MCM 2020: Problem C

April 1, 2020

Contents

1 Introduction

In recent years the wealth of data available to marketers has exploded. Gone are the days of blindly releasing products into a market and hoping that they will succeed. While it's true that one can never be completely certain that success is assured, with the proper analysis companies can greatly increase their chances.

Of particular interest to many companies is the Amazon marketplace. Amazon opened its doors in 1996 and quickly grew to be the largest digital marketplace in the world [4]. In 2019 Amazon saw over 200 million unique page views per month with almost 80% of those coming from within the United States [4]. The success of Amazon's digital marketplace is almost certainly due to it's convenience and ease of use, but also to the fact that Amazon provides an easy way for customers to express their opinions about products in the form of "reviews" together with a "star rating," a score between 1-5 indicating the level of satisfaction with the product. These reviews can then be themselves rated and assigned a "helpfulness rating" by other visitors to the site. Together these factors can act to reassure consumers about their purchase or even dissuade them from purchasing in the negative case. This review system is something unique to digital marketplaces and no doubt influences sales.

These reviews and their implications for the Sunshine Company are the topic of this report. Without access to sales data we consider verified purchase reviews to be a marker of sales and extrapolate that to the larger population of sales that don't leave reviews. We analyze the distribution of the star ratings over time and question whether there is a correlation between the moving average star rating and the volume of verified purchase reviews. We examine whether Amazon's "Vine" program can be correlated with an increase in sales. Then we build a bag-of-words model to establish the similarity of reviews within a product category and answer the question of whether certain key words are associated with a particular rating. Finally we use this model to extract design features that customers deem desirable.

2 Model and Methods

2.1 Variable/Term Definitions

2.2 Assumptions

1. None of the reviews are "fake." Meaning there has been no vote manipulation and no one has been paid to review a product positively. We assume that all reviews are genuine because we have no way to actually prove that a review is not. Without this assumption we can't be sure of any conclusions we draw and without more data we are unable to find a way to verify the level of uncertainty introduced by allowing fake reviews.

2. Products with higher numbers of verified purchase reviews sell more products. This one may seem obvious, but we want to avoid the case where a product is dominating the market but only has two reviews. This allows us to extrapolate from verified purchase review volume to sales volume as a way to measure product success.

3. If a customer says they want something they actually do. We assume that if a customer says, for example, that they dislike the product because the cord is too short it's not because the product was blue. This allows us to extract desirable design features within the product categories.

Variable	**Meaning**
corpus	A collection of documents, in this case review bodies/headlines
vocabulary	The collection of unique words in the *corpus*
N	The number of documents in the *corpus*
n	The number of words in the *vocabulary*
tf	Term frequency
$f_{w,d}$	Number of time word w appears in document d
df_w	Number of documents word w appears in
idf	The inverse document frequency
$tf - idf_{w,d}$	A measure of how important word w is to document d
$tf - idf_{j,normed}$	The j'th unit length $tf - idf_{w,d}$-vector
cos_sim	The cosine similarity measure

Table 1: Variables used in the bag-of-words model.

2.3 The bag-of-words model

We needed to be able to identify two factors from the review text itself: desirable product design features and correlation between key words and star rating. To that end we propose a bag-of-words model and calculate $tf - idf$ vectors for each review to provide a way to both compare review similarity and extract words and phrases that could inform the design process. The $tf - idf$ vectors are computed in two major steps, but before any of that can be done we needed to clean the data and perform a dimensionality reduction.

We begin by answering the question of whether or not certain key words are associated strongly with star ratings. Using Python (see attached code) we group all of the reviews for a given product type, say hair dryers, based on their star rating. This leaves us with five discrete collections of reviews, each associated with a particular star rating. We consider each of these groups to be its own document. The union of these documents forms the corpus and from the corpus we can construct the vocabulary. To do this we first split each document in the union into a collection of single words and punctuation marks, this is henceforth referred to as *tokenization*. For example if our corpus consists of the single string:

$$corpus = \text{"Sally went to the store."}$$

Then we split it into the collection:

["Sally", "went", "to", "the", "store", "."]

