Michael Ventoso
<MichaelVentoso@Gmail.com>

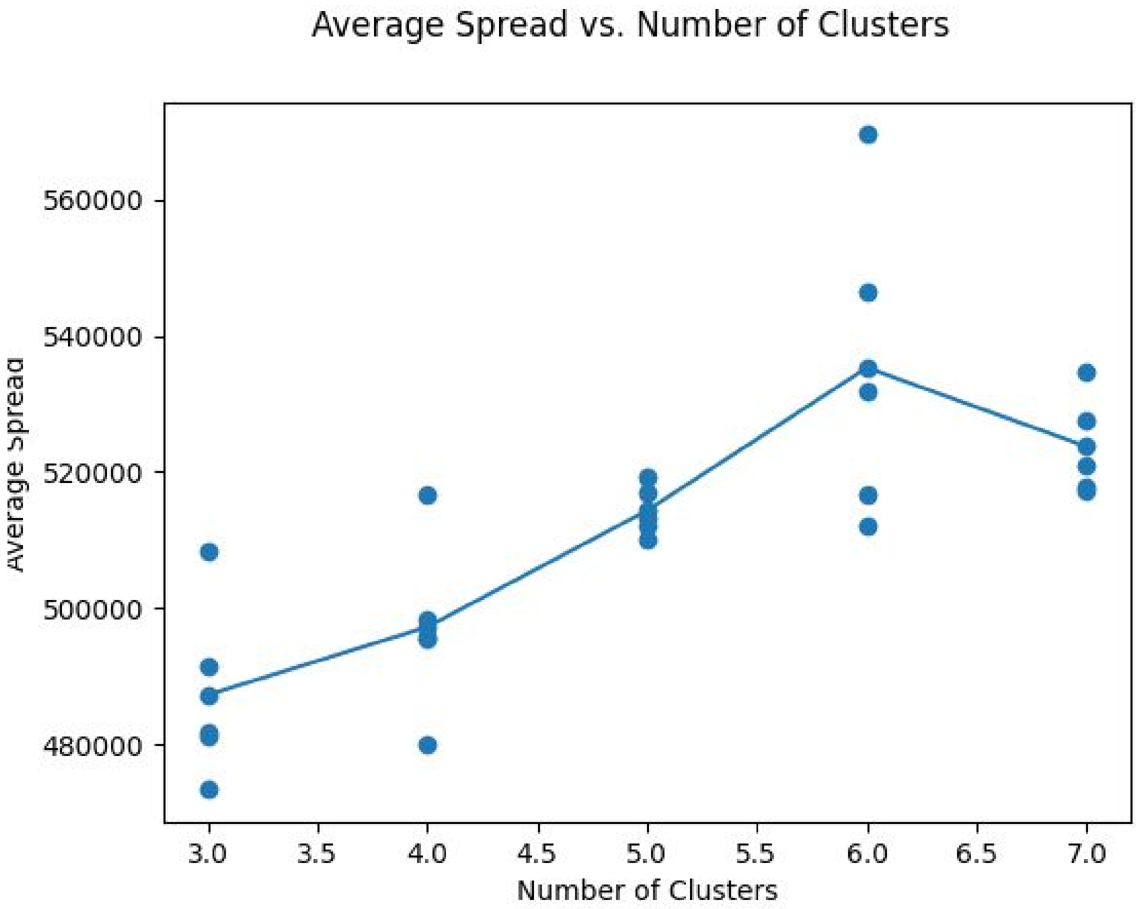
ProtAtOnce Interview Phase 3
Part B: Cell Dataset

Methodology:

When I read the assignment and looked briefly at the dataset, my first thought was that there are definitely some machine-learning algorithms that could be applied to separate the cells into groups. I unfortunately have no experience in machine learning, but had another idea to group the cells. Since the cells' features are all numeric, I conceptualized them as merely points in a 20-dimensional space. My mind immediately went to the Euclidean distance formula, since I know it works in any number of dimensions. Then I had my idea that I would stick to; if I am using distance as a way to compare these points, then I can attempt to cluster them.

My choice for a clustering algorithm is K-Means clustering. I chose it because it is a relatively simple algorithm that scales well to large amounts of data. I preemptively decided I would force a limit on how many iterations it goes through, as doing Euclidean distance in 20 dimensions on thousands of points will already be computationally taxing on my laptop. I remember reading a paper that suggested that any number of iterations past 30-50 was not very significant in most cases, so I chose 50 to be on the safe side. Next, I had to choose my method for starting the clusters. I know there are ways I could do some computing beforehand and avoid the risks of choosing random points as centers, but I also know I am limited on time to do this assignment and that the basic algorithm usually involves randomized starting points anyway. Finally, I just had to choose basic parameters and consider the computational costs. I ended up deciding to test K values of 3, 4, 5, 6, and 7 clusters, and for each one perform 5 trials.

My method of determining which number of clusters was the best out of the set, was to calculate the average spread of each trial for all the clusters. This computation is quite similar to determining which cluster a cell goes in, as they both rely on the Euclidean distance. After an afternoon of running the trials, I was able to analyze the data. It turns out there is not that much difference in the average amount of spread between 3 and 7 groups. What drew my attention though, was how consistent the spreads were in the trials with 5 clusters.



Being so precise meant that the clusters were probably quite similar from trial to trial, so I decided to pick 5 as my best number of groups to separate the cells into. If I had more time available, I would have compared how similar the 5 trials were in terms of grouping the cells. Since there is not much time however, I decided to simply choose my first trial as the best choice grouping. The file for this grouping is called kMeans_k5_n1 and can be found in the folder KMeans_Data.

With the cells in groups, I thought looking at spread could also determine the most distinctive characteristics of each group. Within a cluster, a lower average spread of a given feature shows that the group is generally more consistent in that feature. Therefore, one solution is to find the five features for each cluster that have the lowest spread. I have put the features into a table below for the given trial. The features are listed in descending order starting from the most defining.

	Feature #1	Feature #2	Feature #3	Feature #4	Feature #5
Cluster 1	4	9	8	5	20
Cluster 2	7	4	9	5	12
Cluster 3	7	8	5	19	18
Cluster 4	9	19	4	13	5
Cluster 5	8	14	3	5	9