Exploratory Data Analysis of Amazon.com Book Reviews

by

Timothy Wong

A thesis submitted in fulfillment of the
Requirements for the degree of honors in

Statistics

University of California – Berkeley

2009

UNIVERSITY OF CALIFORNIA - BERKELEY

ABSTRACT

Exploratory Data Analysis of Amazon.com Book Reviews

By Timothy Wong

Advisor: Professor David Aldous
Department of Statistics

Amazon.com is originally found by Jeff Bezos in 1994 and has grown rapidly to become one of the most successful e-commerce businesses in the world. Today, Amazon.com is a Fortune 500 company and is the largest online retailer in the United States. One of the reasons that lead to the company success is the innovative review systems. The structured user friendly system has benefited both the company and customers. This thesis will find out the nature of a dataset from this review system. We would ultimately like to find out whether earlier reviews receive more and better feedbacks than later ones. Based on previous research, we would try to modify the approach in order to give a more precision conclusion to our initial question. Our primary goal is to observe whether earlier reviews tend to receive higher helpfulness ratings because of the duration of the review, instead of the review's content. Also, we would try to explain the nature of the dataset using summary statistics and exploratory data analysis; in particular, we would only focus on perspectives that are related to favorable votes and total votes.

TABLE OF CONTENTS

List of Figures ii
List of Tables iii
Background 1
Data Collection 5
Method 7
Results 9
Conclusion and Further Research 11

LIST OF FIGURES

Number *Page*

1. Amazon.com Layout 12
2. Correlation Table 13
3. Total Favorable Votes Against Reviewer 14
4. Total Votes Against Reviewer Index 15
5. Time Series Plot: Bookmarked For Death 16
6. Time Series Plot: Dream Warrior 16
7. Time Series Plot: Promises in Death 17
8. Time Series Plot: Terminal Freeze 17
9. Time Series Plot: When Giant Fall 18
10. First Type of Time Series Plot (% of Helpfulness over Total Favorable Votes) 19
11. Second Type of Time Series Plot (% of Helpfulness over Total Favorable Votes) 19
12. Bar chart (12-23) 20

LIST OF TABLES

Number *Page*

1. Usable Booklist....12
2. Summary Statistics for Individual Reviewer....13
3. Summary Statistics for Book....14

Background

Amazon.com is originally found by Jeff Bezos in 1994 and has grown rapidly to become one of the most successful e-commerce businesses in the world. Today, Amazon.com is a Fortune 500 company and is one of the largest online retailers in the United States. Unlike other online auction-based companies such as e-bay, Amazon.com focused on retail sale. With a rapid rate, Amazon.com has expanded in the world and has become one of the most popular retailing website in the world. The success is mainly due to its customer friendly website interface and innovative tools that aid the customers such as providing lists of best sellers, popular books and the recommendation system.

The recommendation system has been one of the most evolutionary features in Amazon.com and has been adopted by many other retail websites. The recommendation system allows people to express their opinions and gives ratings to the products that are listed on Amazon.com, including books, music, movies, electronics and more. Reviews are generated in the corresponding product when the customers leave their feedback and rating on the website. A reviewer ranking is introduced in order to monitor the quality of customers' comments. As visitors to the product page read the reviewers' comment about the product, they can also choose to answer the question "Was this review helpful to you" by clicking "Yes" or "No". The reviewer ranking is mainly based on both the amount and the percentage of helpfulness he or she has received. By clicking "Yes", the reviewer would receive one favorable vote and vice versa for clicking "No". Currently, there are two kinds of ranking; New Reviewer Rank and Classic Reviewer Rank. While Classic Reviewer Rank is the original ranking, New Reviewer Rank introduces a weighted average between helpfulness of reviews and the frequency that reviews is written. Top 1000 reviewers would receive a badge. The page that displays all customer reviews

is set by default to list reviews by “Most Helpful First”, but can also be changed to list by “Newest First”. Figure 1 illustrates the basic layout. In this paper, we would like to explore whether earlier reviews receives more votes and favorable votes over time and other related aspects.

We would first look at previous research that is done by Robert Huang in 2008. His objective is to study the relationships between different variables associated with individual customer reviews. His motivation to this research is that he would like to investigate the following:

1) Whether early written reviews for a book get better feedback than later ones of the same quality
2) How other factors affect the type of feedback a review receives
3) What effect different variables such as review rating and reviewer do

His data composed of 20 books, which are all released within the past two years. Only books with between 30 and 45 reviews are used. Throughout the data collection, the following information is collected daily:

1) the date the review was posted
2) the star rating the review gave the book
3) the amount of feedback the review received and how much of it was positive
4) the length of the review (number of words)
5) the reviewer’s rank
6) Two numbers attempting to quantify the quality of the review.

Based on this dataset, he carried out several kinds of exploratory data analysis to verify his hypothesis. First, he fitted a least square line between number of reviews and reviewer index for

every book. This showed that there is a negative relationship between the two variables in almost every case; hence verifies the possibility that earlier posted reviews receive more feedback.

Then, he made bar charts depicting the relationship between median amount of total and positive feedback for each group of 10 successive reviews for each book. The bar chart illustrates a large drop-off in the amount of feedback from the first 10 reviews and the next 10 reviews. Both the amount of positive feedback and amount of total feedback decline over time.

Furthermore, plotting the average rating by reviewers over time also shows that there exists a pattern with first 10 reviews giving higher rating than latter groups. The author deduces that it may be due to publishers attempt to solicit people to give positive responses to their book by leaving positive response to their book.

The author also points out that the reviews giving the book ratings closer to books' average rating usually get better feedback. He explains this by supporting the argument that Amazon visitors give feedback on reviews based on how much they agree.

Overall, the initial objective cannot be fulfilled, because it is still unclear whether early written reviews receive better feedback than later written reviews based on his analysis. The evidence that the author found is not statistically strong enough to give a precise conclusion. The main problem that the author faces is the existence of books that have few reviews, and these books do not give any statistical meaning. However, the author also found out some notable results, such as the relationship between review ratings and positive responses received. In fact, further research can be carried out for a closer look to the subject.

