Analysis of Crime Report Frequency and Reporting Behavior of UCR Part I Crimes in Austin, TX

Carrie Liang (sl54435)   SDS 322E, University of Texas at Austin, Spring 2024

2024-03-31

1. Introduction

Safety has long been an important factor to consider when people move to new places. Being the capital of Texas, a growing technology hub, and home to one of the biggest and most reputable public universities in the United States, Austin is certainly attracting a lot of attention, but it is not immune to the problem of crime. Major cities often have high crime rates; however, empirical and experiential evidence suggest that college campuses are a "safety bubble" of sorts, shielded from the crimes happening in the rest of the city [1]. As a student of the University of Texas at Austin, I am naturally concerned about the state of campus safety and crime in the rest of Austin, which influence my decision to live here after college. This sentiment is reflected by many of my peers, who were propelled by multiple incidents occurring on or near campus in late February and early March to launch discussions surrounding this topic on the UT Austin subreddit [2]. As expressed in this Reddit post, one of the factors causing the most anxiety is the lack of clarity and accessible information regarding crimes around campus.

Therefore, in this report, I will examine trends in crime reporting frequency and behavior in the City of Austin's Crime Reports dataset, with a focus on reports filed by the Austin Police Department (APD) in 2023. Each row in the dataset represents a distinct crime report, identified by incident number. Some types of information included in each report, represented as variables in the dataset, include the Highest Offense Description/Code, Location, Occurred Date, Report Date, and Clearance Date. The analysis will focus on what the Federal Bureau of Investigation (FBI) considers to be "Part I" crimes in its Uniform Crime Reporting (UCR) program, which the FBI acknowledges to be "serious crimes" that "occur with regularity in all areas of the country" and are "likely to be reported to the police" [3].

Questions this analysis will attempt to answer include:

1. How is the type of crime reported distributed in Austin, and how does it differ in areas of the city likely to be populated or visited by students? (What types of crime should Austin residents be worried about, and in which regions, as measured by number of reports?)
2. How quickly are Austin residents reporting crimes they encounter, and does this differ for the type of crime experienced?
3. How do different factors relate to crime rates in the city and, if a crime is encountered, the length of time someone can expect to wait after reporting it for APD to clear the crime?

I do not expect trends in Austin to deviate from people's intuitions regarding where the safer or more dangerous parts of the city are (for example, Downtown commonly being regarded as dangerous) and best practices to deal with crime (for example, reporting as soon as possible). However, I wish to provide in this analysis statistics that are more specific to Austin and examine the strength, not just existence, of relationships between different factors related to crime reports. Knowing how Austin residents tend to report crimes can be a measure of what types of crimes people are most likely to encounter in Austin and useful information in implementing educational programs.

2. Methods

R Packages

This analysis uses the `tidyverse` package for data wrangling and visualization, and the `ggrepel` to customize visualization labeling.

Obtaining Data

The raw data used in this report was obtained from the Crime Reports dataset in the City of Austin Open Data Portal on February 21, 2024 at 1:16 PM. The queries run were:

- `Report Date Time` between **2023 Jan 01 00:00:00 AM** and **2023 Dec 31 12:00:00 AM**
- `Clearance Status` is one of **C**, **N**, **O**

```
# Reading in the dataset
crimereports <- read_csv("C:/Users/CL/Desktop/SDS 322E/Project/Crime_Reports_20240221.csv")

# Examine basic information about dataset
nrow(crimereports)
## [1] 74103
ncol(crimereports)
## [1] 33
head(crimereports,n=3)
## # A tibble: 3 x 33
##   `Incident Number` `Highest Offense Description` `Highest Offense Code`
##               <dbl> <chr>                                          <dbl>
## 1       20239001489 BURGLARY OF RESIDENCE                            500
## 2       20239002739 BURGLARY OF VEHICLE                              601
## 3       20231280121 AUTO THEFT                                       700
## # i 30 more variables: `Family Violence` <chr>, `Occurred Date Time` <chr>,
## #   `Occurred Date` <chr>, `Occurred Time` <dbl>, `Report Date Time` <chr>,
## #   `Report Date` <chr>, `Report Time` <dbl>, `Location Type` <chr>,
## #   Address <chr>, `Zip Code` <dbl>, `Council District` <dbl>,
## #   `APD Sector` <chr>, `APD District` <chr>, PRA <dbl>, `Census Tract` <dbl>,
## #   `Clearance Status` <chr>, `Clearance Date` <chr>, `UCR Category` <chr>,
## #   `Category Description` <chr>, `X-coordinate` <dbl>, ...
# Verify there are no duplicates
n_distinct(crimereports$`Incident Number`)
## [1] 74103
```

In the original dataset downloaded from the portal (renamed as `crimereports`), there are **74103 rows**, each being a crime report uniquely identified by the `Incident Number`, and **33 columns**, each variable being an attribute of each reports. The second filter for clearance status was added to align with the three clearance statuses (C, N, and O) mentioned in the Crime Report dataset's Field Descriptions documentation, to avoid including cases with non-standard clearance statuses not mentioned in the documentation.

It should be noted that assuming the original Crime Reports dataset is updated weekly with the latest reports filed that week, filtering for reports filed in 2023 in theory would ensure that the same query run on a later date would return the same dataset. However, this does not seem to be the case, as the same queries run on March 14, 2024 returned 74301 rows of observations. In order to obtain the most complete results, it is sensible to re-run the code in this analysis when the City releases their Crime Reports 2023 dataset, as they have done in past years. Therefore, results from this analysis are preliminary, and should be used as a reference only.

Creating Outcome Variables

Firstly, I created a variable representing the difference between the `Occurred Date` and the `Report Date`, which is how long (in days) it took since the original crime occurred for it to be reported. I created another variable using the same procedure, representing the difference between the `Clearance Date` and the `Report Date`, which is how long (in days) it took since the original crime was reported to get it cleared.

```
crimereports <- crimereports |>
  # Calculate days it took to report
  mutate(report_time = as.numeric(mdy(`Report Date`) - mdy(`Occurred Date`))) |>
  # Calculate days it took to clear from occurrence, only for crimes with cleared status
  mutate(clear_time = if_else(`Clearance Status` %in% c("C","O"),
          # Else, the report should not have a clearance date, so return NA
          as.numeric(mdy(`Clearance Date`) - mdy(`Report Date`)),NA))

# View summary statistics
crimereports |>
  select(clear_time,report_time) |>
  summary()
##    clear_time        report_time
##  Min.   :-6775.00   Min.   :   0.00
##  1st Qu.:    0.00   1st Qu.:   0.00
##  Median :    1.00   Median :   0.00
##  Mean   :    8.96   Mean   :  18.23
##  3rd Qu.:    5.00   3rd Qu.:   3.00
##  Max.   :  384.00   Max.   :7340.00
##  NA's   :57962
# Calculate proportion of cases with nonnegative clear_time
mean(crimereports$clear_time >= 0,na.rm=TRUE)
## [1] 0.9996902
```

There were a few negative values, less than 0.1% of the data, when calculating the difference between `Clearance Date` and `Report Date`. I hypothesize this means that crimes were reported after they had already been cleared or resolved, but in this analysis, I am more interested in how long it takes for unresolved crimes to be cleared after a report is made to the police. Because there is only a small proportion of cases which had a negative `clear_time`, I will ignore them in the analysis. Additionally, when I was initially examining the dataset, there were some cases marked to be "N" (not cleared) that still had a clearance date. I have attempted to contact the owner of the dataset (Austin Police Department Public Information Office) to request some insight into both of these phenomena, but have not yet gotten a response as of the time of writing.

