Group 1: CM Knox, Ronald Manipol, Traci Garrett, Josh Sniderman, Shahana Shaikh

The Merge:

Analyzing Amazon Kindle Reviews over the span of 7 years.

Introduction:

Amazon first released Kindle in 2007 - changing the way books were purchased and read. While it wasn't the first e-reader ever made, it has proven to be the most successful. As one of Amazon's top-selling items, it allows you to access books, magazines, newspapers, and more. This project focuses on Amazon reviews of books read by Kindle users. While there are limited dimensions in the data set, it will become clear that there is a plethora of analysis to be done on just reviews alone.

Does the number of reviews increase over time?
To find any increase in reviews, it is important to first figure out the time frame that we're working with and the count of reviews. Once we understand the years that the data set provides us, we can plot out the count of reviews over years and track any trends/patterns we may see. The increase/decrease in the number of reviews would tell us either that more people are using Kindle or more people are starting to understand the concept of leaving reviews.

Are there any months in which the reviews increase?
Any sort of increase in a time period triggers us to find the time frame we need. Months with the most reviews can give us many insights. One of the main insights would allow us a marketing edge. By finding the months in which reviews increase it is safe to assume that consumer usage of kindle and/or reading books increases. Either way, if used by a brand like Amazon it would allow for more ad campaign maximization pushes during the popular months as compared to the months with a lower count of reviews.

What is the percentage of reviews based on ratings?
For this project, we want to see the overall trend of ratings as a percentage of the reviews each rating makes up. This pattern shows the trend of consumer ratings, and whether or not they can be influenced by other ratings of similar standing.

Does the number of reviews affect the rating of a book?
When purchasing any product from Amazon, it is common practice to look at the reviews of the product to ensure quality. We would assume the same when purchasing e-books for Kindle. The more reviews a product has the more trustworthy or worth purchasing a product is. However, in some instances, more reviews could also mean that there is a consensus among consumers about the lack of quality in said product.

Methodology:

We started off by pulling the basic statistics to understand the size of the dataset we have.

Kindle was released in 2007, which is why we see the sudden increase in reviews from 2006 to 2007. We assume that the reviews made before that were probably during the testing phase of the product. There was also a drop in reviews from 2013 to 2014 - kindle released their 7th generation model. which allows us to conclude that with user purchases of the new model, it means that consumers would need some time to use it before starting to leave reviews.

The time the reviews were given was presented in the data using a Unix timestamp, this needed to be converted to a readable date so that it could be used for further analysis. To do this we utilized the ‘DateTime’ function within pandas, which gave us a year-month-day format. From there we created new columns to separate the month and year for each review. Once we had the month as a singular value, we created a new data frame that showed the month and a count for the number of reviews during that month through the years as a single sum. This allowed us to see the vast majority of reviews were given April through June, and the least amount of reviews from August through September.

By finding the percentage of reviews per level of rating, we could see the consumer activity. Initially, there were a lot of 5-star reviews and eventually found that certain users would only leave 5-star reviews. To get better data, we drilled it down further by getting rid of the users that only left 5-star reviews to see how the trends of other ratings were doing. We concluded that a large amount of 5-star reviews was skewing the data. By getting rid of repeat reviewers, we were able to analyze the changes in reviews that were affecting other attributes.

Limitations of Data and Analysis:

The dataset we used for the project was one of the first we came across, and though the CSV has almost one million rows of data which include dates, text from reviews, book ratings, and how helpful a review was, there was still much to be desired which provided our first limitation.

This dataset was missing key information such as user demographics – age, sex, and location, and even basic information for the product such as titles, author, price, and genre. We continued to search for other datasets that could be merged and provide a bigger picture, but we came up empty-handed when the other sets we were able to source were rejected due to varying reasons ranging from the realization that our main dataset was based around Kindle books, to fees being charged for access on various websites.

