Análisis exploratorio de datos

Hugo Andrés Dorado

Científico de datos

hugo.doradob@gmail.com

Nociones en el análisis exploratorio de datos

Todo estudio basado en **datos**, sin importar su alcance, debe superar la fase inicial del análisis exploratorio.

"Tabular, graficar, resumir, para identificar patrones y comportamientos regulares y presencia de irregularidades en los datos"

Existen patrones de comportamiento regular en los datos?

- Se presentan datos atipicos?

Que hacer con ellos?

- Como se relacionan las variables de analisis?
- Existen diferencias en el comportamiento de la variable entre grupos de analisis?

Es un paso necesario, que consume tiempo, y que en ocasiones es descuidado por los analistas

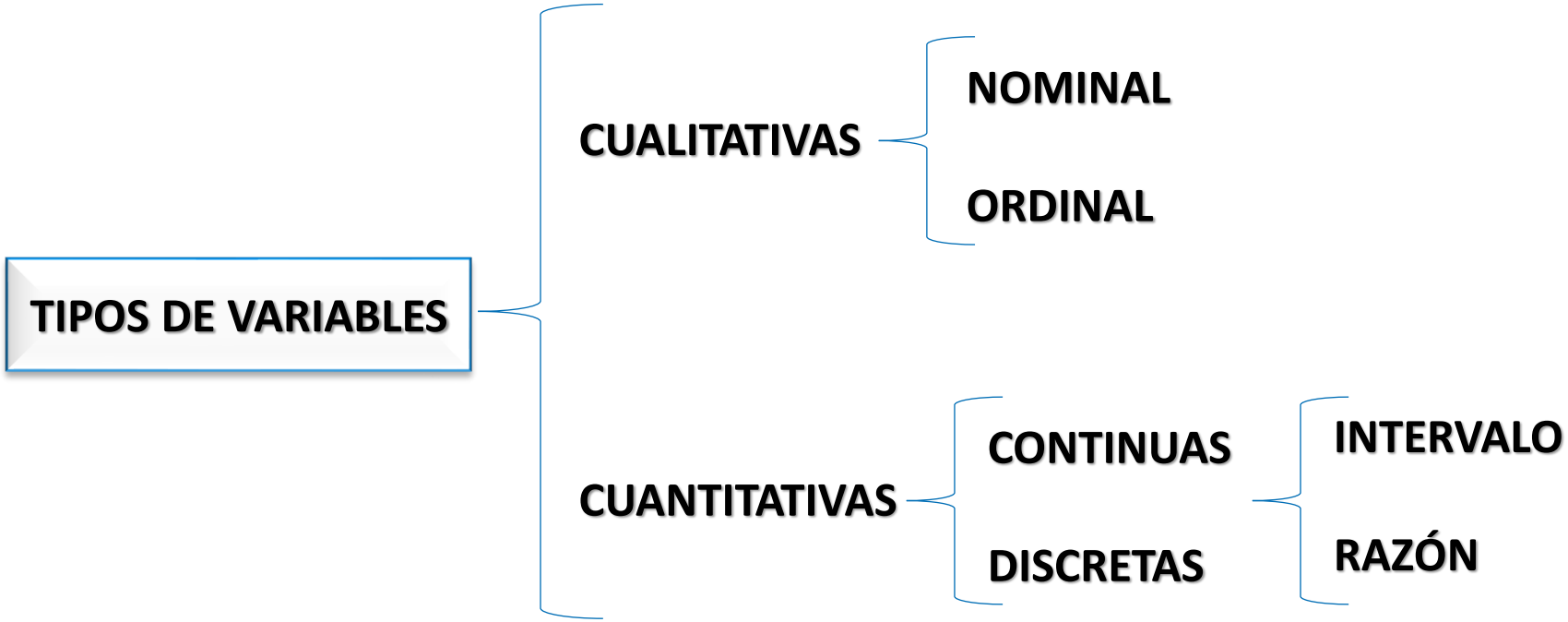
Análisis exploratorio de datos

Proporciona un conjunto de herramientas que intentan descubrir patrones de comportamiento en los datos en un ambiente de variabilidad e incertidumbre.

No siempre se requiere aplicar todas las herramientas exploratorias, cada una se aplica de acuerdo a la necesidad y al propósito de la investigacion.

Hipótesis ----------> Herramientas
(Objetivo) (Plan de Exploración)

VARIABLES


TIPOS DE VARIABLES
CUALITATIVAS
NOMINAL
ORDINAL
CUANTITATIVAS
CONTINUAS
INTERVALO
DISCRETAS
RAZÓN

VARIABLES

CUALITATIVAS

- Si sus valores (modalidades) no se pueden asociar naturalmente a un número.
- No se pueden hacer operaciones algebraicas con ellos.

Nominales:
Si sus valores no se pueden ordenar.
Sexo, Religión, Nacionalidad, Fumar (Si/No)

Ordinales:
Si sus valores se pueden ordenar.
Grado de satisfacción, Intensidad de dolor, Mejoría a un tratamiento.

CUANTITATIVAS

- Si sus valores son numéricos.
- Tiene sentido hacer operaciones algebraicas con ellos.

Discretas:
Si toma valores enteros.
Número de hijos, Número de carros.

Continuas:
Si entre dos valores, son posibles infinitos valores intermedios.
Altura, Temperatura, Duración de una batería, Peso(kg).

Variables

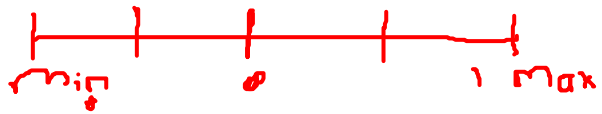
¿Qué tipo de variables tenemos?

¿Cuál es su escala?

Nombre	#Hijos	Genero	Raza	Salario	Cargo
Diego Giraldo	0	Masculino	Blanca	62.100	Directivo
Diana Sánchez	2	Femenino	Blanca	47.350	Técnico
Julián Castro	1	Masculino	Asiática	18.250	Administrativo
Simón Valdés	1	Masculino	Negra	76.600	Directivo

Algunas Medidas descriptivas

CUANTILES: Valores de la distribución que la dividen en partes iguales los mas usados son:



- **Percentiles:** son 99 valores que dividen en cien partes iguales el conjunto de datos ordenados.

