

CS 777: Big Data Analytics

Term Project

LAPD CRIME ANALYSIS

Aditya Maheshwari

Sarthak Pattnaik

Vaidehi Shah

Boston University, Metropolitan College

Fall 2023



I. Introduction

The LAPD police report dataset offers a comprehensive overview of crime incidents within the City of Los Angeles, spanning from 2020 onward. Originating from original crime reports transcribed from paper documents, it's crucial to note potential data inaccuracies stemming from the manual transcription process.

The dataset carefully balances the provision of valuable information on criminal activities like the type and description of crime and the victim demographics; along with the protection of individual privacy, restricting address details to the nearest hundred blocks.

Our analysis addresses critical business problems related to crime rates in different areas and at varying times of the day. This entails identifying both high and low crime rate areas and understanding the prevalence of specific crimes based on temporal patterns. We are also investigating the relationship between the time of occurrence and the reporting of various crime types. By delving into temporal aspects, our goal is to gain insights into the efficiency of crime reporting mechanisms and potentially identify areas for improvement.

Through our examination of the dataset, we aim to contribute valuable insights to enhance public safety strategies and optimize law enforcement efforts in Los Angeles.

Date Rptd	DATE OCC	TIME OCC	AREA	AREA NAME	Rpt Dist No	Part 1-2	Crm Cd	Crm Cd Desc	Mocodes	Vict Age	Vict Sex	Vict Descent	Premis Cd
01/08/2020	01/08/2020	2230	03	Southwest	0377	2	624	BATTERY - SIMPLE ...	0444 0913	36	F	B	501 SINGL
01/02/2020	01/01/2020	0330	01	Central	0163	2	624	BATTERY - SIMPLE ...	0416 1822 1414	25	M	H	102
04/14/2020	02/13/2020	1200	01	Central	0155	2	845	SEX OFFENDER REGI...	1501	0	X	X	726
01/01/2020	01/01/2020	1730	15	N Hollywood	1543	2	745	VANDALISM - MISDE...	0329 1402	76	F	W	502 MULTI
01/01/2020	01/01/2020	0415	19	Mission	1998	2	740	VANDALISM - FELON...	0329	31	X	X	409 BEAL

only showing top 5 rows

Fig 1: Raw Data

II. Data Collection

The raw dataset utilized for our analysis is directly sourced from the official government website, data.gov. This platform serves as a centralized repository for a diverse range of datasets, including comprehensive information on crimes committed in Los Angeles. Given its status as a government-endorsed platform, the dataset is considered reliable and authoritative, ensuring that the information stems from official crime reports.

The dataset's temporal scope spans from the year 2020 to the present, encompassing crime data up to October 2023. This extensive time frame enables a meticulous examination of crime trends and patterns over a substantial period. The inclusion of data up to the present date facilitates a comprehensive analysis of temporal dynamics in criminal activities within the City of Los Angeles.

III. Business Problems

Demographics and Susceptibility to Crime

1. Investigate the influence of demographics on overall susceptibility to crime.
2. Examine whether certain demographics are more susceptible to specific types of crimes.

Crime Rates by Area and Time of Day

1. Identify areas with the highest and lowest crime rates.
2. Analyze the prevalence of different crimes based on the time of day.

Time of Occurrence and Reporting

1. Explore the correlation between the time of occurrence and reporting for various crime types.

IV. Data Cleaning and Preprocessing

In preparation for our analysis, we implemented a comprehensive data cleaning process using PySpark for the Los Angeles crime dataset. The primary objective is to address issues like inconsistent formatting, missing values, and standardizing columns to ensure the dataset's readiness for subsequent analysis. The key steps include:

- **Column Selection and Dropping:**
 - Columns deemed irrelevant, such as "DR_NO," "Weapon Used Cd," and various "Crm Cd" columns, are dropped from the DataFrame using the drop function.
- **Column Formatting:**
 - Date columns like "Date Rptd" and "DATE OCC" undergo formatting to remove irrelevant timestamp information (all 00:00:00). The result is then converted to the appropriate date type.
 - The "LOCATION" and "Cross Street" columns are merged into a new column, "address," eliminating redundant white spaces.
 - All columns are renamed to ensure clarity and consistency with more descriptive names.
 - The 'VictDescent' column undergoes mapping to replace coded values with meaningful labels, enhancing interpretability.
- **Dealing with NULLs and NaNs:**
 - Null values in the 'VictSex' and 'VictDescent' columns are replaced with 'NULL,' and specific values in the 'VictSex' column are transformed.
 - The 'Weapon' column is standardized by replacing null values with 'NO WEAPON INFO.'
 - Rows with null values in the 'PremisCd' or 'PremisDesc' columns are dropped to ensure data integrity.

```

root
|-- DateRptd: date (nullable = true)
|-- DateOcc: date (nullable = true)
|-- TimeOcc: string (nullable = true)
|-- AreaCd: string (nullable = true)
|-- AreaName: string (nullable = true)
|-- RptDistNo: string (nullable = true)
|-- CrimeType: string (nullable = true)
|-- CrmCd: string (nullable = true)
|-- CrmCdDesc: string (nullable = true)
|-- Mocodes: string (nullable = true)
|-- VictAge: string (nullable = true)
|-- VictSex: string (nullable = true)
|-- VictDescent: string (nullable = true)
|-- PremisCd: string (nullable = true)
|-- PremisDesc: string (nullable = true)
|-- Weapon: string (nullable = true)
|-- Status: string (nullable = true)
|-- StatusDesc: string (nullable = true)
|-- Latitude: string (nullable = true)
|-- Longitude: string (nullable = true)
|-- Address: string (nullable = false)

```

Fig 2: Schema after Data cleaning and processing

V. Exploratory Data Analysis

Our journey through Exploratory Data Analysis aimed to unravel the intricacies of our dataset, beginning with a meticulous examination of column descriptions. Understanding the schema laid the groundwork for distinguishing between categorical and numerical columns, a critical step for subsequent quantitative and qualitative analyses. The following was done to understand the schema and column descriptions:

- Identified categorical and numerical columns.
- Made general observations about the categories within each column.
- Utilized `.describe()` for numerical columns to gain a statistical snapshot.
- Replaced NULL and NaN values in relevant columns with the required and appropriate data.

