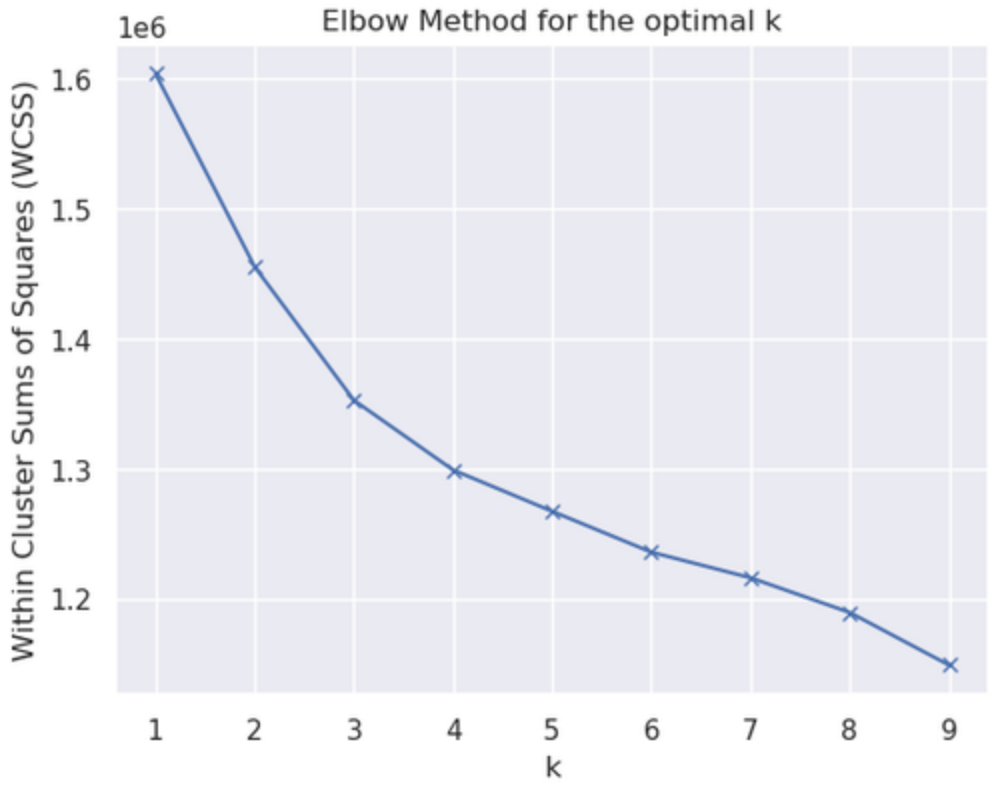

Figure 1: Elbow Method plot indicates that the optimal number of clusters is k = 4. Plots were made by looping ten times over the K-Means model fit function by utilizing the protein data. The optimal value was found at the elbow-shaped inflection point, where the Within Cluster Sums of Squares (WCSS) value dropped to about 1.3. Protein data acquired from BRCA proteomic data.

