Clothing Reviews - Milestone Report 2

Prepared by: Kue Ho

Contact: [kuevho@gmail.com](mailto:kuevho@gmail.com)

July 23, 2019

**Problem Statement:**

The current reality of the apparel industry is that less and less consumers are frequenting brick and mortar establishments. Customers are finding themselves drawn more towards online shopping and e-commerce for its ease and convenience. A unique problem that is arising from e-commerce, particularly for the apparel industry, is how customers are deciding on specific products without physically seeing, touching, or trying on articles of clothing. As a result, reviews are the go-to resource when making decisions on whether to buy something or not. Reviews are not only important for consumers, but also for any online retail. They rely on reviews to generate interest and profit on a particular item.

Reviews can be extensive and oftentimes contain irrelevant information. Can we compress the data to improve the product? Can we take the reviews and predict whether an item will get recommended or not? From reviews, online retailers can improve the quality of their product to satisfy consumers and increase profit and credibility.

**Target Audience:**

* Online Retailers (Amazon, Target, Walmart, Ebay, etc.)
  + Use this report to improve products and increase profit and credibility.
* Online customers
  + Consider this report as a reference for quality reviews and key words for choosing a product or online retailer.

**Data Description:**

The dataset was gathered from Kaggle website: <https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews/version/1> Some of the main features consist of the following: Clothing ID, Age, Review Text, Recommended, Positive Feedback Count, etc.

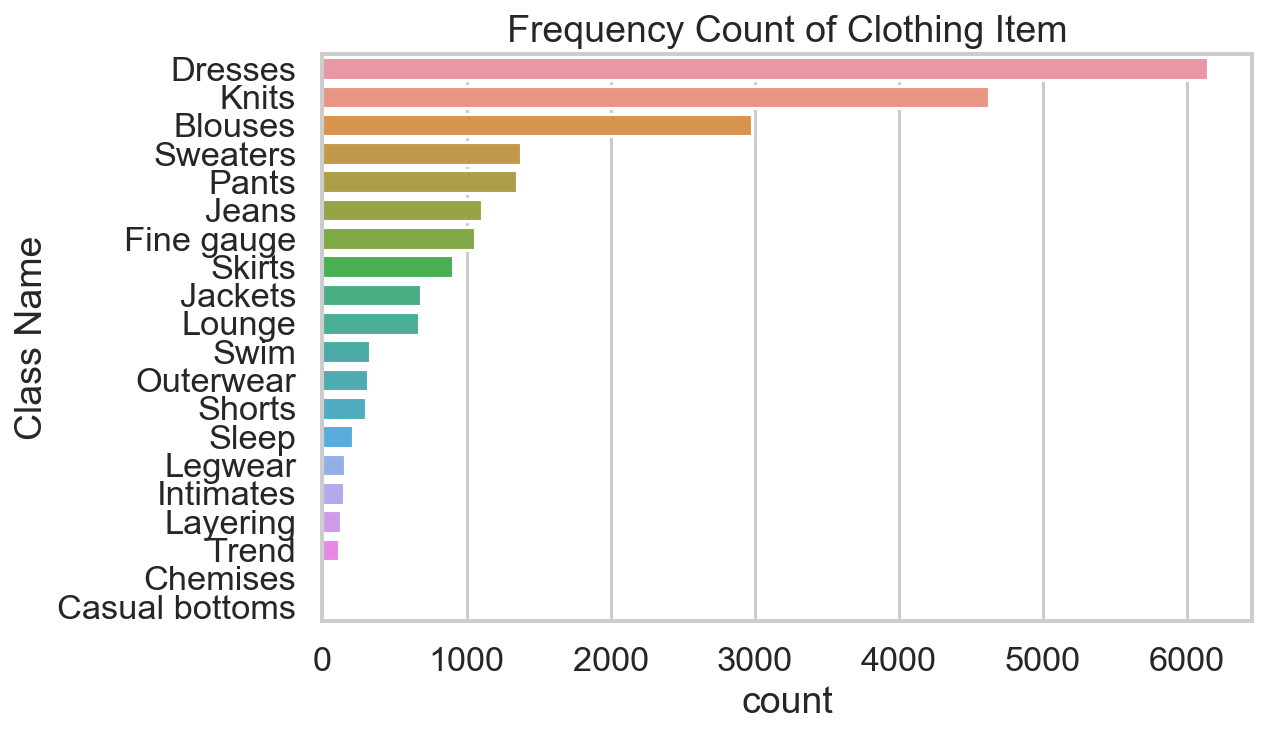
With Python’s pandas package, the data was read into the system as a DataFrame. With the ‘info’ attribute, I found 23,486 rows and 11 columns, with a size of 2.0 MB.

