Overview

Augustin Luna
09 January, 2016

Research Fellow
Department of Biostatistics and Computational Biology
Dana-Farber Cancer Institute

Topics to be Covered

- R: Language Basics, Plotting, Getting Help
- Using the RStudio Editor
- Machine Learning Fundamentals
  - Dimension Reduction
  - Regression
  - Clustering
- Accessing Datasets: CellMiner (cell lines/drugs), CBioPortal (patient samples), Pathway Commons (pathways)
- Developing Web Applications

What is R?

- Free, open source
- Started in 1993
- Geared towards scientific computing
  - Created by Ross Ihaka and Robert Gentleman (statisticians)
- Interpreted; similar to Python and MATLAB

Why is R Popular?

- Free, open source
- Interactive data analysis
  - Script-driven rather than menu-driven helps reproducibility
- Flexible and powerful plotting support
- Excellent package management system
  - Large and growing collection of statistical analysis methods
  - Simple package installation; dependency management
  - R scripts usually portable to other platforms
  - Package repositories ensure functionality, documentation, and interoperability
  - Vignettes (tutorials) provided as runnable analyses

Extending R and Package Repositories

- Comprehensive R Archive Network (CRAN)
  - 5,800 R packages (as of June 2014)
  - Many packages call C, C++, Fortran, or Java code for speedups
- Bioconductor
  - 800+ R packages focused on bioinformatics
  - 50+ packages dedicated to pathway analysis
- Devtools
  - R package that allows package installation from code repositories

