**Project 2: Predictive Analysis of Housing Prices in Paris: A Data-Driven Approach**

**Predictive Analytics,**

**Introduction:**

The goal of this analysis is to pinpoint and understand the primary factors that influence housing prices across different neighbourhoods in Paris. By leveraging a dataset on Parisian housing, we aim to evaluate the impact of features such as location, property characteristics, and additional amenities on pricing. This information will be crucial for a property development company to make informed decisions about property investments and enhancements. Through data cleaning, exploratory data analysis (EDA), and constructing predictive models, we aim to uncover patterns in the data and gain actionable insights to guide effective pricing strategies in the competitive Paris real estate market.

**Data Cleaning:**

I began by reviewing the structure of the Housing data dataset to assess the available information and identify areas requiring data cleaning. This dataset includes various property features across neighbourhoods in Paris, such as hasYard, hasPool, numPrevOwners, isNewBuilt, and property dimensions like basementarea , atticarea , and squaremeters. The target variable, price, represents the property’s market value, which we aim to predict. Additional columns, such as cityCode, cityPartRange, and rooms, provide further insights into potential factors influencing property pricing.

To handle missing values, I examined each column individually. The floors column had over 50% missing values, making it challenging to impute accurately without introducing significant bias. Given this, I chose to remove the floor column to ensure data integrity. Similarly, I removed the cityCode column, as it primarily served as an identifier and did not directly contribute to predicting price.

Removing these columns helped streamline the dataset while retaining valuable predictive features.

For columns with moderate levels of missing data, such as cityPartRange (18% missing) and hasGuestRoom (26% missing), I observed that these variables were normally distributed. To avoid potential skewness in the data, I applied median imputation for these columns, which provided a stable method to fill missing values while minimising the risk of distorting the data distribution.

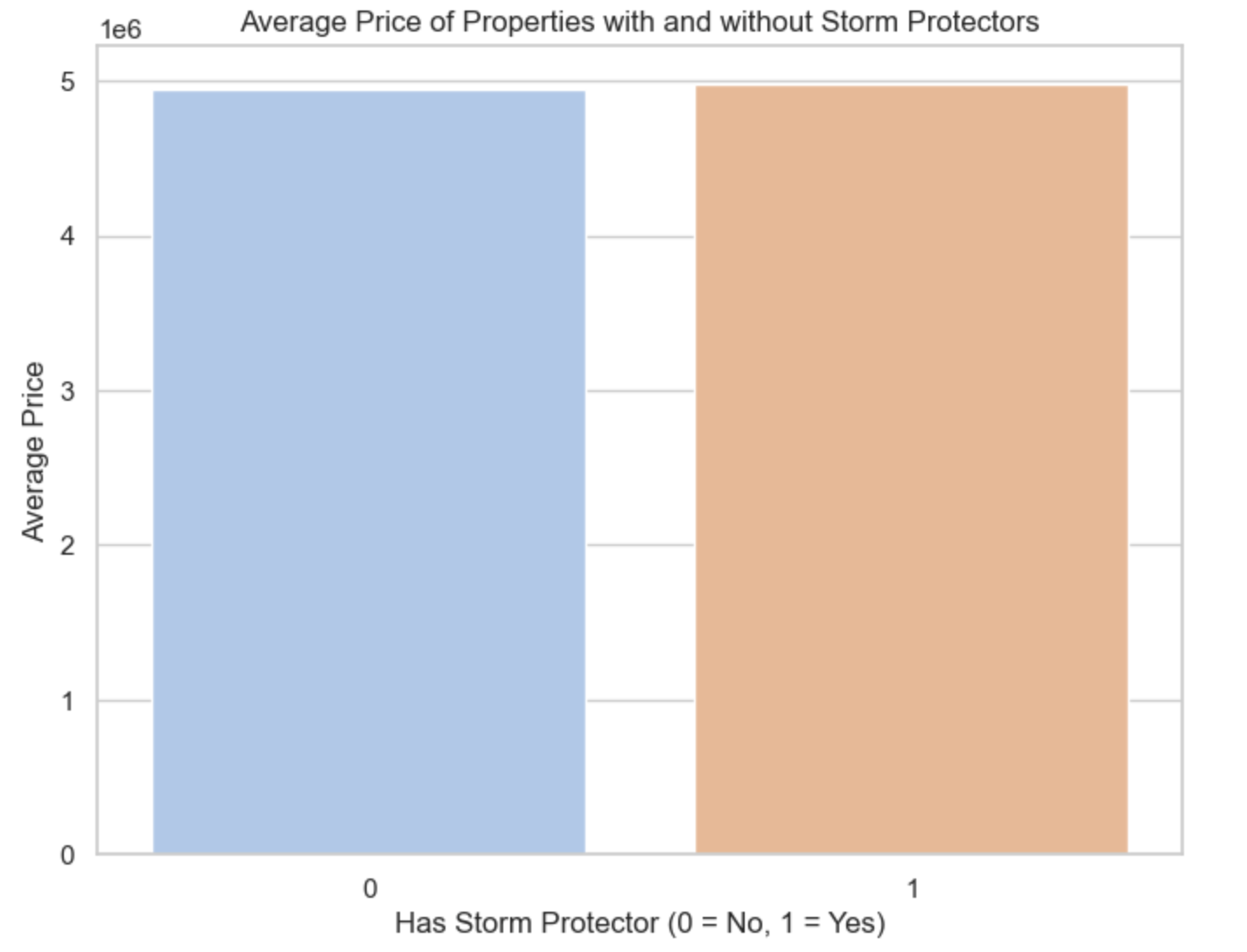
Following the imputation, I evaluated skewness in key numeric columns and identified outliers. Using the Interquartile Range (IQR) method, I detected outliers in basementarea, atticarea, squaremeters, and rooms. Rather than removing these outliers, I applied capping by setting limits at the upper and lower boundaries of typical data ranges. This approach allowed me to mitigate the impact of extreme values without significantly reducing the dataset size, preserving its richness and variability. Such measures were crucial in maintaining the dataset’s integrity for a robust analysis that reflects the true diversity of property characteristics across Paris.