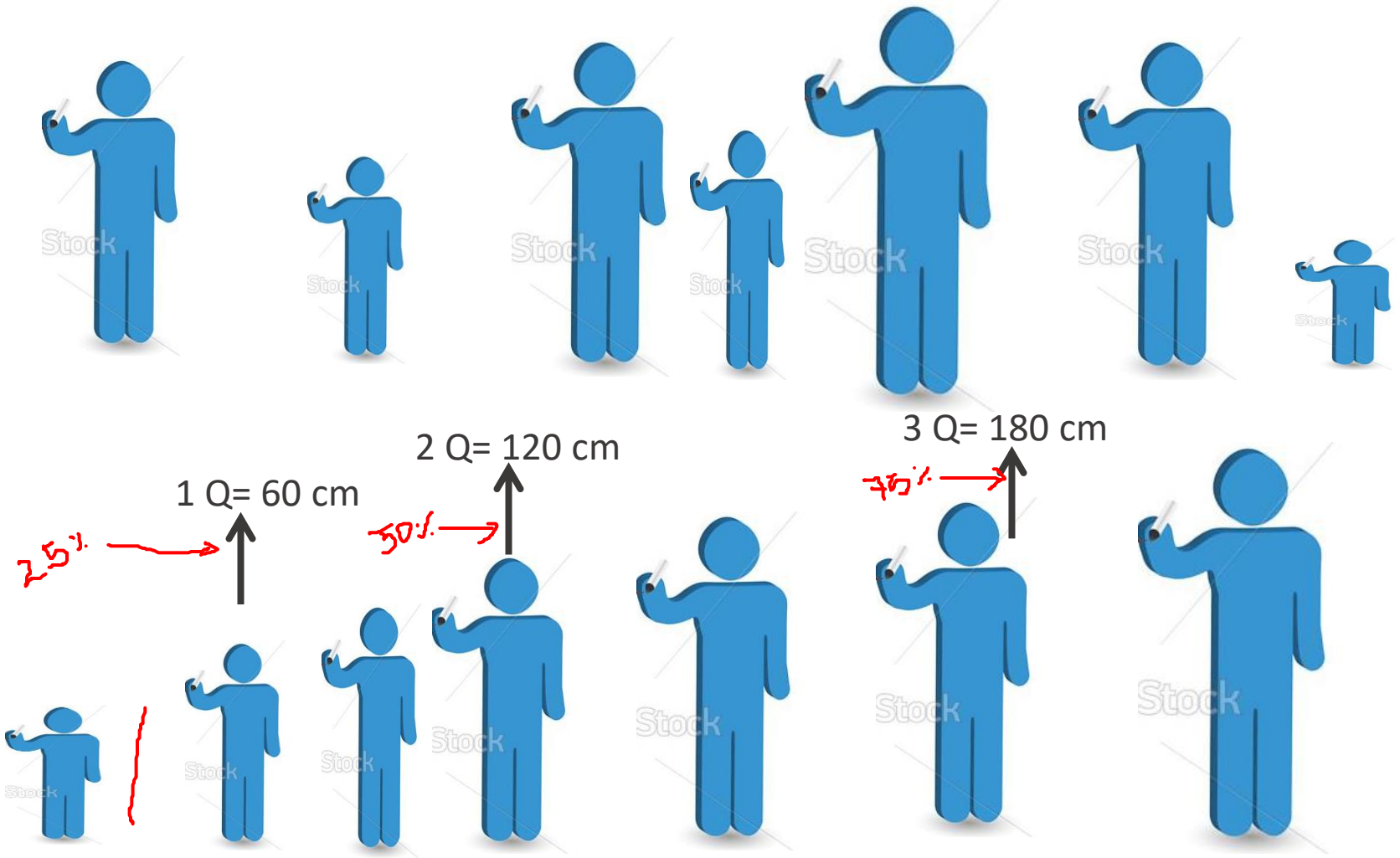
- **Cuartiles:** son los tres valores que dividen al conjunto de datos ordenados en cuatro partes iguales.

- **Deciles:** son los nueve valores que dividen al conjunto de datos ordenados en diez partes iguales.

- Terciles

Medidas descriptivas

Cuartil



Frecuencias – Variables cualitativas

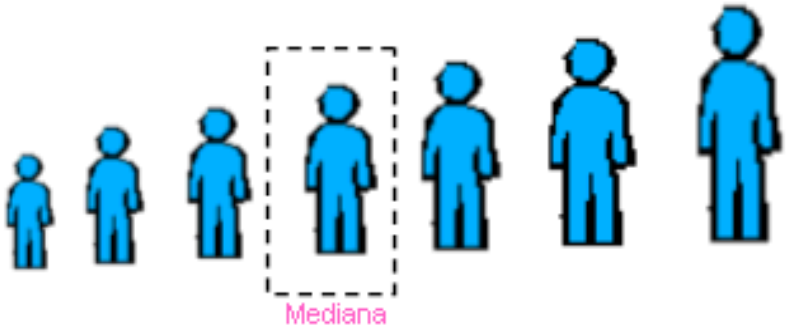
- **Frecuencia absoluta:** Número de veces que se presenta el valor de la variable.
- **Frecuencia relativa:** Cociente entre la frecuencia absoluta y el número total de casos.
- **Porcentaje:** Resultado de multiplicar por 100 la frecuencia relativa. Representado, indica el tanto por ciento de la población que corresponde a ese valor de la variable.

Medidas de Centralización

- **Media:** Es el valor que resulta de la suma de todos los valores dividido el total de datos.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- **Mediana:** es el valor que separa por la mitad las observaciones ordenadas de menor a mayor.

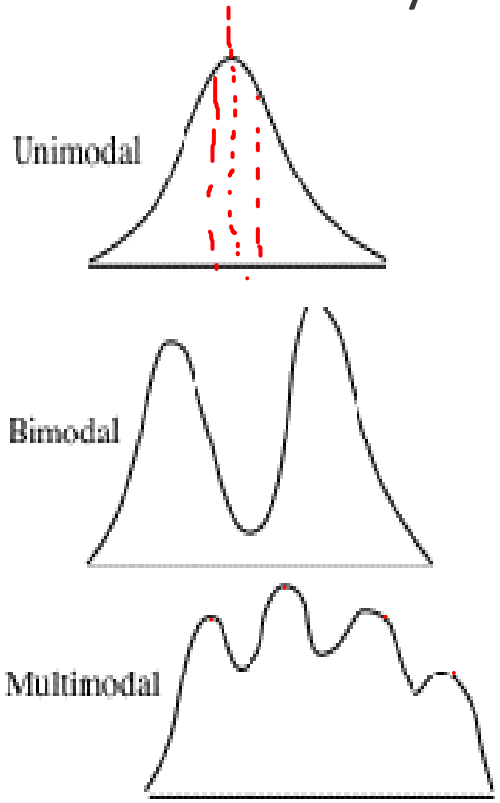


(18, 21, 24, 26, 53)

Me = 24

Mu= 30.8

- **Moda:** es el valor de la variable que más veces se repite, es decir, aquella cuya frecuencia absoluta es mayor.



Medidas de dispersión

- **Varianza:** es el promedio del cuadrado de las distancias entre cada observación y la media aritmética del conjunto de observaciones.

$$V = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

- **Desviación típica:** La varianza viene dada por las mismas unidades que la variable pero al cuadrado.

$$S = \sqrt{V}$$

- **Recorrido o rango muestral :** Es la diferencia entre el valor de las observaciones mayor y el menor.

