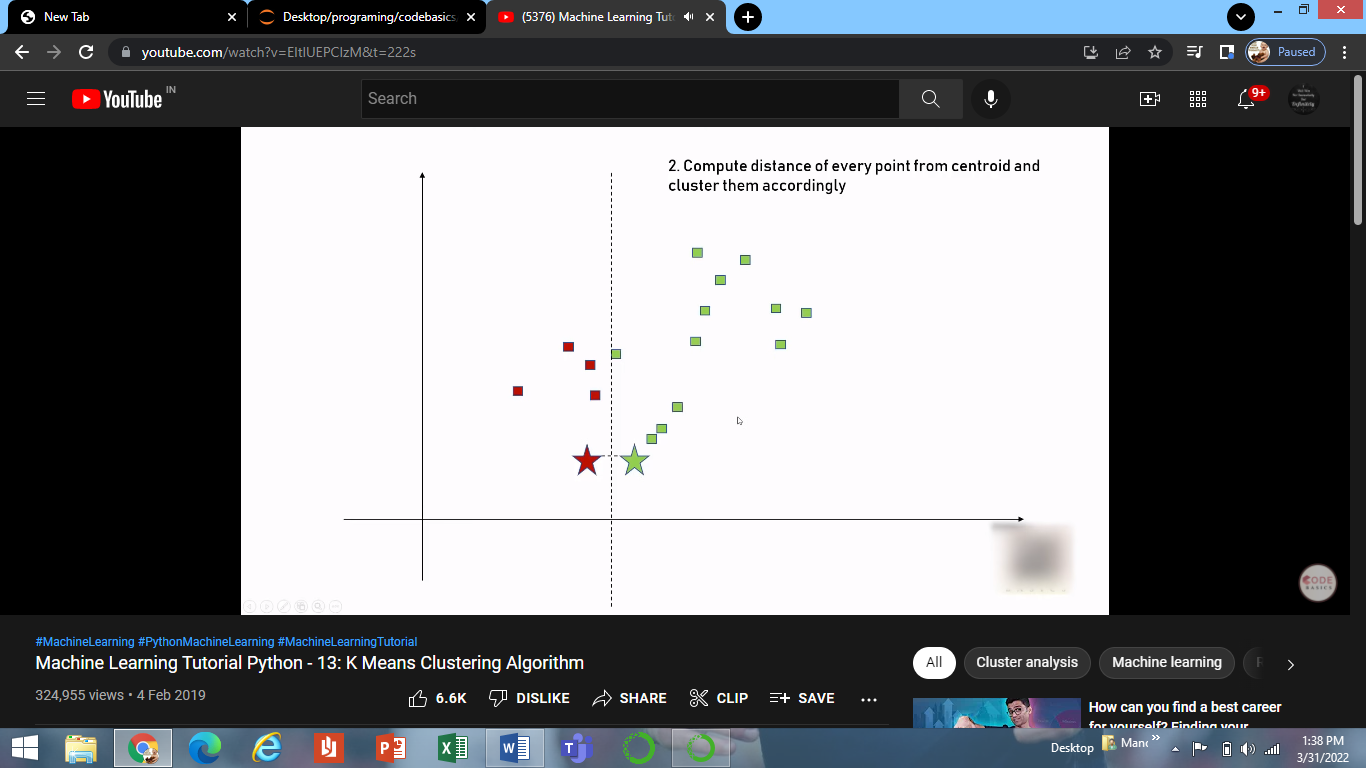
**K means Clustering**

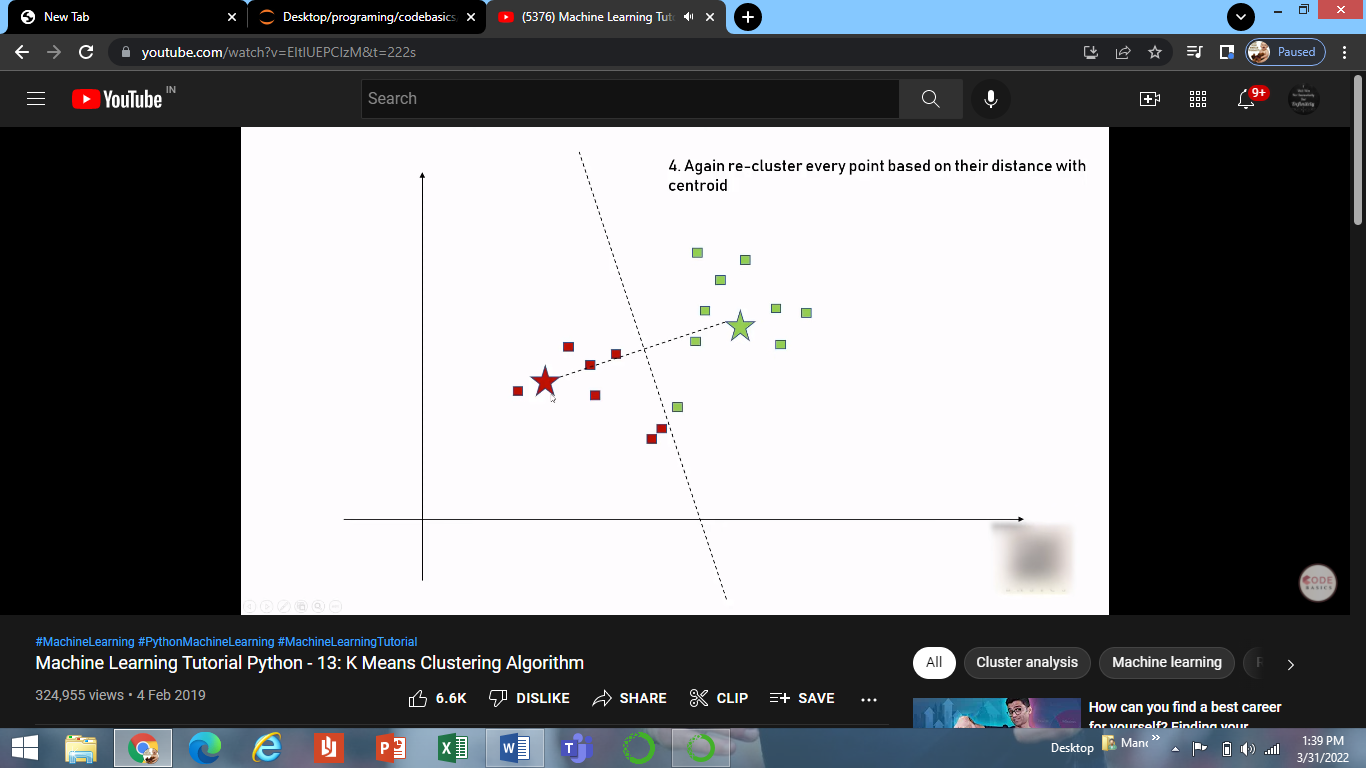
(Unsupervised machine Learning Algorithm)

**Supervised Learning:** We have the input features and target variables present in the dataset. These are used