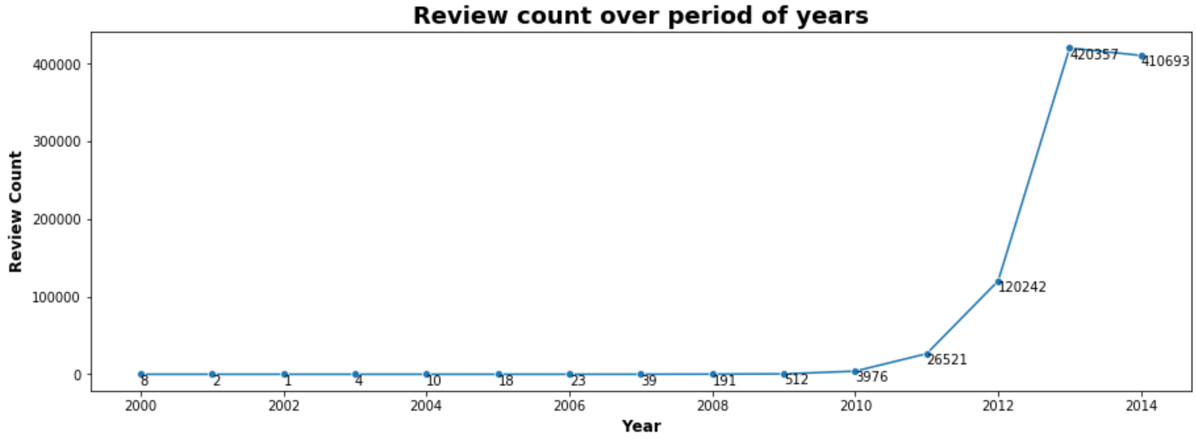
A second glaring limitation that rendered some of our efforts temporarily unsolvable, is the group’s collective coding skill level. We discussed utilizing the text in each review to find a correlation between the words used and the rating given by that specific reviewer. That info would in turn be used in a word cloud graphic to show the most used words based on the rating – 1 through 5 stars. This would have been a success, however, each rating’s most used words were generic ones such as “book”, and “series”.

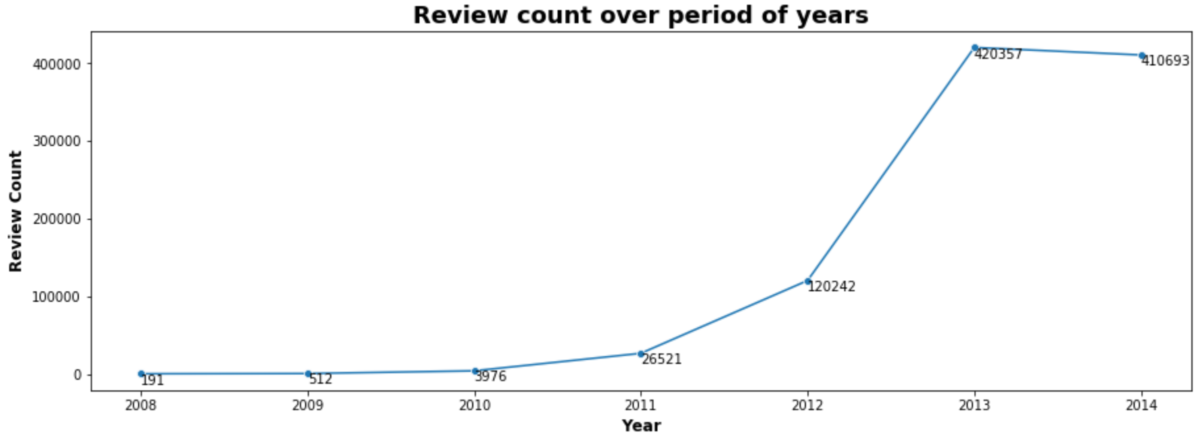
A second challenge also relating to the first limitation, is instead of book titles, the data features ASIN numbers, which are Amazon’s distinct codes used for identifying Kindle books. These numbers would have to be changed to a different format, transcribed to an ISBN, then translated to the actual book title which is a task we could not complete.

