Computation and Visualization
IE6600
Spring 2025

Project 2 Report

Group 1

Barath Keshav Sriram Kumaran
Sam Yuen
Pooja Mishra

sriramkumaran.b@northeastern.edu
yuen.s@northeastern.edu
mishra.po@northeastern.edu

3/15/2025

Contents

Introduction

This dataset we picked from data.gov contains arrests in Los Angeles from 2020 to present day (Feb 2025). The data includes basic descriptions of the arrests and finer details on the location and time.

Crime data analysis plays a vital role in understanding patterns of law enforcement actions and societal trends. This project aims to analyze arrest records in Los Angeles from 2020 to the present using advanced data visualization techniques with the Seaborn library. The objective is to explore and identify trends in arrests based on various parameters, such as year, age, gender, descent, and disposition outcomes.

By leveraging the power of data visualization, we can uncover patterns that might indicate shifts in crime rates, the effectiveness of law enforcement policies, and the influence of socio-economic factors. The study follows a structured approach of data cleaning, preprocessing, exploratory data analysis (EDA), and statistical interpretation.

Through this analysis, we provide insights into yearly arrest trends, demographic distribution of arrests, and possible systemic influences affecting these statistics. This project serves as a case study in advanced data analytics, demonstrating how meaningful insights can be extracted from raw data using Python and Seaborn.

About the Dataset

The dataset utilized in this analysis is the "Arrest Data from 2020 to Present," obtained from data.gov. This dataset is maintained by the Los Angeles Police Department (LAPD) and records information on all arrests made within the jurisdiction of Los Angeles. It provides crucial details about individuals taken into custody, including demographic attributes, location, and legal disposition.

Key Features of the Dataset:

- **Arrest Date & Time**: Provides temporal information regarding arrests, allowing for time-series analysis.
- **Age & Gender**: Demographics of arrested individuals, useful for understanding crime patterns among different age groups and genders.
- **Descent Code**: Encoded ethnic background of individuals, which can be mapped to specific racial or ethnic groups for deeper analysis.
- **Arrest Location (Latitude & Longitude)**: Geospatial information that helps in identifying high-crime areas.
- **Disposition Description**: The legal outcome of the arrest, such as whether the suspect was released, prosecuted, or transferred.

Initial Observations:

- The dataset contains thousands of records, making it a valuable resource for statistical analysis.
- Some columns, such as "Cross Street" and "Booking Location Code," have a significant amount of missing values.
- The "Descent Code" column requires mapping to meaningful labels for better interpretability.
- There are variations in text formatting for certain columns, requiring standardization before visualization.

Dataset Preprocessing

Data Inspection

This 87 MB dataset of multiple years of arrests in LA has 335,485 rows of data and 25 columns. 10 of the columns are either integers or floats. 15 of the columns contain objects. 3 of them are data on the people themselves. 4 of them are about the time of the arrest and the filing. 9 of them are related to location. 6 of them are about the type of arrest that was made, with 2 of them being descriptions of the incident. Many of the columns are nominal data.

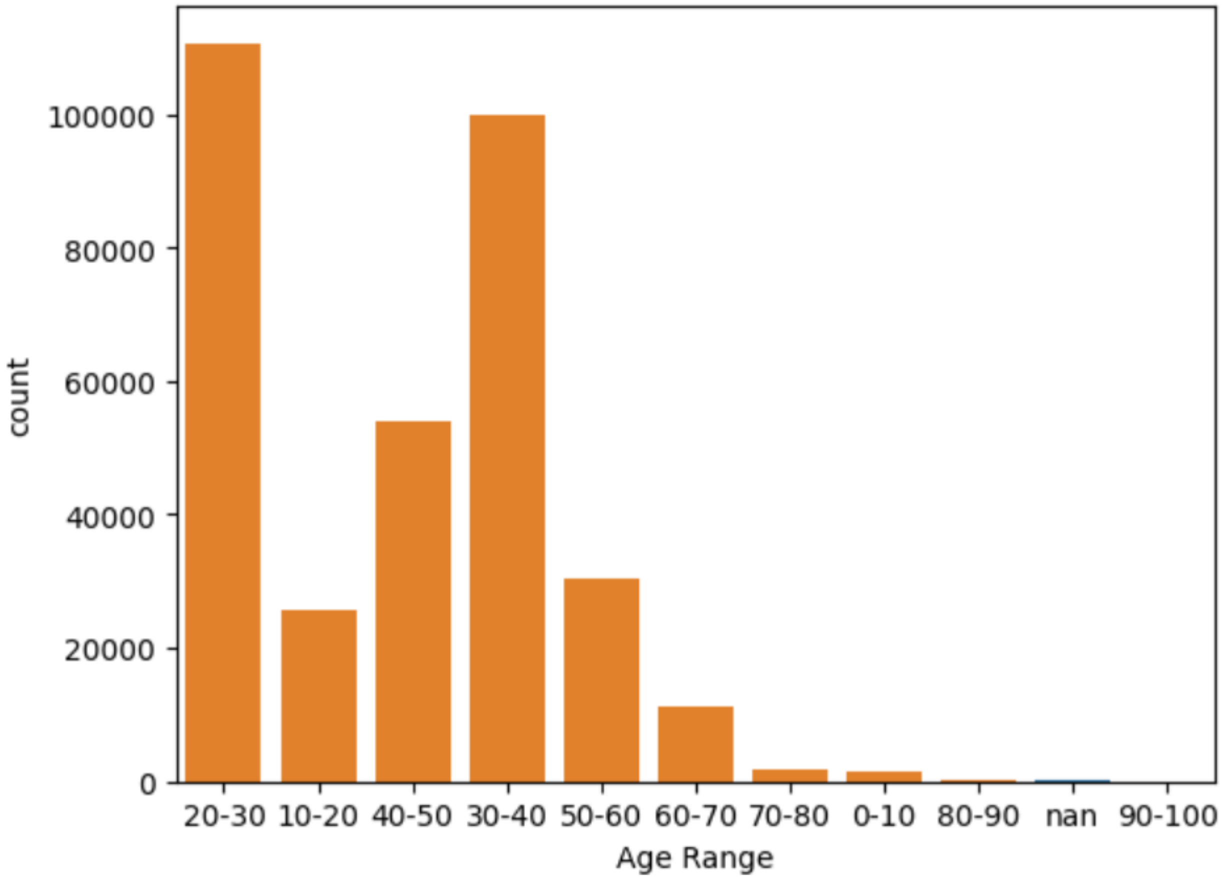
Some of the columns were self-explanatory but there were some columns for codes that those who don't work in law enforcement wouldn't be familiar with. The Report ID column was clearly the primary key for each arrest. Report Type was either "BOOKING" or "RFC" "BOOKING" means that the arrest was formally recorded with the individual's personal information being taken and them being held in custody. "RFC" means they were released from custody without being formally recorded into the system. There are codes for racial descent, charge group, arrest type, and charge. Not only are there addresses for the arrests, but there are also coordinates for the exact locations.

