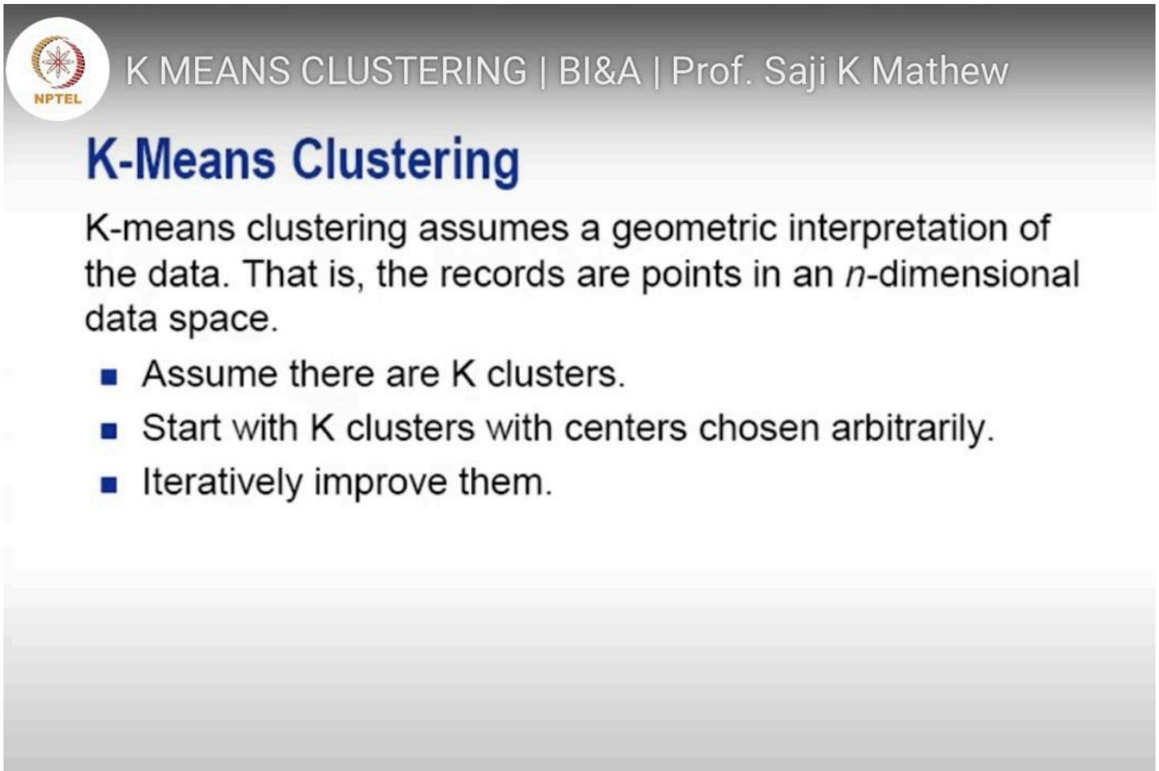
Course Name:Business Intelligence and Analytics
Professor Name:Prof. Saji.K.Mathew
Department Name:Department of Management Studies
Institute Name:Indian Institute of Technology Madras
Week:09
Lecture:35

K MEANS CLUSTERING



K MEANS CLUSTERING | BI&A | Prof. Saji K Mathew

K-Means Clustering

K-means clustering assumes a geometric interpretation of the data. That is, the records are points in an *n*-dimensional data space.

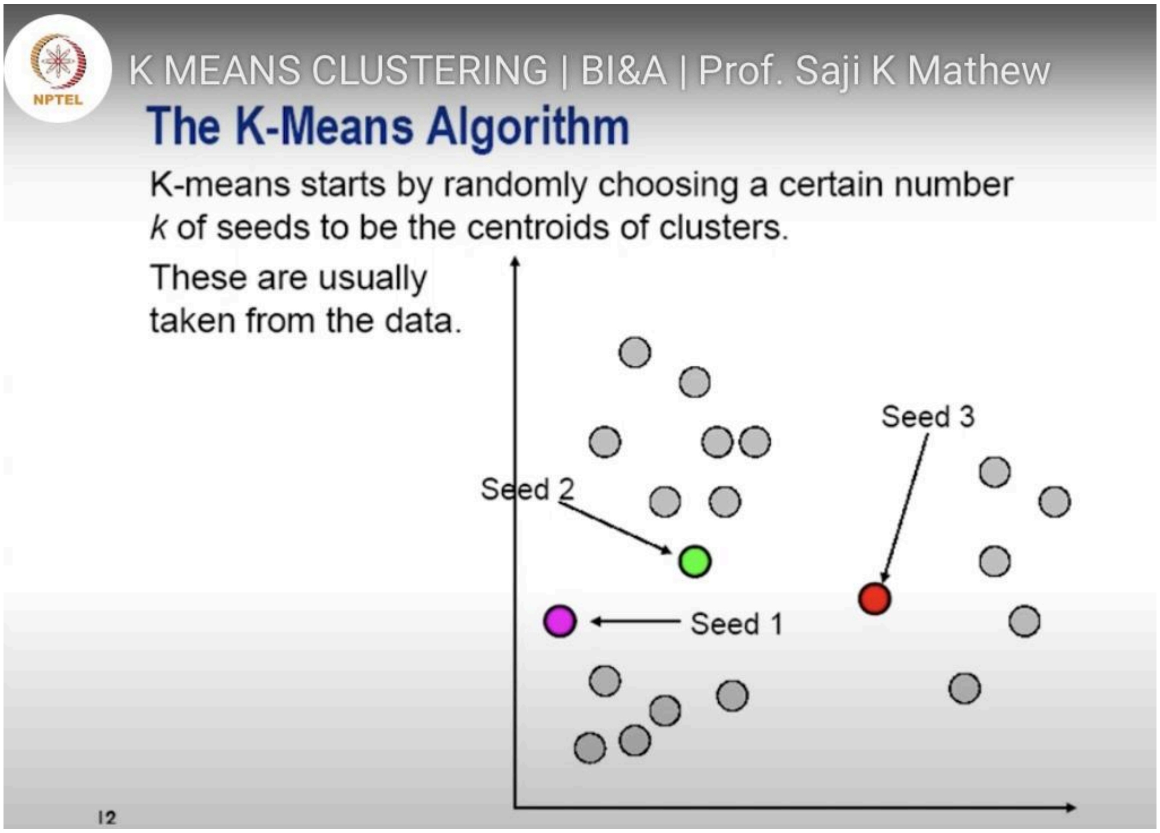
- Assume there are K clusters.
- Start with K clusters with centers chosen arbitrarily.
- Iteratively improve them.

