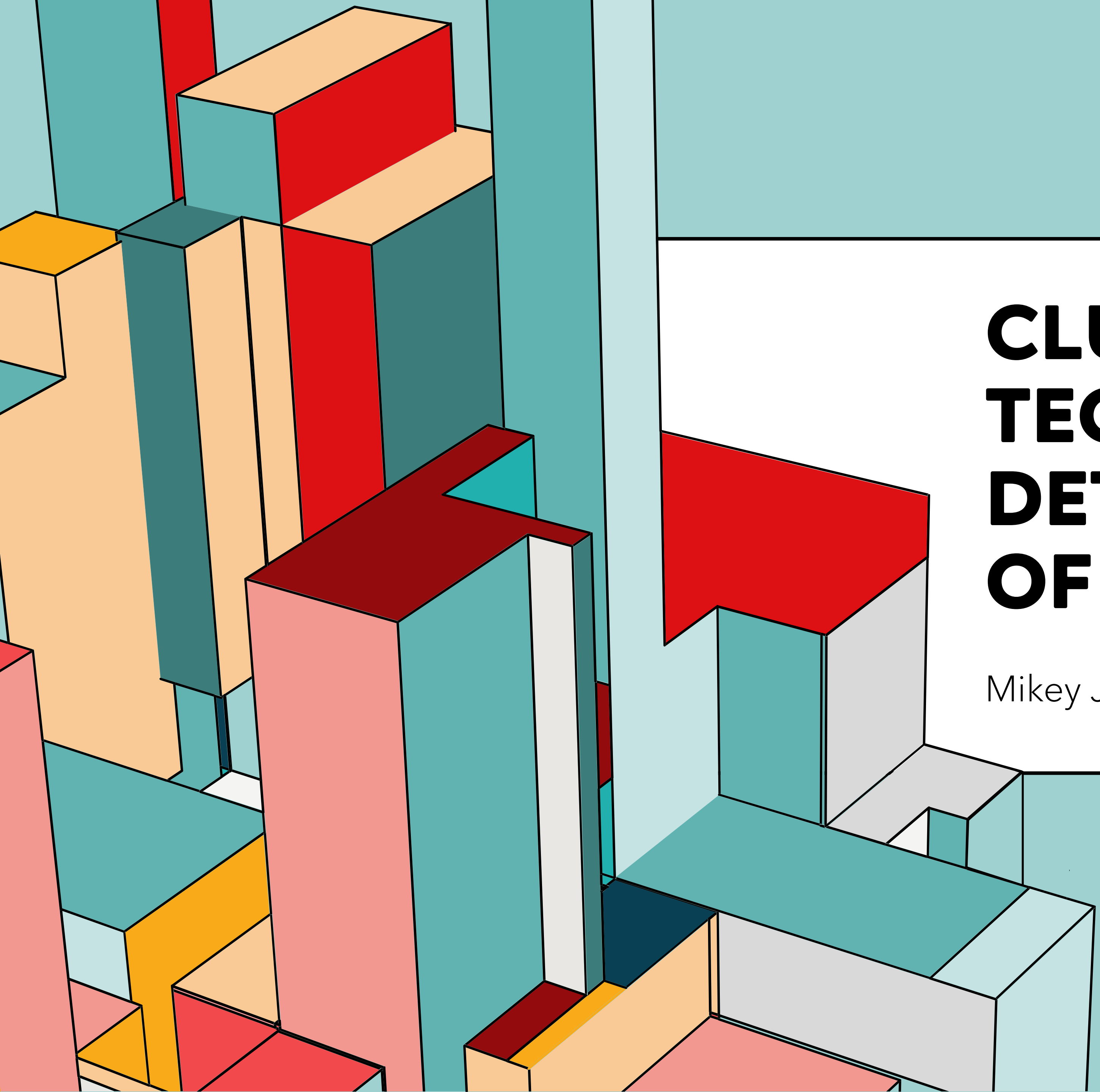


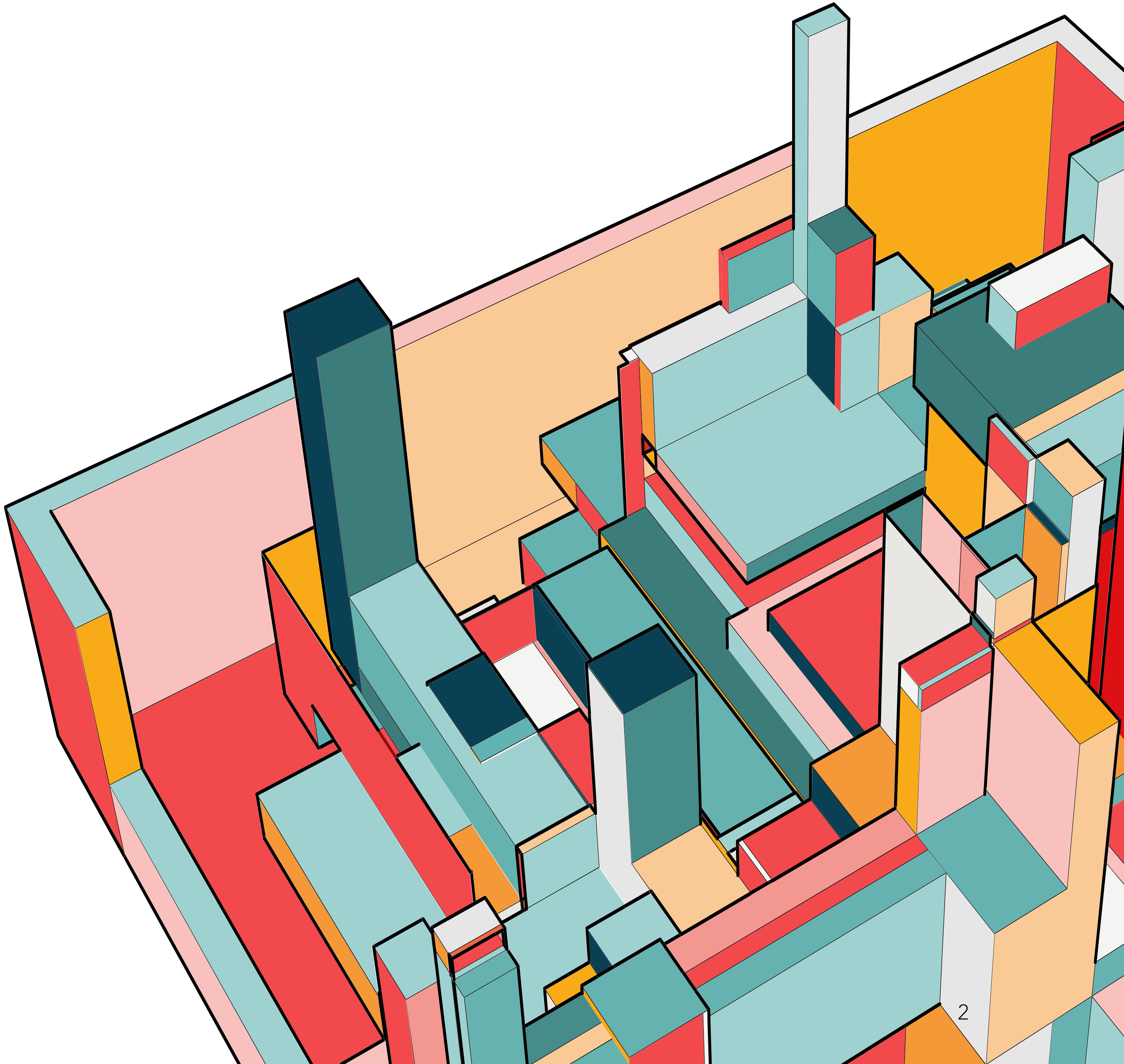
CLUSTERING TECHNIQUES TO DETERMINE ORIGINS OF WINE

Mikey Joyce

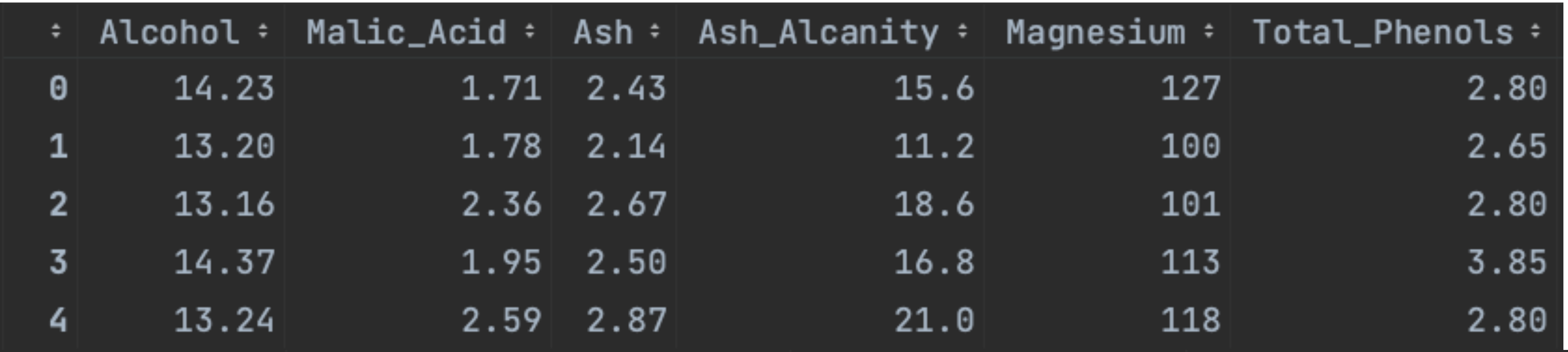
GOAL

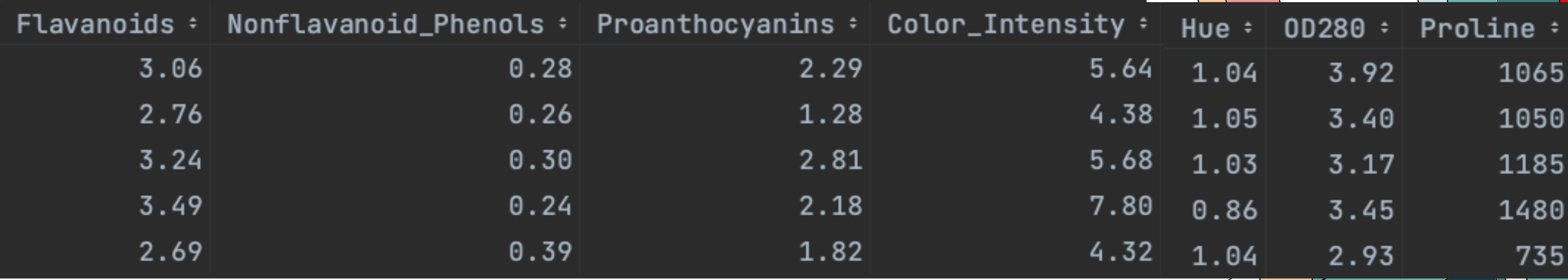
- Analyze wine dataset utilizing clustering techniques to determine which wines should be grouped together

- Create a tool that could be utilized by a sommelier or aspiring sommelier to improve their technique for identifying wines

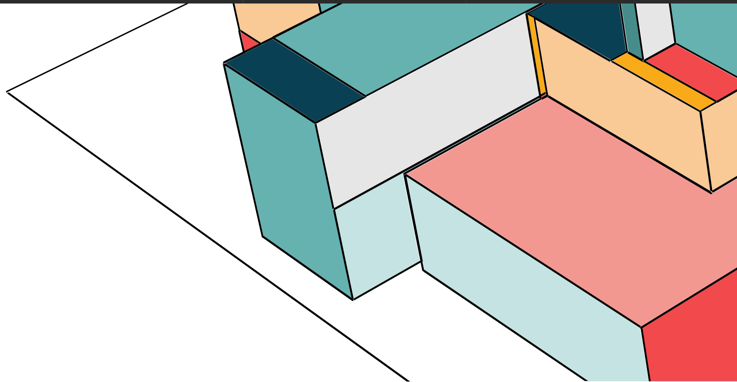


THE DATA





	Alcohol	Malic_Acid	Ash	Ash_Alcanity	Magnesium	Total_Phenols	Flavanoids	Nonflavanoid_Phenols	Proanthocyanins	Color_Intensity	Hue	OD280	Proline
0	14.23	1.71	2.43	15.6	127	2.80	3.06	0.28	2.29	5.64	1.04	3.92	1065
1	13.20	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28	4.38	1.05	3.40	1050
2	13.16	2.36	2.67	18.6	101	2.80	3.24	0.30	2.81	5.68	1.03	3.17	1185
3	14.37	1.95	2.50	16.8	113	3.85	3.49	0.24	2.18	7.80	0.86	3.45	1480
4	13.24	2.59	2.87	21.0	118	2.80	2.69	0.39	1.82	4.32	1.04	2.93	735

