Análisis exploratorio de datos © EDICIONES ROBLE, S.L.

Índice

Análisis exploratorio de datos

I. Introducción

A la hora de entender la estructura de los datos, es imprescindible la representación de los mismos. “Más vale una imagen que mil palabras o mil resultados de fórmulas” sería un enunciado aplicable al caso.

En una EDA (Exploratory Data Analysis), lo que se pretende es realizar varios gráficos que puedan aportar información en forma de relaciones, acumulaciones, anomalías y densidades en las distribuciones de los mismos. Asimismo, tras la EDA, se forman convicciones que son útiles, ya que guían a lo largo del tratamiento y ayudan a la generación de modelos que se hará posteriormente. Se trata de construir una suerte de paseo visual por el conjunto de datos a estudiar que, a través de la creatividad y la curiosidad, aporte información intuitiva sobre la que se construirá, después, formalmente.

Para realizar este “paseo”, se recurrirá a una colección de tipos de gráficos que aportan información de cada variable y de cómo pueden combinarse hasta 2, incluso 3, de las mismas.

II. Objetivos específicos

- Entender la finalidad de un análisis exploratorio de datos y su potencial.
- Aprender el conjunto de tipos de gráficos que existen y qué información aportan.
- Ver y apreciar el potencial de una EDA a través de casos concretos.
- Iniciarse en representación gráfica usando *ggplot*, una librería de representación en R muy potente y versátil.

III. Tipos de gráficos y sus usos

En el proceso de exploración de datos, la representación de los mismos es un paso que facilita su comprensión de una manera intuitiva, ya que revela ideas fundamentales y permite sacar conclusiones estructurales.

Vamos a describir los gráficos esenciales y qué información aportan.

3.1. Gráficos de difusión o scatterplots

Son gráficos que representan dos variables, sirven para analizar la relación que existe entre las mismas y sus distribuciones conjuntas.

En esta gráfica (imagen 2.1.), podemos observar la longitud de pétalo y sépalo en tres especies distintas de planta. Se puede ver que la *setosa* tiene menos en ambas variables y la *versicolor* y *virginica* se mezclan más.

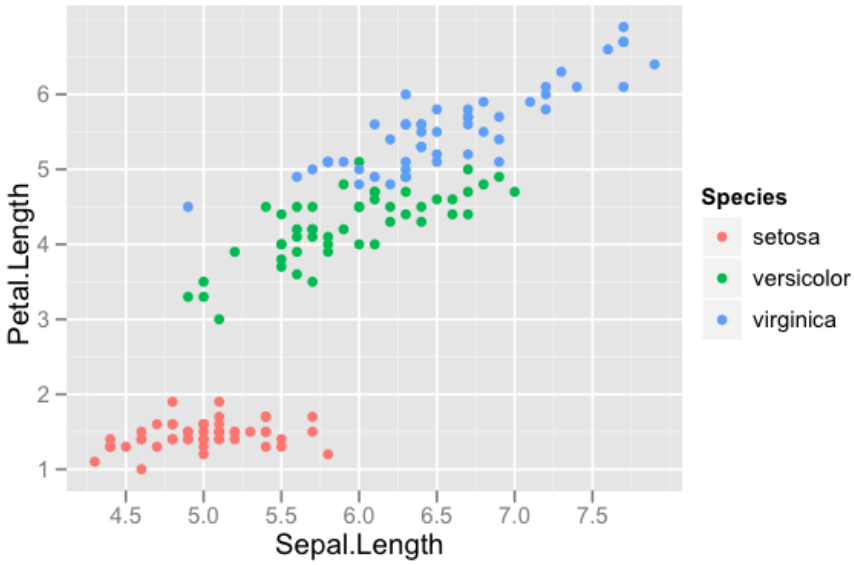

Imagen 2.1. Gráfica de dispersión. *Fuente:* elaboración propia en R.

3.2. Gráfico de barras

Agrupa los valores de una variable respecto a una serie de categorías. Por ejemplo, en el siguiente gráfico de barras observamos la cantidad de propina media que se da según el día de la semana. La línea que se superpone a la barra indica un *intervalo de confianza* de los valores que puede tomar. Parece que en domingo es ligeramente superior.