Okay, first question what is K? K is the number of clusters. So in K means you always start with the number of clusters. You have to input how many clusters you want to start with. And as in the case of agglomerative clustering, that is something that we decided at the end. Therefore you do not have to specify that number to start with. But in partitioning or in K means you have to specify the number of clusters.

And therefore how would you start with? How will you start? How do you know what should be the number of clusters? That is a valid question. So you can keep speculating K should be 3, 4, 5, 6, 7, 8, 9, 10. But that is not fair. There has to be some method to calculate or to estimate the number of clusters.

So we will see that. But assume there are K clusters. Start with K clusters with centres chosen arbitrarily. So there is a random seed. We call it random seeds to start with.

Random seeds are random points in the space. Euclidean space or multidimensional space depending on the number of attributes. You start with those random points and then you follow a method to assign each object to those seeds and iteratively improve the assignment, so that clusters become so low in errors or within cluster distances are minimized. So you follow an iterative process for that. So I will explain the K-means algorithm graphically in the subsequent slides.



K MEANS CLUSTERING | BI&A | Prof. Saji K Mathew

The K-Means Algorithm

K-means starts by randomly choosing a certain number *k* of seeds to be the centroids of clusters.

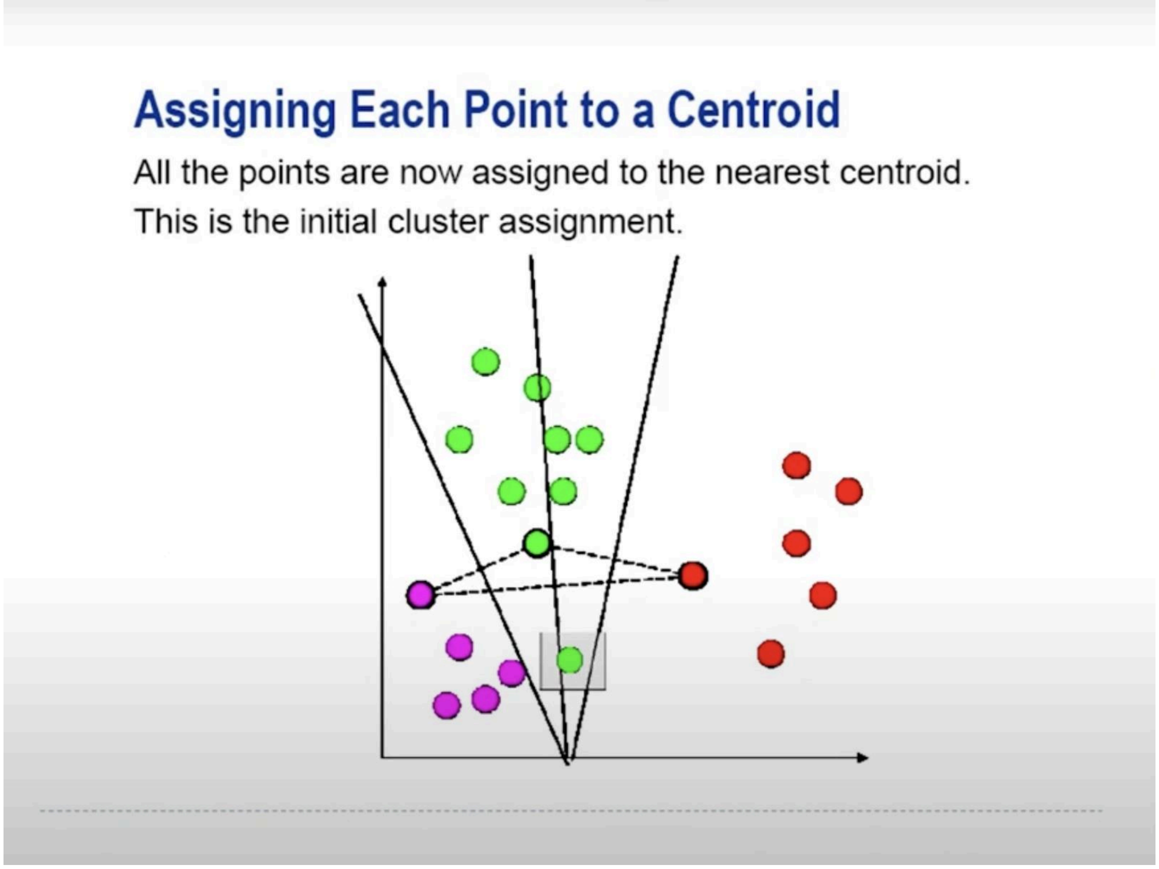
These are usually taken from the data.

You can write it in textual form which is followed, which can be followed, which is there in your textbook. I am trying to explain it graphically. So K-means algorithm, this is a different textbook. So you may benefit from this because it gives you easier visualization of the K-means algorithm. So as I said, assuming that there are only two variables and it is a two dimensional space and you have n number of objects, you can count the number of objects here.

And the starting point is three random seeds. You are saying K equals 3. That is my assumption. I want three clusters. For some reason I assume that, my assumption.

And which are my assumed starting points? There is one point here. What is this color? Magenta and this is blue or green. This is a green, red, m. Let me call it m.

M-G-R. These are random seeds in the Euclidean space. I start there. So you can also assume that if two people do K-means, their solutions can be different because your starting point is different. So it is based on these starting points, you start to continuously iterate going for minimization of total errors and finally arrive at some solution. So what is done is, of course select this random seeds and go to the next slide.



Assigning Each Point to a Centroid

All the points are now assigned to the nearest centroid.

This is the initial cluster assignment.

Here you see the next step in the algorithm. The next step in the algorithm is to assign each object to one seed. Keep in mind if I give you a data set in a spreadsheet, you must be able to do this assignment using, by calculating centroids and continuously doing the assignment. It is an important exercise that you should familiarize yourself with.

