Milestone 1

Group Assignment : Unnati Gaglani, Dia Khosla , Bhoomika Gururaja

College of Professional Studies, Northeastern University
ALY6010: Probability Theory and Introductory Statistics

Instructor - Dr. Selcuk Baran

June 27, 2024

Milestone 1 Report

Introduction :- The dataset Crime_Data provides detailed information about each reported crime incident, including specifics about the incident itself (date, time, location), the nature of the crime (codes and descriptions), details about victims, premises, any weapons involved, and geographical coordinates of the incident location.

I. What is the purpose of the dataset? What is your data source?

Report is on the crime data in the city of Los Angeles :

The analysis focuses on understanding patterns in crime data reported in Los Angeles from 2020 to the present. The dataset used includes information on crime types, occurrence dates, locations, and victim demographics. The data has been picked from Data.gov - City of Los Angeles data.lacity.org.

What kind of data is included? Is it all text data, is it numerical?

The data contains 951608 rows and 28 columns from which it has both textual and numerical value. Here is the structure of the dataset.

Data Types:

- Most fields are either numeric or character (string) data types.
- Numeric fields include incident number, crime codes, age, premises code, etc.
- Character fields include dates, times, area names, crime descriptions, victim information, etc.
- There are also logical (crm_cd_4) and geographical (latitude, longitude) data types present.

Further, the few columns on crm_cd_1", "crm_cd_2", "crm_cd_3", "crm_cd_4", "cross_street", "area" have been removed for data exploration and data cleaning.

Data Preparation and Cleaning

Initially, the dataset underwent cleaning and preparation steps to ensure consistency and usability:

- Data was imported using read_csv() from the readr package.
- Columns were cleaned and standardized using janitor::clean_names().
- Date columns (date_rptd and date_occ) were converted to proper Date format using as.Date().

Summary Statistics

Summary statistics such as mean, median, and quartiles for numeric variables were obtained for the Crime_Data.

```
>   summary(Crime_Data) # finding summary stats
     DR_NO           Date Rptd           DATE OCC           TIME OCC           AREA
Min.   :      817   Length:951608      Length:951608      Length:951608      Length:951608
1st Qu.:210511080   Class :character   Class :character   Class :character   Class :character
Median :220708066   Mode  :character   Mode  :character   Mode  :character   Mode  :character
Mean   :219049633
3rd Qu.:230713397
Max.   :249918669

 AREA NAME         Rpt Dist No          Part 1-2        Crm Cd      Crm Cd Desc         Mocodes
Length:951608      Length:951608      Min.   :1.000   Min.   :110.0   Length:951608      Length:951608
Class :character   Class :character   1st Qu.:1.000   1st Qu.:331.0   Class :character   Class :character
Mode  :character   Mode  :character   Median :1.000   Median :442.0   Mode  :character   Mode  :character
                                      Mean   :1.408   Mean   :500.7
                                      3rd Qu.:2.000   3rd Qu.:626.0
                                      Max.   :2.000   Max.   :956.0

   Vict Age        Vict Sex         Vict Descent         Premis Cd     Premis Desc        Weapon Used Cd
Min.   : -4.00   Length:951608      Length:951608      Min.   :101.0   Length:951608      Min.   :101.0
1st Qu.:  0.00   Class :character   Class :character   1st Qu.:101.0   Class :character   1st Qu.:311.0
Median : 30.00   Mode  :character   Mode  :character   Median :203.0   Mode  :character   Median :400.0
Mean   : 29.42                                        Mean   :306.3                      Mean   :363.7
3rd Qu.: 45.00                                        3rd Qu.:501.0                      3rd Qu.:400.0
Max.   :120.00                                        Max.   :976.0                      Max.   :516.0
                                                      NA's   :11                         NA's   :626761
Weapon Desc           Status          Status Desc           Crm Cd 1        Crm Cd 2        Crm Cd 3
Length:951608      Length:951608      Length:951608      Min.   :110.0   Min.   :210.0   Min.   :310.0
Class :character   Class :character   Class :character   1st Qu.:331.0   1st Qu.:998.0   1st Qu.:998.0
Mode  :character   Mode  :character   Mode  :character   Median :442.0   Median :998.0   Median :998.0
                                                         Mean   :500.5   Mean   :958.1   Mean   :984.1
                                                         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   :883212  NA's   :949318
Crm Cd 4         LOCATION         Cross Street           LAT             LON
Mode:logical   Length:951608      Length:951608      Min.   : 0.00   Min.   :-118.7
NA's:951608    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
```