Next, we remove punctuation and what are known as *stop words*. Stop words are words that appear frequently in a language and carry little information at the word level [2], think of words like "I", "the" and "to". In this instance we use a collection of stop words provided by NLTK [3]. We also make the text lowercase. So our example is now:

["sally", "went", "store"]

The next step is to label each word in the collection with its part of speech. To that end we again employ the NLTK using Python and get this collection:

[("sally", "RB"), ("went", "VBD"), ("store", "NN")]

Each tuple in the collection represents a word and its part of speech tag. In this instance "sally" was tagged "RB" which means the part of speech tagger identified sally as an adverb. Clearly this is incorrect, but fortunately this kind of inconsistency when it comes to handling names shouldn't affect the final outcome when we apply this concept to reviews as we'll see. "Went" and "store" are correctly identified as a past tense verb and noun respectively. The next step towards constructing our vocabulary is to perform a *lemmatization* procedure on the collection of words with their tags. To lemmatize a word means you reduce a word to it's *lemma* or it's dictionary form [5]. For instance our example becomes:

["sally", "go", "store"]

So, we removed the part of speech tag because it's no longer useful, but it's clear that "went" reduced to "go" which is it's correct present tense. By performing these steps we have reduced the size of our corpus from five element down to only three, quite a significant reduction. So we could finally construct the vocabulary.

To see this applied to a real review consider the following hair dryer review sampled from the data provided:

> "Does exactly what i need to do. Perfect! It has a nice texture and nice shape. not too heavy or too light."

Tokenizing, tagging, lemmatizing and cleaning gives:

["exactly", "need", "perfect", "nice", "texture", "nice", "shape", "heavy", "light"]

So we now have in some sense a basis for a space of reviews constructed from the words in the collection. So any other review with the same words could be represented by the elements of this basis. We apply this notion now to the true corpus of our data-sets. Each product type, hair dryer, pacifier, and microwave has its own corpus consisting of all of the reviews given on any product within that category. We apply the method detailed above to construct the vocabulary for each product type. That is we construct the set of n unique words from which we can describe all reviews for a given product type, or in other words we construct a basis from which we can define a vector space of reviews that is isomorphic to $\mathbb{R}^n$. We do this by assigning numerical values to the importance of each word in the vocabulary for a given document.

This importance is measured by what's known as a $tf - idf$ score and we assign each word in the n dimensional vocabulary such a score for a given document. That is we construct an n vector where each element is a real number indicating that words relative importance to the document. To do this we first compute the *term frequency* or tf:

$$tf = log(1 + f_{w,d}) \tag{2.1}$$

In this case we log normalize to account for potentially large numbers, and of course we add 1 since log(x) is undefined at 0 and it's not impossible that certain words might not show up in a given review i.e. $f_{wd} = 0$. Next we need to compute the *inverse document frequency* or idf:

$$idf = log(N/df_w) \tag{2.2}$$

Finally we compute *term frequency-inverse document frequency*:

$$tf - idf_{wd} = tf * idf = log(1 + f_{w,d}) * log(N/df_w) \tag{2.3}$$

Intuitively this is assigning an importance to each word in a given document by counting how many times it appears within the document, but then down-weighting it by the whether or not that word is unique to the given document.

We compute $tf - idf_{w,d}$ for each document in our corpus, in this case for the collections of reviews grouped by star rating, and construct the $tf - idf_{wd}$-vectors for each. In doing so we can identify key words that correlate with the star rating associated to that document. We can also use these vectors to measure similarity between reviews and in particular get at least some sense of how positive or negative a review is without even reading it.

To see how we can measure similarity between reviews we need to normalize the $tf - idf_{w,d}$-vector by dividing by the L2-norm of the vector:

Letting,

$$tf - idf_j = (tf - idf_{i,j}), i = 1, 2, ..., n$$

be the vector of $tf - idf$ scores for document j. Then,

$$tf - idf_{j,normed} = \frac{tf - idf_j}{||\boldsymbol{tf - idf_j}||} \tag{2.4}$$

has unit length. Since $tf - idf_j$ exists within $\mathbb{R}^n$ then the $tf - idf_{j,normed}$ vectors form a unit sphere in $\mathbb{R}^{n-1}$. This means that one possible measure of similarity, and one that is often considered in document retrieval systems, is the *cosine similarity* [1]. Although not a true metric in the mathematical sense (the Cauchy-Schwartz inequality does not hold) it is still a useful measure for our purposes. In a positive space such as this the cosine similarity defined,

$$cos\_sim(a, b) = \frac{a \cdot b}{||\boldsymbol{a}||\,||\boldsymbol{b}||} \tag{2.5}$$

returns a value between 0 and 1 inclusive. The intuition is that a value of 0 indicates the vectors a and b are parallel or very similar and a value of 1 indicates very dissimilar. So we have developed a way to determine if two documents are similar and can thus determine if a newly posted review seems to indicate that the person is happy with the product based on their similarity to the "average" 5 star vector for that product type.

