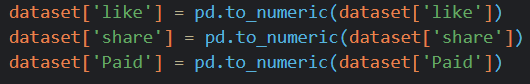
**Pre-Processings:**

* As some of the values in the dataset were string, it was required to convert them into numeric data. I used **panda** function **to\_numeric** to convert types.

****

* Furthermore, there were some missing values marked with an “?”. I replaced them with 0 to make calculations simpler.



* Moreover, I dropped some of the features and chose those which seemed to have more information. This was totally subjective and random and I didn’t perform any comparisons to check if one feature has more information than the other. I used **pandas** function called **drop** to achieve this task.





* Finally I converted the dataset into numpy dataframe for quicker and more powerful calculations. I used **to\_numpy** function to do that.

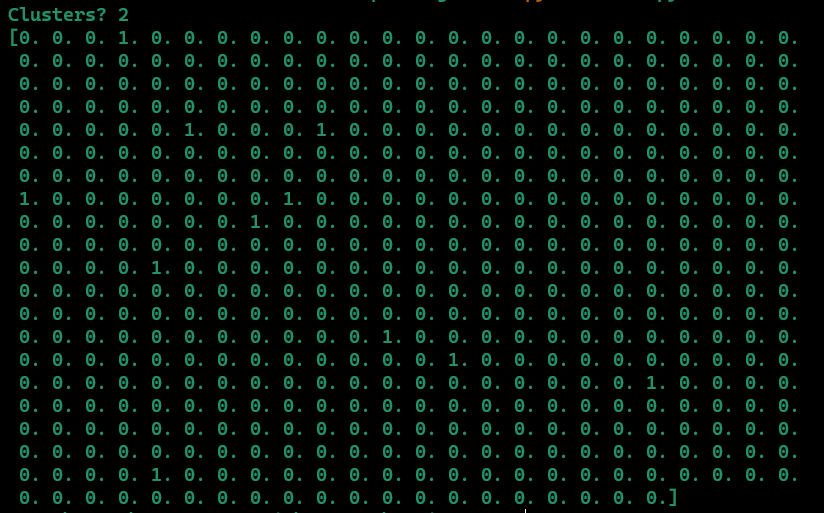


**Clustering Results:**

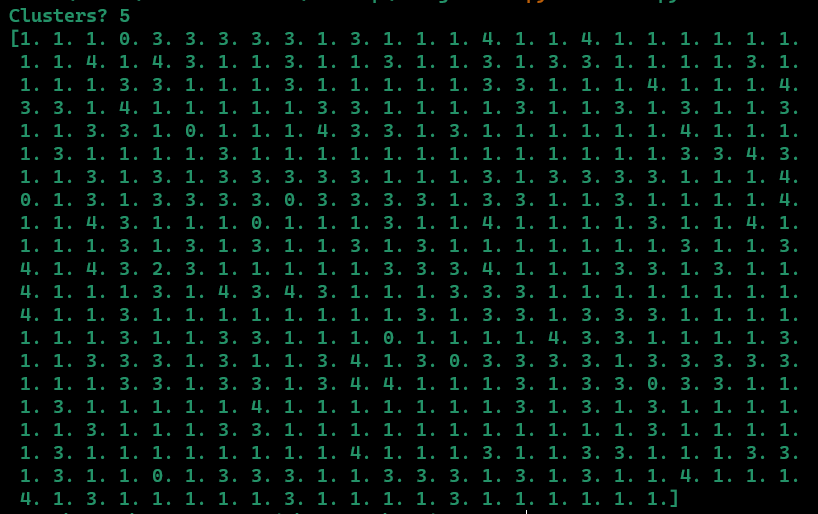
**Maximum Iterations = 500**

There were 500 total samples and each sample was assigned a cluster (centroid) after the iterations.

