Name: Amirtha Ganesh Pugazhendhi **Assignment 3**

Structure & function of the Program:

1. **File used:** The assignment uses only one ".c"file : **"rand_norm_kmeans_cluster.c"**
   - "rand_norm_kmeans_cluster.c " – This file consist of c programming codes for the entire assignment.
2. **Data Structure used: Single dimensional arrays** , **Two-dimensional arrays, int and double** were the data types used.
3. **Function defined:**
   1) "rand_centroid_init( )" : To allocate initial random centroid points for the initial k clusters
   2) "compute_average( )" : To compute the Average across all the columns
   3) "compute_stdev( )" : To compute the Standard Deviation across all the columns
   4) "compute_norm( )" : To perform normalization of data points
   5) "allocate_cluster( )" : To identify which cluster does the data point belongs to
   6) "update_centroid( )" : To determine the new centroid location based on clusters
   7) "main( )" : To execute the entire program

4) Design of Output:

1) "Data Table" : It displays a table of data points, cluster value , last iteration No., RMSE and RSME rate.

2) "Cluster Summary" : It displays the summary of Averages, Standard deviation , Table of Cluster Centroids.

Name: Amirtha Ganesh Pugazhendhi

Unit Testing was found to successful for both varying values of k & input files and Error Cases.
Proof of Test validation screen shots are attached for each test cases

Name of the executable file: ".\test"

Output for varying k and input files:

K= 4 ; input file :"WineData_col2.txt"
(No of Columns =2)

```
Threshold for RMSE = 1x10^-6
;
Average for column 1 = 2.538806

Standard Deviation for column 1 = 1.409487
;
Average for column 2 = 46.467792

Standard Deviation for column 2 = 32.885037
;;
```

Cluster No.	Co. position	Co-ordinate value
1	1	-0.129457
1	2	1.800462
2	1	-0.214119
2	2	-0.715579
3	1	-0.145826
3	2	0.265832
4	1	3.509853
4	2	0.842294

K= 3 ; input file :"WineData_col3.txt"
(No of Columns = 3)

```
Threshold for RMSE = 1x10^-6
;
Average for column 1 = 2.538806

Standard Deviation for column 1 = 1.409487
;
Average for column 2 = 46.467792

Standard Deviation for column 2 = 32.885037
;
Average for column 3 = 10.422983

Standard Deviation for column 3 = 1.065335
;;
```

Cluster No.	Co. position	Co-ordinate value
1	1	0.760819
1	2	1.642524
1	3	-0.578895
2	1	-0.285789
2	2	-0.311866
2	3	-0.587110
3	1	0.014458
3	2	-0.409203
3	3	1.150315

Output for Error cases:

./test WineData_col2.txt

```
|--------------- ERROR ---------------------------------------|
USAGE:./executableName input_file k_value
You must pass your data file  and k value (in that order)as an argument to this program.

|--------------- ERROR ---------------------------------------|
```

./test WineData_col2.txt -6

```
|--------------- ERROR ---------------------------------------|
The k-value input was:  -6
K must be larger than zero, and no larger than 2147483647.

|--------------- ERROR ---------------------------------------|
```

./test WineData.txt 6

```
|--------------- ERROR ---------------------------------------|
Could not open the file:        WineData_col.txt
Failed to open file: No such file or directory

|--------------- ERROR ---------------------------------------|
```

./test output.txt

```
|--------------- ERROR ---------------------------------------|
Error:  File was not in correct format.

|--------------- ERROR ---------------------------------------|
```

Name: Amirtha Ganesh Pugazhendhi

3) Importance of data normalization for K-Means clustering

Normalization is used to **eliminate redundant data** and ensures that **good quality clusters** are generated which can improve the efficiency of clustering algorithms. So it becomes an essential step before clustering as **Euclidean distance** is **very sensitive** to the **changes** in the differences. Without normalization, the variable with the **largest scale** will **dominate the measure**.

