**Exploratory Data Analysis - Part 2 (by Shreya Prabu)**

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.

**Summary of Key Findings**

1. 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.
2. 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.
3. 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.
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 datasets, 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.

**Shreya Prabu’s Contribution**

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 EDA- Part 2 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.