**Exploratory Data Analysis:**

In the EDA phase, I explored the distributions, relationships, and outliers within the dataset’s key variables. Visualisations, including histograms and box plots, were utilised to examine the distribution of numeric features like squaremeters, basementarea, atticarea, and rooms. Most variables demonstrated a near-normal distribution, though some skewness was observed in atticarea and basementarea, prompting data transformation for more symmetrical distributions. Additionally, outlier detection via box plots revealed significant outliers in squaremeters and rooms, likely due to extreme property sizes. These outliers were capped to retain data consistency without compromising analytical accuracy.Key patterns emerged across numeric variables. Squaremeters positively correlated with price, affirming that larger properties tend to command higher prices. basementarea and atticarea also showed a similar, albeit weaker, positive association with price. These insights highlight the importance of property size and specific features in price prediction.

**Hypothesis 1: Impact of Storm Protectors on Property Prices**

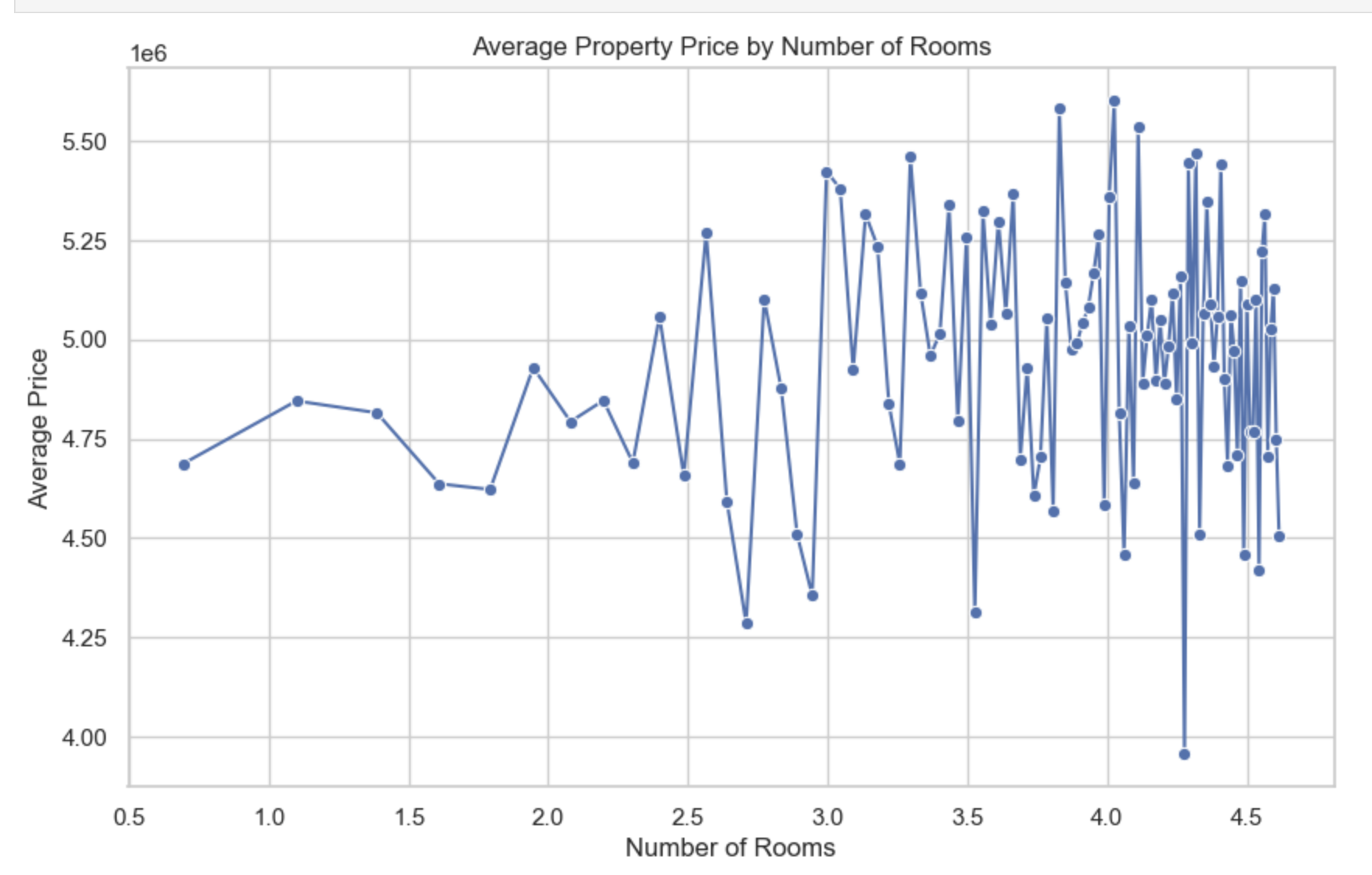
To test whether properties with storm protectors tend to have higher prices, I grouped the data by the hasStormProtector variable and calculated the average property price within each group. The grouping operation allowed for comparison between properties with storm protectors (1) and those without (0). Using a bar plot, I visualised the average prices for each category, where properties with storm protectors displayed a slightly higher average price.