Finally we develop the method for determining which features are deemed desirable by the consumers. To do this we need to establish a measure for what consumers consider desirable. The best measure available for this purpose is the number of helpful ratings a review received. So we gathered all of the reviews with more than 100 helpful votes and performed the steps described above to calculate an "average" $tf - idf$-vector for each of the three product types. We realized however that what we really needed was the document frequency. We would expect that words that appear frequently across all highly helpful documents, after some filtering, could provide insight into the kinds of features. We analyzed df_w on highly helpful reviews within the three product types to find keywords that might clue us in to what consumers want in the products. Then we randomly sampled reviews containing those words and manually parsed them to get a sense of the context the words existed within as a way to verify that our intuition was correct.

2.4 The Vine Program

We wanted to explore the effectiveness of the Amazon Vine Voices Program, therefore we looked for possible indications that the vine program was an effective way of boosting sales in all products, we decided to explore this observation using the Hair-Dryer data, due to the fact that the Microwave Data had only 19 observed vine reviews, which didn't seem conclusive or concrete enough to demonstrate rigorous statistical analysis. To explore this observation we decided to test the normality across the 11 *vine* Hair Dryers. For these 11 products we analyzed Jan 1 2014- Aug 31 2015 and counted the number of verified purchases. We then plotted the occurrences of the number of verified purchases and placed them in appropriate bins 1.The histogram appeared to be normal with a right-skew, we then performed a statistical analysis to verify our claim. We used a χ^2-test in order to test this claim. The expected values were calculated using mean and standard deviation of the entire

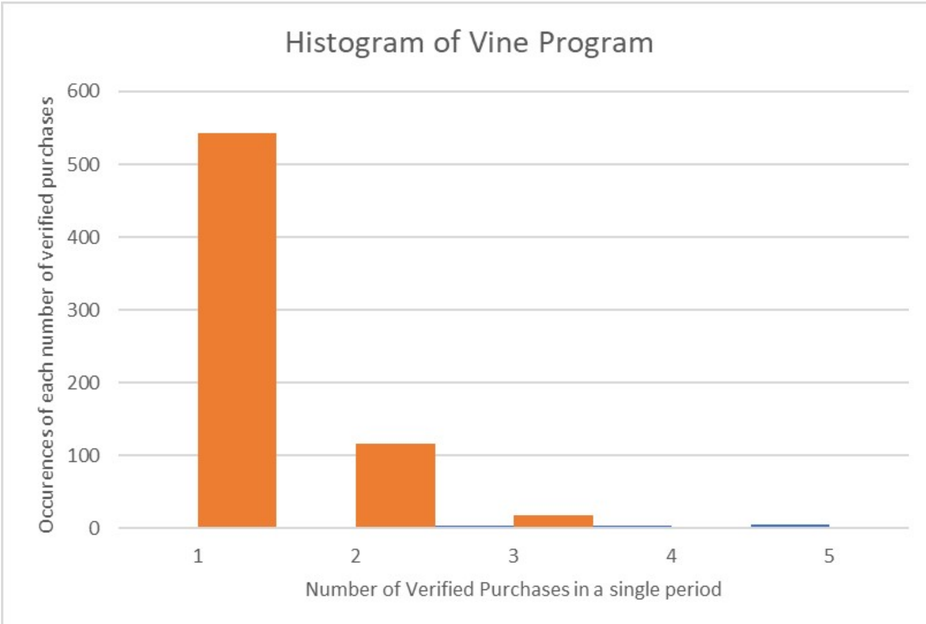

Figure 1: Histogram of Occurrences

sample.

$$H_0 : \text{Vine Product Purchases are normally distributed}$$

$$\text{H}_a : \text{There is not sufficient evidence that vine products are normally distributed} \tag{2.6}$$

Number	**Observed**	**Expected**
1	543	217.64
2	119	419.69
3	18	42.5
4	2	0.14
5	0	0

Table 2: Observed vs. Expected Values for *Vine* products

2.5 Useful Metrics

In deciding which factors a company should consider worthwhile to track, we discerned successful products to have the most and steadily increasing verified purchases in each of the 3 categories. The results are attached below.

