Mukhamejan Assan

Introduction to R

Dr. Marc Kaufmann

April 5, 2023

Report on final project

The Chicago Crime Rate dataset reflects reported incidents of crime that have occurred in the city over the past year, except for the most recent seven days of data. The dataset was extracted from the Chicago Police Department's CLEAR system. The purpose of this EDA was to explore Chicago crime rates and answer the following questions: what are the most frequent crimes? Is there a variation in crime rates among months? Is there a variation in crime rates among days of the week? And, where do crimes usually occur?

The EDA revealed that theft and burglary were the most frequent crimes in Chicago, followed by battery, weapons violation, and narcotics-related crimes, which resulted in arrests the most. There were spikes at midday and midnight. Midday spike could be due to people commuting for lunch and being more exposed to thieves, while thefts peaked on Fridays, perhaps because there are more people moving around to new places like bars, clubs, parks, etc. Battery occurrence peaked at midnight, which is surprising. Sunday seems to be a favorite day for those committing Battery crimes. The trend suggests that there are more crimes when the weather gets warmer, as there is generally more movement in the population when it's warmer.

Raw data description

This dataset includes reported incidents of crime that have occurred in the City of Chicago in the period between April 2021 and April 2022. The data was extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system, and the addresses are only shown at the block level to protect the privacy of crime victims. The dataset contains 209,453 rows and 17 columns. The columns include Case#, Date of Occurrence, Block, IUCR (Illinois Uniform Crime Reporting) code, primary and secondary description of the crime, location description, whether an arrest was made or not, whether the crime was domestic or not, beat number, ward number, FBI code, X and Y coordinates, latitude and longitude, and location.

Data cleaning

The first step in data cleaning was to check for missing values in the dataset. I used the is.na() function to check for missing values in each column of the dataset. This function returns a logical matrix indicating whether each element of the dataset is missing or not. I then used the colSums() function to calculate the total number of missing values in each column of the dataset.

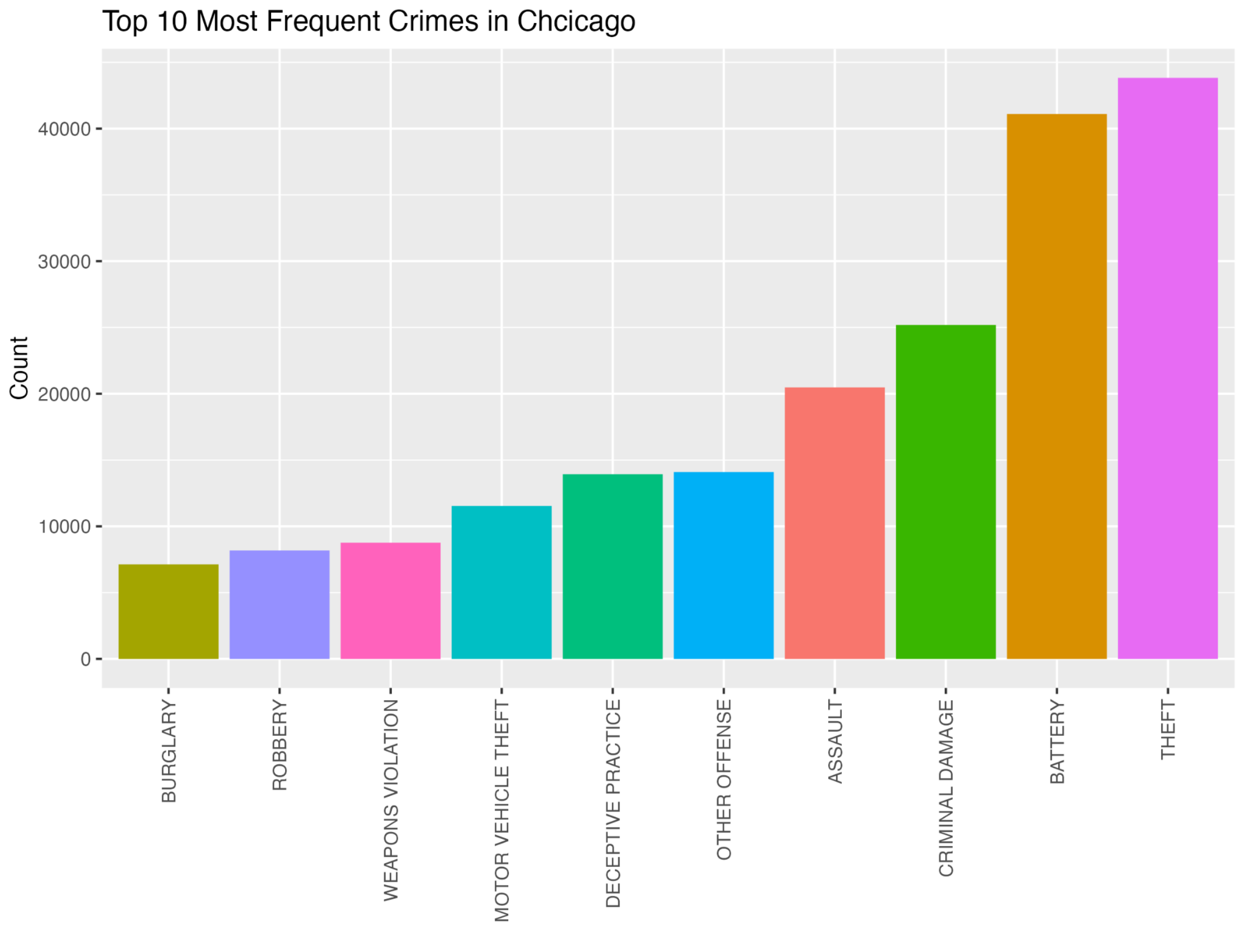
Next, I dropped rows with any missing values using the na.omit() function. This ensures that the dataset contains only complete cases with no missing values. I also checked for duplicate rows in the dataset using the duplicated() function. This function returns a logical vector indicating whether each row of the dataset is a duplicate or not. I stored the duplicate rows in a variable called dup_rows and displayed the first 10 duplicate rows using the head() function.

