Exploratory Data Analysis

DSAN 5200

Marion Bauman

April 21, 2024

Exploratory Data Analysis

This document introduces some EDA for our project. The goal of this analysis is to understand the story underlying our dataset so that we can generate an effective story line for our final project.

Story Context

Based on a viral TikTok track, an easy to trick to break into a Kia or Hyundai car has been circulating using just a USB cord and a screwdriver. Often perpetrated by teenage boys, the so called "Kia Boys" have been stealing cars across the country, largely for the purpose of joyriding.

Load the Data

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.0     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

```
library(arrow)
```

```
Attaching package: 'arrow'

The following object is masked from 'package:lubridate':

    duration

The following object is masked from 'package:utils':

    timestamp
```

```
adult_arrests <- arrow::read_parquet("../../data/clean/adult_arrests_dc.parquet")
crime_dc <- arrow::read_parquet("../../data/clean/crime_dc.parquet")
juvenile_arrests <- arrow::read_parquet("../../data/clean/juvenile_arrests_dc.parquet")
kia_hyundia_thefts <- arrow::read_parquet("../../data/clean/kia_hyundia_thefts.parquet")
timeline <- read_csv("../../data/raw/multiTimeline.csv")
```

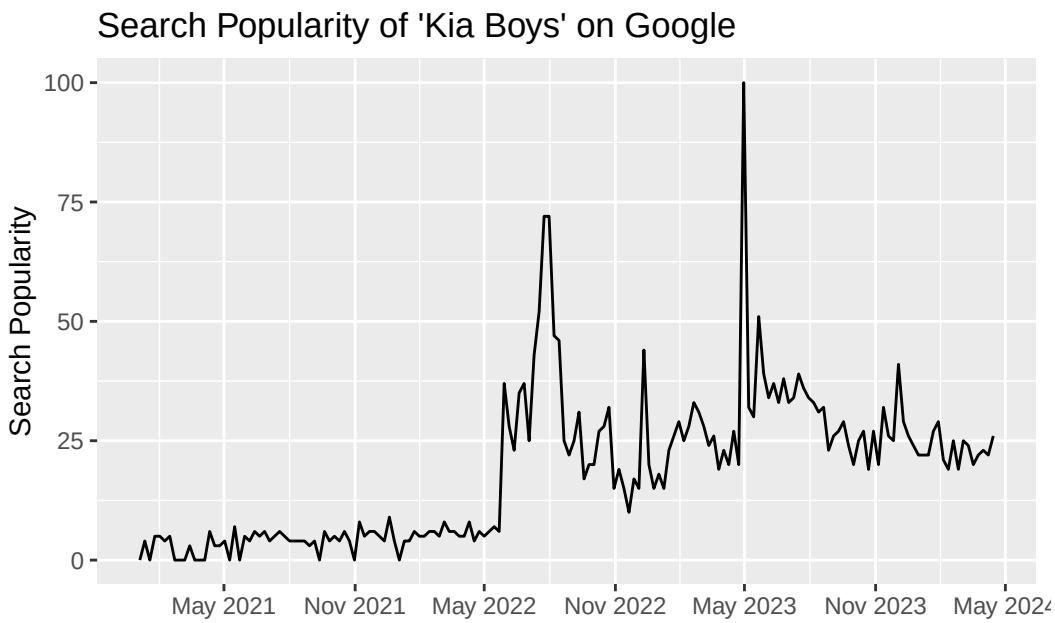
```
Rows: 262 Columns: 2
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl  (1): kia boys
date (1): Week

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Data Overview

Trend Timeline

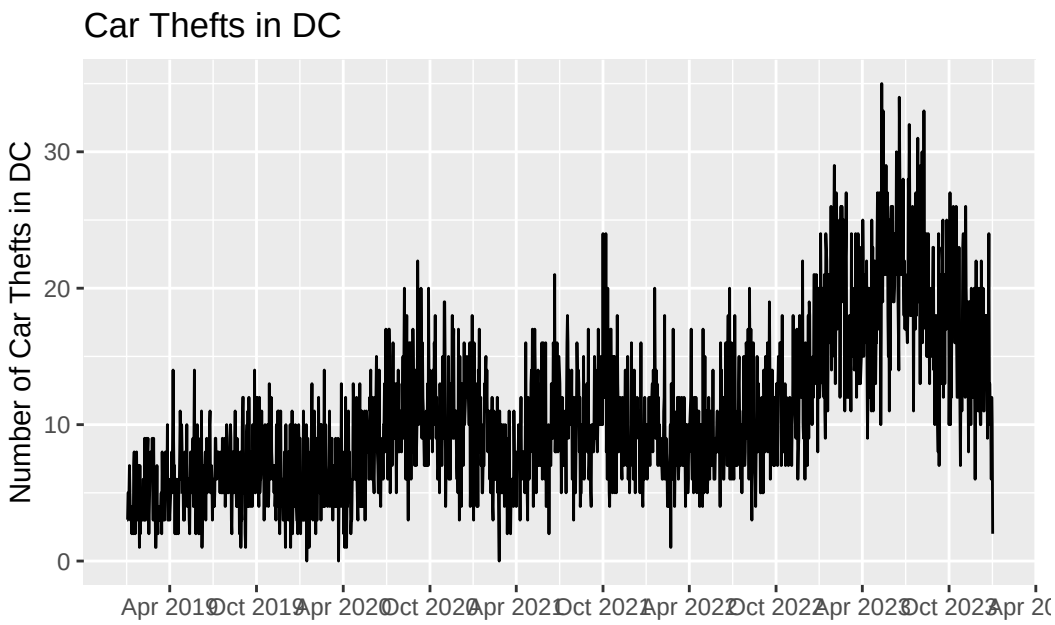
```
timeline |>
    filter(Week >= "2021-01-01") |>
    ggplot(aes(x = Week, y = `kia boys`)) +
    geom_line() +
    scale_x_date(date_labels = "%b %Y", date_breaks = "6 month") +
    labs(
        x = "",
```

