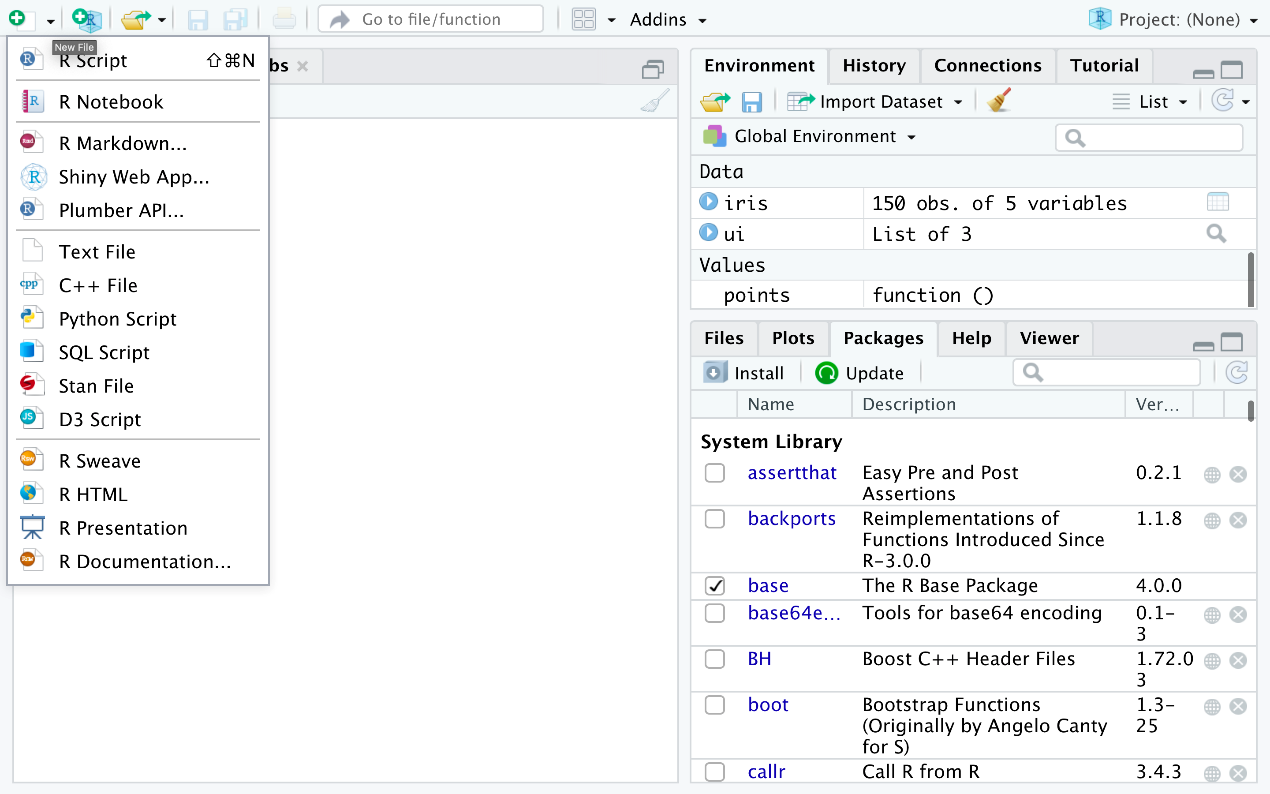
**Objective for Exercise:**

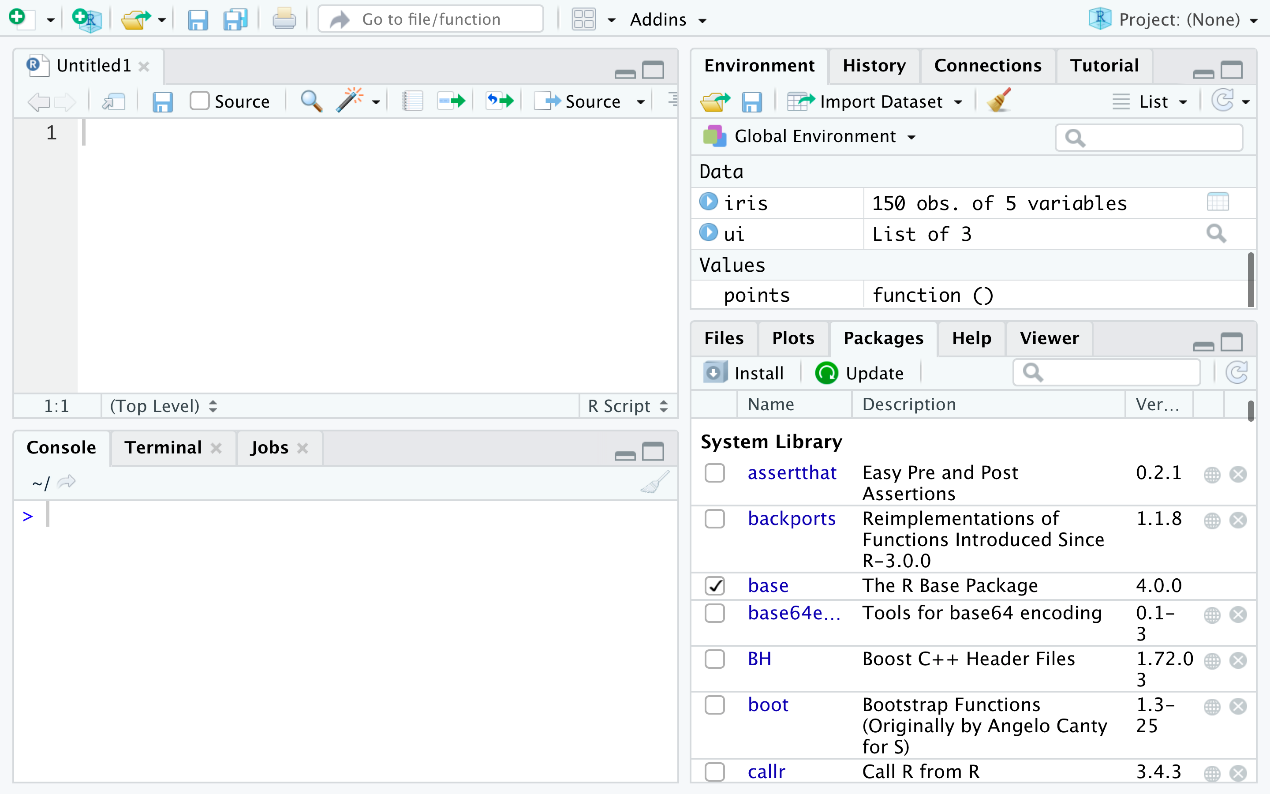
This lab introduces you to R and RStudio