Since the focus of this project is on review text and clothing items, missing observations were deleted from the following columns: division name, department name, class name and review text. This dropped the observations to 22628. The title column has a significant amount of null data, but it will be left alone for now since it has no direct affect on this project.

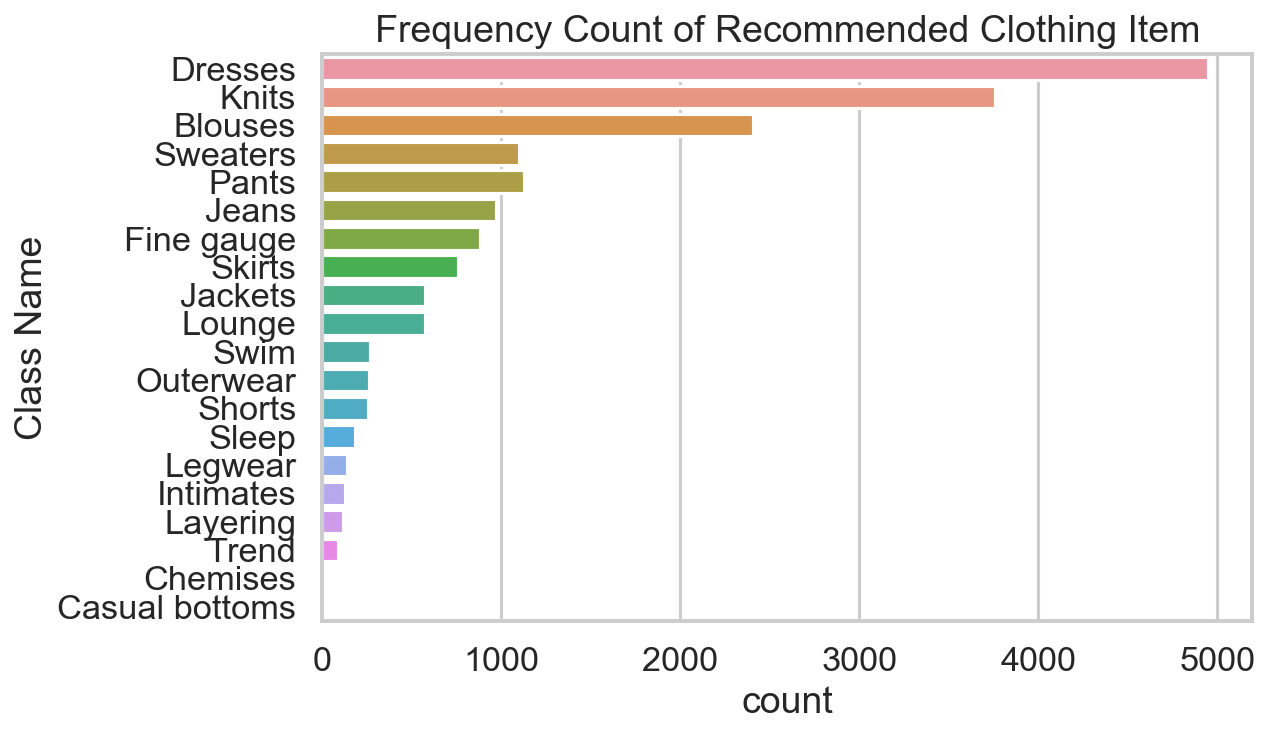
A new column was created to label ratings either negative or positive. A negative label meant the rating was less than 3. A positive then consisted of 3 or above.

