WELCOME

STATISTICAL DATA ANALYSIS ON IRIS DATASET USING R

Project Guide:

K NAVEEN KUMAR

Project by:

G SHIVA RAM

AGENDA

- Introduction to Iris Data
- About data set
- Data collection
- Overview of the dataset
- About software
- Methodology
- Data visualization
- conclusion

Introduction to Iris data

Sample Data

Variables

Observations

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3	1.4	0.2	setosa
4.7	3.2	1.3	0.2	Versicolor
4.6	3.1	1.5	0.2	Virginica
5	3.6	1.4	0.2	Virginica

Why Iris Data?

The Iris Data set was introduced by Robert Fisher. It consists of four measures i.e. Length and Width of Sepals and Petals for three flower species (namely Setosa, Versicolor and Virginica).

1. Easily available and lot of support material available online
2. All types of machine learning algorithms can be easily implemented on this data
3. Caution : Be aware that the results on real-life datasets are not going to be as optimistic.

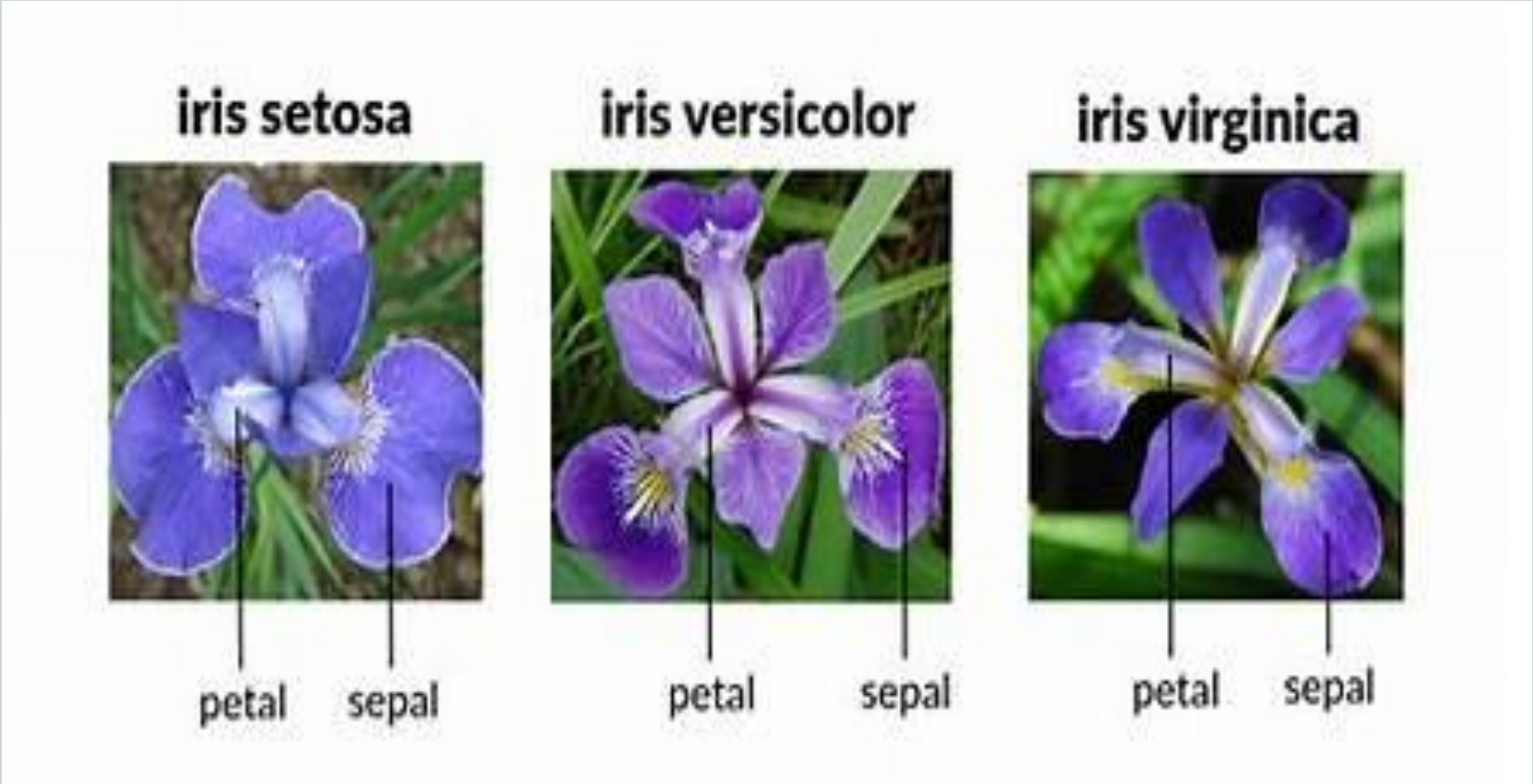
ABOUT THE IRIS DATA

The iris dataset contains the following data

We have 50 samples of 3 different species of iris (150 samples total)

- Here we can see that given 4 features i.e sepal length, sepal width, petal length, and petal width determine whether a flower is Setosa, Versicolor or Virginica.
- Sepal length, Sepal width, Petal length, Petal width are called feature/Independent-variable.
- Species are called Labels/Dependent-variable.

