# Assigment 8

Alexis Kwan, Ethan Jantz, Jordan Evangelista, & Malley Smith

3/8/2021

## Data Cleaning Protocol

Our final dataset should be one object, crime. It should combine 2010 and 2019 data. The columns should be: month, primary type, description, location, arrest, domestic, and type.

In combining the datasets, we will mutate the data values to summarize them as months. We will also group crimes by whether they are violent crimes against persons, crimes against property, or something else. Finally, we will remove all of the columns that are not needed for our analysis.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.0.4
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
crime_2010 <- read_csv("Data/Crimes_2010.csv") %>%
  # convert categories to factors for ease of use later
  mutate(
    across(`Primary Type`:`Location Description`, as.factor))
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
```

```
##   `Case Number` = col_character(),
##   Date = col_character(),
##   Block = col_character(),
##   IUCR = col_character(),
##   `Primary Type` = col_character(),
##   Description = col_character(),
##   `Location Description` = col_character(),
##   Arrest = col_logical(),
##   Domestic = col_logical(),
##   Beat = col_double(),
##   Ward = col_double(),
##   `FBI Code` = col_character(),
##   `X Coordinate` = col_double(),
##   `Y Coordinate` = col_double(),
##   Latitude = col_double(),
##   Longitude = col_double(),
##   Location = col_character()
## )
```

```
crime_2019 <- read_csv("Data/Crimes_2019.csv") %>%
  mutate(across(`Primary Type`:`Location Description`, as.factor))
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   `CASE#` = col_character(),
##   Date = col_character(),
##   Block = col_character(),
##   IUCR = col_character(),
##   `Primary Type` = col_character(),
##   Description = col_character(),
##   `Location Description` = col_character(),
##   Arrest = col_character(),
##   Domestic = col_character(),
##   Beat = col_double(),
##   Ward = col_double(),
##   `FBI Code` = col_character(),
##   `X Coordinate` = col_double(),
##   `Y Coordinate` = col_double(),
##   Latitude = col_double(),
##   Longitude = col_double(),
##   Location = col_character()
## )
```

```
# Violent + Prop Crime Descriptions
violent = c(
  "ASSAULT",
  "BATTERY",
  "CRIM SEXUAL ASSAULT",
  "CRIMINAL SEXUAL ASSAULT",
  "HOMICIDE",
  "HUMAN TRAFFICKING",
  "KIDNAPPING",
  "SEX OFFENSE"
)
```

```
property = c("ARSON",
             "THEFT",
             "MOTOR VEHICLE THEFT",
             "BURGLARY",
             "ROBBERY",
             "CRIMINAL DAMAGE")
```

## Research Question

- Where do violent crimes happen?
- Where do property crimes happen?
- Do certain types of locations have associations with different types of crimes?

```
### --------------------
### Crime data cleaning
###
# Final dataset should be one object, crime. It should combine 2010 and 2019 data
# Columns should be:
# month, primary type, description, location, arrest, domestic, violent, property

clean_crime_2010 <- crime_2010 %>%
  mutate(
    Date = mdy_hm(Date),
    # Convert character type date to date type
    month = month(Date),
    # Pull month
    Arrest = as.logical(Arrest),
    # Ensure logical value is logical for combining data
    Domestic = as.logical(Domestic),
    type = as.factor(
      case_when(
        `Primary Type` %in% violent ~ "violent",
        # Classify crime as violent/property/else
        `Primary Type` %in% property ~ "property",
        TRUE ~ "other"
      )
    )
  ) %>%
  # These are the only values we want to look at
  # renames values for ease of use later
  select(
    Date,
    month,
    primary_type = `Primary Type`,
    desc = Description,
    location = `Location Description`,
    arrest = Arrest,
    domestic = Domestic,
    type
  )

# In a previous version of this script the data was filtered to only include
# dates in 2019, because the 2019 data included 2020 data
```