In this paper, we would ultimately like to find out whether earlier reviews receive more and better feedbacks than later ones. Based on Huang's framework, we would try to modify his

approach in order to give a more precision conclusion to our initial question. Our primary goal is to observe whether earlier reviews tend to receive higher helpfulness ratings because of the duration of the review, instead of the review's content. Also, we would try to explain the nature of the dataset using summary statistics and exploratory data analysis; in particular, we would only focus on perspectives that are related to favorable votes and total votes.

Data Collection

Although Amazon.com has many different kinds of products, here we would only focus on books. From previous research, Huang's dataset consists of books that are released within two years. Since Huang does not collect data at the beginning of the time when the books are released, the dataset may possess some skewness that may lead to a wrong conclusion. For instance, the difference in distribution between recording a book that has released two years and a newly released book would be hard to interpret. Bias may exist as we do not have the data from two years ago for the first book and there could be possible variations. This type of dataset may lack in information on the progression of votes, and the final percentage of favorable votes may become unreliable because of inconsistency of the dataset.

To avoid this problem, we would choose the books that are newly released; hence there would be no prior data regarding favorable votes and total votes, which could totally eliminate the variations that have been mentioned above. Hence, books from our dataset would not be randomly picked from Amazon.com, but were chosen based on the criterion that they were newly released and have no reviews posted. Besides, in order to minimize the possibilities of picking books that have few reviewers, we would look at the number of reviews of previously released books from the same author; hence we can deduce the popularity of the chosen book.

Data collection began on February 3rd, 2009 and ended on March 14th, 2009. The dataset contains a sample of 45 books from Amazon.com. To avoid the variations on voting trend due to specific population, in particular a specific type of genre, books were chosen randomly from various genres including science fiction, biographical, mystery and thrillers, romance, non-fiction, literature and fiction, religion, and history. Data are collected daily and for each of the 45 sampled books, we would collect the following data:

1) Book Title
2) Dates of Data Collection
3) Star Rating from Each Reviewer
4) Reviewer Name
5) Reviewer Index (First Review denoted by “1”)
6) Number of Helpful Votes for Each Review
7) Number of Total Votes for Each Review
8) Total Favorable Votes for the Book
9) Total Votes for the Book

Note that in order to illustrate the reviewer also favors him/her own review, the favorable votes and totals votes for the review are adjusted and would start with 1 instead of 0. Also, we notice that the number of votes for some reviewers has a decrease in quantity (in both favorable votes and total votes), hence we would assume that the quantity that has decreased never exists and pick at one previous vote randomly to delete from the previous data of the reviewer.

After the data collection, we find that some books do not receive an optimum amount of reviews that would give precise result to the analysis. Hence, we would not use books with less than 10 reviews and this leaves a total of 28 books in the dataset. Table 1 summarizes the books that we would use for analysis.

Method

Recall, in this paper, we would like to ultimately find out whether earlier reviews get more favorable votes. Also, we would explore the nature of this 28 books dataset by exploratory data analysis. We would first look at the correlations table among all the variables, which is shown in figure (2). The correlation table clearly visualizes the relationships among all variables, based on this figure; we would summarize some meaningful relationships below. But before that, we would first try to investigate our main objective: whether earlier reviews receive more favorable votes.

To address the research question, we would first fit the data to two simple regression models in order to observe any notable trend:

1) Amount of total votes against reviewer index
2) Amount of favorable votes against reviewer index

For all 28 books, the number of votes and the corresponding reviewer index are plotted on the same plot, and ascending order of reviewer index here would represent time. We would fit a least square line on one plot and we would like to note any notable trends in all books. Ultimately, we would obtain two plots: the amount of total votes across reviews and the amount of favorable votes across the reviews. From these two plots, we would know the relationship between time and the final amount of votes that each reviewer receives.

To show that whether earlier reviews receive more favorable votes, we would use time series plots. Here, although time domain is used, the time variable is eliminated and is replaced by the total votes of book. We can do this because a change in the total votes of book can be interpreted in a change in time. We would create such plot for each book. The horizontal axis

would be the total number of votes, while the vertical axis would be the number of favorable votes for each review. Each line in the plot would represent one particular review. We would expect that earlier reviews receive more favorable votes.

Next, we are interested in exploring the nature of the given dataset using summary statistics and exploratory data analysis. In particular, we would first compute the numerical summary statistics, and try to find if there is any interesting pattern. Based on the correlation table in the previous section, we saw that percentage of helpfulness and favorable votes are highly correlated, hence we would like to investigate the dataset in following perspectives:

1) The relationship between the time of a review posted and the percentage of helpfulness
2) Whether earlier review has higher final percentage of helpfulness

For (1), we would expect that the earlier the review is posted, the higher percentage of helpfulness it will receive, regardless of the content of reviews. To visualize this, we would first compute the percentage of helpfulness: $\frac{Total\ Favorable\ Vote}{Total\ Vote}$. Then, we would use time series plots that are similar to the previous part to look for any trend.

For (2), we would use a bar chart to visualize such relationship. For each book, we would create a bar chart and each bar would represent one review. Our hypothesis is that earlier review will have higher final percentage of helpfulness.

Results

First, we would look at the numerical summary statistics. Table (2) and table (3) respectively show the summary statistics locally (individual reviewer) and globally (Book). We

can see that for popular book, it is possible that one individual reviewer would contribute up to one-third of the votes from the book. Moreover, it is interesting to notice that some reviewers are very popular and left reviews in many books, such as Harriet Klausner. Usually, they are the popular reviewers that usually leave helpful reviews for books.

We would then look at two simple regression plots that are shown in figure 3 and figure 4. Two relations are shown: amount of total favorable votes against reviewer index and amount of total votes against reviewer index. These two plots agree with the previous research that all lines have negative slope, which imply that earlier reviews do receive more votes globally. Yet, we cannot conclude such relationship exists solely based on these two regression results as votes and reviewer index are not the only variables that exist. In fact, it is possible that the inversely proportional relationship that is shown in figure 3 and figure 4 are due to other variables. Hence, the relationship that we notice from the plots may become a casual relationship. Since it is difficult to avoid these lurking variables, to verify whether earlier review receives more votes, we would proceed to the time series plot for favorable votes. We have randomly picked 6 plots and are shown from figure (5) to figure (9). Notice that the bold lines represent the first three reviews, and gray color scheme sort the rest of the reviews from earliest to newest. From all the plots, it is obvious that three bold lines are usually on the top, although some slight variations occur in some books. We can see that most of the lines that are on top are closed to red color and lines that are at the bottom are usually closed to blue color. Hence, based on these plots, we can say that the dataset shows a trend that earlier reviews receive more favorable votes, but since some slight deviations exist, we cannot neglect the possibilities that other factors are controller the number of favorable votes. Also, it is interesting to note that there are more rapid increases in favorable votes in some particular days.

