9. furnishingstatus_semi-furnished

This aligns perfectly with the correlation matrix, suggesting the top features for driving price predictions are:

1. Area
2. Bedrooms
3. Bathrooms
4. Stories
5. Parking

2.7 | Encoding

We need to encode columns that have categorical data in our dataset to perform any dimensionality reduction. Dimensionality reduction techniques (like PCA or t-SNE) expect numerical data.

All of the above top 5 features are numerical columns, thus suggesting that the type of encoding is not such a vital process for our dataset. Target encoding or Frequency encoding do not need to be applied as they would only be applicable for the features outside of these top 5 features. One hot encoding should suffice.

In addition to this, all other categorical features (like 'guestroom, 'mainroad', 'basement' etc) have low cardinality (such as yes or no), meaning one hot encoding will not cause an overwhelming amount of feature additions (sparse encoding) and computational restrictions. Going beyond one hot encoding for this dataset would add complexity without significant improvement.
We applied one-hot encoding to the categorical columns on the dataset.

3 | Dimension Reduction

Dimensionality reduction is an incomparably significant stage of data analysis and modeling and has several benefits. First, it aids in decreasing the expenses that would be involved in computations especially when working with a dataset that is of high dimension. Second, it allows the structures of the data to be better represented through visually appealing and easy-to-interpret 2D graphics which are important when trying to analyze relations and patterns within the data set. Lastly, it helps to expand feature visualization which is important for subsequent modeling tasks.

We considered the following graphs the results of some dimensionality reduction methods later applied to the Housing Prices Dataset after preprocessing, feature selection, and encoding. Also visualized with all features and the Top 5 features.

- PCA of the Housing Prices Dataset
- SVD of the Housing Prices Dataset
- t-SNE of the Housing Prices Dataset
- UMAP of the Housing Prices Dataset
- ISOMAP of the Housing Prices Dataset
- LLE of the Housing Prices Dataset

