DATA 400 Project Proposal 2

Predicting Gentrification in Urban Neighborhoods

Research Question: Can gentrification be reliably predicted using publicly available data, and which socioeconomic or environmental factors serve as the strongest early indicators across different cities and time periods?

Motivation: Gentrification is a complex urban process that often leads to rising property values, increased rent, and the displacement of long-term, lower-income residents. It reshapes neighborhoods socially and economically, sometimes accelerating inequality. Predicting which neighborhoods are at high risk of gentrification allows city planners and local communities to take proactive steps toward equitable development. My project aims to provide data-driven insights to inform urban planning policies, support community advocacy, and guide responsible development, while keeping in mind the ethical risks of predictive models in sensitive urban contexts.

Proposed Data Sources: To build a longitudinal dataset, I will combine multiple open and publicly accessible sources:

- Real estate: Zillow, Redfin, Realtor.com (rental and sale prices, listings, price per square foot)
- Crime: FBI's Uniform Crime Reporting (UCR) Program and open data from local police departments (yearly crime rates, types, clearance rates)
- Demographic: U.S. Census Bureau via American Community Survey (income levels, education, racial/ethnic composition, employment)
- Amenities & Infrastructure: Google Places API, Yelp API, OpenStreetMap (number and type of new businesses, transit access, walkability scores)

My analysis will focus on data from the last 5 years in major U.S. metropolitan areas such as New York City, San Francisco, and Chicago, depending on data availability.

Methodology: The project will use a combination of time-series analysis, feature engineering, and machine learning classification to identify neighborhoods at risk of gentrification:

- Preprocessing & Feature engineering:
  - Normalize neighborhood-level data
  - Calculate multi-year change rates (% rent increase over 5 years)
  - Generate rolling time-window features and lag variables
  - Encode significant events (new transit station openings, zoning changes)
- Define target variable: Use binary classification labels (gentrifying or not) based on threshold criteria (for example >30% rent increase combined with demographic shifts)
- Model:

  - Time series models to analyze longitudinal trends
  - Classification models for neighborhood risk prediction
- Evaluation:
  - Use AUC/ROC, precision/recall scores, and cross-validation to assess predictive performance
  - Develop geospatial visualizations to demonstrate neighborhood-level risk

Challenges: I think one major challenge is defining "gentrification" consistently across neighborhoods and cities, as the process is context-specific and multidimensional. Additionally, some data, particularly real-time housing trends or detailed zoning changes, may be incomplete or lagging. Bias in datasets (crime data reflecting over-policing in minority areas) also poses a risk. To mitigate these issues, I will use transparent modeling and consider multiple definitions for gentrification.