

# NORTHEASTERN UNIVERSITY

**Student** – Sajal Gangrade

**Email** – gangrade.s@northeastern.edu

Module 2 – Paris Housing Prediction

**Class Number** - ALY 6020

**Class Name** – Predictive Analytics

**CRN Number** – 80249

# Professor – Justin Grosz

# Date – June 1, 2025

**2**

# Paris Housing Prediction

# Introduction

The purpose of this analysis is to assist a Paris-based property development company in understanding the key factors that influence residential property prices across different neighborhoods. With real estate representing one of the most substantial investments in urban development, the company seeks to use data-driven insights to guide its future investments, construction priorities, and marketing strategies.

To support this goal, a dataset containing property-level features—such as square footage, room count, amenities, and neighborhood exclusivity—was analyzed. The core objective was to build a robust multiple linear regression model that not only predicts property price but also reveals which specific property characteristics are most impactful. These insights will allow the company to make informed decisions on what types of homes to build, where to invest, and how to effectively price and position their offerings in a competitive market.

This report presents a step-by-step analysis, starting with data cleaning and exploratory analysis, followed by model development, variable significance interpretation, and comparison between modeling approaches. The analysis concludes with actionable recommendations designed to support real estate investment decisions and long-term strategy.

# Dataset Overview

The dataset used in this analysis contains detailed information on 8,000 residential properties across various neighborhoods in Paris. Each row represents a single property and includes attributes such as size, number of rooms, neighborhood exclusivity, presence of amenities (like a yard, pool, garage), and ownership history.

The main goal of this dataset is to help understand how these different features influence property prices. By analyzing this data, we can identify which attributes are most strongly associated with higher property values and develop a model that helps predict prices accurately. This understanding can guide property developers and investors in making more strategic decisions — whether it’s about marketing high-value features or identifying neighborhoods with greater return potential.

**3**

# Data Cleaning & Exploratory Data Analysis (EDA)

### Before building any predictive model, it is essential to ensure the dataset is clean, complete, and ready for analysis. Poor data quality can significantly impact model accuracy and interpretation. In this step, we focused on handling missing values, identifying appropriate encoding, and conducting hypothesis-driven EDA to understand feature relationships with the target variable (price).

### **Handling Missing Values**

### **We identified missing values in three columns:**

### **Floors – 4227 missing values (over 50%)**

### **cityPartRange – 1360 missing values**

### **hasGuestRoom – 2000 missing values**

### **A screenshot of a computer AI-generated content may be incorrect.**

### **Decision 1: Dropping floors**

* **Reason:** Over 50% of the data was missing in the floors column.
* **Justification:** Retaining a column with this much missingness would compromise the reliability of imputation and could introduce bias.
* **Action:** We dropped the floors column entirely from our dataset.

### **Decision 2: Imputing cityPartRange**

* **Reason:** This is a numeric feature representing the exclusivity of a neighborhood. Missingness here is likely random and not structurally tied to price.
* **Action:** We used **median imputation** instead of the mean to avoid influence from outliers.
* **Why Median?**: cityPartRange is a bounded ordinal feature (scale of 1–10), and median better preserves central tendency under skewed distributions.

### **Decision 3: Imputing hasGuestRoom**

* **Reason:** This column reflects the number of guest rooms, a count variable. It had 2000 missing entries.
* **Action:** We used **median imputation,** assuming the typical number of guest rooms is more representative than the mean (which can be inflated by few high values).

**4**

After these operations, we verified that the dataset had **no remaining missing values**

**A screenshot of a computer program

AI-generated content may be incorrect.**

### **Treatment of Categorical Variables**

Most binary variables in the dataset (like hasYard, hasPool, etc.) were already encoded as **0 and 1,** so no transformation was necessary.

We also reviewed the **distribution** of these variables to confirm balance and relevance.

### **Hypothesis-Driven EDA**

We tested two hypotheses to explore the relationship between features and property prices.

#### ****Hypothesis 1:** Properties with a Pool Are More Expensive**

* **Visualization Used:**
  + **Boxplot**: price vs. hasPool
  + **Barplot**: Mean price by pool availability

A graph of a price comparison

AI-generated content may be incorrect.