**Exercise:**

1. Click in the tiny plus symbol top left and select R Script



1. An untitled R Script panel opens. It would look like this.

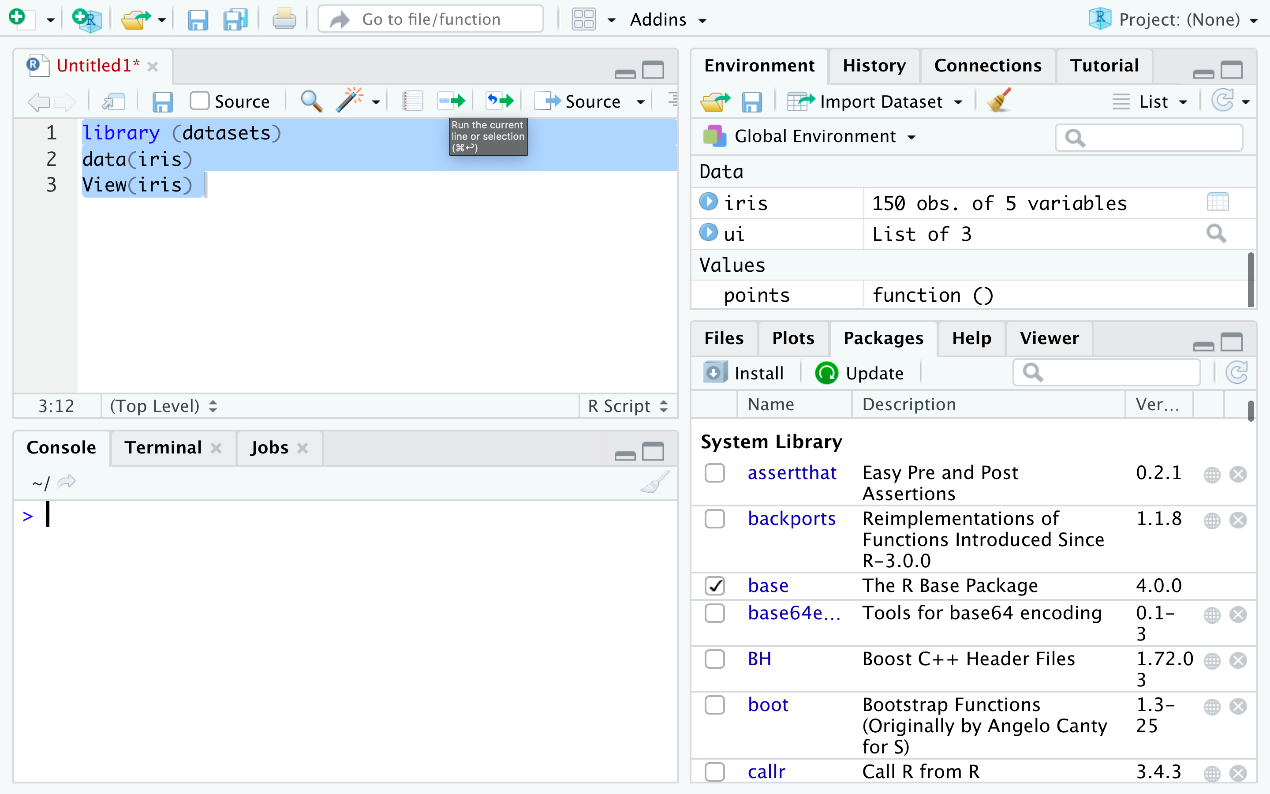


1. Now we load the iris dataset. Please enter the following lines into the editor window which just appeared. Then select them all such that they turn blue. Then click on the tiny Run icon just above the editor window.

