

Visualización de datos en R

La visualización de datos es crucial para la interpretación de estos, ya que permite un entendimiento acertado de los datos.

Los métodos de visualización de datos son múltiples en R:

Tablas

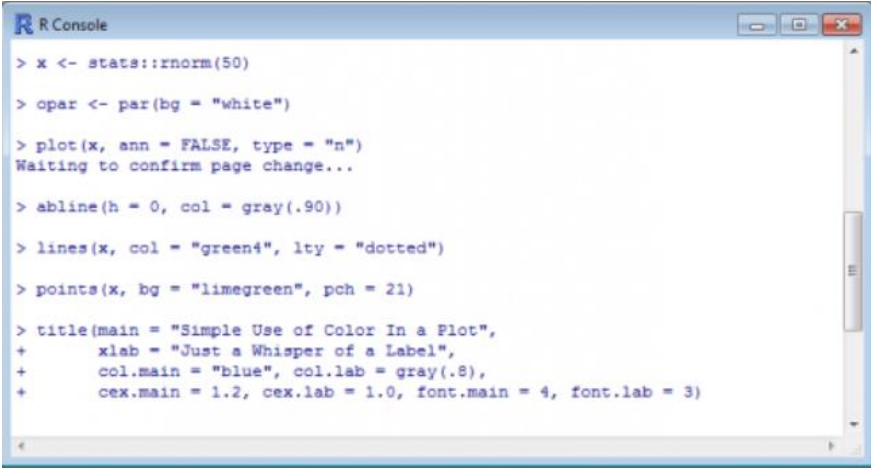
- Tablas de datos
- Tablas de frecuencia

Gráficos

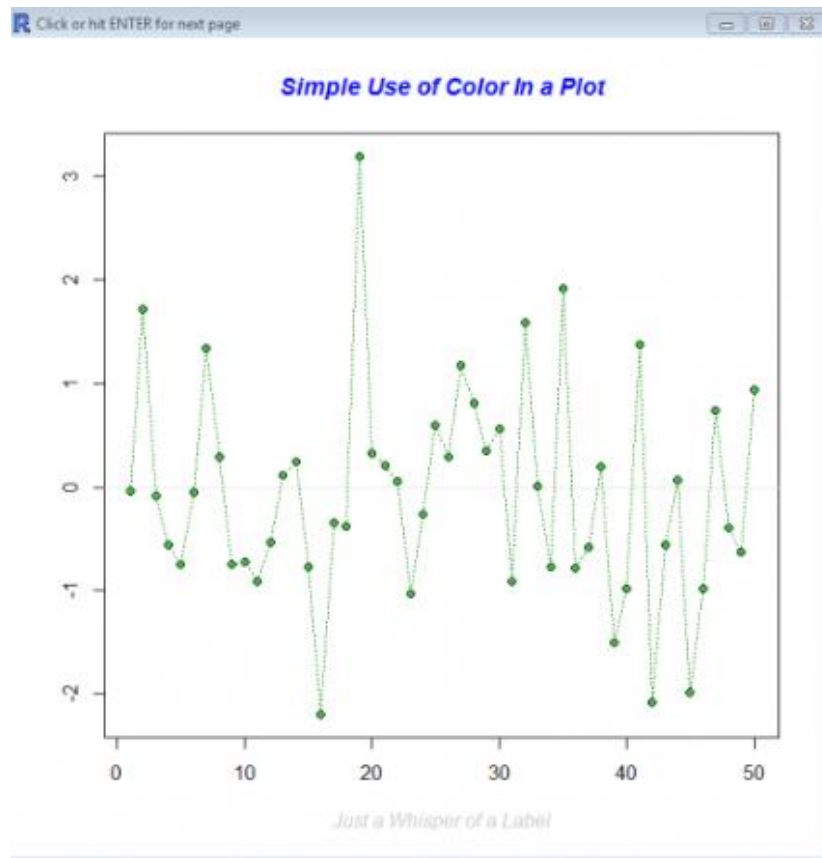
- Gráficos circulares
- Gráficos de dispersión, gráficos de líneas
- Gráficos de barras, Histogramas
- Polígonos de frecuencia
- Gráficos de caja
- Gráficos de contorno, gráficos 3D

Comando demo(graphics)

Este comando muestra algunos ejemplos que se pueden crear "fácilmente" en R usando unas pocas líneas de comando simples. Convenientemente, la consola R muestra comandos mientras construye el gráfico en la ventana de gráfico.

A screenshot of the R Console window. The window has a title bar that says "R Console" and standard Windows window controls (minimize, maximize, close). The console area shows the following R code being executed:

```
> x <- stats::rnorm(50)
> opar <- par(bg = "white")
> plot(x, ann = FALSE, type = "n")
Waiting to confirm page change...
> abline(h = 0, col = gray(.90))
> lines(x, col = "green4", lty = "dotted")
> points(x, bg = "limegreen", pch = 21)
> title(main = "Simple Use of Color In a Plot",
+       xlab = "Just a Whisper of a Label",
+       col.main = "blue", col.lab = gray(.8),
+       cex.main = 1.2, cex.lab = 1.0, font.main = 4, font.lab = 3)
```



La programación en R proporciona amplios conjuntos de herramientas, como funciones incorporadas y una amplia gama de paquetes para realizar análisis de datos, representar datos y construir visualizaciones.

La visualización de datos en R puede realizarse de las siguientes maneras:

- Base Graphics
- Grid Graphics
- Lattice Graphics
- ggplot2

Base Graphics

Hay algunos elementos clave de un gráfico estadístico. Estos elementos son los fundamentos de la gramática de los gráficos. R proporciona algunas funciones incorporadas que se incluyen en el paquete de gráficos para la visualización de datos en R. Vamos a discutir cada uno de los elementos uno por uno para obtener el conocimiento básico de los gráficos.

