# ग्रीष्म/औद्योगिक परियोजना प्रशिक्षण प्रतिवेदन

# Summer/Industrial Project Training Report

on

**<u>Japan Earthquake Hotspot Detection Using K-Means Clustering Method</u>**

**1 जुलाई 2022 – 31 अगस्त 2022**

**1st July 2022 – 31st August 2022**

## (दो माह/Two Months)

**द्वारा प्रस्तुत/Submitted by –**



| | | |
|---|---|---|
| **नाम/Name** | : | **Neerja Tiwari** |
| **पंजी क्रमाक/Roll No.** | : | **07814803120** |
| **कॉलेज आईडी/College ID** | : | **07814803120** |
| **पाठ्यक्रम/Course** | : | **B.Tech** |
| **विश्वविद्यालय/University** | : | **Maharaja Agrasen Institute of Technology** |



**पर्यवेक्षक के अधीन/Under the supervision of –**



| | | |
|---|---|---|
| **नाम/Name** | : | **Mr. Anand Prakash** |
| **पद /Designation** | : | **Scientist ‘D’** |
| **कार्यालय/Office** | : | **ISSA** |

**पद्धति अध्ययन एवं विश्लेषण संस्थान/**

**Institute for Systems Studies and Analyses**

**रक्षा अनुसंधान एवं विकास संगठन/**

**Defence R&D Organisation**

**रक्षा मंत्रालय/Ministry of Defence**

**मेटकाफ भवन, दिल्ली / Metcalfe House, Delhi-110054**

## अभ्यर्थी द्वारा घोषणा/

## DECLARATION BY THE CANDIDATE

I hereby declare that the work that is being presented by me in this project/study entitled "**Japan Earthquake Hotspot Detection Using K-Means Clustering Method**" is an authentic record of my own analysis and theoretical research carried out during the period from 1st July 2022 to 31st August 2022 under the supervision of Mr. Anand Prakash, Scientist 'D'. (Institute of Systems Studies and Analyses, Defence R&D Organisation, Ministry of Defence, Metcalfe House, Delhi 110054).

**द्वारा प्रस्तुत/Submitted by –**

| | | |
|---|---|---|
| **नाम/Name** | : | **Neerja Tiwari** |
| **पंजी क्रमाक/Roll No.** | : | **07814803120** |
| **कॉलेज आईडी/College ID** | : | **07814803120** |
| **पाठ्यक्रम/Course** | : | **B.Tech** |
| **विश्वविद्यालय/University** | : | **Maharaja Agrasen Institute of Technology** |
| **दिनांक/Date** | : | **16 August 2022** |

# अभिस्वीकृति/
# ACKNOWLEDGEMENT

I am grateful to Director, ISSA, and Head of HRD for providing me the opportunity to carry out my project at this esteemed organization. I wish to express my deep gratitude to Mr. Anand Prakash, Scientist ‘D’, ISSA, DRDO for providing guidance and support so far in the project work. This internship report might never have been completed without the necessary practical knowledge, assistance from many books, articles, and websites. I am also grateful to all employees who answered my all questions regarding my study with a smiling face. They helped me in such a way that helped me to feel comfortable there and thus I have completed my report properly. I believe that this endeavour has prepared me for taking up new challenging opportunities in the future.

**द्वारा प्रस्तुत/Submitted by –**

**नाम/Name** : **Neerja Tiwari**

**पंजी क्रमाक/Roll No.** : **07814803120**

**कॉलेज आईडी/College ID** : **07814803120**

**पाठ्यक्रम/Course** : **B.Tech**

**विश्वविद्यालय/University** : **Maharaja Agrasen Institute of Technology**

**दिनांक/Date** : **16 August 2022**

# पद्धति अध्ययन एवं विश्लेषण संस्थान के बारे में/

## About the Institute for System Studies and Analyses (ISSA)

Institute for System Studies and Analyses (ISSA) is a premier institution involved in systems analysis of Defence Systems. It provides analysis support to the top echelons of the three services, SA to RM, and DRDO HQs for scientific decision-making. It also provides systems analysis support to sister labs and other institutions under the Ministry of Defence. ISSA is primarily devoted to systems analysis and specializes in modeling and simulation in wide range of applications.

ISSA adopts state-of-the-art info-technologies such as Computer Networking, Software Engineering, Distributed Database, Distributed Simulation, Web Technologies, Situational Awareness and Soft-Computing Techniques in development of complex simulation products.

# रक्षा अनुसंधान एवं विकास संगठन के बारे में/
# About the Defence Research and Development Organization

DRDO formed in 1958 from the amalgamation of the then, already functioning Technical Development Establishment (TDEs) of the Indian Army and the Directorate of Technical Development & Production (DTDP) with the Defence Science Organization (DSO). Today, DRDO is having more than 50 labs, engaged in developing Defence Technologies covering various disciplines like aeronautics, armaments, electronics, combat vehicles, engineering systems, instrumentation, missiles, advanced computing and simulation, special materials, naval systems, life sciences, training, information systems and agriculture. Over 5000 scientists and about 25,000 other scientific, technical and supporting personnel back DRDO.

**Vision**

Make India prosperous by establishing excellent science and technology base and provide our Defence Services decisive edge by equipping them with internationally competitive systems and solutions.

**Mission**

- Design, develop and lead to production state-of-the-art sensors, weapon systems, platforms, and allied equipment for our Defence Services.
- Provide technological solutions to the Defence Services to optimize combat effectiveness and to promote well-being of the troops.
- Develop infrastructure and committed quality work force and build strong technology base.