```r
# However, on closer inspection, there was an issue in data collection that
# meant that crime_2019 was data from April 2019 until April 2020.
# To capture seasonality we want to include Jan - April in our data
# So while it's not perfect (since the pandemic effects are included)
# I think it is important to ensure a full year of data is captured for
# possible 2010-2019 comparisons
clean_crime_2019 <- crime_2019 %>%
  mutate(Date = mdy_hm(Date),
         month = month(Date),
         # convert values to something usable
         Arrest = as.logical(ifelse(Arrest == "Y", TRUE, FALSE)),
         Domestic = as.logical(ifelse(Domestic == "Y", TRUE, FALSE)),
         type = as.factor(case_when(`Primary Type` %in% violent ~ "violent",
                          `Primary Type` %in% property ~ "property",
                          TRUE ~ "other"))) %>%
  select(Date,
         month,
         primary_type = `Primary Type`,
         desc = Description,
         location = `Location Description`,
         arrest = Arrest,
         domestic = Domestic,
         type)

crime <- bind_rows(clean_crime_2010, clean_crime_2019) # Combine data into one big set
rm(crime_2010, crime_2019, clean_crime_2010, clean_crime_2019) # we only really want the final cleaned
```

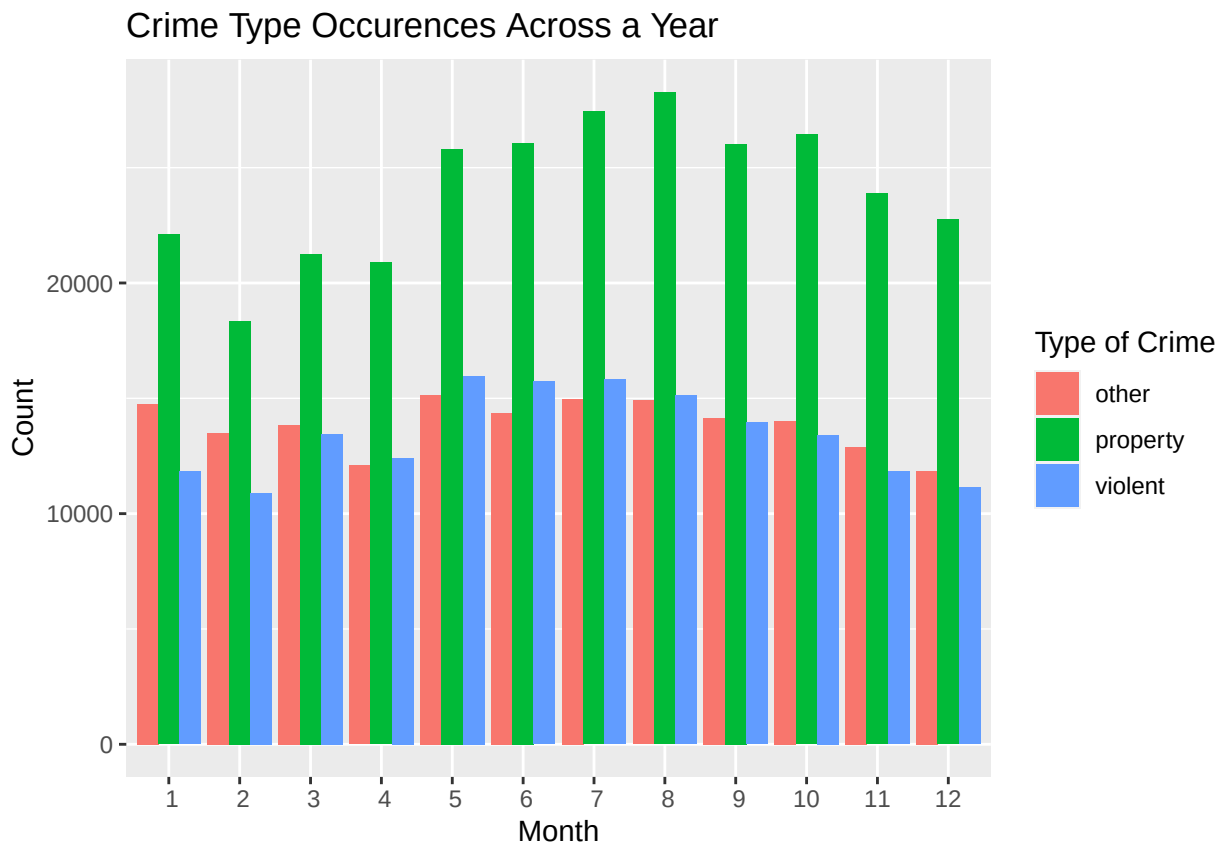
## Exploratory Data Analysis + Data Viz

In our exploratory analysis of the combined dataset we created a number of tables to look at crime seasonality by type and crime by location.

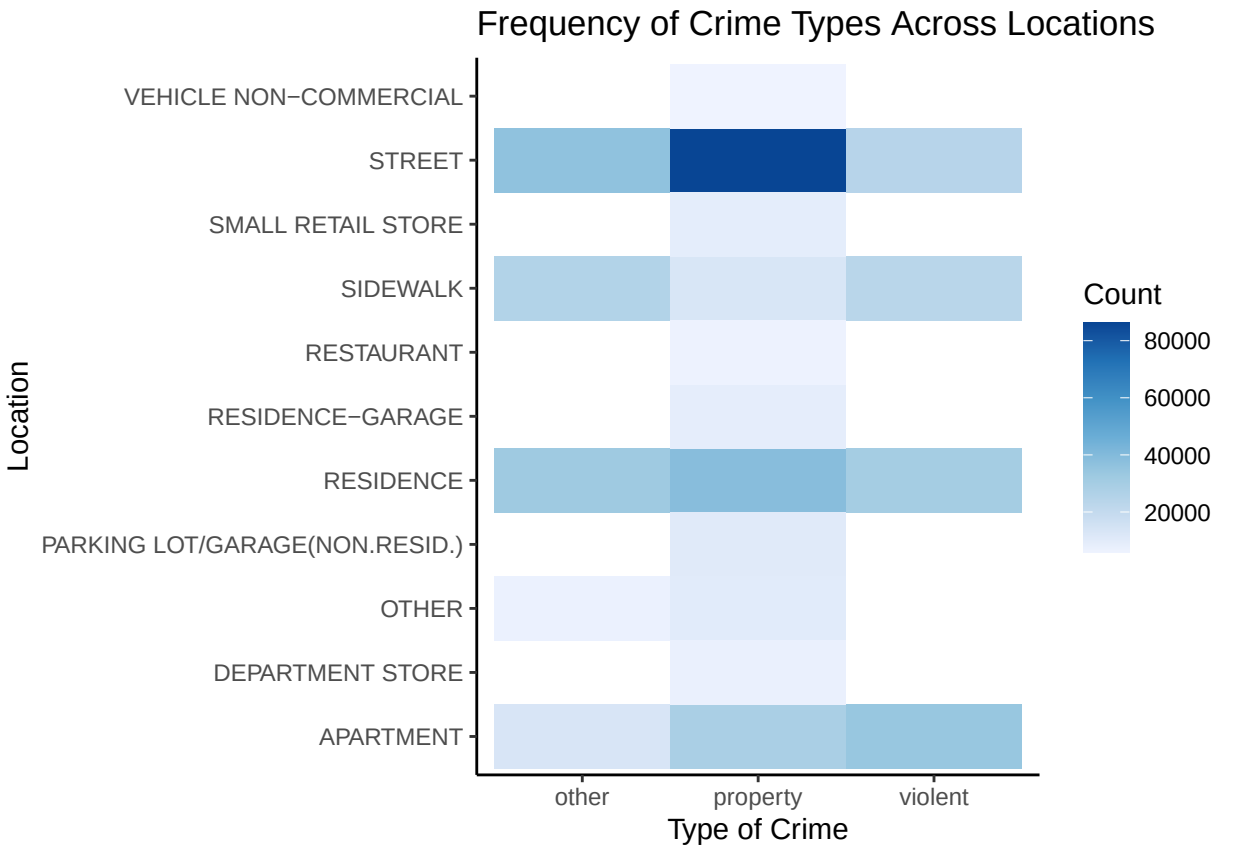
```r
# Grouped geom_bar graph shows the difference in counts across the different types
# but also shows how the numbers evolve over the different months. It also gives
# clear indicators of when certain types of crime happen, signaled by the tallest
# bars, thus addressing the first two research questions.
crime %>%
  mutate(month = stamp("April")(month)) %>%
  ggplot(aes(x = month, fill = type)) +
  geom_bar(position = "dodge") +
  labs(title = "Crime Type Occurences Across a Year", x = "Month", y = "Count", fill = "Type of Crime")
```