Now, we would turn our focus to the relationship between the time of a review posted and the percentage of helpfulness. From the plots, we fail to find a common pattern among all books. Although the lines in some plots remain constant, some plots have lines that are randomly shaped and have no specific pattern. Figure 10 and figure 11 show two main patterns of the plots. Hence, we cannot give any statistical conclusion to this relationship. Yet, further investigation would be needed to verify that the two variables have no relationship.

Then, we would look at the relationship between the reviewer index and the final percentage of favorable votes. Figure 12 to figure 27 give the bar chart for each book. Our hypothesis is that earlier reviews have higher favorable vote percentage. From the plots, we can see that the earliest review usually receive the highest percentage of favorable votes. However, we did not see a trend that earlier reviews have higher favorable vote percentage. Hence, earlier reviews may not give higher favorable vote percentage. One possible reason to explain earliest reviewer receiving highest favorable vote percentage is that these earliest reviewers are the one who have deep interest in the book.

Conclusion and Further Research

In this exploratory data analysis, we explored the relationship between the change of favorable votes and the change of total vote for every reviewer. Our hypothesis is that earlier reviews receive more favorable votes. Despite some exceptions, the time series plots show that most books having a trend those earlier reviewers receive more favorable votes. One possible reason to explain is the structure of Amazon's feedback system. The system displays the reviewer that receives most favorable votes and has highest percentage of favorable votes on top, instead of sorting it by the time of review that is posted. Also, we looked at the dataset in some other aspects, such as time against percentage of favorable votes and reviewers against final percentage of favorable votes. Unlike the trend that we found in favorable votes, we fail to find a trend to represent the relationship between time and percentage of favorable votes although the percentage and favorable votes are highly correlated. Moreover, our hypothesis that earlier reviews get higher percentage of favorable votes is rejected, although the first reviewer in most books have the highest percentage of favorable votes, we cannot observe the same phenomenon in the next few reviews. To conclude, further research can be done in several aspects. For instance, by carrying the same analysis in different period of time, one can observe whether seasonal effect exists. Future researchers can also consider fitting the data to a regression models based on all the variables and determine which variables can best represent the distribution of favorable votes. Based on the correlation table, one can also explore the other highly correlated variables. Also, future researchers should use a larger dataset and collect the data for a longer period of time, so that the analysis will represent the whole population better.

Figure 1: Amazon.com Layout

Author	Book Title
Grant Morrison and Tony Daniel	Batman R.I.P.
Patricia Briggs	Bone Crossed
Sue Ann Jaffarian	Booby Trap
Lorna Barrett	Bookmarked for Death
Shirley Rousseau Murphy	Cat Playing Cupid
Lora Leigh	Coyote's Mate
Jennifer Crusie, Anne Stuart, and Lani Diane Rich	Dogs and Goddesses
Sherrilyn Kenyon	Dream Warrior
Jordan Dane	Evil Without a Face
Thomas P.M. Barnett	Great Powers of America
Taylor Anderson	Maelstrom
Eileen Wilks	Mortal Sins
Robert B. Parker	Night and Day
J.D. Robb	Promises in Death
C. J. Sansom	Revelation
James Patterson and Michael Ledwidge	Run for Your Life
Carolyn Jewel	Scandal (Berkley Sensation)
Connie Brockway	So Enchanting
Alex Irvine	Supernatural
Lincoln Child	Terminal Freeze
Thomas E. Ricks	The Gamble
L. A. Banks	The Thirteenth
Brooke Taylor	Undone
Michael J. Panzner	When Giants Fall
Kim Harrison	White Witch
C. S. Friedman	Wings of Wrath
John Birmingham	Without Warning
Joe Torre and Tom Verducci	Yankee Years

