WOMEN'S E-COMMERCE CLOTHING REVIEWS INITIAL MILESTONE REPORT

1. Problem Statement:

E-commerce is the activity of buying or selling of products on online services or over the Internet. Electronic commerce draws on technologies such as mobile commerce, electronic funds transfer, supply chain management, internet marketing, online transaction processing, electronic data interchange (EDI), inventory management systems, and automated data collection systems.

Modern e-commerce typically uses the World Wide Web for at least one part of the transaction's life cycle although it may also use other technologies such as e-mail. The largest Internet retailer in the world as measured by revenue and market capitalization is Amazon.com, Inc. which is an American e-commerce and cloud computing company. And it became the fourth most valuable public company in the world (behind only Apple, Alphabet, and Microsoft), the largest Internet company by revenue in the world, and after Walmart, the second largest employer in the United States.

But from the individual internet trader point of view, the selling rates mostly depend on the reviews and ratings left by the customers which shows how they are satisfied with the product. That is the reason why it becomes crucial to predict whether customers will leave a good, neutral or bad rating based on their reviews.

2. Who Might Care:

Ecommerce traders whose trade mostly depends on the reviews and ratings.

3. Description of the Data Set:

Women's Clothing E-Commerce dataset revolving around the reviews written by customers. Its nine supportive features offer a great environment to parse out the text through its multiple dimensions. Because this is real commercial data, it has been anonymized, and references to the company in the review text and body have been replaced with "retailer".

This dataset includes 23486 rows and 10 feature variables. Each row corresponds to a customer review, and includes the variables:

- Clothing ID: Integer Categorical variable that refers to the specific piece being reviewed.
- Age: Positive Integer variable of the reviewers age.
- Title: String variable for the title of the review.
- Review Text: String variable for the review body.
- Rating: Positive Ordinal Integer variable for the product score granted by the customer from 1 Worst, to 5 Best.
- Recommended IND: Binary variable stating where the customer recommends the product where 1 is recommended, 0 is not recommended.
- Positive Feedback Count: Positive Integer documenting the number of other customers who found this review positive.
- Division Name: Categorical name of the product high level division.
- Department Name: Categorical name of the product department name.
- Class Name: Categorical name of the product class name.

https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews/home

4. Data Wrangling:

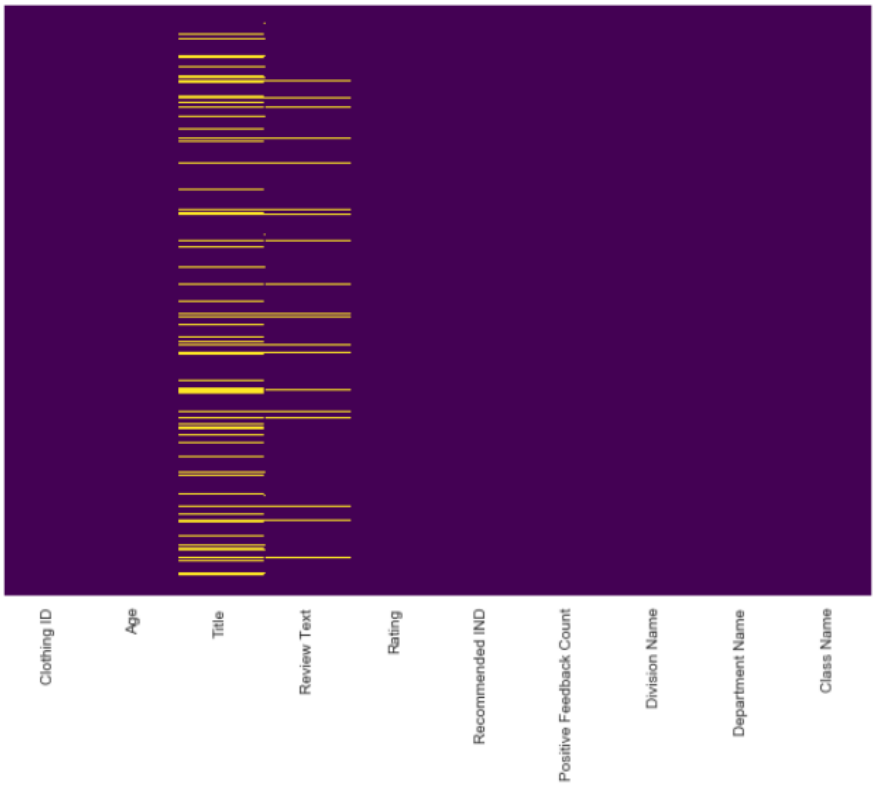
a. Data Set Basic Formatting:

There was no basic formatting on the data set as seen below. The feature names and the whitespace between the word of the names was filled with '_'.

	Unnamed: 0	Clothing ID	Age	Title	Review Text	Rating	Recommended IND	Positive Feedback Count	Division Name	Department Name	Class Name
0	0	767	33	NaN	Absolutely wonderful - silky and sexy and comf...	4	1	0	Initmates	Intimate	Intimates
1	1	1080	34	NaN	Love this dress! it's sooo pretty. i happene...	5	1	4	General	Dresses	Dresses
2	2	1077	60	Some major design flaws	I had such high hopes for this dress and reall...	3	0	0	General	Dresses	Dresses
3	3	1049	50	My favorite buy!	I love, love, love this jumpsuit. it's fun, fl...	5	1	0	General Petite	Bottoms	Pants
4	4	847	47	Flattering shirt	This shirt is very flattering to all due to th...	5	1	6	General	Tops	Blouses

b. Missing Values:

As seen below, mostly null values are cumulated under Title and Review features. Since the texts of Title feature will give us very helpful ideas we merged these two features and created a new one named as 'new_text' which both of them have not null values.



c. Cleaning the new_text Feature:

After creating the merged feature, we applied advanced text cleaning such as:

[1] Lowercase the text

[2] Keep only words

[3] Find URLs

[4] Remove links from posts

[5] Expending contractions

[6] Removing whitespace

[7] Remove apostrophe signs

[8] Remove stop words and stemming

d. Creating a new column consists of the classification of the ratings:

We have created a new column based on the ratings. Rating 4 and 5 are classified as 'Good', Rating 3 as 'Neutral and Rating 1 and 2 as 'Bad'.

e. Creating new column consists of the length of the cleaned reviews:

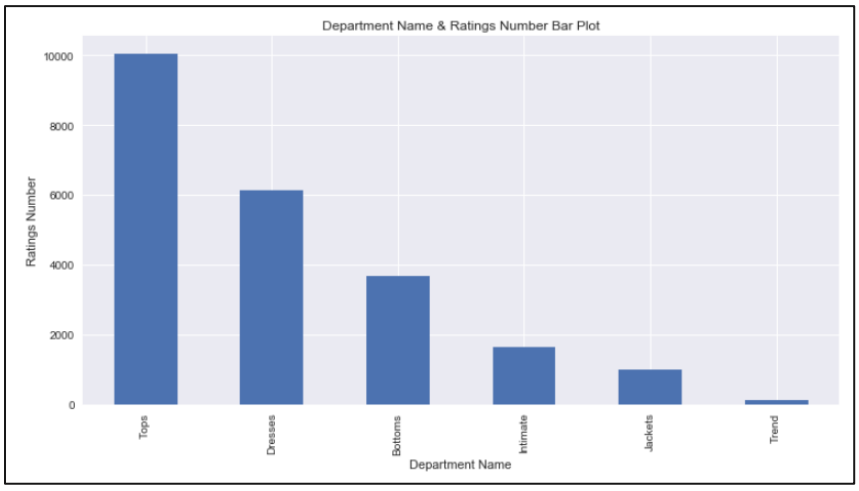
We have created two different columns, first one is the character based the length of the reviews and the second column is the word based by using the tokenizer.

f. Dropping the null values and saving the cleaned data set:

As a final step, we dropped the null values and saved the cleaned data set as 'Cleaned_Women_ECommerce.csv'

