

# Theresa Wohlever | Assignment 02: Data Exploration

## Load packages

```
library(ggplot2)
library(dplyr)
```

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

filter, lag

The following objects are masked from 'package:base':

intersect, setdiff, setequal, union

```
library(stringr)
library(tidyr)
```

## Drug AMP Reporting - Quarterly

Drugs that have been reported under the Medicaid Drug Rebate Program along with an indication of whether or not the required Average Manufacturer Price (AMP) was reported for each drug. All drugs are identified in the file by the 11-digit National Drug Code, product name, labeler name, and reported (R) or not reported (NR).

## Path Preparation

```
setwd("/Users/theresawohlever/git_repos/BQOM-2578_DataMining/BQOM-2578_DataMining_twohlever/")
getwd()
```

```
[1] "/Users/theresawohlever/git_repos/BQOM-2578_DataMining/BQOM-2578_DataMining_twohlever/as"
```

## Environment clean up

```
rm(list = ls())
```

## Load data

Raw data from Medicaid Drug AMP Reporting: <https://data.medicaid.gov/dataset/80956a7d-e343-54f3-94a7-45d41b34fc0b#data-table>

```
base_FILENAME <- "DrugAMPReportingQuarterly022025" ## tiny" ## DrugAMPReportingQuarterly0220
# read the csv into a dataframe, which we can manipulate in R.

csv_FILE <- paste(base_FILENAME, ".csv", sep = "")
raw_amp_df <- read.csv(csv_FILE, stringsAsFactors = FALSE)

csv_OUT_FILE <- paste(base_FILENAME, "_processed.csv", sep = "")
```

## Data Discovery

```
# head displays the first rows
head(raw_amp_df)
```

	Labeler.Name	NDC
1	FLUORITAB CORPORATION	00288110601
2	FLUORITAB CORPORATION	00288110602
3	FLUORITAB CORPORATION	00288110610
4	FLUORITAB CORPORATION	00288110699

5 FLUORITAB CORPORATION  
 6 FLUORITAB CORPORATION

00288220101  
 00288220102

	FDA.Product.Name	Status	Year
1 SODIUM FLUORIDE I.I MG		NR	2013
2 SODIUM FLUORIDE 1.1 MG		NR	2013
3 SODIUM FLUORIDE 1.1MG		NR	2013
4 SODIUM FLUORIDE 1.1MG		NR	2013
5 SODIUM FLUORIDE 2.2MG		NR	2013
6 SODIUM FLUORIDE 2.2 MG		NR	2013

Quarter

1	1
2	1
3	1
4	1
5	1
6	1

```
# tail displays the last rows
tail(raw_amp_df)
```

	Labeler.Name	NDC	FDA.Product.Name	Status	Year
2031672	BAUSCH HEALTH US, LLC	99207030060	ZIANA GEL	R	2025
2031673	BAUSCH HEALTH US, LLC	99207046630	SOLODYN 80MG TABLETS	R	2025
2031674	BAUSCH HEALTH US, LLC	99207052510	VANOS CREAM .1%	R	2025
2031675	BAUSCH HEALTH US, LLC	99207052530	VANOS CREAM .1%	R	2025
2031676	BAUSCH HEALTH US, LLC	99207052560	VANOS CREAM .1%	R	2025
2031677	BAUSCH HEALTH US, LLC	99207085060	LUZU Cream 1% 60gm	R	2025

Quarter

2031672	2
2031673	2
2031674	2
2031675	2
2031676	2
2031677	2

```
# dim tells you how many rows by how many columns you have
dim(raw_amp_df)
```

```
[1] 2031677      6
```

```
# names returns the names of the columns that you have
names(raw_amp_df)
```

```
[1] "Labeler.Name"      "NDC"                "FDA.Product.Name" "Status"
[5] "Year"              "Quarter"
```

```
#summary will give you relevant summary statistics for each variable depending on its type
summary(raw_amp_df)
```

```
Labeler.Name      NDC      FDA.Product.Name      Status
Length:2031677    Length:2031677    Length:2031677    Length:2031677
Class :character  Class :character  Class :character  Class :character
Mode :character   Mode :character   Mode :character   Mode :character
```

```
      Year      Quarter
Min.   :2013   Min.    :1.000
1st Qu.:2016   1st Qu.:2.000
Median :2019   Median :2.000
Mean   :2019   Mean    :2.496
3rd Qu.:2022   3rd Qu.:3.000
Max.   :2025   Max.    :4.000
```

