Homework Assignment

Thank you for continuing on this interview process with us and we are excited to see your results from this task. At a high level, we are looking for how you approach new datasets, perform first order statistical analysis, and go about developing a model for prediction/classification. The attached dataset contains building energy efficiency data and so should be somewhat related to the work that we are both engaged in.

We hypothesize that this dataset contains independent variables that might be used to infer interesting facts about buildings in New York. Of primary interest to us is the Energy Star score, since we believe this score is used as an aggregate measure of a building's energy performance.

Our question is simple to ask, but difficult to answer:

- Does this dataset contain a set of independent variables that correlate to the energy star rating of the building?
- For buildings without energy star scores, can you infer what their score might be?

To answer this question we ask that you prepare a reasoned investigation of this hypothesis in a Jupyter notebook. We believe the best method is to perform visual statistical descriptions of the data using summary tables or a graphing library of your choice. Then train a model and evaluate it. You might regress data to predict the numeric score - or bin the scores into "grades" (A, B, C, D, F) and apply a classifier. You might cluster buildings to determine if similar groups exist.

Create one or more of these models and give us an interpretation of their performance, which you can then use to answer the question of whether this dataset contains enough information to understand the relationship of an energy star score to buildings of different types. For the model performance, we would like to see typical/appropriate evaluation metrics as such as F1 scores, R-squared scores, etc.

We don't want you to spend more than 2-4 hours on this task so please don't look at this as a software engineering evaluation as well. We look forward to seeing your results.