STAT 5243 Project 1

Jiaheng Zhang (jz3855), Yi Lu (yl5735)
Shenghong Wu (sw3962), Mingyan Xu (mx2294)

2025-02-12

1. Introduction

In this project, we focus on the comprehensive data pipeline, including data acquisition, cleaning, exploratory data analysis (EDA), and feature engineering to prepare real-world datasets for meaningful analysis. The dataset will be sourced from public repositories, ensuring accessibility and complexity. A systematic approach will be applied to address inconsistencies, missing values, and formatting issues, followed by an in-depth exploratory analysis to identify key patterns and trends. Furthermore, feature engineering techniques will be applied to enhance the dataset's usability for potential predictive modeling. The final deliverable includes a structured report detailing the methodologies, insights, and challenges encountered throughout the data preparation process.

2. Data Acquisition Methodology

The core dataset for this project is sourced from the Flats Uncleaned Dataset on Kaggle. This dataset includes various features, including price, area, floor and location, offering multiple dimensions for analysis. While the data originates from a single public repository, we particular pay attention to its quality and complexity. The dataset presents substantial data quality challenges such as missing values, inconsistent labels, and formatting discrepancies, reflecting the complexities of real-world data acquisition. The acquisition phase emphasizes data completeness, potential business relevance, and automated data loading with preliminary checks, laying the groundwork for subsequent cleaning and analysis.

The dataset before cleaning consists of **4,525 rows and 11 columns**, containing real estate property listings with various attributes. The table below provides an overview of each column:

Column Name	Description
property_name	The name or identifier of the property listing.
areaWithType	The name of the locality along with its type (e.g., district, suburb, or city).
square_feet	The total area of the property in square feet.
transaction	Type of transaction (e.g., resale, new sale, rental).
status	The current availability status of the property (e.g., available, sold, under construction).
floor	The floor number or range where the property is located.
furnishing	The furnishing status of the property (e.g., furnished, semi-furnished, unfurnished).
facing	The direction the property faces (e.g., North, South, East, West).
description	Additional textual details about the property.
price_per_sqft	The price per square foot of the property.
price	The total listing price of the property.

3. Data Cleaning and Handling Inconsistencies

The raw dataset exhibits multiple uncleaned issues: missing values, duplicate entries, outliers (e.g., unrealistic prices or areas), inconsistent categorical labels (e.g., "2 BHK" vs. "2 Bedroom"), and incorrect data types (e.g., numeric values stored as strings). This phase systematically addresses these challenges: identifying missing values and outliers through descriptive statistics; applying imputation or removal strategies for missing data; standardizing categorical formats using regular expressions; and filtering illogical values via quantile-based thresholds or domain logic. The goal is to produce a structurally consistent, reliable dataset for downstream analysis.

3.1 Deleting Duplicates

To ensure data integrity, duplicate entries (109 entries) were identified and removed, resulting in **4,416 rows and 11 columns.**

3.2 Standardizing Units and Converting Data Types

The dataset contained inconsistencies in unit representations, requiring standardization.

3.2.1 Standardizing Area Units

The `square_feet` column had multiple units (`sqm`, `sqyrd`, `acre`, `sqft`), leading to inconsistencies in area representation. The following steps were performed:

- Extracted the unit suffix from `square_feet` and identified invalid or unexpected units and replaced them with `NaN`.
- Converted all area measurements to **square feet (sqft)**: **1 sqm → 10.7639 sqft, 1 sqyrd → 9 sqft, 1 acre → 43,560 sqft**
- Renamed the column to `area_sqft` for clarity.

3.2.2 Standardizing Price and Price Per Square Foot

The `price` and `price_per_sqft` columns contained variations in format, including Indian Rupee symbols and unit abbreviations (`Lac`, `Cr`). The following transformations were applied:

- Removed the currency symbol and extracted numerical values.
- Standardized price units: **1 Cr → 100 Lac** (converted to a common unit in Lac).
- Converted `price_per_sqft` to numeric values and ensured unit consistency (`per sqft`).

3.3 Cleaning Categorical Variables

Several categorical variables contained inconsistent values, misplaced information, and missing entries. The following preprocessing steps were applied:

- **`areaWithType`:** Retained Carpet Area, Super Area, Plot Area, Built Area, replacing other values with NaN for consistency.
- **`transaction`:** Identified misplaced values and kept only New Property, Resale, with a backup copy (transaction_copy) for reference.
- **`status`:** Standardized "Possession by date" entries as "Not Ready", keeping only Ready to Move, Not Ready categories.
- **`floor`:** Converted mixed-format descriptions (Ground → 0, Upper Basement → -1, Lower Basement → -2), extracted property and total floors, and applied the American numbering method (counting from 1 instead of 0).
- **`furnishing`:** Retained Unfurnished, Semi-Furnished, Furnished, replacing other values with NaN.
- **`facing`:** Standardized direction labels (North, South, West, etc.) and replaced unrecognized values with NaN.

