Had my review w/ Max. Overall, he had a good amount of positive feedback and helped identify some growth areas.

He liked:

* That I sourced my own data
* My data cleaning/prep was thorough and explained well
* That I created and utilized my modularized functions
* Didn’t tear into the readme! Yay!
  + Self-perceived weak point in past
  + Wasn’t super-confident this time around

Areas to improve:

* Readme:
  + Include EDA visualizations
* Functions:
  + Some functions were missing docstring details
* Dataset Background/Explanation/Details:
  + Needed more explanation of why I chose a 4.8/5 threshold (seemed arbitrary)
    - The 4.8 threshold is one of the requirements for the “SuperHost” status
  + Also suggested either changing my target to be whether a property would be a SuperHost; or to drop the feature entirely due to the significant connection to the dataset.
  + RE: low number of scores <= 4.0: explain potential reasoning
    - E.G. strong requirements/aggressive policies penalizing low scores or potentially removing from the dataset
    - Potential limitation of data collection – could only retrieve properties with high scores vs. being able to see all records of all properties
* Lacked proper EDA, visualizations
  + He suggested using the EDA step to identify/create hypotheses about the modeling results
* Modeling - particularly weak area
  + Outline/define target
    - Why macro recall?
      * Why minimize false positives?
    - What is the target value?
      * Eg. Targeting 1.0, but getting .7?
    - Use this specification/goal for evaluation post-modeling
  + Lack of iterative modeling/testing different models
    - Suggested including separate JNB that includes the “unused” modeling
    - Personal idea
      * 1. splitting into cleaning/EDA -> save cleaned data
      * 2. Iterative modeling notebook -> best models, best results
      * 3. Final results and analysis
    - Concern: David didn’t like having the rough/WIP notebooks in the repo
      * This could be considered different as I would want to show the steps taken to get to the final model vs. the WIP parts of the notebook
* Better utilization of GSCV
  + Expand to include pipeline parameters – not just model params
    - Eg. SimpleImputer strategy, missing indicator; etc.
  + Using GSCV for training AND validation
    - Pulling training, test scores from GSCV (didn’t know that existed!)
    - Using best model, params in new model and comparing results to split X\_test/y\_test
* Tweaking reports:
  + “metric” parameter – currently just a placeholder for use in f-string for model’s default scoring
    - Most often “accuracy”, but could vary
    - No impact on actual score; just used for identification purposes
    - CHANGES: GSCV scoring metric
      * eg. If using macro recall, specify “balanced\_accuracy” and have that pass into the classifier
        + if/else statements for changing terms to “average recall”, etc.