**Key Findings:**

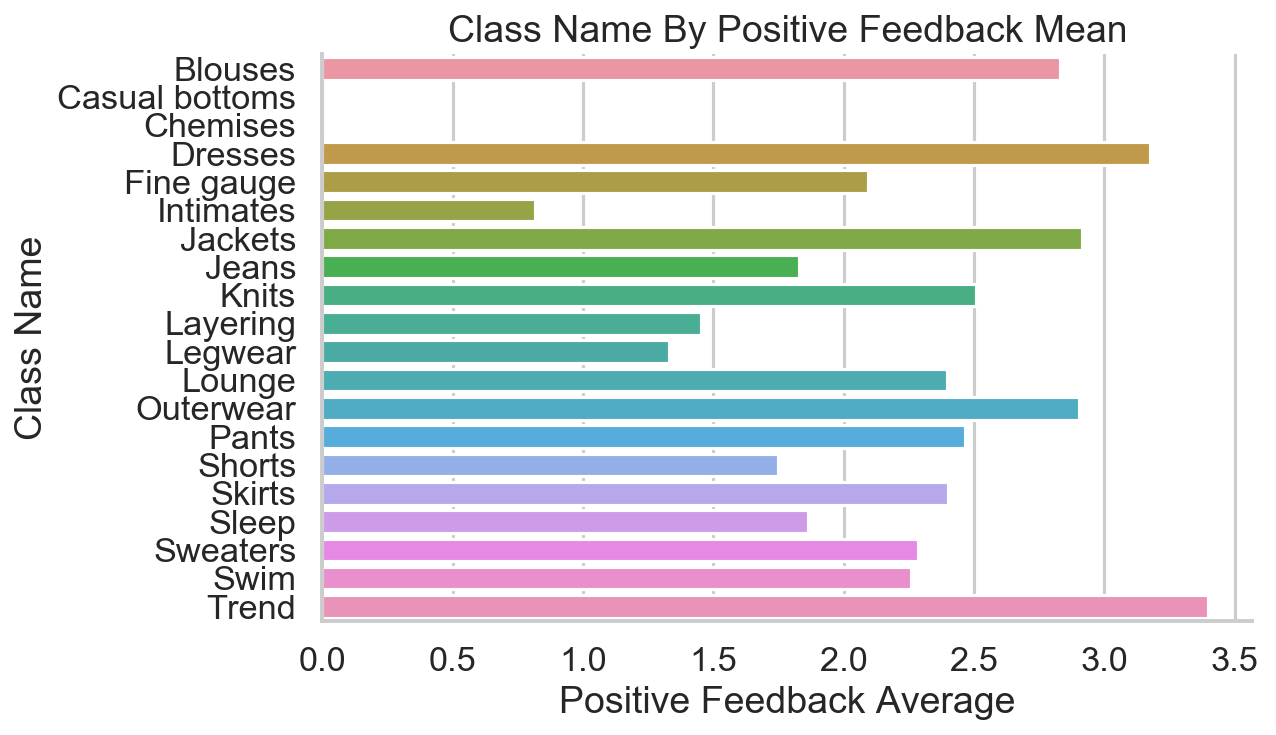
The most reviewed clothing items are dresses, knits and blouses. This indicates that most women buy these items often and they are the most popular items of the online retail. These items help retailers understand what they’re customers want or look for.



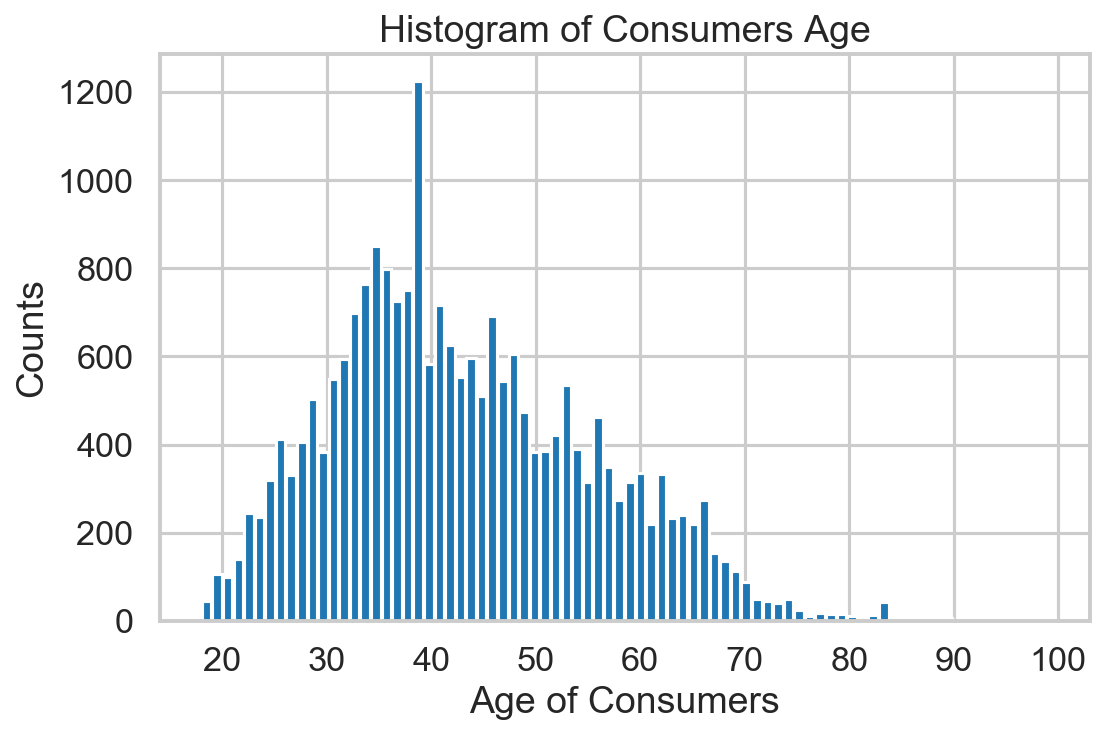
The most reviewed items are also the most recommended items as well. This indicates that most reviewed items also have positive reviews hence the high recommendation of said items.



