Diego Valdes

IST 707

Home Work 4

4/23/2019

**Introduction**

It is human nature to know. To know the answers to questions, to know the events of history, to know the behaviors of the animal kingdom, to know as much as possible about as much as possible. The nature for knowledge can be seen in many examples of society. Museums, documentaries, scientific journals, even religion are all ways that humans seek to gain knowledge from the unknown. A recent example is the unveiling of a picture of a black hole. The resources and technology to achieve creating a single image are beyond compare, but the payoff was knowledge. We know what it looks like now.

The Federalist Papers are comprised of eighty-five essays explaining to the citizens of New York why they should ratify the newly drafted United States Constitution. The writers included Alexander Hamilton, James Madison, and John Jay, but were published under the pen name “Publius”. Their authorship remained secret for thirty-one years. One problem arose from this reveal; a series of the essays, 11 total, had disputed authorship. Madison and Hamilton both claimed authorship of the works. It has been a mystery to historians the true authorship of the essays. They wanted to know.

Authorship attribution has become a more widely used technique. Simply stated, a writer’s use of words and the frequency of use is like a fingerprint. Recently, it was used to attempt to determine a similar conflict between Beetles songs written by either John Lennon or Paul McCartney. Like the Federalist papers, each song writer had written their own songs and they had collaborated on many works. This has been a common source of argument among fans, just as the authorship of the Federalist essays has been among historians.