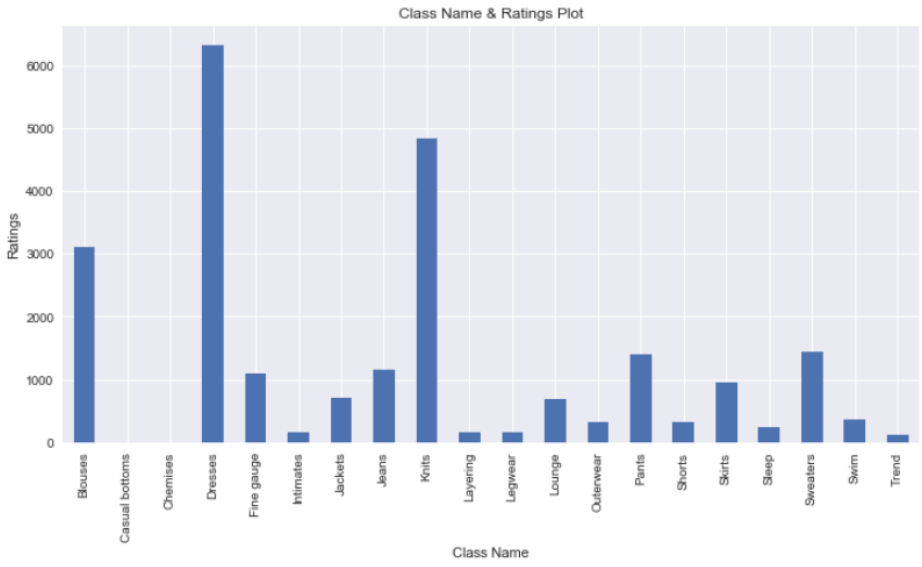
5. Data Storytelling:

a. Department Name and Ratings Number



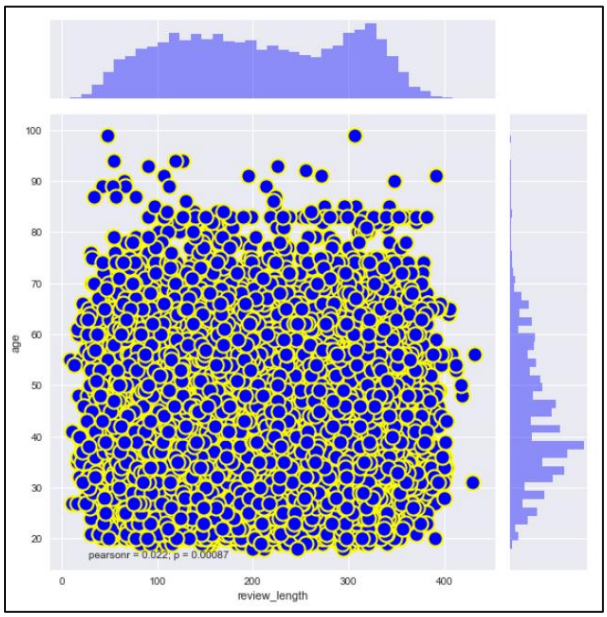
On Department basis, Tops and Dresses are sold mostly. Trend is the weakest sold as seen.

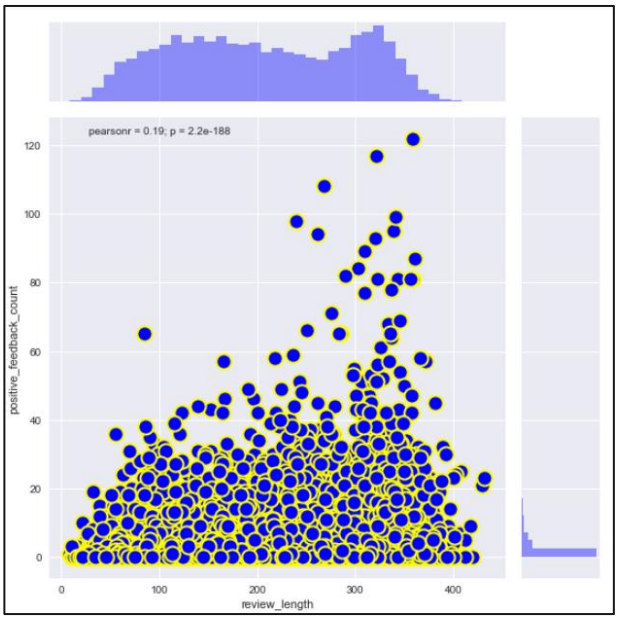
b. Rating Numbers Based on the Class Name



