EDA - Zero to Coke Hero

Michael Tom

April 19, 2024

Introduction

Business Problem and Analytic Approach

Business Problem Statement for Swire Coca-Cola's Innovation Project

Business Problem:

Swire Coca-Cola, a prominent leader in the beverage distribution industry across the Western US, faces the challenge of accurately timing the launch of limited edition "innovation" beverage products. Despite possessing extensive historical sales data and regularly introducing new products, Swire struggles with accurately predicting demand for these innovative, never-before-sold items. These products are pivotal for sustaining customer interest, stimulating demand, and maintaining a competitive edge. However, the risk of overproduction or underestimating demand looms, potentially leading to revenue loss or customer dissatisfaction.

The Benefit of a Solution:

A precise demand forecast for Swire Coca-Cola's innovative products is imperative for the company to meet all potential demand without overproducing, thereby maximizing revenue and elevating customer satisfaction. By strategically launching the right products in suitable regions at optimal times, Swire can attain a market advantage, reinforce its position as an industry trendsetter, and foster brand loyalty among increasingly diverse consumers.

Success Metrics:

The project's success will be gauged based on the accuracy of demand forecasts for the specified innovative products compared to actual sales outcomes. Success metrics encompass the model's capacity to predict location, timing, duration, and units sold for each product launch, furnishing actionable business insights that drive financial gains and amplify market share for essential flavors.

Analytics Approach:

The analytics team will employ Exploratory Data Analysis (EDA) to understand the underlying patterns, Time Series Forecasting techniques (ARIMA and SARIMA models), and Machine Learning algorithms derived from over 24 million observations spanning three years of sales data. Practically, the team must also grapple with the intricacies of this significant market, which encompasses more than 3100 unique products sold across 13 western states with diverse demographics, while considering the costs associated with production overage and underage. This comprehensive approach aims to predict optimal launch periods for new products and discern the attributes and quantities contributing to successful launches. Additionally, the team will adhere to standard data science practices such as data separation into training and test sets to validate model predictions and accurately forecast demand.

Scope:

The project will concentrate on crafting predictive models and generating business insights for seven specified new products. Deliverables include a PowerPoint presentation highlighting key insights, an annotated report delineating the code and model outputs, and all relevant code posted to GitHub for transparency. Future expansions may encompass tailored models for specific brands, locations, or periods to further refine forecasting accuracy.

Details:

The project team comprises four analysts: Ian Donaldson, Michael Tom, Andrew Walton, and Jake Jarrard. The project is slated for completion and presentation to the client on April 11, 2024, with key milestones including exploratory data analysis completion by February 25, model building by March 24, and presentation finalization by April 10. This timeline ensures thorough analysis and model development, culminating in a comprehensive presentation of findings to Swire Coca-Cola.

Conclusion:

By accurately forecasting demand for innovative limited-release beverages, Swire Coca-Cola can make informed decisions to optimize inventory levels, enhance customer satisfaction, and solidify its position as an industry leader in beverage innovation. The collaborative efforts of the analytics team will furnish Swire with the insights needed to navigate the complexities of new product launches successfully.

Purpose of Notebook

The purpose of this EDA notebook is to identify and analyze the Swire Coca-Cola data set looking for hints and clues on guidance of forecasting innovation products in a time series format. Additional feature engineering and data cleaning will be performed to prepare the data for time series forecasting and machine learning models.

Questions about Data to Answer

1. What are the unique features in the data set?
2. What are the data types for each feature?
3. What are the missing values in the data set?
4. Are there unique dates in the data set?
5. Are there unique products in the data set?
6. Are there unique sales territories in the data set?
7. What are Swire's unqiue innovation products (flavors, packaging, etc)?
8. What are top/bottom selling brands?
9. How do competitors stack up in terms of sales?
10. What seasons are key revenue drivers for Swire?
11. Are there intersting financial metrics?
12. How long do innovation products typically run?
13. How to identify week 0/start/tenure for each innovaiton product?

Description of the Data

The data set is a time series data set with 24,000,000+ observations and 13 features. The data set contains sales data for Swire Coca-Cola products across 13 western states. The data set contains sales data for 3 years. The data set contains sales data for 3,100 unique products. The data set contains sales data for 13 unique sales territories. The data set contains sales data for 7 unique

innovation products. The data set contains sales data for 10 unique brands. The data set contains sales data for 3 unique competitors. The data set contains sales data for 4 unique seasons. The data set contains sales data for 3 unique financial metrics. The data set contains sales data for 3 unique tenure metrics.

Discussion of Missing Data

The data set contains NA values for CALORIC_SEGMENT in its raw form comprising 0.2% (59725 missing values) of the data. Using text analysis on ITEM description, imputation was performed to determine the observation's CALORIC_SEGMENT as either diet/light or regular.

Exploratory Visualizations and/or Summary Tables

```
# Load the data
df <- readRDS("swire_no_nas.rds")
#write to csv for other tools
#write.csv(df, "swire_no_nas.csv")
```

> Load the data, packages, and set things up.

```
#season column based on date
df <- df %>%
  mutate(MONTH = month(ymd(df$DATE)), # Extract month from the date
         SEASON = case_when(
           MONTH %in% c(12, 1, 2) ~ "WINTER",
           MONTH %in% c(3, 4, 5) ~ "SPRING",
           MONTH %in% c(6, 7, 8) ~ "SUMMER",
           MONTH %in% c(9, 10, 11) ~ "FALL",
           TRUE ~ NA_character_
         ))

#sales per unit
df <- df %>%
  mutate(SINGLE_PRICE = DOLLAR_SALES / UNIT_SALES)
```

> Months broken up into seasons for potential feature engineering uses, along with sales per unit.

```
#check df class
class(df)
```

```
## [1] "data.frame"
```

```
#check df structure
str(df)
```

```
## 'data.frame':    24461424 obs. of  13 variables:
##  $ DATE           : chr  "2021-08-21" "2022-05-07" "2022-10-22" "2022-08-13" ...
##  $ MARKET_KEY     : int  524 637 628 216 210 278 220 499 754 895 ...
##  $ CALORIC_SEGMENT: chr  "DIET/LIGHT" "REGULAR" "DIET/LIGHT" "REGULAR" ...
##  $ CATEGORY       : chr  "SSD" "SSD" "ING ENHANCED WATER" "SSD" ...
##  $ UNIT_SALES     : num  69 4 1 3 4 112 21 3 19 57 ...
##  $ DOLLAR_SALES   : num  389.74 30.96 2.25 7.55 25.96 ...
##  $ MANUFACTURER   : chr  "SWIRE-CC" "COCOS" "JOLLYS" "COCOS" ...
##  $ BRAND          : chr  "DIET YAWN" "GORGEOUS ORANGEOUS" "DIGRESS FLAVORED" "CHERRY FIZZ"
```

```
...
##  $ PACKAGE        : chr  "12SMALL 12ONE CUP" "12SMALL 12ONE CUP" "20SMALL MULTI JUG" "1L
MULTI JUG" ...
##  $ ITEM           : chr  "YAWN ZERO SUGAR GENTLE DRINK SUPER-JUICE  DURIAN  CUP 12 LIQUID
SMALL X12" "GORGEOUS SUNSET OUS GENTLE DRINK AVOCADO  CUP 12 LIQUID SMALL X12" "DIGRESS ZERO
NUTRIENT ENHANCED WATER BVRG PURPLE  ZERO CALORIE JUG 20 LIQUID SMALL" "KOOL! RED  GENTLE DRINK
RED  COLA CONTOUR JUG 33.8 LIQUID SMALL" ...
##  $ MONTH          : num  8 5 10 8 1 11 3 11 7 4 ...
##  $ SEASON         : chr  "SUMMER" "SPRING" "FALL" "SUMMER" ...
##  $ SINGLE_PRICE   : num  5.65 7.74 2.25 2.52 6.49 ...
```

```
#check df head
head(df)
```

```
##         DATE MARKET_KEY CALORIC_SEGMENT           CATEGORY UNIT_SALES
## 1 2021-08-21        524      DIET/LIGHT                SSD         69
## 2 2022-05-07        637         REGULAR                SSD          4
## 3 2022-10-22        628      DIET/LIGHT ING ENHANCED WATER          1
## 4 2022-08-13        216         REGULAR                SSD          3
## 5 2022-01-01        210         REGULAR                SSD          4
## 6 2021-11-27        278         REGULAR                SSD        112
##   DOLLAR_SALES MANUFACTURER              BRAND           PACKAGE
## 1       389.74     SWIRE-CC          DIET YAWN 12SMALL 12ONE CUP
## 2        30.96        COCOS GORGEOUS ORANGEOUS 12SMALL 12ONE CUP
## 3         2.25       JOLLYS   DIGRESS FLAVORED 20SMALL MULTI JUG
## 4         7.55        COCOS        CHERRY FIZZ      1L MULTI JUG
## 5        25.96        COCOS          RADIANT'S 12SMALL 12ONE CUP
## 6       179.00     SWIRE-CC   ROOT BEER WONDER      2L MULTI JUG
##                                                                                ITEM
## 1          YAWN ZERO SUGAR GENTLE DRINK SUPER-JUICE  DURIAN  CUP 12 LIQUID SMALL X12
## 2                 GORGEOUS SUNSET OUS GENTLE DRINK AVOCADO  CUP 12 LIQUID SMALL X12
## 3 DIGRESS ZERO NUTRIENT ENHANCED WATER BVRG PURPLE  ZERO CALORIE JUG 20 LIQUID SMALL
## 4                   KOOL! RED  GENTLE DRINK RED  COLA CONTOUR JUG 33.8 LIQUID SMALL
## 5                        RADIANT'S GENTLE DRINK GINGER ALE CUP 12 LIQUID SMALL X12
## 6                        JUMPIN JACKS GENTLE DRINK ROOT BEER JUG 67.6 LIQUID SMALL
##   MONTH SEASON SINGLE_PRICE
## 1     8 SUMMER     5.648406
## 2     5 SPRING     7.740000
## 3    10   FALL     2.250000
## 4     8 SUMMER     2.516667
## 5     1 WINTER     6.490000
## 6    11   FALL     1.598214
```

```
#NAs
sum(is.na(df))
```

```
## [1] 0
```

```
#skim df
skim(df)
```

Data summary

Name	df
Number of rows	24461424

Number of columns	13
Column type frequency:	
character	8
numeric	5
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
DATE	0	1	10	10	0	152	0
CALORIC_SEGMENT	0	1	7	10	0	2	0
CATEGORY	0	1	3	18	0	5	0
MANUFACTURER	0	1	5	8	0	8	0
BRAND	0	1	4	56	0	319	0
PACKAGE	0	1	11	26	0	103	0
ITEM	0	1	26	142	0	3692	0
SEASON	0	1	4	6	0	4	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
MARKET_KEY	0	1	593.14	605.88	1.00	260.00	547.00	845.00	6802.00	
UNIT_SALES	0	1	174.37	857.81	0.04	11.00	40.00	126.00	96776.00	
DOLLAR_SALES	0	1	591.14	3040.54	0.01	36.59	135.05	427.14	492591.07	
MONTH	0	1	6.28	3.43	1.00	3.00	6.00	9.00	12.00	
SINLE_PRICE	0	1	4.53	4.79	0.00	2.00	3.18	5.34	224.99	

> There are 59725 missing values in the original dataset. After imputing NAs based on text data (seperate steps from Jake), we are left with 0 missing values. Many features are characters and should be converted to factors. The data is not duplicated. We have 152 unique DATEs (weeks) in the data set, which indicates that there is weekly data for nearly 3 years without any fulling missing weeks. There are 2 CALORIC_SEGMENTS, diet and regular;5 CATEGORIES; 8 MANUFACTURERS; 319 BRANDS; 103 PACKAGE types; 3692 unique ITEM descriptions.

```
#convert date to date type
df$DATE <- as.Date(df$DATE)



#factorize
df <- df %>%
```

```
  mutate(BRAND = as.character(BRAND),
         PACKAGE = as.character(PACKAGE)) %>%
  mutate(across(c(BRAND, PACKAGE, CATEGORY, MANUFACTURER, SEASON), ~as.factor(.)))


#one hot encode CALORIC_SEGMENT as 0 or 1
df <- df %>%
mutate(across(CALORIC_SEGMENT, ~ifelse(. == "REGULAR", 1, 0)))


# Print the result
#validate data types
summary(df)
```

```
##       DATE              MARKET_KEY     CALORIC_SEGMENT
##  Min.   :2020-12-05   Min.   :   1.0   Min.   :0.0000
##  1st Qu.:2021-08-14   1st Qu.: 260.0   1st Qu.:0.0000
##  Median :2022-04-23   Median : 547.0   Median :1.0000
##  Mean   :2022-04-25   Mean   : 593.1   Mean   :0.5023
##  3rd Qu.:2022-12-31   3rd Qu.: 845.0   3rd Qu.:1.0000
##  Max.   :2023-10-28   Max.   :6802.0   Max.   :1.0000
##
##                CATEGORY          UNIT_SALES        DOLLAR_SALES
##  COFFEE            :  145536   Min.   :    0.04   Min.   :     0.0
##  ENERGY            : 5932087   1st Qu.:   11.00   1st Qu.:    36.6
##  ING ENHANCED WATER: 2452456   Median :   40.00   Median :   135.1
##  SPARKLING WATER   : 3019064   Mean   :  174.37   Mean   :   591.1
##  SSD               :12912281   3rd Qu.:  126.00   3rd Qu.:   427.1
##                                Max.   :96776.00   Max.   :492591.1
##
##    MANUFACTURER                          BRAND
##  JOLLYS  :6921978   CROWN                   : 1239010
##  SWIRE-CC:5763809   REAL-TIME EDITIONS      :  827868
##  COCOS   :5595540   DIGRESS FLAVORED        :  731199
##  PONYS   :2259095   MYTHICAL BEVERAGE ULTRA:  718536
##  BEARS   :1593430   BUBBLE JOY              :  535030
##  ALLYS   :1428416   REAL-TIME               :  531911
##  (Other) : 899156   (Other)                 :19877870
##                PACKAGE             ITEM               MONTH
##  16SMALL MULTI CUP: 3065844   Length:24461424    Min.   : 1.000
##  20SMALL MULTI JUG: 2877015   Class :character   1st Qu.: 3.000
##  12SMALL 12ONE CUP: 2870763   Mode  :character   Median : 6.000
##  2L MULTI JUG     : 1896248                      Mean   : 6.284
##  .5L 6ONE JUG     : 1453399                      3rd Qu.: 9.000
##  12SMALL 8ONE CUP : 1408738                      Max.   :12.000
##  (Other)          :10889417
##     SEASON         SINGLE_PRICE
##  FALL  :5215034   Min.   :  0.00333
##  SPRING:6668574   1st Qu.:  2.00429
##  SUMMER:6164348   Median :  3.18000
##  WINTER:6413468   Mean   :  4.52689
##                   3rd Qu.:  5.34105
##                   Max.   :224.99000
##
```

> After feature data type conversion, the data is ready for exploratory data analysis.

```
#echo BRAND name and spacing
cat("\n DIET SMASH \n")
```

```
##
##  DIET SMASH
```

```
#BRAND == DIET SMASH
df %>%
  filter(BRAND == "DIET SMASH") %>%
  select(MANUFACTURER,CALORIC_SEGMENT, CATEGORY, PACKAGE, ) %>%
  summary()
```

```
##    MANUFACTURER   CALORIC_SEGMENT               CATEGORY
##  SWIRE-CC:17483   Min.   :0       COFFEE            :    0
##  ALLYS   :    0   1st Qu.:0       ENERGY            :    0
##  BEARS   :    0   Median :0       ING ENHANCED WATER:    0
##  COCOS   :    0   Mean   :0       SPARKLING WATER   :    0
##  JOLLYS  :    0   3rd Qu.:0       SSD               :17483
##  JORDYS  :    0   Max.   :0
##  (Other) :    0
##               PACKAGE
##  12SMALL 12ONE CUP:11849
##  2L MULTI JUG     : 5630
##  12SMALL 6ONE CUP :    4
##  .5L 12ONE JUG    :    0
##  .5L 24ONE JUG    :    0
##  .5L 4ONE JUG     :    0
##  (Other)          :    0
```

```
#echo BRAND name and spacing
cat("\n SPARKLING JACCEPTABLETLESTER \n")
```

```
##
##  SPARKLING JACCEPTABLETLESTER
```

```
#BRAND == JACCEPTABLETESTER
df %>%
  filter(BRAND == "SPARKLING JACCEPTABLETLESTER") %>%
  select(MANUFACTURER,CALORIC_SEGMENT, CATEGORY, PACKAGE) %>%
  summary()
```

```
##    MANUFACTURER    CALORIC_SEGMENT                CATEGORY
##  SWIRE-CC:299697   Min.   :0.0000   COFFEE            :     0
##  ALLYS   :     0   1st Qu.:0.0000   ENERGY            :     0
##  BEARS   :     0   Median :1.0000   ING ENHANCED WATER:     0
##  COCOS   :     0   Mean   :0.7275   SPARKLING WATER   : 81682
##  JOLLYS  :     0   3rd Qu.:1.0000   SSD               :218015
##  JORDYS  :     0   Max.   :1.0000
##  (Other) :     0
##                       PACKAGE
##  1L MULTI JUG             :65127
##  ALL OTHER ONES           :52678
##  7.5SMALL 6ONE CUP        :50855
```

```
##  10SMALL 6ONE PLASTICS JUG:35845
##  2L MULTI JUG             :27326
##  20SMALL MULTI JUG        :27106
##  (Other)                  :40760
```

```
#echo BRAND name and spacing
cat("\n VENOMOUS BLAST \n")
```

```
##
##  VENOMOUS BLAST
```

```
#BRAND == VENOMOUS BLAST
df %>%
  filter(BRAND == "VENOMOUS BLAST") %>%
  select(MANUFACTURER,CALORIC_SEGMENT, CATEGORY, PACKAGE) %>%
  summary()
```

```
##    MANUFACTURER   CALORIC_SEGMENT                CATEGORY
##  SWIRE-CC:51756   Min.   :0.0000   COFFEE            :    0
##  ALLYS   :    0   1st Qu.:0.0000   ENERGY            :51756
##  BEARS   :    0   Median :1.0000   ING ENHANCED WATER:    0
##  COCOS   :    0   Mean   :0.7449   SPARKLING WATER   :    0
##  JOLLYS  :    0   3rd Qu.:1.0000   SSD               :    0
##  JORDYS  :    0   Max.   :1.0000
##  (Other) :    0
##               PACKAGE
##  16SMALL MULTI CUP:51728
##  8SMALL MULTI CUP :   19
##  16SMALL MULTI JUG:    9
##  .5L 12ONE JUG    :    0
##  .5L 24ONE JUG    :    0
##  .5L 4ONE JUG     :    0
##  (Other)          :    0
```

```
#echo BRAND name and spacing
cat("\n SQUARE \n")
```

```
##
##  SQUARE
```

```
#BRAND == SQUARE
df %>%
  filter(BRAND == "SQUARE") %>%
  select(MANUFACTURER,CALORIC_SEGMENT, CATEGORY, PACKAGE) %>%
  summary()
```

```
##    MANUFACTURER  CALORIC_SEGMENT                CATEGORY
##  SWIRE-CC:7017   Min.   :0.0000   COFFEE            :   0
##  ALLYS   :   0   1st Qu.:1.0000   ENERGY            :   0
##  BEARS   :   0   Median :1.0000   ING ENHANCED WATER:   0
##  COCOS   :   0   Mean   :0.7881   SPARKLING WATER   :7015
##  JOLLYS  :   0   3rd Qu.:1.0000   SSD               :   2
##  JORDYS  :   0   Max.   :1.0000
```

```
##  (Other) :   0
##                PACKAGE
##  20SMALL MULTI JUG  :6641
##  ALL OTHER ONES     : 347
##  2L MULTI JUG       :  27
##  .5L MLT SHADYES JUG:   1
##  1.5L MULTI JUG     :   1
##  .5L 12ONE JUG      :   0
##  (Other)            :   0
```

```
#echo BRAND name and spacing
cat("\n GREETINGLE \n")
```

```
##
##  GREETINGLE
```

```
#BRAND == GREETINGLE
df %>%
  filter(BRAND == "GREETINGLE") %>%
  select(MANUFACTURER,CALORIC_SEGMENT, CATEGORY, PACKAGE) %>%
  summary()
```

```
##    MANUFACTURER    CALORIC_SEGMENT               CATEGORY
##  SWIRE-CC:491300   Min.   :0       COFFEE            :     0
##  ALLYS   :     0   1st Qu.:0       ENERGY            :     0
##  BEARS   :     0   Median :0       ING ENHANCED WATER:491300
##  COCOS   :     0   Mean   :0       SPARKLING WATER   :     0
##  JOLLYS  :     0   3rd Qu.:0       SSD               :     0
##  JORDYS  :     0   Max.   :0
##  (Other) :     0
##               PACKAGE
##  18SMALL MULTI JUG:373131
##  18SMALL 6ONE     : 86750
##  .5L 6ONE JUG     : 23207
##  ALL OTHER ONES   :  7845
##  .5L 12ONE JUG    :   367
##  .5L 24ONE JUG    :     0
##  (Other)          :     0
```

```
#echo BRAND name and spacing
cat("\n DIET MOONLIT \n")
```

```
##
##  DIET MOONLIT
```

```
#BRAND == DIET MOONLIT
df %>%
  filter(BRAND == "DIET MOONLIT") %>%
  select(MANUFACTURER,CALORIC_SEGMENT, CATEGORY, PACKAGE) %>%
  summary()
```

```
##    MANUFACTURER   CALORIC_SEGMENT               CATEGORY
##  SWIRE-CC:75948   Min.   :0       COFFEE            :    0
```

```
##  ALLYS   :    0   1st Qu.:0       ENERGY            :    0
##  BEARS   :    0   Median :0       ING ENHANCED WATER:    0
##  COCOS   :    0   Mean   :0       SPARKLING WATER   :    0
##  JOLLYS  :    0   3rd Qu.:0       SSD               :75948
##  JORDYS  :    0   Max.   :0
##  (Other) :    0
##               PACKAGE
##  2L MULTI JUG     :28140
##  12SMALL 12ONE CUP:28006
##  .5L 6ONE JUG     :12023
##  20SMALL MULTI JUG: 7684
##  12SMALL 6ONE CUP :   94
##  ALL OTHER ONES   :    1
##  (Other)          :    0
```

```
#echo BRAND name and spacing
cat("\n PEPPY \n")
```

```
##
##  PEPPY
```

```
#BRAND == PEPPY
df %>%
  filter(BRAND == "PEPPY") %>%
  select(MANUFACTURER,CALORIC_SEGMENT, CATEGORY, PACKAGE) %>%
  summary()
```

