Project 1 Report

Team 4 - Project 1 Report: Data Acquisition, Cleaning, Preprocessing and Feature Engineering for Exploratory Analysis

Introduction & Dataset Description

Understanding the socioeconomic and housing factors that contribute to violent crime rates is a key objective in crime analysis and public policy. This project explores a subset of the Communities and Crime Dataset, which contains demographic, economic, and crime-related data across various U.S. communities. The target variable for this analysis is:

ViolentCrimesPerPop - The violent crime rate per population.

To enhance the predictive modeling of violent crime rates, we identified and selected 10 key features that are likely to influence crime rates based on socioeconomic and housing conditions. The selected features include:

- PersPerFam - The average number of persons per family.
- PctHousLess3BR - The percentage of houses with less than 3 bedrooms.
- householdsize - The average size of a household.
- NumIlleg - The number of children born out of wedlock.
- state - A numerical identifier for the state.
- PersPerRentOccHous - The number of persons per rent occupied house.
- PctHousOwnOcc - The percentage of houses that are owner-occupied.
- PctWorkMomYoungKids - The percentage of working mothers with young children.
- MedRentPctHousInc - The median rent as a percentage of household income.
- NumUnderPov - The number of individuals living under the poverty line.

To further enhance the predictive power of our dataset, we introduced several engineered features that capture interactions and relationships between key variables. These new features help to account for complex socioeconomic dynamics that may not be fully captured by individual variables. The engineered features include:

- PctHousOwnOcc_PctHousLess3BR - The ratio of owner-occupied houses to houses with fewer than 3 bedrooms.
- PersPerFam_householdsize - The ratio of persons per family to the average household size.
- NumUnderPov_householdsize - The ratio of the number of individuals under poverty to household size.
- NumIlleg_PctWorkMomYoungKids - The ratio of the number of children born out of wedlock to the percentage of working mothers with young children.
- MedRentPctHousInc_PctHousOwnOcc - An interaction term between median rent percentage of household income and the percentage of owner-occupied houses.
- PersPerRentOccHous_PctWorkMomYoungKids - An interaction term between the number of persons per rented occupied house and the percentage of working mothers with young children.

By focusing on these features and their interactions, this project aims to uncover the relationships between housing, economic stressors, and violent crime rates. The dataset provides a foundation for understanding how community-level economic and housing factors contribute to crime and can inform targeted policy interventions.

Data Acquisition Methodology

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

Cleaning & Preprocessing Steps

Data Type Verification

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

- 125 features are floats.
- 1 feature (feature_3) is a string.
- 1 feature is an 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.

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.

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.
- **Elastic Net**
  - 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 R^2.
  - 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 appearances by each method and select the 10 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

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** (The average number of persons 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** (The average size of a household)
    - 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** (The number of children born out of wedlock)
    - 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 children born out of wedlock.
        - 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** (A numerical identifier for the state)
    - 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** (The number of individuals living under the 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.

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

Exploratory Data Analysis (EDA) - Part 1

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 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.

To gain deeper insights, we produced histograms and KDE plots for every numeric column, rather than focusing on just a few variables. With these visualizations we were able 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.

This comprehensive approach allowed us 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, we 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.

Exploratory Data Analysis (EDA) - Part 2

The second part of the exploratory data analysis was prefaced by converting 'state' to a categorical variable, so that it was not confused with the numerical data.

The first exploration conducted was the visualization of the correlation matrix to identify highly correlated features of the 9 features we are focusing on (excluding 'state'). In this matrix, it was observed that the features with the highest correlation (above the absolute value of 0.5) were the following pairs:

- PersPerRentOccHous & PctHousOwnOcc
- NumUnderPov & NumIlleg
- PersPerFam & PersPerRentOccHous
- PctHousOwnOcc & PersPerFam
- PersPerFam & householdsize
- PersPerFam & NumIlleg
- PctHousLess3BR & PersPerRentOccHous
- PctHousLess3BR & PctHousOwnOcc
- NumUnderPov & PersPerFam
- NumIlleg & householdsize
- PersPerRentOccHous & NumUnderPov
- NumUnderPov & PctHousOwnOcc
- PersPerFam & PctHousLess3BR
- NumIlleg & PersPerRentOccHous

The second exploration conducted was the visualization of the 10 highest correlated feature pairs from the aforementioned list in scatterplots. The observations seen in each of these plots were as follows:

- PersPerRentOccHous vs. PctHousOwnOcc: The number of persons per rent occupied house has a strong positive correlation with the percentage of houses that are owner-occupied. This may give insights into highly related housing metrics. This correlation likely exists because both features are directly related to housing occupancy and ownership patterns. As the number of people living in rented occupied housing increases, the percentage of owner-occupied housing also increases, which might suggest an overall growth in housing demand. This may also give insights into economic implications. If homeownership rates are high, it could imply that rental housing is being occupied at a higher density (more people per rental unit). This could indicate affordability challenges, where more individuals share rental units despite high homeownership rates.
- NumUnderPov vs. NumIlleg: The number of individuals living under the poverty line has a positive correlation with the number of children born out of wedlock. This may give insights into economic hardship and family structure. Higher poverty levels may be associated with higher rates of children born out of wedlock. Financial instability can impact family dynamics, potentially leading to fewer marriages and higher numbers of single-parent households. This may also give insights into educational and socioeconomic factors. Poverty may be linked to lower access to education, which in turn can influence family planning decisions. Lack of access to healthcare and reproductive education in low-income areas may contribute to higher birth rates outside of marriage.
- PersPerFam vs. PersPerRentOccHous: The average number of persons per family has a positive correlation with the number of persons per rent occupied house. This may give insights into housing affordability and family size. Areas with higher family sizes may also have higher rental occupancy rates due to housing costs. Families may prefer renting rather than purchasing due to income limitations or high property prices.
- PctHousOwnOcc vs. PersPerFam: The percentage of houses that are owner-occupied has a positive correlation with the average number of persons per family. This might suggest that larger families are more likely to own homes. Households with more people may prefer home ownership over renting to have more space and stability. Larger families may prioritize long-term investments in housing rather than renting. This may also give insights into financial stability and housing decisions. Families who own homes might also have higher financial stability, allowing them to support larger households. On the other hand, rental-heavy areas might see smaller family sizes due to space and cost constraints.
- PersPerFam vs. householdsize: The relationship between the average number of persons per family and the average size of a household is scattered and less structured, suggesting a weaker correlation between these two variables. A stronger correlation was expected since larger families might live in larger households. The scattered nature of the plot suggests that household size does not consistently increase with family size. This might suggest variability in living arrangements. In some cases, multiple small families might live together, increasing household size but not family size. Conversely,

some large families may live in smaller households due to housing constraints or financial limitations. This might also suggest different definitions of household versus family. Household size includes all people living under one roof, which could include non-family members, roommates, or unrelated tenants. Family size refers to biological or legal family members, excluding unrelated occupants. One thing to note is the high density of points near the bottom. This may suggest that some values of householdsize could be discretized or limited in certain cases.

- PersPerFam vs. NumIlleg: The average number of persons per family has a positive correlation with the number of children born out of wedlock. This suggests that larger families may have more children born out of wedlock. In general, areas with larger family sizes may also have higher rates of children born out of wedlock, either due to cultural, economic, or social factors. If family size is increasing, it makes sense that the number of children (including those born out of wedlock) may also increase. One thing to note is possible data discretization or anomalies. The horizontal clustering suggests that NumIlleg might be reported in fixed increments rather than as continuous values. This could indicate that the data is bucketed or rounded, making the relationship appear less fluid than expected.

- PctHousLess3BR vs. PersPerRentOccHous: The percentage of houses with less than 3 bedrooms has a positive correlation with the number of persons per rent occupied house. This may suggest that there is higher rental occupancy in areas with more smaller homes, and lower rental density in regions with larger homes, which may likely be due to housing demand and affordability constraints. One thing to note is the distinct vertical clustering at specific values, which suggests that the data for PctHousLess3BR is discretized, meaning it is a bucketed variable rather than a continuous percentage. This may indicate that areas are classified into low, medium, or high proportions of small homes, rather than having a true percentage distribution.

- PctHousLess3BR vs. PctHousOwnOcc: Refer to previous bullet point for note about distinct vertical clustering with respect to PctHousLess3BR. The percentage of houses with less than 3 bedrooms has a positive correlation with the percentage of houses that are owner-occupied. This suggests that in areas with a high percentage of small homes, there may also be higher homeownership rates. This could be due to affordability as smaller homes may be more affordable, leading to higher ownership. This shows a possible link between housing size and ownership trends. When PctHousLess3BR = 0.0, meaning most houses have 3+ bedrooms, PctHousOwnOcc appears to vary significantly. When PctHousLess3BR = 1.0, meaning most houses are small, PctHousOwnOcc is often high, suggesting high ownership rates in small homes.