To remove the duplicate rows, I used the unique() function. This function removes duplicate rows from the dataset and returns a dataset with only unique rows.

Data exploration

I explored the Chicago crime rates dataset and attempted to answer several questions I mentioned before. Firstly, I investigated the top 10 most frequent crimes in Chicago using the following code:

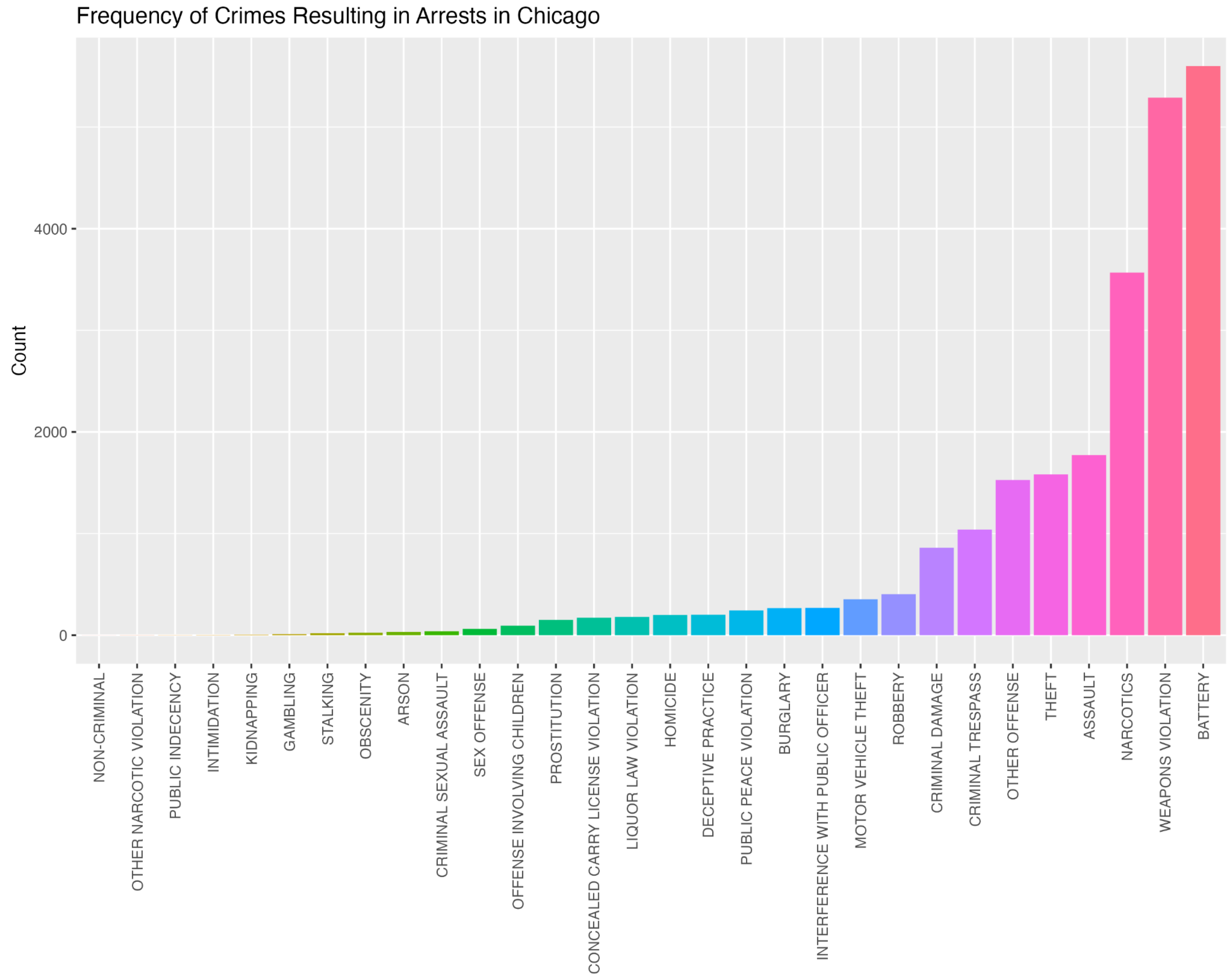
```
df |>
  count(`PRIMARY DESCRIPTION`, sort = TRUE) |>
  head(n = 10) |>
  ggplot(aes(x = fct_reorder(`PRIMARY DESCRIPTION`, n), y = n,
fill = `PRIMARY DESCRIPTION`)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5,
hjust = 1)) +
  labs(y = "Count", x = NULL) +
  ggtitle("Top 10 Most Frequent Crimes in Chicago") +
  guides(fill = "none")
```



