Crime Trend Analysis in the City of Austin using ‘Crime Reports’ Data

SDS322E - John Ahn, Jueun Jeong, Esther Lee

I. Introduction

About the “Crime Reports” dataset:

The **“Crime Reports”** dataset from data.austintexas.gov is specific to the area of Austin, Texas, covering incidents under the jurisdiction of the Austin Police Department (APD). It is a collection of incidents responded to and filed written reports by APD from 2003 to the present, with weekly updates. The dataset focuses on the highest-level offense, meaning only the most severe offense is included in the dataset when multiple offenses are associated with one incident. Limitations include that varying data sources can yield different results due to methodological disparities, and replicating the research on other dates may yield distinct results as the database is updated on a weekly basis.

The “Crime Reports” dataset drew our attention because we are both residents of Austin and students who prioritize safety. This dataset provides us an opportunity to engage with real-world data, specifically the incidents that occur in our city and allows us to explore crime patterns and trends. Through this dataset, we can identify locations that require special attention due to notable crime trends while gaining overall insight into crime patterns and public safety in relation to their geographical areas in Austin.

For a little more background about the context of the dataset, the report below goes over statistics in regards to the location crimes occurred in the city of Austin in the context of the years 2014-2016.

https://www.kaggle.com/code/skirmer/exploratory-analysis-of-austin-crime-with-maps/report
(https://www.kaggle.com/code/skirmer/exploratory-analysis-of-austin-crime-with-maps/report)

A unique row of the dataset represents a single, specific incident that APD responded to and documented written reports about. Each row contains various information regarding the incident including the main variables of our interest and other data.

Our main variables of interest are **“Highest Offense Description,” “Family Violence,” “Location Type,” “Council District,”** and **“Clearance Status.”** We will create two new variables, **“Days Till Clearance”** and **“Time of Day”**, for analysis.

Highest Offense Description: Description of the offense

Family Violence: Incident involves family violence? (Y = yes, N = no)

Location Type: General description of the premise where the incident occurred

Council District: Austin city council district where the incident occurred

Clearance Status: How/whether crime was solved (N = not cleared, C = cleared by arrest, O = cleared by exception)

Days Till Clearance: The number of days it took for the case to be cleared (date cleared - dated reported)

Time of Day: The time of day in which the incident occurred (Early morning, Morning, Noon, Afternoon, Evening, Night)

What are some of the questions that our exploratory data analysis will answer?

1. Do certain types of highest offense take longer or shorter to clear than others?
2. Is there a correlation between the location type of crime and the average days it takes to clear it? Do specific location types consistently take more time to clear than others?
3. Are there specific times of day when family violence incidents occur more or less than others? What about non-family violence incidents?

What trends / relationships do we expect to observe?

1. We predict that the highest offenses that involve violence such as sexual assault would take longer to clear while offenses related to property crimes such as theft would clear quickly than others.

2. We predict that there is a correlation between location type and the mean days it takes to clear it. We expect to see locations, such as hotel/motel/etc., which are locations that are potentially associated with violence or sexual offense to involve complex investigation and, therefore, to take longer time to clear.

3. We expect to see more family violence incidents occurring at night when most family members are likely to be home. In contrast, we predict there is not a distinct pattern in the specific time of day that non-family violence incidents occur more often.

II. Methods

```
# Install packages
install.packages('tidyverse')
install.packages('readxl')
```

Let’s first load `tidyverse` and `readxl` :

```
# Load packages
library('tidyverse')
library('readxl')
```

Then import `crime_report` dataset downloaded from data.austintexas.gov and save it in the environment as `crime` . Use `read_excel` function from `readxl` to read the data.

```
# Import and save the dataset
crime <- read_excel("crime_report.xlsx")

# Take a look at the first six rows of the dataset
head(crime)
```

```
## # A tibble: 6 × 27
##   `Incident Number` `Highest Offense Description`  `Highest Offense Code`
##               <dbl> <chr>                                           <dbl>
## 1        2006471156 FAMILY DISTURBANCE                               3400
## 2       20045044338 TAMPERING WITH ID NUMBER                         2719
## 3        2006960811 FAMILY DISTURBANCE                               3400
## 4        2013851154 SEXUAL ASSAULT OF CHILD/OBJECT                   1707
## 5       20161800084 RAPE OF A CHILD                                   204
## 6        2010701921 RAPE                                              200
## # ℹ 24 more variables: `Family Violence` <chr>, `Occurred Date Time` <dttm>,
## #   `Occurred Date` <dttm>, `Occurred Time` <dbl>, `Report Date Time` <dttm>,
## #   `Report Date` <dttm>, `Report Time` <dbl>, `Location Type` <chr>,
## #   Address <chr>, `Zip Code` <dbl>, `Council District` <dbl>,
## #   `APD Sector` <chr>, `APD District` <dbl>, PRA <dbl>, `Census Tract` <dbl>,
## #   `Clearance Status` <chr>, `Clearance Date` <dttm>, `UCR Category` <chr>,
## #   `Category Description` <chr>, `X-coordinate` <dbl>, `Y-coordinate` <dbl>, …
```

The dataset `crime` is tidy because every observation has its own row and every variable has its own column.

Before we analyze the data, let’s first clean and wrangle the data.

