BIG DATA ANALYTICS (CSCI -720)

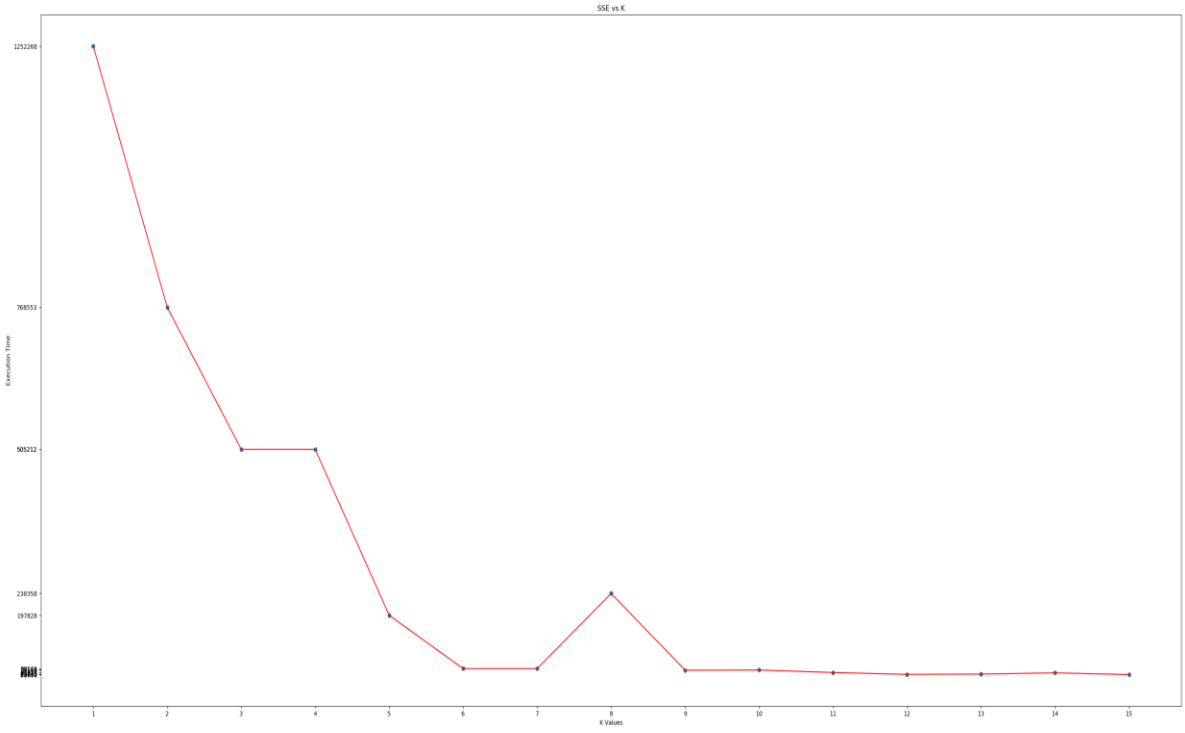
Homework-07 **Name: Rajkumar Lenin Pillai**

For k=15 and No of iterations=10 , the following observations were made.

Question-a.) Plot the SSE versus K for the L2 norm .

Solution:

Plot the SSE versus K for the L2 norm:



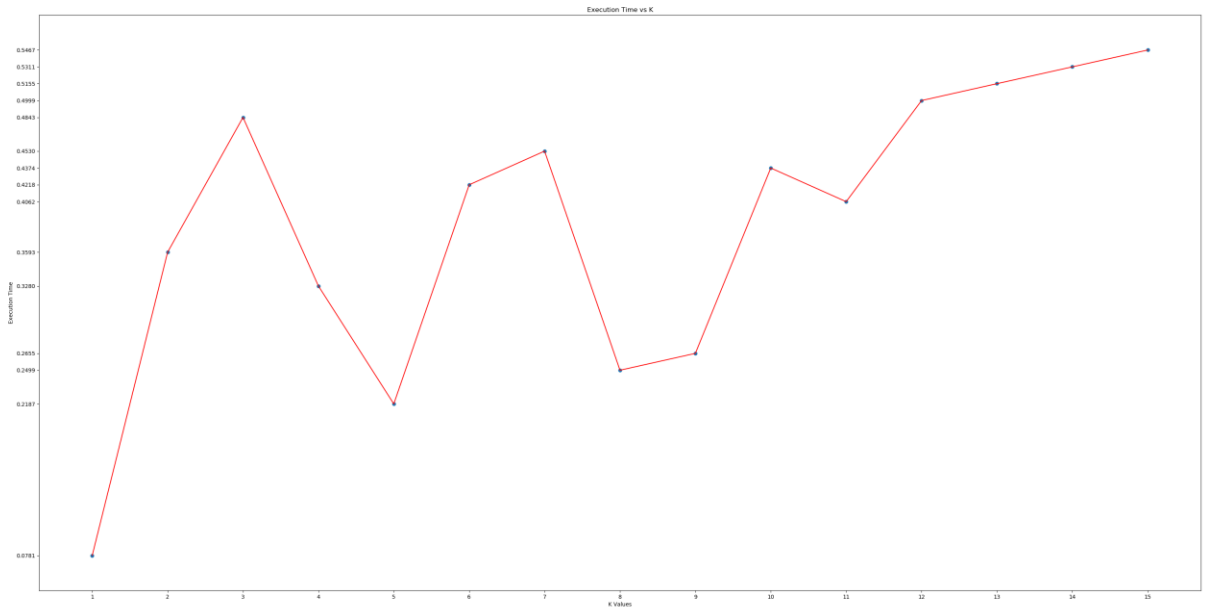
Question-b.) Based on what you observed, what value of K would you use as a knee point? Why did you select this point?