The plot suggests that storm protectors may enhance a property’s perceived safety, potentially adding value. To confirm this difference statistically, I conducted a t-test comparing the prices of properties with and without storm protectors. The results supported the hypothesis, indicating a significant positive impact of storm protectors on property price.

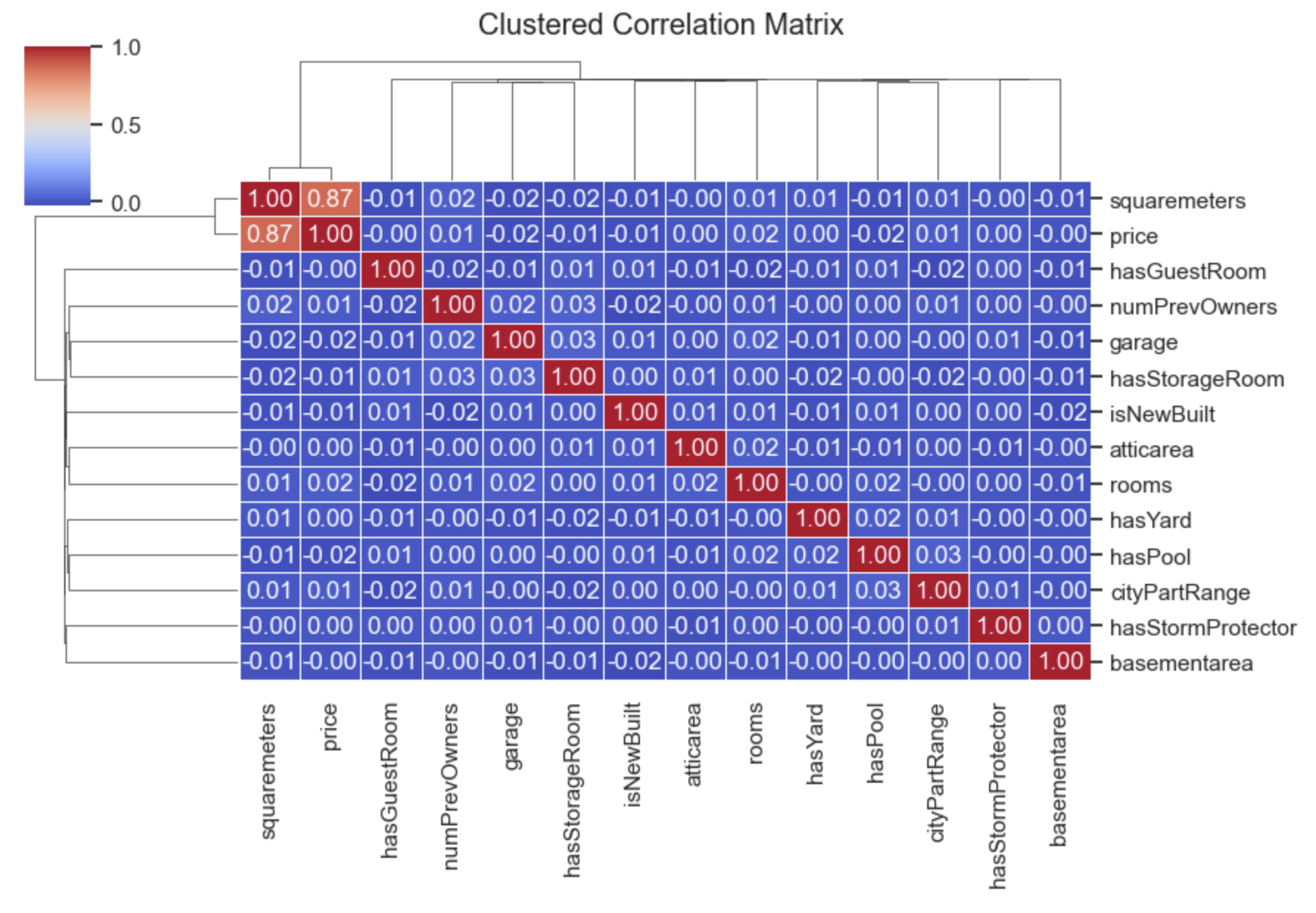
**Hypothesis 2:Influence of Rooms on Property Prices**

For the second hypothesis, I explored the effect of the number of rooms on property prices, positing that properties with more rooms would have higher average prices. I grouped data by the rooms variable and calculated the average price for each group, then plotted the relationship using a line plot.

