Assignment 02 Data Exploration

BQOM 2578 | Data Mining

Theresa Wohlever

2025-09-17

Table of contents

Executive Summary

This dataset is from Medicaid Drug AMP Reporting

The Data Cleaning and Preparation includes creating a Date column generated from the corresponding year and Quarter columns and categorizing drug types.

Of the drugs that can be categorized, the most commonly occurring are those addressing hypertension and cardiac disease. This category is heavily invested in by labeler companies, including 28 of the top 30 companies.

Drug AMP Reporting - Quarterly

The dataset is from Medicaid Drug AMP Reporting and described there:

> *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).*

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

```r
base_FILENAME <- "DrugAMPReportingQuarterly022025" ## tiny

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

raw_amp_df <- read.csv(csv_FILE, stringsAsFactors = FALSE)
```

Data Discovery

Review Head, Tail, Dimensions, Column Headers, and Summary Statistics

```
                            Labeler.Name         NDC
1 FLUORITAB CORPORATION                  00288110601
2 FLUORITAB CORPORATION                  00288110602
3 FLUORITAB CORPORATION                  00288110610
4 FLUORITAB CORPORATION                  00288110699
5 FLUORITAB CORPORATION                  00288220101
6 FLUORITAB CORPORATION                  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
```

```
                 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
```

```
[1] 2031677       6
```

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

```
 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 Cleaning

Random subset to improve analysis execution speed Cleans up labeler company names by removing excessive spacing Cleans up Status

```
df <-raw_amp_df # sample_n(raw_amp_df, 10000)
names(df)[names(df) == "FDA.Product.Name"] <- "Product"


## Clean Up Status
df <- df %>% mutate(Status = case_when(
    grepl("^R\\s*$", Status) ~ "R",
```

```
    grepl("NR", Status) ~ "NR",
    TRUE ~ NA
  ))


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

Data Preparation

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)

```
# 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

```
## Update with Dose
df <- df %>%
  mutate(dose = str_extract(Product, "\\d+(?=\\s*MG)"))


## Prep for Categories
products <- df$Product
escaped_products <- gsub("([][{}()+*^$.|?\\])", "\\\\\\1", toupper(products))
product_regex <- paste(escaped_products, collapse = "|")
has_products <- nzchar(product_regex) # TRUE if pattern not empty

if (product_regex != "") {
  df <- df %>%
   mutate(
      ProductUpper = toupper(Product),
```

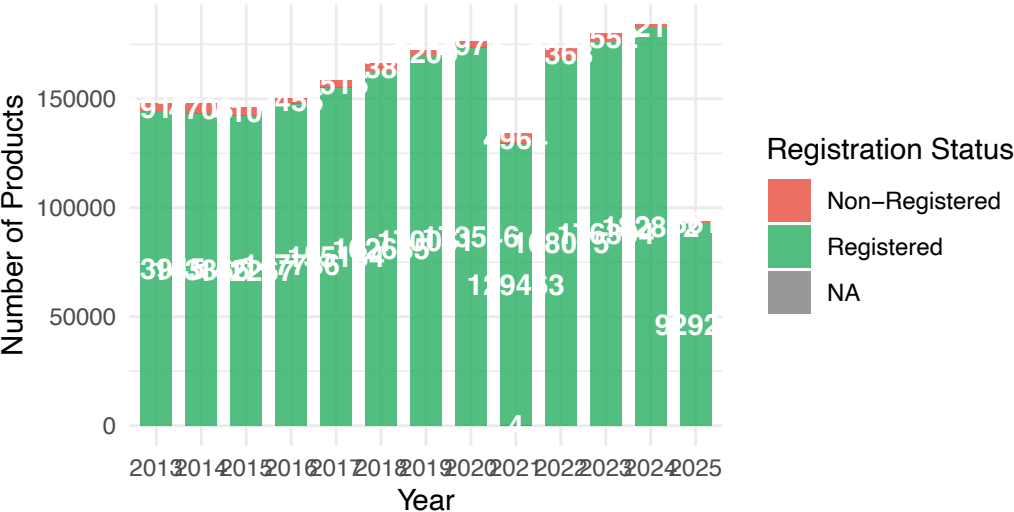
```
      Drug_Category = case_when(
        # All products in the text file
        has_products &

        # Opioid/Combination Analgesic
        str_detect(ProductUpper, "OXYCOD|HYDROCOD|MORPHIN|TRAMADOL|CODEINE|METHADONE|BUPREN

