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 3.3.3 v purrr 0.3.4  
## v tibble 3.0.5 v dplyr 1.0.3  
## 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") %>%  
 mutate(  
 across(`Primary Type`:`Location Description`, as.factor)) # convert categories to factors for ease of use later

##   
## -- 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  
# 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),   
 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 dataset to work with, so I'm removing the transient data objects

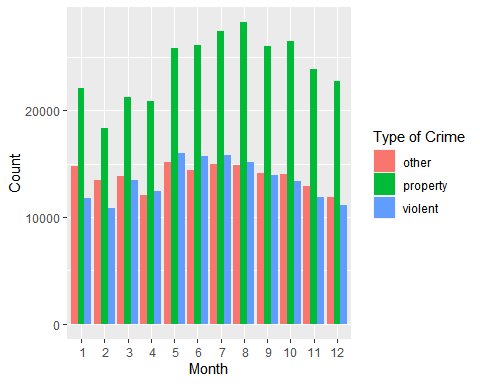
# Exploratory Data Analysis + Data Viz

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.

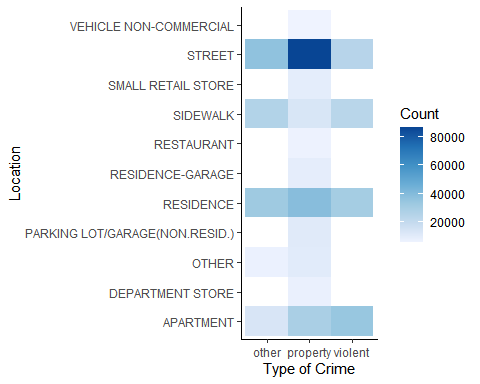
# Grouped geom\_bar graph shows the difference in counts across the different types  
# but also shows how the numbers evolve over the different months. It also gives  
# clear indicators of when certain types of crime happen, signaled by the tallest  
# bars, thus addressing the first two research questions.  
crime %>%  
 mutate(month = stamp("April")(month)) %>%  
 ggplot(aes(x = month, fill = type)) +  
 geom\_bar(position = "dodge") +   
 labs(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 occure 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") +  
 # 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()



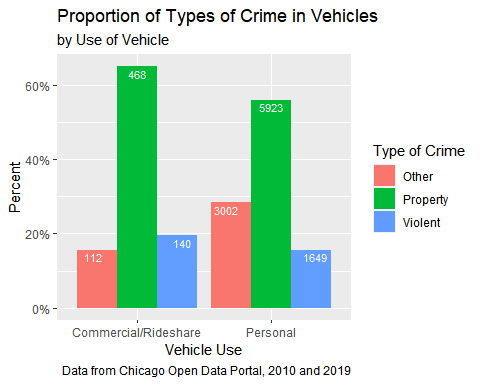
Next, we explored a specific type of location: the vehicle. We wanted to explore two questions.

* 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.case = T)) %>%  
 mutate(location = case\_when(  
 grepl("vehicle-commercial|vehicle - commercial|delivery", location, ignore.case = T) ~ "COMMERCIAL", # grepl checks strings for the presence of a pattern, or set of patterns  
 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() %>%  
 mutate(personal\_vehicle = ifelse(personal\_vehicle, "Personal", "Commercial/Rideshare")) %>%  
 group\_by(personal\_vehicle) %>%  
 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") +  
 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

