Project: Is talking to Alexa enough? Or do we need more?

Amazon's Alexa-enabled Echo started as a smart speaker that served as an in-home hub for the company. There are now many Echo devices with different sizes and form factors, such as those with a screen. Alexa has also come a long way, as she has acquired many new "skills" over the past few years and she can now be the virtual butler of your smart home. Other than playing music, reading news, checking weather, or selling stuff from Amazon, Alexa can now help the user around the house if you have the supported smart home products and services.

This study is going to focus on Echo devices that have a screen – a logical and sensible next-step for the Echo smart speaker product line. These devices are no longer just smart speakers as they can now deliver both audio and visual information to the users. The screen provides another dimension for users to interact with these devices, and Alexa has also evolved to a more capable virtual assistant, who can talk as well show you things.

The key question of this study, as reflected by the title, is to investigate whether these Alexa-enabled devices is better with a screen. This study will examine the user experience to obtain insights into the following questions:
1) For Echo's users, does the added screen and its related functions provide a better experience relative to other screen-less Echo smart speakers?
2) In consideration of the next generation of similar Echo products, what are features or services to include, improve, or exclude?

Data set:

This study will investigate the user experience based on Amazon reviews submitted on five Echo products, including three devices with screen and two screen-less smart speakers:
Echo Show 1st Generation (with screen, #reviews = 4000)
Echo Show 2nd Generation (with screen, #reviews = 2048)
Echo Spot 1st Generation (with screen, #reviews = 4000)
Echo Plus 1st Generation (screen-less, #reviews = 3208)
Echo Plus 2nd Generation (screen-less, #reviews = 1832)

The reviews were downloaded at the end of February, 2019, using the web scrapper implemented in Chrome. The detailed procedures are outlined as follow:
https://www.scrapehero.com/amazon-review-scraper/

The scrapper extracts review information, including author, title, date, review content, and star rating, from each product page, and the reports, with up to 4000 reviews per product, are saved as csv files.

The review data is relatively clean. This study will focus on the review titles, review contents, star ratings and review submission dates. For text data, including titles and contents, punctuations were removed, and lower case was applied. Both reviews and titles were tokenized using the NLTK package, with stop words removed. Bigram and trigram tokens were also generated by linking unigram tokens with underscore. Character count and word count were also

computed for both title and content to measure text length. Star ratings, captured as a string in the format of "X.0 out of 5 stars", was converted to integers. A handful of invalid reviews have a star rating of 0, which were removed from the dataset. No missing data was otherwise found in these key columns.

Data analysis

The first step to explore the reviews data is to examine the satisfaction from each of these productions. Based on the provided star rating, ranging from 1 to 5, the mean star rating was plotted in Figure 1 below. Overall, all devices appear to have good similar ratings. With the exception of Echo Show 1st Generation (mean star rating = 3.7), the other four devices scored more than 4 stars and the Echo Plus 2nd Generation has the best rating of 4.6.

