

# Barcelona Real Estate Price Prediction Model

Applied Probability and Statistics - Section C, Team 44

Shifei Ruan(sr678), Seokho Shin(ss1722), Arayansh Vaish(av323), Lucy Zheng(zz421)

**1 Introduction:** This report outlines the process of developing a predictive model for property prices in Barcelona. The objective was to construct a robust and interpretable statistical model using a dataset of 413 properties and subsequently apply it to predict the prices of 200 new properties.

**2 Data Preparation and Initial Model:** We began with a multiple linear regression model containing all available variables (model1). This initial full model served as a baseline, yielding an  $R^2$  of 0.7418, indicating that the variables explained a substantial portion of the price variance. However, several variables (Kitchen, Type) were highly insignificant (p-values > 0.9), which suggested model overspecification.

## 3 Model Refinement Process

The model was refined through an iterative process.

**3.1 Removing Insignificant Variables:** The first step was to remove the clearly insignificant variables Kitchen and Type to create a more parsimonious model (model2). This simplification did not harm the model's explanatory power (R-squared was still around 0.74) but enhanced its focus on impactful features.

```
Coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept) -41344.1    11258.6   -3.672 0.000273 ***
m2           3195.0      160.3    19.935 < 2e-16 ***
Rooms       -11648.5    4542.2   -2.565 0.010693 *
Bathrooms    25918.4    7403.6    3.501 0.000516 ***
Elevator     4984.7     6785.8    0.735 0.463019
Atico        23165.2    11796.5    1.964 0.050245 .
Terrasse     5341.4     8298.4    0.644 0.520155
Parking      43990.2    12100.5    3.635 0.000313 ***
Yard         33533.2    16412.0    2.043 0.041679 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 64110 on 404 degrees of freedom
Multiple R-squared:  0.7417,    Adjusted R-squared:  0.7366
F-statistic: 145 on 8 and 404 DF, p-value: < 2.2e-16
```

Model 2

## 3.2 Addressing Non-Linearity and Interaction Effect

**3.2.1 Interaction Term:** We tested an interaction between m2 and Rooms (model3). The results showed that this interaction term (m2:Rooms) was highly statistically significant ( $p = 0.000354$ ). This indicates that the effect of size on price is not constant but depends on the number of rooms.

**3.2.2 Logarithmic Transformation:** A major breakthrough was applying a logarithmic transformation to both the dependent variable (Price) and the key independent variable (m2) to create model4. This transformation successfully linearized the relationship.

```
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  50905.37    27905.19   1.824 0.068858 .
m2           2057.83     352.94    5.830 1.13e-08 ***
Rooms       -36637.52    8254.95  -4.438 1.17e-05 ***
Bathrooms    22296.65    7365.08   3.027 0.002626 **
Elevator     7722.08     6730.35   1.147 0.251917
Atico        16062.29    11791.32   1.362 0.173891
Terrasse     8410.14     8222.26   1.023 0.306992
Parking      47076.18    11955.67   3.938 9.70e-05 ***
Yard         31677.87    16182.12   1.958 0.050969 .
m2:Rooms      304.64      84.56    3.603 0.000354 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 63180 on 403 degrees of freedom
Multiple R-squared:  0.7498,    Adjusted R-squared:  0.7442
F-statistic: 134.2 on 9 and 403 DF, p-value: < 2.2e-16
```

Model 3

```
Coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.414823    0.222013   37.902 < 2e-16 ***
log(m2)      0.873776    0.054837   15.934 < 2e-16 ***
Rooms       -0.100428    0.050077   -2.005 0.045581 *
I(Rooms^2)   0.015093    0.008338    1.810 0.071014 .
Bathrooms    0.115981    0.026116    4.441 1.16e-05 ***
Elevator     0.060001    0.023869    2.514 0.012332 *
Atico        0.088991    0.038536    2.309 0.021431 *
Parking      0.164928    0.043035    3.832 0.000147 ***
Yard         0.088315    0.058294    1.515 0.130559
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2277 on 404 degrees of freedom
Multiple R-squared:  0.7011,    Adjusted R-squared:  0.6952
F-statistic: 118.4 on 8 and 404 DF, p-value: < 2.2e-16
```

Model 4

**3.2.3 Non-linear Effect of Rooms:** We identified a potential quadratic relationship for the number of Rooms by adding  $I(\text{Rooms}^2)$ . The negative linear and positive quadratic coefficients in later models suggested a complex, U-shaped relationship where price might decrease slightly from 1 to 2-3 rooms and then increase for larger dwellings.

