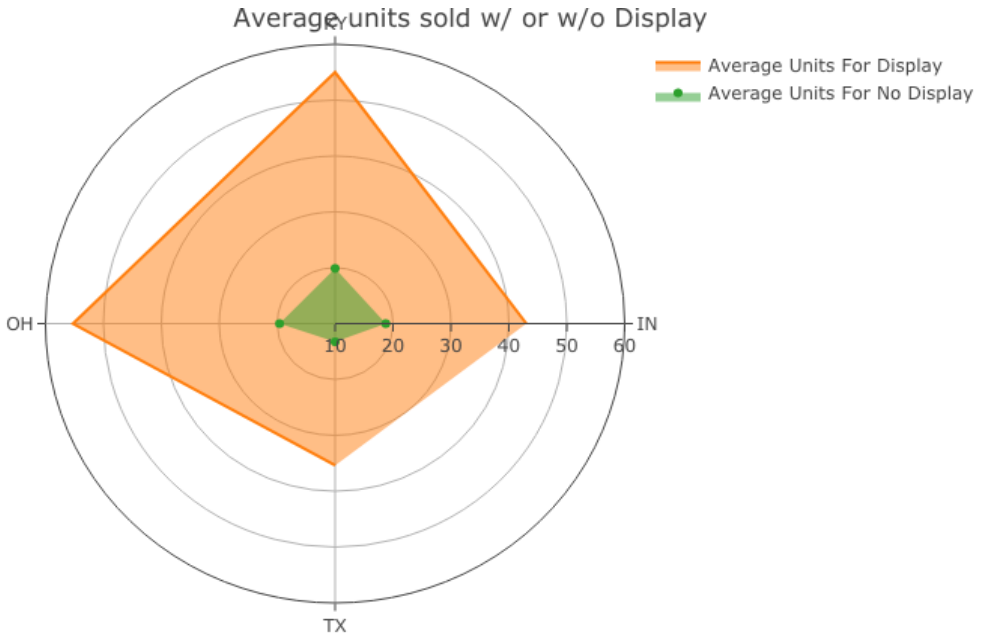
Feature:

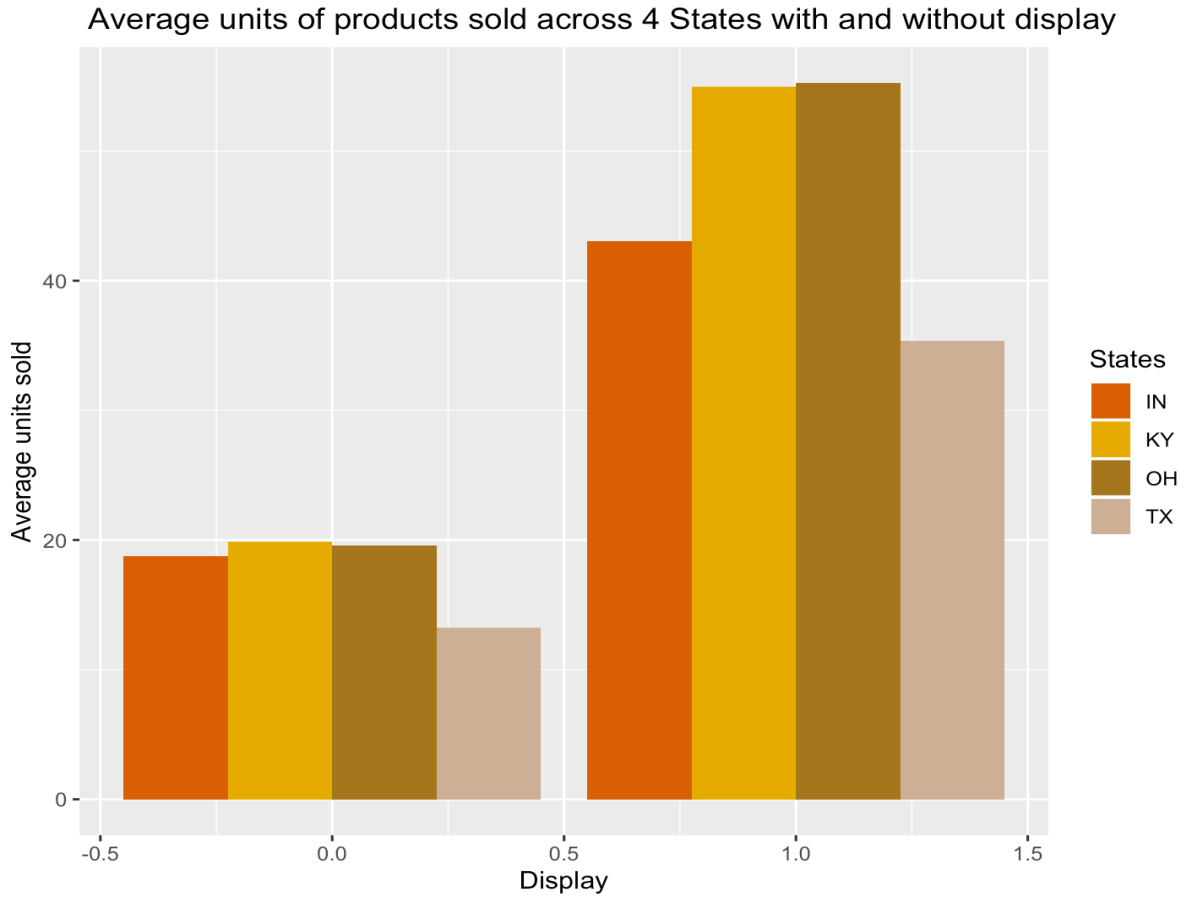




It can be observed that the average units sold is much higher when FEATURE exists since it is the most significant variable found in linear regression. When feature exists, Indiana even surpass Kentucky and Ohio in sales.

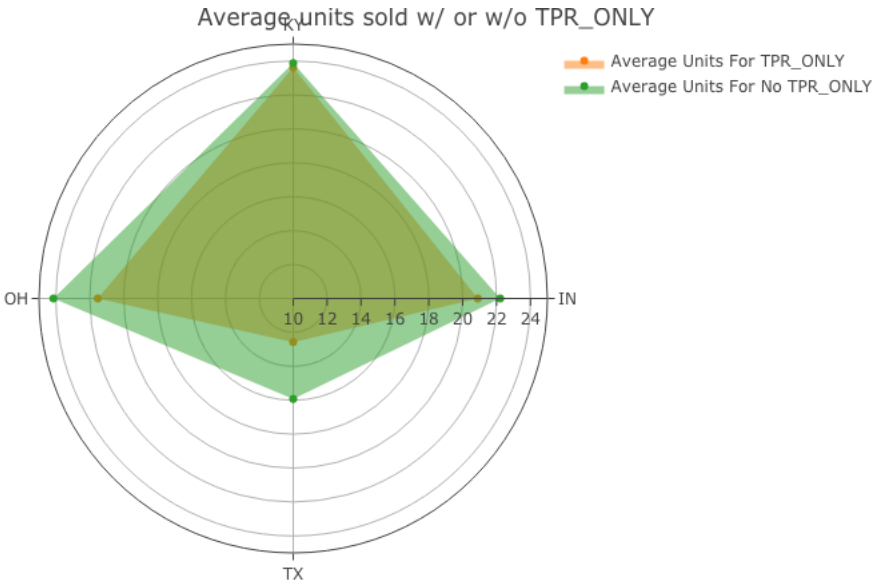
Display:

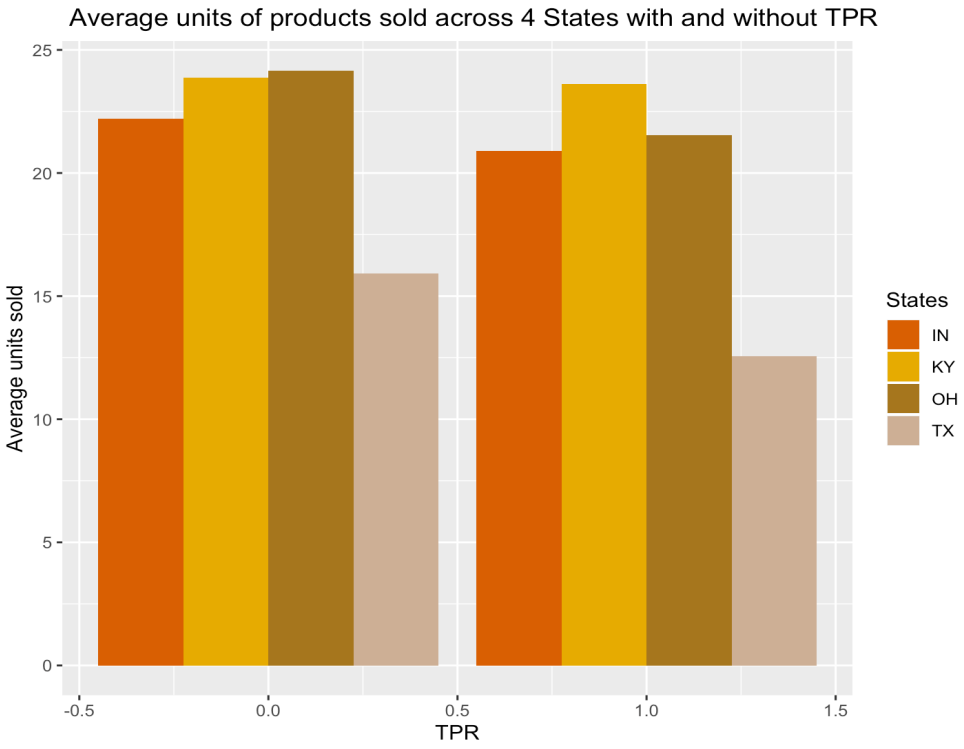




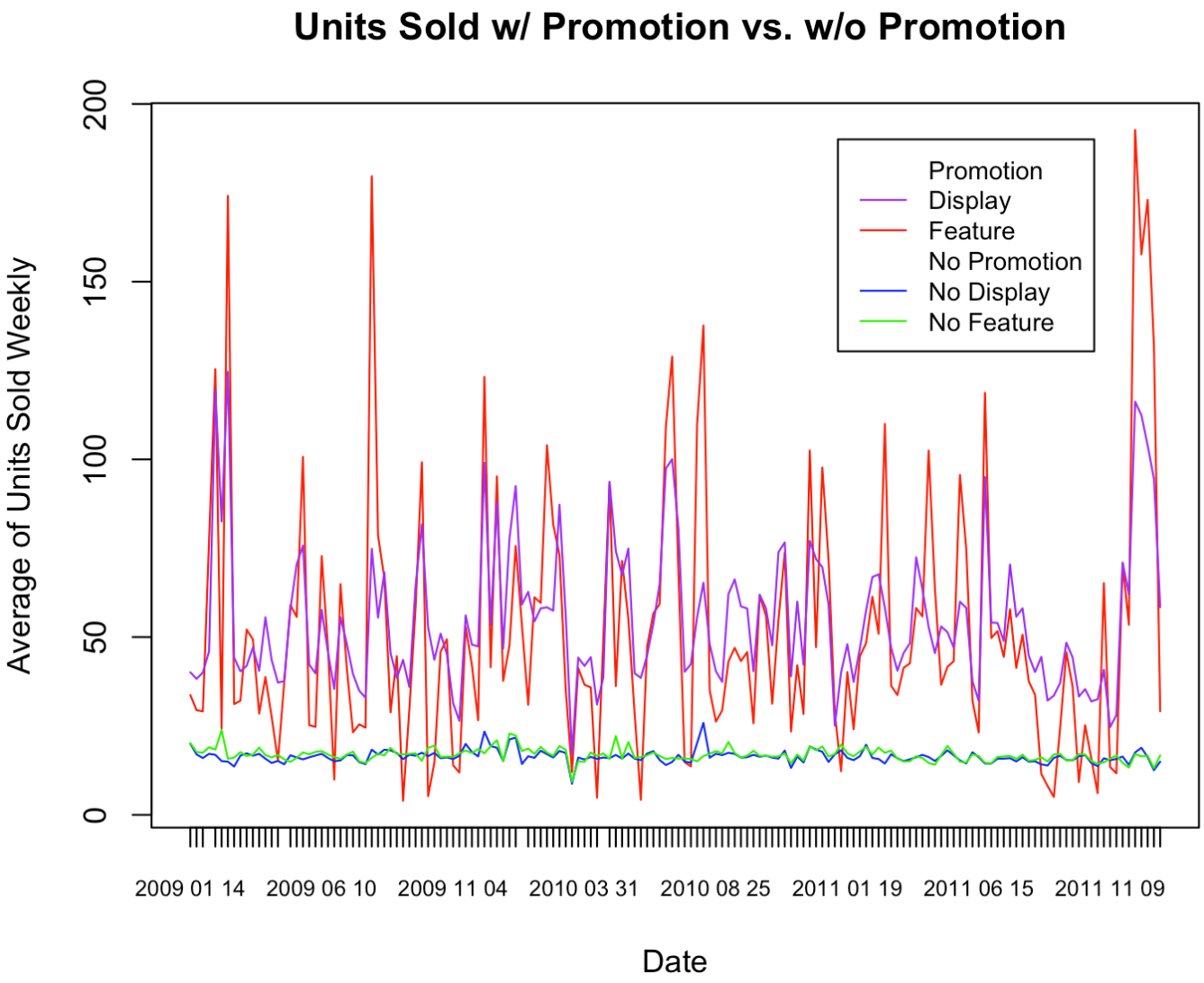
It is possible to see that when products are in display, the sales increase remarkably. For example, the average units sold in Kentucky and Ohio are double when display is included.

TPR only:



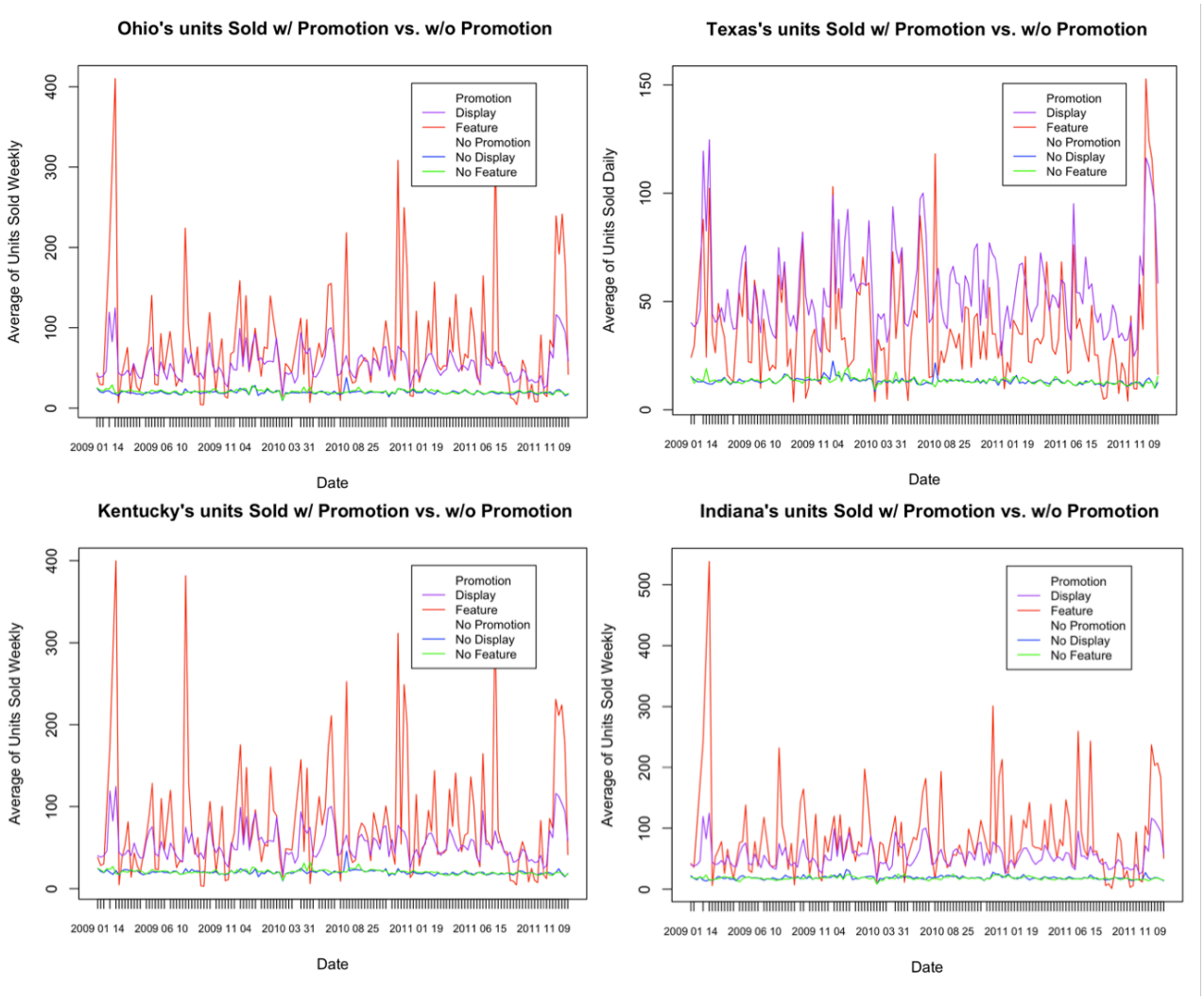


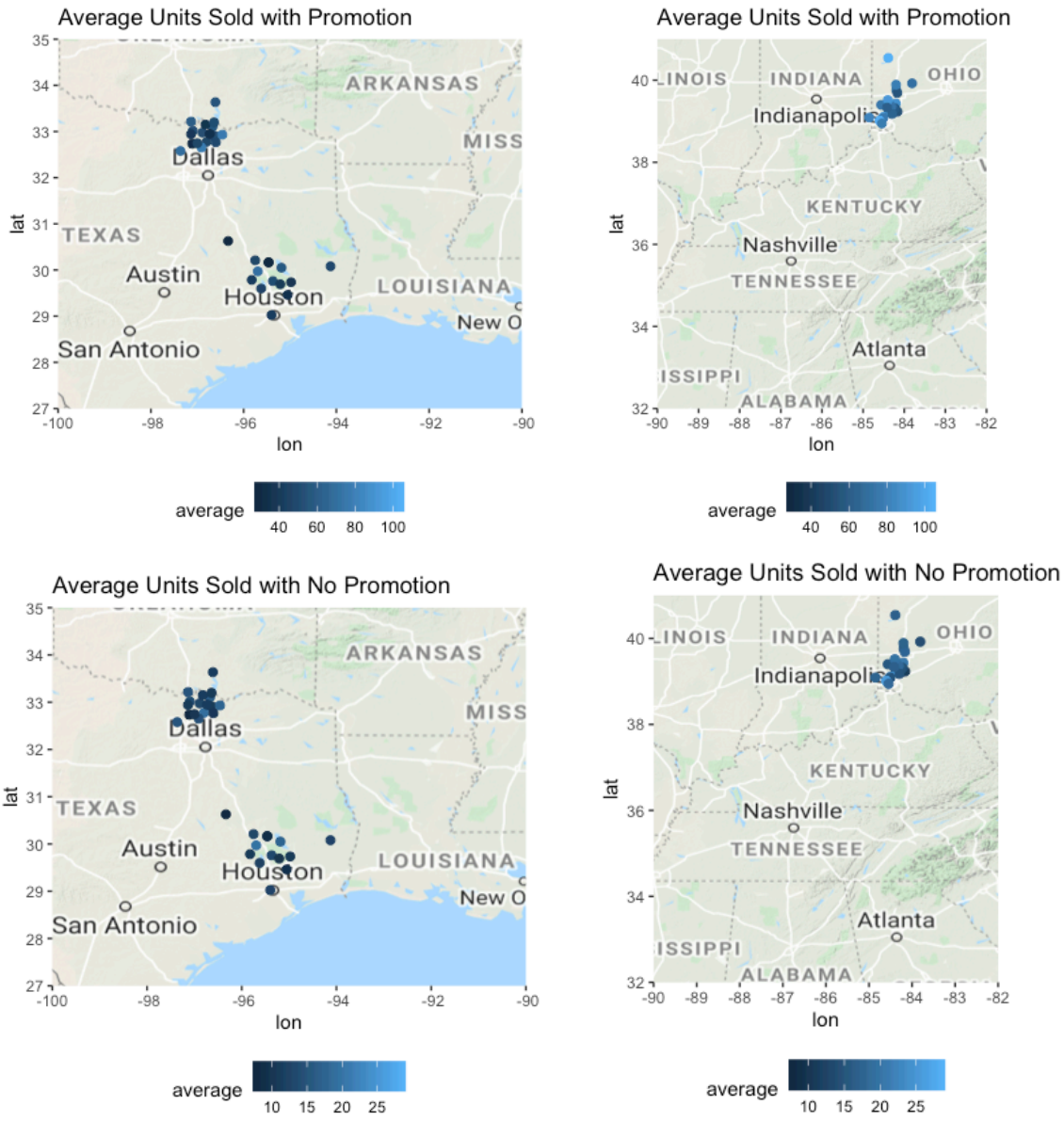
When TPR-only is applied, there is a reduction in sales for each State. Therefore, there is not any increase in demand even when the goods have a price reduction. It might be because TPR-only products are not on display and advertisements, therefore we recommend using other methods of promotions.



The diagram above is a timeline from January 2009 to November 2011 which shows the units of products sold daily when there is a promotion (purple and red) and when there is not (blue and green). We can observe that the impact is very positive all the way through the timeline, obtaining much better results when the promotion is used.

It can be observed that Texas has the smallest interval of fluctuation in sales compared to other states, making it the least affected by promotion. Moreover, the impact of feature on units is more significant than promotion in all states except Texas. Therefore, the managers should choose the most effective method of promotion according to the location to increase sales.





The maps show the average units sold in each city with and without promotion. Promotions include two variables: FEATURE and DISPLAY. In the same way, no promotions include products without FEATURE nor DISPLAY. The darker the point is, the lower the average sales are. It can be observed that products with promotion tend to have higher average sales than non-promoted products, a conclusion which matches previous results obtained.

3. MODELING RESULTS

3.1 Linear Regression: Price Elasticity

	UPC	term	estimate	std.error	statistic	p.value
1	1111009477	log(PRICE)	1.8012762	0.04832328	37.275537	8.769077e-288
2	1111009497	log(PRICE)	1.2392608	0.04395246	28.195483	2.149563e-169
3	1111009507	log(PRICE)	1.8900513	0.05252523	35.983686	1.562602e-268
4	1111035398	log(PRICE)	-0.9384316	0.05116833	-18.340087	7.088252e-74
5	1111038078	log(PRICE)	-1.1572219	0.02561590	-45.175921	0.000000e+00
6	1111038080	log(PRICE)	-1.1004810	0.02771406	-39.708405	1.067182e-321
7	1111085319	log(PRICE)	0.2361616	0.10275029	2.298403	2.155601e-02
8	1111085345	log(PRICE)	-0.4038289	0.08072766	-5.002360	5.743556e-07
9	1111085350	log(PRICE)	-0.8113150	0.08146714	-9.958800	2.843647e-23
10	1111087395	log(PRICE)	-2.0670119	0.04985851	-41.457552	0.000000e+00
11	1111087396	log(PRICE)	-2.0903221	0.04969515	-42.062896	0.000000e+00
12	1111087398	log(PRICE)	-1.5268114	0.05169935	-29.532507	5.178398e-185
13	1600027527	log(PRICE)	-2.5128562	0.03992822	-62.934340	0.000000e+00
14	1600027528	log(PRICE)	-2.1221504	0.03821504	-55.531809	0.000000e+00
15	1600027564	log(PRICE)	-0.9016716	0.04074450	-22.129896	2.173933e-106
16	2066200530	log(PRICE)	-3.0384664	0.12831425	-23.679882	1.084772e-117
17	2066200531	log(PRICE)	-3.0630747	0.15290627	-20.032368	6.447497e-85
18	2066200532	log(PRICE)	-3.1034577	0.15156043	-20.476701	7.681678e-87
19	2840002333	log(PRICE)	-1.6872498	0.08862027	-19.039095	2.625354e-79
20	2840004768	log(PRICE)	-2.7002224	0.08213671	-32.874731	1.142775e-226
21	2840004770	log(PRICE)	-2.7391977	0.08111333	-33.770006	2.927738e-238
22	3000006340	log(PRICE)	-2.7185154	0.04732711	-57.440981	0.000000e+00

23	3000006560	log(PRICE)	-3.2022234	0.05289933	-60.534287	0.000000e+00
24	3000006610	log(PRICE)	-3.7417570	0.05752947	-65.040702	0.000000e+00
25	3500068914	log(PRICE)	-0.2220764	0.04063754	-5.464808	5.493248e-08
26	3700019521	log(PRICE)	-2.1994369	0.04482030	-49.072340	0.000000e+00
27	3700031613	log(PRICE)	-1.6001922	0.05656710	-28.288389	1.038081e-169
28	3700044982	log(PRICE)	-1.2960941	0.08267671	-15.676652	9.226620e-55
29	3800031829	log(PRICE)	-1.3560373	0.05604377	-24.196039	2.674043e-126
30	3800031838	log(PRICE)	-2.9891021	0.04147131	-72.076381	0.000000e+00
31	3800039118	log(PRICE)	-4.4047443	0.03785549	-116.356819	0.000000e+00
32	4116709428	log(PRICE)	-1.4392234	0.10971830	-13.117442	7.253738e-39
33	4116709448	log(PRICE)	-0.9609320	0.10405151	-9.235157	3.283607e-20
34	4116709565	log(PRICE)	-0.9449886	0.08756297	-10.792103	6.019354e-27
35	7027312504	log(PRICE)	-0.6516448	0.06892863	-9.453907	5.086031e-21
36	7027316204	log(PRICE)	-0.5130377	0.06756085	-7.593713	3.680194e-14
37	7027316404	log(PRICE)	-0.5251094	0.06294563	-8.342269	9.306262e-17
38	7110410455	log(PRICE)	-2.5261827	0.10433440	-24.212366	3.291569e-122
39	7110410470	log(PRICE)	-1.9975386	0.11046189	-18.083509	1.234034e-70
40	7110410471	log(PRICE)	-2.7831683	0.10810317	-25.745484	2.260738e-136
41	7192100336	log(PRICE)	-4.0364408	0.04203685	-96.021485	0.000000e+00
42	7192100337	log(PRICE)	-3.5665615	0.04138460	-86.180893	0.000000e+00
43	7192100339	log(PRICE)	-3.3082738	0.04313651	-76.693128	0.000000e+00
44	7218063052	log(PRICE)	-3.8229580	0.04549913	-84.022671	0.000000e+00

44	7218063052	log(PRICE)	-3.8229580	0.04549913	-84.022671	0.000000e+00
45	7218063979	log(PRICE)	-4.0199094	0.04601829	-87.354594	0.000000e+00
46	7218063983	log(PRICE)	-3.3443145	0.04931979	-67.808779	0.000000e+00
47	7797502248	log(PRICE)	-1.9021447	0.07441393	-25.561675	4.791330e-140
48	7797508004	log(PRICE)	-1.7562954	0.10513511	-16.705127	8.112576e-62
49	7797508006	log(PRICE)	-1.9880574	0.09221215	-21.559602	6.203530e-101
50	31254742725	log(PRICE)	-2.1155882	0.04588673	-46.104573	0.000000e+00
51	31254742735	log(PRICE)	-2.3808433	0.06242408	-38.139823	3.735105e-299
52	31254742835	log(PRICE)	-2.4707953	0.06332644	-39.016801	5.265770e-311
53	88491201426	log(PRICE)	-2.0521428	0.06921525	-29.648709	1.004042e-185
54	88491201427	log(PRICE)	-2.5254149	0.06827152	-36.990754	2.364277e-281
55	88491212971	log(PRICE)	-2.8922819	0.04567602	-63.321675	0.000000e+00

The tables above illustrate the Price Elasticity of all 55 products. We added Log in our linear equations as it provides a percentage change in price and quantity, which fully falls within the definition of Price Elasticity of Demand.

