Visualización de datos I

La representación gráfica de las variables de un conjunto de datos es un recurso muy útil para **obtener** y **comunicar** información, de una manera directa y fácil de interpretar. Las gráficas de los estadísticos descriptivos, como los que ven en este curso, son parte esencial del **análisis exploratorio** de datos.

La visualización de datos es un medio imprescindible para hacer ciencia con datos, con diferentes objetivos y usos:

1. **Interpretación.** Las gráficas brindan información más fácil/directa de interpretar que una tabla o un párrafo. Nuestros ojos y nuestro cerebro son especialistas en identificar patrones visuales.
2. **Comunicación.** Así como una gráfica facilita la interpretación de los datos para el analista, también lo hace para el público receptor. Incluso si este público no es especialista en el tema, ni maneja el lenguaje técnico. Para esto, debemos aprender códigos del lenguaje visual, es decir, de cómo percibimos.

1 Tipos de gráficos

En primer lugar, nos centraremos en conocer los diferentes tipos de gráficos disponibles y su aplicación. Los gráficos podemos agruparlos por **tipo y cantidad de variables**, o por **función**. Cuando hablamos de la función de un gráfico nos estamos refiriendo al **tipo de información** que brinda. Según su función las gráficos pueden ser:

- De **distribución**
- De **ranking** o comparación entre grupos
- De **correlación** entre variables
- Series de tiempo
- Mapas

Las gráficas de **distribución** nos permiten conocer la distribución de nuestros datos. Representan gráficamente a una única variable, estas gráficas son soporte (e incluso a veces reemplazo) de las pruebas de normalidad, por ejemplo.

Las gráficas de **ranking** posibilitan la comparación de diferentes aspectos de un conjunto de datos. Son muy útiles para comparar entre datos cuantitativos agrupados en categorías, que pueden ser nominales u ordinales. También podemos realizar gráficos de este tipo a partir de dos o más gráficos de distribución, más adelante veremos un ejemplo.

Los gráficos de **correlación** necesitan de dos o más variables, ya que nos permiten conocer la relación entre ellas. Las **series de tiempo** son un tipo particular de gráfico de correlación, donde siempre se evalúa la relación de una variable con el tiempo.

Los **mapas** son recursos gráficos avanzados, muy útiles para conocer la relación de una o más variables con el espacio. En este curso no aprenderemos a realizar mapas, pero es importante que sepan que existen y cuando nos sería beneficioso implementarlos.

A su vez, los gráficos pueden ser **unidimensionales**, **bidimensionales** o **tridimensionales** según la cantidad de variables representadas. Como sus nombres lo indican, en una gráfica unidimensional estaremos visualizando una única dimensión, es decir, la información referida a una única variable. En cambio, en un gráfico bidimensional tendremos la información referida a dos variables (dos dimensiones).

Para facilitar su aprendizaje, agrupamos las distintas gráficas según la cantidad y tipo de variables, con subcategorías con las funciones de cada gráfico. Recordemos que cuando hablamos de tipo de variable nos estamos refiriendo a que un dato puede ser **cuantitativo** o **cualitativo**. Al mismo tiempo, las variables cuantitativas pueden ser **discretas o continuas**, y las cualitativas pueden ser **nominales u ordinales**.