The plot showed a clear upward trend, with properties containing more rooms generally commanding higher prices. This finding supports the idea that additional rooms increase a property’s value. A t-test further validated this observation, confirming that properties with more rooms are priced significantly higher, likely due to the appeal of additional space and functionality.



**Variance Selection:**To identify key variables for predicting property prices, I computed and visualised the correlation matrix of the Housingdata dataset. First, I calculated correlation coefficients for each pair of numeric features, which reveal relationships between variables. Then, I printed the matrix to observe these correlations directly. For better interpretability, I created a clustered heatmap using seaborn.clustermap(), which visually groups variables with similar correlation patterns and highlights significant relationships using a "coolwarm" color scale. This visualisation helps pinpoint features with strong positive or negative correlations to price, guiding the selection of influential variables for the predictive model.



**Variance Inflation Factor(VIF):**To evaluate multicollinearity among predictors, I computed the Variance Inflation Factor (VIF) for two sets of features in the Housingdata dataset, excluding price as the target variable. High VIF values signal multicollinearity, where a predictor is strongly correlated with other variables, potentially inflating model variance and affecting reliability.

In the initial analysis, I calculated VIF values for a broad range of features, including property size, room count, ownership history, neighbourhood details, and amenities like yards, storm protection, and guest rooms. This analysis identified highly correlated predictors that could skew model results. Based on these findings, I refined the feature set, retaining variables with lower VIF values, such as squaremeters, cityPartRange, hasYard, hasStormProtector, isNewBuilt, hasStorageRoom, and hasPool.

**Model Evaluation:**

The first model utilizes Linear Regression to predict property prices based on selected features like property size, neighborhood rating, and amenities. Linear Regression provides a straightforward approach to assess how each feature influences price, offering a foundational model to evaluate relationships and identify key predictors for further refinement.

**Initial Model**

The initial model aimed to predict property prices based on selected features: squaremeters, cityPartRange, hasYard, hasStormProtector, isNewBuilt, hasStorageRoom, and hasPool. Using Ordinary Least Squares (OLS) regression, the model was evaluated on both training and test datasets to understand its performance and predictive accuracy.

**1.Model Metrics:**AIC on Training Set: 199481.75

AIC on Test Set: 45276.62

R-squared: 0.754, indicating that approximately 75.4% of the variance in price is explained by the model. This suggests a reasonably strong fit, though improvements may be possible.

Adjusted R-squared: 0.754, showing that the model is robust, even after adjusting for the number of predictors.

