Sruthi Kondra

ALY6010: Probability Theory and Introductory Statistics

Final Project

Professor Mykhaylo Trubskyy

Date:- 12/16/23

Introduction

Overview of the Dataset

I have selected the Crime Incident Reports In Boston dataset from data.boston.gov for my Milestone project. I specifically selected and began working with the data from the year 2022. There are 16 columns altogether and 73,852 observations in all. This dataset includes information on the kind of offence, its code, the street where it happened, the precise time, longitude, latitude, and other crucial factors that enable us to conduct a thorough examination of the crime dataset as part of our exploratory data analysis. The primary objective of gathering info is to record various criminal crimes and offer in-depth information about each instance. To facilitate law enforcement and public safety efforts, it is a crucial instrument for understanding criminal trends and patterns in a particular area

Summary of Initial Exploratory Data

The urban landscape is in a constant state of flux, influenced by various social, economic, and political factors. Crime, as a crucial component of urban life, significantly impacts the quality of life and shapes public policy. Drawn to the intricacies of urban dynamics, I was captivated by the "Crime Incident Reports in Boston" dataset sourced from data.boston.gov. Spanning the year 2022, this extensive dataset encompasses 73,852 instances across 16 diverse attributes that marry categorical and numerical data types.

In the preliminary stages of my exploratory data analysis, I undertook the meticulous task of data cleansing to rectify discrepancies and ensure the integrity of subsequent findings. My investigative journey was marked by an array of visual methodologies, including scatter plots, histograms, bar plots, box plots, and summative tables. These tools were instrumental in deciphering complex data structures and uncovering underlying patterns.

Here's what I unearthed through the initial phases of my data exploration:

- I delved into the distribution of crime occurrences over the hours of the day, discerning potential correlations that may guide law enforcement strategies.
- By employing scatter plots, I inspected the relationship between crime frequency and geographical coordinates, unraveling spatial crime trends within the cityscape.
- I charted the prevalence of specific crime types, identifying which offenses were most commonplace across different Boston locales.
- I also discerned the most frequent categories of crime, shedding light on the nature of offenses that predominantly shape Boston's crime narrative.

This data-driven inquiry not only deepened my understanding of Boston's public safety landscape but also underscored the potential of such analysis to inform and refine policy-making, ultimately striving toward a safer and more harmonious urban environment.

Data Cleaning and Preparation

The dataset underwent a thorough cleaning process to ensure the accuracy and reliability of the analysis. This process involved renaming column headers for better readability and standardization, converting data types to their appropriate formats for accurate computation and analysis like converting offense codes to integers, offense description and location as factors and categorizing crimes into distinct groups such as 'Violent', 'Moderate', and 'Fraud'. This categorization aids in understanding the severity and nature of the crimes reported.

First we load and download all the required packages

```
#downloading and loading different packages
library(pacman)
library(dplyr)
library(tidyverse)
library(flextable)
library(janitor)
library(ggplot2)
p_load(tidyverse)
p_load(janitor)
p_load(lubridate)
```

