

```
::page{title="Plotting with RStudio"}
```



Objective of Exercise:

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

Pre Requisite:

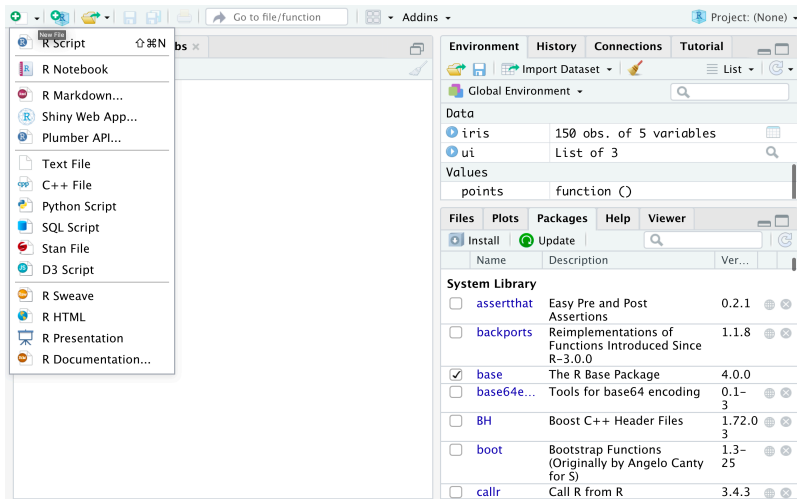
Before loading the `GGally` package, ensure its dependencies are installed. Run the following commands in the Console window, as shown in the screenshot below, and then continue with the steps in the Exercise.

▶ See Screenshot

```
install.packages("https://cran.r-project.org/src/contrib/Archive/patchwork/patchwork_1.1.0.tar.gz", repos = NULL, type = "source", dependencies
install.packages("https://cran.r-project.org/src/contrib/Archive/broom.helpers/broom.helpers_1.4.0.tar.gz", repos = NULL, type = "source", depen
install.packages("https://cran.r-project.org/src/contrib/Archive/ggstats/ggstats_0.5.0.tar.gz", repos = NULL, type = "source", dependencies = TR
```

Exercise:

1. Click the plus symbol on the top left and click **R Script** to create a new R script, if you don't have one open already.



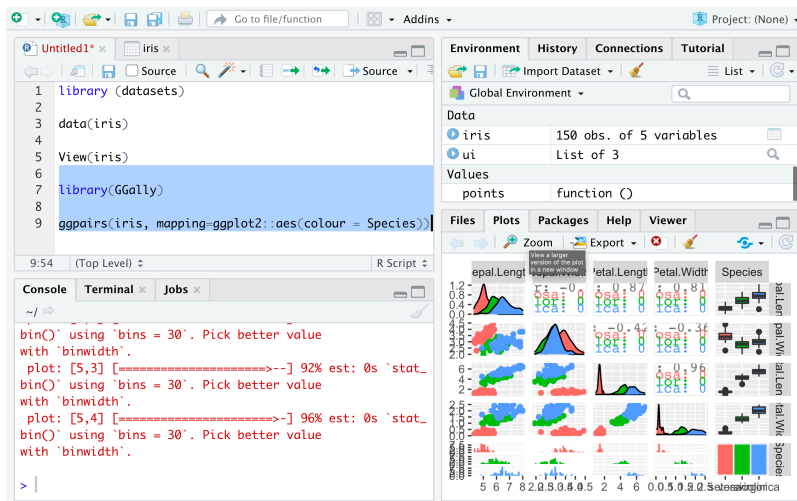
2. You will use the `iris` dataset. If you don't have it loaded, copy and paste the following into your R script file.

```
library(datasets)
data(iris)
```

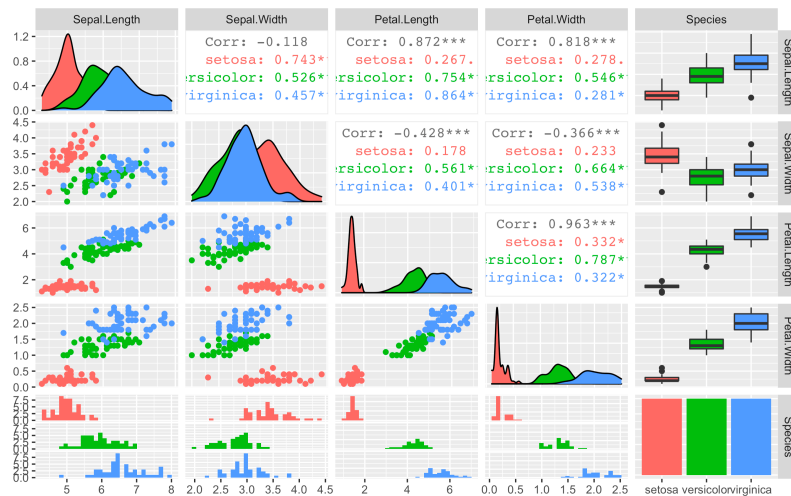
3. In the previous lab, you installed the libraries necessary to create plots, let's execute the following commands:

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

4. Select the commands and click **Run** on the top. You'll see the following plot in the **Plots** window:



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



6. This gives you a lot of information for a single line of code. First, you can see the data distributions per column and species on the diagonal. Then you see all the pair-wise scatter plots on the tiles left to the diagonal, again segregated by color. It is, for example, obvious that a line can be drawn to separate **setosa** against **versicolor** and **virginica**. In later courses, you will also learn how the overlapping species can be separated. This is called supervised machine learning using non-linear classifiers. You can also see the correlation between individual columns in the tiles on the right to the diagonal, which confirms that **setosa** is more different, hence easier to distinguish, than **versicolor** and **virginica**. 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 you will learn about this in a more advanced course in this series.

Author(s)

Romeo

Other Contributor(s)

Lavanya

© IBM Corporation. All rights reserved.