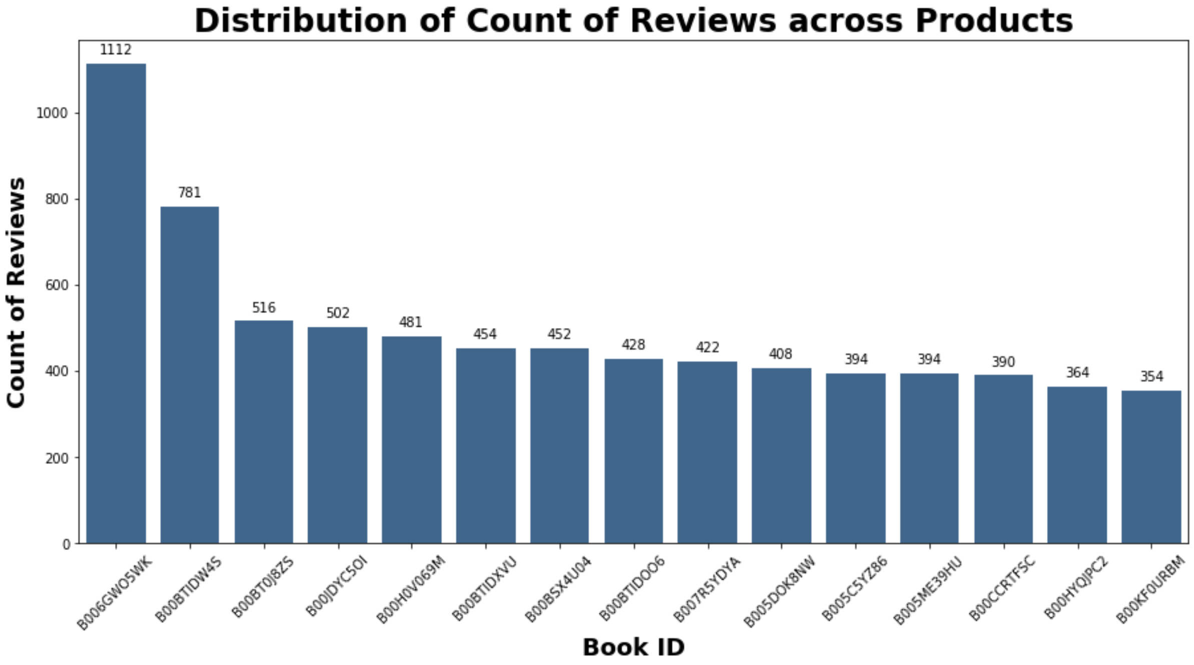
Due to constraints with the project’s deadline, we were unable to spend the time necessary to work through solutions and find workarounds for these issues and others. Time will also provide an indirect window for the growth of skill sets used to complete various tasks in a quicker and more efficient timeframe.