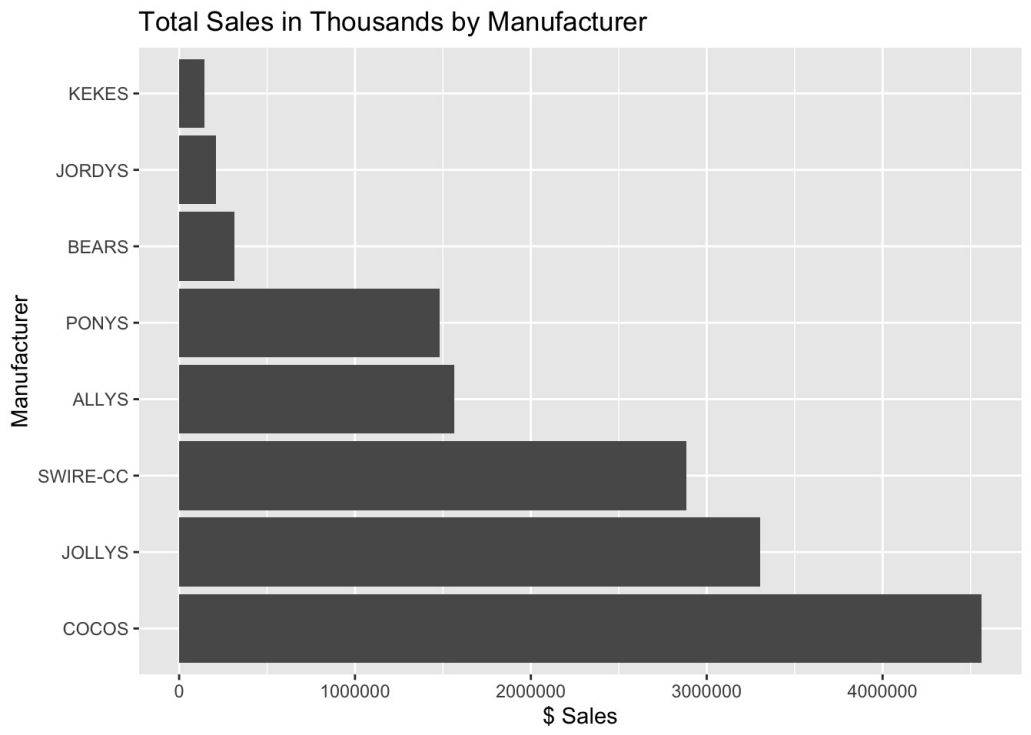
```
##    MANUFACTURER    CALORIC_SEGMENT               CATEGORY
##  SWIRE-CC:399458   Min.   :1       COFFEE            :     0
##  ALLYS   :     0   1st Qu.:1       ENERGY            :     0
##  BEARS   :     0   Median :1       ING ENHANCED WATER:     0
##  COCOS   :     0   Mean   :1       SPARKLING WATER   :     0
##  JOLLYS  :     0   3rd Qu.:1       SSD               :399458
##  JORDYS  :     0   Max.   :1
##  (Other) :     0
##               PACKAGE
##  20SMALL MULTI JUG: 34546
##  2L MULTI JUG     : 29397
##  .5L 6ONE JUG     : 29396
##  12SMALL 12ONE CUP: 29396
##  7.5SMALL 6ONE CUP: 29389
##  1L MULTI JUG     : 29300
##  (Other)          :218034
```

> The Diet Smash product is a diet product, in the energy category, and comes in 3 packaging types (Innovation - Packaging?). The Sparkling Jacceptabletlester brand comes in both diet/regular, straddles both sparkling water and sparkling soda drink categories, and comes in multiple different package types (but not innovation package “16small multi cup”). The Venomous Blast product comes in both diet and regular, in the energy category product, and comes in 3 packaging types (almost exclusively “16small multi cup”, with two short term release sizes, but not future “innovation package?” “16 liquid small”). The Square brand comes in both diet and regular, in the 2 categories (sparkling water and ssd), and comes in 5 packaging types (several have extremely small counts - full innovation packaging?). The Greetingle brand comes in only diet, in the ING Enhanced Water category, and comes in 6 packaging types (all with relatively legit numbers, with the exception of one size). The Diet Moonlit brand comes in only diet, in the ssd category, and comes in multiple packaging types (all with legit numbers). The Peppy brand comes in only regular, in the ssd category, and comes in at least 6 packaging types (all with legit numbers, other could be explored more).

```
# Turn off scientific notation
options(scipen = 999)


#sales in thousands by manufacturer
df %>%
  group_by(MANUFACTURER) %>%
  summarise(TOTAL_SALES = sum(DOLLAR_SALES)) %>%
  arrange(desc(TOTAL_SALES))
```

```
## # A tibble: 8 × 2
##   MANUFACTURER TOTAL_SALES
##   <fct>              <dbl>
## 1 COCOS        4563306476.
## 2 JOLLYS       3301641671.
## 3 SWIRE-CC     2885435787.
## 4 ALLYS        1562675378.
## 5 PONYS        1481611289.
## 6 BEARS         312718094.
## 7 JORDYS        209238104.
## 8 KEKES         143501825.
```

```
#graph sales in thousands by manufacturer

df %>%
  group_by(MANUFACTURER) %>%
  summarise(TOTAL_SALES = sum(DOLLAR_SALES/1000)) %>%
  ggplot(aes(x = reorder(MANUFACTURER, -TOTAL_SALES), y = TOTAL_SALES)) +
  geom_col() +
  coord_flip() +
  labs(title = "Total Sales in Thousands by Manufacturer",
       x = "Manufacturer",
       y = "$ Sales")
```

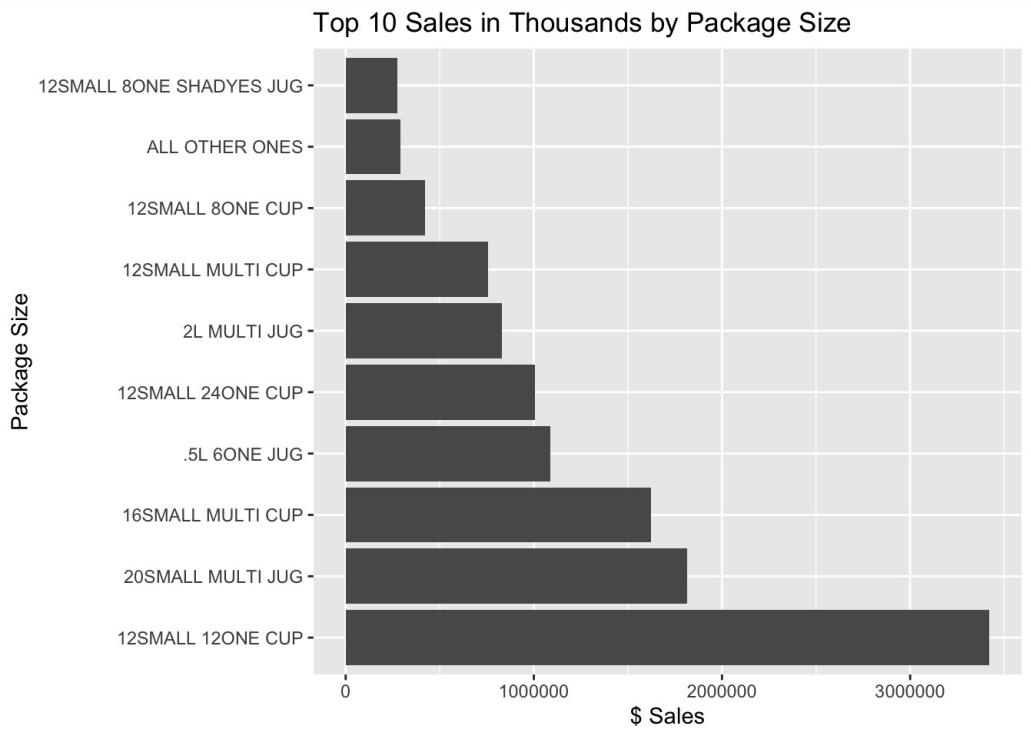



```
#sales in thousands by top 10 package size
```

```
df %>%
  group_by(PACKAGE) %>%
  summarise(TOTAL_SALES = sum(DOLLAR_SALES/1000)) %>%
  arrange(desc(TOTAL_SALES)) %>%
  head(10)
```

```
## # A tibble: 10 × 2
##    PACKAGE                  TOTAL_SALES
##    <fct>                          <dbl>
##  1 12SMALL 12ONE CUP           3422060.
##  2 20SMALL MULTI JUG           1814841.
##  3 16SMALL MULTI CUP           1621832.
##  4 .5L 6ONE JUG                1089477.
##  5 12SMALL 24ONE CUP           1007780.
##  6 2L MULTI JUG                 830755.
##  7 12SMALL MULTI CUP            758168.
##  8 12SMALL 8ONE CUP             422976.
##  9 ALL OTHER ONES               291378.
## 10 12SMALL 8ONE SHADYES JUG     274115.
```

```
#graph sales in thousands by top 10 ten package size
df %>%
  group_by(PACKAGE) %>%
  summarise(TOTAL_SALES = sum(DOLLAR_SALES/1000)) %>%
  arrange(desc(TOTAL_SALES)) %>%
  head(10) %>%
  ggplot(aes(x = reorder(PACKAGE, -TOTAL_SALES), y = TOTAL_SALES)) +
  geom_col() +
  coord_flip() +
  labs(title = "Top 10 Sales in Thousands by Package Size",
       x = "Package Size",
       y = "$ Sales")
```

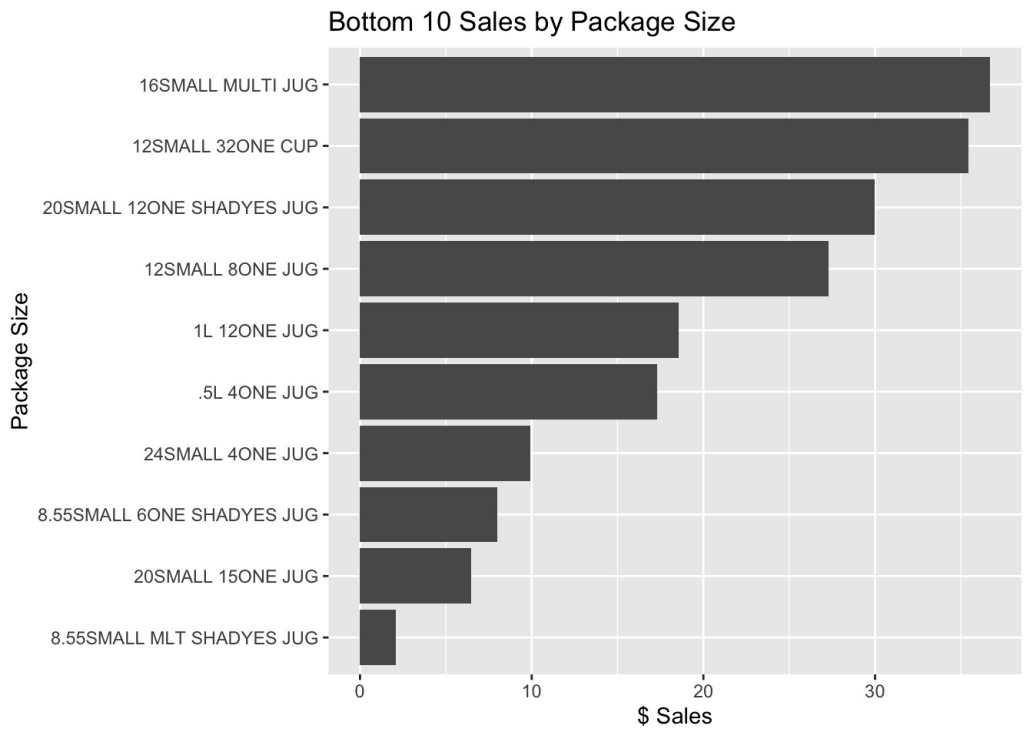



```
#bottom 10 sales in throusands by package size
df %>%
```

```
  group_by(PACKAGE) %>%
  summarise(TOTAL_SALES = sum(DOLLAR_SALES)) %>%
  arrange(TOTAL_SALES) %>%
  head(10)
```

```
## # A tibble: 10 × 2
##    PACKAGE                    TOTAL_SALES
##    <fct>                            <dbl>
##  1 8.55SMALL MLT SHADYES JUG         2.1
##  2 20SMALL 15ONE JUG                 6.49
##  3 8.55SMALL 6ONE SHADYES JUG        8
##  4 24SMALL 4ONE JUG                  9.94
##  5 .5L 4ONE JUG                     17.3
##  6 1L 12ONE JUG                     18.5
##  7 12SMALL 8ONE JUG                 27.3
##  8 20SMALL 12ONE SHADYES JUG        30.0
##  9 12SMALL 32ONE CUP                35.5
## 10 16SMALL MULTI JUG                36.7
```

```
#graph sales in thousands by bottom 10 package size
df %>%
  group_by(PACKAGE) %>%
  summarise(TOTAL_SALES = sum(DOLLAR_SALES)) %>%
  arrange(TOTAL_SALES) %>%
  head(10) %>%
  ggplot(aes(x = reorder(PACKAGE, TOTAL_SALES), y = TOTAL_SALES)) +
  geom_col() +
  coord_flip() +
  labs(title = "Bottom 10 Sales by Package Size",
       x = "Package Size",
       y = "$ Sales")
```

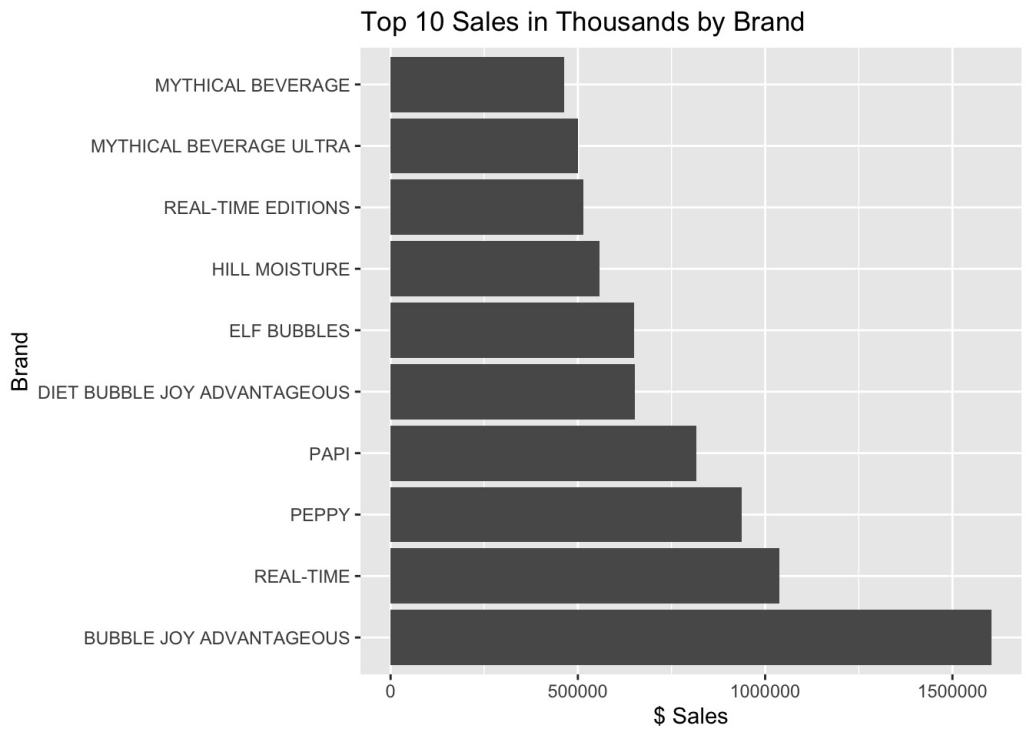



```
#top 10 sales in thousands by brand
df %>%
  group_by(BRAND) %>%
```

```
  summarise(TOTAL_SALES = sum(DOLLAR_SALES/1000)) %>%
  arrange(desc(TOTAL_SALES)) %>%
  head(10)
```

```
## # A tibble: 10 × 2
##    BRAND                         TOTAL_SALES
##    <fct>                               <dbl>
##  1 BUBBLE JOY ADVANTAGEOUS         1604867.
##  2 REAL-TIME                       1037072.
##  3 PEPPY                            937030.
##  4 PAPI                             816414.
##  5 DIET BUBBLE JOY ADVANTAGEOUS     653091.
##  6 ELF BUBBLES                      649144.
##  7 HILL MOISTURE                    557565.
##  8 REAL-TIME EDITIONS               514215.
##  9 MYTHICAL BEVERAGE ULTRA          499398.
## 10 MYTHICAL BEVERAGE                462803.
```

```
#graph sales in thousands by top 10 brand
df %>%
  group_by(BRAND) %>%
  summarise(TOTAL_SALES = sum(DOLLAR_SALES/1000)) %>%
  arrange(desc(TOTAL_SALES)) %>%
  head(10) %>%
  ggplot(aes(x = reorder(BRAND, -TOTAL_SALES), y = TOTAL_SALES)) +
  geom_col() +
  coord_flip() +
  labs(title = "Top 10 Sales in Thousands by Brand",
       x = "Brand",
       y = "$ Sales")
```

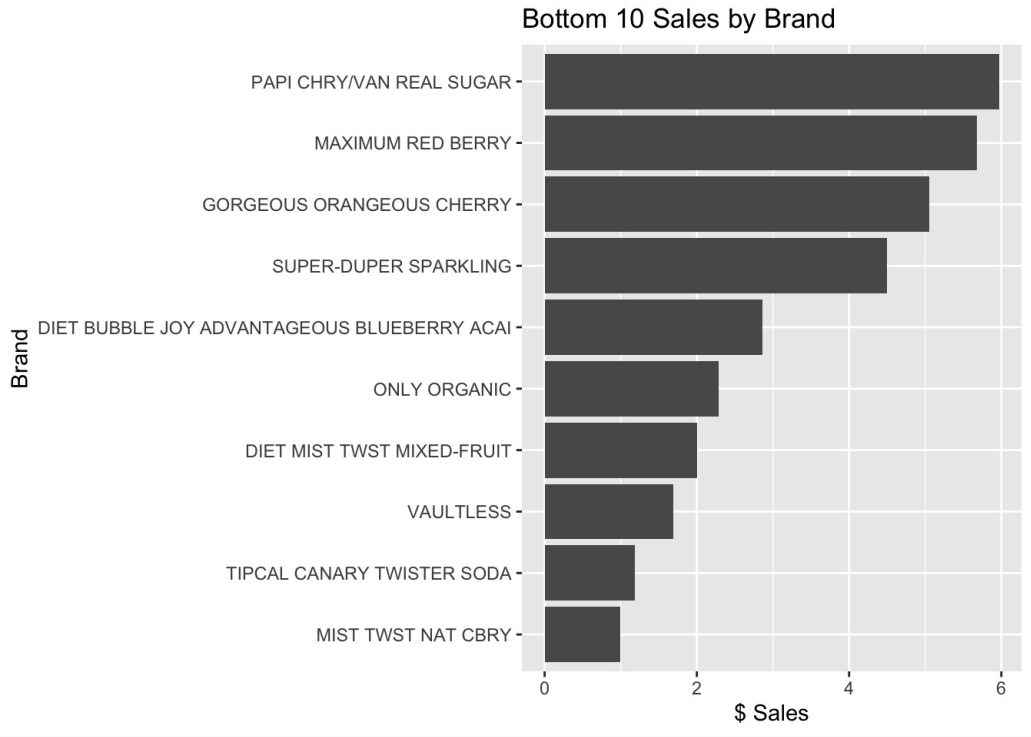



```
#bottom 10 sales in thousands by brand
df %>%
  group_by(BRAND) %>%
  summarise(TOTAL_SALES = sum(DOLLAR_SALES)) %>%
```

```
  arrange(TOTAL_SALES) %>%
  head(10)
```

```
## # A tibble: 10 × 2
##    BRAND                                        TOTAL_SALES
##    <fct>                                              <dbl>
##  1 MIST TWST NAT CBRY                                  0.99
##  2 TIPCAL CANARY TWISTER SODA                          1.19
##  3 VAULTLESS                                           1.69
##  4 DIET MIST TWST MIXED-FRUIT                          2
##  5 ONLY ORGANIC                                        2.29
##  6 DIET BUBBLE JOY ADVANTAGEOUS BLUEBERRY ACAI         2.86
##  7 SUPER-DUPER SPARKLING                               4.5
##  8 GORGEOUS ORANGEOUS CHERRY                           5.05
##  9 MAXIMUM RED BERRY                                   5.68
## 10 PAPI CHRY/VAN REAL SUGAR                            5.97
```

