


Fundamentos de Análisis de Datos

Gráficos

Dra. Andrea Alejandra Rey

Especialización en Ciencia de Datos - ITBA



Visualización de los datos.

Gráficos para datos multivariados

Gráficos para datos bivariados

Gráficos para datos univariados

ANÁLISIS UNIVARIADO

Gráficos

Pregunta

¿Por qué presentar los datos visualmente?

Los gráficos usados para representar datos

- ▢ destacan los aspectos más importantes,
- ▢ permiten describir adecuadamente conjuntos numerosos o complejos,
- ▢ complementan y argumentan información brindada en un texto,
- ▢ muestran claramente la presencia de tendencias o diferencias entre grupos,
- ▢ facilitan la comprensión,
- ▢ son rápidos y directos,
- ▢ convencen al lector,
- ▢ son fáciles de recordar.

Ejemplo

El porcentaje de tiempo dedicado por cada participante a cada una de las actividades de una competencia de triatlón está dado por la siguiente tabla.

Participante	Natación	Ciclismo	Carrera a pie
A	13	50	37
B	32	53	15
C	21	28	51
D	41	14	45
E	9	81	10
F	28	47	25
G	32	40	28
H	38	24	38

Gráfico de barras

- ➡ Está formado por un eje y una serie de barras etiquetadas.
- ➡ Cada barra está asociada a una categoría y representa el valor que una variable tiene para dicha categoría.
- ➡ Las barras pueden ser horizontales o verticales.
- ➡ Cuanto más largas son las barras, mayor es el valor de la variable.
- ➡ Se usa para comparar una sola variable entre varios grupos.
- ➡ Ayuda a reconocer patrones o tendencias.

Gráfico de barras

El porcentaje de tiempo dedicado al ciclismo por cada participante del triatlón se puede representar mediante un **gráfico de barras vertical**.

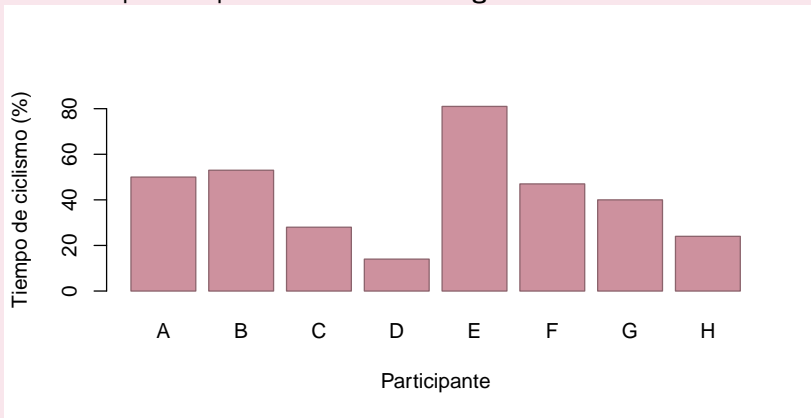


Gráfico de barras

El porcentaje de tiempo dedicado al ciclismo por cada participante del triatlón se puede representar mediante un **gráfico de barras horizontal**.

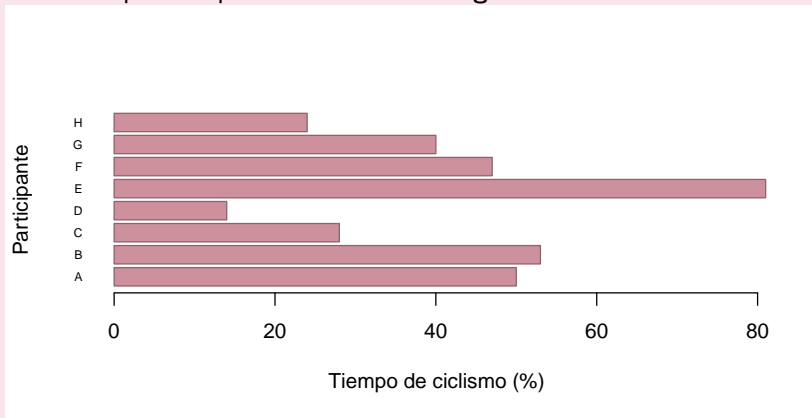


Gráfico de barras

El porcentaje de tiempo dedicado por cada participante a cada una de las actividades del triatlón se puede representar mediante un **gráfico de barras por grupo**.

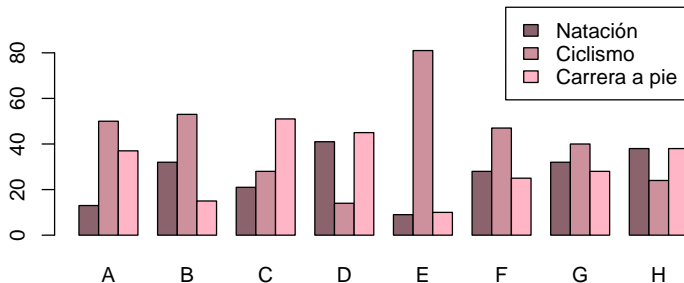


Gráfico de barras

El porcentaje de tiempo dedicado por cada participante a cada una de las actividades del triatlón se puede representar mediante un **gráfico de barras apilado**.