VI. Data Integration with PostgreSQL and Tableau

We seamlessly connected our dataset to PostgreSQL and Tableau using `psycopg2`. This synergy lays the foundation for secure data storage and dynamic visualizations tailored to address key business questions.

- Connection to PostgreSQL and Tableau:
 - Leveraged the Python library `psycopg2` to seamlessly transfer our data into PostgreSQL database.
 - The primary motivation was to establish a connection between Tableau and our dataset where PostgreSQL was a required intermediate. This integration facilitates diverse visualizations in alignment with our three key business questions.
- Focus on Business Problems:

- Selected data specific to each business problem.
- Individually pushed data to PostgreSQL, forming dedicated tables.
- Applied necessary aggregations and manipulations tailored to each business problem.
- Utilized Tableau for grouping, filtering, and visualization.
- Developed Tableau dashboards for each business question.

VII. Machine Learning for Demographic Prediction

This machine learning endeavor, powered by PySpark and scikit-learn, delves into predicting demographic attributes of the victims (descent, gender, and age) from crime data. The process encompasses meticulous data preprocessing, insightful feature engineering, and the application of RandomForest models. Through a detailed analysis, we unravel the challenges and nuances associated with achieving accurate demographic predictions using crime data.

+-----+-----+		
VictAge count		
+-----+-----+		
10 663		
11 1020		
12 1502		
13 2054		
14 2462		
15 2938		
16 3333		
17 3623		
18 5047		
19 7232		
2 339		
20 9083		
21 10311		
22 11906		
23 13205		
24 14897		
25 15861		
26 16045		
27 16702		
28 17398		
+-----+-----+		
only showing top 20 rows		

+-----+-----+		
VictSex count		
+-----+-----+		
F 296253		
M 304743		
X 658		
+-----+-----+		

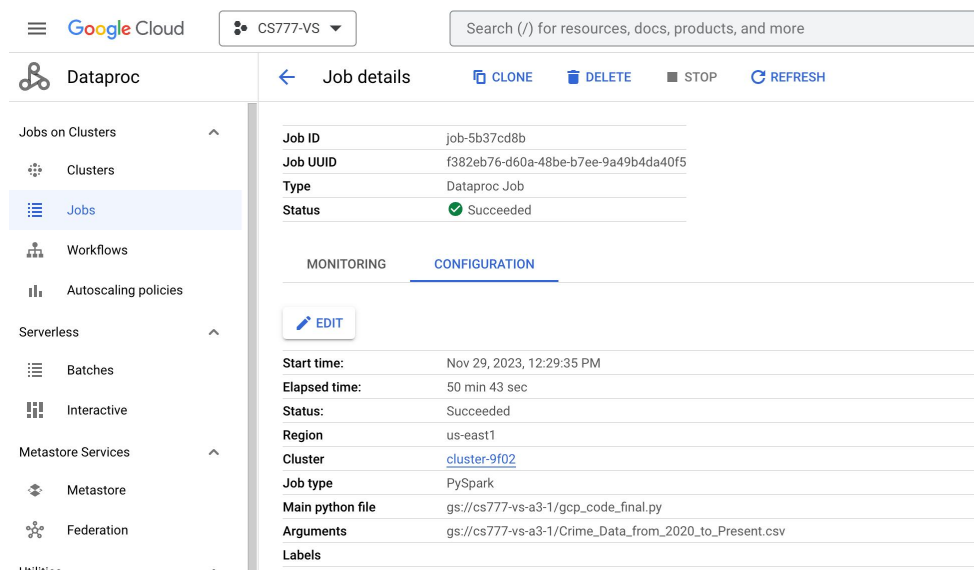
+-----+-----+		
VictDescent count		
+-----+-----+		
Asian Indian 411		
American Indian/A... 770		
Chinese 3104		
Hawaiian 134		
Japanese 1134		
Hispanic/Latin/Me... 246186		
Filipino 3380		
Vietnamese 829		
Other 54019		
Pacific Islander 217		
Samoan 40		
Guamanian 56		
Korean 4307		
Laotian 49		
White 154446		
Cambodian 60		
Black 114952		
Other Asian 17560		
+-----+-----+		

Fig 3: VictAge, VictSex, VictDescent

- Data Preprocessing:
 - Filter the dataset based on conditions related to age ('VictAge'), gender ('VictSex'), and descent ('VictDescent').
 - Create subsets of data satisfying or not satisfying specified conditions.
- Feature Engineering:
 - Calculate time differences and categorize the time of occurrence ('TimeOcc') into four parts - morning, afternoon, evening and night in a column named 'TimeOfDay'.

