Project Summary – Raw House Data Analysis

The **Raw House Data** dataset consists of **5,000 records** related to real estate transactions, containing various attributes such as **sold price, location (latitude & longitude), lot size, property taxes, year built, number of bedrooms and bathrooms, kitchen features, garage details, fireplaces, flooring type, and HOA fees**. The primary goal of this **Exploratory Data Analysis (EDA)** was to clean, preprocess, and analyze the dataset to uncover meaningful insights.

The project was executed in **Google Colab**, leveraging **Python libraries such as Pandas, NumPy, Matplotlib, and Seaborn** to handle missing values, perform type conversions, create new features, and visualize key trends. The data underwent extensive preprocessing, including treating missing values, transforming categorical variables, and identifying outliers. **Various statistical techniques** were applied to understand relationships among different features, particularly focusing on how various factors influence sold_price.

Key visualizations such as **histograms, box plots, pair plots, and correlation matrices** were used to detect trends, distributions, and anomalies. The **final cleaned dataset** was saved in CSV format, ensuring it's ready for further modeling and analysis. Additionally, the **notebook was converted to HTML and text format** for documentation and sharing.

This analysis provides valuable insights into **real estate market trends**, helping stakeholders understand property price variations and key influencing factors.

Key Actions and Technologies Used

- **Google Colab** as the cloud-based development platform for seamless execution.
- **Pandas & NumPy** for data pre-processing, manipulation, and cleaning.
- **Handled missing values** using mean, median, and mode based on data characteristics.
- **Converted data types**, including removing commas from HOA and converting it into numerical format.
- **Feature engineering**: Created num_kitchen_features by counting distinct kitchen attributes.
- **Descriptive statistics** to understand data distributions and anomalies.
- **Seaborn & Matplotlib** for **data visualization**:
- Histograms for **distribution of sold prices**.
- Box plots for **outlier detection** in sold_price, taxes, and lot_acres.
- Pair plots to analyze **relationships among numerical features**.

- Correlation heatmap to explore **strong and weak feature relationships**.
  - **Outlier detection & treatment** using box plots and statistical methods.
  - **Exported the cleaned dataset** to CSV for further analysis.
  - **Converted the notebook to HTML & text formats** using nbconvert.