Final

Claudia Bardales

2025-04-07

```
my_data <- read.csv("my_data_clean_fl.csv")
```

#Libraries

```
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.5.1     ✔ tibble    3.2.1
## ✔ lubridate 1.9.4     ✔ tidyr     1.3.1
## ✔ purrr     1.0.2
## ── Conflicts ──────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force
all conflicts to become errors

library(cowplot)

##
## Attaching package: 'cowplot'
##
## The following object is masked from 'package:lubridate':
##
##     stamp

library(ggradar)
library(tibble)
library(stringr)
library(scales)

##
## Attaching package: 'scales'
##
## The following object is masked from 'package:purrr':
##
##     discard
##
## The following object is masked from 'package:readr':
##
##     col_factor
```

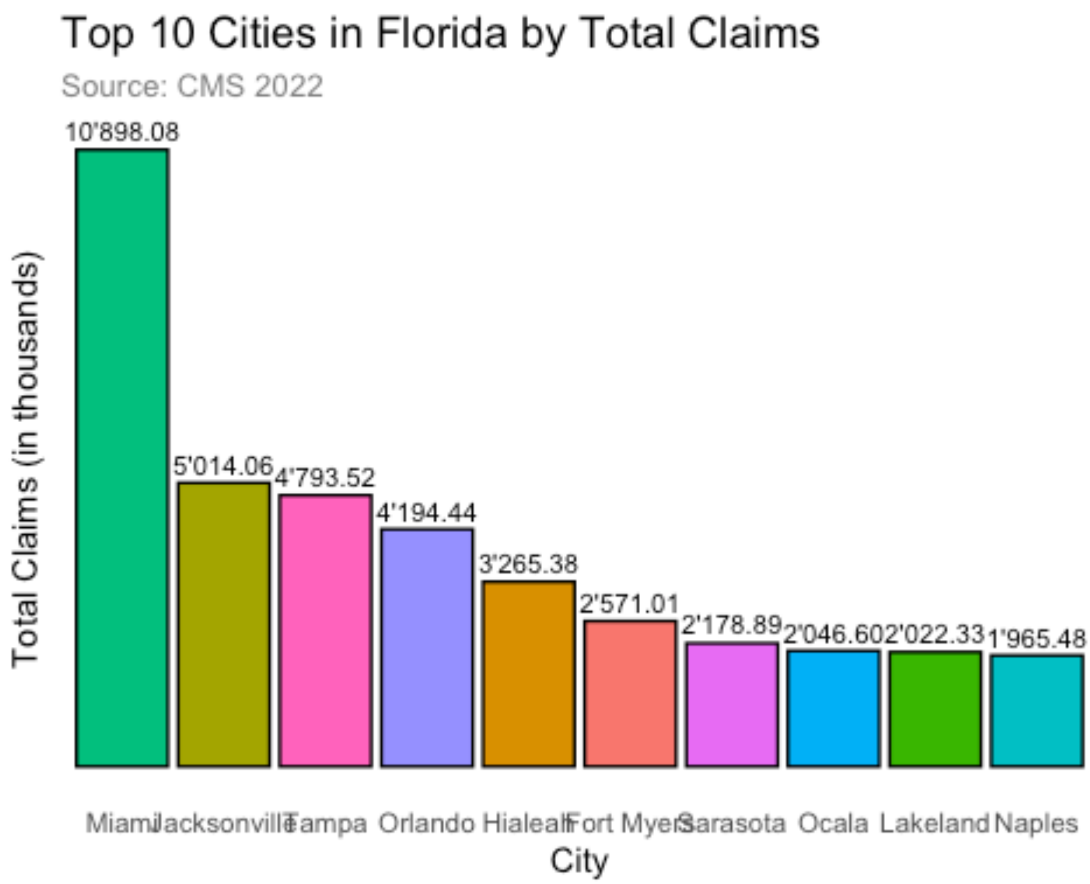
#top 10 cities that made claims

```
# Cities of interest
cities_of_interest <- c("Miami", "Jacksonville", "Tampa", "Orlando",
"Hialeah",
                        "Fort Myers", "Sarasota", "Ocala", "Lakeland",
"Naples")
filtered_data <- my_data |>
  filter(Prscrbr_City %in% cities_of_interest)
```

