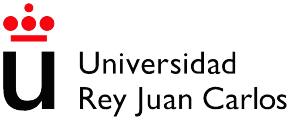
Análisis Exploratorio de Datos

Víctor Aceña - Isaac Martín - Carmen Lancho

DSLAB

2025-09-16



Estadística descriptiva

- Se ocupa de resumir y describir las características de un conjunto de datos mediante herramientas gráficas y numéricas, como tablas, gráficos, medias, medianas, varianzas, etc.
- Su objetivo es proporcionar una visión clara y comprensible de la estructura y características de los datos
- Ejemplo: Tiempo medio que tardan los alumnos de la asignatura de Inferencia Estadística del Grado en Ciencia e Ingeniería de Datos en llegar a la universidad

Datos

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
0	Adelie	Torgersen	39.1	18.7	181	3750	male
1	Adelie	Torgersen	39.5	17.4	186	3800	female
2	Adelie	Torgersen	40.3	18.0	195	3250	female
3	Adelie	Torgersen	36.7	19.3	193	3450	female
4	Adelie	Torgersen	39.3	20.6	190	3650	male
...	...	...	...	...	...	...	...
328	Chinstrap	Dream	55.8	19.8	207	4000	male
329	Chinstrap	Dream	43.5	18.1	202	3400	female
330	Chinstrap	Dream	49.6	18.2	193	3775	male
331	Chinstrap	Dream	50.8	19.0	210	4100	male
332	Chinstrap	Dream	50.2	18.7	198	3775	female

333 rows × 7 columns

Tipos de variables

- **Cualitativas o categóricas**: Describen cualidades. Se dividen en:
  - **Nominales**. Sin orden específico. Ej: color de ojos
  - **Ordinales**. Tienen un orden. Ej: niveles de satisfacción
- **Cuantitativas**: Valores numéricos que se pueden medir. Pueden ser:
  - **Discretas**. Valores contables, como el número de hijos
  - **Continuas**. Pueden tomar cualquier valor dentro de un rango, como la altura o el peso.
- **Marcas de tiempo o identificadores**: Como por ejemplo la fecha y hora de una transacción o el código de un producto o el número de identidad.

Escalas de medición

Una escala de medición define cómo se cuantifican o categorizan las variables recogidas sobre un conjunto de datos, influyendo en el análisis estadístico aplicable:

- **Nominal**: categorización sin orden inherente. Por ejemplo, el género, la nacionalidad o el tipo de sangre
- **Ordinal**: categorización con un orden lógico. Por ejemplo, el nivel educativo, o una clasificación de hoteles
- **Métrica**:
  - **Intervalo**: sin cero verdadero, por ejemplo la temperatura en Celsius.
  - **Razón**: con cero verdadero, por ejemplo los ingresos o la distancia.

Conversión entre escalas de medición

- Fechas a categóricas: convertir fechas exactas en mes, día de la semana, etc.
- Cuantitativas a cualitativas: crear clases o rangos a partir de datos numéricos. Por ejemplo convertir el nivel de ingresos en “bajo”, “medio” y “alto”.
- Variables calculadas: creación de nuevas variables a partir de las existentes. Por ejemplo, se crea el Índice de Masa Corporal (IMC) a partir de peso y altura.

Estadística descriptiva

- Aplicación de técnicas matemáticas para resumir un conjunto de datos
- Objetivo: Presentar los datos de manera clara mediante medidas de tendencia central (media, mediana, moda), medidas de dispersión (desviación estándar, varianza), y visualizaciones (tablas de frecuencia, gráficos de barras, histogramas, etc.).
- Sigue un enfoque más formal y estructurado, centrado en describir las características principales de un conjunto de datos
- Limitaciones: Se centra en los datos de manera resumida, sin necesariamente buscar patrones complejos, relaciones o anomalías.