This dataset seems large, but there are some columns that have a considerable amount of missing data. The "Charge Group Code", "Charge Description" and "Charge Group Description" columns all have 32,135 missing values. "Disposition Description" has 30,900 missing values. "Cross Street" has 182,640 missing values. The "Book Date", "Booking Time", "Booking Location Code" and "Booking Location" columns all have at least 75,400 missing values.

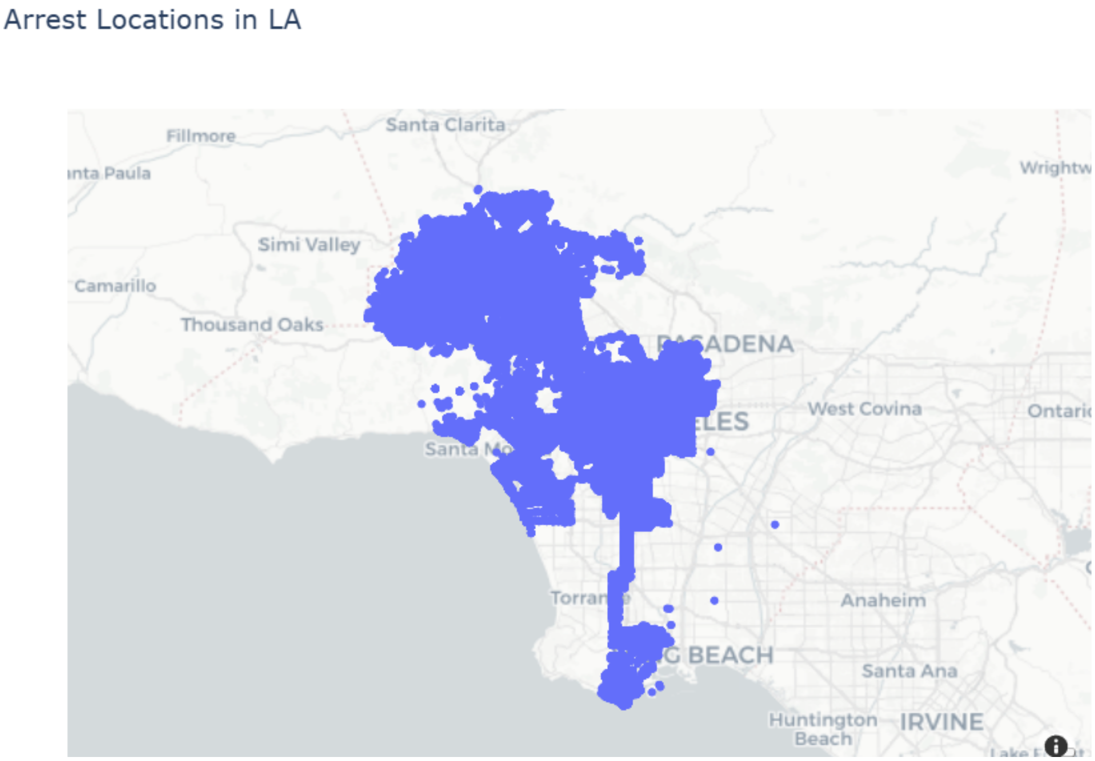
Data Cleaning and Preparation

We started out with removing the following columns we did not need for our analysis: "Charge Group Code", "Charge Group Description", "Charge Description", "Disposition Description", "Cross Street", "Booking Date", "Booking Time", "Booking Location" and "Booking Location Code" These columns had too many missing values, did not contain any insightful data, or were redundant. After that, we made the following columns to make it easier for us to understand and visualize the data: "Arrest Month", "Arrest Year", "Age Range", and "Descent"

