Crime in Dallas

STAT 3355.001

Group 10

Ariel Yong, Ben Wowo, Kyle Evans

April 19th, 2023

Contents

1 Introduction 2

2 Data Cleaning 3

3 Questions and Findings 4

3.1 Does Crime Vary by Month?

3.2 Are Some Locations More Prone to Crime?

3.3 Is There any Relation Between Age of Arrestee and the Arrest Weapon?

3.4 Of Total Crimes, What Share Does Each Type of Drug Comprise?

3.5 How Does the Crime Rate Vary Across Different Days of the Week and Times of Day in Dallas? Are There Any Notable Patterns or Trends?

4 Deadend Graphs 12

4.1 How Has Crime Rate in Dallas Changed Over Time and What Factors May Have Contributed to These Changes?

4.2 Are There Any Unique Factors that Contribute to Crime in Dallas, Such as the City's Demographics or Geography?

4.3 What is the Overall Crime Rate Trend in Dallas and Los Angeles? Which Areas in the Cities Have the Highest and Lowest Crime Rates, and Are There any Common Factors Among Those Areas?

4.4 What is the Distribution of Crimes Across Different Zipcodes in Dallas. Are There Any Notable Findings or Trends?

5 Conclusion 15

6 Code 17

7 Reference 23

1 Introduction

Crime is a pervasive social issue that has far-reaching effects on individuals, communities, and society as a whole. Understanding the nature and patterns of crime incidents is crucial for law enforcement agencies and policymakers to develop effective strategies for crime prevention and intervention, ultimately improving the standard of living in a given area.

In this report, we present an analysis of two datasets detailing crime incidents in Dallas, Texas, as our main focus, and Los Angeles, California. Los Angeles is selected as a comparison due to sharing various qualities. Both cities are economic hubs, have experience with urban sprawl, and possess diverse demographics in terms of ethnic background and income level.

The first dataset examines the number of arrests made and related information in Dallas, Texas, spanning from 2014 to 2022–61,656 observations and attributes, such as incident number, year, date, time, address, zip code, latitude and longitude, city, state, day of the week, location, use of weapon, age and sex of the arrestee, and whether the incident is drug-related and the type of drug involved. The second dataset surveys crime reported in Los Angeles, California, dating from 2010 to 2017, including variables such as DR number, date reported, date occurred, time occurred, area ID, area name, crime code, crime description, and MO code.

Through this analysis, we aim to identify patterns of crime in Dallas and conclude possible trends in comparison to Los Angeles. The report attempts to determine what factors contribute to the frequency of crime, drawing potential relationships with time, day, month, and arrestee-specific information. Fundamentally, we will look at the distribution of incidents as well as the overall crime rate.

2 Data Cleaning

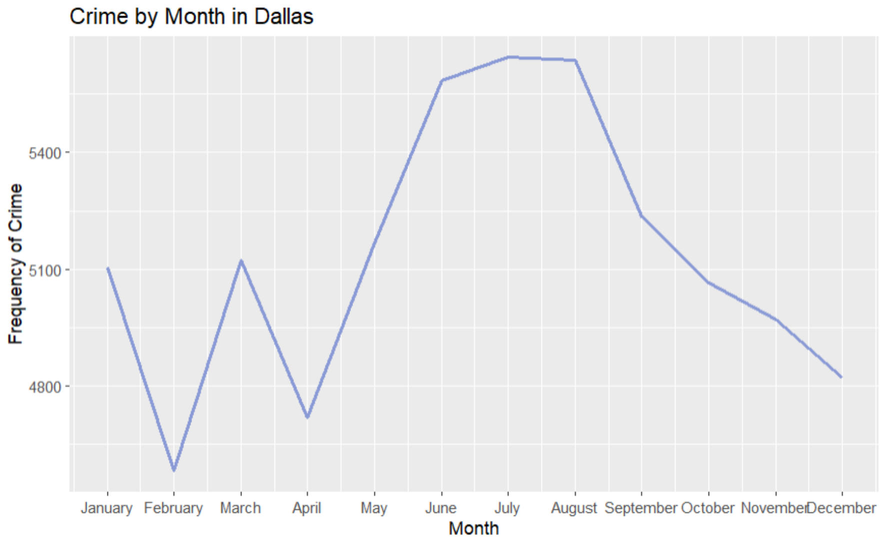
In both datasets, data cleaning was necessary to ensure that the data is suitable for analysis. Our main dataset, Dallas, the state variable was removed as all observations were from Texas. The attribute appeared to be redundant, providing little to no information. Additionally, the variables Arrest.Date and Arrest.Time were converted to the DateTime data type using the as.POSIXct function. Thus, enabling a more accessible format for date manipulation, such as computing time differences between arrests or aggregating arrests by day, week, or month. Similarly, the variables Arrest.Day.of.The.Week, Arrestee.Race, Arrestee.Sex, Drug.Related, and Drug.Type were converted to factor data types. We can then compute the frequency of arrests by race or sex, or examine the relationship between drug-related arrests and the type of drug involved.