```
## Multiple formats matched: "%Om"(1), "%B"(0)
```

```
## Using: "%Om"
```




```
filtered_crime <- crime %>%
  group_by(location, type) %>%
  count(name = "Count") %>%
  arrange(desc(Count))
# Heatmap seemed to be the best choice for answering the third research question.
# The color intensity is a good way to see the relative propensity for the crime
# to occur based on the location split in the y-axis. This matrix like display
# gives the ability to immediately survey the different combinations of scenarios.
# I only chose the top 20 because they were the high frequency and I felt those
# were likely the most relevant because they made up the highest proportion. The
# number 20 was arbitrary.
head(filtered_crime, 20) %>%
  ggplot(aes(x = type, y = location, fill = Count)) +
  geom_tile() +
  scale_fill_distiller(palette = "Blues", direction = 1) +
  xlab("Type of Crime") +
  ylab("Location") +
  ggtitle("Frequency of Crime Types Across Locations") +
  # I changed it from the original gray theme because I felt the gray background
  # may confuse the reader since colors have meaning associated with the data.
  theme_classic()
```




As you can see in the plot, violent crimes tends to be more concentrated in residential spaces and on the street, with almost negligible occurrences elsewhere. Property crime, in comparison, is more evenly distributed across spaces but is still very concentrated in street spaces.

Next, we explored a specific type of location: the vehicle. We wanted to explore two questions.

- Are violent crimes more or less likely to occur in commercial/rideshare vehicles than personal vehicles?
- What types of crimes are happening in vehicles that are not property or violent crime?

The following plots are our findings.

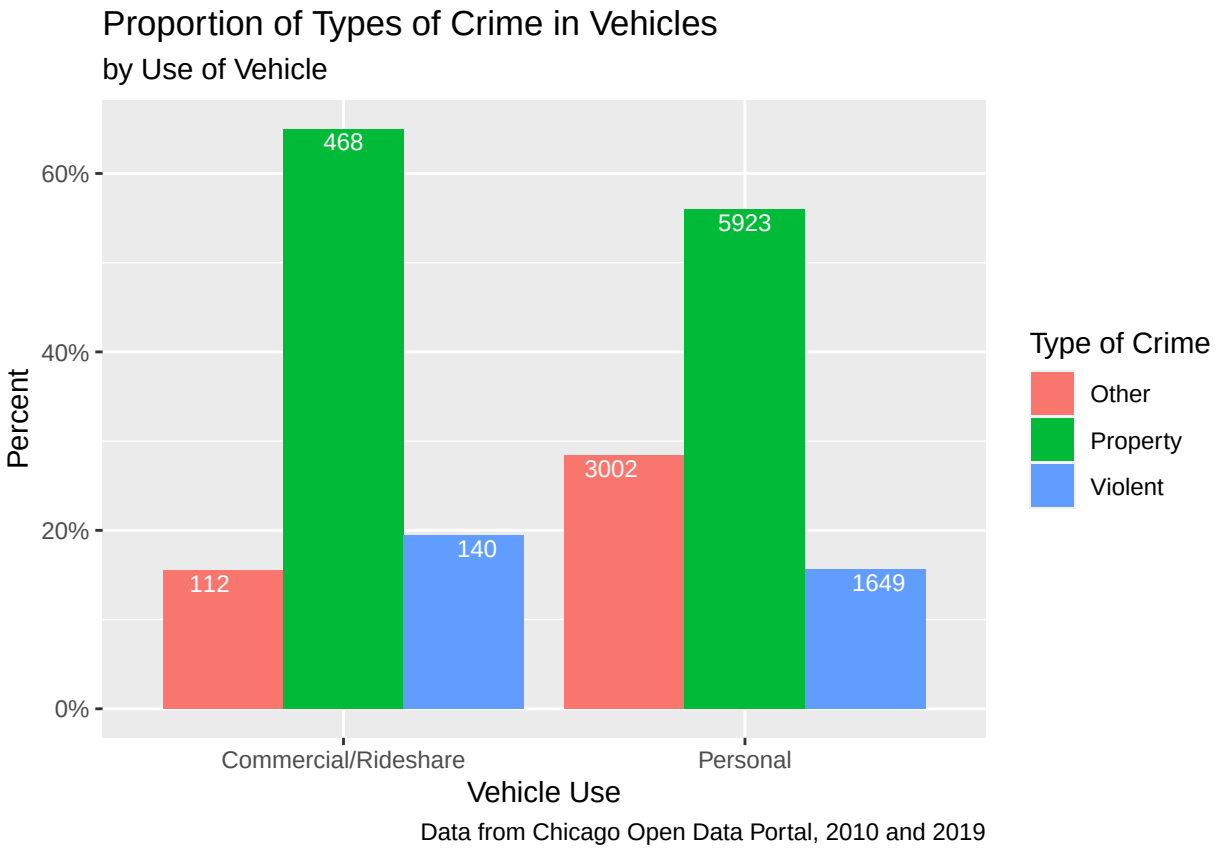
This plot shows us that the there is a distinct difference in the types of crimes that occur in personal and non-personal vehicles.