Vamos a utilizar el conjunto de datos (mtcars) por defecto para la visualización de datos en R.

```
#To load graphics package
library("graphics")
#To load datasets package
```

```
library("datasets")
#To load mtcars dataset
data(mtcars)
#To analyze the structure of the dataset
str(mtcars)
```

La función plot()

La función plot() se utiliza para trazar objetos de R.

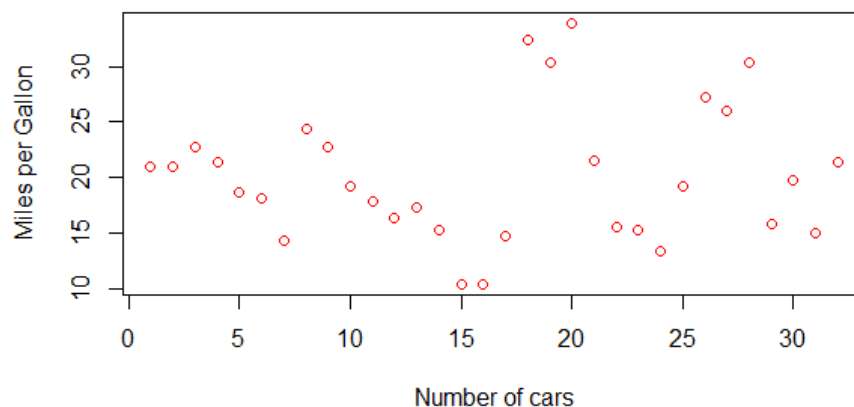
La sintaxis básica de la función plot() es la siguiente:

```
plot(x,y,type,main,sub,xlab,ylab,asp,col,..)
```

- x: La coordenada x del trazado, una estructura única de trazado, una función o un objeto R
- y: La coordenada Y de los puntos del trazado (opcional si la coordenada x es una estructura única)
- type: 'p' para puntos, 'l' para líneas, 'b' para ambos, 'h' para líneas verticales de alta densidad, etc.
- main: Título del gráfico
- sub: Subtítulo del gráfico
- xlab: Título del eje x
- ylab: Título del eje y
- asp: Relación de aspecto (y/x)
- col: Color del gráfico (puntos, líneas, etc.)

Ejemplo:

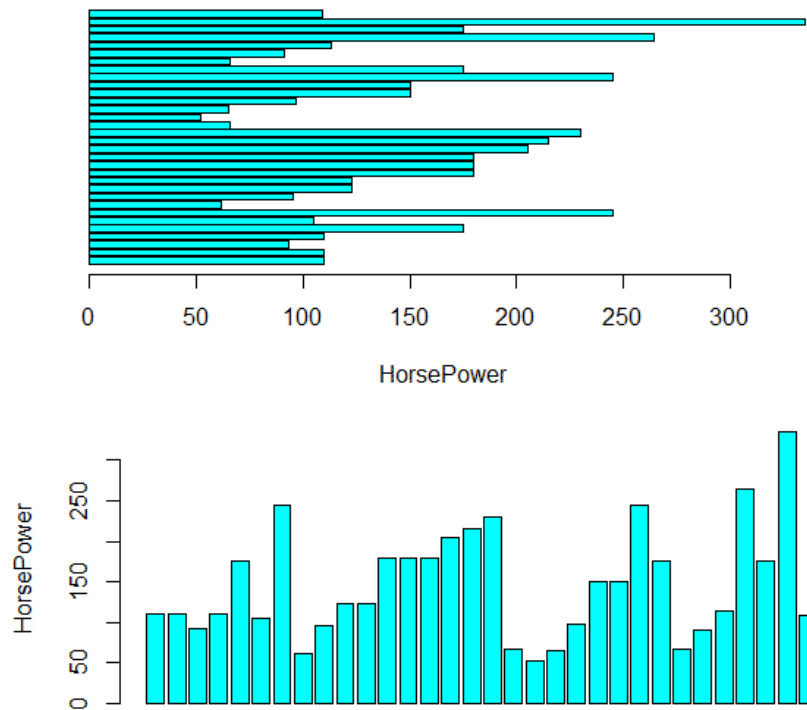
```
#To plot mpg(Miles per Gallon) vs Number of cars
plot(mtcars$mpg, xlab = "Number of cars", ylab = "Miles per Gallon", col = "red")
```



Barplot

Se utiliza para representar los datos en forma de barras rectangulares, tanto en vertical como en horizontal, y la longitud de la barra es proporcional al valor de la variable.

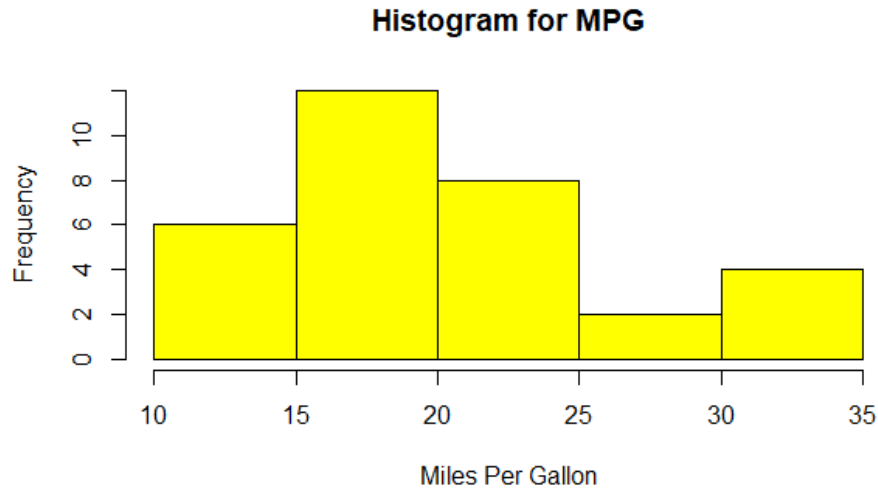
```
#To draw a barplot of hp
#Horizontal
barplot(mtcars$hp,xlab = "HorsePower", col = "cyan", horiz = TRUE)
#Vertical
barplot(mtcars$hp, ylab = "HorsePower", col = "cyan", horiz = FALSE)
```



Histograma

Se utiliza para dividir los valores en grupos de rangos continuos medidos con respecto al rango de frecuencia de la variable.

