Jacky Chan (jc7311), Ethan Chang (elc6696)

CSCI 420

**2a. Which two attributes are most strongly cross-correlated with each other?**
The two attributes most strongly cross-correlated with each other are NonFict and Romance with a correlation coefficient of -0.80 (absolute value = 0.80). This indicates a strong negative correlation - people who buy non-fiction books are very unlikely to buy romance novels, and vice versa.

**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? 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 strongest with Baby_Toddler at -0.67. Since it's negative that means it's NOT likely to buy Baby_Toddler. Horror and Gifts are also not likely to be bought since they include negative values. However, Thrillers with a value of 0.19 are slightly likely to be bought.

**c. What other category is Fiction most strongly correlated with?**
Fiction strongest with Classics at -0.66 meaning they avoid Classics.

**d. What other category is Self Improvement most strongly correlated with?**
SelfImprov strongest with Teen at 0.69 meaning Teen is most likely to be bought.

**e. If someone buys cookbooks, what can you tell about them?**
Cooking strongest negative with Journals at -0.56, then Manga at -0.53; strongest positive with Classics at 0.40. If negative, that means they are avoiding Journals and Manga while focusing on Classics.

**f. If someone buys lots of classic novels, what can you tell about them?**
Classics strong negatives with Fiction (-0.66), Horror (-0.65), Romance (-0.52); positives with NonFict (0.60), Mysteries (0.47). Classic novel buyers buy NonFict and Mysters, while avoiding Fiction and Horror novels.

**g. If someone buys NEWS, what can you tell about them?**
News strong negatives with Romance (-0.67), Teen (-0.56), Horror (-0.56); positive with NonFict (0.68). News novel buyers buy NonFict while avoiding romance, teen and horror novels.

**h. What do you know about people who buy Hairy Pottery?**
HairyPottery strongest with Manga at 0.58; negatives with Horror (-0.52), Romance (-0.50), Cooking (-0.44). HairyPottery are fans of Manga but dislike Horror and Romance novels.

**i. What are Thrillers most strongly associated with, or not associated with?**
Thrillers strongest positive with Mysteries at 0.48; negatives with Games (-0.55). Thriller novel buyers enjoy reading Mysteries while dissociated with Games novels.

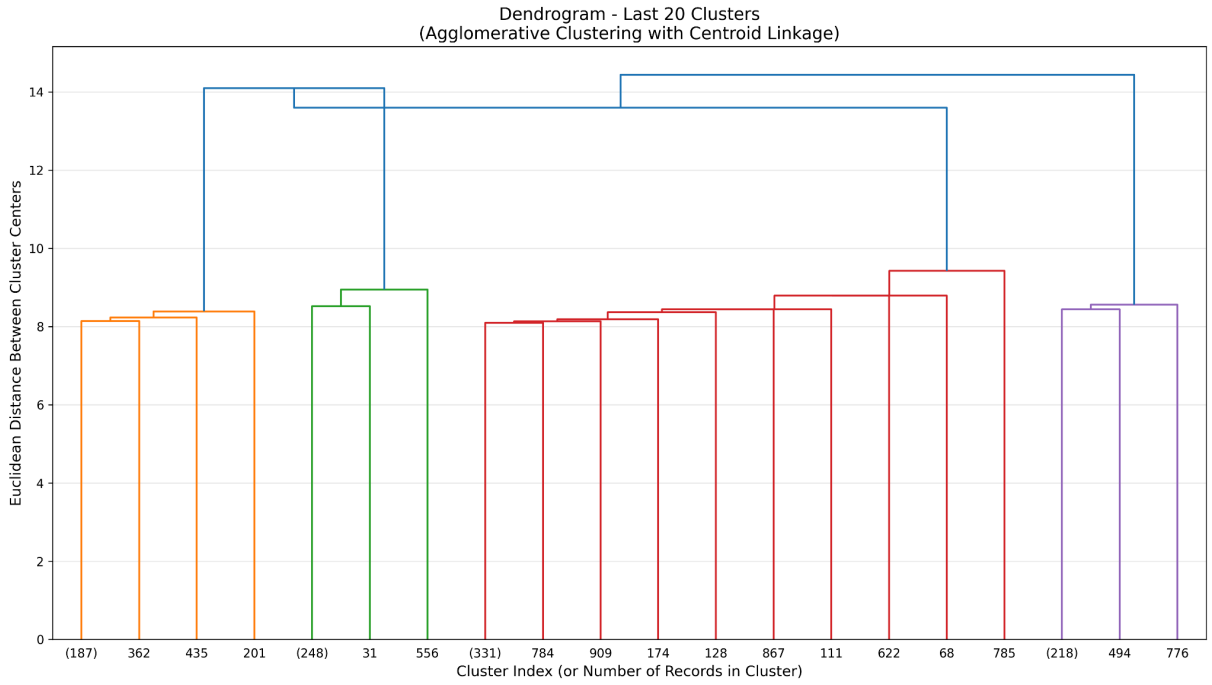
**j. What can we infer about people who buy Art & History books?**
Art&Hist shows near-zero correlations with everything (-0.02 to 0.02). Art&Hist are considered independent since they have no opinions.

**3. If you were to delete three attributes, which would you guess were irrelevant? Why?**
Art & History, Poetry, and Gifts because they don't vary meaningfully with other purchases. The data doesn't help identify distinct customer groups or patterns. They don't help us predict or understand purchasing patterns for other book genres.