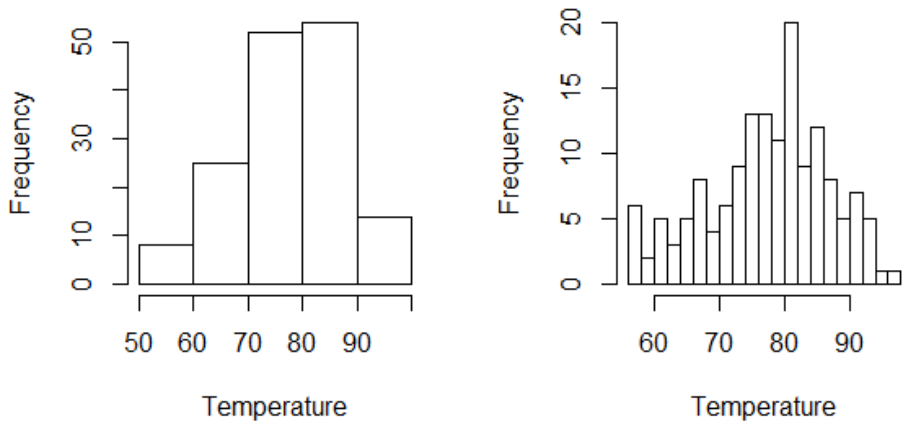
1.1 / Gráficos unidimensionales

Este tipo de gráficos nos permite conocer el conjunto de datos con el que estemos trabajando, qué valores toman sus variables, como se distribuyen estos valores, etc. Cada gráfico unidimensional nos va a estar hablando de una única dimensión. Si pensamos, por ejemplo, en una gráfica que nos muestra la temperatura que hace en determinada ciudad durante el verano, la dimensión que está siendo representada gráficamente es la temperatura.

1.1.1 / Gráficos de distribución

Los gráficos de distribución requieren que la variable a representar sea **cuantitativa y continua**. Existen diferentes gráficos de este tipo, los más utilizados son los histogramas, las gráficas de densidad y los diagramas de cajas (o boxplot).

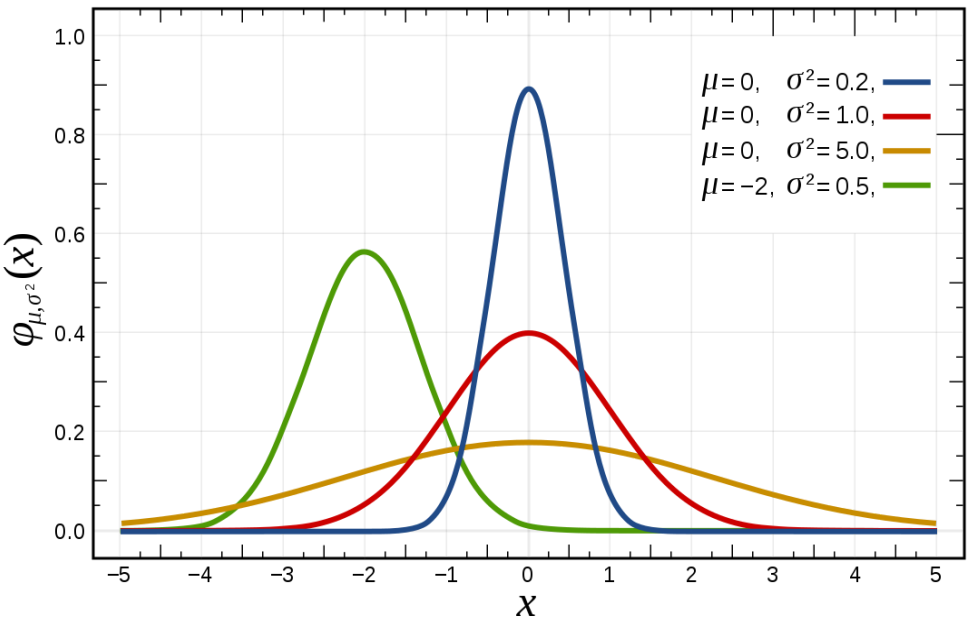
1. Histogramas



Histogramas de frecuencia. Fuente de imagen: www.datamentor.io

Un histograma es una representación gráfica de una variable en forma de barras, donde la superficie de cada barra es proporcional a la frecuencia de los valores representados. Nos permite obtener una primera vista general, o panorama, de la distribución de la población respecto a una característica (nuestra variable cuantitativa y continua). Un histograma puede representar frecuencias o densidad de probabilidad. En este último caso, la suma de las superficies de todas las barras del histograma da 1 (es decir, el 100% de probabilidad).

