An Exploratory Analysis of Customer Review Data

Malik Jackson and John Shueh
CMSC 455 - Numerical Computations
University of Maryland, Baltimore County (UMBC)
{shuj1, mjacks3}@umbc.edu

Abstract
In this paper, we present an analysis of Amazon customer review data using natural language processing and curve fitting techniques. Our approach does not dive too deeply into any particular aspect of the data but rather seeks to provide a broad overview. The analysis is exploratory in nature and seeks to uncover more interesting questions that could be answered through further examination of the data.

Keywords
Customer review data, curve fitting, sentiment analysis, lexical diversity, natural language processing

1. Introduction

In the current information age, e-commerce has grown to become a business comparable to large brick and mortar stores. Understanding consumer purchasing and reviewing trends and how to market to them plays a major role in driving increased sales and overall revenue for a business. Generally, companies, markets and businesses provide various avenues for consumers to provide feedback including : ratings, forums, and commenting systems. Ideally, the responses from these feedback sources should be a primary base for companies'/sellers' marketing analytics.

To gain insight into the data analytics process, this project team will analyze customer review data from Amazon.com and explore relationships between various review factors and product rating trends. We will attempt to answer the following questions: What kinds of reviews give which kind of product ratings? How do reviewers determine what and given sufficient information on the other two questions, why do reviewers rate the way they rate? To answer these questions and analyze relationships in the data we use, our team will primarily utilize curve fitting, some data visualization techniques and lexical analysis.

2. Background

Our team's main motivation behind this analysis was having the opportunity to do an exploratory deep dive into actual customer data and seeing what could be found. While we did not have much background knowledge regarding this field, we did have some applicable numerical methods and natural language processing (NLP) techniques to try out on some datasets.

In brainstorming this project, we chose Amazon.com as it is *the* contemporary premier site in general online shopping and ecommerce. With a variety of categories and even more subcategories, a public forum for any consumer or merchant to buy or sell products and millions of product ratings and reviews, we were confident we would be able to find sufficient data and databases to use.

For our data source, we utilized a collection of reviews provided by Julian McAuley of UCSD used in his paper [1] . McAuley's dataset provided a collection over 10 million reviews broken down into same categories Amazon uses on its site. We sampled 300,000 reviews from 3 distinct categories, beauty, video games and clothing, to create a sufficiently large and varied enough sample size such that any potential correlations we drew from our analysis could be valid enough for further analysis.The collection of reviews are organized as lists of dictionaries, the structure of which is explained later in this report. Additionally, McAuley's collection provides a collection of random product metadata, similarly organized as the reviews, into dictionaries.

Since we could not verify that the product of each review could be matched to a product in the metadata collection, and to limit the scope of our analysis, we decided that we would not be able to make use of McAuley's metadata collection and ultimately decided to focus on the accessible lexical and quantitative portions of the review data.

Using the information available from this set, we chose five factors to analyze as they relate to a product's cumulative score: helpfulness rating, time of review, keyword/frequency tracking, lexical diversity and sentiment analysis.

3. Materials and Methods Overview

3.1. Materials

To make use of the data that we gathered from [2], we made use of the Python programming language and a variety of open-source libraries for manipulating and visualizing that data.

3.1.1. Environment and Libraries

Python was used to read the data into a usable format for further manipulation and visualization. Various libraries also aided us in this endeavor. Examples include: Pandas/Numpy for useful data structures such as data frames alongside optimized operations, Natural Language Toolkit (NLTK) for applying sentiment analysis to the large amounts of text reviews, Scipy for basic curve fitting, and Matlibplot for plotting the data into various charts. All of our code was run on Python 2, using Jupyter notebook and the notebook used to generate the results is linked at the end of this paper.

3.1.2. Data Format

The data that we utilized came from a larger subset hosted at [2]. The dataset contains customer reviews, which contain properties such as item ASIN (alphanumeric label), review text, overall rating, unix review time, etc. An example can be seen in Figure 1 below.

```
{
  "reviewerID": "A2SUAM1J3GNN3B",
  "asin": "0000013714",
  "reviewerName": "J. McDonald",
  "helpful": [2, 3],
  "reviewText": "I bought this for my husband who plays the
piano.  He is having a wonderful time playing these old hymns.
The music  is at times hard to read because we think the book
was published for singing from more than playing from.  Great
purchase though!",
  "overall": 5.0,
  "summary": "Heavenly Highway Hymns",
  "unixReviewTime": 1252800000,
  "reviewTime": "09 13, 2009"
}
```


Figure 1. An example review in the dataset obtained from [2]

3.2. Methods

3.2.1. Curve Fitting

The method of least squares was applied to our data in order to fit a curve. Least squares aims to minimize the sum of the squares of the residuals, or difference between actual dependent variable value and predicted value.

A linear fit was not enough to sufficiently fit our data, so we ended up applying more complex polynomial fits on our data, using values of n from n = 1 to n = 10. For each model, we attempted to significantly minimize the residual error (mean residuals squared) . We defined significantly minimized error as greater than a 1-10% percent difference from n-1's to n's residual error, with particular attention to the trajectory of error values in the entire data set. Adhering to machine learning properties, we were also careful not to select values of n for which the residual errors were too insignificant (generally less than 1-5 percent of n-1's residual error). This was done to keep the chosen regression model general enough to apply to larger or smaller sample sizes and thus, overfitting our particular data set.