**Core Competence**

Department of Defence Research and Development (R&D) is working for indigenous development of weapons, sensors & platforms required by the three wings of the Armed Forces. To fulfill this mandate, Department of Defence Research and Development (R&D), is closely working with academic institutions, Research and Development (R&D) Centers and production agencies of Science and Technology (S&T) Ministries/Depts. in Public & Civil Sector including Defence Public Sector Undertakings & Ordnance Factories.

# विषयसूची/
# CONTENTS


Chapter 1. Introduction - 7

Chapter 2. Earthquake - 8

2.1 History of Earthquake in Japan - 8

2.2 definition - 9

Chapter 3. Related work - 9

3.1 Clustering Vs Classification - 10

3.2 Clustering Techniques - 11

Chapter 4. K-Means Algorithm - 11

4.1 K-Means Clustering Algorithm Applications - 11

4.2 Working of K-Means Clustering Algorithm - 12

4.2.1 purpose of clustering - 13

4.3 Implementation of K-means Clustering Graphical form - 13

Chapter 5. Perform K-Means in Python - 17

5.1 Get Dataset - 17

5.1.1 Importing required libraries - 17

5.2 Exploratory Data analysis - 17

5.2.1 Checking the head of the data - 17

5.2.2 Checking datatype - 19

5.3 Implementing the K-Means clustering - 21

Chapter 6. Visualization of K-Means clustering - 23

Chapter 7. Conclusion - 25

*References* - 25

# Chapter 1

## 1. Introduction

The aim of this study is to apply algorithm like **K-mean Clustering algorithm** on Japan Earthquake (2001-2018) Dataset to detect Hotspot.

In Japan earthquakes of the world's most extreme strength occur. In the case of earthquakes with a magnitude of more than 9.0, destruction and devastation can occur within a range of up to 1000 Kilometers.

Due to the special tectonic situation of the country, there are mote earthquakes than average.

Since 1950 more than 100,000 people died by direct consequences of earthquakes. 96 earthquakes also caused a subsequent tsunami, which claimed further lifes and damages.

Many parts of the country have experienced devastating earthquakes and tsunami in the past. The Great Kanto Earthquake, the worst in Japanese history, hit the Kanto plain around Tokyo in 1923 and resulted in the deaths of over 100,000 people.

In January 1995, a strong earthquake hit the city of Kobe and surroundings. Known as the Southern Hyogo Earthquake or Great Hanshin Earthquake, it killed 6,000 and injured 415,000 people. 100,000 homes were completely destroyed and 185,000 were severely damaged..

On March 11, 2011, the strongest ever recorded earthquake in Japan triggered a massive tsunami along the Pacific Coast of northeastern Japan. Known as the Great East Japan Earthquake, the earthquake and particularly the ensuing tsunami killed nearly 20,000 people and caused a nuclear accident at a power plant in Fukushima Prefecture.

Computer science is one of the fields, like other fields, which paid attention to Japan Earthquake. People have discussed Japan Earthquake in all media outlets including social media. The main process used in computer science for dealing with Japan Earthquake is Clustering and its algorithms to refer to and deal with. Computer intelligence and digital analysis is also one of the fields that has taken this into consideration using clustering algorithms.

Clustering is an important feature that is also applied to learning data. The unsupervised clustering is defined as segmenting the data into clusters that contain data of the same characteristics, which mainly means sorting the data to make homogenous groups. Clustering algorithms are applied in different fields namely, image segmentation, data cleaning and exploratory analysis, information retrieval, web pages grouping, market segmentation and scientific and engineering analysis Clustering can also be used as a preprocessing step to identify pattern classes for subsequent supervised classification .

# Chapter 2

## 2. Earthquake

An **earthquake** (also known as a **quake**, **tremor** or **temblor**) is the shaking of the surface of the Earth resulting from a sudden release of energy in the Earth's lithosphere that creates seismic waves. Earthquakes can range in intensity, from those that are so weak that they cannot be felt, to those violent enough to propel objects and people into the air and wreak destruction across entire cities. The **seismic activity** of an area is the frequency, type, and size of earthquakes experienced over a particular time period. The seismicity at a particular location in the Earth is the average rate of seismic energy release per unit volume. The word *tremor* is also used for non-earthquake seismic rumbling.

At the Earth's surface, earthquakes manifest themselves by shaking and displacing or disrupting the ground. When the epicenter of a large earthquake is located offshore, the seabed may be displaced sufficiently to cause a tsunami. Earthquakes can also trigger landslides.

In its most general sense, the word *earthquake* is used to describe any seismic event—whether natural or caused by humans—that generates seismic waves. Earthquakes are caused mostly by rupture of geological faults but also by other events such as volcanic activity, landslides, mine blasts, and nuclear tests. An earthquake's point of initial rupture is called its hypocenter or focus. The epicenter is the point at ground level directly above the hypocenter.

### 2.1 History of Earthquake in Japan

Japan is situated in the collision zone of at least four lithospheric plates: the Eurasian/Chinese Plate, the North American Plate, the Philippine Plate and the Pacific Plate. The continuous movements of these plates generate a lot of energy released from time to time in earthquakes and tsunamis of varying magnitude and effects (Geologist *Callan Bentley* discusses in great detail the geological setting of the Japanese Islands).

December 31, 1703 Japan was hit by a strong earthquake with a reconstructed intensity of 8 after the Mercalli-scale. In Edo most of the wood buildings collapsed. The earthquake and its aftermath effects, like floods and fires, killed estimated 150.000 people. More than 6.500 people were killed by a flood wave, which caused havoc in the bay of *Sagami* and on the peninsula of *Boso*.
One of the most important historic earthquakes hit Tokyo November 11, 1855, killing 16.000 to 20.000 people. This event and the aftermath are retold by hundreds of woodcuts, especially in the form of a *namazu-e*.