Our secondary dataset, Los Angeles, the attributes Date.Reported and Time.Occurred were converted to the DateTime data type using the as.Date and as.POSIXct functions, respectively. However, Time.Occurred must first be converted to a character data type due to the variable being stored as a numeric value, incompatible .POSIXct function. Furthermore, several variables were removed from the Los Angeles dataset as deemed necessary. The attributes DR.Number, Mo.Code, Area.ID, and Weapon.Used.Code were removed. For example, Area.ID was equivalent to Area.Name but in numeric form, and thus, did not provide any additional information for our analysis–we assumed similarly for Weapon.Used.Code.

3 Questions and Findings

3.1 Does Crime Vary by Month?

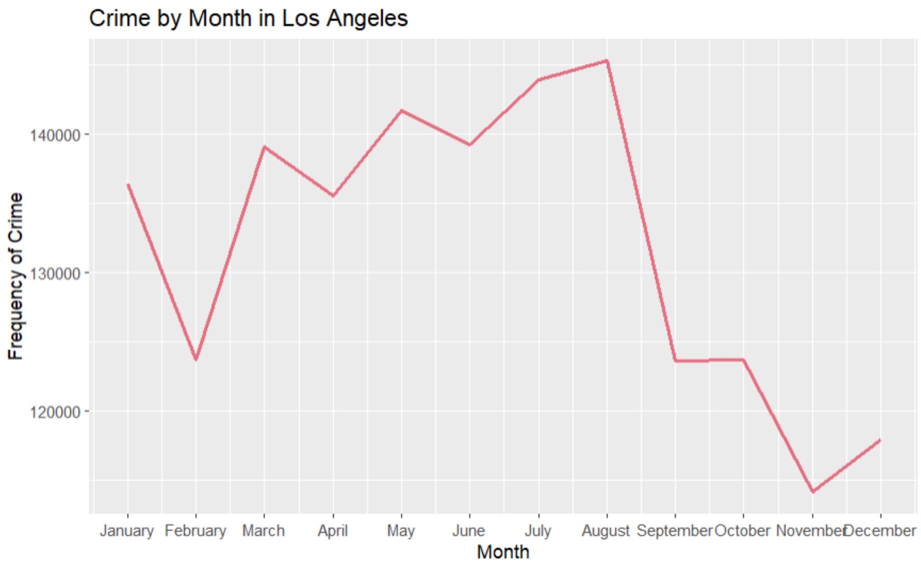
We created a line trend graph using the ggplot2 library to display the total amount of crime from the years 2014 to 2022.varied by the month. During the months of January to April, the rate tends to fluctuate, a significant increase then follows in May and June. Though at an all time high, the overall crime becomes relatively stable during the summer. The trend falls steeply in August, reaching September, and finally, the remaining months share a gradual decrease. February seems to possess the lowest rate in crime whereas July has the greatest. The trend in January, March, and October appear to be comparatively average.