3.2.2. Keyword Frequency

Our keyword/frequency tracking simply involved counting the usage of words from a set of keywords and analyzing how they affected the cumulative product score. In our experimentation with this topic, we initially intended to calculate the "trending" words- the most frequently used words in a corpus- and then plot their frequency in reviews to cumulative product rating. We

later decided since trending words will differ from sample to sample, our results from any one sample may not generalize well. Instead of trending keywords then, we decided to use a set of topical words which we manually supplied that generalized to other samples well. Those words are 'cost','low','high','cheap','expensive','worth','value','money','deal'. In any category, cost should be a concerning factor to some significant subset of reviewers. Using this factor also gave us an additional reason to circumvent needing to link the metadata set, which contained price information for products to the reviews.

3.2.3. Lexical Diversity

Lexical diversity is defined as the number of unique words used in a review over the total number of words used in a review. Initially, we hypothesized the trend for lexical diversity would be inverse to the trend for helpfulness. Our reasoning for this was that reviews with short, succinct responses, might not repeat too many words, but won't provide enough insight to be helpful. Meanwhile, more critical or lengthy responses would naturally repeat more words, like articles and conjunctions. We decided to use this factor to test our hypothesis and because a product rating correlation to lexical diversity does provide us with information about that review.

3.2.4. Sentiment Analysis

Another linguistic technique that we applied to our data was sentiment analysis. Sentiment analysis places a quantitative measure of opinion on a given piece of text in terms of positivity or negativity.the other. Together, this and our other linguistic analyses gave us valuable tools to represent and mathematically explore the text data.

4. Results

For each factor, we followed a set of guidelines for implementation and analysis. First, we computed the numerical values necessary to mathematically represent each factor. Then, we sorted our dataset in ascending order by the value of our numerically represented factor and computed the cumulative product score. Before computing regression statistics, we created a dimensionality reduction plot to give us a relative idea of the distribution of our review scores and values for our observed factor. We then computed polynomial regression at varying degrees for each data set, calculating residual error (residual sum of squares), coefficient values, average error, variance, and standard deviation.

4.1. Helpfulness Rating

The Dimensionality Reduction plot for Review Score Vs. Helpfulness Rating is shown below in Figure 2.

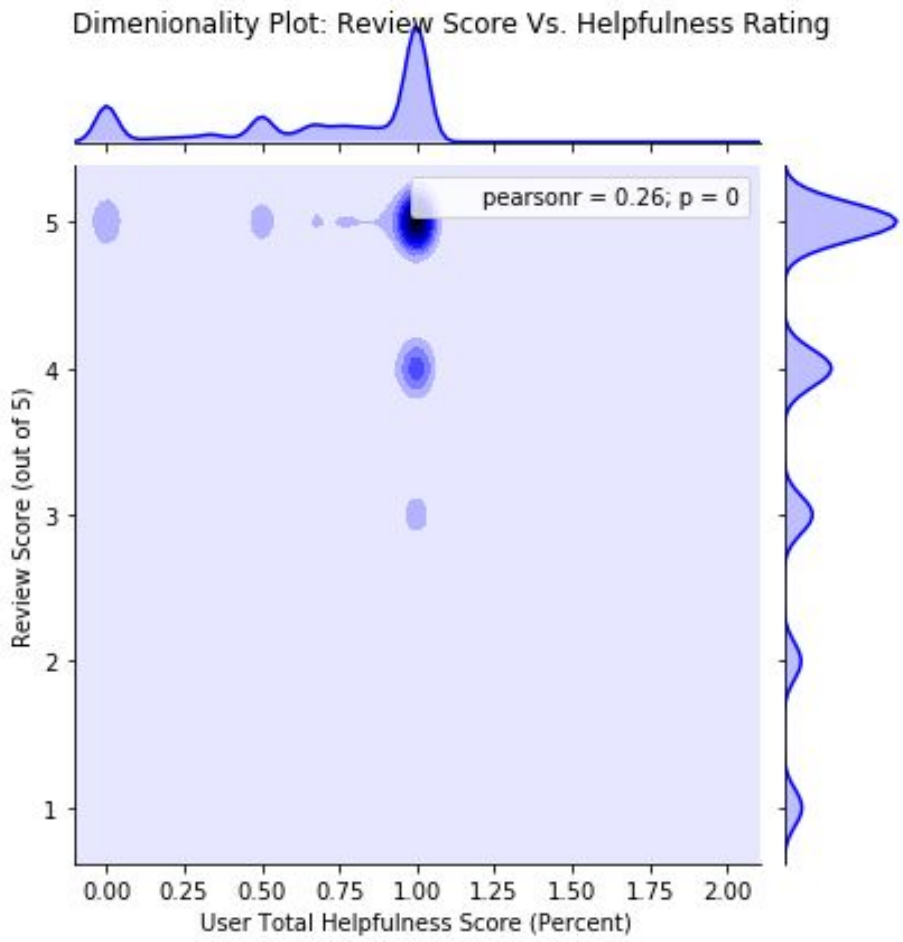

Figure 2. Dimensionality Reduction plot for Review Score vs. Helpfulness Rating

We can see through this plot that our data set contains a significant number of 100 percent helpful rated reviews with most in turn giving 5 and 4 star product reviews. We also have a significant number of 0 percent helpfulness ratings with 5 star product reviews. This will aid in our analysis and discussion section later.

The table of regression statistics for Review Score vs. Helpfulness Rating is shown below.