**Analysis**

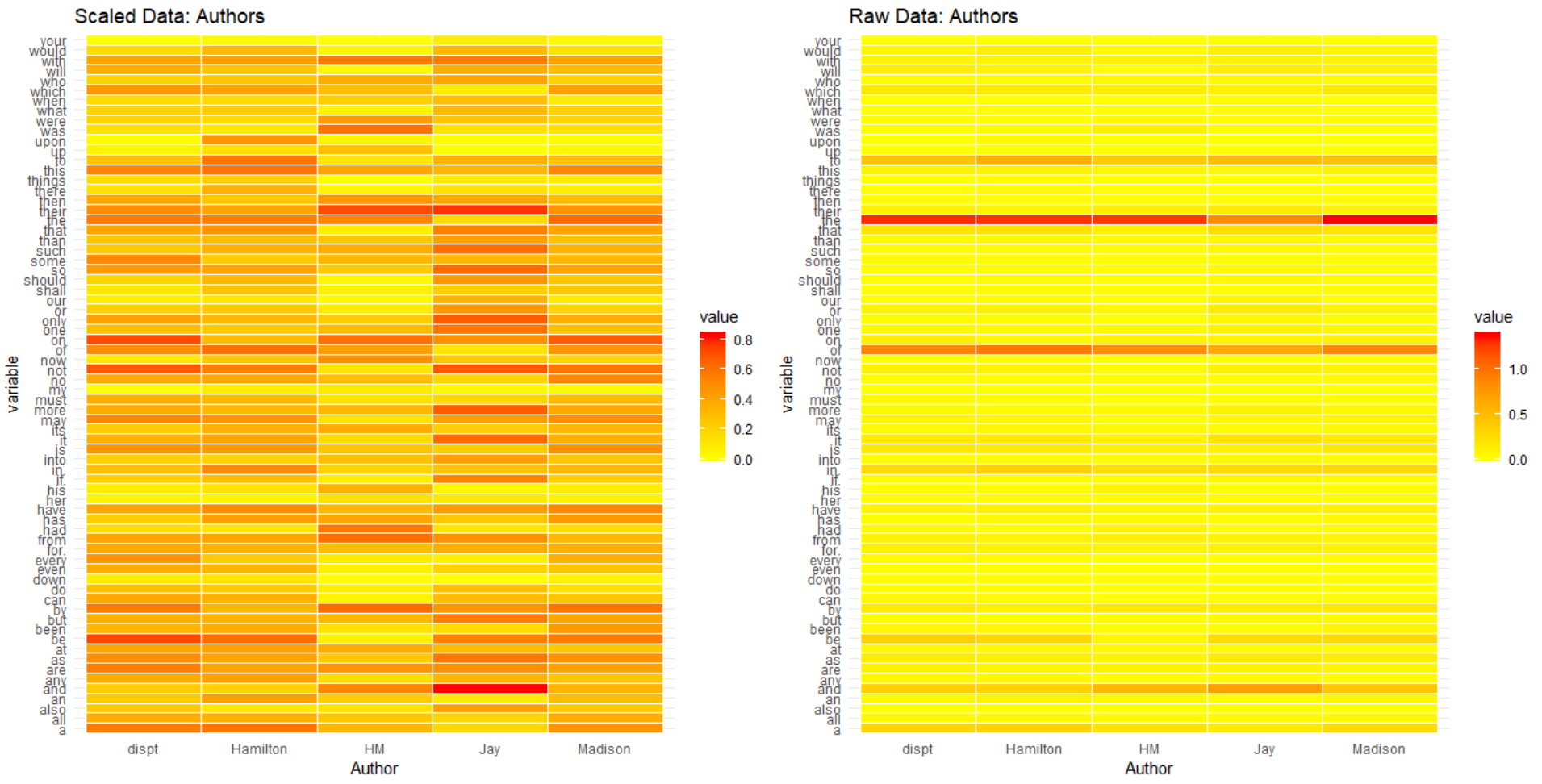
**About the Data**

The dataset is comprised of 85 essays, 74 with identified authorship and 11 disputed, along with the frequency percentage of function words (a, the, then,…) as it appeared in the essay. If the word ‘a’ appeared three times in an essay of one-thousand words, it’s frequency would be 3/1000 = 0.3%. There are a total of 