Analysing Crime, Arrests and Traffic Collisions in Los Angeles during 2021

Preena Darshini .
School of Computing
National College of Ireland
Dublin, Ireland
x22238590@student.ncirl.ie

Priscila Cristina da Silva de Oliveira
School of Computing
National College of Ireland
Dublin, Ireland
x23157003@student.ncirl.ie

Reyna Vargas Antonio
School of Computing
National College of Ireland
Dublin, Ireland
x23127635@student.ncirl.ie

Abstract - **This report provides an extensive analysis of various patterns for crime, arrest and traffic collisions in Los Angeles for the year 2021 across multiple dimensions. All three large datasets were sourced from a government website and these datasets were extracted and transformed to understand the nuances in crime, arrest and traffic collisions activities. Various charts and graphs are employed to extract key findings that reveal important and different variations in all three datasets. The outcomes of this project can help policymakers and community leaders devise strategic policies and plans to prevent and manage crime, arrest offenders on time and help victims and manage traffic collisions better.**

Keywords—XML, JSON, Python, MongoDB, PostgreSQL.

I. Introduction

Our aim is to find related datasets that can be analysed to gain insights into data related to crime, arrest, and traffic collisions in Los Angeles for the year 2021. Crime data [1] is in XML format, while traffic collision [2] and arrest data [3] are in JSON format. All three datasets have been analysed in parallel to gain an understanding of the trends in crimes, arrests and traffic collisions that occurred in Los Angeles in the year 2021. The source for these datasets was gathered from the U.S. Government's Open Data website. The main motivation of this analysis is to facilitate informed decisions concerning public safety and law enforcement strategies. The findings aim to provide useful information for policymakers, law enforcement agencies, and community leaders to increase the safety and well-being of those living in Los Angeles.

II. Methodology

A. Crime Data

The Crime dataset [1] extracted from the U.S. Government's Open Data website is in XML format, which is semi-structured data. It contains 29 columns and 443351 rows. Only rows related to the year 2021 have been extracted and stored in MongoDB, which is an open-source NoSQL database. The number of rows stored in MongoDB for the year 2021 is 208283. MongoDB was chosen to ensure that the data is securely stored and indexed for fast and efficient querying. A large number of missing values were found in modus operandi codes (mocodes), victim sex (vict_sex), victim descent (vict_descent), premises information (premis_cd, premis_desc), weapon information (weapon_used_cd, weapon_desc), cross street (cross_street) and additional crime codes (crm_sd_2, crm_cd_3, crm_cd_4). Data preprocessing and cleaning were done before visualisations and analysis.

B. Arrest Data

The "Arrest Data from 2020 to Present" dataset brings information regarding arrest occurrences in the City of Los Angeles in the United States [3]. Its data is transcribed from original reports and for that reason inaccuracies are expected [3]. The original dataset contains 255787 records distributed in 26 columns corresponding to the period of 2020 to present. This study focuses on the year of 2021 and, for this reason, the data was filtered before being stored in MongoDB. The filtered dataset shows 66773 rows and 26 columns.

Among the data, one can find information regarding the reporting of the arrest, date and district where it was disclosed, age and sex of the offender, charge description, and geographical information such as latitude and longitude of the event and address. It is important to notice that, for legalities related to data protection, the address mentioned corresponds to the nearest hundred block [4].

C. Traffic Collision Data

The Traffic Collision dataset [3] was extracted in JSON format, which is semi-structured data. This format stores two keys, "meta" and "data", in the latter, all the data of this dataset is stored. The information in the key "data" was stored in a list called traffic_json using Python, and contains 600755 sublists. The data extracted is related to the year 2021 and it has been stored in MongoDB. A total of 18992 documents were stored in MongoDB in a document called Traffic_collision. Using Python, these documents were extracted from MongoDB and stored in a data frame, which contains 25 columns and 18992 rows. There were some missing values found in the columns "MO Codes", "Victim Age", "Victim sex", "Victim Descent", "Cross Street", "Census Tracts", "Precinct Boundaries", "La Specific Plans", "Council Districts" and "Neighborhood Council (Certified). Henceforth, this data will be preprocessed and then loaded into PostgreSQL and retrieved to better analyse the data through visualizations.

D. Detailed Description

The flowchart in Fig. 1 gives the step-wise process followed by the analysis. Extract, Transform, and Load (ETL) processes have been followed here.

- ***Extract Phase:*** Data extraction involves retrieving data from various data sources. The Crime data is sourced in XML format. Arrest data and Traffic Collision Data are in JSON format.

- ***Transform Phase***:
  Crime Data: The extracted data is stored in MongoDB. This acts as a preliminary data store. Crime data is cleaned, normalized, and processed to make sure the data is in the correct format and structure for analysis.

Binarization: A new column (vict_age_bin) is created.The victim age (vict_age) is set to 1 if the age is greater than 50; otherwise, it is set to 0.
Discretisation: The get_season function is an example of discretization where the months are divided and categorised into seasons.