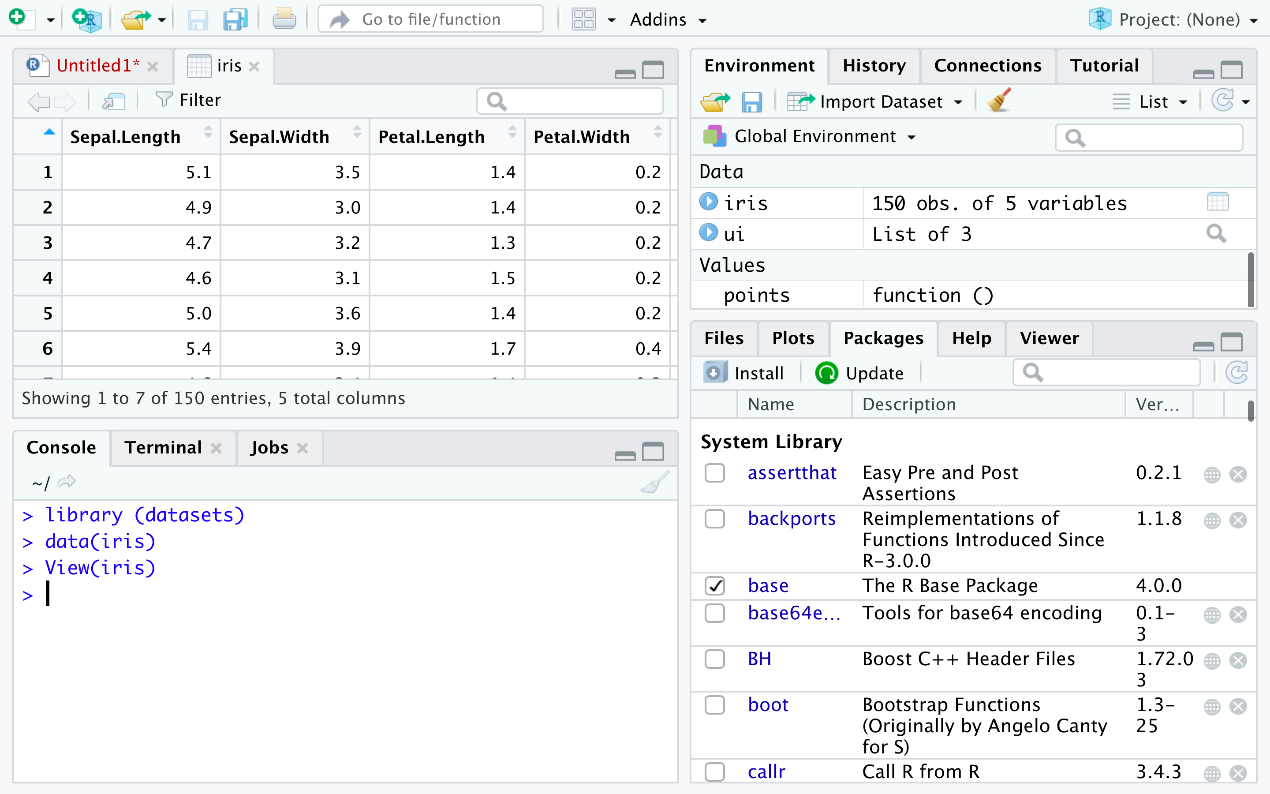
library (datasets)