Polynomial	Residual Error	Average Error	Total Variance	Standard Deviation
n = 1	2620.942557	0.873647519	1.62E-06	0.001271445
n = 2	486.549315	0.162183105	5.71E-06	0.002388742
n = 3	483.2083366	0.161069446	0.000215759	0.014688734
n = 4	481.4636952	0.160487898	0.002009874	0.044831622
n = 5	458.2777749	0.152759258	0.03272643	0.180904478
n = 6	412.8447376	0.137614913	1.241240346	1.114109665
n = 7	408.3573634	0.136119121	40.88729073	6.394317065
n = 8	405.1823773	0.135060792	1771.810755	42.09288248
n = 9	400.0811283	0.133360376	52714.66953	229.5967542
n = 10	395.7793558	0.131926452	1952331.211	1397.258463

We see a significant decrease in residual error per polynomial until about n= 6, so for this set we choose n = 6 for our ideal regression model. Next, we graphed n = 6 (and n = 1 and n = 10 to benchmark our results).

The Graph of Cum Product Rating Vs. Helpfulness Score is shown below in Figure 3.

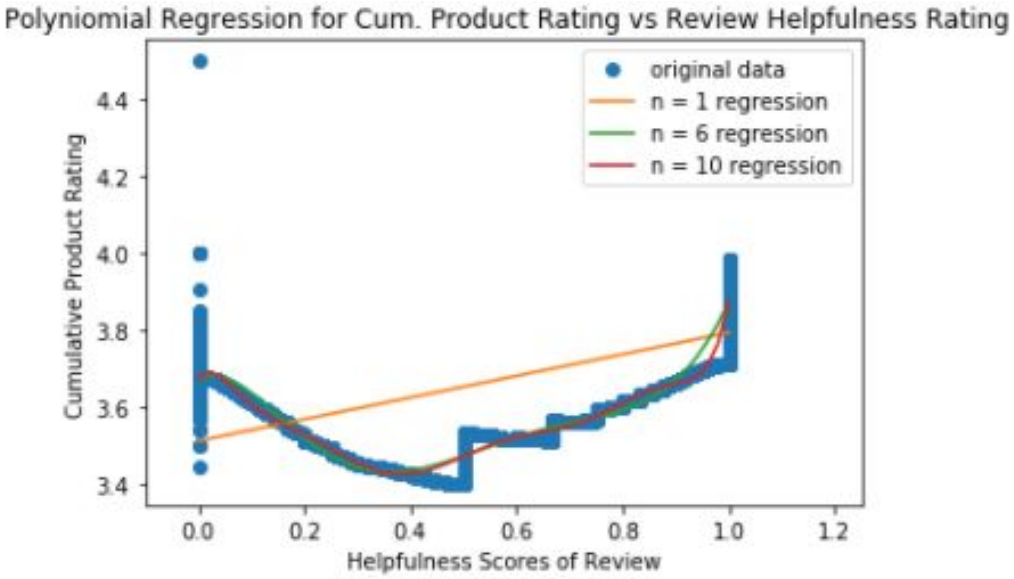

Figure 3. Plot for Cumulative Product Rating vs. Helpfulness Rating

We see a pretty comparable model between n = 6 and n = 10, illustrating the relative insignificant improvement in the model from n = 6 to n = 10, which correlates well with what we see in our data for the residual error between the two. N=6 *does* exponentially improve the residual error from our lower baseline, n = 1. We will continue to see this correlation between our lower and higher benchmarks in the statistics and graphs for *all* of our factors, thus we will omit most of our further analysis of results until the discussion and analysis section.

4.2. Time

The Dimensionality Reduction Plot for Review Score Vs Time is shown below in Figure 4.

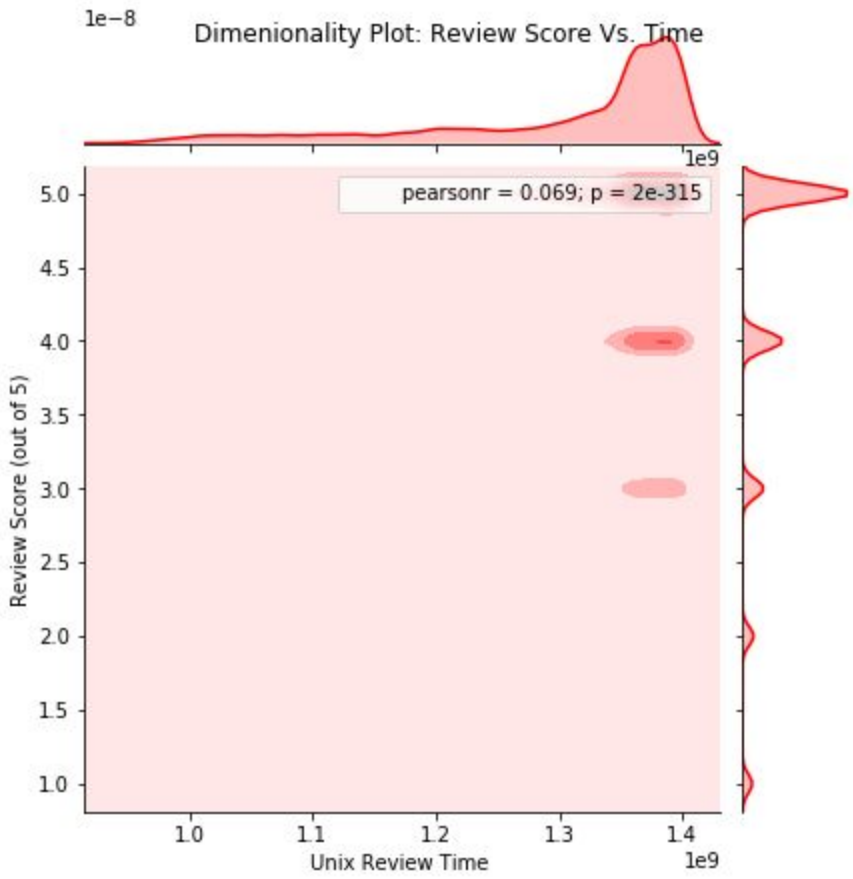

Figure 4. Dimensionality Reduction Plot for Review Score Vs. Time

The x-axis is labelled in unix timestamps, the earliest date in our sample is October 14, 1999. The most recent date in our sample is July 23, 2014. We see a significant portion of 5, 4, and 3 star ratings attributed to reviews made later in the timeframe.