```
        y = "Search Popularity",
        title = "Search Popularity of 'Kia Boys' on Google",
    )
```



DC Car Thefts

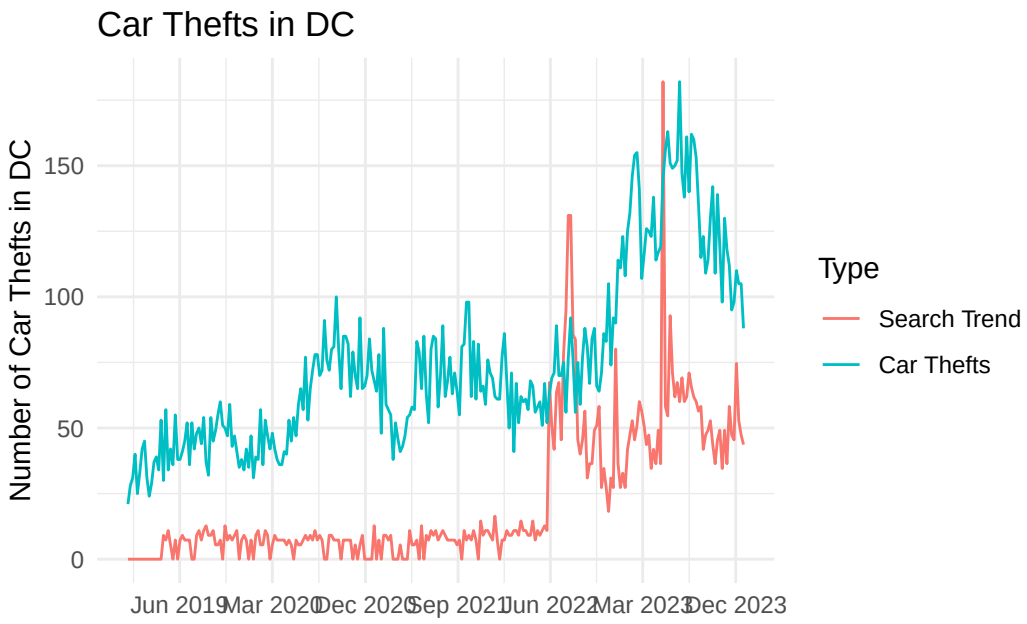
```
crime_dc |>
    mutate(report_date_no_time = as.Date(report_date)) |>
    count(report_date_no_time) |>
    # Add in the missing dates with 0 car thefts
    complete(report_date_no_time = seq.Date(min(report_date_no_time), max(report_date_no_time
    # fill in the missing values with 0
    mutate(n = if_else(is.na(n), 0, n)) |>
    ggplot(aes(x = report_date_no_time, y = n)) +
    geom_line() +
    scale_x_date(date_labels = "%b %Y", date_breaks = "6 month") +
    labs(
        x = "",
        y = "Number of Car Thefts in DC",
        title = "Car Thefts in DC"
    )
```



Let's overlay these two trend lines...

