**Modeling Ideas (prompt)**

• What precisely is the data science problem? (e.g., classification vs regression)

• Is the data science task a supervised or unsupervised task?

• What is a data instance in your data set?

• What might be the target variable?

• What features would be useful in predicting the target variable?

**Modeling Ideas**

This project presents a supervised regression problem, since the goal is to predict the price of a listing based on its attributes. Linear regression, decision trees and nearest neighbors can be used to build different predictive models for the listing price, each of which can be evaluated using nested k-fold cross-validation.

* Each Airbnb listing represents a separate data instance, encoded as distinct rows in the data
* The target variable of interest in this case is the log price of a listing
* A total of 28 relevant features are included in this dataset (excluding id column and log data).
  + When looking at factors that influence the price of a listing, it is likely that the location e.g. city and neighborhood variables will prove useful in the model.
  + Similarly, details on the space available e.g. property type (apartment vs. house) and room type (shared room vs. entire home) will capture meaningful variation in price.
  + Some features will require transformation prior to model inclusion as well. For instance, decoding the ‘amenities’ provided by a listing into wireless, television, etc. will be necessary.
  + Automatic feature selection will be included in certain techniques, while additional feature selection may be implemented as needed for other algorithms, such as k-NN.