70 function words. Authors are represented by their surname and the essays are file names, not the names of the essays (Hamilton, Hamilton\_fed\_28.txt).

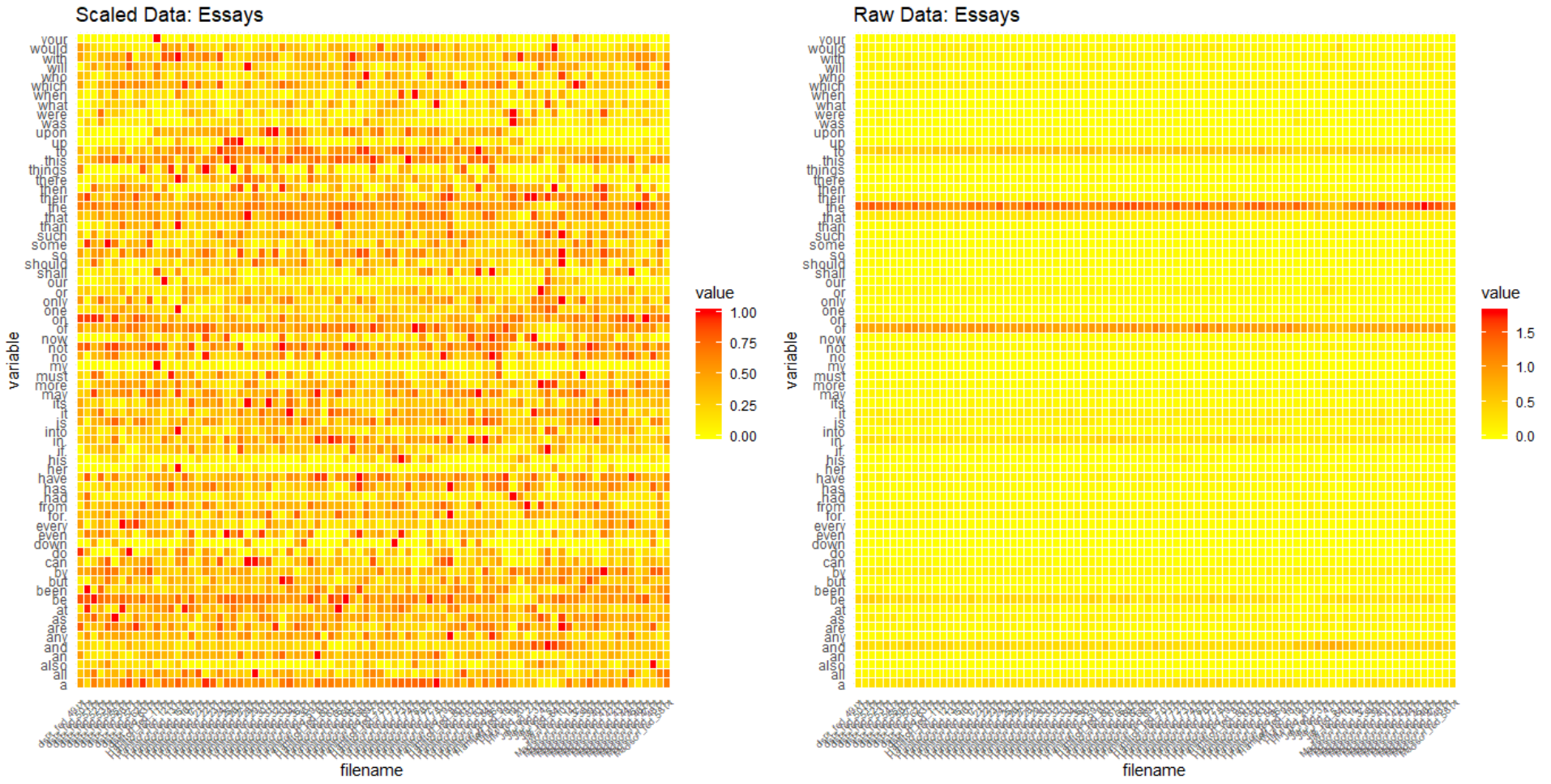
The dataset contained no NA or missing values. To prepare it for analysis, the function word columns were converted to numeric values and the cols with ‘author’ and ‘filename’ converted to factors.