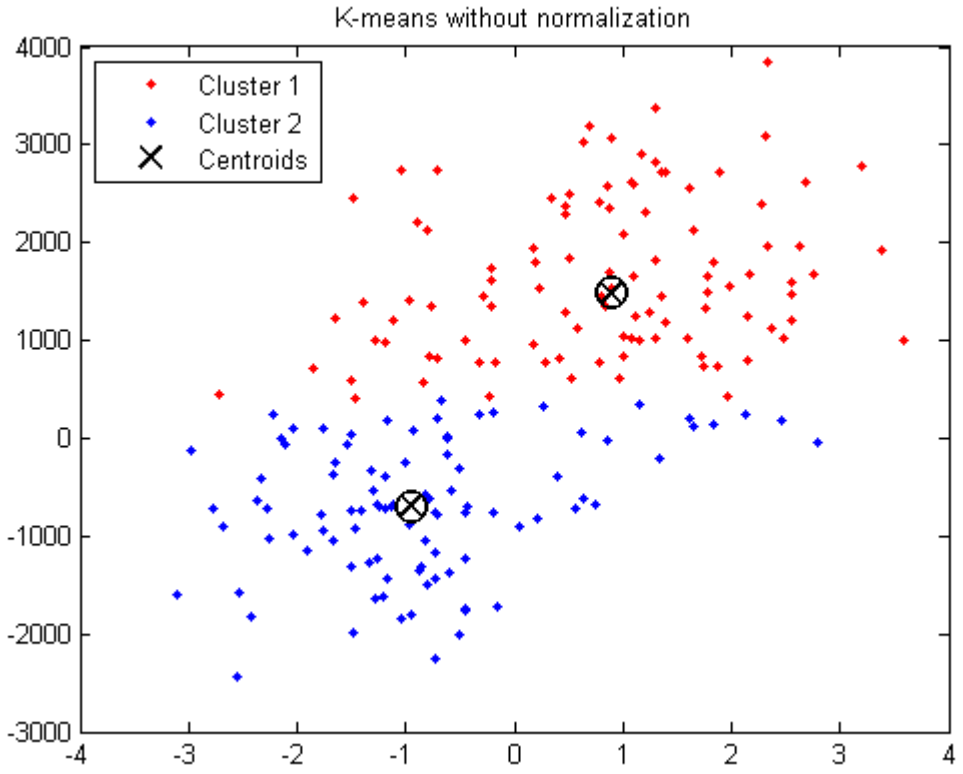


Fig. 1

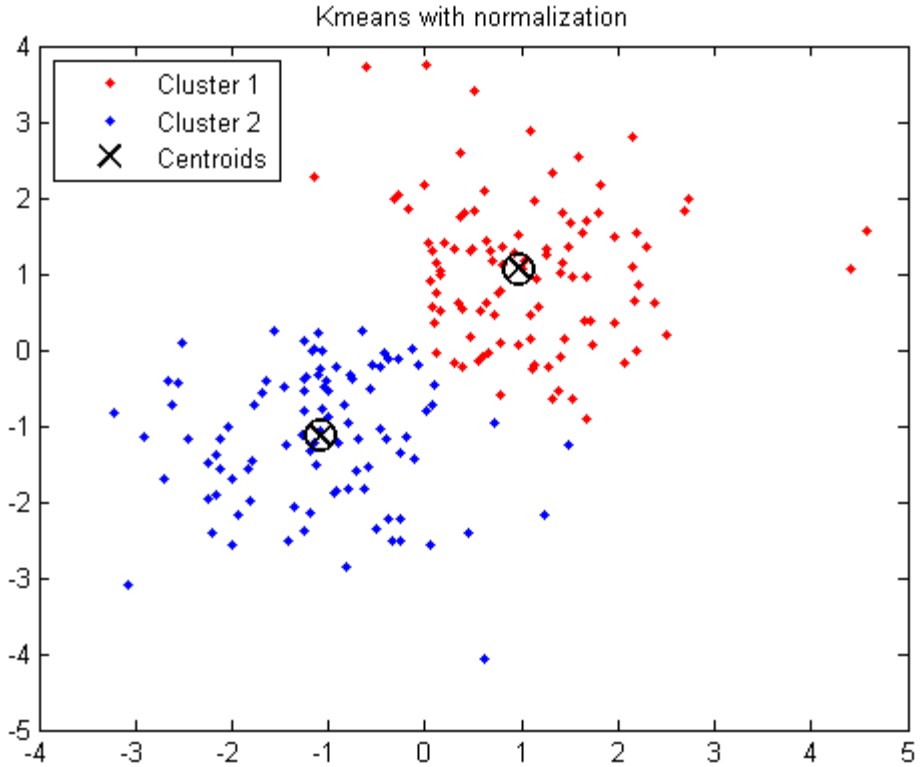


Fig. 2

Name: Amirtha Ganesh Pugazhendhi

4) Effect of Initial Cluster Centroid randomization on K-Means clustering

The main limitation of k-means is that it rarely succeeds in optimizing the centroid locations globally .The reason is that the centroids cannot move between the clusters if their distance is big, or if there are other stable clusters in between preventing the movements, see Fig. 3. The k-means result therefore depends a lot on the initialization. Poor initialization can cause the iterations to get stuck into an inferior local minimum. The Random initialization ensures that data is uniformly spread and preventing the need to move larger distance to find the global optimal.

