Final Exam: December 10, 9:30am - 12:30pm

Note: This is an individual exam that you should complete on your own. Make sure to push your answers to GitLab before 12:30pm. Exams with time stamps after this time will receive penalties. Good luck!

The data contained in the "data" folder is real data from a research company and you should not share it with individuals outside of class. The data are one year of sales data for a a large number of grocery stores for products in the paper towel category. There are three data files:

- **sales.rds**: Each row is data on sales for a (store,week,product) combination. For more information see the salesINFO.txt file.
- **products.rds**: Each row is information on one product (UPC = Universal Product Code. This is a unique bar code for each product). For more information see the productINFO.txt file.
- **stores.rds**: Each row is information on one store. For more information see the storesINFO.txt file. Note that you have information of when each store opened (and whether or not it closed during the 52 week period). Note also that some stores close and then reopen later under a different store chain name. Due to this phenomenon there are more rows in the store file than there are stores (i.e., some stores occur twice).

Manufacturers and grocery store managers use data like this to understand preferences for different types of product in category, how these preferences vary over time and location, and how sales can be impacted by marketing strategies such as price promotions.

Q1 (5 points)

First look at total sales in dollars for all stores combined. Which week (report the week index WEEK) had the highest sales in dollars and what was the dollars sales that week?

```
library(tidyverse)
sales = readRDS('data/sales.rds')
stores = readRDS('data/stores.rds')
products = readRDS('data/products.rds')
sales %>%
  group_by(WEEK) %>%
  summarize(total = sum(DOLLARS)) %>%
  arrange(desc(total)) %>%
  slice(1:5)
```

```
## # A tibble: 5 x 2
##    WEEK    total
##   <int>    <dbl>
## 1  1584 3403374.
## 2  1583 3247411.
## 3  1614 3206245.
## 4  1593 3109939.
## 5  1586 3074152.
```

Week 1584 had the highest sales in dollars and the dollars sales that week was 3403374.

Q2 (5 points)

A **brand** is made up of multiple products (UPCs). The field L5 in the products file denotes the brand of a UPC. How many UPCs are sold under the brand name "BOUNTY" and how many are sold under the brand name "SCOTT"? (note: consider ONLY brands with the name "BOUNTY" and "SCOTT". For example, "BOUNTY BASIC" is a different brand from "BOUNTY").

```
products %>%
    filter(L5=='BOUNTY')%>%
  count('UPC')
```

```
## # A tibble: 1 x 2
##   `"UPC"`     n
##   <chr>   <int>
## 1 UPC       444
```

```
products %>%
    filter(L5=='SCOTT')%>%
  count('UPC')
```

```
## # A tibble: 1 x 2
##   `"UPC"`     n
##   <chr>   <int>
## 1 UPC        85
```

444 UPCs are sold under the brand name "BOUNTY" and 85 are sold under the brand name "SCOTT".

Q3 (6 points)

What are the five biggest brands of paper towels in terms of dollar sales for the whole year?

```
sales_products<-sales %>%
    left_join(products, by='UPC')

sales_products%>%
    group_by(L5)%>%
    summarize(total = sum(DOLLARS)) %>%
    arrange(desc(total)) %>%
    slice(1:5)
```

```
## # A tibble: 5 x 2
##   L5                total
##   <chr>             <dbl>
## 1 BOUNTY        57664342.
## 2 PRIVATE LABEL 40831029.
## 3 KLEENEX VIVA   8605338.
## 4 BRAWNY         8408807.
## 5 SCOTT          7282180.
```

The five biggest brands of paper towels in terms of dollar sales for the whole year are BOUNTY, PRIVATE LABEL, KLEENEX VIVA, BRAWNY and SCOTT.

