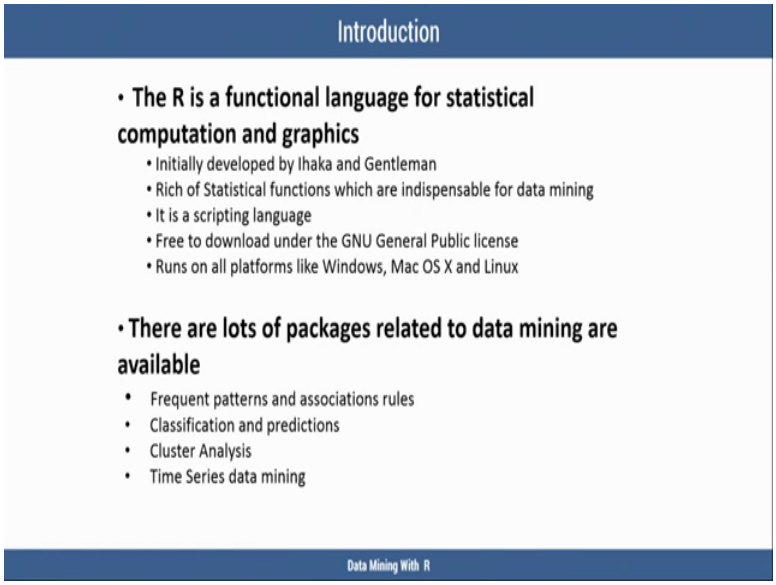
Data Mining
Prof. Arindam Dasgupta
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture – 43
Tutorial

Hello. Welcome to data mining with R course. I am going to present how to apply the R programming language in application of data mining, I am a research scholar in compressors department, my name is Arindam Dasgupta.

(Refer Slide Time: 00:40)

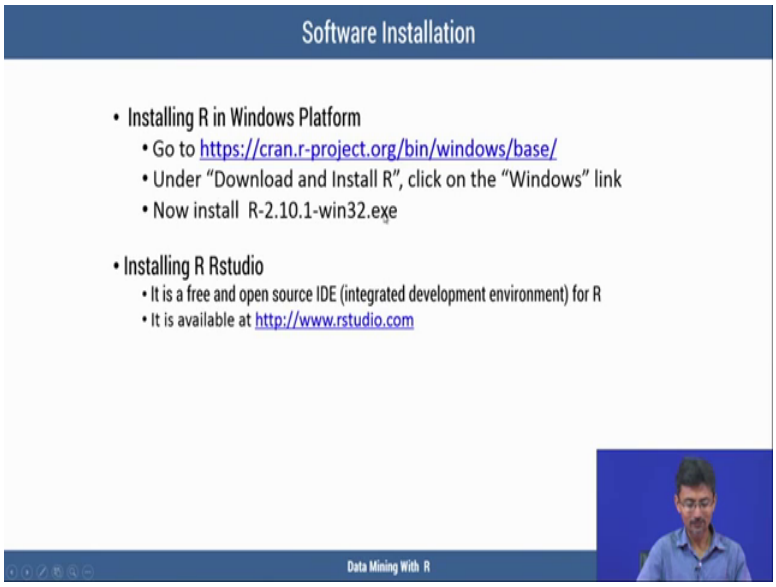


Now, one of what is R? Basically, R is a programming language, it is basically a functional language for statistical computation and graphics, initially, it is developed by Ihaka and gentleman, it is a riched with lots of statistical functions which are very essential for data mining applications. Basically, it is a scripting language and this tool can be downloaded freely from under the GNU license and you can execute run or run the R programming in any platform such as windows MAC, OS, X and Linux because it supports all the platforms. R is very beautiful language because it is basically functional language each functions are already implemented in the R library and we just use that functions for our pass passes.

There are lots of packages data mining available data mining packages are available these packages are frequent pattern analysis and rules as described in the theoretical class

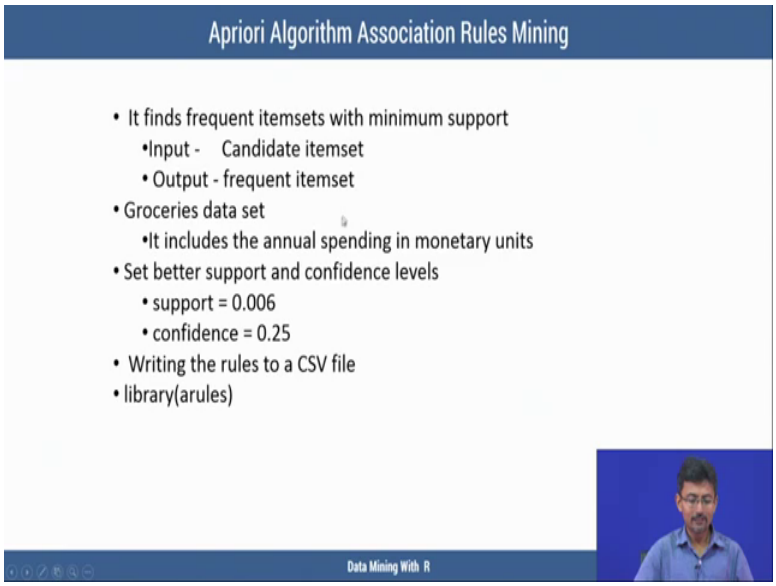
and classification prediction algorithms cluster analysis packages, there are time series data mining packages there are lots of packages available just you have to download the package in real time and just apply the functions on that packages.

(Refer Slide Time: 02:23)



Now, how to get the R tool? To get the R tool, we have to go to the this website is called this website is called cran project org bin windows or base just go to this link at first, then download the R according to your platform in my case, these windows platform that is why I have downloaded R win 32 exe after downloading the software, we have to install the software, it is very easy to install just click on the R exe icon, I will show you the this is the this is the R software, you just double click on; this icon and just install by applying the next I have already installed here.

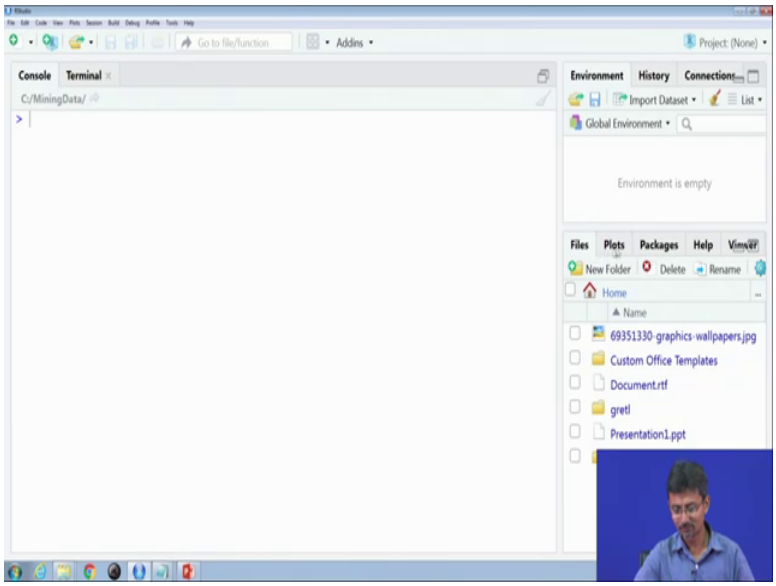
(Refer Slide Time: 03:48)



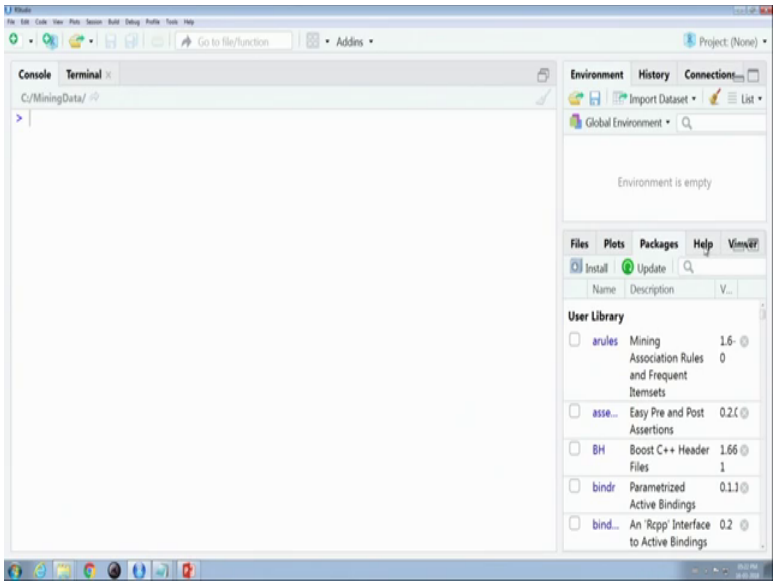
After that after installing the R studio, we have to install another software application that is basically intergrated development environment it is also free and open source id, it is called R studio, it is very user friendly, it provides an user friendly interface to execute your R programming language. I will show you the how to what is the interface of the R studio. These are interface of R studio just I am basically, this part of the top left one part of this area is used to write the programming language R programming language the bottom part is the basically the console of R console of R executing the R script.

Basically, this is the editor and these are R console and right hand side will display, this part top right hand side this part will display the local variables the which are just defined in the R terminal and the bottom of portion you can get the what is this is the files which are in this computer and any plot will be displayed here.

(Refer Slide Time: 05:19)



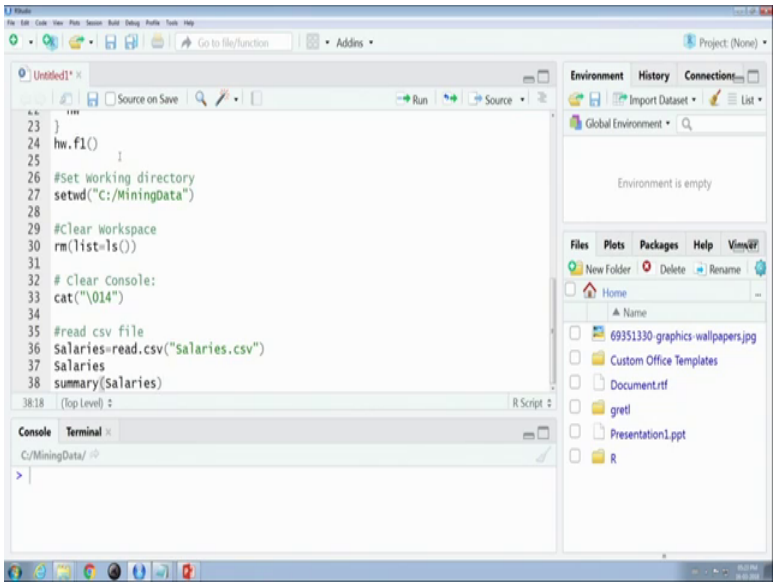
(Refer Slide Time: 05:28)



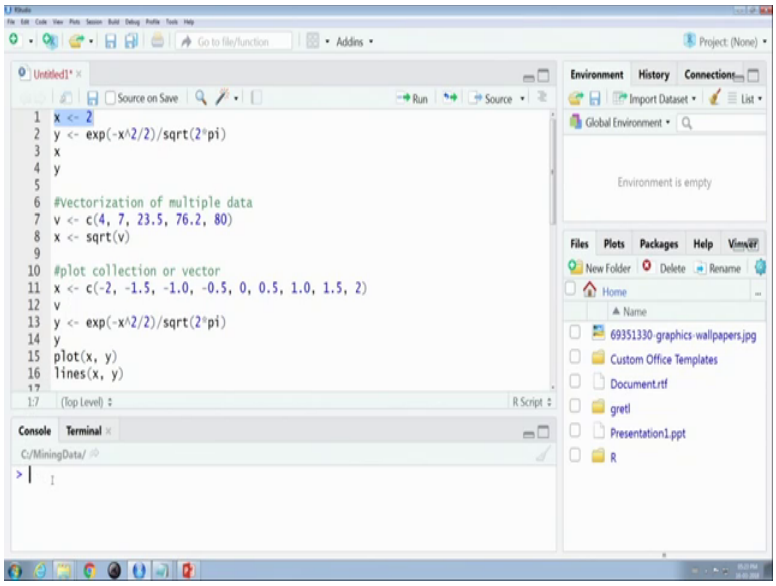
These are package package list; these are which can be used in my program in our program and it the help if any R resource is needed then you can help get help from this link.

And now I want to show you some basic R program at first to click file create a 1 R script new file, I have created one new file and I have just put the script here each scripts are suppose first line in first line of the script, it is.

(Refer Slide Time: 06:31)

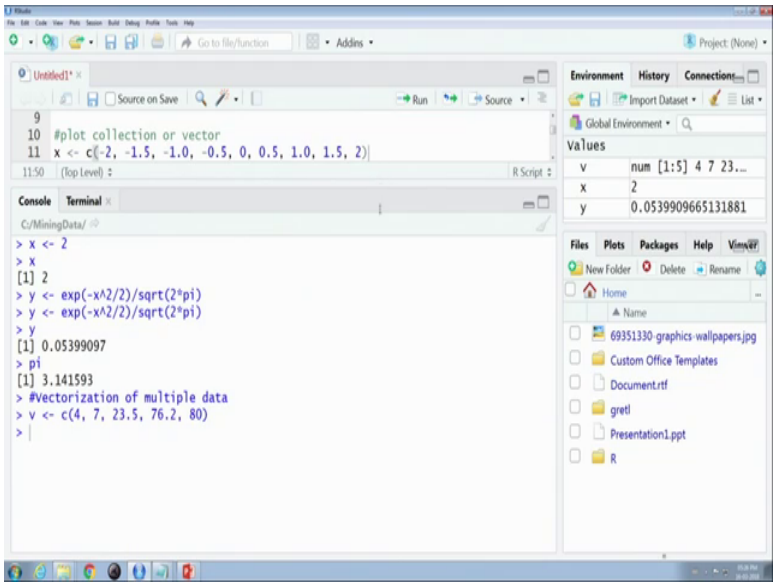


(Refer Slide Time: 06:33)



It is showing x left arrow two here x is the object basically, not is exactly like very well, basically, it is a object it stores the very value 2 in the x variable.

(Refer Slide Time: 07:16)

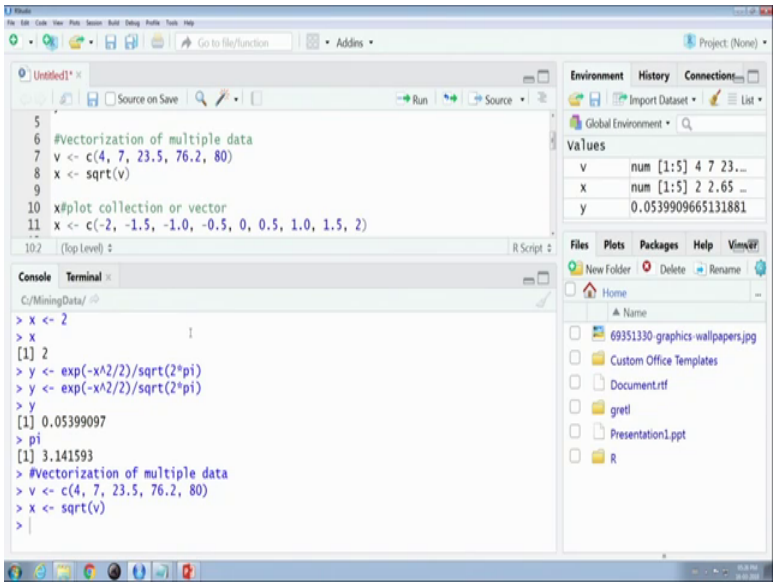


Suppose I am to execute this in R console. Now, I want to execute this after executing this value two is has been assigned to object x, I want to view this value just type x here and write which showing the value of x is 2 ok.