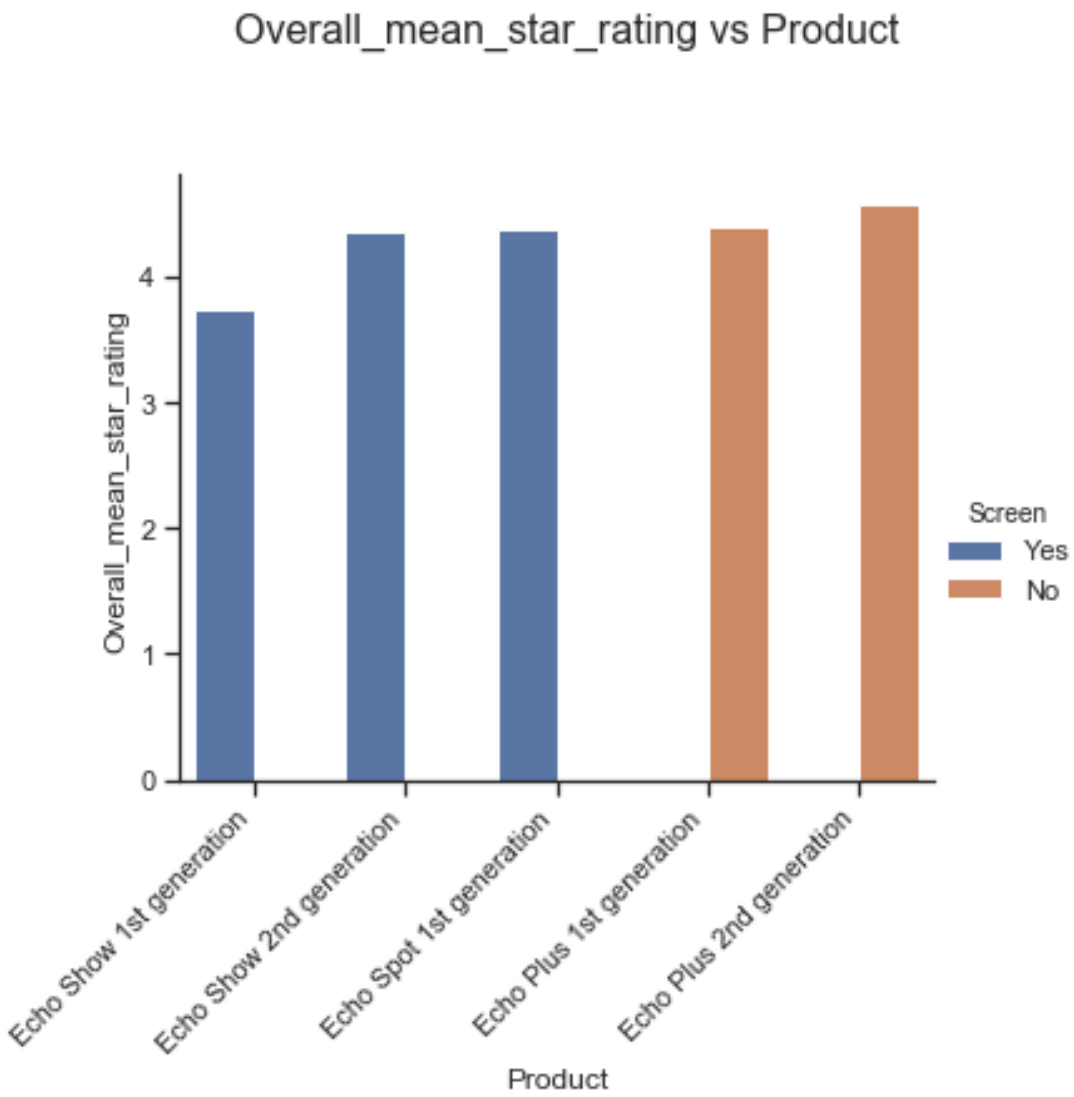

Figure 1. Mean star ratings of selected Echo devices.

The review length was further examined. The mean word counts were calculated from title and contents of reviews for each device. As shown in Figure 2, the titles and review title and contents are shortest for 5-star reviews, and the word counts increase as the star rating decreases. Interestingly, the longest reviews are either 2-star or 3-star reviews, not the 1-star reviews. In addition, reviews for Echo Show and Echo Spot are generally longer than those for Echo Plus. The Echo Show 1st Generation has the longest reviews across all star ratings, while Echo Plus 1st Generation has the shortest reviews. In some cases, the difference is quite significant. For example, the 5-star reviews of Echo Show 1st Generation have 92 words in average, which are 3 times as long as those 5-star reviews from Echo Plus 1st Generation.

