Análisis exploratorio de datos

Minería de Datos - Grado en Matemáticas

Invalid Date

Exploratory Data Analysis (EDA)

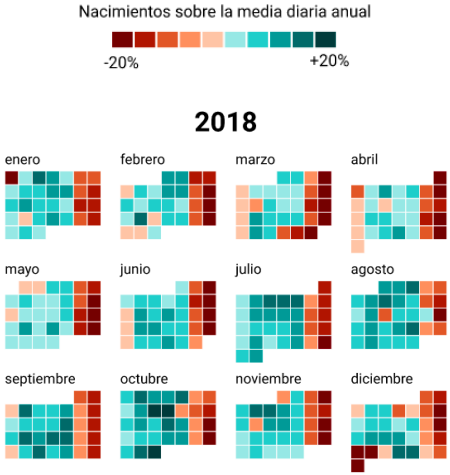
"Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone as the first step"

"Exploratory Data Analysis is detective work"

John Tukey

Importancia del EDA

Estudio natalidad: https://www.eldiario.es/nidos/no-ninos-nacen-toca-dar-luz-semana-21-probable-hacerlo-lunes-viernes_1_6400307.html



Introducción

- Primer análisis sobre cualquier conjunto de datos
- ¡Entender los datos!
- El análisis exploratorio de datos es un conjunto de técnicas que permiten resumir las características más importantes de un conjunto de datos, normalmente con especial énfasis en el uso de métodos de visualización gráfica
- No hay un guión estricto para realizar un EDA (¡somos detectives!)
- Fundamental adquirir conocimiento de los datos antes de usar un modelo de Aprendizaje Automático

Preguntas

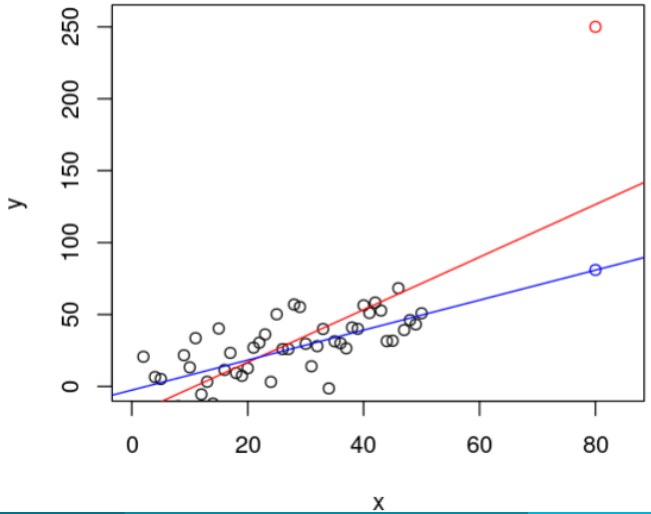
- Objetivo EDA: comprensión profunda de los datos
- ¿Cómo hacerlo? Planteando preguntas:
  - ¿Tamaño de los datos?
  - ¿Tipos de variables?
  - ¿Hay variable objetivo? ¿Cómo es?
  - ¿Hay errores?
  - ¿Hay variables irrelevantes?
  - ¿Están las variables relacionadas?
  - ¿Hay atípicos?
  - ¿Tengo suficiente capacidad de cómputo para procesar los datos?

Datos atípicos (outliers)

- Datos que no están en consonancia con el resto, que destaca por ser distinto del resto
- Causas:
  - Errores de medición (humano o del sistema). Ej: Peso de un paciente: 800 kg o un medidor manipulado
  - Contaminación: la muestra contiene datos de una población distinta a la de interés
  - Desviaciones naturales
- Solo se modifica el dato si es un error. Se busca el valor real y, si no es posible, se pone como missing

Datos atípicos

¿Diferencia entre el valor atípico rojo y el azul?



Datos faltantes (missing)

- Dato vacío, dato perdido, NA
- Causas:
  - Error en la medición, la transcripción
  - No se puede lograr el dato
- Acciones:
  - Trabajar únicamente con los datos sin valores faltantes (representan un % bajo del total de los datos)
  - Imputación de missing (media o mediana de la variable, el valor de los puntos más similares, predicción de un modelo de ML)
  - Agrupar los missings en una nueva categoría ¡fácilmente distinguible! Ej: 9999

Tipos de variables

- **Cualitativa** (también llamada categórica): refleja una cualidad de la realidad, su valor no se representa con un número. Pueden ser:
  - Dicotómicas (Ej: Sí o no) o politómicas (Ej: grupo sanguíneo)
  - Nominales (Ej: color de ojos) u ordinales (Ej: nota de un examen Suspenso-Aprobado-Notable-Sobresaliente)
- **Cuantitativa**: su valor se indica con un número, se corresponde con características que representan cantidades. Ej: distancia en km, nivel de colesterol, temperatura, etc. Pueden ser:
  - Discretas: Toma un número finito o infinito numerable de valores. Ej: números naturales o un recuento
  - Continuas: Puede tomar infinitos valores. Ej: altura, peso