Price elasticity for each state

	ADDRESS_STATE_PROV_CODE	term	estimate	std.error	statistic	p.value
1	IN	log(PRICE)	-1.1308359	0.027061305	-41.78793	0
2	KY	log(PRICE)	-1.2315420	0.014943570	-82.41284	0
3	OH	log(PRICE)	-1.0628398	0.005389001	-197.22392	0
4	TX	log(PRICE)	-0.8942141	0.004222797	-211.75871	0

The table above illustrates the average price elasticity of demand of all 55 products across 4 states. In general, most products are price elastic.

3.2 Linear Regression: Impact of Promotions on units sold

Based on the dataset provided by Dunnhumby, it is logical to measure the impact of promotions by considering three variables: 'DISPLAY', 'FEATURE' and 'TPR_ONLY'. The DISPLAY variable represents the items that are part of the in-store promotional display, the FEATURE variable represents in-store circular, whereas the TPR_ONLY variable represents the items that have temporary price reduction only. TPR_ONLY is different from DISPLAY and FEATURE as the products are not on display and advertisements. Although these variables are different to a small extent, we observed that all 3 variables have the same effect, which is a reduction in prices.

```
Call:
lm(formula = UNITS ~ DISPLAY + TPR_ONLY + FEATURE, data = trans_p_s)

Residuals:
    Min      1Q  Median      3Q     Max
 -58.40  -12.80   -6.93    5.20 1740.60

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.93314    0.04502   331.7   <2e-16 ***
DISPLAY     21.38416    0.13541   157.9   <2e-16 ***
TPR_ONLY     2.86445    0.11412    25.1   <2e-16 ***
FEATURE     23.08582    0.15214   151.7   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27.89 on 524681 degrees of freedom
Multiple R-squared:  0.132,     Adjusted R-squared:  0.132
F-statistic: 2.66e+04 on 3 and 524681 DF,  p-value: < 2.2e-16
```

In the linear regression model above, we estimated the effect of DISPLAY, FEATURE and TPR_ONLY on the number of units sold. As a result, we can examine that when DISPLAY exists, the units of goods sold increased by 21.38, when TPR_ONLY exists, the units of goods sold increased by 2.86 and units increased by 23.09 when FEATURE exists. Besides, we can also interpret that FEATURE has the biggest impact, followed by DISPLAY and TPR_ONLY, although they are all significant as P-value < 0.05. This means that managers are able to decide which marketing strategies will yield higher quantities. In this case, managers should provide more in-store circular and promotions than a temporary price reduction. From this, we may conclude that consumers should be aware of the promotions prior to their visits to the stores. Temporary stock reductions are not made aware of the public and consumers are not able to find out until they are present in the stores. Hence, managers should make early decisions on advertising products instead of adding TPR labels on the product's shelf.

3.3 DEMAND FORECASTING

The purpose of this section is to forecast demand, where the best method will be chosen to make a prediction.

In this section, we use neural networks to find variables that are able to reduce the error significantly and we obtained 5 variables namely: DISPLAY, TPR_ONLY, FEATURE + PRICE + BASE_PRICE. Besides, we listed 8 variables from the dataset in order to understand a bigger picture of the relationship between UNITS and these variables, through logical elimination. These 8 variables have a stronger relationship with units sold.

From these two specific scenarios, we carried out linear regression and regression tree for both in order to determine which model provides the best prediction.

3.3.1 DEMAND FORECASTING: 8 variables

(DISPLAY + TPR_ONLY + FEATURE + PRICE + BASE_PRICE + CATEGORY +MANUFACTURER + ADDRESS_CITY_NAME)

