EDA - Zero to Coke Hero: MSBA Capstone Spring 2024

Ian Donaldson, Michael Tom, Andrew Walton, Jake Jarrard

February 26, 2024

EDA Introduction, Purpose, & Objectives

Business Problem:

> Swire Coca-Cola operates extensively, producing, selling, and distributing beverages across 13 states in the American West. The company is known for its regular introduction of new, often limited-time, products aimed at stimulating market demand. However, predicting accurate demand for these innovative products remains a significant challenge. Historical data provides some insight but is insufficient for precise forecasting due to variations in regional and demographic appeal.

> The firm stands at the forefront of the Western US beverage distribution sector, continually launching limited edition products to maintain consumer interest and market dominance. Yet, the uncertainty in demand forecasting for these unique items presents risks of either overproduction or shortages, each carrying potential financial and reputational impacts. The project aims to leverage demographic and historical sales data to enhance the accuracy of demand predictions for future innovative offerings.

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.

Analytics Approach:

> Our analytics team is set to use Exploratory Data Analysis (EDA), Time Series Forecasting techniques like ARIMA and SARIMA models, and Machine Learning algorithms to analyze over 24 million observations collected over three years. This vast dataset includes sales data for more than 3100 unique products across 13 western states with varied demographics. Our goal is to identify the optimal launch periods for new products and pinpoint the essential attributes contributing to their success, while also managing production-related costs effectively. We will adhere to standard data science methodologies, including the division of data into training and testing sets to ensure the validity of our model predictions and demand forecasts.

> The project presents several analytical challenges. Firstly, we must determine the best approach to integrate external data sources, such as demographic and zip code information, with our primary market demand dataset to enhance regional and demographic-specific demand forecasting. Secondly, the variability in the release durations of historical and future products, ranging from 13 weeks to six months, necessitates tailored analytical approaches to utilize historical data effectively. By addressing these challenges, our team aims to develop robust models that will enable Swire Coca-Cola to accurately forecast demand for innovative products, thereby optimizing inventory levels and enhancing customer satisfaction.

Success Metrics:

> Our success hinges on delivering accurate demand forecasts enabling Swire to balance supply with demand efficiently, ensuring no surplus stock or unmet consumer needs. The effectiveness of our forecasts will be reflected in the company's profitability, evaluated through our modeling outcomes against industry-standard performance indicators.

> We will measure the project's success through the precision of our demand predictions for innovative products against real sales data. Key performance indicators include the model's ability to accurately determine the timing, location, duration, and quantity for each new product launch. These metrics aim to provide Swire with valuable insights, leading to financial benefits and an increased market share for their new offerings.

Description of the Data:

> Our analysis will be supported by one critical dataset, essential for the construction of our predictive models, and a tertiery secondary demographic dataset.

> The primary dataset is Market Demand, encompassing historical sales data across diverse product categories, brands, caloric segments, and packaging types. This dataset is extensive, with over 24 million records, presenting significant considerations for data handling and analysis.

> The secondary dataset comprises demographic details including age, income, gender, household income, and segmentation, coupled with zip code information. This allows for a nuanced approach to forecasting, enabling predictions tailored to specific demographic profiles based on historical sales patterns and census data.

> Collectively, these datasets form a comprehensive time series analysis framework, comprising over 24 million observations across 13 features. This includes detailed sales information spanning three years for Swire Coca-Cola products, covering 13 western states and 3100 unique products, segmented into various sales territories, innovation product lines, and brand categories, alongside metrics for seasonal trends, financial impacts, and market dynamics.

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.

Scope:

> At the outset of this EDA endeavor, the project will attempt to 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. All members of the analytics team contributed equally to the production of this EDA effort and exploration of the data.

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.

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. It would be naive to suppose our EDA captured all of the essential elements we would want to know. Its size and complexity will be nearly impossible to fully understand. Ulimately we will need to break the dataset down into useable sub dataframes that will allow for very boutique feature engineering to capture inherent numeric information in each row, by product, flavor, market region, etc. However, this EDA experience in its entirety answers nearly all essential questions beneath the surface of 24 million rows and billions of dollars of soft drink sales in recent years.

SECTION SUMMARIES

We have provided the following EDA in four sections. Summaries of each follow. Some comments are provided in individual sections when necessary. Most comments are captured inside R code output. Most code has been largely suppressed as has lengthy output.

PART 1: SUMMARY STATISTICS & ESSENTIALS

Several insights can be derived from the comprehensive review of the market demand datasheet. This section captures the essentials any data scientist would want to hear up front in terms of describing the data. Highlights as follows…

Upon examination of the DATE column, we observe approximately three years of data. Certain products within the dataset span the entire three years, while others are cataloged for less than six months. Products with shorter durations are likely to be more pertinent to our predictions regarding innovative products.

The analysis of the DATE attribute proves crucial, suggesting the potential utility of segmenting our models based on varying date ranges, such as 13 weeks or less than six months, to align more closely with the specific queries we aim to resolve.

Regarding CATEGORY, while energy drinks generate the highest revenue per unit, Soft Sodas (SSD) emerge as the predominantly sold category. This dynamic warrants closer inspection.

In comparison with other manufacturers, Swire-Coca-Cola (Swire-CC) presents an average performance in terms of total sales revenue and units sold. The concern arises from their lower unit price relative to competitors, leading to higher volume sales required to match the revenue of other manufacturers, which could potentially erode profit margins. Our analysis will thus aim to identify products with higher average unit prices for enhanced profitability.

Between the two caloric segments, diet/light and regular, both command similar prices per unit. However, regular beverages significantly outperform diet/light in both unit sales and total revenue, indicating a stronger market demand for regular sodas.

The primary challenge lies in sifting through the voluminous data to determine the most relevant fields and questions for focus. Identifying these areas will enable us to harness the extensive dataset effectively, feeding into the models we develop for robust analysis.

PART 2: DEEPER DIVE INTO INNOVATION CHARACTERISTICS

The sales data analysis of Swire Coca-Cola unveils critical insights. Initially, the dataset had 59,725 missing entries, all addressed through text analysis, achieving a state of zero missing values. The dataset covers 152 unique weeks, providing comprehensive weekly data for nearly three years, showcasing consistency in time-series data across two caloric segments, five categories, eight manufacturers, 319 brands, 103 package types, and 3,692 unique product descriptions.

Delving into product specifics, characteristics emerge, such as Diet Smash belonging to the energy category with three types of packaging. The Sparkling Jacceptabletlester brand offers both diet and regular options, available in diverse packaging, while Venomous Blast provides diet and regular choices, predominantly in “16small multi cup” packages. In manufacturer rankings, Swire holds the third position in sales, with Bubble Joy Advantageous leading in brand sales, followed by Real-Time and Peppy.

Data analysis indicates a left skew in brand run times, displaying a median of 137 weeks against a mean of 99.1 weeks, reflecting varied brand life spans. In contrast, package tenure shows right skewness, with longer tenures for established package sizes, evidenced by a median of 147 weeks and a mean of 117 weeks. However, 14 package sizes exhibit durations of less than six months, potentially signaling new packaging innovations.

Seasonal trends present inconsistencies, with missing weeks and notable fluctuations among all analyzed beverages, prompting further examination to clarify whether these arise from data omissions or natural product lifecycle conclusions. This necessitates meticulous validation of product start and end dates, especially for brands with intermittent data.

In conclusion, the detailed analysis sheds light on significant aspects of Swire Coca-Cola's sales trends and product dynamics, laying a foundation for enhanced forecasting accuracy and strategic product launch planning.

PART 3: EVEN DEEPER - MODELING EDA

This segment of the EDA is dedicated to organizing the data into suitable sets for modeling. We aim to identify any data inaccuracies that may lead to incorrect predictions and isolate segments that might introduce noise into the models. The final objective is to discern which attributes of our innovative products currently exist in our dataset, ensuring a clean and relevant data environment for predictive modeling.

During this phase, we undertook a comprehensive review of the datasets intended for model development. Key tasks included identifying multi-category items and filtering out non-essential categories to enhance model clarity. We noted the potential benefit of excluding data on ongoing sales, which could otherwise obscure the models with irrelevant variability. Additionally, our examination of sales tenure distributions by category underscored the need to incorporate these divergences into our modeling strategy effectively.

A crucial part of our evaluation focused on the yearly sales distribution, revealing significant month-to-month and group-to-group variations. Recognizing these temporal differences will be instrumental in shaping our models, especially in predicting seasonal trends and consumer behavior.

Delving into the specifics of modeling our innovation products posed unique challenges, given the sparse data for certain new product launches, including those with novel flavors or categories. Our preliminary analysis highlighted the intricacies and potential obstacles in accurately forecasting demand for these unique items.

Summarizing our findings, the exploration offers critical insights into our data's complexity and the strategic adjustments needed for our models. As we progress, these insights will guide the enhancement of our modeling techniques, aiming to produce more precise and actionable predictive analytics for our innovative product lines.

EDA PART 4: DEMOGRAPHIC DATA

We used several tools to reshape the provided demographic data. We transformed and un-pivoted criteria, segment, and count and created a wide format dataframe from the original XLS and preserved all numbers. We conducted simple calculations to capture the aggregate population and households of each zip code and then we used the zip_to_market_unit_mapping.csv to assign each zip code to its appropriate Market Segment. We removed all Zip codes that were not associated with a Market Segment. We renamed the columns as appropriate. Our dataset contains all 2232 Zip codes with their respective demographic data in 83 total columns. The following exploration captures the essential elements of only the demographic data prior to associating it with the main Swire dataset. Despite learning a great deal about the demographic data, it remains to be seen if it can be married properly to the main Swire dataset for use in modeling.

EDA - PART 1: SUMMARY STATISTICS & ESSENTIALS

Summary of Data Set and Columns

```
#Quick summary of data inspection and manipulation:
#1: Factored a number of columns
#2: Imputed all NA's (The only NAs were in the CALORIC_SEGMENT column and we used text analytics to impute DIET & REGULAR)
#3: Ensured DATE column data type was convered to a date as.Date()
#4: Created a UNIT_PRICE column: swire_df$UNIT_PRICE <- swire_df$DOLLAR_SALES / swire_df$UNIT_SALES

summary(swire_df)
```

```
##       DATE              MARKET_KEY         CALORIC_SEGMENT
##  Min.   :2020-12-05   1811   :  198609   DIET/LIGHT:12174588
##  1st Qu.:2021-08-14   1172   :  185475   REGULAR   :12286836
##  Median :2022-04-23   6802   :  159909
##  Mean   :2022-04-25   602    :  146398
##  3rd Qu.:2022-12-31   1135   :  145933
##  Max.   :2023-10-28   117    :  141489
##                       (Other):23483611
##                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             UNIT_PRICE
##  16SMALL MULTI CUP: 3065844   Length:24461424    Min.   :  0.00333
##  20SMALL MULTI JUG: 2877015   Class :character   1st Qu.:  2.00429
##  12SMALL 12ONE CUP: 2870763   Mode  :character   Median :  3.18000
##  2L MULTI JUG     : 1896248                      Mean   :  4.52689
##  .5L 6ONE JUG     : 1453399                      3rd Qu.:  5.34105
##  12SMALL 8ONE CUP : 1408738                      Max.   :224.99000
##  (Other)          :10889417
```

```
#Outliers:

# UNIT_PRICE: The Max Unit Price seems very high compared to the rest of the observations; no apparent explanation for this as there doesn't seem to be any products in the data set that should sell for that price.

# UNIT_SALES: The Max Unit Sales seems very high compared to the rest of the observations. This could be that they have a certain product that is very popular included in the data set, but it would be worth exploring
```

Exploratory Visualizations, Summary Tables & Essential Takeaways

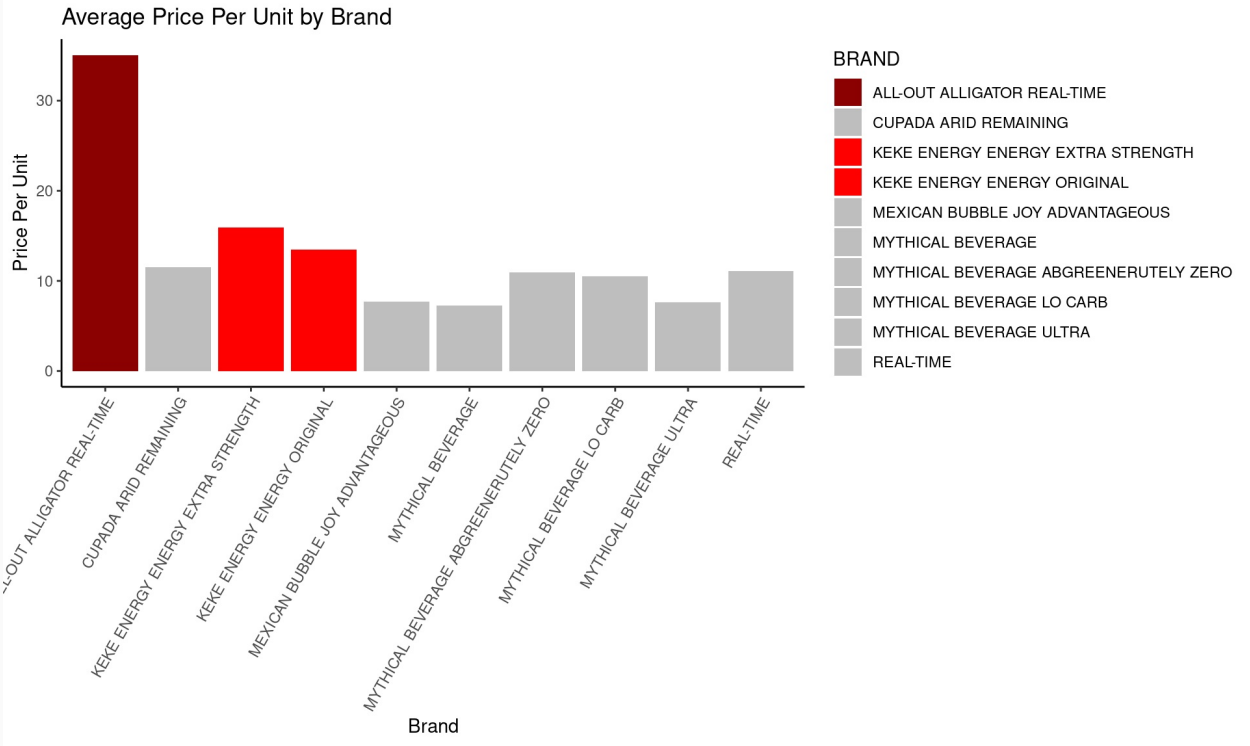
Takeaway 1: Average Price Per Unit

```
brand_ppu_summary <- swire_df %>%
  group_by(BRAND) %>%
  summarise(n = n(),
            avg_price_unit = mean(UNIT_PRICE),
            avg_sales_dollars = mean(DOLLAR_SALES),
            avg_unit_sales = mean(UNIT_SALES)
  ) %>%
  arrange(desc(avg_price_unit)) %>%
  head(10)

brand_ppu_summary
```

```
## # A tibble: 10 × 5
##    BRAND                       n avg_price_unit avg_sales_dollars avg_unit_sales
##    <fct>                   <int>          <dbl>             <dbl>          <dbl>
##  1 ALL-OUT ALLIGATOR REA…  11545          35.1               101.           2.32
##  2 KEKE ENERGY ENERGY EX… 263320          15.9               423.          74.5
##  3 KEKE ENERGY ENERGY OR… 207666          13.5               155.          23.9
##  4 CUPADA ARID REMAINING     488          11.5               763.          66.4
##  5 REAL-TIME              531911          11.1              1950.         505.
##  6 MYTHICAL BEVERAGE ABG…  56240          10.9               544.         173.
##  7 MYTHICAL BEVERAGE LO … 142432          10.5               482.         127.
##  8 MEXICAN BUBBLE JOY AD…  85034           7.69              675.         206.
##  9 MYTHICAL BEVERAGE ULT… 718536           7.60              695.         206.
## 10 MYTHICAL BEVERAGE      271932           7.26             1702.         521.
```

```
ggplot(brand_ppu_summary, aes(x = BRAND )) +
  geom_bar(aes(y = avg_price_unit, fill = BRAND), stat = "identity", position = "dodge") +
  labs(title = "Average Price Per Unit by Brand",
       x = "Brand",
       y = "Price Per Unit") +
   scale_fill_manual(values = c("ALL-OUT ALLIGATOR REAL-TIME" = "darkred", "KEKE ENERGY ENERGY EXTRA STRENGTH" = "red", "KEKE ENERGY ENERGY ORIGINAL" = "red", "CUPADA ARID REMAINING" = "grey", "REAL-TIME" = "grey", "MYTHICAL BEVERAGE ABGREENERUTELY ZERO" = "grey", "MYTHICAL BEVERAGE LO CARB" = "grey", "MEXICAN BUBBLE JOY ADVANTAGEOUS" = "grey", "MYTHICAL BEVERAGE ULTRA" = "grey", "MYTHICAL BEVERAGE" = "grey")) +
  theme_classic() +
```

```
  theme(axis.text.x = element_text(angle = 60, hjust = 1)) +
  theme(axis.text.y = element_text(angle = 0, hjust = 1))
```




```
# Several products have a much higher average unit sales price than the mean ($4.50). This includes one product with and average price of $35.00 and then another tier of products with average prices between $10-$15. These products may not need to sell as many units to be able to reach a high number of total sales dollars due to the high prices associated with the products.

rm(brand_ppu_summary)
```