2. Gráficos de densidad de probabilidad

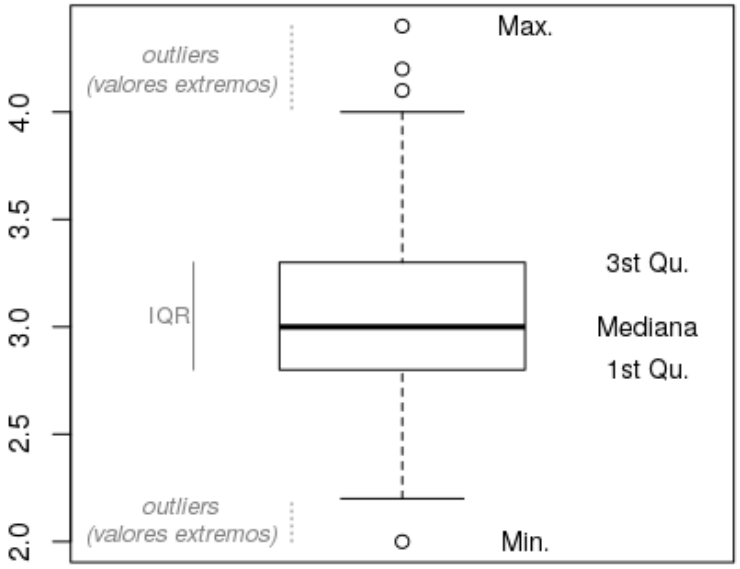


Distribución normal. Fuente de imagen: www.commons.wikimedia.org

Estas gráficas representan la distribución de nuestros datos, son una variante de los histogramas. En lugar de emplear barras, utilizan una línea y el área bajo la curva es siempre igual a 1. El pico en el gráfico (puede haber más de uno) nos indica el valor alrededor del cual se concentran las observaciones. En el ejemplo de la figura vemos diferentes distribuciones normales (curvas simétricas a ambos lados). La curva con μ = 0 (línea azul) corresponde a la distribución normal estándar. Una de las ventajas de este tipo de gráficos es que nos permiten distinguir más fácilmente la forma de la distribución, y si se ajusta, o no, a una función teórica (normal, poisson, etc.).

Al mismo tiempo, en la figura se presentan tres gráficos de densidad juntas. Esto nos podría hacer pensar que se trata de un gráfico tridimensional, ya que estamos mostrando la distribución de tres variables diferentes. Sin embargo, se trata de un gráfico **unidimensional múltiple**. Podemos pensarlo como un agrupamiento de gráficas, es un recurso muy útil para simplificar la interpretación de los datos.

3. Diagramas de caja



Boxplot o diagrama de caja. Fuente de imagen: www.picandoconr.wordpress.com

Los diagramas de cajas también brindan información sobre la distribución de nuestra variable de interés. Son gráficos muy útiles ya que permiten identificar visualmente la mediana, el mínimo, el máximo, el primer y tercer cuartil, e incluso los valores extremos (outliers).

Como vemos en la figura, la línea gruesa del medio de la caja es la mediana. Este valor no siempre va a estar en el medio de la caja, esto nos indicará si la distribución de nuestros datos es simétrica o sesgada hacia uno de los lados. Los bordes inferior y superior de la caja están formados por el primer y tercer cuartil respectivamente (Q1 y Q3). Recordemos que los cuartiles surgen de ordenar nuestros valores de menor a mayor, y representan las cuatro porciones en que podemos dividirlos. Q1 es la mediana de la primera mitad de los datos, el segundo cuartil es la mediana de todo el conjunto de datos, y Q3, análogo al primero, es la mediana de la segunda mitad de los datos. La diferencia entre Q1 y Q3 se denomina Rango Intercuartílico (IQR, de sus siglas en inglés) y nos da una idea de la dispersión de los datos. Cajas más achatadas indican baja dispersión, mientras que cajas más largas verticalmente son señal de una mayor dispersión.

