K-Means Clustering In Excel

Amal Thomas, *CB.EN.P2DSC21001*, Anand Vinod, *CB.EN.P2DSC21002*
Ist $year\ MTECH - DataScience$,
Batch(2021-2023), Course: 21DS602(21-22(Odd)),
cb.en.p2dsc21001@cb.students.amrita.edu
cb.en.p2dsc21002@cb.students.amrita.edu

Abstract

Data clustering is a technique for grouping data into distinct groups based on their similarities. This grouping gives the data a sense of order, making further processing of the data easier. The K-Means clustering algorithm is used in this paper to explain the clustering process on 2 datasets.The datasets taken into consideration are the IRIS dataset as well as the Mall customer dataset.The main highlight of this paper is that it performs k-means clustering algorithm on the dataset in Microsoft Excel itself. No macros were taken into consideration for this project.The paper also emphasizes on the add-in feature the Excel Solver which helps to find the minimum or maximum value depending on the experiment.For this experiment,the Excel Solver takes the minima of the objective function. This research ensures that Microsoft Excel can be used for implementing machine learning algorithms

Index Terms

Machine learning,K-means Clustering,Microsoft Excel,Excel Solver

I. Introduction

IN the modern realm of computers, billions of data undergo manipulation for the extraction of information.The information obtained can be either labeled or unlabelled.If the dataset is unlabelled,then it becomes a herculean task to manage and perform operations on such data.For such unlabelled datasets,we utilize the potential of the clustering technique which can help to bring an insight of information based on the data taken. Clustering is a process/technique which is performed by dividing a data set into different clusters/groups based on similarities.Unlabelled dataset is used to perform such clustering algorithms. This method offers a plethora of applications in every aspect of life.It provides a convenient way to discover the categories of the groups without the need for any training. Clustering is heavily regarded as one of the most significant unsupervised learning issues.Data mining is one of the most important applications of clustering

Data mining helps to find enormous amounts of data in order to find crucial correlations, patterns, and trends utilizing pattern recognition tools as well as statistical and mathematical methodologies by scouring huge amounts of data. One of the initial phases in data mining analysis is clustering. Here we have explained how to implement a simple clustering algorithm, the K-Means in Excel without using Macros and just by using Solver.

The main aim is to implement the k-means clustering technique in Microsoft Excel and find the optimum number of clusters with the help of the Excel add-in Solver package for 2 different datasets.

Dr.Sowmya V,
Assistant professor (Sr. Grade),
Computational Engineering and Networking,Amrita School of Engineering,
Amrita Vishwa Vidhyapeetham, Coimbatore, India.

II. Literature Review

There are many researches conducted based on clustering and other machine learning algorithms.But there are only few projects or researches which consist of machine learning algorithms being implemented in Microsoft Excel[1]-[2]-[3]. Microsoft Excel has come into fruition in the last decade due its potential in handling,manipulating,organizing data etc as well as the availability of add-ins[2]-[3].The Excel add-ins interact with Excel objects, read and write Excel data.It helps the user to have a richer interaction within the add-ins using a dialog window. Researches have been implemented in excel for image segmentation,customer segmentation[4] as well as for biomedical analysis[3]-[5].However,now there are innumerable amounts of research being conducted in Excel due to its ability to analyze thousands of rows containing text data in a scalable,consistent manner.

Authors[1]Ragsdale, C. T. and Zobel, C. W. presents an easy method for constructing as well as training artificial neural networks using Microsoft Excel by depending on both the front end as well as the back-end features of spreadsheets in Excel.

Author[2] Tang describes how to use the XLMiner add-in to perform data mining in Microsoft Excel. Data preparation in Excel may be done quickly and easily using this method.describes an approach of data mining with Excel using the XLMiner add-in. Using this approach, data preparation can be easily accomplished in Excel.

Author[3] Soman,Aravind and Rajgopal also developed an approach on clustering using K-means by using an add-in from Excel called “loadImageArray” inorder to obtain the values referencing to RGB from respective images.After taking in the values,k-means clustering is performed for R,G,and B respectively.

Author [4] Asith Ishanta implemented clustering of customer segmentation data using a python code which enabled in learning market basket analysis.This in turn allowed marketing staffs of companies to devise a strategy for introducing new products to customers who share their interests.This was achieved in python and not in Excel.

Author [5]Peter Ako Larbi,Daniel Asah Larbi demonstrated how Excel is a powerful tool in the field of biomedical signal and image analysis.The paper sheds light on how optical imaging in single spectra or multiple spectra can be performed using various biomedical imaging techniques.

.

III. Objective

This research paper maintains a critical focus on the technique/method used known as the k-means clustering and its implementation in Microsoft Excel along with an additional package in Excel known as the Excel Solver.The K-means clustering is performed on 2 different datasets such as the IRIS dataset and the Mall Customer dataset.The respective optimum k value is to be determined with using Elbow method and the cluster graphs are to be plotted.

IV. Theoretical Background

A. Clustering

Clustering is a process/technique which is performed by dividing a data set into different clusters/groups based on similarities.It is the same as keeping like objects together and separating different objects. So,data points are allocated to their respective cluster centers based on certain conditions or the type of clustering that is used. Here,K-means clustering technique which comes under partitioning clusters is implemented. Partitioning clustering is one of the most popular choices for analysts to create clusters. In this, the clusters are partitioned based upon the characteristics of the data points. We need to specify the number of clusters to be created for this clustering method. These clustering algorithms follow an iterative process to reassign the data points between clusters based upon the distance.

B. K-means algorithm

It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters. The k-means clustering algorithm mainly performs two tasks: Determines the best value for K center points or centroids by an iterative process. Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster. The working of the K-Means algorithm is explained in the below steps:

Step-1: Select the number K to decide the number of clusters.

Step-2: Select random K points or centroids. (It can be different from the input dataset).

Step-3: Assign each data point to their closest centroid, which will form the predefined K clusters.

Step-4: Calculate the variance and place a new centroid of each cluster.Hence each cluster has data points with some commonalities, and it is away from other clusters/the respective data points within the cluster resides to the center of the cluster.

K means algorithm can be applied to any form of data – as long as the data has numerical (continuous) entities and is much faster than other algorithms.This algorithm fails for non-linear data.It is based on the user to decide on the number of clusters before the start of the algorithm.

C. Euclidean Distance:

Euclidean distance is a straight line distance between two points. it is not preferred as a distance measure in high dimensional data.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2} \tag{1}$$

D. K-means loss function:

The K-means clustering process is basically an optimization process whose goal is to minimize the sum of squared error (SSE) or in other terms the loss function.

$$SSE = \sum_{i=1}^{K} \sum_{x\varepsilon c_i} ||m_i - x||^2 \tag{2}$$

where m$_i isthecenterofthei-thclusterwithdataitems, isdefinedby:$

$$m_i = \frac{1}{n_i} \sum_{x\varepsilon C_i} X \tag{3}$$

E. Elbow Method

For K-means clustering, the ”elbow” method is used for finding the optimal number of clusters used (k-value).The graph is plotted with the number of clusters(K) in the X-axis and and the Cluster sum of square(WCSS) in the Y-axis.The ”elbow” (the point of inflection on the curve) is the best value of k if the line chart resembles an arm.

