CSCI 420
HW 06
Lindsay Cagarli
Anna Kurchenko

Part A: Using Cross-Correlation for Feature Rejection and Selection:

1. Using a package, compute the cross-correlation coefficients of all attributes. Your resulting matrix should be n by n where n is the number of attributes. All values computed should be in the range [-1, 1]. Record these values with two digits past the decimal point.
   Recorded in output file cross_correlations_READABLE.tsv

2. Report:(1/10) point each)
   a. **Which two attributes are most strongly cross-correlated with each other? Hint: It is the absolute value of the cross-correlation that matters, not how positive it is. [And don't forget to repeat these questions.]**

      ('NONFICT', 'ROMANCE'): -0.82

   b. **If someone buys lots of Manga books, else are they likely to buy (or not buy, depending on what is strongest.)? In other words, what is the strongest Absolute value of CC with any other category?**

      ('BABY_TODDLER', 'MANGA'), -0.6833434800123567)

      Likely to buy BABY_TODDLER

      They are very unlikely to buy SCIFI (cross correlation of 0.03)

      i. **Are they also likely to buy Horror or Gifts? Or, are they NOT likely to buy Thrillers? Do the analysis and report the answer here.**
         Manga and Horror have a -0.30 CC, Manga and Gifts have a -0.00 CC. Manga and Thriller has 0.21 CC.
         So they have no shot of buying from gifts, and an uncommon level likelihood of buying from Horror and Thriller, with Thriller less likely.

   c. **What other category is Fiction most strongly correlated with?**
      Fiction is most strongly correlated with Classics, CC of -0.67

   d. **What other category is Self-improvement most strongly correlated with?**

      ('TEEN', 'SELFIMPROV'), 0.7055040773771842)

   e. **If someone buys cookbooks, what can you tell about them?**
      Following notable CC's:

Fiction -0.52
Manga -0.57
Journals -0.60
Arthist 0.00
Gifts 0.00
Horror -0.00

If someone likes cookbooks, they are more likely to also buy fiction, manga and journals. They are extremely unlikely to buy anything from art history, gifts or horror.

f. If someone buys lots of classic novels, what can you tell about them?

Following notable CC's:

Fiction -0.67
Games -0.62
Horror -0.64
Gifts -0.03
Poetry -0.01
Hairypottery 0.04

If someone buys classics we can tell they are more likely to buy fiction games and horror, but very unlikely to buy gifts poetry and hairypottery

g. If someone buys NEWS, what can you tell about them?

Following notable CC's:
scifi 0.07
ArtiHist 0.01
NonFiction 0.69
Romance -0.66

If someone buys news we can tell they probably will also buy nonfiction and romance and will very likely NOT buy science fiction or art history books.

h. What do you know about people who buy Hairy Pottery?

Following notable CC's
Manga 0.56
Arthist 0.01
Gifts 0.02
Horror -0.55

They may also like Manga and or horror. They probably do not like gifts and art history.

i. What are Thrillers most strongly associated with, or not associated with?

Thrillers are least strongly associated to poetry (-0.01) and Harry Potter (0.00). Most strongly associated with Games (-0.58).

j. What can we infer about people who buy Art & History books?

They only really shop for Art & History books, they have very low cross correlations with any other categories.

3. **If you were to delete three attributes, which would you guess were irrelevant? Why? (1)**