Q4 (10 points)

What are the five biggest brands of paper towels in terms of volume, i.e., amount of paper towel? Note: Different products/UPCs contain different amounts of paper towel, i.e., if you sell one unit of a UPC with

12 rolls of paper towel, you sell more volume than if you sell a unit of a UPC with only 2 rolls of paper towel. The field VOL_EQ in the products file allows you to compute a normalized quantity of volume across products. The bigger VOL_EQ is the more paper towel the product contain. Therefore, you can think of UNITS*VOL_EQ as the sales volume for a product in "standardized" units that you can compare across UPCs.

```
sales_products <- sales_products %>%
    mutate(sales.volumne = UNITS*VOL_EQ)

sales_products%>%
    group_by(L5)%>%
    summarize(total = sum(sales.volumne)) %>%
    arrange(desc(total)) %>%
    slice(1:5)
```

```
## # A tibble: 5 x 2
##   L5                total
##   <chr>             <dbl>
## 1 PRIVATE LABEL 23941675.
## 2 BOUNTY        22518344.
## 3 SCOTT          3900713.
## 4 BRAWNY         3822108.
## 5 SPARKLE        3411591.
```

PRIVATE LABEL, BOUNTY, SCOTT, BRAWNY and SPARKLE are the five biggest brands of paper towels in terms of volume.

Q5 (5 points)

How would explain the fact that Bounty is biggest in dollar sales but "Private Label" is biggest in volume? (a one line answer is enough)

Bounty is more expensive than Private Label.

Q6 (10) points)

EST_ACV in the stores file is a measure of the size of a store, basically a proxy for store sales (in total across all product categories) with bigger EST_ACV implying higher overall store sales. Can you verify this using the paper towel sales data?

```
sales_stores <- sales %>%
    left_join(stores, by='IRI_KEY')

top_EST_ACV<-stores%>%
    arrange(desc(EST_ACV))%>%
    slice(1:5)%>%
    select(IRI_KEY,EST_ACV)

top_store_sales<-sales_stores%>%
    group_by(IRI_KEY)%>%
    summarize(total = sum(DOLLARS))

left_join(top_EST_ACV,top_store_sales)
```

```
## # A tibble: 5 x 3
```

```
##   IRI_KEY EST_ACV   total
##     <int>   <dbl>   <dbl>
## 1  262860    146. 511582.
## 2  205512    124. 617593.
## 3  258282    103. 439347.
## 4  931001    103. 455002.
## 5  283639    101. 426800.
```

Comparison of EST_ACV and overall store dollar sales in paper towel for the five stores with the highest EST_ACV shows no evidence that bigger EST_ACV implies higher overall store sales in paper towel.

Q7 (7 points)

What are the top 20 UPCs in terms of volume (where volume again is measured in standardized units (UNITS*VOL_EQ))?

```
top20_UPC_volume<-sales_products%>%
    group_by(UPC)%>%
    summarize(total = sum(sales.volumne)) %>%
    arrange(desc(total)) %>%
    slice(1:20)
```

Q8 (10 points)

What are the five most and five least expensive UPCs among the top 20 you found in Q7?

- Note: Here you should measure the "expensiveness" of a UPC as the **average price per standardized unit** over all weeks and stores. (Hint: The price of a UPC will change from week to week in a given store and vary between stores within a week. You should calculate the average price that was charged across all units sold of the UPC - taking into account that stores with low prices may sell more units than stores with high prices. For example, if one store sells 250 standardized units at a price of \$1 per standardized unit and another store sells 100 standardized units at \$1.5, then the average price charged across all units sold is NOT $0.5\$1 + 0.5\$1.5 = \$1.25$ since there were many more units sold at \$1 than at \$1.5)

```
sales_products <- sales_products %>%
    mutate(price_per_su = DOLLARS/sales.volumne)

top20_UPC_volume%>%left_join(sales_products)%>%
    group_by(UPC)%>%
    summarize(average = mean(price_per_su)) %>%
    arrange(desc(average)) %>%
    slice(1:5)
```

```
## # A tibble: 5 x 2
##   UPC              average
##   <chr>              <dbl>
## 1 0_1_37000_28831    2.97
## 2 0_1_37000_28848    2.92
## 3 0_1_37000_28839    2.80
## 4 0_1_37000_28876    2.79
## 5 0_1_37000_28857    2.68
```

```
top20_UPC_volume%>%left_join(sales_products)%>%
    group_by(UPC)%>%
    summarize(average = mean(price_per_su)) %>%
```

```
    arrange(average) %>%
    slice(1:5)
```