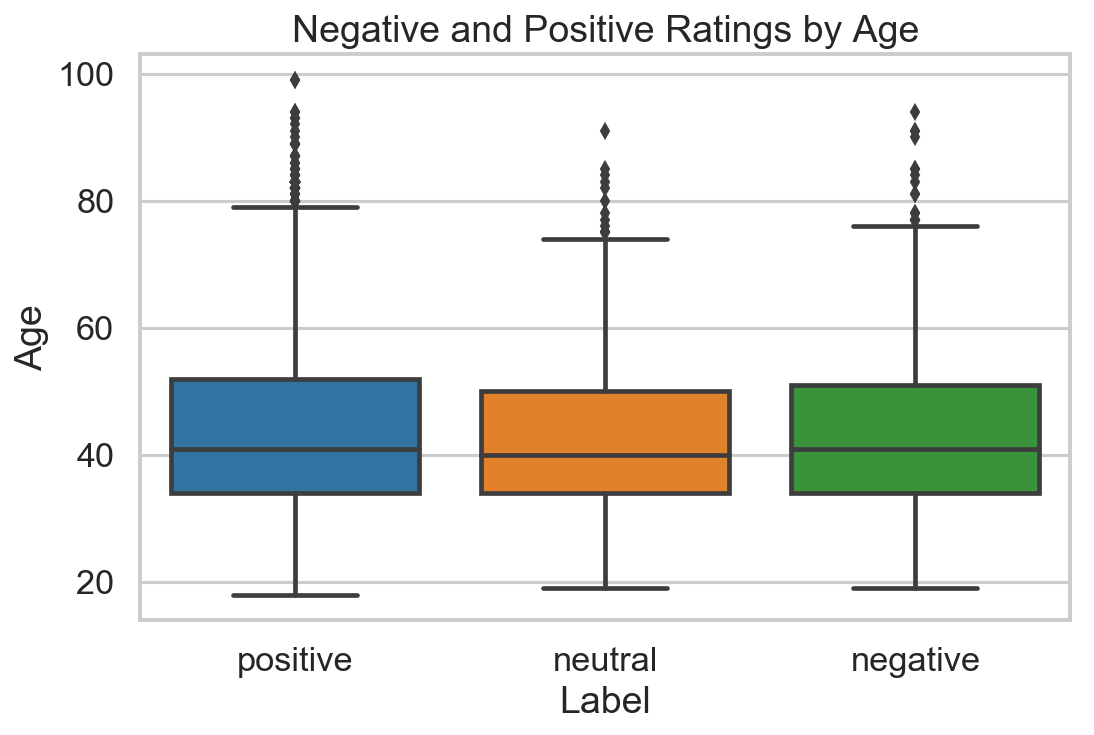
Positive feedback is an indication of how other consumers felt about the review text. If other consumers felt the review was useful, then the review received a positive feedback. Here we look at the average positive feedback of each clothing item. We see that the most reviewed and recommended items (dresses, knits, and blouses) received good average positive feedback. But we also notice that trend received the most positive feedback on average, although it did not get a lot of reviews or recommendations.

