Modeling Product #7

Michael Tom

April 19, 2024

Set up

Taking a sample of the whole dataset

```
df <- readRDS("swire_no_nas.rds")  #inject the data and we will sub-sample
```

```
regions_joinme <- read.csv("states_summary.csv")

unique(regions_joinme$REGION)
```

```
##  [1] "NORTHERN"    "DESERT_SW"   "PRAIRIE"     "CALI_NEVADA" "MOUNTAIN"
##  [6] "SOCAL"       "ARIZONA"     "NEWMEXICO"   "NOCAL"       "COLORADO"
## [11] "KANSAS"
```

```
# "NORTHERN"    "DESERT_SW"   "PRAIRIE"     "CALI_NEVADA"  "MOUNTAIN"    "SOCAL"   "ARIZONA"
"NEWMEXICO"   "NOCAL"    "COLORADO"    "KANSAS"

str(regions_joinme)
```

```
## 'data.frame':    200 obs. of  2 variables:
##  $ MARKET_KEY: int  13 70 179 197 272 352 32 33 44 50 ...
##  $ REGION    : chr  "NORTHERN" "NORTHERN" "DESERT_SW" "DESERT_SW" ...
```

```
# Perform a left join using the merge() function
df <- merge(df, regions_joinme[, c("MARKET_KEY", "REGION")], by = "MARKET_KEY", all.x = TRUE)
rm(regions_joinme)
```

Quick imputations

```
# Update CALORIC_SEGMENT values: 0 if 'DIET/LIGHT', otherwise 1
df$CALORIC_SEGMENT <- ifelse(df$CALORIC_SEGMENT == "DIET/LIGHT", 0, 1)
```

```
df$MARKET_KEY <- as.character(df$MARKET_KEY)
df <- df %>%
  mutate(
    MONTH = as.numeric(substr(DATE, 6, 7)),  # Extract the month from YYYY-MM-DD format
    SEASON = case_when(
      MONTH %in% c(12, 01, 02) ~ "WINTER",
      MONTH %in% c(03, 04, 05) ~ "SPRING",
      MONTH %in% c(06, 07, 08) ~ "SUMMER",
      MONTH %in% c(09, 10, 11) ~ "FALL",
      TRUE ~ NA_character_  # This is just in case there are any undefined values
    )
  )
```

```
str(df)
```

```
## 'data.frame':    24461424 obs. of  13 variables:
##  $ MARKET_KEY     : chr  "1" "1" "1" "1" ...
##  $ DATE           : chr  "2021-10-16" "2022-06-04" "2022-02-05" "2022-10-08" ...
##  $ CALORIC_SEGMENT: num  0 0 1 0 0 1 0 0 1 0 ...
##  $ CATEGORY       : chr  "ENERGY" "SSD" "SSD" "SSD" ...
##  $ UNIT_SALES     : num  434 28 42 1 26 161 6 5 68 90 ...
##  $ DOLLAR_SALES   : num  924.04 147.77 25.13 0.99 94.56 ...
##  $ MANUFACTURER   : chr  "PONYS" "SWIRE-CC" "COCOS" "JOLLYS" ...
##  $ BRAND          : chr  "MYTHICAL BEVERAGE ULTRA" "DIET PEPPY CF" "HANSENIZZLE'S ECO" "DIET
PAPI" ...
##  $ PACKAGE        : chr  "16SMALL MULTI CUP" "12SMALL 12ONE CUP" "12SMALL 6ONE CUP" "12SMALL
6ONE CUP" ...
##  $ ITEM           : chr  "MYTHICAL BEVERAGE ULTRA SUNRISE ENERGY DRINK UNFLAVORED ZERO SUGAR
CUP 16 LIQUID SMALL" "DIET PEPPY CAFFEINE FREE GENTLE DRINK RED  PEPPER COLA DIET CUP 12 LIQUID
SMALL X12" "HANSENIZZLE'S ECO GENTLE DRINK MANDARIN DURIAN  CUP 12 LIQUID SMALL" "DIET PAPI
GENTLE DRINK COLA DIET CUP 12 LIQUID SMALL" ...
##  $ REGION         : chr  "NORTHERN" "NORTHERN" "NORTHERN" "NORTHERN" ...
##  $ MONTH          : num  10 6 2 10 7 9 9 6 10 5 ...
##  $ SEASON         : chr  "FALL" "SUMMER" "WINTER" "FALL" ...
```

Making a 10% sample of the data to shrink it

```
# Assuming df is your dataframe
set.seed(123) # Set a random seed for reproducibility
sampled_df <- df[sample(1:nrow(df), 2446143), ]
rm(df)
```

```
df <- sampled_df
rm(sampled_df)
```

```
#skim(df)
```

```
summary(df)
```

```
##   MARKET_KEY            DATE           CALORIC_SEGMENT    CATEGORY
##  Length:2446143     Length:2446143     Min.   :0.0000   Length:2446143
##  Class :character   Class :character   1st Qu.:0.0000   Class :character
##  Mode  :character   Mode  :character   Median :1.0000   Mode  :character
```

```
##                                         Mean   :0.5025
##                                         3rd Qu.:1.0000
##                                         Max.   :1.0000
##    UNIT_SALES        DOLLAR_SALES      MANUFACTURER          BRAND
##  Min.   :    0.04   Min.   :     0.0   Length:2446143     Length:2446143
##  1st Qu.:   11.00   1st Qu.:    36.5   Class :character   Class :character
##  Median :   40.00   Median :   135.1   Mode  :character   Mode  :character
##  Mean   :  173.43   Mean   :   587.4
##  3rd Qu.:  126.00   3rd Qu.:   427.4
##  Max.   :91778.00   Max.   :409159.3
##    PACKAGE              ITEM              REGION              MONTH
##  Length:2446143     Length:2446143     Length:2446143     Min.   : 1.000
##  Class :character   Class :character   Class :character   1st Qu.: 3.000
##  Mode  :character   Mode  :character   Mode  :character   Median : 6.000
##                                                           Mean   : 6.283
##                                                           3rd Qu.: 9.000
##                                                           Max.   :12.000
##     SEASON
##  Length:2446143
##  Class :character
##  Mode  :character
##
##
##
```

