Project Title: Real Estate Price Prediction in Bangladesh

Project Overview

Description:
This project involves the exploration, cleaning, and analysis of a real estate dataset from various cities in Bangladesh, including Dhaka, Chattogram, Cumilla, Narayanganj, and Gazipur. The goal is to build predictive models to estimate property prices based on various features such as bedrooms, bathrooms, floor area, and property type.

Several machine learning techniques were employed, starting with linear regression and progressing through advanced models like Ridge, Lasso, and Elastic Net regression. Regularization techniques were used to improve model performance and prevent overfitting.

The dataset is sourced from Kaggle and provides details on property features such as bedrooms, bathrooms, floor area, and corresponding prices in Bangladeshi Taka (৳).

Key Analyses:

1. Data Cleaning & Preprocessing:
   - Removal of duplicates.
   - Cleaning non-numeric data in the "Floor_no" column.
   - Creating meaningful features such as "Type of Property for Sale" from the title data.
   - Filling missing values using medians and feature engineering.
2. Exploratory Data Analysis (EDA):
   - Correlation analysis to identify key predictors.
   - City-wise property price analysis.
   - Visualizing price distribution by city and examining price vs. floor area and floor number.
3. Predictive Modeling:
   - Multiple regression models (simple and multiple linear regression).
   - Advanced techniques using Ridge, Lasso, and Elastic Net regression for regularization.
   - Model performance evaluation through cross-validation with R^2 and Mean Absolute Error (MAE) metrics.

Data Cleaning and Preparation

1. Initial Dataset Overview

After downloading the dataset from Kaggle, I began by examining the data to understand its structure and identify potential issues. This included:

- Previewing the data to observe the columns and types of entries.
- Checking the dataset size, which initially contained 3865 rows and 9 columns.
- Reviewing summary statistics to detect obvious outliers or irregularities.
- Investigating the data types of each column and looking for discrepancies, such as Floor_no and Price_in_taka, which were stored as text instead of numeric values.

Key Findings:

1. The dataset had 934 duplicate rows, which were removed to ensure data integrity, as duplicate entries do not provide additional insights for property datasets.

2. Missing values were present in several key columns:

    - Bedrooms
    - Bathrooms
    - Floor_no
    - Occupancy_status
    - Floor_area

3. The Floor_no column had mixed formats and inconsistent entries, requiring extensive cleaning.

2. Cleaning and Standardizing the Floor_no Column

The Floor_no column contained various irregularities:

- Numeric entries like 1, 2, ..., 18.
- Ordinal strings such as 1st, 8th.
- Complex strings like G+7, 4th to 8th backside, and A1, A2, ..., A7.
- Unusable entries such as "Merin City – Purbach".

Steps Taken:

1. **Dropped Unusable Entries**:
    - Rows with entries like "Merin City – Purbach" were removed, as they lacked information to infer the floor number.
2. **Handled Ranges**:
    - For entries like "4th to 8th backside", I interpreted these as properties spanning multiple floors. The single row was replaced with multiple rows, each representing an individual floor.
3. **Converted Ordinal Strings**:
    - Entries such as "1st" and "8th" were standardized by removing text suffixes and converting them into numeric values.
4. **Resolved "G+7" and Similar Entries**:
    - Entries like "G+7" and "0+7" referred to total floors in a building. These were converted to numeric values representing the total number of floors, e.g., 8.
5. **Assumptions for Special Cases**:
    - For entries such as "A1, A2, ..., A7", it was inferred that this represented a building with seven apartments. Based on the data, I assumed the building had four floors with two apartments per floor and updated the floor numbers accordingly.
6. **Handled Missing Values**:
    - I clarified the ambiguity in the Floor_no column by analyzing the Title column. Using this, I added a new feature called Property Type, categorizing properties into Building, Apartment, Commercial, or House.
    - Missing values in Floor_no were then filled using the median floor number for each property type, ensuring logical consistency.

3. Addressing Missing Values

Bedrooms and Bathrooms:

- Missing values in these columns were filled based on medians calculated for each combination of Property Type and Floor_no. This ensured that the imputed values aligned with the typical configurations of similar properties.

Floor Area:

- Rows with missing values in Floor_area were removed, as this feature is essential for property valuation and could not be reliably estimated.

Location:

- Only one row had a missing value in the Location column. This was filled by referencing the Title column to infer the property's location.

Final Steps:

- After addressing the missing data, I removed rows with insufficient information in critical features such as Bedrooms, Bathrooms, Floor_no, and Floor_area.