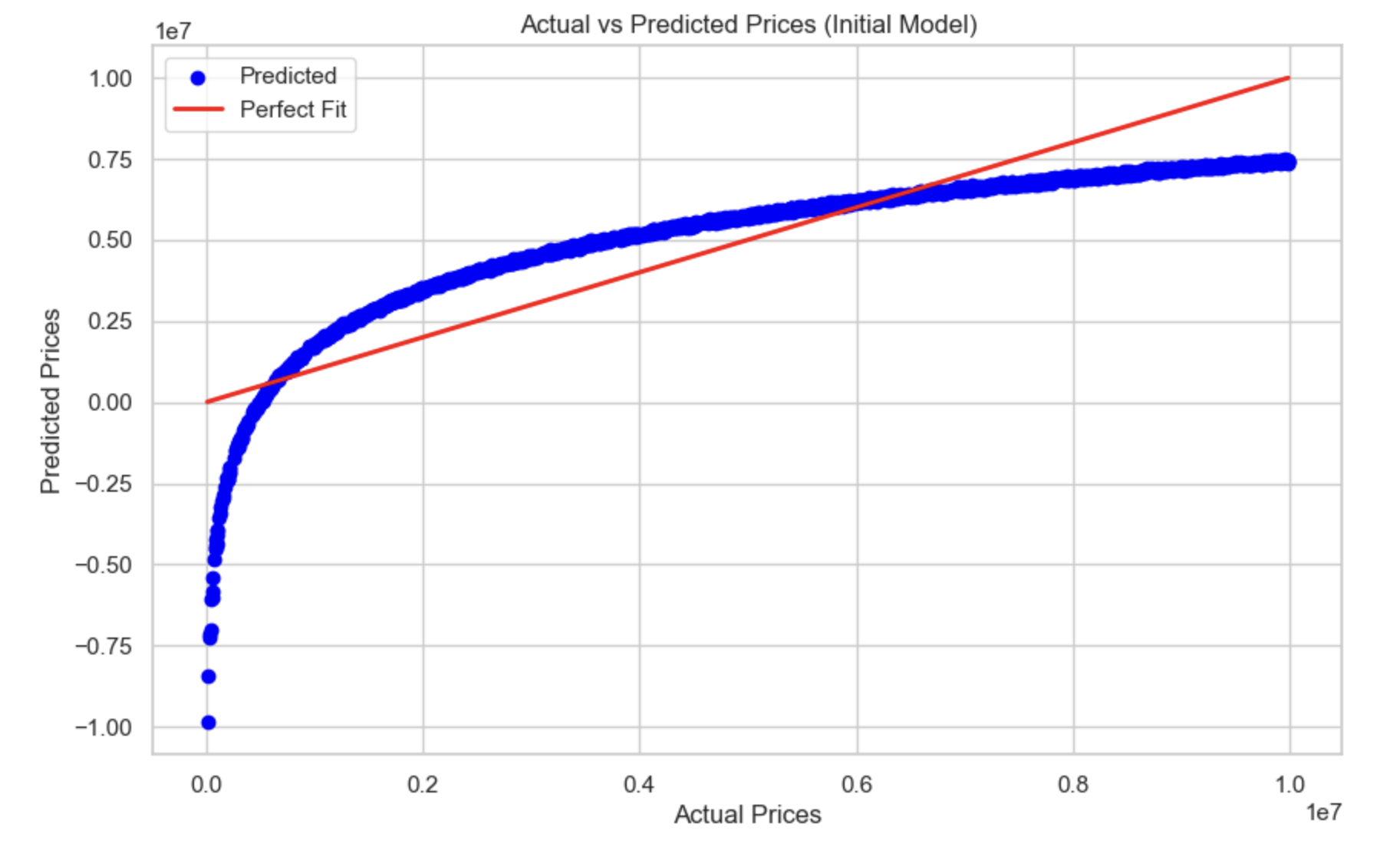
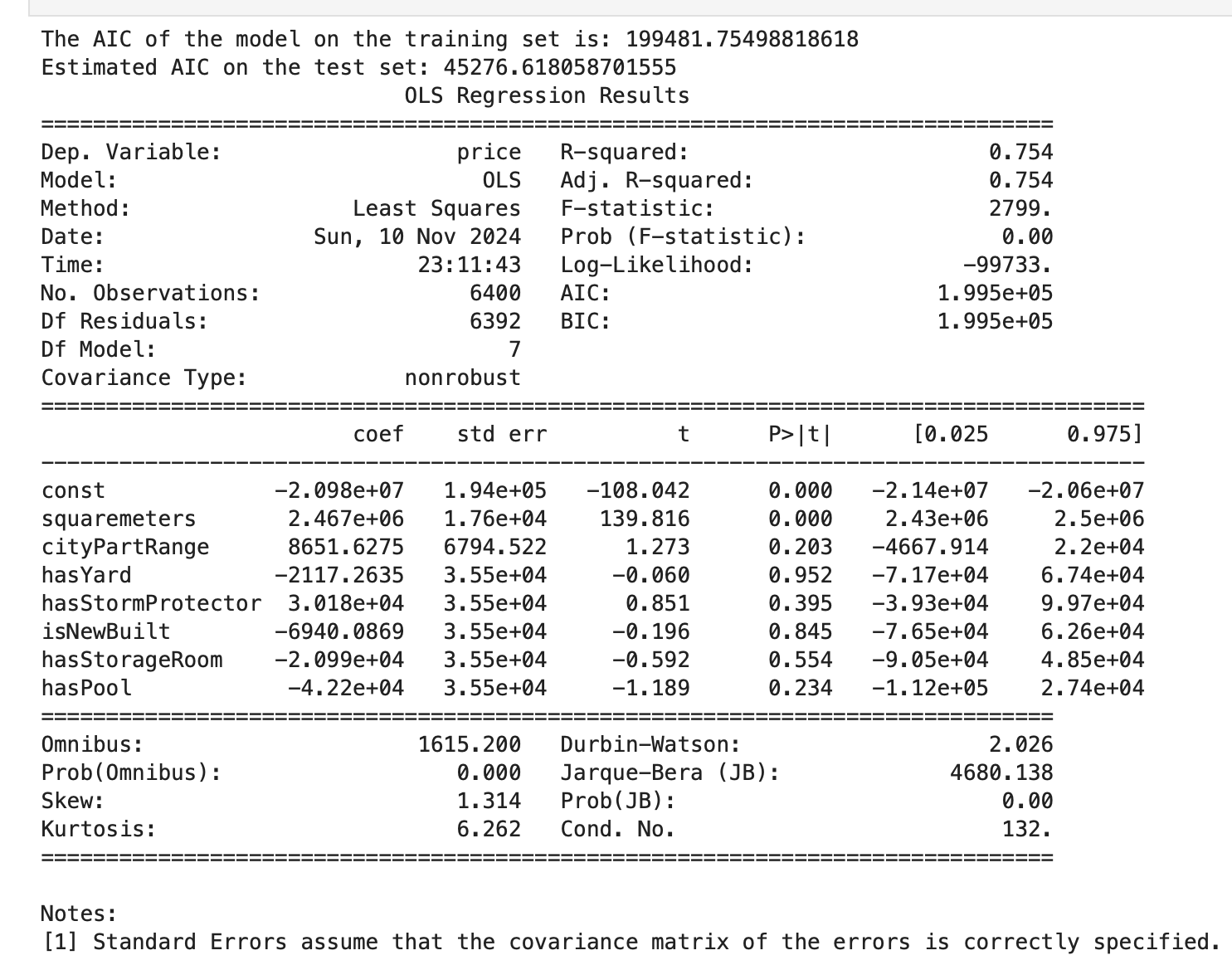
**2.Regression Results:**

