PM566 Report: Los Angeles Crime

Misha Khan

2022-12-08

Introduction

For my final project, I selected Los Angeles crime dataset starting from 2020. Because Los Angeles is known for having higher crime rates than most populated cities, I thought it would be interesting to examine local areas with the highest crime, most common crime, and victim demographics. The original dataset is a .csv file that comes from Los Angeles Open Data website. The dimensions are 586,295 observations and 28 variables that describe the date, time, crime, weapon, victim, and location.

Data source: https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8

Data Cleaning

Removing and Reformatting Columns

First, I renamed the columns and dropped 10 columns that were not relevant for analysis. Then, I reformatted date and time variables to extract month and year and to use them as variables. Since the current year is not over, I only removed 2022 for certain time related EDA/analysis.

Examining Coordinates

For coordinates, there were 2,266 coordinates that were (0,0) due to privacy reasons. I dropped these rows since the dataset is already large and it would not affect analysis. I kept these coordinates for my leaflet plots.

Empty Cells and Negative Values

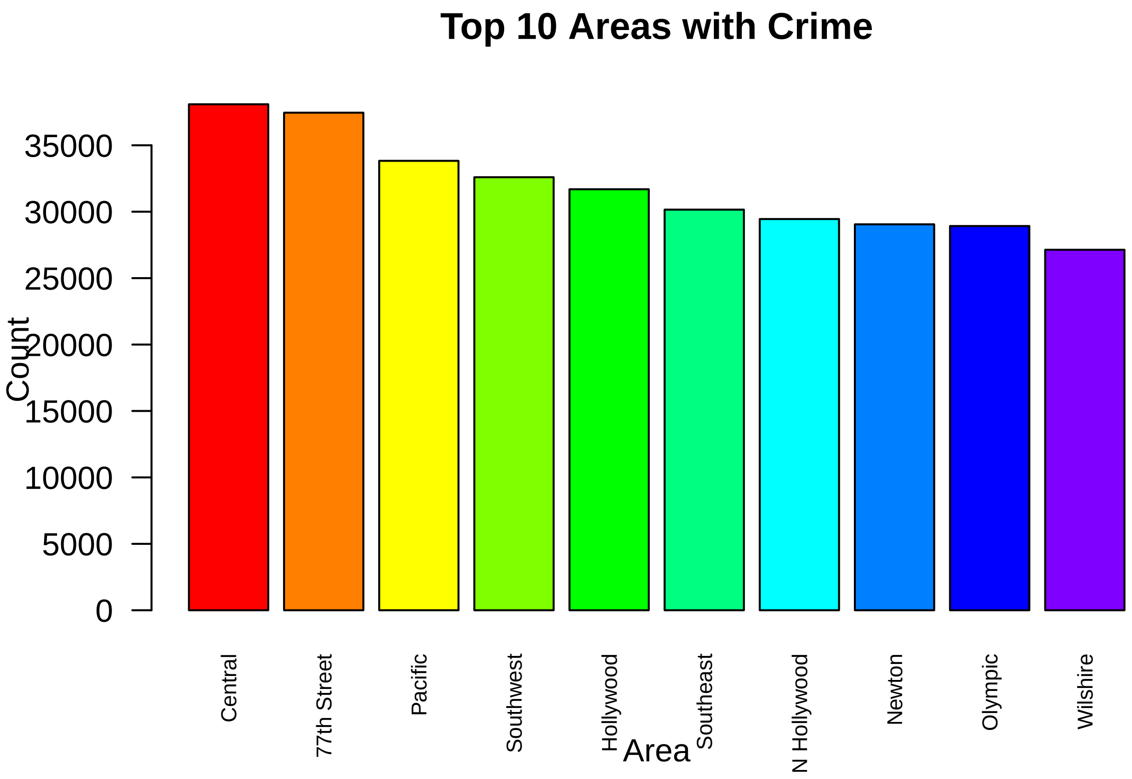
Victim Age had negative values, 0's (meaning not available), and an oddly high value of 120. For simplicity, I filtered age from 0-100 to exclude negative values. This step discarded 25 rows. The variables Weapon, Victim Sex, and Victim Ethnicity had thousands of empty cells. For Victim Sex and Ethnicity, I replaced it with NA instead of dropping it. I assumed that the empty cells meant that the information was not given. As for weapon, I assigned the empty cells to "NONE" meaning that no weapon was used.

