Final Project

Motivation

Since the pandemic began, crime has increased dramatically across the United States. According to the New York Times, the murder rate increased 25% alone in 2020. This means the United States had surpassed 20,000 murders in a year, a level not seen since 1995. More specifically, the murders per 100,000 people rate was 6.22 in 2020, with even higher rates in large metropolitan cities. Looking at Los Angeles, the homicide rate is up 24.2% compared to last year and 37.8% compared to 2019 according to statistics published in CompStat from Chief of the Los Angeles Police Department, Michael Moore.

In this project, I chose to analyze the homicide/crime problem in Los Angeles County because violent crime is an extremely important issue that affects everyone. In addition, crime has multifaceted causes that must be analyzed deeper to understand why it occurs and where it occurs most. Better statistics and data collection help police departments in both response times and allocating appropriate resources to high crime areas.

Brief of Data Sources

1. First Dataset: https://homicide.latimes.com/

This is the homicide page on the Los Angeles Times. This site is a collection of homicide victims in Los Angeles County from present to last year. Below is a snippet:
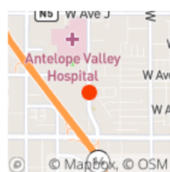
From this dataset, I scraped the victims' name, age, neighborhood and address of homicide, and date of death using Beautiful Soup and Selenium for scrolling the page. Most of 2021 crimes will be scraped.

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

Second, is the Los Angeles public crime data from 2020 to Present. This data is transcribed from the Los Angeles Police Department to a variety of formats and accessible by API.

The dataset contains the following info: "This data is transcribed from original crime reports that are typed on paper and therefore there may be some inaccuracies within the data. Some location fields with missing data are noted as (0°, 0°). Address fields are only provided to the nearest hundred block in order to maintain privacy."

Below is a snippet:

DR_NO	Date ...	DATE...	TIME ...	AREA	AREA...	Rpt D...	Part ...	Crm ...	Crm ...	Moco...	Vict ...	Vict
010304468	2020 Jan ...	2020 Jan ...	2230	03	Southwest	0377	2	624	BATTERY ...	0444 0913	36	F
190101086	2020 Jan ...	2020 Jan ...	0330	01	Central	0163	2	624	BATTERY ...	0416 182...	25	M
191501505	2020 Jan ...	2020 Jan ...	1730	15	N Hollyw...	1543	2	745	VANDALI...	0329 1402	76	F
191921269	2020 Jan ...	2020 Jan ...	0415	19	Mission	1998	2	740	VANDALI...	0329	31	X
200100501	2020 Jan ...	2020 Jan ...	0030	01	Central	0163	1	121	RAPE, FO...	0413 182...	25	F
200100502	2020 Jan ...	2020 Jan ...	1315	01	Central	0161	1	442	SHOPLIFT...	1402 200...	23	M
200100504	2020 Jan ...	2020 Jan ...	0040	01	Central	0155	2	946	OTHER M...	1402 0392	0	X
200100507	2020 Jan ...	2020 Jan ...	0200	01	Central	0101	1	341	THEFT-GR...	1822 034...	23	M
200100509	2020 Jan ...	2020 Jan ...	2200	01	Central	0192	1	330	BURGLAR...	1822 141...	29	M
200100510	2020 Jan ...	2020 Jan ...	0955	01	Central	0111	2	930	CRIMINAL...	0421 0906	35	M
200100514	2020 Jan ...	2020 Jan ...	1355	01	Central	0162	1	341	THEFT-GR...	1822 034...	41	M
200100515	2020 Jan ...	2020 Jan ...	1638	01	Central	0162	1	648	ARSON	1402 150...	0	X
200100520	2020 Jan ...	2020 Jan ...	1805	01	Central	0128	1	442	SHOPLIFT...	0325 140...	24	F
200100535	2020 Jan ...	2020 Jan ...	1330	01	Central	0152	1	210	ROBBERY	0416 041...	66	M

I scraped this using the Socrata API and inputted it into a SQL database. There are lots of important columns given with each crime such as "Crm Cd Desc", which provides the description of the crime and "Lat" and "Lon.", which provides information about the location of the crime.