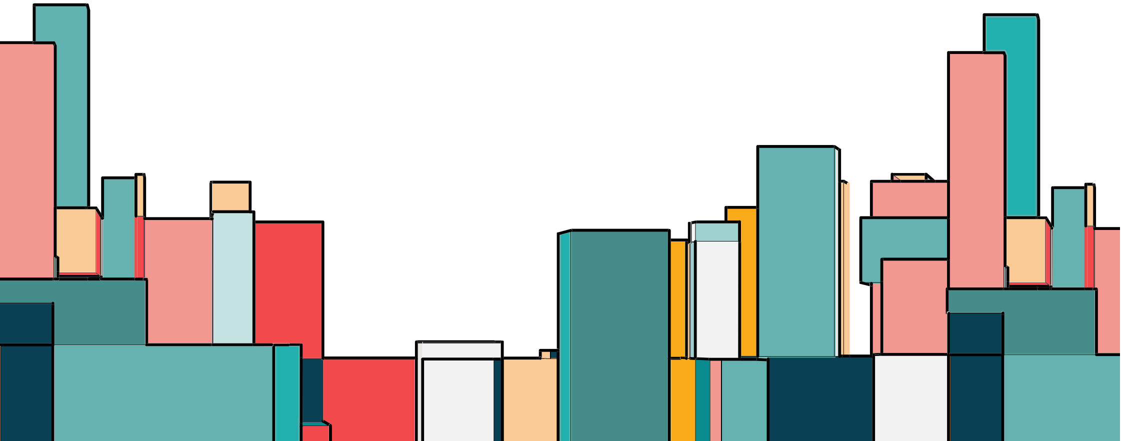


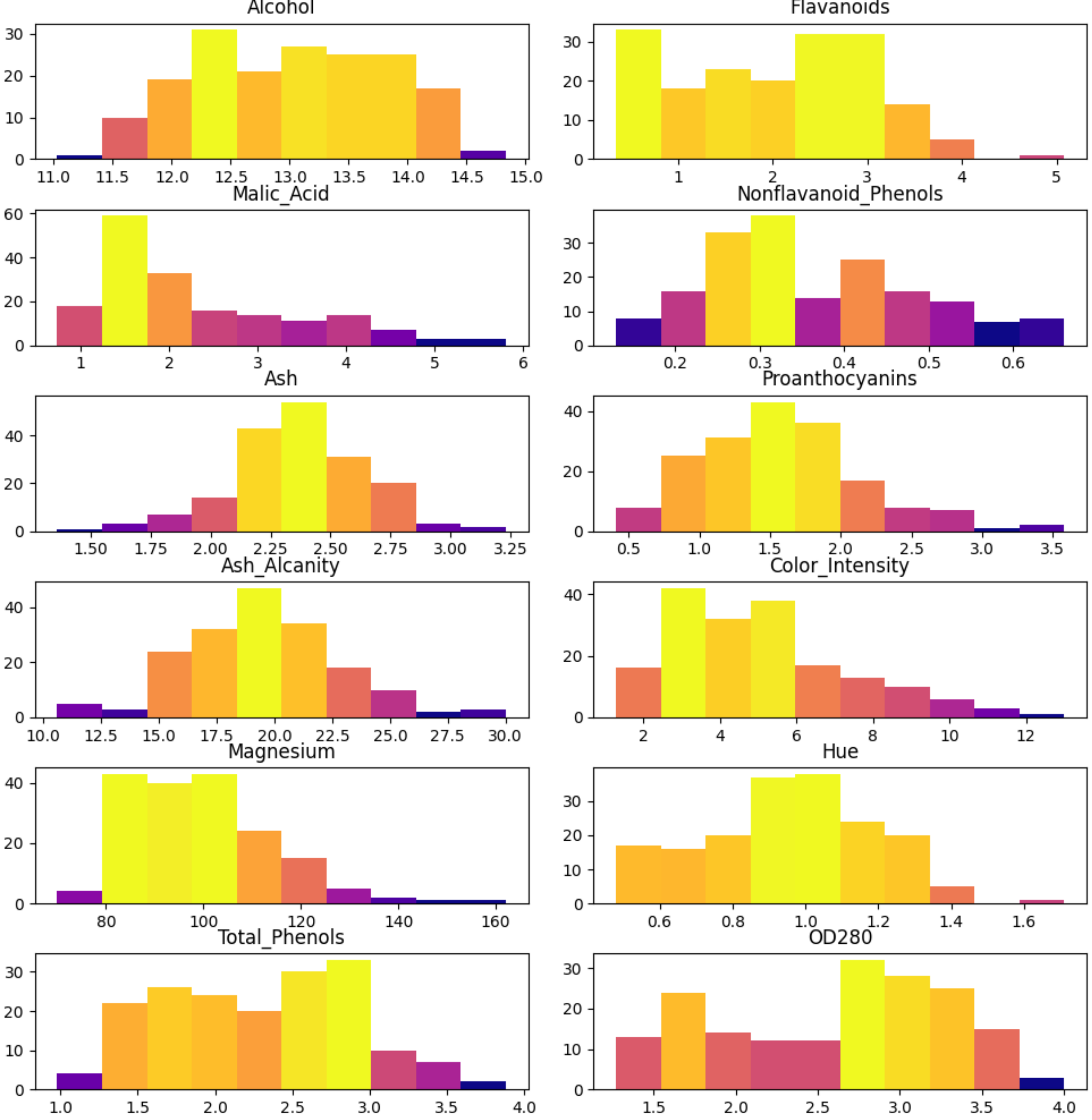
- 13 features contained within the dataset
- Will drop Proline because it is the only categorical feature and feels out of place

RESULTING FEATURE VECTOR

- Alcohol ←→ Falvanoids
- Malic_Acid ←→ Nonflavanoid_Phenols
- Ash ←→ Proanthocyanins
- Ash_Alcanity ←→ Color_Intensity
- Magnesium ←→ Hue
- Total_Phenols ←→ OD280

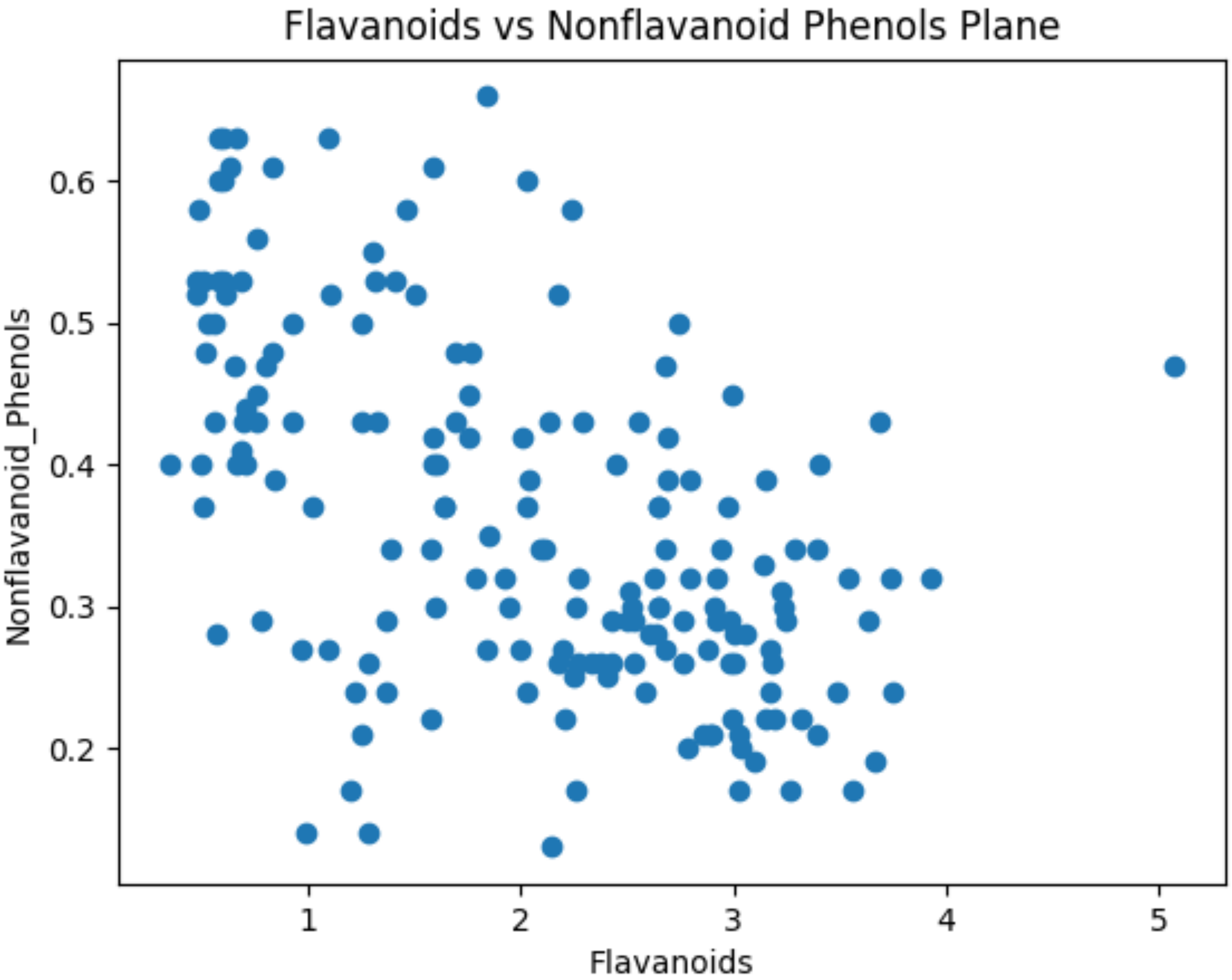
12 features within the vector





ARE CLUSTERS OBVIOUS?

