Orange Juice Regression Model

Ahmad Mudrik

2024-06-03

Orange Juice Regression Model

Introduction

Data set showcases orange juice sales from Dominick's grocery stores. The data includes weekly prices, sales, and feat (Boolean indicator if brand was advertised in store or flyer) for 3 brands of orange juice: Tropicana, Minute Maid and Dominick's. This study aim to to evaluate the sale performances for orange juice brands at 3 different price points (cheap, mid, and expensive)

Hypothesis

1. Customers are more prone to buy mid-priced orange juice than the cheap cheap and expensive orange juices.
2. The price of the product affects customer's purchasing behavior.

Objective

1. To evaluate prices of every brands using a scatter plot plot a regression model of orange juice sales for 3 brands of orange juice

Install packages, import and inspect data set

Import Dataset

```
Orange_Juice <- read.csv("D:\\2. Analytics Data Sets\\BDS Datasets\\oj.csv")
head(Orange_Juice)
```

```
##   sales price     brand feat
## 1  8256  3.87 tropicana    0
## 2  6144  3.87 tropicana    0
## 3  3840  3.87 tropicana    0
## 4  8000  3.87 tropicana    0
## 5  8896  3.87 tropicana    0
## 6  7168  3.87 tropicana    0
```

- The data set includes weekly prices for 3 orange juice brands: Tropicana, Minute Maid, and Dominick's.

Level brands

```
Orange_Juice$brand <- factor(Orange_Juice$brand)
levels(Orange_Juice$brand)
```

```
## [1] "dominicks"   "minute.maid" "tropicana"
```

```r
head(Orange_Juice)
```

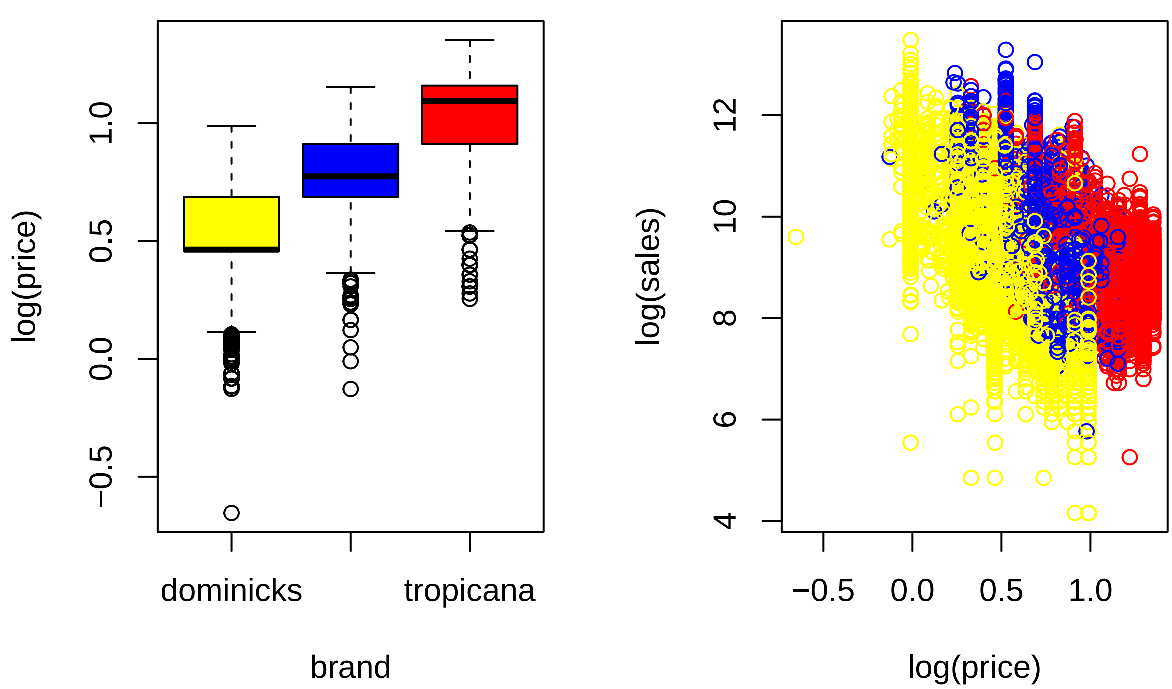
```
##   sales price     brand feat
## 1  8256  3.87 tropicana    0
## 2  6144  3.87 tropicana    0
## 3  3840  3.87 tropicana    0
## 4  8000  3.87 tropicana    0
## 5  8896  3.87 tropicana    0
## 6  7168  3.87 tropicana    0
```

```r
brandcol <- c("yellow","blue","red")
```