F. Microsoft Excel

Microsoft Excel is a powerful and easy-to-use learning tool. Excel shines in numerical calculations, data organization,comparison,data visualization etc. It is possible to implement machine learning algorithms using Excel, which allows the user to obtain a better understanding of the algorithm's operation. . The Excel package has a significant number of capabilities, the majority of which are available as add-in packages within Excel itself.Most of the packages will be present/installed while installing excel.Here,for this experiment we have used the Solver add-in package.

G. Solver

Solver is a Microsoft Excel add-in program that is used to find an optimal (maximum or minimum) value for a formula in one cell — called the objective cell — subject to constraints, or limits, on the values of other formula cells on a worksheet. Solver works with a group of cells, called decision variables or simply variable cells that are used in computing the formulas in the objective and constraint cells. Solver adjusts the values in the decision variable cells to satisfy the limits on constraint cells and produce the result you want for the objective cell.

In general,the Excel Solver is a sophisticated optimization program that enables users to find the solutions to complex problems which require very high-level mathematical analysis.

V. METHODOLOGY

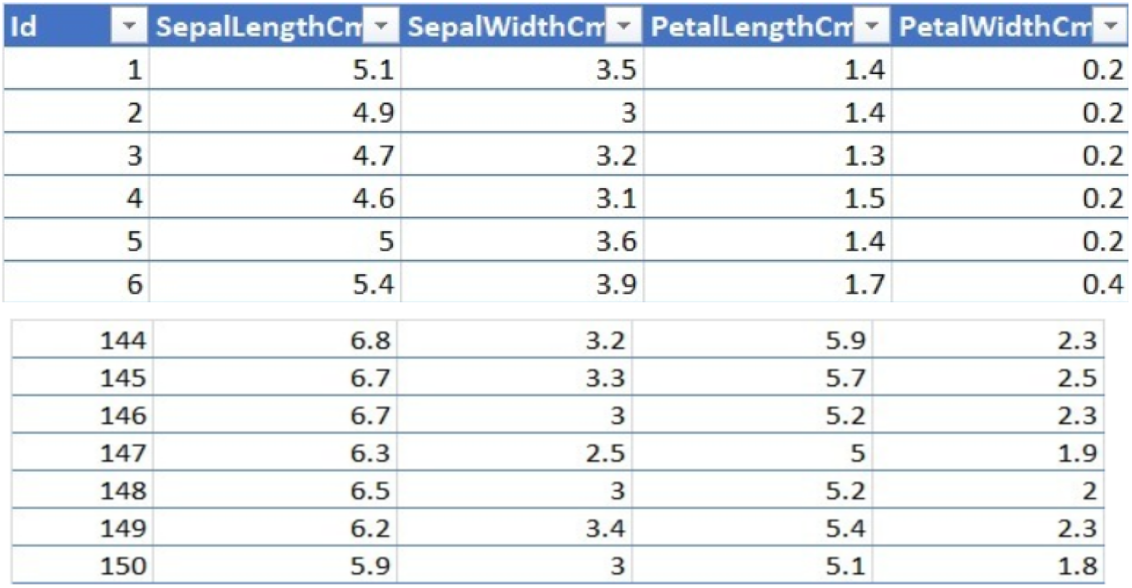
A. Dataset

The dataset which is taken for performing any ml algorithm whether supervised or unsupervised as rows and columns. The rows represent the no of instances/data points ,the columns of the dataset represent the features of the dataset. 2 datasets are taken into consideration for performing the K-means algorithm in excel.

.

1) Iris Dataset: The dataset consists of data of 3 plant species with the following data:

1.Sepal length
2.Sepal width
3.Petal length
4.Petal width



Id	SepalLengthCn	SepalWidthCn	PetalLengthCn	PetalWidthCn
1	5.1	3.5	1.4	0.2
2	4.9	3	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5	3.6	1.4	0.2
6	5.4	3.9	1.7	0.4
144	6.8	3.2	5.9	2.3
145	6.7	3.3	5.7	2.5
146	6.7	3	5.2	2.3
147	6.3	2.5	5	1.9
148	6.5	3	5.2	2
149	6.2	3.4	5.4	2.3
150	5.9	3	5.1	1.8

Fig. 1. Iris Dataset.

2) Mall Customer Dataset: The dataset consists of data of customers in mall with the following data:

1.Customer ID
2.Age
3.Gender
4.Annual Income
5.Spending Score

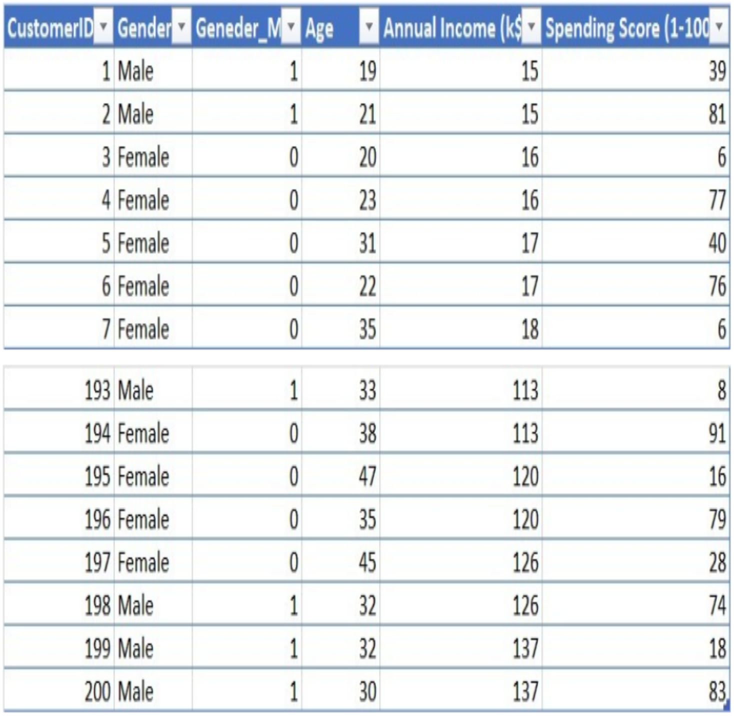
3) Dataset 1: Here we take the iris dataset which has four features.The 4 features are mainly Sepal length ,Sepal width,Petal length and Petal width.K means algorithm starts by assuming a value for K. Lets assume K as 3. Since the value of K is 3, there will be 3 cluster centers. Let the initial centroids be (5,3.6,1.4,0.2) , (5.5,2.4,3.7,1) , (6.3,2.8,5.1,1.5). These points are entered as separate cells in the spreadsheet.

