Output Analysis

Ans. a) Performed K-means clustering on powerusage data with k=6 between global_active_power, global_reactive_power and global_intensity variables.

Cluster Summary

Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	7	1.0698	3.5437		4	5.4027
2	82	0.8294	2.9015		4	4.9825
3	683	0.7644	3.1264		5	5.1425
4	29	0.7601	2.5912		2	4.9825
5	440	0.7938	2.8084		3	5.1425
6	199	0.9519	3.1262		3	6.2198

Ans. b) Before removing the outliers 5 iterations are performed. The initial centroid for Cluster 2 is: (4.9680, 0.0, 20.80) and the final centroid is: (4.61, 0.10, 19.62).

Initial Seeds

Cluster	Global_active_power	Global_reactive_power	Global_intensity
1	7.70600000	0.00000000	33.20000000
2	4.96800000	0.00000000	20.80000000
3	1.79800000	0.47000000	7.80000000
4	6.47400000	0.14400000	27.80000000
5	3.35400000	0.23800000	14.40000000
6	0.20600000	0.00000000	0.80000000

Cluster Means

Cluster	Global_active_power	Global_reactive_power	Global_intensity
1	6.93257143	0.08971429	29.74285714
2	4.61846341	0.10346341	19.62439024
3	2.10293704	0.12840703	8.86676428
4	5.71172414	0.26455172	24.48275862
5	3.27419545	0.11561364	13.87409091
6	0.56529648	0.08516583	2.84020101

After removing the outliers 7 iterations are performed. The initial centroid for Cluster 2 is: (2.038, 0.068, 9.4) and the final centroid is: (2.321, 0.1066, 9.691)

Output Analysis

Initial Seeds			
Cluster	Global_active_power	Global_reactive_power	Global_intensity
1	0.20600000	0.00000000	0.80000000
2	2.03800000	0.06800000	9.40000000
3	3.28200000	0.04600000	13.60000000
4	0.85600000	0.00000000	5.20000000
5	4.17000000	0.26800000	18.00000000
6	5.28200000	0.15000000	22.40000000

Cluster Means			
Cluster	Global_active_power	Global_reactive_power	Global_intensity
1	0.40880000	0.08547586	1.99034483
2	2.32101721	0.10664245	9.69101338
3	3.23945000	0.09091875	13.70937500
4	1.46320362	0.12618100	6.61176471
5	3.76564286	0.17404762	16.06904762
6	4.74150000	0.06750000	20.09687500

Ans. c) Initially before removing the outliers from the data the Cubic Clustering Criterion was 1.268. But we know that values of the cubic clustering criterion greater than 2 or 3 indicate good clusters. Values between 0 and 2 indicate potential clusters. Hence after removing the outliers from the data the Cubic Clustering Criterion is 23.803 which is good. One good suggestion to get a good quality cluster is to remove the outliers from the data.

Cubic Clustering Criterion =	1.268

Cubic Clustering Criterion =	23.803

Ans. d) After removing the outliers from the data (steps suggested in Ans.c) we have performed K-means clustering with K=6 and the quality of clustering is good based on Cubic Clustering Criterion which is 23.803 (more than 2).
Following is the comparison before and after removing outliers :-

Output Analysis

Before removing outliers	After removing outliers
1) Number of records distributed among clusters is highly uneven. For Eg :-	1) Number of records distributed among clusters is even. For Eg :-

Cluster	Frequency	RMS Std Deviation
1	7	1.0698
2	82	0.8294
3	683	0.7644
4	29	0.7601
5	440	0.7938
6	199	0.9519

As we can observe that in above snapshot cluster 1 have only 7 elements whereas cluster 3 have 683. Hence number of elements are higly uneven

Cluster	Frequency	RMS Std Deviation
1	145	0.5353
2	523	0.5033
3	320	0.5451
4	218	0.6309
5	84	0.4982
6	64	0.6408

As we can observe that all clusters have elements evenly distributed.

2) Standard deviation for elements in a cluster is higher.

2) Standard deviation for elements in a cluster is lower.

3) Cubic clustering criterion value without removing outliers is 1.268, which indicates that the cluster formed is potential cluster.

Cubic Clustering Criterion =	1.268

3) Cubic clustering criterion value without removing outliers is 23.812, which indicates that the cluster formed is of a good quality.

Cubic Clustering Criterion =	23.812

4) F-test value for clusters created is lower.

Pseudo F Statistic =	3637.00

4) F-test value for clusters created is higher.

Pseudo F Statistic =	6124.91