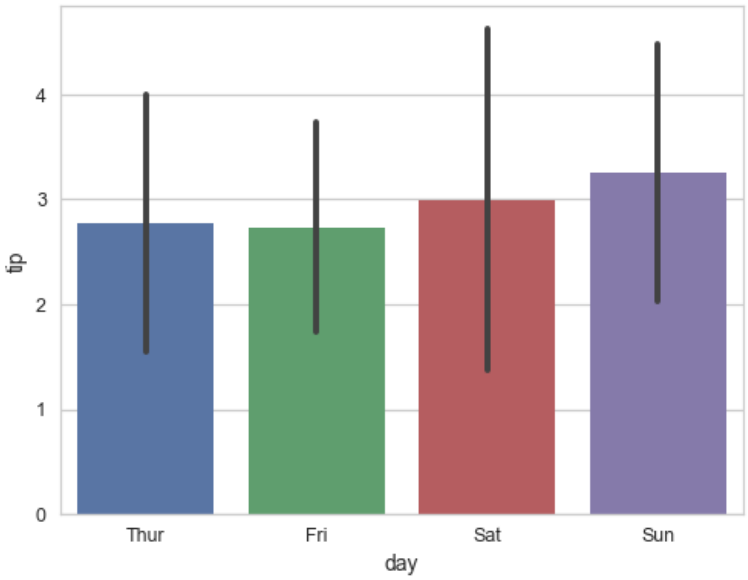

Imagen 2.2. Gráfico de barras. *Fuente:* www.seaborn.pydata.org.

3.3. Histograma

Es una gráfica que representa la acumulación de valores de una variable, indicando en qué zonas existe una mayor densidad de ocurrencias. Se disponen en rectángulos y pueden ser representados junto a una curva de densidad que completa la gráfica.

En la siguiente imagen, se observa que los valores de densidad de ozono en el aire son mayoritariamente de entre 0 y 50 partes por billón.

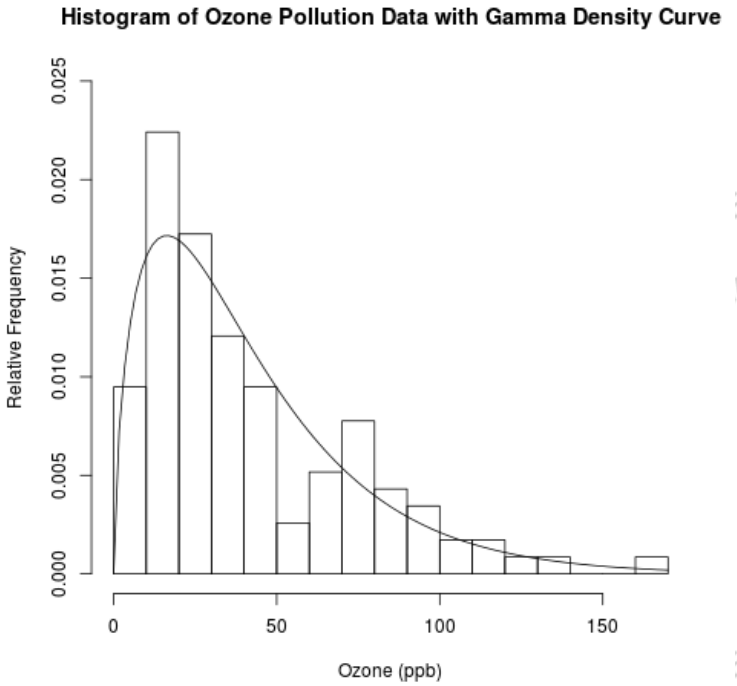

Imagen 2.3. Histograma. *Fuente:* elaboración propia con R.

En el siguiente histograma, que representa una variable respecto a dos categorías (hombre y mujer), observamos que el peso de los hombres tiene valores más altos que el de las mujeres.

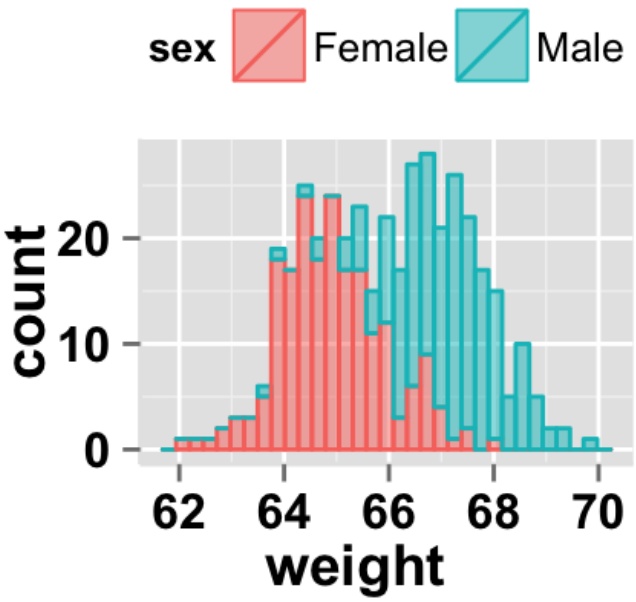

Imagen 2.4. Histograma. *Fuente:* elaboración propia con R.

