LA Crime Analysis

Arun Mathew

Data Preparation

Loading Packages

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(readr)
library(lubridate)
library(stringr)
```

Loading Dataset

```
crime_data=read_csv("https://media.githubusercontent.com/media/ArunMathew77777/LA-Crime-analysis/main/C
colnames(crime_data) <- c('dr_no','date_rep','date_occ','time_occ','area','area_name','rep_dis_no','part
                          'crime_code','crime_desc','mocodes','victim_age',
                          'victim_sex','victim_descent','premise_cd','premise_desc',
                          'weapon_cd','weapon_desc','status','status_desc','crime_cd_1',
                          'crime_cd_2','crime_cd_3','crime_cd_4','location','cross_street','lat','lon')
glimpse(crime_data)
```

```
## Rows: 901,357
## Columns: 28
## $ dr_no          <dbl> 190326475, 200106753, 200320258, 200907217, 220614831, ~
## $ date_rep       <chr> "03/01/2020 12:00:00 AM", "02/09/2020 12:00:00 AM", "11~
## $ date_occ       <chr> "03/01/2020 12:00:00 AM", "02/08/2020 12:00:00 AM", "11~
## $ time_occ       <chr> "2130", "1800", "1700", "2037", "1200", "2300", "0900",~
## $ area           <chr> "07", "01", "03", "09", "06", "18", "01", "03", "13", "~
## $ area_name      <chr> "Wilshire", "Central", "Southwest", "Van Nuys", "Hollyw~
## $ rep_dis_no     <chr> "0784", "0182", "0356", "0964", "0666", "1826", "0182",~
## $ part_1_2       <dbl> 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 1, 2, 2, 1, 2, 2, 2, 2~
## $ crime_code     <dbl> 510, 330, 480, 343, 354, 354, 354, 354, 354, 624, 354, ~
```

```
## $ crime_desc     <chr> "VEHICLE - STOLEN", "BURGLARY FROM VEHICLE", "BIKE - ST~
## $ mocodes        <chr> NA, "1822 1402 0344", "0344 1251", "0325 1501", "1822 1~
## $ victim_age     <dbl> 0, 47, 19, 19, 28, 41, 25, 27, 24, 26, 26, 8, 7, 8, 0, ~
## $ victim_sex     <chr> "M", "M", "X", "M", "M", "M", "M", "F", "F", "M", "M", ~
## $ victim_descent <chr> "O", "O", "X", "O", "H", "H", "H", "B", "B", "H", "B", ~
## $ premise_cd     <dbl> 101, 128, 502, 405, 102, 501, 502, 248, 750, 502, 501, ~
## $ premise_desc   <chr> "STREET", "BUS STOP/LAYOVER (ALSO QUERY 124)", "MULTI-U~
## $ weapon_cd      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, 400, NA, 400, 400, ~
## $ weapon_desc    <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, "STRONG-ARM (HANDS,~
## $ status         <chr> "AA", "IC", "IC", "IC", "IC", "IC", "IC", "IC", "IC", "~
## $ status_desc    <chr> "Adult Arrest", "Invest Cont", "Invest Cont", "Invest C~
## $ crime_cd_1     <dbl> 510, 330, 480, 343, 354, 354, 354, 354, 354, 624, 354, ~
## $ crime_cd_2     <dbl> 998, 998, NA, NA, NA, NA, NA, NA, NA, NA, NA, 821, 860,~
## $ crime_cd_3     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ crime_cd_4     <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ location       <chr> "1900 S  LONGWOOD                      AV", "1000 S  FLO~
## $ cross_street   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ lat            <dbl> 34.0375, 34.0444, 34.0210, 34.1576, 34.0944, 33.9467, 3~
## $ lon            <dbl> -118.3506, -118.2628, -118.3002, -118.4387, -118.3277, ~
```

Introduction

Crime is something that many people worry about on a daily basis. Whether it's ensuring your door is locked when you leave the house, avoiding a rough Neighbourhood or be Vigiliant or installing a security alarm prevention of crime takes up a significant part of our lives.

The most important aspect of our lives is 'safety'.My interest in this dataset stems from its potential to analyze crime trends, identify vulnerable victims, discern crime timings, and track crime occurrences over time. Such analyses can raise awareness among the public and enable law enforcement to remain vigilant and alert in specific locations . This dataset reflects incidents of Crime in the City of Los Angeles dating back to 2020.This data is transcribed from original crime reports that are typed on paper,With columns detailing 'Report_Date' capturing incident dates, 'Age_group', providing demographic insights, 'Crime_Type' showcasing various crime categories like theft, assault, vandalism, and 'Citywide_count' quantifying the occurrences of each crime, this dataset serves as a pivotal resource for understanding crime patterns. Analyzing trends over time allows for a nuanced comprehension of crime prevalence across different age demographics and crime types. This dataset's depth enables investigations into factors influencing criminal activities and aids in the development of strategic interventions to address community safety concerns.

Data

Cases

This Dataset consists of 820599 Rows(Use Cases) and 28 Columns before cleaning.

Variables

AREA NAME, Part 1-2, Crm Cd Desc, Mocodes, Premis Desc, Weapon Desc, Status Desc, LOCATION, Cross Street, Vict Sex, Vict Descent, DR_NO, DATE OCC, TIME OCC, AREA, Rpt Dist No, Crm Cd, Weapon Used Cd, Crm Cd 1, Crm Cd 2, Crm Cd 3, Crm Cd 4 (assuming they are numerical crime codes), LAT, LON, Vict Age, Premis Cd,

Data collection

This is a realtime data from LAPD, which is collected from data.gov

Type of study

Its an Observational study.

Data Source

This data has been collected from data.gov.

https://catalog.data.gov/dataset/crime-data-from-2020-to-present

Describe your variables?

Interpreting the Columns DR_NO: This column likely represents a unique identifier for each incident or report.

Date Rptd: This column appears to represent the date when the incident was reported.

DATE OCC: This column seems to represent the date when the incident occurred.

TIME OCC: This column appears to represent the time when the incident occurred.

AREA: This column represents a numerical code or identifier for a specific area.

AREA NAME: This column represents the name or description of the area.

Rpt Dist No: This column represents a numerical code or identifier for the reporting district.

Part 1-2: This column's meaning is not clear without additional context.

Crm Cd: This column represents a numerical code for the crime.

Crm Cd Desc: This column contains the description of the crime corresponding to the Crm Cd.

Mocodes: This column contain additional codes or descriptions related to the mode of operation for the crime.

Premis Desc: This column contains the description of the premise corresponding to the Premis Cd.

Weapon Used Cd: This column represents a numerical code for the weapon used in the crime.

Weapon Desc: This column contains the description of the weapon corresponding to the Weapon Used Cd.

Status: This column likely represents the status of the incident or case.

Status Desc: This column contains the description of the status corresponding to the Status.

Crm Cd 1, Crm Cd 2, Crm Cd 3, Crm Cd 4: These columns might represent additional crime codes or identifiers.

LOCATION: This column contains the location or address of the incident.

Cross Street: This column might contain information about a nearby cross street or location.

LAT and LON: These columns likely represent latitude and longitude coordinates related to the incident location.

Vict Age: This column represents the age of the victim.

Vict Sex: This column represents the gender of the victim.

Vict Descent: This column likely represents the ethnicity or descent of the victim.

Premis Cd: This column represents a numerical code for the premise where the crime occurred.

If you are are running a regression or similar model, which one is your dependent variable?

TIME OCC will be our dependent variable.

Relevant summary statistics

