**Big Data – Amazon Vine Review Analysis**

**Overview**

Amazon provides user reviews for products sold through their site to aid consumers in making purchases. While this data is made publicly available, the sheer size of the review datasets make them impossible for the average consumer to analyze and find meaningful insights without leveraging big data analysis tools to parse and organize the data.

For the purposes of this analysis, we are reviewing us book data (dataset linked below).

To aid with this process, we are leveraging AWS RDS, Google Colab and Pyspark to host, parse, organize and filter the data and derive insights into the use of paid vs unpaid reviews and whether there is any bias in the reviews given by paid reviewers (vine). To do this, we’ll look at the percentage of vine reviews that are 5 star versus the number of non-vine reviews that give the same rating.

**Related Analysis Files**

Amazon source dataset: <https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Books_v1_02.tsv.gz>

Deliverable 1

1. IPYNB file to Extract and create tables in postgres <https://github.com/Roland791/Amazon_Vine_Analysis-GitHub/blob/main/Amazon_Reviews_ETL.ipynb>
2. Images of Tables in postgres: <https://github.com/Roland791/Amazon_Vine_Analysis-GitHub/tree/main/Deliverable%201%20-%20pgAdmin%20Table%20images>

Deliverable 2

1. Vine Table Analysis: <https://github.com/Roland791/Amazon_Vine_Analysis-GitHub/blob/main/Vine_Review_Analysis.ipynb>

**Results**

The results of this analysis were actually compelling in a very unexpected way. Upon filtering the data to remove any entries where the total votes was less than 20, the results returned absolutely no vine reviews within the dataset:

Table

Description automatically generated

So, while non-vine reviews totaled over 400,000, with 242,889 5-star reviews which account for 60% of the total review scores:

Graphical user interface, text, application

Description automatically generated with medium confidence

Vine reviews had zero five star reviews out of zero total reviews:

Graphical user interface, text, application, email

Description automatically generated

The conclusion here is that there is no bias in vine reviews simply because there were no vine reviews.

Bonus:

In the modern age of the internet and product reviews, it is easy for companies and individuals to review just about anything, even if they don’t have experience with the thing they are reviewing. Bots or agencies creating fake accounts to boost the perception of a product, users protesting products or movies by review-bombing things before they are released and similar tactics can make the reliability of reviews less than ideal. To combat this, some companies such as amazon have included “verified purchase” tags on reviews where they can confirm that a product was purchased by the reviewer.

As with the vine review bias question, we want to investigate whether reviews from verified purchasers and those who haven’t been verified show any variation, and whether unverified reviews might be skewing the data.

To do this we first look at the statistics for the two groups, verified and unverified:

Graphical user interface, application

Description automatically generated

As demonstrated in the figure above, the number of unverified reviews is at 357,607 while the verified is only 46,200. The sheer size of the number of unverified reviews could easily shift the overall rating of the product based on the unverified and therefore less reliable group if there is a skew between the two.

However, when we determine the percentage of views for each group, we can see that both groups have a 5-star rating percentage of ~60%, with nearly identical means and deviations. This would indicate that in the vast majority of cases, unverified and verified reviews are consistently reflective and are therefore likely reliable. A secondary analysis should be done on one star reviews to determine whether the consistency holds true at the other end of the rating spectrum.