DATASCIENCE - DAY 1

Tutorial ggplot2

Germán Leandro Contreras Sagredo

5 de Agosto, 2016

Pontificia Universidad Católica de Chile - SocVis

INTRODUCCIÓN

OBJETIVOS

Los objetivos de este tutorial consisten en:

- Conocer los comandos básicos de ggplot2.
- Añadir estilos a un gráfico básico (histograma).

¿QUÉ ES GGPLOT2?

ggplot2 adquiere su nombre de **'The Grammar of Graphics'** (la gramática de los gráficos), pues su creación fue basada en el libro del mismo nombre, de Leland Wilkinson (estadístico). Corresponde a una librería dentro del programa R, por lo que es necesario tener al menos un manejo básico con este. Está enfocada en la realización de gráficos de diversos tipos (por ejemplo, gráficos de barras, gráficos de densidad, entre otros).

CONSIDERACIÓN IMPORTANTE

Esta librería **no se enfoca en**:

- Gráficos tridimensionales.
- Grafos.
- Gráficos interactivos.

Para ellos existen otras librerías especializadas.

ELEMENTOS BÁSICOS

Como toda librería en R, se debe **instalar**. Esto se puede hacer a través del siguiente comando:

```
install.packages("ggplot2")
```

Luego, para su uso:

```
library("ggplot2")
```

Ahora se pueden utilizar todos los comandos sin problemas.

TIPO DE DATOS

ggplot2 trabaja con marcos de datos (data frames), los que son similares a una matriz, salvo por la diferencia que este puede tener distintos tipos de datos en cada columna. Suponga la siguiente tabla:

Cuadro: Charlistas de DataScience Day 1 (nombre de la variable: çharlistas")

Charlista	Charla	Asistentes
Denis	R 101	50
Germán	ggplot2	39
Daniela	Dashboards	47
Vicente	MatplotLib	42

EJEMPLO - DATA FRAME

La tabla anterior se puede definir como sigue:

```
# Notar que c(x1,x2,...,xn) corresponde a un vector de n variables.
charlistas = data.frame(
              charlista = c("Denis","German","Daniela","Vicente"),
              charla = c("R 101", "ggplot2", "Dashboards", "MatplotLib"),
              asistentes = c(50,39,47,42)
              )
```

Luego, se tienen distintos accesos:

```
charlistas\$charlista
# [1] Denis   German  Daniela Vicente
charlistas\$charlista[3]
# [1] German
charlistas[2]
#         charla
# 1      R 101
# 2    ggplot2
# 3 Dashboards
# 4 MatplotLib
```

EJEMPLO - LECTURA DE ARCHIVOS

Si los datos se encontraran en un archivo de texto (un CSV, por ejemplo), se pueden pasar a una variable de forma directa:

```
# Con head=TRUE se guardan los nombres de cada columna.
# Al ser un CSV, la separación de cada dato se da por una coma, he ahí
# sep=",".
read.table(file="Charlistas.csv",
         head=TRUE,
         sep=","
         )

# Esto se puede simplificar con read.csv, que configura de forma
# predeterminada las variables "head" y "sep".
read.csv(file="Charlistas.csv")

# También se puede hacer con otros tipos. Por ejemplo, un TSV sin encabezado.
# La separación cambia, y añadimos manualmente el encabezado.
read.table(file="Charlistas.tsv",
           head=FALSE,
           sep="\t",
           col.names = c("charlista","charla","asistentes")
          )
```

EJEMPLO - OTRO TIPO DE DATOS

Se puede dar el caso en el que el conjunto de datos esté en un formato similar, pero diferente finalmente al necesitado (por ejemplo, una matriz). En este caso, basta con utilizar cualquiera de los siguientes comandos:

```
# Matriz de 3x2 con números del 1 al 6.
x = matrix(data = 1:6, nrow = 3, ncol = 2)
# Transformación a marco de datos - Forma 1
y = as.data.frame(x)
#   V1 V2
# 1  1  4
# 2  2  5
# 3  3  6

# Transformación a marco de datos - Forma 2
z = data.frame(x)
#   X1 X2
# 1  1  4
# 2  2  5
# 3  3  6
```

EJEMPLO - OTRO TIPO DE DATOS

En este caso, la diferencia radica en el nombre estándar que adquieren las columnas (en una matriz, puntualmente, se les puede dar un nombre también). as.data.frame suele ser más rápido, pues el otro método hace uso de este.

