Supermarket Shrink

Amelia Farrell

November 11th 2021

Supermarket Shrink - How supermarkets can reduce waste by ordering the right amount at the right time.

Introduction

The problem with shrinkage in supermarkets can be seen in every department. From transport to checkout, dairy to house-hold goods. According to research by the FMI and The Retail Control Group, 64% of store shrink can be traced back to ineffective store operating practices (wheresmyshrink.com, 2012). The highest percentage of this comes from Ordering and Production Planning inefficiencies. Meaning that if we can simply order better and plan to stock the right items, we could reduce 64% of supermarket shrink! However, we know it is not as simple as that. Every department has their own inefficiencies and reasons behind their lost goods. Maybe the deli department is taking too much out of the freezer too soon. Maybe the in store bakery is making too many cakes around Christmas time (when less folks are buying cakes). Due to the shear complexity of this task, we will be addressing one of the departments with the most amount of shrink. The produce department. This will give us a great starting place to solving a wide spread problem. Since this is a world wide problem steaming from all levels of the supply chain, lets review the scope of this analysis below;

Scope:

- Location - U.S.A
- Supply Chain - Retail (grocery store floor level)
- Department - Produce
- Produce - Fresh fruit and vegetables (no frozen, package, or pe-prepared)

Definitions:

- Shrink - loss of inventory in retail
- Retail - the exchange of goods to an end customer
- Customers - individuals making purchase at a grocery store (on-line orders are excluded)

Problem statement - How can we decearse waste (shrink) in the produce department?

As you can still see, even our problem statement seems a bit broad. If you think about all that goes into making a single produce department run smoothly, you can see where this may get complicated. We have the farmers growing/harvesting the goods, truckers transporting, purchasing making these orders, delivery handing the goods, retail workers setting up the displays. We won't pretend that these factors don't make an impact. However, we need to start somewhere and that will be the goods lost on the retail floor.

How we addressed this problem

Due to the complexity of this problem, we wanted to focus on the top items in produce contributing to the waste/shrink. In order to pin point the top items of interest, we gathered data on food loss provided by the The Food and Agriculture Organization (FAO). This data set provided us insight to the top fruits and vegetables with the highest % loss at the grocery store (Retail level of the value chain).

This data base contained much more information than we needed for this analysis. Therefore, we selected data based off the following pentameters;

- *Year Range* - 2000 through 2017 (data set ends in 2017 and we need to go back to year 2000 in order to have enough data to work with for the average loss per item).
- *Aggregation* - Country
- *Country* - United States
- *Basket Items* - Fruit & Vegetables
- *Value Chain* - Retail
- *Commodity* - All
- *Method of Data Collection* - All