How many rows of data are there? how many fields?

Number of rows and columns were checked using nrow() and ncol(). Here total rows are 951608 and columns are 28.

Describe any data cleaning you did?

Data Preparation and Cleaning

Columns were cleaned and standardized using janitor::clean_names().

```
library(janitor) # data exploration and cleaning
Crime_Data <- Crime_Data %>% clean_names()
  head(Crime_Data)
  names(Crime_Data)
```

Describe the data fields including the title, the data type, the data description, etc.

dr_no: Incident number (numeric)
date_rptd: Date reported (character, formatted as MM/DD/YYYY)
date_occ: Date occurred (character, formatted as MM/DD/YYYY)

- **time_occ**: Time occurred (character, HHMM)
- **area**: Area code (character)
- **area_name**: Area name (character)
- **rpt_dist_no**: Reporting district number (character)
- **part_1_2**: Part 1 or Part 2 crime indicator (numeric)

- **crm_cd**: Crime code (numeric)
- **crm_cd_desc**: Crime code description (character)
- **mocodes**: Modus operandi codes (character)
- **vict_age**: Victim's age (numeric)
- **vict_sex**: Victim's sex (character)
- **vict_descent**: Victim's descent (character)
- **premis_cd**: Premises code (numeric)
- **premis_desc**: Premises description (character)
- **weapon_used_cd**: Weapon used code (numeric)
- **weapon_desc**: Weapon description (character)
- **status**: Status of the report (character)
- **status_desc**: Status description (character)
- **crm_cd_1, crm_cd_2, crm_cd_3, crm_cd_4**: Additional crime codes (numeric)
- **location**: Location description (character)
- **cross_street**: Cross street (character)
- **lat**: Latitude (numeric)
- **lon**: Longitude (numeric)

```
>   names(Crime_Data)
 [1] "dr_no"          "date_rptd"      "date_occ"       "time_occ"       "area"           "area_name"
 [7] "rpt_dist_no"    "part_1_2"       "crm_cd"         "crm_cd_desc"    "mocodes"        "vict_age"
[13] "vict_sex"       "vict_descent"   "premis_cd"      "premis_desc"    "weapon_used_cd" "weapon_desc"
[19] "status"         "status_desc"    "crm_cd_1"       "crm_cd_2"       "crm_cd_3"       "crm_cd_4"
[25] "location"       "cross_street"   "lat"            "lon"
>
```

Date columns (date_rptd and date_occ) were converted to Date format (as.Date() function).

Date Rptd	DATE OCC
03/01/2020 12:00:00 AM	03/01/2020 12:00:00 AM
02/09/2020 12:00:00 AM	02/08/2020 12:00:00 AM
11/11/2020 12:00:00 AM	11/04/2020 12:00:00 AM
05/10/2023 12:00:00 AM	03/10/2020 12:00:00 AM

date_rptd	date_occ
2020-03-01	2020-03-01
2020-02-09	2020-02-08
2020-11-11	2020-11-04
2023-05-10	2020-03-10
2022-08-18	2020-08-17
2023-04-04	2020-12-01

