Lab: RStudio – The Basics

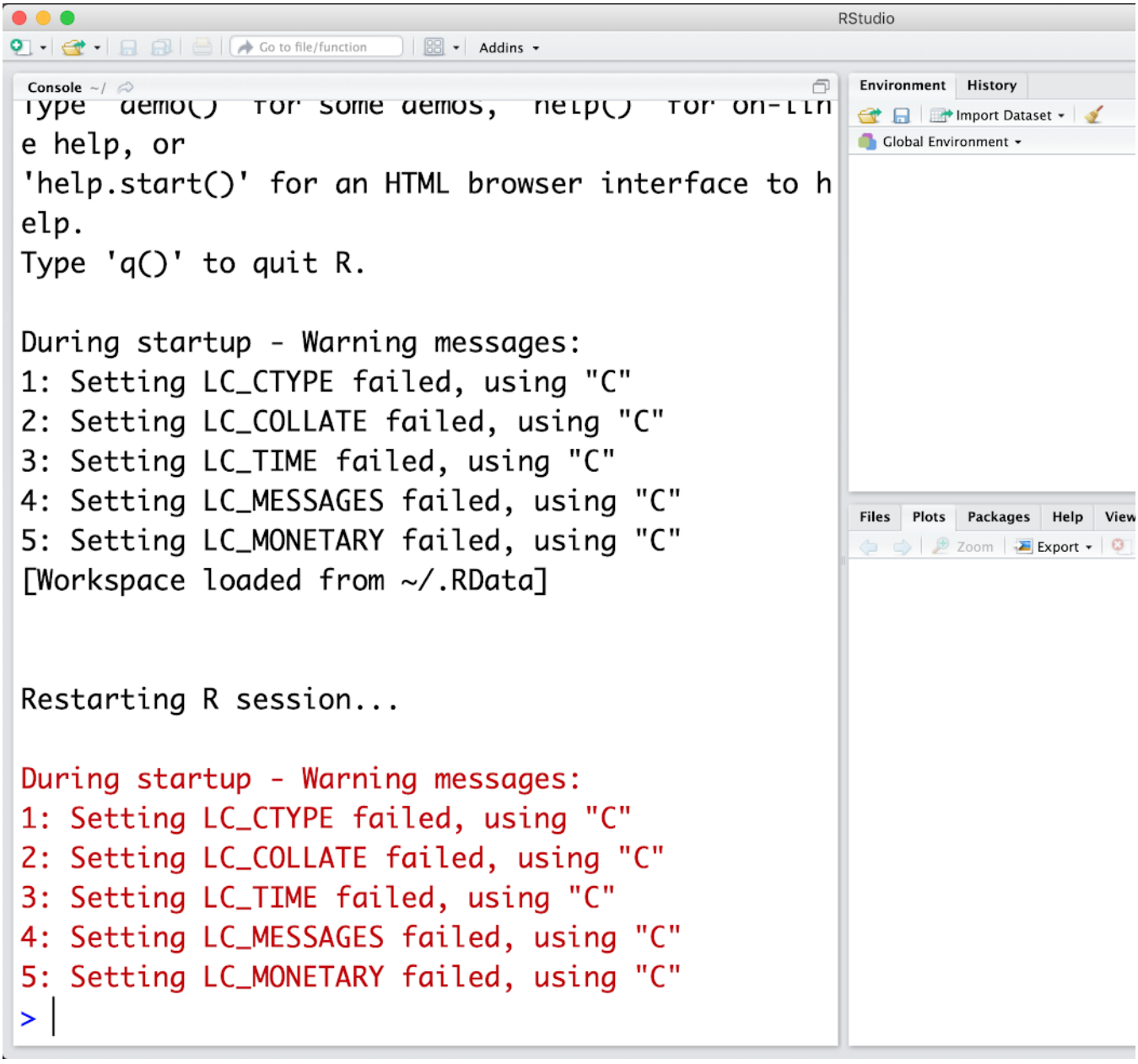
coursera.org/learn/open-source-tools-for-data-science/supplement/eeoJY/lab-rstudio-the-basics

This lab introduces you to R and RStudio. First, we need to install it.

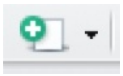
Please install R appropriate for your operating system. You can obtain it here: https://cran.rstudio.com/

Once you've installed R, please download and install RStudio appropriate for your operating system. You can find it here: https://rstudio.com/products/rstudio/download/#download

1. Start RStudio. You should see something like this:

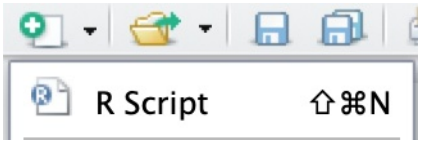


2. Now click in the tiny "plus" symbol top left.



and select "Rscript" . This gives you the following:

3. Now we load the iris data set which you already should be familiar with from the previous labs. Please enter the following lines into the editor window which just appeared.



library (datasets)

data(iris)

View(iris)

Then select them all such that they turn blue. Then click on the tiny run icon just above the editor window.