Por debajo y por encima de la caja se extienden los bigotes (las líneas punteadas), por fuera de estos caerán los valores atípicos, si es que los hay. Se consideran atípicos los valores inferiores a Q1–1.5*IQR o superiores a Q3+1.5*IQR. Aquellos valores máximos y mínimos que no son valores atípicos serán los extremos de los bigotes.

1.1.2 / Gráficos de ranking

Los gráficos de distribución requieren que la variable a representar sea **cuantitativa** y que esté agrupada por **categorías**. Las categorías pueden ser otra variable dentro de nuestro conjunto de datos, por lo general se tratará de un dato de tipo **cualitativo**, tanto nominal como ordinal. Es entre estas categorías que nos interesa establecer **comparaciones**. Existen diferentes gráficas de este tipo, los más utilizados son los de barras, los de torta, y los gráficos de distribución múltiple.

1. Gráficos de barra

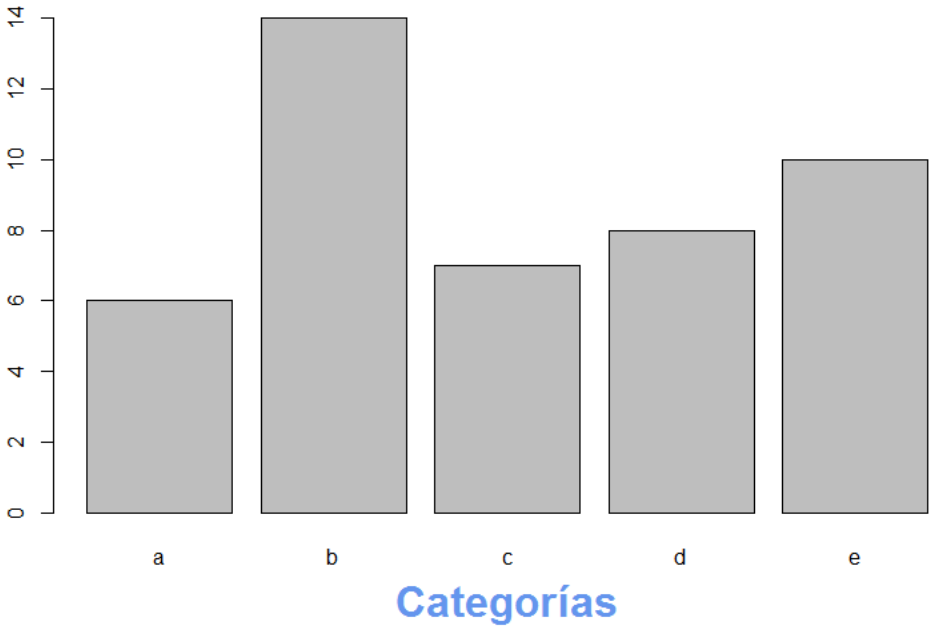


Gráfico de barras con categorías nominales.

A simple vista, los gráficos de barras pueden confundirse con los histogramas. En este caso, la longitud de las barras es proporcional a la magnitud que estemos representando. Un ejemplo fácil: cantidad de observaciones de cada categoría, en el ejemplo de la figura vemos que la categoría b presenta el mayor número de observaciones, o registros, dentro del conjunto de datos.

A su vez, en estos gráficos las barras se encuentran separadas con una distancia entre ellas (que debe ser constante), al contrario que los histogramas donde las barras están unidas. Esta distancia entre barras indica que los datos son discretos, y no continuos. Son de los gráficos más empleados, fáciles y directos de interpretar!

