**Data description**

In this final project, we are using Amazon Review Data. Current data includes reviews in the range May 1996 - Oct 2018. The data comes in gzip files in the form of json. We downloaded the data from the following website: <https://nijianmo.github.io/amazon/index.html>. We used the 5-core data, whom have been reduced to extract the 5-core, such that each of the remaining users and items have 5 reviews each.

We found two shortcomings of the data:

● Customer bias: those customers who write reviews might usually post overly positive or overly negative reviews, because they wouldn’t come up and post otherwise.

● Fake reviews: some reviews may not from authenticate customers: positive reviews with short and repeated sentences may come from merchants’ partners, negative reviews may due to

merchants’ competitors malicious frame.

Right now, we use python to read it into DataFrames and we found the dimensions as below:

* reviewerID - ID of the reviewer, e.g. A2SUAM1J3GNN3B
* asin - ID of the product, e.g. 0000013714
* reviewerName - name of the reviewer
* vote - helpful votes of the review
* style - a disctionary of the product metadata, e.g., "Format" is "Hardcover"
* reviewText - text of the review
* overall - rating of the product
* summary - summary of the review
* unixReviewTime - time of the review (unix time)
* reviewTime - time of the review (raw)
* image - images that users post after they have received the product

**Project objectives**

**Methodology:**

Describe your methodology and discuss the pros and cons for using this methodology to achieve your objectives.

**Results and discussion**

**Conclusion:**

Provide recommendations for the parties interested in your results;

Indicate possible shortcomings of your work and ways to overcome them.