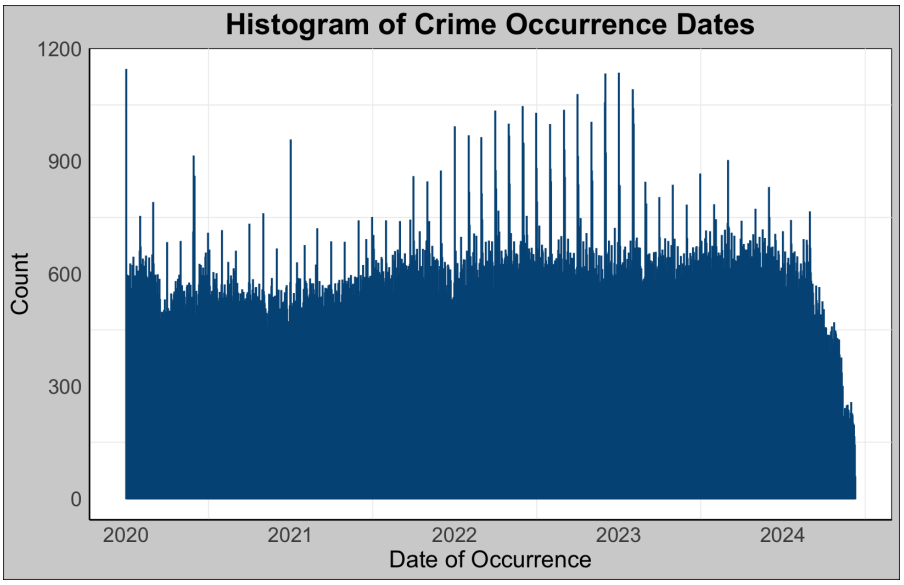
II. Data Analysis: Top Crimes by Activity Counts :-

Provide visualizations of the key data and subset data of interest. This should be done for categorical data, discrete data and continuous data.

Crime Frequency Analysis :-

A table and a bar plot were generated to visualize the top 20 crimes by activity count. The histogram illustrates the distribution of crime occurrences across different years in Los Angeles. From the year

2020, till 2023 Crimes rates have persisted to have an increasing trend, whereas from year 2023 to 2024, crimes rate show significant downfall trend , meaning crimes rates started to decrease From year 2023.



Heatmap and Bar Plot Creation

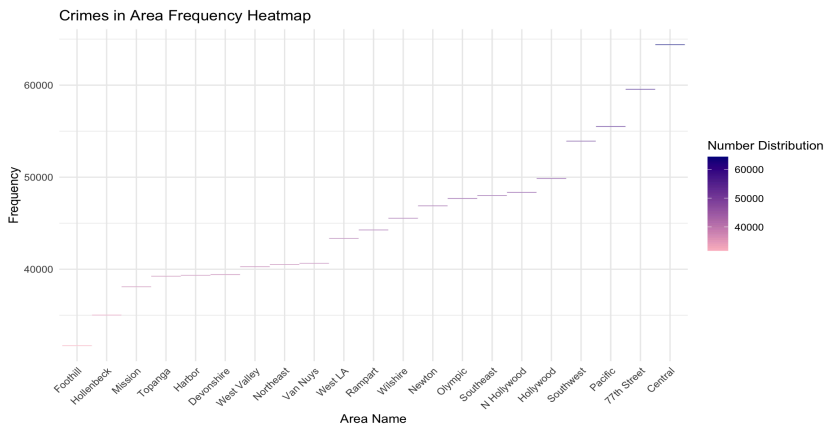
The heatmap and Bar plot was generated using ggplot2, emphasizing the frequency of crimes across different areas.