Table 1: Usable Booklist

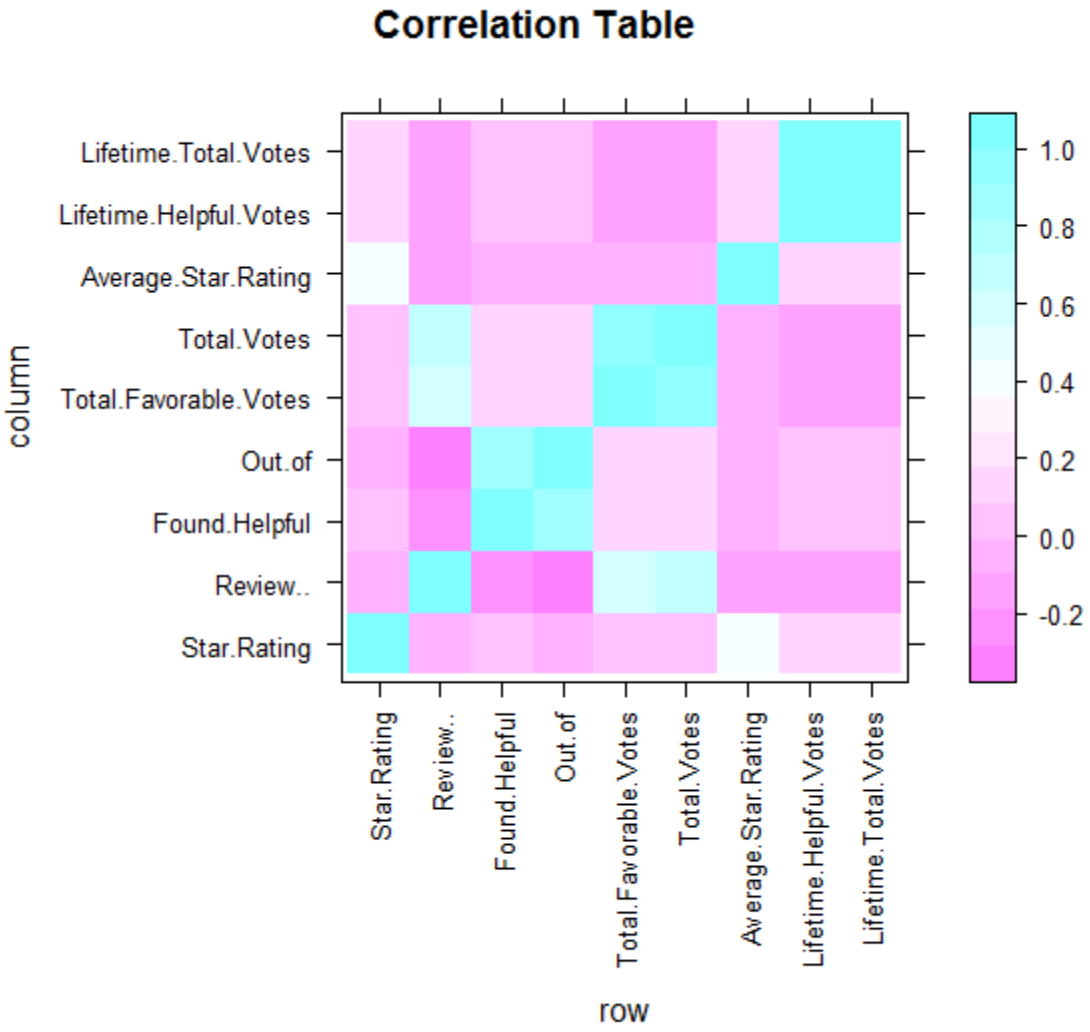


Figure 2: Correlation Table

Individual	Total Favorable Votes	Total Votes	Percentage of Favorable Votes (%)
Minimum	1	1	3.57
1st Quantile	1	2	50
Median	2	4	75
Mean	4.5	6.98	72.1
3rd Quantile	5	8	100
Maximum	83	89	100

Table 2: Summary Statistics for Individual Reviewer

Book	Total Favorable Votes	Total Votes	Percentage of Favorable Votes (%)
Minimum	0	0	0
1st Quantile	29	46	51.1
Median	83	134	62.3
Mean	101.9	176.9	63.2
3rd Quantile	176	304	71.3
Maximum	296	510	100