**5**

* **Finding:**  
  Properties with pools had **slightly higher average prices**, but the difference wasn't dramatic. The distribution was wide in both cases, indicating other factors also contribute significantly.
* **Conclusion:**  
  There is a **weak positive relationship**, but having a pool alone does not guarantee a higher price.

#### ****Hypothesis 2:** Homes in More Exclusive Neighborhoods (Higher** cityPartRange**) Are Priced Higher**

* **Visualization Used:**
  + **Boxplot** of price by cityPartRange
  + **Line plot** of average price across different cityPartRange values
* **Finding:**  
  We observed a **positive trend** in average prices as exclusivity increased. While not linear, the general direction supports the hypothesis.
* **Conclusion:**  
  This hypothesis is **supported**, making cityPartRange a strong candidate for inclusion in the regression model.

**A graph and chart of a property price

AI-generated content may be incorrect.**

**6**

# Model Development and Evaluation

# The main goal of this step was to build a Multiple Linear Regression model that could explain and predict housing prices in Paris using various features. We aimed not only for predictive power but also for clarity — helping the real estate company understand what drives price differences.

**1. Model Setup**

We used the cleaned dataset, where all missing values had been handled and categorical variables were appropriately encoded.

* **Target Variable**: price
* **Predictor Variables**: All remaining columns except price
* **Regression Method**: Ordinary Least Squares (OLS)

We added a constant term to account for the intercept and fitted the model using the stats model’s library in Python, which allows access to p-values, confidence intervals, and AIC values for model evaluation.

### **2. Key Evaluation Metrics**

#### ****R-squared (R²)****

* Our model achieved an **R² of 0.756**, indicating that approximately **75.6% of the variability in housing price** is explained by the included features.
* This is considered a **strong model fit** for real-world property data, which often contains noise due to location, historical factors, or unobserved features.

#### ****Akaike Information Criterion (AIC)****

* The model's **AIC was 249,300.** AIC is a relative measure used to compare models — lower values are better.
* This value serves as a benchmark when we compare it with the **stepwise-selected model** in the next section.

### **3. Observations from the Initial Model**

* The only strongly significant predictor (p-value < 0.05) was squaremeters (total area of the home), which had a **large positive coefficient.**
* Most other features (like hasYard, garage, hasPool, etc.) had high p-values, indicating they may not be statistically significant in this model.
* This doesn't mean those features are unimportant, but it does suggest **they may not explain price variation well when considered alongside other features.**

### **4. Visual Diagnostic Consideration**

* **Multicollinearity Check (VIF):** All predictor VIF values were **close to 1,** suggesting no multicollinearity issues. This gave us confidence in the model's stability.
* **Residuals and Assumptions:** Though not included in the model code, we visually examined residual plots to verify assumptions of **normality and constant variance,** which appeared reasonably met.

**7**

# Analysis of Variable Significance

Once the regression model was built, we analyzed the impact of each variable using the model's coefficients and their respective p-values. This step helped us determine which features significantly influence housing prices in Paris and which do not.

#### Interpreting Significance Using p-values

In our model, a p-value less than 0.05 generally indicates a variable is **statistically significant** — meaning it has a measurable impact on the target variable (in this case, property price). Conversely, higher p-values suggest that any observed impact may be due to random chance.

From the output of the **full linear regression model**, we observed the following:

* **Statistically Significant Variable:**
* **Square meters (p < 0.001)** — This was by far the most impactful predictor. The model shows that as the size of the property increases, so does its price. This aligns with real-world expectations: larger properties tend to command higher market values.
* **Other Variables (e.g., hasYard, hasPool, cityCode, isNewBuilt)** had high p-values (> 0.05), indicating they were **not statistically significant** in this model. Their contribution to predicting price was minimal or inconsistent when controlling for other factors.

#### Real-World Interpretation

The significance of **square meters** underscores its central role in pricing decisions. For developers or investors, this suggests that property size is the most reliable indicator of market value across different neighborhoods — regardless of the presence of other features like pools or storage rooms.