- Each brand occupies a price range: Dominick's (cheap), Minute Maid (mid), and Tropicana (expensive).
- Relationship is defined by assigning colors for each brands
- yellow, blue, and red from cheapest to most expensive

Study 1: Brand Prices

```r
par(mfrow=c(1,2))
plot(log(price) ~ brand, data=Orange_Juice, col=brandcol)
plot(log(sales) ~ log(price), data=Orange_Juice, col=brandcol[Orange_Juice$brand])
```



Findings from Study 1

- Each brand occupies a well-defined range
- Dominick's a cheap orange juice option, Minute Maid is mid, and Tropicana is the expensive option.

- log sales has a linear relationship with log price.
- Charging more means selling less.

Study 2: Regression Model

```
reg <- glm(log(sales) ~ log(price) + brand, data=Orange_Juice)
summary(reg)
```

```
##
## Call:
## glm(formula = log(sales) ~ log(price) + brand, data = Orange_Juice)
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      10.82882    0.01453  745.04   <2e-16 ***
## log(price)       -3.13869    0.02293 -136.89   <2e-16 ***
## brandminute.maid  0.87017    0.01293   67.32   <2e-16 ***
## brandtropicana    1.52994    0.01631   93.81   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.6296804)
##
##     Null deviance: 30079  on 28946  degrees of freedom
## Residual deviance: 18225  on 28943  degrees of freedom
## AIC: 68765
##
## Number of Fisher Scoring iterations: 2
```

```
coef(reg)
```

```
##      (Intercept)       log(price) brandminute.maid   brandtropicana
##       10.8288216       -3.1386914        0.8701747        1.5299428
```

```
beta <- coef(reg)
beta
```

```
##      (Intercept)       log(price) brandminute.maid   brandtropicana
##       10.8288216       -3.1386914        0.8701747        1.5299428
```