1. Microwaves This was helpful in determining the most successful products, however we found an outlier and decided to compare the top 5 microwaves without the outlier.

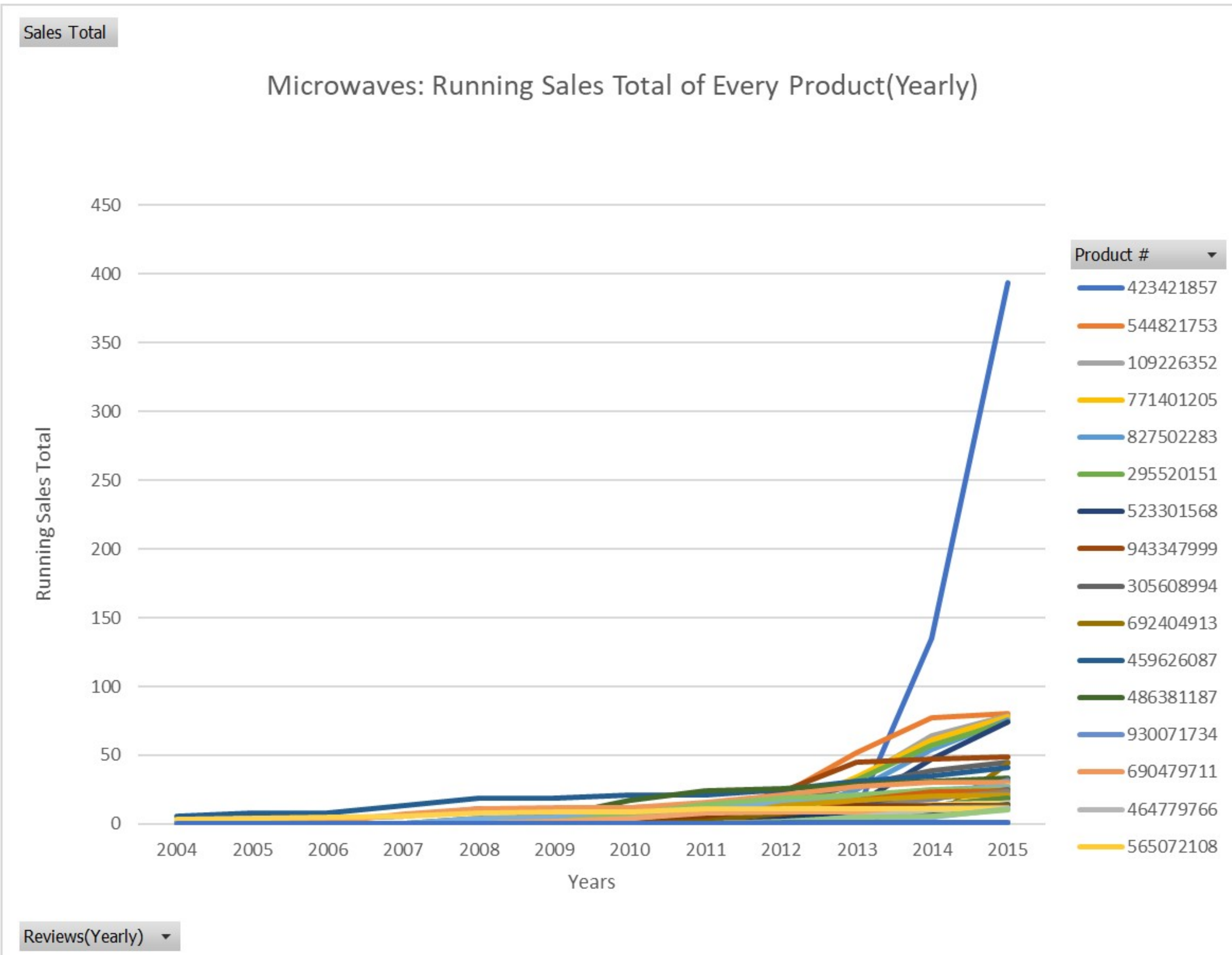

Figure 2: Microwave Running Sales Total: Every Product

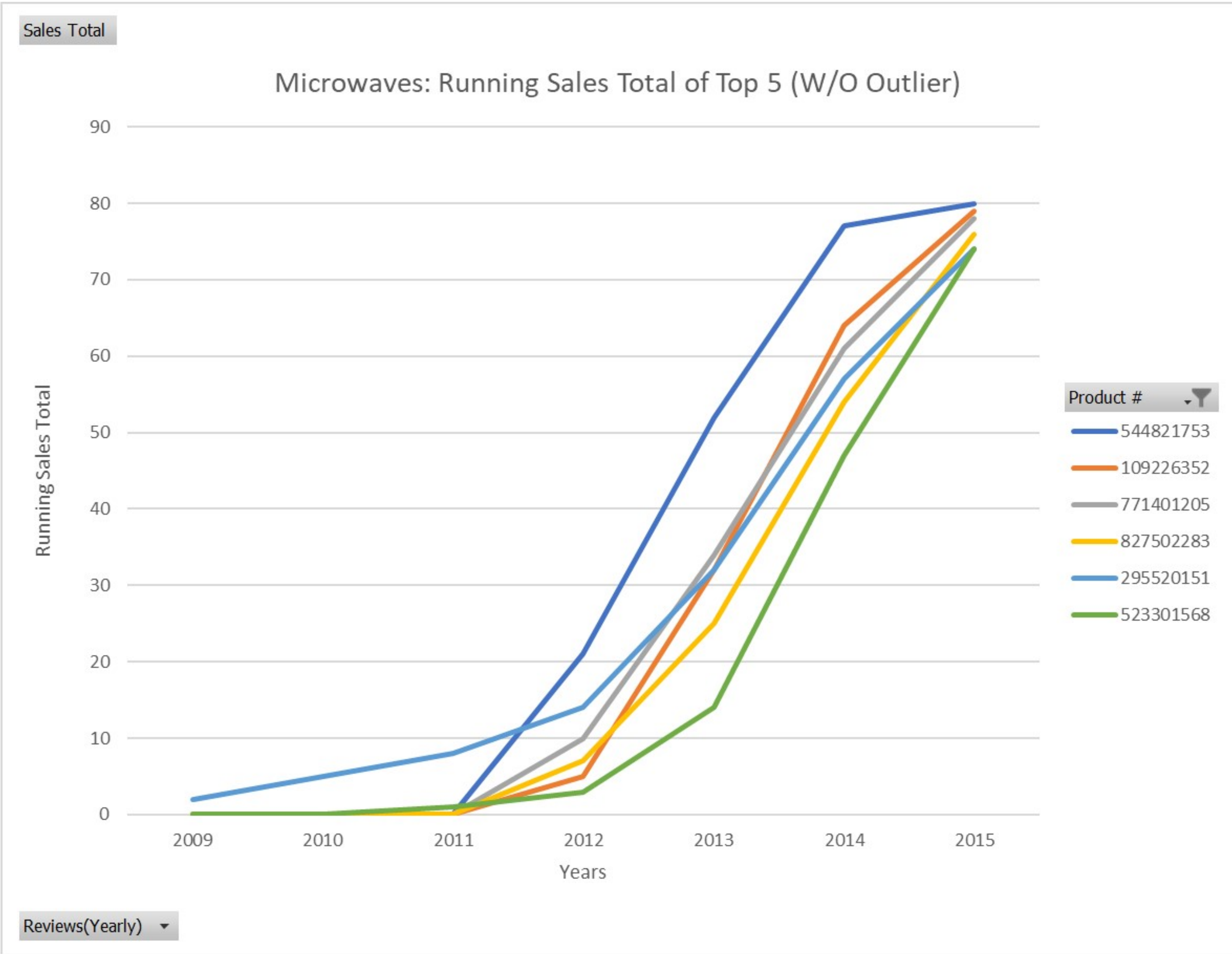

Figure 3: Running Sales Total Top 5 Microwaves

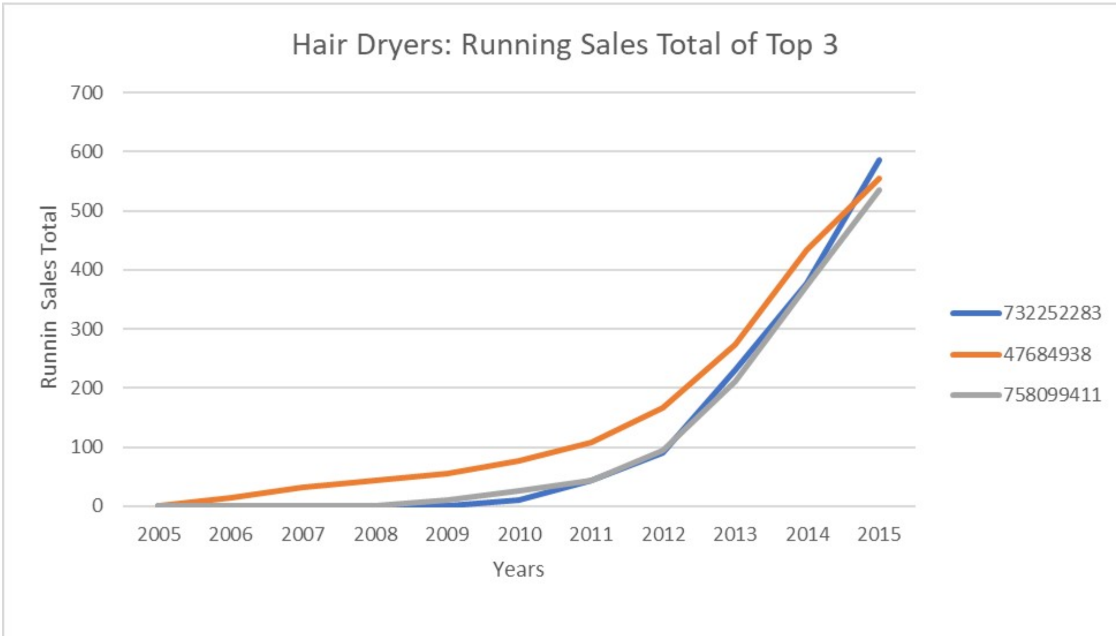

Figure 4: Running Sales Total of Top 3 Hair Dryers

2. Hair Dryers
3. Pacifiers

3 Results

3.1 Review body analysis

The method described in 2.3 works phenomenally. First, we were able to verify that certain key words directly correlate to high or low star ratings. Second, we have identified several key features that we believe consumers consider most important in each product type.

Tables 3, 4, and 5 demonstrate clearly that words with strong positive connotations correlate strongly with a 5 star rating. Using our similarity measure described one could certainly compare new reviews against the average 5 star review to get a sense of how likely it is the person is happy with the product. One expects then that the most dissimilar reviews to the 5 star average are in fact the 1 star reviews.

To get a sense of which words are prominent in the highly helpful review body corpus we considered the *df* scores for each word in each product's respective vocabulary, these can be seen in detail in the Python code [CODE REF], but the most interesting are summarized in Tables 6, 7, 8. These are just key words so cross referencing with the reviews themselves was necessary. We can say with certainty that these truly identify features we believe customers

