Exploratory Data Analysis (EDA)

In this phase, I first reviewed the structure of our dataset with `df.info()`, confirming the number of rows, columns, and any missing values. Using `df.describe()`, I then explored central tendencies (mean, median) and measures of dispersion (standard deviation, min, and max) across all numeric features. These summary statistics quickly highlighted how variables like poverty rate, median rent as a percentage of household income, and percentage of owned housing compared in terms of typical values and extreme cases.

To gain deeper insights, I produced **histograms and KDE plots** for **every numeric column**, rather than focusing on just a few variables. This comprehensive approach allowed me to observe important distributional characteristics—such as skewness, the presence of multiple peaks, or any clustering patterns. In a few instances, certain variables showed heavy tails and potential outliers, suggesting additional scrutiny might be needed.

Furthermore, I created **boxplots** for each numeric feature, which made it much easier to spot observations lying well beyond the interquartile range. This step confirmed that while most features clustered within a reasonable range, several outlier points stood out for variables like poverty rate and rent burden. Whether these points reflect true extremes or data-entry anomalies remains an open question for further investigation. Ultimately, these findings informed how we might handle outliers—such as through transformations or capping—and underscored which variables might require scaling to ensure stable results in subsequent modeling. By clarifying these distributional nuances, the EDA step has laid a stronger foundation for the next phases of our analysis.