PARIS HOUSING PREDICTION

AKHILA SINGARAJU

**INTRODUCTION**

In this assignment, I analyzed property data to identify key variables that drive property prices, aiming to provide actionable recommendations for the business. Through data exploration, I assessed several features including property size, neighborhood attributes, and amenities using linear regression and backward elimination techniques. By examining the statistical significance of each variable and evaluating model performance through metrics such as R-squared and AIC, I identified the most impactful factors. This process allows me to determine where the business should focus its efforts to optimize property values effectively.

**DATA CLEANING**

I first examined the structure of the dataset to understand the available data and identify areas needing cleanup. The dataset includes various columns describing property features across Paris neighborhoods, such as hasYard, hasPool, numPrevOwners, isNewBuilt, and property dimensions like basementarea, atticarea, and squaremeters. The target column, price, represents the home price we aim to predict. Additionally, the dataset contains neighborhood identifiers, cityCode and cityPartRange, along with features like rooms, all of which provide insights into potential factors influencing property prices.

To address missing data, I first reviewed each column for missing values. I found that the floors column had over 50% missing values, making it difficult to fill in without risking accuracy. To maintain data integrity, I removed the floors column entirely, as filling such a large portion would likely distort the dataset. Similarly, I removed cityCode since it served primarily as an identifier without directly contributing to price prediction, streamlining the dataset without losing valuable insights. For columns with some missing data, like cityPartRange (about 18% missing) and hasGuestRoom (around 26% missing), I checked their distribution and found it to be nearly normal. This meant that both mean and median imputation were options.

I chose median imputation because it is unaffected by outliers, making it a more stable option than the mean for filling in missing values. Outliers can skew the mean, leading to an unbalanced dataset if mean imputation is used, whereas the median provides a more accurate representation of typical values. This was especially important for cityPartRange, which reflects neighborhood ratings in set ranges, and for hasGuestRoom, which shows the number of guest rooms. After imputing missing values, I assessed each column for skewness to determine if any adjustments were needed to create a more balanced distribution.

I then turned to handling outliers in numerical columns. Using the Interquartile Range (IQR) method, I identified outliers as values that fell significantly beyond the central range of the data. Columns like basementarea, atticarea, squaremeters, and rooms had noticeable outliers. Rather than removing these data points, I chose to cap them by setting upper and lower limits. This approach allowed me to adjust extreme values without reducing the dataset’s size or variability, preserving data richness while minimizing the influence of outliers on our analysis.

**EXPLORATORY DATA ANALYSIS**

**Hypothesis 1:** Newly built properties tend to have higher prices compared to older properties.

To test this hypothesis, I compared the average prices for new and old properties by grouping the data based on the isNewBuilt attribute. I then visualized the price distributions for each category using a violin plot.