Regrouping Categories

In the dataset, each crime report has an associated highest offense and Uniform Crime Reporting (UCR) category/description. Because there are too many specific highest offense types (as categorized by APD) to easily compare, I chose to look at the broader UCR categories instead. According to the documentation of the dataset, the `UCR Category` variable represents "code for the most serious crimes identified by the FBI as part of its Uniform Crime Reporting program", and `Category Description` has the associated descriptions for each UCR category code. Thus, utilizing the UCR categories is a good way to compare major, broader types of crime.

These "most serious" crimes align with what the FBI calls "Part I" crimes in its UCR program (which is the terminology I will use from now on in this report). Below is code summarizing how `Category Description` is distributed as it originally appeared in the data:

```
crimereports |>
  group_by(`Category Description`) |>
  summarize(count=n()) |>
  arrange(desc(count))
## # A tibble: 8 x 2
##   `Category Description` count
##   <chr>                  <int>
## 1 <NA>                   40003
## 2 Theft                  20670
## 3 Auto Theft              5179
## 4 Burglary                4286
## 5 Aggravated Assault      2639
## 6 Robbery                  818
## 7 Rape                     465
## 8 Murder                    43
```

For the purposes of this project, which require at most 4 categories for a categorical variable, I decided to do the following:

1. combine "Theft", "Auto Theft", and "Robbery" (these are similar in nature, involving stealing, may involve the use of force. Burglary, on the other hand, involves breaking and entering.)
2. ignore "Murder" cases (there are very little cases compared to the total number of cases in all of Austin, and thus may be harder to compare with other categories)

This will keep "Theft/Robbery","Aggravated Assault","Burglary", and "Rape" as four categories of UCR crime reports to be analyzed.

The first part of the code below is for recoding these variables, just keeping reports associated with the four (merged) UCR categories, and storing the result in a dataframe `allUCR`. The line afterwards stores all reports in `allUCR` that *have been cleared* in a dataframe `clearedUCR`.

```
allUCR <- crimereports |>
  mutate(UCR=recode(`Category Description`, # Recode UCR categories
                    "Theft" = "Theft/Robbery",
                    "Auto Theft" = "Theft/Robbery",
                    "Robbery" = "Theft/Robbery")) |>
  filter(!is.na(UCR),UCR!="Murder")

clearedUCR <- allUCR |> filter(!is.na(clear_time))

# Count number of observations
nrow(allUCR)
## [1] 34057
nrow(clearedUCR)
## [1] 3924
```

It turns out that `allUCR` has 34057 rows and `clearedUCR` has 3924 rows. Both dataframes are considered "tidy" because each variable has its proper column, and each observation (crime report) is in its own row.

Joining Demographic Information

To add additional insight to the analysis, I joined information from the 2023 Austin Council District Demographic Data dataset about the total population (from the 2020 census), median household income, and average closing rent, by Council District (numbered 1-10 in Austin) to `allUCR`.

```r
demographics <- read_csv("C:/Users/CL/Desktop/SDS 322E/Project/2023_Austin_Council_District_Demographic

councilJoined <- allUCR |>
  group_by(`Council District`,UCR) |>
  summarize(count = n()) |>
  # Creates count of crimes by UCR category for each council district
  pivot_wider(names_from=UCR,values_from=count) |>
  mutate(total=`Aggravated Assault`+`Burglary`+`Rape`+`Theft/Robbery`) |>
  # Join demographics data
  left_join(demographics,by=c(`Council District`="District"))
```

The `councilJoined` dataset has 11 rows, representing features of each of the 10 city council districts, with one row for `NA` values.

Finally we will be using a colorblind-friendly color palette for all visualizations.

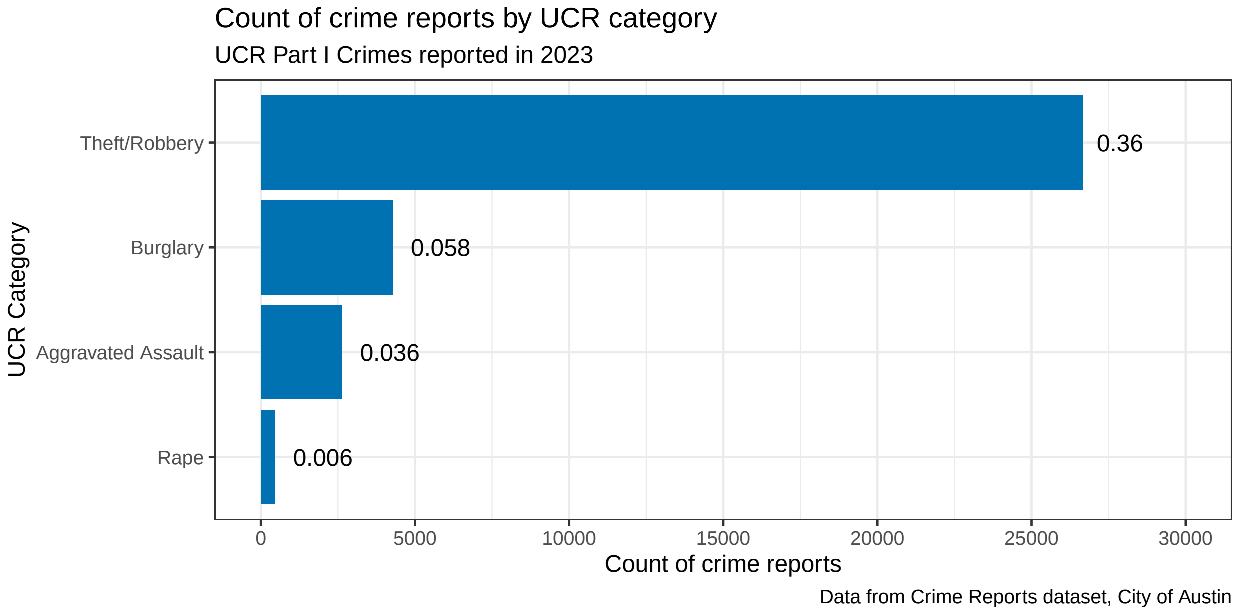
```r
color_blind_friendly <- c("#56B4E9","#F0E442","#CC79A7","#009E73")
```