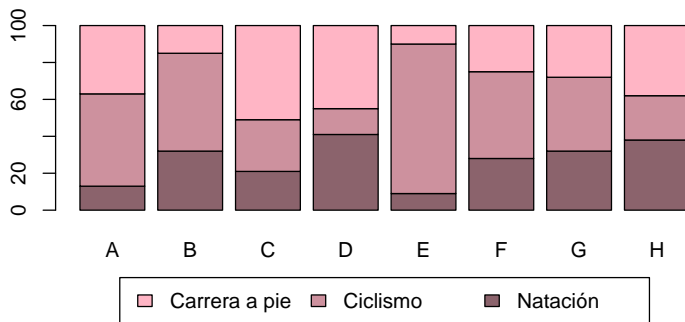


Gráfico circular

- ➡ Muestra grupos de datos -porciones- en proporción al conjunto completo de datos -torta-.
- ➡ Para calcular el valor del ángulo θ de un sector se aplica la regla de tres simple: $\theta = (\text{porcentaje del sector} \times 360^\circ)/100$.
- ➡ Resulta útil indicar el porcentaje correspondiente a cada porción.
- ➡ Son simples de usar.
- ➡ Son recomendables para un número reducido de categorías.
- ➡ No es aconsejable usarlos cuando las diferencias entre las regiones son muy pequeñas.

Gráfico circular

El porcentaje de tiempo dedicado por el participante A a cada actividad del triatlón se puede representar mediante un **gráfico circular**.

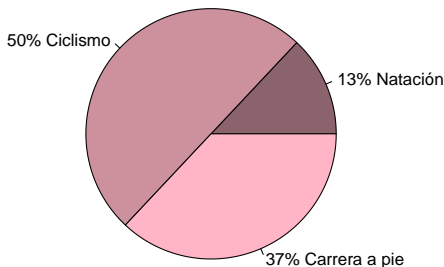


Gráfico circular

El porcentaje de tiempo dedicado por el participante A a cada actividad del triatlón se puede representar mediante un **gráfico circular “explotado”**.

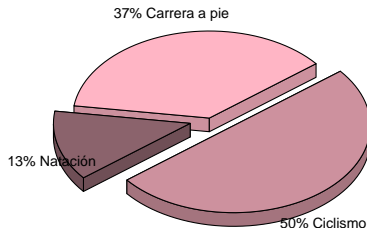


Gráfico circular

Si queremos comparar el porcentaje de tiempo dedicado por dos participantes a cada actividad del triatlón, debemos usar dos gráficos circulares.

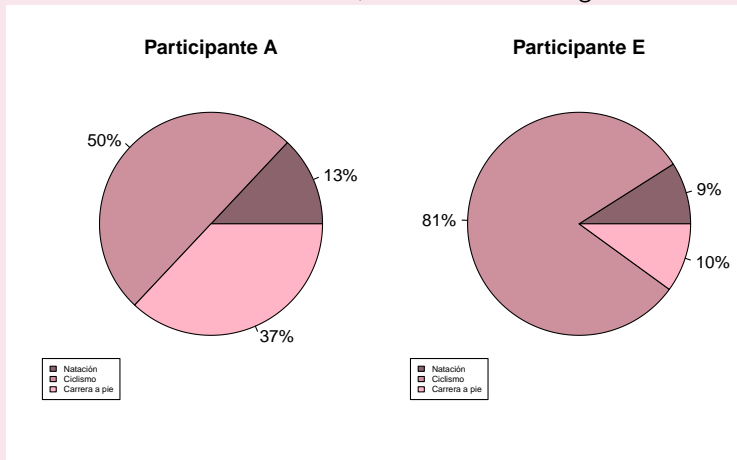


Gráfico de cajas

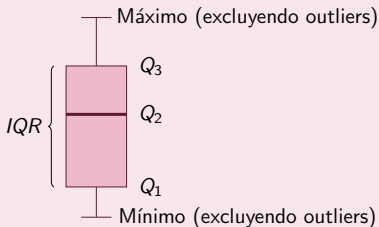
- ➡ Se usa para mostrar un resumen estadístico de una variables, incluyendo la mediana, el primer y el tercer cuartil, y los valores mínimo y máximo.
- ➡ Puede ser construido con tan sólo 5 datos.
- ➡ Puede ser horizontal o vertical.
- ➡ Permite identificar valores atípicos, como observaciones por debajo del límite inferior o por encima del límite superior.
- ➡ Los espacios entre las diferentes partes de la caja indican el grado de dispersión de los datos.
- ➡ Aporta más detalles acerca de las colas de la distribución.
- ➡ Presenta mayores ventajas para comparar la distribución en distintos grupos.
- ➡ Se suele utilizar su equivalente en inglés *boxplot*.

Gráfico de cajas

Pregunta

¿Cómo interpretamos un boxplot?

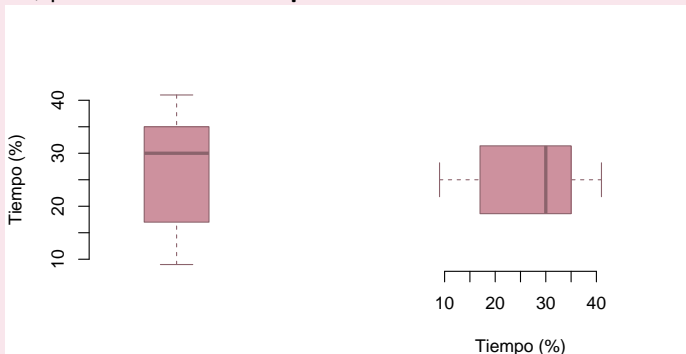
--- Límite superior: $Q_3 + 1.5/IQR$



--- Límite inferior: $Q_1 - 1.5/IQR$
• Outlier

Gráfico de cajas

Si queremos la distribución del tiempo dedicado por los participantes a la natación, podemos usar un **boxplot vertical** u **horizontal**.