Resumen numérico de variables cualitativas

- **Moda**: valor más frecuente de la distribución
- **Tabla de frecuencias** o **tabla de contingencia**: Muestra, para cada valor que tome una variable categórica, o para cada combinación de valores de dos o más variables categóricas, el número de casos que aparecen con dicho valor o combinación de valores

```
   Adelie Chinstrap    Gentoo
      152        68       124
```

```
   Adelie Chinstrap    Gentoo
0.4418605 0.1976744 0.3604651
```

Medidas de centralidad

- **Media**. Dada la variable $\mathbf{x} = (x_1, \dots, x_n)$ medida en n observaciones, su media es

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

  Muy afectada por valores atípicos/extremos

- **Mediana**: valor que ocupa la posición central de los datos, i.e., deja el 50% de los puntos a su izquierda (por debajo de él) y el otro 50% a la derecha (por encima de él). Sea $\mathbf{x}$ una variable con n observaciones, ordenados de menor a mayor, entonces:
  - Si n es impar, la mediana es justamente el valor que ocupa justamente la posición central $\lfloor n/2 \rfloor + 1$, $Med(\mathbf{x}) = x_{(\lfloor n/2 \rfloor + 1)}$
  - Si n es par, la mediana será la media de los dos valores centrales, esto es, $Med(\mathbf{x}) = \frac{x_{(n/2)} + x_{(n/2+1)}}{2}$

Medidas de posición

- Valores **mínimo** y **máximo** de la variable: x_{min}, x_{max}
- Primer y tercer **cuartil**: Los valores que dejan por debajo un $p\%$ de los datos, siendo $p = 25\%$ en el caso del primer cuartil (Q_1) y $p = 75\%$ en el caso del tercer cuartil (Q_3). El segundo cuartil es la mediana.
- **Deciles**: Mismo concepto que los cuartiles pero de 10 en 10

Resumen numérico de variables cuantitativas

Medidas de dispersión: ¿Cómo varían los datos en torno a los valores centrales?

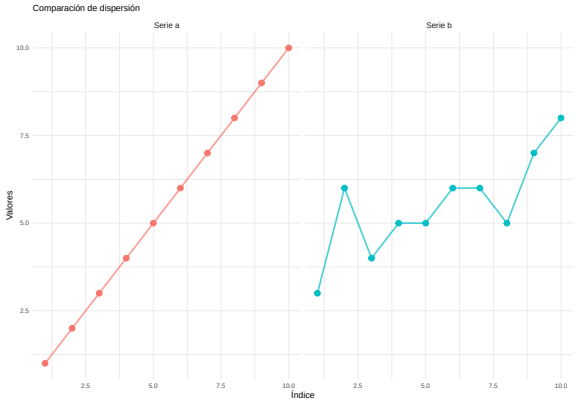
```r
a <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
b <- c(3, 6, 4, 5, 5, 6, 6, 5, 7, 8)
```

Serie a:

- Media: 5.5
- Desviación típica: 3.03

Serie b:

- Media: 5.5
- Desviación típica: 1.43



Ambas tienen la misma media pero diferente variabilidad

Medidas de dispersión

- **Rango** o **recorrido**: $Rango = x_{max} - x_{min}$
- **Varianza**: Mide la dispersión de los valores de la variable respecto a la media
  - Varianza **muestral**: $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$
  - Varianza **poblacional**: $\sigma^2 = \frac{1}{N} \sum_{i=1}^{n} (x_i - \mu)^2$, siendo N el tamaño de la población y μ su media
- **Desviación típica**: raíz cuadrada de la varianza muestral o poblacional.
  - Desviación típica **muestral**: $s = \sqrt{s^2}$
  - Desviación típica **poblacional**: $\sigma = \sqrt{\sigma^2}$

Interpretación más sencilla al medir la dispersión en las mismas unidades que la variable

Medidas de dispersión

- **Rango intercuartílico**: diferencia entre el tercer y el primer cuartil $IQR = Q_3 - Q_1$
- **Coeficiente de variación**: representa la desviación típica en unidades de la media $CV = s/\bar{x}$. Se suele expresar en porcentaje. Por ejemplo, CV = 60% indica que el valor de la desviación típica es 0.6 veces la magnitud de la media.

Diagrama de barras

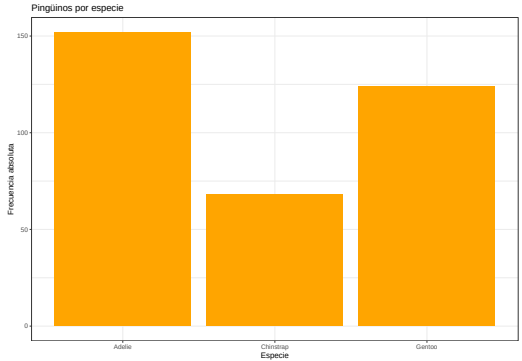
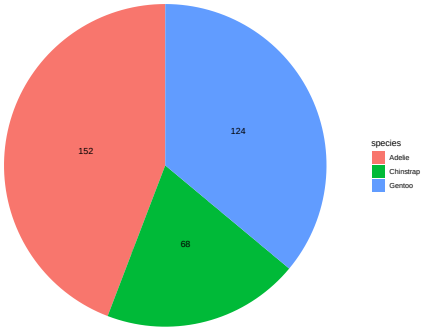


Gráfico de tarta



Problema: ojo humano tiene problemas para percibir correctamente diferencias en sectores angulares

Gráficos de gofre (*waffle*)