3.4. Gráfico de caja o boxplot

Los gráficos de caja o bigote permiten visualizar dónde se encuentra el 50 % de los datos de una variable indicando el rango de valores central que queda delimitado por el rectángulo o caja. Además, generan un rango razonable donde cabe esperar que se encuentren las lecturas de la misma y marcan como *valores atípicos* los que se salen de ese rango.

En la siguiente gráfica, vemos los valores de factura totales en distintos días de la semana. Se indican con rombos los valores que están fuera del ámbito esperable. Observamos que los días con más facturas atípicas son los jueves.

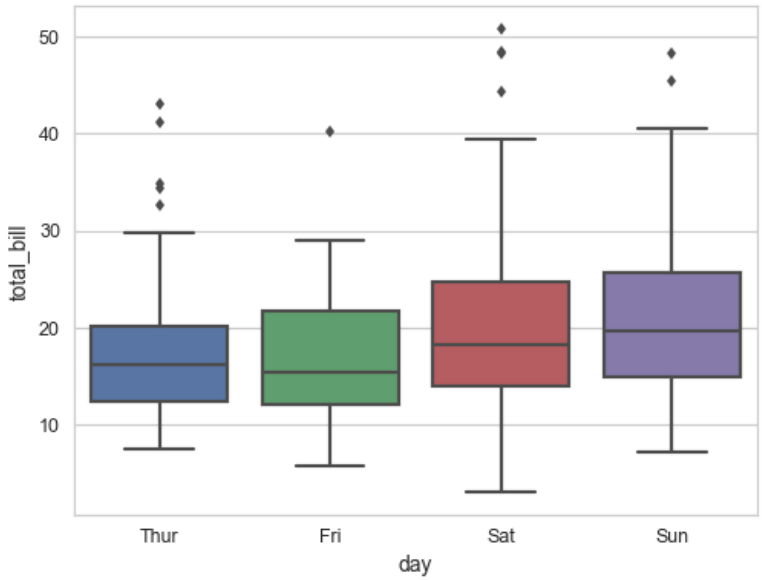

Imagen 2.5. Gráfica de caja. *Fuente:* www.seaborn.pydata.org

3.5. Gráfico de serie temporal

Una serie temporal es una cadena de valores tomados en distintos instantes. Representarlas es útil para analizar tendencias y puntos de cambio.

En la siguiente gráfica, podemos ver que lknd supera en un valor (desconocido) a ibm después de mediados de 2013; sin embargo, sufrió una caída posterior que los igualó.

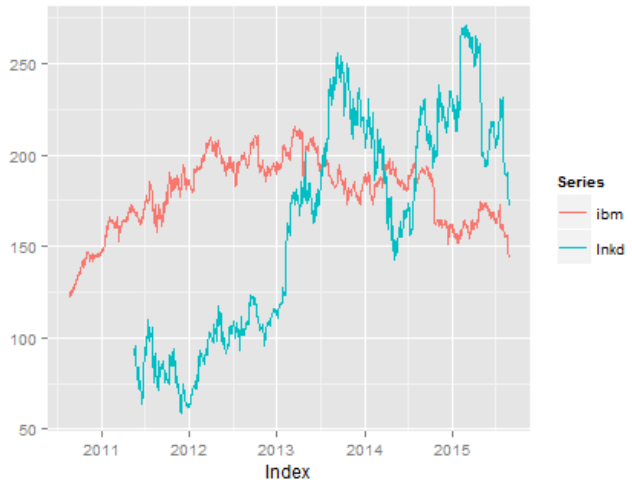

Imagen 2.6. Gráfico de series temporales. *Fuente:* elaboración propia con R.

3.6. Gráficos combinados

Existen gráficos más elaborados que combinan los tipos anteriores y otros, no tan básicos, que omitimos.