October 28, 1891, the agricultural region of *Nobi* experienced an earthquake of magnitude 8. Modern buildings and traditional houses were damaged or collapsed, thousands of people lost their homes and 7.000 people were killed.

The English geologists *John Milne* (1849-1913), who in 1880 founded the *Seismologists Society of Japan*, studied the effects of this earthquake and published an important monographic work "*The great earthquake in Japan, 1891*". The Japanese geologist *Bunjiro*

*Koto* observed a superficial dislocation during the same event and recognized a fundamental principle in seismology - faults are not the result of an earthquake, but movements along a fault cause the seismic waves.

During the second half of the 19th and beginning of the 20th century scientific research on earthquakes became rapidly established in Japan.
*Fusakichi Omori* (1868-1923), director of the *Seismological Institute of Japan*, studied the occurrence of earthquakes around Tokyo and noted in 1922:
"*Currently the immediate area of Tokyo is seismically quiet, while in the mountains around Tokyo, in a distance of about 60 kilometres, there are often triggered earthquakes, which - although they are may felt in the capital - are in fact harmless, because the affected areas are not part of a larger destructive seismic zone.*

Machine learning, similar to other methods, played an important role in finding out more about the causes and the conditions of the earthquake. That was an attempt to clean the noisy data spread worldwide in order to inform biological fields where researchers were making every effort to know how to counter such disaster, and to know impact of the different variables like the temperature, populations, due to Earthquake in Japan.

## 2.2 Definition

The clustering and classification problems are essential and admired topics of research in the area of pattern recognition and data mining.

The conventional binary and multiclass classifiers are surely not suitable for this target-task mining task because, in this unsupervised learning mode, it is always possible that some of the clusters may not be assigned with any target-class, whereas an increase in the number of target-tasks to solve this problem leads to generation of duplicate information.

These conventional classifiers work fine in the presence of at least two well-defined classes but may become biased, if the dataset suffers from data irregularity problems (imbalanced classes, small disjunct, skewed class distribution, missing values, etc.). Specially, when a class is ill-defined, the classifier may give biased outcome.

# Chapter 3

# 3 Related work

Clustering analysis or simply Clustering is basically an Unsupervised learning method that divides the data points into a number of specific batches or groups, such that the data points in the same groups have similar properties and data points in different groups have different properties in some sense.

Clustering is a set of techniques used to partition data into groups, or clusters. **Clusters** are loosely defined as groups of data objects that are more similar to other objects in their cluster than they are to data objects in other clusters. In practice, clustering helps identify two qualities of data:

1. Meaningfulness
2. Usefulness

**Meaningful** clusters expand domain knowledge. For example, in the medical field, researchers applied clustering to gene expression experiments. The clustering results identified groups of patients who respond differently to medical treatments.

**Useful** clusters, on the other hand, serve as an intermediate step in a data pipeline. For example, businesses use clustering for customer segmentation. The clustering results segment customers into groups with similar purchase histories, which businesses can then use to create targeted advertising campaigns.

With the rising use of the Internet in today's society, the quantity of data created is incomprehensibly huge. Even though the nature of individual data is straightforward, the sheer amount of data to be analyzed makes processing difficult for even computers.

To manage such procedures, we need large data analysis tools. Data mining methods and techniques, in conjunction with machine learning, enable us to analyze large amounts of data in an intelligible manner. k-means is a technique for data clustering that may be used for unsupervised machine learning. It is capable of classifying unlabeled data into a predetermined number of clusters based on similarities (k).

## 3.1 Clustering Vs Classification

Before starting our discussion on k-means clustering, I would like point out the difference between clustering and classification.

- Samples in a classification task have labels. Each data point is classified according to some measurements. Classification algorithms try to model the relationship between measurements (features) on samples and their assigned class. Then the model predicts the class of new samples.

- Samples in clustering do not have labels. We expect the model to find structures in the data set so that similar samples can be grouped into clusters. We basically ask the model to label samples

**Clustering is an** unsupervised learning **technique where we try to group the data points based on specific characteristics. There are various clustering algorithms with** K-Means **and** Hierarchical **being the most used ones. Some of the use cases of clustering algorithms include:**

- Document Clustering
- Recommendation Engine
- Image Segmentation
- Market Segmentation
- Search Result Grouping
- and Anomaly Detection.

## 3.2 Clustering Techniques

Clustering is an unsupervised learning approach used iteratively to create groups of relatively similar samples from the population. In this research, K-means clustering techniques have been used to create the clusters of the available samples (research articles):

# Chapter 4

# 4 K-Means Algorithm

The K-means clustering algorithm computes centroids and repeats until the optimal centroid is found. It is presumptively known how many clusters there are. It is also known as the flat clustering algorithm. The number of clusters found from data by the method is denoted by the letter 'K' in K-means.

In this method, data points are assigned to clusters in such a way that the sum of the squared distances between the data points and the centroid is as small as possible. It is essential to note that reduced diversity within clusters leads to more identical data points within the same cluster.

## 4.1 K-Means Clustering Algorithm Applications

The performance of K-means clustering is sufficient to achieve the given goals. When it comes to the following scenarios, it is useful:

- To get relevant insights from the data we're dealing with.
- Distinct models will be created for different subgroups in a cluster-then-predict approach.

