**Data Acquisition**

**Data Sources:** The dataset was procured from UC Irvine Machine Learning Repository and other open APIs, primarily imported from the UCI Machine Learning Repository.

**2. Detection and Rectification of Data Inconsistencies**

**2.1 Data Type Verification**

After downloading the data, we check the type of each feature and find that:

* 125 features are float.
* 1 feature (feature\_3) is string.
* 1 feature is integer.
* Combining with the data description on the website:
  + state (feature\_0), county (feature\_1), community (feature\_2), and fold (feature\_4) are categorical features.

122 features are predictive, 5 are non-predictive, and 1 is the goal.

**2.2 Handling Missing Values**

Imputing missing values is vital to maintain dataset integrity.

* **Categorical Features:** Used K-Nearest Neighbors (KNN) to handle missing data, ensuring accuracy by inferring relationships from nearby data points.
* **Numerical Features:** Applied mean imputation, replacing missing values with the mean of the respective feature.

**3. Data Selection**

We divided features into two parts:

* **Predictive Features:** (feature\_0 : feature\_125)
* **Goal Feature:** (feature\_126)

To simplify future machine learning methods, we applied feature selection methods including:

**Lasso Regression**

* **Selection Criteria:** L1 regularization shrinks some coefficients exactly to zero, allowing variable selection and complexity control.
* **Selected Variables:** feature\_6, feature\_47, feature\_53

**Ridge Regression**

* **Selection Criteria:** L2 regularization compresses coefficients but does not shrink them to zero, effective for multicollinearity.
* **Selected Variables:** Ridge does not explicitly select variables but emphasizes features similar to Lasso.

**ElasticNet**

* **Selection Criteria:** Combines L1 and L2 regularization, suitable for correlated features.
* **Selected Variables:** feature\_6, feature\_46, feature\_47, feature\_53, feature\_74

**Best Subset Selection**

* **Selection Criteria:** Evaluates all feature subsets and selects the best based on adjusted R².
* **Selected Variables:** feature\_6, feature\_47, feature\_74

**Stepwise Selection**

* **Selection Criteria:** Combines forward selection and backward elimination, optimizing based on statistical significance.
* **Selected Variables:** feature\_0, feature\_5, feature\_13, feature\_14, feature\_16, feature\_18, feature\_21, feature\_28, feature\_31, feature\_34, feature\_41, feature\_44, feature\_47, feature\_51, feature\_53, feature\_71, feature\_74, feature\_77, feature\_89, feature\_90, feature\_91, feature\_93, feature\_111

**Ordinary Least Squares (OLS)**

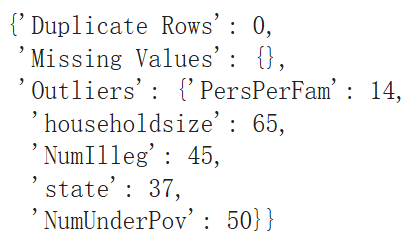
* **Selection Criteria:** Features with p-values below 0.05 are considered significant.
* **Selected Variables:** feature\_0, feature\_14, feature\_24, feature\_31, feature\_42, feature\_51, feature\_69, feature\_71, feature\_77, feature\_78

We count the frequency of feature appears by each method and select 10 the most significant features. Selecting the most frequently occurring features as the final variables for the model is based on several key considerations. First, features consistently chosen across multiple methods demonstrate their importance under different statistical criteria and modeling techniques, reducing the likelihood of selection bias and enhancing their robustness. Second, these features help mitigate overfitting, as their significance is less likely to be a result of model-specific assumptions and more likely to represent genuine, generalizable relationships within the data. Additionally, incorporating features selected from multiple analytical approaches provides a more comprehensive understanding of the data, allowing for a well-balanced trade-off between model complexity and predictive power. Lastly, this approach typically leads to improved model performance and accuracy, particularly when applied to unseen data. By leveraging the consistency of feature selection across different methods, this strategy ensures the development of a model that is both reliable and interpretable.

