Analysis of Housing Price Data and Model Implementation

### Housing Price Prediction: Data Analysis and Model Implementation

#### Objective

The objective of this analysis was to clean, preprocess, and analyze housing price data to develop machine learning models capable of predicting property prices accurately. Several regression models were evaluated to determine the best-performing approach for price prediction.

Key Critical Points Addressed

1. What is the distribution of housing prices across different states?
   * Pennsylvania and New York had the highest representation in the dataset.
   * The District of Columbia had the highest median house prices.
2. Are there missing values or duplicates in the dataset?
   * All missing values were removed, and no duplicates were found.
3. How does the dataset change after preprocessing?
   * The original dataset had 250k+ rows, which was reduced to 24,362 rows after filtering for 2018-2023.
   * The number of features increased to 4,227 columns after encoding categorical variables.
4. Which regression model performs best for predicting house prices?
   * Gradient Boosting Regression performed best with an R² score of 0.7411 and the lowest Mean Squared Error (MSE).
5. How does feature scaling impact model performance?
   * Standardizing the features improved model performance by ensuring numerical stability.

Data Cleaning and Preprocessing

* Missing Values: Rows with NaN values were removed, resulting in 86,633 rows.
* Duplicate Rows: No duplicates were found in the dataset.
* Column Dropping: The status column was removed as it was not relevant to predictive modeling.
* Datetime Conversion: The previously sold date column was converted to datetime format.
* Data Filtering: The dataset was filtered to include only records from 2018-2023, reducing it to 24,362 rows.
* Feature Engineering: Categorical columns (city, state) were converted into dummy variables, increasing the dataset to 4,227 columns.
* Missing Values Check: After preprocessing, no missing values remained.

Exploratory Data Analysis (EDA)

State Distribution:

- Pennsylvania (22.6%) and New York (21.8%) had the highest representation in the dataset. NJ & ML the 4th most populated accounts for 23 % of the sales. ​

- Most Populated Estates have also the largest numbers of transactions​

​- Tri-state (NY, PA, VI ) which are the most populated, accounts for 62 % of the volume of sales​

- By City the top 10 cities with the largest listing of properties are Philly, Agawarn, Baltimore, Charlotesville and Richmond

Supply:

- Per price, the highest amount of houses listed are in the range of $200K to $400K

- Supply drops almost to half for the next bucket $400K –$600K and then maintains for lower than $200K​

- $1M -$2M supply is higher than $800k - $1M​

​Price Analysis:

- The District of Columbia holds the highest average price of the set, being significantly higher than Massachusetts 33% ​

​Pennsylvania holds the lowest average price of the NE​

- Price per state has a high dispersion, especially in New York, Connecticut, New Jersey, Pennsylvania, and Virginia​

- Most of the prices range below $1MM. However, the District of Columbia sets prices way above the rest of the states.

​Size and Price

- Average price seems to have a direct relationship with the size of the property

However, prices vary widely, especially in 1,000-3,000 sf buckets; in these buckets, prices can be as high as the average price of 6,000 sf. ​

- Prices can be impacted by area, but there definitely can be other variables impacting them ​.

Model Implementation and Performance Evaluation

1. Data Preparation
   * The feature matrix (X) was created by dropping price and prev sold date.
   * The target variable (y) was set as price.
   * The dataset was split into 80% training and 20% testing.
   * Features were standardized using StandardScaler.
2. Regression Models
   * Decision Tree Regression
     + MSE: 438,165,276,568.75
     + R² Score: 0.3686
   * K-Nearest Neighbors (KNN) Regression
     + MSE: 256,608,374,796.27
     + R² Score: 0.6303
   * Gradient Boosting Regression
     + MSE: 179,675,782,813.27
     + R² Score: 0.7411
   * Stacking Regression (combining Decision Tree, KNN, and Gradient Boosting)
     + Process interrupted before completing evaluation.

Findings and Conclusion

* Gradient Boosting Regression performed the best, achieving the highest R² score (0.7411) and the lowest Mean Squared Error.
* Decision Tree Regression had the weakest performance with the lowest R² score (0.3686).
* KNN Regression performed better than the Decision Tree model but was outperformed by Gradient Boosting.
* The stacking model was interrupted, but it aimed to combine multiple regressors for better predictive performance.
* The final dataset was saved for further optimization.

Limitations

1. Computational Complexity:
   * The dataset grew significantly after one-hot encoding, increasing processing time and resource requirements.
2. Feature Selection:
   * Some features may not contribute meaningfully to price prediction and could be further analyzed.
3. Model Performance:
   * While Gradient Boosting performed best, the R² score of 0.7411 indicates room for improvement.
4. Incomplete Stacking Model:
   * The stacking approach could not be fully executed, limiting insights into potential model improvements.

Next Steps

* Hyperparameter tuning for Gradient Boosting to further improve performance.
* Re-attempt stacking model execution to evaluate its predictive power.
* Feature selection optimization to reduce dimensionality and computational burden.
* Explore additional factors like economic indicators or location-based features for better modeling.

This analysis provides a solid foundation for house price prediction, demonstrating strong performance with Gradient Boosting and highlighting opportunities for further optimization.