**ADEI: Spain Real Estate Analysis**

Reports 1,2 & 3

2024-2025

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

**Index**

[**1. Preprocessing 3**](#_ospwnqv0yrii)

[**1.1 Data Source 3**](#_f9fi31f3bvuw)

[**1.2 Data Description 3**](#_25djlq2nkde3)

[1.2.1 Description of Variables 3](#_3xsclg2qf7wo)

[1.2.2 Descriptive Statistics of Variables 5](#_gs99yx8c0uls)

[1.2.3 Frequency of Variables 10](#_6giq3v321ovz)

[**1.3. Data Preprocessing 12**](#_4x21r0gnlb1v)

[1.3.1 Data Renaming 12](#_s5nxw59ld6ip)

[1.3.2 Handling Missing Data 13](#_1fiidx3u73vj)

[1.3.3 Changing Variable Type 14](#_axi8s02f5rbh)

[**1.4. Description of the preprocessed dataset 15**](#_st89rqlof6a)

[1.4.2 Descriptive Statistics of Preprocessed Variables 16](#_vf5ngui6rhn1)

[1.4.3 Frequency of Preprocessed Variables 18](#_dkhpux12pda)

[**1.5. Plot Comparison 20**](#_sch05gz0omd)

[**2. Model Building 22**](#_m1q63nc1gr8s)

[**2.1 First steps to create our models 22**](#_jxeshohjx1pn)

[2.1.1. Importing libraries 22](#_8ecr5a4iesvw)

[2.1.2 Filtering data 23](#_yso80sn1hyka)

[2.1.3 Creating our models 23](#_1w9gnqoovow6)

[**2.2 Linear regression model 25**](#_ov1nhi4grht9)

[2.2.1 Model Output and Interpretation 25](#_opgcieu59te0)

[2.2.2 Diagnostic Plots and Assumption Checks 27](#_xixh3a2n88xe)

[2.2.3 Interpretation and Analysis of Results 30](#_odgs0mmizhdn)

[2.2.4 Limitations and Recommendations 30](#_jhk5ake5encm)

[2.2.5 Conclusions 31](#_4dfk04ggfegl)

[**2.3 Multiple variable linear regression model 32**](#_p94tbo8z9d08)

[**2.4 Categorical Models 34**](#_oesn5bxf82hk)

[2.4.1 Model Output and Interpretations 34](#_e8ql8pch38pg)

[2.4.2 Analysis of Model Comparisons and Results 36](#_jvqatag35eas)

[2.4.3 Conclusions and Implications 37](#_n4idazymi89w)

[**3. PCA, Clustering and Profiling 40**](#_s4visn9ql6aa)

[**3.1 Principal Component Analysis 40**](#_akaes8w95u5i)

[3.1.1 Objective 40](#_bttr6hpsekoi)

[3.1.2 Data Preparation 40](#_lww04m8nzw20)

[3.1.3 Results 40](#_ugo4r3ecw4ka)

[**3.2 Clustering 44**](#_jd22v822f6vi)

[3.2.1 Objective 44](#_g6g5a7o4o8dc)

[3.2.2 Methodology 44](#_7xkcx0gnt4ee)

[3.2.3 Results 44](#_6y15jzmnsyd1)

[3.2.4 Interpretation 46](#_v06a75m2xknz)

[**3.3 Profiling 47**](#_j3a9ev5f9b9d)

[3.3.1 Objective 47](#_w2tekicdi905)

[3.3.2 Methodology 47](#_7o1lz6qkdhlz)

[3.3.3 Results 47](#_bwyquq9g5ccp)

[3.3.4 Graphs and additional analysis 47](#_y81noxdgfnpj)

[3.3.5 Conclusions 49](#_dbuh36qc0kks)

[**3.4 Conclusions 49**](#_mbkgn2gis9xu)

# 

# 

# **1. Preprocessing**

## 1.1 Data Source

The data for our study is available at the following link provided by Data Market:  
<https://datamarket.es/media/samples/inmuebles-sample.csv>

The primary objective is to predict the type of operation ("operation") of the properties.

## 1.2 Data Description

The selected dataset contains 1000 instances and 24 variables that provide information about different types of properties.

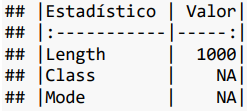
### 1.2.1 Description of Variables

* **website**Variable type: "character".  
  Description: Website where the property is advertised.
* **url**Variable type: "character".  
  Description: Web link to the property advertisement.  
  Additional information: Initially, this variable does not provide useful information as each instance will be unique, and it does not provide any information about the property.
* **reference**Variable type: "character".  
  Description: Advertisement identifier.
* **country**Variable type: "character".  
  Description: Country code of the property.  
  Additional information: 100% of the instances have the same value "ES," meaning all the properties are from Spain. Therefore, this variable does not provide meaningful information for the current dataset unless properties from other countries are added.
* **province**Variable type: "character".  
  Description: Province where the property is located.
* **location**Variable type: "character".  
  Description: Location of the property (street, avenue, etc.).
* **title**Variable type: "character".  
  Description: Title of the property advertisement.
* **description**Variable type: "character".  
  Description: Description of the property advertisement.
* **price**Variable type: "numeric".  
  Description: Property price in euros.
* **operation**Variable type: "character".  
  Description: Type of operation on the property (sale or rent).  
  Additional information: This is the target variable of our dataset.
* **property\_type**Variable type: "character".  
  Description: Type of property (shops, land, etc.).
* **rooms**Variable type: "integer".  
  Description: Number of bedrooms in the property.
* **baths**Variable type: "integer".  
  Description: Number of bathrooms in the property.
* **area**Variable type: "integer".  
  Description: Property area in square meters.
* **floor**Variable type: "integer".  
  Description: Number of floors in the property.
* **elevator**Variable type: "logical".  
  Description: Whether the property has an elevator or not.
* **outside**Variable type: "logical".  
  Description: Whether the property has an outdoor space or not.
* **floor.1**Variable type: "integer".  
  Description: Number of floors in the property.  
  Additional information: This variable is a duplicate of the "floor" variable.
* **images**Variable type: "integer".  
  Description: Number of images of the property on the advertiser's website.
* **latitude**Variable type: "numeric".  
  Description: Latitude of the property.
* **longitude**Variable type: "numeric".  
  Description: Longitude of the property.
* **dealer**Variable type: "character".  
  Description: Name of the dealer (companies).
* **dealer\_url**Variable type: "character".  
  Description: Web link to the dealer.
* **dealer\_is\_professional**Variable type: "character".  
  Description: Indicates whether the dealer is a professional or not.

### 1.2.2 Descriptive Statistics of Variables

We do not show descriptive statistics for all non-numeric variables, as they are all similar to the example shown below for “website.”

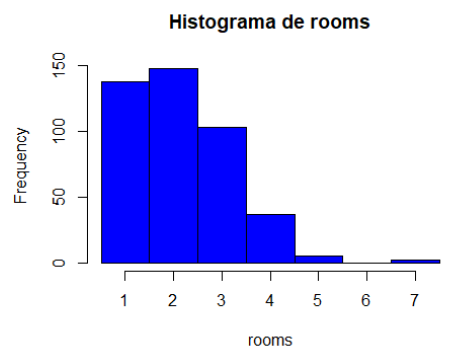
* **website**

****

* **price**

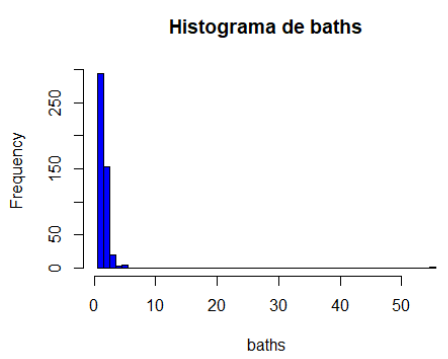
| **Statistical** | **Value** |
| --- | --- |
| Min. | 374.0 |
| 1st Qu. | 1200.0 |
| Median | 36500.0 |
| Mean | 156967.6 |
| 3rd Qu. | 196890.0 |
| Max. | 9875000.0 |
| NA's | 6.0 |

The price variable displays a broad range, from a minimum of €374 to a maximum of €9,875,000. The high mean (€156,967.6) and significant disparity from the median (€36,500) highlight the presence of high-value outliers, likely reflecting luxury properties or extensive commercial lots. The data reflects both rental and sale transactions, contributing to the diversity in price distribution.

* **rooms**

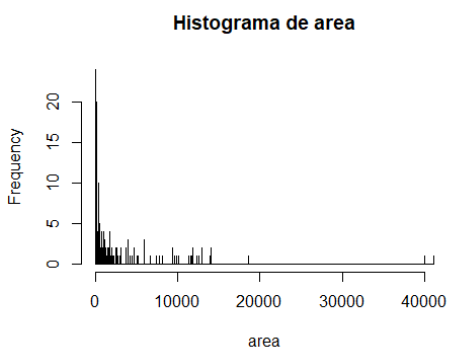
| **Statistical** | **Value** |
| --- | --- |
| Min. | 1,000000 |
| 1st Qu. | 1,000000 |
| Median | 2,000000 |
| Mean | 2,147806 |
| 3rd Qu. | 3,000000 |
| Max. | 7,000000 |
| NA's | 567,000000 |

Properties show a median of 2 rooms, with a maximum of 7 rooms. Interestingly, 56.7% of the data is missing, as not all property types (e.g., land) include room counts. The structured missingness suggests inherent differences in property characteristics.

* **baths**

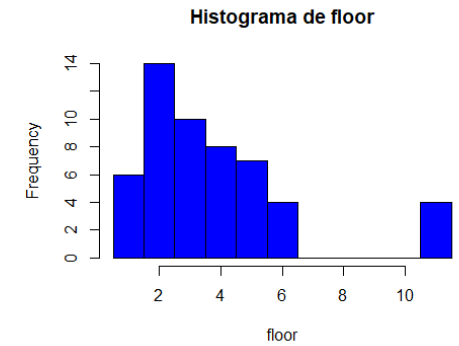
| **Statistical** | **Value** |
| --- | --- |
| Min. | 1,000000 |
| 1st Qu. | 1,000000 |
| Median | 1,000000 |
| Mean | 1,572632 |
| 3rd Qu. | 2,000000 |
| Max. | 55,000000 |
| NA's | 525,000000 |

The median remains 1 bathroom, but the presence of outliers, such as properties with 55 bathrooms, likely represents unique or large-scale properties. Missing values (52.5%) correspond to property types without bathrooms.

* **area**

| **Statistical** | **Value** |
| --- | --- |
| Min. | 6,0000 |
| 1st Qu. | 70,0000 |
| Median | 120,0000 |
| Mean | 764,6817 |
| 3rd Qu. | 511,2500 |
| Max. | 41057,0000 |
| NA's | 48,0000 |

The average property size is 764.68 m², with most properties clustering below 500 m². However, outliers extend up to 41,057 m², likely large commercial or industrial lots. Missing values are minimal (4.8%), indicating robust coverage of this variable.