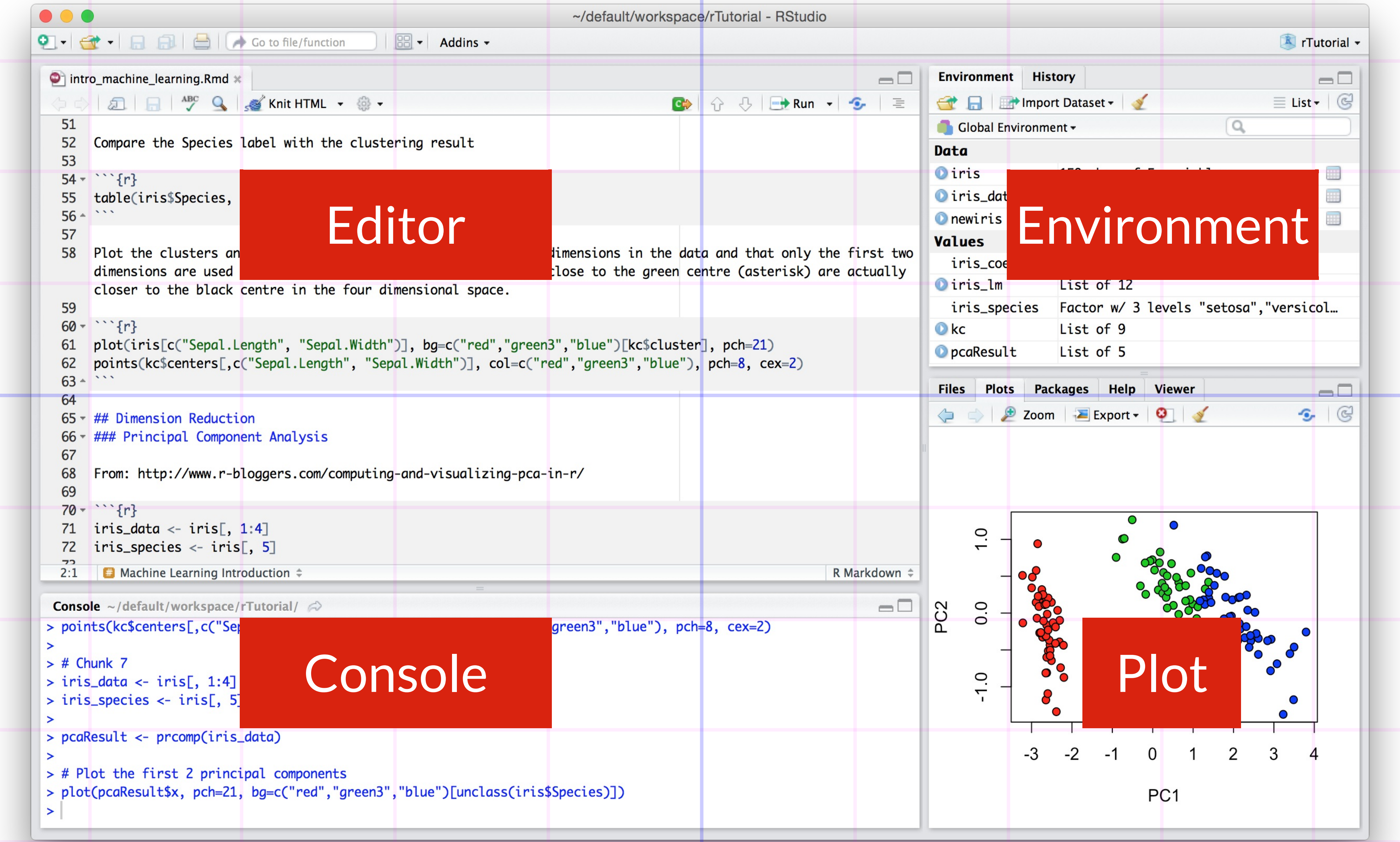
RStudio

- https://www.rstudio.com/
- Available for Windows, OSX, and Linux
- Simplifies common tasks: plotting, package installation, accessing files, viewing variables, etc.

Installing R and RStudio

- Install R
  - https://cran.rstudio.com/
- Install RStudio
  - https://www.rstudio.com/products/rstudio/download/
- RStudio does not come with R and R must be installed first

RStudio Overview



~~~
Compare the Species label with the clustering result

```{r}
table(iris$Species,
~~~

Plot the clusters an   dimensions in the data and that only the first two
dimensions are used   close to the green centre (asterisk) are actually
closer to the black centre in the four dimensional space.