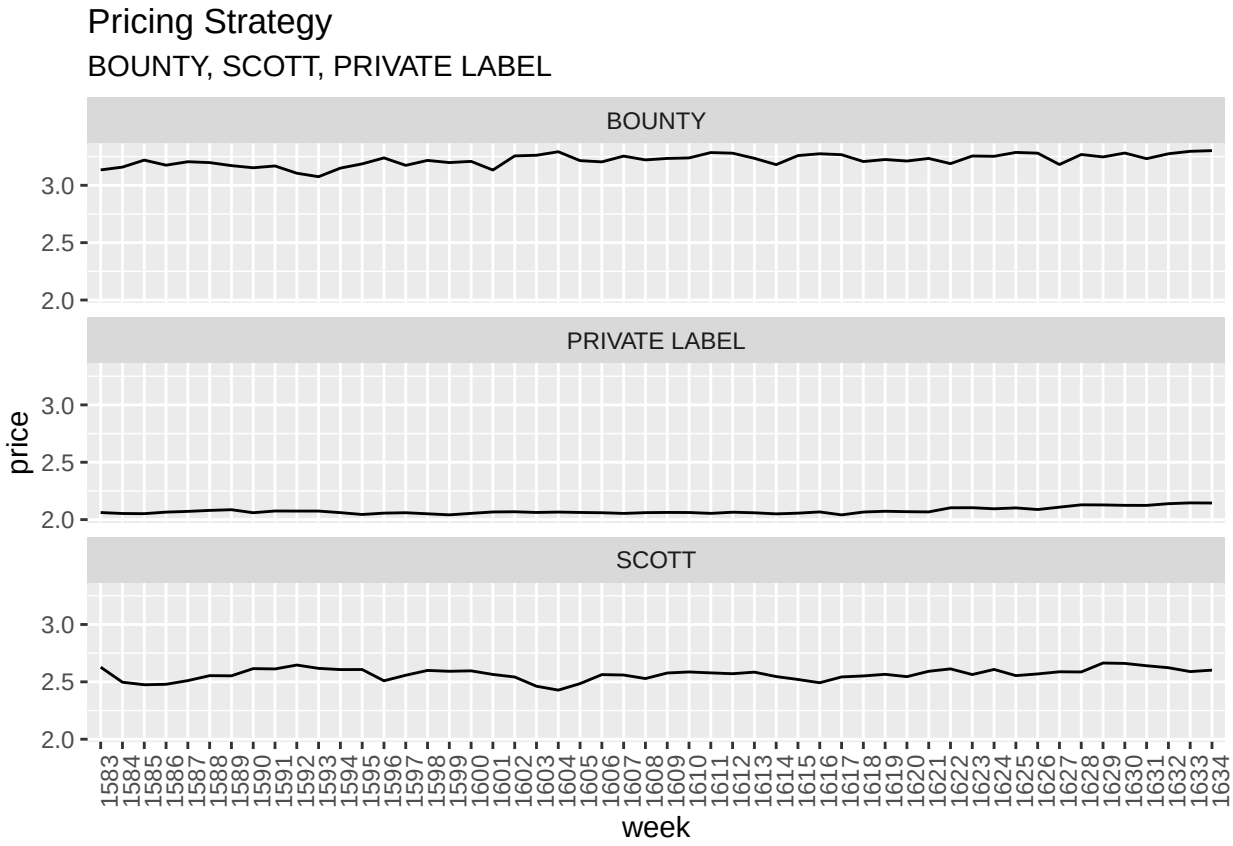
```
## # A tibble: 5 x 2
##   UPC              average
##   <chr>              <dbl>
## 1 88_2_99998_79047    1.83
## 2 0_1_30400_1380      1.90
## 3 0_1_54000_16447     2.12
## 4 0_1_30400_21473     2.15
## 5 0_1_37000_28316     2.21
```

Q9 (10 points)