```
joined <- crime_dc |>
    mutate(report_date_no_time = as.Date(report_date)) |>
    count(report_date_no_time) |>
    # Add in the missing dates with 0 car thefts
    complete(report_date_no_time = seq.Date(min(report_date_no_time), max(report_date_no_time
    # fill in the missing values with 0
    mutate(n = if_else(is.na(n), 0, n)) |>
    # Summarise by week
    mutate(Week = floor_date(report_date_no_time, "week")) |>
    group_by(Week) |>
    summarise(n = sum(n)) |>
    left_join(timeline, by = c("Week" = "Week")) |>
    mutate(`kia boys` = if_else(is.na(`kia boys`), 0, `kia boys`)) |>
    # scale the trend line to the same scale as the car theft
    mutate(`kia boys` = `kia boys` * max(n) / max(`kia boys`)) |>
    pivot_longer(
        cols = c(n, `kia boys`),
        names_to = "Type",
        values_to = "n"
    )
```

```
joined |>
    filter(Week < as.Date("2023-12-31")) |>
    ggplot(aes(x = Week, y = n, color = Type)) +
    geom_line() +
    scale_x_date(date_labels = "%b %Y", date_breaks = "9 month") +
    scale_color_discrete(labels = c("Search Trend", "Car Thefts")) +
    labs(
        x = "",
        y = "Number of Car Thefts in DC",
        title = "Car Thefts in DC"
    ) +
    theme_minimal()
```



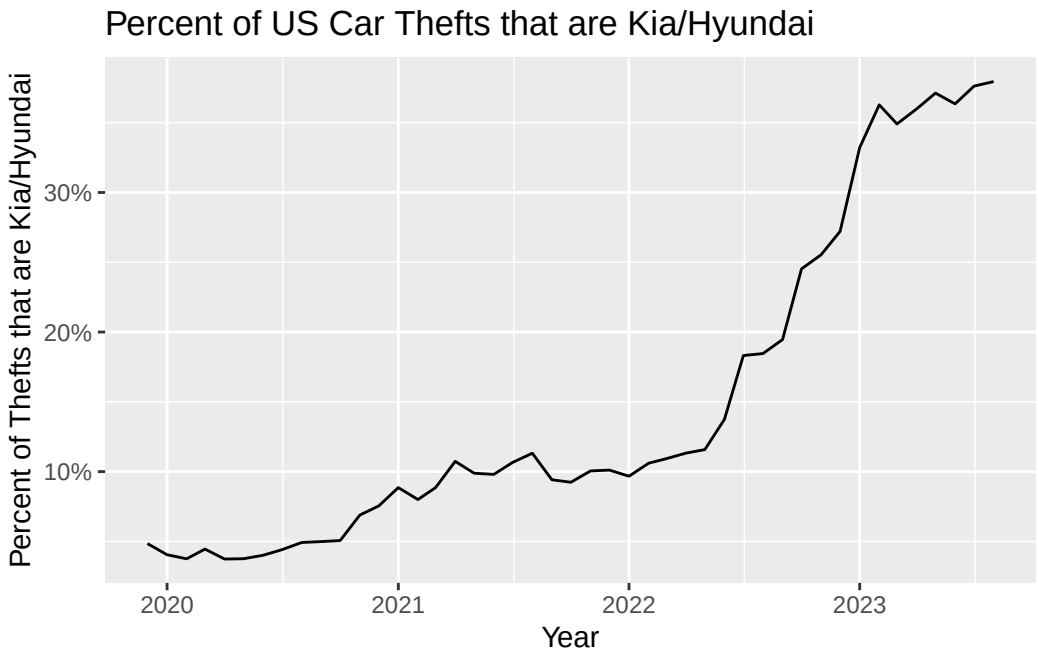
There definitely seems to be a connection here! Let's find the story and communicate it :)

Digging Deeper

Kia and Hyundai Thefts

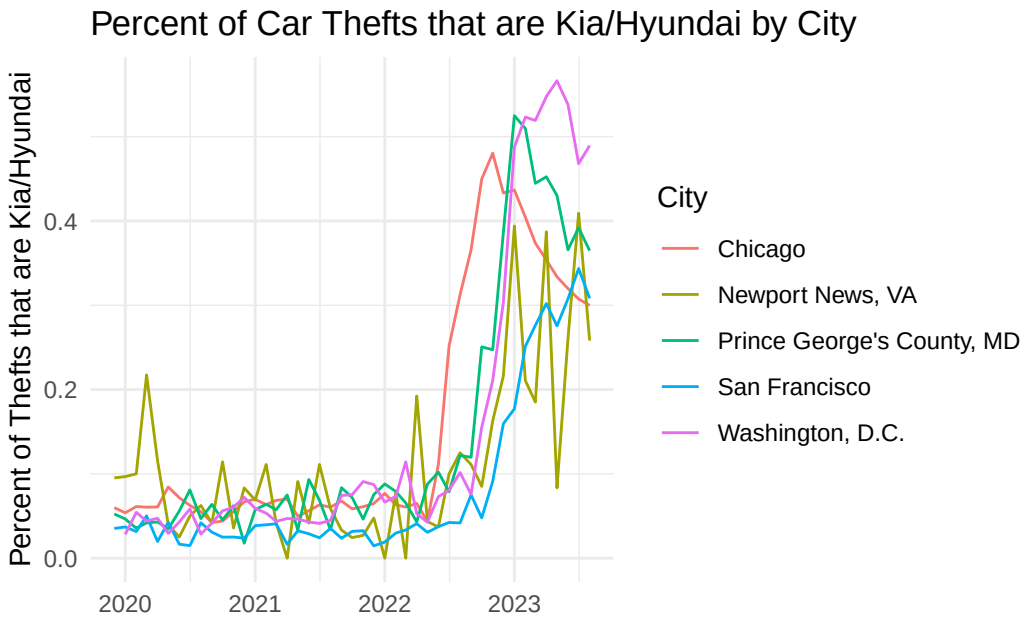
```
kia_hyundia_thefts |>
    summarise(
```

