Judging a book by its cover:
Analyzing the relation between a disordered environment and crime in Los Angeles.

1. PROJECT DESCRIPTION

This project analyzes whether more crimes are committed in disordered areas than in ordered ones, using MyLA311 reports and Los Angeles Police Department crime data from 2023. MyLA311 is a platform for reporting issues like graffiti, illegal dumping, and homeless encampments. Crime data includes location details for incidents in 2023.

2. QUESTION

Did reported signs of disorder, such as graffiti or littering, correlate with crime in Los Angeles in 2023?

3. DATA SOURCES

The project uses data from MyLA311 and the Los Angeles Police Department, as well as additional data about street names and ZIP code regions.

All datasets are made available under the public domain license CC0. Hence, they can be copied, modified and distributed for any purpose [1] and can be used without restrictions in this project. An overview is given in Table 1.

Name	Source	License	Publisher
MyLA311 Service Request Data	[2]	CC0	Information Technology Agency of the City of Los Angeles
Crime Data	[3]	CC0	L.A. Police Department
Street Names	[4]	CC0	Bureau of Engineering of L.A.
ZIP Code Regions	[5]	CC0	City of Los Angeles

Table 1 Data Sets

3.1 MyLA311 Service Request Data

To measure the degree of disorder, reports from MyLa311 from 2023 are used. The data set was chosen because LA311 is the main platform to report signs of disorder in Los Angeles. Moreover, rows contain location information, which is of great importance for this project.

Important columns such as CREATEDDATE, REQUESTTYPE, ZIPCODE and LONGITUDE/LATITUDE are of good quality, as each of them is valid and filled in more than 99.8 % of the rows. In addition, the format of dates is consistent throughout the dataset. Moreover, GPS coordinates and ZIP codes are plausible as they are within the region of L.A., indicating a good accuracy. However, only 82.49 % of the rows have a valid and filled STREETNAME.

3.2 Crime Data

The Crime Data includes incidents of crime, including location data, in Los Angeles since 2020. The data set was chosen as it provides a comprehensive list of crimes committed in the relevant time range of 2023

The format of dates is consistent throughout the dataset. Moreover, important columns, such as DATE OCC (Date when the crime occurred) and CRM CD DESC (description of the crime) are valid and filled for 100% of the rows. Furthermore, 99.79 % of LAT/LON coordinates are plausible as they are within the region of L.A., indicating good accuracy.

Nevertheless, the LOCATION column, which contains the address where the crime happened, is not consistent, as some entries have additional whitespace or are missing important information such as the ZIP Code.

3.3 Street Names

To enhance the Crime data, missing street suffixes are looked up using an L.A. street name dataset, focusing on the Street Name and Suffix columns. This dataset was chosen for its overall quality and the publisher's reputation.

Street names are 100% complete, while 3.36% of suffixes are missing. However, the available suffix data is consistent and accurate based on random samples.

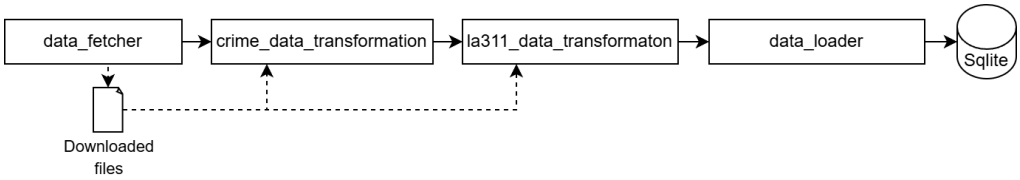
3.4 ZIP Code Regions

To map GPS coordinates to ZIP code regions, a shapefile with ZIP Codes and areas for L.A. is used. The dataset was preferred over the actively maintained version at the L.A. City data hub, as an automated downloading of the files from the portal via Selenium does not work reliably.

The visual representation of the file indicates complete and accurate data. The ZIP Codes are all consistent, following the expected structure of ZIP Codes for L.A.

4. DATA PIPELINE

The data pipeline is implemented using Python and the libraries Selenium, pandas, geopandas, numpy and retry. It consists of the DataFetcher, the CrimeDataTransformation and the La311DataTransformation. Each of the Components is located in its own class.