Most marketers believe that big premium national brands use a pricing strategy called HI-LO while private label brands use a strategy called EDLP. Here HI-LO means a high price that occasionally is discounted, while EDLP means "every day low price" (a constant low price). Using the national brands BOUNTY and SCOTT and the PRIVATE LABEL brand, can you find evidence for this claim in this data? Hint: calculate the average price per brand for each week and plot this as a function of week.

```
pricing_strategy <- sales_products%>%
    filter(L5=='BOUNTY'|L5=='SCOTT'|L5=='PRIVATE LABEL')%>%
    group_by(WEEK,L5)%>%
    summarise(average_price_week=mean(price_per_su))%>%
    as.data.frame()

pricing_strategy %>%
  ggplot(aes(x = factor(WEEK),y = average_price_week, group = L5)) +
  geom_line() +
  facet_wrap(~L5, nrow =3)+
  labs(title = 'Pricing Strategy',
       subtitle = 'BOUNTY, SCOTT, PRIVATE LABEL',
       y = 'price',
       x = 'week')+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Yes. While SCOTT and BOUNTY both charge a relatively high price that fluctuates, PRIVATE LABEL charges a fairly constant low price.

Q10 (5 points)

Some stores close or open during the 52 week period. Create a data frame consisting of the subset of stores that were open throughout the 52 week period (remember the first week of the 52 weeks for which we have sales data is 1583 and the last is 1634) (Hint 1: the store file has information on when each store opened and closed) (Hint 2: You should get a data frame consisting of 1760 stores).

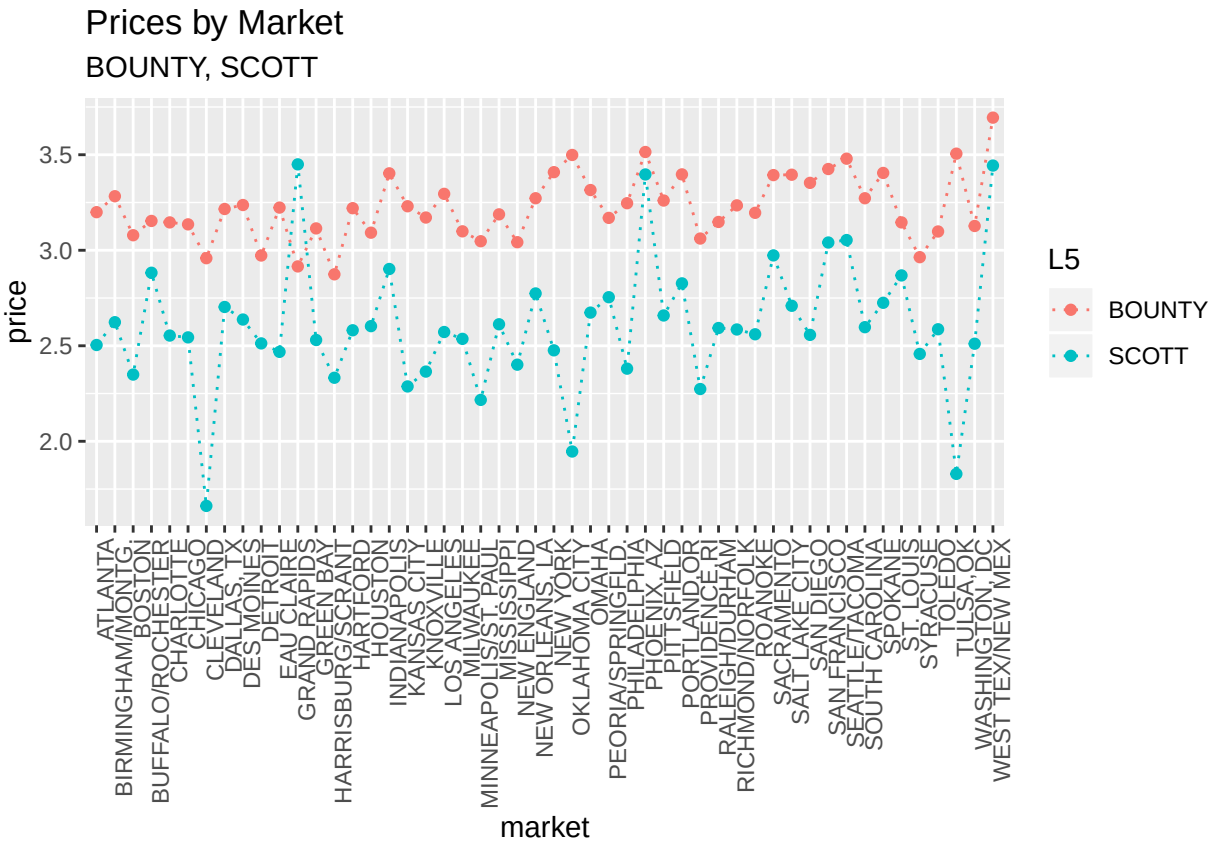
```
openstores<-stores%>%
    filter(Open<=1583&Clsd>=1634)%>%
    as.data.frame()
```

Q11 (7 points)

Let's investigate how prices vary by market. Each store (IRI_KEY) can be matched to a market (Market_Name) in the stores file. We will focus on the two brands BOUNTY and SCOTT and one the stores that were open in all 52 weeks (what you found in Q10). Calculate the average brand price for BOUNTY and SCOTT in each of the 50 markets. Are markets where BOUNTY is expensive also markets where SCOTT is expensive?