- Data Transformation:
 - Convert the PySpark DataFrame to a Pandas DataFrame for further analysis and encoding.
 - One-hot encoding categorical columns as they do not hold any ordinal value.
- Machine Learning Models:
 - Train separate models for predicting 'VictDescent,' 'VictSex,' and 'VictAge'. Utilizes RandomForestClassifier for classification tasks ('VictDescent' and 'VictSex') and RandomForestRegressor for the regression task ('VictAge').
- Evaluation:
 - Assesses model performance through metrics such as accuracy score, classification report, Mean Squared Error (MSE), and R-squared.
- Challenges and Insights:
 - The categorical variables within one of our columns boasted over 1000 categories (with others being near 10-15 per column), presenting a substantial challenge during the one-hot encoding process.
 - This explosion in dimensions not only complicated model training but also demanded significant computational resources. To overcome this, we applied Principal Component Analysis (PCA) to reduce dimensionality.
 - Recognize challenges stemming from the complexity of predicting demographic information based on the provided features.
- Prediction Accuracy:
 - Achieved accuracy levels of 40-50% for each of the models. This indicates challenges in accurately predicting demographic attributes.
- Improvement Areas:
 - Explores factors influencing prediction accuracy and identifies potential areas for enhancement.
 - Implement ensemble machine learning techniques to predict the victim demographics all at once.

This machine learning approach lays the foundation for demographic prediction, underscoring the importance of continuous refinement and exploration to enhance model performance and glean meaningful insights.



The screenshot shows the Google Cloud Dataproc console. The left sidebar contains navigation links: Dataproc, Jobs on Clusters, Clusters, Jobs (selected), Workflows, Autoscaling policies, Serverless, Batches, Interactive, Metastore Services, Metastore, Federation, and Utilities. The main panel displays 'Job details' for a job with ID 'job-5b37cd8b'. The job status is 'Succeeded'. Below the job details, there are tabs for 'MONITORING' and 'CONFIGURATION'. The 'CONFIGURATION' tab is active, showing an 'EDIT' button and various job parameters.

Job ID	job-5b37cd8b
Job UUID	f382eb76-d60a-48be-b7ee-9a49b4da40f5
Type	Dataproc Job
Status	✓ Succeeded

CONFIGURATION	
Start time:	Nov 29, 2023, 12:29:35 PM
Elapsed time:	50 min 43 sec
Status:	Succeeded
Region:	us-east1
Cluster:	cluster-9f02
Job type:	PySpark
Main python file:	gs://cs777-vs-a3-1/gcp_code_final.py
Arguments:	gs://cs777-vs-a3-1/Crime_Data_from_2020_to_Present.csv
Labels:	

Fig 4 : GCP Job Success

Accuracy: 0.5118049380458901

Classification Report:

					precision	recall	f1-score	support
American Indian/Alaskan Native					1.00	0.01	0.01	157
				Asian Indian	0.00	0.00	0.00	80
				Black	0.48	0.33	0.39	22992
				Cambodian	0.00	0.00	0.00	7
				Chinese	0.35	0.01	0.02	640
				Filipino	0.00	0.00	0.00	696
				Guamanian	0.00	0.00	0.00	11
				Hawaiian	0.00	0.00	0.00	26
				Hispanic/Latin/Mexican	0.54	0.74	0.63	49217
				Japanese	0.25	0.00	0.01	227
				Korean	0.19	0.02	0.03	849
				Laotian	0.00	0.00	0.00	12
				Other	0.24	0.03	0.05	10766
				Other Asian	0.20	0.01	0.03	3528
				Pacific Islander	0.00	0.00	0.00	47
				Samoan	0.00	0.00	0.00	4
				Vietnamese	0.00	0.00	0.00	163
				White	0.48	0.56	0.52	30909
				accuracy			0.51	120331
				macro avg	0.21	0.09	0.09	120331
				weighted avg	0.47	0.51	0.47	120331

Fig 5 : Model I - Classification - Predicting VictDescent

Accuracy: 0.5711329582568083

Classification Report:

			precision	recall	f1-score	support
		F	0.57	0.56	0.56	59256
		M	0.58	0.58	0.58	60928
		X	0.00	0.00	0.00	147
		accuracy			0.57	120331
		macro avg	0.38	0.38	0.38	120331
		weighted avg	0.57	0.57	0.57	120331

Fig 6 : Model II - Classification - Predicting VictSex

Linear Regression with PCA MSE: 236.2902424485539
 Decision Tree Regression with PCA MSE: 562.0412958005537
 Decision Tree Regression with PCA MSE: 243.05583133485314
 Random Forest Regression R-squared: 0.005272755796456696
 Decision Tree Regression R-squared: -1.300203151801954
 Linear Regression R-squared: 0.03296151994304031

Fig 7 : Model III - Regression - Predicting VictAge

VIII. Geospatial Analysis

In addition to our comprehensive analysis, we conducted geospatial analysis using the Python library *Folium* to visualize the spatial distribution of crimes within Los Angeles. This additional layer of analysis provides a geographical context to our findings and helps identify crime hotspots. The steps included:

1. Extracted geographical coordinates (latitude and longitude) from the dataset to accurately plot crime locations on the map.
2. Downloaded the geojson file for LA from data.gov.
3. Plotted crime locations on the map, with each point representing an incident. This visual representation allows for a quick and intuitive understanding of where crimes are concentrated.
4. Generated heatmaps to visually represent the intensity of crime in different areas. Heatmaps provide a smooth gradient, allowing for the identification of areas with the highest concentration of criminal activities.