TYPES OF IRIS FLOWERS



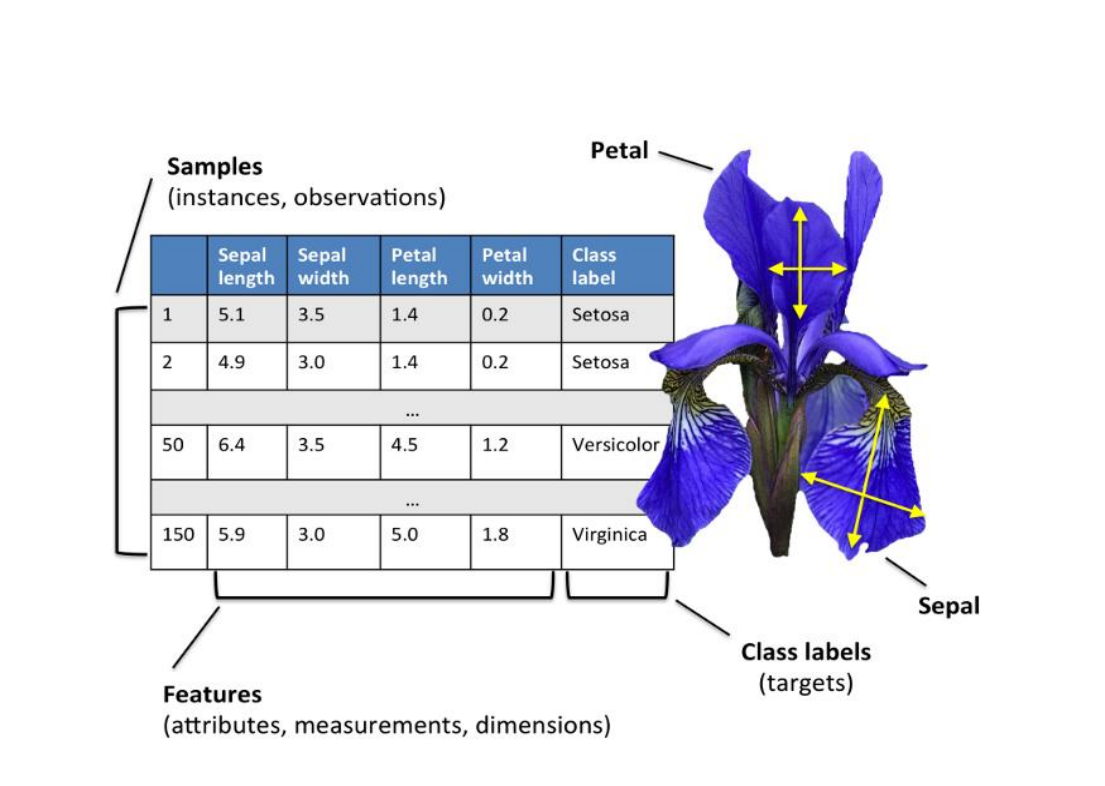
ATTRIBUTE INFORMATION

- Four features of flower: **length** and the **width** of sepal and petal

Sepal length in cm
Sepal width in cm
Petal length in cm
Petal width in cm

<u>class:</u>

1) Iris Setosa
2) Iris Versicolour
3) Iris Virginica



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

Data Collection

- **UC Irvine Machine Learning Repository**

If you want to download the data set instead of using the one that is built into R, you can go to the UC Irvine Machine Learning Repository and look up the Iris data set.

Overview Of The Data Set

- First, you can already try to get an idea of your data by making some graphs, such as boxplots. In this case, however, scatter plots can give you a great idea of what you’re dealing with: it can be interesting to see how much one variable is affected by another.
- In other words, you want to see if there is any correlation between two variables.
- Review basic descriptive statistics in R
- To display summary statistics for each feature available in dataset.
- The linear combination of original variables that provide the best possible separation between the groups.

About Software

What is R & Why we use it.

- It's a tool : Open-Source, cross platform, free programming language designed to build statistical solutions
- Powerful : Gives access to CRAN repository containing over 10,000 packages with pre-defined functions for almost every purpose
- Stays Relevant : Constantly being updated by users (Scientists, Statisticians, Researchers, Students!)
- More: Makes beautiful graphs, can create custom functions or modify existing ones, can be integrated into many environments and platforms such as Hadoop etc.

Installing R

- Download the version compatible with your OS
- Can be downloaded for free from

  http://www.r-project.org/

- Simple/Standard installation process

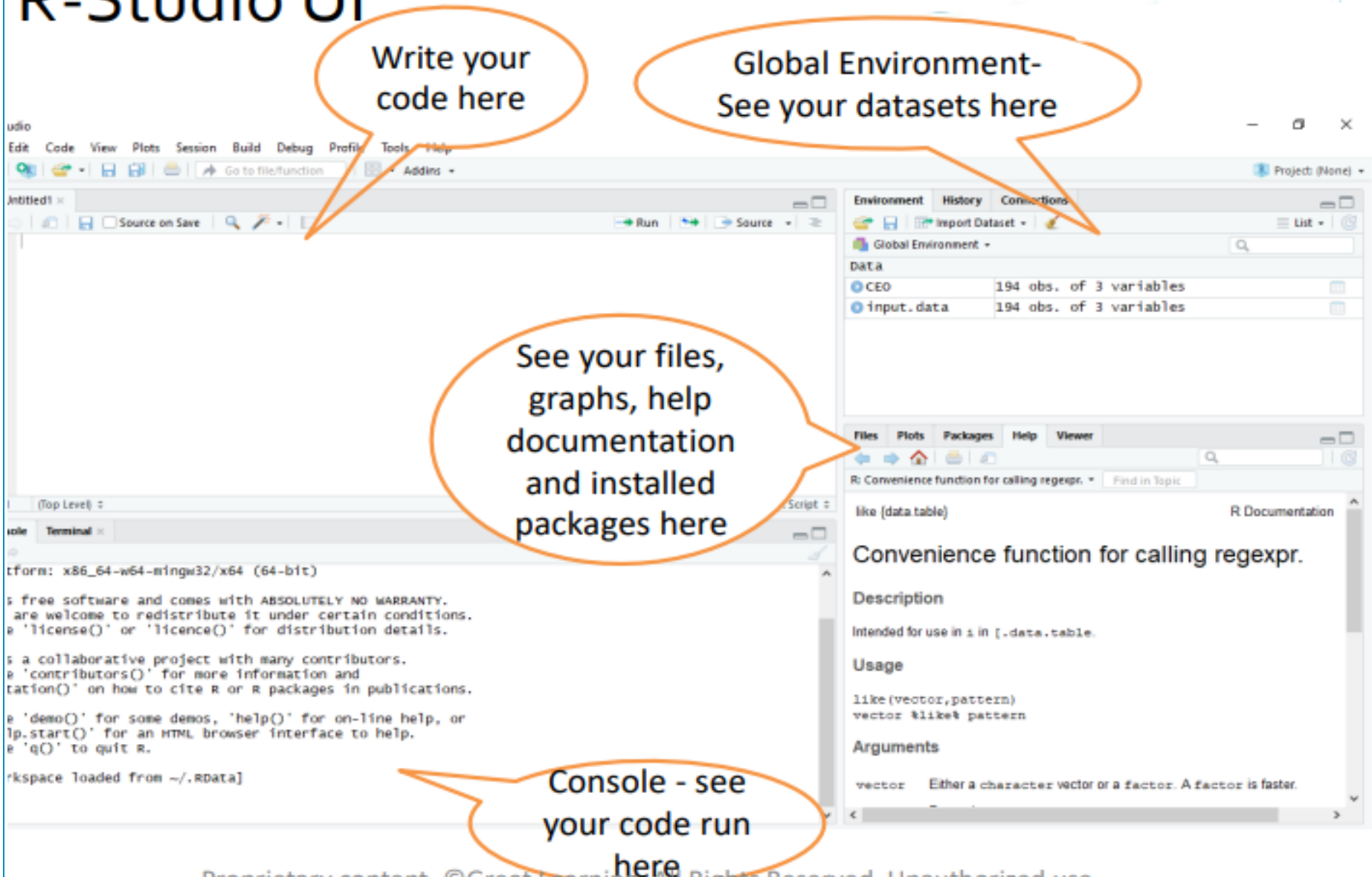
Installing R -Studio :