3. Results

a. Distribution of UCR Part 1 crimes

We first examine the overall distribution of UCR Part I Crimes (`UCR`), out of the `allUCR` dataframe.

```r
# Create a bar graph of UCR Part I Crimes ("UCR" variable)
allUCR |>
  group_by(UCR) |>
  summarize(count=n(),prop=n()/nrow(crimereports)) |>
ggplot(aes(x=count,y=reorder(UCR,count))) +
  geom_bar(stat="identity",fill="#0072B2") +
  geom_text(aes(label = round(prop,3)), hjust = -0.3) +
  scale_x_continuous(limits=c(0,30000),breaks=seq(0,30000,5000)) +
  theme_bw() +
  labs(x="Count of crime reports",y="UCR Category",
       title="Count of crime reports by UCR category",
       subtitle="UCR Part I Crimes reported in 2023",
       caption="Data from Crime Reports dataset, City of Austin")
```



```
# Calculate proportion of total cases this represents
nrow(allUCR) / nrow(crimereports)
## [1] 0.45959
```

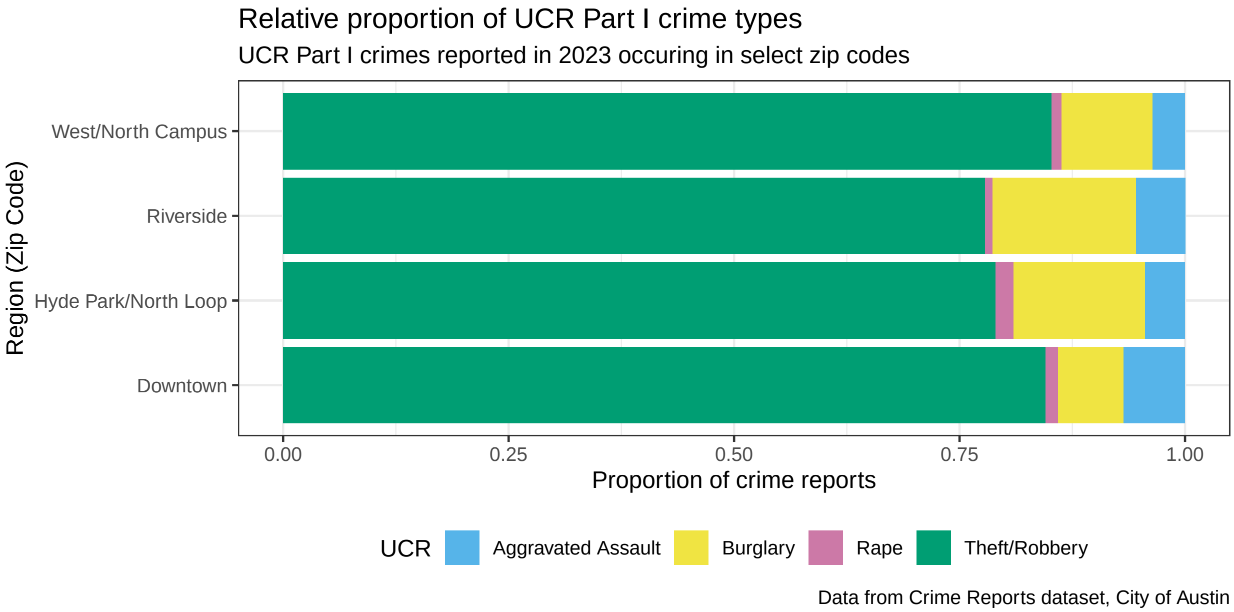
This bar graph compares the counts of crime reports associated with each Part I Crime type. To the right of each bar is the proportion of total 2023 crime reports each type represents. Theft and Robbery cases combined are overwhelmingly the most reported type of Part I case, and they form 36% of all Austin crime reports made in 2023. Burglary cases are reported at the second highest frequency, followed by Aggravated Assault cases. Rape cases are the least reported type of Part I crime, totaling only 0.6% of all crime reports. These “most serious” crimes (excluding murder) represented in this chart form nearly 46% of all Austin crime reports in 2023.

b. Comparing relative frequencies of types of Part I crimes occuring in selected zip codes near UT Austin

Next, we compare four regions likely to be populated or visited by UT Austin students—Downtown (78701), Riverside (78704), West/North Campus (78705), and Hyde Park/North Loop (78751) —and whether there are differences in types of crime reported in these locations.

```
centralCrimes <- allUCR |>
  # Select reports from four zip codes near UT Austin
  filter(`Zip Code` %in% c(78705,78751,78701,78704)) |>
  rename(Region = `Zip Code`) |>
  mutate(Region = recode(Region,
         `78701` = "Downtown",
         `78704` = "Riverside",
         `78705` = "West/North Campus",
         `78751` = "Hyde Park/North Loop"))

# Create a segmented bar graph to compare
ggplot(centralCrimes) +
  geom_bar(aes(y=Region,fill=UCR),position="fill") +
  theme_bw() +
  theme(legend.position="bottom") +
  scale_fill_manual(values=color_blind_friendly) +
  labs(x="Proportion of crime reports",y="Region (Zip Code)",
       title="Relative proportion of UCR Part I crime types",
       subtitle="UCR Part I crimes reported in 2023 occuring in select zip codes",
       caption="Data from Crime Reports dataset, City of Austin")
```



