

# **A Study of Amazon User Review Data using Visualization**

by

**Preeti Bamane**

A Project Report Submitted  
in  
Partial Fulfillment of the  
Requirements for the Degree of  
Master of Science  
in  
Computer Science

Supervised by

Dr. Rajendra Raj

Department of Computer Science

B. Thomas Golisano College of Computing and Information Sciences  
Rochester Institute of Technology  
Rochester, New York

December 2016

# **Abstract**

## **A Study of Amazon User Review Data using Visualization**

**Preeti Bamane**

**Supervising Professor: Dr. Rajendra Raj**

Reviews are heavily used by customers when purchasing products from any e-commerce website. Improving it is essential for retaining the existing customers enticing new customers. This project aims at finding relations between different aspects of a review, including the length of the review, overall star rating of the review, helpfulness of the review, frequency of the reviewer, price of the product, etc.

This project studies the various statistical trends attributed by these aspects of the reviews by using advanced visualization techniques. The relations were analysed using Tableau. An amazon review dataset containing 7.8 million reviews for 498k electronic products was used for the purpose of analysis.

The project focuses on performing empirical analysis on the amazon review data as opposed to sentiment analysis of the actual review text.

# Contents

<b>Abstract</b> . . . . .	<b>ii</b>
<b>1 Introduction</b> . . . . .	<b>1</b>
1.1 Dataset Description . . . . .	2
<b>2 Design &amp; Implementation</b> . . . . .	<b>3</b>
2.1 Visualization tool exploration . . . . .	3
2.2 Data Collection and Preparation . . . . .	4
2.3 Data Cleaning . . . . .	5
<b>3 Visualizations and Analysis</b> . . . . .	<b>7</b>
3.1 The distribution of ratings across the total number of reviews . . . . .	7
3.2 The growth in the number of reviews over the years . . . . .	8
3.3 Length of the review and the number of reviewers . . . . .	9
3.4 Number of helpful votes and the number of reviews . . . . .	9
3.5 The average ratings across the number of years . . . . .	10
3.6 Returning reviewers . . . . .	11
3.7 Reviews and the price of the product . . . . .	12
3.8 The number of reviews and helpfulness . . . . .	13
3.9 The distribution of ratings among infrequent and frequent reviewers . . . . .	14
3.10 The distribution of helpfulness percentage among infrequent and frequent reviewers . . . . .	17
3.11 A comparison between ratings and helpfulness . . . . .	19
3.12 Common words used in reviews . . . . .	19
3.13 The length of the review and the ratings . . . . .	21
3.14 The length of the review and its helpfulness . . . . .	23
3.15 The length of the review and the frequency of reviewers . . . . .	24
3.16 The growth of the length of the reviews over the years . . . . .	26
3.17 The review length and the price of the product . . . . .	27

3.18 The price of the product vs the frequency of the reviewers . . . . .	28
3.19 The price of the product vs Helpfulness . . . . .	29
<b>4 Conclusions . . . . .</b>	<b>31</b>
<b>5 Future Work . . . . .</b>	<b>33</b>
<b>Bibliography . . . . .</b>	<b>34</b>

# List of Figures

1.1	Sample Review . . . . .	2
2.1	Schema structure . . . . .	6
3.1	Ratings vs Number of reviews . . . . .	7
3.2	Month,year vs Number of reviews . . . . .	8
3.3	Length of the review vs the number of reviewers . . . . .	9
3.4	Number of helpful votes vs the number of reviews . . . . .	10
3.5	Month vs The average number of ratings . . . . .	11
3.6	Number of reviewers vs Number of reviews . . . . .	12
3.7	Price of the product vs Number of reviews . . . . .	12
3.8	Helpfulness vs Rating(for Chromecast) . . . . .	13
3.9	Rating vs Number of infrequent reviewers . . . . .	14
3.10	Rating vs Number of frequent reviewers . . . . .	15
3.11	Rating vs Frequency of reviewers . . . . .	16
3.12	Helpfulness vs Number of infrequent reviewers . . . . .	17
3.13	Helpfulness vs Number of frequent reviewers . . . . .	18
3.14	Helpfulness vs Rating . . . . .	19
3.15	Word cloud for good ratings . . . . .	20
3.16	Word cloud for bad ratings . . . . .	21
3.17	Review length vs Ratings . . . . .	22
3.18	Review length vs Helpfulness . . . . .	23
3.19	Review length vs Reviewer frequency . . . . .	25
3.20	Review length vs time . . . . .	26
3.21	Review length vs Price groups . . . . .	27
3.22	Price groups vs Frequency of reviewers . . . . .	28
3.23	Price groups vs Helpfulness percentage . . . . .	29

# Chapter 1

## Introduction

In today's world, the significance of online shopping is growing day by day. The business ideas have been refashioned and completely transformed by making it so easy for the customers to purchase anything they want at just one click of a mouse button. It is becoming even more popular due to its high level of convenience. The only thing customers must have is Internet and appropriate method of payment. Amazon.com is one such widely known E-commerce website and it is being used world-wide. It was initially known for its huge collection of books but later it was expanded to sell electronics and other home appliances and consumer products. At present, Amazon is known to sell millions of products. This growth of E-commerce gave importance to customer needs and opinions which in turn gave rise to an important aspect of online shopping known as 'User Reviews'.

User reviews are customer suggestions and opinions about the product which help other customers make decisions about that product. Such review systems form the backbone of E-commerce. The goal of this project is to understand and analyze the Amazon User Review Dataset with the help of different visualization techniques. These visualization techniques will help showcase various informative statistical trends which will provide us with insights about the Amazon Review system. These insights will help in exploring the possible improvements that can be done to satisfy the customers. Major work will involve empirical analysis for data understanding and exploration by taking into consideration, the various metrics related to the user reviews as opposed to sentimental analysis on the review text which aims at understanding the overall emotion of the reviews which has been done previously [3] [2].

## 1.1 Dataset Description

In order to understand the structure and schema of the dataset, let us look at a sample Amazon Review.

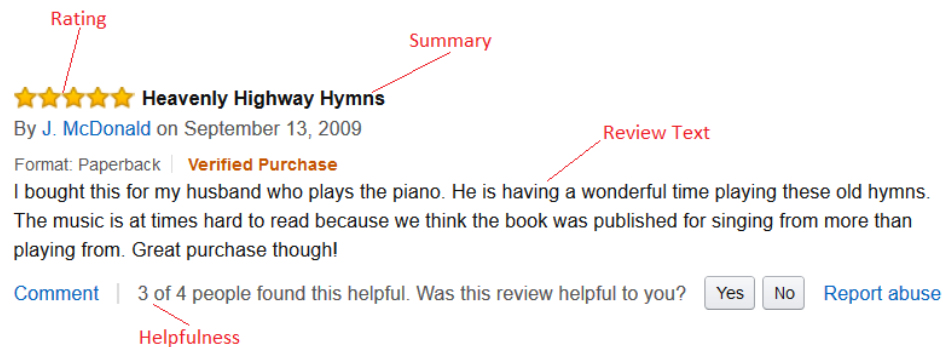


Figure 1.1: Sample Review

As seen in Fig. 1.1, an Amazon User Review consists of four important aspects:

- Summary : The title of the review
- Review text : The actual content of the review.
- Rating : User rating of the product on a scale of 1 to 5.
- Helpfulness : The number of people who found the review useful.

These aspects will help us understand and analyze the reviews in order to derive insights.

The Amazon User Review dataset [6] [5] is freely available on request from the Stanford University data repository [4]. This dataset contains approximately 35 million user reviews spanning approximately 18 years. This dataset is divided into two parts. One part contains all the information about the reviews and reviewers. The second part contains information about the products and companies. All together, this dataset contains information for about 9 million products.

## Chapter 2

# Design & Implementation

The workflow of this project involves four stages:

- Visualization tool exploration
- Data collection and preparation
- Analysis from visualizations
- Conclusions derived from visualizations

### 2.1 Visualization tool exploration

The tools considered for visualization were Tableau, R, Leaflet, Charted. Out of these, the tool selected for visualization was Tableau. Tableau is considered to be one of the best tools for data visualization in recent times [1]. Tableau is capable of turning humungous amounts of data into visualizations which are interactive and also visually attractive. Moreover, this can be achieved in very less amount of time as compared to other data visualization tools. Another plus point of this tool is that it has a user friendly and highly interactive drag-and-drop user interface. It is compatible across various well-known platforms. Tableau is a winner of multiple awards in the business analytics world for its highly attractive yet functional data visualization capability. Due to these features, Tableau has revolutionized the business analytics world by allowing the analysts to spend more time on visual analysis rather than development. Moreover, it has a free license for students.



## 2.2 Data Collection and Preparation

The Amazon Review Dataset [4] is freely available online on the SNAP datasets website. This dataset contains user reviews of products sold on the Amazon website. This dataset consists of two parts. The first part contains about approximately 35 million reviews spanning 13 years of product review data. It contains information related to the user providing the review, the time related information and the characteristics of the review like the review length, summary etc. The second part consists of metadata of about 9 million products which consists of product description, brand name etc. In order to obtain conclusive results, a subset of this enormous dataset was used for analysis. This subset consists of the reviews of electronic items sold on the Amazon website.

The electronic dataset in itself contains 7.8 million reviews and its relevant data. The meta-dataset contains details of all the electronic products stored on Amazon. Following is the description of the attributes of both parts of the dataset [4] :

- Review dataset
  - reviewerID - the id of the user who has given the review
  - asin - Amazon product id
  - reviewerName - name of the user who has given the review
  - helpful - the number of times the review was voted helpful
  - reviewText - the actual content of the review
  - overall - the rating of the product ranging from 1 to 5
  - summary - the title of the review
  - unixReviewTime - the time of the review in Unix format
  - reviewTime - the time of the review
- Meta dataset
  - asin - Amazon product id

- title - product name
- price - the price of the product
- imageUrl - the link to the image of the product
- related - other products related to this product which were either viewed or bought by other customers
- salesRank - information relevant to the rank of the sales
- brand - the name of the brand of the product
- categories - the possible categories the product can belong to

## 2.3 Data Cleaning

The freely available Amazon User Review dataset was originally in JSON format. There was no standard schema for this dataset. For e.g. some of the review texts were more than 20 kb (~10 pages of text) while some other reviews were as short as one word. There was no definite way to figure out the schema. The schema structure of the data is as given in 2.1

Since the tool chosen for visualization was Tableau, certain factors were taken into consideration. First, Tableau doesn't support data in JSON format. For this reason, data was stored in MySQL. MySQL is an open source database and it works seamlessly with huge amounts of JSON data. In order to read the data into MySQL, a script was written in python. Following is the part of the script that was used to parse through the data.

```
def parse(fpath):
    print("Loading compressed file " + fpath + "..")
    gZippedFile = gzip.open(fpath, 'r')
    for line in gZippedFile:
        yield eval(line)
```

Since there was no definite schema, a significant portion of time was spent on figuring out the size of each attribute with respect to its datatype. As mentioned above, the size of