Linear model on sampled data looks the same largely

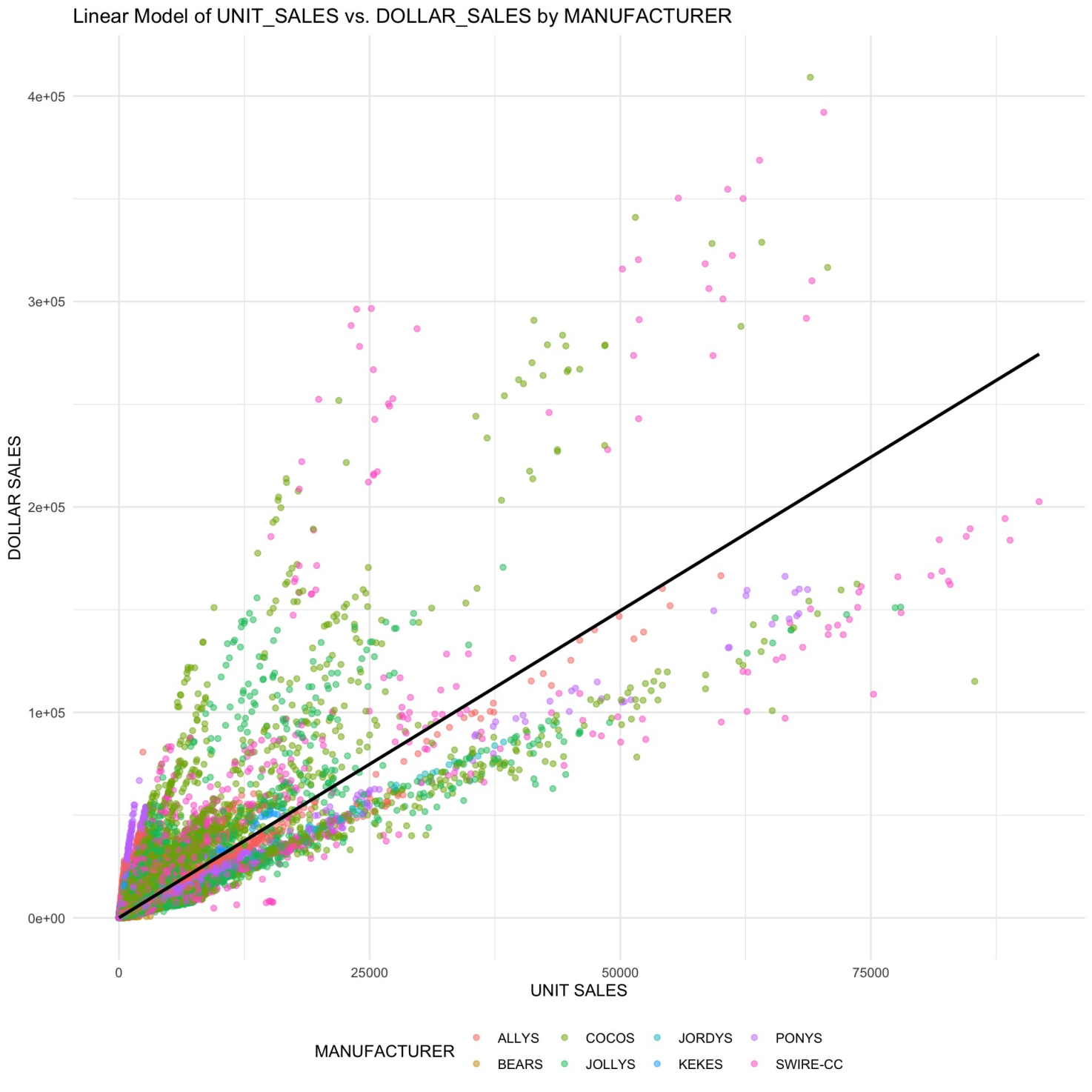
```
# Perform a linear regression with UNIT_SALES as the dependent variable
# and PRICE (or your chosen variable) as the independent variable
linear_model <- lm(DOLLAR_SALES ~ UNIT_SALES, data = df)

# Print the summary of the linear model to see the results
summary(linear_model)
```

```
##
## Call:
## lm(formula = DOLLAR_SALES ~ UNIT_SALES, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -140089    -117     -68      -3  225329
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 69.056096   1.023439   67.47   <2e-16 ***
## UNIT_SALES   2.989060   0.001201 2489.17   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1567 on 2446141 degrees of freedom
## Multiple R-squared:  0.717,  Adjusted R-squared:  0.717
## F-statistic: 6.196e+06 on 1 and 2446141 DF,  p-value: < 2.2e-16
```

```
# Create a scatter plot with the regression line, colored by MANUFACTURER
ggplot(df, aes(x = UNIT_SALES, y = DOLLAR_SALES, color = MANUFACTURER)) +
```

```
  geom_point(alpha = 0.5) +  # Adjust alpha to avoid overplotting, if necessary
  geom_smooth(method = "lm", color = "black", se = FALSE) +  # Add linear regression line
without confidence band for clarity
  labs(title = "Linear Model of UNIT_SALES vs. DOLLAR_SALES by MANUFACTURER",
       x = "UNIT SALES",
       y = "DOLLAR SALES") +
  theme_minimal() +
  theme(legend.position = "bottom")  # Adjust legend position if needed
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# create a table of total values by brand
brand_summary <- df %>%
  group_by(BRAND) %>%
  summarise(
    total_units_sold = sum(UNIT_SALES),
```

```
    total_revenue = sum(DOLLAR_SALES),
    avg_price = total_revenue / total_units_sold,
    total_days_sold = n() # Count the number of rows for each brand
  ) %>%
  arrange(desc(total_units_sold)) %>%  # Order by revenue in descending order
  mutate(rank = row_number())

summary(brand_summary)
```

```
##     BRAND           total_units_sold   total_revenue         avg_price
##  Length:288         Min.   :       1   Min.   :        1   Min.   : 0.5315
##  Class :character   1st Qu.:    2310   1st Qu.:     7563   1st Qu.: 2.0861
##  Mode  :character   Median :   94691   Median :   266075   Median : 3.0291
##                     Mean   : 1473003   Mean   :  4989427   Mean   : 3.2661
##                     3rd Qu.:  651385   3rd Qu.:  2161764   3rd Qu.: 3.7252
##                     Max.   :40414038   Max.   :159387186   Max.   :42.9378
##  total_days_sold         rank
##  Min.   :     1.0   Min.   :  1.00
##  1st Qu.:   121.8   1st Qu.: 72.75
##  Median :  1988.0   Median :144.50
##  Mean   :  8493.5   Mean   :144.50
##  3rd Qu.:  8075.8   3rd Qu.:216.25
##  Max.   :124603.0   Max.   :288.00
```

```
print(brand_summary[brand_summary$BRAND == "PEPPY", ])
```

```
## # A tibble: 1 × 6
##   BRAND total_units_sold total_revenue avg_price total_days_sold  rank
##   <chr>            <dbl>         <dbl>     <dbl>           <int> <int>
## 1 PEPPY        24118678.     89515431.      3.71           39613     3
```

> Peppy is a hugly popular brand coming in at 3rd in total revenue and 3rd in total units sold

Take a look at your brand..

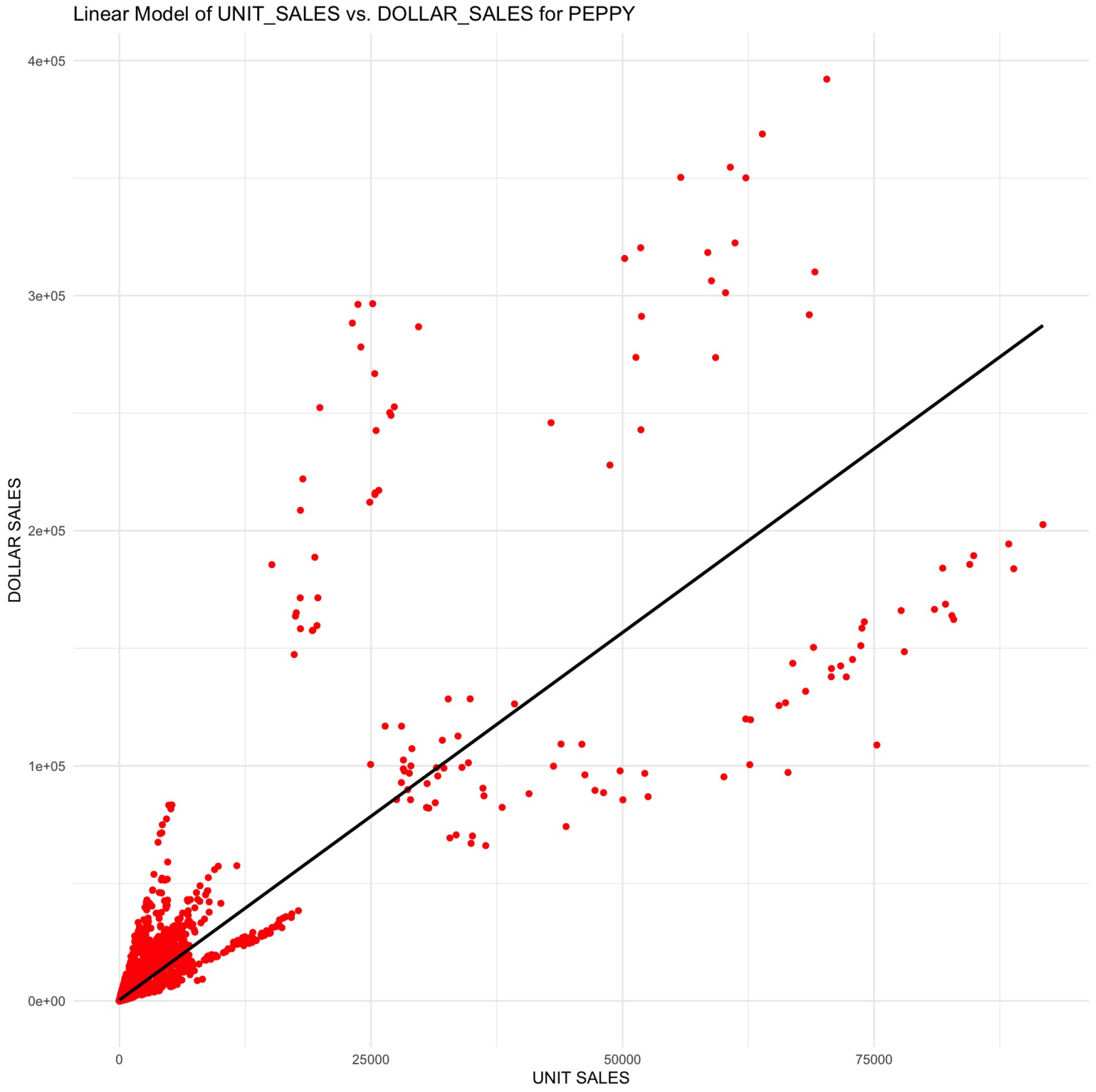
```
# Filter the dataframe for only 'PEPPY'
filtered_df <- df %>%
  filter(BRAND == "PEPPY")



summary(filtered_df)
```

```
##   MARKET_KEY            DATE           CALORIC_SEGMENT   CATEGORY
##  Length:39613       Length:39613       Min.   :1       Length:39613
##  Class :character   Class :character   1st Qu.:1       Class :character
##  Mode  :character   Mode  :character   Median :1       Mode  :character
##                                        Mean   :1
##                                        3rd Qu.:1
##                                        Max.   :1
##    UNIT_SALES        DOLLAR_SALES      MANUFACTURER          BRAND
##  Min.   :    0.17   Min.   :     0.2   Length:39613       Length:39613
##  1st Qu.:   56.00   1st Qu.:   190.7   Class :character   Class :character
##  Median :  154.00   Median :   588.6   Mode  :character   Mode  :character
##  Mean   :  608.86   Mean   :  2259.7
```

```
##  3rd Qu.:  481.00   3rd Qu.:  1824.8
##  Max.   :91778.00   Max.   :392062.7
##    PACKAGE              ITEM              REGION              MONTH
##  Length:39613       Length:39613       Length:39613       Min.   : 1.000
##  Class :character   Class :character   Class :character   1st Qu.: 3.000
##  Mode  :character   Mode  :character   Mode  :character   Median : 6.000
##                                                           Mean   : 6.306
##                                                           3rd Qu.: 9.000
##                                                           Max.   :12.000
##     SEASON
##  Length:39613
##  Class :character
##  Mode  :character
##
##
##
```

```
# Create the plot
ggplot(filtered_df, aes(x = UNIT_SALES, y = DOLLAR_SALES)) +
  geom_point(color = "red", alpha = 1) +  # Bright red points with full opacity
  geom_smooth(method = "lm", color = "black", se = FALSE) +  # Add linear regression line
without confidence band
  labs(title = "Linear Model of UNIT_SALES vs. DOLLAR_SALES for PEPPY",
       x = "UNIT SALES",
       y = "DOLLAR SALES") +
  theme_minimal() +
  theme(legend.position = "none")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



>PEPPY only has multiple sizes of their one drink “GENTLE DRINK RED PEPPER COLA”

