Custom python clustering project:

1. In the jupyter notebook, I have a keys list called keys\_less\_than

These are actually the cluster labels who have a probability less than 5/2638. In total I have 2638 data points labeled into 455 clusters. Out of the 455 clusters, 420 of them contain only a single point. Therefore, I have 420 clusters with a probability of 1/2638. My goal is to cluster these 420 data points to create groups of at least 5 points per group (meaning each cluster should have a minimum of 5/2638 probability). Maybe we will have 40 final clusters, maybe 100, maybe 150, but no less than 35 because those major clusters have a probability of already > 5/2638 so we do not need to merge them unless other points from keys\_less\_than merge into them. Because those single points might be closer to these big ones. However, priority is given to merge closest ones with smallest probabilities first.

To achieve this, I will outline the following rules.

1. Compute the distance between the islands. I provided a sample code to do that. Since my data points are in the form of a 7-dimensional integer grid, lets stick to the Manhattan (cityblock) distance.
2. Pick a cluster from keys\_less\_than and find the island closest to it. The preference and priority should be given to another cluster inside keys\_less\_than.
3. Now merge these two clusters and add their probabilities. Make note of this merging because later we want to label these points.
4. The next point to choose is another cluster that only has one point with a probability of 1/2638. If not available then 2/2638. The goal is to do the minimum merging until all clusters have at least 5 points in them.
5. Merge. Store merged labels.
6. Repeat.
7. Do not choose clusters with probability > 5/2638
8. We should end up with more than 35 final clusters
9. If a cluster is equal distance from more than one cluster, how do we choose to which cluster we merge? --- we need to think about this because many of the clusters are equal distance from each other. Either merge with the cluster that has the smallest probability. I am thinking maybe Pearson correlation in PCA space?

If I have 3 clusters. One cluster has 75 data points. one cluster has 2 data points. And another cluster has 21 data points. How can I compare the similarity between these clusters in PCA space? Choose to merge based on this score.