```{r}
plot(iris[c("Sepal.Length", "Sepal.Width")], bg=c("red","green3","blue")[kc$cluster], pch=21)
points(kc$centers[,c("Sepal.Length", "Sepal.Width")], col=c("red","green3","blue"), pch=8, cex=2)
```

Dimension Reduction
Principal Component Analysis

From: http://www.r-bloggers.com/computing-and-visualizing-pca-in-r/

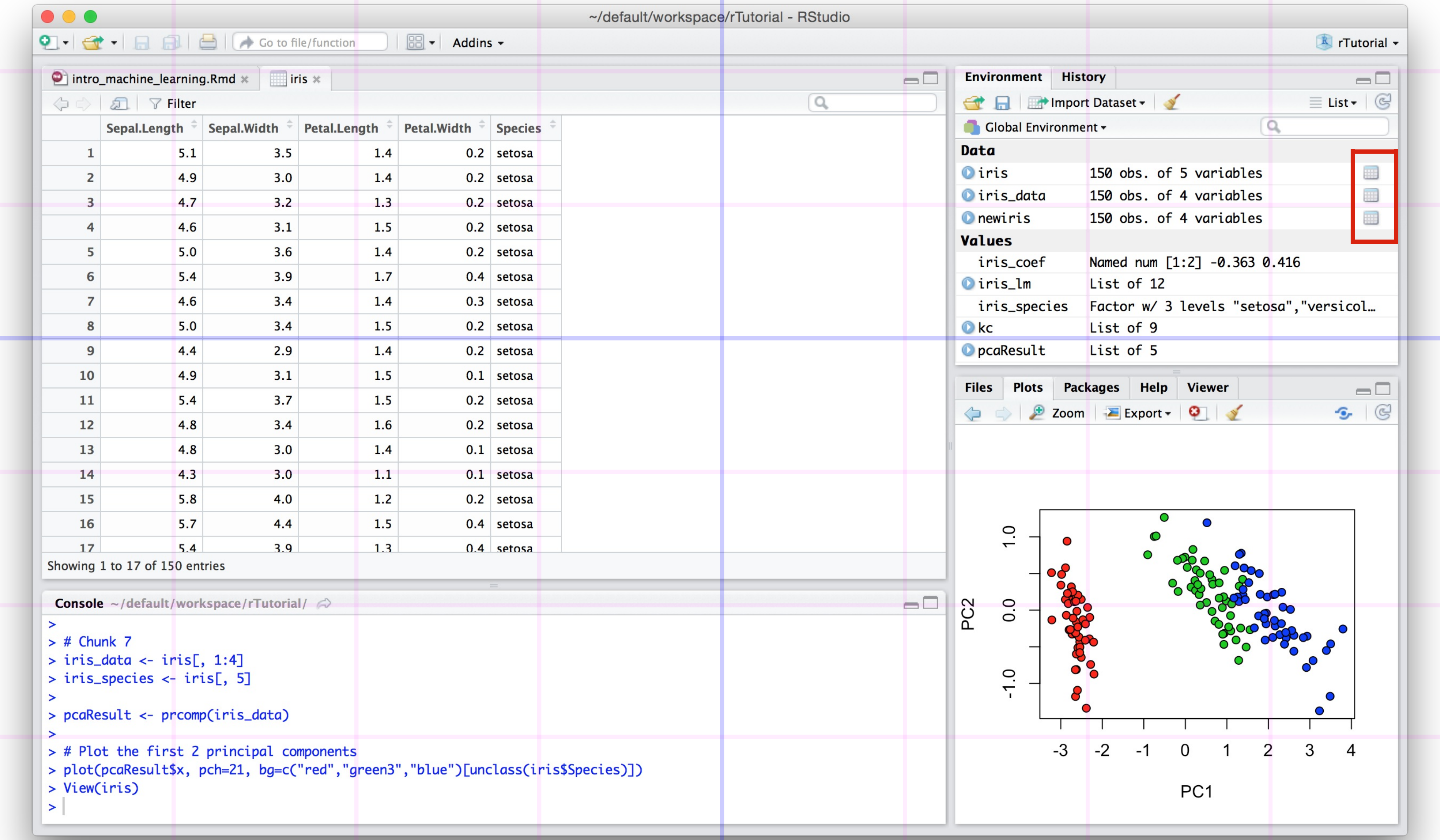
```{r}
iris_data <- iris[, 1:4]
iris_species <- iris[, 5]
```

Data	
iris	
iris_dat	
newiris	
Values	
iris_coe	
iris_lm	List of 12
iris_species	Factor w/ 3 levels "setosa","versicol…
kc	List of 9
pcaResult	List of 5