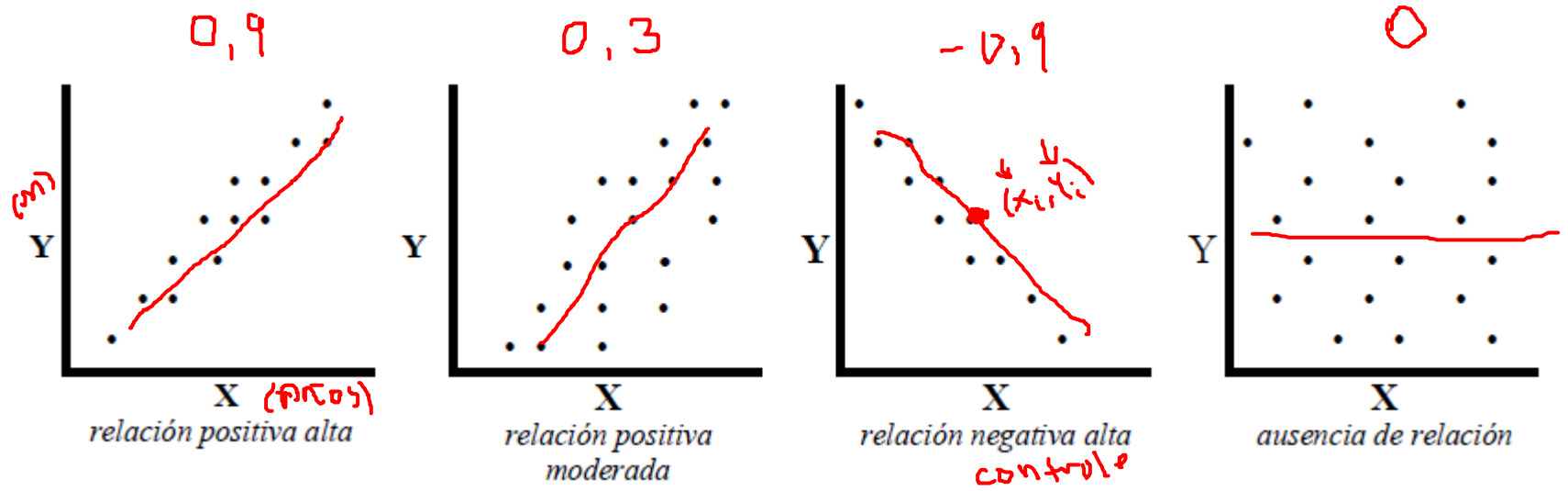
$$R_e = x_{max} - x_{min}$$

- **Coeficiente de variación de pearson:** Representa el número de veces que la desviación típica contiene a la media aritmética.

$$CV = \frac{S}{|\bar{x}|} \times 100$$

Medidas de covariación y correlación

Relación existente entre dos o mas variables cuantitativas.



Gráficos en R

Definición de un gráfico ‘bueno’ según Winer 1990 (Investigador estadístico)

“Un gráfico fuertemente bueno muestra todo lo que queremos conocer sólo con mirarlo.”

“Un gráfico débilmente bueno nos muestra lo que necesitamos conocer observándolo, una vez sepamos como mirarlo.”

Gráficos estadísticos con R

Gráficos base

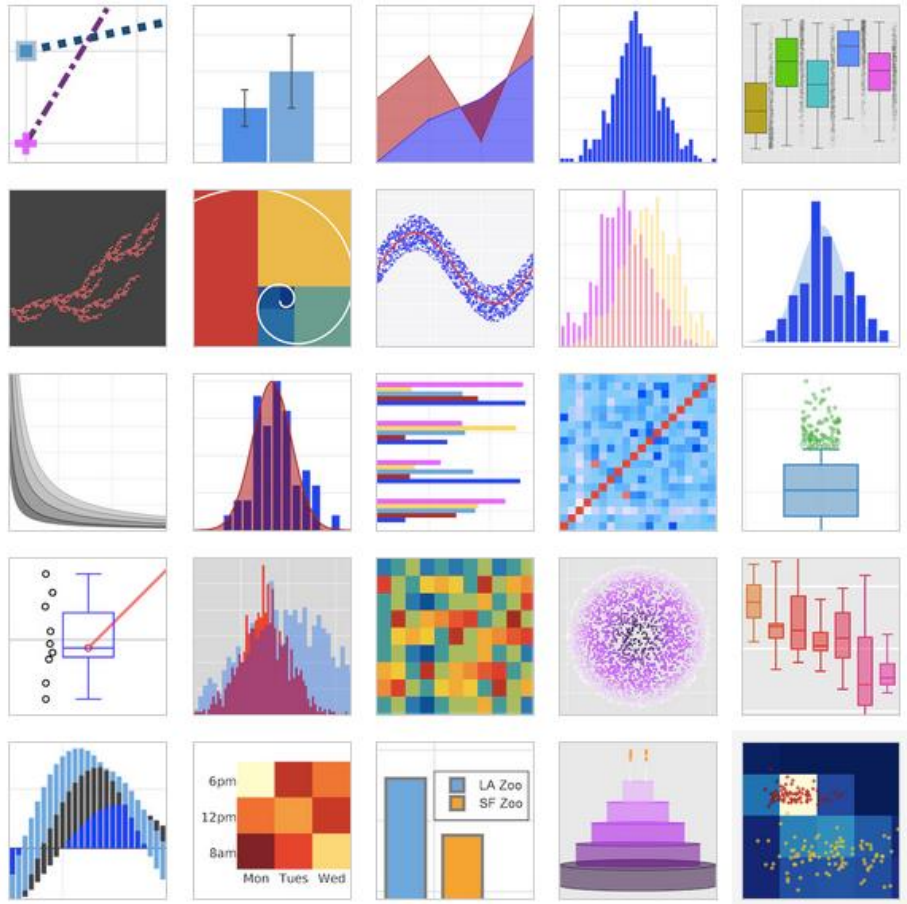
- Solo producen un resultado.
- No requiere paquetes adicionales.
- Varios parámetros para especificar.

ggplot

- Combinación de funciones que proporcionan los componentes del gráfico.
- Permiten crear varios datos simultáneos divididos por una o varias categorías.
- Menos parámetros.
- Ideal para presentaciones, informes, exploración.

lattice

- Utilizan una estructura matricial de paneles definida a partir de una forma
- Permiten crear varios datos simultáneos divididos por una o varias categorías.
- Ideal para publicaciones.



Img. datasciencecentral.com

Características de gráficos

William Playfair el pionero de la estadística gráfica, realizó su investigación basado en los siguientes principios.

1. El método gráfico es una forma de simplificar lo tedioso y lo complejo.
2. Las personas ocupadas necesitan alguna clase de ayuda visual.
3. Un gráfico es más accesible que una tabla
4. El método gráfico es concordante con los ojos.
5. El método gráfico ayuda al cerebro, ya que permite entender y memorizar mejor.

Fuente: https://cran.r-project.org/doc/contrib/grafi3.pdf

Principios de un gráfico

- Entendibilidad, ¿Nos permite ver la relaciones entre variables?
- Claridad, ¿Son los elementos del gráfico distinguibles?
- Consistencia, ¿Es el gráfico consecuente con gráficos anteriores?
- Eficiencia, ¿Están todos los elementos del grafico eficientemente representados?
- Necesidad, ¿Cada elemento es realmente necesario?
- Confiabilidad, ¿Están los datos realmente representados por la escala y sobre la región del gráfico?