**3.3 Feature Engineering:** The binary variables Yard and Terrasse were merged into a new variable Outdoor (model5). This was a logical step to create a more generalized feature for any type of private outdoor space, simplifying the model without losing predictive power.

```

Coefficients:
(Intercept)  8.49989  0.22499  37.779  < 2e-16 ***
log(m2)      0.85332  0.05553  15.366  < 2e-16 ***
Rooms       -0.10452  0.05000  -2.090  0.0372 *
I(Rooms^2)   0.01602  0.00833  1.924  0.0551 .
Bathrooms    0.11627  0.02604  4.465  1.04e-05 ***
Elevator     0.05391  0.02397  2.249  0.0250 *
Atico        0.05274  0.04150  1.271  0.2045
Parking      0.17290  0.04186  4.130  4.41e-05 ***
Outdoor      0.06003  0.02798  2.146  0.0325 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2271 on 404 degrees of freedom
Multiple R-squared:  0.7028, Adjusted R-squared:  0.6969
F-statistic: 119.4 on 8 and 404 DF, p-value: < 2.2e-16

```

Model 5

```

Coefficients:
(Intercept)  8.538117  0.223139  38.264  < 2e-16 ***
log(m2)      0.847020  0.055354  15.302  < 2e-16 ***
Rooms       -0.113070  0.049580  -2.281  0.02309 *
I(Rooms^2)   0.017115  0.008292  2.064  0.03964 *
Bathrooms    0.121350  0.025751  4.712  3.37e-06 ***
Elevator     0.053021  0.023978  2.211  0.02758 *
Parking      0.167191  0.041653  4.014  7.12e-05 ***
Outdoor      0.073527  0.025905  2.838  0.00476 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2273 on 405 degrees of freedom
Multiple R-squared:  0.7016, Adjusted R-squared:  0.6964
F-statistic: 136 on 7 and 405 DF, p-value: < 2.2e-16

```

Model 6

**3.4 Final Variable Selection:** The variable Atico remained statistically insignificant (p-value = 0.204) even in the log-transformed model, so it was removed to yield a cleaner, more efficient model (model6).

**4 Geographic Fixed Effects:** The most significant improvement came from incorporating City Zone as a fixed effect (model7). Location is a paramount factor in real estate valuation. By adding factor('City Zone'), we controlled for all unobserved, time-invariant characteristics of each district. Adjusted R-squared jumped dramatically and coefficients for geographic areas were highly significant.

Crucially, after controlling for location, the variables Rooms and  $I(\text{Rooms}^2)$  became statistically insignificant. This suggests that the apparent nonlinear effect of rooms was actually a proxy for location-specific housing characteristics.

```

Elevator     0.045861  0.020717  2.214  0.027420 *
Parking      0.148346  0.035705  4.155  3.99e-05 ***
Outdoor      0.086857  0.022350  3.886  0.000119 ***
factor('City Zone')Eixample  0.127883  0.044348  2.884  0.004146 **
factor('City Zone')Gràcia    -0.106433  0.046160  -2.306  0.021642 *
factor('City Zone')Horta - Guinardó -0.043036  0.044127  -0.975  0.330022
factor('City Zone')Les Corts  0.070028  0.064810  1.081  0.280572
factor('City Zone')Nou Barris -0.216522  0.042873  -5.050  6.74e-07 ***
factor('City Zone')Sant Andreu -0.142134  0.043721  -3.251  0.001249 **
factor('City Zone')Sant Martí  -0.157312  0.046708  -3.368  0.000831 ***
factor('City Zone')Sants - Montjuïc -0.111816  0.056453  -1.981  0.048317 *
factor('City Zone')Sarrià - Sant Gervasi 0.326622  0.061043  5.351  1.49e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.193 on 396 degrees of freedom
Multiple R-squared:  0.7894, Adjusted R-squared:  0.7809

```

Model 7

**5 Final Model Selection:** Therefore, the final model (final\_simple) was simplified by removing the room-related variables, resulting in a powerful and interpretable specification:

**Final Model:**

$\log(\text{Price}) \sim \log(m2) + \text{Bathrooms} + \text{Elevator} + \text{Parking} + \text{Outdoor} + \text{factor('City Zone')}$

