So guys, we are going to start our first unsupervised machine learning algorithm that is called as K-means clustering algorithm. And uh, in this video we are going to understand the geometric intuition behind K-means clustering algorithm. Now, first of all, let's consider that I have some data points over here. And obviously I can show you in two dimensions. So over here let's say that I have all my data points which is basically spread like this. Let's say between x and y axis. And suppose if I have this kind of data points, obviously directly from seeing this particular data points, you can see that, okay, there are two groups of data points together. Our final aim is that whenever we have this kind of data points, this entire data points after applying K-means clustering will become something like this. So let's say if these are my data points, I'm just trying to draw the same one. So don't uh, worry if 1 or 2 points are not missed, right? So after applying K-means. We are going to get something output which looks like this. This will be one group and this will be the other group. Okay. And remember, um, this this will be my one cluster. Let's say this will be my cluster one, and this will be my cluster two. Along with this clusters, you will also be getting one important thing which is called as centroids. Now since here I have got two clusters out of it. So I am going to get two centroids of this clusters. Okay. This will be one centroid and this will be the other centroid. Similarly if let's say I have data points which may probably having three different groups. Suppose let's say this is my one group. This is my another group. And this may be my another group. Right? So after I probably apply K-means again, I'm going to get an output which will look something like this. Okay, so this is the entire geometric intuition behind K-means clustering. And we'll try to understand how this entire step will basically happen. Okay. So let's say these are all my points. So here also I'm going to get one centroid over here. Sorry one one group over here. Second group over here. Third group over here. But different different centroids. Okay. So in short what we are able to do is that we are able to cluster similar kind of points together okay. That is the main aim behind K-means clustering. So now once you have understood geometric intuition, let's go ahead and let's see that how this entire group is basically getting created. Okay. So again uh, I'm just going to draw the same points over here. Let's consider what. Groups like this. Okay, let's say I have some of the data points over here. And some of the data points over here. And we will go step by step in creating these groups. Okay. So the first step is that we initialize some centroids. Okay. Initialize some centroids, that is some k value. That is the reason we say k means okay. And whenever I say about centroids I'm basically talking about k value okay. This is nothing. But this is called as centroids. Now in this particular case let's consider that okay I know there are two groups. So I'm going to initialize two different centroids okay. So in this particular case I can definitely draw one centroid over here. And this is randomly initialized. And the other centroid somewhere over here okay. Over here, or let me draw it over here so that it will be easy. Right. So this is the first step. We initialize some centroids. Now you may be seeing Chris why k is equal to two. You have taken. Okay. We'll try to understand how to select the k value later on. But let's understand that okay. I'm just going to know that I definitely know that for this particular thing you definitely require two centroids. But later on I'll show you how you can select the k values okay. Now the second step that is and it is a super important step, is that we will try to find out. The distance from this particular centroid to all the points. Let's say for this particular point, if I say this is, we'll just try to find out for which centroid this point is near, whether this centroid it is near or this centroid it is near. Here you can definitely see that obviously the distance between the centroid and this point will be near when compared to this. So if this particular centroid is near to this we are going to just mark it as orange color okay. Similarly you will be able to see this points will be sent near to this. This point will be near to this in a simpler way. If I really want to find out, uh, how do I do it? I'll just try to create a straight line. Let's say this is my straight line, and I'll try to create another perpendicular line which will be passing through this line. Okay. So

in this particular case, let's say this is not a exactly perpendicular line. But yes. Now in this particular scenario you can see this is a perpendicular line to this particular line. And obviously to find out the distance, uh, we will be using something called as Euclidean distance or Manhattan distance. Okay. We will also understand about how does Euclidean distance actually work and Manhattan distance actually work. Okay. So we will be using either one of this. And we'll also try to understand that when should we use Euclidean distance or when should we use Manhattan distance. Okay. All those points will be covered. Now here you can see that all the points that are below this line will basically be nearer to this particular yellow centroid. Okay. So most of the points that you will be seeing over here, this will become yellow, this will become yellow. This will become yellow. This will become yellow. This will become yellow. This will become yellow. And similarly this will become yellow. Okay. And uh, with respect to all the other points here, what will happen is that here this will become orange, this will become orange, this will become orange, orange, orange, orange, orange, orange. That basically means this points are nearer to the centroid. Okay. And this will become yellow okay. Now in the second step I have already seen we need to find out the points