Crime Data Analysis

Candace Gostinski

2024-07-19

Introduction

In a world of increasing criminal activity, many regions are interested in studying patterns of such incidents in order to adjust their approach to combating crime. Our client, based in Los Angeles, CA, provided us with the last four years of their data: 2020-2024, as well as three main questions they would like to have answered. The questions are listed below.

1. How did the rate of crime change between 2020 and 2023?
2. What was the most common crime committed in the last four years?
3. Was there a particular group that was targeted?

The following analysis will answer these questions in detail.

Setting Up Analysis Environment

For the purposes of this examination, our analysis was completed in RStudio Cloud. The following packages were used: tidyverse, dplyr, ggplot, and reshape2. These packages would serve to organize and display our data for easy review. We also downloaded the QuantPsyc package just in case we needed to do some statistical analysis.

```
install.packages("tidyverse")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
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   3.5.1     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 become errors
```

```
library(dplyr)
library(ggplot2)
install.packages("QuantPsyc")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
library(QuantPsyc)
```

```
## Loading required package: boot
## Loading required package: MASS
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
##
##
## Attaching package: 'QuantPsyc'
##
## The following object is masked from 'package:base':
##
##     norm
```

Next, we imported our csv file: "Crime_Data_from_2020_to_Present.csv," and assigned it to the variable "LA_crime_data."

```
LA_crime_data <- read.csv("Crime_Data_from_2020_to_Present.csv", nrows = 1000)
```

Cleaning Our Data

In previous projects, we cleaned our data using Google Sheets or SQL. This time, however, we decided to clean the data directly in RStudio. Upon examining it further, we noticed that the main things that needed to be corrected were removing duplicate rows as well as the extra spaces in the location column. In addition, we also converted the date type to a more analysis-friendly format of YY-MM-DD.

```
clean_location <- function(location) {
  location <- gsub("\\s+", " ", location)
  location <- trimws(location)
  return(location)
}
crime_data_cleaned <- LA_crime_data %>%
  mutate(LOCATION = sapply(LOCATION, clean_location)) %>%
  distinct(LOCATION, .keep_all = TRUE)

crime_data_cleaned$Date.Rptd <- as.Date(crime_data_cleaned$Date.Rptd, format="%Y-%m-%d")
crime_data_cleaned$DATE.OCC <- as.Date(crime_data_cleaned$DATE.OCC, format="%Y-%m-%d")
```

Analysis

Crime Rate Difference: 2020-2023 To begin our analysis, our client wanted to know the difference in crime rate between 2020 and 2023. Results show that there was a dramatic decrease in the years following 2020. 2020 totaled 39,221 crimes reported, while 2023 shows only 5,028. That is an **87% decrease** from the first year to the last.

Highest Crime Committed Our client next wanted to know which category received the highest number of crimes committed in the last four years. In analyzing the results, we found that **vehicle stolen** was the most common crime, with **8,786** reported incidents. Behind it was identity theft at 5,259, and burglary at 5,255.