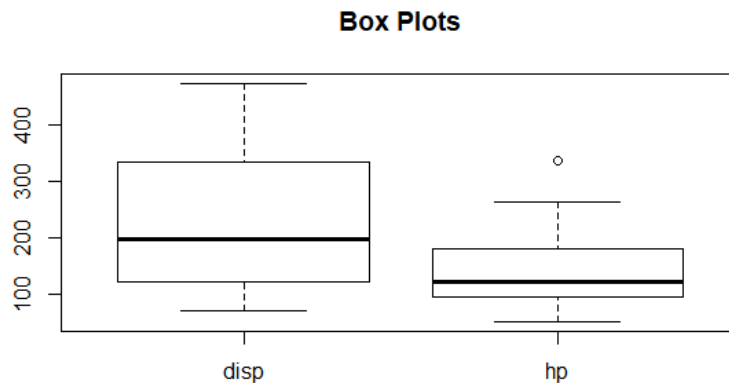
```
#To find histogram for mpg (Miles per Gallon)
hist(mtcars$mpg,xlab = "Miles Per Gallon", main = "Histogram for MPG", col = "yellow")
```



Boxplot

Se utiliza para representar las estadísticas descriptivas de cada variable en un conjunto de datos. Representa el mínimo, el primer cuartil, la mediana, el tercer cuartil y los valores máximos de una variable

```
#To draw boxplots for disp (Displacement) and hp (Horse Power)  
boxplot(mtcars[,3:4])
```



Visualización de datos en R con el paquete ggplot2

El paquete ggplot2 en R se basa en la gramática de los gráficos, que es un conjunto de reglas para describir y construir gráficos. Al dividir los gráficos en componentes semánticos, como escalas y capas, ggplot2 implementa la gramática de los gráficos.

La gramática de los gráficos de ggplot2 se compone de lo siguiente:

- Data
- Layers
- Scales

- Coordinates
- Faceting
- Themes

ggplot2 es uno de los paquetes más sofisticados de R para la visualización de datos, y ayuda a crear los más elegantes y versátiles gráficos con calidad de impresión con mínimos ajustes. Es muy sencillo crear gráficos de una o varias variables con la ayuda del paquete ggplot2.

Los tres componentes básicos para construir un ggplot son los siguientes:

Data Conjunto de datos a graficar

- Aesthetics: Mapeo de los datos a la visualización
- Geometry/Layers: Elementos visuales utilizados para los datos

La sintaxis básica de ggplot es la siguiente:

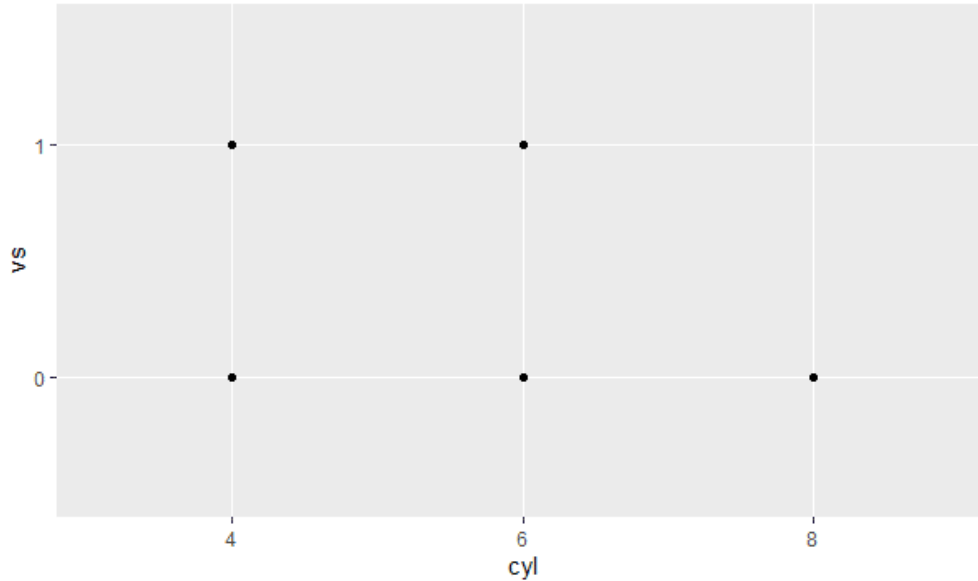
```
ggplot(data = NULL, mapping = aes()) + geom_function()

#To Install and load the ggplot2 package
install.packages("ggplot2")
library(ggplot2)
```