**data**(iris)

**View**(iris)

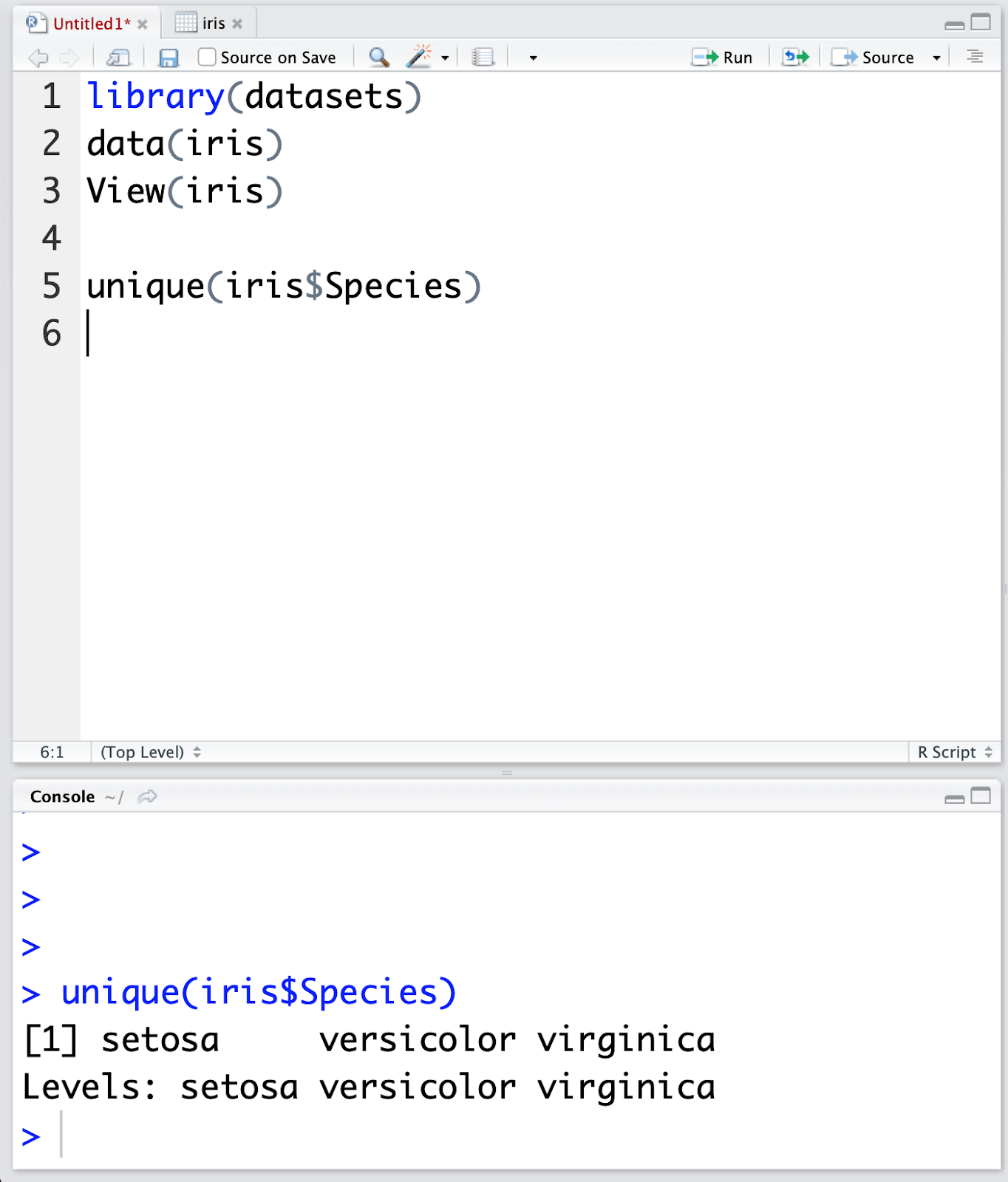


1. You are directly taken to the data view tab to inspect your dataset. We can see that there are five columns in this data set and that the first four are floating point and the last one is label of data type string which contains the category value of our data set. We also see that we have 150 entries in total of which we are seeing the first 19.



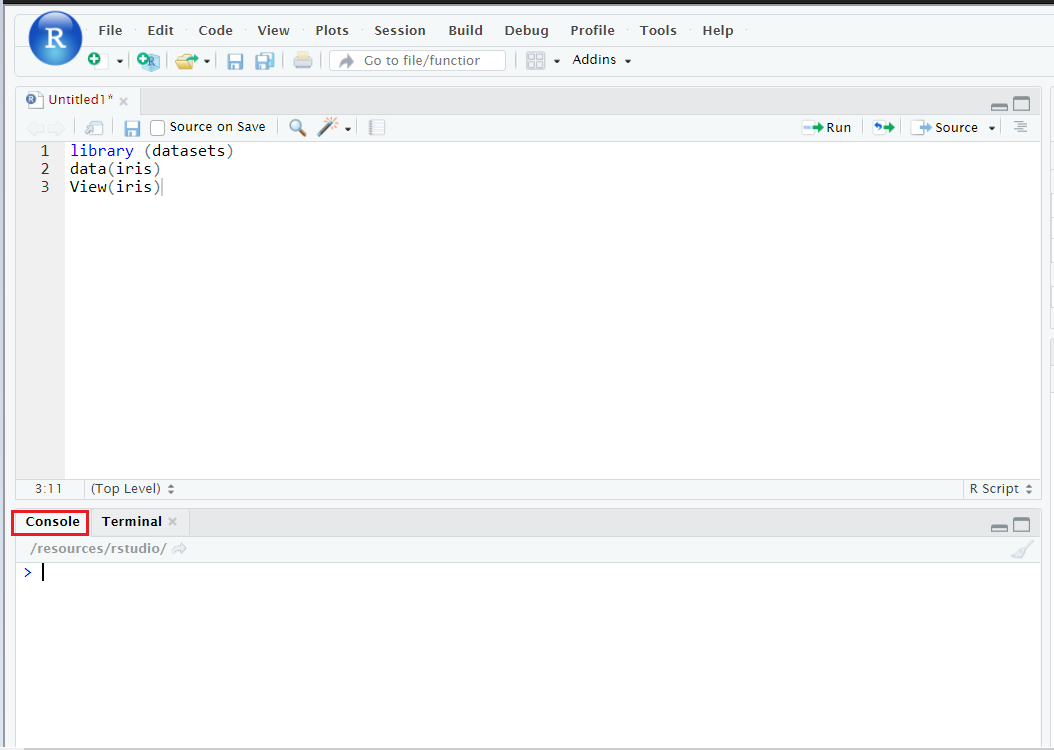
1. Now let's find how many different species there are present in the data set. Type the following command into the editor window and click the run.