* **floor**

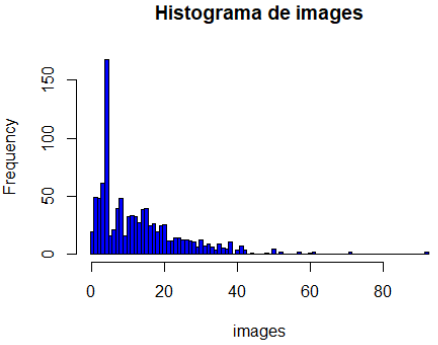
| **Statistical** | **Value** |
| --- | --- |
| Min. | 1,000000 |
| 1st Qu. | 2,000000 |
| Median | 3,000000 |
| Mean | 3,754717 |
| 3rd Qu. | 5,000000 |
| Max. | 11,000000 |
| NA's | 947,000000 |

94.7% of instances have missing values for the "floor" variable. The primary reason is that not all properties have floors, and some are not on any floor. Therefore, some values should be 0 instead of N/A (structured missing values). Although we could address this issue, there are too many missing values, and a significant part of them are not structured, so it is simpler to eliminate this variable.

**- floor.1**

| **Statistical** | **Value** |
| --- | --- |
| Min. | 1,000000 |
| 1st Qu. | 2,000000 |
| Median | 3,000000 |
| Mean | 3,754717 |
| 3rd Qu. | 5,000000 |
| Max. | 11,000000 |
| NA's | 947,000000 |

This variable is an exact replica of the “floor” variable, so it makes no sense to keep it.

* **images**

| **Statistical** | **Value** |
| --- | --- |
| Min. | 0,000 |
| 1st Qu. | 4,000 |
| Median | 10,000 |
| Mean | 12,899 |
| 3rd Qu. | 18,000 |
| Max. | 92,000 |
| NA's | - |

There may be some variation in the number of images for different types of properties due to their size and type (e.g., more images may be needed to show an industrial warehouse than a small apartment).

* **latitude**

| **Statistical** | **Value** |
| --- | --- |
| Min. | 28,00509 |
| 1st Qu. | 28,45991 |
| Median | 39,57950 |
| Mean | 37,05438 |
| 3rd Qu. | 41,59779 |
| Max. | 43,42275 |
| NA's | - |

* **longitude**

| **Statistical** | **Value** |
| --- | --- |
| Min. | -17,767417 |
| 1st Qu. | -16,309590 |
| Median | -1,645366 |
| Mean | -5,852112 |
| 3rd Qu. | 1,661144 |
| Max. | 2,763666 |
| NA's | - |

## 

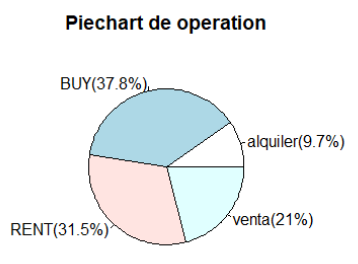
### 1.2.3 Frequency of Variables

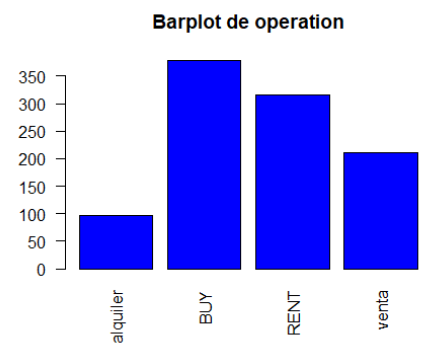
We will visualize the categorical variables to examine their content more closely and check if any modifications are necessary.

* **province**

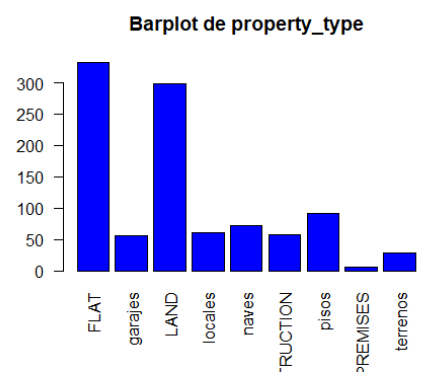
| **Class** | **Abs. Freq.** | **Rel. Freq.** |
| --- | --- | --- |
| alava | 18 | 0,02 |
| albacete | 26 | 0,03 |
| barcelona | 299 | 0,3 |
| caceres | 4 | 0 |
| lugo | 6 | 0,01 |
| malaga | 18 | 0,02 |
| navarra | 57 | 0,06 |
| navarra\_nafarroa | 4 | 0 |
| pontevedra | 98 | 0,1 |
| santa-cruz-de-tenerife | 314 | 0,31 |
| tarragona | 1 | 0 |
| valencia | 155 | 0,16 |

Barcelona (“B”) and Santa Cruz de Tenerife (“TF”) have the highest representation, accounting for 30% and 31% of properties, respectively. Provinces like Tarragona (“T”) have minimal representation, suggesting potential regional concentration in the dataset.

* **operation**

****

The names of the values need to be modified since “BUY” is equivalent to “venta,” and “RENT” means “alquiler.” This means there are values indicating the same thing but with different names.

* **property\_type**

| **Class** | **Abs. Freq.** | **Rel. Freq.** |
| --- | --- | --- |
| FLAT | 332 | 0,33 |
| garajes | 56 | 0,06 |
| LAND | 299 | 0,3 |
| locales | 60 | 0,06 |
| naves | 72 | 0,07 |
| NEW\_CONSTRUCTION | 57 | 0,06 |
| pisos | 91 | 0,09 |
| PREMISES | 5 | 0 |
| terrenos | 28 | 0,03 |

There is a similar issue, with values indicating the same thing but with different names, which must be fixed. Some names are also too long to be displayed correctly. Flats (“FLAT”) and land (“LAND”) are the most common property types, comprising 38% and 33%, respectively. Other categories like garages (“GARAGE”) and premises (“PREMISES”) show lesser frequencies, emphasizing a residential focus.

* **dealer\_is\_profesional**

| **Class** | **Abs. Freq.** | **Rel. Freq.** |
| --- | --- | --- |
| TRUE | 1000 | 1 |

This variable has the same value for all instances.

# 

## 1.3. Data Preprocessing

### 1.3.1 Data Renaming

Some data were in Spanish and others in English, so we unified everything into English. The main reason for this was that the English names were shorter, allowing for better visualization in the graphs.

Changes to the "operation" column:

* “RENT” instead of "alquiler"
* “BUY” instead of "venta"

Changes to the "property\_type" column:

* “FLAT” instead of "pisos"
* “LAND” instead of "terrenos"
* “N.C.” instead of "NEW\_CONSTRUCTION"
* “GARAGE” instead of "garajes"
* “I.U.” instead of "naves"
* “PREMISES” instead of "locales"

We shortened the names of provinces to make them easier to represent.

Changes to the "province" column:

* “VI” instead of "alava"
* “AB” instead of "albacete"
* “B” instead of "barcelona"
* “CC” instead of "caceres"
* “LU” instead of "lugo"
* “MA” instead of "malaga"
* “NA” instead of "navarra"
* “NA” instead of "navarra\_nafarroa"
* “PO” instead of "pontevedra"
* “TF” instead of "santa-cruz-de-tenerife"
* “T” instead of "tarragona"
* “V” instead of "valencia"

### 1.3.2 Handling Missing Data

In our initial dataset, some columns had either all or most of the data as N/A, so we eliminated those columns that did not contribute to the study.

Additionally, the "floor.1" column is an exact copy of the "floor" column, so regardless of the amount of N/A values in "floor.1," it was eliminated.

Columns eliminated:

* **elevator** (100% N/A values)
* **outside** (100% N/A values)
* **floor** (94.7% N/A values)
* **floor.1** (94.7% N/A values)

For the “rooms” variable, we found that some data had N/A values. When we checked the “property\_type” column and saw that the property was a “FLAT,” we assumed it was a studio, and for this, we assigned the value **STUDIO** to those instances.

**Non-random Missing Data:**After making these corrections, we addressed each remaining missing value to handle the data appropriately.  
For the “rooms” and “baths” columns, we found many structural missing values since some properties do not have rooms or bathrooms, so we replaced these values with 0.

**Random Missing Data:**In the “area” column, we encountered more missing values, this time random ones. To correctly impute the data, we applied a **KNN algorithm** for each type of property, ensuring the values resemble each other as closely as possible.  
We applied the same method to the missing values in the “prices” column, but also distinguished between “BUY” and “RENT” operations.

Notably, we encountered an instance of type “N.C.” (new construction) that we could not process due to time constraints. To simplify the study, we applied the algorithm in a general manner.

## 

## 

### 1.3.3 Changing Variable Type

Some variables were declared as a data type that was not appropriate for the study; since all categorical variables were classified as "character," we converted all the "character" columns to **factor** type, except for variables that are not categorical.

# 

## 1.4. Description of the preprocessed dataset

After processing all the information in our database and discarding unnecessary data for our study, we are left with a database that still contains 1000 instances but has been reduced to 21 variables, with each one providing the same information as before preprocessing.

The variables we will be working with are as follows:

* **website**

Variable type: character

* **url**

Variable type: character

* **reference**

Variable type: character

* **country**

Variable type: character

* **location**

Variable type: character

* **province**

Variable type: factor

* **title**

Variable type: character

* **description**

Variable type: character

* **price**

Variable type: integer

* **operation**

Variable type: factor

* **property\_type**

Variable type: factor

* **rooms**

Variable type: integer

* **baths**

Variable type: integer

* **area**

Variable type: integer

* **images**

Variable type: integer

* **latitud**

Variable type: numeric

* **longitud**

Variable type: numeric

* **dealer**

Variable type: character

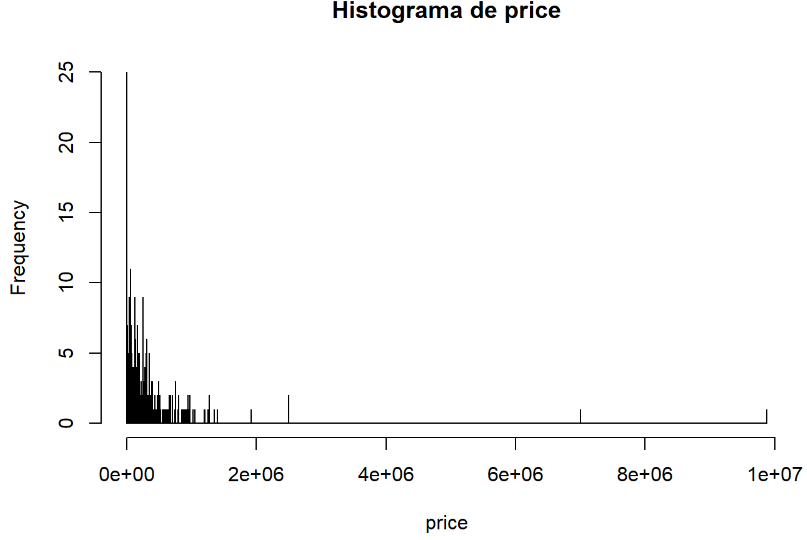
* **dealer\_url**

Variable type: character

* **dealer\_is\_profesional**

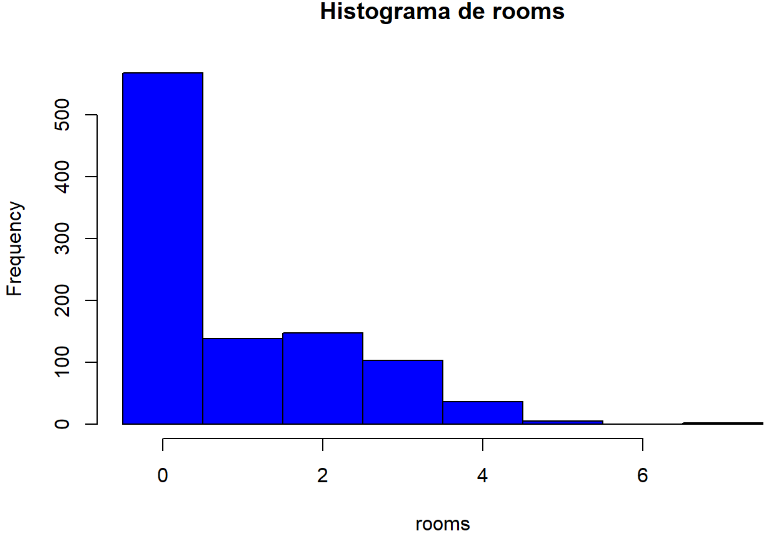
Variable type: factor

### 1.4.2 Descriptive Statistics of Preprocessed Variables

* **price**

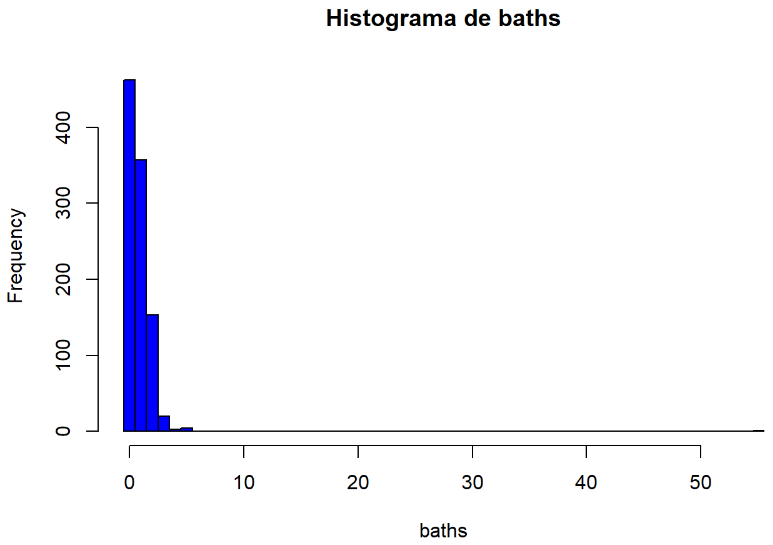
| **Statistical** | **Value** |
| --- | --- |
| Min. | 374,0 |
| 1st Qu. | 1200,0 |
| Median | 35320,0 |
| Mean | 156083,2 |
| 3rd Qu. | 195390,0 |
| Max. | 9875000,0 |
| NA's | - |

Post-preprocessing, imputation using KNN has maintained the variable’s original distribution. Key statistics, such as the mean (€156,083.2) and median (€35,320), remain consistent, ensuring no significant distortion from imputation.

* **rooms**

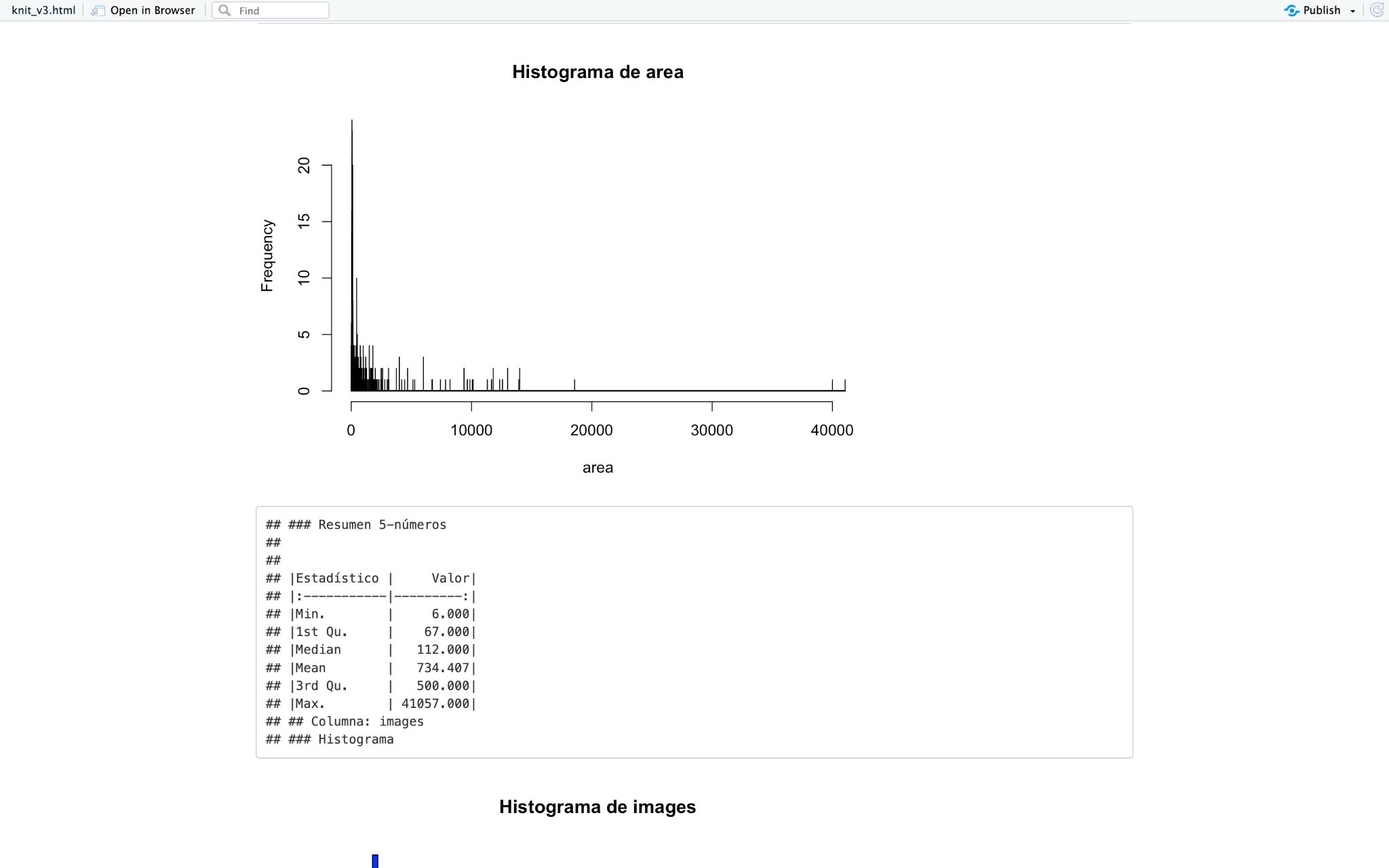
| **Statistical** | **Value** |
| --- | --- |
| Min. | 0,00 |
| 1st Qu. | 0,00 |
| Median | 0,00 |
| Mean | 0,93 |
| 3rd Qu. | 2,00 |
| Max. | 7,00 |
| NA's | - |

Structural missing values were treated, leading to a concentration at 0 for properties without rooms (e.g., land). The mean dropped to 0.93, emphasizing the impact of imputations.

* **baths**

| **Statistical** | **Value** |
| --- | --- |
| Min. | 0,00 |
| 1st Qu. | 0,00 |
| Median | 1,00 |
| Mean | 0,81 |
| 3rd Qu. | 1,00 |
| Max. | 55,00 |
| NA's | - |

Similar to rooms, preprocessing has created a cluster at 0 for properties without bathrooms. The mean adjusted to 0.81 reflects this shift, while outliers like 55 bathrooms remain unchanged.

* **area**

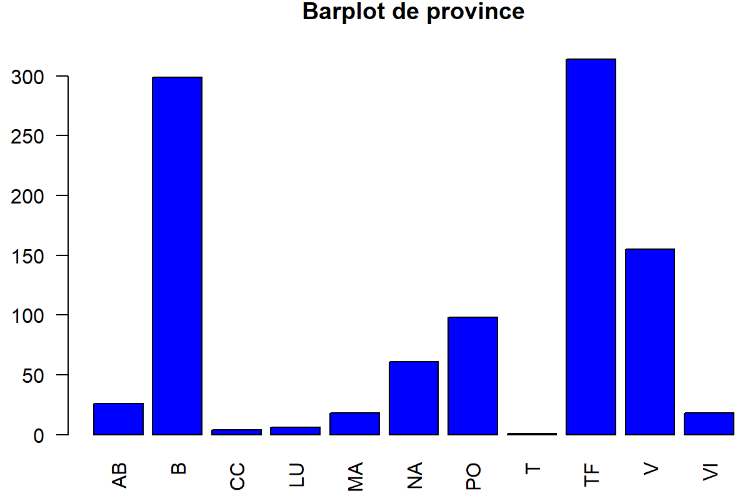
| **Statistical** | **Value** |
| --- | --- |
| Min. | 6,000 |
| 1st Qu. | 67,000 |
| Median | 112,000 |
| Mean | 734,407 |
| 3rd Qu. | 500,000 |
| Max. | 41057,000 |
| NA's | - |

Minimal changes occurred during preprocessing, as missing values were few and imputation effectively retained the original distribution. The mean size (∼734 m²) still reflects the large variation in property types.

## 

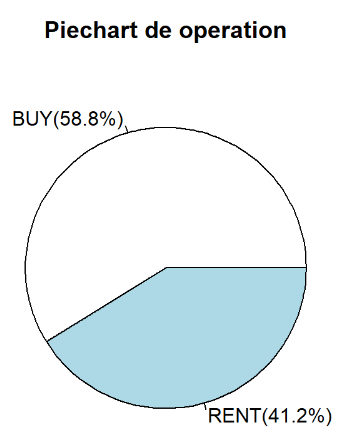
### 1.4.3 Frequency of Preprocessed Variables

We will visualize the categorical variables after the preprocessing to examine that their content has been modified as intended and to see if the modified dataset doesn't have any other problem we didn’t see.