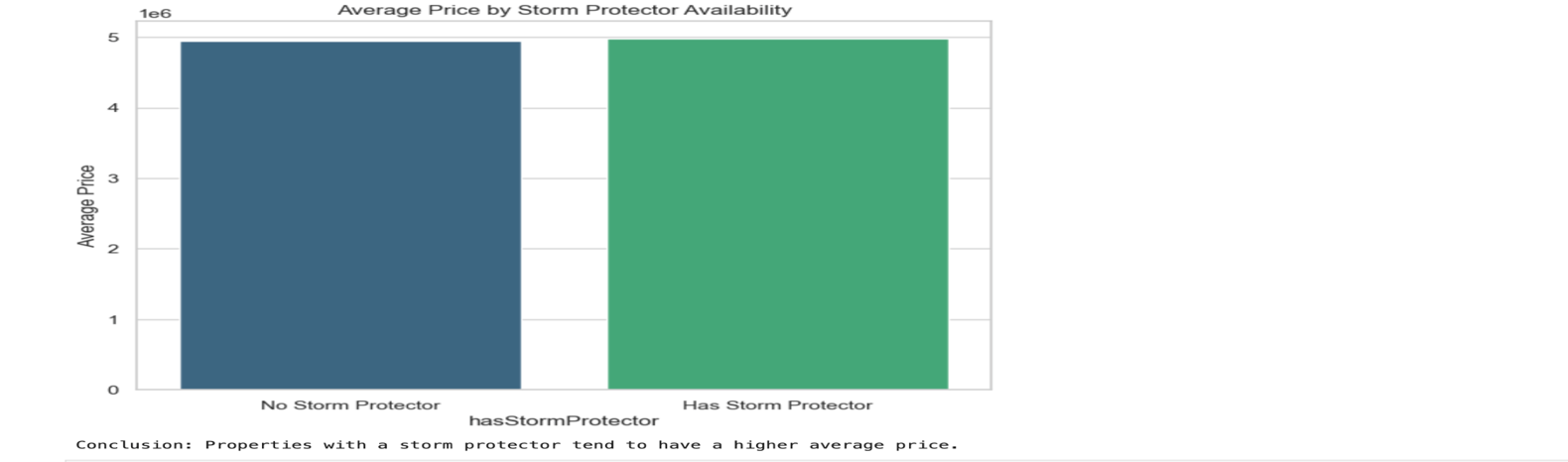
A diagram of a leaf

Description automatically generated with medium confidence

The plot illustrates that there is no significant difference in the price distribution between new and old properties, as the ranges overlap considerably. The data does not support the hypothesis, as being a new property does not necessarily lead to a higher price in this dataset. . Therefore, we conclude that being a new property does not necessarily lead to higher prices in this dataset.

**Hypothesis 2:** Properties with storm protectors have higher prices than those without.

Next, I investigated whether properties equipped with storm protectors tend to have higher prices. I grouped the dataset by the hasStormProtector column and calculated the average price for properties with and without storm protection.



Using a bar plot to visualize these averages, the results showed that properties with storm protectors indeed have a slightly higher average price compared to those without. This finding suggests that having a storm protector may add some value to a property, likely due to the perceived added safety and durability benefits. Thus, we conclude that properties with storm protectors tend to be valued higher on average.

**VARIABLE SELECTION**

For variable selection, I began by analyzing the correlation matrix to understand the relationships between each feature and the target variable, price. Correlation provided insights into which features had stronger or weaker relationships with price, guiding the selection process by highlighting variables with potential influence. Squaremeters has highest co-relation with price.

A screenshot of a computer screen

Description automatically generated

Following this, I calculated the Variance Inflation Factor (VIF) for the remaining features to assess multicollinearity, which can affect model stability. The initial VIF analysis revealed high values for several features, including squaremeters, rooms, basementarea, and atticarea, indicating strong multicollinearity among these size-related features. To address this, I retained only squaremeters as the primary indicator of property size because it had the highest correlation with price compared to the other size-related features. I removed rooms, basementarea, and atticarea to reduce potential multicollinearity issues. This adjustment lowered the VIF for squaremeters to below 10, enhancing model stability. I also removed numPrevOwners, garage, and hasGuestRoom after the VIF analysis because they had lower correlations with the target variable, price, and thus contributed less to predicting property prices. The final variables squaremeters, cityPartRange, hasYard, hasStormProtector, isNewBuilt, hasStorageRoom, and hasPool—included variables with both acceptable VIF values.

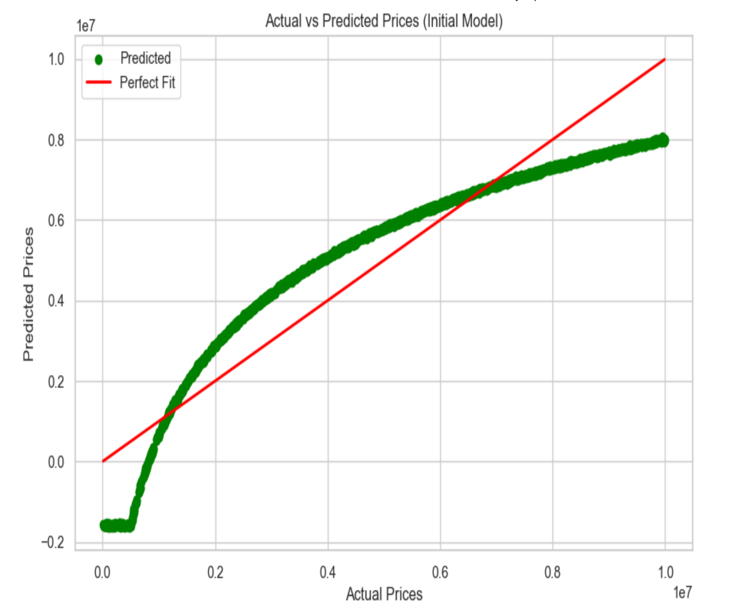
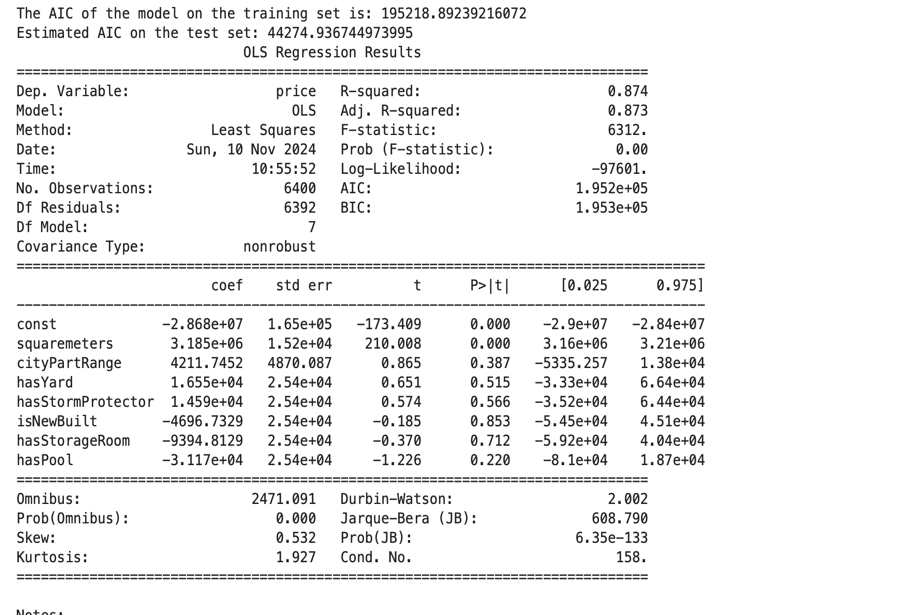
**MODEL DEVELOPMENT AND EVALUATION**

