Columns Dropped: REASON

1] COBALT\_JOBID: All job IDs are Unique, hence no need to consider this as a parameter for cluster formation.

2] DARSHAN\_LIB\_VERSION & DARSHAN\_LOG\_VERSION: These could not be considerable parameters

3] END\_TIME & START\_TIME: Instead of these, we considered RUN\_TIME parameter as this parameter is

Actual time elapsed between operations. (END\_TIME – START\_TIME)

4] EXE\_NAME\_GENID: This is too long string of number which affecting algorithm outputs.

5] MACHINE\_NAME: This parameter is a String so cannot be considered in numerical analysis-based algorithms

6] RUN\_DATE\_ID: This parameter has many identical values.

7] TOTAL\_MPIIO\_COLL\_READS: This parameter has summation value of another columns.

8] TOTAL\_MPIIO\_COLL\_WRITES: This parameter has summation value of another columns.

9] TOTAL\_READ\_OPS: This parameter has summation value of another columns.

10] TOTAL\_WRITE\_OPS: This parameter has summation value of another columns.

Summation values are too big, they are changing the result of algorithms.

Columns Selected

1] NPROCS

2] RUN\_TIME

3] TOTAL\_IO\_TIME

4] TOTAL\_POSIX\_F\_META\_TIME

5] TOTAL\_POSIX\_F\_READ\_TIME

6] TOTAL\_POSIX\_F\_WRITE\_TIME

7] TOTAL\_IO\_PER\_PROC

8] TOTAL\_IO\_OPS

9] TOTAL\_MD\_OPS

10] TOTAL\_READ\_TIME

11] TOTAL\_WRITE\_TIME

12] TOTAL\_MD\_TIME

After selecting these columns, we removed the outliers based on quantile values.

Then, we separated real & debug apps based on following criterion:

1. Lower 25% values among all selected columns would be debug apps.
2. The common rows amongst all the lower 25% values rows calculated in (a).
3. Remaining 75% would be real apps.

Calculated elbow value using Distortions & Inertia: we got elbow value 3.

We run the KMeans( ) with number of clusters 3, and got results of 3 clusters.

We are doing cluster analysis now, [What the obtained clusters do mean?]