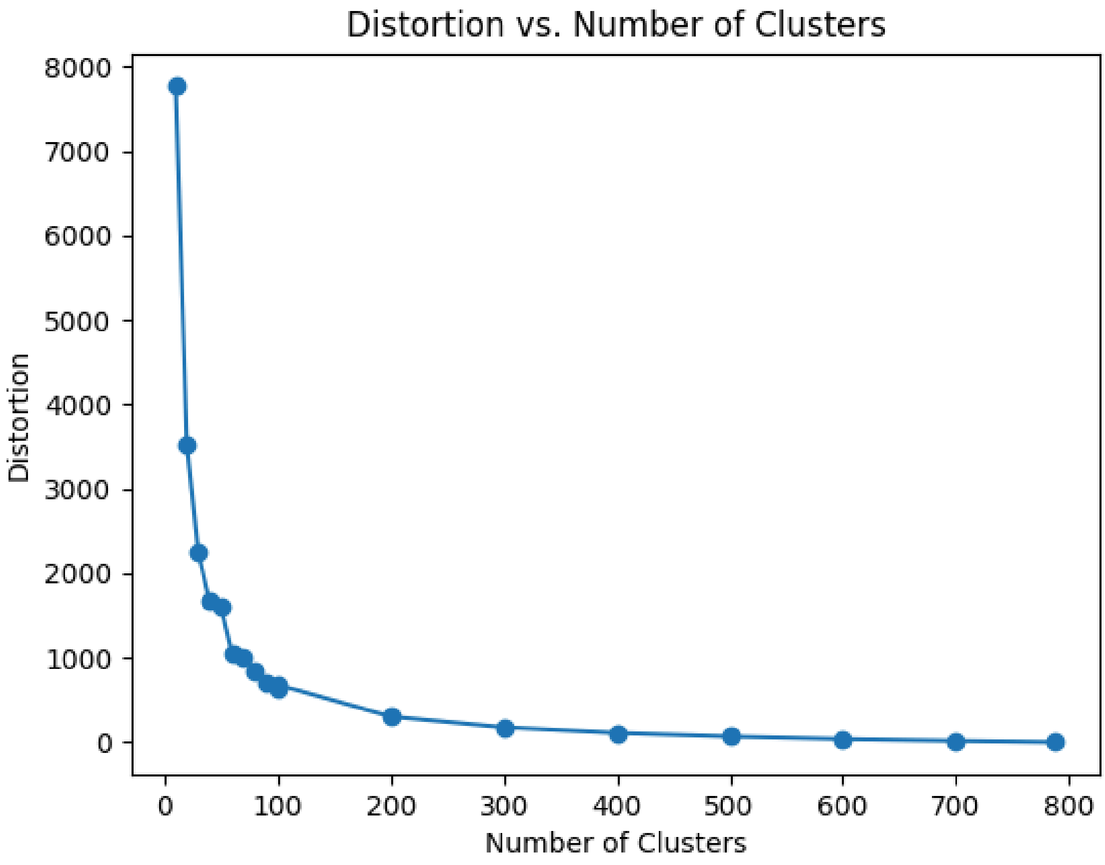
1.What’s the minimum and maximum number of possible clusters when you have n points?

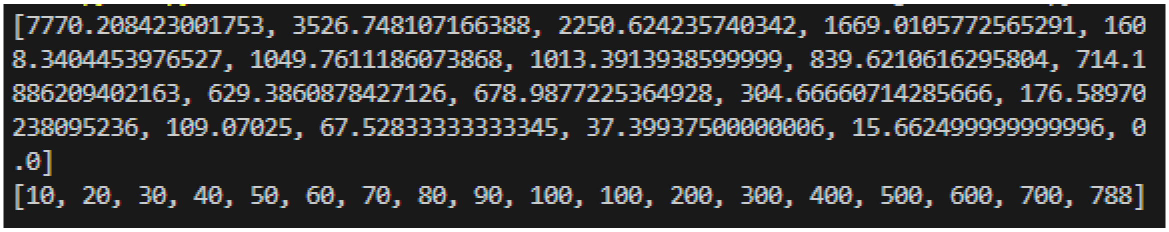
Minimum number is 1;
Maximum number is n;

2.Experiment with different numbers of clusters, from minimum to maximum. Record the distortions while you change the parameter for kmeans. Draw a graph to show how distortion changes with different numbers of clusters.

This is the plot of the distortion changes with different numbers of clusters.



And this one is the output of distortion with the specific number of clusters below.



```
[7770.208423001753, 3526.748107166388, 2250.624235740342, 1669.0105772565291, 160
8.3404453976527, 1049.7611186073868, 1013.3913938599999, 839.6210616295804, 714.1
886209402163, 629.3860878427126, 678.9877225364928, 304.66660714285666, 176.58970
238095236, 109.07025, 67.52833333333345, 37.39937500000006, 15.662499999999996, 0
.0]
[10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 100, 200, 300, 400, 500, 600, 700, 788]
```

