Exploratory Data Analysis (EDA) - Part 1 (by Xinyi Gui)

Introduction

The first part of our exploratory data analysis focused on understanding the overall distribution and basic statistical properties of our dataset, particularly the numeric features. This step was essential to identify skewness, outliers, and underlying patterns that might impact subsequent analysis or modeling.

The First Exploration: Summary Statistics

The initial approach involved converting any categorical variables (like `state`) to the correct data type, ensuring they were excluded from quantitative analyses. We then generated **summary statistics** (`df.describe()`) for the remaining numeric columns. The main observations included:

- **Means and Medians**: Several features (e.g., `PersPerFam`) had close mean and median values, suggesting relatively symmetrical distributions.
- **Range of Values**: Certain variables, such as `householdsize` and `PctHousOwnOcc`, showed a wide range—indicating possible right-skew or the presence of extreme values (e.g., 100% owner occupancy).
- **Missing or Anomalous Data**: We confirmed that no major columns had a large portion of missing values, which simplified further analysis.

By understanding the dataset structure, we identified which features required deeper examination due to potential skewness or abnormal ranges.

The Second Exploration: Distribution Visualizations

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.

Histograms & KDE Plots

Next, we created **histograms** and **KDE plots** for each numeric variable. This allowed us to:

- **Spot Skewness**: A few variables, like `householdsize`, were heavily right-skewed.
- **Check for Multiple Peaks**: For instance, `PovLess3BR` hinted at possible bimodal behavior in some areas, suggesting distinct subgroups in the data.

- **Confirm Summary Statistics**: Where means and medians aligned, the distribution tended to be more or less unimodal and symmetric.

Boxplots

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.

- **Potential Outliers**: Variables like PctHousOwnOcc at 1.0 (100%) might be genuine (e.g., certain communities) or questionable data points.
- **IQR and Whisker Span**: Most features had a modest Interquartile Range (IQR), but the whiskers revealed a handful of regions or observations well outside the typical range.
- **Data Integrity Checks**: Understanding whether extreme values are errors or rare but valid cases is critical. We flagged these for future validation.

These distribution visualizations gave us a clearer view of how each feature behaves, where clusters occur, and which observations might be outliers.

Summary of Key Findings

1. **Right-Skewed Distributions**: Several variables (e.g., householdsize, PctHousOwnOcc near 1.0) showed evidence of long tails, which may require transformations if used in parametric models.
2. **Outlier Candidates**: Boxplots revealed distinct outliers that could bias statistical results or model training if not addressed.
3. **Potential Subgroup Patterns**: Certain features hinted at more than one peak in their KDE plots, suggesting the data might not be a single homogeneous group.
4. **Data Integrity**: The dataset appeared mostly consistent and complete, with few missing values; however, the validity of extreme observations remains to be confirmed.

Challenges Faced & Future Recommendations

- **Skewed Data Handling**: For variables with strong skewness, consider transformations (log, Box-Cox) or outlier mitigation strategies (winsorizing, robust scaling).
- **Validating Extreme Values**: Investigate whether the high or low extremes (e.g., 100% homeownership) represent genuine cases or coding anomalies.
- **Segmentation Possibility**: If the data truly has bimodal or multimodal distributions, exploring clusters or state-level groupings might yield more nuanced insights.

- **Further Feature Engineering**: Future analysis could benefit from combining or deriving new variables (e.g., ratio of NumUnderPov to local population) to capture deeper socioeconomic nuances.

Xinyi Gui 's Contribution

For **EDA Part 1**, I conducted the **summary statistics** review and created the **histograms, KDE plots, and boxplots** for all numeric columns. I identified skewness, outliers, and potential anomalies, establishing a foundation for more advanced correlation and segmentation analyses in Part 2.