First, let’s rename the variables given in the dataset for easier access.

```
# Rename the variables
names(crime)[names(crime) == "Highest Offense Description"] <- "highest_offense"
names(crime)[names(crime) == "Family Violence"] <- "family_violence"
names(crime)[names(crime) == "Location Type"] <- "location_type"
names(crime)[names(crime) == "Council District"] <- "council_district"
names(crime)[names(crime) == "Clearance Status"] <- "clearance_stat"
```

Then, change the format of the variables `clearance_date` and `occurrence_date` so that we can subtract the two to create a new numeric variable, `days_till_clearance` .

```
# Change the format of the variables 'clearnce_date' and 'occurence_date'
clearance_date <- as.Date(crime$`Clearance Date`, format = '%m/%d/%Y')
occurence_date <- as.Date(crime$`Occurred Date`, format = '%m/%d/%Y')

# Create a new variable 'days_till_clearnace' as a numeric variable
crime$days_till_clearance <- as.numeric(clearance_date - occurence_date)
```

Create another new variable `time_of_day` , this time categorical, by grouping the numeric variable `occurred_time` into six categories: Early Morning, Morning, Noon, Afternoon, Evening, and Night.

```
# Create a new variable 'time_of_day' using ifelse statements
crime$time_of_day <- ifelse(crime$`Occurred Time` < 400, "Early Morning",
                            ifelse(crime$`Occurred Time` < 1100, "Morning",
                                   ifelse(crime$`Occurred Time` < 1400, "Noon",
                                          ifelse(crime$`Occurred Time` < 1700, "Afternoon",
                                                 ifelse(crime$`Occurred Time` < 2000, "Evening",
                                                        ifelse(crime$`Occurred Time` <= 2400, "N
ight")
                                                 )
                                          )
                                   )
                            )
                        )
```

The original dataset included 2.42M rows and 27 columns. After data wrangling, we ended up with 100 rows and 29 columns. The 100 rows were selected at random when we imported the dataset from the database.

III. Results

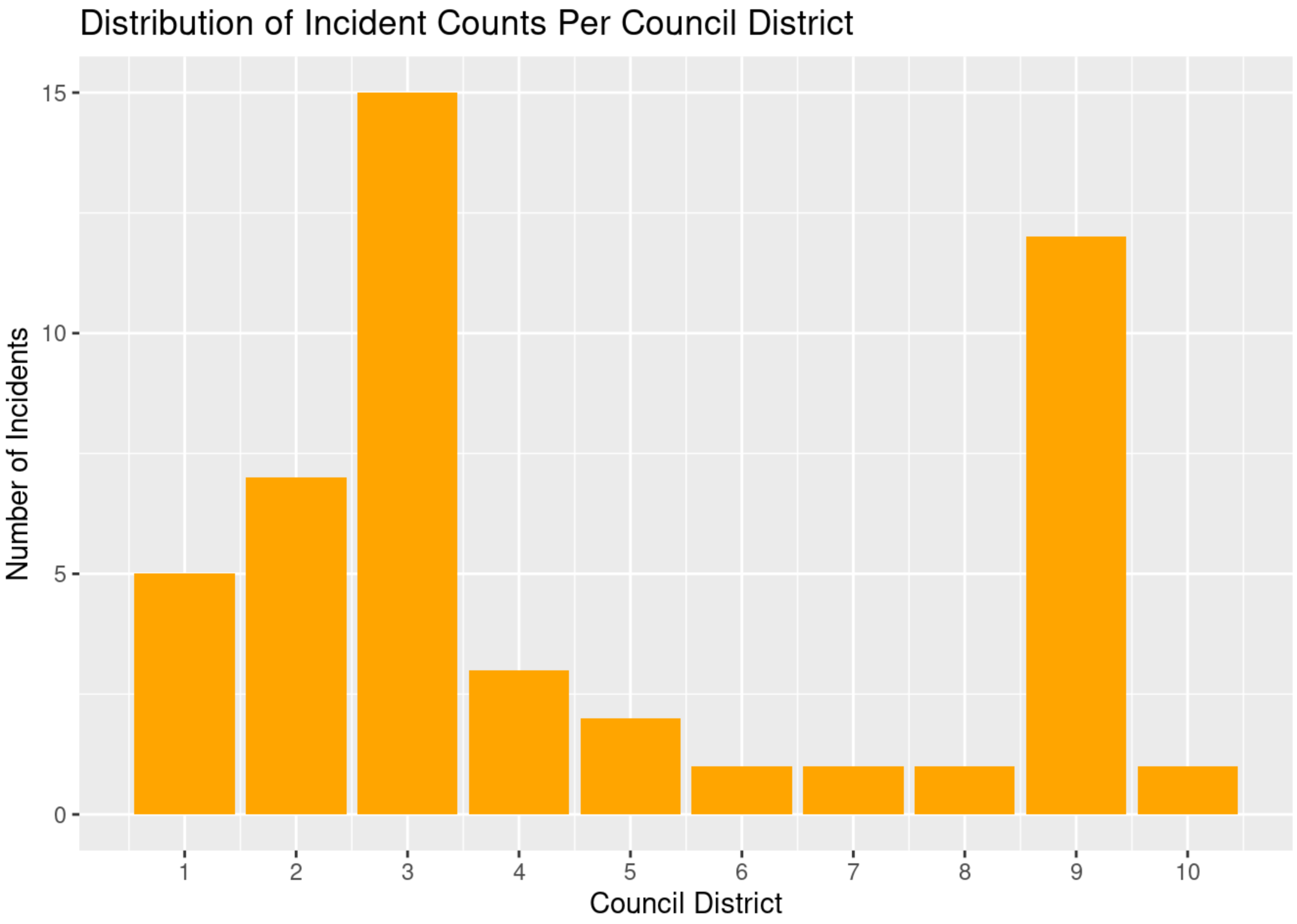
It’s time to explore and visualize the data!

a. Univariate Distributions

Let’s first take a look at three univariate distributions: `council_district` , `clearance_stat` , and `family_violence` . For each distribution, appropriate statistics will be calculated.

