
AD699: Data Mining for Business Analytics
Individual Assignment #5
Spring 2021
Due Date: 19APR, 11:59 p.m.

You will submit two files:

(1) **a PDF with your write-up**, along with
(2) **the script you used** to generate your results.

## Task 1: Hierarchical Clustering

1. Read the dataset Country-data.csv into your R environment.

2. Build a hierarchical clustering model for the dataset, using the "average" method for inter-cluster dissimilarity. Do this without altering the numerical data values in any way.
    a. Create and display a dendrogram for your model.
    b. By looking at your dendrogram, how many clusters do you see here? (There is not a single correct answer to this question, and not all people will answer it the same way-- just describe the number of clusters that seem to be showing here).
    c. Use the cutree function to assign the records to clusters. Specify your desired number of clusters, and show the resulting cluster assignments for each country.
    d. Attach the assigned cluster numbers back to the original dataset (you may wish to use something similar to what we did with the in-class student clustering exercise). Use groupby() and summarize() from dplyr to generate per-cluster summary stats, and write 2-3 sentences about what you find. Which variables seemed to most strongly impact the cluster assignments?

3. If this data should be scaled, why? In a couple of sentences, make the case for putting these variables onto a common scale.

4. Use the scale() function to standardize your data, so that each variable has a mean of zero and a standard deviation of one.

5. Create another hierarchical clustering model with your scaled data, and use another dendrogram to view it. Is your new dendrogram very different from the old one? Write a few sentences that identify anything noteworthy about the differences between the dendrograms (no domain knowledge is required, and you don't need to be comprehensive here).
    a. Pick any single country. Who was it grouped with the first time, and what changed with the second model iteration? Say a little bit about why its cluster changed. (2-3 sentences should be enough here. A very good answer will demonstrate that you understand what's going on with the model, and why this record's 'neighbors' are now different).

6. In a previous step, you made the case *for* standardizing the variables. Now they're all on equal footing… but look at what this dataset is about (and keep in mind that the help function can help to explain the variables' exact meaning). Why might it be problematic to view these variables with equal weight?

7. Now it's time to fix that problem! Come up with your own weighting system for these variables, and apply it here. Be sure that you are working with a dataframe (if the data is not a dataframe, you can quickly fix that with as.data.frame(). Multiply each column by the weight that you have assigned to it.
    a. Explain the weighting system in a short paragraph. In some ways, this is a personal decision. There is of course no "right" answer to your weighting decision, but you should show here that you've made some effort to put some thought into this.

8. Now, generate one more dendrogram, using your newly-rescaled set of variables. Once more, provide some description of what you see, and whether there are any noteworthy changes between this and the other two dendrograms.
    a. Just as you did after the first hierarchical clustering, use the cutree() function one more time to assign the records to clusters. Specify your desired number of clusters, and show the resulting cluster assignments for each state.
    b. Attach the cluster assignments back to the original dataset (you may wish to use something similar to what we did with the in-class student clustering exercise). Use groupby() and summarize() from dplyr to generate per-cluster summary stats, and write 2-3 sentences about what you find.
    c. Let's check back in on that country that you selected during a previous step. Where is that country now, with this new model? Who else is in its cluster? In a few sentences, talk about what changed, and why, regarding this country's cluster assignment.

**Task 2: Text Mining**

1. Load the *gutenbergr* package into your R environment. Bring the text whose number is the same as twice your seed into your R environment (if no title in the Gutenberg library has a value that's twice your seed value, that's okay -- you can just pick another title with a nearby value). If the book does not contain any words, the same thing applies -- just pick a title with a nearby number.

2. Use the gutenberg_download() function to bring the text into your environment. Save your text as an object in your environment, using any variable name that you wish to call it.

3. Call the View() function on the object you created in the previous step. In a sentence, how would you describe what you see?

4. Now, let's get this text into a tidy format. We can use the unnest_tokens() function to help us out with this.
    a. Run the unnest_tokens() function now, and be sure to reassign the results to a new object.
    b. View this object -- what do you see? Describe it in a sentence, and explain what changed from this step to what you saw in Step 2.

5. What were the 10 most frequently used words in your book? Show the code that you used to answer this question, along with your results.
    a. Now, use the anti_join() function to remove stopwords. Show the code that you used to do this. With the stopwords removed, what are the 10 most common words in your text? Show them here.
    b. Do this again, but instead, do it with bigrams instead of unigrams.
        i. How are bigrams different from unigrams?
        ii. How might bigram analysis yield different results than unigram analysis?
    c. Write 1-2 sentences that speculate about why it might be useful/interesting to see this list of the most frequently-used words from your text. What could someone do with it? Use your imagination and creativity to answer this.

6. Next, let's do some sentiment analysis. We will use the bing lexicon for this purpose.
    a. What 10 words made the biggest sentiment contributions in your text? Show the code that you used to find this, along with your results.
    b. Of these top 10 words, how many were positive? How many were negative?
    c. In a sentence or two, speculate about what this list suggests about your text.

7. Create a barplot that shows both the 10 negative words and 10 positive words that contributed the most to the sentiment of your text.
    a. Use 2-3 sentences to describe what you see in your barplot.

8. Now let's take a look at how a different sentiment lexicon would view your text. Bring the afinn lexicon into your environment, and join it with the text from your book. Show the step(s) you used to do this.
    a. Sum all the values for your text. What was the total?
    b. What does this suggest about your text? Why might this be helpful...but why might it also be incomplete or even misleading?