Statistical Analysis of Denver Crime Data

Hadiseh Mesbah

February 15, 2024

Abstract

Our project analyzes Denver's crime data using various visualization and summarization techniques to explore and understand the dataset's variables. Each feature within the dataset serves a specific purpose in painting a comprehensive picture of the crime landscape in Denver. Below, we describe our approach to visualizing and summarizing these variables.

1 Introduction

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum. Our project analyzing Denver's crime data uses various visualization and summarization techniques to explore and understand the dataset's variables. Each feature within the dataset serves a specific purpose in painting a comprehensive picture of the crime landscape in Denver. Below, we describe our approach to visualizing and summarizing these variables:

- **Incident ID and Offense ID:** These identifiers are unique to each crime incident and offense. While not directly applicable to visualization, they are crucial for data integrity checks and ensuring accurate record linkage.
- **Offense Code and Offense Code Extension:** These numerical codes represent the specific type of offense committed. We use histograms to visualize the frequency of different offense codes and their extensions, highlighting Denver's most common types of crimes.

- **Offense Type ID and Offense Category ID:** These text descriptions provide more detail on the nature of the offense. Bar charts summarize the frequency of various offense types and categories, offering insight into the prevalent crime categories in the city.
- **First Occurrence Date, Last Occurrence Date, and Reported Date:** These dates indicate the timing of criminal activities and their reporting to the police. Time series plots analyze trends, including seasonal variations and potential delays between occurrences and reporting.
- **Incident Address, GEO X, GEO Y, GEO LAT, and GEO LON:** Geographic information is pivotal for spatial analysis. We use geographic plotting tools to map the locations of crimes, identifying hotspots and patterns across different neighborhoods. Heatmaps are particularly useful in visualizing the concentration of crimes in specific areas.
- **District ID, Precinct ID, and Neighborhood ID:** These identifiers provide a hierarchical geographical context of crimes. Pie charts and choropleth maps depict the distribution of crimes across different districts, precincts, and neighborhoods, revealing areas with higher crime rates.
- **Is Crime and Is Traffic:** These binary variables indicate whether an incident is a criminal or traffic offense. Stacked bar charts compare the proportion of crime-related incidents versus traffic-related incidents, both overall and within specific areas or times.

The Dataset is like this:

offense_category_id	first_occurrence_date	last_occurrence_date	reported_date	incident_address	geo_x	geo_y	geo_lon	geo_lat	district_id	precinct_id
public-disorder	2/10/2022 2:50:00 AM	NaN	2/10/2022 3:16:00 AM	1107 N SANTA FE DR	3140929.0	1692612.0	-104.998910	39.733957	1	123
public-disorder	7/7/2021 9:02:00 PM	NaN	7/8/2021 12:55:00 AM	815 16TH ST	3142470.0	1697098.0	-104.993342	39.746248	6	611
public-disorder	10/29/2020 1:30:00 AM	NaN	10/29/2020 4:31:00 AM	4745 N FEDERAL BLVD	3133352.0	1710396.0	-105.025520	39.782888	1	111
public-disorder	9/6/2018 5:00:00 PM	9/6/2018 11:00:00 PM	9/7/2018 9:58:00 AM	65 S FEDERAL BLVD	3133534.0	1685797.0	-105.025330	39.715357	4	411
public-disorder	5/8/2020 5:00:00 AM	5/8/2020 6:30:00 PM	5/13/2020 10:00:00 AM	12295 E ALBROOK DR	3184065.0	1710782.0	-104.845074	39.783082	5	521

Figure 1: Dataset sampels

We also have a complementary dataset named code, which is about the crimes with the following columns:

- **OBJECTID:** This is likely a unique identifier for each ”Codes” dataset entry. It serves as a key for database management and ensures data integrity but might not be directly useful for analytical purposes.
- **OFFENSE_CODE and OFFENSE_CODE_EXTENSION:** These fields correspond to the offense codes in the primary crime dataset. They are crucial for linking the detailed offense descriptions and categories in the ”Codes” dataset to each crime incident. Analysis can leverage these links to decode the numerical offense codes into more meaningful descriptions.

- **OFFENSE_TYPE_ID and OFFENSE_TYPE_NAME:** These provide specific descriptions of offenses. By mapping these to the primary dataset, you can perform a more nuanced analysis of crime types, identifying patterns and trends in particular offenses across different areas and times.
- **OFFENSE_CATEGORY_ID and OFFENSE_CATEGORY_NAME:** These fields generalize the offenses into broader categories, such as theft, assault, or traffic violations. They are essential for thematic analyses, allowing for examining more general crime trends and the effectiveness of law enforcement strategies against different crime categories.
- **IS_CRIME and IS_TRAFFIC:** Indicating whether an offense is classified as a crime or a traffic incident, these variables align with the primary dataset's similar fields. They confirm the nature of each offense and can be used to filter and compare crime and traffic incident trends, contributing to targeted policy interventions.

A small sample of this dataset is like this:

	OBJECTID	OFFENSE_CODE	OFFENSE_CODE_EXTENSION	OFFENSE_TYPE_ID	OFFENSE_TYPE_NAME	OFFENSE_CATEGORY_ID	OFFENSE_CATEGORY_NAME	IS_CRIME	IS_TRAFFIC
0	1	2804	1	stolen-property-possession	Possession of stolen property	all-other-crimes	All Other Crimes	1	0
1	2	2804	2	fraud-possess-financial-device	Possession of a financial device	all-other-crimes	All Other Crimes	1	0
2	3	2901	0	damaged-prop-bus	Damaged business property	public-disorder	Public Disorder	1	0
3	4	2902	0	criminal-mischief-private	Criminal mischief to private property	public-disorder	Public Disorder	1	0
4	5	2903	0	criminal-mischief-public	Criminal mischief to public property	public-disorder	Public Disorder	1	0

Figure 2: Crime Dataset Sampels

Since there are many missing values in this dataset, we first start by fixing these values, and then we will visualize the data.

2 Parametric Inference and Estimation

There are lots of missing values for the last_occurance_date. There are missing values for a location where the crime happened and some missing values for district_ids, precinct_ids, and incident_address.

This dataset has 386864 rows, and based on the above plot, it is evident that more than half of last_occurrence_date is missing. For handling different missing values, we have many approaches:

2.1 Deleting Records

- **Listwise Deletion:** Remove all records (rows) missing any value. This method is simple but can lead to significant data loss, especially if missingness is widespread across your dataset.

incident_id	0
offense_id	0
offense_code	0
offense_code_extension	0
offense_type_id	0
offense_category_id	0
first_occurrence_date	0
last_occurrence_date	0
reported_date	0
incident_address	15503
geo_x	15503
geo_y	15503
geo_lon	15769
geo_lat	15769
district_id	57
precinct_id	0
neighborhood_id	689
is_crime	0
is_traffic	0
victim_count	0

dtype: int64

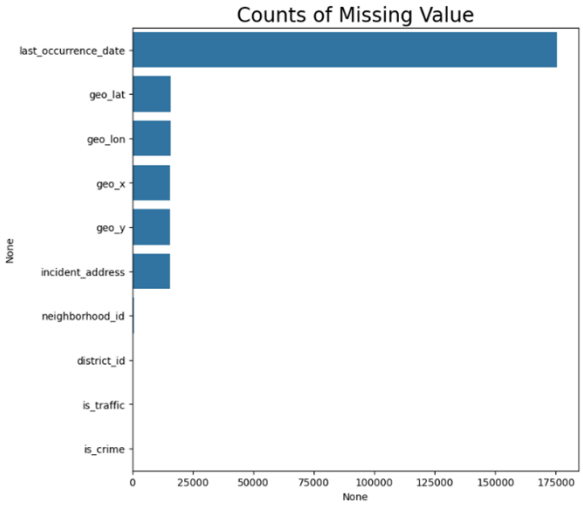


Figure 3: Missing Values

- **Pairwise Deletion:** Used primarily in statistical analyses, where calculations are performed only on cases with complete data for the variables of interest. It maximizes data use but can introduce bias if the patterns of missingness vary across variables.

2.2 Imputation

- **Mean/Median/Mode Imputation:** Replace missing values with the non-missing data's mean, median, or mode (most frequent value) in the same column. This method is straightforward but can reduce variability and potentially introduce bias.
- **Predictive Imputation:** Use statistical models (e.g., linear regression, decision trees) to predict and fill in missing values based on other variables in the dataset. This method considers correlations between variables but can be computationally intensive and may overfit the data.
- **K-Nearest Neighbors (KNN) Imputation:** Replace missing values using the nearest neighbors' values in the multidimensional feature space. It's useful when data patterns are complex, but the choice of k (the number of neighbors) can significantly affect the imputation quality.
- **Hot Deck Imputation:** Randomly select a non-missing value from a similar item (e.g., within the same cluster) to fill in the missing value. This method maintains the distribution of data but requires careful definition of "similar" items.
- **Multiple Imputation:** Create multiple imputed datasets, analyze each separately, and then combine the results. This approach reflects the uncertainty about the missing data but is more complex to implement and interpret.

2.3 Handling Missing Values for `last_occurrence_date`

For this variable, there are three approaches:

1. **Dropping the Entire Column:** Since this column does not contain any critical information and we will not use it in our analysis, we can drop the column.
2. **Replacing:** We can replace the missing values with the `first_occurrence_date` or with the `reported_date`.
3. **Replacing with Unknown:** We opted for the latter approach and replaced all missing values with the `first_occurrence_date`.

2.4 Handling Missing Values for `incident_address`

For this variable, there are three approaches:

1. **Dropping the Entire Column:** Since this column does not contain any critical information and we will not use it in our analysis, we can drop the column.
2. **Replacing with values:** We can replace the missing values with the combination of `district_id`, `precinct_id`, and `neighborhood`. However, this adds a little redundancy to data, but it is better than losing data.
3. **Replacing with Unknown:** Replace the address with unknown.

We opted for the second approach and replaced all missing values with the incident_address.

2.5 Handling Missing Values for `district_id`

let's see a join of precinct and neighborhood and district:

And map of Denver is like this

Based on this map, we have some information about Denver City and how the districts and precincts are divided. Firstly, there is 7 districts named from 1 to 7, and there is no U district; therefore, the U in the plot for districts is for missing values. Second, the starting digit of the precinct shows the district that the precinct is located in. Because there is no 9 district, the precinct named 999 values are for missing values precinct. Third, for the place where we have the district number since the position is danced around a specified location, we can calculate the mean geo_lon and geo_lat for a district.

For this variable, there are two approaches:

1. **Replacing with values:**

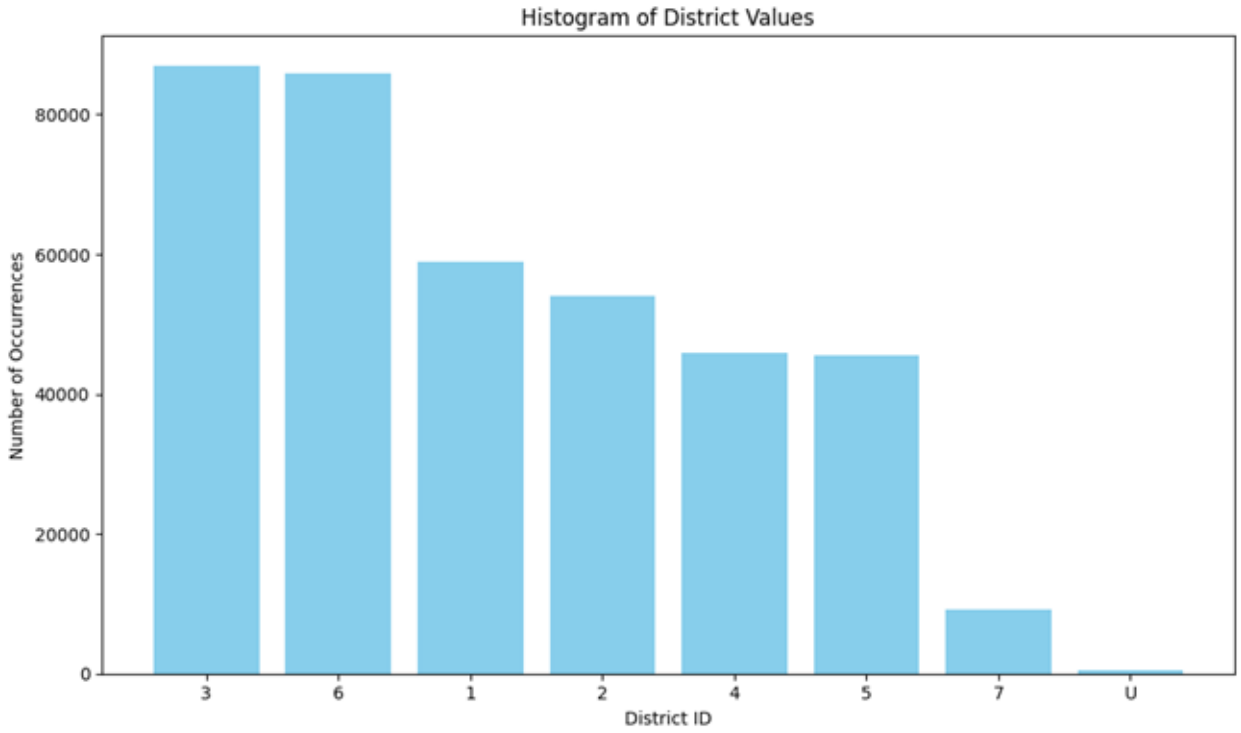


Figure 4: Missing Values For district_id

	precinct_id	district_id	neighborhood_id	count
0	111	1	berkeley	3305
1	111	1	chaffee-park	1527
2	111	1	highland	29
3	111	1	regis	1879
4	111	1	sunnyside	2539
5	111	1	west-highland	30
6	112	1	chaffee-park	283
7	112	1	elyria-swansea	39
8	112	1	globeville	4419
9	112	1	highland	40
10	112	1	regis	1
11	112	1	sunnyside	1761
12	112	1	west-highland	1
13	113	1	central-park	2
14	113	1	highland	6292
15	113	1	jefferson-park	54
16	113	1	sloan-lake	18
17	113	1	union-station	2
18	113	1	west-highland	3585
19	121	1	auraria	60
20	121	1	gateway-green-valley-ranch	1
21	121	1	hale	1
22	121	1	jefferson-park	2577

Figure 5: join of precinct and neighborhood and district

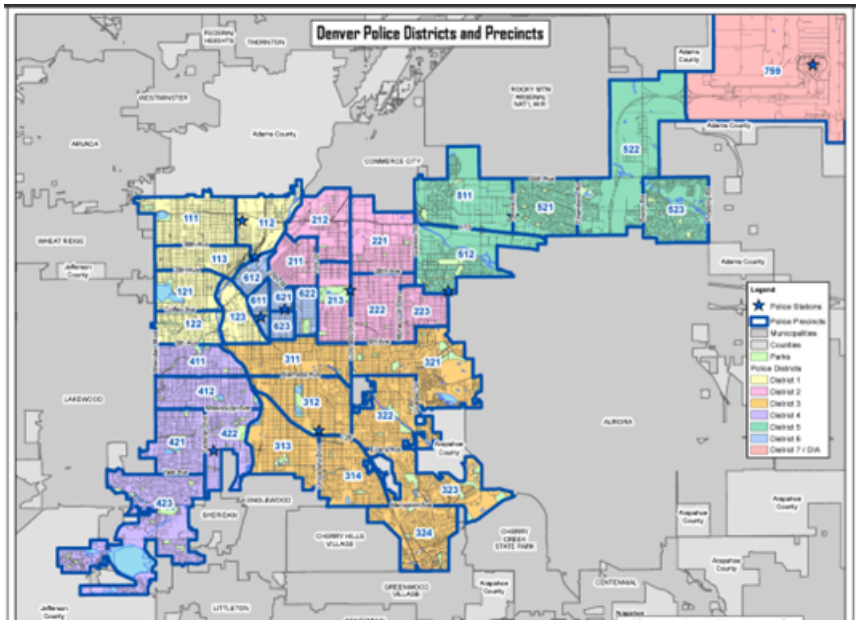


Figure 6: Denvor map

(a) Since the `precint_id` is derived from the `district_id`, we can use the first digit of `precint_id` to get the `district_id`.

(b) Since we have the geo-location of all the places and their corresponding districts, we can calculate a mean geo_location for all the districts, and for the rows that are missing the `district_id`, we can assign the data based on their geolocation to the nearest district.

2. **Replacing with Unknown:** Replace the district with Unknown.

We opted for the second approach, and we tried to replace all the values with the help of precinct ID; however, for any of the missing district IDs, the precinct ID was missing too, and therefore, it did not help us in replacing the value. Therefore, we were forced to use the calculating the mean approach, calculate the mean location of each district, and assign accordingly.

2.6 Handling Missing Values for `geo_location`

For this variable, there are three approaches:

1. **Dropping the Entire Column:** This data has important information, but since calculating the missing value for this data is pretty hard, it might be better to drop these data.

