STAT 4355.001

Spring 2025

Dr. Chuan-Fa Tang

May 15, 2025

THE DALLAS DETECTIVES

Dallas 2014 Crime Data — Key Insights and Analysis

Prepared by

Vaishnavi Josyula, Esther Kim, Rayyan Nour

Introduction:

Crime is a significant concern for many communities and impacts multiple areas of society, including but not limited to public safety, economic growth, and quality of life. Dallas had the third-highest crime rate among major Texas metropolitan areas in 2024, according to Greco Neyland PC, a leading Texas law firm. According to the Rocky Mountain Collegian, the daily student newspaper of Colorado State University, Dallas experiences 38% of violent crimes attributed to gang activity, while property crimes occur at a rate of 32.21 incidents per 1,000 residents.

To enhance law enforcement strategies and gain a deeper understanding of crime trends in Dallas, we conducted a thorough analysis of the city's crime data to observe and understand significant underlying patterns. To effectively address this need for informed decision-making, our team analyzed the "Dallas Crimes" dataset. This dataset was acquired from the City of Dallas Open Data Portal (dallasopendata.com), which provides public access to a wide range of civic information. We chose the year 2014 due to its broad accessibility and the abundance of available data points and information. The "Dallas Crimes" dataset records reported criminal incidents within the geographical boundaries of Dallas. It has a substantial volume of information, containing over 81,000 individual entries, each representing a criminal event. Additionally, each entry is characterized by 45 distinct variables, giving us a comprehensive view of the circumstances surrounding each incident. These variables include critical details such as the date and time of the crime's occurrence, a specific categorization of the type of offense committed, the ZIP code where the incident took place, the designated police beat assigned to the location, and the classification of the premises involved. By organizing, cleaning, and analyzing this dataset, our primary objective is to look into the underlying patterns and identify key factors that significantly influence the distribution of crime across various geographical regions within the city and throughout different timeframes. Ultimately, this in-depth analysis aims to provide actionable insights that can contribute to more effective crime prevention and resource allocation strategies within the Dallas Police Department and other relevant civic agencies.

As such, our analysis will answer the following research questions:

RQ 1. At what time of day do most crimes occur?

RQ 2. What types of crimes are most common in each neighborhood?

RQ 3. Does the location, particularly ZIP code, affect the number of crimes reported?

RQ 4. Can we predict the number of crimes based on time, location, and crime type?

Data Description & Display

Data Description

We analyzed the 2014 “Dallas Crimes” dataset from the City of Dallas Open Data Portal, which includes over 81,000 records and 45 variables. For our analysis, we selected a set of key variables from the dataset that would help us understand and predict crime patterns across Dallas. The central variable in our study is CrimeCount, a continuous, engineered response variable that represents the number of crimes occurring within a given ZIP code during a weekly interval. This was created by grouping the data and counting the number of reported incidents per group.

To explore temporal trends, we used the offensedate field to extract the day of the week and month, and the offensestarttime field to extract the hour of the day, which allowed us to examine when crimes are most likely to occur. We also considered the offensedescription column, which provides a textual classification of the crime type, to compare offense trends across neighborhoods.

Geographic information was drawn from the offensezip column, which identifies the ZIP code of the crime, and the offensebeat variable, which corresponds to the police patrol beat or division, giving insight into regional law enforcement patterns. To understand the physical context of crimes, we used offensepremises, which indicates the type of location where the crime occurred, such as a street, residence, or parking lot. Overall, these primary variables formed the foundation of our data analysis and modeling.

Data preprocessing

Our data preprocessing phase started with the creation of the “CrimeCount” variable. This new feature aimed to provide a temporal quantification of criminal activity by aggregating criminal incidents based on three fundamental dimensions: the ZIP code where the crime occurred, the hour of the day when the offense commenced (derived from the offensestarttime column), and the specific category of the crime committed. This aggregation allowed for a more nuanced understanding of crime patterns, especially when comparing different areas (i.e., zip codes) in Dallas.

Next, for time-based analysis, we performed feature extraction on the offensestarttime column. We recognized that the temporal distribution of crime can vary significantly throughout the day, so we isolated the hour of the day as a distinct analytical unit. This transformation allowed us to investigate potential correlations between the time of day and the frequency or type of criminal activity.

Then, to make sure that the columns we chose were fit for regression in our scenario, we addressed the nature of certain key variables. While the zip code variable was represented numerically in the original dataset, it inherently functions as a categorical variable. As such, treating it as a continuous variable would lead to incorrect interpretations by the regression algorithm. Therefore, we explicitly converted it into a categorical data type as this conversion enabled the model to recognize and process the distinct levels within each category appropriately and as desired.

Moving on, we noticed that some values in certain columns were invalid. For instance, some zip codes were 0 and some columns (especially those with time) had inconsistent formatting. To make sure that our dataset was consistent, we identified and removed null and invalid entries for columns of interest. Our particular focus was placed on the zip code variable, where we identified and eliminated rows containing faulty or meaningless entries, such as the numerical value 0 or completely empty strings.

Lastly, rather than employing imputation techniques (which involve estimating and filling in missing values with modes and medians), we opted for the complete removal of rows containing null or invalid data points. Our rationale for this approach was based on the observation that the small proportion of missing data meant that their removal would not result in a significant loss of information since our dataset had a lot of rows as we had columns. Furthermore, removing problematic entries directly avoids the potential introduction of bias or artificial data patterns that can sometimes arise from imputation methods. After performing these steps, we moved on to exploring our dataset.

Data exploration

Our initial exploration of the dataset focused on its temporal characteristics, specifically the distribution of criminal activity across the 24 hours of a day. To visualize this, we generated a histogram (Figure 1) where the x-axis represented the hour of the day (ranging from 0 to 23) and the y-axis indicated the corresponding frequency, or count, of reported crimes. The resulting distribution revealed a clear pattern in the temporal occurrence of crime. The histogram highlighted a significant concentration of criminal incidents occurring during the evening hours, with a peak around 5:00 p.m. This suggests that factors prevalent during this time, such as increased social interaction, reduced visibility as daylight begins to fade, or the end of the traditional workday, may contribute to a higher likelihood for criminal activity. Conversely, the early morning hours, spanning approximately from 4:00 a.m. to 7:00 a.m., showed the lowest frequency of reported crimes. This period, characterized by reduced general activity and a larger proportion of the population engaged in sleep, appears to be a time of comparatively lower criminal incident rates.

