**Exploratory Data Analysis Report: House Prices Dataset**

* **Sahiti Lakkoju (21WU0101028)**

**Dataset Overview**

The analysis focuses on the **House Prices Dataset** obtained from Kaggle. This dataset contains various features influencing house prices, such as location, size, number of rooms, and distances to nearby amenities.

**Objectives**

1. Understand the data structure and identify key patterns.
2. Clean the data to handle missing or inconsistent values.
3. Explore feature distributions and relationships.
4. Generate new features and remove outliers for better modeling.

**Data Cleaning**

* **Handling Missing Values**:
  + Imputed missing values for garage with the mean value.
  + Replaced null values in build\_year with its mean.
  + Dropped columns with excessive null values, such as nearest\_sch\_rank and postcode.
* **Column Renaming**:
  + Converted column names to lowercase for consistency.

**Data Exploration**

* **Statistical Summary**:
  + Extracted summary statistics (mean, median, standard deviation).
  + Assessed distributions for numerical and categorical features.
* **Null Value Analysis**:
  + Quantified missing data and addressed it systematically.
* **Visualizations**:
  + Scatter plots showed relationships among key numerical features.
  + Distribution plots highlighted skewness in variables like price and lot\_size.

**Feature Engineering**

* Converted all distance-related columns to a uniform unit (kilometers).
* Extracted temporal data from the date\_sold column.
* Generated interaction terms to capture relationships between existing features.

**Outlier Detection and Removal**

* Visualized scatter plots to identify extreme values.
* Removed outliers based on statistical thresholds.
* Replotted distributions post-removal to confirm improvements.

**Observations**

1. **Feature Relationships**:
   * Strong positive correlation between lot\_size and price.
   * Moderate negative correlation between distance\_from\_city and price.
2. **Improvements Post-Cleaning**:
   * Reduced variance in key features after imputing missing values.
   * Enhanced clarity of feature relationships through engineered columns.

**Conclusion**

The dataset is now prepared for advanced modeling. Data cleaning and feature engineering have addressed inconsistencies and improved feature relevance. Future steps include applying machine learning models to predict house prices accurately.