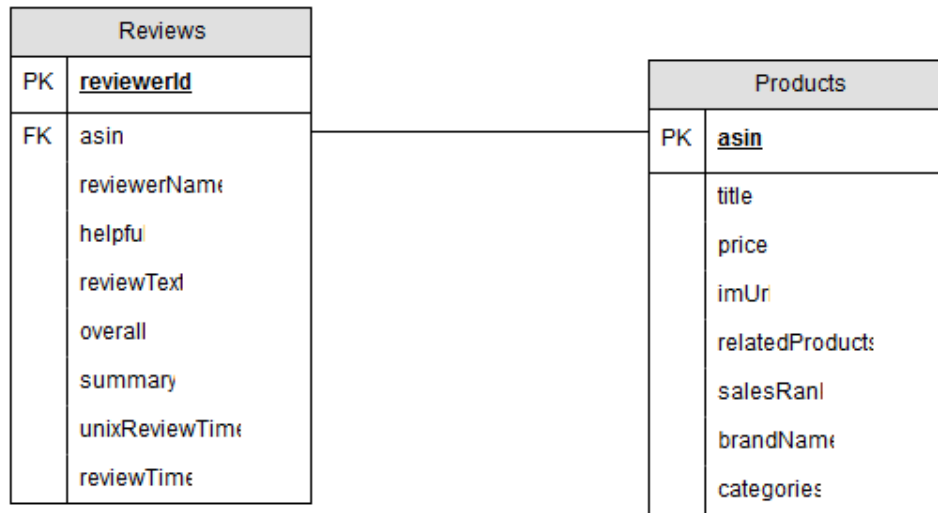


Figure 2.1: Schema structure

the review texts was very indefinite. The date and time related attributes also needed some fixing. Some of the time information in the data was inconclusive. For eg., the value ‘18, 2011’ doesn’t make any sense and needed to be scraped out from the data. Other attributes like reviewerName and summary were taking up extra space. For eg., no reviewerName cannot have a size of 100 characters. All such attributes were given a predefined maximum length for the purpose of space conservation.

The review dataset contained some duplicate review items. The reason for this was that Amazon combines the reviews of the products which are almost identical. For eg., the book review for the book “Deception Point” was duplicated due to the presence of multiple versions of the book namely paperback, kindle etc. The reviews of these various versions were combined by Amazon and thus repeated in the dataset in different categories. These duplicate records were scraped off from the dataset to avoid data redundancy which will increase the accuracy of visualization analysis.

## Chapter 3

# Visualizations and Analysis

### 3.1 The distribution of ratings across the total number of reviews

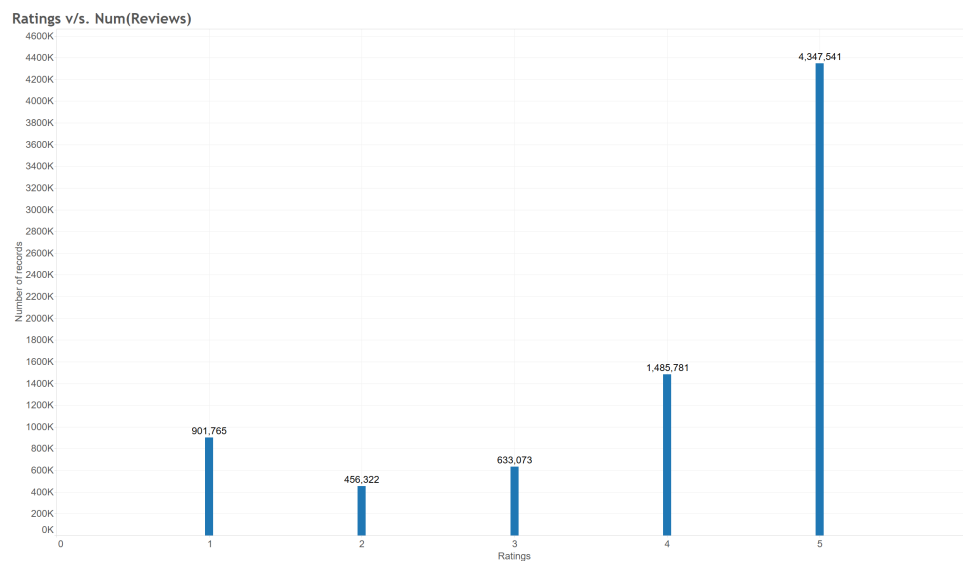


Figure 3.1: Ratings vs Number of reviews

The above figure shows how the user ratings of electronic products are spread among the total number of reviews. As it can be seen from the above distribution, a major portion of the reviews have received a 5 star rating. After the perfect 5 star rating, reviewers tend to rate products as 4 star or 1 star. This means most reviewers give a full 5 star if they're extremely happy with the product without giving it a second thought. The same explanation

goes for a 1 star rating where the reviewer is extremely unhappy with the product. The 4 star rating is often considered as a near perfect rating which is why it is the second highest rating given by the reviewers. Very few reviewers give 2 or 3 star rating as compared to other rating levels which leads to the conclusion that not many people like to analyze the product describing its plus and minus points. Due to this extreme rating given by reviewers, most of the products show a high end average star rating.

### 3.2 The growth in the number of reviews over the years

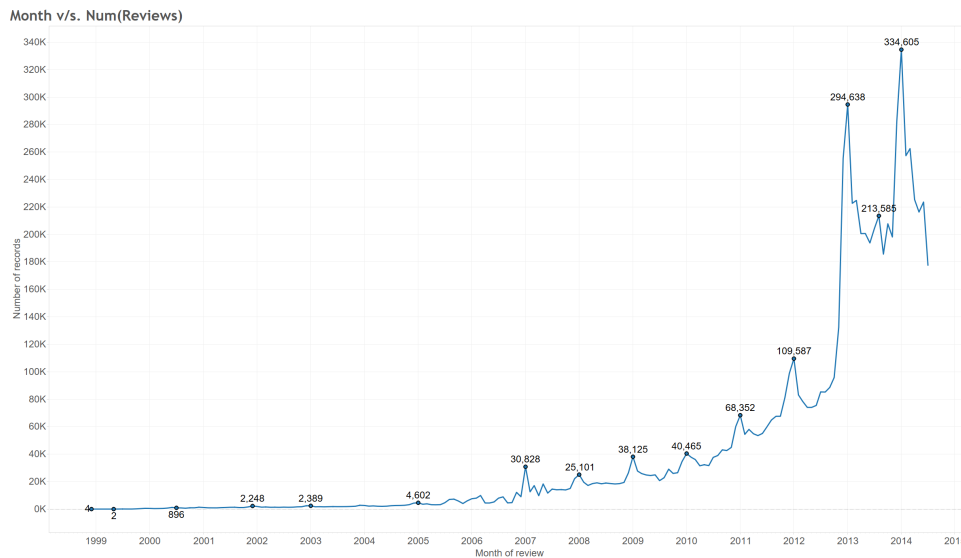


Figure 3.2: Month,year vs Number of reviews

As it is seen from the above graph, the number of reviews is clearly on the rise. In the year 1999, it showed the lowest number of reviews which was as low as 2. The number of reviews is on the increase there on. This clearly confirms the growing importance of reviews over the years and also tells us that the reviews are and will remain an important factor in online shopping for a long time until some new technology shows up.

### 3.3 Length of the review and the number of reviewers

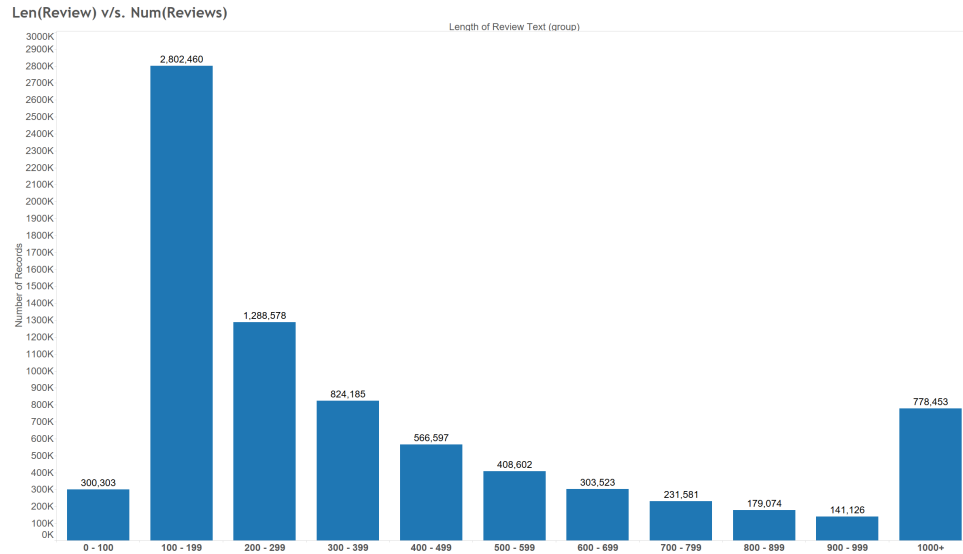


Figure 3.3: Length of the review vs the number of reviewers

As we can see from the above graph, a major chunk of the reviewers have given the review in about 100-200 characters. As the number of characters, i.e. the length of the review, increases, the number of reviewers decreases consistently. This just shows that users simply don't find the need as well as time to give in a detailed description of their experience with the new product which is quite understandable.

### 3.4 Number of helpful votes and the number of reviews

There is an attribute in the dataset called helpfulness. It indicates the usefulness of the review in buying the product and not the product itself. It represented as the [the number of helpful votes/the total number of votes]. For eg. [2/3] means three reviewers have voted for the review of the product, out of which two were found helpful by buyers. So, for this particular example we can say that the percentage of helpfulness of the review was 66.66 percent. The above graph shows the distribution of helpful reviews among the total number

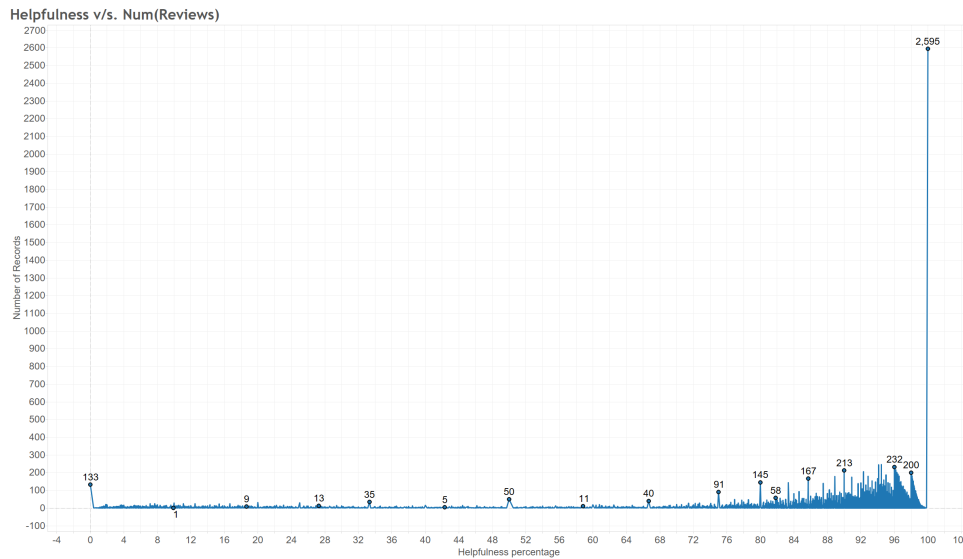


Figure 3.4: Number of helpful votes vs the number of reviews

of reviews. As can be seen, a major chunk of the total number of review was found to be helpful by the buyers. So this again just shows the growing importance of reviews in online shopping. The number of reviews being found helpful proves that reviews are indeed legit. A number of reviews have also been found less helpful, although this number is very small when compared to the number of reviews that were found helpful. This says that buyers take the reviews seriously and base their decisions on the ratings and reviews of products.