        # NSAIDs/Non-opioid Analgesic/Antipyretic
        str_detect(ProductUpper, "IBUPROFEN|NAPROXEN|ACETAMINOPHEN|APAP|ASPIRIN|DICLOFENAC|

        # Antidiabetic
        str_detect(ProductUpper, "METFORMIN|GLIMEPIRIDE|GLIPIZIDE|GLYBURIDE|JANUMET|GLUCOVA

        # Statins/Cholesterol
        str_detect(ProductUpper, "ATORVASTATIN|SIMVASTATIN|LOVASTATIN|PRAVASTATIN|FLUVASTAT

        # Antihypertensive/Cardiac
        str_detect(ProductUpper, "LISINOPRIL|ENALAPRIL|LOSARTAN|VALSARTAN|QUINAPRIL|TELMISA

        # Diuretic/Electrolyte/Laxative/IV Fluid
        str_detect(ProductUpper, "FUROSEMIDE|BUMETANIDE|HYDROCHLOROTHIAZIDE|HCTZ|CHLORTHALI

        # Psychotropic/Antidepressant/Neuro
        str_detect(ProductUpper, "SERTRALINE|FLUOXETINE|ESCITALOPRAM|CITALOPRAM|PAROXETINE|

        # Antipsychotic/Neuroleptic
        str_detect(ProductUpper, "QUETIAPINE|SEROQUEL|OLANZAPINE|ZYPREXA|RISPERIDONE|RISPER

        # Anticonvulsant/Epilepsy/Neurologic
        str_detect(ProductUpper, "LEVETIRACETAM|LAMOTRIGINE|TOPIRAMATE|DIVALPROEX|VALPROIC|

        # Antimicrobial/Antiinfective/Oncology/Immune
        str_detect(ProductUpper, "AMOXICILLIN|CLAVULANATE|PENICILLIN|OXACILLIN|NAFCILLIN|AM

        # Respiratory/ENT/Antihistamine
        str_detect(ProductUpper, "ALBUTEROL|IPRATROPIUM|LEVOSALBUTAMOL|FORMOTEROL|BUDESONID

        # Hormonal/Endocrine
        str_detect(ProductUpper, "LEVOTHYROXINE|LIOTHYRONINE|THYROID|ESTRADIOL|PROGESTERONE

        # GI/GERD/Acid/IBD
        str_detect(ProductUpper, "OMEPRAZOLE|PANTOPRAZOLE|LANSOPRAZOLE|ESOMEPRAZOLE|RABEPRA
```

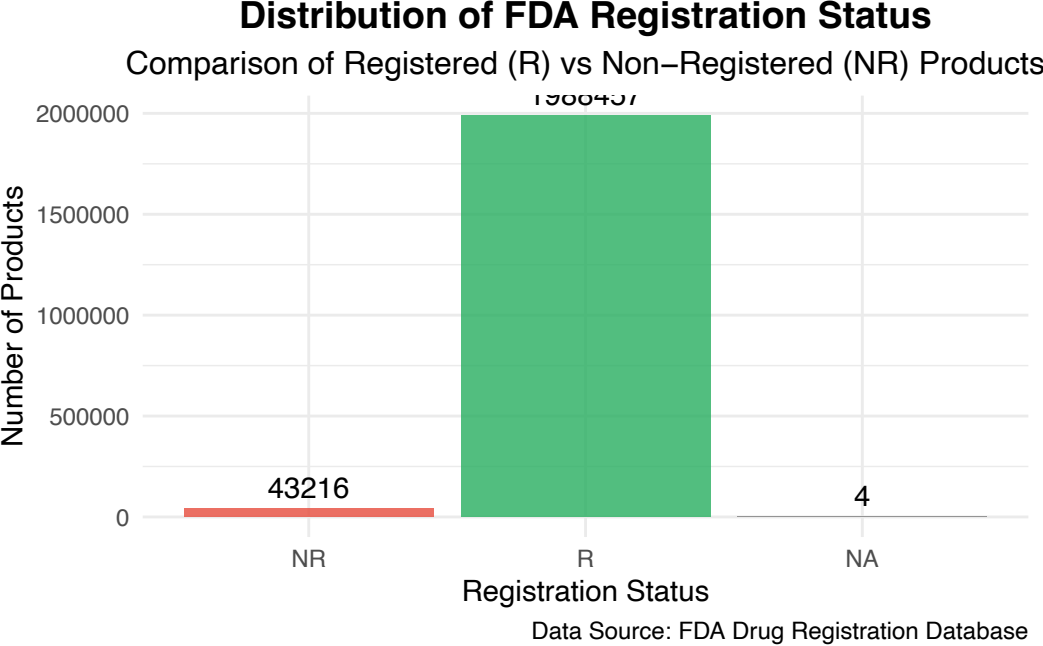
```
        # Anticoagulant/Platelet/Thrombolytic
        str_detect(ProductUpper, "WARFARIN|COUMADIN|HEPARIN|ENOXAPARIN|MRXABAN|APIXABAN|XAR

