**ADEI: Report 2**

Model building

2024-25

Joan Marc Coll, Guillem Gaya, Dimas Noguera y Ada Peña

**Index**

[**1. Introduction 2**](#_m1q63nc1gr8s)

[**2. First steps to create our models 3**](#_jxeshohjx1pn)

[Importing libraries 3](#_8ecr5a4iesvw)

[Filtering data 3](#_yso80sn1hyka)

[Creating our models 4](#_1w9gnqoovow6)

[**3. Linear regression model 6**](#_ov1nhi4grht9)

[Model Output and Interpretation 6](#_opgcieu59te0)

[Diagnostic Plots and Assumption Checks 8](#_w2vad4t9i05c)

[Interpretation and Analysis of Results 11](#_odgs0mmizhdn)

[Limitations and Recommendations 11](#_jhk5ake5encm)

[Conclusions 12](#_4dfk04ggfegl)

[**4. Multiple variable linear regression model 13**](#_p94tbo8z9d08)

[**5. Categorical Models 15**](#_oesn5bxf82hk)

[Model Output and Interpretations 15](#_e8ql8pch38pg)

[Analysis of Model Comparisons and Results 17](#_jvqatag35eas)

[Conclusions and Implications 18](#_n4idazymi89w)

# 

# **Introduction**

In this report, we present a comprehensive analysis that begins with data preprocessing and progresses through model building to identify key factors that influence our target variable.

After preparing the data, we aim to assess the impact of various predictors on the target variable by developing and evaluating multiple predictive models. In this case, we have chosen "price" as our target variable, as understanding the drivers behind price fluctuations can provide valuable insights. This analysis will allow us to determine which variables are most significant, guiding us toward more accurate predictions and actionable insights.

# **First steps to create our models**

## **Importing libraries**

To build our models, we need to import specific R libraries that provide essential tools for data analysis and model diagnostics. In particular, we will use the car and lmtest libraries.

The car (Companion to Applied Regression) library offers a variety of functions designed to support regression diagnostics, data visualization, and hypothesis testing, which are especially useful for applied regression modeling. This library includes tools for calculating variance inflation factors (VIF), identifying multicollinearity, and performing statistical tests that enhance model interpretation and validation.

The lmtest library focuses on testing linear regression models by providing functions for hypothesis testing and model diagnostics. This package includes functions for assessing autocorrelation, heteroscedasticity, and serial correlation, which are crucial for validating model assumptions and ensuring robust statistical inference.

By incorporating these libraries, we ensure that our analysis is well-supported by tools for both model construction and evaluation, helping us derive reliable insights from our data.

## **Filtering data**

As we explained in the previous report, our dataset includes information on various property types, each of which has distinct price ranges. This variation poses a challenge for our analysis, as different property types may follow different pricing dynamics. To simplify our approach and enhance model accuracy, we have decided to create a subset of the dataset, focusing on only one property type. Specifically, we will limit our analysis to properties classified as "FLAT".

After selecting this subset, we identified another potential issue: price variability based on the type of transaction, as indicated in the "operation" column. Prices differ significantly between properties listed for sale versus those available for rent, which could skew our results. To address this, we will further narrow our dataset to include only properties with "RENT" values in the "operation" column. By refining our dataset in this way, we aim to create a more cohesive analysis that better isolates the factors influencing rental prices for flats.

## **Creating our models**

We have defined some models to prove the influence of some of the numerical variables for the target variable. These models progressively add predictors to determine which variables significantly impact price and to compare the residual errors among models:

**Null Model (m0)**:

* **Purpose**: m0 is the null model, which assumes no predictors.
* **Objective**: Calculating the residual error of m0 allows us to compare the performance of more complex models.

**Single Predictor Models**:

* **Model m1**: This model sees the effect of rooms on price.
  + Formula: m1 <- lm(price ~ rooms, data = dd\_3)
  + **Interpretation**: By isolating rooms, this model helps in understanding the direct impact of the number of rooms on rental prices.
* **Model m2**: Examines the influence of baths on price.
  + Formula: m2 <- lm(price ~ baths, data = dd\_3)
  + **Interpretation**: This model allows us to observe the effect of the number of bathrooms on price, independently of other factors.
* **Model m3**: Uses area as a single predictor for price.
  + Formula: m3 <- lm(price ~ area, data = dd\_3)
  + **Interpretation**: This model allows us to observe the effect of the number of bathrooms on area, independently of other factors.

