

# Data-Driven Exploration of Airbnb Listings Price

Data Mining and Machine Learning

MSc in Artificial Intelligence and Data Engineering

2023/2024

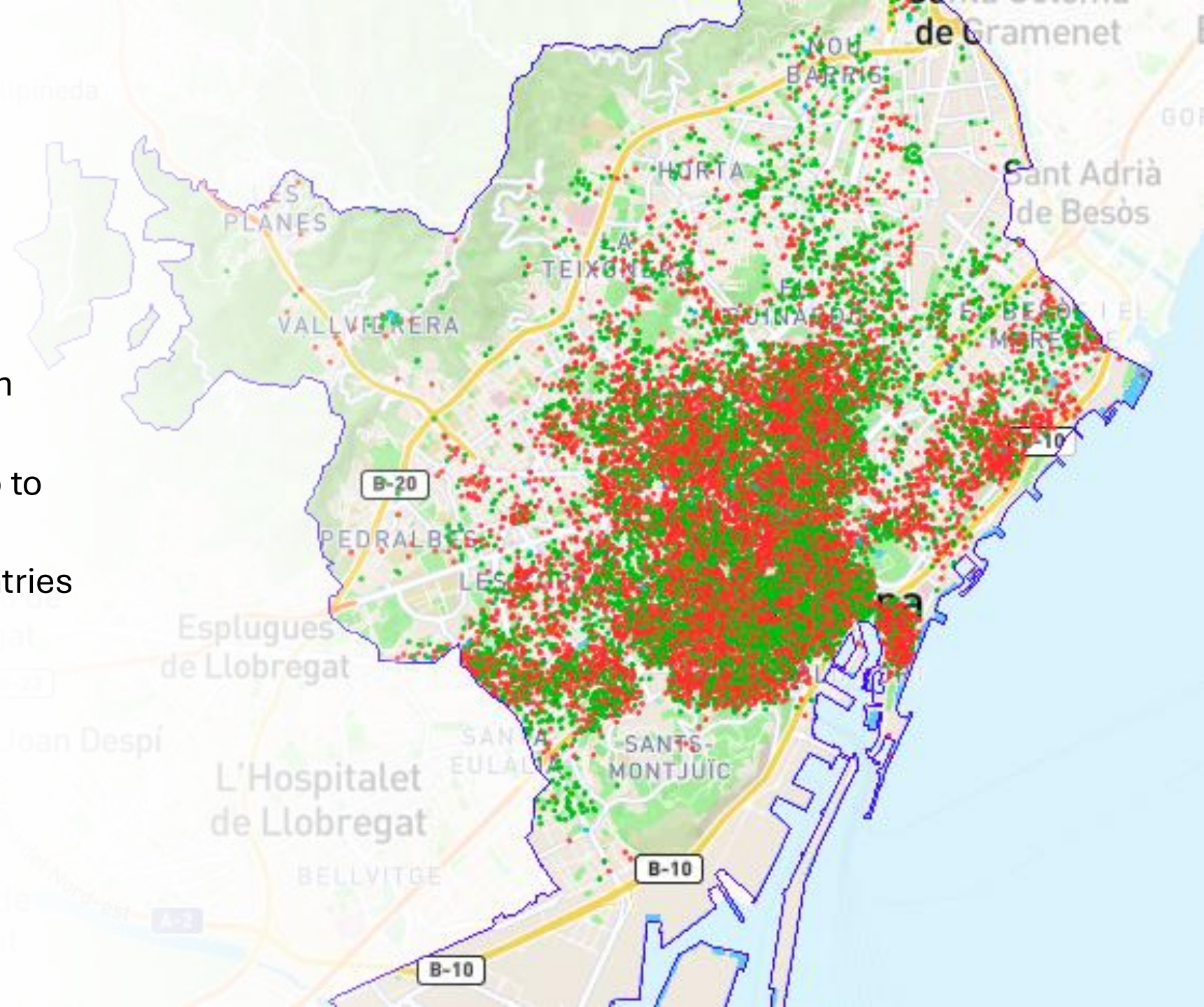


UNIVERSITÀ DI PISA

*Christian Petruzzella*

# Dataset

- The data collected is taken from *insideairbnb.com*
- Barcelona, information up to date on 5 July 2024
- *Listings* dataset: 18925 entries
- *Reviews* dataset: 882263 entries