to train the model

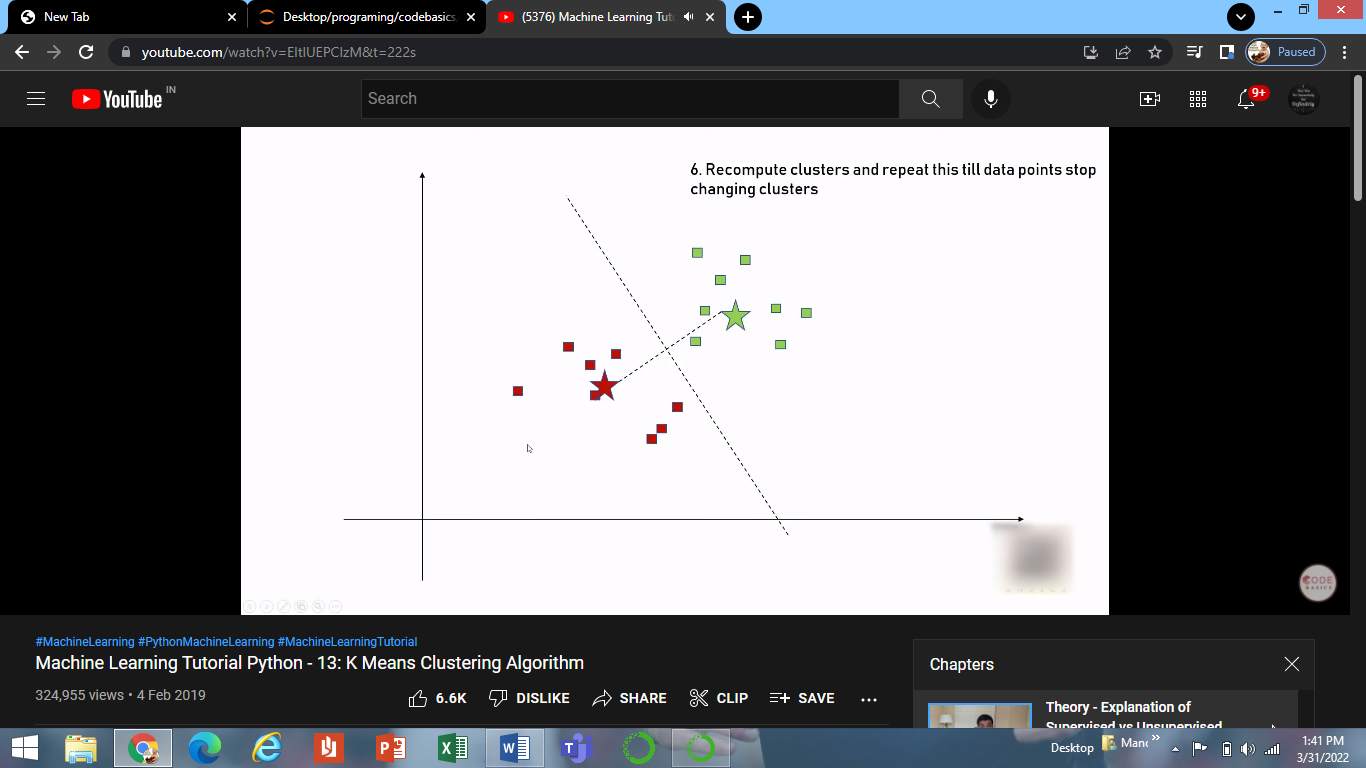
**Unsupervised Learning:** All we have is a set of features, We do not have any target variable. We try to find underlying clusters (or structure) in the data that may have some common characteristics