#A chart that shows the distribution of a single categorical variable

```
#total claims by city

filtered_data_by_city<-filtered_data|>
  select(Prscrbr_City,Tot_Clms)|>
  group_by(Prscrbr_City)|>
  summarise(total_claims_per_city=sum(Tot_Clms)/1000)|>
  arrange(desc(total_claims_per_city))

#set min and max values

min_val <- min(filtered_data_by_city$total_claims_per_city, na.rm = TRUE)
max_val <- max(filtered_data_by_city$total_claims_per_city, na.rm = TRUE)

#plot

a <- ggplot(data = filtered_data_by_city,
            aes(x = reorder(Prscrbr_City, -total_claims_per_city),
                y = total_claims_per_city,
                fill = Prscrbr_City)) +
  geom_bar(stat = "identity", color = "black") +
  #add labels
  geom_text(aes(label = label_number(accuracy = 0.01, big.mark =
"'")(total_claims_per_city)),
            vjust = -0.5, size = 3) +
  labs(title = "Top 10 Cities in Florida by Total Claims",
       subtitle = "Source: CMS 2022",
       x = "City",
       y = "Total Claims (in thousands)") +
  #theme
  theme_minimal() +
  theme(
        plot.title = element_text(size = 14),
        plot.subtitle = element_text(size = 10, color = "gray50"),
        legend.position = "none",
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.ticks.y = element_blank(),
        axis.text.y = element_blank()
```

```
        ) +
  guides(fill = "none")

a
```



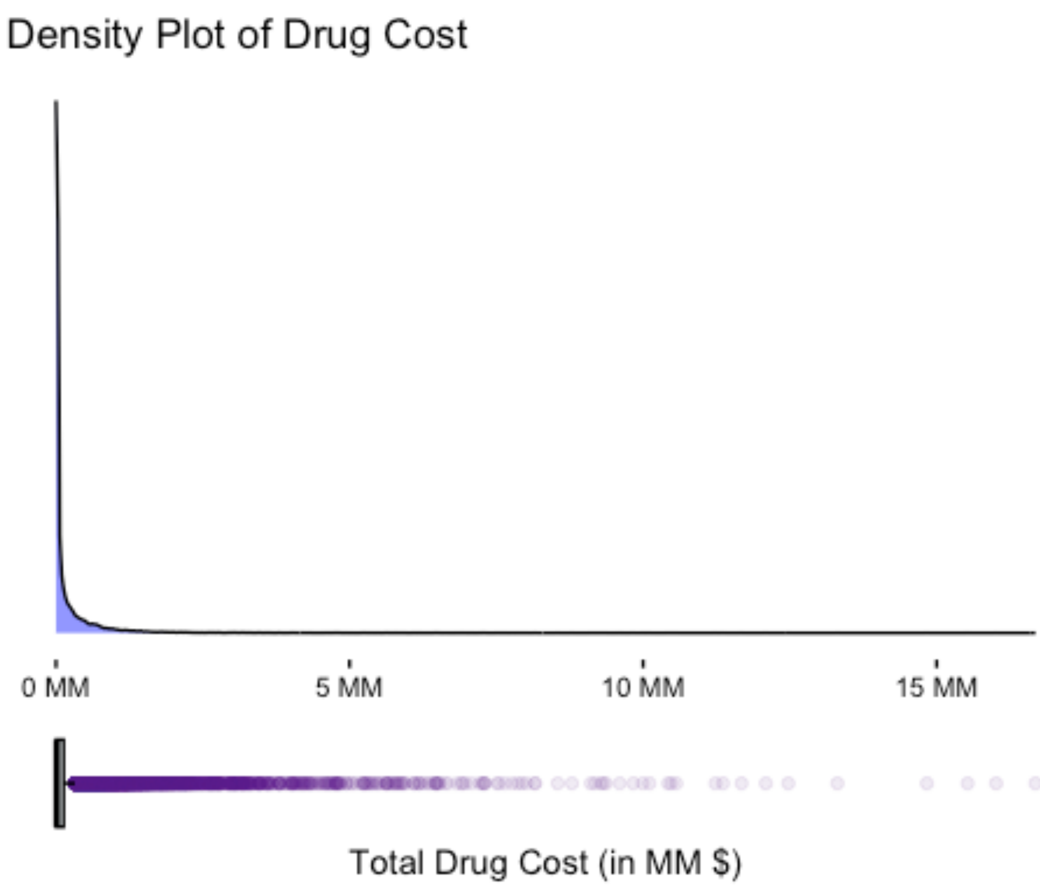

```
#Top 10 cities filing Medicare claims related to drug medication for Medicare beneficiaries. In first place is the city of Miami, with around 10 million accumulated claims in 2022, followed by Jacksonville and Tampa with 5 million and 4 million claims, respectively. These top ten cities collectively account for around 38 million claims, out of Florida's total of 118 million claims.
```