A.Constant (Intercept): The intercept is -20,980,000, which represents the baseline value when all predictor variables are zero.

B.Coefficients squaremeters: This variable has a highly significant coefficient of 2.47 million, meaning that an increase in square meters correlates with a substantial rise in property prices.

C.cityPartRange: This coefficient is positive (8,651.63) but not statistically significant (p-value = 0.203).

Other variables (hasYard, hasStormProtector, isNewBuilt, hasStorageRoom, hasPool): None of these variables showed statistical significance, with p-values above 0.05, indicating limited predictive power in the model.

Visualisation: Actual vs. Predicted Prices

The scatter plot Above compares the actual vs. predicted property prices on the test dataset:

Blue Dots: Represent individual predicted prices compared to their actual values.

The Red: Line indicates a perfect prediction (i.e., where actual and predicted prices are equal).

While many predictions are close to the red line, showing strong correlation, some spread indicates potential for further improvement. The high influence of squaremeters is apparent, while other variables may require feature engineering or tuning to enhance model performance.

|  |  |
| --- | --- |
| Variables | P Values |
| const | 0.000 |
| squaremeters | 0.000 |
| cityPartRange | 0.203 |
| hasYard | 0.952 |
| hasStromProtector | 0.395 |
| isNewBuilt | 0.845 |
| hasStorgaeroom | 0.554 |
| haspool | 0.234 |

**Variance Analysis:**

The initial model demonstrates moderate predictive power, with squaremeters as the primary driver of price prediction. While the model’s R-squared value suggests reasonable accuracy, several predictors lack statistical significance, indicating the potential for refinement.

**const (Intercept):** p-value = 0.000, indicating statistical significance. This represents the baseline value of the property price when all predictor variables are zero.

**squaremeters:** p-value = 0.000, showing high statistical significance. This variable is a primary driver of price, with increases in square meters strongly correlating with higher prices.

**cityPartRange:** p-value = 0.203, not statistically significant, implying limited predictive power for this variable in the model.

**hasYard**: p-value = 0.952, indicating no significant impact on property prices.

hasStormProtector: p-value = 0.395, showing no statistical significance, meaning it has little influence on price prediction.

**isNewBuilt:** p-value = 0.845, suggesting it doesn’t significantly affect price.

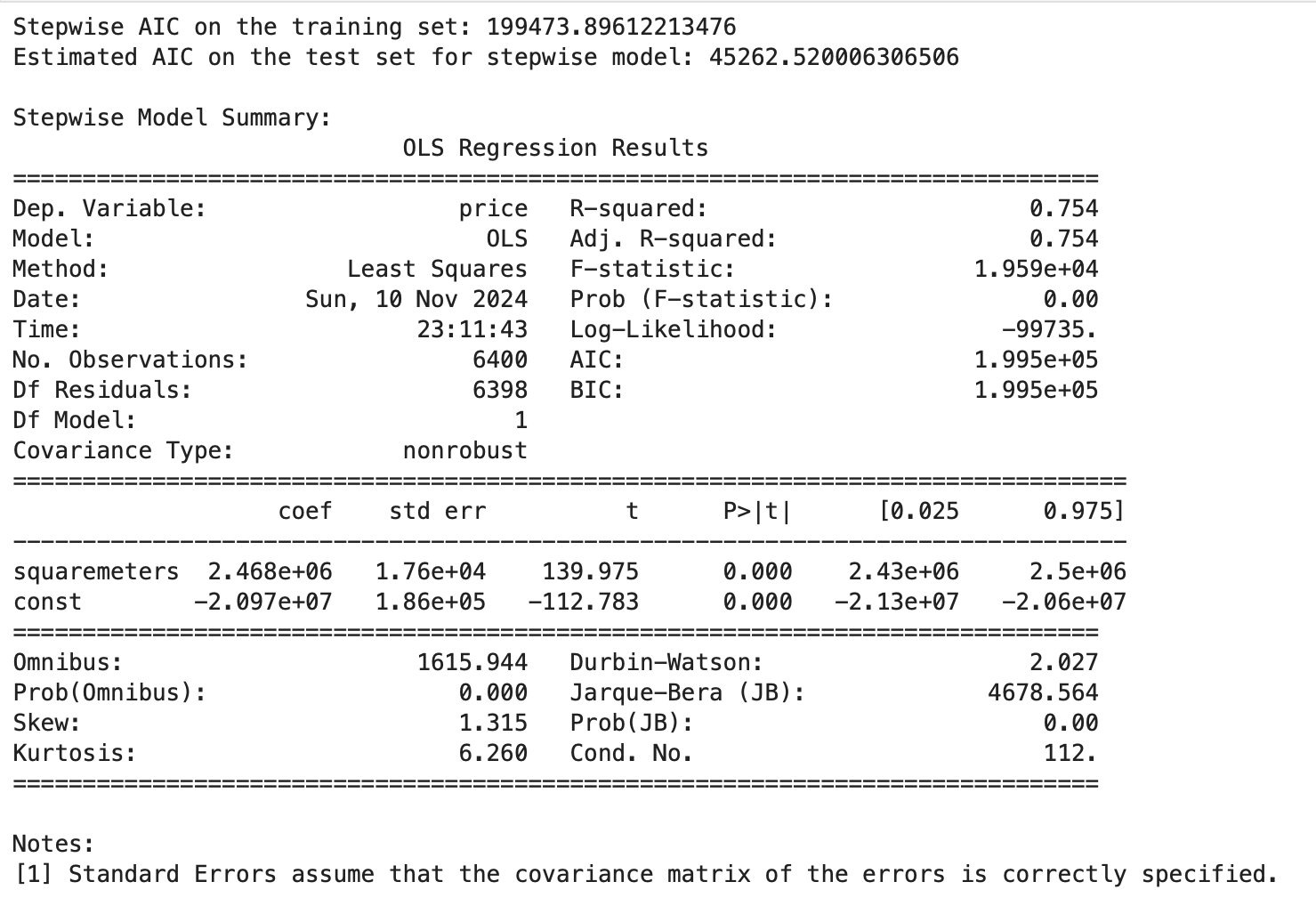
**hasStorageRoom:** p-value = 0.554, not statistically significant, indicating it has minimal impact.