2. **Replacing with values:** We can use the district or precinct ID to fill the geo-location with an approximation of this data.

3. **Replacing with Unknown:** Replace the address with unknown.

For this feature, we used the second approach and assigned the geo-location of the district to any row with a missing geo-location.

2.7 Handling Missing Values for `neighborhood_id`

For this variable, there are two approaches:

1. **Dropping the Entire Column:** This data has important information, but since calculating the missing value for this data is pretty hard, it might be better to drop these data.

2. **Replacing with Unknown:** Replace the address with unknown.

We used the second approach for this feature and assigned any neighborhood with missing value with Unknown.

Now, after all these missing values handling there were some rows that still had missing values:

```
incident_id               0
offense_id                0
offense_code              0
offense_code_extension    0
offense_type_id           0
offense_category_id       0
first_occurrence_date     0
last_occurrence_date      0
reported_date             0
incident_address          0
geo_x                     8
geo_y                     8
geo_lon                   8
geo_lat                   8
district_id               0
precinct_id               0
neighborhood_id           0
is_crime                  0
is_traffic                0
victim_count              0
dtype: int64
```

Figure 7: Remaining missing values

We drop these rows for these data because there is not enough information to get what these data are.

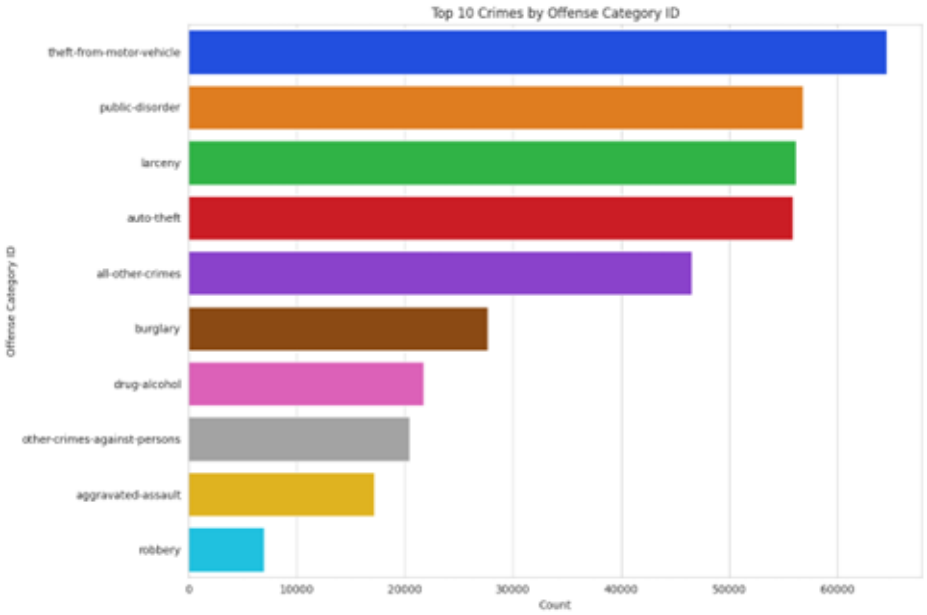


Figure 8: Most prevalent crimes

3 Data Visualization

Let's see what the most prevalent crimes in Denver are:

1. **Theft from Motor Vehicle:** Stealing items from parked cars.
2. **Public Disorder:** Offenses disrupting public peace, like public intoxication.
3. **Larceny:** Theft of personal property without force or breaking in.
4. **Auto Theft:** Stealing motor vehicles.
5. **All Other Crimes:** Various offenses not categorized separately.
6. **Burglary:** Illegally entering buildings to commit theft.
7. **Drug and Alcohol:** Crimes related to controlled substances or alcohol, like possession or DUI.
8. **Other Crimes Against Persons:** Offenses directly harming individuals, not classified elsewhere.
9. **Aggravated Assault:** Causing severe injury or using a deadly weapon during an assault.
10. **Robbery:** Taking items from others using force or threat.

3.1 Monthly Crime Analysis

The monthly crime trends in Denver exhibit a seasonal effect, with higher crime rates observed during the summer months and lower rates in the colder months. The peak in July and the sharp decrease in November are particularly noteworthy. While the data provides an overview of crime fluctuations, further investigation is needed to understand the underlying causes of these patterns. Factors such as weather conditions, economic trends, population movements, and law enforcement policies should be considered for a comprehensive analysis.

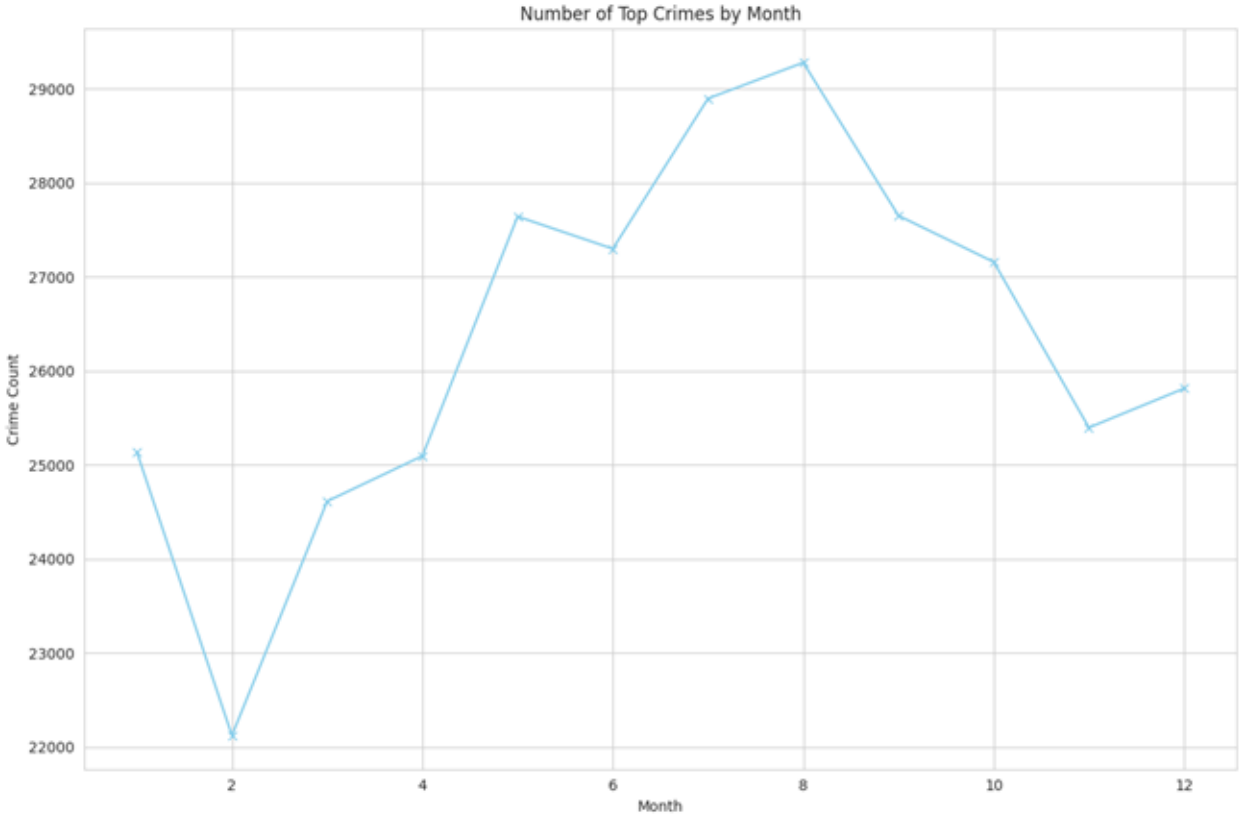


Figure 9: Monthly Crime

January (Month 1): The year begins with the lowest crime count, indicating a relatively calm period for the city in terms of top crimes reported.

February to March (Months 2 to 3): A sharp rise in crime rates suggests an escalation in criminal activity as the winter season progresses.

April (Month 4): There is a notable decrease in crime during this month. This could reflect the effectiveness of crime prevention strategies or other seasonal factors that deter criminal behavior.

May to July (Months 5 to 7): The crime rate experiences a steady and significant increase, peaking in July. This uptrend may correlate with the increase in outdoor activities during the warmer summer months, which can lead to higher instances of certain crimes.

August (Month 8): Following the peak, a slight decrease is observed. This reduction could be associated with the culmination of summer activities and the beginning of the school season.

September to October (Months 9 to 10): An increase during these months reverses the previous downward trend, which may be influenced by social and economic factors unique to this period.

November (Month 11): A dramatic fall in crime rates is seen in November. This drop could be attributed to enhanced law enforcement efforts or changes in public behavior as the weather cools.

December (Month 12): The year concludes with a minor increase in crime rates, potentially linked to the holiday season and its associated factors, such as increased retail activity and opportunities for theft-related crimes.

3.2 Yearly Crime Analysis

This report examines the annual crime data for the city of Denver over a five-year period from 2018 through 2022. The data is visualized through a bar graph indicating the number of top crimes for each year. The overall pattern indicates fluctuations in the crime rate, with a notable increase in the latter part of the observed period. The sharp rise in 2022 may reflect various socio-economic changes, policy alterations, or other factors not captured within the scope of this data.

3.3 Hourly Crime Analysis

The overall pattern suggests that crime is closely tied to the patterns of human activity in the city, with fewer incidents occurring in the very early morning hours and higher incidents correlating with times of high social and commercial activity.

Lowest Crime Period: The chart shows the lowest count of crimes occurs between 3 AM and 5 AM, with the absolute lowest point just before 5 AM. This could be attributed to fewer people being outside or active during these hours.

Rise in Crime During the Day: Starting from around 5 AM, there is a sharp increase in crime count, which continues to rise steadily until it peaks at 4 PM. This increase might correspond with the general increase in population activity as the day progresses, including both the opening hours of businesses and increased public presence.

Peak Crime Hours: The peak crime period is between 3 PM and 6 PM, with the highest point at 4 PM. This could be related to the conclusion of the workday, high traffic volumes, and increased opportunities for crimes such as theft or assault.

Evening Decrease: After the peak, there is a gradual decline in the number of crimes reported, which could be due to the end of business hours for many establishments and a decrease in pedes-

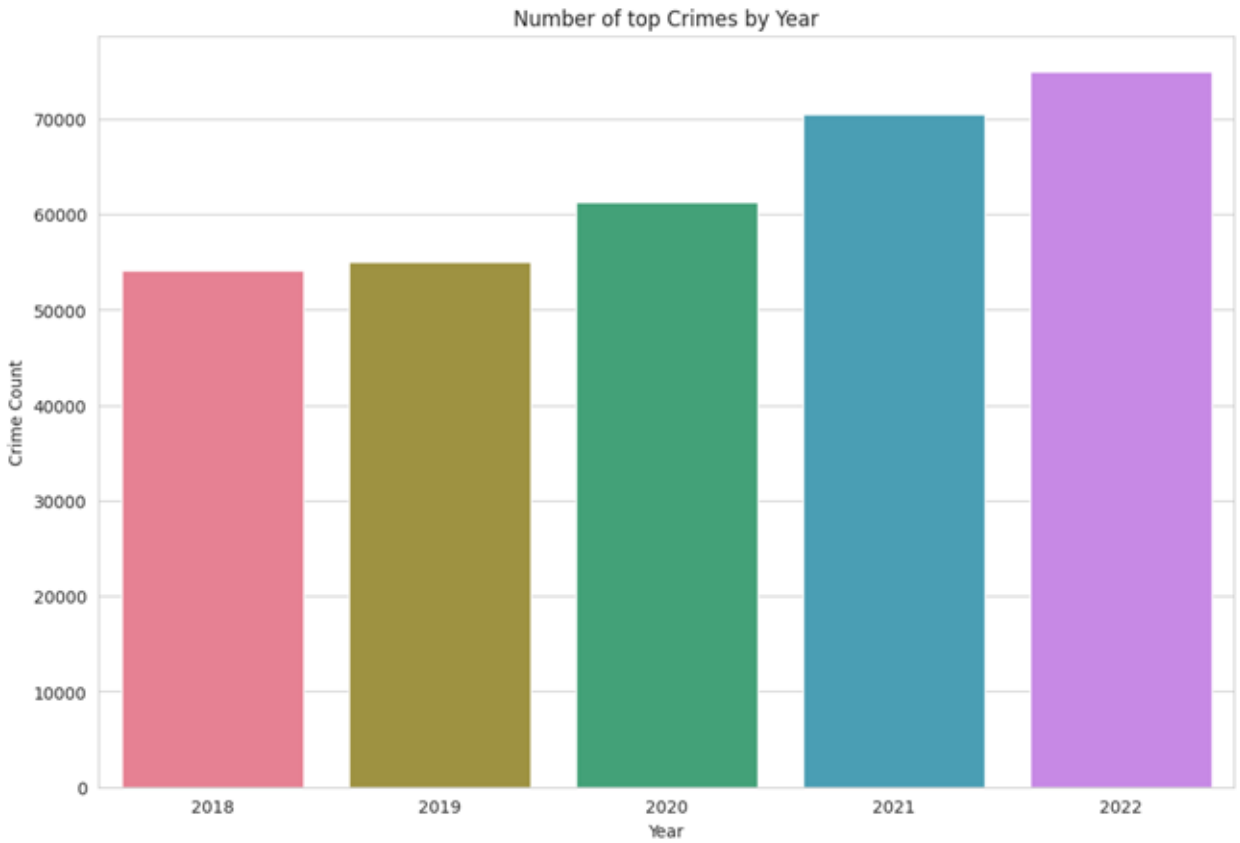


Figure 10: Yearly Crime

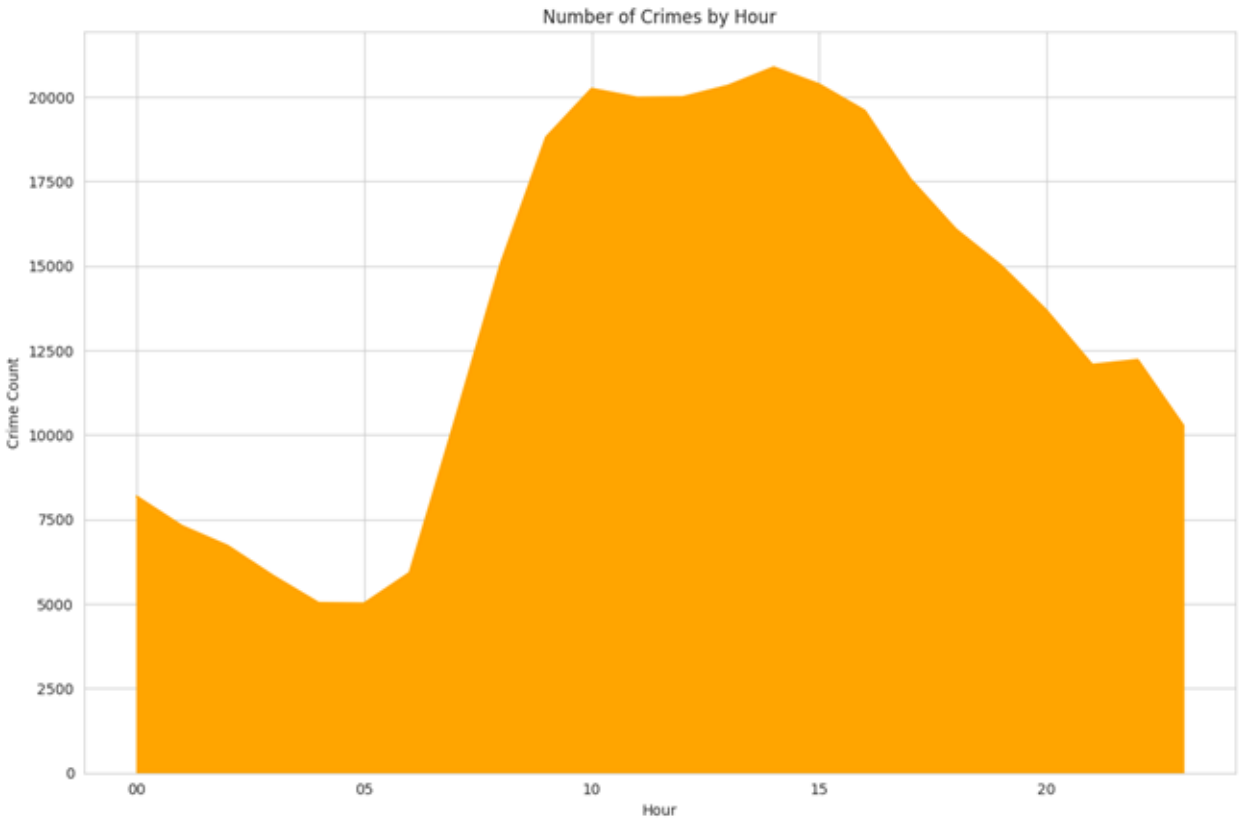


Figure 11: Hourly Crime

trian and vehicle traffic as people return to their homes.