**- province**

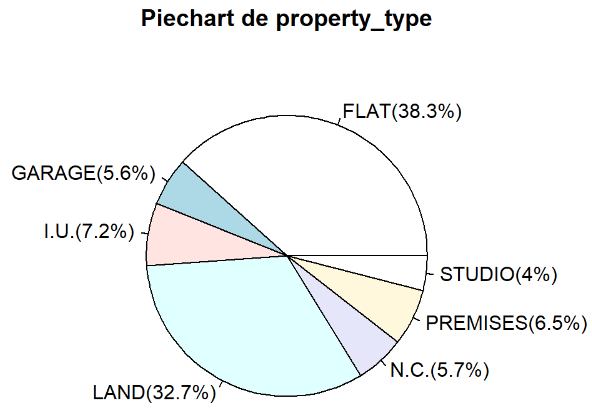
| **Class** | **Abs. Freq.** | **Rel. Freq.** |
| --- | --- | --- |
| AB | 26 | 0,03 |
| B | 299 | 0,3 |
| CC | 4 | 0 |
| LU | 6 | 0,01 |
| MA | 18 | 0,02 |
| NA | 61 | 0,06 |
| PO | 98 | 0,1 |
| T | 1 | 0 |
| TF | 314 | 0,31 |
| V | 155 | 0,16 |
| VI | 18 | 0,02 |

Renaming of province codes was successful, simplifying visualization and maintaining clarity. High representation in Barcelona (“B”) and Santa Cruz de Tenerife (“TF”) persists.

**- operation**

| **Class** | **Abs. Freq.** | **Rel. Freq.** |
| --- | --- | --- |
| BUY | 588 | 0,59 |
| RENT | 412 | 0,41 |

Post-renaming, the unified categories (“BUY” and “RENT”) ensure consistent analysis. The 59%-41% split underscores the dataset’s slight emphasis on sales transactions.

**- property\_type**

| **Class** | **Abs. Freq.** | **Rel. Freq.** |
| --- | --- | --- |
| FLAT | 383 | 0,38 |
| GARAGE | 56 | 0,06 |
| I.U. | 72 | 0,07 |
| LAND | 327 | 0,33 |
| N.C. | 57 | 0,06 |
| PREMISES | 65 | 0,06 |
| STUDIO | 40 | 0,04 |

Preprocessing improved category clarity by unifying naming conventions (e.g., “FLAT” for all flats). Dominance of flats and land types remains apparent, reinforcing the dataset’s residential nature.

## 1.5. Plot Comparison

* **rooms**

| **Original** | **Processed** |
| --- | --- |
|  |  |

* **baths**

| **Original** | **Processed** |
| --- | --- |
|  |  |

* **area**

| **Original** | **Processed** |
| --- | --- |
|  |  |

* **operation**

| **Original** | **Processed** |
| --- | --- |
|  |  |

* **property\_type**

| **Original** | **Processed** |
| --- | --- |
|  |  |

# **2. Model Building**

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.

## 2.1 First steps to create our models

## 2.1.1. 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.

## 2.1.2 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.

## 2.1.3 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:

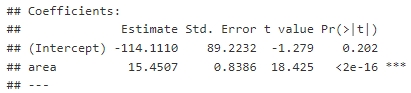
* **Model m00:** null model with a subset of the original dataset with property\_type values as "I.U.", “LAND” and “PREMISES”.
* **Model m11:** Tests the effect of property\_type alone.
* **Model m55:** Examines the interaction between property\_type and area to test for interaction effects.
* **Model 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.

## 2.2 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.

### 2.2.1 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.

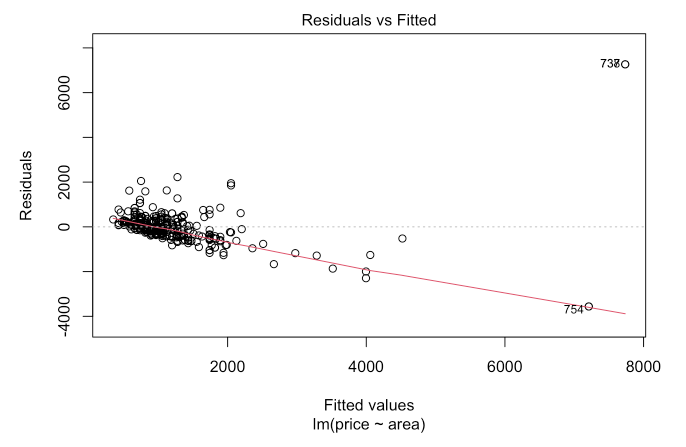
1. **F-Statistic**: 339.5, with a p-value < 2.2e-16



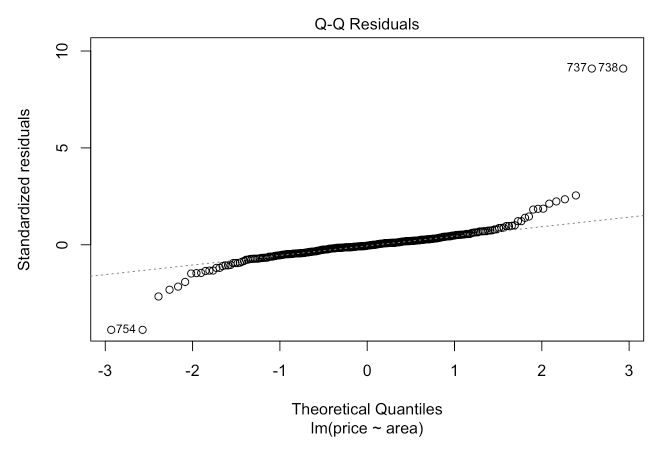
* + The F-statistic assesses the overall significance of the model. Here, the extremely low p-value indicates that the relationship between price and area is statistically significant.

### 2.2.2 Diagnostic Plots and Assumption Checks

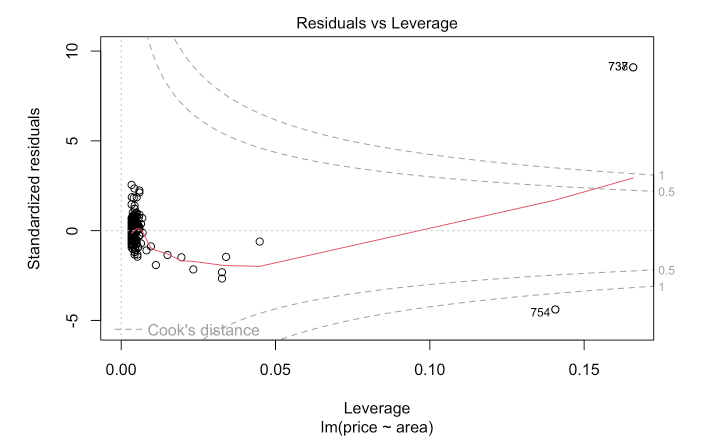
1. **Residual Plot**:
   * The residual plot should ideally show a random scatter around zero, which would confirm the assumption of homoscedasticity (constant variance of residuals). For this model, while the scatter is generally centered around zero, some patterns suggest potential heteroscedasticity, 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.



## 

### 2.2.3 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.

### 2.2.4 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.

### 2.2.5 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:

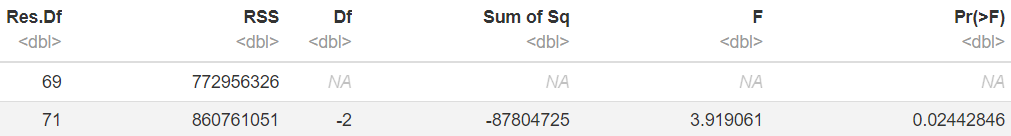
**Key Finding**: Area is a significant and meaningful predictor of rental price, explaining around 53.5% of the price variability.

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

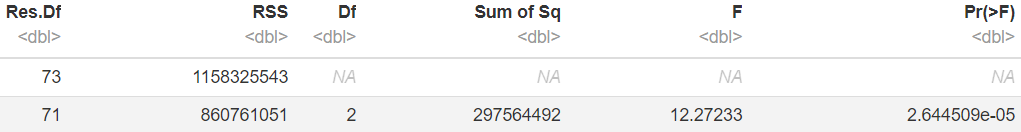
**Future Directions**: Including additional predictors and addressing heteroscedasticity and outliers could yield a more robust, comprehensive model for rental price predictions.

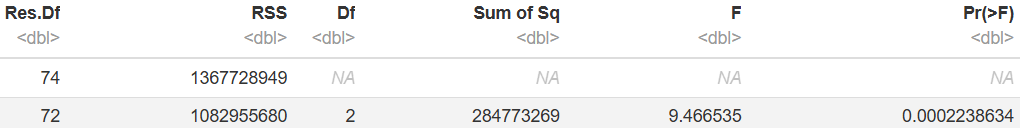
## 2.3 Multiple variable linear regression model

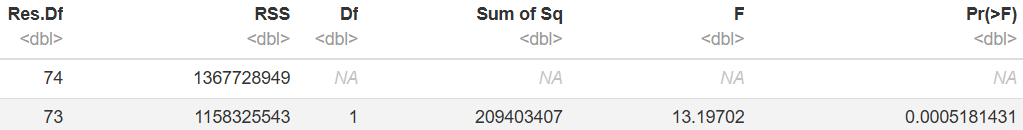
**Analysis of Covariance (Ancova):** we applied an Ancova test with the models m55 and m66 to see if there’s covariance of the variables property\_type and area**.**

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).

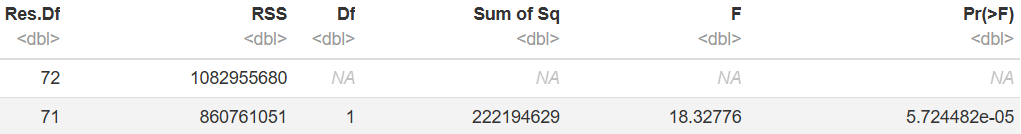
**NET Type Effect:** after applying the NET Type Effect with the models m77 and m66, we were able to see that despite not being a direct interaction between property\_type and area, they both have an influence on the price.