4. Final Dataset Summary

After the cleaning process, the dataset was reduced from its original size to 2,759 rows and 18 columns. The following transformations were completed:

1. **Duplicates Removed**: Eliminated 934 duplicate rows.
2. **Floor_no Standardized**: Cleaned and formatted the column to ensure all entries were numeric and logical.
3. **Missing Values Addressed**: Missing data in critical features was filled using appropriate techniques, or rows were removed if imputation was not feasible.
4. **Property Types Defined**: Created a new feature, Property Type, to provide clarity and context for other variables.

5. Reflection on Data Cleaning

This cleaning process ensured the dataset was accurate, consistent, and ready for analysis. Special attention was given to:

- Retaining meaningful data by carefully handling missing values and irregular entries.
- Making reasonable assumptions, supported by context from the dataset (e.g., using property titles to clarify ambiguous features).
- Ensuring that the dataset's integrity was preserved while making it suitable for further analysis and modeling.

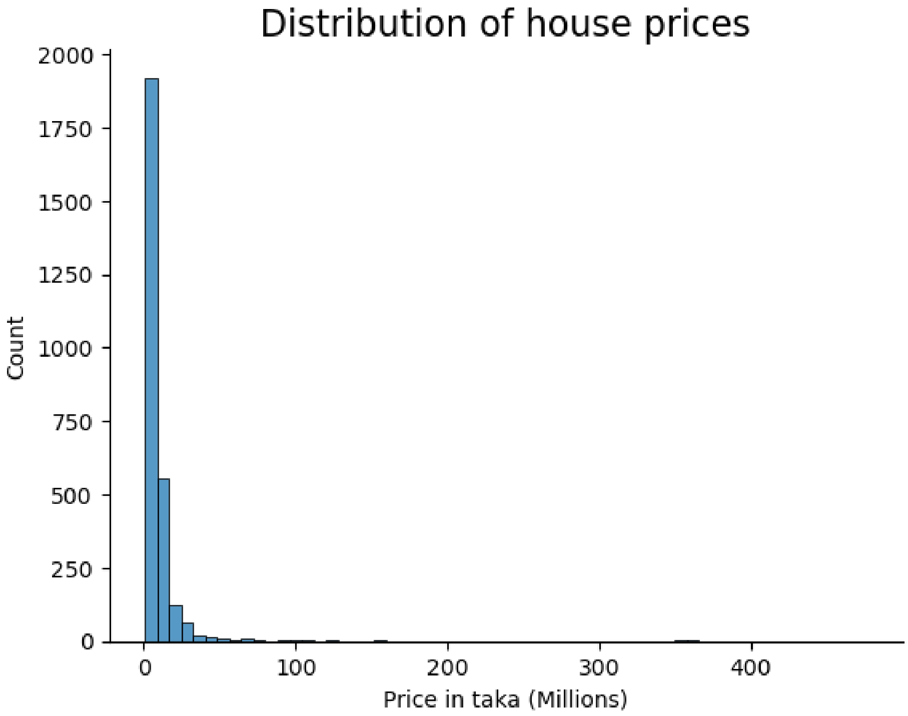
The cleaned dataset now provides a solid foundation for exploring trends, developing predictive models, and drawing actionable insights about the real estate market in Bangladesh.

Exploratory Data Analysis (EDA)

1. Target Variable: Price_in_taka

To understand the distribution of house prices, I began by exploring the target variable, Price_in_taka.

- **Distribution Analysis**:
  - I visualized the distribution of house prices using a histogram, which revealed that the data was left-skewed. This was a logical observation, as property prices generally follow a distribution where most properties are priced lower, with fewer high-priced properties pulling the distribution to the right. This left-skewed pattern aligns with expectations in the real estate market.