Elementos mínimos de un gráfico

- Titulo Principal
- Región de Datos y Símbolos
- Eje Horizontal y Escala
- Eje Vertical y Escala
- Leyenda

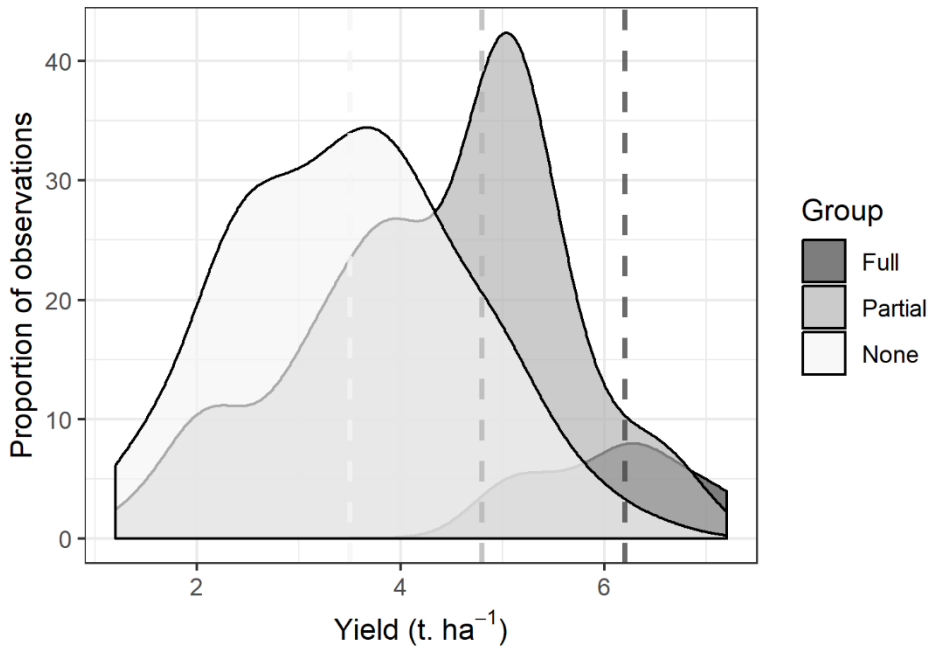


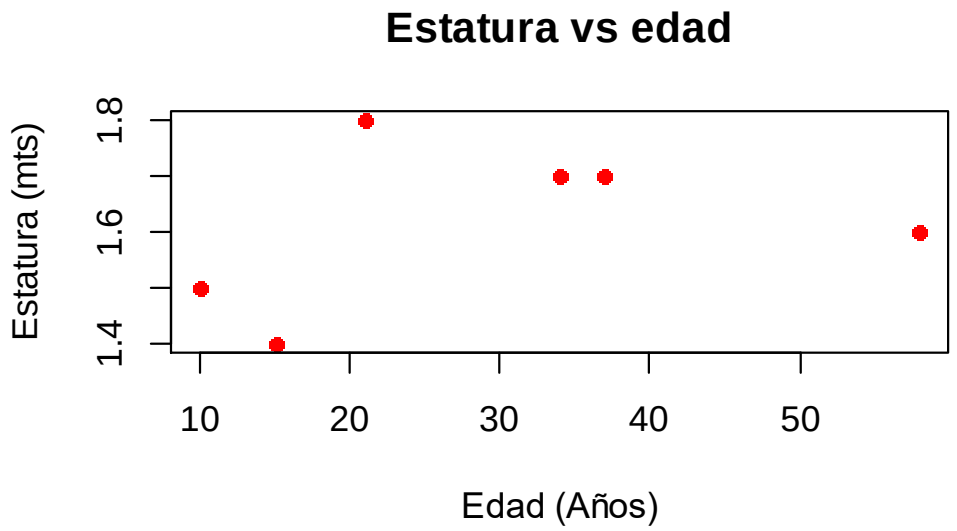
Fig.3. Observed yield distribution for the three different groups of farmers using data-driven recommended practices

Gráficos en R paquete base

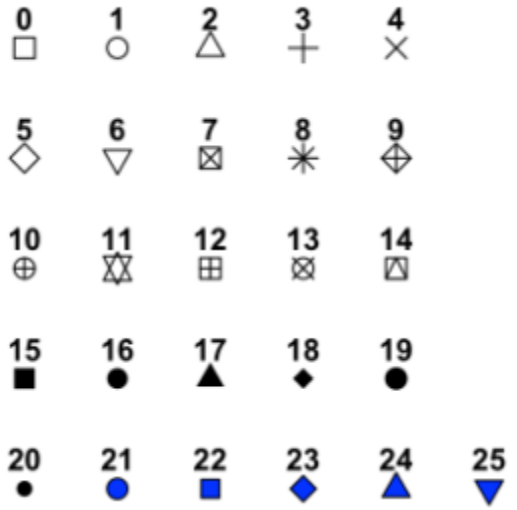
Grafico de dispersión

```
edad     <- c(21,34,10,15,37,58)
estatura <- c(1.8,1.7,1.5,1.4,1.7,1.6)

plot(edad,estatura,col="red",pch=16,main= 'Estatura vs edad',  xlab =
'Edad (Años)', ylab = 'Estatura (mts)')
```



Símbolos para plot (pch)



Fuente: http://www.sthda.com/english/wiki/r-plot-pch-symbols-the-different-point-shapes-available-in-r

2.1.2 Algunos Parámetros para Graficar en R

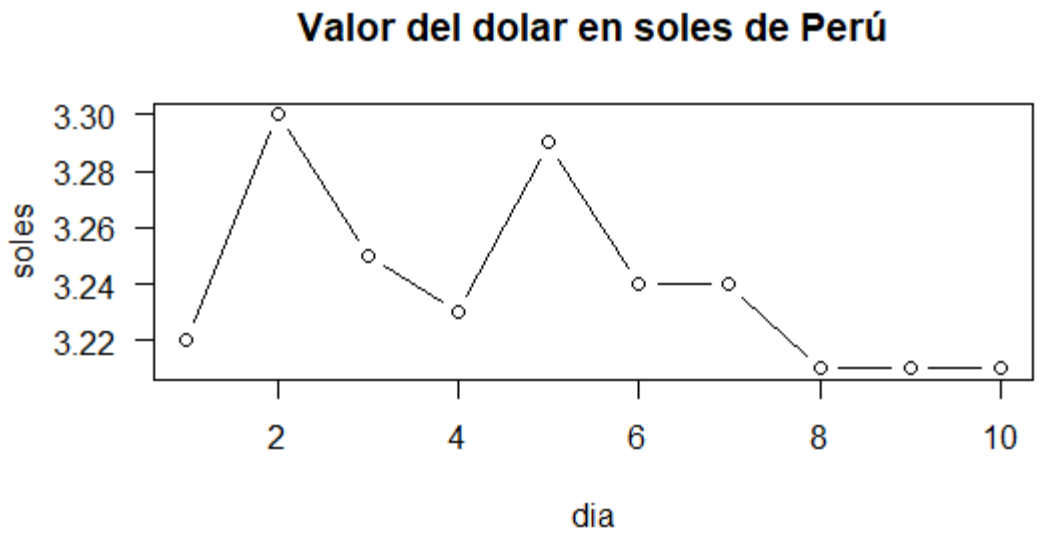
log='<x|y|xy>'	Ejes Logarítmicos
main='título'	
new=<logical>	Adiciona sobre el gráfico actual
sub='título de abajo'	
type='<l|p|b|n>'	Línea, puntos, ambos, ninguno
lty=n	Tipo de Línea
pch='.'	Caracter de dibujo
xlab='Nombre del eje x'	
ylab='Nombre del eje y'	
xlim=c(x_{minimo}, x_{maximo})	
ylim=c(y_{minimo}, y_{maximo})	