Figure 3 shows the graph consisting of data points and initial centroids for two features Sepal width and Petal length plotted using the XY Plot available in Microsoft Excel.

The same is done and plotted against the remaining features as well. In the K -means algorithm, we need to minimize the distance .Therefore we need to find the euclidean distance between each data point and the cluster centroids. For that we assign 3 columns each for calculating the distances between the cluster centroids and data points. For example, in the spreadsheet the column Dist1 consists of the distances between the data points and the cluster centroid 1.

For finding the distance between the cluster centroid 1 and the first data point we type the euclidean distance formula :
SQRT((B2-$D$154)∧2 $+(C2-\$E\$154)\wedge 2+(D2-\$F\$154)\wedge 2+(E2-\$G\$154)\wedge 2)$

The second cell always remains a constant as it is the cluster centroid in the formula for aall rows in Dist 1.Distance is calculated in the first row of the Dist 1 column in the spreadsheet. Similarly we type the corresponding formulas for distances between the cluster centroid 2 and cluster centroid 3 with the first data point in the first rows of Dist 2 and Dist 3 columns respectively:
=SQRT((B2-$D$155)∧2 $+(C2-\$E\$155)\wedge 2+(D2-\$F\$155)\wedge 2+(E2-\$G\$155)\wedge 2)$
$=SQRT((B2-\$D\$156)\wedge 2+(C2-\$E\$156)\wedge 2+(D2-\$F\$156)\wedge 2+(E2-\$G\$156)\wedge 2).$



CustomerID	Gender	Geneder_M	Age	Annual Income (k$	Spending Score (1-100
1	Male	1	19	15	39
2	Male	1	21	15	81
3	Female	0	20	16	6
4	Female	0	23	16	77
5	Female	0	31	17	40
6	Female	0	22	17	76
7	Female	0	35	18	6
193	Male	1	33	113	8
194	Female	0	38	113	91
195	Female	0	47	120	16
196	Female	0	35	120	79
197	Female	0	45	126	28
198	Male	1	32	126	74
199	Male	1	32	137	18
200	Male	1	30	137	83

Fig. 2. Mall Customer Dataset.

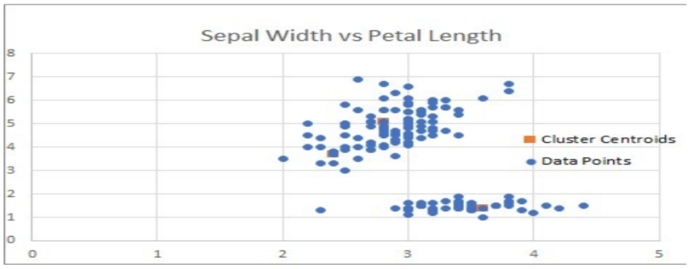
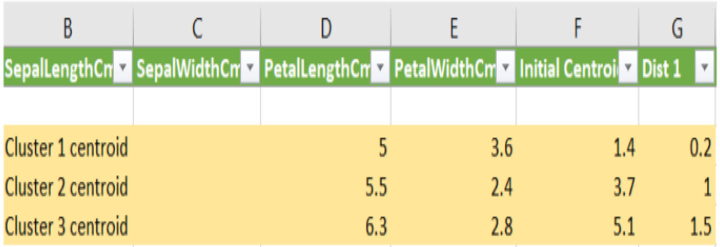


Fig. 3. Sepal width vs Petal length

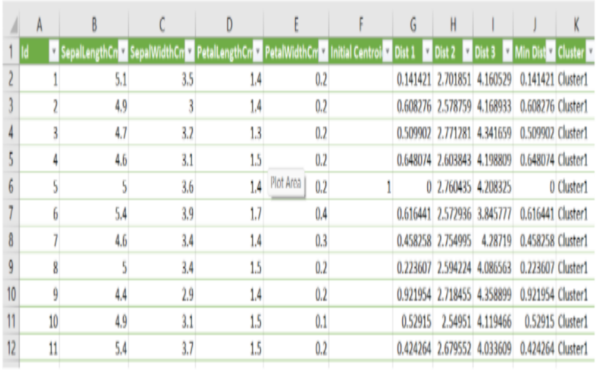


B	C	D	E	F	G
SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Initial Centroi	Dist 1
Cluster 1 centroid		5	3.6	1.4	0.2
Cluster 2 centroid		5.5	2.4	3.7	1
Cluster 3 centroid		6.3	2.8	5.1	1.5

Fig. 4. Cluster centroid of dataset1

In the next column,we find the minimum of these distances for that in case for the first row we use the formula
=MIN(G2:I2)
In the next column we assign the class in which the respective datapoint belongs. In case for our first data point we use the formula:
=IF(MIN(G2:I2)=G2,”Cluster1”,IF(MIN(G2:I2)=H2,”Cluster2”,”Cluster 3”)),

This formula assigns the corresponding clusters for the minimum distance. For example if the minimum distance comes from the Dist 1 column, then the data point gets assigned to Cluster 1.If not ,then it will be assigned to Cluster 2 else Cluster 3.Then we use Excel’s drag and drop option to get the corresponding values for the other data points which is represented in Fig.5..



	A	B	C	D	E	F	G	H	I	J	K
1	Id	SepalLengthCn	SepalWidthCn	PetalLengthCn	PetalWidthCn	Initial Centroi	Dist 1	Dist 2	Dist 3	Min Dist	Cluster
2	1	5.1	3.5	1.4	0.2		0.141421	2.701851	4.160529	0.141421	Cluster1
3	2	4.9	3	1.4	0.2		0.608276	2.578759	4.168933	0.608276	Cluster1
4	3	4.7	3.2	1.3	0.2		0.509902	2.771281	4.341659	0.509902	Cluster1
5	4	4.6	3.1	1.5	0.2		0.648074	2.603843	4.198809	0.648074	Cluster1
6	5	5	3.6	1.4	0.2	1	0	2.760435	4.208325	0	Cluster1
7	6	5.4	3.9	1.7	0.4		0.616441	2.572936	3.845777	0.616441	Cluster1
8	7	4.6	3.4	1.4	0.3		0.458258	2.754995	4.28719	0.458258	Cluster1
9	8	5	3.4	1.5	0.2		0.223607	2.594224	4.086563	0.223607	Cluster1
10	9	4.4	2.9	1.4	0.2		0.921954	2.718455	4.358899	0.921954	Cluster1
11	10	4.9	3.1	1.5	0.1		0.52915	2.54951	4.119466	0.52915	Cluster1
12	11	5.4	3.7	1.5	0.2		0.424264	2.679552	4.033609	0.424264	Cluster1