Then I want to assign one expression, I am executing directly from my source script by clicking here just you have to just put the cursor here and click run, it will execute a particular command you see these executing this particular command is written here from this portion to this portion and now I want to view; what is the value of y, this is the value of y, you see, this is a basically expression e to the power minus x square by 2 and by sqrt 2 star pi, if I want to know that what is the value of pi you just type the pi value it is constant; it is already defined in R and then what is the;

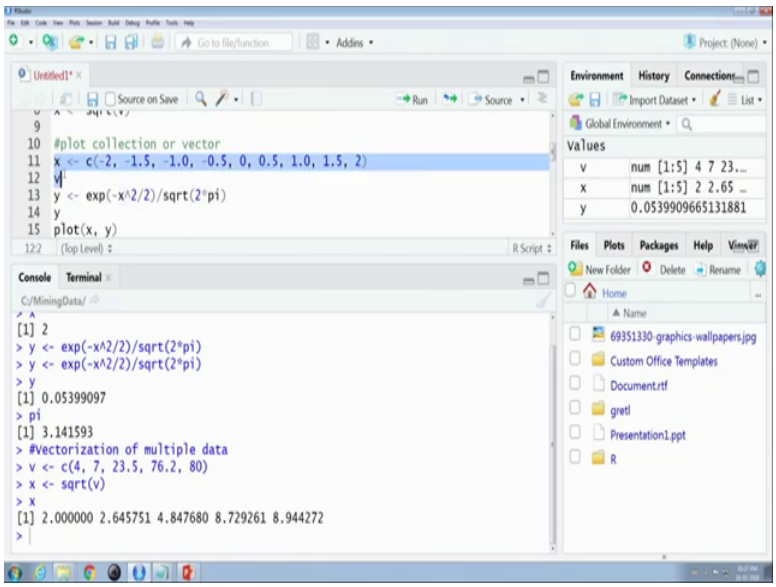
Now, I have to show how to store the multiple datas into a particular object suppose I am to executes this v these vectorization, here see first bracket start within the bracket, there are lots of values 4 7; these are the values different types of values here; basically it it is a it sees for collection it is a collection of datas; this data has been stored into the object v.

(Refer Slide Time: 09:54)



Then after then after storing the value into the v, I want to display the value of v into the x into x. Now see now each of the value has been square rooted. Now check the value of x here, it is showing the each value is square rooted; that means, the vectorization on multiple data and the square root function has been applied in each of the data into the collections.

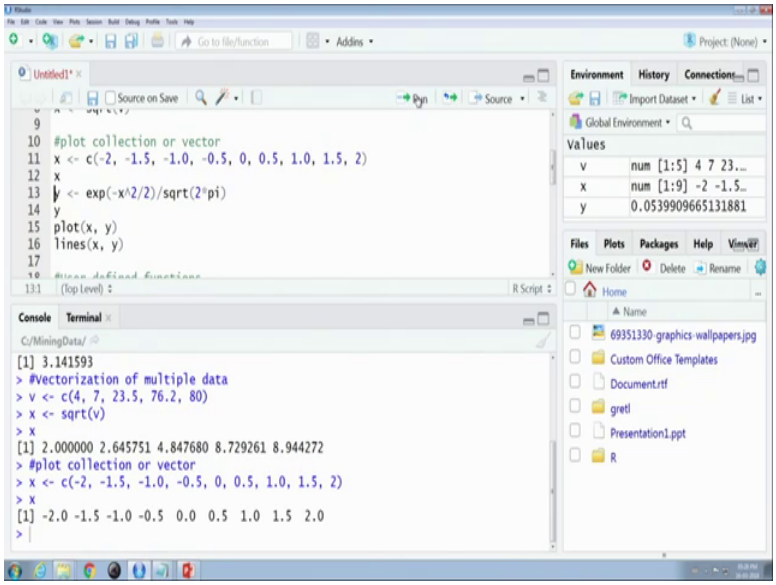
(Refer Slide Time: 10:49)



Now, I want to plot how to plot multiple vector data; suppose, first I store a collection of data into the x variable these the collection of data, then I will display the x variable, then

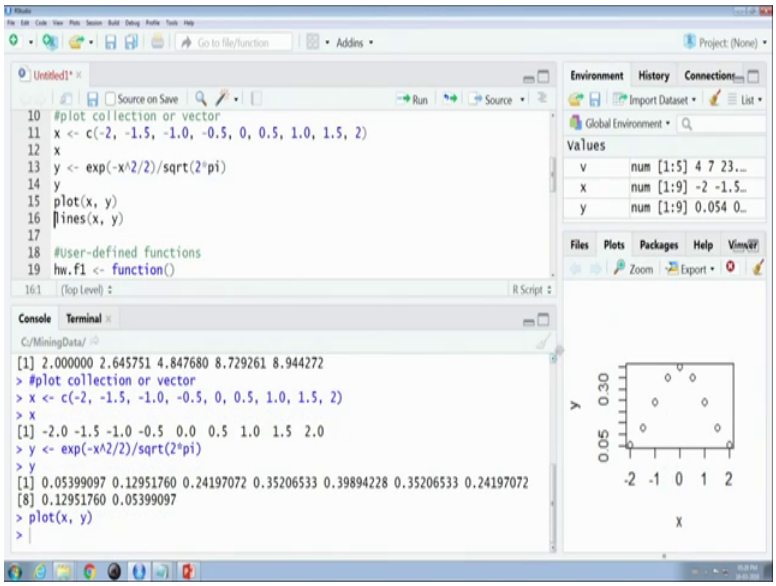
I apply this formula into each of data set here in x then plot, then I will display the value of y, then I plot the data set.

(Refer Slide Time: 11:47)



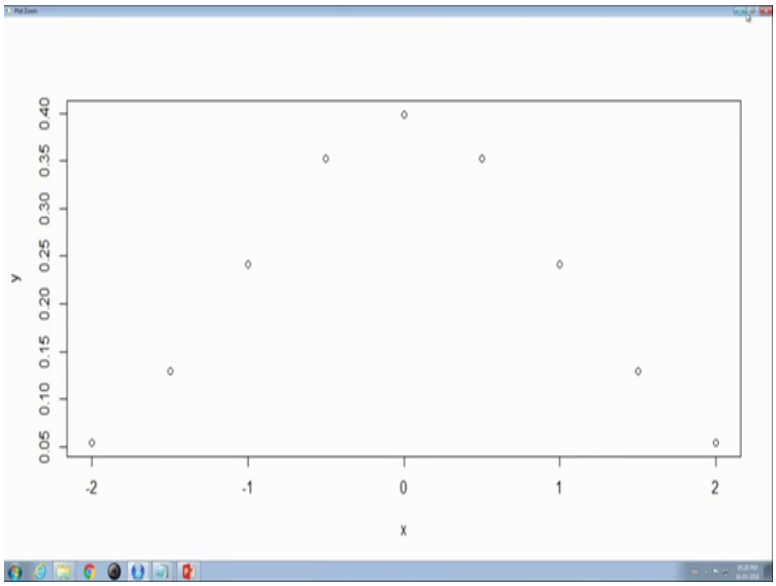
Now, start now it is already stored, then it is displaying the value of x object. Now, it is displaying, it is computing the values of y object, then it is displaying the values of y objects, I want to plot x data of x and y just for plotting the data of x and y just use plot within x axis and y axis.

(Refer Slide Time: 12:23)



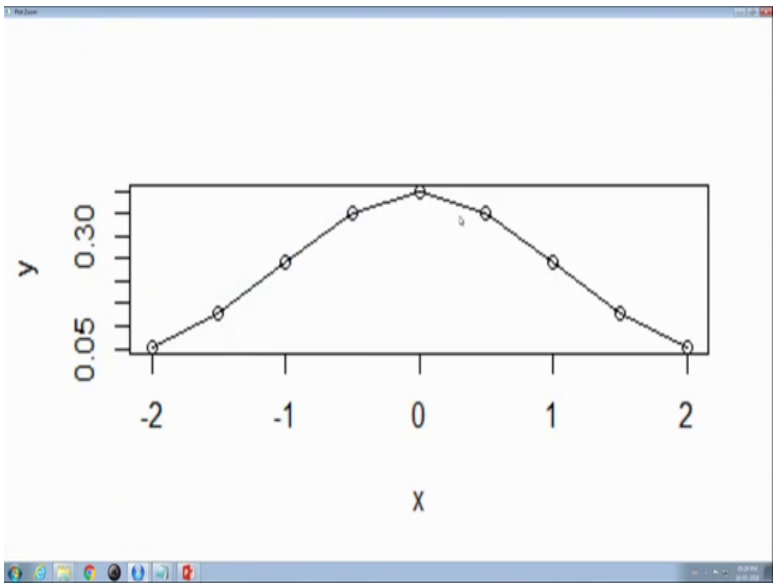
Now, it is plotting here.

(Refer Slide Time: 12:30)



It is showing; these are x axis and these are the y axis, I want to know I want to connect these points with line just typed lines there is a command lines. Now it is connected.

(Refer Slide Time: 13:00)

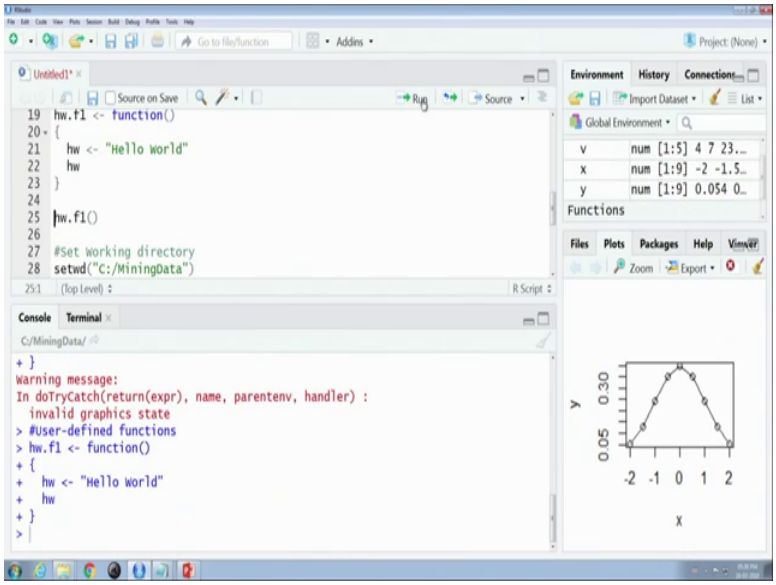


These are the very basic implementation of plotting which will be used in data mining application.

Now, we can define functions in R programming; these are the structure just write a function name here and then define it is a function the function name is this and function definition, I just generate use one object hw, then type insert one string here in this object

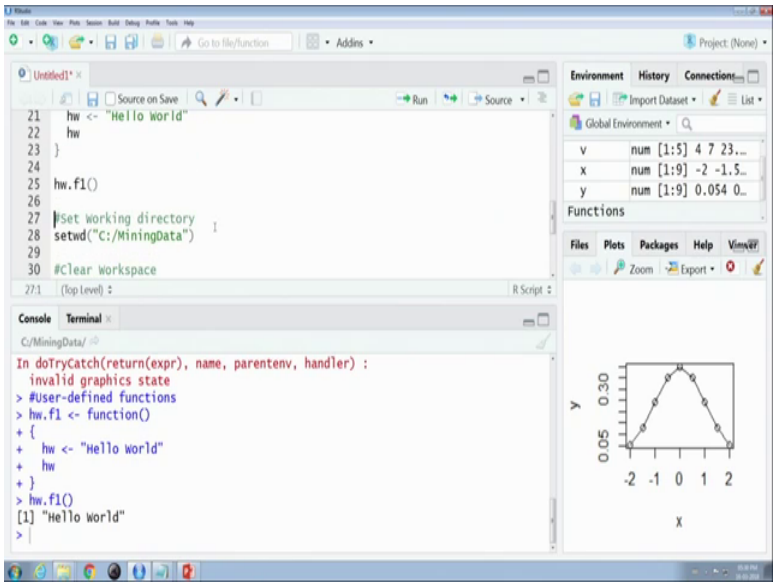
and display the string and this is the this is the calling of function. Now at first I have to define the function. Now function is defined. Now I want to call the function.

(Refer Slide Time: 14:08)



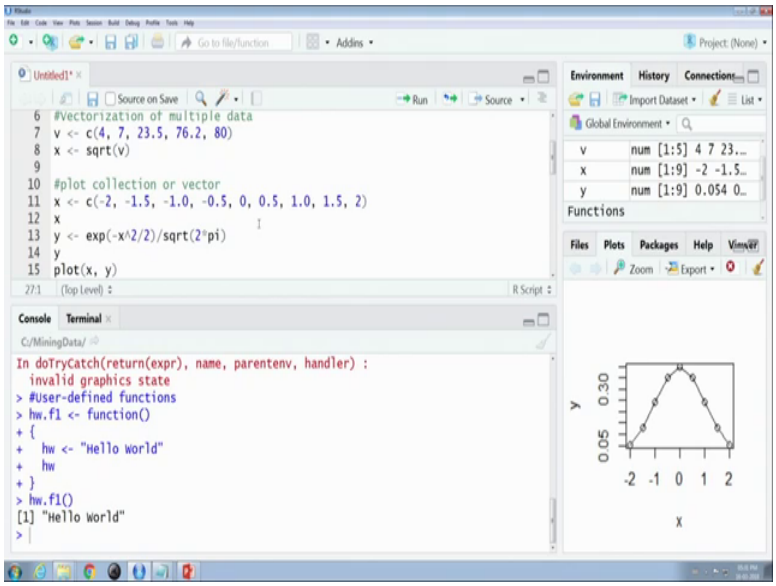
Here it is showing the hello world after by calling the function.