Most ratings were given to Dresses, Knits and Blouses.

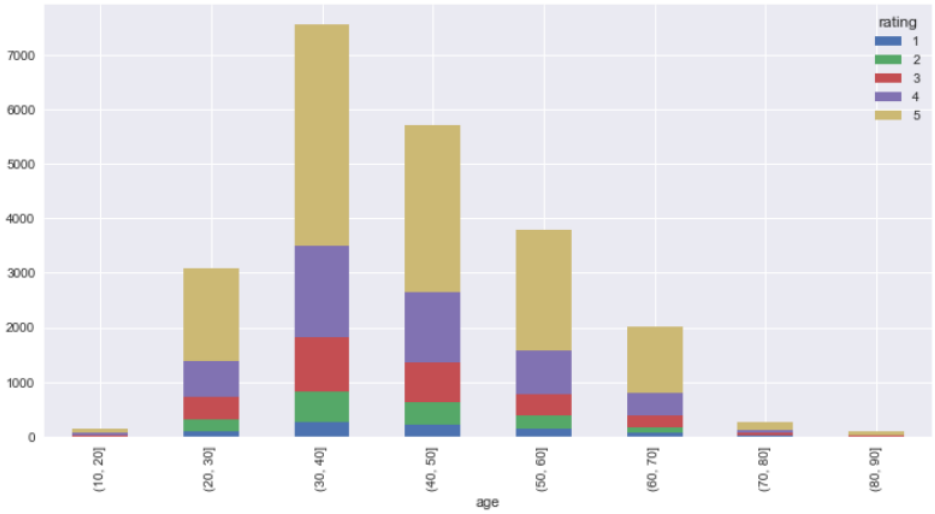
c. Review Length & Age/Positive Feedback Count





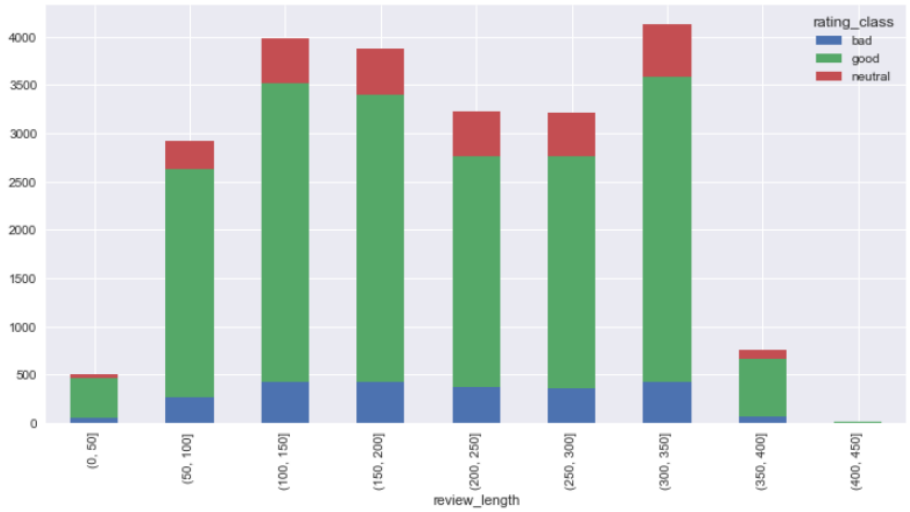
Mostly the customers between 20 and 65 ages left reviews and the review length is between 50 to 350 characters.