- NumUnderPov vs. PersPerFam: The number of individuals living under the poverty line has a weak to moderate positive correlation with the average number of persons per family. This may suggest that larger families may be more common in low-income areas. This might also suggest that economic stress leading o larger household sizes. In economically struggling regions, families may have more dependents per household,

increasing the average number of persons per family. One thing to note is the vertical clustering of points, suggesting data discretization in regards to NumUnderPov, which may have been recorded in fixed increments rather than as a continuous variable.

- NumIlleg vs. householdsize: The number of children born out of wedlock has a weak or no strong correlation with the average household size. This suggests that other factors, like economic conditions and home ownership rates, may be stronger predictors or household size. One thing to note is the vertical clustering of points, which suggests data discretion in regards to NumIlleg, which may have been recorded in fixed increments rather than as a continuous variable.

The third exploration conducted was creating bar charts of the data count across states as well as the mean values of the 9 features (excluding 'state') that are focused on. The first bar chart showed how many data points were collected for all states in the data. This ranged from the highest amount of data points in a state being right under 80, and the lowest being 1. The second bar chart showed the mean values of the 9 features calculated across all data points from all states. To see if there was a significant difference between the mean values of the highlighted features between the states with the highest counts of data points and the states with the lowest counts of data points, the following four bar charts were subsequently created: Top 10 States by Data Count, Mean Values of Features for Top 10 States, Bottom 10 States by Data Count, Mean Values for Features for Bottom 10 States. The following was observed for the mean values of each feature:

- PersPerFam: Overall had a mean value of ~0.62. Top 10 had a mean value of ~0.65 and Bottom 10 had a mean value of ~0.57.
- PersPerRentOccHous: Overall had a mean value of ~0.56. Top 10 had a mean value of ~0.58 and Bottom 10 had a mean value of ~0.58.
- PctHousOwnOcc: Overall had a mean value of ~0.55. Top 10 had a mean value of ~0.56 and Bottom 10 had a mean value of ~0.56.
- PctWorkMomYoungKids: Overall had a mean value of ~0.49. Top 10 had a mean value of ~0.49 and Bottom 10 had a mean value of ~0.51.
- MedRentPctHousInc: Overall had a mean value of ~0.45. Top 10 had a mean value of ~0.47 and Bottom 10 had a mean value of ~0.33.
- PctHousLess3BR: Overall had a mean value of ~0.32. Top 10 had a mean value of ~0.35 and Bottom 10 had a mean value of ~0.35.
- householdsize: Overall had a mean value of ~0.12. Top 10 had a mean value of ~0.11 and Bottom 10 had a mean value of ~0.15.
- NumUnderPov: Overall had a mean value of ~0.03. Top 10 had a mean value of ~0.02 and Bottom 10 had a mean value of ~0.06.

- NumIlleg: Overall had a mean value of ~0.02. Top 10 had a mean value of ~0.01 and Bottom 10 had a mean value of ~0.03.

The key takeaways from the bar chart observations are as follows:

- States with fewer data points tend to have lower rent burdens, higher poverty rates, larger household sizes, and higher rates of non-marital births.
- States with more data points tend to have higher rent burdens, lower poverty rates, and slightly larger families but smaller household sizes.
- Most homeownership and rental-related features remain stable, suggesting consistent housing trends across different states.

Feature Engineering Process & Justification

In our feature engineering process, we created several ratio-based and interaction features to better capture relationships between key socioeconomic variables. These features provide more nuanced insights into housing, income distribution, and family structure, which can improve model performance by enhancing predictive power.

Ratio-Based Features

```
PctHousOwnOcc_PctHousLess3BR = PctHousOwnOcc / (PctHousLess3BR + 1e-5)
```

- Captures the relationship between the proportion of owner-occupied housing and the percentage of houses with fewer than three bedrooms.
- A high value may indicate a preference for smaller owner-occupied homes, while a low value may suggest ownership is more common in larger homes.
- Helps assess housing market dynamics and affordability trends.

```
PersPerFam_householdsize = PersPerFam / (householdsize + 1e-5)
```

- Normalizes the number of people per family by the overall household size.
- Helps differentiate between family-based households and shared living arrangements.
- A higher value may indicate a concentration of family members within households, whereas a lower value may suggest more non-family occupants.