Nighttime Activity: The chart shows a secondary, smaller peak around 10 PM, which might be associated with nightlife when there might be an increase in assaults or DUIs.

3.4 Worst Neighborhoods in Denver

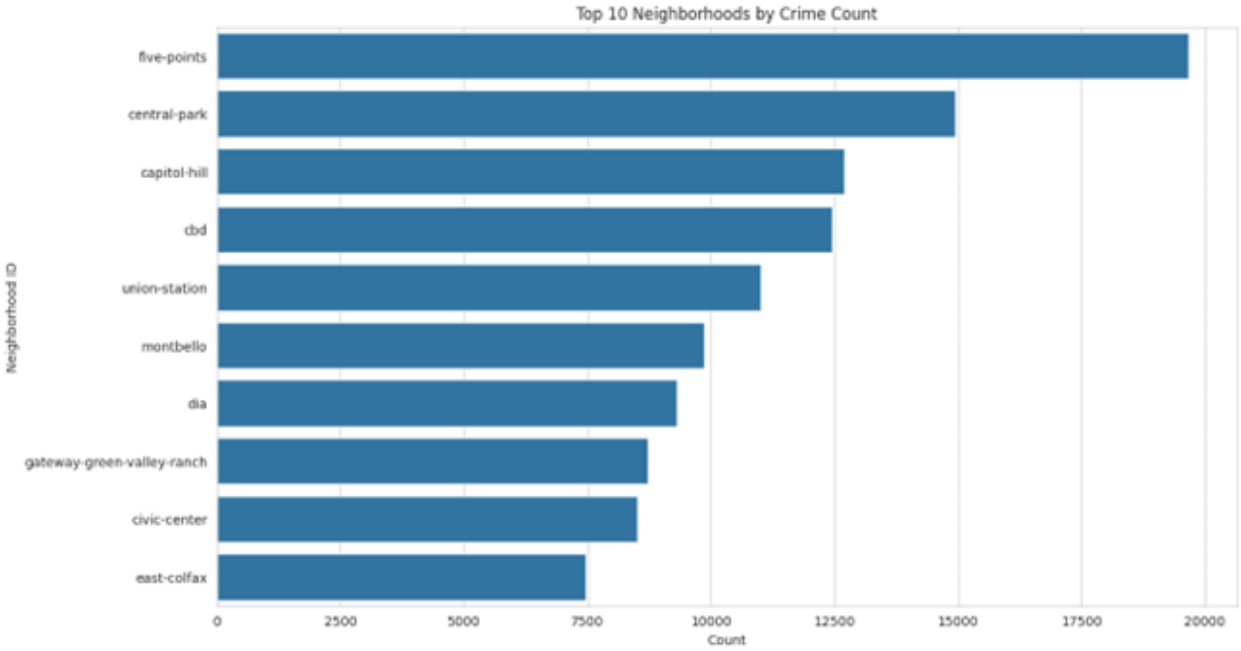


Figure 12: High Crime Neighborhoods

The data is visualized through a horizontal bar chart, which allows for a clear comparison of crime incidence across different areas.

High Crime Neighborhoods

Five Points: This neighborhood has the highest crime count, significantly more than any other neighborhood, with just under 20,000 reported incidents. This may indicate a need for increased law enforcement presence and community safety programs.

Mid-Level Crime Neighborhoods

Central Park, Capitol Hill, CBD, and Union Station: These neighborhoods exhibit a medium level of crime count, ranging from approximately 7,500 to 12,500 incidents. These areas may benefit from targeted crime prevention strategies.

Lower Crime Neighborhoods

Montbello, DIA (Denver International Airport), Gateway-Green Valley Ranch, Civic Center, and East Colfax: These neighborhoods have lower crime counts, with East Colfax showing the fewest incidents among the top 10, at around 5,000 crimes.

Crime distribution in Denver varies significantly across different neighborhoods. While Five Points stands out with the highest number of reported crimes, other neighborhoods show varying levels of crime incidence. A multifaceted approach tailored to each neighborhood's needs is essential for effective crime prevention and reduction.

3.5 Neighborhood Crime Distribution

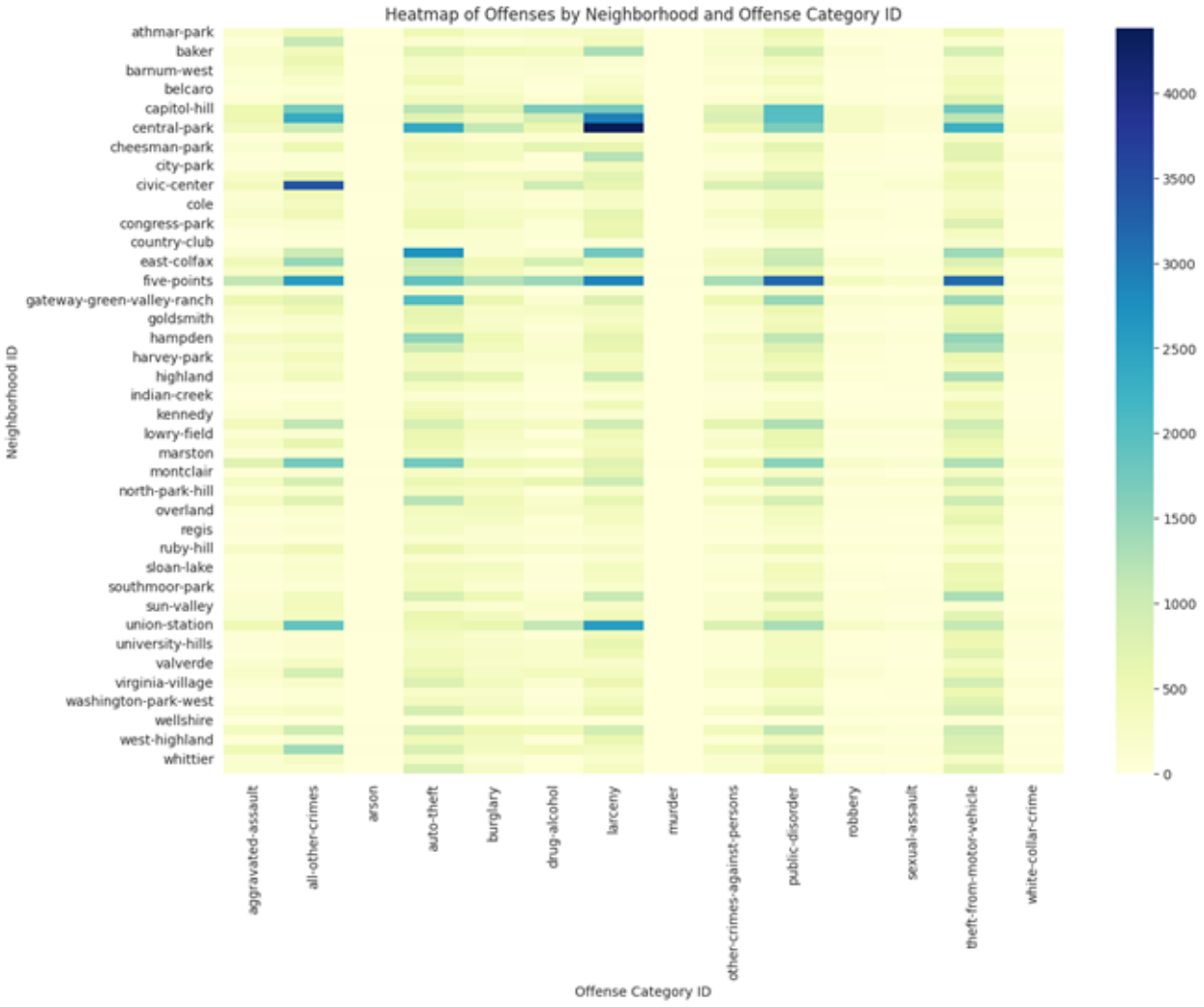


Figure 13: Neighborhood Crime Distribution

High Incidence Areas: Certain neighborhoods, such as Five Points and CBD (Central Business District), display a higher incidence across multiple categories of crime, especially in categories like public disorder and drug/alcohol offenses.

Targeted Crimes: Some neighborhoods show a concentration in specific types of crime. For example, Capitol Hill shows a notably higher incidence of public disorder offenses.

Crime Categories:

- **Public Disorder and Drug/Alcohol:** These categories are the most prominently represented across the majority of neighborhoods, suggesting they are the most common offenses city-wide.
- **Property Crimes:** Auto theft and burglary also appear to be common across many neighborhoods, with specific areas showing higher incidences, indicating potential targets for increased security and law enforcement measures.
- **Violent Crimes:** Aggravated assault shows up in certain neighborhoods, though less frequently than disorderly conduct or drug-related offenses. Homicides (murder) are relatively rare across the board but do appear in some neighborhoods.

3.6 Timing of Crime Reports by Offense Category

The chart presents a boxplot for each offense category, displaying the range and median of reported hours for crimes within that category.

Public Disorder: Reports for public disorder offenses are concentrated in the late-night to early-morning hours, with a median reporting time around 2 AM.

Drug/Alcohol Offenses: Similar to public disorder, drug and alcohol-related offenses tend to be reported predominantly during nighttime hours.

Sexual Assault: The reporting times for sexual assault are spread throughout the day but with a slight concentration in the early morning hours.

Other Crimes Against Persons: This category shows a wider range of reported hours, with incidents occurring fairly consistently throughout the day and night.

White Collar Crime: White-collar crimes have a narrow reporting distribution, primarily during standard business hours.

Murder: Reports of murder have outliers indicating atypical reporting times, but generally, the occurrences span across the day and night with a wide distribution.

Robbery: Robbery reports show a concentration in the late evening to early morning hours.

Aggravated Assault: This category has a wide range but a median reporting time in the evening.

Arson, Burglary, Larceny, Theft from Motor Vehicle, Auto Theft: Property crimes tend to have a broader distribution of reporting times, with medians generally in the afternoon or evening hours.

The timing of crime reports can be indicative of the nature of the crime and can inform policing strategies. For instance, the concentration of public disorder and drug/alcohol offenses at night might necessitate increased night patrols or after-hours social services.

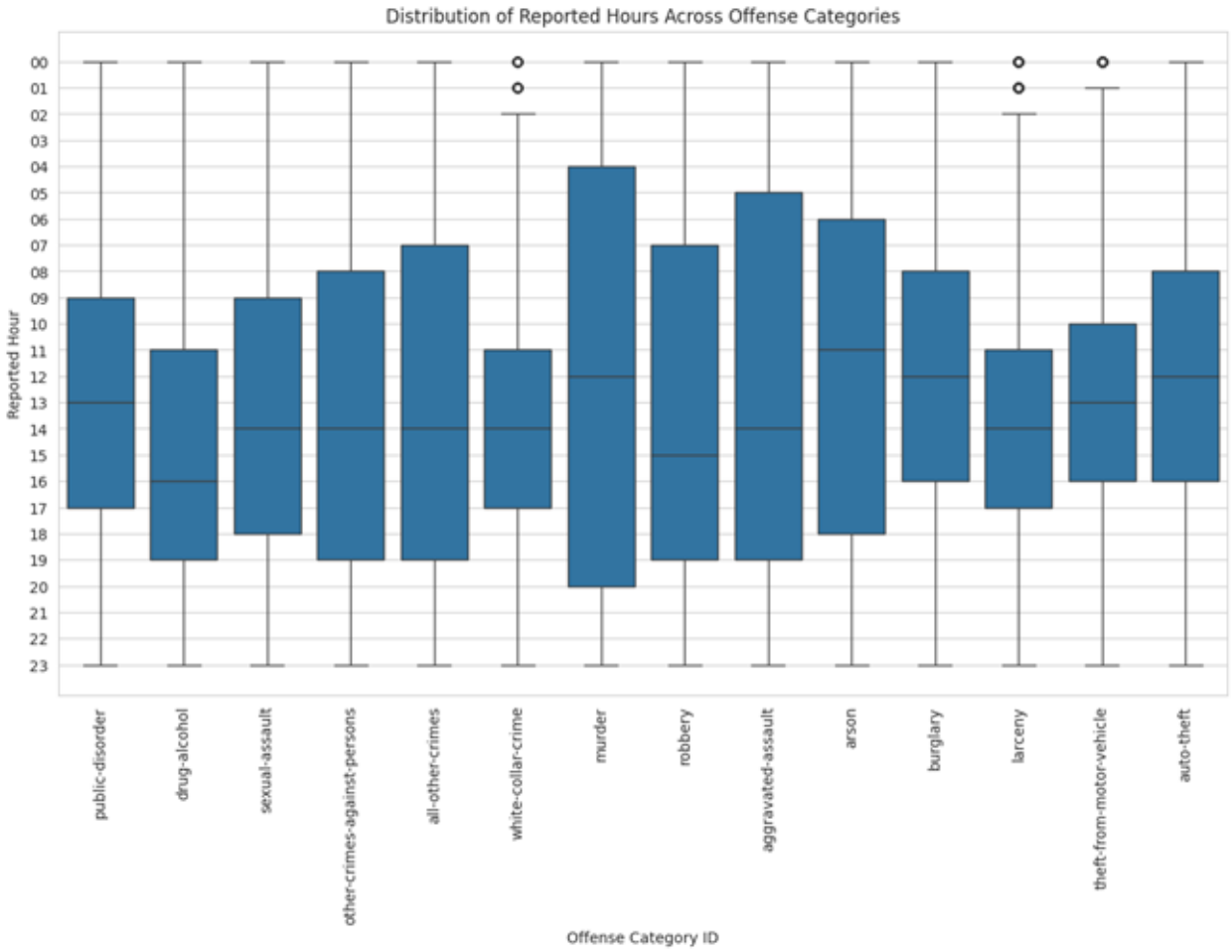


Figure 14: Timing of Crime Reports

3.7 Analysis of offenses by month

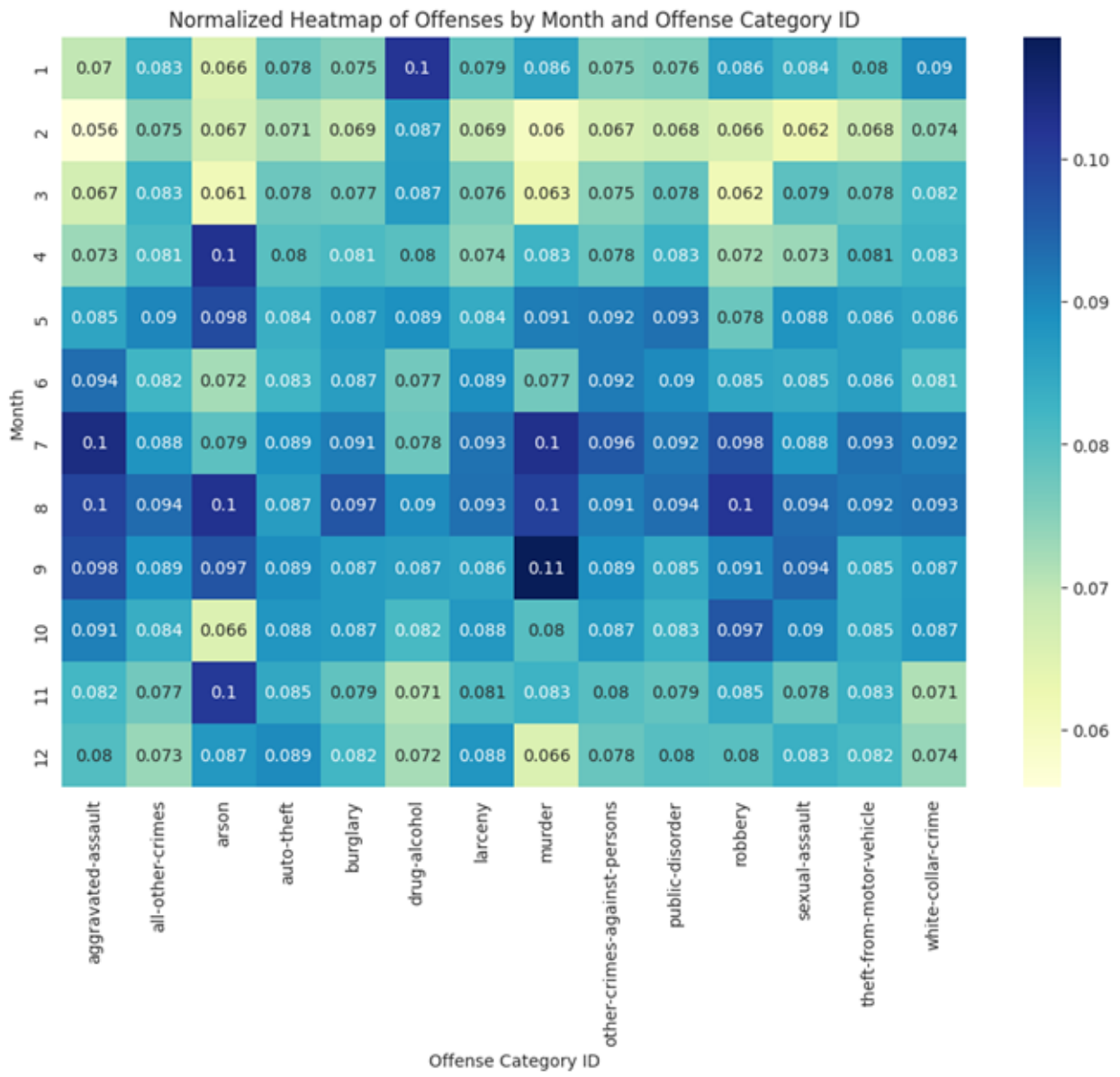


Figure 15: offenses by month

This image shows a normalized heatmap of offenses by month and offense category in Denver. The heatmap visualizes the frequency of crimes, providing an understanding of crime trends over different months of the year.