# Preprocessing

- removed columns where all values were the same
- removed uninformative features like URLs (*listings\_url*, *host\_url*), host details (*host\_id*, *host\_name*, *host\_about*)
- converted some features into binary format
- removed special characters like dollar signs and percentages
- converted date columns such as *host\_since* into numeric values representing the number of days since the date of reference (2024-07-04)


host_since	host_response_rate	host_is_superhost
2011-11-16	100%	f
2011-11-16	98%	f
2010-01-19	95%	f



host_since	host_response_rate	host_is_superhost
4614.0	100.0	0.0
4614.0	98.0	0.0
5280.0	95.0	0.0

host_response_time
within an hour
within an hour
within an hour

Handled categorical features using the *OneHotEncoder*, converting them into binary columns



host_response_time_within a day	host_response_time_within a few hours	host_response_time_within an hour
0.0	0.0	1.0
0.0	0.0	1.0
0.0	0.0	1.0
0.0	0.0	0.0
0.0	0.0	1.0

# Handling amenities



# Sentiment Analysis on the Reviews

listing_id	id	date	reviewer_id	reviewer_name	comments
18674	4808211	2013-05-27	4841196	Caron	Great location. Clean, spacious flat. Would re...
18674	10660311	2014-03-02	11600277	Juan Carlos	Mi mejor recomendación para este departamento....
269467	223325774	2018-01-01	41885149	Anita	Perfect place in a perfect area of Barcelona
269467	247939103	2018-03-30	120056532	Michael	Nice apartment



listing_id	sentiment
10003263	0.405025
1000432235531291341	0.336934
1000447810456915898	0.376677
1000466852301593784	0.432986
1000514588274707195	0.364167

# Experiments and Discussion

A thick, hand-drawn style orange line underlining the text.