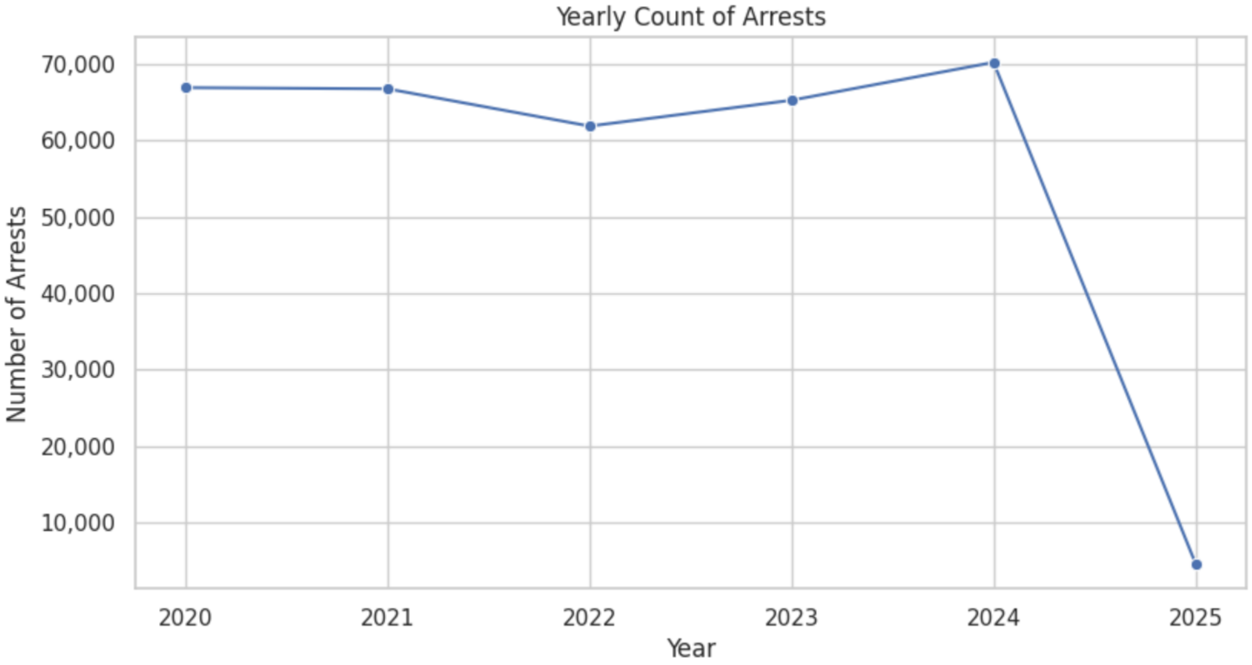
Exploratory Data Analysis (EDA)



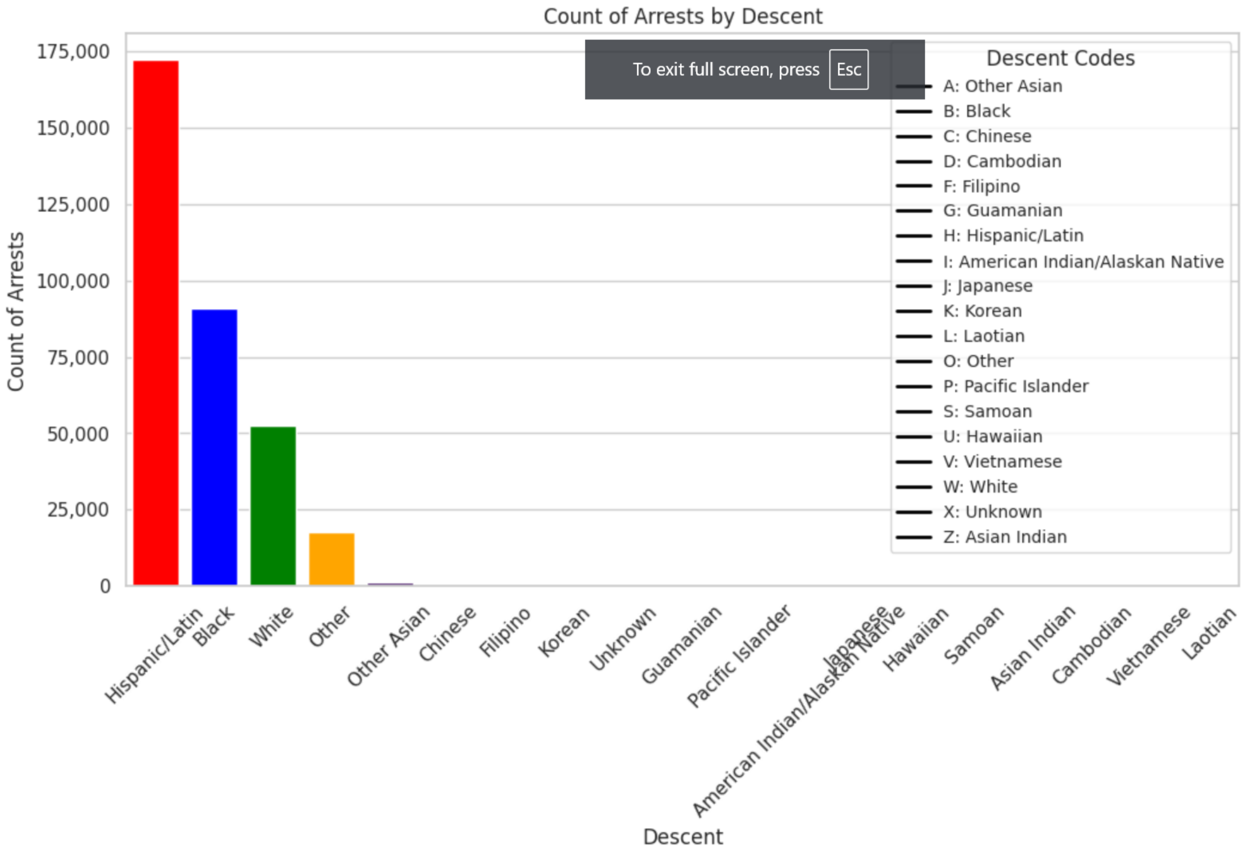
Most arrests occurred among individuals aged 20 to 30, significantly higher than other age groups. This trend suggests that people in their 20s are the most likely to be arrested, potentially due to factors such as increased risk-taking behavior, higher engagement in certain crimes, or greater police focus on this demographic. The second highest number of arrests was recorded in the 30-40 age group, reinforcing the idea that young adulthood is a period with a higher likelihood of arrests. Beyond the age of 50, there is a clear decline in arrest numbers, which could reflect a decrease in criminal activity, changes in police focus, or other societal influences. The extremely low arrest numbers among older age groups indicate that arrests are rare in the elderly population.