```
> points(kc$centers[,c("Se                    green3","blue"), pch=8, cex=2)
>
> # Chunk 7
> iris_data <- iris[, 1:4]
> iris_species <- iris[, 5]
>
> pcaResult <- prcomp(iris_data)
>
> # Plot the first 2 principal components
> plot(pcaResult$x, pch=21, bg=c("red","green3","blue")[unclass(iris$Species)])
>
```

Table View of Variables

- Highlighted boxes open a table view of variable contents



	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

Showing 1 to 17 of 150 entries

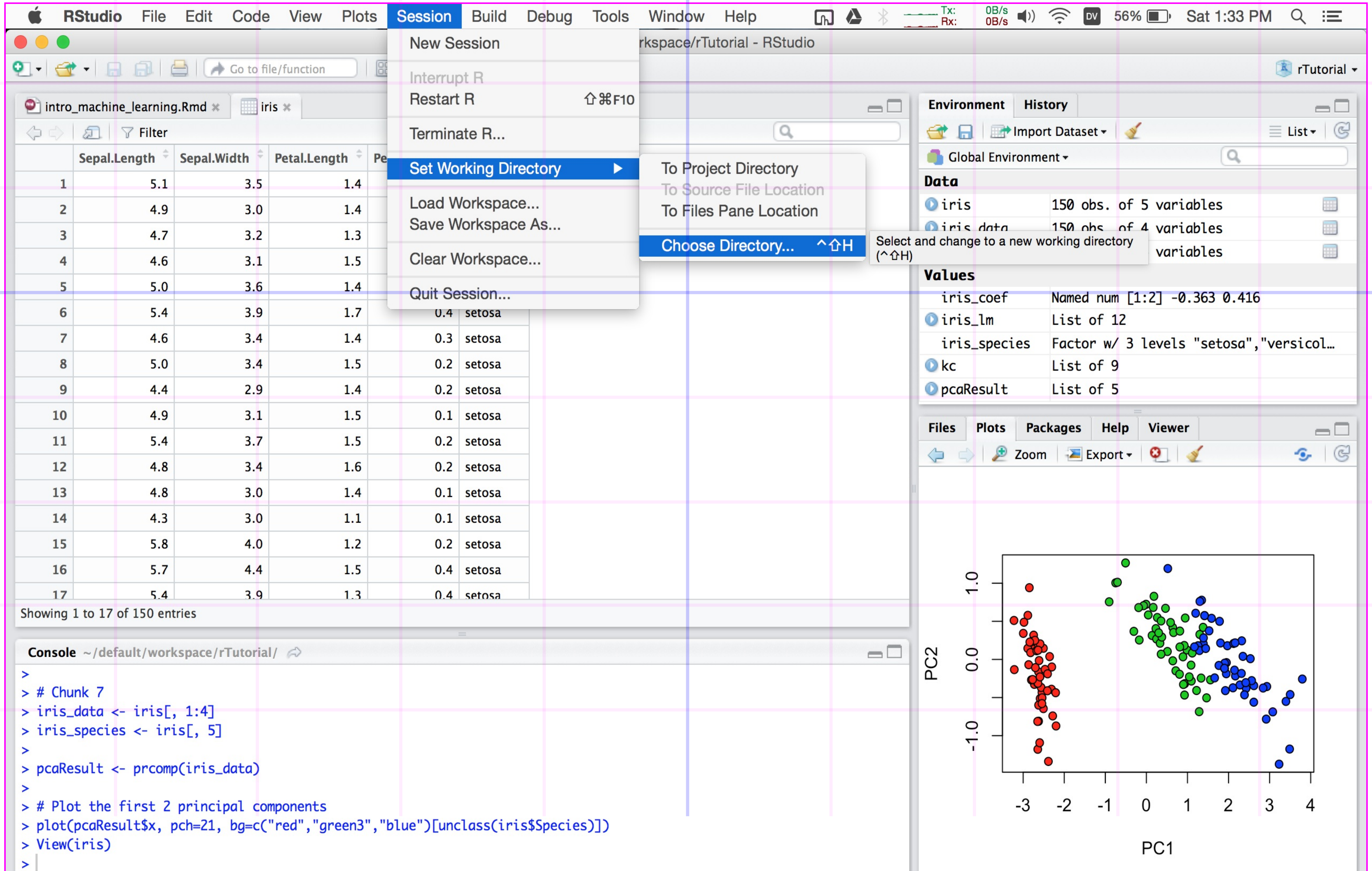
Data	
iris	150 obs. of 5 variables
iris_data	150 obs. of 4 variables
newiris	150 obs. of 4 variables
Values	
iris_coef	Named num [1:2] -0.363 0.416
iris_lm	List of 12
iris_species	Factor w/ 3 levels "setosa","versicol…
kc	List of 9
pcaResult	List of 5

Console ~/default/workspace/rTutorial/