Relative proportion of UCR Part I crime types

UCR Part I crimes reported in 2023 occuring in select zip codes

Data from Crime Reports dataset, City of Austin

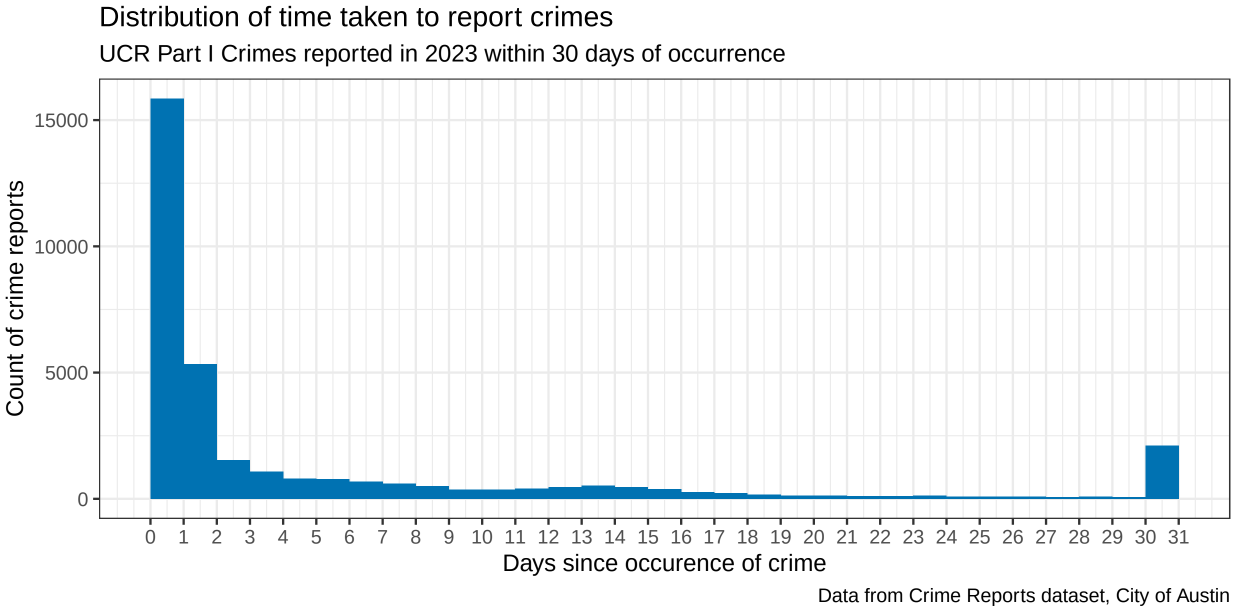
```
# Create count of crimes by UCR category
centralCrimes |>
  group_by(Region,UCR) |>
  summarize(count = n()) |>
  pivot_wider(names_from=UCR,values_from=count) |>
  mutate(total=`Aggravated Assault` + Burglary + Rape + `Theft/Robbery`)
## # A tibble: 4 x 6
## # Groups:   Region [4]
##   Region               `Aggravated Assault` Burglary  Rape `Theft/Robbery` total
##   <chr>                               <int>    <int> <int>           <int> <int>
## 1 Downtown                              148      156    30            1823  2157
## 2 Hyde Park/North Loop                   31      102    14             552   699
## 3 Riverside                             131      383    20            1874  2408
## 4 West/North Campus                      33       91    10             771   905
```

In specific regions, the frequencies of different crime types reported are still similar to the overall distribution found in (a). In all four regions, theft/robbery cases are reported at the highest frequency. In Downtown, it seems that Aggravated Assault and Burglary are reported at similar rates, while in the other three regions, Burglary is reported more often. In West/North Campus, where residents are mainly UT Austin students, cases related to stealing of property (theft, robbery, burglary) combined are slightly more frequently reported than in other regions. From the summary table, Downtown and Riverside which are known to be “dangerous” areas of the city indeed have higher combined counts of crime reports than in the two other regions, at approximately triple the number.

c. Distribution of time taken to report UCR Part I crimes

We examine the distribution of time (`report_time`) taken to report UCR Part 1 crimes, again out of the `allUCR` dataframe.

```
# Create a histogram representing distribution of "report_time"
allUCR |>
  # Adjust values by a little bit for cleanliness of visualization
  # Put 30+ values into the 30-31 bin
  mutate(report_time = if_else(report_time > 30,30.1,report_time+0.1)) |>
ggplot(aes(x=report_time)) +
  geom_histogram(binwidth=1,center=0.5,fill="#0072B2") +
  scale_x_continuous(breaks=seq(0,31,1)) +
  theme_bw() +
  labs(x="Days since occurence of crime",y="Count of crime reports",
       title="Distribution of time taken to report crimes",
       subtitle="UCR Part I Crimes reported in 2023 within 30 days of occurrence",
       caption="Data from Crime Reports dataset, City of Austin")
```



```
# Calculate summary statistics
allUCR |>
  mutate(onemonth = if_else(report_time<=30,"Within one month","After one month")) |>
  group_by(onemonth) |>
  summarize(min_days = min(report_time),
            median_days = median(report_time),
            max_days = max(report_time),
            iqr_days = IQR(report_time),
            prop_cases = round(n()/nrow(allUCR),4)) |>
  # Add a row reporting statistics for all UCR reports (unseparated)
  rbind(c("ALL",min(allUCR$report_time),median(allUCR$report_time),
          max(allUCR$report_time),IQR(allUCR$report_time),1))
## # A tibble: 3 x 6
##   onemonth         min_days median_days max_days iqr_days prop_cases
##   <chr>            <chr>    <chr>      <chr>    <chr>    <chr>
## 1 After one month  31       57         7340     85       0.0595
## 2 Within one month 0        1          30       4        0.9405
## 3 ALL              0        1          7340     6        1
```

This histogram displays the distribution of the days it took since the occurrence of a crime for it to be reported. For crimes which were reported after one month (30 days), they were included in the 30-day bin in the histogram, effectively making that bin show the proportion of cases reported in 30+ days, which represent approximately 6% of crime reports made in 2023.

The histogram is extremely skewed right—among all crime reports, the median value of `report_time` was 1, as indicated by the table of summary statistics (units reported in days). As the number of days since the occurrence of a crime increases, the count of crime reports generally decreases approximately exponentially, excluding a small bump at around 12-14 days. The interquartile range is 6 days. The maximum number of days was 7340 (more than 20 years), so not regrouping these extreme values would have made the histogram appear even more stretched out and skewed right.

