# Women Clothing - Predicting Ratings based on Reviews

Finance Group 3

# Agenda

1. Background & Purpose
2. Analytical Techniques
3. Results - Exploratory Analysis, Sentiment Analysis, Predictive Modelling, Cluster Analysis
4. Conclusion and Future Outlook

# Background

- Obtained a dataset from Kaggle detailing customer reviews of women's clothing in e-commerce.
- Data set includes variables such as Clothing ID, Age, Title, Review text, Rating, Recommended IND, Positive feedback count, Division name, Department name, and Class name.

# Purpose

- Improve e-commerce marketing and sales strategy and increase revenue by:
  - Understand the characteristics of reviews
  - Building a predictive model to predict ratings based on customer reviews

# Analytical Techniques

- Variables - age, product, dept
- Reviews
- Ratings

## Exploratory Analysis

## Sentiment Analysis

- Bing lexicon
- NRC sentiment polarity table - lexicon
- NRC emotion lexicon
- Sentiment score - Afinn lexicon
- Word clouds - top 100, (+) & (-) words

- Data preparation
- Predictive Model

## Predictive Modelling

## Clustering

- K-mean
- Hierarchical

# Results - Exploratory Analysis

## Variables

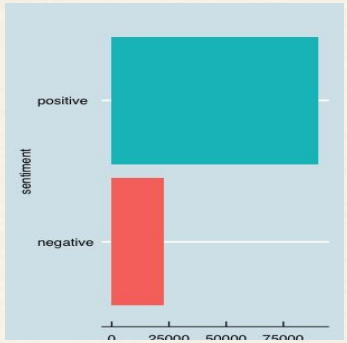
| | |
|---|---|
| Age group with highest number of reviews | **21 to 40** |
| Frequently purchased products | **Dresses, knits & blouses** |
| Most reviewed product | **Dresses, knits & blouses** |
| Most reviewed department | **Tops** |

## Reviews & Ratings

| | |
|---|---|
| Average rating | **4.18** |
| Review length | **Low correlation between length of review and rating.** |
| Uppercase | **Low correlation between uppercase letters and rating.** |
| Exclamation Mark ! | **Low correlation between ! and rating, but has a higher impact than uppercase.** |
| Common words | **Dress, Size, Love, Fit, Top, Wear, etc. (Excl stop words)** |

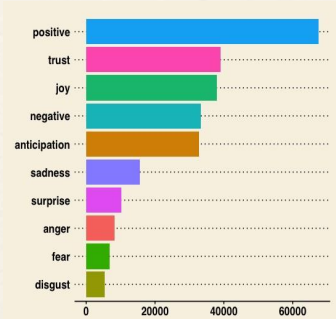
# Results - Sentiment Analysis

## Bing Lexicon



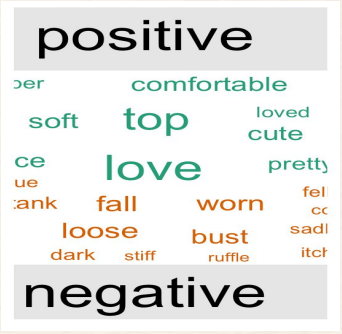

80 % of the words used in the reviews are positive.

## NRC Emotion Lexicon



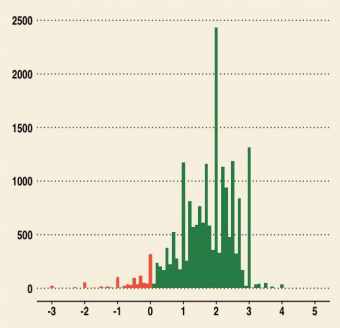

Positive & Trust emotions have the highest count.