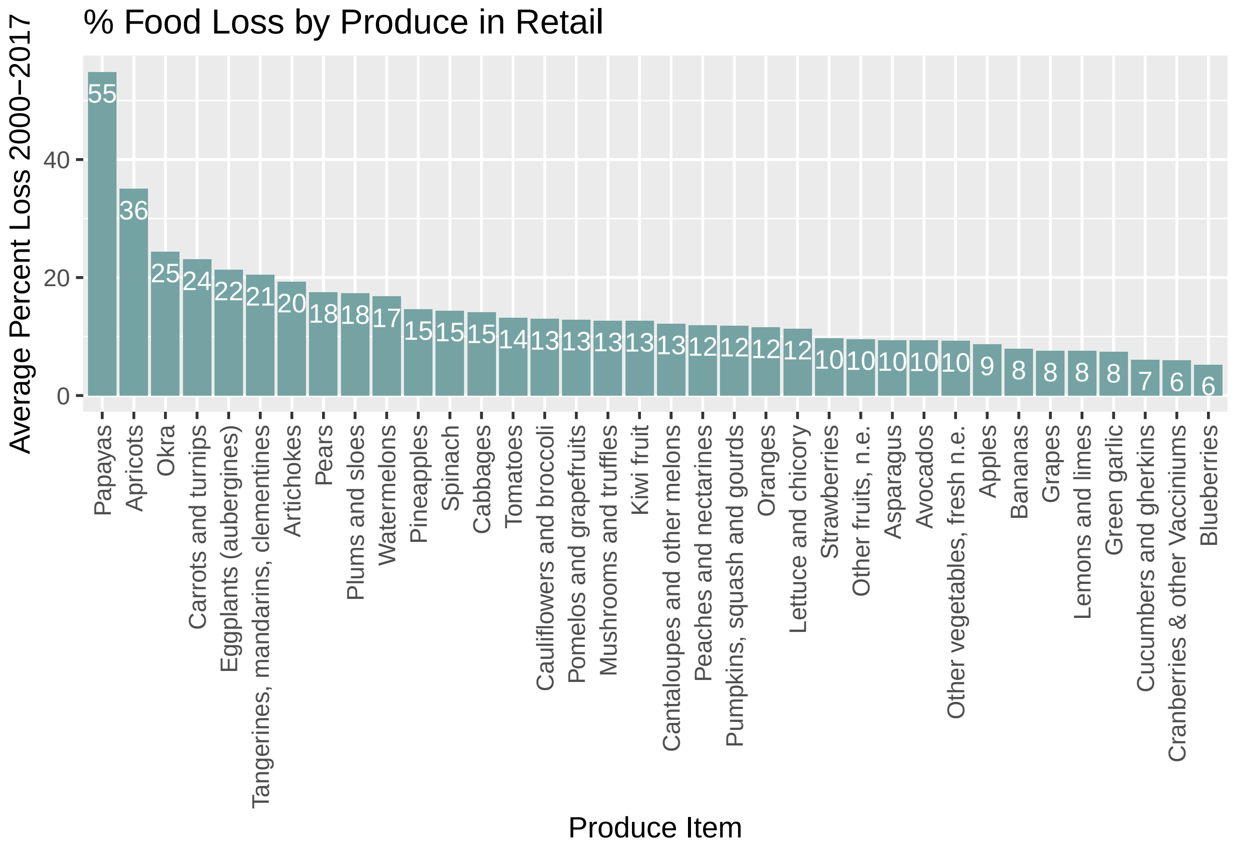
Once we had our top items, we knew the top fruits and vegetables that would make the biggest impact if waste was reduced.

However, his does not solve our problem. Even if we know the top contributes to the shrink/waste, how to reduce it? This is where we want to look into the buying patterns of customers. If we can identify trends in buying habits, we can determine when we would want to buy more or less of our top shrink items. To see how we did this, feel free to continue reading and learn more about this analysis.

Analysis

Step 1 - Top contributors to produce waste

As stated in the problem statement, we will first identify the top contributors to produce waste in the U.S.A by calculating the averages from data published by The Food and Agriculture Organization (FAO) (2000-2017). After calculating the average percent loss from 2000-2017, we can plot this data to visualy identify the top items that contribute to loss.



From the above plot we can clearly know what produce to focus on first. Papayas. Based off the data provided by The Food and Agriculture Organization (FAO), on average, 55% Papayas put on the shelf in retail grocery stores are thrown away before being purchased by the customer. 55%?! That is a significant percentage in comparison to many of the other items in our data set. However, we do need to note that we do not know what 55% is in regards to pounds of food lost.

> Example: Say we by 10 pounds of papayas a week and lose 50% due to a short product shelf life. We also buy 500 pounds of strawberries per week and lose 10% consistently. First glace you may think that we are losing more pounds of papayas each week (50%) when in reality we are only losing 5 pounds of papayas and 50 pounds of strawberries. If we look at this from a dollar loss perspective, the impact could shift again. It is important to note this in our analysis and realize our limitations. We will gain insights with this analysis but we must be aware of other factors at play that we have no visibility to.

Based off the above we will want to focus on buying Papayas and Apricots at the right time to have the most impact on produce shrink. But how do we know how to adjust our stocking/buying patterns? That is where th next step in this analysis comes into play.

