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

Eric Walsh

I. Introduction

WHEN the Broken Window Theory was published in 1982, the relationship between a disordered environment and crime became subject to debate in the public sphere. Since then, various case studies have been conducted which analyze how restoring order in a certain area affects the number of crimes.

The goal of this project is to investigate whether more crimes are committed in disordered areas than in ordered ones. It therefore aims to answer the following question:

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

The results of this analysis could provide insights into whether the common perception that more crimes occur in bad-looking areas holds true. This information could be valuable for both individuals and businesses, indicating whether avoiding disordered environments might mean avoiding crime hotspots.

II. Used Data

All data used is provided under the Creative Commons Zero License and can therefore be copied, modified and distributed for any purpose in this project.

To measure the degree of disorder, reports from MyLA311 in 2023 are used. MyLA311 is a platform provided by the City of Los Angeles that allows citizens to report issues such as graffiti, illegal dumping, and homeless encampments. To conduct the analysis, the columns described in Table I are used after being preprocessed by the data pipeline. Only rows deemed valid by the ETL pipeline are used, resulting in 510,477 rows (99.73% of the original dataset) being considered. The RequestType column contains the type of reported disorder. The signs of disorder in this project are graffiti, waste dumping, homeless encampments, and broken streetlights.

Information about reported incidents of crime in L.A. is provided by the Los Angeles Police Department and includes location data. Only crimes that occurred in 2023 are considered. The columns that are used after the preprocessing of the data are described in Table II. About 99.8% of the original data is considered valid and used for this project, which equals 214,564 rows. The column 'Crm Cd' contains the numeric value that identifies the type of crime with a numeric code, whereas 'Crm Cd Desc' contains the textual description. For example, the 'Crm Cd' 210 corresponds to the textual description 'Robbery'.

TABLE I
Description of MyLA311 Data Columns

Column Name	Description
`Zipcode`	The ZIP code of the region where the disorder was reported.
`RequestType`	The type of disorder reported, such as illegal dumping or graffiti.

TABLE II
Description of Crime Data Columns

Column Name	Description
`Crm Cd`	The code representing the type of crime.
`Crm Cd Desc`	The textual description corresponding to the crime code.
`Zipcode`	The ZIP code of the region where the crime was reported.

TABLE III
Description of Population Data Columns

Column Name	Description
`Zipcode`	The ZIP code of the region where the disorder was reported.
`Population`	The number of people reported to live in the area.

In addition, information about the most recent population data is used, as described in Table III. It stems from the 2020 Census and all relevant rows and columns in the original data set are considered valid.

III. Analysis

For the analysis, the data about crimes and disorders are grouped by ZIP code, categorized and standardized. Moreover, the general data structure and the correlation between disorder rate and crime rate are inspected.

A. Preparation

The crime data is split into five disjoint categories depending on the type of the crime as described in Table IV. Disorder-related crimes are excluded from further analysis, as reported signs of disorder usually also lead to a report to the police. The category 'other' contains all crimes that do not fit in any of

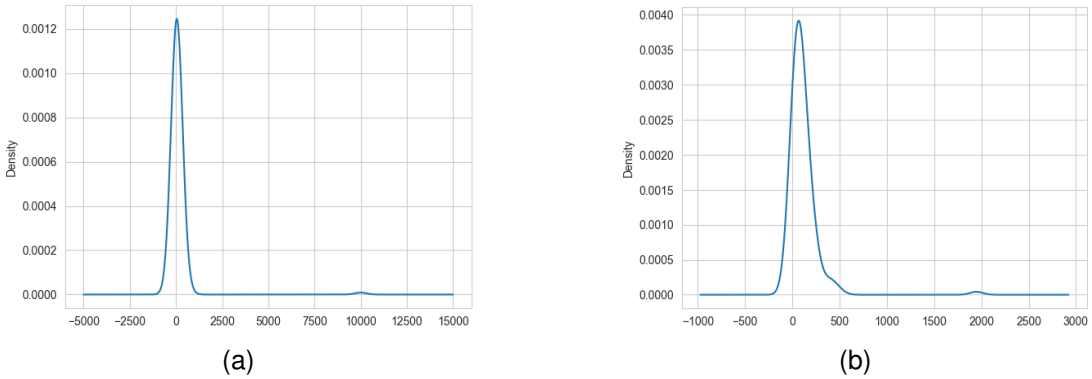

Fig. 1. Density Estimate Plots of (a) Total Crime Rate and (b) Disorder Rate

the categories, such as bribery. The categorization of the rows is based on the crime code, which follows the UCR Offense Classifications [1].