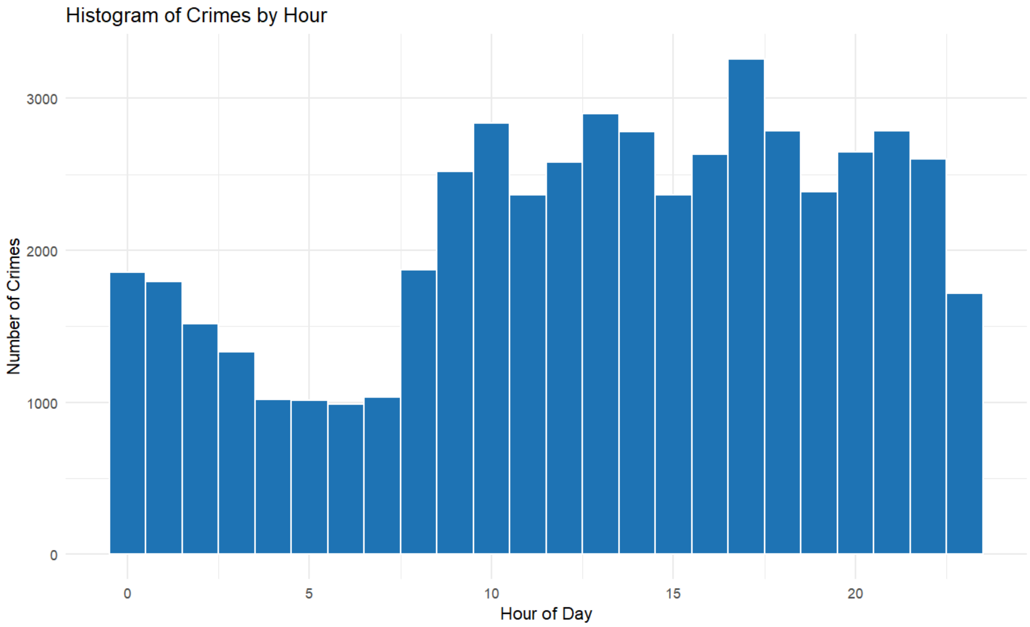


Figure 1: Number of crimes based on the hour of the day

Next, we analyzed the day of each committed crime. As illustrated in Figure 2, the frequency of reported crimes varies considerably across the week. The X-axis of the figure represents the days of the week, while the Y-axis displays the total number of crimes, calculated as the sum of all reported incidents for each respective day. We noticed that the data indicates a lower incidence of criminal activity on Sundays and Mondays. Following this initial lull, there is a gradual increase in crime occurring on Tuesdays, Wednesdays, and Thursdays. The peak in criminal activity is observed on Fridays and Saturdays, with significantly higher numbers of reported crimes compared to the earlier days of the week. This suggests a potential correlation between day-of-week patterns and criminal behavior. Some factors could include increased social activity, changes in alcohol consumption patterns, or variations in law enforcement presence across different days.

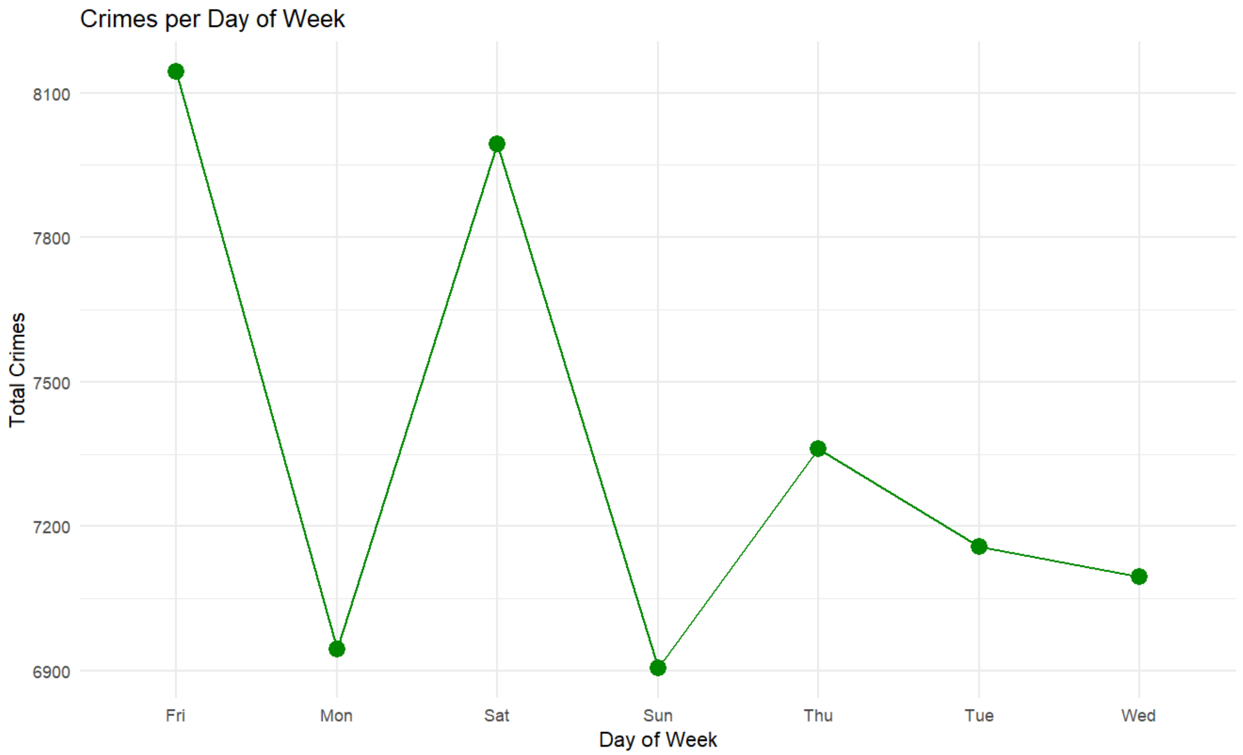


Figure 2: Number of crimes based on day of the week

Within the ten Dallas zip codes exhibiting the highest reported crime rates, a focused analysis revealed the five most prevalent crime types: unauthorized use of a motor vehicle (UUMV), theft, found property, criminal mischief, and burglary of a motor vehicle (BMV). First, we notice that vehicle related crime is the most frequent in the Dallas area, since two of the three offense type categories are directly related to motor vehicles. Next, we see that Figure 3 shows the distribution of these five offenses, and highlights a significant concentration of criminal mischief in zip code 75217, situated in East Dallas, which recorded the highest frequency of this particular crime. The adjacent zip code 75243, located near southwest Richardson, demonstrated the second-highest incidence of criminal mischief, suggesting a potential localized trend for this type of offense. Notably, when considering the aggregate frequency of all five identified crime categories, zip code 75217 also registered the highest overall crime count among the top ten high-crime areas. This suggests that East Dallas, specifically within the 75217 region, experiences a broader spectrum and higher volume of the most common criminal activities compared to other zip codes with elevated crime rates. To gain a more comprehensive understanding of the factors contributing to this observed pattern, we could analyze socioeconomic indicators such as poverty rates and literacy rates, alongside other potentially influential demographic and environmental variables.

