Los Angeles Crime Data: EDA

Group 13: Titus(tra29), Revanth(rg361), Suim(sp699), Meixiang(md480)

1. Introduction

1.1. Data Overview

The data used originates from the crime records of the Los Angeles Police Department(LAPD) from 2020 up to present day. The data is created by transcribing original crime reports that are typed on paper and is updated on a `weekly` basis.

The Dataset contains **820,599 observations**, encompassing **28 variables** as of the latest update on **18-Oct-2023**.

Each row of the Dataset represents a crime reported in Los Angeles, and contains the following categories of information(with selected examples):

1. `Location`: Latitude, Longitude, Area, Street, District
2. `Victim Demographic`: Age, Gender, Ethnicity
3. `Crime Description`: Type of Crime, Investigation Outcomes, Weapon Usage
4. `Date and Time`: Date Reported, Date Occurred, Time Occurred
5. `Identifier/Classifier`: Crime Record Identifier, Mocodes

1.2. Research Questions

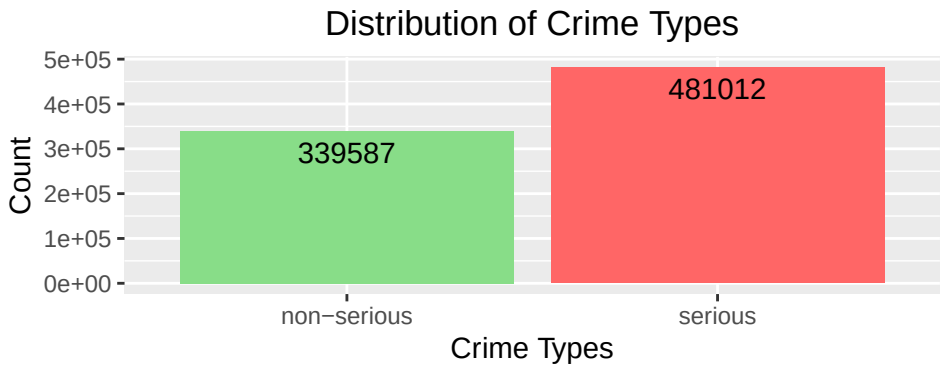
We aim to answer the following two questions using the data:

1. What are the strongest indicating factors that influence the seriousness of crime committed(categorical outcome).
2. What are the factors which influence the number of crimes committed(continuous outcome).

2. Outcome Variables

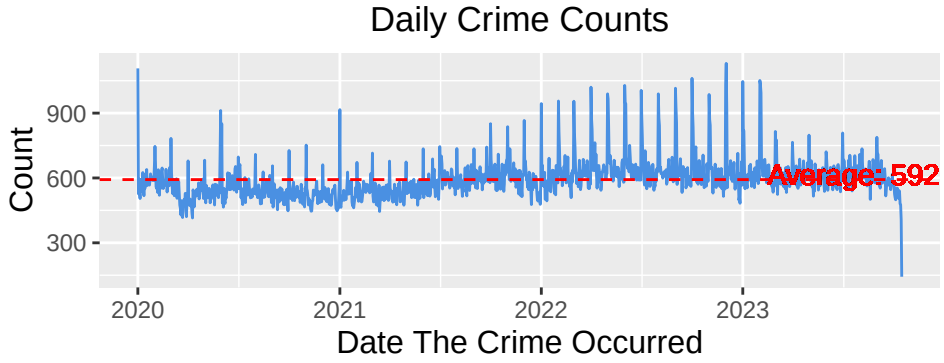
1. A new variable `crime_type` is derived from the 'Part 1-2' column in the original dataset which classifies crime committed into two categories; `serious` for crimes such as felony offenses like criminal homicide, forcible rape, etc; and `non-serious` for less severe crimes.

   The bar plot below represents the count of the seriousness of crime by category in the dataset, about **60%** of the crimes committed are serious crimes.



2. The second outcome variable is the count of crimes reported on daily basis derived by grouping the original dataset by the 'Date Occurred' column.

