Big Data Report

Christine Li

In this big-data-challenge homework, I have completed level 1 and level 2 with these two datasets:

1. https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Major_Appliances_v1_00.tsv.gz
2. https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Luggage_v1_00.tsv.gz

Level 1:

Q: Count the number of records (rows) in the dataset

Ans: There are **96,902** rows in the first data set, and **348,613** rows in the second data set.

Level 2:

I choose to use SQL to perform the data gathering from the tables after extracting and transforming the data and inserting them to RDS.

The queries are included in the file *level2/query.sql*

There are 3 criterias that I am looking at to find out whether the vine reviews are truly trustworthy.

1. The number of reviews of vine program (paid) and non-vine (unpaid)
2. The number of 5 stars reviews of vine program (paid) and non-vine (unpaid)
3. The average rating of vine program (paid) and non-vine (unpaid)

Here are the data summary for the first dataset:

- The number of reviews
  - vine program (paid): **248**
  - non-vine (unpaid): **96,640**
- The number of 5 stars reviews
  - vine program (paid): **112**
  - non-vine (unpaid): **49588**
- The average rating:
  - vine program (paid): **4.2540**
  - non-vine (unpaid): **3.7150**

The percentage of 5 stars reviews of the vine program is 112/248 = **45.16%**. The percentage of 5 stars reviews of the non-vine program is 49,588/96,640 = **51.31%**

Here are the data summary for the second dataset:

- The number of reviews
    - vine program (paid): **904**
    - non-vine (unpaid): **347,709**
- The number of 5 stars reviews
    - vine program (paid): **472**
    - non-vine (unpaid): **216,003**
- The average rating:
    - vine program (paid): **4.3584**
    - non-vine (unpaid): **4.2231**

The percentage of 5 stars reviews of the vine program is 472/904 = **52.21%**. The percentage of 5 stars reviews of the non-vine program is 216,003/347,709 = **62.12%**

From the data above, we can see that the average rating of the vine program is higher than the one of the non-vine program, which means that the people in the vine program tend to give a higher rating than the people in the non-vine program.

However, we can also see that more people in the non-vine program give 5 stars review than the vine program.

Therefore, we can conclude that Vine reviews are truly trustworthy because the difference of average rating of vine and non-vine is less than 0.2, which is pretty close. And the same as the difference of percentage of 5 stars reviews, which is within 10%.