- Can be downloaded for free from:

https://www.rstudio.com/products/rstudio/download/

- Download the free version compatible with your OS
- R needs to be installed before installing R- Studio

R-Studio UI



METHODOLOGY

- Enter the data in excel sheet & save the data with the **read.csv()**– It is used to read csv files and create a data frame from it.
- We import iris data by giving path of data file of “iris.csv”.
- Iris = **read.csv**(“iris.csv”)
- Import the data from excel to R & Explore the *iris* Dataset with R.
- Let's get started by importing all the libraries that we are going to need.
- Importing Packages: **"ggplot2","dplyr",”MASS”.**

DATA EXPLORATION AND VISUALISATION

Some basic function in R to examine iris dataset:

- **data**(iris)

	A	B	C	D	E
1	sepal_length	sepal_width	petal_length	petal_width	species
2	5.1	3.5	1.4	0.2	setosa
3	4.9	3	1.4	0.2	setosa
4	4.7	3.2	1.3	0.2	setosa
5	4.6	3.1	1.5	0.2	setosa
6	5	3.6	1.4	0.2	setosa
7	5.4	3.9	1.7	0.4	setosa
8	4.6	3.4	1.4	0.3	setosa
9	5	3.4	1.5	0.2	setosa
10	4.4	2.9	1.4	0.2	setosa
11	4.9	3.1	1.5	0.1	setosa
12	5.4	3.7	1.5	0.2	setosa
13	4.8	3.4	1.6	0.2	setosa
14	4.8	3	1.4	0.1	setosa
15	4.3	3	1.1	0.1	setosa
16	5.8	4	1.2	0.2	setosa
17	5.7	4.4	1.5	0.4	setosa
18	5.4	3.9	1.3	0.4	setosa
19	5.1	3.5	1.4	0.3	setosa
20	5.7	3.8	1.7	0.3	setosa
21	5.1	3.8	1.5	0.3	setosa

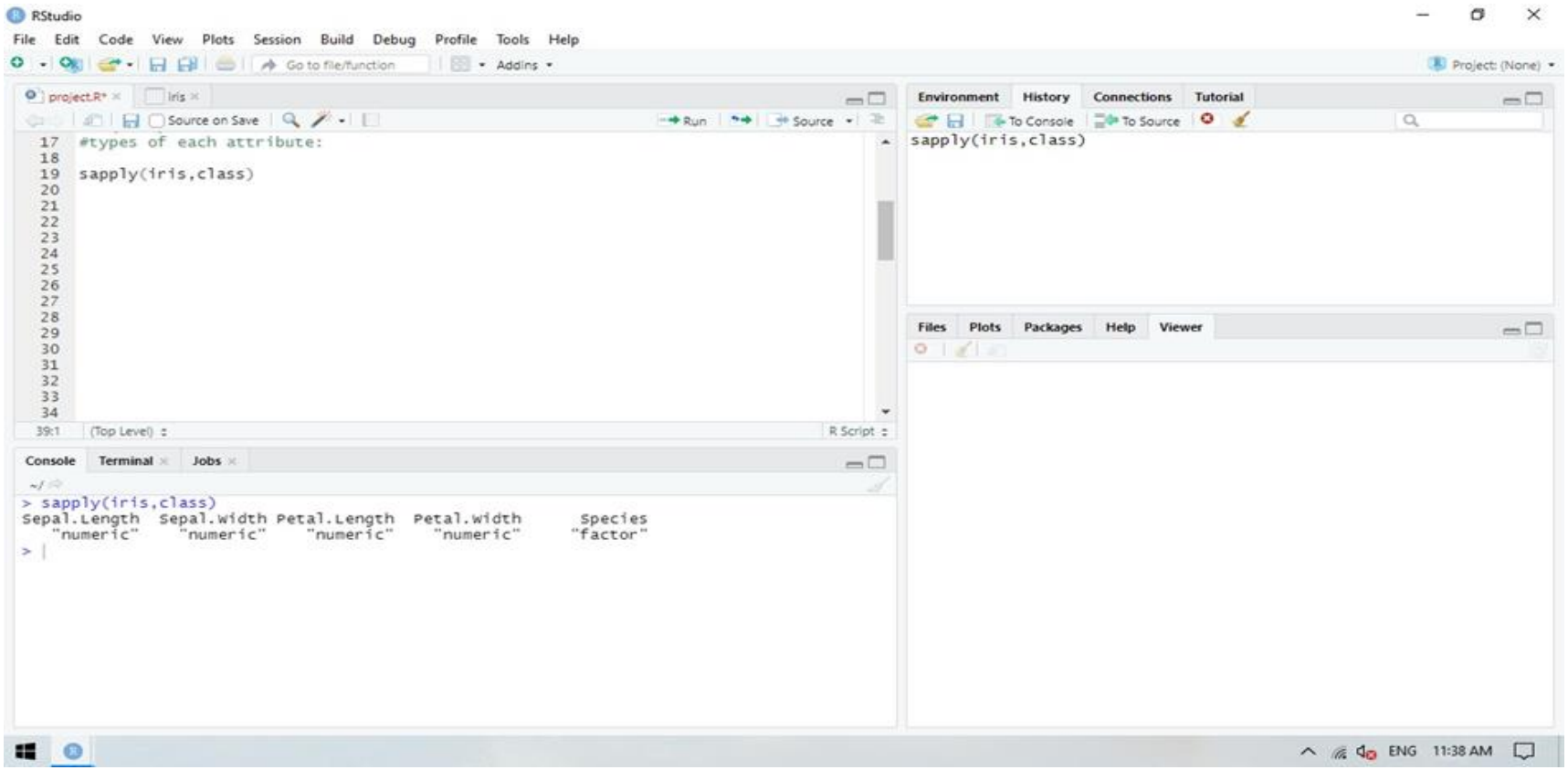
	A	B	C	D	E
1	Sepal Length	Sepal Width	Petal Length	Petal Width	Class
2	5.1	3.5	1.4	0.2	Iris-setosa
3	4.9	3	1.4	0.2	Iris-setosa
4	4.7	3.2	1.3	0.2	Iris-setosa
5	4.6	3.1	1.5	0.2	Iris-setosa
6	5	3.6	1.4	0.2	Iris-setosa
7	5.4	3.9	1.7	0.4	Iris-setosa
8	4.6	3.4	1.4	0.3	Iris-setosa
9	5	3.4	1.5	0.2	Iris-setosa
10	4.4	2.9	1.4	0.2	Iris-setosa
11	4.9	3.1	1.5	0.1	Iris-setosa
12	5.4	3.7	1.5	0.2	Iris-setosa
13	4.8	3.4	1.6	0.2	Iris-setosa
14	4.8	3	1.4	0.1	Iris-setosa
15	4.3	3	1.1	0.1	Iris-setosa
16	5.8	4	1.2	0.2	Iris-setosa
17	5.7	4.4	1.5	0.4	Iris-setosa
18	5.4	3.9	1.3	0.4	Iris-setosa
19	5.1	3.5	1.4	0.3	Iris-setosa
20	5.7	3.8	1.7	0.3	Iris-setosa
21	5.1	3.8	1.5	0.3	Iris-setosa
22	5.4	3.4	1.7	0.2	Iris-setosa
23	5.1	3.7	1.5	0.4	Iris-setosa
24	4.6	3.6	1	0.2	Iris-setosa
25	5.1	3.3	1.7	0.5	Iris-setosa

