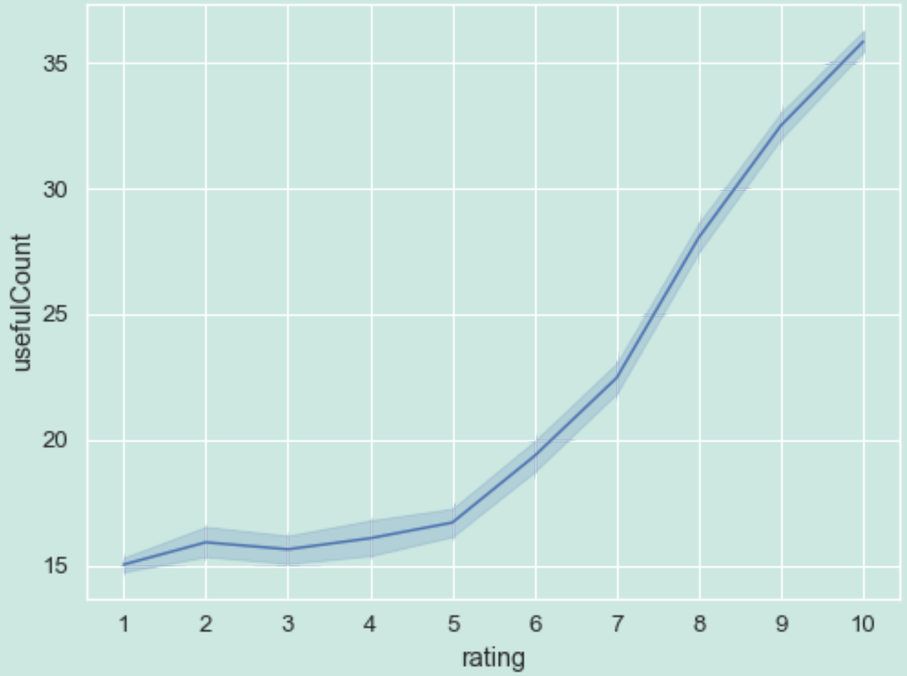
Presentation Outline

Duplication alert!

- Scraping error in data collection caused the dataset to have a large amount of duplicates (not addressed in the research paper!).
- Duplicates = same text input repeated twice => biased model.
- After removal, dataset shrank almost by half.

Usefulness votes

Average of UsefulCount per Rating



Comments associated with positive rating tend to receive more 'useful comment' votes.

However, data is imbalanced towards positive reviews (see next slide):

- Removing less useful ones would decrease negative reviews
- i.e. **increase imbalance**.

=> No removing of records