Scatter Plots

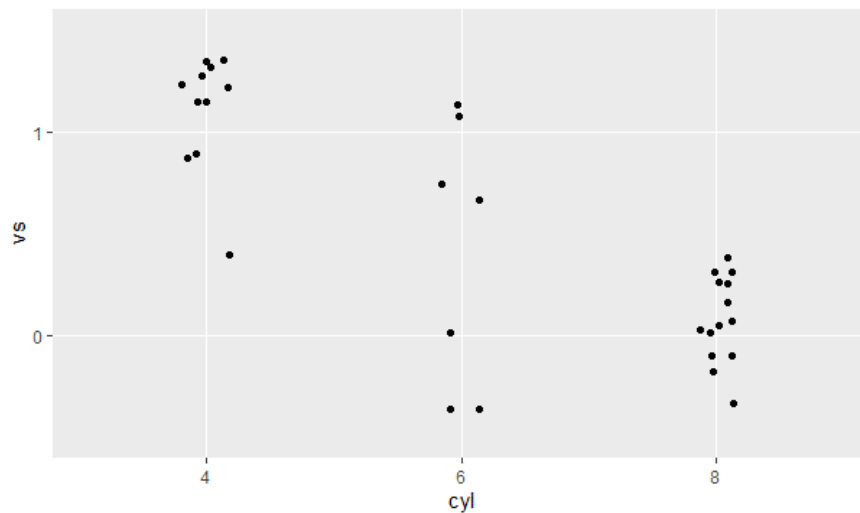
Para dibujar un gráfico de dispersión de cyl(Número de cilindros) y vs(Tipo de motor(0 = en forma de V, 1 = recto)), ejecute el siguiente código:

```
#Since the following columns have discrete(categorical) set of values, So we can
convert them to factors for optimal plotting
mtcars$am <- as.factor(mtcars$am)
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$vs <- as.factor(mtcars$vs)
mtcars$gear <- as.factor(mtcars$gear)
#To draw scatter plot
ggplot(mtcars, aes(x= cyl , y= vs)) + geom_point()
```



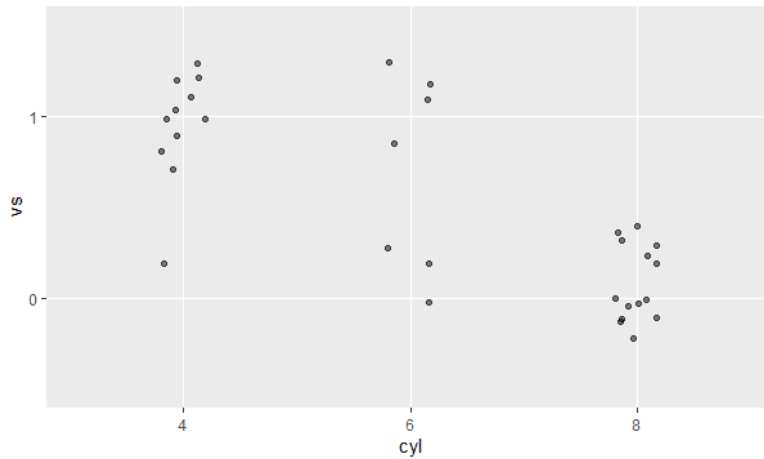
Como este gráfico tiene muchos valores superpuestos, lo que se conoce como overplotting, utilizaremos la función `geom_jitter()` para añadir una cierta cantidad de ruido para evitarlo.

```
#Here width argument is used to set the amount of jitter
ggplot(mtcars, aes(x= cyl , y= vs)) + geom_jitter(width = 0.1)
```



Aquí, también podemos usar el argumento `alpha` para establecer la transparencia de los puntos para reducir aún más el sobretrazado para la visualización de datos en R.

```
#Transparency set to 50%
ggplot(mtcars, aes(x= cyl , y= vs)) + geom_jitter(width = 0.1, alpha = 0.5)
```

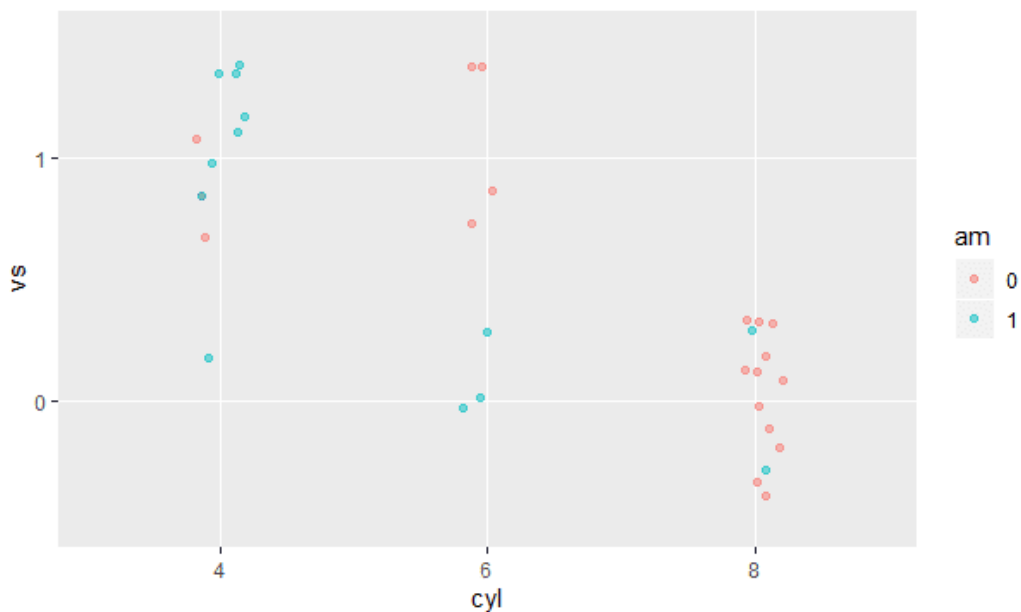


Con ggplot2, podemos trazar gráficos multivariantes de forma eficaz.

