Here's a comprehensive analysis writeup that documents the process and the conclusions we reached during our dataset comparison and potential merging process:

Comprehensive Analysis and Process Summary

In this analysis, we aimed to explore how different housing price datasets could be harmonized and complemented to enhance the overall quality of the data. We compared the cleaned house price dataset with three other datasets: the houseprice dataset, the housing prices dataset, and the USA house prices dataset. Each dataset provided different structures, but common features allowed for potential data merging. Below is the step-by-step process and findings from this analysis:

1. Initial Dataset Comparisons

- **Cleaned Data vs. Houseprice Data**:
    - The cleaned data and the houseprice data share similar columns, such as "Price," "Square Footage (Sqr Ft)," "Beds," and "Baths." However, discrepancies in formatting (e.g., dollar signs, square footage abbreviations) required harmonization before merging.
    - **Conclusion**: These datasets can complement each other once formatting and unit discrepancies are resolved.
- **Housing Prices Dataset**:
    - This dataset contains additional property features such as "stories," "mainroad," "guestroom," "basement," and "furnishing status," which are not present in the cleaned dataset. These additional details provide more comprehensive information on property characteristics, making it a valuable addition to the cleaned dataset.
    - **Conclusion**: The housing prices dataset can enrich the cleaned dataset by adding more detailed property features, potentially matching through other identifiers (such as area or number of bedrooms).
- **USA House Prices Dataset**:
    - This dataset focuses on specific geographic regions (particularly Los Angeles) and includes a date column. The cleaned dataset covers multiple states. Therefore, this dataset could add more region-specific information, especially for temporal analysis.
    - **Conclusion**: This dataset could complement the cleaned data by adding region-specific or time-related insights.

2. Address and Coordinate Matching

- **No Common Addresses**:
    - Upon attempting to match datasets using property addresses, no common entries were found between the housing prices dataset and the houseprice dataset. This led to the conclusion that address matching was not a viable option for combining these datasets.
    - **No Common Coordinates**:
        - Similar attempts were made to match the datasets based on geographic coordinates (longitude and latitude). However, no common coordinates were found either.
    - **Conclusion**: Address or coordinate-based merging is not feasible. Alternative strategies, such as matching based on similar features

(e.g., square footage, bedrooms, bathrooms), must be explored.

3. Feature-Based Matching Attempts

- We adjusted the approach to compare properties based on features such as square footage, number of bedrooms, and bathrooms, allowing for a tolerance in values.
- After iterating through both datasets, we identified potential matches where properties shared similar features. These matches suggested that properties across the datasets could complement one another by filling in missing data or enhancing detail.

4. Visualization of Matched Properties

- **Bedrooms Comparison**: The scatterplot indicated a general alignment between the number of bedrooms in the matched properties across the datasets.
- **Bathrooms Comparison**: While there was some variation in bathroom counts, the matches were consistent enough to consider merging.
- **Area Comparison**: A strong correlation was observed in square footage between the datasets, confirming that area was a solid feature for matching the properties.
- **Conclusion**: The visualizations provided evidence that the datasets share enough similarities in key features to justify a merger.

5. Unmatched Properties

- We also identified properties that did not have matches based on our criteria. These unmatched properties were examined separately and may require further analysis to determine if they can be merged based on different criteria or additional sources.

6. Conclusion and Recommendations for Merging

Based on the strong correlations observed between the datasets, particularly in square footage, we concluded that it is feasible to merge the datasets to enhance the overall data quality. Below are some recommended steps for the merging process:

1. **Match Properties by Key Features**: Given the strong alignment in square footage, bedrooms, and bathrooms, merging properties based on these criteria should yield consistent results.
2. **Fill in Data Gaps**: Use missing information from one dataset (such as longitude, price per square foot) to complement the other dataset, providing a more comprehensive property profile.
3. **Regionally Specific Information**: Since the USA House Prices dataset focuses on specific regions like Los Angeles, separating properties by region could improve the accuracy of the match and provide more detailed regional analysis.
4. **Use Visualizations for Quality Control**: As seen from the scatterplots, visual checks can help ensure that the merged data is consistent across key features like area, bedrooms, and bathrooms.

By following this approach, we can combine these datasets into a unified, enriched dataset that captures the key property features more comprehensively and provides the basis for further detailed analysis.