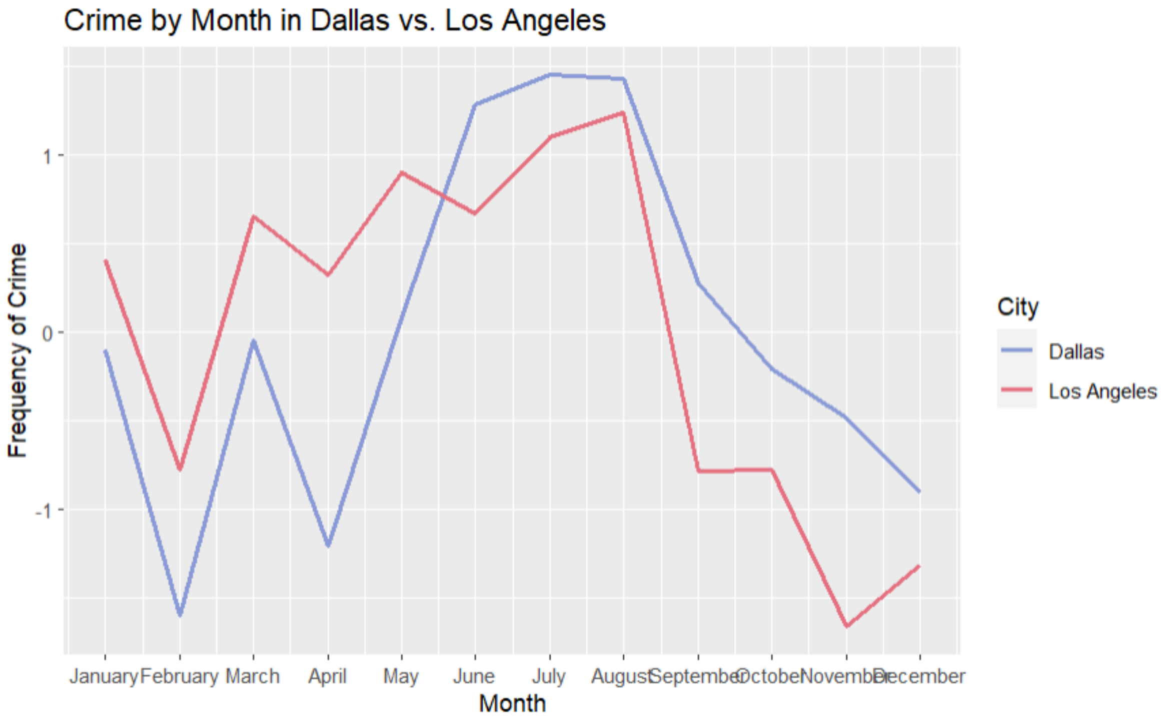
An assumption can be concluded to account for such high crime during June, July, and August. Tourism in Dallas most likely rises during the summer season, causing an increase in the current population. We can also assume, there is a direct correlation between population growth and crime rate. As for the overall decrease in incidents in the months of September, October, November, and December–colder temperatures could deter any incentive to commit an offense. Lastly, the beginning months tend to fluctuate due to extreme weather. January and February typically experience the lowest temperatures, becoming prone to winter storms. March and April are known for rainy weather thus, thunderstorms and tornado season.

How Does Crime Vary by Month in Los Angeles? How Does it Compare to Dallas?

We then looked how crime varied by month in Los Angeles, California–using a similar process to the Dallas graph, we created a line trend plot. Similarly, the rate of incidents fluctuates in the beginning months, following the greatest amount of crime in August. After, a large drop is seen in September with no significant change in October. The remaining months boast the lowest amount of crime compared to the rest of the year.

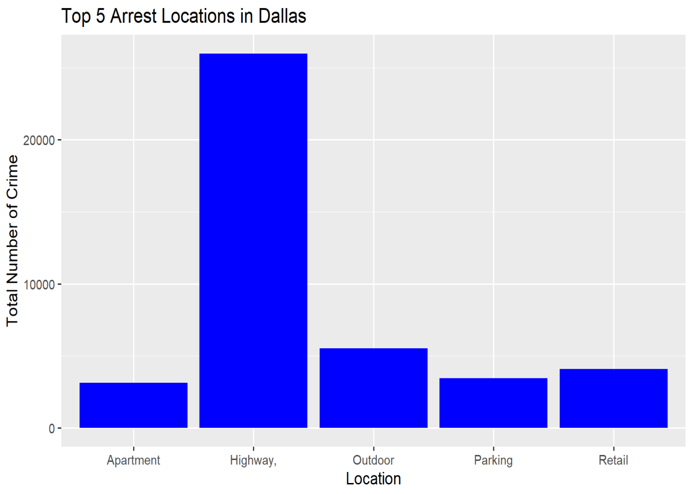


We plotted the two line trends in the same graph to obtain a closer comparison. Dallas and Los Angeles, appear to possess alike trends. Using the original values comparison would be difficult as the variation of crime frequency between the two cities was too large thus, we normalize the data points using log transformation. We can conclude that Dallas and Los Angeles follow similar trends.



3.2 Are Some Locations More Prone to Crime?

Locations in Dallas



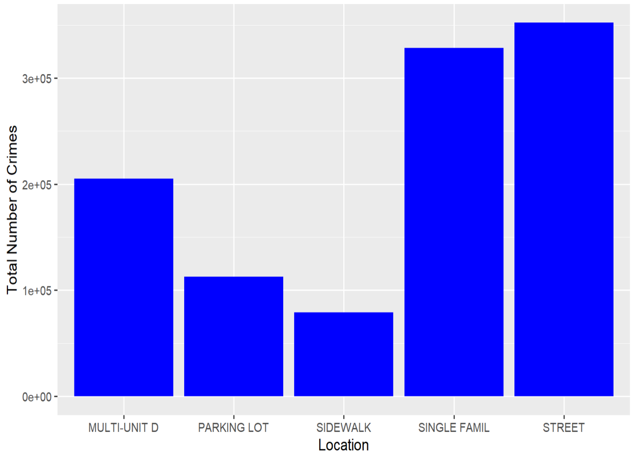
We wanted to identify if there are any locations in Dallas that are more prone to crime than others. We first counted the number of observations for each location, obtained the top five locations, shortened the location names to a single word, and created a barplot using the ggplot2 library.