## Data Structure

### Date Column Creation

- Combines Year and Quarter columns into a proper Date column for better temporal analysis
- Converts quarters to actual dates (Q1 = January 1st, Q4 = October 1st)

```
df <- raw_amp_df

# Create a meaningful Date column by combining Year and Quarter
# Convert quarter to actual dates for better temporal analysis
df$Date <- as.Date(paste(df$Year, (df$Quarter - 1) * 3 + 1, "01", sep = "-"))
```

## Drug Category Classification

- Creates meaningful drug categories by analyzing FDA Product Names
- Categories include: Fluoride Supplements, Pain Management, Antibiotics, Topical Treatments, Respiratory, OTC Pain Relief, and Other
- Cleans up labeler company names by removing excessive spacing

```
# Create drug category classification from FDA Product Name
# Extract drug categories and clean up labeler names
df$Drug_Category <- case_when(
  str_detect(toupper(df$FDA.Product.Name), "SODIUM FLUORIDE|FLUORITAB") ~ "Fluoride Supplement",
  str_detect(toupper(df$FDA.Product.Name), "VICODIN|PAIN") ~ "Pain Management",
  str_detect(toupper(df$FDA.Product.Name), "ANTIBIOTIC|OXACILLIN|PENICILLIN") ~ "Antibiotics",
  str_detect(toupper(df$FDA.Product.Name), "CREAM|LOTION|OINTMENT") ~ "Topical Treatments",
  str_detect(toupper(df$FDA.Product.Name), "COUGH|EXPECTORANT") ~ "Respiratory",
  str_detect(toupper(df$FDA.Product.Name), "ASPIRIN|IBUPROFEN|ACETAMINOPHEN|NAPROXEN") ~ "OTC Pain Relief",
  TRUE ~ "Other"
)

# Clean up labeler names (remove excessive spacing and formatting)
df$Labeler_Clean <- str_trim(str_replace_all(df$Labeler.Name, "\\s+", " "))
```

## Grouping

```
CategoryAndStatus <- group_by(df, Status, Drug_Category) %>%
  summarise(count = n())
```

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

## C. Cleaning: Data Types and Missing Values

### Handling Missing Values

```
# We'll drop all Markdown values with select( - ...)
# and replace NA values with the mean of Unemployment and CPI.
# features2<-features%>%select(-c(MarkDown1:MarkDown5))%>%
  # mutate(Unemployment = replace_na(Unemployment,mean(Unemployment, na.rm = TRUE)),
    #CPI = replace_na(CPI,mean(CPI, na.rm = TRUE)))

# summary(features2)
```

## F. Publishing

```
summary(df)
```

Labeler.Name	NDC	FDA.Product.Name	Status
Length:2031677	Length:2031677	Length:2031677	Length:2031677
Class :character	Class :character	Class :character	Class :character
Mode :character	Mode :character	Mode :character	Mode :character

