**Milestone 1 - Project Proposal**

**Group Members & Github Username**

Jingle Wang, [gillianw@seas.upenn.edu](mailto:gillianw@seas.upenn.edu) (gillianwang0325)

Ziqi He, [ziqihe@seas.upenn.edu](mailto:ziqihe@seas.upenn.edu) (HeidiHe2001)

Lingpei Luo, [luol@seas.upenn.edu](mailto:luol@seas.upenn.edu) (JLmm123)

Guangqiuse Hu, [violahu@seas.upenn.edu](mailto:violahu@seas.upenn.edu) (violahu930)

**Application/Website Idea**

We want to create a searchable crime database platform named "CrimeSearchLA". This platform primarily serves as a robust search engine for crime records, enabling users, researchers, law enforcement, and journalists to access and analyze detailed crime data. The primary aim is to offer detailed insights into crime patterns, assist in academic and professional research, and improve public awareness.

**Dataset**

a. A 1-2 sentence description of the dataset

The dataset consists of detailed crime records for Los Angeles from 2020 onwards, with fields including dates, times, locations described to the nearest hundred block, types of crimes, weapon used, victim demographics, and status of cases. Each incident is identified by a unique record number (DR\_NO), with additional information about the crime's circumstances and the involved parties.

b. A link to where you found the dataset

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

c. If you’re not scraping the data:

i. relevant size statistics (e.g. For a table, mb/gb, number of rows, and number of

attributes. For a graph, mb, number of nodes, and number of edges)

* Number of rows: 901,357
* Number of columns (attributes): 21
* Size: 235.5 mb

ii. summary statistics of several attributes (e.g. report mean, standard deviation)

The only attribute with meaningful mean and standard deviation:

* Attribute: Vict Age, Mean: 29.6, Standard Deviation: 21.8

Other categorical variables with number of unique values, top value and count:

* Attribute: Date Rptd, Unique: 1511, Top: “02/03/2023 12:00:00 AM”, Top Count: 925
* Attribute: DATE OCC, Unique: 1511, Top: “12/02/2022 12:00:00 AM”, Top Count: 1132
* Attribute: AREA NAME, Unique: 21, Top: “Central”, Top Count: 61416
* Attribute: Crm Cd Desc, Unique: 139, Top: “VEHICLE - STOLEN”, Top Count: 96751
* Attribute: Status, Unique: 6, Top: “IC”, Top Count: 721221
* Attribute: LOCATION, Unique: 65040, Top: “800 N ALAMEDA ST”, Top Count: 1717

**Queries**

A list of at least 5 queries (in natural language) you could write for your datasets. Some

of these should require complex SQL (aggregations, subqueries, joins, etc.)

1. Find the Number of Specific Crimes by Neighborhood over a Given Time Period:
   * “Show the total number of burglary incidents in each neighborhood of Los Angeles for the year 2020.”
   * This would require aggregation (COUNT) and grouping (GROUP BY) by neighborhood, filtered by crime type and date.
2. Impact of Crime on Different Days of the Week:
   * “For each type of crime, calculate the average number of incidents on weekends versus weekdays in 2022.”
3. Compare Crime Rates Before and After a Specific Event or Policy Implementation:
   * “Compare the monthly average of all crimes six months before and six months after the implementation of the new public safety measure on January 1, 2021.”
   * This query requires dividing the data set into two periods, then using aggregation to calculate averages.
4. Identify Trends in Weapon Use in Violent Crimes:
   * “Identify the trend in the use of firearms in violent crimes from 2012 to 2022, broken down by quarter.”
   * Requires filtering by crime type to include only violent crimes, extracting the quarter from the date, and counting incidents involving firearms, followed by a trend analysis over the quarters.
5. Correlate Crime Types with Victim Demographics:
   * “Determine the most common type of crime against seniors (age 65 and older) in Los Angeles, including a breakdown by gender of the victim.”
   * This involves a join between the crime records and victim demographics, filtered by victim age, and requires aggregation to find the most common crime types, with an additional grouping by victim gender.