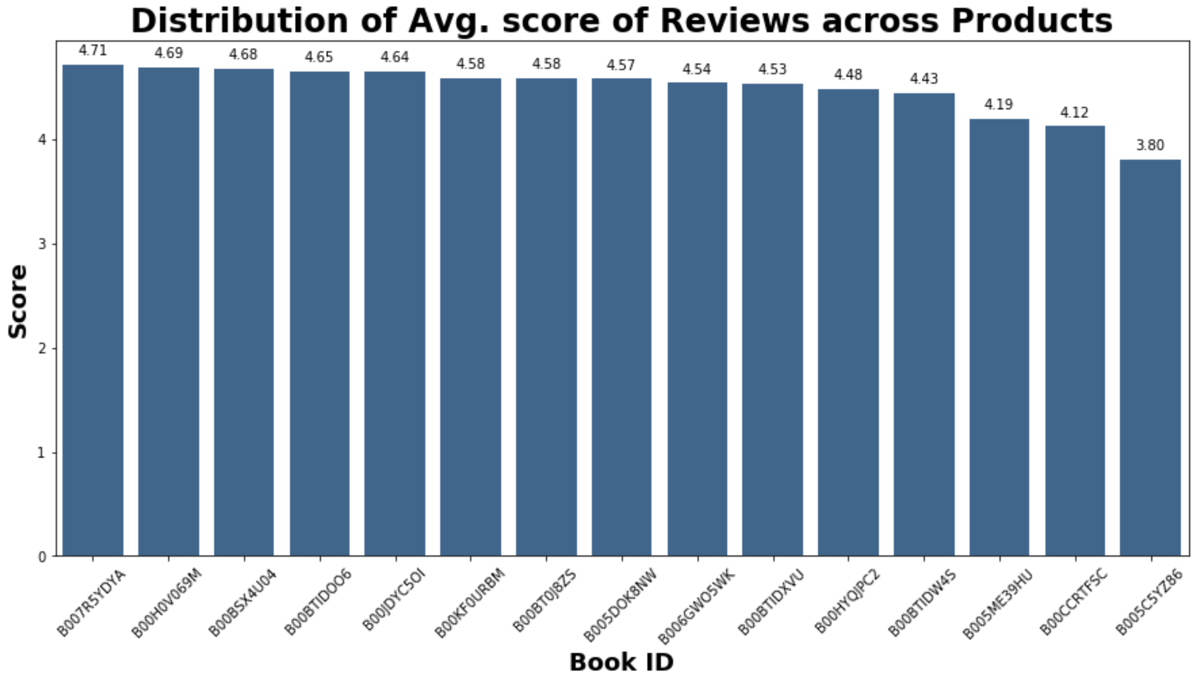
Conclusion:

We started with the knowledge that our data set was limited to just book reviews. The expectations weren't very high in terms of results. However, as we dove into the data, we saw different trends and patterns that allowed us a deep analysis of book reviews. We could see how the number of reviews increase over time, revenue increases due to reviews in certain months, and how the number of reviews affect the rating. This lays a strong foundation that we can eventually build upon.