```
vehicle_comparison_crime <- crime %>%
  filter(grepl("ride share|vehicle-commercial|vehicle - commercial|non-commercial", location, ignore.cas
  mutate(location = case_when(
    grepl("vehicle-commercial|vehicle - commercial|delivery", location,  ignore.case = T) ~ "COMMERCIAL
    grepl("ride share", location, ignore.case = T) ~ "RIDESHARE",
    grepl("non-commercial", location, ignore.case = T) ~ "PERSONAL"
  ))

vehicle_type_plot <- vehicle_comparison_crime %>%
  mutate(personal_vehicle = ifelse(location == "PERSONAL", T, F)) %>%
  group_by(personal_vehicle, type) %>%
  count() %>%
  group_by(personal_vehicle) %>%
  mutate(freq = n / sum(n)) %>% # creates a frequency table
  ungroup() %>%
```

```
  mutate(personal_vehicle = ifelse(personal_vehicle, "Personal", "Commercial/Rideshare")) %>%
  ggplot(aes(x = personal_vehicle, y = freq)) +
  geom_col(aes(fill = str_to_title(type)), position = "dodge") +
  geom_text(
    aes(label = n, group = type),
    colour = "white",
    size = 3,
    vjust = 1.25,
    hjust = "center",
    position = position_dodge(width = 1)
  ) +
  scale_y_continuous(labels = scales::percent_format(accuracy = 2)) +
  labs(
    title = "Proportion of Types of Crime in Vehicles",
    subtitle = "by Use of Vehicle",
    caption = "Data from Chicago Open Data Portal, 2010 and 2019",
    x = "Vehicle Use",
    y = "Percent",
    fill = "Type of Crime"
  ) # It looks like the answer is yes. Let's try and take this a step further

