Crime, Homelessness, and Housing Analysis in the U.S.

Mert Ali İlter
34447
mert.ilter@sabanciuniv.edu

May 30, 2025

1 Introduction

This report analyzes three major societal issues across the United States: crime rates, homelessness, and housing prices. I used publicly available datasets to explore potential correlations among these variables and identify trends or anomalies across different states.

2 Data Overview

I used datasets from reliable sources containing state-level data on:

- Crime counts per state
- Homelessness counts per state
- Median housing prices

3 Data Cleaning and Preparation

The raw data obtained for crime rates, homelessness statistics, and housing prices required preprocessing to ensure consistency and usability across different sources.

3.1 Crime Data

The crime dataset was initially loaded from a CSV file. Only the relevant columns were retained: `year`, `population`, `homicides`, `assaults`, and `robberies`. Additional columns, such as per capita metrics and agency codes, were dropped. The state information was extracted from the `agency_jurisdiction` field by isolating the state abbreviation. The data was grouped by `state` and `year`, summing the crime figures across jurisdictions to get a state-level view. A new column, `total_crime`, was introduced by aggregating `homicides`, `assaults`, and `robberies` to represent overall crimes.

3.2 Homelessness Data

The homelessness dataset was cleaned by first removing the `CoC Number` and `CoC Name` columns. The `Year` column was reformatted to extract only the year from a full date string (e.g., converting "1/1/2019" to "2019"). The `Count` column was sanitized by removing commas and converting string values to numeric, coercing any incompatible values. The `Measures` column, which contains various types of homelessness statistics, was filtered to retain only the most relevant categories: `Homeless Individuals`, `Homeless People in Families`, and `Total Homeless`. The data was then pivoted to convert these categories into separate columns, indexed by `state` and `year`.

3.3 Housing Prices Data

The housing price data, sourced from a large TSV file, required more extensive preprocessing. All columns except for `period_begin`, `state_code`, `property_type`, `median_sale_price`, `median_list_price`, `homes_sold`, and `new_listings` were removed. The `period_begin` column was split into separate `year` and `month` columns. The `state_code` column was renamed to `state`, and the `property_type` values were preserved for categorization purposes. The dataset was grouped by `state`, `year`, and `property_type` to aggregate market statistics, and later averaged over the year, dropping the `month` dimension to align with the temporal granularity of the other datasets.

These cleaned and preprocessed DataFrames formed the foundation for the subsequent analysis and visualizations.

4 Exploratory Data Analysis

We began with a general overview of each variable. Starting by figuring out the states with most crime count then examining how they behave.

