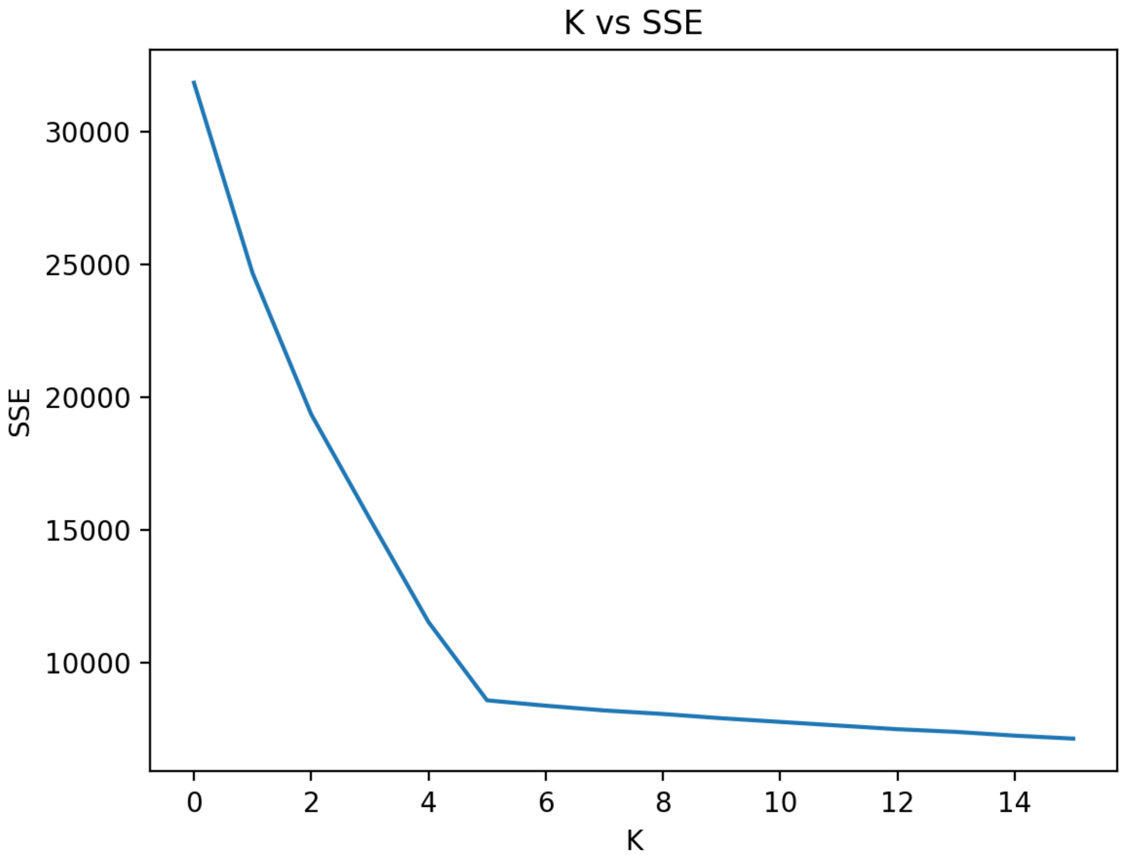
Pratishta Prakash Rao

Big Data Analytics - HW_07

1. Plot the SSE vs K for the L2 norm.

Solution:

The plot for SSE vs K is as follows:



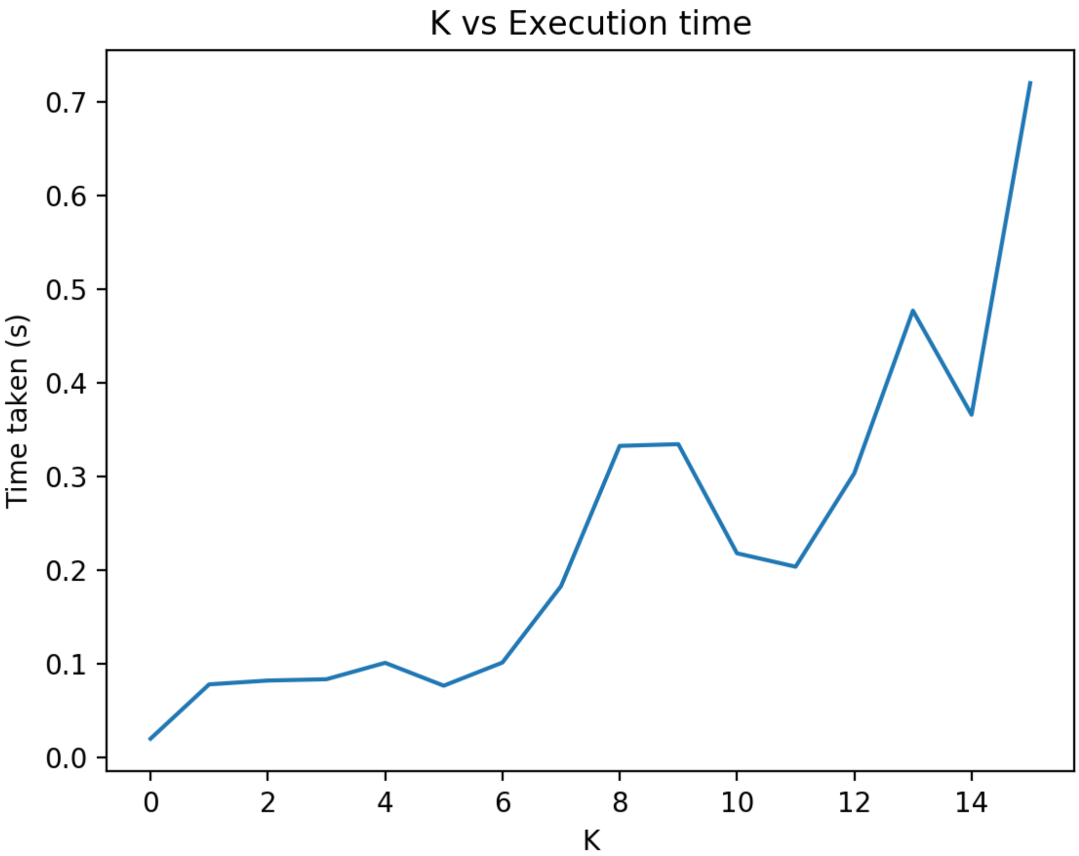
2. Based on what you observed, what value of K would you use as knee point? Why did you select this point?

Solution:

The value of the k obtained is 5.

3. Plot the time required for completion vs K, for all values of K. What can you say about this? Can you model this time mathematically?

<u>Solution:</u>



4. Cluster Statistics:

<u>Solution:</u>

Because of the large data output for cluster statistics for k = 1 to 15. I have listed the cluster statistics of only the best K(= knee point).

Pratishta Prakash Rao

Table 1 : Cluster statistics for k = 5

Cluster Number	A1	A2	A3	A4	No of points in the cluster	Cluster SSE
1	83.0	39.9	66.8	59.6	167	1806.66
2	83.0	75.4	33.3	48.9	153	1575.07
3	39.2	66.3	70.0	53.1	257	5555.66
4	52.5	20.5	41.0	53.5	119	1156.65
5	12.9	50.3	39.9	51.8	140	1429.37

Table 2 : Cluster statistics for k = 6

Cluster Number	A1	A2	A3	A4	No of points in the cluster	Cluster SSE
1	52.5	20.5	41.0	53.5	119	1156.65
2	83.0	39.9	66.8	59.6	167	1806.66
3	12.8	50.1	39.8	51.6	139	1396.31
4	34.4	73.5	62.7	71.2	118	1180.15
5	43.2	60.2	76.0	38.0	140	1450.43
6	83.0	75.4	33.3	48.9	153	1575.07

5. What stopping criteria did you use for the inner loop?

Solution:

The stopping criteria for the inner loop is when the old centroid points is equal to the newly computed centroids.

