10.3 Final Project Step 2

Amelia Farrell

November 1st 2021

Supermarket Shrink

Importing and cleaning the data

As described in Project Step 2, we will be using three different data sets to answer our problem statement (aka Business Case).I will review how we pulled and cleaned each data set in preparation for our final set.

1. *Food Loss and Waste Database*

Purpose of data set - this data (provided by the The Food and Agriculture Organization (FAO)) set will give us insight to the top fruits and vegetables with the highest % loss at the grocery store (Retail level of the value chain). Our goal is to obtain the average % loss per Fruit/Vegetable over the selected time frame.

This data base contains much more information than we need for this analysis. Therefore, we will be downloading 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

After downloading the data set with the above perameters from the The Food and Agriculture Organization site, we can see that it has 22 variables.

```
##  [1] "geographicaream49"          "country"
##  [3] "region"                     "measureditemcpc"
##  [5] "crop"                       "timepointyears"
##  [7] "loss_per_clean"             "percentage_loss_of_quantity"
##  [9] "loss_quantity"              "loss_qualitiative"
## [11] "loss_monetary"              "activity"
## [13] "fsc_location1"              "periodofstorage"
## [15] "treatment"                  "causeofloss"
## [17] "samplesize"                 "units"
## [19] "method_datacollection"      "tag_datacollection"
## [21] "reference"                  "url"
```

As you can see there a lot of variables that we do not need. We already choose the country, region, etc. So we will drop any of the variables that are known and unneeded.

```
## [1] "crop"           "loss_per_clean"
```

Final Variables

- crop - The specific fruit or vegetable name
- loss_per_clean - % Loss at retail (by observation)

These will provide us with the information on the top contributor to produce waste.

2. *Groceries dataset*

Purpose of data set - The Groceries dataset from Kaggle will be the main data set used in this analysis. This data set will give us insight to customer buying patterns and let us see whether or not there is relationships between the types of groceries purchased and when. Based off this information, we may be able to detrime when to stock less produce leading to less waste (less going bad/rotten before it is purchased by the end customer)

This data base only contains 3 varibles which are all important to this analysis. Therefore, we will be using all the variables below;

```
## [1] "Member_number"   "Date"            "itemDescription"
```

- Member_number - Unique customer ID
- Date - Date of transaction
- itemDescription - High level grouping of product (this data is not split out by individual product names)

However, it is important to note that "itemDescription" is a categorical variable. So if we want to use this to build a prediction model, we will have to code it to numeric values.

2. *The Food Keeper Data Set*

Purpose of data set - The Food Keeper Data, put to gather by The Food Marketing Institute & Cornell University. Will provide us with an additional variable to add to our Groceries dataset. It will let us look at relationships related to the average shelf-life of a category (e.g. citrus fruit, tropical fruit, etc.).

This data has been scrapped from the PDF provided by fightbac.org. After scrapping the shelf-life of fruits and vegetables (kept refrigerated), we categorized them in the same categories from the Groceries data set (Tropical fruit, pip fruit, etc.) then calculated the average shelf-life in days for each category. Our final set contains the variables below,

```
## [1] "Produce.Category"   "Average.Shelf.Life"
```

What does the final data set look like?

In the end we will be using two data sets. One for looking to identify customer buying patterns to better plan for the stocking of perishable produce and the other looking at the fruit and vegetables that produce the most waste in retail, giving us specific items to focus on.

Lets combine the Groceries and ShelfLife data sets and review the final two below,

Combining the Groceries and ShelfLife (Note: This will leave NAs for any Grocery category that we did not calculate shelf-life for (all non-perishables, meat, etc.). We dont need to worry since we only care about the shelf life of the produce).

```
## [1] "itemDescription"    "Member_number"      "Date"
## [4] "Average.Shelf.Life"
```

GroceriesFinal

- Member_number
- Date
- itemDescription
- Average.shelf.life

Food_Loss2

- crop
- loss_per_clean

Questions for future steps.

After we understand the data sets we are working with and concatenate *Groceries* and *ShelfLife* (as seen above). We can start manipulating/summarizing/visualizing out data to answer some of the key questions we laid out in Project Step One.