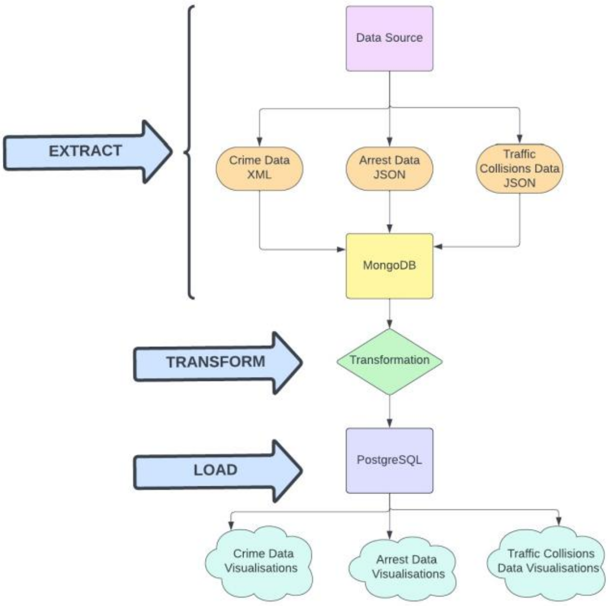
EXTRACT
Data Source
Crime Data XML
Arrest Data JSON
Traffic Collisions Data JSON
MongoDB
TRANSFORM
Transformation
LOAD
PostgreSQL
Crime Data Visualisations
Arrest Data Visualisations
Traffic Collisions Data Visualisations

Fig. 1. Flowchart

Arrest Data: the columns "Report ID", "Charge Description", "Address" and "Cross Street" were removed from the dataset for not being relevant to the analysis that will be later proposed. The data were verified for missing values. The columns "Disposition Description", "Charge Group Code", "Charge Group Description", "Booking Date", "Booking Time", "Booking Location Code" and "Booking Location" presented several missing values. It was found that some of the missing data does not correspond to inaccuracies, given that in many cases the person is cited and released from custody, not being booked at a detention facility. The missing values were replaced by "NA" for object data types and by "NaN" when dealing with non-object data types. The dataset does not have duplicated values. The columns "Lat" and "Lon" were respectively renamed to "Latitude" and "Longitude". Additionally, all column names were converted to lowercase and spaces were replaced with the underscore symbol. The columns with time information were previously in the format HHMM and were converted to HH: MM in order to improve the user experience.
Traffic Collision Data: The data were stored in MongoDB, which allows manipulation, iteration and printing of objects retrieved through PyMongo in Python, because it is compatible with dictionaries and lists. The Traffic Collision data was stored in Python as a data frame, where it was cleaned, and transformed into columns in the correct format. Outliers and missing values were detected and handled efficiently.

- ***Load Phase:*** The processed Crime data, Arrest data, and Traffic Collision data were loaded into PostgreSQL, where it was queried and used for additional analysis. ACID properties i.e. Atomicity, Consistency, Isolation, and Durability are applied in the case of PostgreSQL.

E. *Tools and Technologies Used*

This project utilizes a variety of technologies and programming language techniques.

- Python: xml.etree.ElementTree, pandas, numpy, matplotlib, plotly and seaborn libraries are used. The "datetime" module is also used for manipulating date and time.
- MongoDB: This NoSQL database is used to store the data that is extracted after parsing the XML file and JSON files. The "pymongo" library enables the connection to MongoDB database.
- PostgreSQL: This is a relational database. The transformed data is stored and retrieved for further analysis using psycopg2 and SQLalchemy libraries allowing data insertion and retrieval in Python for deeper analysis.

III. Literature Review

By using data mining and statistical methods, new algorithms were developed for the analysis of crime patterns and crime prediction in different regions [5]. It focuses on the capacity of Big Data Analytics to process large and complex datasets and offers insights into urban dynamics, especially public safety. Recent studies related to cities like San Francisco, Chicago, and Philadelphia have applied these technologies to increase the predictive accuracy of crime trends in urban cities. There has been a recent increase in data visualisation techniques, especially in the field of crime analysis [6]. This marks a great shift in criminological studies. It is argued that traditional crime analysis falls short when it comes to handling large datasets [6]. This study aims to combine many aspects of crime, arrests and traffic collisions in Los Angeles to contribute to evolving criminology and offer scalable and actionable insights for law enforcement and government agencies in Los Angeles.

Based on the Los Angeles Times newsletter [7], this city has become the traffic violence capital of the nation. There were more people killed on traffic accidents than gun-related events. From 2020, it increased around 30% the number of traffic deaths. In order to provide a careful analysis of roadway traffic data, it is important to identify variables related to fatal accidents. Data mining techniques could help in understanding the data available and provide insights in how to reduce fatality rate[8].

IV. RESULTS AND EVALUATION

A. *Crime Data Visualisations*

The Seaborn python library is mainly used for visualisations.

- Correlation Heat Map

The map in Fig. 2 gives details about the strong predictors and the correlation coefficients between different predictors. Latitude and Area show a positive correlation of 0.34 and Latitude and Longitude have a strong negative correlation. This is mainly useful for modelling and feature selection.

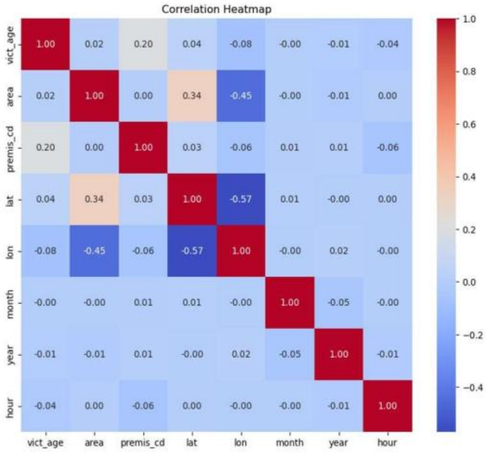

Fig 2. Correlation Heat Map

- Top 20 Crime Code Distribution

This horizontal bar chart in Fig. 3 shows the frequency of different crime types. Each bar represents the crime type, and the length of the bar indicates the frequency of that crime. Vehicle theft is the crime with the highest frequency. Trespassing has the lowest frequency.

