Final Project Brief

**Part 3 – Analysis Code**

* Due Monday 24/05/2021 23:59 - This component comprises the code you use for your final analysis as will be included in your report, following exploratory work. It does not need to include code or methods that you worked on and then abandoned or ruled out, code for exploratory analysis, or code for creating graphs/tables/etc for your report.
* For this component please upload
  + the code you have used for your analysis as both an R file and a txt file
  + the dataset that you use for your analysis
* even if you have previously uploaded this dataset, please also upload it here
* your code should run on the dataset you upload, without need for any additional code not included in your R file.

**Part 4 – Report**

* This component composes your final report, where you present in full your dataset, research question, methods, results, and conclusions.
* Your report should not exceed 8 pages in length (12-point Calibri/equivalent - including exploratory analysis and any appendices).
* For this component, please upload your report as a pdf.
* Look at the PowerPoint on report for guidance

**Exploratory Analysis: E-Commerce Reviews for Women’s Clothing**

**Final Dataset Description (20685 observations of 8 variables)**

The data variables were all initially coded as characters. Variables were re-coded as necessary, and the final types are visible in the table below.

|  |  |  |
| --- | --- | --- |
| Variable | Description | Type |
| Age | Age of buyer | Numeric |
| Title | Title of review | Char |
| Review\_Text | Text of review | Char |
| Rating | Rating 1 (worst) to 5 (best) | Factor |
| Recommended\_IND | 1 (recommended) or 0 (not recommended). | Factor |
| Positive\_Feedback\_Count | Number of positive reviews for clothing item. Our dataset only differentiated between types of clothing e.g. “Knits”. This variable is a count of a specific garment e.g.: Beige woollen knit (which we did not have access to). | Numeric |
| Department\_Name | 5 departments such as ‘dresses’ or ‘trend’. | Factor |
| Class\_Name | 20 classes that categorise clothing within departments such as ‘Layering’ or ‘Sleep’ | Factor |

**Analysis Aim**

* Predict whether a customer will recommend an item using regression.
* Sentiment analysis on titles and/or ratings and analyse relationship to recommendations
* Potentially cluster similar buyers

82% of observations recommended the clothing item they bought.

**Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedCChart, bar chart

Description automatically generated**

Rating is highly correlated to recommendation as expected.

Most reviews are from Tops department and generally the same proportion (approx. 5/6) of items from a department are recommended. Class of clothing (a more specific categorisation than department) exemplifies the same behaviour (see appendix a)- recommendations are generally evenly proportioned. Therefore, I may not use department or class as a prediction tool. Specific clothing items are not more likely to be returned depending on specific type.

**Chart, bar chart

Description automatically generated**

Most common age group was 35-39. Most buyers were between 25 to 60. Age distribution mimics right-skewed distribution. Average buyer was 43. Recommendation by age looks to be slightly less proportionate, so age might play a factor in recommendations. This will have to be analysed further.

**Conclusion**

* Ongoing analysis will mostly focus on sentiment analysis of review text and title. New variables such as count of particular words could be created as a prediction for recommendation.
* Majority of continuing analysis will focus on Review text, number of existing recommendations and age as prediction variables.
* Recommendations are evenly proportioned by class and department – might not play a huge part in continuing analysis. Prior to exploratory analysis it was assumed some clothing items would have higher recommendation rates than others.
* Recommendations by age and number of positive reviews (appendix b) are not as evenly proportioned. These variables could potentially be used as predictors.
* Some variables seem to be following distributions – Age (normal with right skew), Positive Feedback Count (exponential – see appendix b)

**Appendix**

Chart, bar chart

Description automatically generated

*Appendix a – recommendation by Class of item*

*Chart

Description automatically generated*

*Appendix b – recommendation by number of existing positive reviews*

**What is sentiment Analysis? (“Opinion Mining or Emotion AI - wiki”)**

Wiki – natural language processing, text analysis etc. used to extract subjective information.. widely used in reviews and surveys, online and social media, marketing customer service

Get attitude

**Sentiment analysis in R**

[Text Mining: Sentiment Analysis in R - YouTube](https://www.youtube.com/watch?v=BqNTcewq0k0) – SQL data, wordcloud, nice graph,