2. Correlation Analysis

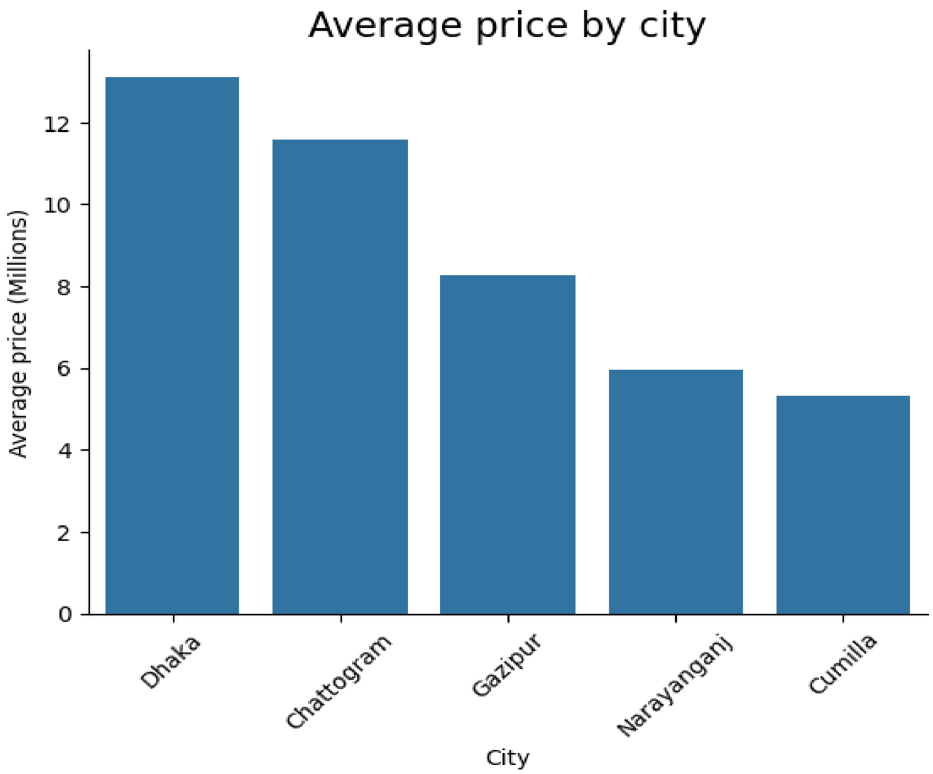
Next, I performed a correlation analysis to identify relationships between the target variable (Price_in_taka) and other features in the dataset.

- **Correlation with Key Features**:
  - **Floor Area**: There was a moderate positive correlation (0.5) between Price_in_taka and Floor_area, suggesting that larger properties tend to be more expensive.
  - **Bedrooms and Bathrooms**: Both features showed a modest positive correlation (0.19 and 0.2, respectively) with price, indicating that properties with more bedrooms and bathrooms tend to be priced higher, though the correlation was not very strong.

3. City-Wise Price Analysis

I next explored the data by city to identify regional pricing trends.

- **Average Price by City**:
  - I used a barplot to visualize the average price of properties across different cities. The analysis revealed that **Dhaka** and **Chattogram** had the highest average property prices, followed by **Gazipur**, **Cumilla**, and **Narayanganj**, which had the lowest average prices.