3. Third Dataset: https://www.google.com/maps

I use Google Maps API and the following APIs: Geocoding API, Maps Javascript API, and Maps Static API. The LA Times homicide page does not give latitude and longitude. So, I input the address and neighborhood into the Geocoding API to reverse geocode and receive the latitude and longitude for each homicide. Then, I append lat and lon to the existing LA Times homicide dataset. I also create a separate SQL table for the reverse geocoded data: (lat, lon).
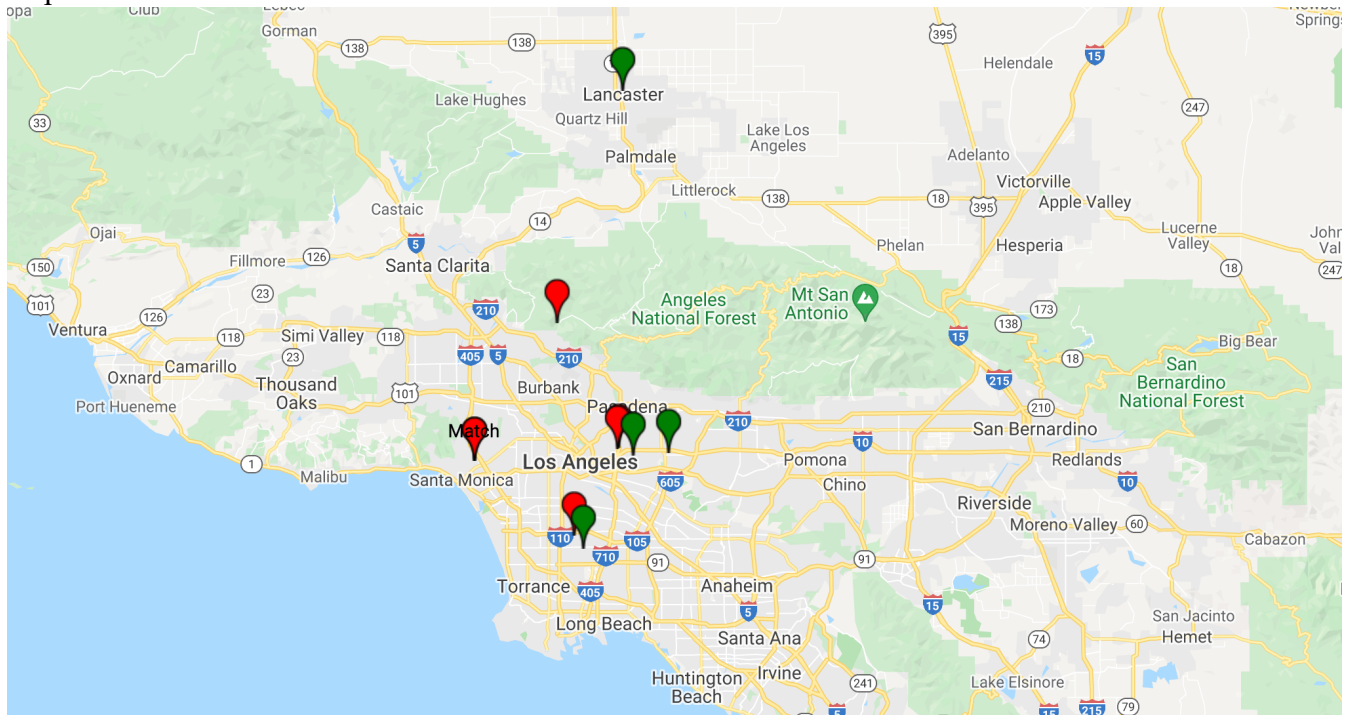
Analysis Performed

1. Google Maps Visualization

Using the Google Maps API for reverse geocoding, I use it (as mentioned above) to receive the latitude and longitude for each crime in the LA Times homicide page. If the LA Times homicide dataset lacks address or neighborhood information, the reverse geocoder is skipped for these rows.

Then, every single one of the homicides in the LA Times dataset is cross referenced against all of entries in the LA Public Crime database (250,000 entries) to see if there is a match based on the distance between the two crimes. A match is the distance between the two crimes where it is less than 0.4. If there is a match, it is plotted on Google Maps with the crime number (dr_no), so it can be referenced in the LA Public Crime database. If there is no match, the nearest crime is plotted with its crime number(dr_no) from the LA Public Crime database in addition to the LA Times homicide.

For simplicity and readability, I plotted the first 5 homicides from the LA Times. There is only 1 match from the first 5 homicides. See below for the visualization:
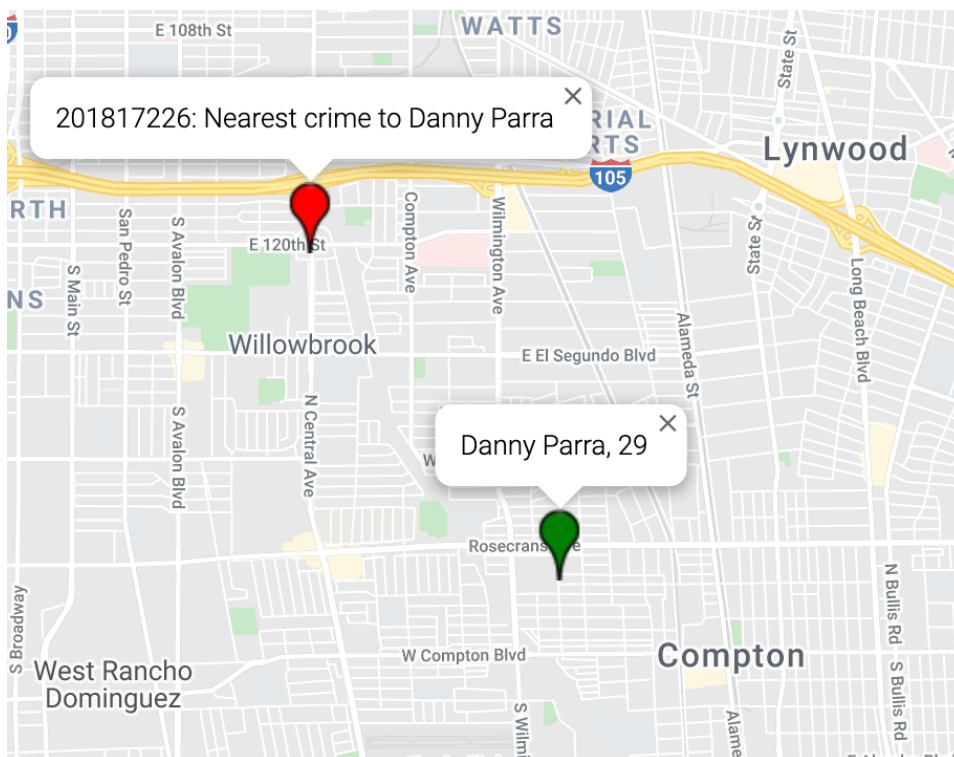
map1.html:

Match:
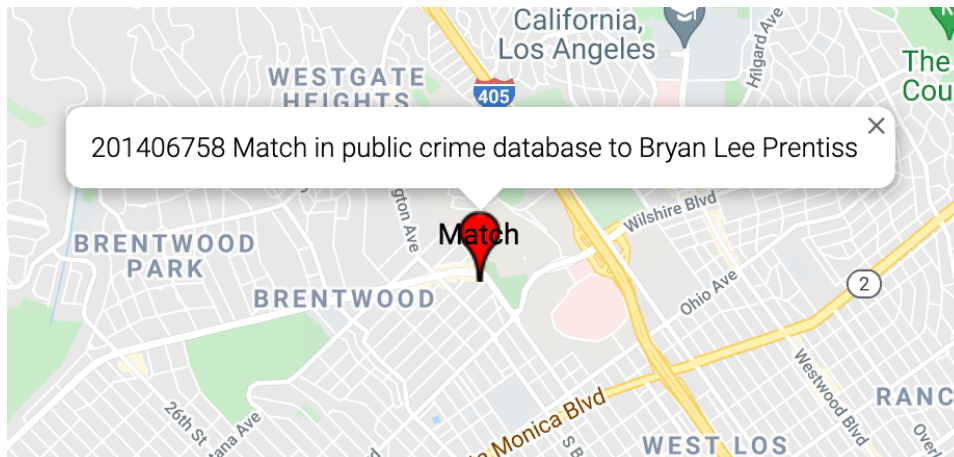
Nearest crime:

2. Statistical Analysis

The other analysis performed is the following: For all crimes with distance>0.4, the distance to the nearest crime is documented in a list as well as the average distance of all crimes. In addition, it checks whether the nearest crime is also a homicide. If not, it checks whether it is a violent crime and what specific type of violent crime it is. Based on crime databases, rape, robbery, assault, and any crimes involving force and weaponry are considered violent crimes. Thus, the crimes are checked against these types based on the description. Lastly, the distance is outputted on the terminal for all crimes that are not matched and have a crime nearby.

Conclusion

Out of the 254 entries scraped from the LA Times homicide page, 95 are a match About 40% of crimes are matched: (95/254). This means the statement: "there may be some inaccuracies within the data" from LAPD public crime data is accurate. I expected more crimes to be a match since they are homicides and generally homicide data provides an accurate location, even if they are transcribed incorrectly.

The distance from the nearest crime also varies. The average distance of the nearest crime is: 234.76311660110693. I also use a trimmed mean which takes out 10% of the outliers from both the leftmost and rightmost scores. Accounting for outliers, the average distance of the nearest crime is: 5.528804102094232. This means, on average, crimes are happening in the same neighborhoods as the LA Times homicides. A distance of 5.52 indicates a bit more than that of Danny Parra above, whose distance is 3.10. This proves crimes are happening within a few streets of each other. Out of the 159 crimes that are not a match, 51 are violent crimes but none are homicides. This means approximately 30% of the nearest crimes to homicides are other violent crimes.

The analysis may vary in scraper.py as the number of data points from scraping is subject to how many scrolls Selenium performs.

Visualizations/Analysis Results

Sample of analysis section from --static:

```
ANALYSIS SECTION:

The number of LA Times homicides that match with the LAPublicCrime database are: 95 out of 254

Number of homicides near other homicides: 0

Number of shootings near other homicides: 16

Number of assaults near other homicides: 20

Number of robberies near other homicides: 5

Number of crimes involving threat of weapons near other homicides: 10

Number of violent crimes near other 159 homicides from LA Times: 51

32.08% of nearest crimes to homicides are other violent crimes

First visualization based on analysis titled: Type of crimes nearest to homicides

Mean distance is: 234.76311660110693
Trimmed mean distance, accounting for 10% outliers is : 5.528804102094232

Displaying distance for 2 nearest crimes of non-matched crimes for: 141 crimes
Distance between 2 nearest crimes of 200412573 and Zhaiming Li is 3.2295437644558813
Distance between 2 nearest crimes of 200401155 and Miguel Meza Quintero is 9.888105921985659
Distance between 2 nearest crimes of 201817226 and Danny Parra is 3.1018299442679282
```

Matplotlib visualization for types of violent crimes near homicides from --static