Por ejemplo:

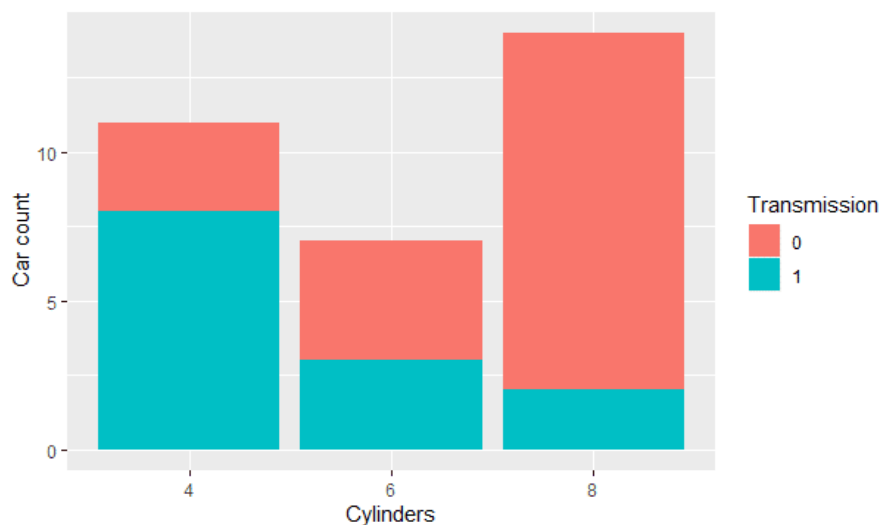
Para dibujar un gráfico de dispersión de cyl(Número de cilindros) y vs(Tipo de motor(0 = en forma de V, 1 = recto)) según am Transmisión (0 = automática, 1 = manual), ejecute el siguiente código:-

```
#We use the color aesthetic to introduce third variable with a legend on the right side
ggplot(mtcars, aes(x= cyl,y= vs,color = am)) + geom_jitter(width = 0.1, alpha = 0.5)
```

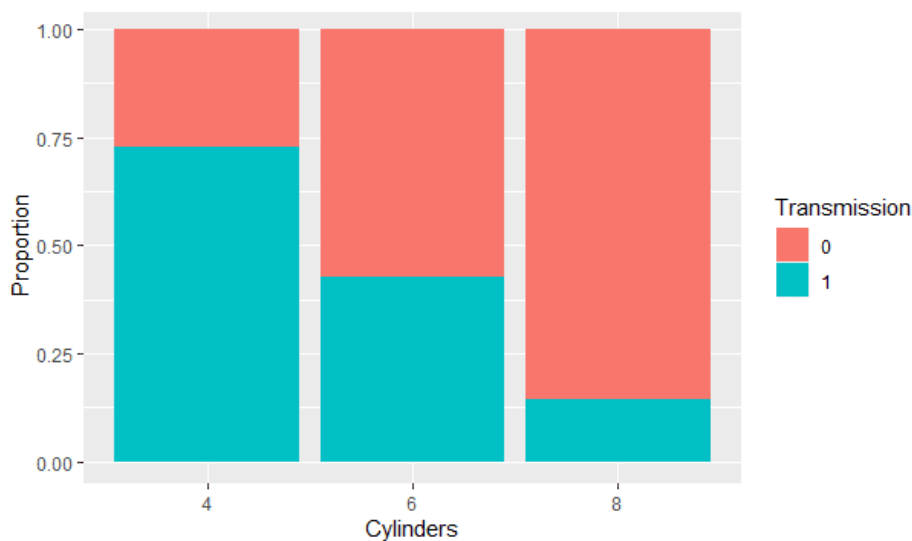


Bar Plots

```
#To draw a bar plot of cyl(Number of Cylinders) according to the Transmission type
using geom_bar() and fill()
ggplot(mtcars, aes(x = cyl, fill = am)) +
  geom_bar() +
  labs(x = "Cylinders", y = "Car count", fill = "Transmission")
```

```
#To find the proportion, we use position argument,as follows:
ggplot(mtcars, aes(x = cyl, fill = am)) +
  geom_bar(position = "fill") +
  labs(x = "Cylinders",y = "Proportion",fill = "Transmission")
```



Themes

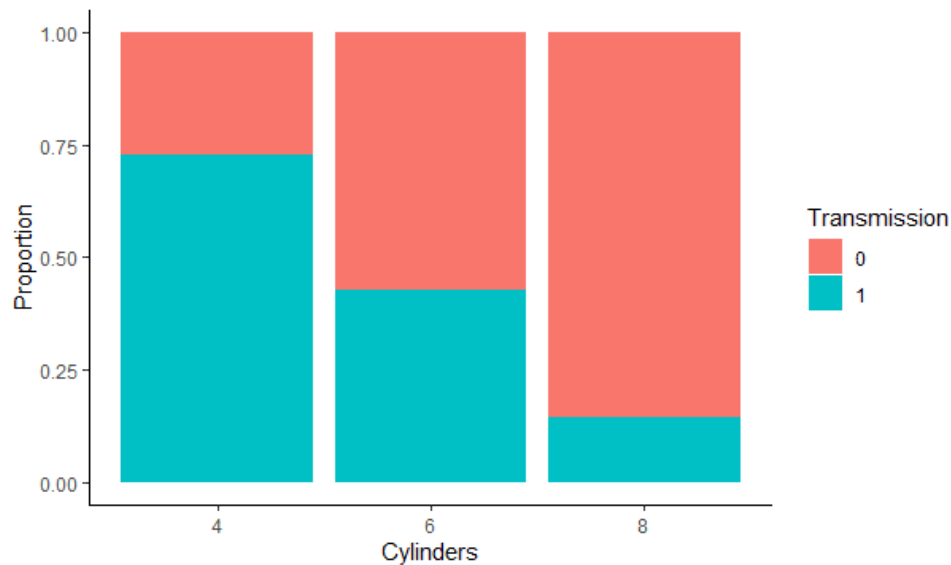
Se utiliza para cambiar los atributos de los elementos que no son datos de nuestro gráfico como el texto, las líneas, el fondo, etc. Utilizamos la función `theme_function()` para realizar cambios en estos elementos para la visualización de datos en R.