Análisis exploratorio de datos (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

Análisis exploratorio de datos

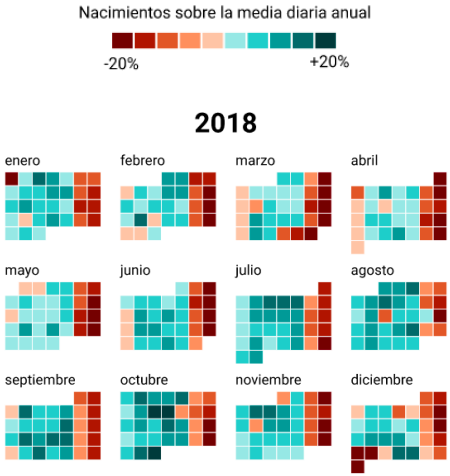
- Enfoque más amplio y flexible que combina herramientas matemáticas y estadísticas son técnicas gráficas para descubrir patrones, tendencias y relaciones en los datos
- Objetivo: Busca explora y entender los datos de manera más profunda e interactiva, generando hipótesis y obteniendo información antes de aplicar técnicas de modelado más formales
- Sigue un enfoque más experimental y visual, permitiendo descubrir patrones inesperados o relaciones ocultas. No hay un guión estricto para realizar un EDA (¡somos detectives!)
- Limitaciones: Para sacar conclusiones generales (no limitadas a la muestra) debe ser seguida por análisis más formales o inferenciales

Importancia del EDA

Supongamos que tenemos los datos de natalidad desde 1977 hasta 2018. ¿Qué haríais con ellos?

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

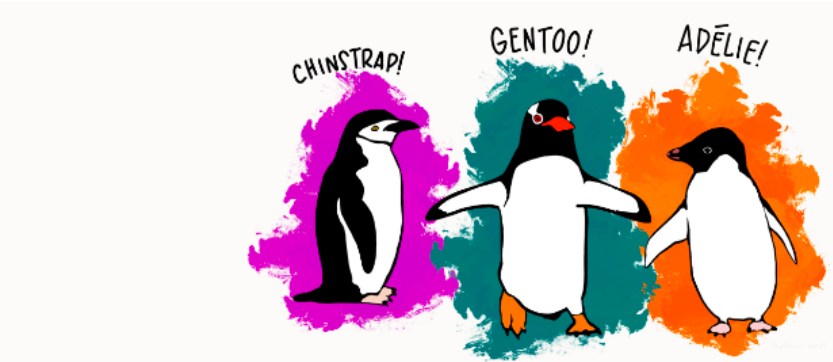


Estadística descriptiva

- Estudiemos algunas de las herramientas de la estadística descriptiva y el EDA
- Herramientas
  - Resúmenes numéricos: media, moda, mediana, cuantiles, tablas de frecuencia, etc
  - Métodos gráficos. diagramas de barras, histograma, boxplot, etc.
- En función de los datos (categóricos o continuos), se usarán unos métodos u otros

Dataset Palmer penguin

https://allisonhorst.github.io/palmerpenguins/articles/intro.html



The Palmer Archipelago penguins. Artwork by @allison_horst.

Dataset Palmer penguin

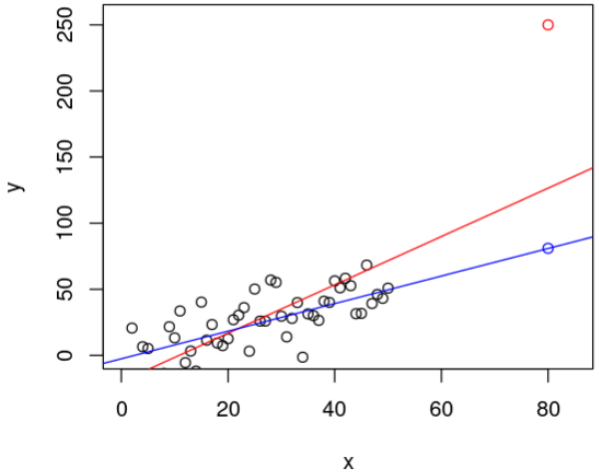
```
Rows: 344
Columns: 8
$ species           <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Adel~
$ island            <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torgerse~
$ bill_length_mm    <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34.1, ~
$ bill_depth_mm     <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 18.1, ~
$ flipper_length_mm <int> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190, 186~
$ body_mass_g       <int> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675, 3475, ~
$ sex               <fct> male, female, female, NA, female, male, female, male~
$ year              <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007~
```

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

Tabla de frecuencias

- **Tabla de frecuencias** o **tabla de contingencia**: Muestra el número de casos que aparecen para cada valor de una variable categórica o combinación de valores de dos o más variables categóricas

```
table(penguins$species)
```