We can conclude the reason for certain areas possessing higher amounts of crime because such areas have an overall greater population density. We can also assume that some locations may have higher concentrations of poverty or unemployment, often encouraging residents to resort to crime due to desperation or a lack of resources.. Moreover, some areas may have less police presence than others, causing increased vulnerability to crime.

As for why highways might be the place with the most occurrences of crime, one reason could be that highways are often used as a means of transportation for criminals who are looking to escape quickly after committing a crime.. Highways also tend to be less populated than other areas, which can make them more attractive targets for criminals who are looking to avoid detection.

3.2 Locations in Los Angeles

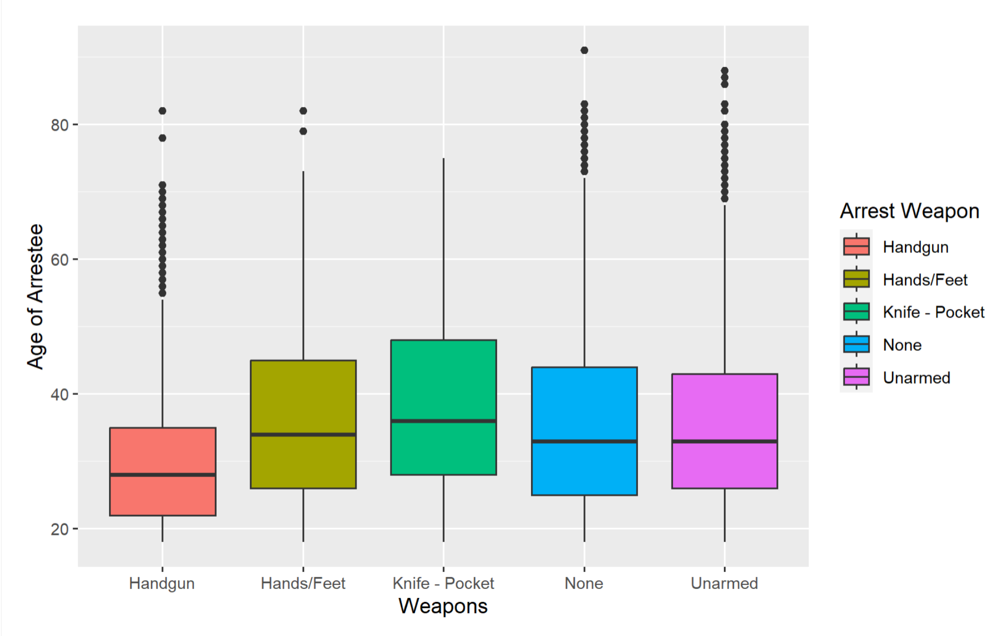
This summarizes the data to get the total number of observations for each location. This is done using the summarize() function from the dplyr package in R. Summarizing data is important because it provides a clear idea about the dataset and helps to identify patterns and trends in the data1. In this case, summarizing the data allows us to see which locations have more crime than others. The code then arranges the summarized data in descending order of count and selects only the top 5 locations with the most crime observations. The locations are then abbreviated to make them one word using str_to_title() and str_replace_all() functions from the stringr package2. Finally, a bar plot is created using the ggplot() function from the ggplot2 package to visualize the total number of crimes for each location.



In Los Angeles, streets and single-family dwellings are the places with the most occurrences of crime because they are more accessible and easier targets for criminals. Criminals can easily blend in with crowds on streets and can easily escape after committing a crime. Single-family dwellings are also easier targets because they are often located in isolated areas and have fewer people around them.

3.3 Is There any Relation Between Age of Arrestee and the Arrest Weapon?

There are many reasons why some age groups in Dallas might have more crime than others. One reason is that young people commit a disproportionate amount of street crime, in part because of the influence of their peers and their lack of stakes in conformity1. Another reason is that neighborhoods with high residential turnover might have more crime than neighborhoods with a stable residential community.