Gráficos de series temporales

```
dia <- c(1:10)

soles <- c(3.22,3.3,3.25,3.23,3.29,3.24,3.24,3.21,3.21,3.21)

plot(dia,lempira,type='b',las=1,main='Valor del dolar en soles de Perú')
```

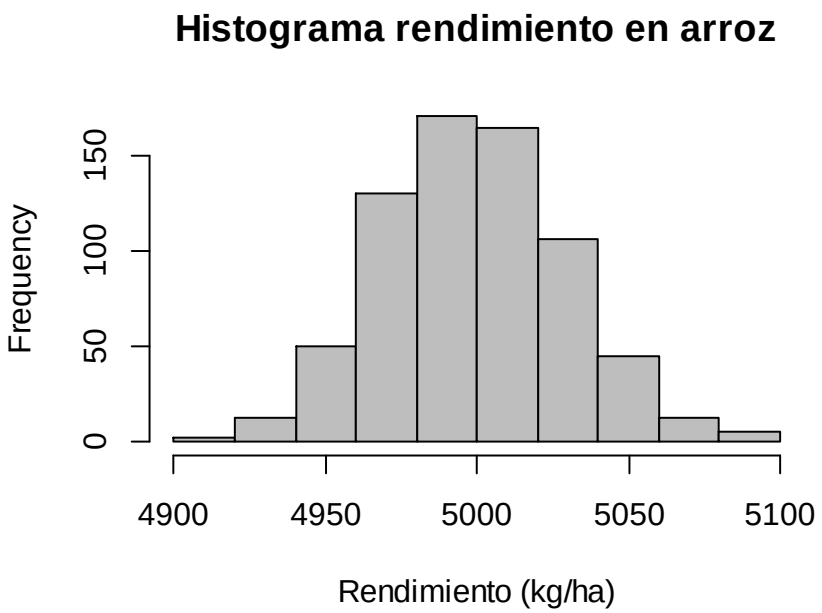


Histograma de frecuencia

```
x <- rnorm(700,mean = 5000, sd = 30 )

hist(x,xlab = 'Rendimiento (kg/ha)',main = 'Histograma
rendimiento' ,col='gray')
```

Variables cuantitativas




```
barplot(tabcultivos, ylim=c(0,25), main = 'Diagrama de barras de
cultivos',ylab='Frecuencia')

box()
```

Variables cualitativas

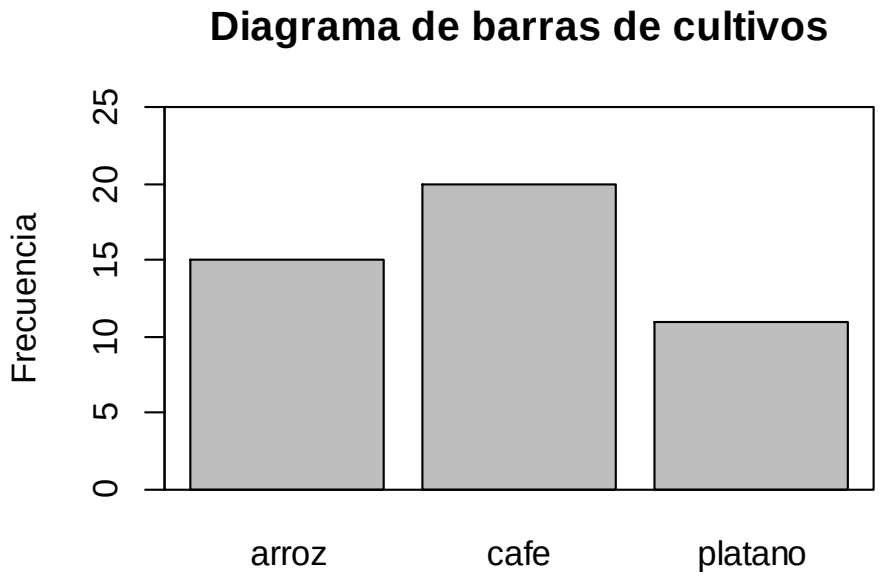
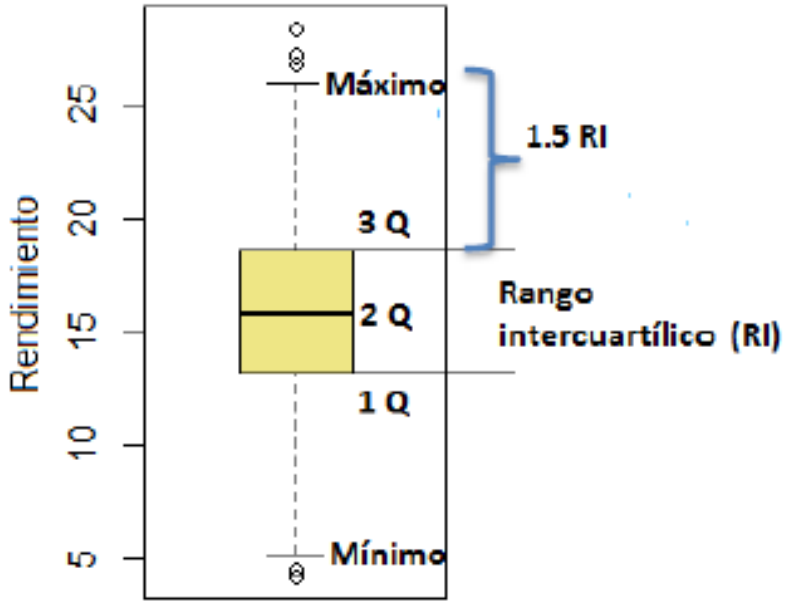
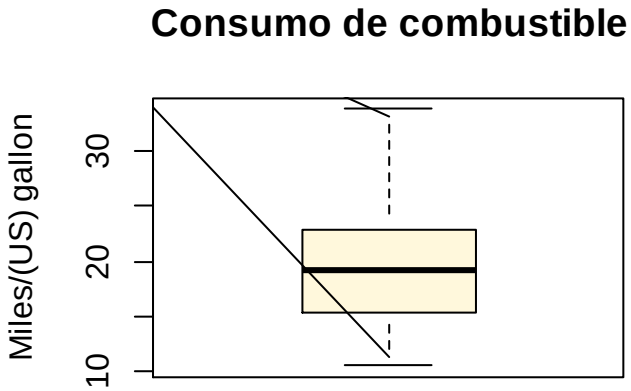