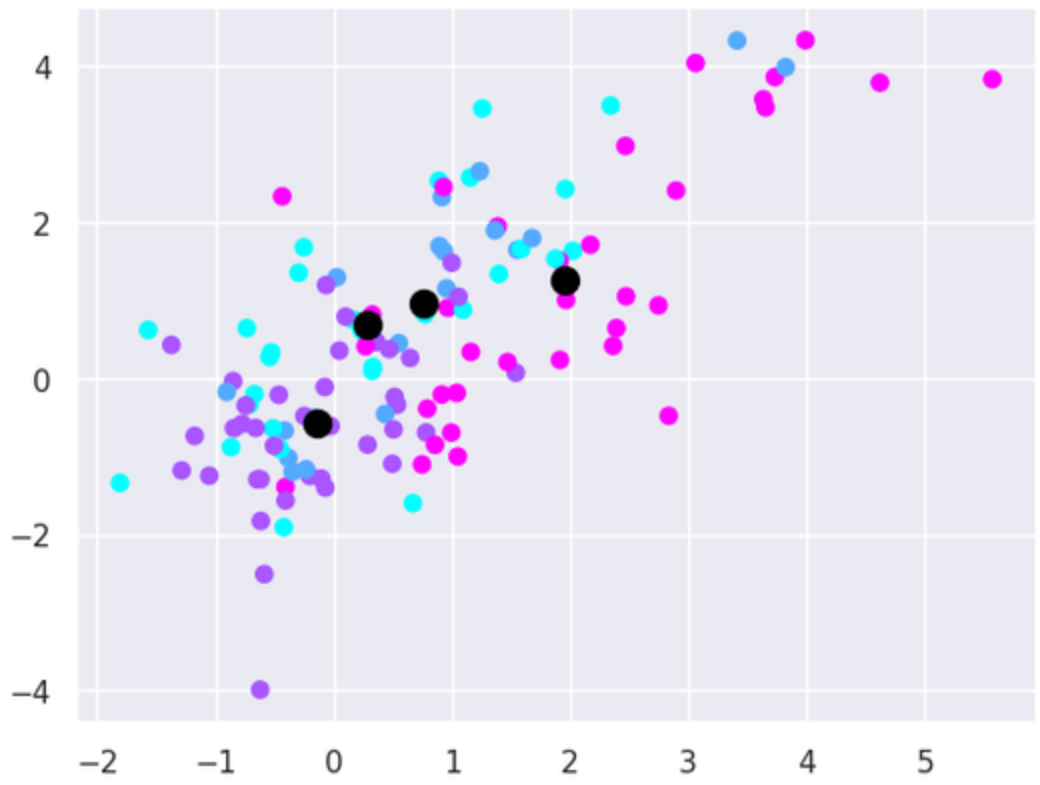

Figure 2: K-Means clustering on protein data indicates four centroids. A total of four clusters were defined from the elbow plot, and the respective centroids appear to be hard to distinguish within the clusters. The significant overlap and lack of distinct clusters implies a lack of correlation within the protein data. Protein data acquired from BRCA proteomic data.

PCA graph was not able to be acquired

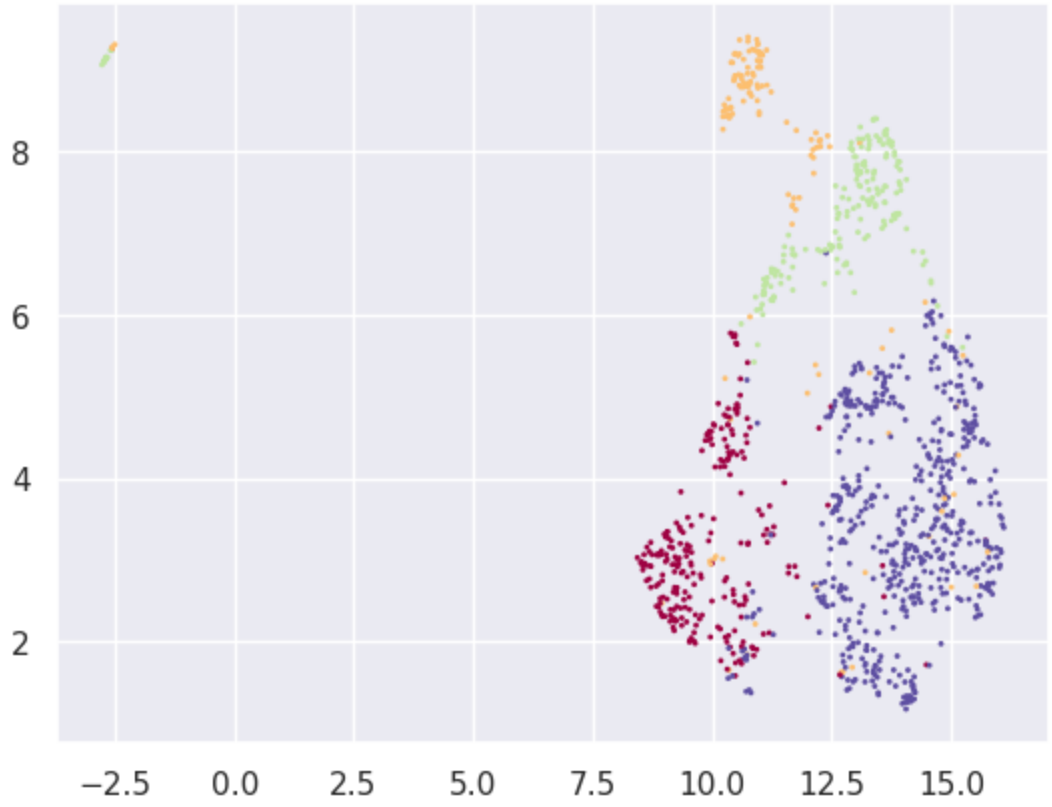

Figure 4: UMAP for RNA data, designated by 4 K-Means Clusters. Four distinct clusters are apparent, with minimal overlap across the four clusters. The orange group appears to span the largest spatial region, and there are two distinct outlier groups in the upper left region from the orange and green clusters.

