Relationship between the quantity of ratings and rating score on Yelp

Data Set

- https://www.yelp.com/dataset/challenge

Problem Statement

- Does knowing the amount of reviews a reviewer has already submitted to Yelp and the number of reviews a business already has on Yelp help predict the likely rating that the reviewer will end up giving for a particular business when writing a new rating?

Hypothesis / Assumptions

- That there are clear data relationships between both the amount of reviews a reviewer has already submitted to Yelp and the number of reviews a business already has received on Yelp, and that these help predict at least partially the likely rating that the reviewer will end up giving for a particular business.
- That segments of reviewers can be identified that show patterns of ratings scoring based on frequency.

Goals and success metrics

- That a predictive model can found that beats baseline accuracy.

Risks and Limitations

- Yelp may not be providing the full data set which could invalidate the predictive model.
- The data set is large and analyzing the data set is not performant when being accessed from a local drive.
- There may be too much noise in the data to find clear correlations to test the hypothesis.