        # Smoking Cessation
        str_detect(ProductUpper, "NICOTINE|NICORETTE|NICODERM|NICOTROL|GUM|LOZENGE|PATCH|VA

        # Reproductive/Contraceptive
        str_detect(ProductUpper, "DROSPIRENONE|NORETHINDRONE|NORGESTIMATE|LEVONORGESTREL|ET

        # Vitamins/Nutrition
        str_detect(ProductUpper, "VITAMIN|FOLIC|FOLATE|FOLIVANE|CITRANATAL|FERROUS|CYANOCOB

        # Dermatologic/Topical
        str_detect(ProductUpper, "HYDROCORTISONE|CLOBETASOL|BETAMETHASONE|TRIAMCINOLONE|DES

        # Urologic/Bladder
        str_detect(ProductUpper, "OXYBUTYNIN|TOLTERODINE|SOLIFENACIN|MIRABEGRON|TAMSULOSIN|

        # Miscellaneous/Other
        TRUE ~ "Other"
      )
    ) %>%
    select(-ProductUpper)

} else{ stop("Product Regex Empty!") }

Others <- df %>% filter(Drug_Category == "Other") %>% distinct(Product) %>% pull(Product)
df_noOther <- df %>% filter(Drug_Category != "Other")
```

```
# Graph 2: Drug Categories by Registration Status
p2 <- ggplot(df_noOther, 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"),
```

Pharmaceutical Product Registration Timeline

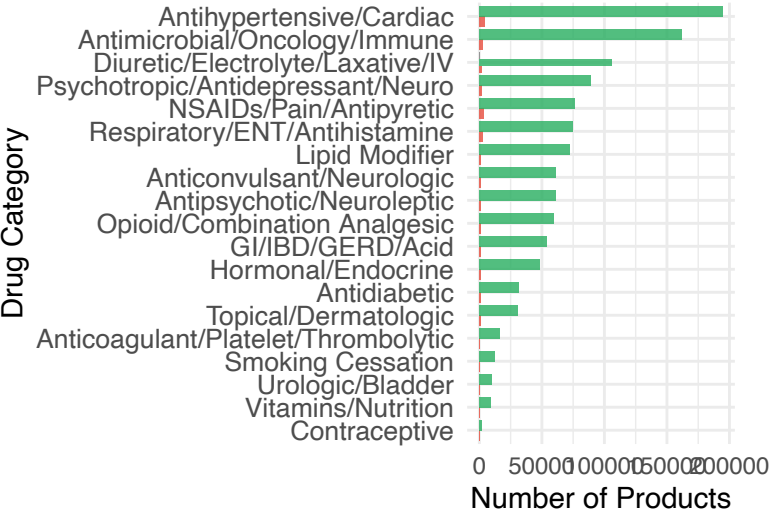
Yearly distribution of registered vs non–registered products




Distribution of FDA Registration Status
Comparison of Registered (R) vs Non−Registered (NR) Products