3.4 Handling Missing Values

After resolving data inconsistencies and formatting issues, we handled missing values using **statistical imputation**, where categorical variables were filled with the **mode** and numerical variables with the **median**. This approach ensures data completeness while preserving the underlying distribution and minimizing bias.

3.4.1 Filling Missing Categorical Values

For categorical variables, missing values were imputed using the **mode** (most frequently occurring value), as this method preserves the most common category while maintaining the distribution of the dataset. The following categorical columns were processed:

- `transaction`: Filled with the most frequent transaction type.
- `furnishing`: Replaced missing values with the most common furnishing status.
- `areaWithType`: Imputed based on the most frequent area type.
- `facing`: Missing orientations were assigned the most commonly observed value.
- `status`: Standardized using the most frequent property status.

This approach ensures that missing categorical values are replaced with the most representative option while preventing distortions in the dataset.

3.4.2 Filling Missing Numerical Values

For numerical variables, missing values were replaced with the **median** instead of the mean to avoid the influence of extreme outliers. The following numerical columns were processed:

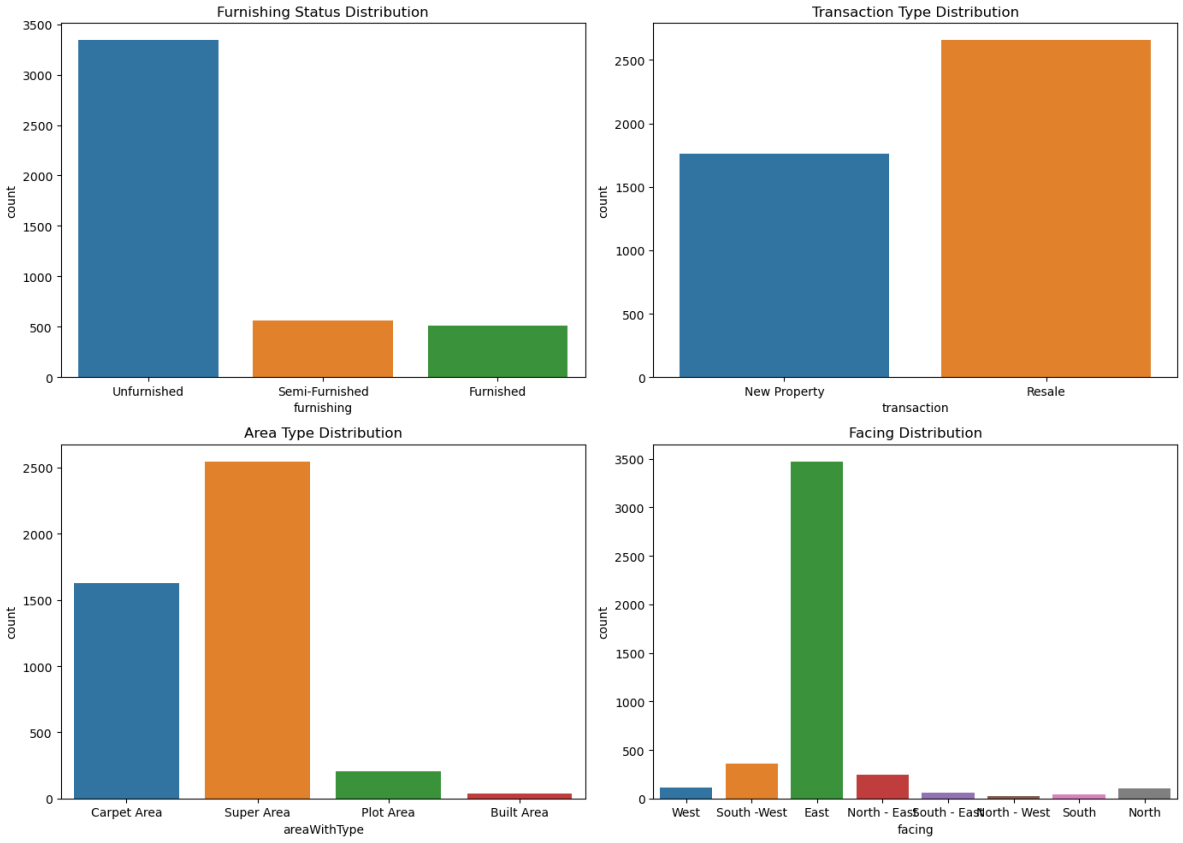
- `area_sqft`: Imputed using the median property area to maintain consistency.
- `property_floor`: Rounded median values were assigned to missing entries.
- `total_floor`: Median-based imputation was applied to preserve structural integrity.
- `price_per_sqft`: Imputed using the median value to align with market pricing trends.
- `price_Lac`: Median-based imputation was performed for missing property prices.

4. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to understand the structure, distribution, and relationships within the dataset. Various visualizations were utilized to examine **numerical and categorical variables**, detect **outliers**, and analyze **correlations**.

4.1 Distribution of Categorical Variables

The four bar charts illustrate the distribution of key categorical variables: **Furnishing Status, Transaction Type, Area Type, and Facing Direction**.



- **Furnishing Status:** The majority of properties are **unfurnished**, while semi-furnished and fully furnished properties are significantly less common. This suggests that most buyers prefer to furnish properties themselves.
- **Transaction Type: Resale properties** dominate the dataset, indicating a more active secondary market compared to new property sales.
- **Area Type: Super Area** is the most frequently reported measurement, followed by **Carpet Area**, while Plot Area and Built Area account for a smaller portion of the listings.
- **Facing Direction: East-facing properties are the most common**, possibly due to cultural or environmental preferences, while other directions like South and West are less frequently listed.

These distributions provide valuable insights into **buyer preferences and market trends**, helping to understand the composition of the dataset.

4.2 Price per Square Foot Across Transaction Types

This boxplot compares the `price_per_sqft` between **new properties and resale properties**. The **x-axis** represents transaction types, while the **y-axis (log-scaled)** shows price per square foot, helping to visualize differences despite large value variations.

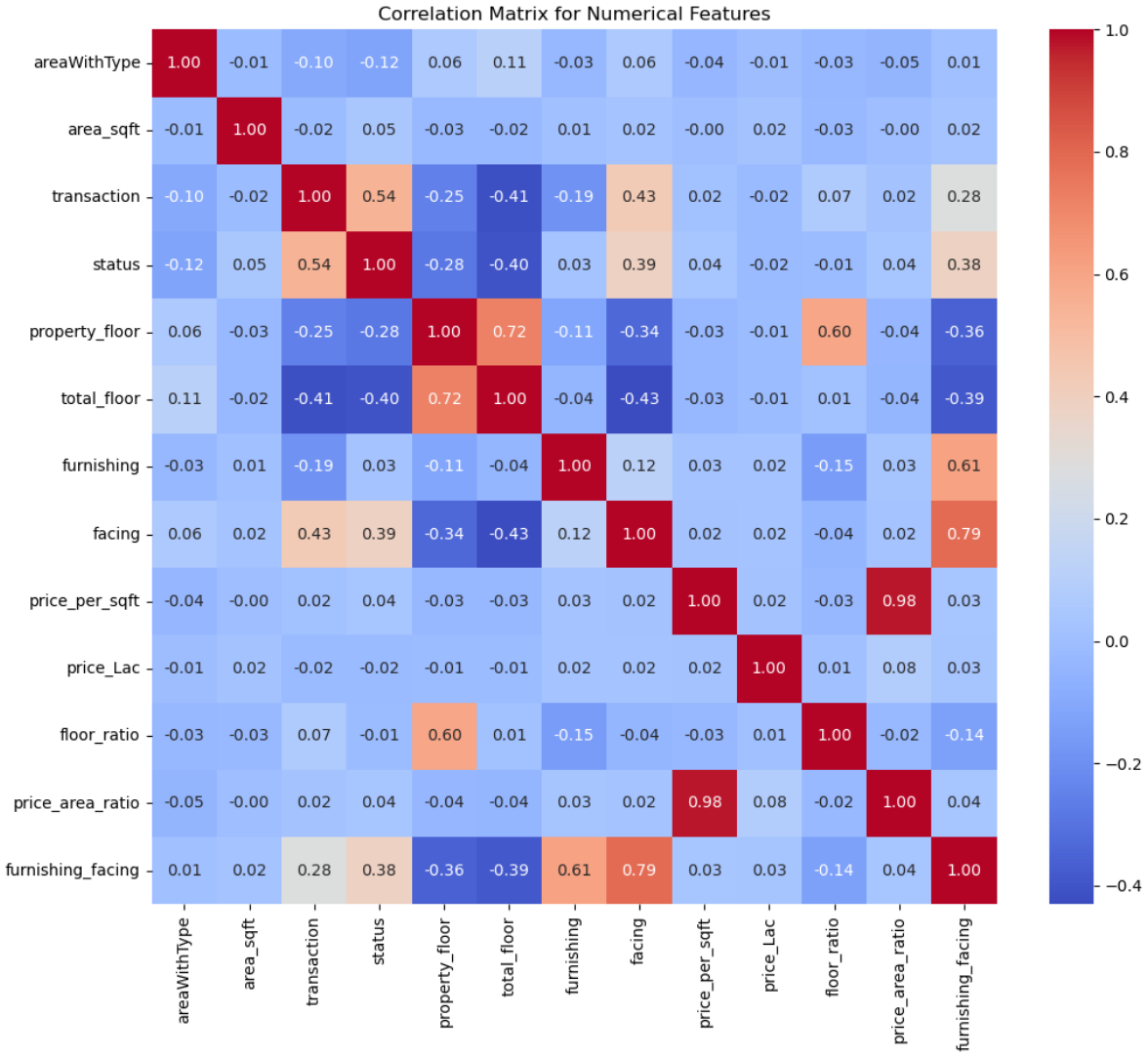
Both categories exhibit a **similar median price**, but resale properties display **greater variability and more extreme outliers**, suggesting a wider range of pricing in the resale market. The log scale ensures that properties with exceptionally high or low prices remain visible for comparison.

