BIO 4022. Análisis y manipulación de datos en R

Derek Corcoran

2018-07-18

Contents

Parte I

Chapter 1

Requerimientos

La última versión de RStudio y R (R Core Team 2018), también se requiere de los paquetes *tidyverse* y *tinytex*. El código para la instalación de esos paquetes es el siguiente

```
install.packages("tidyverse", "tinytex")
```

Si necesitan ayuda para la instalación contactarse con elinstructor del curso.

1.1 Si nunca has usado R antes

Si nunca han usado `R` antes de este curso, porfavor installar el paquete Swirl y realizar los primeros 7 modulos del programa *R Programming: The basics of programming in R* que incluye:

- Basic Building Blocks
- Workspace and Files
- Sequences of Numbers
- Vectors
- Missing Values
- Subsetting Vectors
- Matrices and Data Frames

Pueden ver un video explicativo de como usar swirl en el siguiente link

1.2 Objetivo del curso

Aprender los principios de investigación reproducible y tidy data a través del aprendizaje de programación y uso de R. Los principios de este curso están explicados en los siguientes libros gratuitos.

- Gandrud, Christopher. Reproducible Research with R and R Studio. CRC Press, 2013. Available for free in the following link
- Stodden, Victoria, Friedrich Leisch, and Roger D. Peng, eds. Implementing reproducible research. CRC Press, 2014. Available for free in the following link

1.3 Contenidos

- En el Capítulo 2 aprenderemos que es una base de datos *tidy*, y como manipular estas bases de datos con el paquete *dplyr* (Wickham et al. 2018)

Chapter 2

Tidy Data y manipulación de datos

En este capítulo explicaremos que es una base de datos *tidy* y aprenderemos a usar funciones del paquete *dplyr* (Wickham et al. 2018) para manipular los datos.

Chapter 3

Investigación reproducible

Here is a review of existing methods.

Chapter 4

El Tidyverso

We describe our methods in this chapter.

Chapter 5

Visualización de datos

En este capítulo aprenderemos a usar el paquete *ggplot2* (Wickham 2016), parte del paquete *tidyverse* (Wickham 2017).

5.1 El esqueleto

El esqueleto de una visualización usando *ggplot2* es la siguiente

```
ggplot(data.frame, aes(nombres de columna)) + geom_algo(argumentos, aes(columnas)) + theme_algo()
```

Como ejemplo para discutir usaremos el siguiente código que genera la figura 5.1:

```
library(tidyverse)
data("diamonds")
ggplot(diamonds, aes(x = carat, y=price)) + geom_point(aes(color = cut)) + theme_classic()
```

En este caso general, lo primero que ponemos después de ggplot es el data.frame desde el cuál graficaremos algo, en el ejemplo de la figura 5.1 usamos la base de datos *diamonds* del paquete *ggplot2* (Wickham 2016). Luego dentro de `aes` ponemos las columnas que graficaremos como x y/o y, en nuestro ejemplo dentro de aes ponemos como eje x los kilates de los diamantes (caret) y como y el precio de los mismos (price). La necesidad de poner `aes` en ggplot2 (algo que no había sido necesario cuando usamos *dplyr* o *tidyr*) es que ggplot2 es el paquete mas antiguo del *tidyverse*.

5.2 geom_algo

Luego de especificar una base de datos, esto viene seguido de un `geom_algo`, esto nos indicará que tipo de gráfico usaremos, estos pueden ser combinados como veremos en ejemplos futuros

5.2.1 Una variable categórica una continua

Primero veremos algunos de los *geom* que podemos utilizar con una variable categórica y una continua