that are. Nearest to the centroids. Nearest to the centroids. And we will mark them in that particular group. Okay. So right here you can see over here the centroids color is basically given over here just to indicate in this way. Now coming to the third point which is a super important point. Now after the marking is done, what we are going to find out is that we're just going to move the centroids. Now how do we move the centroids. This is super important. Now you know that you have got your group of points that are nearer to the centroid and nearer to the centroid. We are just going to find out the average of all these points. My average of all these parts. That basically means we are going to consider all this point and find out the average, right? So I will just remove this. Okay. So let's say these are my centroids that are near to this points. So I'm just going to find out the average and whatever average will come this centroid will get updated. So if I go to my next diagram then what may happen is that. So these are my points over here. These are my points over here. Okay I'm again going to use this okay. These are my points over here. And here you can see some of the points that are probably yellow in color. So these are some other points that I'm just marking so that you'll be able to understand. Now when my centroid is getting updated then what will happen? It will probably move from this point to this point. From the older, older point to this point, we are just going to try to find out the average. So here average will be computed. The average point will basically be the new centroid location. Similarly with respect to this all points you will be able to see. Some other points are here and some of the points are here. And initially my previous centroid was here. Now let's say it was here. Now it is going to move to a new location. Let's say the location is somewhere. Since most of the points are away it is going to move it over here okay. So I'm just going to rub it and I'm just going to use the new position. So here it is my new position of my centroid okay. Perfect. So this was my third step. Move the centroids by finding out the average. Again we will be repeating this all steps. Okay that is my second step okay. Then again we really need to find out which are the points nearer to the centroid again. So here you will be able to see again if I probably try to create two lines that are perpendicular to each other. Okay, one point over here and one point over here. Suppose let's say this is exactly center perpendicular to each other. Now what is going to happen in this particular case here you can clearly see that some of the points have again got mismatch. Right now this yellow point, this yellow point will become an orange point because this is nearer to this particular centroid. What about this point. This will basically become my. Yellow parts, right? Then again, what will happen again? The same step will happen. We'll try to move the centroids now by finding out the average. Now once I find out the average, my new points will move from here probably to here okay. Similarly, my yellow points will move from here to

probably here. Okay. We are just trying to calculate. It again. The next step. What will happen again? This line will get created and finally you'll be able to see. Okay, so once we move from here to here, let me draw it once again for you so that you will be able to understand this points right. This is a super important thing. So I really want to draw it again. But if you understand the geometric intuition this is how the internal working will basically happen. So now my new points are over here, here, here, here, here, here, here here here here okay here. And this is my new centroid. Similarly to this you will be able to see I'll be able to see points like this. And the new centroid will probably be over here. Now here you can see that again. We'll start the second process by finding out whether there is a change in the points or not. So I'm just going to create two lines which will be perpendicular to each other. Now definitely you can see that now I'm clearly getting two groups. This is my one group and this is my another group. This is my one group and this is obviously my next group. Okay. So this is how we have clearly created two clusters. This is my cluster one and this is my cluster two. Okay. And obviously by just seeing this initially my data points was like this, right, with two groups. And obviously by following this three points, three steps, we were able to get this right. Now your question may be Chris. Fine. Uh, this was clear, right. This data points had just two groups. What about data points which may have three four groups. So how do we exactly select the k value. This k value is super super important okay. So uh understanding how to select the k value will try to do it in the next video. Uh but understand what are the steps. Initially some k centroids is basically initialized. It can be randomly initialized here and there. In our case we have initialized k is equal to two. Then which are the nearest points. Uh, we will just try to group that particular point, or we'll try to mark that point with respect to that specific centroid. And then we'll move the centroid by computing the average with respect to the new points. Now how do we know we have to stop over here. Because here we can definitely see there is no movement of points. Everything is correctly grouped when compared to the previous stage. Right. So here we are going to basically stop. And we will be clearly getting a two different clusters okay. So this particular steps is super super. The step that we will be performing in order to get the K-means clustering. And this is super easy. Now, the next thing that we are going to discuss, how do we select this k value. So let me write down the question over here itself. How do we select the k value. How do we. Select the k value. So let's go ahead and let's discuss this in the next video. Thank you.