(Refer Slide Time: 14:21)



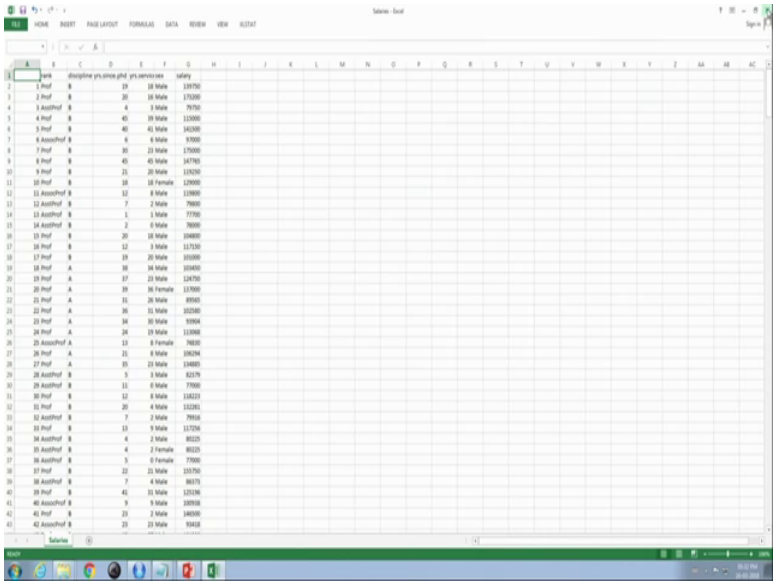
Now, we have learned how to store the values into the objects.

(Refer Slide Time: 14:26)



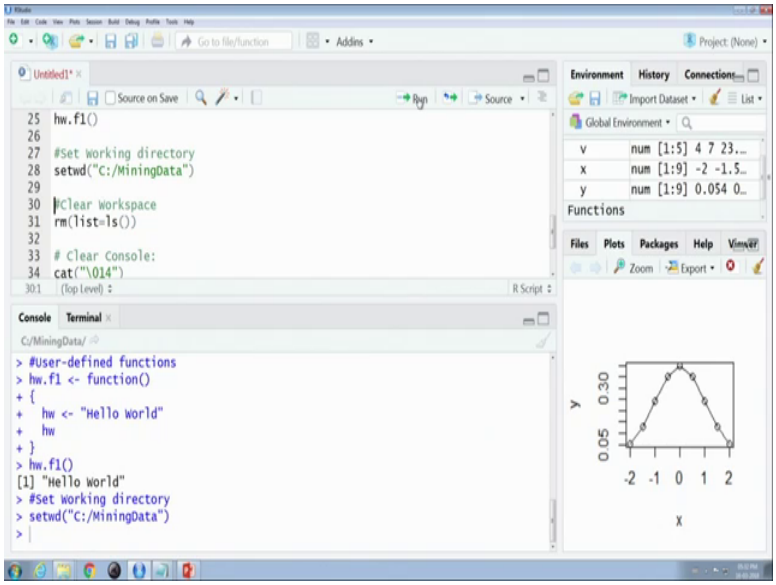
And how to plot data, then I want to access one data file for data mining, you have to access multiple data files or a single file for the analysis purpose for that we have to set one working directory in warren in in my working directory I have stored multiple data files in c drive, I have created one folder data mining data and within these folder I have stored lots of csv files for analysis these are the csv files. Now, I want to retrieve the csv files through R programming.

(Refer Slide Time: 15:28)



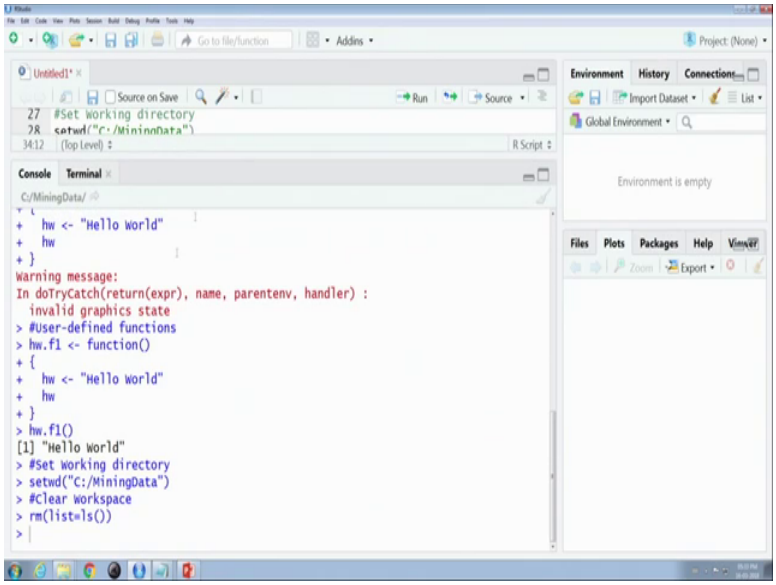
At first, I have to set the working directory because my data files are stored in mining data folder under c drive. Now I execute this instruction.

(Refer Slide Time: 15:59)



Then I have to clear my workspace these workspace these clear these variables and for that you have to use rm list equals to ls now execute this query now my environment is cleared after that clear my this console these R console.

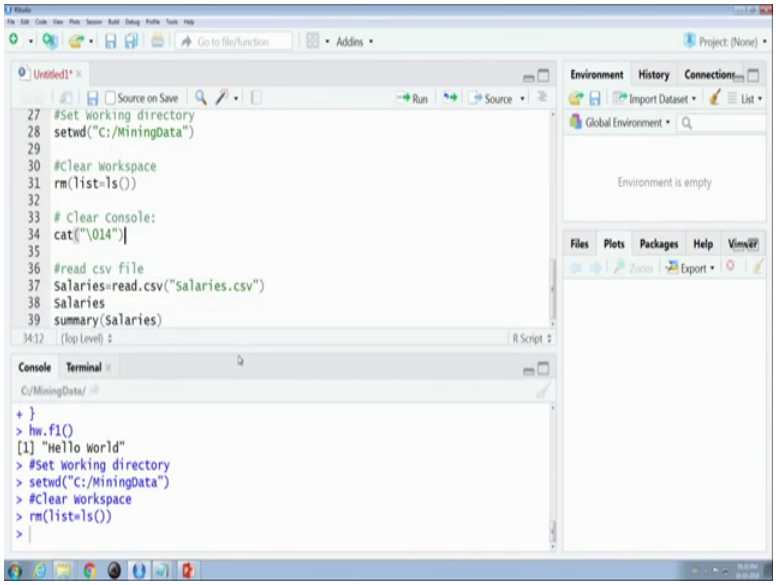
(Refer Slide Time: 16:28)



Clear this console, you have to use this command is cat slash 0 1 4 within string, just execute this command for clearing your console. Now my console is cleared now.

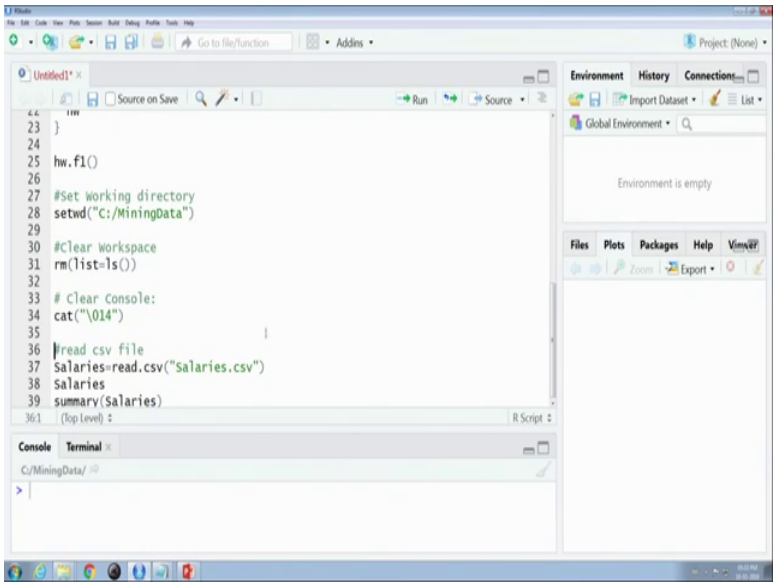
Now, second type we have to read the there is a file salaries dot csv in my data mining directory we have to read this file.

(Refer Slide Time: 16:31)



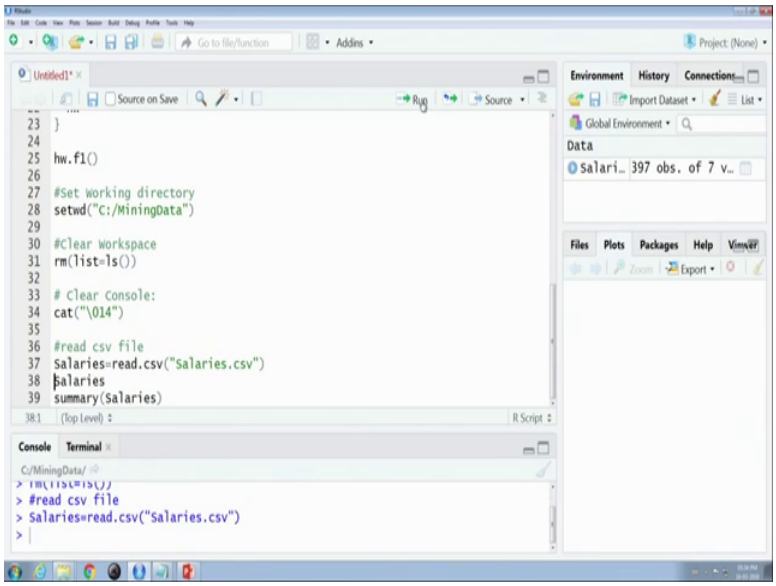
Through this command there is inbuilt function read dot csv.

(Refer Slide Time: 16:44)



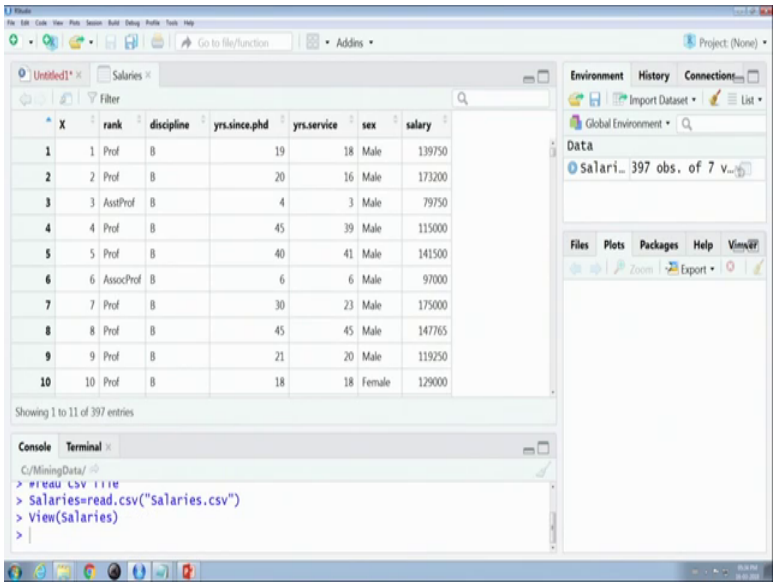
It is meant for accessing the csv file and after executing this command entire data will be stored in salary objects this salary object now salary objects is.

(Refer Slide Time: 17:21)



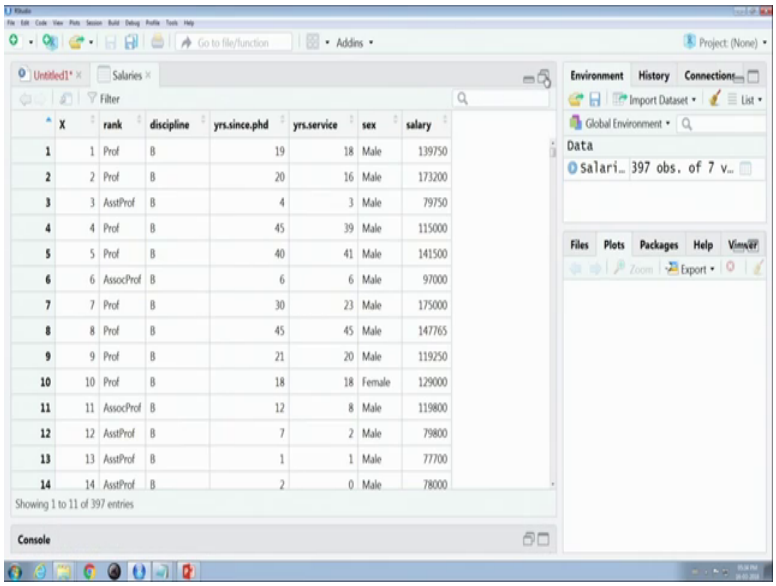
These salary objects you see destroying the tables.

(Refer Slide Time: 17:25)



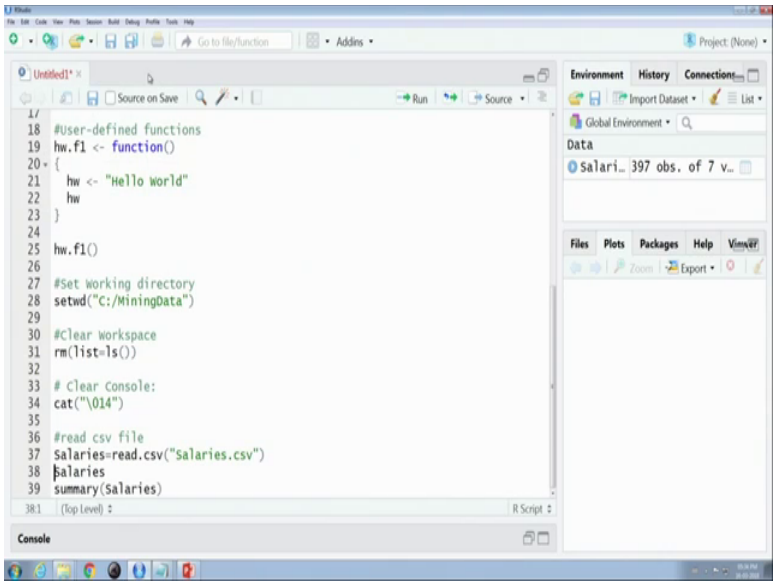
Now, I want to display the salary objects in my R console.