`NumUnderPov_householdsize = NumUnderPov / (householdsize + 1e-5)`

- Highlights the proportion of individuals living under the poverty line relative to total household size.
- Provides a more precise measurement of poverty density, which may be more informative than absolute poverty counts.
- Captures localized socioeconomic distress and its potential impact on the community.

`NumIlleg_PctWorkMomYoungKids = NumIlleg / (PctWorkMomYoungKids + 1e-5)`

- Assesses the number of children born out of wedlock in relation to the percentage of working mothers with young kids.
- A higher value may indicate social or economic pressures that contribute to higher rates of single-parent households.
- Useful for analyzing child welfare, family stability, and labor market participation among mothers.

Interaction Features

`MedRentPctHousInc_PctHousOwnOcc = MedRentPctHousInc * PctHousOwnOcc`

- Captures the interaction between median rent as a percentage of household income and the percentage of owner-occupied homes.
- Helps analyze how rental affordability correlates with home ownership rates.
- A higher value might suggest that in areas where rent is a significant burden, home ownership rates tend to be either relatively low (due to affordability constraints) or high (due to an incentive to purchase rather than rent).

`PersPerRentOccHous_PctWorkMomYoungKids = PersPerRentOccHous * PctWorkMomYoungKids`

- Combines the number of people per rented occupied household with the percentage of working mothers with young kids.
- Helps capture living density in rental housing and how it interacts with workforce participation among mothers.
- A higher value may indicate areas where working mothers are more likely to live in densely populated rental homes, which could have implications for childcare access and economic well-being.

We also applied various transformation techniques to improve model performance.

Target Encoding

- State: Applied target encoding with cross-validation.

Standardization

The following features were standardized to ensure consistent scaling:

- PersPerFam (Persons per Family)
- PersPerRentOccHous (Persons per Rented Occupied House)
- PctHousOwnOcc (Percentage of Owner-Occupied Houses)
- PctWorkMomYoungKids (Percentage of Working Mothers with Young Kids)
- MedRentPctHousInc (Median Rent as a Percentage of Household Income)

Log Transformation

To normalize skewed distributions, log transformation was applied to:

- NumIlleg (Number of children born out of wedlock)
- NumUnderPov (Number of People Under Poverty Line)

Binarization

To create binary indicators, binarization was applied to:

- PctHousLess3BR (Percentage of Houses with Less than 3 Bedrooms)

Summary of Key Findings

1. Distribution Analysis
    - 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.
    - Outlier Candidates: Boxplots revealed potential outliers that could bias statistical results or model training if not addressed. The validity of these extreme observations remains to be confirmed.
    - 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.

2. Correlation Analysis

   - Highly correlated feature pairs were identified, with 14 pairs showing correlation above 0.5 (absolute value).
   - Key positive correlations and insights:
     - Housing trends:
       * PersPerRentOccHous & PctOwnOcc: Suggests a link between rental occupancy rates and home ownership trends, possibly indicating affordability challenges.
       * PctHousLess3BR & PctHousOwnOcc: In areas with more small homes, home ownership rates tend to be higher, likely due to affordability.
     - Socioeconomic factors:
       * NumUnderPov & NumIlleg: Higher poverty levels correlate with higher non-marital birth rates, possibly due to economic instability affecting family structure.
       * PersPerFam & NumIlleg: Larger families also tend to have higher numbers of children born out of wedlock, which could relect cultural and economic influences.
     - Household size trends:
       * PersPerFam & householdsize: Expected to be strongly correlated, but the scatterplot showed weak correlation, suggesting diverse living arrangements rather than a direct link.

3. Scatter Plot Observations

   - Key insights into housing & economic trends:
     - Larger families correlate with home ownership rates, suggesting that families may prioritize buying homes over renting for stability and space.
     - Higher rental occupancy in areas with small homes, indicating housing constraints and affordability issues.
     - Weak correlation between household size and family size, possibly due to non-family members (roommates, tenants) being included in households.
     - Data discretization observed in some features (NumIlleg, NumUnderPov, PctHousLess3BR), suggesting grouped or bucketed values rather than continuous distributions.

4. Bar Chart Analysis Across States

- Key differences between states with the most and least data points:
  - States with fewer data points tend to have:
    - Lower rent burdens
    - Higher poverty rates
    - Larger household sizes
    - Higher rates of non-marital births
  - States with more data points tend to have:
    - Higher rent burdens
    - Lower poverty rates
    - Slightly larger families but smaller household sizes
  - Home ownership and rental-related features remained stable across different states.