```
head(crime_data)
```

```
## # A tibble: 6 x 28
##       dr_no date_rep       date_occ time_occ area  area_name rep_dis_no part_1_2
##       <dbl> <chr>          <chr>    <chr>    <chr> <chr>     <chr>         <dbl>
## 1 190326475 03/01/2020 12~ 03/01/2~ 2130     07    Wilshire  0784              1
## 2 200106753 02/09/2020 12~ 02/08/2~ 1800     01    Central   0182              1
## 3 200320258 11/11/2020 12~ 11/04/2~ 1700     03    Southwest 0356              1
## 4 200907217 05/10/2023 12~ 03/10/2~ 2037     09    Van Nuys  0964              1
## 5 220614831 08/18/2022 12~ 08/17/2~ 1200     06    Hollywood 0666              2
## 6 231808869 04/04/2023 12~ 12/01/2~ 2300     18    Southeast 1826              2
## # i 20 more variables: crime_code <dbl>, crime_desc <chr>, mocodes <chr>,
## #   victim_age <dbl>, victim_sex <chr>, victim_descent <chr>, premise_cd <dbl>,
## #   premise_desc <chr>, weapon_cd <dbl>, weapon_desc <chr>, status <chr>,
## #   status_desc <chr>, crime_cd_1 <dbl>, crime_cd_2 <dbl>, crime_cd_3 <dbl>,
## #   crime_cd_4 <lgl>, location <chr>, cross_street <chr>, lat <dbl>, lon <dbl>
```

Summary Statistics

```
summary(crime_data)
```

```
##      dr_no             date_rep           date_occ           time_occ
##  Min.   :      817   Length:901357      Length:901357      Length:901357
##  1st Qu.:210320927   Class :character   Class :character   Class :character
##  Median :220416646   Mode  :character   Mode  :character   Mode  :character
##  Mean   :217818489
##  3rd Qu.:230320234
##  Max.   :249904551
##
##      area            area_name          rep_dis_no          part_1_2
##  Length:901357      Length:901357      Length:901357      Min.   :1.000
##  Class :character   Class :character   Class :character   1st Qu.:1.000
##  Mode  :character   Mode  :character   Mode  :character   Median :1.000
##                                                           Mean   :1.411
##                                                           3rd Qu.:2.000
##                                                           Max.   :2.000
##
##    crime_code     crime_desc          mocodes            victim_age
##  Min.   :110.0   Length:901357      Length:901357      Min.   : -4.00
##  1st Qu.:331.0   Class :character   Class :character   1st Qu.:  0.00
##  Median :442.0   Mode  :character   Mode  :character   Median : 31.00
##  Mean   :500.8                                         Mean   : 29.63
```

```
##  3rd Qu.:626.0                                          3rd Qu.: 45.00
##  Max.   :956.0                                          Max.   :120.00
##
##   victim_sex        victim_descent       premise_cd    premise_desc
##  Length:901357      Length:901357      Min.   :101.0   Length:901357
##  Class :character   Class :character   1st Qu.:101.0   Class :character
##  Mode  :character   Mode  :character   Median :203.0   Mode  :character
##                                        Mean   :306.5
##                                        3rd Qu.:501.0
##                                        Max.   :976.0
##                                        NA's   :10
##    weapon_cd      weapon_desc           status          status_desc
##  Min.   :101.0    Length:901357      Length:901357      Length:901357
##  1st Qu.:310.0    Class :character   Class :character   Class :character
##  Median :400.0    Mode  :character   Mode  :character   Mode  :character
##  Mean   :363.4
##  3rd Qu.:400.0
##  Max.   :516.0
##  NA's   :589089
##    crime_cd_1      crime_cd_2       crime_cd_3     crime_cd_4
##  Min.   :110.0   Min.   :210.0    Min.   :310.0    Mode:logical
##  1st Qu.:331.0   1st Qu.:998.0    1st Qu.:998.0    NA's:901357
##  Median :442.0   Median :998.0    Median :998.0
##  Mean   :500.6   Mean   :957.9    Mean   :983.9
##  3rd Qu.:626.0   3rd Qu.:998.0    3rd Qu.:998.0
##  Max.   :956.0   Max.   :999.0    Max.   :999.0
##  NA's   :11      NA's   :835674   NA's   :899144
##    location         cross_street            lat             lon
##  Length:901357      Length:901357      Min.   : 0.00   Min.   :-118.7
##  Class :character   Class :character   1st Qu.:34.01   1st Qu.:-118.4
##  Mode  :character   Mode  :character   Median :34.06   Median :-118.3
##                                        Mean   :33.99   Mean   :-118.1
##                                        3rd Qu.:34.16   3rd Qu.:-118.3
##                                        Max.   :34.33   Max.   :   0.0
##
```

changing the format of date_rep and date_occ from chr to dttm so that we can extract day of the week, month, year.

```
crime_data$date_occ=mdy_hms(crime_data$date_occ)
crime_data$date_rep=mdy_hms(crime_data$date_rep)
```

creating new columns Month, Year, Day of the Week and Month_Year which will contain the data of Month and Year in which the crime was committed.

```
crime_data$month=format(as.Date(crime_data$date_occ),'%b')
crime_data$year=format(as.Date(crime_data$date_occ),'%Y')
crime_data$day=format(as.Date(crime_data$date_occ),'%d')
crime_data$day_of_week=format(as.Date(crime_data$date_occ),'%A')
crime_data$month_year=format(as.Date(crime_data$date_occ),'%b %Y')
glimpse(crime_data)
```

```
## Rows: 901,357
```

```
## Columns: 33
## $ dr_no          <dbl> 190326475, 200106753, 200320258, 200907217, 220614831, ~
## $ date_rep       <dttm> 2020-03-01, 2020-02-09, 2020-11-11, 2023-05-10, 2022-0~
## $ date_occ       <dttm> 2020-03-01, 2020-02-08, 2020-11-04, 2020-03-10, 2020-0~
## $ time_occ       <chr> "2130", "1800", "1700", "2037", "1200", "2300", "0900",~
## $ area           <chr> "07", "01", "03", "09", "06", "18", "01", "03", "13", "~
## $ area_name      <chr> "Wilshire", "Central", "Southwest", "Van Nuys", "Hollyw~
## $ rep_dis_no     <chr> "0784", "0182", "0356", "0964", "0666", "1826", "0182",~
## $ part_1_2       <dbl> 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 1, 2, 2, 1, 2, 2, 2, 2~
## $ crime_code     <dbl> 510, 330, 480, 343, 354, 354, 354, 354, 354, 624, 354, ~
## $ crime_desc     <chr> "VEHICLE - STOLEN", "BURGLARY FROM VEHICLE", "BIKE - ST~
## $ mocodes        <chr> NA, "1822 1402 0344", "0344 1251", "0325 1501", "1822 1~
## $ victim_age     <dbl> 0, 47, 19, 19, 28, 41, 25, 27, 24, 26, 26, 8, 7, 8, 0, ~
## $ victim_sex     <chr> "M", "M", "X", "M", "M", "M", "M", "F", "F", "M", "M", ~
## $ victim_descent <chr> "O", "O", "X", "O", "H", "H", "H", "B", "B", "H", "B", ~
## $ premise_cd     <dbl> 101, 128, 502, 405, 102, 501, 502, 248, 750, 502, 501, ~
## $ premise_desc   <chr> "STREET", "BUS STOP/LAYOVER (ALSO QUERY 124)", "MULTI-U~
## $ weapon_cd      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, 400, NA, 400, 400, ~
## $ weapon_desc    <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, "STRONG-ARM (HANDS,~
## $ status         <chr> "AA", "IC", "IC", "IC", "IC", "IC", "IC", "IC", "IC", "~
## $ status_desc    <chr> "Adult Arrest", "Invest Cont", "Invest Cont", "Invest C~
## $ crime_cd_1     <dbl> 510, 330, 480, 343, 354, 354, 354, 354, 354, 624, 354, ~
## $ crime_cd_2     <dbl> 998, 998, NA, NA, NA, NA, NA, NA, NA, NA, NA, 821, 860,~
## $ crime_cd_3     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ crime_cd_4     <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ location       <chr> "1900 S  LONGWOOD                      AV", "1000 S  FLO~
## $ cross_street   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ lat            <dbl> 34.0375, 34.0444, 34.0210, 34.1576, 34.0944, 33.9467, 3~
## $ lon            <dbl> -118.3506, -118.2628, -118.3002, -118.4387, -118.3277, ~
## $ month          <chr> "Mar", "Feb", "Nov", "Mar", "Aug", "Dec", "Jul", "May",~
## $ year           <chr> "2020", "2020", "2020", "2020", "2020", "2020", "2020",~
## $ day            <chr> "01", "08", "04", "10", "17", "01", "03", "12", "09", "~
## $ day_of_week    <chr> "Sunday", "Saturday", "Wednesday", "Tuesday", "Monday",~
## $ month_year     <chr> "Mar 2020", "Feb 2020", "Nov 2020", "Mar 2020", "Aug 20~
```

Formatting time_occ to contain data in Hours instead of hours and minutes by changing the minutes to 0 in whichever hours they were present in. The reason to do this is that it'll make it easier to analyze data and to find out information specific to a particular hour of the day.