### 3.5 The average ratings across the number of years

As we see in figure 3.2, number of reviews started growing consistently after the year 2003. Although the growth has been consistent and gradual, the reviews showed a very high rise after the year 2013. So, for figure 3.5 to make sense, it is necessary to ignore the first few years where we do not have enough data to make conclusions. So as the above grows shows the average ratings have been fairly constant in the period between 2003 and 2007 with a slight increase in the next period between 2007 and 2015. These ratings not only help the buyers in making decisions but also the production companies change or enhance

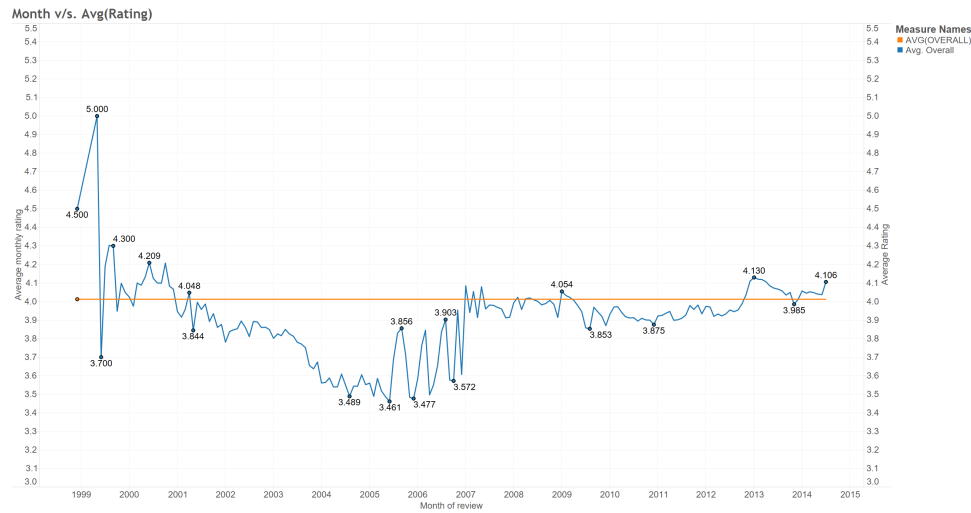


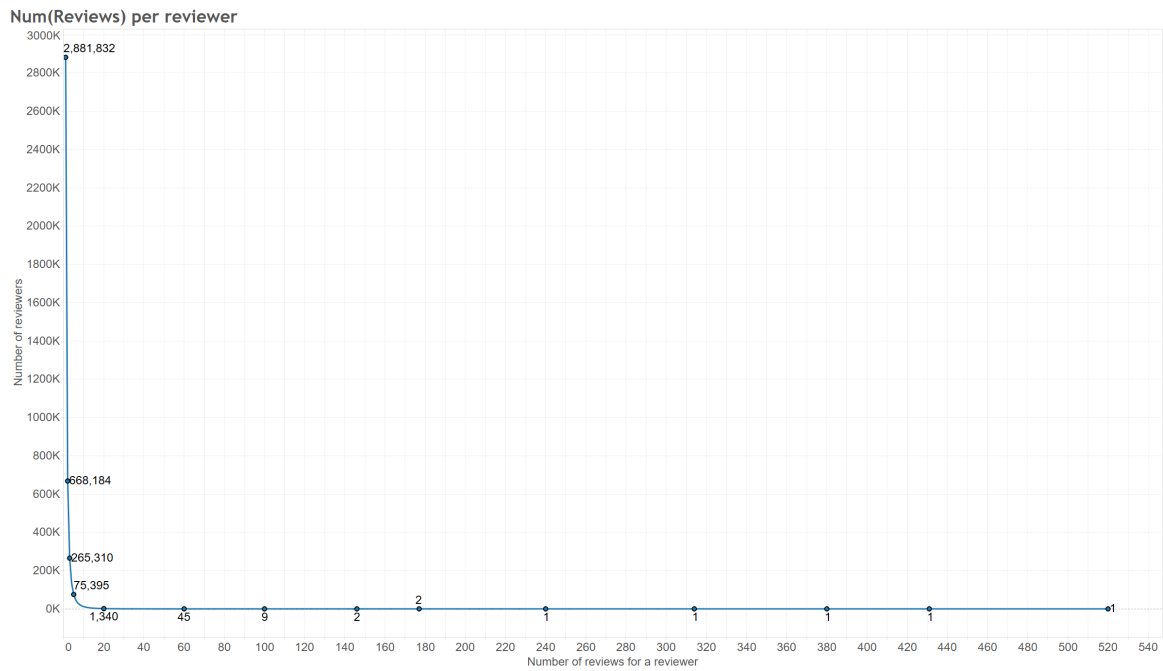
Figure 3.5: Month vs The average number of ratings

the product. This can be seen from the increase in average ratings.

### 3.6 Returning reviewers

As seen from figure 3.6, there is major portion of reviewers who have reviewed only once. Although, there is a considerable amount of reviewers who have reviewed more than once. Certain users are frequent customers and are used to reviewing the product after using it. But the number of reviews for returning reviewers goes as high as 500 which suggests possible indication of fake reviews because one user giving 500 reviews seems rather impossible.





As we clearly see in figure 3.7, there are more number of reviews for the products with a low price range. As the price range of the products increases, the number of reviews decreases. This clearly indicates that less number of users buy expensive products online. Almost 60 percent of the total reviews are for the products that have a price range \$0-50. This shows that users prefer to buy products which are not extremely expensive online.

### 3.8 The number of reviews and helpfulness

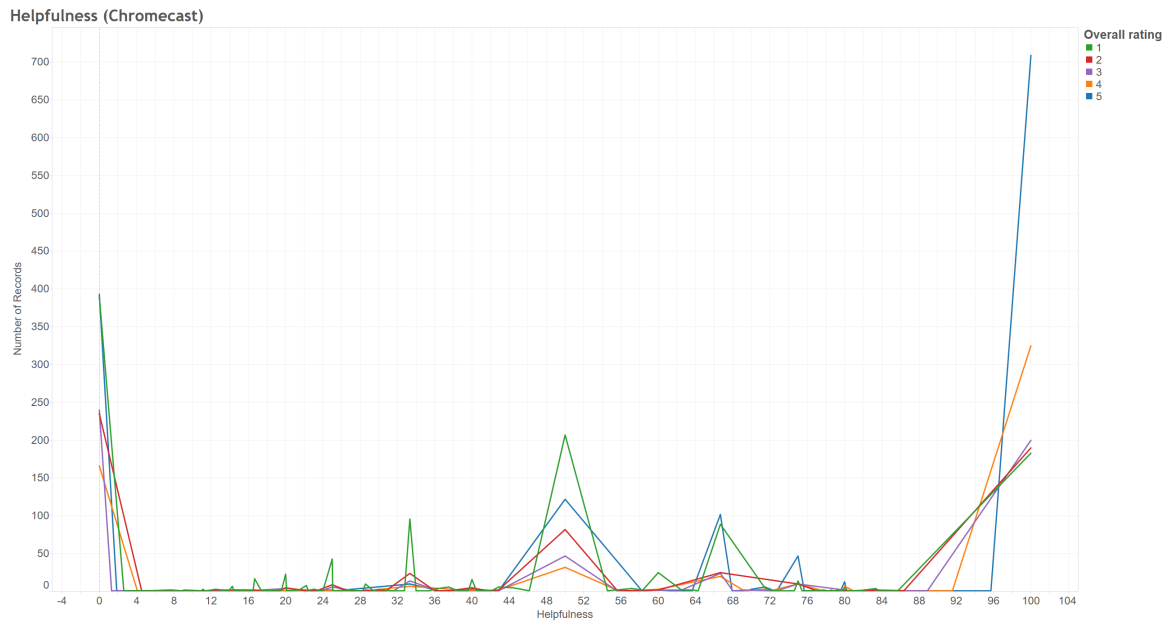


Figure 3.8: Helpfulness vs Rating(for Chromecast)

The figure 3.8 shows the relationship between rating of a review and the number of helpfulness votes a review has received. The individual ratings from 1 to 5 are represented with different colors as can be seen from the legend at the top right corner of the graph. As it can be clearly seen from the graph, the rating 1 has the highest percentage of 0 percent helpfulness. The percentage of helpfulness goes on increasing with the rating. The rating 5 shows the highest percentage of helpfulness. When the users are extremely happy with some product, reviewers give a 5 star. As seen from the graph, more than 75 percent of

the reviews were found helpful meaning that reviewers give an in detail review when they are extremely happy with the product which helps prospective buyers in making decisions about the product.

