Ang, Justin Clifford
Villanueva, Howard
Bruce, Nikki
Dellosa, Gio

CS 129.1- A
Big Data Project

Big Data Problem:

Given the data on the different reviews of people on amazon for video games and toys, how do we determine products received generally positive reviews and negative reviews?

This study aims to determine the ranked percentage of each Amazon product in order rate and classify them accordingly. Given this, customers would be able to easily look through and view which product has the highest positive rank and which has the lowest negative ranked percentage.

Description of Data Source:

This dataset was gathered by *Shantanu Acharya* at and is available to for download at Kaggle.com at the following link: https://www.kaggle.com/shanwizard/amazon-reviews/data. Dataset used was: Video_Games_5.json. The group only used one out of the two datasets provided as analyzing two different datasets for the same output would be repetitive. All datasets was provided in JSON format and were imported into MongoDB for the group to collect and organize.

Description of the output:

After organizing and categorizing the data based on its ranking by Amazon customers. The products with the highest average rating are the following by order from highest total score to lowest: Game Case, Playstation Network Account, Donkey Kong, Street of Rage 2, Megaman Legend. All of these games hold an average rating of 5, as all votes for these products were 5 star ratings. The products with the lowest average rating are the following: E-Shop Card, Lord of the Rings: Fellowship, Tycoon, Frisbee! Silent Hunter: Battle of the Atlantic Chris, and “Unknown.” All these products contain an average star rating of 1 star.

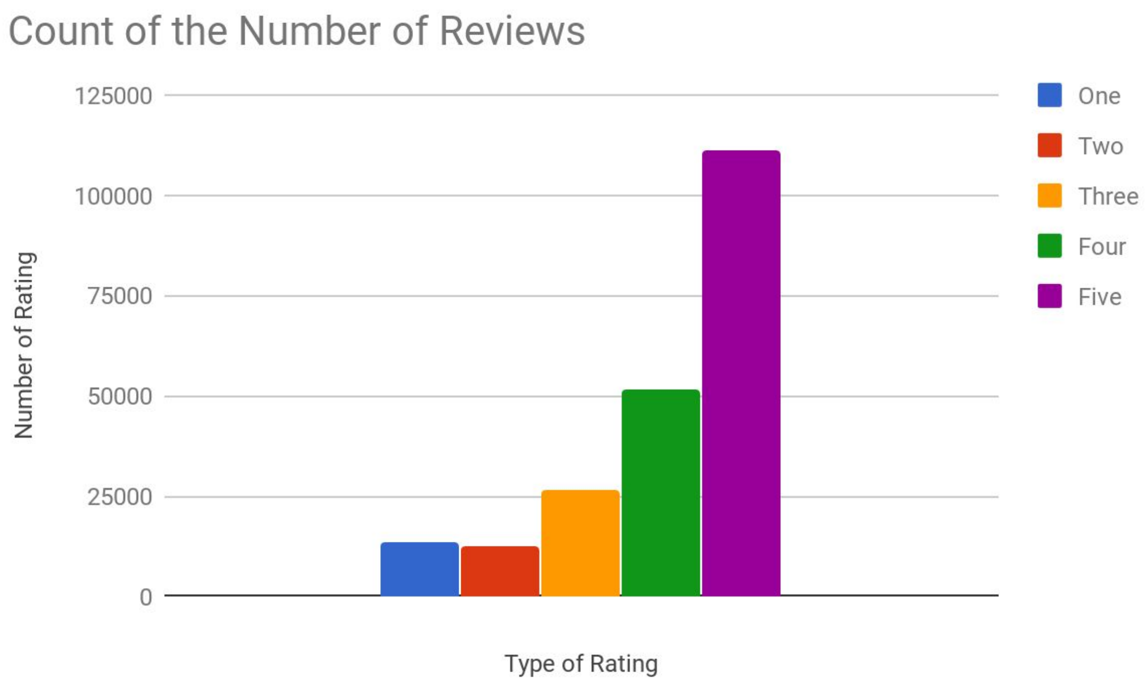
The product with the highest total score from the dataset given is the “PS4.” To determine this, the group developed the formula of “Total Score.” The formula is as follow [Total Score = Average Rating x Value Count]. Average rating is determined by averaging the total amount of star rating a product has. Value Count is determined by the number of ratings a product receives. It has an average ranking of 4.38 from its total voting count of 802 ratings with 571 votes or 71 % of its total votes being ratings of 5. On the other hard, product “Drake of the 99 Dragons” ranked an average of 1.2 from its voting count of only 5 ratings.

There were 110933 items ranked at five, 51428 items ranked at four, 26452 items ranked at three, 12727 ranked at two, and 13690 items ranked at one. Comparing the highest ranked positive output from the lowest output gives us difference of 3.18. As seen in the case of the PS4, there was a total count of 802 rating versus the ratings of the “Drake of the 99 Dragons” which only had 5 total ratings. For the PS4, there were 571 votes that were counted at 5, which is about 71% of its total rating. Ratings of 4 were at 13%, while ratings of 3 were at 4%.

