1. Why do we use 10 visits instead of just keeping records of every single visit?

We use 10 visits instead of just keeping records of every single visit to reduce the noise and capture meaningful patterns. Individual visits would contain high variability and random purchasing decisions, but averaging over 10 visits reveal consistent preferences and shopping habits. This aggregation also reduces dataset size to a manageable scale while maintaining enough information to identify customer segments.

2b. Often students get a covariance matrix that is 21x21 instead of 20x20! What did they forget to do?

They forgot to remove the ID column.

5. Why does this tell you about the attributes? Which attributes are most important? Which can be ignored? Explain your answers.

The attributes vary together and contribute most to the variance in the dataset. Large absolute values in the eigenvector component indicate attributes that strongly influence that component. In the first eigenvector, the most important attributes are horror, romance, and games which have the largest positive values. Classics and nonfict also contribute. This component seems to separate entertainment reading (hence the horror, romance, games) from the more serious ones (classics, nonfict).

In the second eigenvector, the most important attributes are baby_toddler, teen, selfimprov, sci-fict, and nonfict. This component seems to distinguish a time of life purchases (baby_toddler, teen).

Attributes that can be ignored are art&hist, gifts, and poetry which have values near zero in both eigenvectors.

8. What do they tell you about each cluster? Do they tell you anything?

The cluster centers show four distinct segments.
Cluster 0 (5.31,1.91) demonstrates moderate positive values on both components
Cluster 1 (-3.83, -8.55) demonstrates strong negative on the second component
Cluster 2 (-1.74, 7.17) demonstrates strong positive on the second component
Cluster 3 (12.20, -1.04) demonstrates strong positive on the first component

9. What prototype amounts do you get back? What are the relative amounts? One of the clusters might buy a lot of Horror. What else do they buy? There is loss of information in this process. Do you notice anything odd? Did anything go negative?

Cluster 0 (Romance/Horror buyers): High Romance (2.36) and Horror (2.71), moderate Games (1.19) and Teen (1.44). Negative NonFict (-2.50) and Classics (-1.97)

Cluster 1 (Baby/NonFiction buyers): High Baby_Toddler (3.42), NonFict (3.74), and Sci-Fict (1.50). Negative SelfImprov (-3.55), Teen (-4.16), and Romance (-3.35)
Cluster 2 (Teen/Self-Improvement buyers): High Teen (2.91) and SelfImprov (3.08), Mysteries (1.98). Negative Baby_Toddler (-3.57) and Games (-2.74)
Cluster 3 (Entertainment/Games buyers): Very high Horror (6.01), Games (4.37), and Romance (4.18). Negative NonFict (-4.28) and Classics (-4.63)

Many values went negative. This is odd because the negative values result from information loss during PCA dimension reduction. Using 2 of 20 dims, the reconstruction is incomplete and can possibly produce impossible negative purchase counts.

10. If you projected the data onto all 20 of the eigenvectors, why would this not help you with your data understanding?

Projecting all 20 would reconstruct the original 20 dimension data without dimensionality reduction which would defeat the entire purpose of PCA. With 20 dimensions we'd have too many variables to visualize or interpret easily.

11. Conclusion

This assignment demonstrated the usage of PCA and K-means clustering. Using the 20 book category attributes, we used PCA to reduce dimensionality while retaining most variance. Projecting onto the first two principal components revealed clear cluster structure that would be impossible to visualize in the original 20 dimensions.
The K-means algorithm identified four distinct customer segments with interpretable patterns like cluster 0 (romance/horror), cluster 1 (baby/nonfiction), cluster 2 (teen/self-improvement), and cluster 3 (entertainment/games). The cumulative eigenvalue showed that dimensionality reduction is effective for the dataset with the first two components capturing substantial variance.
The reprojection revealed limitations of using two components. Negative values appeared in the reconstructed data, highlighting information loss. However, this trade off is acceptable given the interpretability that was received.