1. What produce produces the most amount of waste? - In order to answer this question we will be using the *Food_Loss2* data set. However, this data set has multiple observations per "crop" (fruit/vegetable). So we will need to group each crop to get the total food loss for all of the observations combined. We will then need to count the number of observations per crop. Once we have the total loss and count we can divide the loss by count to get the average loss per crop. This will give us the data we need to summarize and visualize the crops with the highest average loss from 2000-2017 (based on the parameters we set in section 1 of this discussion)
2. Is there any relationship between the categories of items bought and the time of year? Seasonality. - This question will also require some data engineering of the concatenated *Groceries* and *ShelfLife* data set (GroceriesFinal). This data set lists out the category of items bought by customer by day. This is not ideal, in a perfect world we would like to see the Qty of each good purchased by day and customer. However, in reality we will never have the "perfect" data to work with. So we will need to make it work. In order to do so, we will need to count the number of transactions under each category per day. There are many ways to transform our data to our liking in R, for ease of use, we will be using the tabyl() function from the janitor library to make this transformation and assign it to a new data frame (details shown below)

```
GroceriesCountbyDate <- tabyl(Groceries, Date, itemDescription)
str(GroceriesCountbyDate[1:4])
```

```
## Classes 'tabyl' and 'data.frame':    728 obs. of  4 variables:
##  $ Date            : chr  "01-01-2014" "01-01-2015" "01-02-2014" "01-02-2015" ...
##  $ abrasive cleaner: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ artif. sweetener: num  0 0 0 1 0 0 0 0 0 0 ...
##  $ baby cosmetics  : num  0 0 0 0 0 0 0 0 0 0 ...
```

We can also use this function to look at the data by percentage.

```
GroceriesCountbyDatePercent <- tabyl(Groceries, Date, itemDescription) %>%
  adorn_percentages("col") %>%
  adorn_pct_formatting(digits = 1)
str(GroceriesCountbyDatePercent[1:4])
```

```
## Classes 'tabyl' and 'data.frame':    728 obs. of  4 variables:
##  $ Date            : chr  "01-01-2014" "01-01-2015" "01-02-2014" "01-02-2015" ...
##  $ abrasive cleaner: chr  "0.0%" "0.0%" "0.0%" "0.0%" ...
##  $ artif. sweetener: chr  "0.0%" "0.0%" "0.0%" "3.4%" ...
##  $ baby cosmetics  : chr  "0.0%" "0.0%" "0.0%" "0.0%" ...
```

As you can see, our "Date"s are not in order. We can use the sort() function to set these in the right order for plotting.

3. Are there any correlations between the amount of produce sold and other non-perishable goods? - To answer this question we can also use the *Groceries* data set. However, for this, we may want to control for the date variable (controlling for any seasonality).

What information is not self-evident?

There are many questions left unanswered about our data sets.

- Is there any missing data?
- How is the data distributed?
- Are there typos in the data?
- Are there any patterns or this the data completely random?

Thankfully R provides us with countless tools and packages to help us answer these questions. We can use some of these tools with our data sets below;

- Missing data? We can use this to varify the known missing values under Average.Shelf.Life

```
GroceriesFinalMissing <- GroceriesFinal[!complete.cases(GroceriesFinal),]
head(GroceriesFinalMissing, 3)
```

```
##    itemDescription Member_number       Date Average.Shelf.Life
## 1 abrasive cleaner          2421 21-11-2015                 NA
## 2 abrasive cleaner          3390 21-02-2015                 NA
## 3 abrasive cleaner          4569 24-12-2014                 NA
```