#A chart that shows the distribution of a single quantitative variable

```
filtered_data_by_city_Cost<-filtered_data|>
  select(Prscrbr_City, Tot_Drug_Cst)
#filtered_data_by_city_Cost

min_a<-min(filtered_data_by_city_Cost$Tot_Drug_Cst)
min_a

## [1] 0

max_a<-max(filtered_data_by_city_Cost$Tot_Drug_Cst)
max_a
```

```
## [1] 16683069

summary(filtered_data_by_city_Cost$Tot_Drug_Cst)

##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##        0     1901    12585   199564   128693 16683069

#outlier
Q1 <- quantile(filtered_data_by_city_Cost$Tot_Drug_Cst, 0.25, na.rm = TRUE)
Q3 <- quantile(filtered_data_by_city_Cost$Tot_Drug_Cst, 0.75, na.rm = TRUE)

IQR_value <- Q3 - Q1

lower_bound <- Q1 - 1.5 * IQR_value
upper_bound <- Q3 + 1.5 * IQR_value

# Count outliers
left_outliers <- sum(filtered_data_by_city_Cost$Tot_Drug_Cst < lower_bound,
na.rm = TRUE)
right_outliers <- sum(filtered_data_by_city_Cost$Tot_Drug_Cst > upper_bound,
na.rm = TRUE)

#There is no outlier to the left, there are 4843 outiler to the right

library(scales)
dens <- ggplot(filtered_data_by_city_Cost, aes(x = Tot_Drug_Cst)) +
  geom_density(fill = "blue", alpha = 0.5) +
  scale_x_continuous(
    #limits = c(0, 1e6),
    labels = label_number(scale = 1e-6, big.mark = "'", suffix = " MM")  #
Format labels
  ) +
  labs(
    title = "Density Plot of Drug Cost",
    x = NULL,
    y = NULL
  ) +
  theme_minimal() +
  theme(
    axis.ticks.x = element_line(),   # Ensure tick marks are visible
    axis.text.x = element_text(color = "black"),  # Keep tick labels visible
    axis.ticks.y = element_blank(),
    axis.text.y = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank()
  )

box<-ggplot(filtered_data_by_city_Cost, aes(x = Tot_Drug_Cst)) +
  geom_boxplot(fill = "skyblue", color = "black", alpha = 0.1, outlier.color
= "purple4")  +
```

```r
  labs(
       x = "Total Drug Cost (in MM $)") +
  theme_minimal()+
  theme(
    axis.ticks.y = element_blank(),
    axis.text.y = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    axis.text.x = element_blank(),
    axis.line.x = element_blank())

#join graphs
plot_grid(dens,box, ncol=1, rel_heights = c(0.80,0.20) , align='v',
axis='lr')
```

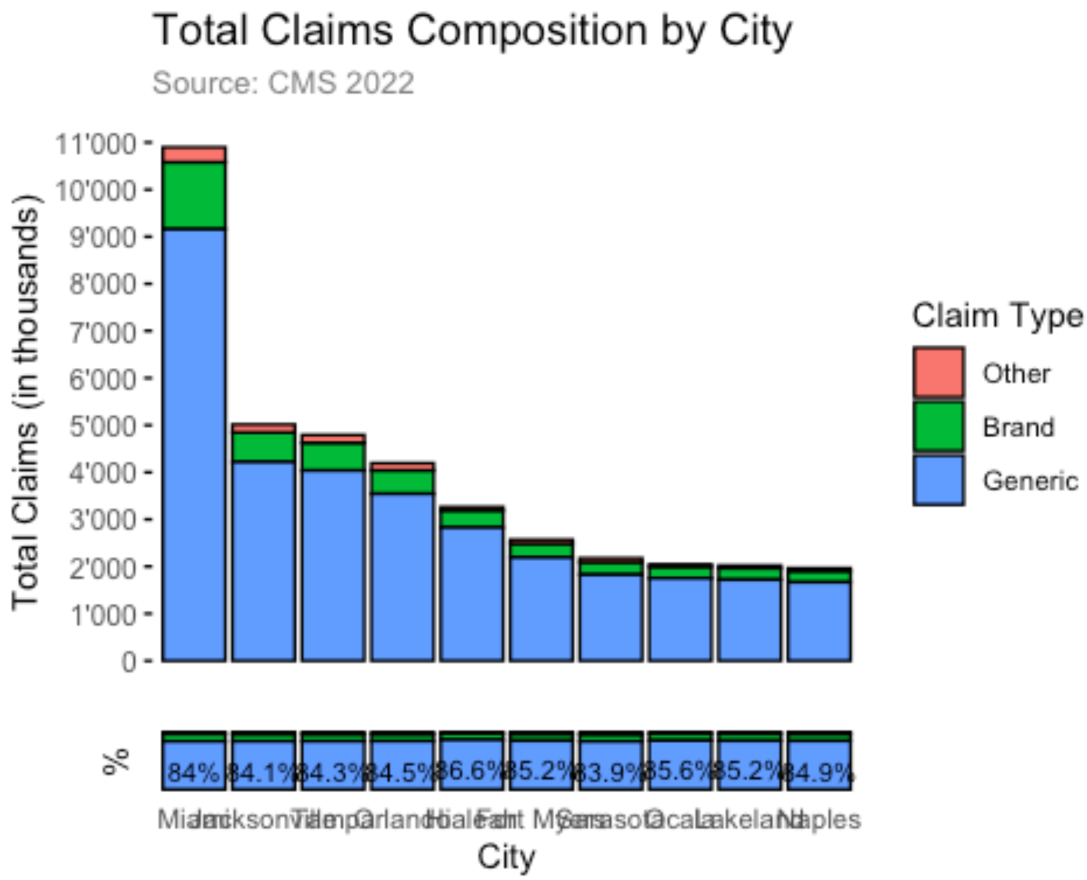



```r
#This density plot represents the distribution of total drug costs under
Medicare Part D. On the x-axis, we can observe that the total cost ranges
from 0 to 15 million, while the density shows the frequency of total drug
costs. We can see high density near 0 million, with several outliers as the
cost increases. This graph displays a right-skewed distribution.
```

