DECISION 518Q
Section B Team29
Aarush Ambasht, Jinghua He, Sam MacArthur, Sanjna Pandhi, Skylar Qiu
aa901, jh982, sem150, sp806, zq67

Barcelona Real Estate Case:

Overview and Approach:

Using the 413 residential properties included in Barcelona Real Estate data and their respective amenities/features, we built a linear regression model to estimate the price of the additional 200 "mystery" properties. After reviewing the data on face value, we decided to approach the task using the modelling techniques we have learned so far in the course. Moreover, we used the price of the 413 listed properties as the dependent variable and the amenities/features as independent variables to provide insight into the value of the "mystery" properties.

Modelling Choices:

We used R to prep the data, apply transformations, and develop a useful model for predicting prices. Below is a brief synopsis of our process:

After installing the necessary packages and loading our data, we first performed a log-transformation on the target variable (price). We are using the log(price) transformation as it results in a more normal distribution of the typically right-skewed real-estate prices which improves linear regression performance. It also converts multiplicative relationships (common in real estate) into additive ones that linear models handle better. The coefficients then represent percentage changes rather than absolute dollar changes, making them more interpretable across different price ranges. Then, after viewing a summary of the transformation, we created a log-transformed price variable and rebuilt the model several times to exclude non-significant variables (after checking the summary). More specifically, we used a trial-and-error style approach, systematically removing "non-significant" variables based on p-values. We addressed limitations with categorical variable levels by removing what we considered to be non-significant city zones. By doing so, we could focus the model to reflect zones with relevant price differences. Our final model includes the statistically significant city zones (with refined factor levels), property size, number of rooms, elevator feature, and the presence of a terrace.

```
model_refined01 <- lm(lnPrice ~
                        `City Zone` +
                        `m^2` +
                        Rooms +
                        Elevator +
                        Terrasse,
                      data = data.BRE)
```

Overall, we wanted the model to demonstrate a balanced perspective on Barcelona real estate, maintaining only statistically significant predictors while keeping with practical interpretability. For a more accurate and sophisticated reflection of Barcelona's real estate prices, we would need significantly more data points.

DECISION 518Q
Section B Team29
Aarush Ambasht, Jinghua He, Sam MacArthur, Sanjna Pandhi, Skylar Qiu
aa901, jh982, sem150, sp806, zq67

Takeaways:

We predicted prices for 200 Barcelona residential properties using data from 413 similar properties. The R code reveals a *systematic approach to building a predictive model*.

→ **Location (City Zone)** is a **critical predictor** but *not all zones are statistically significant*.

→ **Property size and room count** are *important predictors*

→ The **iterative refinement process** shows the importance of statistical significance in model building

Success is measured by minimizing the average squared prediction errors across the 200 test properties, emphasizing accuracy over complexity. **This reflects real-world data science practices where simpler, statistically sound models often outperform complex ones**.