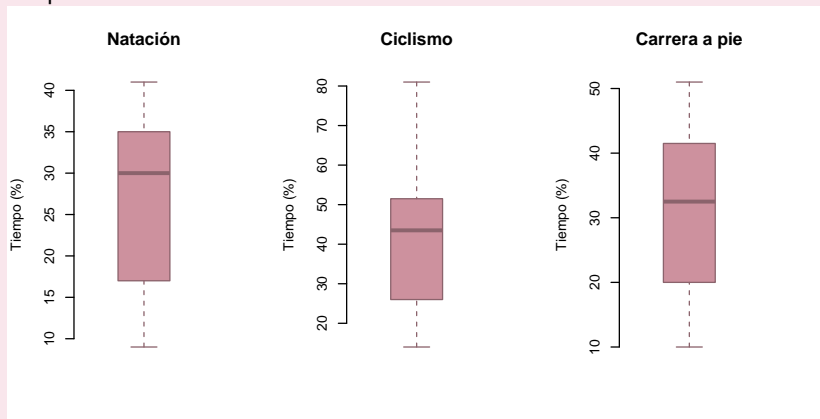
Mínimo: 9

Mediana: 30

Máximo: 41

Gráfico de cajas

Si queremos comparar la distribución del tiempo dedicado por los participantes a cada una de las actividades del triatlón, podemos usar varios boxplots.

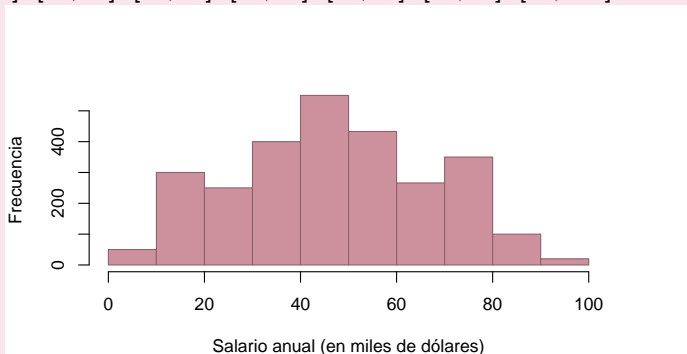


Histograma

- ➡ Se usa para representar datos medidos en un intervalo de escala.
- ➡ Luego de dividir el rango de valores posibles en grupos, se construye un rectángulo por cada grupo con una base de longitud igual al rango de valores en ese grupo y una altura de longitud igual al número de observaciones que caen en ese grupo.
- ➡ La frecuencia está dada por el área de la columna.
- ➡ Permite detectar observaciones atípicas o faltantes.
- ➡ Se utiliza para verificar si los datos siguen una distribución Normal.
- ➡ Es muy útil para conjuntos con un gran número de datos.

Histograma

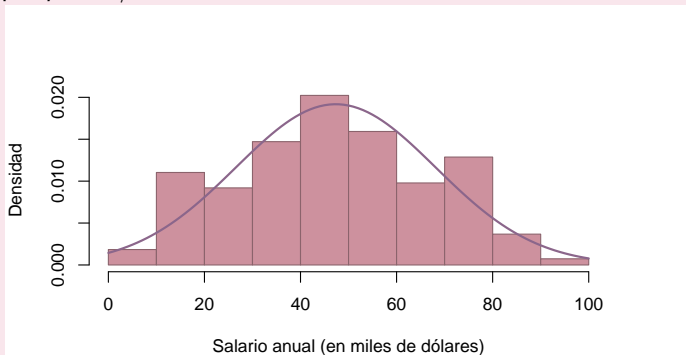
Una empresa estadounidense con 2719 empleados desea estudiar la distribución del salario anual de sus trabajadores calculado en miles de dólares. Supongamos que las clases están determinadas por los salarios comprendidos en cada uno de los siguientes intervalos: $[0, 10]$, $[11, 20]$, $[21, 30]$, $[31, 40]$, $[41, 50]$, $[51, 60]$, $[61, 70]$, $[71, 80]$, $[81, 90]$, $[91, 100]$.



Histograma

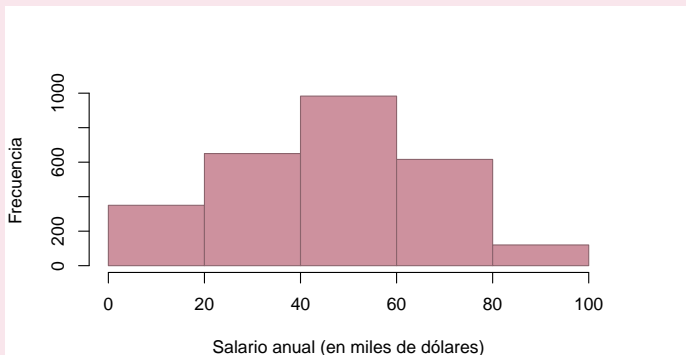
Para analizar “normalidad”, convertimos la frecuencia en densidad calculando probabilidades empíricas.

Es decir, si N es la cantidad total de datos y un grupo G_i tiene frecuencia f_i , la probabilidad de que una observación corresponda al grupo G_i está dada por $p_i = f_i/N$.



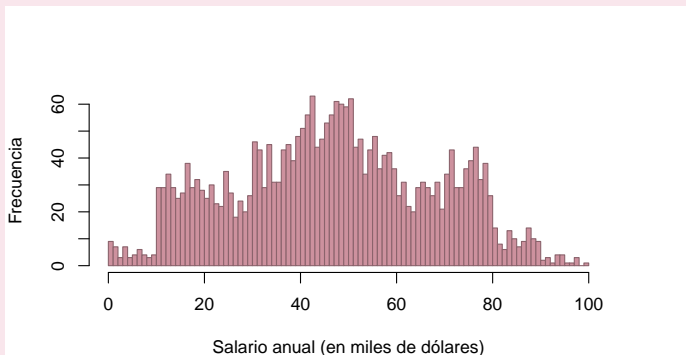
Histograma

Supongamos ahora que el conjunto de datos se divide en 5 clases.



