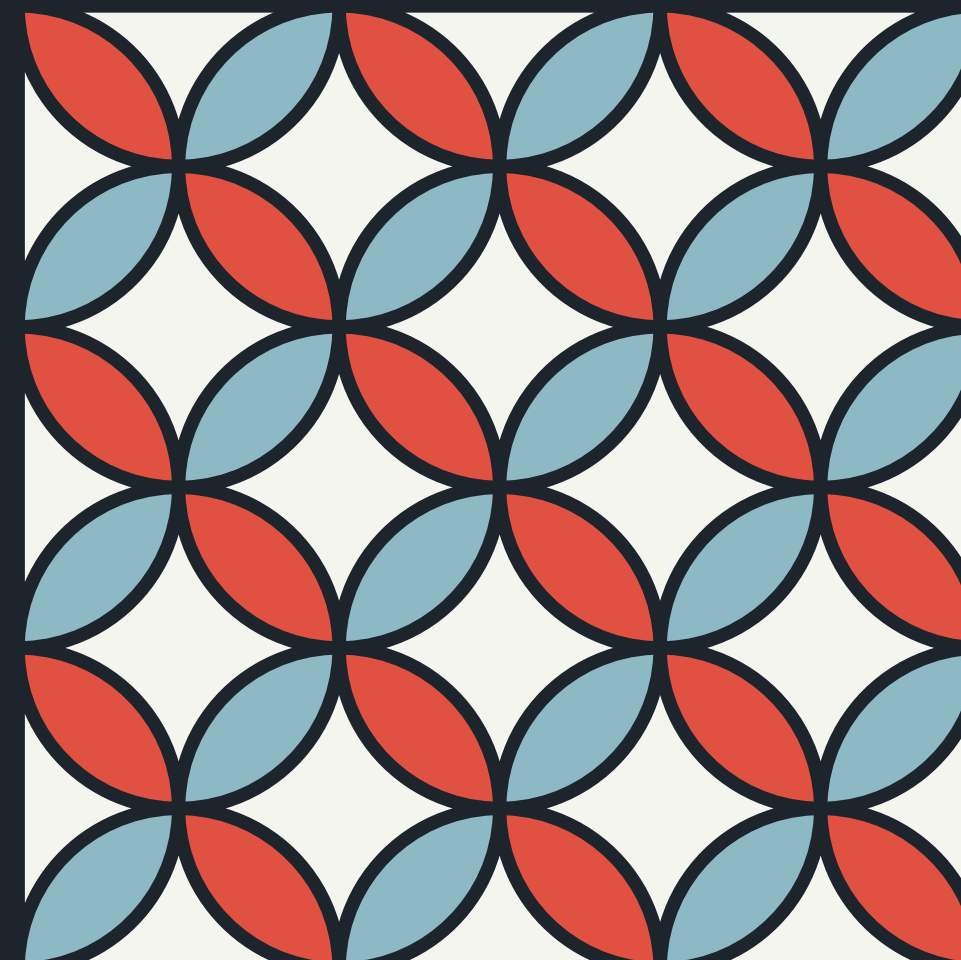
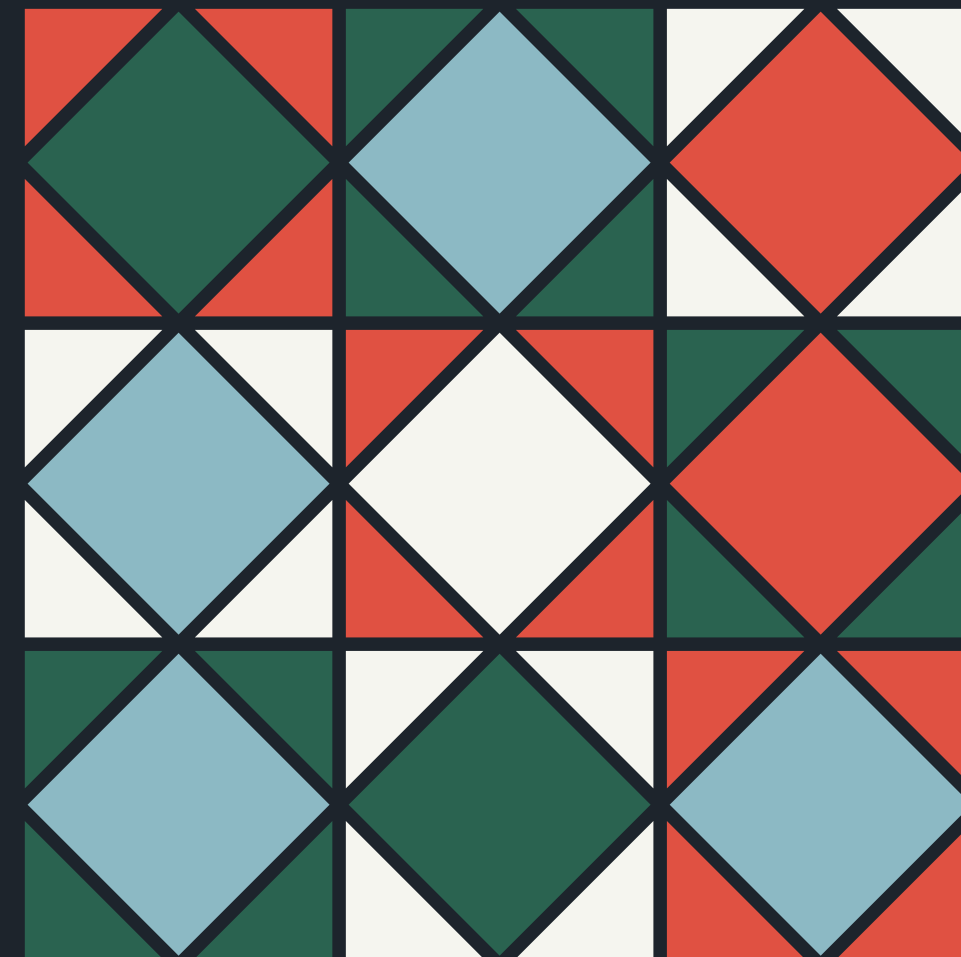
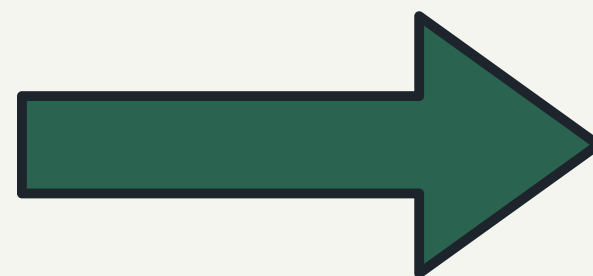