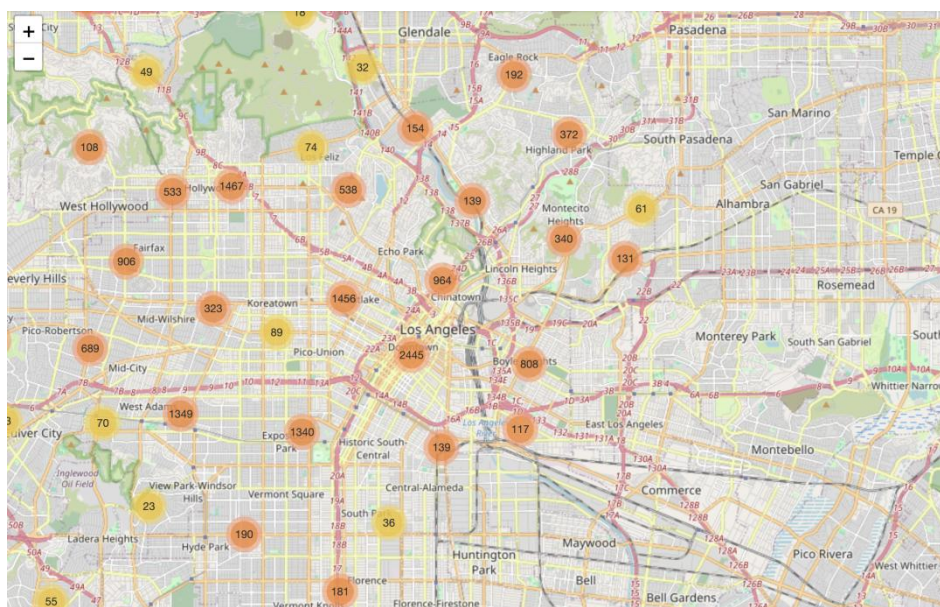
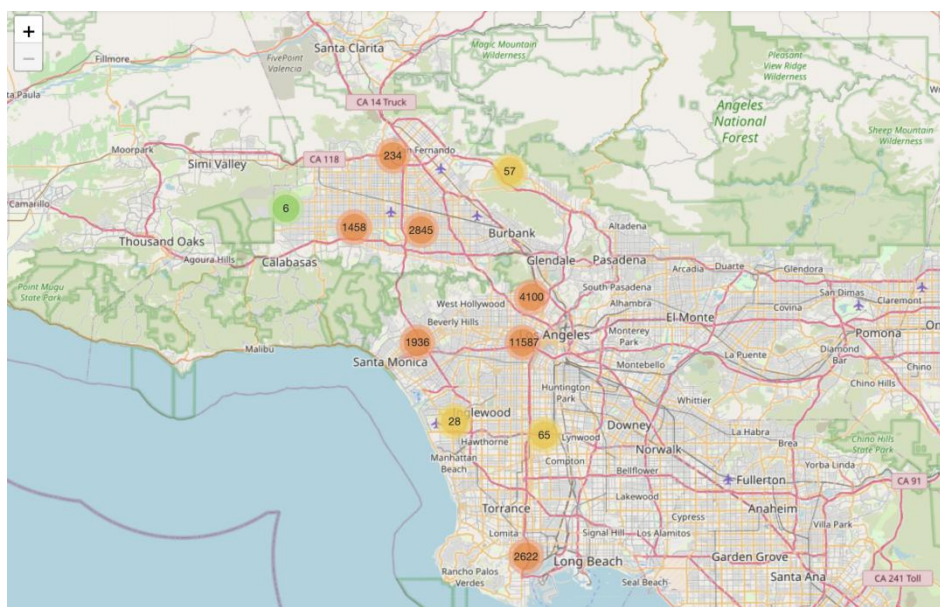