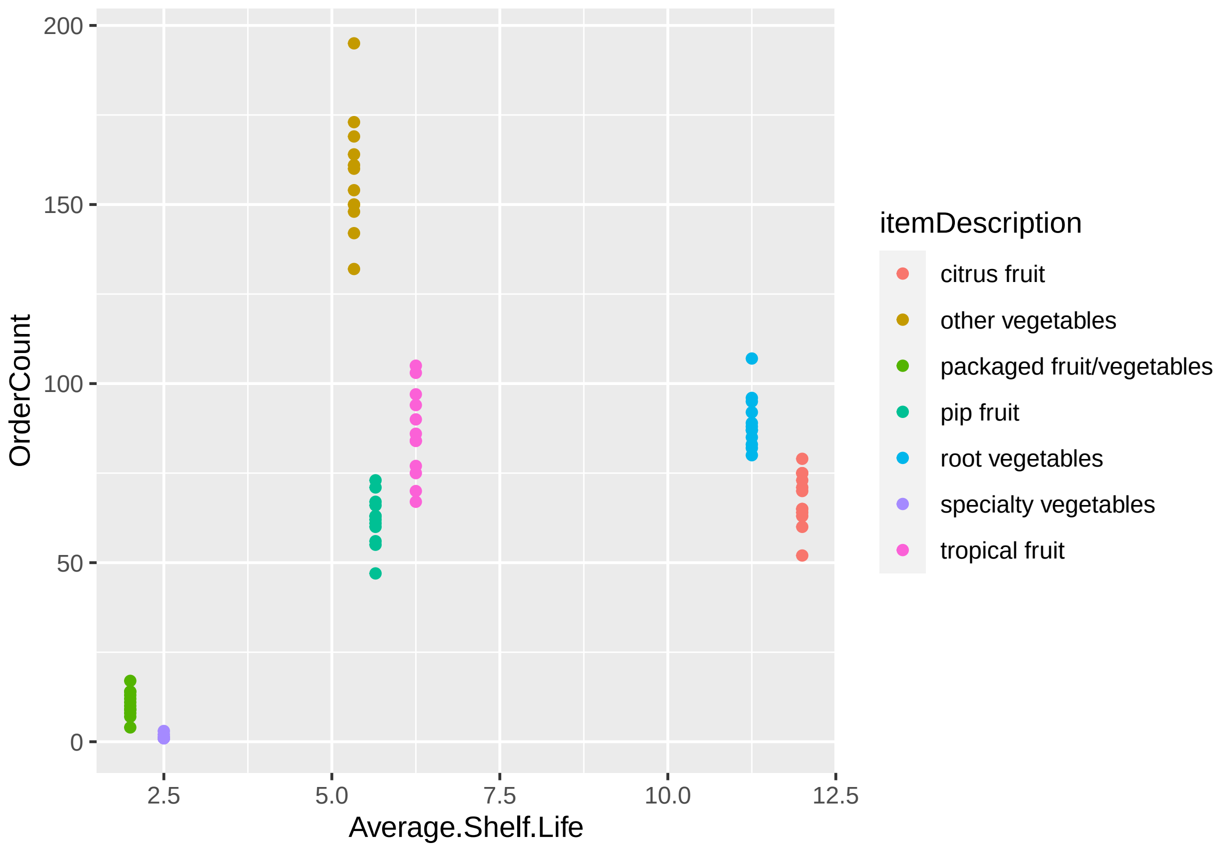
Step 2 - Customer buying patterns - Shelf Life

Now why are Papayas and Apricots not flying off the shelf in time for us to not throw them away? Is that because customers also have issues with consuming these types of foods before they go bad? To answer this question, we combined a Groceries data set from Kaggle and shelf life data from fightbac.org (gathered by The Food Marketing Institute & Cornell University. Added the average shelf life by produce categories to a table with the total number of orders for each (from 2014-2015). We then ran a correlation matrix to see if the average shelf has a relationship to the number of orders by produce category.

```r
#assign a 1 for each observation in our data set
Groceries$observation <- 1

#Adding shelf life and filering on only produce categories
# (since we know that every produce category has an Average.Shelf.Life, we can exclude those without)
GroceriesFinal <- merge(Groceries, ShelfLife,
                        by.x = "itemDescription",
                        by.y = "Produce.Category", all.x=TRUE)

GroceriesProduce <- GroceriesFinal %>% filter(complete.cases(.))

#group/sum number of total orders by month then category
Groceriesgroupbymonth <- GroceriesProduce %>%
  group_by((format(as.POSIXct(GroceriesProduce$Date),"%m")), itemDescription) %>%
  dplyr::summarize(gr_sum = sum(observation))
```

```
## `summarise()` has grouped output by '(format(as.POSIXct(GroceriesProduce$Date), "%m"))'. You can ove
```

```r
names(Groceriesgroupbymonth)[1] <- 'Month'
names(Groceriesgroupbymonth)[3] <- 'OrderCount'

# Adding back Shelf Life to summary
Groceriesgroupbymonth2 <- merge(Groceriesgroupbymonth, ShelfLife,
                                by.x = "itemDescription",
                                by.y = "Produce.Category", all.x=TRUE)

# plot correlation
ggplot(Groceriesgroupbymonth2, aes(x=Average.Shelf.Life,
                                   y=OrderCount,
                                   color=itemDescription)) +
  geom_point()
```



```
# run correlation
cor(Groceriesgroupbymonth2$OrderCount, Groceriesgroupbymonth2$Average.Shelf.Life ,
    method = "pearson")
```

```
## [1] 0.4141675
```