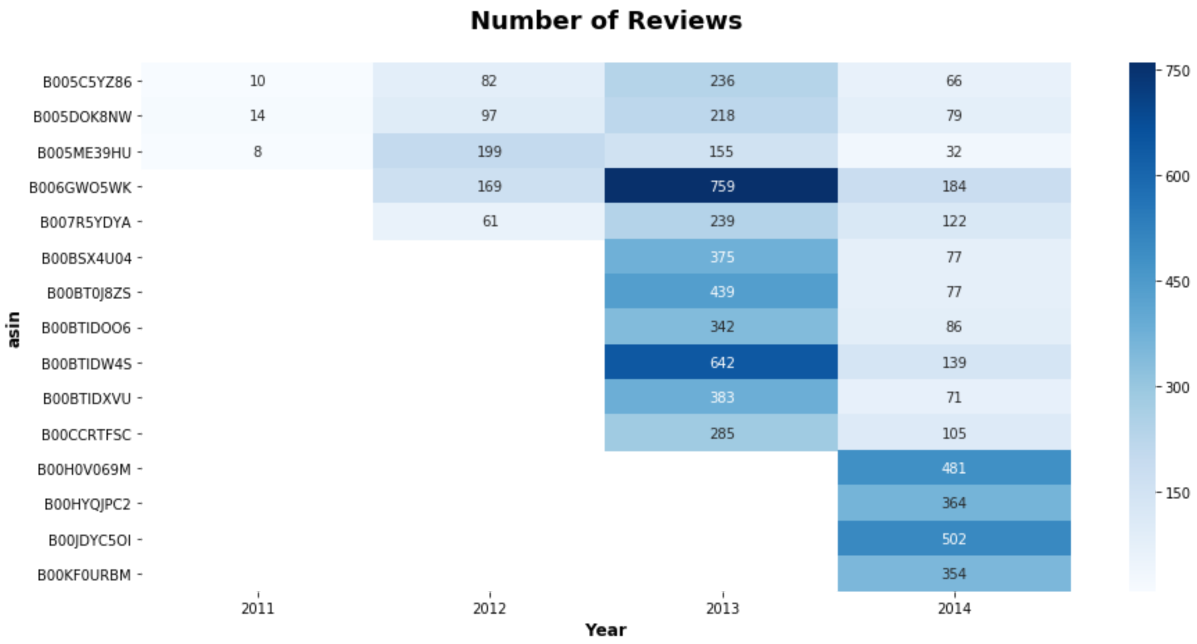

Review count over period of years

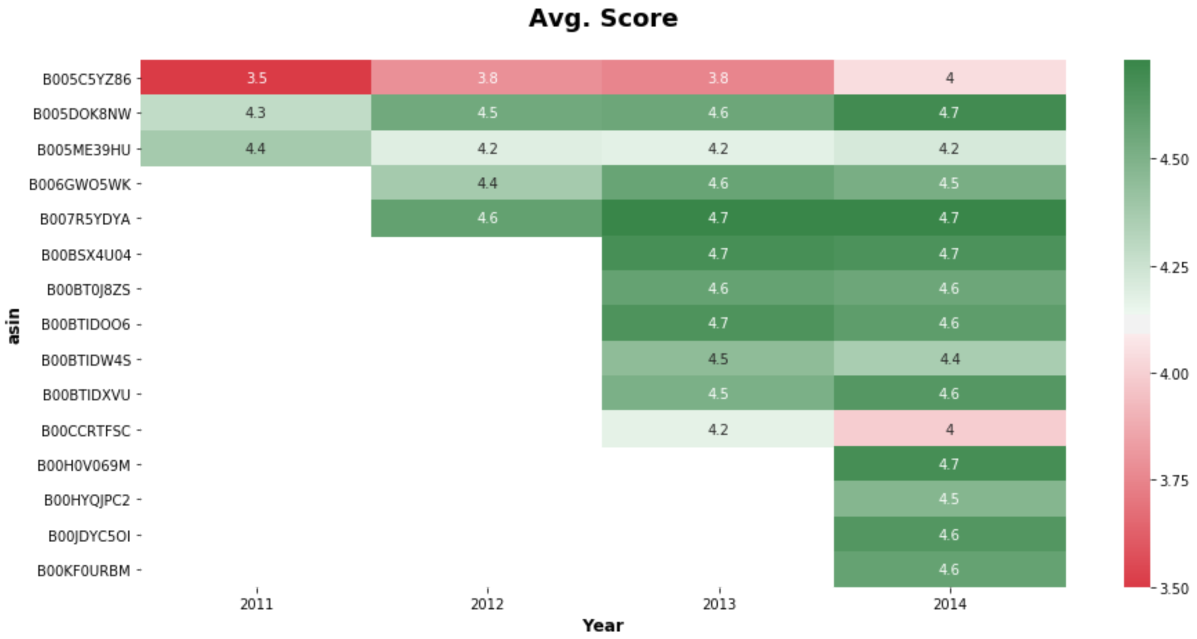

Review count over period of years


Distribution of Count of Reviews across Products
Count of Reviews
Book ID
0
200
400
600
800
1000
1112
781
516
502
481
454
452
428
422
408
394
394
390
364
354
B006GWO5WK
B00BTIDW4S
B00BT0J8ZS
B00JDYC5OI
B00H0V069M
B00BTIDXVU
B00BSX4U04
B00BTIDOO6
B007R5YDYA
B005DOK8NW
B005C5YZ86
B005ME39HU
B00CCRTFSC
B00HYQJPC2
B00KF0URBM