(Refer Slide Time: 17:28)



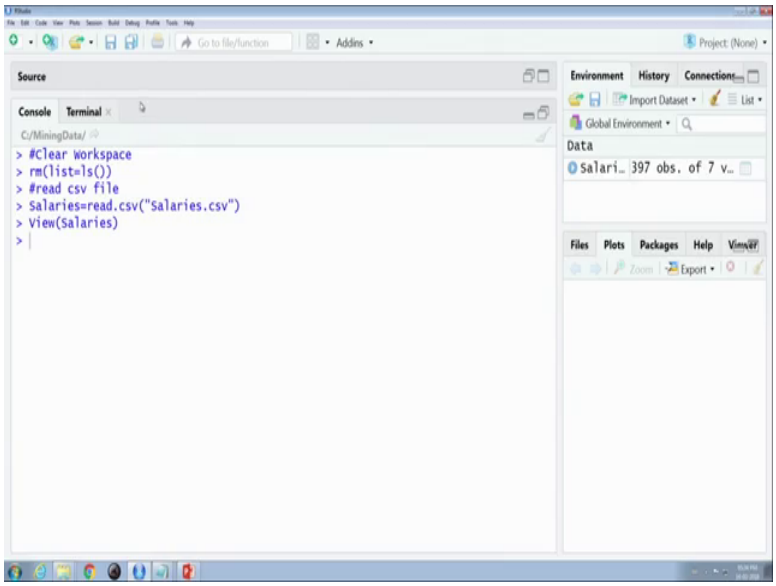
	X	rank	discipline	yrs.since.phd	yrs.service	sex	salary
1	1	Prof	B	19	18	Male	139750
2	2	Prof	B	20	16	Male	173200
3	3	AsstProf	B	4	3	Male	79750
4	4	Prof	B	45	39	Male	115000
5	5	Prof	B	40	41	Male	141500
6	6	AssocProf	B	6	6	Male	97000
7	7	Prof	B	30	23	Male	175000
8	8	Prof	B	45	45	Male	147765
9	9	Prof	B	21	20	Male	119250
10	10	Prof	B	18	18	Female	129000
11	11	AssocProf	B	12	8	Male	119800
12	12	AsstProf	B	7	2	Male	79800
13	13	AsstProf	B	1	1	Male	77700
14	14	AsstProf	B	2	0	Male	78000

(Refer Slide Time: 17:48)

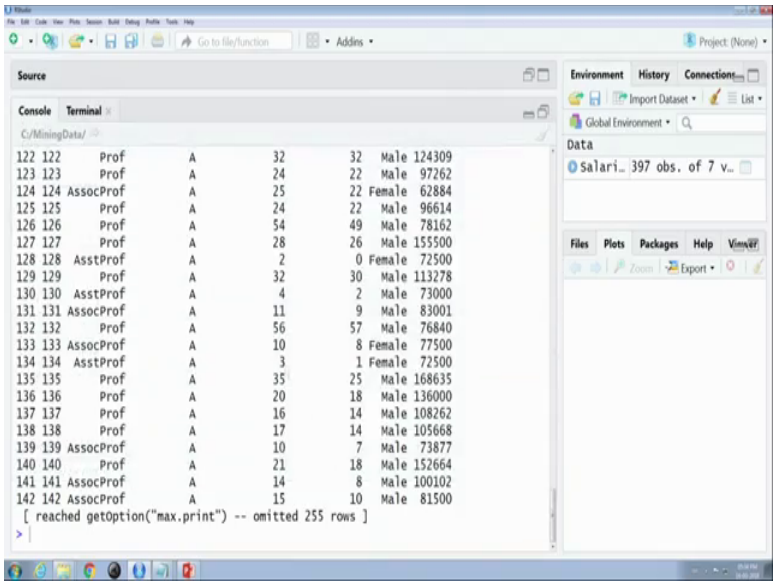


```
#User-defined functions
hw.f1 <- function()
{
  hw <- "Hello World"
  hw
}

hw.f1()

#Set Working directory
setwd("C:/MiningData")

#Clear Workspace
rm(list=ls())

# Clear Console:
cat("\014")

#read csv file
Salaries=read.csv("Salaries.csv")
Salaries
summary(Salaries)
```

(Refer Slide Time: 18:04)

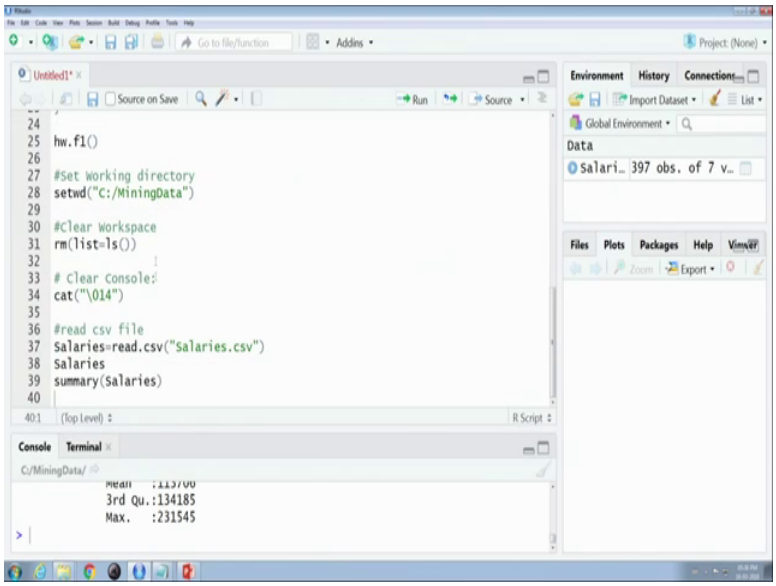


(Refer Slide Time: 18:13)



Now, it is showing in R console suppose, I want to summarize this salary summarize means, it will display the; suppose, there is a very well called discipline in this very well, there are 2 values A and B, it contains 181 values of A and 2 16 values of B. Similarly, these are basically nominal value and the numeric values in numeric values, it is showing the what is the minimum value what is the maximum value what is a quartile what is the median value? Mean value of each attributes in salary the minimum salary is this maximum salary is this and mean salary is this and it is it is it is showing all the summary information summary information of each attribute.

(Refer Slide Time: 19:38)

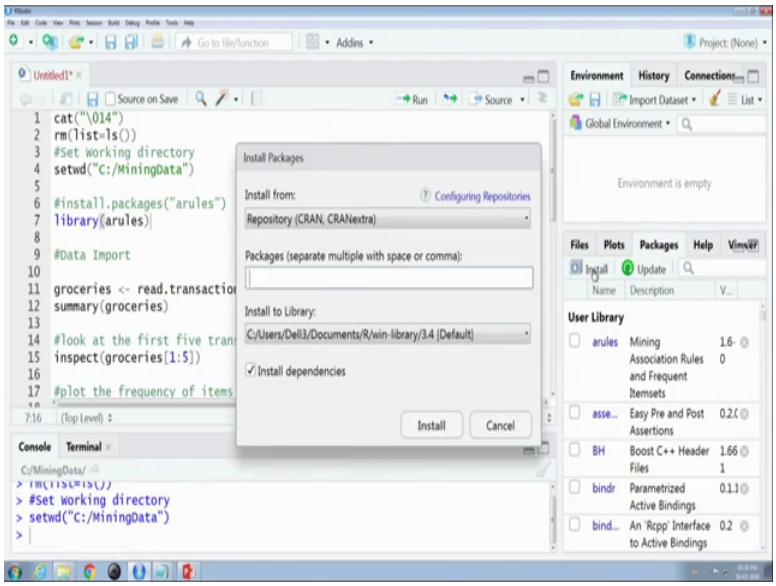


Now, I will leave the basics of R programming which are needed to needed in any time any type of determining algorithms enough.

Now, at first I will discuss about how to use apriori algorithm for accessing association rules from a data set basically apriori algorithm, we have learned it finds the frequent data items with minimum support in I will show, I have a grocery data set and from that grocery data set, we just access the rules with support 0.006 and confidence 2.5 after generating the rules, we will store the rules into a csv file and for apriori algorithm there are the input is one file it is called candidates itemset and output is the frequent itemset ok.

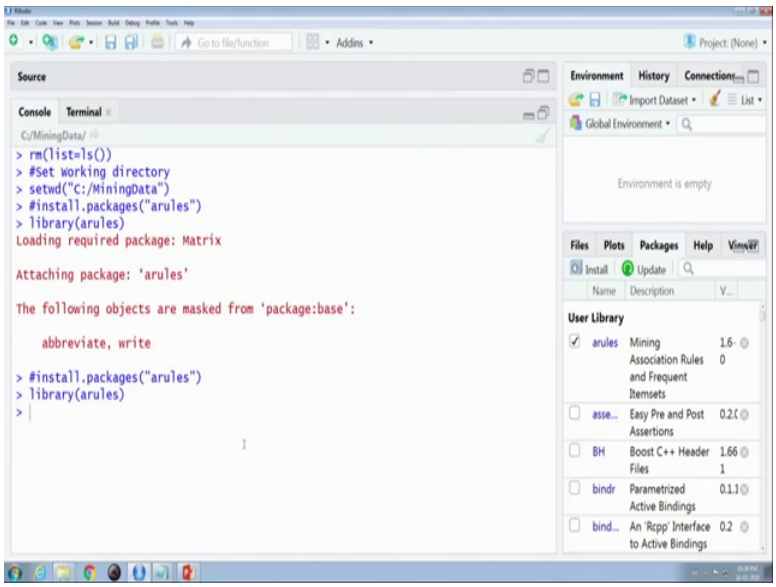
To execute apriori algorithm we have to use one package it is very essential it is called arules to download this package you just go to at right hand side click on install.

(Refer Slide Time: 21:24)



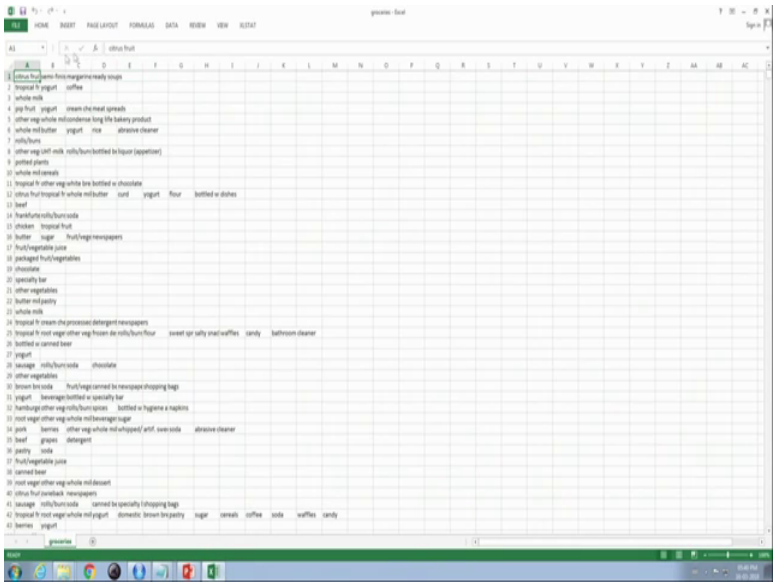
And type arules here arules here, it is displaying I selected the arules here. Now in install the arules packets from this terminal, after it is already installed in my system after installing the library installing the package, sorry, we have to include the library in R program for including library, just write library and type the package name and now I am including the library.

(Refer Slide Time: 22:24)



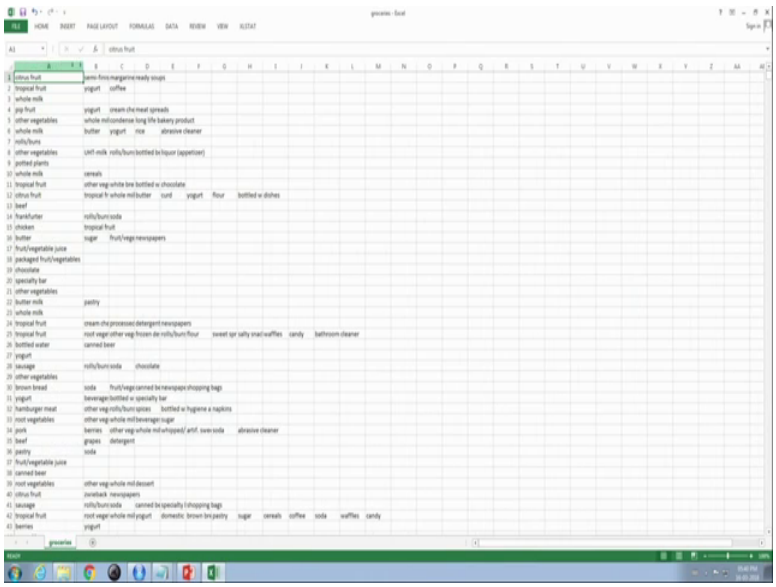
It is included; it is included smoothly

(Refer Slide Time: 22:52)



Then I have to read the read the grocery csv file, the grocery csv file is like that these are grocery csv file.

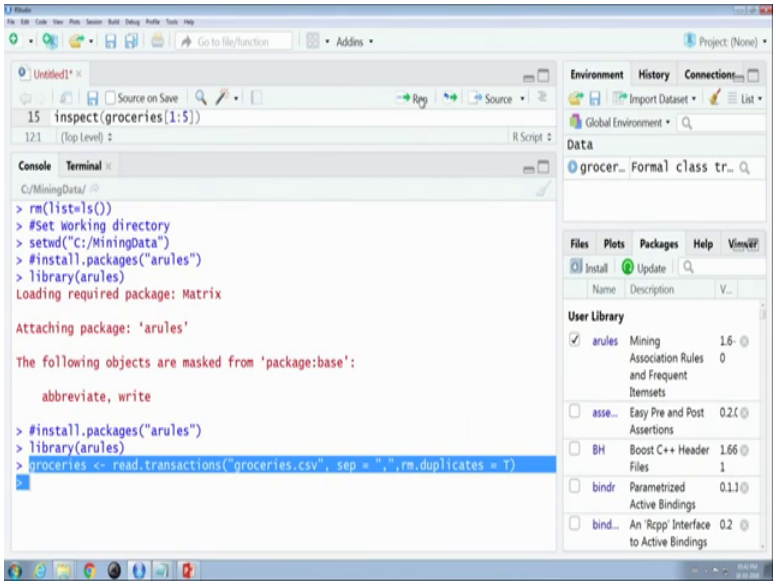
(Refer Slide Time: 22:53)



These are the items which are bought together these are transaction informations it contains the transactions of items this citrus fruit semi finished bread tropical food yogurt coffee are brought together in next transaction, only whole milk in next transaction pip fruit yogurt cream cheese these are the datasets my I have to find the item sets which are

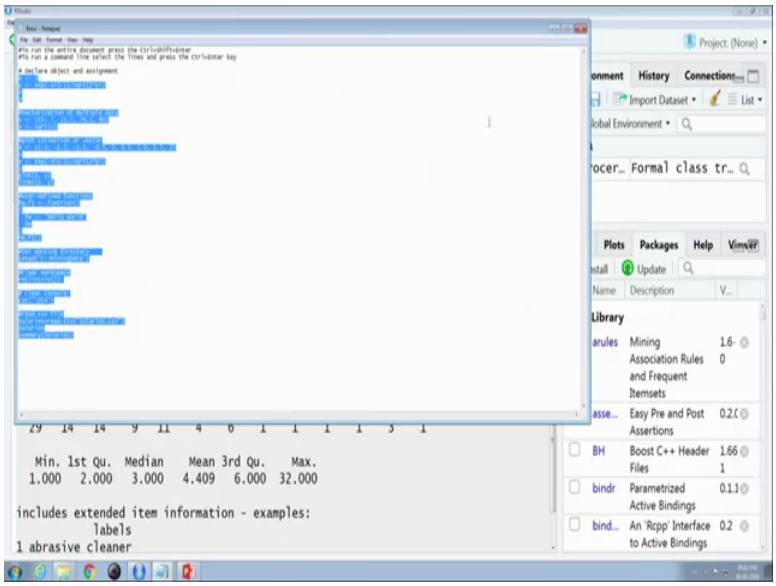
bought together next in most of time most frequently. Now I have to read the data set. Now dataset is readed the dataset is stored into the grocery.