#A chart that shows the distribution of two categorical variables

```r
cat_cat <- filtered_data |>
  select(Prscrbr_City, Brnd_Tot_Clms, Gnrc_Tot_Clms, Tot_Clms) |>
  group_by(Prscrbr_City) |>
  summarise(
    Brand = sum(Brnd_Tot_Clms, na.rm = TRUE),
    Generic = sum(Gnrc_Tot_Clms, na.rm = TRUE),
    Total = sum(Tot_Clms, na.rm = TRUE),
    .groups = "drop"
  ) |>
  mutate(Other = Total - Brand - Generic)|>
  pivot_longer(cols=c(Brand, Generic, Other),
              names_to = "Claim_Type",
              values_to = "Total_Claims")

cat_cat<- cat_cat|>
  select(Prscrbr_City,Claim_Type,Total_Claims)

#maintain order:
cat_cat$Claim_Type <- factor(cat_cat$Claim_Type, levels = c( "Other",
"Brand", "Generic"))

#porcentage
cat_cat_percent <- cat_cat |>
  group_by(Prscrbr_City) |>
  mutate(
    Percent_Claims = Total_Claims / sum(Total_Claims) * 100
  ) |>
  ungroup()

cat_cat_thousands <- cat_cat |>
  mutate(Total_Claims = Total_Claims / 1000)

max_val <- 11000
y_breaks <- pretty(c(0, max_val), n = 10)

c <- ggplot(cat_cat_thousands,
            aes(x = reorder(Prscrbr_City, -Total_Claims),
                y = Total_Claims,
                fill = Claim_Type)) +
  geom_bar(stat = "identity", color = "black") +

  labs(
    title = "Total Claims Composition by City",
    subtitle = "Source: CMS 2022",
 #   x = "City",
    y = "Total Claims (in thousands)",
    fill = "Claim Type"
  ) +
```

```r
  scale_y_continuous(
    breaks = y_breaks,
    limits = c(0, max(y_breaks)),
    labels = label_number(accuracy = 1, big.mark = "'")
  ) +

  theme_minimal() +
  theme(
    plot.title = element_text(size = 14),
    plot.subtitle = element_text(size = 10, color = "gray50"),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    axis.ticks.y = element_line(),
    axis.text.y = element_text(size = 9),
    axis.text.x = element_blank(),
    axis.title.x = element_blank()
  )

d <- ggplot(cat_cat_percent,
            aes(x = reorder(Prscrbr_City, -Total_Claims),
                y = Percent_Claims,
                fill = Claim_Type)) +
  geom_bar(stat = "identity", color = "black") +
  labs(x = "City",
       y= "%")+
  geom_text(
    data = cat_cat_percent |> filter(Claim_Type %in% c("Generic")),
    aes(label = paste0(round(Percent_Claims, 1), "%")),
    position = position_stack(vjust = 0.4),
    size = 3, color = "black"
  ) +

  theme_minimal() +
  theme(
    plot.title = element_text(size = 14),
    plot.subtitle = element_text(size = 10, color = "gray50"),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    axis.ticks.y = element_blank(),
    axis.text.y = element_blank(),
    axis.ticks.x = element_blank(),
    #axis.text.x = element_blank(),
    #axis.title.x = element_blank(),  # remove x-axis title
    #axis.title.y = element_blank(),
    legend.position = "none"
  )

plot_grid(c,d, ncol=1, rel_heights = c(0.80,0.20) , align='v', axis='lr')
```