- Market segmentation
- Document Clustering
- Image segmentation
- Image compression
- Customer segmentation
- Analyzing the trend on dynamic data

## 4.2 Working of K-Means clustering Algorithm

The following stages will help us understand how the K-Means clustering technique works-

- ***Step 1:*** First, we need to provide the number of clusters, K, that need to be generated by this algorithm.
- ***Step 2:*** Next, choose K data points at random and assign each to a cluster. Briefly, categorize the data based on the number of data points.
- ***Step 3:*** The cluster centroids will now be computed.
- ***Step 4:*** Iterate the steps below until we find the ideal centroid, which is the assigning of data points to clusters that do not vary.
- 4.1 The sum of squared distances between data points and centroids would be calculated first.
- 4.2 At this point, we need to allocate each data point to the cluster that is closest to the others (centroid).
- 4.3 Finally, compute the centroids for the clusters by averaging all of the cluster's data points.

K-means implements the **Expectation-Maximization** strategy to solve the problem. The Expectation-step is used to assign data points to the nearest cluster, and the Maximization-step is used to compute the centroid of each cluster.

***When using the K-means algorithm, we must keep the following points in mind:***

- It is suggested to normalize the data while dealing with clustering algorithms such as K-Means since such algorithms employ distance-based measurement to identify the similarity between data points.
- Because of the iterative nature of K-Means and the random initialization of centroids, K-Means may become stuck in a local optimum and fail to converge to the global optimum. As a result, it is advised to employ distinct centroids' initializations.

### 4.2.1 purpose of clustering

- The **goal of cluster** analysis or **clustering** is to group a collection of objects in such a way that objects in the same group (called a **cluster**) are more like each other (in some sense) than objects in other groups (**clusters**)
- In many areas, cluster analysis is used, including pattern recognition, image analysis, retrieval of information, bioinformatics, compression of data, computer graphics, and machine learning.

## 4.3 Implementation of K-Means Clustering Graphical Form

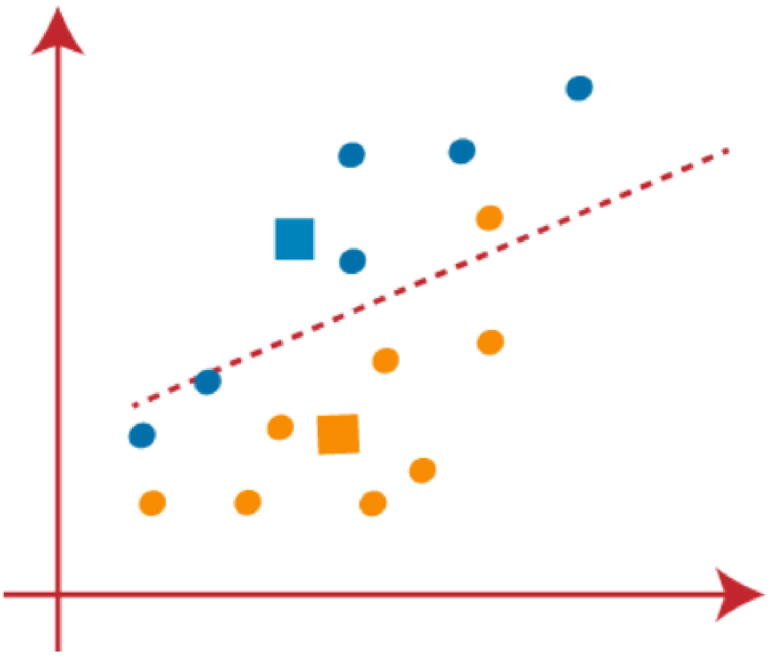
STEP 1: Let us pick k clusters, i.e., K=2, to separate the dataset and assign it to its appropriate clusters. We will select two random places to function as the cluster's centroid.

STEP 2: Now, each data point will be assigned to a scatter plot depending on its distance from the nearest K-point or centroid. This will be accomplished by establishing a median between both centroids. Consider the following illustration:

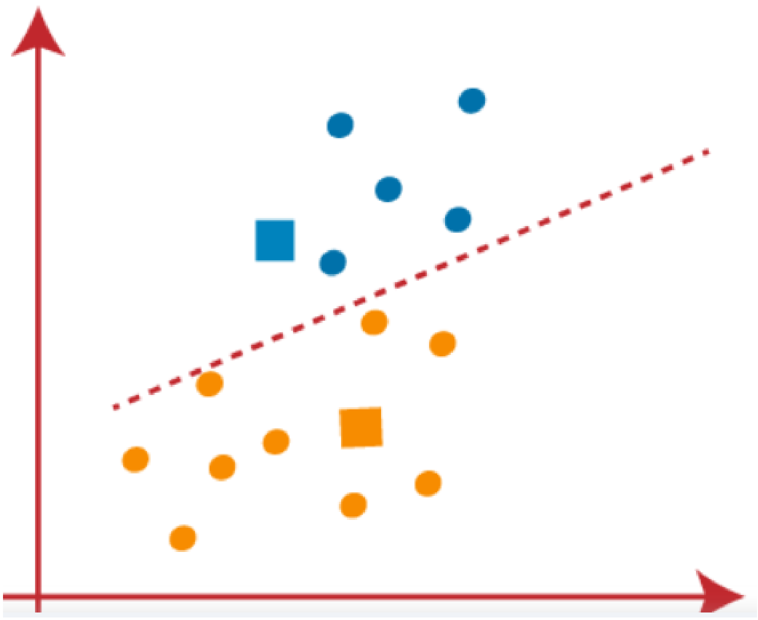
STEP 3: The points on the line's left side are close to the blue centroid, while the points on the line's right side are close to the yellow centroid. The left Form cluster has a blue centroid, whereas the right Form cluster has a yellow centroid.