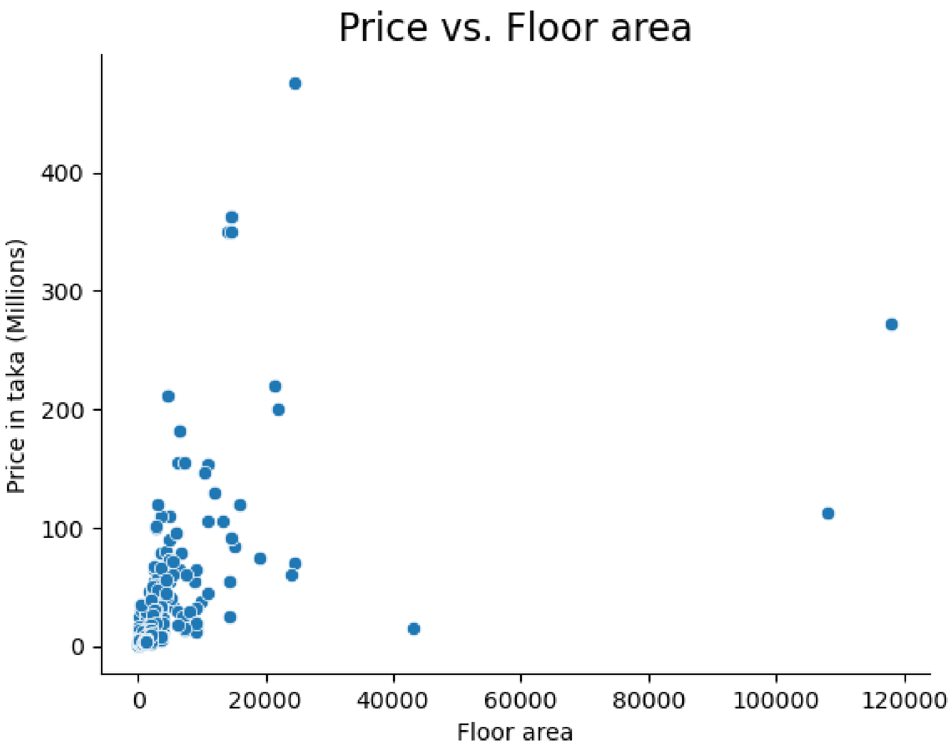


- **Price Distribution by City**:
  - I used a boxplot to visualize the range and variation in property prices across different cities.
    - **Dhaka** had the widest price range, with medium outliers reaching up to **200 million Taka**.
    - **Chattogram** exhibited a relatively broad price range, with outliers extending up to **500 million Taka**.
    - **Cumilla** had the narrowest price range, with outliers clustered between **200,000 Taka** and **300,000 Taka**.
    - **Narayanganj** and **Gazipur** followed a similar pattern, with **Gazipur** having a few outliers reaching up to **300 million Taka**.

4. Price vs. Floor Area

To explore the relationship between property size and price, I plotted a scatterplot of Price_in_taka versus Floor_area.

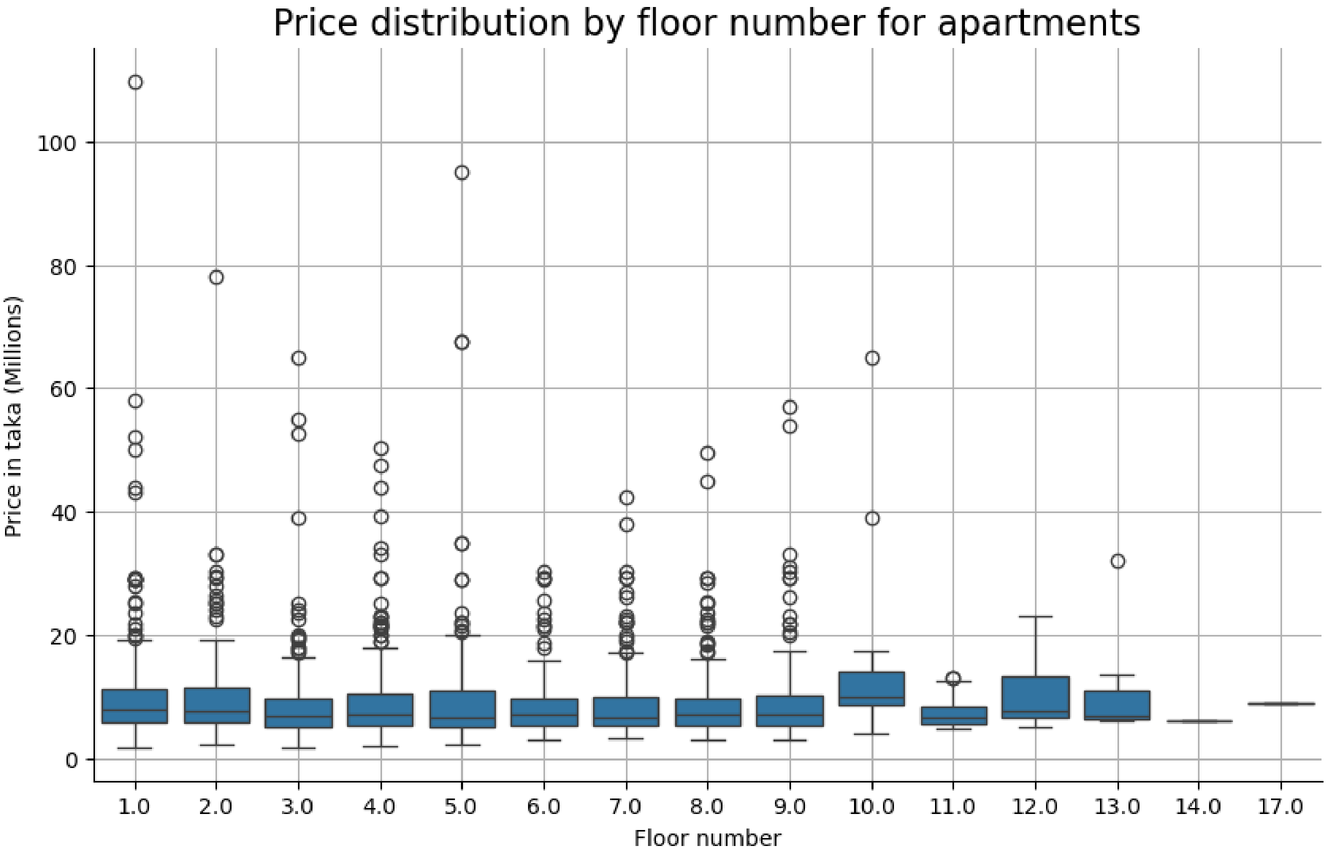
- **Findings**:
  - There is a positive relationship between Price_in_taka and Floor_area, which aligns with the expectation that larger properties tend to have higher prices.
  - However, the scatterplot also shows a spread of data points, indicating that the relationship is not strictly linear. Some properties with smaller floor areas commanded higher prices, possibly due to factors like location, property type, or additional amenities.