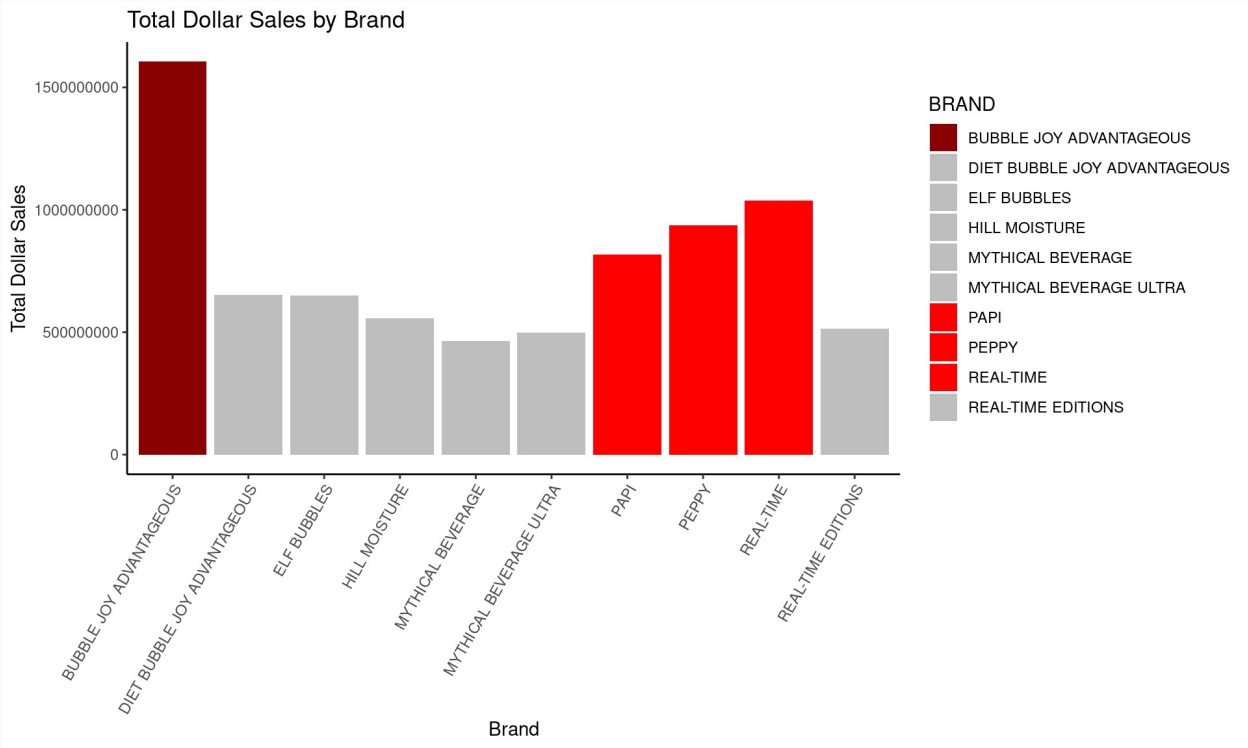
Takeaway 2: Large Total Dollar Sales

```
brand_summary <- swire_df %>%
  group_by(BRAND) %>%
  summarise(n = n(),
            total_sales_dollars = sum(DOLLAR_SALES),
            total_unit_sales = sum(UNIT_SALES)
  ) %>%
  arrange(desc(total_sales_dollars)) %>%
  head(10)

brand_summary
```

```
## # A tibble: 10 × 4
##    BRAND                          n total_sales_dollars total_unit_sales
##    <fct>                      <int>               <dbl>            <dbl>
##  1 BUBBLE JOY ADVANTAGEOUS   443625         1604866764.       404310942.
##  2 REAL-TIME                 531911         1037072280.       268351647.
##  3 PEPPY                     399458          937030445.       251184816.
##  4 PAPI                      454567          816413505.       216457130.
##  5 DIET BUBBLE JOY ADVANTAGEOUS 369807       653091204.       168687013.
##  6 ELF BUBBLES               402084          649144309.       177671835.
##  7 HILL MOISTURE             505286          557565468.       167352606.
##  8 REAL-TIME EDITIONS        827868          514215183.       172718827.
##  9 MYTHICAL BEVERAGE ULTRA   718536          499397532.       147757133.
## 10 MYTHICAL BEVERAGE         271932          462803408.       141581921.
```

```
ggplot(brand_summary, aes(x = BRAND )) +
  geom_bar(aes(y = total_sales_dollars, fill = BRAND), stat = "identity", position = "dodge") +
  labs(title = "Total Dollar Sales by Brand",
       x = "Brand",
```

```
       y = "Total Dollar Sales") +
   scale_fill_manual(values = c("BUBBLE JOY ADVANTAGEOUS" = "darkred", "REAL-TIME" = "red", "PEPPY" = "red",
"PAPI" = "red", "DIET BUBBLE JOY ADVANTAGEOUS" = "grey", "ELF BUBBLES" = "grey", "HILL MOISTURE" = "grey", "REAL-
TIME EDITIONS" = "grey", "MYTHICAL BEVERAGE ULTRA" = "grey", "MYTHICAL BEVERAGE" = "grey")) +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 60, hjust = 1)) +
  theme(axis.text.y = element_text(angle = 0, hjust = 1))
```




```
# A handful of brands exceed $1B in total sales dollars, and several more above $500M. We are obviously dealing
with very large and well known brands in this data set.

rm(brand_summary)
```

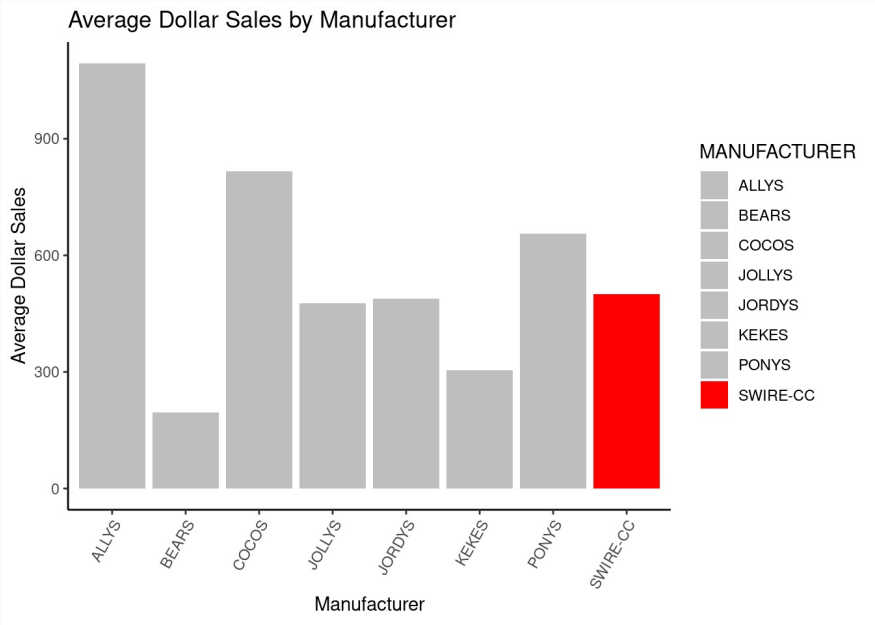
Takeaway 3: Manufacturers Average Sales

```
man_avg_summary <- swire_df %>%
  group_by(MANUFACTURER) %>%
  summarise(n = n(),
            avg_price_unit = mean(UNIT_PRICE),
            avg_sales_dollars = mean(DOLLAR_SALES),
            avg_unit_sales = mean(UNIT_SALES)
  ) %>%
  arrange(desc(avg_sales_dollars))

man_avg_summary
```

```
## # A tibble: 8 × 5
##   MANUFACTURER       n avg_price_unit avg_sales_dollars avg_unit_sales
##   <fct>          <int>          <dbl>             <dbl>          <dbl>
## 1 ALLYS        1428416           8.26             1094.          312.
## 2 COCOS        5595540           3.92              816.          229.
## 3 PONYS        2259095           6.19              656.          210.
## 4 SWIRE-CC     5763809           3.76              501.          145.
## 5 JORDYS        428170           5.45              489.          188.
## 6 JOLLYS       6921978           3.82              477.          151.
## 7 KEKES         470986          14.8               305.           52.2
## 8 BEARS        1593430           3.50              196.           51.4
```

```
ggplot(man_avg_summary, aes(x = MANUFACTURER )) +
  geom_bar(aes(y = avg_sales_dollars, fill = MANUFACTURER), stat = "identity", position = "dodge") +
  labs(title = "Average Dollar Sales by Manufacturer",
```

```
       x = "Manufacturer",
       y = "Average Dollar Sales") +
  scale_fill_manual(values = c("SWIRE-CC" = "red", "JOLLYS" = "grey", "COCOS" = "grey", "ALLYS" = "grey", "PONYS"
= "grey", "BEARS" = "grey", "JORDYS" = "grey", "KEKES" = "grey")) +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 60, hjust = 1)) +
  theme(axis.text.y = element_text(angle = 0, hjust = 1))
```




```
# Swire-CC is 4th in average sales when compared to other manufacturers in the data set. Their average sales are
only a little more than half of the average sales of the ALLYS. Although, they rank #2 in the number of
observations in the data behind Jollys. There are numerous inferences that can be made from the table above.
```

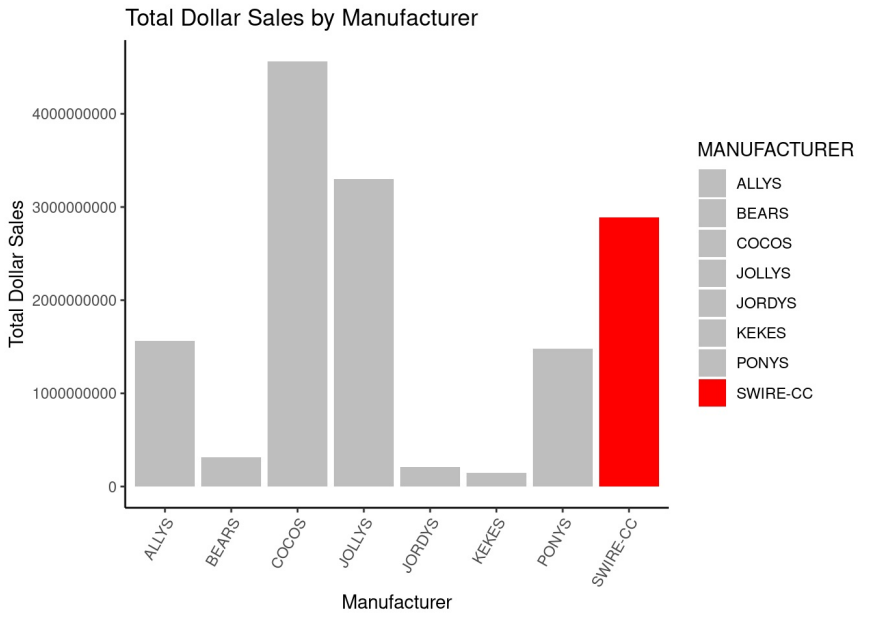
Takeaway 4: Manufacturers Total Sales

```
man_ts_summary <- swire_df %>%
  group_by(MANUFACTURER) %>%
  summarise(n = n(),
            total_sales_dollars = sum(DOLLAR_SALES),
            total_unit_sales = sum(UNIT_SALES)
  ) %>%
  arrange(desc(total_sales_dollars))

man_ts_summary
```

```
## # A tibble: 8 × 4
##   MANUFACTURER       n total_sales_dollars total_unit_sales
##   <fct>          <int>               <dbl>            <dbl>
## 1 COCOS        5595540         4563306476.      1279323647.
## 2 JOLLYS       6921978         3301641671.      1043704245.
## 3 SWIRE-CC     5763809         2885435787.       834941867.
## 4 ALLYS        1428416         1562675378.       445086050.
## 5 PONYS        2259095         1481611289.       475149320.
## 6 BEARS        1593430          312718094.        81974082.
## 7 JORDYS        428170          209238104.        80589148.
## 8 KEKES         470986          143501825.        24594243.
```

```
ggplot(man_ts_summary, aes(x = MANUFACTURER )) +
  geom_bar(aes(y = total_sales_dollars, fill = MANUFACTURER), stat = "identity", position = "dodge") +
  labs(title = "Total Dollar Sales by Manufacturer",
       x = "Manufacturer",
       y = "Total Dollar Sales") +
  scale_fill_manual(values = c("SWIRE-CC" = "red", "JOLLYS" = "grey", "COCOS" = "grey", "ALLYS" = "grey", "PONYS"
= "grey", "BEARS" = "grey", "JORDYS" = "grey", "KEKES" = "grey")) +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 60, hjust = 1)) +
```

```
  theme(axis.text.y = element_text(angle = 0, hjust = 1))
```



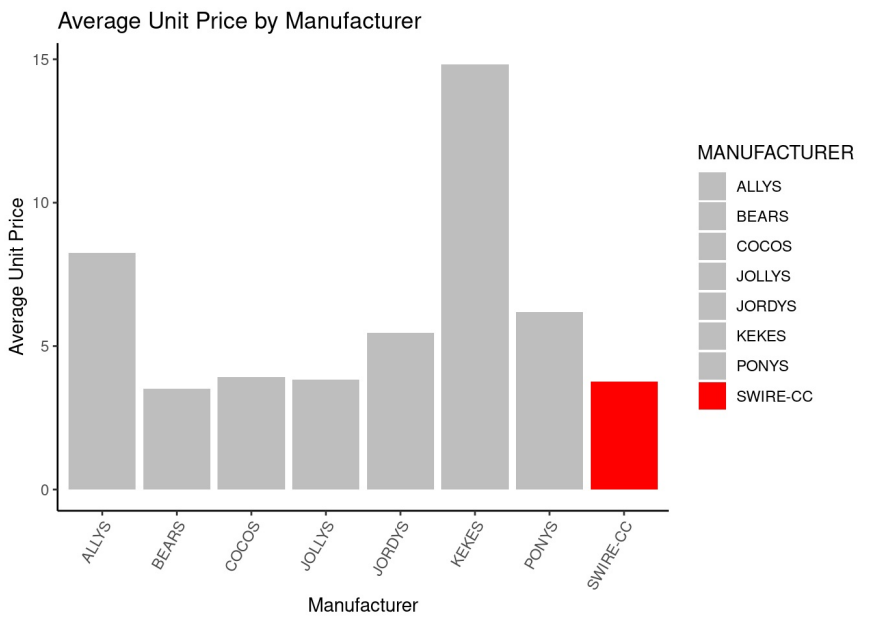

```
# Swire-CC is 3rd total revenues when compared to other manufacturers. COCOS has nearly twice as many total sales as SWIRE-CC, though COCO's also has higher priced packaged items than Swire.
```

Takeaway 5: Manufacturers Average Unit Price

```
man_up_summary<- swire_df %>%
  group_by(MANUFACTURER) %>%
  summarise(n = n(),
            total_sales_dollars = sum(DOLLAR_SALES),
            average_unit_price = mean(UNIT_PRICE)
  ) %>%
  arrange(desc(average_unit_price))

man_up_summary
```

```
## # A tibble: 8 × 4
##   MANUFACTURER       n total_sales_dollars average_unit_price
##   <fct>          <int>               <dbl>              <dbl>
## 1 KEKES         470986          143501825.              14.8
## 2 ALLYS        1428416         1562675378.               8.26
## 3 PONYS        2259095         1481611289.               6.19
## 4 JORDYS        428170          209238104.               5.45
## 5 COCOS        5595540         4563306476.               3.92
## 6 JOLLYS       6921978         3301641671.               3.82
## 7 SWIRE-CC     5763809         2885435787.               3.76
## 8 BEARS        1593430          312718094.               3.50
```

```
ggplot(man_up_summary, aes(x = MANUFACTURER )) +
  geom_bar(aes(y = average_unit_price, fill = MANUFACTURER), stat = "identity", position = "dodge") +
  labs(title = "Average Unit Price by Manufacturer",
       x = "Manufacturer",
       y = "Average Unit Price") +
  scale_fill_manual(values = c("SWIRE-CC" = "red", "JOLLYS" = "grey", "COCOS" = "grey", "ALLYS" = "grey", "PONYS" = "grey", "BEARS" = "grey", "JORDYS" = "grey", "KEKES" = "grey")) +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 60, hjust = 1)) +
  theme(axis.text.y = element_text(angle = 0, hjust = 1))
```




```
# The average unit price has a large range when grouped by manufacturers, and SWIRE-CC is near the bottom of this list; since we know that Coca Cola is the largest soft drink manufacturer in the world, it isn't a surprise that the overall price per ounce of soda is likely less than its competitors.

#KEKES, ALLYS, and PONYS all do significantly less total dollar sales and total unit sales than SWIRE-CC, but all 3 are selling their units at a much higher price; this may prove to be a point worth further discovery as smaller companies that produce in smaller quantities may prove to be a good study for short term product production.

rm(man_avg_summary)
rm(man_up_summary)
rm(man_ts_summary)
```

Takeaway 6: Caloric Segment

```
man_cs_summary <- swire_df %>%
  group_by(CALORIC_SEGMENT) %>%
  summarise(n = n(),
            avg_price_unit = mean(UNIT_PRICE),
            avg_sales_dollars = mean(DOLLAR_SALES),
            avg_unit_sales = mean(UNIT_SALES),
            total_sales_dollars = sum(DOLLAR_SALES)
  ) %>%
  arrange(desc(avg_sales_dollars))

man_cs_summary
```