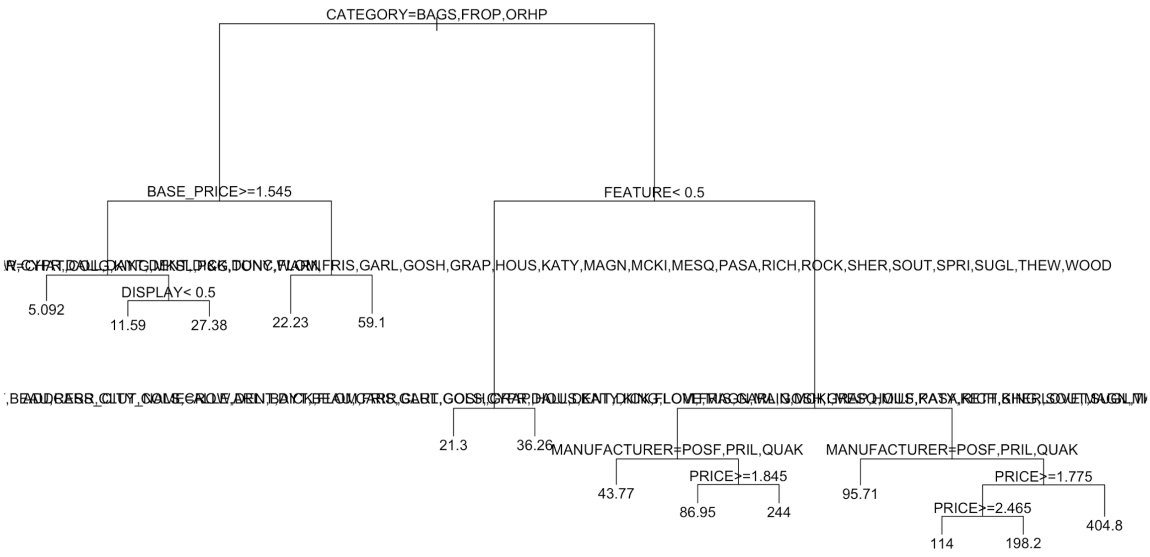
1. Linear Regression

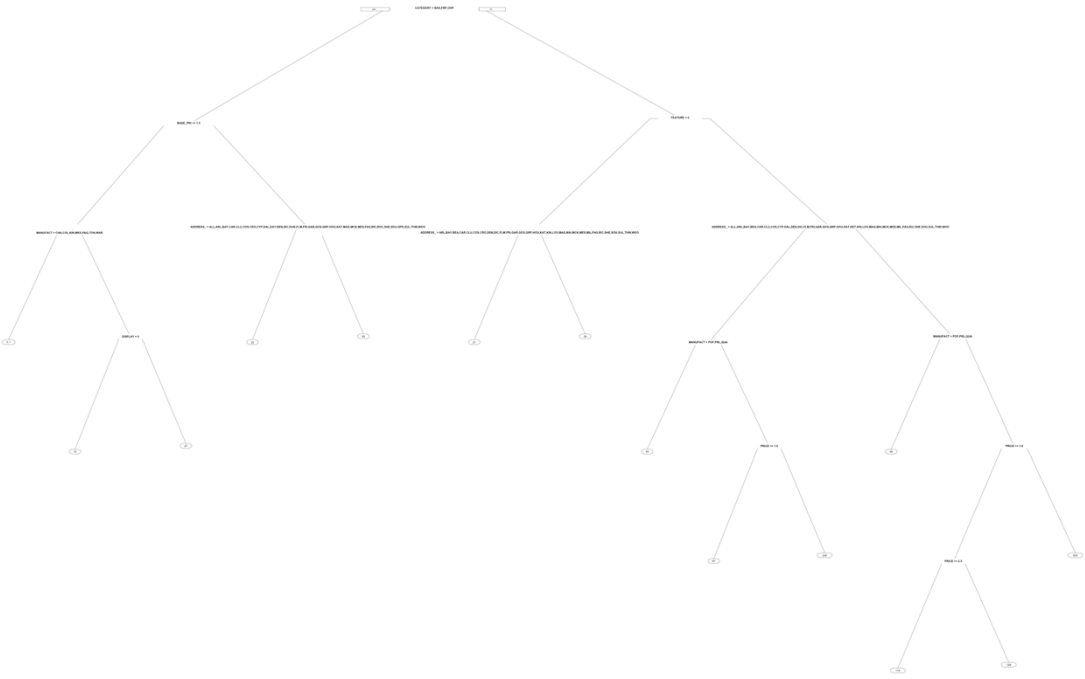
```
Call:
lm(formula = UNITS ~ PRICE + BASE_PRICE + FEATURE + DISPLAY +
    TPR_ONLY + CATEGORY + MANUFACTURER + ADDRESS_CITY_NAME, data = train_p_s_copy)

Residuals:
    Min      1Q  Median      3Q     Max
 -88.32   -9.93   -1.00    6.74 1700.54

Coefficients:
                                Estimate Std. Error  t value Pr(>|t|)
(Intercept)                      53.9101     0.4562  118.173  < 2e-16 ***
PRICE                            -3.6073     0.1089  -33.128  < 2e-16 ***
BASE_PRICE                       -1.1348     0.1159   -9.790  < 2e-16 ***
FEATURE                          22.3336     0.1598  139.760  < 2e-16 ***
DISPLAY                          19.0388     0.1220  156.089  < 2e-16 ***
TPR_ONLY                          1.7874     0.1260   14.184  < 2e-16 ***
CATEGORYCOLD CEREAL              -4.3903     0.1814  -24.199  < 2e-16 ***
CATEGORYFROZEN PIZZA            -15.6181     0.2454  -63.636  < 2e-16 ***
CATEGORYORAL HYGIENE PRODUCTS   -29.7127     0.1921 -154.680  < 2e-16 ***
MANUFACTURERCOLGATE              14.0139     0.6940   20.194  < 2e-16 ***
MANUFACTURERFRITO LAY           -30.3841     0.2807 -108.243  < 2e-16 ***
MANUFACTURERGENERAL MI           10.1355     0.2785   36.398  < 2e-16 ***
MANUFACTURERKELLOGG               0.4291     0.2851    1.505 0.132332
MANUFACTURERKING                 -4.1681     0.3232  -12.898  < 2e-16 ***
MANUFACTURERMKSL                -42.1899     0.3331 -126.643  < 2e-16 ***
MANUFACTURERP & G                -4.3524     0.2091  -20.814  < 2e-16 ***
MANUFACTURERPOST FOODS           -9.1327     0.2903  -31.458  < 2e-16 ***
MANUFACTURERPRIVATE LABEL       -10.5622     0.2714  -38.925  < 2e-16 ***
MANUFACTURERQUAKER              -17.4428     0.2997  -58.202  < 2e-16 ***
MANUFACTURERSHULTZ              -28.7387     0.3395  -84.638  < 2e-16 ***
MANUFACTURERSNYDER S            -32.1595     0.2836 -113.400  < 2e-16 ***
MANUFACTURERTOMBSTONE            -2.3142     0.2721   -8.504  < 2e-16 ***
MANUFACTURERTONYS                -7.0119     0.2806  -24.989  < 2e-16 ***
MANUFACTURERWARNER               -0.5052     0.2049   -2.466 0.013665 *
ADDRESS_CITY_NAMEARLINGTON      -12.9323     0.4135  -31.279  < 2e-16 ***
ADDRESS_CITY_NAMEBAYTOWN         -8.9907     0.4076  -22.060  < 2e-16 ***
ADDRESS_CITY_NAMEBEAUMONT        -3.7919     0.3992   -9.500  < 2e-16 ***
ADDRESS_CITY_NAMEBLUE ASH        15.5705     0.3817   40.797  < 2e-16 ***
ADDRESS_CITY_NAMECARROLLTON      -5.0293     0.3971  -12.665  < 2e-16 ***
ADDRESS_CITY_NAMECINCINNATI       7.9669     0.2939   27.105  < 2e-16 ***
ADDRESS_CITY_NAMECLUTE           -5.9222     0.3984  -14.864  < 2e-16 ***
ADDRESS_CITY_NAMECOLLEGE STATION -12.0308     0.4081  -29.483  < 2e-16 ***
ADDRESS_CITY_NAMECOVINGTON        0.3699     0.3431    1.078 0.280918
ADDRESS_CITY_NAMECROWLEY         -1.6887     0.3991   -4.231 2.32e-05 ***
ADDRESS_CITY_NAMECYPRESS          2.5832     0.3934    6.566 5.18e-11 ***
```

```
ADDRESS_CITY_NAMEDALLAS              1.5938     0.3947    4.038 5.38e-05 ***
ADDRESS_CITY_NAMEDAYTON              1.0522     0.3419    3.077 0.002090 **
ADDRESS_CITY_NAMEDENTON             -1.6902     0.3990   -4.236 2.27e-05 ***
ADDRESS_CITY_NAMEDICKINSON         -10.1647     0.4031  -25.214  < 2e-16 ***
ADDRESS_CITY_NAMEDUNCANVILLE        -2.0991     0.4061   -5.168 2.36e-07 ***
ADDRESS_CITY_NAMEERLANGER           11.9744     0.3843   31.158  < 2e-16 ***
ADDRESS_CITY_NAMEFLOWER MOUND       -2.8965     0.3964   -7.306 2.75e-13 ***
ADDRESS_CITY_NAMEFRISCO             -7.1446     0.3951  -18.084  < 2e-16 ***
ADDRESS_CITY_NAMEGARLAND            -7.6230     0.4071  -18.725  < 2e-16 ***
ADDRESS_CITY_NAMEGOSHEN             -5.6751     0.3950  -14.368  < 2e-16 ***
ADDRESS_CITY_NAMEGRAND PRAIRIE     -10.5983     0.4254  -24.914  < 2e-16 ***
ADDRESS_CITY_NAMEHAMILTON            1.3056     0.3379    3.864 0.000111 ***
ADDRESS_CITY_NAMEHOUSTON            -5.3434     0.2971  -17.983  < 2e-16 ***
ADDRESS_CITY_NAMEINDEPENDENCE        2.1882     0.3898    5.614 1.98e-08 ***
ADDRESS_CITY_NAMEKATY               -7.7079     0.3455  -22.307  < 2e-16 ***
ADDRESS_CITY_NAMEKETTERING           1.6378     0.3457    4.738 2.16e-06 ***
ADDRESS_CITY_NAMEKINGWOOD            0.8557     0.3934    2.175 0.029612 *
ADDRESS_CITY_NAMELAWRENCEBURG        2.6869     0.3854    6.971 3.15e-12 ***
ADDRESS_CITY_NAMELEBANON            10.8313     0.3829   28.291  < 2e-16 ***
ADDRESS_CITY_NAMELOVELAND           -1.0750     0.3376   -3.185 0.001449 **
ADDRESS_CITY_NAMEMAGNOLIA           -4.6056     0.3991  -11.541  < 2e-16 ***
ADDRESS_CITY_NAMEMAINEVILLE          0.1304     0.3363    0.388 0.698158
ADDRESS_CITY_NAMEMASON               5.1345     0.3836   13.384  < 2e-16 ***
ADDRESS_CITY_NAMEMCKINNEY           -7.5089     0.3477  -21.597  < 2e-16 ***
ADDRESS_CITY_NAMEMESQUITE           -9.1778     0.4119  -22.282  < 2e-16 ***
ADDRESS_CITY_NAMEMIDDLETOWN          6.0693     0.3186   19.052  < 2e-16 ***
ADDRESS_CITY_NAMEMILFORD            -0.1301     0.3912   -0.333 0.739428
ADDRESS_CITY_NAMEPASADENA          -12.0655     0.4227  -28.545  < 2e-16 ***
ADDRESS_CITY_NAMERICHARDSON         -9.1371     0.4027  -22.688  < 2e-16 ***
ADDRESS_CITY_NAMEROCKWALL            2.5219     0.3947    6.389 1.67e-10 ***
ADDRESS_CITY_NAMESAINT MARYS         5.5399     0.3906   14.182  < 2e-16 ***
ADDRESS_CITY_NAMESHERMAN            -7.2353     0.4071  -17.774  < 2e-16 ***
ADDRESS_CITY_NAMESOUTHLAKE          -7.1978     0.3969  -18.134  < 2e-16 ***
ADDRESS_CITY_NAMESPRINGFIELD        -0.9301     0.3893   -2.389 0.016899 *
ADDRESS_CITY_NAMESUGAR LAND         -6.1288     0.3441  -17.809  < 2e-16 ***
ADDRESS_CITY_NAMETHE WOODLANDS      -5.6259     0.3961  -14.202  < 2e-16 ***
ADDRESS_CITY_NAMEVANDALIA            5.7624     0.3835   15.027  < 2e-16 ***
ADDRESS_CITY_NAMEWEST CHESTER        1.8501     0.3851    4.804 1.55e-06 ***
ADDRESS_CITY_NAMEWOODLANDS         -10.6018     0.4016  -26.402  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23.17 on 520726 degrees of freedom
Multiple R-squared:  0.4032,    Adjusted R-squared:  0.4031
F-statistic:  4820 on 73 and 520726 DF,  p-value: < 2.2e-16
```

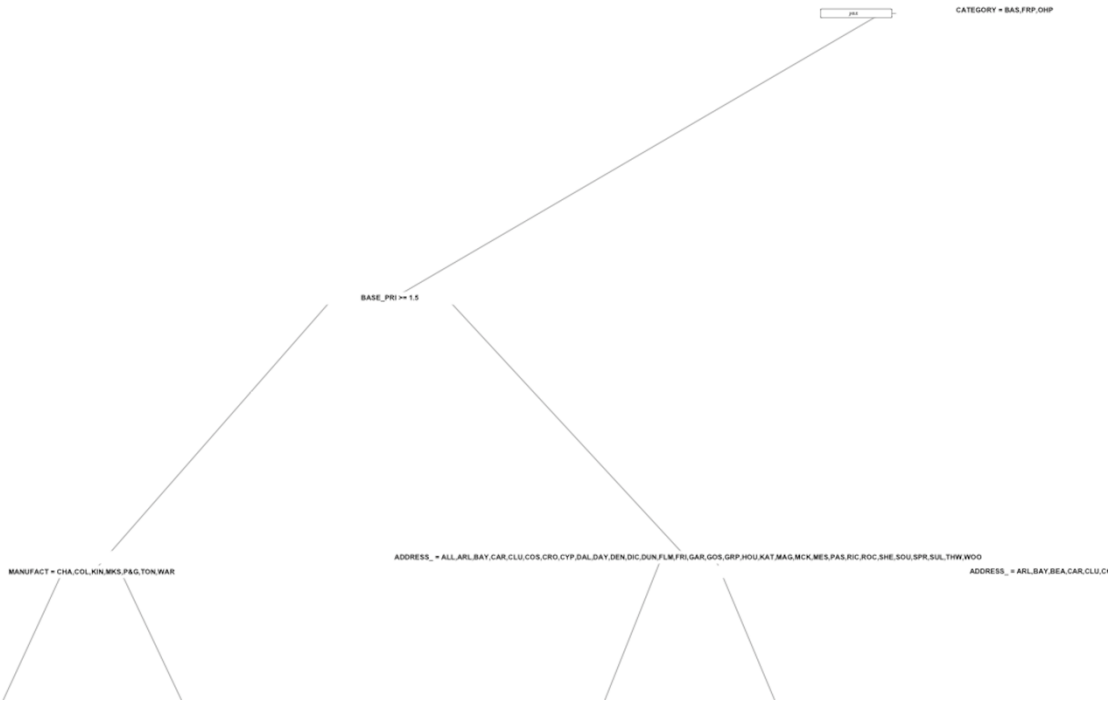
The above regression is hard to interpret because some variables have too many values, but we kept it because it performed better than other models in-sample.

