Project #2 Analysis

Intro:

At the start of this project we struggled to understand what was required of us. We were unsure of how to implement k-means clustering in a c program and found the concept of multiple dimensions difficult to grasp. After thorough research, we better understood the concept and began implementing this in our program.

General explanation of code:

In the first part of our code, we read the input from the file in order to retrieve the number of clusters, number of dimensions, and number of data. Next we stored the data in a 2-D array, then chose a random data point as our centroids. We calculated the distance of each data from the centroid and from this, we were able to assign each data to a cluster. In order to calculate the position of the new centroids, we had to get the average distance of the data within each cluster. We used a do while loop to repeat the process until the centroid position did not change anymore - this meant that all the data were sorted into their respective clusters.

Analysis and testing (including utilising various data sets):

We tested our code throughout the development process by utilising print statements within blocks of code to ensure they were producing the intended output. We have left these print statements in our program to allow testers and code reviewers to easily see how each part of the code functions, and how it all works in conjunction to accomplish the purpose of this project. To ensure our code worked with diverse inputs, we created multiple, varying test text files and ran them through our code. We found that with smaller data sets, we saw less iterations of the program before the data points inside the clusters were unchanging and definite.

Applications of program:

There are many applications of k-means clustering, such as image segmentation (can be used in medical imaging and pattern recognition), market research, creation of recommendation engines and much more. This ability to get a meaningful structure of the data we are working with means we can better understand the significance of the data we are working with. We look forward to applying what we've learnt in future projects and endeavors.

Resources:

This is the video that helped us understand how to implement kmeans clustering in a c program:

- https://www.youtube.com/watch?v=_S5tvagaQRU

Team Members:
2019067574 Janae Fariñas
2019073681 Sophie Nugent
2019041349 Joshua O'Leary

Results

Data Set 1	Terminal Output
5 4 30 10 19 24 22 28 4 20 4 9 11 24 5 3 13 8 29 16 17 20 7 27 10 3 30 8 17 1 23 28 1 24 10 19 3 15 3 29 18 23 18 2 13 7 6 17 21 18 23 10 13 24 19 21 11 1 9 18 14 16 26 21 22 12 6 1 16 17 6 10 8 29 23 9 24 22 15 5 24 21 11 14 18 19 23 3 1 11 29 12 22 24 21 22 11 12 15	1 8 9 14 15 16 24 25 4 7 11 22 27 2 3 5 17 26 28 6 10 30 1 12 13 18 19 20 21 23 29 2 2 8 9 14 16 24 25 4 7 11 22 27 3 5 17 26 28 6 10 15 30 1 12 13 18 19 20 21 23 29 3 2 8 9 14 16 24 25 4 7 22 27 3 5 11 17 20 26 28 6 15 30 1 10 12 13 18 19 21 23 29 4 2 8 9 14 16 24 25 4 7 22 27 3 5 11 17 20 28 6 15 30 1 10 12 13 18 19 21 23 26 29

Data Set 2	Terminal Output
3 2 15 2 14 21 23 23 14 9 23 8 7 13 4 12 18 8 13 7 13 11 6 18 7 9 19 23 20 20 9 11 25	1 1 5 6 8 9 10 4 7 12 15 2 3 11 13 14

Team Members:
2019067574 Janae Fariñas
2019073681 Sophie Nugent
2019041349 Joshua O'Leary

Data Set 3	Terminal Output
4 5 20 41 45 7 83 8 49 2 91 56 61 33 20 26 8 98 49 88 5 78 35 12 14 93 60 47 37 11 69 22 71 81 89 75 11 48 28 79 36 52 87 70 97 32 74 50 38 8 49 58 5 17 7 55 22 74 17 55 7 10 60 82 86 88 16 99 42 97 60 21 46 71 96 42 17 97 56 89 9 62 8 61 90 12 16 67 58 9 16 85 29 39 20 1 71 4 92 39 49 75 85	1 7 8 13 14 15 17 1 4 9 16 3 6 11 12 20 2 5 10 18 19

Data Set 4	Terminal Output
4 3 40 63 62 45 34 39 90 39 49 84 52 79 64 22 19 17 42 89 97 8 2 97 31 56 11 92 18 14 51 68 29 19 62 60 60 52 14 12 51 86 54 80 92 43 27 47 92 84 9 81 3 74 5 46 68 90 7 87 64 66 97 71 84 64 11 98 20 51 26 1	1 2 3 4 6 7 13 14 17 19 20 21 29 31 34 35 39 5 8 11 18 22 24 27 28 40 9 15 23 33 36 1 10 12 16 25 26 30 32 37 38 2 2 3 6 7 13 14 19 20 29 31 34 39 8 11 18 22 24 25 27 28 35 40 5 9 15 17 23 30 33 36 1 4 10 12 16 21 26 32 37 38 3 2 3 6 7 13 14 20 29 31 34 39 8 11 18 22 24 25 27 28 35 38 40 5 9 15 17 19 23 30 33 36 1 4 10 12 16 21 26 32 37 4 2 3 6 7 13 14 20 27 29 31 34 39 8 11 18 22 24 25 28 35 38 40 5 9 15 17 19 23 30 33 36 1 4 10 12 16 21 26 32 37 5 2 3 6 7 13 14 18 20 27 29 31 34 39 8 11 22 24 25 35 38 40 5 9 15 17 19 23 28 30 33 36 1 4 10 12 16 21 26 32 37 6 2 3 6 7 11 13 14 18 20 25 27 29 31 34 39 8 22 24 35 38 40 5 9 15 17 19 23 28 30 33 36

Team Members:
2019067574 Janae Fariñas
2019073681 Sophie Nugent
2019041349 Joshua O'Leary