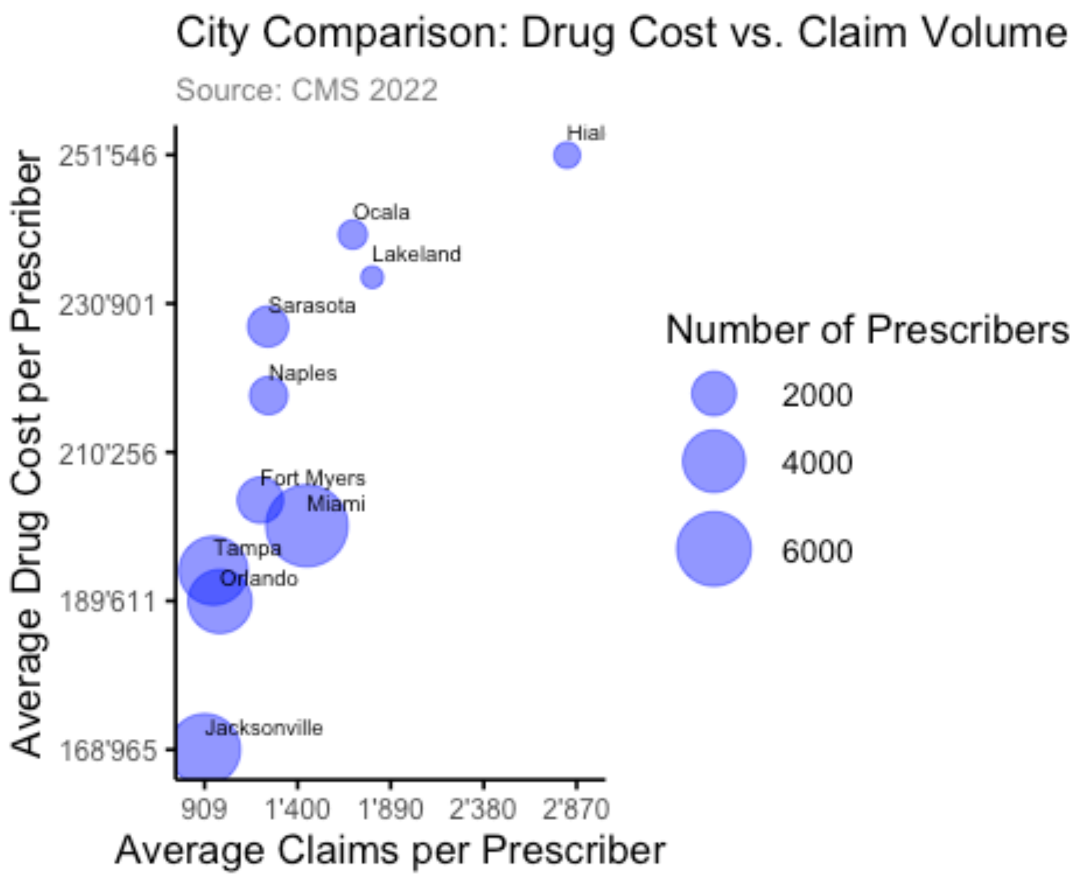



```
#In this chart, we can see the drug claims per city divided by drug type. In
all the cities, the majority of drug claims are for Generic drugs, which
account for between 83.9% and 86.6% of the total claims.
```

#A chart that shows the relationship between two quantitative variables

```
city_summary_data <- filtered_data |>
  select(Prscrbr_NPI, Prscrbr_City, Tot_Drug_Cst, Tot_Clms) |>
  group_by(Prscrbr_City) |>
  summarise(
    Drug_Cost_mean = mean(Tot_Drug_Cst, na.rm = TRUE),
    Total_Claims_mean = mean(Tot_Clms, na.rm = TRUE),
    Number_Prescribers=n(),
    .groups = 'drop')

x_max<-max(city_summary_data$Total_Claims_mean)
x_min<-min(city_summary_data$Total_Claims_mean)
y_max<-max(city_summary_data$Drug_Cost_mean)
y_min<-min(city_summary_data$Drug_Cost_mean)

lgd_min<-min(city_summary_data$Number_Prescribers)
lgd_max<-max(city_summary_data$Number_Prescribers)
```

```r
ggplot(city_summary_data, aes(
  x = Total_Claims_mean,
  y = Drug_Cost_mean,
  size = Number_Prescribers,
  label = Prscrbr_City
)) +
  geom_point(alpha = 0.5, color = "blue", stroke = 0.5) +
  geom_text(vjust = -1, size = 2.5, hjust=-0.001) +
  scale_size(range = c(3, 12)) +
  labs(
    title = "City Comparison: Drug Cost vs. Claim Volume",
    subtitle = "Source: CMS 2022",
    x = "Average Claims per Prescriber",
    y = "Average Drug Cost per Prescriber",
    size = "Number of Prescribers"
  ) +
  scale_x_continuous(
    limits = c(x_min-50, x_max + 100),
    breaks = seq(x_min, x_max + 50, length.out = 5),
    labels = label_number(big.mark = "'")
  ) +
  scale_y_continuous(
    limits = c(y_min-50, y_max + 50),
    breaks = seq(y_min, y_max + 50, length.out = 5),
    labels = label_number(big.mark = "'")
  ) +
  theme_minimal(base_size = 13) +
  theme(
    plot.title = element_text(size = 14),
    plot.subtitle = element_text(size = 10, color = "gray50"),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    axis.ticks = element_line(),
    axis.text = element_text(size = 9),
    axis.line.x = element_line(color = "black"),
    axis.line.y = element_line(color = "black")
  )
```




```
#This chart helps visualize the relationship between claims and drug cost,
based on prescriber volume across different cities.
```