Monthly Crime Trends

Higher Frequency in Summer: Offenses such as aggravated assault, drug/alcohol-related crimes, and public disorder tend to peak in the summer months (June, July, and August).

Winter Decrease: There is a noticeable decrease in the frequency of most crime categories

during the winter months (November, December, and January).

Offense Categories

Drug/Alcohol Offenses: These offenses exhibit higher frequencies during the summer months, with the peak in September (0.11).

Theft-Related Crimes: Auto theft, burglary, and theft from motor vehicle offenses are relatively consistent throughout the year, with slight increases in the warmer months.

Violent Crimes: Aggravated assault and murder offenses show an increase in frequency during the summer months, aligning with general trends for violent crimes.

Seasonal Variation: The heatmap suggests a seasonal pattern in crime rates, with higher overall crime rates during the summer.

Consistent Offenses: White-collar crime remains relatively consistent throughout the year, likely due to its less spontaneous nature compared to other crime types.

3.8 Yearly Trends in Neighborhood Crime Rates

The provided heatmap visualizes normalized yearly incident counts for the top 10 neighborhoods in Denver from 2018 to 2022. This analysis seeks to identify trends and changes in incident rates over the five-year period.

Yearly Trends

- *Stability in Crime Rates:* Most neighborhoods show relative stability in crime rates over the five-year period, with minor fluctuations.
- *Decrease in Certain Areas:* Some neighborhoods like CBD and Union Station have seen a slight decrease in normalized incident counts over the years.

Neighborhood Comparison

- *Five Points and CBD:* These neighborhoods consistently show higher normalized incident counts compared to others, suggesting they have higher crime rates.
- *East Colfax and Gateway-Green Valley Ranch:* These neighborhoods typically have lower normalized incident counts, which could indicate lower crime rates.

Observations Over Time

- *2018 to 2019:* A slight increase in incident counts in neighborhoods such as Central Park and Five Points.

Figure 16: Yearly Trends in Neighborhood Crime Rates

- *2020 Stability:* Despite the pandemic, crime rates remained relatively stable in most neighborhoods compared to previous years.
- *2021 to 2022:* There is a notable decrease in some neighborhoods, such as CBD, which could be due to various factors, including changes in policing or community initiatives.

3.9 Analysis of Offense Category Trends

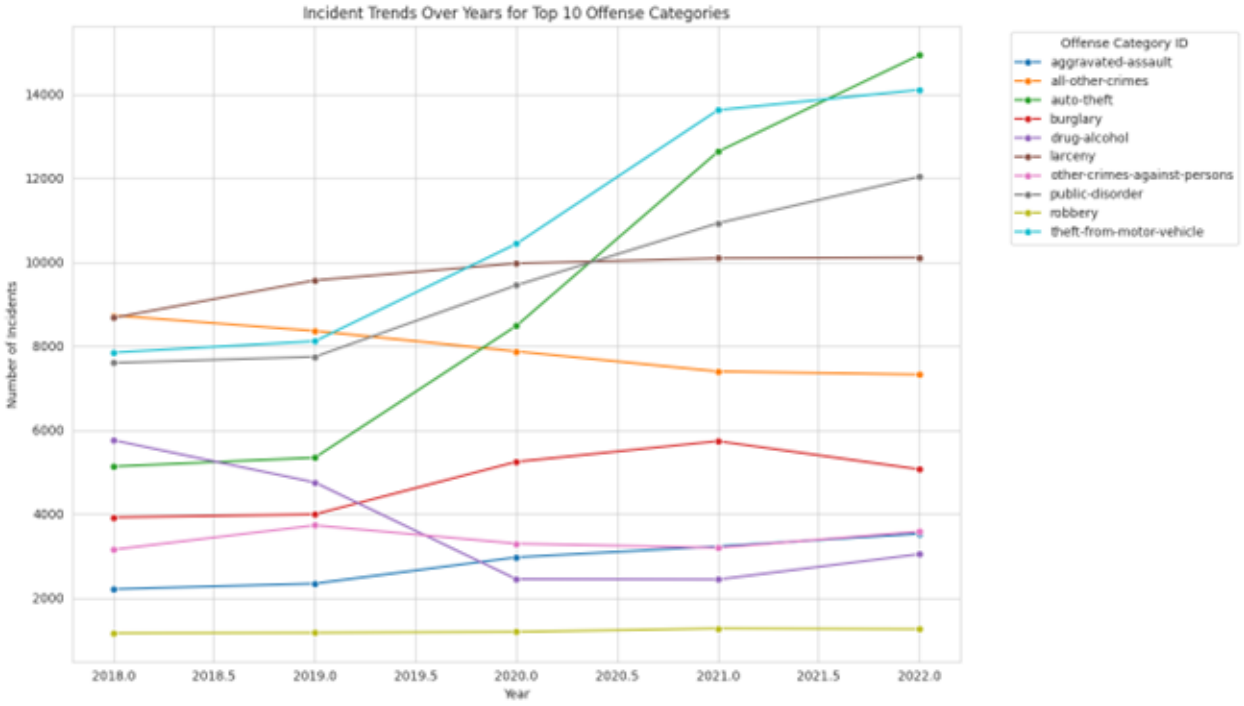


Figure 17: Analysis of Offense Category Trends

An analysis of the line graph provided reveals the trends in incident reports for the top 10 offense categories in Denver over a five-year period from 2018 to 2022.

General Trends

- *Increase in Auto-Theft:* There is a notable upward trend in incidents of auto-theft, with numbers significantly rising from 2018 to 2022.
- *Rise in Drug-Alcohol Related Crimes:* Incidents related to drug and alcohol use show a steady increase over the five-year span.
- *Stable or Decreasing Trends:* Categories such as burglary and larceny exhibit either a stable trend or a slight decrease in the number of incidents over the years.

Specific Offense Trends

- *Aggravated Assault:* This category has shown a fluctuating trend with a slight increase towards 2022.
- *All Other Crimes:* A category that possibly aggregates less common offenses shows a moderate increase over time.
- *Burglary:* The incidents of burglary show a general decline, particularly noticeable from 2020 to 2022.
- *Public Disorder:* There's a consistent upward trend in public disorder incidents, indicating a growing issue in this area.
- *Robbery and Theft from Motor Vehicle:* These categories show an initial increase followed by a plateau or slight decline in recent years.

The data suggests an overall increase in crime rates for certain categories, particularly auto-theft and drug-alcohol related offenses. This could be indicative of broader social issues, such as increased drug use or economic factors affecting property crime rates. The decline in burglary may reflect improved security measures or changes in policing strategy.

3.10 Analysis of Crime Distribution by District

The variation in crime rates across districts suggests differing levels of safety, law enforcement activity, or population density. Districts 3 and 6 may require additional resources and strategic planning to address the high incidence of crimes. District 7's notably low crime rate may be due to various factors such as effective community policing strategies, lower population density, or socioeconomic variables.

High Crime Districts

- *District 3:* This district shows the highest number of crimes, surpassing 60,000 incidents, indicating it is a hotspot for criminal activity within the city.
- *District 6:* Following closely, District 6 also exhibits a high crime count, with incidents approaching the 60,000 mark.

Moderate Crime Districts

- *Districts 1, 2, and 5:* These districts have a moderate number of crimes, with each reporting between 40,000 to just under 50,000 incidents.

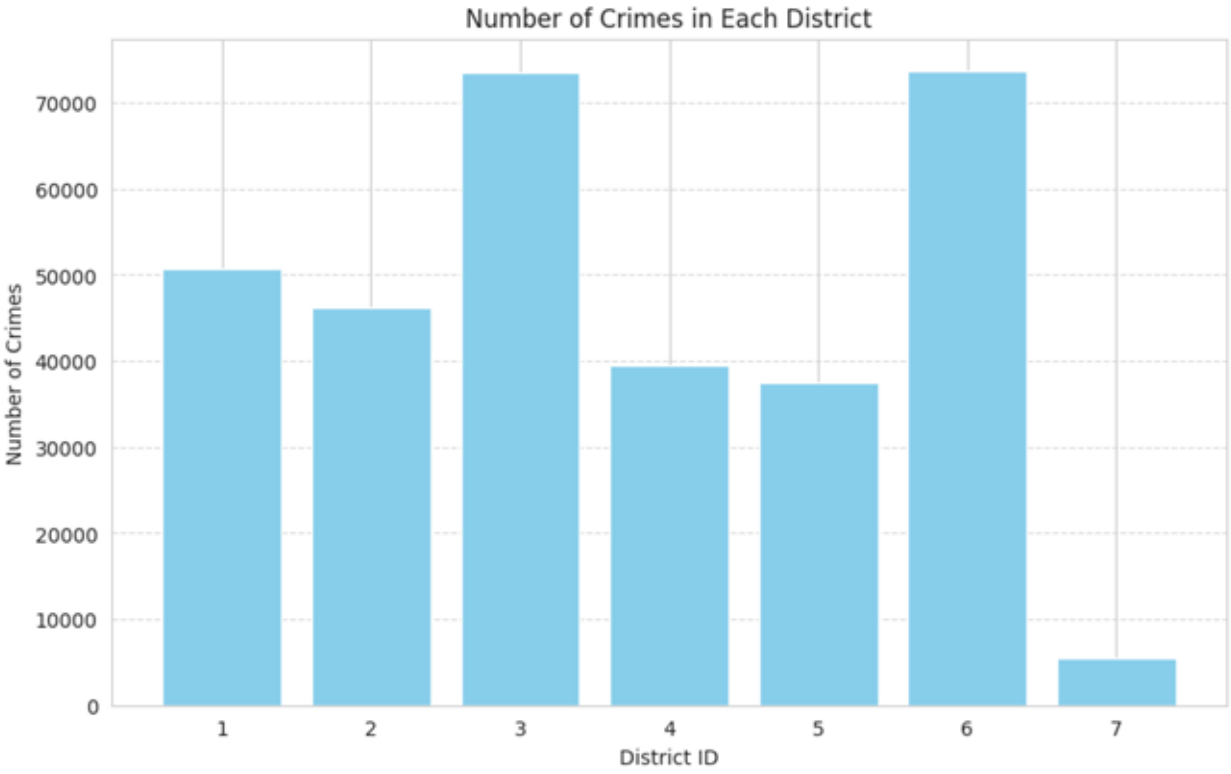


Figure 18: Crime Distribution by District

Low Crime Districts

- *District 4:* It is notable for having a substantially lower crime count compared to other districts, with incidents around the 30,000 mark.
- *District 7:* This district reports significantly fewer crimes, with the number of incidents being substantially lower than all other districts.

3.11 Analysis of Crime Distribution by Geographic Coordinates

This report analyzes a scatter plot depicting the distribution of the top 10 types of crimes in Denver, based on geographic coordinates (latitude and longitude).

Findings

Crime Hotspots

- *Concentration of Incidents:* The scatter plot indicates a high concentration of crimes in central areas, with less frequency on the outskirts. This is consistent with urban crime patterns where higher population density correlates with higher crime rates.

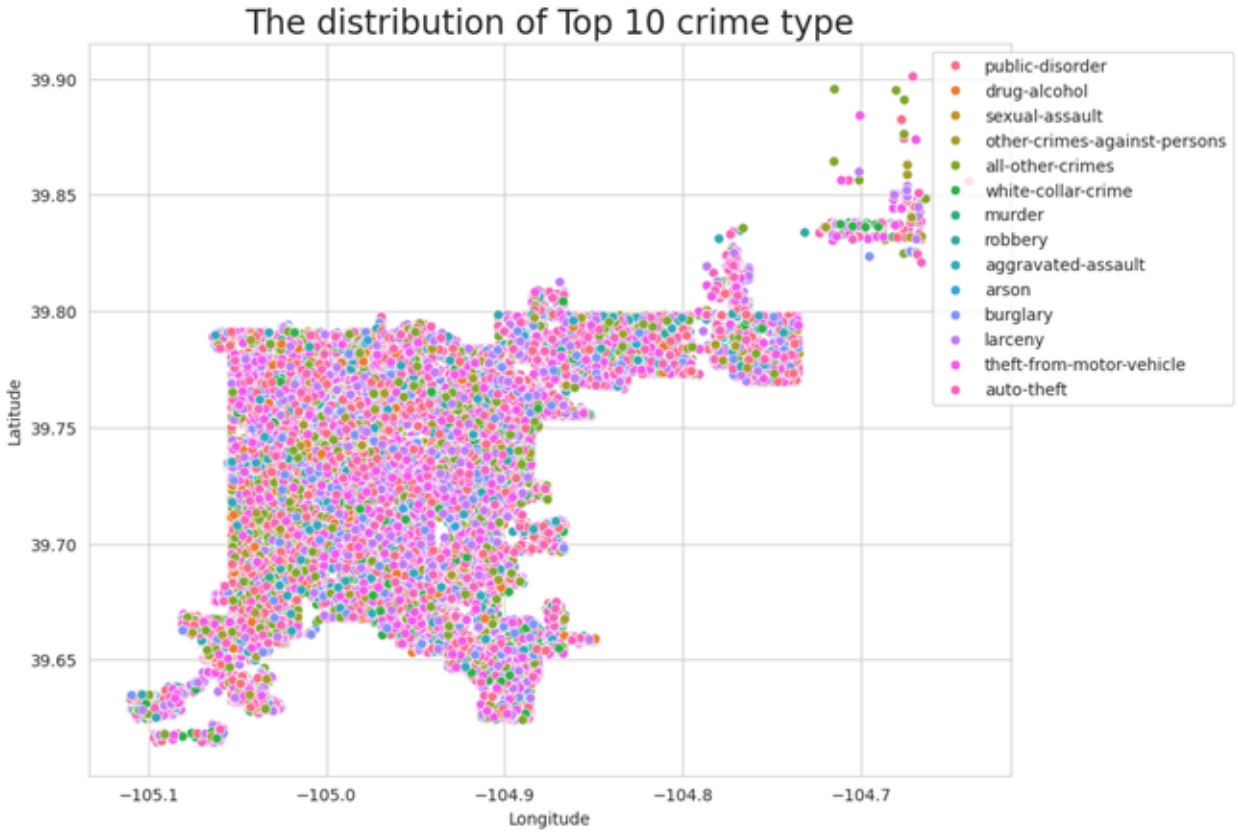


Figure 19: Crime Distribution by Geographic Coordinates

- *Diverse Crime Types:* There is a wide variety of crimes occurring throughout Denver, with no single type of crime dominating the overall distribution.

Specific Crime Observations

- *Public Disorder:* Represented by red dots, public disorder incidents are widely distributed across the city, suggesting it's a common issue throughout Denver.
- *Drug and Alcohol Offenses:* These incidents, shown in green, also appear widespread, indicating that substance-related issues are prevalent across multiple districts.
- *Property Crimes:* Theft-related crimes, including larceny (purple), theft from motor vehicles (light purple), and auto theft (pink), are prominent and well-distributed, pointing to a city-wide issue with property crimes.
- *Violent Crimes:* Aggravated assault (dark blue) and robbery (blue) are somewhat more concentrated in specific areas, which could be indicative of particular neighborhoods with higher rates of violent crime.

Analysis

The distribution of crimes across Denver is not uniform, with certain areas showing higher instances of specific crime types. The central regions show a greater mix of crime types, possibly due to higher commercial activity and nightlife. In contrast, the outskirts experience fewer incidents, which might be attributed to lower population density or different socio-economic factors.

Also, we can use this map on code to interact with location and crimes.

4 Method and result

In this phase, we formulate various hypotheses and conduct regressions on different variables. Subsequently, we meticulously analyze the outcomes to derive insights and conclusions.

4.1 Hypotheses Test

4.1.1 Hypotheses for Neighborhood Crime Distribution:

- *Null Hypothesis (H_0):* There is no association between neighborhood and category of offense.
- *Alternative Hypothesis (H_1):* There is a significant association between neighborhood and category of offense.