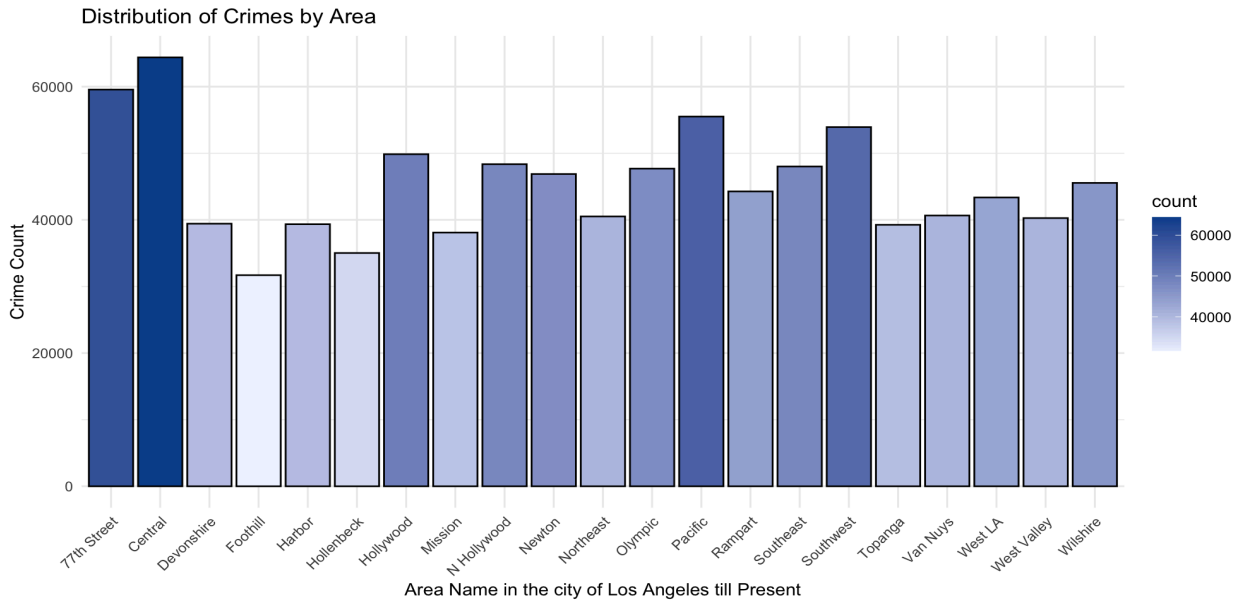
Visual Insights

- **Frequency Distribution:** The heatmap highlights areas with higher crime frequencies using a gradient scale (pink to darkblue).
- **Area Representation:** Each area is represented horizontally, sorted by descending crime frequency (n).

Key Findings

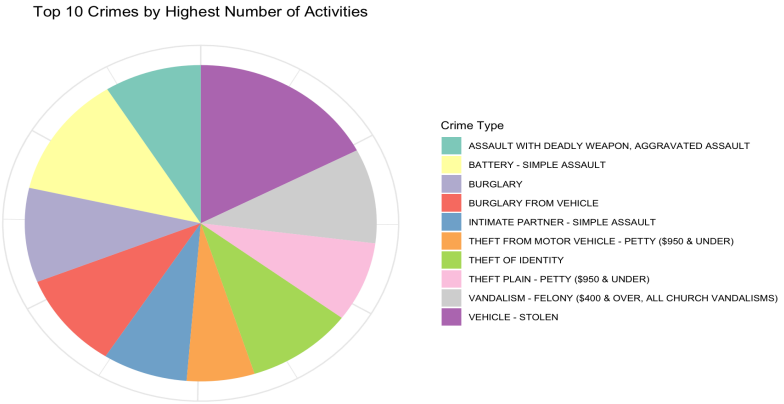
The area Central has highest crime rates (above 60000), where as in area Foothill in the city of Los Angeles has lowest crime rates (below 3000).