ANÁLISIS EXPLORATORIO DE DATOS



¿QUIEN SOY YO?

ALBERTO CARRILLO

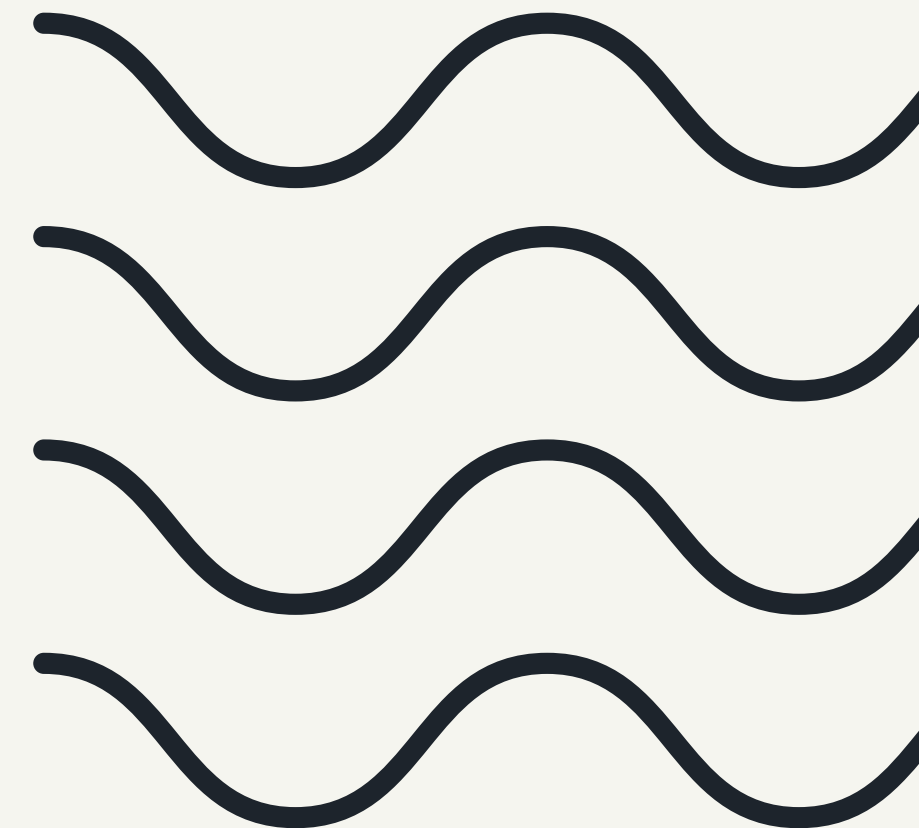
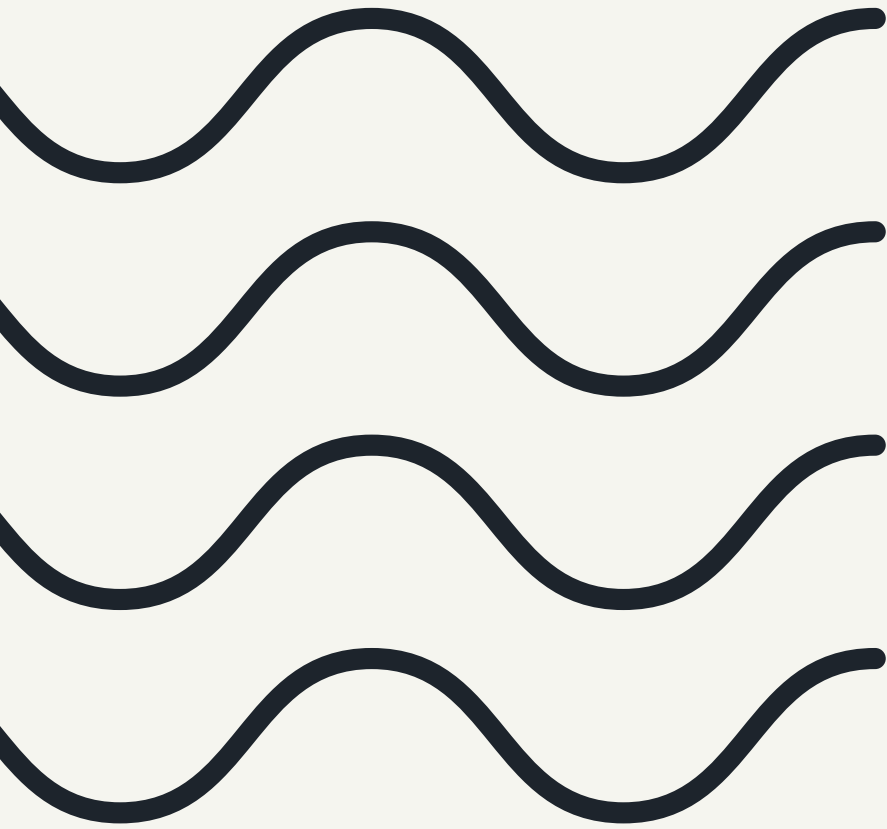
METODOLOGO - PSICOLOGO - AI ENGINEER





CONTENIDO

- INTRODUCCIÓN
- HERRAMIENTAS DE PYTHON
- TIPOS DE DATOS
- DESCRIPCIÓN DE LOS DATOS
- PRIMERAS VISUALIZACIONES
- MISSING DATA
- OUTLIERS
- CORRELACIONES
- CONCLUSIONES Y FUTUROS PASOS





BREVE INTRODUCCIÓN

PROCESO PARA RESUMIR LAS PRINCIPALES CARACTERÍSTICAS QUE PRESENTAN. NOS PERMITE IDENTIFICAR PATRONES EN LOS DATOS, RELACIONES ENTRE VARIABLES, DISTRIBUCIONES, ANOMALÍAS, ETC.

NOS PERMITE EVITAR ERRORES, GUIAR DECISIONES TÉCNICAS Y PREPARAR EL DATO.



REPOSITORIO DE TRABAJO

HERRAMIENTAS DE PYTHON



★ **NUMPY**

Cálculos numéricos, trabajos con arrays, etc.

★ **PANDAS**

Gestión de conjuntos de datos ("tablas"),
lectura de csv, excel. Transformaciones.

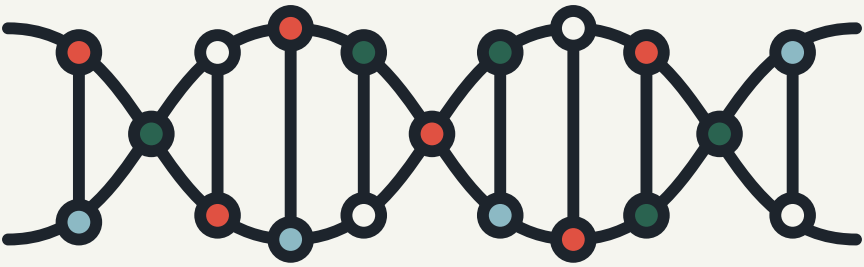
★ **MATPLOTLIB/
SEABORN**

Visualización de datos, creación de gráficos.