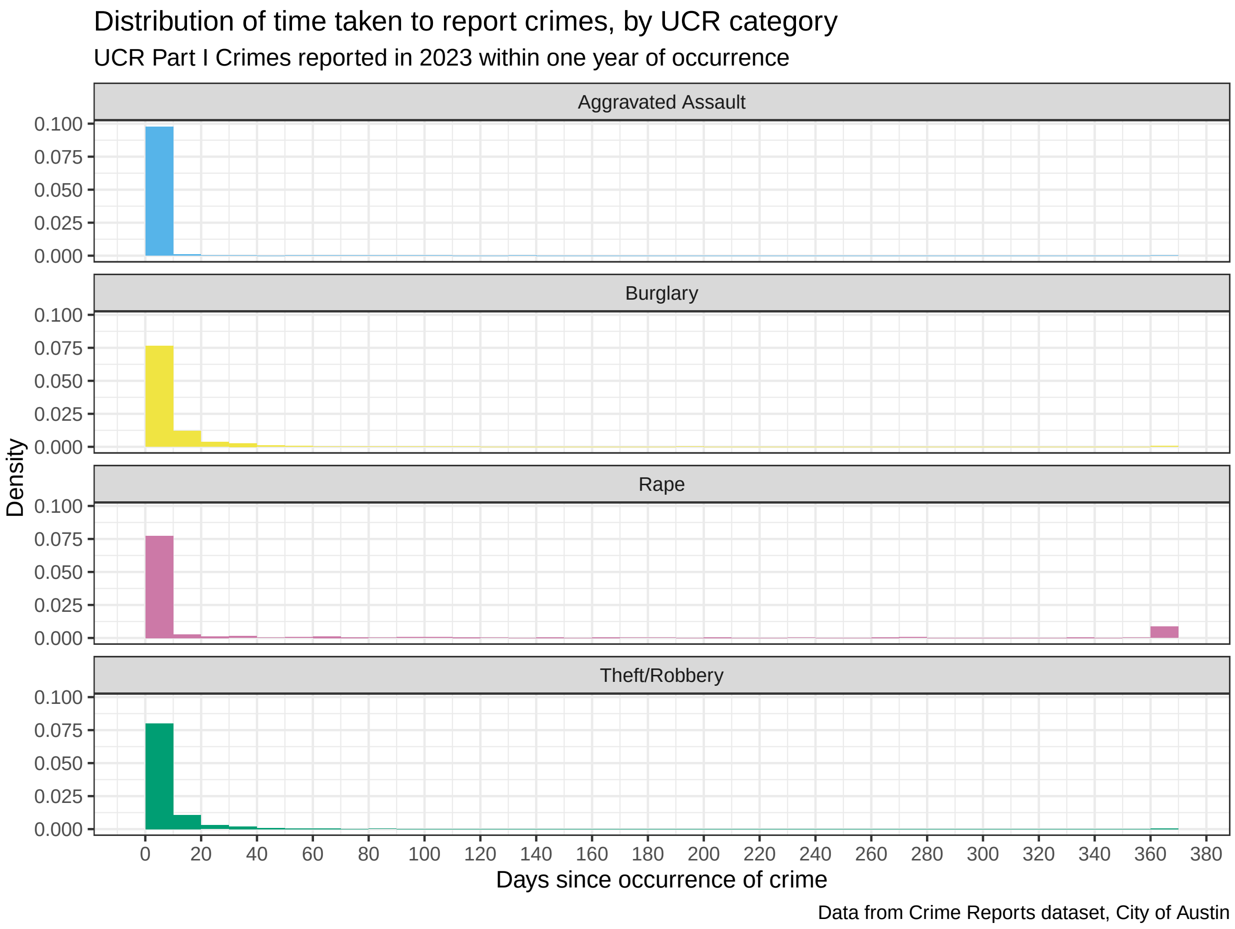
d. Relationship between type of UCR Part 1 crime and days taken to report the crime

Upon first glance, the Rape category has a significantly higher proportion (8%) of crimes that are reported long after the crime occurs (here, the statistic is calculated with 1 year = 365 days):

```
# Calculate summary statistics
allUCR |>
  group_by(UCR) |>
  summarize(min_days = min(report_time),
            median_days = median(report_time),
            max_days = max(report_time),
            iqr_days = IQR(report_time),
            prop_year = mean(report_time > 365))
## # A tibble: 4 x 6
##   UCR                min_days median_days max_days iqr_days prop_year
##   <chr>                 <dbl>       <dbl>    <dbl>    <dbl>     <dbl>
## 1 Aggravated Assault        0           0     1060        0   0.00114
## 2 Burglary                  0           1     7340        8   0.00537
## 3 Rape                      0           0     6567        6   0.0817
## 4 Theft/Robbery             0           1     7307        6   0.00480
```

We compare the actual distributions of report time by UCR category with density histograms:

```
# Create density histograms, facet-wrapped by UCR category
allUCR |>
  # Adjust values by a little bit for cleanliness of visualization
  # Put 365+ values into the 360-370 bin
  mutate(report_time = if_else(report_time > 365,365.1,report_time+0.1)) |>
  ggplot(aes(x=report_time)) +
  geom_histogram(aes(y=after_stat(density),fill=UCR),binwidth=10,center=5,show.legend=FALSE) +
  facet_wrap(~UCR,ncol=1) +
  scale_x_continuous(breaks=seq(0,380,20)) +
  theme_bw() +
  scale_fill_manual(values=color_blind_friendly) +
  labs(x="Days since occurrence of crime",y="Density",
      title="Distribution of time taken to report crimes, by UCR category",
      subtitle="UCR Part I Crimes reported in 2023 within one year of occurrence",
      caption="Data from Crime Reports dataset, City of Austin")
```