So guys, now let's continue the discussion with respect to K-means clustering. And here we are going to answer this particular question. How do we select the k value. You already know that here in my previous problem statement I selected k is equal to two because from my eyes only I could see that okay, there are two different groups, but in the real world scenario you'll be finding a lot of overlap points. Right. So how do we select the k values. Now in order to

understand this. First of all I want to bring a new notation which is called as w cs w CS basically means within cluster sum of squares. Within cluster sum of squares. Okay. Now what does this basically mean? Okay, first of all, uh, in order to find out the k value, we will start with a process will initialize k value from one to some, some value, let's say 20. Let's, let's consider that we are going to take we are going to play with 20 different, uh, centroids uh, like in a at a time first. Initially we'll take one centroid. In my second iteration I'll take two centroids and all for as I initialize okay. Probably k to 1 to 20. We will just initialize this or we will just traverse this. Let's say in my data set initially I have this specific points okay. When my k is equal to one, that basically means I'm going to consider that many number of centroids okay. So let's say these are my data points. And obviously from this you can definitely see right now, if I consider k is equal to one, then what I'm going to do, I'm just going to initialize one centroid and it can be initialized anywhere okay. And we'll also be learning about a new initialization technique as we go ahead. Now as soon as I initialize with k is equal to one. In this particular case it k is equal to one. Now what will happen is that I will just go and compute the distance from every point to this particular centroid okay. So I'm just going to compute the distance from every point to this particular centroid. Okay. Once I compute it, you obviously know that this distance will be quite high. Okay. This distance will be quite high. Right. So let's say I computed this and this is my w x w x is nothing. But here I'm going to do the summation of one to n. And we are going to find out the distance between every point. Between every point. Between the points. Two. Nearest. Centroid okay. Nearest centroid right now since I had just k is equal to one. So this is my nearest centroid. So let's say obviously in this particular scenario. And this will be square okay within cluster sum of squares. So we are going to do the squaring part also okay. Now when I obviously this x value will be high. So what I'm going to do is that on the left hand side I'm just going to plot a graph. And this particular graph will be with respect to k value and x value. So some values will initialize over here. And obviously when the k value is one. When it is two. When it is three, we are just going to plot it. Okay for the first instance, obviously my value will be high when k is equal to one, right? So my k value when it is one, it will be very, very high. Now coming to the second scenario when my k value is equal to two. Now when my k value is equal to two, let's say in this particular scenario my k is equal to two, right. So for this particular data points I'm just going to consider two. This all data points over here. Same data points I'm just trying to mostly write in the same way. Now in this particular scenario what will happen. Two centroids will be initialized. Let's consider this is one centroid. And this is my another centroid okay. Now whichever points are nearer to this particular centroid will just try to calculate the distance. Now in this particular case let's say this point is nearest to this okay. So what we are going to do we are going to compute the distance right. So this distance this distance this distance we are just going to compute it. And finally we are going to do the summation right. And obviously you know that in this particular scenario all these points will be nearer to this right. Nearer to this. Now when I compute the distance here, obviously in this particular case I will again be getting some x value. But when compared to k is equal to one, k is equal to 2WX value will reduce. Why? Because there are two centroids. Initially I just had one centroid. If you see the distance right from all the points, it is obviously going to increase in this particular case since we are dividing this data points between two centroids. So if you go and compute the distance it is obviously going to reduce. So let's say with respect to k is equal to two. This point will come somewhere here. And when we continue to do this with k is equal to three k is equal to four. You will be seeing that my value will keep on decreasing like this. And after some point it will be almost stable. Right now when I combine all this sponge together. This point this way, this this is basically called as elbow method. Okay. This is called as elbow method. Now what does elbow method basically say that we have to find out a point wherein there is an abrupt decrease in