(Refer Slide Time: 24:29)



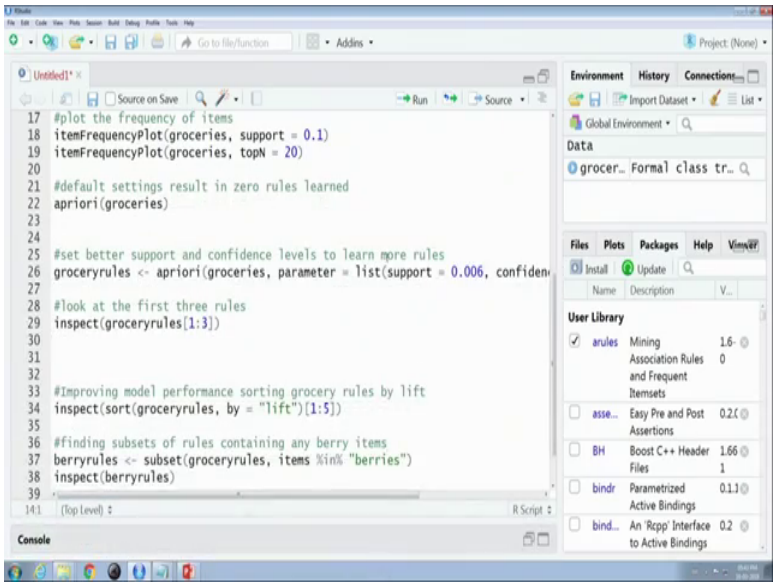
Then next it is showing the summary of the data set.

(Refer Slide Time: 24:53)

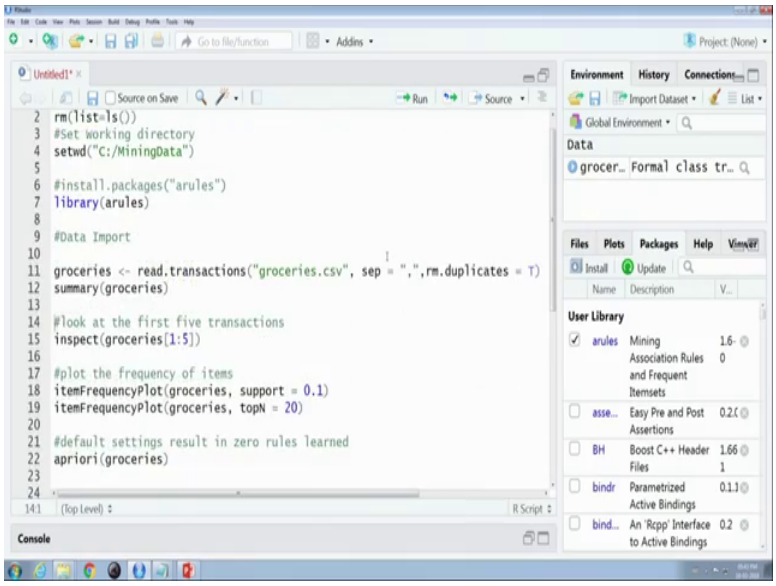


It contains 9835 rows and 169 columns because somebody in transaction, there are they have bought 169 items that is why it is showing here these are the most frequent items whole milk vegetables rolls and soda; these are more most frequently items and these are the size of distributions transactions.

(Refer Slide Time: 25:48)

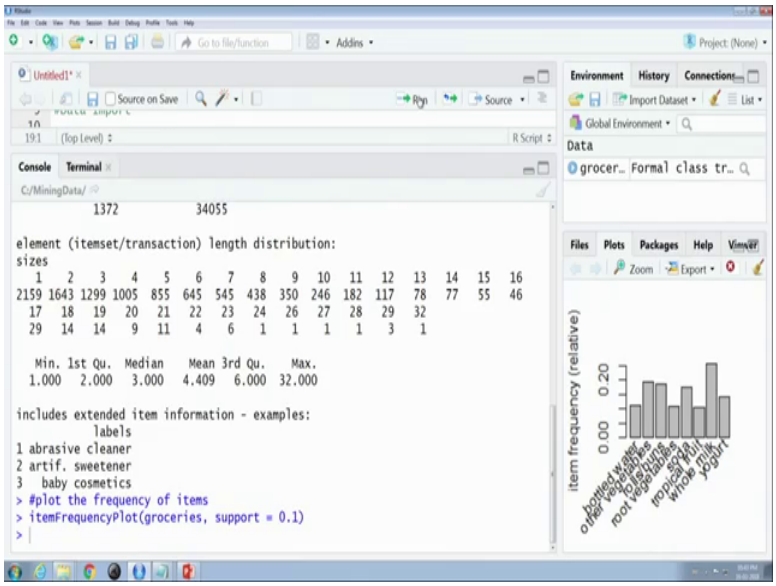


(Refer Slide Time: 25:49)

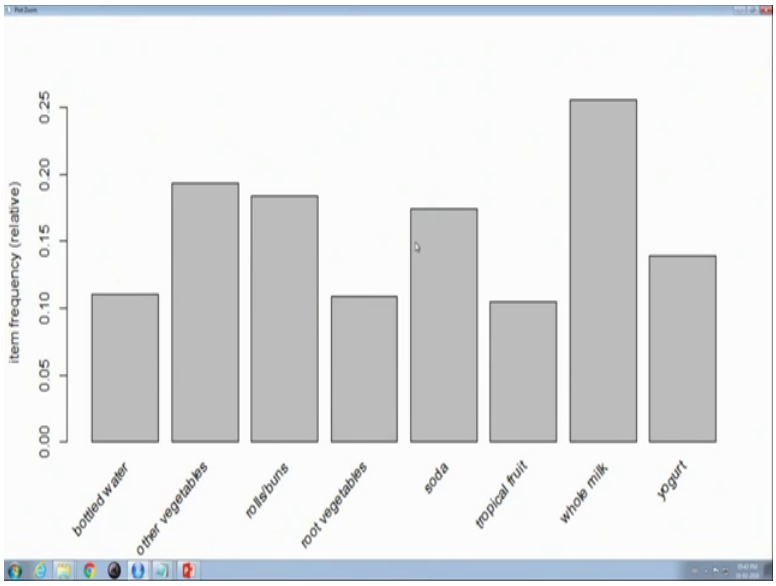


Now, and to plot the frequency of items these are frequency of items.

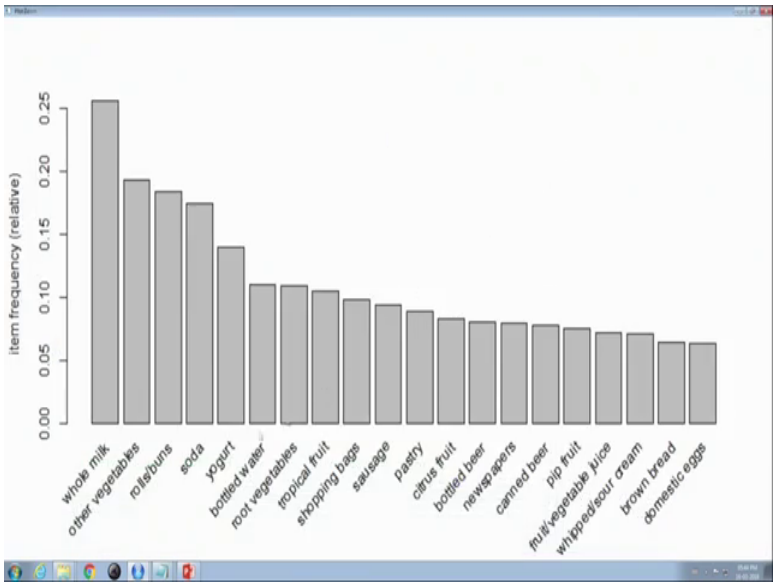
(Refer Slide Time: 26:01)



(Refer Slide Time: 26:05)



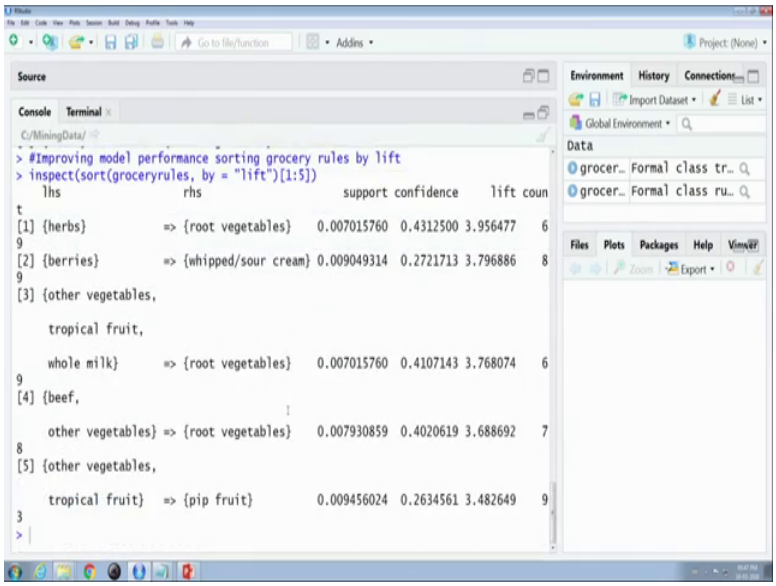
(Refer Slide Time: 27:20)



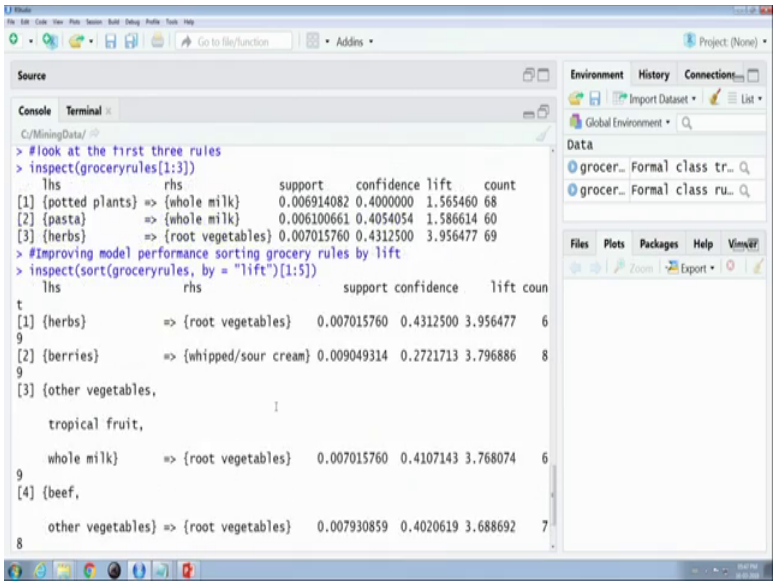
Item frequency bottled water other vegetables roll buns root vegetables these are the showing the frequency of items and to display frequency of plot this frequency of item top twenty items here is the top 20 items. Frequency of top 20 items, it is basically relative frequency in 0 to 1 scale, the most the whole milk is maximum frequency has a maximum frequency, the domestic is the minimum frequency. Now I apply the apriori. Now, apply now I have applied the apriori algorithm, it has confidence level initially the confidence is 0 8 minimum interval is 00.5 and now no rules has been generated in first iteration.

Now, we have to improve the grocery rules which support 0.6 and confidence 0.25 and minimum length is 2. Now the rules has been created, I have to inspect the rules. Now, it is showing the rules left hand side potted plants whole milk pasta depends on whole milk which support these are support. This is confidence and these are lift value, these are the rules. Now, this showing first 3 rules and to inspect by lift inspect the last last 5 rules by lift operation.

(Refer Slide Time: 29:50)

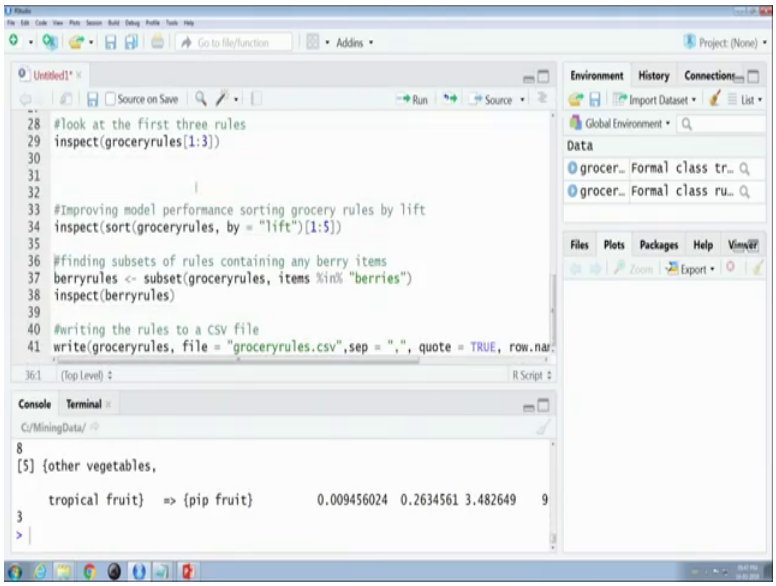


(Refer Slide Time: 29:51)

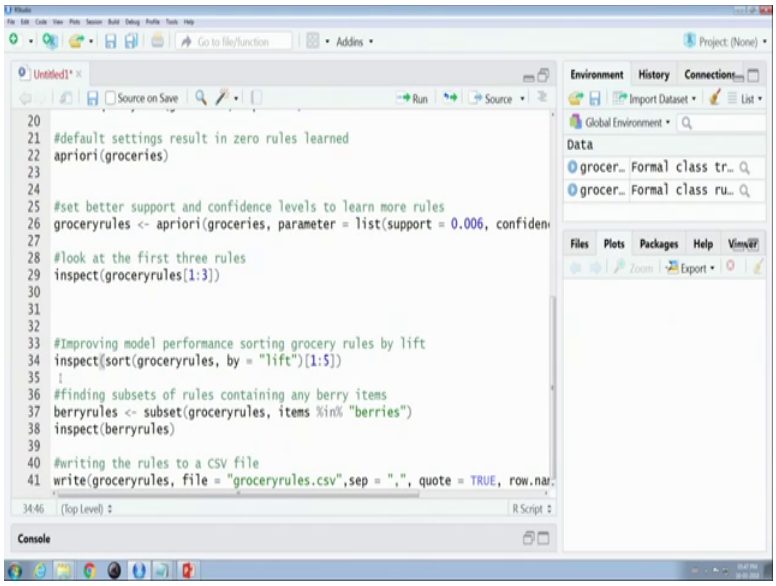


Now, it is showing herbs; if it is herbs, then root vegetables will be in right hand side, it is berries whipped and sour cream and tropical other vegetables tropical fruit whole milk root vegetables and so on.