Arrest Locations in LA

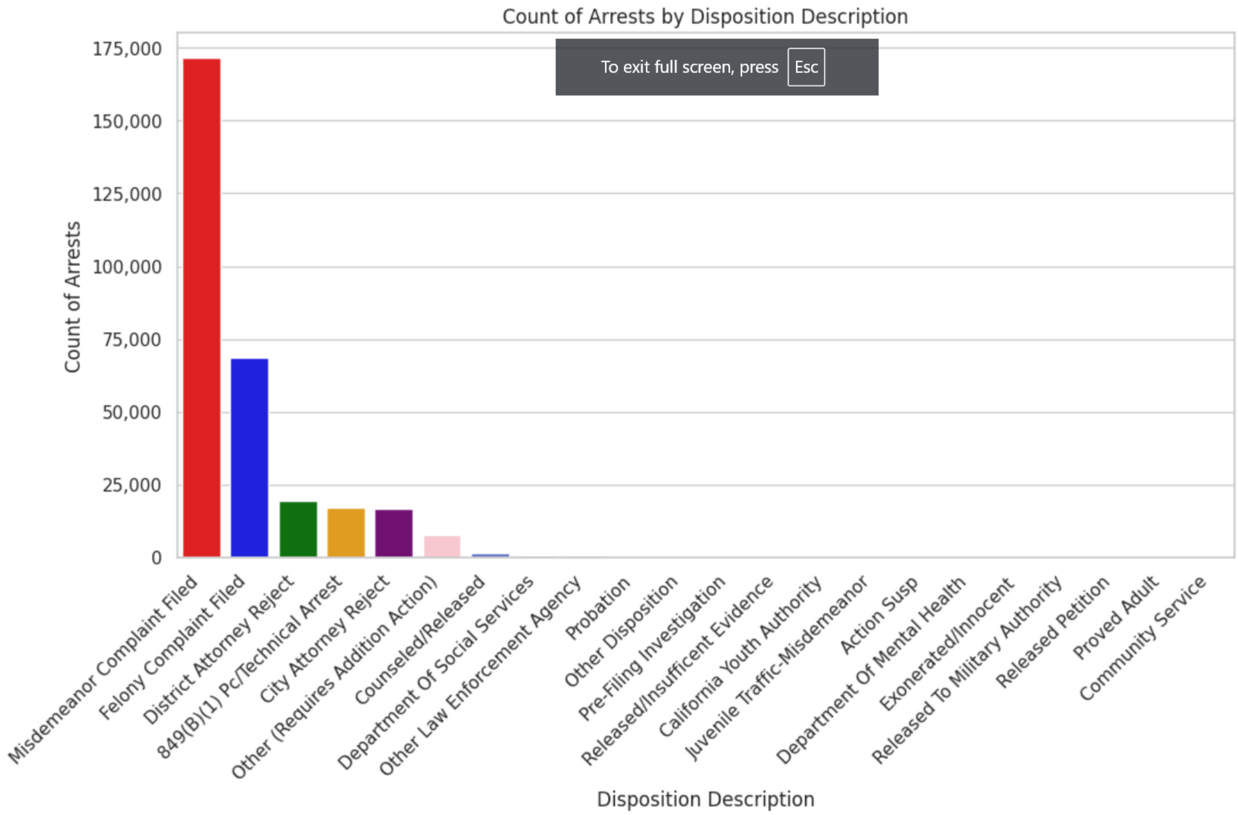
Arrest data indicates a dense concentration of arrests in the central area of Los Angeles, particularly in and around downtown LA, suggesting significantly higher crime rates in this region compared to surrounding areas. Additionally, there is a notable cluster of arrests along the coastline, especially in the South Bay area, including Torrance and Long Beach, which may be influenced by factors such as population density, tourism, or specific crime patterns. In contrast, outlying areas like Santa Clarita, Simi Valley, and Thousand Oaks exhibit significantly fewer arrests, possibly reflecting lower crime rates or different policing strategies in these suburban and rural regions. Meanwhile, areas east of downtown LA, such as West Covina and Ontario, show a more dispersed pattern of arrests, indicating a different crime distribution compared to the central and coastal areas.