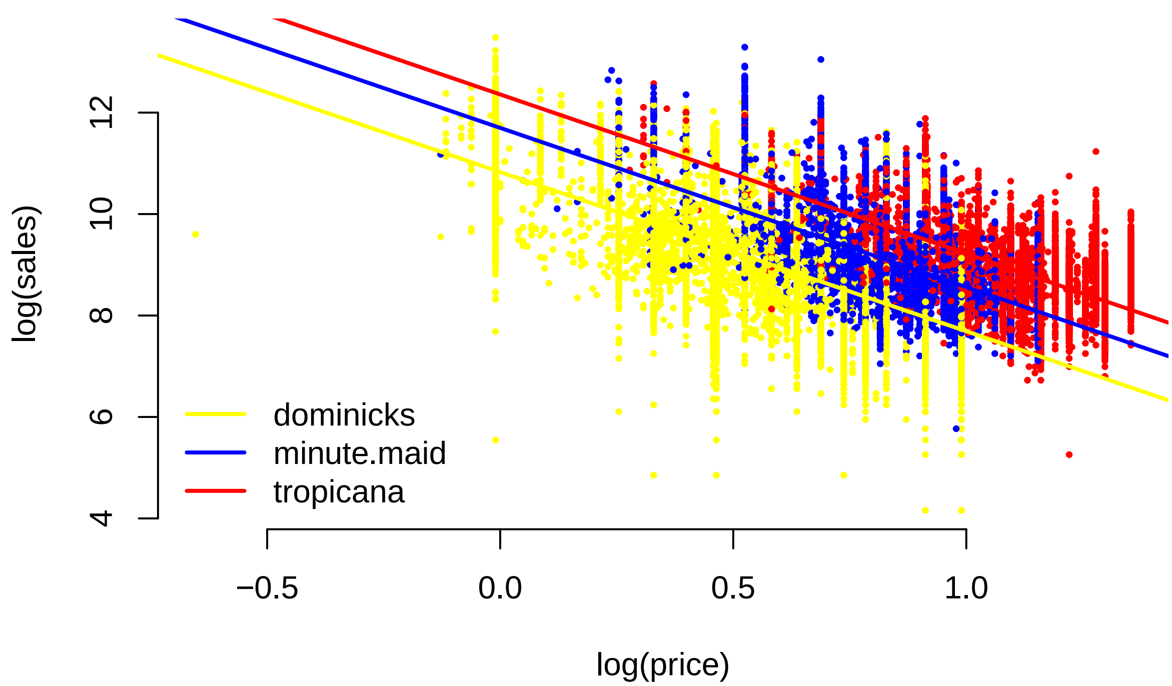
Findings from Study 2

glm (generalized linear model) was used, and log(sales) vs. log(price) was plotted, brand were assigned colors - used **model.matrix** to create model matrix that defines the numeric inputs x - used summary(reg) to access coefficient, test and fit - used coef(reg) to obtain coefficients - beta = -3.13 for the log price effect - sales drop by about 3.13% for every 1% price hike across all brands

Study 3: Same Price Sensitivity

```
plot(log(sales) ~ log(price),
     data=Orange_Juice,
     col=brandcol[Orange_Juice$brand],
     cex=.5,
     pch=20,
     bty="n")
```

```r
abline(a=beta[1],
       b=beta[2],
       col=brandcol[1],
       lwd=2)
abline(a=beta[1]+ beta[3],
       b=beta[2],
       col=brandcol[2],
       lwd=2)
abline(a=beta[1]+beta[4],
       b=beta[2],
       col=brandcol[3],
       lwd=2)
legend("bottomleft",
       bty="n",
       lwd=2,
       col=brandcol,
       legend=levels(Orange_Juice$brand))
```



Findings from Study 3

- 3 lines shifted according to brand identity
- At the same price, Tropicana (expensive) sells more than Minute Maid (mid), and Minute maid sells more than Dominick's (cheap)
- This makes sense: Tropicana is a luxury product that is preferable at the same price.
- All lines have the same slope
- In economic terms, the model assumes consumers of the three bands have the same price sensitivity.

- This is unrealistic; money is less of an issue for Tropicana consumers than it is for Dominick's consumer.
- Next is to build this information into Study 4 by having **log(price)** interact with **brand**