Interestingly enough there does look to be a pattern. You can see that the items with the shortest shelf life appear at the bottom left corner of our scatter plot. Showing that the items with the shortest shelf life are bought the least. Even though we can clearly see this in our plot, our correlation coefficient (0.369) indicates a weak positive relationship. Can you see what may me skewing our correlation coefficient? "Other vegetables" appear to be the outlires here. Since we do not know what these "Other vegetables" are, lets try running our correlation coefficient without the "Other vegetables" in our data set.

```
Groceriesgroupbymonth3 <- Groceriesgroupbymonth2 %>% filter(itemDescription != "other vegetables")
cor(Groceriesgroupbymonth3$OrderCount, Groceriesgroupbymonth3$Average.Shelf.Life , method = "pearson")
```

```
## [1] 0.7748305
```

Now we get correlation coefficient 0.757. Indicating a strong positive relationship between shelf life and number of orders. So is shelf life important in predicting the number of orders? Idealy we want to know the exzact number of orders bought and shelf life per fruit/vegetable. However, since we lack this data, we will make the best use of what we have and preform a One-Way ANOVA test to test the significance of our first correlation coefficient. The output of this One-Way ANOVA can be seen below.

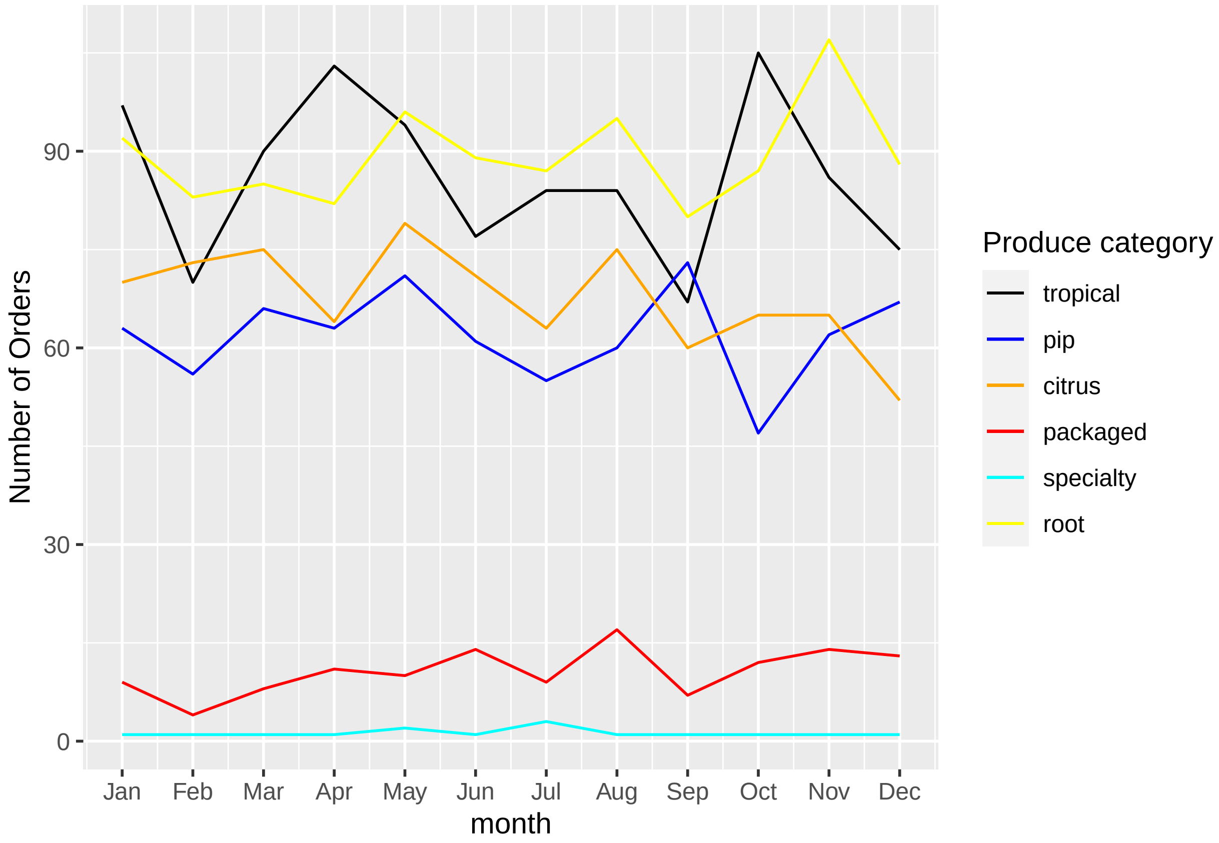
```
#One-Way ANOVA
Shelf.Life.one.way <- aov(OrderCount ~ Average.Shelf.Life, data = Groceriesgroupbymonth2)
summary(Shelf.Life.one.way)
```

```
##                    Df Sum Sq Mean Sq F value   Pr(>F)
## Average.Shelf.Life  1  35469   35469   16.98 8.99e-05 ***
## Residuals          82 171307    2089
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the looks of this One-Way ANOVA, we can conclude that our Average.Shelf.Life variable does make an impact on the number of sales! ($p<0.5$). Even though our correlation coefficient was weak with the unidentified vegetables, our hypothesis was correct and shelf life does impact the total number of sales. Now does shelf life correlate with food loss? Sadly these two data sets a grouped differently and we would really need data at am item level to answer this question with certainty. For now, we will save this question for when we have gather better data. However, we can say that the two high loss items we are focusing on (Papayas and Apricots) fall in very close groupings when it comes to average shelf life (tropical and pip fruit respectively).

Step 3 - Customer buying patterns - Seasonality

Now we want to know if there are other factors such as the time of years (aka Seasonality) is influencing our customers buying patterns as well. To look at this, we grouped the data by month and plotted the total number of orders by produce category.



From looking at the above line chart, we do not see much seasonality taking place. If there was significant seasonality, than we would see patterns in the line chart, such as increases during the summer/spring months.

We do not see any clear trends above and this may be due to the limited amount of data we are working with (this is only 2 years worth of orders).

Step 4 - Customer buying patterns - Market Basket Analysis

As you can see in the last step, we were not be able to predict how many Papayas and Apricots to stock based off the historical sales of tropical and pip fruit alone. Could we make any inferences based off other items customers bought along side tropical and pip fruit? To answer this question we used the 'arules' library and apriori() function/algorithm to do a bit of association rule mining (aka Market Basket Analysis) help us identify trends in customers shopping carts.