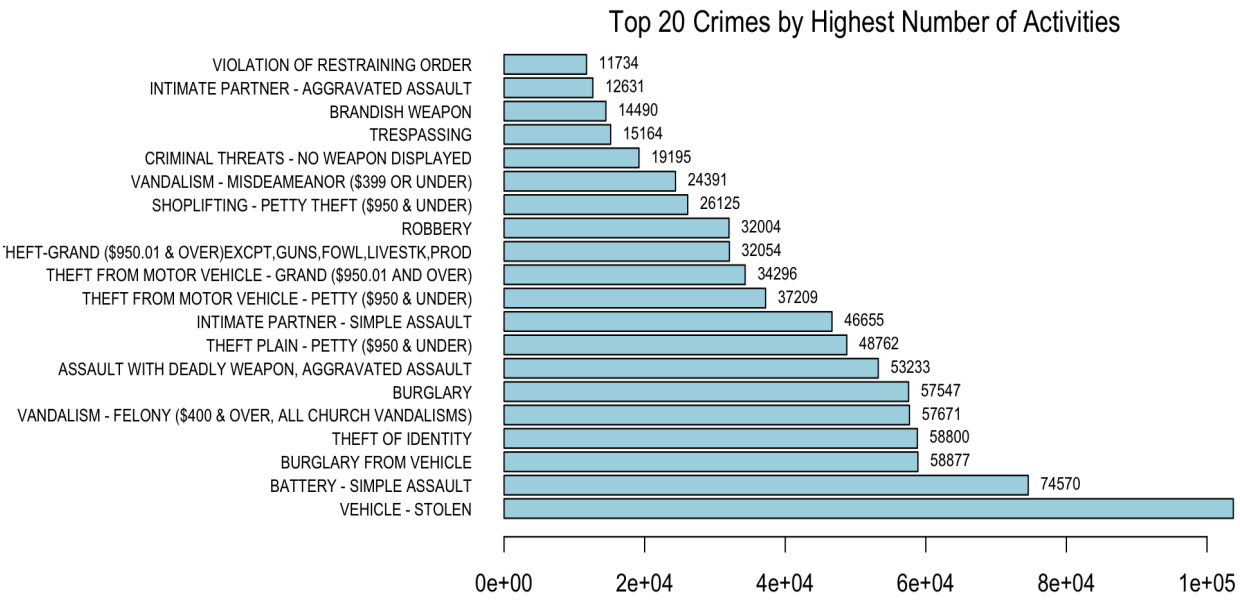
Top 20 Crime by No of activity from 2020-till present:-

Horizontal bar plot and pie chart was created to visualize effectively to identify and rank the most frequent crimes, to understand the distribution and prevalence of criminal activities. The bar plot demonstrated that , Vehicle Stolen has been the one the the crime with the highest number of activities more than 80000 cases, whereas Violation of Restraining order has been reported cases



upto 11734.

1. Vehicle - Stolen: 103,745 occurrences
2. Battery - Simple Assault: 74,570 occurrences
3. Burglary from Vehicle: 58,877 occurrences
4. Theft of Identity: 58,800 occurrences
5. Vandalism - Felony ($400 & Over, All Church Vandalisms): 57,671 occurrences
6. Burglary: 57,547 occurrences
7. Assault with Deadly Weapon, Aggravated Assault: 53,233 occurrences
8. Theft Plain - Petty ($950 & Under): 48,762 occurrences
9. Intimate Partner - Simple Assault: 46,655 occurrences
10. Theft from Motor Vehicle - Petty ($950 & Under): 37,209 occurrences