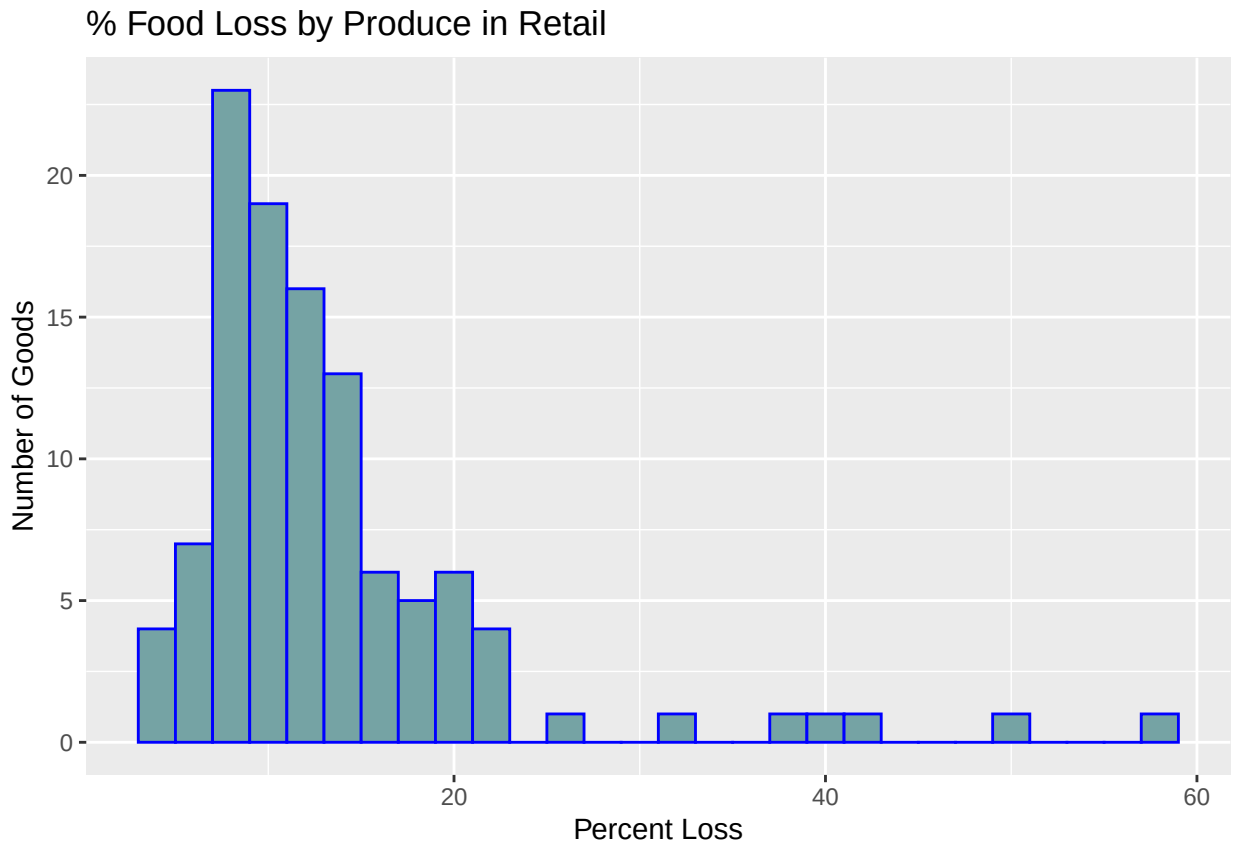
We can also use to check for missing vaules in the Food_Loss data set.

```
Food_Loss2Missing <- Food_Loss2[!complete.cases(Food_Loss2),]
head(Food_Loss2Missing, 3)
```

```
## [1] crop           loss_per_clean
## <0 rows> (or 0-length row.names)
```

- Data distribution - histograms, *stat.desc()* (for skewness, Kurtosis, etc.) Lets start by creating some simple histograms to look at our two data sets starting with Food_Loss.

```
ggplot(Food_Loss2, aes(loss_per_clean)) + geom_histogram(bins = 10, binwidth=2, fill="#75a3a4", color="
  ggtitle("% Food Loss by Produce in Retail") + xlab("Percent Loss") + ylab("Number of Goods")
```



We can see from above that many of our produce falls below 20%. Therefore, our distribution is not normal. However, this lets us know that there are indeed outlires (items falling above 20%) that we can focus on. These items will have the greatest impact when there loss is reduced.

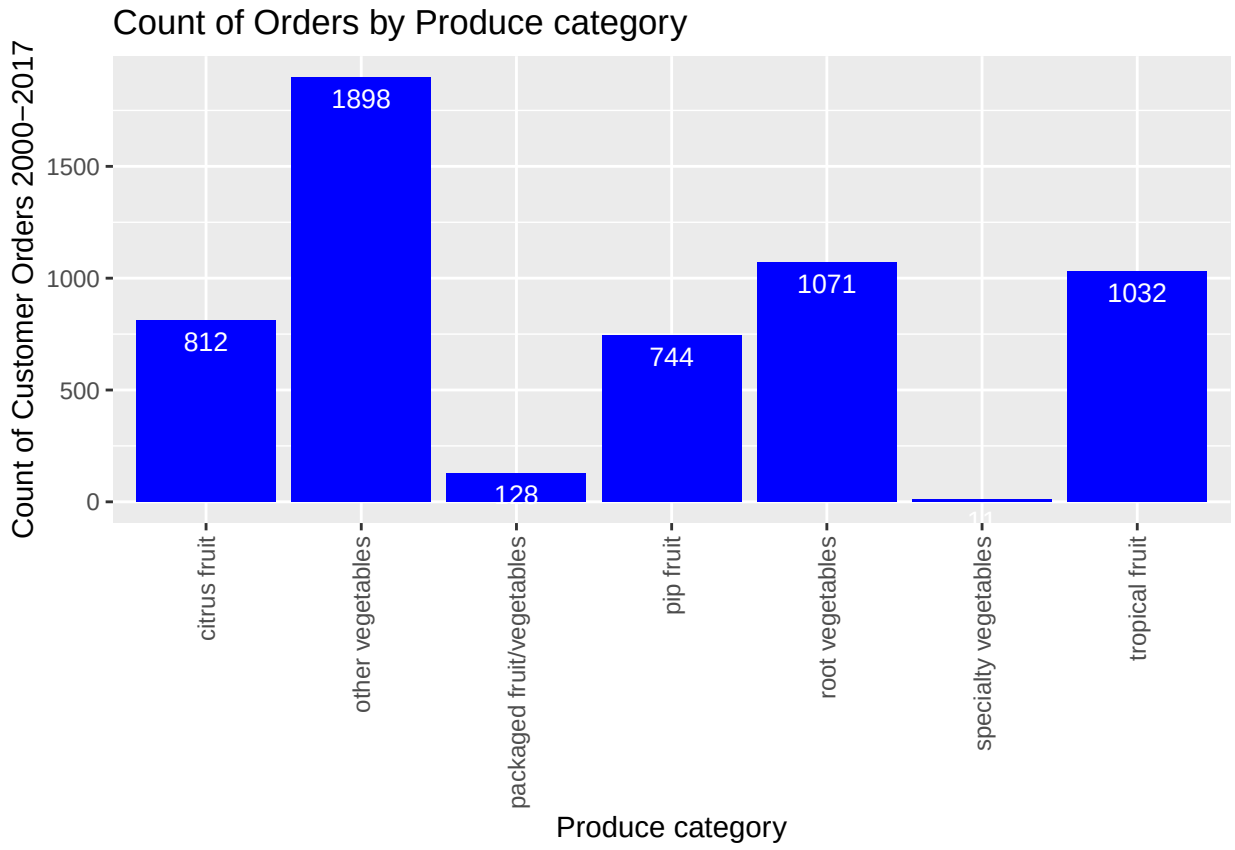
Next lets look at bar chart of the produce categories in our GroceriesFinal data set. This will let us see what group of produce customers are buying the most of. First wee need to assign a 1 for each observation in our data set (since the categories are categorical variables we need something for R to measure them by). Then filter the other non-produce categories and lastly filer group by category to plot.