2. Gráfico de torta

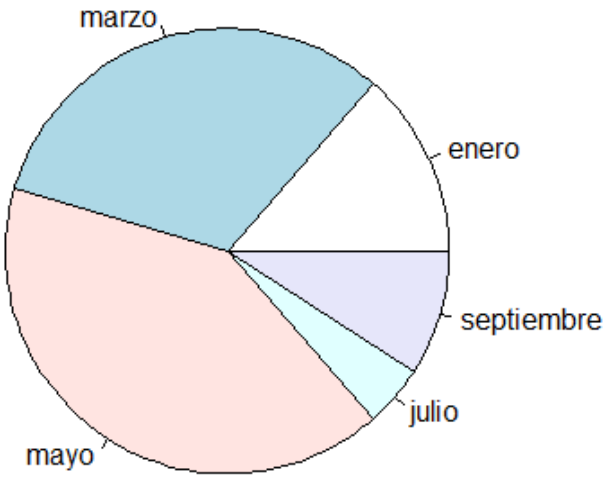


Gráfico de torta con categorías ordinales.

Los clásicos gráficos de torta se utilizan a menudo para representar porcentajes o proporciones. Cada una de las porciones en el círculo representa una categoría, y el tamaño de cada porción es relativo a su relación con el todo. El total de categorías constituye el 100% de nuestros datos. En el ejemplo vemos categorías mensuales, y los datos cuantitativos que están asociados a los meses podrían ser, por ejemplo, la cantidad de turistas que llega a una dada ciudad, en cada mes.

Si bien los gráficos de torta son muy conocidos y comunes, su uso no es recomendable. Nuestra capacidad para estimar relaciones de proporción, o diferencias, entre **áreas** es mucho menor que, por ejemplo, entre **longitudes o posiciones**, como ocurre en un gráfico de barras.

3. Gráficos de distribución múltiple

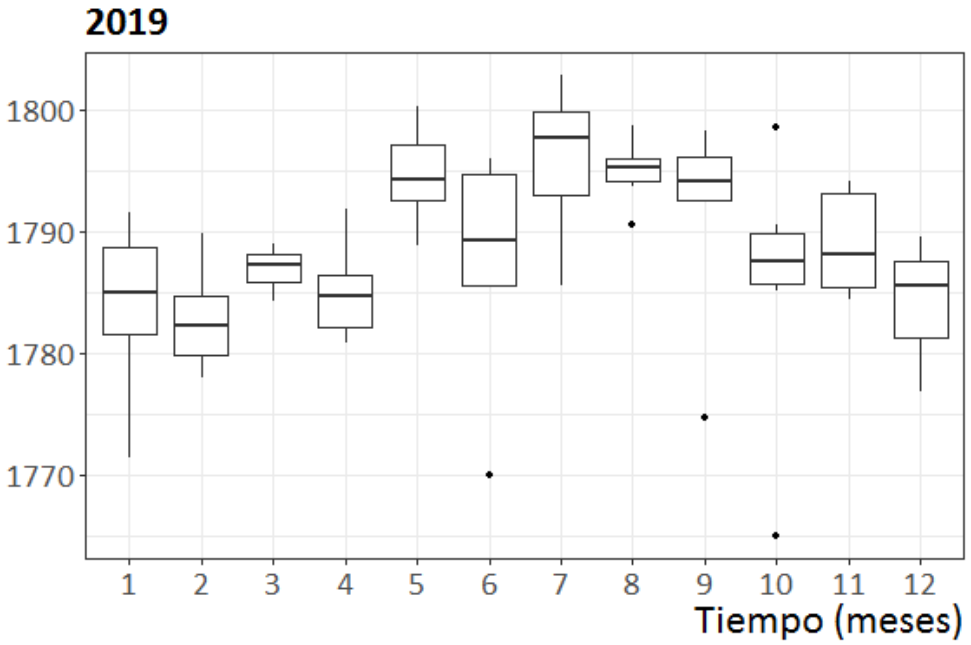


Gráfico de torta con categorías ordinales.

Los gráficos unidimensionales de distribución que vimos, como los histogramas, las gráficas de densidad y los boxplot, pueden ser presentados de forma simultánea para facilitar la comparación entre los agrupamientos o categorías. El requisito indispensable, para que la interpretación no sea errónea, es que todos los datos a representar en el gráfico compartan unidad y escala. Es decir, que compartan exactamente los mismos ejes vertical y horizontal.