Histograma

Supongamos finalmente que el conjunto de datos se divide en 100 clases.



Histograma

Resulta evidente que el número de clases, intervalos o *bins*, debe ser elegido correctamente de manera de visualizar la distribución real de la variable de interés.

Para lo que sigue, consideramos la **función techo** que asigna a un número real x , el menor de los enteros mayores o iguales a x . La notación es $\lceil x \rceil$. Supongamos que tenemos una muestra de tamaño n : $X = \{x_1, x_2, \dots, x_n\}$. Si k denota el número de bins y h el ancho de cada bin, entonces:

$$k = \left\lceil \frac{\max(X) - \min(X)}{h} \right\rceil.$$

Pregunta

¿Cómo elegimos el número de clases?
Equivalentemente, ¿cómo elegimos el ancho de cada columna?

Métodos para la selección de número de bins

Regla Mosteller–Tukey (1977)

$$k = \lceil \sqrt{n} \rceil$$

Regla de Velleman (1976)

$$k = \lceil 2\sqrt{n} \rceil$$

Regla de Rice

$$k = \lceil 2\sqrt[3]{n} \rceil$$

Regla de Sturges (1926)

$$k = 1 + \lceil \log_2(n) \rceil$$

Métodos para la selección de número de bins

Regla de Dixon-Kronmal (1965)

$$k = \lceil 10 \log_{10}(n) \rceil$$

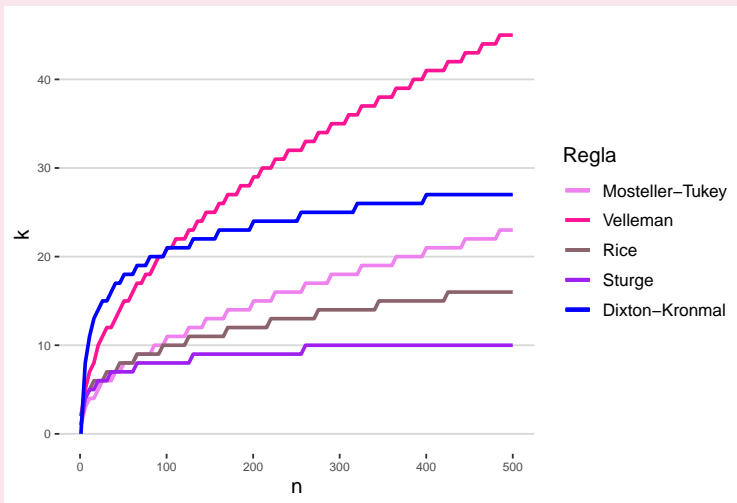
Regla de Doane (1976)

$$k = \left\lceil 1 + \log_2(n) + \log_2 \left(1 + g_1 \sqrt{\frac{(n+1)(n+3)}{6(n-2)}} \right) \right\rceil,$$

donde g_1 es la estimación de la asimetría dada por:

$$g_1 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^3}{\left[\frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^2 \right]^{3/2}}.$$

Comparación de métodos que dependen sólo del tamaño muestral



Comparación de métodos

Observación

- ➡ La regla de Doane es la única que considera la distribución de los datos, mientras que las demás sólo el tamaño muestral.
- ➡ La regla de Sturges funciona bien para conjuntos con una cantidad menor a 100 datos y para datos distribuidos normalmente y simétricos.
- ➡ La regla de Rice se presenta como una alternativa más simple de la regla de Sturges.
- ➡ La regla de Doane es una modificación de la regla de Sturges que mejora su rendimiento en el caso de datos no normales.
- ➡ La regla de Dixon-Kronmal resulta bastante efectiva en la práctica si $n < 100$.
- ➡ La regla de Mosteller-Tukey, también conocida como regla empírica, es la que usan algunos programas como Excel.

Métodos para la selección del ancho de bin

Regla de Scott (1979)

$$h = \frac{3.49s_X}{\sqrt[3]{n}}$$

Regla de Freedman-Diaconis (1981)

$$h = \frac{2(Q_3 - Q_1)}{\sqrt[3]{n}}$$

Comparación de métodos

Observación

- ➡ La regla Freedman-Diaconis funciona bien para conjuntos con una cantidad menor a 100 datos y para datos distribuidos normalmente y simétricos.
- ➡ La regla Freedman-Diaconis es poco sensible a la presencia de outliers y, a diferencia de la regla de Scott, arroja intervalos un poco más pequeños.
- ➡ Los valores atípicos pueden agrandar sobremanera el rango, aumentando el tamaño de los intervalos obtenido a partir de la regla de Strudge.

Histograma versus gráfico de barras

Ítem a comparar	Gráfico de barras	Histograma
Uso	Comparar diferentes categorías de datos.	Mostrar la distribución de variables.
Tipo de variable	Categóricas	Numéricas
Representación	Cada punto se representa en una barra separada.	Los puntos son agrupados en intervalos disjuntos y representados por un valor de <i>bin</i> .
Espaciado entre barras	Puede existir.	No existe.
Ordenamiento	Las barras pueden ser reordenadas.	Las barras no pueden ser reordenadas.

Ejemplo*



*Base de datos sleep de R

ANÁLISIS BIVARIADO

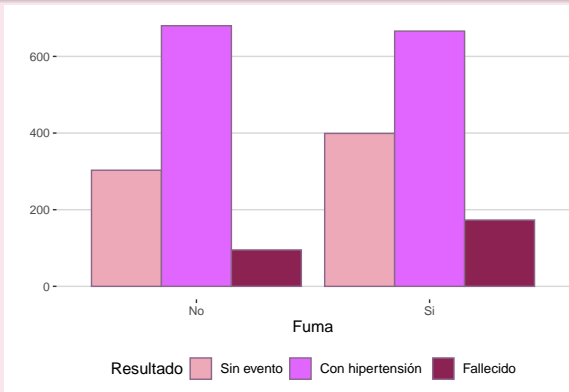
Gráfico de barras agrupadas

- ▶ Permite observar cómo se relacionan dos variables categóricas.
- ▶ Es equivalente al gráfico de barras univariado, pero ahora las barras tienen distintos colores asociados a la segunda variable categórica.