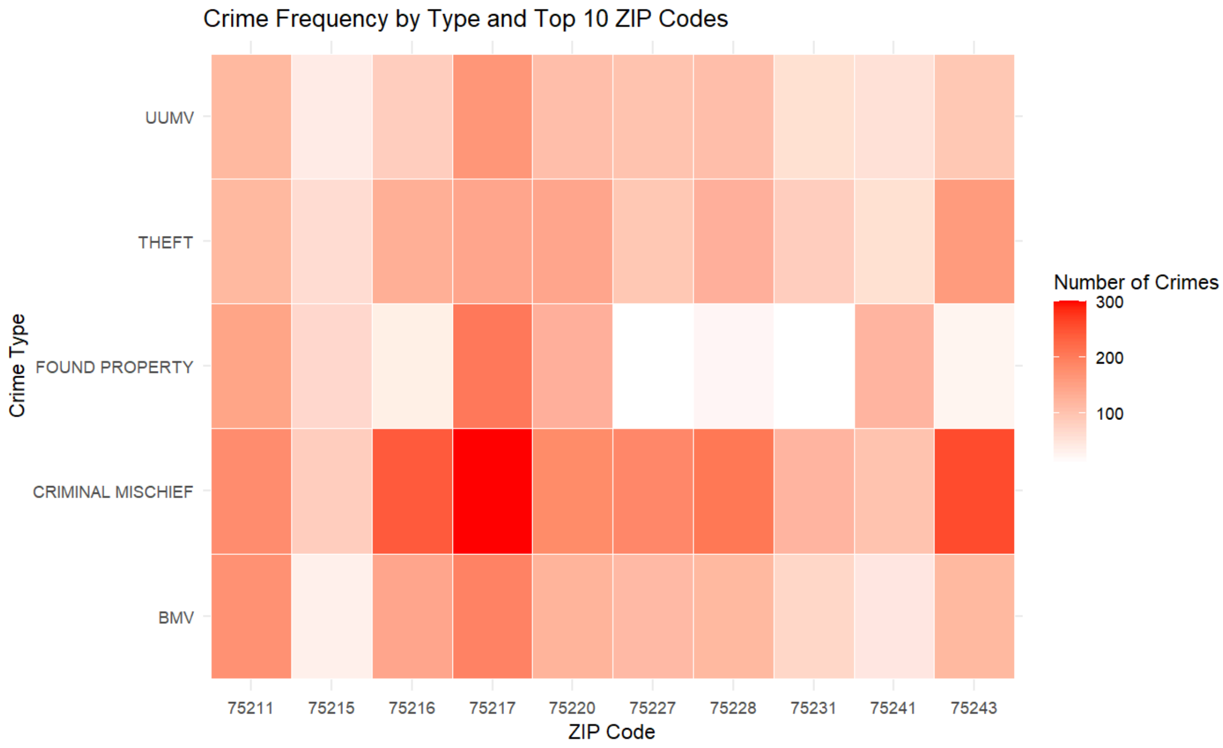


Figure 3: Crime frequency for the top 10 zip codes and top 5 crime types

To provide a more in-depth understanding of the crime distribution across Dallas zip codes, we generated two complementary visual representations: a static map providing an overview of crime counts across all zip codes and a dynamic, interactive map allowing for granular exploration of specific areas. The static map offers an immediate visual grasp of the overall spatial patterns of crime within Dallas by highlighting areas with higher and lower crime occurrences. However, we noticed that there was a lot of visual clutter in densely populated areas when displaying data for all zip codes simultaneously, so we developed an interactive dynamic map as a more refined analytical tool.

This dynamic map uses a GeoJSON file obtained from OpenDataDE on GitHub, which contains detailed geographical boundaries and zip code information for Texas. By integrating this geospatial data with our crime count dataset, we established a link between each zip code and its corresponding crime frequency. This integration was achieved by performing a join operation based on matching zip code identifiers. The resulting dataset was then used within the R programming environment. We ultimately used the Plotly library to construct the interactive map. The interactive capabilities of this map allow users to hover over specific geographic areas, revealing the associated zip code and the corresponding crime count. This feature enables a more detailed examination of crime hotspots and their immediate surroundings, which overcomes the limitations of a static map when analyzing localized patterns.

Like with the previous figure, our analysis of both the static and dynamic visualizations reveals a consistent trend: the south-eastern region of Dallas, specifically zip code 75217, exhibits the highest concentration of reported crimes. Figure 4 strongly supports the quantitative results obtained from our previous statistical analysis. Furthermore, the maps illustrate that the highest crime rates in the surrounding areas form a distinct U-shaped, almost circular pattern centered around zip code 75217, suggesting a potential spatial correlation or influence that warrants further investigation. Like before, a further and deeper analysis into socio-economic factors at these zip codes could help us understand why this pattern is prominent.

