BANL 6900 - Business Analytics CAPSTONE PROJECT - Crime Data Analysis

Pavani Badam, Antara Moundekar, Shravanthi Logishetti

2024-05-04

## Introduction

Crime is a widespread social issue that affects communities all over the world and poses serious risks to economic stability, individual well-being, and public safety. As one of the biggest and most populated cities in the country, Los Angeles has a complicated terrain of criminal activity, which calls for a sophisticated knowledge of crime patterns, trends, and contributing factors. This project uses a complete dataset provided by the LAPD to analyze crime trends in Los Angeles from 2020 to March 1st, 2024. The dataset includes several important attributes, such as the location coordinates, victim demographics (age, sex, and descent), crime code description, victim demographics recorded, date of occurrence, time of occurrence, and geographic area. This project intends to find patterns, correlations, and insights in the data using sophisticated analytical techniques and visualization tools, providing useful information to support law enforcement agencies, lawmakers, and community stakeholders in their efforts to prevent and mitigate crime. Ultimately, the findings of this analysis can inform evidence-based decision-making processes and contribute to fostering safer communities in Los Angeles.

## Business Problem Identification

In the current era, where data is a potent instrument for well-informed decision-making, the goal of our study is to show how data analytics may enable business users to choose safe neighborhoods in Los Angeles. In addition, we aim to demonstrate how companies that are committed to creating safer communities may benefit from using data insights, which is in line with the goals of both promoting economic growth and maintaining public safety. Our primary focus placement andbusinesses in real estate, security services, and insurance by leveraging crime data analytics to provide valuable insights for informed decision-making in property investments, optimizing security resource placement, and assessing insurance premiums based on location-specific risks.

*Real estate* and property management companies can leverage crime data analytics to give their clients information about neighborhood safety, which in turn helps them make well-informed investment decisions.

*Security Services* - Businesses that provide security services can maximize the deployment of resources and staff in accordance with crime hotspots by utilizing crime data analytics, which will increase overall efficacy.

*Insurance Companies* - Based on the risk involved with the location of insured properties, insurance companies can evaluate and modify insurance premiums by using crime data analytics.

## Literature Review

The LAPD crime dataset, which covers the years 2020 to 2024, is an important yet complex tool for studying the evolution of urban crime. This dataset addresses a number of difficulties related to privacy, temporal elements, and crime data accuracy that have been brought up in the literature. These issues include reporting problems, temporary update modifications and privacy concerns. The LAPD’s dedication to addressing these issues is consistent with continuous efforts within academia to improve the accuracy of crime databases, guaranteeing their usefulness in shaping urban safety policies.

Publisher data.lacity.org. (2024, April 19). *City of Los Angeles - crime data from 2020 to present*. Catalog. https://catalog.data.gov/dataset/crime-data-from-2020-to-present

Department, L. A. P. (2024, May 1). *Crime data from 2020 to present: Los Angeles - open data portal*. Crime Data from 2020 to Present | Los Angeles - Open Data Portal. https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8/about\_data

## Methodology

The study will employ a combination of descriptive and predictive analytics to explore crime patterns.

**Descriptive Analytics (Cluster Analysis)**

Past crime data will be examined to identify general trends, patterns, and variances in crime rates over time and between various Los Angeles neighborhoods. Examining aggregate statistics such as the overall number of crimes, the crime rate by time, and the distribution of different crime types among different neighborhoods and regions will be part of this. Descriptive analytics will also give light on temporal changes in the incidence of crime, including patterns related to the day of the week and times of day when crime is more likely to occur.

Cluster Analysis will make it possible to identify regions where crime is more likely to occur, which will facilitate the proactive deployment of law enforcement resources and the application of focused crime prevention tactics.

**Data Collection**

The study leverages Los Angeles city crime data sourced from Los Angeles Database from January 2020 onwards, containing comprehensive details such as date, time, crime types, and incident locations. Additionally, demographic information on victims, where available, enriches the dataset. This data collection approach ensures a diverse and robust analysis across 901357 crime records. The variables in the dataset are listed below, each with a description:

1. Date Rptd - Date when crime got reported
2. DATE OCC - Date when crime occured
3. TIME OCC - Time when crime occured
4. AREA - The LAPD has 21 Community Police Stations referred to as Geographic Areas within the department. These Geographic Areas are sequentially numbered from 1-21.
5. AREA NAME - The 21 Geographic Areas or Patrol Divisions are also given a name designation that references a landmark or the surrounding community that it is responsible for.
6. Rpt Dist No - A four-digit code that represents a sub-area within a Geographic Area.
7. Crm Cd Desc - Defines the Crime Code provided.
8. Vict Age - Age of victim
9. Vict Sex - Sex of victim
10. Vict Descent - Descent Code: A - Other Asian B - Black C - Chinese D - Cambodian F - Filipino G - Guamanian H - Hispanic/Latin/Mexican I - American Indian/Alaskan Native J - Japanese K - Korean L - Laotian O - Other P - Pacific Islander S - Samoan U - Hawaiian V - Vietnamese W - White X - Unknown Z - Asian Indian
11. Premis Cd - The type of structure, vehicle, or location where the crime took place.
12. Premis Desc - Defines the Premise Code provided.
13. Status - Status of the case. (IC is the default)
14. Status Desc - Defines the Status Code provided.
15. LAT - Latitude
16. LON – Longitude