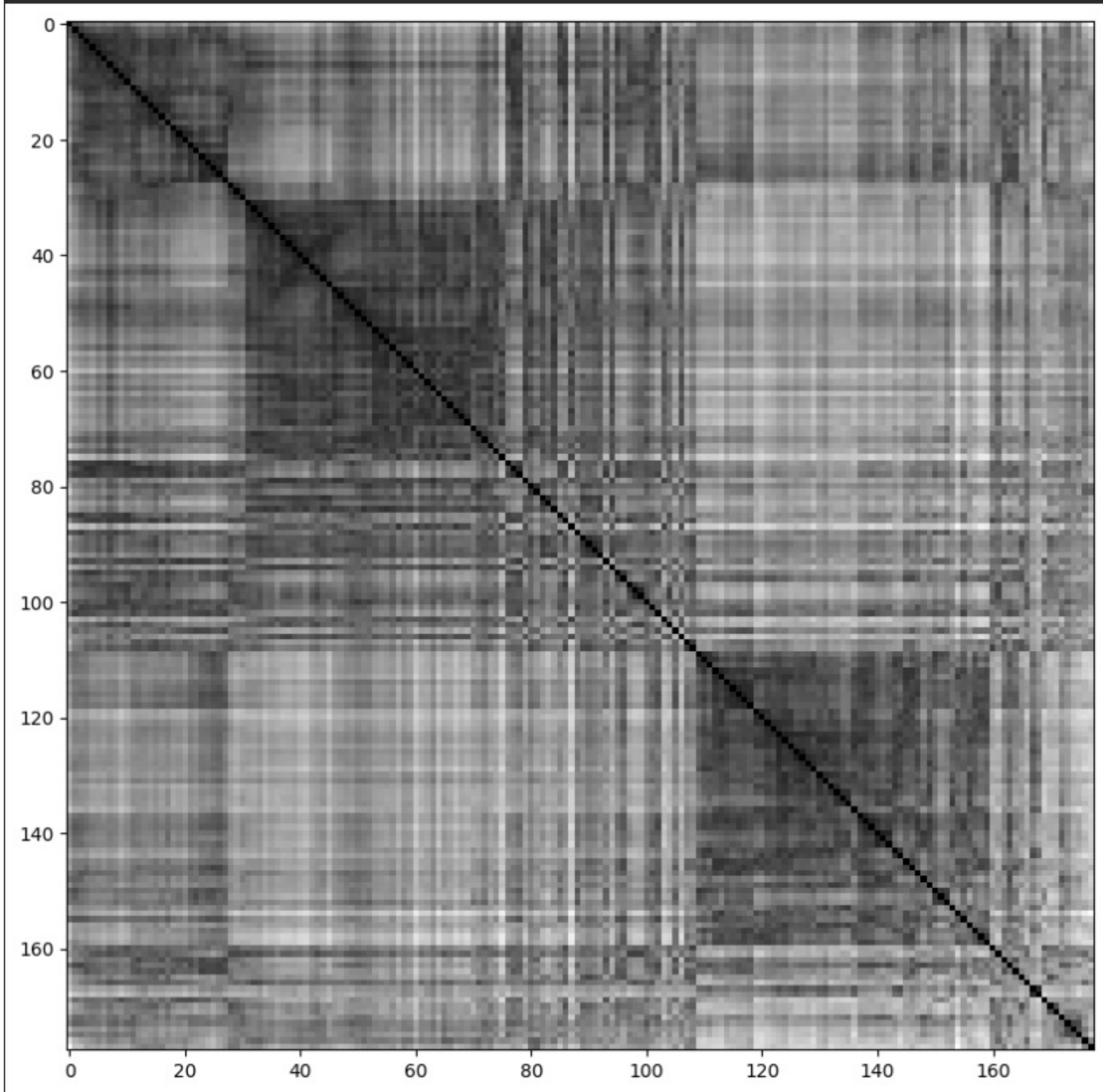
- Clusters are not obvious to the eye on any of the plots



DETERMINING NUMBER OF CLUSTERS

- Decided to run the vat and ivat algorithm to visualize if there is a cluster structure within the data

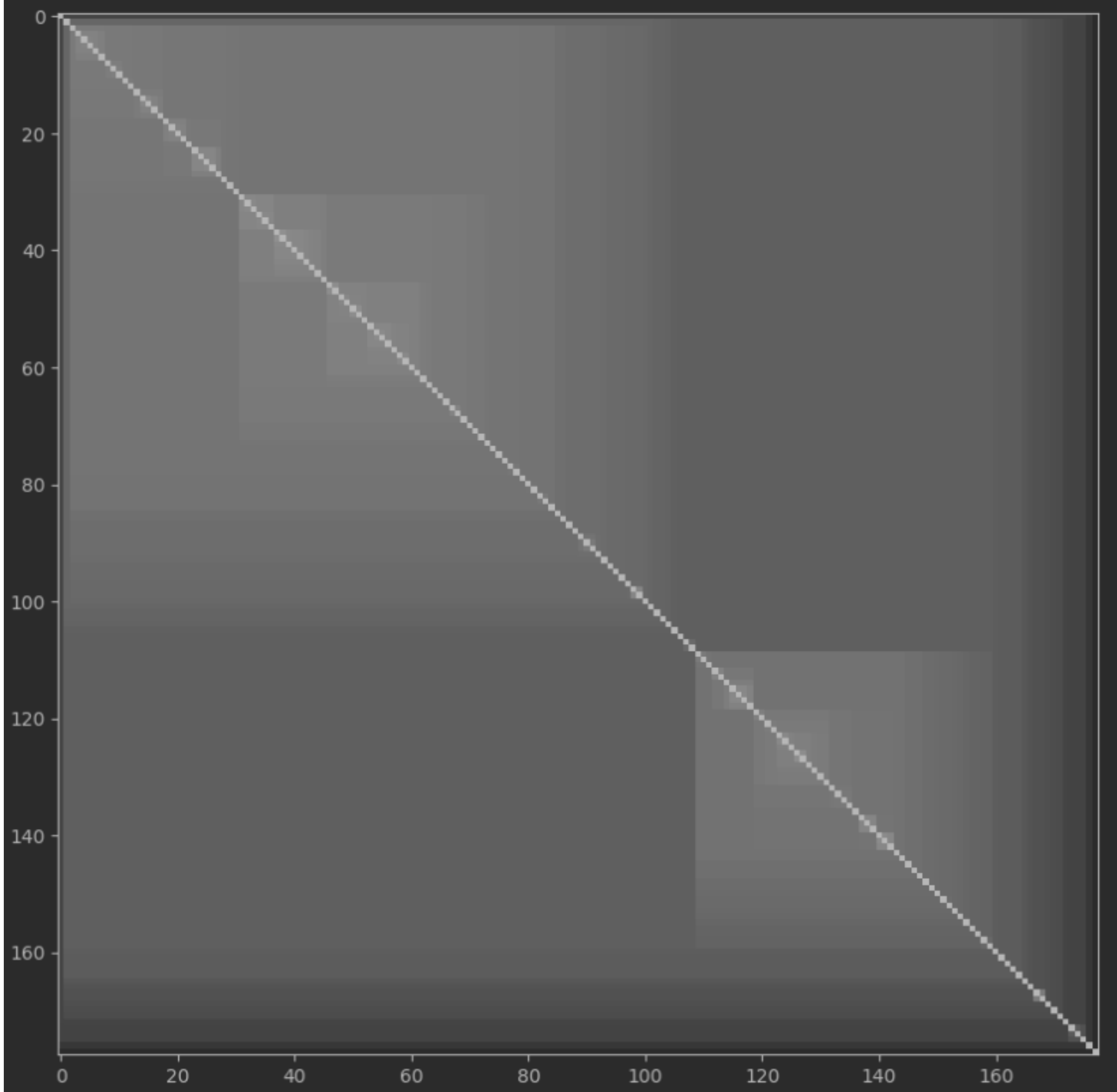
- There seems to be a structure with 2 main clusters and the larger cluster may have 2 sub clusters within