Steps involved in Data Cleaning and Preparation and its outcomes

```
#reads the dataframe into parcel_dataset
crime_dataset <- read.csv("tmpdfeo3qy2.csv")
crime_dataset

#cleaning the column names and printing them
crime_dataset <- clean_names(crime_dataset)
names(crime_dataset)

glimpse(crime_dataset)

# Changing data types
crime_dataset$offense_code <- as.integer(crime_dataset$offense_code)
crime_dataset$offense_description <- as.factor(crime_dataset$offense_description)
crime_dataset$location <- as.factor(crime_dataset$location)

# Dropping the offense_code_group column
crime_dataset <- crime_dataset %>%
  select(-offense_code_group)

# View the cleaned dataset
head(crime_dataset)

summary(crime_dataset)

str(crime_dataset)
```

```
>
> #cleaning the column names and printing them
> crime_dataset <- clean_names(crime_dataset)
> names(crime_dataset)
 [1] "incident_number"     "offense_code"       "offense_code_group"  "offense_description" "district"
 [6] "reporting_area"      "shooting"           "occurred_on_date"    "year"                "month"
[11] "day_of_week"         "hour"               "ucr_part"            "street"              "lat"
[16] "long"                "location"
>
> glimpse(crime_dataset)
Rows: 73,852
Columns: 17
$ incident_number     <chr> "222076257", "222053099", "222039411", "222011090", "222062685", "222040307", "222023700", "22…
$ offense_code        <int> 619, 2670, 3201, 3201, 3201, 3115, 2670, 3114, 1109, 423, 2647, 3201, 3201, 3201, 1402, 3114, …
$ offense_code_group  <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…
$ offense_description <chr> "LARCENY ALL OTHERS", "HARASSMENT/ CRIMINAL HARASSMENT", "PROPERTY - LOST/ MISSING", "PROPERTY…
$ district            <chr> "D4", "A7", "D14", "B3", "B3", "A1", "D4", "B3", "C6", "D4", "A1", "A1", "A15", "A1", "E5", "A…
$ reporting_area      <int> 167, NA, 778, 465, 465, 954, NA, NA, 200, NA, 77, 77, 41, NA, NA, NA, NA, 93, NA, 519, 319, 13…
$ shooting            <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0…
$ occurred_on_date    <chr> "2022-01-01 00:00:00", "2022-01-01 00:00:00", "2022-01-01 00:00:00", "2022-01-01 00:00:00", "2…
$ year                <int> 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022…
$ month               <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
$ day_of_week         <chr> "Saturday", "Saturday", "Saturday", "Saturday", "Saturday", "Saturday", "Saturday", "Saturday"…
$ hour                <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1…
$ ucr_part            <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…
$ street              <chr> "HARRISON AVE", "BENNINGTON ST", "WASHINGTON ST", "BLUE HILL AVE", "BLUE HILL AVE", "FULTON ST…
$ lat                 <dbl> 42.33954, 42.37725, 42.34906, 42.28483, 42.28483, 42.36294, 42.33954, 42.28089, 42.34129, 42.3…
$ long                <dbl> -71.06941, -71.03260, -71.15050, -71.09137, -71.09137, -71.05254, -71.06941, -71.08037, -71.05…
$ location            <chr> "(42.33954198983014, -71.06940876967543)", "(42.37724638479816, -71.03259708041278)", "(42.3490…
>
> # Changing data types
> crime_dataset$offense_code <- as.integer(crime_dataset$offense_code)
> crime_dataset$offense_description <- as.factor(crime_dataset$offense_description)
> crime_dataset$location <- as.factor(crime_dataset$location)
>
> # Dropping the offense_code_group column
> crime_dataset <- crime_dataset %>%
+   select(-offense_code_group)
>
```

```
> # View the cleaned dataset
> head(crime_dataset)
  incident_number offense_code             offense_description district reporting_area shooting    occurred_on_date year
1       222076257          619              LARCENY ALL OTHERS       D4            167        0 2022-01-01 00:00:00 2022
2       222053099         2670 HARASSMENT/ CRIMINAL HARASSMENT       A7             NA        0 2022-01-01 00:00:00 2022
3       222039411         3201         PROPERTY - LOST/ MISSING      D14            778        0 2022-01-01 00:00:00 2022
4       222011090         3201         PROPERTY - LOST/ MISSING       B3            465        0 2022-01-01 00:00:00 2022
5       222062685         3201         PROPERTY - LOST/ MISSING       B3            465        0 2022-01-01 00:00:00 2022
6       222040307         3115                INVESTIGATE PERSON       A1            954        0 2022-01-01 00:00:00 2022
  month day_of_week hour ucr_part        street      lat      long                 location
1     1    Saturday    0       NA  HARRISON AVE 42.33954 -71.06941 (42.33954198983014, -71.06940876967543)
2     1    Saturday    0       NA BENNINGTON ST 42.37725 -71.03260  (42.37724638479816, -71.0325970804128)
3     1    Saturday    0       NA WASHINGTON ST 42.34906 -71.15050 (42.34905600030506, -71.15049849975023)
4     1    Saturday    0       NA BLUE HILL AVE 42.28483 -71.09137 (42.28482576580488, -71.09137368938802)
5     1    Saturday    0       NA BLUE HILL AVE 42.28483 -71.09137 (42.28482576580488, -71.09137368938802)
6     1    Saturday    0       NA      FULTON ST 42.36294 -71.05254  (42.36293610909294, -71.0525379472723)
>
> summary(crime_dataset)
 incident_number    offense_code                          offense_description   district         reporting_area
 Length:73852       Min.   : 111   INVESTIGATE PERSON                 : 8070    Length:73852       Min.   :  1.0
 Class :character   1st Qu.:1107   SICK ASSIST                        : 5484    Class :character   1st Qu.:167.0
 Mode  :character   Median :3006   M/V - LEAVING SCENE - PROPERTY DAMAGE: 4562    Mode  :character   Median :355.0
                    Mean   :2377   INVESTIGATE PROPERTY               : 3539                       Mean   :371.7
                    3rd Qu.:3207   TOWED MOTOR VEHICLE                : 3190                       3rd Qu.:522.0
                    Max.   :3831   ASSAULT - SIMPLE                   : 2972                       Max.   :962.0
                                   (Other)                            :46035                       NA's   :44448
    shooting        occurred_on_date       year          month         day_of_week            hour      ucr_part
 Min.   :0.000000   Length:73852       Min.   :2022   Min.   : 1.000   Length:73852       Min.   : 0.00   Mode:logical
 1st Qu.:0.000000   Class :character   1st Qu.:2022   1st Qu.: 4.000   Class :character   1st Qu.: 9.00   NA's:73852
 Median :0.000000   Mode  :character   Median :2022   Median : 7.000   Mode  :character   Median :13.00
 Mean   :0.009925                      Mean   :2022   Mean   : 6.574                      Mean   :12.67
 3rd Qu.:0.000000                      3rd Qu.:2022   3rd Qu.: 9.000                      3rd Qu.:18.00
 Max.   :1.000000                      Max.   :2022   Max.   :12.000                      Max.   :23.00

    street              lat             long                                        location
 Length:73852       Min.   :42.21   Min.   :-71.35                                      : 3808
 Class :character   1st Qu.:42.30   1st Qu.:-71.10   (42.29755532959655, -71.05970910242573) : 2666
 Mode  :character   Median :42.33   Median :-71.08   (42.33954198983014, -71.06940876967543) : 2207
                    Mean   :42.32   Mean   :-71.08   (42.34905600030506, -71.15049849975023) : 1464
                    3rd Qu.:42.35   3rd Qu.:-71.06   (42.28482576580488, -71.09137368938802) : 1345
```

