

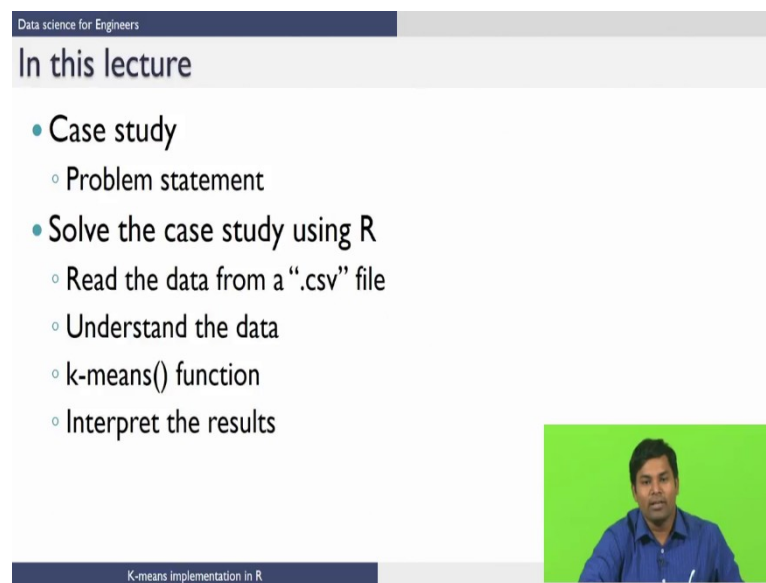
**Data science for Engineers**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Madras**

**Lecture - 49**  
**K-means implementation in R**

Welcome to this lecture on Implementation of K-means Algorithm in R. In the previous lectures Professor Raghu would have given a brief introduction about this K-means clustering algorithm and the mathematical details of how this K-means algorithm works.

In this lecture what we are going to do is to introduce you to a case study which we use as a means to explain how K-means algorithm can be implemented in R.

(Refer Slide Time: 00:34)

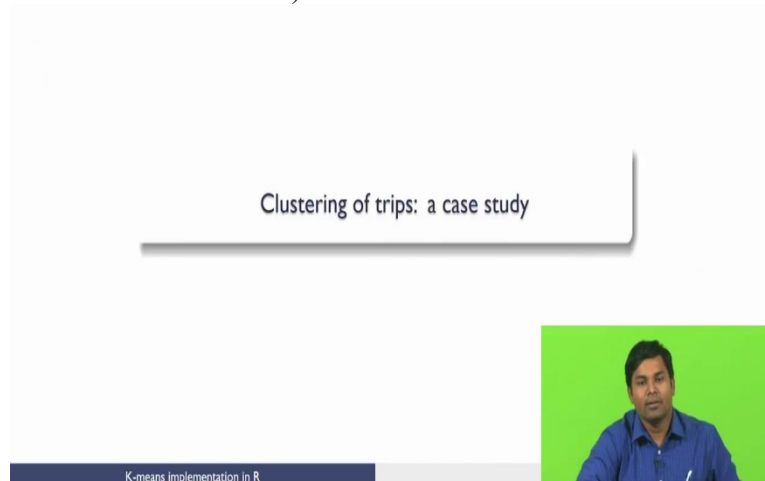


The slide is titled "In this lecture" and is part of a presentation on "Data science for Engineers" and "K-means implementation in R". It contains a bulleted list of topics to be covered in the lecture. A video inset in the bottom right corner shows a man in a blue shirt speaking against a green background.

- Case study
  - Problem statement
- Solve the case study using R
  - Read the data from a ".csv" file
  - Understand the data
  - k-means() function
  - Interpret the results

First will start with the problem statement of the case study followed by how to solve the case study using R. As a part of the solution methodology we will also introduce the following aspects such as how to read the data from dot csv file, how to understand the already read data which is in the workspace of your R, details about this K-means function and how to interpret the result that are given by this K-means function. Let us first look at the case study.

(Refer Slide Time: 01:30)



We have main this case study as clustering of trips, the reason will become clear when you see the problem statement.

(Refer Slide Time: 01:41)

A presentation slide titled "Clustering of trips: Problem statement". The slide has a white background with a blue header bar at the top containing the text "Data science for Engineers". A video inset in the bottom right corner shows a man in a blue shirt speaking. A dark blue footer bar at the bottom contains the text "K-means implementation in R".