STEP 4: Repeat the procedure, this time selecting a different centroid. To choose the new centroids, we will determine their new center of gravity, which is represented below:

STEP 5: After that, we'll re-assign each data point to its new centroid. We shall repeat the procedure outlined before (using a median line). The blue cluster will contain the yellow data point on the blue side of the median line.

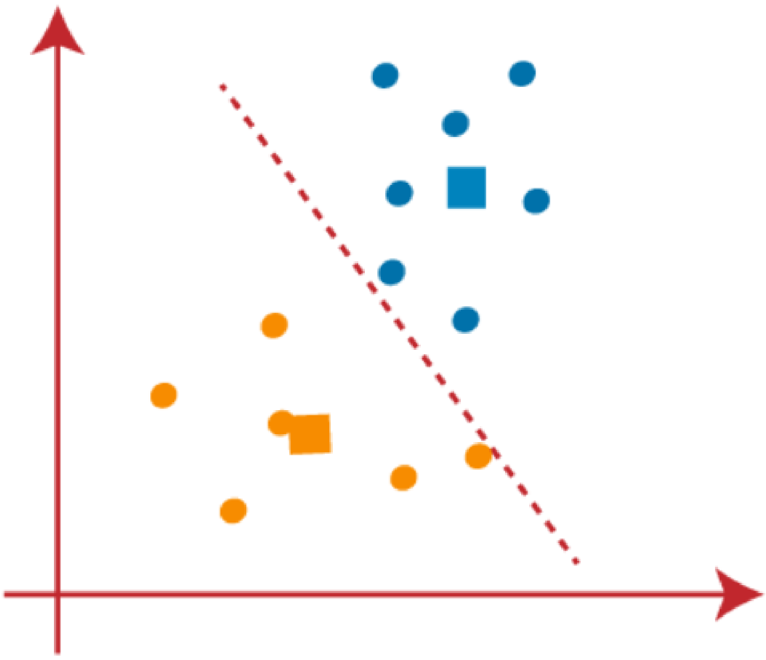


STEP 6: Now that reassignment has occurred, we will repeat the previous step of locating new centroids.

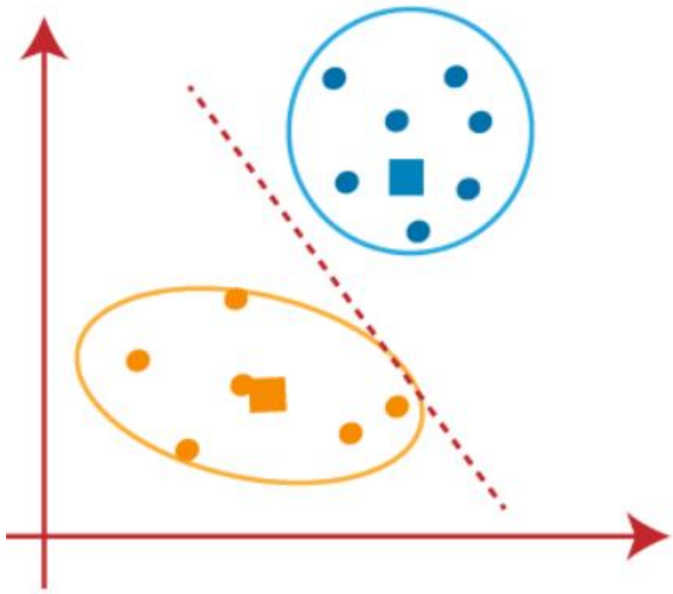


STEP 7: We will repeat the procedure outlined above for determining the center of gravity of centroids, as shown below.

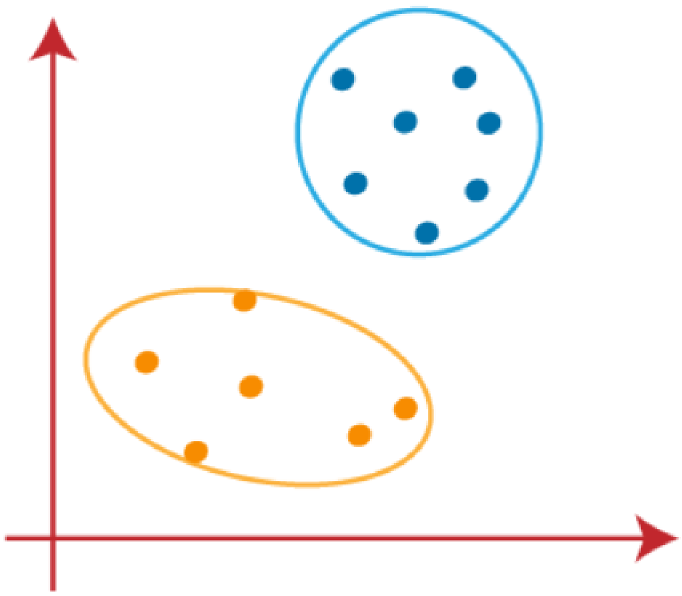
STEP 8: Similar to the previous stages, we will draw the median line and reassign the data points after locating the new centroids.



STEP 9: We will finally group points depending on their distance from the median line, ensuring that two distinct groups are established and that no dissimilar points are included in a single group.



The final Cluster is as follows:



**How to Implementing K-Means Clustering ?**

- Choose the number of clusters k
- Select k random points from the data as centroids
- Assign all the points to the closest cluster centroid
- Recompute the centroids of newly formed clusters
- Repeat steps 3 and 4

## Chapter 5

# 5. Perform K-Means in Python

To perform K-Means clustering in Python, you will require to install sklearn, pandas, and matplotlib Python packages.