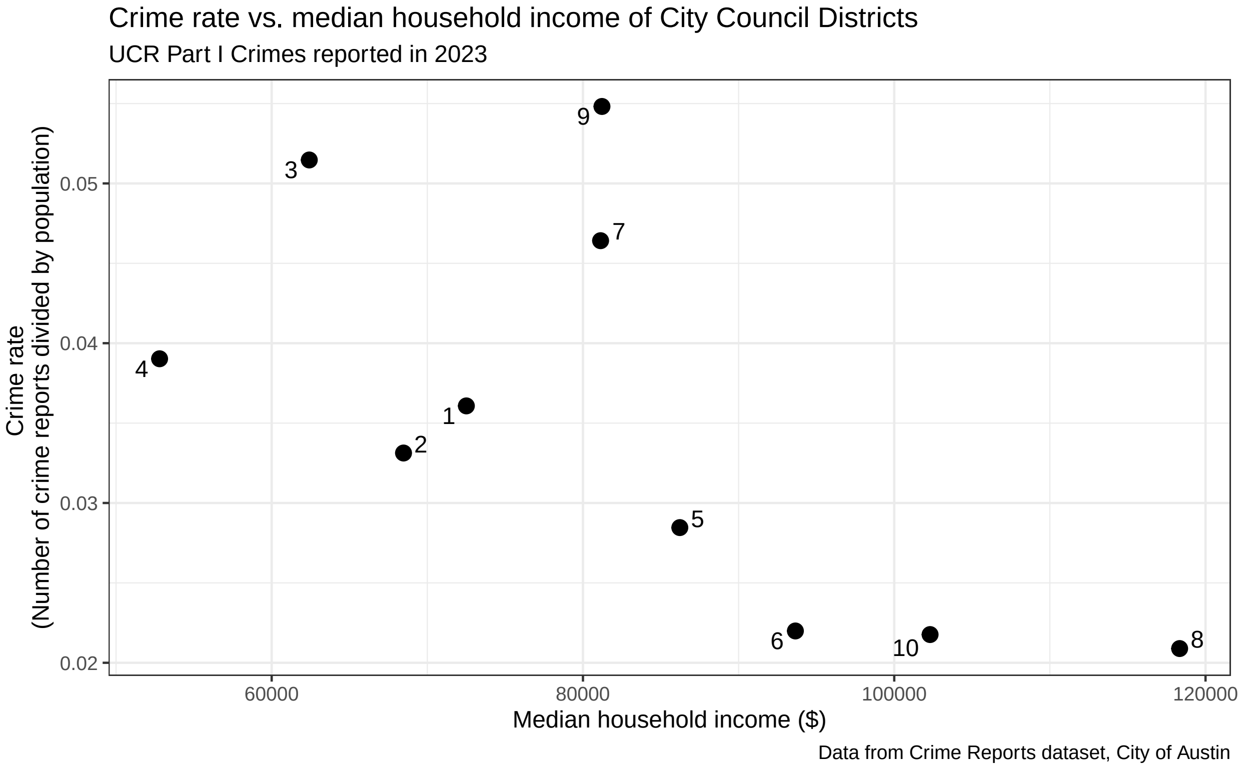


All histograms are skewed right, just as in the histogram in (b), with median `report_time` either 0 or 1 and very high outlier values. Aggravated Assault cases had the lowest maximum value for `report_time`, the interquartile range is 0, its histogram does not have a visible tail like the other ones, and approximately 0.1% of cases were reported after 1 year of occurrence, indicating that reports of this type of case tend to be filed the quickest. The shape of Burglary and Theft/Robbery cases' histograms are similar, and for both, about 0.5% of cases were reported after 1 year of occurrence. Rape cases have the highest rate (8%) of cases reported after 1 year, but the interquartile range is less than that for Burglary. Thus, it appears that there was still a substantial portion of Rape cases reported quickly (within about the first 10 days) compared to Burglary cases, which can also be seen in their histograms.

e. Relationship between median household income and Part I crime occurrence grouped by City Council district

Now that we know how Austin residents tend to report crimes, we take a step back to examine one factor that may make them more likely to experience crime, which is median household income. This demographic information was joined previously with `allUCR` to create the dataframe `councilJoined`.

```r
# Make scatterplot of median household income vs. crime rate
ggplot(councilJoined,aes(x=Median_HH_Income,y=total/Total_Population_2020_Census)) +
  geom_point(size=3) +
  geom_text_repel(aes(label = `Council District`), size = 4) +
  theme_bw() +
  labs(x="Median household income ($)",
       y="Crime rate \n(Number of crime reports divided by population)",
       title="Crime rate vs. median household income of City Council Districts",
       subtitle="UCR Part I Crimes reported in 2023",
       caption="Data from Crime Reports dataset, City of Austin")
```



As we are using a dataframe with only the UCR Part I crimes excluding murder, the crime rate in this case is calculated as the number of crime reports categorized under UCR Part I divided by the population of the district. There seems to be a slightly negative nonlinear correlation between the median household income and crime rate (counting only non-murder Part I crimes) of a district, for districts with median household income above $80,000. However, overall, there is no correlation between the two factors—it seems more plausible that there is a group of districts with high median household income and low crime rate, and another group of districts with lower median household income and high crime rate, rather than median household income directly influencing crime rate. As each City Council district represents a large number of people, it would be illogical to regard any of the points as an outlier.

Notably, District 9 has the highest UCR Part I crime rate at more than 5.5%, but is in the middle of all districts in terms of median household income. Districts 6, 8, and 10 all have similar UCR Part I crime rate at about 2%, and have the three highest median household incomes.

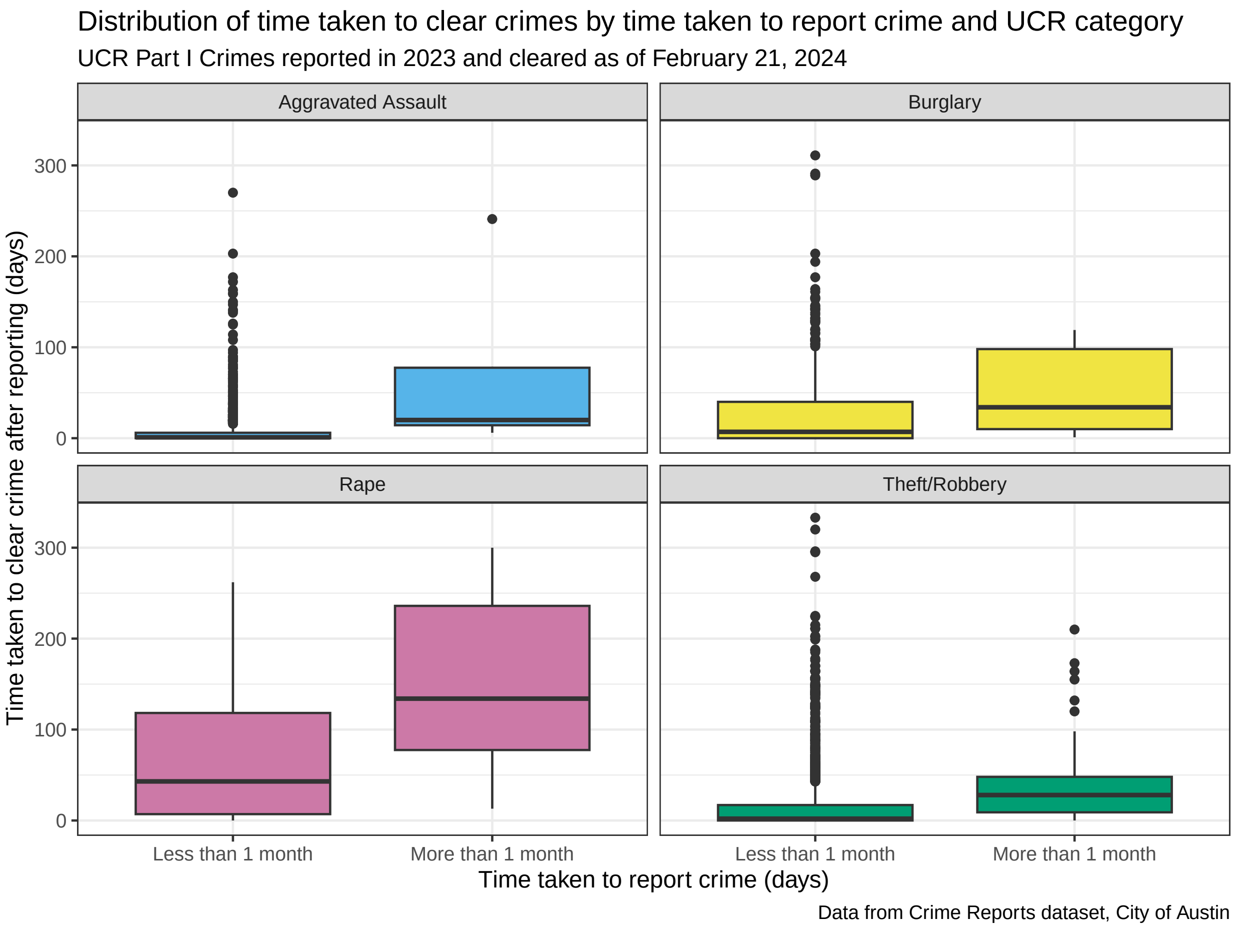
f. Relationship between whether a crime was reported within a month and distribution of time taken to clear the report