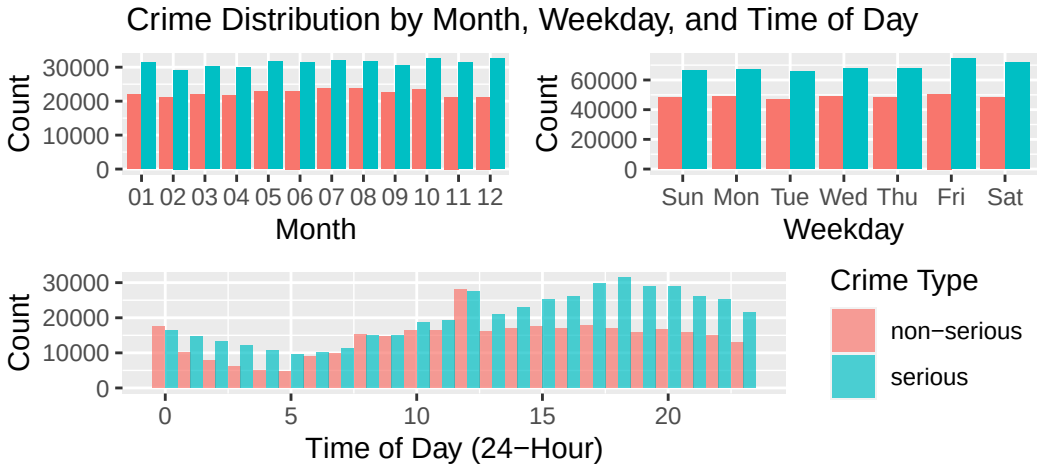
   The time series plot below displays the number of crimes for each day in the dataset along with the average number of crimes per day. There is a slight uptrend in the number of crimes per day over time with a few days having very high number of crimes. There is a drastic drop at the end as the lastest crimes may not have been reported/updated yet.



3. Primary Relationships of Interest

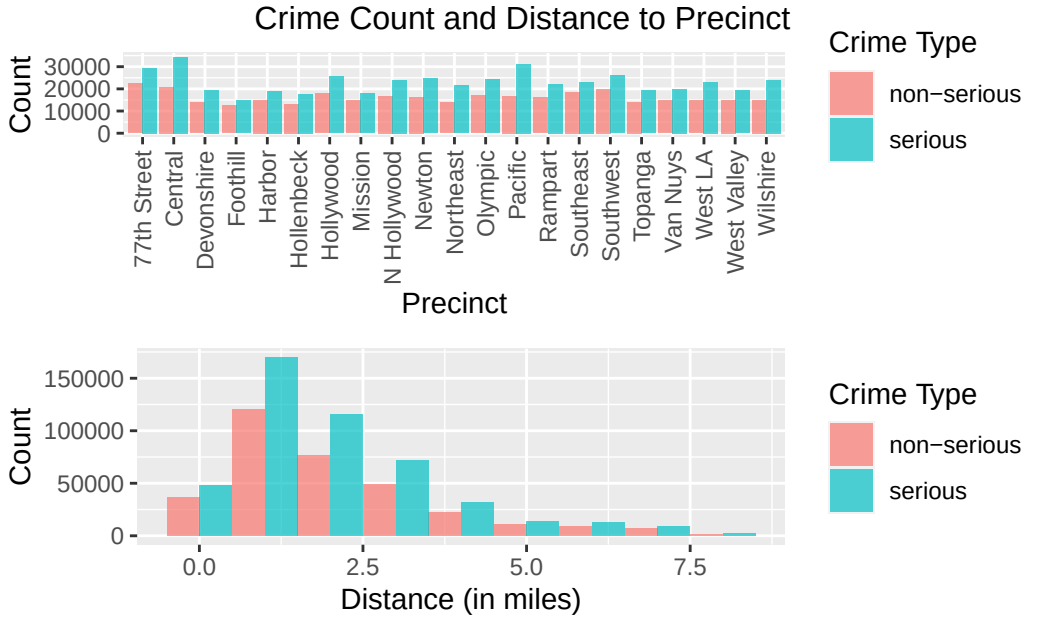
3.1. Time and Date Variable

There is a minor variation in the proportion and count of crimes when compared on a monthly and weekday level. The variation is more prominent at the 'Time of Day' level.