Algunas de las funciones temáticas más utilizadas son las siguientes:

- `theme_bw()` : Para el fondo blanco y las líneas grises de la cuadrícula
- `theme_gray`: Para fondo gris y líneas de cuadrícula blancas
- `theme_linedraw`: Para líneas negras alrededor del gráfico

- `theme_light`: Para líneas y ejes de color gris claro
- `theme_void`: Un tema vacío, útil para trazados con coordenadas no estándar o para dibujos
- `theme_dark()`: Un fondo oscuro diseñado para hacer resaltar los colores

```
ggplot(mtcars, aes(x = cyl, fill = am)) +
  geom_bar(position = "fill") +
  theme_classic()+
  labs(x = "Cylinders", y = "Proportion", fill = "Transmission")
```

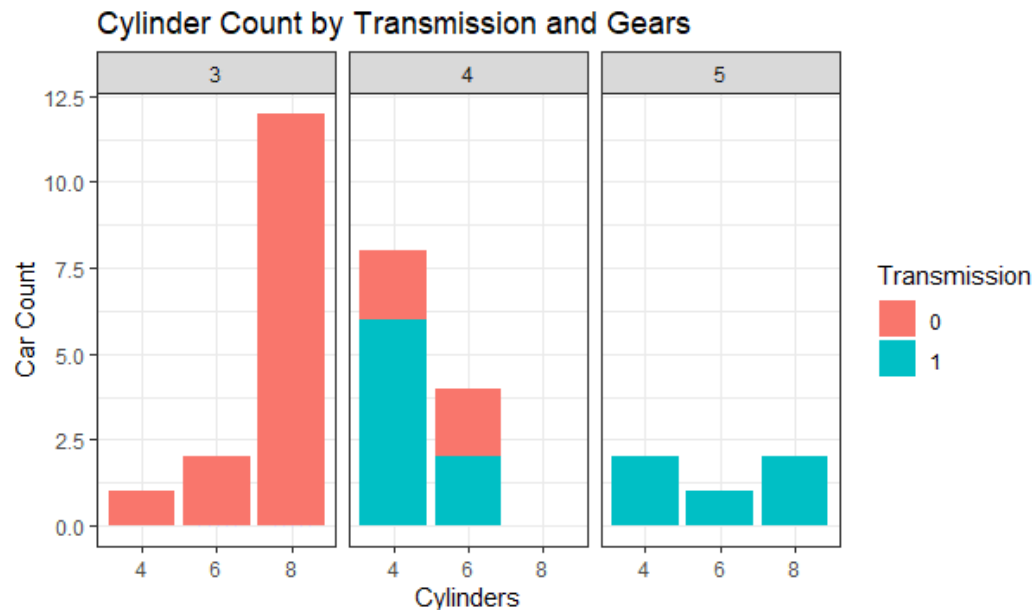


Faceting

Se utiliza para profundizar en los datos y dividirlos por una o más variables, y luego trazar los subconjuntos de los datos en conjunto para una óptima visualización de los datos en R.

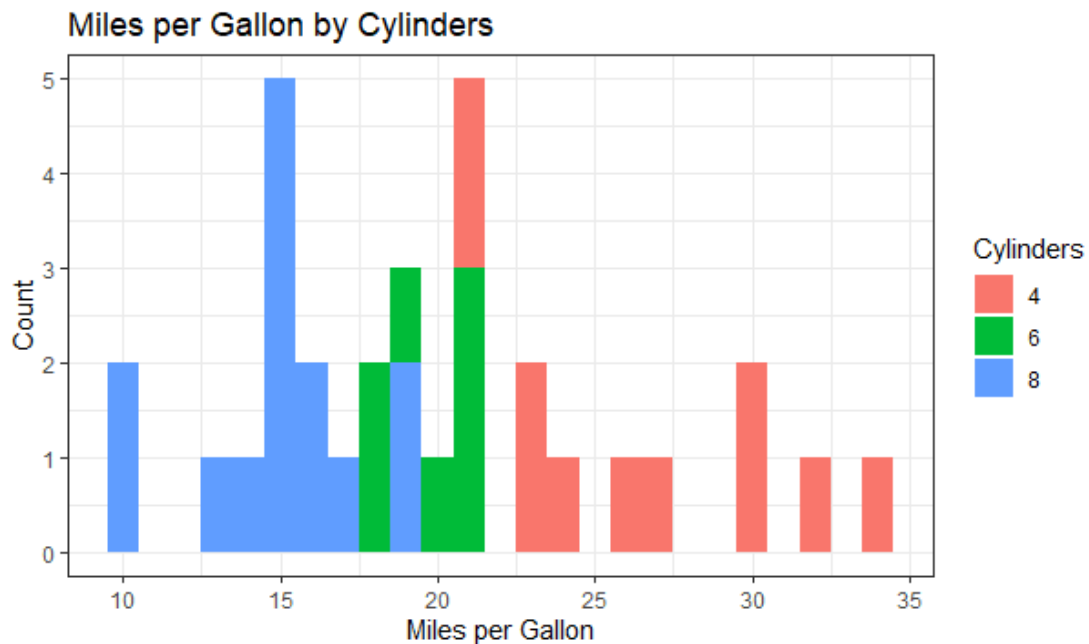
Por ejemplo

```
#To facet the following plot according to gear(Number of Gears(3,4,5)), we use
facet_grid() function as follows:
ggplot(mtcars, aes(x = cyl, fill = am)) +
  geom_bar() +
  facet_grid(.~gear)+
  #facet_grid(rows ~ columns) theme_bw() + labs(title = "Cylinder count by transmission
and Gears", x = "Cylinders", y = "Count", fill = "Transmission")
```