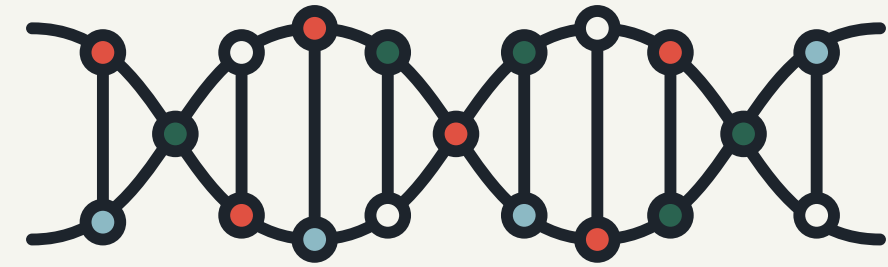
★ **SCIKIT-LEARN**

Librería "toolbox" para análisis estadísticos,
ML, etc.





TIPOS DE DATOS



# peso	# número de hijos	# porcentaje de algo	# salario mensual
97.45	0	13.95	1682.1
87.93	0	29.21	1260.21
99.72	1	36.64	4795.54
112.85	1	45.61	4862.53
86.49	0	78.52	4233.59
86.49	0	19.97	2218.46
113.69	0	51.42	1390.69
101.51	2	59.24	3736.93
82.96	2	4.65	2760.61
98.14	2	60.75	1488.15

nivel de gravedad	glucosa en sangre	fumador
Missing: 0 (0%) Distinct: 3 (30%)	Missing: 0 (0%) Distinct: 3 (30%)	Missing: 0 (0%) Distinct: 2 (20%)
red	40% low	40% yes
blue	30% medium	30% no
green	30% high	30% no
red	low	yes
blue	medium	no
green	high	yes
red	low	no
blue	high	yes
green	medium	no
red	low	yes
blue	high	no
green	medium	yes
red	low	no

fecha_completa	año-mes	timestamp
Missing: 0 (0%) Distinct: 10 (100%)	Missing: 0 (0%) Distinct: 10 (100%)	Missing: 0 (0%) Distinct: 10 (100%)
Min: 2023-01-01 00:00:00 Max: 2023-01-10 00:00:00	Min: 2023-01-31 00:00:00 Max: 2023-10-31 00:00:00	Min: 2023-01-01 00:00:00 Max: 2023-01-03 06:00:00
0	2023-01-01 00:00:00	2023-01-31 00:00:00
1	2023-01-02 00:00:00	2023-02-28 00:00:00
2	2023-01-03 00:00:00	2023-03-31 00:00:00
3	2023-01-04 00:00:00	2023-04-30 00:00:00
4	2023-01-05 00:00:00	2023-05-31 00:00:00
5	2023-01-06 00:00:00	2023-06-30 00:00:00
6	2023-01-07 00:00:00	2023-07-31 00:00:00
7	2023-01-08 00:00:00	2023-08-31 00:00:00
8	2023-01-09 00:00:00	2023-09-30 00:00:00
9	2023-01-10 00:00:00	2023-10-31 00:00:00

NUMERICOS

- Continuos: Toman cualquier valor en un rango (Altura, Salario Medio, Peso)
- Discretos: Solo pueden tomar valores enteros (número de hijos, recuentos)