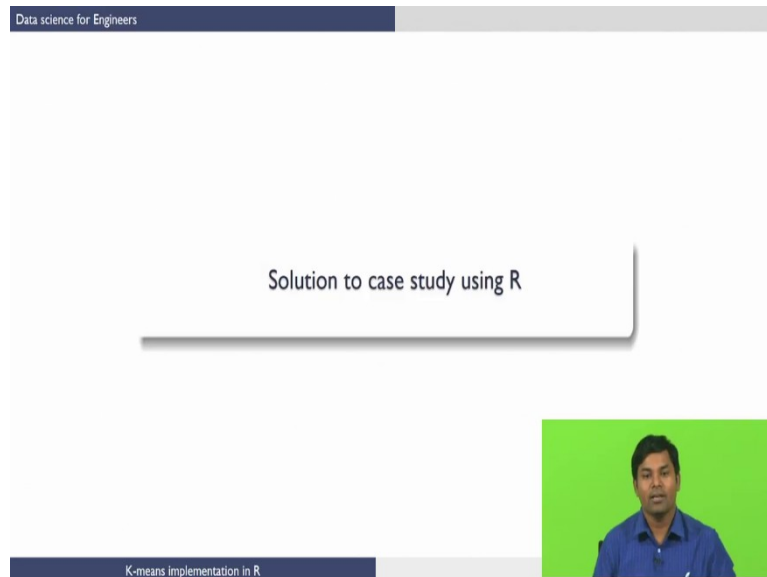
An Uber cab driver has attended 91 Trips in a week (5 days). He has a facility which continuously monitors the following parameters for each trip  
Trip length, Max speed, Most frequent speed, Trip duration, number of times brakes are used, idling time and number times the horn is being honked.

Uber wants to group the trips in to certain number of categories based on the details collected during the trip for some business plan. They have consulted Mr. Sam, a data scientist to perform this job and the details of trips are shared in a ".csv" format file with name "tripDetails.csv"

Let us look at the problem statement of the case study. An Uber cab driver has attended 91 trips in a week. He has a facility in the car which continuously monitors the following parameters for each trip such as trip length, maximum speed, most frequent speed, trip duration, number of times the brakes are used, idling time and number of times the horn is being honked. Uber wants to group this trips into certain number of categories based on the details collected during the trips for some business plan. They have consulted Mister Sam, a data scientist, to perform this job and the details of the trips are shared to him in a dot

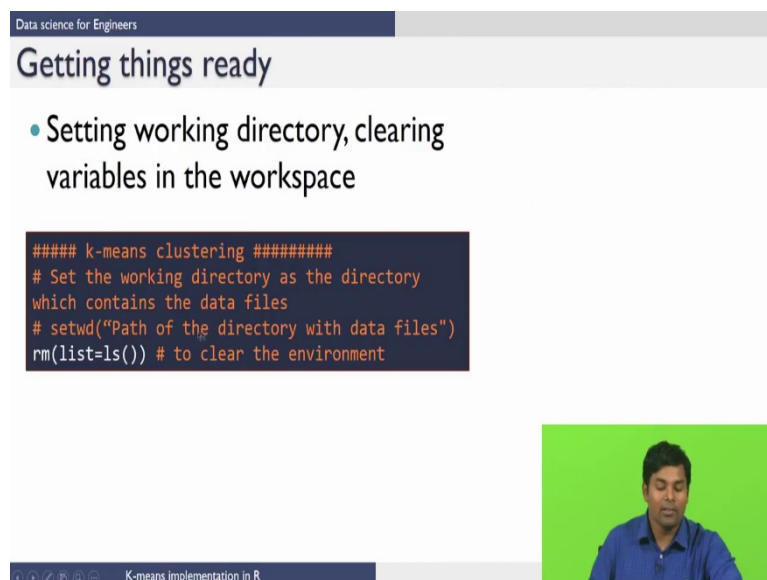
csv format file with the name trip details dot csv. This is a problem statement let us look how to solve this case study using R.

(Refer Slide Time: 02:35)



So, to solve this case study first we need to set up our R studio work space.

(Refer Slide Time: 02:41)



You need to copy the file which we have shared with you into the working directory and clear the variables in the workspace. You can set the working directory using the set working directory command and you can pass the path of the directory which contains this data file as


an argument to the set working directory function. Or you can use the GUI as we have specified in the R module to set the working directory. And this command here removes all the variables that are in the workspace and clear the R environment. You can very well use the brush button to clear all the variables from the environment.

(Refer Slide Time: 03:28)

Data science for Engineers

Reading the data

- Data for this case study is provided to you file with name "tripDetails.csv"
- To read the data from a ".csv" file we use `read.csv()` function



K-means implementation in R

The next step is to read the data from the given file. And data for this case study is provided in a file with name trip details dot csv. If you notice the extension of this file is dot csv which means comma separated value file. In R to read the data from a dot csv file we use read.csv function.