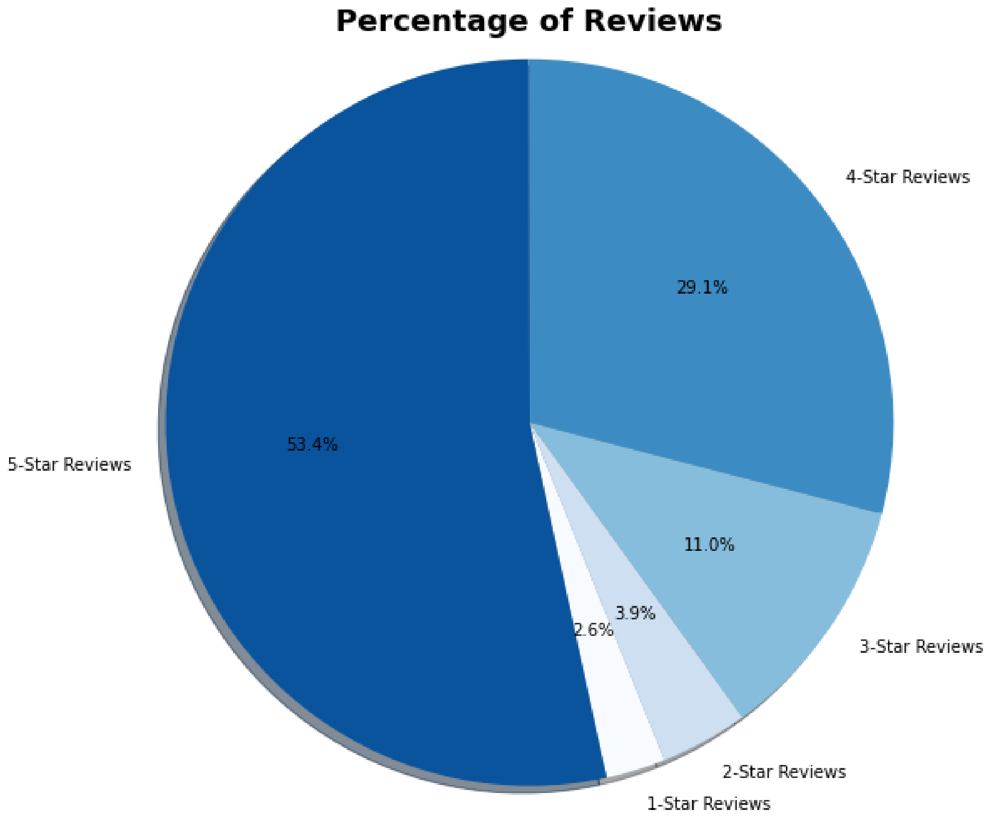

Distribution of Avg. score of Reviews across Products
Score
Book ID
0
1
2
3
4
4.71
4.69
4.68
4.65
4.64
4.58
4.58
4.57
4.54
4.53
4.48
4.43
4.19
4.12
3.80
B007R5YDYA
B00H0V069M
B00BSX4U04
B00BTIDOO6
B00JDYC5OI
B00KF0URBM
B00BT0J8ZS
B005DOK8NW
B006GWO5WK
B00BTIDXVU
B00HYQJPC2
B00BTIDW4S
B005ME39HU
B00CCRTFSC
B005C5YZ86