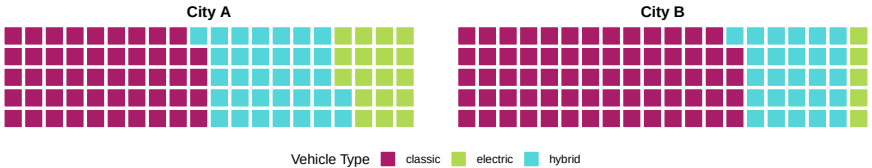


Diagrama de cajas y bigotes (boxplot)

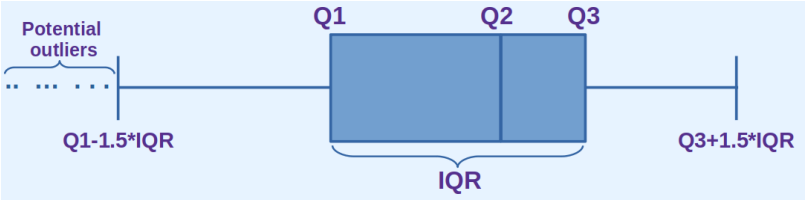
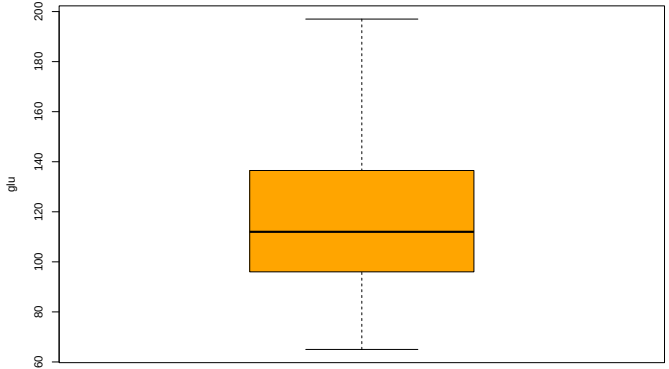


Diagrama de cajas y bigotes (boxplot)



Histograma

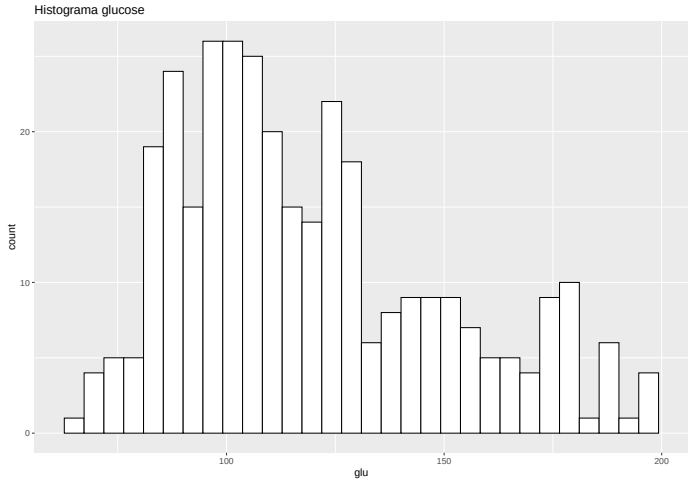
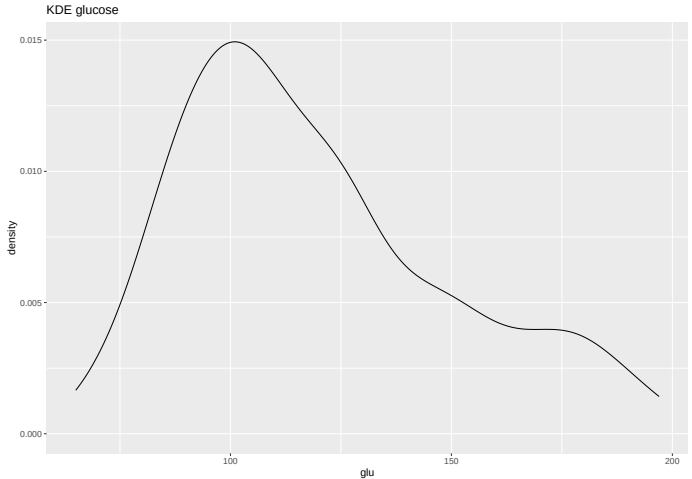