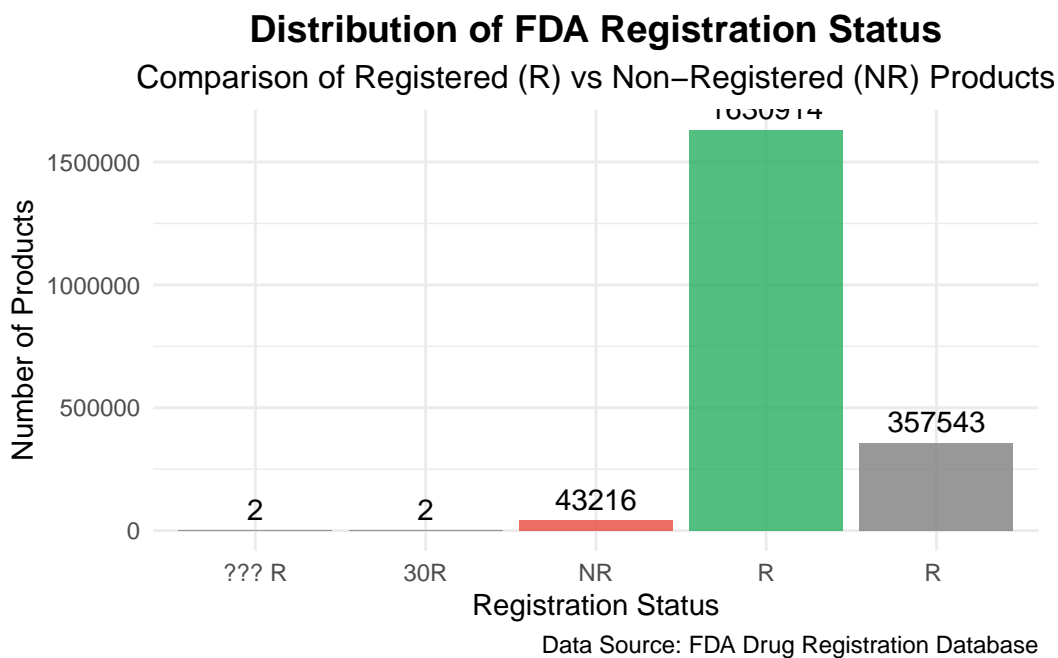
Year	Quarter	Date	Drug_Category
Min. :2013	Min. :1.000	Min. :2013-01-01	Length:2031677
1st Qu.:2016	1st Qu.:2.000	1st Qu.:2016-04-01	Class :character
Median :2019	Median :2.000	Median :2019-07-01	Mode :character
Mean :2019	Mean :2.496	Mean :2019-05-20	
3rd Qu.:2022	3rd Qu.:3.000	3rd Qu.:2022-07-01	
Max. :2025	Max. :4.000	Max. :2025-04-01	
Labeler_Clean			
Length:2031677			
Class :character			
Mode :character			

```
df%>%write.csv(csv_OUT_FILE,row.names = FALSE)
```

## E. Verifying / Exploring

### Distribution of Registration Status

```
p1 <- ggplot(df, aes(x = Status, fill = Status)) +  
  geom_bar(stat = "count", alpha = 0.8) +  
  geom_text(stat = "count", aes(label = after_stat(count)), vjust = -0.5) +  
  labs(title = "Distribution of FDA Registration Status",  
       subtitle = "Comparison of Registered (R) vs Non-Registered (NR) Products",  
       x = "Registration Status",  
       y = "Number of Products",  
       caption = "Data Source: FDA Drug Registration Database") +  
  scale_fill_manual(values = c("NR" = "#E74C3C", "R" = "#27AE60")) +  
  theme_minimal() +  
  theme(plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),  
        plot.subtitle = element_text(hjust = 0.5, size = 12),  
        legend.position = "none")  
  
print(p1)
```