Table 3: Summary Statistics for Book

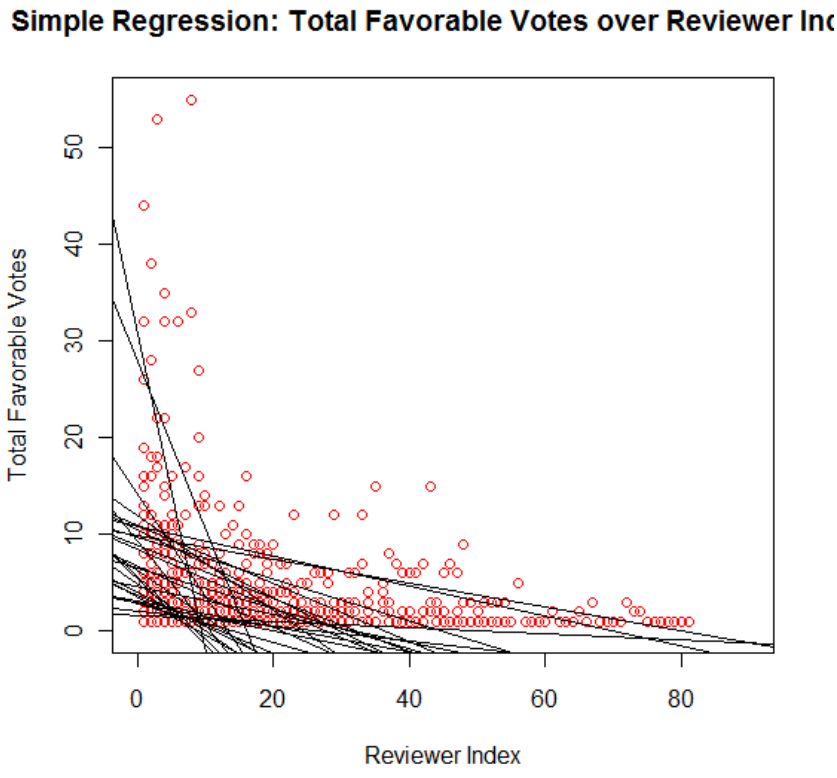


Figure 3: Total Favorable Votes against Reviewer Index

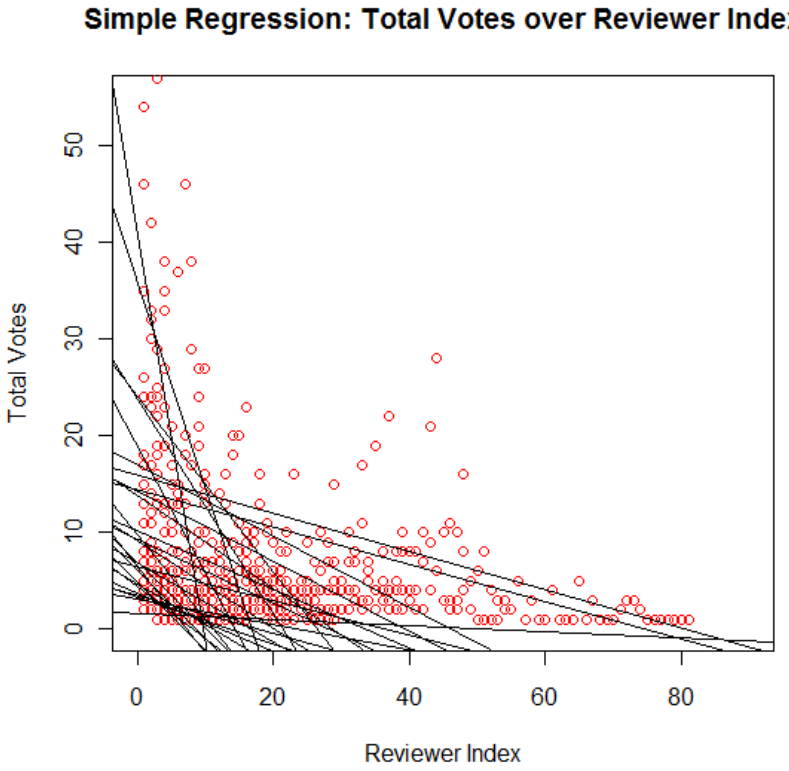


Figure 4: Total Votes against Reviewer Index

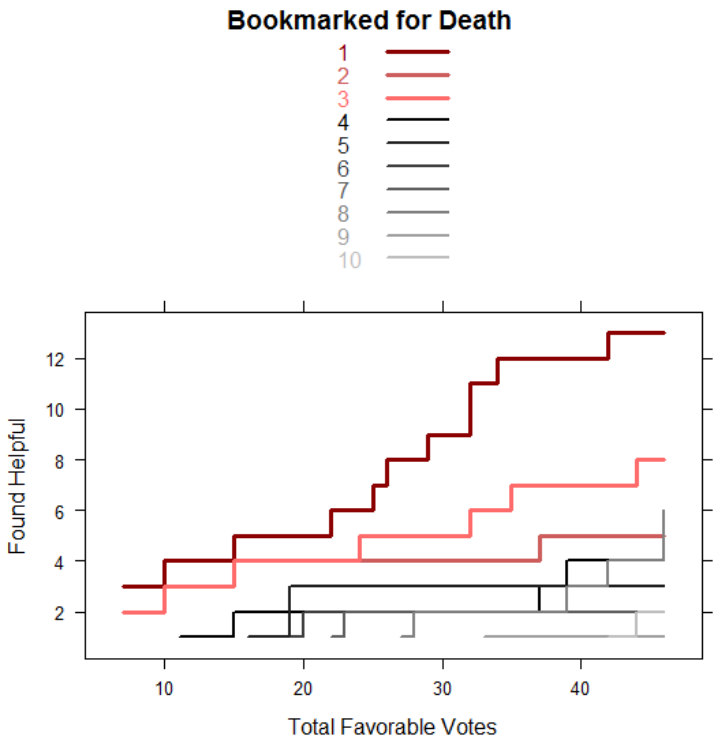


Figure 5: Time Series Plot: Bookmarked for Death

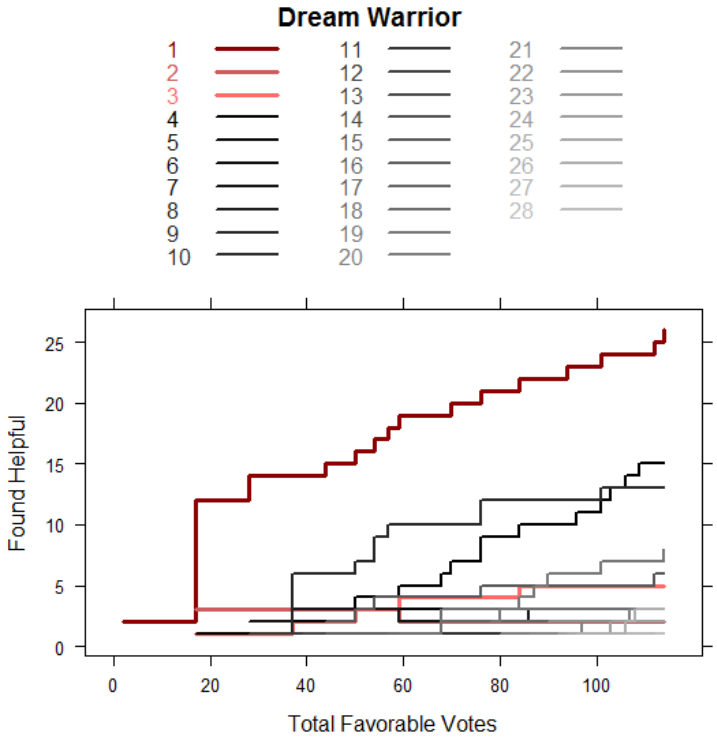


Figure 6: Time Series Plot: Dream Warrior