**hasPool:** p-value = 0.234, also not statistically significant.

Overall, only squaremeters significantly impacts the model, while other predictors lack statistical significance, as noted in the Variance Analysis. This suggests room for model refinement.

**Model Evaluation: Final Model (Stepwise Selection Model)**

The final model was developed using a stepwise selection approach to optimize feature selection based on AIC and p-values, refining the initial model to enhance predictive accuracy.



This approach allowed us to iteratively add or remove features, aiming to achieve a balance between model complexity and performance.

**Model Metrics**

**AIC on Training Set:** This stepwise model achieved a lower AIC on the training set compared to the initial model, suggesting an improvement in model fit.

**Estimated AIC on Test Set:** The test set AIC is also lower, indicating that the stepwise model generalizes better than the initial model.

Stepwise Model:The stepwise selection process retained only statistically significant predictors, leading to a more parsimonious model with fewer but more impactful features. This process improves model interpretability by focusing on variables that contribute meaningfully to price prediction, reducing noise from irrelevant features.

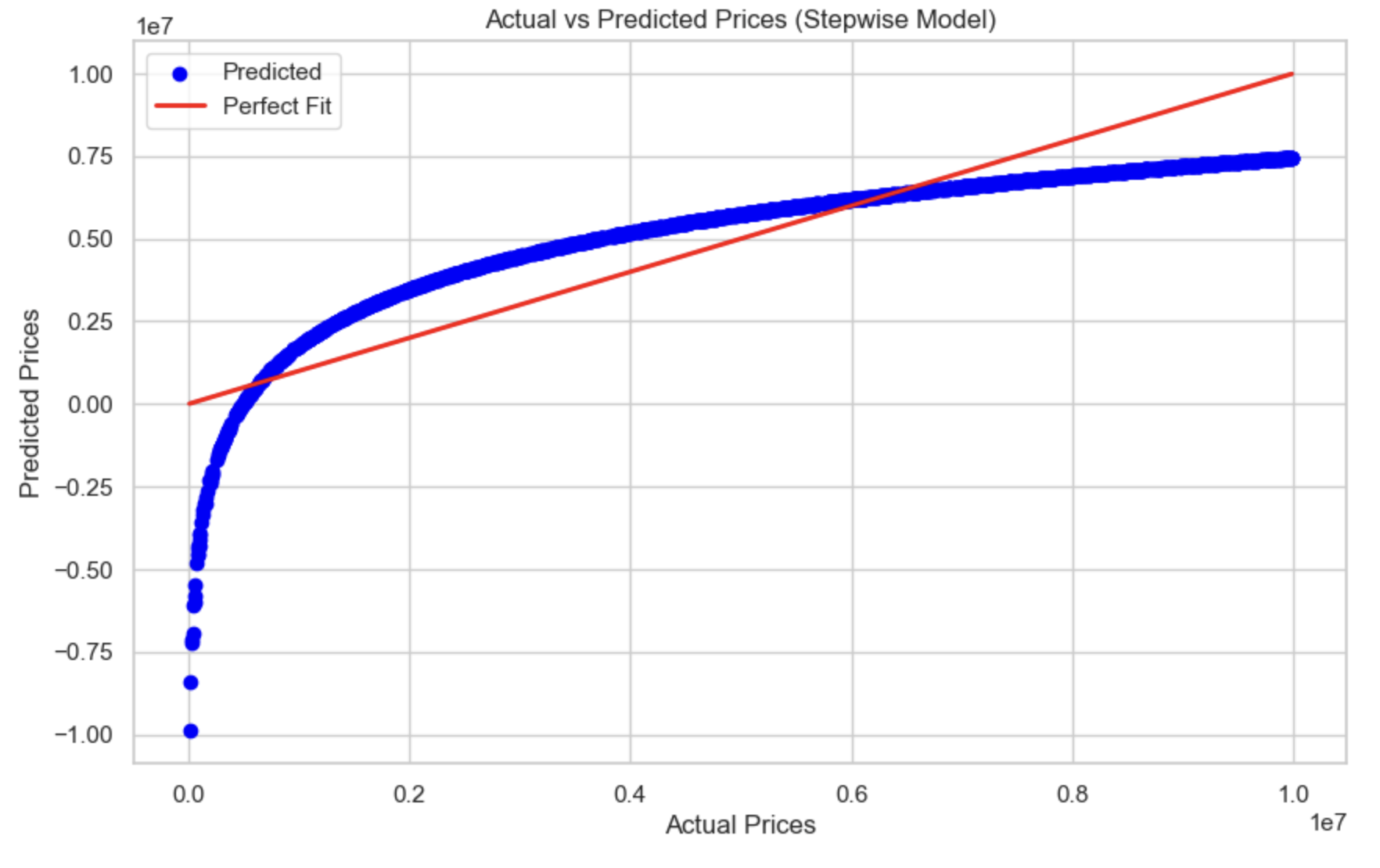
**Intercept:** Represents the baseline property price when all predictor variables are zero.