Initial Steps in Analysis

Creating a data set for the offense code and the offense description from the original crime dataset called offense details

```
> offense_details <- crime_dataset %>%
+   select(offense_code, offense_description) %>%
+   arrange(offense_code) %>%
+   distinct()
> View(offense_details)
```

	offense_code	offense_description
1	111	MURDER, NON-NEGLIGIENT MANSLAUGHTER
2	111	MURDER, NON-NEGLIGENT MANSLAUGHTER
3	121	MANSLAUGHTER - VEHICLE - NEGLIGENCE
4	301	ROBBERY
5	423	ASSAULT - AGGRAVATED
6	520	BURGLARY - RESIDENTIAL
7	530	BREAKING AND ENTERING (B&E) MOTOR VEHICLE
8	531	BREAKING AND ENTERING (B&E) MOTOR VEHICLE (NO PRO...
9	540	BURGLARY - COMMERICAL
10	611	LARCENY PICK-POCKET

We are now defining the offense code into 3 crime categories such as violent, moderate and fraud according to the description and put all the remaining crimes into other

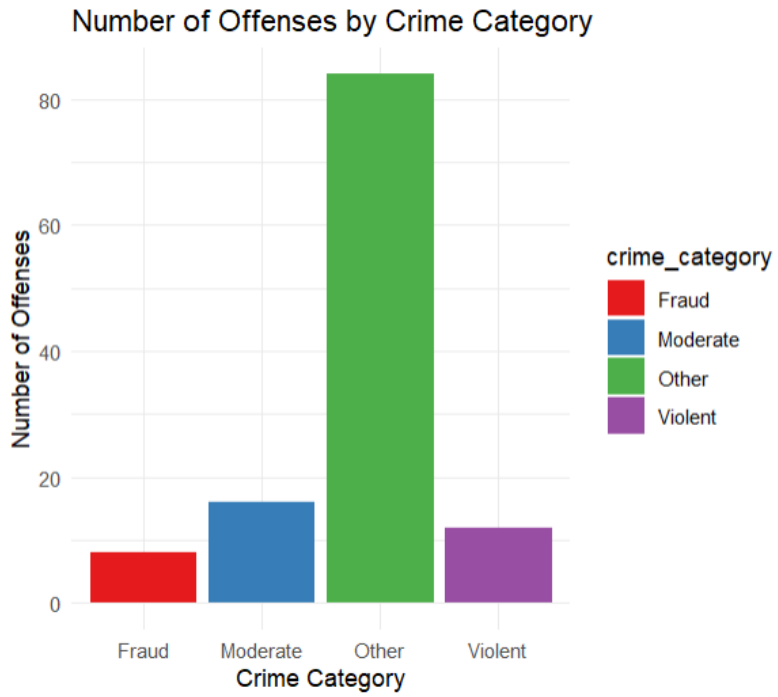
```
> # Define offense codes for each crime category
> violent_crimes <- c(111,121,301,423,520,530,531,540,2500,2511,2604,2618,2622)
> moderate_crimes <- c(611,612,613,614,615,616,617,618,619,706,724,727,801,900,2610,2670)
> fraud_crimes <- c(1001,1102,1106,1107,1108,1109,1201,2636)
>
```

Once the categorization is done we then add this column crime category to our offense details subset

```
> # Categorize each offense
> offense_details$crime_category <- case_when(
+   offense_details$offense_code %in% violent_crimes ~ "Violent",
+   offense_details$offense_code %in% moderate_crimes ~ "Moderate",
+   offense_details$offense_code %in% fraud_crimes ~ "Fraud",
+   TRUE ~ "Other"
+ )
>
> head(offense_details)
  offense_code                offense_description crime_category
1          111 MURDER, NON-NEGLIGIENT MANSLAUGHTER        Violent
2          111  MURDER, NON-NEGLIGENT MANSLAUGHTER        Violent
3          121 MANSLAUGHTER - VEHICLE - NEGLIGENCE        Violent
4          301                            ROBBERY        Violent
5          423                ASSAULT - AGGRAVATED        Violent
6          520              BURGLARY - RESIDENTIAL        Violent
>
```