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

```
prop.table(table(penguins$species))
```

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

Tabla de frecuencias

Dada una variable X con dominio $\{X_1, \dots, X_k\}$ que ha sido medida en una muestra de tamaño n y denotemos por n_i el número de elementos en la muestra que toman el valor X_i. La tabla de frecuencias correspondiente es:

Valores de X	Frec. absoluta	Frec. relativa	Frec. acumulada	Frec. relativa acumulada
X_1	n_1	$fr_1 = \frac{n_1}{n}$	$F_1 = n_1$	$Fr_1 = \frac{F_1}{n}$
X_2	n_2	$fr_2 = \frac{n_2}{n}$	$F_2 = F_1 + n_2$	$Fr_2 = \frac{F_2}{n}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_k	n_k	$fr_k = \frac{n_k}{n}$	$F_k = F_{k-1} + n_k$	$Fr_k = \frac{F_k}{n}$

Tabla de frecuencias

¿Para qué tipos de variable sirve?

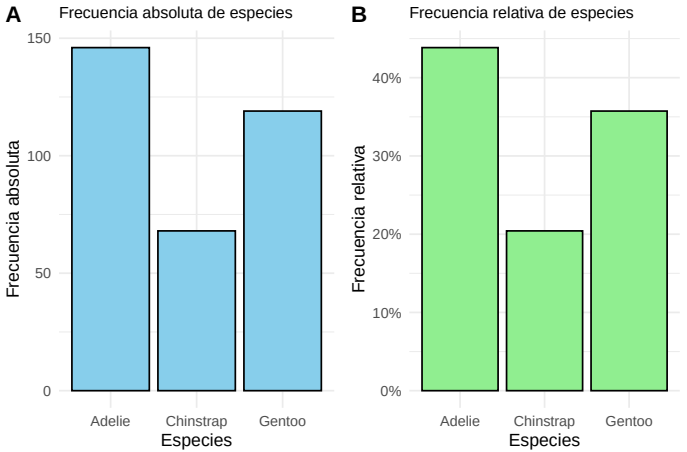
Tabla de frecuencias

¿Para qué tipos de variable sirve?

- Categóricas nominales. Color de ojos.
- Categóricas ordinales. Grado de satisfacción
- Continuas. Altura.
  - Dividimos en intervalos: (0,140],(140,155],...,(210,225]

Diagrama de barras

- Para variables cualitativas



No hay diferencia entre frecuencias relativas o absolutas en este caso

Diagrama de barras

¿Y si quisiéramos comparar a nuestros pingüinos con un grupo de otra isla?

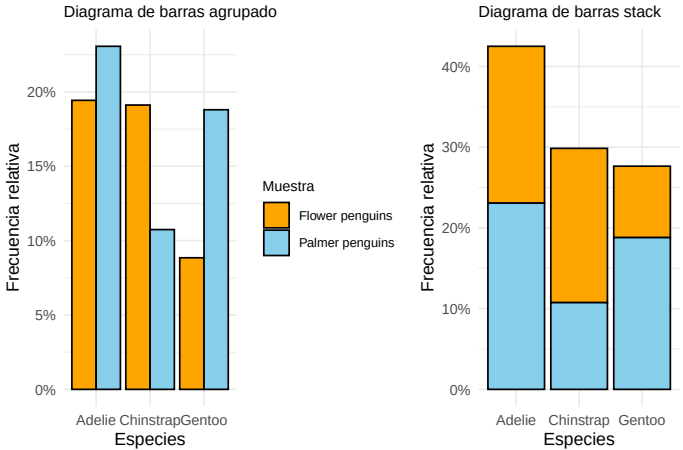


Tabla de frecuencias (contingencia)