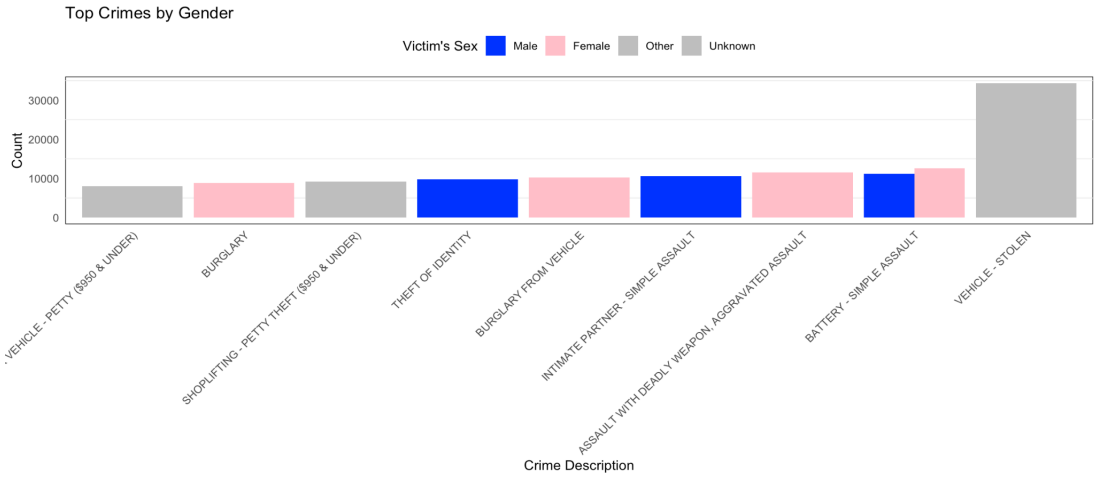


Grouping and Summarizing Crime Counts by Description and Victim Sex:-

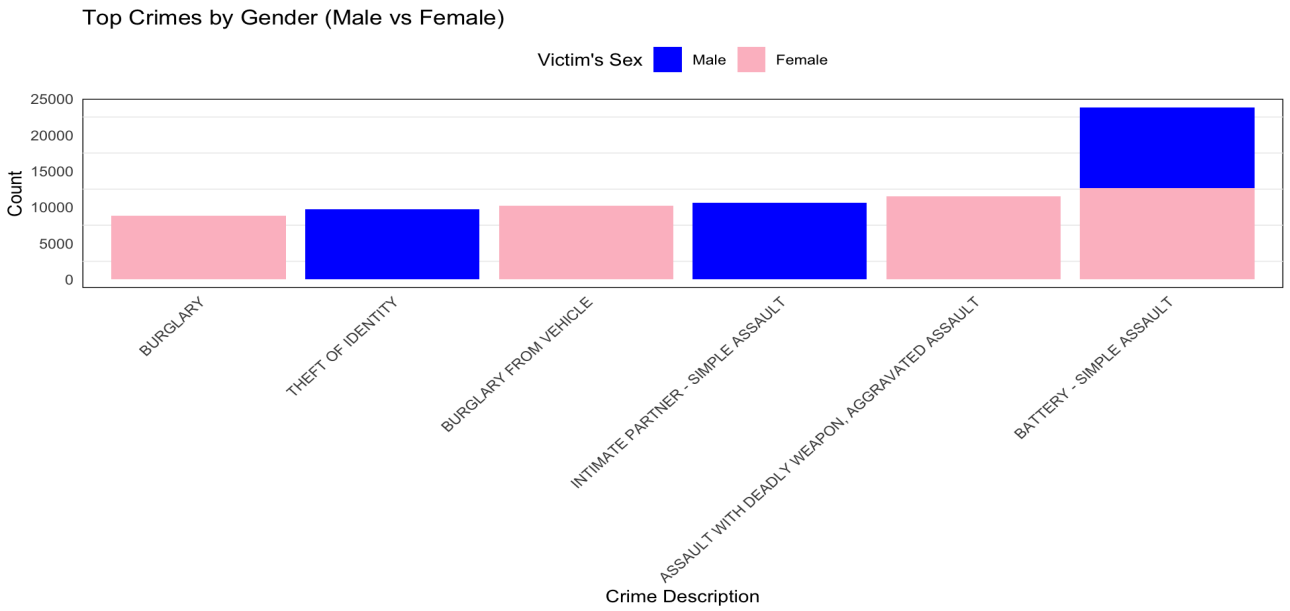
Crime_counts groups were filtered using function (filtered_data by crm_cd_desc) for crime description and victim sex. So as to calculate the count of occurrences (n()) for each combination of victim sex as well. The results were then summarized and arranged in descending order of count (arrange(desc(count))). Only the top 10 crimes by count (top_n(10, wt = count)) are selected.