Wcss. And then after that it is going to stabilize. Let's say that in this particular case right here you can see there is an abrupt decrease. And after this it is almost becoming stable. Right. Just see that elbow. Right. You have probably seen our elbow. So in that elbow which will be the middle part right. That you can consider this specific part right where your wcss is abruptly decreasing. And after that specific point you can see that it is stable. So we have to go ahead and select that particular k value okay. So that is the concept of elbow method. And how do we select the k value. Okay, now, this is a super important technique. Uh, and in practicals also, I'll try to show you is that we will try to construct this elbow method wherein we'll play with different, different k values and wherever there is an abrupt decrease. And after that, when it is stabilizing, we are going to select that particular k value, because from here you'll be able to see that the distance between the within cluster sum of square is basically getting reduced. Okay. So this is how we basically find out the k value. Now let's discuss about two things. One is Euclidean distance. How do we compute the distance between two points in Euclidean distance? Okay. Now, in Euclidean distance, what we do is that suppose let's say there are two points. P1 and P2. Let's say there are two points, p1 and p2, and obviously this point is denoted by X1Y1 let's say. And this point is denoted by X2Y2. Okay, now if I really want to apply a Euclidean distance in order to find out the distance between these two points, the formula is super, super simple. We'll just compute by using the Euclidean distance formula that is root of x two minus x one plus y two minus y one whole square. Okay so this is what is the formula with respect to the Euclidean distance. Uh again let me write it down in a clear way so that you'll be able to understand it. So Euclidean distance is nothing but. The formula is root of. X two minus x one whole square plus y two minus y one whole square. Right? Suppose if it is a three dimensional point, then we have to continue writing like z two minus z one also. Right. So this is basically the Euclidean distance. Now in case of Manhattan distance what we do is that. In case of Manhattan distance, you'll be seeing that our points will be like this. Right now, we have to compute the distance from here to here and then here to here. Okay. So in this particular way we will be able to compute the Manhattan distance okay. So suppose if this is a b c I will probably be taking a b plus b c okay. By that specific way we'll be able to compute it. Now here you can see that if I probably plot this point over here. Okay, so this is basically my X1. This is basically my x2. Okay. Similarly this is my y1 and this is my y2 right. So if I really want to apply to find out this a b a b is nothing, but it is nothing but s2 minus x1. And this bc is nothing but y2 minus y1. So the overall formula with respect to Manhattan distance can be written by absolute value of x2 minus x1 plus y2 minus y1. So if this use this particular formula we will be able to compute the Manhattan distance. Now when should we use Euclidean distance or Manhattan distance? I think everybody has seen Iron Man movie right. So in Iron Man movie, uh, in us. Right. Consider countries, many of the states in many of the states in US write in us. What happens is that your cities, how they are planned, they are planned in blocks, right? Like this. They are planned in blocks. So here you will be having buildings in the middle, you will be having road and all the other things. Right. So let's say if if your entire city is planned like this, I want to go from this particular point. Let's say I want to hire a Uber from this particular point to this particular point, then how do I go? I cannot go directly like this, right? Obviously there will be buildings. So usually the technique that I follow is basically to calculate the distance will be used where as Manhattan distance, Manhattan distance. Right now similarly in in the case of Euclidean, let's say uh, in a traffic control, right. In a traffic control, if I probably say I want to go from point A to point B, obviously in this particular case, I'll not be going like this. Instead, I can go directly like this. So based on this the distance will get computed. Uh, computed. So in this particular case I can definitely use something called as Euclidean distance okay. Euclidean distance. So this is the basic difference between Euclidean and Manhattan distance. And I've

also showed you how to initialize the k value. Uh over here okay. Now as we go ahead, uh we are going to understand one very important thing. Because see, at every k value we initialize this points okay. What if the initialization is done in such a way that both the points are very near to each other? Let's say if I have two centroids that are very much near to each other, if it is initialized in this particular way, then what may happen? What is the problem that you may face? That is what we are going to discuss in the next video. But in this video we have discussed about how like how do you find out this k value and uh, what is the importance or what is the meaning of Euclidean distance and Manhattan distance. So yes, I will see you all in the next video.

