Deriving a "truer" aggregate rating for services on Yelp

Problem Statement

- When submitting a new review on Yelp, some reviewers may give a rating that is strongly influenced by the reviews and scoring that he or she has already read on Yelp about that service beforehand. Those users who are influenced by the reviews they read before they create their reviews may end up creating a review that becomes a form of "noise" among the body of reviews for a service as a whole that make it harder for general yelp users to determine the "true" rated quality of the service.
- If these kinds of users could be clearly identified and potentially filtered out of an aggregate score for a service, other Yelp users may be able to get a "truer" understanding of what is the actual underlying rating of the service and make a better decision on whether or not to use the service.
- [One other problem I may look at is recalibrating the scores of different segments of reviewers who share the same sentiment but may score more extremely than other segments]

Hypothesis / Assumptions

- That there are a significant number of reviewers who may give a rating that is strongly influenced by the reviews and scoring that he or she has already read on Yelp about that service.
- That these users can be identified and segmented via EDA or predicted via logistical probability.
- That when these users taken are taken out of the aggregate review calculation for a service there a measurable change in the aggregate review for a service in a significant number of cases.
- That followers and influencers can be identified in the Yelp data.
- That patterns of scoring are distinct to particular segments of reviewers and that for certain segments, how the user will score can be predicted more accurately than the baseline.

Goals and success metrics

- Creating of a segmentation of "followers" and "influencers" of users who create Yelp reviews
- Ability to successfully predict the likely rating score of a "follower" for a service he/she has not yet reviewed.
- A recalculated aggregate rating once followers are filtered out.

Risk and Limitations

- The data set is large and analyzing the data set is not performant when being accessed from a local drive.

- The data set is likely going to need to be housed on Big Query or in a similar platform to be analyzed in a performant way
- Even then, given the size of the data set, I still may not be able to do analysis in a performant way.
- After analysis, the Yelp data may not show that some reviewers may give a rating that is strongly influenced by the reviews and scoring that he or she has already read on Yelp about that service. In which case, I would need to alter my problem statement to find a positive insight.