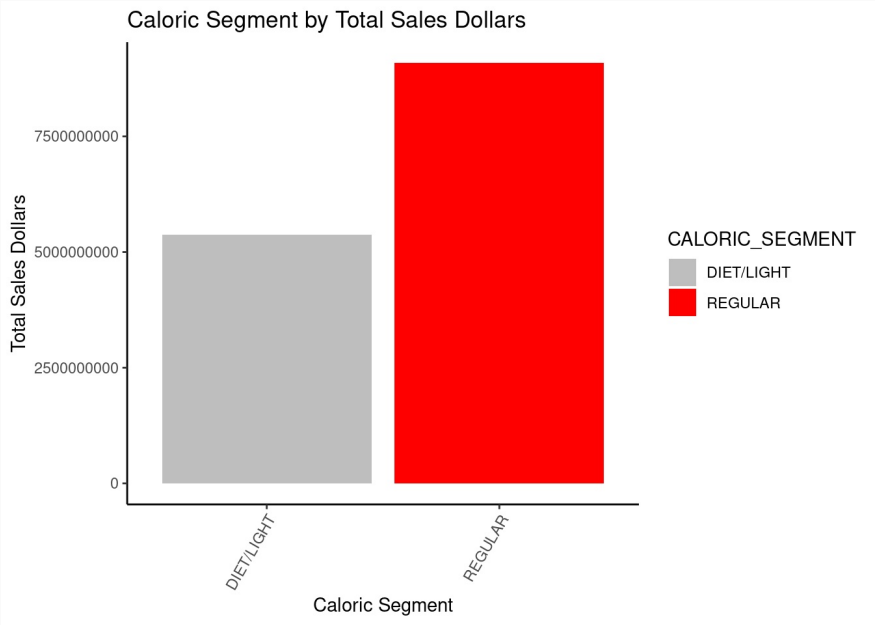
```
## # A tibble: 2 × 6
##   CALORIC_SEGMENT        n avg_price_unit avg_sales_dollars avg_unit_sales
##   <fct>              <int>          <dbl>             <dbl>          <dbl>
## 1 REGULAR         12286836           4.29              739.           220.
## 2 DIET/LIGHT      12174588           4.77              442.           128.
## # ℹ 1 more variable: total_sales_dollars <dbl>
```

```
swire_df %>%
  group_by(CALORIC_SEGMENT) %>%
  summarise(n = n(),
            total_sales_dollars = sum(DOLLAR_SALES),
            total_unit_sales = sum(UNIT_SALES)
  ) %>%
  arrange(desc(total_sales_dollars))
```

```
## # A tibble: 2 × 4
##   CALORIC_SEGMENT        n total_sales_dollars total_unit_sales
##   <fct>              <int>               <dbl>            <dbl>
```

```
## 1 REGULAR         12286836         9081956949.      2702820349.
## 2 DIET/LIGHT      12174588         5378171675.      1562542253.
```

```
ggplot(man_cs_summary, aes(x = CALORIC_SEGMENT )) +
  geom_bar(aes(y = total_sales_dollars, fill = CALORIC_SEGMENT), stat = "identity", position = "dodge") +
  labs(title = "Caloric Segment by Total Sales Dollars",
       x = "Caloric Segment",
       y = "Total Sales Dollars") +
  theme_classic() +
  scale_fill_manual(values = c( "DIET/LIGHT" = "grey", "REGULAR" = "red")) +
  theme(axis.text.x = element_text(angle = 60, hjust = 1)) +
  theme(axis.text.y = element_text(angle = 0, hjust = 1))
```




```
# There is a large difference in total unit sales and total dollar sales between regular soft drinks and
diet/light.  Regular drinks show about double the unit sales and dollar sales even though total observations are
split nearly down the middle 50/50 between regular and diet.

rm(man_cs_summary)
```

Takeaway 7: Category SSD is by far the Biggest Category

```
cat_summary <- swire_df %>%
  group_by(CATEGORY) %>%
  summarise(n = n(),
            avg_price_unit = mean(UNIT_PRICE),
            avg_sales_dollars = mean(DOLLAR_SALES),
            avg_unit_sales = mean(UNIT_SALES),
            total_sales_dollars = sum(DOLLAR_SALES)
  ) %>%
  arrange(desc(avg_sales_dollars))

cat_summary
```

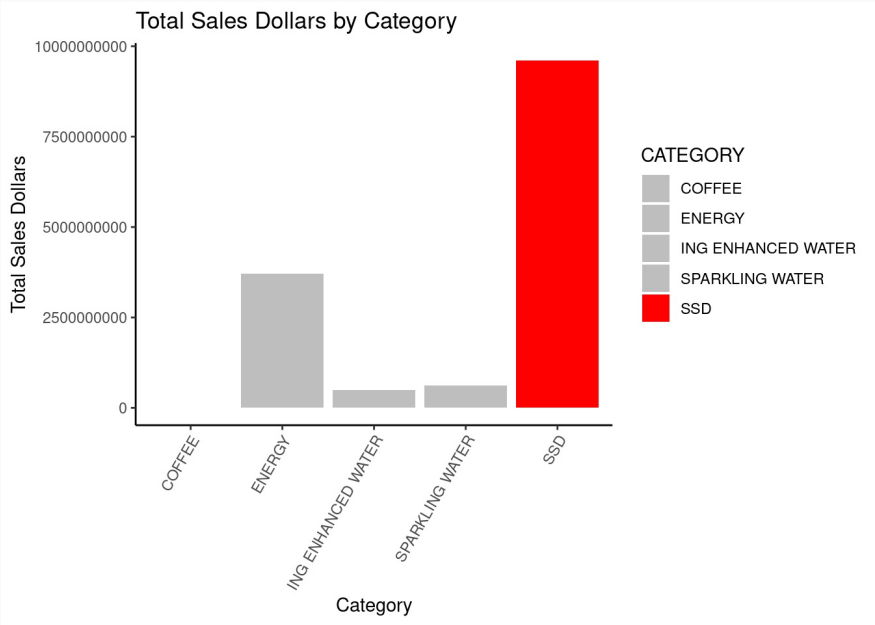
```
## # A tibble: 5 × 6
##   CATEGORY                 n avg_price_unit avg_sales_dollars avg_unit_sales
##   <fct>                <int>          <dbl>             <dbl>          <dbl>
## 1 SSD                12912281           3.91              744.           209.
## 2 ENERGY              5932087           6.60              627.           202.
## 3 SPARKLING WATER     3019064           3.74              208.            55.0
## 4 ING ENHANCED WATER  2452456           3.80              202.            80.4
## 5 COFFEE               145536           2.91               99.0           45.1
## # i 1 more variable: total_sales_dollars <dbl>
```

```
swire_df %>%
```

```
  group_by(CATEGORY) %>%
  summarise(n = n(),
            total_sales_dollars = sum(DOLLAR_SALES),
            total_unit_sales = sum(UNIT_SALES)
  ) %>%
  arrange(desc(total_sales_dollars))
```

```
## # A tibble: 5 × 4
##   CATEGORY                 n total_sales_dollars total_unit_sales
##   <fct>                <int>              <dbl>            <dbl>
## 1 SSD                12912281         9606514926.      2696996876.
## 2 ENERGY              5932087         3718445029.      1198324442.
## 3 SPARKLING WATER     3019064          626470020.       166177968.
## 4 ING ENHANCED WATER  2452456          494293460.       197299015.
## 5 COFFEE               145536           14405189.         6564301
```

```
ggplot(cat_summary, aes(x = CATEGORY )) +
  geom_bar(aes(y = total_sales_dollars, fill = CATEGORY), stat = "identity", position = "dodge") +
  labs(title = "Total Sales Dollars by Category",
       x = "Category",
       y = "Total Sales Dollars") +
  theme_classic() +
  scale_fill_manual(values = c( "SPARKLING WATER" = "grey", "SSD" = "red", "ING ENHANCED WATER" = "grey", "COFFEE"
= "grey", "ENERGY" = "grey")) +
  theme(axis.text.x = element_text(angle = 60, hjust = 1)) +
  theme(axis.text.y = element_text(angle = 0, hjust = 1))
```




```
# SSD accounts for the vast majority of total sales in both units and dollars. This is likely to be interepreted
as standard soda, both diet and regular - the classics we know and love. Swire-CC generates most of their revenue
from this category. Though the "Energy Drink" category may prove to be useful to use in modeling scenarios, it may
prove more useful for our team to focus on SSD products when we consider how Swire-CC would launch innovation
products.

rm(cat_summary)
```

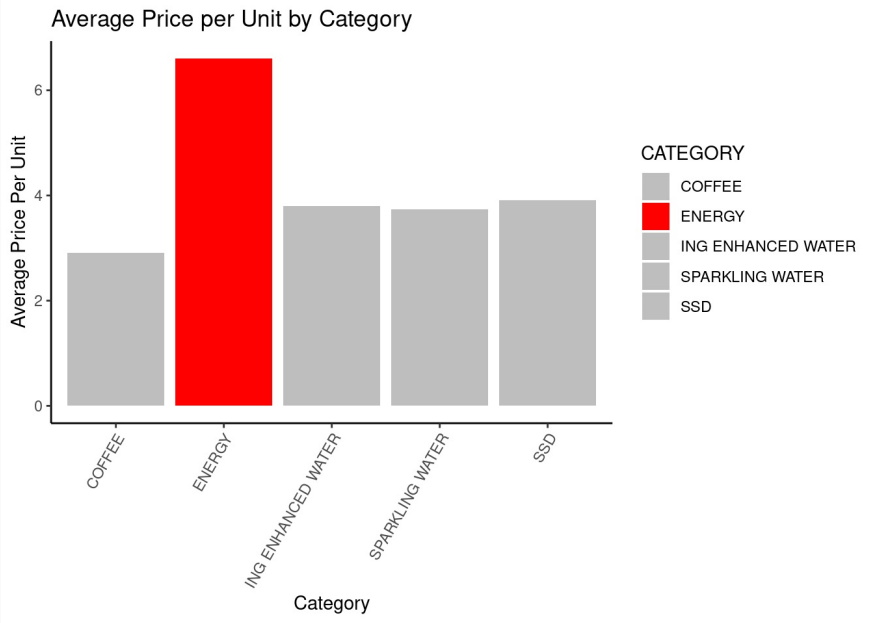
Takeaway 8: Energy Drinks have the highest Average Sale Price by Category

```
cat_up_summary <- swire_df %>%
  group_by(CATEGORY) %>%
  summarise(n = n(),
            avg_price_unit = mean(UNIT_PRICE),
            avg_sales_dollars = mean(DOLLAR_SALES),
            avg_unit_sales = mean(UNIT_SALES)
  ) %>%
  arrange(desc(avg_price_unit))
```

```
cat_up_summary
```

```
## # A tibble: 5 × 5
##   CATEGORY                 n avg_price_unit avg_sales_dollars avg_unit_sales
##   <fct>                <int>          <dbl>             <dbl>          <dbl>
## 1 ENERGY             5932087           6.60             627.           202.
## 2 SSD               12912281           3.91             744.           209.
## 3 ING ENHANCED WATER  2452456           3.80             202.            80.4
## 4 SPARKLING WATER     3019064           3.74             208.            55.0
## 5 COFFEE               145536           2.91              99.0           45.1
```

```
ggplot(cat_up_summary, aes(x = CATEGORY )) +
  geom_bar(aes(y = avg_price_unit, fill = CATEGORY), stat = "identity", position = "dodge") +
  labs(title = "Average Price per Unit by Category",
       x = "Category",
       y = "Average Price Per Unit") +
  theme_classic() +
  scale_fill_manual(values = c( "SPARKLING WATER" = "grey", "SSD" = "grey", "ING ENHANCED WATER" = "grey",
"COFFEE" = "grey", "ENERGY" = "red")) +
  theme(axis.text.x = element_text(angle = 60, hjust = 1)) +
  theme(axis.text.y = element_text(angle = 0, hjust = 1))
```




```
# Energy drinks sell at a much higher price per unit than the other categories. This means that assuming their
cost to produce is not much different from the other caloric segments, producing more energy drinks would be an
easy way to increase margins and could prove useful in comparing and contrasting innovation products against
current sellers in this emerging category. It is also no surprise that the typical energy drink costs 2x at most
locations than standard bottled soda products.

rm(cat_up_summary)
```

Takeaway 9: Initial Observations of Innovation Product Characteristics

```
# All Brand and Package combinations
swire_df %>%
  group_by(BRAND, PACKAGE) %>%
  summarise(MIN_DATE = min(DATE),
            MAX_DATE = max(DATE),
            DURATION = MAX_DATE - MIN_DATE) %>%
  arrange(desc(DURATION))
```

```
## `summarise()` has grouped output by 'BRAND'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 1,720 × 5
## # Groups:   BRAND [319]
##    BRAND                       PACKAGE            MIN_DATE   MAX_DATE   DURATION
##    <fct>                       <fct>              <date>     <date>     <drtn>
##  1 ABCS                        12SMALL 24ONE PLA… 2020-12-05 2023-10-28 1057 da…
##  2 ABCS                        12SMALL 4ONE PLAS… 2020-12-05 2023-10-28 1057 da…
##  3 ABCS                        12SMALL MLT PLAST… 2020-12-05 2023-10-28 1057 da…
##  4 ALL-OUT ALLIGATOR REAL-TIME 12SMALL 24ONE CUP  2020-12-05 2023-10-28 1057 da…
##  5 ANCESTOR PEPPY              12SMALL 12ONE CUP  2020-12-05 2023-10-28 1057 da…
##  6 ANCIENT ELIXIR              12SMALL 6ONE CUP   2020-12-05 2023-10-28 1057 da…
##  7 ANCIENT ELIXIR              2L MULTI JUG       2020-12-05 2023-10-28 1057 da…
##  8 AZURE HORIZON               12SMALL 6ONE CUP   2020-12-05 2023-10-28 1057 da…
##  9 BEAUTIFUL GREENER           .5L 6ONE JUG       2020-12-05 2023-10-28 1057 da…
## 10 BEAUTIFUL GREENER           12SMALL 12ONE CUP  2020-12-05 2023-10-28 1057 da…
## # i 1,710 more rows
```

```
# launch DATE, end DATE, and DURATION of brand or package less than 6 months, or 26 weeks, in duration
swire_df %>%
  group_by(BRAND, PACKAGE) %>%
  summarise(MIN_DATE = min(DATE),
            MAX_DATE = max(DATE),
            DURATION = MAX_DATE - MIN_DATE) %>%
  filter(DURATION < 180) %>%
  filter(DURATION > 30) %>%
  arrange(desc(DURATION))
```

```
## `summarise()` has grouped output by 'BRAND'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 95 × 5
## # Groups:   BRAND [78]
##    BRAND                        PACKAGE           MIN_DATE   MAX_DATE   DURATION
##    <fct>                        <fct>             <date>     <date>     <drtn>
##  1 CROWN                        ALL OTHER ONES    2023-05-06 2023-10-28 175 days
##  2 DIET BUBBLE JOY ADVANTAGEOUS 7.5SMALL 4ONE CUP 2023-05-06 2023-10-28 175 days
##  3 DIET ELF BUBBLES ZERO        7.5SMALL 4ONE CUP 2023-05-06 2023-10-28 175 days
##  4 ELF BUBBLES LYMONADE LEGACY  20SMALL MULTI JUG 2023-05-06 2023-10-28 175 days
##  5 FANTASMIC                    7.5SMALL 4ONE CUP 2023-05-06 2023-10-28 175 days
##  6 FRESH GRAPEFRUIT             7.5SMALL 4ONE CUP 2023-05-06 2023-10-28 175 days
##  7 KOOL! ZERO SUGAR             12SMALL 8ONE CUP  2022-04-16 2022-10-08 175 days
##  8 KOOL! ZERO SUGAR             7.5SMALL 4ONE CUP 2023-05-06 2023-10-28 175 days
##  9 SUPER-DUPER PUNCHED          12SMALL 6ONE CUP  2023-05-06 2023-10-28 175 days
## 10 ASTRAL BEVERAGE              .5L 12ONE JUG     2023-05-13 2023-10-28 168 days
## # i 85 more rows
```

```
# As an initial foray into determining the characteristics of "Innovation Products" some of the data surrounding
launch and packaging caught our attention. If we were to define innovation products as being in the market for at
least 30 days but less than 6 months, we see there are 95 different brand packaging combinations that fit this
criteria in our historic data set out of a total of 1720 Package and Brand Combinations. Does packaging matter for
an innovation product?
```