The findings showed that theft, criminal battery, criminal damage, and assault are the most common crimes in Chicago.

Next, I examined the frequency of crimes resulting in arrest with the following code:

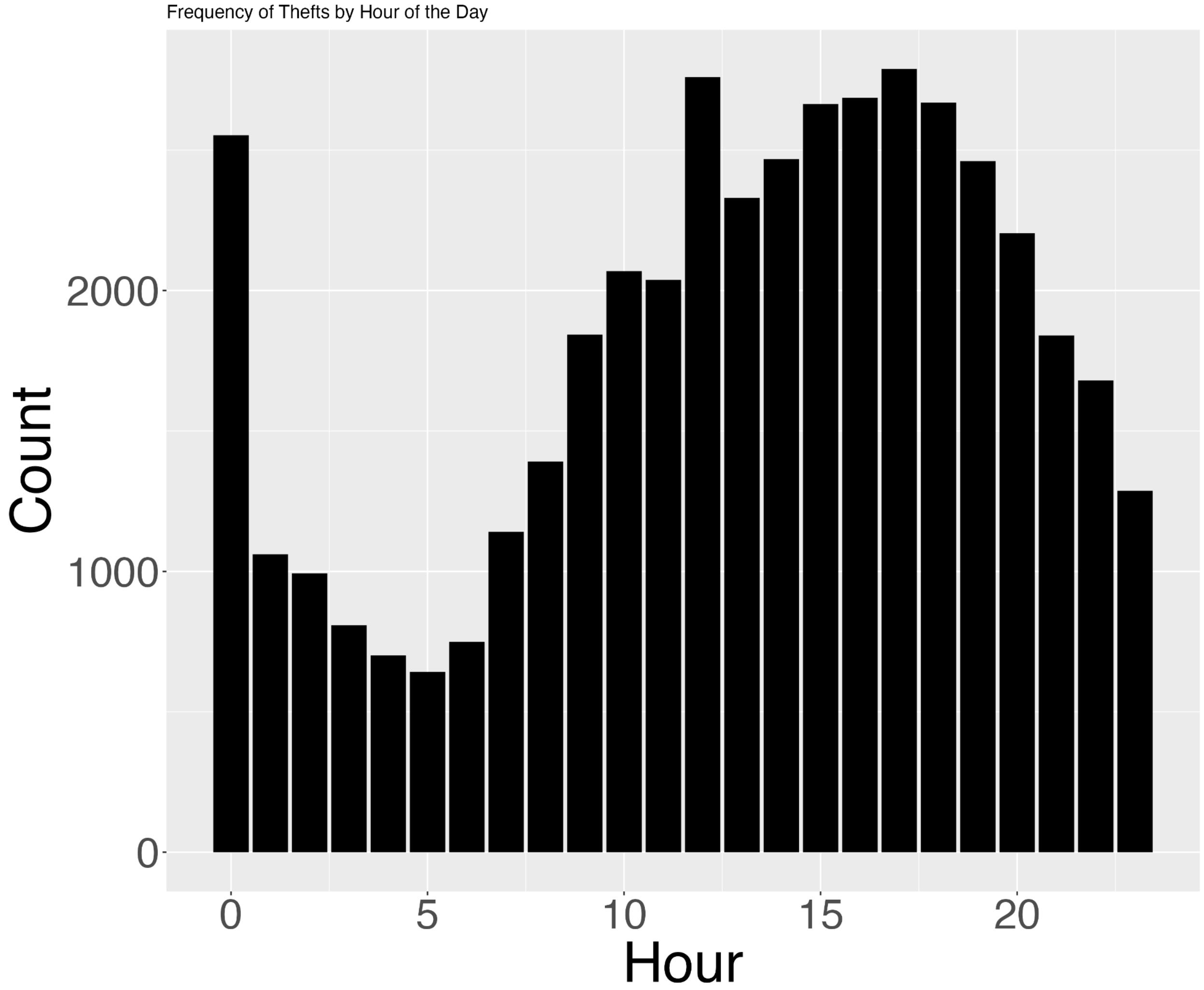
```
# Filter rows where the "ARREST" column has a value of "Y"
df |> filter(ARREST == "Y") |>
  # Count the number of occurrences of each unique value of
"PRIMARY DESCRIPTION"
  count(`PRIMARY DESCRIPTION`) |>
  # Sort the resulting counts in descending order
  arrange(desc(n)) |>
  # Reorder the "PRIMARY DESCRIPTION" column according to the
sorted counts
  mutate(`PRIMARY DESCRIPTION` = fct_reorder(`PRIMARY
DESCRIPTION`, n)) |>
  # Create a bar plot using "PRIMARY DESCRIPTION" as the x-axis
variable, and the count as the y-axis variable
  ggplot(aes(x = `PRIMARY DESCRIPTION`, y = n, fill = `PRIMARY
DESCRIPTION`)) +
  # Create a bar chart with each bar representing the count of a
unique "PRIMARY DESCRIPTION" value
  geom_bar(stat = "identity") +
  # Rotate the x-axis labels by 90 degrees for better
readability
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5,
hjust = 1)) +
  # Set the y-axis label to "Count"
  labs(y = "Count", x = NULL) +
  # Set the plot title to "Frequency of Crimes Resulting in
Arrests in Chicago"
  ggtitle("Frequency of Crimes Resulting in Arrests in Chicago")
+
  # Remove the legend
  guides(fill = "none")
```



The findings indicated that battery, weapons violation, and narcotics related crimes result in arrests the most.

Then, I analyzed the hourly frequency of thefts with the following code:

```
ggplot(df %>% filter(`PRIMARY DESCRIPTION` == "THEFT"), aes(x =
hour)) +
  geom_bar(fill = "black") +
  labs(x = "Hour", y = "Count") +
  ggtitle("Frequency of Thefts by Hour of the Day") +
  theme(axis.text.x = element_text(size = 30), axis.text.y =
element_text(size = 30), axis.title = element_text(size = 40))
```