Study 4:: Price Sensitivity Interacts with Brand Indicator

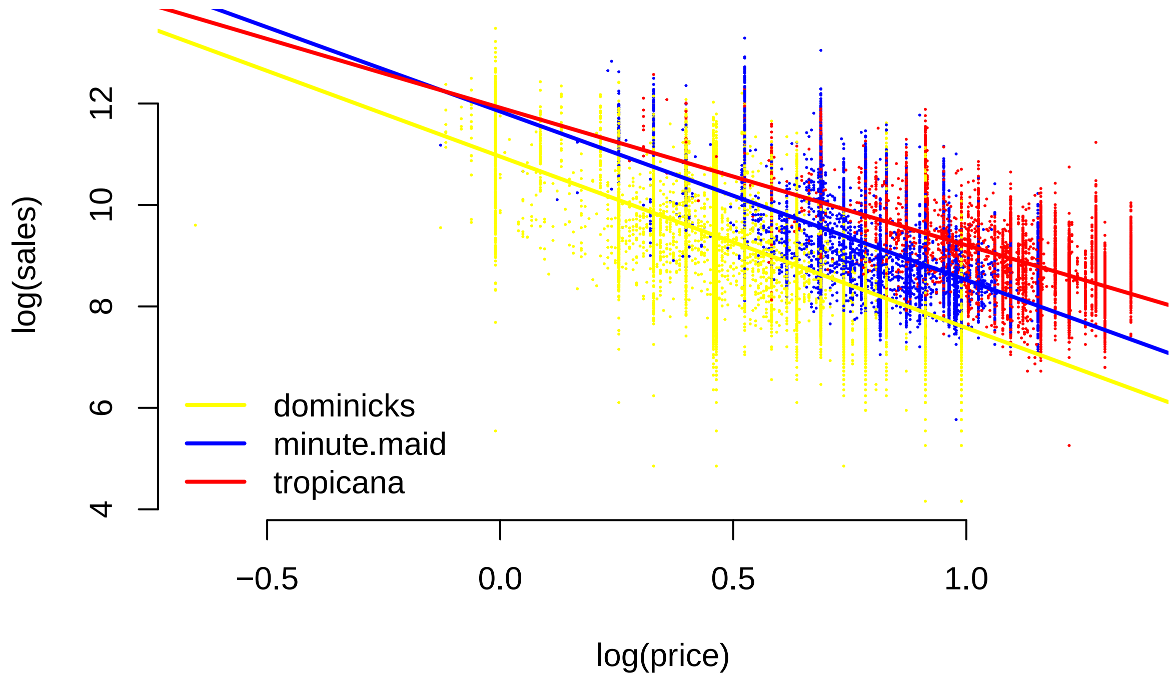
```
reg_interact <- glm(log(sales) ~ log(price)*brand,
                    data = Orange_Juice)
coef(reg_interact)
```

```
##                 (Intercept)                  log(price)
##                 10.95468173                 -3.37752963
##             brandminute.maid              brandtropicana
##                  0.88825363                  0.96238960
## log(price):brandminute.maid   log(price):brandtropicana
##                  0.05679476                  0.66576088
```

```
beta <- coef(reg_interact)
beta
```

```
##                 (Intercept)                  log(price)
##                 10.95468173                 -3.37752963
##             brandminute.maid              brandtropicana
##                  0.88825363                  0.96238960
## log(price):brandminute.maid   log(price):brandtropicana
##                  0.05679476                  0.66576088
```

```
plot(log(sales) ~ log(price),
     data=Orange_Juice,
     col=brandcol[Orange_Juice$brand],
     cex=.1,
     pch=20,
     bty="n")
abline(a=beta[1],
       b=beta[2],
       col=brandcol[1],
       lwd=2)
abline(a=beta[1]+beta[3],
       b=beta[2]+beta[5],
       col=brandcol[2],
       lwd=2)
abline(a=beta[1]+beta[4],
       b=beta[2]+beta[6],
       col=brandcol[3],
       lwd=2)
legend("bottomleft",
       bty="n",
       lwd=2,
       col=brandcol,
       legend=levels(Orange_Juice$brand))
```



Findings from Study 4

- Adding " '*' " inside the "glm" function interacts log price with brand, which influences the main slope
- log price for brand Tropicana, Minute Maid, and Dominick's is -2.7, -3.3 and -3.4, respectively.
- Tropicana brand are less price sensitive, meaning that the Tropicana core customers are less bothered by price changes