4.3 Price per Square Foot Across Transaction Types (Violin Plot)

This violin plot compares **price per square foot** across different transaction types, using a **log scale** to handle extreme price variations. Resale properties exhibit a **wider spread**, indicating greater price variability compared to new properties. The plot also highlights **outliers**, suggesting that some properties are significantly more expensive per square foot.

4.4 Correlation Analysis

The heatmap visualizes the **correlation between numerical features**, where darker red indicates **strong positive correlations**, and darker blue represents **strong negative correlations**.



- Strong correlation: **property_floor and total_floor (0.72)**, **price_per_sqft and price_area_ratio (0.98)**. However, the correlation between property_floor and total_floor is not particularly meaningful, as it simply reflects the structural relationship between a building's floor count and individual unit floors rather than providing any analytical insight. **facing and furnishing_facing (0.79)** indicate that a property's orientation may influence its furnishing style, possibly due to regional or market preferences.
- Additionally, **transaction type and total_floor (-0.41)** implies that taller buildings tend to have different transaction patterns, which could be influenced by market trends or buyer preferences.

5. Data Preprocessing and Feature Engineering

5.1 K-Nearest Neighbors (KNN) Imputation for Enhanced Accuracy

To further refine missing value imputation, a **K-Nearest Neighbors (KNN) imputer** was applied to numerical and categorical variables after encoding categorical values into numeric representations. The following steps were undertaken:

- **Label Encoding**: Categorical columns (`transaction`, `furnishing`, `areaWithType`, `facing`, `status`) were transformed into numerical form using **Label Encoding**.
- **KNN Imputation**: A **KNN Imputer with k=2** was applied, leveraging nearest neighbors to predict missing values based on similar properties.
- **Restoration of Original Categorical Labels**: Once imputation was complete, numerical values were mapped back to their categorical labels using the stored encodings.

Final Dataset Reconstruction: The imputed dataset was reconstructed with original non-imputed columns (`property_name`, `description`) reintroduced for completeness.

5.2 Feature Engineering

In this part, we use several feature engineering techniques, including ratio calculations, text vectorization, and data normalization, to enhance the dataset and improve model performance.

5.2.1 Ratio Feature

To provide deeper insights into property characteristics, the following derived features were created:

- **Floor Ratio (`floor_ratio`)**: Represents a property's relative position within the building.

$$floor\_ratio = \frac{property\_floor}{total\_floor}$$

  This helps distinguish **lower vs. higher-floor properties**.

- **Price per Area Ratio (`price_area_ratio`)**: Indicates the total cost per unit area.

$$price\_area\_ratio = \frac{price\_Lac}{area\_sqft}$$

  Useful for **analyzing pricing trends** based on property size.

- **Furnishing-Facing Interaction (`furnishing_facing`)**: Captures the combined effect of furnishing status and property orientation.

$$furnishing\_facing = furnishing \times facing$$

This may highlight **how furnishing and direction together impact valuation**.

5.2.2 Text Feature Extraction with TF-IDF

We want to convert property descriptions into numerical features, allowing machine learning models to utilize textual information. By applying **TF-IDF (Term Frequency-Inverse Document Frequency)** , we extract key terms from property descriptions, such as "luxury" or "pool," and combine them with numerical features (e.g., area, price) to enhance model predictions. This minimizes information loss, makes the dataset more comprehensive, and enables the model to learn property characteristics more effectively, improving prediction accuracy.