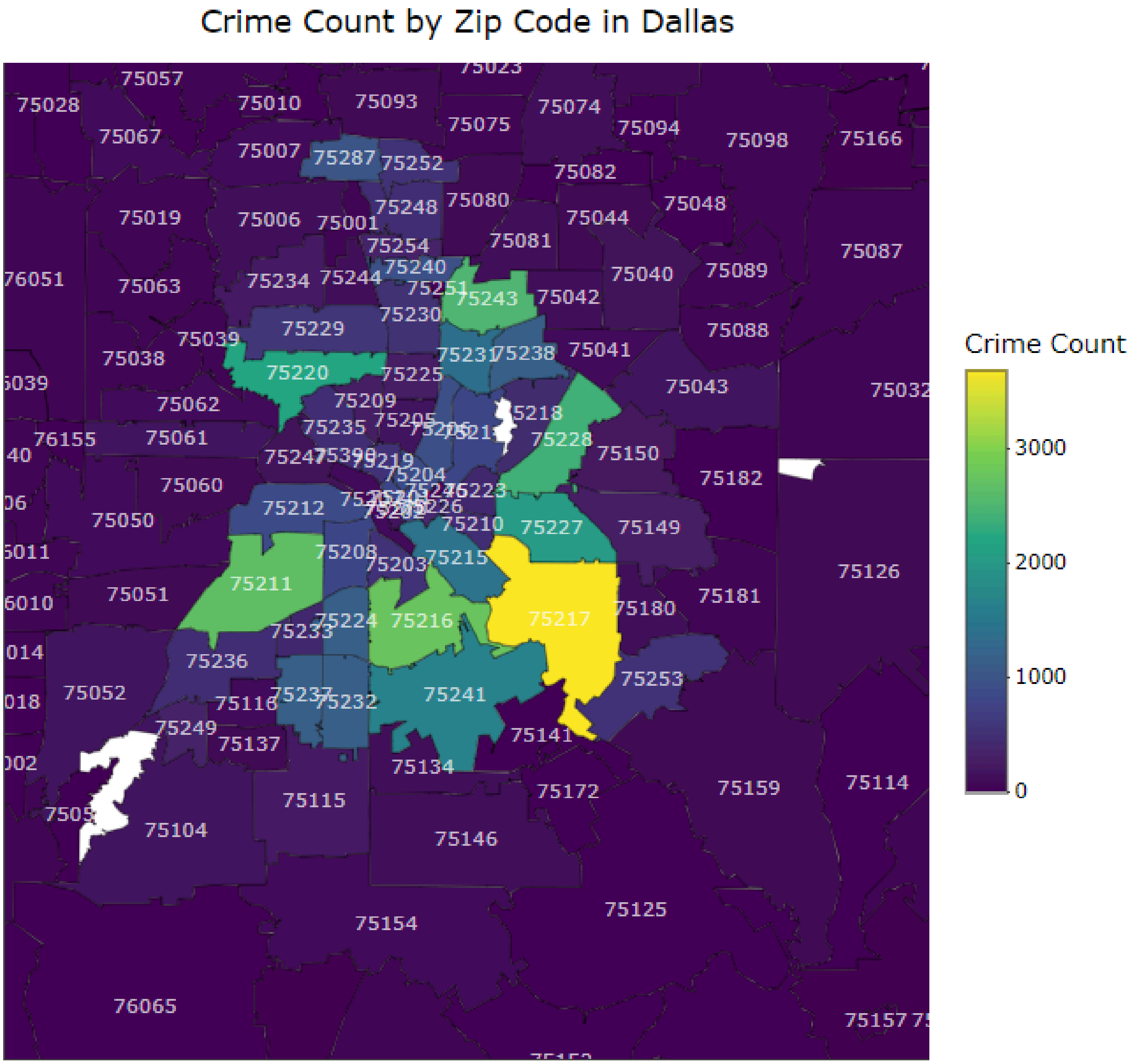


Figure 4: Crime count based on zip code: static map visualization

Since it is not possible to embed a video in the pdf submission, I have attached images (Figure 5) from the dynamic visualization. Additionally, here is a video showcasing the dynamic interaction with map visualization.

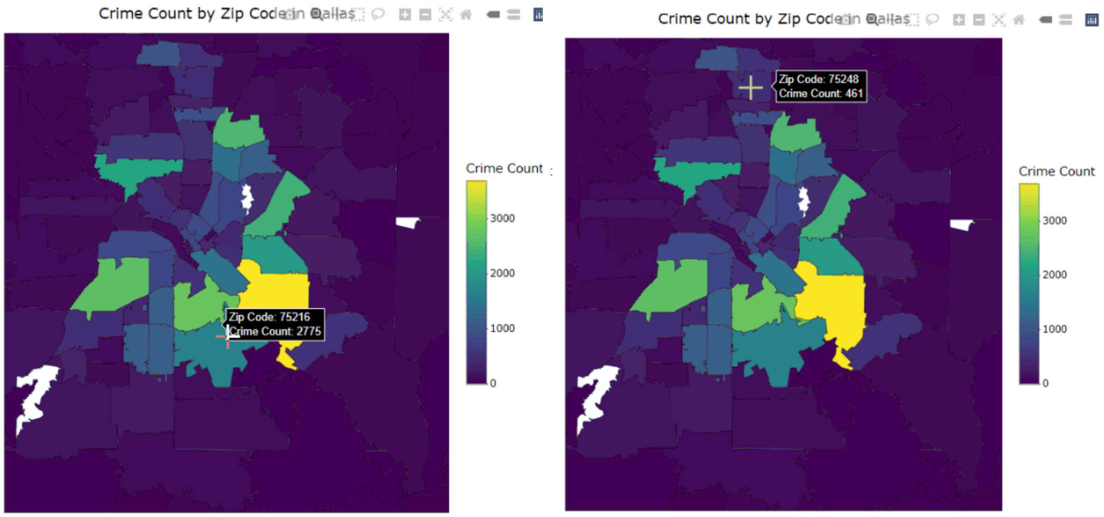


Figure 5: The image displays crime counts for two zip codes from the dynamic map visualization. On the left, zip code 75216 shows a crime count of 2775. On the right, zip code 75248 shows a crime count of 461.