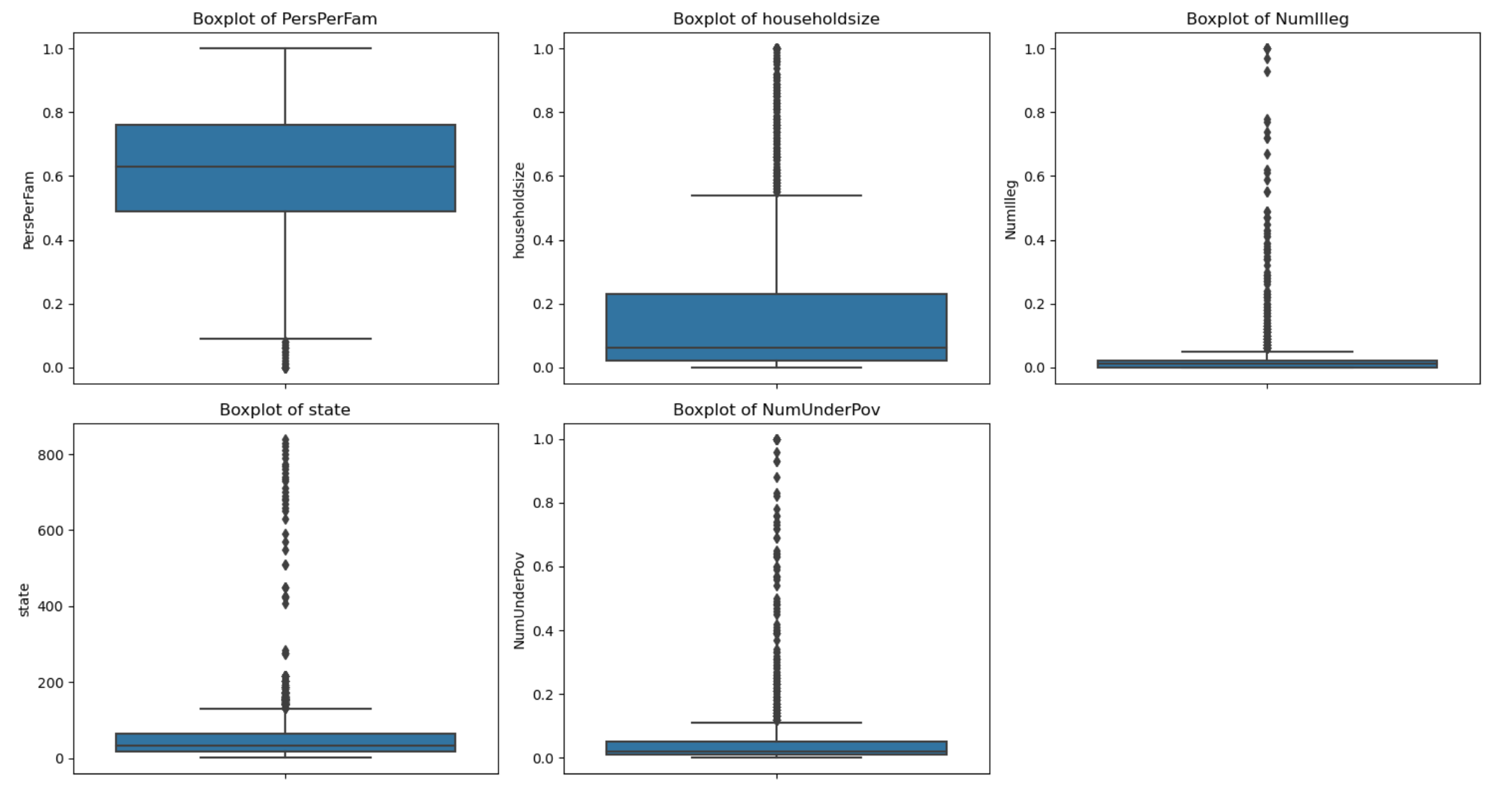
Based on frequency across methods, we selected the 10 most significant features:

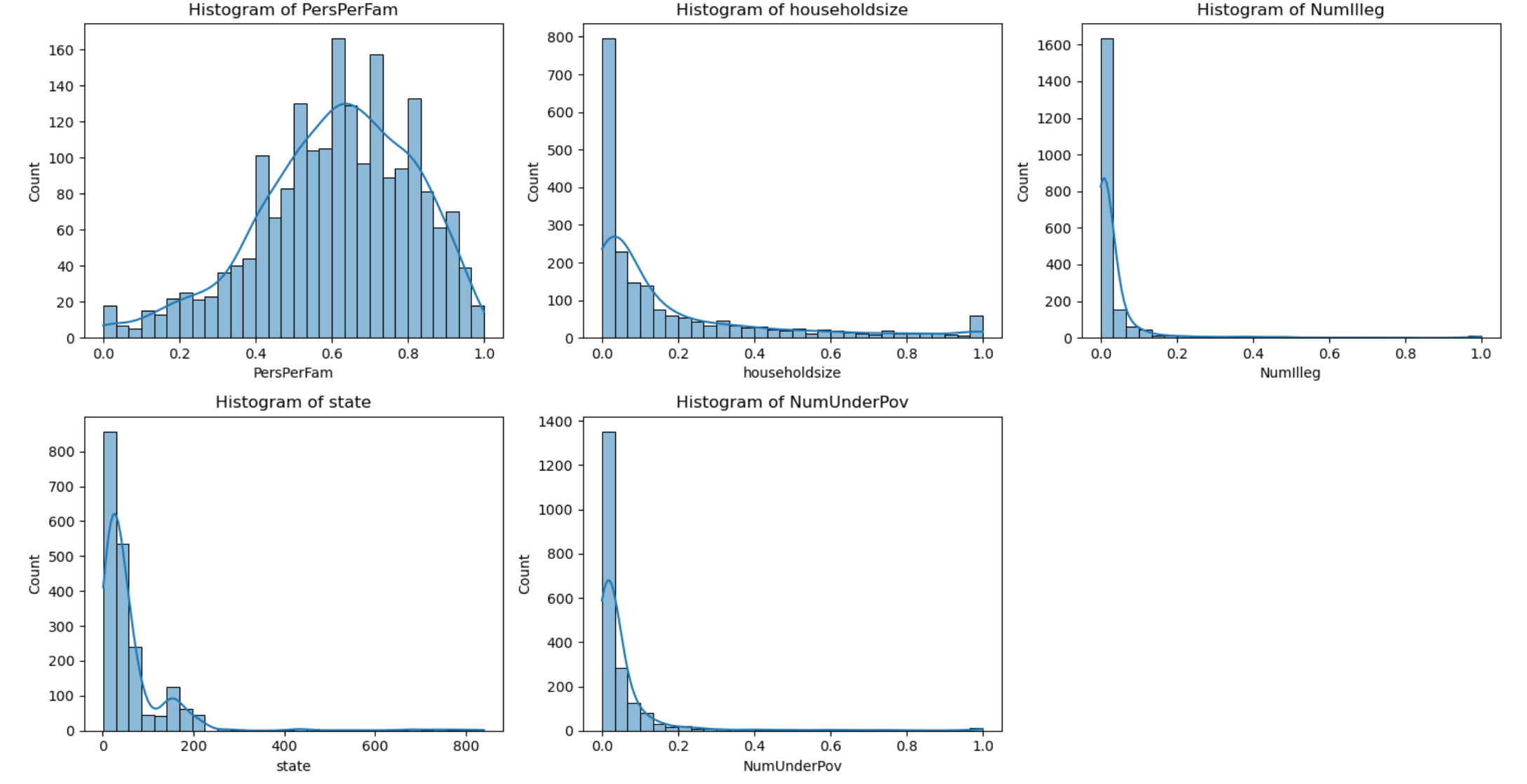
1. **PersPerFam** (feature\_47) - float
2. **PctHousLess3BR** (feature\_74) - float
3. **householdsize** (feature\_6) - float
4. **NumIlleg** (feature\_53) - float
5. **state** (feature\_0) - categorical
6. **PersPerRentOccHous** (feature\_71) - float
7. **PctHousOwnOcc** (feature\_77) - float
8. **PctWorkMomYoungKids** (feature\_51) - float
9. **MedRentPctHousInc** (feature\_91) - float
10. **NumUnderPov** (feature\_31) - float

**4. Identify and Handle Outliers** After removing duplicate data, we identified outliers using the Z-score method.



Only 5 features contained outliers, for which we applied appropriate handling techniques:





1. PersPerFam (Average Household Members per Family)

**Method Applied**: Winsorization (Capping at 5th and 95th percentiles)

**Rationale:**

* This variable represents the average number of household members per family. While most values fall within a reasonable range, extreme values may occur due to data entry errors or unusual cases.
* Winsorization replaces extreme values below the 5th percentile and above the 95th percentile with the nearest valid values, ensuring that extreme outliers do not disproportionately influence the model.