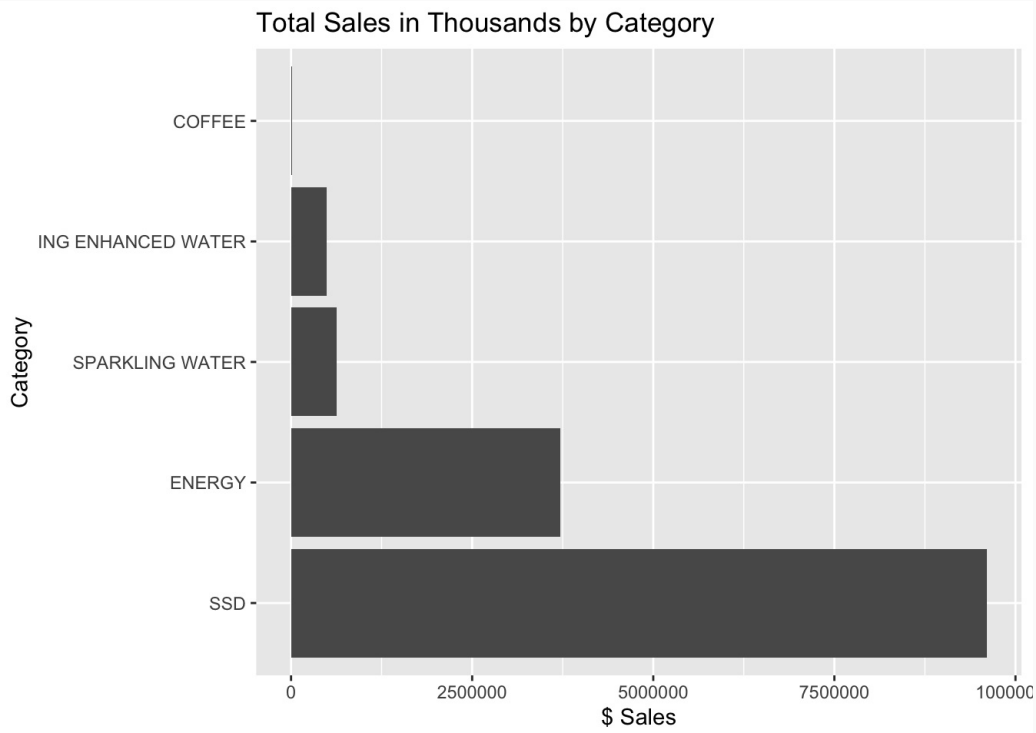
```
#graph sales in thousands by bottom 10 brand
df %>%
  group_by(BRAND) %>%
  summarise(TOTAL_SALES = sum(DOLLAR_SALES)) %>%
  arrange(TOTAL_SALES) %>%
  head(10) %>%
  ggplot(aes(x = reorder(BRAND, TOTAL_SALES), y = TOTAL_SALES)) +
  geom_col() +
  coord_flip() +
  labs(title = "Bottom 10 Sales by Brand",
       x = "Brand",
       y = "$ Sales")
```




```
#sales in thousands by category
df %>%
  group_by(CATEGORY) %>%
  summarise(TOTAL_SALES = sum(DOLLAR_SALES/1000)) %>%
  arrange(desc(TOTAL_SALES))
```

```
## # A tibble: 5 × 2
##   CATEGORY           TOTAL_SALES
##   <fct>                    <dbl>
## 1 SSD                   9606515.
## 2 ENERGY                3718445.
## 3 SPARKLING WATER        626470.
## 4 ING ENHANCED WATER     494293.
## 5 COFFEE                  14405.
```

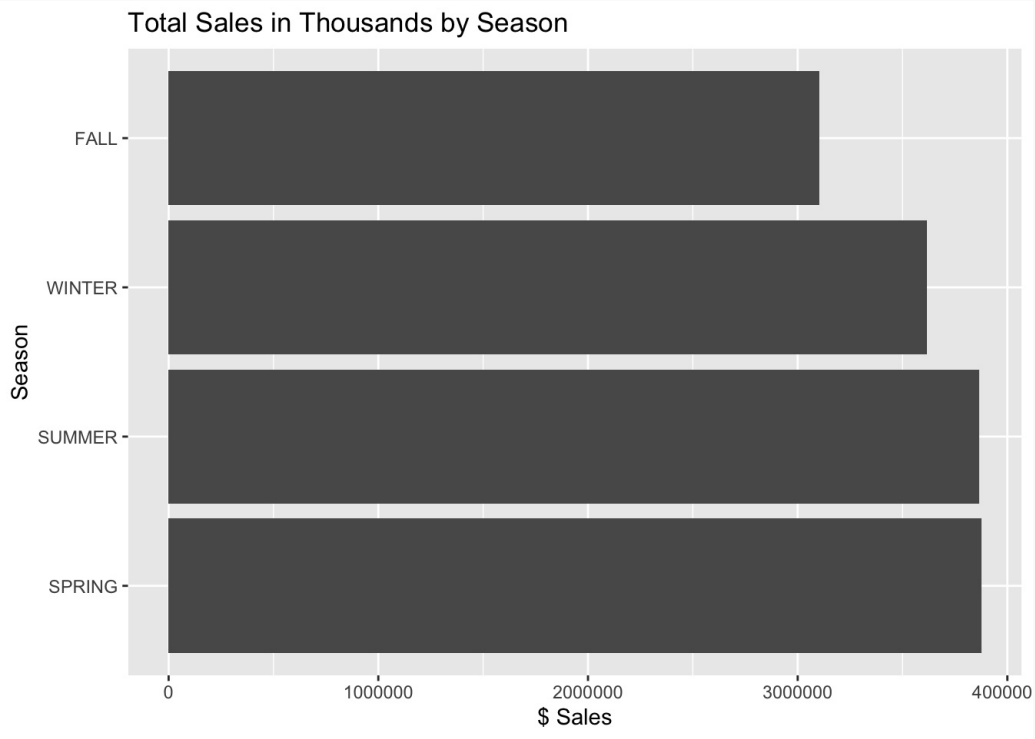
```
#graph sales in thousands by CATEGORY
df %>%
  group_by(CATEGORY) %>%
  summarise(TOTAL_SALES = sum(DOLLAR_SALES/1000)) %>%
  ggplot(aes(x = reorder(CATEGORY, -TOTAL_SALES), y = TOTAL_SALES)) +
  geom_col() +
  coord_flip() +
  labs(title = "Total Sales in Thousands by Category",
       x = "Category",
       y = "$ Sales")
```




```
#sales in thousands by season
df %>%
  group_by(SEASON) %>%
  summarise(TOTAL_SALES = sum(DOLLAR_SALES/1000)) %>%
  arrange(desc(TOTAL_SALES))
```

```
## # A tibble: 4 × 2
##   SEASON TOTAL_SALES
##   <fct>        <dbl>
## 1 SPRING    3874929.
## 2 SUMMER    3865217.
## 3 WINTER    3616622.
## 4 FALL      3103362.
```

```
#graph sales by SEASON
df %>%
  group_by(SEASON) %>%
  summarise(TOTAL_SALES = sum(DOLLAR_SALES/1000)) %>%
  ggplot(aes(x = reorder(SEASON, -TOTAL_SALES), y = TOTAL_SALES)) +
  geom_col() +
  coord_flip() +
  labs(title = "Total Sales in Thousands by Season",
       x = "Season",
       y = "$ Sales")
```



> Swire is in 3rd place for overall sales behind Jollys and Cocos for overall sales by manufacturer. The top 3 package sizes range from 12small 12one cup, 20small multi jug, and 16small multi cup. Bottom 3 sales by package size are 8.55small mlt shadyes jug, 20 small 15one jug, and 8.55small 6one shadyes jug. Bottom packges sizes are likely innovation package that did not do so well based on extremely small sales (<$40K). Bubble Joy Advantageous (a Coco's regular soda) is a clear winner in sales by brand, followed by real-time (Ally's primarily energy drink in diet/regular), and peppy (Swire-CC's regular soda). Bottom 10 sales by brand all falls into extremely small buckets of sales less than $6k, are these innovation product failures? The bottom three are Mist Twst Nat Cbry (Jolly's 1 single sale of regular soda), Tipcal Canary Twister Soda (1 single sale of Jolly's regular soda), and Vaultless (1 single sale of Coco's diet soda). The sparkling soda drink category is more than double the next (energy). ING Enhanced Water and Sparkling Water register as notable bottled drinks, but coffee barely scratches the surface in terms of sales. Summer and Spring are roughly about the same in terms of sales, followed by a slight drop in Winter, and a more noticeable drop for Fall.

```
#top 10 longest running brands
df %>%
  group_by(BRAND) %>%
  summarise(LENGTH = n_distinct(DATE)) %>%
  arrange(desc(LENGTH)) %>%
  head(10)
```

```
## # A tibble: 10 × 2
##    BRAND                       LENGTH
##    <fct>                        <int>
##  1 CARBONATE STREAM               152
##  2 CUPADA ARID                    152
```

```
##  3 RADIANT'S                       152
##  4 SINGLE GROUP                    152
##  5 SPARKLING JACCEPTABLETLESTER    152
##  6 BUBBLE JOY                      148
##  7 CARBONATE STREAM WATERS         148
##  8 CROWN                           148
##  9 DIGRESS FLAVORED                148
## 10 EXCLAMATION SODA                148
```

```
#shortest 10 running brands
df %>%
  group_by(BRAND) %>%
  summarise(LENGTH = n_distinct(DATE)) %>%
  arrange(LENGTH) %>%
  head(10)
```

```
## # A tibble: 10 × 2
##    BRAND                                                LENGTH
##    <fct>                                                 <int>
##  1 LAUGHING MYTHICAL BEVERAGE CHAI HAI                       1
##  2 MIST TWST NAT CBRY                                        1
##  3 ONLY ORGANIC                                              1
##  4 PAPI VANILLA REAL SUGAR                                   1
##  5 TIPCAL CANARY TWISTER SODA                                1
##  6 VAULTLESS                                                 1
##  7 DIET BUBBLE JOY ADVANTAGEOUS BLUEBERRY ACAI               2
##  8 DIET BUBBLE JOY ADVANTAGEOUS PLANT-BASED SWEENTENERS      2
##  9 DIET MIST TWST MIXED-FRUIT                                2
## 10 DIET MUTANT                                               2
```

```
#median length of time for a brand
df %>%
  group_by(BRAND) %>%
  summarise(LENGTH = n_distinct(DATE)) %>%
  summarise(median(LENGTH))
```

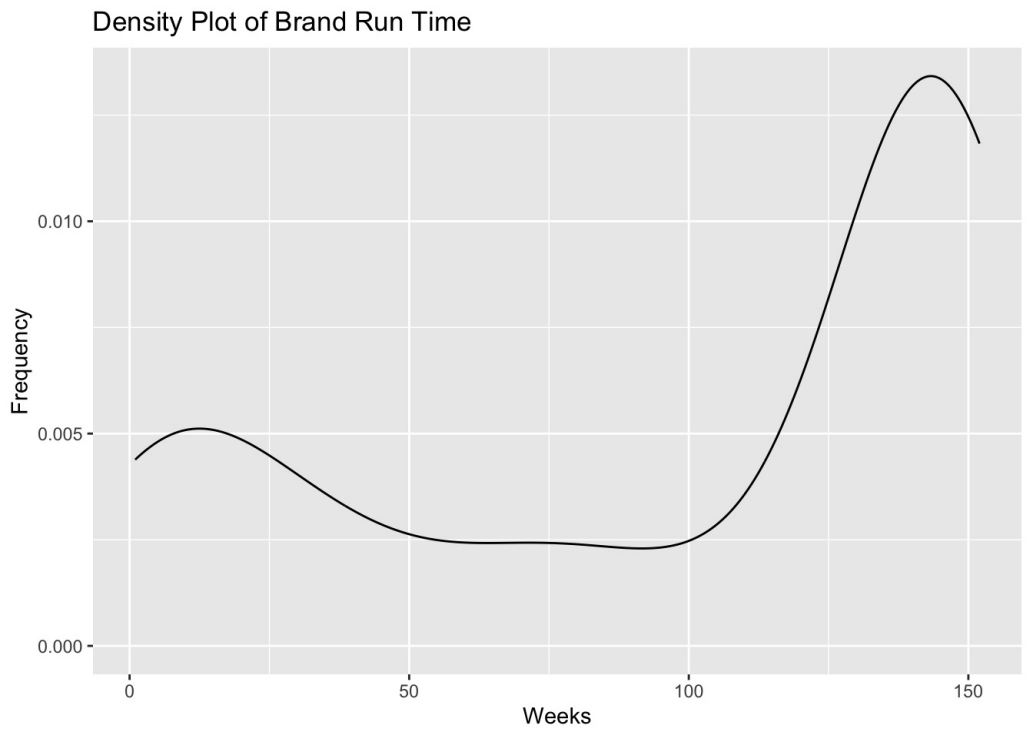
```
## # A tibble: 1 × 1
##   `median(LENGTH)`
##              <int>
## 1              137
```

```
#mean length of time for a brand
df %>%
  group_by(BRAND) %>%
  summarise(LENGTH = n_distinct(DATE)) %>%
  summarise(mean(LENGTH))
```

```
## # A tibble: 1 × 1
##   `mean(LENGTH)`
##            <dbl>
## 1           99.1
```

```
#density plot of brand run time
df %>%
```

```
  group_by(BRAND) %>%
  summarise(LENGTH = n_distinct(DATE)) %>%
  ggplot(aes(x = LENGTH)) +
  geom_density() +
  labs(title = "Density Plot of Brand Run Time",
       x = "Weeks",
       y = "Frequency")
```




```
#brands that run for less than 6 months
df %>%
  group_by(BRAND) %>%
  summarise(LENGTH = n_distinct(DATE)) %>%
  filter(LENGTH < 26)
```

```
## # A tibble: 63 × 2
##    BRAND                                              LENGTH
##    <fct>                                               <int>
##  1 BARS                                                    4
##  2 BUBBLE JOY ADVANTAGEOUS W/LIME                         14
##  3 CLEAR RADIANCE PAPI                                    24
##  4 CUPADA ARID REMAINING                                  24
##  5 DIET BUBBLE JOY ADVANTAGEOUS BLUEBERRY ACAI             2
##  6 DIET BUBBLE JOY ADVANTAGEOUS PLANT-BASED SWEENTENERS    2
##  7 DIET BUBBLE JOY ADVANTAGEOUS STRAWBERRY GUAVA           3
##  8 DIET BUBBLE JOY ADVANTAGEOUS W/LIME                    21
##  9 DIET DROPTOP                                           10
## 10 DIET HILL MOISTURE ELECTRICITY                         21
## # ℹ 53 more rows
```

```
#summarize features of brands that run for less than 6 months
df %>%
  group_by(BRAND) %>%
  summarise(LENGTH = n_distinct(DATE)) %>%
  filter(LENGTH < 6) %>%
  left_join(df, by = "BRAND") %>%
```

```
  select(BRAND, CATEGORY, SEASON, PACKAGE, MANUFACTURER) %>%
  distinct()
```

```
## # A tibble: 61 × 5
##    BRAND                                    CATEGORY SEASON PACKAGE MANUFACTURER
##    <fct>                                    <fct>    <fct>  <fct>   <fct>
##  1 BARS                                     SSD      FALL   2L MUL… COCOS
##  2 DIET BUBBLE JOY ADVANTAGEOUS BLUEBERRY … SSD      SUMMER 12SMAL… COCOS
##  3 DIET BUBBLE JOY ADVANTAGEOUS BLUEBERRY … SSD      WINTER 12SMAL… COCOS
##  4 DIET BUBBLE JOY ADVANTAGEOUS PLANT-BASE… SSD      SUMMER 7.5SMA… COCOS
##  5 DIET BUBBLE JOY ADVANTAGEOUS STRAWBERRY… SSD      SPRING 12SMAL… COCOS
##  6 DIET BUBBLE JOY ADVANTAGEOUS STRAWBERRY… SSD      WINTER 12SMAL… COCOS
##  7 DIET BUBBLE JOY ADVANTAGEOUS STRAWBERRY… SSD      WINTER 12SMAL… COCOS
##  8 DIET MIST TWST MIXED-FRUIT               SSD      WINTER 2L MUL… JOLLYS
##  9 DIET MIST TWST MIXED-FRUIT               SSD      SUMMER 2L MUL… JOLLYS
## 10 DIET MUTANT                              SSD      SPRING 20SMAL… PONYS
## # ℹ 51 more rows
```

> The top 10 brands that run 148 or 152 weeks do not include some of the top sales by brands (are there missing weeks?). The top 10 shortest running brands include bottom brands by sales and only ran for 1 or two weeks, which makes sense if something only registered one single sale. The median length of a brand is 137 weeks, with mean brand run time falling in a 99.1 weeks. Therefore, the data likely exhibits left skewness, indicating that there are some brands with very short run times (which pull down the mean), while the median is higher due to the influence of some brands with longer run times. The histogram of brand run time shows a left skew with several stalwart brands running for a very long time, and a lot of brands, including innovation types, running for a shorter duration. As can be seen in the density plot there are two main humps or modes, one that clusters between 0 and ~25 weeks (6 months) trending down to a flatter line between week 50 and week 100, and another much larger cluster presumably dominated by the always on the shlef types cluster between 125 and 152 weeks. It may be worth looking for week 0/launch/tenure date spikes for those products that runs less than 6 months in order to determine start/stop times for any ARIMA/Time-Series models later.

```
#top 10 longest running packages
df %>%
  group_by(PACKAGE) %>%
  summarise(LENGTH = n_distinct(DATE)) %>%
  arrange(desc(LENGTH)) %>%
  head(10)
```

```
## # A tibble: 10 × 2
##    PACKAGE           LENGTH
##    <fct>              <int>
##  1 .5L 12ONE JUG        152
##  2 .5L 24ONE JUG        152
##  3 .5L 6ONE JUG         152
##  4 12SMALL 12ONE CUP    152
##  5 12SMALL 15ONE CUP    152
##  6 12SMALL 24ONE CUP    152
##  7 12SMALL 6ONE CUP     152
##  8 16SMALL MULTI CUP    152
##  9 1L MULTI JUG         152
## 10 20SMALL 24ONE JUG    152
```

```
#shortest 10 running packages
df %>%
  group_by(PACKAGE) %>%
  summarise(LENGTH = n_distinct(DATE)) %>%
```

```
  arrange(LENGTH) %>%
  head(10)
```

```
## # A tibble: 10 × 2
##    PACKAGE                    LENGTH
##    <fct>                       <int>
##  1 1L 12ONE JUG                    1
##  2 20SMALL 15ONE JUG               1
##  3 8.55SMALL 6ONE SHADYES JUG      1
##  4 24SMALL 4ONE JUG                2
##  5 20SMALL 12ONE SHADYES JUG       3
##  6 8.55SMALL MLT SHADYES JUG       3
##  7 .5L 4ONE JUG                    5
##  8 12SMALL 32ONE CUP               6
##  9 12SMALL 8ONE JUG                6
## 10 .5L 8ONE SHADYES JUG            7
```

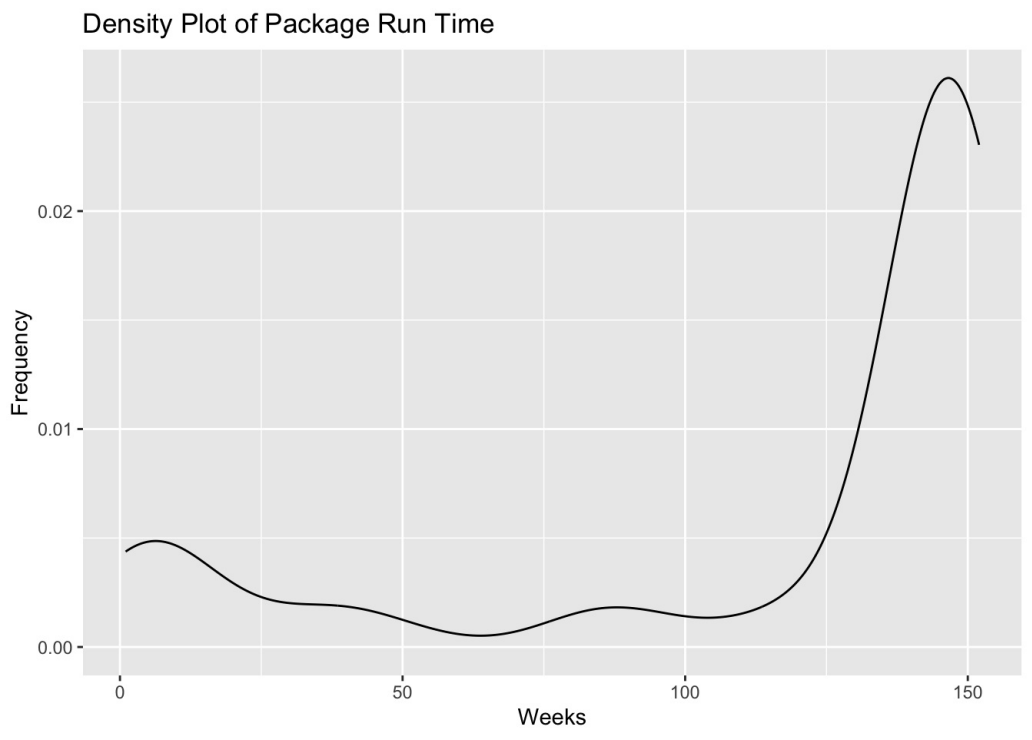
```
#median length of time for a package
df %>%
  group_by(PACKAGE) %>%
  summarise(LENGTH = n_distinct(DATE)) %>%
  summarise(median(LENGTH))
```

```
## # A tibble: 1 × 1
##   `median(LENGTH)`
##              <int>
## 1              147
```

```
#mean length of time for a package
df %>%
  group_by(PACKAGE) %>%
  summarise(LENGTH = n_distinct(DATE)) %>%
  summarise(mean(LENGTH))
```

```
## # A tibble: 1 × 1
##   `mean(LENGTH)`
##            <dbl>
## 1           117.
```

```
#density plot of package run time
df %>%
  group_by(PACKAGE) %>%
  summarise(LENGTH = n_distinct(DATE)) %>%
  ggplot(aes(x = LENGTH)) +
  geom_density() +
  labs(title = "Density Plot of Package Run Time",
       x = "Weeks",
       y = "Frequency")
```




```
#packages that run for less than 6 months
df %>%
  group_by(PACKAGE) %>%
  summarise(LENGTH = n_distinct(DATE)) %>%
  filter(LENGTH < 26)
```

```
## # A tibble: 14 × 2
##    PACKAGE                    LENGTH
##    <fct>                       <int>
##  1 .5L 4ONE JUG                    5
##  2 .5L 8ONE SHADYES JUG            7
##  3 12SMALL 12ONE BUMPY CUP        17
##  4 12SMALL 32ONE CUP               6
##  5 12SMALL 8ONE JUG                6
##  6 16SMALL MULTI JUG              12
##  7 1L 12ONE JUG                    1
##  8 20SMALL 12ONE SHADYES JUG       3
##  9 20SMALL 15ONE JUG               1
## 10 22SMALL MULTI JUG              15
## 11 24SMALL 4ONE JUG                2
## 12 7.5SMALL 10ONE                 12
## 13 8.55SMALL 6ONE SHADYES JUG      1
## 14 8.55SMALL MLT SHADYES JUG       3
```