Gráfico de densidad



Histograma + gráfico de densidad

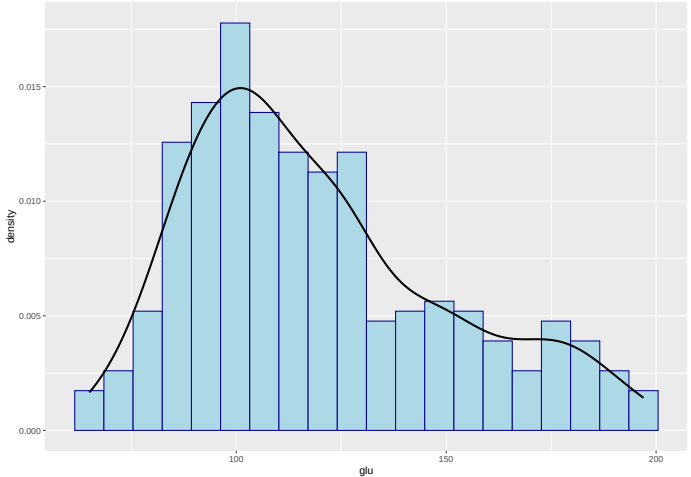


Diagrama de barras

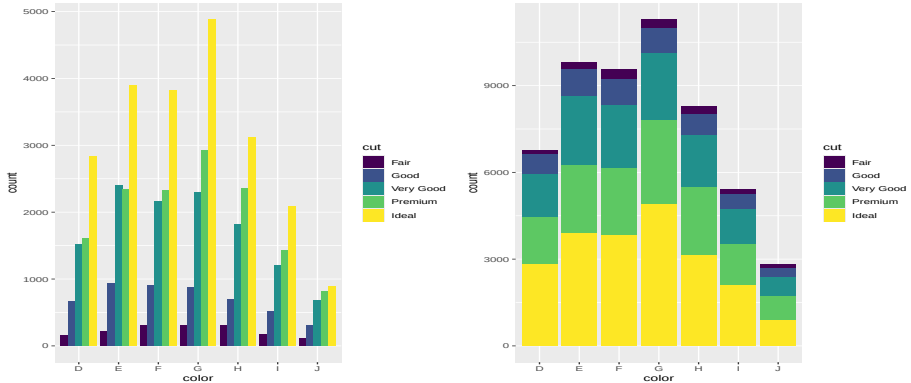
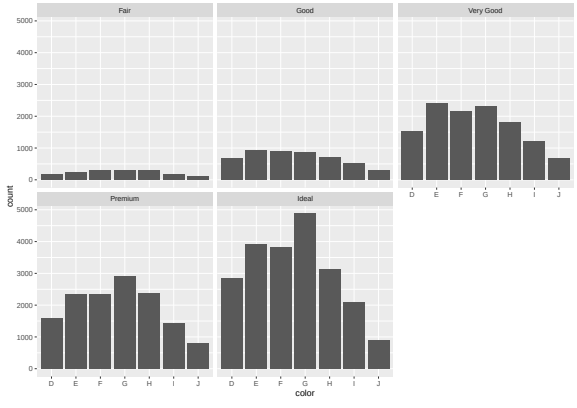
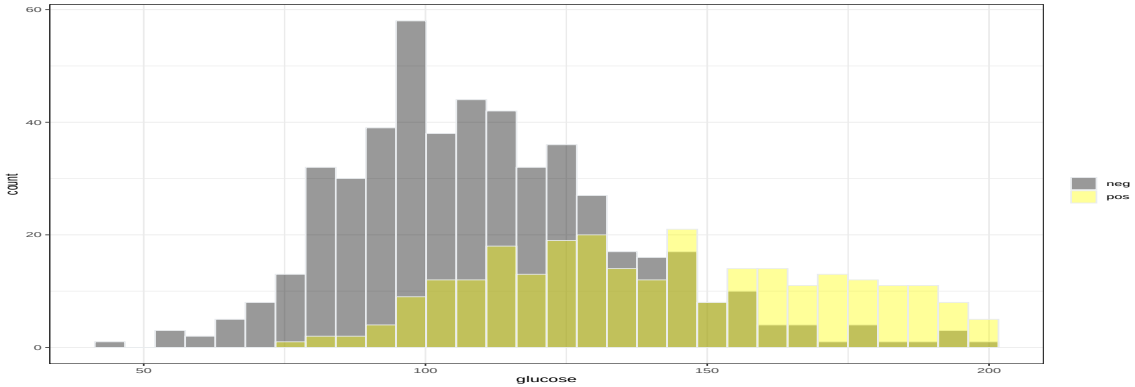