Challenges Faced & Future Recommendations

Challenges Faced

1. Data Discretization & Grouping:
   - Several features, such as NumIlleg, NumUnderPov, and PctHousLess3BR, appeared to have discretized or bucketed values rather than continuous distributions.
   - This made it difficult to interpret smooth relationships in scatter plots, as many data points were clustered at specific values instead of forming natural trends.
2. Correlation Complexity:
   - While some feature pairs had high correlation values, their scatter plots did not always display a clear linear trend.
   - This suggests that certain relationships may be influenced by underlying variables or non-linear patterns, which simple correlation analysis may not fully capture.
3. State-Level Data Distribution Variability:
   - The number of data points per state varied significantly, with some states having as few as 1 data point while others had nearly 80.

- This imbalance may have skewed the mean feature values for states with fewer data points, potentially making it difficult to draw generalizable conclusions across states.

4. Interpreting Socioeconomic Indicators:

   - Relationships between variables such as poverty levels, family size, and home ownership are often influenced by multiple factors (e.g. policy, economic conditions, regional differences).
   - Without additional external datasets (e.g. employment rates, education levels, or geographic indicators, it was challenging to fully understand the causation behind observed trends.

Future Recommendations

1. Improve Data Quality & Granularity:

   - Investigate whether features like NumIlleg and NumUnderPov can be obtained with more precise values rather than fixed increments.
   - Check for potential data transformations (e.g. scaling or binning) that could make relationships more interpretable.
   - Investigate whether the high or low extremes (e.g. 100% home ownership) represent genuine cases or coding anomalies.

2. Explore Non-Linear & Interaction Effects:

   - Some feature relationships might be better explained using polynomial regression, log transformations, or interaction terms rather than simple correlations.
   - Conduct clustering or segmentation analysis to explore whether certain trends vary based on urban vs. rural states or economic zones.

3. Balance Data Representation Across States:

   - If possible, consider downsampling states with excessive data points or aggregating insights for states with limited data.
   - Explore whether regional grouping (e.g. Midwest, South, West, Northeast) rather than state-level analysis provides a more balanced view.

4. Integrate Additional Contextual Data:

   - Incorporate external data sets, such as median income, unemployment rates, or education levels, or enrich the analysis.

- This could help validate assumptions and explain why certain correlations exist beyond numerical associations.

5. Further Investigate Key Socioeconomic Trends:

   - Deeper analysis on why certain states have higher poverty levels, non-marital birth rates, or rental occupancy challenges could yield meaningful insights for policymakers.
   - Conduct a time-series analysis (if data is available) to understand how these trends have evolved over time.

Link to GitHub Repository

https://github.com/LuckyJH2024/5243_project1-Team4/tree/main

Each Member's Contribution

Jinze Shi conducted a comprehensive data preprocessing pipeline for the Communities and Crime dataset. First, he verified and corrected data types, ensuring categorical and numerical features were properly identified. To handle missing values, he applied K-Nearest Neighbors (KNN) imputation for categorical features and mean imputation for numerical features to maintain data integrity. Next, he 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, he 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. He then also wrote the Data Acquisition Methodology and Cleaning & Preprocessing Steps portions of the report.

Xinyi Gui contributed by doing EDA - Part 1 which consisted of conducting the summary statistics review and creating the histograms, KDE plots, and box plots for all numeric columns. She identified skewness, outliers, and potential anomalies for her part, establishing a foundation for more advanced correlation and segmentation analyses for Part 2. She then also wrote the EDA - Part 1 portion of the report, and assisted in writing the summary of key findings and challenges/future recommendations related to her portion of the work in the report.

Shreya Prabu contributed by doing EDA - Part 2 which consisted of the correlation matrix, scatter plots of highly correlated features, and bar charts of data distributions across states and mean values of features. She then also wrote the Introduction & Dataset Description portion of the report, the EDA - Part 2 portion of the report, the README file, assisted in writing the summary of key findings and challenges/future recommendations related to her portion of the work in the report, and integrating all parts of the report.

Jiangao Han was primarily responsible for feature engineering including designing and constructing new features to enhance the predictive power of our model. Additionally, he refined the originally selected features by applying various transformation techniques to improve their interpretability and effectiveness in subsequent analysis. He also contributed to code integration and modification.