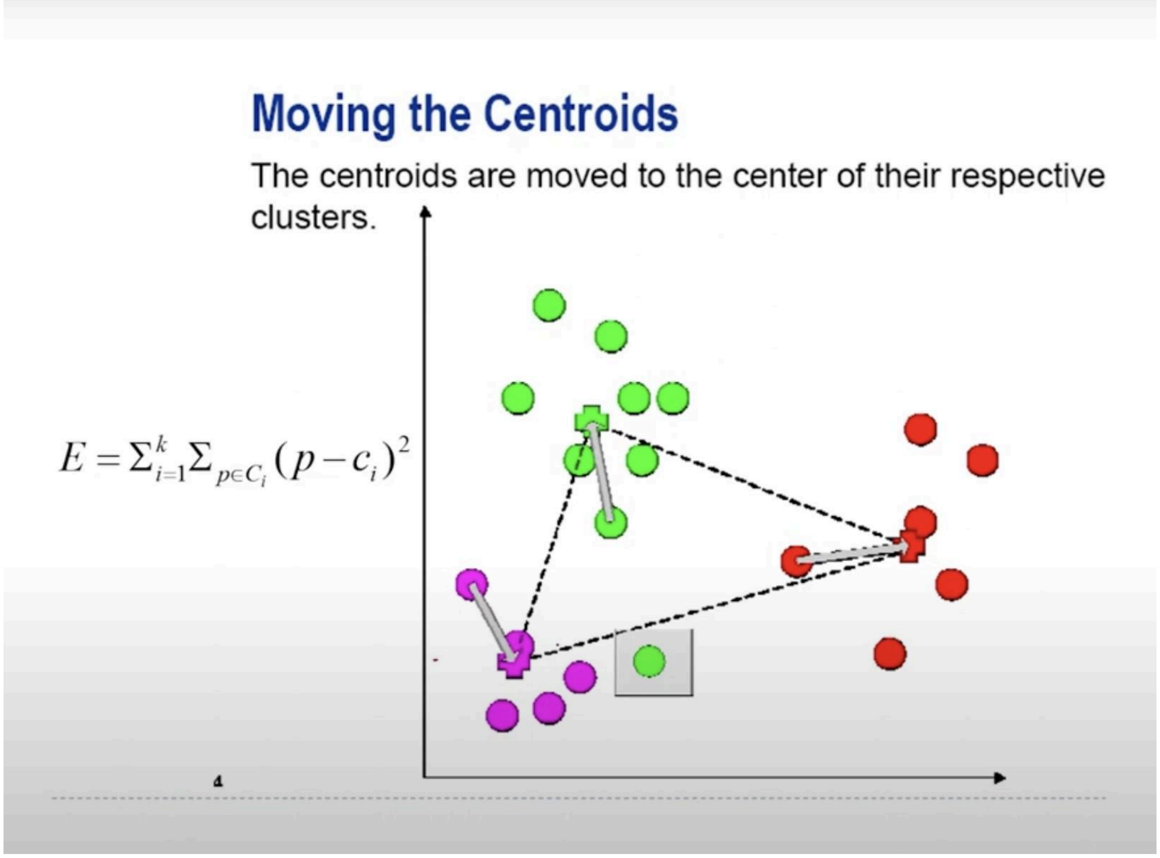
I am explaining that here. So you have the random seeds. So and, what is done is you fixed those seeds here and here. And then what you do you, calculate the distance of each object to the, each object to each of the three random seeds. For example, there is one object here.

I calculate the distance of this object to this. I calculate the distance of this object to this seed. I also calculate the distance of this. Now where does this guy belong? This guy belongs to that seed with the shortest distance. That is where I belong.

That is where the distance is the shortest or that is where the similarity is the highest. And therefore, visually you can see that this guy should belong to this point. This is the shortest distance. And similarly you assign each object to that seed where the distance is the shortest. That is the first iteration, first level of assignment.

Each object gets assigned to one of the seeds to which it is closest. And then you see, you notice one interesting point here. That is, there is an object here, this guy. Usually because the seeds are placed here, when you look at the distances, distance to this, distance to this and distance to, sorry, distance to this. We actually calculate, it appears that this is a green object because this distance is the shortest.

So this fellow got assigned to this group, that is the green group and not the magenta group. Although visually you will see that this guy belong here. Now, the next step is this assignments are done in the first step. Second, you calculate the centroids of each group. Now, there are three clusters formed.



Moving the Centroids

The centroids are moved to the center of their respective clusters.

$$E = \Sigma_{i=1}^{k} \Sigma_{p \in C_i} (p - c_i)^2$$

Calculate the centroids. The centroids is not the same as the seeds. Centroids got shifted here or now the centroids are here. These were the initial random points but when the assignment happened, groups got formed. Now you find the centroids of the groups.

That is the central point of the cluster. That is the next step. And now also, there is a scope for reassignment now. Now calculate the distance of each object to the three centroids. Calculate the distance of each object to the three centroids.

Now for example, take this guy who was assigned to the green group. Calculate the distance of the centroids. Which is the shortest distance? This fellow actually is closest to the magenta group and therefore there is a scope for reassignment. So in the second iteration, you reassign each object based on distance to centroids. And therefore, he changes group, from one party to another.

From AIADMK to DMK, or you know MGR to something else. Just for fun. So the, based on distance you actually change or reassign each object to another group. Now it does not end here. At this stage, now again since the reassignment happened.

This is a very simple case where you say that one person changed party or group or cluster. Now what happens? The centroids change again. Because reassignment happened, the centroids again change. So you recalculate the centroid. When you recompute the centroid again, there is a chance that some objects would change the clusters because the distances again change.

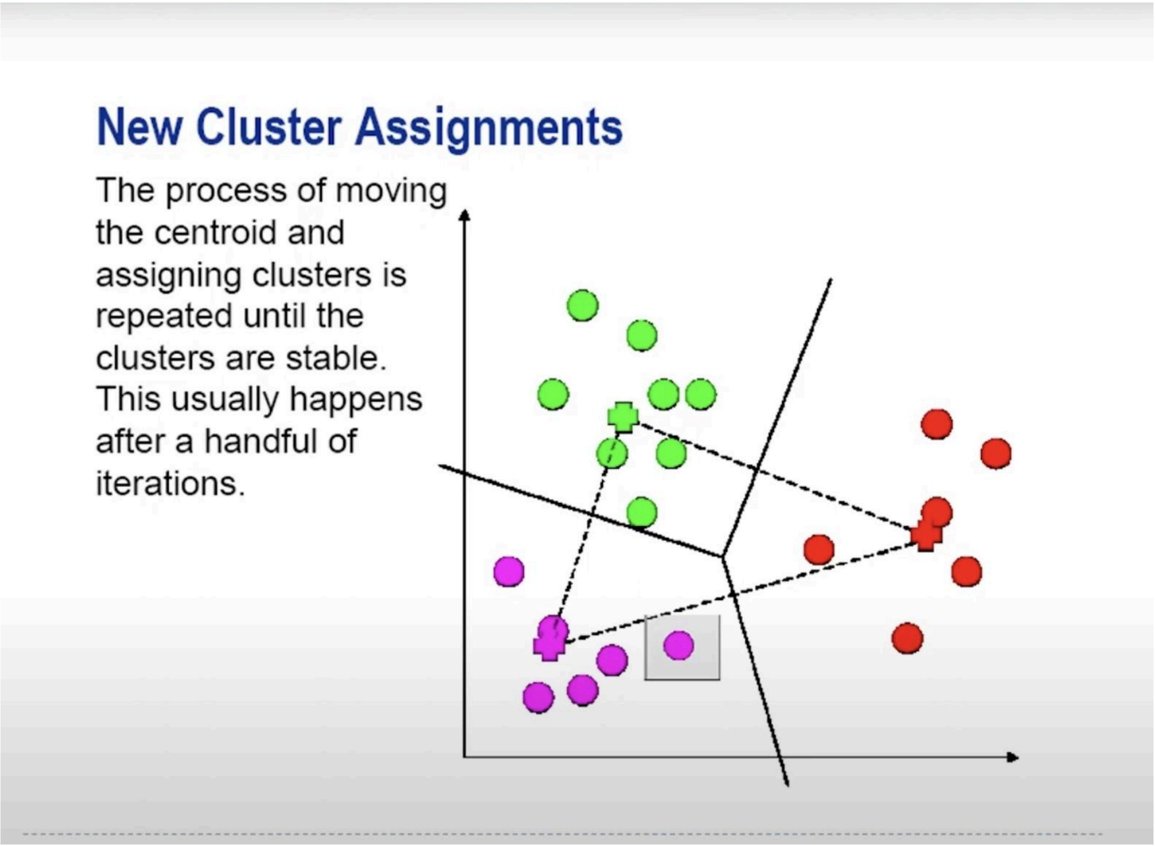
So you see the iteration going on. So in each iteration, you compute the total E is the total error or total within cluster distance. The function, the purpose is to minimize this. Minimize total within cluster distance and you expect that the error should be, if this is the error, error should be falling with each iteration.