## Word Cloud



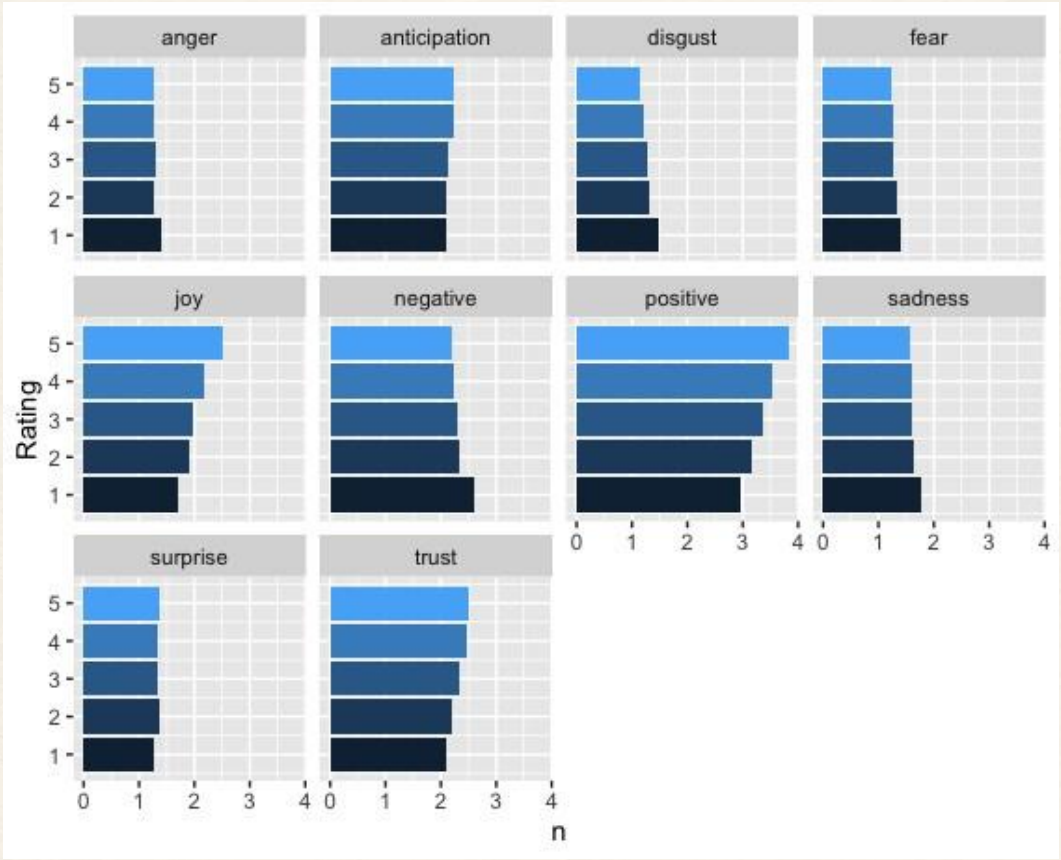

Word cloud shows green as positive & red as negative words

## Affin Lexicon




Sentiment score:
Mean = 1.71
Median = 1.85

# Results - Sentiment Analysis (cont.)




As ratings increase, there is a rise in the number of positive words and emotions and a drop in the number of negative words and emotions.

# Predictive Modelling - Data Preparation

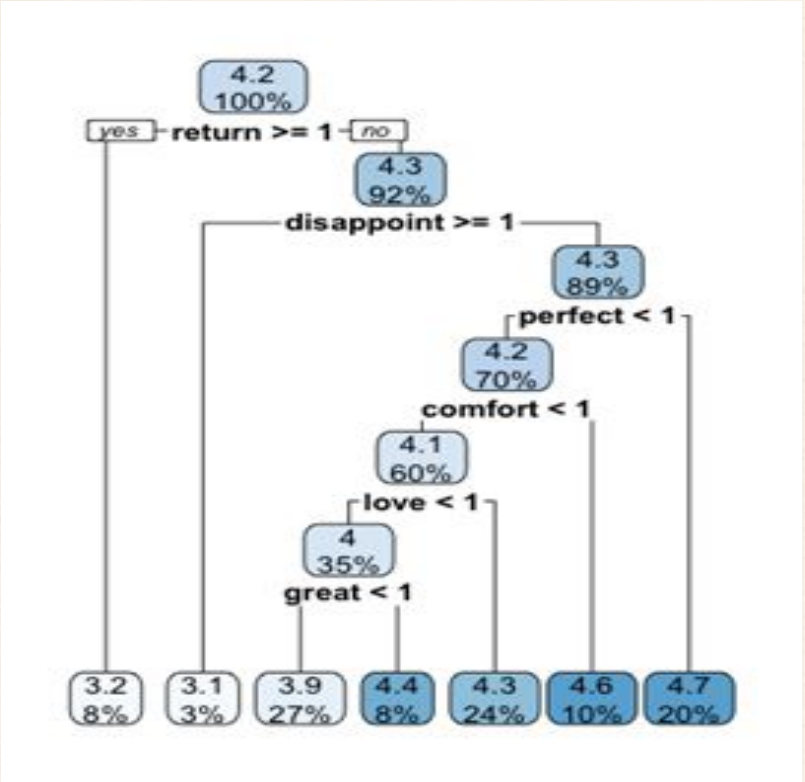
1. Checked for astronomical variables
2. Checked for outliers
3. Determined correlation among variables
4. Removed NA values – low percentage of NA entries, therefore did not impute values
5. Prepared data for sentiment analysis and rating prediction by:
    - Created a corpus from the variable 'Review.text' and 'Title'
    - Used tm_map to transform text to lower case
    - Removed stop words
    - Removed punctuation
    - Removed whitespace
6. Created a dictionary
7. Used tm_map to stem words
8. Created a document term matrix
9. Removed sparse items (words that appeared in less than 3% of the reviews)