```
# Grouping catagoires by customer order and date
basket <- ddply(Groceries, c("Member_number","Date"),
                function(df1)paste(df1$itemDescription,collapse = ","))
head(basket,5)

# Removing member and date, leaving only the grouped lists (baskets)
basket$Member_number <- NULL
basket$Date <- NULL
colnames(basket) <- c("basket")

write.csv(basket,"basket.csv", quote = FALSE, row.names = TRUE)
head(basket)

# Converting to a S4 object for apriori() to run correctly
baskets = read.transactions(file="basket.csv",
                            rm.duplicates= TRUE,
                            format="basket",sep=",",cols=1);
print(baskets)

# Creating our set of rules. We did not set a number of rules. We are curious as to how many rules stem
basketrules <- apriori(baskets, parameter = list(minlen=2, sup = 0.01, conf = 0.05, target="rules"))

# How many rules were created?
print(length(basketrules))

# What rules were created?
inspect(basketrules[1:10])
```

We first ran a Market Basket Analysis on our entire data set. This left us with quite a number of rules that were unrelated to produce. See the first 5 rules below;

```
basketrules_by_lift <- sort(basketrules, by = "lift")
inspect(basketrules[1:5])
```

```
##     lhs             rhs                support    confidence coverage
## [1] {yogurt}     => {whole milk}       0.01115714 0.12996109 0.08584981
## [2] {whole milk} => {yogurt}           0.01115714 0.07067287 0.15787012
## [3] {soda}       => {whole milk}       0.01162480 0.11975224 0.09707376
## [4] {whole milk} => {soda}             0.01162480 0.07363521 0.15787012
## [5] {rolls/buns} => {other vegetables} 0.01055585 0.09599028 0.10996793
##     lift      count
## [1] 0.8232152 167
## [2] 0.8232152 167
## [3] 0.7585491 174
## [4] 0.7585491 174
## [5] 0.7864163 158
```

As you can see there are many here that are unrelated to produce. In order to exlude these we ran a for loop to identify the baskets with a least one produce item and re-ran our association rules.