```
openstores%>%
    left_join(sales_products)%>%
    filter(L5=='BOUNTY'|L5=='SCOTT')%>%
    group_by(Market_Name,L5)%>%
    summarise(average_brand_price=mean(price_per_su))%>%
```

```
    ggplot(aes(x = factor(Market_Name),y = average_brand_price,color=L5,group=L5)) +
    geom_point() +
    geom_line(linetype='dotted')+
    labs(title = 'Prices by Market',
       subtitle = 'BOUNTY, SCOTT',
       y = 'price',
       x = 'market')+
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



There are markets where BOUNTY and SCOTT are both expensive, but there are also markets where pricing strategy of the two brands deviates from each other.

Q12 (10 points)

UPCs vary in absorbency level, i.e., how good the product is at soaking up liquids. Do markets prefer different absorbency levels in their paper towl product? To answer this question, calculate the share of each absorbency level out of total sales for each market (where sales is measure in volume of standardized units). You can drop UPCs that have missing absorbency level for this calculation (so the share is out of all UPCs with absorbency level information). Create a visualization of absorbency level shares that highlights any potential differences in shares across markets. As above, limit your focus to stores that were open in all 52 weeks.

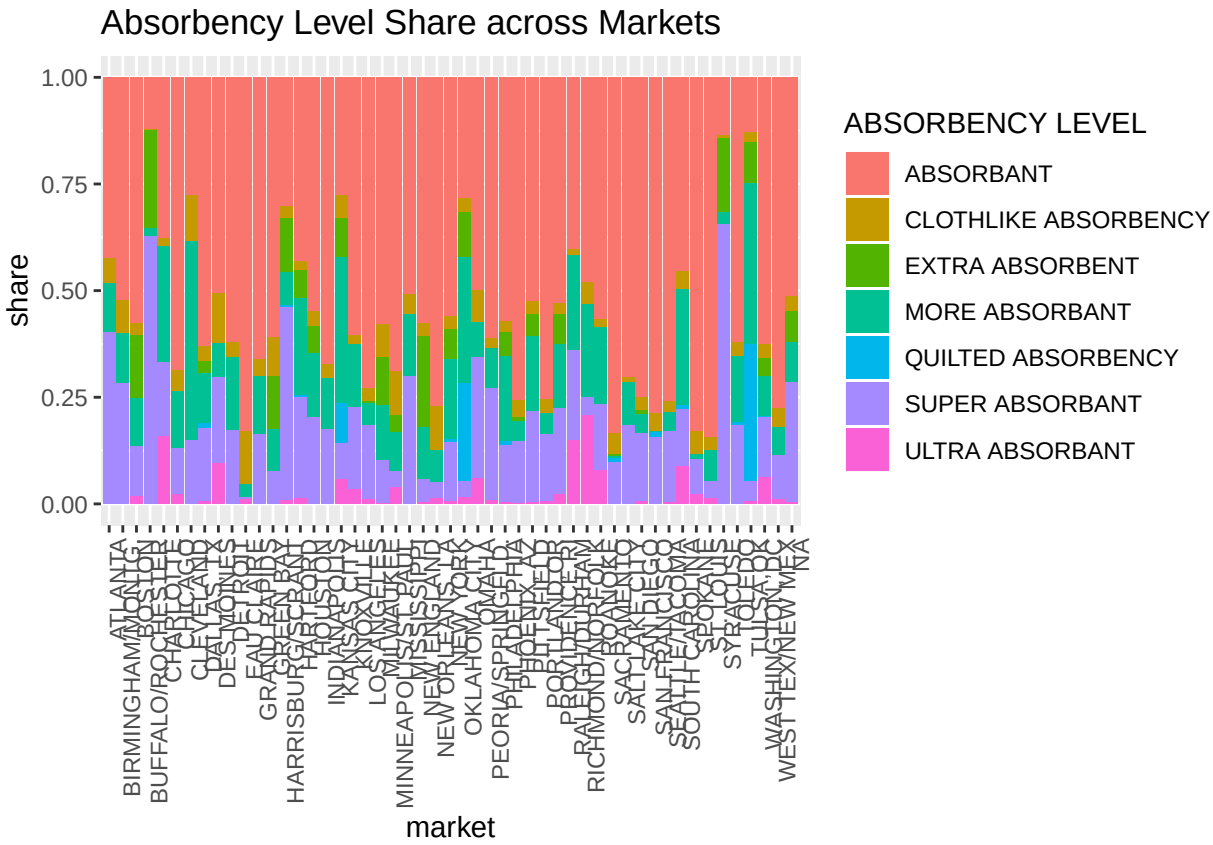
```
sales_products_store<-sales_products%>%
     left_join(openstores)%>%
     filter(`ABSORBENCY LEVEL`!='MISSING')

sum_market<-sales_products_store%>%
```

```
    group_by(Market_Name)%>%
    summarise(sum_market=sum(sales.volumne))
sales_products_store<- sales_products_store%>%
    left_join(sum_market)

ABSORBENCY<- sales_products_store%>%
    group_by(Market_Name,`ABSORBENCY LEVEL`)%>%
   summarise(sum=sum(sales.volumne))%>%
   left_join(sum_market)%>%
    mutate(pct=sum/sum_market)

ABSORBENCY %>%
  ggplot(aes(x=Market_Name,y=pct, fill=`ABSORBENCY LEVEL`)) + geom_bar(stat='identity') +
    labs(title = 'Absorbency Level Share across Markets',
       y = 'share',
       x = 'market')+
   theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Q13 (10 points)

Marketers are often interested in "market structure", i.e., which brands compete with which other brands. In this part you are asked to investigate the market structure in the paper towel category. Focus attention only on the brands BOUNTY, SCOTT and PRIVATE LABEL. Then develop a sales model for each brand at the (Market_Name,WEEK) level. This model should predict sales of a brand in a given week and market. So to develop your model you should first aggregate the data up to this level and then generate any other variables that you think might impact demand for a certain brand.