Gráfico de barras agrupadas

Pregunta

¿Influye el hecho de fumar en el resultado obtenido en un estudio?*



*Base de datos framingham disponible en el paquete LocalControl de R.

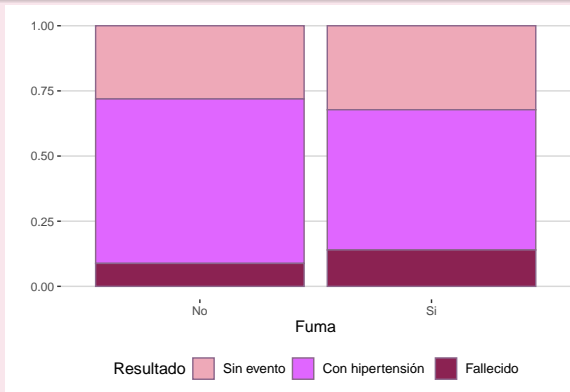
Gráfico de barras segmentadas

- ▶ Permite observar cómo se relacionan dos variables categóricas.
- ▶ Cada una de las barras suma el 100% de las observaciones correspondientes a la categoría representada por la barra.
- ▶ Cada barra se separa en partes que representan el porcentaje de observaciones dentro de esa categoría que pertenecen a cada grupo de la otra categoría.
- ▶ Permite comparar cómo se distribuyen los valores de una variable en cada categoría de la otra variable.

Gráfico de barras segmentadas

Pregunta

¿Influye el hecho de fumar en el resultado obtenido en un estudio?*



*Base de datos framingham disponible en el paquete LocalControl de R.

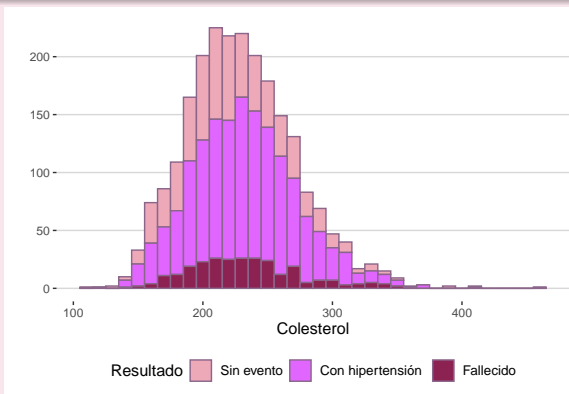
Histogramas agrupados

- ▶ Permiten observar cómo se relacionan una variable categórica y una variable cuantitativa.
- ▶ Si bien la cantidad de observaciones dentro cada grupo de la variable categórica puede variar, los histogramas agrupados permiten observar si las distribuciones son o no similares en cada grupo.

Histogramas agrupados

Pregunta

¿Se puede asociar el nivel de colesterol con el resultado de un estudio?*



*Base de datos framingham disponible en el paquete LocalControl de R.

Gráficos de densidad agrupados

- ▶ Permiten observar cómo se relacionan una variable categórica y una variable cuantitativa.
- ▶ Permiten observar si las distribuciones son o no similares en cada grupo de manera más clara que usando histogramas agrupados puesto que no dependen de la cantidad de observaciones.

Gráficos de densidad agrupados

Pregunta

¿Se puede asociar el nivel de colesterol con el resultado de un estudio?*



*Base de datos `framingham` disponible en el paquete `LocalControl` de R.

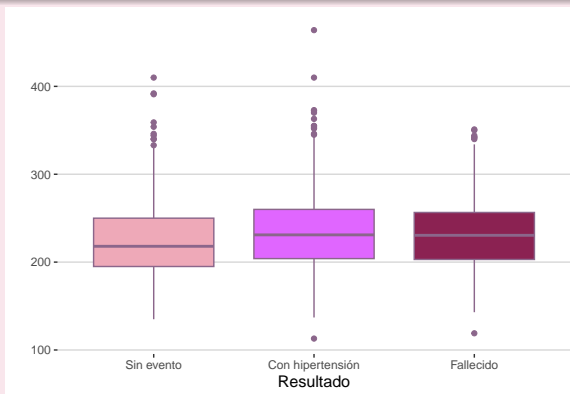
Boxplots por grupos

- ▶ Permiten observar cómo se relacionan una variable categórica y una variable cuantitativa.
- ▶ Permiten observar si las distribuciones son o no similares en cada grupo, siendo más útil para visualizar si las medias o medianas difieren entre los grupos.

Boxplots por grupos

Pregunta

¿Se puede asociar el nivel de colesterol con el resultado de un estudio?*



*Base de datos framingham disponible en el paquete LocalControl de R.

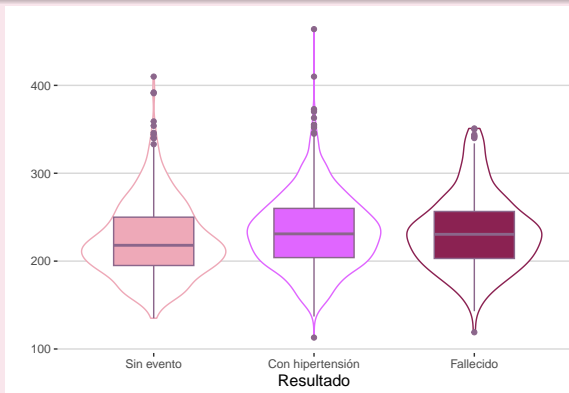
Gráficos de violín

- ▶ Permiten observar cómo se relacionan una variable categórica y una variable cuantitativa.
- ▶ Es una combinación de un boxplot y un gráfico de densidad superpuesto.