```
# Check Tenure of all items
filtered_df %>%
  mutate(DATE = as.Date(DATE)) %>%
    group_by(ITEM) %>%
  summarize(Date_Difference = difftime(max(DATE), min(DATE), units = "weeks"),
            Total_Unit_Sales = sum(UNIT_SALES)) %>%
  arrange(Date_Difference)
```

```
## # A tibble: 38 × 3
##    ITEM                                        Date_Difference Total_Unit_Sales
##    <chr>                                       <drtn>                     <dbl>
##  1 PEPPY GENTLE DRINK RED  PEPPER COLA CUP 8 L…   0 weeks                      1
##  2 PEPPY GENTLE DRINK RED  PEPPER COLA JUG 12.…   0 weeks                      1
##  3 PEPPY GENTLE DRINK RED  PEPPER COLA CUP 12 …  71 weeks                      2
##  4 PEPPY GENTLE DRINK RED  PEPPER COLA JUG 16 …  74 weeks                     18
```

```
##  5 PEPPY GENTLE DRINK RED  PEPPER COLA JUG 12 …  83 weeks                    326
##  6 PEPPY GENTLE DRINK RED  PEPPER COLA JUG 13.… 104 weeks                  17234
##  7 PEPPY GENTLE DRINK RED  PEPPER COLA JUG 67.… 113 weeks                   7818
##  8 PEPPY GENTLE DRINK RED  PEPPER COLA CUP 7.5… 120 weeks                  20545
##  9 PEPPY GENTLE DRINK RED  PEPPER COLA CUP 12 … 128 weeks                    524
## 10 PEPPY GENTLE DRINK RED  PEPPER COLA JUG 12 … 134 weeks                    104
## # i 28 more rows
```

> Peppy only has 2 products that could be considered innovation products, but they each were only sold once and each only sold 1 unit.

```
#check for Pink woodsy flavor sales
sales_by_pink_woodsy <- df %>%
  filter(str_detect(ITEM, "PINK") & str_detect(ITEM, "WOODSY"))
```

> there are no current sales around pink woodsy flavored data.

```
#check for packaging
df %>%
  filter(CATEGORY == "SSD",
         CALORIC_SEGMENT == 1,
         str_detect(PACKAGE, "JUG") & str_detect(PACKAGE, "\\.5L") & str_detect(PACKAGE,
"MULTI")) %>%
  mutate(PACKAGE = as.factor(PACKAGE))
```

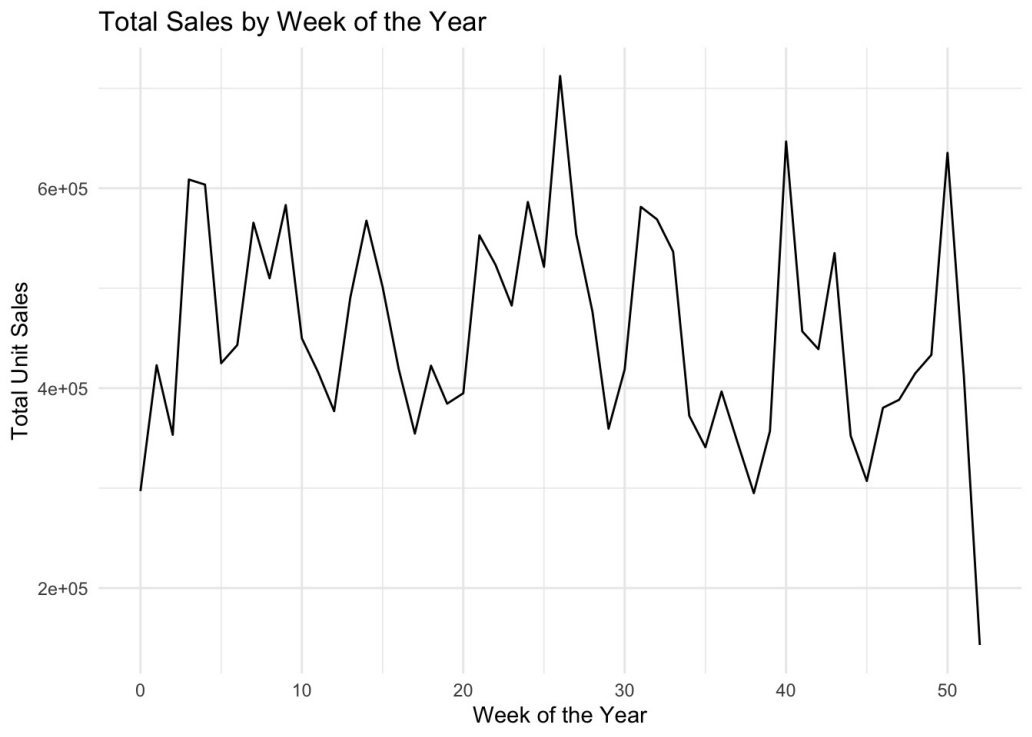
```
##    MARKET_KEY       DATE CALORIC_SEGMENT CATEGORY UNIT_SALES DOLLAR_SALES
## 1         132 2023-01-28               1      SSD          1         1.50
## 2           6 2023-07-08               1      SSD          1         1.50
## 3        1172 2023-02-04               1      SSD          1         1.50
## 4        1172 2022-02-12               1      SSD          1         1.25
## 5        1135 2021-04-10               1      SSD          1         1.00
## 6         817 2023-02-18               1      SSD          1         1.50
## 7        1172 2022-08-27               1      SSD          1         1.25
## 8         399 2023-01-28               1      SSD          2         3.00
## 9         535 2021-10-16               1      SSD          2         2.00
## 10        216 2022-04-09               1      SSD          1         1.25
##    MANUFACTURER         BRAND        PACKAGE
## 1        JOLLYS HILL MOISTURE 1.5L MULTI JUG
## 2        JOLLYS          PAPI 1.5L MULTI JUG
## 3        JOLLYS HILL MOISTURE 1.5L MULTI JUG
## 4        JOLLYS HILL MOISTURE 1.5L MULTI JUG
## 5      SWIRE-CC         SMASH 1.5L MULTI JUG
## 6        JOLLYS HILL MOISTURE 1.5L MULTI JUG
## 7        JOLLYS          PAPI 1.5L MULTI JUG
## 8        JOLLYS HILL MOISTURE 1.5L MULTI JUG
## 9        JOLLYS          PAPI 1.5L MULTI JUG
## 10       JOLLYS HILL MOISTURE 1.5L MULTI JUG
##                                                   ITEM      REGION MONTH SEASON
## 1  RAINING GENTLE DRINK AVOCADO  JUG 50.7 LIQUID SMALL CALI_NEVADA     1 WINTER
## 2         PAPI GENTLE DRINK COLA JUG 50.7 LIQUID SMALL    NORTHERN     7 SUMMER
## 3  RAINING GENTLE DRINK AVOCADO  JUG 50.7 LIQUID SMALL      KANSAS     2 WINTER
## 4  RAINING GENTLE DRINK AVOCADO  JUG 50.7 LIQUID SMALL      KANSAS     2 WINTER
## 5     SMASH GENTLE DRINK SUNSET  JUG 50.7 LIQUID SMALL     PRAIRIE     4 SPRING
## 6  RAINING GENTLE DRINK AVOCADO  JUG 50.7 LIQUID SMALL    COLORADO     2 WINTER
## 7         PAPI GENTLE DRINK COLA JUG 50.7 LIQUID SMALL      KANSAS     8 SUMMER
## 8  RAINING GENTLE DRINK AVOCADO  JUG 50.7 LIQUID SMALL    MOUNTAIN     1 WINTER
```