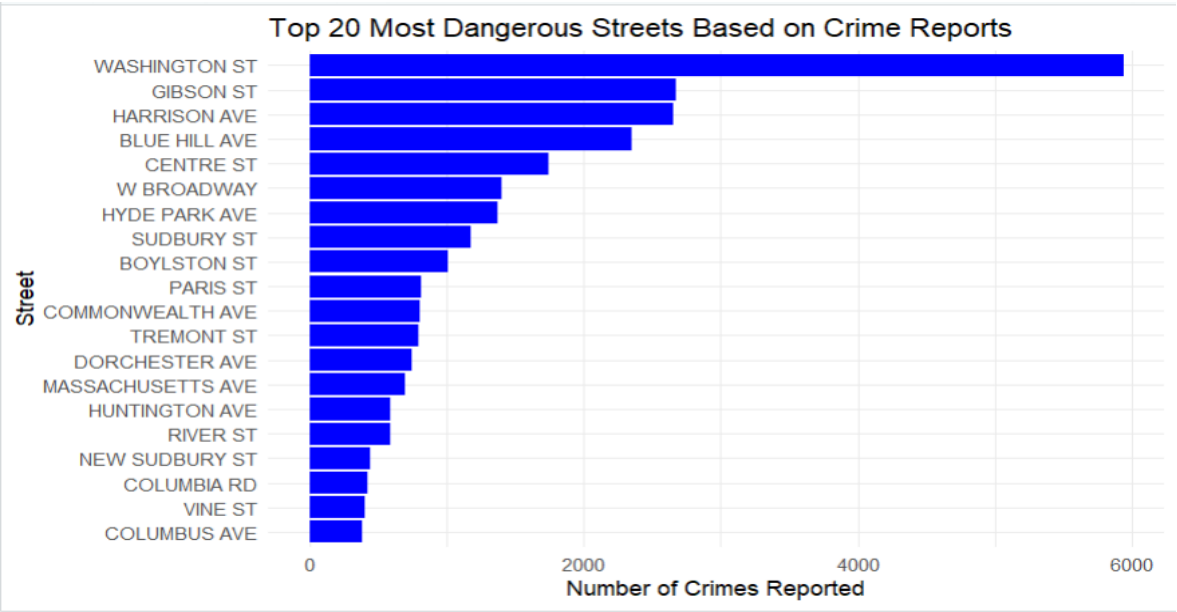
Visualizations from Initial Analysis

1.Bar chart to display the number of offenses by the different crime categories



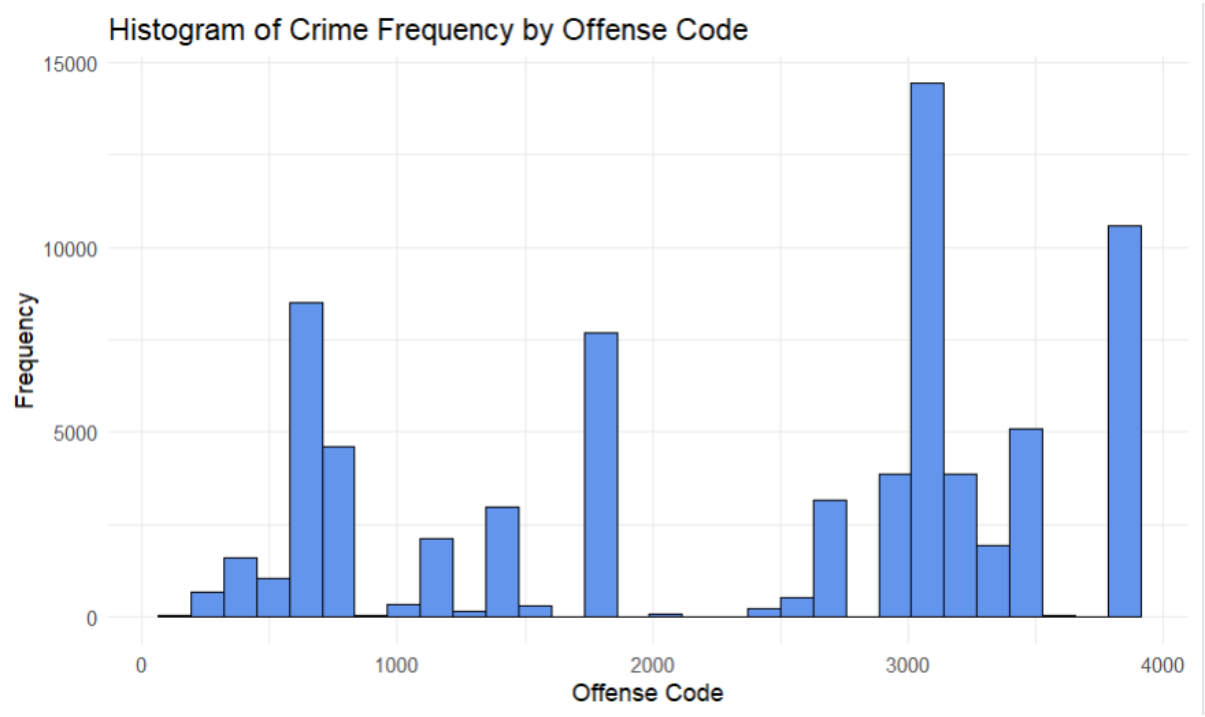
The 'Other' category of crimes significantly surpasses 'Violent', 'Moderate', and 'Fraud' in frequency, suggesting a broad range of criminal activities beyond major categories that could require additional categorization or analysis.

2.Bar chart of the top 20 most dangerous streets in boston



Washington Street stands out as the most reported location for crime, indicating a hotspot that may require targeted law enforcement and community safety interventions.