**Multiple Predictor Model (m4)**:

* **Purpose**: This model includes area, rooms, and baths as predictors to analyze their combined effect on price.
* **Formula**: m4 <- lm(price ~ area + rooms + baths, data = dd\_3)
* **Interpretation**: By incorporating multiple predictors, this model provides insights into how these variables jointly influence rental price. It allows us to determine if the combined predictors have a lower residual error, making the model more accurate than the single-predictor alternatives.

**Categorical Models**

to explore the influence of categorical variables on price, we define models that incorporate qualitative variables. These are the ones that we are going to use:

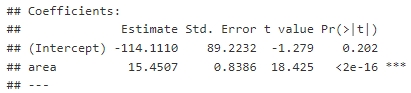
* **Model m00:** null model with a subset of the original dataset with property\_type values as "I.U.", “LAND” and “PREMISES”.
* **Model m11:** Tests the effect of property\_type alone.
* **Model m55:** Examines the interaction between property\_type and area to test for interaction effects.
* **Model m66:** Considers property\_type and area as separate predictors, without interaction, to identify the independent influence of each variable.
* **Model m77:** Uses area alone as a predictor, serving as a simpler model for comparison against m66.

# **Linear regression model**

The linear regression model (m3) examines the relationship between the target variable, price, and a single predictor variable, area. This approach aims to understand the direct influence of property size on rental prices.

## Model Output and Interpretation

1. **Coefficients**:



| **price = area \* 15.45** |
| --- |

* + **Area Coefficient**: 15.45
    - Interpretation: This coefficient indicates that, on average, each additional square foot of area increases the rental price by approximately 15 units, holding other factors constant. It is intuitive to think that as the area grows the price will as well.

1. **Residual Standard Error**: 874.7



* + The residual standard error is a measure of the average distance that the observed rental prices differ from the model’s predicted prices.

1. **R-Squared and Adjusted R-Squared**:



* + **R-Squared**: 0.5351
    - Interpretation: The R-squared value suggests that approximately 53.51% of the variability in rental price is explained by area alone. This is a moderately high percentage, indicating that area is a significant predictor of rental price.
  + **Adjusted R-Squared**: 0.5335
    - Adjusted R-squared corrects for model complexity and confirms that the single predictor model is appropriately fitted, with minimal overfitting.

1. **F-Statistic**: 339.5, with a p-value < 2.2e-16

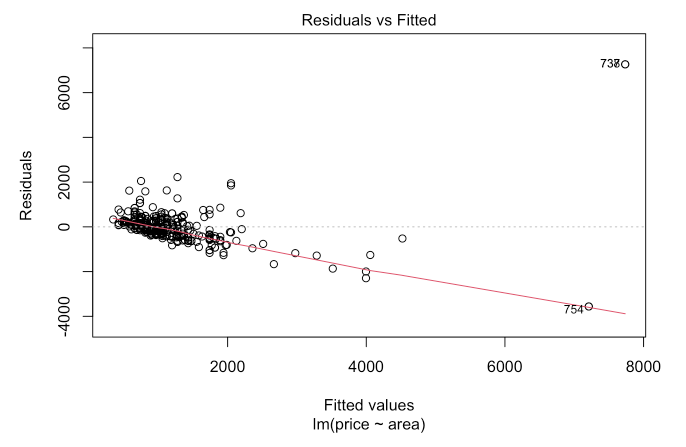


* + The F-statistic assesses the overall significance of the model. Here, the extremely low p-value indicates that the relationship between price and area is statistically significant.