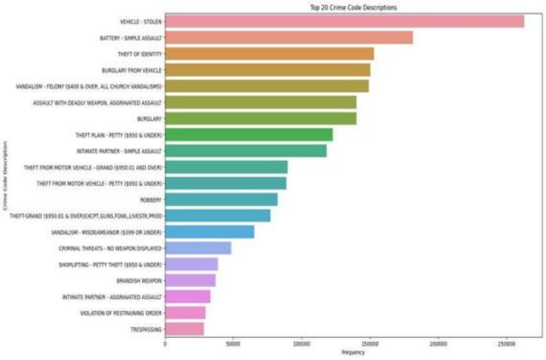

Fig. 3. Top 20 Crime Code

- Victim's Gender Distribution

Fig. 4 shows two charts for victim gender distribution. In the bar chart, each bar represents different categories of genders with the frequency of the victims. The pie chart displays the same; however, the proportion is shown in percentage. Males constitute the largest percentage. "X" might be non-binary or unspecified genders. There is no "H", i.e., Hermaphrodites.

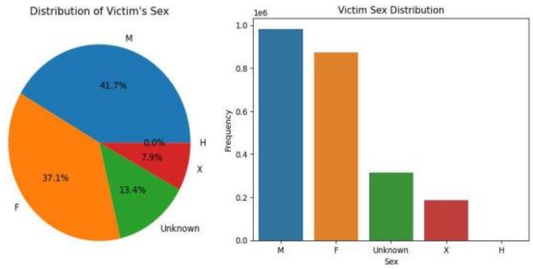

Fig. 4. Victim's Gender Distribution

- Top 20 Victim Descent Distribution

Fig. 5. shows a horizontal bar chart for the top 20 Victim Descent. The letters represent the abbreviations for different racial and ethnic groups. The "H" has the longest bar followed by "W" and "B". The unknown category suggests that the descent was not recorded.

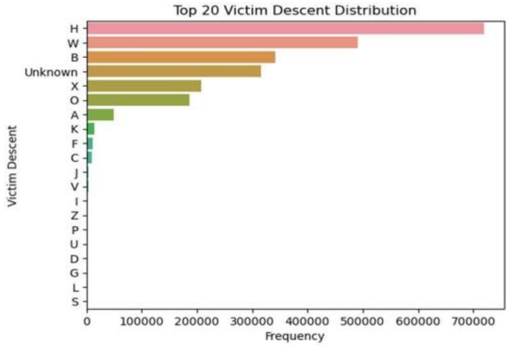

Fig. 5. Victim Descent Distribution

- Monthly Crime Trends

Fig. 6 shows a line chart that displays the number of crimes per month in the year 2021. 1 to 12 on the x-axis indicates the months and the y-axis shows the number of crimes. From the graph, it can be inferred that there is a decrease in crimes in February and an increase in March. After March, there is a continuous increase and a gradual decrease from July and peaks in October.

- Crime Distribution by Area

Fig. 7 is a vertical bar chart indicating the number of crimes in different areas of Los Angeles. "Central" has the highest number of crimes, followed by "77th Street". "Foothill" has the lowest crime numbers.

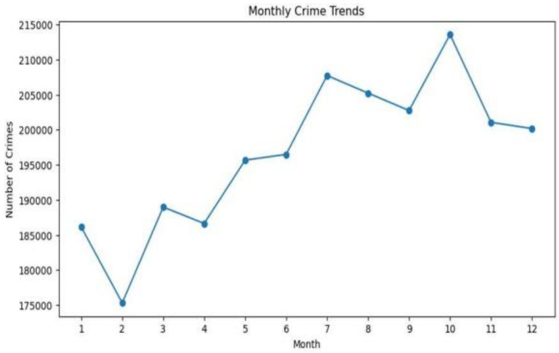

Fig. 6. Monthly Crime Trends

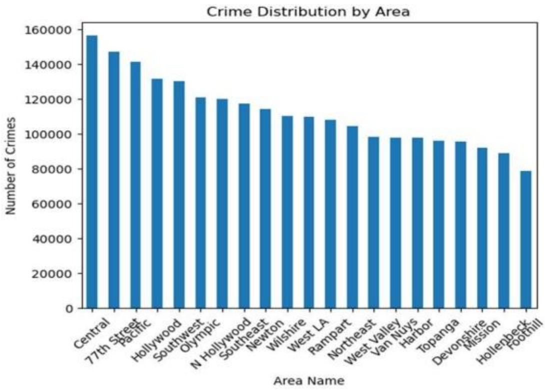

Fig. 7. Crime Distribution by Area

- Monthly Crime Distribution in Central

Fig. 8 is a line graph that shows the number of crimes that occurred in Central. There has been a substantial increase in crimes in 2021.