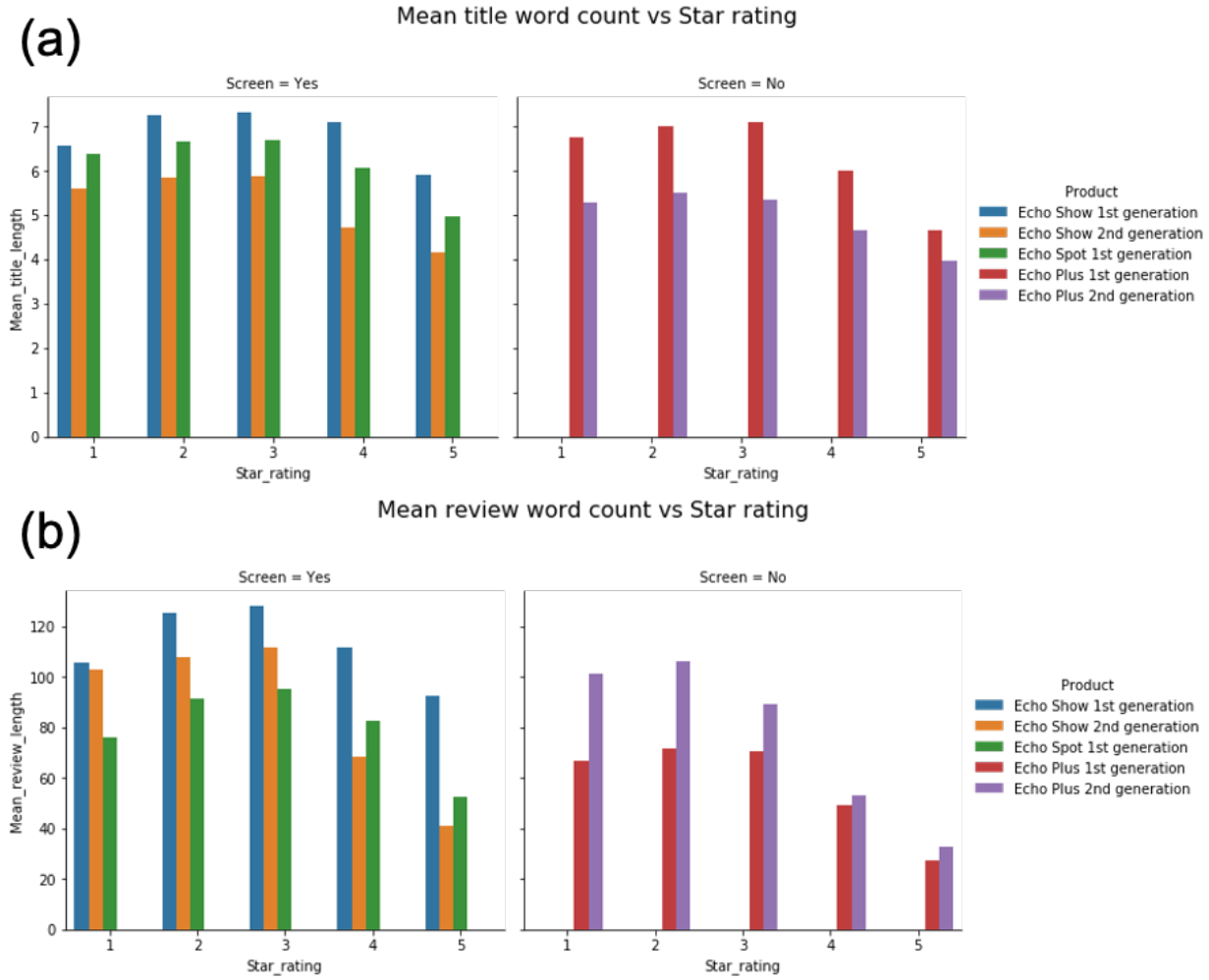

Figure 2. (a) Mean title lengths (word count) of reviews grouped by star rating. (b) Mean review content lengths (word count) of reviews grouped by star rating.

In Figure 3, the mean review length was plotted against the mean star rating, showing that reviewers, when they are not as satisfied with the device as reflected by the lower star rating, would write more. Interestingly, the two screen-less devices, with their relatively higher mean star ratings, have shorter reviews than the three devices with screen.