I chose linear regression in this case because it offers a simple, interpretable way to analyze the relationship between property features and price, allowing us to quantify the impact of each attribute on price. Given the continuous nature of the target variable (price) and our objective to identify key factors influencing it, linear regression is a suitable model to provide initial insights while keeping model complexity manageable.

INITIAL MODEL

The objective of this modeling process was to develop a multiple linear regression model that accurately predicts property prices based on a set of relevant features. I selected seven variables likely to influence price: squaremeters, cityPartRange, hasYard, hasStormProtector, isNewBuilt, hasStorageRoom, and hasPool. These features capture various property attributes, from size and location to amenities, providing a well-rounded basis for price prediction.

To begin, I split the data into training and test sets, using 80% of the data for model training and 20% for testing. I added a constant term to the dataset, as required by the statsmodels library, to account for the intercept in the regression model. I then fitted the model on the training data and used R-squared and Akaike Information Criterion (AIC) to evaluate its performance.



Upon examining the model’s coefficients, squaremeters emerged as the most significant predictor, with a statistically significant positive impact on price (p < 0.001). This aligns with expectations, as larger properties tend to have higher values. However, other features such as cityPartRange, hasYard, hasStormProtector, isNewBuilt, hasStorageRoom, and hasPool did not show significant effects, with higher p-values, suggesting that their influence on price is limited within this dataset. A plot of actual vs. predicted prices demonstrated a generally strong fit, though some deviations were observed at higher price levels, potentially indicating the need for further refinement or more complex modeling techniques. While the R-squared value is high, the insignificant coefficients for some variables suggest potential for model improvement.  
 **ANALYSIS OF VARIABLE SIGNIFICANCE**

| **Variable** | **P-value** |
| --- | --- |
| const | 0.000000 |
| squaremeters | 0.000000 |
| cityPartRange | 0.387170 |
| hasYard | 0.515158 |
| hasStormProtector | 0.566032 |
| isNewBuilt | 0.853422 |
| hasStorageRoom | 0.711760 |
| hasPool | 0.220373 |

* **const (Intercept)**: The p-value is 0.000, indicating that the intercept is statistically significant. This constant represents the base price when all other variables are zero, providing a starting point for interpreting the effect of other variables.
* **squaremeters**: With a p-value of 0.000, this variable is highly significant, confirming that property size has a strong and direct impact on price. A positive coefficient on squaremeters implies that larger properties are associated with higher prices, which aligns with expectations in the property market. For the property development company, focusing on increasing property size could be an effective strategy for boosting prices.
* **cityPartRange**: This variable has a p-value of 0.387, indicating it is not statistically significant at conventional levels (e.g., 0.05). Although we might expect more exclusive neighborhoods to command higher prices, the lack of significance here suggests that neighborhood exclusivity may not be a primary driver of price in this dataset.
* **hasYard**: With a p-value of 0.515, hasYard is not statistically significant, suggesting that whether a property has a yard does not meaningfully impact its price in this dataset. For the development company, this implies that prioritizing yard space may not necessarily lead to higher pricing.
* **hasStormProtector**: This feature has a p-value of 0.566, showing no significant effect on price. Despite potential perceived safety benefits, the presence of storm protection does not appear to command price premium here.
* **isNewBuilt**: The p-value for isNewBuilt is 0.853, making it statistically insignificant. This finding suggests that the age of a property does not have a strong effect on price, which could mean that, in this dataset, both new and older properties are valued similarly.
* **hasStorageRoom**: This variable has a p-value of 0.711, also showing no statistical significance. Storage rooms do not appear to contribute significantly to price, which could inform design choices if the development company is focused on features that increase value.
* **hasPool**: With a p-value of 0.220, hasPool is also statistically insignificant in this model. Although pools are often seen as luxury features, their addition does not necessarily lead to a significant increase in property price in this context.