```
crime_data$time_occ=as.integer(crime_data$time_occ)
crime_data$time_occ=as.POSIXct(sprintf("%04.0f", crime_data$time_occ), format='%H')
crime_data$time_occ=as_datetime(crime_data$time_occ)
crime_data$time_occ=format(crime_data$time_occ,'%H:%M')
glimpse(crime_data)
```

```
## Rows: 901,357
## Columns: 33
## $ dr_no          <dbl> 190326475, 200106753, 200320258, 200907217, 220614831, ~
## $ date_rep       <dttm> 2020-03-01, 2020-02-09, 2020-11-11, 2023-05-10, 2022-0~
## $ date_occ       <dttm> 2020-03-01, 2020-02-08, 2020-11-04, 2020-03-10, 2020-0~
## $ time_occ       <chr> "21:00", "18:00", "17:00", "20:00", "12:00", "23:00", "~
## $ area           <chr> "07", "01", "03", "09", "06", "18", "01", "03", "13", "~
## $ area_name      <chr> "Wilshire", "Central", "Southwest", "Van Nuys", "Hollyw~
```

```
## $ rep_dis_no     <chr> "0784", "0182", "0356", "0964", "0666", "1826", "0182",~
## $ part_1_2       <dbl> 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 1, 2, 2, 1, 2, 2, 2, 2~
## $ crime_code     <dbl> 510, 330, 480, 343, 354, 354, 354, 354, 354, 624, 354, ~
## $ crime_desc     <chr> "VEHICLE - STOLEN", "BURGLARY FROM VEHICLE", "BIKE - ST~
## $ mocodes        <chr> NA, "1822 1402 0344", "0344 1251", "0325 1501", "1822 1~
## $ victim_age     <dbl> 0, 47, 19, 19, 28, 41, 25, 27, 24, 26, 26, 8, 7, 8, 0, ~
## $ victim_sex     <chr> "M", "M", "X", "M", "M", "M", "M", "F", "F", "M", "M", ~
## $ victim_descent <chr> "O", "O", "X", "O", "H", "H", "H", "B", "B", "H", "B", ~
## $ premise_cd     <dbl> 101, 128, 502, 405, 102, 501, 502, 248, 750, 502, 501, ~
## $ premise_desc   <chr> "STREET", "BUS STOP/LAYOVER (ALSO QUERY 124)", "MULTI-U~
## $ weapon_cd      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, 400, NA, 400, 400, ~
## $ weapon_desc    <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, "STRONG-ARM (HANDS,~
## $ status         <chr> "AA", "IC", "IC", "IC", "IC", "IC", "IC", "IC", "IC", "~
## $ status_desc    <chr> "Adult Arrest", "Invest Cont", "Invest Cont", "Invest C~
## $ crime_cd_1     <dbl> 510, 330, 480, 343, 354, 354, 354, 354, 354, 624, 354, ~
## $ crime_cd_2     <dbl> 998, 998, NA, NA, NA, NA, NA, NA, NA, NA, NA, 821, 860,~
## $ crime_cd_3     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ crime_cd_4     <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ location       <chr> "1900 S  LONGWOOD                      AV", "1000 S  FLO~
## $ cross_street   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ lat            <dbl> 34.0375, 34.0444, 34.0210, 34.1576, 34.0944, 33.9467, 3~
## $ lon            <dbl> -118.3506, -118.2628, -118.3002, -118.4387, -118.3277, ~
## $ month          <chr> "Mar", "Feb", "Nov", "Mar", "Aug", "Dec", "Jul", "May",~
## $ year           <chr> "2020", "2020", "2020", "2020", "2020", "2020", "2020",~
## $ day            <chr> "01", "08", "04", "10", "17", "01", "03", "12", "09", "~
## $ day_of_week    <chr> "Sunday", "Saturday", "Wednesday", "Tuesday", "Monday",~
## $ month_year     <chr> "Mar 2020", "Feb 2020", "Nov 2020", "Mar 2020", "Aug 20~
```

Location column contains data with a lot of unnecessary spaces. To remove them str_squish is used which combines to multiple spaces to only a single one.

```
crime_data$location=str_squish(crime_data$location)
glimpse(crime_data)
```

```
## Rows: 901,357
## Columns: 33
## $ dr_no          <dbl> 190326475, 200106753, 200320258, 200907217, 220614831, ~
## $ date_rep       <dttm> 2020-03-01, 2020-02-09, 2020-11-11, 2023-05-10, 2022-0~
## $ date_occ       <dttm> 2020-03-01, 2020-02-08, 2020-11-04, 2020-03-10, 2020-0~
## $ time_occ       <chr> "21:00", "18:00", "17:00", "20:00", "12:00", "23:00", "~
## $ area           <chr> "07", "01", "03", "09", "06", "18", "01", "03", "13", "~
## $ area_name      <chr> "Wilshire", "Central", "Southwest", "Van Nuys", "Hollyw~
## $ rep_dis_no     <chr> "0784", "0182", "0356", "0964", "0666", "1826", "0182",~
## $ part_1_2       <dbl> 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 1, 2, 2, 1, 2, 2, 2, 2~
## $ crime_code     <dbl> 510, 330, 480, 343, 354, 354, 354, 354, 354, 624, 354, ~
## $ crime_desc     <chr> "VEHICLE - STOLEN", "BURGLARY FROM VEHICLE", "BIKE - ST~
## $ mocodes        <chr> NA, "1822 1402 0344", "0344 1251", "0325 1501", "1822 1~
## $ victim_age     <dbl> 0, 47, 19, 19, 28, 41, 25, 27, 24, 26, 26, 8, 7, 8, 0, ~
## $ victim_sex     <chr> "M", "M", "X", "M", "M", "M", "M", "F", "F", "M", "M", ~
## $ victim_descent <chr> "O", "O", "X", "O", "H", "H", "H", "B", "B", "H", "B", ~
## $ premise_cd     <dbl> 101, 128, 502, 405, 102, 501, 502, 248, 750, 502, 501, ~
## $ premise_desc   <chr> "STREET", "BUS STOP/LAYOVER (ALSO QUERY 124)", "MULTI-U~
## $ weapon_cd      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, 400, NA, 400, 400, ~
```