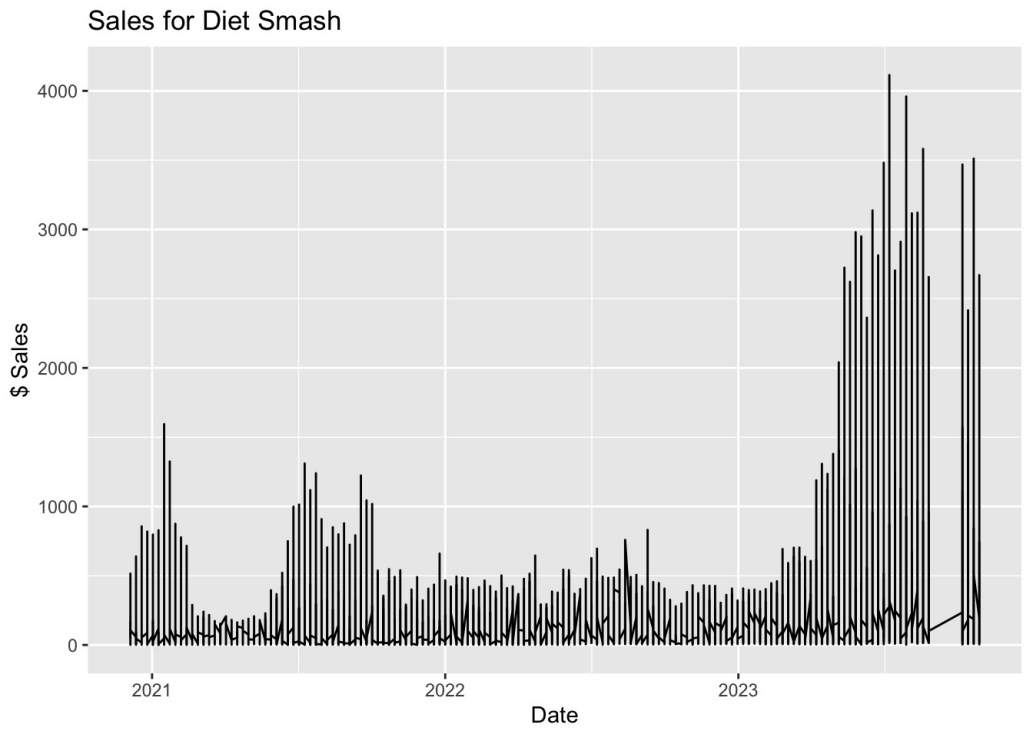
```
#summarize features of packages that run for less than 6 months
df %>%
  group_by(PACKAGE) %>%
  summarise(LENGTH = n_distinct(DATE)) %>%
  filter(LENGTH < 6) %>%
  left_join(df, by = "PACKAGE") %>%
  select(PACKAGE, CATEGORY, SEASON, BRAND, MANUFACTURER) %>%
  distinct()
```

```
## # A tibble: 15 × 5
##    PACKAGE                    CATEGORY          SEASON BRAND       MANUFACTURER
```

```
##    <fct>                        <fct>              <fct>  <fct>       <fct>
##  1 .5L 4ONE JUG                 ING ENHANCED WATER SUMMER VITAMINAL … COCOS
##  2 .5L 4ONE JUG                 ING ENHANCED WATER WINTER VITAMINAL … COCOS
##  3 .5L 4ONE JUG                 ING ENHANCED WATER FALL   VITAMINAL … COCOS
##  4 .5L 4ONE JUG                 ING ENHANCED WATER SPRING VITAMINAL … COCOS
##  5 1L 12ONE JUG                 SPARKLING WATER    WINTER INTELLIGEN… COCOS
##  6 20SMALL 12ONE SHADYES JUG    SSD                WINTER DIET BUBBL… COCOS
##  7 20SMALL 12ONE SHADYES JUG    SSD                SUMMER DIET BUBBL… COCOS
##  8 20SMALL 12ONE SHADYES JUG    SSD                SPRING DIET BUBBL… COCOS
##  9 20SMALL 15ONE JUG            ING ENHANCED WATER WINTER VITAMINAL … COCOS
## 10 24SMALL 4ONE JUG             SSD                SPRING HILL MOIST… JOLLYS
## 11 24SMALL 4ONE JUG             SSD                SUMMER HILL MOIST… JOLLYS
## 12 8.55SMALL 6ONE SHADYES JUG   SSD                FALL   DIET BUBBL… COCOS
## 13 8.55SMALL MLT SHADYES JUG    SSD                FALL   FANTASMIC   COCOS
## 14 8.55SMALL MLT SHADYES JUG    SSD                SPRING FANTASMIC   COCOS
## 15 8.55SMALL MLT SHADYES JUG    SSD                WINTER FANTASMIC   COCOS
```

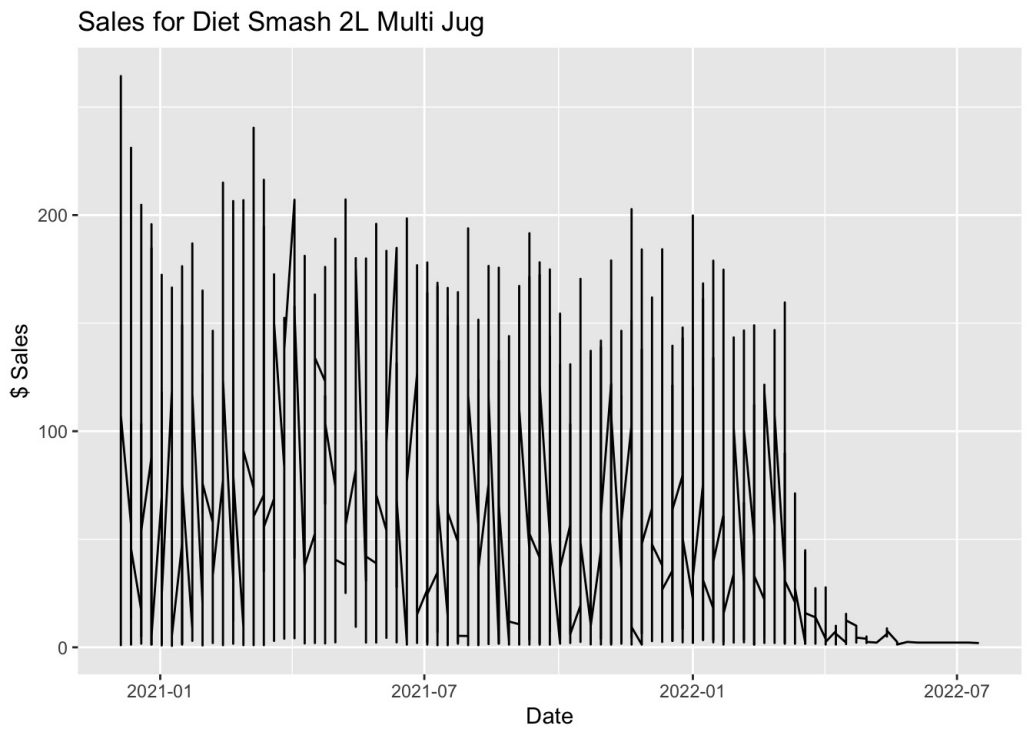
> Top 10 packages run the entire length of the dataset, which is 152 weeks. These are tried and true packaging sizes that we have all likely grown up with, know and love to enjoy our beverages from, whether from a gas station or grocery store. The top 10 shortest packages range from 1 week to 7 weeks. The median package run length is 147, indicating that tried and true packing overwhelmingly dominates distribution sales. The mean package tenure is 117 weeks. Therefore, the data likely exhibits right skewness, indicating that there are some packages with very high tenures (which push up the median), but the mean is lower due to the influence of some packages with shorter tenures. The density plot shows two primary modes, one smaller < 12 weeks, likely innovation dominated, and another larger > 140 weeks, legacy package sizes. There are 14 package sizes that run for less than 6 months or 26 weeks.

```
#graph DOLLAR_SALES by DATE for BRAND == "DIET SMASH"
df %>%
  filter(BRAND == "DIET SMASH") %>%
  ggplot(aes(x = DATE, y = DOLLAR_SALES)) +
  geom_line() +
  labs(title = "Sales for Diet Smash",
       x = "Date",
       y = "$ Sales")
```

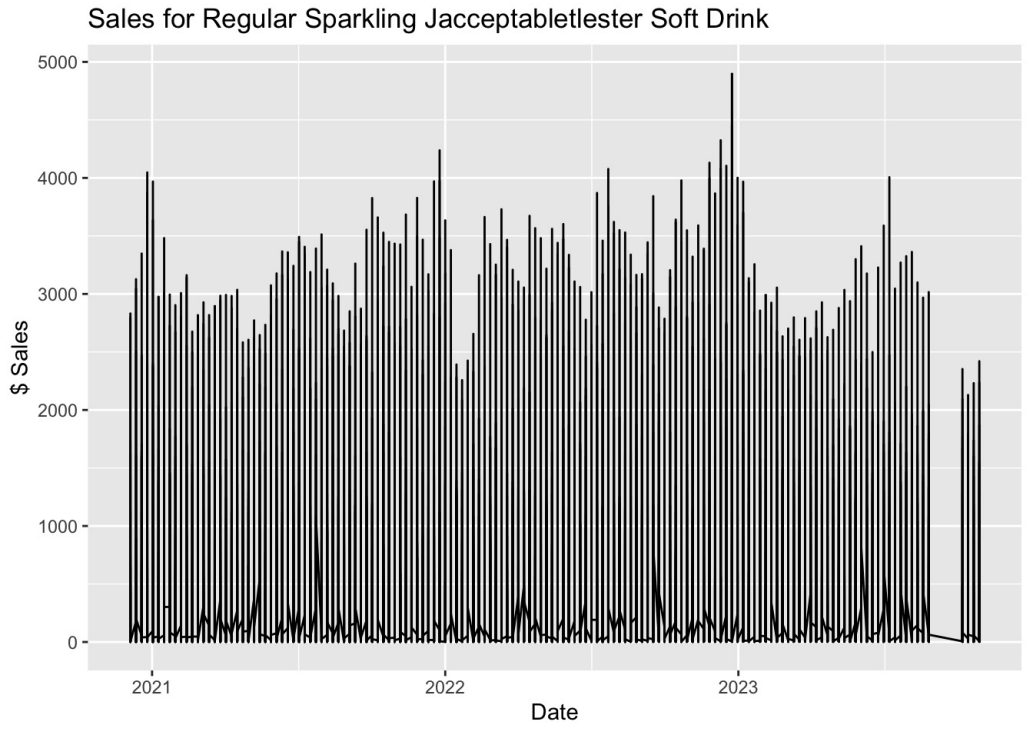



```
#graph DOLLAR_SALES by DATE for BRAND == "DIET SMASH" - INNOVATION PACKAGE == "2L MULTI JUG"
```

```
df %>%
  filter(BRAND == "DIET SMASH", PACKAGE == "2L MULTI JUG") %>%
  ggplot(aes(x = DATE, y = DOLLAR_SALES)) +
  geom_line() +
  labs(title = "Sales for Diet Smash 2L Multi Jug",
       x = "Date",
       y = "$ Sales")
```



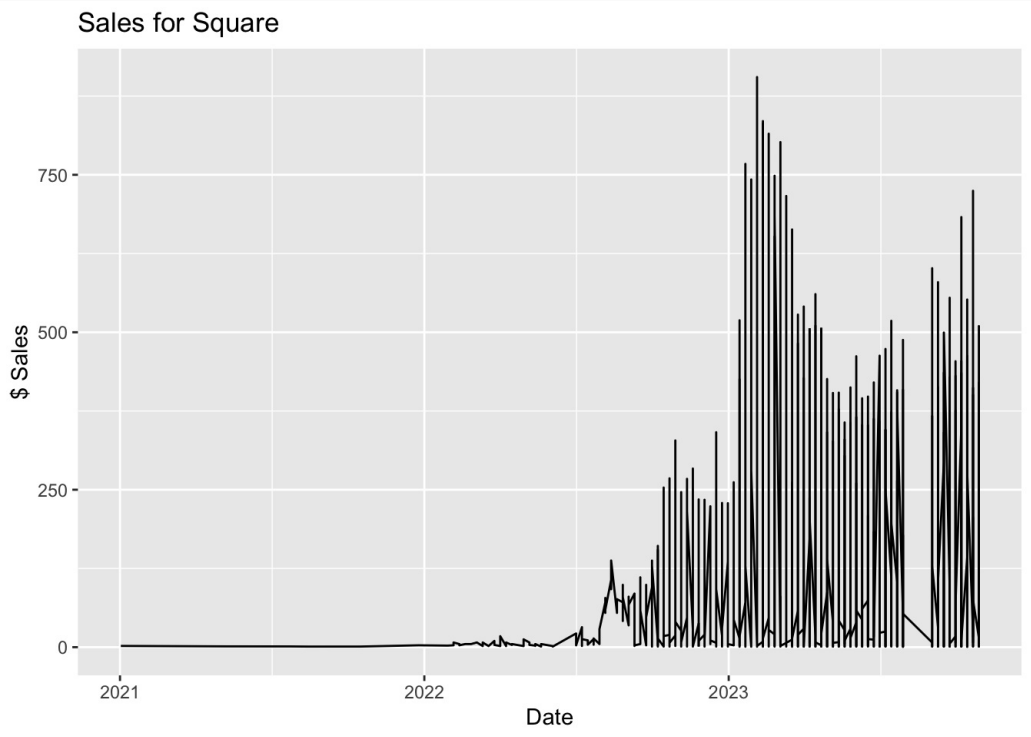

```
#graph DOLLAR_SALES by DATE for BRAND == "SPARKLING JACCEPTABLETLESTER",
#CATEGORY == "SSD", CALORIC_SEGMENT == REGULAR
df %>%
  filter(BRAND == "SPARKLING JACCEPTABLETLESTER", CATEGORY == "SSD",
         CALORIC_SEGMENT == "1" ) %>%
  ggplot(aes(x = DATE, y = DOLLAR_SALES)) +
  geom_line() +
  labs(title = "Sales for Regular Sparkling Jacceptabletlester Soft Drink",
       x = "Date",
       y = "$ Sales")
```



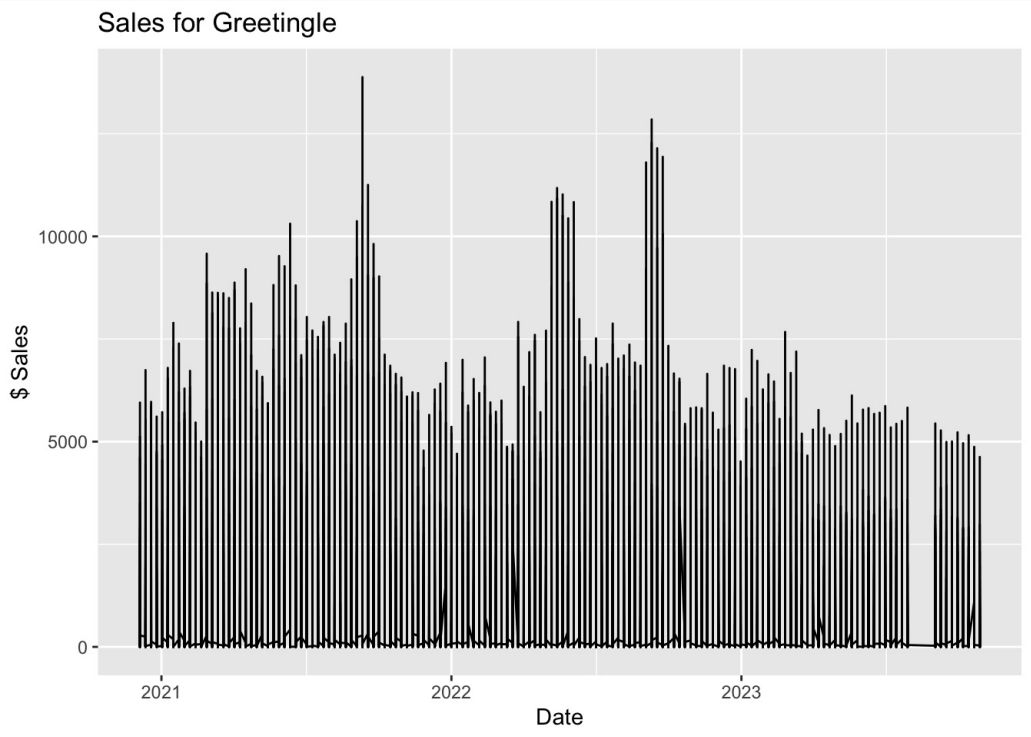

```
#graph DOLLAR_SALES by DATE for BRAND == "VENOMOUS BLAST" and CATEGORY == DIET/LIGHT
df %>%
  filter(BRAND == "VENOMOUS BLAST", CATEGORY == "DIET/LIGHT") %>%
  ggplot(aes(x = DATE, y = DOLLAR_SALES)) +
  geom_line() +
  labs(title = "Sales for Venomous Blast",
       x = "Date",
       y = "$ Sales")
```


```
#graph DOLLAR_SALES by DATE for BRAND == "SQUARE"
df %>%
  filter(BRAND == "SQUARE") %>%
  ggplot(aes(x = DATE, y = DOLLAR_SALES)) +
  geom_line() +
  labs(title = "Sales for Square",
```

```
       x = "Date",
       y = "$ Sales")
```

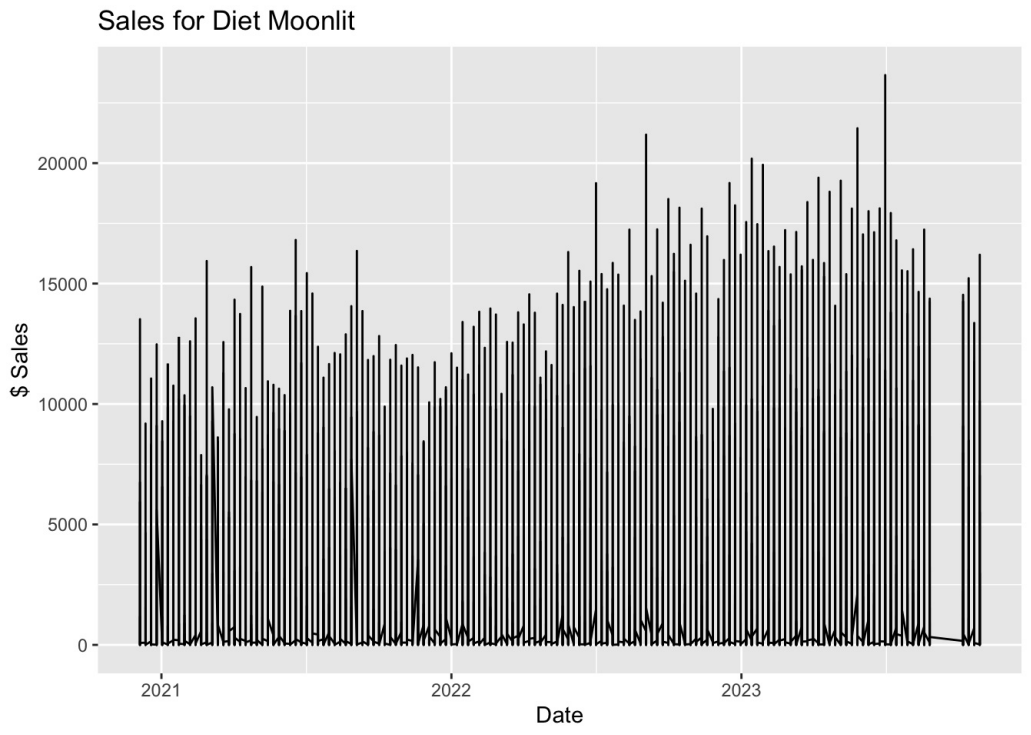



```
#graph DOLLAR_SALES by DATE for BRAND == "GREETINGLE"
df %>%
  filter(BRAND == "GREETINGLE") %>%
  ggplot(aes(x = DATE, y = DOLLAR_SALES)) +
  geom_line() +
  labs(title = "Sales for Greetingle",
       x = "Date",
       y = "$ Sales")
```