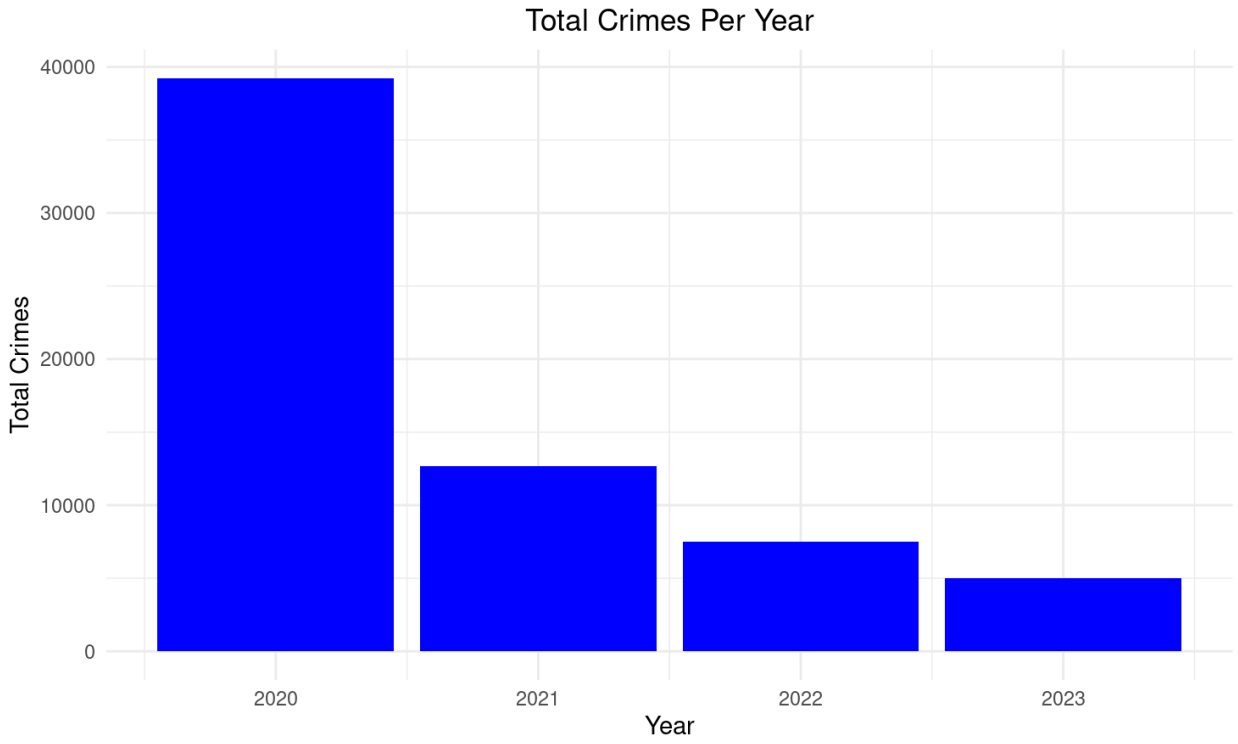

Figure 1: crime data

Targeted Groups Our client later informed us that one of the reasons they are examining this data is to develop action plans to reduce the crime rate in LA, as well as analyze what groups and/or categories of people are most often targeted. We focused on the top three crimes reported in the last four years: vehicle stolen, identity theft, and burglary. Our client also specifically asked that we examine domestic violence as well.

Vehicle Stolen In reviewing the data for vehicles stolen, a limitation in our analysis was that most of the reported incidents did not specify the gender or race of the victim. Therefore, we cannot say conclusively which group was targeted the most. However, we were able to see that **Northeast County** had the highest number of reported incidents at 691. Hollenbeck County followed with 686.

In order to get a rough idea of who may be the majory sex and race in these counties, we filtered the data to show the totals for each. For Northeast County, **white males make up nearly 20% of the population** - the highest percentage of an individual race. That is about one-fifth of the county, which while by itself is not extremely significant, when combined with white females it brings the total percentage up to **37%.** (White females are behind them at 17%.) That is considerably more significant. For Hollenbeck County, the majority of the population appears to be hispanic males and females. Both **hispanic males and females** each make up roughly 25% of the population. That is a **combined average of almost 50%**, which is very significant.

Therefore, based on the data, it is **possible** that in the majority of the vehicle stolen incidents in Northeast County, the victims were either a white male or white female, but since they only account for 37% of the population, we would need far more data to support this theory. Also, it is possible that in the majority of the incidents in Hollenbeck County, the victims were either Hispanic males or females. Again, though, we would need more data to support that theory.

Identity Theft For our examination into identity theft, we were able to find race and gender for the vast majority of the reported incidents. We found that **white males were the most common victim**

(23%) and white females were behind them at 20%. Together, that totals roughly **43% of the reported incidents**, which is considerably significant. Therefore, it is likely that white males and females are the most common victims of identity theft in LA county. Below is an example of the code used to determine this.

```
theft_identity_noblanks <- crime_data_cleaned %>%
  filter(Crm.Cd.Desc == "THEFT OF IDENTITY" &
           Vict.Sex != "" & Vict.Descent != "")

theft_identity_vics <- theft_identity_noblanks %>%
  group_by(Vict.Descent, Vict.Sex) %>%
  summarise(count = n()) %>%
  arrange(desc(count))
```

Burglary Our analysis found that **white males had the highest count of reported incidents by far** - 1,479 - or **28%**. White females were behind them at a distant 788 (15%). Together, roughly **43%** of all reported cases between 2020 and 2024 were white men and women. When running a comparison check with the percentage of all other races/descents, the next highest is "Other" with 17%, which is considerably lower. Below is an example of the code used to determine this.

```
burg_descent_percentage <- burglary_vics_noblanks %>%
  group_by(Vict.Descent) %>%
  summarise(Count = n()) %>%
  mutate(Percentage = (Count / sum(Count)) * 100) %>%
  arrange(desc(Percentage))
```

Therefore, it can reasonably be said that the main group targeted in burglary incidents between 2020 and 2024 were white males and females. However, this definitely does not account for all situations and may be a misleading statement. Gender and race discrimination, fear of retaliation, as well as access to resources, may play a role in why victims of other races did not *report* as many incidents - not necessarily that more incidents didn't occur. In addition, the demographics of each county may affect the proportion of victims of one race versus another. Additional data and analysis would be needed.

Domestic Violence Our analysis for domestic violence shows that **Hispanic females** had by far the **highest number of reported incidents (52%)**. By comparison, the group with the second highest number, African American women, only had 413 reported incidents. That is roughly a 64% decrease from the highest to the second highest, implying that the majority of the cases fall under Hispanic women. This, of course, has its limitations. There may very well be more incidents that have taken place among other ethnicities; however, location, fear of retaliation, county demographics, and access to resources may affect how many are reported.