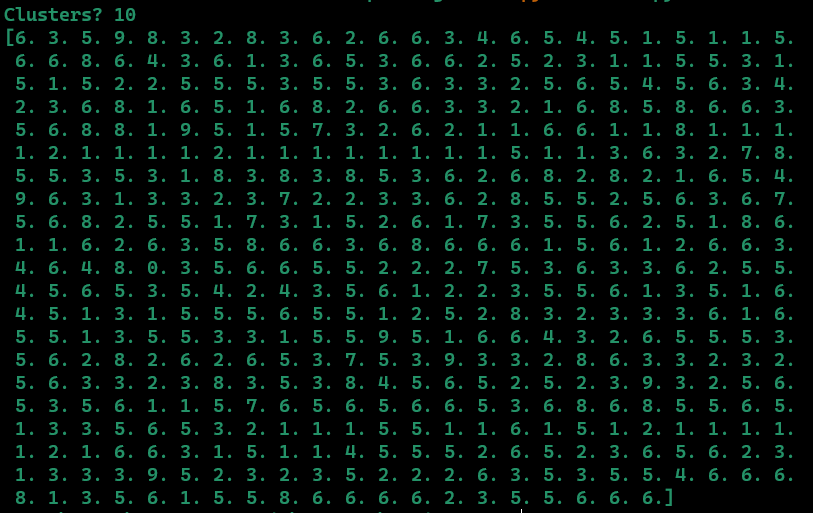
**K = 2**

****

**K = 5**

****

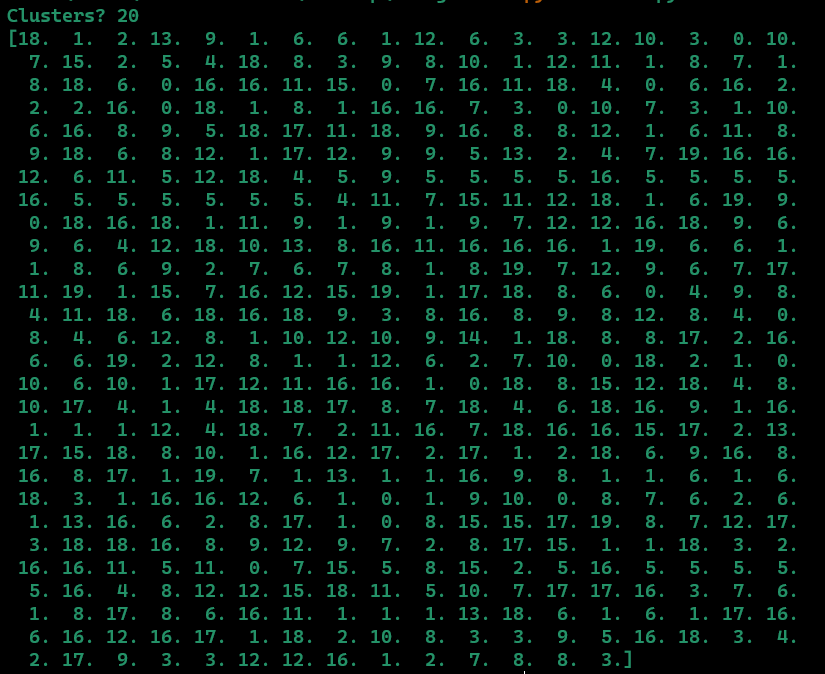
**K = 10**

****

**K = 15**

****

**K = 20**

****

**Findings Against Different Values of K**:

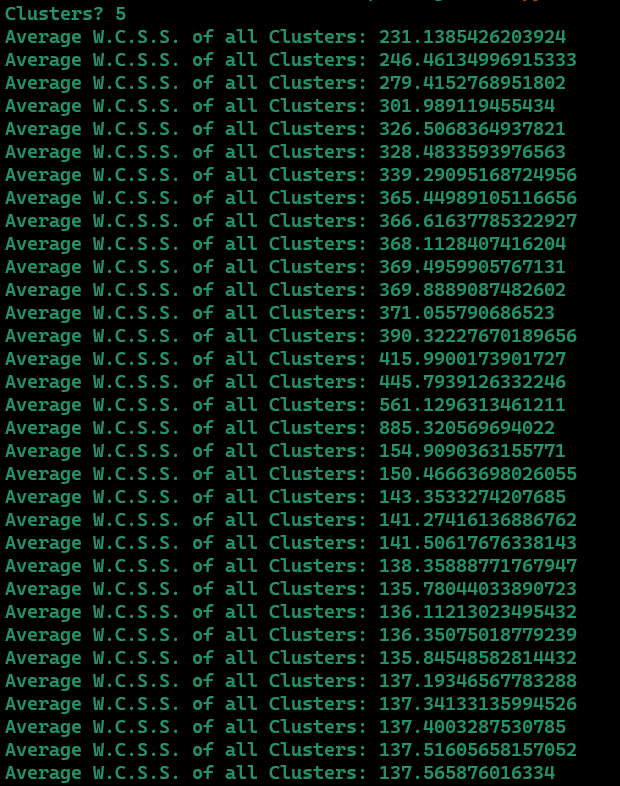
**Maximum Iterations = 500**

* I my opinion the k = 20 gave quite well balanced results. The area under each cluster was quite balanced.
* Furthermore k = 2 produced the worst results because there were **outliers** which seemed to lie in a 3rd cluster but because there were only 2 so they had to be assigned a farther centroid. The average W.C.S.S. Error was quite inconsistent too with k = 2.
* K = 50 took the longest time to converge because it had a lot of updation and calcultions involved.

**Results of Within Cluster Sum of Squares (W.C.S.S.):**

***Note: I printed out the average error of all clusters instead of printing it for each individual cluster. The reason was the screenshot of individual ones was too large to document here.***

