We started trying out the following models:

1. Model 1: m^2 , Rooms, Bathrooms as the independent variables and Price as the dependent variable: Adjusted R^2: 0.7221

2. Model 2: m^2 as the independent variable: Adjusted R^2: 0.709

3. Model 3: Rooms as the independent variable: Adjusted R^2: 0.2493

4. Model 4: Bathrooms as the independent variable: Adjusted R^2: 0.393

5. Model 5: m^2, Bathrooms as the independent variables: Adjusted R^2: 0.7167

6. Model 6: m^2, Room as the independent variables: Adjusted R^2: 0.7121

7. Model 7: Bathrooms, Rooms as the independent variables: Adjusted R^2: 0.4319

8. Model 8: All variables as the independent variables, with dummy variable set for 9 city zones: Adjusted R^2: 0.8235

From Model 8, we decided to drop the variables Rooms, Elevator, Terrasse and Type because they have significantly high p values.

We also decided to use a random function on excel to split the data into parts of 80 and 20. This way we used 80% of the data to train the model and tested the model on the remaining 20% of the data.

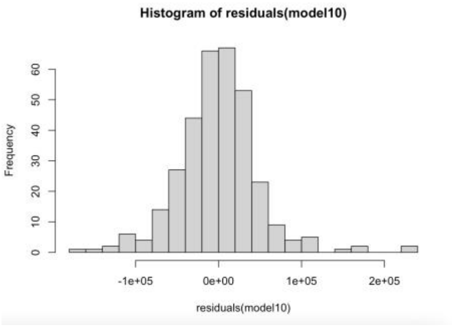
9. Model 9: All variables as the independent variables excluding Rooms, Elevator, Terrasse and Type, including each of the city zone dummy variable multiplied with the area m^2 In Model 9 we were trying to study the combined effects of the city zones with the area.
   Adjusted R^2: 0.6939
   This model used the interactions between the area of an apartment and the location of it. On further inspection we noted that there were some areas where the parameters desired were different (eg: for City Zone Eixample, elevators were significant) On accommodating those into the model we noticed that this reduced the R^2 significantly. Hence we had to pick a different model.

10. Model 10: All variables, with dummy variable set for 9 city zones, except Rooms, Elevator, Terrasse and Type
    Adjusted R^2:0.8244

After comparing the adjusted R^2, we have decided to use model 10 as it has the highest adjusted R^2 and hence the best fit.

We did the following checks on this model:

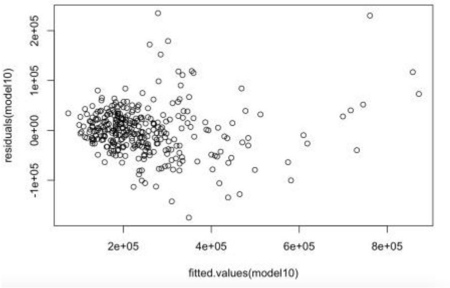
a. Normality assumption check:



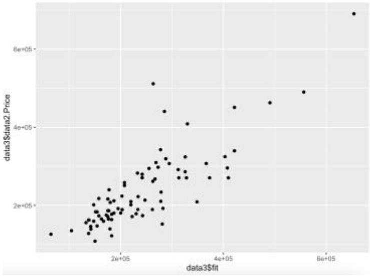
The histogram is bell-shaped and rather symmetrical. Hence we can say that the error is normally distributed.

b. Constant variance check:

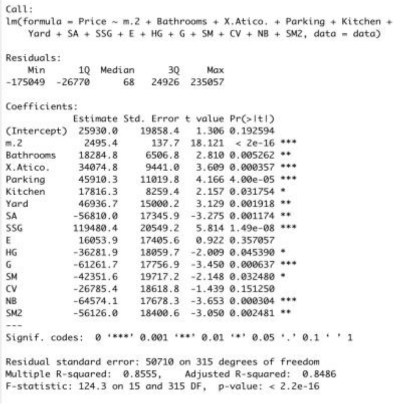
(i) Residuals vs Fitted values
The points on the plot are scattered.



c. Linear fit check:
Price vs fit: the points look pretty linear.



d. P-value check:



```
Call:
lm(formula = Price ~ m.2 + Bathrooms + X.Atico. + Parking + Kitchen +
    Yard + SA + SSG + E + HG + G + SM + CV + NB + SM2, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-175049  -26770      68   24926  235057

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  25930.0    19858.4   1.306 0.192594
m.2           2495.4      137.7  18.121  < 2e-16 ***
Bathrooms    18284.8     6506.8   2.810 0.005262 **
X.Atico.     34074.8     9441.0   3.609 0.000357 ***
Parking      45910.3    11019.8   4.166 4.00e-05 ***
Kitchen      17816.3     8259.4   2.157 0.031754 *
Yard         46936.7    15000.2   3.129 0.001918 **
SA          -56810.0    17345.9  -3.275 0.001174 **
SSG         119480.4    20549.2   5.814 1.49e-08 ***
E            16053.9    17405.6   0.922 0.357057
HG          -36281.9    18059.7  -2.009 0.045390 *
G           -61261.7    17756.9  -3.450 0.000637 ***
SM          -42351.6    19717.2  -2.148 0.032480 *
CV          -26785.4    18618.8  -1.439 0.151250
NB          -64574.1    17678.3  -3.653 0.000304 ***
SM2         -56126.0    18400.6  -3.050 0.002481 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 50710 on 315 degrees of freedom
Multiple R-squared:  0.8555,    Adjusted R-squared:  0.8486
F-statistic: 124.3 on 15 and 315 DF,  p-value: < 2.2e-16
```

We noted that the P-Value was low for most variables and the Adjusted R-squared was a value of 0.8486 after segmentation of the data.

We were unable to remove the dummy variables E (for City Zone Eixample) with a high P-value of 0.357 because that would change the interpretation of all the other variables that would be affected by it.
Hence, we decided to go with this model.