Fig 8 : LA Crime Incident Co-ordinates

- a) Zoomed out
- b) Zoomed in

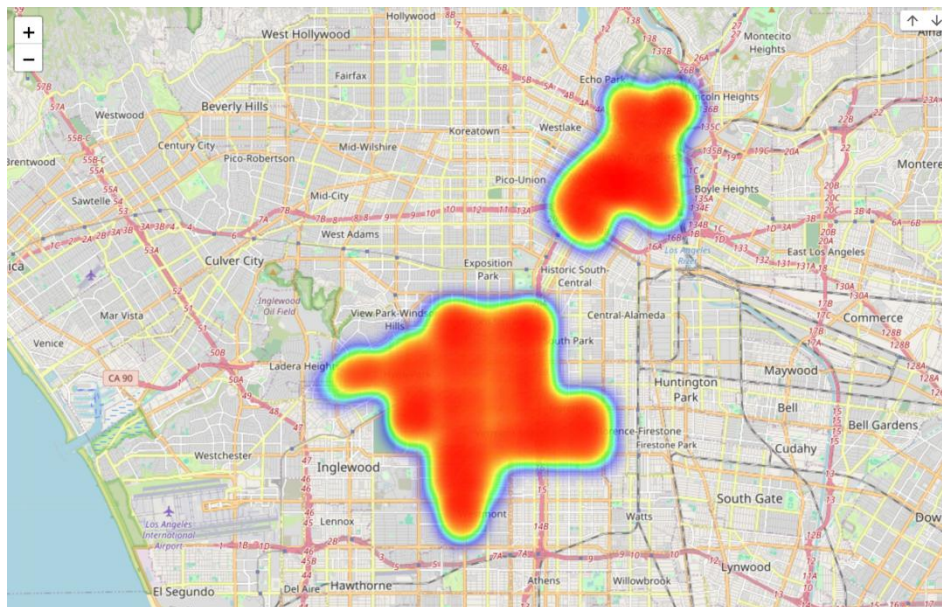
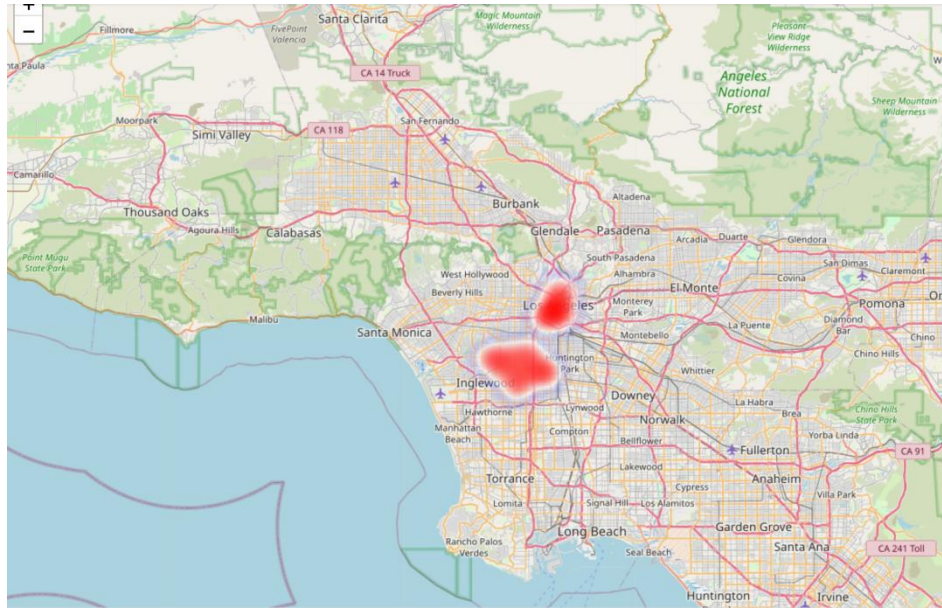


Fig 9 : LA Crime Incident Hotspots - Central & 77th Street

a) Zoomed out

b) Zoomed in

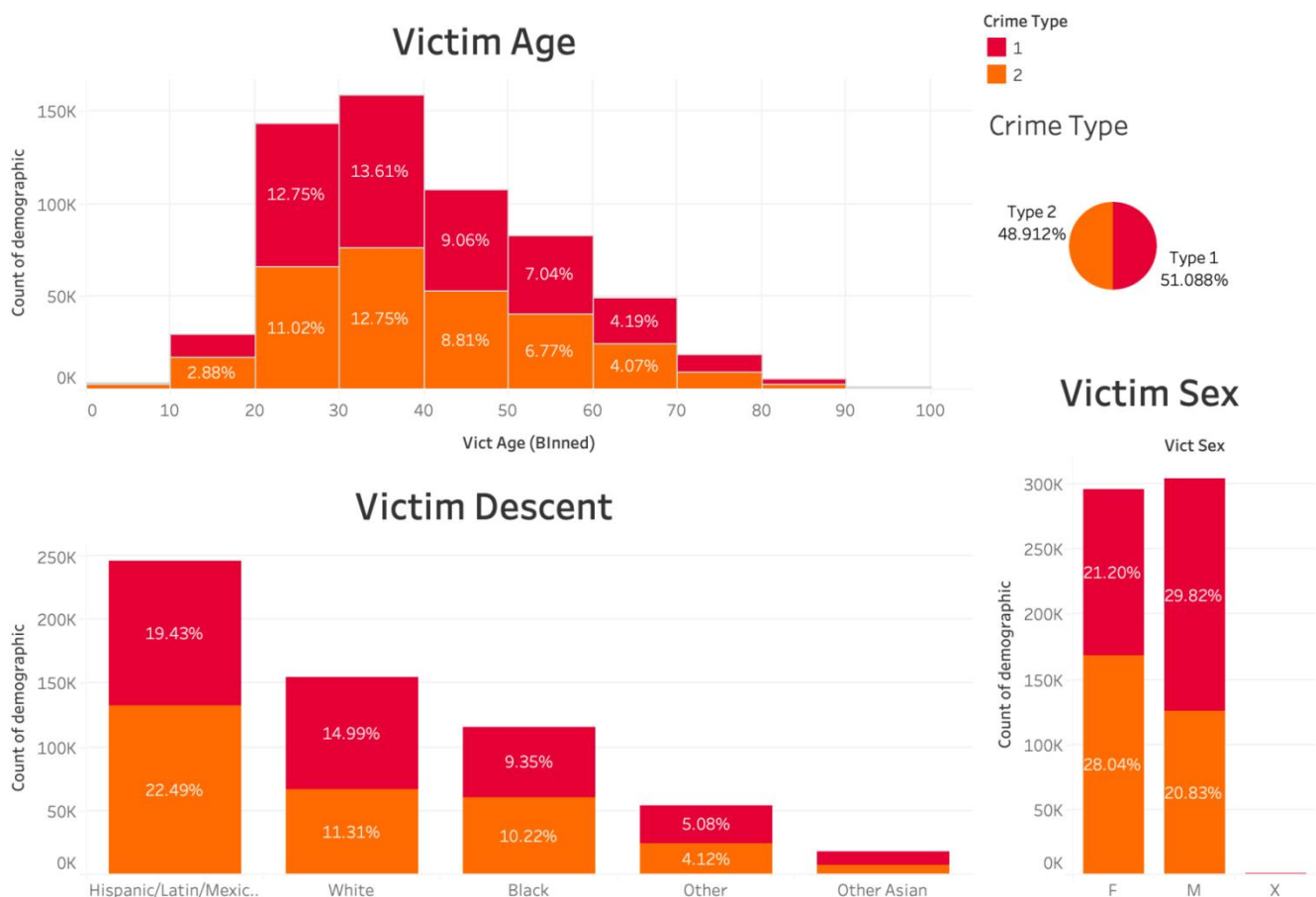
With the help of this geospatial map, we observed that Central LA has the highest concentration of crimes.