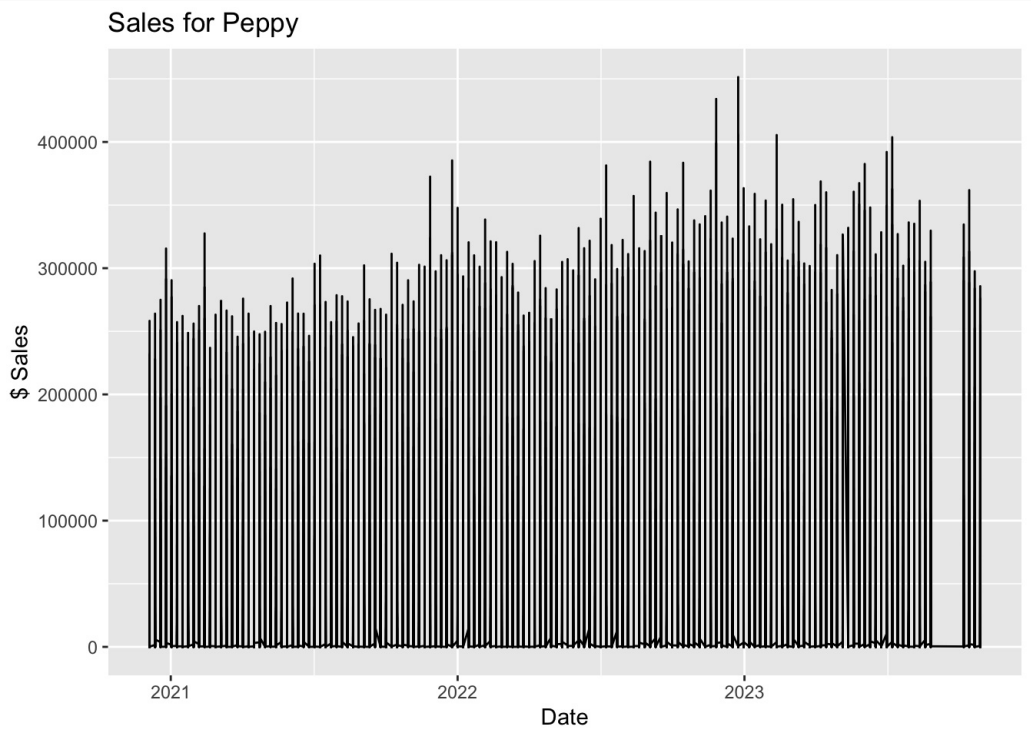



```
#graph DOLLAR_SALES by DATE for BRAND == "DIET MOONLIT"
df %>%
  filter(BRAND == "DIET MOONLIT") %>%
```

```
  ggplot(aes(x = DATE, y = DOLLAR_SALES)) +
  geom_line() +
  labs(title = "Sales for Diet Moonlit",
       x = "Date",
       y = "$ Sales")
```




```
#graph DOLLAR_SALES by DATE for BRAND == "PEPPY"
df %>%
  filter(BRAND == "PEPPY") %>%
  ggplot(aes(x = DATE, y = DOLLAR_SALES)) +
  geom_line() +
  labs(title = "Sales for Peppy",
       x = "Date",
       y = "$ Sales")
```



> Some seasonlity is observed in all drinks analyzed, with most including missing weeks and spikes. Are missing weeks, missing

> data, end of product lifecycle, or something else? The missing period seems common to several products analyzed.

```
#What are the common gaps between DATE where there is no weekly data for BRAND == "DIET SMASH"
df %>%
  filter(BRAND == "DIET SMASH") %>%
  arrange(DATE) %>%
  mutate(DIFF = DATE - lag(DATE)) %>%
  filter(DIFF > 7) %>%
  count(DIFF) %>%
  arrange(desc(n)) %>%
  head(20)
```

```
##      DIFF n
## 1 42 days 1
```

```
#What are the common gaps between DATE where there is no weekly data for BRAND == "DIET SMASH"
and PACKAGE == "2L MULTI JUG"
df %>%
  filter(BRAND == "DIET SMASH", PACKAGE == "2L MULTI JUG") %>%
  arrange(DATE) %>%
  mutate(DIFF = DATE - lag(DATE)) %>%
  filter(DIFF > 7) %>%
  count(DIFF) %>%
  arrange(desc(n)) %>%
  head(20)
```

```
##      DIFF n
## 1 35 days 1
```

```
#What are the common gaps between DATE where there is no weekly data for BRAND == "SPARKLING
JACCEPTABLETLESTER",
#CATEGORY == "SSD", CALORIC_SEGMENT == REGULAR
df %>%
  filter(BRAND == "SPARKLING JACCEPTABLETLESTER", CATEGORY == "SSD",
         CALORIC_SEGMENT == 1) %>%
  arrange(DATE) %>%
  mutate(DIFF = DATE - lag(DATE)) %>%
  filter(DIFF > 7) %>%
  count(DIFF) %>%
  arrange(desc(n)) %>%
  head(20)
```

```
##      DIFF n
## 1 42 days 1
```

```
#What are the common gaps between DATE where there is no weekly data for BRAND == "VENOMOUS
BLAST" and CALORIC_SEGMENT == DIET/LIGHT
df %>%
  filter(BRAND == "VENOMOUS BLAST", CALORIC_SEGMENT == 0 ) %>%
  arrange(DATE) %>%
  mutate(DIFF = DATE - lag(DATE)) %>%
  filter(DIFF > 7) %>%
  count(DIFF) %>%
  arrange(desc(n)) %>%
```

```
  head(20)
```

```
##      DIFF n
## 1 35 days 1
## 2 70 days 1
```

```
#What are the common gaps between DATE where there is no weekly data for BRAND == "SQUARE"
df %>%
  filter(BRAND == "SQUARE") %>%
  arrange(DATE) %>%
  mutate(DIFF = DATE - lag(DATE)) %>%
  filter(DIFF > 7) %>%
  count(DIFF) %>%
  arrange(desc(n)) %>%
  head(20)
```

```
##       DIFF n
## 1  35 days 3
## 2  14 days 2
## 3  28 days 2
## 4  21 days 1
## 5  70 days 1
## 6  77 days 1
## 7 126 days 1
```

```
#What are the common gaps between DATE where there is no weekly data for BRAND == "GREETINGLE"
df %>%
  filter(BRAND == "GREETINGLE") %>%
  arrange(DATE) %>%
  mutate(DIFF = DATE - lag(DATE)) %>%
  filter(DIFF > 7) %>%
  count(DIFF) %>%
  arrange(desc(n)) %>%
  head(20)
```

```
##      DIFF n
## 1 35 days 1
```

```
#What are the common gaps between DATE where there is no weekly data for BRAND == "DIET
MOONLIT"
df %>%
  filter(BRAND == "DIET MOONLIT") %>%
  arrange(DATE) %>%
  mutate(DIFF = DATE - lag(DATE)) %>%
  filter(DIFF > 7) %>%
  count(DIFF) %>%
  arrange(desc(n)) %>%
  head(20)
```

```
##      DIFF n
## 1 42 days 1
```

```
#What are the common gaps between DATE where there is no weekly data for BRAND == "PEPPY"
```

```
df %>%
  filter(BRAND == "PEPPY") %>%
  arrange(DATE) %>%
  mutate(DIFF = DATE - lag(DATE)) %>%
  filter(DIFF > 7) %>%
  count(DIFF) %>%
  arrange(desc(n)) %>%
  head(20)
```

```
##      DIFF n
## 1 42 days 1
```

> The “Diet Smash” brand has a gap in dates of 42 days. If packaging such as “2L Multi Jug” is added in, Diet Smash has a gap of 35 days. The “regular Sparkling Jacceptabletler brand in the ssd category” has a gap of 42 days. The “Venomous Blast” brand has 2 gaps, one of 35 days and the other of 70 days. The “Square” brand has 7 gap lengths in dates, with the most common of 35 days occuring 3 times, and the longest being 126 days. The “Greetingle” brand has a single gap length of 35 days. The “Diet Moonlit” has a single gap of 42 days. The “Peppy” brand has a single gap of 42 days. Further analysis should be done to ensure start/stop dates for tenure are accurate and that we are not missing weekly data.

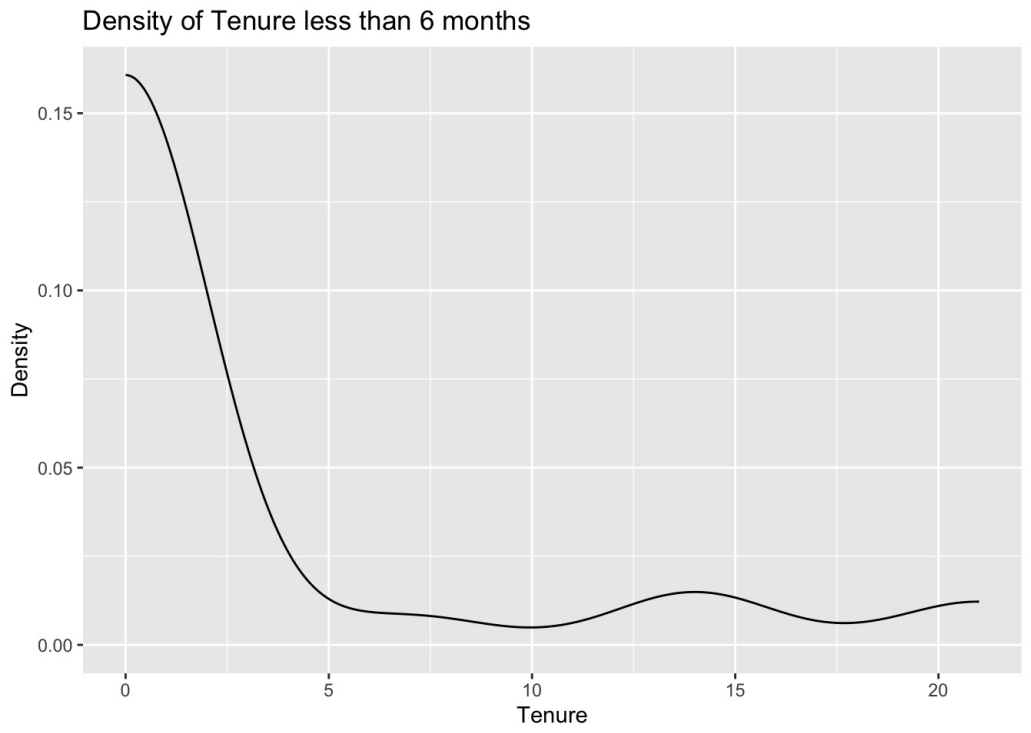
```
#most common launch DATE, end DATE, and TENURE of brand or package less than 6 months, or 26
weeks, in duration
df %>%
  group_by(BRAND, PACKAGE) %>%
  summarise(MIN_DATE = min(DATE),
            MAX_DATE = max(DATE),
            TENURE = MAX_DATE - MIN_DATE) %>%
  filter(TENURE < 26) %>%
  arrange(desc(TENURE))
```

```
## # A tibble: 145 × 5
## # Groups:   BRAND [102]
##    BRAND                                  PACKAGE MIN_DATE   MAX_DATE   TENURE
##    <fct>                                  <fct>   <date>     <date>     <drtn>
##  1 BARS                                   2L MUL… 2022-10-29 2022-11-19 21 da…
##  2 BUBBLE JOY ADVANTAGEOUS MOON           7.5SMA… 2022-02-26 2022-03-19 21 da…
##  3 JUICY SQUIRREL                         3L MUL… 2023-10-07 2023-10-28 21 da…
##  4 KOOL! READY-TO-GO                      20SMAL… 2023-10-07 2023-10-28 21 da…
##  5 KOOL! READY-TO-GO                      7.5SMA… 2023-10-07 2023-10-28 21 da…
##  6 KOOL! ZERO SUGAR READY-TO-GO           7.5SMA… 2023-10-07 2023-10-28 21 da…
##  7 ORANGE VANILLA BUBBLE JOY ADVANTAGEOUS … 12SMAL… 2021-06-12 2021-07-03 21 da…
##  8 PAPI ZERO SUGAR                        24SMAL… 2020-12-19 2021-01-09 21 da…
##  9 YAWN TROP                              20SMAL… 2023-10-07 2023-10-28 21 da…
## 10 DIET BUBBLE JOY ADVANTAGEOUS PLANT-BASE… 7.5SMA… 2023-08-05 2023-08-19 14 da…
## # ℹ 135 more rows
```

```
#summarize top 10  brand or package less than 6 months, or 26 weeks, in duration for innovation
set
df %>%
  group_by(BRAND, PACKAGE) %>%
  summarise(MIN_DATE = min(DATE),
            MAX_DATE = max(DATE),
            TENURE = MAX_DATE - MIN_DATE) %>%
  filter(TENURE < 26) %>%
  arrange(desc(TENURE)) %>%
  head(10)
```

```
## # A tibble: 10 × 5
## # Groups:   BRAND [9]
##    BRAND                                    PACKAGE MIN_DATE   MAX_DATE   TENURE
##    <fct>                                    <fct>   <date>     <date>     <drtn>
##  1 BARS                                     2L MUL… 2022-10-29 2022-11-19 21 da…
##  2 BUBBLE JOY ADVANTAGEOUS MOON             7.5SMA… 2022-02-26 2022-03-19 21 da…
##  3 JUICY SQUIRREL                           3L MUL… 2023-10-07 2023-10-28 21 da…
##  4 KOOL! READY-TO-GO                        20SMAL… 2023-10-07 2023-10-28 21 da…
##  5 KOOL! READY-TO-GO                        7.5SMA… 2023-10-07 2023-10-28 21 da…
##  6 KOOL! ZERO SUGAR READY-TO-GO             7.5SMA… 2023-10-07 2023-10-28 21 da…
##  7 ORANGE VANILLA BUBBLE JOY ADVANTAGEOUS … 12SMAL… 2021-06-12 2021-07-03 21 da…
##  8 PAPI ZERO SUGAR                          24SMAL… 2020-12-19 2021-01-09 21 da…
##  9 YAWN TROP                                20SMAL… 2023-10-07 2023-10-28 21 da…
## 10 DIET BUBBLE JOY ADVANTAGEOUS PLANT-BASE… 7.5SMA… 2023-08-05 2023-08-19 14 da…
```

```
#density plot of TENURE less than 6 months, or 26 weeks.
df %>%
  group_by(BRAND, PACKAGE) %>%
  summarise(MIN_DATE = min(DATE),
            MAX_DATE = max(DATE),
            TENURE = MAX_DATE - MIN_DATE) %>%
  filter(TENURE < 26) %>%
  ggplot(aes(x = TENURE)) +
  geom_density() +
  labs(title = "Density of Tenure less than 6 months",
       x = "Tenure",
       y = "Density")
```



> 145 Brand/Package sets under 6 months, with start and stop date to set as week 0. Will need to guarantee that each one is not missing week/date data for time series analysis. 9 of the top 10 brand/package sets from this set run 21 days, with the 10th running 14 days. October 7th is an extremely popular launch date. The most common tenure of the 145 brand/package combos with tenure less than 6 months run for less than 5 weeks in duration.

Summary and Results

> The analysis of Swire Coca-Cola's sales data reveals several key insights. Initially, there were 59,725 missing values in the dataset, which were subsequently imputed using text data analysis, resulting in zero missing values. The dataset comprises 152 unique weeks, indicating weekly data availability for nearly three years without any missing weeks. Notably, there are 2 caloric segments, 5 categories, 8 manufacturers, 319 brands, 103 package types, and 3,692 unique item descriptions.

> Examining specific products, the analysis distinguishes characteristics such as the Diet Smash product being in the energy category with three packaging types. Similarly, the Sparkling Jacceptabletlester brand offers both diet and regular options across various package types, while the Venomous Blast product is available in both diet and regular variants, primarily in the "16small multi cup" packaging. Moreover, Swire ranks third in overall sales among manufacturers, with Bubble Joy Advantageous leading in sales by brand, followed by Real-Time and Peppy.

> The data exhibits left skewness regarding brand run times, with a median of 137 weeks and a mean of 99.1 weeks, indicating varying brand durations. Conversely, package tenure demonstrates right skewness, with a median of 147 weeks and a mean of 117 weeks, suggesting longer tenure for tried and true packaging sizes. Notably, 14 package sizes run for less than six months, indicating potential innovation packages.

> Seasonal analysis reveals missing weeks and spikes across all drinks analyzed, raising questions about the cause, whether due to missing data or product lifecycle endings. Further investigation is warranted to ensure accurate start and stop dates for product tenures, particularly for brands with gaps in dates.

> In summary, the analysis provides valuable insights into Swire Coca-Cola's sales data, highlighting key trends, product characteristics, and areas for further investigation to optimize forecasting and product launch strategies.

Modeling EDA

> This section of the EDA will focus on potential ways of breaking up the data in to modeling sets. It will check for any potential errors in the data that could cause incorrect prediction, look at sections of the data that could be removed as they will cause noise in the model and the last parts will zero in on what factors of our new inovation products are in our current data set.

Category Check by Item

```
### Create table counting each item and how many brand

categories_count <- df %>%
  group_by(ITEM) %>%
  summarize(
    num_manufacturers = n_distinct(MANUFACTURER),
    num_category = n_distinct(CATEGORY),
    num_market_key = n_distinct(MARKET_KEY),
    num_caloric_segment = n_distinct(CALORIC_SEGMENT),
    num_brand = n_distinct(BRAND),
    num_package = n_distinct(PACKAGE)
  )
summary(categories_count)
```

```
##      ITEM           num_manufacturers  num_category num_market_key
##  Length:3692        Min.   :1         Min.   :1     Min.   :  1.00
##  Class :character   1st Qu.:1         1st Qu.:1     1st Qu.:  3.00
##  Mode  :character   Median :1         Median :1     Median : 40.00
##                     Mean   :1         Mean   :1     Mean   : 82.14
##                     3rd Qu.:1         3rd Qu.:1     3rd Qu.:189.00
##                     Max.   :1         Max.   :1     Max.   :200.00
##  num_caloric_segment   num_brand      num_package
##  Min.   :1.000       Min.   :1.000   Min.   :1.00
##  1st Qu.:1.000       1st Qu.:1.000   1st Qu.:1.00
##  Median :1.000       Median :1.000   Median :1.00
```

```
##  Mean   :1.013       Mean   :1.026   Mean   :1.01
##  3rd Qu.:1.000       3rd Qu.:1.000   3rd Qu.:1.00
##  Max.   :2.000       Max.   :3.000   Max.   :3.00
```

```
categories_count %>%
  summarise(
    count_more_than_one_caloric_segment = sum(num_caloric_segment > 1),
    count_more_than_one_brand = sum(num_brand > 1),
    count_more_than_one_package = sum(num_package > 1)
  )
```

```
## # A tibble: 1 × 3
##   count_more_than_one_caloric_se…¹ count_more_than_one_…² count_more_than_one_…³
##                              <int>                  <int>                  <int>
## 1                               49                     88                     34
## # ℹ abbreviated names: ¹count_more_than_one_caloric_segment,
## #   ²count_more_than_one_brand, ³count_more_than_one_package
```

```
#88 items fall into 2 or more brands
#49 items fall into 2 or more categories
#34 Items with 2 or more packages

 df %>%
  inner_join(categories_count %>%
               group_by(ITEM) %>%
               filter(sum(num_brand) > 1) %>%
               select(ITEM),
             by = "ITEM") %>%
   select(ITEM, BRAND) %>%
   arrange(ITEM) %>%
   distinct(ITEM, BRAND) %>%
   head(5)
```

```
##                                                                    ITEM
## 1  AZURE HORIZON GENTLE DRINK SUPER-JUICE  DURIAN  CUP 12 LIQUID SMALL
## 2  AZURE HORIZON GENTLE DRINK SUPER-JUICE  DURIAN  CUP 12 LIQUID SMALL
## 3 CUPADA ARID GENTLE DRINK GINGER ALE AND ADE  CUP 12 LIQUID SMALL X12
## 4 CUPADA ARID GENTLE DRINK GINGER ALE AND ADE  CUP 12 LIQUID SMALL X12
## 5              CUPADA ARID GENTLE DRINK GINGER ALE CUP 12 LIQUID SMALL
##                BRAND
## 1      AZURE HORIZON
## 2 DIET AZURE HORIZON
## 3        CUPADA ARID
## 4   DIET CUPADA ARID
## 5        CUPADA ARID
```

```
# This DF shows that many of the items with 2 types of brand are both in Diet and Regular
category. If we are using these items as filters in our model building we will need to remember
this.

df %>%
  inner_join(categories_count %>%
               group_by(ITEM) %>%
               filter(sum(num_caloric_segment) > 1) %>%
               select(ITEM),
```

```
            by = "ITEM") %>%
   select(ITEM, CALORIC_SEGMENT) %>%
   arrange(ITEM) %>%
   distinct(ITEM, CALORIC_SEGMENT) %>%
   head(5)
```

```
##                                                                                ITEM
## 1 AAE LIQUORICE REVITALIZING BOOST LIQUID FREEZE BURN FAT LS ENRGY TCHNL JUG 8 LIQUID SMALL
## 2 AAE LIQUORICE REVITALIZING BOOST LIQUID FREEZE BURN FAT LS ENRGY TCHNL JUG 8 LIQUID SMALL
## 3                      AZURE HORIZON GENTLE DRINK SUPER-JUICE  DURIAN  CUP 12 LIQUID SMALL
## 4                      AZURE HORIZON GENTLE DRINK SUPER-JUICE  DURIAN  CUP 12 LIQUID SMALL
## 5                CARBONATE STREAM ENERGY DRINK CONCENTRATE UNFLAVORED JUG 14.8 LIQUID SMALL
##   CALORIC_SEGMENT
## 1               1
## 2               0
## 3               1
## 4               0
## 5               1
```

```
# As with the brand caloric segment has the same duplicate items in both segments. We will need
to remember this in modeling

df %>%
  inner_join(categories_count %>%
               group_by(ITEM) %>%
               filter(sum(num_package) > 1) %>%
               select(ITEM),
             by = "ITEM") %>%
   select(ITEM, PACKAGE) %>%
   arrange(ITEM) %>%
   distinct(ITEM, PACKAGE) %>%
   head(5)
```

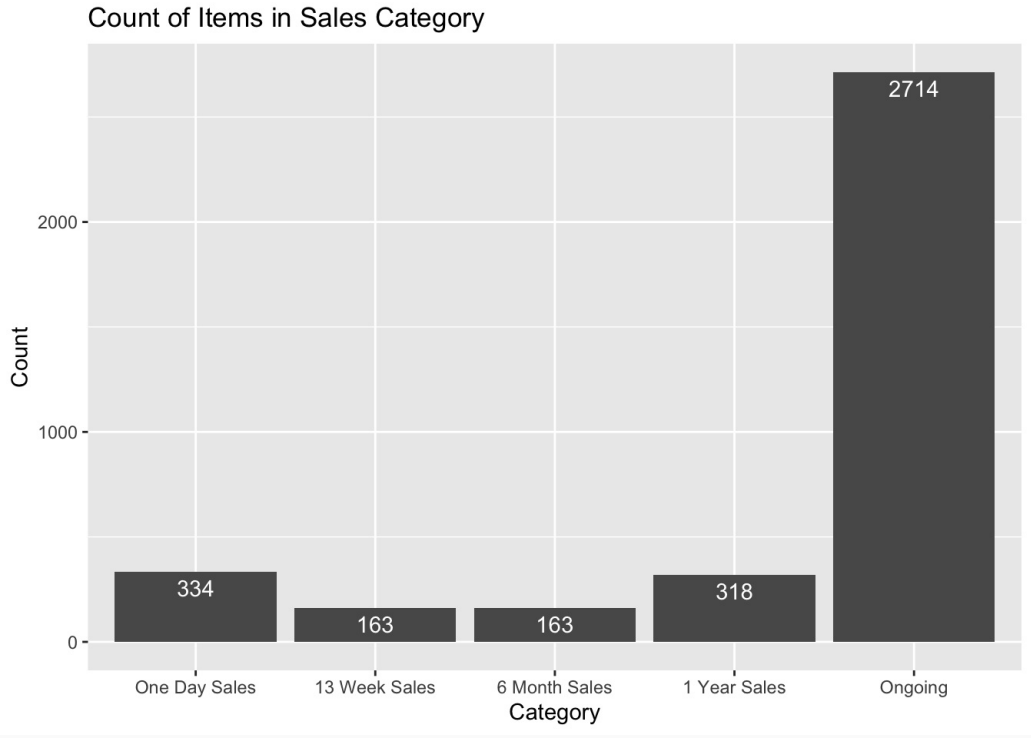
```
##                                                        ITEM
## 1 BUBBLE JOY WATER-JUGD-CARBONATED CONTAINER 288 LIQUID SMALL
## 2 BUBBLE JOY WATER-JUGD-CARBONATED CONTAINER 288 LIQUID SMALL
## 3  CUPADA ARID GENTLE DRINK GINGER ALE JUG 10 LIQUID SMALL X6
## 4  CUPADA ARID GENTLE DRINK GINGER ALE JUG 10 LIQUID SMALL X6
## 5   CUPADA ARID TONIC WATER UNFLAVORED JUG 10 LIQUID SMALL X6
##                     PACKAGE
## 1         12SMALL 24ONE CUP
## 2            ALL OTHER ONES
## 3            ALL OTHER ONES
## 4 10SMALL 6ONE PLASTICS JUG
## 5            ALL OTHER ONES
```

```
#Even though these items have a package in the item description we their packaging category
changes
```

> In this section we see that there are products that fall into multiple Caloric Segments, Brands and Packages. We will need to keep these items in mind when building models. When creating our samller data sets for modeling we will want to assure that we filter by these 3 categories.

