FINAL PROJECT OUTLINE

1. Problem Formulation and Significance

Problem statement

The project studies how property and host characteristics influence the trailing twelve-month revenue per available room for Airbnb listings in Stockholm. The research question is how strongly features like *room type, beds, average rate, occupancy, reserved days, available days* and *superhost status* explain variation in *ttm_revpar*. The task is multiple linear regression, chosen because it isolates each predictor's contribution to revenue while controlling for all others.

Significance

Anyone who has ever booked an Airbnb has wondered why one place is cheap, another is insanely expensive, and a third has the same price. Hosts wonder the same things from the opposite side. They need to know which parts of their listing actually drive revenue and which details barely matter. Intuitively one would guess that room size and location is what drives revenue, but by modeling these relationships with other features the project aims to explore how different features translate into earnings, which reduces uncertainty for both hosts and guests and makes the whole market function better.

The findings will be useful for both hosts and guests, as the modeling increases knowledge in how to both set a reasonable price and to recognize a good deal.

Dataset description

The dataset includes 300 Stockholm Airbnb listings from public AirDNA data. It contains both qualitative and quantitative features: *room type, beds, average rate, occupancy, reserved days, available days and superhost status*. The target is *ttm_revpar*. Missing values are only found in beds, and there are no duplicate rows found. The dataset is appropriate for regression because the predictors directly relate to pricing, demand and property characteristics.

2. Exploratory Data Analysis

Data exploration

The dataset was first inspected to understand the distribution and behavior of each feature before modeling. Early exploration focused on checking how the features behave individually and how they relate to revenue, where linear regression was tested on for example Occupancy vs RevPar.

Missing and imbalanced data

The dataset was checked for missing values and that only the beds column had incomplete entries. The missing values rows were removed by utilizing .dropna(). The distribution across room types and superhost status is adequate for modeling and does not require balancing. The absence of significant imbalance indicates no further need for additional preprocessing.

Data visualization

Planned visuals include scatterplots, histograms, and boxplots of linear regression, residuals, and noise.

3. Model Selection, Application, and Evaluation

Justification for model choice

Multiple linear regression is selected because the target variable is continuous and the goal is to quantify how each predictor influences revenue when all others are held constant. The dataset includes meaningful numeric variables and a small number of categorical variables that can be encoded without issues. This model is consistent with class topics and fits the interpretation goals. OBJ

Method application

You will implement your own linear regression code from scratch, using matrix algebra or gradient descent, following course rules. You may verify results with a library but not use library regression as the primary model. Hyperparameters to discuss include any regularization (if added) and how you encoded categorical variables like room type. You also need to describe how you selected which features to keep, how you normalized data, and how you split training versus testing.

Model evaluation

Metrics include R-squared and RMSE. R-squared evaluates how much of revenue variation the model explains. RMSE measures prediction error in actual currency units.

4. Results, Conclusions, and Real-World Implications

Results presentation

Tables that summarize coefficients, RMSE and R-squared. Plot of predicted versus actual ttm_revpar. Highlight which predictors have statistically meaningful impact, e.g. occupancy rate, beds, superhost. All visuals will be reproducible through the notebook provided in the repository.

Conclusions

Summarizes what the coefficients mean in practical terms. Argue for which features matter, why they matter and what the limitations are, including any multicollinearity or low-sample issues. Conclude and answer the research question about which characteristics drive revenue differences in Stockholm Airbnb listings.

Workforce or graduate school preparation

Reflect on skills gained: implementing regression from scratch, debugging numerical issues, working with real-world messy data, interpreting model outputs, producing reproducible visualizations and writing a structured research paper. Discuss how these skills apply to future careers in consulting/banking.