```
        .by = date,
        count_kia_hyundai = sum(count_kia_hyundai, na.rm = TRUE),
        count_all = sum(count_all, na.rm = TRUE)
    ) |>
    mutate(percent_kia_hyundai = count_kia_hyundai / count_all) |>
    ggplot(aes(x = date, y = percent_kia_hyundai)) +
    geom_line() +
    scale_y_continuous(labels = scales::percent) +
    labs(
        x = "Year",
        y = "Percent of Thefts that are Kia/Hyundai",
        title = "Percent of US Car Thefts that are Kia/Hyundai"
    ) +
    theme(
        legend.position = "none"
    )
```



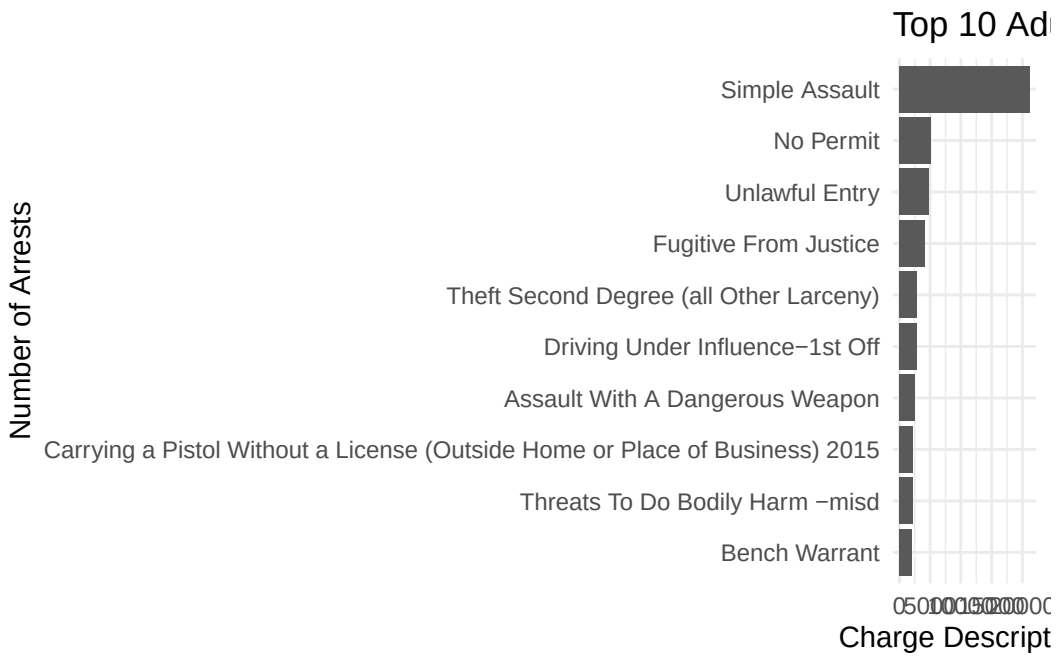