attributes() :It shows attributes of iris data

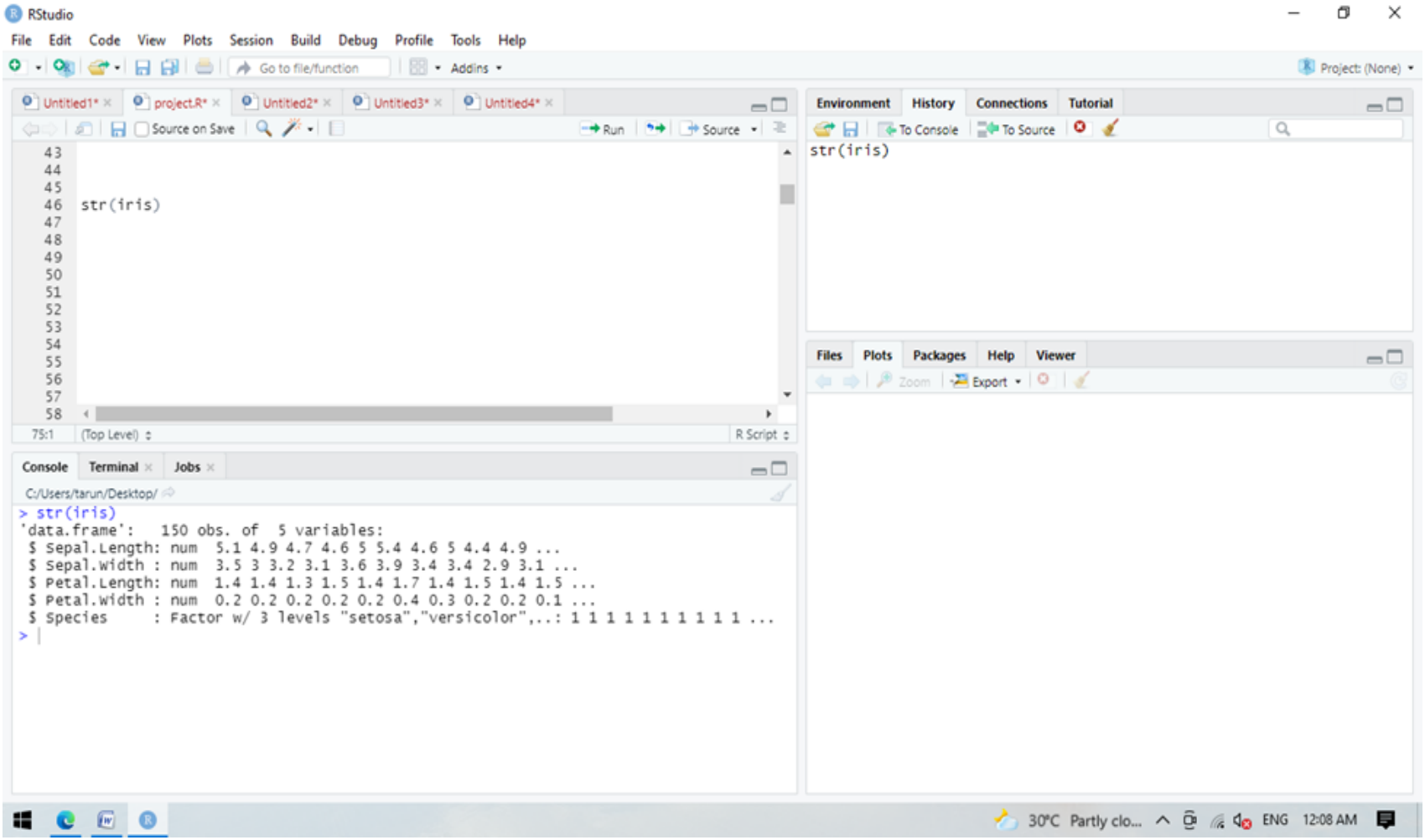
```
> attributes(iris)
$names
[1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"  "Species"

$class
[1] "data.frame"

$row.names
  [1]   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28
 [29]  29  30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56
 [57]  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84
 [85]  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99 100 101 102 103 104 105 106 107 108 109 110 111 112
[113] 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140
[141] 141 142 143 144 145 146 147 148 149
```

sapply ():