That is what you expect. And the minimization function is the total within cluster distance. What is c_i? Suppose there are K clusters, for each cluster you calculate the total within cluster distance which is the distance of each object to the centroid. Distance of each object to the centroid, you sum it up for each cluster and you sum it for all the clusters, you get the total within cluster distance. So the idea is, through each of the iterations, the solution should be converging or it should be, the error should be falling till it reaches an acceptable value called the minimum error acceptable. You know in an algorithm, you can set a minimum value for error.

You can also set a minimum value for tolerance. Tolerance is the incremental change in error that happens between iterations. So these are values or hyper parameters you can tune within an algorithm, within a clustering algorithm. So why does it need to be squared? Squared distance. Squared, square has certain mathematical properties for minimization, that is for minimization, that is one explanation and that is the only explanation I can use, squared error is used.

Otherwise you can use a sum of absolute errors. But for a minimization algorithm, you typically use the squared error. So for this particular cluster, the p-c_i would be the distance of this cluster to distance this one, this one, all squared. That would be the sum of squares for one cluster, one cluster K equals 1 for K equals 2 you go here,find the total error for K equals 3, you go here find the total error with respect to the centroid.

Sum that up and that is the total error. You are breaking down, you are actually going to, you are looking at whether the features, each feature would be, you know what data type. That challenge would be there, so in terms we have to establish some way of calculating or transforming the data type into one standardized format and then do this. This can work only if it is continuous valued, for interval, ordinal and continuous value data this is a simple formula. Otherwise it becomes a challenge. That is the detailed level of computation, but essentially at the level I am explaining it calculates the total error with respect to, total squared error with respect to centroid.



New Cluster Assignments

The process of moving the centroid and assigning clusters is repeated until the clusters are stable. This usually happens after a handful of iterations.

Okay, so we have seen this method goes on until the final solution converges. Now there is a challenge here, that is with respect to determining the number of clusters, we talked about it right. What is the value of K? K as I said, there is a practical consideration you ask the manager or your project sponsor, how many number of clusters would be acceptable? That is a practical consideration. Then there are heuristic techniques. One heuristic is, how many number of clusters, if n is the number of number of objects or number of records, square root of n by 2 is one suggested number of clusters.

And there should be a square root of 2 n data points in each cluster, that is a minimum cluster size. This is heuristics, one recommendation. This may conflict with the practical consideration. So you know, these are sort of suggestions. Another widely used method for determining the value of K is the ELBOW method.

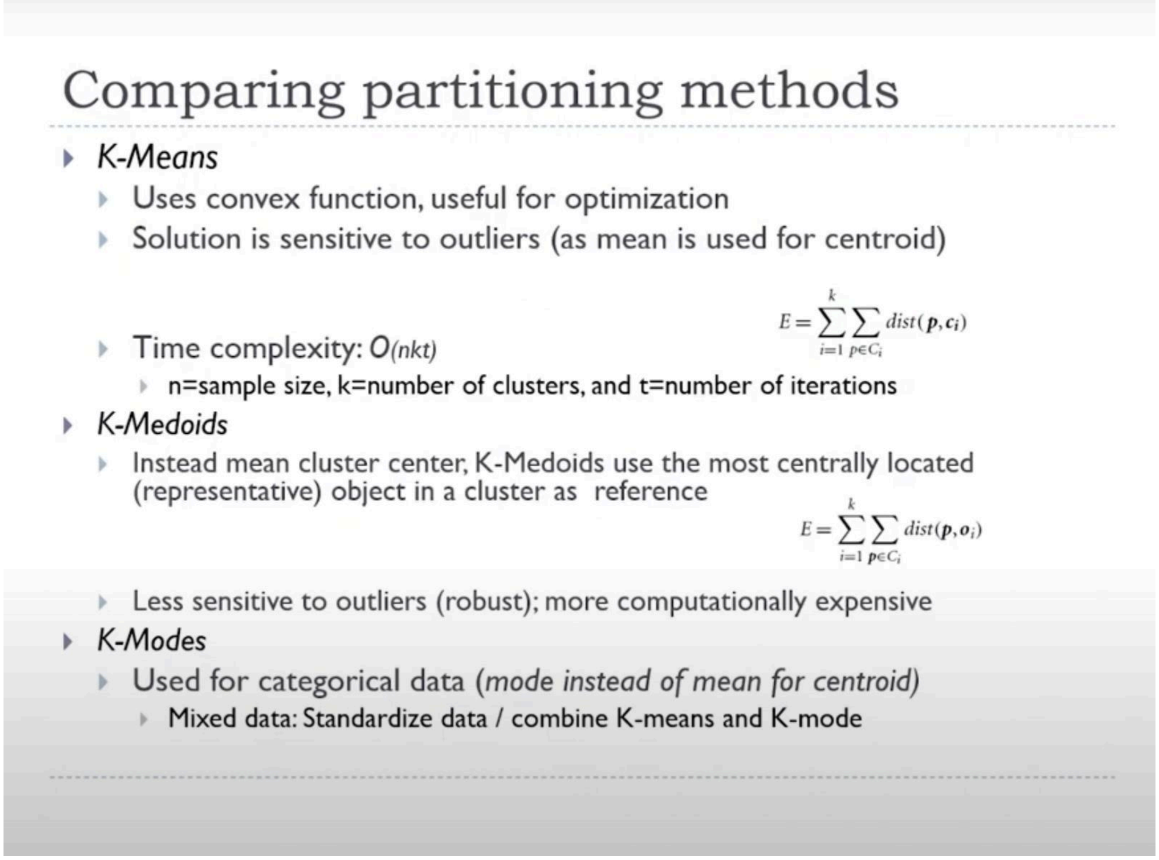
ELBOW method is very intuitive and very useful because the researcher can choose the number of clusters as a range also. You can also, you know try out a 3 or a 4 or a 5 and see how good the cluster solution is and decide. But where to choose that range from? It is something that can be determined using a ELBOW method. The ELBOW method, the x axis is the number of clusters and the y axis is nothing but the error. The error can be the total within cluster distance.