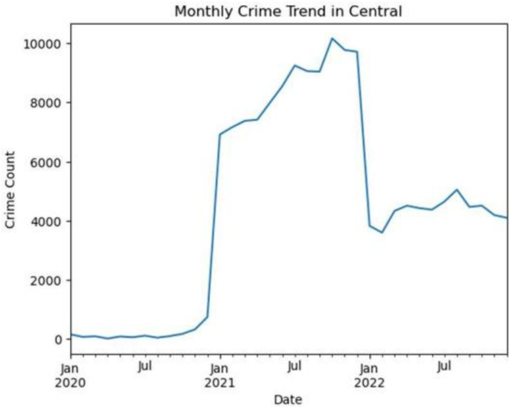

Fig. 8. Monthly Crime Trend in Central (2021)

- Crime Distribution by Time of the Day

Fig. 9 depicts a vertical bar graph that displays the frequency of crimes that occurred at different times during the day. The x-axis represents a 24-hour clock. Most crimes occur at noon. There is a general increase in the number of crimes in the afternoons and evenings.

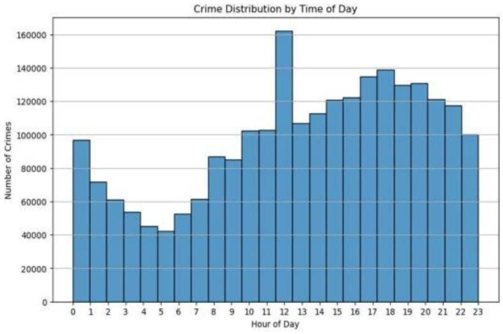

Fig. 9. Crime Distribution by Time of Day

- Top 10 Crime Types

Fig. 10 shows a vertical bar graph which displays the top 10 crime types that took place in 2021 in the City of Los Angeles. Vehicle stolen is the most frequent crime followed by battery-simple assault and other types of crime. There is a visual decrease in height suggesting a descending order in frequency for the top 10 crime types.

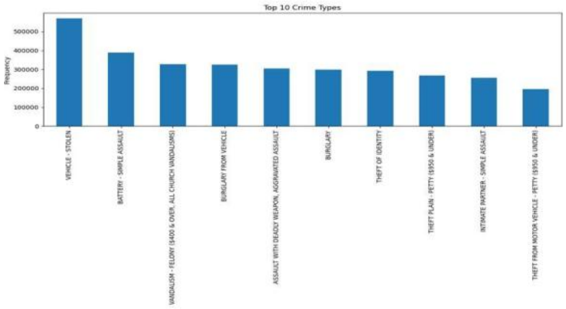

Fig. 10. Top 10 Crime Types

- Top 10 Weapons Used in Crimes

Fig. 11 shows the bar graph for different types of weapons used in crimes. Most of the weapons were not reported. The second highest type of weapon used is Strong-Arm followed by other weapons, verbal threats, handguns, semi-automatic pistols, firearms, knives and so on.

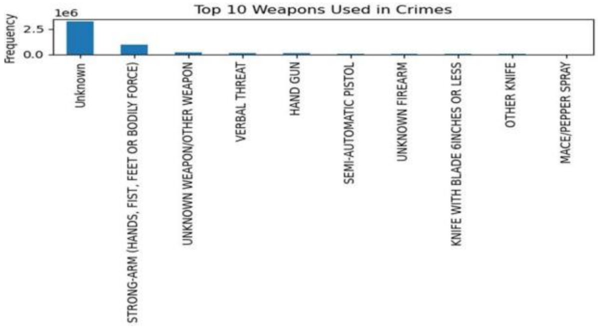

Fig. 11. Top 10 Weapons

- Victim Age Distribution

Fig. 12 is a histogram that illustrates the victim age distribution. The x-axis represents the age of victims and the highest frequency for age group 0 is shown which indicates that the age is unknown and unspecified. There is an increase in younger adult age ranges. This trend gradually decreases with elderly victims. This graph can be used to understand which age groups are more vulnerable to crime.

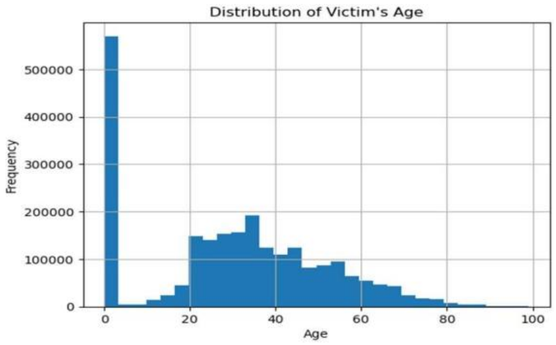

Fig. 12. Distribution of Victim’s Age

- Crime Distribution by Day of the Week

Fig. 13 shows the crime distribution by day of the week. It depicts the frequency of crime each day. The chart indicates that the crime occurrences are fairly distributed. There is a slight increase in the number of crimes committed on Fridays.