In summary, **squaremeters** is the only significant predictor in this model, suggesting that property size is the key driver of price. This insight is valuable for the property development company, as it highlights the importance of focusing on property size rather than features like yards, pools, or storm protection, which do not show a significant impact on price in this dataset.

**FINAL MODEL – STEP-WISE SELECTION MODEL**

For the final model, I used a backward elimination process to select the most statistically significant features. Starting with all relevant features (`squaremeters`, `cityPartRange`, `hasYard`, `hasStormProtector`, `isNewBuilt`, `hasStorageRoom`, and `hasPool`). In each step, the model is evaluated, and the variable that contributes least to the model’s performance (e.g., the highest p-value or smallest AIC reduction) is removed. This process continues until only the most significant variables remain, leading to a more efficient and interpretable model.

A screenshot of a computer

Description automatically generatedA graph with a green line

Description automatically generated

The final model showed an R-squared value of 0.874, indicating a strong explanatory power, as it accounted for 87.4% of the variance in property prices. The AIC for the final model on the training set was 195210, slightly lower than the initial model, suggesting improved model efficiency. The estimated AIC on the test set was approximately 44263, confirming the model's stability when applied to new data. A plot comparing actual and predicted prices further demonstrated the model's performance, with predicted values closely aligning with actual values, particularly in the lower price ranges.

**COMPARISION**

| **Model** | **R-squared (Training)** | **R-squared (Testing)** | **AIC (Training)** | **AIC (Testing)** |
| --- | --- | --- | --- | --- |
| Initial Model | 0.874 | 0.870 | 195218.892 | 44274.937 |
| Final Model | 0.874 | 0.872 | 195210.032 | 44263.171 |

` Comparing the initial and final models, I observe that both have similar R-squared values on the training set (0.874), indicating they explain about the same proportion of variance in property prices. However, the final model slightly outperforms the initial model on the test set, with an R-squared of 0.872 compared to 0.870, suggesting a marginally better fit on unseen data. Additionally, the final model has lower AIC values on both training and testing data indicating a more efficient model. The initial model included additional variables like `cityPartRange`, `hasYard`, and `hasPool`, but backward elimination based on AIC retained only `squaremeters` in the final model. Despite this simplification, the final model achieves nearly the same R-squared value as the initial model, indicating that `squaremeters` alone explains most of the variance in property prices. The lower AIC in the final model further supports its efficiency, suggesting that property size is the primary driver of price, while other factors have minimal impact.

**BUSINESS FOCUS AND RECOMMENDATIONS**

Based on the analysis, I recommend focusing on two key variables to optimize property prices: squaremeters (property size) and hasPool (presence of a pool). The objective is to leverage these variables in property development and marketing to maximize property values. Property size emerged as the most significant predictor of price, while amenities like a pool, though less significant statistically, can enhance property appeal. By prioritizing these factors, the business can target areas that align with buyer preferences, helping to drive higher prices.

Recommendation:

Focusing on property size as a primary factor and the addition of appealing amenities like pools offers clear benefits to the business. Property size has a direct and substantial impact on price, with larger properties consistently commanding higher values. Additionally, while not statistically dominant, features like pools add to a property's perceived luxury and lifestyle appeal, making it more attractive in competitive markets. Emphasizing these attributes aligns with market demand and positions properties to meet buyer expectations for both space and premium amenities.

Impact Explanation:

Attention to property size and targeted amenities can significantly enhance the company’s value proposition. Larger properties not only attract a higher price point but also meet demand for spacious living, which can be a crucial differentiator in the market. The addition of sought-after amenities, such as pools, can further appeal to buyers looking for premium features, informing marketing strategies and attracting a more affluent customer base. This focus will guide investment in properties that maximize value and inform targeted marketing efforts, ultimately leading to increased profitability and a stronger market position.

**CONCLUSION**

From this analysis, I found that property size, measured by squaremeters, emerged as a primary driver of price, while other features, such as neighborhood exclusivity, showed limited impact within this dataset. By focusing on property size and possibly enhancing select amenities, the business can target the factors most likely to increase property values. This emphasis will allow the company to tailor development and marketing strategies to meet buyer demands, attract higher-paying clients, and ultimately boost profitability, aligning business decisions with data-driven insights.

**REFERENCES:**

1. GeeksforGeeks. (2024f, October 14). *Detecting Multicollinearity with VIF Python*. GeeksforGeeks. https://www.geeksforgeeks.org/detecting-multicollinearity-with-vif-python