Because each function word value is a percentage of frequency, a second dataset was created where all these values where scaled. Moving forward, all exploration, models, and analysis of the dataset would be completed on each set to examine which produced the best results.

When comparing the scaled and unscaled datasets visually, it becomes much easier to see the frequency distribution of words among the authors in the scaled plot. It is especially interesting to compare the columns of each author with the disputed works. It is also of note the similarities in usage that Hamilton and Madison share, but how dissimilar either author is when their efforts are combined.

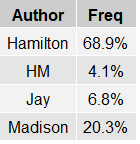


As with the authorship comparison, comparing essays is more visually distinct between the essays in the scaled dataset.



While the scaled dataset plots provide a more distinct portrait of the author or the essay, the unscaled plots do also provide relevant information. In both the author and essay plots, the darker line spanning across the plot are the words, ‘the’, ‘of’, ‘and’, and ‘to’. Across all the works and authors, these three words have the highest frequency, but because of how similar the unscaled dataset is when compared to the contrast of authorship and essays provided by the scaled dataset, it is unclear as to how valuable this will be when it is used to cluster for authorship determination.

Authorship frequency is not evenly spread.

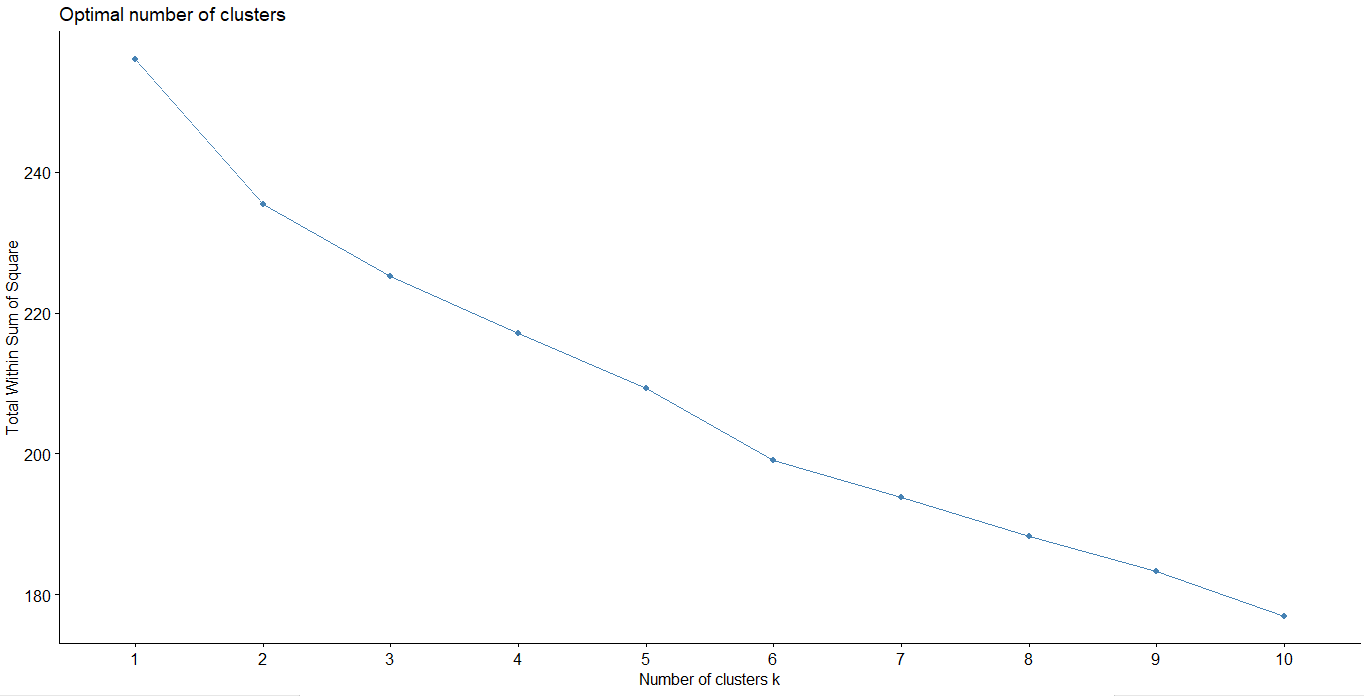
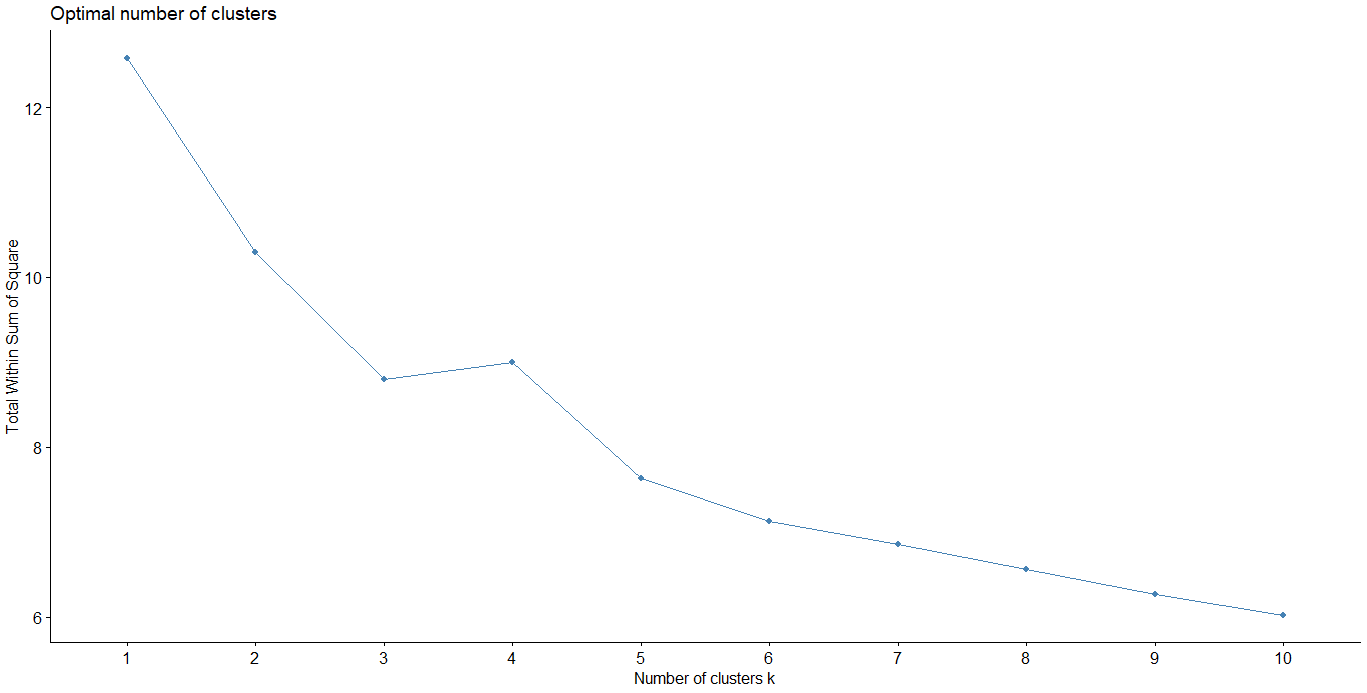