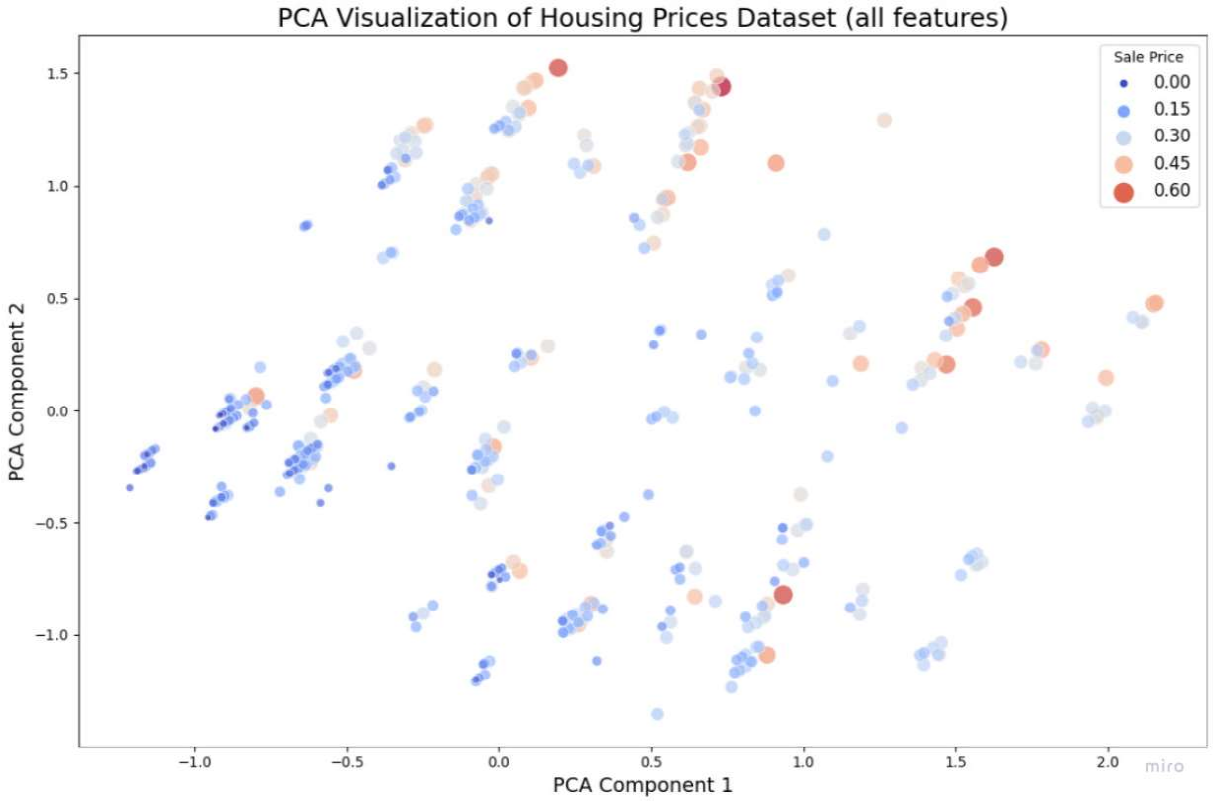


Figure 3.1: PCA Analysis

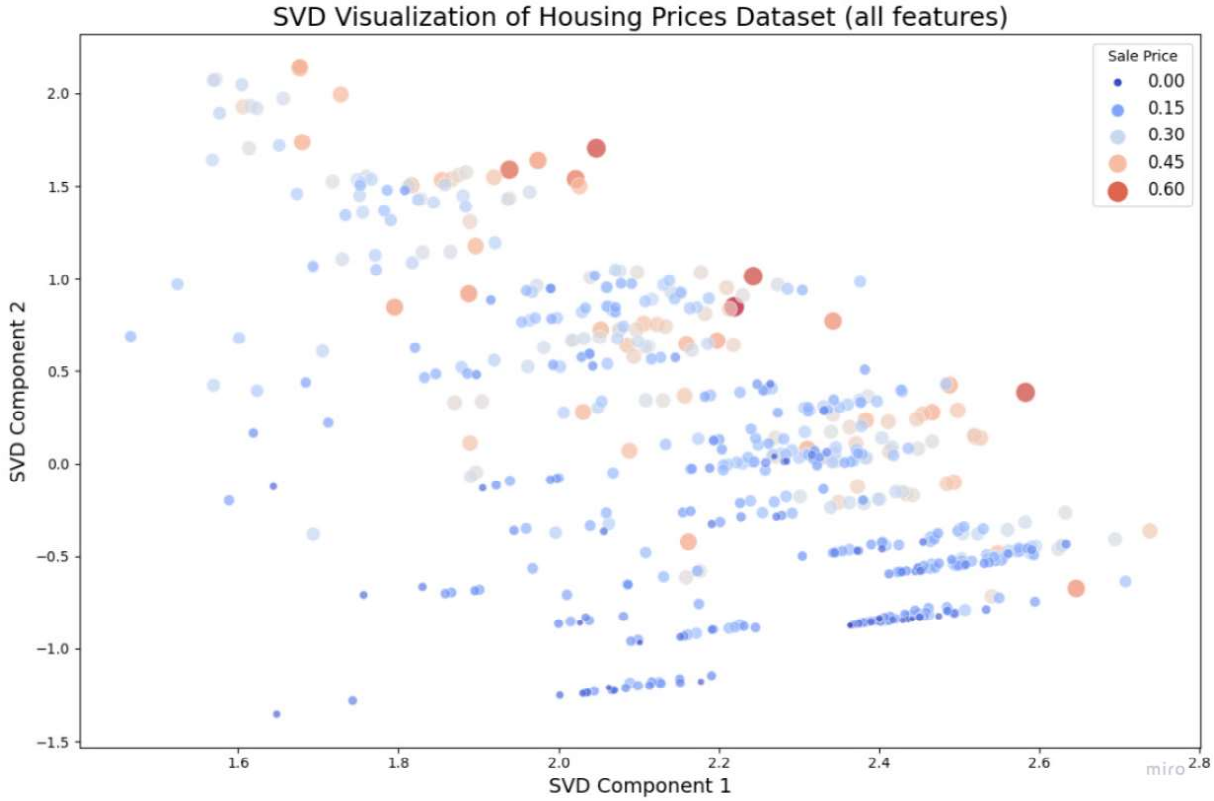
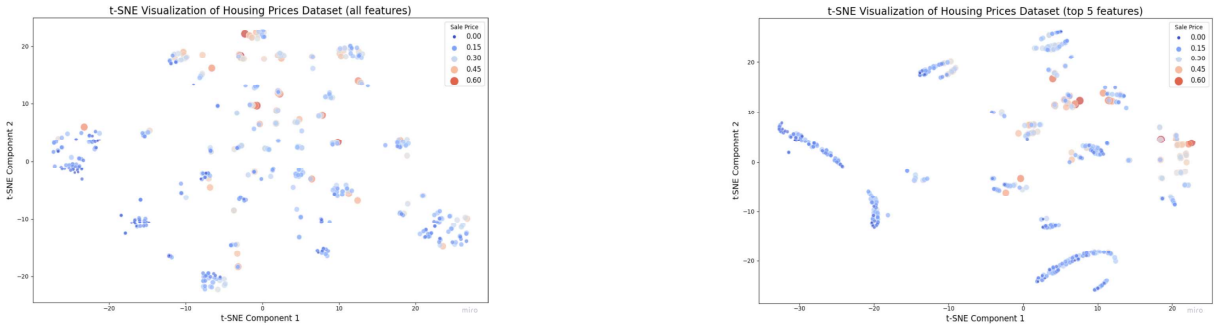
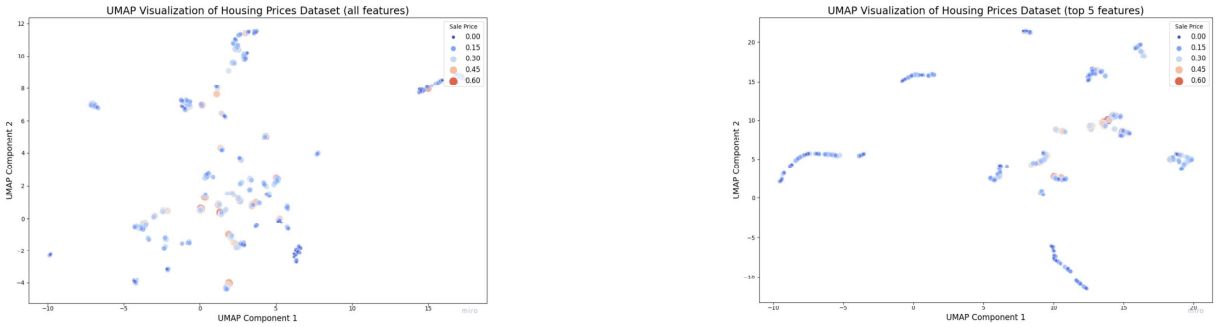