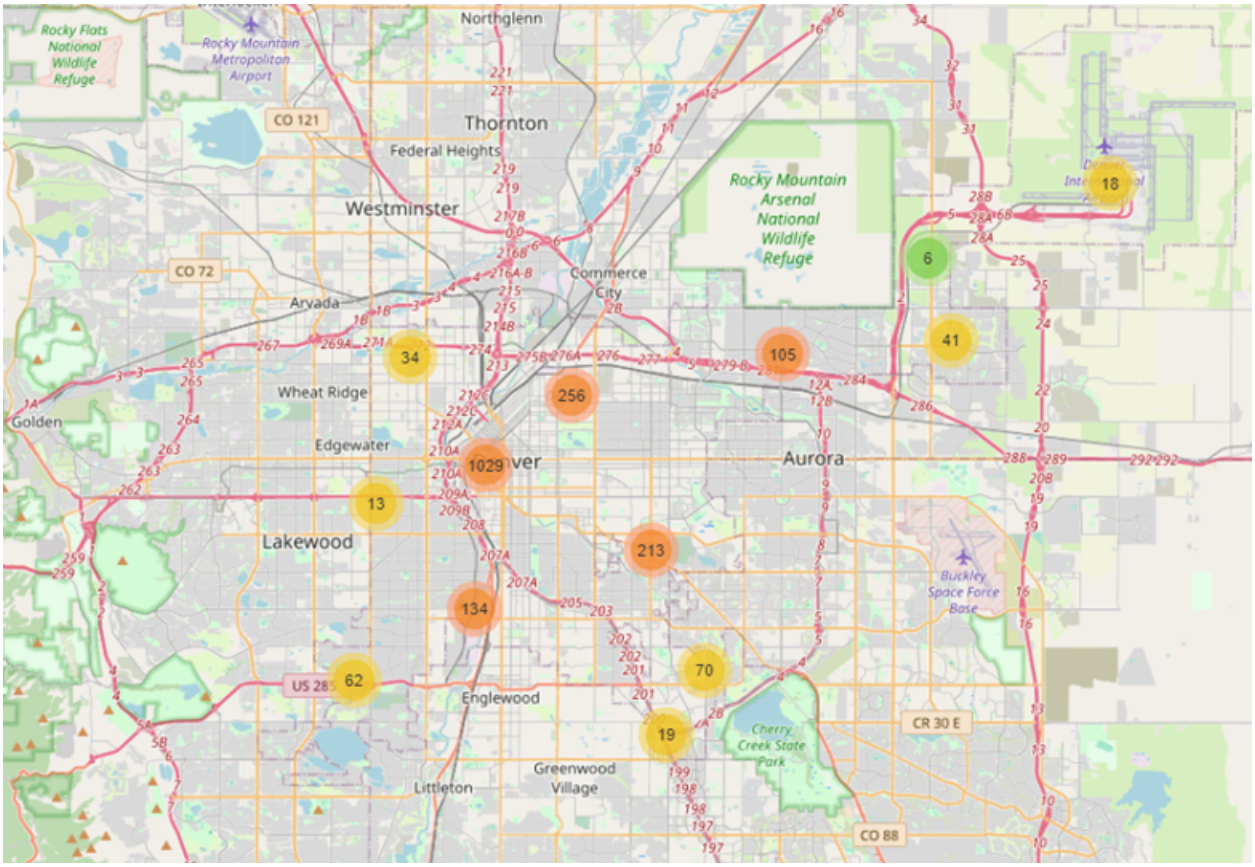


Figure 20: Interactive Map

2. Test and Rationale

Chi-square Test of Independence

- **What is the Test?** The Chi-square test of independence is a statistical test used to determine if there is a significant association between two categorical variables.
- **Why and When Can We Use This Test?** This test is appropriate when you have two categorical variables from a single population and want to determine whether there is a significant association between them. In this scenario:
  - Both neighborhood_id and offense_category_id are categorical variables.
  - The data in the contingency table represent frequencies of occurrences, suitable for a Chi-square test.
  - The test is non-parametric and does not assume a normal distribution of the data.
  - It's suitable for analyzing whether the distribution of incident categories varies by neighborhood, which could imply specific patterns or tendencies in crime or offenses in different areas.

3. Result

- *Chi-square Statistic:* 617.92
- *p-value:* 0.0111

The p-value of 0.0111 is less than the significance level of 0.05, indicating that we reject the null hypothesis. Therefore, there is a statistically significant association between neighborhood and category of offense. This result suggests that the distribution of offense categories is not uniform across neighborhoods, with certain types of offenses being more prevalent in specific areas. This insight can inform targeted law enforcement and community safety initiatives.

4.1.2 Hypotheses and Test for Geographical Correlation

- *Null Hypothesis (H_0):* There is no linear correlation between the geo_x and geo_y variables.
- *Alternative Hypothesis (H_1):* There is a linear correlation between the geo_x and geo_y variables.

2. Test and Rationale

Pearson's Correlation Coefficient

- **What is the Test?** Pearson's correlation coefficient (r) measures the strength and direction of a linear relationship between two continuous variables.
- **Why and When Can We Use This Test?** This test is used to examine the linear correlation between two continuous variables when the relationship is expected to be linear and the data is approximately normally distributed. In this scenario:
  - Both geo_x and geo_y are continuous variables representing geographical coordinates.
  - The test investigates if there's a linear relationship between these coordinates across incidents in the dataset.

3. Result

- *Pearson Correlation Coefficient:* 0.281
- *p-value:* 0.00465

The Pearson correlation coefficient of 0.281 suggests a low to moderate positive linear relationship between the geo_x and geo_y variables. The p-value of 0.00465 indicates statistically significant evidence of this linear relationship, albeit weak. These results imply that while there is a statistically significant positive association between the geographical coordinates of incidents, the relationship is not strong. Other factors or variables may also influence the geographical patterns observed in the dataset.

4.1.3 Hypotheses and Test for District-wise Victim Counts

1. Hypotheses

- *Null Hypothesis (H_0):* There is no significant difference in the mean victim counts across different districts.
- *Alternative Hypothesis (H_1):* There is a significant difference in the mean victim counts across different districts.

2. Test and Rationale
One-way ANOVA

- **What is the Test?** One-way ANOVA compares the means of three or more independent groups to determine if at least one group mean is significantly different from the others.
- **Why and When Can We Use This Test?** This test is used when:

- You have a categorical independent variable (district_id) with three or more levels/groups.
- The dependent variable (victim_count) is continuous.
- The groups are independent of each other, and the dependent variable follows a normal distribution within groups or has large enough sample sizes for the Central Limit Theorem to apply.
- Homogeneity of variances is observed among the groups.

3. Result

- *F-statistic:* 0.720
- *p-value:* 0.634

The F-statistic measures the ratio of variance between the group means to variance within the groups. With a p-value of 0.634, failing to reject the null hypothesis, there is no statistically significant difference in the mean victim counts across different districts. This suggests that variations in victim counts between districts may be due to random chance rather than a specific effect of the district itself. These results imply that other factors may be more influential in determining variations in victim counts across different areas.

4.1.4 Hypotheses and Test for Neighborhood-wise Victim Counts

1. Hypotheses

- *Null Hypothesis (H_0):* There is no significant difference in the mean victim counts across different neighborhoods.
- *Alternative Hypothesis (H_1):* There is a significant difference in the mean victim counts across different neighborhoods.

2. Test and Rationale
One-way ANOVA

- **What is the Test?** One-way ANOVA compares the means of three or more independent groups to determine if there is a statistically significant difference among them.
- **Why and When Can We Use This Test?** This test is used when:
  - The independent variable is categorical with multiple levels or groups (different neighborhoods).

- The dependent variable is continuous.
- The observations are independent, and the distribution of the dependent variable is approximately normal within each group or has large enough sample sizes for the Central Limit Theorem to apply.
- Homogeneity of variances is observed among the groups.

3. Result

- *F-statistic:* 0.447
- *p-value:* 0.998

The F-statistic measures the ratio of variance between the group means to the variance within the groups. With a p-value of 0.998, failing to reject the null hypothesis, there is no statistically significant difference in the mean victim counts across different neighborhoods. This suggests that any observed differences in victim counts among neighborhoods are likely due to random chance rather than a systematic difference attributable to the neighborhood itself. These results imply that other factors not considered in this analysis may be responsible for variations in crime impact across different areas.

4.1.5 Hypotheses and Test for Victim Counts Across Offense Categories

1. Hypotheses

- *Null Hypothesis (H_0):* There is no significant difference in the mean victim counts across different offense categories.
- *Alternative Hypothesis (H_1):* There is a significant difference in the mean victim counts across different offense categories.

2. Test and Rationale
One-way ANOVA

- **What is the Test?** One-way ANOVA compares the means of three or more independent groups to determine if there is a statistically significant difference among them.
- **Why and When Can We Use This Test?** This test is used when:
  - The independent variable is categorical with multiple levels or groups (different offense categories).

- The dependent variable is continuous.
- The observations are independent, and the distribution of the dependent variable is approximately normal within each group or has large enough sample sizes for the Central Limit Theorem to apply.
- Homogeneity of variances is observed among the groups.

3. Result

- *F-statistic:* 6.653
- *p-value:* 1.43×10^{-11}

The F-statistic measures the ratio of variance between the group means to the variance within the groups. With a p-value much lower than 0.05, rejecting the null hypothesis, there is significant evidence that the mean victim count is different across offense categories. This suggests that certain types of offenses are consistently associated with higher or lower numbers of victims than others. These findings can inform law enforcement and public safety organizations in allocating resources effectively and tailoring prevention strategies to the nature of the offenses.

4.1.6 Hypotheses and Test for Reported Hour of Crimes

1. Hypotheses

- *Null Hypothesis (H_0):* There is no significant difference in the mean reported hour of crimes between weekdays and weekends.
- *Alternative Hypothesis (H_1):* There is a significant difference in the mean reported hour of crimes between weekdays and weekends.

2. Test and Rationale

Independent Samples t-test

- **What is the Test?** The independent samples t-test compares the means of two independent groups to determine if there is a statistically significant difference between them.
- **Why and When Can We Use This Test?** This test is used when:
  - You have two independent groups.
  - The dependent variable is continuous.

- The data is approximately normally distributed, or the sample sizes are large enough for the Central Limit Theorem to apply.
- The variances between the groups are approximately equal.

3. Result

- *T-Statistic:* 0.230
- *p-value:* 0.818

The T-statistic measures the difference between the mean reported hour of crimes on weekdays and weekends relative to the variability of the data. With a p-value much higher than 0.05, we fail to reject the null hypothesis, suggesting that there is no significant evidence that the mean reported hour of crimes differs between weekdays and weekends. This implies that, based on the reported hour, there doesn't appear to be a significant difference in the timing of crime reports between weekdays and weekends, indicating relatively consistent patterns throughout the week.

4.1.7 Hypotheses and Test for Seasonal Variation in Crime Reports

1. Hypotheses

- *Null Hypothesis (H_0):* There is no significant difference in the mean number of reported crimes between summer months and winter months.
- *Alternative Hypothesis (H_1):* There is a significant difference in the mean number of reported crimes between summer months and winter months.

2. Test and Rationale

Independent Samples t-test

- **What is the Test?** The independent samples t-test compares the means of two independent groups to determine if there is a statistically significant difference between them.
- **Why and When Can We Use This Test?** This test is used when:
  - You have two independent groups.
  - The dependent variable is continuous.
  - The data should ideally follow a normal distribution, although the t-test is robust to normality violations, especially with larger sample sizes.

- Equal variances between the two groups are assumed, although variations of the t-test exist that do not require this assumption.

3. Result

- *T-Statistic:* -0.615
- *p-value:* 0.539

The T-statistic measures the difference in the mean number of reported crimes between summer and winter, in terms of the standard error of the difference. With a p-value greater than 0.05, we fail to reject the null hypothesis. This suggests that there is no significant evidence that the mean number of reported crimes differs between summer months and winter months. This implies that, at least in terms of the raw number of reported crimes, seasonal changes from summer to winter do not have a statistically significant impact on crime rates. Other factors might play a more significant role in influencing crime patterns.

4.1.8 Hypotheses and Test for Variations in Reported Crime Hours by Neighborhood

1. Hypotheses

- *Null Hypothesis (H_0):* There is no significant difference in the mean reported hour of crimes across neighborhoods.
- *Alternative Hypothesis (H_1):* There is a significant difference in the mean reported hour of crimes across neighborhoods.

2. Test and Rationale

One-way ANOVA (Analysis of Variance)

- **What is the Test?** The one-way ANOVA test determines if there are statistically significant differences between the means of three or more independent groups.
- **Why and When Can We Use This Test?** This test is appropriate when:
  - The independent variable is categorical with multiple levels or groups.
  - The dependent variable is continuous.
  - Observations across groups are independent.
  - The dependent variable should be approximately normally distributed within each group, or the sample sizes should be large enough for the Central Limit Theorem to apply.

- Homogeneity of variances among the groups is assumed.

3. Result

- *F-Statistic:* 1.037
- *p-value:* 0.397

The F-statistic measures the ratio of the variance between the group means to the variance within the groups. With a p-value greater than 0.05, we fail to reject the null hypothesis. This suggests that there is no significant evidence that the mean reported hour of crimes differs across neighborhoods. This finding implies that the timing of crime reports is relatively consistent across different areas, indicating that neighborhood-specific factors do not significantly influence the timing of crime reporting.

4.1.9 Hypotheses and Test for Geographical Differences in Reported Crime Latitudes

1. Hypotheses

- *Null Hypothesis (H_0):* There is no significant difference in the mean latitudes of reported crimes between auto thefts and white-collar crimes.
- *Alternative Hypothesis (H_1):* There is a significant difference in the mean latitudes of reported crimes between auto thefts and white-collar crimes.

2. Test and Rationale

Independent Samples t-test (Welch's t-test)

- **What is the Test?** The independent samples t-test compares the means of two groups to determine if there is a statistically significant difference between them.
- **Why and When Can We Use This Test?** This test is appropriate when:
  - The two groups being compared are independent.
  - The dependent variable is continuous.
  - The data does not necessarily need to have equal variances between groups, especially when specifying `equal_var=False` for Welch's t-test.

3. Result

- *T-Statistic:* -0.561

- *p-value:* 0.620

The T-statistic indicates the direction and magnitude of the difference between the mean latitudes of the two crime types. With a p-value greater than 0.05, we fail to reject the null hypothesis, suggesting that there is no significant evidence to suggest a difference in the latitudinal locations of auto thefts versus white-collar crimes. This implies that, at least in terms of latitude, these crimes do not show a significant spatial preference or pattern that distinguishes one from the other.

4.1.10 Hypotheses and Tests for Offense Types and Victim Counts

1. Hypotheses

- *Null Hypothesis (H_0) for ANOVA:* There is no significant difference in the mean victim count per crime across offense types.
- *Alternative Hypothesis (H_1) for ANOVA:* There is a significant difference in the mean victim count per crime across offense types.

2. Tests and Rationale

One-way ANOVA Test

- **What is the Test?** One-way ANOVA checks for significant differences between the means of three or more independent groups.
- **Why and When Can We Use This Test?** This test is appropriate when:
  - The independent variable is categorical with multiple levels or groups.
  - The dependent variable is continuous.
  - Observations are independent.
  - The data ideally meets assumptions of normality and homogeneity of variances.

Tukey's HSD Post-hoc Test

- **What is the Test?** Tukey's HSD test identifies which pairs of groups have significant differences in their means after a significant result in ANOVA.
- **Why and When Can We Use This Test?** This test is appropriate when multiple comparisons are needed without inflating the type I error rate associated with performing many t-tests.

3. Result

- *F-Statistic:* 2.275
- *p-value:* 0.0022

The F-statistic indicates a significant variance among the group means compared to within the groups. With a p-value below 0.05, we reject the null hypothesis, suggesting significant evidence that the mean victim count per crime is different across offense types.

Following this, Tukey's HSD test would provide specific pairwise comparisons to identify which offense types have significantly different mean victim counts. This further analysis enables more targeted interventions or resource allocation based on the types of crimes with differing victim counts.

4.2 Regression

4.2.1 Relationship between Year and Crime Count

Objective: To examine the relationship between the year (normalized) and the count of crimes reported.

Results:

- The regression model has an R-squared of 0.756, indicating that approximately 75.6% of the variability in the crime count is explained by the year.
- The adjusted R-squared is 0.674, which takes into account the number of predictors in the model and the number of observations, still indicating a strong fit.
- The F-statistic is 9.281 with a p-value of 0.056, which is marginally above the conventional alpha level of 0.05, suggesting that the model is on the borderline of being statistically significant.
- The regression coefficient for the constant term is 15.4000 with a standard error of 1.849, and it's statistically significant (p ¡ 0.05). This value represents the estimated number of crimes when the year is at its minimum value (after normalization).
- The regression coefficient for the year (X1) is 9.2000 with a standard error of 3.020. The t-statistic for this coefficient is 3.046, and its p-value is 0.056, which is marginally above the conventional alpha level of 0.05, suggesting that there is a positive, but not statistically significant at the 0.05 level, relationship between year and crime count.