Group 1: CM Knox, Ronald Manipol, Traci Garrett, Josh Sniderman, Shahana Shaikh

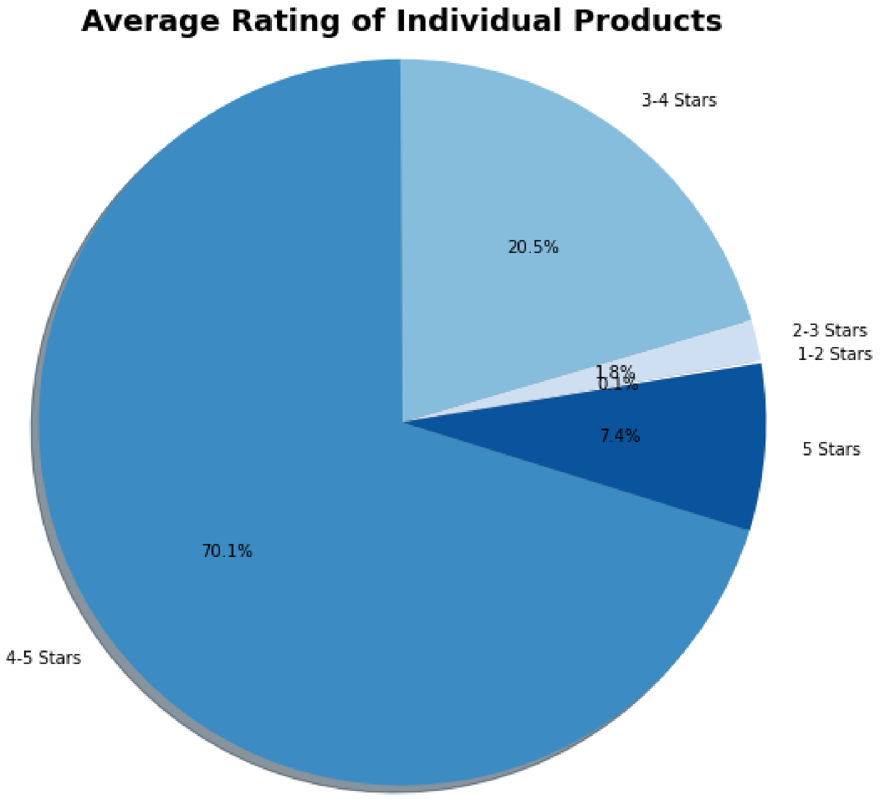




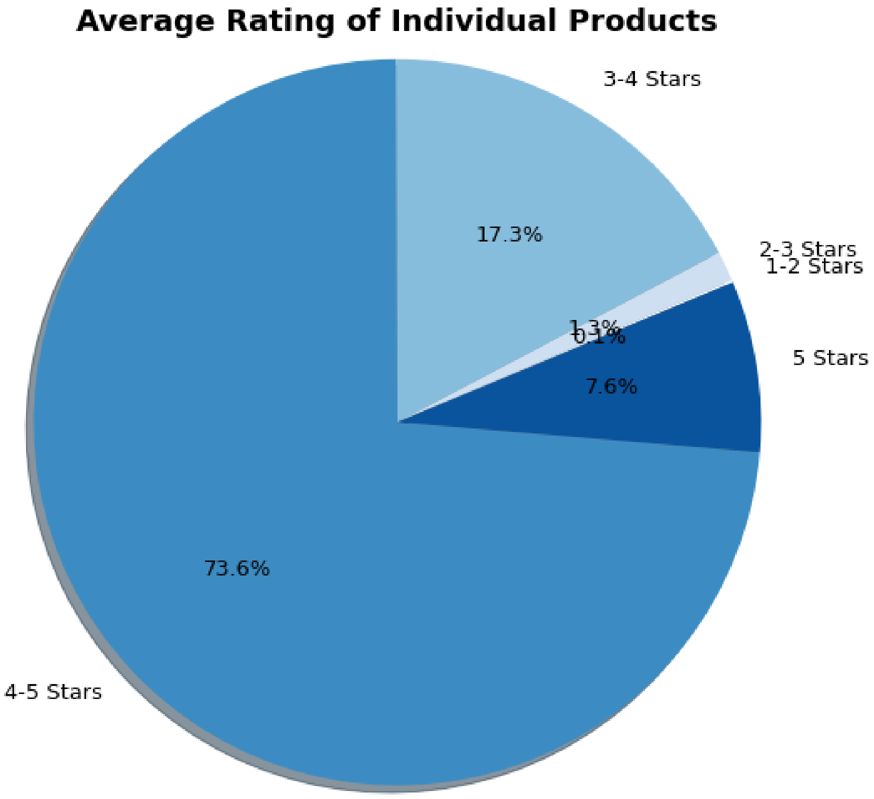
Group 1: CM Knox, Ronald Manipol, Traci Garrett, Josh Sniderman, Shahana Shaikh