After Cleaning

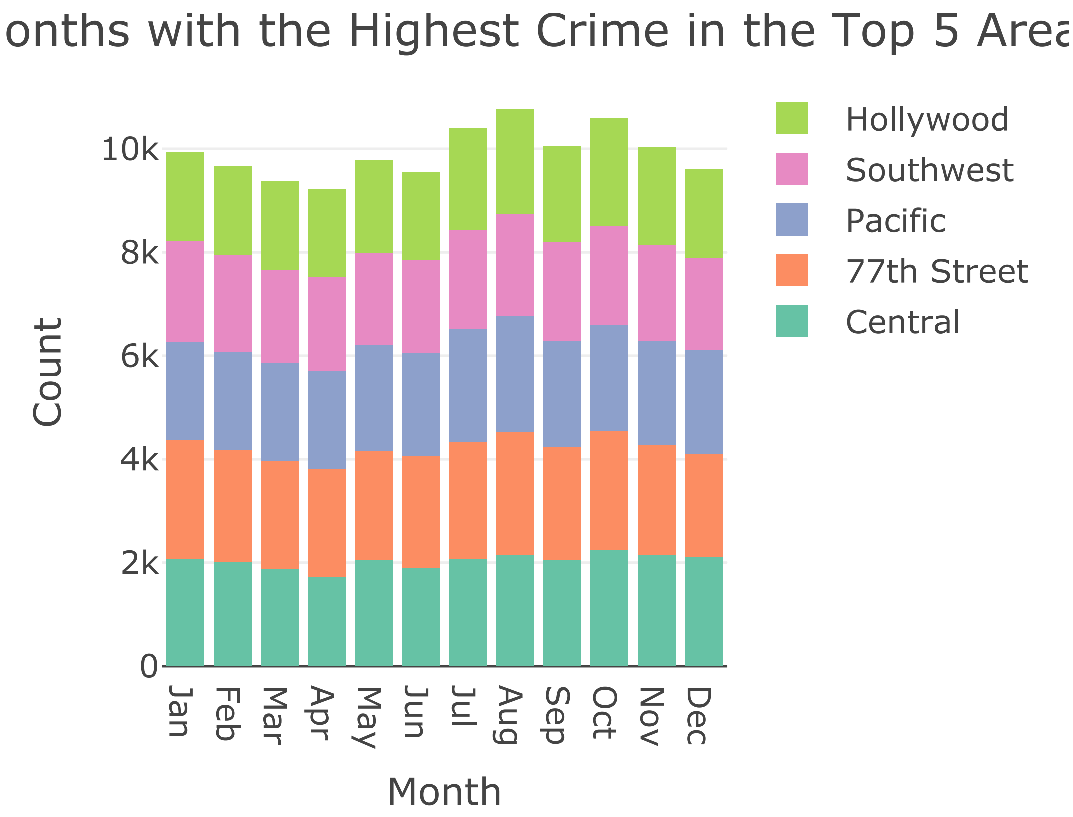
After data cleaning, the process removed 2,291 rows. The crimedat dataset now has 584,004 observations and 21 variables. For further analysis, I subsetted the data to top 5 areas with the highest crime because of how large the dataset is.

EDA

Where and when does the most crime occur?