d. Age and Ratings



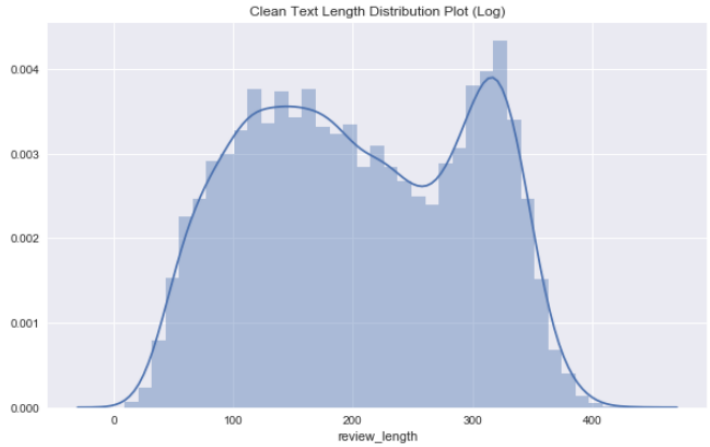
The most satisfied age group is between 30 and 50. Customers mostly left positive reviews.

e. Review Length and Rating Class:

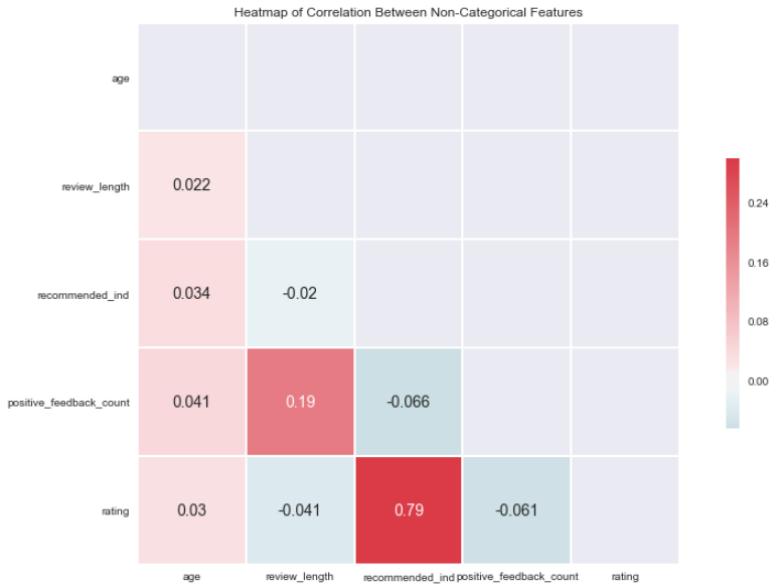


Good reviews have the longest review lengths since Good reviews has the overwhelming majority in our data set.

f. Distribution of Review Length:

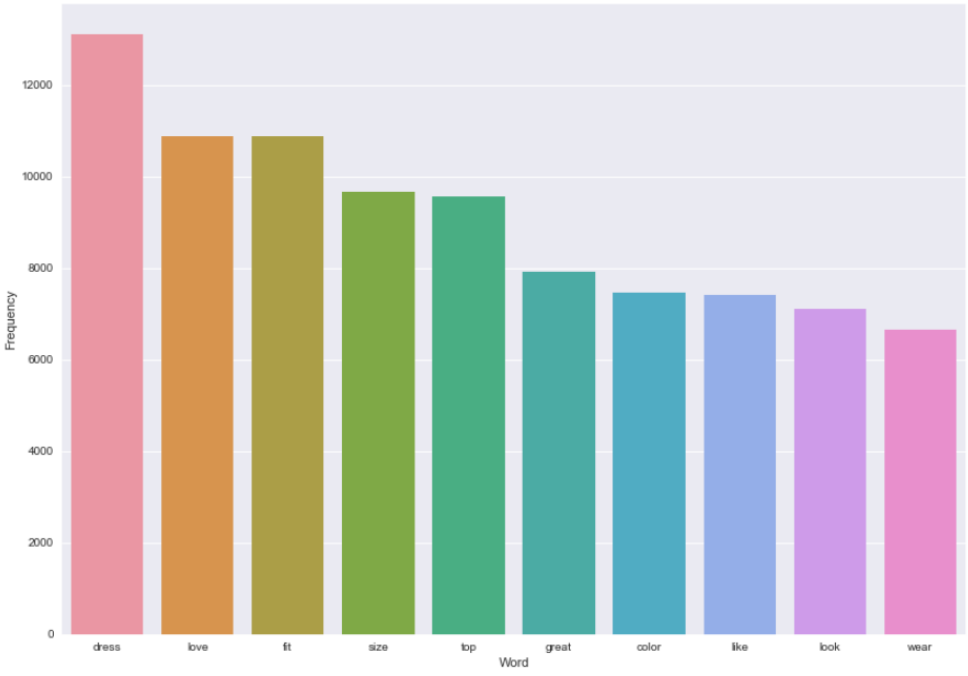


g. Numerical Features Correlation Heatmap:



There is a strong correlation between rating and the recommendation. And also a positive correlation between review length and the positive feedback count.

h. Most Common Words:

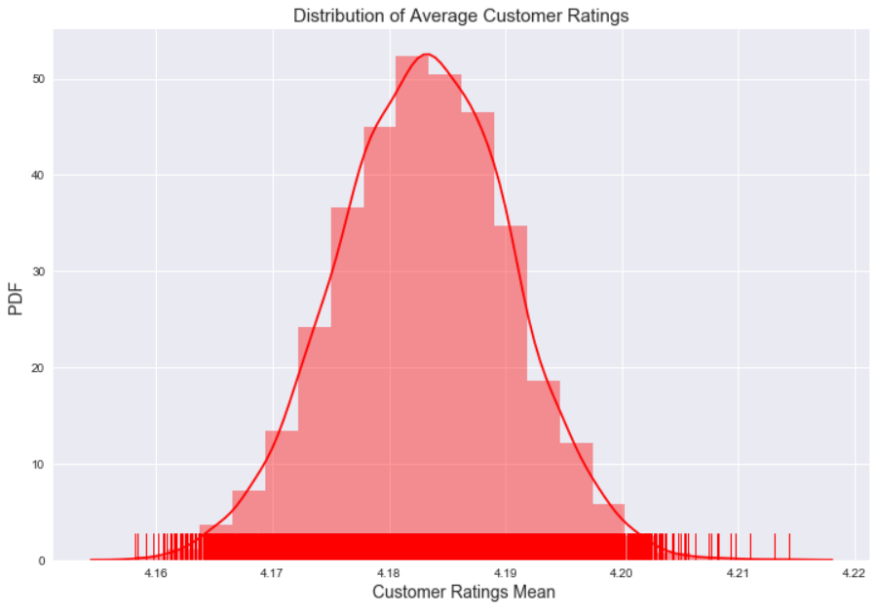


The most common words used are dress, love, fit, size, top etc.

5. Initial Findings from Exploratory Analysis:

a. Ratings:

There appears that Ratings are not normally distributed, Rating 5 is overwhelming the other Ratings. The 25 percentile is on Rating 4 and the rests was cumulated on rating 5 which shows that Rating consists of more than %50 Rating 5. Rating 4 and Rating 5 are more than %75 of the all ratings.



But the distribution of average customer rating tends to show a normal distribution.

b. Hypothesis Application:

Null Hypothesis: Null Hypothesis is: There is no significant difference between Good and Neutral Rating in average(mean) Costumer Age. We apply 2_sample test.

Result: Since p-value is 0.0, we should reject null hypothesis which means that there is significant difference between Good Rating and Neutral Rating in average of Customer Age.

6. Next Steps:

Brief Description:

I will build a baseline Logistic Regression, SVM and Naïve Bayes models to predict rating classes based on the customer reviews, I will apply different models to increase the prediction accuracy.