Figure 3.2: SVD Analysis



(a) All Features **(b)** Top 5 Features

Figure 3.3: t-SNE Analysis



(a) All Features **(b)** Top 5 Features

Figure 3.4: UMAP Analysis

3.1 | Analysis of Dimensionality Reduction Techniques

3.1.1 | Why Nonlinear (tSNE and UMAP) compared to Linear (PCA and SVD)?

Data appears to be correlated in a non-linear fashion. This can be noticed because PCA and SVD show less succinct clusters and are more spread across the graph. tSNE and UMAP on the other hand show more distinct sections, which indicate different interwoven relationships.

3.1.2 | Why does UMAP (all features) show us more than tSNE (all features)?

UMAP clusters the data with a clearer representation of both local and global structures. tSNE rather focuses more on the insights for local structures. UMAP preserves global relationships better than t-SNE and therefore shows how clusters are related across the dataset. UMAP will capture the broader trend and faster computational time compared to t-SNE for larger datasets.

tSNE shows dense pockets of similar homes, emphasizing small-scale variations between observations. This also has its benefits for deeper insights into the tighter clusters. However, clusters seem more scattered, making it harder to identify clear transitions between sale price ranges.

UMAP shows more distinct regions for low and high prices, tSNE is more scattered and harder to interpret. In addition, UMAP's clusters suggest stronger interactions between numerical and categorical features (e.g., area and furnishing status) compared to t-SNE, where clusters are more fragmented.

- Pros of top features
  - Give us more nuanced shaping and groups that were might otherwise miss
  - Refined patterns and detailed regions

- □ Including all features provides a richer understanding but increases complexity.
- □ For models that thrive on interactions between numerical and categorical data (e.g., decision trees, gradient boosting), the lack of categorical features might reduce predictive power.

- Pros of top 5 features
  - □ t-SNE and UMAP are now easier to interpret with fewer features.
  - □ Demonstrate the emphasis and impact of the top 5 features
  - □ These features are the most impactful, reducing noise and computational complexity.

3.1.3 | Why 'all features' are better than 'top 5' at this stage in the process?

While it is insightful to see the differences between performing dimensionality reduction on all features and only on the top 5 features, all features will be considered the most important aspect at this time. Perhaps the top 5 features might prove useful later upon hyperparameter tuning and creating model variations to compare for computational efficiency, however on the onset all features should perform sufficiently with nonlinear algorithms. Furthermore, the insights we gain from the global understanding are more impactful at this early stage, whereas perhaps later on the local subclusters could be more insightful.

3.1.4 | What does UMAP tell us?

While it is insightful to see the differences between performing dimensionality reduction on all features and only on the top 5 features, all features will be considered the most important aspect at this time. Perhaps the top 5 features might prove useful later upon hyperparameter tuning and creating model variations to compare for computational efficiency, however on the onset all features should perform sufficiently with nonlinear algorithms. Furthermore, the insights we gain from the global understanding are more impactful at this early stage, whereas perhaps later on the local subclusters could be more insightful.
Global overview of data. Key for understanding our whole dataset: its spread and interconnected features.