**4d. Note: At each step of clustering, two clusters are merged together.**
**Track the size of the smallest of the two clusters that are merged together.**
**There are questions about this later. Write down the size of the smallest cluster in the last 20 merges. For example, if we merge a cluster of size 30 with a cluster of size 10, you remember that a 10 was merged in. Cluster to completion. Record and report the size of the last 10 smallest clusters merged.**

LAST 20 - [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 190, 340, 220]
LAST 10 - [1, 1, 1, 1, 1, 1, 1, 190, 340, 220]
What this tells us is that the first 17 of the last 20 merges involved outliers (size 1) being absorbed into main clusters. Only the final 3 merges combined substantial clusters, indicating 4 stable groups existed before the final consolidation.

4e. **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?**
4 clusters because of the last 10 smallest sizes. The dendrogram shows 4 distinct colored clusters.




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

Cluster 3: 190 members
Cluster 1: 220 members
Cluster 2: 250 members
Cluster 0: 340 members

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

**Cluster 0:**

High: Teen (7.94), SelfImprov (7.96), Classics (6.94), Mysteries (6.05), Cooking (5.12), Romance (5.13)
Low: Games (0.99), Art&Hist (0.97), Journals (1.45), Horror (2.01)

**Cluster 1:**

High: NonFict (9.03), Baby_Toddler (8.52), Classics (7.11), Sci-Fi (5.85), News (5.27), Games (5.05)
Low: Romance (0.50), Horror (0.00), Teen (0.94), Manga (1.01), SelfImprov (1.04)

**Cluster 2:**

High: Horror (9.00), Romance (8.48), Baby_Toddler (8.40), Games (8.09), Sci-Fi (6.05), Teen (5.92)
Low: Manga (0.22), NonFict (1.03), News (1.04), HairyPottery (0.52)

**Cluster 3:**

High: Fiction (7.04), Sci-Fi (6.87), Teen (6.93), HairyPottery (6.57), Journals (6.54), Manga (5.93)
Low: Cooking (1.03), Baby_Toddler (1.33), Classics (1.76), News (1.95)

**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?**

Cluster 0 consists of middle-aged parents (35-55) with teens, focused on personal growth
Typified by: High self-help (7.96) + teen books (7.94) + classics (6.94) + mysteries (6.05)

Cluster 1 consists of parents with young children (0-5 years), education-focused
Typified by: Extreme non-fiction (9.03) + baby/toddler books (8.52), no horror

Cluster 2 consists of young adults (20-35) who love escapist fiction, likely buying baby gifts
Typified by: Horror (9.00) + romance (8.48) + games (8.09), but also baby/toddler (8.40)

Cluster 3 consists of teens/young adults (15-25), imaginative fantasy enthusiasts
Typified by: Fiction (7.04) + sci-fi (6.87) + Harry Potter (6.57) + manga (5.93) + journals (6.54)

9. Write a conclusion about what you learned overall.

**(Jacky)**

We gained a much deeper understanding of how correlation analysis and hierarchical clustering complement each other in segmenting customers based on their underlying similarities. Through this process, we learned how to interpret relationships between variables and how those relationships can shape the quality of clustering results.

A key lesson from the correlation analysis was the importance of identifying which variables contribute useful information and which ones add noise. We learned how to interpret strong positive and negative correlations as indicators of customer tendencies, and how near-zero correlations can reveal irrelevant attributes that weaken clustering accuracy. This taught us that effective feature selection can be just as valuable as the clustering itself because it improves both efficiency and interpretability.

Working on agglomerative clustering helped me understand not just how the algorithm functions conceptually, but also how to implement and optimize it in practice. Building it from scratch forced us to think carefully about time complexity, data structures, and mathematical precision. We learned that a naive implementation can be extremely slow (had an initial O(n^3) time complexity) due to repeated distance calculations, and that using data structures like priority queues can dramatically improve performance (improved to O(n^2logn). At the same time, we discovered that optimization must be balanced with correctness, shortcuts that simplify computation can compromise the integrity of the results. This reinforced the idea that algorithmic efficiency and statistical accuracy must go hand in hand.

Beyond the technical side, we learned how to interpret clustering results in a meaningful way. Visual tools like dendrograms showed us how to determine natural groupings by observing patterns in merge distances. This experience helped us understand how to choose an appropriate number of clusters and how to evaluate whether they make sense both statistically and contextually. We also learned that clustering is inherently exploratory, there is no single “correct” answer, and interpretation often depends on the business context or research goals.

**(Ethan)**

We learned a lot about how agglomerative clustering functions and how it processes data. It’s important to pay attention to the small details of how this algorithm operates and computes clusters. For example, including the Guest ID can significantly impact the results, so it’s essential to handle and clean the data properly before executing the algorithm. Not only did we learn about agglomerative clustering, but we also gained a deeper understanding of how it begins with each record representing its own individual cluster. As the clusters merge, we explored how linkage methods and distance metrics play a major role in determining how groups form, which

ultimately leads to the dendrogram results. We also recognized the importance of runtime optimization when handling large datasets to ensure faster and more efficient computations.

Implementing algorithms like this can be very helpful for problem-solving and allows us to explore new Python tools and packages such as pandas, NumPy, SciPy, and matplotlib. Pandas was used for managing data frames, NumPy for performing mathematical operations on arrays and vectors, and SciPy for creating dendrograms. I found it interesting how we could integrate these packages with our own functions, such as Euclidean distance, giving us hands-on experience applying math to vectors.

Some challenges we faced included working with new data structures like heapq, which implements a priority queue, and learning how to handle data efficiently. We were also new to dendrograms, so it took some time to understand how to implement them correctly. From the dendrogram, we learned how to interpret the data and the resulting plot. It showed how certain customers naturally grouped together and revealed purchasing patterns such as how some book genres attract similar buyers, allowing businesses to gain insights into what products to market next.