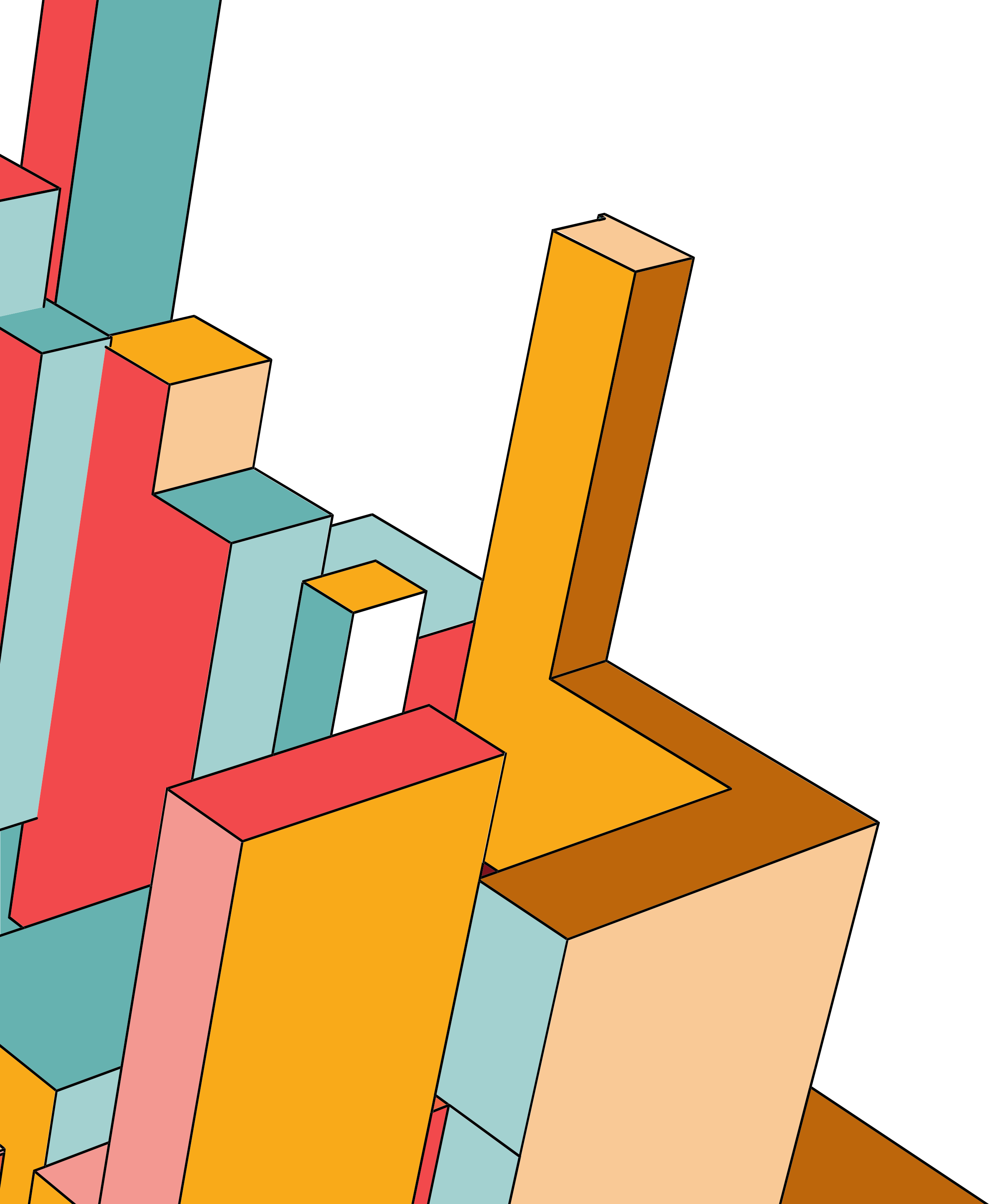


DETERMINING NUMBER OF CLUSTERS

- It is more clear in the ivat that there are two main clusters and the larger cluster has two sub clusters

- This observation will influence my approach. I will not try to cluster using 3 clusters as num clusters. Instead I will try and find two main clusters and then with the larger cluster I will try and find the sub clusters.





CLUSTERING PLAN

DIMENSIONALITY REDUCTION ROUND #1

Reduce the 12-feature set with t-SNE into a 2D map to aid in clustering

CLUSTERING ROUND #1

Utilize the map given from dimensionality reduction to identify two clusters with spectral clustering

DIMENSIONALITY REDUCTION ROUND #2

Remove the data points from the smaller cluster from the dataset and then reduce the 12-feature set into a 2D map with t-SNE for clustering

CLUSTERING ROUND #2

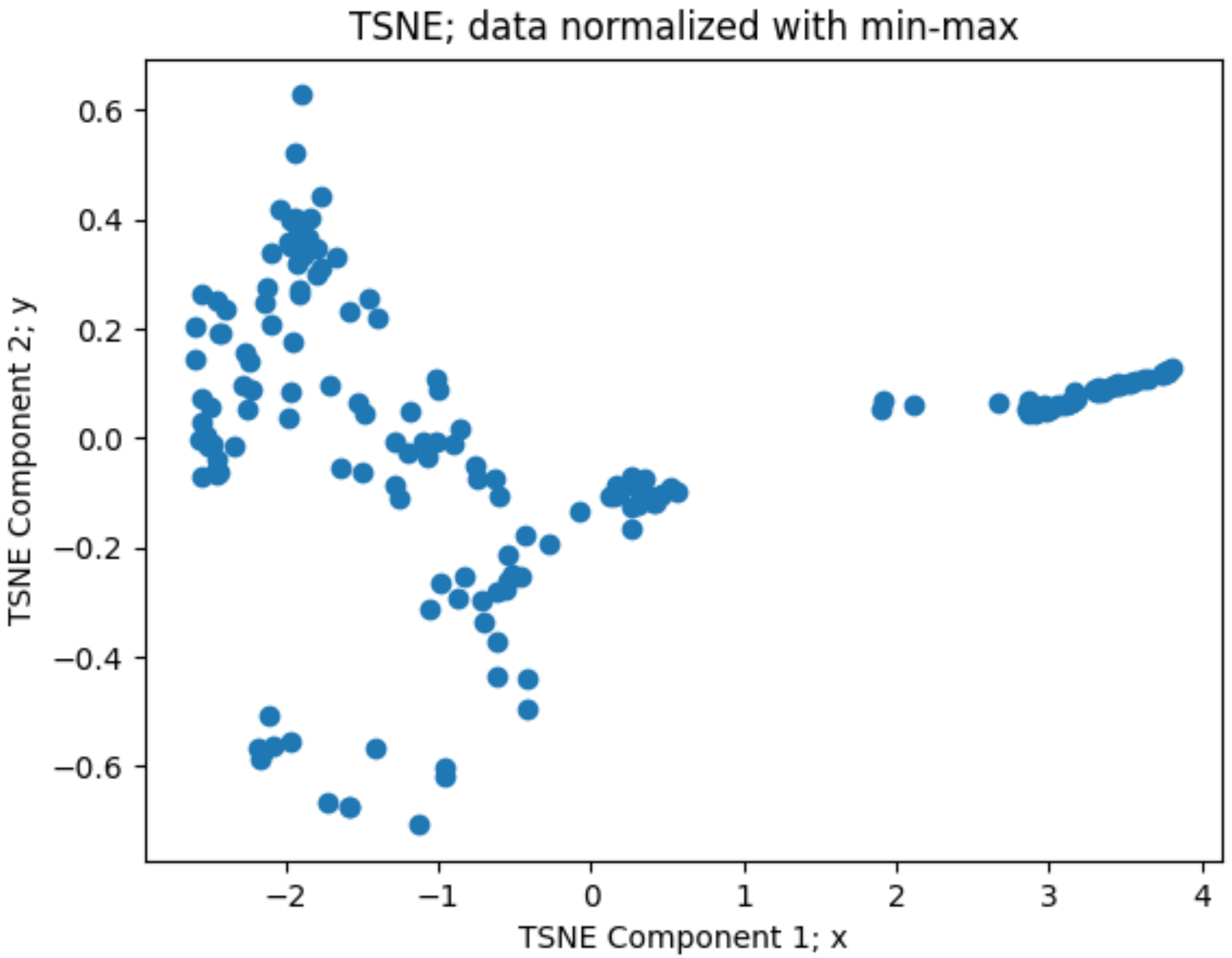
With the final map apply fuzzy c–means clustering over it to obtain two subclusters. Harden the memberships and calculate the final clusters

DIMENSIONALITY REDUCTION ROUND 1: T-SNE

- PCA didn't give ideal results

- Applied t-SNE on the data and there appears to be more separability between the data

- This t-SNE implementation allows for the saving of embeddings as well, so new data can be mapped to this exact embedding