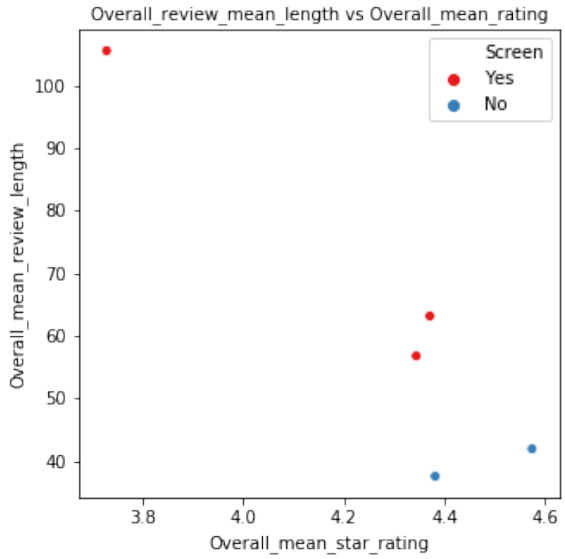

Figure 3. Mean review length vs. mean star rating of selected devices.

Further examination of the data shows that the reviews for all the five selected devices are highly skewed with mostly 5-star reviews. The number of reviews and their relative proportion for each star-rating group are plotted in Figures 4 and 5, respectively. Figure 4 shows that 5-star reviews made up for about 70% of the reviews for all devices but Echo Show 1st Generation, from which ~50% are 5-star reviews.

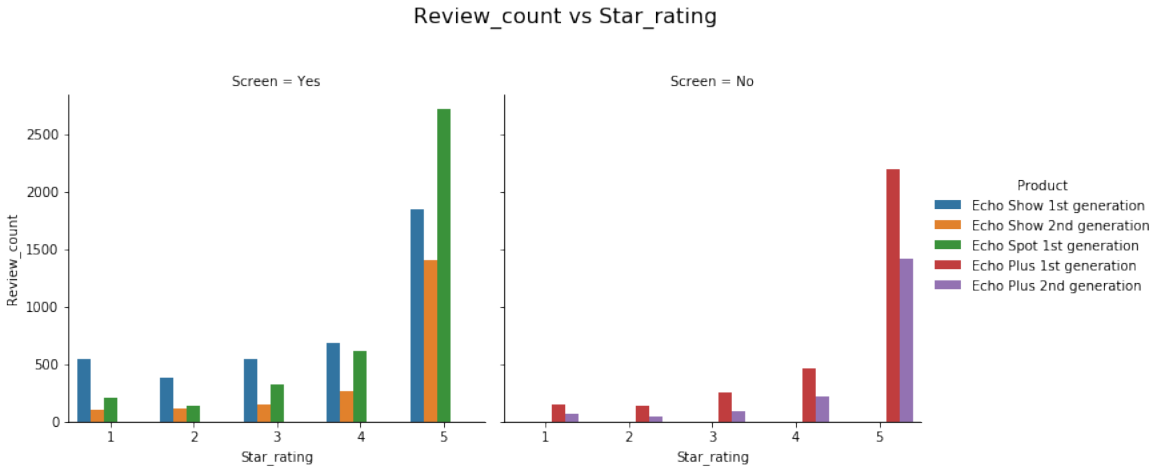

Figure 4. Number of reviews from each star rating group.

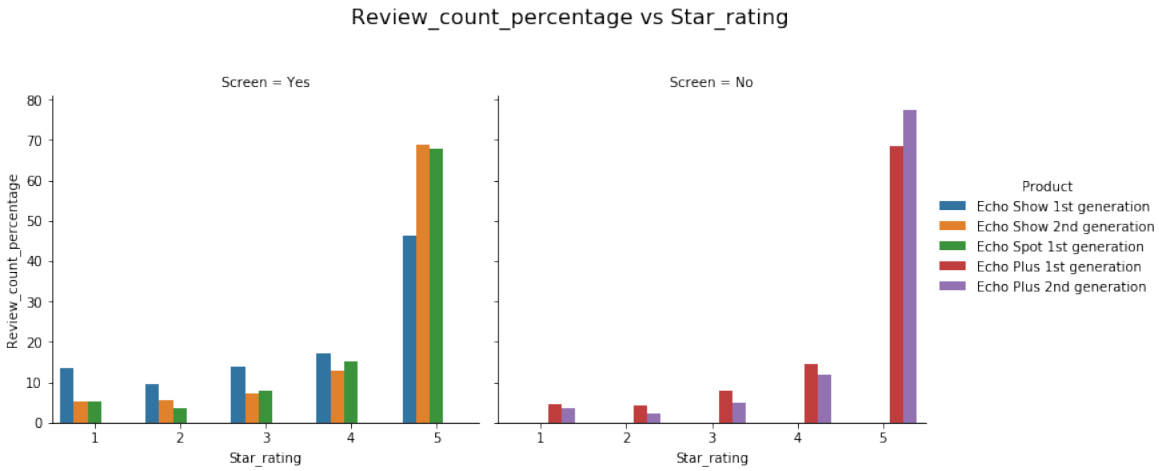

Figure 5. Relative proportion of review count from each star rating group.