**‘K’ needs to specified while creating the model. K = No. of clusters required in the model**

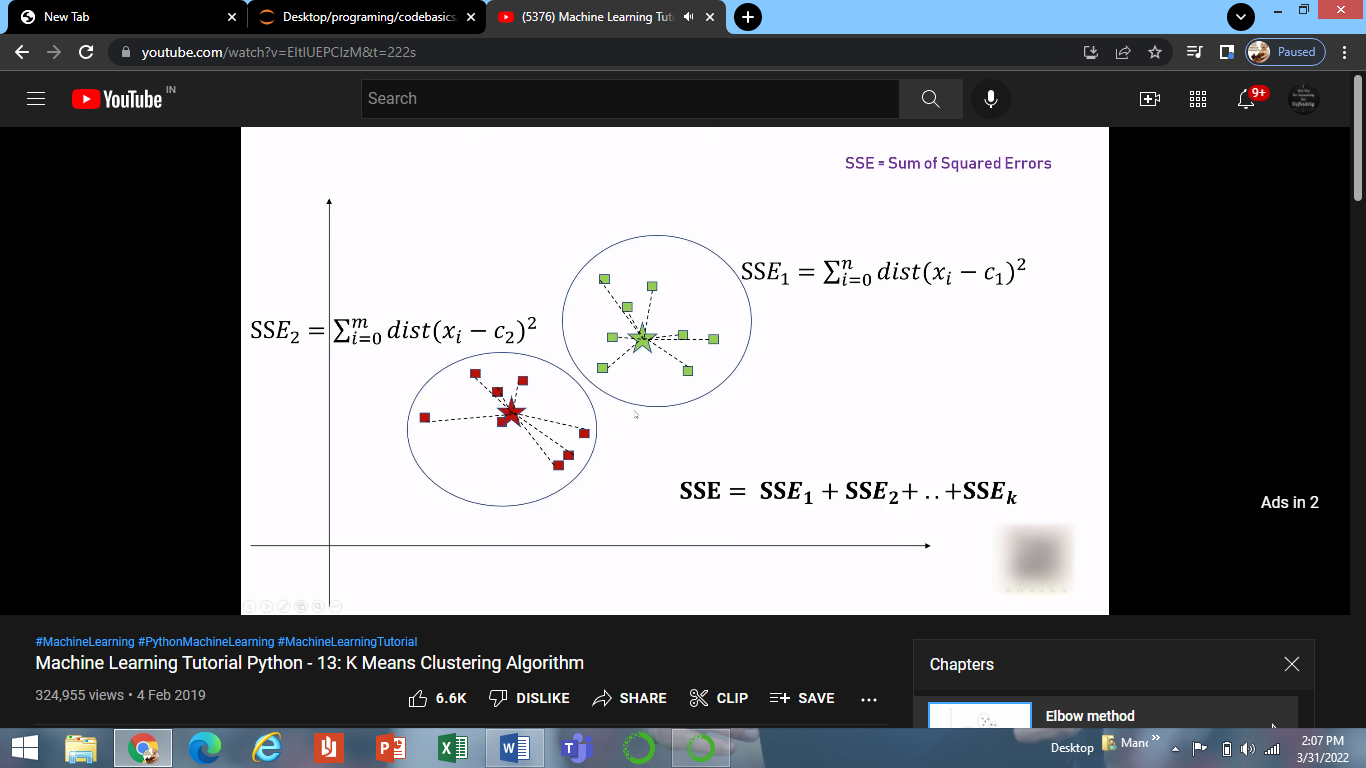
The model creates ‘K’ centre points and calculates the distance of each datapoint from the centres. It assigns the datapoint to any of the centres depending on the shortest distance. After assignment of each datapt to centres, the clusters are formed and the centres are placed at the center of gravity (i.e. center) of the cluster.