Diagrama de barras

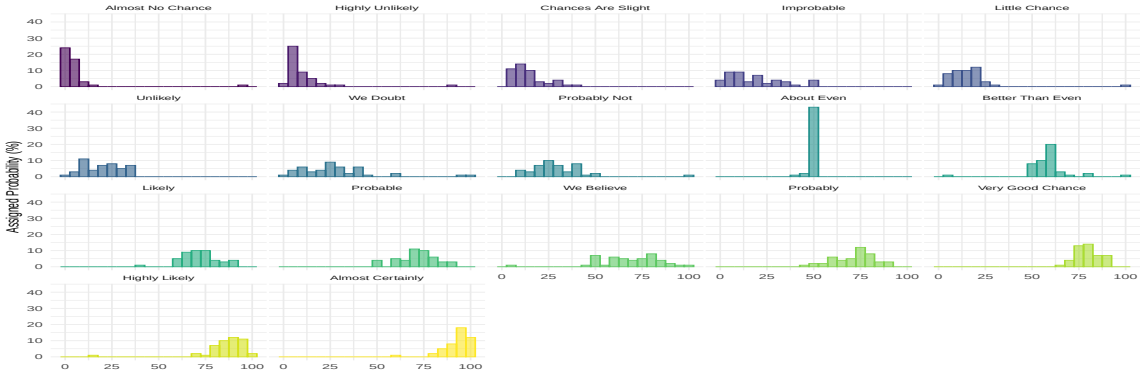


Histogramas conjuntos

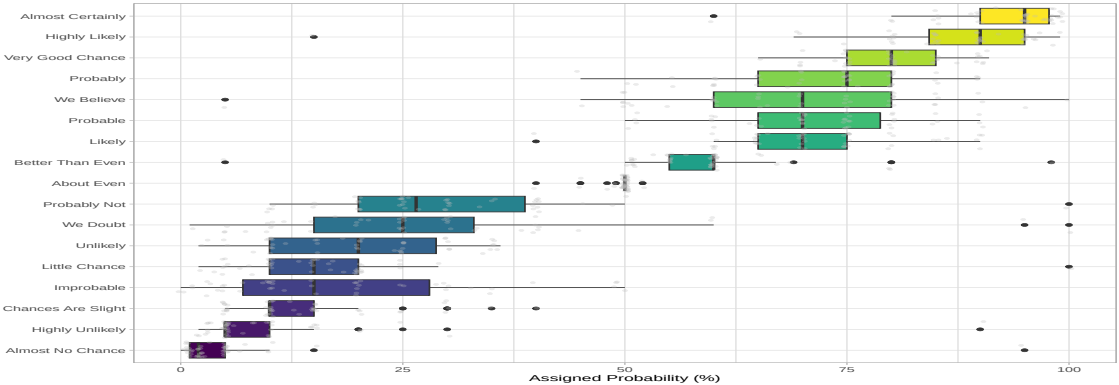


Histogramas conjuntos

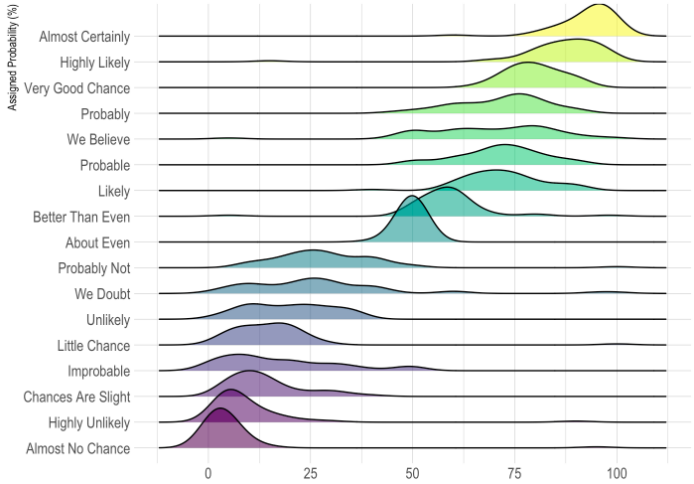
Origen del gráfico: From Data to Viz



Origen del gráfico: From Data to Viz



Origen del gráfico: From Data to Viz



Corrplot

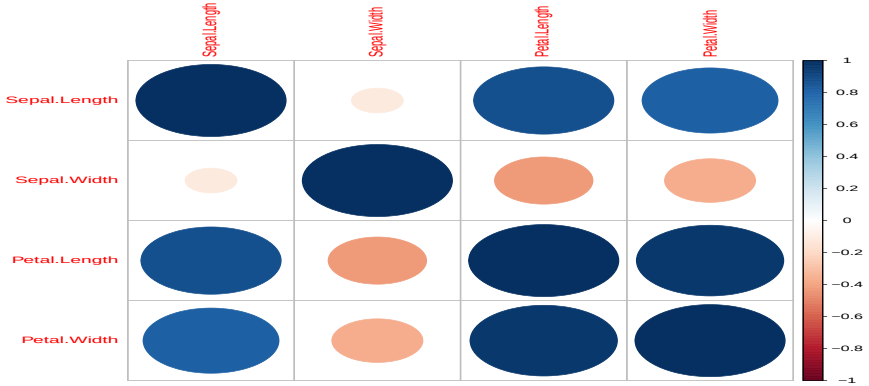

Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