Conclusion

In examining this dataset, we were able to find several key observations about crime in Los Angeles County over the last four years. First, we found that the crime rate actually decreased by 87% from 2020 to 2023, implying that the current measures in place are positively affecting the crime rate.

Next, we discovered that stolen vehicles received the most reported incidents, followed by identity theft and burglary. Our client was interested in seeing which gender/ethnic groups were most targeted. For vehicles stolen, we were unable to find data about many of the victims' gender or race. Therefore, we studied the affected counties' demographic information as an estimate, and determined that in Northeast County, it is possible that a large portion of the victims were white males or females. However, in Hollenbeck County, 50% of the population was found to be Hispanic. As a result, there is a high chance that the majority of incidents occurred with Hispanic victims. Of course, there are certainly limitations to this data and those statements, and more research would need to be conducted.

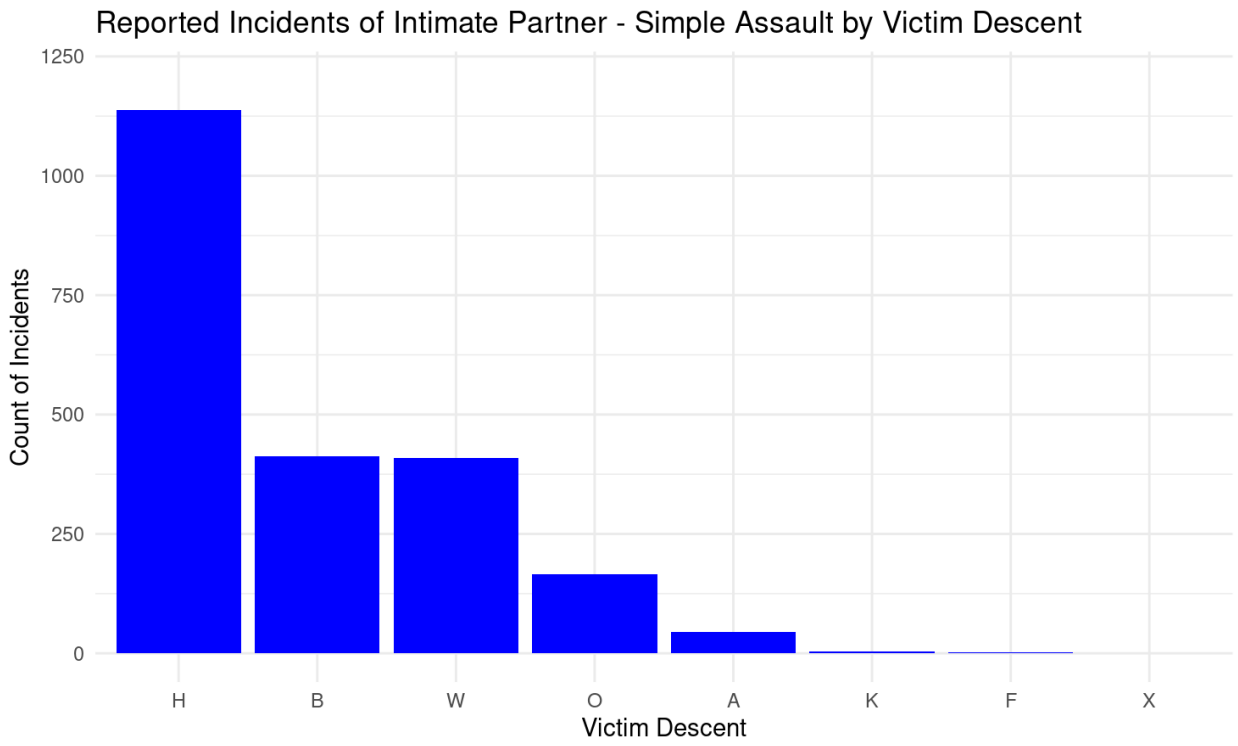

Figure 2: domestic violence data

For identity theft and burglary, we learned that 43% of all reported cases involved white male or female victims. Hispanic females totaled the highest percentage of domestic violence cases (52%). For all of these results, social and economic status, location, fear of retaliation, and access to resources may impact the exact results.

Below is a summary of the analysis results.

- **87% decrease** in crime from 2020 to 2023.
- **Vehicle stolen** was the most common crime reported during the four year span, followed by **identity theft** and **burglary**.
- Due to data limitations, we could not obtain gender/race information for the majority of vehicle theft victims. However, 50% of the population in Hollenbeck is Hispanic, leading to the possibility that the majority of the victims of this crime were Hispanic. Additional information and research would be required to be certain.
- **43%** of identity theft victims were **white**.
- **43%** of burglary victims were **white**.
- **52%** of domestic violence victims were **Hispanic females**.