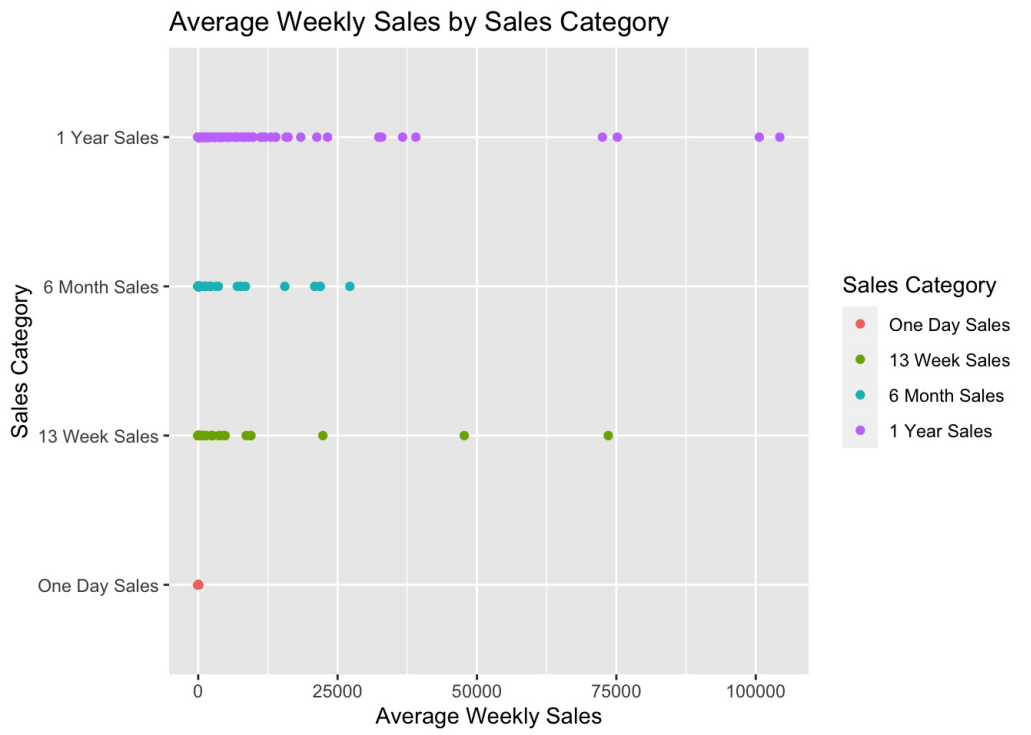
Breaking out Items by Tenure

```
#Create table to summarize total sale days

sales_summary <- df %>%
  group_by(ITEM) %>%
  summarize(first_date = min(DATE), last_date = max(DATE), total_sales = sum(UNIT_SALES),
total_revenue = sum(DOLLAR_SALES), total_sale_days = n_distinct(DATE))

#Calculate Total window of days sold

sales_summary <- sales_summary %>%
  mutate(
        duration_days = last_date - first_date,
        duration_weeks = ceiling(as.numeric(duration_days) / 7),
        launch13week_date =first_date + lubridate::weeks(13),
        launch6month_date = first_date + months(6),
        launch1year_date = first_date + lubridate::years(1),
        avg_sales_per_week = ifelse(duration_weeks == 0, total_sales, total_sales /
duration_weeks),
)

#Batch data in to categories 1 day sales, upto 13 week sales, 13 week to 6 months, 6 months to
a year, more than a year = ongoing

sales_summary$sales_category <- cut(sales_summary$duration_weeks, breaks = c(-Inf, 0, 13, 26,
52, Inf),
                        labels = c("One Day Sales", "13 Week Sales", "6 Month Sales", "1 Year
Sales", "Ongoing"))

# Create plot couting unique itmes in each sales category

ggplot(sales_summary, aes(x = sales_category)) +
  geom_bar() +
  geom_text(aes(label = after_stat(count)), stat = "count", vjust = 1.5, colour = "white")+
  labs(title = "Count of Items in Sales Category",
       x = "Category",
       y = "Count")
```



> In the section above we added sales summeries to our items. We see that when breaking up our itmes into sales groups we have a majority of ites that are ongoing. These will most likely not be helpful when running our 7 questions based around limited sales.

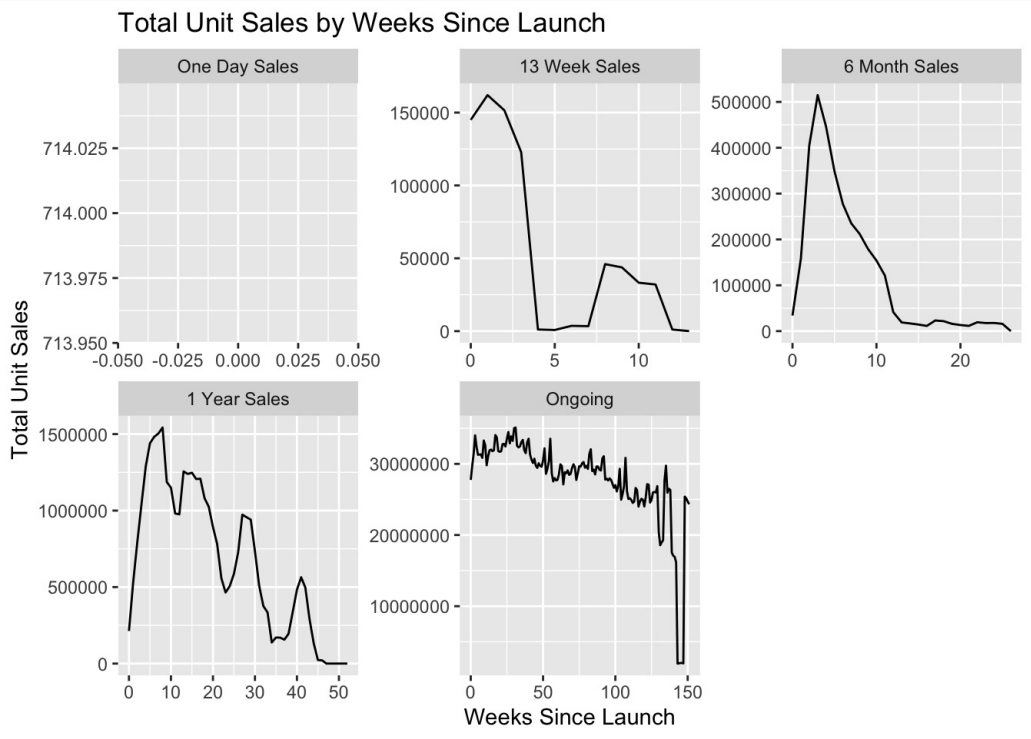
```
#Create plot of sales category with average weekly sales

sales_summary %>%
filter(sales_category != "Ongoing") %>%
ggplot( aes(x = avg_sales_per_week, y = sales_category, color = sales_category)) +
  geom_point() +
  labs(title = "Average Weekly Sales by Sales Category",
       x = "Average Weekly Sales",
       y = "Sales Category",
       color = "Sales Category")
```



> This plot shows us that in our short sale items we are highly weighted to the left with a few large outliars when creating our modeling sets we will want to do more outliar analysis on each group to see how to address each one.

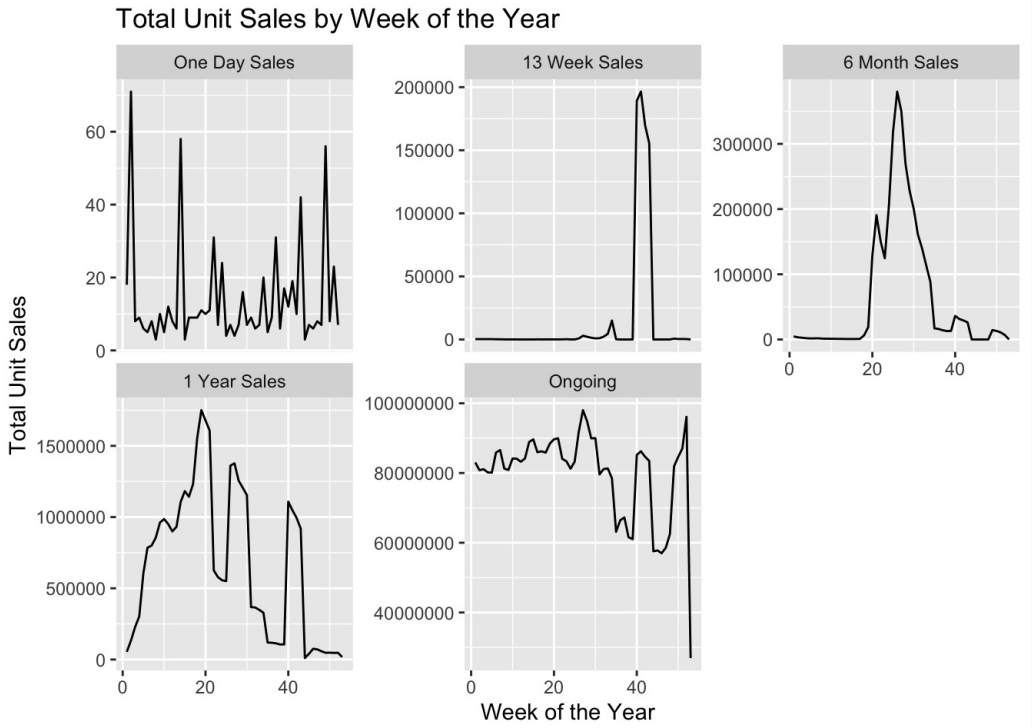
```
df <- left_join(df,sales_summary %>% select(ITEM, sales_category, duration_days,
duration_weeks, total_sale_days, first_date, launch13week_date, launch6month_date,
launch1year_date), by = "ITEM")

# Calculate days since launch

df <- df %>%
  mutate(days_since_launch = as.numeric(DATE - first_date),
         weeks_since_launch = ceiling(as.numeric((DATE - first_date)/7)))

# Group by sales category and create separate line graphs for each sales category

df %>%
  group_by(sales_category, weeks_since_launch) %>%
  summarize(total_unit_sales = sum(UNIT_SALES)) %>%
  group_by(sales_category) %>%
  filter(total_unit_sales > 0) %>%
ggplot( aes(x = weeks_since_launch, y = total_unit_sales)) +
  geom_line() +
```

```
  labs(title = "Total Unit Sales by Weeks Since Launch",
       x = "Weeks Since Launch",
       y = "Total Unit Sales") +
  facet_wrap(~ sales_category, scales = "free")
```



> In these 5 line graphs we are given the shape of sales from launch date. We see that with the short temr porducts we see a large spike at the start that tapers off as time goes on. With the ongoing products they start strong and then slowly start to fall over time. This demenstrates further that when modeling breaking our items into categories to model will help us be more accurate when looking at forecasting newe products for shorter amounts of time. It also reinforces that we should be able to exclude our ongoing products from the data.

```
# Group by sales category and create separate line graphs
df %>%
  mutate(week_of_year = week(DATE)) %>%
  group_by(sales_category, week_of_year) %>%
  summarize(total_unit_sales = sum(UNIT_SALES)) %>%
  group_by(sales_category) %>%
  filter(total_unit_sales > 0) %>%
ggplot(aes(x = week_of_year, y = total_unit_sales)) +
  geom_line() +
  labs(title = "Total Unit Sales by Week of the Year",
       x = "Week of the Year",
       y = "Total Unit Sales") +
  facet_wrap(~ sales_category, scales = "free_y")
```



> These graphs show us when forecasting for specific weeks of the year we will have large amounts of variance based on past sales. We may need to look at outliears in the 13 week and 1 year category.

Question 1 Parameters

Item Description: Diet Smash Plum 11Small 4One Caloric Segment: Diet Market Category: SSD Manufacturer: Swire-CC Brand: Diet Smash Package Type: 11Small 4One Flavor: 'Plum' Which 13 weeks of the year would this product perform best in the market? What is the forecasted demand, in weeks, for those 13 weeks?

```
library(stringr)
# Matching parameters for Q1
df %>%
  filter(sales_category == "13 Week Sales",
         CALORIC_SEGMENT == 0,
         CATEGORY == "SSD") %>%
         #str_detect(ITEM, "PLUM")) %>%
  group_by(ITEM) %>%
  summarize(distinct_items = n_distinct(ITEM))
```

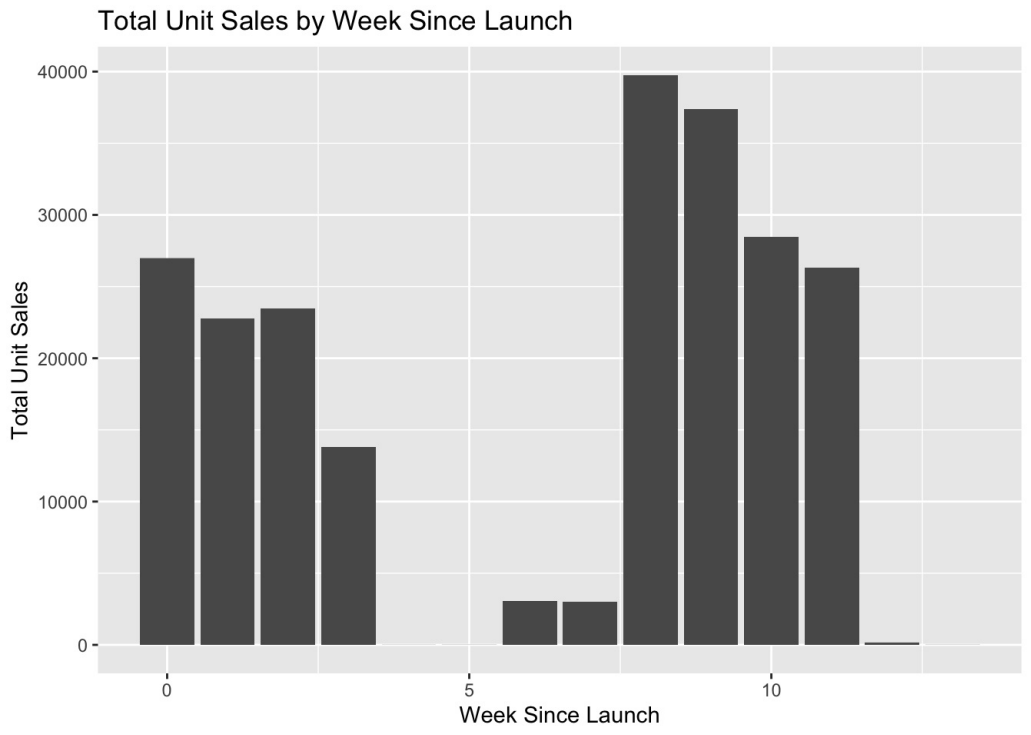
```
## # A tibble: 37 × 2
##    ITEM                                                      distinct_items
##    <chr>                                                              <int>
##  1 AZURE HORIZON FREE GENTLE DRINK SUPER-JUICE  DURIAN  CALORIE …           1
##  2 CAFFEINE FREE DIET PAPI GENTLE DRINK COLA DIET JUG 16 LIQUID …           1
##  3 CAFFEINE FREE DIET RAINING GENTLE DRINK AVOCADO  DIET CUP 12 …           1
##  4 CUPSHIELD'S GENTLE DRINK POWDER  FUDYNAMOE DIET CUP 12 LIQUID…           1
##  5 CUPSHIELD'S TONIC WATER UNFLAVORED DIET JUG 33.8 LIQUID SMALL             1
##  6 DIET BUBBLE JOY ADVANTAGEOUS GENTLE DRINK COLA DIET JUG 12 LI…           1
##  7 DIET BUBBLE JOY ADVANTAGEOUS GENTLE DRINK COLA PINK  GUAVA DI…           1
##  8 DIET BUBBLE JOY ADVANTAGEOUS GENTLE DRINK COLA PINK  GUAVA DI…           1
##  9 DIET BUBBLE JOY ADVANTAGEOUS GENTLE DRINK DURIAN  COLA DIET C…           1
## 10 DIET GORGEOUS SUNSET OUS GENTLE DRINK AVOCADO  DIET CUP 12 LI…           1
## # ℹ 27 more rows
```

```
# There are 37 items that match time period caloric segment, category. flavor and packaging
```

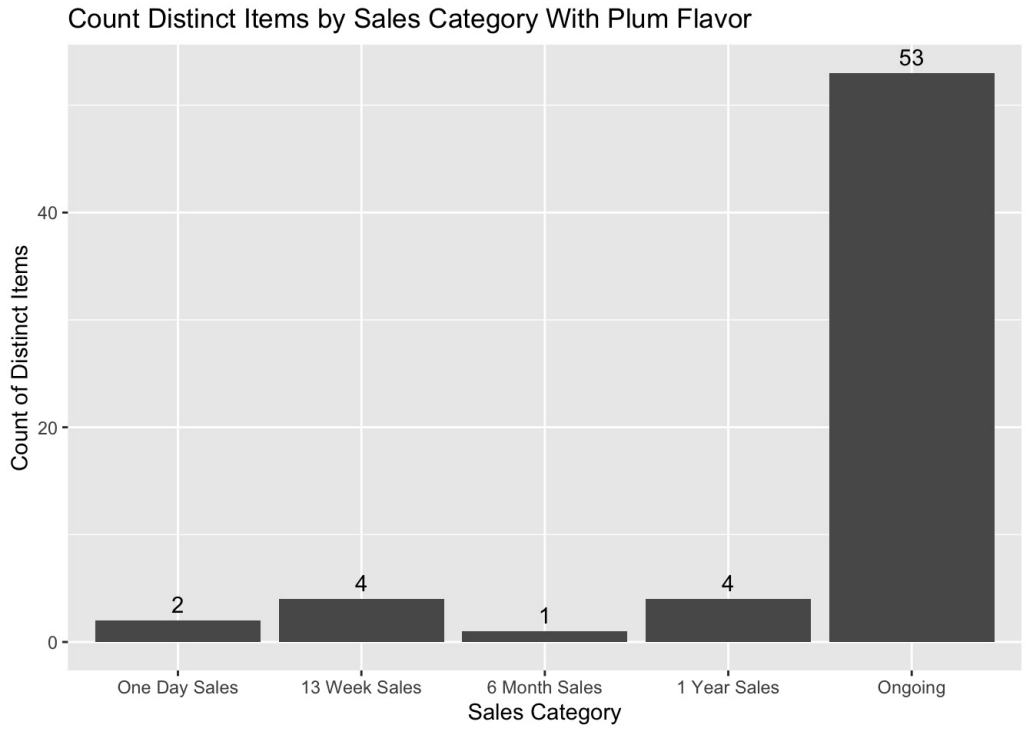
```
dont exist

df %>%
  filter(str_detect(ITEM, "PLUM")) %>%
  summarize(distinct_items = n_distinct(ITEM))
```

```
##   distinct_items
## 1             64
```

```
#64 Plum Flavored items

#Distribution of matching items
df %>%
  filter(sales_category == "13 Week Sales",
         CALORIC_SEGMENT == 0,
         CATEGORY == "SSD") %>%
  group_by(ITEM, weeks_since_launch) %>%
  summarize(total_unit_sales = sum(UNIT_SALES)) %>%
  ggplot(aes(x = weeks_since_launch, y = total_unit_sales)) +
  geom_bar(stat = "identity") +
  labs(title = "Total Unit Sales by Week Since Launch",
       x = "Week Since Launch",
       y = "Total Unit Sales")
```




```
# Create Sales Category distribution of plum items
df %>%
  filter(str_detect(ITEM, "PLUM")) %>%
  group_by(sales_category) %>%
  summarize(distinct_items = n_distinct(ITEM)) %>%
  ggplot(aes(x = sales_category, y = distinct_items)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = distinct_items), vjust = -.5, color = "black") +
  labs(title = "Count Distinct Items by Sales Category With Plum Flavor",
       x = "Sales Category",
       y = "Count of Distinct Items")
```



> In this we found that there is a potential 37 itmes that have matching paramaters minus the flavor and packageing size. There are no itmes with the package size and 64 total plum flavored items.