The table of regression statistics for Review Score vs. Time is shown below:

Polynomial	Residual Error	Average Error	Total Variance	Standard Deviation
n = 1	461.6997	0.153899891	7.06E-07	0.000840303
n = 2	90.9553	0.030318435	1.61E-05	0.004010969
n = 3	27.34839	0.009116129	0.0005232	0.02287447
n = 4	22.89399	0.007631331	0.0428302	0.20695464
n = 5	20.87677	0.006958925	4.0374626	2.009343821
n = 6	20.1661	0.006722033	63.093731	7.943156258
n = 7	11.06316	0.00368772	85.437158	9.243222293
n = 8**				

** numpy's polyfit function failed to provide coefficients for n values higher than 7.

We choose n = 4. The regression graph is shown below in Figure 5.

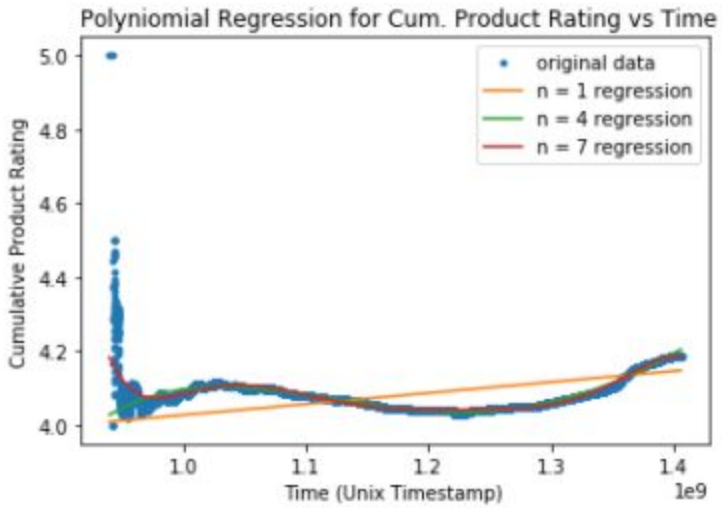

Figure 5. Plot of Cumulative Product Rating vs. Time

4.3. Keyword Frequency

The Dimensionality Reduction plot for Review Score Vs Keyword Frequency is shown below in Figure 6.

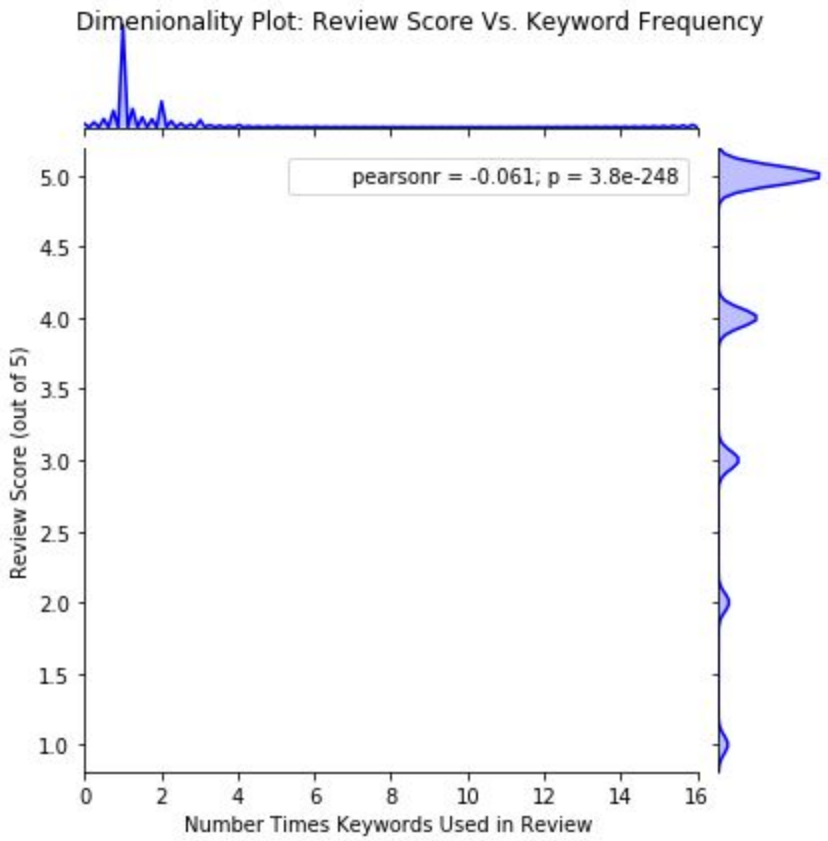

Figure 6. Dimensionality Reduction Plot for Review Score vs. Keyword Frequency

Type issues in calculations for keyword frequency prevented the center portion of the graph from displaying, but we can discern the relative intensity without it. We see a majority of reviews for every star rating on the lower end of the frequency axis. As we'll discuss reason for in detail later, this is indicative of a insignificant change on the y axis throughout most of the x-axis. One potiential reason that' we'll mention now is the presence of outlier data that influences the model.

The table of regression statistics for Review Score vs. Keyword Frequency is shown below.