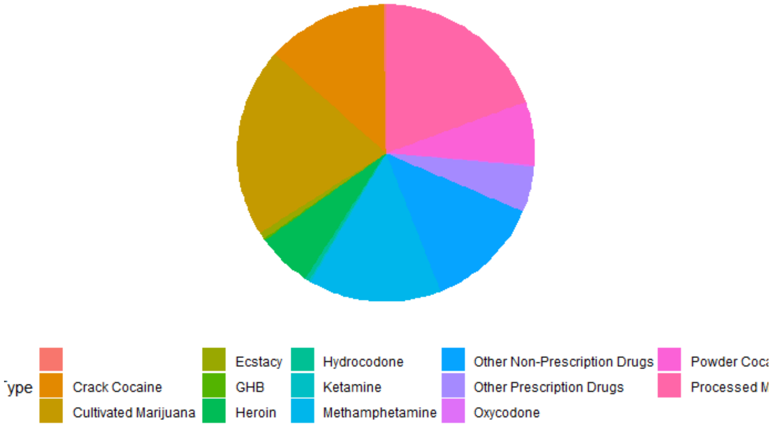
It’s important to note that crime rates vary by location and time, and that there are many factors that can contribute to crime rates such as poverty, unemployment, drug use, and social disorganization

The chart that we thought was most useful was a box plot.A boxplot is a type of plot that displays the five number summary of a dataset, which includes: The minimum value, the first quartile (the 25th percentile), the median value, the third quartile (the 75th percentile), and the maximum value1. Boxplots are useful when comparing distributions across groups. In this case, we are comparing age and arrest weapon. The boxplot shows us how age varies across different arrest weapons. The boxplot also shows us if there are any outliers in our data.

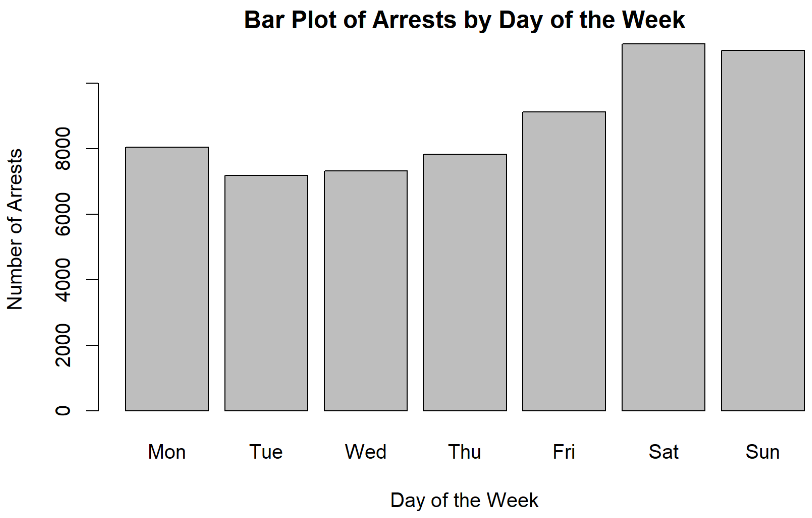
The code reads in the data from a file called “crime.csv” and filters it to only include the top 5 arrest weapons by number of observations. It then creates a boxplot using ggplot2 library with Arrest.Weapon on x-axis, Arrestee.Age.At.Arrest.Time on y-axis and Arrest.Weapon as fill.

3.4 Of Total Crimes, What Share Does Each Type of Drug Comprise?

We created a pie chart to display the makeup of drug-related crimes and the kind of drug used. The light pink depicts non drug-related incidents, comprising the largest percentage of the overall crime. We wanted to determine the distribution of drug use. A conclusion can be drawn claiming the amount of drug type is relatively similar.



3.5 How does the crime rate vary across different days of the week and times of day in both cities? Are there any notable patterns or trends?



We chose a bar plot to visualize our data. Bar plots are commonly used to display and compare the counts or frequencies of categorical data. We compared the days of the week with its corresponding frequency of crime

Other types of graphs, such as line graphs or scatter plots, are better suited for displaying continuous data or for showing the relationship between two continuous variables. Since the data in this case is categorical, days of the week, and not continuous, a bar plot is a more appropriate choice.

There could be many reasons why there is more crime on the weekends compared to weekdays. One possible explanation is that people have more free time on the weekends and may engage in activities that increase their likelihood of committing or being a victim of a crime. For example, people may go out more on the weekends and consume alcohol, which can increase the likelihood of criminal behavior. Additionally, many businesses are closed on the weekends, which could make them more vulnerable to crimes such as burglary. However, without further information and analysis, it is difficult to determine the exact reasons for the observed pattern in your data.