```r
# Creating a new data frame so we can exclude anything basket that does not contain produce.
list2 = c()

for (i in 1:nrow(basket)) {
  search <- c("pip fruit", "packaged fruit/vegetables","citrus fruit",
              "specialty vegetables","root vegetables","other vegetables")
  out <- str_contains(basket[i, ], search, ignore.case = TRUE, logic = "or")
  list2 <- c(list2, out)
}
basket5 <- cbind(basket, list2)
names(basket5)[1] <- 'newbasket'
names(basket5)[2] <- 'logic'

# Checking names of new list
names(basket5)

# Filtering for baskets with produce
basket6 <- filter(basket5, logic == "TRUE")

# Removing logical column "list2"
basket6$logic <- NULL

# writing new file for the apriori to read
write.csv(basket6,"basket6.csv", quote = FALSE, row.names = TRUE)
head(basket6)

# Converting to a S4 object for apriori() to run correctly
basket6 = read.transactions(file="basket6.csv", rm.duplicates= FALSE, format="basket",sep=",",cols=1);
```

```
## Warning in asMethod(object): removing duplicated items in transactions
```

```r
print(baskets)

# Creating new set of rules
producerules <- apriori(basket6, parameter = list(minlen=2, sup = 0.01, conf = 0.3, target="rules"))

# How many rules were created?
print(length(producerules))
```

Lets take a look at our top 5 rules related to produce but first sorting by the amount of lift.

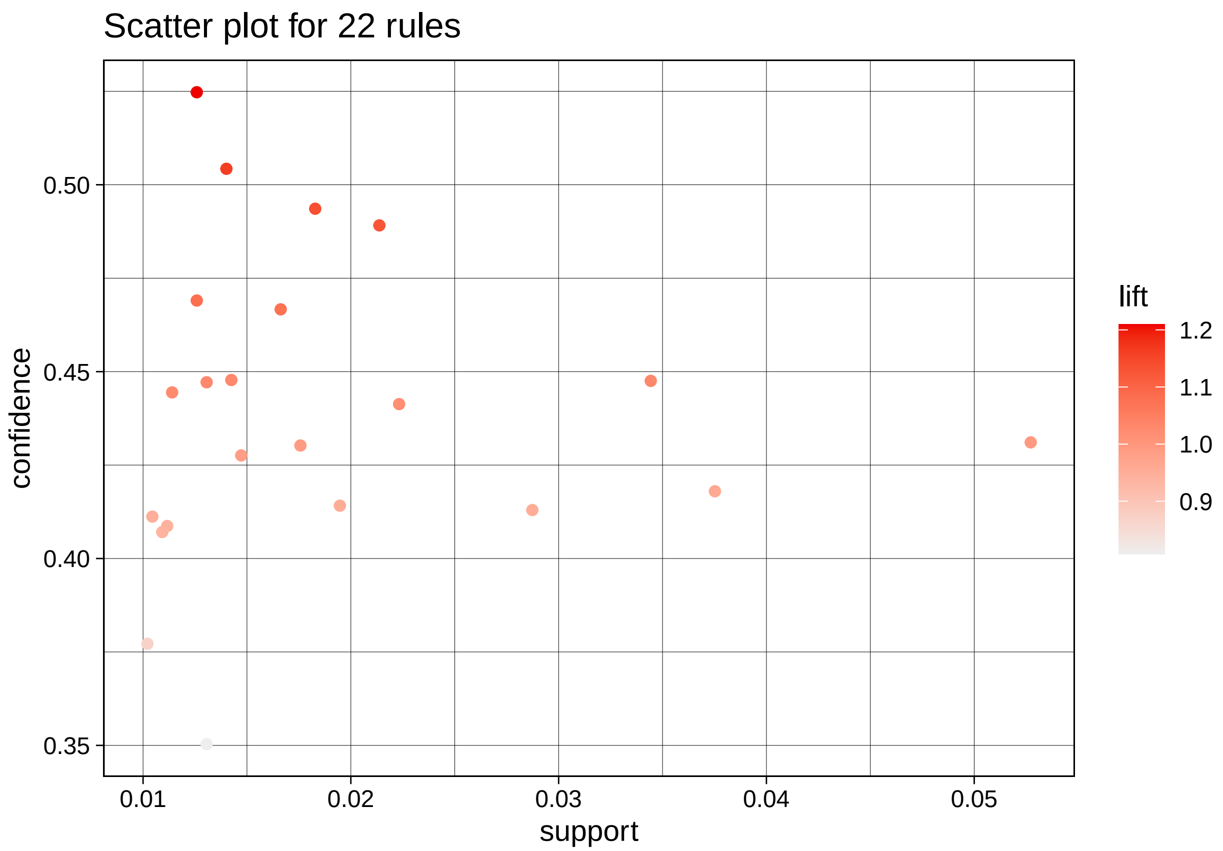
```r
# Sorting by lift
rules_by_lift <- sort(producerules, by = "lift")
inspect(rules_by_lift[1:5])
```