(Refer Slide Time: 04:01)

Data science for Engineers

read.csv()

Reads a file in table format and creates a data frame from it

SYNTAX

```
read.csv(file,row.names=1)
```

file	the name of the file which the data are to be read from. Each row of the table appears as one line of the file.
row.names	a vector of row names. This can be a vector giving the actual row names, or a single number giving the column of the table which contains the row names, or character string giving the name of the table column containing the row names.

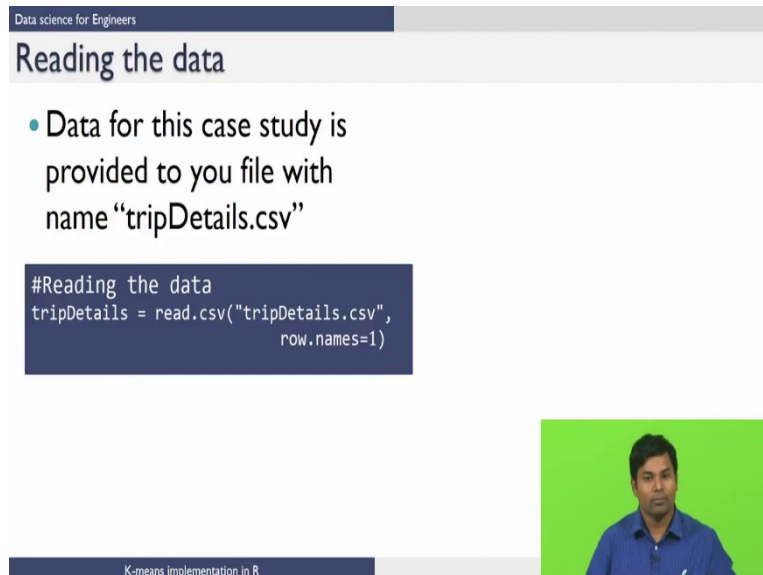


K-means implementation in R

Let us look what does the read.csv function takes as input argument and what it gives us an output. read.csv reads a file in the table format and creates a data frame from it. The syntax of the read.csv is as follows; read dot csv it takes two arguments the first argument is a file name and the second argument is the row names we will see what this individual arguments are about.

The file is the name of the file from which the data has to be read, row names is a vector of row names. This can be a vector giving the actual row names are a single number specifying which column of the table contains this row names. So, essentially the syntax is read dot csv, the filename and if you have a column which species the row names you can give that particular column as row names. In this case we have the first column as the row indices that is the reason why we have given this row dot names as 1.

(Refer Slide Time: 05:08)



Data science for Engineers

## Reading the data

- Data for this case study is provided to you file with name "tripDetails.csv"

```
#Reading the data
tripDetails = read.csv("tripDetails.csv",
                      row.names=1)
```

K-means implementation in R

Let us see how to read the data from the trip details dot csv. The data from this tripdetails.csv can be read by executing this following command here. I am using read dot csv command and I am trying to read the data from tripdetails.csv and I know that in the first column of the csv file I have the row names. Therefore, I have specified row dot names as one and this is the filename from which I want to have read the data.

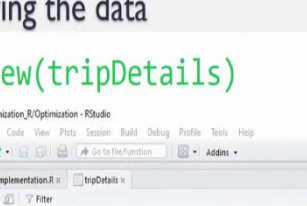
And assigning the data which is read from this read dot csv to the object called trip details. As mentioned in the help of read.csv, it reads the data from this tabular format and then assign it to object call trip details which is of type data frame.

(Refer Slide Time: 06:07)

# Data science for Engineers

## Viewing the data

- View(tripDetails)



The screenshot shows the RStudio interface with the 'tripDetails' data frame loaded. The data is displayed in a table with the following columns: TripLength, MaxSpeed, MostFreqSpeed, TripDuration, Brakes, IdlingTime, and Honking. The data is sorted by TripLength in descending order.