```
## 9          PAPI GENTLE DRINK COLA JUG 50.7 LIQUID SMALL   NEWMEXICO    10   FALL
## 10 RAINING GENTLE DRINK AVOCADO  JUG 50.7 LIQUID SMALL   DESERT_SW     4 SPRING
```

> there is currently no sales of .5LJUG

Sales by Week of the year

```
filtered_df %>%
  mutate(DATE = as.Date(DATE)) %>%
  mutate(WEEK = as.integer(format(DATE, "%U"))) %>%
  group_by(WEEK) %>%
  summarise(total_sales = sum(UNIT_SALES)) %>%
  ggplot(aes(x = WEEK, y = total_sales)) +
  geom_line(color = "black") +  # Blue line connecting points
  labs(title = "Total Sales by Week of the Year",
       x = "Week of the Year",
       y = "Total Unit Sales") +
  theme_minimal()
```



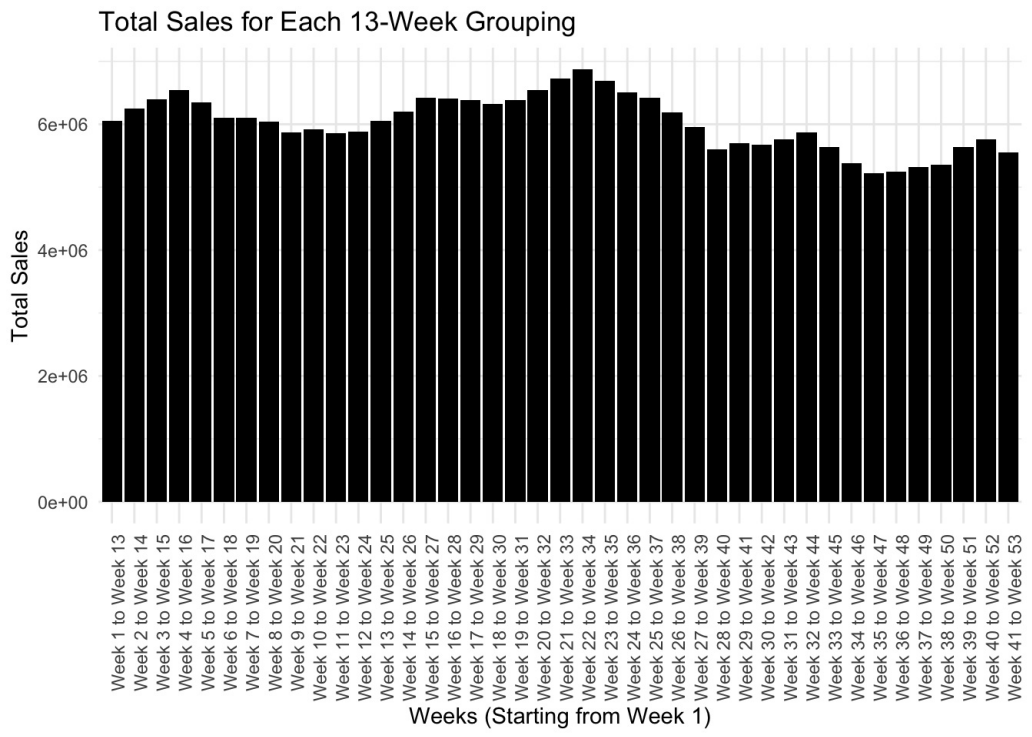

```
#find the best 13 weeks
library(zoo)
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
# Calculate total sales for each group of 13 consecutive weeks
sales_by_group <- filtered_df %>%
  mutate(DATE = as.Date(DATE)) %>%
  mutate(WEEK = as.integer(format(DATE, "%U"))) %>%
  group_by(WEEK) %>%
```

```
  summarise(total_sales = sum(UNIT_SALES)) %>%
  mutate(sales_in_group = rollsum(total_sales, 13, align = "left", fill = NA)) %>%
  mutate(week_label = paste0("Week ", WEEK + 1, " to Week ", WEEK + 13)) %>%
  arrange(WEEK) %>%  # Order by WEEK
  filter(!is.na(sales_in_group))  # Remove rows with sales_in_group = NA

# Plot the bar chart
sales_by_group$week_label <- factor(sales_by_group$week_label, levels =
sales_by_group$week_label[order(sales_by_group$WEEK)])
ggplot(sales_by_group, aes(x = factor(week_label), y = sales_in_group)) +
  geom_bar(stat = "identity", fill = "black") +
  labs(title = "Total Sales for Each 13-Week Grouping",
       x = "Weeks (Starting from Week 1)",
       y = "Total Sales") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



> From this graph we see that weeks 21 to 34 historically have the highest unit sales of PEPPY

