

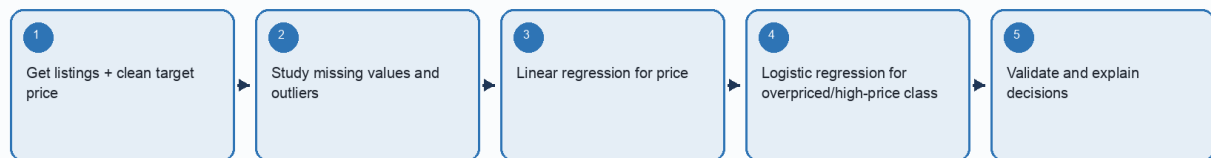
Machine Learning

Practical Project 1 - Supervised Learning

Dataset	Airbnb listings.csv.gz ; suggested city: Lisbon. You may use another city if the same variables are available.
Main question	Can nightly price be predicted, and can unusually expensive listings be detected, using interpretable supervised models?
Core models	Linear regression, polynomial regression, regularized variants, and logistic regression (classification).
Data difficulty	Missing values, categorical variables, skewed prices, spatial effects, and extreme outliers are expected.

Project 1 scheme

Supervised learning with messy structured listings



Central idea: compare what the model learns from data alone with what changes when representation, objective, or physics constraints are changed.

Project Highlight

Build a complete supervised-learning pipeline for real-estate/Airbnb listings. The project is not only about achieving low error; you must justify every preprocessing choice and explain how missing values and outliers influence both regression and classification.

Major goals / objectives

- Predict the nightly listing price from structured variables using linear regression and meaningful feature engineering.
- Define and predict a binary target such as high-price (overpriced) using logistic regression.
- Compare how different missing-value and outlier strategies affect predictive performance and model interpretation.
- Produce an interpretable report that connects model coefficients, residuals, and classification errors to listing characteristics.

Minor goals / checkpoints

- Convert price, host response rate, amenities, neighbourhood, room type, and property type into usable numeric/categorical features.
- Map missing values and distinguish unknown values from meaningful absence, for example no reviews yet.
- Detect extreme prices and abnormal minimum nights/accommodates values using at least two robust criteria.
- Implement train/validation/test split and k-fold validation without leakage from preprocessing.

- Compare raw price, log-price, polynomial terms, and selected interaction features.
- Use threshold-independent classification measures such as ROC-AUC or precision-recall analysis.

Implementation requirements

1. Create a `data_preparation.py` module that loads the selected city, cleans the target price, encodes categorical variables, scales numeric variables, and saves the final design matrix.
2. Create a `linear_regression.py` module implementing gradient descent for linear regression. Include batch or mini-batch training, convergence plots, and L2 regularization as an extension.
3. Create a `logistic_regression.py` module implementing binary logistic regression with gradient descent. The positive class must be defined clearly, for example top 20 percent price within comparable room type/neighbourhood groups.
4. Compare at least three preprocessing strategies: simple imputation, grouped imputation, and dropping/capping selected outliers.

Experiments to report

- Regression: MAE, RMSE, R2, residual plots, and error analysis by neighbourhood and room type.
- Classification: accuracy, precision, recall, F1, confusion matrix, ROC-AUC, and threshold discussion.
- Outlier study: performance with no treatment, capping/winsorization, and explicit robust filtering.
- Missing-value study: compare at least two imputation policies and explain where each policy is defensible.
- Interpretability: coefficient table for the final linear/logistic models and a short discussion of signs/magnitudes.

What to submit

- Python scripts or notebook with reproducible execution instructions.
- A short report with preprocessing decisions, model comparison tables, and final conclusions.
- Figures showing missingness, outliers, regression residuals, and classification confusion matrix/ROC curve.
- A README explaining the dataset version/city used and how to reproduce the experiments.

Data Source

Inside Airbnb Get the Data: <https://insideairbnb.com/get-the-data/>

Suggested workflow

- Download the selected city, clean prices, map missingness/outliers, and define regression/classification targets.
- Implement preprocessing plus linear/logistic baselines; validate train/validation/test separation.
- Run feature engineering, imputation/outlier ablations, polynomial/interactions, and model comparison.
- Prepare residual/error analysis, coefficient interpretation, final report, and reproducible README.