Univariate Visualization #1. How is incident counts per `council_district` distributed?

```
# Create a visualization for the distribution of incident counts per council district
crime |>
  ggplot(aes(x = council_district)) +
  geom_bar(na.rm = TRUE, fill = "orange") +
  # Adjust scale
  scale_x_continuous(breaks = 1:10) +
  # Add labels
  labs(title = "Distribution of Incident Counts Per Council District",
       x = "Council District",
       y = "Number of Incidents")
```



```
# Calculate appropriate summary statistics
# Find frequencies
table(crime$council_district)
```

```
##
##  1  2  3  4  5  6  7  8  9 10
##  5  7 15  3  2  1  1  1 12  1
```

```
# Find proportions
prop.table(table(crime$council_district))
```

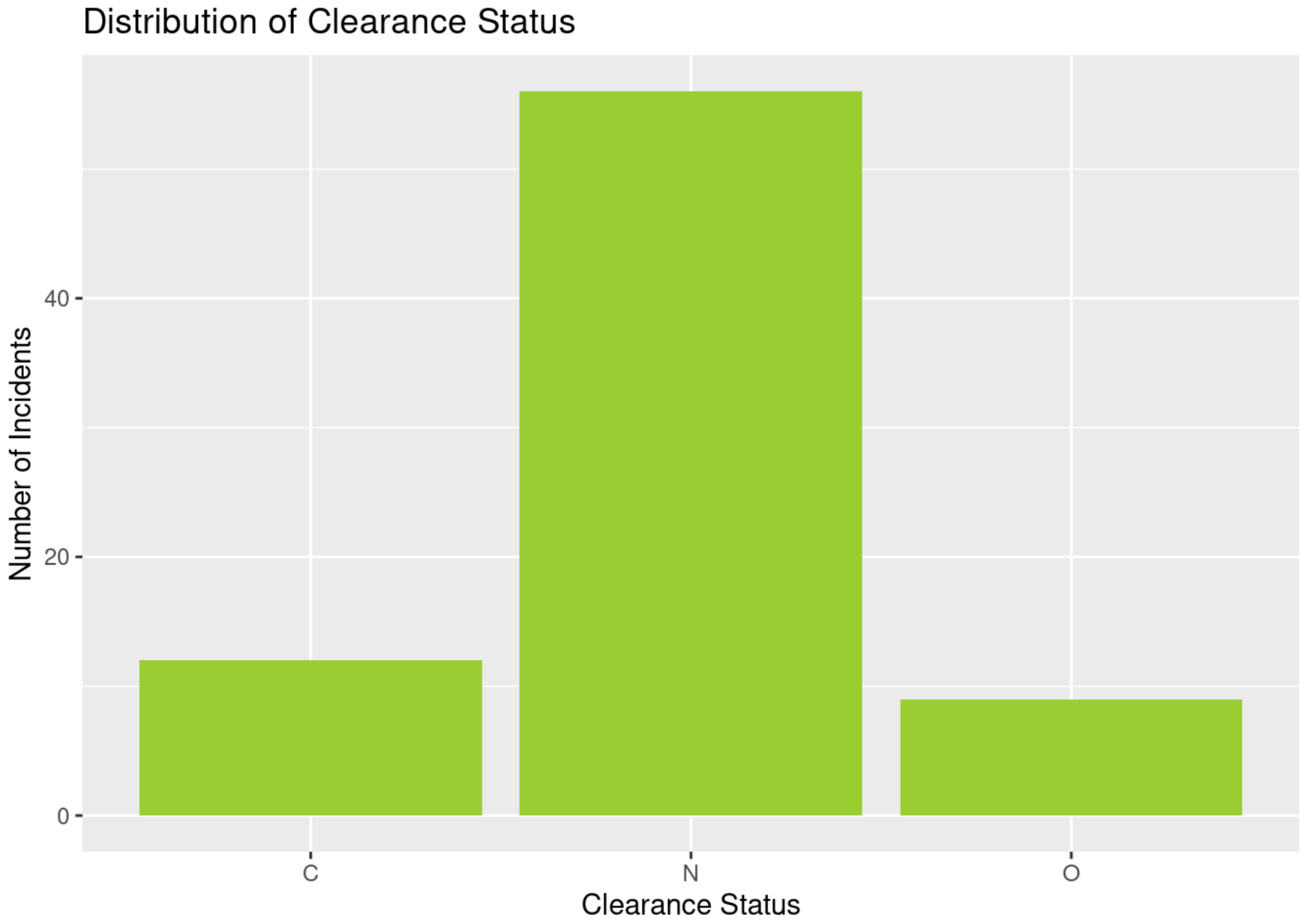
```
##
##          1          2          3          4          5          6          7
## 0.10416667 0.14583333 0.31250000 0.06250000 0.04166667 0.02083333 0.02083333
##          8          9         10
## 0.02083333 0.25000000 0.02083333
```

There are no recognizable trends or relationships in the number of reported crimes across the 10 different districts. Of all the council districts, district 3 and 9 stand out due to their relatively high number of reported incidents, which contribute to 31.25% and 25% of total incidents, respectively. On the other hand, council districts 1, 2, 4, and 5 have lower number of reported incidents, each district accounting for about 10.41%, 14.58%, 6.25%, and 4.17% of all incidents, respectively. Finally, council districts 6, 7, 8, and 10 have the least reported incidents, each contributing to about 2.08% of all incidents.

Univariate Visualization #2. How is incident counts per `clearance_stat` distributed?

```
# Filter out missing values
crime_filtered_clearance <- crime[!is.na(crime$clearance_stat),]

# Create a visualization for the distribution of clearance status
crime_filtered_clearance |>
  ggplot(aes(x = clearance_stat)) +
  geom_bar(na.rm = TRUE, fill = "yellowgreen") +
  # Add labels
  labs(title = "Distribution of Clearance Status",
       x = "Clearance Status",
       y = "Number of Incidents")
```



```
# Calculate appropriate summary statistics
# Find frequencies
table(crime_filtered_clearance$clearance_stat)
```

```
##
##  C  N  O
## 12 56  9
```