COMENZANDO A GRAFICAR

EJEMPLO - HISTOGRAMA

Primero, se hará un histograma simple con los datos que se tienen a disposición, y de a poco se le irán añadiendo más elementos gráficos.

HISTOGRAMA - PARTE 1

```
# Histograma básico, con eje X rating y eje Y la cantidad de tiendas
# con dicho rating.
# geom_histogram es el que hace que el tipo de gráfico sea un histograma.
plot = (ggplot(data = ejemplo1, aes(x = rating))
       + geom_histogram())
plot
```

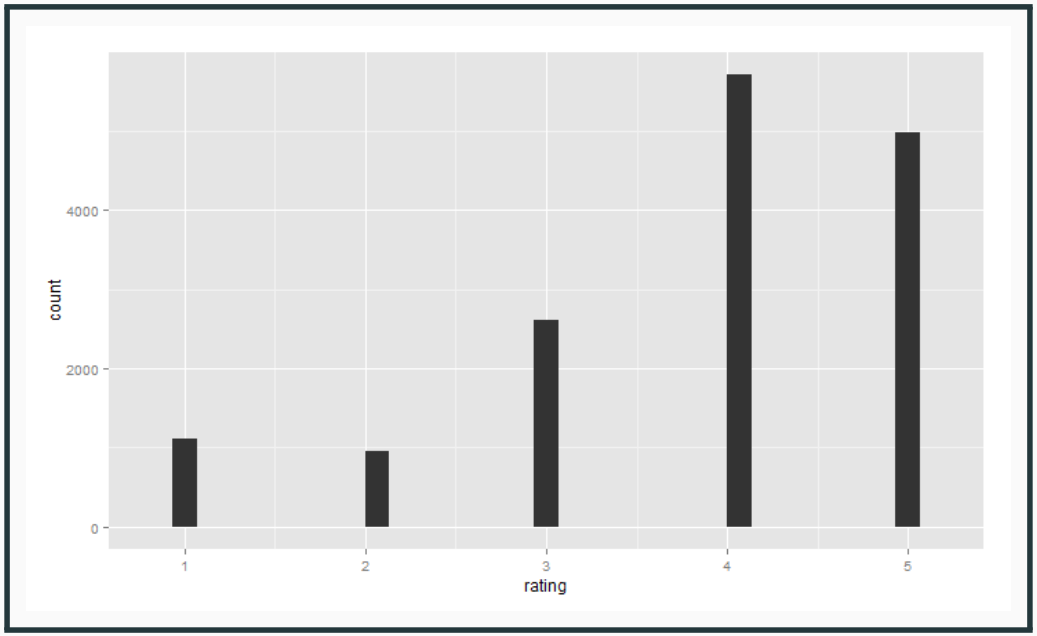

Figura: La visualización resultante: Un histograma simple.

HISTOGRAMA - PARTE 2

```
# Vamos a editar la información básica. Cambiaremos los nombres de los ejes
# y le daremos un nombre a nuestro gráfico.
plot = (ggplot(data = ejemplo1, aes(x = rating))
        + geom_histogram()
        + xlab("Rating")
        + ylab("Cantidad de tiendas")
        + ggtitle("Frecuencia de tiendas por rating"))
plot
```