Takeaway 10: Potential Innovation Products - Brand & Packaging Combinations ~13 Weeks

```
swire_df %>%
  group_by(BRAND, PACKAGE) %>%
  summarise(n = n(),
            MIN_DATE = min(DATE),
            MAX_DATE = max(DATE),
            DURATION = MAX_DATE - MIN_DATE) %>%
  filter(DURATION < 100 ) %>%
  filter(DURATION > 81) %>%
  arrange(desc(DURATION))
```

```
## `summarise()` has grouped output by 'BRAND'. You can override using the
```

```
## `.groups` argument.
```

```
## # A tibble: 15 × 6
## # Groups:   BRAND [15]
##    BRAND                             PACKAGE     n MIN_DATE   MAX_DATE   DURATION
##    <fct>                             <fct>   <int> <date>     <date>     <drtn>
##  1 HILL MOISTURE ZERO SUGAR COASTA… 12SMAL…    12 2023-07-15 2023-10-21 98 days
##  2 KOOL! ZERO SUGAR MOON            7.5SMA…    11 2022-02-26 2022-06-04 98 days
##  3 PAPI ZERO SUGAR                  12SMAL…     6 2022-06-25 2022-10-01 98 days
##  4 RADIANT'S                        12SMAL…    13 2022-01-22 2022-04-30 98 days
##  5 DIET HANSENIZZLE'S               12SMAL…     4 2022-11-26 2023-02-25 91 days
##  6 HILL MOISTURE ZERO SUGAR         16SMAL…     4 2021-01-09 2021-04-10 91 days
##  7 DIET BUBBLE JOY ADVANTAGEOUS     12SMAL…     4 2022-10-08 2022-12-31 84 days
##  8 DIET PEPPY                       12SMAL…     3 2022-03-26 2022-06-18 84 days
##  9 ELF BUBBLES                      8.5SMA…     2 2021-06-12 2021-09-04 84 days
## 10 ELF BUBBLES SUMMER SPICED MIXED… 12SMAL…    18 2022-10-29 2023-01-21 84 days
## 11 KOOL! ZERO SUGAR SHREK           7.5SMA…     3 2022-08-20 2022-11-12 84 days
## 12 MYTHICAL BEVERAGE REHAB          16SMAL…    10 2020-12-12 2021-03-06 84 days
## 13 PAPI WILD CHERRY                 7.5SMA…   323 2023-08-05 2023-10-28 84 days
## 14 PEPPY ZERO SUGAR                 12SMAL…    16 2023-08-05 2023-10-28 84 days
## 15 PLEASURE & LIGHT                 12SMAL…     3 2021-07-17 2021-10-09 84 days
```

```
# 3 of the 7 questions Swire raised were related to 13 week launches. 13*7 = 91 days, ergo this table presents a glimpse at products on the market between 80-100 days and also released around the same time frames as Swire-CC anticipates releasing a handfull of Innovation Products.
```

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

EDA - PART 2: DEEPER DIVE INTO INNOVATION CHARACTERISTICS

```
# Team added a few more columns and features to the dataset for a deeper dive.

#season column based on date
#Months broken up into seasons for potential feature engineering uses, along with sales per unit.

df <- swire_df
rm(swire_df)
names(df)[names(df) == "UNIT_PRICE"] <- "SINGLE_PRICE"


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_
         ))
```

```
#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
# After feature data type conversion, the data is ready for exploratory data analysis.
str(df)
```

```
## 'data.frame':    24461424 obs. of  13 variables:
##  $ DATE           : Date, format: "2021-08-21" "2022-05-07" ...
##  $ MARKET_KEY     : Factor w/ 200 levels "1","6","7","13",..: 98 119 117 47 45 59 48 93 138 165 ...
##  $ CALORIC_SEGMENT: num  0 1 0 1 1 1 0 0 0 1 ...
##  $ CATEGORY       : Factor w/ 5 levels "COFFEE","ENERGY",..: 5 5 3 5 5 5 4 5 5 5 ...
##  $ 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   : Factor w/ 8 levels "ALLYS","BEARS",..: 8 3 4 3 3 8 4 4 4 4 ...
##  $ BRAND          : Factor w/ 319 levels "ABCS","ALL-OUT ALLIGATOR REAL-TIME",..: 93 131 96 33 257 262 11 164
241 151 ...
##  $ PACKAGE        : Factor w/ 103 levels ".5L 12ONE JUG",..: 14 14 69 62 14 79 38 4 14 4 ...
##  $ 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" ...
##  $ SINGLE_PRICE   : num  5.65 7.74 2.25 2.52 6.49 ...
##  $ MONTH          : int  8 5 10 8 1 11 3 11 7 4 ...
##  $ SEASON         : Factor w/ 4 levels "FALL","SPRING",..: 3 2 1 3 4 1 2 1 3 2 ...
```

Exploratory Visualizations, Summary Tables & Essential Takeaways

Swire Innovation Brands Summary

```
#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
```

> Comments: 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).

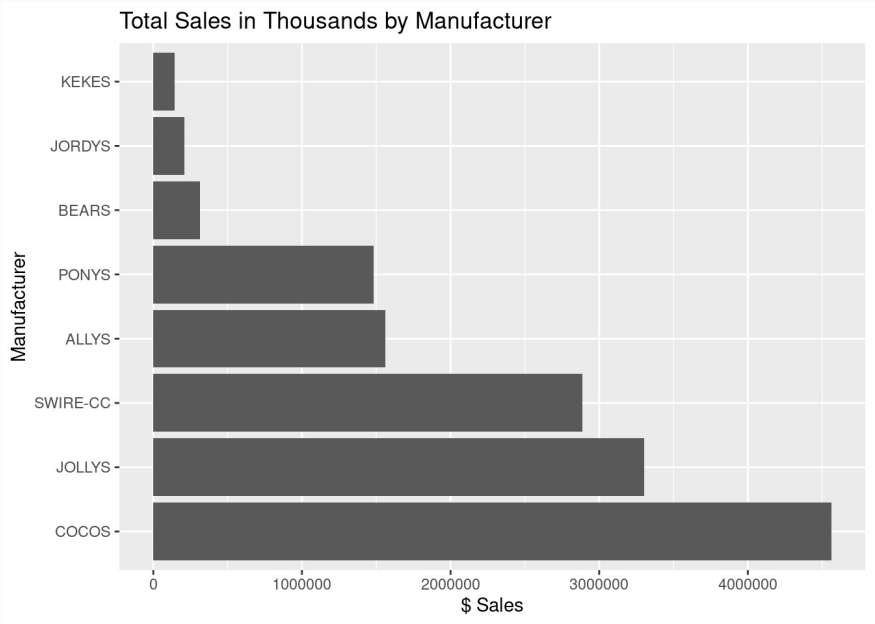
Swire Innovation Brands Sales

```
#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")
```

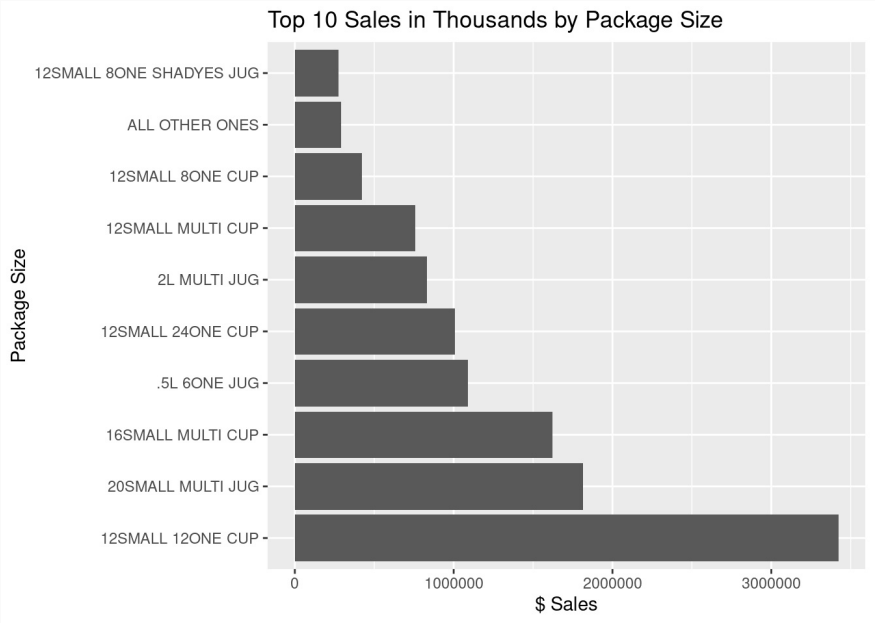



```
#sales in thousands by top 10 package size
df %>%
  group_by(PACKAGE) %>%
  summarise(TOTAL_SALES = sum(DOLLAR_SALES/1000)) %>%
  arrange(desc(TOTAL_SALES)) %>%
  head(10)
```

```
## # 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
```

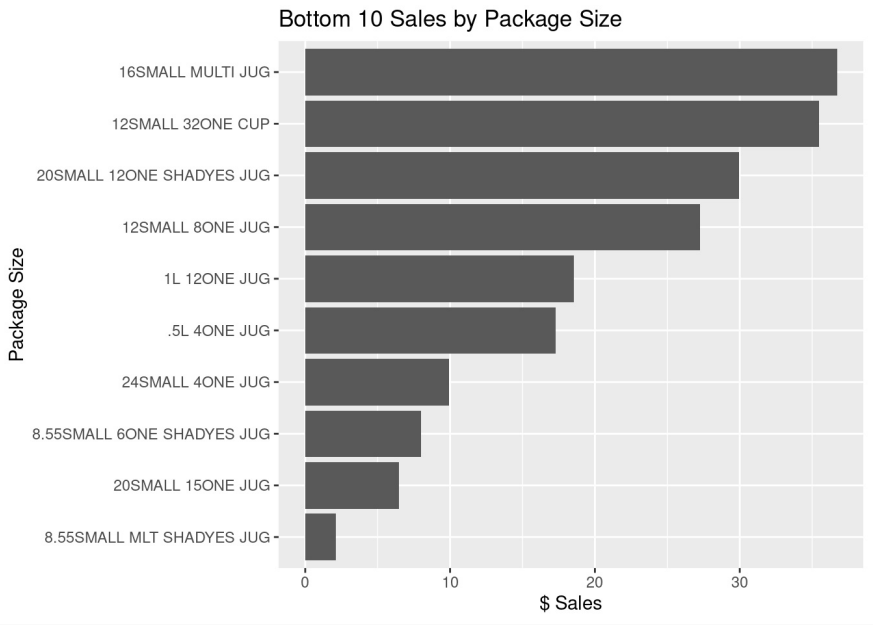
```
df %>%
  group_by(PACKAGE) %>%
  summarise(TOTAL_SALES = sum(DOLLAR_SALES/1000)) %>%
  arrange(desc(TOTAL_SALES)) %>%
  head(10) %>%
  ggplot(aes(x = reorder(PACKAGE, -TOTAL_SALES), y = TOTAL_SALES)) +
  geom_col() +
  coord_flip() +
  labs(title = "Top 10 Sales in Thousands by Package Size",
  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

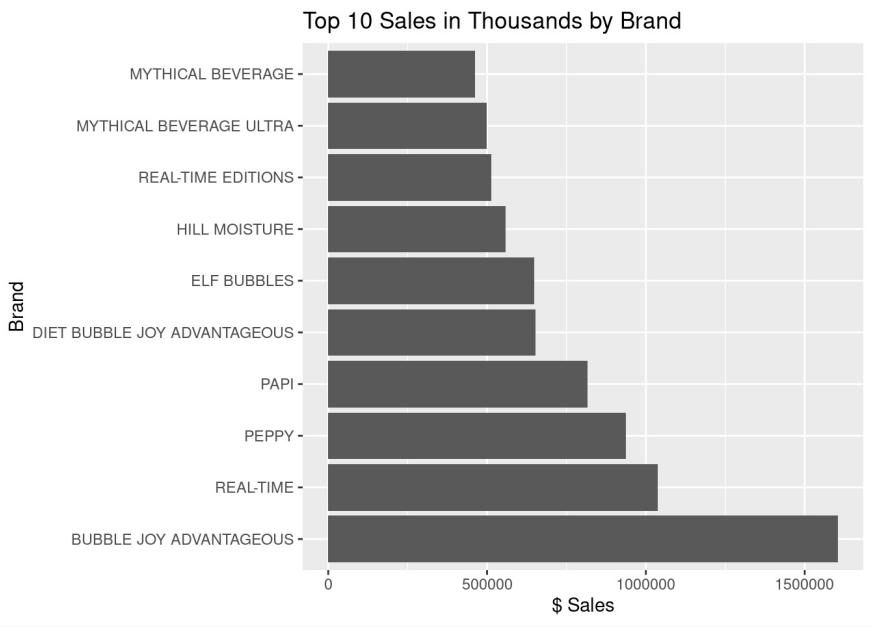
```
#graph sales in thousands by bottom 10 package size
df %>%
  group_by(PACKAGE) %>%
  summarise(TOTAL_SALES = sum(DOLLAR_SALES)) %>%
  arrange(TOTAL_SALES) %>%
  head(10) %>%
  ggplot(aes(x = reorder(PACKAGE, TOTAL_SALES), y = TOTAL_SALES)) +
  geom_col() +
  coord_flip() +
  labs(title = "Bottom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

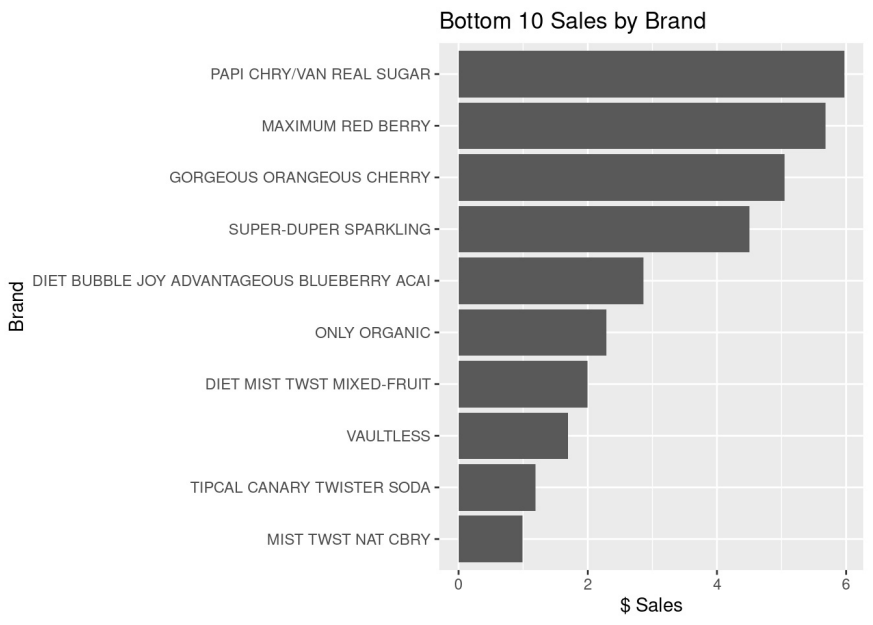
```
#graph sales in thousands by top 10 brand
df %>%
  group_by(BRAND) %>%
  summarise(TOTAL_SALES = sum(DOLLAR_SALES/1000)) %>%
  arrange(desc(TOTAL_SALES)) %>%
  head(10) %>%
  ggplot(aes(x = reorder(BRAND, -TOTAL_SALES), y = TOTAL_SALES)) +
  geom_col() +
  coord_flip() +
  labs(title = "Top 10 Sales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

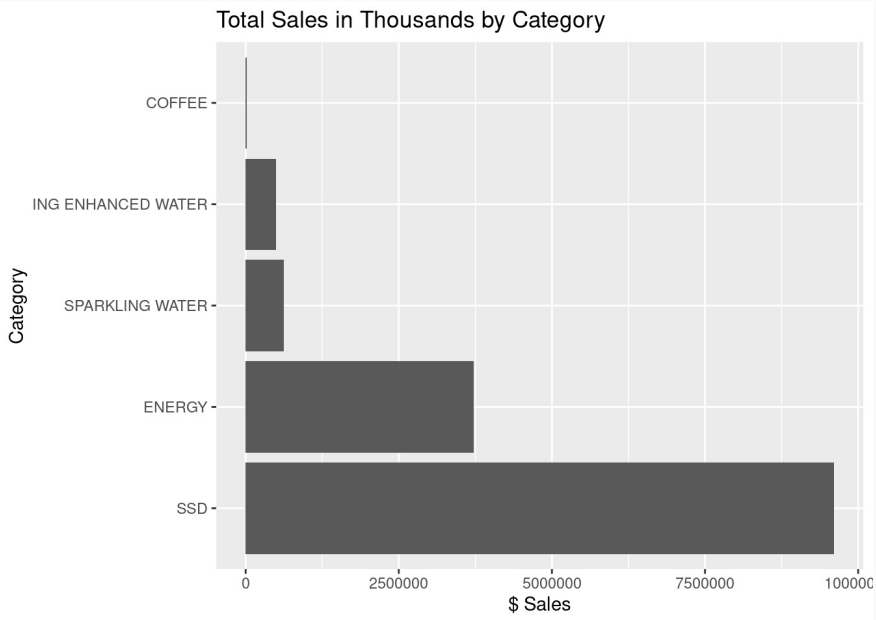
```
#graph sales in thousands by bottom 10 brand
df %>%
  group_by(BRAND) %>%
  summarise(TOTAL_SALES = sum(DOLLAR_SALES)) %>%
  arrange(TOTAL_SALES) %>%
  head(10) %>%
  ggplot(aes(x = reorder(BRAND, TOTAL_SALES), y = TOTAL_SALES)) +
  geom_col() +
  coord_flip() +
  labs(title = "Bottom 10 Sales by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

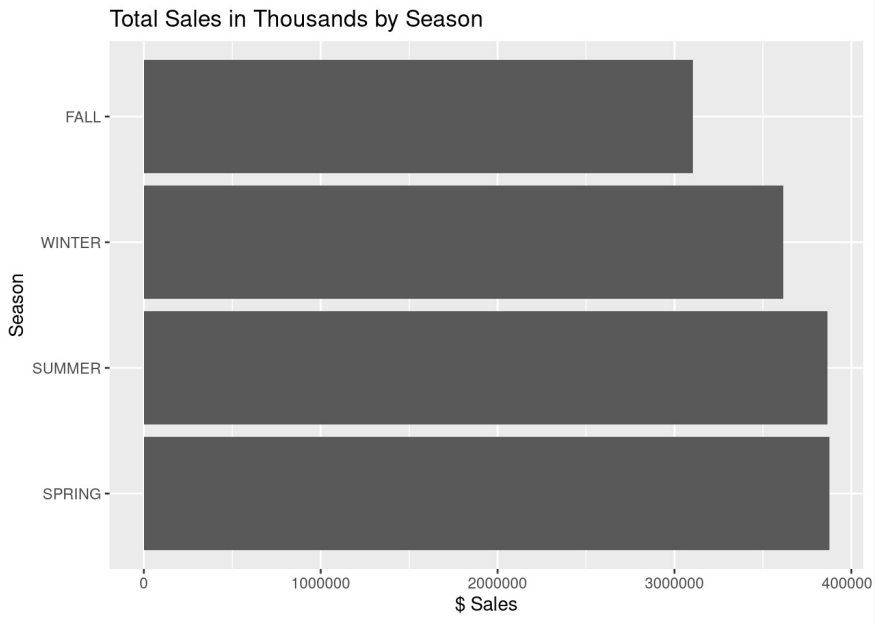
```
#graph sales in thousands by CATEGORY
df %>%
  group_by(CATEGORY) %>%
  summarise(TOTAL_SALES = sum(DOLLAR_SALES/1000)) %>%
  ggplot(aes(x = reorder(CATEGORY, -TOTAL_SALES), y = TOTAL_SALES)) +
  geom_col() +
  coord_flip() +
  labs(title = "Total Sales in Thousands by Category",
       x = "Category",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



COMMENTS: 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.

Product Observations