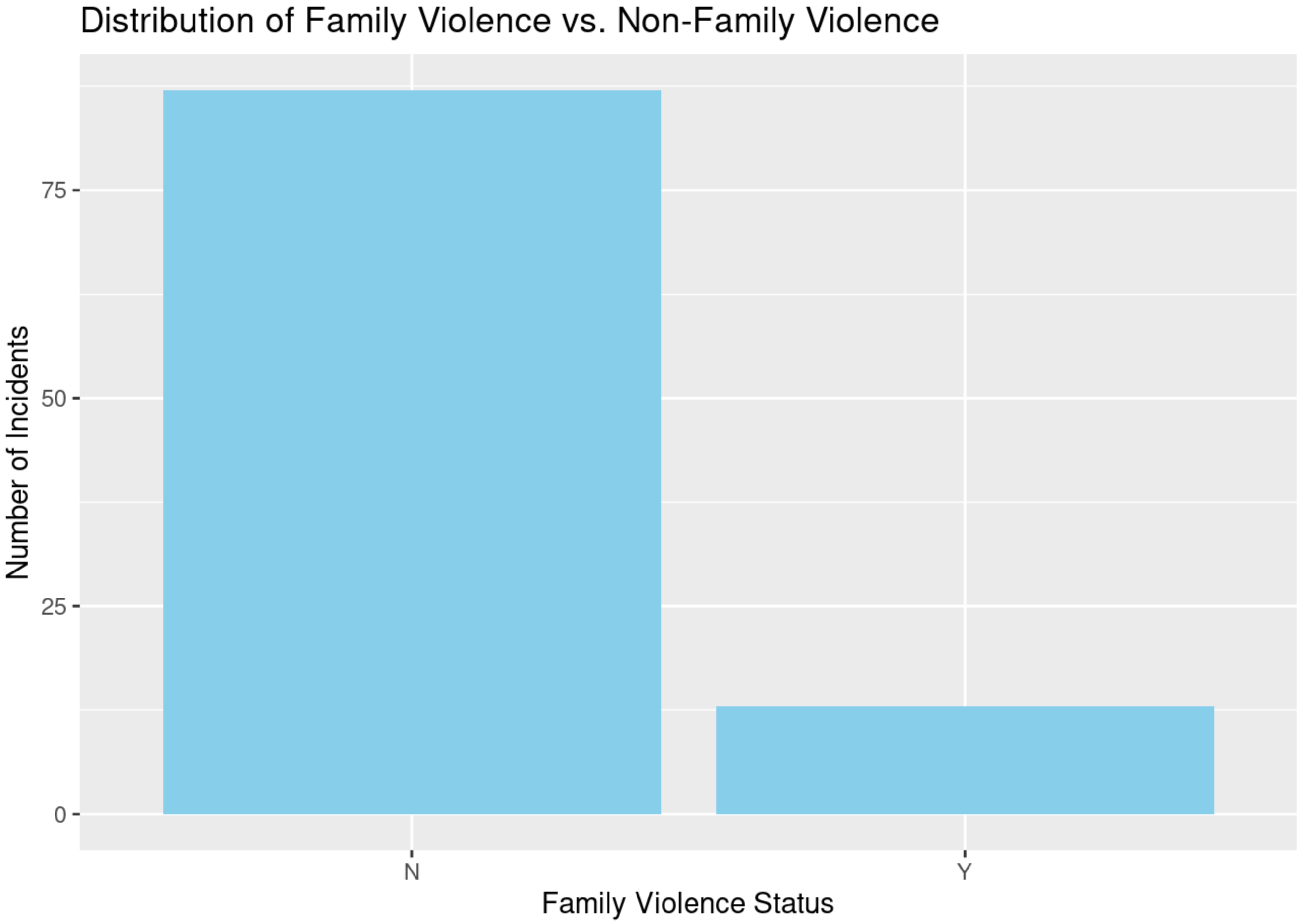
```
# Find proportions
prop.table(table(crime_filtered_clearance$clearance_stat))
```

```
## 
##         C         N         O 
## 0.1558442 0.7272727 0.1168831
```

The most prevalent clearance status in Austin is “N”, not cleared. This accounts for about 72.73% of all incidents, meaning that a significant majority of cases have not been resolved yet. We find this to be surprising. The next highest clearing status is “C”, cleared by arrest, making up approximately 15.58% of the total incidents. The remaining 11.69% of all incidents are “O”, which have been cleared by exception.

Univariate Visualization #3. How is incident counts per `family_violence` distributed?

```
# Create a visualization for the distribution of family violence vs. non-family violence
crime |>
  ggplot(aes(x = family_violence)) +
  geom_bar(na.rm = TRUE, fill = "skyblue") +
  # Add labels
  labs(title = "Distribution of Family Violence vs. Non-Family Violence",
       x = "Family Violence Status",
       y = "Number of Incidents")
```



```
# Calculate appropriate summary statistics
# Find frequencies
table(crime$family_violence)
```

```
## 
##  N  Y 
## 87 13
```

```
# Find proportions
prop.table(table(crime$family_violence))
```

```
## 
##    N    Y 
## 0.87 0.13
```