4 Deadend Graphs

4.1 How Has Crime Rate in Dallas Changed Over Time and What Factors May Have Contributed to These Changes?

We were unable to conclude the change in crime rates in Dallas over time because the data set for the year 2022 was incomplete. However, factors such as changes in demographics, economic conditions, law enforcement strategies, community involvement, and political and social factors may influence crime rates. For the given set, we constricted our variables to include only the time and date of the crime. We compressed our set to include the sex, race, and involvements of drugs in such cases. It's also important to note that the relationship between the factors and crime rates are complex and may vary depending on the specific circumstances of a community or economy of a country.

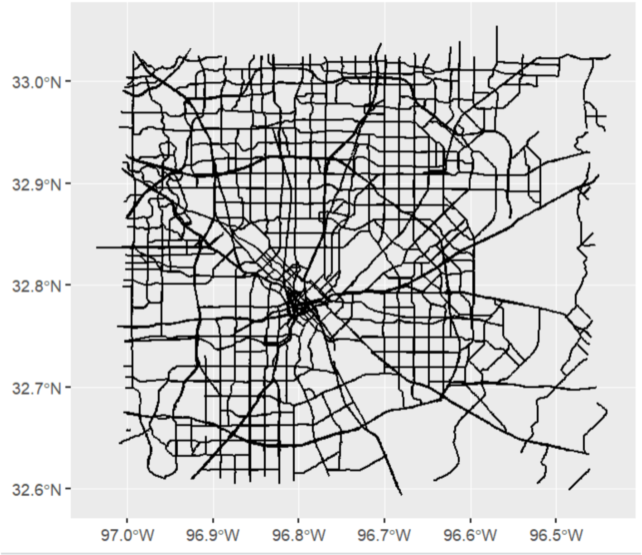
4.2 Are There Any Unique Factors that Contribute to Crime in Dallas, Such as the City's Demographics or Geography?

We could not conclude a graph to gather any unique factors that could contribute to the overall crime rate. Due to cleaning our data, we downsized the variables to include the arrest time, day, and month, the arrestee's race and sex, and whether the incident involved crime and the kind of drug used. Moreover, we previously determined if any relation existed between the frequency of crime and the month. We cannot determine if the city’s demographics or geography played a role in the crime rate from our dataset alone, outside resources would have to be consulted.

4.3 What is the Overall Crime Rate Trend in Dallas and Los Angeles? Which Areas in the Cities Have the Highest and Lowest Crime Rates, and Are There any Common Factors Among Those Areas?

We were unable to create a graph showing the overall crime rate in Dallas and Los Angeles. This is because the dates were over different periods in time.

4.4 What is the Distribution of Crimes Across Different Zipcodes in Dallas. Are There any Notable Findings or Trends?



Zip codes are useful in analyzing crime statistics because they provide a way to group crimes by location and compare crime rates between different areas1. By analyzing crime data by zip code, law enforcement can identify patterns and trends in criminal activity that can help them allocate resources more effectively.

Zip codes can also be used to identify areas with high crime rates and help law enforcement develop strategies to reduce crime in those areas3. Additionally, zip codes can be used to identify areas with low crime rates, which can help law enforcement identify factors that contribute to lower crime rates and develop strategies to replicate those factors in other areas

5 Conclusion

In our report, we analyzed two datasets of crime incidents in Dallas, Texas, being our primary focus and Los Angeles, California, major cities that share economic, demographic, and urban characteristics. Our goal was to understand the nature and patterns of crime in Dallas and compare them with Los Angeles. We wanted to uncover any possible trends that exist among the crime rate in Dallas and whether any similarities can be found in the Los Angeles dataset. Our analysis focused on variables such as date, time, location, weapon use, drug involvement, and arrestee information and how such varies with the frequency of crime.

In summary, there is no direct correlation between the attributes mentioned previously and the overall amount of incidents. However, we can make assumptions in regards to the arrestee's incentives and behaviors as well as the role of population density. Crime is influenced by numerous factors and thus, third-party resources must be studied to draw a thorough conclusion.

Future Work

We may include additional datasets, providing us with further information about the population background of a district. For example, given an area of higher rates of crime, we take into account the average and median household income. We can also study the demographic makeup of the arrestees. Rather than drawing assumptions on a specific racial group contributing to a large portion of the crime, we can study potential trends we see in said group's culture to explain such phenomena. Furthermore, we may check for relationships between a census district's average age, average income, or employment rate and the incidence of crime there.

The dates of crimes might also be contrasted with other date-based factors, such the weather. Such comparisons might be used to address issues like "Do crimes occur less frequently on rainy days?".

