Vine Program Bias Analysis via Pet Products

1. **Overview of the Vine analysis:** Explain the purpose of this analysis.

Analysis of large amounts of data can be stressful to not only the user, but the computer. There are many resources to extract data, transform said data into the format you desire, and loading the final product into a medium that you can further alter or visualize while it is at a manageable size. In this case, a company that pays a small fee to Amazon Vine member for their review of products, is inquiring on a concept that may be affecting honest product reviews. The company would like to explore the idea that there may be bias on the positive votes for a product based on the fact that they are receiving a monetary incentive. Through the Amazon RDS services we are able to extract the data on a given shopping category, in this case pet products, and then create data frames through PySpark that can ultimately provide the answer to this companies question.

1. **Results:**

Prior to answering these questions, we removed some of the data that may cause issues in future calculations. First, we filtered our Vine Table to only include products that had 20 votes or more. The purpose of this is to reduce outliers of products that might not be as readily bought, giving a more accurate representation of whether there may be bias or not. We also reduced our data set to only products that a 50% or greater ratio of helpful votes. Because humans will be humans, some review may not even be related to the product or spam. Helpful votes help determine the credibility of a review, and will also aid in our bias analysis.

How many Vine reviews and non-Vine reviews were there?

In order to obtain the number of votes placed by paid and unpaid individuals, the filtered data was further filtered into two data frames representing each monetary category. In regards to the number of reviews that were submitted by paid Vine members, there were 170, while the unpaid reviews totaled 37,840.

How many Vine reviews were 5 stars? How many non-Vine reviews were 5 stars?

In order to decrease variables, we will only analyze the reviews that were 5 stars. There was a total of 65 5 start reviews made by paid individuals, and 20,612 reviews by those who were not paid.

What percentage of Vine reviews were 5 stars? What percentage of non-Vine reviews were 5 stars?

Finally, we are able to determine the percentage of helpful 5-star reviews made by paid members and those who were unpaid. Approximately 38% of the paid reviews were 5-star, while 54% of the unpaid votes were 5-stars.

1. **Summary:** State if there is any positivity bias for reviews in the Vine program. Provide one additional analysis that you could do with the dataset to support your statement.

While some may think there would be a positive bias in the review of products due to the fact that there was a monetary return, this data does not support that theory. As we can see, the unpaid positive reviews were around 15% higher than that of the paid reviews. In order to really form a reliable analysis, we might want to do the same process for other products. Another bit of information that we are provided with is whether the reviewer actually bought the product. We could further filter the data to only include those who were verified purchasers providing accurate feedback of the item.