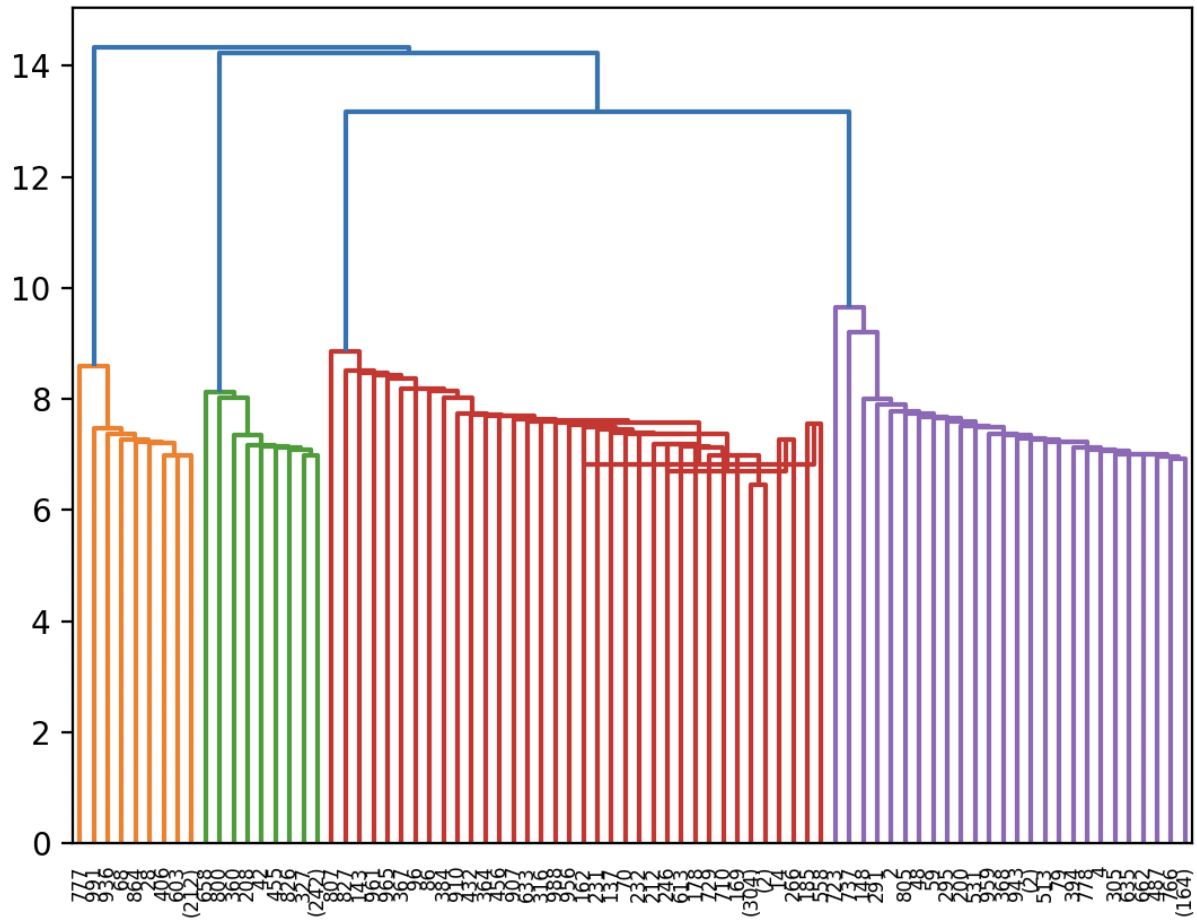
   Arthistory, Gifts, and poetry, which all have extremely low cross correlation with any other categories. All are floating around or near 0.

Part B: Agglomeration:

4:

e. **Based on agglomeration, how many clusters do you think are in the data? Why did you reach this conclusion? Support your guess. Can you support this guess with a dendrogram?**

We think there's ~5 clusters in this data. This is because after 5 the clusters start to really fall apart into ones with way fewer data members. This is supported by our dendrogram because using the technique learned from class, we can count about 5-7 significant clusters.



Discussion Questions – Copy and paste all questions so that you can understand the context of your answers later on, and so that the grader knows which question you are answering.

5. Report the size of each suspected cluster, from smallest to largest size.

[1, 189, 220, 250, 340]

6. Report the average prototype of each of these the clusters.

Average Prototypes of Each Cluster: [np.float64(3.9367)]

7. What typifies each of the clusters? What typical names should we give each of these prototypes? Is there a family group? Is there a gift-giving group? What typifies each group?

The strongest identified clusters were:

A family cluster with baby_todder and teen intersection

An academic type with: Journals, classics, fiction, mysteries

The jaded realist with games, news, harry-potter, non-fiction, horror

Unpretentious bookworm: Romance, non-fiction, horror

Part C: Write a summary and conclusion overall.

This assignment helped us understand clustering better, especially with how to figure out when the clustering tapers off and how many clusters the end result of your data should have by analyzing the cluster members and distances in your clusters. Merging based on the smallest distances helped us tighten our clusters and get better results.

A huge part of this assignment and getting a good understanding of the data was playing around with the covariances of all the attributes and studying relationships between them. It gave us an impression of what the end clusters should look like and helped steer us in the right direction when we faced mistakes. Also, being able to identify non-relevant data like the poetry and gifts attributes made some of the results like singular clusters make more sense.

Using the dendrogram was another good visual queue to gauge how our clustering was behaving. It gave more insight into the merging process which we were able to reverse engineer along with printing out all the merges to a separate file.

It is important to note that there were no alterations made to the explicit dataset to remove outliers or anomalies that could be present. If this did occur alongside other data cleaning processes, we would be able to have more accurate predictions of customers' wants.

Looking at our resulting data, we saw that many of the clusters ended up being of size 1 suggesting that the dataset is spare. We can support this further since there are few commonalities across purchases which could mean the data consists of a unique customer base or book store purchases were more infrequent.