Question 2 Parameters

Item Description: Sparkling Jacceptabletlester Avocado 11Small MLT Caloric Segment: Regular Market Category: SSD Manufacturer: Swire-CC Brand: Sparkling Jacceptabletlester SPARKLING JACCEPTABLETLESTER Package Type: 11Small MLT Flavor: 'Avocado' Swire plans to release this product 2 weeks prior to Easter and 2 weeks post Easter. What will the forecasted demand be, in weeks, for this product?

```
# Matching parameters for Q2
df %>%
  filter(
         month(first_date) %in% c(3, 4),
         CALORIC_SEGMENT == 1,
         CATEGORY == "SSD") %>%
         #str_detect(ITEM, "AVOCADO") %>%
  summarize(distinct_items = n_distinct(ITEM))
```

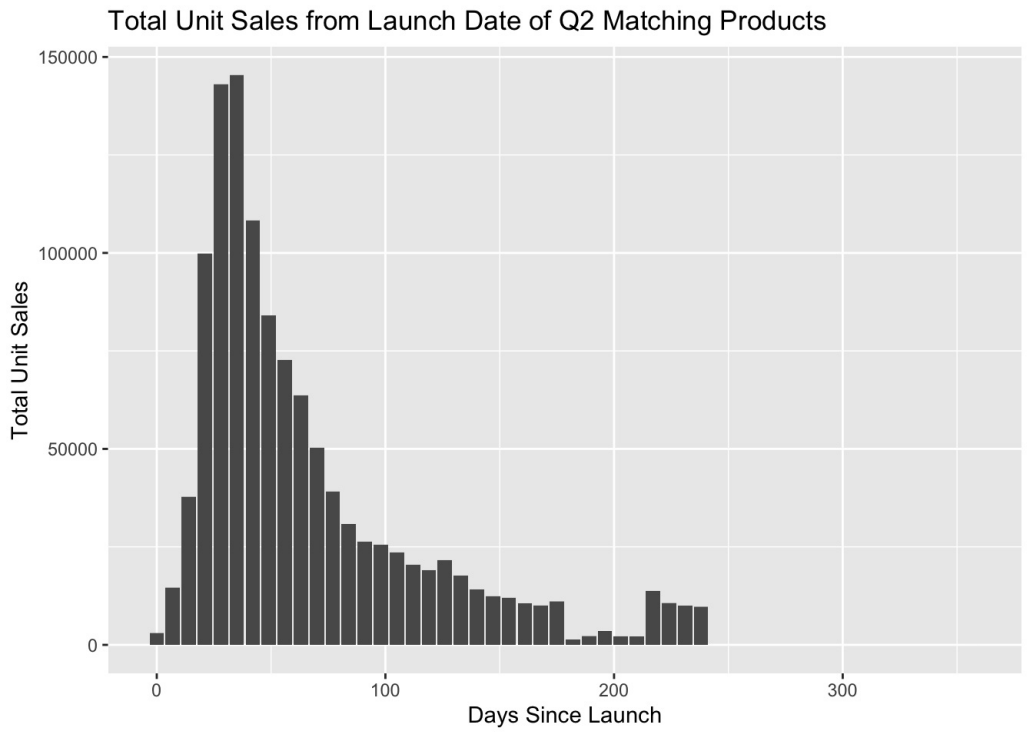
```
##   distinct_items
## 1            148
```

```
#148 items match launching in either March or April, regular and SSD category. There are non
with Avocado in this group

df %>%
  filter(str_detect(ITEM, "AVOCADO")) %>%
  summarize(distinct_items = n_distinct(ITEM))
```

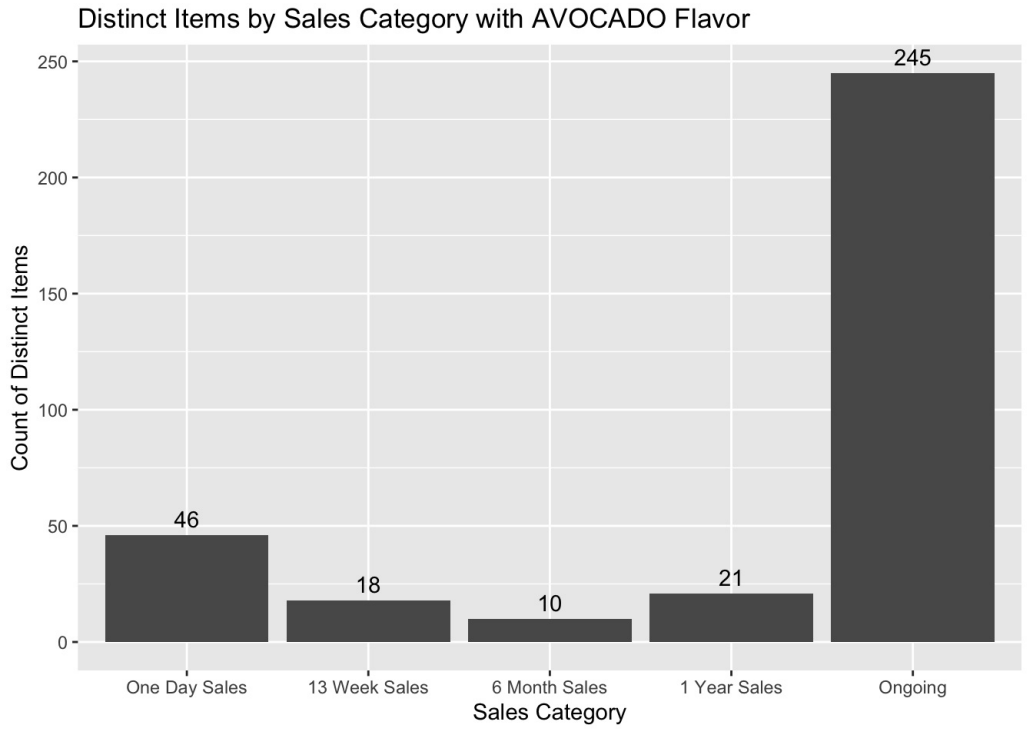
```
##   distinct_items
## 1            340
```

```
#340 AVOCADO Flavored items
```

```
#Distribution of matching items

df %>%
  filter(
         month(first_date) %in% c(3, 4),
         CALORIC_SEGMENT == 1,
         CATEGORY == "SSD",
         sales_category != 'Ongoing') %>%
  group_by(ITEM, days_since_launch) %>%
  summarize(total_unit_sales = sum(UNIT_SALES)) %>%
  ggplot(aes(x = days_since_launch, y = total_unit_sales)) +
  geom_bar(stat = "identity") +
  labs(title = "Total Unit Sales from Launch Date of Q2 Matching Products",
       x = "Days Since Launch",
       y = "Total Unit Sales")
```




```
# Create Sales Category distribution of AVACADO items

df %>%
  filter(str_detect(ITEM, "AVOCADO")) %>%
  group_by(sales_category) %>%
  summarize(distinct_items = n_distinct(ITEM)) %>%
  ggplot(aes(x = sales_category, y = distinct_items)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = distinct_items), vjust = -.5, color = "black") +
  labs(title = "Distinct Items by Sales Category with AVOCADO Flavor",
       x = "Sales Category",
       y = "Count of Distinct Items")
```



> From this section we are able to find a group of 148 items that launched either in March or April that have some of the matching features of product 2. We also have a much larger group or products that have avocado with 340 itmes.

Question 3 Parameters

Item Description: Diet Venomous Blast Energy Drink Kiwano 16 Liquid Small Caloric Segment: Diet Market Category: Energy Manufacturer: Swire-CC Brand: Venomous Blast Package Type: 16 Liquid Small Flavor: 'Kiwano' Which 13 weeks of the year would this product perform best in the market? What is the forecasted demand, in weeks, for those 13 weeks?

```
# Matching parameters for Q3

df %>%
  filter(
    sales_category == "13 Week Sales",
    CALORIC_SEGMENT == 0,
    CATEGORY == "ENERGY",
    #BRAND == "VENOMOUS BLAST",
    #str_detect(ITEM, "KIWANO")
  ) %>%
  group_by(ITEM) %>%
  summarize(distinct_items = n_distinct(ITEM))
```

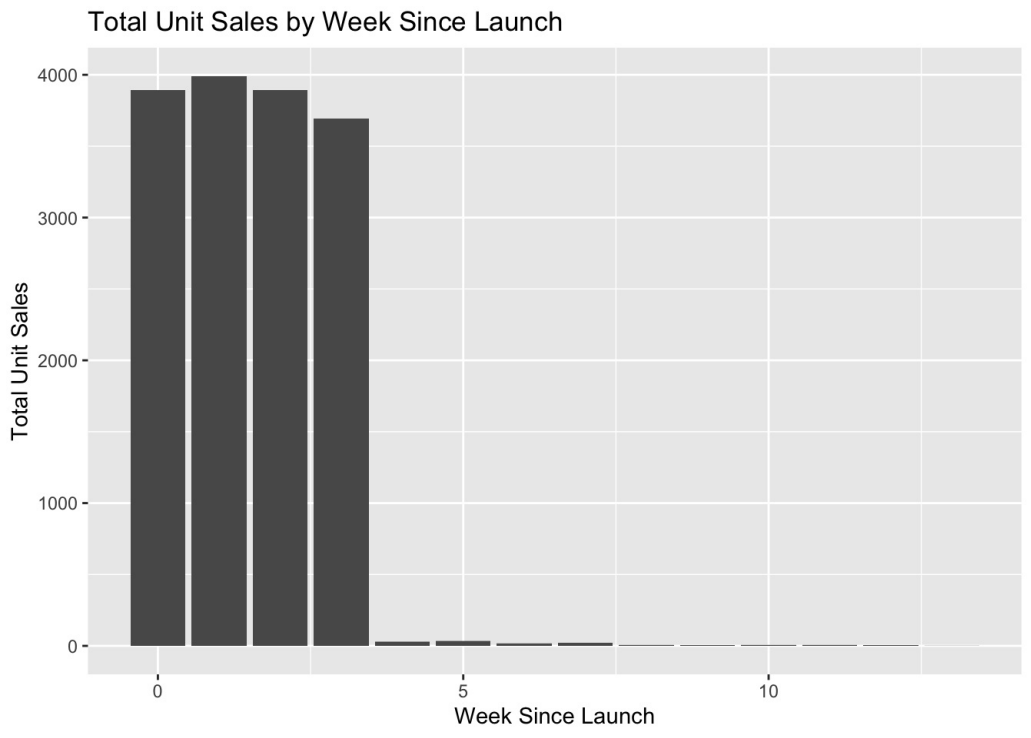
```
## # A tibble: 16 × 2
##    ITEM                                                      distinct_items
##    <chr>                                                              <int>
##  1 AAE LIQUORICE REVITALIZING BOOST LIQUID FREEZE AND BURN JUG 8…         1
##  2 AUTHENTIC SIP SPARKLES WATER KIWANO  MIST ZERO CALORIES CUP 1…         1
##  3 AUTHENTIC SIP SPARKLES WATER MANDARIN YUZU ZERO CALORIES CUP …         1
##  4 AUTHENTIC SIP SPARKLES WATER WHITE POPPIN  KEEN  ZERO CALORIE…         1
##  5 KEKE ENERGY ENERGY REVITALIZING BOOST LIQUID JUG 1.93 LIQUID …         1
##  6 KEKE ENERGY ENERGY REVITALIZING BOOST LIQUID NUTRIENTS JUG 1.…         1
##  7 MYTHICAL BEVERAGE LO-CARB ENERGY DRINK UNFLAVORED CUP 8.3 LIQ…         1
##  8 MYTHICAL BEVERAGE MAXX ENERGY DRINK RAD RED ZERO SUGAR CUP 12…         1
##  9 MYTHICAL BEVERAGE REHAB DRINK  FLAVORED ENERGY DRINK DRINK  A…         1
## 10 MYTHICAL BEVERAGE REHAB ENERGY DRINK KIWANO  CUP 15.5 LIQUID …         1
## 11 MYTHICAL BEVERAGE ULTRA SUNRISE ENERGY DRINK ULTRA SUNRISE ZE…         1
```

```
## 12 MYTHICAL BEVERAGE ZERO ULTRA ENERGY DRINK UNFLAVORED CUP 12 L…              1
## 13 POW-POW DIETARY HEALTH SUPPLEMENT LIQUID POTENT BRAIN AND BOD…              1
## 14 REAL-TIME THE KEEN  EDITION ENERGY DRINK CRISP KEEN  SUGAR FR…              1
## 15 RULE TEMPEST REVITALIZING BOOST LIQUID CLEAN ENERGY CUP 12 CO…              1
## 16 RULE TEMPEST REVITALIZING BOOST LIQUID CLEAN ENERGY CUP 12 LI…              1
```

```
# There are only 16  items that match time period caloric segment, category. flavor, packaging
and brand dont exist

df %>%
  filter(str_detect(ITEM, "KIWANO")) %>%
  summarize(distinct_items = n_distinct(ITEM))
```

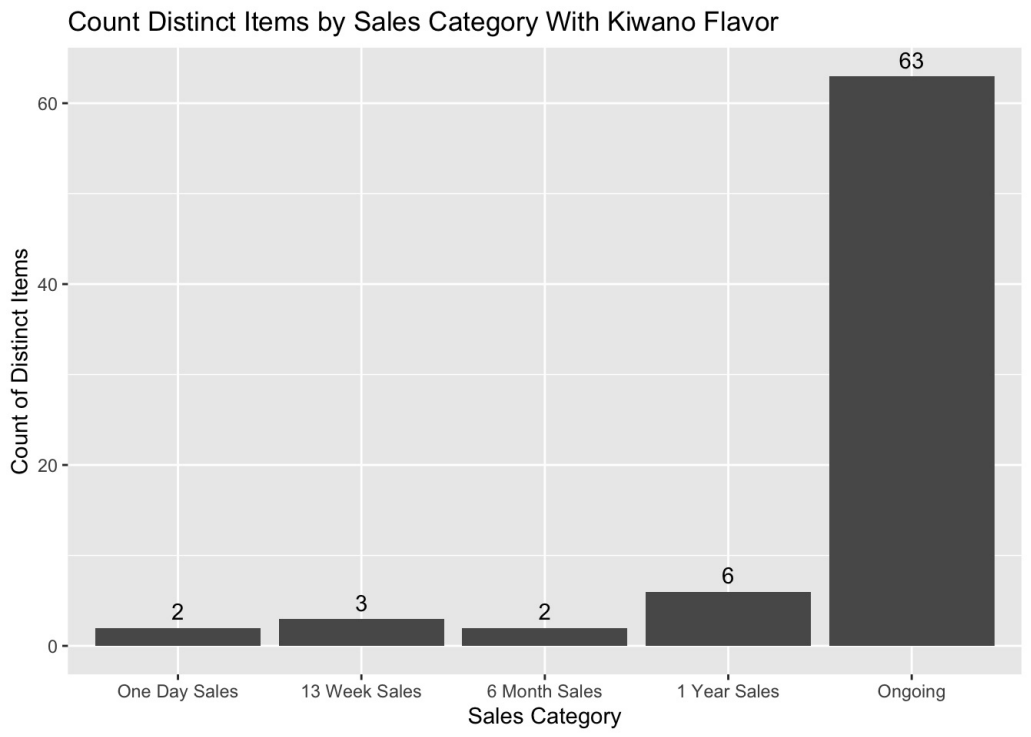
```
##   distinct_items
## 1             76
```

```
#76 Kiwano Flavored items

#Distribution of matching items

df %>%
  filter(sales_category == "13 Week Sales",
         CALORIC_SEGMENT == 0,
         CATEGORY == "ENERGY") %>%
  group_by(ITEM, weeks_since_launch) %>%
  summarize(total_unit_sales = sum(UNIT_SALES)) %>%
  ggplot(aes(x = weeks_since_launch, y = total_unit_sales)) +
  geom_bar(stat = "identity") +
  labs(title = "Total Unit Sales by Week Since Launch",
       x = "Week Since Launch",
       y = "Total Unit Sales")
```




```
# Create Sales Category distribution of Kiwano items

df %>%
```

```
  filter(str_detect(ITEM, "KIWANO")) %>%
  group_by(sales_category) %>%
  summarize(distinct_items = n_distinct(ITEM)) %>%
  ggplot(aes(x = sales_category, y = distinct_items)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = distinct_items), vjust = -.5, color = "black") +
  labs(title = "Count Distinct Items by Sales Category With Kiwano Flavor",
       x = "Sales Category",
       y = "Count of Distinct Items")
```



> In this section we find that there is signficatnly less data around energy drinks that match our category. Overall there is 76 current historical KIWANO flavored items with only 3 of those being sold for more than one day but less than 13 weeks.

Question 4 Parameters

Item Description: Diet Square Mulberries Sparkling Water 10Small MLT Caloric Segment: Diet Market Category: Sparkling Water Manufacturer: Swire-CC Brand: Square Package Type: 10Small MLT Flavor: 'Mulberries Swire plans to release this product for the duration of 1 year but only in the Northern region. What will the forecasted demand be, in weeks, for this product?

```
# Matching parameters for Q4

df %>%
  filter(
         sales_category == "1 Year Sales",
         CALORIC_SEGMENT == 0,
         CATEGORY == "SPARKLING WATER") %>%
         #str_detect(ITEM, "MULBERRIES")) %>%
  summarize(distinct_items = n_distinct(ITEM))
```

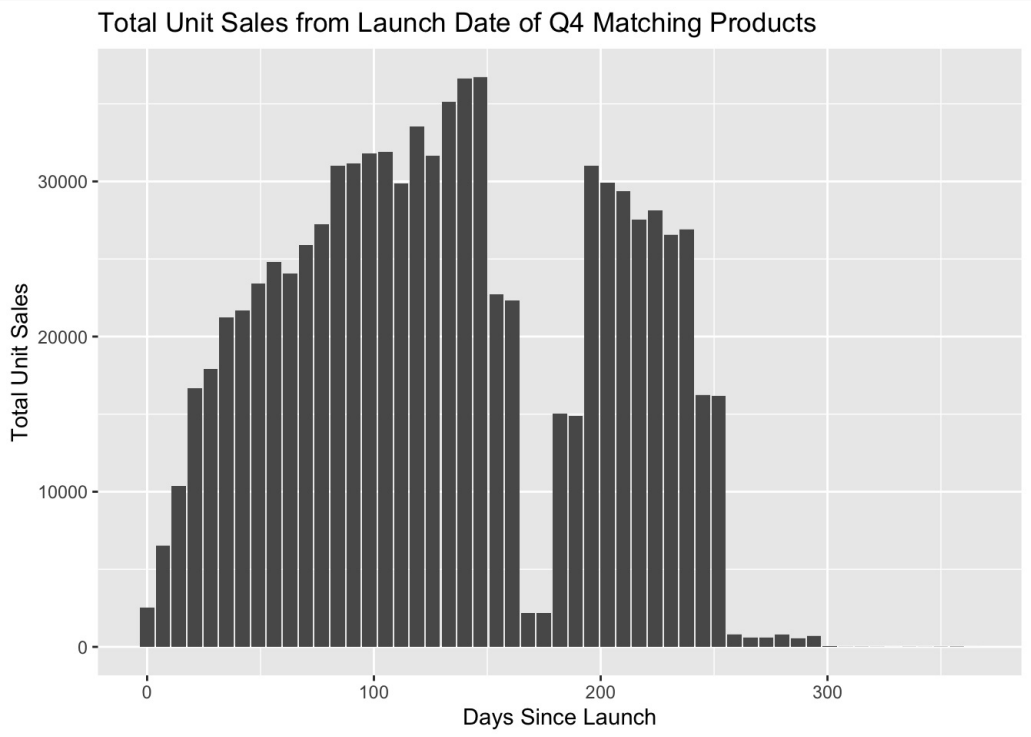
```
##   distinct_items
## 1             34
```

```
#34 items match 1 year launch sales, diet and Sparkling Water category. There are non with
Mulberries in this group
```

```
df %>%
  filter(str_detect(ITEM, "MULBERRIES")) %>%
  summarize(distinct_items = n_distinct(ITEM))
```

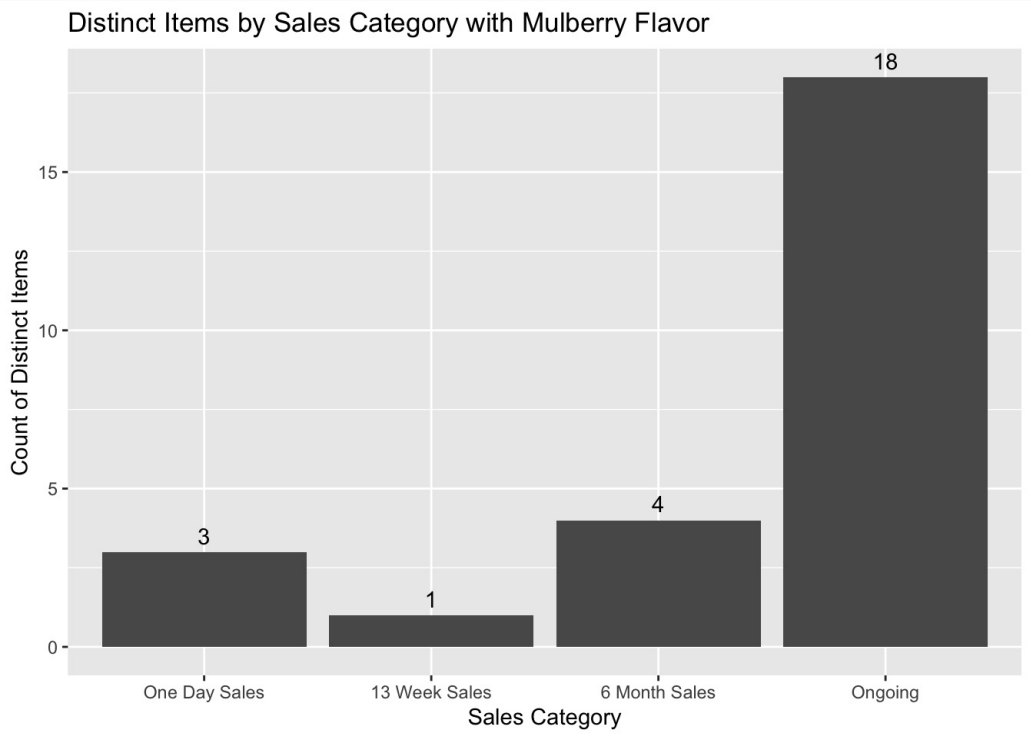
```
##   distinct_items
## 1             26
```

```
#26 Mulberries Flavored items

#Distribution of matching items

df %>%
  filter(
         sales_category == "1 Year Sales",
         CALORIC_SEGMENT == 0,
         CATEGORY == "SPARKLING WATER",
         sales_category != 'Ongoing') %>%
  group_by(ITEM, days_since_launch) %>%
  summarize(total_unit_sales = sum(UNIT_SALES)) %>%
  ggplot(aes(x = days_since_launch, y = total_unit_sales)) +
  geom_bar(stat = "identity") +
  labs(title = "Total Unit Sales from Launch Date of Q4 Matching Products",
       x = "Days Since Launch",
       y = "Total Unit Sales")
```




```
# Create Sales Category distribution of Mulberries items

df %>%
  filter(str_detect(ITEM, "MULBERRIES")) %>%
  group_by(sales_category) %>%
  summarize(distinct_items = n_distinct(ITEM)) %>%
  ggplot(aes(x = sales_category, y = distinct_items)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = distinct_items), vjust = -.5, color = "black") +
  labs(title = "Distinct Items by Sales Category with Mulberry Flavor",
```

```
       x = "Sales Category",
       y = "Count of Distinct Items")
```



> In this section as with the energy drinks, we find there is much less data around sparkling water with only 34 items matching category, segment and sales category. We will need to add in demographic data for just sales in the norther region when setting up modeling. Also, of intreset there currently has been no Mulberry products placed on the market for only 1 year.