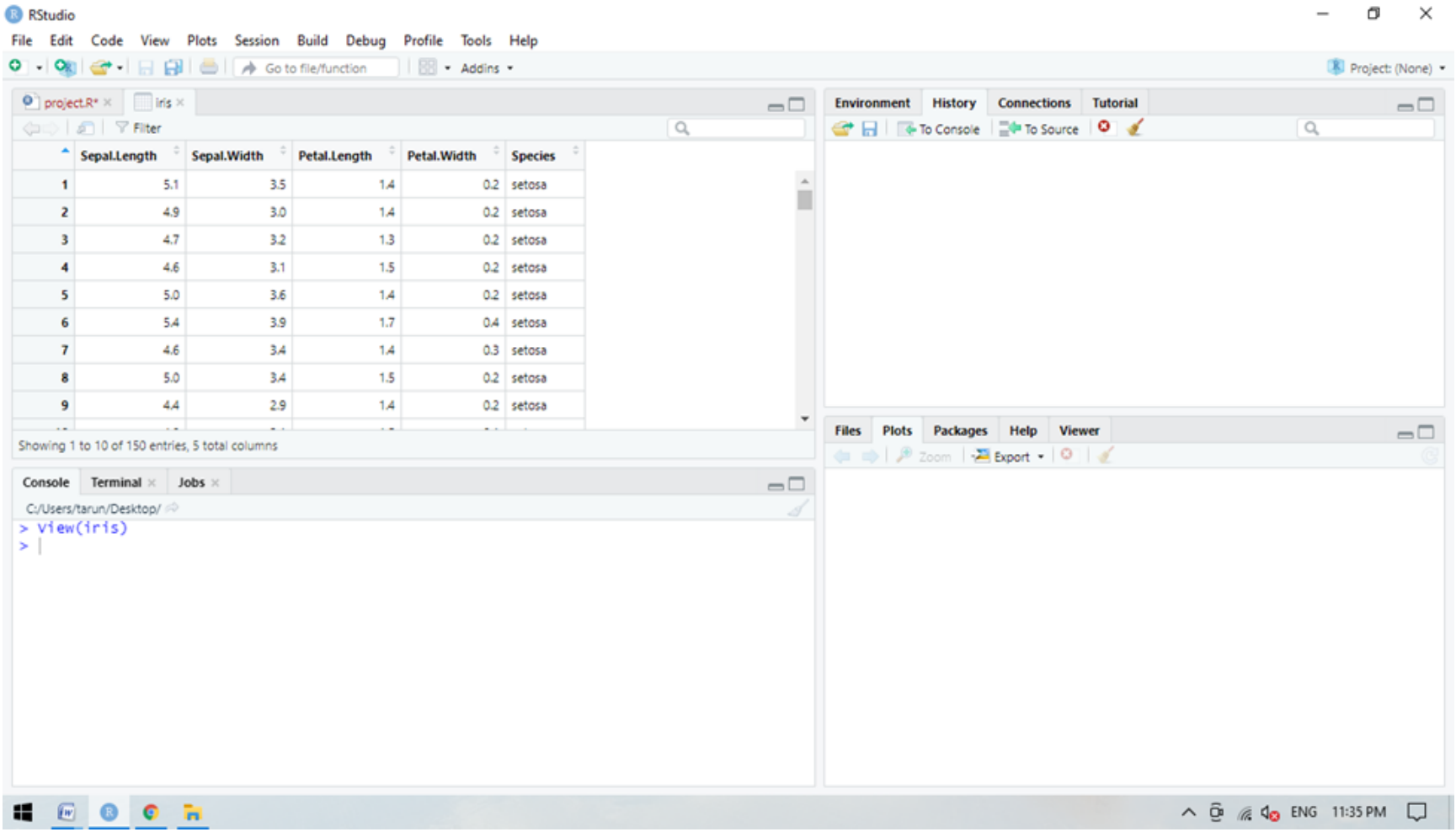
sapply() function takes list, vector or data frame as input and gives output in vector or matrix. It is useful for operations on list objects and returns a list object of same length of original set **: sapply (iris, class)**



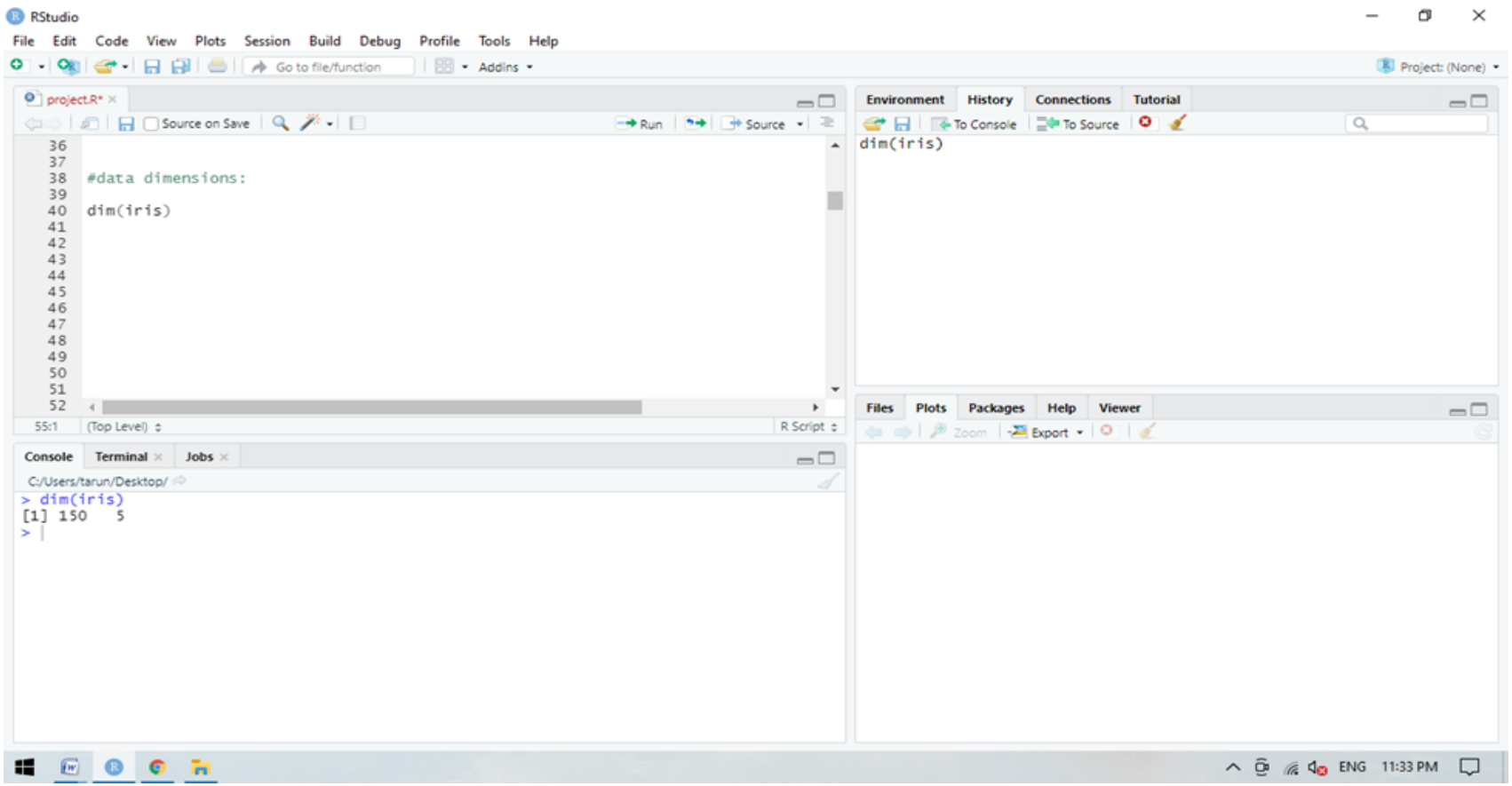
<u>str(iris)</u>: It is used to give the structure of the data



