Product Review Analysis

Summary

The objective of this project is to use PySpark, SQLAlchemy, and Pandas to extract, transform, and load Amazon product review data into an AWS RDS database and to investigate whether Amazon Vine reviews are biased. Customers who have a reputation for writing helpful reviews and for expertise in specific product categories are invited to participate in Amazon's Vine program. These reviewers receive free products in exchange for reviews.

For this project, product reviews for mobile electronics and major appliances were examined by comparing the average star rating, number of five-star reviews, number of helpful votes, and total number of votes between Vine and non-Vine reviews. For both categories, although the average star rating was slightly higher among Vine reviews, there were many more non-Vine reviews with five-star ratings. This suggests that even though Vine reviewers may give higher star ratings overall, they are less likely to give five-star ratings than other reviewers. Additionally, the average number of helpful votes and the total number of votes were much higher for Vine reviews suggesting that Amazon's customers find Vine reviews more helpful than other reviews.

Product Review Summary

Mobile Electronics

vine	avg_star_rating	num_five_star_reviews	avg_helpful_votes	avg_total_votes
N	3.763141	52249	1.237345	1.607125
Y	3.888889	6	41.888889	50.500000

Major Appliances

vine	avg_star_rating	num_five_star_reviews	avg_helpful_votes	avg_total_votes
N	3.714805	49592	4.315779	5.173528
Y	4.254032	112	14.274194	16.338710