IX. Results and Analysis

Business Problem 1: Demographics and Crime

- The rates of type-1 and type-2 crimes are comparable.
- Women are much more susceptible to type-2 crimes while males are more susceptible to type-1 crimes.
- People in the age group of 30-40 years, specifically Hispanic/Latino/Mexican males, are most likely victims of Type-1 crime.
- link to dashboard: <https://public.tableau.com/BQ-1victimdemographics>

BUSINESS QUESTION 1: Does an individual's demographic profile, including age, ethnicity, and gender, correlate with and potentially indicate a higher susceptibility to specific crime type?



INFERENCE:

The data indicates that 30-40 year old Hispanic/Latin/Mexican males are most likely to be victims of Type 1 crimes.

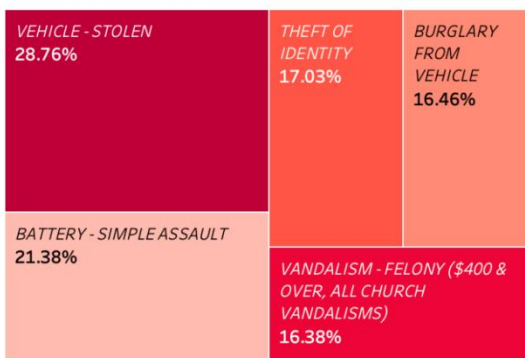
Fig 10 : Tableau Dashboard 1

Business Problem 2: Crime Rates by Area and Crime Type

- The most prevalent crime in LA is vehicle stealing followed by identity theft and battery - simple assault.
- These were also among the top 10 crimes committed in the areas in LA where the crime rates are the highest.
- Central LA faced a concentrated threat with a notable shift towards targeted property crimes like Burglary from vehicles, while Foothill experienced overall lower crime rates. The rarity of inciting a riot suggests a generally stable societal environment.
- link to dashboard: <https://public.tableau.com/BQ-2topcrimesandareas>

BUSINESS QUESTION 2: How to leverage crime data to understand most prevalent crimes and crime rates based on the areas?

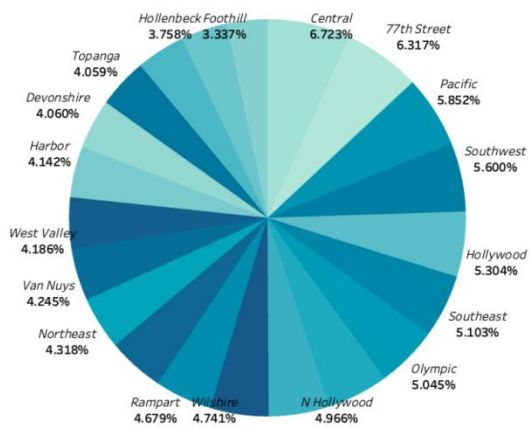
Top 5 Crimes Committed



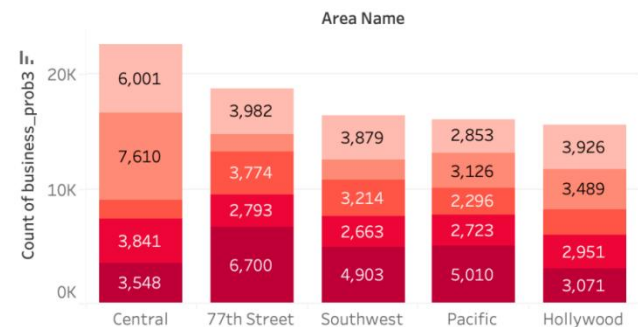
Least 5 Crimes Committed



Area-Wise Incidence Rates



Crime Hotspots: Top 5 Offenses Across Key Areas



INFERENCE:

Stolen vehicles dominate crime in Los Angeles, notably in Central LA, where property crimes like burglary from vehicles are on the rise. Despite lower overall crime rates in Foothill, the challenge of stolen vehicles persists citywide. The rarity of inciting a riot suggests a generally stable societal environment.

