

# CHAPTER 5

## Implementing $k$ -Means Clustering in Excel

### OBJECTIVE

In this chapter, you will learn how to implement  $k$ -Means clustering in Excel. First, I will guide you through the manual steps of the  $k$ -means procedure using the Excel workbook *knn.xlsx*. This will help you understand how the algorithm works. Then, I will introduce the Excel template *Multivariate.xlsm*, which automates the clustering process for greater efficiency. The *knn.xlsx* workbook provides insight into the inner workings of the  $k$ -means algorithm, while the *Multivariate.xlsm* template streamlines the implementation.

### Contents

5.1	<i>Introduction</i>	81
5.2	<i>Manual <math>k</math>-Means Clustering of the Iris15 Dataset</i>	82
5.2.1	<i>Description of the Iris15 Dataset</i>	83
5.2.2	<i>Initial Cluster Centers</i>	85
5.2.3	<i>The Iterative Process</i>	89
5.2.4	<i>Visual Exploration of Clusters</i>	93
5.3	<i>Cluster Analysis with <i>Multivariate.xlsm</i></i>	97
5.4	<i>Concluding Remarks</i>	109

---

## 5.1 Introduction

---

In this chapter, I will provide a detailed guide on how to implement the  $k$ -means procedure using Excel. Each step of the procedure, as described in Section 4.3 of Chapter 4, will be reviewed and implemented in Excel. We will first perform the process manually in Section 5.2 to give you a clear understanding of how cluster analysis works. Later, in Section 5.3, I will introduce the *Multivariate.xlsm* Excel template, which automates the clustering process and facilitates the analysis of larger datasets more efficiently.

No prior Excel knowledge is required for this chapter, though any previous experience with Excel will certainly be beneficial. Specifically, in the manual implementation of cluster analysis, I aim to create a dynamic worksheet where changes to input data or parameters automatically update relevant cells. To achieve this, I will make extensive use of Excel formulas, explaining their functionality and purpose along the way. If you're not familiar with Excel formulas, this will be a great opportunity to learn.

The *Multivariate.xlsm* template does not require any advanced Excel skills. It is a user-friendly, menu-driven, point-and-click program that simplifies the clustering process. All you need to do is provide the data for analysis and specify the type of analysis you want to perform.

Let's get started with the manual implementation of  $k$ -means clustering.

## 5.2 Manual $k$ -Means Clustering of the Iris15 Dataset

---

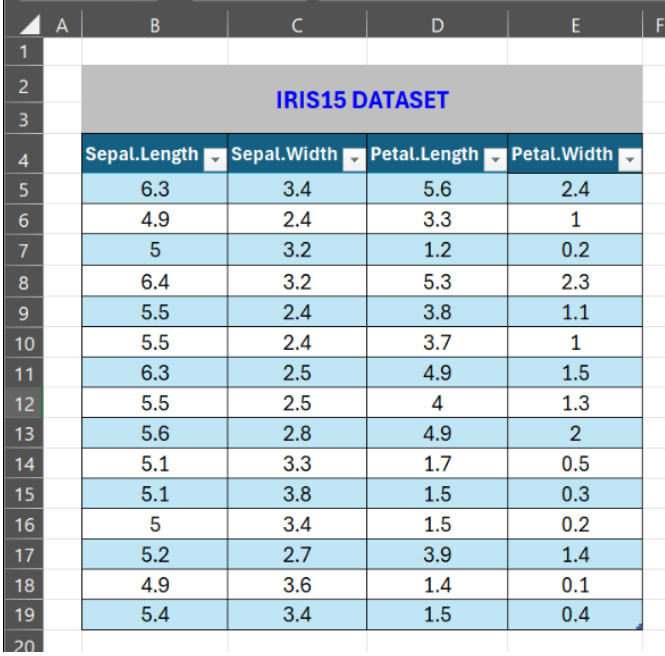
Before proceeding with this section, I recommend downloading the Excel file we'll be working with from the following link: <https://agreestat.com/books/knn/datasets/knn.xlsx>. The file contains a single worksheet organized into several tables, which are divided into the following seven sections.

- **IRIS15 DATASET:** This section includes a table with 15 records from the *Iris15* dataset. This dataset will serve as input for the exercises and examples throughout the section.
- **INITIALIZATION:** This is where the initial cluster centers are calculated. These baseline values will serve as starting points for the  $k$ -Means algorithm, which will refine them through the iterative process.
- **ITERATION#1-ITERATION#5:** Each of the next five sections ITERATION#1 through ITERATION#5 presents two tables corresponding to the a given iteration of the  $k$ -Means algorithm. The first table shows the distances from each data point to the three cluster centers, while the second table displays the updated cluster centers after the completion of the respective iteration (1, 2, 3, 4, or 5).