OLS Regression Results

Dep. Variable:	is_crime	R-squared:	0.756
Model:	OLS	Adj. R-squared:	0.674
Method:	Least Squares	F-statistic:	9.281
Date:	Wed, 14 Feb 2024	Prob (F-statistic):	0.0556
Time:	10:46:35	Log-Likelihood:	-10.169
No. Observations:	5	AIC:	24.34
Df Residuals:	3	BIC:	23.56
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	15.4000	1.849	8.327	0.004	9.515	21.285
x1	9.2000	3.020	3.046	0.056	-0.411	18.811

Omnibus:	nan	Durbin-Watson:	3.132
Prob(Omnibus):	nan	Jarque-Bera (JB):	0.614
Skew:	-0.798	Prob(JB):	0.736
Kurtosis:	2.368	Cond. No.	3.61

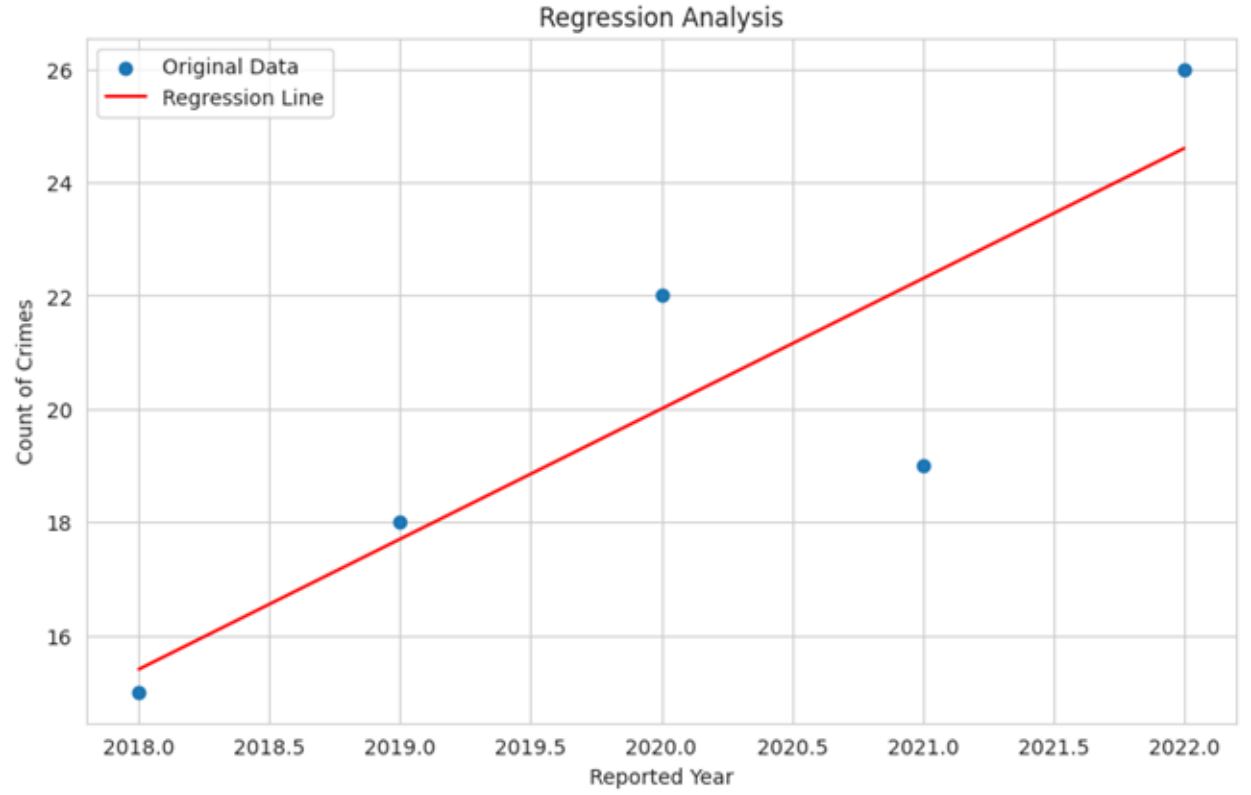


Figure 21: Relationship between Year and Crime Count

- The model exhibits a Durbin-Watson statistic of 3.132, suggesting that there is no autocorrelation in the residuals.

Diagnostics:

- The skewness of the residuals is -0.798, indicating a slight asymmetry in the distribution of residuals.
- The kurtosis is 2.368, which is relatively close to the normal distribution's kurtosis of 3, indicating that the tails of the residual distribution are neither particularly heavy nor light.
- The Jarque-Bera test has a p-value of 0.736, suggesting that the residuals are normally distributed.

Conclusions:

- The model indicates a positive trend in the number of crimes reported over the years, though the trend is not statistically significant at the 0.05 level.
- The high R-squared value suggests that year is a strong predictor for the number of crimes reported.
- Given the small sample size (5 observations), caution should be exercised when interpreting the results. A larger dataset might provide more definitive insights.
- Further analysis with additional data and possibly incorporating more variables could enhance the understanding of the factors influencing crime rates over the years.

4.2.2 Relationship between Normalized Year and Total Victim Count

Results:

- The model has an R-squared of 0.735, suggesting that about 73.5% of the variance in the victim count is accounted for by the year.
- The adjusted R-squared is 0.646, reflecting the model's explanatory power after adjusting for the number of predictors.
- The F-statistic is 8.308 with a p-value of 0.0634, indicating that the model is close to being statistically significant at the conventional 0.05 alpha level.
- The coefficient for the constant is 14.6000 with a standard error of 2.634, which is statistically significant ($p = 0.012$). This indicates the expected victim count when the year is at its minimum normalized value.

OLS Regression Results

Dep. Variable:	is_crime	R-squared:	0.756
Model:	OLS	Adj. R-squared:	0.674
Method:	Least Squares	F-statistic:	9.281
Date:	Wed, 14 Feb 2024	Prob (F-statistic):	0.0556
Time:	10:46:35	Log-Likelihood:	-10.169
No. Observations:	5	AIC:	24.34
Df Residuals:	3	BIC:	23.56
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	15.4000	1.849	8.327	0.004	9.515	21.285
x1	9.2000	3.020	3.046	0.056	-0.411	18.811

Omnibus:	nan	Durbin-Watson:	3.132
Prob(Omnibus):	nan	Jarque-Bera (JB):	0.614
Skew:	-0.798	Prob(JB):	0.736
Kurtosis:	2.368	Cond. No.	3.61

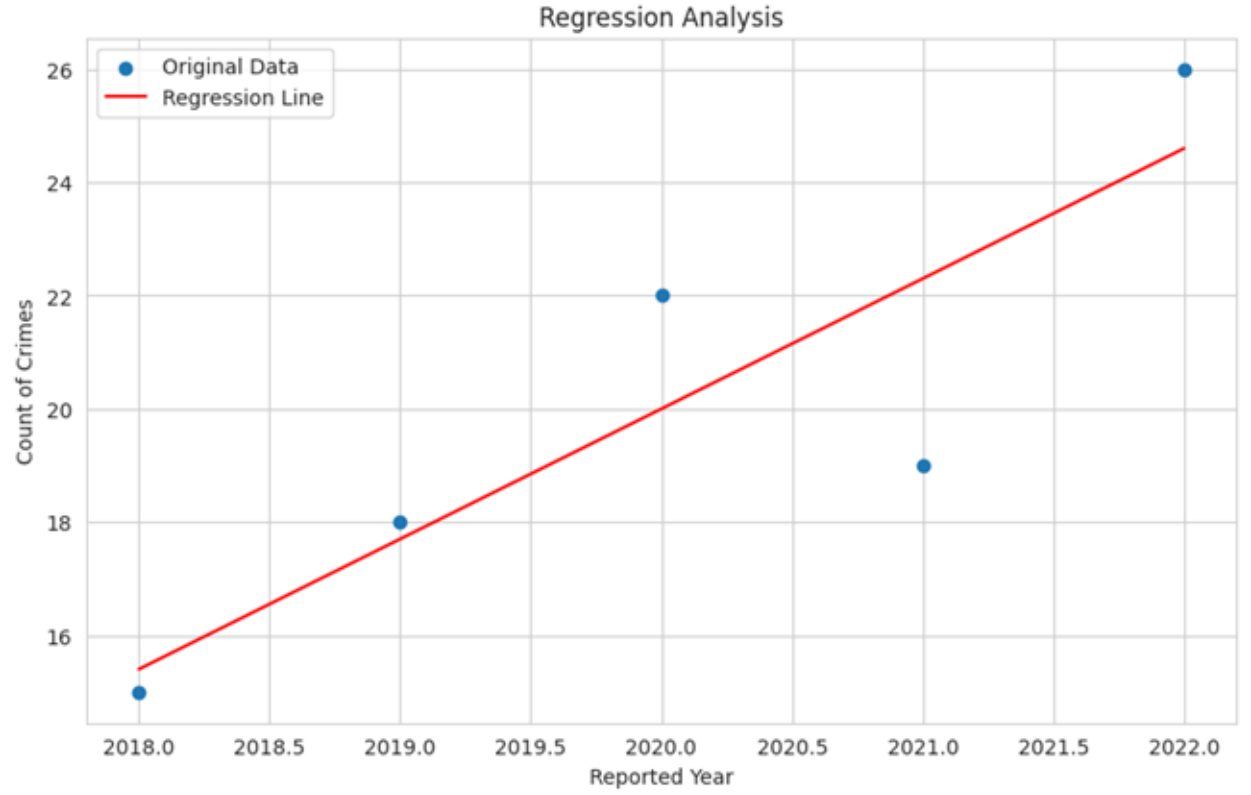


Figure 22: Relationship between Normalized Year and Total Victim Count

- The coefficient for the year (X_1) is 12.4000 with a standard error of 4.302. The t-statistic for this coefficient is 2.882, and the p-value is 0.063, which means the relationship between the year and victim count is positive but not statistically significant at the 0.05 level.
- The Durbin-Watson statistic is 2.895, suggesting that there is no autocorrelation in the residuals of the model.

Diagnostics:

- The skewness of the residuals is -0.972, indicating a moderate negative skew.
- Kurtosis is 2.739, suggesting that the distribution of residuals is slightly less peaked than a normal distribution.
- The Jarque-Bera test yields a p-value of 0.670, indicating that the residuals could be normally distributed.

Conclusions:

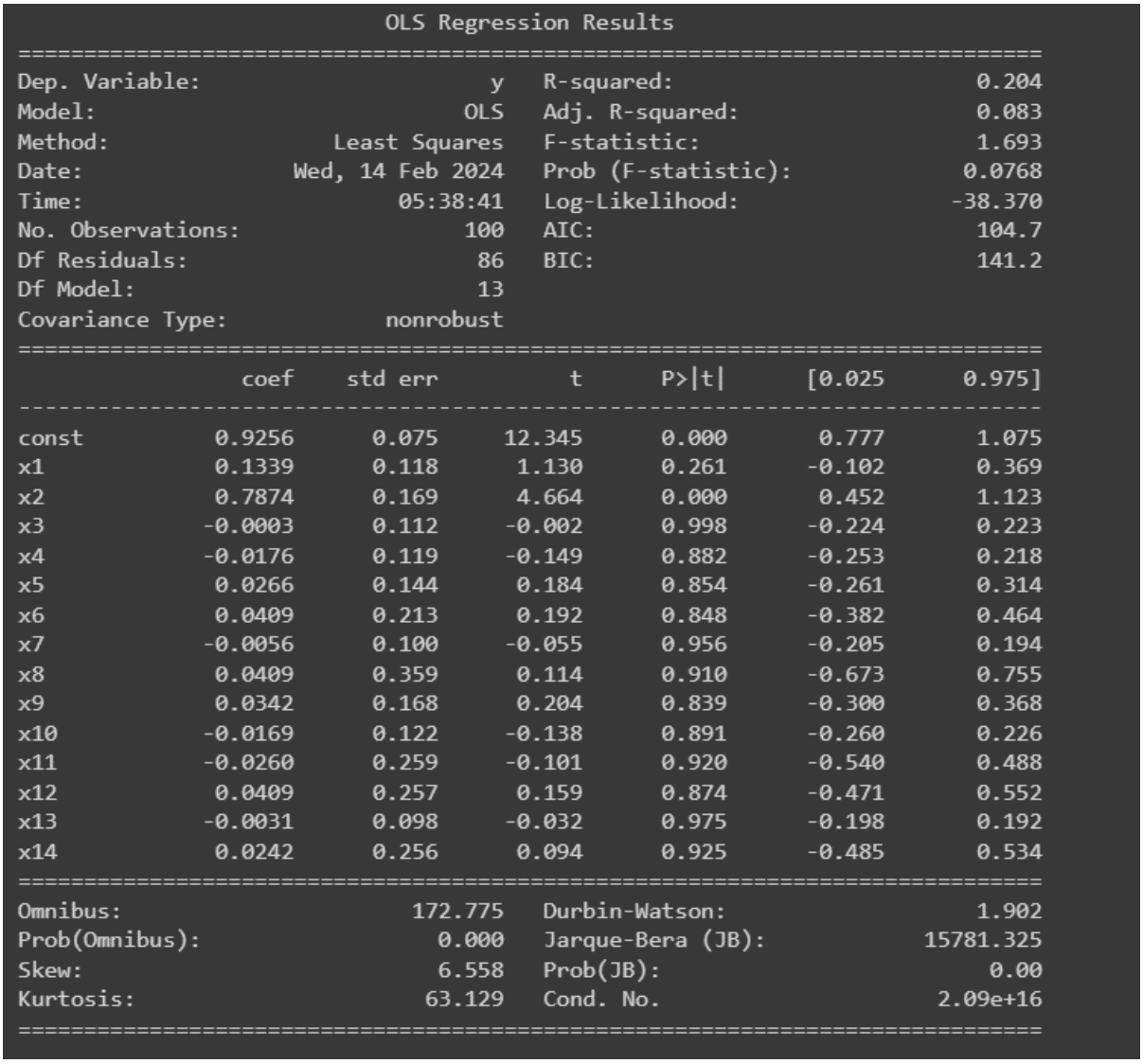
- There is a positive association between the year and the total victim count, although this relationship is not statistically significant at the 0.05 level.
- The R-squared value indicates that the normalized year is a strong predictor of the victim count.
- Similar to the previous model on crime counts, the small sample size is a limitation and may affect the power of the statistical tests.

4.2.3 Impact of Reported Year and Offense Categories on Victim Count

Regression Results:

- The R-squared of the model is 0.204, suggesting that approximately 20.4% of the variation in the victim count is explained by the combined effect of the reported year and offense categories.
- The adjusted R-squared is 0.083, significantly lower than the R-squared, which may indicate that some of the independent variables do not contribute significantly to the model.
- The F-statistic is 1.693 with a p-value of 0.0768, indicating that the overall model is on the borderline of statistical significance at the 0.05 alpha level.

Coefficients:



OLS Regression Results

Dep. Variable:	y	R-squared:	0.204
Model:	OLS	Adj. R-squared:	0.083
Method:	Least Squares	F-statistic:	1.693
Date:	Wed, 14 Feb 2024	Prob (F-statistic):	0.0768
Time:	05:38:41	Log-Likelihood:	-38.370
No. Observations:	100	AIC:	104.7
Df Residuals:	86	BIC:	141.2
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.9256	0.075	12.345	0.000	0.777	1.075
x1	0.1339	0.118	1.130	0.261	-0.102	0.369
x2	0.7874	0.169	4.664	0.000	0.452	1.123
x3	-0.0003	0.112	-0.002	0.998	-0.224	0.223
x4	-0.0176	0.119	-0.149	0.882	-0.253	0.218
x5	0.0266	0.144	0.184	0.854	-0.261	0.314
x6	0.0409	0.213	0.192	0.848	-0.382	0.464
x7	-0.0056	0.100	-0.055	0.956	-0.205	0.194
x8	0.0409	0.359	0.114	0.910	-0.673	0.755
x9	0.0342	0.168	0.204	0.839	-0.300	0.368
x10	-0.0169	0.122	-0.138	0.891	-0.260	0.226
x11	-0.0260	0.259	-0.101	0.920	-0.540	0.488
x12	0.0409	0.257	0.159	0.874	-0.471	0.552
x13	-0.0031	0.098	-0.032	0.975	-0.198	0.192
x14	0.0242	0.256	0.094	0.925	-0.485	0.534

Omnibus:	172.775	Durbin-Watson:	1.902
Prob(Omnibus):	0.000	Jarque-Bera (JB):	15781.325
Skew:	6.558	Prob(JB):	0.00
Kurtosis:	63.129	Cond. No.	2.09e+16

Figure 23: Impact of Reported Year and Offense Categories on Victim Count

- The constant term's coefficient is significant ($p < 0.05$), with a value of 0.9256, interpreted as the baseline victim count when all other variables are at their reference level.
- The normalized reported year (x_1) has a coefficient of 0.1339, but it is not statistically significant ($p = 0.261$).
- The offense category represented by x_2 has a significant positive coefficient (0.7874, $p < 0.05$), suggesting a strong relationship with the victim count.
- All other offense categories (x_3 to x_{14}) do not show statistically significant coefficients.