- The TF-IDF matrix was created using **unigrams and bigrams** (`ngram_range=(1,2)`) with a **maximum of 200 features** to retain the most informative terms.
- The transformed TF-IDF features were concatenated with the original dataset, ensuring that textual information contributes to predictive modeling.

5.2.3 Data Normalization and Standardization

To ensure numerical stability and comparability across features, different scaling techniques were applied:

- **MinMax Scaling (0-1 range)** was applied to `area_sqft` and `price_per_sqft` to preserve relative **differences**while constraining values within a fixed range.
- **Standardization (Mean = 0, Standard Deviation = 1)** was applied to `price_Lac` and `total_floor` to **normalize** distributions and **handle varying scales.**

5.2.4 Outlier Detection and Removal

To remove extreme values that could skew model training, we applied two methods:

- **Z-Score Method**: Removed properties with `price_per_sqft` values **more than 3 standard deviations away** from the mean.
- **Interquartile Range (IQR) Method**: Identified and removed outliers in `price_Lac`, ensuring **better data distribution balance**.

6. Summary of Key Findings

This project involved **data cleaning, exploratory data analysis (EDA), and feature engineering** to prepare a real estate dataset for further analysis. The key findings are as follows:

6.1 Data Cleaning and Preprocessing

- **Standardized** area and price units, converting all area values to square feet and price values to Lacs.
- Removed 109 duplicate records and handled missing values using **mode imputation (categorical)** and **median imputation (numerical)**.
- **Outliers in `price_per_sqft`** were detected using **Z-score and IQR methods**, and extreme values were removed.

6.2 Dataset Composition and Distribution

- Most properties are resale rather than newly developed, indicating a more **active secondary market**.
- **Unfurnished properties dominate** (~60%), with fewer fully furnished listings, suggesting buyers prefer to furnish properties themselves.
- Super Area (57.7%) is the **most common area type**, meaning most listings include shared spaces in their measurements.
- **East-facing** properties are the most common, possibly due to cultural preferences or sunlight exposure benefits.
- Price and area distributions are **highly skewed**, with a few extreme values affecting the dataset.
- Strong correlation (0.98) between `price_per_sqft` and `price_area_ratio`, making one feature **redundant for modeling**.
- Facing direction and furnishing status (0.79 correlation) suggest that certain orientations are preferred for furnished homes.
- Weak correlation between `area_sqft` and `price_Lac`, indicating that location and other factors have a greater impact on price than size alone.

6.3 Feature Engineering

- **Created derived features**:
  - **Floor Ratio (`floor_ratio`)**: Indicates a unit's position within a building.
  - **Price per Area Ratio (`price_area_ratio`)**: Normalizes price across different property sizes.
  - **Furnishing-Facing Interaction (`furnishing_facing`)**: Captures combined effects of furnishing and property orientation.
- **Applied TF-IDF vectorization** to extract key terms from property descriptions, making text-based information usable in modeling.
- **Standardized numerical features** to ensure comparability across different scales.

7. Challenges and Recommendations

Challenges

- The data is very dirty, and we don't know which parts of the data are problematic. Some problem data may not be identified, thereby destroying the consistency of the data.
- Selecting the most effective charts to represent price distributions and transaction trends was challenging. Some visualizations (like violin plots) were harder to interpret for non-technical audience.
- This dataset contains many extreme values in price and area created difficulties in visualization and modeling.
- Although feature analysis of the varaibles in the table can be utilized to help predict the listing price for the property like total area, floor, furnishing and so on, some other factors may also play a great role in determining the price like CPI, policy, resident purchasing power.

Recommendations:

- Before processing the data, we can first visualize the data to help us better discover problematic parts of the data. For the KNN part, we arbitrarily set K=2. In the future, we can try more values of K to better fill in the missing data.
- After visualization, we can check the graph to do the cleaning again so that we can remove or modify the extreme values or outliers.
- Use boxplots for price comparisons (helps show outliers clearly).Use scatterplots with transparency (alpha blending) to improve readability. Use annotations to highlight key trends or anomalies in the data.

- In further study, more variables can be imported to help build a better prediction model. We can also set a index for some suppotive or negative policies.

8. Group Distribution

Yi Lu: Data Cleaning and Handling Inconsistencies

Shenghong Wu: Exploratory Data Analysis

Jiaheng Zhang: Data Preprocessing and Feature Engineering

Mingyan Xu: Final compilation and report writing

Our project, including code and dataset, is available in Github.