```
##     lhs                  rhs                  support    confidence coverage  
## [1] {curd}            => {other vegetables} 0.01258608 0.5247525  0.02398480
## [2] {pork}            => {other vegetables} 0.01401092 0.5042735  0.02778437
## [3] {frankfurter}     => {other vegetables} 0.01828544 0.4935897  0.03704583
## [4] {sausage}         => {other vegetables} 0.02137260 0.4891304  0.04369508
## [5] {domestic eggs} => {other vegetables} 0.01258608 0.4690265  0.02683448
##     lift     count
## [1] 1.209487 53   
## [2] 1.162286 59   
## [3] 1.137661 77   
## [4] 1.127383 90   
## [5] 1.081046 53   
```

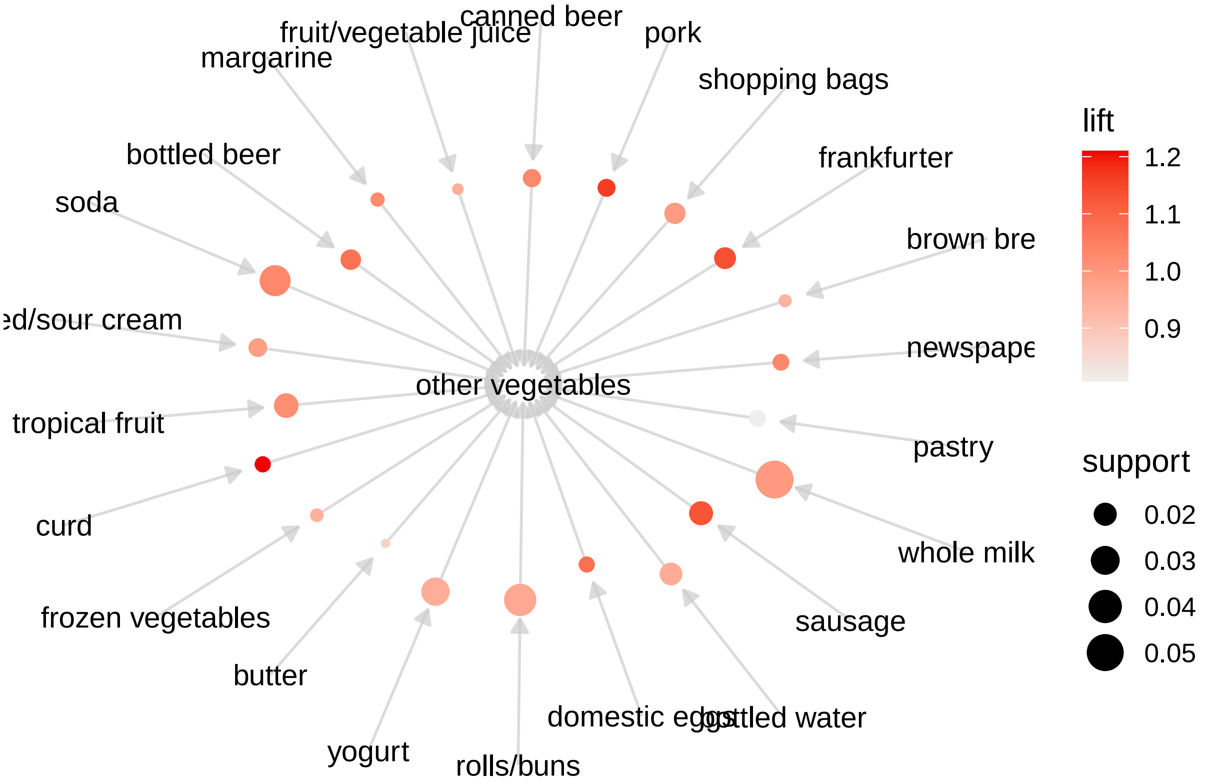
Looking at the top rule, we can say there that from the customers buying produce there is a 52% chance that a customer buying curd will buy "other vegetables". The remaining follow th same logic. However, this is not helpful for identifying where our top shrink vegetables lie (Papayas/tropical and Apricots/pip).

Lets also plot these rules shown below