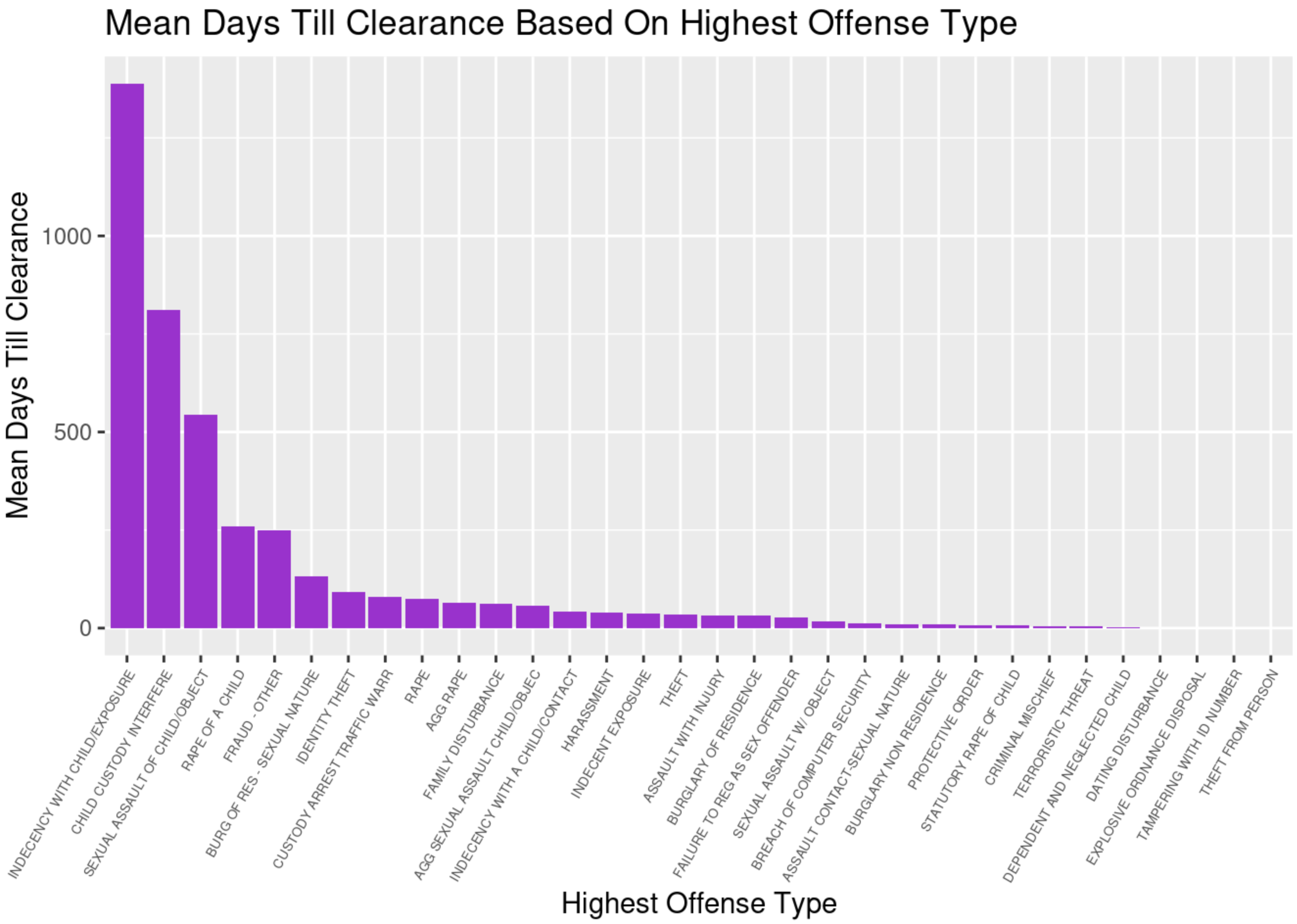
In this dataset, “Y” denotes an incident that involves family violence while “N” represents an incident that does not involve family violence. About 87% of the reported incidents have no connection to family violence while only 13% of all incidents do involve family violence. Only a minority proportion of cases are related to family violence.

b. Bivariate Distributions

Next, we are going to explore relationships between two variables! Appropriate statistics will be calculated for each distribution.

Bivariate Visualization #1. `highest_offense` vs. `days_till_clearance`:

```
# Create a visualization that represents the relationship between highest offense type and the m
ean number of days it took for clearance
crime |>
  # Filter out missing values
  filter(highest_offense != "NA") |>
  group_by(highest_offense) |>
  # Find the mean days till clearnace for each highest offense type
  summarise(mean_days = mean(coalesce(days_till_clearance,0),na.rm = TRUE)) |>
  # Add a ggplot -- bar graph
  ggplot(aes(x = reorder(highest_offense, -mean_days), y = mean_days)) +
  geom_bar(stat = 'identity', na.rm = TRUE, fill = "darkorchid") +
  # Adjust the x-axis
  theme(axis.text.x = element_text(angle = 60, hjust = 1, size = 5)) +
  # Add labels
  labs(title = "Mean Days Till Clearance Based On Highest Offense Type",
       x = "Highest Offense Type",
       y = "Mean Days Till Clearance")
```



```
# Find appropriate summary statistics
crime |>
  group_by(highest_offense) |>
  summarise(mean_days = mean(days_till_clearance, na.rm = TRUE)) |>
  arrange(desc(mean_days))
```