Polynomial	Residual Error	Average Error	Total Variance	Standard Deviation
n = 1	605.97033	0.2019901	2.54E-08	0.0001593
n = 2	594.21584	0.1980719	5.96E-08	0.0002442
n = 3	590.91151	0.1969705	1.34E-07	0.0003667
n = 4	590.02377	0.1966746	3.45E-07	0.0005877
n = 5	589.82866	0.1966096	1.02E-06	0.0010114
n = 6	589.79526	0.1965984	3.94E-06	0.0019845
n = 7	589.7904	0.1965968	1.79E-05	0.0042309
n = 8	589.7897	0.1965966	0.000105	0.0102459
n = 9	589.78957	0.1965965	0.0006633	0.0257549
n = 10	589.78955	0.1965965	0.0059509	0.0771419

We choose n = 3. The graph of regression statistics for Review Score vs. Keyword Frequency is shown below in Figure 7:

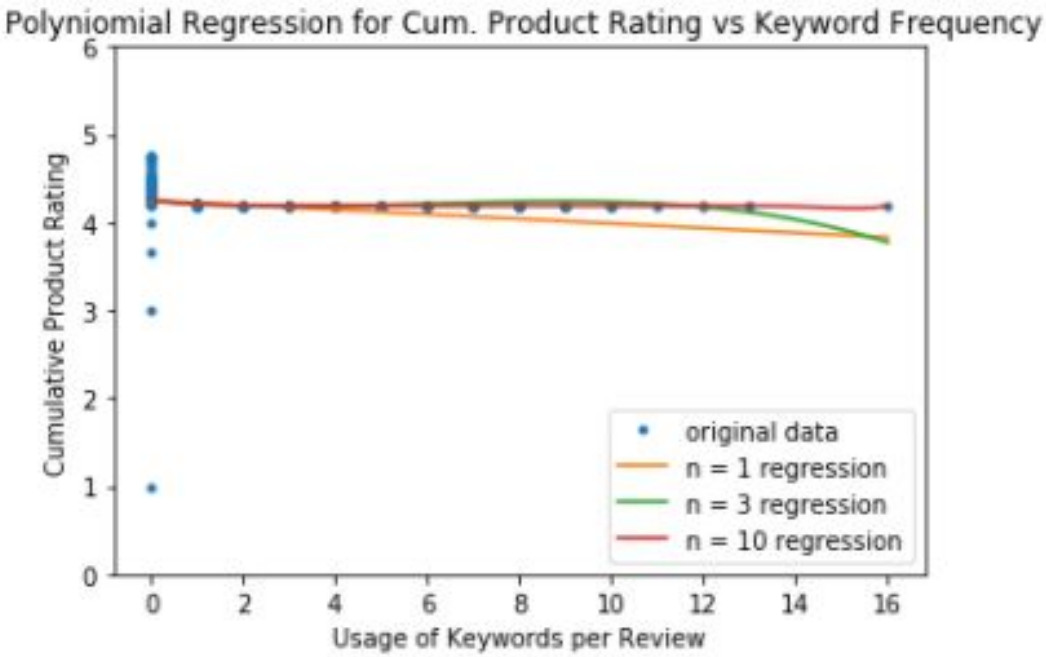

Figure 7. Plot for Cumulative Product Rating vs. Keyword Frequency

4.4. Lexical Diversity

The Dimensionality Reduction plot for Review Score Vs Lexical Diversity is shown below in Figure 8.

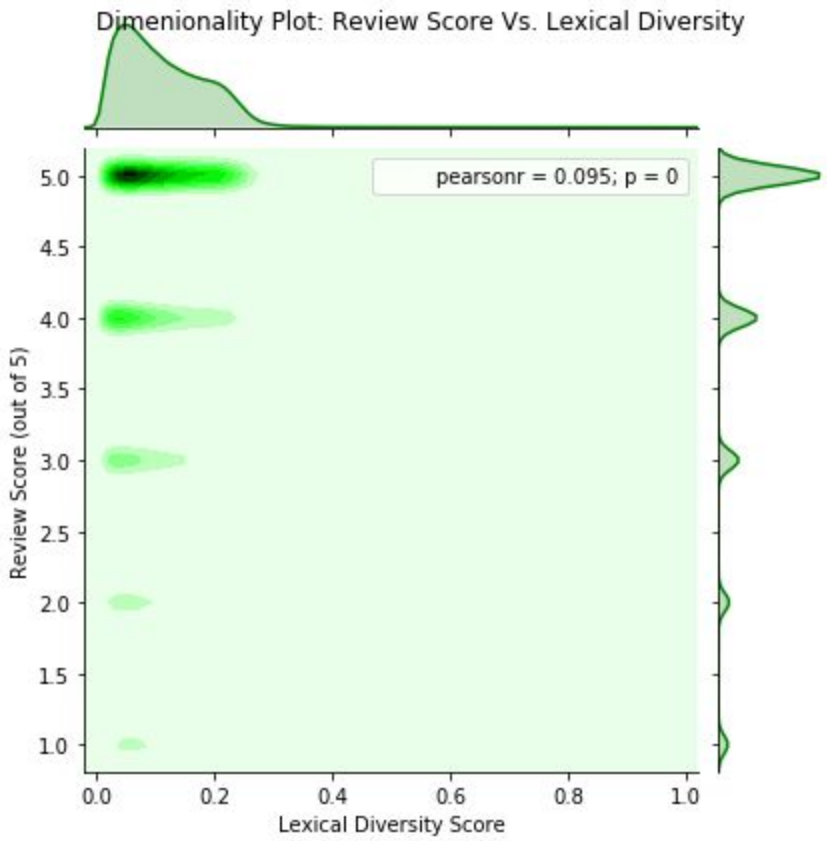

Figure 8. Dimensionality Reduction Plot for Review Score vs. Lexical Diversity

We similarly see a majority of the data clustered around the earlier values of the x axis.Our table of regression statistics for Review Score vs. Lexical Diversity is shown below:

Polynomial	Residual Error	Average Error	Total Variance	Standard Deviation
n = 1	95.97852335	0.031992841	1.80E-07	0.00042377
n = 2	28.27002838	0.009423343	9.72E-07	0.000986111
n = 3	28.26978034	0.00942326	3.21E-05	0.0056671
n = 4	20.26173057	0.00675391	0.000498961	0.022337444
n = 5	15.01734501	0.005005782	0.012358124	0.111167098
n = 6	14.48680375	0.004828935	0.3926744	0.626637375
n = 7	14.24529004	0.00474843	12.29317883	3.506162978
n = 8	13.66763257	0.004555878	441.6159157	21.01465954
n = 9	13.54324468	0.004514415	15735.49004	125.4411816
n = 10	13.54313825	0.004514379	568916.5635	754.2655789

Overall we see a low amount of residual error reduction after n = 2. Instead of choosing n =2 , however, we'll look at our data holistically and choose n = 5, as there is still a significant reduction from n=4 to n = 5. We see this information represented similarly for our regression plot shown below in Figure 9.

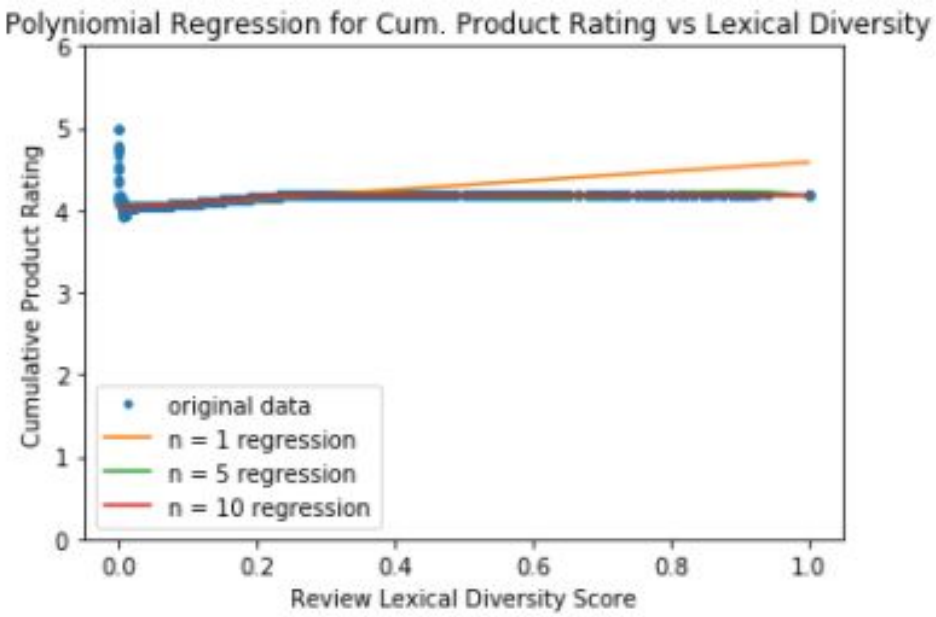

Figure 9. Plot for Cumulative Product Rating vs. Lexical Diversity

4.5. Sentiment

The Dimensionality Reduction plot for Review Score Vs Sentiment is shown below in Figure 10.