Gráficos de violín

Pregunta

¿Se puede asociar el nivel de colesterol con el resultado de un estudio?*



*Base de datos `framingham` disponible en el paquete `LocalControl` de R.

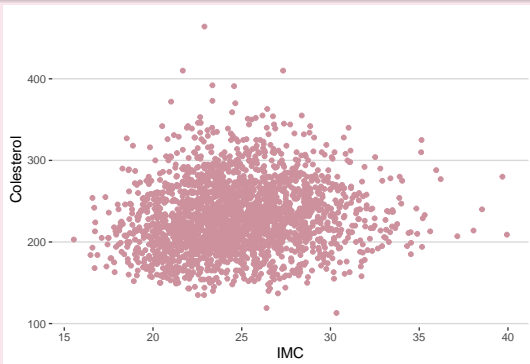
Gráfico de dispersión

- ▶ Permite observar cómo se relacionan dos variables cuantitativas.
- ▶ Cada observación se representa como un punto ubicado en el plano en base a los valores que toma cada variable, donde cada uno de los ejes corresponde a cada una de las variables.
- ▶ Los puntos se trazan pero no se unen.
- ▶ Se conoce por su equivalente en inglés **scatterplots**.

Gráfico de dispersión

Pregunta

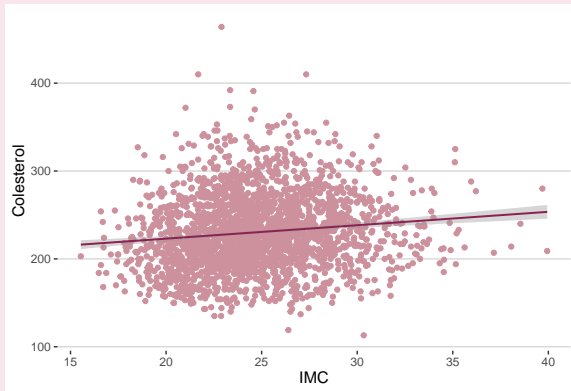
¿Existe una asociación entre el índice de masa corporal (IMC) y el nivel de colesterol?*



*Base de datos `framingham` disponible en el paquete `LocalControl` de R.

Gráfico de dispersión

Podemos agregar una línea de tendencia para facilitar la comprensión.



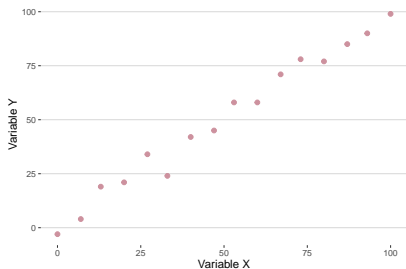
En este caso se sugiere que existe una asociación positiva leve entre las variables.

Gráfico de dispersión

Pregunta

¿Qué podemos observar en un scatterplot?

Relación lineal



Relación no lineal

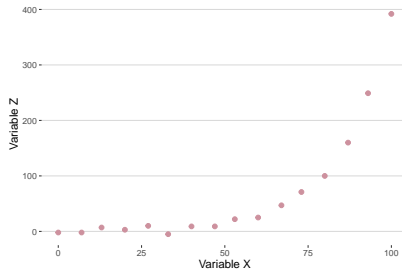
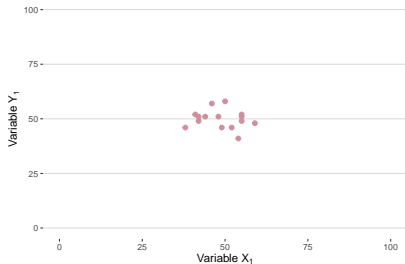


Gráfico de dispersión

Datos concentrados



Datos ampliamente dispersados

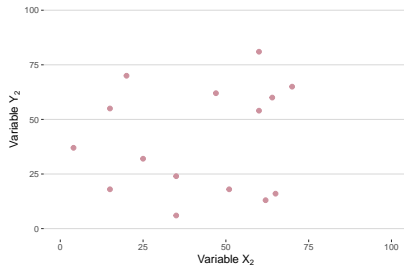


Gráfico de dispersión

Presencia de outliers

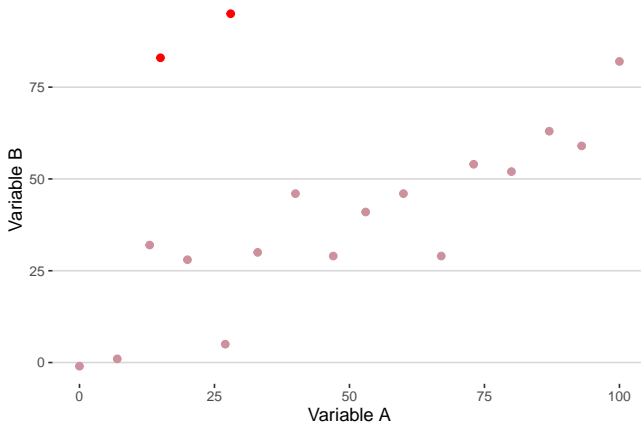


Gráfico de línea

- ▶▶▶ Permite observar cómo se relacionan dos variables.
- ▶▶▶ En general el eje vertical representa la cantidad o porcentaje de una variable, mientras que el eje horizontal suele representar unidades de medidas temporales.
- ▶▶▶ Facilita la visualización de características de los datos que indican una tendencia.
- ▶▶▶ Resulta fácil de construir dibujando una línea continua entre todos los puntos de la grilla.
- ▶▶▶ Puede utilizarse para comparar tendencias entre colecciones diferentes de datos.
- ▶▶▶ Suele verse como el gráfico de una serie de tiempo.