En el ejemplo de arriba, vemos 12 boxplots diferentes, con datos agrupados por mes. Todos los boxplot muestran la distribución de los valores del eje vertical, no hay ningún mes que tenga datos en diferentes unidades o con algún factor de escala aplicado. Esto nos permite distinguir, por ejemplo, que en el mes 7 la mediana de los valores fue máxima. Hasta qué punto consideramos que la diferencia entre distribuciones es significativa o no es un tema para profundizar, existen test estadísticos para evaluar estas cuestiones. De forma muy general, en el caso de los boxplot podemos considerar al solapamiento entre cajas como un indicio de que la diferencia no es significativa.

1.2 / Gráficos bidimensionales

Estos gráficos nos permiten conocer la **relación** que existe, o no, entre dos variables de nuestro conjunto de datos. En este caso estaremos mostrando información referida a **dos dimensiones**, por ejemplo, los valores de temperatura y las lluvias que ocurren en un dado lugar durante un tiempo determinado. En este ejemplo, las variables a graficar son la temperatura y las lluvias. Cada variable es asignada a uno de los ejes. Por convención, la **variable dependiente** (la que modifica su valor a partir de un cambio en el valor de la otra) se representa en el eje vertical, y la **variable independiente** en el eje horizontal.

1.2.1 / Gráficos de correlación

Los gráficos de correlación nos permiten evaluar la relación entre nuestras dos variables. En este caso, ninguna de las variables es una magnitud temporal o espacial (par de coordenadas). El objetivo general es relacionar la variación de una característica de interés con factores de causa potenciales, para explicar cómo cada factor contribuye a esa variación. Existen diferentes gráficas de correlación, las principales son los **gráficos de dispersión**, de **densidad 2D** y los **mapas de calor**. Estos últimos, sin embargo, son gráficos multidimensionales por lo que no los daremos en este curso.

1. Gráficos de dispersión

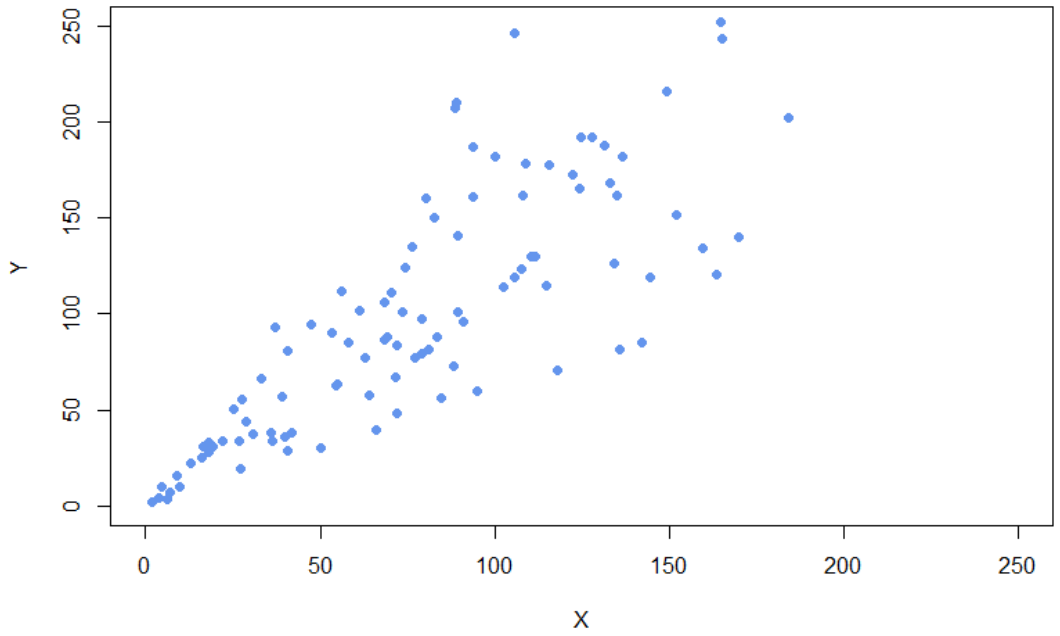


Gráfico de dispersión tradicional.