Once you have built your model answer the question: Which brands compete with which other brands?

Since this is a longer question you will be given some hints in a set of sub-questions below.

Step 1

Start by calculating sales and prices at the (market,week) level for the three brands. So for each (market,week,brand) combination you should calculate total sales volume (in standardized units as above) and the average price (as you did above).

```
sales_products_store<-sales_products%>%
     left_join(stores)
train<-sales_products_store%>%
  filter(L5=='BOUNTY'|L5=='SCOTT'|L5=='PRIVATE LABEL')%>%
  group_by(WEEK,Market_Name,L5)%>%
  summarise(pricemean=mean(price_per_su),
            salestotal=sum(sales.volumne))%>%
  rename(Brand=L5)

train2<-train%>%select(Brand, pricemean)%>%spread(Brand, pricemean)%>%left_join(train)#%>%drop_na()
```

Step 2

Set-up three predictive models where you predict the sales volume (or log(sales volume)) of each of the three brands (in each week and market) using a set of features that you think might predict sales volume. Note that if you include the price (or log(price)) of each of the three brands as features, then you can see how the price of other brands impact sales of a given brand. This will tell you which brands compete with each other (hint: just use a linear model where you use log prices, market and week. Then look at the estimated weights for log price. You do not need to divide the data into training and validation - just train the model on the full sample).

which brands compete with which other brands. generate any other variables that you think might impact demand for a certain brand. Set-up three predictive models where you predict the sales volume (or log(sales volume)) of each of the three brands (in each week and market) using a set of features that you think might predict sales volume.Note that if you include the price (or log(price)) of each of the three brands as features, then you can see how the price of other brands impact sales of a given brand. This will tell you which brands compete with each other (hint: just use a linear model where you use log prices, market and week. Then look at the estimated weights for log price.