Each word into a row seems to be very common… use gsub or csplit maybe

Separating by spaces, new lines, careful with special characters

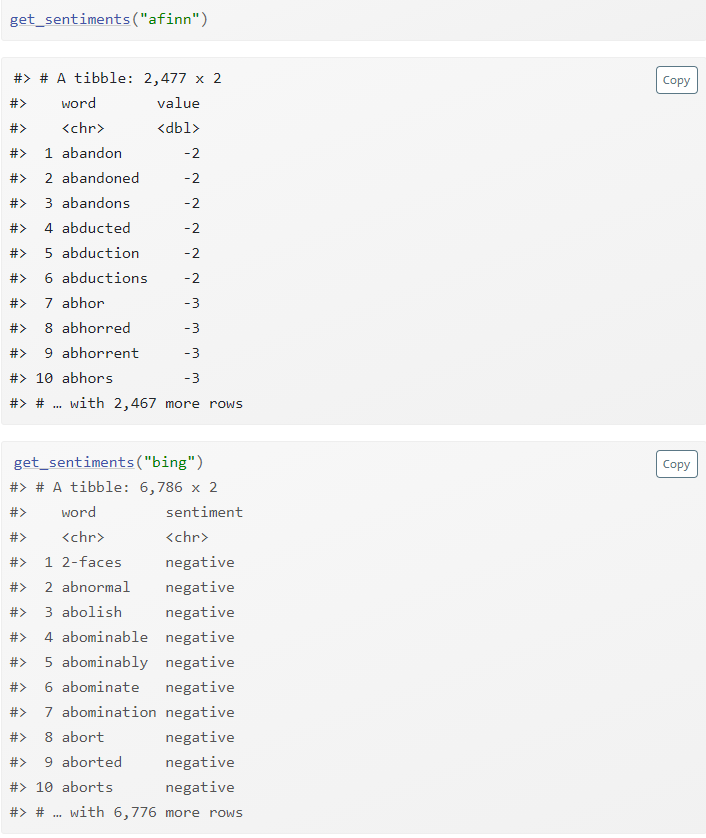
Look at most common words

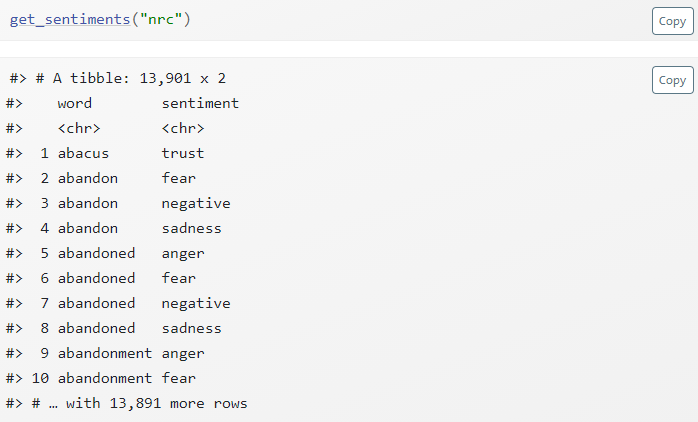
* Tail(names(sort(table(CareData$word))),10)
* You can make a wordcloud of the words might be nice..; colour by sentiment…

Might have to keep ID in

Use sentiment librarys… get\_sentiments(“bing”)

1. Afin – use for mathematical assignments eg/: nice is 3… might be good to assign a score to a review
2. Bing – assigns sentiment as negative or positive, excludes neutral words
3. Nrc – emotion sentiment package – assigns sentiment as negative, sadness, anger, fear, helpful, etc…. will duplicate words eg: wonderful as surprise, joy and positive





Possible Problems

Negations will be tricky… i.e. not well, not bad. Or…. “I doubt anyone could find a fault with this Top”

Should we use the review title or text or both?

Should be in tidy text format - [1 The tidy text format | Text Mining with R (tidytextmining.com)](https://www.tidytextmining.com/tidytext.html)

* A table with one token per row, a token is a meaningful unit of text, such as a word
* Tibbles are modern data frames,,, convenient print methods for strings especially
* Great for use with tidy tools

**Tuesday 18th Summary**

Might use afin for the actual regression, Bing is useful for visualisations. Will have to explain how some reviews are misinterpreted.