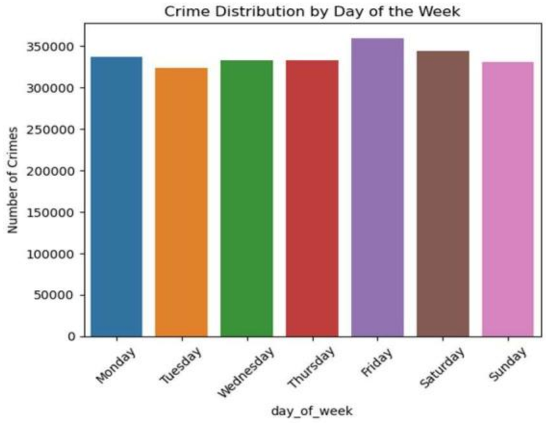

Fig. 13. Crime Distribution by Day of the Week

- Crime Distribution by Season

Fig. 14 shows a pie chart that displays the percentage of crimes that took place in each of the four seasons, namely Summer, Winter, Autumn, and Spring. Autumn has the highest percentage of crimes at 26.2%, followed by Summer with 25.8%, Spring with a lower percentage of 24.2% and, finally Winter with 23.8%.

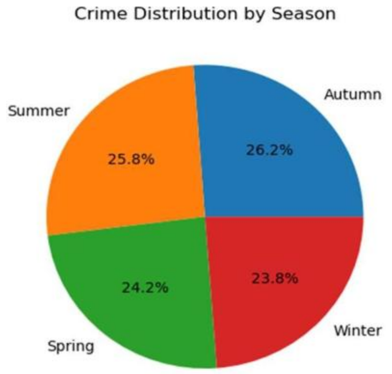

Fig. 14. Crime Distribution by Season

B. Arrest Data Visualisations

- Age Distribution

The histogram shown in Fig. 15 illustrates the frequency of the offender’s age when arrested. The range corresponding to 25 to 35 years is the one with the biggest number of offenders. It is also possible to observe that the ranges 5 to 15 years and 75 to 85 years are not significant, however, they could inspire social solutions related to those groups. Employing actions related to the younger group, for instance, could contribute with valuable insights for effective strategies in crime prevention.

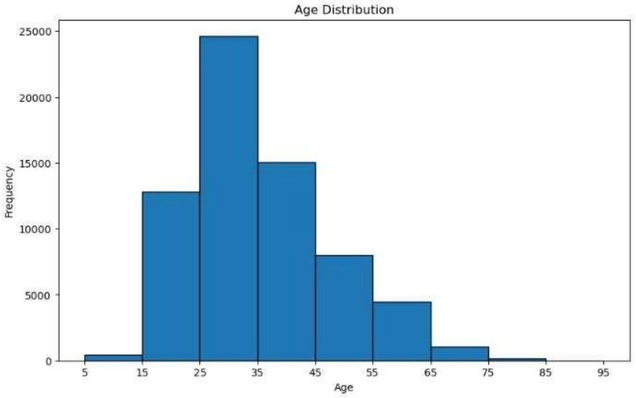

Fig. 15. Age Distribution

- Report Type by Area

Fig. 16 illustrates a stacked bar chart for Report Type by Area. The two types of report present in the dataset are “Booking” meaning that a person was booked in a detention facility and “RFC” for those who receive a written notice or ticket and are then released [4]. Area 14 presents the highest total of reports, and area 12 stands out with the biggest number of bookings.

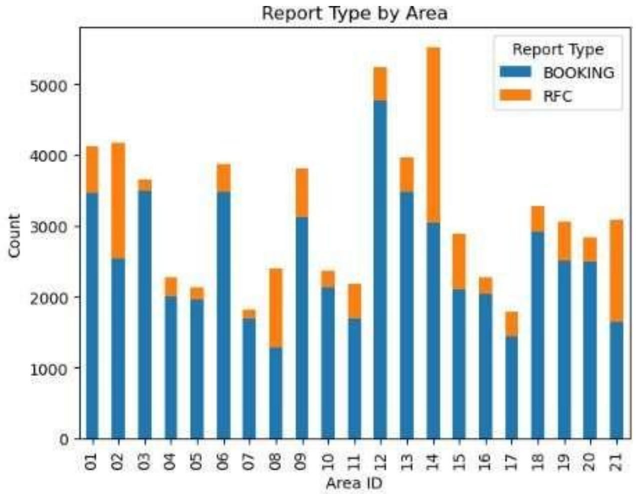

Fig. 16. Report Type by Area

- Number of Arrests by Area

The column graph depicted in Fig. 17 highlights the main city areas where arrests are registered. It could be used when allocating the police force since those areas are more prone to witness incidents.

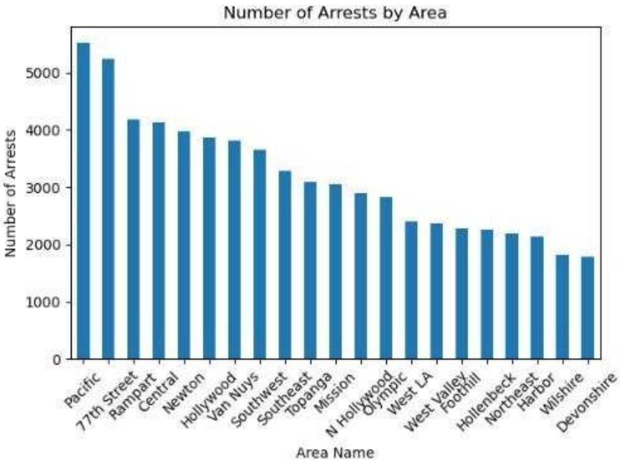

Fig. 17. Number of Arrests by Area

- Arrests by Charge Group

When analysing the charges associated with arrests in the City of Los Angeles, it was observed that most of the charges are associated with violations that do not fall into a major group ("Miscellaneous Other Violations") and Aggravated Assaults as can be seen in Fig. 18.