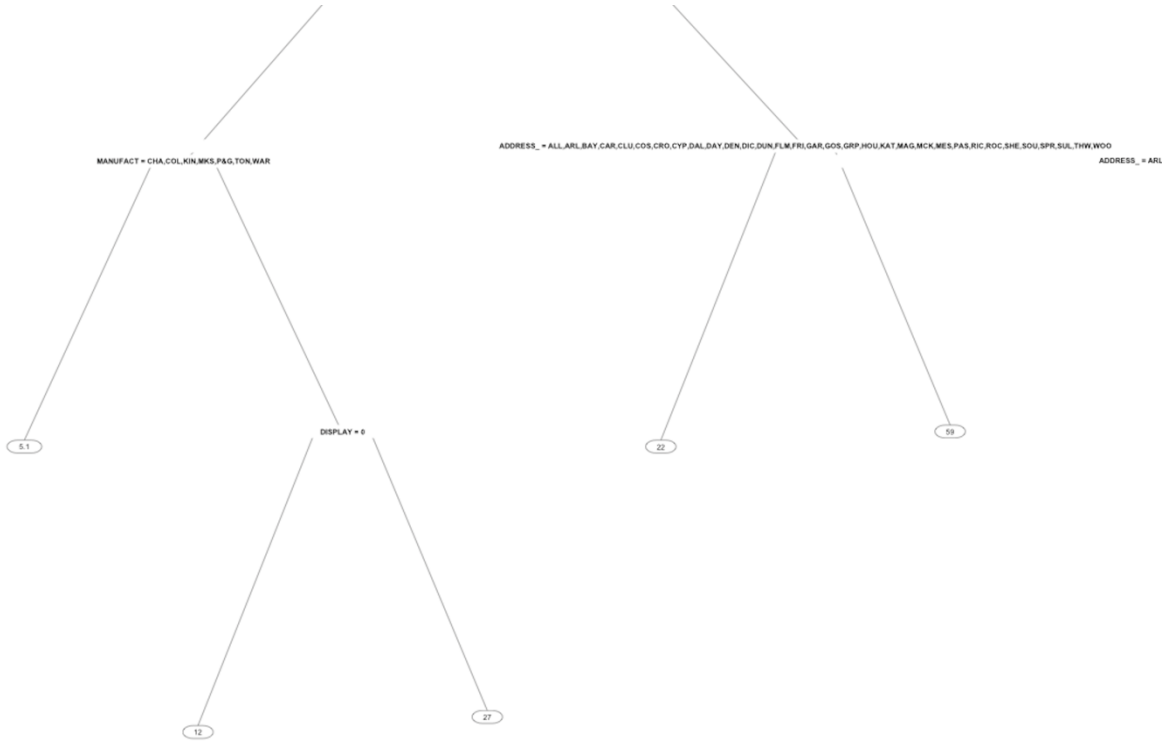
2. Regression Tree

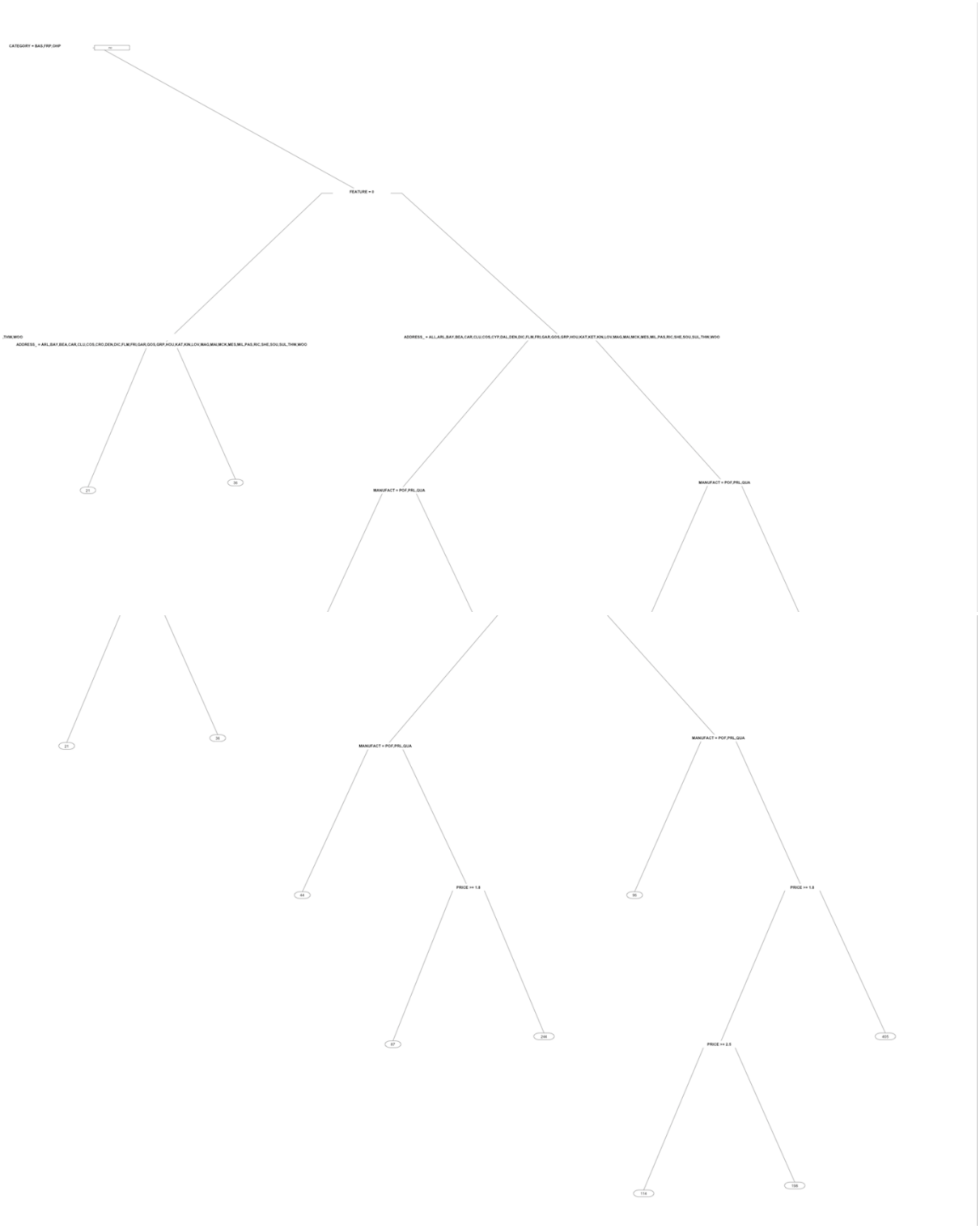




(As in the first tree some values were overlapping, we created the second one. Then we zoomed in to see clearly.)


CATEGORY = BAS,FRP,OHP
BASE_PRI >= 1.5
MANUFACT = CHA,COL,KIN,MKS,P&G,TON,WAR
ADDRESS_ = ALL,ARL,BAY,CAR,CLU,COS,CRO,CYP,DAL,DAY,DEN,DIC,DUN,FLM,FRI,GAR,GOS,GRP,HOU,KAT,MAG,MCK,MES,PAS,RIC,ROC,SHE,SOU,SPR,SUL,THW,WOO
ADDRESS_ = ARL,BAY,BEA,CAR,CLU,CI


MANUFACT = CHA,COL,KIN,MKS,P&G,TON,WAR
ADDRESS_ = ALL,ARL,BAY,CAR,CLU,COS,CRO,CYP,DAL,DAY,DEN,DIC,DUN,FLM,FRI,GAR,GOS,GRP,HOU,KAT,MAG,MCK,MES,PAS,RIC,ROC,SHE,SOU,SPR,SUL,THW,WOO
ADDRESS_ = ARL
5.1
DISPLAY = 0
22
59
12
27



From these results, we can interpret that CATEGORY, BASE_PRICE and FEATURE are the important variables affecting UNITS, and CATEGORY being the root node. It is different from the linear regression as the linear regression shows FEATURE as the most significant variable, whereas the regression tree above shows CATEGORY as the most significant variable.