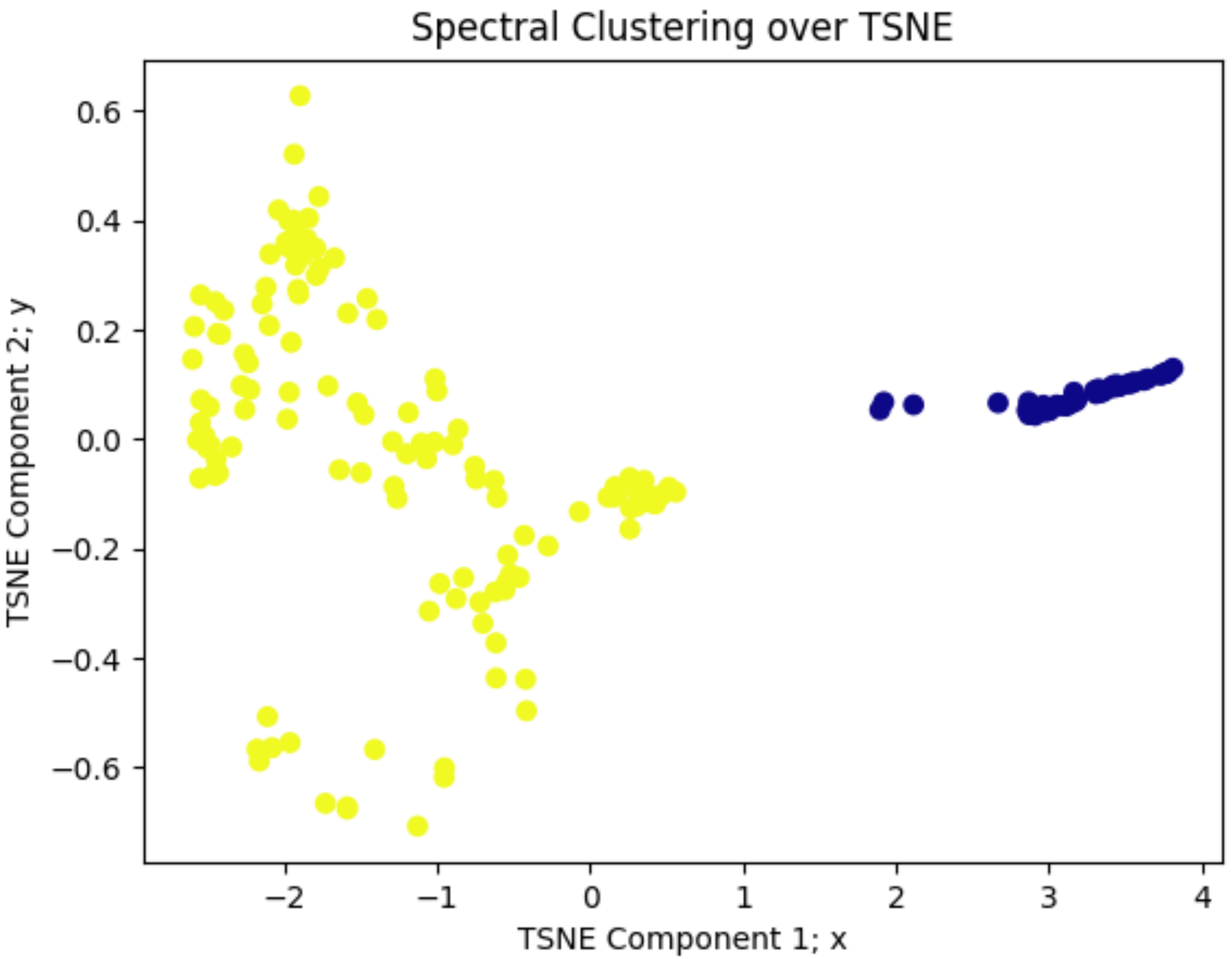


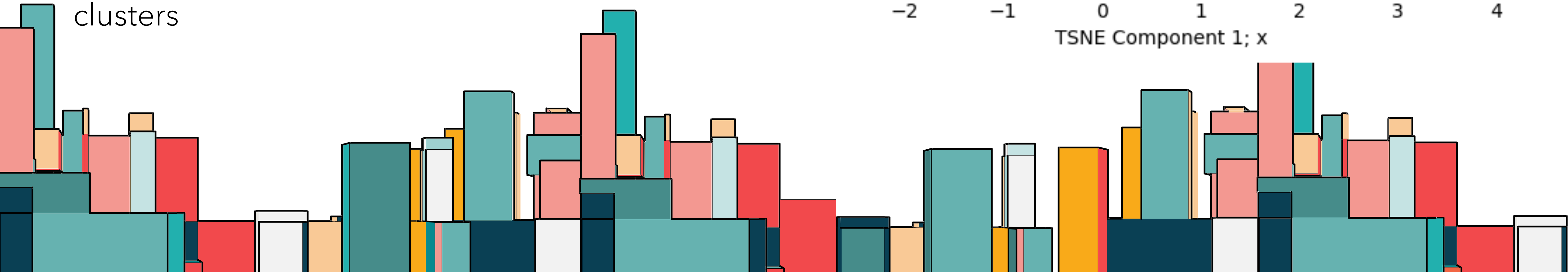
CLUSTERING ROUND 1: SPECTRAL

- Since we know one cluster is likely larger than the other cluster, I decided to opt for **spectral clustering** to deal with the imbalance of cluster sizes

- Spectral clustering also seemed like a good fit because the clusters each have structured shapes that are not the same

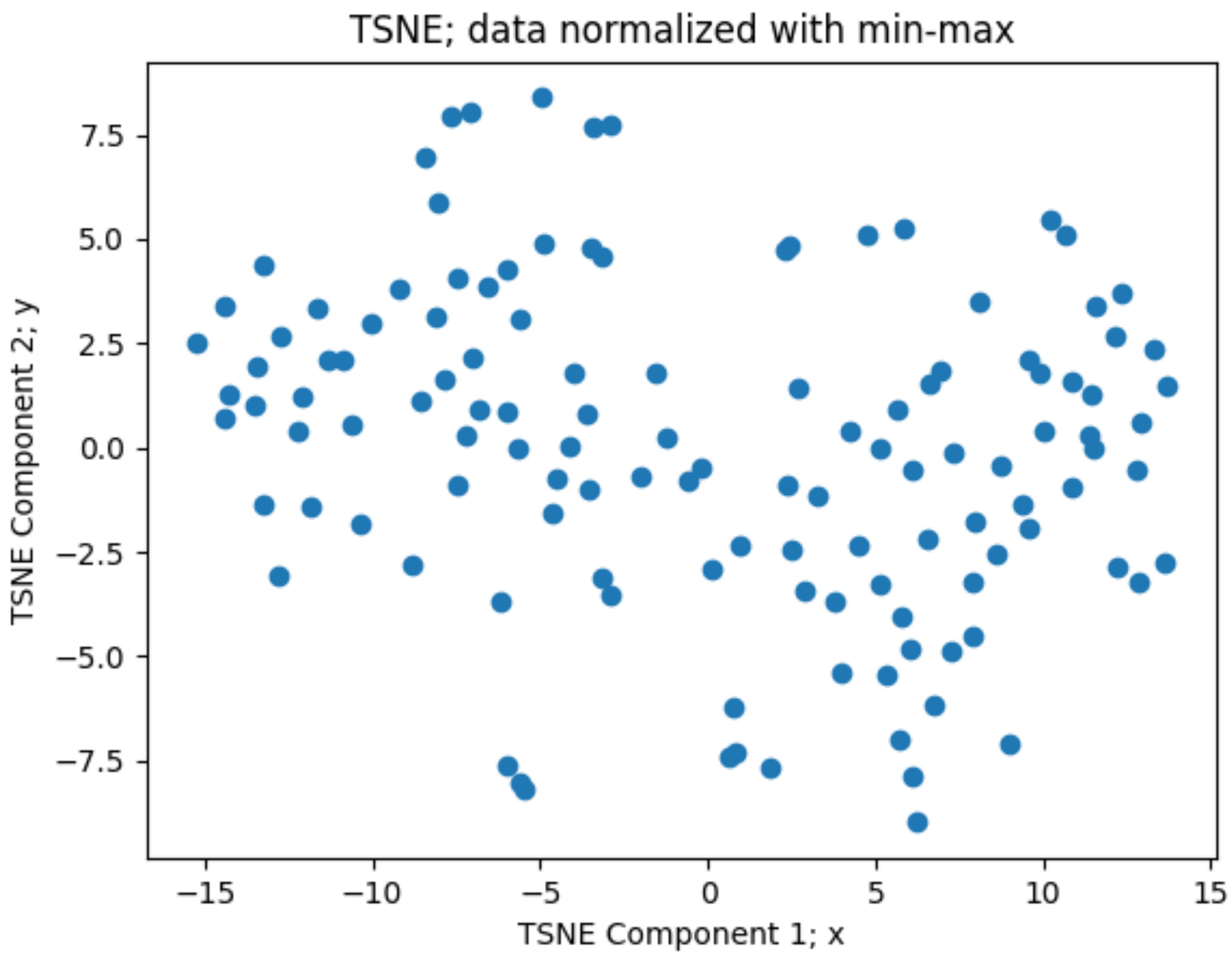
- Was able to obtain two clear and concise clusters