	TripLength	MaxSpeed	MostFreqSpeed	TripDuration	Brakes	IdlingTime	Honking
1	21	51	14	93	307	27	112
2	148	130	106	156	226	5	114
3	18	38	16	100	351	26	107
4	22	43	48	38	17	4	5
5	183	108	90	171	88	5	29
6	18	43	13	64	136	25	21
7	20	37	15	85	121	26	23
8	21	38	14	69	114	25	20
9	181	90	108	155	86	5	25
10	174	100	92	133	106	5	34
11	177	130	85	152	210	5	128
12	17	67	41	30	33	4	17
13	19	42	14	102	429	27	97

Clipboard 1 to 13 of 31 samples

### K-means implementation in R

Once you get this data onto your R environment you can view the data frame using `view` function. Notice this is capital V and once you run this command it will pop up a tabular column in your editor window which shows the variables in the data frame and the number of entries in the data frame.

In this case we have 7 variables and around 91 entries. This is how you can view the data frame once it is loaded into your workspace.

(Refer Slide Time: 06:50)

# Data science for Engineers

## Understanding the data


### Variables

	Trip length	Max. Speed	Most Freq. speed	Trip duration	Brakes	Idling time	Honking
1	21	51	14	93	307	27	112
2	148	130	106	156	226	5	114
3	18	38	16	100	351	26	107
4	22	43	48	36	17	4	5
5	183	108	90	171	88	5	29
6	18	43	13	64	136	25	21
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...

Data contains 91 Trips where 7 variables (columns) named

Trip length, Max speed, Most Freq. speed, Trip duration, Brakes, Idling time and Honking are noted for each trip

91 observations



K-means implementation in R

So, to make it very clear, we have this 7 variables and there are 91 observations in this data frame. And the 7 variables are trip length,

maximum speed, most frequent speed, trip duration, brakes, idling time and honking are noted for each trip.

(Refer Slide Time: 07:13)

Data science for Engineers

Structure of the data


- Structure of data
  - Variables and their data types
- `str()`  
Compactly display the internal structure of an R object

SYNTAX

```
str(object)
```

object	any R object about which you want to have some information.
--------	-------------------------------------------------------------

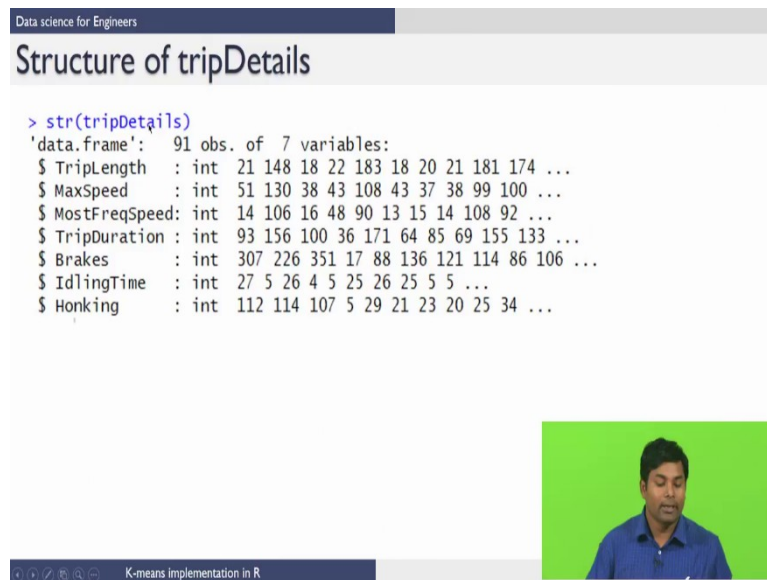
K-means implementation in R



Now, that we have seen how a data frame looks, it is time to see what are the data types of each variable that is available in the data frame. What is the way one can do that? One has to use the structure function to do that. The structure function compactly displays the internal structure of an R object.

The syntax of the structure function is as follows, structure and the argument it takes place an R object. This object is an any R object about which you want to have some information.

(Refer Slide Time: 08:00)



The slide is titled "Structure of tripDetails" and is part of a presentation on "Data science for Engineers". It displays the output of the R command `str(tripDetails)`. The output shows that `tripDetails` is a data frame with 91 observations and 7 variables, all of which are integers. The variables and their first few values are: `TripLength` (21, 148, 18, 22, 183, 18, 20, 21, 181, 174, ...), `MaxSpeed` (51, 130, 38, 43, 108, 43, 37, 38, 99, 100, ...), `MostFreqSpeed` (14, 106, 16, 48, 90, 13, 15, 14, 108, 92, ...), `TripDuration` (93, 156, 100, 36, 171, 64, 85, 69, 155, 133, ...), `Brakes` (307, 226, 351, 17, 88, 136, 121, 114, 86, 106, ...), `IdlingTime` (27, 5, 26, 4, 5, 25, 26, 25, 5, 5, ...), and `Honking` (112, 114, 107, 5, 29, 21, 23, 20, 25, 34, ...). In the bottom right corner, there is a small video inset showing a man in a blue shirt speaking against a green background. The bottom of the slide has a navigation bar with icons and the text "K-means implementation in R".

```
> str(tripDetails)
'data.frame':  91 obs. of  7 variables:
 $ TripLength  : int  21 148 18 22 183 18 20 21 181 174 ...
 $ MaxSpeed    : int  51 130 38 43 108 43 37 38 99 100 ...
 $ MostFreqSpeed: int  14 106 16 48 90 13 15 14 108 92 ...
 $ TripDuration : int  93 156 100 36 171 64 85 69 155 133 ...
 $ Brakes      : int  307 226 351 17 88 136 121 114 86 106 ...
 $ IdlingTime  : int   27  5 26  4  5 25 26 25  5  5 ...
 $ Honking     : int  112 114 107  5 29 21 23 20 25 34 ...
```