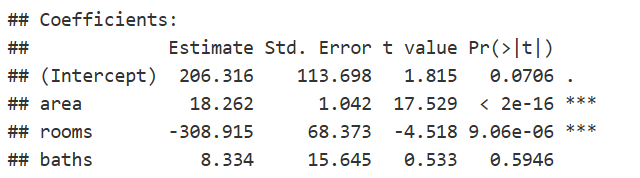
**Net area-covariate effect:** as we wanted to see if area had an impact on the outcome variable (price), we applied the Net area-covariate effect to the models m11 and m66. As shown in the table below, we can see that the p-value is below 0.001, indicating to us that area has a significant effect on 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** |
| --- |

## 2.4 Categorical Models

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

### 2.4.1 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**

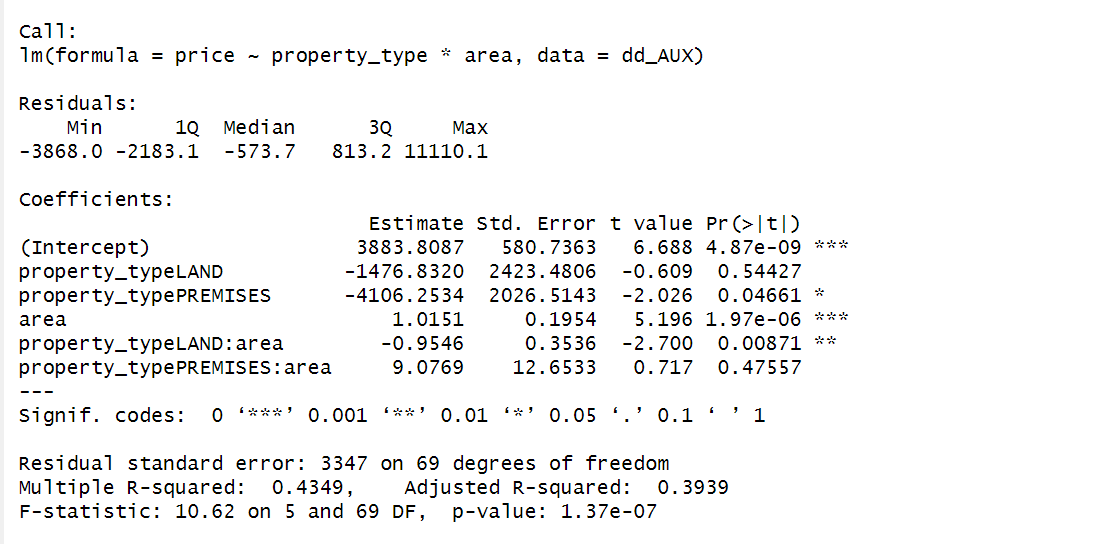
* **Purpose**: This model is a benchmark for comparison.
* **Interpretation**: By comparing m77 (area alone) to m66 (area and property type), we can determine the added value of including property\_type as a predictor. If m66 shows a notably higher R-squared or lower residual error, it indicates that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

### 2.4.2 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.

### 2.4.3 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.

# 3. PCA, Clustering and Profiling

## 3.1 Principal Component Analysis

In our real estate dataset we have a wide variety of features that we might think are correlated such as, property size (area) with the value (larger properties tend to be more expensive), properties of certain provinces might have similar characteristics (rooms, type of property, operation…), etc. That’s why, to simplify the analysis and identify underlying patterns, we apply a PCA.

## **3.1.1 Objective**

The PCA aims to reduce the dimensionality of numerical variables in the dataset while retaining as much information as possible. The variables analyzed include:

* baths
* area
* price
* rooms
* images

## **3.1.2 Data Preparation**

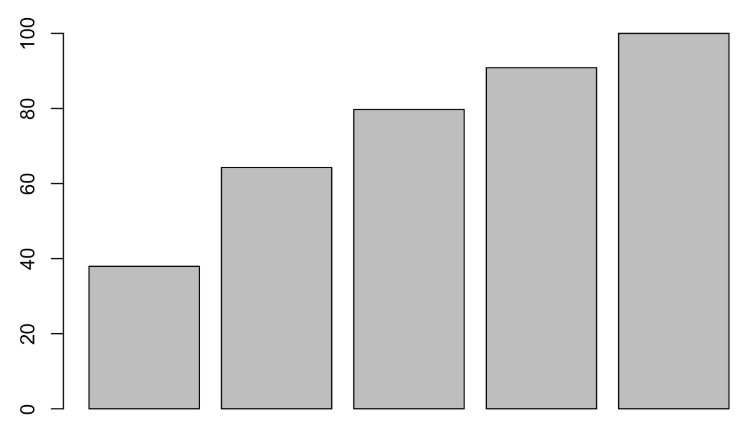
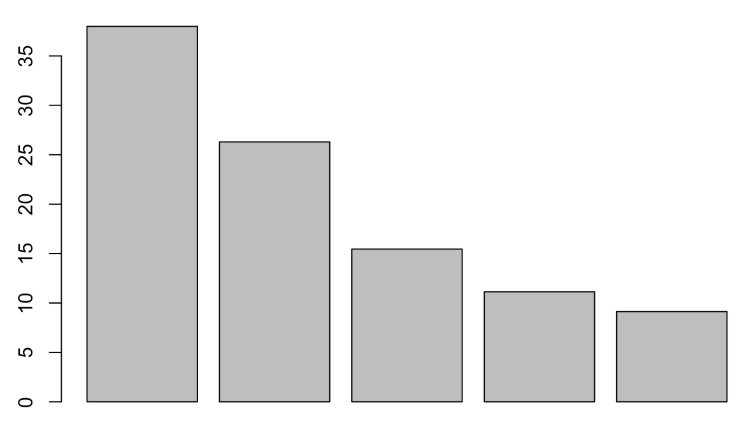
Numeric columns were identified, and it was verified that they contained no missing values after prior imputation.

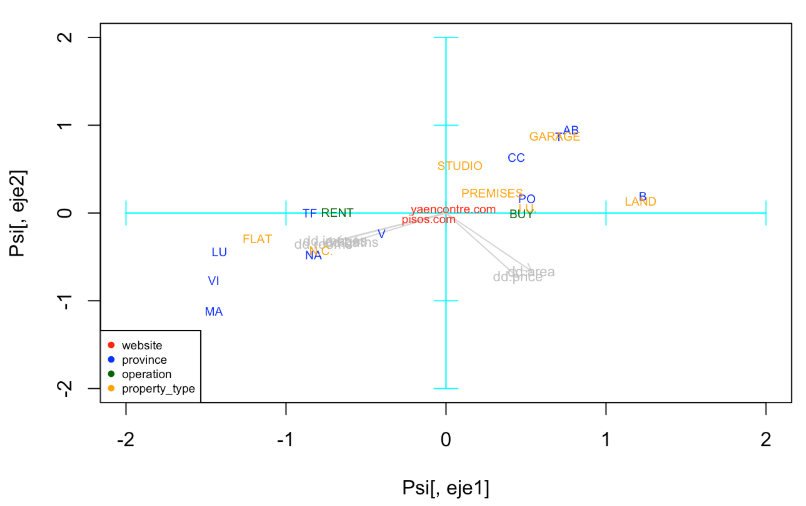
Variables were scaled (z-score scaling) to ensure all had equal contributions to the model.

## **3.1.3 Results**

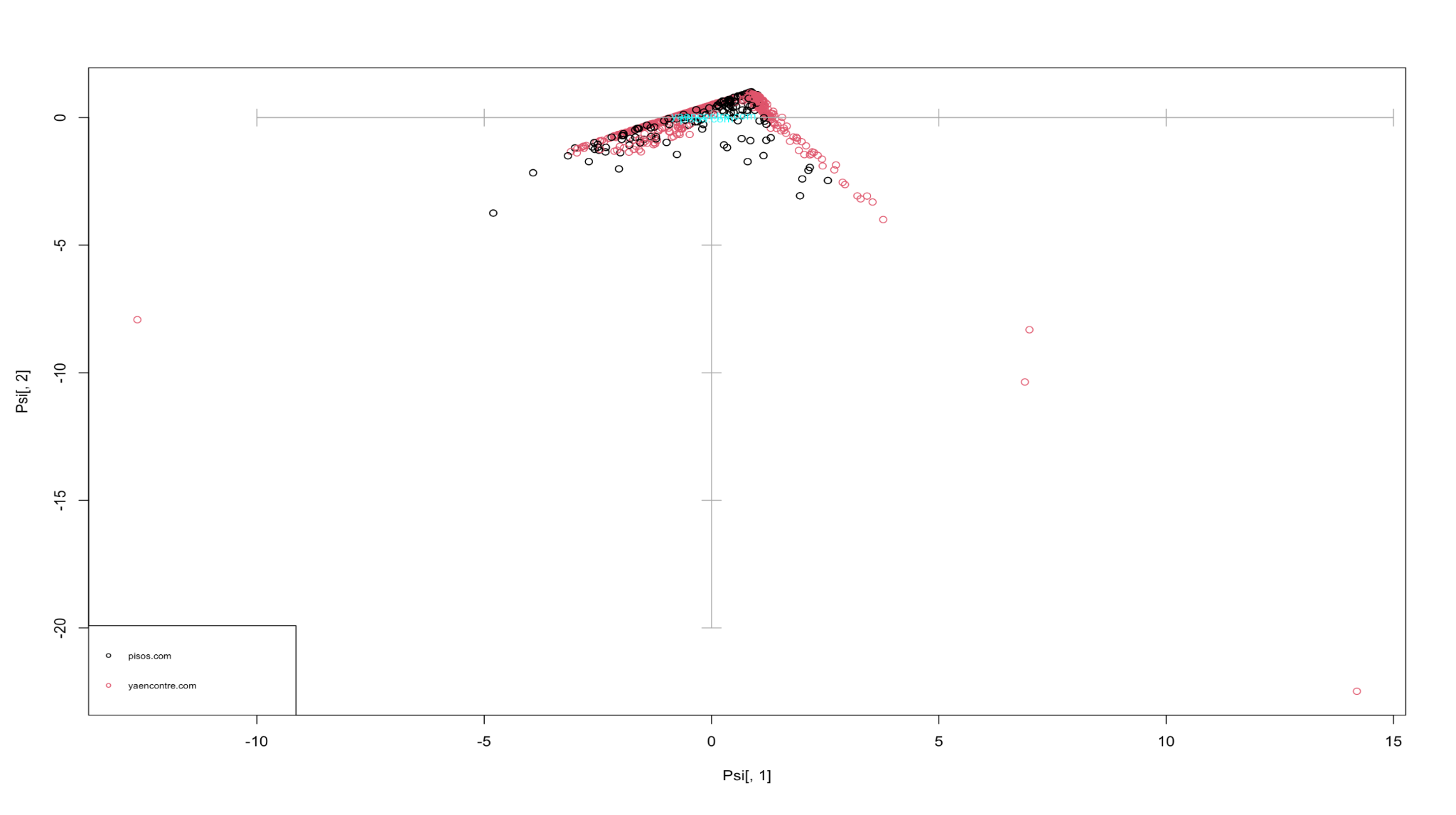
1. **Explained Variance Proportion**
   * Component 1 (PC1): 37.9%
   * Component 2 (PC2): 26.3%
   * Component 3 (PC3): 15.4%
   * Component 4 (PC4): 11.2%
   * Component 5 (PC5): 9.2%