```
kia_hyundia_thefts |>
    summarise(
        .by = c(date, city),
        count_kia_hyundai = sum(count_kia_hyundai, na.rm = TRUE),
        count_all = sum(count_all, na.rm = TRUE)
```

```
    ) |>
    mutate(count_kia_hyundai = if_else(is.na(count_kia_hyundai), 0, count_kia_hyundai)) |>
    mutate(count_all = if_else(is.na(count_all), 0, count_all)) |>
    mutate(percent_kia_hyundai = count_kia_hyundai / count_all) |>
    mutate(percent_kia_hyundai = if_else(is.na(percent_kia_hyundai), 0, percent_kia_hyundai)) |>
    drop_na() |>
    filter(count_all > 0) |>
    filter(city %in% c("Washington, D.C.", "Prince George's County, MD", "Newport News, VA",
    ggplot(aes(x = date, y = percent_kia_hyundai, color = city)) +
    geom_line() +
    scale_x_date(date_labels = "%Y", date_breaks = "1 year") +
    labs(
        x = "",
        y = "Percent of Thefts that are Kia/Hyundai",
        title = "Percent of Car Thefts that are Kia/Hyundai by City",
        color = "City"
    ) +
    theme_minimal()
```



Arrests

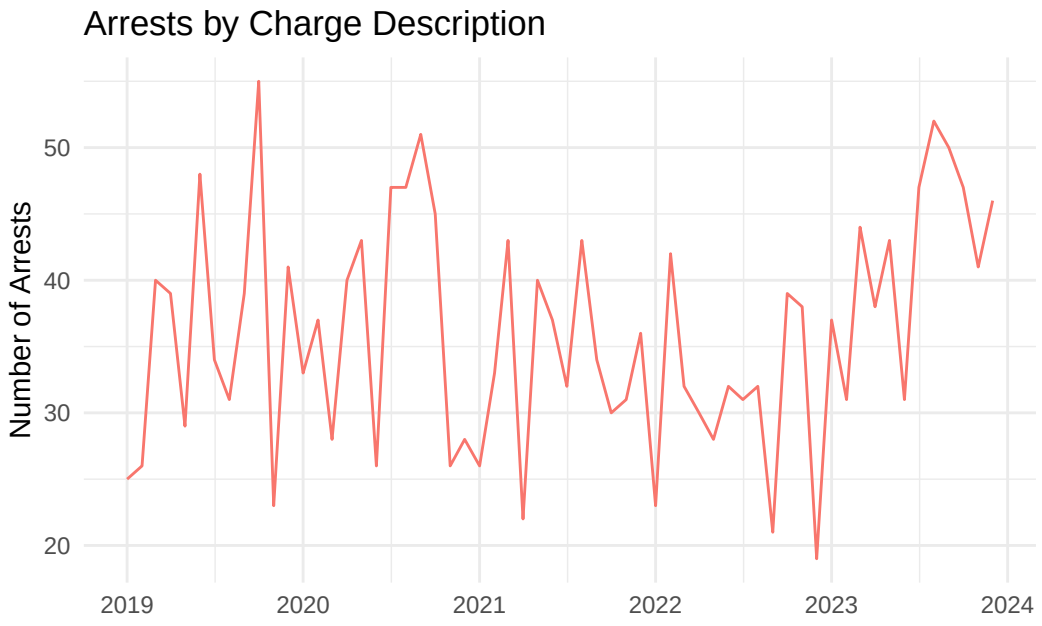
```
adult_arrests |>
    count(charge_description) |>
    arrange(desc(n)) |>
    head(10) |>
    ggplot(aes(x = reorder(charge_description, n), y = n)) +
    geom_col() +
    coord_flip() +
    labs(
        x = "Number of Arrests",
        y = "Charge Description",
        title = "Top 10 Adult Arrests in DC"
    ) +
    theme_minimal()
```




```
adult_arrests |>
    mutate(day = case_when(
        str_detect(arrest_date, "/") ~ str_split_fixed(arrest_date, "/", 3)[, 2],
        TRUE ~ str_split_fixed(arrest_date, "-", 3)[, 3]
    )) |>
```