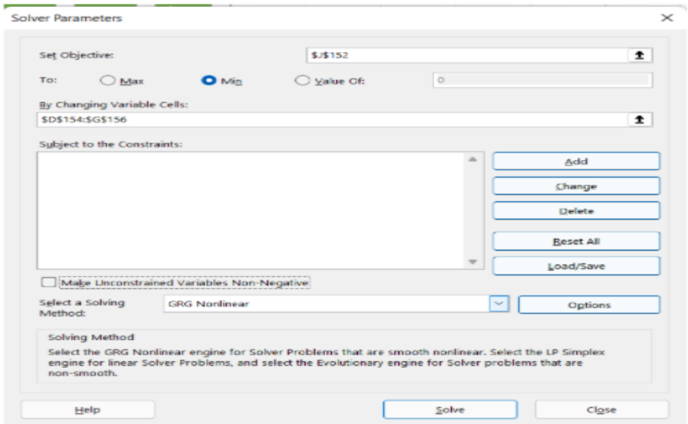
Fig. 5. Dataset 1 consising of minimum dist and cluster

The next step is to find out the sum of the values of the Min Dist column by using Excel’s AutoSum function In K means clustering, we need to minimize the sum of minimum distances. For that we use the Solver add-in feature of Excel. Solver is used to find an optimal (maximum or minimum) value for a formula in one cell subject to constraints, or limits, on the values of other formula cells on a worksheet.

The figure 6 shows the solver dialogue box. In the figure the box corresponding to the label ‘Set Objective’ is assigned for the cell address whose value is to be minimized. In this case it is $J$152 (objective function). The box corresponding to the label ‘By Changing Variable Cells’ is assigned for the cell address whose values must be changed to minimize the objective function. Here in this case it is $D$154:$G$156.

By clicking the Solve button the Solver gives the optimal centroid values. This is performed with the help of iterations.

The first five iteration values of the cluster centroids are as follows in Fig 7:
Above steps are followed for K values ranging from 2 to 8 . A graph is plotted between K values and the corresponding sum of minimum distances in Fig . An elbow curve is formed in the graph and it is seen that optimum K value comes at the elbow/knee of the graph.Fig 17 shows the elbow method for the dataset.



Solver Parameters

Set Objective: $J$152

To: ○ Max ● Min ○ Value Of: 0

By Changing Variable Cells: $D$154:$G$156

Subject to the Constraints:

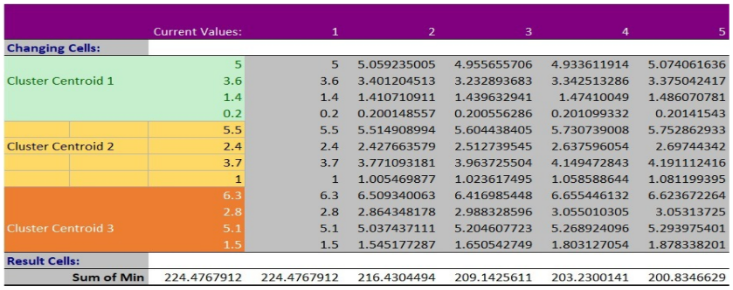
□ Make Unconstrained Variables Non-Negative

Select a Solving Method: GRG Nonlinear

Solving Method

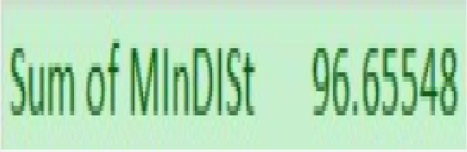
Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

Fig. 6. Solver



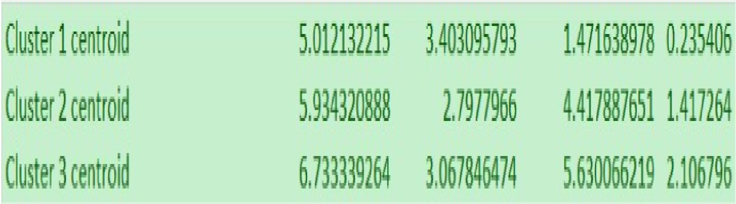
	Current Values:	1	2	3	4	5
Changing Cells:						
	5	5	5.059235005	4.955655706	4.933611914	5.074061636
Cluster Centroid 1	3.6	3.6	3.401204513	3.232893683	3.342513286	3.375042417
	1.4	1.4	1.410710911	1.439632941	1.47410049	1.486070781
	0.2	0.2	0.200148557	0.200556286	0.201099332	0.20141543
	5.5	5.5	5.514908994	5.604438405	5.730739008	5.752862933
Cluster Centroid 2	2.4	2.4	2.427663579	2.512739545	2.637596054	2.69744342
	3.7	3.7	3.771093181	3.963725504	4.149472843	4.191112416
	1	1	1.005469877	1.023617495	1.058588644	1.081199395
	6.3	6.3	6.509340063	6.416985448	6.655446132	6.623672264
	2.8	2.8	2.864348178	2.988328596	3.055010305	3.05313725
Cluster Centroid 3	5.1	5.1	5.037437111	5.204607723	5.268924096	5.293975401
	1.5	1.5	1.545177287	1.650542749	1.803127054	1.878338201
Result Cells:						
Sum of Min	224.4767912	224.4767912	216.4304494	209.1425611	203.2300141	200.8346629

Fig. 7. Iteration of dataset1



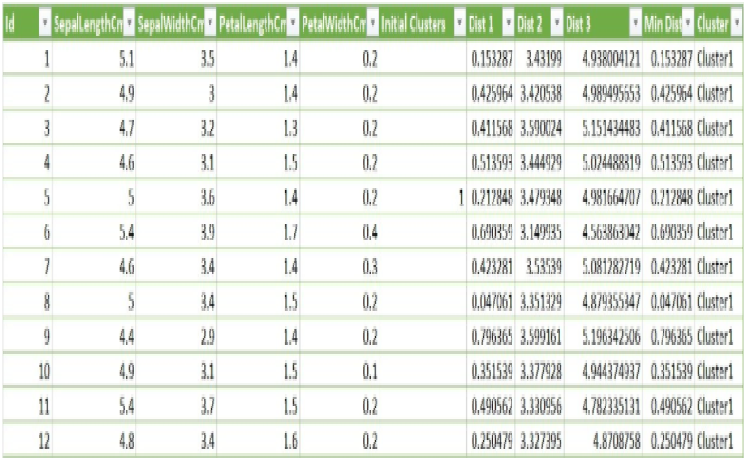
Sum of MInDISt	96.65548

Fig. 8. Min Dist



Cluster 1 centroid	5.012132215	3.403095793	1.471638978	0.235406
Cluster 2 centroid	5.934320888	2.7977966	4.417887651	1.417264
Cluster 3 centroid	6.733339264	3.067846474	5.630066219	2.106796