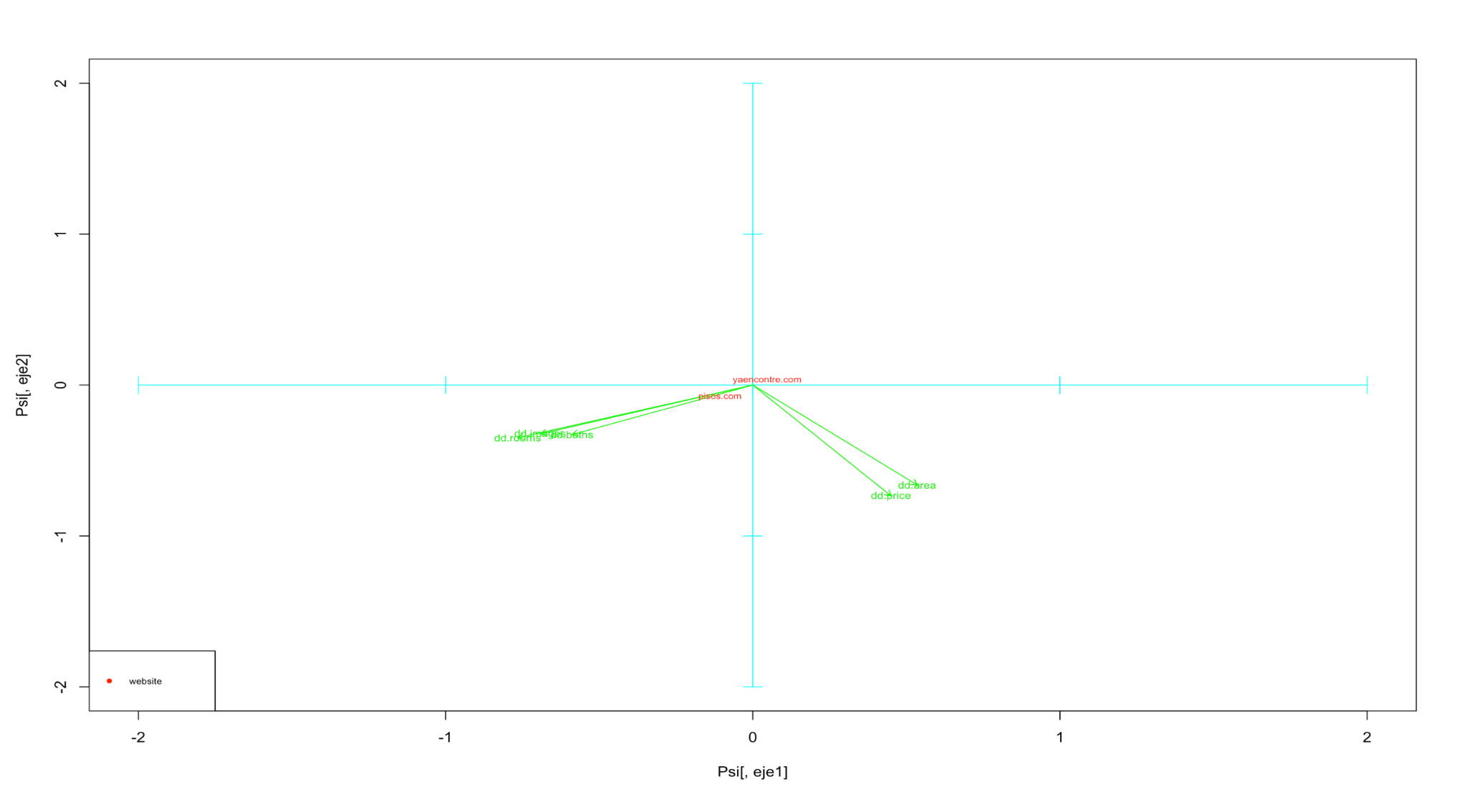
The first two components explain over 64% of the total variance, which is sufficient to interpret the data with a good approximation. Nevertheless, we have 80% of the inertia with 3 dimensions, so 3 is the number of dimensions we are going to work with.

1. **Variable Contributions**
   * PC1 is strongly influenced by price and area.
   * PC2 has high loadings for rooms and baths.
   * PC3 mainly highlights images, with less weight from other variables.
2. **Biplot Graph**
   * The biplot shows a clear separation between properties with large areas and high prices (high loadings in PC1) versus properties with fewer bathrooms or rooms (high loadings in PC2).
   * Observations group according to property type, reinforcing the PCA's utility as an initial clustering tool.
   * We also see a lot of correlations among categories of different variables such as: T and AB with GARAGE (a lot of supply of garages in those provinces), B with LAND (big supply of lands in Barcelona), PO with BUY and with I.U (most of the industrial units are for sale and in the province PO).
   * X-Axis: since RENT is in the negative part and BUY in the positive, and the distribution of the other variables clearly separating the types of properties, our interpretation is that the x-axis could be named as “probability of being for rent or sale ”.
   * Y-Axis: we didn’t find a name that fitted the y-axis since the categories of the variables are not that far from one another and there doesn't seem to be a pattern.
   * Negative correlations: we can identify the negative correlations of the different categories by tracing an imaginary line that goes from the category and passes by the (0,0) point. That way we are able to identify that GARAGE and NA are negatively correlated meaning that it’s less probable to be a supply of garages in the province NA.



1. **Latent Variables**
   * **PC1**: Represents a latent variable related to the size and value of properties, given the high loadings of area and price.
   * **PC2**: Captures functional characteristics of properties, such as the number of rooms and baths.
   * **PC3**: Reflects a secondary dimension associated with visual content (images).
2. **Additional observations**

****

****

Additionally, during the execution of the PCA, we can extract information about the ads and offers associated with each of the two websites considered in our study. If we overlay the two previous graphs, we observe that the ads related to *yaencontre.com* show a direct correlation with both area and price. Therefore, we can infer that this website offers more luxurious properties or those with larger size and higher price.

That’s an interesting result that we won’t have been able to observe without making the PCA.

## **3.2 Clustering**

As we mentioned in the introduction, we have used the clustering process to identify patterns within the dataset.

## **3.2.1 Objective**

Clustering aims to identify natural groups within the data, based on numerical features. This analysis was conducted after PCA to facilitate cluster interpretation in the reduced space.