En los gráficos de dispersión cada par de valores (x, y) se presenta como un punto. Las variables tendrán una correlación positiva si son directamente proporcionales, es decir que, cuando una aumenta, la otra también lo hace proporcionalmente. En cambio, la correlación será negativa si las variables son inversamente proporcionales (cuando una disminuye, la otra aumenta en igual proporción).

Estas gráficas son muy útiles para distinguir el **tipo de relación** que existe entre las variables. Esta puede ser nula, lineal, exponencial, logarítmica, etc. Para distinguir visualmente el tipo de relación nos fijamos en la **forma** que toman los puntos en el gráfico. Si la relación es lineal, tanto positiva como negativa, los puntos caerán distribuídos sobre y alrededor de una recta imaginaria. Mientras mejor dibujen los puntos una recta, mejor será el ajuste lineal. Para otro tipo de relaciones el criterio es el mismo, por ejemplo, en una relación exponencial los puntos dibujaran una curva exponencial imaginaria. Luego, también podemos probar diferentes modelos y evaluar si el ajuste con los datos es bueno, o no, numéricamente.

2. Gráficos de densidad 2D

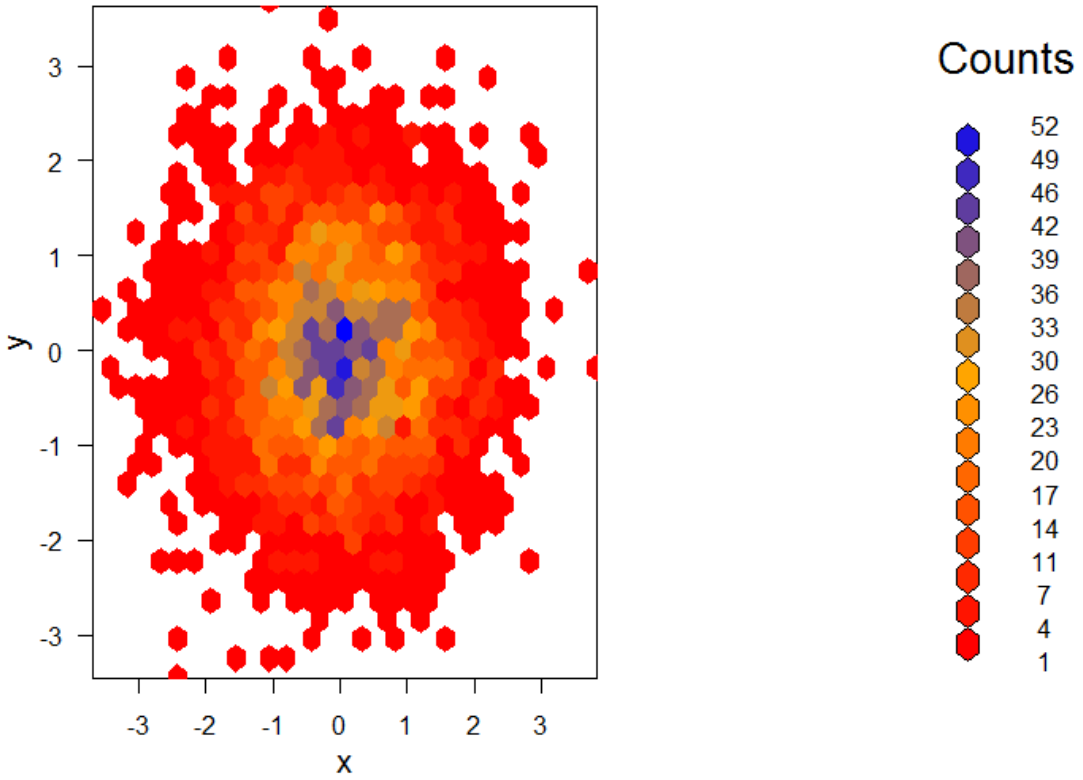


Gráfico de densidad 2D con celdas hexagonales.