**Models**

Both the scaled and unscaled datasets were used in creating models. In total, six models were created using k-Means, HAC, and EM approaches. The ideal number of clusters for each model was four, one for each other authorship. The columns ‘author’ and ‘filename’ were not included as variables to design the clusters.

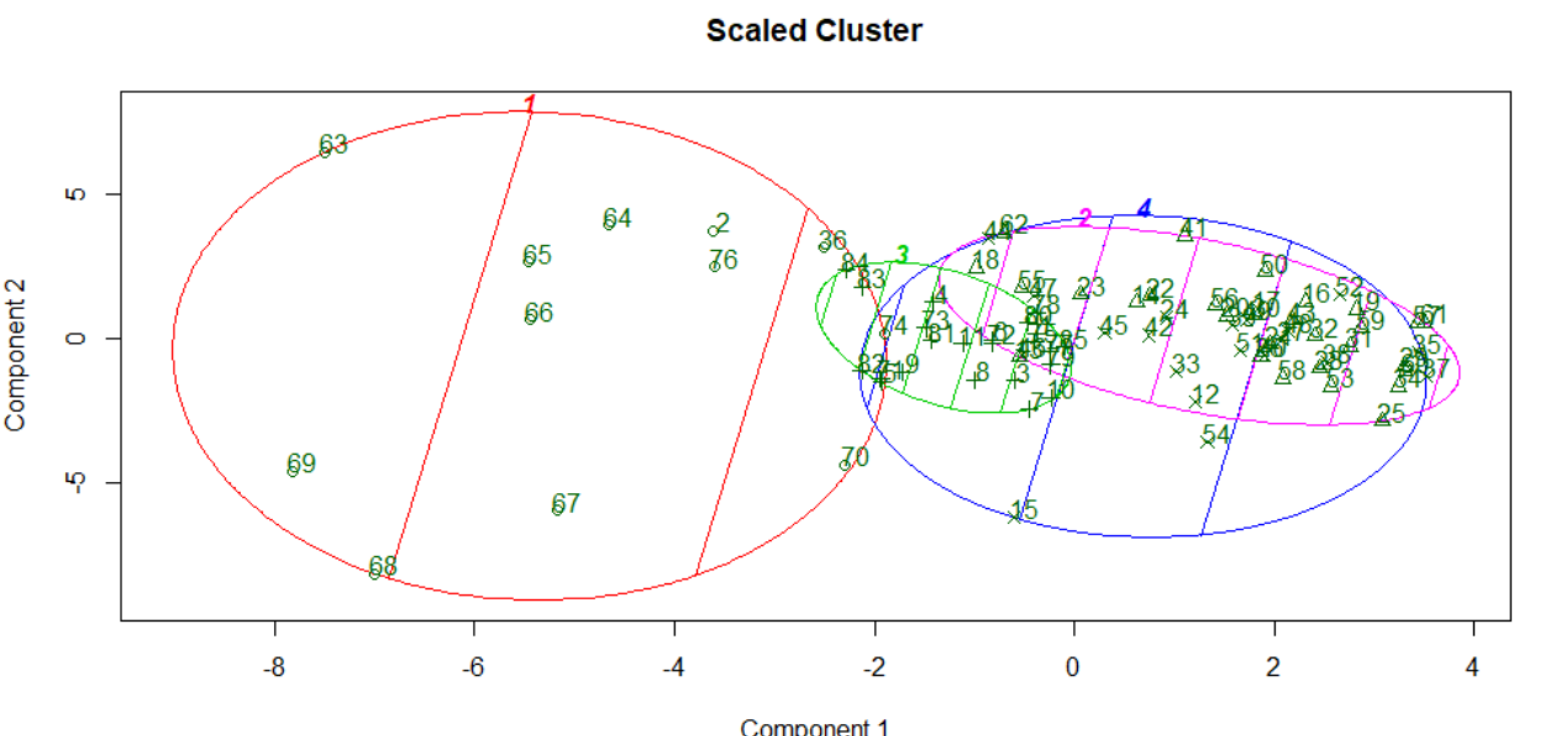
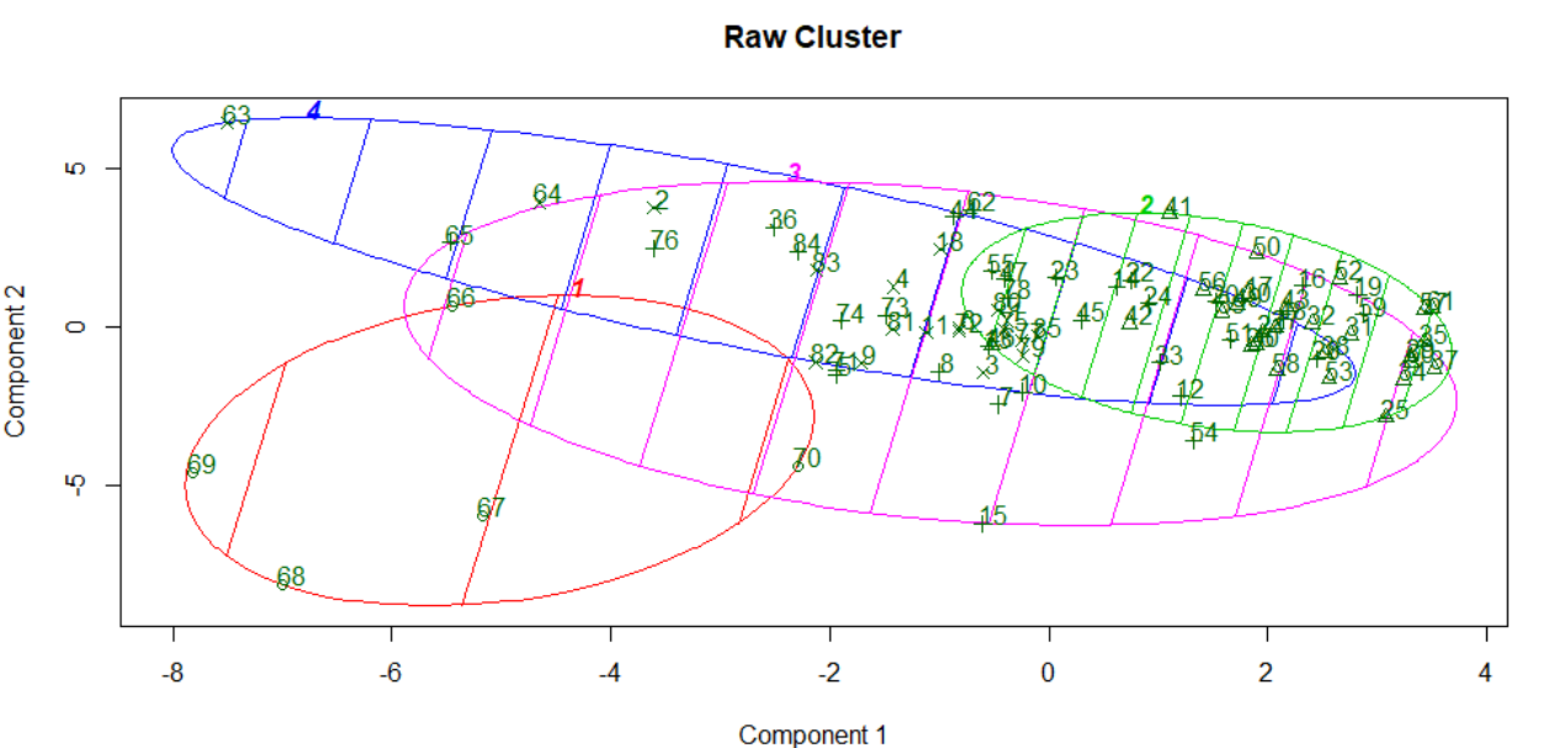
**K-Means**

Although the cluster number was determined by authorship, the within sum of squares method was still used to determine if four was the optimal number based on the algorithm.

