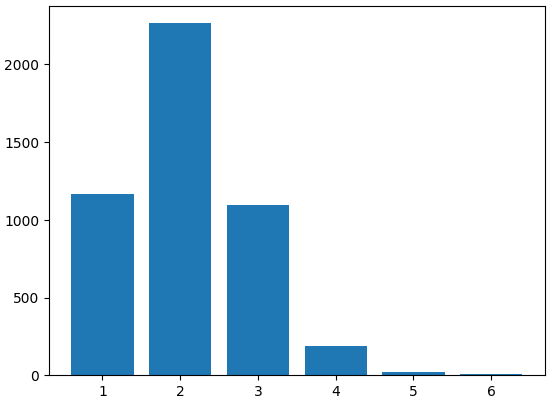
**Introduction**

The rental cost of a house is influenced by various factors such as size, number of bathrooms, furnishing status, and more. This complexity poses challenges for real estate owners in accurately pricing their properties, particularly in a competitive market. The primary goal of this project is to develop a model capable of determining the rental price of a property using information derived from housing data. The essential steps undertaken to achieve this objective include:

1. Data cleaning and exploration
2. Feature engineering
3. Train-test split validation
4. Scaling and modeling
5. Evaluation of the model’s performance

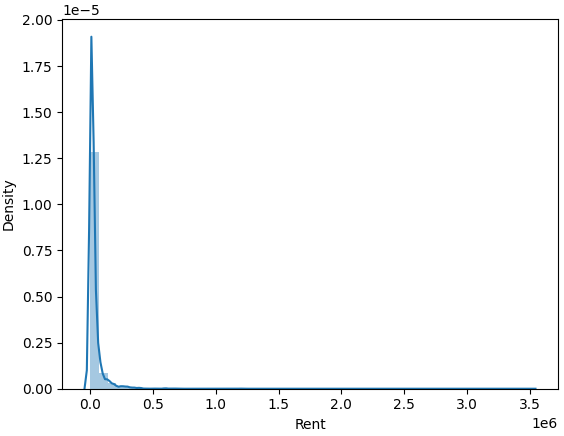
**Data Exploration**

The dataset includes a total of 4,746 entries with a total of 12 features. Delving into the specifics, the distribution of the ‘BHK’ (Bedrooms, Hall, Kitchen) in the dataset reveals a diverse composition of housing configurations. The most common being 2 BHK, constituting 2,265 instances, followed by 1 BHK with 1,167 occurrences. The dataset also encompasses 1,098 instances of 3 BHK, 189 instances of 4 BHK, 19 instances of 5 BHK, and 8 instances of 6 BHK.



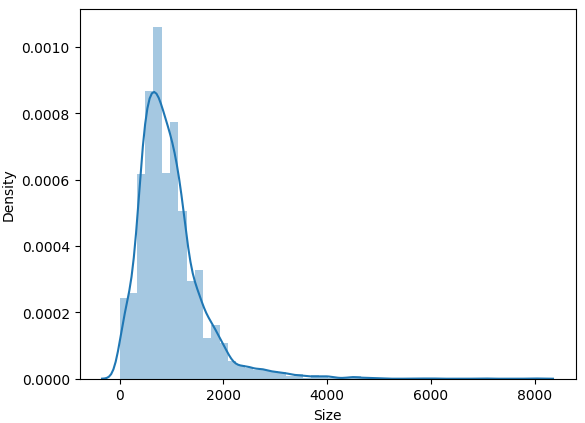
**Figure 1.** Distribution of BHK

Following BHK is the Rent feature. It has an average of approximately ₹34,993.45 with the values ranging from a minimum of ₹1,200 to a maximum of ₹3,500,000. The rent distribution exhibits a large standard deviation of approximately ₹78,106.41, highlighting significant rent variability.



**Figure 2.** Rent Distribution Plot

Another feature is the size of the house. Based on the dataset, the average size of these residential units is around 967.49 square feet, exhibiting diversity as sizes vary from a minimum of 10 square feet to a maximum of 8,000 square feet. The distribution of sizes is characterized by a notable standard deviation of approximately 634.20, highlighting significant variability in the size data.



**Figure 3.** Size Distribution Plot

The fourth feature is Floor, which has a total of 480 unique entries. The most prevalent one being ‘1 out of 2’ with 379 occurrences, closely followed by 'Ground out of 2' at 350 instances. Other common configurations include '2 out of 3' with 312 occurrences, '2 out of 4' with 308 occurrences, and '1 out of 3' with 293 occurrences. Following this is the Furnishing Status. It showcases three distinct categories with semi-furnished being the most common at 2,251 instances. This is followed by unfurnished at 1,815 occurrences and furnished at 680 instances.

**Methodology**

A structured methodology was adopted in the creation of this project. The initial step involved observation and exploration of data. This step is essential in laying the foundation for an informed decision-making throughout the modeling process. Following this is data cleaning. This addresses the missing and inconsistent values. This is evident on the values under the Floor feature, specifically, the naming convention for each floor differs per building. Subsequently, feature engineering is employed to enhance the model. This encompasses the creation of new features, transformations, and categorical variable encoding. The dataset is then partitioned into training and testing sets using the train-test split validation technique, ensuring the model's training and evaluation on independent subsets. With the data prepared, feature scaling was applied to standardize numerical features, and machine learning models, suitable for regression tasks, are implemented. Linear Regression was utilized to learn the patterns and relationships within the data. The final step involves the critical evaluation of the model's performance. Metrics such as Mean Squared Error (MSE) and R-squared were employed to assess accuracy and reliability. Predicted values are compared with actual values from the test set, facilitating iterative improvements to the model or consideration of alternative algorithms for optimization.