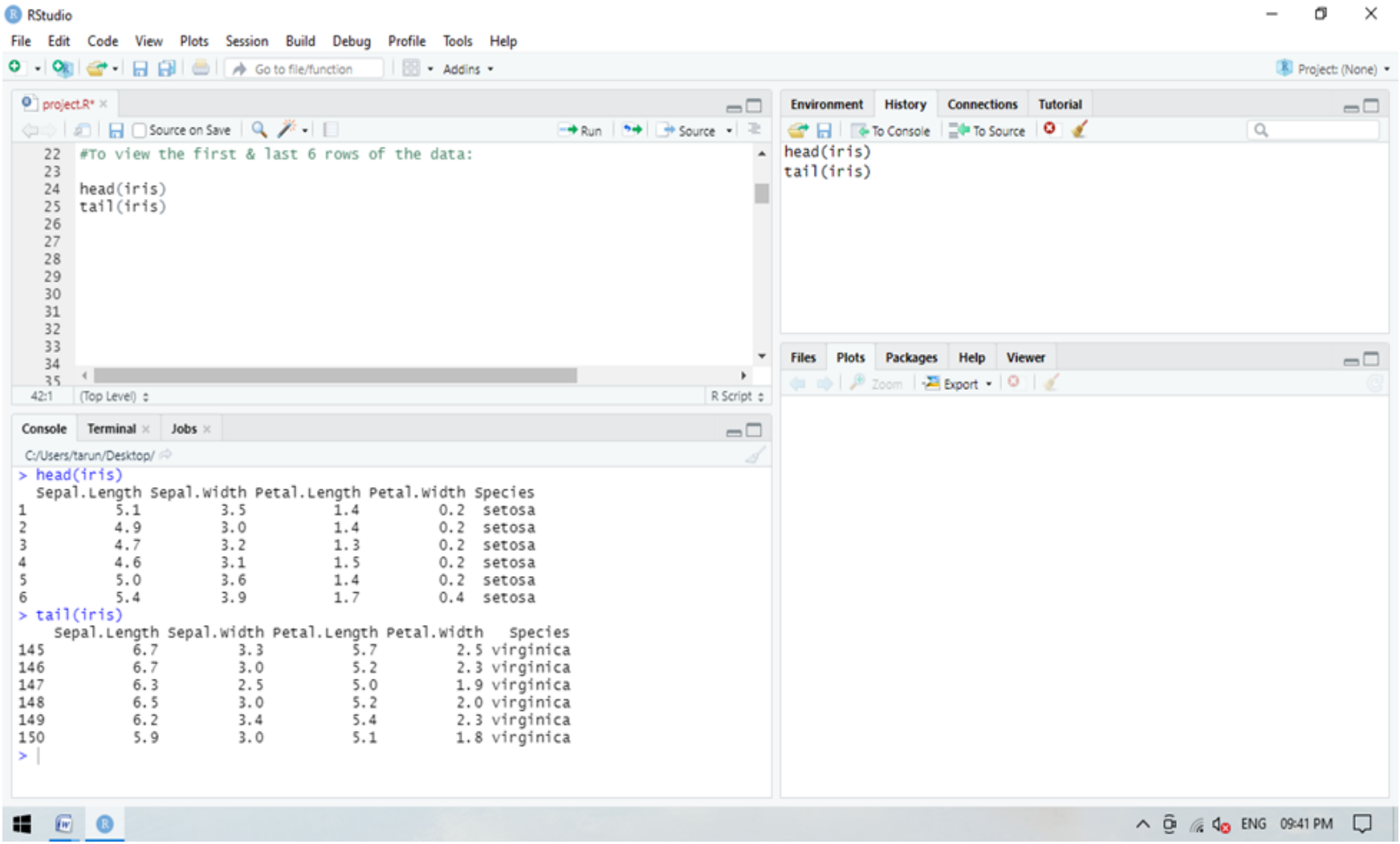
View(): To View the data **: View**(iris)



dim(): We can get a quick idea of how many instances (rows) and how many attributes (columns) the data contains with the **dim** function **:** dim(iris)



head(iris); **tail**(iris)



summary ():

The summary( ) function in R is used to obtain the summary statistics of the dataset, including minimum value, 1st quantile, median, mean 3rd quantile, maximum value for each numerical variable and the count for each level of the only categorical variable “species”.

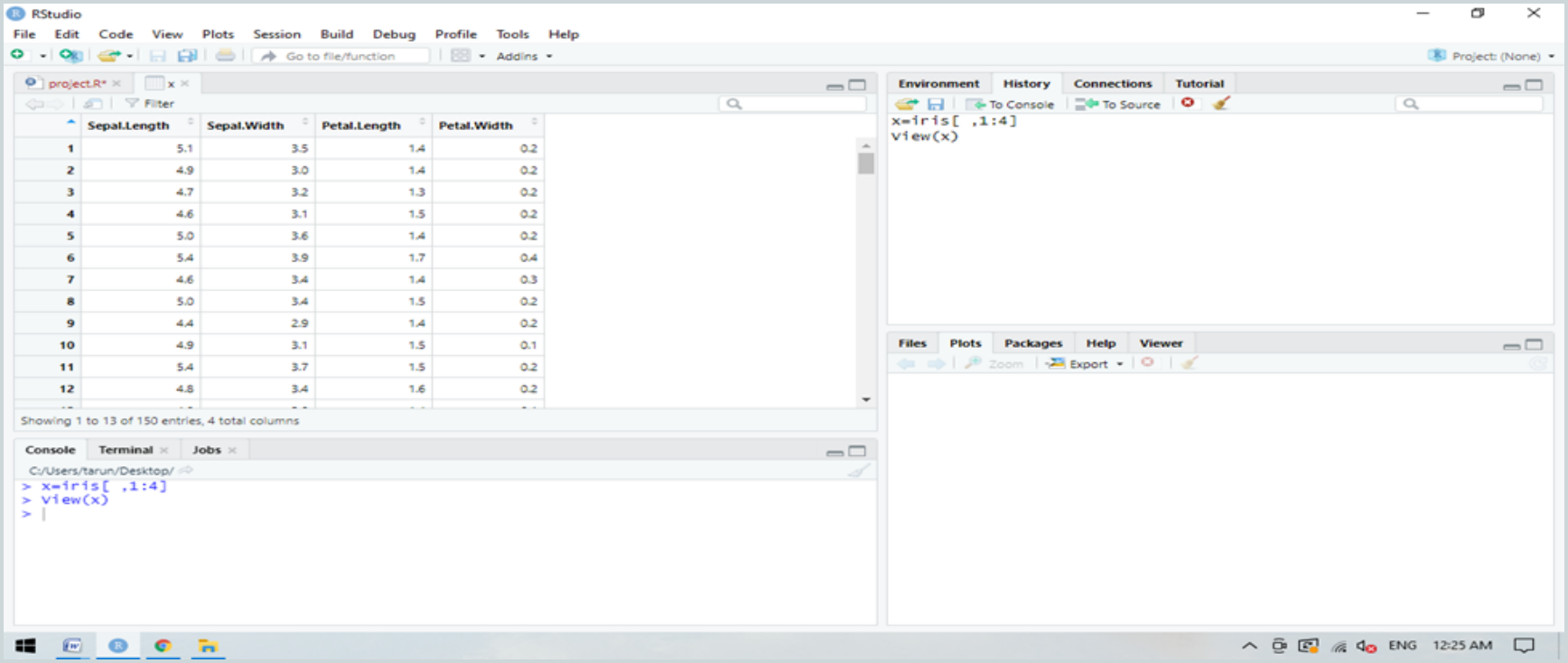
which is displayed in output : **summary**(iris)

```
Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
Median :5.800   Median :3.000   Median :4.350   Median :1.300
Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
Species
setosa    :50
versicolor:50
virginica :50
```