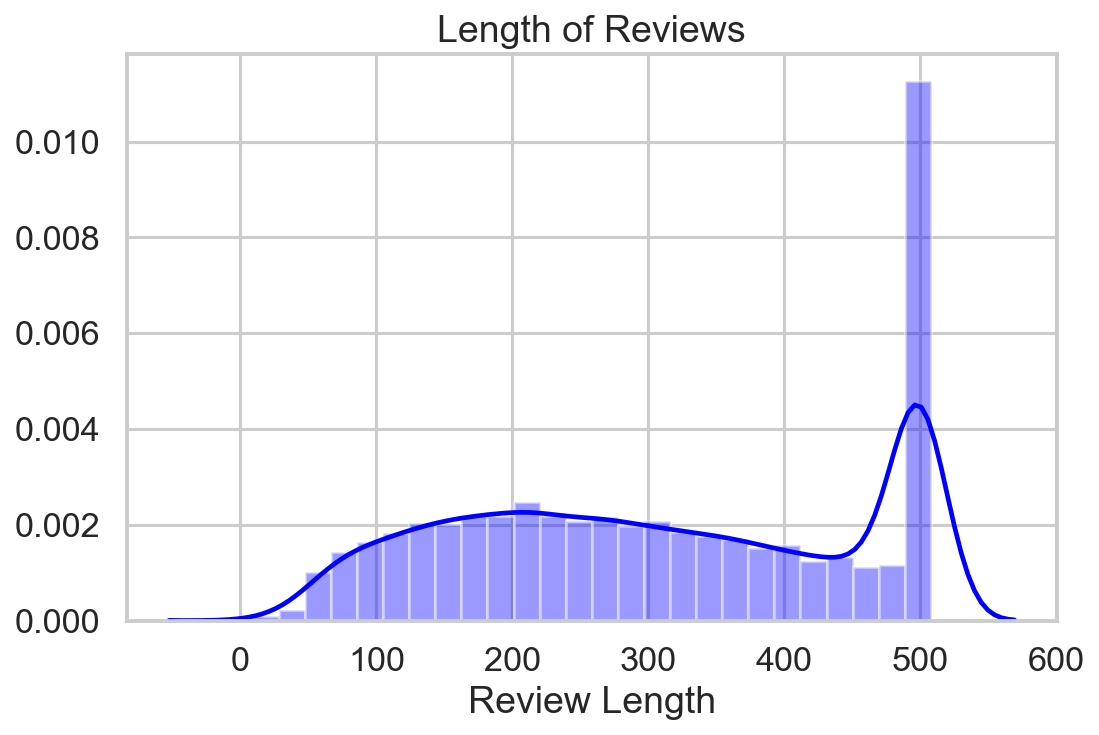


Overall, we see that most consumers range between age 30-40. The consumer target for this online retail consists of working women who may or may not be starting families (huge generalization).



The breakdown of review ratings based on age shows that rating sentiment is similar across median age. These ratings are skewed left. The minimum positive rating starts at the youngest age versus neutral and negative ratings.



Exploring the review text, we see that the length of review share mainly a unimodal curve except for the hard spike at 500. This indicates that there is a max length of characters allowed for each review and that is 500 characters total.

**Sentiment Analysis:**

Sentiment analysis aims to identify and classify opinions or emotions of each review text. There are two common sentiment analysis models that was used in this project: Textblob and Afinn. Textblob determines sentiment by looking at polarity and subjectivity of a statement. Afinn also looks at polarity to determine sentiment. It contains over 3,300 words with polarity associated to each word.

Assuming that rating is a good representation of an individual’s feeling of the clothing item, rating was converted to positive, neutral and negative sentiment. Then it was used as the ground truth table to compare both models to. Using accuracy and classification metrics, we find that Afinn was better at determining sentiment closest to rating, with an accuracy score of \_\_%.

**Predictions:**

**Summary:**

As we explore this data set, we see that consumers favor dresses, blouses and knits. We also learn that most of the consumers are 30-40 years old. The breakdown of ratings into sentiment compared to age shows that there is no significant difference between the two variables. A 40 year old woman who shares a positive sentiment about an item most likely will have a negative sentiment about a similar item. Lastly, we see that the review text maximizes at 500 characters whereas most reviews end at 200 characters.