The overall high mean star ratings can probably be explained by the skewed reviews. The above analysis also suggests that reviewer's satisfaction might influence how much they write. Understanding the relationship between the reviewer's sentiment and the review content would be of interest. However, being able to extract information regardless of the review's length would be important. The next steps will include examining the reviews further by word cloud analysis, which would help identifying features that are liked and disliked by the reviewers, and by performing a sentiment analysis, which would allow us to measure and analyze the sentiments expressed quantitatively.

Word cloud analysis

Monogram, bigram and trigram word cloud analyses were performed on both review titles and review contents of each star-rating group from all selected Echo devices. Aiming to identify the most liked and disliked features, the following analysis will focus on the 5-star and 1-star reviews. Word clouds were generated based on monogram, bigram and trigram frequencies, which are shown respectively in Figures 6 and 7 from 5-star and 1-star reviews. In each figure, the 20 most frequent words/phrase are shown, and the word size based on the computed frequency.

The word cloud analyses from the 5-star reviews reveal some of the most popular features (Figure 6). Reviewers, who gave 5-star reviews to Echo Show 1st and 2nd Gen, did mention the screen. When Echo Show was released, Amazon promoted the device's capability of making video calls via the Drop-in feature. Amazon also advertised Echo Show as a companion device in the kitchen. Both of these were present in the word clouds. Interestingly, reviewers of Echo Show 2nd Generation have mentioned "sound" a lot, which is similar to the reviews for the screen-less speakers. Reviews of Echo Spot, which is advertised as a smart alarm clock, show buyers do enjoy using the device as alarm clock. Unsurprisingly, Amazon's smart voice assistant, Alexa, appears to be a prominent feature for all devices. Non-Amazon services that are compatible with Echo, such as Spotify, however, are not present, suggesting that these third-party services are overshadowed by the Amazon's services.

Figure 6. Word clouds from 5-star review contents.

Figure 7. Word clouds from 1-star review contents.

Word clouds from 1-star reviews were generated to examine reviews at the other end of the spectrum. Interestingly, some features appear in both unigram word clouds from 5-star and 1-star reviews, such as Alexa, Amazon, music, screen (for Echo Shows and Echo Spot) and speaker (for Echo Pluses). This may suggest that 1-star reviewers also like these features or the 1-star reviewers actually dislike these features liked by many 5-star reviewers. The bigram and trigram word clouds do provide a bit more insights into reviews. For Echo Show devices, the reviews mentioned about YouTube videos and flash player, both of which is specific to screen-enabled devices. This likely relates to the rollout of Echo Show's own YouTube app at its release. The Echo Show 1st generation was first released with a YouTube app, however, the app was then blocked by Google for some time. The "Alexa app" and "phone" also appear here for multiple devices, which could reflect on interactions with the Echo devices via the Alexa app in the phone. This analysis also helps identifying couple device-specific issues. Some Echo Plus 2nd generation devices might have a problem with WiFi signal (from bigram word cloud). Some 1-star reviewers also reported screen issues, such as screen flickering and lines on the screen, for Echo Show 1st generation. These issues, while do not appear in the word clouds of the review contents, was present the word clouds generated from the review titles as shown in Figure 8.

Figure 8. Word clouds from 1-star review titles of Echo Show 1st Generation. Words and phrases related to screen flickering is highlighted in pink.

The next step is to carry out sentiment analysis to quantify the sentiments expressed in these reviews. The goal would be to connect the results from sentiment analysis and the features/issues identified from the word cloud analysis, so we could profile and compare user experience of these Echo devices.