#A chart that shows the distribution of a quantitative variable across categories of a categorical variable

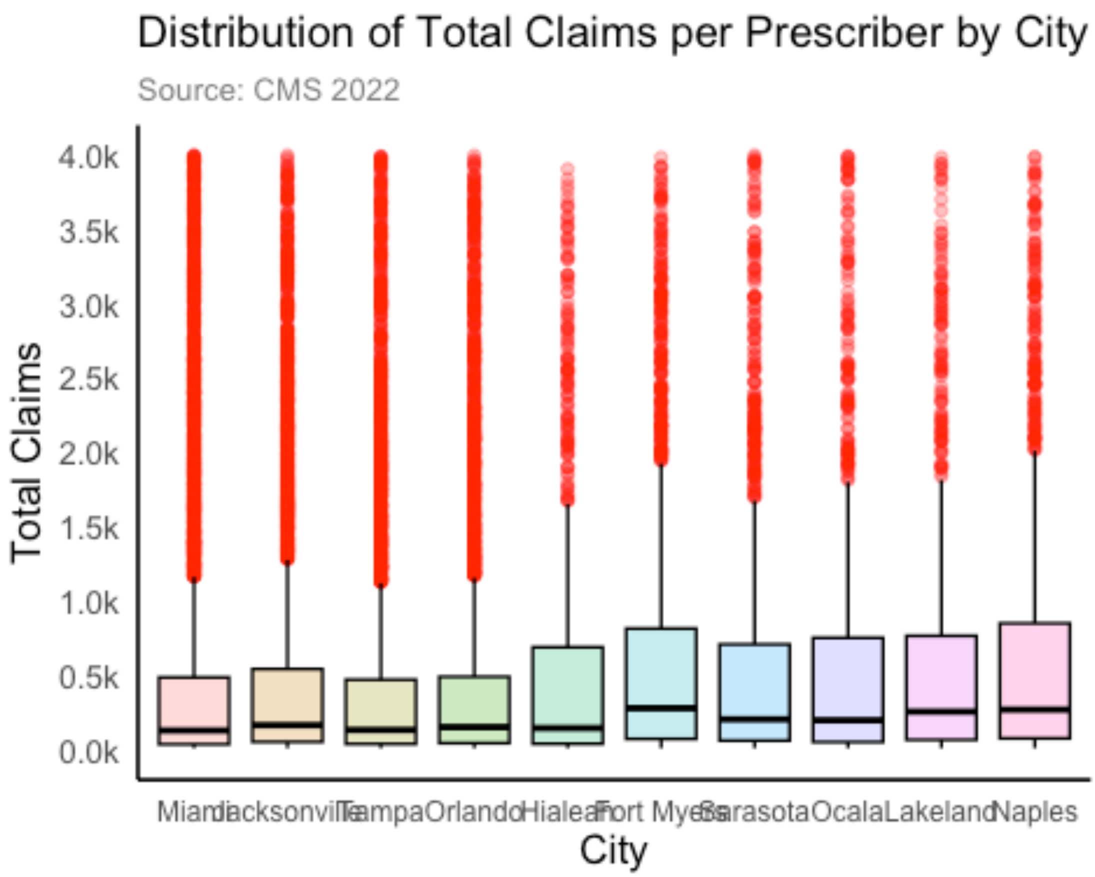
```
ordered_cities <- filtered_data_by_city$Prscrbr_City

eff_per_city_2 <- filtered_data |>
  select(Prscrbr_NPI, Prscrbr_City, Tot_Clms) |>
  mutate(Prscrbr_City = factor(Prscrbr_City, levels = ordered_cities))

eff_per_city_summary <- eff_per_city_2 |>
  group_by(Prscrbr_City) |>
  summarise(
    min = min(Tot_Clms, na.rm = TRUE),
    q1 = quantile(Tot_Clms, 0.25, na.rm = TRUE),
    q2 = quantile(Tot_Clms, 0.50, na.rm = TRUE),
    q3 = quantile(Tot_Clms, 0.75, na.rm = TRUE),
    max = max(Tot_Clms, na.rm = TRUE),
    iqr = IQR(Tot_Clms, na.rm = TRUE),
    .groups = "drop"
```

```
  )
eff_per_city_summary
```

```
## # A tibble: 10 × 7
##    Prscrbr_City   min    q1    q2    q3   max   iqr
##    <fct>        <int> <dbl> <dbl> <dbl> <int> <dbl>
##  1 Miami           11  44    170    889  67440  845
##  2 Jacksonville    11  58    190    714  47564  656
##  3 Tampa           11  45    159    673  37877  628
##  4 Orlando         11  50    184    720. 32798  670.
##  5 Hialeah         11  51.2  300   2612. 52658 2560.