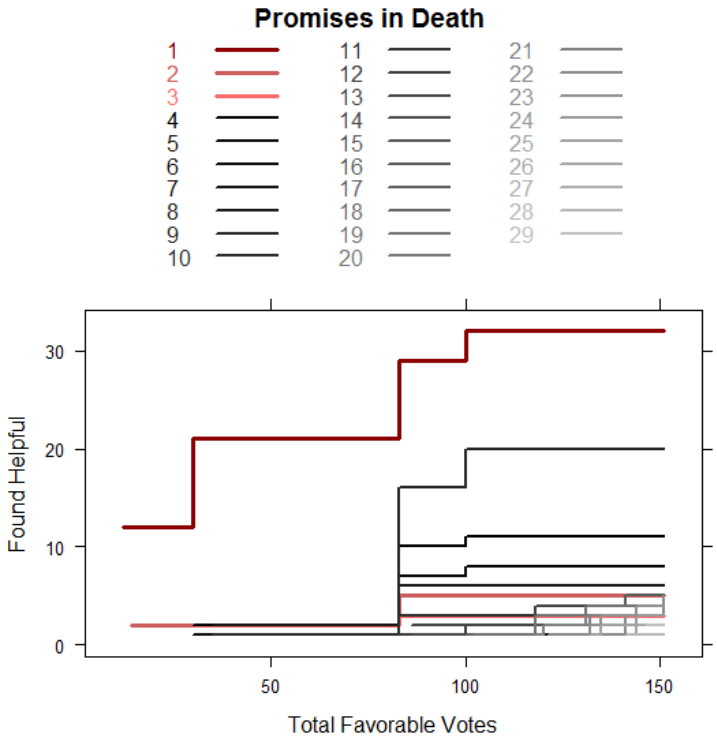


Figure 7: Time Series Plot: Promises in Death

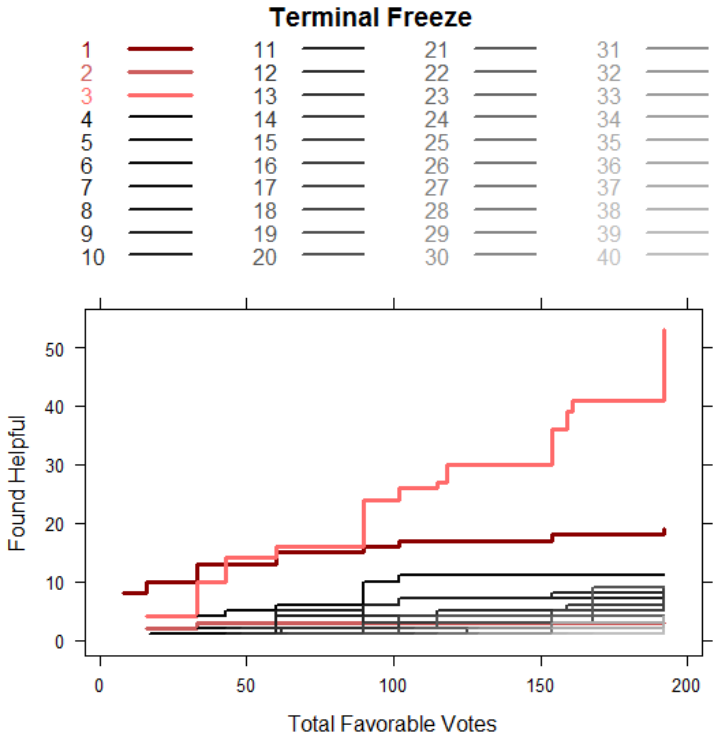


Figure 8: Time Series Plot: Terminal Freeze

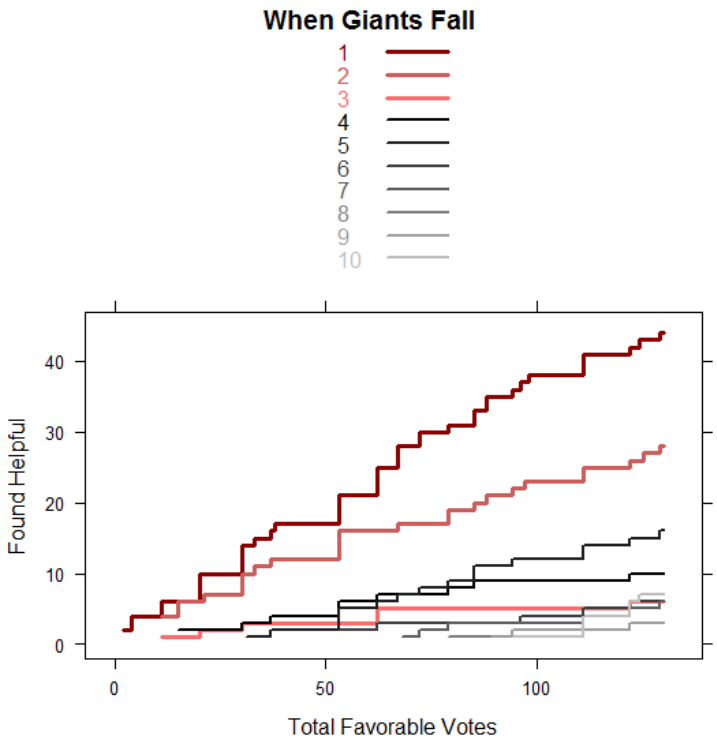


Figure 9: Time Series Plot: When Giants Fall

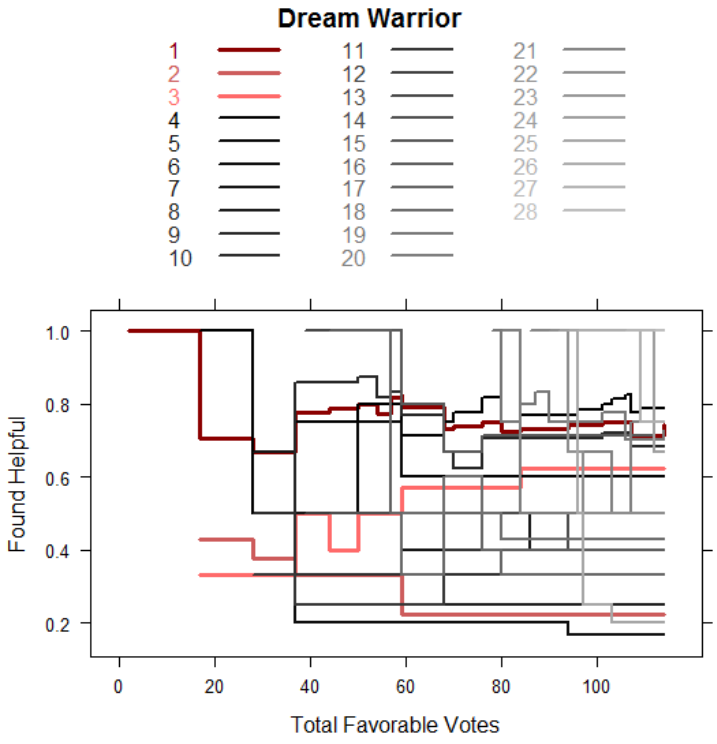


Figure 10: Percentage of helpfulness over Total Favorable Votes (First Kind- No Pattern)