Made a new smaller “innovation” data fram

```
innovation <- df %>%
  filter(CATEGORY == "SSD",
         CALORIC_SEGMENT == 1,
         str_detect(ITEM, "PINK") | str_detect(ITEM, "WOODSY"))

print(unique(innovation$ITEM))
```

```
##  [1] "FANTASMIC GENTLE DRINK PINK  CUP 12 LIQUID SMALL X12"
##  [2] "KOOL! FLUFFY GENTLE DRINK KIWANO  PINK  CUP 12 LIQUID SMALL"
##  [3] "MOONLIT GENTLE DRINK PINK  ADE  CUP 7.5 LIQUID SMALL X6"
##  [4] "FANTASMIC GENTLE DRINK PINK  JUG 12 LIQUID SMALL"
##  [5] "FANTASMIC GENTLE DRINK PINK  JUG 20 LIQUID SMALL"
##  [6] "MOONLIT GENTLE DRINK PINK  ADE  JUG 20 LIQUID SMALL"
##  [7] "MOONLIT GENTLE DRINK PINK  ADE  CUP 12 LIQUID SMALL X12"
```

```
##  [8] "SMASH GENTLE DRINK PINK  JUG 12 LIQUID SMALL X4"
##  [9] "MOONLIT GENTLE DRINK PINK  JUG 20 LIQUID SMALL"
## [10] "RAINING COASTAL BLAST GENTLE DRINK WOODSY  DURIAN  CUP 7.5 LIQUID SMALL X10"
## [11] "ZIZZLES GENTLE DRINK PINK  JUG 67.6 LIQUID SMALL"
## [12] "RAINING KICK DARK ENERGY DRINK PINK  SUPER-JUICE  CUP 16 LIQUID SMALL"
## [13] "SMASH GENTLE DRINK PINK  CUP 12 LIQUID SMALL X12"
## [14] "SMASH GENTLE DRINK PINK  JUG 67.6 LIQUID SMALL"
## [15] "FANTASMIC GENTLE DRINK PINK  JUG 67.6 LIQUID SMALL"
## [16] "MOONLIT GENTLE DRINK PINK  JUG 67.6 LIQUID SMALL"
## [17] "RAINING COASTAL BLAST GENTLE DRINK WOODSY  DURIAN  AVOCADO  CUP 12 LIQUID SMALL X12"
## [18] "SMASH GENTLE DRINK PINK  290 CALORIES PER JUG JUG 20 LIQUID SMALL"
## [19] "MOONLIT GENTLE DRINK PINK  ADE  JUG 67.6 LIQUID SMALL"
## [20] "RAINING COASTAL PITAYA  GENTLE DRINK WOODSY  DURIAN  AVOCADO  CUP 12 LIQUID SMALL X12"
## [21] "ZIZZLES GENTLE DRINK KEKE  PINK  CUP 12 LIQUID SMALL X12"
## [22] "RAINING COASTAL PITAYA  GENTLE DRINK WOODSY  DURIAN  AVOCADO  JUG 20 LIQUID SMALL"
## [23] "ELF BUBBLES GENTLE DRINK SUPER-JUICE  DURIAN  WOODSY  MIX JUG 20 LIQUID SMALL"
## [24] "RAINING COASTAL BLAST GENTLE DRINK WOODSY  DURIAN  AVOCADO  JUG 20 LIQUID SMALL"
## [25] "RAINING COASTAL BLAST GENTLE DRINK WOODSY  DURIAN  AVOCADO  JUG 16.9 LIQUID SMALL X6"
## [26] "MOONLIT GENTLE DRINK PINK  CUP 12 LIQUID SMALL X12"
## [27] "RAINING COASTAL BLAST GENTLE DRINK WOODSY  DURIAN  AVOCADO  CUP 12 LIQUID SMALL"
## [28] "RAINING COASTAL BLAST GENTLE DRINK WOODSY  DURIAN  AVOCADO  CUP 12 LIQUID SMALL X36"
## [29] "GO-DAY GENTLE DRINK PINK  JUG 20 LIQUID SMALL"
## [30] "RAINING COASTAL BLAST GENTLE DRINK WOODSY  DURIAN  AVOCADO  JUG 33.8 LIQUID SMALL"
## [31] "FANTASMIC GENTLE DRINK PINK  CUP 12 LIQUID SMALL"
## [32] "WILDWOOD GENTLE DRINK PINK  JUG 20 LIQUID SMALL"
## [33] "ZIZZLES GENTLE DRINK KEKE  PINK  CUP 12 LIQUID SMALL"
## [34] "WOODSY  YAWN GENTLE DRINK WOODSY  CUP 12 LIQUID SMALL X12"
## [35] "WOODSY  YAWN GENTLE DRINK WOODSY  JUG 20 LIQUID SMALL"
## [36] "SMASH GENTLE DRINK PINK  170 CALORIES PER CUP CUP 12 LIQUID SMALL"
## [37] "ZIZZLES GENTLE DRINK PINK  JUG 84.5 LIQUID SMALL"
## [38] "SMASH GENTLE DRINK PINK  JUG 16.9 LIQUID SMALL X6"
## [39] "GO-DAY RED POP! GENTLE DRINK PINK  JUG 24 LIQUID SMALL"
## [40] "FANTASMIC GENTLE DRINK PINK  CUP 7.5 LIQUID SMALL X6"
## [41] "GO-DAY RED POP! GENTLE DRINK PINK  CUP 12 LIQUID SMALL"
## [42] "HANSENIZZLE'S ECO GENTLE DRINK KEKE  PINK  CUP 12 LIQUID SMALL"
## [43] "RAINING COASTAL BLAST GENTLE DRINK WOODSY  DURIAN  AVOCADO  CUP 12 LIQUID SMALL X24"
## [44] "PINK TINGLE GENTLE DRINK SPARKLING PURPLE TROPPY  JUG 10.14 LIQUID SMALL"
## [45] "ZIZZLES GENTLE DRINK PINK  CUP 12 LIQUID SMALL"
## [46] "WOODSY  YAWN GENTLE DRINK SUPER-JUICE  DURIAN  WOODSY  CUP 16 LIQUID SMALL"
## [47] "ZIZZLES GENTLE DRINK PINK  CUP 12 LIQUID SMALL X12"
## [48] "GO-DAY RED POP! GENTLE DRINK PINK  JUG 12 LIQUID SMALL"
## [49] "MOONLIT GENTLE DRINK PINK  ADE  CUP 12 LIQUID SMALL"
## [50] "RAINING COASTAL BLAST GENTLE DRINK WOODSY  DURIAN  AVOCADO  JUG 24 LIQUID SMALL"
## [51] "GO-DAY RED POP! GENTLE DRINK PINK  JUG 84.5 LIQUID SMALL"
## [52] "FANTASMIC GENTLE DRINK PINK  JUG 12 LIQUID SMALL X24"
## [53] "ELF BUBBLES GENTLE DRINK SUPER-JUICE  DURIAN  WOODSY  JUG 20 LIQUID SMALL"
## [54] "PAPI SUMMER MIX GENTLE DRINK WOODSY  TROPPY  COLA JUG 20 LIQUID SMALL"
## [55] "GO-DAY GENTLE DRINK KEKE  PINK  CUP 12 LIQUID SMALL"
## [56] "AZURE HORIZON GENTLE DRINK WILD PINK  CUP 12 LIQUID SMALL"
## [57] "GO-DAY GENTLE DRINK PINK  CUP 12 LIQUID SMALL X12"
## [58] "GO-DAY RED POP! GENTLE DRINK PINK  JUG 67.6 LIQUID SMALL"
## [59] "FANTASMIC GENTLE DRINK PINK  JUG 16 LIQUID SMALL"
## [60] "HANSENIZZLE'S ECO CUPE REFRESHER  GENTLE DRINK KEKE  PINK  CUP 12 LIQUID SMALL"
```

```
#there are 60 items with SSD, Regular, and PINT OR WOODSY, but none of them are from PEPPY.
```

```
#Add a month Date factor
library(dplyr)
library(lubridate)

innovation <- innovation %>%
  mutate(
    MONTH = month(ymd(DATE)),  # Extract month using lubridate's ymd function
    MONTH = as.factor(MONTH)   # Convert the extracted month into a factor
  )

str(innovation)
```

```
## 'data.frame':    31460 obs. of  13 variables:
##  $ MARKET_KEY     : chr  "187" "583" "61" "32" ...
##  $ DATE           : chr  "2022-09-10" "2022-09-17" "2022-12-24" "2021-03-20" ...
##  $ CALORIC_SEGMENT: num  1 1 1 1 1 1 1 1 1 1 ...
##  $ CATEGORY       : chr  "SSD" "SSD" "SSD" "SSD" ...
##  $ UNIT_SALES     : num  65 6 5 31 119 20 12 1 5 13 ...
##  $ DOLLAR_SALES   : num  348.7 15.9 20.9 40 274.6 ...
##  $ MANUFACTURER   : chr  "COCOS" "COCOS" "SWIRE-CC" "COCOS" ...
##  $ BRAND          : chr  "FANTASMIC" "FLUFFY'S LIMITED EDITION KOOL!" "MOONLIT" "MEXICAN
FANTASMIC" ...
##  $ PACKAGE        : chr  "12SMALL 12ONE CUP" "12SMALL MLT BUMPY CUP" "7.5SMALL 6ONE CUP"
"12SMALL MLT PLASTICS JUG" ...
##  $ ITEM           : chr  "FANTASMIC GENTLE DRINK PINK  CUP 12 LIQUID SMALL X12" "KOOL!
FLUFFY GENTLE DRINK KIWANO  PINK  CUP 12 LIQUID SMALL" "MOONLIT GENTLE DRINK PINK  ADE  CUP 7.5
LIQUID SMALL X6" "FANTASMIC GENTLE DRINK PINK  JUG 12 LIQUID SMALL" ...
##  $ REGION         : chr  "NORTHERN" "NOCAL" "NORTHERN" "NORTHERN" ...
##  $ MONTH          : Factor w/ 12 levels "1","2","3","4",..: 9 9 12 3 8 8 8 1 4 4 ...
##  $ SEASON         : chr  "FALL" "FALL" "WINTER" "SPRING" ...
```

```
# Assuming 'innovation' is your data frame
model <- lm(DOLLAR_SALES ~ UNIT_SALES + CALORIC_SEGMENT + PACKAGE + SEASON + REGION, data =
innovation)
summary(model)
```

```
##
## Call:
## lm(formula = DOLLAR_SALES ~ UNIT_SALES + CALORIC_SEGMENT + PACKAGE +
##     SEASON + REGION, data = innovation)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9653.6   -92.1    14.0    66.4 24621.9
##
## Coefficients: (1 not defined because of singularities)
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       5.089e+01  1.625e+01   3.132 0.001735 **
## UNIT_SALES                        2.542e+00  9.985e-03 254.532  < 2e-16 ***
## CALORIC_SEGMENT                          NA         NA      NA       NA
## PACKAGE12SMALL 12ONE CUP          1.249e+02  1.556e+01   8.029 1.02e-15 ***
## PACKAGE12SMALL 24ONE CUP          2.942e+02  1.405e+02   2.093 0.036342 *
## PACKAGE12SMALL 24ONE PLASTICS JUG -5.699e+01  2.421e+02  -0.235 0.813940
## PACKAGE12SMALL 36ONE CUP          1.396e+03  5.216e+01  26.768  < 2e-16 ***
## PACKAGE12SMALL 4ONE PLASTICS JUG  -9.412e+01  2.591e+01  -3.633 0.000280 ***
```

```
## PACKAGE12SMALL 6ONE CUP           -1.556e+02  2.574e+01  -6.044 1.52e-09 ***
## PACKAGE12SMALL MLT BUMPY CUP      -1.124e+02  2.146e+01  -5.239 1.62e-07 ***
## PACKAGE12SMALL MLT PLASTICS JUG   -1.041e+02  1.963e+01  -5.305 1.14e-07 ***
## PACKAGE16SMALL MLT SHADYES JUG    -5.426e+01  4.187e+02  -0.130 0.896897
## PACKAGE16SMALL MULTI CUP          -2.802e+02  7.710e+01  -3.634 0.000279 ***
## PACKAGE1L MULTI JUG               -1.647e+02  7.232e+01  -2.277 0.022792 *
## PACKAGE20SMALL MULTI JUG          -1.518e+02  1.565e+01  -9.698  < 2e-16 ***
## PACKAGE24SMALL MLT SHADYES JUG    -1.060e+02  5.313e+01  -1.994 0.046151 *
## PACKAGE2L MULTI JUG               -1.676e+02  1.574e+01 -10.643  < 2e-16 ***
## PACKAGE7.5SMALL 10ONE CUP         -4.654e+01  3.437e+01  -1.354 0.175692
## PACKAGE7.5SMALL 6ONE CUP          -6.977e+01  1.832e+01  -3.808 0.000140 ***
## PACKAGEALL OTHER ONES             -6.006e+01  5.974e+01  -1.005 0.314734
## SEASONSPRING                       1.773e+00  7.034e+00   0.252 0.800953
## SEASONSUMMER                       4.245e+01  6.569e+00   6.462 1.05e-10 ***
## SEASONWINTER                      -6.942e+00  7.104e+00  -0.977 0.328497
## REGIONCALI_NEVADA                  1.121e+01  1.334e+01   0.840 0.400718
## REGIONCOLORADO                     3.355e+01  8.397e+00   3.996 6.46e-05 ***
## REGIONDESERT_SW                    9.157e+00  9.602e+00   0.954 0.340272
## REGIONKANSAS                       7.939e+01  1.551e+01   5.119 3.10e-07 ***
## REGIONMOUNTAIN                     6.719e+01  1.079e+01   6.229 4.75e-10 ***
## REGIONNEWMEXICO                    2.700e+01  1.140e+01   2.368 0.017875 *
## REGIONNOCAL                        8.286e+00  1.174e+01   0.706 0.480380
## REGIONNORTHERN                     3.742e+01  7.222e+00   5.182 2.21e-07 ***
## REGIONPRAIRIE                      5.414e+01  1.468e+01   3.689 0.000226 ***
## REGIONSOCAL                       -1.091e+01  9.029e+00  -1.209 0.226720
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 418.4 on 31428 degrees of freedom
## Multiple R-squared:  0.7225, Adjusted R-squared:  0.7222
## F-statistic:  2639 on 31 and 31428 DF,  p-value: < 2.2e-16
```