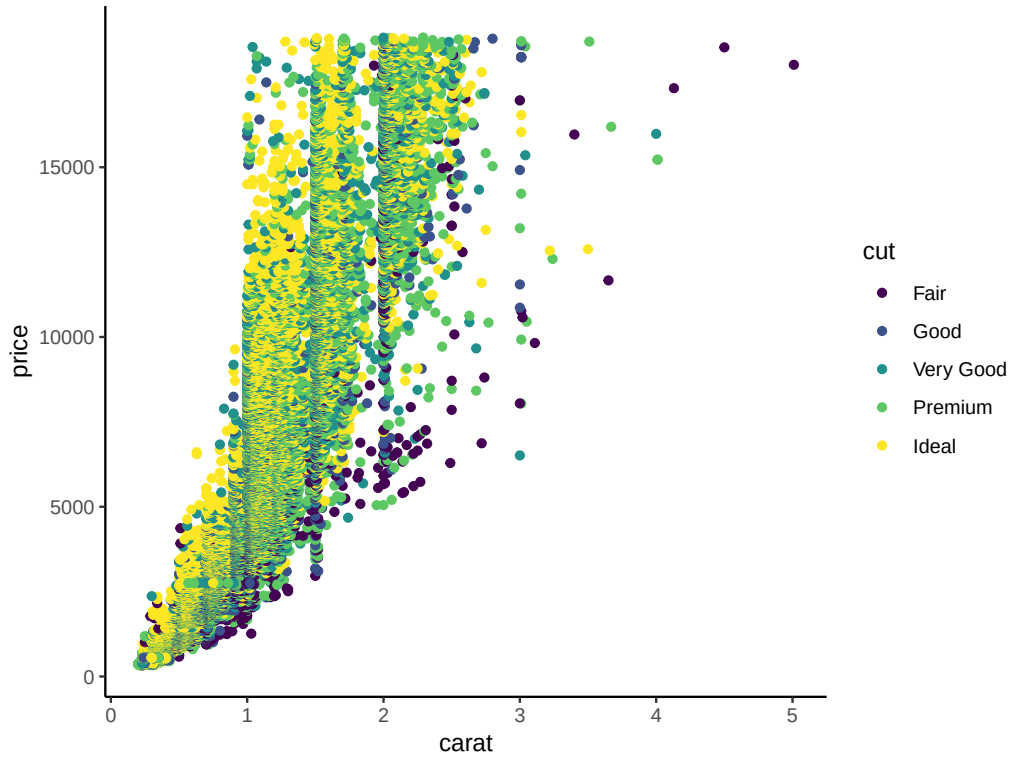

Figure 5.1: Gráfico en el cual gráficamos los quilates de diamantes versus su precio, con el corte del diamante representado por el color

5.2.1.1 geom_boxplot

En la figura 5.2, generado a partir del código a continuacón con la base de datos iris presente en `R` (Anderson 1935).

```
data("iris")
ggplot(iris, aes(x = Species, y = Sepal.Length)) + geom_boxplot()
```

Los boxplots muestran una linea gruesa central (la mediana), una caja, que delimita el primer y tercer cuartil, y los bigotes, los cuales se extienden hasta los valores extremos. A menos que estos esten por sobre 1.5 veces la distance entre el primer y tercer cuartil, en cuyo caso se consideran outlyers, y estos son representados por puntos. En la figura 5.2, solo *Iris virginica* presenta un outlayer en cuanto a las medidas del largo del sepalo.

Los boxplots, como todos los gráficos pueden ser personalizados usando otros argumentos, los cuales son detallados en la sección 5.3, pero en los ejemplos que mostraremos en esta sección los iremos introduciendo de a poco. Si quisieramos por ejemplo que el color de las cajas del *boxplot* fuera deacuerdo a la especie, cambiamos el llenado (**fill**) de la caja, como vemos en el siguiente ejemplo y figura 5.3

```
ggplot(iris, aes(x = Species, y = Sepal.Length)) + geom_boxplot(aes(fill = Species))
```

Dos cosas a notar en este ejemplo, por un lado la leyenda se genera de forma automática, y por otro lado, vemos que es necesario poner *Species* dentro de `aes`, esto es debido a que Species es una columna y como se explicó al principio de este capítulo, todas las columnas deben ser incuidas dentro de la función `aes` para poder ser referenciadas.