**unique**(iris$Species)



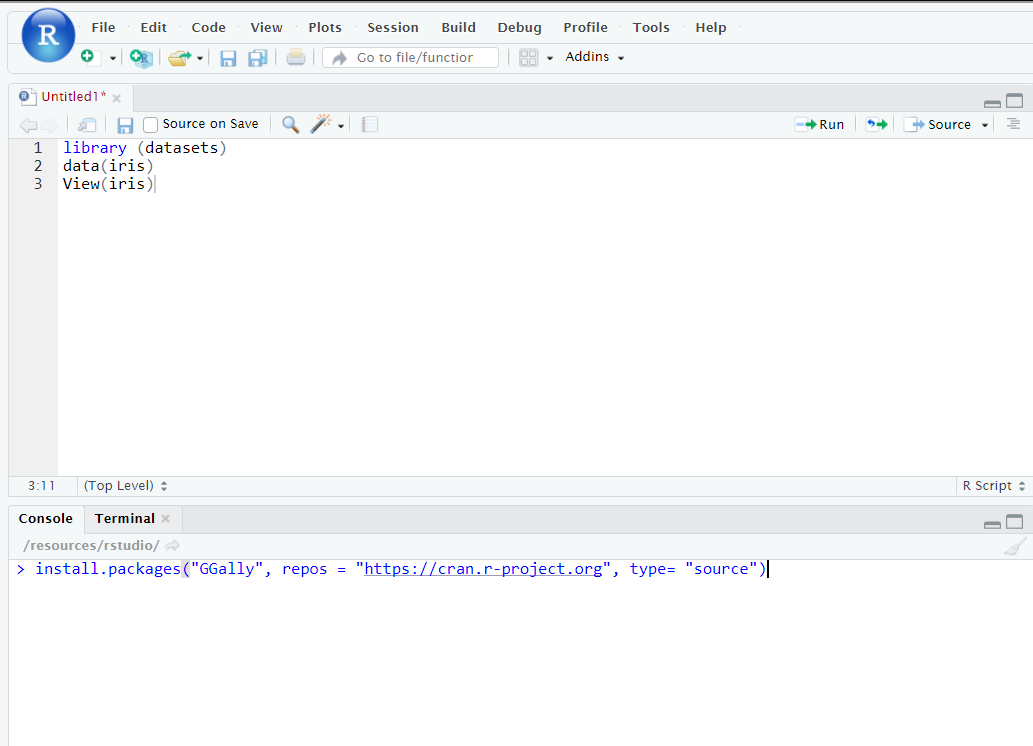
In the Console window at the bottom you’ll see the result of the executed command and will know that there are only three different species present in the data set.

1. Now it’s time to look into the data set in more detail.
2. Open a Console.



1. Run the following command in the console.

install.packages("GGally", repos = "https://cran.r-project.org", type= "source")



1. Click *Enter* to install the packages.

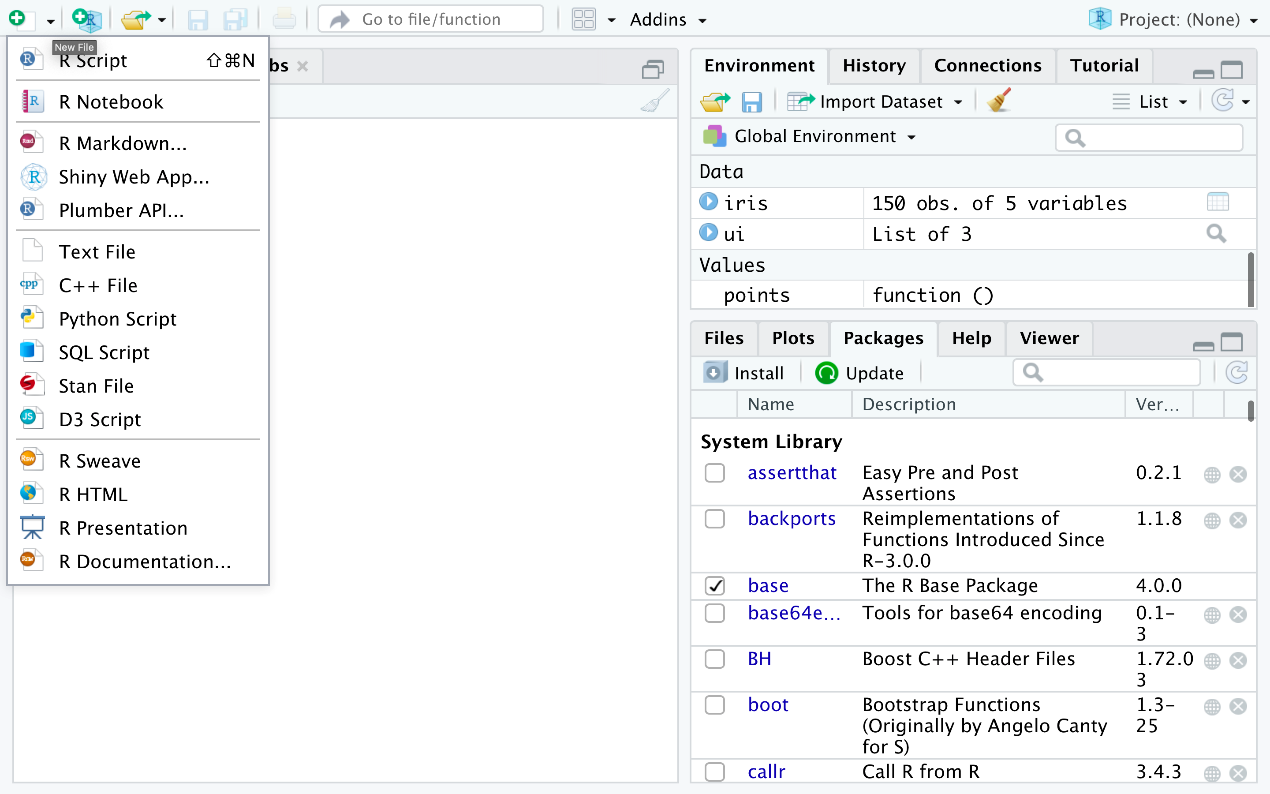
This concludes this lab, I hope you’ve enjoyed it!