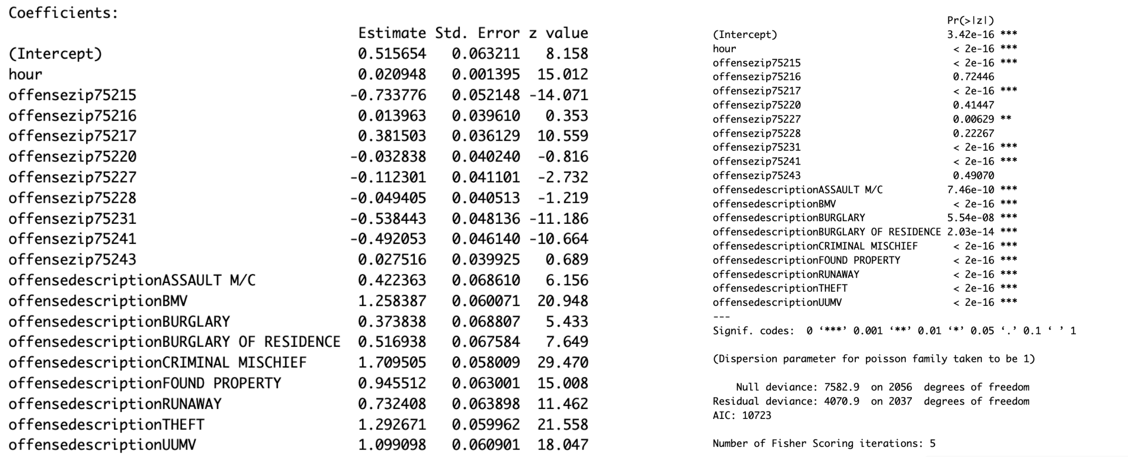
Overall, our analysis of the dataset's temporal dimensions revealed significant patterns in crime occurrences. Specifically, Friday consistently exhibits the highest frequency of reported incidents. Furthermore, the evening hours, particularly around 5 p.m., show a notable peak in crime counts. Geographically, the zip code 75217 stands out as having the largest number of recorded crimes. These findings provide valuable insights into when and where criminal activity is most concentrated within the dataset. A deeper understanding of these temporal and spatial patterns can inform resource allocation and targeted crime prevention strategies to local law enforcement and civic agencies.

Data analysis

Model fitting

After visualizing and understanding the dataset, we began with data analysis. To predict crime counts, we used a Poisson regression model since our response variable was count-based. We included hour, ZIP code, and crime type as predictors.

After fitting the model, we checked the summary and found that all three variables were statistically significant. The model showed a good fit, which means that it was able to explain a large part of the variation in crime counts.



```
Coefficients:
                                        Estimate Std. Error z value
(Intercept)                             0.515654   0.063211   8.158
hour                                    0.020948   0.001395  15.012
offensezip75215                        -0.733776   0.052148 -14.071
offensezip75216                         0.013963   0.039610   0.353
offensezip75217                         0.381503   0.036129  10.559
offensezip75220                        -0.032838   0.040240  -0.816
offensezip75227                        -0.112301   0.041101  -2.732
offensezip75228                        -0.049405   0.040513  -1.219
offensezip75231                        -0.538443   0.048136 -11.186
offensezip75241                        -0.492053   0.046140 -10.664
offensezip75243                         0.027516   0.039925   0.689
offensedescriptionASSAULT M/C           0.422363   0.068610   6.156
offensedescriptionBMV                   1.258387   0.060071  20.948
offensedescriptionBURGLARY              0.373838   0.068807   5.433
offensedescriptionBURGLARY OF RESIDENCE 0.516938   0.067584   7.649
offensedescriptionCRIMINAL MISCHIEF     1.709505   0.058009  29.470
offensedescriptionFOUND PROPERTY        0.945512   0.063001  15.008
offensedescriptionRUNAWAY               0.732408   0.063898  11.462
offensedescriptionTHEFT                 1.292671   0.059962  21.558
offensedescriptionUUMV                  1.099098   0.060901  18.047
```

```
                                        Pr(>|z|)
(Intercept)                             3.42e-16 ***
hour                                     < 2e-16 ***
offensezip75215                          < 2e-16 ***
offensezip75216                          0.72446
offensezip75217                          < 2e-16 ***
offensezip75220                          0.41447
offensezip75227                          0.00629 **
offensezip75228                          0.22267
offensezip75231                          < 2e-16 ***
offensezip75241                          < 2e-16 ***
offensezip75243                          0.49070
offensedescriptionASSAULT M/C           7.46e-10 ***
offensedescriptionBMV                    < 2e-16 ***
offensedescriptionBURGLARY              5.54e-08 ***
offensedescriptionBURGLARY OF RESIDENCE 2.03e-14 ***
offensedescriptionCRIMINAL MISCHIEF      < 2e-16 ***
offensedescriptionFOUND PROPERTY         < 2e-16 ***
offensedescriptionRUNAWAY                < 2e-16 ***
offensedescriptionTHEFT                  < 2e-16 ***
offensedescriptionUUMV                   < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 7582.9  on 2056  degrees of freedom
Residual deviance: 4070.9  on 2037  degrees of freedom
AIC: 10723

Number of Fisher Scoring iterations: 5
```

Figure 6: Results from Poisson Regression

Variable selection

- Hour - numeric
- Offensezip - categorical
- Offensedescription - categorical

It was based on our research questions. We made sure each variable added useful information to the model. After testing, all three variables were statistically significant and helped improve prediction accuracy.

Transformations

We didn’t need to make any transformations since the residuals didn’t show any major patterns, indicating a reasonable model fit.

Residual analysis

After building our Poisson regression model, we checked how well it fits the data by looking at the residuals. We plotted the deviance residuals against the model’s predicted values. The points looked fairly random and centered around zero, which means the model was not missing any obvious patterns.

We also checked for outliers or skewed patterns in a separate residuals plot. There were no extreme outliers or biases. Overall, the residuals confirmed that our model did a good job explaining the variation in crime counts.