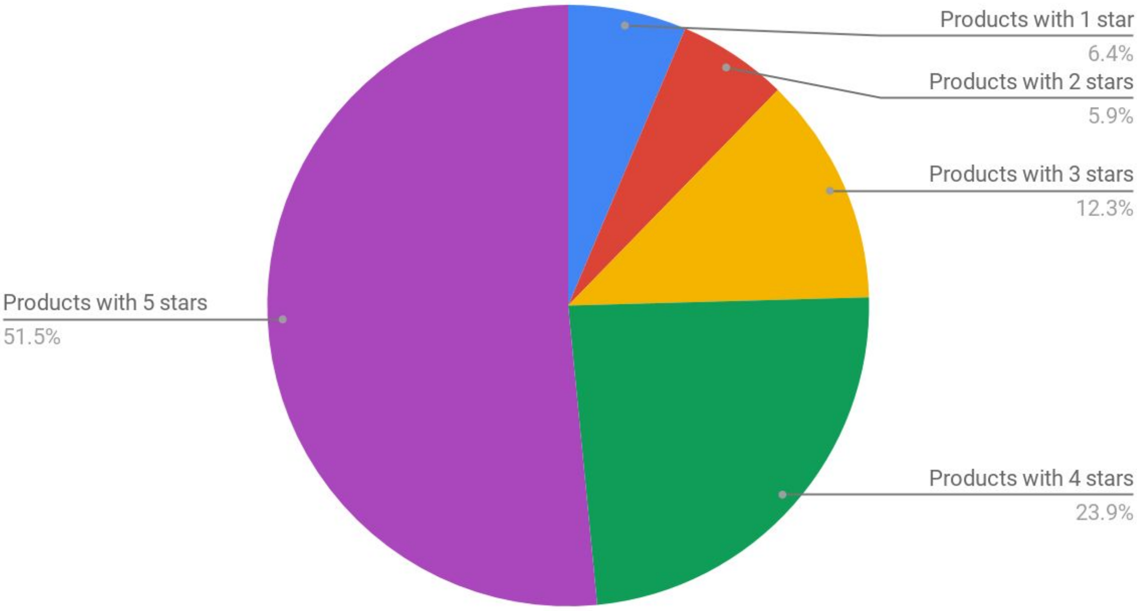
Overall, there was about 5970 products which were voted between 4 to 5 the rating, 4201 products rating between 3 to 4 the rating, 805 products voted between 2 to 3 the rating, and 90 products voted between 1 to 2 the rating.

The group can also conclude that the number of votes can become a significant factor when determining the rankings of the product. The usage of big data is required to understand the profitability of large amount of Amazon products, dating back to the early 2000s. From the research, it becomes obvious that majority of the Amazon products are performing well due to their proper screening of products. However, there are those products that do not perform so good and constant negative reviews further lead to the decrease in sales of this products. As such, with this data readily available to producers and consumers, they can easily change their tactics in choosing which products to sell and buy respectively.

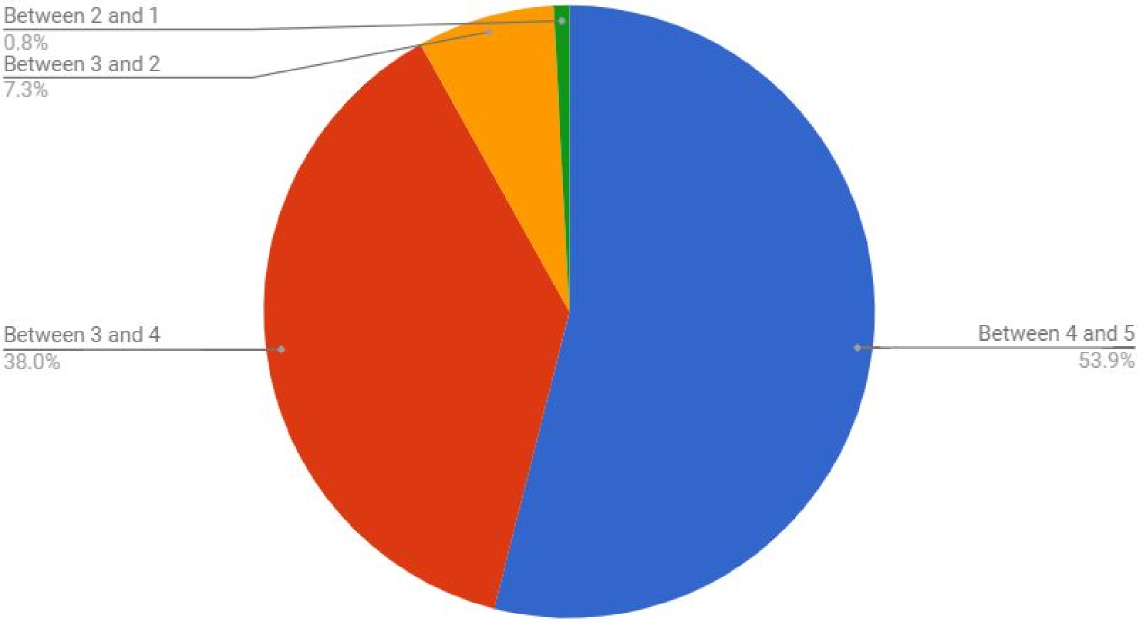
Visualization of the output:



Points scored



Rating's Proportion


Between 2 and 1
Between 3 and 2
Between 3 and 4
Between 4 and 5