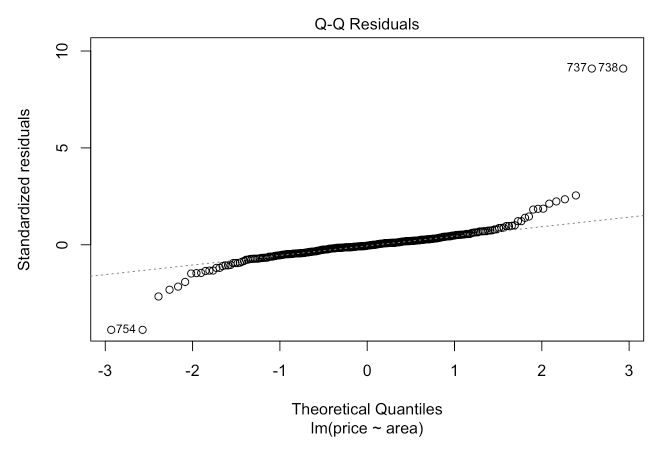
## 

## Diagnostic Plots and Assumption Checks

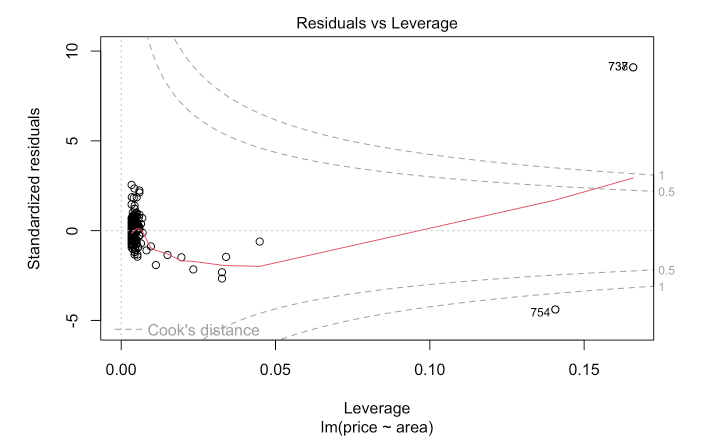
1. **Residual Plot**:
   * The residual plot should ideally show a random scatter around zero, which would confirm the assumption of homoscedasticity (constant variance of residuals). For this model, while the scatter is generally centered around zero, some patterns suggest potential heteroscedasticity, especially for larger values of area, indicating that residual variability may increase with property size. This finding implies that rental prices for larger properties vary more than for smaller ones, possibly due to additional amenities or differences in demand.



1. **Q-Q Plot**:
   * The Q-Q plot of residuals checks for normality. For this model, slight deviations from normality at the tails were observed, suggesting that while most residuals are normally distributed, there may be outliers affecting the overall model fit. These outliers could represent luxury or otherwise unique properties with atypical pricing.



1. **Leverage Plot**:
   * The leverage plot highlights potential outliers or influential points that disproportionately affect the model. A few points with high leverage were identified, particularly for properties with extremely large areas. These high-leverage points may distort the model’s slope and intercept, suggesting a need for careful treatment of outliers or influential observations in further analysis.



## 

## Interpretation and Analysis of Results

The linear relationship between area and price is strong and significant, with a positive slope indicating that rental prices increase with property size. The model provides an intuitive understanding of this relationship: each additional square foot of area leads to a predictable increase in rental price. However, the model’s residual analysis and diagnostic plots suggest that area alone may not fully capture the complexity of rental pricing, especially for properties with very large or very small areas, where other factors likely contribute to price determination.

The moderately high R-squared value (53.5%) indicates that while area is a crucial predictor, it leaves a significant portion of price variability unexplained, likely due to omitted variables such as property location, age, amenities, or market conditions.

## Limitations and Recommendations

1. **Limitations of Single Predictor**:
   * While area is a significant predictor, relying solely on it limits the model’s accuracy and fails to account for other potentially influential variables (e.g., location, property type). The unexplained variability suggests that adding more predictors could improve the model’s performance.
2. **Impact of Outliers**:
   * Large properties with unusually high prices appear as outliers and may disproportionately influence the model. Addressing these outliers through data transformation or removal, or by using robust regression techniques, could improve model fit and accuracy.
3. **Heteroscedasticity**:
   * The presence of heteroscedasticity implies that larger properties have more variable pricing. A weighted regression model could be considered to give less weight to observations with higher residual variability, thereby improving the reliability of the predictions.