5. Price vs. Floor Number for Apartments

I analyzed the distribution of property prices across different floor numbers for apartments specifically.

- **Findings**:
  - The **median price** across floor numbers remained relatively stable, hovering around **0.2 million Taka** for most floor numbers.
  - For floor numbers **10 and above**, the price distribution narrowed, indicating a more stable price range for these properties.
  - Outliers were observed for all floor numbers, particularly between **floors 3-9**, where high outliers were recorded, likely due to luxury apartments or unique properties (e.g., larger units or better locations).
  - There was no clear upward or downward trend in prices as the floor number increased, suggesting that factors beyond just the floor number (such as building amenities or views) may influence prices.

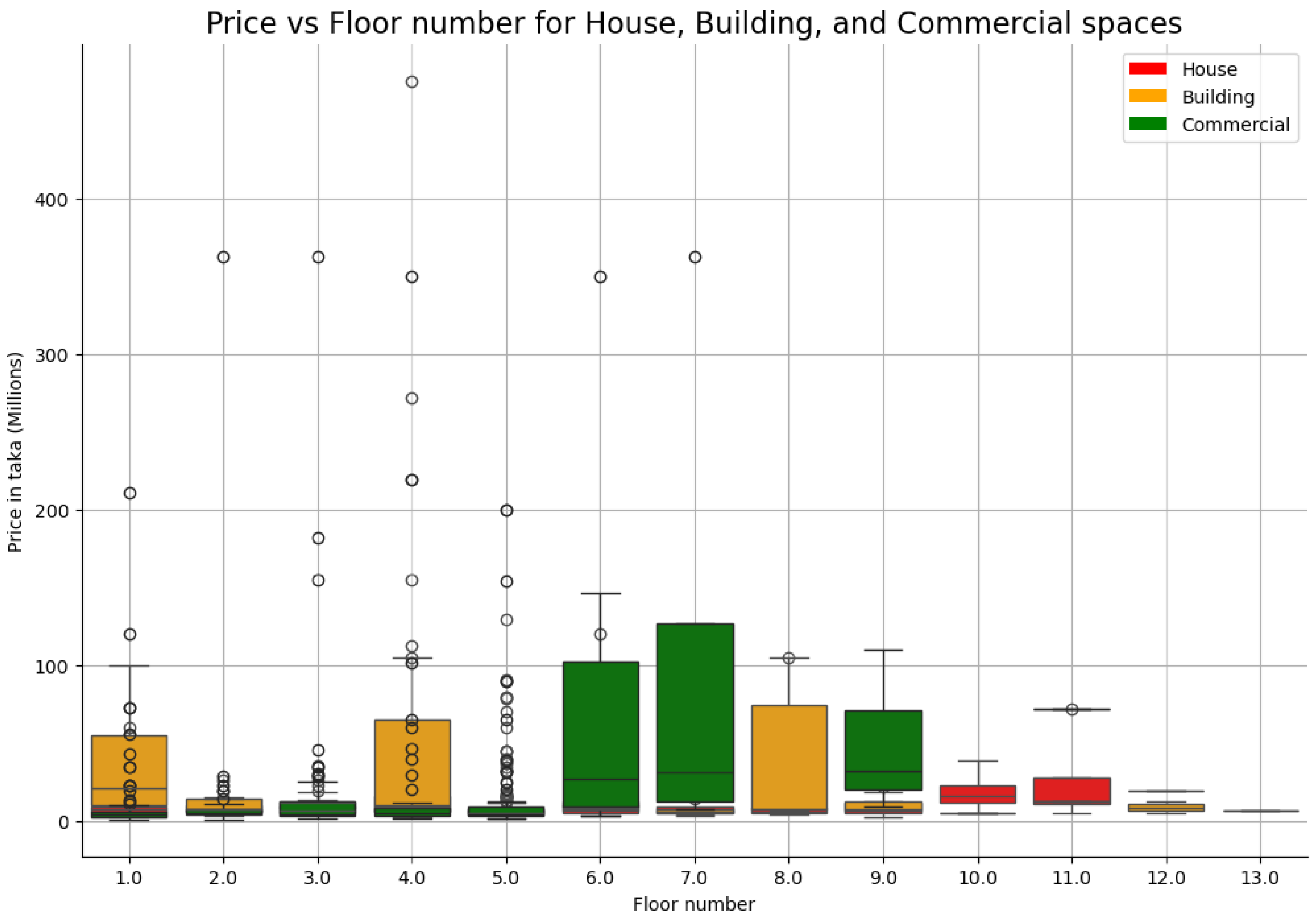


6. Price vs. Floor Number for Houses, Buildings, and Commercials

I then examined how price distributions differed across property types (Houses, Buildings, and Commercial Spaces) as floor numbers increased.

- **Houses**:
  - Prices were relatively low and consistent across floor numbers. The majority of houses were priced in the lower range, with few high-price outliers. This indicates that houses tend to be more affordable compared to other property types.