```
## $ weapon_desc    <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, "STRONG-ARM (HANDS,~
## $ status         <chr> "AA", "IC", "IC", "IC", "IC", "IC", "IC", "IC", "IC", "~
## $ status_desc    <chr> "Adult Arrest", "Invest Cont", "Invest Cont", "Invest C~
## $ crime_cd_1     <dbl> 510, 330, 480, 343, 354, 354, 354, 354, 354, 624, 354, ~
## $ crime_cd_2     <dbl> 998, 998, NA, NA, NA, NA, NA, NA, NA, NA, NA, 821, 860,~
## $ crime_cd_3     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ crime_cd_4     <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ location       <chr> "1900 S LONGWOOD AV", "1000 S FLOWER ST", "1400 W 37TH ~
## $ cross_street   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ lat            <dbl> 34.0375, 34.0444, 34.0210, 34.1576, 34.0944, 33.9467, 3~
## $ lon            <dbl> -118.3506, -118.2628, -118.3002, -118.4387, -118.3277, ~
## $ month          <chr> "Mar", "Feb", "Nov", "Mar", "Aug", "Dec", "Jul", "May",~
## $ year           <chr> "2020", "2020", "2020", "2020", "2020", "2020", "2020",~
## $ day            <chr> "01", "08", "04", "10", "17", "01", "03", "12", "09", "~
## $ day_of_week    <chr> "Sunday", "Saturday", "Wednesday", "Tuesday", "Monday",~
## $ month_year     <chr> "Mar 2020", "Feb 2020", "Nov 2020", "Mar 2020", "Aug 20~
```

Exploratory Data Analysis

1) looking at what hour of the day crimes are committed the most

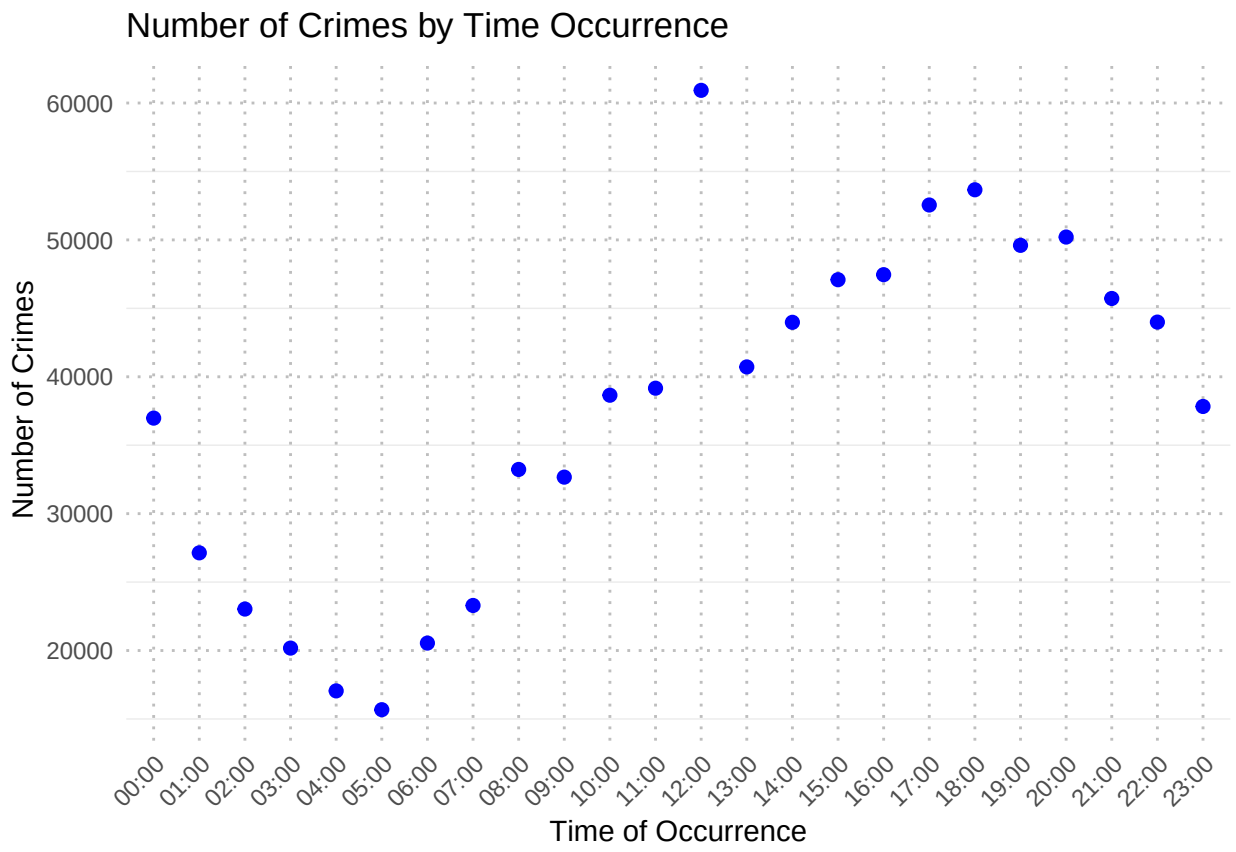
```
crime_time= crime_data%>%
group_by(time_occ)%>%
count(time_occ)

crime_time
```

```
## # A tibble: 24 x 2
## # Groups:   time_occ [24]
##    time_occ     n
##    <chr>    <int>
##  1 00:00    36980
##  2 01:00    27137
##  3 02:00    23031
##  4 03:00    20178
##  5 04:00    17052
##  6 05:00    15677
##  7 06:00    20544
##  8 07:00    23292
##  9 08:00    33230
## 10 09:00    32668
## # i 14 more rows
```

```
ggplot(crime_time, aes(x = time_occ, y = n)) +
  geom_point(size = 2, color = "blue") +  # Increase point size and change color
  geom_smooth(method = "loess", se = FALSE, color = "red") +  # Smooth the line
  labs(title = "Number of Crimes by Time Occurrence",
       x = "Time of Occurrence",
       y = "Number of Crimes")  +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),  # Rotate x-axis labels
        panel.grid.major = element_line(color = "gray", linetype = "dotted"))  # Add gridlines
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



From the graph, we can observe certain patterns:

Peak Hours: There appears to be variations in the frequency of crimes throughout the day. Certain time intervals exhibit higher crime rates compared to others. For example, there might be peaks during evening or early morning hours when criminal activities tend to be more prevalent.

Trend Analysis: The line connecting the points helps in identifying any overall trends in crime occurrence throughout the day. For instance, if the line slopes upwards towards the evening or early morning hours, it suggests that crime rates tend to increase during those times.

Identifying Hotspots: Certain time intervals may stand out as hotspots for criminal activities. These intervals might warrant closer attention from law enforcement agencies or community stakeholders to implement targeted crime prevention strategies during those times.

2) Finding the day on which most crimes are committed.