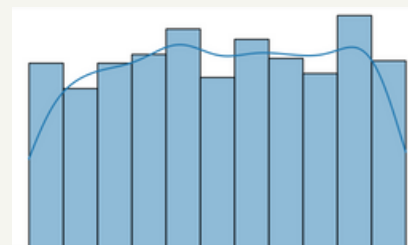
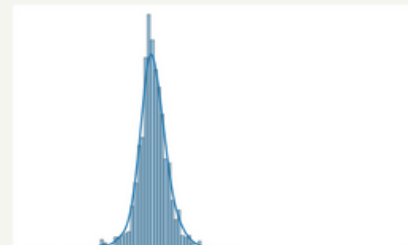
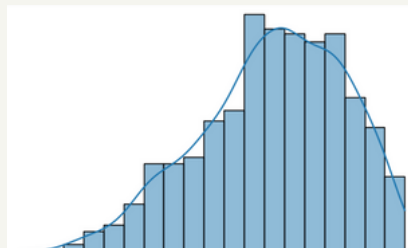
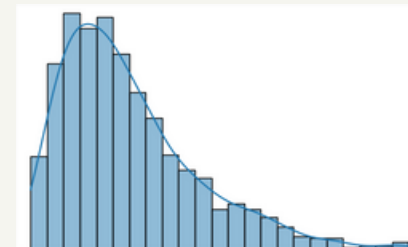
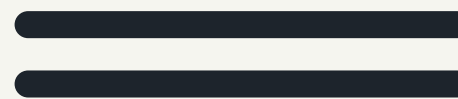
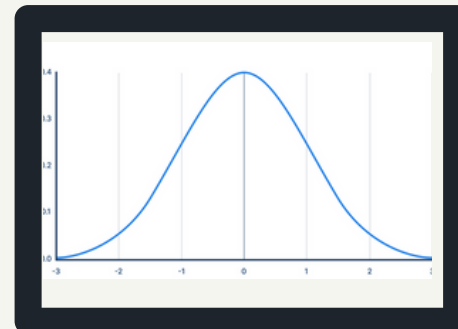
CATEGORICOS

- Nominales: Categorías sin orden natural (color de ojos, género).
- Ordinales: Categorías con un orden natural (Satisfacción, Nivel Educativo).
- Dicotómicas: Solo toman dos valores (si/no)

FECHAS

Técnicamente pueden considerarse continuas (punto en el tiempo), pero en Python requieren tratamiento especial por las diferencias en los formatos y transformaciones.