Pregunta

¿Cómo y para qué podemos aplicar gráficos de línea?

Gráfico de líneas

Gráfico de una tendencia a lo largo del tiempo

Estudio de la fluctuación de la fuerza laboral de marzo a septiembre en un grupo de estudiantes de un colegio de educación media.

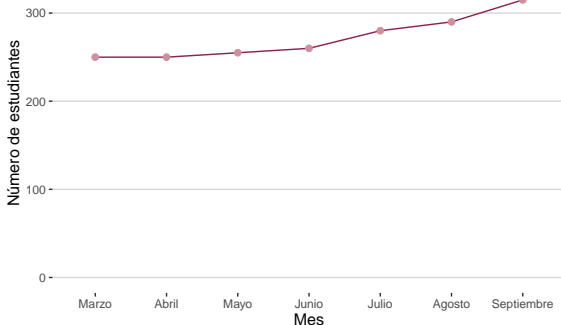


Gráfico de líneas

Comparación de dos variables relacionadas

Estudio del promedio de dólares donado en función de la edad de los donates.

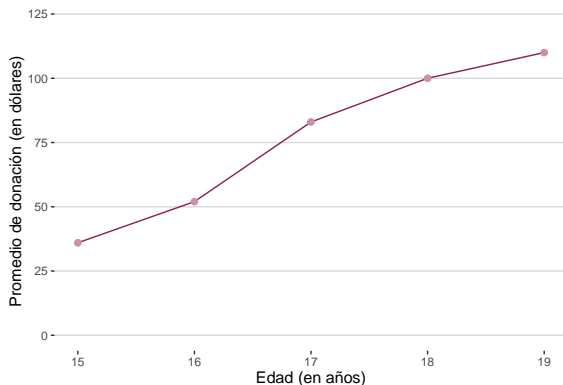
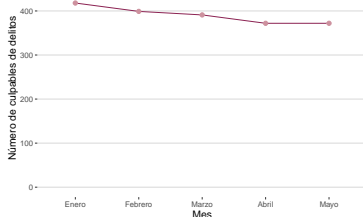
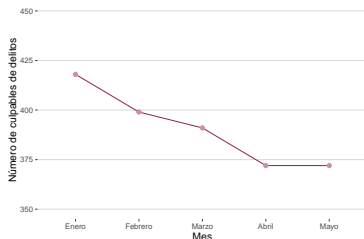


Gráfico de líneas

Uso correcto de escalas

Número de culpables de delitos en cierta localidad.



El gráfico de la izquierda se enfoca en un rango pequeño de valores exagerando la tendencia. Sin embargo, la escala elegida en el gráfico de la derecha muestra mejor cuán pequeña fue realmente la disminución en el número de delincuentes culpables. Ambos gráficos pueden ser útiles según el contexto y siempre teniendo en cuenta la escala que se utiliza al interpretar un gráfico.

Gráfico de líneas

Comparación de varios ítems en un mismo período de tiempo

Estudio del uso de teléfonos celulares en dos ciudades.



Ejemplo*



*Bases de datos HairEyeColor de R

Ejemplo*



* Bases de datos ChickenWeight de R

ANÁLISIS MULTIVARIADO

Desafíos de la visualización de datos multivariados

- ➡ El objetivo es encontrar un mapeo adecuado de un conjunto de datos multivariados de gran dimensión en un espacio 2D.
- ➡ La asociación entre los atributos de los datos debe ser presentada de tal forma que no abrume la capacidad visual.
- ➡ Es importante que los diferentes atributos puedan verse al mismo tiempo de manera holística para un análisis integrado, y de manera separada e independiente para el estudio de cada dimensión.
- ➡ Dependiendo de la técnica elegida, se pueden obtener diferentes conclusiones. Sin embargo, no hay una regla general que establezca un orden de preferencia.

Desafíos de la visualización de datos multivariados

- ➡ La visualización puede proporcionar una descripción general cualitativa de un conjunto grande de datos complejos con el fin de encontrar estructuras, características, patrones, tendencias y relaciones de manera más efectiva.
- ➡ Debido a la alta dimensionalidad de los datos multivariados, a veces es necesario sacrificar la capacidad de mostrar los detalles de cada atributo.
- ➡ Debe haber un equilibrio entre la cantidad de información, la simplicidad y la precisión.
- ➡ Puesto que no sabemos qué conocimiento valioso está presente en los datos, la visualización pretende encontrar información.

Técnicas de visualización

Proyecciones
geométricas

Matriz de scatterplots

Icónicas

Caras de Chernoff - Glifos de estrellas

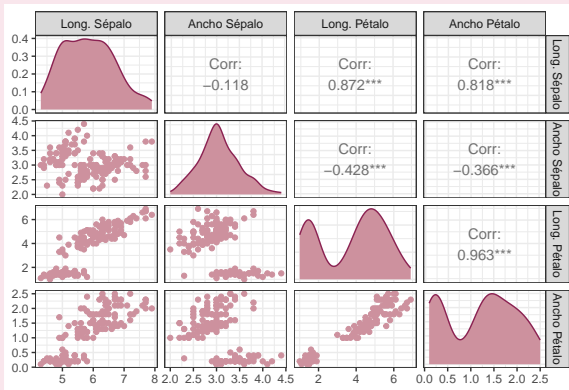
Jerárquicas

Mapa de calor

Matriz de scatterplots

- Organiza todos los diagramas de dispersión por pares de variables en una grilla cuadrada de m^2 celdas donde m es la cantidad de variables.
- La celda ubicada en la fila i y la columna j representa la relación entre las variables X_i y X_j .
- La grilla es simétrica respecto de la diagonal.
- Algunos programas usan uno de los triángulos para mostrar el coeficiente de correlación.