### 3.9 The distribution of ratings among infrequent and frequent reviewers

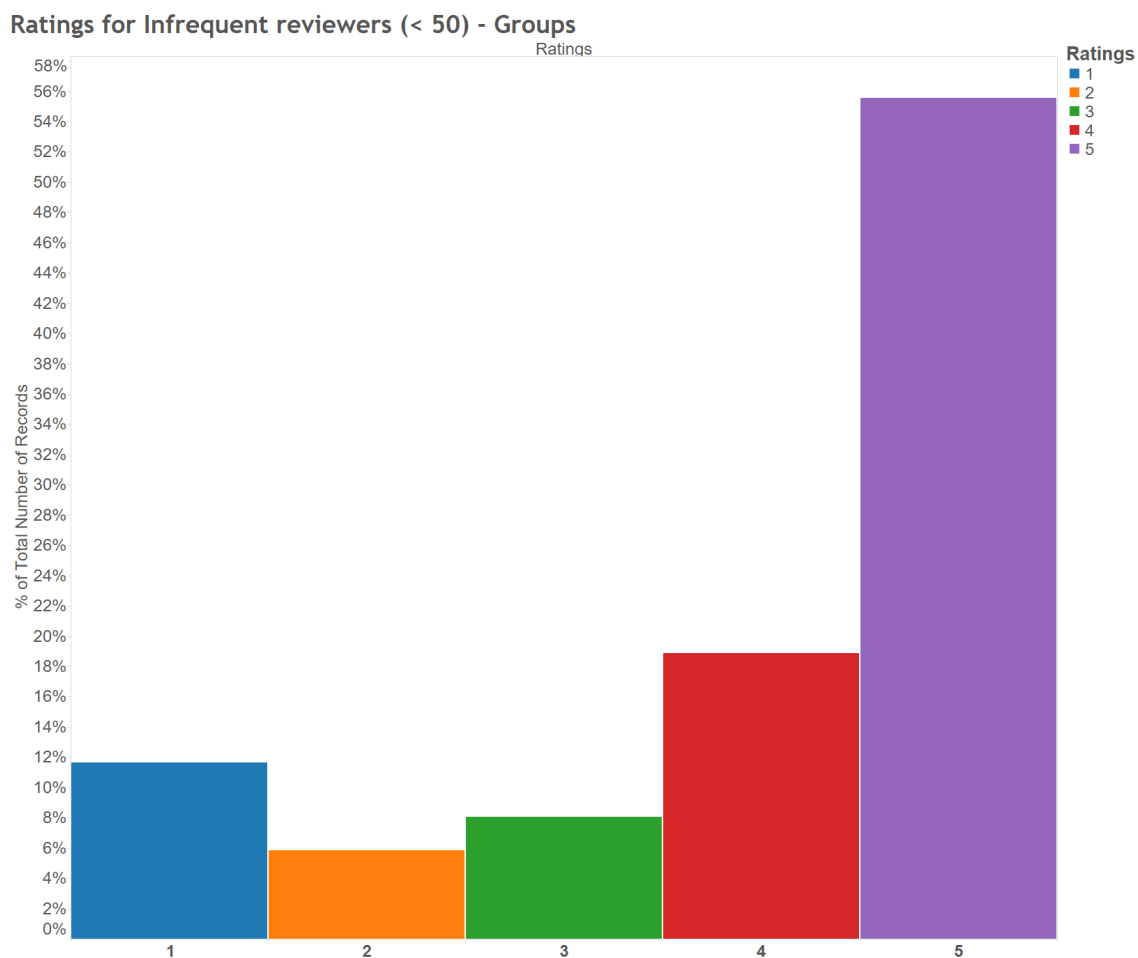


Figure 3.9: Rating vs Number of infrequent reviewers

For better understanding, let's assume that reviewers who have given less than fifty reviews are the infrequent reviewers. As we can see from 3.9, more than 75 percent of the

infrequent users give a star rating 4 or 5. The percentage of infrequent reviewers who have given low ratings of 1,2 or 3 is comparatively very less. The same trend can be observed for the ratings of frequent reviewers. But the percentage of frequent reviewers who give very high ratings of 4 or 5 goes as high as 86 percent as can be seen from figure 3.10. Only 8 percent of the frequent reviewers give 1 or 2 star rating whereas for infrequent reviewers, it is 20 percent.

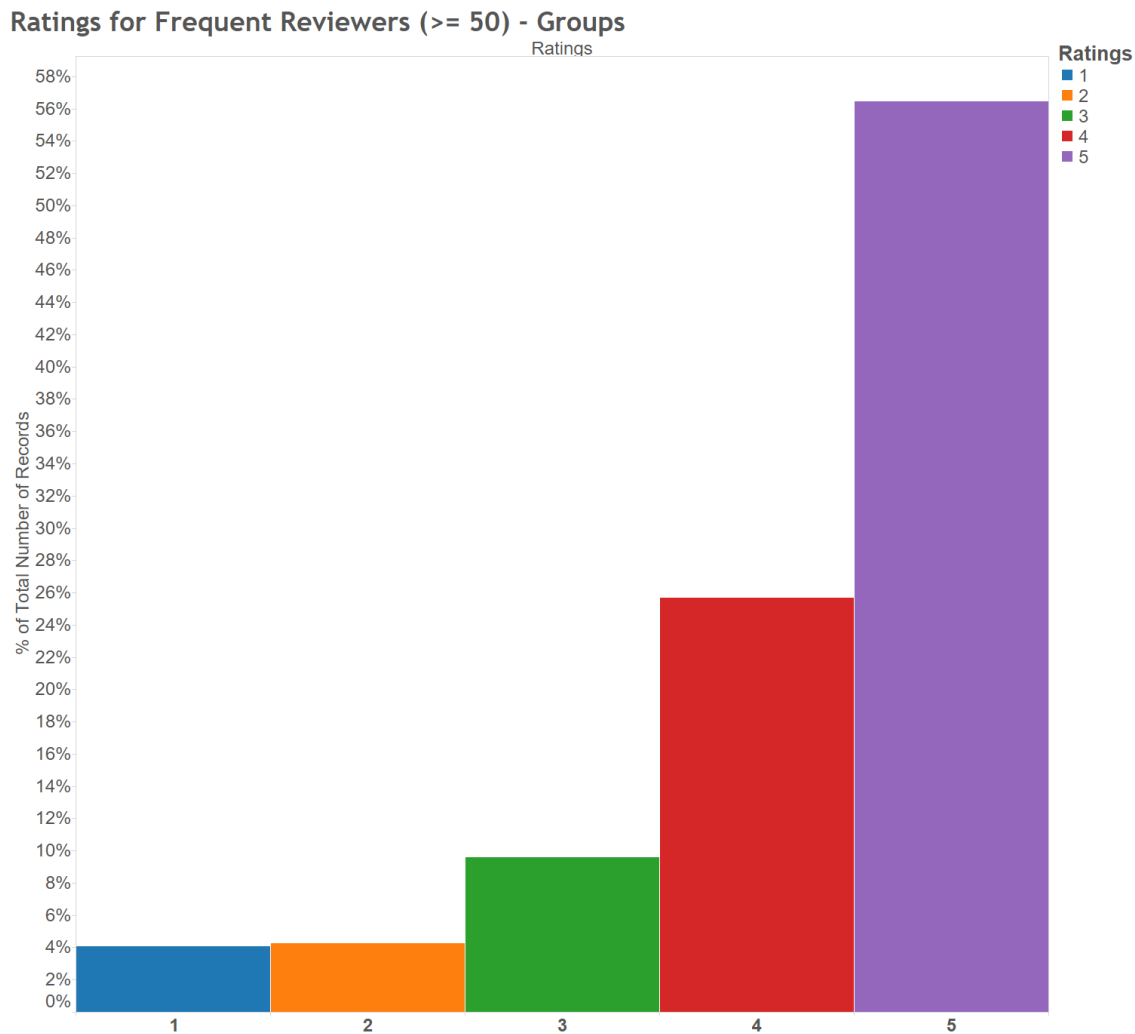


Figure 3.10: Rating vs Number of frequent reviewers

A side by side view as shown in figure 3.11 would give a clear understanding of this.