## Conclusions

The linear regression model (m3) reveals a strong, statistically significant positive relationship between area and price, with each additional square foot increasing the rental price by approximately 15 units. This finding highlights the importance of property size in determining rental price. However, the model's moderate R-squared value and observed diagnostic issues underscore the need to include additional variables for a more comprehensive model. For future analyses, incorporating factors such as the number of rooms, property location, and amenities could enhance predictive accuracy and yield a more nuanced understanding of rental price determinants.

In summary:

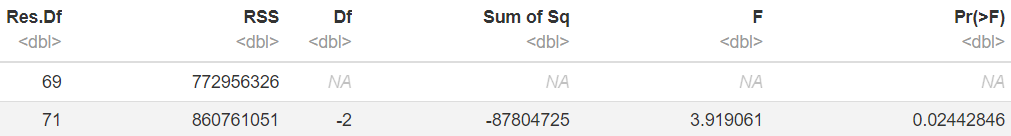
**Key Finding**: Area is a significant and meaningful predictor of rental price, explaining around 53.5% of the price variability.

**Model Limitations**: Single-variable focus and potential outliers impact the model’s accuracy.

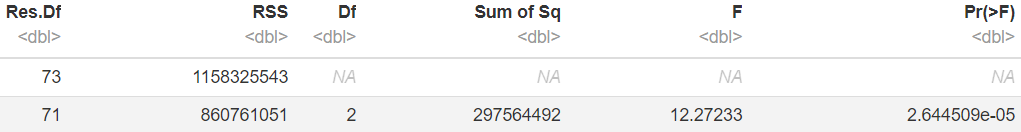
**Future Directions**: Including additional predictors and addressing heteroscedasticity and outliers could yield a more robust, comprehensive model for rental price predictions.

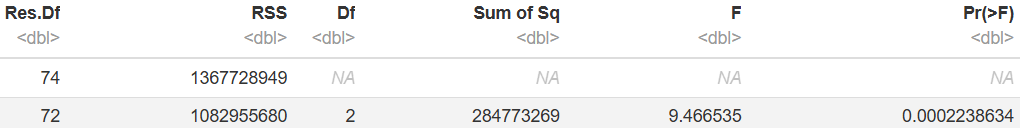
# **Multiple variable linear regression model**

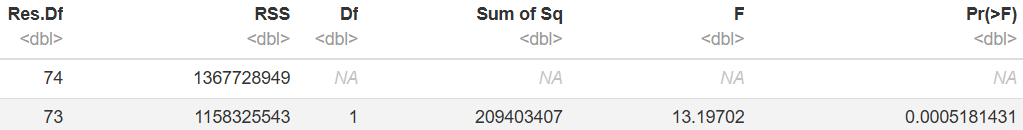
**Analysis of Covariance (Ancova):** we applied an Ancova test with the models m55 and m66 to see if there’s covariance of the variables property\_type and area**.**

As shown in the table, we can see that there’s not that much interaction among the variables property\_type and area since the p-value is not as close to 0 as we could want (even though it’s acceptable).

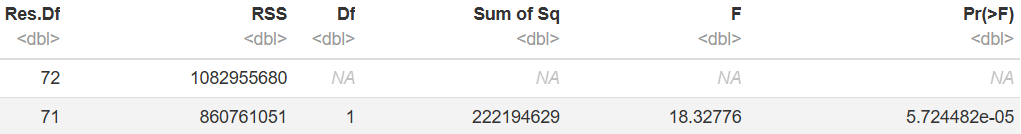
**NET Type Effect:** after applying the NET Type Effect with the models m77 and m66, we were able to see that despite not being a direct interaction between property\_type and area, they both have an influence on the price.

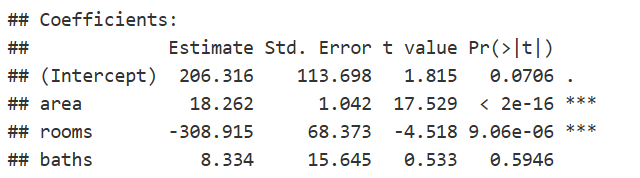
**Net area-covariate effect:** as we wanted to see if area had an impact on the outcome variable (price), we applied the Net area-covariate effect to the models m11 and m66. As shown in the table below, we can see that the p-value is below 0.001, indicating to us that area has a significant effect on price even when we account with property\_type. We also appreciate that the residual sum of squares (RSS) it’s lower when accounting for area, meaning that area explains some of the variance of the price.