Study 5:: Price Sensitivity Interacts with Brand Indicator and Feat

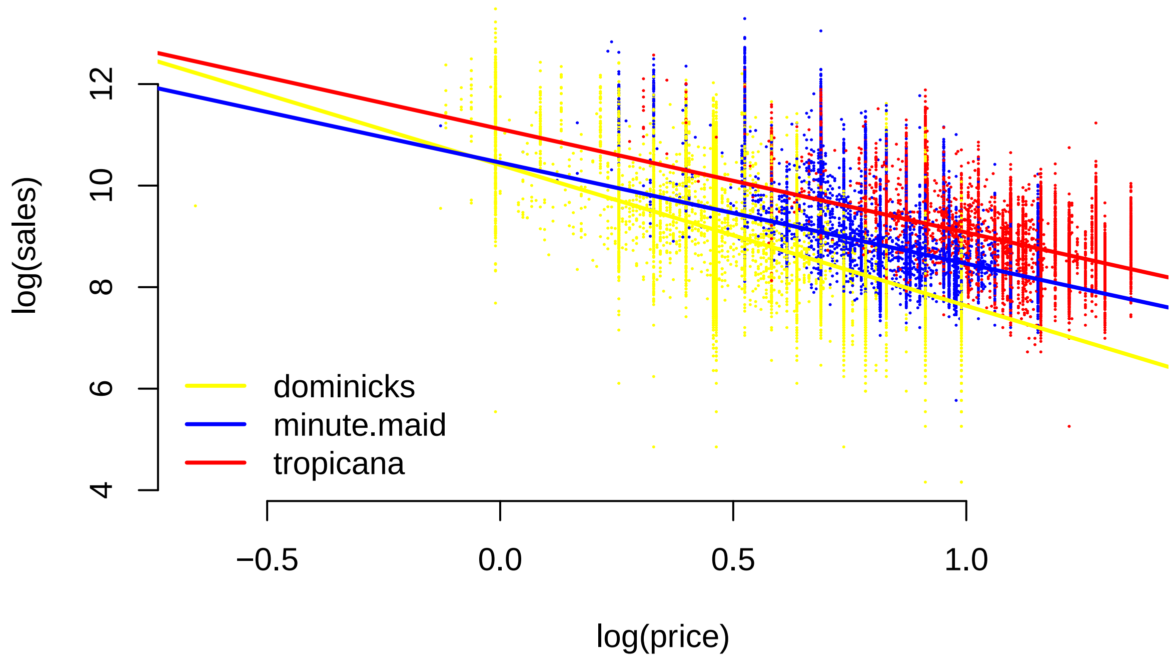
```
ojreg <- glm(log(sales) ~ log(price)*brand*feat, data=Orange_Juice)
coef(ojreg)
```

```
##                     (Intercept)                      log(price)
##                     10.40657579                     -2.77415436
##                 brandminute.maid                  brandtropicana
##                      0.04720317                      0.70794089
##                            feat      log(price):brandminute.maid
##                      1.09440665                      0.78293210
##       log(price):brandtropicana                 log(price):feat
##                      0.73579299                     -0.47055331
##           brandminute.maid:feat             brandtropicana:feat
##                      1.17294361                      0.78525237
## log(price):brandminute.maid:feat  log(price):brandtropicana:feat
##                     -1.10922376                     -0.98614093
```