Now, let us look at the structure of the data frame which we have extracted from the trip details dot csv. The data frame which we have extracted from trip details dot csv is trip details and I am passing that data frame as an argument to this structure function. When I execute this command, it will show that trip details is a data frame which contains 91 observations of 7 variables and the 7 variables are trip length, maximum speed, most frequent speed and so on.

And correspondingly it gives what is the data type of these variables. If you can notice, all of them are integer type data variables. Keep this in mind because the K-means wants all the variables in the data matrix as the numeric variables or integer variables. We will see that when we discuss about the K-means function as we go along.

Now, structure gives you type of the object and variables that are there in the object and their data types.



(Refer Slide Time: 09:12)

Data science for Engineers

## Summary of the data

- Summary of data
  - Five point summary of the numeric variables
- `summary()`


Summary is a generic function used to produce result summaries of the results of various model fitting functions and five point summaries of numeric R objects

SYNTAX

```
summary(object)
```

object	any R object about which you want to have some information.
--------	-------------------------------------------------------------

K-means implementation in R



There is another command which is summary which gives you the five point summary of the numeric variables. This is the function summary. Summary function is a generic function used to produce results summaries of the results of various models and five point summaries of numeric R objects. The syntax for the summary function is as follows.

The summary function takes one argument which is an R object. This object is again any R object about which you want to know some information. Let us see the summary for our data frame trip details.

(Refer Slide Time: 09:50)

Data science for Engineers

## Summary of tripDetails

```
> summary(tripDetails)
```


tripLength	MaxSpeed	MostFreqSpeed
Min. : 16.00	Min. : 35.00	Min. : 12.00
1st Qu.: 20.00	1st Qu.: 42.00	1st Qu.: 15.50
Median : 21.00	Median : 54.00	Median : 42.00
Mean : 70.77	Mean : 70.36	Mean : 50.65
3rd Qu.:163.00	3rd Qu.:105.50	3rd Qu.: 89.00
Max. :210.00	Max. :138.00	Max. :118.00

TripDuration	Brakes	IdlingTime
Min. : 22.00	Min. : 14.0	Min. : 4.00
1st Qu.: 34.50	1st Qu.: 36.5	1st Qu.: 5.00
Median : 88.00	Median :100.0	Median : 5.00
Mean : 87.37	Mean :135.4	Mean :11.59
3rd Qu.:133.00	3rd Qu.:198.0	3rd Qu.:24.00
Max. :171.00	Max. :429.0	Max. :32.00

Honking

Min. : 4.00
1st Qu.: 20.00
Median : 25.00
Mean : 49.92
3rd Qu.: 97.50
Max. :155.00

K-means implementation in R



When you execute the summary command on your data frame trip details, it will give you five point summary for all the 7 variables of your data frame.

(Refer Slide Time: 10:02)

Data science for Engineers

## K-means clustering

- Given the dataset of trip details, Mr. Sam's job is to segregate these trips into clusters
  - We seek an answer through k-means clustering
- Using k-means clustering on data
  - k-means clustering in R can be applied on data using "`kmeans()`" function

K-means implementation in R

Now, let us look at our primary task of implementing K-means clustering on the data frame. So, we have been given this data set of trip details and we have to segregate the trips into clusters.

We are seeking an answer through the K-means clustering algorithm. If you would have noticed in the data the trips are not labeled as short trip, long trip and so on. This means the data what we have is an unlabeled data and when one wants to learn from this unlabeled data one has to go for unsupervised learning technique. K-means is one such unsupervised learning technique and this K-means clustering in R can be implemented by using this K-means function.

Let us look at what this K-means function takes as input arguments and what does it return.