**Data Cleaning and Preparation**

A careful data cleaning process is applied to address missing values, inconsistencies, and anomalies in the dataset. This involves managing null values, standardizing data formats, and ensuring uniformity across records. Preprocessing steps prepare and structure the data appropriately for analysis.

* **Handling Missing Values**

**The data cleaning process begins by identifying and addressing missing values in attributes like crime types, victim demographics, or location details. Methods such as imputation or exclusion of incomplete records ensure data integrity and reliability for analysis.**

* **Resolving Inconsistencies and Anomalies**

**The dataset undergoes scrutiny to address inconsistencies and anomalies such as date format discrepancies, using anomaly detection algorithms and data profiling to correct these issues and uphold data reliability.**

* **Standardizing Formats and Data Uniformity**

**Formats across various attributes are standardized for coherent analysis, harmonizing date and time fields to ensure consistency and streamline further analysis.**

**Exploratory Data Analysis (EDA)**

This analysis leverages statistical tools in R for exploratory data examination, distribution, and variance across neighborhoods and types. Visualizations are essential in revealing crime patterns such as which area, at what time and which sex.

Visualization techniques include histograms to represent crime frequency distributions, providing insights into common occurrence times. Maps use geographical coordinates to display crime clustering in specific neighborhoods or regions. Temporal plots in power Bi highlight temporal trends and recurring crime patterns, pinpointing peak periods and seasonal fluctuations.

* **Crime by Area**

**Visualizations reveal spatial crime patterns within the dataset, using geospatial analyses to pinpoint hotspots and areas of high incidence. These analyses identify peak crime areas such as 77th Street, Central, Hollywood, and specific neighborhoods with elevated crime rates, while Foothill and West Valley consistently exhibit lower crime rates.**

A graph of pink and blue bars

Description automatically generated

* **Victim Profiling**

**Victim analysis examines demographic factors such as age, gender, ethnicity, and socioeconomic status to explore relationships between victim profiles and crime types. Understanding the age and gender of victims provides deeper insight, revealing that females aged 31 to 65 constitute the most common victim demographic, highlighting a potential focus area for tailored crime prevention measures.**

A blue yellow and green circle with text

Description automatically generated

A graph of a number of people

Description automatically generated

* **Crime at hourly intervals**

The Crime by hourly Interval in a day provides a concise overview of crime trends in Los Angeles over specific time of the day. This offers valuable insights into peak hours of criminal activity and throughout the day.

A graph of a number of red bars

Description automatically generated

**Findings and Discussions**

From the above exploration data analysis, we can say that the Central area has more number of crimes followed by 77th street. We can also infer that Victims age ranging in between (31-65) at 12:00 pm are most targeted people. The correlation coefficient between AREA and Rpt.Dist.No is approximately 1. This indicates a very strong positive linear relationship between these two variables. LAT and LON have a strong negative correlation coefficient of approximately -1. This indicates a strong negative linear relationship between latitude and longitude, which is expected as they represent geographic coordinates.

A graph with blue circles and red text

Description automatically generated

**Conclusion**

**Combining exploratory data analysis, victim profiling, and cluster analysis highlights areas of high crime incidence and peak periods of activity. The identification of crime hotspots using cluster analysis allows insurers to allocate resources effectively and focus on high-risk zones. This proactive approach helps prevent losses and enhances fraud detection processes. Insurers can work with law enforcement, community organizations, and advocacy groups to address crime trends and promote safety. By incorporating data-driven strategies into policy formulations and interventions, insurers play a role in strengthening communities which might help realtors to improve their performance by building safer communities. This alignment between data analytics and community well-being shapes a future where insurers contribute to safer, more informed, and resilient communities.**The Los Angeles crime map analysis shows that East Los Angeles, Pasadena, and Glendale are particularly safe communities with low crime rates.

A map of a city with orange circles

Description automatically generated

## References

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

landing page (description of column) -<https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8/about_data>

**Group Member Responsibilities**

1. **Business Analyst – Shravanthi Logishetti**
2. **Coder – Pavani Badam**
3. **Visualization Analyst – Antara Moundekar**