Diagrama de dispersión

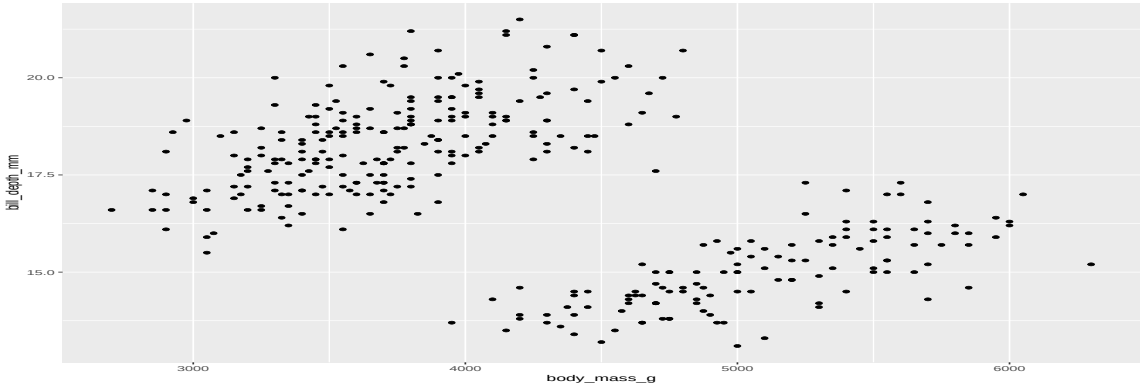


Diagrama de dispersión

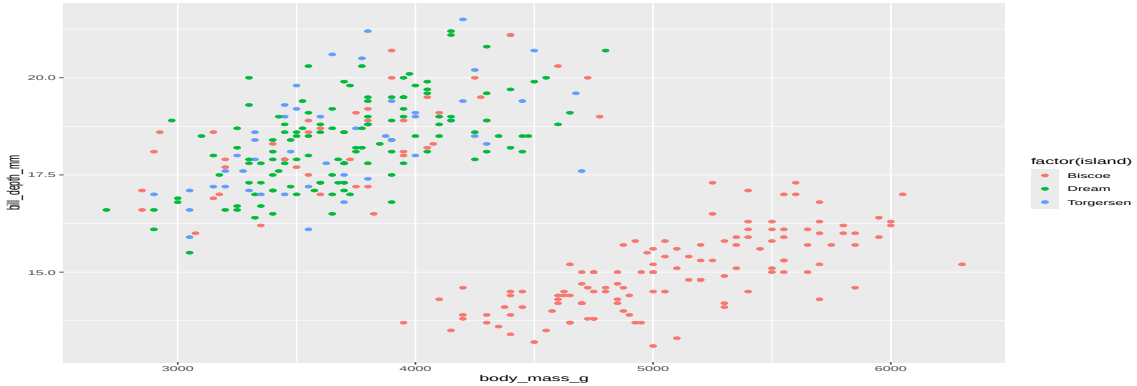


Diagrama de dispersión

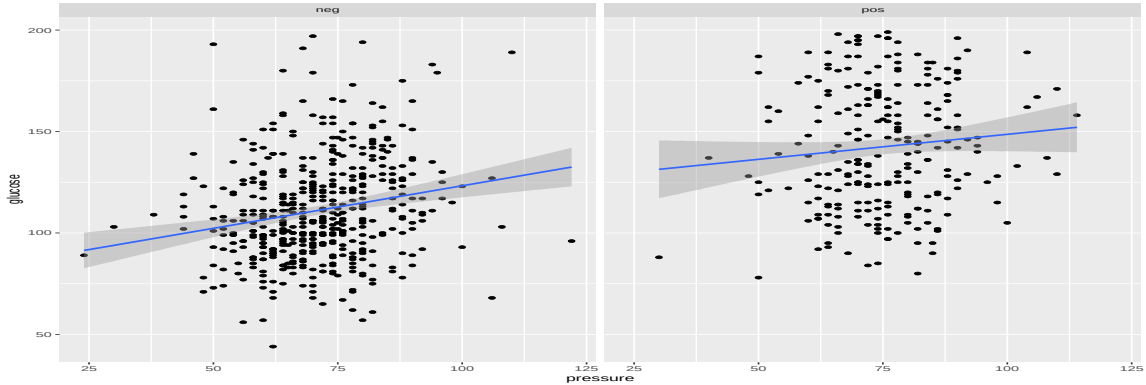


Diagrama de puntos (dotplot): categórica vs continua

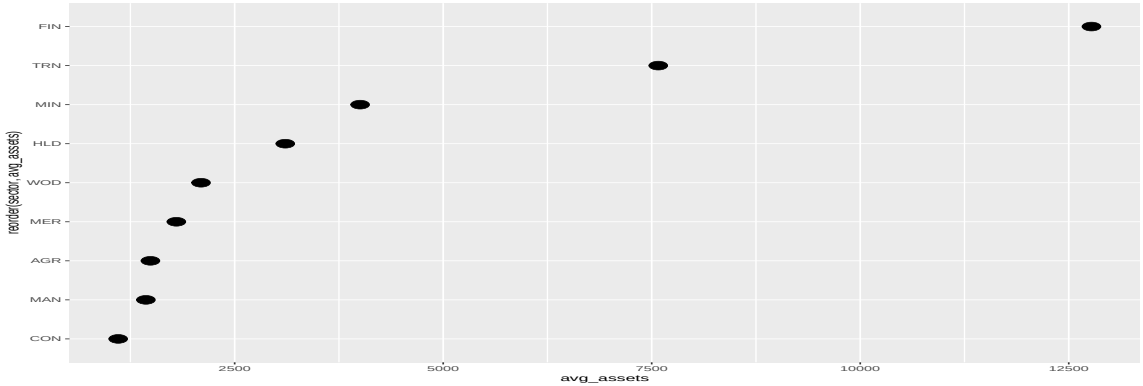
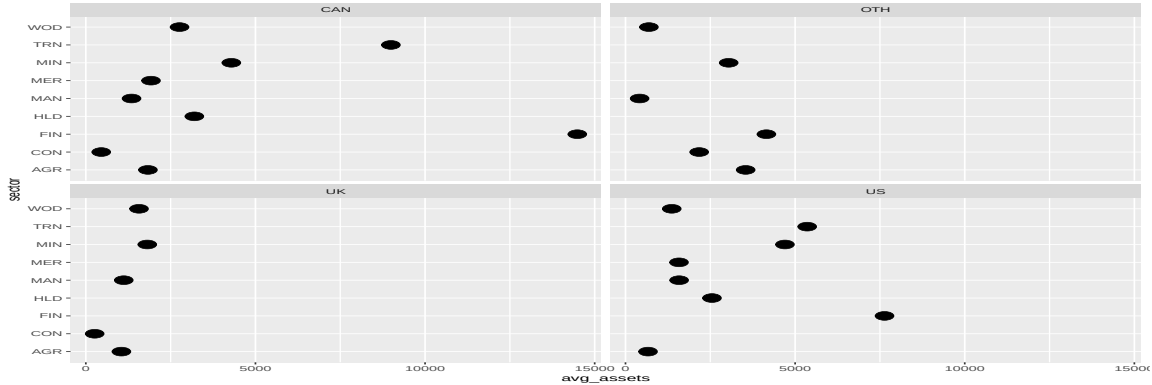
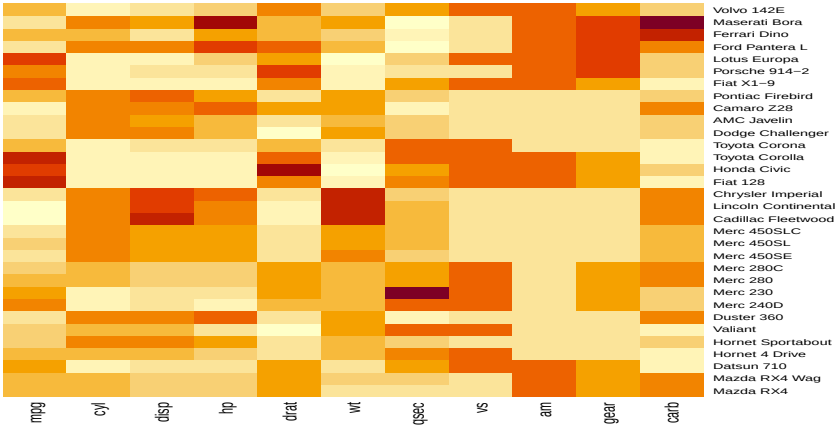