4. You are directly taken to the data view tab to inspect your data set:

Untitled1* | iris

Filter

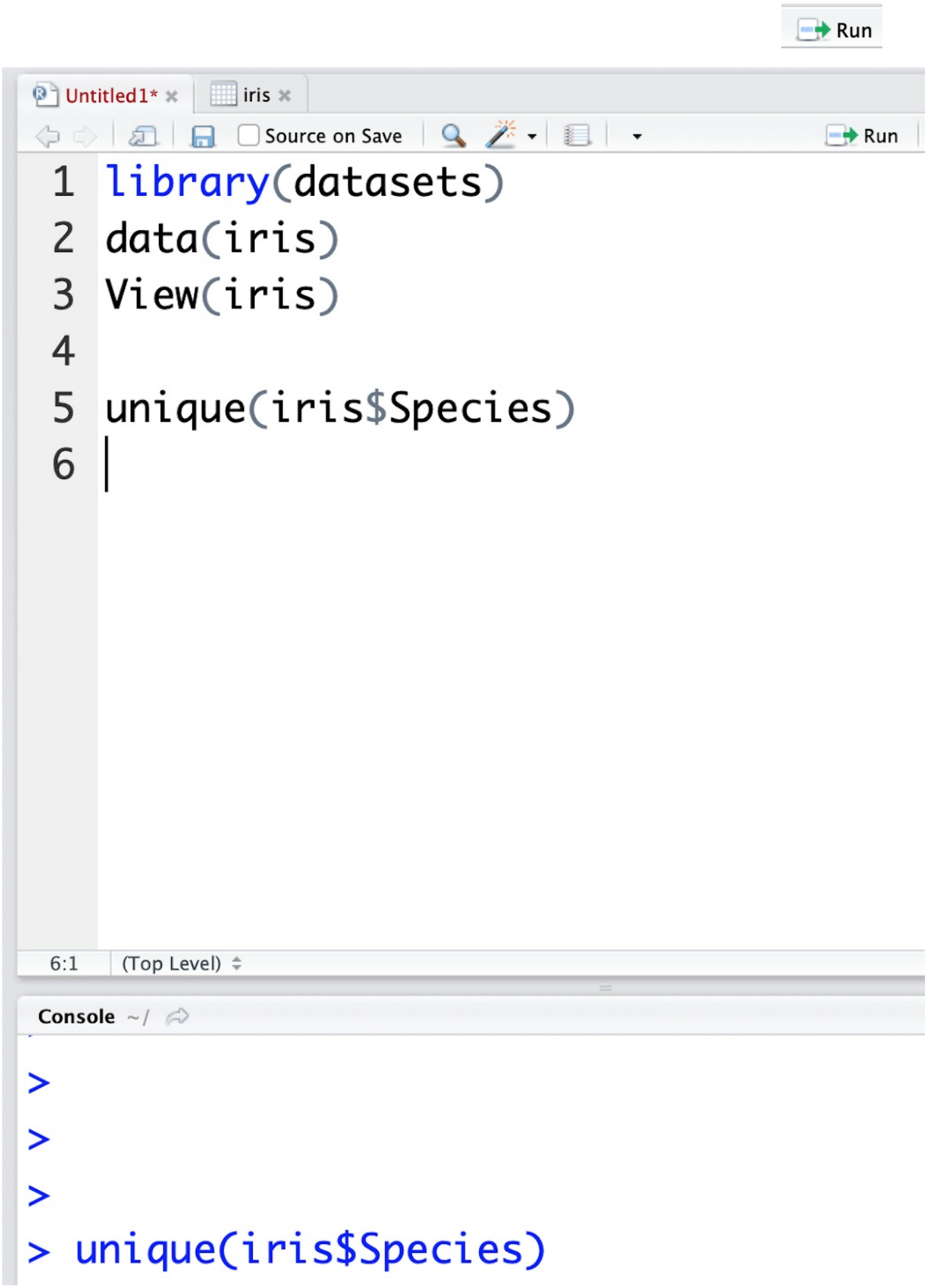
	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3.0	1.4	0.1	setosa
14	4.3	3.0	1.1	0.1	setosa
15	5.8	4.0	1.2	0.2	setosa
16	5.7	4.4	1.5	0.4	setosa
17	5.4	3.9	1.3	0.4	setosa
18	5.1	3.5	1.4	0.3	setosa
19	5.7	3.8	1.7	0.3	setosa