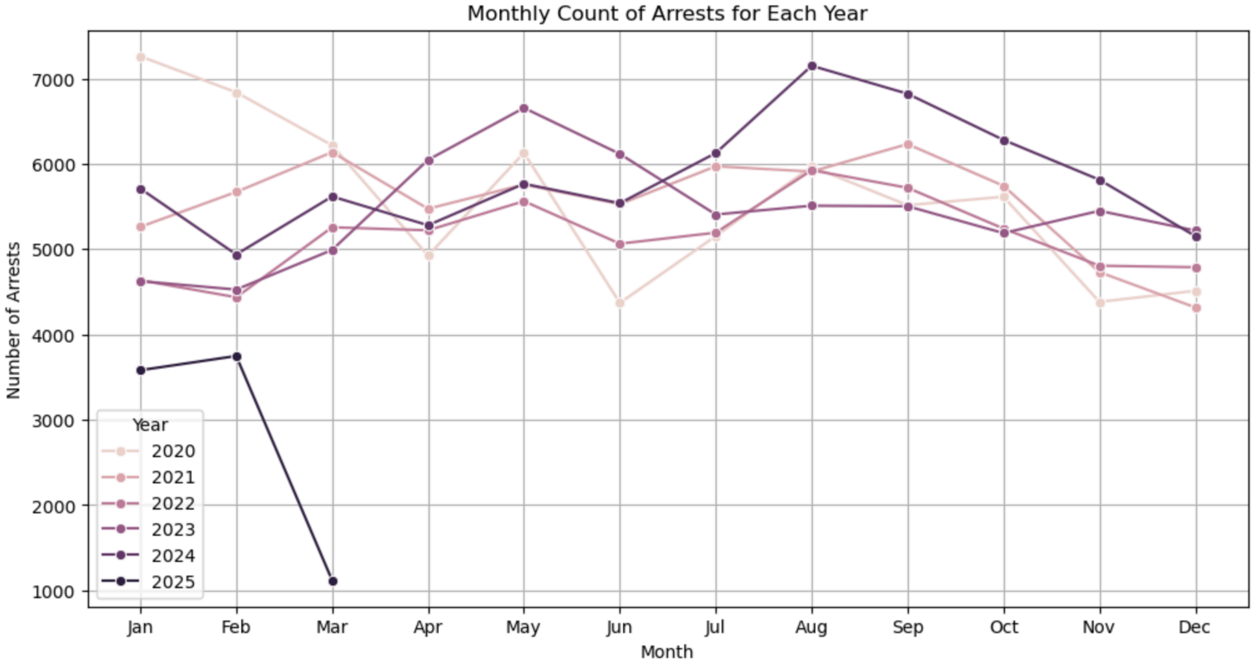
An upward trend in arrests over time may indicate stricter law enforcement, population growth, or rising crime rates. Conversely, a downward trend could suggest improved crime prevention strategies, policy changes, or declining crime rates. If the data shows sharp fluctuations in certain years, external factors such as new laws, social unrest, or changes in police policy may be influencing arrest rates.



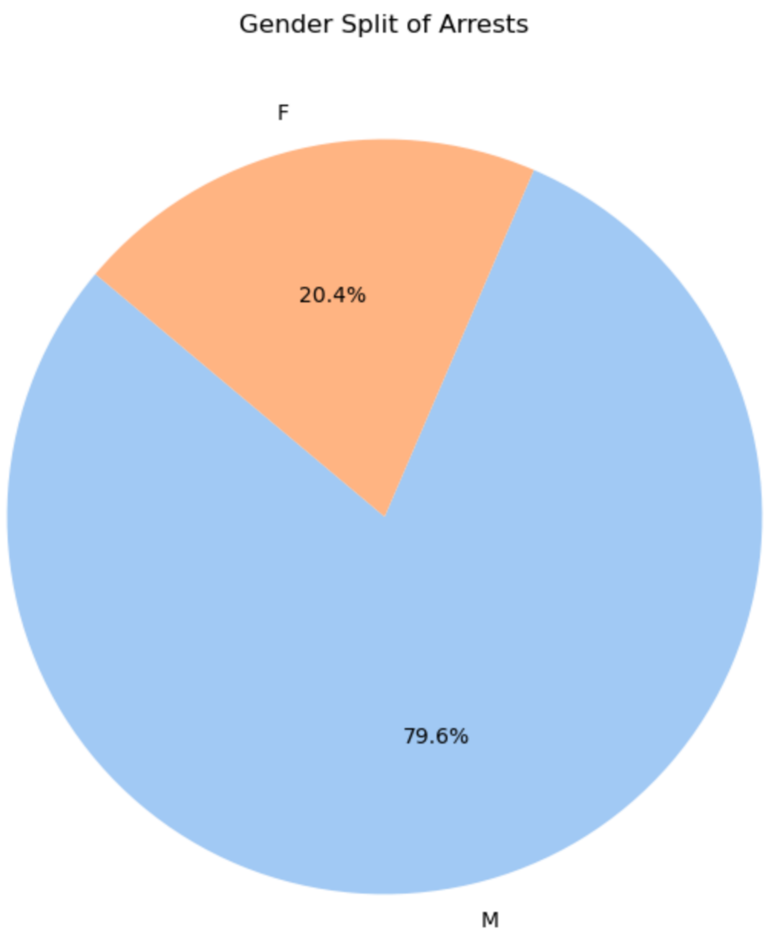
Some descent groups may have significantly higher arrest counts than others, indicating a potential overrepresentation in the dataset. Conversely, groups with very few recorded arrests might be underrepresented or have less recorded data. These variations can highlight disparities and trends in law enforcement interactions with different communities.