Now what is done here is you employ one algorithm, one clustering algorithm, it could be the K-means algorithm. You try you run K-means with two clusters, three clusters, four clusters, five clusters, six clusters and so on. And plot this graph, the total within cluster distance versus number of clusters. And then you find that after some stage the error actually does not change much.

It is like the ELBOW. There is a steep fall here then it becomes almost same, it does not improve much. So therefore the ELBOW is the point that one should fix as the optimum number of clusters. So that way some graphs may give an exact ELBOW like this one 4. But sometimes you do not find a clear ELBOW.

Sometimes the hand may be little stretched and so on. So therefore you can try, if I am working on this problem I will not just go by this recommendation of 4 alone. I would try this one. I would try a 5 also or a 6 also and see how interpretable the cluster is. A cluster should be interpretable and useful.

It is for a purpose and I would choose that solution. I will combine an ELBOW method with a practical consideration. And that is where, sort of you can fix the number of clusters. You can decide the number of clusters. Now, since partitioning techniques are widely used, there are different choices of clustering techniques within the partitioning method.



Comparing partitioning methods

- *K-Means*
  - Uses convex function, useful for optimization
  - Solution is sensitive to outliers (as mean is used for centroid)

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(\boldsymbol{p}, \boldsymbol{c}_i)$$

  - Time complexity: *O(nkt)*
    - n=sample size, k=number of clusters, and t=number of iterations
- *K-Medoids*
  - Instead mean cluster center, K-Medoids use the most centrally located (representative) object in a cluster as reference

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(\boldsymbol{p}, \boldsymbol{o}_i)$$

  - Less sensitive to outliers (robust); more computationally expensive
- *K-Modes*
  - Used for categorical data (*mode instead of mean for centroid)*
    - Mixed data: Standardize data / combine K-means and K-mode

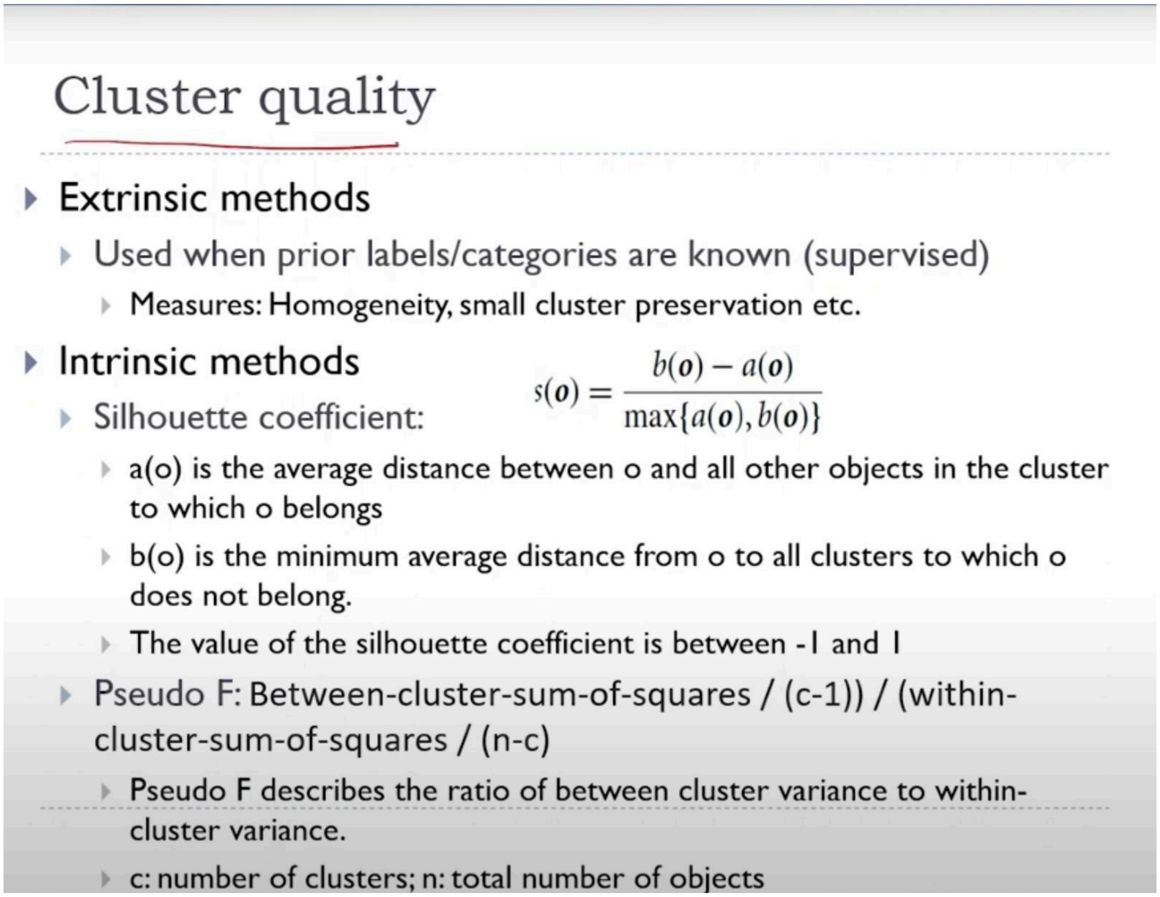
One is of course the K-means which we discussed already. And there is something called time complexity or cost of computation that is involved in an algorithm which is minimization algorithm. And time complexity is given by sample size, number of clusters and number of iterations that determine the time complexity. And one challenge with the K-means, since it is actually for centroid, it is averaging. The centroid is a average of the feature values and therefore solution is sensitive to outliers. A centroid can lie at one corner of the cluster if there are actually outliers.