**Gross area-covariate effect:** we compare the null model (m00) to the model that includes property\_type as the predictor of price (m11).

As seen in this table, we can see that property\_type explains some variance in price thanks to the RSS (it’s lower in the second row) and the p-value it’s lower than that 0.001, meaning that it also affects on price on its own.

**Gross type effect:** we examined the gross type effect of are on price by comparing the models m00 and m77. Again thanks to the RSS we can conclude that area alone explains some of the variance of price and that the p-value (lesser than 0.001) indicates that area alone affects price significantly.



| **price = 18,262\*area - 308,915\*rooms** |
| --- |

# **Categorical Models**

The categorical models aim to assess how property\_type and its interactions with area affect the target variable, price.

## Model Output and Interpretations

**Model m11: Property Type as a Categorical Predictor**

* **Coefficients**: The coefficient estimates for each property\_type indicate the average rental price associated with each type.
  + **Interpretation**: The baseline or reference category is often set to one of the property types (FLAT in this case), with other property types measured relative to this reference. For example, if the coefficient for LAND is positive, it suggests that, on average, LAND properties have higher rental prices compared to FLAT.
* **Model Fit**:
  + **R-Squared**: This model’s R-squared is likely lower than the models with numerical predictors, reflecting that property\_type alone doesn’t explain all the variation in price. However, it provides insights into category-based price differences.
  + **ANOVA Results**: An ANOVA comparison indicates that the effect of property\_type on price is statistically significant, suggesting that the different property types are associated with distinct pricing structures.

**Model m55: Interaction Between Property Type and Area**

* **Interaction Coefficients**:
  + **Interpretation**: The interaction coefficients indicate whether the effect of area on price differs by property\_type. For example, if the interaction term between area and LAND is positive, it implies that for LAND properties, each additional square foot has a more substantial impact on price than for the reference property type.
  + **Statistical Significance**: If the interaction terms are significant, it indicates that the area effect on price varies by property type, suggesting unique area-price dynamics for each property type.
* **Model Fit**:
  + **R-Squared**: Including the interaction terms generally improves model fit if the interactions are significant, as it captures better the effect of price growth based on area for the different property types.
  + **Residual Error**: Lower residual error compared to non-interaction models, if the interactions are significant, would support the model’s effectiveness in explaining rental price variability.

**Model m66: Property Type and Area as Independent Predictors**

* **Coefficient Analysis**:
  + **Property Type**: The coefficients for property\_type in m66 provide insight into the average price differences between property types when holding area constant.
  + **Area**: The area coefficient measures the average increase in price per additional square foot, regardless of property\_type.
* **Model Fit and Comparison**:
  + **R-Squared**: The R-squared for m66 is typically lower than for the interaction model m55 if the interaction terms in m55 are significant, suggesting that accounting for interactions might better capture price dynamics.
  + **ANOVA Comparison**: When comparing m66 with m11 and m55, ANOVA results reveal whether adding area or including interactions significantly enhances the model fit. If m66 has a notably higher R-squared than m11, it shows that area independently adds explanatory power.