Each of these topics will be discussed in detail throughout this section, guiding you step-by-step through the workings of the  $k$ -Means algorithm.

### 5.2.1 Description of the Iris15 Dataset

Figure 5.1 shows the *Iris15* dataset, which will be used to demonstrate the  $k$ -Means algorithm. The objective is to classify all 15 cases into 3 clusters based on four numeric variables: *Sepal.Length*, *Sepal.Width*, *Petal.Length*, and *Petal.Width*.



	A	B	C	D	E	F
1						
2		IRIS15 DATASET				
3						
4		Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	
5		6.3	3.4	5.6	2.4	
6		4.9	2.4	3.3	1	
7		5	3.2	1.2	0.2	
8		6.4	3.2	5.3	2.3	
9		5.5	2.4	3.8	1.1	
10		5.5	2.4	3.7	1	
11		6.3	2.5	4.9	1.5	
12		5.5	2.5	4	1.3	
13		5.6	2.8	4.9	2	
14		5.1	3.3	1.7	0.5	
15		5.1	3.8	1.5	0.3	
16		5	3.4	1.5	0.2	
17		5.2	2.7	3.9	1.4	
18		4.9	3.6	1.4	0.1	
19		5.4	3.4	1.5	0.4	
20						

Figure 5.1: Extract of 15 records randomly selected from the Iris dataset

This dataset is structured as an Excel “data table,” which I’ve named *Iris*. Data tables offer more functionality than regular tables in Excel. I prefer using them because they can be easily expanded without affecting the rest of the spreadsheet. For more information on Excel data tables, refer to Sections A.1 and A.2 in Appendix A.

*Finding the Variable with the Highest Variation*

As mentioned earlier, the initial selection of centroids (or cluster centers) is crucial. The three initial cluster centers must be chosen from the input dataset, ideally from data points that are likely to end up in separate clusters. A good initial choice will require fewer iterations of the *k*-means algorithm, while a poor choice may fail to create distinct clusters, potentially leading to unsatisfactory results. The strategy I propose relies on the variable with the highest variation. If your input dataset is made up of principal components, then the variable with the highest variation will always be the first principal component.

Selecting the initial cluster centers is often a subjective process, with no universally accepted method. While some approaches suggest randomly picking three records, I am cautious about relying solely on randomness, as it can lead to suboptimal outcomes or significant errors.

Here's my recommended approach: first, identify the variable with the greatest variation. Then, divide the dataset into three clusters based on the 25th, 50th, and 75th percentiles of this variable<sup>1</sup>. To measure variation, I used Excel's standard deviation function<sup>2</sup>, `STDEV.S()`, which produced the values highlighted in yellow and written in blue in row #22 as shown in Figure 5.2. Based on these standard deviations, *Petal.Length*, with a stan-

---

<sup>1</sup>This method is most effective if the clustering is based on PCA composite scores, as the first composite variable typically exhibits the highest variation.

<sup>2</sup>The standard deviation measures the extent to which any given number from a data series is expected to stray away from its average value.

dard deviation of 1.6, is the variable with the highest variation. Once this variable is identified, you can delete the “Standard Deviation” row, which is not part of the input data table.

IRIS15 DATASET				
	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
5	6.3	3.4	5.6	2.4
6	4.9	2.4	3.3	1
7	5	3.2	1.2	0.2
8	6.4	3.2	5.3	2.3
9	5.5	2.4	3.8	1.1
10	5.5	2.4	3.7	1
11	6.3	2.5	4.9	1.5
12	5.5	2.5	4	1.3
13	5.6	2.8	4.9	2
14	5.1	3.3	1.7	0.5
15	5.1	3.8	1.5	0.3
16	5	3.4	1.5	0.2
17	5.2	2.7	3.9	1.4
18	4.9	3.6	1.4	0.1
19	5.4	3.4	1.5	0.4
Standard Deviation	0.514	0.490	1.600	0.771

Figure 5.2: Standard deviations of the 4 variables in the Iris dataset.

### 5.2.2 Initial Cluster Centers

Figure 5.3 displays two tables. The first of these tables, titled “Pctiles / *Petal.Length*,” lists the three percentiles of the *Petal.Length* variable you want to calculate, along with the corresponding row numbers in the dataset where they are located. The