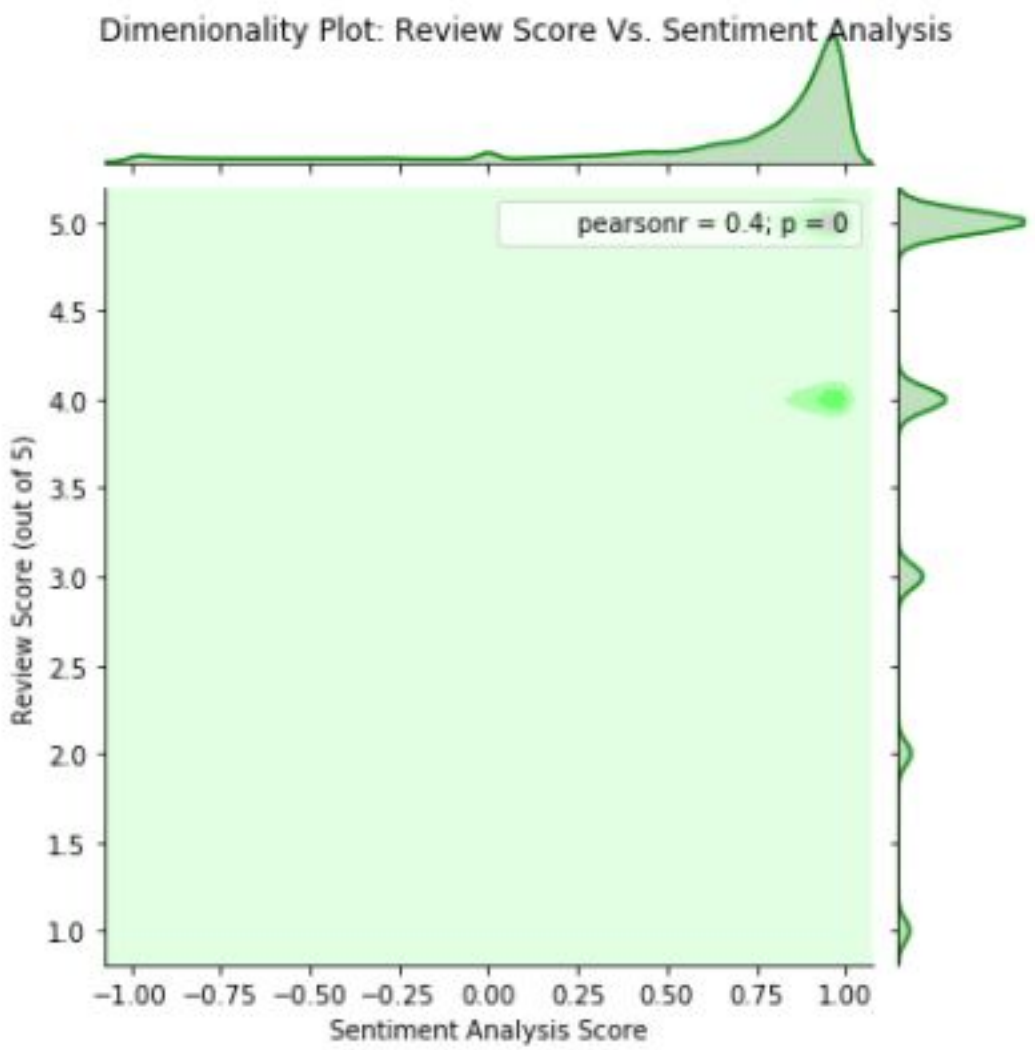

Figure 10. Dimensionality Reduction Plot for Review Score vs. Sentiment Analysis Scores

The table of regression statistics for Review Score vs. Sentiment Score is shown below.

Polynomial	Residual Error	Average Error	Total Variance	Standard Deviation
n = 1	7321.2972	2.4404324	5.38E-07	0.0007333
n = 2	573.23154	0.1910772	1.72E-07	0.0004142
n = 3	251.65223	0.0838841	3.49E-07	0.0005904
n = 4	33.019451	0.0110065	2.30E-07	0.0004791
n = 5	32.997766	0.0109993	1.20E-06	0.0010947
n = 6	32.878583	0.0109595	6.31E-06	0.0025118
n = 7	32.789476	0.0109298	3.41E-05	0.0058389
n = 8	30.682699	0.0102276	0.0001731	0.0131575
n = 9	30.648762	0.0102163	0.0009678	0.0311103
n = 10	23.474697	0.0078249	0.0040117	0.063338

We choose n = 5. The regression plot is shown below in Figure 11.

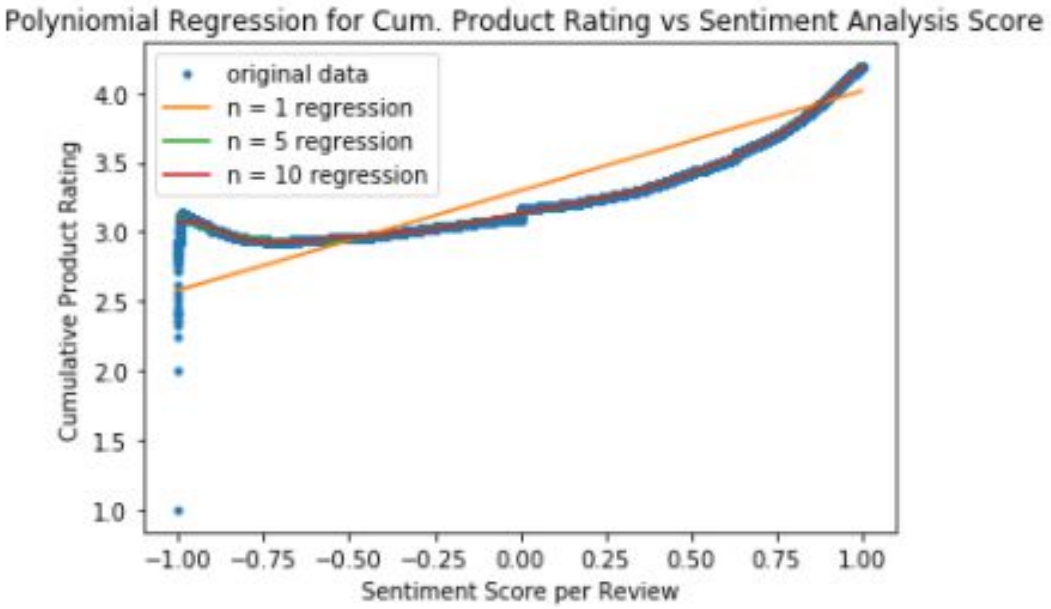

Figure 11. Plot for Cumulative Product Rating vs. Sentiment Analysis Score

5. Discussion

5.1. Analysis and Thoughts

In our implementation, we observed positive correlations on product rating for helpfulness rating, time and sentiment, while we saw generally non-correlative relationships for Lexical Diversity and Keywords. We'll take a look at each of the factors individually, evaluating their correlations and discussing factor specific limitations and strengths.