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

```
#filer on only produce categories (since we know thart every produce category has an Average.Shelf.Life
GroceriesProduce <- GroceriesFinal %>% filter(complete.cases(.))

#group by category
GroceriesCount <- GroceriesProduce %>%
  group_by(itemDescription) %>%
  summarise(sum = sum(observation))

#bar chart
GroceriesCount = GroceriesCount[with(GroceriesCount, order(-sum)), ]
ggplot(data=GroceriesCount, aes(x=itemDescription, y=sum)) +
  geom_bar(stat="identity", fill="blue")+
  geom_text(aes(label=sum), vjust=1.6, color="white", size=3.5)+
  theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))+
  ggtitle("Count of Orders by Produce category")+
  ylab("Count of Customer Orders 2000-2017")+
  xlab("Produce category")
```



As we can see above, "other vegetables" is the most poplar category of produce sold. As we ca see "specialty vegetables" are the least likely to be in a customers order. Is that because they have a shorter shelf-life? Lets do a quick correlation matrix to test our theory.

```
GroceriesCount1 <- merge(GroceriesCount, ShelfLife, by.x = "itemDescription", by.y = "Produce.Category"
cor(GroceriesCount1$sum, GroceriesCount1$Average.Shelf.Life)
```

```
## [1] 0.4214029
```

As you can see, there is a slight positive correlation, between the number of orders and longer shelf-life. However, more tests will need to be done in order so wee if there is a true relationship.

- Patterns - we can run numbers of statistical tests (AVONA, T-Test, Z-Test, etc.) in order to tests hypothesis on patterns that we think may exists in our data.

What are different ways you could look at this data (plots, tables, etc.)?

As you can see from the above, there are numerous ways that we can look at this data. We have only just begun scratching the surface. We can crate box plots and histograms of the number of items sold by year, by customer. We can create line charts to look at seasonality. The options are endless and we just began. We could look at our data in a number of different whys using R. From simple tables to complex dynamic/interactive visualizations.The options are endless and we just began

How do you plan to slice and dice the data?

We plan to slice and dice the data based on similar methodology as above. We already summed up the total order by produce type and merged two data sets together. Next we may want to summarize by month or year to look more at the buying patterns of customers (zooming in on the timing of produce purchases in relation to the month od the year)

How could you summarize your data to answer key questions?

Just to reiterate, after summarizing our data by category, we will also be summarizing it by month of the year in the next step.

Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

This is a tough question to answer. Since we have only just began exploring our data, it is difficult to say whether or not our data will fit into a machine learning model. If we find that our data does indeed have seasonality, we could test out a Holt-Winters forecasting model to predict the number of produce puchases over time.

Questions for future steps.

Our next steps will be to fully answer the questions we were not able to above, - Is there seasonality in our GroceriesFinal data set? - What items are sitting above 50% loss?

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