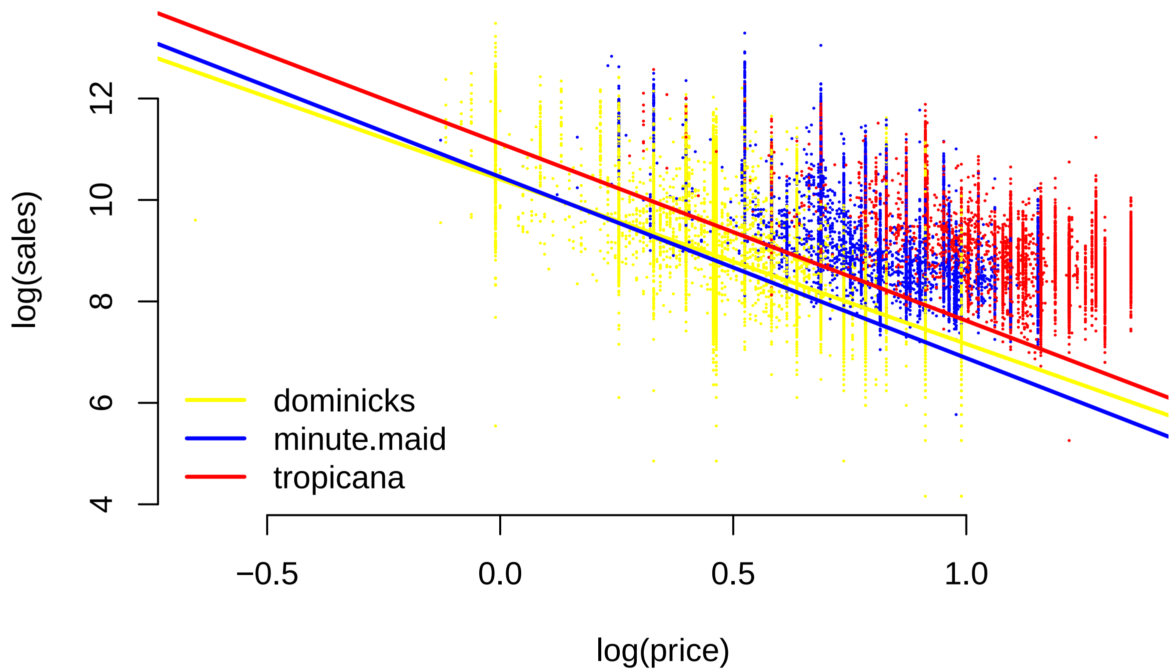
```r
beta <- coef(ojreg)
beta
```

```
##                     (Intercept)                      log(price) 
##                     10.40657579                     -2.77415436 
##                brandminute.maid                  brandtropicana 
##                      0.04720317                      0.70794089 
##                            feat      log(price):brandminute.maid 
##                      1.09440665                      0.78293210 
##       log(price):brandtropicana                 log(price):feat 
##                      0.73579299                     -0.47055331 
##           brandminute.maid:feat             brandtropicana:feat 
##                      1.17294361                      0.78525237 
## log(price):brandminute.maid:feat   log(price):brandtropicana:feat 
##                     -1.10922376                     -0.98614093
```

```r
# Without factoring feat
plot(log(sales) ~ log(price),
     data=Orange_Juice,
     col=brandcol[Orange_Juice$brand],
     cex=.1,
     pch=20,
     bty="n")
abline(a=beta[1],
       b=beta[2],
       col=brandcol[1],
       lwd=2) #Dominick's
abline(a=beta[1]+beta[3],
       b=beta[2]+beta[6],
       col=brandcol[2],
       lwd=2) # Minute Maid
abline(a=beta[1]+beta[4],
       b=beta[2]+beta[7],
       col=brandcol[3],
       lwd=2) # Tropicana
legend("bottomleft",
       bty="n",
       lwd=2,
       col=brandcol,
       legend=levels(Orange_Juice$brand))
```




```
# With factoring feat
plot(log(sales) ~ log(price),
     data=Orange_Juice,
     col=brandcol[Orange_Juice$brand],
     cex=.1,
     pch=20,
     bty="n")
abline(a=beta[1],
       b=beta[2] + beta[8],
       col=brandcol[1],
       lwd=2) #Dominick's
abline(a=beta[1]+beta[3],
       b=beta[2]+beta[6] + beta[8] + beta[11],
       col=brandcol[2],
       lwd=2) # Minute Maid
abline(a=beta[1]+beta[4],
       b=beta[2]+beta[7] + beta[8]+ beta[12],
       col=brandcol[3],
       lwd=2) # Tropicana
legend("bottomleft",
       bty="n",
       lwd=2,
       col=brandcol,
       legend=levels(Orange_Juice$brand))
```



Findings from Study 5

- Note that feat is Boolean, indicating if product had been advertised or not. Adding feat into the "glm" equation evaluates the influence of advertising in the relationship between sales and prices.
- Featured products leads to price sensitivity
- Minute Maid elasticity dropped from -2.0 to -3.5 with ads
- Tropicana elasticity dropped from -2.0 to -3.6
- Dominick's elasticity dropped from -2.8 to -3.2

Marketing efforts can expand consumer base to include people who are more price sensitive,attracts consumers beyond brand loyalists. Therefore ad campaigns must be accompanied by **price cuts**. Since featured products are often discounted, it could be that the demand curve is non-linear. At lower price points, the average customers is more price sensitive.