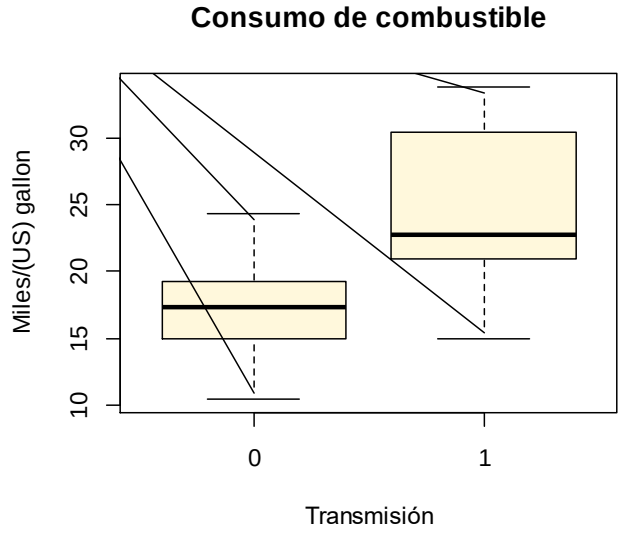
Gráfico de boxplot





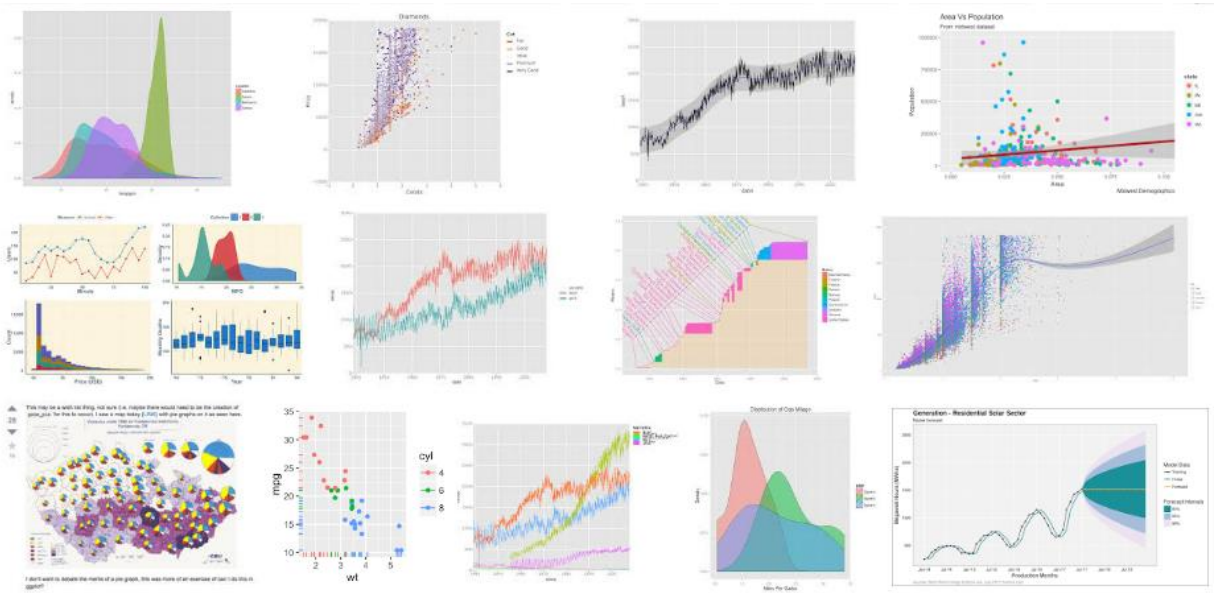

```
boxplot(mtcars$mpg,main='Consumo de combustible',
        ylab='Miles/(US) gallon',col='cornsilk')
```

```
boxplot(mpg~am,data=mtcars,main='Consumo de
combustible por tipo de trasmisión',
        ylab='Miles/(US)
gallon',col='cornsilk',xlab='Transmisión')
```



Gráficos con ggplot

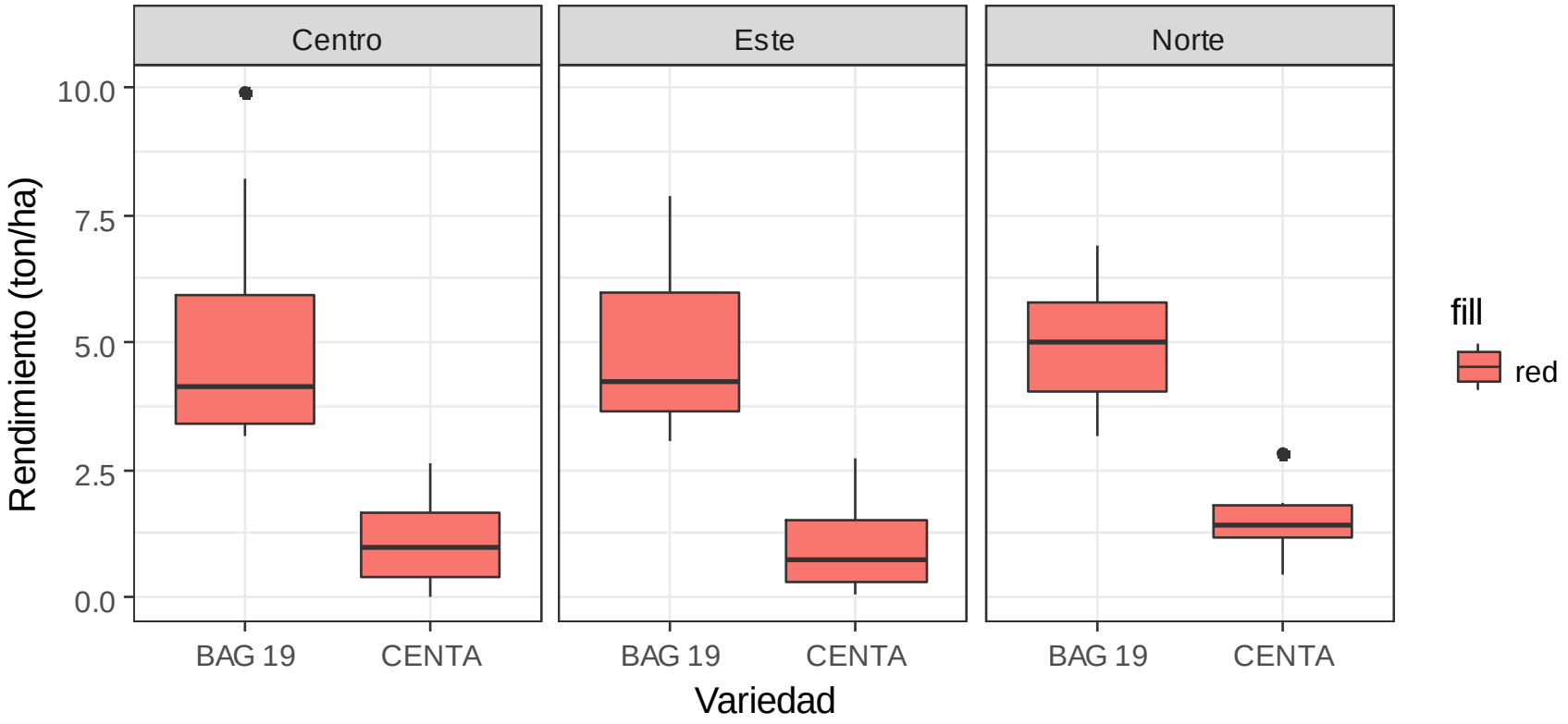
- El orden de los comando es inspirado en una estructura gramatical de gráficos (Wilkinson, 2005)
- Especificación de los gráficos con un alto nivel de abstracción.
- Muy flexible
- Distintos temas y apariencias.
- Muchos usuarios activos



