DS Project 1
Data Preprocessing Report

Guichen Zheng

February 17, 2026

Section 1: Overview

The objective of this project is to construct a clean ZIP-level monthly panel dataset integrating multiple data sources:

- Zillow housing price data (ZHVI)
- ACS socioeconomic variables
- NYC crime data
- FRED macroeconomic indicators

The final dataset is structured at the **ZIP** $\times$ **Month** level (from January 2023 onward) and prepared for downstream statistical analysis.

Section 2: Zillow Data Restructuring

The original Zillow dataset was in wide format, with each month represented as a separate column.

Processing

- Identified monthly date columns.
- Standardized ZIP codes to five-digit string format.
- Converted wide format into long format using `melt()`.
- Converted dates to `YYYY-MM`.
- Filtered data to retain observations from January 2023 onward.

Result

A clean long-format panel:

```
zip | month | zhvi
```

This serves as the base structure of the project.

Section 3: Integration of ACS Socioeconomic Data

The ACS dataset provides ZIP-level demographic and economic characteristics.

Selected Variables

- Population
- Median household income
- Poverty base and poverty count
- Labor force and unemployment
- Educational attainment levels

Processing

- Renamed coded ACS variables to readable feature names.
- Standardized ZIP format.
- Constructed rate variables: poverty rate, unemployment rate, higher education count.
- Merged ACS into Zillow dataset using a left join on ZIP.

Each ZIP-month observation now includes static socioeconomic characteristics.

Section 4: Spatial Integration of Crime Data

The NYC crime dataset contains incident-level records without ZIP codes.

Spatial Assignment

- Converted latitude and longitude into geographic point objects.
- Parsed MODZCTA polygon geometries from WKT format.
- Performed a point-in-polygon spatial join to assign each crime to a ZIP area.

Temporal Alignment

- Converted crime dates to datetime format.
- Aggregated to monthly (`YYYY-MM`) level.

Aggregation

Crime records were grouped by ZIP and month to compute:

$$\texttt{crime\_count}_{zip,month}$$

Missing values after merging were replaced with 0.

Section 5: Integration of Macroeconomic Data (FRED)

FRED provides national-level macroeconomic indicators.

Processing

- Converted dates to monthly format.
- Aggregated daily/weekly data to monthly averages.
- Merged on month into the main dataset.

Since FRED variables are national-level indicators, values are identical across ZIP codes within the same month.

Section 6: Missing Value Handling

Missing ratios were computed for all variables. All variables exhibited missing rates below 1%.

Rows containing missing values were removed. A sentinel value (-66666666) in `median_income` was identified and treated as missing.

After cleaning, the dataset contains no missing values.

Section 7: Outlier Detection

Outliers were evaluated using descriptive statistics, logical validation, and statistical detection methods (IQR and Z-score).

Many statistical outliers reflect genuine cross-sectional heterogeneity across ZIP codes rather than data errors. Therefore, no statistical outliers were removed.

Section 8: Final Dataset Structure

The final dataset is a clean panel:

```
zip | month | zhvi | socioeconomic variables | crime_count | macro variables
```

The dataset is spatially and temporally aligned, logically consistent, and free of missing values.