5.1.1. Helpfulness Rating

The correlation from our helpfulness rating factor seemed one of the strongest and most likely to be similar in an analysis with a larger sample set. We see a range of diverse cumulative scores along the 0 percent helpfulness score. As we move across the x axis to nonzero low scores, we see a definite decrease in product review scores. This could be due to lower rated reviews providing insufficient feedback for low scores, articulating poorly or providing misinformation. As we move towards higher-helpful rated helpful reviews, we see an increase in cumulative product score. While this may *seem t*o indicate that reviews with higher helpfulness ratings correlate with higher product scores, we must also take into account that our data in this experiment was not evenly distributed. As we can see from the dimensionality plot, a majority of our 5 star reviews had 100 percent helpfulness rating. However, given that this significant selection of 300,000 reviews were arbitrarily selected and displayed this skew, we might expect

that most highly “helpful” rated reviews generally give positive product ratings on Amazon. Therefore the significant sample size of the data are a strength to this particular experiment. The next steps for exploring this relationship

5.1.2. Time

We see a similar skew in the data with time as we did with helpfulness rating, but with a less definite correlation. From left to right our data and regression plot seems to just show the curve of the data arbitrarily rising and falling in cumulative product score, with the end of the graph leading positive just barely due to the slight significant amount of 5star reviews that occur later in the timeframe. Overall, the model does a much job of limiting residual error and being more evenly distributed, but falls short in not not being definite enough for us to claim a distinct correlation. Judging by the nature of factor itself, an intuitive guess that time might uncorrelated seems to be correct, but was still a valid hypothesis to experiment based upon, regardless.

5.1.3. Keyword Frequency

Our regression graph seems to paint the problem with our keyword frequency factory being an outlier that disrupts the model. That intuition would be partially correct; there are outliers, but the outliers are most of the data above about 2 occurrences of our keyword in a review, as our dimensionality graph indicates. In general , there is insufficient data for this experiment. In our regression model we see far too few data points for frequencies above 2, causing low change in cumulative product score and an almost linear model for every value of n. While the idea to study trends/ or a set of keywords, like cost-relating words, sounds pretty interesting we might need to rethink our implementation for this factor relative to the results we’ve seen here. We could try increasing the set of keywords and re-running the model or we could use a binary class on the x axis and compare product review scores of those who utilize the keyword once or more time, versus those who don’t utilize the keyword at all.

5.1.4. Lexical Diversity

At first glance the results from our lexical diversity dimensionality reduction plot seem to illustrate the same conclusion as keyword frequency: that there is insufficient data for this factor’s experiment. Most majority values for each star rating are clustered at low score of lexical diversity. Studying the regression plot over, we see a much better, more even dispersion of data points along the lexical diversity axis. Instead of insufficient data, it seems that there is no significant correlation between lexical diversity and product review score that we can draw. If we try to analyze *why* we see this curve we might be able to answer why. If we assume posts with high lexical diversity are more wordy, one reason for their wordiness might be their critique of the products. To fairly critique a product one needs to weigh its pros and cons. Doing so might

increase the length of your review considerably. If weighing a product's weaknesses and strengths, one might feel more compelled to vote somewhere in the middle of 1 star and 5. Next steps for this factor would be to either consider testing a new factor or use a larger sample.

5.1.5. Sentiment

Finally, judging by its dimensionality plot, sentiment has the smoothest distribution of ratings all factors in this project. We also see an interesting set of relationships in the regression plot. There is a sharp increase in cumulative product rating starting at the most negative sentiment. Perhaps the Vader Sentiment dictionary misinterpreted the meaning intent of a significant portion of reviews? Between the initial incline and the later incline beyond 50 percent, we see a much more level graph, though the relationship is still positively correlative. We might surmise that most of the same critical reviews that we observed in the lexical diversity trend populate this area as well. Cumulative product score around those ratings are close to how they appear here, though slightly above the scores in the middle dip of the sentiment graph. We could also argue that more critical reviews use both positive and negative language, nearly "evening out the score". Finally, at the latter end of the regression graph, we see a stepper positive correlation. As one might expect from this end, extremely positive sounding reviews give extremely positive ratings. The results from this particular experiment seem largely irrefutable, due to experimental semantics like even distribution compared to the others, thus there are few limitations in this particular experiment beyond being able to see " how" extremely negative or positive scoring words were used.

5.2. General Strengths and Limitations

Recapping from the individual factors and analyzing holistic strengths and weaknesses, we can surmise having that significant sample size is a general strength for for this project, validating our correlations, but identifying uneven samples by variable value is a severe limitation in confidently making those correlations in the first place. The dimensionality reduction plot also proved to a be valuable tool for the analysis and perspective of our results. Other possible limitations for this project include market base and resources. On market basis, we chose 3 random categories from amazon, but the general consumer of each category might each rate products slightly differently that how we've seen in their converged grouping. Finally, while significant, our project represented a small subset of the amount of reviews available on Amazon.com. This project could have benefitted from using a larger sample, but doing so might have impacted our abilities to compute the values on our machines. Having the capability to utilize larger datasets containing items such as metadata would have expanded the depth of exploratory analysis that we could've done.

6. Next Steps and Conclusions

General next steps for this project would be to continue to explore influential factors. Once enough significant factors have been identified, an interesting next step from our results would be to use machine learning to test the kinds of regression models we generated on actual datasets: classifying test data based on chosen factors and trying to guess the product's rating. If any particularly unique factors or correlations are discovered, the tools used to discover and analyze them could be sold to companies like Amazon or provided to companies as a marketing tool.

Overall, this project allowed us to give at least some definite answers to our research question of who, how and why do people buy and review amazon products. From our helpfulness ratings results, we say that the most influential reviews tend to rate a product positively. From our sentiment and lexical analysis, We know that more wordy or critical reviews collectively tend to rate products somewhere in between 5 stars and 2 stars, while shorter, highly positive reviews tend to give high ratings.

Finally. we believe that the kind of analysis performed for this project has been a worthwhile practical application of real world data analytics and numerical computations.

7. References

[1] R. He, J. McAuley. *Modeling the visual evolution of fashion trends with one-class collaborative filtering*. WWW, 2016
[2] J. McAuley. *Amazon Product Data*. UCSD

8. Code

A Jupyter Notebook containing code for the project can be found at
https://github.com/Mjacks3/CMSCDEMO