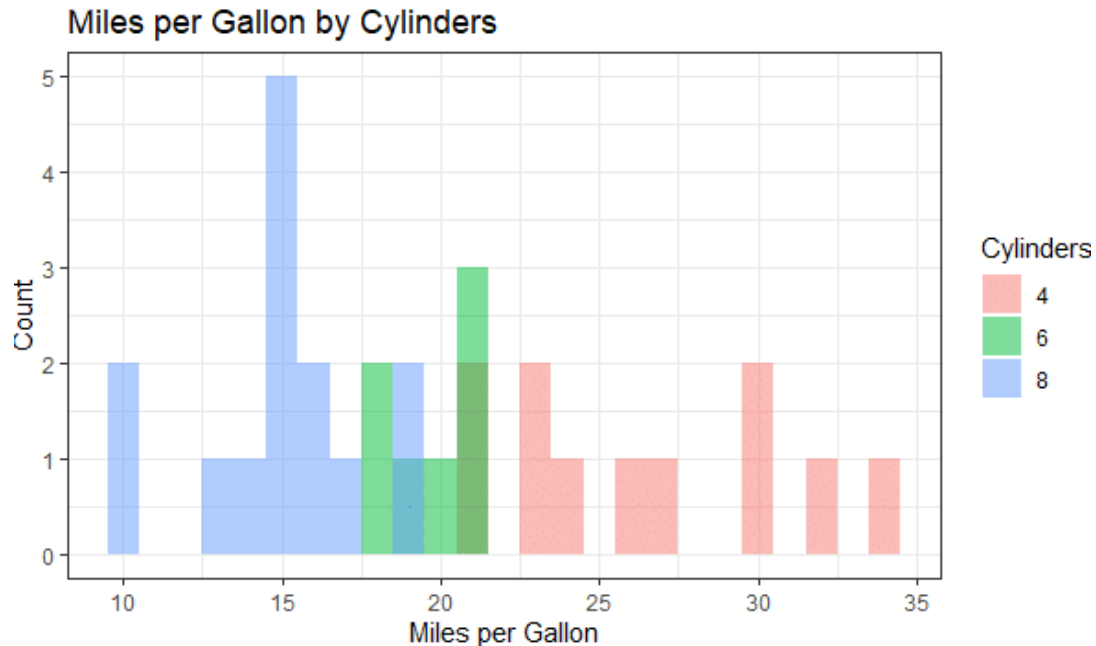
Histograms

```
#To plot a histogram for mpg (Miles per Gallon),
according to cyl(Number of Cylinders), we use the geom_histogram() function
ggplot(mtcars, aes(mpg, fill = cyl)) +
  geom_histogram(binwidth = 1) +
  theme_bw() +
  labs(title = "Miles per Gallon by Cylinders", x = "Miles per Gallon", y =
"Count", fill = "Cylinders")
```



```
#To show overlapping, we set position to identity and alpha to 0.5
ggplot(mtcars, aes(mpg, fill = cyl)) +
```

```
geom_histogram(binwidth = 1, position = "identity", alpha = 0.5)+
theme_bw()+
labs(title = "Miles per Gallon by Cylinders", x = "Miles per Gallon", y = "Count", fill =
"Cylinders")
```



<https://intellipaat.com/blog/tutorial/r-programming/data-visualization-in-r/>

Estadística indiferencia

La estadística inferencial se utiliza para sacar conclusiones de los datos

El objetivo

Formular hipótesis y ponerlas a prueba para poder hacer generalizaciones sobre poblaciones a partir de muestras.

El procedimiento

Utilizar las prácticas de muestreo aleatorio y los procedimientos de comprobación de hipótesis para juzgar la validez de las hipótesis previamente establecidas hipótesis previamente establecidas

La estadística inferencial suele invocar medidas de significación estadística

La información se entrega a la estadística inferencial en una multitud de formas diferentes (por ejemplo, vectores, matrices, marcos de datos). Esta información se utiliza para:

Hipótesis establecidas

Hipótesis probadas

- Nula/ Alternativas
- (No) Direccional
- No específica
- Diferencial
- Equivalente
- Relación

Test no parametrizado

- Nominal y de correlación
- Ordinal y métrica

Test parametrizado

- t-test
- ANOVA

La estadística inferencial permite generalizar más allá de los datos disponibles.

Las hipótesis y su importancia

¿Qué es una hipótesis?

En el caso de la estadística inferencial, una hipótesis presenta algún razonamiento sobre los patrones del mundo natural y, por tanto, de los datos.

¿Cuál es el problema?

Las hipótesis son simplificaciones de las posibles normas de los procesos naturales y hacen que las cosas sean comprobables.

¿Y?

Obtener las respuestas correctas siempre se reduce a formular las preguntas correctas.

Las hipótesis son, más o menos, conjeturas.

Hipótesis nula y alternativa (teoría)

Este es el formato más básico de las hipótesis sobre el que se construye cualquier otro tipo de hipótesis se construye.

Hipótesis nula:

- Representa una hipótesis de base ($X = Y$)
- Puede aceptarse o rechazarse

Hipótesis alternativa:

- Representa la negación de la hipótesis nula ($X \neq Y$) Se aceptará o rechazará en función de si la hipótesis nula es resulta ser correcta o no.

Ejemplo

- Nuestra expectativa nula, si el nicho climático se expande al azar o por igual en todas las periferias del nicho, [...]. Esto daría como resultado un valor de EI (Índice de Expansión) de 0. (Ralston, J. et al. (2016) 'Population trends influence species ability to track climate change', *Global Change Biology*, pp. 1-10. doi: 10.1111/gcb.13478).

- [...] detectar la cubierta vegetal [...]. Se aplicó la prueba del chi al cuadrado para comprobar la hipótesis nula de ausencia de efectos. [...] el modelo de regresión logística funciona mejor que el modelo nulo [...]. (Nioti, F. et al. (2015) 'A Remote Sensing and GIS Approach to Study the Long-Term Vegetation Recovery of a Fire Affected Pine Forest in Southern Greece', *Remote Sensing*, 7(6), pp. 7712-7731. doi: 10.3390/rs70607712).

Hipótesis de diferencia

Este formato de hipótesis se basa en las diferencias postuladas en los parámetros de cada variable dentro de las muestras.