##  6 Fort Myers      11  83    325   1157  33388 1074
##  7 Sarasota        11  73    256.  1082. 27685 1008.
##  8 Ocala           11  67    296   1408  49732 1341
##  9 Lakeland        11  80    340   1360. 29843 1280.
## 10 Naples          11  90.8  329   1283. 19975 1192.
```

```
ggplot(eff_per_city_2, aes(x = Prscrbr_City, y = Tot_Clms, fill =
Prscrbr_City)) +
  geom_boxplot(color = "black", alpha = 0.3, outlier.color = "red") +
  scale_y_continuous(
    breaks = seq(0, 4000, by = 500),
    limits = c(0, 4000),
    labels = label_number(scale = 1e-3, suffix = "k", big.mark = "'")
  ) +
  labs(
    title = "Distribution of Total Claims per Prescriber by City",
    subtitle = "Source: CMS 2022",
    x = "City",
    y = "Total Claims"
  ) +
  theme_minimal(base_size = 13) +
  theme(
    plot.title = element_text(size = 14),
    plot.subtitle = element_text(size = 10, color = "gray50"),
    axis.text.x = element_text(size = 9, vjust = 0.5),
    axis.line = element_line(color = "black"),
    panel.grid = element_blank()
  ) +
  guides(fill = "none")
```

```
## Warning: Removed 2710 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



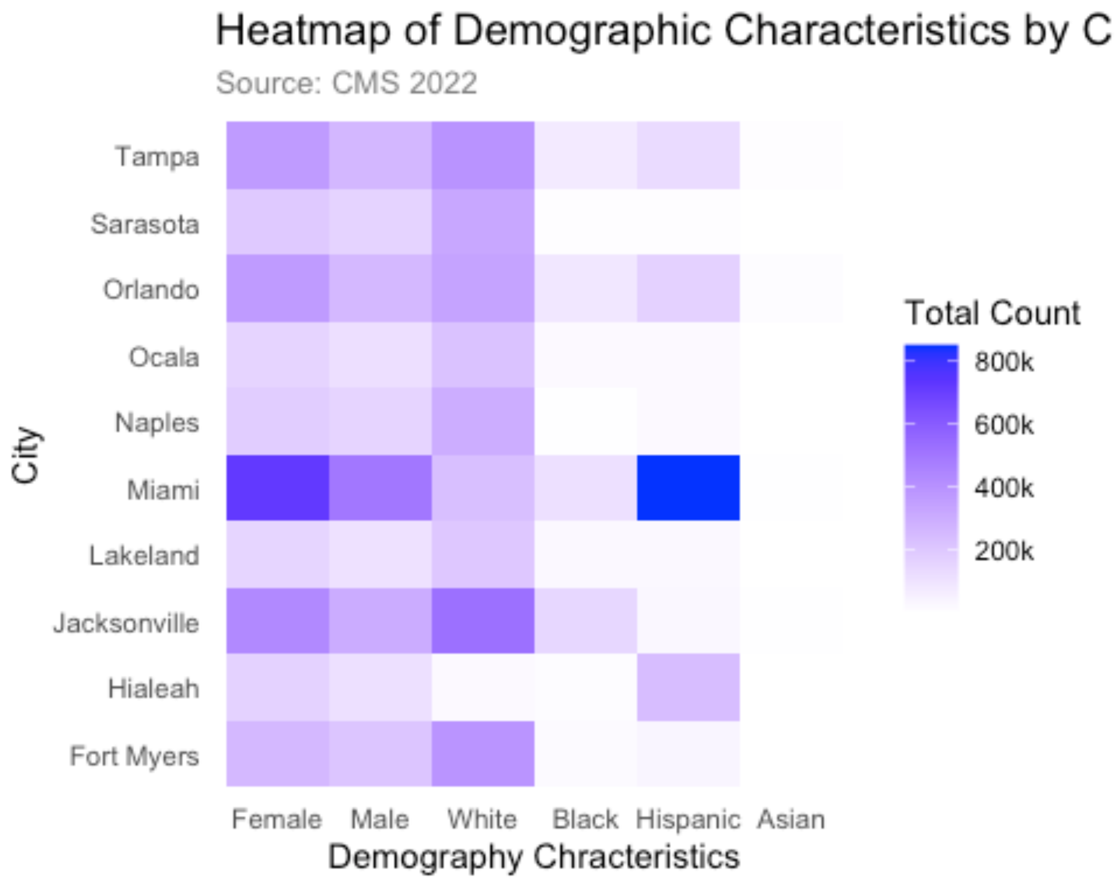