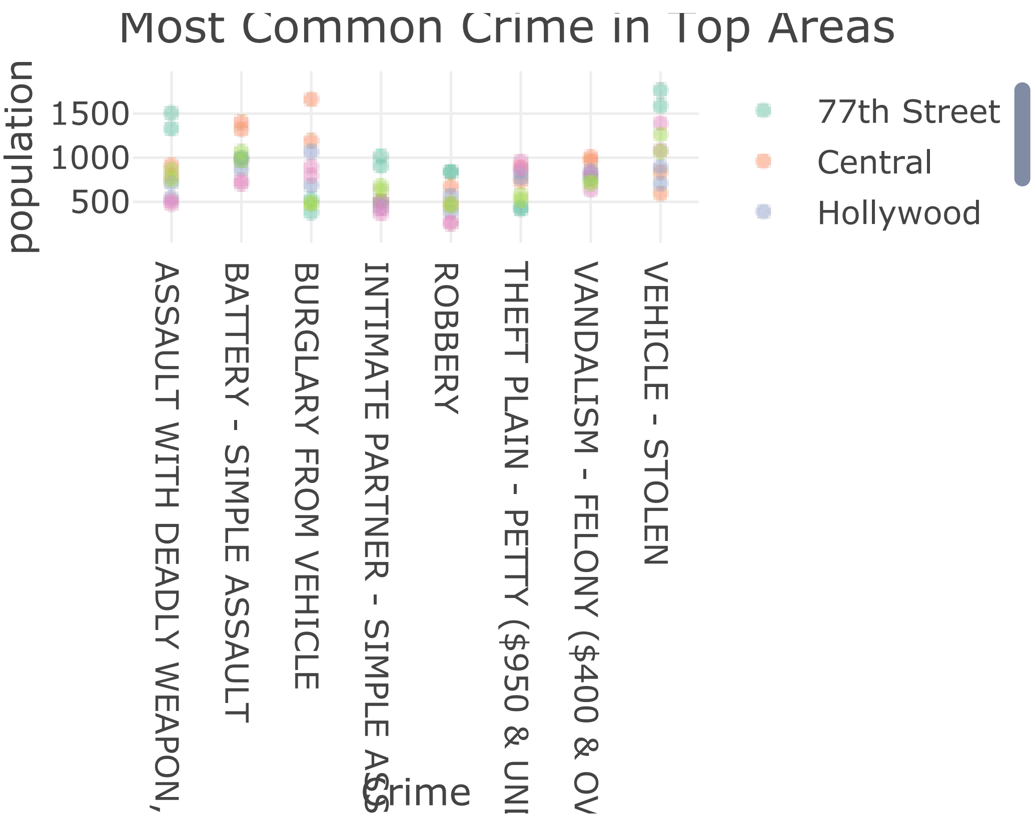
Bar Chart



Like mentioned before, 2022 is excluded from this analysis since the year is not over. This bar chart displays the proportion of crime in the top 10 areas in Los Angeles. Central and 77th Street are ranked significantly higher than the rest. As for when does crime occur the most, the EDA above contains two outputs per area for the total number of reports in 2020 and 2021. Focusing on the top 5 areas with the highest crime, I saw that January to June there are a little under 10,000 reports. However, in the warmer months, the reports increase over 10,000 and then drop again in November and December.

What are the most common crimes?