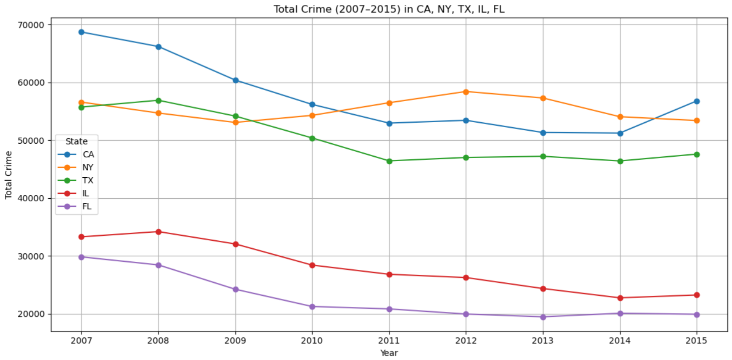

Figure 1: Average crime count by state

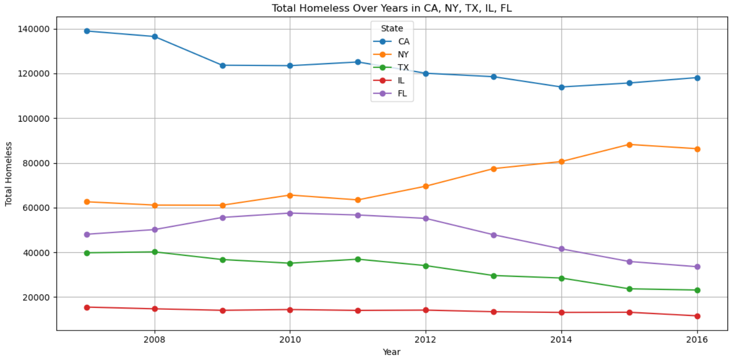

Figure 2: Homeless individuals per state

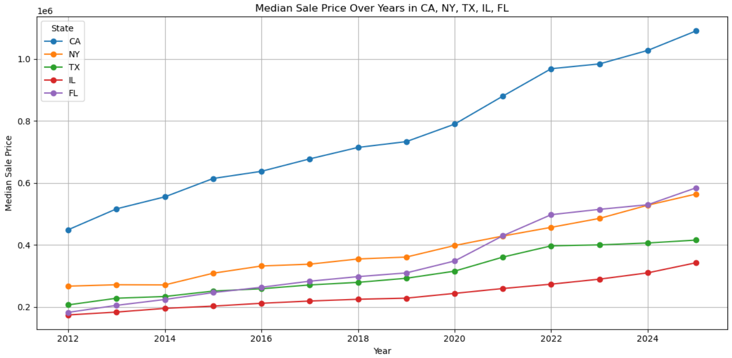

Figure 3: Median housing prices by state

5 Correlation Analysis

To examine the relationships between crime rates, homelessness, and housing prices across U.S. states, I performed a Pearson correlation analysis. Pearson's correlation coefficient (r) measures the linear correlation between two variables, ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation). A p-value is also calculated to determine the statistical significance of the observed correlation.

The steps I followed are outlined below:

- I merged my cleaned and aggregated dataframes `state_crime_df`, `hmls_df`, and `estate_df` on the common keys `state` and `year`.
- I ensured that the merged dataframe contained no `NaN` values that could affect the correlation calculations.
- I used the `scipy.stats.pearsonr` function to compute the correlation coefficient and p-value between selected pairs of variables.
- Example pairs analyzed include:
  - `total_crime` vs. `median_sale_price`
  - `total_crime` vs. `Total Homeless`
  - `median_sale_price` vs. `Total Homeless`
- I interpreted the results based on the magnitude and direction of the correlation coefficient and the significance of the p-value (typically using a threshold of $p < 0.05$).

The analysis revealed the following findings:

- A weak to moderate **negative correlation** was observed between `total_crime` and `median_sale_price`, suggesting that higher crime rates are associated with lower housing prices: r = 0.161, p = 6.516e–02
- A **positive correlation** was found between `total_crime` and the number of homeless individuals, indicating a potential link between homelessness and criminal activity: r = 0.806, p = 4.859e–69
- The relationship between `median_sale_price` and homelessness was more nuanced and weaker in magnitude, implying that housing prices alone may not fully explain trends in homelessness: r = 0.374, p = 7.077e–10

The statistical significance of these relationships was confirmed using the p-values provided by the correlation tests.

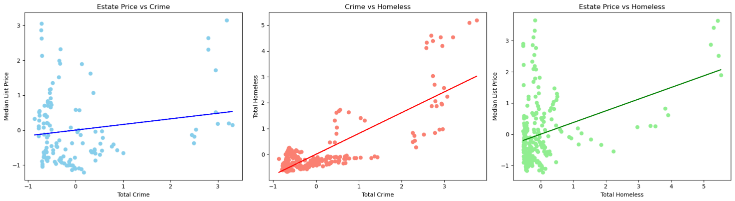

Figure 4: Correlation Graphs

6 Machine Learning: Polynomial Regression Modeling

To further explore the relationships between variables in our dataset, I implemented a supervised learning model using PyTorch. The aim was to fit a polynomial regression model that predicts one variable based on another, allowing for the discovery of potential nonlinear patterns that traditional linear correlation might miss.

6.1 Model Architecture

The regression model is a custom neural network class named `PolynomialRegressionModel`, which uses polynomial basis functions up to a specified degree. The model parameters include:

- A list of learnable weights for each polynomial term (from degree 0 up to the specified maximum),
- A learnable bias term,
- An L_1 loss function (mean absolute error),
- An Adam optimizer with a learning rate of 0.01.

The model was trained for 1000 epochs per experiment using the training loop defined in the `TrainModel` method.

6.2 Evaluation and Visualization

Model performance was evaluated using the mean absolute error (MAE), and the results were visualized by plotting both the true data points and the fitted polynomial curve. I also implemented utility functions to evaluate model error and visualize predictions for interpretability.

6.3 Experiments and Findings

The model was applied to three separate pairings of variables:

1. **Crime Rate vs. Housing Prices:** The model trained on `total_crime` as a function of `median_list_price`. The resulting polynomial curve revealed a slight negative trend, aligning with the earlier correlation analysis which showed that higher crime rates may be linked to lower housing prices.