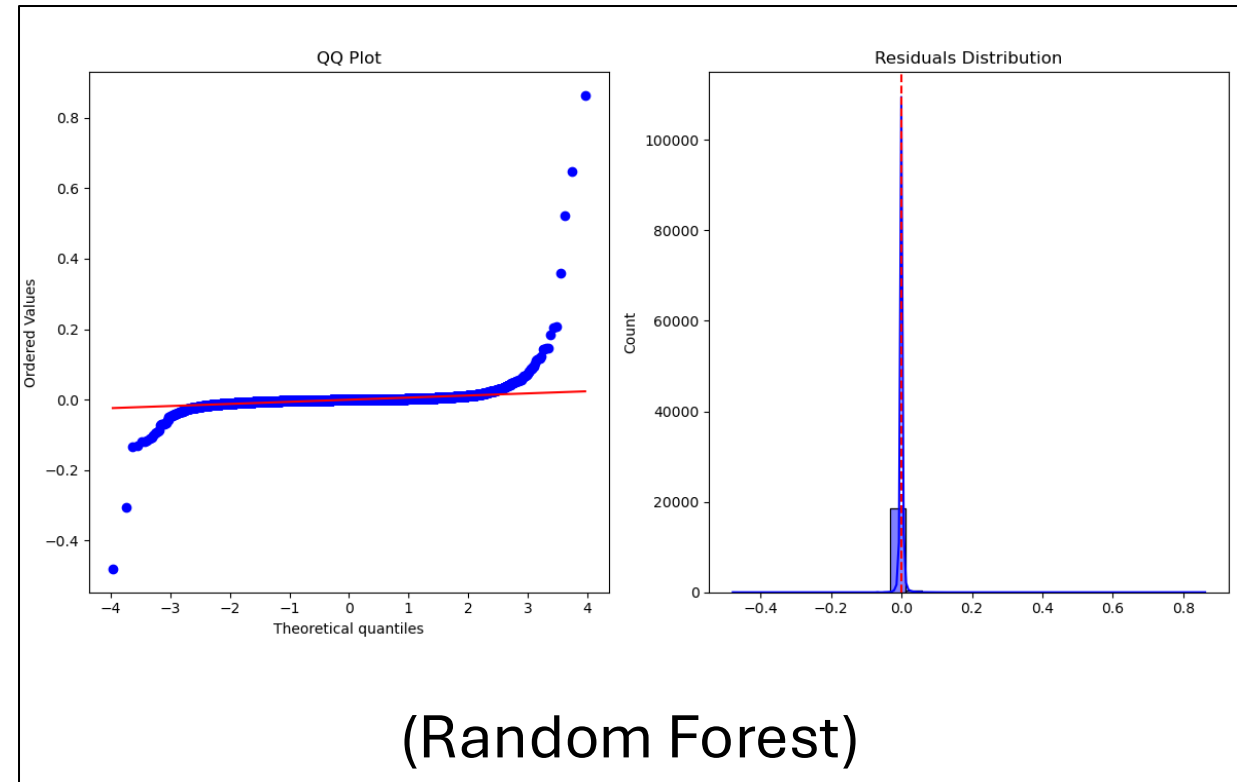
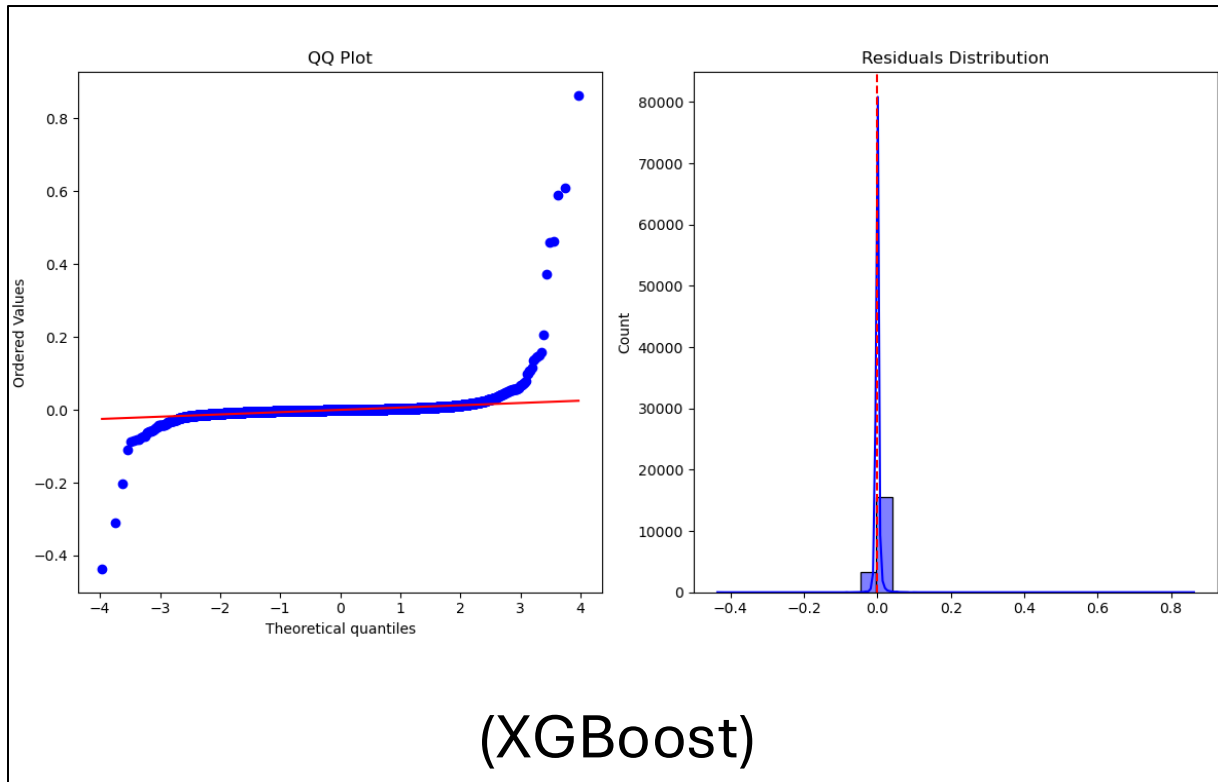