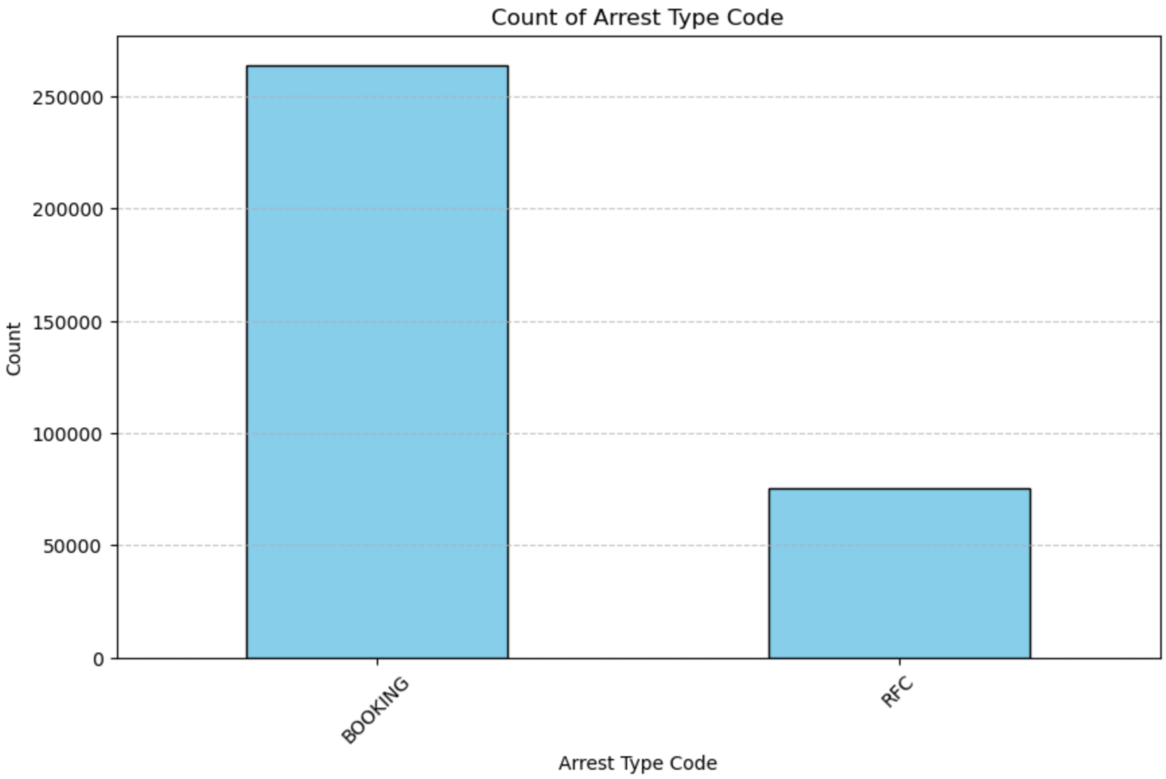
This visualization provides a clearer comparison of arrests per year. Unlike a line plot, which emphasizes overall trends, the bar chart distinctly highlights year-to-year variations. Years with exceptionally high or low arrest numbers stand out more prominently, making it easier to identify irregularities or emerging trends.



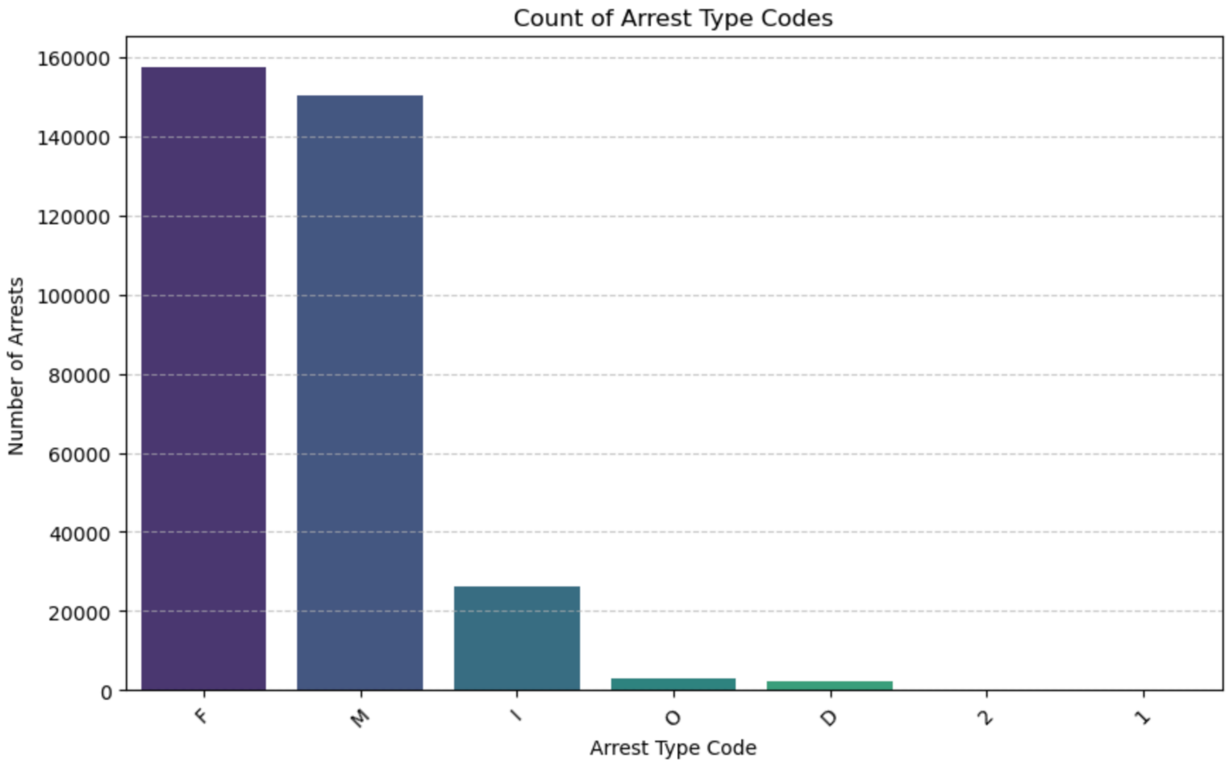
This visualization is a **line plot** that illustrates the **monthly count of arrests for each year**. The x-axis represents the months from January to December, while the y-axis shows the number of arrests. Different years are color-coded to highlight yearly trends, making it easier to compare fluctuations in arrests across different periods. This visualization helps in identifying seasonal patterns, such as increases or decreases in arrests during certain months.