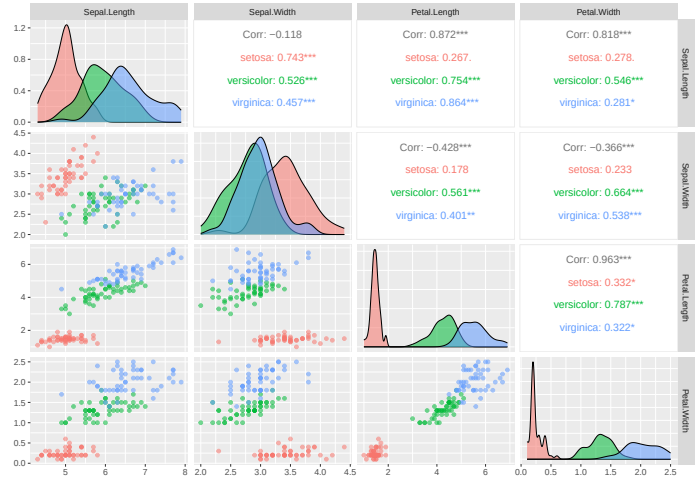
Diagrama de dispersión por categorías



Heatmap (mapa de calor)



https://r-charts.com/es/correlacion/ggpairs/

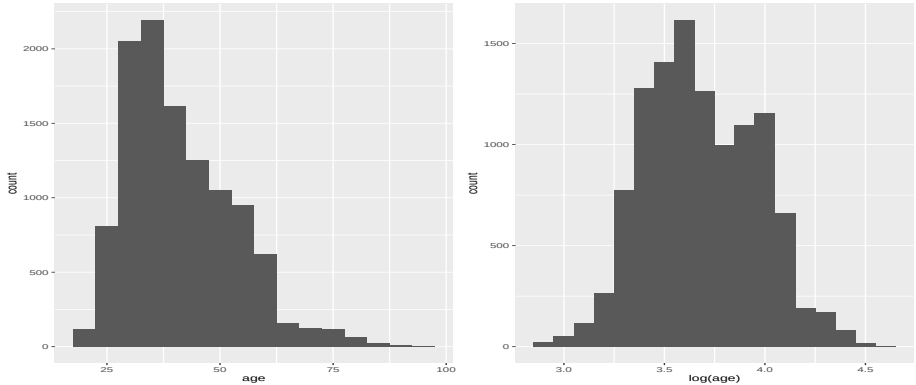


Transformación de variables

- Distribución adecuada de la variable (Ej: distribución Normal)
- Relación con otras variables y visualización
- Igualar dispersión entre variables –> variables en escalas comparables
- Variables cuantitativas –> variables categóricas

Transformación de variables

- Distribución adecuada de la variable (Ej: distribución Normal)



