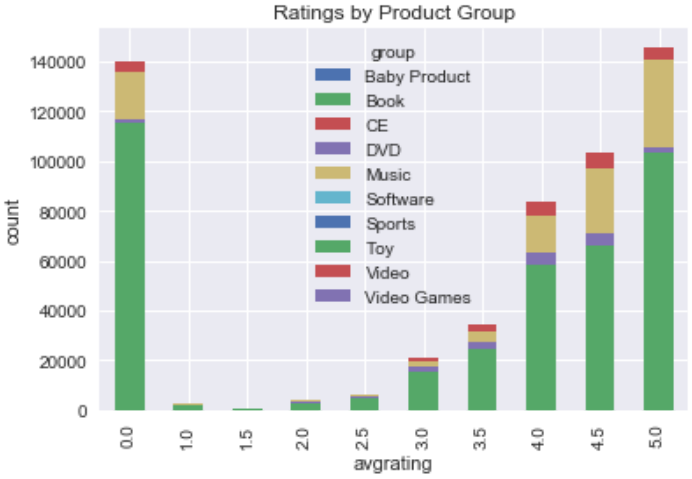
# Take Home Challenge - Amazon Product Metadata

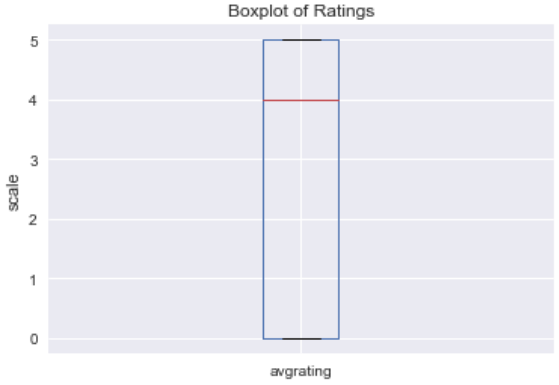
## Part A - Exploratory Data Analysis

*Trustworthiness of Ratings*

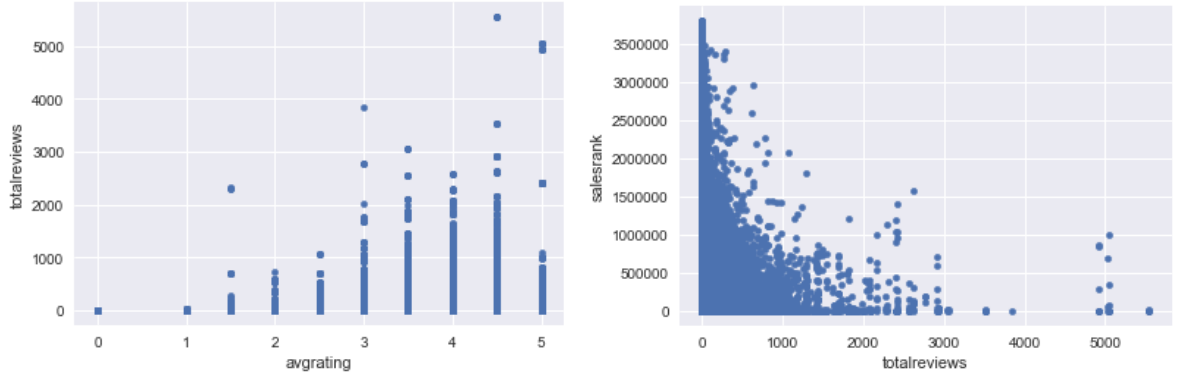
The ratings provided in this dataset are quite varied. The following bar chart shows the ratings by product group.




We see here that there is a large chunk of products that have been given a rating of zero. This is likely a sign of bias or some external factor beyond the product's control. Also it would appear that music is a product customers either love or hate. It is consistently rated above the median as per the following box plot.




There are other correlations that would be of interest for further study. For example we see that the more a product is reviewed the higher its rating.