```
    # If no leading 0, add one
    mutate(day = if_else(nchar(day) == 1, paste0("0", day), day)) |>
    mutate(month = case_when(
        str_detect(arrest_date, "/") ~ str_split_fixed(arrest_date, "/", 3)[, 1],
        TRUE ~ str_split_fixed(arrest_date, "-", 3)[, 2]
    )) |>
    # If no leading 0, add one
    mutate(month = if_else(nchar(month) == 1, paste0("0", month), month)) |>
    mutate(year = case_when(
        str_detect(arrest_date, "/") ~ str_split_fixed(arrest_date, "/", 3)[, 3],
        TRUE ~ str_split_fixed(arrest_date, "-", 3)[, 1]
    )) |>
    mutate(new_cat = case_when(
        charge_description == "Armed Carjacking" ~ "Vehicle Theft",
        charge_description == "Carjacking" ~ "Vehicle Theft",
        charge_description == "Unarmed Carjacking" ~ "Vehicle Theft",
        charge_description == "Unauthorized Use Of A Vehicle" ~ "Vehicle Theft",
        charge_description == "Unauthorized Use Of A Vehicle - Crime Of Violence" ~ "Vehicle
        charge_description == "Unauthorized Use Of A Vehicle - Prior Conviction" ~ "Vehicle
        charge_description == "Unlawful Entry Of A Motor Vehicle" ~ "Vehicle Theft",
        charge_description == "Theft First Degree (Stolen Auto)" ~ "Vehicle Theft",
        charge_description == "Theft Second Degree (Stolen Auto)" ~ "Vehicle Theft",
        str_detect(str_detect(charge_description, "(T|t)heft"), "(A|a)uto") ~ "Vehicle Theft
        TRUE ~ "Other"
    )) |>
    filter(new_cat != "Other") |>
    mutate(date = ymd(paste(year, month, day, sep = "-"))) |>
    summarise(
        .by = c(month, year, new_cat),
        n = n()
    ) |>
    mutate(date = ymd(paste(year, month, "01", sep = "-"))) |>
    drop_na() |>
    ggplot(aes(x = date, y = n, color = new_cat)) +
    geom_line() +
    scale_x_date(date_labels = "%Y", date_breaks = "1 year") +
    labs(
        x = "",
        y = "Number of Arrests",
        title = "Arrests by Charge Description"
    ) +
    theme_minimal() +
```

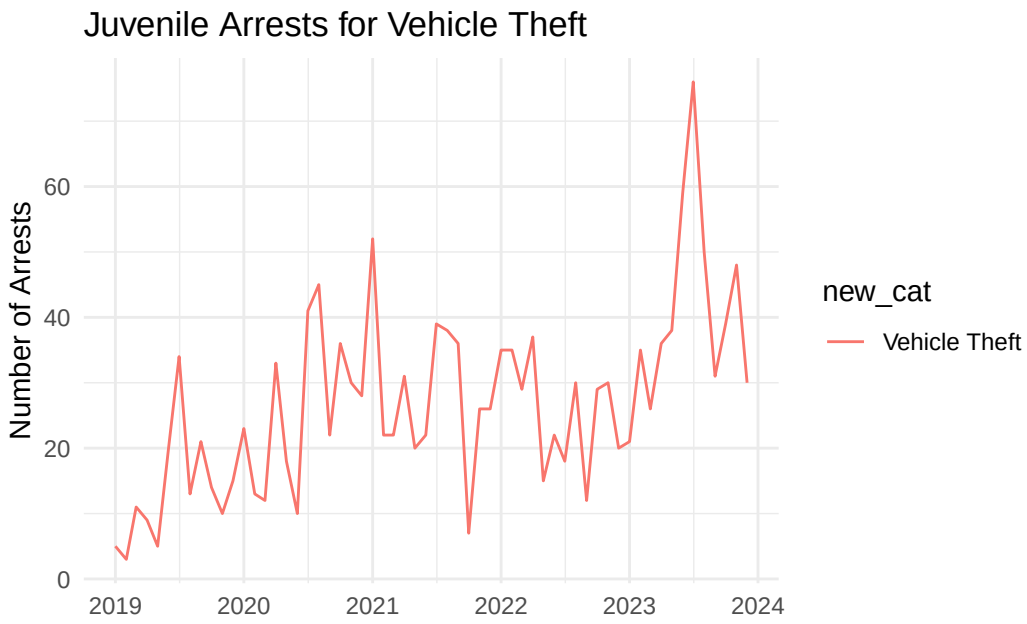
```
    theme(
        legend.position = "none"
    )
```



Juvenile Arrests