Ejemplos:

- [...] diferencia en la tasa de eventos de eventos hemorrágicos [...] entre [...] profilaxis (grupo A) y [...] sin profilaxis (grupo B) [...]. (Oldenberg, J. et al. (2017) 'Emicizumab prophylaxis in hemophilia A with inhibitors', *N.Engl.J Med.*, pp. 1-10. doi: 10.1056/NEJMoa1703068)
- [...] podría permitir a la planta reaccionar de manera diferente a la próxima helada (Walter, J. et al. (2013) 'Ecological stress memory and cross stress tolerance in plants in the face of climate extremes', *Environmental and Experimental Botany*. Elsevier B.V., 94, pp. 3-8. doi:10.1016/j.envexpbot.2012.02.009).

Hipótesis de equivalencia

Este formato de hipótesis se basa en la equivalencia postulada de los parámetros de las variables dentro de las muestras

Ejemplos:

- Los umbrales equivalen a puntos de inflexión puntos de inflexión [...] (Angeler, D. G. and Allen, C. R. (2016) 'Quantifying Resilience', *Applied Ecology*, pp. 617-624. doi: 10.1111/1365-2664.12649).
- Así como el LAI es el equivalente del dosel del área foliar, ϵ_g^* es el equivalente de la cubierta del rendimiento cuántico (Prince, S. D. and Goward, S. N. (1995) 'Global primary production: a remote sensing approach', *Journal of Biogeography*, pp. 815-835. doi: 10.2307/2845983)

Hipótesis de relación

Este formato de hipótesis se basa en las relaciones postuladas de las variables dentro de una población.

Ejemplos:

- [...] producen relaciones significativas entre la GPP y la diversidad de árboles (Nightingale, J. M. et al. (2008) 'PREDICTING TREE DIVERSITY ACROSS THE UNITED STATES AS A FUNCTION OF MODELED GROSS PRIMARY PRODUCTION', *Ecological Applications*, 18(1), p. 93. Available at: <http://dx.doi.org/10.1890/07-0693.1>)
- [...] probar una serie de relaciones hipotéticas relaciones (es decir, lineales hasta umbral) entre la variable de respuesta variable de respuesta y el medio ambiente [...]

(Seddon, A. et al. (2014) 'A quantitative framework for analysis of regime shifts in a Galapagos coastal lagoon', *Ecology*, 95(11), pp. 3046-3055. doi: 10.1890/13-1974.1)

Hipótesis direccionales y no direccionales

Este formato de hipótesis se construye sobre las conexiones y/o diferencias postuladas de las variables dentro de las muestras.

Hipótesis direccional:

Afirmación sobre la dirección en la que se postula que funcionan estas conexiones o diferencias a lo largo.

$X > Y$; $X \leq Y$; $X < Y$; $X \geq Y$

Hipótesis no direccional:

Ninguna afirmación sobre la dirección estas conexiones o diferencias se postula que funcionan a lo largo de

$X \neq Y$ con $X ? Y$

Ejemplos:

- [...] una variación de diez veces en las tasas de mineralización de las dunas de arena a las praderas fertilizadas (Ellenberg 1977) se asoció con un aumento de 12 veces en la ANPP (Poorter & de Jong 1999) [...] (Lavorel, S. and Garnier, E. (2002) 'Predicting changes in community composition and ecosystem functioning from plant traits: revisiting the Holy Grail', *Functional Ecology*, 16(Essay Review), pp. 545-556. doi: Doi 10.1046/J.1365-2435.2002.00664.X)
- Los árboles tropicales individuales muestran increíblemente fuertes y persistentes variación en las tasas de crecimiento a largo plazo, lo que resulta en una variación de cuatro veces en las edad de los árboles de tamaño similar (Brienen, R. J. W., Sch, J. and Zuidema, P. A. (2016) 'Tree Rings in the Tropics: Insights into the Ecology and Climate Sensitivity of Tropical Trees', in *Tropical Tree Physiology*. doi: 10.1007/978-3-319-27422-5)

Hipótesis especificadas y no especificadas

Este formato se basa en los tamaños de los efectos postulados de los tratamientos/grupos en experimental/observacional.

Hipótesis especificada:

Afirmación sobre un tamaño/intensidad del efecto esperado dentro de un conjunto de variables de respuesta basadas en un conjunto de variables predictoras.

$X = \beta * Y$, siendo β un coeficiente predefinido

Hipótesis no especificada:

Afirmación sobre un efecto esperado efecto esperado dentro de un conjunto de respuesta basado en un conjunto de variables predictoras sin una noción de un tamaño/intensidad del efecto.

$X = \beta * Y$ siendo β algún coeficiente indefinido

Ejemplos:

- El tamaño del efecto de la diversidad (log natural de respuesta; LRR) se basa en la comparación de niveles de riqueza de especies altos y bajos riqueza de especies [...] (De Boeck, H. J. et al. (2017) 'Patterns and drivers of biodiversity-stability relationships under climate extremes', *Journal of Ecology*, (October), pp. 1-13. doi: 10.1111/1365-2745.12897.)
- [...] una muestra de 51 participantes con una tasa de retirada del 10% en el grupo de control proporcionaría una potencia de más del 95% a un nivel de significación de dos lados de 0,05 para detectar un tamaño del efecto de $4/18 = 0,22$ (hipótesis nula: relación de tasas = 1). (Oldenberg, J. et al. (2017) 'Emicizumab prophylaxis in hemophilia A with inhibitors', *N.Engl.J Med.*, pp. 1-10. doi: 10.1056/NEJMoa1703068).

Referencias

- [1] Naveen (2022, 15 Junio) Data Visualization in R [Online]. Available: <https://intellipaat.com/blog/tutorial/r-programming/data-visualization-in-r/>
- [2] Erik Kusch AARHUS UNIVERSITY, Center for biodiversity and dynamics in a changing world, "Descriptive Statistics", 2019.