Solution:

The value of k=6 can be used as a knee point as it closely resembles to a knee point and it is the lowest value of k among all other values for which the SSE is very low which can be observed from the above figure.

Question-c.) Plot the time required for completion versus K, for all values of K. What can you say about this? Can you model the time mathematically?

Solution:



The execution time varies significantly and there is no general trend that can be observed from this. For some higher value of k execution time decreases compared to previous value of k.

This may be due to the initial centroid which are randomly considered and since no data points belonging to that cluster the centroid was removed and no of clusters was reduced in the program.So,it cannot be modeled mathematically.

Question-d) Cluster statistics: Sort the cluster centers from smallest to largest. Print the centers two one significant digit in this order. This makes life easy for the grader.

Run: kmeans kmeans

Computing kmeans for k= 15

ClusterID	A1	A2	A3	A4	Num.Points in this cluster	SSE For this cluster
1	[54.0,	54.0,	54.0,	54.0]	836	1252267.8098600477
2	[75.0,	47.0,	48.0,	54.0]	439	768552.8121916814
2	[30.0,	61.0,	59.0,	53.0]	397	768552.8121916814
3	[30.0,	61.0,	59.0,	53.0]	397	505211.7236255379
3	[75.0,	47.0,	48.0,	54.0]	286	505211.7236255379
4	[35.0,	52.0,	55.0,	53.0]	119	366019.8854738484
4	[83.0,	56.0,	51.0,	54.0]	397	366019.8854738484
5	[13.0,	51.0,	41.0,	53.0]	119	197827.53222662545
5	[70.0,	32.0,	56.0,	57.0]	257	197827.53222662545
5	[56.0,	70.0,	56.0,	51.0]	153	197827.53222662545
6	[30.0,	61.0,	59.0,	53.0]	257	197827.53222662574
6	[83.0,	57.0,	51.0,	54.0]	167	197827.53222662574
6	[53.0,	21.0,	41.0,	54.0]	140	197827.53222662574
7	[53.0,	21.0,	41.0,	54.0]	257	197827.5322266257
7	[83.0,	75.0,	33.0,	49.0]	153	197827.5322266257
7	[83.0,	40.0,	67.0,	60.0]	140	197827.5322266257
7	[30.0,	61.0,	59.0,	53.0]	119	197827.5322266257
8	[83.0,	40.0,	67.0,	60.0]	167	91918.53911730366
8	[35.0,	73.0,	63.0,	71.0]	139	91918.53911730366
8	[28.0,	55.0,	58.0,	45.0]	65	91918.53911730366
8	[83.0,	75.0,	33.0,	49.0]	78	91918.53911730366
8	[53.0,	21.0,	41.0,	54.0]	118	91918.53911730366
9	[83.0,	57.0,	51.0,	54.0]	140	99166.10294329426
9	[13.0,	50.0,	40.0,	52.0]	167	99166.10294329426
9	[39.0,	66.0,	70.0,	53.0]	139	99166.10294329426
9	[53.0,	21.0,	41.0,	54.0]	153	99166.10294329426
10	[53.0,	21.0,	41.0,	54.0]	167	95504.97364475316
10	[83.0,	57.0,	51.0,	54.0]	140	95504.97364475316
10	[30.0,	61.0,	59.0,	53.0]	119	95504.97364475316
11	[65.0,	49.0,	71.0,	50.0]	49	87124.98403354186
11	[31.0,	37.0,	41.0,	52.0]	66	87124.98403354186
11	[62.0,	75.0,	46.0,	59.0]	140	87124.98403354186
12	[83.0,	40.0,	67.0,	60.0]	74	82174.32679393479
12	[83.0,	75.0,	33.0,	49.0]	62	82174.32679393479
12	[53.0,	21.0,	41.0,	54.0]	66	82174.32679393479
12	[43.0,	60.0,	76.0,	38.0]	76	82174.32679393479
12	[23.0,	61.0,	50.0,	61.0]	77	82174.32679393479
13	[39.0,	66.0,	70.0,	53.0]	118	91899.73445527659
13	[83.0,	57.0,	51.0,	55.0]	61	91899.73445527659
13	[31.0,	36.0,	40.0,	53.0]	38	91899.73445527659
14	[53.0,	21.0,	41.0,	54.0]	76	86466.68707651942

Question-e) What stopping criterion did you use for your inner loop?

Solution: The stopping criteria used for the inner loop was if the new centroids which are computed are similar to the old centroid then the inner loop stops.

Question-f.) What was the hardest part of getting all this working? Did anything go wrong?

Solution: The hardest part was removing the centroid and the reducing the cluster value when no data points are associated with some the Initial randomly generated centroids .Calculation of Euclidean distance was observed as wrong so later on the Euclidean distance from Scipy package was used which yielded the correct results.

Question-g.) Conclusion: What did you learn about all of this? Will you remember the kMeans algorithm for the next quiz? Will you remember it for a job interview? Did you try using the L1 norm to compare to? Did you compare your results to a standard package, such as sci-kit learn or R… did your results match?

Solution:

Yes K-means algorithm can be remembered for next quiz. Yes it can be remembered for Job interview. No , L1 Norm was never used for comparison. Results were not compared with other packages