Finally we compare, for different categories of crime, the distribution of time taken to clear a crime report based on whether the crime was reported within a month. This section uses `clearedUCR` which has only crime reports marked with a cleared status ("C" or "O"). Due to extreme outliers, it is easier to investigate this relationship by grouping cases into those cleared in less or more than 1 month and making a boxplot, rather than creating a scatterplot of `clear_time` and `report_time`.

```
# Calculate summary statistics
clearedUCR |> filter(clear_time >= 0) |>
  mutate(month=if_else(report_time<=30,"Less than 1 month","More than 1 month")) |>
  group_by(UCR,month) |>
  summarize(min_days = min(clear_time),
            median_days = median(clear_time),
            max_days = max(clear_time),
            iqr_days = IQR(clear_time))
## # A tibble: 8 x 6
## # Groups:   UCR [4]
##   UCR                month             min_days median_days max_days iqr_days
##   <chr>              <chr>                <dbl>       <dbl>    <dbl>    <dbl>
## 1 Aggravated Assault Less than 1 month        0           1      270      6
## 2 Aggravated Assault More than 1 month        6          20      241     63.2
## 3 Burglary           Less than 1 month        0           7      311     40
## 4 Burglary           More than 1 month        1          34      119     88
## 5 Rape               Less than 1 month        0          43      262    111.
## 6 Rape               More than 1 month       13         134      300    158.
## 7 Theft/Robbery      Less than 1 month        0           2      333     17
## 8 Theft/Robbery      More than 1 month        0          28      210     39
```

From these statistics, it seems like for every category of crime, the median days in `clear_time` is higher for the group of crimes reported after 1 month. We examine this data more visually with grouped boxplots:

```
# Create grouped boxplot comparing distribution of time taken to clear crime
clearedUCR |> filter(clear_time>=0) |>
  # for whether crime was reported within a month
  mutate(month=if_else(report_time<=30,"Less than 1 month","More than 1 month")) |>
ggplot() +
  geom_boxplot(aes(x=month,y=clear_time,fill=UCR),show.legend=FALSE) +
  facet_wrap(~UCR) +
  scale_fill_manual(values=color_blind_friendly) +
  theme_bw() +
  labs(x="Time taken to report crime (days)",y="Time taken to clear crime after reporting (days)",
      title="Distribution of time taken to clear crimes by time taken to report crime and UCR category"
      subtitle="UCR Part I Crimes reported in 2023 and cleared as of February 21, 2024",
      caption="Data from Crime Reports dataset, City of Austin")
```


Distribution of time taken to clear crimes by time taken to report crime and UCR category

UCR Part I Crimes reported in 2023 and cleared as of February 21, 2024

Aggravated Assault

Burglary

Rape

Theft/Robbery

Data from Crime Reports dataset, City of Austin

For all categories of crime, the middle 50% of `clear_time` is slightly higher for crimes reported after 1 month than for crimes reported within 1 month, suggesting that crimes reported soon after occurrence are typically cleared quicker. For crimes reported after 1 month, the IQR of `clear_time` is greater; there is greater variance in how long it will take for crimes to be cleared. Overall, Rape cases have significantly higher variance in `clear_time` compared to other categories and take longer to clear—50% of cases reported within 1 month are cleared in more than about 50 days, and 50% of cases reported after 1 month are cleared in more than about 130 days. Meanwhile, Aggravated Assault and Theft/Robbery have the least variance in `clear_time` and for both, about 75% of cases reported within 1 month are cleared within about 20 days.

4. Discussion

Summary of findings

APD labels UCR Category for its crime reports, for the "most serious crimes identified by the FBI as part of its Uniform Crime Reporting program", which are also known as Part I crimes. Theft, robbery, and burglary cases combined are by far the most common types of UCR Part I crimes reported in Austin. Victims of aggravated assault cases are more likely to report them quickly, but a significant proportion of rape victims may wait a long time before reporting their case. It is generally better for residents to report cases sooner rather than later, as cases reported within 1 month of occurrence were on average cleared in a shorter period of time than cases reported after 1 month. The City of Austin may consider encouraging education surrounding crime and best practices when reporting crimes to the Austin Police Department, especially for categories of crime like rape where victims tend to wait longer before reporting and thus experience a higher wait time before the crime is cleared. More specifically, for each research question:

1. *How is the type of crime reported distributed in Austin, and how does it differ in areas of the city likely to be populated or visited by students?*

   Austin residents should be most worried about cases related to the stealing of property (theft, robbery, and burglary)—they make up about 42% of all crime reports made in 2023 (Figure (a)). For UT Austin students, who are likely to live in or visit West/North Campus, Riverside, Hyde Park/North Loop, and Downtown regions, this is no different, and they should be most cautious against theft/robbery cases, which have the highest number of reports (Figure (b)). Riverside and Downtown which are reputed to have high crime rates indeed have more of these "most serious" crime reports than the more residential West/North Campus and Hyde Park/North Loop regions, at about triple the amount in 2023 (Table/Figure (b)). Students should be warned that these regions further from campus have a higher risk of crime, and make necessary preparations.

2. *How quickly are Austin residents reporting crimes they encounter, and does this differ for the type of crime experienced?*

   Most UCR Part I crimes are reported within a day of it occurring; as the number of days since occurrence of crime increases, the number of cases reported generally decreases in an approximately exponential manner (Figure (c)). This pattern is the same when categorizing cases by type of Part I crime experienced, but Rape cases tend to have a higher proportion of cases reported long after they occur, while Aggravated Assault cases tend to be reported the quickest (Figure (d)). More research should be done into potential causes of why victims of certain types of crimes may take longer to report the crime, as this is related to answers to the third research question below.

3. *How do different factors relate to crime rates in the city and, if a crime is encountered, the length of time someone can expect to wait after reporting it for APD to clear the crime?*

   It is often believed that richer areas are safer, but when comparing the median household income and Part I crime rate by City Council districts, there was not a significant correlation. The most that can be said is that some richer districts tend to be safer, but it is not always the case that the richer a district is (as measured by median household income), the lower the crime rate is (Figure (e)). All residents should be educated on what to do if they do encounter a crime, and should not automatically assume that some districts richer than others are automatically less susceptible to crime. This plot can be made better if data for each zip code rather than City Council district was plotted.