**Creating Data Visualizations using ggplot**

**Objective for Exercise**

We will create different data visualizations using the ggplot package using the inbuilt dataset in R called mtcars

1. Click on the + symbol on the top left and choose R Script from the menu to open a new R edit window in RStudio:



1. Read and view the first 5 rows of the Data using the following:

**library**(datasets)

# Load Data

**data**(mtcars)

# View first 5 rows

**head**(mtcars, 5)

1. Type this ?mtcars to get information about the variables. This will print the information at the bottom right panel, on the Help tab
2. Copy and paste the following code to load the ggplot package and create a scatterplot of disp and mpg.

#load ggplot package

**library**(ggplot2)

# create **a** scatterplot of displacement (disp) and miles per gallon (mpg)

**ggplot**(aes(x=disp,y=mpg,),data=mtcars)+geom\_point()

1. Use the following code to add a title.

# Add a title

**ggplot**(**aes**(x=disp,y=mpg,),data=mtcars)+**geom\_point**()+**ggtitle**("displacement vs miles per gallon")

1. Use the following code to change the name of the x-axis and y-axis

# change axis name

**ggplot**(aes(x=disp,y=mpg,),**data**=mtcars)+geom\_point()+ggtitle("**displacement** **vs** **miles** **per** **gallon**") + labs(**x** = "Displacement", **y** = "Miles **per** Gallon")

1. Use the following to create a boxplot of the the distribution of mpg for the individual Engine types vs Engine (0 = V-shaped, 1 = straight)

To do this you have to make vs a string or factor.

#make vs a factor

mtcars$vs <- as.factor(mtcars$vs)

# create boxplot of the distribution for v-shaped and straight Engine

ggplot(aes(**x**=vs, **y**=mpg), data = mtcars) + geom\_boxplot()

1. Add color to the boxplots to help differentiate:

ggplot(aes(x=vs, y=mpg, fill = vs), data = mtcars) +

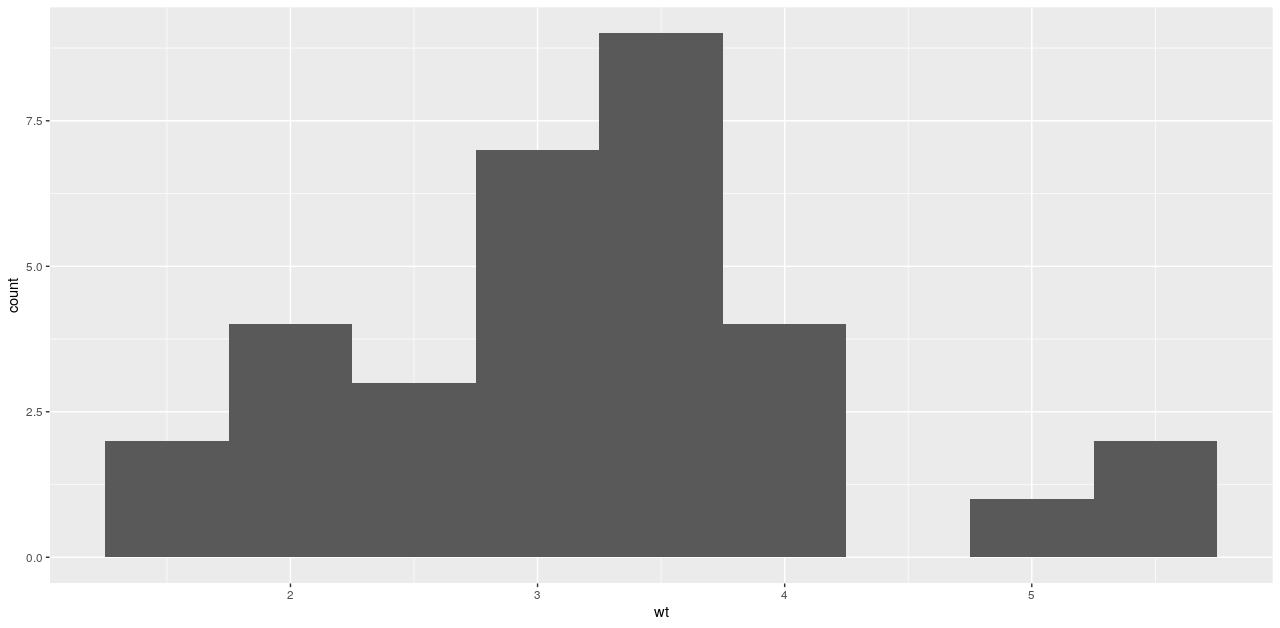
geom\_boxplot(alpha=0.3) +

theme(legend.position="none")

1. Finally, let us create the histogram of weight wt.

**ggplot**(**aes**(x=wt),data=mtcars) + **geom\_histogram**(binwidth=0.5)

This concludes this lab, we hope that you had fun!