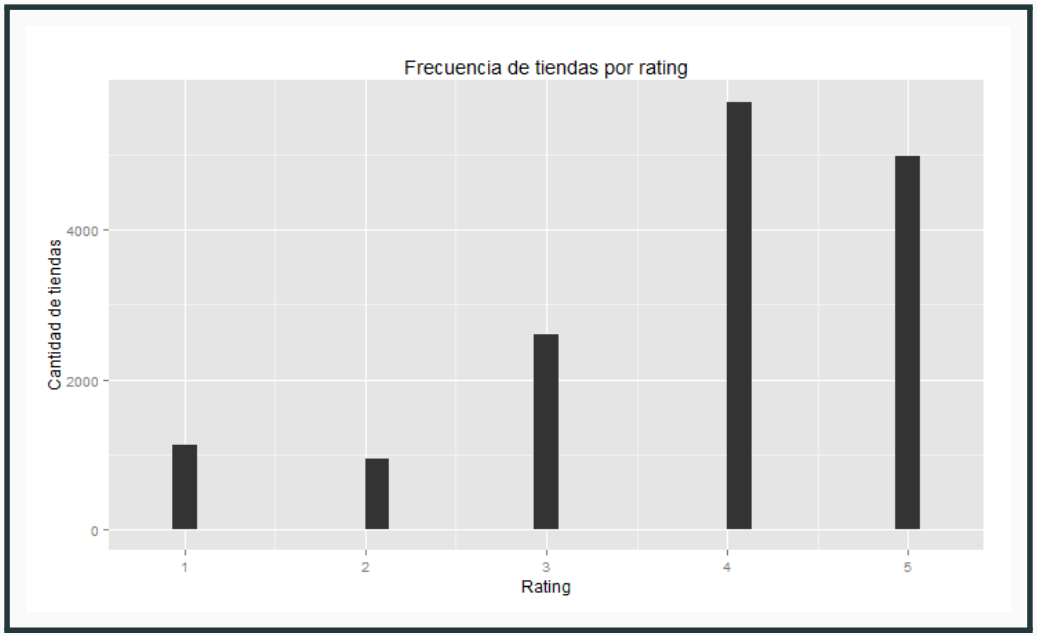

Figura: La visualización resultante: Un histograma simple con etiquetas y título.

HISTOGRAMA - PARTE 3

```
# Ahora, le añadiremos estilo al gráfico. Haremos que cada barra posea un color similar al del logo de
# Second Life, además de añadirle un marco. Por otra parte, aumentaremos el ancho de nuestras barras,
# además de alinearlas con su respectiva etiqueta.
plot = (ggplot(data = ejemplo1, aes(x = rating))
        + geom_histogram(colour = "darkslategray", fill = "darkslategray4",
                         binwidth = 0.5)
        + scale_x_discrete(limits = c(1,5), expand = c(0.05,0.05))
        + xlab("Rating")
        + ylab("Cantidad de tiendas")
        + ggtitle("Frecuencia de tiendas por rating"))
  plot
```

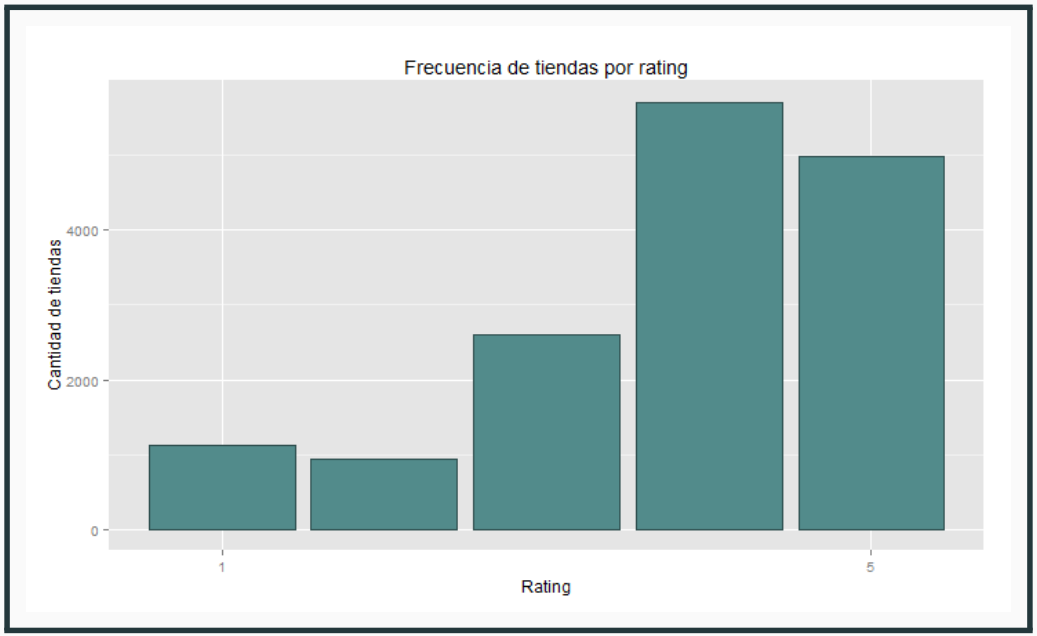

Figura: La visualización resultante: Un histograma a color.