# Regression Models Evaluation

A 10-fold cross-validation of the data has been performed with various regression approaches.

Model	MSE	MAE	RMSE	R <sup>2</sup>
Lasso	0.00037	0.0079	0.01823	-0.0011
Ridge	0.00027	0.00500	0.01501	0.3635
Bayesian Regressor	0.00027	0.00502	0.01501	0.3627
K-Nearest Neighbors	0.00025	0.00475	0.01474	0.3651
Decision Tree	0.0002	0.00513	0.01342	0.3947
XGBoost	0.00018	0.00367	0.01211	0.5611
Random Forest	0.00016	0.00334	0.01163	0.5443

# Residuals Analysis

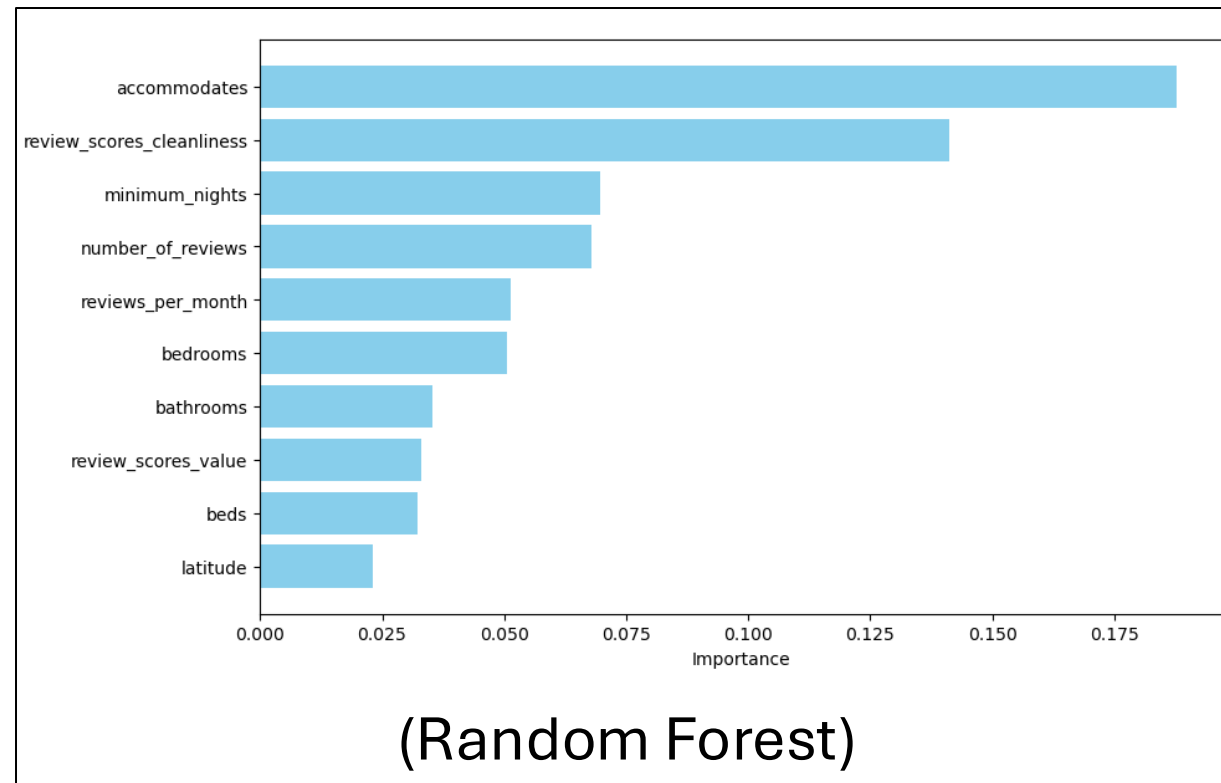
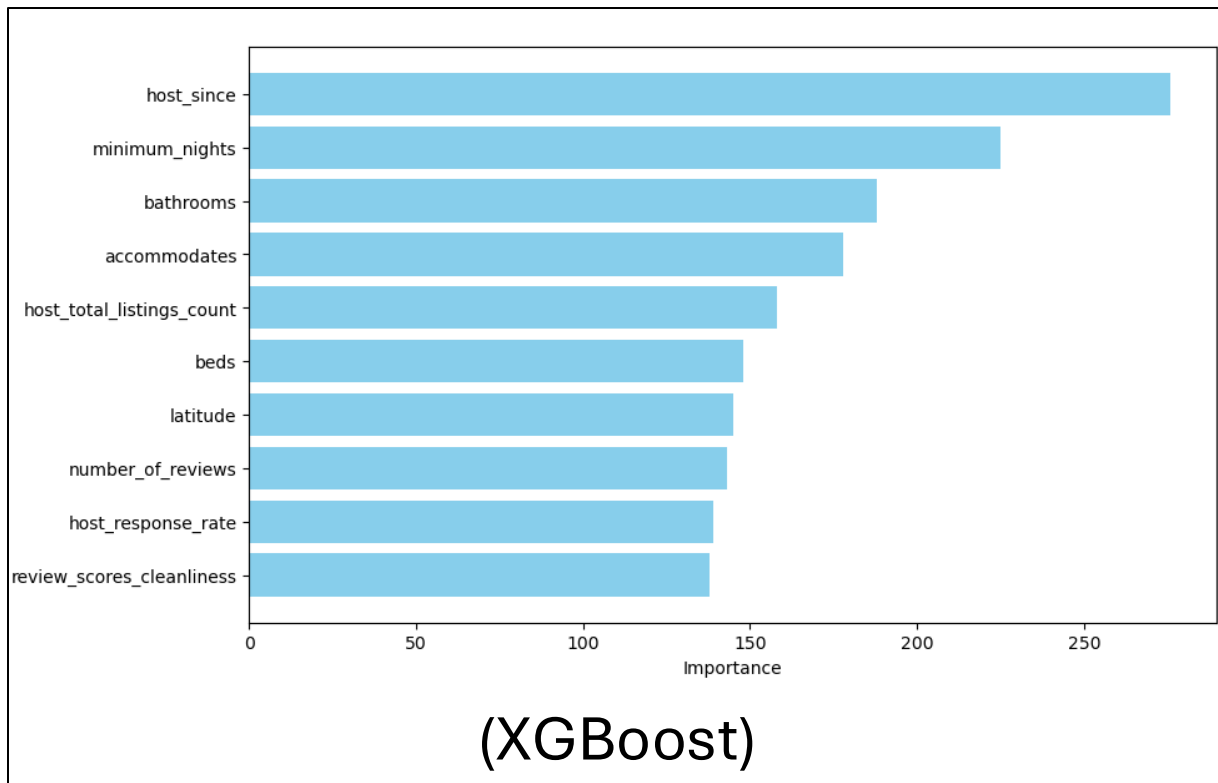


Wilcoxon statistic: 82220337.0

p-value: 1.9616517018557887e-22

Reject null hypothesis: There is a significant difference between the two models' residuals.

# Feature Importance Analysis



The property size (*accommodates*, *beds*, *bedrooms*), guest feedback (*number of reviews*, *review score*) and minimum stay requirements emerge as the most important factors in determining listing prices.