En el siguiente gráfico vemos un histograma por cada variable en la diagonal y gráficos de dispersión que relacionan las variables en el resto de las posiciones. Además, está coloreado según el tipo de flor que representa (el dataset es *Iris*, un dataset famoso en análisis de datos).

Se puede observar que la flor *setosa*, coloreada de azul, tiene una distribución de longitud y anchura de pétalo diferenciada de las otras dos, que se entremezclan.

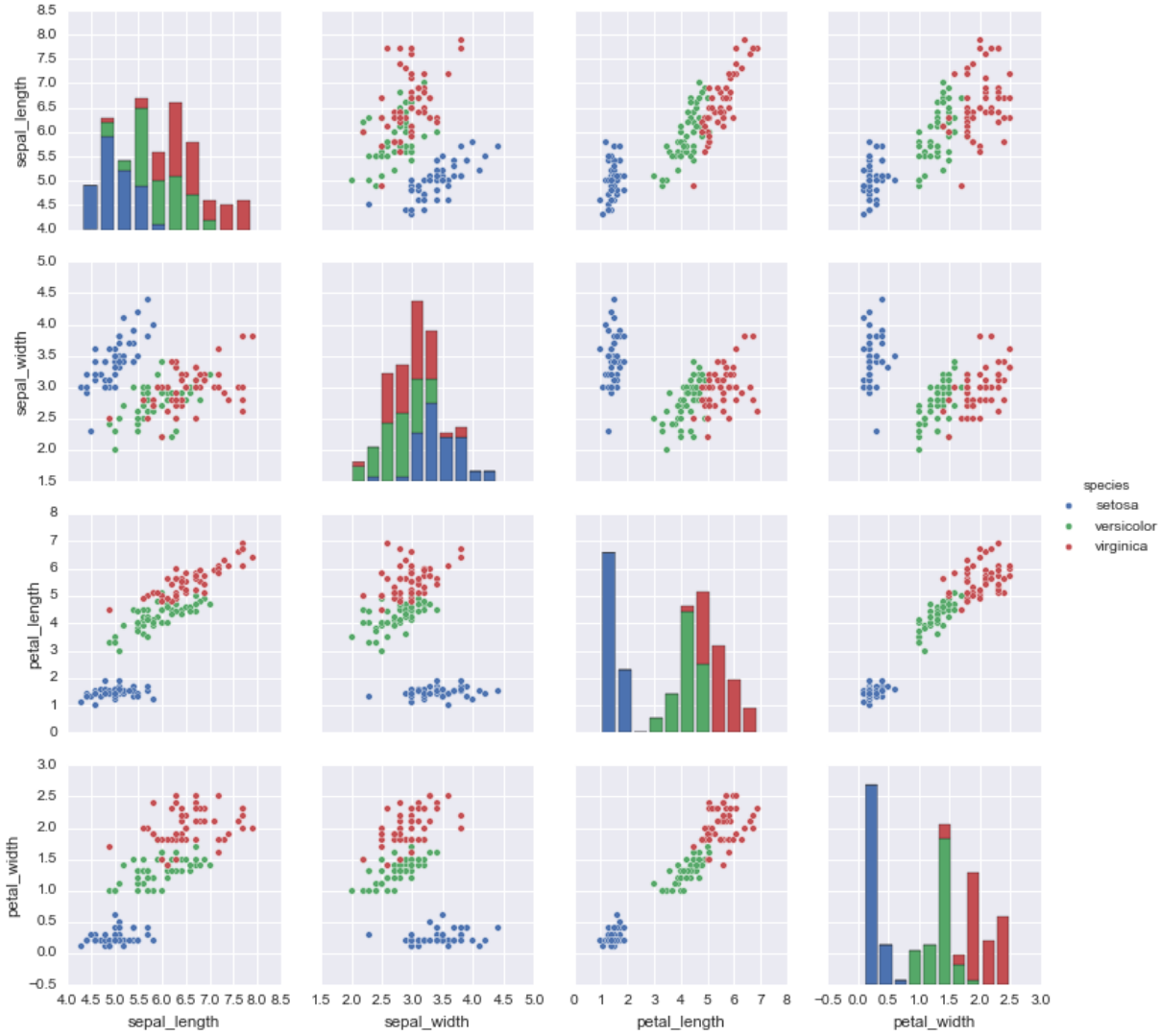

Imagen 2.7. Gráficos por pares. *Fuente:* www.seaborn.pydata.org

IV. Librería ggplot

En R existe un módulo muy potente para representación de datos denominado ggplot, de hecho, es tan bueno que el resto de lenguajes pretenden, en general, emular su sencillez de sintaxis y la calidad de sus gráficas.