vehicle_type_plot
```



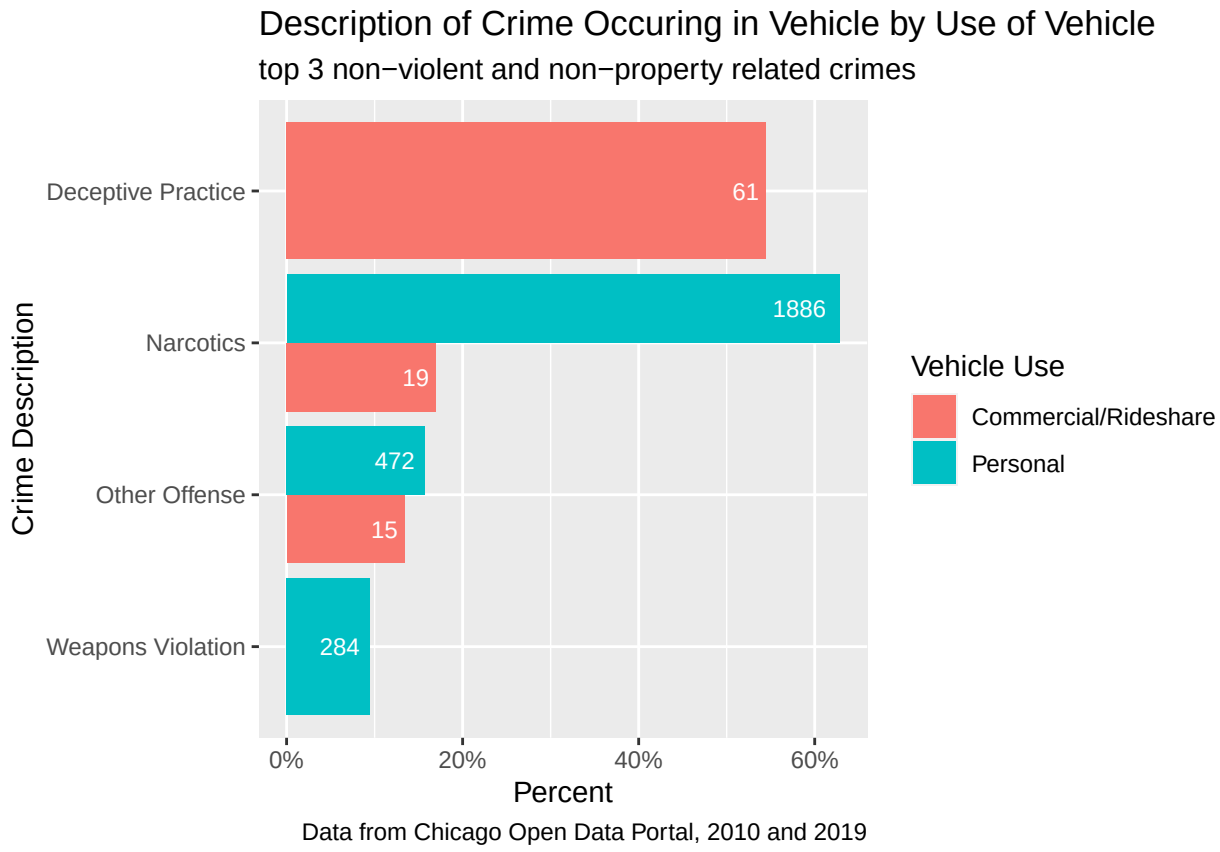

We can also see that police are used much more often for property crimes.

The above plot had us curious as what "other" crimes were happening, since violent and property crimes are

more straightforward.

```
other_plot <- vehicle_comparison_crime %>%
  mutate(personal_vehicle = ifelse(location == "PERSONAL", T, F)) %>%
  filter(type == "other") %>%
  count(personal_vehicle, primary_type) %>%
  group_by(personal_vehicle) %>%
  mutate(freq = n / sum(n)) %>% # creates a frequency table
  ungroup() %>%
  # clean the category, and reduce non-personal categories
  mutate(personal_vehicle = ifelse(personal_vehicle,
                                   "Personal", "Commercial/Rideshare")) %>%
  group_by(personal_vehicle) %>%
  # get the top 3 of every group
  slice_max(order_by = freq, n = 3) %>%
  ggplot(aes(x = reorder(str_to_title(primary_type), freq),
             y = freq, fill = personal_vehicle)) +
  geom_col(position = "dodge") +
  # make the y axis more readable
  scale_y_continuous(labels = scales::percent_format(accuracy = 2)) +
  coord_flip() +
  geom_text(aes(label = n), colour = "white", size = 3,
            hjust = 1.25, position = position_dodge(.9)) +
  labs(title = "Description of Crime Occuring in Vehicle by Use of Vehicle",
       subtitle = "top 3 non-violent and non-property related crimes",
       caption = "Data from Chicago Open Data Portal, 2010 and 2019",
       x = "Crime Description",
       y = "Percent",
       fill = "Vehicle Use")

other_plot
```


Description of Crime Occuring in Vehicle by Use of Vehicle
top 3 non−violent and non−property related crimes
Crime Description
Deceptive Practice
Narcotics
Other Offense
Weapons Violation
61
1886
19
472
15
284
0%
20%
40%
60%
Percent
Vehicle Use
Commercial/Rideshare
Personal
Data from Chicago Open Data Portal, 2010 and 2019