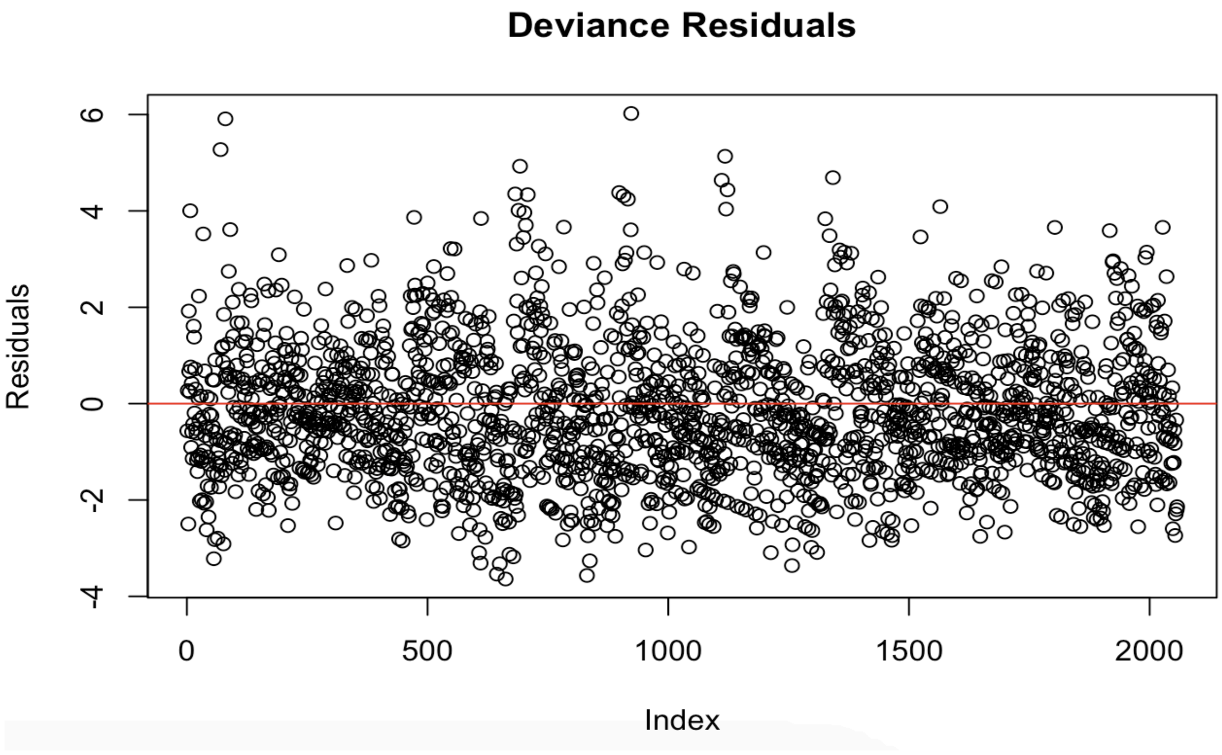


Figure 7: Residuals Plot

Conclusion

Our analysis of the Dallas Crimes dataset revealed that crime in the city is not a random occurrence but follows identifiable patterns driven by both time and location. By creating the crime count variable and analyzing multiple predictors, we found that ZIP code 75217 consistently reported the highest number of offenses, confirming that geographic location plays a major role in crime distribution. Time of day also emerged as a significant factor, with crime activity peaking in the mid-afternoon hours, particularly between 2:00 PM and 4:00 PM. Additionally, certain crime types were more prevalent in specific neighborhoods, suggesting that local context and environment influence the nature of offenses. Our multiple regression model showed that it is possible to predict the number of crimes in an area using a combination of time, ZIP code, and offense type. These findings support our motivation: Dallas continues to face serious public safety challenges, with high rates of violent and property crimes. By identifying when and where crimes are most likely to occur, this project contributes data-driven insights that can inform prevention strategies, law enforcement resource allocation, and community awareness initiatives. Ultimately, understanding these patterns brings us closer to the goal of safer neighborhoods throughout Dallas.

Reflection

One of the key strengths of our project was collaboration within our team. From the beginning, we established clear communication channels and made it a priority to meet weekly to discuss progress, share updates, and address challenges together. This regular interaction helped us build trust and allowed us to divide tasks efficiently. We used Google Docs, Google Slides, and Discord to keep track of our progress and communicate with each other.

That said, not everything went as smoothly as planned. One of the main difficulties we encountered was related to time management. Although we had set up a rough timeline in the early stages, several components of the analysis ended up being pushed close to the final deadline. This was partly due to the time it took to clean and explore such a large dataset, and partly due to scheduling conflicts that made it hard for everyone to consistently stay ahead of schedule. In hindsight, breaking the work into smaller, weekly milestones with clearer deadlines might have helped us better distribute the workload and avoid the last-minute crunch. We also learned the importance of starting the writing process earlier so that our conclusions could be shaped in parallel with our analysis, rather than afterward.

For future study, there are several promising directions we would be interested in exploring. First, increasing the temporal granularity of the data by analyzing crime patterns on a daily or hourly basis, rather than weekly, could lead to more precise and responsive predictions. Second, integrating external datasets, such as weather data, socioeconomic indicators (such as poverty, income, and unemployment rates), or school locations, could reveal broader contextual factors that influence crime trends across Dallas. Finally, while we used multiple linear regression for this project, future work could involve testing more advanced machine learning methods, such as decision trees, random forests, or neural networks. These models may be better able to capture the relationship between crime count and other variables better, and could improve prediction accuracy. With additional time and resources, these extensions could provide even more actionable insights for policy makers, law enforcement, and community leaders.

Materials & Appendix

Group member roles

To ensure our project progressed smoothly and efficiently, we divided the work among ourselves in a way that played to each team member’s strengths and interests.

- Vaishnavi took on several key responsibilities, including writing the introduction and conclusion sections, overseeing the data preprocessing process, creating the interactable visualization for crime count based on zip code, and ensuring that all deliverables and presentations were submitted on time.

- Esther focused on the core of our analysis, handling the model fitting, variable selection, and residual analysis. She took the lead in building our regression model, applying techniques such as poisson regression and analyzing residuals to ensure the model met assumptions and performed optimally.
- Rayyan focused on data exploration, creating visualizations, and interpreting the results. He was responsible for uncovering insights through exploratory data analysis, creating graphs, histograms, and thematic maps to visualize crime trends, and interpreting the results of these visualizations to help inform the model and conclusions.

References

1. City of Dallas. Dallas Open Data Portal. Retrieved from https://www.dallasopendata.com/dataset/Dallas-Crimes/pumt-d92b
2. Greco Neyland, PC. (2024, November 27). Texas Crime Rate by City – Latest Statistics. Retrieved from https://www.greconeylandtx.com/blog/texas-crime-rate-by-city/
3. The Rocky Mountain Collegian. (2025, February 19). Is Dallas, TX A Safe Place To Live? Retrieved from https://collegian.com/sponsored/2025/02/is-dallas-tx-a-safe-place-to-live/
4. OpenDataDE. (n.d.). State-zip-code-GeoJSON [GitHub repository]. Retrieved from https://github.com/OpenDataDE/State-zip-code-GeoJSON/blob/master/tx_texas_zip_codes_geo.min.json