- **Buildings**:
  - Prices for buildings exhibited a moderate range, with the distribution expanding as the floor number increased. The price range particularly widened between **floors 4 and 7**, with a few premium units at higher floors.
- **Commercial Spaces**:
  - Commercial spaces showed the highest price variability. The price distribution widened significantly starting from **floor 6**, with extreme outliers observed at higher floors (7+). This suggests that commercial properties, especially at higher floors, tend to have higher prices, likely due to their location or intended use.
- **General Trends**:
  - **Commercial spaces** exhibited the most significant price variability, with some very high prices at higher floors.

- **Buildings** also showed a larger price distribution as the floor number increased, while **houses** remained consistently low-priced with fewer high-price outliers.



7. Price vs. Floor Area with City Interaction

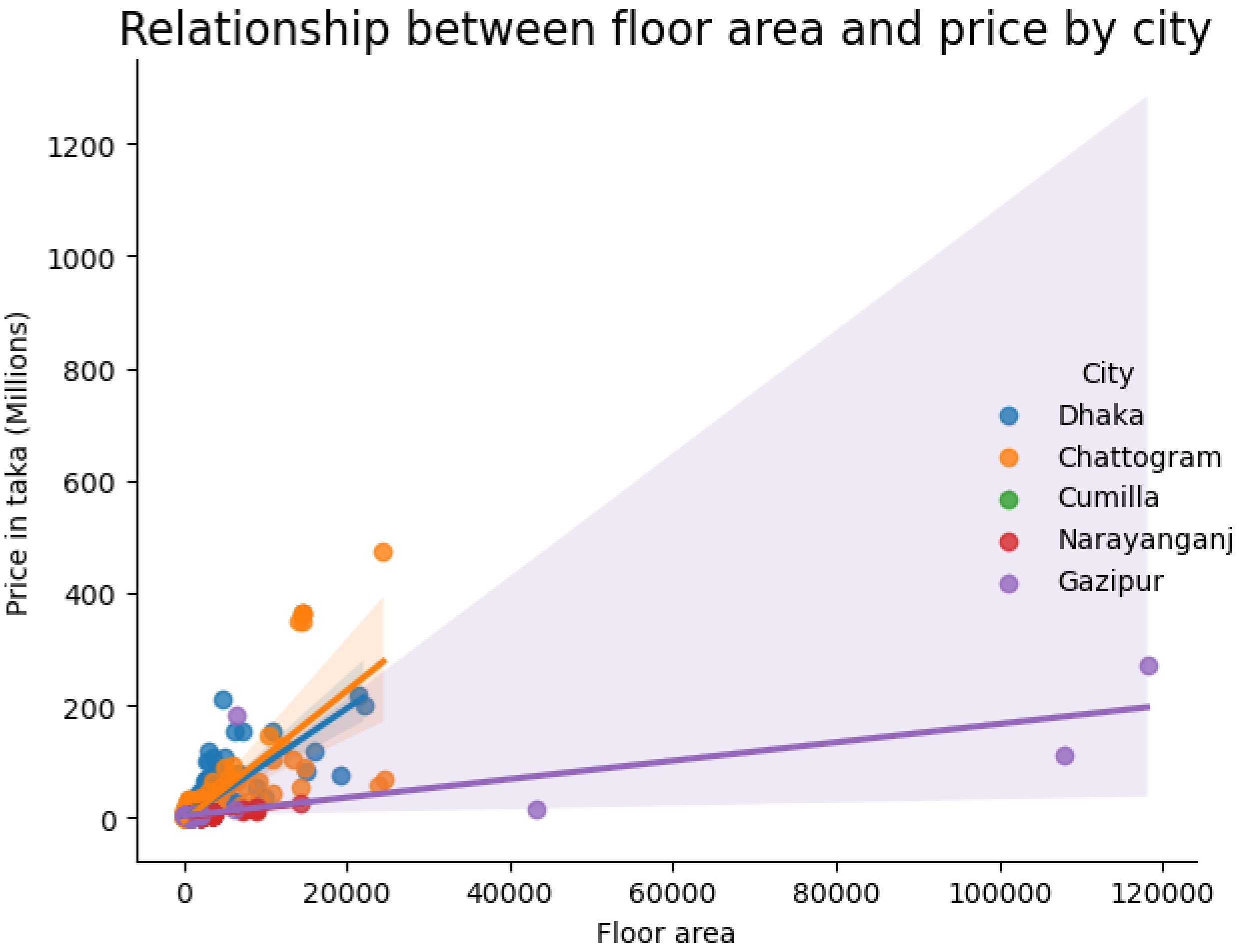
I further explored the relationship between floor area and price, while incorporating the city as an interaction term to see if trends varied across different locations.