Fig. 9. Updated centroids for dataset 1

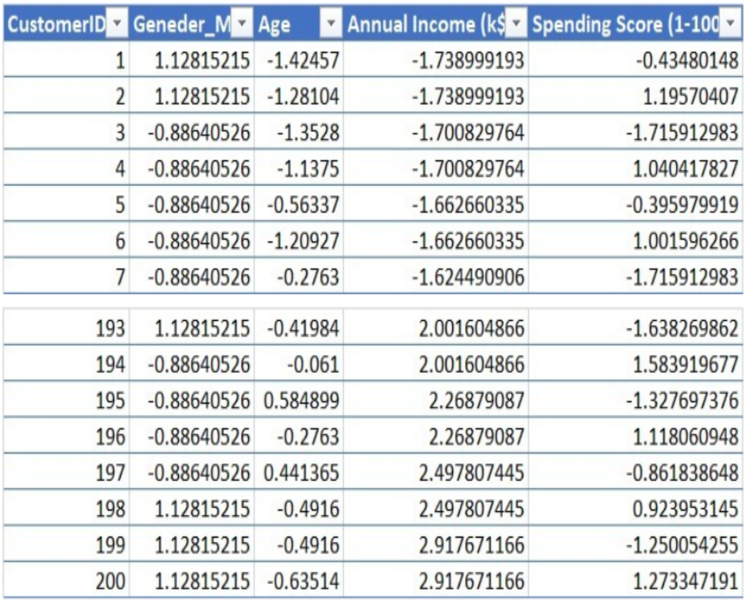


Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Initial Clusters	Dist 1	Dist 2	Dist 3	Min Dist	Cluster
1	5.1	3.5	1.4	0.2		0.153287	3.43199	4.938004121	0.153287	Cluster1
2	4.9	3	1.4	0.2		0.425964	3.420538	4.989495653	0.425964	Cluster1
3	4.7	3.2	1.3	0.2		0.411568	3.590024	5.151434483	0.411568	Cluster1
4	4.6	3.1	1.5	0.2		0.513593	3.444929	5.024488819	0.513593	Cluster1
5	5	3.6	1.4	0.2	1	0.212848	3.479348	4.981664707	0.212848	Cluster1
6	5.4	3.9	1.7	0.4		0.690359	3.149935	4.563863042	0.690359	Cluster1
7	4.6	3.4	1.4	0.3		0.423281	3.53539	5.081282719	0.423281	Cluster1
8	5	3.4	1.5	0.2		0.047061	3.351329	4.879355347	0.047061	Cluster1
9	4.4	2.9	1.4	0.2		0.796365	3.599161	5.196342506	0.796365	Cluster1
10	4.9	3.1	1.5	0.1		0.351539	3.377928	4.944374937	0.351539	Cluster1
11	5.4	3.7	1.5	0.2		0.490562	3.330956	4.782335131	0.490562	Cluster1
12	4.8	3.4	1.6	0.2		0.250479	3.327395	4.8708758	0.250479	Cluster1

Fig. 10. Dist of dataset1

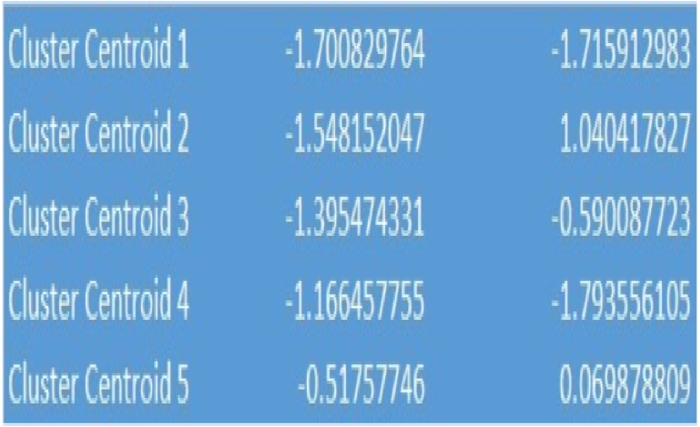
B. For Dataset 2

Here we take the Mall Customer Dataset ,having 2 features within the dataset.The features are annual income and spending score respectively.But the values in the column for both features have different ranges,so we standardize the dataset within a specific range with mean taken as 0 and the standard deviation as 1. Fig 11 shows the standardized data of the Mall customer dataset.



CustomerID	Geneder_M	Age	Annual Income (k$	Spending Score (1-100
1	1.12815215	-1.42457	-1.738999193	-0.43480148
2	1.12815215	-1.28104	-1.738999193	1.19570407
3	-0.88640526	-1.3528	-1.700829764	-1.715912983
4	-0.88640526	-1.1375	-1.700829764	1.040417827
5	-0.88640526	-0.56337	-1.662660335	-0.395979919
6	-0.88640526	-1.20927	-1.662660335	1.001596266
7	-0.88640526	-0.2763	-1.624490906	-1.715912983
193	1.12815215	-0.41984	2.001604866	-1.638269862
194	-0.88640526	-0.061	2.001604866	1.583919677
195	-0.88640526	0.584899	2.26879087	-1.327697376
196	-0.88640526	-0.2763	2.26879087	1.118060948
197	-0.88640526	0.441365	2.497807445	-0.861838648
198	1.12815215	-0.4916	2.497807445	0.923953145
199	1.12815215	-0.4916	2.917671166	-1.250054255
200	1.12815215	-0.63514	2.917671166	1.273347191

Fig. 11. Standardized Mall Customer Dataset



Cluster Centroid 1	-1.700829764	-1.715912983
Cluster Centroid 2	-1.548152047	1.040417827
Cluster Centroid 3	-1.395474331	-0.590087723
Cluster Centroid 4	-1.166457755	-1.793556105
Cluster Centroid 5	-0.51757746	0.069878809

Fig. 12. Initial Cluster Centroids for Dataset 2

We assume the value of K as 5. Initial cluster centroids are randomly selected.

Figure 13 shows the plot between features consisting of the data points and the initial centroids.