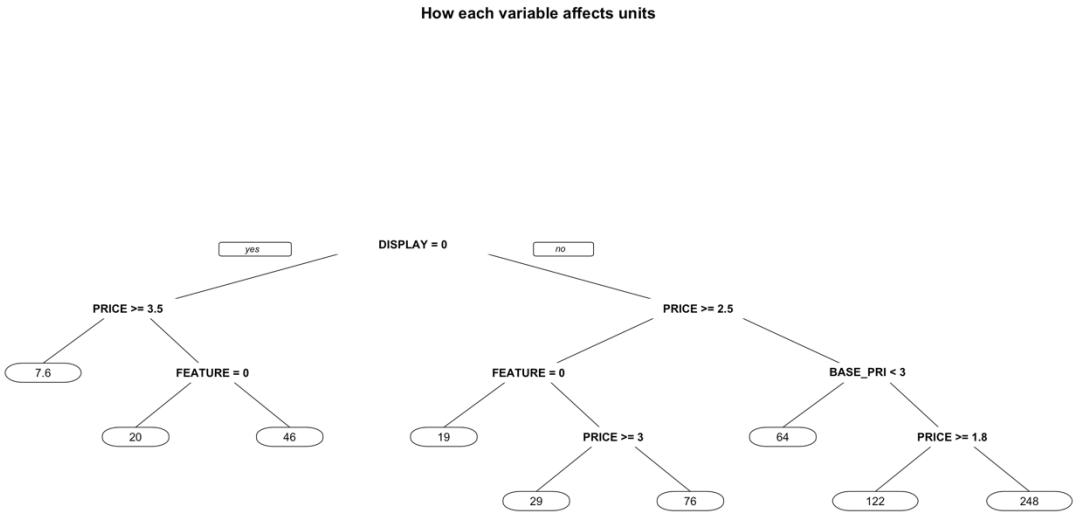
3.3.2 DEMAND FORECASTING: 5 variables
(DISPLAY + TPR_ONLY + FEATURE + PRICE + BASE_PRICE)

1.Linear Regression

```
Call:
lm(formula = UNITS ~ DISPLAY + TPR_ONLY + FEATURE + PRICE + BASE_PRICE
    data = train_p_s)

Residuals:
    Min      1Q  Median      3Q     Max
 -70.12  -12.78   -4.69    6.48 1733.76

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.04354    0.09767 338.321   <2e-16 ***
DISPLAY     19.94041    0.13839 144.091   <2e-16 ***
TPR_ONLY     0.26193    0.14288   1.833   0.0668 .
FEATURE     25.79246    0.18121 142.335   <2e-16 ***
PRICE       -2.38350    0.12377 -19.257   <2e-16 ***
BASE_PRICE  -2.70418    0.11731 -23.052   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26.79 on 520794 degrees of freedom
Multiple R-squared:  0.2024,    Adjusted R-squared:  0.2024
F-statistic: 2.643e+04 on 5 and 520794 DF,  p-value: < 2.2e-16
```

From these results, we can interpret that PRICE and BASE_PRICE have a negative relationship with UNITS, which supports the law of demand. Besides, TPR_ONLY is not significant for demand as the P-value is greater than 0.05. The key insight this regression provides is that Feature affects demand the most.

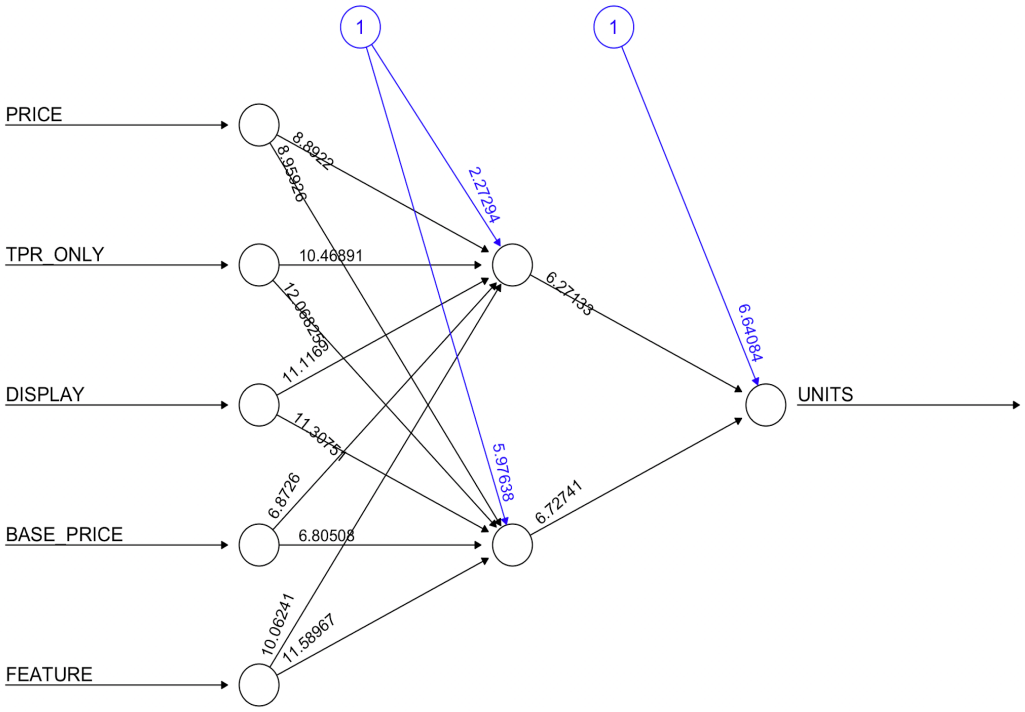
2. Regression Tree



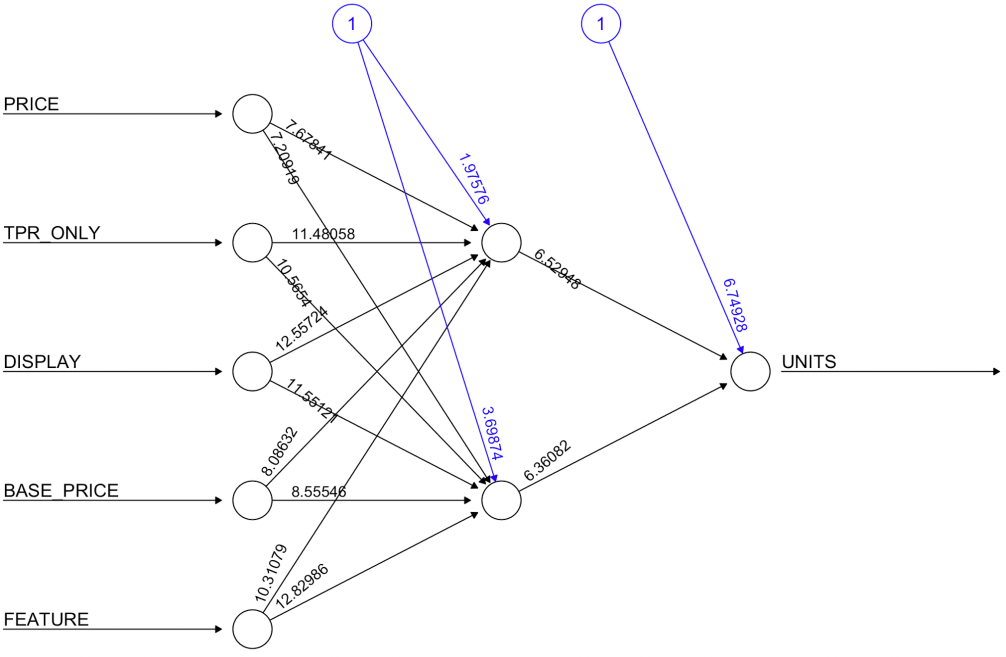
From these results, we can interpret that DISPLAY, PRICE, BASE_PRICE and FEATURE are the important variables affecting UNITS, and DISPLAY being the root node. A major difference can be compared with the 8-variables regression tree, which shows CATEGORY as the most important variable.