Matriz de scatterplots



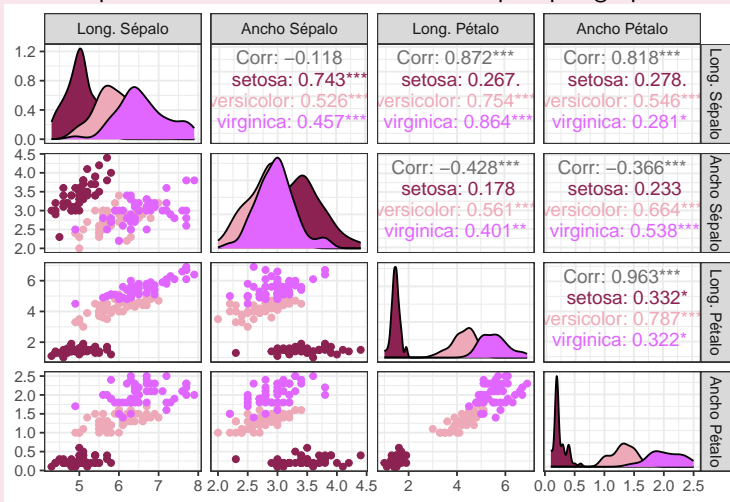
Ejemplo

Sobre la base de datos iris disponible en R.

- Diagonal: gráfico de densidad de cada variable.
- Triángulo inferior: scatterplot de las variables involucradas.
- Triángulo superior: coeficiente de correlación de cada par de variables.
- Los asteriscos indican significancia estadística.

Matriz de scatterplots

También se puede realizar una matriz de scatterplot por grupos.

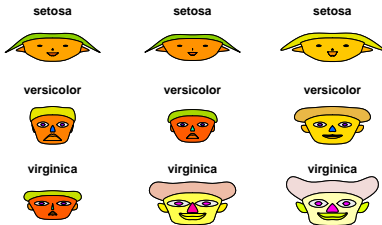


Caras de Chernoff

- ➡ Su nombre se debe a su creador Herman Chernoff, en 1973.
- ➡ Se usan caras humanas simples para la representación.
- ➡ Cada atributo es mapeado en una característica facial diferente.
- ➡ Se basan en el supuesto de que como las personas podemos leer fácilmente las expresiones en los rostros, deberíamos ser capaces de reconocer pequeñas diferencias que se presenten en los datos.

Caras de Chernoff

Sobre la base de datos iris.



Parámetros de características

- 1 altura de la cara
- 2 ancho de la cara
- 3 forma de la cara
- 4 altura de la boca
- 5 ancho de la boca
- 6 curva de la sonrisa
- 7 altura de los ojos
- 8 ancho de los ojos
- 9 altura del cabello
- 10 ancho del cabello
- 11 estilo del cabello
- 12 altura de la nariz
- 13 ancho de la nariz
- 14 ancho de las orejas
- 15 altura de las orejas

Glifos de estrellas

- ➡ Cada registro se representa por una figura con forma de estrella con un rayo por cada variable.
- ➡ La longitud de cada rayo es proporcional al valor de la variable correspondiente.
- ➡ Generalmente, cada variable es normalizada en el intervalo $[0, 1]$.
- ➡ Los extremos de cada intervalo se suelen conectar mediante líneas.
- ➡ En la medida que el número de rayos aumenta, se hace más difícil su separación. Se sugiere una separación de al menos 30° .

Glifos de estrellas

Sobre la base de datos iris.



setosa



setosa



setosa



versicolor



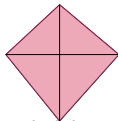
versicolor



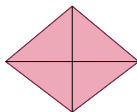
versicolor



virginica



virginica



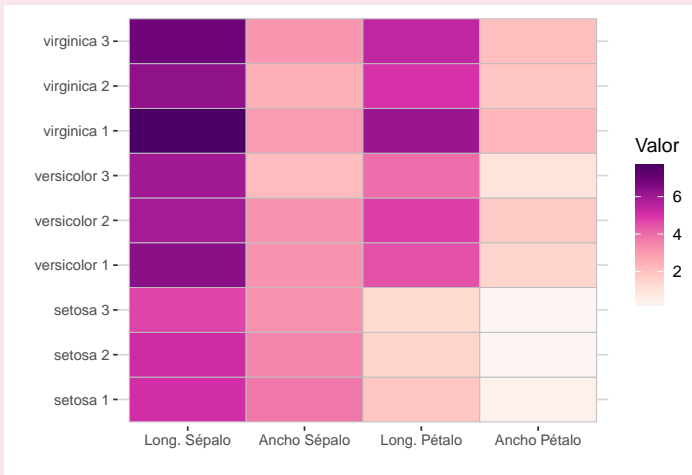
virginica

Mapa de calor

- ▶ Permite visualizar la agrupación jerárquica, donde los valores de los datos se transforman a escala de color.
- ▶ La magnitud del valor de una variable se representa en un código de colores que va de menor a mayor intensidad.
- ▶ El primer agrupamiento jerárquico se realiza tanto para las filas como para las columnas de la matriz de datos.
- ▶ Las columnas/filas de la matriz de datos se reordenan de acuerdo con el resultado del agrupamiento jerárquico, colocando observaciones similares cerca unas de otras.
- ▶ Ayuda a encontrar las variables que parecen ser características de cada conglomerado de muestra.

Mapa de calor

Sobre la base de datos iris.



Ejemplo*



*Base de datos Salaries del paquete carData de R