En estos gráficos, una variable se asigna al eje vertical y otra al horizontal, al igual que en una gráfica de dispersión. Luego, el número de observaciones es representado por un gradiente de color dentro del espacio bidimensional formado por los ejes cartesianos. De esta forma, al igual que en un gráfico de densidad unidimensional, podemos ver alrededor de qué valores se centran la mayoría de las observaciones. En este caso, al ser una gráfica bidimensional, este valor central es un par ordenado (x, y). El gradiente de color y la leyenda asociada nos permite identificarlo. En el ejemplo de arriba, las celdas de color azul intenso indican alrededor de 49-52 osbervaciones, mientras que las celdas rojas indican entre 1 y 4 observaciones. Podemos ver que la mayoría de nuestros datos se centran en el par (0,0).

Una forma intuitiva y fácil de interpretar estos gráficos es pensar que estamos observando algo similar a una montaña vista desde arriba. El pico de la montaña son los valores donde se concentra la mayor cantidad de datos, análogo a lo que nos muestra una gráfica de densidad de probabilidad simple.

2 Seleccionando un gráfico adecuado

Las gráficas que vimos son una gran parte del total de gráficos y recursos visuales disponibles para evaluar un conjunto de datos, o para explicar algo en particular del mismo. Antes de ponernos a graficar, es útil que nos hagamos un par de preguntas:

1. **¿Qué es lo que quiero saber de mi conjunto de datos?** Una forma directa de responder esto es a través de preguntas como:

    - **¿En cuál o cuáles variables pongo mi atención?** Si nos interesa una única variable, entonces buscaremos un gráfico unidimensional. En cambio si queremos conocer la relación entre dos variables, nos centraremos en los gráficos bidimensionales de correlación.

    - **¿De qué tipo de variables se trata? ¿Son cuantitativas? ¿Continuas o discretas?** Si nos interesa una única variable cuantitativa y continua, vamos a necesitar un gráfico de distribución. En cambio, si nos interesa evaluar una variable cualitativa (pensemos en una columna denominada “frutas”, adentro los registros pueden ser cosas como “pera”, “frutilla”, “manzana”, etc.) vamos a limitarnos a un gráfico de ranking (atención, debe haber un valor cualitativo asociado a las variables cuantitativas para poder hacer el gráfico).

2. **¿Cuál es el gráfico que mejor mostrará esa información?** Esto va a depender de varios factores. Si quiero conocer la distribución de una variable continua, y al mismo tiempo la dispersión de sus valores, me va a convenir un diagrama de cajas que muestra ambos factores con claridad. Pero si me interesa evaluar si mis datos tienen una distribución normal, por ejemplo, me va a convenir un gráfico de densidad de probabilidad. A su vez, si estamos en la instancia del análisis explicativo (donde queremos transmitir un mensaje determinado a una audiencia dada) tendremos que tener en cuenta otras cuestiones, como el mensaje en sí, las características de nuestra audiencia, etc. Esta parte la cubriremos en el segundo encuentro de Visualización de Datos!

3 Bibliografía y referencias recomendadas

- Charte Ojeda, Francisco. (2014) Análisis exploratorio y visualización de datos con R. https://fcharte.com/assets/pdfs/ExploraVisualizaConR-FCharte.pdf
- Correa, Juan Carlos y González, Nelfi. (2002). Gráficos estadísticos con R. https://cran.r-project.org/doc/contrib/grafi3.pdf
- Lillis, David Alexander. (2014) R Graph Essentials. Packt Publishing, Birmingham: ISBN: 9781783554553.
- ggplot2 Cheatsheet: https://raw.githubusercontent.com/rstudio/cheatsheets/main/pngs/data-visualization.png