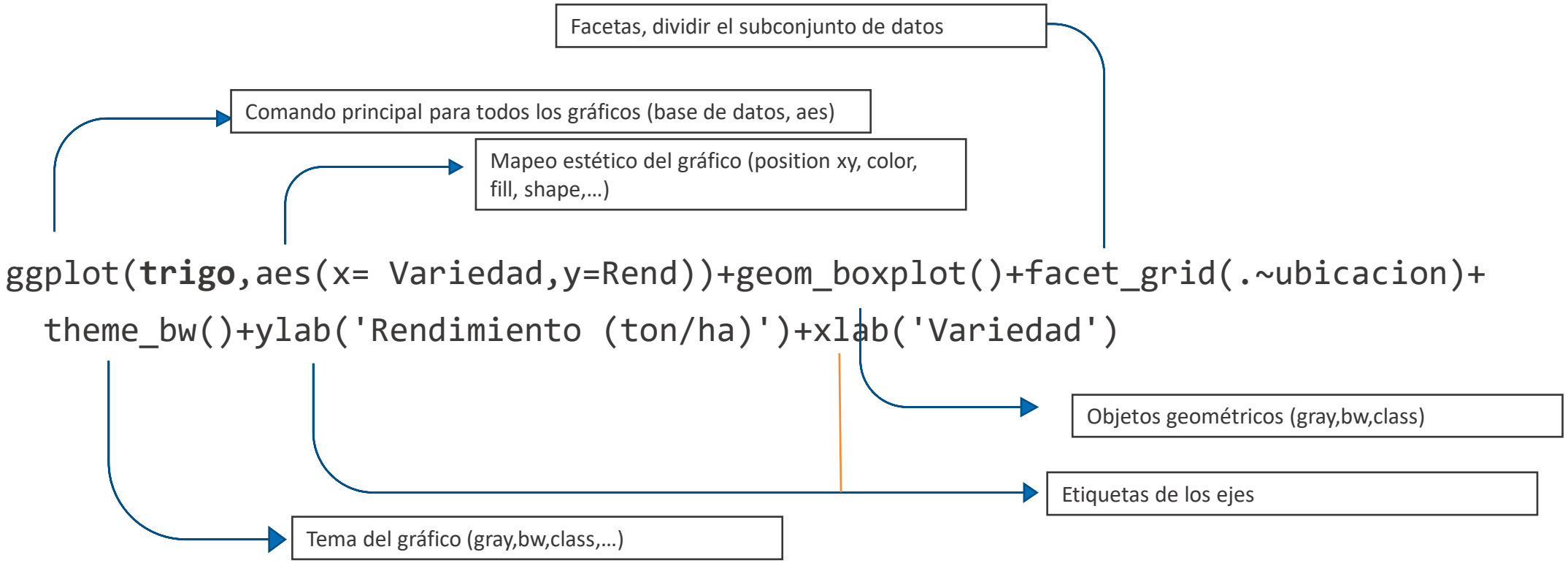
Fuente: https://tutorials.iq.harvard.edu/R/Rgraphics/Rgraphics.html

Gráficos con ggplot

```
ggplot(trigo,aes(x= Variedad,y=Rend,fill='red'))+geom_boxplot()+facet_grid(.~ubicacion)+
  theme_bw()+ylab('Rendimiento (ton/ha)')+xlab('Variedad')
```