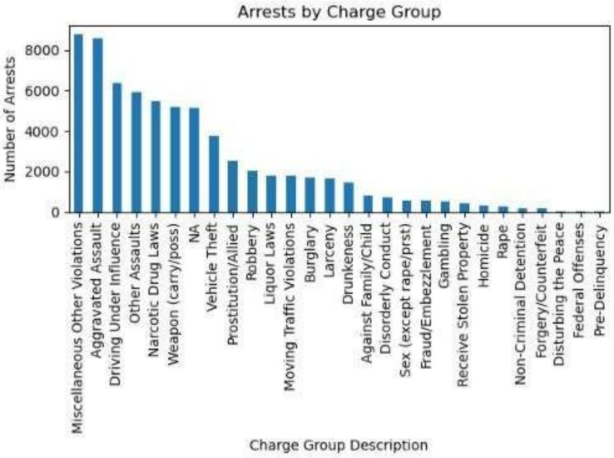

Fig. 18. Arrests by Charge Group

- Arrest Type Code Frequency

Arrests can be classified into five categories: F (felony), M (misdemeanor), I (infraction), O (other) and D (dependent) [4]. Felonies and misdemeanor are the main types of charge categorizing arrests as shown in Fig. 19.

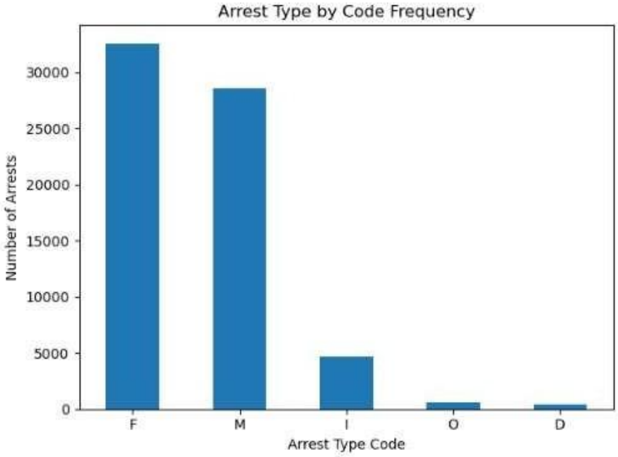

Fig. 19. Arrests Type by Code Frequency

- Percentage of Arrests by Month

Fig. 20 shows the percentage of arrests by month. The distribution over the year is quite balanced. However, it is possible to notice December being the month with the smallest percentage of arrests and

September the month with the highest number.

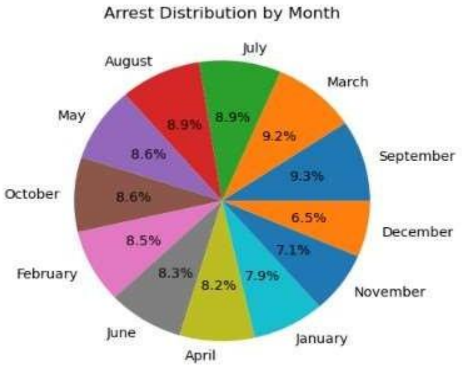

Fig. 20. Arrest Distribution by Month

- Arrests by Hour of the Day

The column distribution seen in Fig. 21, shows that most arrests happen during nighttime in hours where the pedestrian traffic is smaller when compared to daytime hours.

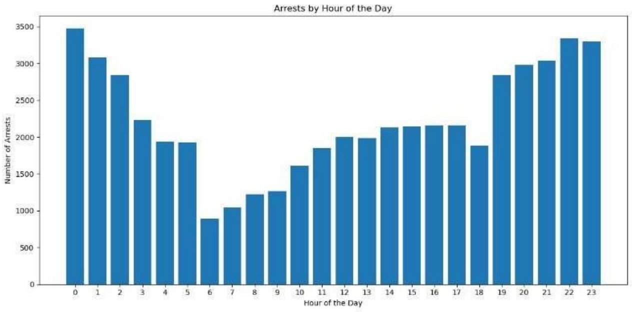

Fig. 21: Arrests by Hour of the Day

C. *Traffic Collision Data Visualisations*

- Number of Traffic Collisions by Month

The following chart shows that there were more accidents in January than in the rest of the months. The accidents reported in January were 1767; meanwhile, from February to December, there were an average of 1000 reports per month.

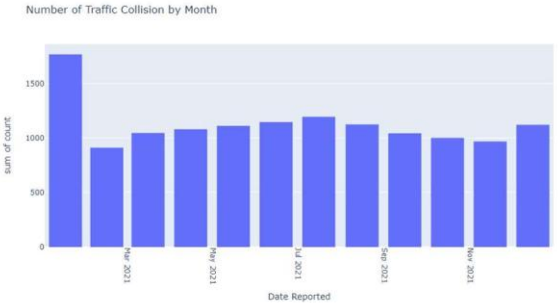

Fig. 22: Number of Traffic Collisions by Month

- Traffic Collisions by Area and Gender

The chart in Fig. 23 displays the distribution of traffic collisions depending on the Area in Los Angeles and Victim Sex; there were more cases of men involved in vehicle collisions than women. The probability that this happens is two or three times greater.