Tomorrow assign each review a score, might not separate them into tidy format, and use as a variable in regression. Could also cluster reviews perhaps?

May have to stem words… convert dancing to dance etc.

[Sentiment Analysis using Logistic Regression and Naive Bayes | by Atharva Mashalkar | Towards Data Science](https://towardsdatascience.com/sentiment-analysis-using-logistic-regression-and-naive-bayes-16b806eb4c4b#:~:text=Sentiment%20analysis%20%28also%20known%20as%20opinion%20mining%20or,So%2C%20let%27s%20start%20sentiment%20analysis%20using%20Logistic%20Regression)

**Friday**

review\_sentiment – could use this for glm e.g. 1anger +0disgust+3happiness

using sentiment() avoids negations from sentiment library

“**sentimentr** attempts to consider valence shifters (i.e., negators, amplifiers (intensifiers), de-amplifiers (downtoners), and adversative conjunctions) while maintaining speed. Simply put, **sentimentr** is an augmented dictionary lookup”

“This dress was clearly made very well! The whole thing fell apart in the wash”

“Couldn’t be happier”

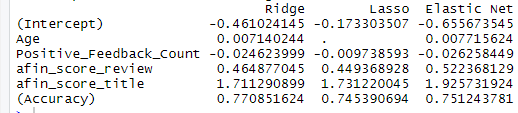
[How to do negation-proof sentiment analysis in R | R-bloggers](https://www.r-bloggers.com/2019/07/how-to-do-negation-proof-sentiment-analysis-in-r/)

[Negators and modifiers with Syuzhet vs. SentimentR for sentiment analysis in R - Stack Overflow](https://stackoverflow.com/questions/53696254/negators-and-modifiers-with-syuzhet-vs-sentimentr-for-sentiment-analysis-in-r)

Whether to use linear/lasso/ridge etc.

* The simplest form of regression is linear regression, which assumes that the predictors have a linear relationship with the target variable. The input variables are assumed to have a Gaussian distribution and are not correlated with each other (a problem called multi-collinearity).
* Ridge regression is an extension of linear regression where the loss function is modified to minimize the complexity of the model. This modification is done by adding a penalty parameter that is equivalent to **the square of the magnitude of the coefficients.**
* Lasso regression, or the Least Absolute Shrinkage and Selection Operator, is also a modification of linear regression. In lasso, the loss function is modified to minimize the complexity of the model by limiting the **sum of the absolute values of the model coefficients** (also called the l1-norm).
* Elastic net regression combines the properties of ridge and lasso regression. It works by penalizing the model using both the 1l2-norm1 and the 1l1-norm1. The model can be easily built using the caret package, which automatically selects the optimal value of parameters alpha and lambda.

**Background –** Data subject, methods(appropriate methods, techniques for missing data, suitability of different techniques), performance metrics(ROC Curve is well known but some pitfalls include… especially for this dataset….), critical analysis (opposing views, pick a side, )

1. **Introduction**
   * State the objectives of the project, give the background to it and generally orientate the reader.
   * This section can include a brief review of any relevant literature if beneficial. In longer reports, the Literature Review might be its own section.
2. **Methods**
   * Describe what you did. Justify/explain your choice of methods.
   * Data Exploration, Cleaning
   * Analysis methods… why they were appropriate
   * Evaluation Methods … why they were appropriate. Why did you choose them
3. **Results**
   * Present what you found. The results need to be presented in enough detail for your reader to understand them. Objective information. Empirical findings of your research with factual detail/criticism - eg an association was found and it is insignificant/strong/weak...
   * Discussion plus visuals
   * Graphs are key point… you had a guest lecture on this so she’ll be expecting some top-notch stuff
4. **Discussion**
   * What do your results mean? How do they relate to the literature/existing knowledge? Unrolls the main results, explain their meanings. Put there the new questions and perspectives, describe the most interesting points for the entire field. Subjective interpretation or explanation. \*Why\* these relationships are insignificant/weak/ strong etc. No new results/figures. Refer back to results section. In hindsight, should you have made different choices?
   * Conclusion: A summary of the discussion or the whole work. Re-state the main points in a new concise way that you want your readers to remember. Make recommendations.
   * Possible future work.
   * Challenges – large dataset so run-time was very long, especially when code was inefficient
   * , Methods tried and abandoned,
5. 
6. 