DESCRIBIR DATOS

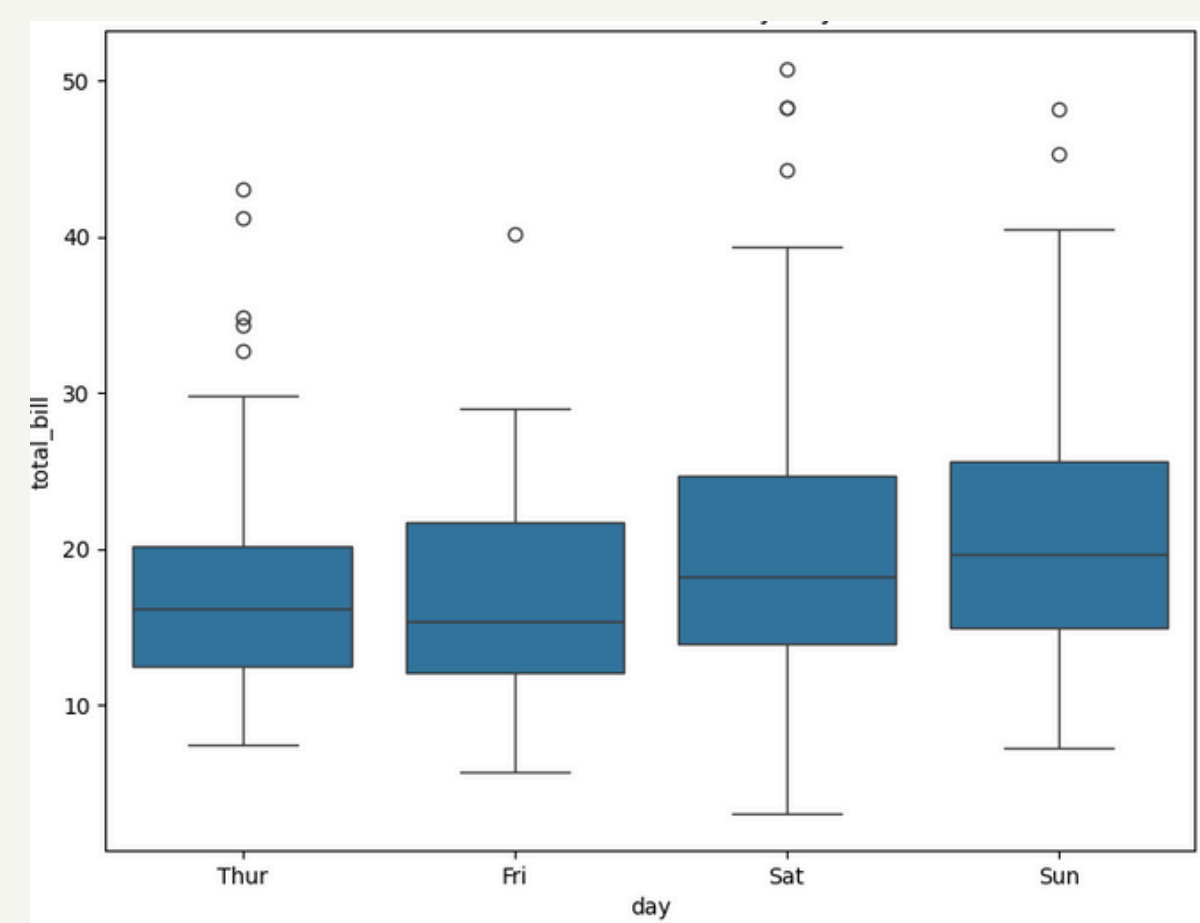
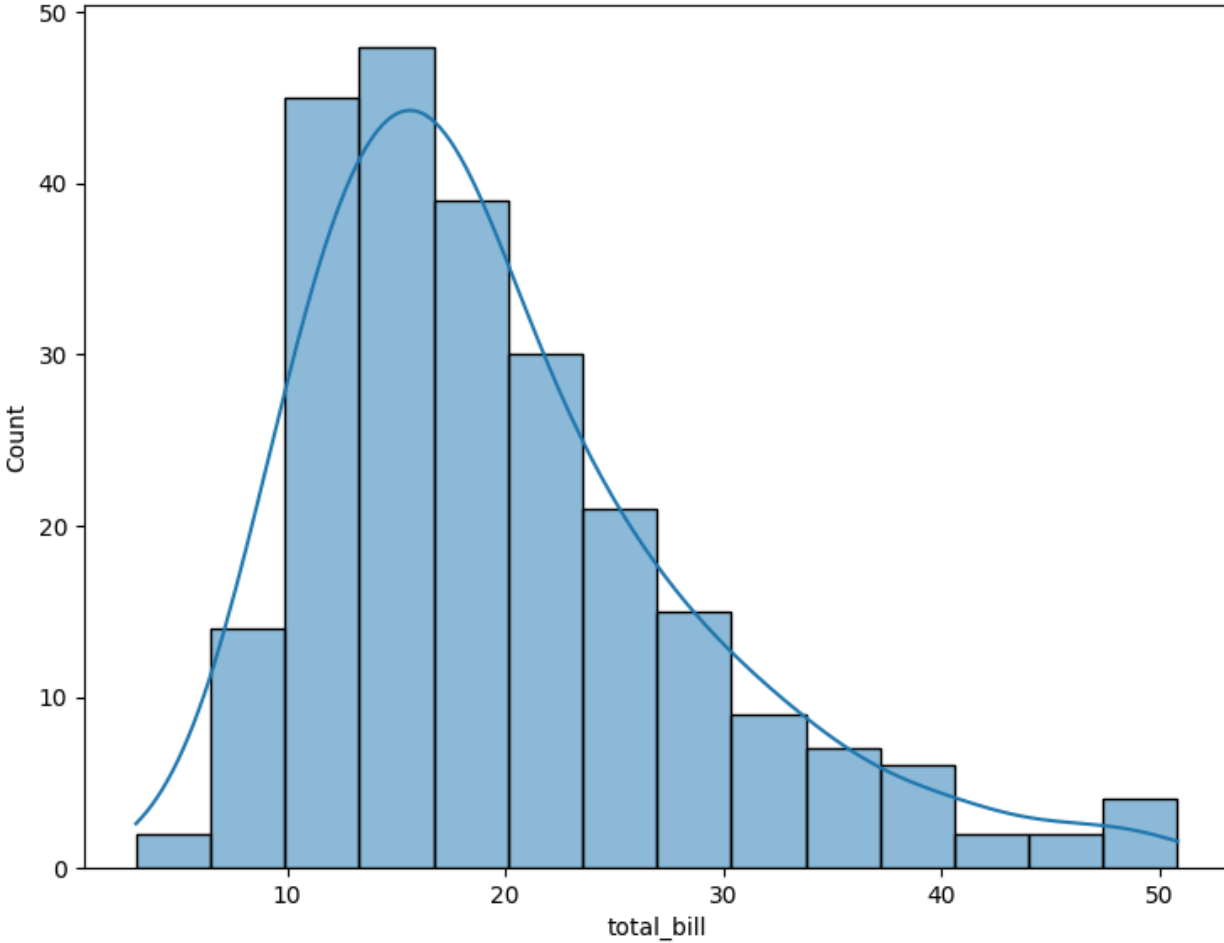


EN QUE CONSISTE

Como hemos mencionado antes, buscamos encontrar las características clave de los datos, para entender la estructura, distribución y posibles problemas de nuestro conjunto de datos.