TABLE IV
DESCRIPTION OF THE CRIME CATEGORIES

Crime Category	Description
`Violent`	Victims are threatened or harmed with violence.
`Property`	Involves private property but no violence.
`Fraud`	Fraud such as Identity theft.
`Disorder-related (Excluded)`	Related to disorder, such as graffiti. Excluded from analysis.
`Other`	Crimes that do not match the other categories.

The crime data and the MyLA311 data are grouped by ZIP code and standardized using population data to account for differences in population size. The result of the standardization is the 'rate', which gives the number of occurrences per 1000 inhabitants of the area.

B. General Data Structure

The density estimation of the total crime rate as well as the total disorder rate roughly follows a normal distribution with a slight right skew. The graph, which is shown in Figure 1, also shows outliers with very high rates. These are caused by regions with a very low population. For instance, the area with ZIP code 90095 is primarily associated with the University of Los Angeles and reports an anomalous rate of 10,000 crimes per 1000 people living in that area.

The scatter plot shown in Figure 2, which compares the total crime rate and the total disorder rate, suggests a possible positive correlation. It can also be perceived that outliers are often observed in areas with a low population.

C. Correlation

To identify correlations between the data, the Pearson correlation coefficient (PCC) and the Spearman's correlation coefficient are used. Both yield a value between -1 and 1. An absolute value of 1 indicates a strong correlation, and the sign indicates whether the correlation is positive or negative.

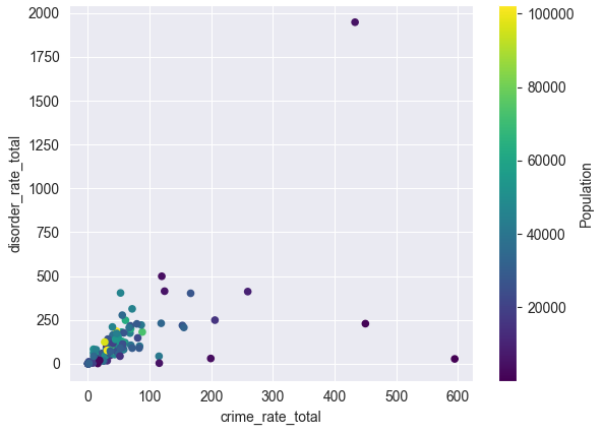

Fig. 2. Total disorder rate and total crime rate per area (truncated for outliers over 1000)

PCC measures the linear relationship between two variables and works well for normally distributed data. However, it's sensitive to outliers. Spearman's correlation coefficient measures the monotonic association and is robust against outliers.

The PCC between the total crime rate and total disorder rate is 0.54 and implies a moderate to strong correlation between these two factors. In addition, Spearman's Correlation indicates a high correlation, with a value of 0.77.

Due to Pearson's sensitivity to outliers, Spearman's correlation is used to conduct the analysis of the correlation between different categories of disorder and crimes. As can be seen in Figure 3, the coefficient of 0.05 indicates that there's no significant correlation between the population and the total crime rate. However, there's a weak correlation (0.3) between the total disorder rate and the population. Furthermore, there's a strong correlation (0.77) between the rates of total crime and total disorder. This holds true for all categories of crime except fraud, where only a weak correlation (0.13) exists. As fraud is often committed remotely via phone or internet, it seems reasonable that it is not directly affected by the disorder of the environment on site. The category of the observed disorders