- Higher-priced homes form distinct clusters, possibly reflecting premium features, like larger house areas and extra amenities.
- Including only categorical features gave us more distinct clusters, suggesting some noise from other features.

The findings most importantly tell us that the data we are dealing with has more nonlinear relationships compared to linear relationships. This means we should use nonlinear algorithms and not linear ones for best results:
Suggested algorithms for models:
Nonlinear algorithms that would work well on the data:

- Decision Trees
- Random Forest
- Gradient Boosting
- Neural Networks
- K-nearest neighbors

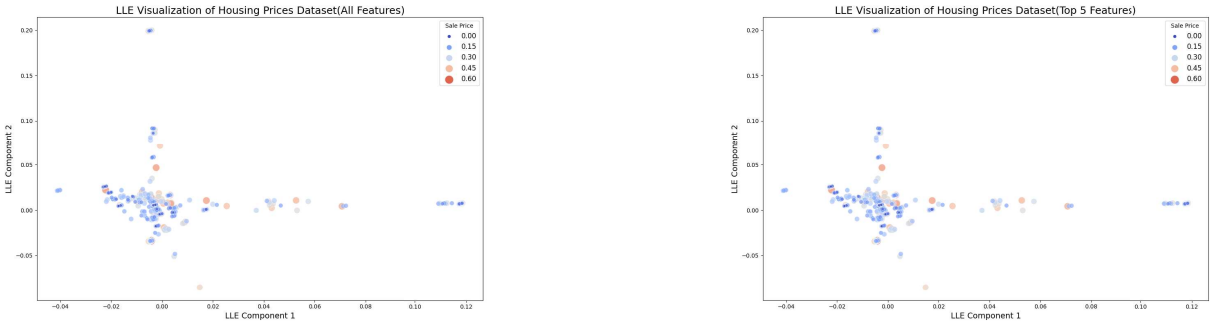
Linear algorithms that will not work so well on this data:

- Linear regression
- Logistic regression

3.1.5 | Manifold Learning

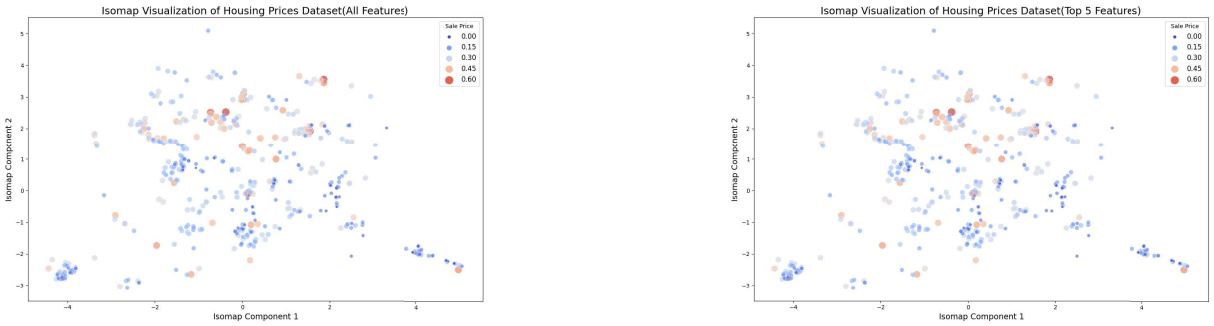
Manifold learning attempts to generalize PCA to perform dimensionality reduction on all sorts of dataset structures, with the main idea that manifolds, or curved, continuous surfaces, should be modeled by preserving and prioritizing local over global distance.[6]

- **ISOMAP** tries to preserve geodesic distance, or the distance measured not in Euclidean space but on the curved surface of the manifold.
- **Locally Linear Embedding(LLE)** can be thought of as representing the manifold as several linear patches, in which PCA is performed.



(a) All Features **(b)** Top 5 Features

Figure 3.5: Locally Linear Embedding (LLE) Analysis



(a) All Features **(b)** Top 5 Features

Figure 3.6: ISOMAP Analysis

4 | Data Ethics and Responsible AI

There are several ethical considerations when considering using house price data samples for analytical and research purposes, namely privacy, fairness, biases,and the potential impact on vulnerable groups.

4.1 | Privacy and confidentiality

Many house price data samples are generally made publicly available and often contain personally identifiable information (PII), either explicitly or implicitly. For example, names, addresses, or any other sort of personally identifiable information...etc. While examining the house price data set in hand, we noted that it didn't include any such information, implying that this data set is not subject to any data privacy regulations similar to GDPR for example. That being said, it imposes certain limitations in terms of how and what the data would be used for, and potential challenges if it is meant to produce accurate house price predictions, ie. missing the house addresses which could have a significant impact on the house prices.