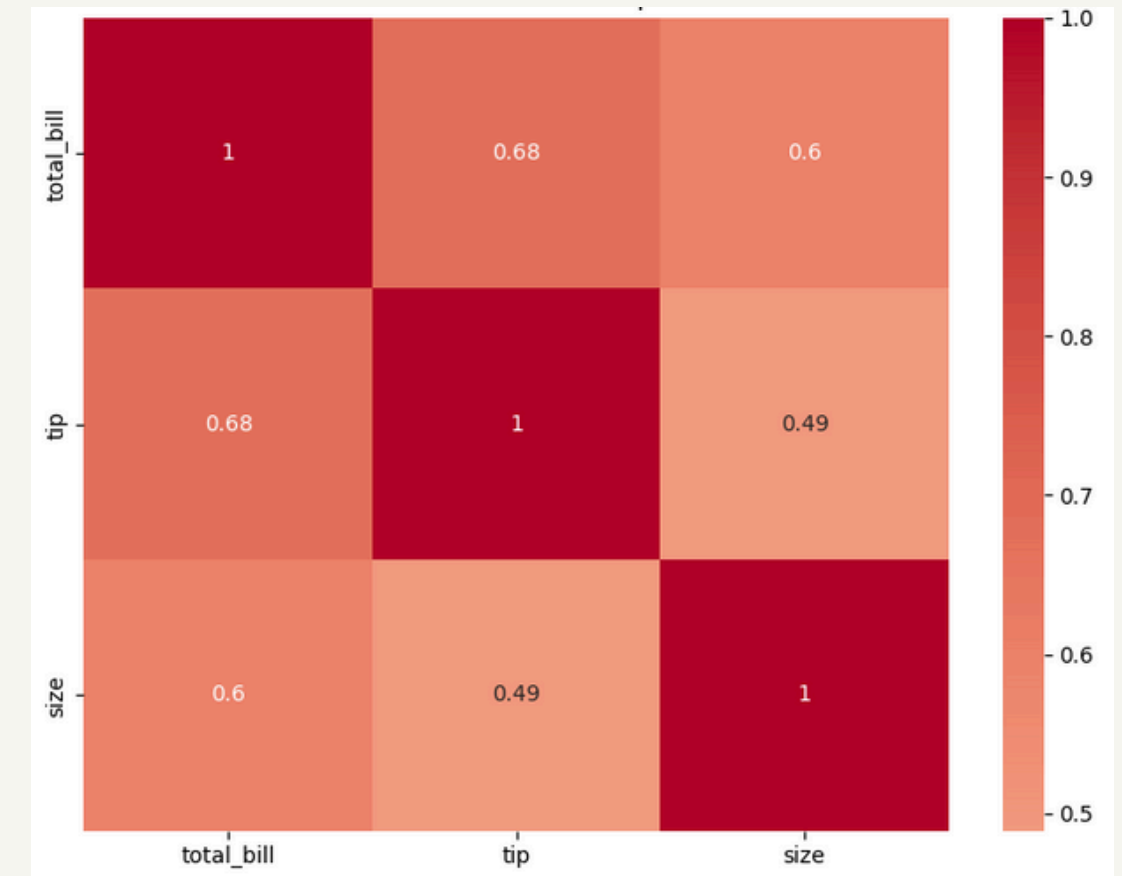
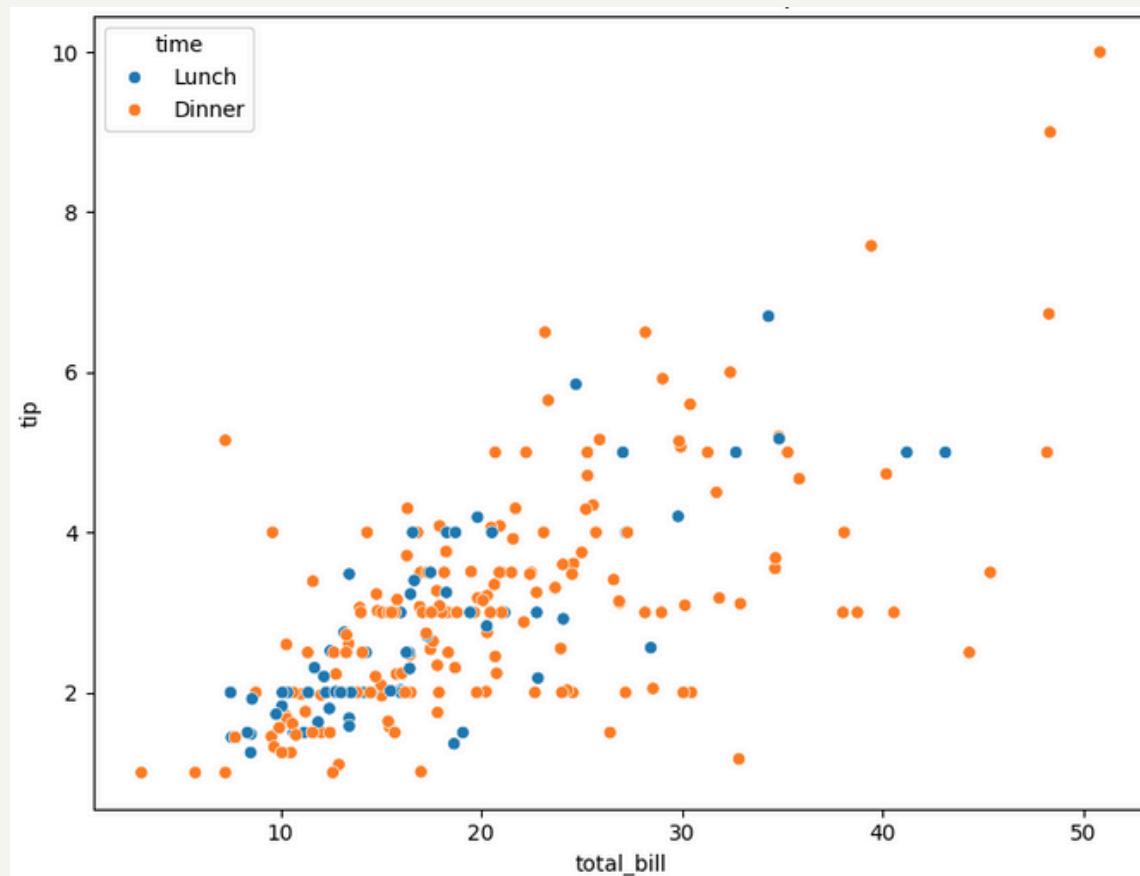
ESTADISTICOS CLAVE