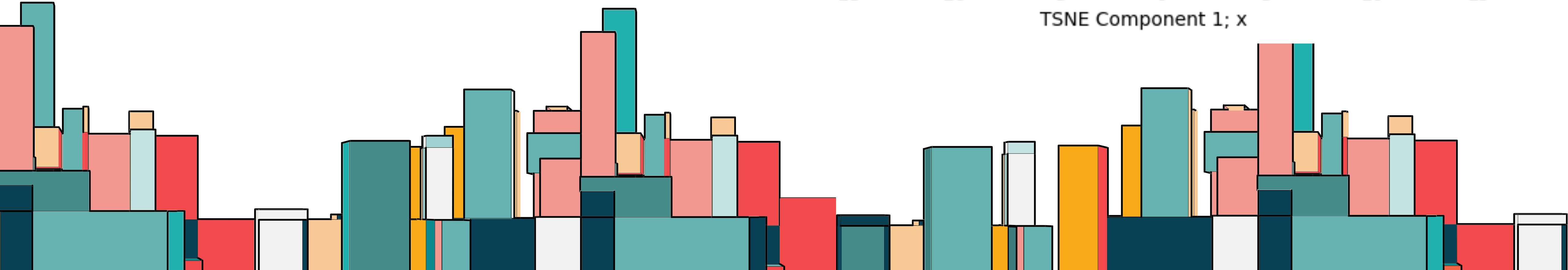




DIMENSIONALITY REDUCTION ROUND #2: T-SNE

- Now to start this round, I deleted the data points from the smaller cluster from the input vector
- Applied t-SNE on the larger cluster to get a smaller feature set

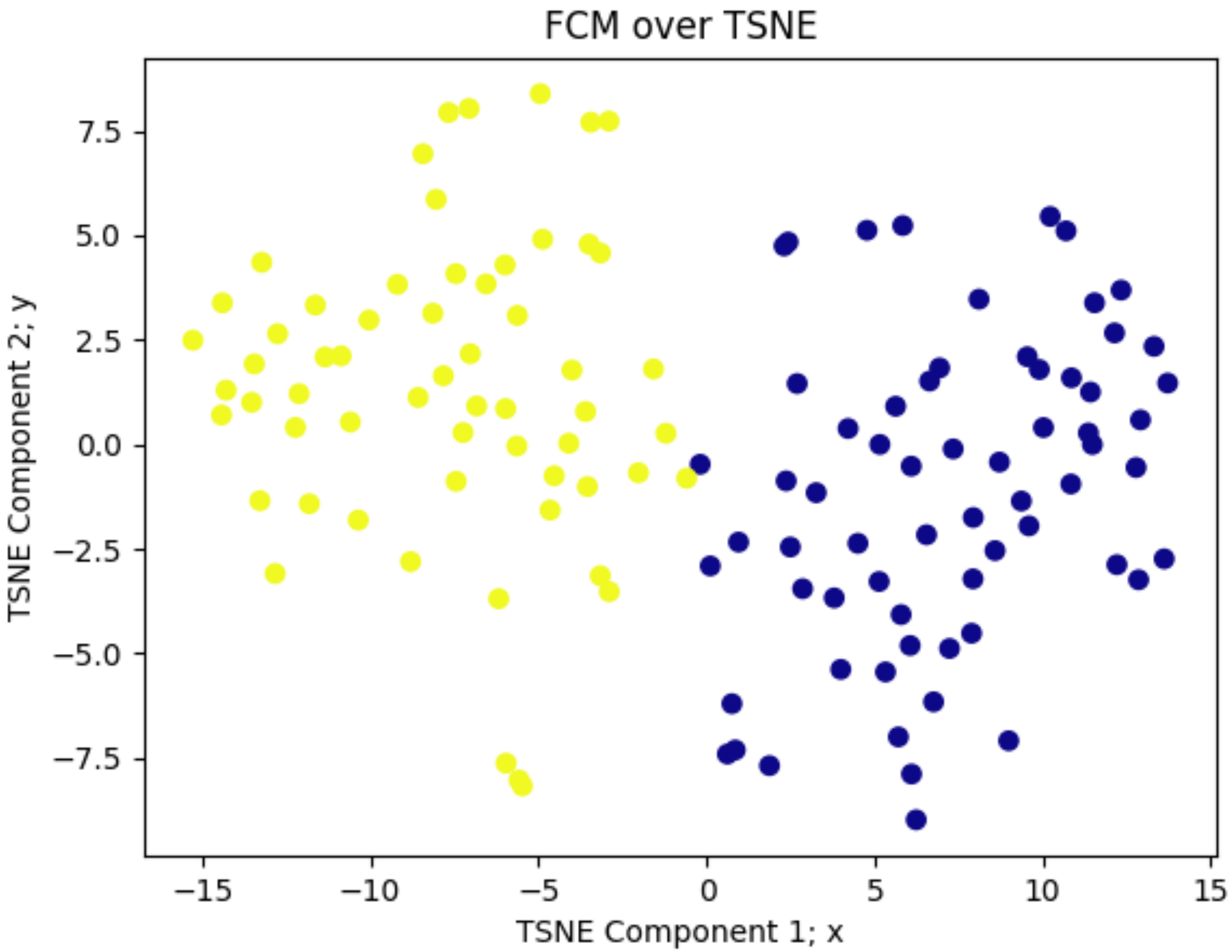




CLUSTERING ROUND 2: FUZZY C-MEANS

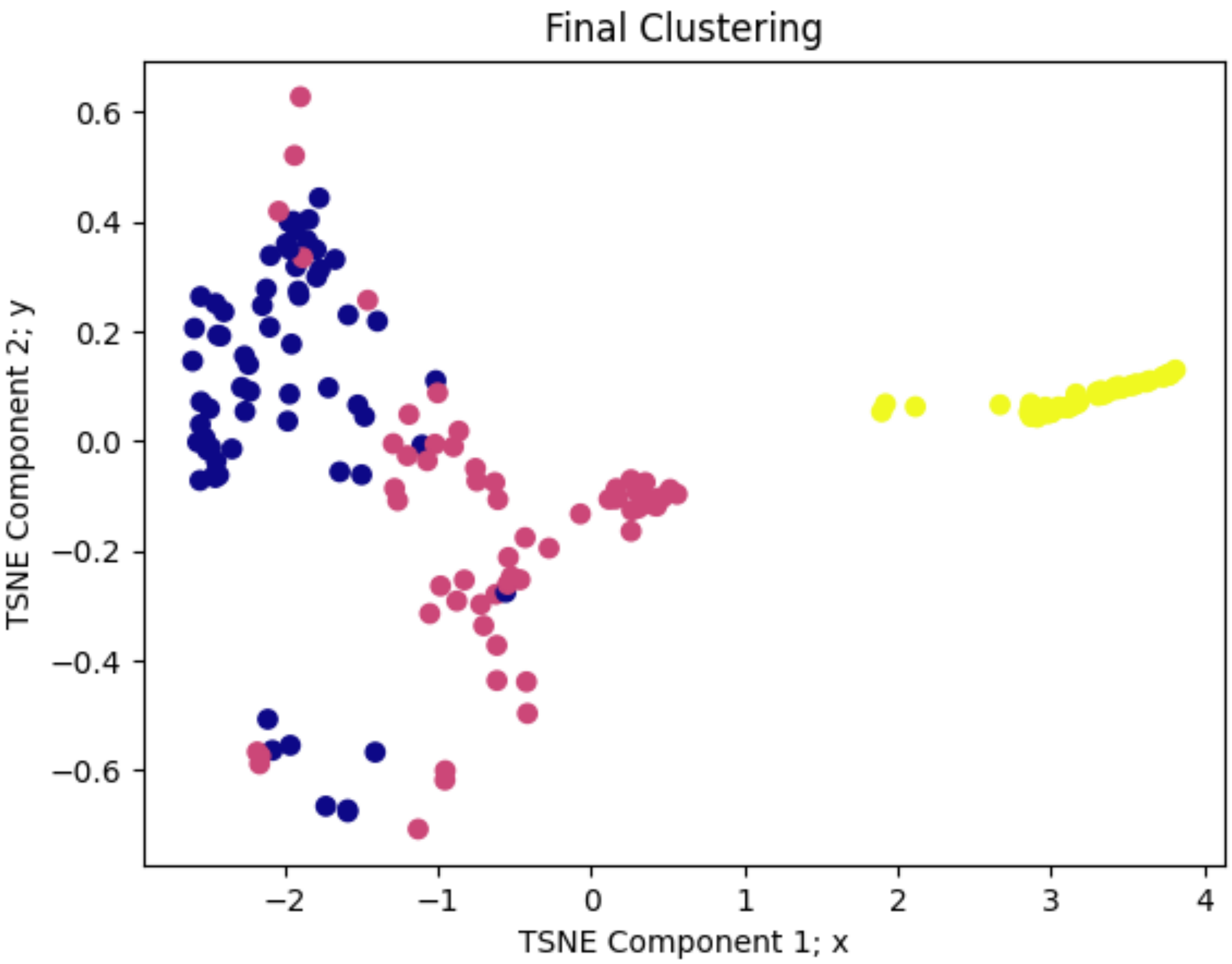
- The t-SNE plot obtained from the larger cluster did not seem to have as crisp of clusters as the first t-SNE plot.