(Refer Slide Time: 11:03)

Data science for Engineers


kmeans()

object = kmeans(x, centers, iter.max = 10, nstart = 1)

Arguments

x	numeric matrix of data, or an object that can be coerced to such a matrix (such as a numeric vector or a data frame with all numeric columns).
centers	either the number of clusters, say k, or a set of initial (distinct) cluster centers. If a number, a random set of (distinct) rows in x is chosen as the initial centers.
iter.max	the maximum number of iterations allowed.
nstart	if centers is a number, how many random sets should be chosen?
object	an R object of class "kmeans", typically the result "ob" of ob <- kmeans(.).

K-means implementation in R



K-means takes several input arguments. I have given few of important arguments here x stands for numeric matrix of data or the objects that can be coerced to a matrix and this centers we can either give the number of clusters you want to generate out of this data or you can give a set of initial cluster centers.

So, if you specify this k as a number let us say 3 clusters, what it does is it will choose a random set of distinct rows in this numeric matrix x as the initial centers. And as you know K-means is an iterative algorithm. You can set a maximum iteration limit using this iter max argument.

The K-means clustering algorithm depends upon the initial cluster centers, this option here nstart will help you to specify how many sets of different cluster centers has to be used to come up with the model. And finally, the output argument is object which returns an R object which is of class K-means that is basically is a result of a function which is as shown here that is exactly what we have here. When you execute this command, it will take the data matrix and it will take number of clusters you want to build from this data and returns you a K-means R object. Let us now implement this K-means algorithm on our data.

(Refer Slide Time: 12:45)


Data science for Engineers

Implementing K-means

- Clustering data using k-means and seeing the clusters details

```
# k-means clustering using kmeans command
tripCluster <- kmeans(tripDetails,3)
```

K-means implementation in R



This is what we are essentially doing. We are clustering our data using the K-means and we can see what are the details that this K-means gives as an output. This command here takes this data frame trip details and this argument here specifies I want to divide the data into 3 clusters. When I execute this command it will divide the data into 3 clusters and it will assign the result as a trip cluster R object which is essentially a list. Let us see what information does this trip cluster has.

(Refer Slide Time: 13:28)

Data science for Engineers

Results

tripCluster has the following information


```
> tripCluster
K-means clustering with 3 clusters of sizes 46, 15, 30

Cluster means:
  TripLength  MaxSpeed MostFreqSpeed TripDuration  Brakes
1  19.91304  48.21739    32.82609    50.13043  59.93478
2  20.26667  45.06667    14.46667    88.73333 350.13333
3 174.00000 116.96667    96.06667   143.80000 143.86667

  IdlingTime  Honking
1 11.413043 15.60870
2 25.400000 97.73333
3  4.966667 78.63333

Clustering vector:
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
2 3 2 1 3 1 1 1 3 3 3 1 2 1 1 3 1 1 2 1 3
22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42
1 2 2 2 1 1 2 3 3 1 1 1 3 1 3 3 3 1 2 3 1
43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63
1 1 3 1 2 1 3 1 1 1 1 1 2 3 1 1 3 1 1 1 3
64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84
3 2 1 1 1 3 1 1 2 1 1 1 3 3 3 1 1 3 3 3
85 86 87 88 89 90 91
1 1 2 3 2 3 1
```

K-means implementation in R



Trip cluster has the following information. The first line essentially gives you it has clustered the data into 3 cluster which are of sizes 46, 16 and 30 and if you sum this up, you will end up with your total number of rows in your data frame that is 91 you can verify that. And for each variable it will give the means of the clusters. So, because you wanted to divide the whole data into 3 clusters the first row is the cluster one information where the mean of the trip length is 19.9 and the mean of the maximum speed is 48.21 and so on.

Similarly the lines 2 and 3 represents the information about the cluster. So, what can one infer from this cluster means? You can actually see for example, mean trip length is 19.9, I can actually say that this is of shortest trip and of the mean trip length is 174 I can treat this as a long trips. So, this information can essentially help the people who wanted to do this analysis and then categorize these clusters into meaningful information depending upon the problem they are looking at.

The next means of information that trip cluster contains is, it will say among 91 rows what does each row belong to. For example, look at here the first element that means, the first row belongs to the cluster 2 and the second element belongs to the cluster 3, total means belongs cluster 2 and so on it will identify each row into one of this clusters. And as you know K-means is a hard clustering algorithm that means, each point has to be allotted to any of the clusters and no two clusters contain similar data point otherwise you cannot have a single data point belonging into 2 clusters.