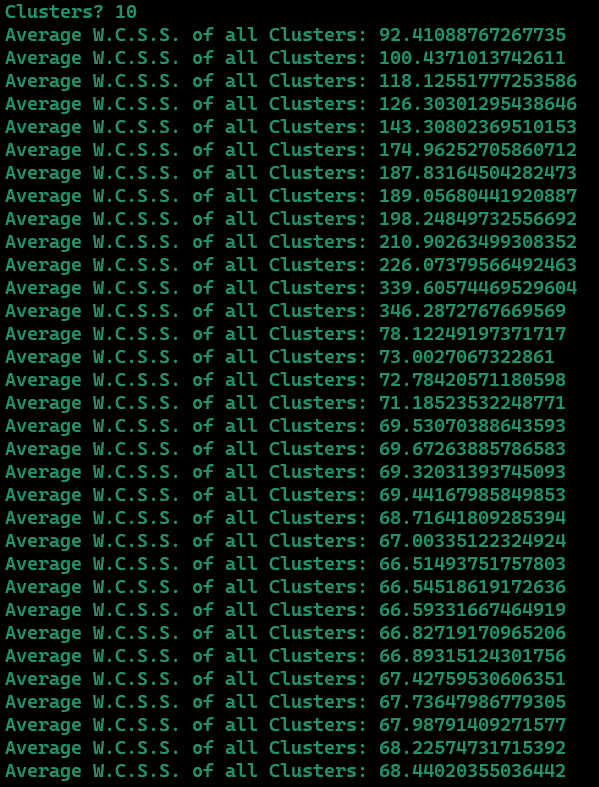
**For k = 5, Iterations Taken = 33**

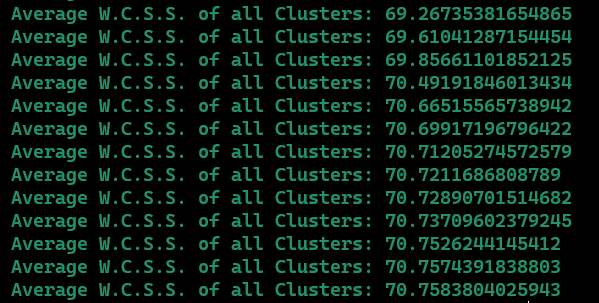
****

**For k = 2, Iterations Taken = 10**

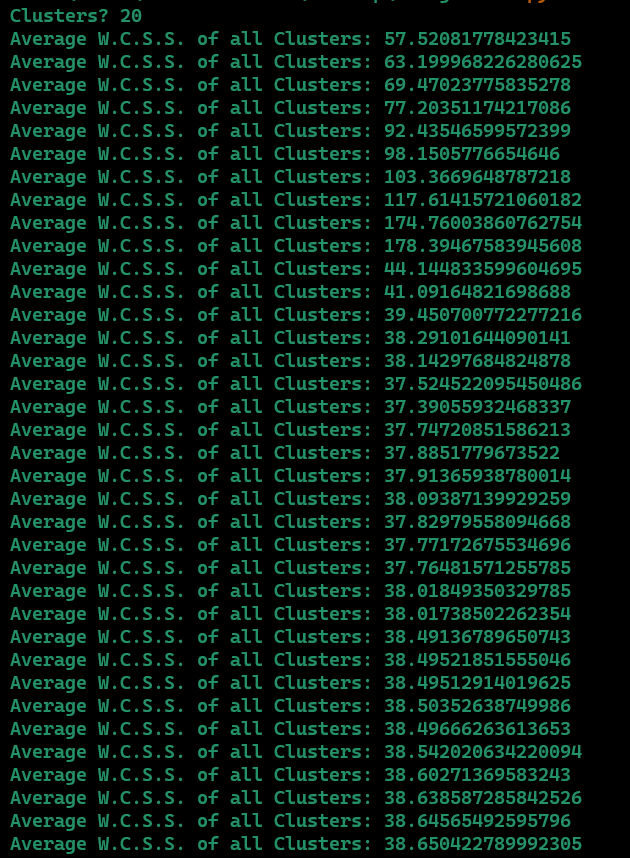


**For k = 10, Iterations Taken = 46**





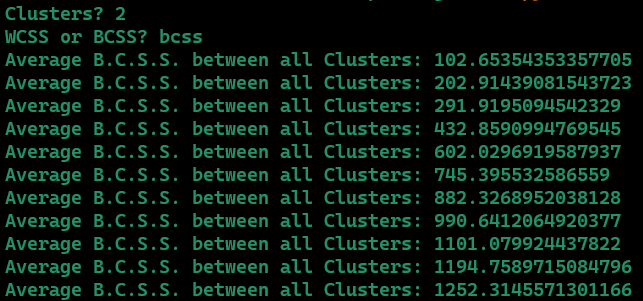