**We have used two text columns for predictive modelling –**

- **Review.Text**
- **Title**

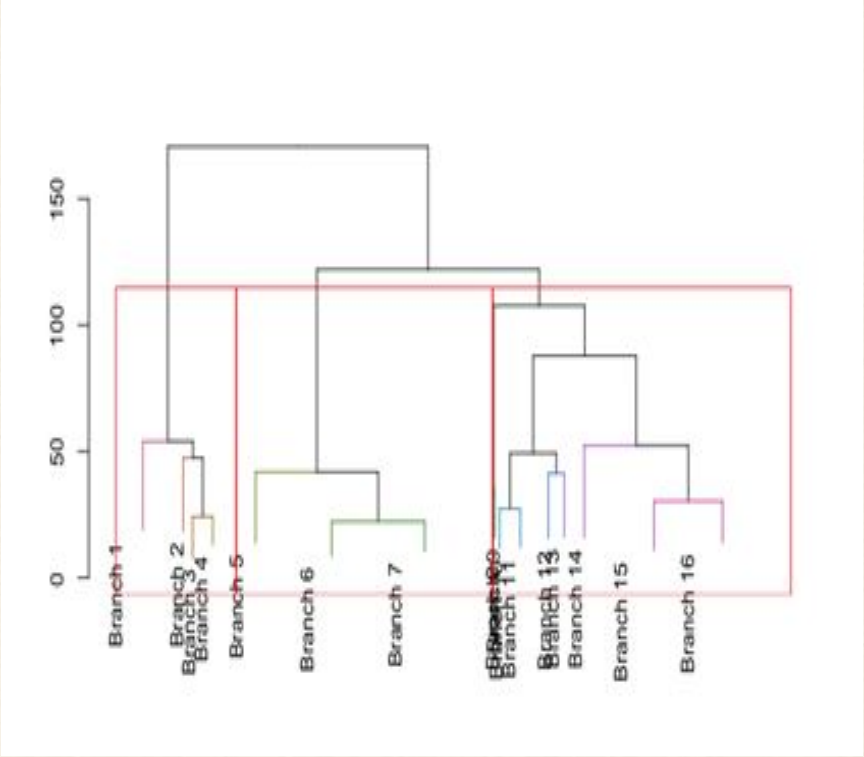
# Predictive Modelling - Results

- Using the CART method, RMSE is 1.009915 for Reviews.Text and 1.075686 for Title.
- Using regression analysis, RMSE is 0.9013822 for Reviews.Text and 1.06697 for Title.
- **Review.Text** is a better predictor of the ratings given it has a lower RMSE in both instances.
- **Regression tree** shows the words and how the words used in reviews impact whether the item was returned or not. For example, if the review has the word 'disappoint', there is a 92% chance the item was returned.



# Clustering & Predictive Modelling

- For text variables, simple regression was run, the data was normalized.
- **For numerical columns**, a hierarchical and k-means cluster analysis was run. Based on the plot, a **3-cluster solution** looks good.
- After doing the k-means clustering, total SSE plots, ratio plot and silhouette plot, the results were applied to the test set to compare the results.
- The prediction was done for each cluster and then the results were combined.
- The results indicated the following: SSE on Entire data = 2262.32, SSE on Clusters = 1643.99.
- Hence, **prediction using clusters is more accurate**, as the standard error is less.



# Conclusion

## Exploratory & Sentimental Analysis

Successfully developed:

- numerous graphical representations of text reviews
- determined correlations between text characteristics and ratings/reviews
- determined positive and negative words and most common words/character.

## Predictive Modelling

- Built a predictive model using TF, TF-IDF, Regression, Cart and Trees methods
- Predicted individual ratings for reviews using all the methods

## Cluster Analysis

- Predict using tree clusters method as it yields the lowest RMSE and standard error vs non-clusters.

# Thanks!

## Any questions?


Harsh Dhanuka - hd2457@columbia.edu

Arik Shinkarevsky - as5997@columbia.edu

Emmanuel Gyeng - emg2232@columbia.edu

Rafael Kazandjis - rk3083@columbia.edu

Umay Ayyub - ura2001@columbia.edu

Qiao Zhou - qz2395@columbia.edu

Harshitha Tummalapalli - ht2522@columbia.edu