(Refer Slide Time: 30:18)

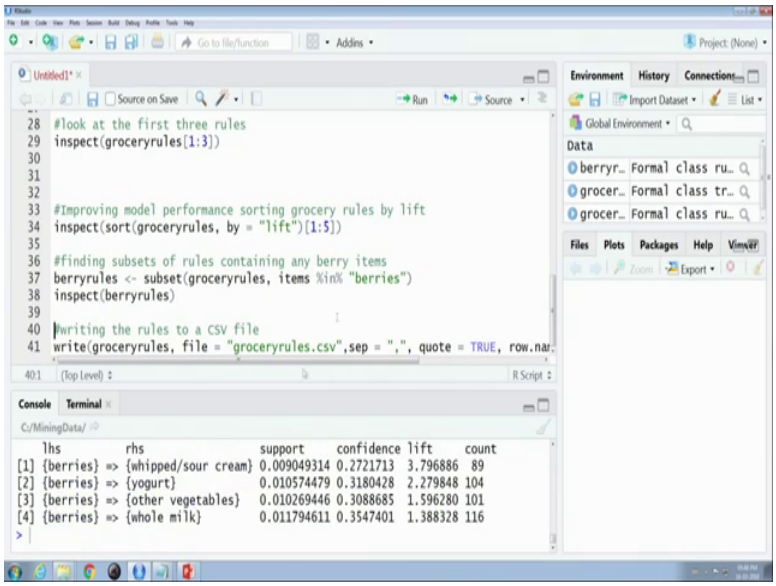


(Refer Slide Time: 30:25)



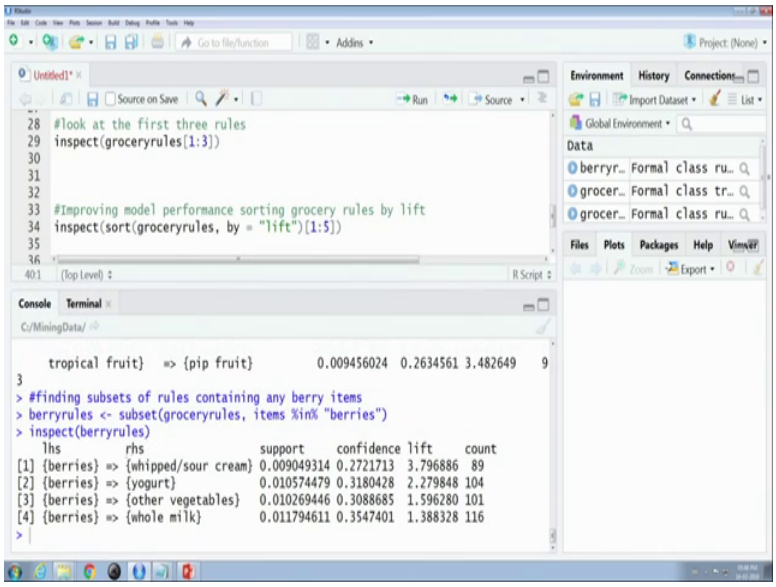
Now, I want to find a subset of rules containing any various items and how for that I have to use the subset function these is the it contains the rules and item is varies. Now I just execute this command after executing this command, I have to inspect this command.

(Refer Slide Time: 30:50)



Now it is showing rules with berries.

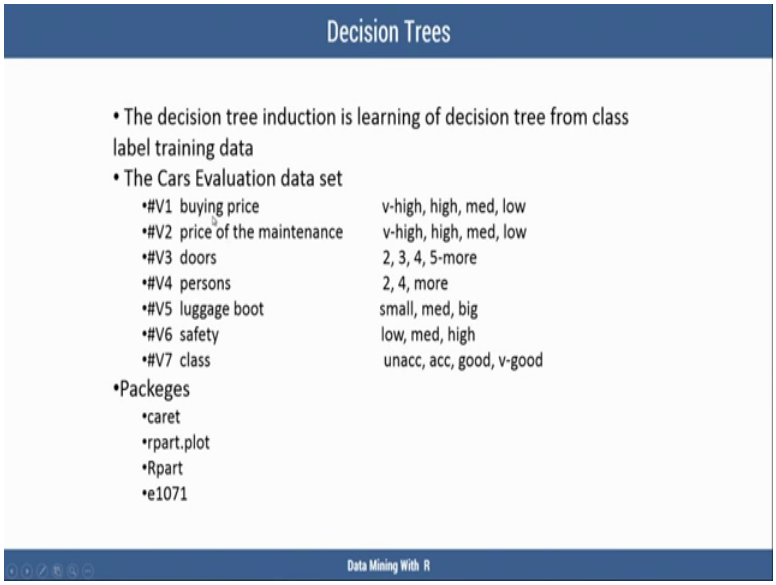
(Refer Slide Time: 30:53)



Berries whipped and sour cream berries and yogurt with confidence level these berries and other vegetables and berries and whole milk, these are the most frequently bought items with berries. Now I have to write the all the rules into a particular file for that I have to use right command, right the grocery rules into the file grocery rules csv. Now I am now trying to execute this. Now the file has been created. Now I have to check the file this is the rule files here inter rule files has been saved.

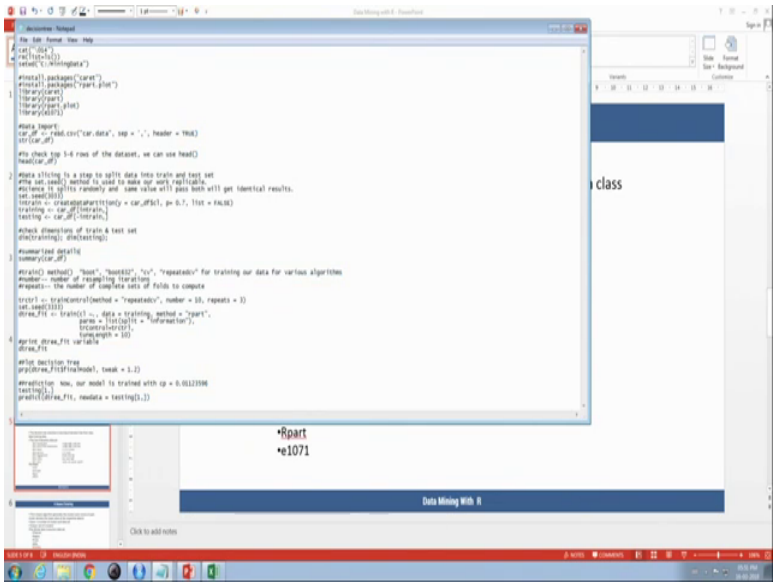
Now, this is the procedure of how to generate association rules and save into the csv file the library is arules.

(Refer Slide Time: 32:53)



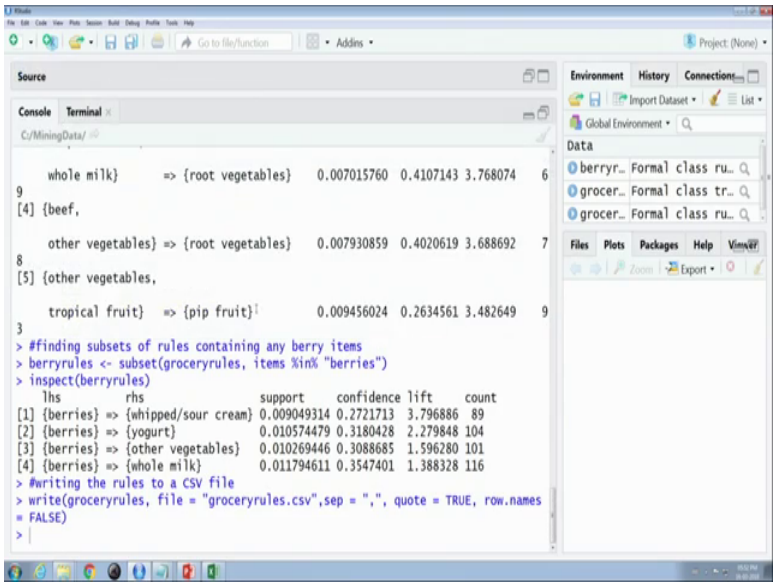
Now, I am going to show; how to generate the decision tree for creating decision tree we have to use 4 packeges; one is caret another is Rpart dot plot and for decision tree generation e1071 and Rpart e1071 is the most important packege for creating decision trees in I am using one data set; this name is car evaluation data sets, it contains 5 7 attributes price of the car maintenance value of the car doors of the car number of persons can sit into the car the booted the luggage area of the car the safety matters of the car and class; the class is either unsatisfactory, it is satisfactory good or very good these are the class now you want to make decision tree based on this information.

(Refer Slide Time: 34:05)

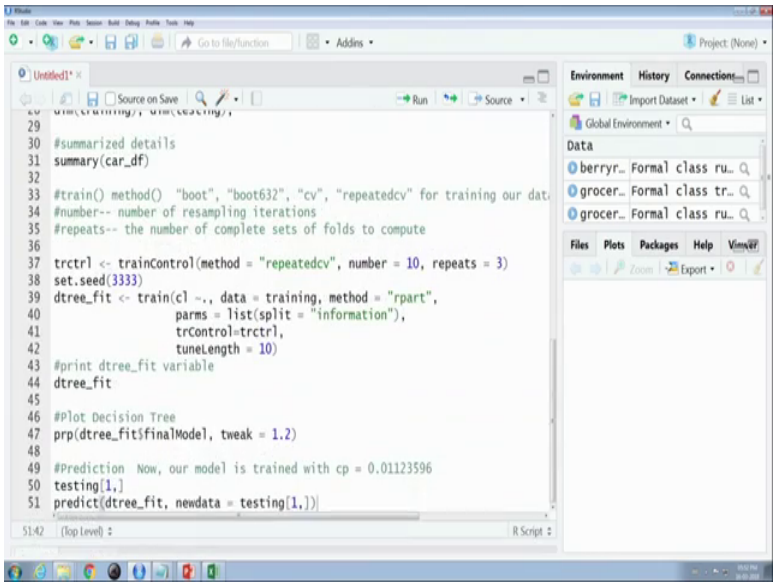


After generating the decision tree; we are we will predict one value.

(Refer Slide Time: 34:36)

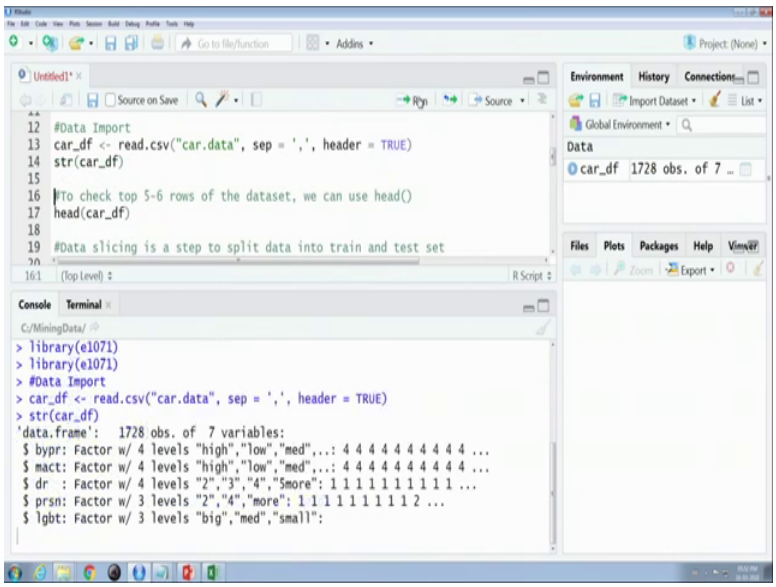


(Refer Slide Time: 34:42)



Now, I want to add the library caret Rpart Rplot and e701 e1071. Now, I have to read the data from card data, then I have to view the structure of the data set.

(Refer Slide Time: 35:26)



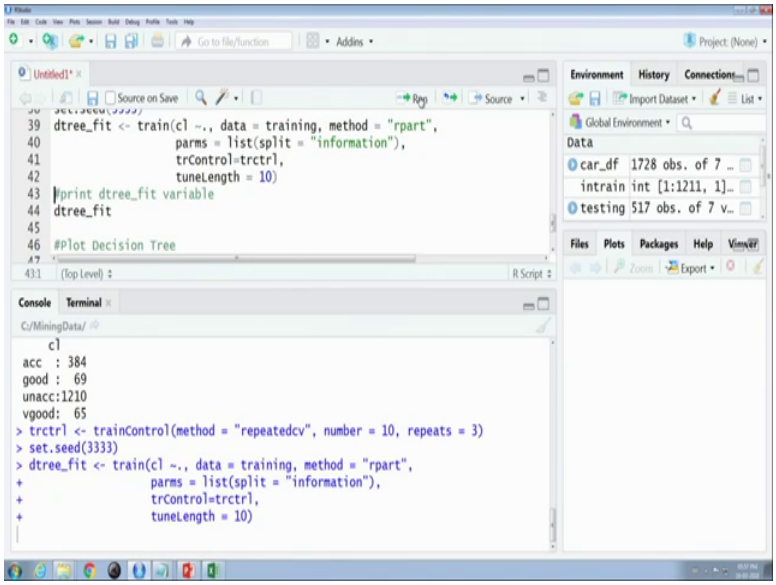
It is showing the structure of the data set it has 4 levels bypr means buying price mact means the maintaining cost, it has 4 levels door number of doors peers number of persons here these are the structure of the data set, then I am to display, the first few data set and it is displaying the data set 6 data sets, then I have to divides the data set into 2 parts 30 percent for testing the decision tree and 70 percent for training purpose that is

why I am creating a party one partition based on class column; cl is class based on class column and 0.7 means seventy percent of data will be used for training purpose list is false.

Now, at at first, I have to make a seed on random value, then I have generated a training data set. Now, training data set has been created. Now test data set; these are the test data set. Now I have to check the dimension of the training data set. Now, it is show showing the dimension of the training data set and testing data set the training data set contains 1211 records and test data set has 5 517 records with 7 attributes.

Here is the summary of the car data sets. Now, I have to apply the training with number of 10 folds number of 10 folds number of resampling iterations, these are number and repeats means the number of complete sets of folds to compute I am here using repeated cv method for training, there are lots of method available boot method boot six 32 method cv method repeat cv method for training our data. Now execute training make one random variable, then then I have to create the tree these function training these are training function and it will create store the tree into the tree feet this is a class level data is training data I am using Rpart method is the transaction control tune length is 10; 10 is the number of resampling iterations.