- **General Trends**:
  - A positive correlation between floor area and price was observed across all cities, as expected. Larger properties tended to be more expensive.
- **City-wise Observations**:
  - **Dhaka**: The relationship between floor area and price was relatively strong, with a tighter clustering of data points and a steep upward trend, indicating that larger properties in Dhaka tend to have higher prices.
  - **Chattogram**: The trend was similar to Dhaka, but with more variability in prices, especially for smaller floor areas.
  - **Cumilla**: Data points were more scattered, and while the positive correlation still existed, the prices were generally lower compared to Dhaka and Chattogram.
  - **Narayanganj**: A weaker correlation was found, with prices showing greater variability across floor areas.

  - **Gazipur**: The trend here was flatter, with prices increasing gradually with floor area, but at a lower overall price range.

- **Comparison**:

  - **Dhaka** and **Chattogram** had the strongest positive relationship between floor area and price, with steeper regression lines indicating a more defined trend. Smaller cities like **Cumilla**, **Narayanganj**, and **Gazipur** exhibited more price variability and weaker correlation trends, with less consistent price increases as floor area grew.



8. Conclusion of EDA

The exploratory data analysis provided valuable insights into the relationships between property prices and various factors such as floor area, floor number, and city. Key findings include:

- **Price Distribution**: Prices in Dhaka and Chattogram are generally higher, with significant variability observed in commercial properties.

- **Correlations**: Price correlates positively with floor area, though the relationship is not strictly linear, with some smaller properties fetching higher prices due to other factors.

- **Regional Trends**: Dhaka and Chattogram exhibit stronger positive correlations between floor area and price, while smaller cities like Cumilla and Narayanganj show more variability.

- **Floor Number Influence**: For apartments, floor numbers didn’t directly correlate with higher or lower prices, suggesting that other factors, such as location and amenities, play a significant role in determining property prices.

This EDA sets the stage for further modeling and feature engineering, offering a deeper understanding of the property market dynamics.

Modeling Overview:

To predict house prices, I implemented a series of regression models, beginning with simpler models and progressively incorporating advanced techniques to improve predictive performance. The key steps included simple linear regression, multiple linear regression, and regularization techniques such as Ridge, Lasso, and Elastic Net Regression. Below is a detailed breakdown of each stage:

1. Simple Linear Regression

- **Objective**: To establish a baseline model using floor area, the feature most strongly correlated with price.
- **Data Transformation**: To address potential non-linearity and skewness in the data, I applied log transformations to both the target variable (price_in_taka) and the independent variable (floor_area).
- **Evaluation**:
  - **Cross-validation R^2**: 0.341 ± 0.044
  - **Cross-validation MAE**: 5,662,440 ± 896,207
- **Conclusion**: The simple linear regression model showed limited predictive power, with R^2 value indicating that a substantial portion of the variance in house prices was not captured by floor area alone.