```
juvenile_arrests |>
    mutate(date = ymd(arrest_date)) |>
    # get the month and year
    mutate(month = month(date), year = year(date)) |>
    mutate(new_cat = case_when(
        top_charge_description == "Armed Carjacking" ~ "Vehicle Theft",
        top_charge_description == "Carjacking" ~ "Vehicle Theft",
        top_charge_description == "Unarmed Carjacking" ~ "Vehicle Theft",
        top_charge_description == "Unauthorized Use Of A Vehicle" ~ "Vehicle Theft",
        top_charge_description == "Unauthorized Use Of A Vehicle - Crime Of Violence" ~ "Vehi
        top_charge_description == "Unauthorized Use Of A Vehicle - Prior Conviction" ~ "Vehic
        top_charge_description == "Unlawful Entry Of A Motor Vehicle" ~ "Vehicle Theft",
        top_charge_description == "Theft First Degree (Stolen Auto)" ~ "Vehicle Theft",
        top_charge_description == "Theft Second Degree (Stolen Auto)" ~ "Vehicle Theft",
```

```
        TRUE ~ "Other"
    )) |>
    summarise(
        .by = c(month, year, new_cat),
        n = n()
    ) |>
    filter(new_cat != "Other") |>
    mutate(date = ymd(paste(year, month, "01", sep = "-"))) |>
    ggplot(aes(x = date, y = n, color = new_cat)) +
    geom_line() +
    scale_x_date(date_labels = "%Y", date_breaks = "1 year") +
    theme(
        legend.position = "none"
    ) +
    labs(
        x = "",
        y = "Number of Arrests",
        title = "Juvenile Arrests for Vehicle Theft"
    ) +
    theme_minimal()
```

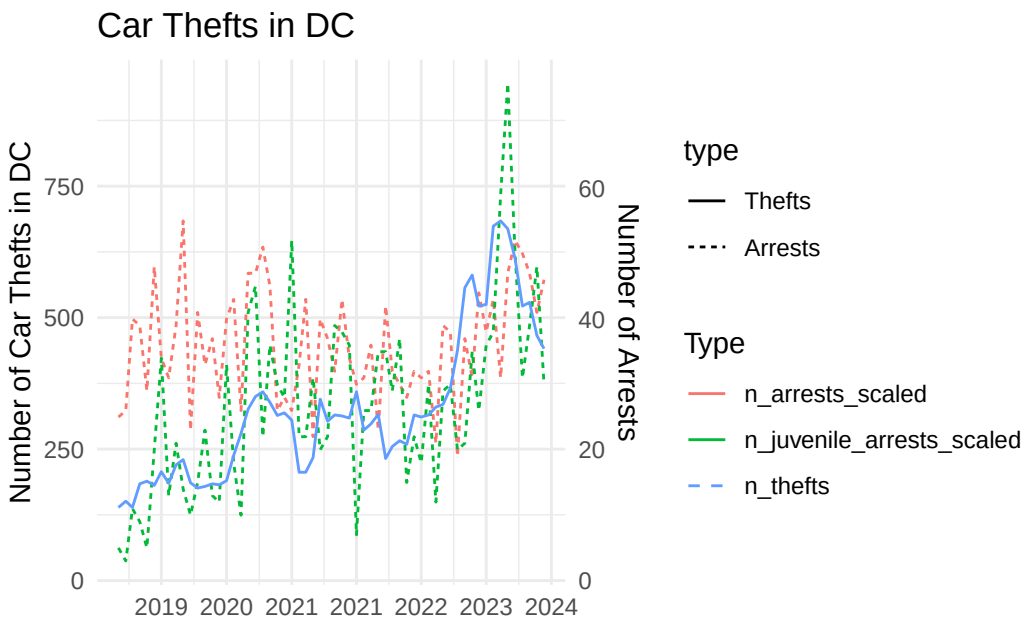


Arrests Linked to Car Thefts