(Refer Slide Time: 15:51)


Data science for Engineers


## Results

tripCluster has the following information

```
Within cluster sum of squares by cluster:
[1] 160740.2 25986.8 194647.0
(between_SS / total_SS = 83.3 %)

Available components:
[1] "cluster" "centers" "totss" "withinss"
[5] "tot.withinss" "betweenss" "size" "iter"
[9] "ifault"
```



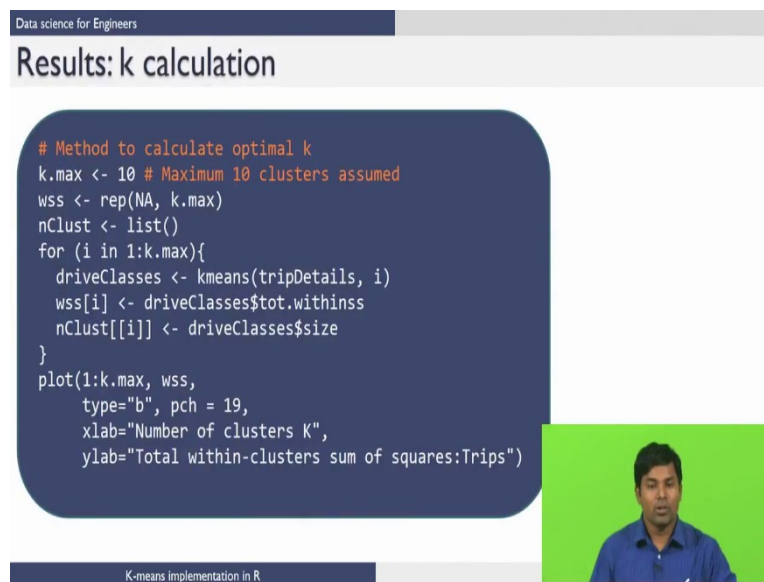


K-means implementation in R

This trip cluster has some more information which is known as within clusters sum of squares. And if you see this contains 3 elements. That means, how much is the variance in each of these clusters is what this within cluster sum of the squares uses. The lesser the variance the good are the clusters and it will also give what are the other components that you can look from the trip clusters.

Essentially when you build a K-means algorithm it will give you the following information. How many clusters it has built and how many data points are there in each of these clusters and what are the means of each of these clusters and how are each points are categorized into 1 of the 3 clusters which you want to build and the other information.

(Refer Slide Time: 16:59)



The slide is titled "Results: k calculation" and is part of a presentation on "Data science for Engineers". It features a code block with R code for calculating the optimal number of clusters (k) using the elbow method. The code defines a function that iterates through different values of k (from 1 to k.max), performs K-means clustering on the 'tripDetails' dataset, and calculates the within-cluster sum of squares (WSS). The results are plotted as a scatter plot with the x-axis labeled "Number of clusters K" and the y-axis labeled "Total within-clusters sum of squares: Trips". A small video inset in the bottom right corner shows a man in a blue shirt speaking.

```
# Method to calculate optimal k
k.max <- 10 # Maximum 10 clusters assumed
wss <- rep(NA, k.max)
nClust <- list()
for (i in 1:k.max){
  driveClasses <- kmeans(tripDetails, i)
  wss[i] <- driveClasses$tot.withinss
  nClust[[i]] <- driveClasses$size
}
plot(1:k.max, wss,
     type="b", pch = 19,
     xlab="Number of clusters K",
     ylab="Total within-clusters sum of squares:Trips")
```

K-means implementation in R

So, biggest problem with this K-means algorithm is to figure out what number of clusters has to be given. So, there is one method which is called as the elbow method which can help us to calculate the optimal number of clusters K. For that all you need to do is you have to write one for loop which does lot of K-means algorithms and get the metric which is within sum of squares and when you plot this within sum of squares with the number of clusters you will find out after certain number of clusters the decrease in this within sum of squares value is low where you say look this is the optimal number of clusters into which your data has to be divided.