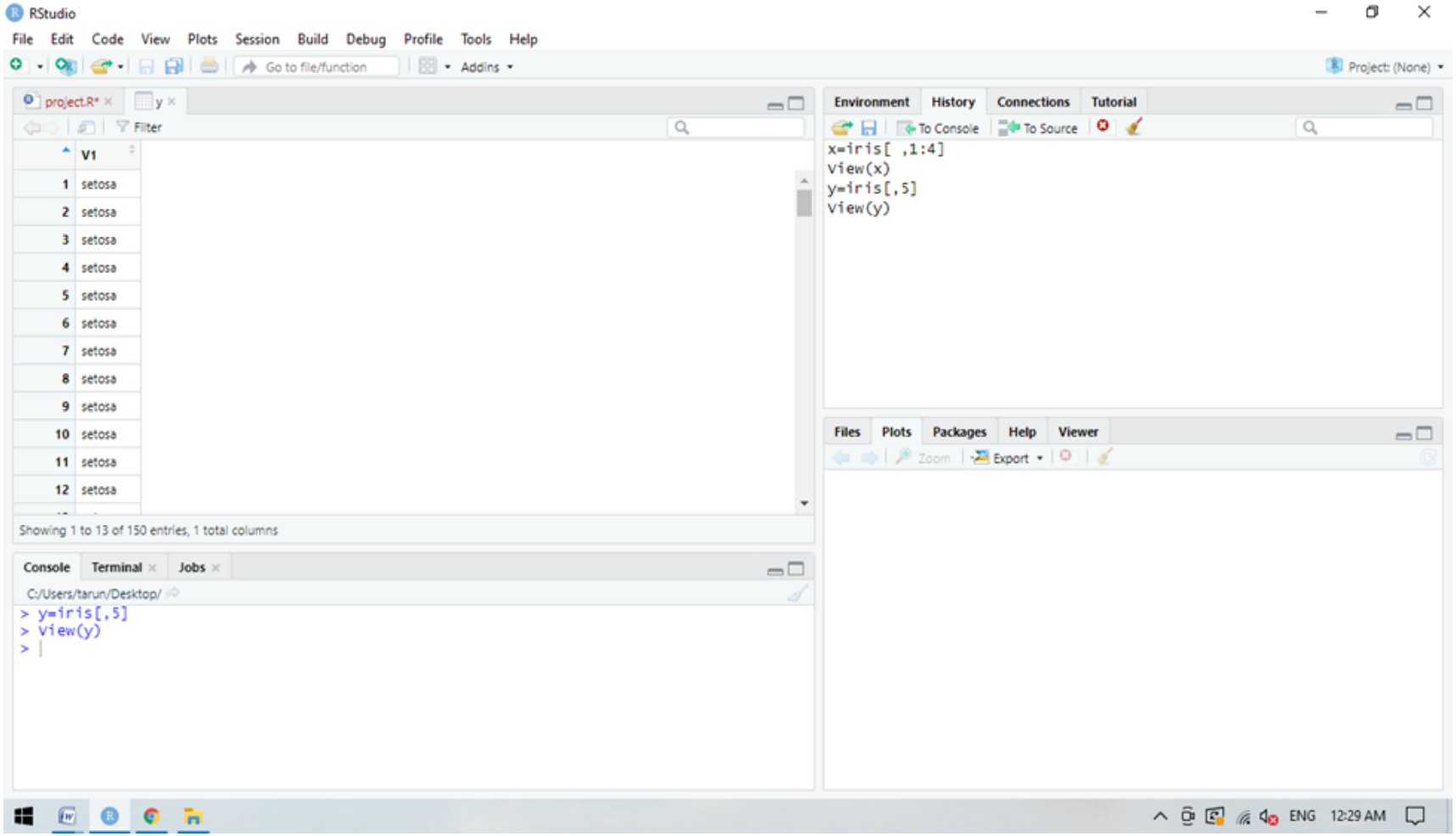
VISUALIZE DATASET, SPLIT THE DATA

x=iris[,1:4]

View(x)

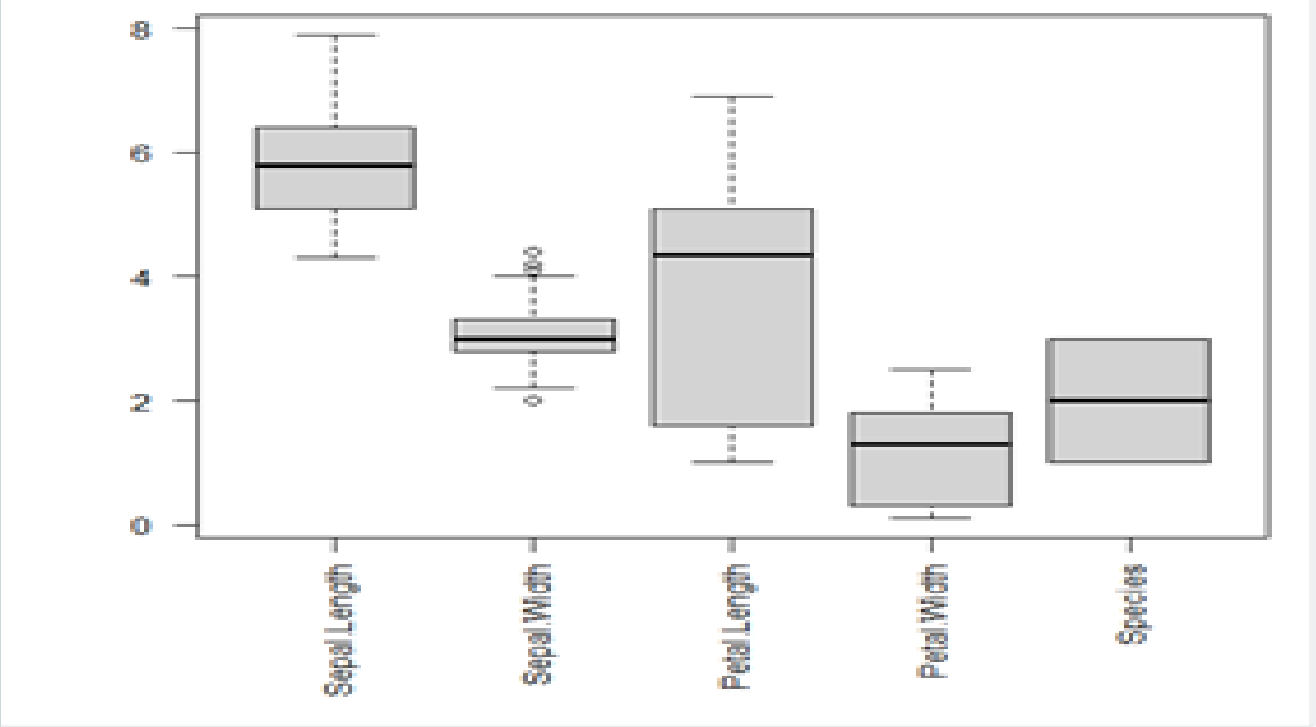


y=iris[,5]
View(y)