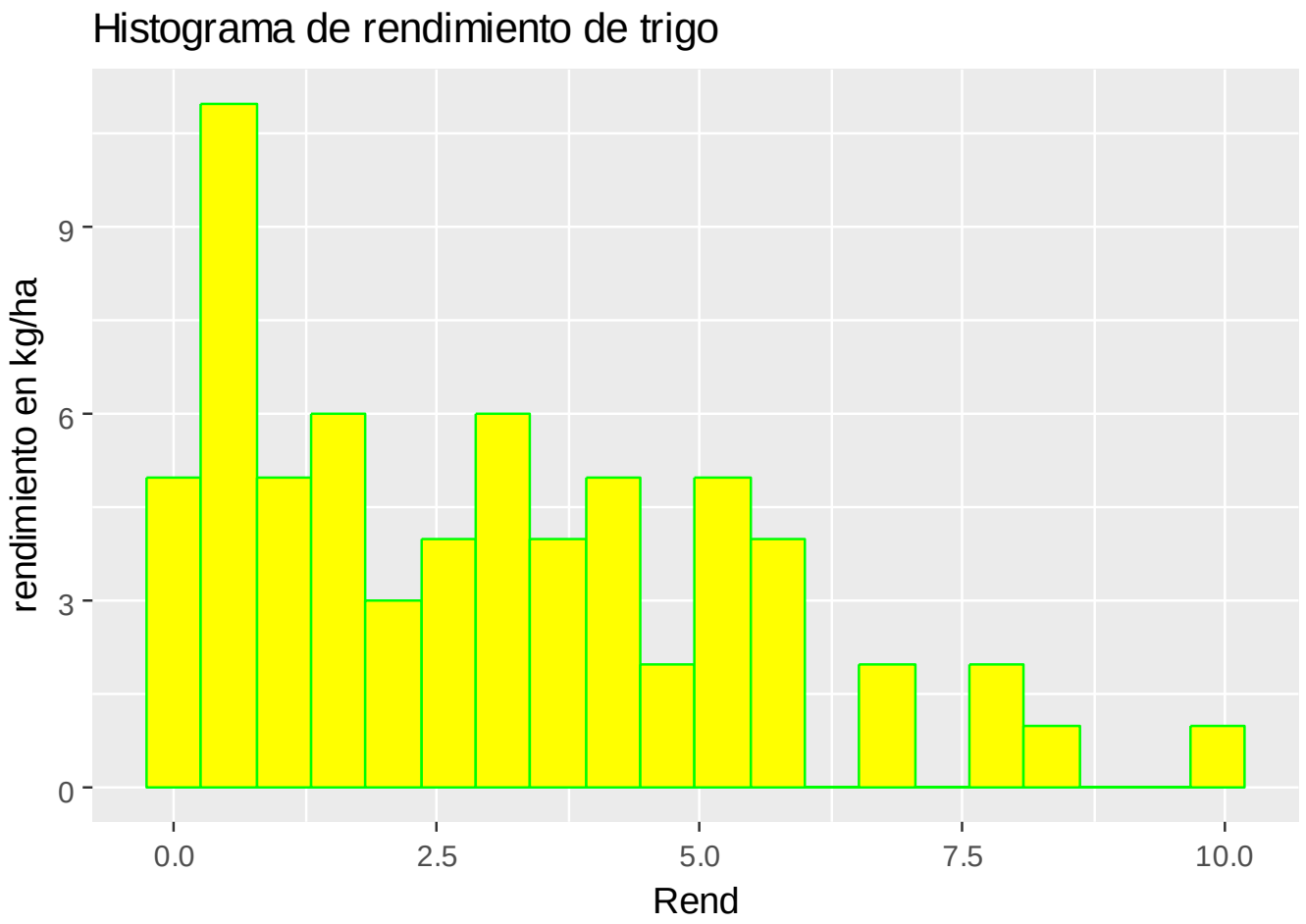
Gramática de graficos



https://tutorials.iq.harvard.edu/R/Rgraphics/Rgraphics.html#scales

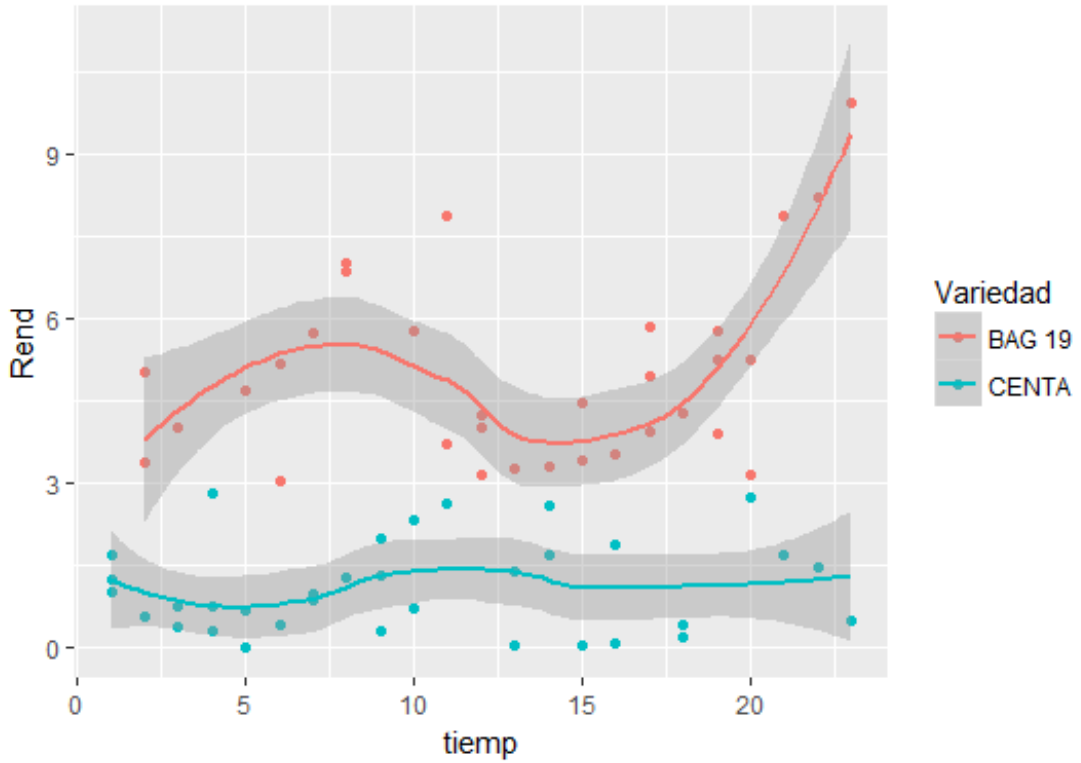
Comando básico de graficos

```
ggplot(trigo,aes(x=Rend))+geom_histogram(bins = 20,colour='green',fill='yellow')+ggtitle('Histograma de rendimiento de
trigo')+ylab('rendimiento en kg/ha')
```



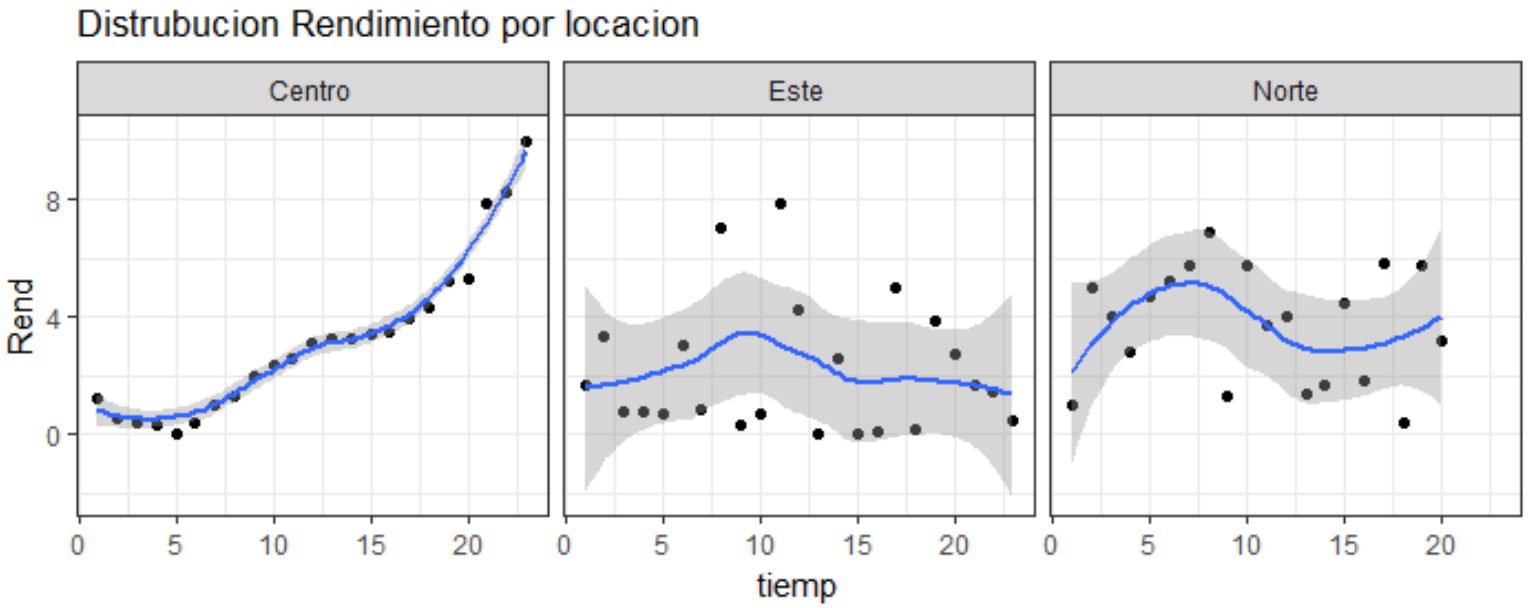
Colores y tendencia

```
ggplot(trigo,aes(x=tiemp,y =Rend ))+geom_point(aes(colour=Variedad))+geom_smooth(aes(colour=Variedad))
```



Facetas

```
ggplot(trigo,aes(x=tiemp,y =Rend ))+geom_point()+geom_smooth()+facet_grid(.~ubicacion) +theme_bw()+
  ggtitle('Distrubucion Rendimiento por locacion')
```



Tamaños, colores y transparencia

```
ggplot(trigo,aes(x=tiemp,y =Rend ))+geom_point()+geom_smooth()+facet_grid(.~ubicacion) +theme_bw()+
  ggtitle('Distrubucion Rendimiento por locacion')
```