```
#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))
```

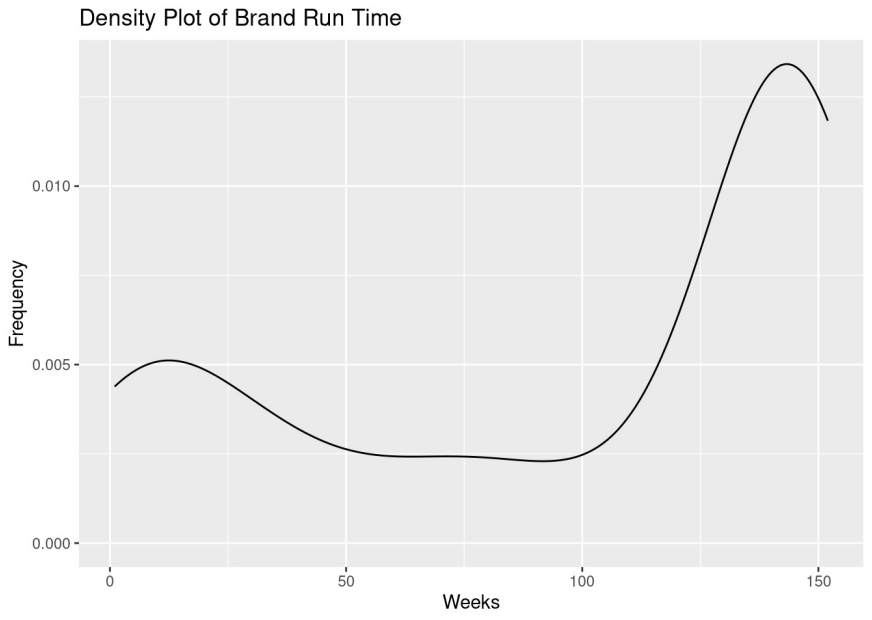
```
## # A tibble: 1 × 1
##   `median(LENGTH)`
##              <int>
## 1              137
```

```
#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
```

> COMMENT: 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.

Packaging Details

```
#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))
```

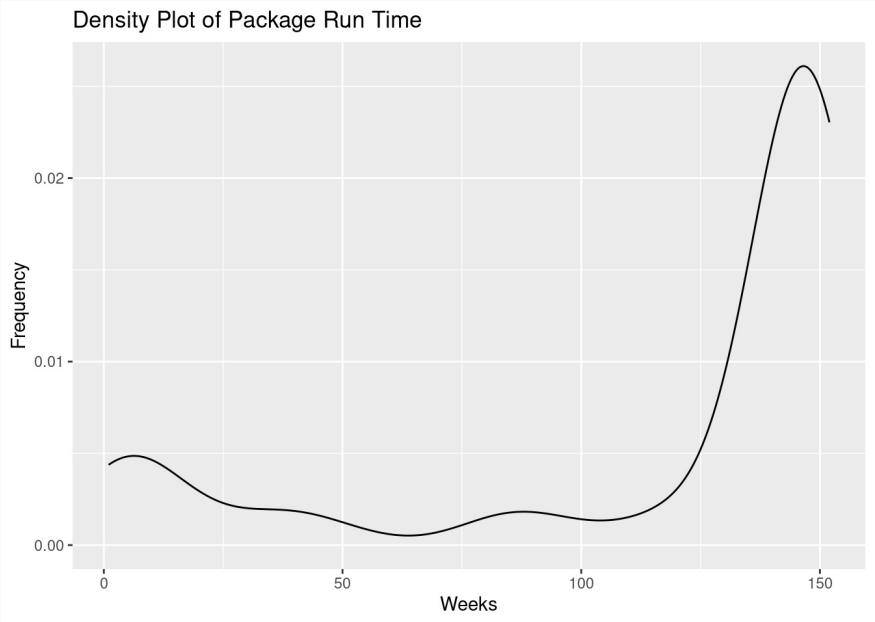
```
## # A tibble: 1 × 1
##   `median(LENGTH)`
##              <int>
```

```
## 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
```

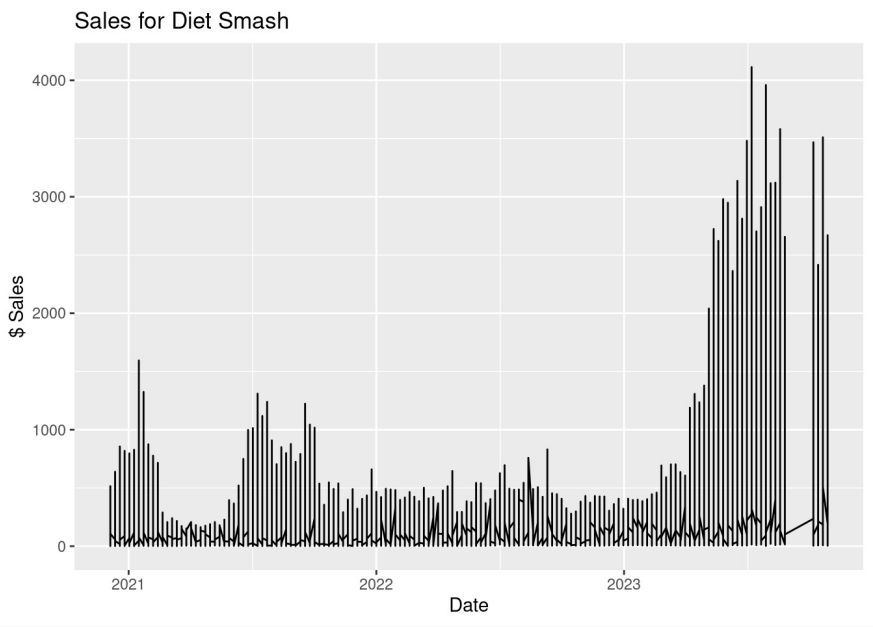
```
#summarize features of packages that run for less than 6 months
df %>%
  group_by(PACKAGE) %>%
  summarise(LENGTH = n_distinct(DATE)) %>%
  filter(LENGTH < 6) %>%
  left_join(df, by = "PACKAGE") %>%
  select(PACKAGE, CATEGORY, SEASON, BRAND, MANUFACTURER) %>%
  distinct()
```

```
## # 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
```

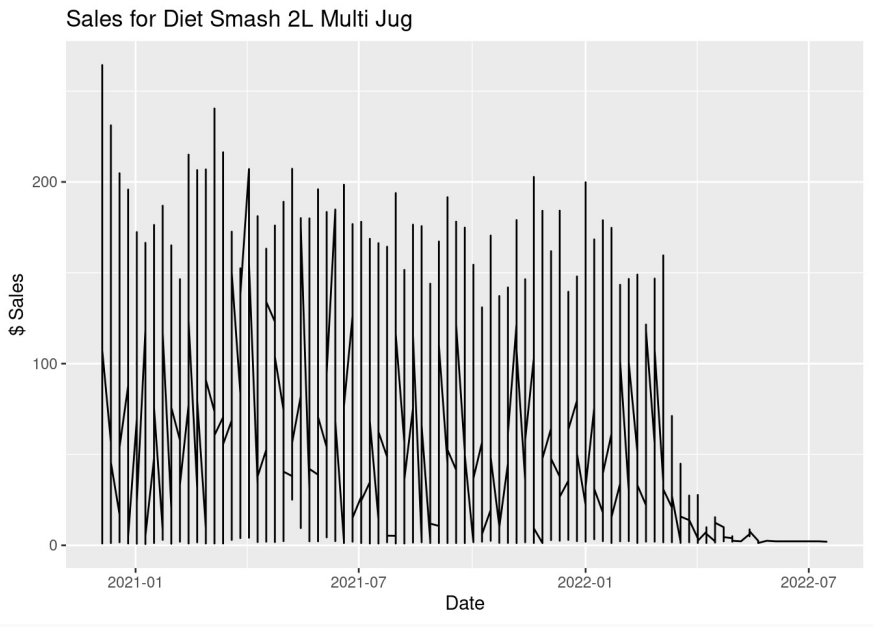
> COMMENT: Top 10 packages run the entire length of the dataset, which is 152 weeks. These are tried and true packaging sizes that we have all likely grown up with, know and love to enjoy our beverages from, whether from a gas station or grocery store. The top 10 shortest packages range from 1 week to 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.

Innovation Characteristics

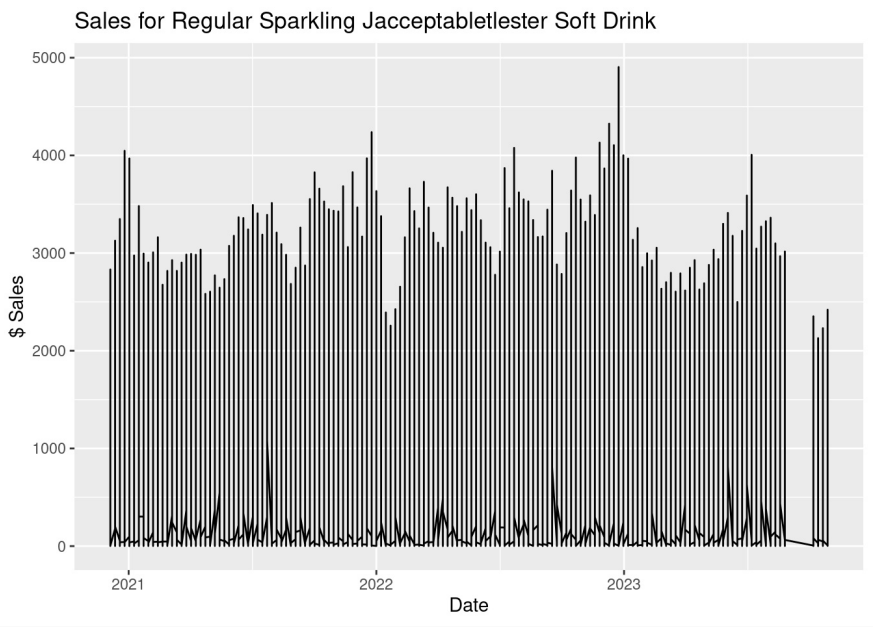
```
#graph DOLLAR_SALES by DATE for BRAND == "DIET SMASH"
df %>%
  filter(BRAND == "DIET SMASH") %>%
  ggplot(aes(x = DATE, y = DOLLAR_SALES)) +
  geom_line() +
  labs(title = "Sales for Diet Smash",
       x = "Date",
       y = "$ Sales")
```



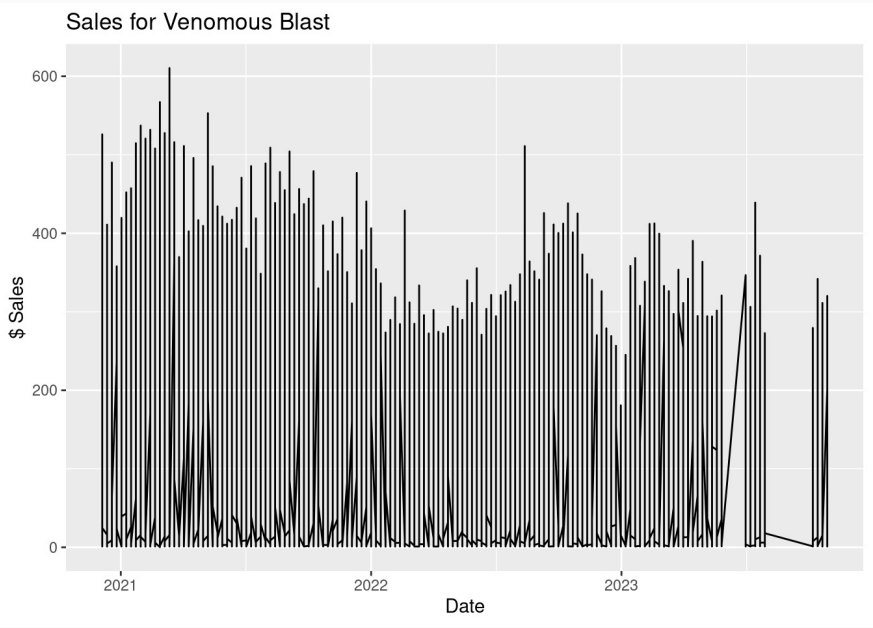

```
#graph DOLLAR_SALES by DATE for BRAND == "DIET SMASH" - INNOVATION PACKAGE == "2L MULTI JUG"
df %>%
  filter(BRAND == "DIET SMASH", PACKAGE == "2L MULTI JUG") %>%
  ggplot(aes(x = DATE, y = DOLLAR_SALES)) +
  geom_line() +
  labs(title = "Sales for Diet Smash 2L Multi Jug",
       x = "Date",
  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



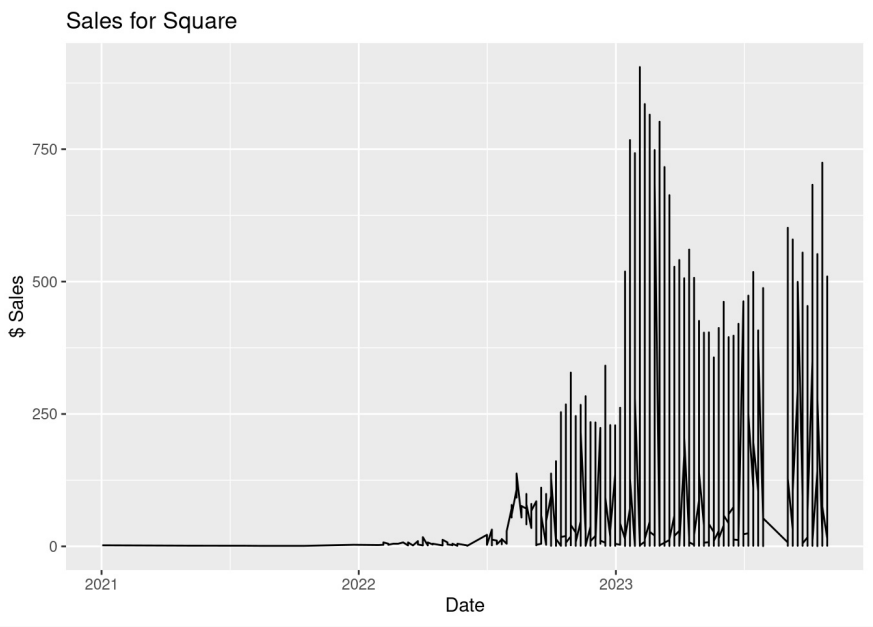

```
#graph DOLLAR_SALES by DATE for BRAND == "SPARKLING JACCEPTABLETLESTER",
#CATEGORY == "SSD", CALORIC_SEGMENT == REGULAR
df %>%
  filter(BRAND == "SPARKLING JACCEPTABLETLESTER", CATEGORY == "SSD",
         CALORIC_SEGMENT == "1" ) %>%
  ggplot(aes(x = DATE, y = DOLLAR_SALES)) +
  geom_line() +
  labs(title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



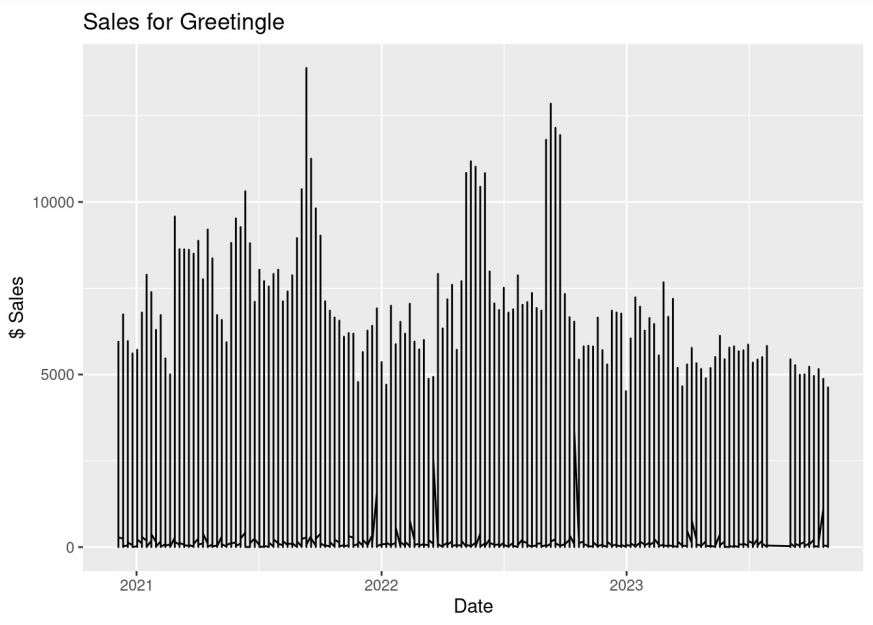