	female	male
Adelie	73	73
Chinstrap	34	34
Gentoo	58	61

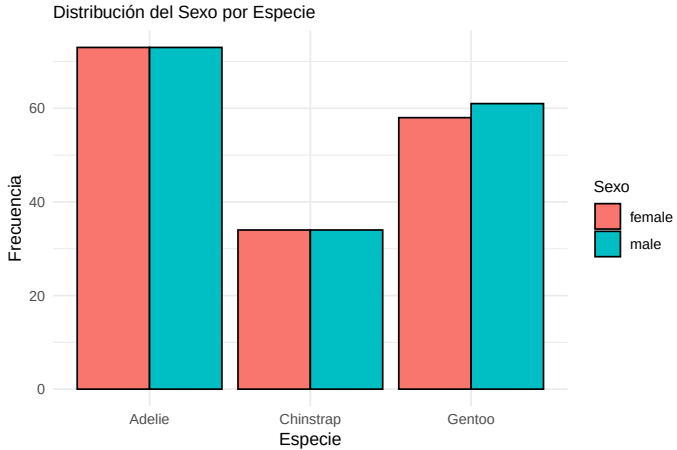
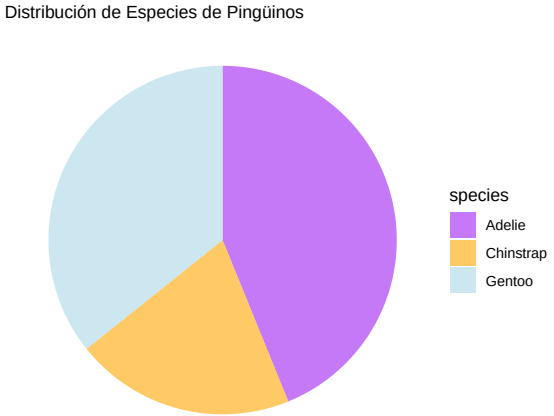


Gráfico de tarta

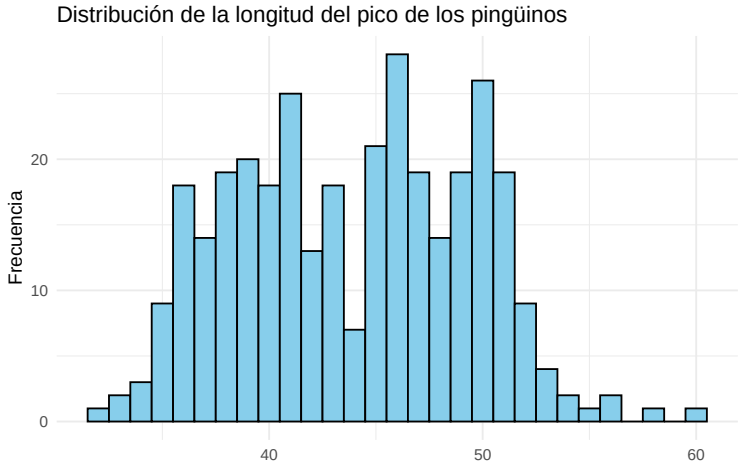
- Para variables cualitativas
- También llamado diagrama de sectores



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

Histograma

- Para variables cuantitativas
- Refleja la distribución de los datos



Resumen numérico de variables

- Medidas de centralidad
- Medidas de posición
- Medidas de dispersión

Resumen numérico de variables

Medidas de centralidad

- **Moda**: valor más frecuente de la distribución
- **Media**. Dada una muestra de n observaciones $\mathbf{x} = (x_1, \dots, x_n)$ de la variable X su media es

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

  Muy afectada por valores atípicos/extremos

Medidas de centralidad

- **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 muestra 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, $Med(\mathbf{x}) = \frac{x_{(n/2)} + x_{(n/2+1)}}{2}$