```
>
> # Chunk 7
> iris_data <- iris[, 1:4]
> iris_species <- iris[, 5]
>
> pcaResult <- prcomp(iris_data)
>
> # Plot the first 2 principal components
> plot(pcaResult$x, pch=21, bg=c("red","green3","blue")[unclass(iris$Species)])
> View(iris)
>
```

Change Current Directory

- Highlighted boxes open a table view of variable contents



Select and change to a new working directory (^⇧H)

	Sepal.Length	Sepal.Width	Petal.Length	Pe	
1	5.1	3.5	1.4		
2	4.9	3.0	1.4		
3	4.7	3.2	1.3		
4	4.6	3.1	1.5		
5	5.0	3.6	1.4		
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

Showing 1 to 17 of 150 entries

Data

iris	150 obs. of 5 variables
iris_data	150 obs. of 4 variables
	variables

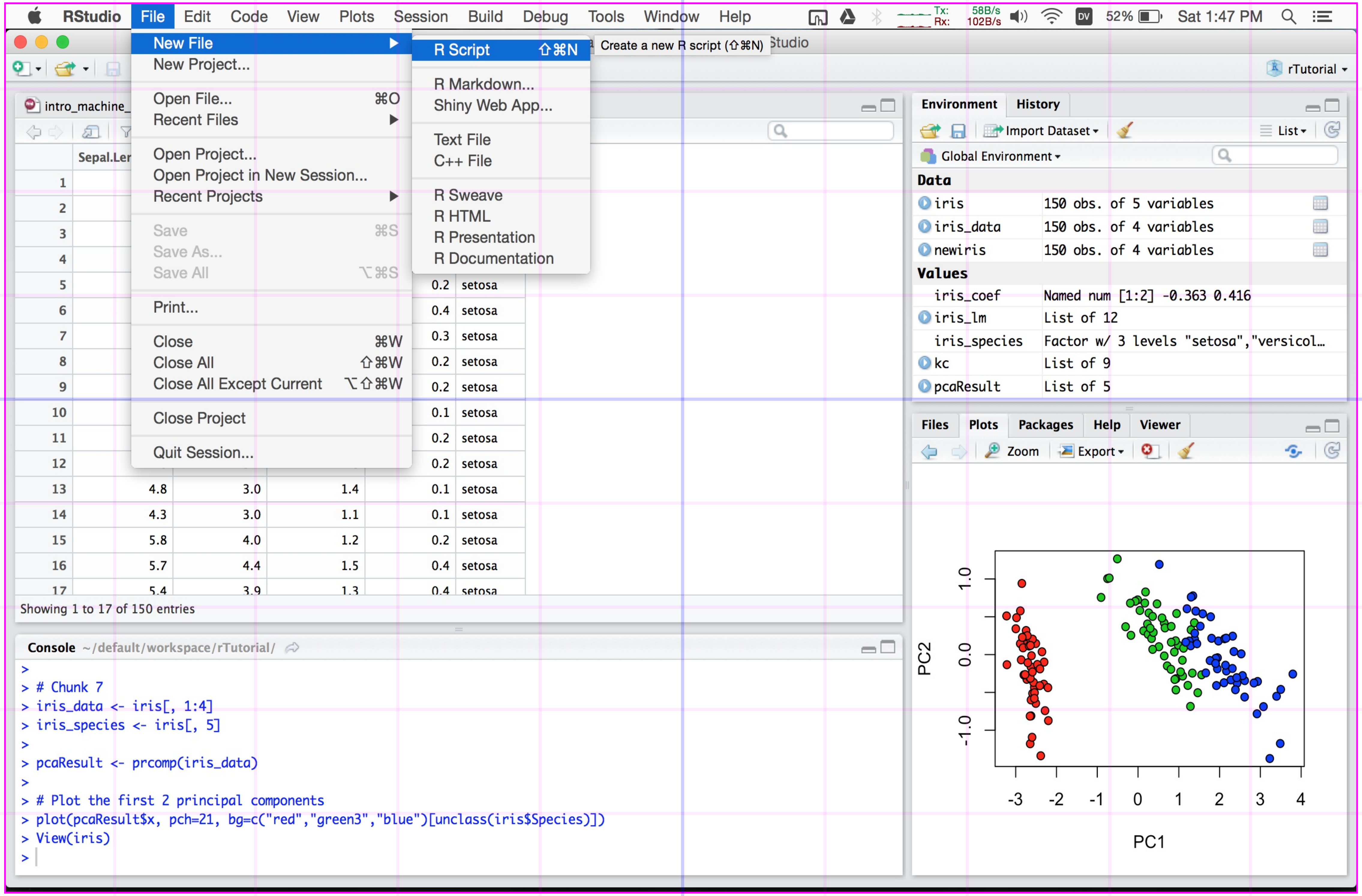
Values

iris_coef	Named num [1:2] -0.363 0.416
iris_lm	List of 12
iris_species	Factor w/ 3 levels "setosa","versicol…
kc	List of 9
pcaResult	List of 5

Console ~/default/workspace/rTutorial/