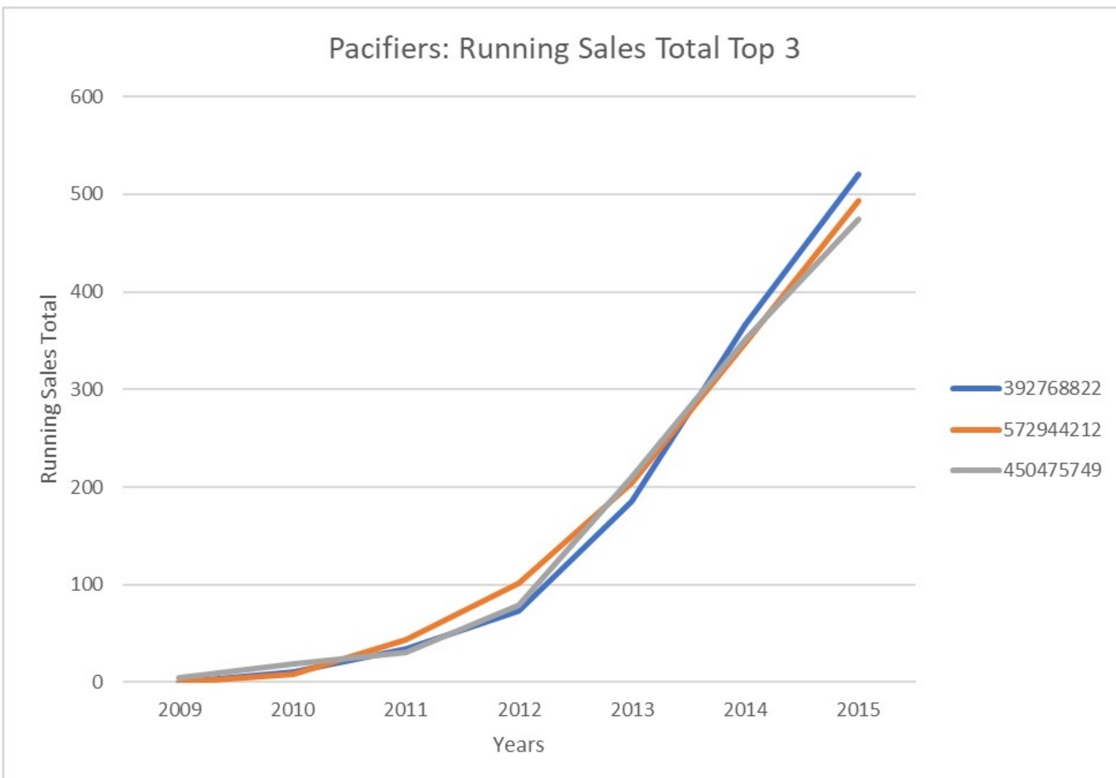

Figure 5: Running Sales Total Top 3 Pacifiers

Word	**tf-idf score**
great	0.242754
love	0.234865
good	0.187971

Table 3: Top 3 words by tf-idf score for the average hair dryer 5 star review

Word	**tf-idf score**
great	0.256225
work	0.188830
good	0.186215

Table 4: Top 3 words by tf-idf score for the average microwave 5 star review

Word	**tf-idf score**
love	0.280501
great	0.215560
baby	0.187513

Table 5: Top 3 words by tf-idf score for the average pacifier 5 star review

Word	**df score**
cord	365
quickly	201
heavy	193

Table 6: Selected hair dryer feature words

Word	**df score**
price	86
small	74
looks	71

Table 7: Selected microwave feature words

Word	**df score**
mouth	192
easy	143
little	139

Table 8: Selected pacifier feature words

most value in the respective products. For hair dryers customers want to see a long cord, ideally retractable, it needs to be powerful and dry your hair quickly and it can't be heavy. For microwaves people look for a low price point that fits on their counter and looks sleek. Finally, for pacifiers we determined that consumers most value a pacifier that is small enough that it fits in a baby's mouth and can be easily grasped by their small hands.

3.2 The Vine Program

We used a χ^2-test in order to test whether our claim was correct, using a confidence level of 95% we resulted in an actual value of $1.32 * 10^{-159}$). This result means that our data is statistically significant and therefore we reject the null hypothesis claiming that, **There is not sufficient evidence to claim that the purchases of vine products is normal**. This could suggest that the vine program may not be the most valuable factor to consider when entering a new market place.

3.3 Useful Metrics

We knew that it would be useful to determine a metric that a company could follow in order to quickly discern the success or failure of a product. The most useful metric we found was to track the running sales total for the top products in each category. We then suggest the company to look at these successful products and analyze their reviews in order to discern successful features. The trend we noticed in every category was that there was a steady increase of verified products in the years ranging from 2012-2014. Further analysis needs to be performed in order to understand why this is exactly the case, however it's something that we find would be useful for a company to know and understand.

3.4 Strengths and weaknesses

3.4.1 Normality of Vine Program

Even using the hair dryer data, we discovered that we didn't have the ideal amount of data, we conclude effectively, however to be more rigorous, more data would be needed.If we had more time, we would have analyzed the vine program with more metrics other than verified purchases, such as star ratings, or helpful and total votes on specific reviews.