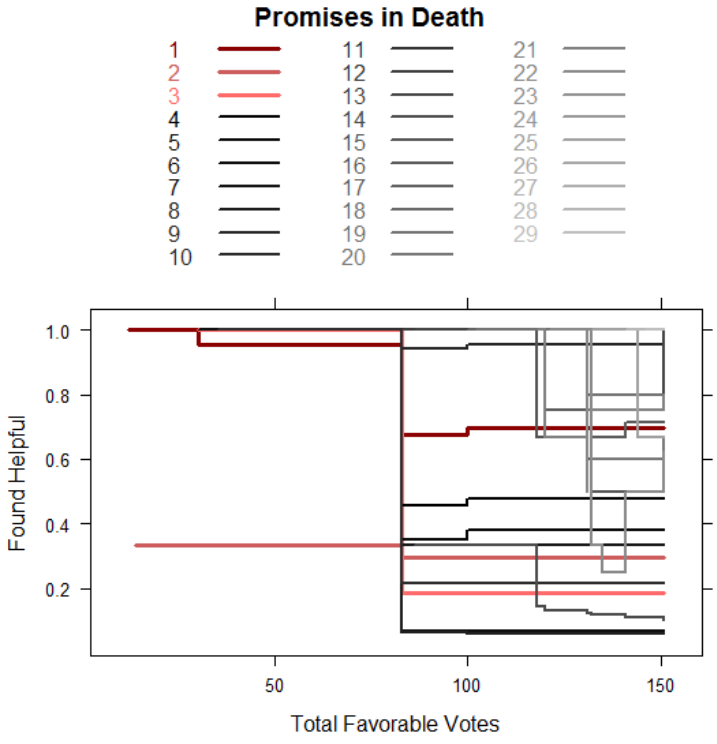
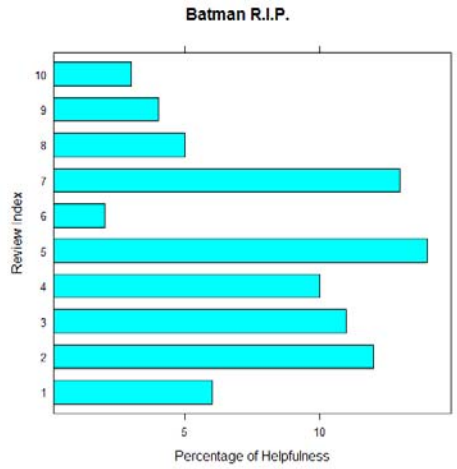
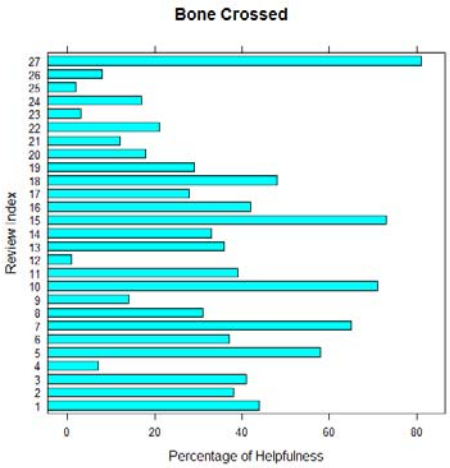
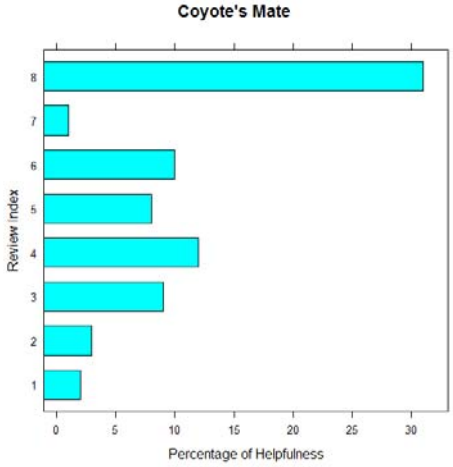
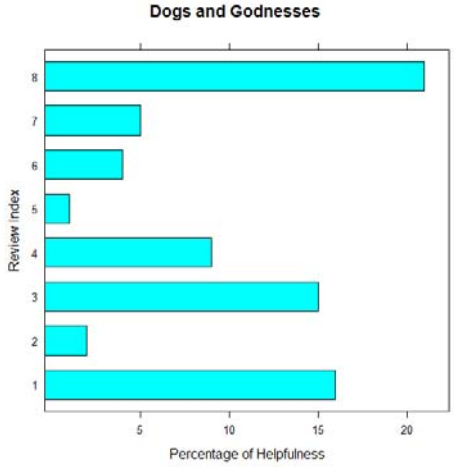


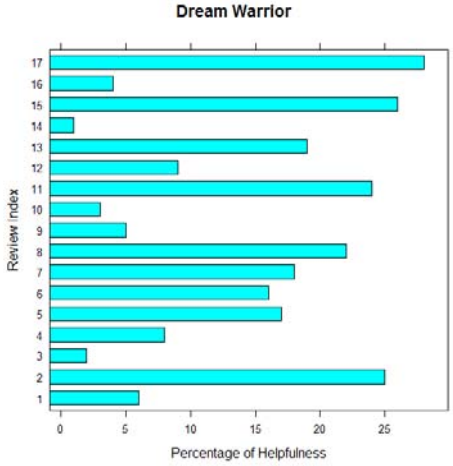
Figure 11: Percentage of helpfulness over Total Favorable Votes (Second Kind – With Pattern)