```
car_arrests_month <- adult_arrests |>
    # Get number of adult car theft arrests by week
    mutate(day = case_when(
        str_detect(arrest_date, "/") ~ str_split_fixed(arrest_date, "/", 3)[, 2],
        TRUE ~ str_split_fixed(arrest_date, "-", 3)[, 3]
    )) |>
    # If no leading 0, add one
    mutate(day = if_else(nchar(day) == 1, paste0("0", day), day)) |>
    mutate(month = case_when(
        str_detect(arrest_date, "/") ~ str_split_fixed(arrest_date, "/", 3)[, 1],
        TRUE ~ str_split_fixed(arrest_date, "-", 3)[, 2]
    )) |>
    # If no leading 0, add one
    mutate(month = if_else(nchar(month) == 1, paste0("0", month), month)) |>
    mutate(year = case_when(
        str_detect(arrest_date, "/") ~ str_split_fixed(arrest_date, "/", 3)[, 3],
        TRUE ~ str_split_fixed(arrest_date, "-", 3)[, 1]
    )) |>
    mutate(Month = ymd(paste(year, month, day, sep = "-")) %>% floor_date("month")) |>
    filter(charge_description %in% c("Armed Carjacking", "Carjacking", "Unarmed Carjacking",
    count(Month)

car_theft_month <- crime_dc |>
    mutate(report_date_no_time = as.Date(report_date)) |>
    count(report_date_no_time) |>
    # Add in the missing dates with 0 car thefts
    complete(report_date_no_time = seq.Date(min(report_date_no_time), max(report_date_no_tim
    # fill in the missing values with 0
    mutate(n = if_else(is.na(n), 0, n)) |>
    # Summarise by month
    mutate(Month = floor_date(report_date_no_time, "month")) |>
    group_by(Month) |>
    summarise(n = sum(n))

car_arrests_juvenile_month <- juvenile_arrests |>
    # Get number of juvenile car theft arrests by week
    mutate(date = ymd(arrest_date)) |>
    # get the month and year
    mutate(month = month(date), year = year(date)) |>
    mutate(Month = ymd(paste(year, month, "01", sep = "-")) %>% floor_date("month")) |>
```

```
    filter(top_charge_description %in% c("Armed Carjacking", "Carjacking", "Unarmed Carjacki
    count(Month)

joiner <- car_theft_month |>
    full_join(car_arrests_month, by = c("Month" = "Month")) |>
    rename(
        "n_thefts" = n.x,
        "n_arrests" = n.y
    ) |>
    full_join(car_arrests_juvenile_month, by = c("Month" = "Month")) |>
    rename(
        "n_juvenile_arrests" = n
    ) |>
    mutate(n_arrests = if_else(is.na(n_arrests), 0, n_arrests)) |>
    mutate(n_thefts = if_else(is.na(n_thefts), 0, n_thefts)) |>
    mutate(n_juvenile_arrests = if_else(is.na(n_juvenile_arrests), 0, n_juvenile_arrests)) |
    # Scale the juvenile and adult arrests to the same as the car thefts
    mutate(
        n_arrests_scaled = n_arrests * max(n_thefts) / max(n_arrests),
        n_juvenile_arrests_scaled = n_juvenile_arrests * max(n_thefts) / max(n_arrests)
    ) |>
    select(-n_arrests, -n_juvenile_arrests) |>
    pivot_longer(
        cols = c(n_thefts, n_arrests_scaled, n_juvenile_arrests_scaled),
        names_to = "Type",
        values_to = "n"
    )

# Get max value for scaling
max_n_arrests <- max(car_arrests_month |> select(n))
max_n_thefts <- max(car_theft_month |> select(n))

joiner |>
    filter(Month < as.Date("2023-12-31")) |>
    mutate(type = case_when(
        Type == "n_thefts" ~ "Thefts",
        Type == "n_arrests_scaled" ~ "zArrests",
        Type == "n_juvenile_arrests_scaled" ~ "zArrests"
    )) |>
    ggplot(aes(x = Month, y = n)) +
    geom_line(
        aes(color = Type, linetype = type)
```

```
    ) +
    scale_x_date(date_labels = "%Y", date_breaks = "9 month") +
    # add second scale for arrests
    scale_y_continuous(
        sec.axis = sec_axis(~ . * max_n_arrests / max_n_thefts, name = "Number of Arrests")
    ) +
    labs(
        x = "",
        y = "Number of Car Thefts in DC",
        title = "Car Thefts in DC"
    ) +
    theme_minimal() +
    scale_linetype(
        breaks = c("Thefts", "zArrests"),
        labels = c("Thefts", "Arrests")
    ) +
    # Can't figure out how to combine the legends
    guides(
        color = guide_legend(title = "Type", override.aes = list(linetype = c("solid", "solic
    )
```



Interesting... DC arrests do not show this same theft trend. Maybe DC police are not arresting the perpetrators?