## 5.1 Get Dataset

For clustering using K-Means, I am using Japan Earthquake Dataset.

### 5.1.1 Importing required libraries

Let us begin by importing the required libraries for implementation on the algorithm.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

data = pd.read_csv("japan.csv")
```

## *5.2 Exploratory data analysis*

This is the process of investigating the available data and determining inconsistencies in patterns and other anomalies with the help of graphical representations and statistical summaries.

### 5.2.1. Checking the head of the data.

```
df.head()
```

- Output

| | time | latitude | longitude | depth | mag | magType | nst | gap | dmin | rms | ... | updated | place | type | horizontalError |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 2018-11-27T14:34:20.900Z | 48.3780 | 154.9620 | 35.00 | 4.9 | mb | NaN | 92.0 | 5.044 | 0.63 | ... | 2018-11-27T16:06:33.040Z | 269km SSW of Severo-Kuril'sk, Russia | earthquake | 7.6 |
| **1** | 2018-11-26T23:33:50.630Z | 36.0733 | 139.7830 | 48.82 | 4.8 | mww | NaN | 113.0 | 1.359 | 1.13 | ... | 2018-11-27T16:44:22.223Z | 3km SSW of Sakai, Japan | earthquake | 6.0 |
| **2** | 2018-11-26T13:04:02.250Z | 38.8576 | 141.8384 | 50.56 | 4.5 | mb | NaN | 145.0 | 1.286 | 0.84 | ... | 2018-11-26T23:52:21.074Z | 26km SSE of Ofunato, Japan | earthquake | 8.4 |
| **3** | 2018-11-26T05:20:16.440Z | 50.0727 | 156.1420 | 66.34 | 4.6 | mb | NaN | 128.0 | 3.191 | 0.62 | ... | 2018-11-26T08:13:58.040Z | 67km S of Severo-Kuril'sk, Russia | earthquake | 9.7 |
| **4** | 2018-11-25T09:19:05.010Z | 33.9500 | 134.4942 | 38.19 | 4.6 | mb | NaN | 104.0 | 0.558 | 0.61 | ... | 2018-11-25T23:24:52.615Z | 9km SW of Komatsushima, Japan | earthquake | 3.4 |

### 5.2.2. Checking datatype

```
data.dtypes
```

```
time                object
,latitude           float64
,longitude          float64
,depth              float64
,mag                float64
,magType             object
,nst                float64
,gap                float64
,dmin               float64
,rms                float64
,net                 object
,id                  object
,updated             object
,place               object
,type                object
,horizontalError    float64
,depthError         float64
,magError           float64
,magNst             float64
,status              object
,locationSource      object
,magSource           object
,dtype: object
```

```
data.shape
```

```
(14092, 22)
```

Additionally, we can get the total number of rows and the total number of columns in this dataset by doing the following:

```python
import numpy as np
```

```python
df = data.select_dtypes([np.float64])
```

```python
df.shape
```

```
(14092, 12)
```

According to the result, this dataset has 14092 samples with 12 features.