**For k = 20, Iterations Taken = 36**

****

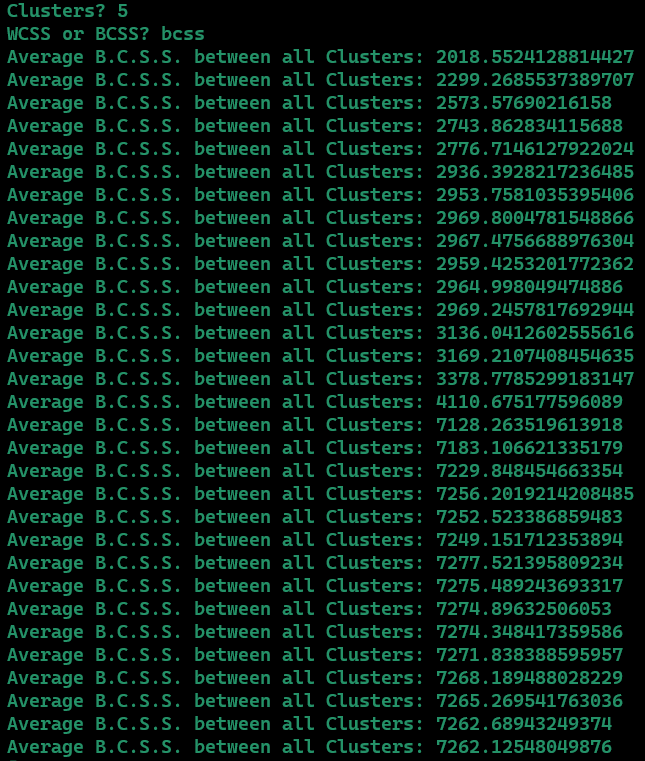
**Results of Between Cluster Sum of Squares (B.C.S.S.):**

***Note: I printed out the average error of all clusters instead of printing it for each individual cluster. The reason was the screenshot of individual ones was too large to document here.***