```
>
> # Chunk 7
> iris_data <- iris[, 1:4]
> iris_species <- iris[, 5]
>
> pcaResult <- prcomp(iris_data)
>
> # Plot the first 2 principal components
> plot(pcaResult$x, pch=21, bg=c("red","green3","blue")[unclass(iris$Species)])
> View(iris)
>
```

Making a New R Script



	Sepal.Ler				
1					
2					
3					
4					
5				0.2	setosa
6				0.4	setosa
7				0.3	setosa
8				0.2	setosa
9				0.2	setosa
10				0.1	setosa
11				0.2	setosa
12				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

Showing 1 to 17 of 150 entries

Data	
iris	150 obs. of 5 variables
iris_data	150 obs. of 4 variables
newiris	150 obs. of 4 variables
Values	
iris_coef	Named num [1:2] -0.363 0.416
iris_lm	List of 12
iris_species	Factor w/ 3 levels "setosa","versicol…
kc	List of 9
pcaResult	List of 5

Console ~/default/workspace/rTutorial/

```
>
> # Chunk 7
> iris_data <- iris[, 1:4]
> iris_species <- iris[, 5]
>
> pcaResult <- prcomp(iris_data)
>
> # Plot the first 2 principal components
> plot(pcaResult$x, pch=21, bg=c("red","green3","blue")[unclass(iris$Species)])
> View(iris)
>
```

First Script: Hello World!

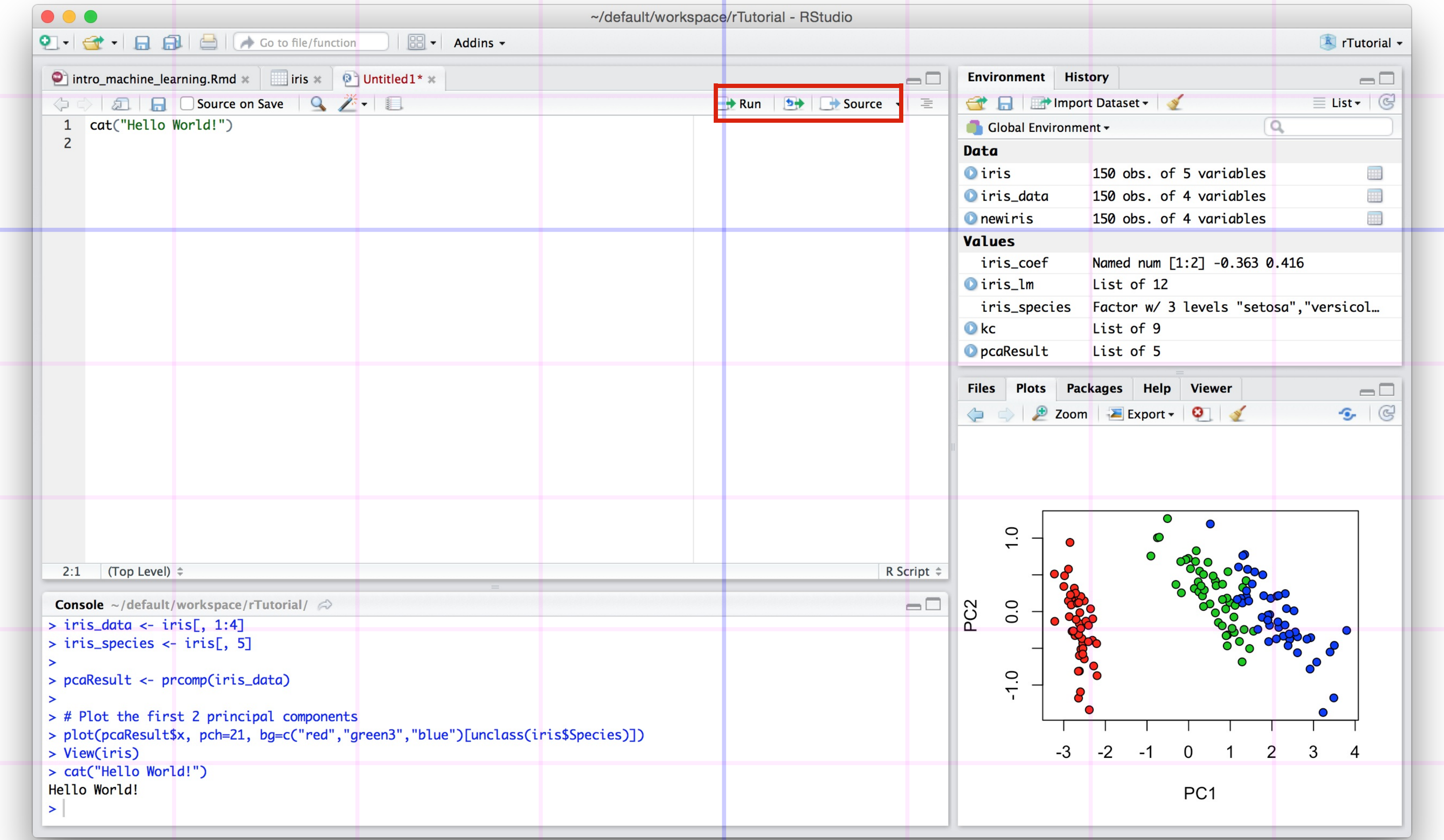
- `cat()` prints a simple message in the console

```r
cat("Hello World!")
```

```
Hello World!
```

Running Hello World Script

- “Run” button runs current line or selected lines
- “Source” button runs all lines in file