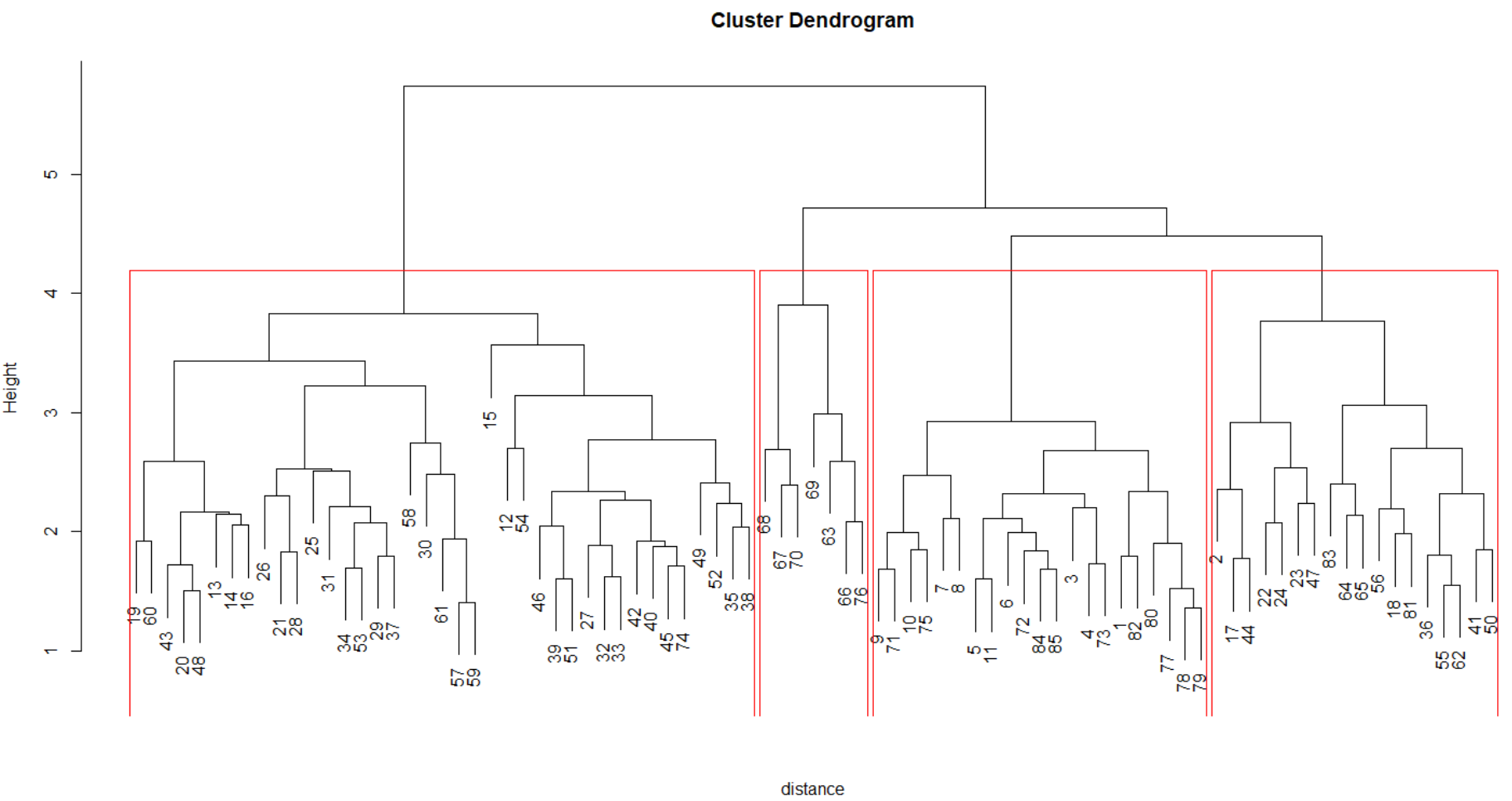
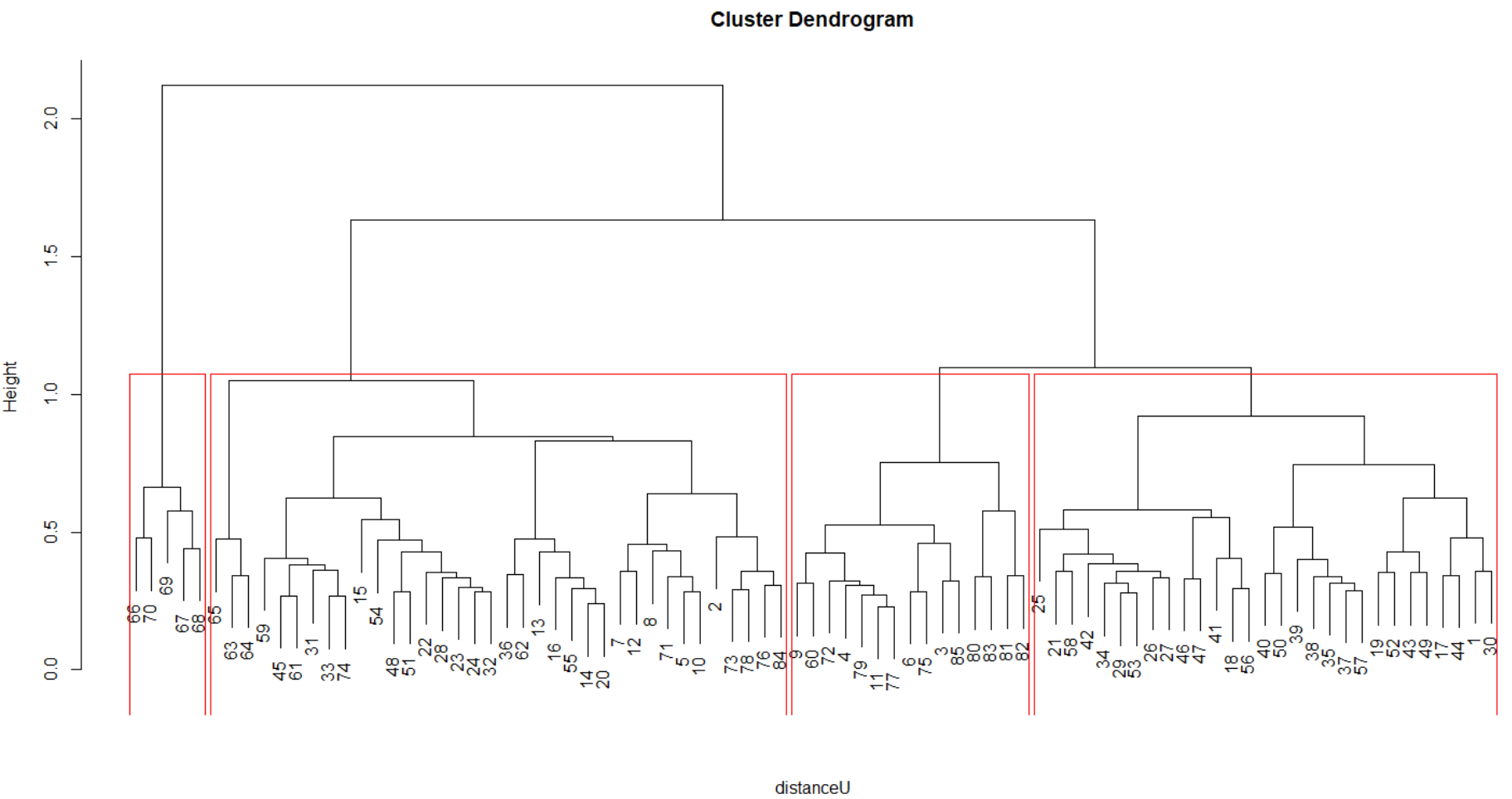
In the scaled data plot (left) the optimal clusters aren’t as defined as the unscaled data plot(right). Four is the optimal cluster in the raw dataset. If examining the scaled data without the knowledge of the purpose, five would have been the cluster option that seems most appropriate.