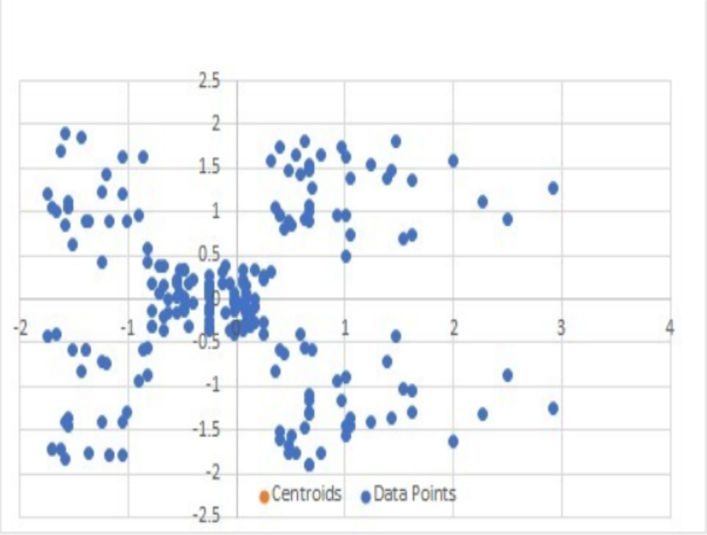
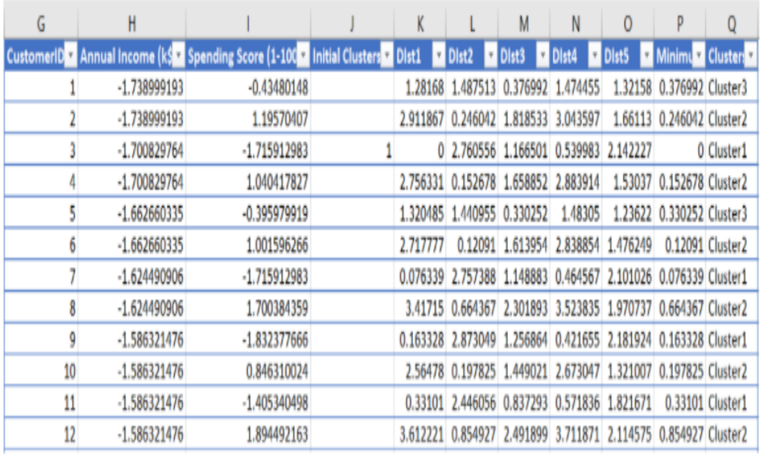


Fig. 13. Plot of data points with initial centroids

We then find the euclidean distance between each data point and the centroids in the same way we did for the IRIS dataset. For example in case for first row of the Dist 1 column we find the distance between the first data point and the first centroid by using the euclidean formula i.e.
=SQRT((H2-$B$3)∧2 + (I2 − $C$3) ∧ 2).
Then, we take the minimum distance among them by using the formula
=MIN(K2:O2)

and corresponding to that minimum distance we allocate/assign the data point to that particular cluster .

Figure 14 shows computed minimum distances and the assigned clusters

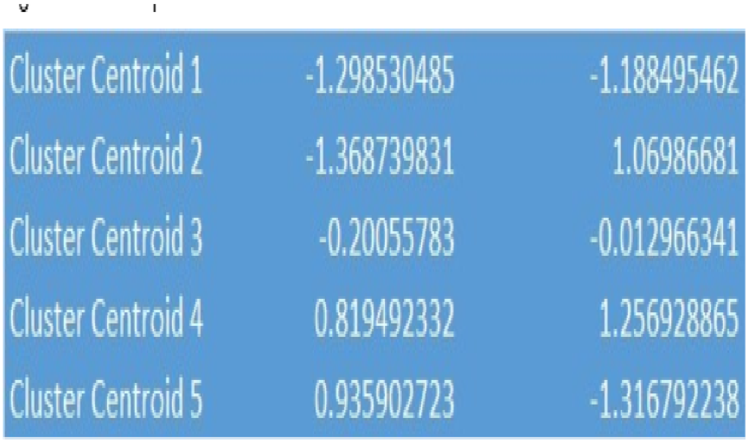


G	H	I	J	K	L	M	N	O	P	Q
CustomerID	Annual Income (k$	Spending Score (1-100	Initial Clusters	Dist1	Dist2	Dist3	Dist4	Dist5	Minimu	Cluster
1	-1.738999193	-0.43480148		1.28168	1.487513	0.376992	1.474455	1.32158	0.376992	Cluster3
2	-1.738999193	1.19570407		2.911867	0.246042	1.818533	3.043597	1.66113	0.246042	Cluster2
3	-1.700829764	-1.715912983	1	0	2.760556	1.166501	0.539983	2.142227	0	Cluster1
4	-1.700829764	1.040417827		2.756331	0.152678	1.658852	2.883914	1.53037	0.152678	Cluster2
5	-1.662660335	-0.395979919		1.320485	1.440955	0.330252	1.48305	1.23622	0.330252	Cluster3
6	-1.662660335	1.001596266		2.717777	0.12091	1.613954	2.838854	1.476249	0.12091	Cluster2
7	-1.624490906	-1.715912983		0.076339	2.757388	1.148883	0.464567	2.101026	0.076339	Cluster1
8	-1.624490906	1.700384359		3.41715	0.664367	2.301893	3.523835	1.970737	0.664367	Cluster2
9	-1.586321476	-1.832377666		0.163328	2.873049	1.256864	0.421655	2.181924	0.163328	Cluster1
10	-1.586321476	0.846310024		2.56478	0.197825	1.449021	2.673047	1.321007	0.197825	Cluster2
11	-1.586321476	-1.405340498		0.33101	2.446056	0.837293	0.571836	1.821671	0.33101	Cluster1
12	-1.586321476	1.894492163		3.612221	0.854927	2.491899	3.711871	2.114575	0.854927	Cluster2

Fig. 14. Dataset 2 consisting of Min dist and cluster

In the same way we did for the iris dataset, we will minimize the value of sum of minimum distance by using the solver add in. The Solver will provide the optimum values for the cluster centroids and correspondingly the data points are assigned to their respective clusters. Similarly,the same has been done for 8 k values and the corresponding cluster centroids were updated.

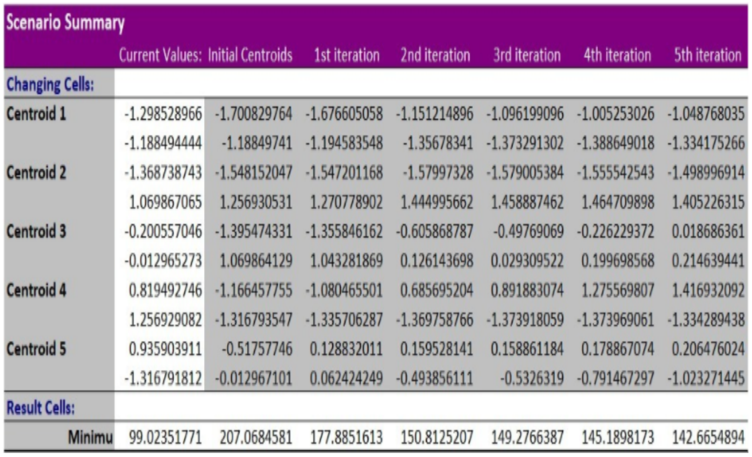
Figure 15 shows the updated Cluster Centroids.



