Candidate Datasets

Dataset Name & Source	URL	Main Variables	Join Keys / Merge Fields	How It Adds Value	Fairness / Bias Considerations
U.S. Census Bureau / American Community Survey (ACS)	https://data.census.gov Links to an external site.	Median household income, race/ethnicity, education level, unemployment rate, poverty rate, population density	Census tract, ZIP code, or neighborhood	Adds socioeconomic context that can help explain spatial variation in crime rates (e.g., property thefts concentrated in high-density or low-income areas).	ACS data may reflect systemic inequities (e.g., segregation, income inequality). Must interpret carefully to avoid implying causation between demographics and crime.
City of Los Angeles Open Data Portal	https://data.lacity.org Links to an external site.	Neighborhood names, area boundaries, LAPD division codes, population estimates	Area or neighborhood name; LAPD division	Enables geographic visualization of crime hot spots and allows linking of model predictions to real patrol areas.	Geographic boundaries may mask within-neighborhood diversity; risk of reinforcing “dangerous area” labels.
NOAA Climate Data Online	https://www.ncei.noaa.gov/cdo-web Links to an external site.	Daily temperature, precipitation, extreme weather events	Date and ZIP code (or nearest weather station)	Can assess whether weather conditions (rain, heat waves) correlate with crime occurrence patterns.	Climate data itself is neutral, but may not evenly represent micro-climates across Los Angeles.

Step 4: One-Page Write-Up

Dataset Overview
I used the American Community Survey (ACS) dataset, which is maintained by the US Census Bureau. It offers extensive demographic, housing, and socioeconomic data for each census tract and ZIP code in the United States. Key factors include median household income, educational attainment, unemployment rate, racial makeup, poverty rates, and population density. These indicators are updated on a yearly basis, ensuring a constant and trustworthy source of neighborhood-level context.

Integration Potential

The ACS dataset may be physically combined with the LAPD crime dataset using shared place identifiers like ZIP codes or census tracts. Each recorded incident in the LAPD dataset has geographic coordinates or an area name (for example, "Wilshire," "Central," etc.), which may be used to cross-reference with ACS tracts or neighborhoods via a spatial join. This integration allows us to include socioeconomic indicators to our model, linking individual incident data with wider contextual characteristics.

Added Value

Including ACS variables can greatly improve model interpretability and fairness. For example, socioeconomic background may explain why property thefts occur more frequently in regions with crowded rental housing or lower average earnings. It also allows for more in-depth fairness analysis: we can see how forecast accuracy varies among areas with different demographics or wealth levels, ensuring that model outputs do not consistently penalize specific populations. In layman's words, it switches the focus from "where crime happens" to "why certain areas experience more reported incidents."

Limitations

Despite its virtues, the ACS dataset has certain drawbacks. First, demographic and economic indicators can be used as proxies for race or class, posing ethical concerns if read incorrectly. Correlations between socioeconomic characteristics and crime should not be interpreted as causality, since they may represent greater structural disparities such as enforcement methods or historical segregation. Furthermore, ACS statistics are aggregated and updated once a year, so they may not capture short-term changes like

as gentrification or transitory population movements. Finally, census tracts differ in size and population density, which may result in discrepancies when aligned with LAPD divisions.