**Impact:**

* Prevents extreme values from distorting the model while preserving the overall distribution.
* Maintains data integrity by avoiding unnecessary data loss.

**2. householdsize (Household Size)**

**Method Applied: Log Transformation + Winsorization**

**Rationale:**

* The distribution of householdsize is heavily right-skewed, meaning that while most values are low, there are a few extremely high values.
* Log transformation helps normalize the distribution by reducing the range of extreme values, making the data more symmetric and improving its suitability for regression models.
* Winsorization further limits the effect of outliers by capping extreme values at reasonable thresholds.

**Impact:**

* Reduces skewness, making the variable more suitable for linear modeling.
* Minimizes the influence of extreme values while maintaining a meaningful distribution.

**3. NumIlleg (Number of Undocumented Immigrants)**

**Method Applied: Log Transformation + Winsorization**

**Rationale:**

* This variable also exhibits a highly right-skewed distribution, with most values being low while a few areas have an exceptionally high number of undocumented immigrants.
* Log transformation compresses large values, reducing their disproportionate impact.
* Winsorization ensures that extreme values, even after log transformation, remain within a controlled range.

**Impact:**

* Prevents extreme values from dominating the dataset.
* Improves linearity, making the variable more effective in predictive modeling.

**4. state (State Code)**

**Method Applied: Removal of Out-of-Range Values**

**Rationale:**

* The state variable should represent categorical values corresponding to US state codes, typically ranging from 1 to 100.
* If values exceed this range, they are likely due to data entry errors or irrelevant records.
* Removing values greater than 100 ensures that the variable remains valid and meaningful.

**Impact:**

* Maintains categorical consistency by removing invalid entries.
* Prevents erroneous data from influencing state-based analyses.

**5. NumUnderPov (Number of People Below Poverty Line)**

**Method Applied: Log Transformation + Winsorization**

**Rationale:**

* The distribution is highly skewed, with a few areas having disproportionately high poverty numbers.
* Log transformation reduces the effect of extreme values while preserving relative differences between observations.
* Winsorization further limits the impact of extreme outliers to ensure stable modeling.

**Impact:**

* Improves data distribution, making it more suitable for statistical models.
* Ensures that extreme cases do not disproportionately affect the model’s results.

| **Feature** | **Method Applied** | **Rationale & Impact** |
| --- | --- | --- |
| **PersPerFam** | Winsorization (5th & 95th Percentile Capping) | Prevents extreme values from skewing the model while retaining data integrity. |
| **householdsize** | Log Transformation + Winsorization | Reduces skewness and minimizes extreme value effects. |
| **NumIlleg** | Log Transformation + Winsorization | Prevents extreme values from dominating the dataset while preserving key patterns. |
| **state** | Removal of Invalid Categorical Values (>100) | Ensures categorical consistency and removes incorrect entries. |
| **NumUnderPov** | Log Transformation + Winsorization | Smooths out extreme values while improving distribution normality. |

By implementing these outlier handling techniques, we ensure a more robust and interpretable model.

Contribution：

In this project, I conducted a comprehensive data preprocessing pipeline for the **Communities and Crime** dataset. First, I verified and corrected data types, ensuring categorical and numerical features were properly identified. To handle missing values, I applied **K-Nearest Neighbors (KNN) imputation for categorical features** and **mean imputation for numerical features** to maintain data integrity. Next, I performed **feature selection** using multiple methods, including **Lasso, Ridge, ElasticNet, OLS, Best Subset Selection, and Stepwise Selection**, ultimately identifying **10 key features** to enhance model performance. Additionally, I removed duplicate data and detected outliers using **Z-score analysis** and applied appropriate treatments such as **Winsorization, log transformation, and removal of invalid categorical values** to ensure a well-distributed dataset.