### Ratings v/s Frequency

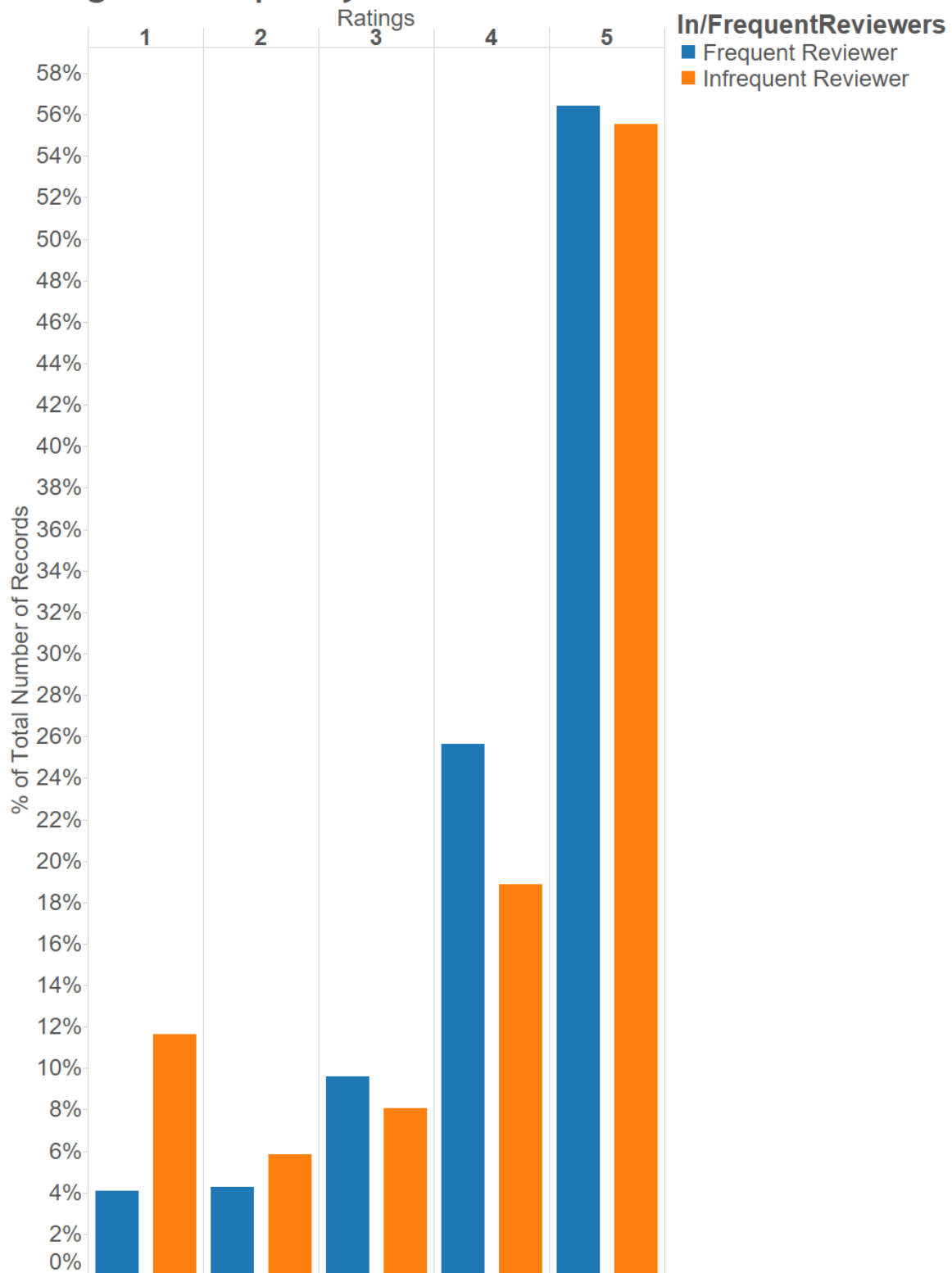


Figure 3.11: Rating vs Frequency of reviewers

### 3.10 The distribution of helpfulness percentage among infrequent and frequent reviewers

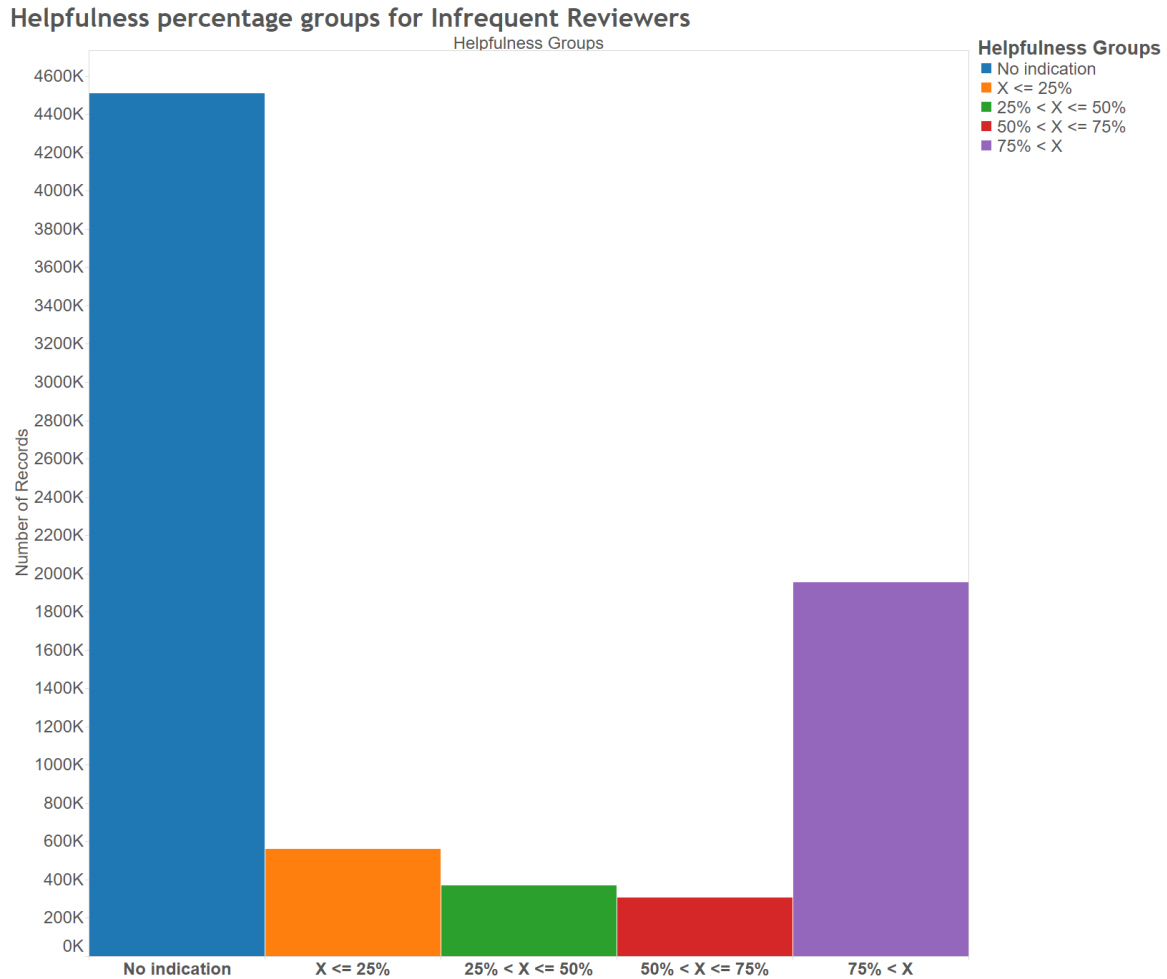


Figure 3.12: Helpfulness vs Number of infrequent reviewers

The distribution of the percentage of helpfulness among infrequent and frequent reviewers is shown in the figures 3.12 and 3.12 respectively. As we can clearly see, the reviews given by frequent reviews have more number of helpful or not helpful votes when compared with those of the frequent reviewers. The helpfulness percentage is divided into four groups - less than or equal to 25 percent helpful, 25 to 50 percent helpful, 50 to 75

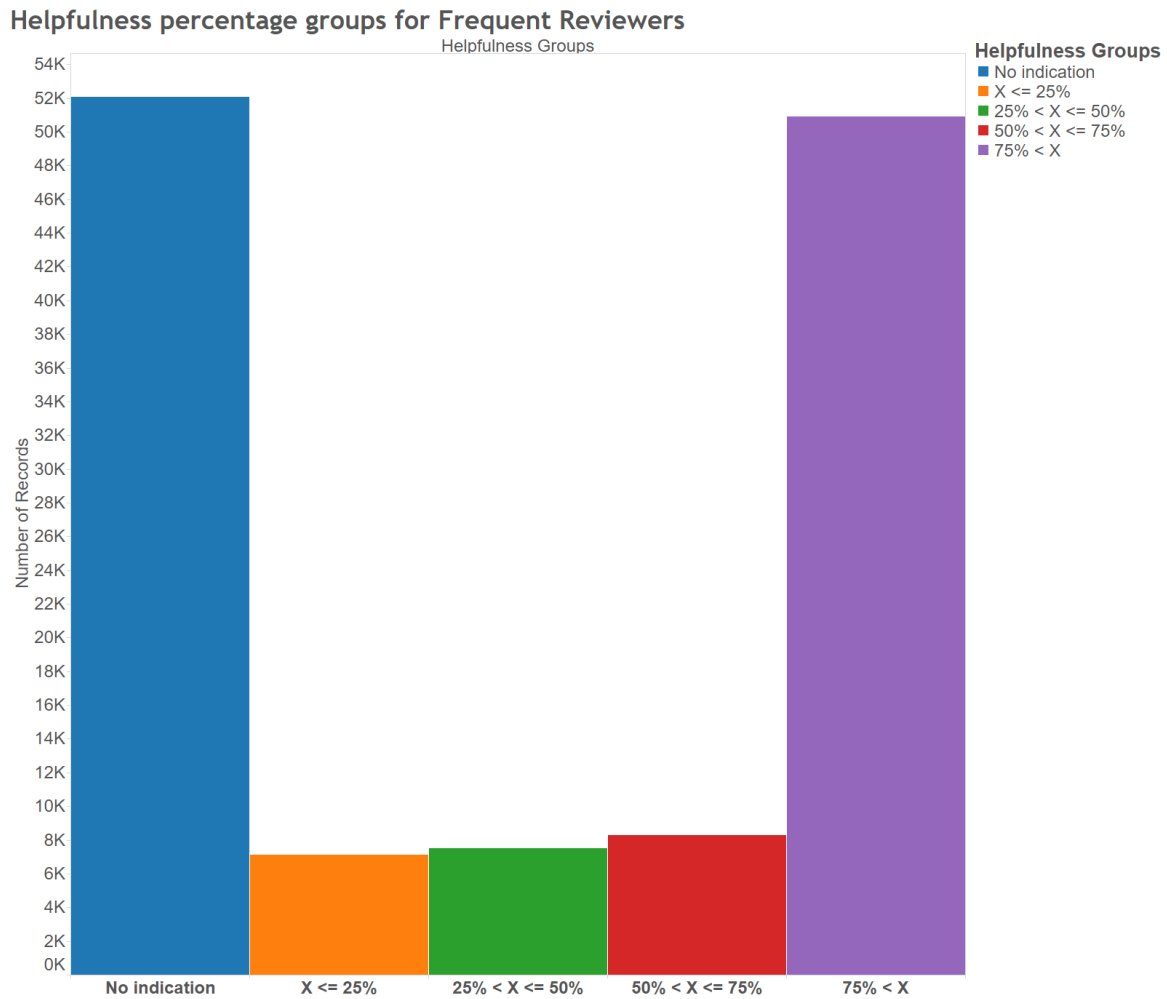


Figure 3.13: Helpfulness vs Number of frequent reviewers

percent helpful and more than 75 percent helpful. On comparing the trends of both frequent and infrequent reviewers, it is seen that the proportion of reviews that belonged to the more than 75 percent helpful group for frequent reviews was significantly higher as compared to insignificant reviewers. This indicates that the reviews given by frequent reviewers are usually more upvoted which means that they prove more helpful.

### 3.11 A comparison between ratings and helpfulness

Figure 3.14 shows the distribution of ratings for each of the helpfulness groups mentioned in section 3.10. The proportion of reviews with rating 5 are highest in the group with a helpfulness of more than 75%. The reviews with a rating 4 show a similar trend. But the proportion of reviews with ratings 1, 2, and 3 are the lowest in that group. This indicates that the reviewers find the reviews with extremely positive ratings more helpful which suggests that more of the potential buyers would consider voting the review as helpful.

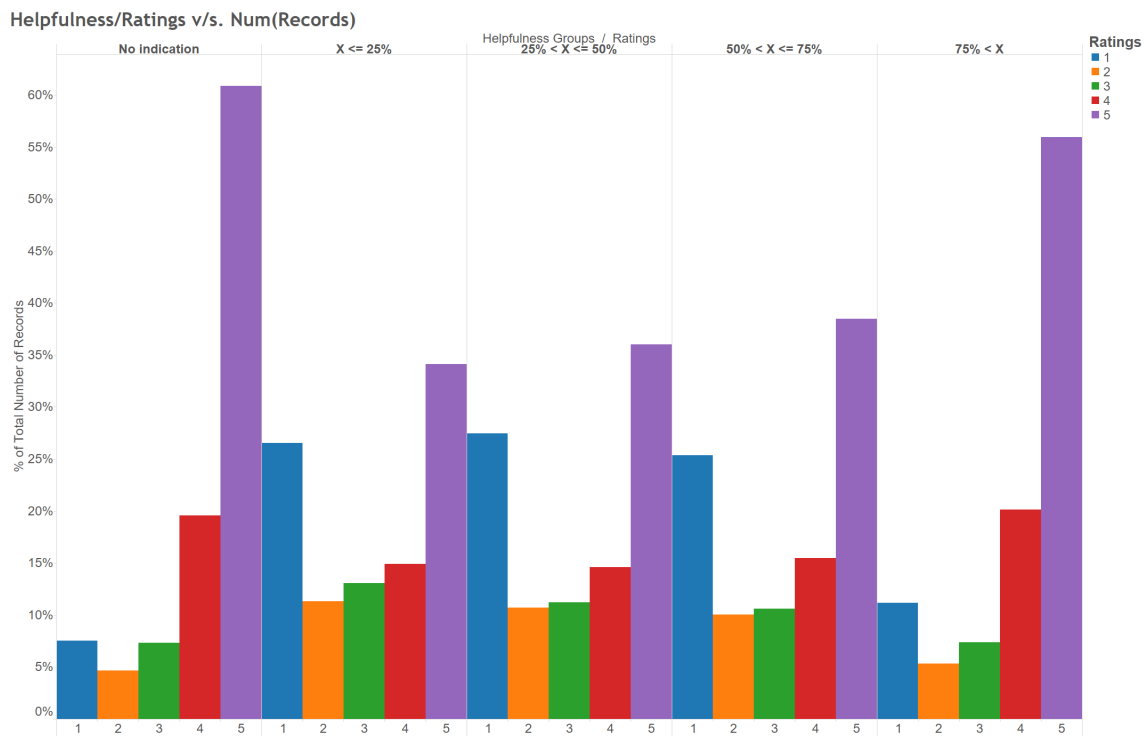


Figure 3.14: Helpfulness vs Rating

### 3.12 Common words used in reviews

Figure 3.15 shows the common words used in the reviews which have good ratings. The words with higher frequency are greater in font size and are darker in color. The font size of the word decreases and color becomes lighter as the frequency of the words in the reviews



Word cloud for good ratings



Figure 3.15: Word cloud for good ratings

decreases. As can be seen, the words great, good, well, easy, very are common in reviews with positive ratings(4 - 5 stars).