Now, the model again calculates the distance of each datapoint from the cluster centres and reassigns the points to respective centres and readjusts the position of each cluster centre.



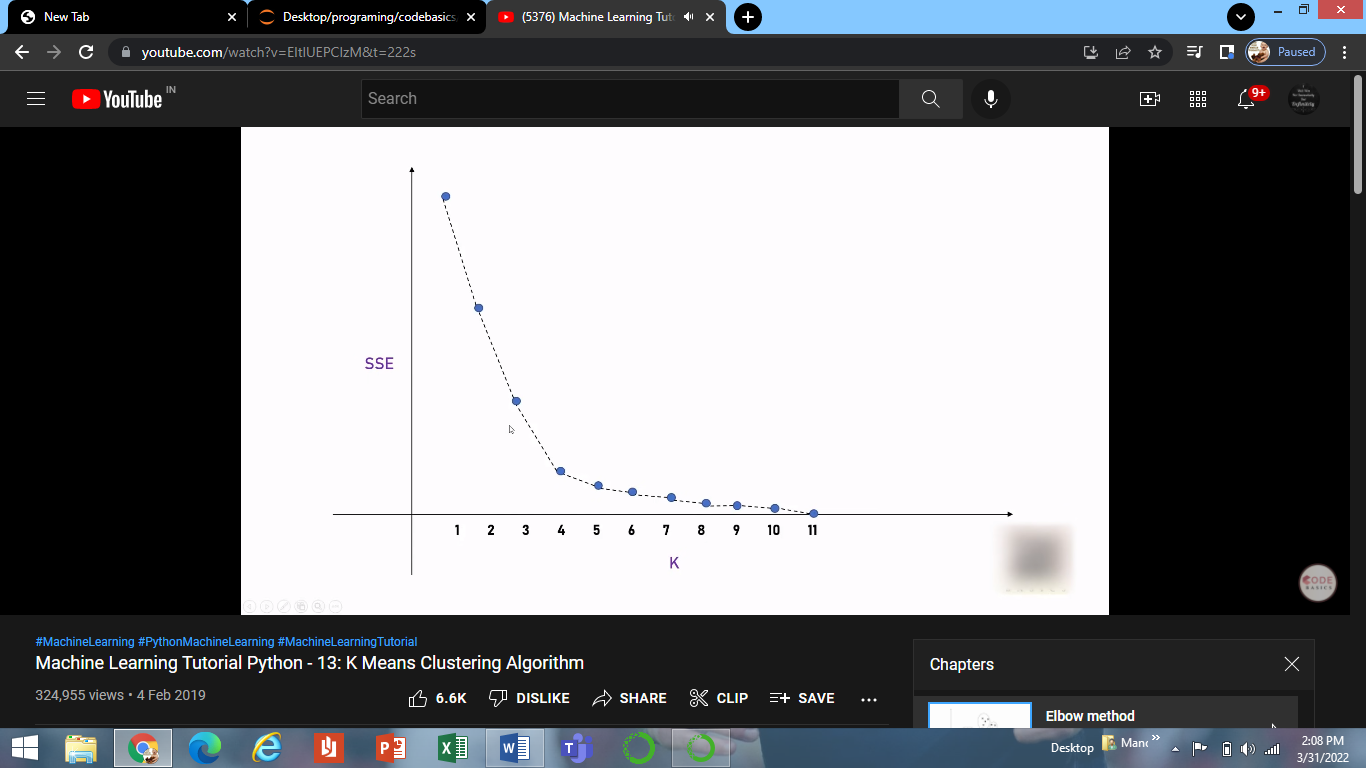
This process is repeated iteratively till the datapoints stop changing clusters

**How to determine the best value of ‘K’? …………. (Using Elbow Method)**

Start with some value of ‘k’

Find the Sum Squared Error (SSE) for each cluster (square of distance between cluster center and each point)

Find the Sum of SSE for each cluster… (Total SSE)



Our objective is to minimize this Total SSE

Note: The optimal value for ‘K’ may differ for each model

By inituition and common sense, the error will reduce as ‘K’ increases and eventually become 0 when ‘K’ = No. of datapoints. (i.e. each datapoint will be a separate cluster in itself)

So the Optimal Value of ‘K’ is at the elbow of the chart representing SSE on Y axis and K on X axis

(K = 4 is a good K value in the given example)