```
cat("Hello World!")
```

Data

iris	150 obs. of 5 variables
iris_data	150 obs. of 4 variables
newiris	150 obs. of 4 variables

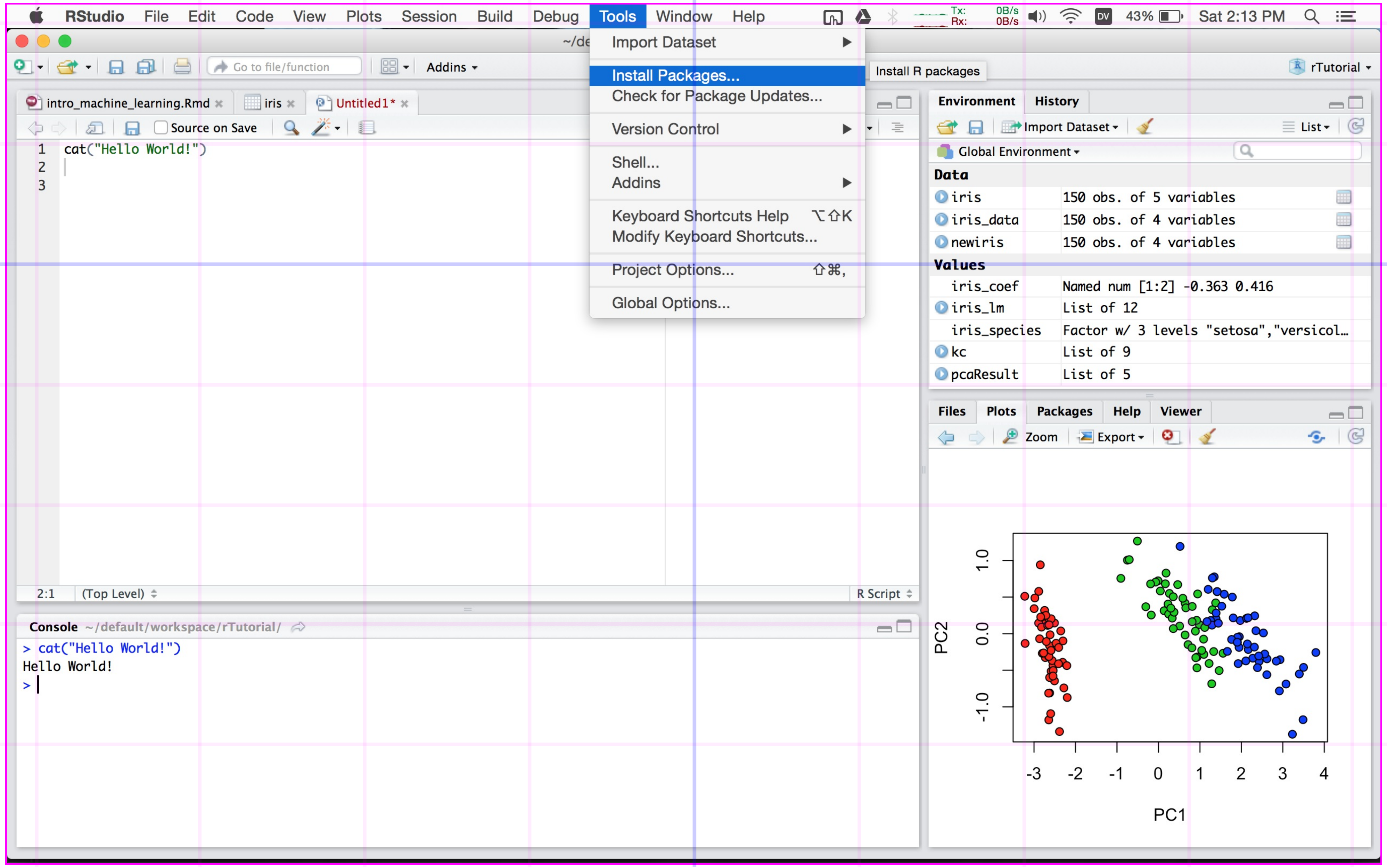
Values

iris_coef	Named num [1:2] -0.363 0.416
iris_lm	List of 12
iris_species	Factor w/ 3 levels "setosa","versicol…
kc	List of 9
pcaResult	List of 5

```
> iris_data <- iris[, 1:4]
> iris_species <- iris[, 5]
>
> pcaResult <- prcomp(iris_data)
>
> # Plot the first 2 principal components
> plot(pcaResult$x, pch=21, bg=c("red","green3","blue")[unclass(iris$Species)])
> View(iris)
> cat("Hello World!")
Hello World!
> 
```

Installing Packages

- CRAN packages can be installed using RStudio or `install.packages()`



```
cat("Hello World!")
```

Console ~/default/workspace/rTutorial/

```
> cat("Hello World!")
Hello World!
>
```

```
Data
iris            150 obs. of 5 variables
iris_data       150 obs. of 4 variables
newiris         150 obs. of 4 variables
Values
iris_coef       Named num [1:2] -0.363 0.416
iris_lm         List of 12
iris_species    Factor w/ 3 levels "setosa","versicol…
kc              List of 9
pcaResult       List of 5
```