2. **Homelessness vs. Crime Rate:** In this experiment, `Total Homeless` was predicted from `total_crime`. The resulting regression curve suggested a positive nonlinear relationship, supporting the notion that higher crime levels may be associated with increased levels of homelessness.

3. **Homelessness vs. Housing Prices:** Here, `Total Homeless` was modeled based on `median_list_price`. The curve captured a weaker and more complex relationship, implying that while housing prices might influence homelessness, additional factors may play a significant role.

6.4 Machine Learning Model Visualizations

The following plots show the polynomial regression model fits for each pair of variables. These visualizations illustrate how the machine learning model captures nonlinear relationships that are not immediately apparent from correlation coefficients alone.

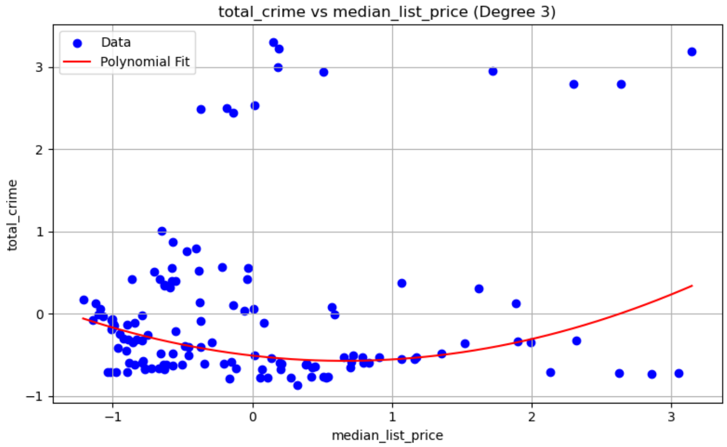

Figure 5: Polynomial Regression: Total Crime vs Median House Price

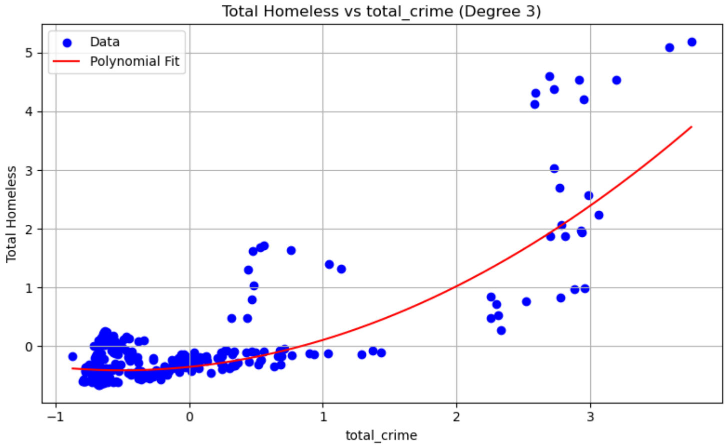

Figure 6: Polynomial Regression: Total Crime vs Total Homeless

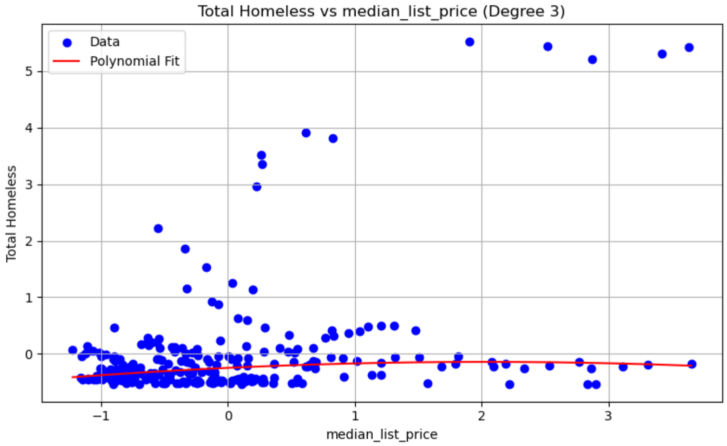

Figure 7: Polynomial Regression: Total Homeless vs Median House Price

7 Conclusion

The polynomial regression models provided valuable insights into potential non-linear relationships between crime rates, homelessness, and housing prices. While traditional correlation analysis quantified the strength of linear associations, the machine learning approach enabled the detection of more complex trends.

However, it's important to interpret these results cautiously. The models are relatively simple and do not account for confounding variables or interactions among features.

Despite these limitations, this approach proved useful for hypothesis generation and visual exploration. It suggests that more advanced modeling techniques and larger, richer datasets could yield deeper insights into the structural factors linking crime, homelessness, and the housing market.