6. What was the hardest part of getting all of it working?
Did anything go wrong?

Solution:

I thinking working with numpy was a little challenging I hadn’t worked with it before.

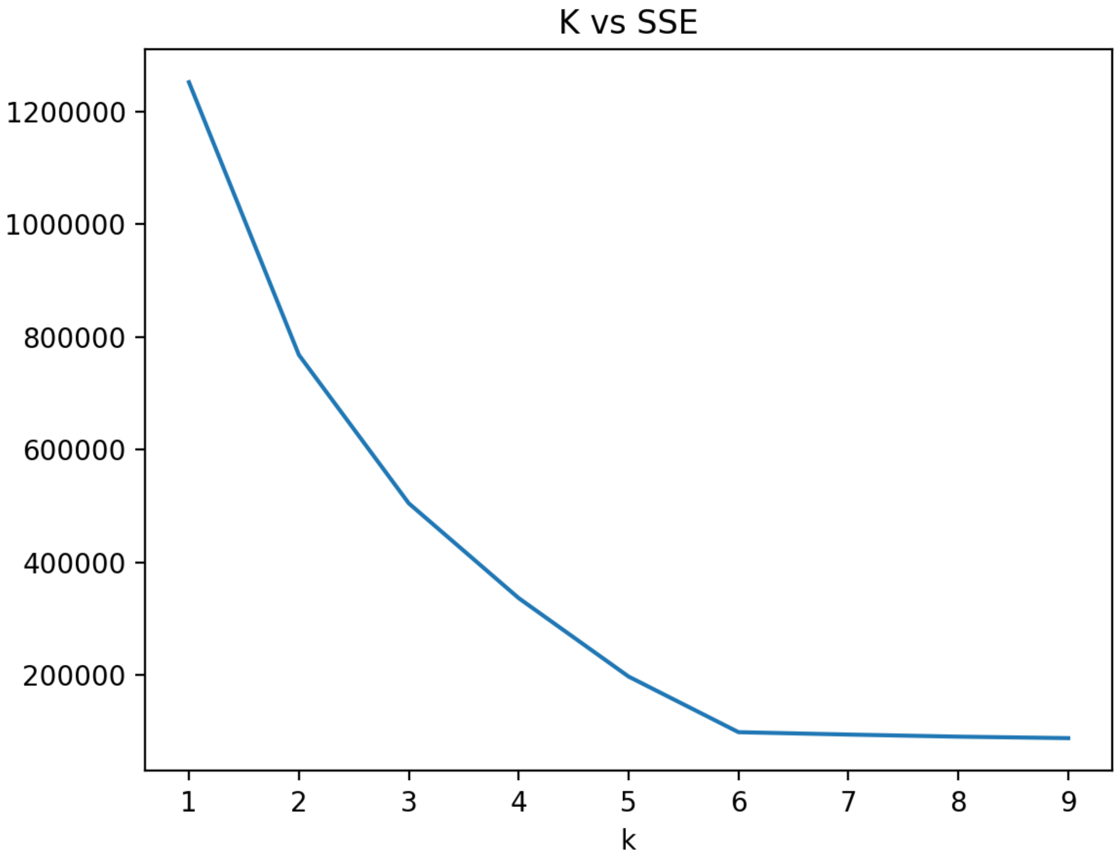
Also, calculating the cluster statistics for each iteration of k was slightly challenging. I had issues with indexing while keeping track of multiple data point, centroids , SSEs and cluster labels.

5. What did I learn about all this?
Solution:

Lessons learnt:

- Working with numpy
- K-means clustering
- Comparing doubles using numpy

Yes I will Kmeans for the quiz and my job interview.
I did compare it with the built in package and the following is the graph obtained:


K vs SSE
k