```python
from sklearn.cluster import KMeans
```

```python
df.isnull().sum()
```

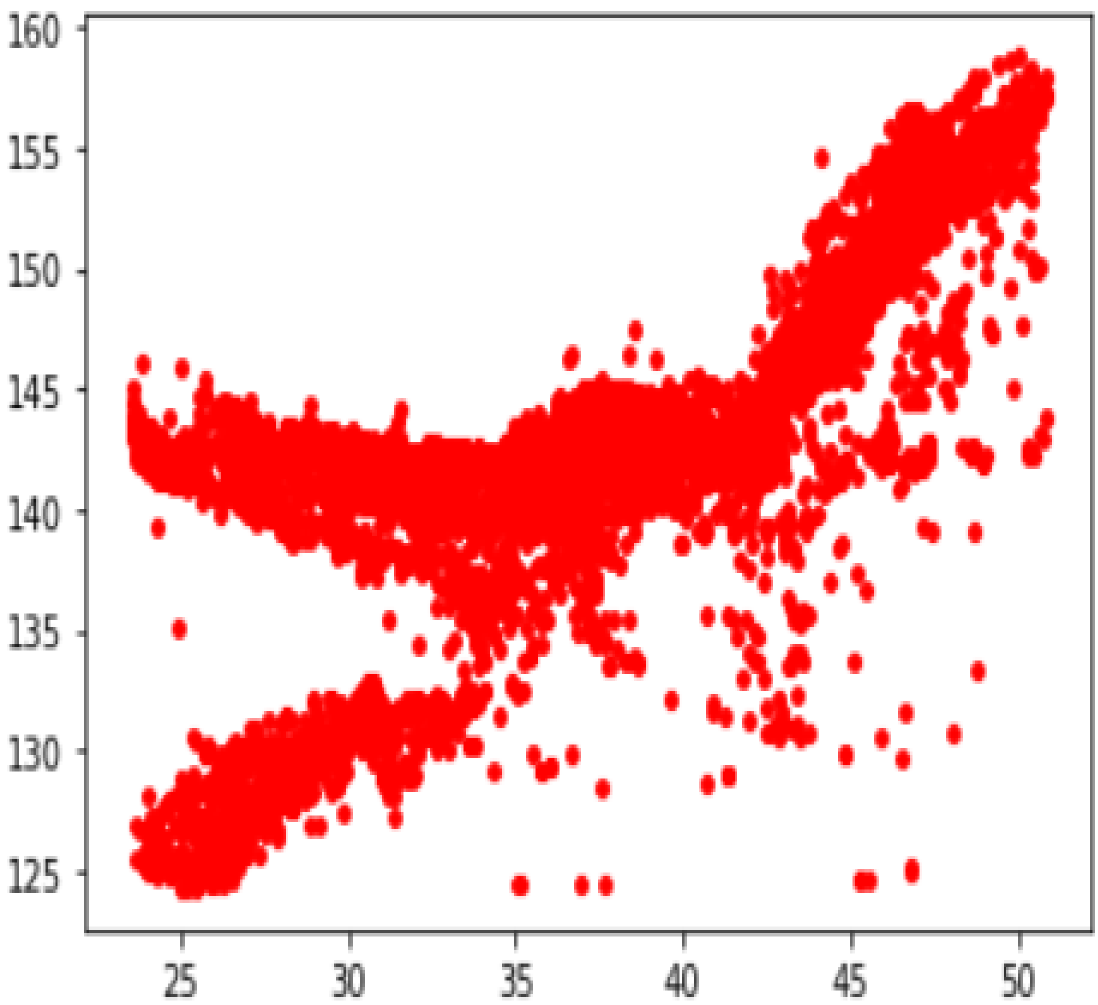
```
latitude               0
,longitude             0
,depth                 0
,mag                   0
,nst                3609
,gap                 782
,dmin              10485
,rms                  78
,horizontalError   11292
,depthError         5052
,magError          10661
,magNst             3044
,dtype: int64
```

## 5.3. Implementing the K-Means clustering

**Following that, the code below will assist us in visualizing the dataset.**

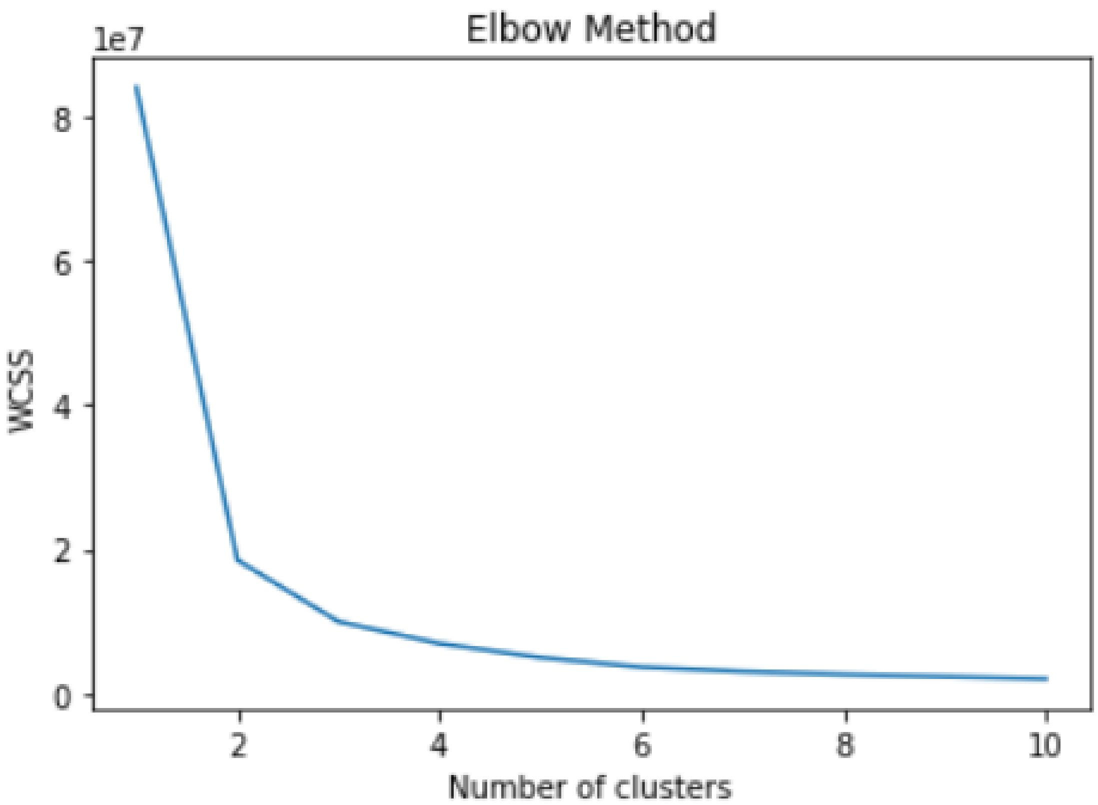
```
plt.scatter(df.iloc[:,0], df.iloc[:,1],s=15, c= "red")
```

```
<matplotlib.collections.PathCollection at 0x21278554d60>
```



**Using the elbow method to determine the optimal number of clusters for k-means clustering**