2. Multiple Linear Regression

- **Objective**: To enhance the model by incorporating additional features that could capture more complex relationships between the target variable and predictors.
- **Feature Engineering**:
  - Applied log transformations to the target and the floor_area to address skewness.
  - Included interaction terms between Bedrooms, Bathrooms, and Floor_area to account for combined effects on price.
  - Introduced polynomial features for Bedrooms and Bathrooms to model non-linear relationships.
  - Created bins for Bedrooms (e.g., 1-3 small, 4-6 medium, 7-10 large, 11+ extra-large) to capture categorical variations.
- **Performance**:
  - **Cross-validation R^2**: 0.686 ± 0.039
  - **Cross-validation MAE**: 4,013,192 ± 696,519
- **Conclusion**: The multiple linear regression model significantly outperformed the simple model, with a nearly 35% increase in R^2. The inclusion of engineered features allowed the model to better capture the complexity of the dataset.

3. Advanced Regression Techniques

After establishing a robust multiple linear regression model, I explored several advanced regression techniques to further enhance predictive performance and handle potential issues such as multicollinearity and overfitting.

Ridge Regression

- **Objective**: To regularize the model and control overfitting by adding a penalty term to the loss function based on the magnitude of coefficients.

- **Implementation**: Initially, I applied Ridge Regression with a subset of the most predictive features. Later, I incorporated all engineered features into the Ridge model to fully leverage the regularization effect.
- **Performance**: I observed a notable performance improvement (~7%) after including all features, which demonstrated Ridge Regression's ability to prevent overfitting while improving accuracy.

Lasso Regression

- **Objective**: To perform feature selection by shrinking less significant coefficients to zero, thus eliminating unimportant features and simplifying the model.
- **Implementation**: I applied Lasso Regression after Ridge to identify the most important features by examining the coefficients that remained non-zero.
- **Outcome**: Lasso Regression provided insight into which features contributed most to predicting house prices. It effectively reduced the model complexity by removing redundant or less informative variables.

Elastic Net Regression

- **Objective**: To combine the strengths of both Ridge and Lasso regression by utilizing both L2 (Ridge) and L1 (Lasso) regularization, allowing for a balance between feature selection and regularization.
- **Implementation**: Elastic Net is particularly beneficial when dealing with multicollinearity, as it both regularizes the model and performs feature selection.
- **Outcome**: Elastic Net provided a hybrid approach that helped stabilize the model and manage multicollinearity more effectively than Ridge or Lasso alone.

4. Model Performance Evaluation

To determine the best-performing model, I compared the performance of all regression techniques based on the following evaluation criteria:

- **R^2 Score**: This metric indicates the proportion of variance in the target variable (house price) explained by the model's features. Higher R^2 values suggest better model fit.
- **Mean Absolute Error (MAE)**: This metric quantifies the average absolute difference between the predicted and actual values. A lower MAE indicates better predictive accuracy.

The models were tested using cross-validation, and their performance was analyzed to identify which model provided the best balance of predictive accuracy and generalizability.

5. Final Model Selection

After evaluating all models, Ridge Regression emerged as the best choice due to its strong predictive accuracy and its ability to control overfitting. The model delivered the best balance between complexity and interpretability, making it the most reliable option for predicting house prices.

6. Model Assumptions Check

To ensure the validity of the chosen regression models, I checked the following key assumptions:

- **Equal Variance of Errors (Homoscedasticity)**: I visually inspected the residuals, which were evenly spread along the x-axis, confirming that the assumption of equal variance of errors was met.

- **Normality of Errors**: A normal probability plot indicated that the errors followed a normal distribution, as most points fell within -2 to +2 standard deviations, satisfying this assumption.
- **No Perfect Multicollinearity**: I checked for multicollinearity using Variance Inflation Factors (VIF). All VIF values were below 5, indicating that multicollinearity was not a major concern in the model. Some feature engineering, such as polynomial terms, naturally introduced some correlations, but they did not lead to problematic levels of multicollinearity.

7. Final Model Evaluation

After selecting Ridge Regression, I evaluated the model on the test set to assess its performance in a real-world scenario:

- **Test R^2**: 0.7759
- **Test Mean Absolute Error**: 3,392,546

8. Feature Selection and Regularization

Using Lasso Regression for feature selection, I shrunk the coefficients of unimportant features to zero, leaving only the most significant predictors. I then applied Ridge Regression to these selected features to refine the model further. This process helped ensure that the final model was both accurate and efficient, with fewer variables contributing to the prediction of house prices.

Conclusion

The Ridge Regression model, after incorporating feature engineering, regularization, and careful assumption testing, proved to be the most effective approach for predicting house prices. It offered a reliable balance between accuracy and simplicity, making it the best model for this dataset. Regularization techniques helped prevent overfitting, and the careful selection of features ensured the model's robustness and interpretability.