Diagnostics:

- The Omnibus test has a p-value of 0.000, indicating the residuals are not normally distributed, which could be a concern for the validity of some statistical tests.
- The Durbin-Watson statistic is 1.902, close to 2, suggesting there is no serious autocorrelation issue.
- The model displays high skewness (6.558) and kurtosis (63.129), confirming the non-normality of residuals.
- The Jarque-Bera test confirms the non-normality with a highly significant p-value ($p < 0.05$).

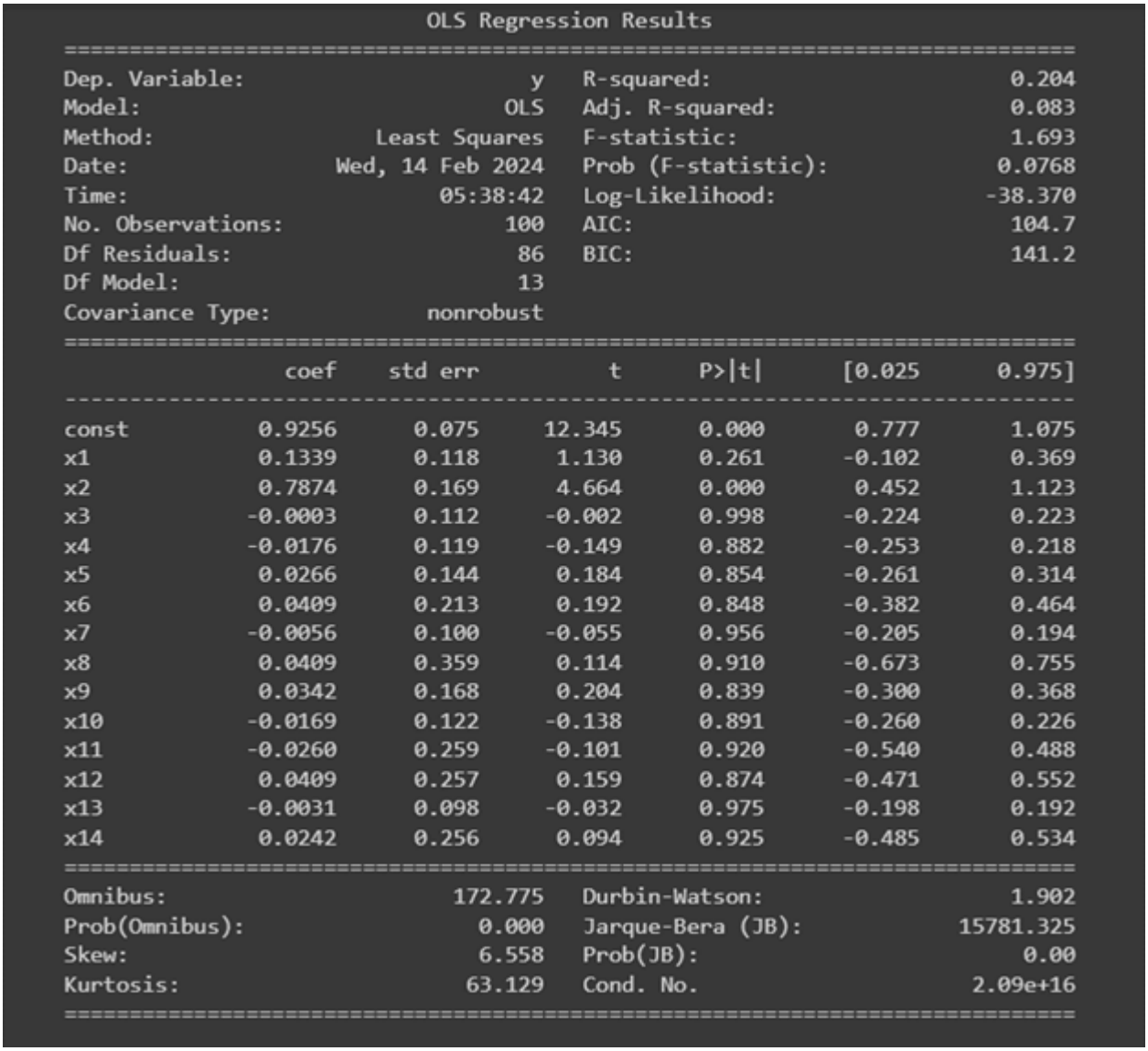
Conclusions:

- The model has limited explanatory power, with only one offense category showing a significant relationship with the victim count.
- The lack of significance in most coefficients, coupled with the low adjusted R-squared, suggests that additional relevant predictors are needed to improve the model's explanatory power.
- The diagnostics indicate issues with the distribution of residuals, which can affect the reliability of the coefficient estimates and the overall model.

4.2.4 Combined Effects of Multiple Variables on Victim Count

Regression Results:

- The R-squared is 0.204, indicating that around 20.4% of the variability in victim count can be explained by the model.



OLS Regression Results

Dep. Variable:	y	R-squared:	0.204
Model:	OLS	Adj. R-squared:	0.083
Method:	Least Squares	F-statistic:	1.693
Date:	Wed, 14 Feb 2024	Prob (F-statistic):	0.0768
Time:	05:38:42	Log-Likelihood:	-38.370
No. Observations:	100	AIC:	104.7
Df Residuals:	86	BIC:	141.2
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.9256	0.075	12.345	0.000	0.777	1.075
x1	0.1339	0.118	1.130	0.261	-0.102	0.369
x2	0.7874	0.169	4.664	0.000	0.452	1.123
x3	-0.0003	0.112	-0.002	0.998	-0.224	0.223
x4	-0.0176	0.119	-0.149	0.882	-0.253	0.218
x5	0.0266	0.144	0.184	0.854	-0.261	0.314
x6	0.0409	0.213	0.192	0.848	-0.382	0.464
x7	-0.0056	0.100	-0.055	0.956	-0.205	0.194
x8	0.0409	0.359	0.114	0.910	-0.673	0.755
x9	0.0342	0.168	0.204	0.839	-0.300	0.368
x10	-0.0169	0.122	-0.138	0.891	-0.260	0.226
x11	-0.0260	0.259	-0.101	0.920	-0.540	0.488
x12	0.0409	0.257	0.159	0.874	-0.471	0.552
x13	-0.0031	0.098	-0.032	0.975	-0.198	0.192
x14	0.0242	0.256	0.094	0.925	-0.485	0.534

Omnibus:	172.775	Durbin-Watson:	1.902
Prob(Omnibus):	0.000	Jarque-Bera (JB):	15781.325
Skew:	6.558	Prob(JB):	0.00
Kurtosis:	63.129	Cond. No.	2.09e+16

Figure 24: Combined Effects of Multiple Variables on Victim Count

- The adjusted R-squared is 0.083, suggesting limited explanatory power when adjusting for the number of predictors.
- The F-statistic is 1.693 with a p-value of 0.0768, showing that the model is not statistically significant at the 0.05 level.

Coefficients:

- The constant term's coefficient is statistically significant with a value of 0.9256, suggesting a baseline victim count when all predictors are at their reference levels.
- The coefficient for the normalized reported year (x_1) is not significant, nor are the coefficients for most other predictors, with the exception of x_2, which is statistically significant and indicates a strong positive relationship with the victim count.

Diagnostics:

- The Omnibus test has a p-value of 0.000, indicating non-normality in the residuals.
- The Durbin-Watson statistic is 1.902, suggesting no major issues with autocorrelation.
- High skewness and kurtosis values suggest that the residuals are not normally distributed.
- The Jarque-Bera test result supports the presence of non-normality in residuals.

Conclusions:

- The model does not have strong predictive power, as evidenced by the low adjusted R-squared and the lack of significance in most coefficients.
- Non-normality of residuals points to potential model specification issues or the presence of outliers or influential points that could be affecting the results.

4.2.5 Impact of Reported Hour on Victim Count

Regression Results:

- The R-squared is 0.009, indicating that only 0.9% of the variance in the victim count is explained by the reported hour, which is very low.
- The adjusted R-squared is -0.001, suggesting that the independent variable does not explain any of the variability of the response data around its mean.

OLS Regression Results

Dep. Variable:	victim_count	R-squared:	0.009
Model:	OLS	Adj. R-squared:	-0.001
Method:	Least Squares	F-statistic:	0.8609
Date:	Wed, 14 Feb 2024	Prob (F-statistic):	0.356
Time:	10:51:59	Log-Likelihood:	-49.325
No. Observations:	100	AIC:	102.6
Df Residuals:	98	BIC:	107.9
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.1246	0.100	11.289	0.000	0.927	1.322
reported_hour	-0.0063	0.007	-0.928	0.356	-0.020	0.007

Omnibus:	214.458	Durbin-Watson:	2.030
Prob(Omnibus):	0.000	Jarque-Bera (JB):	37853.314
Skew:	9.718	Prob(JB):	0.00
Kurtosis:	96.311	Cond. No.	36.8

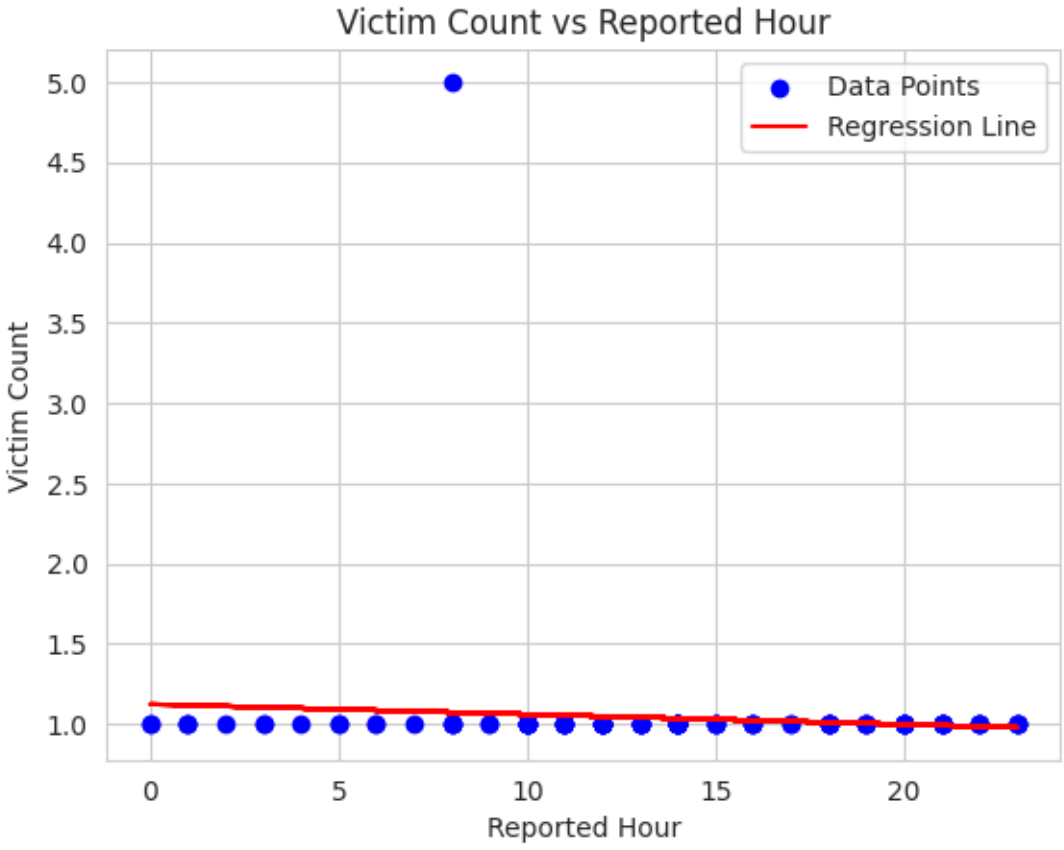


Figure 25: Impact of Reported Hour on Victim Count

- The F-statistic is 0.862 with a p-value of 0.356, showing that the model is not statistically significant.

Coefficients:

- The coefficient for the constant term is 1.1246 with a standard error of 0.100, and it is statistically significant (p ¡ 0.05).
- The coefficient for the reported hour is -0.0063 with a standard error of 0.067, and it is not statistically significant (p = 0.356).

Diagnostics:

- The Omnibus test gives a p-value of 0.000, indicating that the residuals of the model are not normally distributed.
- The Durbin-Watson statistic is 2.030, suggesting that there is no autocorrelation in the residuals.
- The residuals have high skewness (9.718) and kurtosis (96.311), indicating a non-normal distribution.
- The Jarque-Bera test has a p-value of 0.00, confirming the non-normal distribution of the residuals.

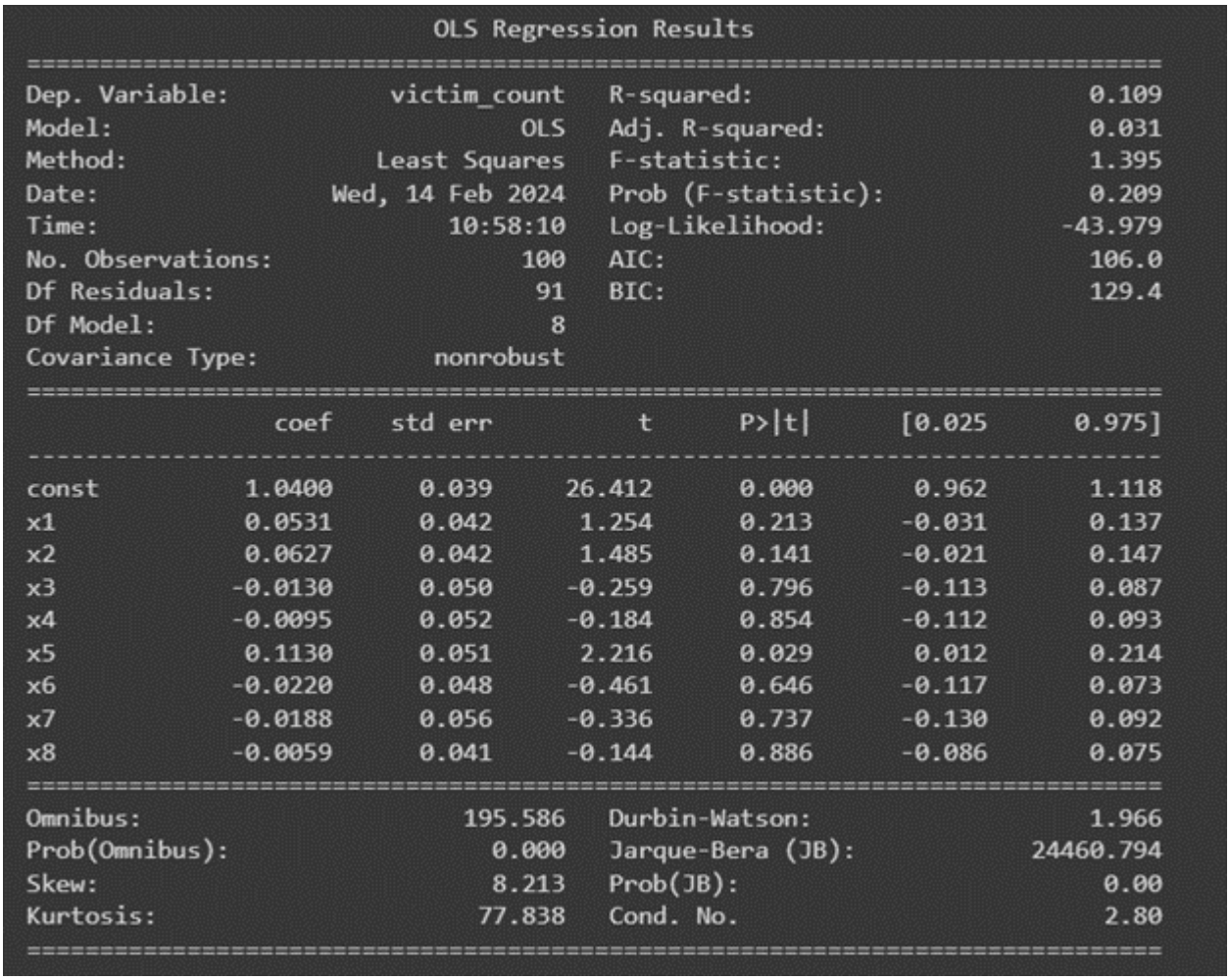
Conclusions:

- There is no significant relationship between the reported hour and the victim count based on the current dataset and the model used.
- The low R-squared values indicate that the reported hour is not a good predictor of the victim count.
- The statistical tests for normality indicate potential issues with the model, as the assumptions for OLS regression are not fully met.