3. Histogram of Crime Frequency by Offense Code



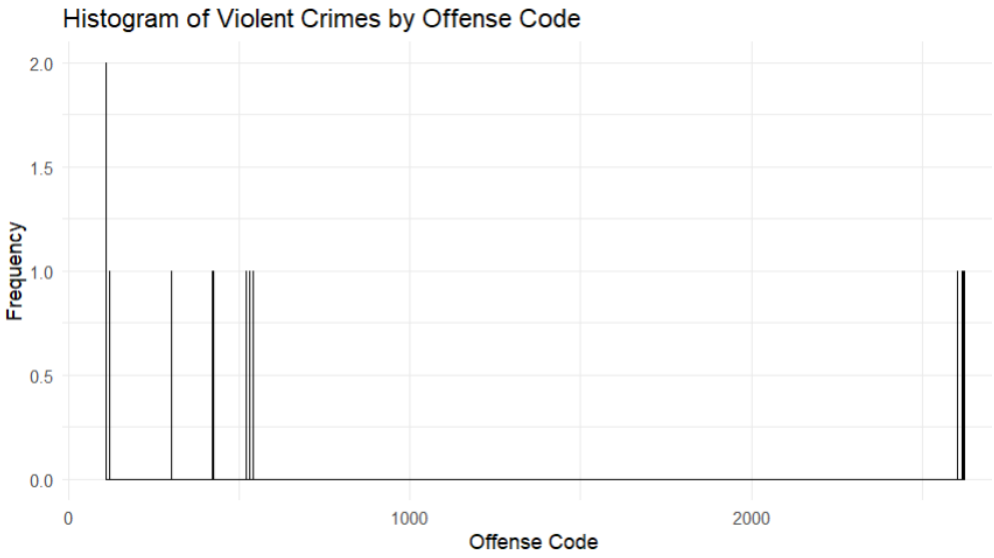
A few offense codes show a particularly high frequency of crimes, pointing towards specific types of criminal activities that are more prevalent and may need focused prevention strategies.

4.Descriptive Statistics Table for Each Crime Category

crime_category	count	mean_offense_code	median_offense_code
Fraud	8	1,296.2500	1,107.5
Moderate	16	917.0625	618.5
Other	84	2,843.2381	3,023.5
Violent	12	919.3333	525.0

The above table is used to obtain descriptive statistics for each crime category, such as mean, median, and count of the offense code and display it in a table

5.Histogram of Violent Crimes by Offense Code



Violent crimes are concentrated within specific offense codes, which could be pivotal in understanding and addressing the root causes of violence in the community.

Analysis of Crime Dataset

Questions Explored and why we choose these specific questions

The specific questions in the exploratory data analysis (EDA) of the Boston crime dataset were chosen to uncover underlying patterns that the initial EDA suggested might exist. This includes examining the relationship between crime frequency and time of day, geographic coordinates, and monthly trends. These facets were explored to provide actionable insights for public safety strategies. Since offense codes are categorical with wide-ranging numerical values, using their means for analysis is less meaningful. Instead, aggregating data into time slots, geographic units, and temporal segments like months offers a more insightful perspective, allowing for a more informed and strategic approach to understanding crime patterns.

Question 1:

Is there a significant relationship between the time of day (hour) and the frequency of violent crimes?

Variables:

- Dependent Variable: Frequency of violent crimes
- Independent Variable: Time of day (hour)

Hypothesis Testing

1. **Null Hypothesis (H0)**: There is no significant relationship between the time of day and the frequency of violent crimes. This means that the hour of the day does not have a statistically significant impact on how often violent crimes occur.
2. **Alternative Hypothesis (H1)**: There is a significant relationship between the time of day and the frequency of violent crimes. This implies that certain hours of the day might have higher or lower frequencies of violent crime incidents.

The type of testing used in the analysis of the relationship between the time of day and the frequency of violent crimes is linear regression analysis. The test results from the linear regression analysis are used to evaluate the significance of this relationship.

Methodology:

Data is filtered for violent crimes and aggregated by hour. A linear regression is performed to test the hypotheses.

```
>
> #___QUESTION 1____
> #Is there a significant relationship between the time of day (hour) and the frequency of violent crimes?
>
> # Filtering for violent crimes
> violent_crime_data <- crime_dataset %>%
+   filter(offense_code %in% violent_crimes)
>
> # Aggregating data by hour
> hourly_crime_frequency <- violent_crime_data %>%
+   group_by(hour) %>%
+   summarize(frequency = n())
>
> # Linear regression analysis
> lm_result <- lm(frequency ~ hour, data = hourly_crime_frequency)
> summary(lm_result)
```

Results:

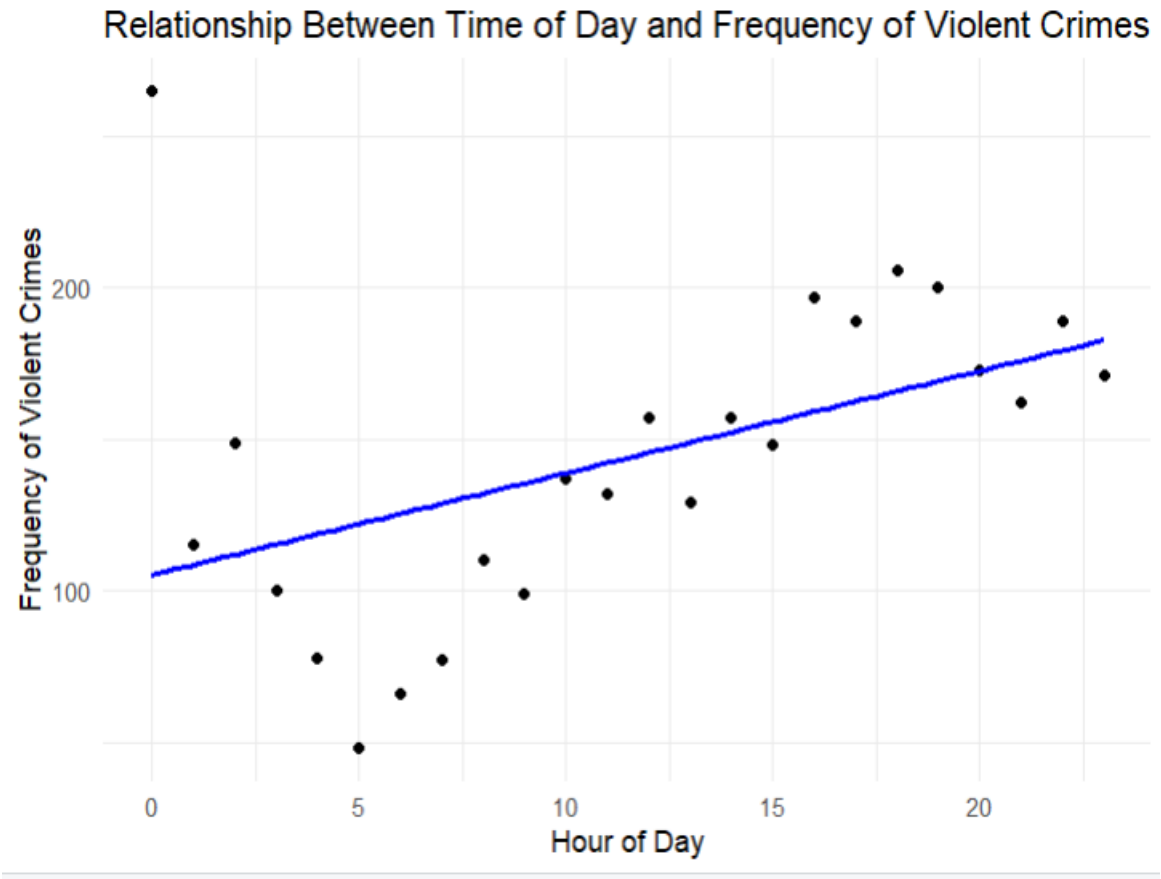
```
Call:
lm(formula = frequency ~ hour, data = hourly_crime_frequency)

Residuals:
    Min      1Q  Median      3Q     Max
-74.003 -20.509  -4.788  15.184 159.853

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  105.147     18.527   5.675 0.0000105 ***
hour           3.371      1.380   2.442    0.0231 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 46.81 on 22 degrees of freedom
Multiple R-squared:  0.2133,    Adjusted R-squared:  0.1776
F-statistic: 5.965 on 1 and 22 DF,  p-value: 0.02308
```

The linear regression analysis reveals that as each hour passes, violent crimes rise by about 3.371 incidents. With a p-value at 0.0231, this trend is statistically meaningful, as it falls below the standard cutoff of 0.05, allowing us to assert the influence of time of day on crime frequency with confidence.

Interpretation

The data shows that violent crimes tend to increase as the day progresses. The low p-value leads to the rejection of the null hypothesis and acceptance of the alternative, indicating a significant relationship between the time of day and the frequency of violent crimes

Visualization



This scatterplot shows the relationship between the time of day (hours) and the frequency of violent crimes. Each dot represents an hour of the day and the corresponding number of violent crimes reported during that hour. The line across the plot is the result of a linear regression analysis and indicates a positive trend, suggesting that the frequency of violent crimes increases as the day progresses. The slope of the line is upward, supporting the conclusion that there is a statistically significant relationship between the time of day and the occurrence of violent crimes.

Question 2:

Is there a significant correlation between the latitude of a crime incident and the frequency of crimes in that area?

Variables:

- Dependent Variable: Frequency of crimes(frequency)
- Independent Variable: Latitude(lat)

Hypothesis Testing:

1. **Null Hypothesis (H0)**: There is no correlation between latitude and the total number of crime incidents. This implies that the geographical location (north to south) does not significantly influence the crime rate.
2. **Alternative Hypothesis (H1)**: There is a significant correlation between latitude and the total number of crime incidents. This suggests that as you move either north or south, there is a noticeable change in the crime rate, either increasing or decreasing.

For this hypothesis testing we used a correlation test called the Pearson's correlation coefficient. This test assesses the strength and direction of the relationship between two continuous variables in this case, latitude and total crime incidents.

Methodology:

Data was grouped by latitude to determine how crime frequency varies with location. Pearson's correlation test was used to assess the strength and direction of the relationship.

```
> #___QUESTION 2___
> #Is there a significant correlation between the latitude of a crime incident and the frequency of crimes in th
at area?
>
> # Aggregating data by latitude
> latitude_crime_frequency <- crime_dataset %>%
+   group_by(lat) %>%
+   summarize(frequency = n())
>
> # Pearson's correlation test
> cor_test_result <- cor.test(latitude_crime_frequency$lat, latitude_crime_frequency$frequency, method = "pearso
n")
>
> # Output the result of the correlation test
> print(cor_test_result)
```

Results:

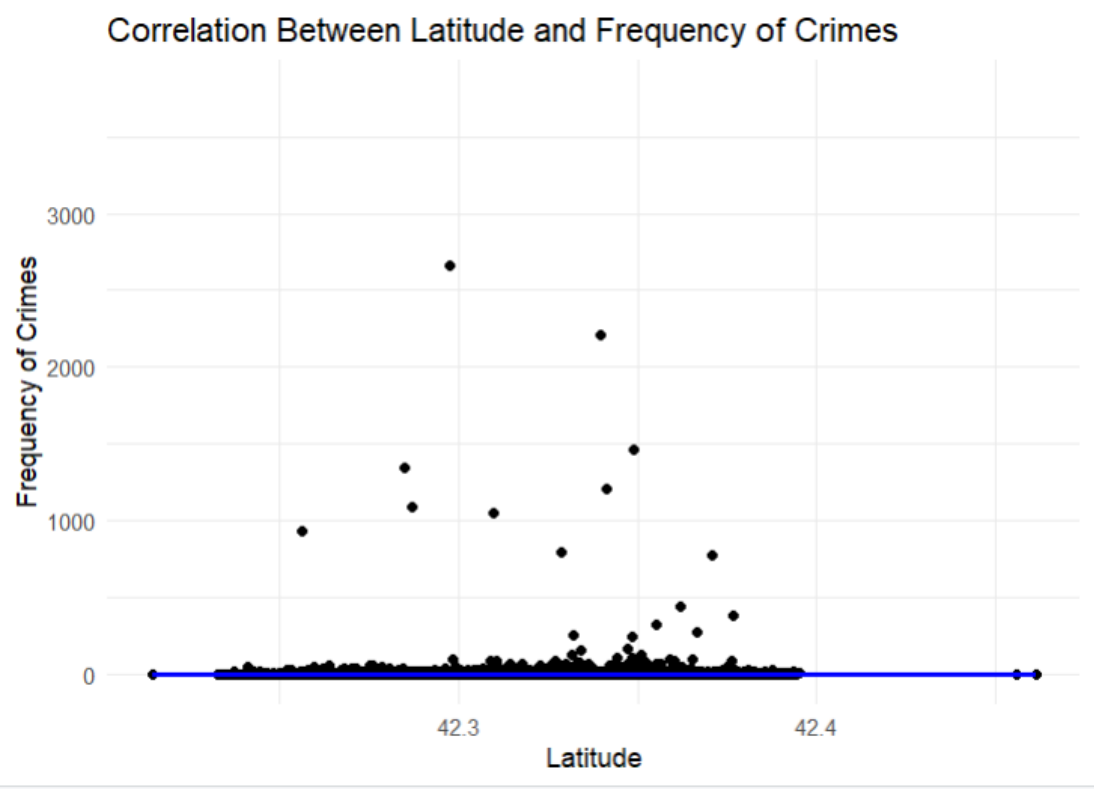
```
        Pearson's product-moment correlation

data:  latitude_crime_frequency$lat and latitude_crime_frequency$frequency
t = 0.92448, df = 12540, p-value = 0.3553
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.009247437  0.025752938
sample estimates:
        cor
0.008255279
```

The correlation coefficient (cor) is approximately 0.0083, which is very close to zero. This indicates a very weak linear relationship between latitude and frequency of crimes.The p-value is 0.3553, which is much higher than the commonly used significance level of 0.05. This high p-value suggests that the observed correlation is not statistically significant, meaning that any linear relationship between latitude and frequency of crimes in the data could be due to random chance rather than an actual association.

Interpretation:

The analysis suggests that there is no significant correlation between latitude and crime frequency. The high p-value supports maintaining the null hypothesis, indicating that the latitude of a crime incident does not show a measurable impact on the frequency of crimes in that area.

Visualization:



The visualization is a scatter plot with the title "Correlation Between Latitude and Frequency of Crimes". The X-axis represents latitude, and the Y-axis represents the frequency of crimes. The scatterplot visualizes each point's latitude against the frequency of crimes. The scatter plot shows a large cluster of points near the bottom of the graph, indicating that for most latitudes in the data set, the frequency of crimes is low. There are a few outliers with much higher crime frequencies. However, there is no clear upward or downward trend visible, suggesting that there may not be a strong linear relationship between latitude and crime frequency. The relatively flat regression line and the spread of points suggest no strong correlation between latitude and crime frequency, aligning with the statistical findings.

Question 3:

Does the frequency of total crimes significantly vary across different months of the year?

Variables:

- Dependent Variable: Total frequency of crimes(frequency)
- Independent Variable: Month of the year(month)

Hypothesis Testing:

1. **Null Hypothesis (H0):** The frequency of total crimes does not significantly vary across different months of the year. This suggests that the month has no effect on the overall crime rate.
2. **Alternative Hypothesis (H1):** The frequency of total crimes significantly varies across different months of the year. This implies that certain months may have a higher or lower crime rate compared to others.

A correlation test was conducted to analyze the relationship between the month and crime frequency.