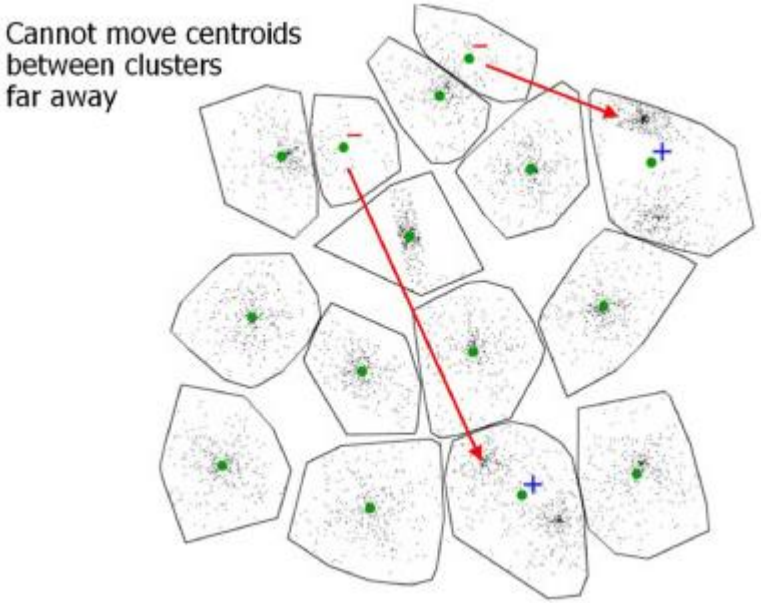
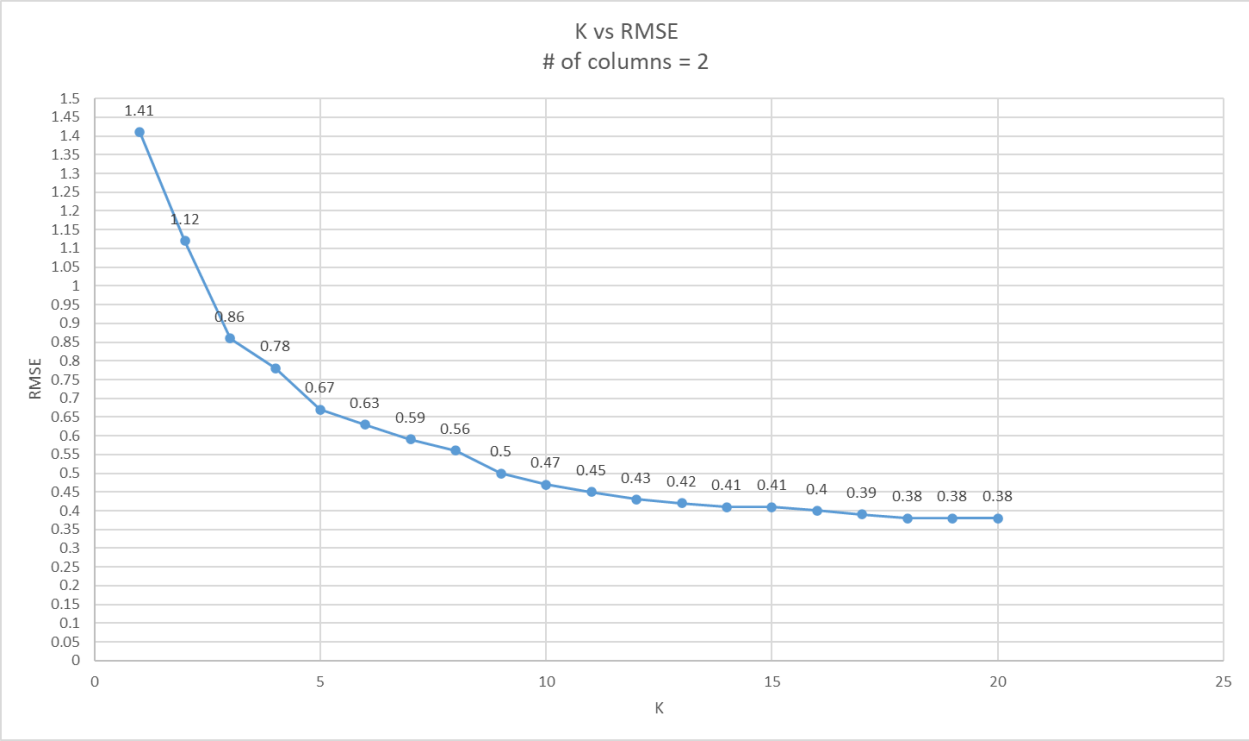