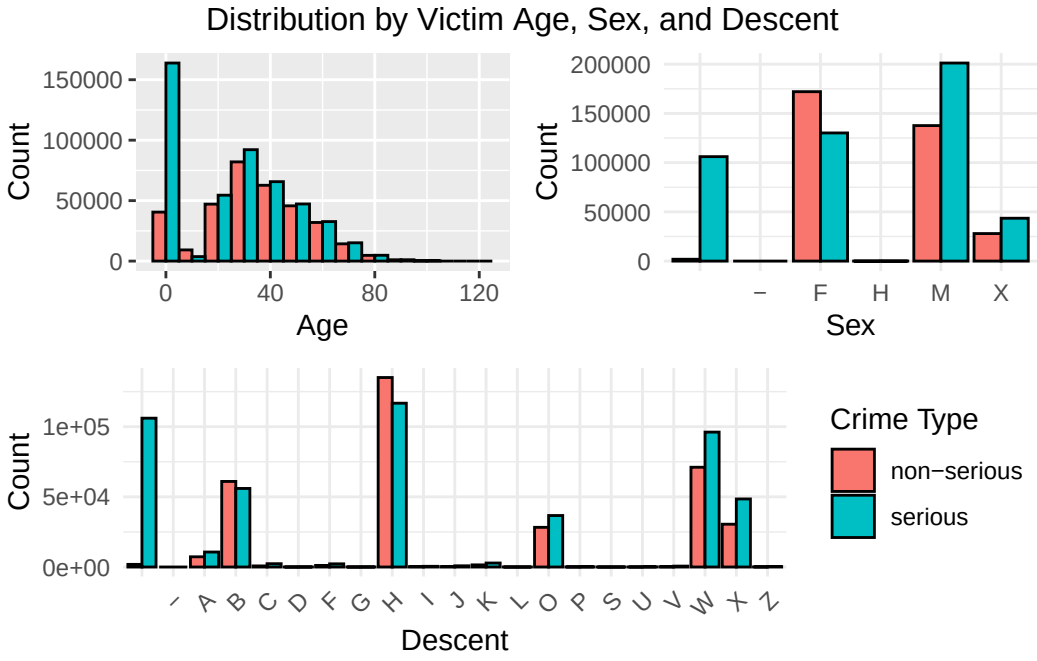
3.2 Variation with Location

There is a minor variation in the proportion and count of crimes at different precincts.



3.3 Victim Demographics

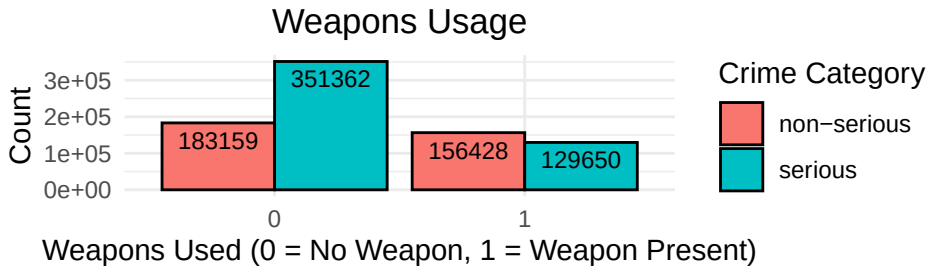
The victim demographic data - 'Age', 'Sex' and 'Descent' have a large number of missing values(systematic). There is a clear observable variation of the number and proportion of the type of crimes with respect to these variables. (Refer to Appendix)



*Victim Sex: F - Female, M - Male, X - Unknown; Victim Descent Code: A - Other Asian, B - Black, H - Hispanic/Latin/Mexican, W - White, X - Unknown, O - Other.

3.4. Weapons Usage

Against common intuition the proportion of serious crimes is lesser when a weapon is involved.



4. Other Characteristics

Some of the variables in the Dataset were excluded from the analysis due to the following reasons(with selected examples):

1. `Absence of a connection`: Certain ID-related variables such as crime record identifier.
2. `Duplicated Information`: Columns which contained duplicated or redundant information as Crime Code 1.
3. `Missing Values`: Columns which had a high number of NAs which couldn't be imputed e.g. Crime Code 2, 3, 4.
4. `Out of Scope`: Variables which are not inline with the research interest such as Mocodes and detailed crime description.

5. Potential challenges

The following are the potential challenges we may face during the modelling phase:

1. `Missingness`: The absence of data, particularly in victim demographic variables such as "age," "sex," and "decent," can introduce bias and significantly impact the model's performance. Depending on the extent of missingness, it might be necessary to employ suitable strategies such as imputation techniques or cautious exclusion of incomplete cases to ensure the integrity and accuracy of the analysis.
2. `Data`: While we have sufficient size of Data (~800k Records) to use for the analysis, the dynamic nature of the Data (updated weekly) might posses a challenge as we may see new discrepancies which were not present in earlier during EDA. TO overcome this, we plan to use the same Dataset we used for the EDA phase and not update it to the latest one at the time of modelling.
3. `Categorical Variable Handling`: Certain categorical variables, such as crime description comprise a substantial number of levels, potentially complicating the modeling process. Certain categorical variables, including "Area," comprise a substantial number of levels, potentially complicating the modeling process.
4. `Outlier`: Some variables such as the distance to precinct have outlines which need to be investigated and handled accordingly.