The three areas with the least risk of traffic collisions are:

1. West LA: there were 248 men associated with traffic collisions versus 140 women.
2. North East: In this area, men involved were 296 versus 169 women and 13 people that are of unknown gender.
3. Pacific: there were only 170 women implicated versus 329 men

Conversely, the most dangerous area is:

1. 77th Street: A total of 1251 cases, implicating 741 men, 452 women, and 58 people that are of unknown gender.

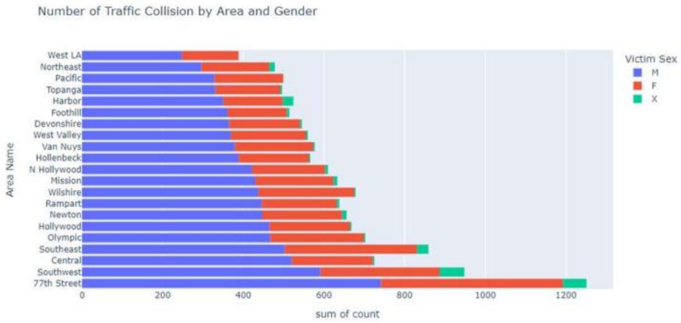

Fig. 23 : Traffic Collisions by Area and Gender

- Traffic Collision according to where an incident took place

The Los Angeles Police Department has classified these incidents in 23 places where they occur frequently. Most of the incidents occurred on the Street with 12932 cases; the second place most reported with 307 incidents was a Parking Lot; and the third was on a sidewalk with 153 incidents. There were random incidents reported in Medical Marijuana Facilities / Business, Church/Chapel, Restaurant/ Fast Food and Yard.

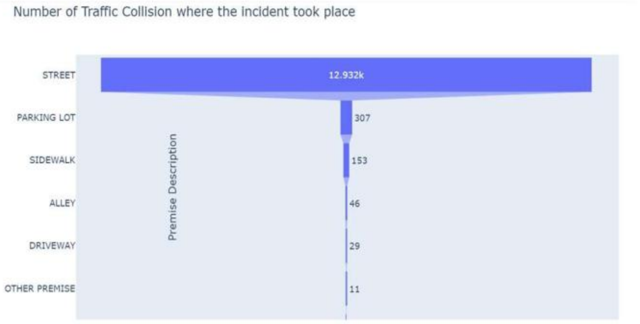

Fig. 24. Traffic Collision according to where an incident took place.

- Percentage of victims by their Descent

Los Angeles is one of the most culturally diverse and largest regions in the United States, where there is the highest probability that a victim is a descendant of another country. Fig. 25 displays that 45.7% of victims are descended from Hispanic/Latin Mexican people, 20.4% are white people, and 17.6% correspond to black people. Relative to the United States Census Bureau [9] 48.1.3% of the population is Hispanic or Latino and 28.1% percentage is white alone (not Hispanic or Latino) in Los Angeles, California, which explains the percentage of incidents involved in traffic collisions.

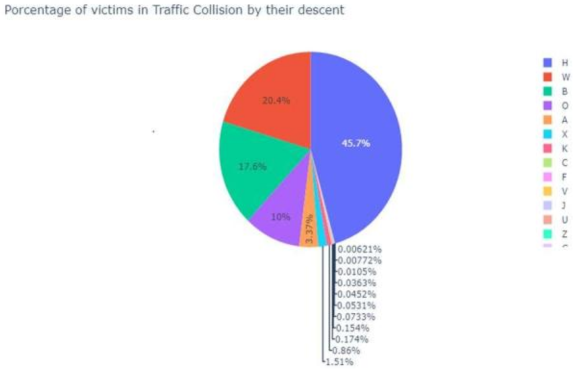

Fig. 25. Percentage of victims by their Descent.

- Time Occurred of Traffic Collision according to Victim Descent

Fig. 26 illustrates the time occurred obtained in 24 hour military hours, for which the range is between 0 and 2399. The histogram showed in Fig. 26 displays the time when an incident occurred from 00:00 hours to 24:00 hours and how many people reported the incident. It can be seen that from 04:00 to 04:49 hours, it was the time when there were fewer incidents; from 17:00 to 17:49, it was the time with more incidents, even though there was not substantial difference between 16:00 and 21:00 hours because at this time there were 600 reported incidents or more. The graph also shows that there are more descendants involved, like Pacific Islanders (red colour) and Asian people green colour).

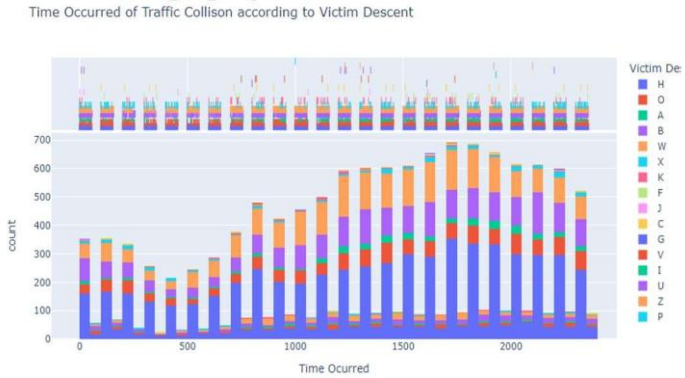

Fig. 26. Time Occurred of Traffic Collision according to Victim descent.