Bar Plot by Crime Description and Victim Sex:-

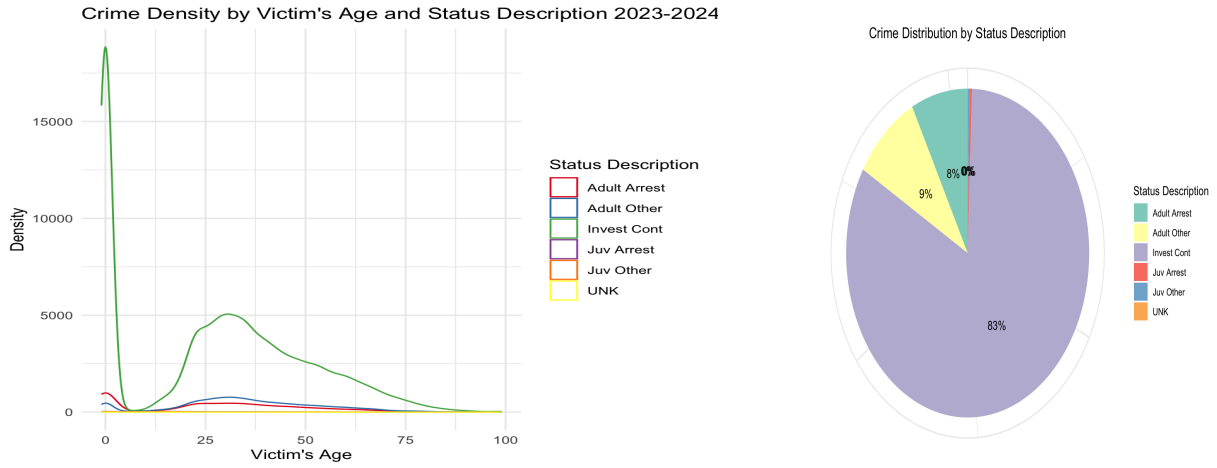
- Male vs. Female: It highlights the distribution of top crimes where gender data (vict_sex) is available.
- Crime Distribution: Shows how certain crimes are more prevalent among males or females based on the count of occurrences. Insights: Vehicle Stolen has been done by Other people, rather than male and female.



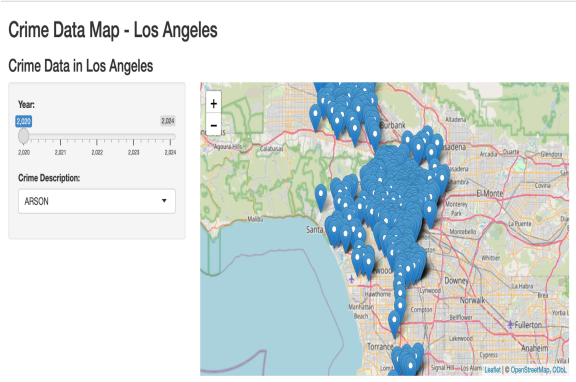
Further, I have categorized bar plot to visualize crimes committed by gender Male and Female only, where Battery- simple assault has been committed more by Male rather than Female in this stacked Bar plot.