3.4.2 Useful Metrics

We would have liked to find more metrics a company would find useful to track, for example we had a claim that tracking star ratings before and after vine reviews would directly influence consumer behavior. However, we did not have time to test this directly. In addition, we would have tried to understand how each piece of data is important and tells the entire story of the success of a specific product, not just the most successful, rather the products that seem to be successful in the short term vs. long term. We noticed a few trends that we specifically wanted to explore more, the trends we noticed were a major-die off of products in 2007-2009 year range and major increase in popularity of specific in 2012-2014 range. We think that these trends were a direct result of the American economy in those times, however we did not have data supporting this claim. We would have liked to explore this specific claim further. Due to the time-constraints we were not able to completely understand how all of the metrics factored into the usefulness of a specified product.

3.5 Future Work

We initially tried to establish a correlation between star rating and number of verified purchases in order to conclude that star rating could be a significant influence in a customer's decision making process, however the data we used was not normal, meaning that linear regression was not an effective way of testing the correlation. If we had more time we would explore this concept more and use other metrics other than number of verified purchases as well as for all categories of data given.

We would also recommend future work be done on the topic of text classification and sentiment analysis. The bag-of-words model presented works well in some cases, but it loses a lot of information that comes with context. More sophisticated models could be employed in the future to improve performance. Consider perhaps any of the various machine learning algorithms such as doc2vec that exist for this exact purpose.

4 Letter/Memo

Dear whom it may concern,

We understand the transition to an online market place is not an easy one especially in the Microwave, Hair Dryer and Pacifier markets, we hope to make this transition as seamless as possible. Here at Team 2018233, we identified some useful metrics to consider as well as some metrics that aren't necessary when accessing the success of your products. We used data provided by Amazon and performed rigorous statistical analysis in order to present with you the most useful and accessible information.

The first thing to consider would be how the verified purchases change. This means that the most useful metric we found to track the success of a product is the total number of verified purchases as well as how quickly that number increases or decreases, large periods of increasing verified purchases could indicate that a product is performing well in the desired marketplace, conversely a failing product could exhibit long periods of either constant verified purchases or decreasing verified purchases. While this is a useful metric in discerning the success of a product verified purchases doesn't always give the entire story of the products performance.

Additionally, we built a tool to quickly analyze text in reviews. Especially in popular online market places like Amazon, there are 1,000s of reviews every day, while this can be cumbersome to track for your products, our text analyzing tool will help. Our text analyzing tool uses rigorous comparison in order to discern the the most positive/negative feedback as well as important descriptors that can effectively indicate the healthiness of a specific product. With our text analyzing tool, you could have the ability to discern whether a specific review is satisfied or dissatisfied with the specific product, this should in turn help your decision making department determine the best plan of action in order to save a failing product or promote a growing product. With our tools and analysis we hope you will be ahead of the decision making curve of the entire market.

Lastly, we know the Amazon Vine Program may seem like an attractive option, especially for traditionally smaller/ mid-size companies, however our statistical analysis concluded that the Amazon Vine program is not guaranteed to increase sales. Because the Amazon Vine program costs a significant amount of money, we would advise to save one's money and focus on other metrics such as customer's satisfaction with the specific product or the changing of the number of verified purchases.

While we can't give guarantee that specific products will succeed or fail in a market place, we are able to give you the tools in order to discern whether a product is performing well or poorly. We wish the best of luck in this competitive market place and we hope with our advice and tools one can be ahead of the decision making-curve and take the most appropriate plan of action, before the market follows.

Yours truly, Team 2018233

References

[1] *Dot products*, 2008 (accessed March 6, 2020). https://nlp.stanford.edu/IR-book/html/htmledition/dot-products-1.html.

[2] *Dropping common terms: stop words*, 2008 (accessed March 6, 2020). https://nlp.stanford.edu/IR-book/html/htmledition/dropping-common-terms-stop-words-1.html.

[3] *Natural language toolkit*, 2020 (accessed March 6, 2020). https://www.nltk.org/.

[4] *Amazon.com february 2020 overview*, 2020 (accessed March 9, 2020). https://www.similarweb.com/website/amazon.com.

[5] *Lemma*, 2020 (accessed March 9, 2020). https://www.merriam-webster.com/dictionary/lemma.

A Code Appendix

Attached pdf version of Jupyter notebooks contain all code.