Study 6: Elasticity Table

```
b <- coef(ojreg)
b["log(price)"]
```

```
## log(price)
##  -2.774154
```

```
b["log(price)"] + b["log(price):brandminute.maid"]
```

```
## log(price)
##  -1.991222
```

```
b["log(price)"] + b["log(price):brandtropicana"]
```

```
## log(price)
##  -2.038361
```

```
b["log(price)"] + b["log(price):feat"]
```

```
## log(price)
##  -3.244708
```

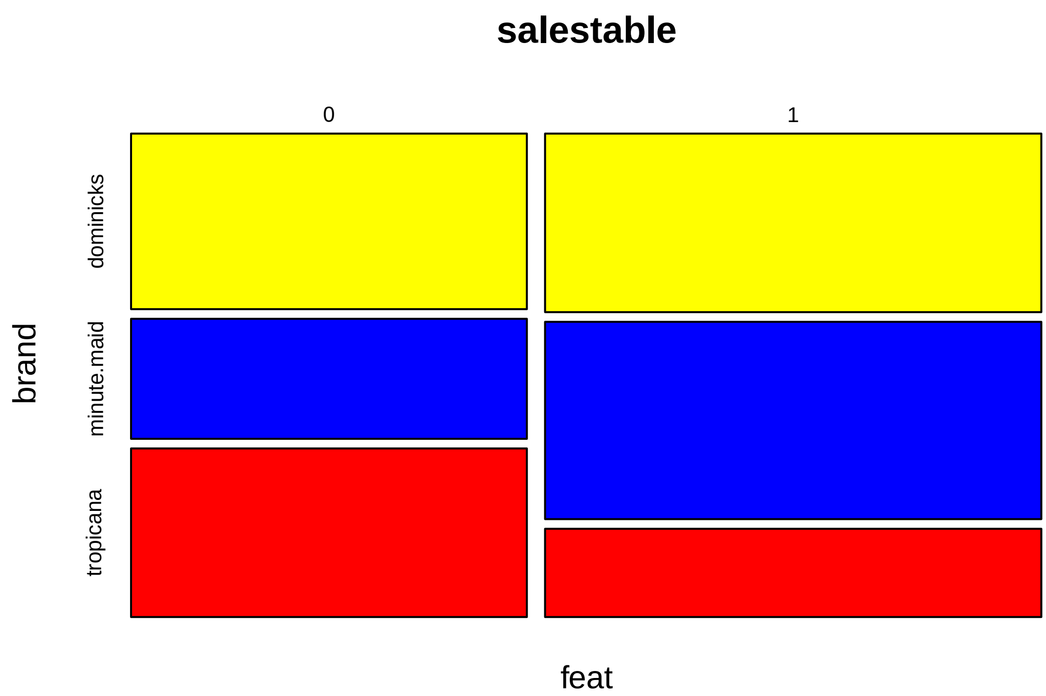
```
b["log(price)"] + b["log(price):brandminute.maid"] + b["log(price):feat"] + b["log(price):brandminute.ma
```

```
## log(price)
##  -3.570999
```

```
b["log(price)"] + b["log(price):brandtropicana"] + b["log(price):feat"] + b["log(price):brandtropicana::
```

```
## log(price)
##  -3.495056
```

```
salestable <- tapply(Orange_Juice$sales, Orange_Juice[,c("feat","brand")], sum)
mosaicplot(salestable,col=brandcol)
```



Findings from Study 6

- Mosaic plot of the amount of advertisement by brand
- Minute Maid is featured more than Dominick's and Tropicana (Minute Maid < Dominick's < Tropicana)
- Minute Maid had similar price elasticity as Tropicana; it behaved like an expensive product

- This is due to contradicting effect between brands and advertisement.