1988457
2000000
1500000
1000000
500000
0
Number of Products
43216
4
NR
R
NA
Registration Status
Data Source: FDA Drug Registration Database

Drug Categories by Registration Status

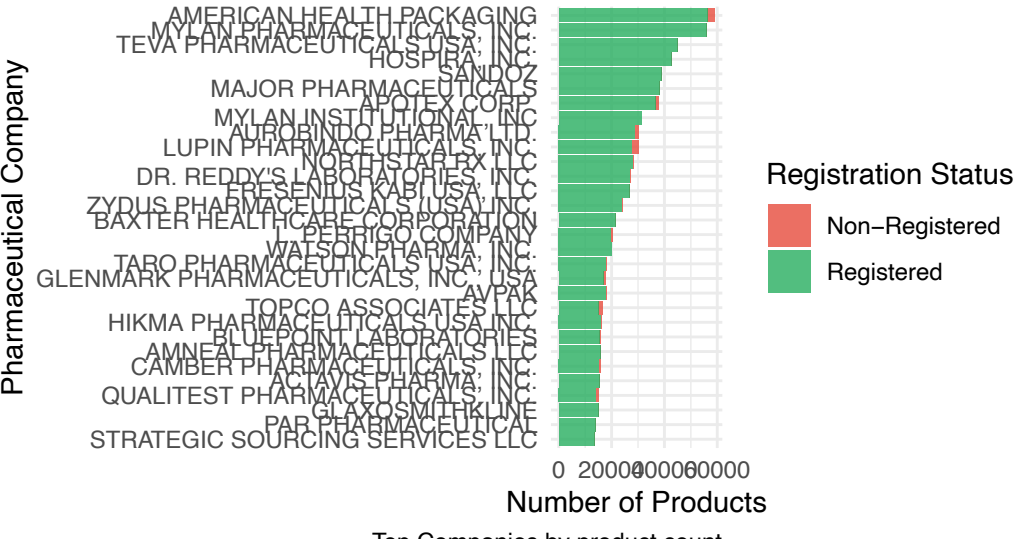
Distribution of product categories and their registration status



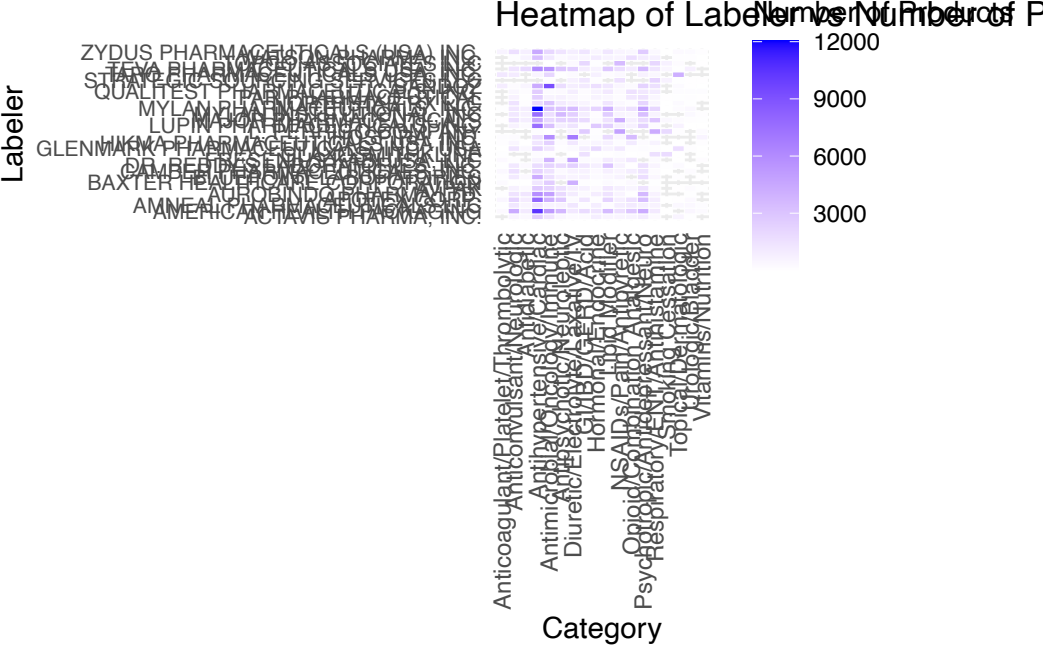
Categories derived from FDA Product Names

Product Count by Top Pharmaceutical Companie

Leading companies by number of products in the database



Top Companies by product count


Heatmap of Labeler vs Number of P
Number Of Products
12000
9000
6000
3000
Labeler
ZYDUS PHARMACEUTICALS (USA) INC.
TEVA PHARMACEUTICALS USA INC
STAT RX USA LLC
QUALITEST PHARMACEUTICALS INC
MYLAN PHARMACEUTICALS INC.
LUPIN PHARMACEUTICALS INC.
HIKMA PHARMACEUTICALS USA INC.
GLENMARK PHARMACEUTICALS INC., USA
DR. REDDY'S LABORATORIES INC
CAMBER PHARMACEUTICALS, INC.
BAXTER HEALTHCARE CORPORATION
AMNEAL PHARMACEUTICALS
AMERICAN HEALTH PHARMA, INC.
ACTAVIS PHARMA, INC.
Anticoagulant/Platelet/Thrombolytic
Antidiabetic
Antihypertensive/Vasodilator
Antimicrobial/Antiviral
Diuretic/Electrolyte/Acid/Base
Hormone/Hormone Modulator
NSAID's/Pain/Anti-Inflammatory
Opioid
Psychotherapeutic/Antidepressant/Antipsychotic
Respiratory/Cough/Cold
Topical/Dermatologic
Vitamins/Nutrition
Category