Figure 3.16 shows the common words used in the reviews which have bad ratings. The words with higher frequency are greater in font size and are darker in color. The font size of the word decreases and color becomes lighter as the frequency of the words in the reviews decreases. As can be seen, the word not is the most common word in reviews with negative ratings(1 - 2 stars) followed by other words like no, don't. The words great, good, easy which had a high frequency for reviews with positive ratings have a comparatively lower frequency in reviews with bad ratings.



Figure 3.16: Word cloud for bad ratings

### 3.13 The length of the review and the ratings

Figure 3.17 shows the number of characters present in each kind of review based on its ratings. The reviews with the rating 1 and 2 show maximum 29948 characters and 24091 characters respectively. The number of characters i.e. the length of the reviews is increasing as the rating increases from here on. This trend suggests that the reviewers/users tend to express themselves more or write more about the product when they are extremely happy with the product or if the product has a very high rating of 4 or 5.

## Review Text Length statistics v/s. Ratings

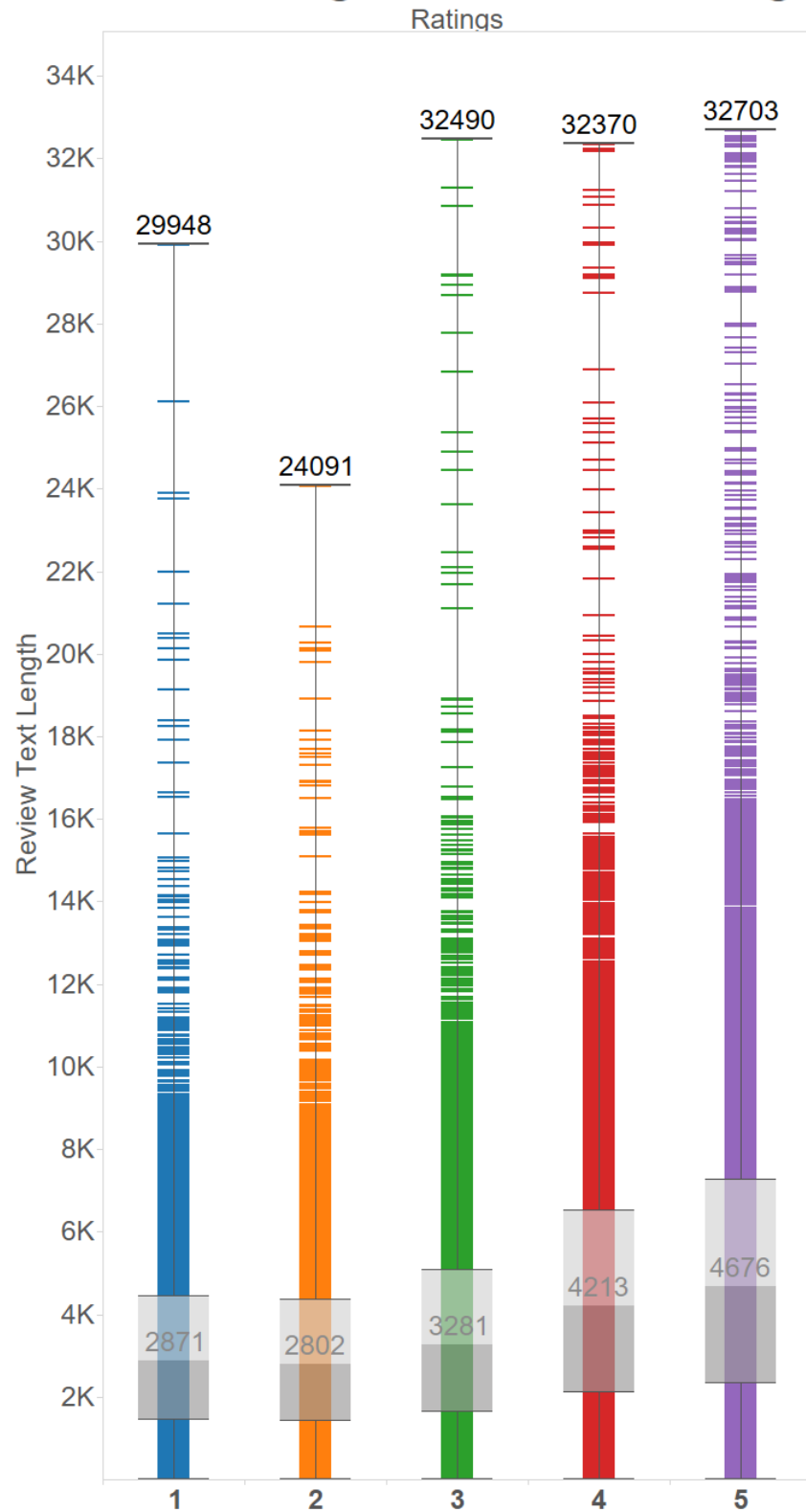


Figure 3.17: Review length vs Ratings

### 3.14 The length of the review and its helpfulness

#### Review Text Length statistics v/s. Helpfulness

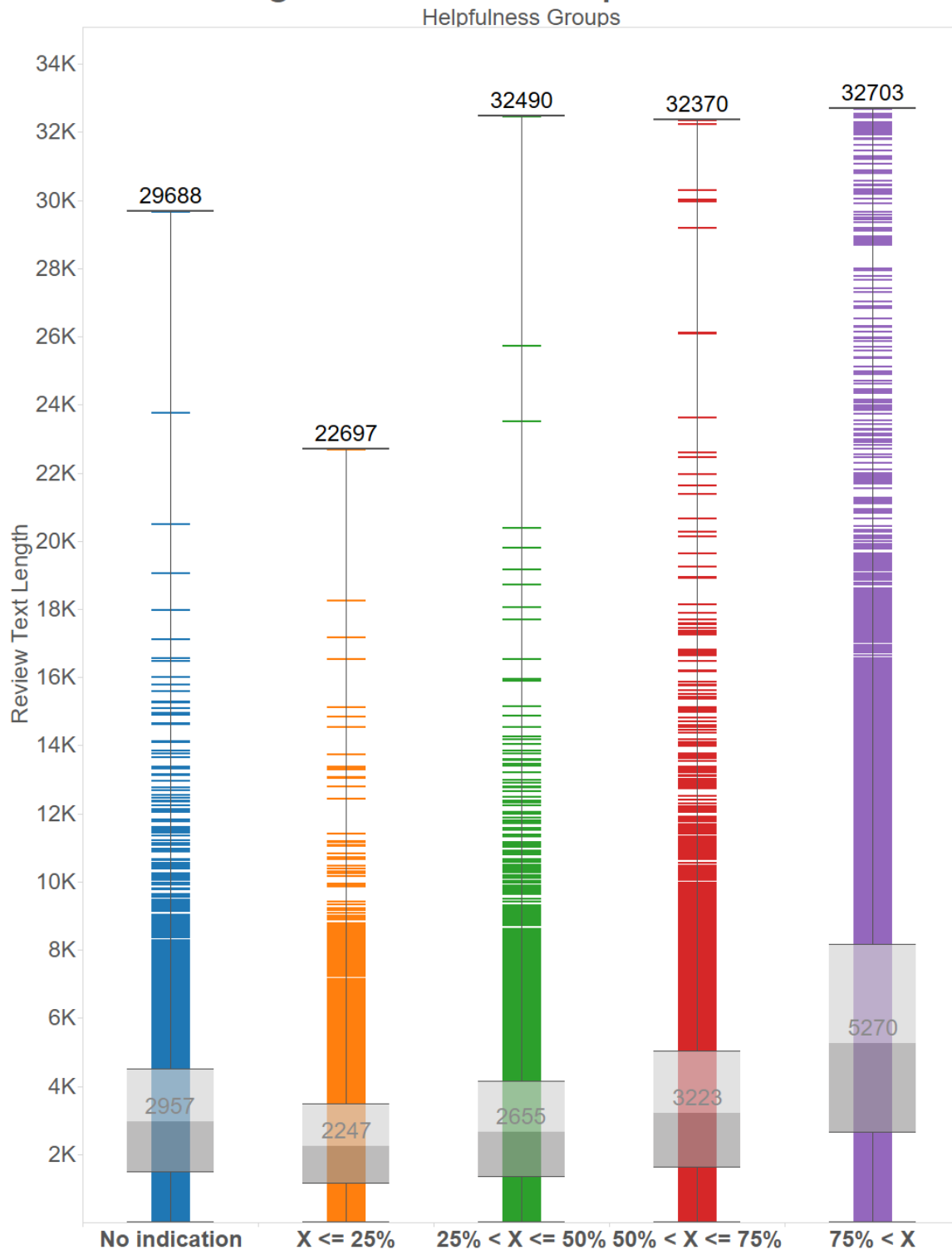


Figure 3.18: Review length vs Helpfulness

Figure 3.18 shows a similar trend to that of figure 3.17. Again, the helpfulness here is divided into four groups based on the percentage of people who found the review helpful - ( $\leq 25\%$ ,  $\leq 50\%$ ,  $\leq 75\%$  and  $\leq 100\%$ ). There is major chunk of reviewers who did not take part in the voting for helpfulness. As we see from the Figure 3.18, the number of characters i.e. the length of the review increases with the increase in helpfulness percentage. This suggests that longer reviews are indeed more helpful than the shorter ones.

### **3.15 The length of the review and the frequency of reviewers**

As can be clearly seen from figure 3.19, the maximum as well as median review length is very large for the reviews given by frequent reviewers as compared to that of the reviews given by infrequent reviewers. This suggests that users who actively give product reviews give a more detailed experience of the product and its pros and cons. Infrequent users are more likely to keep it short.