R code

```
# Data preprocessing code:
# Load necessary libraries
library(dplyr)
library(lubridate)

dallas_crimes <-read.csv("C:/Users/vaish/Downloads/Dallas_Crimes_Dataset_4_28.csv", header = TRUE) # read  data
summary(dallas_crimes) # Load data and print summary

# Basic Dataset Dimensions
num_rows <- nrow(dallas_crimes)
num_cols <- ncol(dallas_crimes)
print(num_rows)

cat("Dataset Summary:\n\n")
cat(paste("Number of Rows (Records):", num_rows, "\n"))
cat(paste("Number of Columns (Features/Variables):", num_cols, "\n\n"))

# Data Types of Columns (Overview)
column_types <- sapply(dallas_crimes, class)
unique_types <- unique(column_types)
cat("Column Data Types (Overview):\n")
for (type in unique_types) {
  cols_of_type <- names(column_types[column_types == type])
  cat(paste("-", type, ":", paste(head(cols_of_type, 5), collapse = ", "),
            ifelse(length(cols_of_type) > 5, "...", ""), "\n"))
}
cat("\n")

# Unique Values in Key Categorical Columns
key_categorical_cols <- c("offensedescription", "offensebeat", "offensezip")
cat("Number of Unique Values in Key Categorical Columns:\n")
for (col in key_categorical_cols) {
  if (col %in% names(dallas_crimes)) {
    num_unique <- length(unique(dallas_crimes[[col]]))
    cat(paste("-", col, ":", num_unique, "\n"))
  } else {
    cat(paste("-", col, ": Column not found\n"))
  }
}
cat("\n")

# Missing Values (Overview)
missing_values <- sum(is.na(dallas_crimes))
cat("Missing Values (Overview):\n")
cat(paste("Total Number of Missing Values:", missing_values, "\n"))

# Proportion of Missing Values per Column (Top 5)
missing_by_column <- colSums(is.na(dallas_crimes))
missing_proportion <- sort(missing_by_column / num_rows, decreasing = TRUE)
cat("\nProportion of Missing Values per Column (Top 5):\n")
if (any(missing_proportion > 0)) {
  for (i in 1:min(5, length(missing_proportion))) {
    cat(paste("-", names(missing_proportion)[i], ":", round(missing_proportion[i], 3), "\n"))
  }
} else {
  cat("No missing values found.\n")
}
cat("\n")

# Basic Descriptive Statistics for Numerical Columns (if any)
numerical_cols <- names(dallas_crimes)[sapply(dallas_crimes, is.numeric)]
if (length(numerical_cols) > 0) {
  cat("Basic Descriptive Statistics for Numerical Columns (First 3):\n")
  print(summary(dallas_crimes[, head(numerical_cols, 3)]))
  cat("\n")
}
```

```
# Top Offense Descriptions (Top 5)
if ("offensedescription" %in% names(dallas_crimes)) {
  top_offenses <- head(sort(table(dallas_crimes$offensedescription), decreasing = TRUE), 5)
  cat("Top 5 Most Common Offense Descriptions:\n")
  print(top_offenses)
  cat("\n")
}

# Answering RQs
df <- read.csv("~/Downloads/Dallas_-_Crimes_20250429.csv", stringsAsFactors = FALSE, header =
TRUE)
names(df) <- trimws(names(df))
print(names(df))
df$hour <- as.numeric(substr(df$offensestarttime, 1, 2))

regression_df <- df[, c("offensezip", "offensedescription", "hour")]
regression_df <- na.omit(regression_df)
head(regression_df)

# Question 1 (what time)
hourly_crimes <- table(regression_df$hour)

sorted_crimes <- sort(hourly_crimes, decreasing = TRUE)
print(sorted_crimes)



# Question 2 (high crime rate ZIP code)
library(dplyr)

crime_counts <- regression_df %>%
  group_by(offensezip, offensedescription) %>%
  summarise(count = n(), .groups = "drop")
```

```
crime_counts <- crime_counts %>% filter(offensezip >= 10000)

zip_total <- crime_counts %>%
  group_by(offensezip) %>%
  summarise(total = sum(count), .groups = "drop") %>%
  arrange(desc(total))

top_zips <- head(zip_total$offensezip, 10) # for top 10

top_crime_by_zip <- crime_counts %>%
  filter(offensezip %in% top_zips) %>%
  group_by(offensezip) %>%
  slice_max(order_by = count, n = 1) %>%
  ungroup()

top_crime_by_zip <- top_crime_by_zip %>%
  arrange(desc(count))

top_crime_by_zip #results

# Question 3
clean_df <- regression_df %>%
  filter(!is.na(offensezip), offensezip >= 10000)

zip_crime_counts <- clean_df %>%
  group_by(offensezip) %>%
  summarise(crime_counts = n(), .groups = "drop") %>%
  arrange(desc(crime_counts))

head(zip_crime_counts)
# modeled crime counts using both ZIP code and time of day.
# The results show that location(ZIP code) has a statistically
# significant influence on the number of crimes reported.
# Question 4
top_zips <- regression_df %>%
  count(offensezip) %>%
  filter(offensezip >= 10000) %>%
  arrange(desc(n)) %>%
  slice(1:10) %>%
  pull(offensezip)

top_crimes <- regression_df %>%
  count(offensedescription) %>%
  arrange(desc(n)) %>%
  slice(1:10) %>%
  pull(offensedescription)

crime_model_df <- regression_df %>%
  filter(offensezip %in% top_zips,
         offensedescription %in% top_crimes) %>%
  group_by(offensezip, hour, offensedescription) %>%
  summarise(crime_counts = n(), .groups = "drop")

crime_model_df$offensezip <- as.factor(crime_model_df$offensezip)
crime_model_df$offensedescription <- as.factor(crime_model_df$offensedescription)

model <- glm(crime_counts ~ hour + offensezip + offensedescription,
             data = crime_model_df,
             family = poisson())

summary(model)

# Residual Analysis
res <- residuals(model, type = "deviance")
```

```
fitted_vals <- fitted(model)

plot(fitted_vals, res,
     xlab = "Fitted values",
     ylab = "Deviance residuals",
     main = "Residuals vs Fitted Values")
abline(h = 0, col = "red")

plot(res, main = "Deviance Residuals", ylab = "Residuals")
abline(h = 0, col = "red")

# Visualization
# Interactable graph: crime count by zip count (dynamic/animated/interactive)
# Load libraries
library(sf)
library(ggplot2)
library(dplyr)
library(plotly)
library(hms)

crime_df <- read.csv("C:/Users/vaish/Downloads/dallas_crimes_processed.csv", stringsAsFactors =