HISTOGRAMA - PARTE 4

```
# Ahora, en vez de tener en el eje Y la cantidad de tiendas, tendremos
# la densidad de tiendas. Además, pondremos sobre el histograma el gráfico de la densidad.
plot = (ggplot(data = ejemplo1, aes(x = rating, y = ..density..))
        + geom_histogram(colour = "darkslategray", fill = "darkslategray4",
                         binwidth = 0.5)
        + geom_density(colour = "darkturquoise")
        + scale_x_discrete(limits = c(1,5), expand = c(0.05,0.05))
        + xlab("Rating")
        + ylab("Densidad de tiendas")
        + ggtitle("Frecuencia de tiendas por rating"))
plot
```

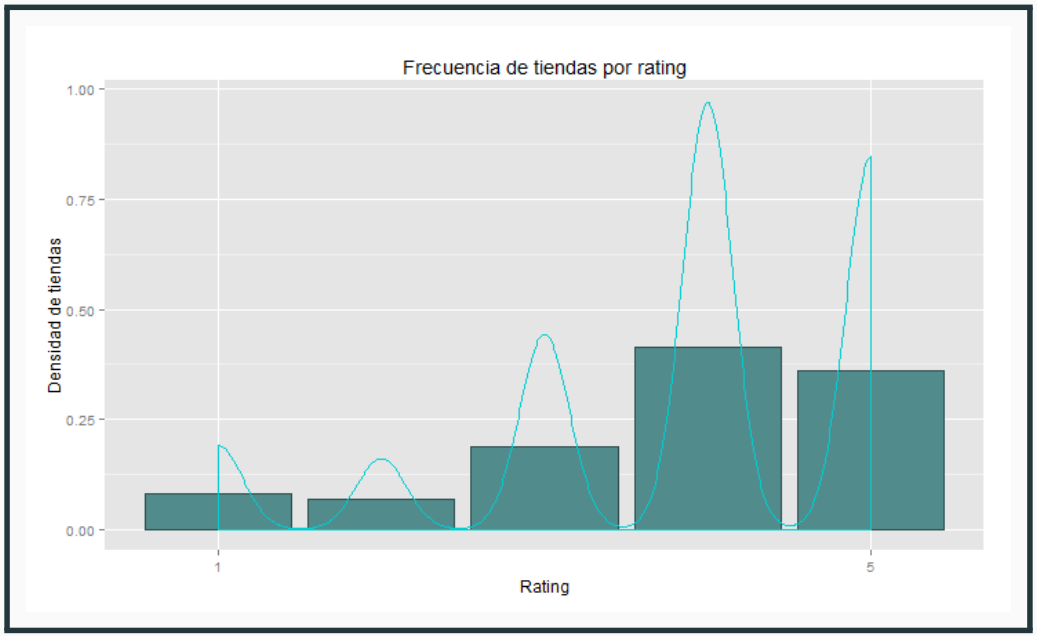

Figura: La visualización resultante: Un histograma a color con su densidad.

HISTOGRAMA - PARTE 5

```
# Si no quisiéramos perder la frecuencia obtenida, podemos dejarla sobre cada barra.
plot = (ggplot(data = ejemplo1, aes(x = rating, y = ..density..))
        + geom_histogram(colour = "darkslategray", fill = "darkslategray4",
                         binwidth = 0.5)
        + geom_text(stat = "bin", aes(label = ..count..), vjust = -1, fontface = "bold",
                    colour = "darkslategray4")
        + geom_density(colour = "darkturquoise")
        + scale_x_discrete(limits = c(1,5), expand = c(0.05,0.05))
        + xlab("Rating")
        + ylab("Densidad de tiendas")
        + ggtitle("Frecuencia de tiendas por rating"))
plot
```