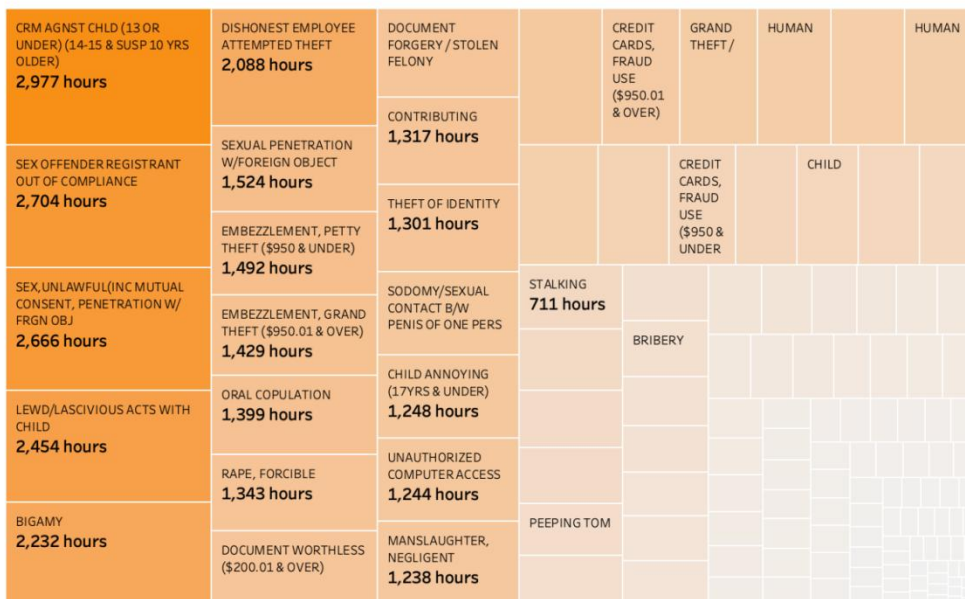
Fig 11 : Tableau Dashboard 2

Business Problem 3: Time of Occurrence and Reporting

- On average, most time was taken to report type-1 crimes (more serious offenses such as violent crimes and major property crimes), which ranged from around 1000 - 2500 hours difference between the crimes being committed and reported.
- Smaller crimes (type-2) like purse-snatching, theft from coin machines and bomb scares were reported in less than an hour or a day.
- Crimes against children and sex offenses, requiring more time for reporting, underscore potential barriers in victim disclosure or societal reluctance. Conversely, swift reporting of school disruptions and purse snatching suggests heightened awareness and prompt community responsiveness to immediate threats.
- link to dashboard: <https://public.tableau.com/BQ-3reportingtimeofcrimes>

BUSINESS PROBLEM 3: Is there a relation between the difference in the time of occurrence and report of a crime dependent on the crime?

Temporal Discrepancy Heatmap: Occurrence vs. Reporting Time



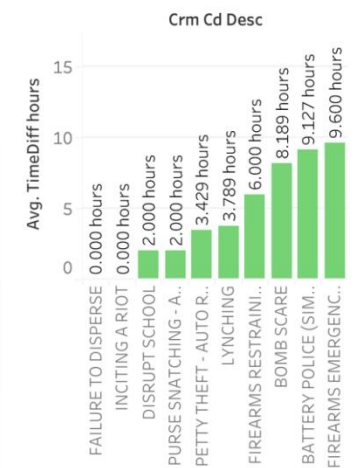
Average Time Difference in Hours



INFERENCE:

Crimes against children and sex offenses, requiring more time for reporting, underscore potential barriers in victim disclosure or societal reluctance. Conversely, swift reporting of school disruptions and purse snatching suggests heightened awareness and prompt community responsiveness to immediate threats

Analyzing Minimum Crime Reporting Times



Analyzing Maximum Crime Reporting Times

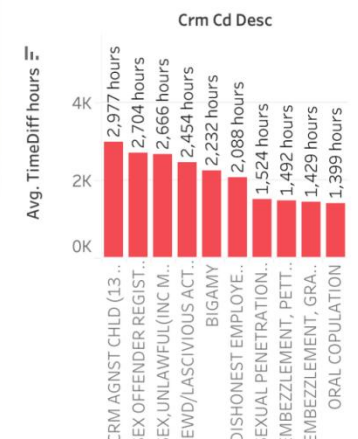


Fig 12 : Tableau Dashboard 3

Tableau Interactive Dashboard

- The data reveals a peak in crime rates in 2022 followed by a sudden dip in 2023 because of lack of data post Sept 2023. Nighttime stands out as the predominant period for criminal activities, with nights and evenings witnessing a higher occurrence of more severe Type 1 crimes, while mornings and afternoons exhibit a prevalence of less severe Type 2 crimes, highlighting nuanced temporal patterns in criminal behavior.
- The dashboard reveals monthly variations in crime rates, enabling law enforcement to anticipate seasonal shifts in criminal activities.
- By categorizing crime rates based on the time of day, the dashboard provides actionable intelligence for law enforcement patrols. Strict patrolling during peak crime hours can enhance public safety and deter criminal activities.
- Noteworthy distinctions in the prevalence of type-1 and type-2 crimes during specific times of the day underscore the importance of targeted strategies. Enhanced vigilance during the night for type-1 crimes and afternoon for type-2 crimes can optimize law enforcement efforts.
- link to dashboard: <https://public.tableau.com/LACrimeAnalytics>