This visualization is a **pie chart** that represents the **gender distribution of arrests**. The chart is divided into segments corresponding to different genders, with each section labeled with a percentage. This visualization provides an easy-to-understand breakdown of how arrests are distributed between male and female individuals, offering insights into demographic patterns in the dataset.

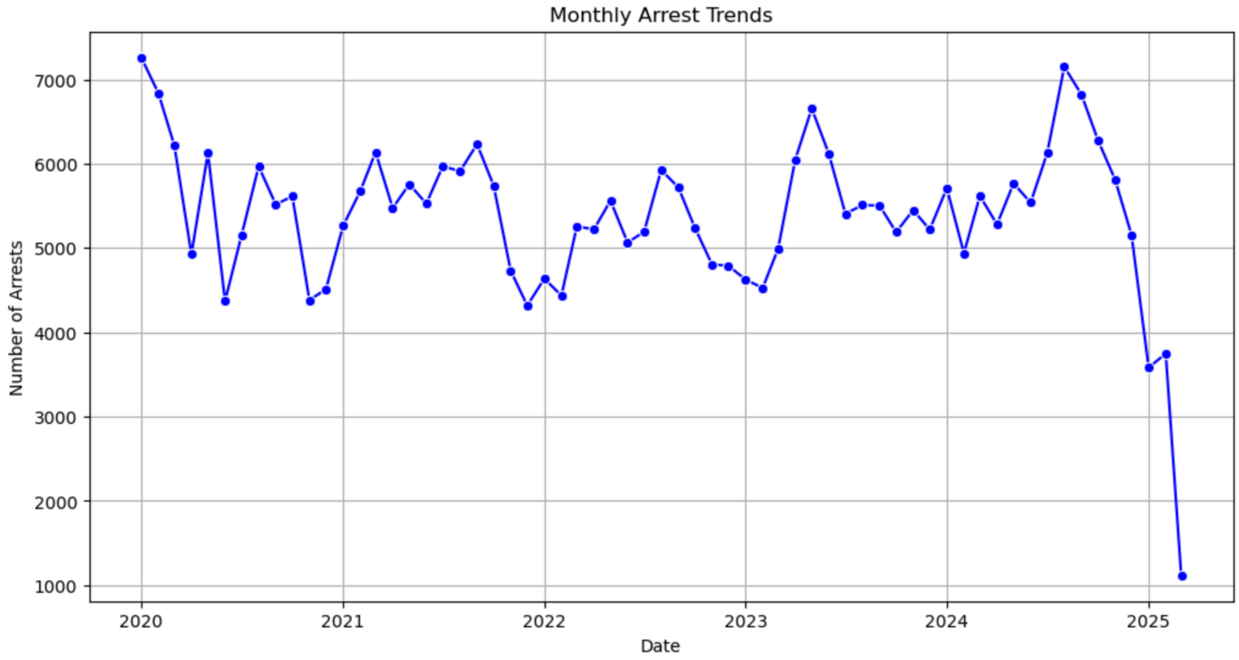


The third visualization is a **bar chart** displaying the **count of arrest type codes**. Each bar represents a specific type of arrest, with its height corresponding to the number of occurrences. The chart uses distinct colors to differentiate arrest types and is structured to provide a clear comparison of the most and least common arrest categories. This helps in understanding which types of arrests are most prevalent