```
library(broom)
#just one price
lmR <- lm(log(salestotal)~log(pricemean)+Market_Name + WEEK,data=filter(train,Brand=='BOUNTY'))
#just one price
lmR1 <- lm(log(salestotal)~log(pricemean)+Market_Name + WEEK,data=filter(train,Brand=='BOUNTY'))
lmR2 <- lm(log(salestotal)~log(pricemean)+Market_Name + WEEK,data=filter(train,Brand=='SCOTT'))
lmR3 <- lm(log(salestotal)~log(pricemean)+Market_Name + WEEK,data=filter(train,Brand=='PRIVATE LABEL'))
# three prices
lmR11 <- lm(log(salestotal)~log(BOUNTY)+log(SCOTT)+log(`PRIVATE LABEL`)+Market_Name + WEEK,data=filter(
lmR22 <- lm(log(salestotal)~log(BOUNTY)+log(SCOTT)+log(`PRIVATE LABEL`)+Market_Name + WEEK,data=filter(
lmR33 <- lm(log(salestotal)~log(BOUNTY)+log(SCOTT)+log(`PRIVATE LABEL`)+Market_Name + WEEK,data=filter(

effects <- select(tidy(lmR1),term, estimate) %>%
  rename(BOUNTY=estimate)
effects$SCOTT = tidy(lmR2)$estimate
```

```
effects$PRIVATE = tidy(lmR3)$estimate
effects
```

```
## # A tibble: 52 x 4
##    term                           BOUNTY  SCOTT PRIVATE
##    <chr>                           <dbl>  <dbl>   <dbl>
##  1 (Intercept)                   10.4     18.6    12.2
##  2 log(pricemean)                -5.82    -4.14   -1.76
##  3 Market_NameBIRMINGHAM/MONTG. -0.785   -0.294  -0.679
##  4 Market_NameBOSTON              1.25     1.30    1.49
##  5 Market_NameBUFFALO/ROCHESTER   0.00539  0.243   1.32
##  6 Market_NameCHARLOTTE          -0.116    0.724  -0.292
##  7 Market_NameCHICAGO             0.795    1.42    0.894
##  8 Market_NameCLEVELAND          -1.46    -2.55   -1.68
##  9 Market_NameDALLAS, TX          0.0893   0.494   0.482
## 10 Market_NameDES MOINES         -1.53    -1.28   -1.38
## # ... with 42 more rows
```

```
effects2 <- select(tidy(lmR11),term, estimate) %>%
  rename(BOUNTY=estimate)
effects2$SCOTT = tidy(lmR22)$estimate
effects2$PRIVATE = tidy(lmR33)$estimate
effects2
```

```
## # A tibble: 54 x 4
##    term                           BOUNTY  SCOTT PRIVATE
##    <chr>                           <dbl>  <dbl>   <dbl>
##  1 (Intercept)                   11.4     19.4    12.8
##  2 log(BOUNTY)                   -5.78     2.05    0.402
##  3 log(SCOTT)                     0.155   -4.09   -0.0436
##  4 log(`PRIVATE LABEL`)           0.552    1.31   -1.85
##  5 Market_NameBIRMINGHAM/MONTG. -0.876   -0.547  -0.675
##  6 Market_NameBOSTON              1.17     1.17    1.52
##  7 Market_NameBUFFALO/ROCHESTER -0.0798   0.105   1.34
##  8 Market_NameCHARLOTTE          -0.203    0.552  -0.273
##  9 Market_NameCHICAGO             0.700    1.25    0.919
## 10 Market_NameCLEVELAND          -1.41    -2.40   -1.67
## # ... with 44 more rows
```

Bounty is hugely influenced by Private Label Scott is slightly influenced by Private Label Private Label is not influenced by the other two