Cluster Centroid 1	-1.298530485	-1.188495462
Cluster Centroid 2	-1.368739831	1.06986681
Cluster Centroid 3	-0.20055783	-0.012966341
Cluster Centroid 4	0.819492332	1.256928865
Cluster Centroid 5	0.935902723	-1.316792238

Fig. 15. Updated Centroid cluster Dataset 2

This is done using iterations within the solver.Fig 16 shows the iteration of the 5 cluster centroids.A graph is plotted between K values and the corresponding sum of minimum distances. An elbow curve is formed in the graph and it is seen that optimum K value comes at the elbow/knee of the graph. Fig 24 shows the elbow method of the dataset.



Scenario Summary

	Current Values:	Initial Centroids	1st iteration	2nd iteration	3rd iteration	4th iteration	5th iteration
Changing Cells:							
Centroid 1	-1.298528966	-1.700829764	-1.676605058	-1.151214896	-1.096199096	-1.005253026	-1.048768035
	-1.188494444	-1.18849741	-1.194583548	-1.35678341	-1.373291302	-1.388649018	-1.334175266
Centroid 2	-1.368738743	-1.548152047	-1.547201168	-1.57997328	-1.579005384	-1.555542543	-1.498996914
	1.069867065	1.256930531	1.270778902	1.444995662	1.458887462	1.464709898	1.405226315
Centroid 3	-0.200557046	-1.395474331	-1.355846162	-0.605868787	-0.49769069	-0.226229372	0.018686361
	-0.012965273	1.069864129	1.043281869	0.126143698	0.029309522	0.199698568	0.214639441
Centroid 4	0.819492746	-1.166457755	-1.080465501	0.685695204	0.891883074	1.275569807	1.416932092
	1.256929082	-1.316793547	-1.335706287	-1.369758766	-1.373918059	-1.373969061	-1.334289438
Centroid 5	0.935903911	-0.51757746	0.128832011	0.159528141	0.158861184	0.178867074	0.206476024
	-1.316791812	-0.012967101	0.062424249	-0.493856111	-0.5326319	-0.791467297	-1.023271445
Result Cells:							
Minimu	99.02351771	207.0684581	177.8851613	150.8125207	149.2766387	145.1898173	142.6654894

Fig. 16. Updated Dataset 2

VI. Result

The optimum value “k” is obtained as shown below with the help of elbow method graphs and the graphs consisting of the data points along with the cluster centroids have been plotted for the 2 datasets respectively.

The plotted graphs for the dataset 1 is as follows:

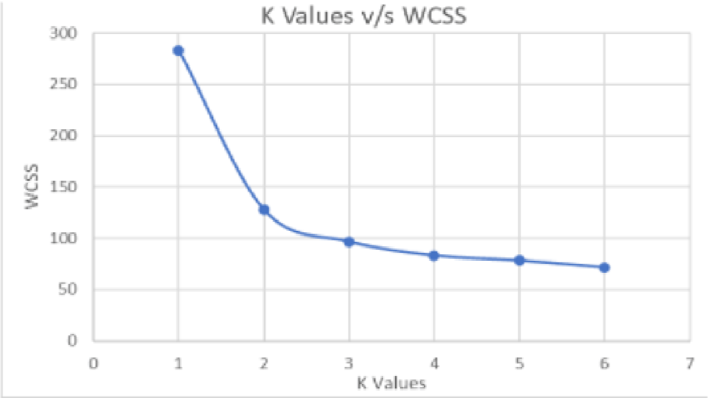


Fig. 17. Elbow curve for Dataset 1

It is clear from the elbow point that the optimum k value is 3. Since the dataset 1 has four features we plot the graph between two features.
The graphs are shown below:

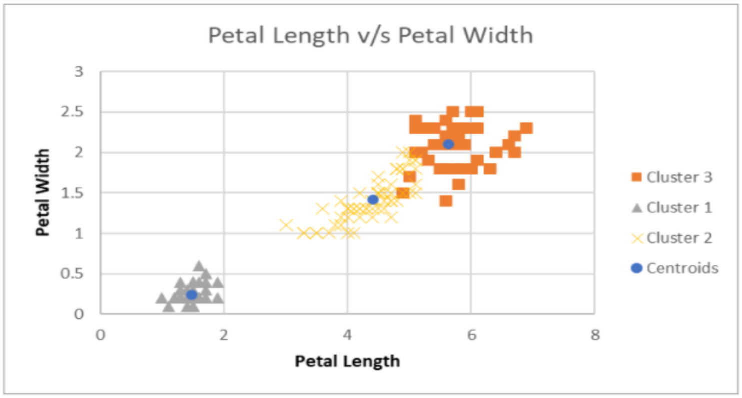


Fig. 18. Petal length vs Petal Width

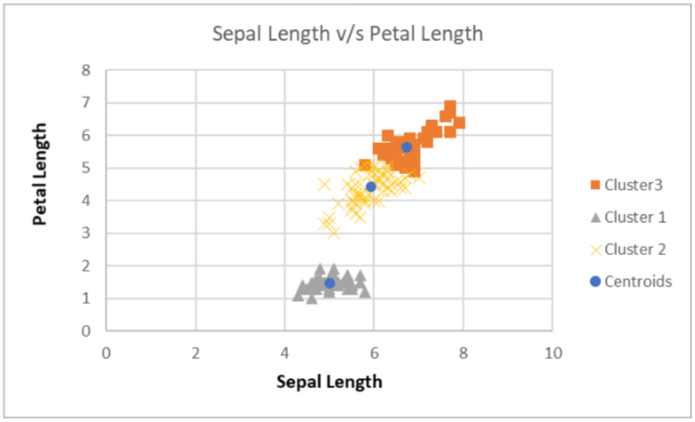


Fig. 19. Sepal Length vs Petal Length

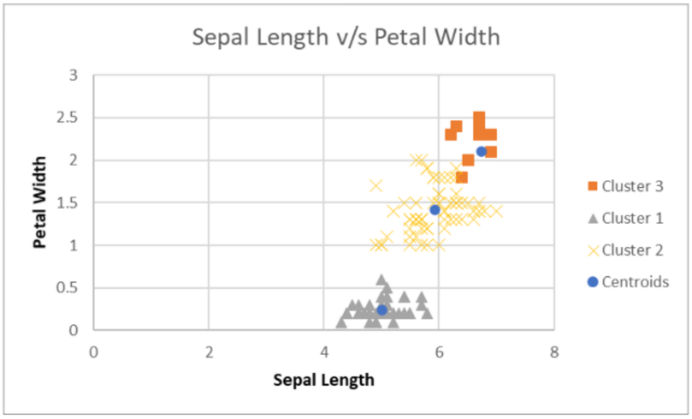


Fig. 20. Sepal Length vs Petal Width

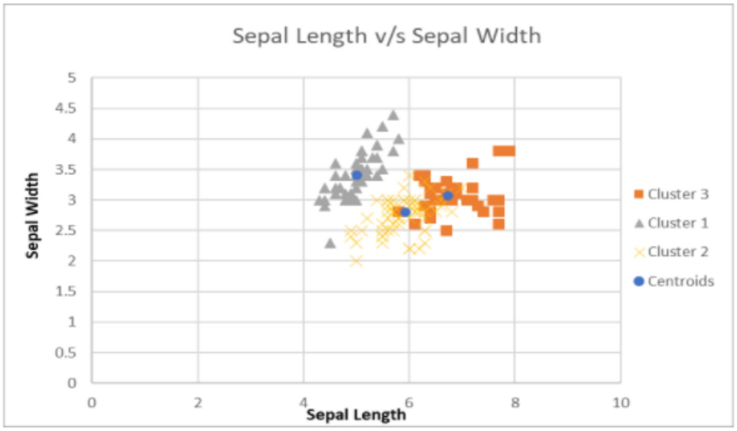


Fig. 21. Sepal Length vs Sepal Width

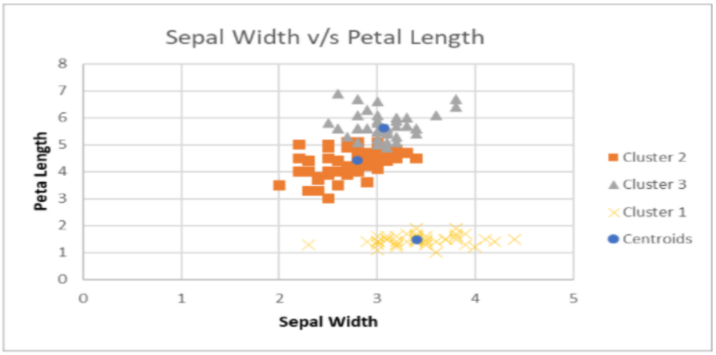


Fig. 22. Sepal Width vs Petal Length

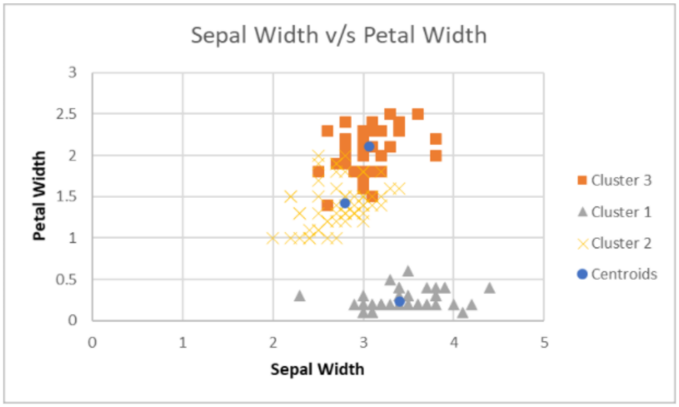


Fig. 23. Sepal Width vs Petal Width

For dataset 2: Mall customer dataset Here the elbow or the knee point comes at k=5 i.e; 5 clusters. Here,the 2 features are the annual income and spending score

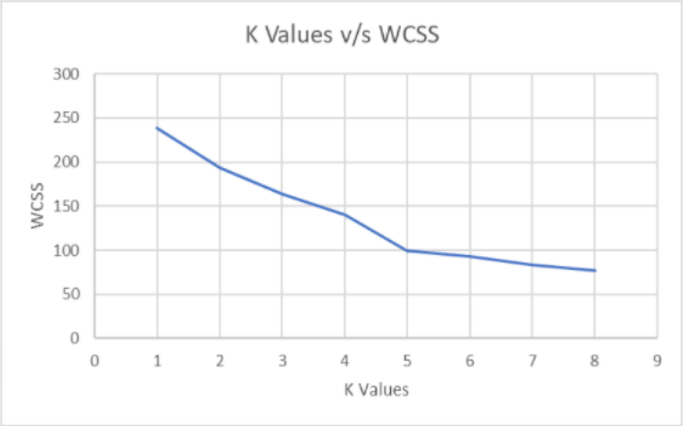


Fig. 24. Elbow Curve for Dataset 2

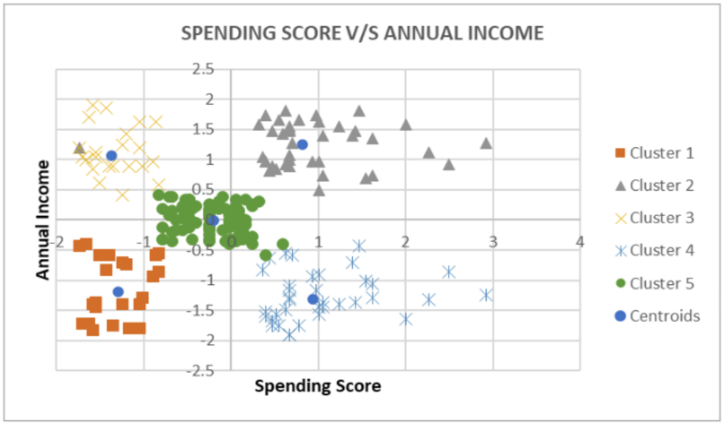


Fig. 25. Graph consisting data points and final centroids

VII. Conclusion

Clustering is very useful for unlabelled data and helps to obtain information it.Here,implementation of K-means clustering has been done for 2 datasets i.e;IRIS and Mall customer dataset respectively .The optimum value was obtained and the cluster centroids along with the respective data points are plotted.Thus, this experiment shows that an unsupervised machine learning algorithm(in this case k-means) can be implemented in Microsoft Excel along with the use of Excel Solver to find the optimum value and such graphs can be plotted.

References

[1] Ragsdale, C. T. and Zobel, C. W. (2010), *”A Simple Approach to Implementing and Training Neural Networks in Excel”*, in Decision Sciences Journal of Innovative Education, 8: 143–149. doi: 10.1111/j.1540-4609.2009.00249.x

[2] Tang, H. (2008), “A Simple Approach of Data Mining in Excel”,IEEE Fourth International Conference Wireless Communications, Networking and Mobile Computing, doi :10.1109/WiCom.2008.2679

[3] Aravind H, Rajgopal C, Soman KP.(2010) ,“A simple approach to clustering in Excel” ,International Journal of Computers and Applications. 7, 2010;11:19-25

[4] Asith Ishanta .(2021),”Mall customer segmentation using clustering algorithm”,Research gate

[5] Peter Ako Larbi,Daniel Asah Larbi.(2018),”Adopting Microsoft Excel for Biomedical Signal and Image Processing”,IntechOpen,DOI:https://dx.doi.org/10.5772/intechopen.81732

.