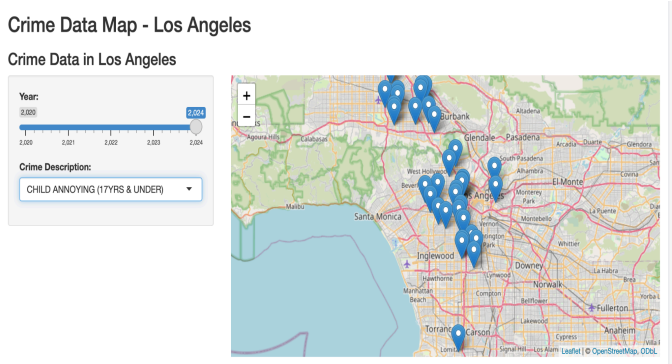


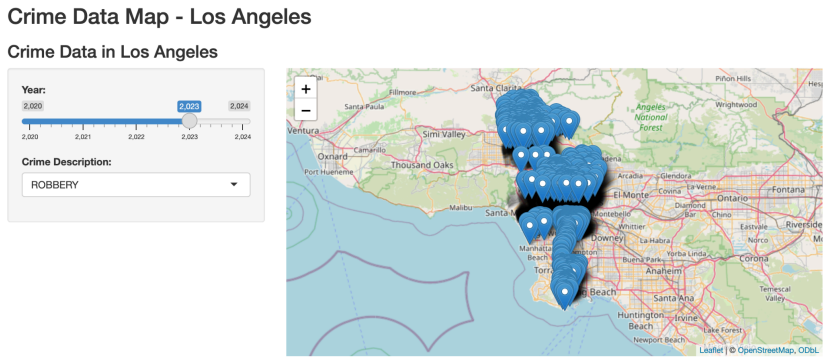
Further, the age factor indicates which age group has reported the highest number of crimes. According to the line plot chart, the age group 25-30 has reported the highest number of crimes, while investigations are still ongoing. Investigations have been reported for 83% of all crimes that occurred in the years 2023-2024.



3D visualization using Leaflet and Shiny for year 2020- till present :- We have used Leaflet and Shiny package to take all the factors of crime name w.r.t to crime occurrence from the year 2020 till 2024 (present) with a option to change crime positions and locate on map crime areas.







Summary :-

- Crime Frequency Analysis: Visualized trends through bar plots and histograms, noting an initial increase and subsequent decrease in crime rates.
- Top 20 Crimes by Activity Counts: Highlighted Vehicle Theft as the most prevalent crime, with detailed analysis of other top offenses.
- Grouping and Summarizing Crime Counts: Utilized bar plots to explore crime distributions by description and victim sex.
- Age Factor Analysis: Identified the 25-30 age group as reporting the highest number of crimes.
- 3D Visualization using Leaflet and Shiny: Mapped crime occurrences across Los Angeles by type and location using interactive tools.

The report concludes with references to R packages (ggplot2, janitor, Leaflet, Shiny) and for further exploration into crime dynamics and demographics in subsequent milestones.

Reference -

- Amir, I. (n.d.). US Crime EDA. https://rpubs.com/isal_amir/US_Crime_EDA
- Gambogi, R. (2020). Introduction to the janitor package. https://cran.r-project.org/web/packages/janitor/vignettes/janitor.html
- U.S. General Services Administration. (n.d.). Crime data from 2020 to present. https://catalog.data.gov/dataset/crime-data-from-2020-to-present
- Minn, M. (n.d.). Analyzing crime data with R.https://michaelminn.net/tutorials/r-crime/index.html
- GeeksforGeeks. (n.d.). R - Bar Charts. https://www.geeksforgeeks.org/r-bar-charts/
- Stack Overflow. (2015). How to shorten data frame values to first character in R? https://stackoverflow.com/questions/28984483/r-how-to-shorten-data-frame-values-to-first-character
- Psychological Statistics. (n.d.). Bar Charts.https://advstats.psychstat.org/book/graphr/bar.php
- Stack Overflow. (2012). Rotating x-axis labels in R for barplot. from https://stackoverflow.com/questions/10286473/rotating-x-axis-labels-in-r-for-barplot
- Baishya, M. (2023). LA Crime Data 2010 to 2023. Kaggle. https://www.kaggle.com/datasets/manjitbaishya001/la-crime-data-2010-to-2023/code?datasetId=4192235
- RStudio. Using Leaflet with Shiny. https://rstudio.github.io/leaflet/articles/shiny.html
- Programming Historian. Shiny and Leaflet for interactive maps: A tutorial for historians. Programming Historian. https://programminghistorian.org/en/lessons/shiny-leaflet-newspaper-map-tutorial
- Stack Overflow. (2023). Create a Shiny module that creates a Leaflet map in Shiny app. https://stackoverflow.com/questions/70550397/create-a-shiny-module-that-creates-a-leaflet-map-in-shiny-app