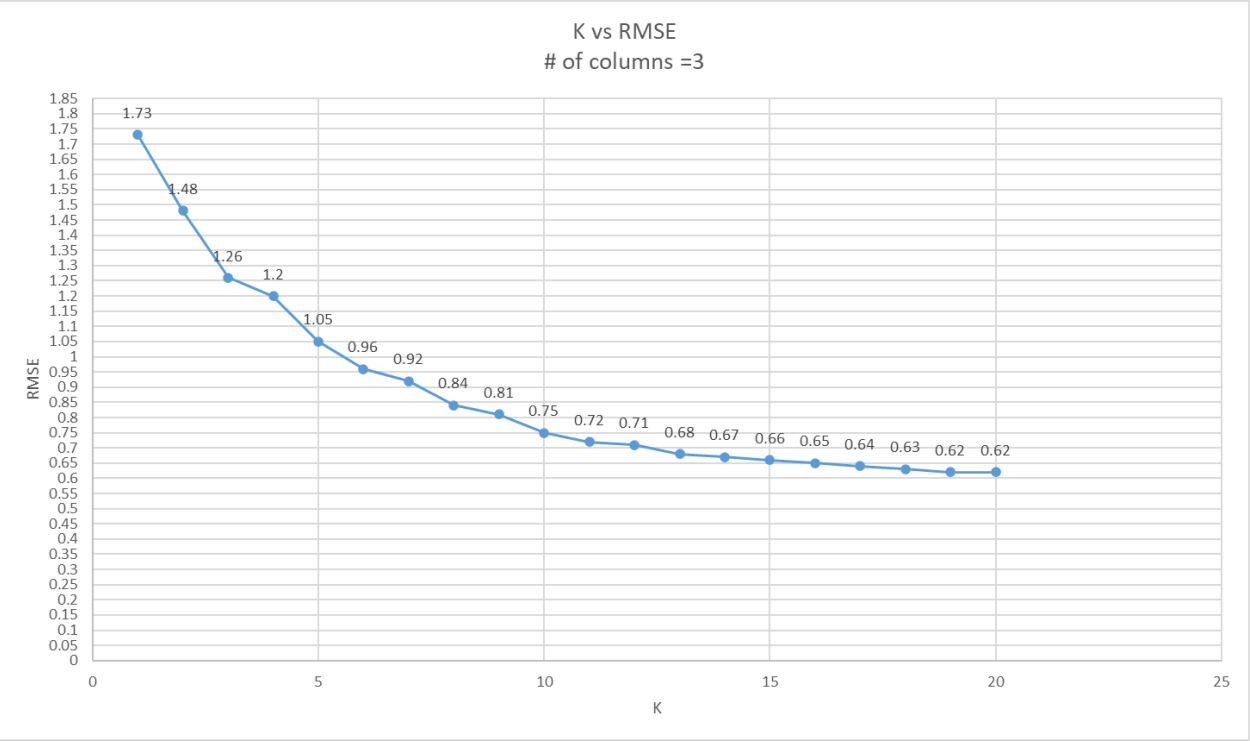
Fig. 3

Reference: “How much can k-means be improved by using better initialization and repeats?” , Pasi Fränti, Sami Sieranoja ,Elsevier, Pattern Recognition 93 (2019) 95–112

Name: Amirtha Ganesh Pugazhendhi

5) Summary and Inference of Results:





1 Better trade-off values of k lie between **10** and **15. k=12** seems to be most important as drop in RSME is not so significant beyond the 12 .

2 The algorithm produced meaningful clusters for these two datasets because both the case the

- Cluster centroids were randomly initialized , this ensure that there is no need to move larger distance to find the global optimal.
- Since the data points were normalized, it reduces the dominance of one particular column. This ensures better quality of clustering.