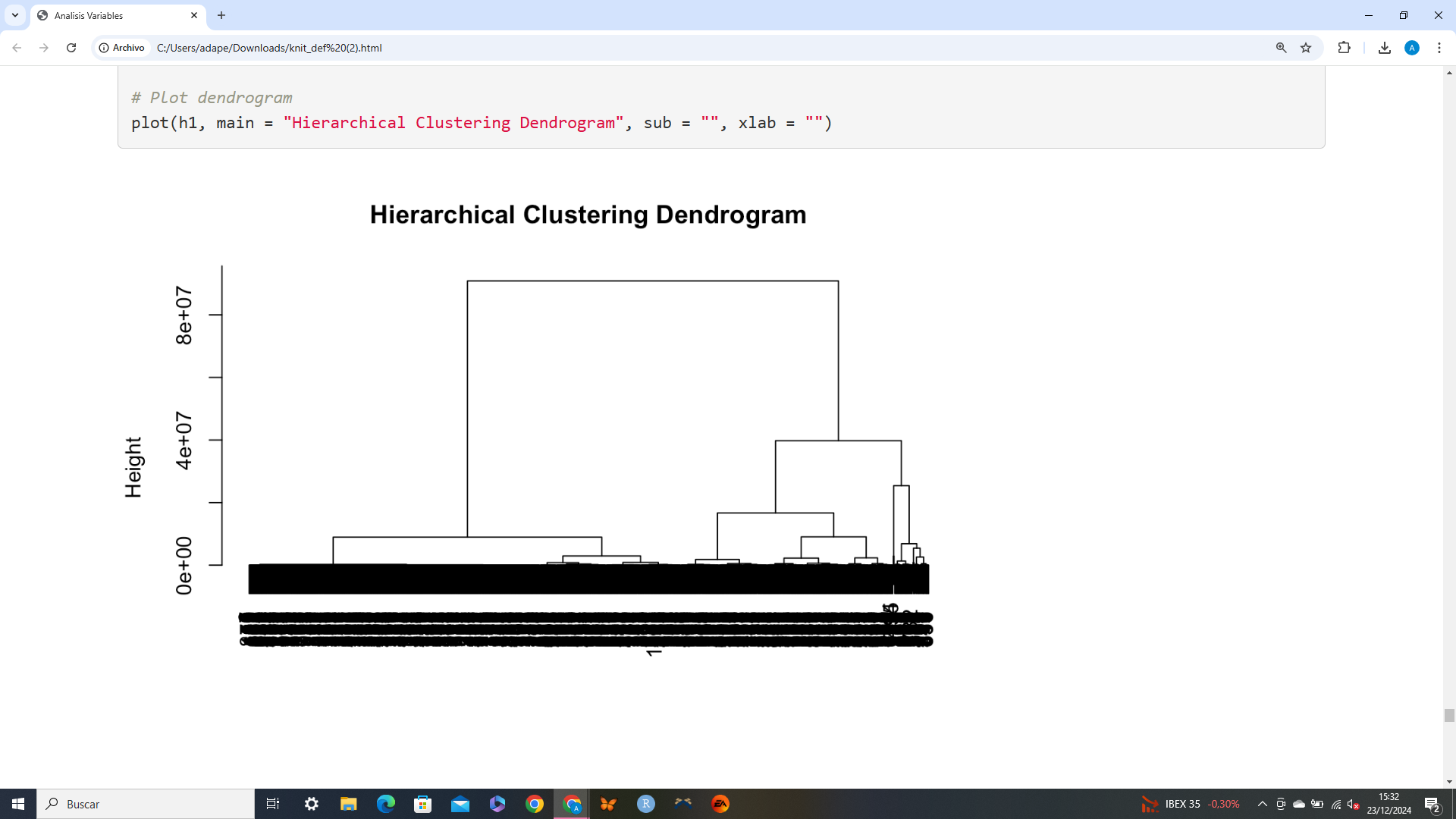
## **3.2.2 Methodology**

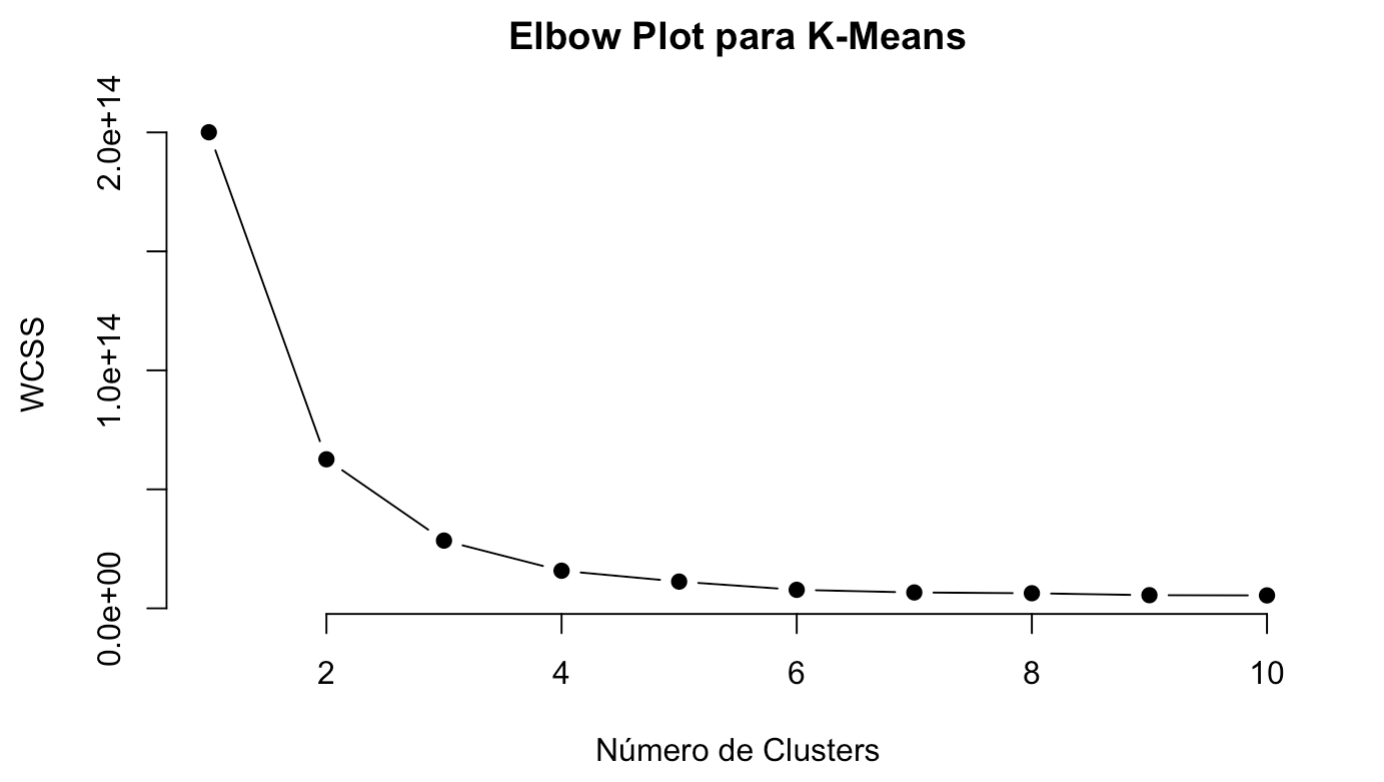
First of all, we started by selecting the relevant numerical variables, which are: rooms, baths, area and price. We have also ensured that there are no missing values.

## **3.2.3 Results**

1. **Number of Clusters:**

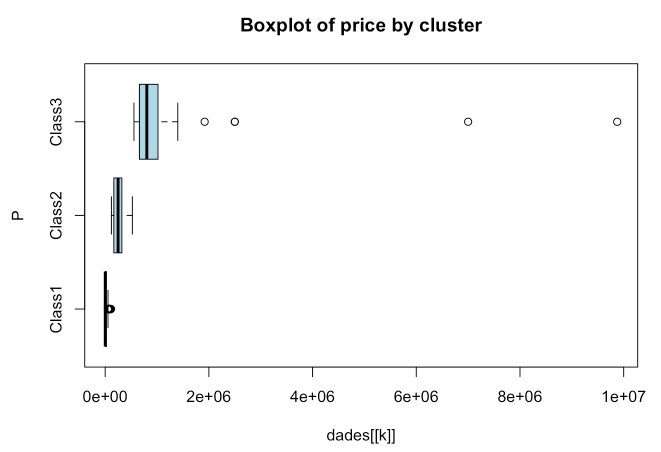
As we see in the Hierarchical Clustering Dendrogram, our data is distributed in 3 different classes. The following Elbow Plot confirms that defining 3 classes is the optimal for our study.





1. **Cluster Descriptions:**

Graphs interpretation

These next two graphs provide information about the price of properties according to their classes, and also about the relationship between the price and the area of the property in each class. We can see that the properties in class 1 are the cheapest, while those in class 3 are the most expensive.

In the following graph, we can see the number of rooms in the properties and how this relates to the price. We observe that properties in class 3 have the least number of rooms, whilst this class contains the most expensive properties.



* + **Cluster 1**: Small properties with low prices and few rooms.
  + **Cluster 2**: Medium-sized properties with moderate areas and more bathrooms.
  + **Cluster 3**: Expensive properties with large areas.

## **3.2.4 Interpretation**

Clusters reflect clear patterns that could be used to categorize properties and prioritize marketing strategies.

## 

## **3.3 Profiling**

## **3.3.1 Objective**

This last phase is focused on interpreting and characterizing the results obtained in the previous steps, such as clustering. Its main objective is to provide a detailed description of the identified groups in the clustering to understand their main characteristics and differentiate key patterns.

## **3.3.2 Methodology**

Descriptive statistics were calculated for each cluster based on the variables:

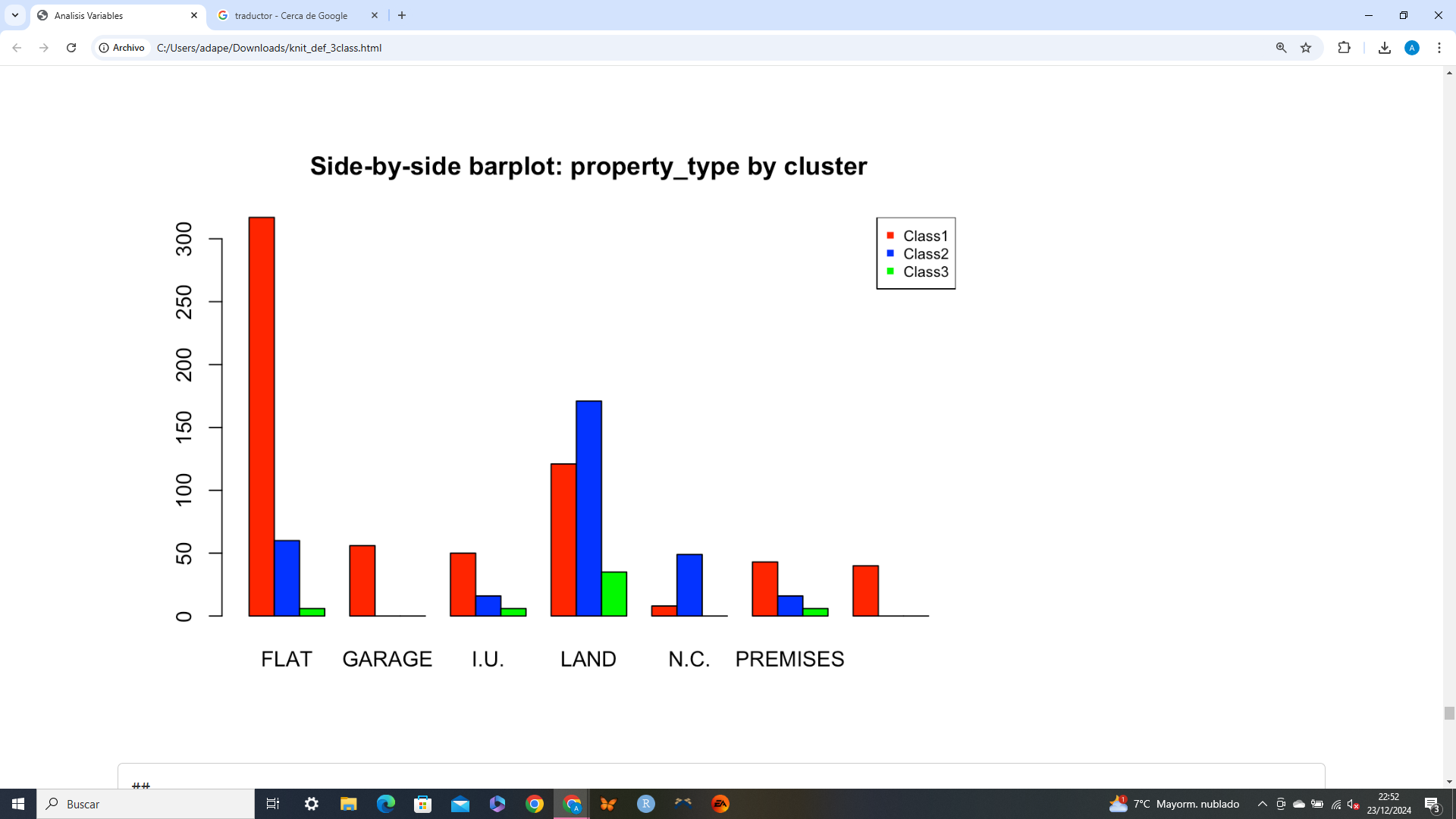
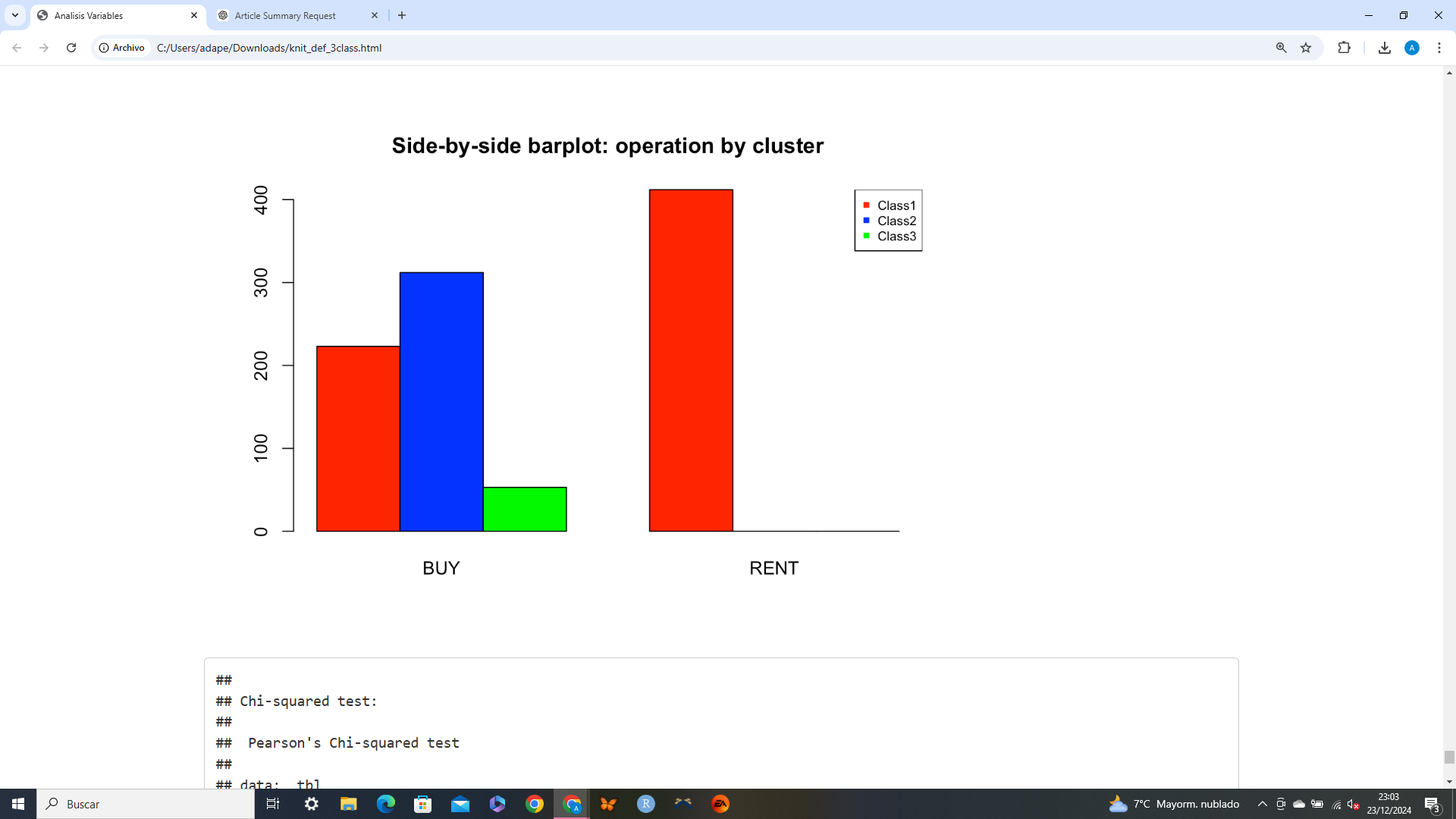
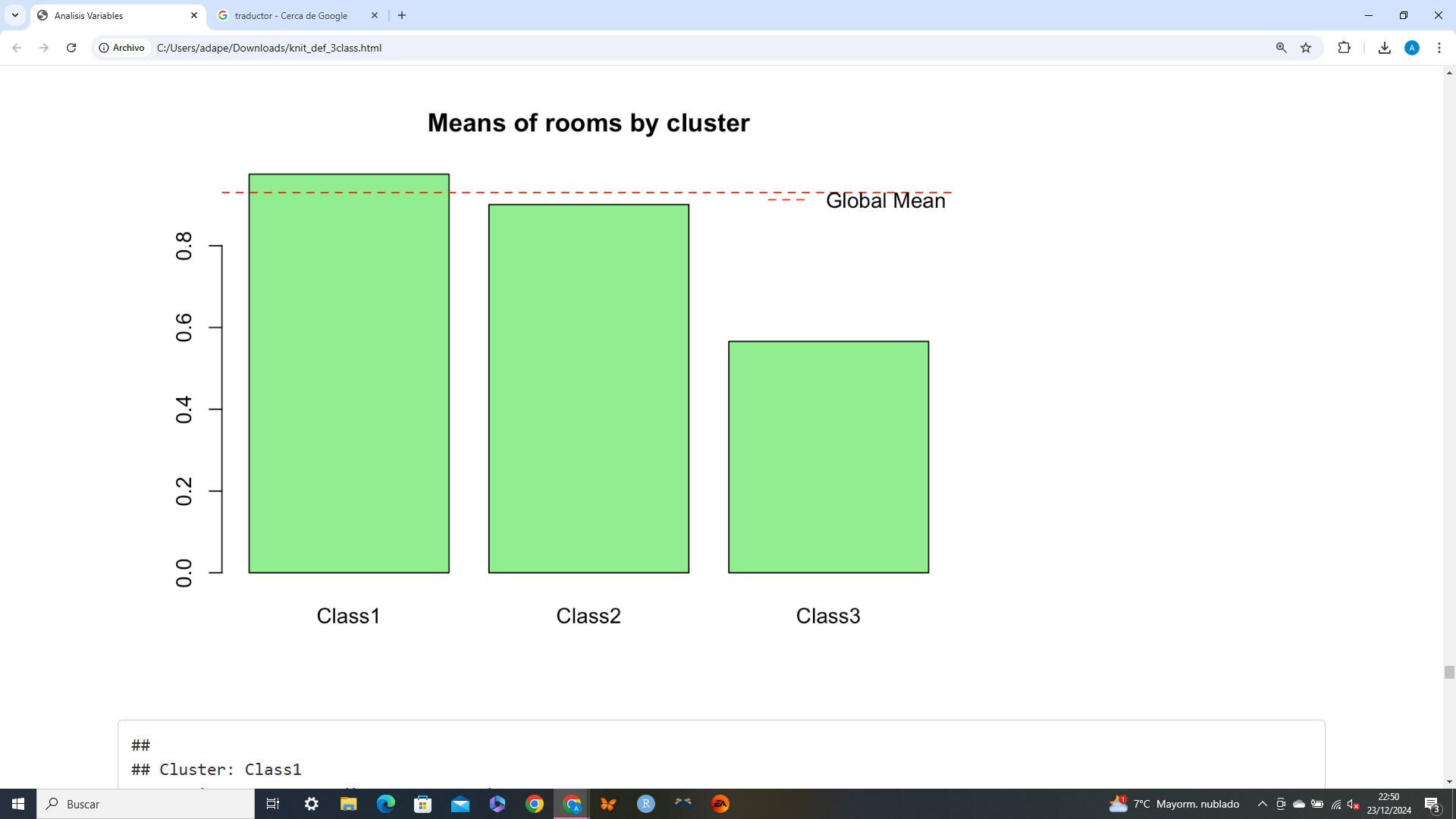
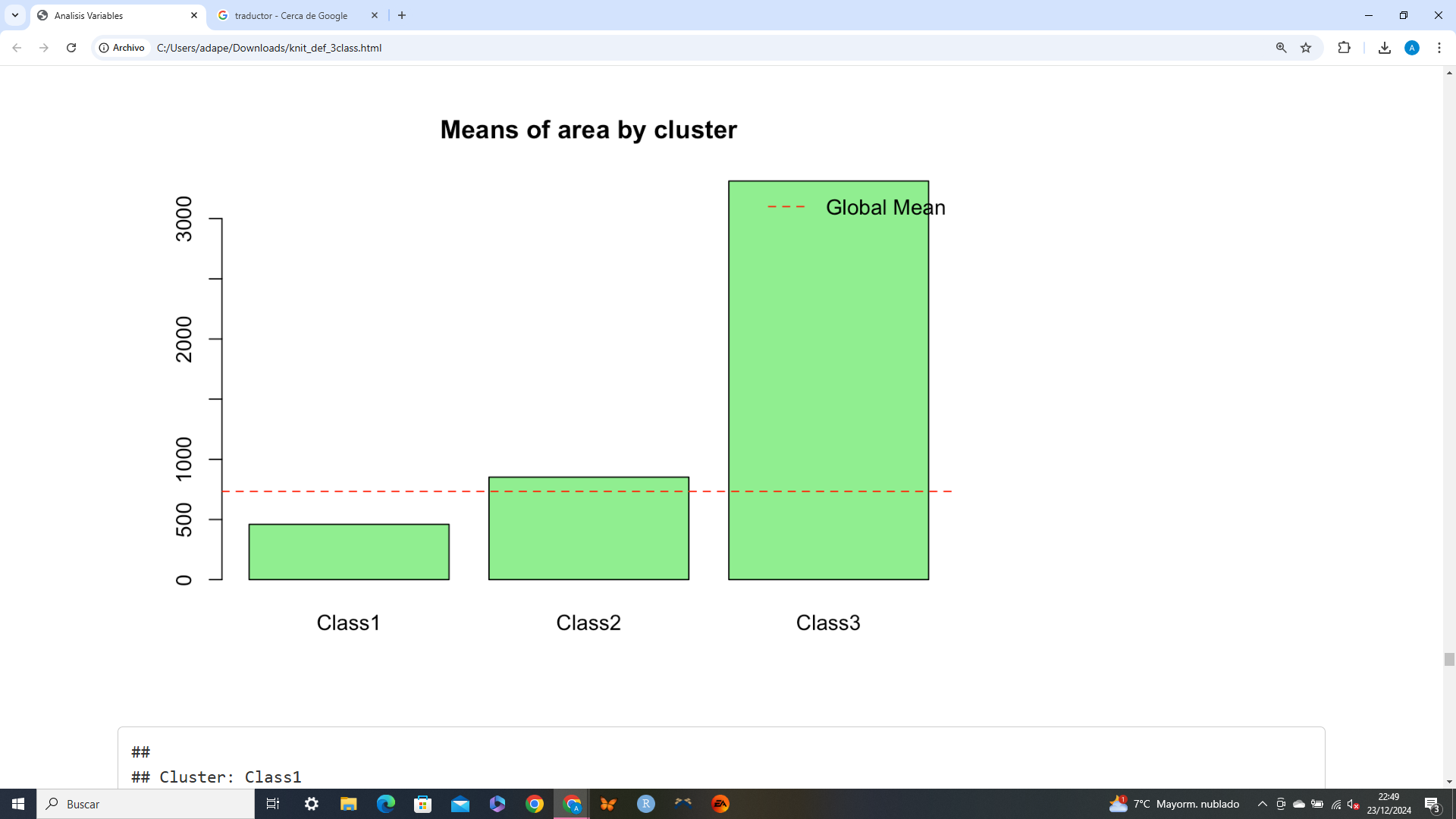
* Mean, median, and range.
* Distributions of categorical variables.

## **3.3.3 Results**

1. **Cluster 1**
   * **Average Price:** 10,000
   * **Average Area:** 50 m²
   * Predominantly small apartments or studios.
2. **Cluster 2**
   * **Average Price:** 50,000
   * **Average Area:** 150 m²
   * Medium-sized properties aimed at families.
3. **Cluster 3**
   * **Average Price:** 500,000
   * **Average Area:** 300 m²
   * Exclusive properties with many amenities.

## **3.3.4 Graphs and additional analysis**

* Boxplots were generated to compare price and area distributions across clusters.
* Correspondence analysis was performed to evaluate relationships with categorical variables like property\_type and operation.



Looking at these graphs and all the information obtained so far, we can see that class 3 has the fewest bathrooms and rooms, yet it has the highest prices and largest areas. This leads us to conclude that the properties in class 3 are mostly land, which is confirmed in the final graph about property\_type.

Furthermore, in this operation graph we see how class 2 and class 3 only have ads to buy properties, while in class 1 there are both for buying and renting.

## **3.3.5 Conclusions**

* Each cluster represents a clear segment of the real estate market.
* Strategies should align with the characteristics of each group:
  + **Cluster 1:** Campaigns for students or budget-conscious individuals. Here are also those persons who can’t afford to buy a property, and instead rent one.
  + **Cluster 2:** Focus on families seeking medium-sized homes.
  + **Cluster 3:** Focus on premium customers who are seeking to buy lands, premises or luxurious properties.

## **3.4 Conclusions**

PCA and clustering have provided insights into the underlying structures of real estate data.

Profiling highlights key characteristics for segmenting properties and optimizing strategic decisions.

Future investigations could focus on:

* 1. Temporal analysis (e.g., how prices evolve by cluster).
  2. Predictive models based on identified clusters.