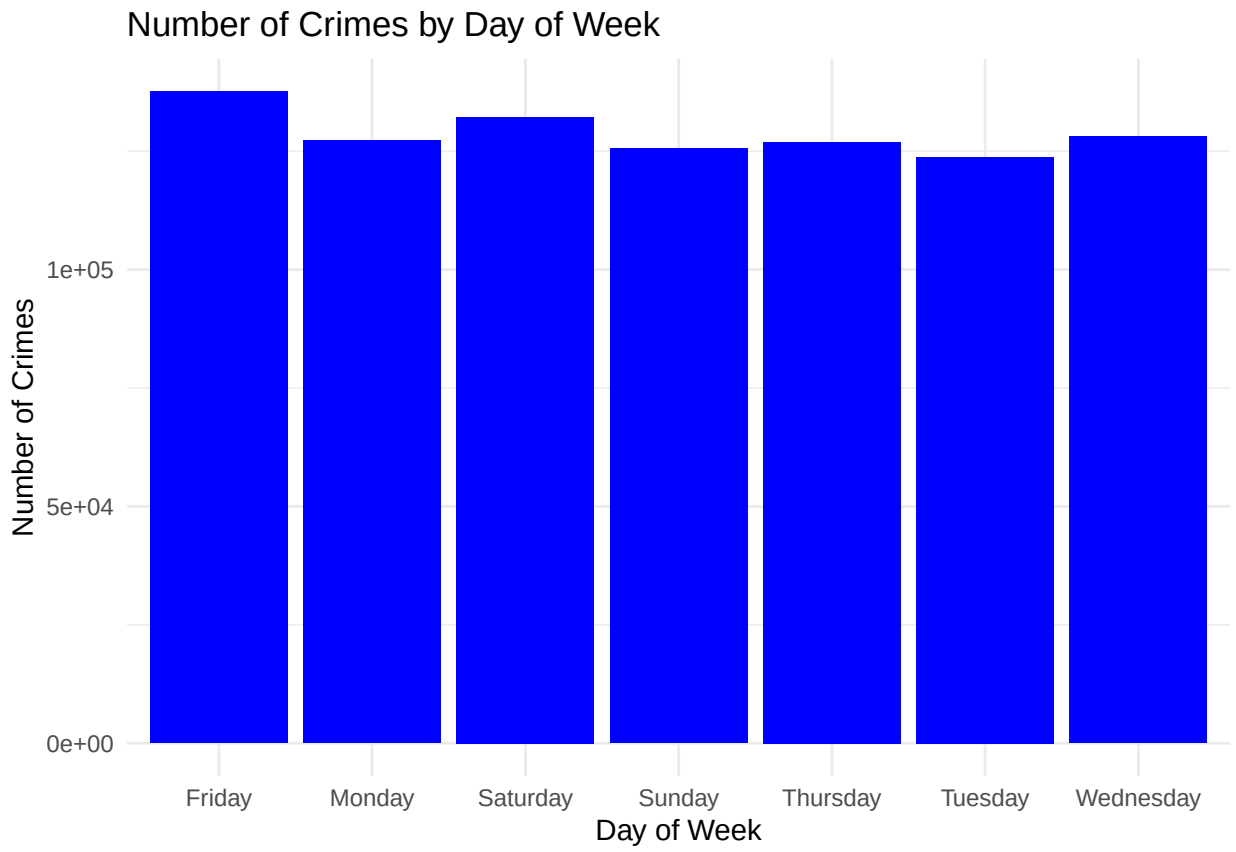
```
crimes_by_day_of_week=crime_data%>%
  group_by(day_of_week)%>%
  count(day_of_week)

crimes_by_day_of_week
```

```
## # A tibble: 7 x 2
## # Groups:   day_of_week [7]
```

```
##   day_of_week      n
##   <chr>        <int>
## 1 Friday      137622
## 2 Monday      127202
## 3 Saturday    132191
## 4 Sunday      125591
## 5 Thursday    126962
## 6 Tuesday     123709
## 7 Wednesday   128080
```

```r
ggplot(crimes_by_day_of_week, aes(x = day_of_week, y = n)) +
  geom_bar(stat = "identity", fill = "blue") +
  labs(title = "Number of Crimes by Day of Week",
       x = "Day of Week",
       y = "Number of Crimes") +
  theme_minimal()
```



Here are some observations from the graph:

Variation Across Days: There is variation in the number of crimes reported throughout the week. Some days exhibit higher crime rates compared to others, as indicated by taller bars in the graph.

Weekday vs. Weekend: There seems to be a noticeable difference between weekdays (Monday to Friday) and weekends (Saturday and Sunday). Generally, weekdays tend to have higher crime rates, possibly due to factors such as increased economic activity, commuter traffic, or routine patterns of criminal behavior.

Patterns and Trends: Analyzing the distribution of crimes across different days can reveal patterns or trends that might inform law enforcement strategies or community interventions. For example, if certain days

consistently experience higher crime rates, targeted policing or prevention efforts could be implemented during those times.

3) Finding the month in which the crimes are committed the most

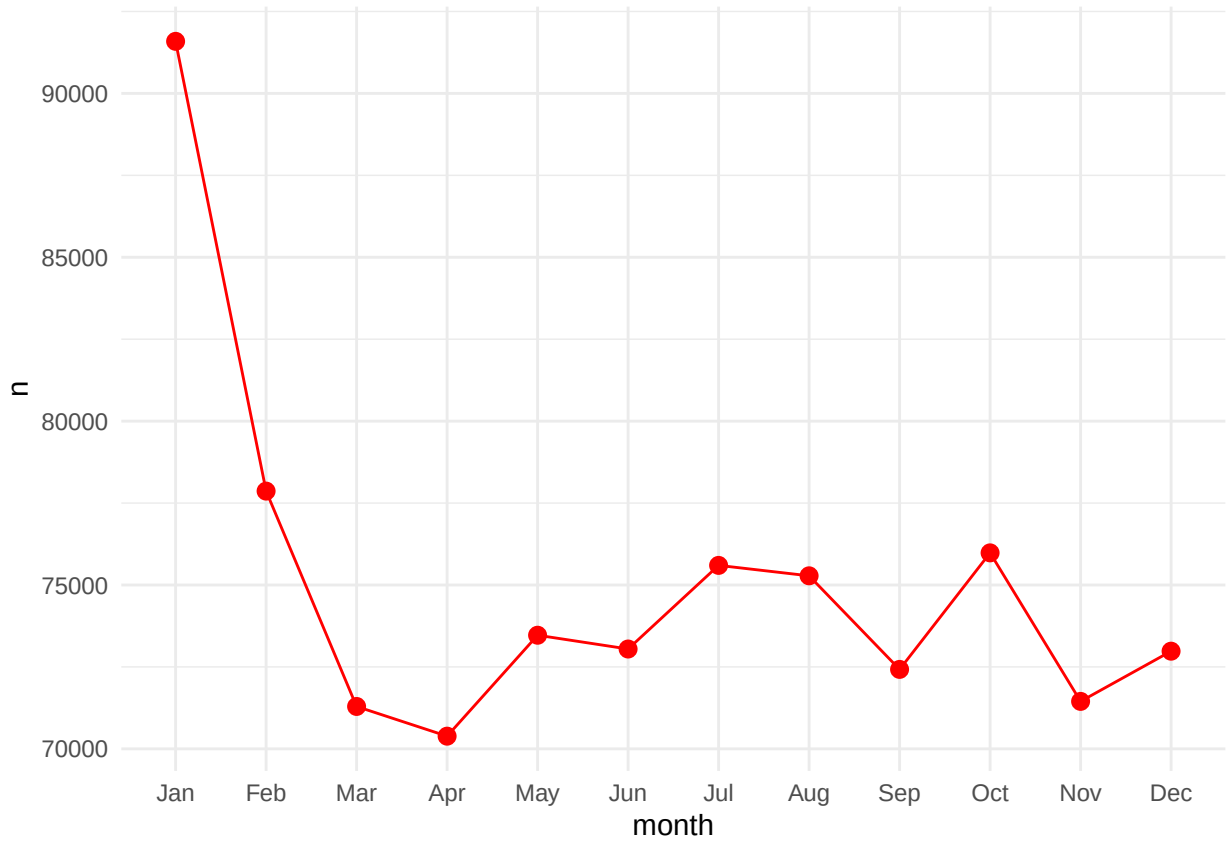
```
crimes_by_month=crime_data%>%
  group_by(month)%>%
  count(month)

crimes_by_month
```

```
## # A tibble: 12 x 2
## # Groups:   month [12]
##    month     n
##    <chr> <int>
##  1 Apr   70387
##  2 Aug   75280
##  3 Dec   72980
##  4 Feb   77865
##  5 Jan   91589
##  6 Jul   75598
##  7 Jun   73046
##  8 Mar   71293
##  9 May   73467
## 10 Nov   71448
## 11 Oct   75980
## 12 Sep   72424
```

```
# Reorder months chronologically
crimes_by_month$month <- factor(crimes_by_month$month, levels = c("Jan", "Feb", "Mar", "Apr", "May", "Ju

# Plot with enhancements
ggplot(crimes_by_month, aes(x = month, y = n, group = 1)) +
  geom_line(color = "red") +
  geom_point(shape = 16, size = 3, color = "red", fill = "red") +  # Change point shape and size
  theme_minimal()
```



Here are some key observations from the graph:

Seasonal Trends: The graph reveals potential seasonal patterns in crime occurrence. There may be months with higher crime rates, possibly associated with factors such as holidays, weather conditions, or social dynamics.