This also means that **amenities alone don't drive price** in Paris — size trumps other characteristics unless combined with location or luxury indicators, which may not be fully captured by the available dataset.

**Optional Analysis: Log-Transformed Model (Brief Mention)**

To further validate our results and explore model stability, we also ran a **log-transformed regression model** by applying a natural log to the price variable (log\_price). This transformation is commonly used to stabilize variance and improve interpretability in skewed price distributions.

* **Result**: The model yielded an R-squared of 1.000, indicating a nearly perfect fit — but this is likely an overfit to the data due to the transformation and scale of the dataset.
* **Observation**: While technically strong, the log model did not offer additional meaningful insight compared to the original linear model.
* **Conclusion**: Since this analysis goes beyond the assignment scope, we included it only as a side check — our final recommendation is still based on the original linear regression model and stepwise comparison.

# 8

# Comparison with Stepwise Selection Model

### **Objective**

The property development company is looking to identify which features most strongly influence home prices in Paris, enabling them to make more informed investment, development, and marketing decisions.

### **Key Takeaways from the Model**

After evaluating both the full regression model and the stepwise selection model (based on AIC), we observed that:

* **Square Meters** was consistently the most impactful predictor, with a highly significant coefficient.
* **CityPartRange** (a proxy for neighborhood exclusivity) also showed a positive correlation with price, aligning with business intuition, although its significance level was moderate.

These two variables not only make logical sense — larger homes and those in more exclusive areas command higher prices — but they also held their importance across both models and maintained low multicollinearity scores (confirmed via VIF analysis).

### **Final Recommendation**

We recommend the company focus on these two key variables:

|  |  |
| --- | --- |
| Variable | Why It Matters |
| Square Meters | A direct and statistically strong driver of pair. Larger properties consistently yield higher returns |
| CityPartRange | Represents exclusivity and neighborhood desirability. Higher values correlate with increased property prices. |

By **targeting larger homes in more exclusive neighborhoods**, the company can:

* Maximize profit margins on property investments.
* Tailor development efforts toward premium markets.
* Position marketing around space and location — two factors buyers are consistently willing to pay more for.

**9**

# Conclusion

This analysis aimed to identify the key factors influencing housing prices in various neighborhoods of Paris. Through a structured process of data cleaning, exploratory data analysis, and regression modeling, we were able to uncover both statistically significant variables and practical insights for business decision-making.

We ensured that the dataset was of high quality by handling missing values appropriately, verifying the structure of categorical variables, and confirming the absence of multicollinearity among predictors. Using multiple linear regression, we found that the **square meters of the property** had the strongest and most consistent relationship with price. Additionally, our comparison with a stepwise selection model reinforced the robustness of this predictor and helped confirm that no unnecessary variables were inflating the model’s complexity.

While we briefly explored a log-transformed model to address skewness and compare performance, our focus remained aligned with the core objective: building a model that is interpretable and actionable for the business. The findings ultimately suggest that by prioritizing larger homes and properties in more exclusive neighborhoods, the development company can strategically optimize pricing and marketing investments.

This report not only delivers a predictive framework for estimating housing prices but also offers data-backed recommendations to support future development planning and business growth in the Paris real estate market.

**10**

# References

1. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: With applications in R (Vol. 103). Springer. <https://doi.org/10.1007/978-1-4614-7138-7>
2. Seaborn. (n.d.). Seaborn documentation. Retrieved May 31, 2025, from <https://seaborn.pydata.org/>
3. Statsmodels. (n.d.). Statsmodels: Statistical modeling and econometrics in Python. Retrieved May 31, 2025, from <https://www.statsmodels.org/>
4. McKinney, W. (2022). Pandas documentation. Retrieved May 31, 2025, from <https://pandas.pydata.org/>
5. VanderPlas, J. (2016). Python data science handbook: Essential tools for working with data. O’Reilly Media.
6. Paris Housing Dataset. (n.d.). Provided by the instructor through course materials on the Northeastern University portal.