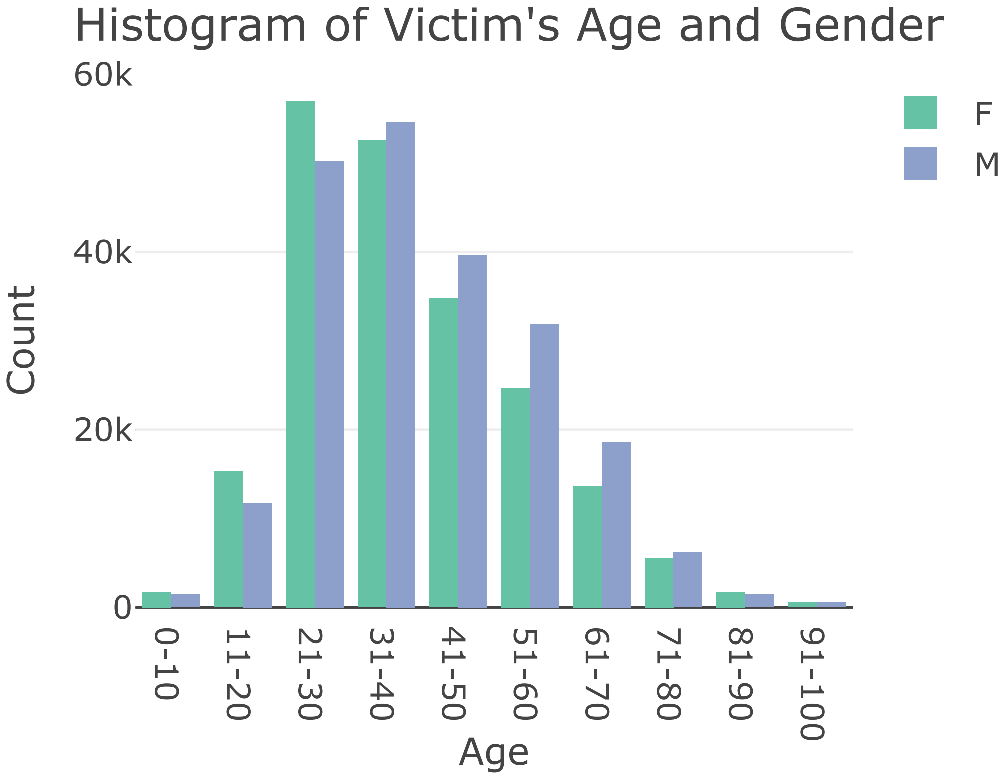
Scatter Plot



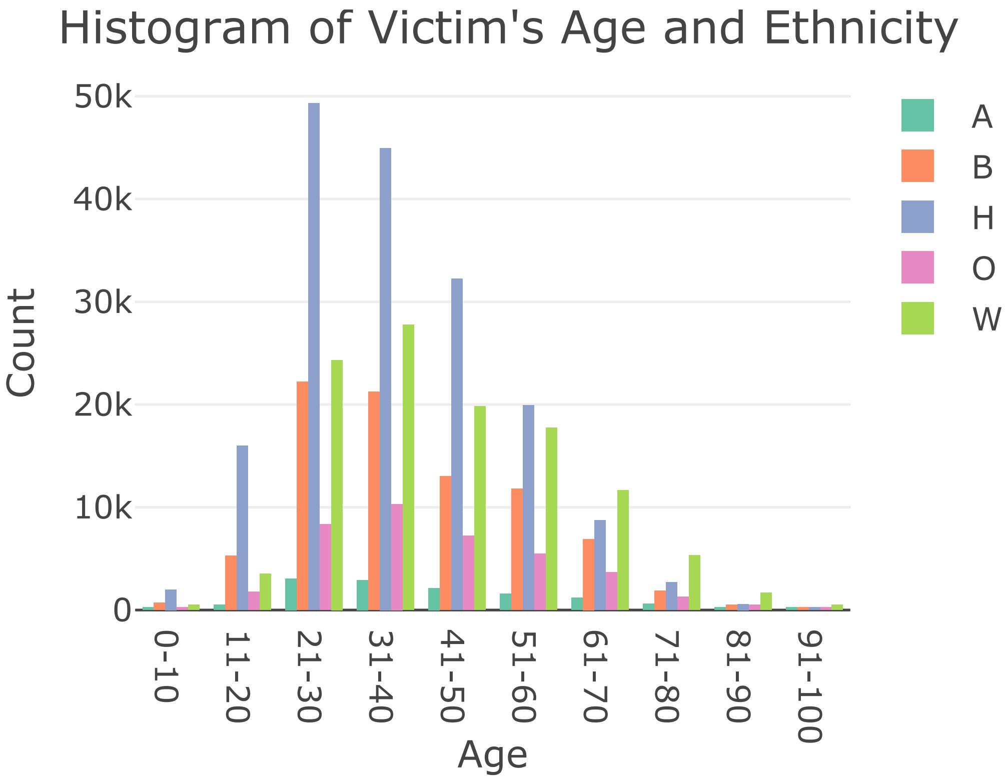
There are two dots per area for each year (2020 and 2021). Across the plot, the highest points (blue and orange) are from the areas 77th Street and Central. There are over 1,500 reports of burglary from vehicle and vehicle stolen.

What is the demographic breakdown of victims?

Victim Sex



Victim Ethnicity



As for victim demographics, younger women, older men, and Hispanics report more crime.

Conclusion

The areas with the highest crime reports are 77th Street and Central.For when the does the most crime occur, there are a higher number of crime reports in the warmer months than colder months. The highest number of reports is in August while the lowest number of reports is in Aprils. This would be interesting to look into as for what possible factors could be possibly contributing to that. Correlation is not causation but it is an pattern to note.

The areas with the highest crime rate in Los Angeles are 77th Street, Central, Hollywood, Pacific, and Southwest. In these areas, theft from vehicle and vehicle theft are the most common crimes. The highest rates of crime occur in 77th Street and Central areas. In 2021, 77th Street has 1,765 reports of vehicle stolen and Central has 1,660 reports of theft from vehicle. This can be useful to keep in mind if one is in the area.

From the histogram of victim's age, there is a noticeable spike in women in their 20s reporting crimes compared to men. However after 30s, there are more men reporting crime than women. This is an interesting observation that they are a trend of younger women and older men reporting crimes.

When examining victim's ethnicity, there are high number of reports by Hispanic residents in their 20s through 40s. This is probably due to Los Angeles having a mostly Hispanic population. The second highest number of victims is White people then Black people. With further analysis, it would be compelling to see the breakdown of the type of crime being reported by each demographic.