Monthly Fluctuations: The line plot shows fluctuations in crime counts from month to month. Some months exhibit peaks, indicating higher crime rates, while others have troughs, representing lower crime activity.

Identifying Hotspots: Analyzing crime trends by month can help identify potential hotspots or periods of increased criminal activity. Law enforcement agencies and policymakers can use this information to allocate resources effectively and implement targeted interventions in areas or during times of heightened criminal activity.

4) Finding the crime that is committed the most in each area.

```
find_mode <- function(x) {
  u <- unique(x)
  tab <- tabulate(match(x, u))
  u[tab == max(tab)]
}

crime_new=crime_data%>%
  group_by(area_name)%>%
  summarize(crime_desc=find_mode(crime_desc))
```

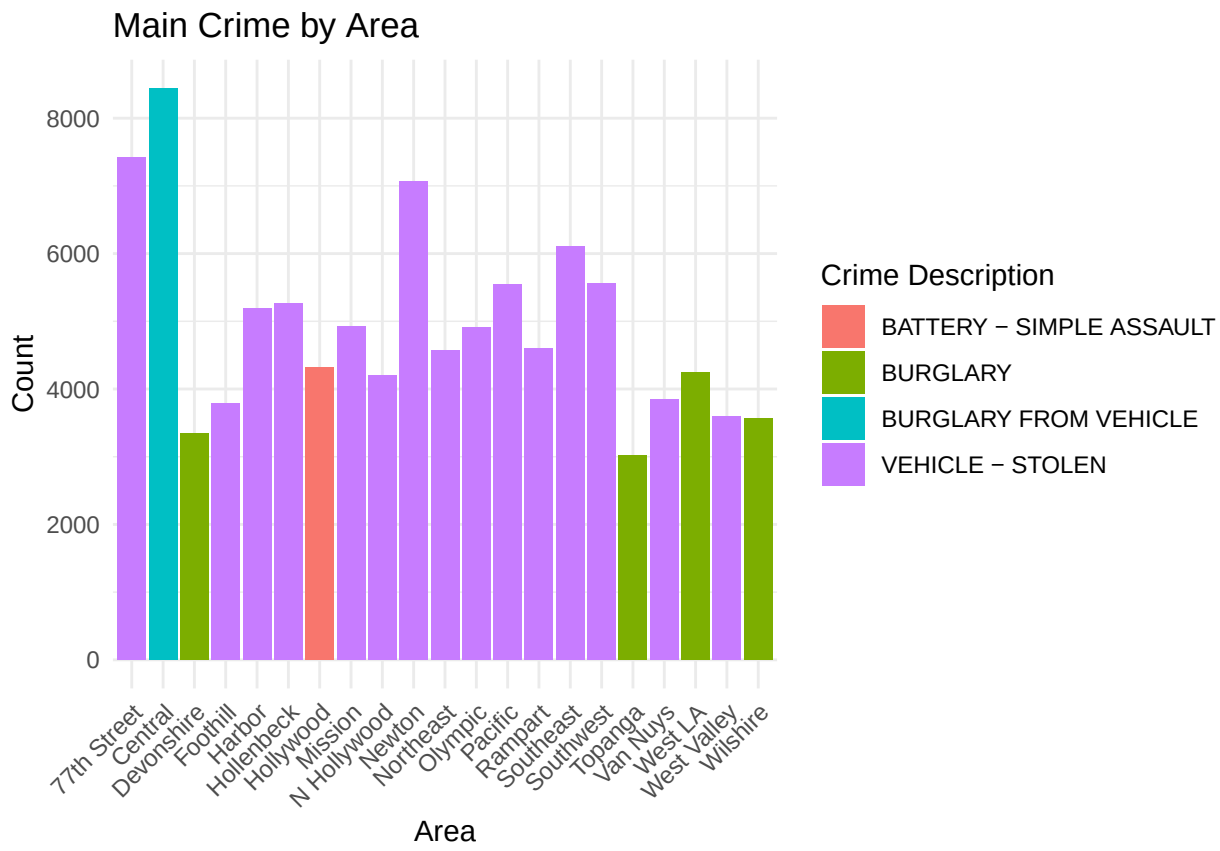
```r
main_crime_by_area=data.frame(area_name=c(),crime_desc=c(),n=c())

for(i in 1:21){
  crime_new_2=crime_data%>%
    filter(crime_data$area_name==crime_new$area_name[i] & crime_data$crime_desc==crime_new$crime_desc[i]
    group_by(area_name)%>%count(crime_desc)
  main_crime_by_area=rbind(main_crime_by_area,crime_new_2)
}

main_crime_by_area
```

```
## # A tibble: 21 x 3
## # Groups:   area_name [21]
##    area_name   crime_desc                  n
##    <chr>       <chr>                   <int>
##  1 77th Street VEHICLE - STOLEN         7428
##  2 Central     BURGLARY FROM VEHICLE    8441
##  3 Devonshire  BURGLARY                 3341
##  4 Foothill    VEHICLE - STOLEN         3795
##  5 Harbor      VEHICLE - STOLEN         5184
##  6 Hollenbeck  VEHICLE - STOLEN         5272
##  7 Hollywood   BATTERY - SIMPLE ASSAULT 4325
##  8 Mission     VEHICLE - STOLEN         4933
##  9 N Hollywood VEHICLE - STOLEN         4204
## 10 Newton      VEHICLE - STOLEN         7072
## # i 11 more rows
```

```r
ggplot(main_crime_by_area, aes(x = area_name, y = n, fill = crime_desc)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Main Crime by Area",
       x = "Area",
       y = "Count",
       fill = "Crime Description") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Here are some key observations from the graph:

Crime Distribution Across Areas: The chart allows us to compare the prevalence of different crime types across various geographic areas. By examining the height of the bars, we can identify which crimes are most common in each area.

Variability in Crime Types: The graph illustrates that certain crime types dominate in specific areas, while others may be less frequent or absent altogether. This variability suggests that crime patterns can vary significantly from one location to another.

Insights for Law Enforcement: Law enforcement agencies can leverage this information to prioritize resources and develop targeted strategies for crime prevention and intervention. Understanding the prevalent crime types in each area enables law enforcement to tailor their approach to address local needs effectively.

Geographic Trends: Analyzing crime distribution by area can also reveal geographic trends and patterns. Identifying areas with higher concentrations of specific crimes may indicate underlying social, economic, or environmental factors contributing to criminal activity.

5) Now we'll look at the count of crimes in each areas.

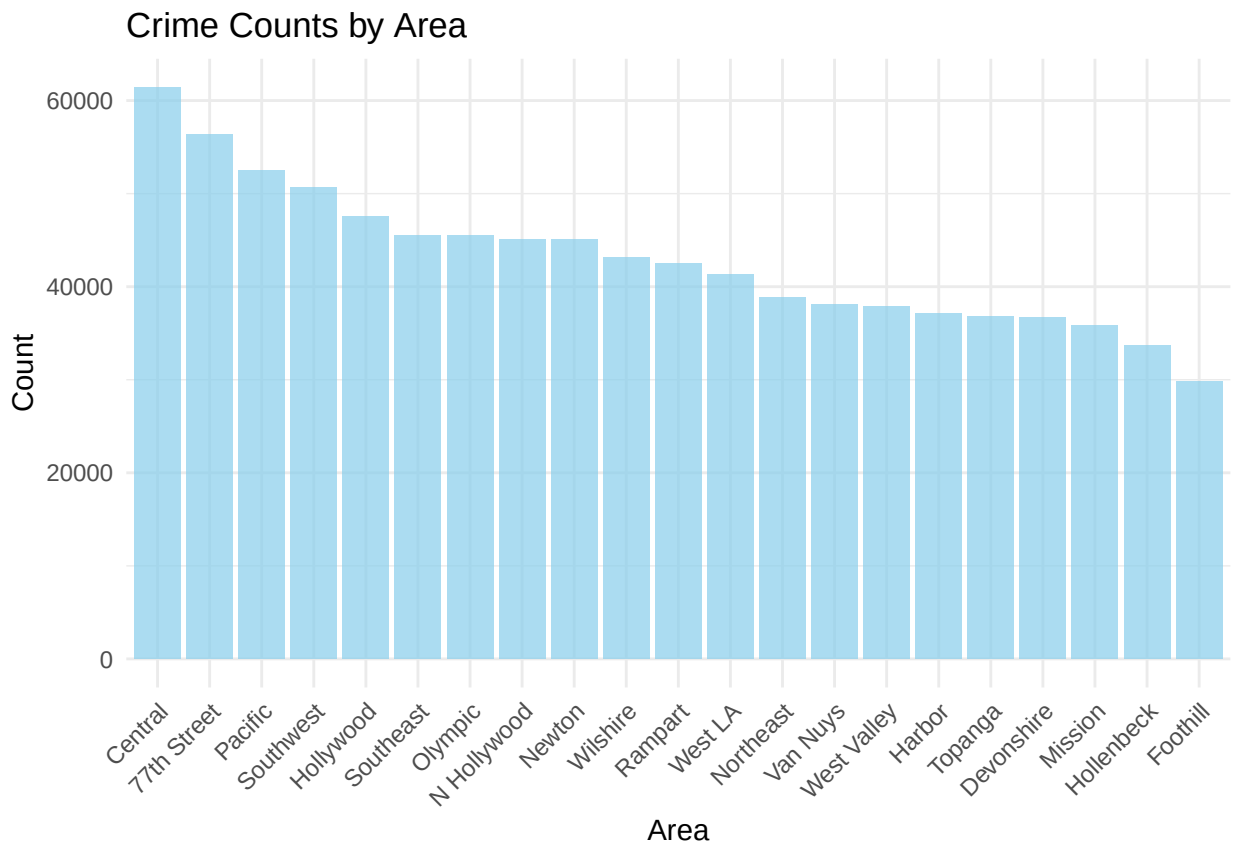
```
area_crimes_count=crime_data%>%
  count(area_name)%>%
  arrange(-n)

