IST 707 Applied Machine Learning
HW4: Use Clustering to Solve a Mystery in History
Name: Jainish Savaliya

- In this assignment first I loaded all the necessary libraries for the clustering such facto extra and grid extra even though we can do it with out visualization we can get the proper understanding by the cluster visualization.

- So, I read the CSV file and distributed the data in main three sets like Hamilton, Madison and disputed.

- After that, I unlabeled the data and removed author names and filenames from the csv. To get the unlabeled numeric values. I scaled the data to prevent it from taking the arbitrary units. After that I measured the Euclidian distance using factoextra library and performed some distance visualization.

- I performed K means clustering on the scaled data and assigned them according to their word of appearance. Printed the centroids & tried some clustering variations by choosing different number of iterations and centers & tried to get better squared sum of errors.

- As per asked in the assignment the main goal is to find out who wrote the disputed papers, so I divided the data into two clusters and created the gap stat and visualized the clustering.

- After creating the table of gap stat I got that in the cluster one there was no distribution as all the data points in the cluster one which in my case was 46 was written papers by Hamilton and in the cluster two there were 11 disputed papers, 5 Hamilton and 15 Madison. So here number of Hamilton papers is very low so that Is considered an error so we can say the disputed papers were written by Madison .
  Conclusion: Disputed Paper were written by Madison.