   It is generally better to report the crime as soon as possible, for example within one month, for faster clearance times after reporting (Figure (f)). Aggravated Assault cases tend to be cleared quickly while Rape cases on average take much longer than any other category to be cleared (Figure (f)). Combining this with results above, Rape victims in particular should be aware that their cases not only tend to wait the longest to report, but also are slowest to be cleared if a long time has passed since the crime.

Ethical considerations

Crime is a sensitive issue and victims may not always wish to have their information revealed publicly. The most identifying piece of information in each report of the Crime Reports dataset is the location and longitude/latitude; the victims/aggressors' personal information such as name is not included. While including other attributes of victims/aggressors in each crime report would make it possible to do more detailed analysis of how different demographic information influences crime, this information is not always possible to collect or publicize, and emphasizing certain trends may lead to hostilities against certain social groups for being more correlated with crime. Additionally, any extreme outliers in this data in terms of report or clearance time have still reflected real, living people, so it is important to not exclude them and keep in mind that each crime report is not merely a data point.

In this report, the findings about report counts by categories are somewhat of a measure of crimes Austin residents are more likely to encounter, so this information lets them take any precautions as necessary. It also reveals that crimes reported after a month of occurrence generally take longer to clear (counting from when the report was made). This suggests to Austin residents that if they experience a crime, it is better to report it to APD as soon as possible (at least for UCR Part I crimes), for potentially shorter clearance times. One potential reason for this phenomenon is lack of evidence—it becomes harder to gather evidence the longer it has been since the crime has occurred, which is a situation often reported in the news for Rape cases. The City of Austin may consider increasing efforts to educate residents regarding how to recognize crime (especially Rape) and obtain evidence, so that they can report the case quickly and have a higher chance of clearing the case sooner.

Reflections on the state of the Crime Reports dataset

Overall, the City of Austin's Crime Reports dataset is tidy and has a lot of information about each crime report. It is understandable that there is some missing information for certain crime reports (for example, those which have occurred long ago and are impossible to remember the location or time of). However, I think it would be more useful for people looking at this dataset to have more detailed information in the dataset documentation to reference. For example, as I mentioned above, I could not find a clear explanation as to why some reports were marked as "N" for not cleared yet still had a clearance date, and how reports' clearance dates could be before their reporting dates. It would also be good to have more publicly available context regarding APD's methodology for making reports (for example, are reports in the dataset ever updated, does the report date refer to when a report was made to APD or when APD created a write-up, etc.). This information would change how I, and others, would interpret any statistics or visualizations generated from the data.

Potential problems

One issue with figure (e) is that there are only 10 districts, so each district is very large and thus not as specific as regions marked by zip code. For example, district 9 has the highest Part I crime rate, but it encompasses both the UT Austin region and Downtown Austin, which vary greatly in level of crime, so district 9's crime rate does not accurately reflect these differences within the region. I think that Austin residents would be most concerned with which general areas of the city are more dangerous, but "area" can be vague and not always correspond perfectly with how location data/demographics are structured in City datasets.

Some other things to note are that these represent only crimes which have been reported to APD. There may be a significant portion of cases that go unreported. It is very likely that unreported cases will change the statistics in this report. Additionally, UCR's Part I "Rape" category does not include sexual offenses that are not rape or human trafficking/commercial sex acts. Even though there are relatively fewer Rape cases compared to Theft/Robbery/Burglary, this does not mean that offenses of a sexual nature are uncomon in Austin. Finally, definitions of crime and which crimes are most serious are not always consistent, so these results only pertain to what was marked with UCR categories in the City of Austin's Crime Reports dataset for reports made in 2023. UCR categories were chosen as so to make it easier to compare crime uniformly across the nation, and do not necessary encompass all categories/nuances of crime [4].

5. Reflection, Acknowledgements, and References

Reflection

The most challenging part of any exploratory data analysis project, including this one, is the "exploration" part—learning what the dataset I am working with is like, finding inconsistencies, coming up with ways to appropriately treat those inconsistencies. I spent the most time in the initial stages of this process, figuring out how to interpret the Crime Reports dataset in context and link my research questions together. There were many outliers in terms of reporting and clearance time, so tinkering with different regroupings to make the cleanest visualizations required some effort. It was important for me to refrain from wrangling the dataset in a way that it would show results that I wanted to see (for example, very clear trends between variables), and let the data itself demonstrate what patterns were hidden inside.

I learned from this process that real-world data is much more harder to deal with than the clean data we are given in the classroom for instructional purposes. Seeing extremely clear or significant relationships as we do in class is rare, and as data scientists, we often need to settle for less pretty results. While my data did not reveal much more than what we commonly know about crime, it provided some specific statistics closer to home, which I think is still valuable to know. In the process, I also learned some additional techniques to customize visualizations with labels, and I believe I have improved at summarizing data to make visualizations. This project was a valuable in-depth dive into a large, relevant real-world dataset and really tested the skills I have learned so far in SDS 322E. I really enjoyed seeing how data science has the potential to reveal insights that could have a real impact on improving our society.

Acknowledgements

All data used in this report was obtained from the City of Austin's Open Data Portal. The Crime Reports dataset is maintained by the Austin Police Department Office of Public Information. The 2023 Austin City Council District Demographic Data dataset is owned by Dan Brooks.

This project was completed with guidance from the instructional team of SDS 322E Section 56665, Spring 2024 at the University of Texas at Austin, including professor Layla Guyot and all teaching assistants.

References

Crime Reports dataset, APD-PIO. https://data.austintexas.gov/Public-Safety/Crime-Reports/fdj4-gpfu. (See the associated CrimeReportFieldDescriptions.pdf attachment for details on what each variable in the dataset means).

2023 Austin City Council District Demographic Data dataset, Dan Brooks. https://data.austintexas.gov/City-Government/2023-Austin-Council-District-Demographic-Data/puux-7swp.

[1] Henson, Verna A, and William E Stone. "Campus Crime." *Journal of Criminal Justice* 27, no. 4 (July 1999): 295–307. https://doi.org/10.1016/s0047-2352(99)00003-3.

[2] https://www.reddit.com/r/UTAustin/comments/1b5xgyk/ut_hasnt_been_feeling_very_safe_nor_transparent/.

[3] "Offense Definitions." FBI, September 13, 2019. https://ucr.fbi.gov/crime-in-the-u.s/2019/crime-in-the-u.s.-2019/topic-pages/offense-definitions.

[4] "The Nation's Two Crime Measures." FBI, September 13, 2019. https://ucr.fbi.gov/crime-in-the-u.s/2019/crime-in-the-u.s.-2019/topic-pages/nations-two-crime-measures.