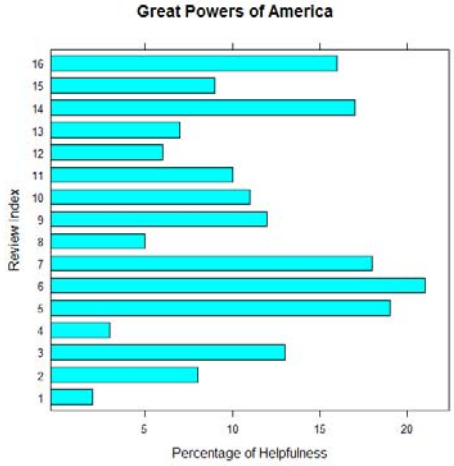

Batman R.I.P.
Review Index
Percentage of Helpfulness

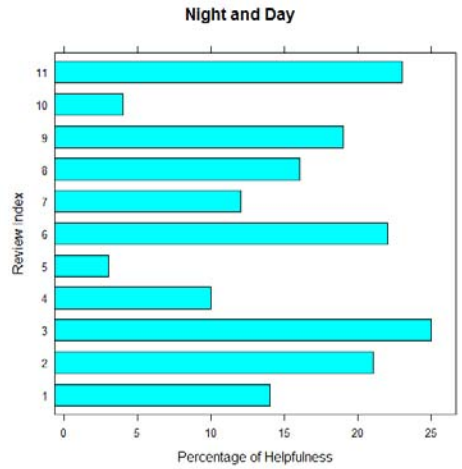

Bone Crossed
Review Index
Percentage of Helpfulness

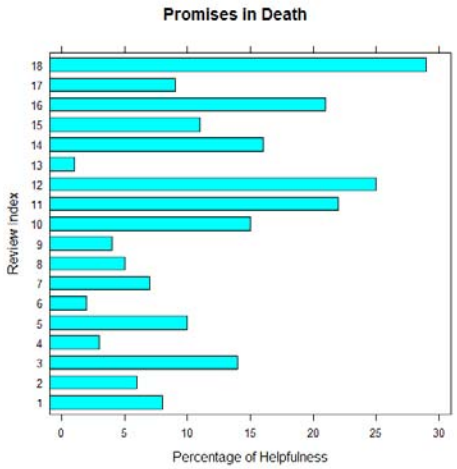

Coyote's Mate
Review Index
Percentage of Helpfulness

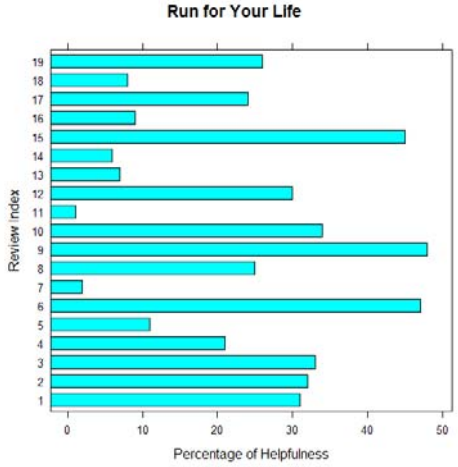

Dogs and Godnesses
Review Index
Percentage of Helpfulness

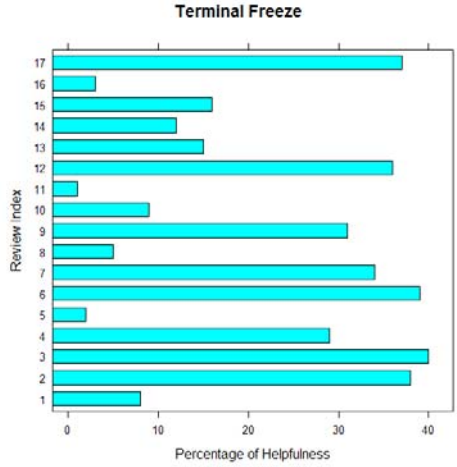

Dream Warrior
Review Index
Percentage of Helpfulness

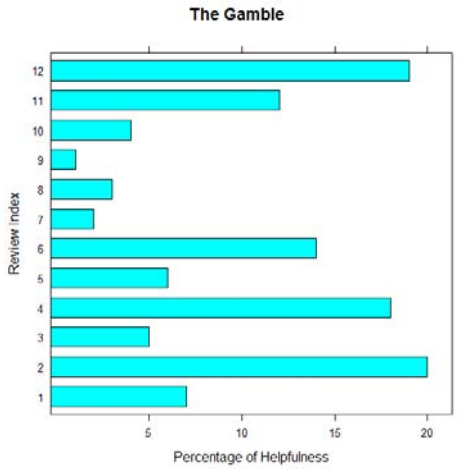

Great Powers of America
Review Index
Percentage of Helpfulness

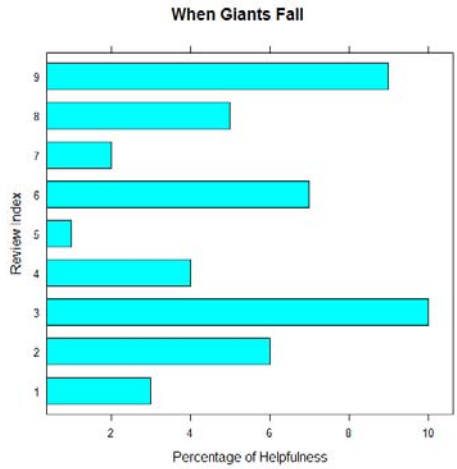

Night and Day
Review index
Percentage of Helpfulness

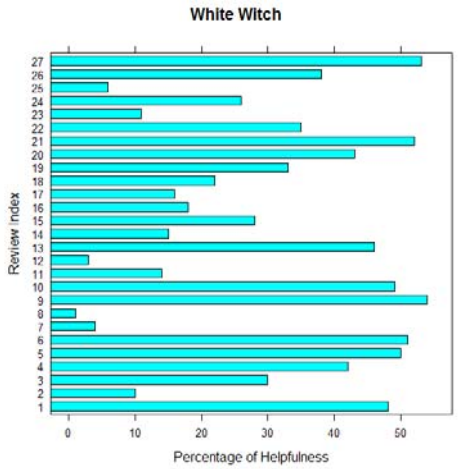

Promises in Death
Review index
Percentage of Helpfulness

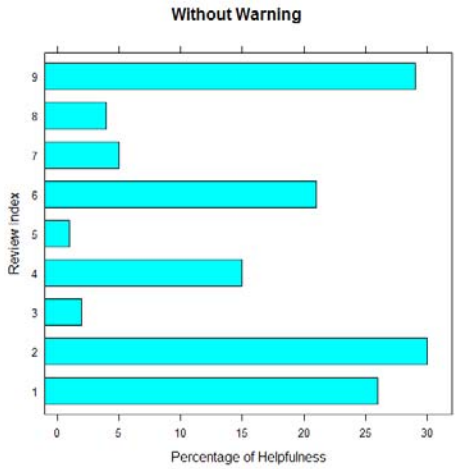

Run for Your Life
Review index
Percentage of Helpfulness

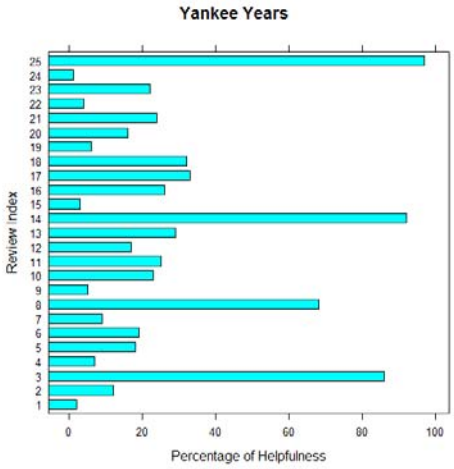

Terminal Freeze
Review index
Percentage of Helpfulness


The Gamble
Review index
Percentage of Helpfulness


When Giants Fall
Review index
Percentage of Helpfulness


White Witch
Review Index
Percentage of Helpfulness


Without Warning
Review Index
Percentage of Helpfulness


Yankee Years
Review Index
Percentage of Helpfulness