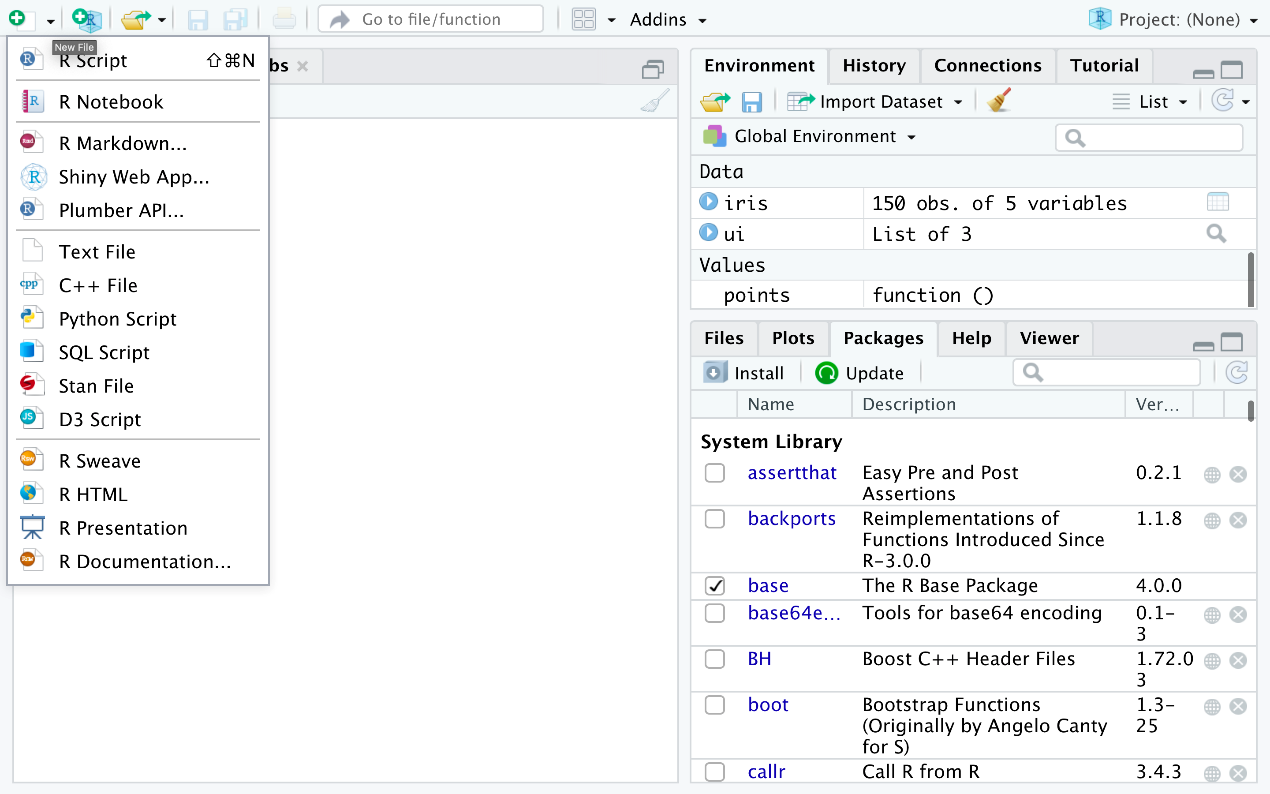


**Objective for Exercise:**

This lab introduces you to plotting in R with ggplot and GGally. GGally is an extension of ggplot2

**Exercise:**

1. Click in the tiny plus symbol top left and select R Script to create a new R script if you dont have one open already.



1. We will use the iris dataset. If you don't have it loaded, please copy and paste the following into your R script file.

**library**(datasets)

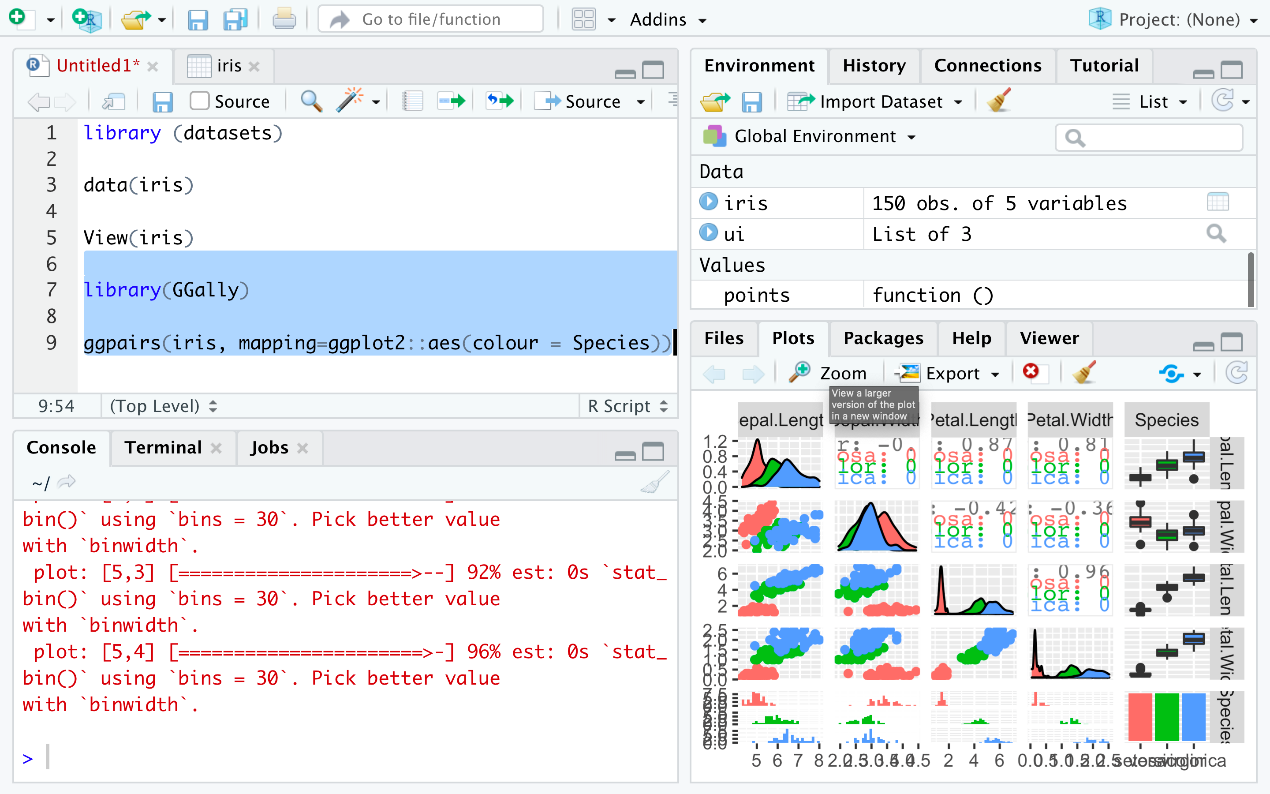
**data**(iris)