//Lack of axis labels is a function of technical issues. //

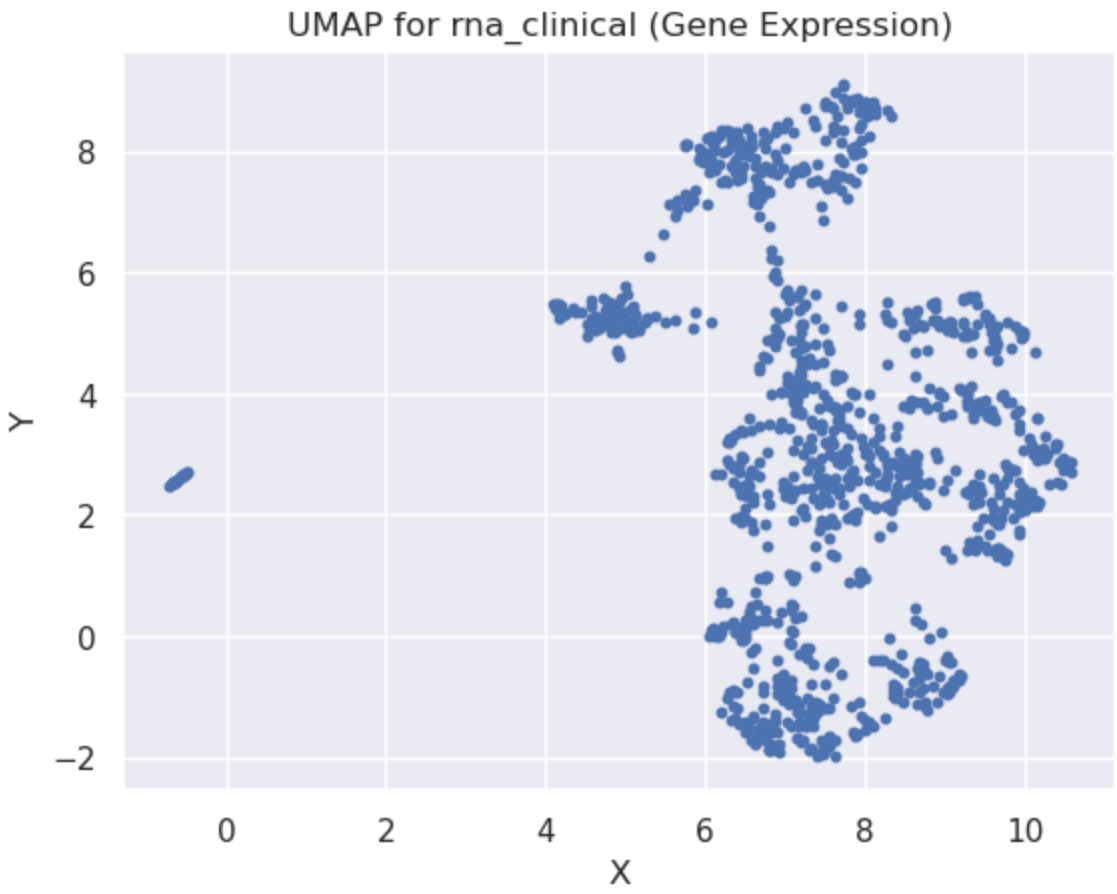

Figure 4: UMAP for RNA data, designated by gene expression. There appears to be some broad regions of clustering, but the data does pinpoint clusters in practice. This may indicate potential error, as the dispersion is somewhat reminiscent of the K-means clustering, which had requisite clustering identification.