We would also like to compare crime rates in Dallas to other major cities in the country. We can uncover whether such cities follow similar trends and patterns. Moreover, we can compare the population density and income level amongst the areas, diving deeper into how the concentration of residents per square foot relates to overall crime rate. In terms of income level, we can ask questions like “are wealthier areas more vulnerable to crime” or expand on how residents of poorer districts are more incentivised to commit an offense.

What is the Importance of Studying Crime Rates?

First, it provides a clear understanding of the prevalence and nature of criminal activities in the area, which can help in developing effective law enforcement strategies. Second, crime statistics can help identify trends and patterns of criminal behavior, enabling law enforcement agencies to allocate resources more effectively and target their efforts towards specific areas or types of crime. Third, analyzing crime statistics can assist in identifying the root causes of criminal activity, such as poverty, unemployment, or drug use, which can inform policies aimed at reducing crime in the long term. Finally, crime statistics can provide valuable information for citizens, businesses, and policymakers to make informed decisions about where to live, work, and invest in a city. All in all, studying the statistics of crime in any city is a critical step in promoting a safe and desirable environment.

6 Code

```
## Data Cleaning
## DALLAS DATA CLEANING

# read the data
dcrime = read.csv("crime.csv")
lcrime = read.csv("crime_la.csv")

# Convert the date column to Date type
as.Date(dcrime$Arrest.Date, format = "%m/%d/%Y")

# Convert the time column to POSIXct
as.POSIXct(dcrime$Arrest.Time, format = "%I:%M:%S %p")

# Remove state column
dcrime$Arrest.State <- NULL

# Convert day of week to factor
dcrime$Arrest.Day.of.The.Week <- as.factor(dcrime$Arrest.Day.of.The.Week)

# Convert race to factor
dcrime$Arrestee.Race <- as.factor(dcrime$Arrestee.Race)

# Convert sex to factor
dcrime$Arrestee.Sex <- as.factor(dcrime$Arrestee.Sex)

# Convert drug related to factor
dcrime$Drug.Related <- as.factor(dcrime$Drug.Related)

# Convert drug type to factor
dcrime$Drug.Type <- as.factor(dcrime$Drug.Type)

## LA data cleaning
```

```
# Convert Date Occurred
lcrime$Date.Occurred <- as.Date(lcrime$Date.Occurred, format = "%m/%d/%Y")

# Convert Date Reported
lcrime$Date.Reported <- as.Date(lcrime$Date.Reported, format = "%m/%d/%Y")

# Convert Time Occurred
lcrime$Time.Occurred <- as.character(lcrime$Time.Occurred)
as.POSIXct(lcrime$Time.Occurred, format = "%H%M")


## Data analysis section
TX <- map_data("state", region="texas")
ggplot(TX, aes(x=long, y=lat))+geom_polygon()

# Question 3
# "pattern between time of day when arrests are made"
# create histogram with different time intervals and count frequency

# Question 4
# Distribution of crimes across zip codes
# create map of zip codes and color darkness prop to amount of crime
# or
# …

## Figure 3.1     Month vs. Frequency of Month Graph
## CRIME BY MONTH

# read data
setwd("~/3355")
dcrime <- read.csv("crime_dallas.csv")
lcrime <- read.csv("crime_la.csv")

# get packages
library(ggplot2)
library(lubridate)
```

```
# extract month from date
l_months <- format(as.Date(lcrime$Date.Reported, format = "%m/%d/%Y"), "%m")
d_months <- format(as.Date(mdy(dcrime$Arrest.Date), format = "%Y-%m-%d"),
"%m")

# get table of months and its frequency
dfreq <- data.frame(count = table(d_months))
lfreq <- data.frame(count = table(l_months))

# relabel
colnames(dfreq) <- c("month", "frequency")
colnames(lfreq) <- c("month", "frequency")

# display plot for dallas
ggplot() +
  geom_line(data = dfreq, aes(x = as.numeric(month), y = frequency),  color =
"#899AD6", size = 1) +
  ggtitle("Crime by Month in Dallas") +
  xlab("Month") + ylab("Frequency of Crime") +
  scale_x_continuous(name = "Month", breaks = 1:12, labels =
month.name[1:12])

# display plot for la
ggplot() +