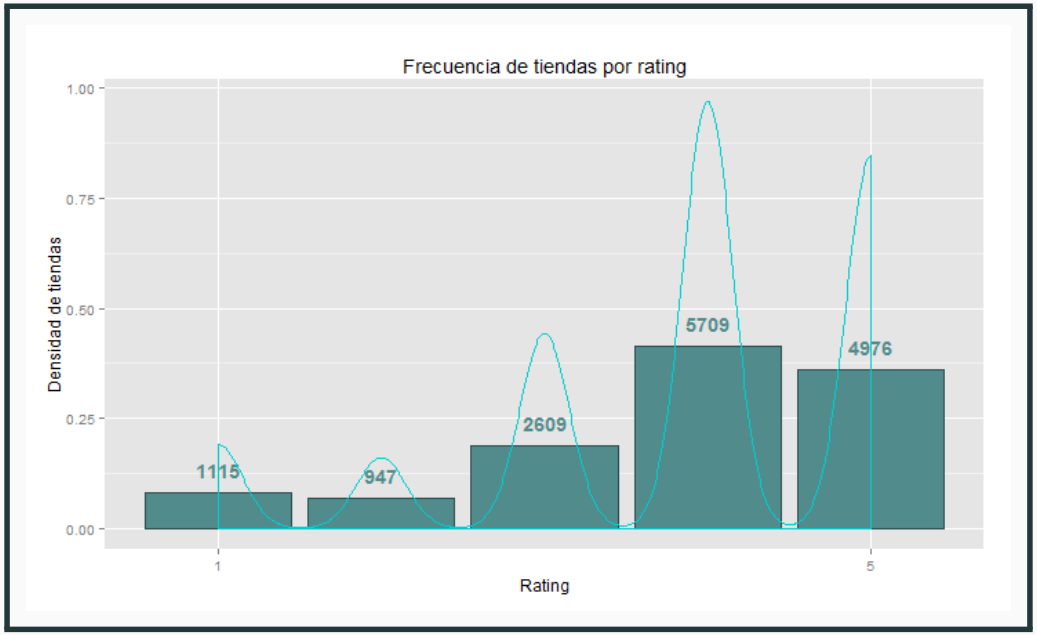

Figura: La visualización resultante: Un histograma a color con su densidad y frecuencias.

HISTOGRAMA - PARTE 6

```
# Ahora, usaremos el segundo conjunto de datos, para poder mostrar múltiples datos a la vez.
# Estos datos ahora contienen un año. Veremos dos formas de visualizar esto.
plot = (ggplot(data = ejemplo2, aes(x = rating, y = ..density..,
                                    fill = as.factor(year)))
        + geom_histogram(colour = "darkslategray", binwidth = 0.5, position = "dodge")
        + scale_x_discrete(limits = c(0,5), expand = c(0.05,0.05))
        + xlab("Rating")
        + ylab("Densidad de tiendas")
        + ggtitle("Frecuencia de tiendas por rating"))
plot
```