  Medida robusta

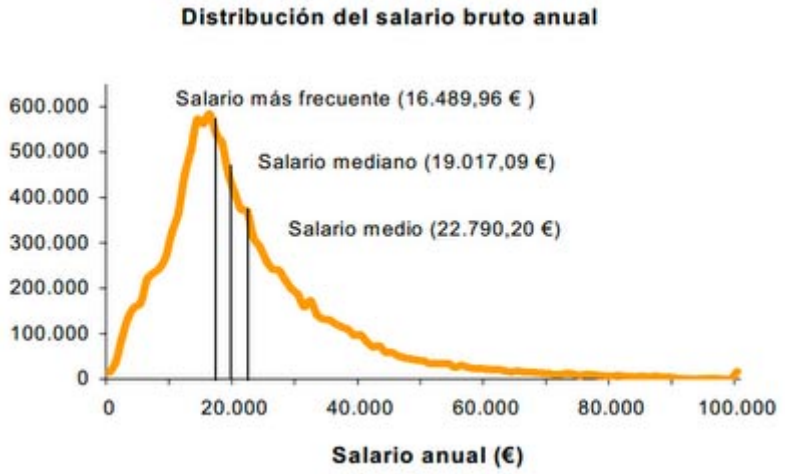


Figure 1: Microsiervos

Resumen numérico de variables

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

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

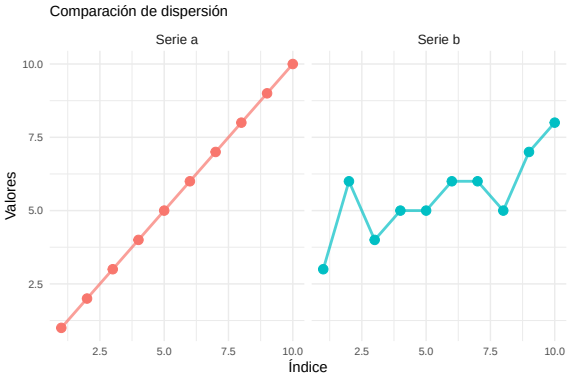
```r
a <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
b <- c(3, 6, 4, 5, 5, 6, 6, 5, 7, 8)
```

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

Resumen numérico de variables cuantitativas

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 cajas y bigotes (boxplot)

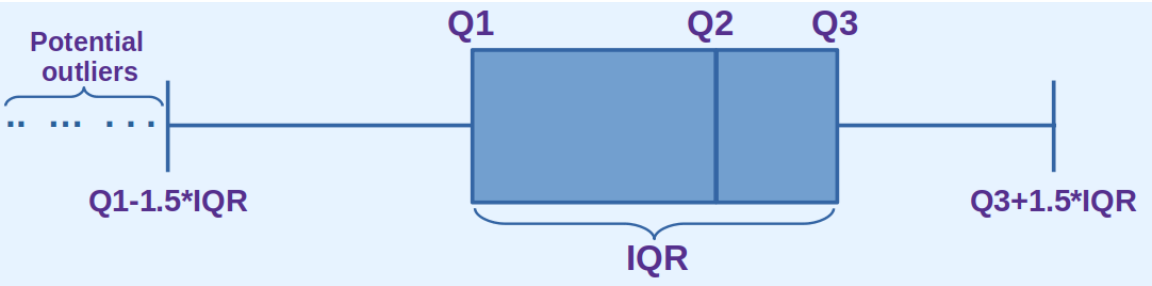
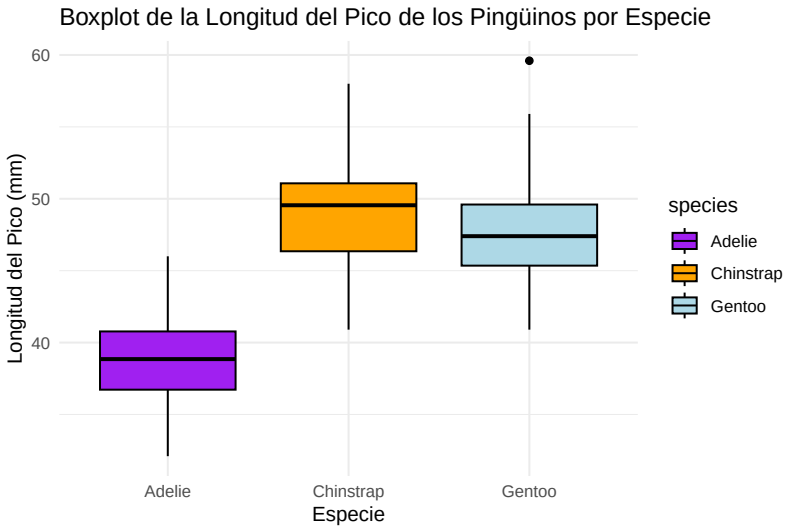


Diagrama de cajas y bigotes (boxplot)



