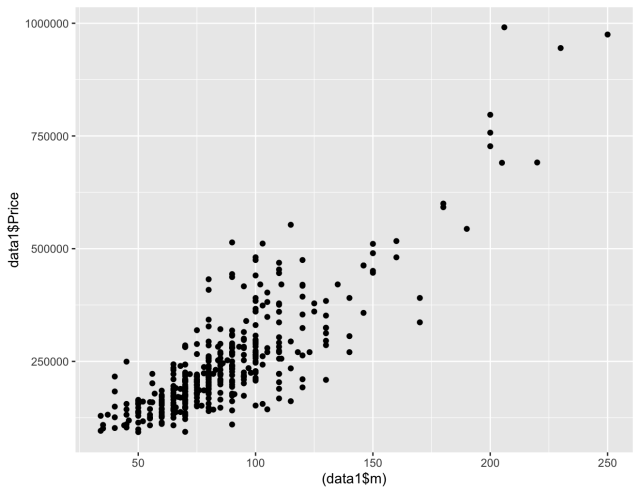
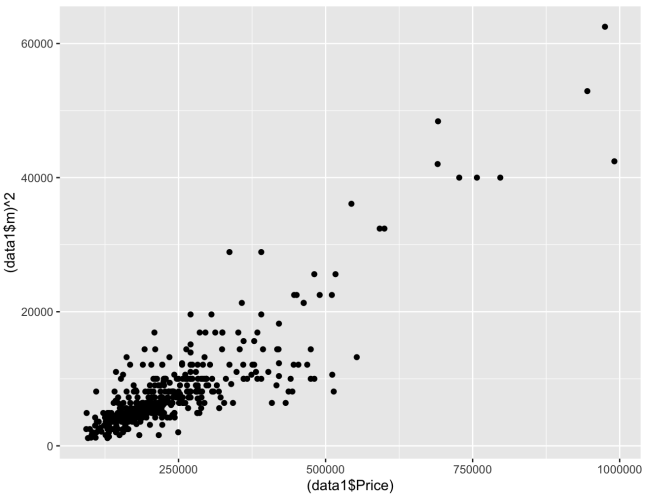
Barcelona Real Estate Price Prediction Case Study
Section A - Team 50

Kavya Gupta: kg355
Jimmy Li: jl1341
Syed Irtiza Mehdi: sm1106
Yaxuan Qi: yq117

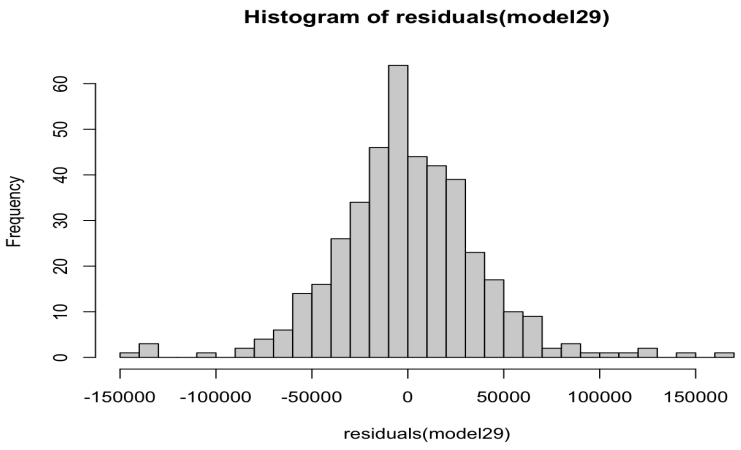
Using a dataset of 413 properties for which prices and various characteristics were known, our team worked on estimating the prices of 200 residential homes in Barcelona. In order to find the best fit, we built several models in order to address this problem.

We started by examining the price-to-property size (m2) connection and found it to be non-linear. In order to solve this, we squared the property size variable, which helped to make the connection linear and stabilise the variance.





Zones were found to be among the most important variables affecting property prices, and in order to incorporate them into our regression model, we set up dummy variables, where the Zone "Sarria - Sant Gervasi" was assumed as the reference category. We hypothesised that the impact of

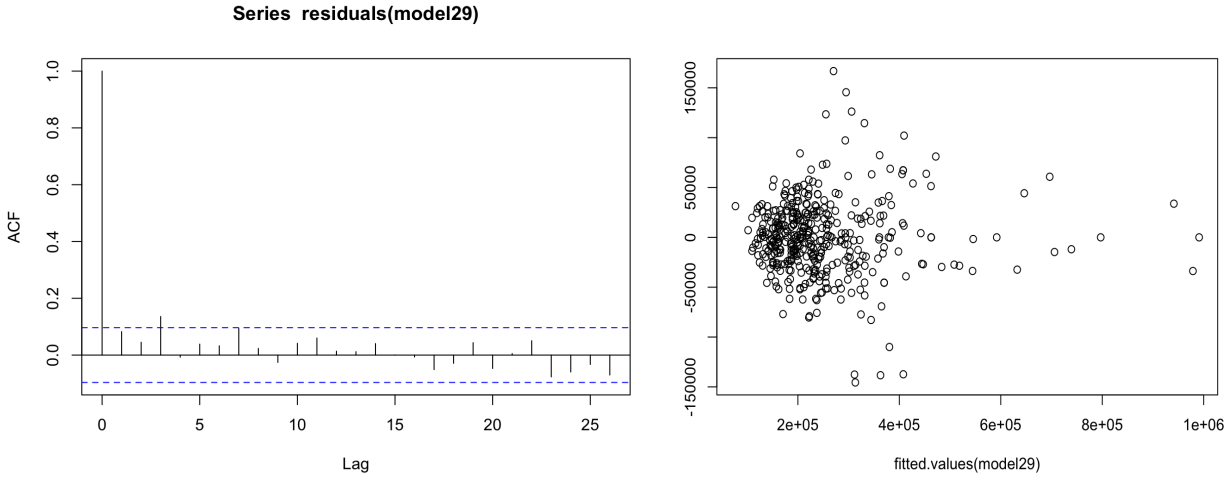


additional elements (such as rooms, size, etc.) would change based on the zone, and introduced interaction variables between Zone and other parameters in order to capture this.

We constructed thirty different models, improving our model in each round by introducing and eliminating variables and interaction terms. For instance, we first thought of incorporating a logarithmic modification of the property size, but we decided against it because the link with pricing didn't match up well with theory.

We kept a careful eye on the p-values of our coefficients as we refined the model. Making sure the intercept's p-value was negligible—that is, that it suggested a greater likelihood that its impact may be zero—was a top priority because it was consistent with our theoretical analysis.

After running the models, we carefully examined a variety of graphs, paying special attention to the residuals and their distribution. We also looked at the autocorrelation function (ACF) to make sure no trends went unnoticed.



During the modelling process, we encountered challenges such as syntax errors while creating dummy variables and other transformations. Iterative testing and meticulous debugging were used to fix this.

The squared term for property size and other interaction factors with Zone were included in the final model. We select this model, because it offers a balance between predicted accuracy and complexity. A strong fit was shown by the residual standard error and R-squared values, and the important factors such as zone, size, and bathrooms had substantial and theoretically consistent coefficients.

```
Residual standard error: 44640 on 298 degrees of freedom
Multiple R-squared:  0.9076,    Adjusted R-squared:  0.8723
F-statistic: 25.68 on 114 and 298 DF,  p-value: < 2.2e-16
```