So therefore you have the K-medoid. I think I wrongly printed it as m-medoid. It is K-medoid. Medoid uses not the mod. Please see the difference between medoid and K-modes.

K-medoid is different from K-mod. In K-medoid you find the center value, not center value, the most central object. Suppose there are, say q number of objects within a cluster. You find out that object which is most centrally located. Instead of just going for the centroid which is average value.

The mean can be influenced by the outliers. Instead of that, use the most central object as the reference point to calculate errors. And then use that, that is the K-medoid technique. And this solution you expect to be better than the K-means if you have, if you are aware of too many outliers. That is second technique.

And K-mode works for data when the data are categorical. You can use the mode instead of the mean. So that is the K-modes technique. So K-means, K-medoids and K-modes are variants of the partitioning technique. Depending on the data type, that is K-means and K-mode.



Cluster quality

- Extrinsic methods
  - Used when prior labels/categories are known (supervised)
    - Measures: Homogeneity, small cluster preservation etc.
- Intrinsic methods
  - Silhouette coefficient: $s(\boldsymbol{o}) = \dfrac{b(\boldsymbol{o}) - a(\boldsymbol{o})}{\max\{a(\boldsymbol{o}), b(\boldsymbol{o})\}}$
    - a(o) is the average distance between o and all other objects in the cluster to which o belongs
    - b(o) is the minimum average distance from o to all clusters to which o does not belong.
    - The value of the silhouette coefficient is between -1 and 1
  - Pseudo F: Between-cluster-sum-of-squares / (c-1)) / (within-cluster-sum-of-squares / (n-c)
    - Pseudo F describes the ratio of between cluster variance to within-cluster variance.
    - c: number of clusters; n: total number of objects

And depending on the reference point for calculating errors. So K-medoid is an improvement on K-means. So cluster quality. We will, the last 10 minutes we will try to have problem. But we will try today or we will do it in the next session.

It is better to understand concepts well and then go to a problem. So cluster quality. At the end what is the quality of the cluster that you obtain. Of course, you can look at the errors and then speculate about the cluster is good or not good. But it is better to have a ratio always. To comment on something, instead of going by some absolute value of error.

So there are certain ratios that are available known as, one is known as the Silhouette coefficient other is known as the F, pseudo F. Pseudo F and Silhouette coefficients are widely used to report quality of clusters that were formed. Now the extrinsic methods which is given in your textbook, refers to those methods where the cluster classes or the labels are already given, like that is more close to a decision tree kind of exercise where the class labels are not.

In most cases, when you use clustering that there is no label. The clusters get formed based on closest distances, you have to label them. In our exercise you will see, you are the person, you as researcher or analyst would finally give a name to a cluster. This is a Punjabi cluster for example or this is DMK, AIADMK or this is BJP or Congress or whatever party you know. You name them based on their characteristics or their affinity or affiliation.

There are attributes that we capture, group and then name them. So, take it all in a lighter sense. Do not take it too hard. This all, you know just how we form social groups.

That is also based on certain social attributes. All right. Now so what is the Silhouette distance? It is clearly explained taken just from Hahn and Kamber. So, suppose a(o) is the average distance between o, which is an object o and all other objects in the cluster to which o belongs. And b(o) is the minimum average distance from o to all clusters to which o does not belong. The value of Silhouette distance is given by b(o) - a(o) divided by maximum of a(o) or b(o). But intuitively what is this formula trying to tell you? There are three clusters and there are objects within clusters.

o is here. The fellow o is here. o belongs here. So what is a(o)? Is the average distance between o and all other objects in the cluster to which o belongs. o's distance to other, his friends, his or her friends. All his friends are in that group. What is the distance of me to my friends? That is a(o). But what is b(o)? What is the distance of me to my rivals, if it is political party? So there are people that belong to other parties.

What is my distance to other objects? So o's distance to all other objects. That is what b(o) is. So that is, what you desire is o should be close to his or her friends. But o should be far from the enemies or rivals in a competition or in that scenario. Your clustering has to be taken that way, in the sense the cluster should be close, cluster object should be close within but they should be very different.

They should discriminate from other clusters. So it is a measure of how well a cluster discriminates versus how well a cluster is homogeneous. So homogeneity versus heterogeneity. So b(o) is a measure of heterogeneity across clusters. The distance of o from other, a(o) is a measure of homogeneity within that group. And therefore what do you want? What should be maximum? a(o) should be maximum or b(o) should be maximum? There you fail the test.

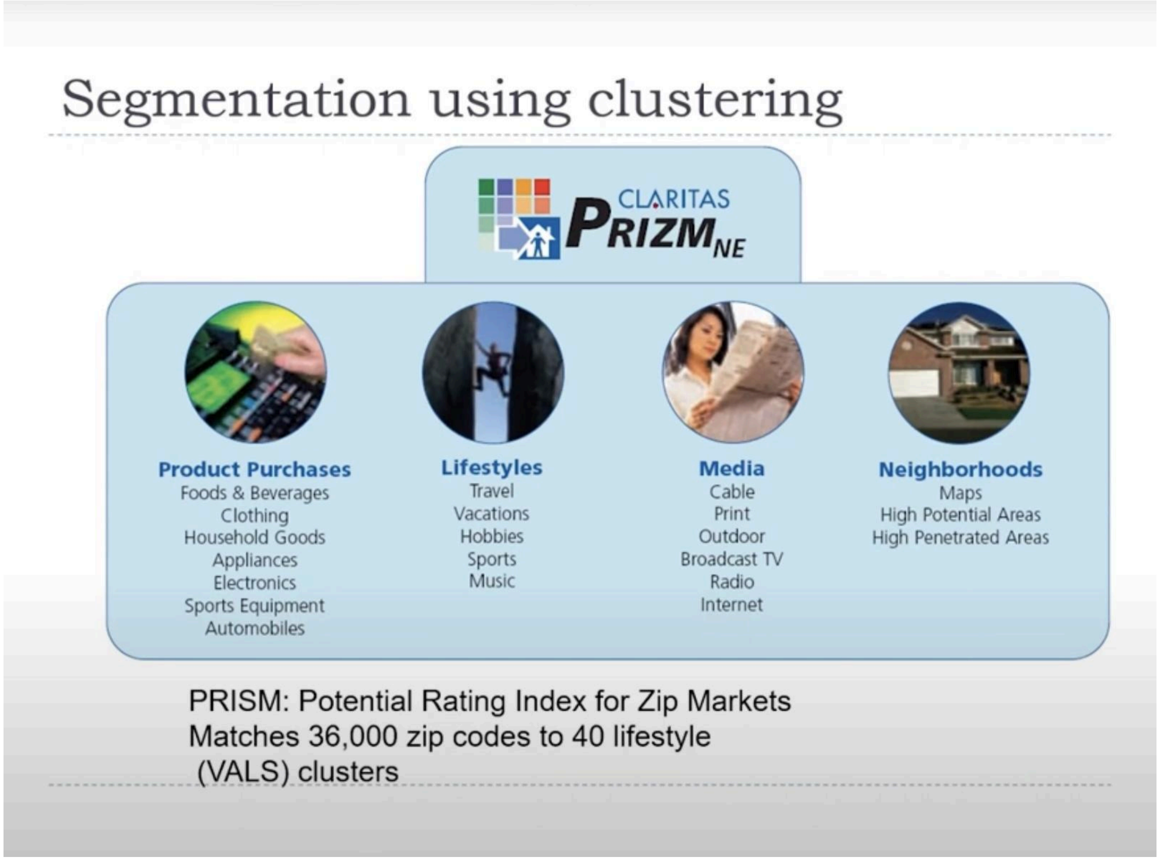
You want b(o) the error should be maximum when I consider my distance from others. So b(o) should be maximum. And b(o) should be, when b(o) is maximum, a(o) should be minimum. At that points, the ratio will become 1. But in a scenario where a(o), b(o) is not maximum, a(o) is maximum then b(o) is small. b(o) is small. Therefore denominator becomes very high and b(o) is small. Therefore this will have a negative value. Therefore it will tend to -1. That is what the range is going to be. Silhouette distance will have a range of this kind.