Diagramas de dispersión

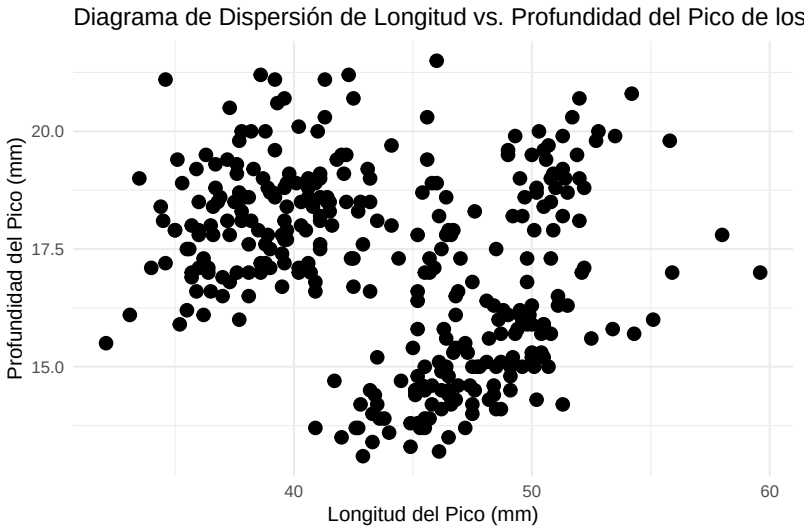
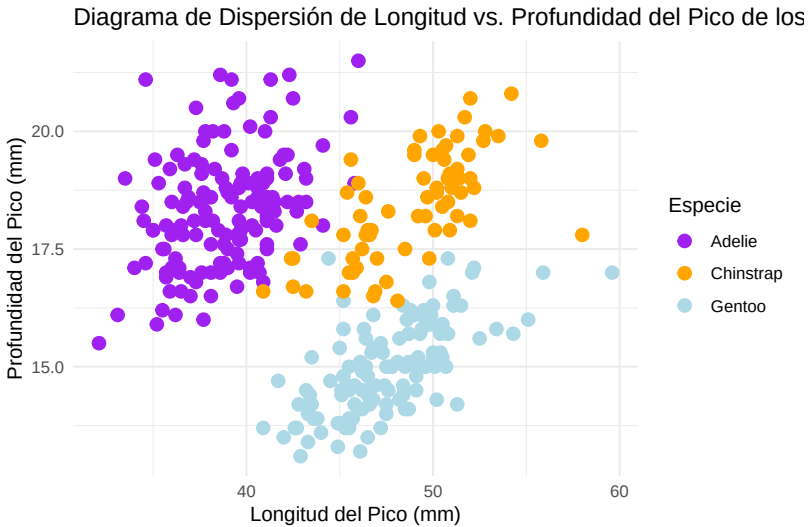