K-Means algorithm was ran with both datasets with a k value of four. Clusters 1, 2, and 4 represent Jay, Hamilton and Madison, respectively. The scaled cluster identifies HM (Hamilton & Madison authorship) as cluster 3, while the raw plot does have a cluster 3, it is not as defined as in the scaled version. The raw cluster 3 places most of HM authorship in with Hamilton.

**Hierarchical Clustering**

To use the Hierarchical (HAC) algorithm (pvclust::hclust), a distance measure (vector) is required. Euclidean distance was taken for both scaled and raw data to pass into the algorithm. The result is a tree that can be cut to the desired number of clusters. A similarity between both sets of results is that Jay’s writings are all on their own branch, but do include writings of other authorship; 4th branch in both results. It seems his writings are much more unique and easier for the algorithm to cluster together, even if some of Hamilton and Madison’s writings have made it through. This is also true in the k-Means algorithms. Where the lines start to blur is with the other two author’s and their combined writings.

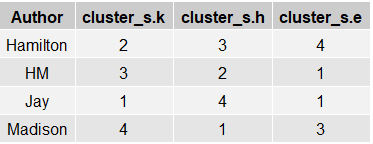
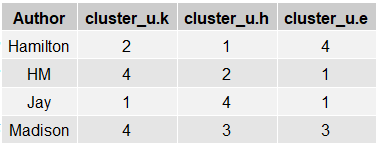
 

**Expectation Maximization**

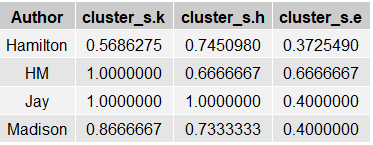
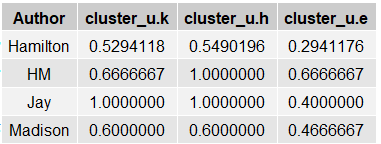
Just as k-Means, the Expectation Maximization algorithm needs a predefined number of clusters. Both results provided an overlap in cluster authorship. Scaled results had much of HM essays clustered with Hamilton’s other writings, and the raw results had HM clustered with Jay’s writings. Unlike the previous clustering models, Jay’s writings did not stand on their own in any version. The cluster scale was numbered 0 – 3, to stay consistent with the other models, 1 was added to the results to provide a scale 1 – 4.

**Results**

Each algorithm produced a vector of cluster assignments for each row (essay). To define which cluster is associated most with an author, the dataset was aggregated by author and the mode was found. The mode for each author dataset was determined to be the author’s cluster. As an example, in the scaled k-Means dataset, Hamilton’s writings appear most in cluster 2, therefore, cluster 2 are writings most associated with his work. This method of evaluation was applied to each model and author. The results for scaled models (left) and raw models (left) are listed below. It is of note, that not every model clearly defines each author within their own cluster.

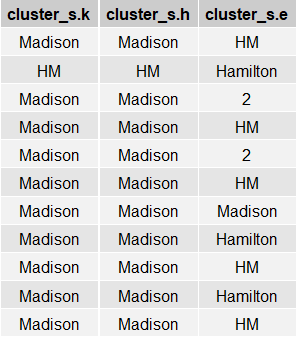
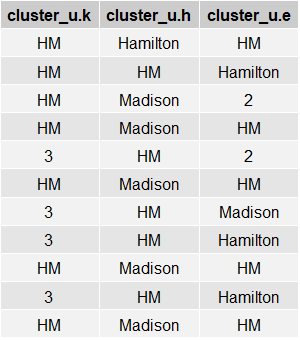
 

With each cluster being assigned to an author, the percentage of writings within the cluster belonging to the author was determined. It is important not to consider this as a measure of how ‘right’ the model is at predicting authorship. The disputed works are included in the building of the model with the idea that these works would be clustered within the correct authorship. Within most of the models, Jay’s writings were most distinguished enough to be in their own cluster.

Each model assigned the disputed essays to one of the clusters. This is the ‘prediction’ of authorship by the model. Raw models of k-Means (cluster 3) and both models of EM (cluster 2) didn’t produce a clear cluster assignment for authorship. These clusters had essays spread across multiple authorships. Interesting observations in the predictions include:

* The vast difference produced by scaled and raw data. HM comprises 4% of the observations, but is the most predicted authorship in unscaled results (17/33: 52%) compared with scaled (7/33: 21%).
* Hamilton has the most essays in the dataset (68.9%) and he is the least predicted authorship across every model
* Scaled results for k-means and HAC produced the identical predictions.
* The trend of Jay’s writing being much more distinct appears in the predictions as well. He is never part of any of the predictions, and historians do not dispute him as a possible author. In the clusters that were not able to be assessed to an author, Jay’s writing appeared in 2, both EM models.

Based on the performance of each model to be able to place essays in clusters by authorship, k-Means and HAC using scaled data, and HAC using raw data performed the best. They also share many of the same results, sharing 6 predictions. When looking at the remaining predictions, it is much more difficult to determine which model to trust.

**Conclusion**

It is often said that writers have a voice. There is a way they sound, command the language, and use punctuation to control the beats and flow of their work. With the authorship of these essays, the work of Jay stands out as the most unique. It is more often then not identified by algorithms and isn’t clustered with the works of Hamilton and Madison. He is also the only of the three writers that doesn’t share a byline for any of the works.

When looking at Madison and Hamilton, they have shared authorship on three essays. Their works are often part of the same clusters. The most reliable models all share the commonality of giving sole or joint authorship in ten out of the eleven essays. Only one model grants Hamilton sole authorship of one essay. With this in mind, it is easy to make the assessment that Madison is at least a joint author of all of the disputed essays.

There is, however, another possibility that these results do raise. Because both men’s works were often clustered together, and they did work jointly on a few essays, there may be the chance that they edited each other’s work. Considering that both men had the same goal, to persuade New Yorkers into ratifying the Constitution, it isn’t unreasonable to believe that when one writer completed a work, he shared it with the other for advice, comments, general editing. If this is the case, it would explain why their works are often clustered together. Perhaps their unique voices began to merge with each collaboration, to the point that each writer grew and adapted based on the other’s feedback.

There is no way of validating this, as both Hamilton and Madison claimed sole authorship and provided no insight to whether works were shared before publishing. Make no mistake, as any professional writer will testify that before publishing, their work undergoes an editing process, where multiple voices can be inserted into the original piece.