4.1 DataFetcher

At first, an instance of the DataFetcher retrieves all required files. To download the files from the L.A. data catalog, selenium is used, while files from kaggle.com are downloaded using the standard library urllib. There are up to three retries before the program fails with an exception. Retries can be triggered when the predefined timeout is exceeded or an issue with the network or Selenium occurrs. If all retries fail, the program exits with an exception and an error message, as the data is required for the next steps.

4.2 CrimeDataTransformation

After retrieving the files, an instance of the CrimeDataTransformation filters rows by selecting only those from 2023. To do so, the DATE OCC column is transformed to a date. Rows with invalid dates are discarded, reported as invalid and a warning is raised.

Furthermore, the LOCATION column is split into the columns STREET NAME and SUFFIX. As the suffix is missing for some street names, missing suffixes are looked up using a list of L.A. street names. However, this process only works for unambiguous street names. Street names with multiple possible suffix matches are not enriched. Since Street Names are not the prime target of this project's analysis, missing values are reported but no rows discarded.

A main issue is the fact that ZIP codes are often missing in the LOCATION column, despite being important for further analysis steps. Therefore, the most important data enrichment step is the assignment of ZIP Codes based on given position data. To do so, the shapefile that contains the ZIP code regions is used and combined with the crime data using a spatial join, which is provided by geopandas. Missing ZIP codes are reported as invalid and an error is thrown if the number of invalid rows exceeds a predefined threshold of 10 %.

In addition, column names are renamed for consistency, additional whitespaces removed and invalid values for the column VICT SEX (sex of the victims) are changed to X ("unknown").

4.3 La311DataTransformation

After fetching the datasets, the LA311 dataset is processed through the La311DataTransformation. The data is filtered by request type, retaining only the relevant types: "Single Streetlight Issue," "Graffiti Removal," "Illegal Dumping Pickup," and "Homeless Encampment."

Afterwards, the column ADDRESSVERIFIED is cleaned up by setting invalid inputs to N ("not Verified"). Due to the importance of ZIP codes, all rows are checked to have a filled column for ZIP codes. Rows with missing ZIP codes are reported as invalid and a warning is raised if the number exceeds a threshold of 10 %.

4.4 Loading and Meta-Quality measures

After running each of the Transformations a short report is printed that shows the percentage of invalid rows that were found during the process. If the threshold for invalid rows is exceeded, a warning is raised.

The transformed data is loaded into a SQLite database. This allows an easier retrieval of the data in the next steps and makes it easier to check the data manually, as opening the data in Excel or a text editor takes some time due to the amount of data.

4. RESULTS AND LIMITATIONS

The resulting data contains only rows of the relevant year 2023 for both data sets. The completeness of the crime data set was greatly improved by adding related ZIP Codes to all crime incidents, which is needed for further analyzing and comparing both datasets area-wise. GPS coordinates appear to be generally accurate and complete for both datasets.

The consistency is improved by removing whitespace and adjusting invalid values. Furthermore, the LOCATION column of the crime data set is split in Street Name and Suffix, as it is in the LA311 dataset. Even though the street suffixes were partly added using lookups, not all of them could be identified.

As of the September 27 2024, about 0.029 % of the rows in crime data and 0.27% of the rows in LA311 were reported as invalid during the transformations described in Chapter 3.

Limitations arise from the bad data quality of the LOCATION column in the crime data set and missing Street Names in the LA311 data set. Additionally, citizens in "bad neighborhoods" might hesitate to report issues at LA311 due to less trust in public agencies, even though it can be reported anonymously via various communication channels.

REFERENCES

[1] Creative Commons, "CC0 1.0 Universal," [Online]. Available: https://creativecommons.org/publicdomain/zero/1.0/legalcode.en.

[2] City of Los Angeles, "MyLA311 Service Request Data," [Online]. Available: https://data.lacity.org/City-Infrastructure-Service-Requests/MyLA311-Service-Request-Data-2023/4a4x-mna2/about_data.

[3] City of Los Angeles, "Crime Data from 2020 to Present," [Online]. Available: https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8/about_data.

[4] City of Los Angeles, "Street Names," [Online]. Available: https://data.lacity.org/City-Infrastructure-Service-Requests/Street-Names/hntu-mwxc/about_data.

[5] Kaggle, "Los Angeles County Shapefiles," [Online]. Available: https://www.kaggle.com/datasets/cityofLA/los-angeles-county-shapefiles/data.