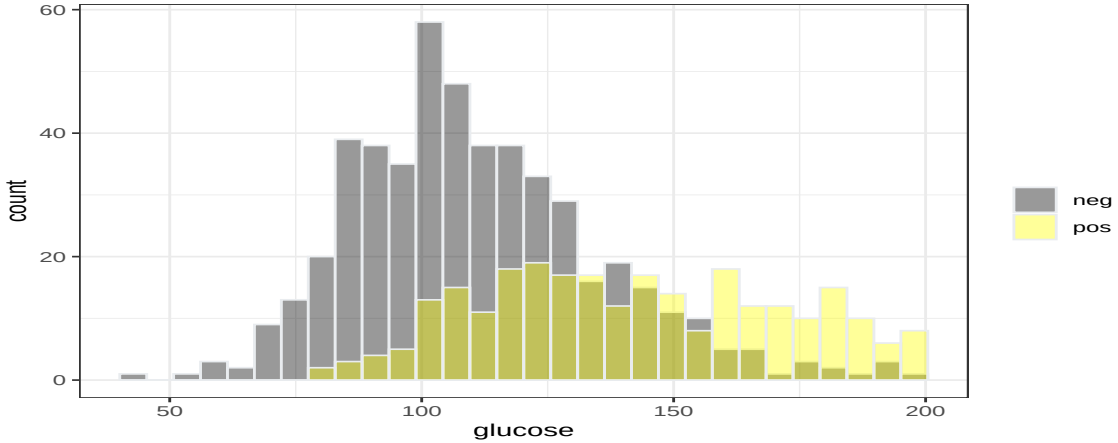
Diagrama de dispersión



Más y más y más

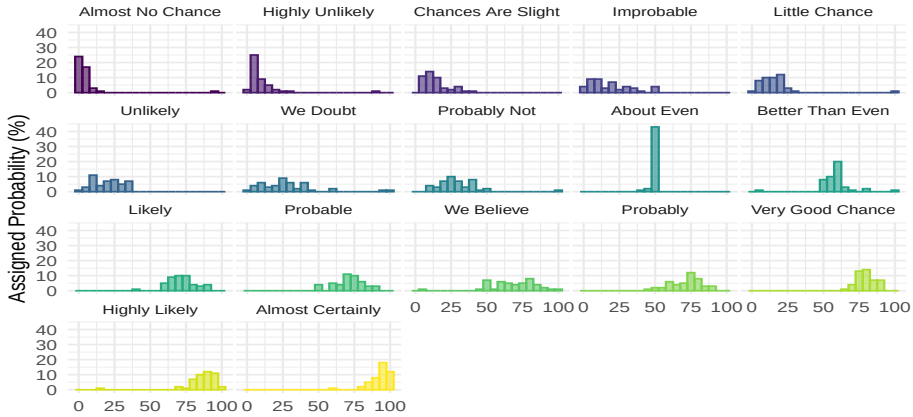
- R graph gallery https://r-graph-gallery.com/
- R Gallery book https://bookdown.org/content/b298e479-b1ab-49fa-b83d-a57c2b034d49/
- ¿El mejor gráfico hecho? The Minard map https://bigthink.com/strange-maps/229-vital-statistics-of-a-deadly-campaign-the-minard-map/

Histogramas conjuntos

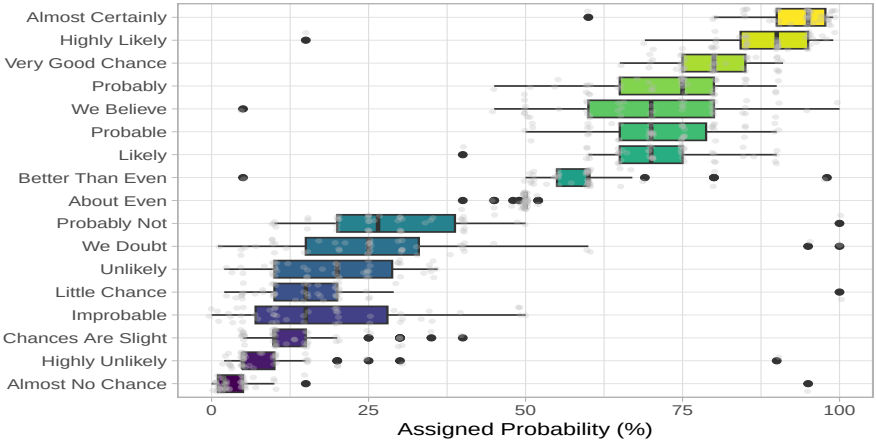


Histogramas conjuntos

Origen del gráfico: From Data to Viz

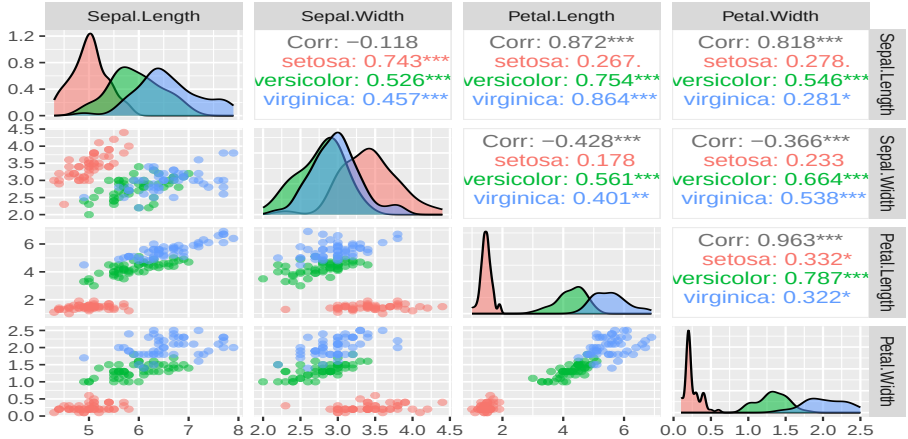


Origen del gráfico: From Data to Viz



ggpairs

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



Más y más

- Gráficos multivariantes
- Gráficos de correlación
- Series temporales
- Mapas
- 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/