So it gives you a sense of homogeneity and heterogeneity. Heterogeneity across and homogeneity within. Pseudo F is another ratio. Look at that ratio. It is very intuitive.

Between cluster sum of squares divided by within cluster sum of squares. And like degrees of freedom, it is also trying to consider number of clusters. So simply look at that ratio between cluster sum of squares divided by within cluster sum of squares. It is actually again similar to what Silhouette distance is trying to do.

Between cluster distance should be high. Within cluster distance should be low. And therefore you want an F value which should be high. A high, higher F value is desirable. The denominator should be low and the numerator should be high. Just like the F ratio in your statistical test.

And this gets, sort of moderated by the C-1 divided, N- C divided by C-1 ratio. For example when the number of clusters is equal to the number of objects, what happens to this ratio? What N-C becomes? C is equal to N. So it becomes 0 or F value will become 0. So therefore, it does you, know it is for number of clusters, it moderates the value. So it should be C should be much lower than N.



Segmentation using clustering

Product Purchases
Foods & Beverages
Clothing
Household Goods
Appliances
Electronics
Sports Equipment
Automobiles

Lifestyles
Travel
Vacations
Hobbies
Sports
Music

Media
Cable
Print
Outdoor
Broadcast TV
Radio
Internet

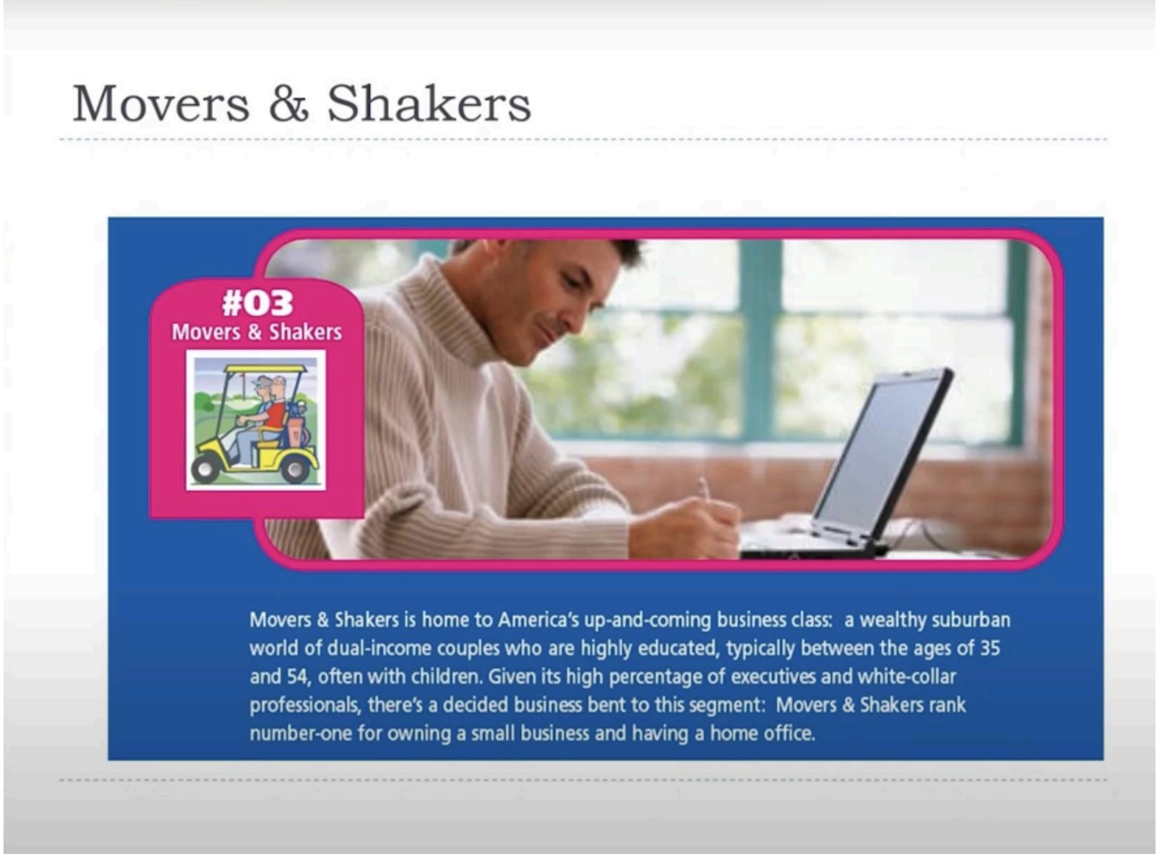
Neighborhoods
Maps
High Potential Areas
High Penetrated Areas

PRISM: Potential Rating Index for Zip Markets
Matches 36,000 zip codes to 40 lifestyle
(VALS) clusters

Only then this ratio would give a good value or reasonable value. So that is the measure of F or pseudo F in cluster, in assessing cluster quality. Two measures Silhouette distance and pseudo F to measure the cluster quality at the end. And if you are doing a clustering exercise, you should report this cluster quality measures also. At the end since you are students of business, when you do a cluster analysis it is not for satisfying yourself with how you have done a nice data analysis, you know nice values of F and Silhouette distance and so on.