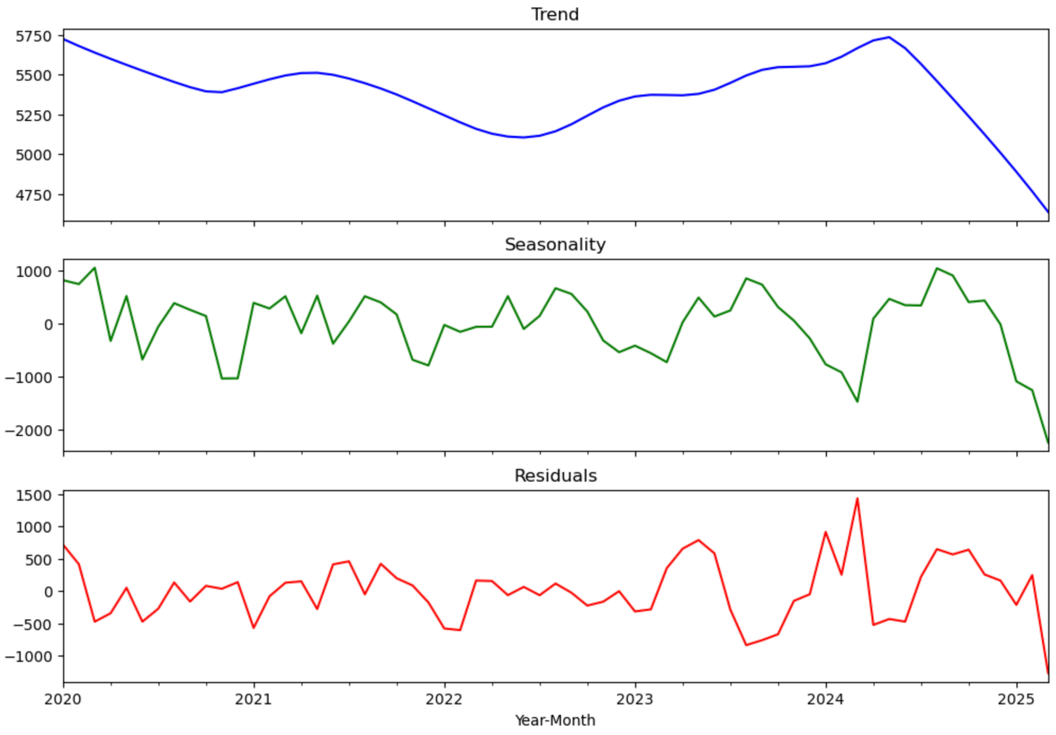


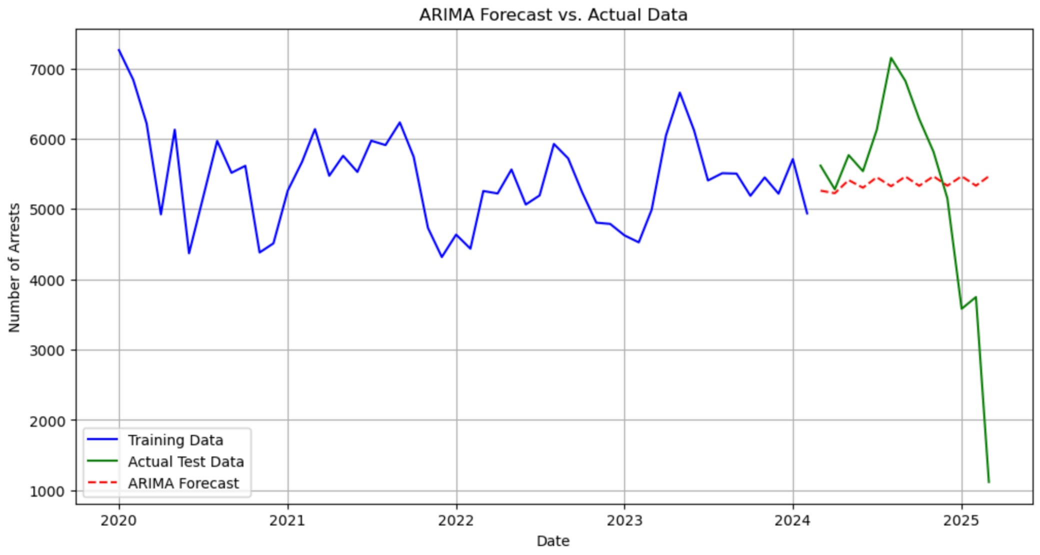
This visualization is another **line plot**, which shows **monthly arrest trends over time**. The x-axis represents the date in a "Year-Month" format, while the y-axis indicates the total number of arrests. Data points are connected with a line to illustrate fluctuations in arrests over time. This visualization is useful for analyzing long-term trends and identifying periods of increase or decrease in arrest activities

Advanced Analysis



This advanced visualization is a **Seasonal-Trend Decomposition (STL) plot** for time-series analysis. This plot breaks down the monthly arrest data into three components: **trend, seasonality, and residuals**. The **trend component** highlights long-term increases or decreases in arrests, while the **seasonality component** captures recurring patterns across months. The **residuals component** represents random fluctuations that are not explained by trend or seasonality. This analysis helps in understanding underlying patterns in the arrest data.


Trend
5750
5500
5250
5000
4750
Seasonality
1000
0
−1000
−2000
Residuals
1500
1000
500
0
−500
−1000
2020
2021
2022
2023
2024
2025
Year-Month


ARIMA Forecast vs. Actual Data
Number of Arrests
7000
6000
5000
4000
3000
2000
1000
Training Data
Actual Test Data
ARIMA Forecast
2020
2021
2022
2023
2024
2025
Date

Next, there is an **ARIMA (AutoRegressive Integrated Moving Average) model visualization**, which is used for forecasting arrests based on past data. The plot displays three key lines: **training data (blue), actual test data (green), and the ARIMA forecast (red, dashed line)**. The training data represents the portion used to build the model, while the test data is used to evaluate its accuracy. The forecasted values provide insights into potential future trends in arrest numbers. This model is useful for predicting future patterns and making data-driven decisions.

These advanced analyses add depth to the dataset by identifying long-term trends, seasonal fluctuations, and making predictive insights based on historical data. Let me know if you'd like a more detailed explanation of any specific aspect!

Conclusion

This project successfully demonstrated the importance of data-driven analysis in law enforcement trends using the Seaborn library. Through the visual exploration of arrest records in Los Angeles, we identified critical insights related to crime patterns, demographics, and judicial outcomes.

Key takeaways from the study include:

1. **Temporal Trends:** Arrest numbers fluctuated over the years, which could be linked to external influences such as social movements, economic conditions, or policing policies.
2. **Demographic Patterns:** A large proportion of arrests involved young adults, particularly those in their 20s. Certain ethnic groups had disproportionately higher arrests, raising questions about social and systemic factors.
3. **Geographical Distribution:** Arrests were not randomly distributed but rather concentrated in specific neighborhoods, indicating high-crime areas that might require targeted interventions.
4. **Judicial Processing:** Different types of disposition descriptions revealed trends in how cases are handled post-arrest.

By leveraging data visualization, this project provided a deeper understanding of arrest patterns, helping policymakers and researchers assess law enforcement strategies. Future studies could incorporate predictive analytics to forecast arrest trends and identify factors contributing to crime hotspots.