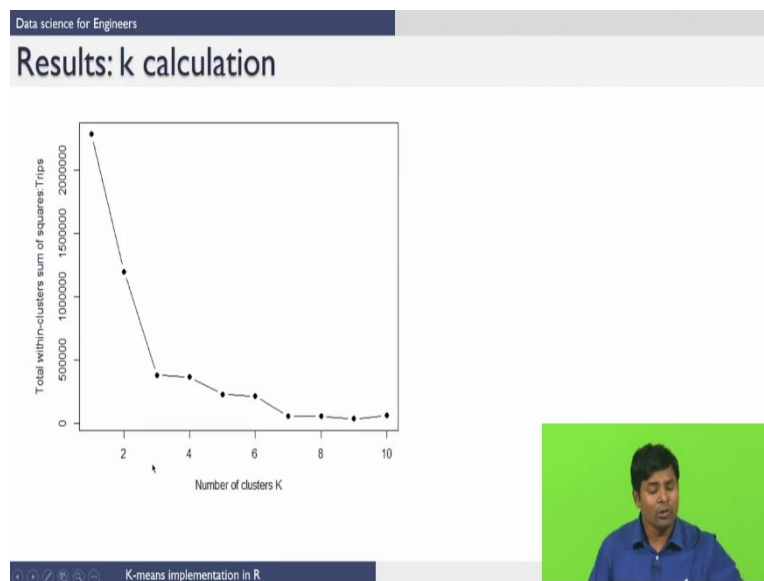
Let us illustrate this code line by line. I am assigning a value of ten to the k max and I am pre allocating a memory which is, so this many number of clusters us there and I am repeating NA's for this many number of times. This is within sum of squares value. Essentially I am

creating a vector with NA's which is of the size 10 by 1. And I am initializing in empty list for the number of clusters.

This for loop will run from 1 to number of maximum clusters that is in this case its 10 and for each value of i this K-means algorithm is implemented, the objects are being stored into this drive classes and in the drive classes the total within sum of squares distance is been allocated into this wss of i. And the size of the cluster is allocated into the components of the list which you have created.

So, once this for loop get executed you will get a vector of the within sum of those errors and a list of the number of clusters. Now, we can plot that with number of clusters in x axis and within sum of squares value in y axis and type = b represents both line and points has to be there and pch = 19 specify the symbol that has to be used along with the plot and x lab and y lab has their normal meaning which are x label and the y label.

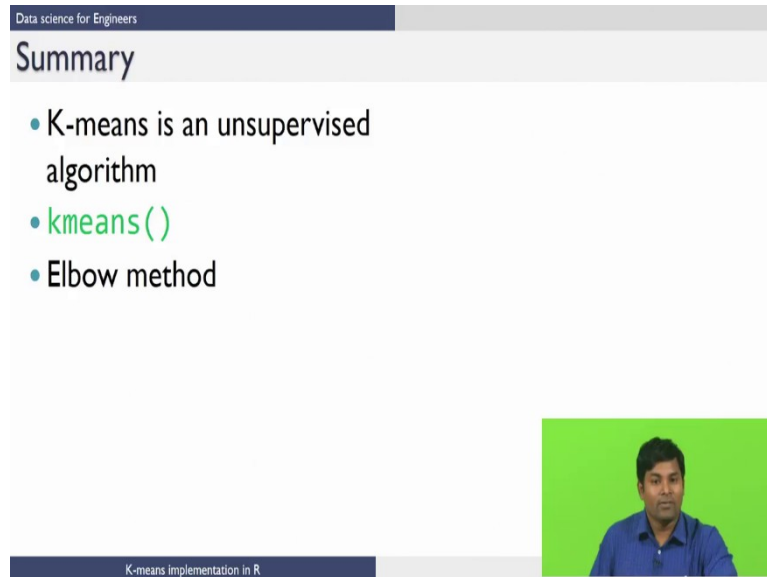
(Refer Slide Time: 19:27)



Let us look at the plot. So, this is the x axis which is number of clusters which is varied from 0 to 10 because the maximum number of clusters we had is 0 to 10 and this y axis is total within sum of squares values with respect to trips and if you can see this total within sum of square value drastically decreased when one moved from 1 cluster to 2 clusters and from 2 clusters to 3 clusters and after that decrease in total within sum of squares clusters is not much when compared to the earlier ones.

That is the reason why I can choose this  $K = 3$  as my optimal number of clusters.

(Refer Slide Time: 20:13)



The video player interface displays a slide titled "Summary" under the heading "Data science for Engineers". The slide contains three bullet points: "K-means is an unsupervised algorithm", "`kmeans()`", and "Elbow method". At the bottom of the slide, it says "K-means implementation in R". A small video inset in the bottom right corner shows a man in a blue shirt.

In summary we have seen that K-means algorithm is an unsupervised learning algorithm. That means, we have the data which is not labeled. In this cases we have to use one of the unsupervised learning algorithms, in this case we have use K-means algorithm and we have seen how to implement this K-means algorithm in R, and we have seen one method to find the optimal number of clusters for K-means which is elbow method.

Thank you.