  geom_line(data = lfreq, aes(x = as.numeric(month), y = frequency), color =
"#E47080", size = 1) +
  ggtitle("Crime by Month in Los Angeles") +
  xlab("Month") + ylab("Frequency of Crime") +
  scale_x_continuous(name = "Month", breaks = 1:12, labels =
month.name[1:12])

# normalize data using log transformation
# amount of total crime too large of a gap -> must normalize for better
comparison
dfreq$dnorm <- scale(dfreq$frequency)
```

```
lfreq$lnorm <- scale(lfreq$frequency)

# display plot dallas vs la
ggplot() +
  geom_line(data = dfreq, aes(x = as.numeric(month), y = dnorm, color =
"Dallas"), size = 1) +
  geom_line(data = lfreq, aes(x = as.numeric(month), y = lnorm, color = "Los
Angeles"), size = 1) +
  ggtitle("Crime by Month in Dallas vs. Los Angeles") +
  xlab("Month") +
  ylab("Frequency of Crime") +
  scale_color_manual(name = "City", values = c("Dallas" = "#899AD6", "Los
Angeles" = "#E47080")) +
  #theme(axis.text = element_text(size = 7)) +
  scale_x_continuous(name = "Month", breaks = 1:12, labels =
month.name[1:12])

## Figure 3.2 Crime by Location in Dallas

library(ggplot2)

# Load your data
dcrime <- read.csv("C:/Users/mathg/Downloads/crime (1).csv")

# Count the number of observations for each location
location_counts <- as.data.frame(table(dcrime$Arrest.Location))

# Get the top 5 locations
top_locations <- location_counts[order(location_counts$Freq, decreasing =
TRUE),][1:5,]

# Abbreviate the location names to one word
top_locations$Var1 <- sapply(strsplit(as.character(top_locations$Var1), " "),
`[`, 1)

# Load ggplot2 library
```

```
library(ggplot2)

# Create a barplot using ggplot
ggplot(data = top_locations, aes(x = Var1, y = Freq)) +
  geom_bar(stat = "identity", fill = "blue") +
  ggtitle("Top 5 Arrest Locations in Dallas") +
  xlab("Location") +
  ylab("Total Number of Crime")

## Figure 3.2     Crime by Location in Los Angeles

library(ggplot2)
lacrime <- read.csv("C:/Users/mathg/Downloads/crime_la.csv")
lacrime$Premise.Description <- substr(lacrime$Premise.Description, 1, 12)
top5 <- names(sort(table(lacrime$Premise.Description), decreasing =
TRUE)[1:5])
lacrime_top5 <- lacrime[lacrime$Premise.Description %in% top5, ]
ggplot(lacrime_top5, aes(x = Premise.Description)) +
  geom_bar(fill = "blue") +
  xlab("Location") +
  ylab("Total Number of Crimes")

## Figure 3.3     Age of Arrestee vs. Weapon in Use

# Load the necessary libraries
library(ggplot2)
library(dplyr)

# Read in the data
dcrime <- read.csv("C:/Users/mathg/Downloads/crime (1).csv")

# Find the top 5 arrest weapons by number of observations
top_weapons <- dcrime %>%
  group_by(Arrest.Weapon) %>%
  summarize(count = n()) %>%
  arrange(desc(count)) %>%
  head(5) %>%
  pull(Arrest.Weapon)

# Filter the data to only include the top 5 arrest weapons
dcrime_filtered <- dcrime %>%
  filter(Arrest.Weapon %in% top_weapons)

# Create the boxplot
ggplot(dcrime_filtered, aes(x = Arrest.Weapon, y =
Arrestee.Age.At.Arrest.Time, fill = Arrest.Weapon)) +
  geom_boxplot() +
  labs(x = "Weapons", y = "Age of Arrestee") +
  scale_fill_discrete(name = "Arrest Weapon")

## Figure 3.4     Drug-related Crime Pie Chart
df = as.data.frame(table(dcrime[which(dcrime$Drug.Related == "Yes"), ]$Drug.Type))


ggplot(df, aes(x="", y=Freq, fill=Var1)) +
  geom_bar(stat="identity", width=1) +
  coord_polar("y", start=0) +
  theme_void() +
  labs(title="Drug Types", fill="Category")

## Figure 3.5     Crime by Day of the Week
# Load the data
dcrime <- read.csv("C:/Users/mathg/Downloads/crime (1).csv")

# Create a factor variable for the days of the week
dcrime$Arrest.Day.of.The.Week <- factor(dcrime$Arrest.Day.of.The.Week, levels
= c("Mon", "Tue", "Wed", "Thu", "Fri", "Sat", "Sun"))

# Create a table of counts for each day of the week
counts <- table(dcrime$Arrest.Day.of.The.Week)

# Create the bar plot
```

```
barplot(counts, xlab = "Day of the Week", ylab = "Number of Arrests", main =
"Bar Plot of Arrests by Day of the Week")
```

7 Reference

Los Angeles Dataset: https://www.kaggle.com/datasets/cityofLA/crime-in-los-angeles