> This model returned an R2 of .7225, which is one of the lowest of our innovation products. The strongest predictors are the differnt sized items and where they are selling.

More exploration

```
library(dplyr)

small_group <- df %>%
  filter(UNIT_SALES < 76000, DOLLAR_SALES < 500000)

skim(small_group)
```

Data summary

Name	small_group
Number of rows	2446128
Number of columns	13

Column type frequency:	
character	9
numeric	4

Group variables	None

Variable type: character

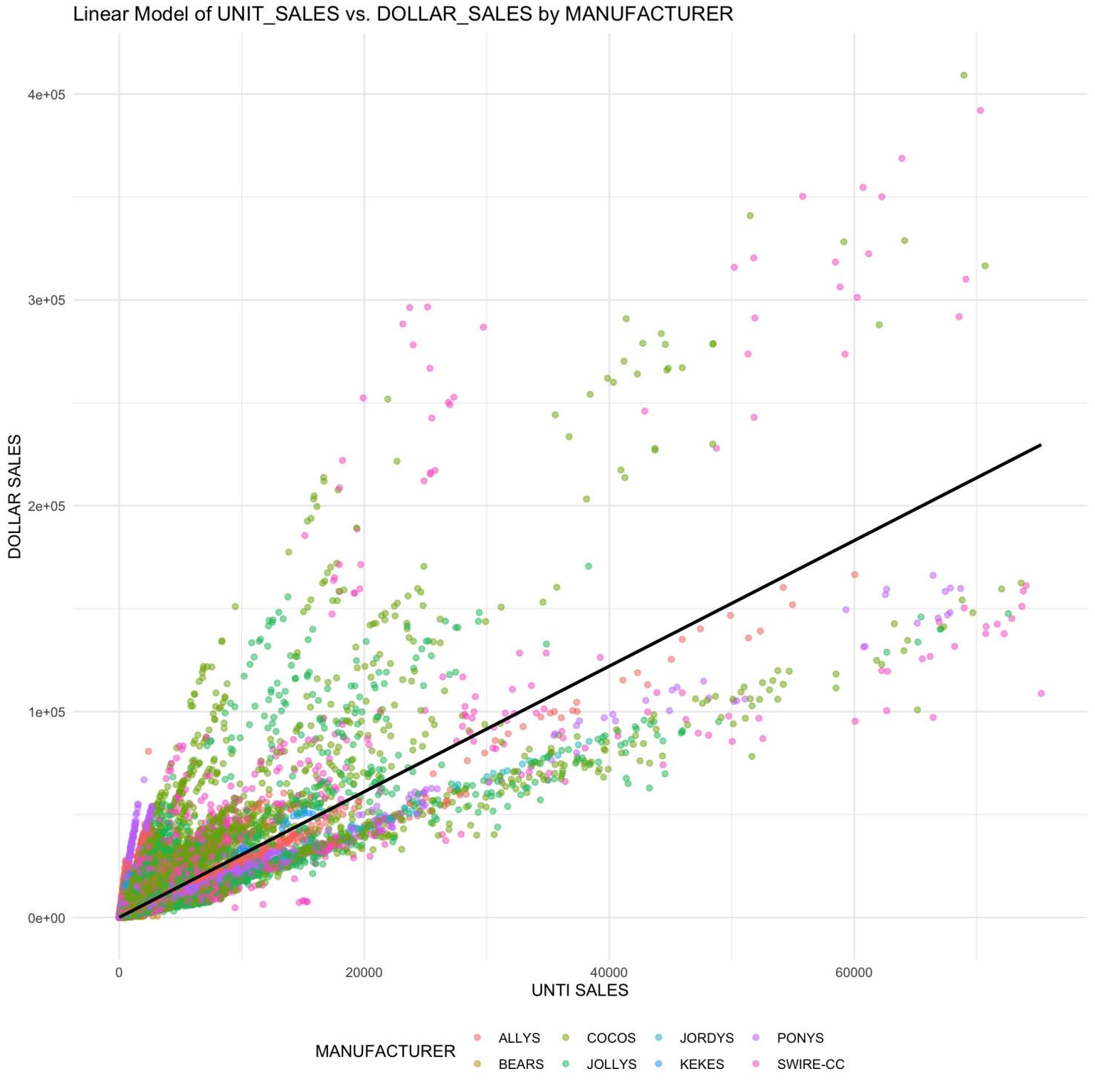
skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
MARKET_KEY	0	1	1	4	0	200	0
DATE	0	1	10	10	0	152	0
CATEGORY	0	1	3	18	0	5	0
MANUFACTURER	0	1	5	8	0	8	0
BRAND	0	1	4	56	0	288	0
PACKAGE	0	1	11	26	0	95	0
ITEM	0	1	26	142	0	2999	0
REGION	0	1	5	11	0	11	0
SEASON	0	1	4	6	0	4	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
CALORIC_SEGMENT	0	1	0.50	0.50	0.00	0.00	1.00	1.00	1.0	
UNIT_SALES	0	1	172.92	808.77	0.04	11.00	40.00	126.00	75266.0	
DOLLAR_SALES	0	1	586.41	2915.60	0.01	36.47	135.06	427.39	409159.3	
MONTH	0	1	6.28	3.43	1.00	3.00	6.00	9.00	12.0	

```
# Create a scatter plot with the regression line, colored by MANUFACTURER
ggplot(small_group, aes(x = UNIT_SALES, y = DOLLAR_SALES, color = MANUFACTURER)) +
  geom_point(alpha = 0.5) +  # Adjust alpha to avoid overplotting, if necessary
  geom_smooth(method = "lm", color = "black", se = FALSE) +  # Add linear regression line
without confidence band for clarity
  labs(title = "Linear Model of UNIT_SALES vs. DOLLAR_SALES by MANUFACTURER",
       x = "UNTI SALES",
       y = "DOLLAR SALES") +
  theme_minimal() +
  theme(legend.position = "bottom")  # Adjust legend position if needed
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



> This is where Peppy lives, as they are one of the highest in unit sales and revenue it is grabbing essentially everything.

#Make the small pink woodsy