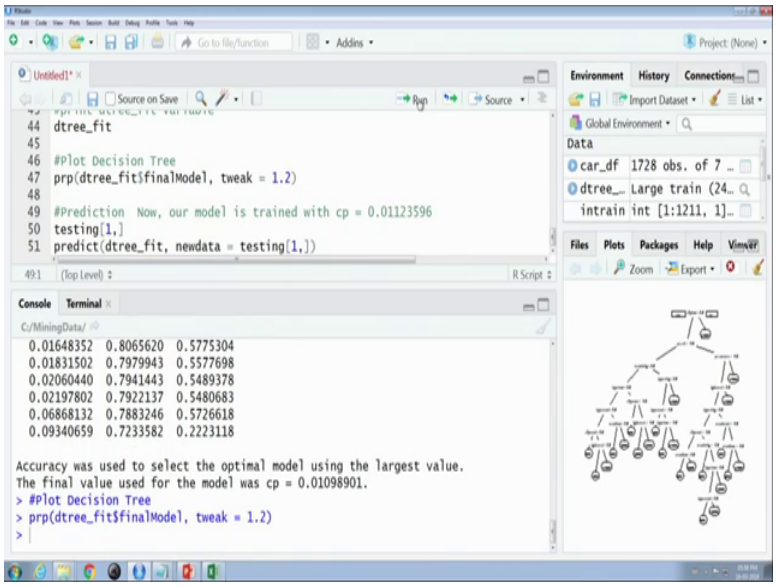
(Refer Slide Time: 39:52)



Now, you have to execute the train data to create the tree. Now, tree is created; now I have to get that tree information; these are the data of a this is the complexity parameters

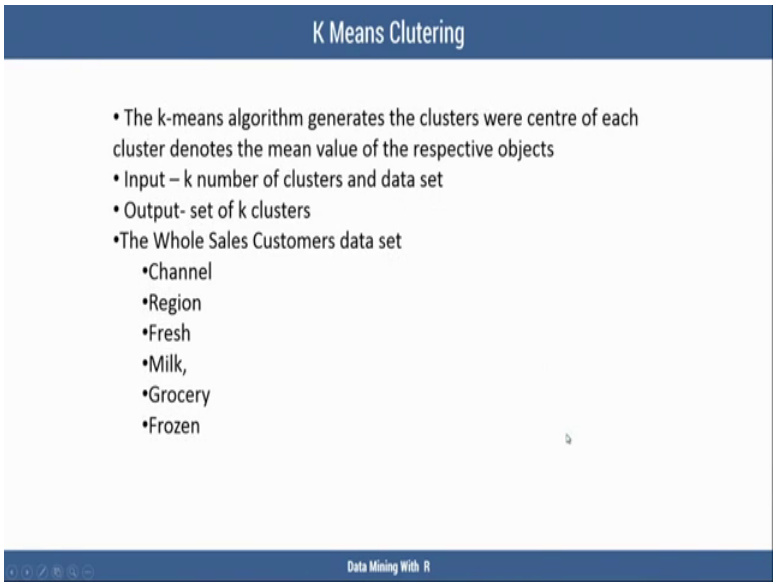
this is the accuracy and kappa value and it is showing it is using cart algorithm, there are 6 predictors and 4 classes. Now I want to plot the tree plot the decision tree; these are decision tree, but it will be due to graphics problem in this machine, it will not displaying properly. Now we have to test one dataset. Suppose, these one testing data set the testing dataset showing buying price is very high maintenance cost is very high and class is unsatisfactory these my testing data whether these data is correctly predicted or not now.

(Refer Slide Time: 40:41)



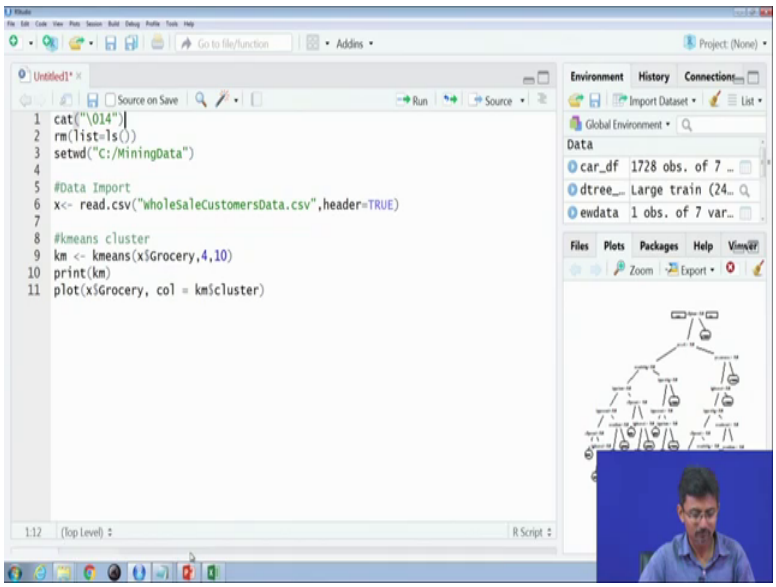
I am now I am predicting the same data here new data set predict dtree the data record is this now; now it is predicting correctly because this unsatisfactory and in my dataset it is also unsatisfactory here these both are same.

(Refer Slide Time: 42:31)



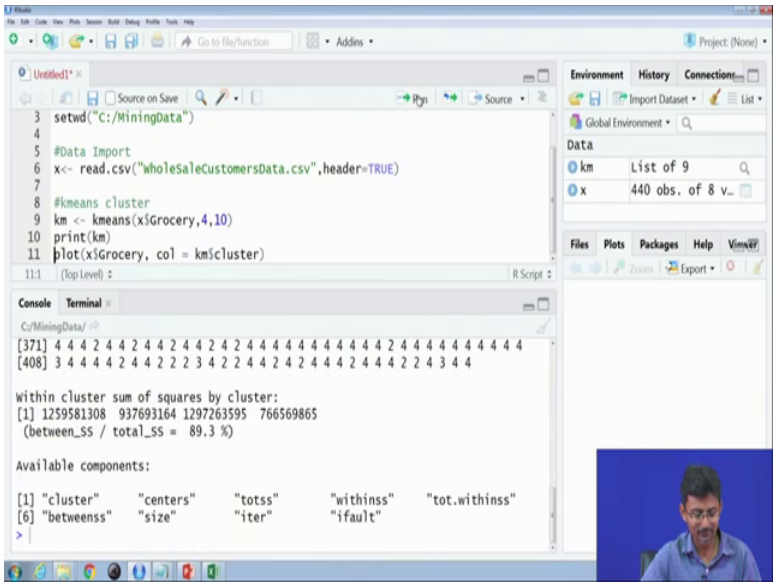
Now, I want to show how to apply R program to use k means clustering the k means clustering generates cluster where the centre of each cluster denotes mean value of respective objects the input set of k means algorithm the number of clusters you want to generate and the data set and it will create k k clusters in my data set, I am using 1 wholesale customers data set in that data set, there are lots of attributes; I am I am trying to the attributes are channel regions fresh milk grocery I will generate the clusters on grocery attribute.

(Refer Slide Time: 43:59)



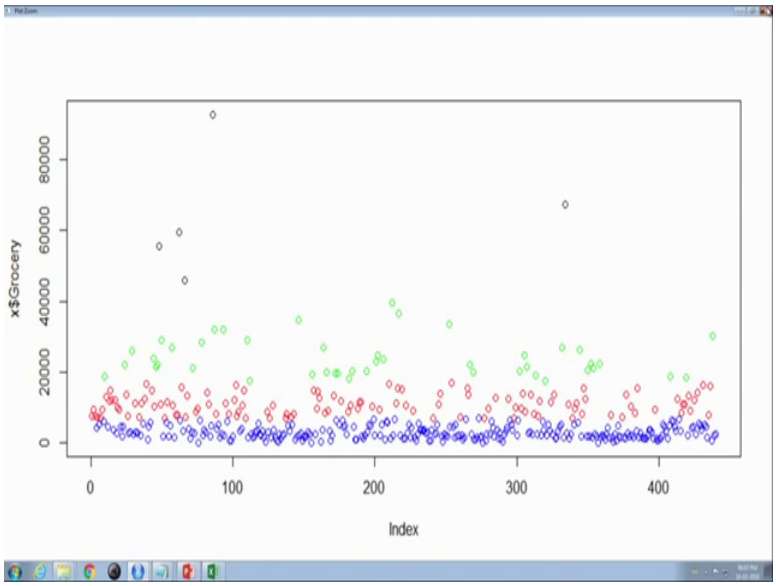
It is very simple program at first, we have to read the data from the wholesale customer and data is stored in x very well now and to generate k means cluster on grocery data with 4 clusters and with number of iterations will be 10. Now, I am executing the k means algorithm.

(Refer Slide Time: 45:06)



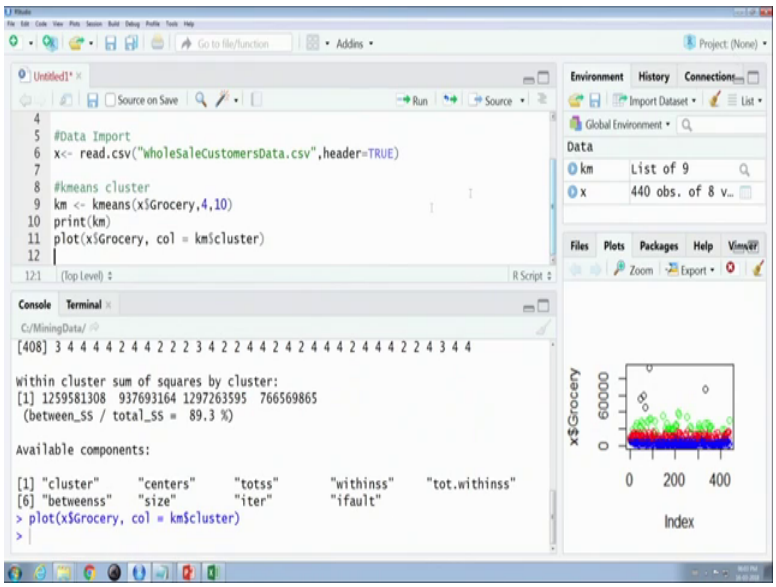
It will be the result will be stored in km objects and I want to print the [Noise] data it contains 4 clusters with these are the mean values of 4 clusters, these are the cluster vectors within cluster some squares by clusters are this some of the objects of these clusters these are the available components.

(Refer Slide Time: 45:54)



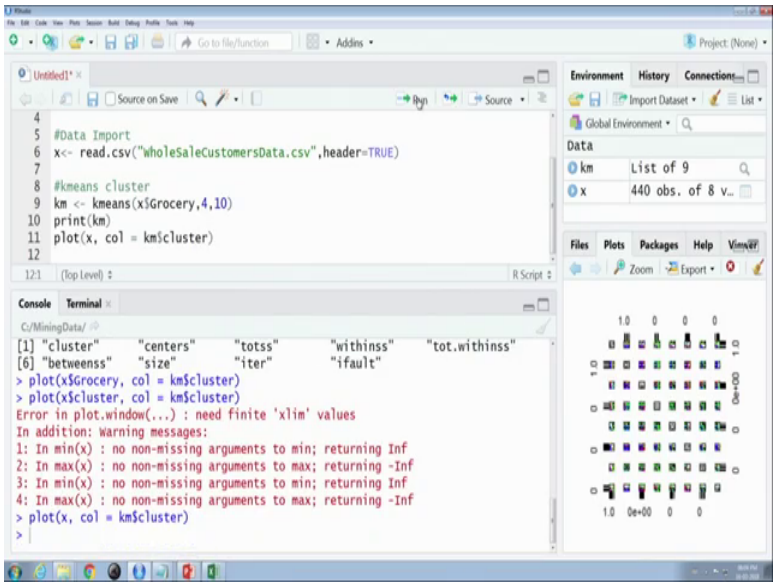
Now, we want to generate the cluster on grocery. Now it is showing the clusters on grocery; there are 4 clusters this blue, one is one cluster, red one is another cluster and green one is another cluster and very small; these are very small cluster, it scattered basically.

(Refer Slide Time: 46:17)

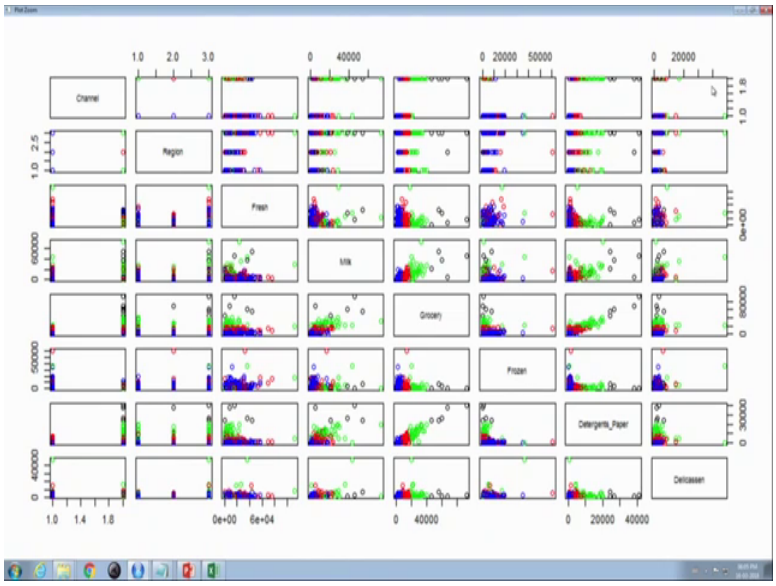


Suppose, I want to display inter cluster all the clusters of all attributes now it is showing all the cloud clusters of all attributes here these are clusters of channel attributes these are clusters of where is the grocery is the clusters of grocery attributes.