Transformación de variables

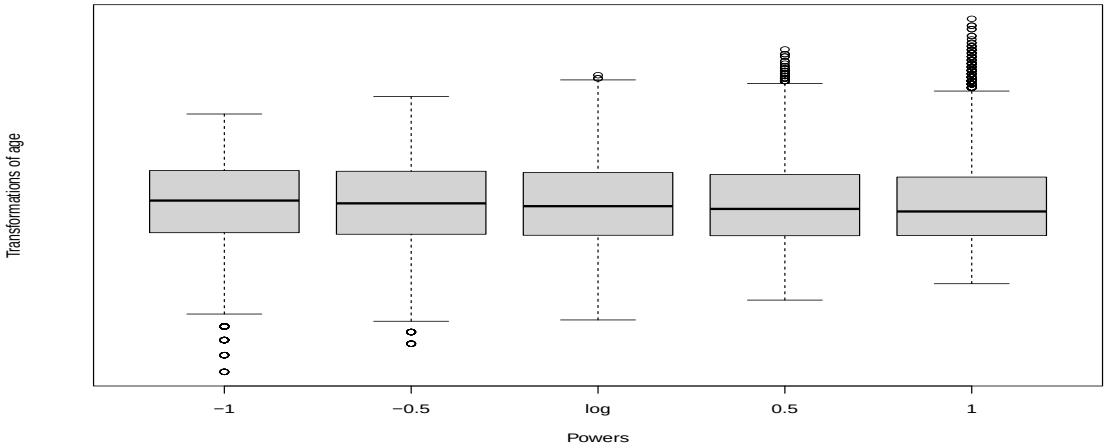
Obtener una variable cuya distribución de valores sea:

- Más simétrica y con menor dispersión que la original
- Más semejante a una distribución normal (e.g. para algunos modelos lineales)
- Restringida en un intervalo de valores (e.g. [0,1])

Transformaciones de escala-potencia o transformaciones Box-Cox:

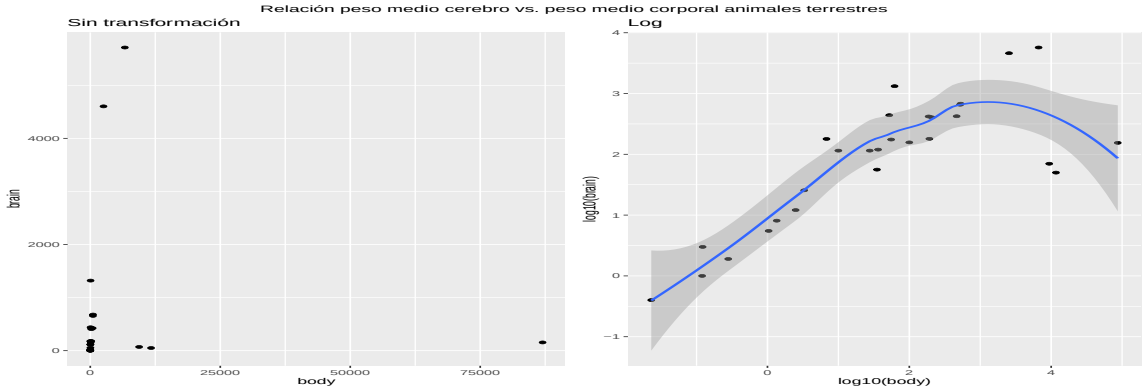
$$x(\lambda) = \begin{cases} \frac{x^{\lambda}-1}{\lambda}, & \text{cuando } \lambda \neq 0, \\ log_e(x), & \text{cuando } \lambda = 0 \end{cases}$$

Transformación de variables



Transformación de variables

- Relación con otras variables y visualización



Transformación de variables

Igualar dispersión entre variables –> variables en escalas comparables

- **Reescalado o cambio de escala**: Sumar o restar una constante a un vector, y luego multiplicar o dividir por una constante. Por ejemplo, para transformar la unidad de medida de una variable (grados Farenheit $\rightarrow$ grados Celsius).
- **Normalización**: Dividir por la norma de un vector, por ejemplo para hacer su distancia euclídea igual a 1.
- **Estandarización**: Consiste en restar a un vector una medida de localización o nivel (e.g. media, mediana) y dividir por una medida de escala (dispersión). Sea X una variable aleatoria con media $\bar{x}$ y desviación típica s:

  Estandarización $\rightarrow Y = \frac{X - \overline{x}}{s}$

  Distribución con media 0 y desviación típica 1

  Escalado$min - max \rightarrow Y = \frac{X - min_x}{max_x - min_x}$

Transformación de variables

Variables cuantitativas –> variables categóricas:

- Agrupación de datos numéricos. Ej: Edad -> “menos de 18 años”, “18-30 años”, “31-45 años”, “46-60 años”, “mayor de 60”
- Creación de variables binarias. Ej: clientes satisfechos / insatisfechos

Más y más

- Series temporales
- Mapas
- Gráficos específicos de clustering
- Pirámides de población
- QQplot
- etc

Más y más

https://elartedeldato.com/

https://rkabacoff.github.io/datavis/ Modern Data Visualization with R

https://r-graph-gallery.com/ggplot2-package.html

https://r-graph-gallery.com/

https://www.data-to-viz.com/

Bibliografía de utilidad para el lector

- "Fundamentos de ciencia de datos con R" coordinado por Gema Fernández-Avilés y José-María Montero: https://cdr-book.github.io/
- Weiss, N. A., & Weiss, C. A. (2017). *Introductory statistics*. London: Pearson.
- "Estadística Aplicada a las Ciencias y la Ingeniería" escrito por Emilio L. Cano. https://emilopezcano.github.io/estadistica-ciencias-ingenieria/index.html
- R for Data Science: https://r4ds.hadley.nz/eda
  - Primera versión en castellano: https://es.r4ds.hadley.nz/