```
pinkwoodsy_small <- df[grep("pink|woodsy", df$ITEM, ignore.case = TRUE), ]
```

```
skim(pinkwoodsy_small)
```

Data summary

Name	pinkwoodsy_small
Number of rows	136043
Number of columns	13

Column type frequency:

character	9
numeric	4

Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
MARKET_KEY	0	1	1	4	0	200	0
DATE	0	1	10	10	0	152	0
CATEGORY	0	1	3	18	0	4	0
MANUFACTURER	0	1	5	8	0	7	0
BRAND	0	1	5	45	0	55	0
PACKAGE	0	1	11	26	0	38	0
ITEM	0	1	45	112	0	165	0
REGION	0	1	5	11	0	11	0
SEASON	0	1	4	6	0	4	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
CALORIC_SEGMENT	0	1	0.57	0.50	0.00	0.0	1.00	1.0	1.00	
UNIT_SALES	0	1	105.66	298.79	0.04	9.0	31.00	97.0	17037.00	
DOLLAR_SALES	0	1	263.76	799.55	0.01	27.3	88.41	243.8	46442.23	
MONTH	0	1	6.37	3.37	1.00	4.0	6.00	9.0	12.00	

```
# Assuming 'innovation' is your data frame
model <- lm(DOLLAR_SALES ~ UNIT_SALES + CALORIC_SEGMENT + PACKAGE + CATEGORY + SEASON + REGION,
data = pinkwoodsy_small)
summary(model)
```

```
##
## Call:
## lm(formula = DOLLAR_SALES ~ UNIT_SALES + CALORIC_SEGMENT + PACKAGE +
##     CATEGORY + SEASON + REGION, data = pinkwoodsy_small)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8326.5   -50.0     3.7    55.5 25354.4
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       4.624e+02  1.199e+01  38.569  < 2e-16 ***
## UNIT_SALES                        2.450e+00  3.054e-03 802.428  < 2e-16 ***
```

```
## CALORIC_SEGMENT                    -6.249e+01  2.535e+00 -24.647  < 2e-16 ***
## PACKAGE.5L 6ONE JUG                -1.631e+02  5.820e+00 -28.024  < 2e-16 ***
## PACKAGE.5L MULTI JUG               -3.273e+02  1.146e+01 -28.561  < 2e-16 ***
## PACKAGE12SMALL 12ONE CUP           -8.108e+01  6.486e+00 -12.502  < 2e-16 ***
## PACKAGE12SMALL 24ONE CUP           -3.706e+02  1.580e+01 -23.452  < 2e-16 ***
## PACKAGE12SMALL 24ONE PLASTICS JUG -2.835e+02  1.664e+02  -1.703   0.0885 .
## PACKAGE12SMALL 36ONE CUP            1.222e+03  3.493e+01  34.968  < 2e-16 ***
## PACKAGE12SMALL 4ONE PLASTICS JUG  -2.833e+02  1.567e+01 -18.083  < 2e-16 ***
## PACKAGE12SMALL 6ONE CUP            -2.675e+02  1.208e+01 -22.139  < 2e-16 ***
## PACKAGE12SMALL 8ONE BUMPY CUP      -4.375e+02  2.881e+02  -1.518   0.1289
## PACKAGE12SMALL 8ONE CUP            -2.609e+02  1.212e+01 -21.518  < 2e-16 ***
## PACKAGE12SMALL MLT BUMPY CUP       -3.360e+02  9.705e+00 -34.626  < 2e-16 ***
## PACKAGE12SMALL MLT PLASTICS JUG    -3.006e+02  1.049e+01 -28.663  < 2e-16 ***
## PACKAGE12SMALL MULTI CUP           -1.950e+02  1.270e+01 -15.350  < 2e-16 ***
## PACKAGE15SMALL MLT                 -5.264e+02  1.542e+01 -34.143  < 2e-16 ***
## PACKAGE16SMALL 12ONE CUP           -4.707e+02  2.241e+01 -21.007  < 2e-16 ***
## PACKAGE16SMALL 24ONE CUP           -4.416e+02  1.922e+01 -22.978  < 2e-16 ***
## PACKAGE16SMALL MLT SHADYES JUG     -2.771e+02  2.881e+02  -0.962   0.3362
## PACKAGE16SMALL MULTI CUP           -4.584e+02  1.186e+01 -38.633  < 2e-16 ***
## PACKAGE18SMALL 6ONE                -1.642e+02  9.041e+00 -18.166  < 2e-16 ***
## PACKAGE18SMALL MULTI JUG           -2.256e+02  5.655e+00 -39.900  < 2e-16 ***
## PACKAGE1L MULTI JUG                -2.266e+02  1.172e+01 -19.340  < 2e-16 ***
## PACKAGE20SMALL 12ONE JUG           -1.370e+02  2.241e+01  -6.112 9.88e-10 ***
## PACKAGE20SMALL MULTI JUG           -3.505e+02  5.214e+00 -67.216  < 2e-16 ***
## PACKAGE24 - 25SMALL MULTI JUG      -2.753e+02  6.331e+00 -43.485  < 2e-16 ***
## PACKAGE24SMALL MLT SHADYES JUG     -3.066e+02  3.546e+01  -8.646  < 2e-16 ***
## PACKAGE24SMALL MULTI CUP           -4.019e+02  1.796e+01 -22.377  < 2e-16 ***
## PACKAGE26-32SMALL MLT              -1.264e+02  6.365e+00 -19.856  < 2e-16 ***
## PACKAGE2L MULTI JUG                -3.635e+02  6.716e+00 -54.123  < 2e-16 ***
## PACKAGE3L MULTI JUG                -3.482e+02  1.441e+02  -2.416   0.0157 *
## PACKAGE7.5SMALL 10ONE              -3.412e+02  2.883e+02  -1.184   0.2366
## PACKAGE7.5SMALL 10ONE CUP          -2.188e+02  2.216e+01  -9.875  < 2e-16 ***
## PACKAGE7.5SMALL 6ONE CUP           -2.690e+02  9.324e+00 -28.853  < 2e-16 ***
## PACKAGE8SMALL 12ONE CUP            -2.623e+02  1.332e+01 -19.686  < 2e-16 ***
## PACKAGE8SMALL 24ONE CUP            -4.076e+02  2.341e+01 -17.412  < 2e-16 ***
## PACKAGE8SMALL 4ONE CUP             -2.936e+02  1.283e+01 -22.878  < 2e-16 ***
## PACKAGE8SMALL MULTI CUP            -4.524e+02  1.274e+01 -35.508  < 2e-16 ***
## PACKAGEALL OTHER ONES              -2.509e+02  1.236e+01 -20.304  < 2e-16 ***
## CATEGORYING ENHANCED WATER         -2.728e+02  1.088e+01 -25.079  < 2e-16 ***
## CATEGORYSPARKLING WATER            -1.181e+02  5.437e+00 -21.721  < 2e-16 ***
## CATEGORYSSD                        -1.247e+02  1.089e+01 -11.452  < 2e-16 ***
## SEASONSPRING                       -3.895e+00  2.240e+00  -1.739   0.0820 .
## SEASONSUMMER                        3.456e+00  2.247e+00   1.538   0.1242
## SEASONWINTER                       -2.096e+00  2.284e+00  -0.918   0.3589
## REGIONCALI_NEVADA                   8.901e+00  4.534e+00   1.963   0.0496 *
## REGIONCOLORADO                      1.785e+01  2.791e+00   6.394 1.62e-10 ***
## REGIONDESERT_SW                     2.866e+00  3.315e+00   0.865   0.3873
## REGIONKANSAS                        1.159e+02  5.992e+00  19.343  < 2e-16 ***
## REGIONMOUNTAIN                      1.330e+01  3.104e+00   4.286 1.82e-05 ***
## REGIONNEWMEXICO                     1.652e+01  4.059e+00   4.069 4.72e-05 ***
## REGIONNOCAL                         7.681e+00  4.255e+00   1.805   0.0710 .
## REGIONNORTHERN                      1.120e+01  2.287e+00   4.895 9.82e-07 ***
## REGIONPRAIRIE                       2.146e+01  4.975e+00   4.313 1.61e-05 ***
## REGIONSOCAL                        -6.204e+00  3.245e+00  -1.912   0.0559 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 288 on 135987 degrees of freedom
## Multiple R-squared:  0.8703, Adjusted R-squared:  0.8703
## F-statistic: 1.659e+04 on 55 and 135987 DF,  p-value: < 2.2e-16
```

> r2 even higher than before of .8703. This one had about 60k more ovbservations to train on.

Cleaning

> Rework pinkwoodsy for more features