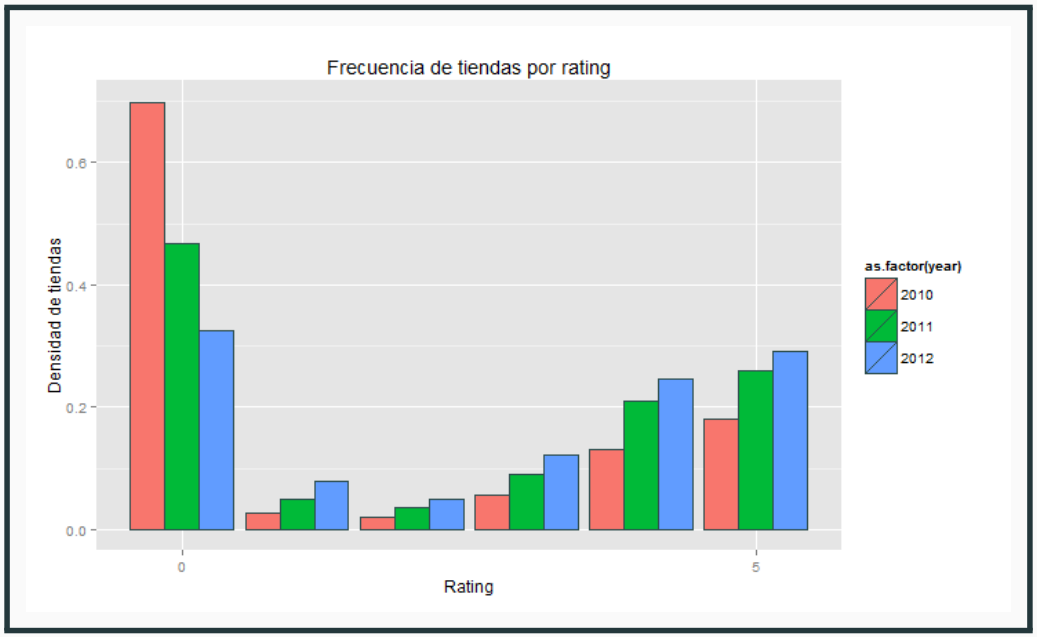

Figura: La visualización resultante: Un histograma a color, siendo el color el representante de otra variable del conjunto de datos.

HISTOGRAMA - PARTE 7

```
# Pasamos al otro método.
plot = (ggplot(data = ejemplo2, aes(x = rating, y = ..density..))
        + geom_histogram(colour = "darkslategray", fill = "darkslategray4",
                         binwidth = 0.5)
        + geom_density(colour = "darkturquoise")
        + scale_x_discrete(limits = c(0,5), expand = c(0.05,0.05))
        + xlab("Rating")
        + ylab("Densidad de tiendas")
        + ggtitle("Frecuencia de tiendas por rating")
        + facet_wrap(~ year))
plot
```

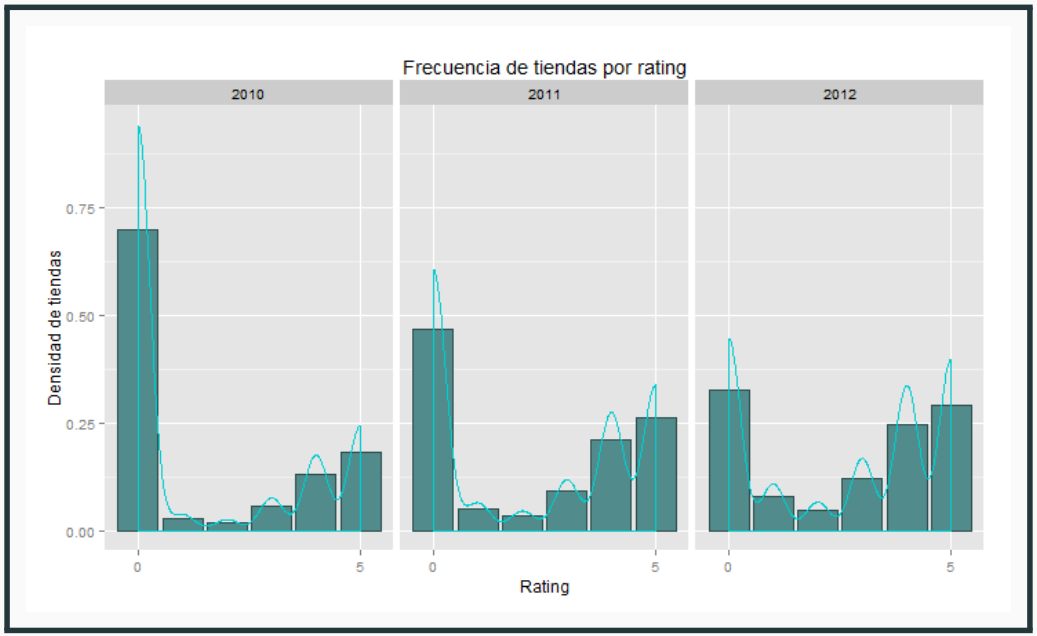

Figura: La visualización resultante: Un histograma a color con su densidad, separado en otros gráficos a partir de otra variable.

ENLACES Y RECURSOS ÚTILES

LIBRERÍAS Y TUTORIALES

- ggplot2 - Página oficial
- Cookbook for R - Graphs

FIN