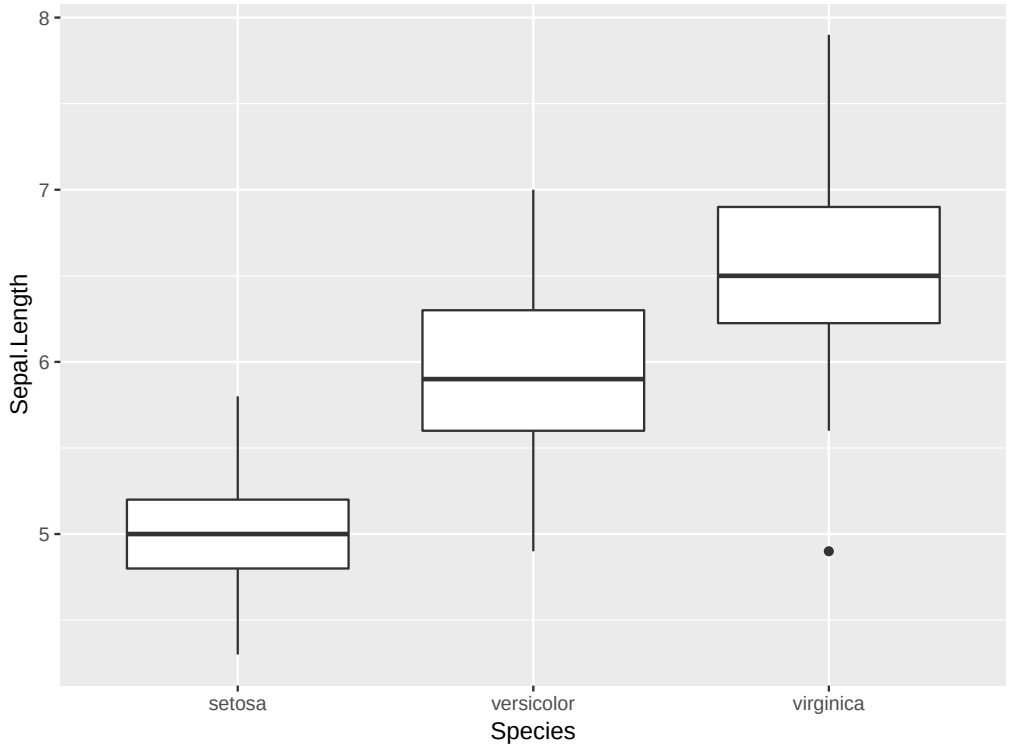

Figure 5.2: Boxplot que representa los largos del sépalo de tres especies del género Iris