```
#graph DOLLAR_SALES by DATE for BRAND == "VENOMOUS BLAST" and CATEGORY == DIET/LIGHT
df %>%
  filter(BRAND == "VENOMOUS BLAST", CALORIC_SEGMENT == 0) %>%
  ggplot(aes(x = DATE, y = DOLLAR_SALES)) +
  geom_line() +
  labs(title = "Sales for Venomous Blast",
       x = "Date",
       y = "$ Sales")
```



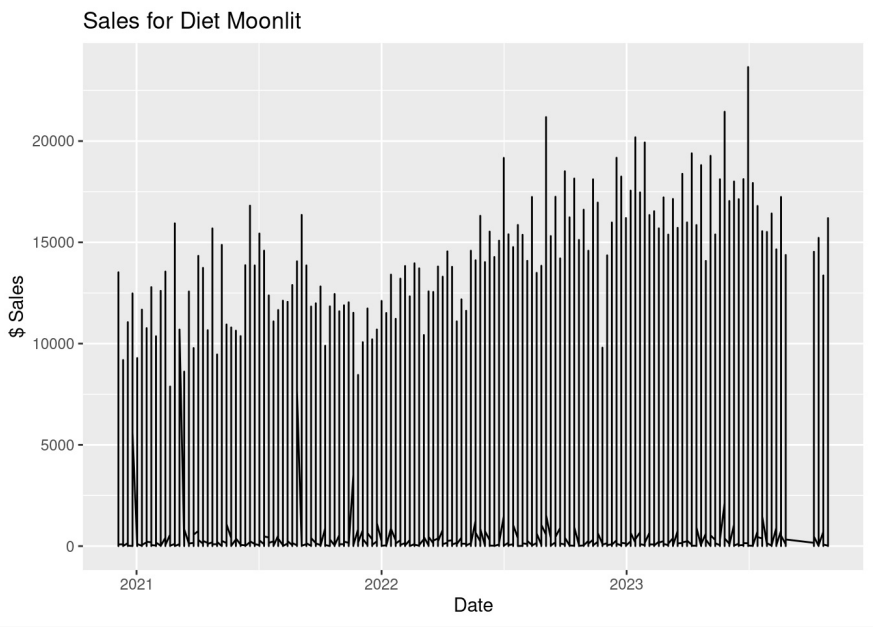

```
#graph DOLLAR_SALES by DATE for BRAND == "SQUARE"
df %>%
  filter(BRAND == "SQUARE") %>%
  ggplot(aes(x = DATE, y = DOLLAR_SALES)) +
  geom_line() +
  labs(title = "Sales for Square",
       x = "Date",
       y = "$ Sales")
```



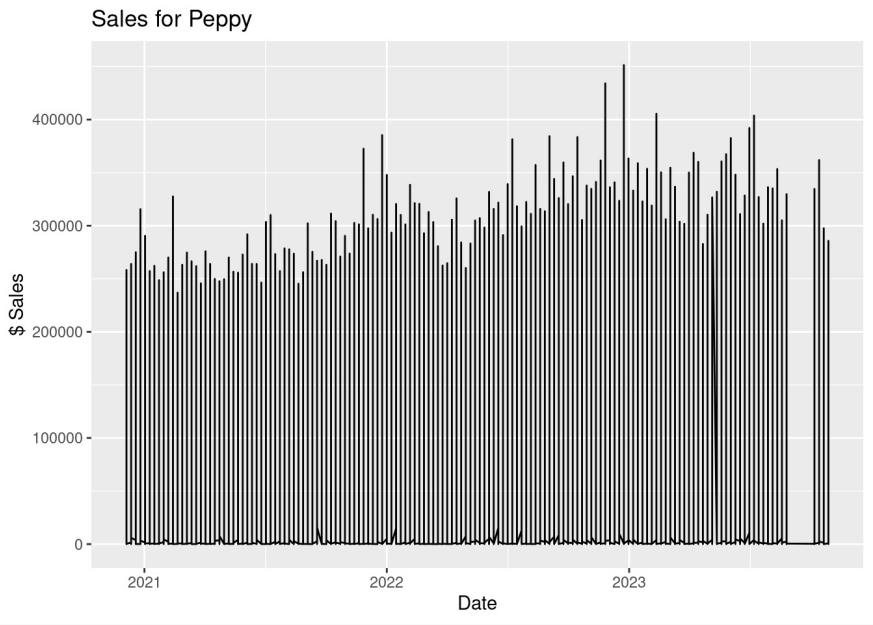

```
#graph DOLLAR_SALES by DATE for BRAND == "GREETINGLE"
df %>%
  filter(BRAND == "GREETINGLE") %>%
  ggplot(aes(x = DATE, y = DOLLAR_SALES)) +
  geom_line() +
  labs(title = "Sales for Greetingle",
       x = "Date",
       y = "$ Sales")
```




```
#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")
```



> COMMENT: 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.

Missing Date Analysis

```
#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
```

> COMMENT: 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. The question is whether or not this constitutes missing data OR product taken off market and put back into the market based on supply, supply chain, or other business issues.

Innovation Focus

```
#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))
```

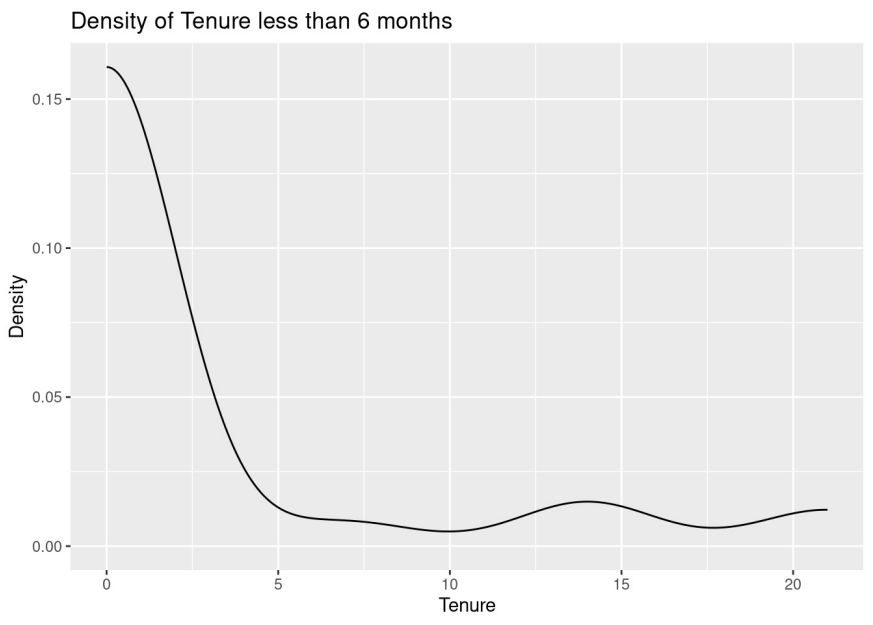
```
## # A tibble: 145 × 5
## # Groups:   BRAND [102]
##    BRAND                                PACKAGE MIN_DATE   MAX_DATE   TENURE
##    <fct>                                <fct>   <date>     <date>     <drtn>
##  1 BARS                                 2L MUL… 2022-10-29 2022-11-19 21 da…
##  2 BUBBLE 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.

EDA - PART 3: EVEN DEEPER - MODELING EDA

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)
```

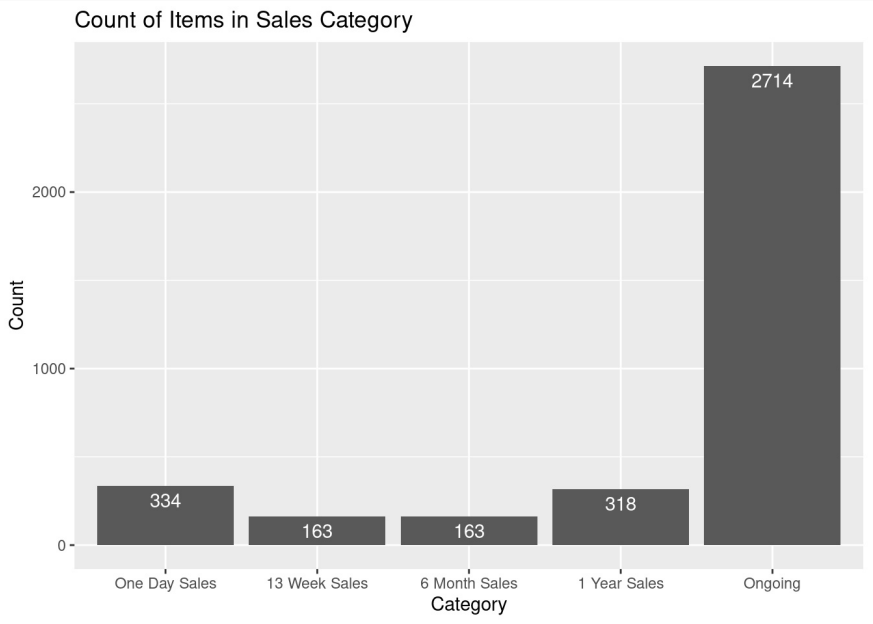
```
##                                                       ITEM
## 1 BUBBLE JOY WATER-JUGD-CARBONATED CONTAINER 288 LIQUID SMALL
## 2 BUBBLE JOY WATER-JUGD-CARBONATED CONTAINER 288 LIQUID SMALL
## 3  CUPADA ARID GENTLE DRINK GINGER ALE JUG 10 LIQUID SMALL X6
## 4  CUPADA ARID GENTLE DRINK GINGER ALE 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.
# 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 smaller data sets for modeling we will want to assure that we filter by these 3 categories.
```

Breaking out Items by Tenure

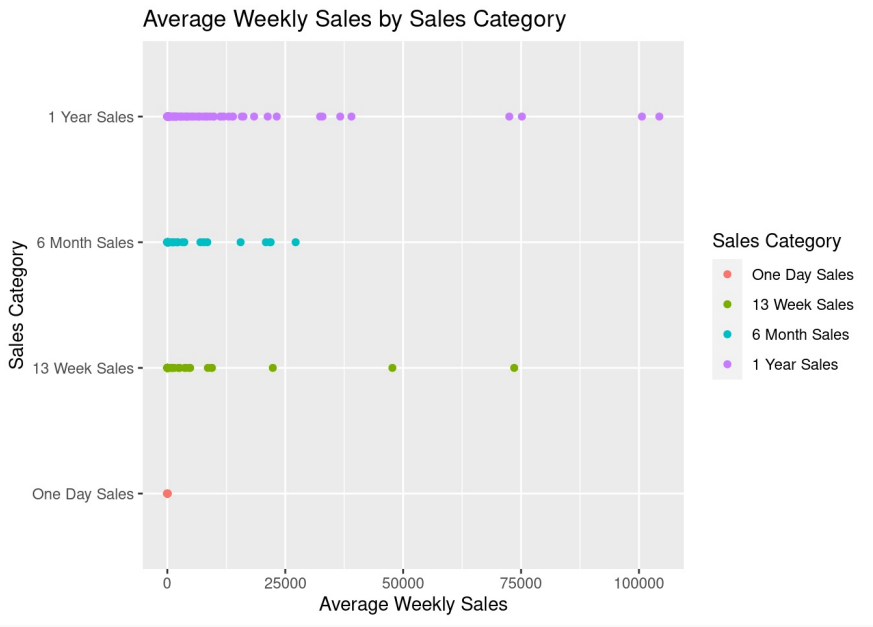
```
#Create table to summarize total sale days

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"))

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

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")
```



Sales Groups

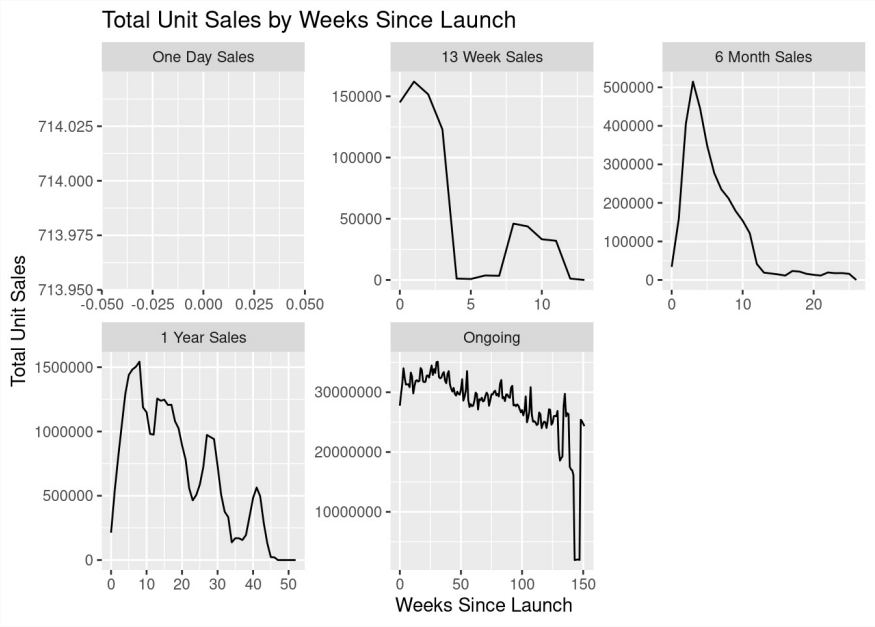
```
#Create plot of sales category with average weekly sales

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 outliers when creating our modeling sets we will want to do more outlier analysis on each group to see how to address each one.
```

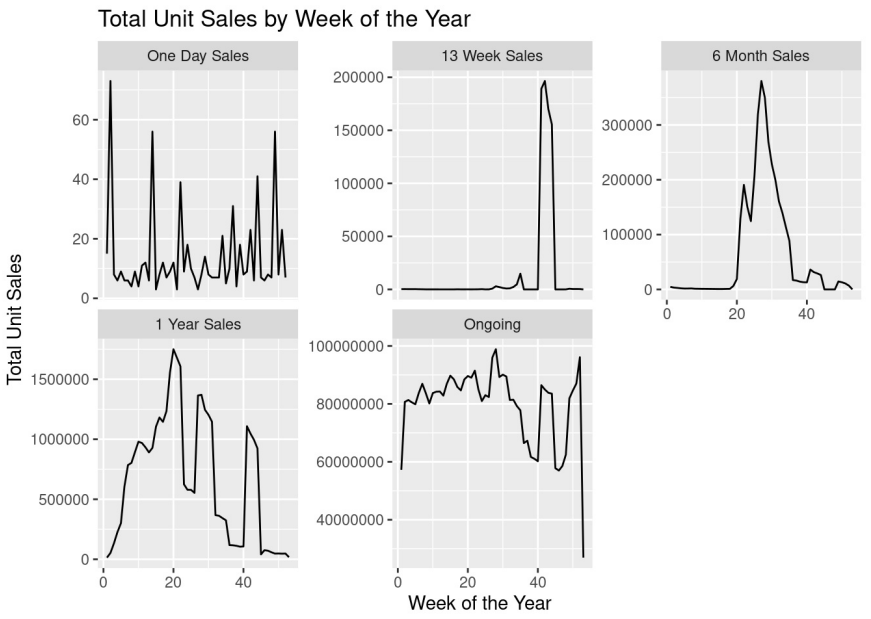
```
df <- left_join(df,sales_summary %>% select(ITEM, sales_category, duration_days, duration_weeks, total_sale_days, first_date, launch13week_date, launch6month_date, launch1year_date), by = "ITEM")

# 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")
```



> COMMENT: In these 5 line graphs we are given the shape of sales from launch date. We see that with the short term products a large spike starts and tapers off as time goes on. “Ongoing” products tend to 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.

```
# 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 outliers in the 13 week and 1 year category.