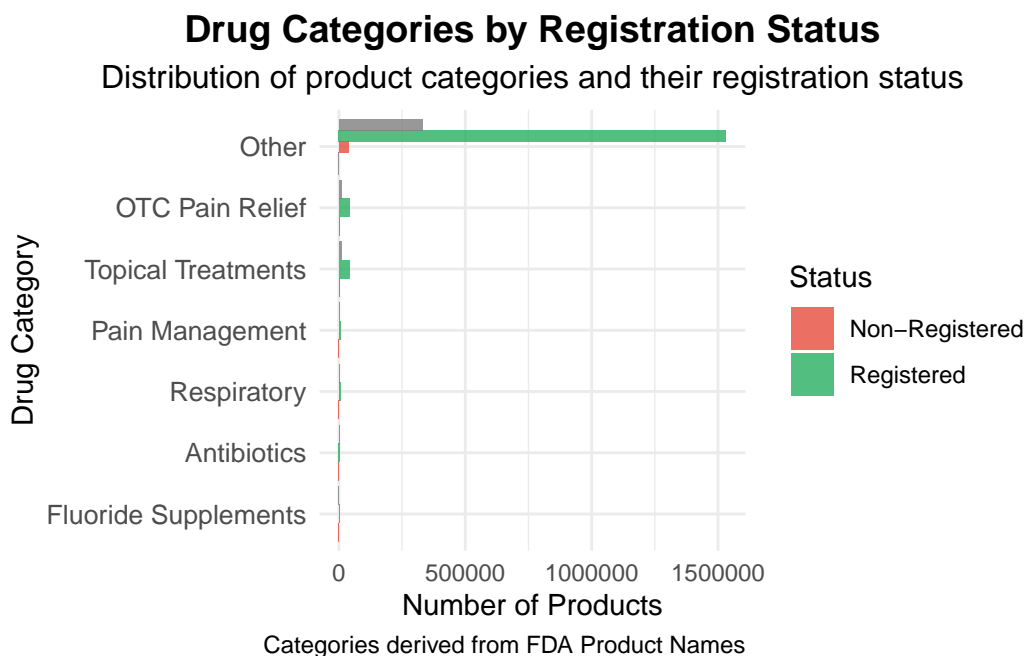
```
# Graph 2: Drug Categories by Registration Status  
p2 <- ggplot(df, aes(x = reorder(Drug_Category, Drug_Category, function(x) length(x)),
```

```

      fill = Status)) +
geom_bar(position = "dodge", alpha = 0.8) +
coord_flip() +
labs(title = "Drug Categories by Registration Status",
      subtitle = "Distribution of product categories and their registration status",
      x = "Drug Category",
      y = "Number of Products",
      fill = "Status",
      caption = "Categories derived from FDA Product Names") +
scale_fill_manual(values = c("NR" = "#E74C3C", "R" = "#27AE60"),
                  labels = c("NR" = "Non-Registered", "R" = "Registered")) +
theme_minimal() +
theme(plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),
      plot.subtitle = element_text(hjust = 0.5, size = 12),
      axis.text.y = element_text(size = 10))

print(p2)

```



```

# Graph 3: Top Pharmaceutical Companies (Top 8)
top_labelers <- names(head(sort(table(df$Labeler_Clean), decreasing = TRUE), 8))
df_top <- df[df$Labeler_Clean %in% top_labelers, ]

p3 <- ggplot(df_top, aes(x = reorder(Labeler_Clean, Labeler_Clean, function(x) length(x)),

```



```

    fill = Status)) +
geom_bar(stat = "count", alpha = 0.8) +
coord_flip() +
labs(title = "Product Count by Top Pharmaceutical Companies",
      subtitle = "Leading companies by number of products in the database",
      x = "Pharmaceutical Company",
      y = "Number of Products",
      fill = "Registration Status",
      caption = "Top 8 companies by product count") +
scale_fill_manual(values = c("NR" = "#E74C3C", "R" = "#27AE60"),
                  labels = c("NR" = "Non-Registered", "R" = "Registered")) +
theme_minimal() +
theme(plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),
      plot.subtitle = element_text(hjust = 0.5, size = 12),
      axis.text.y = element_text(size = 9))

print(p3)

```



```

# Graph 4: Quarterly Registration Timeline
quarterly_summary <- df %>%
  group_by(Date, Status) %>%
  summarise(count = n(), .groups = 'drop')

```

```
p4 <- ggplot(quarterly_summary, aes(x = Date, y = count, fill = Status)) +
  geom_col(position = "stack", alpha = 0.8, width = 50) +
  geom_text(aes(label = count), position = position_stack(vjust = 0.5),
    color = "white", size = 4, fontface = "bold") +
  labs(title = "Pharmaceutical Product Registration Timeline",
    subtitle = "Quarterly distribution of registered vs non-registered products in 2013",
    x = "Quarter",
    y = "Number of Products",
    fill = "Registration Status",
    caption = "Data shows Q1 and Q4 of 2013") +
  scale_fill_manual(values = c("NR" = "#E74C3C", "R" = "#27AE60"),
    labels = c("NR" = "Non-Registered", "R" = "Registered")) +
  scale_x_date(date_labels = "%Y Q%q", date_breaks = "3 months") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),
    plot.subtitle = element_text(hjust = 0.5, size = 12),
    axis.text.x = element_text(angle = 45, hjust = 1))

print(p4)
```

## Pharmaceutical Product Registration Timeline

Quarterly distribution of registered vs non-registered products in 2013