So in this video I'm going to discuss about something called as random initialization trap in K-means clustering. And uh, there's a initialization technique which is called as k means plus plus. Okay. So this is what we are going to discuss about this. Let's say guys I have this kind of data points, you know, and according to you probably if we apply K-means clustering, you know, we should be able to get this kind of, this kind of clustered groups. And obviously it looks correct because all the groups are over here. And here you can see some of the data points group is here. So definitely three clusters I'm able to get with k is equal to three. Right now let's say that what if and obviously you know that whenever we initialize centroids I told you that okay we randomly initialize some of the centroids in this particular space. What if that random initialization goes wrong. You know. So in that particular way let's let's consider that we are just going to initialize two centroids that may be very near to each other, that may be very near to each other. And one centroid which may be over here somewhere here. Now in this particular scenario, you will be able to see that even though in this scenario, if I try to see all the nearest point, you know, you'll be able to see that. Okay, perfectly. We are able to get all the points that I initialized right. So this will be one group, this will be the other group. And this entire thing will be the other group. Because these are the points that are near to this. Right. So in this particular scenario, when this kind of initialization of the centroid happens, you know, the final output that you will probably be able to see, is that how you'll be getting the output over here. So in this case you'll be seeing that these are my points. Uh, and uh obviously here my yellow centroid is there. So this will be created as an yellow group. And uh here you have other points. Right. Other points with this orange one. And obviously this will be created as another group over here. And finally, all these points that you see on the right hand side, right. You will be able to see that there will be one point in between and. But all this will be grouped together and this group will be basically displayed in another, another, another cluster. Now see from the initial right we we directly from our eyes we could see that okay, they are specific two clusters on the right hand side. And this should definitely be belonging to one cluster. But since my initialization my initialization of the sand uh, of the centroids, you know, it did not happen properly, it was randomly initialized, you know, so some points got here, some points came over here. And this

was my third centroid. This is my centroid one. Centroid two. Centroid three. In this kind of initialization you could see that my output is coming like this right. And obviously from the algorithm that we have learned yes this clustering looks fine. Yes we are able to group together. But understand what was the main thing we had to get in this particular way. But because of the initialization technique of the centroids gone wrong, you know, it was randomly used, right? We used random random initialization, right? We use this initialization technique. And we got one centroid over here, one centroid over here one centroid. And because of this we got this specific output which is definitely wrong. Right. Because this is how you had to actually get. So in this particular scenario we basically say that this is called as random initialization trap. So there are some problems. Whenever you randomly initialize all the centroids you may be getting a different cluster of groups. You know that may not work well with respect to your data set. So for this particular case what we do is that we. Use k means plus plus initialization technique. Now you may be thinking what exactly is this? K means initialization technique. K means plus plus initialization technique. Now in this technique, what we do is that we initialize all the centroids in such a way that at least it should be at max distance that it can write. Let's say this if if this particular point is initialized away, if this centroid is initialized away, the next centroid that is getting initialized is quite far from the previous centroid. Similarly, the next centroid. Suppose if I'm considering k is equal to three, then the next centroid that will be initialized will be quite far from there. Right in this sample space the centroids that are going to get initialized it will be initialized in a far away. Right. That is what this random K-means plus plus will basically do. And in practical this I'll try to show you, uh, you know, and random initialization. It will not happen for each and every data set. Yes, but for some of the data sets. Suppose this kind of initialization basically has happened at that point of time. We will be. Right? So definitely try to use k means plus plus initialization technique whenever you're using k means clustering. And again if someone asks you in an interview why specifically it is used, it is just to initialize the centroid in such a way that it is completely at a farther distance between each other. So after this, when we apply all this techniques right, we will be able to get this kind of clustering techniques. Basically the cluster groups. Right. So I hope you like this specific video. Uh, in this video we have discussed about the random initialization trap. And in our previous video we have understood about Euclidean distance. We have understood how to find out the k value. And finally, we also understood that how the working of the k means clustering will work, right. And we will try to see more practical implementation as we go ahead. So yes, I will see you all in the next video. Thank you.