area_crimes_count
```

```
## # A tibble: 21 x 2
```

```
##    area_name       n
##    <chr>       <int>
##  1 Central     61416
##  2 77th Street 56381
##  3 Pacific     52537
##  4 Southwest   50654
##  5 Hollywood   47508
##  6 Southeast   45503
##  7 Olympic     45486
##  8 N Hollywood 45124
##  9 Newton      45032
## 10 Wilshire    43191
## # i 11 more rows
```

```
ggplot(area_crimes_count, aes(x = reorder(area_name, -n), y = n)) +
  geom_bar(stat = "identity", fill = "skyblue", alpha = 0.7) +
  labs(title = "Crime Counts by Area",
       x = "Area",
       y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Here are some key insights from the graph:

Variation in Crime Counts: The heights of the bars vary, indicating differences in the number of reported crimes across different areas. Some areas have higher crime counts, while others have relatively lower counts.

High-Crime Areas: The graph highlights areas with the highest reported crime counts, as evidenced by the tallest bars. These areas may require additional attention and resources from law enforcement and community initiatives to address crime prevention and intervention.

Insights for Resource Allocation: By visualizing crime counts by area, stakeholders can identify areas with the greatest need for crime-fighting resources, such as increased patrols, community policing efforts, or targeted intervention programs.

Comparative Analysis: Comparing the heights of the bars allows for a quick assessment of relative crime levels between different areas. Areas with similar crime counts can be further analyzed to identify commonalities or differences contributing to crime rates.

Inference

H0: There is no increase in number of crime as the time goes forward.

HA: There is an increase in number of crime as the time goes forward

Lets try to predict the variation in number of crime by hour

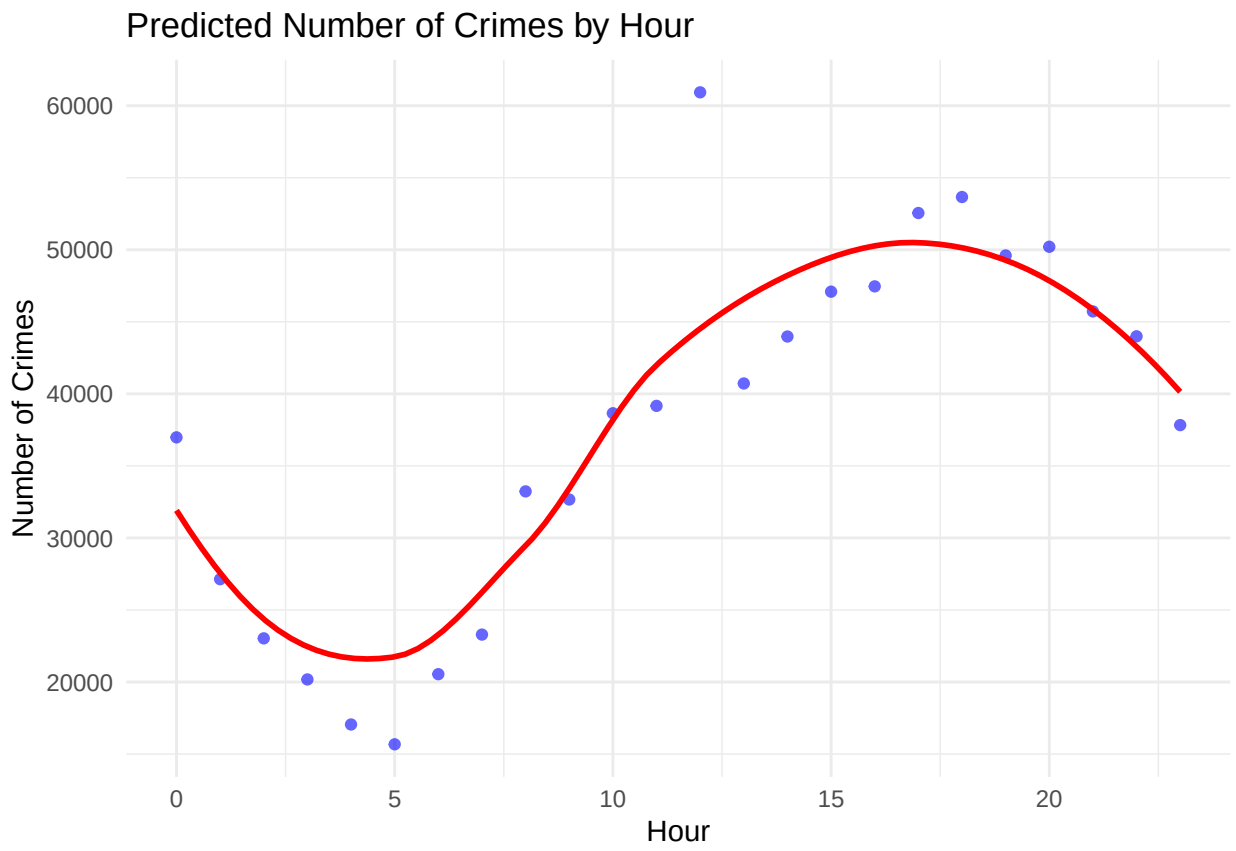
```
# Load necessary libraries
library(dplyr)
library(ggplot2)

# Convert time_occ to numeric hours
crime_data <- crime_data %>%
  mutate(hour = as.numeric(substr(time_occ, 1, 2)))

# Group data by hour and count the number of crimes
crime_by_hour <- crime_data %>%
  group_by(hour) %>%
  summarize(crime_count = n())

# Plot predicted number of crimes by hour using LOESS regression
ggplot(crime_by_hour, aes(x = hour, y = crime_count)) +
  geom_point(color = "blue", alpha = 0.6) +
  geom_smooth(method = "loess", se = FALSE, color = "red") +  # LOESS regression
  labs(title = "Predicted Number of Crimes by Hour",
       x = "Hour",
       y = "Number of Crimes") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# Fit cubic model
lm_cubic <- lm(crime_count ~ poly(hour, 3), data = crime_by_hour)

# Calculate R-squared value
rsquared <- summary(lm_cubic)$r.squared

# Print R-squared value
print(paste("R-squared value:", round(rsquared, 4)))
```

```
## [1] "R-squared value: 0.8354"
```