```
wcss = []
for i in range(1,11):
    kmeans = KMeans(n_clusters=i, init = "k-means++",
                    max_iter= 300, n_init=10)
    kmeans.fit(df)
    wcss.append(kmeans.inertia_)
plt.plot(range(1,11), wcss)
plt.title("Elbow Method")
plt.xlabel("Number of clusters")
plt.ylabel("WCSS")
plt.show()
```



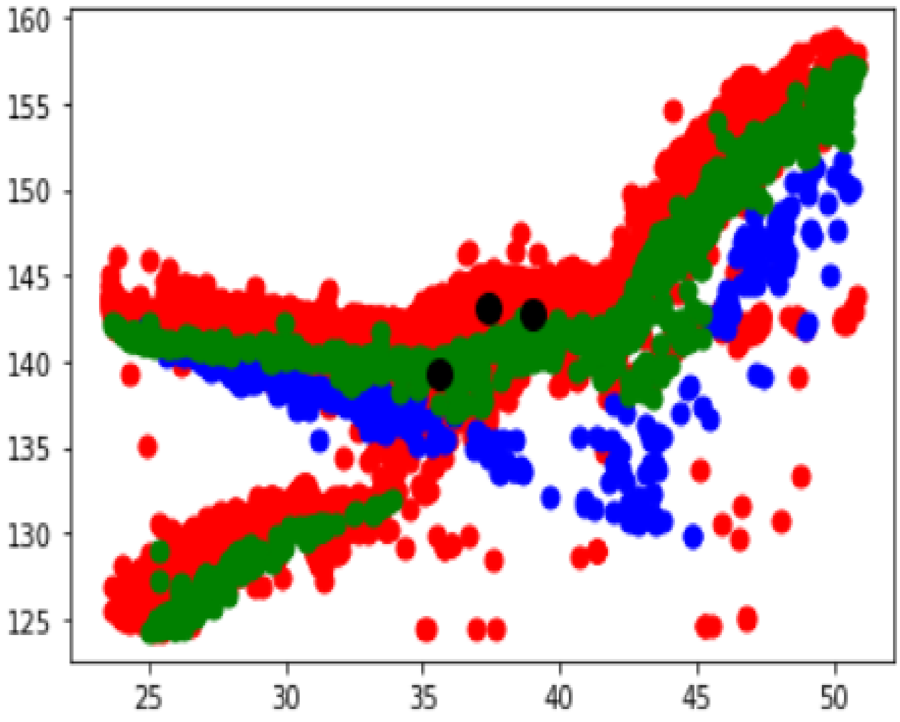
## Chapter 6

# 6. Visualization of K-Means clustering

Now, using the code below, we can plot and visualize the cluster's centers as determined by the k-means Python estimator-

```
plt.scatter(df.iloc[labels==0,0], df.iloc[labels==0,1], s=50, c= "red")
plt.scatter(df.iloc[labels==1,0], df.iloc[labels==1,1], s=50, c= "blue")
plt.scatter(df.iloc[labels==2,0], df.iloc[labels==2,1], s=50, c= "green")

plt.scatter(kmeans.cluster_centers_[:,0],
            kmeans.cluster_centers_[:,1], s=100, c= "black")
plt.title("")
plt.xlabel("")
plt.ylabel("")
plt.show()
```



## Pros and Cons K-Means

**Pros:**

- Easy to interpret
- Relatively fast
- Scalable for large data sets
- Able to choose the positions of initial centroids in a smart way that speeds up the convergence
- Guarantees convergence

**Cons:**

- Number of clusters must be pre-determined. K-means algorithm is not able to guess how many clusters exist in the data. Determining number of clusters may well be a challenging task.
- Can only draw linear boundaries. If there is a non-linear structure separating groups in the data, k-means will not be a good choice.
- Slows down as the number of samples increases because at each step, k-means algorithm accesses all data points and calculates distances. An alternative way is to use a

subset of data points to update the location of centroids (i.e. sklearn.cluster.MiniBatchKMeans)

- Sensitive to outliers

# Chapter 7

## 7. conclusion

*K-Means* clustering is a simple yet very effective unsupervised machine learning algorithm for data clustering. It clusters data based on the Euclidean distance between data points. K-Means clustering algorithm has many uses for grouping text documents, images, videos, and much more. K-means clustering is a widely used approach for clustering. Generally, practitioners begin by learning about the architecture of the dataset. K-means clusters data points into unique, non-overlapping groupings. It works very well when the clusters have a spherical form. However, it suffers from the fact that clusters' geometric forms depart from spherical shapes.

Additionally, it does not learn the number of clusters from the data and needs that it be stated beforehand. It's always beneficial to understand the assumptions behind algorithms/methods in order to have a better understanding of each technique's strengths and drawbacks. This will assist us in determining when and under what conditions to utilize each form.

In this Report, we have covered the K-Means algorithm. First, we looked at its key parameters and how this algorithm clusters the data points. We also learned about data points associated with the K-Means algorithm. Later we looked at how we implement this algorithm on Japan Earthquake Dataset then by using K-Means, We have detected Japan Earthquake hotspot.

For more detail about this **Japan Earthquake Hotspot Detection Using K-Means Clustering Method Method** Report and Dataset and Codes, I am mentioning my Github repository link below:

➔ https://github.com/neerjatiwari/Japan-Earthquake-hotspot-detection

# *References*


1. https://www.analyticsvidhya.com/blog/2021/11/understanding-k-means-clustering-in-machine-learningwith-examples/

2. https://www.kaggle.com/code/khotijahs1/k-means-clustering-of-iris-dataset

3. https://towardsdatascience.com/a-practical-guide-on-k-means-clustering-ca3bef3c853d

4. Dataset Source: https://www.kaggle.com/datasets/aerodinamicc/earthquakes-in-japan?resource=download&select=Japan+earthquakes+2001+-+2018.csv
5. https://stackabuse.com/k-means-clustering-with-scikit-learn/
6. https://rpubs.com/AnanyaDu/361293#:~:text=Iris%20Dataset%20%2D%20Clustering%20using%20K%20means&text=The%20data%20gives%20the%20measurements,to%20cluster%20them%20into%20groups.
7. https://www.japan-guide.com/e/e2116.html
8. http://eprints.lse.ac.uk/62495/1/Hunter_Earthquakes%20in%20Japan.pdf