Showing 1 to 19 of 150 entries

5. 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 a 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. Now we want to know how many different species there are present in the data set. Therefore, please type the following command into the editor window:

unique(iris$Species)

and click the run icon:


Run
Untitled1*
iris
Source on Save
Run
1 library(datasets)
2 data(iris)
3 View(iris)
4
5 unique(iris$Species)
6
6:1
(Top Level)
Console ~/
>
>
>
> unique(iris$Species)

```
[1] setosa     versicolor virginica
Levels: setosa versicolor virginica
>
```

6. 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. Now it's time to look into the data set in more detail. We'll create a similar plot as we've already done in the JupyterLab. To do so we first need to install the ggplot2 package. Please type and execute the following command:

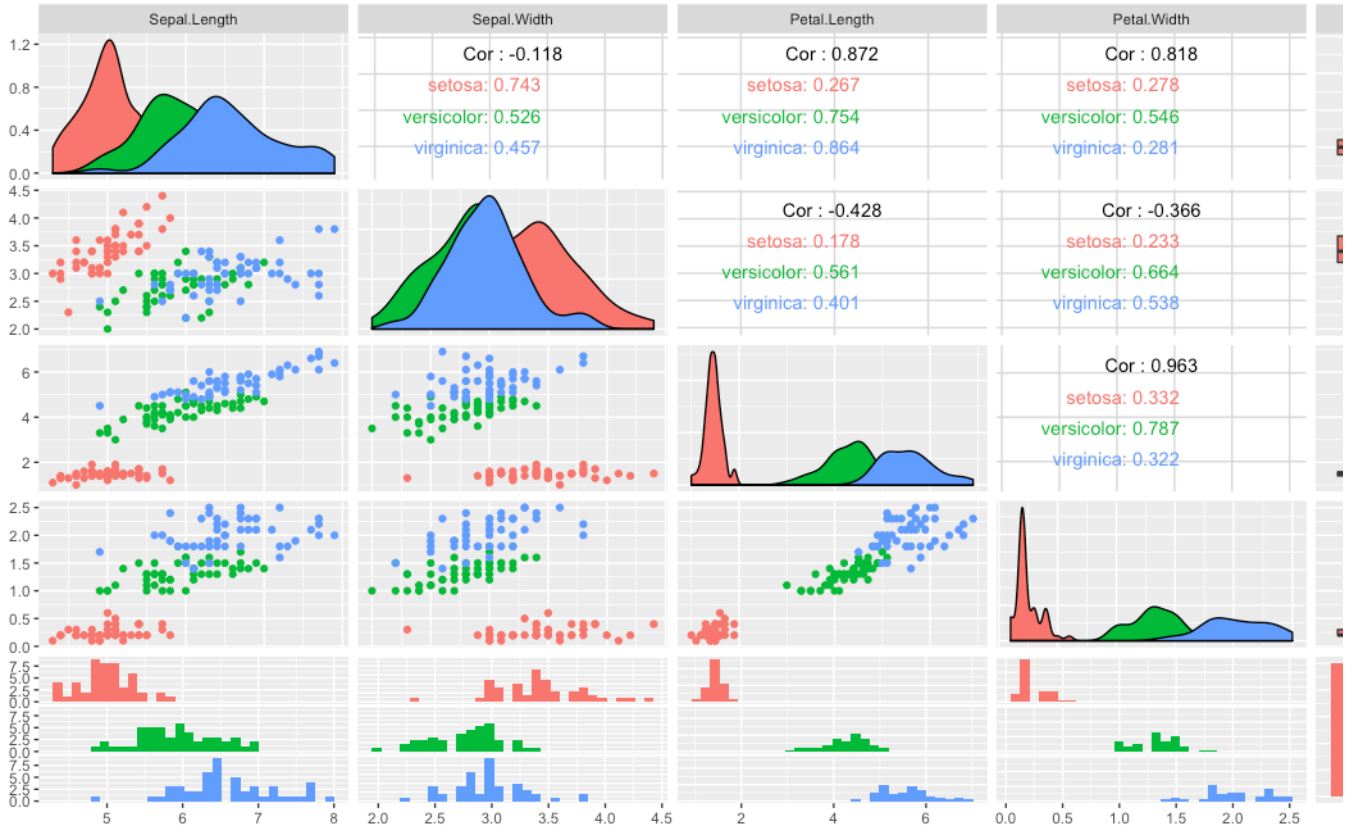
install.packages(c("GGally", "ggplot2"))

7. Now that you have installed the libraries necessary to create some nice plots let's execute the following commands:

library(GGally)

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

8. You'll now see the following plot in the **Plots** window:



9. 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 will 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", but we won't go into detail here, rather we will save this for a more advanced course in this series.

This concludes the lab, I hope you've enjoyed it!