### Review Text Length statistics v/s. Frequency of Reviewer

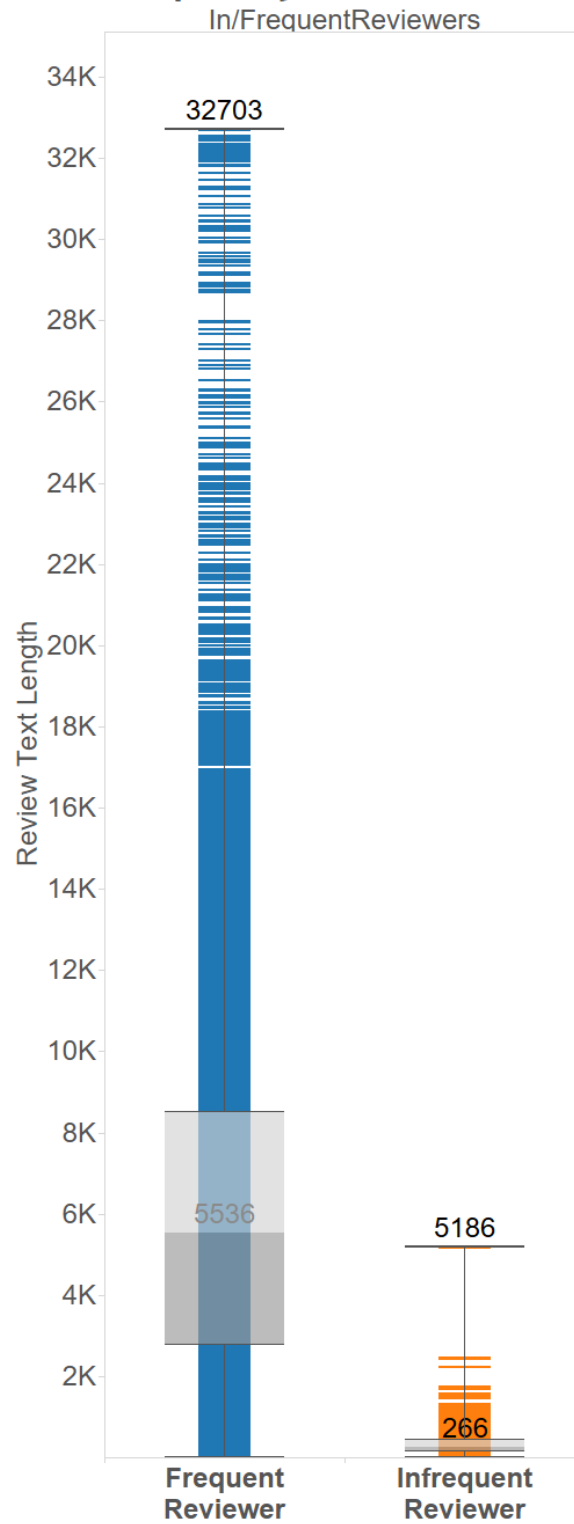


Figure 3.19: Review length vs Reviewer frequency

### 3.16 The growth of the length of the reviews over the years

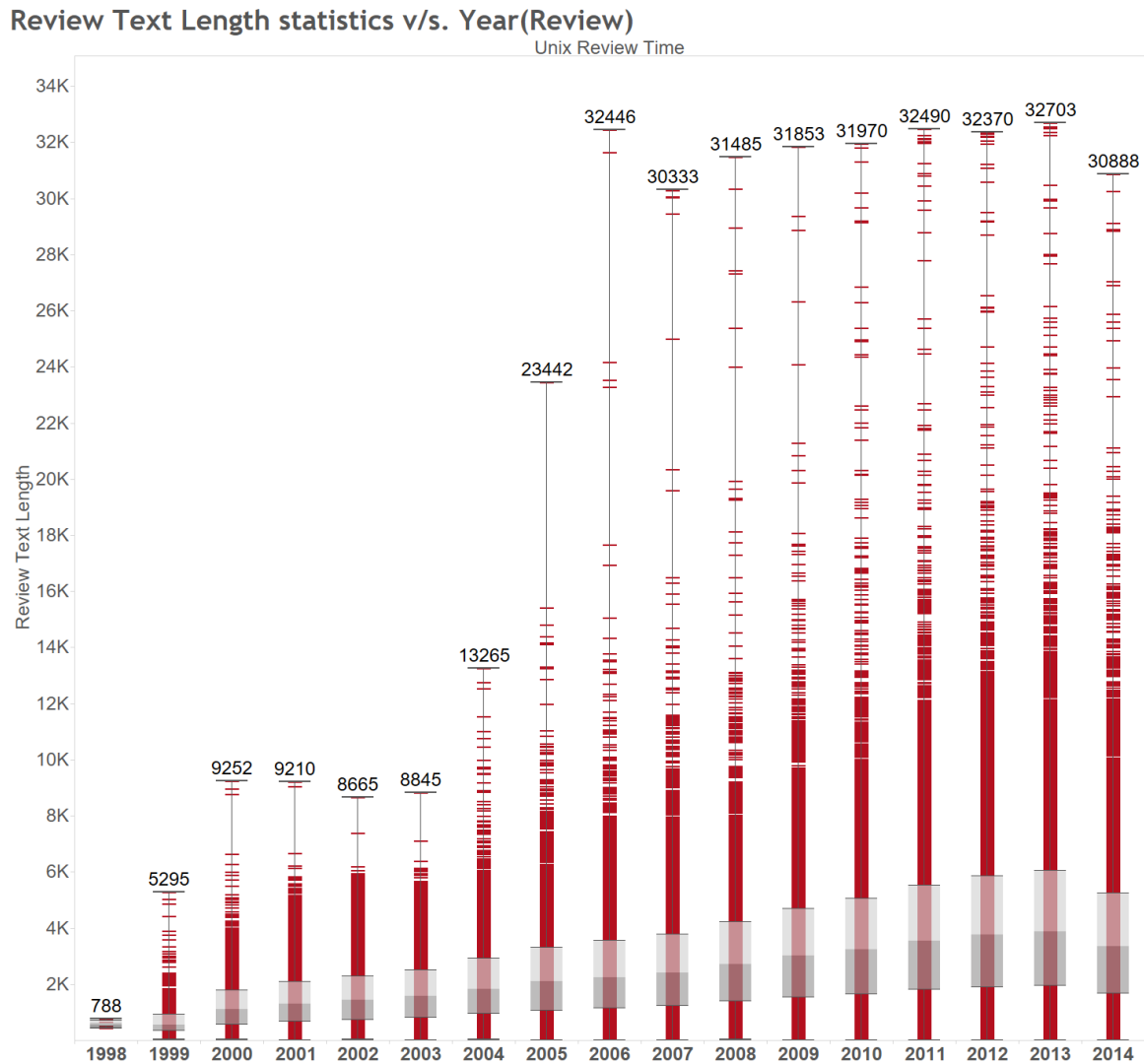


Figure 3.20: Review length vs time

As can be clearly seen from figure 3.20, there has been a gradual increase in the length of the review along with the span of time. If we consider a span of 5 years, the difference between the maximum length of reviews is quite significant. The median length of reviews is growing every year which is suggestive of two things - the increase in the number of reviews every year and the also, the increase in the size of the reviews over the years. This

again is indicative of the growing importance of writing reviews for products in the field of online shopping.

### 3.17 The review length and the price of the product

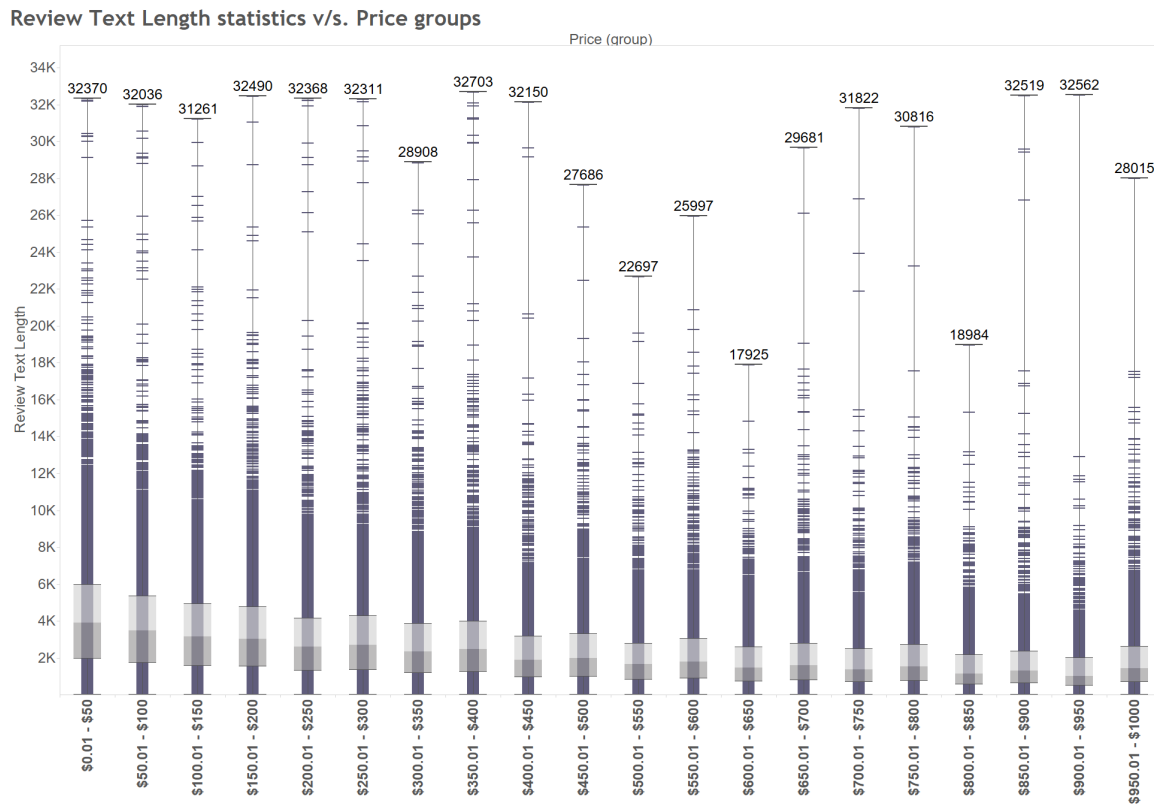


Figure 3.21: Review length vs Price groups

In the figure 3.21, the prices of the products have been grouped together for better visualization purposes. The products with the lowest price range have the highest median length of the reviews. This length goes on decreasing as the price range of the product increases. This trend is related to that of the number of users buying more low price products online than the high price products. Since there are more number of reviews for products with a lower price range, it may have been just by chance that these products have reviews with a



greater length than the higher range products.