Question 5 Parameters

Item Description: Greetingle Health Beverage Woodsy Yellow .5L 12One Jug Caloric Segment: Regular Market Category: ING Enhanced Water Manufacturer: Swire-CC Brand: Greetingle Package Type: .5L 12One Jug Flavor: 'Woodsy Yellow' Swire plans to release this product for 13 weeks, but only in one region. Which region would it perform best in?

```
# Matching parameters for Q5

df %>%
  filter(
    #sales_category == "13 Week Sales",
    CALORIC_SEGMENT == 1,
    CATEGORY == "ING ENHANCED WATER",
    #BRAND == "GREETINGLE",
    #str_detect(ITEM, "WOODSY YELLOW")
  ) %>%
  group_by(ITEM) %>%
  summarize(distinct_items = n_distinct(ITEM))
```

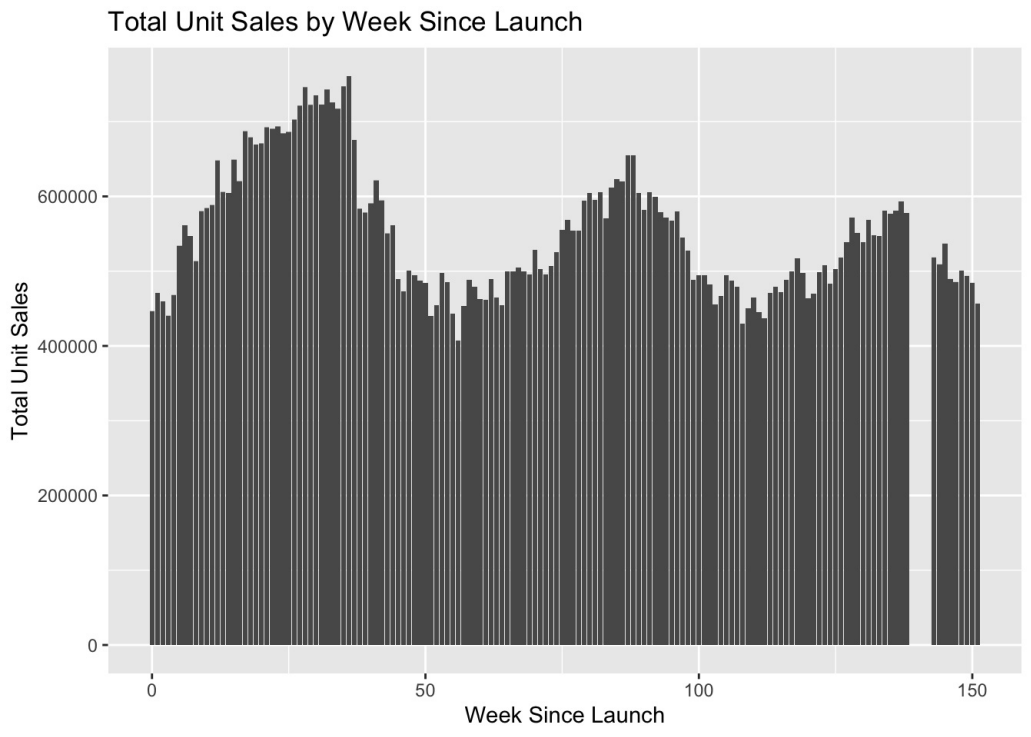
```
## # A tibble: 55 × 2
##    ITEM                                                    distinct_items
##    <chr>                                                            <int>
##  1 MYTHICAL BEVERAGE HYDRO ENERGY DRINK BLUE ICE JUG 25.4 LIQUID…       1
##  2 MYTHICAL BEVERAGE HYDRO ENERGY DRINK MANIC CANES  CUP 16.9 LI…       1
##  3 MYTHICAL BEVERAGE HYDRO ENERGY DRINK MANIC CANES  JUG 25.4 LI…       1
##  4 MYTHICAL BEVERAGE HYDRO ENERGY DRINK MEAN CUSTARD APPLE  CUP …       1
##  5 MYTHICAL BEVERAGE HYDRO ENERGY DRINK MEAN CUSTARD APPLE  JUG …       1
##  6 MYTHICAL BEVERAGE HYDRO ENERGY DRINK PURPLE EXCITEMENT  JUG 2…       1
##  7 MYTHICAL BEVERAGE HYDRO ENERGY DRINK WOODSY  THUNDER CUP 16.9…       1
##  8 MYTHICAL BEVERAGE HYDRO ENERGY DRINK WOODSY  THUNDER JUG 25.4…       1
```

```
##  9 MYTHICAL BEVERAGE HYDRO ENERGY WATER BLUE ICE JUG 20 LIQUID S…              1
## 10 MYTHICAL BEVERAGE HYDRO ENERGY WATER BLUE ICE JUG 20 LIQUID S…              1
## # ℹ 45 more rows
```

```
# There are only 55 items that match Caloric Segment and Category. This one the sales cateogry
also does not have matches.

df %>%
  filter(str_detect(ITEM, "WOODSY YELLOW")) %>%
  summarize(distinct_items = n_distinct(ITEM))
```

```
##   distinct_items
## 1              0
```

```
#0 Flavored items

#Distribution of matching items

df %>%
  filter(#sales_category == "13 Week Sales",
         CALORIC_SEGMENT == 1,
         CATEGORY == "ING ENHANCED WATER") %>%
  group_by(ITEM, weeks_since_launch) %>%
  summarize(total_unit_sales = sum(UNIT_SALES)) %>%
  ggplot(aes(x = weeks_since_launch, y = total_unit_sales)) +
  geom_bar(stat = "identity") +
  labs(title = "Total Unit Sales by Week Since Launch",
       x = "Week Since Launch",
       y = "Total Unit Sales")
```



> This product proved to be the least common to our current data. The enhanced water segement is very small and has not had any limited releases like this before. Also there has been no sales of this flavor in the past. This will be one where our estimate will have many larget assumptions especially once we bring in the Demographic of only selling in one region.

Question 6 Parameters

Item Description: Diet Energy Moonlit Casava 2L Multi Jug Caloric Segment: Diet Market Category: Energy Manufacturer: Swire-CC Brand: Diet Moonlit Package Type: 2L Multi Jug Flavor: ‘Cassava’ Swire plans to release this product for 6 months. What will the forecasted demand be, in weeks, for this product?

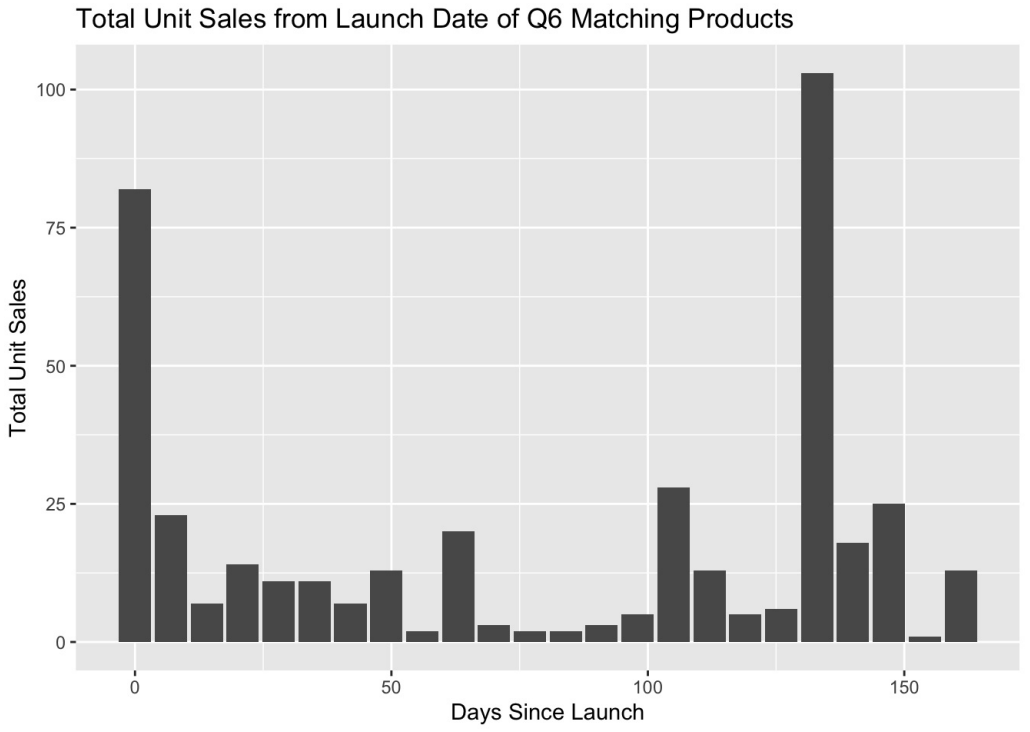
```
# Matching parameters for Q6
df %>%
  filter(
         sales_category == "6 Month Sales",
         CALORIC_SEGMENT == 0,
         CATEGORY == "ENERGY"
         #str_detect(ITEM, "CASSAVA")
         ) %>%
  summarize(distinct_items = n_distinct(ITEM))
```

```
##   distinct_items
## 1             10
```

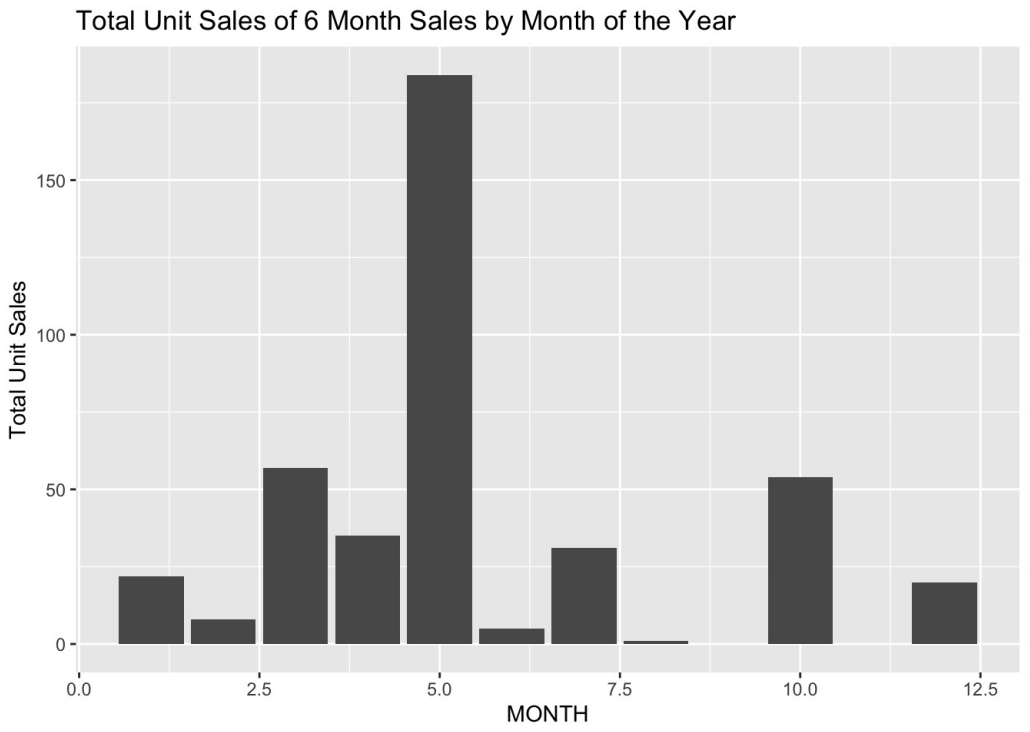
```
#10 items match 6 Month launch sales, diet and Energy category. There are non with Cassava in
this group

df %>%
  filter(str_detect(ITEM, "CASSAVA")) %>%
  summarize(distinct_items = n_distinct(ITEM))
```

```
##   distinct_items
## 1              0
```

```
# 0 Cassava Flavored items

#Distribution of matching items

df %>%
  filter(
         sales_category == "6 Month Sales",
         CALORIC_SEGMENT == 0,
         CATEGORY == "ENERGY",
         sales_category != 'Ongoing') %>%
  group_by(ITEM, days_since_launch) %>%
  summarize(total_unit_sales = sum(UNIT_SALES)) %>%
  ggplot(aes(x = days_since_launch, y = total_unit_sales)) +
  geom_bar(stat = "identity") +
  labs(title = "Total Unit Sales from Launch Date of Q6 Matching Products",
       x = "Days Since Launch",
       y = "Total Unit Sales")
```




```
# Distribution of Sales by month of the year

df %>%
  filter(
         sales_category == "6 Month Sales",
         CALORIC_SEGMENT == 0,
         CATEGORY == "ENERGY",
         sales_category != 'Ongoing') %>%
  group_by(ITEM, MONTH) %>%
  summarize(total_unit_sales = sum(UNIT_SALES)) %>%
  ggplot(aes(x = MONTH, y = total_unit_sales)) +
  geom_bar(stat = "identity") +
  labs(title = "Total Unit Sales of 6 Month Sales by Month of the Year",
       x = "MONTH",
       y = "Total Unit Sales")
```



> This product grouping as with the last will need to be built on many assuptions. There currently is no other products with CASSAVA in the market. There are 10 itmes that have been sold for 6 months at a time that are in the energy drink category. For this we will need to expand out our data set to hone in on which 6 months of the year would be best. In a graph of sales we do see a window of time that these matching products have sold the most, Feb - July.

Question 7 Parameters

Item Description: Peppy Gentle Drink Pink Woodsy .5L Multi Jug Caloric Segment: Regular Type: SSD Manufacturer: Swire-CC Brand: Peppy Package Type: .5L Multi Jug Flavor: 'Pink Woodsy' Swire plans to release this product in the Southern region for 13 weeks. What will the forecasted demand be, in weeks, for this product?

```
# Matching parameters for Q3

df %>%
  filter(
    sales_category == "13 Week Sales",
    CALORIC_SEGMENT == 1,
    CATEGORY == "SSD",
    #BRAND == "PEPPY",
    #str_detect(ITEM, "PINK WOODSY")
  ) %>%
  group_by(ITEM) %>%
  summarize(distinct_items = n_distinct(ITEM))
```

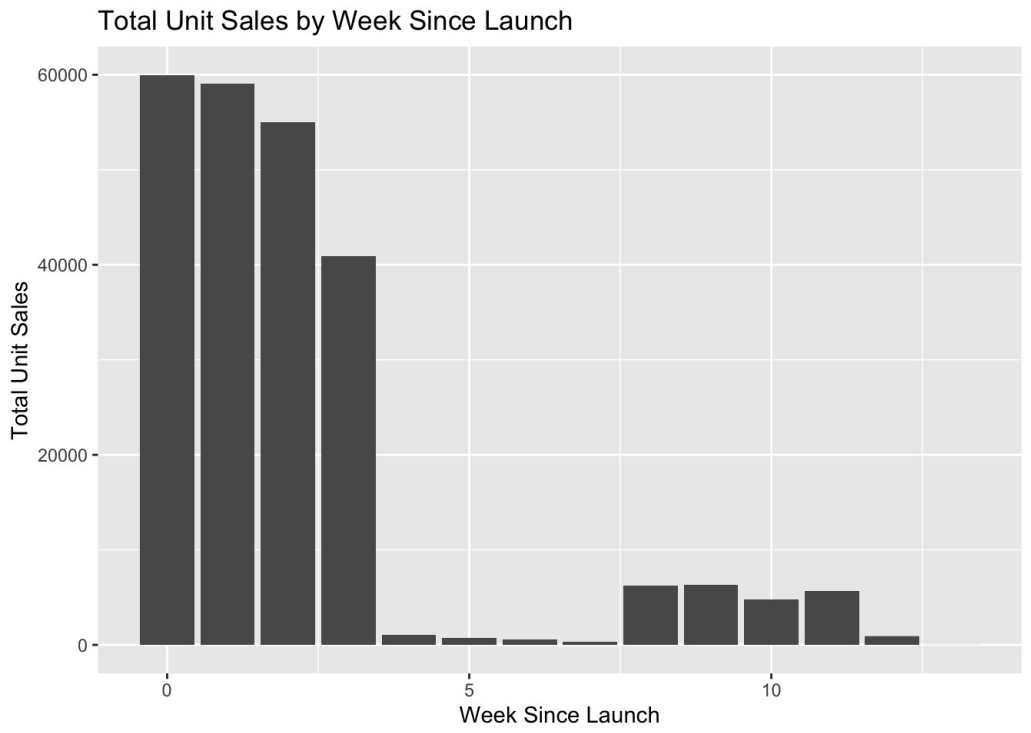
```
## # A tibble: 62 × 2
##    ITEM                                                    distinct_items
##    <chr>                                                            <int>
##  1 AZURE HORIZON GENTLE DRINK WILD PINK  CUP 12 LIQUID SMALL                1
##  2 BARS GENTLE DRINK PINA JUG 67.6 LIQUID SMALL                            1
##  3 BARS GENTLE DRINK PONCHE WOODSY  JUG 67.6 LIQUID SMALL                  1
##  4 CUPSHIELD'S TONIC WATER UNFLAVORED JUG 10 LIQUID SMALL                  1
##  5 DESERT REFRESHMENT GENTLE DRINK SUNSET  CASAVA  BLAST CUP 12 …          1
##  6 ELF BUBBLES GENTLE DRINK MELLOW D MIXED-TROPPY  JUG 16.9 LIQU…          1
##  7 ELF BUBBLES GENTLE DRINK RED  SUNSET  SUPER-JUICE  DURIAN  JU…          1
##  8 ELF BUBBLES GENTLE DRINK SUMMER MELLOW D MIXED-TROPPY  SUPER-…          1
##  9 FANTASMIC GENTLE DRINK BERRY CUP 12 LIQUID SMALL X12                    1
## 10 FANTASMIC GENTLE DRINK CASAVA  JUG 12 LIQUID SMALL X4                   1
## # ℹ 52 more rows
```

```
# There are 62 items that match time period caloric segment, category. flavor, packaging and
brand combination do not exist

df %>%
  filter(str_detect(ITEM, "PINK WOODSY")) %>%
  summarize(distinct_items = n_distinct(ITEM))
```

```
##   distinct_items
## 1              0
```

```
#0 items have Pink Woodsy Flavored items

#Distribution of matching items

df %>%
```

```
  filter(sales_category == "13 Week Sales",
         CALORIC_SEGMENT == 1,
         CATEGORY == "SSD") %>%
  group_by(ITEM, weeks_since_launch) %>%
  summarize(total_unit_sales = sum(UNIT_SALES)) %>%
  ggplot(aes(x = weeks_since_launch, y = total_unit_sales)) +
  geom_bar(stat = "identity") +
  labs(title = "Total Unit Sales by Week Since Launch",
       x = "Week Since Launch",
       y = "Total Unit Sales"
  )
```



> This product does have more compariable data since it is in the SSD category. We should be able to create a good data set once linked with the region data around a 13 week forecast. The distribution of the 13 week sales products in the SSD category follows the general distribution with large sales in week 0 and 1 very small sales in the middle and a bump in the final weeks.

Summary

> In this section, we conducted an extensive analysis of our modeling datasets, considering various factors that may impact our modeling outcomes. We encountered several considerations, including the presence of items falling into multiple categories and the need to filter out irrelevant categories. Particularly noteworthy was the potential advantage of discarding ongoing sales data, as these products may introduce unnecessary noise into our models. Our analysis of sales tenure distributions revealed significant differences across categories, emphasizing the importance of understanding these variations in our modeling efforts.

> One significant aspect of our analysis was the examination of sales distributions throughout the year, highlighting substantial variations between product groups and months. This temporal variation is expected to be a crucial variable in our modeling endeavors.

> Furthermore, we explored the challenges associated with modeling our innovation products. We discovered limited data points meeting the criteria for these products, with some featuring unique flavors that have never been sold before and others representing entirely new product types within specific selling categories (e.g., 13 weeks, 6 months, 1 year). These initial findings shed light on the potential challenges and opportunities in accurately modeling these innovative products.

> In conclusion, our analysis provides valuable insights into the complexities of our datasets and the considerations necessary for effective modeling. Moving forward, we will leverage these insights to refine our modeling approaches and develop more accurate predictive models.

```
stopImplicitCluster()
```