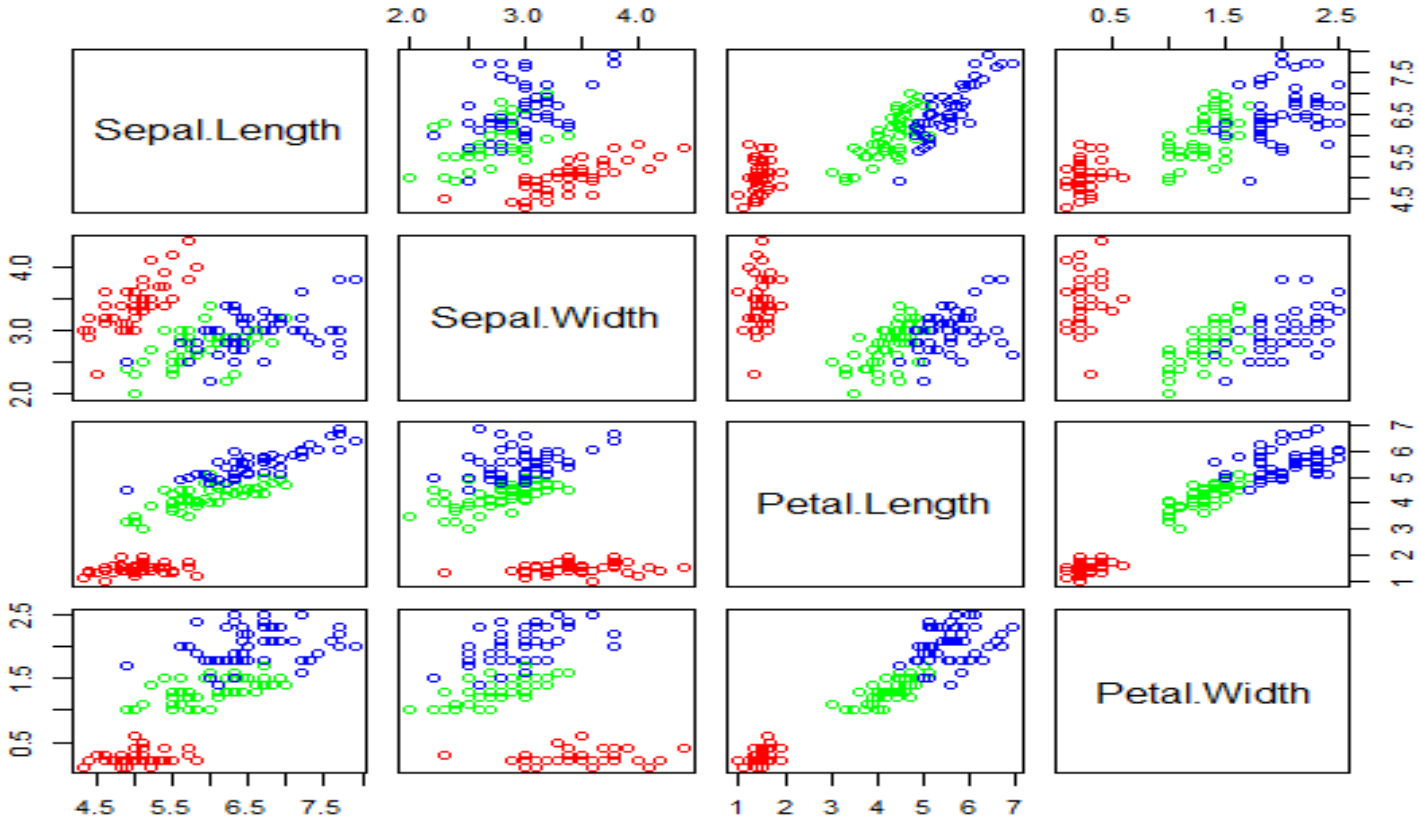
BOX PLOTS

- Box plot, also known as a box and whisker plot, displays a summary of a large amount of data in five numbers - minimum, lower quartile(25th percentile), median(50th percentile), upper quartile(75th percentile) and maximum data values.



SCATTER PLOT

Scatter plots are used to plot data points on a horizontal and a vertical axis in the attempt to show how much one variable is affected by another. The relationship between two variables is called their correlation.



From the below plot, there seems to be a positive correlation between the length and width of all the species, however there is a distinguishing strong correlation and relationship between petal length and petal width.

LINEAR DISCRIMINANT ANALYSIS

Linear Discriminant Analysis(LDA) Linear discriminant methods group images of the same classes and separates images of the different classes. To identify an input test image, the projected test image is compared to each projected training image, and the test image is identified as the closest training image.

- Let’s create a training dataset and test dataset for prediction and testing purposes. 60% dataset used for training purposes and 40$ used for testing purposes.

```
> linear=lda(Species~.,training);linear
Call:
lda(Species ~ ., data = training)

Prior probabilities of groups:
    setosa versicolor  virginica
 0.3837209  0.3139535  0.3023256

Group means:
           Sepal.Length Sepal.Width Petal.Length Petal.Width
setosa         4.975758    3.357576     1.472727   0.2454545
versicolor     5.974074    2.751852     4.281481   1.3407407
virginica      6.580769    2.946154     5.553846   1.9807692

Coefficients of linear discriminants:
                   LD1        LD2
Sepal.Length  1.252207 -0.1229923
Sepal.Width   1.115823  2.2711963
Petal.Length -2.616277 -0.7924520
Petal.Width  -2.156489  2.6956343

Proportion of trace:
   LD1    LD2
0.9937 0.0063
```

- Based on the training dataset, 38% belongs to setosa group, 31% belongs to versicolor groups and 30% belongs to virginica groups.
- The proportion of trace tells us how well each discriminant distinguishes between the species, and given the very high size of the first discriminant (0.9937), we see that LD1 explains 99% of the variance and so LD2 is contributing very little to the distinction of species.

CONCLUSION

- While Setosa can be easily identified (Linearly separable), Virginica and Versicolor have some overlap (almost linearly separable).
- The percentage separation achieved for the first discriminant function is 99.3%. so we have clear separation among the three species.
- The separation percentage achieved by LD2 is not good. That is quietly less than one percent.

THANK YOU