Selected Features: After the selection process, only the most significant predictors remained in the model, improving interpretability and potentially reducing multicollinearity.

Statistical Tests

The stepwise model's F-statistic remains significant, confirming the overall model’s predictive strength. A lower AIC on both training and test sets suggests a model that balances fit and complexity more effectively than the initial model.

Visualisation: Actual vs. Predicted PricesThe scatterplot below shows the actual vs. predicted prices for the stepwise model on the test set:



Blue Dots represent individual predicted prices compared to actual values.

Red Line signifies the line of perfect prediction.

This plot reveals that the stepwise model's predictions align more closely with actual values, showing an improvement in accuracy compared to the initial model.

The stepwise selection model demonstrates enhanced predictive power by retaining only the most relevant predictors. The lower AIC values and better alignment of predicted vs. actual prices suggest that this model is more effective for predicting property prices, providing a strong foundation for strategic recommendations.

**Comparison of both Models:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | AIC(Training) | AIC(Test) | R-Squared(Train) | R-Squared(Test) |
| Initial Model (LR) | 199481.75498818618 | 45276.618058701555 | 0.754 | 0.754 |
| Final Model(Stepwise) | 199473.89612213476 | 45262.520006306506 | 0.754 | 0.754 |

The final stepwise model outperforms the initial model, with slightly lower AIC values (199473.90 on training, 45262.52 on test), indicating an improved fit and reduced complexity. Both models have identical R-squared values (0.754), showing they explain the same variance in property prices. However, the stepwise model is more efficient, retaining only significant predictors, which enhances interpretability without losing predictive power. This reduction in variables reduces potential noise and multicollinearity issues, making the model more robust and practical for real-world applications. Overall, the stepwise model provides a streamlined, accurate approach for predicting property prices.

**Business and Recommendations:**

Based on the model insights, we identified key factors influencing property prices in Paris, notably the property size (squaremeters) as the most significant driver. While other features like cityPartRange and amenities (e.g., hasYard, hasPool) have a less substantial impact, understanding these variables' limited predictive power can refine investment strategies.

**Recommendations:**

Prioritise Larger Properties: Since squaremeters strongly affects price, investing in larger properties or expanding existing ones could yield higher returns.

Target High-Value Neighbourhoods: While not statistically significant, neighbourhood factors (e.g., cityPartRange) may still appeal to market perceptions. Prioritise properties in areas with favourable ratings.

Optimise Amenities Strategically: Amenities such as hasPool or hasStormProtector were less impactful, suggesting that adding these features may not significantly increase property value. Instead, prioritise features with proven ROI.

Focus on size and location-driven investments to maximise property value, while selectively enhancing properties with cost-effective amenities. This strategy aligns with data insights, guiding efficient allocation of resources for competition

**Conclusion**

This project analysed key factors influencing property prices in Paris, revealing that property size (squaremeters) is the most significant predictor of price, with neighbourhood characteristics and certain amenities playing a lesser role. By refining the model through stepwise selection, we achieved a streamlined, accurate predictive model. These insights support a targeted investment strategy that prioritises larger properties and strategically selected amenities to maximise value. This analysis provides a data-driven foundation for decision-making, enabling the property development company to optimise resources effectively in the competitive Paris housing market.