3. Neural Network

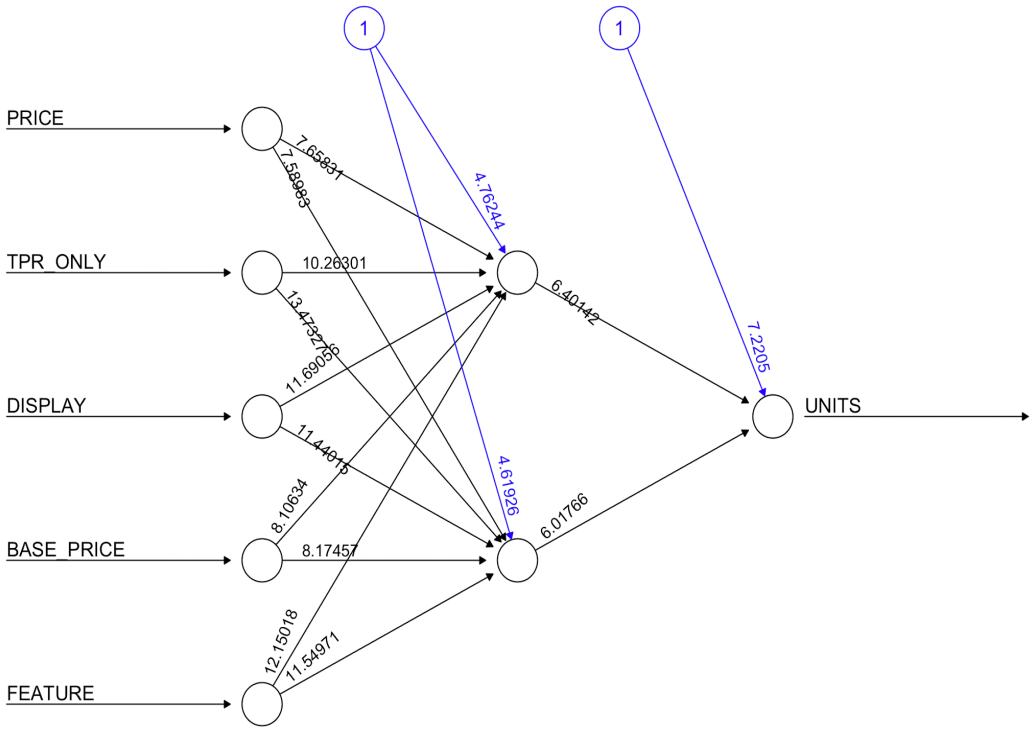
After multiple tests on the neural networks, we found that these 5 variables are significant in lowering the error. We limit the neural network to 5 variables as it takes an extremely long time to generate in R. And we choose numeric variables as it is easier to process than characters. Although we successfully generated rep=8, it only varies the error by a few decimal points and in order to prevent excessive time on loading the neural network, we decided to present it with rep=3. As we increase the number of hidden layers and neurons, the data took a long processing time where we would only focus on 1 hidden layer and 2 neurons.



Error: 234264059.772953 Steps: 116



Error: 234264059.77235 Steps: 113



Error: 234264059.773782 Steps: 145

3.3.3 DEMAND FORECASTING: Results and Predictions

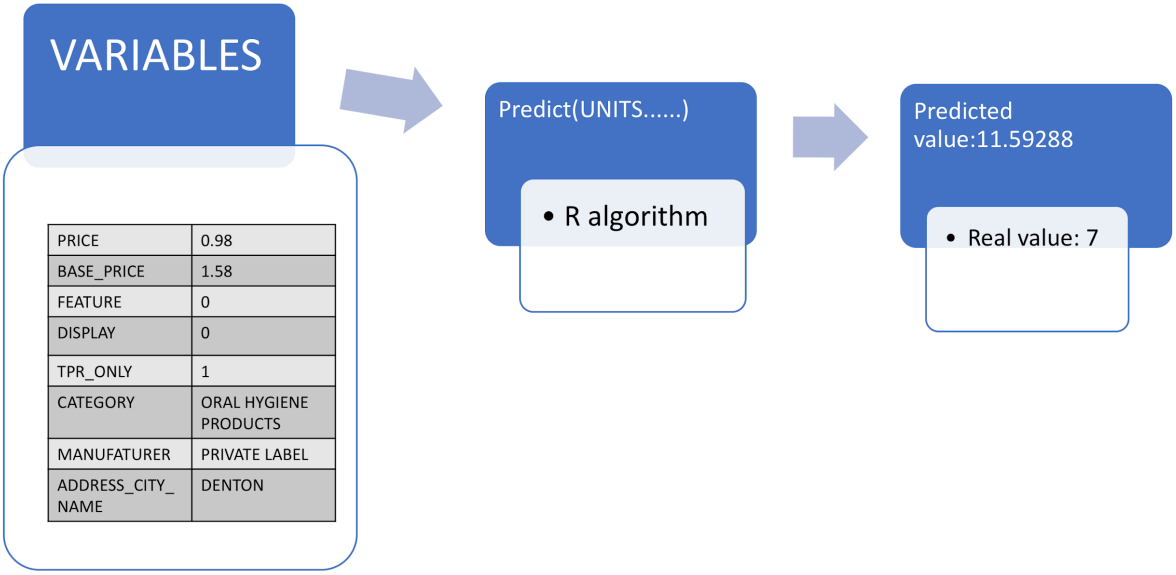
Initially, we had 3 models: linear regression, regression tree with 8 variables and neural network with 5 variables. Then, we included linear regression and regression tree with 5 variables to check the importance of the number of variables in model performance and to compare models with the same number of variables. We kept the models with 8 variables as well because they were the most accurate.

	In-sample RMSE	Predictive RMSE
Linear regression (8 variables)	23.17051	15.38453
Regression tree (8 variables)	21.7764	16.36717
Linear regression (5 variables)	26.78756	18.55336
Regression tree (5 variables)	25.51073	19.27102
Neural network (5 variables)	29.99386	20.7308

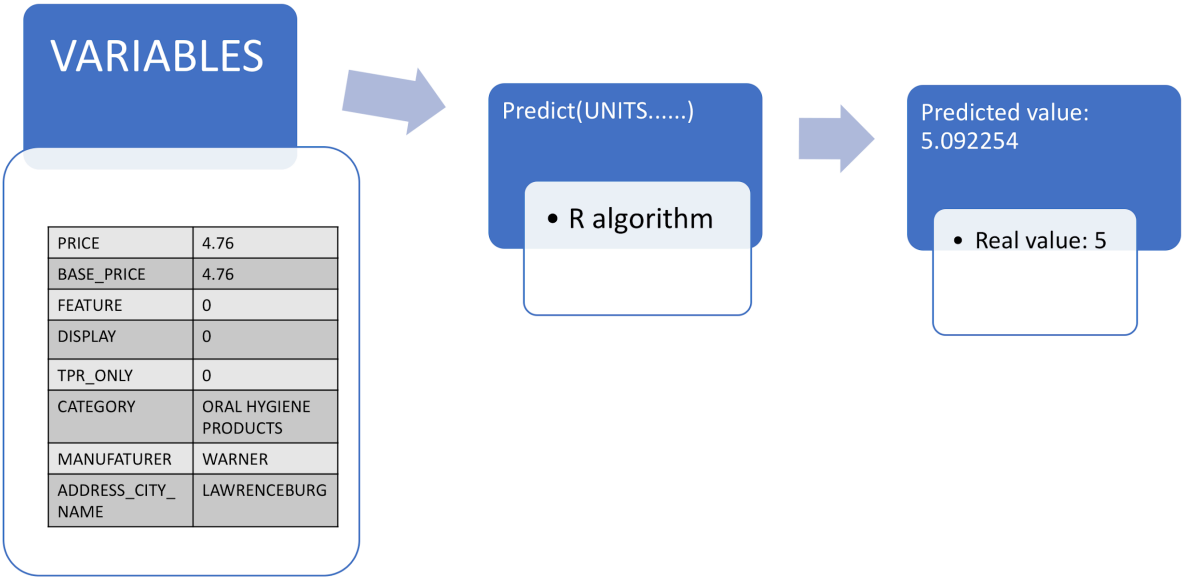
From the above table we can see that the best model (overall and in-sample) is the regression tree with 8 variables. The best model out-of-sample is the linear regression with 8 variables. One explanation could be that a prediction with 8 variables is more reliable than a prediction with 5 variables if the 3 added variables are significant. The worst model (overall) is the neural-network. So advanced models do not always perform better than the basic models.

Best model: Regression tree (8 variables)
We can apply the model to generate insights for managers. If we introduce values for the 8 variables, the model will predict the UNITS (demand) generated from those.
We can see below 2 examples:

Select a random row from **train** and introduce the variables:



Select a random row from **test** and introduce the variables:



APPENDIX:

DICTIONARY:

Variables explained:

dh Transaction Data

Description: This table contains a sample of 156 weeks of mouthwash, pretzels, frozen pizza, and boxed cereal transactions, at the product level by store, by week.	
# of Records: 524,950	
VARIABLE NAME	**DESCRIPTION**
BASE_PRICE	base price of item
DISPLAY	product was a part of in-store promotional display
FEATURE	product was in in-store circular
HHS	# of purchasing households
PRICE	actual amount charged for the product at shelf
SPEND	total spend (i.e., $ sales)
STORE_NUM	store number
TPR_ONLY	temporary price reduction only (i.e., shelf tag only, product was reduced in price but not on display or in an advertisment)
UNITS	units sold
VISITS	number of unique purchases (baskets) that included the product
WEEK_END_DATE	week ending date
UPC	(Universal Product Code) product specific identifier

dh Product Lookup

Description: Provides detailed product information for each upc in 'dh Transaction Data'.	
# of Records: 58	
VARIABLE NAME	**DESCRIPTION**
CATEGORY	category of product
DESCRIPTION	prodcut description
MANUFACTURER	manufacturer
PRODUCT_SIZE	package size or quantity of product
SUB_CATEGORY	sub-category of product
UPC	(Universal Product Code) product specific identifier

dh Store Lookup

Description: Provides detailed store information for each store in 'dh Transaction Data'.	
# of Records: 79	
VARIABLE NAME	**DESCRIPTION**
ADDRESS_CITY_NAME	city
ADDRESS_STATE_PROV_CODE	state
AVG_WEEKLY_BASKETS	average weekly baskets sold in the store
MSA_CODE	(Metropolitan Statistical Area) geographic region with a high core population density and close economic ties throughout the surrounding areas
PARKING_SPACE_QTY	number of parking spaces in the Kroger parking lot
SALES_AREA_SIZE_NUM	square footage of Kroger store
STORE_APPEAL	Kroger's designated store appeal
STORE_NUM	store number