En ggplot hay múltiples estilos y tipos de gráficos, incluyendo los ya vistos en la sección anterior.

Para más información de estilos se puede consultar la web de los desarrolladores de la librería ggplot.

Descarga: Consulta el notebook UD2 N01

Descárgate el archivo UD2 N01 y ejecútalo en R. También puedes verlo en.html

Descarga: Consulta el notebook UD2 N02

Descárgate el archivo UD2 N02 y ejecútalo en R. También puedes consultar la siguiente imagen y verlo en .html

V. Casos de análisis exploratorio de datos

Un análisis exploratorio de datos consiste en realizar una serie de representaciones, agrupaciones y sumarios descriptivos de un dataset, de modo que permita obtener intuiciones y contrastar convicciones sobre la información contenida.

Vamos a ver dos notebooks con ejemplos de exploración de datos:

- El primero está basado en un histórico de ventas de videojuegos de varias compañías en distintas plataformas desde el inicio del mercado de este producto. Se pretende obtener tendencias, deducir quién controla el mercado, puntos de crisis o cambio de consumo, y cuáles son las plataformas que dominan el mercado actual.
- El segundo analiza datos de la temperatura por mes en España desde 1750, aportando, a su vez, información de la incertidumbre de temperatura. En esta exploración pretendemos contrastar la convicción generalizada sobre el cambio climático.

Descarga: Consulta el notebook UD2 N03

Descárgate el archivo UD2 N03 en R y este csv. También puedes verlo en .html

A continuación, se facilitan los siguientes dos notebooks alternativos con más ejemplos de análisis que enriquecen el Notebook 3 de esta unidad:

- ANEXO_Ud2_vgsales2.Rmd
- ANEXO_Ud2_vgsales_analisis.Rmd

Descarga: Consulta el notebook UD2 N04

Descárgate el archivo UD2 N04 en R y este csv. También puedes verlo en .html

VI. Resumen final

Se han explicado las interacciones y las variables de los conjuntos de datos a través de las gráficas. Hay que ser conscientes de las limitaciones de la representación: solo recoge interacciones de variables dos a dos, como máximo, lo que significa que las dependencias más complejas, que escapan a una dimensión superior, no quedan reflejadas en una EDA; por lo tanto, es un método útil y potente para explorar, pero no es definitivo. Así pues, se considera una primera aproximación necesaria.

Por otra parte, la librería ggplot es relativamente compleja en su uso. Esto se debe a su enorme versatilidad.

En los casos de uso expuestos se ha analizado un dataset de ventas de empresas de videojuegos, así como la temperatura de la Tierra en los últimos tres siglos. Se invita al lector a investigar gráficas de desarrollo propio para obtener más convicciones de estos datasets.

Ejercicios

Caso práctico

Como repaso del tema y preparación para el Caso práctico final, se presenta el siguiente caso práctico. Consiste en explorar los datos de los supervivientes del accidente del Titanic para extraer patrones y verdades del famoso hito.

Descárgate el archivo ACTIVIDAD2_UD2 en R. También puedes verlo en.html

Cuando lo hayas realizado, puedes descargar su solución y comprobar tus resultados.

Solución

En los siguientes archivos dispones de la solución de la actividad propuesta:

- Solución en .html
- Solución en R.

Recursos

Enlaces de Interés

http://ggplot.yhathq.com/docs/index.html
http://ggplot.yhathq.com/docs/index.html
Librería ggplot

Glosario.

- **EDA**: Análisis exploratorio de datos. Consiste en una búsqueda de información visual de los datos realizando representaciones de los mismos.
- **Gráfico de barras**: Representación de una variable numérica continua respecto a varias categorías.
- **Gráfico de caja**: Representación de una variable numérica continua en una caja en la que se aglomeran la mayor parte de los datos, indicando los elementos que quedan fuera del rango como atípicos.
- **Gráfico de dispersión**: Representa en puntos los elementos del dataset a lo largo de dos ejes que indican dos variables numéricas elegidas. Se obtiene el dibujo de dependencia entre las mismas.
- **Histograma**: Representación de la distribución de una variable numérica continua.