FALSE, header = TRUE)

# Load Dallas Zip Code Boundaries from JSON:
dallas_zips <- st_read("C:/Users/vaish/Downloads/dallas/tx_texas_zip_codes_geo.min.json")

# Rename the Zip Code Column in dallas_zips
dallas_zips <- dallas_zips %>%
  rename(offensezip = ZCTA5CE10)

# Dallas crime dataset:
crime_df$offensezip <- as.character(crime_df$offensezip)

# Aggregate Crime Counts by Zip Code
```

```
crime_counts_by_zip <- crime_df %>%
  group_by(offensezip) %>%
  summarise(CrimeCount = n())

# Merge Crime Data with Zip Code Boundaries
merged_data <- dallas_zips %>%
  left_join(crime_counts_by_zip, by = "offensezip") %>%
  mutate(CrimeCount = ifelse(is.na(CrimeCount), 0, CrimeCount))

# Find Centroids for Label Placement
# Calculate the centroid of each zip code polygon
zip_centroids <- st_centroid(merged_data) %>%
  mutate(centroid_lon = st_coordinates(.)[,1],
         centroid_lat = st_coordinates(.)[,2])

# Create the Interactive Map using plotly
merged_data_sf <- st_as_sf(merged_data)
p <- ggplot(merged_data_sf) +
  geom_sf(aes(fill = CrimeCount,
              text = paste("Zip Code:", offensezip, "<br>Crime Count:", CrimeCount)), # Keep CrimeCount in
hover
          color = "black", linewidth = 0.1) +
  scale_fill_viridis_c(option = "viridis", name = "Crime Count") +
  labs(title = "Crime Count by Zip Code in Dallas") + # Changed title
  theme_void() +
  theme(plot.title = element_text(hjust = 0.5)) +
  coord_sf(xlim = c(-97.0, -96.5), ylim = c(32.5, 33.0)) # Removed geom_text

# Convert to plotly
p <- ggplotly(p, tooltip = "text")

# Display the plot
p
```

```
# static graph:

# Load libraries
library(sf)
library(ggplot2)
library(dplyr)
library(plotly)
library(hms)

crime_df <- read.csv("C:/Users/vaish/Downloads/dallas_crimes_processed.csv", stringsAsFactors = FALSE, header = TRUE)

# --- Step 1: Load Dallas Zip Code Boundaries from JSON ---
dallas_zips <- st_read("C:/Users/vaish/Downloads/dallas/tx_texas_zip_codes_geo.min.json")

# --- Rename the Zip Code Column in dallas_zips ---
dallas_zips <- dallas_zips %>%
  rename(offensezip = ZCTA5CE10)

# --- Step 2: Prepare Your Crime Data ---
# Ensure 'offensezip' is character
crime_df$offensezip <- as.character(crime_df$offensezip)

# --- Step 3: Aggregate Crime Counts by Zip Code ---
crime_counts_by_zip <- crime_df %>%
  group_by(offensezip) %>%
  summarise(CrimeCount = n())

# --- Step 4: Merge Crime Data with Zip Code Boundaries ---
merged_data <- dallas_zips %>%
  left_join(crime_counts_by_zip, by = "offensezip") %>%
  mutate(CrimeCount = ifelse(is.na(CrimeCount), 0, CrimeCount))

# --- Step 5: Find Centroids for Label Placement ---
# Calculate the centroid of each zip code polygon
zip_centroids <- st_centroid(merged_data) %>%
  mutate(centroid_lon = st_coordinates(.)[,1],
         centroid_lat = st_coordinates(.)[,2])

# --- Step 6: Create the Interactive Map using plotly ---
# Create a simple feature collection
merged_data_sf <- st_as_sf(merged_data)

# Create the base plot
p <- ggplot(merged_data_sf) +
  geom_sf(aes(fill = CrimeCount,
              text = paste("Zip Code:", offensezip, "<br>Crime Count:", CrimeCount)),
          color = "black", linewidth = 0.1) +
  scale_fill_viridis_c(option = "viridis", name = "Crime Count") +
  labs(title = "Crime Count by Zip Code in Dallas") + # Changed title
  theme_void() +
  theme(plot.title = element_text(hjust = 0.5)) +
  coord_sf(xlim = c(-97.0, -96.5), ylim = c(32.5, 33.0)) +
  # Add labels for all zip codes
  geom_text(data = zip_centroids, # Use all centroids
            aes(x = centroid_lon, y = centroid_lat, label = offensezip),
            size = 3, color = "white", fontface = "bold", alpha = 0.7)

# Convert to plotly
p <- ggplotly(p, tooltip = "text")

# Display the plot
p
```

```
# ======== Load Libraries ========
library(ggplot2)
library(readr)
library(dplyr)



# ======== Load Data ========

setwd ("C:/Users/rayya/Downloads")
df <- read_csv("dallas_crimes_processed.csv")

# ======== 1. Histogram of Crimes by Hour ========
# Convert start time to hour
df$hour <- hour(hms(df$offensestarttime))

ggplot(df, aes(x = hour)) +
  geom_histogram(binwidth = 1, fill = "#1f77b4", color = "white") +
  labs(title = "Histogram of Crimes by Hour",
       x = "Hour of Day",
       y = "Number of Crimes") +
  theme_minimal()

# ======== 2. Bar Chart of Top Crime Types by ZIP ========
# Group by day of week and count crimes
weekday_counts <- df %>%
  group_by(OffenseDayOfWeek) %>%
  summarise(CrimeCount = n(), .groups = "drop")

# Plot the line chart
ggplot(weekday_counts, aes(x = OffenseDayOfWeek, y = CrimeCount)) +
  geom_point(size = 4, color = "green4") +
  geom_line(group = 1, color = "green4") +
  labs(title = "Crimes per Day of Week",
       x = "Day of Week",
       y = "Total Crimes") +
  theme_minimal()
# ======== 3. Bar Chart of Total Crimes per ZIP ========
# Heatmap-style ZIP vs. Crime Type
# Get top 10 ZIPs and top 5 crime types
top_zips <- df %>% count(offensezip, sort = TRUE) %>% head(10) %>% pull(offensezip)
top_crimes <- df %>% count(offensedescription, sort = TRUE) %>% head(5) %>%
pull(offensedescription)

# Prepare heatmap data
df_tile <- df %>%
  filter(offensezip %in% top_zips, offensedescription %in% top_crimes) %>%
  group_by(offensezip, offensedescription) %>%
  summarise(n = n(), .groups = "drop")

# Plot heatmap
ggplot(df_tile, aes(x = factor(offensezip), y = offensedescription, fill = n)) +
  geom_tile(color = "white") +
  scale_fill_gradient(low = "white", high = "red") +
  labs(title = "Crime Frequency by Type and Top 10 ZIP Codes",
       x = "ZIP Code",
       y = "Crime Type",
       fill = "Number of Crimes") +
  theme_minimal()
```