Insightful analysis of the Boston Housing dataset.

Target Variable Analysis 🎯

The **target_analysis.jpg** image provides a deep dive into the distribution of the target variable, **MEDV** (Median value of owner-occupied homes).

- **Distribution:** The histogram and KDE plot show that the distribution of housing prices is **right-skewed**, meaning there's a long tail of higher-priced homes. Most properties cluster in the **$20k to $45k** range, with the peak around **$35k to $40k**.
- **Normality:** The Q-Q plot confirms this non-normality. The data points deviate from the red diagonal line, especially at the lower and higher ends. This suggests that a simple linear regression model might not perform optimally without a transformation of the target variable (e.g., using a logarithmic scale).
- **Outliers:** The box plot highlights several **outliers** at the lower end of the price spectrum, indicating a few unusually cheap properties. However, a significant number of properties are priced near the maximum value of $50k, which is likely a **capped value** in the dataset rather than a natural outlier, affecting the overall distribution.
- **Categorization:** The pie chart categorizes housing prices, revealing that the **majority of properties are in the "High" price category (46.2%)**, with "Medium-High" being the next largest group (39.5%). This further underscores the right-skewed nature of the data and suggests a market with a concentration of valuable homes.

Feature Distributions 📈

The **distributions.jpg** image visualizes the individual distributions of all 14 features.

- **Skewed Distributions:** Several features, such as **CRIM**, **ZN**, and **DIS**, are highly **right-skewed**, with most values concentrated near zero. This indicates that most areas have very low crime rates, residential land zoning, and distance to employment centers, respectively.
- **Bimodal/Multimodal Distributions:** **RAD** (index of accessibility to radial highways) and **TAX** (full-value property-tax rate) show what appear to be **bimodal or even multimodal distributions**. This might suggest distinct groups or clusters within the data that have different accessibility or tax rate characteristics.
- **Approximately Normal:** **RM** (average number of rooms per dwelling) and **LSTAT** (percentage of lower status population) are more evenly distributed, resembling a more **normal-like curve**, with their peaks near the average values. This suggests a consistent range of values across the dataset for these features.

Correlation Analysis 🤝

The **correlation_heatmap.jpg** shows the correlation matrix between all features and the target variable, **MEDV**.

- **Strongest Correlations with MEDV:** The target variable **MEDV** has the strongest correlations with **LSTAT** and **RM**.
    - **LSTAT** has a **strong negative correlation (-0.74)**. This is a crucial insight, indicating that as the percentage of the lower-status population increases, the median home value tends to decrease significantly.
    - **RM** has a **strong positive correlation (+0.70)**. This is an intuitive finding: as the average number of rooms per dwelling increases, the home value also tends to increase.
- **Other Significant Correlations:**
    - **PTRATIO** and **INDUS** have **moderately negative correlations** with MEDV, suggesting that a higher pupil-teacher ratio and a greater proportion of non-retail business acres are associated with lower property values.
    - **NOX**, **CRIM**, and **TAX** all show **negative correlations** with MEDV, which aligns with expectations—higher pollution, crime, and tax rates are generally associated with lower property values.
- **Feature-to-Feature Correlations:** There are also strong correlations between some independent variables, such as **NOX** and **INDUS** (+0.76), and **LSTAT** and **PTRATIO** (-0.39). These collinearities should be considered when building a predictive model.

Feature Relationships 📊

The **feature_relationships.jpg** provides a visual scatter plot analysis of the key relationships identified by the heatmap.

- **MEDV vs. LSTAT:** The plot for MEDV vs. LSTAT clearly shows a **strong inverse relationship**. As LSTAT increases, the median home value MEDV decreases, confirming the strong negative correlation seen in the heatmap.
- **MEDV vs. RM:** The MEDV vs. RM plot shows a clear **positive linear relationship**. Homes with more rooms generally have higher values. The cluster of points at the top of the graph (around MEDV = $50k) again highlights the data capping issue.
- **MEDV vs. PTRATIO:** The scatter plot for MEDV vs. PTRATIO shows a **moderate negative relationship**, with home values tending to decrease as the pupil-teacher ratio increases.
- **MEDV vs. INDUS:** The plot for MEDV vs. INDUS confirms the **negative correlation**, though with more spread. Higher proportions of non-retail business land seem to correspond with lower home values.

Outlier Analysis 🔍

The **outlier_boxplots.jpg** image uses box plots to identify outliers for each feature.

- **Features with Outliers:** Several features, including **CRIM**, **ZN**, **B**, and **DIS**, contain a significant number of **outliers**on the high end. This is consistent with their right-skewed distributions.
- **CHAS:** The box plot for CHAS (Charles River dummy variable) shows no outliers, as it is a binary categorical feature with values of 0 and 1.
- **MEDV:** The box plot for **MEDV** clearly shows outliers on both the lower and higher ends. The high-end outliers are the aforementioned data points capped at $50k, while the low-end outliers represent a few exceptionally low-priced properties. These outliers could be significant for a model's performance and may need to be addressed through data cleaning or transformation.