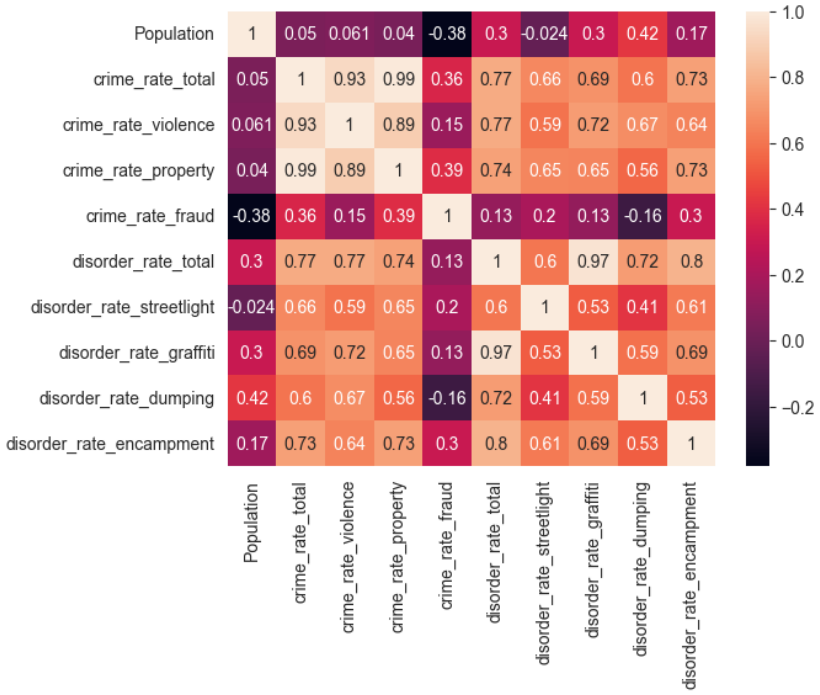

Fig. 3. Correlation matrix (Spearman's correlation)

does not seem to have a big impact on the total crime rate, with broken streetlights having the lowest impact (0.66) and homeless encampments having the highest impact (0.73).

D. Geographic Interpretation

High rates of crime and disorder were especially prominent in business districts. This includes, for instance, the Arts District (ZIP code 90071) and the downtown area (90013), which contains the Skid Row area of L.A., notable for its high population of homeless individuals. The crime rates for these areas are shown in Figure 4. Other areas of high crime and disorder rates, such as the Hollywood Studios and the University of Los Angeles, are likely a statistical effect as the rates are very high because of few registered inhabitants in that area.

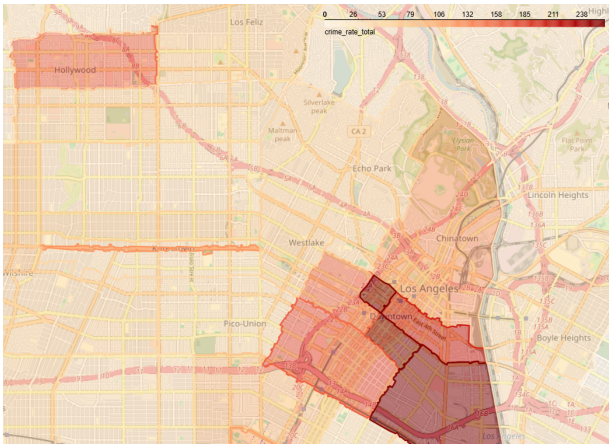

Fig. 4. Regions of L.A. with high crime rates (excerpt)

IV. Conclusions

It can be shown that there is a correlation between signs of disorder and all considered crimes, except fraud, in Los Angeles in 2023. This supports the idea that there's a link between visible disorder and crime.

However, it has to be considered that total disorder was treated as the sum of disorder cases, which means that a broken streetlight and homeless encampments are weighted equally. This often does not reflect the perception of disorder in real life. In addition, the rates are calculated using the number of people who officially live in a region. This does not include tourists or visitors and creates comparably high rates in highly frequented areas with a low population. Further research may be done that takes the difference between housing districts and business districts into account.

Furthermore, there are limitations that are inherent to the nature of the data. Only crimes and signs of disorder that are reported can be considered.

References

[1] "UCR VS COMPSTAT" Accessed: Jan. 2025. [Online]. Available: https://data.lacity.org/api/views/63jg-8b9z/files/fff2caac-94b0-4ae5-9ca5-d235b19e3c44?download=true&filename=UCR-COMPSTAT062618.pdf