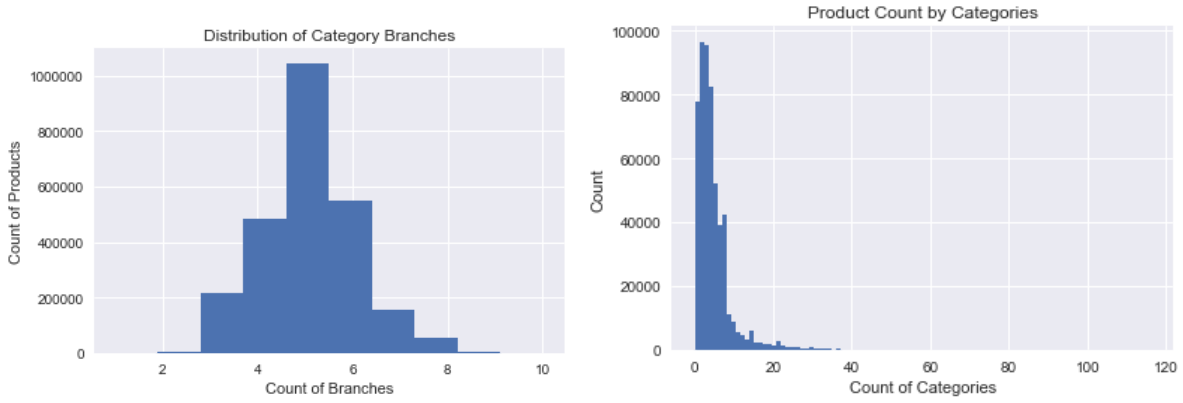



Typically a product is reviewed ~14 times however as can be seen above, some products receive thousands of reviews. It is worth understanding how sales rank is impacted by the number of reviews. While not always indicative of a high ranking, the number of reviews may give the consumer an idea of how well the product is liked.

*Category Bloat*

The categories in the dataset are quite interesting. They presented a challenge in extracting, transforming and loading into a usable format. I was able to make some approximation of their specificity by reading them into a list and counting the number of branches. In the future I would probably perform some more advanced regular expressions analysis, perhaps combined with Spark.

I found that most of the products have a category branch count of ~5 with a standard deviation of ~1.




Based on this approximation it is likely that 5 +- 1 standard deviation would capture most of the products adequately. This is just an initial estimate, but the categorization could be drastically simplified to even 10% of current categories. Note that there are 35,406 unique categories in the original dataset, this seems a massive number considering all the possible branch permutations. It is highly likely that the dataset is overspecified for category.

Take note of the chart on the right. It shows that there are many products belonging to more than a dozen categories. One product in particular lands in 116 categories but cannot be seen on the plot due to scale.

Identifying where the redundancy occurs would involve the following strategy:

- Identify duplicates and consolidate, i.e. ‘General’:[1234,4567,789], doesn’t need 3 values
- Eliminate the tails of the distribution, cut out any branches <4 and >6
- Experiment with removing the final branch/node of the categorization as this is where the largest multiplier will be, adding many extra categories that are overly specific
- Do not allow a product to belong to more than 6 different categories

## Part B - Exploratory Data Analysis

*Algorithm Thinking - Categories from Co-Purchase*

The items that were purchased with another item can serve as a basis for clustering similar types of products. I would try to cluster using the ASIN and the ‘similar’ fields. The dataset would likely be pretty large so I would probably use some form of dimensionality reduction and projection to map the clusters back to the original data. This would produce some interesting results and likely test the integrity of the ‘similar’ data. The product groups could also be incorporated to provide some general categorization ahead of the clustering.

*Product Thinking*

As a user I normally make use of the the categorization hierarchy when shopping. Being able to drill up or down on a family of products helps me compare. I think it is important that this be available as long as it does not get too specific. I would ensure that a user is provided access to this hierarchy regardless of how they landed on the product page.

Similar and co-purchased items is a great way to force the impulse buy. I would promote this to the user as much as possible in order to drive volume of sales. It can also help the user in determining similar items that might be ranked or reviewed better. I would ensure this kind of information is available on the margin of the page.

Nowadays, the ‘review’ has become all powerful. This is a very loaded question because a true analysis would have to incorporate the economic, psychological and philosophical viewpoints. A common thread amongst all three is that the review system can homogenize a marketplace. People can become reduced to the pure act of consumerism versus appreciating a product for other merits. Reviewers can also suffer from groupthink in which they emulate what others have said or done. For example, the website ‘Rotten Tomatoes’ is a phenomena of the last 10 years that has changed cinema and film forever. People do not take the time to appreciate the details and craft of filmmaking. It is reduced to a simple fact, is it good or not, according to the masses. I would test the impact of showing reviews to some users and not to others. This would be an A/B style test and serve to highlight the impact of reviews on variety, volumes and average ratings.