Prediction of pattern of crime daily from 2020

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

# Convert date_occ to numeric format (to be used as the independent variable)
crime_by_day <- crime_data %>%
  group_by(date_occ) %>%
  summarize(crime_count = n()) %>%
  mutate(date_numeric = as.numeric(date_occ))

# Perform cubic regression
cubic_model <- lm(crime_count ~ poly(date_numeric, 3, raw = TRUE), data = crime_by_day)
```

```r
# Summary of the cubic regression model
summary(cubic_model)
```

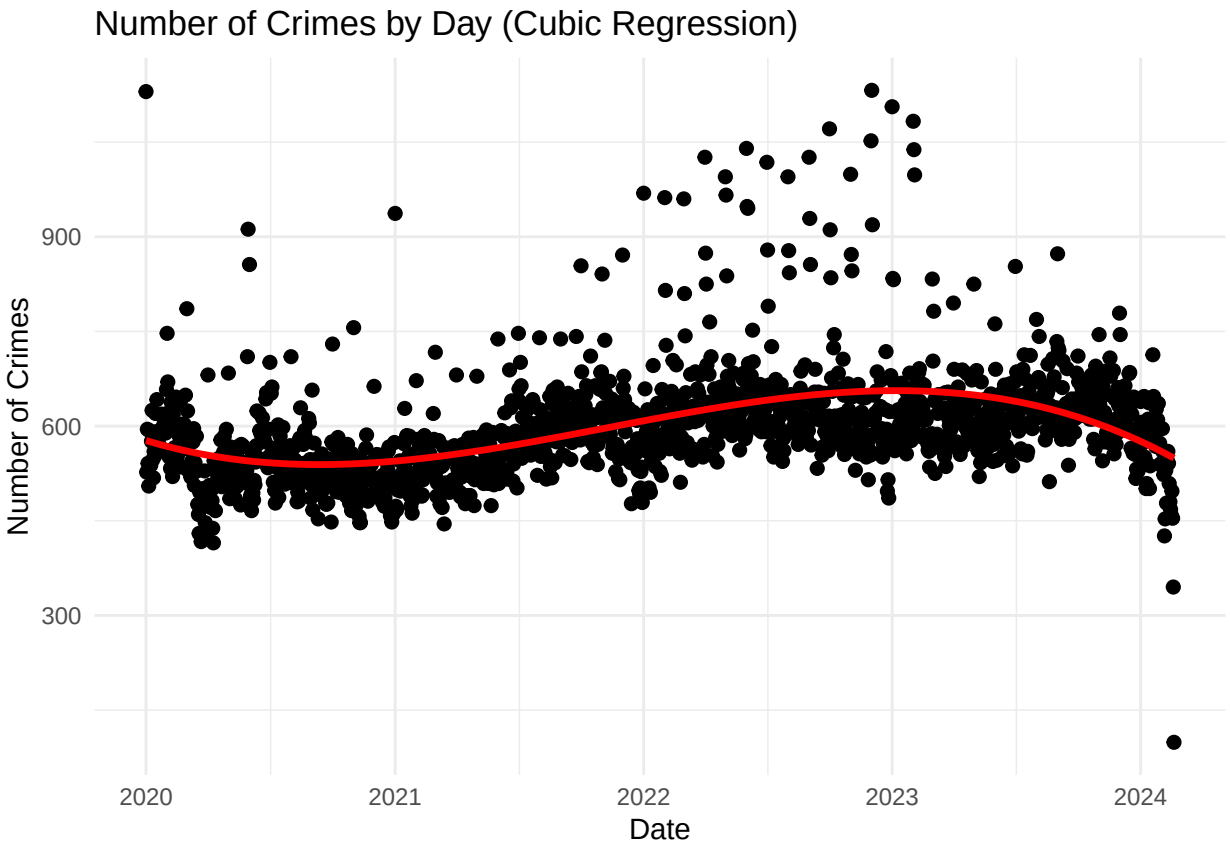
```
##
## Call:
## lm(formula = crime_count ~ poly(date_numeric, 3, raw = TRUE),
##     data = crime_by_day)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -450.24  -43.18   -8.98   22.45  552.85
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         2.606e+06  2.109e+05   12.36   <2e-16 ***
## poly(date_numeric, 3, raw = TRUE)1 -4.781e-03  3.853e-04  -12.41   <2e-16 ***
## poly(date_numeric, 3, raw = TRUE)2  2.923e-12  2.346e-13   12.46   <2e-16 ***
## poly(date_numeric, 3, raw = TRUE)3 -5.954e-22  4.759e-23  -12.51   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 77.78 on 1507 degrees of freedom
## Multiple R-squared:  0.2251, Adjusted R-squared:  0.2236
## F-statistic:   146 on 3 and 1507 DF,  p-value: < 2.2e-16
```

```r
# Generate predicted values using the cubic model
predicted_values <- predict(cubic_model, newdata = list(date_numeric = crime_by_day$date_numeric))

# Create a data frame for plotting
plot_data <- data.frame(date_occ = crime_by_day$date_occ, crime_count = crime_by_day$crime_count, predi

# Plot actual data and regression line
ggplot(plot_data, aes(x = date_occ)) +
  geom_point(aes(y = crime_count), color = "black", size = 2, alpha = 1) +  # Actual data points
  geom_line(aes(y = predicted_count), color = "red", size = 1.2, alpha = 1) +  # Regression line
  labs(title = "Number of Crimes by Day (Cubic Regression)",
       x = "Date",
       y = "Number of Crimes") +
  theme_minimal()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



From the plot, we can observe fluctuations in crime counts over different days. The regression line helps visualize the general trend, showing whether crime counts are increasing, decreasing, or remaining stable over time. Additionally, any notable peaks or dips in the data can be identified, which may correspond to specific events or periods of interest. Overall, the plot provides insight into the temporal dynamics of crime occurrence within the dataset.

Conclusion

The analysis conducted on the variation in crime occurrence by hour and by day provides valuable insights into the temporal patterns of criminal activity.

For the analysis by hour, the linear regression model and plot indicate a clear trend: the number of crimes tends to increase throughout the day, with peaks typically occurring in the late afternoon or evening hours. This suggests that certain times of the day may be more susceptible to criminal activity than others.

Similarly, the analysis by day reveals fluctuations in crime counts over time, with the regression line helping to visualize the overall trend. While there may be variations from day to day, the regression analysis provides insight into whether crime counts are generally increasing, decreasing, or remaining stable over the period examined.

In conclusion, understanding the temporal patterns of crime occurrence can be crucial for law enforcement agencies and policymakers to allocate resources effectively and implement targeted interventions to prevent and address criminal activity. The insights gained from these analyses can inform strategies aimed at enhancing public safety and reducing crime rates in communities.

References

Wikipedia - Crime in Los Angeles:This page provides an overview of crime statistics and trends in Los Angeles, covering historical data, prevalent types of crimes, law enforcement efforts, and initiatives to address crime-related issues.

https://en.wikipedia.org/wiki/Crime_in_Los_Angeles

Data.gov - Crime Data from 2020 to Present:This dataset contains crime-related data from 2020 onwards, likely including information such as types of crimes reported, locations, timestamps, and other relevant details, providing a comprehensive resource for analyzing recent crime statistics.

https://catalog.data.gov/dataset/crime-data-from-2020-to-present