4.2.6 Effects of District ID, Reported Year, and Reported Month on Victim Count

Regression Results:

- The R-squared of the model is 0.109, indicating that approximately 10.9% of the variability in the victim count is explained by the model.



OLS Regression Results

Dep. Variable:	victim_count	R-squared:	0.109
Model:	OLS	Adj. R-squared:	0.031
Method:	Least Squares	F-statistic:	1.395
Date:	Wed, 14 Feb 2024	Prob (F-statistic):	0.209
Time:	10:58:10	Log-Likelihood:	-43.979
No. Observations:	100	AIC:	106.0
Df Residuals:	91	BIC:	129.4
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.0400	0.039	26.412	0.000	0.962	1.118
x1	0.0531	0.042	1.254	0.213	-0.031	0.137
x2	0.0627	0.042	1.485	0.141	-0.021	0.147
x3	-0.0130	0.050	-0.259	0.796	-0.113	0.087
x4	-0.0095	0.052	-0.184	0.854	-0.112	0.093
x5	0.1130	0.051	2.216	0.029	0.012	0.214
x6	-0.0220	0.048	-0.461	0.646	-0.117	0.073
x7	-0.0188	0.056	-0.336	0.737	-0.130	0.092
x8	-0.0059	0.041	-0.144	0.886	-0.086	0.075

Omnibus:	195.586	Durbin-Watson:	1.966
Prob(Omnibus):	0.000	Jarque-Bera (JB):	24460.794
Skew:	8.213	Prob(JB):	0.00
Kurtosis:	77.838	Cond. No.	2.80

Figure 26: Effects of District ID, Reported Year, and Reported Month on Victim Count

- The adjusted R-squared is 0.031, which is quite low, suggesting that few of the independent variables meaningfully contribute to the model after adjusting for the number of predictors.
- The F-statistic is 1.635 with a p-value of 0.209, implying that the model is not statistically significant.

Coefficients:

- The coefficient for the constant term is statistically significant (p ¡ 0.05), with a value of 1.0400.
- None of the coefficients for the independent variables (x1 through x8) are statistically significant, with p-values all above the conventional threshold of 0.05.

Diagnostics:

- The Omnibus test has a p-value of 0.000, suggesting non-normality in the residuals.
- The Durbin-Watson statistic is close to 2 (1.966), indicating no serious autocorrelation concerns.
- There is a high skewness (8.213) and kurtosis (77.838), indicating that the residuals are not normally distributed.
- The Jarque-Bera test corroborates the non-normality of the residuals with a significant p-value (p ¡ 0.05).

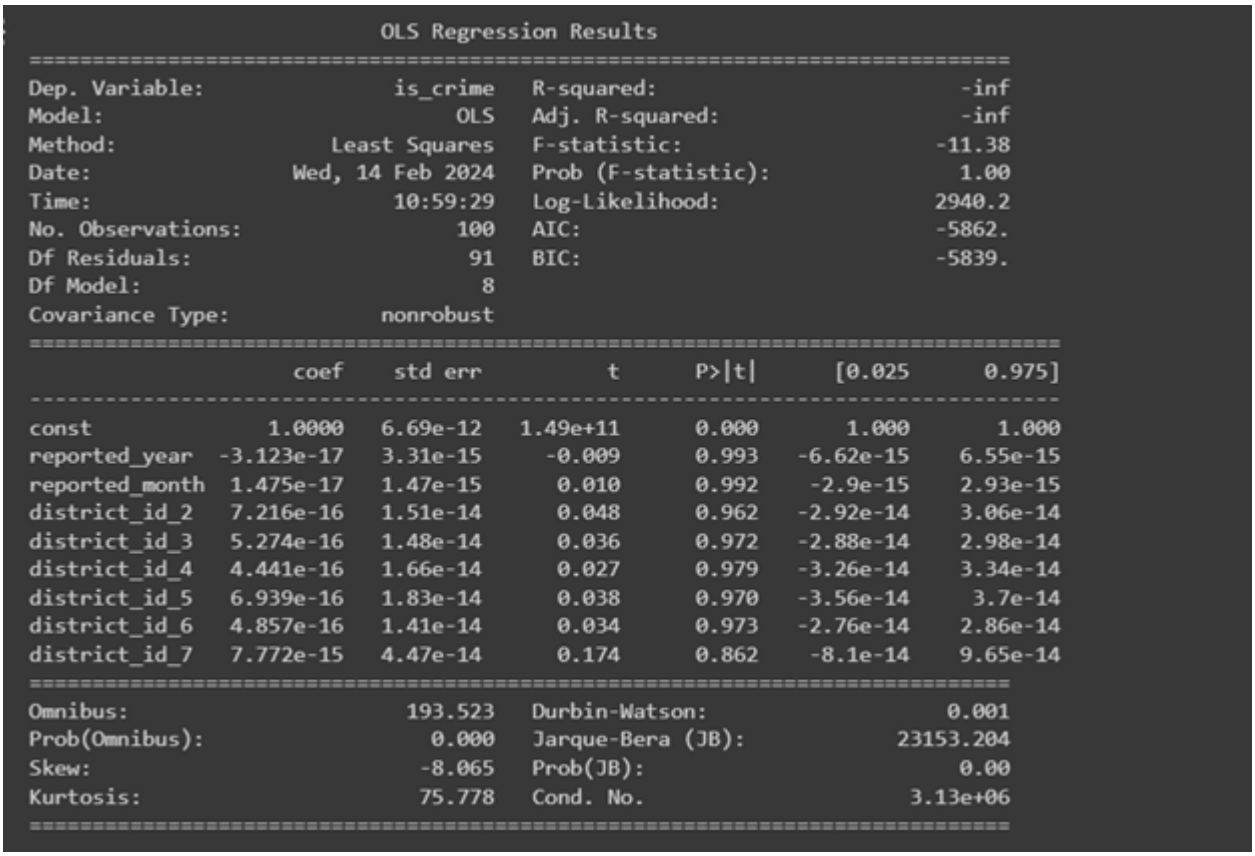
Conclusions:

- The model does not provide a statistically significant explanation of the variation in victim count.
- The low R-squared values indicate that the included predictors have limited predictive power.
- The diagnostics suggest that the assumptions necessary for OLS regression, such as normality of residuals, are not met.

4.2.7 Probability of Incident Classification as a Crime

Regression Results:

- The R-squared and adjusted R-squared are both reported as negative infinity, indicating a potential issue with the model.



OLS Regression Results

Dep. Variable:	is_crime	R-squared:	-inf
Model:	OLS	Adj. R-squared:	-inf
Method:	Least Squares	F-statistic:	-11.38
Date:	Wed, 14 Feb 2024	Prob (F-statistic):	1.00
Time:	10:59:29	Log-Likelihood:	2940.2
No. Observations:	100	AIC:	-5862.
Df Residuals:	91	BIC:	-5839.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.0000	6.69e-12	1.49e+11	0.000	1.000	1.000
reported_year	-3.123e-17	3.31e-15	-0.009	0.993	-6.62e-15	6.55e-15
reported_month	1.475e-17	1.47e-15	0.010	0.992	-2.9e-15	2.93e-15
district_id_2	7.216e-16	1.51e-14	0.048	0.962	-2.92e-14	3.06e-14
district_id_3	5.274e-16	1.48e-14	0.036	0.972	-2.88e-14	2.98e-14
district_id_4	4.441e-16	1.66e-14	0.027	0.979	-3.26e-14	3.34e-14
district_id_5	6.939e-16	1.83e-14	0.038	0.970	-3.56e-14	3.7e-14
district_id_6	4.857e-16	1.41e-14	0.034	0.973	-2.76e-14	2.86e-14
district_id_7	7.772e-15	4.47e-14	0.174	0.862	-8.1e-14	9.65e-14

Omnibus:	193.523	Durbin-Watson:	0.001
Prob(Omnibus):	0.000	Jarque-Bera (JB):	23153.204
Skew:	-8.065	Prob(JB):	0.00
Kurtosis:	75.778	Cond. No.	3.13e+06

Figure 27: Probability of Incident Classification as a Crime

- The F-statistic is extremely negative (-11.38), and the p-value is 1.00, suggesting fundamental issues with the regression model.

Coefficients:

- The coefficient for the constant term is 1.0000 with a standard error on the order of 10^{-12}, and it is statistically significant (p ¡ 0.05).
- All other coefficients for reported year, reported month, and district IDs are extremely small and not statistically significant.

Diagnostics:

- The Omnibus test has a p-value of 0.000, suggesting that the residuals of the model are not normally distributed.
- The Durbin-Watson statistic is 0.001, indicating potential problems with the model.
- The skewness and kurtosis values are extremely high, further suggesting issues with the residual distribution.
- The Jarque-Bera test result is significantly large, indicating that the residuals do not follow a normal distribution.

Conclusions:

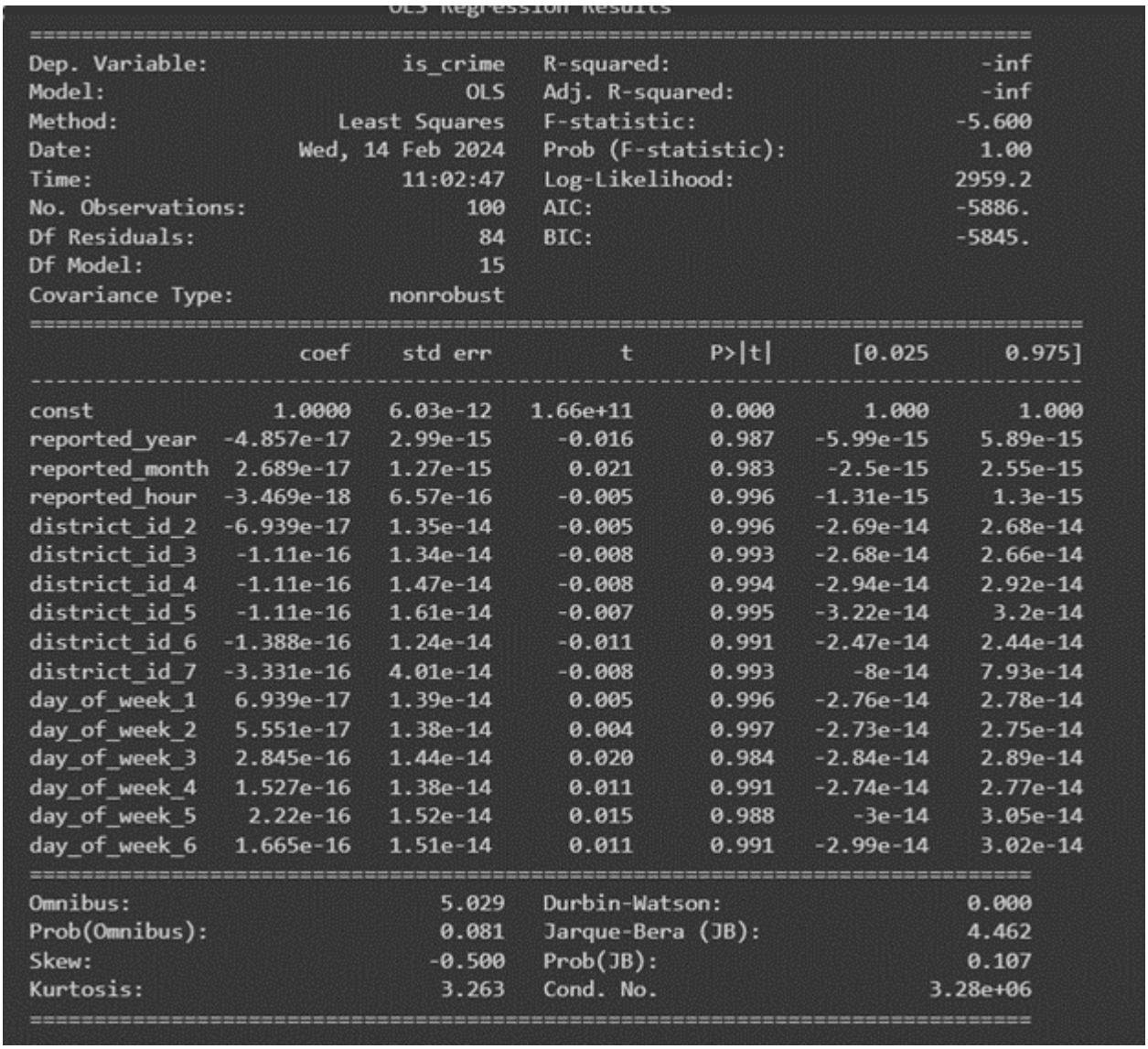
- The regression output indicates serious issues with the model, which may be due to the use of OLS regression with a binary dependent variable or errors in data entry or processing.
- The predictors do not appear to have a significant relationship with the dependent variable 'is_crime' based on this model.

4.2.8 Likelihood of Incident Classification as a Crime

Regression Results:

- The R-squared and adjusted R-squared values are reported as negative infinity, indicating potential issues with the model.
- The F-statistic is -5.600 with a p-value of 1.00, suggesting a lack of overall statistical significance.

Coefficients:



OLS Regression Results

Dep. Variable:	is_crime	R-squared:	-inf
Model:	OLS	Adj. R-squared:	-inf
Method:	Least Squares	F-statistic:	-5.600
Date:	Wed, 14 Feb 2024	Prob (F-statistic):	1.00
Time:	11:02:47	Log-Likelihood:	2959.2
No. Observations:	100	AIC:	-5886.
Df Residuals:	84	BIC:	-5845.
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.0000	6.03e-12	1.66e+11	0.000	1.000	1.000
reported_year	-4.857e-17	2.99e-15	-0.016	0.987	-5.99e-15	5.89e-15
reported_month	2.689e-17	1.27e-15	0.021	0.983	-2.5e-15	2.55e-15
reported_hour	-3.469e-18	6.57e-16	-0.005	0.996	-1.31e-15	1.3e-15
district_id_2	-6.939e-17	1.35e-14	-0.005	0.996	-2.69e-14	2.68e-14
district_id_3	-1.11e-16	1.34e-14	-0.008	0.993	-2.68e-14	2.66e-14
district_id_4	-1.11e-16	1.47e-14	-0.008	0.994	-2.94e-14	2.92e-14
district_id_5	-1.11e-16	1.61e-14	-0.007	0.995	-3.22e-14	3.2e-14
district_id_6	-1.388e-16	1.24e-14	-0.011	0.991	-2.47e-14	2.44e-14
district_id_7	-3.331e-16	4.01e-14	-0.008	0.993	-8e-14	7.93e-14
day_of_week_1	6.939e-17	1.39e-14	0.005	0.996	-2.76e-14	2.78e-14
day_of_week_2	5.551e-17	1.38e-14	0.004	0.997	-2.73e-14	2.75e-14
day_of_week_3	2.845e-16	1.44e-14	0.020	0.984	-2.84e-14	2.89e-14
day_of_week_4	1.527e-16	1.38e-14	0.011	0.991	-2.74e-14	2.77e-14
day_of_week_5	2.22e-16	1.52e-14	0.015	0.988	-3e-14	3.05e-14
day_of_week_6	1.665e-16	1.51e-14	0.011	0.991	-2.99e-14	3.02e-14

Omnibus:	5.029	Durbin-Watson:	0.000
Prob(Omnibus):	0.081	Jarque-Bera (JB):	4.462
Skew:	-0.500	Prob(JB):	0.107
Kurtosis:	3.263	Cond. No.	3.28e+06

Figure 28: Likelihood of Incident Classification as a Crime

- The coefficient for the constant term is exactly 1.0000, which is abnormal for OLS regression with a binary dependent variable.
- All other coefficients are extremely small and statistically insignificant.

Diagnostics:

- The Omnibus test indicates a possibility of non-normality in the residuals (p = 0.081).
- The Durbin-Watson statistic is 0.000, suggesting potential issues with the model such as autocorrelation.
- Skewness and kurtosis indicate a deviation from the normal distribution but are not as extreme as in the previous models.

Conclusions:

- The OLS regression model is not appropriate for the binary dependent variable 'is_crime'. The coefficients and statistics suggest significant issues with the model.
- The variables included do not provide significant predictive power for the outcome based on this model.

5 Conclusion

This comprehensive study on Denver's crime dynamics, utilizing advanced data analytics and regression models, reveals intricate patterns of criminal activity influenced by temporal, spatial, and socio-economic factors. Through meticulous data preprocessing, visualization, and statistical analysis, it uncovers seasonal crime trends, district-specific crime rates, and the impact of various predictors on crime and victim counts. While some models demonstrate strong predictive power, others highlight the complex nature of crime forecasting. The findings suggest targeted interventions could be more effective, emphasizing the need for ongoing analysis and adaptation of strategies to address urban crime effectively.