**Experiments**

In this section, the experimentation and trials conducted are discussed. The trials were carried out using Jupyter Notebook and Python. In the experimentation phase, a comprehensive approach was undertaken to prepare and optimize the dataset for rent price prediction. Initially, the data underwent a meticulous cleaning process, specifically targeting the 'Floors' section. This involved assigning numerical representations to string names such as 'Ground' and 'Lower Basement' to facilitate subsequent analyses.

A screenshot of a computer

Description automatically generated

**Figure 4.** Before Data Cleaning

A screenshot of a computer

Description automatically generated

**Figure 5.** After Data Cleaning

To ensure data integrity, a crucial step involved validating and correcting the variables to their appropriate data types, including datetime, integers, and other relevant formats. Following this data cleaning phase, the focus shifted to feature engineering. New columns were created, deemed relevant for aiding the machine in identifying patterns related to rent pricing. One notable feature engineering technique employed was the conversion of categorical variables to One-Hot Encoding, enhancing the model's ability to comprehend and utilize categorical information effectively. Subsequently, the dataset was partitioned using the train-test split validation technique, with a test size of 0.2 and a fixed random state of 42, ensuring consistency in subsequent evaluations.

A white card with black text

Description automatically generated

**Figure 6.** One-Hot Encoding of Categorical Variables

Standardization of values was implemented using the StandardScaler to harmonize the numerical features and eliminate potential biases. For model selection, a Linear Regression model was employed, serving as the foundation for subsequent analyses. The evaluation process encompassed both qualitative and quantitative assessments, utilizing metrics such as Mean Squared Error (MSE) and R-squared. This comprehensive approach ensured a robust exploration of various preprocessing and modeling techniques, setting the stage for refining the rent price prediction model.

**New Columns Added**

|  |  |
| --- | --- |
| Floor Number | This is the cleaned version of the value under Floor. The string/digit that is found before ‘out of’ goes here. |
| Total Floors | This is the cleaned version of the value under Floor. The string/digit that is found after ‘out of’ goes here. |
| FloorRatio | The ratio between FloorNumber and TotalFloors. This is useful in knowing how high up the unit is. |
| FloorDiff | This shows the difference between TotalFloors and FloorNumber. |
| SizePerBath | This shows the correlation between the size of the unit and the number of bathrooms in it. |
| BHKSize | This shows the correlation between the size of the unit and the number of bedrooms, halls, and kitchens in it. |
| BathPerBHK | This shows the correlation between the number of bathrooms and bedrooms, halls, and kitchens. |
| SizeAreaAve | This shows the average size of the unit depending on the area type. |
| SizeRatioAve | This shows the average size of the unit depending on the ratio of its floor number and the building’s total number of floors. |
| SizeCityAve | This shows the average size of the unit depending on the city. |
| BathroomSizeAve | This shows the average number of bathrooms depending on the unit’s size. |
| FloorRatioAreaAve | This shows the average floor ratio depending on the area type. |
| FloorRatioCityAve | This shows the average floor ratio depending on the city. |

**Results and Analysis**

In the initial phase of model evaluation, a coefficient of determination of 0.92 was achieved, accompanied by a mean squared error of 192,337,765.23. It's important to note that this performance was obtained while utilizing features that directly incorporated the original 'Rent' values in their calculation. Subsequently, a critical realization was found regarding the potential risk of data leakage. This phenomenon occurs when information from the target variable is unintentionally included in the features used to construct the model, resulting in a seemingly robust performance during training but compromising efficacy on new, unseen data.

A screenshot of a computer

Description automatically generated

**Figure 7.** Quantitative Evaluation of First Trial

A screenshot of a computer program

Description automatically generated

**Figure 8.** Qualitative Evaluation of First Trial

To address this concern, an attempt was made to remove any features that involved the 'Rent' variable, leading to a revised model with a coefficient of determination of 0.62 and a mean squared error of 1434696539.59. While these metrics reflected a comparative decline, it is crucial to emphasize that this model is designed to exclude the target variable, ensuring resilience against data leakage. By doing so, the model maintains its ability to generalize effectively to novel, unseen data scenarios where 'Rent' values may not be available. This strategic adjustment prioritizes the model's reliability in real-world predictive applications, acknowledging the importance of preventing data leakage for robust and accurate predictions.

A screenshot of a computer

Description automatically generated

**Figure 9.** Quantitative Evaluation of Second Trial

A screenshot of a computer program

Description automatically generated

**Figure 10.** Qualitative Evaluation of Second Trial

**Conclusions & Recommendations**

Predicting the rental cost of a house is a complex task influenced by a multitude of factors. The challenge is heightened by the constraint of not utilizing the 'Rent' variable during the model's training, necessitating a nuanced approach to feature selection and engineering. The absence of direct access to the target variable demands careful consideration of various aspects influencing rent pricing. Throughout the process, it became evident that the intricacies of rental cost prediction are multifaceted, and despite numerous attempts, there were continuous discoveries of overlooked nuances.

Moving forward, several areas for improvement have been identified. Meticulous cleaning of ‘Floor’ values could enhance the accuracy of the model by addressing potential inconsistencies in the dataset. Additionally, the exploration of other relevant features could contribute to a more comprehensive understanding of the factors influencing rent pricing. Finally, considering alternative linear models such as Ridge and Lasso could provide valuable insights into whether regularization techniques would yield improvements in model performance. These recommendations offer avenues for further exploration and refinement, ensuring the ongoing enhancement of the rent price prediction model.