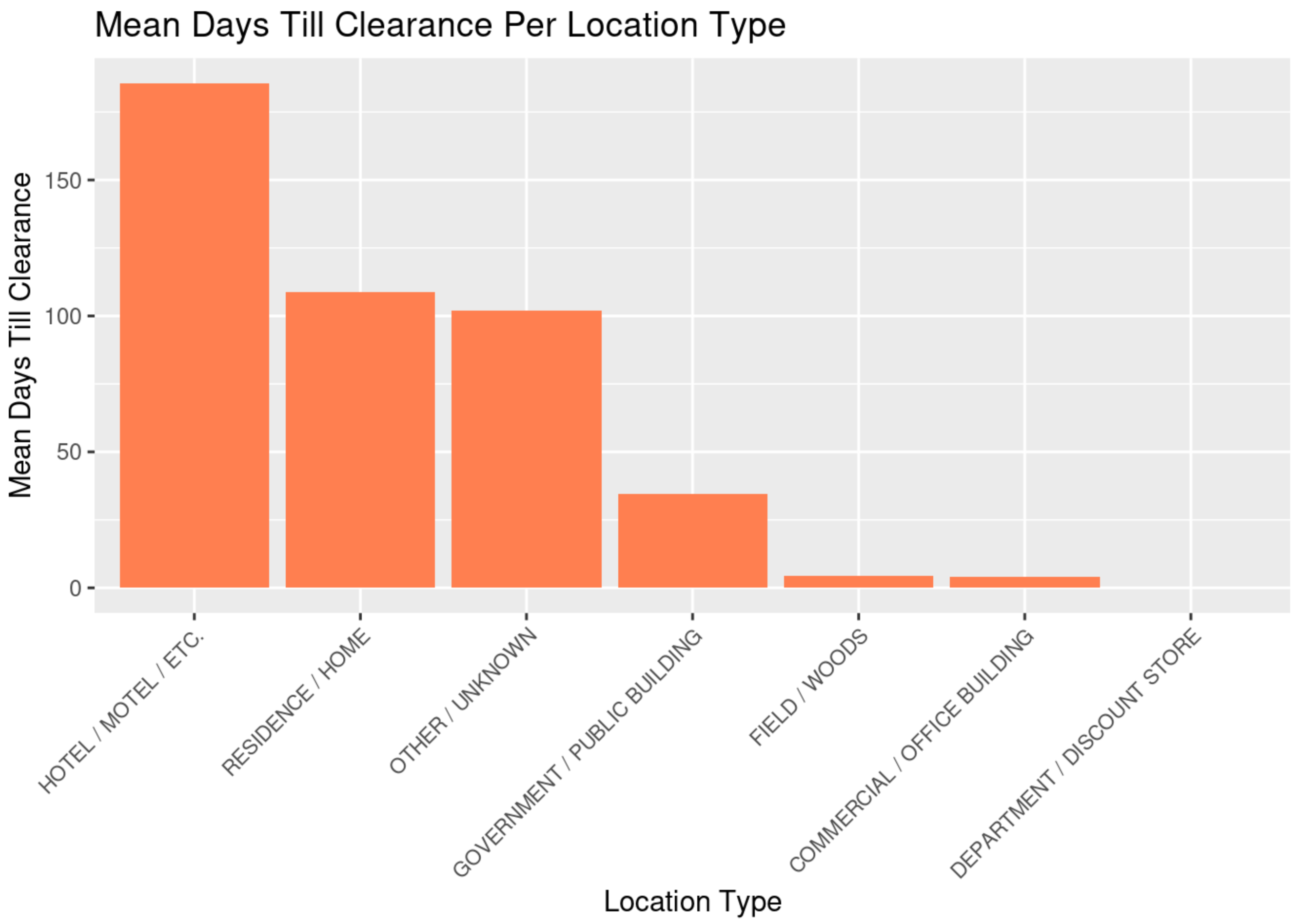
```
## # A tibble: 32 × 2
##    highest_offense                mean_days
##    <chr>                              <dbl>
##  1 INDECENCY WITH CHILD/EXPOSURE     1388
##  2 CHILD CUSTODY INTERFERE            810
##  3 SEXUAL ASSAULT OF CHILD/OBJECT     543.
##  4 FRAUD - OTHER                      501
##  5 RAPE OF A CHILD                    258.
##  6 BURG OF RES - SEXUAL NATURE        131
##  7 HARASSMENT                         119
##  8 IDENTITY THEFT                      92.7
##  9 CUSTODY ARREST TRAFFIC WARR         80.5
## 10 RAPE                                74.6
## # ℹ 22 more rows
```

Of the many, 5 highest offenses stand out in their high mean number of days till clearance. Cases with indecency with child/exposure have the greatest mean time for clearance, followed by cases involving child custody interference, sexual assault of child/object, fraud-other, and rape of a child with mean days of clearance of 1388, 810, 543, 501, and 258 days, respectively. It is notable that the majority of the top 5 highest offenses with the highest mean days till clearance involve children and/or sexual misconduct.

Bivariate Visualization #2. `location_type` vs. `days_till_clearance`:

```
# Filter out missing values
crime_filtered_location <- crime[!is.na(crime$location_type),]

# Create a visualization that represents the relationship between location type and the mean num
ber of days it took for clearance
crime_filtered_location |>
  group_by(location_type) |>
  # Find the mean days till clearance for each location type
  summarise(mean_days = mean(coalesce(days_till_clearance, 0),na.rm = TRUE))|>
  # Add a gg plot -- bar graph
  ggplot(aes(x = reorder(location_type, -mean_days), y = mean_days)) +
  geom_bar(stat = 'identity', na.rm = TRUE, fill = "coral") +
  # Adjust the x-axis
  theme(axis.text.x = element_text(angle = 45, hjust = 1,size = 8)) +
  # Add labels
  labs(title = "Mean Days Till Clearance Per Location Type",
       x = "Location Type",
       y = "Mean Days Till Clearance")
```


