Report for Task 1 – s1558717

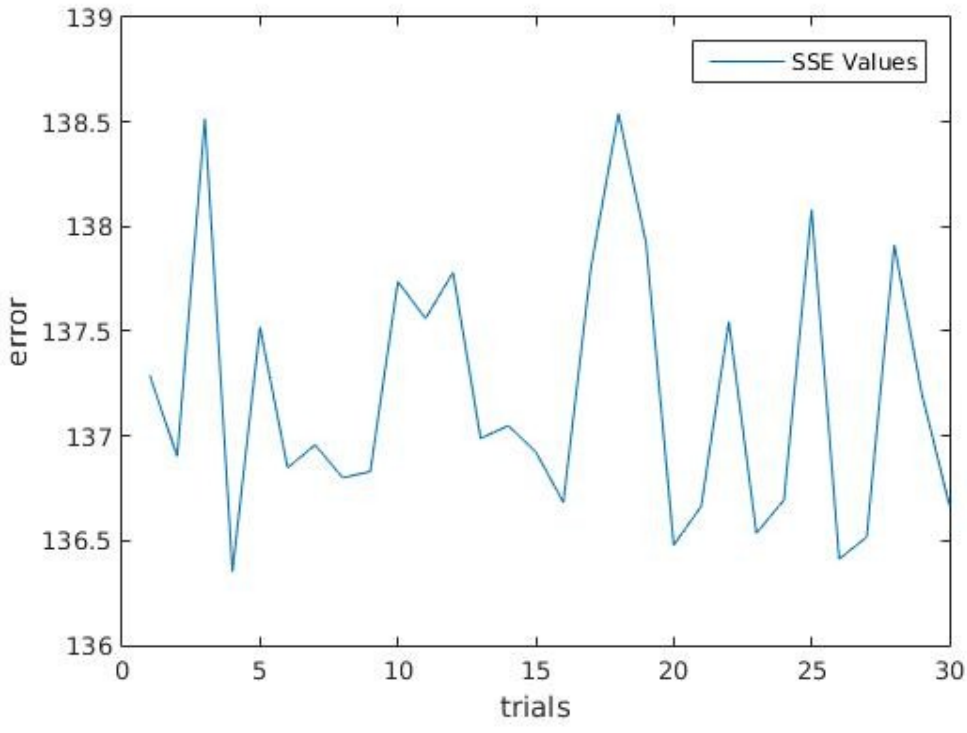
556	99	122	112	47	750	232	82
269	34	378	295	433	132	391	68
249	364	31	103	94	75	378	706
467	148	259	248	145	206	419	108
131	641	193	570	32	86	112	235

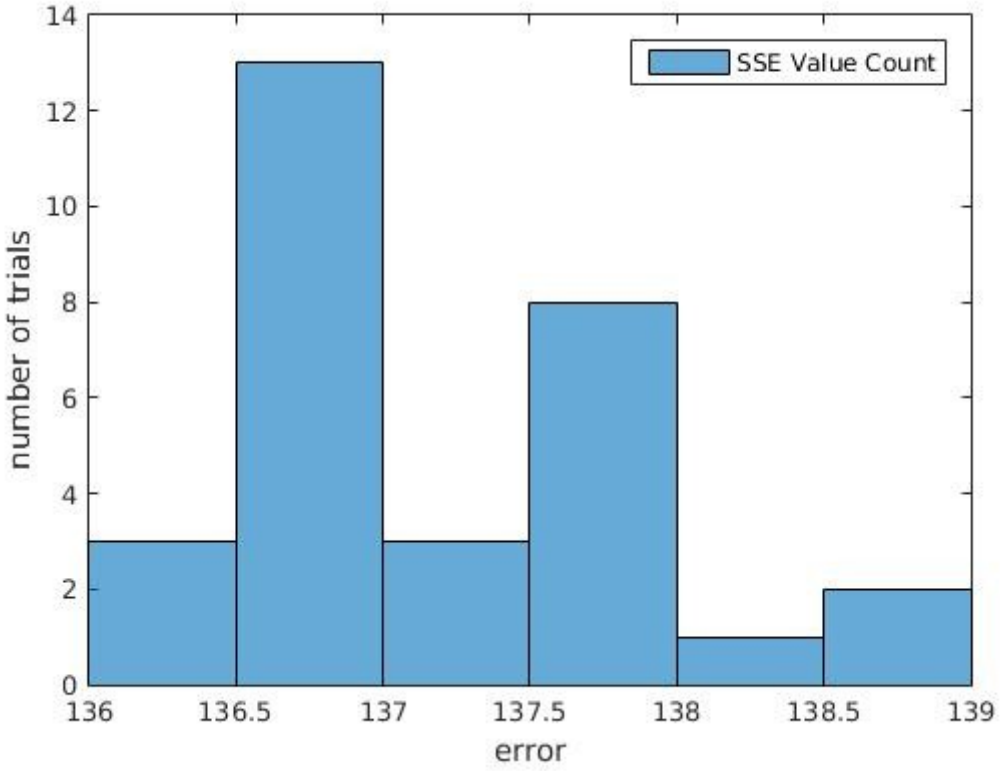
The confusion matrix is a visualisation tool used in supervised training. The columns of the matrix represent the values in a predicted class, whereas the rows represent the values in the true class.

For example, the value in the cell (1,1) indicates the many data points that were supposed to be clustered under class 1 and have actually been clustered under class 1.

An ideal confusion matrix represents all values along the leading diagonal with the remaining values being zero. We have already found out from our large Mean Squared Error value that our correct classification rate isn't very high.

The final SSE that we obtained after 10 iterations was 137.178.





As we can see here, most of the SSE values in the 30 trials were between 136.5 and 137. This was quite close to the final SSE obtained from the 10 iterations using the initial 8 data points as cluster centres.

However, in these 30 trials of 10 iterations each, we used random cluster centres each time and hence the SSE value was very high in the first iteration of each trial (more than 180 as observed in some cases).

However, using random cluster centres wasn't very effective and only lowered the average final SSE in every trial by a small relative value if any.

Clustering performance could possibly be improved by finding the mean for the whole data seta and then perturbing this into K means. This would help decrease the SSE and number of trials needed to reach convergence.