- Victim Age by Gender

The box chart shown in Fig. 27 displays the distribution of victims involved in incidents. The concentration of victims' ages ranges from 28 to 50 years for men and 27 to 50 years for women. The median age differs slightly: 38 years for men and 36 years for women. For people of unknown gender, the youngest was 19 years old and the oldest was 71 years old. This indicates that age is not a limiting factor; anyone can be involved in an incident, regardless of age, even while simply walking around.

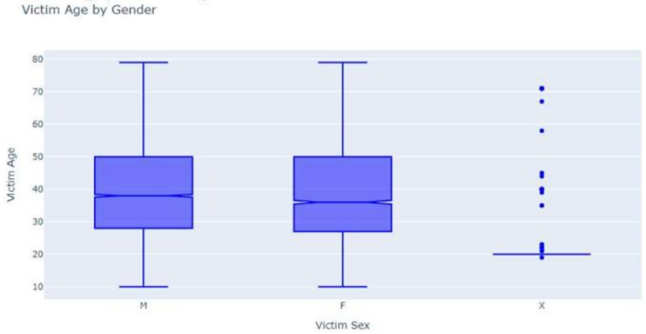

Fig. 27: Victim by Gender

- Probability of being in Traffic Collision by Gender

Fig. 28 displays the probability of men and women participating in an incident. The cumulative distribution for both men and women is similar, indicating that the probability of someone being involved in an incident is 32% when a person is in their 30s. This probability increases proportionally with age. In contrast, the case of unknown gender individuals is atypical, as they had an 84% probability of involvement in incidents at the age of 20s.

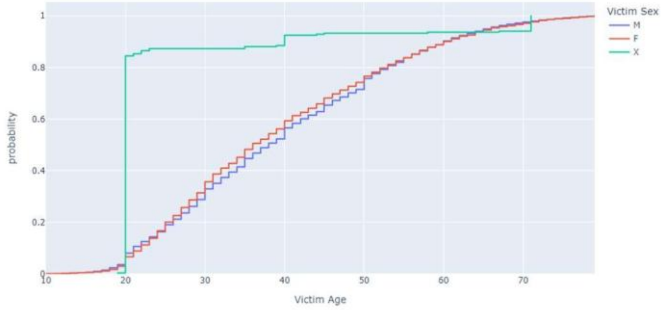

Fig. 28: Probability of being in traffic collision by gender.

V. Conclusion and Future Work

It can be concluded that, on analysing the above graphs and charts, there are some common findings. From Fig. 12 and Fig. 15 it can be seen that the most vulnerable to crime is the young adult to middle age category, and the younger adults are involved in incidents that lead to arrests. From Fig. 7, Fig. 17, and Fig. 23, it can be inferred that despite the Pacific Area leading in arrests, it does not have high number of crimes. This could indicate proactive arrest policies or low crime reporting rates in that area. The Pacific area does not have very high traffic collisions as well. Comparing Fig. 5 and Fig. 25, the "H" racial group has the greatest number of victims by descent, and at the same time, it is also the race that is most affected by traffic collisions.

Machine Learning models and Advanced Statistical analysis can be used to predict crime, arrests, and traffic collisions in and their application are suggested as future work. Further advanced analysis can be performed to gain deeper insights.

REFERENCES

[1] Data.gov,Crime Data from 2020 to Present, 2023. [Online] Available: https://catalog.data.gov/dataset/crime-data-from-2020-to-present [Accessed on: Nov. 01, 2023].

[2] Data.gov,Arrest Data from 2020 to Present, 2023. [Online] Available: https://catalog.data.gov/dataset/arrest-data-from-2020-to-present [Accessed on: Nov. 01, 2023].

[3] Data.gov,Traffic Collision Data from 2010 to Present. [Online] Available:https://catalog.data.gov/dataset/traffic-collision-data-from-2010-to-present [Accessed on: Nov. 01, 2023].

[4] About this dataset, 2023. [Online] Available: https://data.lacity.org/Public-Safety/Arrest-Data-from-2020-to-Present/amvf-fr72/about_data [Accessed on: Nov. 01, 2023].

[5] "Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data Mingchen Feng;Jiangbin Zheng", Jinchang Ren; Amir Hussain; Xiuxiu Li; Yue Xi; Qiaoyuan Liu, 2019.

[6] S. Walker and S. Sharma, "Data Visualization Tool for Covid-19 and Crime Data," 2021 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2021, pp. 238-242, doi: 10.1109/CSCI54926.2021.00111.

[7] "Traffic violence continues to surge in L.A. Why can't (or won't) we make streets safer?," Los Angeles Times, Jan. 19, 2023. https://www.latimes.com/california/newsletter/2023-01-19/essential-california-2022-traffic-violence-essential-california [Accessed on: Nov. 16, 2023].

[8] L. Li, S. Shrestha, and G. Hu, "Analysis of road traffic fatal accidents using data mining techniques," 2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA), Jun. 2017, doi: https://doi.org/10.1109/sera.2017.7965753.

[9] "U.S. Census Bureau QuickFacts: Los Angeles city, California," www.census.gov, 2022. https://www.census.gov/quickfacts/fact/table/losangelescitycalifornia/PST045222 [Accessed on: Nov. 16, 2023].