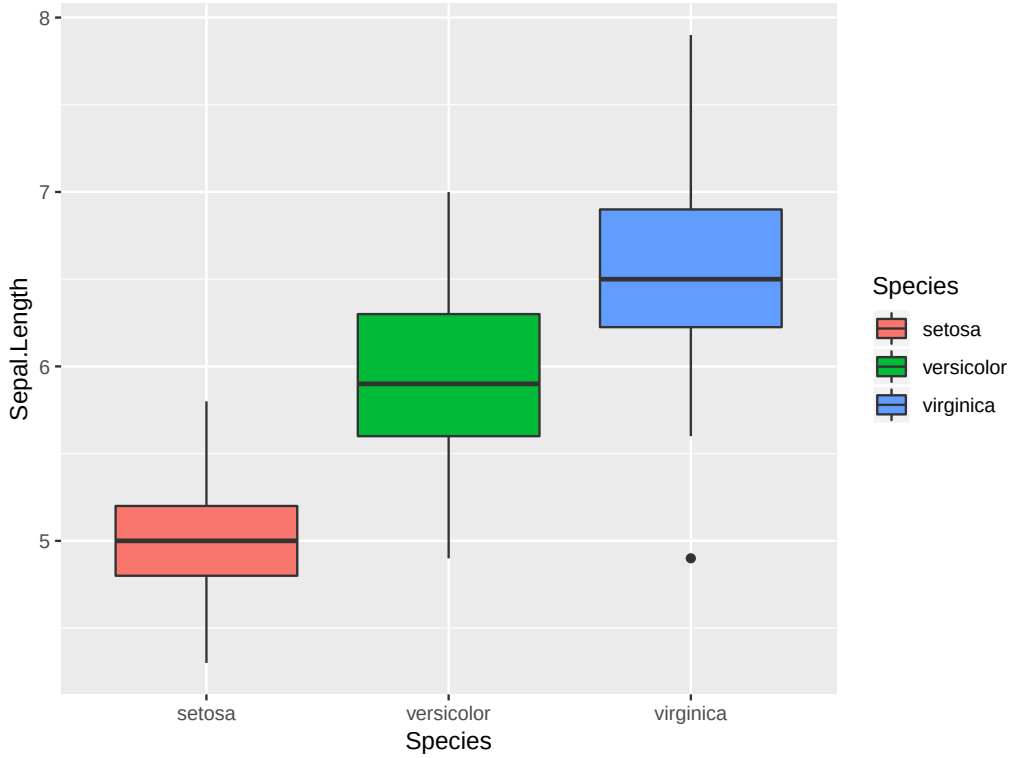

Figure 5.3: Boxplot que representa los largos del sépalo de tres especies del género Iris, en este caso el color de la caja representa la especie

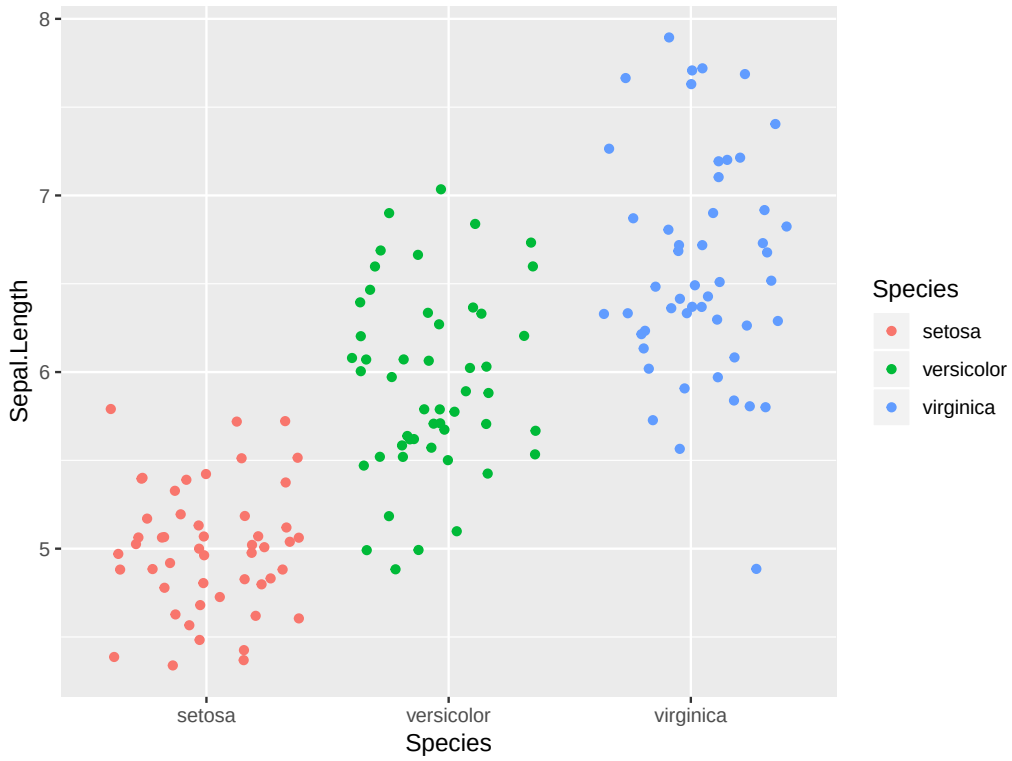

Figure 5.4: Boxplot que representa los largos del sépalo de tres especies del género Iris, en este caso el color de la caja representa la especie

5.2.1.2 geom_jitter

```
ggplot(iris, aes(x = Species, y = Sepal.Length)) + geom_jitter(aes(color = Species))
```

5.3 Argumentos

Chapter 6

Modelos en R

We have finished a nice book.

Chapter 7

Loops (purrr) y bibliografía (rticles)

Chapter 8

Presentaciones en R

Anderson, Edgar. 1935. "The Irises of the Gaspe Peninsula." *Bulletin of the American Iris Society* 59: 2–5.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. http://ggplot2.org.

———. 2017. *Tidyverse: Easily Install and Load the 'Tidyverse'*. https://CRAN.R-project.org/package=tidyverse.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2018. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.