Methodology:

Crime data was grouped by month, and each month was treated as a numerical variable to assess potential patterns

```
>
> # Creating a bar chart to visually represent crime frequency by month
> ggplot(monthly_crime_frequency, aes(x = factor(month), y = frequency)) +
+   geom_bar(stat = "identity", fill = "blue") +
+   labs(title = "Total Crime Frequency by Month",
+        x = "Month",
+        y = "Frequency of Total Crimes") +
+   theme_minimal()
> # Aggregating data by month
> monthly_crime_frequency <- crime_dataset %>%
+   group_by(month) %>%
+   summarize(frequency = n())
>
> # Convert month to a numerical variable
> monthly_crime_frequency$month_num <- as.numeric(as.factor(monthly_crime_frequency$month))
>
> # Correlation test
> cor_test_result <- cor.test(monthly_crime_frequency$month_num, monthly_crime_frequency$frequency)
>
> # Output the result of the correlation test
> print(cor_test_result)
```

Results:

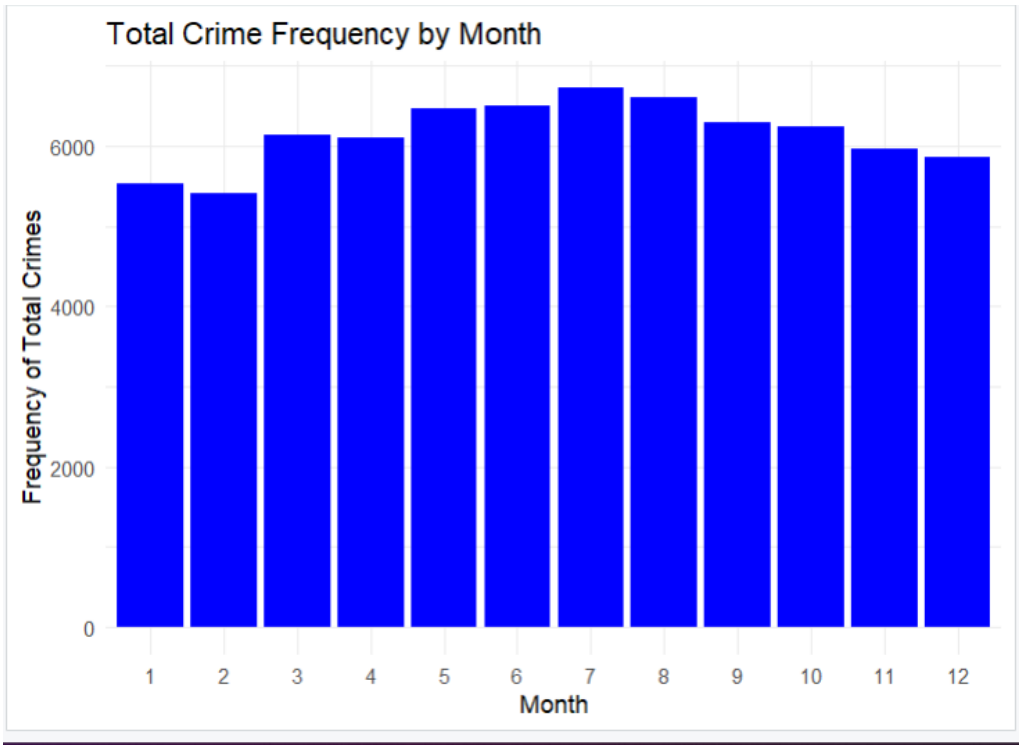
```
        Pearson's product-moment correlation

data:  monthly_crime_frequency$month_num and monthly_crime_frequency$frequency
t = 1.1232, df = 10, p-value = 0.2876
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2960781  0.7621929
sample estimates:
      cor
0.3346948
```

The correlation coefficient (cor) is 0.3346498, indicating a moderate positive linear relationship between the month number and the frequency of crimes. This means that as the month number increases, there tends to be a slight increase in crime frequency. The p-value is 0.2876, which is above the common alpha level of 0.05 used to determine statistical significance. Therefore, we would not reject the null hypothesis and conclude that there is no statistically significant linear relationship between the month and crime frequency based on the data provided.

Interpretation:

The data indicates that there is no significant variation in crime frequency across different months. The statistical output suggests a moderate positive relationship, but the p-value indicates that this relationship is not statistically significant. The higher p-value leads us to retain the null hypothesis, suggesting that the month of the year does not have a measurable influence on the overall frequency of crimes.

Visualization:



The image shows a bar chart titled "Total Crime Frequency by Month," with the X-axis representing the month of the year (from 1 to 12) and the Y-axis representing the frequency of total crimes.The bar chart displays the total number of crimes for each month. The heights of the bars appear relatively consistent, with slight variations, suggesting that the total crime frequency does not drastically change from month to month. There is no clear trend showing a significant increase or decrease over the months.

Conclusion

In Conclusion, through detailed exploratory data analysis and hypothesis testing, we investigated the relationships between crime occurrence and various factors such as time of day, geographical coordinates, and monthly trends. Our objective was to not only interpret these relationships but also to provide actionable insights that could potentially guide public safety measures and law enforcement strategies. Our investigation revealed a significant trend where the frequency of violent crimes increases throughout the day, suggesting targeted approaches for crime prevention during specific hours might be beneficial. However, the analysis did not find substantial evidence to link crime frequency with latitude or distinct monthly patterns. This highlights the intricate nature of crime incidents and suggests that while some factors like time of day can inform preventive tactics, other aspects such as geographical and monthly variations may require more nuanced investigation.

Reference

- Dataset link: https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system
- McNulty, K. (2021). Handbook of Regression Modeling in People Analytics: With Examples in R and Python. United States: CRC Press.
- Harrell, F. E. (2013). Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. United States: Springer New York.
- Matloff, N. (2017). Statistical Regression and Classification: From Linear Models to MachineLearning. United Kingdom: CRC Press.