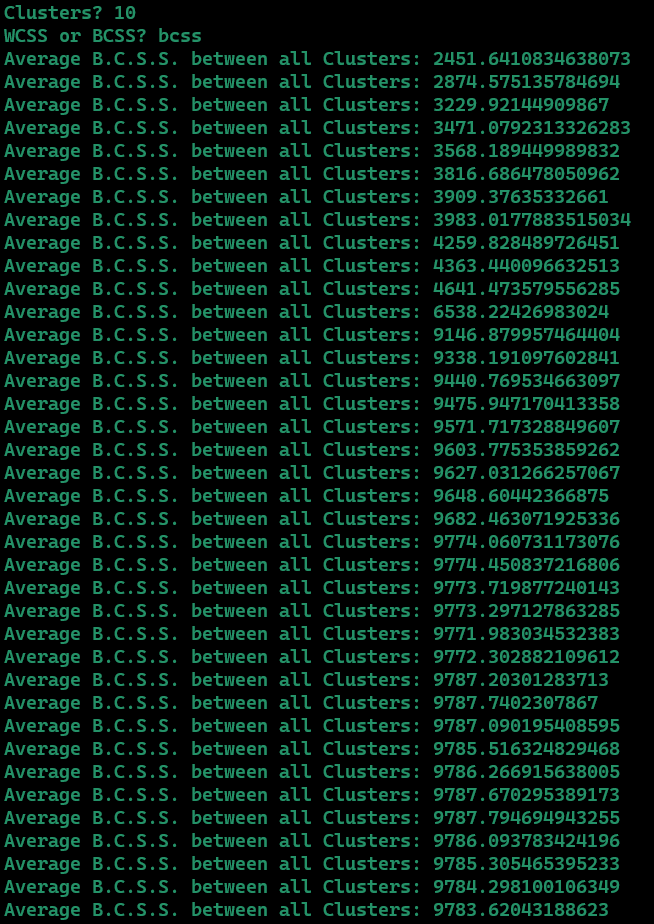
**For k = 2, Iterations Taken = 11**

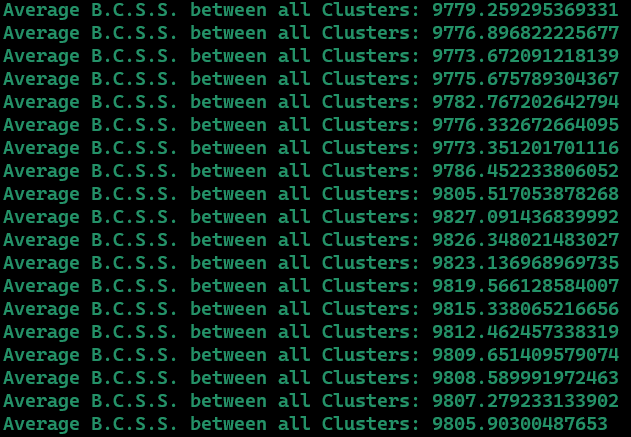
****

**For k = 5, Iterations Taken = 31**

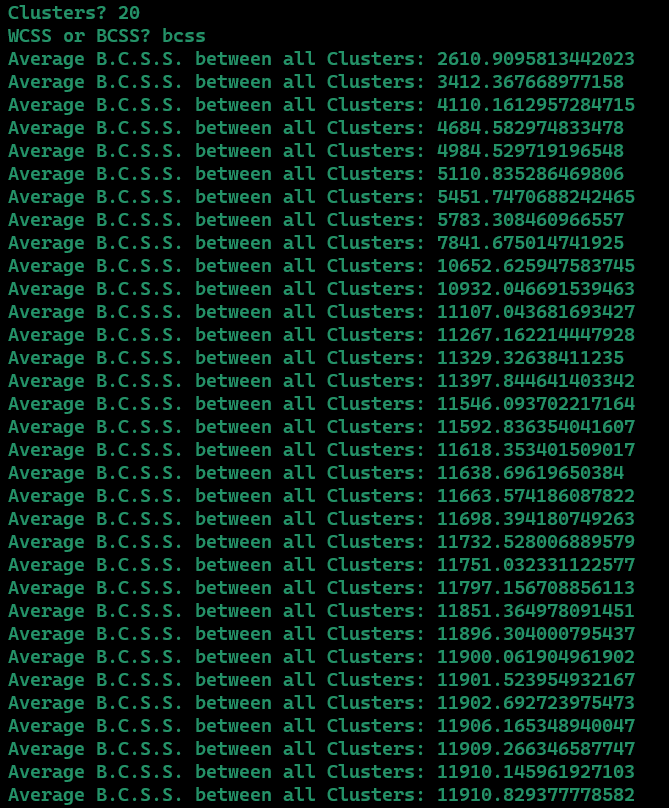
****

**For k = 10, Iterations Taken = 58**

****

****

**For k = 20, Iterations Taken = 33**

****