```r
# plot of rules
plot(producerules, jitter = 0)
```



```r
plot(rules_by_lift[1:22], method = "graph")
```



We can easily see tropical and pip fruit in the second chart. However they do not have any rules applied to them other than "other vegetables" associated with their purchase. Knowing that customers who purchase "other vegetables" will likely also buy tropical or pip fruit does not help answer our question. We were hoping to find a strong association between a non perishable good and tropical/pip fruit purchase. This would have given us another variable to help predict the sales of tropical and pip fruit. Since we did not find a strong association, so far we only know that average shelf life of produce may play a role in customers buying habits of produce. In short, the above chart shows that our second apriori algorithm could not associate tropical or pip fruit to any other non produce purchase with over 30% confidence.We could drop our confidence lower and see what we get but this would mean that we are lowing our expectations on the model we want to build.

Implications

There was a lot covered in this analysis and many directions this analysis could go from here. As we saw in the beginning of the analysis, super market shrink and produce shrink in particular, take a toll on super markets bottom lines. We have not even begun to discuss the impact this has on the environment. In order to have a place to start, we focused on finding the produce items with the highest amount of waste at the retail level of the super market. We quickly discovered that Papayas and Apricots were indeed the top culprits. With just this information alone, we could began taking action by looking at how much revenue these two items bring to the store and how much they are losing (margin). Are these two items worth keeping on the shelves? Do customers come to the store for these high shrink items in particular or are customers making these purchases on a whim? We wouldn't want to remove Papayas and Apricots if they significantly drove up the sales of another high margin good. These follow up questions would need more data and further analysis. However, we now have a place to start. We also uncovered a relationship between short shelf life items and fewer customer purchases. Based off these findings we know to take a closer look at all produce items with short shelf life. Why are customers buying these less often? Are most of the produce in this short shelf life "basket" also high shrink? In order to fully answer these additional questions, we really would prefer to have a more complete data set with item names and not groups of items like "pip fruit" etc.

Limitations

As we stepped through this analysis, we uncovered many limitations. We did not start this analysis with the best data sets. The main problem stemmed from our groceries data set. This data set was grouped by category and not by the specific item that was sold. This made it very difficult to tie back our findings on the high shrink items (Papayas and Apricots) to customer purchases. We could only look at their groupings of tropical and pip fruit respectively. This severely limited our capabilities to fully understand customer buying patters.

Concluding Remarks

After reviewing the implications and limitations of the analysis, you can easily see how many directions this analysis could go. However, we would like to continue on the path of reducing waste at the retail level. This way we can take our findings to the retailer and help them implement changes that would help their bottom line and environment at the same time. The next steps would be to collect the line item data we truly needed from as many retailers as possible. We would also design and implement customer surveys asking exactly why they made the purchases they did. With both of these data sets in hand, we would easily take this analysis to the next level and run some linear regression models to see what variables are associated with customers purchasing Papayas and Apricots.

References

wheresmyshrink.com, 2012. Executive Summary. http://wheresmyshrink.com/executivesummary.html?fbclid=IwAR0w7KKjS-4Lr1wJ3JuJ2ZYbsZGZbc57Go4NuBinNwytYNG5911QUBtXXYE

FAO, 2021. Food Loss and Waste Database. The Food and Agriculture Organization (FAO). https://www.fao.org/platform-food-loss-waste/flw-data/es/.

Dedhia H., 2020. Groceries dataset. Kaggle.com. https://www.kaggle.com/heeraldedhia/groceries-dataset

Food Marketing Institute & Cornell University, 2020. The Food Keeper. fightbac.org. https://lee.ces.ncsu.edu/wp-content/uploads/2012/12/TheFoodKeeper.pdf?fwd=no&fbclid=IwAR2QE_yWd_E6kzD7Sp18AnLN36h7uLPpmM7CrsUZC91OQz_pHi_hT3jZvBU