# 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")
```



Swire Directed Questions:

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))
```

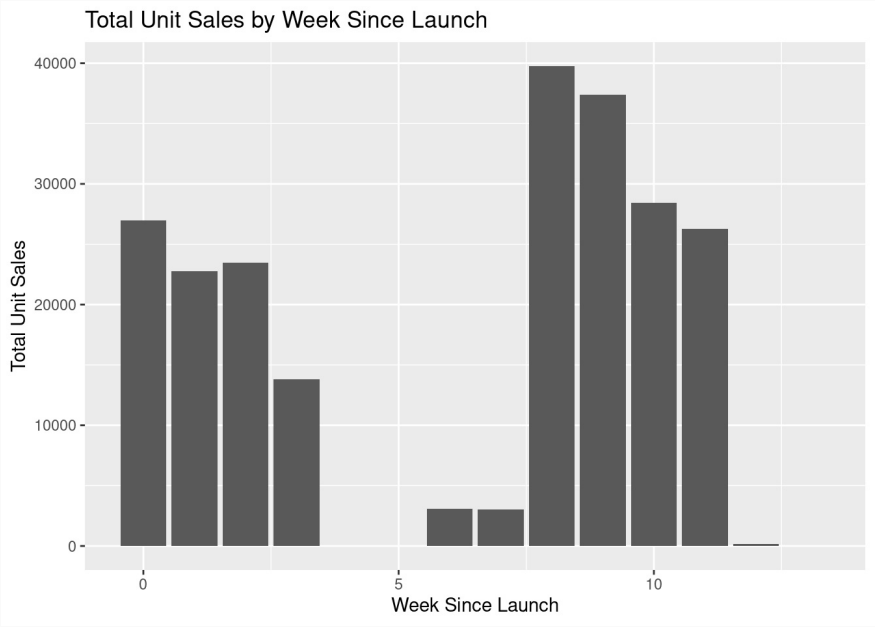
```
## # A tibble: 37 × 2
##    ITEM                                                        distinct_items
##    <chr>                                                                <int>
##  1 AZURE HORIZON FREE GENTLE DRINK SUPER-JUICE  DURIAN  CALORIE …            1
##  2 CAFFEINE FREE DIET PAPI GENTLE 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))
```

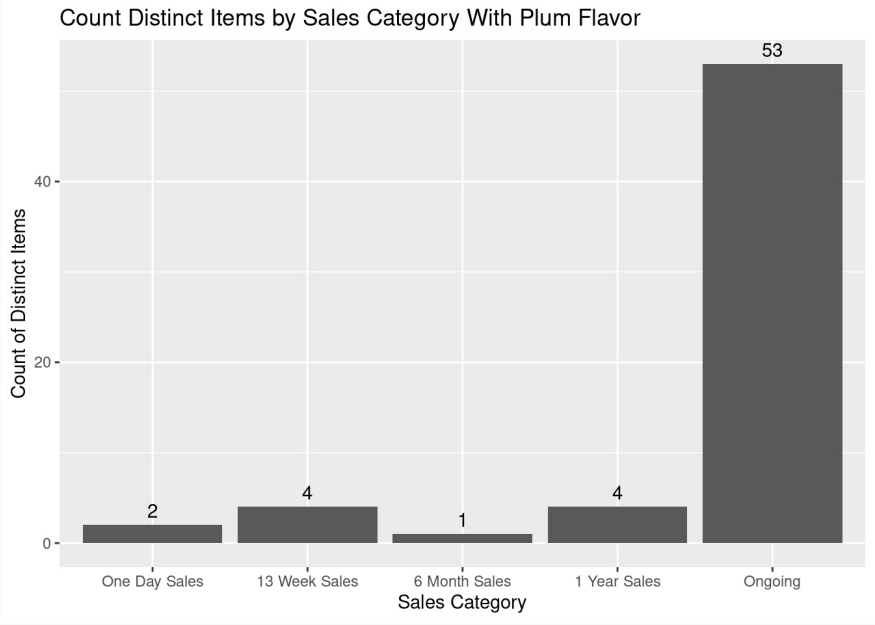
```
##   distinct_items
## 1             64
```

```
#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?

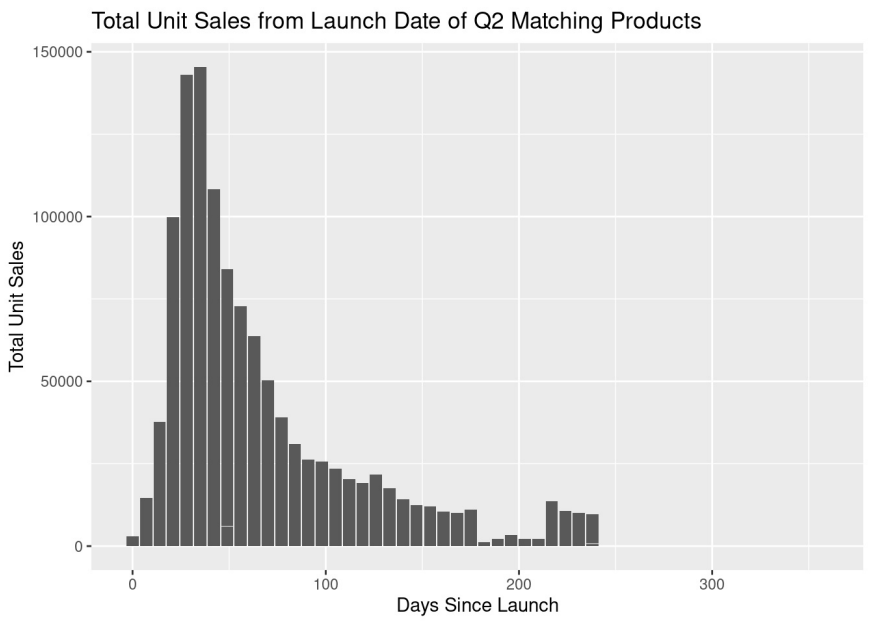
```
# Matching parameters for Q2
df %>%
  filter(
         month(first_date) %in% c(3, 4),
         CALORIC_SEGMENT == 1,
         CATEGORY == "SSD") %>%
         #str_detect(ITEM, "AVOCADO") %>%
  summarize(distinct_items = n_distinct(ITEM))
```

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

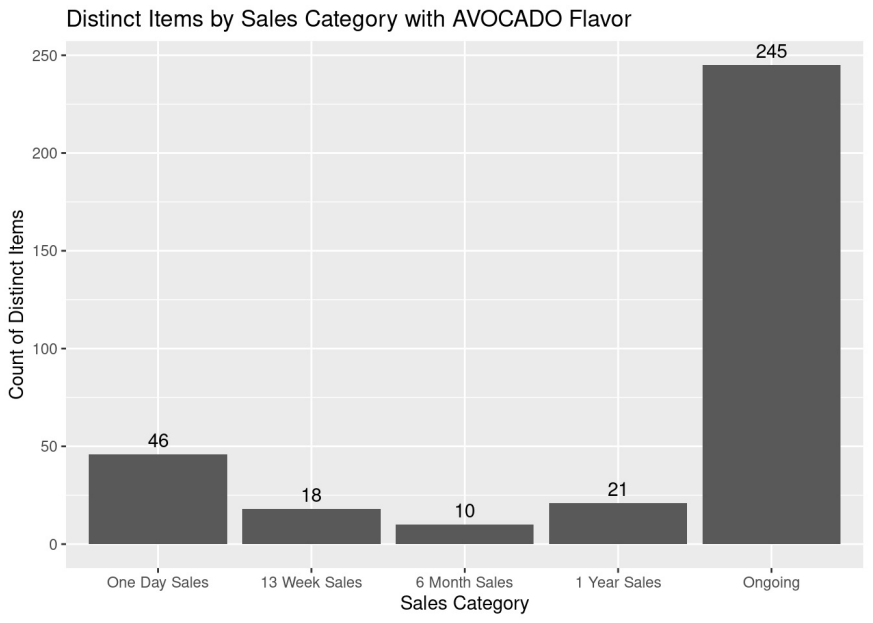
```
#148 items match launching in either March or April, regular and SSD category. There are non with Avocado in this
group

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",
```

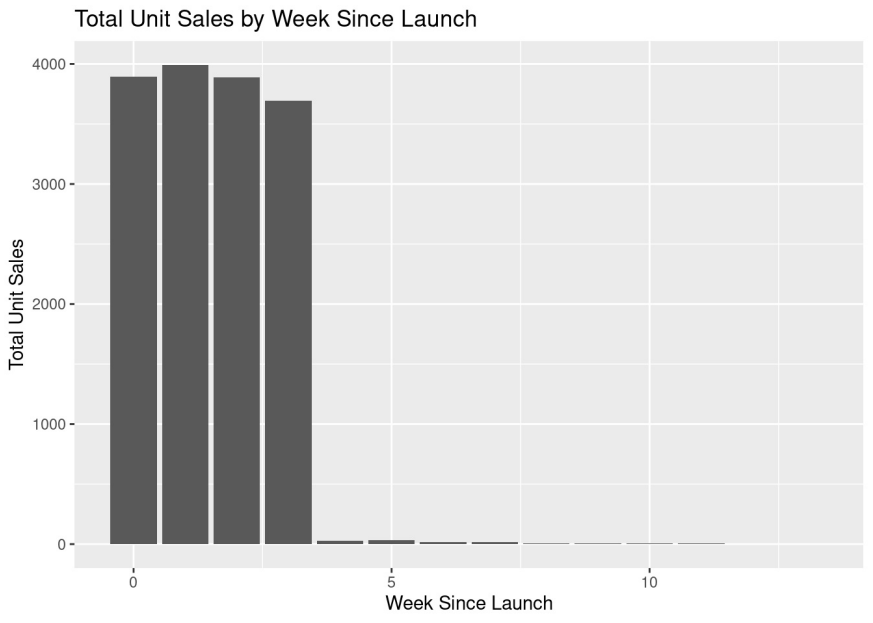
```
    CALORIC_SEGMENT == 0,
    CATEGORY == "ENERGY",
    #BRAND == "VENOMOUS BLAST",
    #str_detect(ITEM, "KIWANO")
  ) %>%
  group_by(ITEM) %>%
  summarize(distinct_items = n_distinct(ITEM))
```

```
## # 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
```

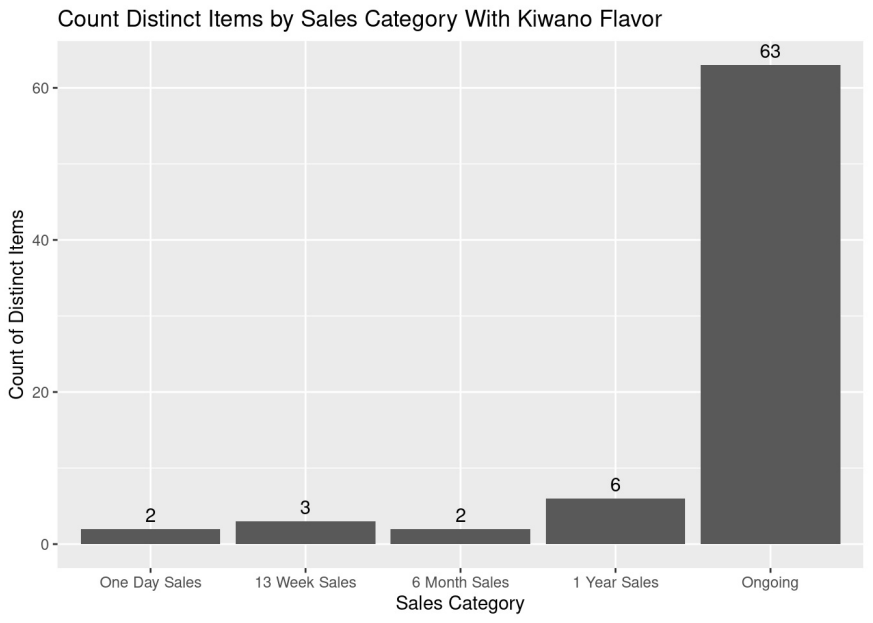
```
# There are only 16  items that match time period caloric segment, category. flavor, packaging and brand dont
exist

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",
```

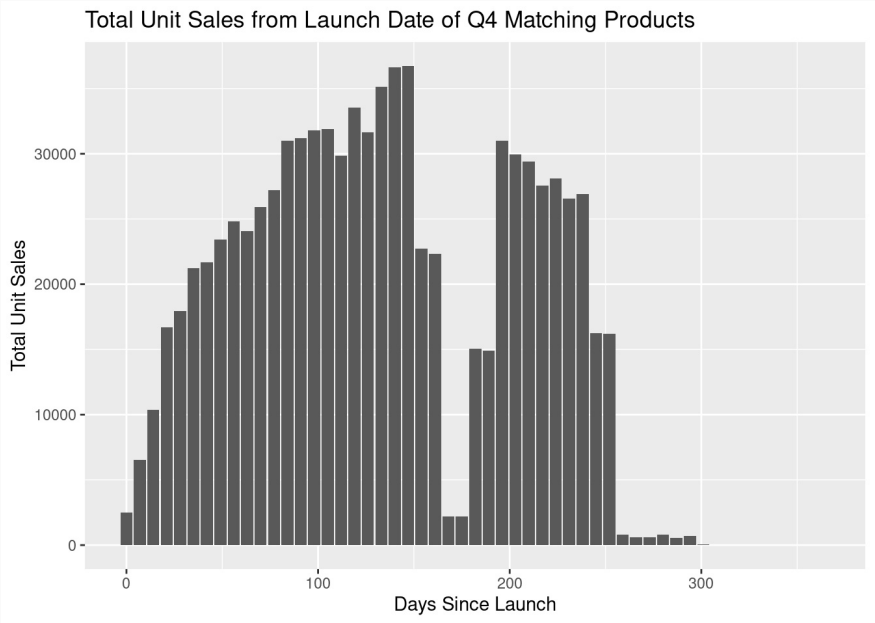
```
        CALORIC_SEGMENT == 0,
        CATEGORY == "SPARKLING WATER") %>%
        #str_detect(ITEM, "MULBERRIES")) %>%
  summarize(distinct_items = n_distinct(ITEM))
```

```
##   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))
```

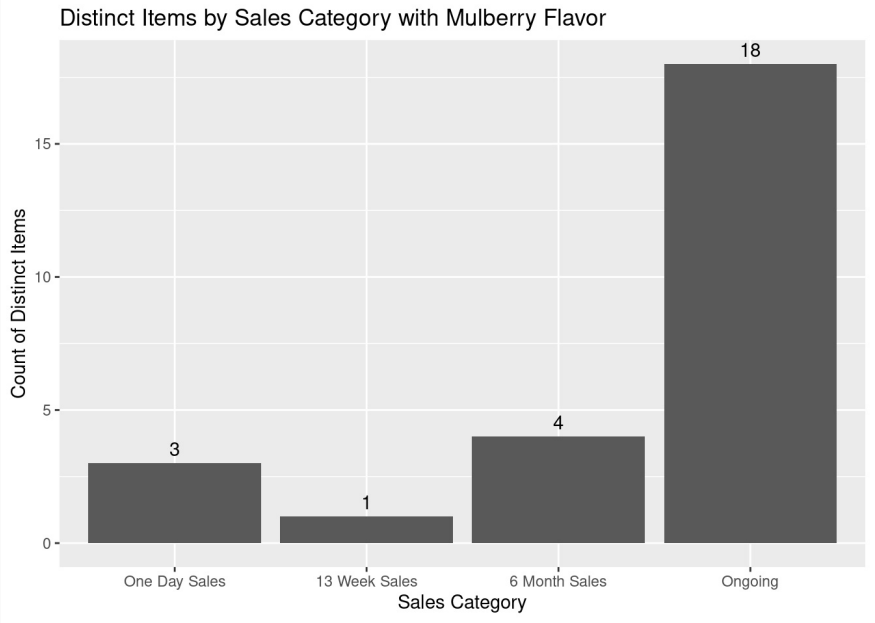
```
##   distinct_items
## 1             26
```

```
#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 northern region when setting up modeling. Also, of interest 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))
```

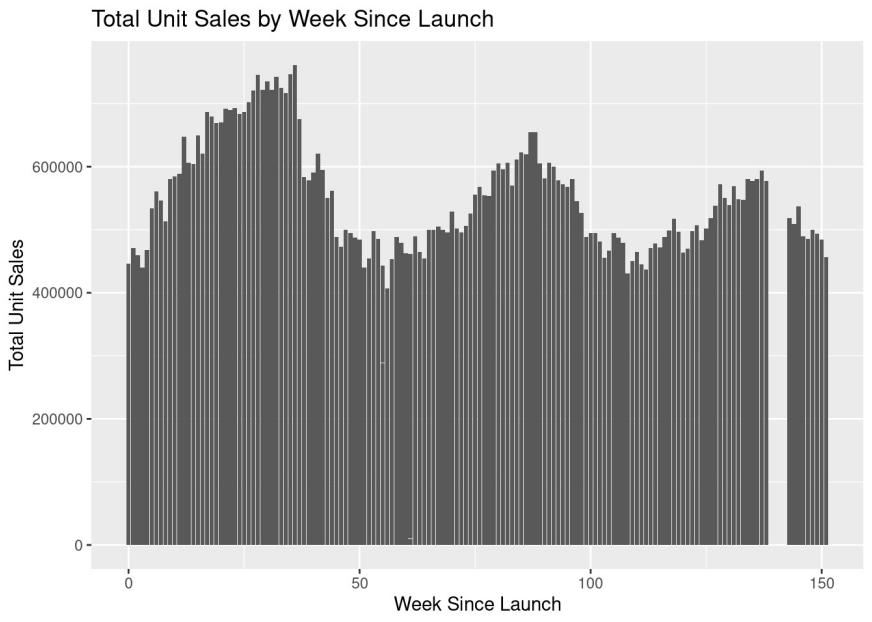
```
## # A tibble: 55 × 2
##    ITEM                                                    distinct_items
##    <chr>                                                            <int>
##  1 MYTHICAL BEVERAGE HYDRO ENERGY DRINK BLUE ICE JUG 25.4 LIQUID…              1
##  2 MYTHICAL BEVERAGE HYDRO ENERGY 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))
```

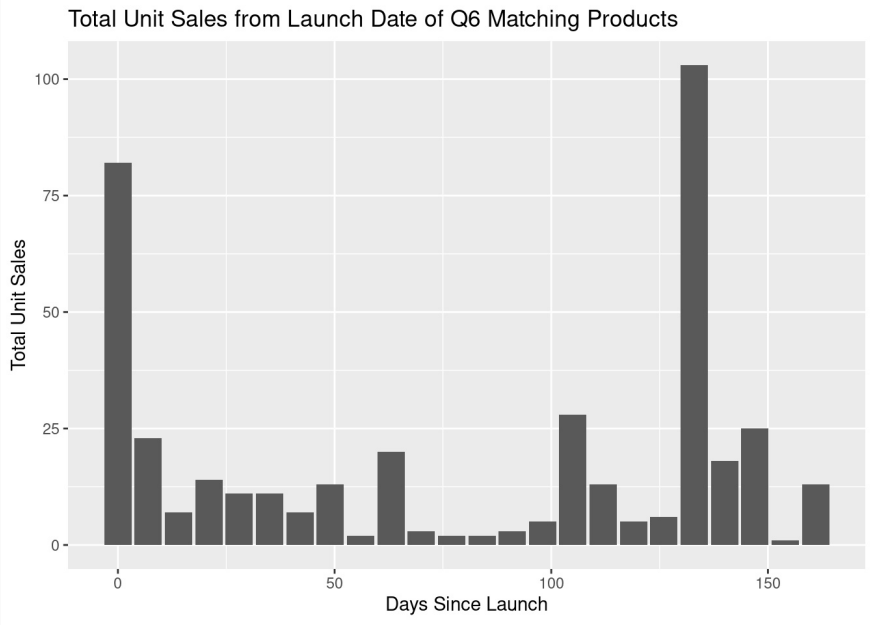
```
##   distinct_items
## 1             10
```

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

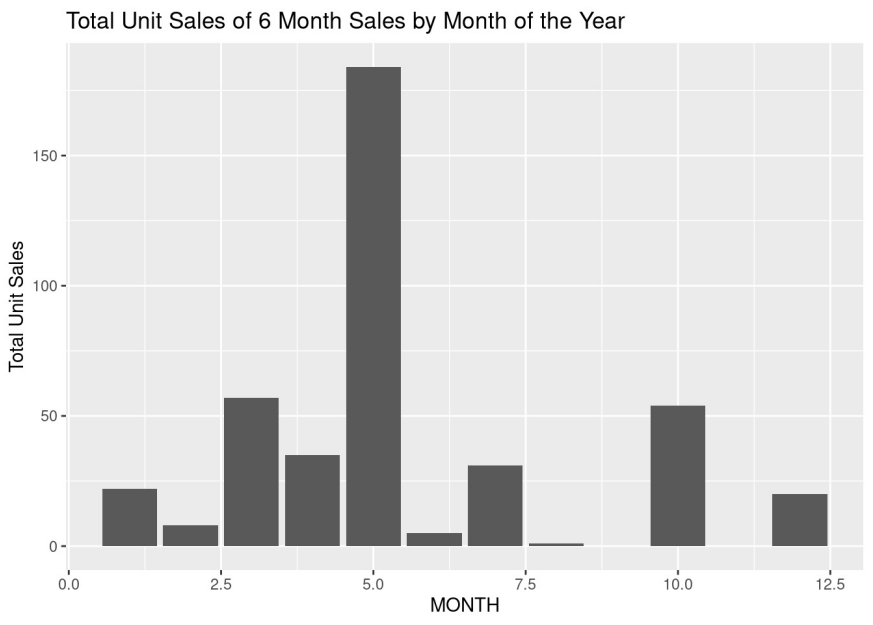
```
df %>%
  filter(str_detect(ITEM, "CASSAVA")) %>%
  summarize(distinct_items = n_distinct(ITEM))
```