Close your project. Now your project actually starts here, in analytics. You did this exercise to present your solution to a project sponsorer . Project sponsorer has sponsored your project for solving some business problem. So you have to take it to the sponsorer and satisfy yourself with the measures of quality, but present it for practical application.

So segmentation using clustering. In such exercises, cluster profiling, profiling a cluster is a very important exercise. You describe the cluster. If you formed five clusters what are the clusters known for? This is an example where PRIZM, Claritas PRIZM which is a professional agency which does segmentation in the United States and maybe other countries too using zip codes and lifestyle attributes.



Movers & Shakers

Movers & Shakers is home to America's up-and-coming business class: a wealthy suburban world of dual-income couples who are highly educated, typically between the ages of 35 and 54, often with children. Given its high percentage of executives and white-collar professionals, there's a decided business bent to this segment: Movers & Shakers rank number-one for owning a small business and having a home office.

They form geographical clusters. See how they do this job and how they sell their product. Their clustering product is sold as well profiled clusters, geographical clusters. You see what they do? Movers and shakers. This is class label or cluster label. Movers and shakers, looking at the characteristic of a cluster they give a name.

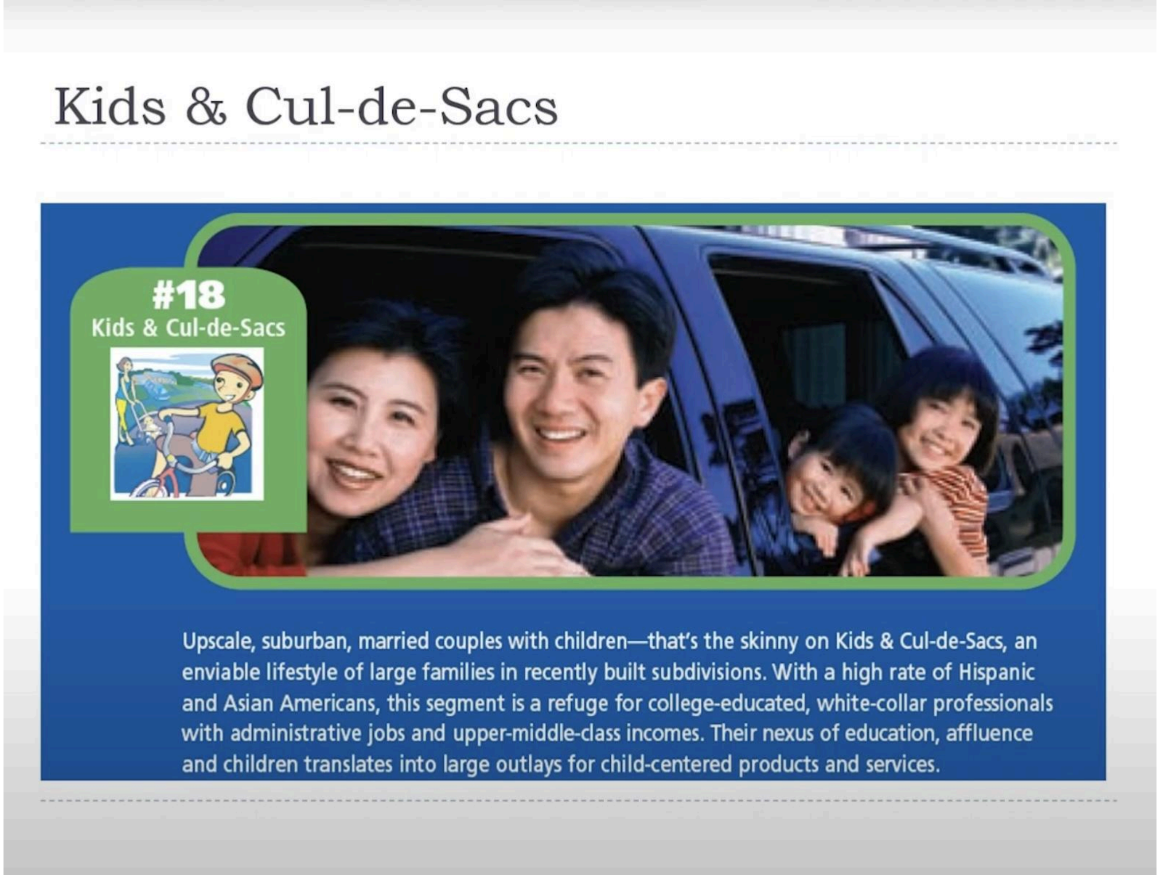
This cluster can be named as movers and shakers. In decision trees you know that the cluster labels or the class labels were given, like buyers and non buyers. Here there is no label. You look at the cluster characteristics well, this cluster consists of people who have these characteristics high. How do they describe the cluster? You see it is for application.

Movers and shakers is a cluster, is home to America's up and coming business class. Therefore the business class attribute is having high value here, whichever attribute they used for that. Wealthy, they have used income as a attribute. And this cluster has average income which is high. You have to look at the each cluster and compare based on some measure of average.

And comparatively this is a wealthy group. They are suburban. Dual income, all are attributes. Dual income couples. I think this is becoming common in our country also.

Dual income couples. So they have used that as an attribute. Highly educated. So education is one attribute. And it is high because average value is high. Typically between ages. Age bracket is another attribute. But you see the, how they position it for a user or a decision maker.

When you cluster, got good solution, go for a presentation, take something of this kind. Your back end analysis is with you. You have control on that. But in presentation, look at the cluster profiles and describe them in a way they can appreciate it for application. Well, is this cluster useful for me or is this segment useful for me? Depends what product or service I have. And they would make that recommendation as you know as consultants.



Kids & Cul-de-Sacs

#18
Kids & Cul-de-Sacs

Upscale, suburban, married couples with children—that's the skinny on Kids & Cul-de-Sacs, an enviable lifestyle of large families in recently built subdivisions. With a high rate of Hispanic and Asian Americans, this segment is a refuge for college-educated, white-collar professionals with administrative jobs and upper-middle-class incomes. Their nexus of education, affluence and children translates into large outlays for child-centered products and services.

Another group or another cluster which they have profiled and you know name is given by them. PRIZM Claritas. That is a clustering exercise for analytics. Clustering for classroom work, you can do up till your F value and silhouette distance.

But if you have to make a presentation or make it useful for analytics, then you also have to look at the profile. So in our example which I am going to use in the class, I will try to profile the clusters using graph so that you get to understand what each cluster is and then subsequently you can name the clusters and then describe them using language.