Data 100, Spring 2022 **Name:**

Discussion #14

Clustering

1. (a) Describe the difference between clustering and classification.

   (b) Given a set of points and their labels (or cluster assignments) from a K-Means clustering, how can we compute the centroids of each of the clusters?

   (c) The process of fitting a K-means model outputs a set of k centers. We can compute the quality of the output by computing the distortion on the dataset. A Data 100 student suggests that distortion is not well-defined when evaluating the output of any agglomerative clustering algorithm because the algorithm doesn't return centers, but simply labels each point individually. Is the student correct?

   (d) Describe qualitatively what it means for a data point to have a negative silhouette score.

   (e) Suppose that no two points have the same distance from each other. Are the cluster labels computed by K-means always the same for a given dataset? What about for max agglomerative clustering?