```
#On the x-axis, we can find the breakdown of the number of claims filed by
prescribers per city. The boxplot represents the interquartile range. Based
on the previous chart, the minimum number of claims submitted was 11 across
all cities, with varying maximum numbers. The largest outlier is in Miami,
with a total of 67,440 claims annually, followed by Hialeah and Fort Myers
with 52658 and 33388.
```

#Heatmaps

```
demo<-filtered_data|>
  select(Prscrbr_City,
    Bene_Age_LT_65_Cnt,
    Bene_Age_65_74_Cnt,
    Bene_Age_75_84_Cnt,
    Bene_Age_GT_84_Cnt,
    Bene_Feml_Cnt,
    Bene_Male_Cnt,
    Bene_Race_Wht_Cnt,
    Bene_Race_Black_Cnt,
    Bene_Race_Api_Cnt,
    Bene_Race_Hspnc_Cnt,
    Bene_Race_Natind_Cnt,
    Bene_Race_Othr_Cnt
```

```r
  )


# Summarize the data with better names
demo_3 <- demo |>
  group_by(Prscrbr_City) %>%
  summarise(
    Age_LT_65_Cnt_Sum = sum(Bene_Age_LT_65_Cnt, na.rm = TRUE),
    Age_65_74_Cnt_Sum = sum(Bene_Age_65_74_Cnt, na.rm = TRUE),
    Age_75_84_Cnt_Sum = sum(Bene_Age_75_84_Cnt, na.rm = TRUE),
    Age_GT_84_Cnt_Sum = sum(Bene_Age_GT_84_Cnt, na.rm = TRUE),
    Female_Cnt_Sum = sum(Bene_Feml_Cnt, na.rm = TRUE),
    Male_Cnt_Sum = sum(Bene_Male_Cnt, na.rm = TRUE),
    White_Cnt_Sum = sum(Bene_Race_Wht_Cnt, na.rm = TRUE),
    Black_Cnt_Sum = sum(Bene_Race_Black_Cnt, na.rm = TRUE),
    Asian_Pacific_Islander_Cnt_Sum = sum(Bene_Race_Api_Cnt, na.rm = TRUE),
    Hispanic_Cnt_Sum = sum(Bene_Race_Hspnc_Cnt, na.rm = TRUE),
    Native_Indian_Cnt_Sum = sum(Bene_Race_Natind_Cnt, na.rm = TRUE),
    Other_Race_Cnt_Sum = sum(Bene_Race_Othr_Cnt, na.rm = TRUE))

demo_4 <- demo_3 |>
  pivot_longer(
    cols = -Prscrbr_City,
    names_to = "Category",
    values_to = "Count"
  )

demo_4$Category <- recode(demo_4$Category,
                          "Age_LT_65_Cnt_Sum" = "Age < 65",
                          "Age_65_74_Cnt_Sum" = "Age 65-74",
                          "Age_75_84_Cnt_Sum" = "Age 75-84",
                          "Age_GT_84_Cnt_Sum" = "Age > 84",
                          "Female_Cnt_Sum" = "Female",
                          "Male_Cnt_Sum" = "Male",
                          "White_Cnt_Sum" = "White",
                          "Black_Cnt_Sum" = "Black",
                          "Asian_Pacific_Islander_Cnt_Sum" = "Asian",
                          "Hispanic_Cnt_Sum" = "Hispanic",
                          "Native_Indian_Cnt_Sum" = "Native Indian",
                          "Other_Race_Cnt_Sum" = "Other Race")



demo_4$Category <- str_wrap(demo_4$Category, width = 5)



demo_4$Category <- factor(demo_4$Category,
                           levels = c(
                             "Age < 65", "Age 65-74", "Age 75-84", "Age >
84",
```

```
                             "Female", "Male",
                             "White", "Black",  "Hispanic", "Asian",
                             "Native Indian", "Other Race"
                           ))

demo_4 <- demo_4 |>
  filter(!is.na(Category))

ggplot(demo_4, aes(x = Category, y = Prscrbr_City, fill = Count)) +
  geom_tile() +
  scale_fill_gradient(
    low = "white", high = "blue",
    name = "Total Count",
    labels = label_number(scale = 1e-3, suffix = "k", big.mark = "'")
  )   +
    labs(title = "Heatmap of Demographic Characteristics by City",
       subtitle = "Source: CMS 2022",
       x = "Demography Chracteristics",
       y = "City",
       fill = "Count") +
  theme_minimal() +
    theme(
    plot.title = element_text(size = 14),
    plot.subtitle = element_text(size = 10, color = "gray50"),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank())
```



#Miami stands out among all cities, as it has the darkest color in all categories, indicating a larger number of claims in this city. Additionally, Miami has the largest Hispanic population compared to all other cities, where there is a consistency of White populations. We can also see that the female population tends to be predominant in all cities.