Mean Days Till Clearance Per Location Type

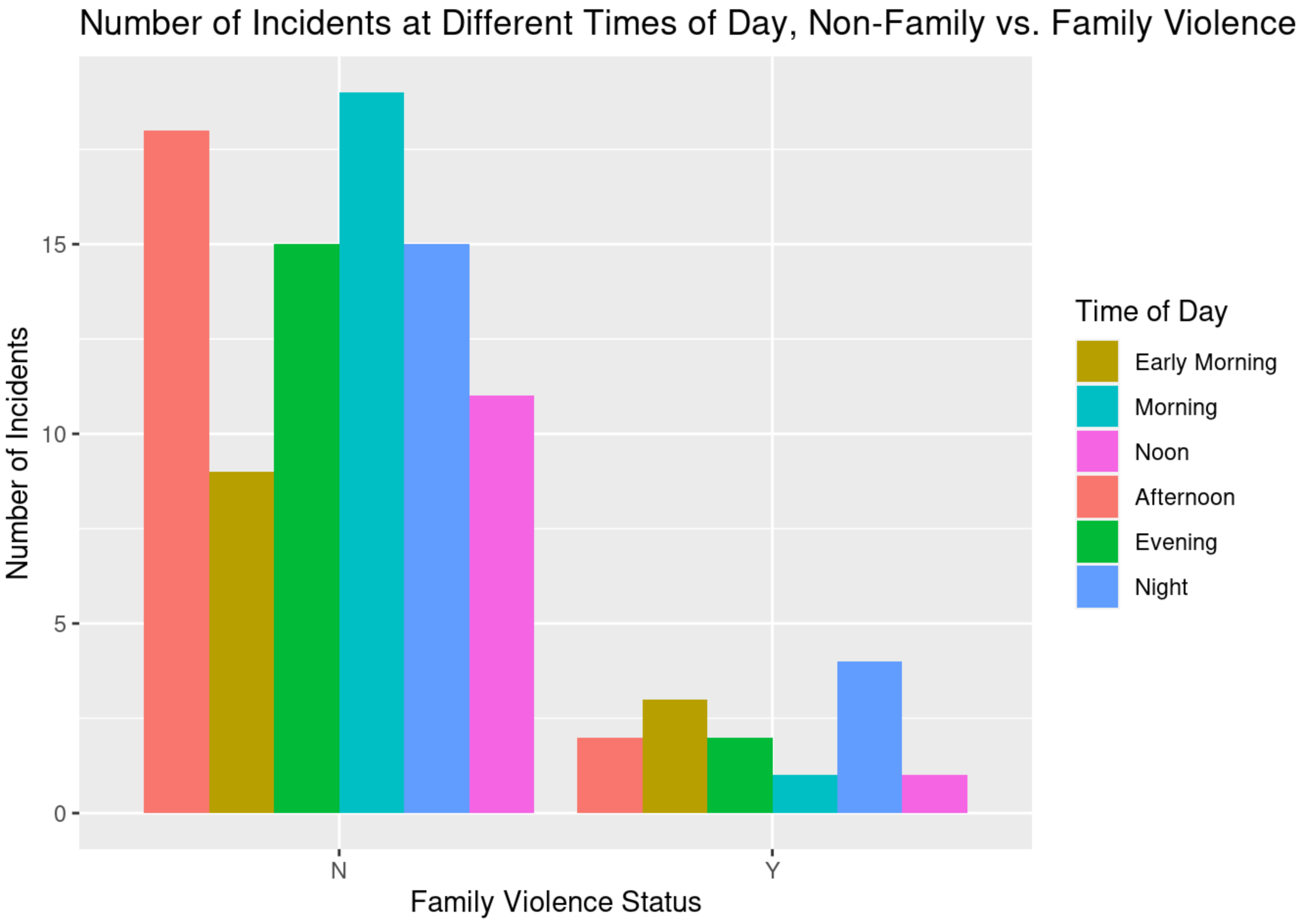
```
# Find appropriate summary statistics
crime_filtered_location |>
  group_by(location_type) |>
  summarise(mean_days = mean(days_till_clearance, na.rm = TRUE)) |>
  arrange(desc(mean_days))
```

```
## # A tibble: 7 × 2
##   location_type                mean_days
##   <chr>                            <dbl>
## 1 HOTEL / MOTEL / ETC.             186. 
## 2 RESIDENCE / HOME                 121. 
## 3 OTHER / UNKNOWN                  119. 
## 4 GOVERNMENT / PUBLIC BUILDING      34.5
## 5 FIELD / WOODS                      4.5
## 6 COMMERCIAL / OFFICE BUILDING       4  
## 7 DEPARTMENT / DISCOUNT STORE      NaN
```

Among the different location types, incidents that occurred at hotel/motel/etc. have the longest mean clearance time of 186 days, followed by residence/home, other/unknown, and government/public building having mean clearance times of 121, 119, and 34.5 days, respectively. On the other hand, incidents that occurred at field/woods and commercial/office buildings have relatively shorter mean clearance times of 4.5 and 4 days, respectively. The mean clearance time for incidents that occurred at department/discount stores is marked as “NaN” because these incidences have not been cleared yet based on the most recent update of “Crime Reports” data.

Bivariate Visualization #3. `family_violence` vs. `time_of_day`:

```
# Create a visualization that represents the relationship between family violence status and tim
e of day
crime |>
  # Add a gg plot -- bar graph
  ggplot() +
  geom_bar(aes(x = family_violence, fill = time_of_day), position = "dodge") +
  # Add labels
  labs(title = "Number of Incidents at Different Times of Day, Non-Family vs. Family Violence",
       x = "Family Violence Status",
       y = "Number of Incidents") +
  # Adjust the legend
  scale_fill_discrete(name = "Time of Day",
                      breaks=c('Early Morning', 'Morning', 'Noon', 'Afternoon','Evening','Night'
))
```



```
# Calculate appropriate summary statistics
table(crime$family_violence, crime$time_of_day)
```

```
## 
##     Afternoon Early Morning Evening Morning Night Noon
##   N        18             9      15      19    15   11
##   Y         2             3       2       1     4    1
```

```
# Find frequencies
prop.table(table(crime$family_violence, crime$time_of_day))
```

```
## 
##     Afternoon Early Morning Evening Morning Night Noon
##   N      0.18          0.09    0.15    0.19  0.15 0.11
##   Y      0.02          0.03    0.02    0.01  0.04 0.01
```

Overall, there are more proportions of incidents that are non-family violence (“N”) than there are family violence (“Y). The proportions of non-family violence incidents that occur in the early morning, morning, noon, afternoon, evening, and night are 9%, 19%, 11%, 18%, 15%, and 15%, respectively. There is the highest proportion of incidents in the morning (19%) and the lowest proportion of incidents in the early morning (9%). The proportions of family violence incidents that occur in the early morning, morning, noon, afternoon, evening, and night are 3%, 1%, 1%, 2%, 2%, and 4%, respectively. The incidents spike at night (4%) and the lowest incident of time of day is noon (1%).

IV. Discussion

Research Question 1: Do certain types of highest offenses take longer or shorter to clear than others?