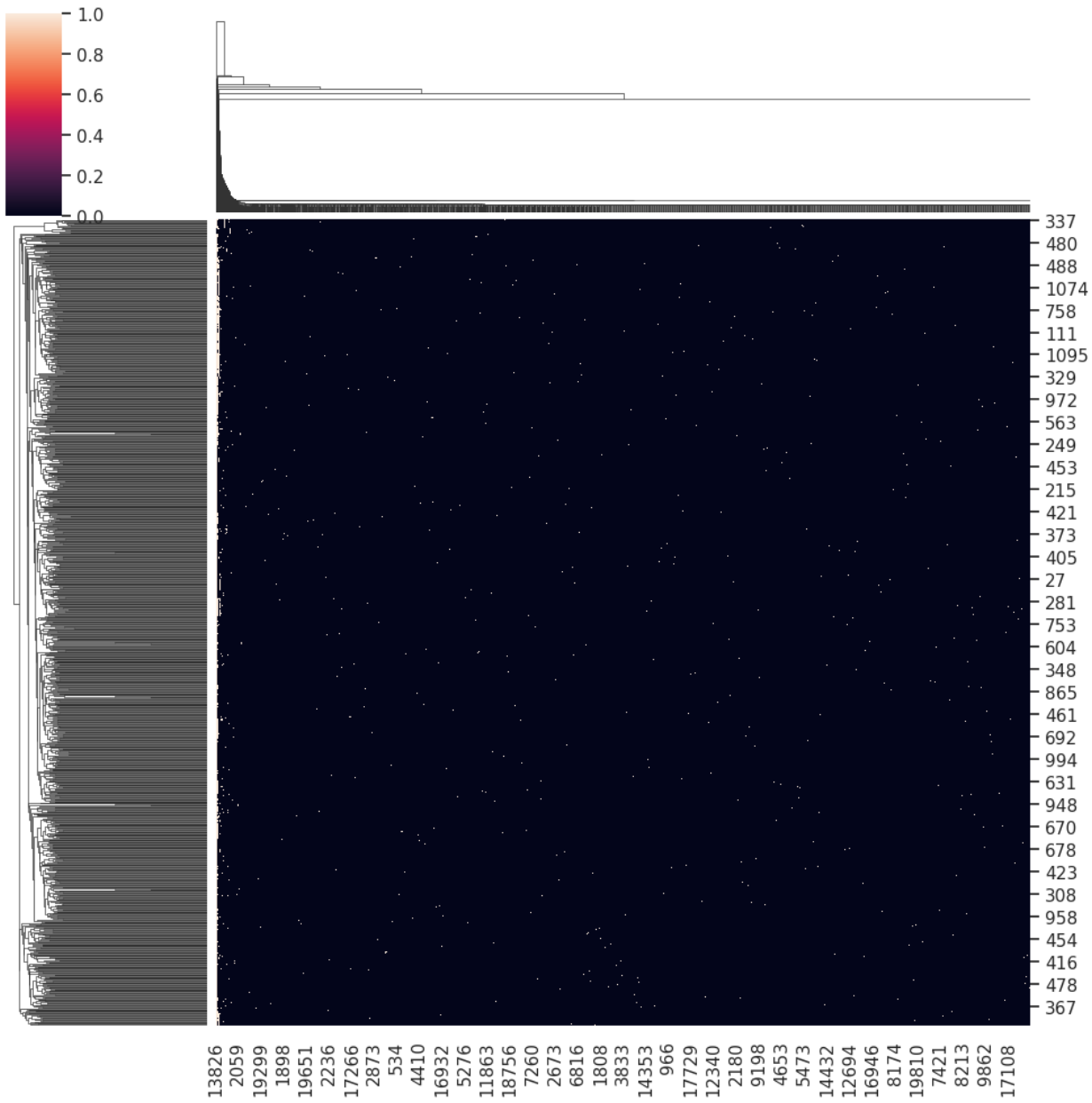

Figure 5: Hierarchical clustering of RNA data indicates minimal relationships between features. The analysis shows that there is limited relatability between the various features within the plot, as the majority of the inter-relationships demonstrate about zero relatability.

5.

a) **The biggest thing that surprised me within my figures was the relative variability between each of the graphs. The exact code that I used in one time span led to a different result in another time span, and it even simply stopped outputting values after a while. When the elbow plot did work, I was surprised how obvious it was to delineate the correct number of clusters at the elbow.**

b) **I would specifically want to understand the edge cases, as there seems to be a bit of overlap between each of the clusters. More astutely defining the boundaries would limit variability and potential long-term ramifications in the ideal treatment and prognosis of individuals at the edges. I could get this information simply by analyzing the parameters and taking a sample of colliding edge cases to define the effectiveness of clustering.**

c)