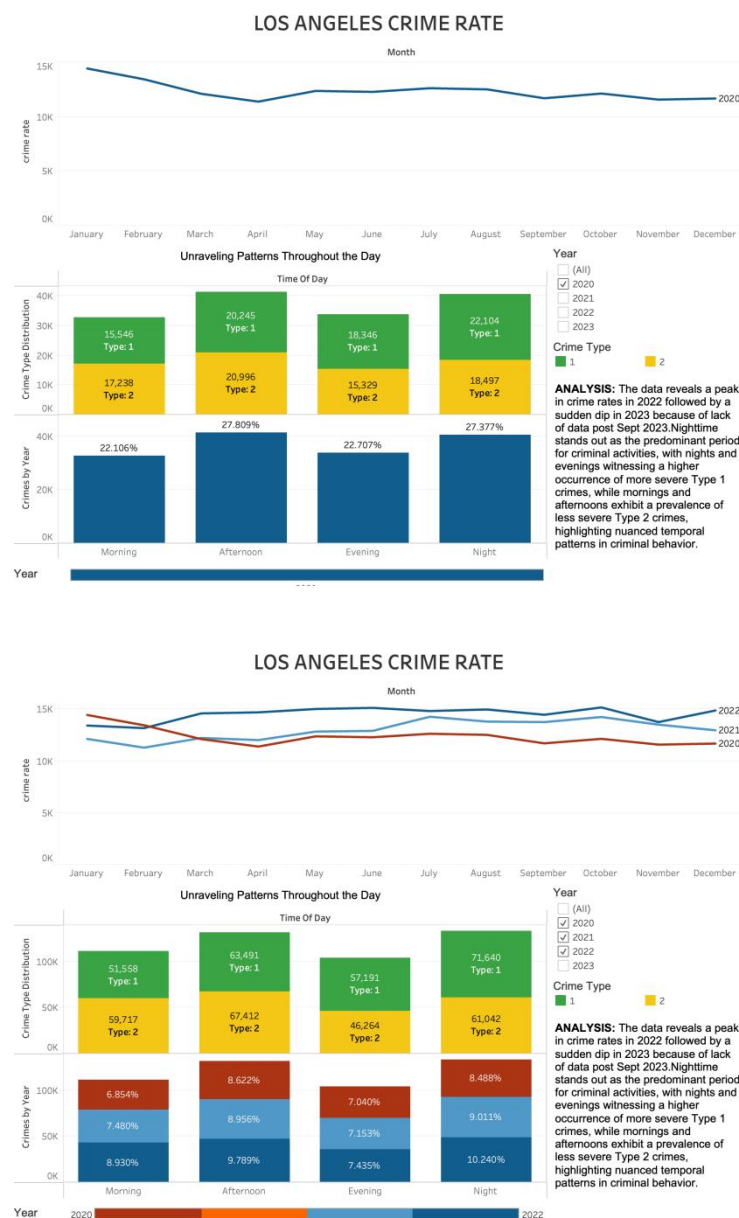


Fig 13 : Tableau Dashboard 4

- a) Year 2020
- b) Years 2020-2022

X. Conclusion

Our comprehensive analysis of the LAPD police report dataset has provided valuable insights into crime trends and patterns within the City of Los Angeles. By addressing critical business problems related to crime rates, demographics, and temporal aspects, we aimed to contribute to public safety strategies and law enforcement optimization.

Through data cleaning and preprocessing, we ensured the dataset's integrity and readiness for analysis, addressing issues like inconsistent formatting and missing values. The exploratory data analysis (EDA) phase allowed us to understand the dataset's structure and identify key trends, helping us formulate targeted business problems.

The integration of our dataset with PostgreSQL and Tableau enabled dynamic visualizations tailored to each business question. The interactive Tableau dashboard, in particular, offers a user-friendly interface for law enforcement to gain actionable intelligence on crime rates, areas with high or low crime rates, and temporal patterns.

Our machine learning endeavor focused on predicting demographic attributes of crime victims, revealing challenges associated with the complexity of predicting such information accurately. While achieving accuracy levels of 40-50%, we recognize the need for continuous refinement and exploration, suggesting areas for improvement such as ensemble machine learning techniques.

The results and analysis provided insights into the relationship between demographics and crime, prevalent crime types, and the time lapse between crime occurrence and reporting. Law enforcement can use this information to tailor their strategies, allocate resources efficiently, and enhance public safety.

In conclusion, our analysis serves as a foundation for ongoing efforts to refine and expand upon these findings. Continuous collaboration between data scientists, law enforcement, and policymakers is essential for leveraging data-driven insights to address evolving challenges in crime prevention and public safety.

XI. Limitations and Future Work

- **Refining Machine Learning Models:** Future efforts will center on refining machine learning models through the strategic use of ensemble techniques and hyper-parameter tuning, aiming to enhance predictive accuracy.
- **Exploration of Deep-Learning Integration:** Consideration will be given to integrating deep-learning models into our approach, seeking to capitalize on their potential for improved analytical capabilities.
- **Continuous Data Collection for Trend Detection:** The continuous collection of new crime data is crucial for identifying emerging trends not visible in the current dataset, ensuring the ongoing relevance and effectiveness of the project. Additionally, harnessing suspect data holds the potential to further enhance our model predictions, contributing to the ongoing improvement of our analytical capabilities.

XII. References

1. <https://data.gov/>
2. https://youtu.be/y5P90Qme13k?si=rsCCV6wrzP_G0w9j
3. <https://www.youtube.com/watch?v=S8ROhkhpQQE&t=86s>
4. <https://www.youtube.com/watch?v=M2NzvnfS-hI>
5. https://help.tableau.com/current/pro/desktop/en-us/examples_postgresql.htm
6. <https://stats.stackexchange.com/questions/258938/pca-before-random-forest-regression-provide-better-predictive-scores-for-my-data>
7. <https://www.analyticsvidhya.com/blog/2020/03/one-hot-encoding-vs-label-encoding-using-scikit-learn/>
8. <https://hevodata.com/learn/tableau-postgres/>
9. <https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.mllib.tree.RandomForest.html>
10. <https://spark.apache.org/docs/3.1.3/api/python/reference/api/pyspark.ml.classification.RandomForestClassifier.html>
11. <https://medium.com/edureka/tableau-dashboards-3e19dd713bc7>