**Model m77: Area Alone as a Predictor**

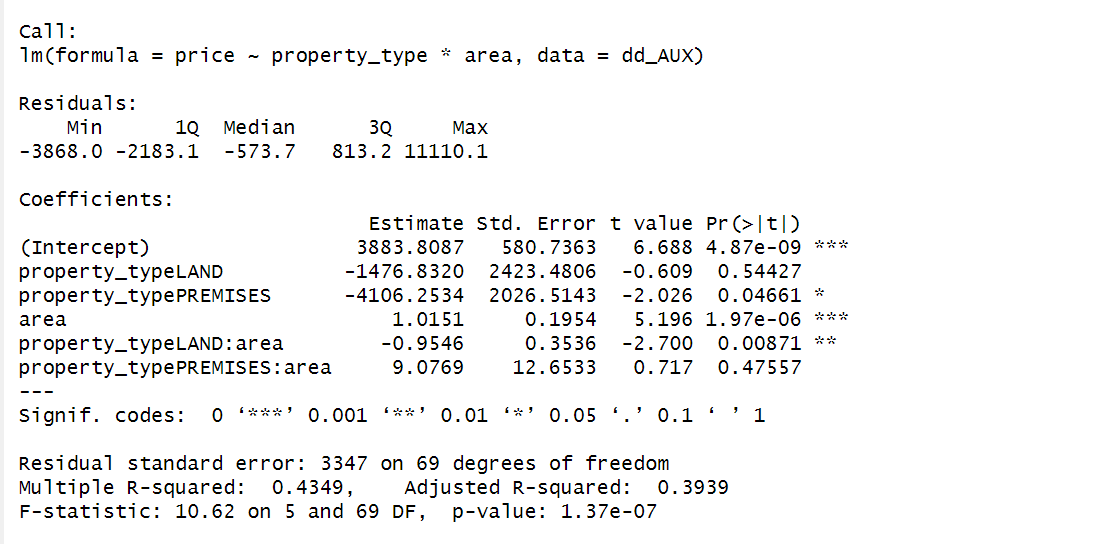
* **Purpose**: This model is a benchmark for comparison.
* **Interpretation**: By comparing m77 (area alone) to m66 (area and property type), we can determine the added value of including property\_type as a predictor. If m66 shows a notably higher R-squared or lower residual error, it indicates that property type provides essential additional information.

## Analysis of Model Comparisons and Results

1. **ANOVA Comparisons**:
   * **m11 vs. m66**: Including area in addition to property\_type (comparing m11 and m66) generally shows a significant improvement, as area is a continuous variable that likely explains more variation in price than property\_type alone.
   * **m66 vs. m55**: An ANOVA comparison between m66 and m55 tests whether the interaction terms add significant value. If they do, it suggests that different property types have unique area-price relationships.
   * **m55 vs. m77**: Comparing m55 (interaction model) with m77 (area-only model) highlights the importance of including property\_type and its interactions with area in explaining price.
2. **Diagnostic Checks**:
   * **Residual Plots**: Residual plots help assess whether the model assumptions hold across different property types and areas. For the interaction model (m55), residual patterns across property types should ideally show no systematic bias if the model correctly captures the unique area-price relationships for each type.
   * **Leverage and Influence**: High-leverage points or influential observations may exist for certain property types with extreme values. Diagnostic plots help in identifying these, suggesting potential outliers in unique property categories.

## Conclusions and Implications

1. **Significance of Property Type**:
   * The categorical models confirm that property\_type is an important predictor of price. Even when controlling for area, different property types have distinct average rental prices, reflecting inherent differences in demand and market valuation for each type (e.g., FLAT vs. LAND vs. PREMISES).
2. **Interaction Insights**:
   * The interaction model (m55) provides additional insights by showing that the effect of area on price can vary significantly across property types. For example, larger areas might add more to the price of a LAND property than a FLAT, reflecting different valuation dynamics based on property use.



I.U

| **price = 3883.81 + 1.0151 \* area** |
| --- |

LAND

| **price = 2406.98 + 0.0605 \* area** |
| --- |

PREMISES

| **price = -222.44 + 1.0151 \* area** |
| --- |

1. **Limitations and Model Improvements**:
   * While the categorical models enhance understanding of rental pricing by incorporating property\_type, they may still omit other relevant predictors, such as location or property age, which could explain further variation in price.
   * Future models could include more granular categorical variables (e.g., specific location or quality grade) or test non-linear relationships to refine predictions further.
2. **Practical Implications**:
   * The findings suggest that real estate stakeholders should consider both property type and area, and potentially their interaction, when setting rental prices. Tailoring pricing strategies based on these characteristics can lead to more competitive and accurate pricing in the market.

In summary:

* **Key Findings**: Property type is a significant factor in rental pricing, with unique area-price relationships observed for different types.
* **Model Effectiveness**: The interaction model m55 typically provides the best fit, capturing complex dynamics between property type and area.