1. In the previous lab, you installed the libraries necessary to create some nice plots let’s execute the following commands:

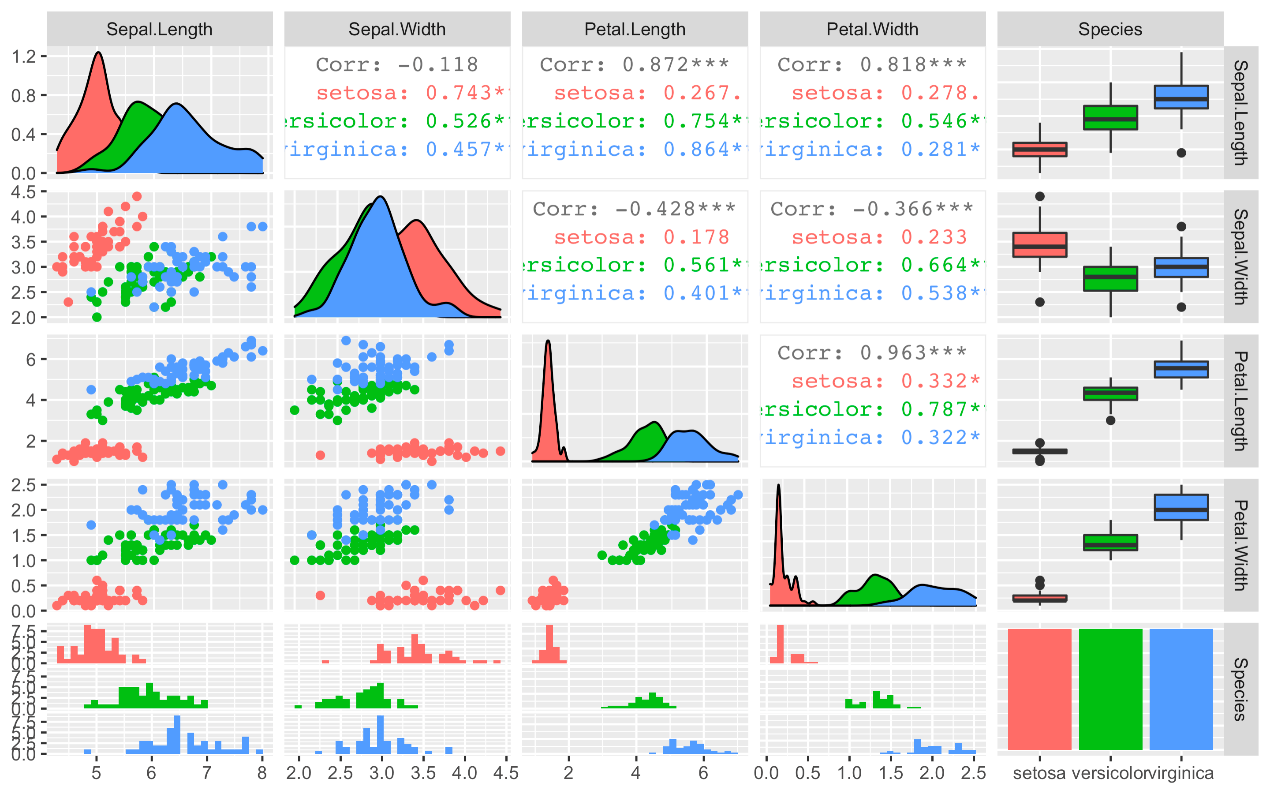
**library**(GGally)

**ggpairs**(iris, mapping=ggplot2::aes(colour = Species))

1. Select the commands and click on run on the top. You’ll now see the following plot in the **plots** window:



1. Click on **Zoom** icon on the plot window to zoom and see the plot.



1. This gives us a lot of information for a single line of code. First, we see the data distributions per column and species on the diagonal. Then we see all pair-wise scatter plots on the tiles left to the diagonal, again broken down by color. It is, for example, obvious to see that a line can be drawn to separate **setosa** against **versicolor** and **virginica**. In later courses, we’ll of course teach how the overlapping species can be separated as well. This is called supervised machine learning using non-linear classifiers by the way. Then you see the correlation between individual columns in the tiles right to the diagonal which confirms our thoughts that **setose** is more different, hence more easy to distinguish, than **versicolor** and **virginica** since a correlation value close to one signifies high similarity whereas a value closer to zero signifies less similarity. The remaining plots on the right are called **box-plots** and the ones at the bottom are called **histograms** bit we won’t go into detail here and save this for a more advanced course in this series.