Big Data Challenge homework report

The amazon reviews data was analyzed for any bias in the vine program reviews. The reviews are stored in the gzipped tar files at <https://s3.amazonaws.com/amazon-reviews-pds/tsv/index.txt>

The files were downloaded from the S3 bucket using pyspark and then analyzed.

The analyzed data was written back into the Amazon RDS postgress database.

The data that was used for analysis was from the Home Improvement product category.

The percentage of five star reviews for the vine and non-vine reviews were compared.

Of the 10779 reviews from the paid vine category 5557 reviews were Fivestars. This amounts to a 52% of fivestar reviews.

Of the 2623758 non paid reviews 1650924 were Fivestar which amounts to 63% of fivestar reviews.

Based on this information there is no favorable bias in the paid vine reviews. In fact it looks like there is a negative bias amont vine reviewers.

To analyze the data further the percentages for fivestar reviews were compared for reviews with votes and surprisingly only 26% of the reviews from the paid category received votes and only 20% of the unpaid reviews received votes. It looks like there is some manipulation in the unpaid reviews because eventhough a high number (63%) rated the products at fivestars only 20% of those reviews received a vote.