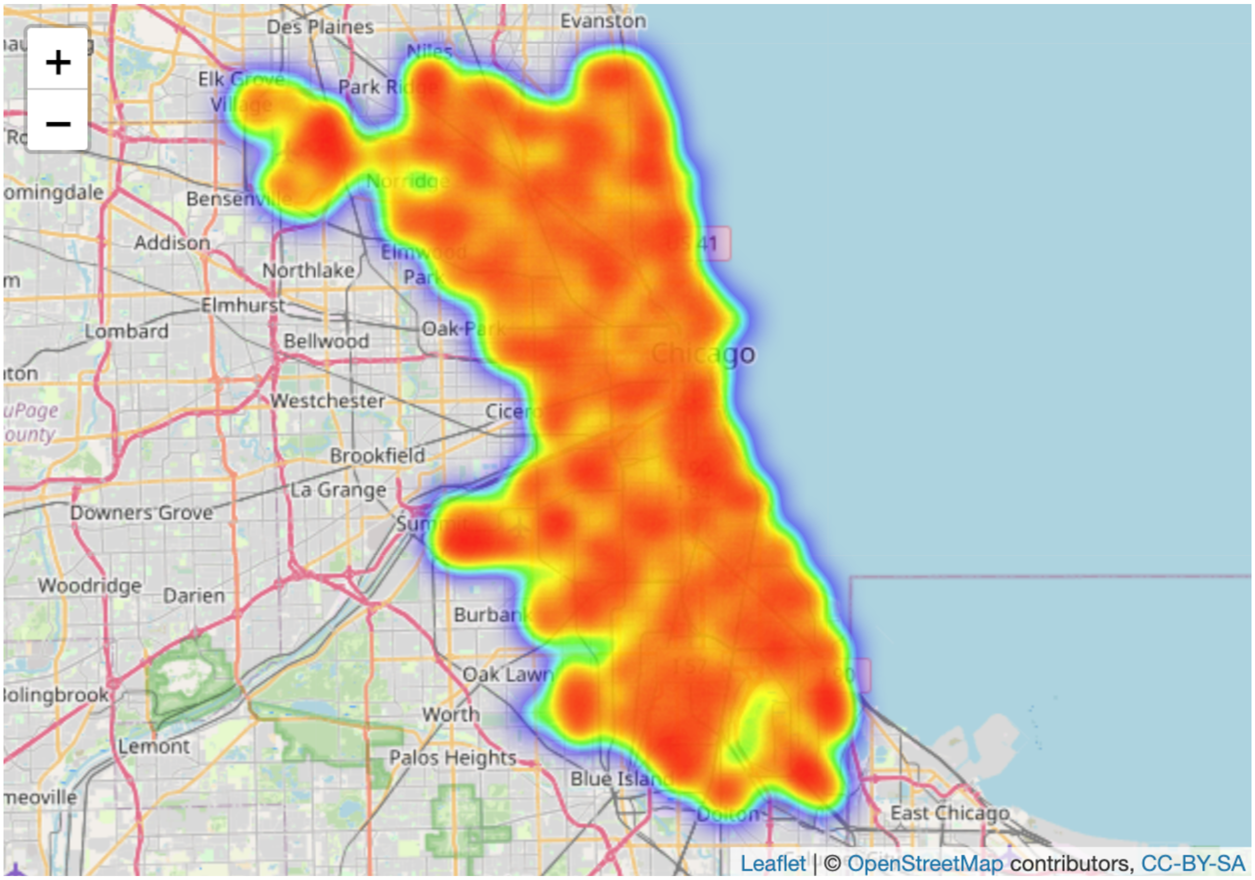
The findings revealed an unusual spike in thefts at midnight and then again at 1 am, followed by a decrease to around 600 at 5 am. The frequency of thefts then rose to almost 2000 by 10 am and another spike was observed at noon. The trend continued, peaking at around 2700 at 5 pm. The peaks may be attributed to the absence or irrelevance of the exact time of crime occurrence. Overall, the trend coincided with people's daily rhythms.

Finally, I wanted to see if I can find spatial patterns in crime occurrence with the following code:

```
# Create leaflet map
crime_map <- leaflet(df_filtered) %>%
  setView(lng = -87.7098, lat = 41.8836, zoom = 11)

# Add heatmap
crime_map <- crime_map %>% addHeatmap(
  data = df_filtered,
  lng = ~ LONGITUDE,
  lat = ~ LATITUDE,
  radius = 10
)

# Display map
crime_map
```



Based on the map plot, it does not appear to show any instantly available insights. However, one possible hypothesis is that the crime activity might coincide with the density of population in that particular area.

Reference

Chicago Crime Rate dataset. City of Chicago Data Portal. Retrieved April 4, 2023, from https://data.cityofchicago.org/Public-Safety/Chicago-Crime-Rate/q6ys-f7ek.