```
pinkwoodsy_small <- df %>%
  filter(CATEGORY == "SSD",
         CALORIC_SEGMENT == 1, # Specify each pattern separately
         REGION %in% c("NEWMEXICO", "ARIZONA", "DESERT_SW"))

pinkwoodsy_small <- pinkwoodsy_small %>%
  mutate(
    PACKAGE2 = str_extract(ITEM, "(CUP|JUG).*"),  # Extracts the part from CUP or JUG to the
end.
    ITEM = str_replace(ITEM, "(CUP|JUG).*", "")  # Replaces the CUP/JUG and everything after it
with empty string in ITEM.
  )
```

```
pinkwoodsy_small <- pinkwoodsy_small %>%
  mutate(
    TEMP = str_extract(ITEM, "\\d+\\.?\\d*.*"), # Extracts the part from the first number to
the end.
    PACKAGE2 = if_else(is.na(PACKAGE2), TEMP, paste(PACKAGE2, TEMP)), # Combines existing
PACKAGE2 with new extraction if needed.
    ITEM = str_replace(ITEM, "\\d+\\.?\\d*.*", ""), # Removes the numeric part and everything
after it from ITEM.
    TEMP = NULL  # Removes the temporary column.
  )
```

```
na_rows <- pinkwoodsy_small %>%
  filter(is.na(PACKAGE2))
#na_rows
#the above steps excised all packaging out of ITEM column
```

```
pinkwoodsy_small <- pinkwoodsy_small %>%
  mutate(
    GENTLE_DRINK = if_else(str_detect(ITEM, "GENTLE DRINK"), 1, 0), # Assigns 1 if "GENTLE
DRINK" exists, otherwise 0.
    ITEM = str_replace(ITEM, "GENTLE DRINK", "") # Removes "GENTLE DRINK" from ITEM.
  )
```

```
pinkwoodsy_small <- pinkwoodsy_small %>%
  mutate(
    ENERGY_DRINK = if_else(str_detect(ITEM, "ENERGY DRINK"), 1, 0), # Assigns 1 if "ENERGY
DRINK" exists, otherwise 0.
    ITEM = str_replace(ITEM, "ENERGY DRINK", "") # Removes "ENERGY DRINK" from ITEM.
  )
```

```
library(stringr)
```

```
# Define the pattern as a regular expression
pattern <- "ZERO CALORIES|ZERO CALORIE|ZERO SUGAR|SUGAR FREE|NO CALORIES"

pinkwoodsy_small <- pinkwoodsy_small %>%
  mutate(
    CALORIC_SEGMENT_TEXT = str_extract(ITEM, pattern), # Extracts matching text based on the
pattern.
    ITEM = str_replace_all(ITEM, pattern, "") # Removes extracted text from ITEM.
  )
```

```
pinkwoodsy_small <- pinkwoodsy_small %>%
  mutate(
    CALORIC_SEGMENT_TEXT = if_else(str_detect(ITEM, "\\bDIET\\b"),
                                   if_else(is.na(CALORIC_SEGMENT_TEXT), "DIET",
paste(CALORIC_SEGMENT_TEXT, "DIET", sep=", ")),
                                   CALORIC_SEGMENT_TEXT)
  )
```

```
# Function to remove the second instance of any repeating word
remove_second_instance <- function(item) {
  words <- unlist(str_split(item, "\\s+")) # Split item into words
  unique_words <- unique(words) # Get unique words to check for repeats
  for (word in unique_words) {
    word_indices <- which(words == word) # Find all indices of the current word
    if (length(word_indices) > 1) { # If there is more than one occurrence
      words[word_indices[2]] <- "" # Remove the second occurrence
    }
  }
  return(paste(words, collapse = " ")) # Reconstruct sentence without the second instance
}

# Apply the function to the 'ITEM' column
pinkwoodsy_small <- pinkwoodsy_small %>%
  mutate(ITEM = sapply(ITEM, remove_second_instance))


# Remove specific columns
pinkwoodsy_small <- select(pinkwoodsy_small, -PACKAGE2, -GENTLE_DRINK, -ENERGY_DRINK, -
CALORIC_SEGMENT_TEXT)
```

```
head(pinkwoodsy_small)
```

```
##   MARKET_KEY       DATE CALORIC_SEGMENT CATEGORY UNIT_SALES DOLLAR_SALES
## 1        893 2022-02-26               1      SSD         37        60.74
## 2        794 2023-07-01               1      SSD        512      2854.85
## 3        882 2020-12-12               1      SSD         36        52.48
## 4        733 2023-04-22               1      SSD         38        86.28
## 5        585 2022-04-16               1      SSD        203       715.54
## 6        197 2021-07-10               1      SSD        329       380.65
##   MANUFACTURER         BRAND           PACKAGE                            ITEM
## 1     SWIRE-CC         SMASH      2L MULTI JUG                    SMASH PURPLE
## 2     SWIRE-CC   CUPADA ARID 12SMALL 12ONE CUP
## 3       JOLLYS HILL MOISTURE 16SMALL MULTI CUP                 RAINING AVOCADO
## 4     SWIRE-CC      RESIDENT      2L MULTI JUG                 RESIDENT GINGER
## 5        COCOS   ELF BUBBLES      .5L 6ONE JUG ELF BUBBLES SUPER-JUICE DURIAN
```

```
## 6        COCOS   ELF BUBBLES   1.25L MULTI JUG ELF BUBBLES SUPER-JUICE DURIAN
##      REGION MONTH SEASON
## 1   ARIZONA     2 WINTER
## 2   ARIZONA     7 SUMMER
## 3   ARIZONA    12 WINTER
## 4   ARIZONA     4 SPRING
## 5   ARIZONA     4 SPRING
## 6 DESERT_SW     7 SUMMER
```

```
write.csv(pinkwoodsy_small, "pinkwoodsy_small.csv", row.names = FALSE)
```

FINAL THOUGHTS

> Thorugh our analysis of a “Pink Woodsy” flavored launch, there was very little evidance that further modeling would create a reliable prediction. As the historical data is missing many accurate features we would like to see in order do explaine variation. A few of the features from this specific innovation product that are missing are: 1. Lack of compariabel flavors. Though there have been products in the past with Pink or Woodsy, there has never been any items with this combination. 2. Brand “Peppy” having no innovation product data. In our research of the brand we found they do not have any innovation data that would give us indications of how a new product would compete if launched. 3. Lack of deffinition of which regions or areas would be considered “South.” For this launch. With these crucical factors either being excluded from modeling or using best estimates on the “closest” items we do not believe moving forward with prediciton of this would be advised. With a product such as this any type of trials data or directions on which items would be most comparable would help assure accuracy.