The top 5 highest offenses with the highest mean days till clearance were indecency with child/exposure, child custody interference, sexual assault of child/object, fraud-other, and rape of a child. According to our **bivariate visualization #1**, the mean days of clearance were 1388, 810, 543, 501, and 258 days, respectively. Excluding the highest offense that has “NaN” mean number of days to clear, the highest offenses with the lowest mean days till clearance were tampering with ID number, dependent and neglected child, and criminal mischief, Their mean days of clearance were 0, 3, 4 days, respectively. We conclude that there are certain types of highest offenses that have a lengthier/shorter time to clear than others. Incidents involving children and/or sexual offenses take longer to clear while incidents involving identity theft, child protection, and property crime (criminal mischief) take shorter time to clear on average.

Research Question 2: Is there a correlation between the location type of crime and the average days it takes to clear it? Do specific location types consistently take more time to clear than others?

Based on our **bivariate visualization #2**, incidents taking place at hotel/motel/etc. have the highest mean clearance time of 186 days. The next highest mean clearance time is incidents taking place at residence/home, other/unknown, and government/public building having mean clearance times of 121, 119, and 34.5 days, respectively. In comparison, incidents that occur at field/woods and commercial/office buildings have shorter clearance times, averaging 45. and 4 days, respectively. It is important to note that incidents that occurred at department/discount stores are marked as “NaN”, meaning that as of the most recent update of the “Crime Reports” data, these incidences have not been cleared yet.

Research Question 3: Are there specific times of day when family violence incidents occur more or less than others? What about non-family violence?

There are specific times of day when family violence incidents occur more than others. However, for non-family violence incidents, there is not a notable time of frequent occurrence. There is more variability in the specific times of the day that incidents occur. Based on our **bivariate visualization #3**, for family violence, incidences occur more at night (between 12 a.m. to 4 a.m.) at 0.04% and low at noon (between 2 p.m. and 5 p.m.) with 0.01%. For non-family violence, incidences occurring in the morning (between 11 a.m. to 2 p.m.) have the highest proportion at 0.19% and incidences occurring in the early morning (between 4 a.m. to 11 a.m.) have the lowest proportion at 0.09%.

Was our expectations met?

We predicted that incidents involving violence such as sexual assault would, on average, take a longer time to clear while incidents involving crimes such as theft would take a shorter time to clear. The data partially aligns with our expectations. We were surprised that the top offenses with the highest mean number of days to clear involve children and/or sexual offenses. The data matched our expectation that incidents related to property crime are usually quicker to clear. It takes 34 days on average to clear theft based on our first bivariate visualization, which falls on the lower mean days to clear.

The data supports our expectation as we predicted that incidents that occur at hotel/motel/etc. would take longer to clear as incidences have the potential to involve violence and/or sexual misconduct, requiring complex investigation.

The data supports our expectation as we predicted that incidents that occur at hotel/motel/etc. would take longer to clear as incidences have the potential to involve violence and/or sexual misconduct, requiring complex investigation.

Ethical concerns – what impacts/implications could our results have on the community?

It is important to note that only the highest offense is reflected in this dataset though multiple offenses may have been involved. One must take into consideration that conducting the same study with this data may be variable as the dataset is updated on a weekly basis.

Our results on crime trends in Austin can be a potential resource and an influence in helping law enforcement agencies take additional proactive measures in protecting certain locations with high crime patterns. Further, our findings can enhance public safety and raise awareness of certain crime patterns. Public access to data can allow the Austin community to get involved in safety enhancement, crime prevention efforts, and initiation of policies. On the other hand, inaccurate interpretation and analysis of data may erode public trust in the community’s safety and in law enforcement.

Final Thoughts / Main Takeaways

We did not find any inconsistencies or typos in the dataset. However, there were a lot of missing values, especially for non-major variables (e.g. APD District, Census Tract, UCR Category). The percentage of filed cases that were not cleared was surpriginsly high (72.73%). It was also surprising and sad how it takes more than a year, on average, for incidents involving child and sexual assault to be cleared. Based on our analysis, hotels and motels seem to be a blind spot for crimes, which perhaps could be improved by installing more security cameras.

V. Reflection

One of the key challenges encountered in data management involved the meticulous coordination of variables to optimize the relevance of statistical comparisons. This task was particularly intricate due to the bivariate observations made during dataset sorting decisions. It is worth noting that this complexity was heightened by the presence of a sole numeric value within the dataset, derived from the interplay of existing variables, namely the clearance date and occurrence date, resulting in the creation of a novel variable known as “days to clearance.”

However, despite the challenges, it was clear that the ability to manipulate the data allowed for more focused analysis of what statistics of the dataset are representative of and provided a more in-depth observation that was more difficult with the existing variables of the dataset before data wrangling was employed.

VI. Acknowledgements

https://ggplot2.tidyverse.org/reference/ (https://ggplot2.tidyverse.org/reference/)a

https://dplyr.tidyverse.org/reference/index.html#data-frames (https://dplyr.tidyverse.org/reference/index.html#data-frames)

The provided links were instrumental in optimizing the utilization of ggplot2 and dplyr functions for efficiently organizing and visualizing the data during the initial variable assessments.

Thank you to Dr. Guyot who clarified what appropriate statistics we need to find for bivariate distributions.

Contributions:

Code – John Ahn; Statistical analysis – Jueun Jeong; R markdown – Esther Lee

VII. References

Background Context:

Skirmer. (2017, October 1). Exploratory analysis of Austin crime with maps. Kaggle. https://www.kaggle.com/code/skirmer/exploratory-analysis-of-austin-crime-with-maps/report (https://www.kaggle.com/code/skirmer/exploratory-analysis-of-austin-crime-with-maps/report)

Dataset Link:

https://data.austintexas.gov/Public-Safety/Crime-Reports/fdj4-gpfu (https://data.austintexas.gov/Public-Safety/Crime-Reports/fdj4-gpfu)