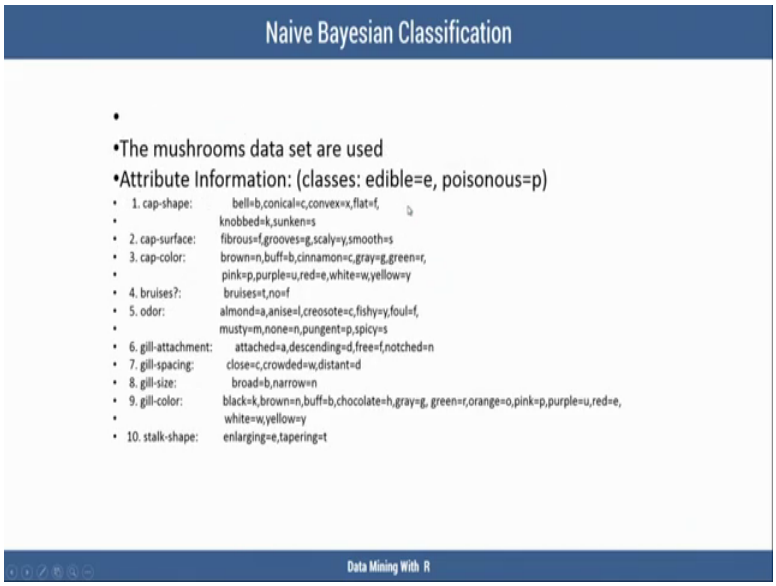
(Refer Slide Time: 47:27)



(Refer Slide Time: 47:31)

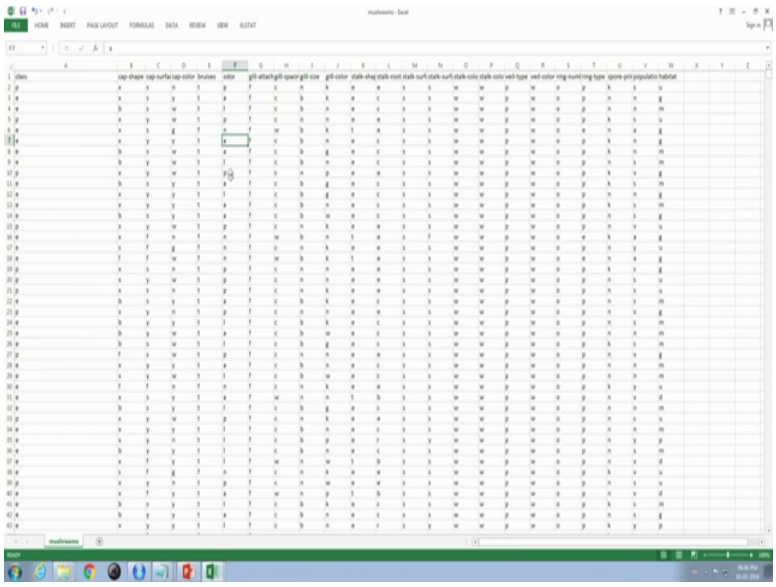


(Refer Slide Time: 48:03)



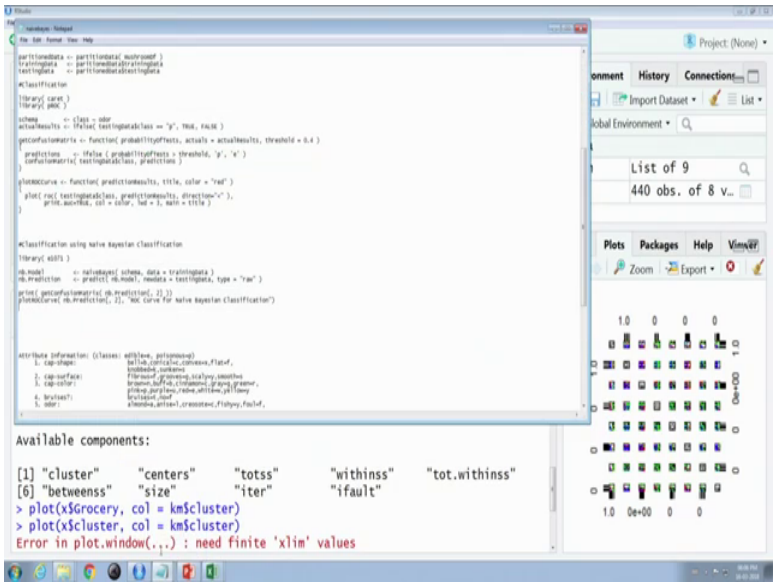
Now, Naive Bayesian classification Naive Bayesian classification, we classify the data set based on the probability Naive based theorem and I am using one table it contains mushroom data set its very big table it contains one class.

(Refer Slide Time: 48:36)

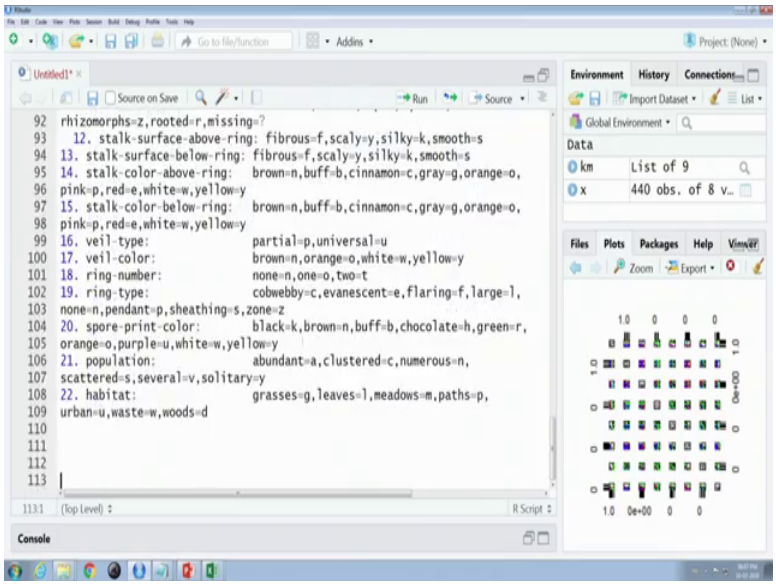


Whether; it is eatable or poisonous the class are e and p whether it is eatable and poisonous these are the characteristics of mushrooms these are the characteristics values just like odor; odor means odor odor type of odor.

(Refer Slide Time: 49:17)

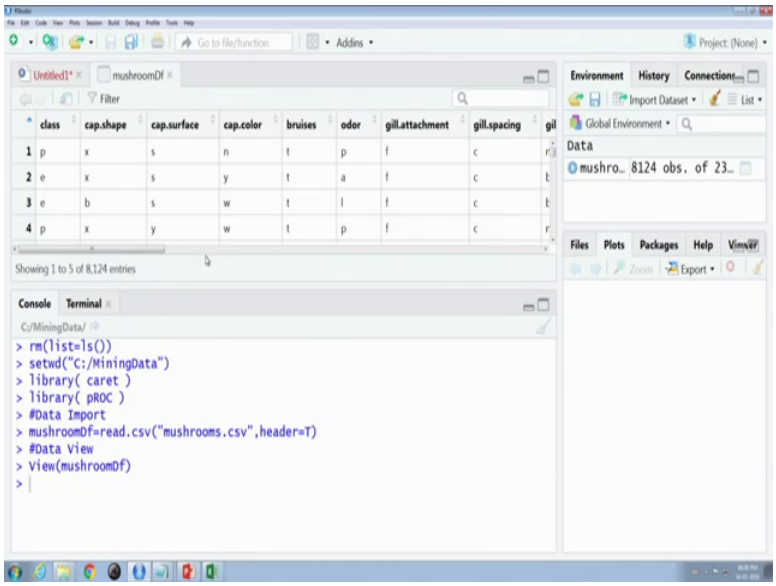


(Refer Slide Time: 49:32)



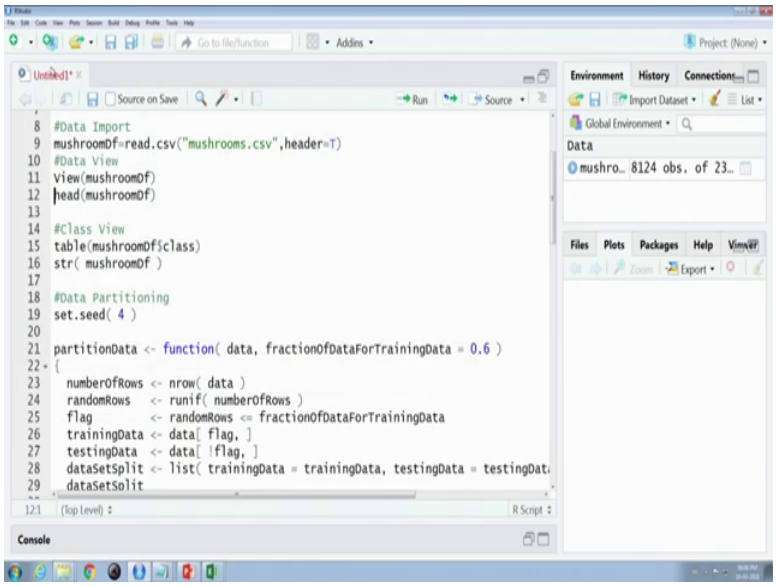
It contains lots of attributes cap shape cap surface cap colour, these are the abbreviations of the table, these are the abbreviations of the tables and two classes eatable or poisonous to execute Naive Bayesian algorithm, we need to include two libraries one is Carret and the other one is proc library, then I have to import data same way.

(Refer Slide Time: 50:45)



Similarly, then want to view the mushroom datasets this is the same data.

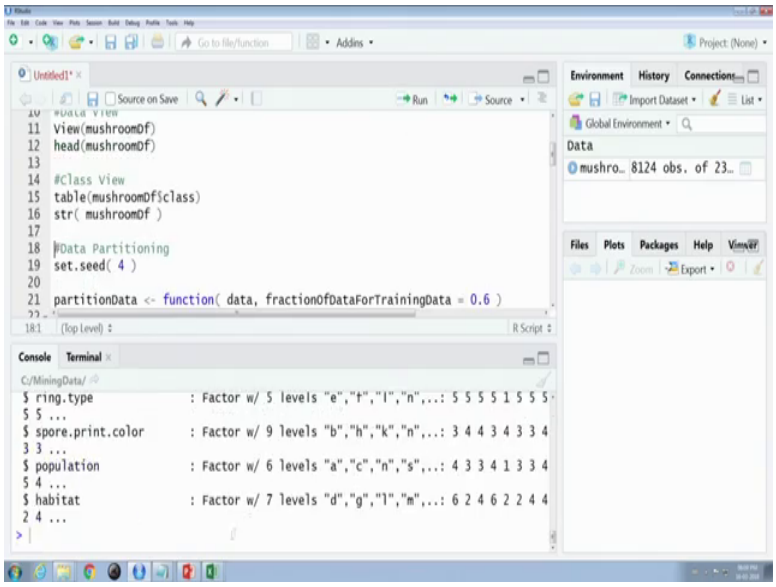
(Refer Slide Time: 50:59)



These are the first few first 6 data sets, it is very big table that is why it is displaying this way first six in second second set of attributes these are third set of attributes these are the fourth set of attributes this because there are lots of attributes that is why it is displaying this way.

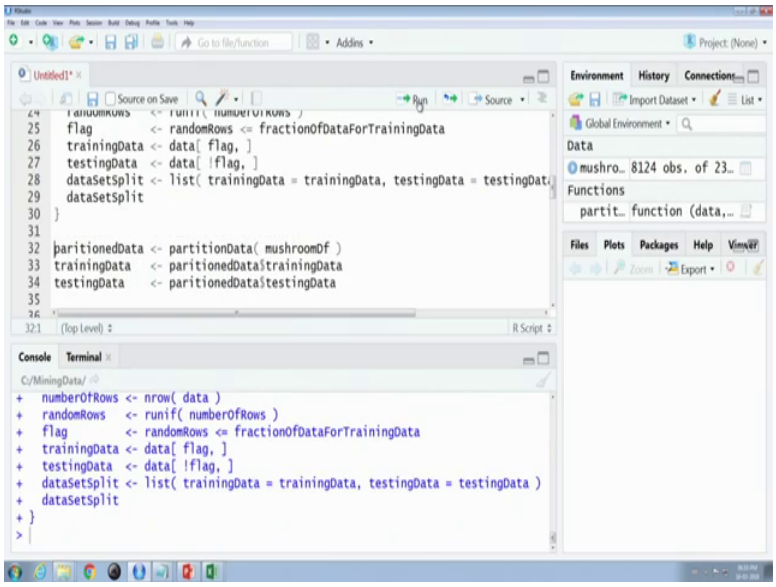
Now, I want to view the class table; there are 2 classes eatable and poisonous; there are 4000; around 4000 classes mushrooms are edible around 3000 classes are poisonous mushrooms.

(Refer Slide Time: 52:05)



what is the structure of the; these are these is the structure of the data sets there are twenty three variables in the data sets and around eight thousand observations, then for again have to partition the data for training set and test set.

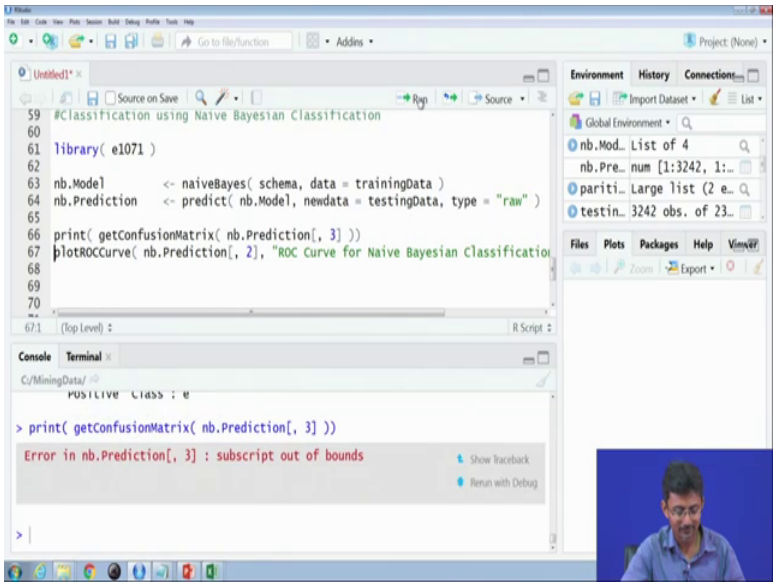
(Refer Slide Time: 52:33)



Now training set is being generated now test set is generated now to make a schema class and odor class versus odor schema. Now testing the data set on class equal to p actual results now have to develop the confusion matrix with the probability test with the function probability test. Now confusion matrix has been generated. Now I have to make prediction curve. Now, execute the naive Bayesian algorithm on training data and it will be stored in nb dot model and this will be the predicted data.

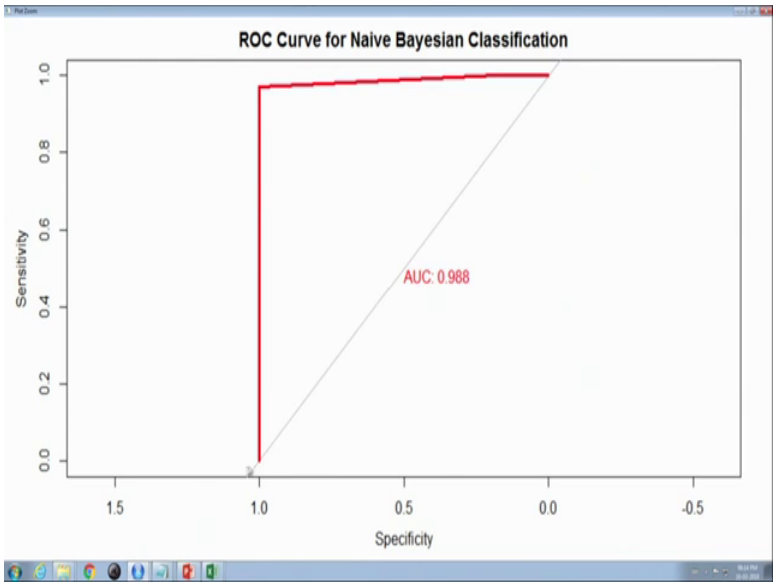
Now, prediction on any model testing data type is raw now and to print the confusion matrix the prediction two prediction data two see for prediction data 2, this is the confusion matrix references is e to e is and with accuracy 0.9851; the class is edible here.

(Refer Slide Time: 55:33)



Suppose, I want to check another data this is the first data no there is no; now I want to plot the ROC curve or Bayesian matrix prediction matrix. Now this is the curve these ROC curve for Bayesian classification is a specificity and sensitivity curve.

(Refer Slide Time: 56:32)



Therefore, for executing the Naive Bayesian algorithm at first, we have to partition the data into training set and test set then make a schema in this class with an particular attribute then the actual results based on poisonous here then generate the confusion matrix after that execute the Naive Bayesian algorithm these are procedure ok.

Thank you.