```
##   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))
```

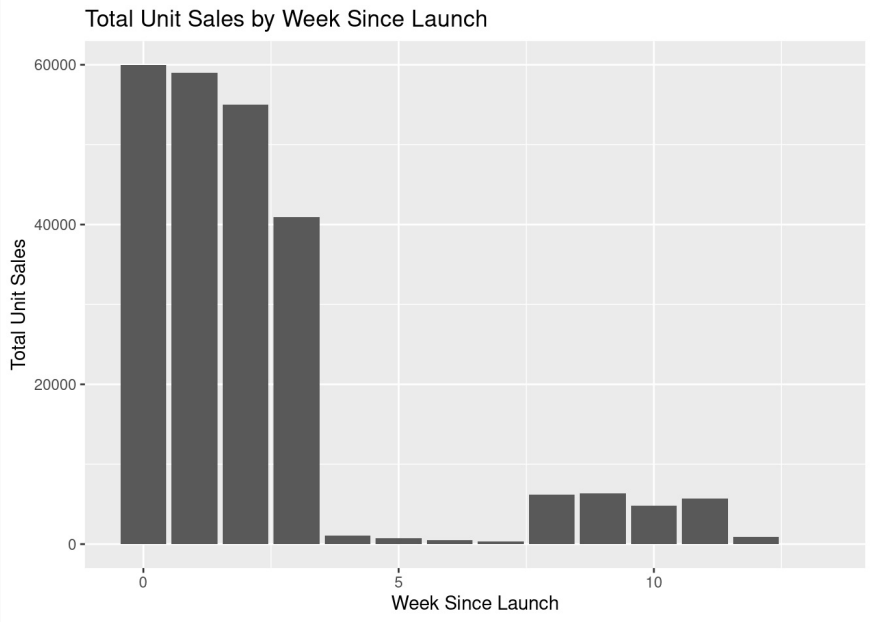
```
## # A tibble: 62 × 2
##    ITEM                                                        distinct_items
##    <chr>                                                                <int>
##  1 AZURE HORIZON GENTLE DRINK WILD PINK  CUP 12 LIQUID SMALL                1
##  2 BARS GENTLE DRINK PINA JUG 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.

```
rm(categories_count)
rm(sales_summary)
rm(df)
```

EDA - PART 4: DEMOGRAPHIC DATA

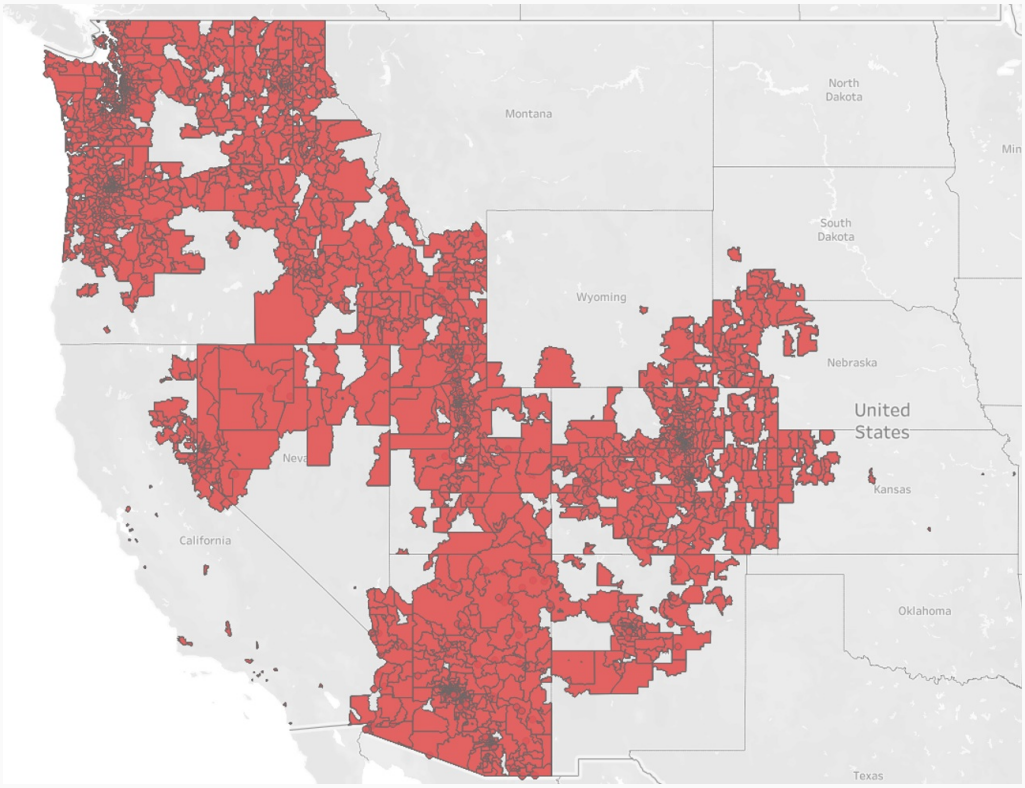
Quick View of the Demographic Data

Summary Statistics

State	Total Market Pop	Household Count	# of Market Segments	# of Zip Codes	# of Adult Males	# of Adult Females
WA	5,626,119	2,896,229	33	477	2,779,456	2,846,661
AZ	5,411,596	2,670,174	56	352	2,637,239	2,774,348
CO	4,351,983	2,259,753	35	395	2,160,955	2,191,051
OR	2,934,725	1,507,859	23	281	1,434,336	1,500,381
UT	2,209,746	1,004,513	14	211	1,098,924	1,110,831
ID	1,385,341	699,690	12	221	684,024	701,329

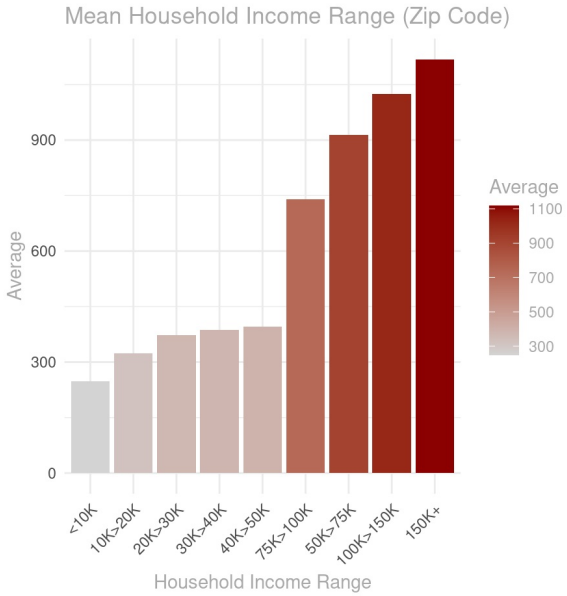
NM	864,766	444,419	13	80	417,194	447,574
CA	846,163	357,260	28	61	417,686	428,482
NV	618,654	311,145	10	70	311,641	307,009
WY	162,927	90,222	4	23	82,158	80,771
KS	66,474	37,490	6	22	32,329	34,147
NE	59,441	33,767	2	29	29,097	30,345
SD	40,253	20,262	1	10	19,711	20,539
Total	**24,578,188**	**12,332,783**	**200**	**2,232**	**12,104,750**	**12,473,468**

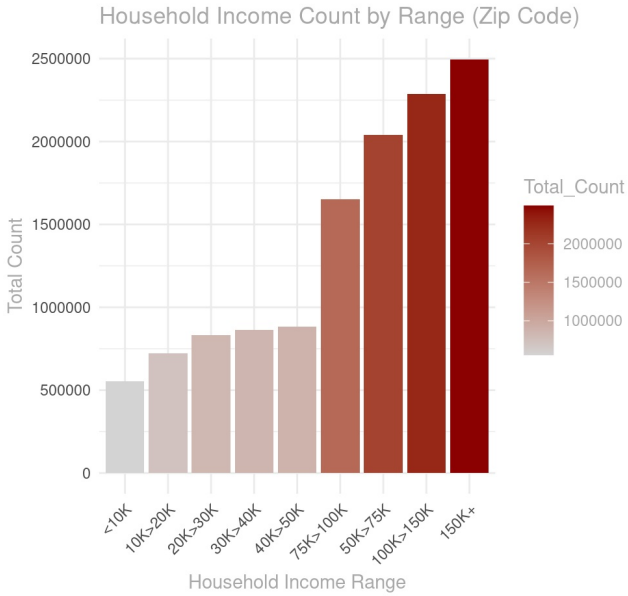
Regional Market Map (Zip Codes)



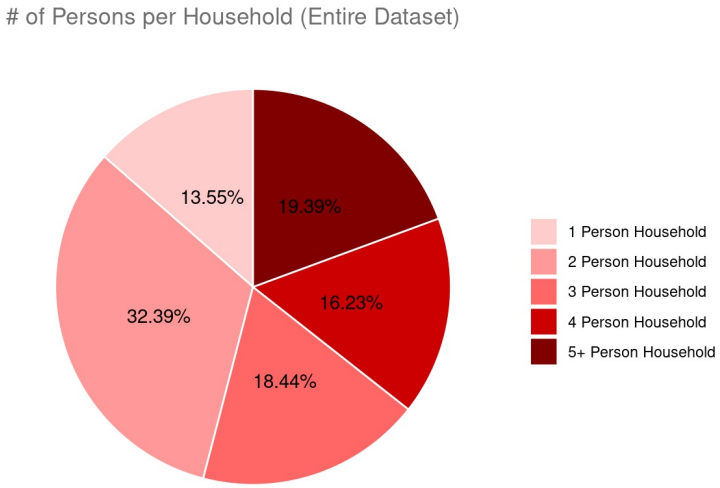
Demographics Drill Down

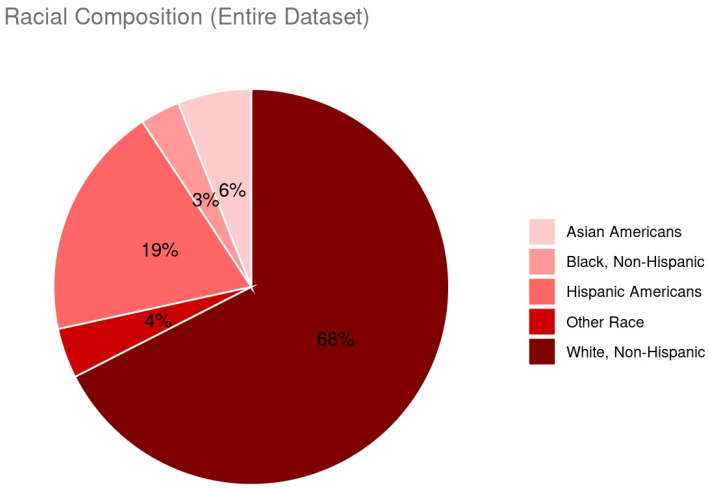
```
#The data collected on household income is poorly distributed.
#Makes it appear that most people are wealthy, in the 150K+ group.
#We will consider combining household income groups into 50K segments to capture a more realistic distribution.
combined_plots #aggregated by Household Income (averages and counts)
```



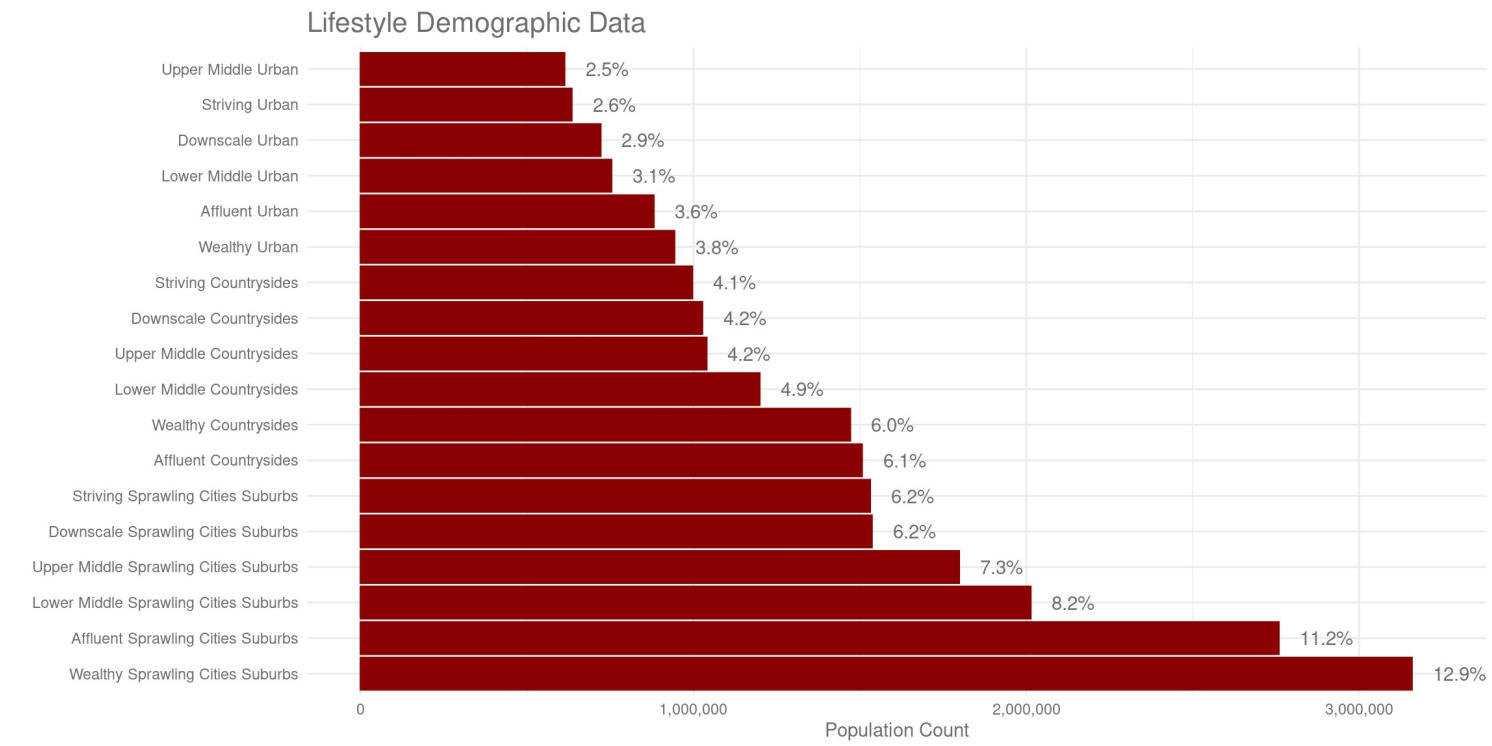



```
combined_pie_charts
```






```
# The Lifestyle plot gives us an idea of how the population is spread amongst the various Zip codes that belong to all of the Swire (& Competitors) market areas. More than half of all consumers live in the suburban sprawl. As a reminder the total number of adults included in the market area was 24,578,188.
lifestyle_plot
```




```
# Typical Men - more in the earlier years, dying off by middle age faster than women.
age_group_plot
```