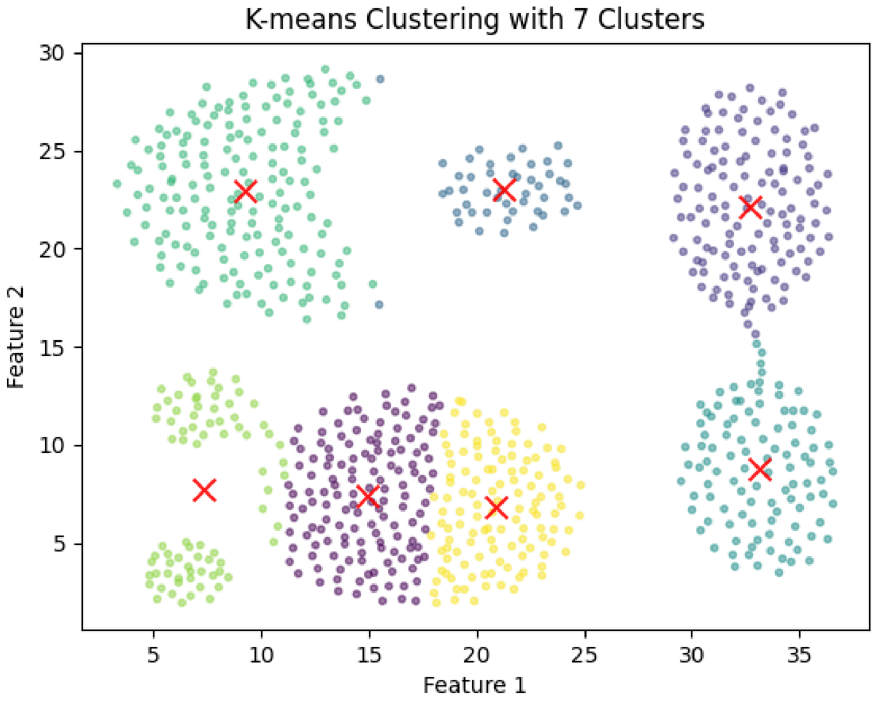
3.What’s the lowest possible distortion? When will this happen? Explain your answers.

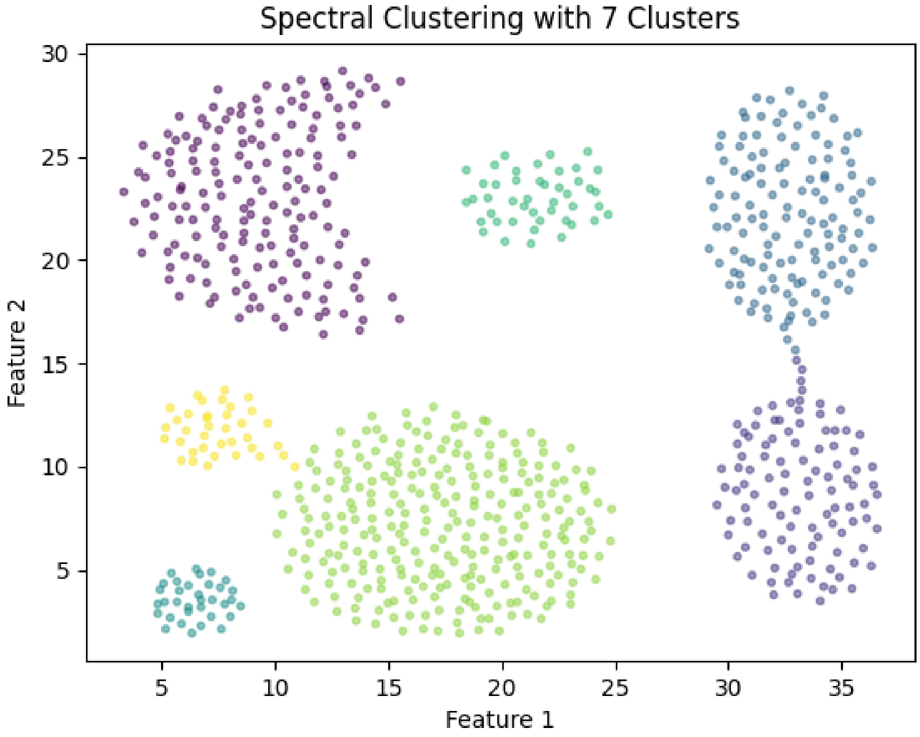
We can see in the images above,the lowest distortion is 0, since the number of data points is 788,and when the number of clusters is the same as the number of data points,every cluster has only one data point,therefore, the total distortion is 0.

4.What’s the optimal number of clusters? How can we get this number in a program (without any human intervention/interaction)?

We can use the Sum of Squares Method;it chooses the optimal number of clusters by minimizing the within-cluster sum of squares (a measure of how tight each cluster is) and maximizing the between-cluster sum of squares (a measure of how separated each cluster is from the others).Here is the reference:https://towardsdatascience.com/10-tips-for-choosing-the-optimal-number-of-clusters-277e93d72d92

Part2b:


K-means Clustering with 7 Clusters
Feature 2
Feature 1
30
25
20
15
10
5
5
10
15
20
25
30
35


Spectral Clustering with 7 Clusters
Feature 2
Feature 1
30
25
20
15
10
5
5
10
15
20
25
30
35

Part2c:

(1)Compare the results of the two clustering methods.Comment on why one method is performing better/worse than the other.

Spectral clustering is performing better in this scenario because it can handle the non-linear separability of clusters. It considers the global relationships among data points, making it more adept at identifying clusters that are not well-separated by linear boundaries.
K-means clustering, relies on Euclidean distance, which can be a limitation in cases where clusters have irregular shapes or densities. This can lead to less accurate clustering when the data does not meet these assumptions, as potentially seen in the provided plots.

(2)Suggest a way to find the optimal sigma

We can use a heuristic to find the optimal sigma.
We can do Sigma Estimation: The first step involves estimating an optimal sigma value for the "Ng kernel." This can be done by evaluating the clustering outcome for a range of sigma values and selecting the one that results in the best clustering performance. And then do the Automatic Tuning.

reference:chrome-extension://cdonnmffkdaoajfknoeeecmchibpmkmg/assets/pdf/web/viewer.html?file=https%3A%2F%2Fcran.r-project.org%2Fweb%2Fpackages%2FSpectrum%2Fvignettes%2FSpectrum_vignette.pdf