### 3.18 The price of the product vs the frequency of the reviewers

Frequency(Reviewers) v/s. Price groups

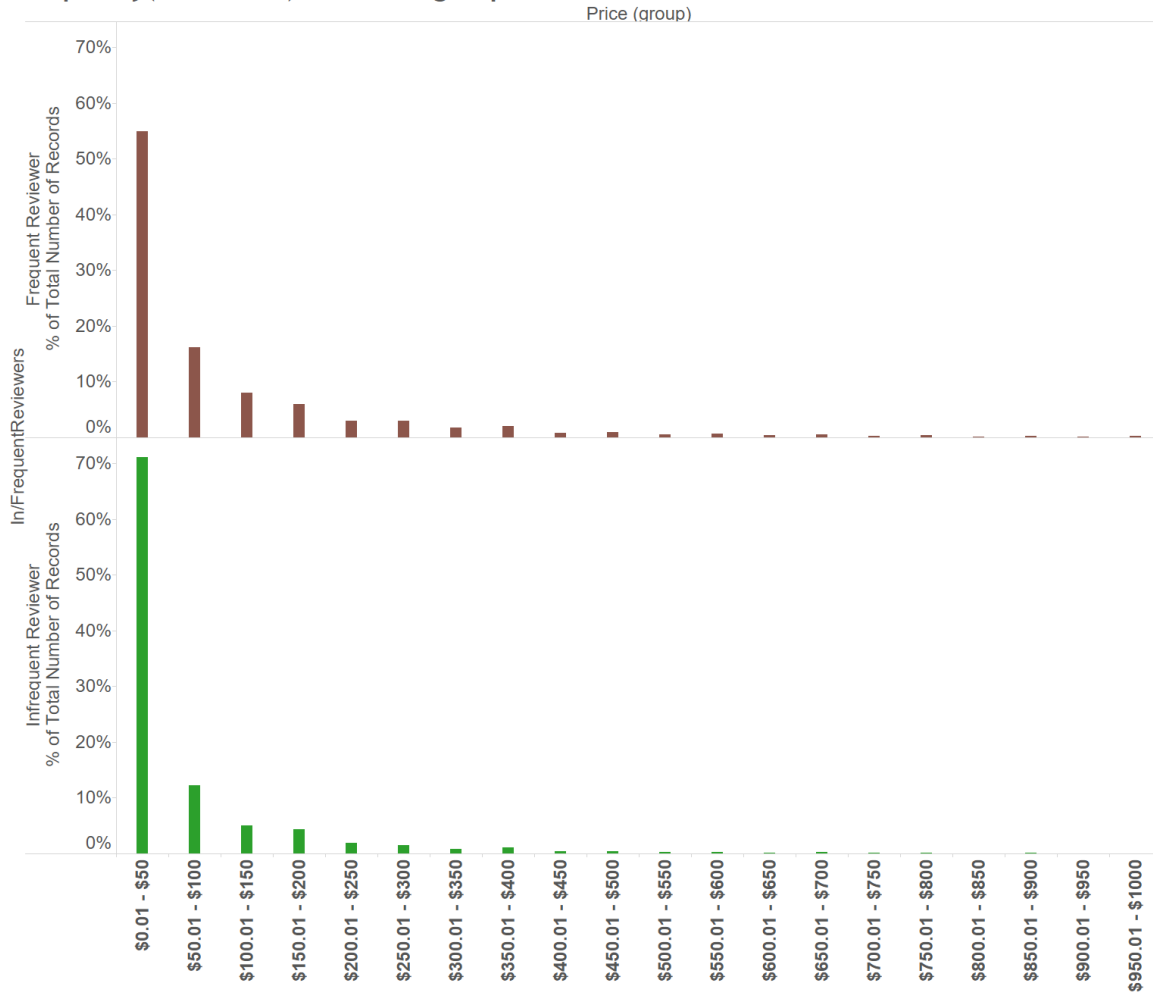


Figure 3.22: Price groups vs Frequency of reviewers

As seen from figure 3.22, out of the total number of frequent as well as infrequent reviewers, major chunk of the portion is for the products whose price ranges from \$0 to \$50. 55% of the total frequent users have given reviews for the products with the price range of \$0 to

\$50. 70% of the total infrequent users have given reviews for the same. This percentage goes on decreasing with the increase in price range.

### 3.19 The price of the product vs Helpfulness

Helpfulness v/s. Price group

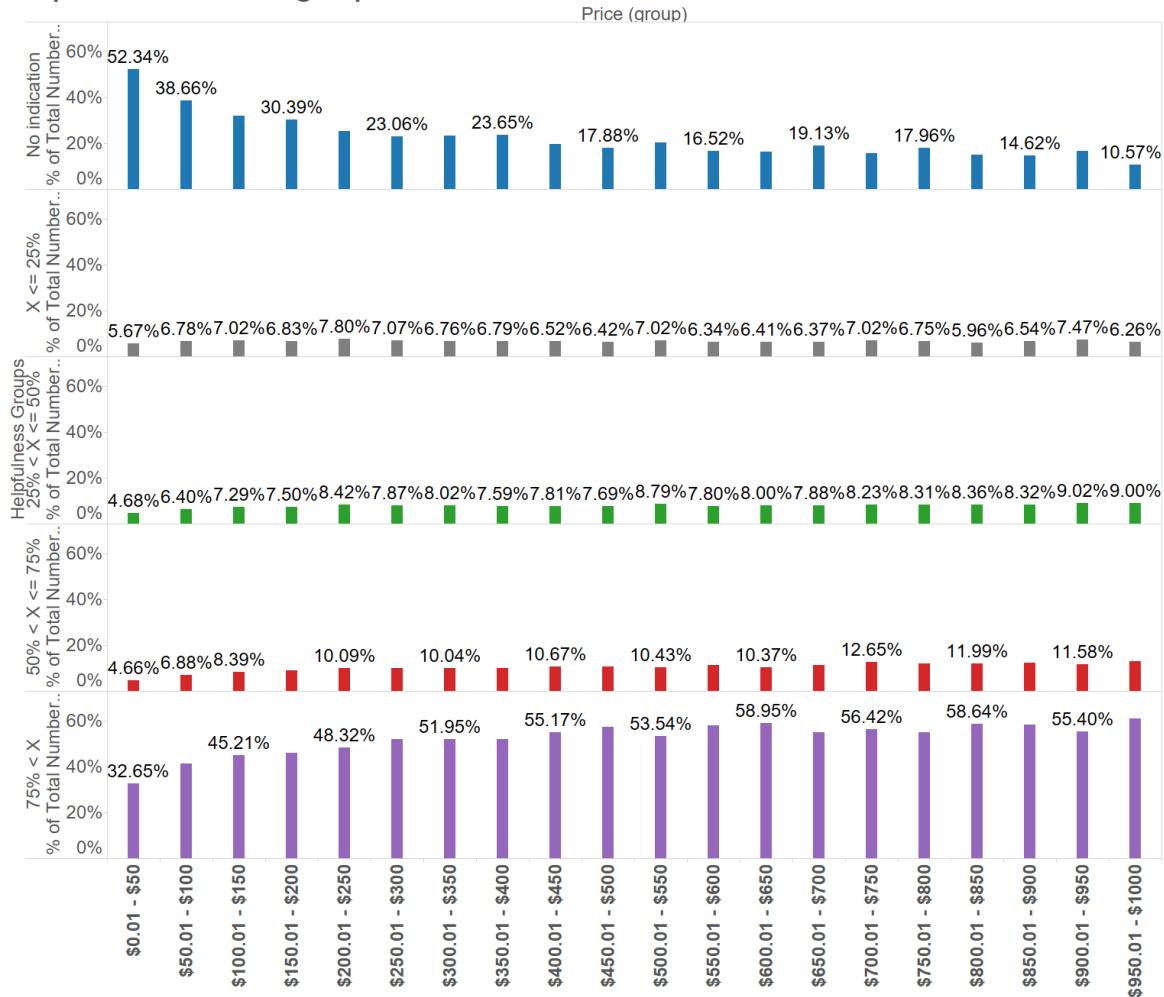


Figure 3.23: Price groups vs Helpfulness percentage

For analysis purposes, the helpfulness percentage and the price range of products have been divided into the same groups as mentioned earlier. As can be seen from figure 3.23, the helpfulness percentage increases with the increase in price range. This shows that people

especially pay more attention to reviews when they are buying high end products than the reviews of the cheaper products.

## Chapter 4

### Conclusions

A number of useful insights have been derived from the visualizations and analysis which may help in improving the existing review system of Amazon to make it better for the customers as well as the sellers. While some of them are pretty obvious, the rest of them are instinctive and can help build a better review system. The prominent features considered while analyzing the review system are the timestamp of the review, the length of the reviews, the frequency of the reviewers, the number of reviews, the price of the product, the helpfulness votes given to the review and the rating given to the product.

There has been a consistent growth in the number of reviews as well as the reviewers over the time span considered in the analysis. This obviously shows that more and more buyers are relying on the reviews of the customers who have already bought the product and are also writing reviews. This implies that the review system is an important feature of the Amazon online shopping system and improving the system will definitely enhance the shopping experience.

Most reviewers prefer to write concise reviews as opposed to very large or extremely short reviews. However, lengthy reviews are considered more helpful overall. Another interesting trend that was observed is that 5-star reviews tend to be more lengthy. This suggests that the detailed description of reviews should be promoted which can be done by possibly asking product specific questions when asking for review. For e.g. when writing reviews for a phone, the review system may ask for the battery consumption, camera quality and so on.

A vast majority of the total reviewers have written less than 5 reviews. Any users giving

more than 50 reviews (referred to as frequent reviewers) seems to be suspiciously high and might hint towards fake reviews. There was one user who posted more than 500 reviews. It is also interesting to note that the frequent reviewers give more reviews with a higher star rating as compared to the infrequent reviewers. Whereas the infrequent reviewers give more reviews with a lower star rating as compared to the frequent reviewers. Also, the reviews given by infrequent reviewers are more helpful than those by the frequent reviewers. These points might also suggest that the frequent reviewers are getting paid to write fake good reviews.

Many users avoid buying expensive items from Amazon. Relatively low priced products have a higher number of reviews. Infrequent reviewers have a relatively higher number of reviews for lower priced items as compared to the frequent reviewers who have a higher percentage of reviews for expensive items. However, higher priced items have more helpful reviews. This might suggest that frequent customers don't hesitate to buy expensive items and that not all frequent reviewers would be fake.

Higher star rating (4 to 5 stars) is pretty common across all the reviewers, and have a overall high helpfulness rating as well. 1-star rating is also quite frequent, but are relatively less helpful. This trend is more visible for infrequent reviewers. This might point that having a binary rating would be helpful, but having more precision in rating makes comparison between products easier.

These are some of the interesting trends that were uncovered throughout the course of this project. Further analysis and supporting user studies may be helpful before updating the review system according to the observations.

## **Chapter 5**

### **Future Work**

The analysis done in this project is limited to electronic products only. Something similar can be done for other categories as well or for all the categories together. Another way to look at it would be to have a finer grained analysis for specific product categories, like cell phones, T.V., etc.

Another possible direction this project could grow in would be using this analysis as a starting point for detection of fake reviews based on the reviewers statistics, length of the review, and duplicates.

All of the analyses done in this project will be helpful in understanding the review system, thus paving a way for a better review system and hence improving the overall online shopping experience.

# Bibliography

- [1] Tableau. <http://interworks.co.uk/business-intelligence/why-tableau/>, 2016. [Online; accessed 4-March-2016].
- [2] A. Bhatt, A. Patel, H. Chheda, and K. Gawande. Amazon Review Classification and Sentiment Analysis. *International Journal of Computer Science and Information Technologies*, 6(6):5107–5110, 2015.
- [3] X. Fang and J. Zhan. Sentiment analysis using product review data. *Journal of Big Data*, 2:1–14, 2015.
- [4] J. McAuley. Amazon product data. <http://jmcauley.ucsd.edu/data/amazon/>, 2016. [Online; accessed 4-March-2016].
- [5] J. McAuley, R. Pandey, and J. Leskovec. Inferring Networks of Substitutable and Complementary Products. *Knowledge Discovery and Data Mining*, 2015.
- [6] J. McAuley, C. Targett, and A. Van Den Hengel. Image-based Recommendations on Styles and Substitutes. *SIGIR*, 2015.