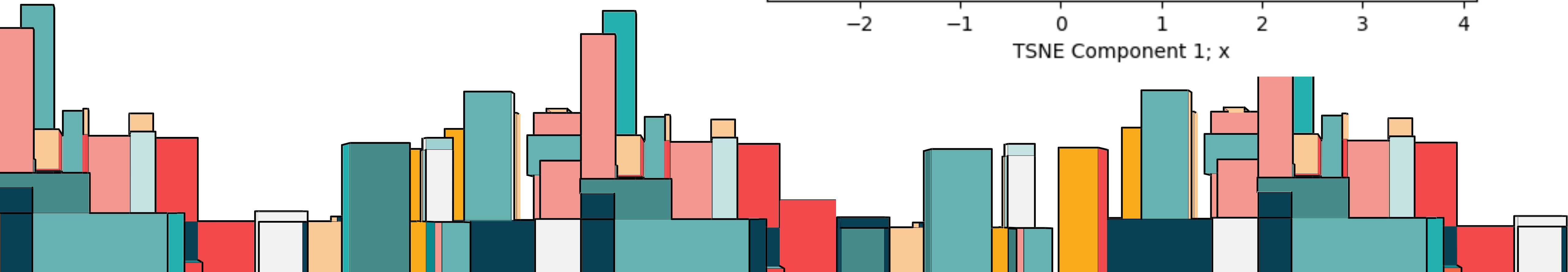
- Because of this I decided to employ the Fuzzy C-Means clustering algorithm to account for uncertainty within the clusters

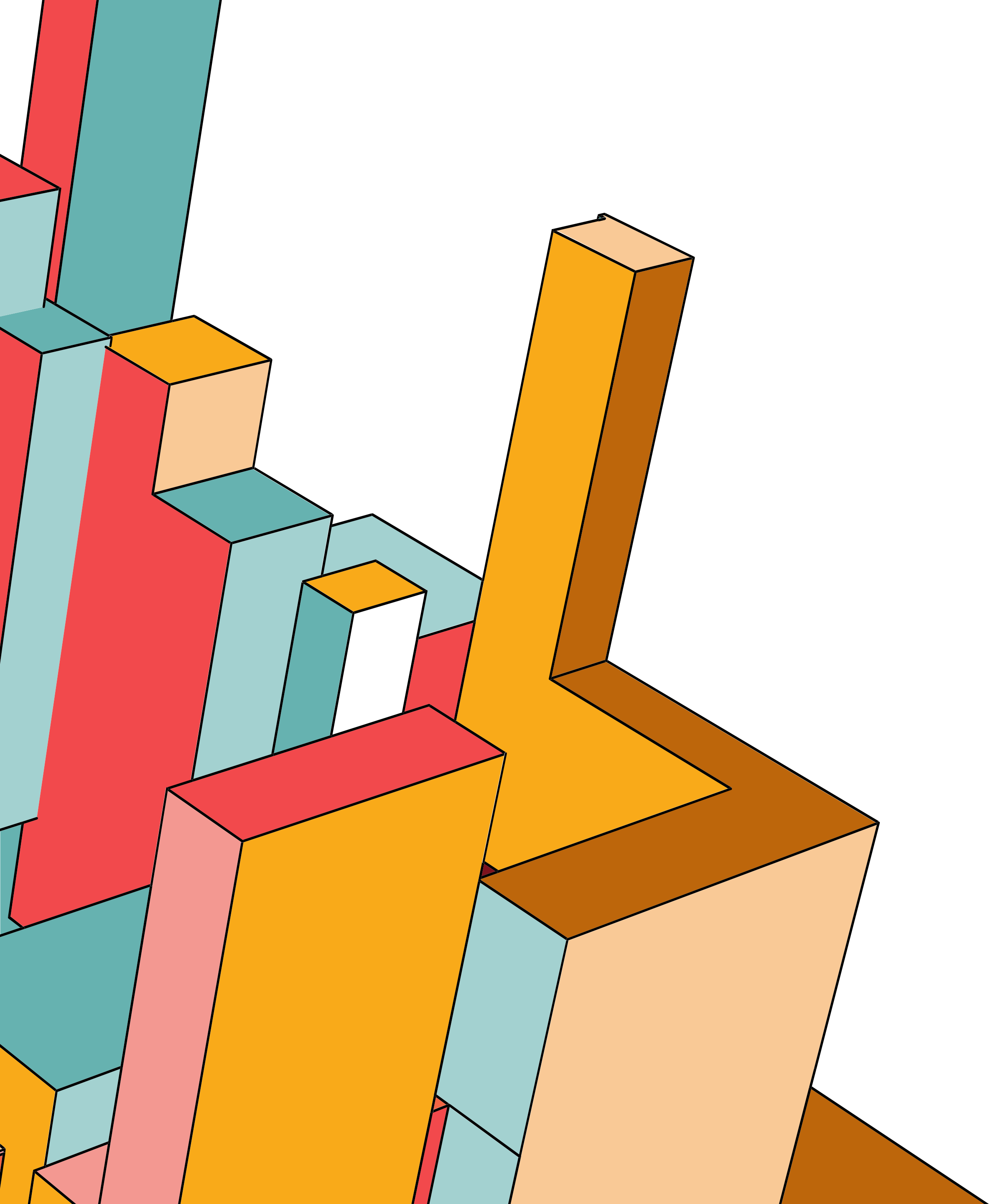


FINAL CLUSTERING RESULT

- The results are visualized on the first t-SNE embedding that was produced
- My dataset that I found online had a counterpart with the same data for supervised learning, so I was able to obtain the labels, my method performs at **93%** accuracy when applied against the labels as a benchmark







CITATIONS

- Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. 2001. On spectral clustering: analysis and an algorithm. In Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic (NIPS'01). MIT Press, Cambridge, MA, USA, 849–856.
- Bezdek, James C., et al. "FCM: The fuzzy C-means clustering algorithm." *Computers &amp; Geosciences*, vol. 10, no. 2–3, 1984, pp. 191–203, https://doi.org/10.1016/0098-3004(84)90020-7.
- Havens, Timothy & Bezdek, James. (2012). An Efficient Formulation of the Improved Visual Assessment of Cluster Tendency (iVAT) Algorithm. Knowledge and Data Engineering, IEEE Transactions on. 24. 1 - 1. 10.1109/TKDE.2011.33.
- Maaten, Laurens van der and Geoffrey E. Hinton. "Visualizing Data using t-SNE." Journal of Machine Learning Research 9 (2008): 2579-2605.
- Wang, Harry. "Wine Dataset for Clustering." *Kaggle*, 29 Apr. 2020, www.kaggle.com/datasets/harrywang/wine-dataset-for-clustering.