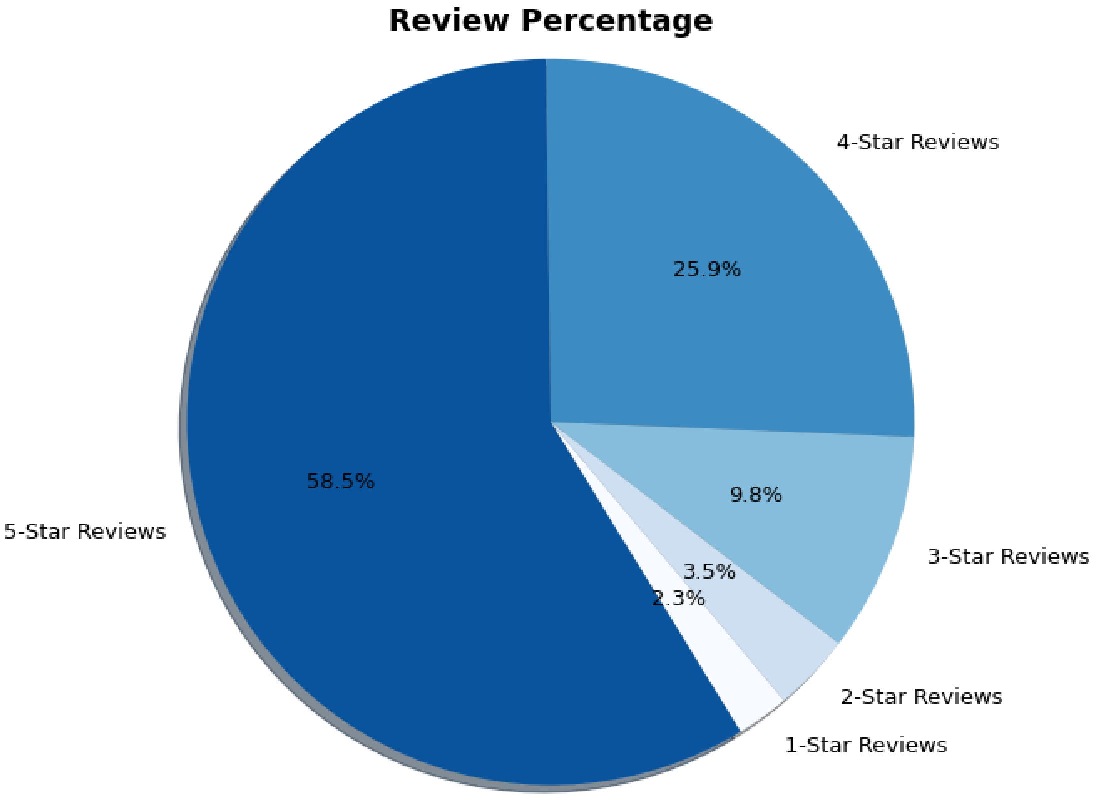
Group 1: CM Knox, Ronald Manipol, Traci Garrett, Josh Sniderman, Shahana Shaikh



Group 1: CM Knox, Ronald Manipol, Traci Garrett, Josh Sniderman, Shahana Shaikh


Average Rating of Individual Products

Group 1: CM Knox, Ronald Manipol, Traci Garrett, Josh Sniderman, Shahana Shaikh


Review Percentage
4-Star Reviews
5-Star Reviews
3-Star Reviews
2-Star Reviews
1-Star Reviews