- Tendencia Central: Valor de la variable más probable de encontrar.
 - Media: Promedio de todos los valores. `[df.mean()]`
 - Mediana: Valor central que divide el conjunto de datos en dos partes iguales. `[df.median()]`
 - Moda: Valor que aparece con más frecuencia. `[df.mode()]`
- Dispersión: Concentración de los valores en torno al centro.
 - Desviación estándar: Cuán alejados están los valores de la media. `[df.std()]`
- Forma: Que pinta tiene la distribución de valores.
 - Asimetría: Cómo de simétrica es la distribución y hacia que lado tiende. `[df.skwe()]`
 - Kurtosis: Como de 'apuntada' es la distribución. `[df.kurtosis()]`
- Frecuencias. Permiten ver si hay desbalance en los datos, cuántos valores únicos hay, si hay datos perdidos...



PRIMERAS VISUALIZACIONES

¿QUIEN PUEDE DECIR PARA QUE SIRVEN ALGUNAS DE ELLAS?





OUTLIERS

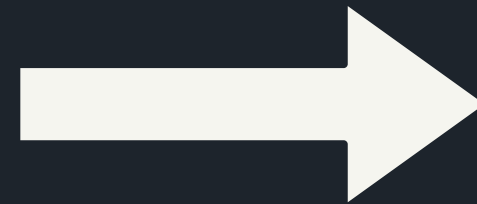
VALORES EN UNA VARIABLE QUE SE ALEJAN EN GRAN MEDIDA DEL VALOR ESPERADO.

DEPENDEN DEL TIPO DE VARIABLE CON LA QUE ESTAMOS TRABAJANDO PUEDE TENER DIFERENTES SIGNIFICADOS:

1. ERROR EN LA RECOGIDA DEL DATO.
2. ES UN DATO REAL PERO POCO COMÚN.
3. DATO PERDIDO.

SOLUCIONES:

1. ELIMINAR
2. TRABAJAR CON ELLOS
3. RECUPERAR EL DATO



IDENTIFICACIÓN

BOXPLOT

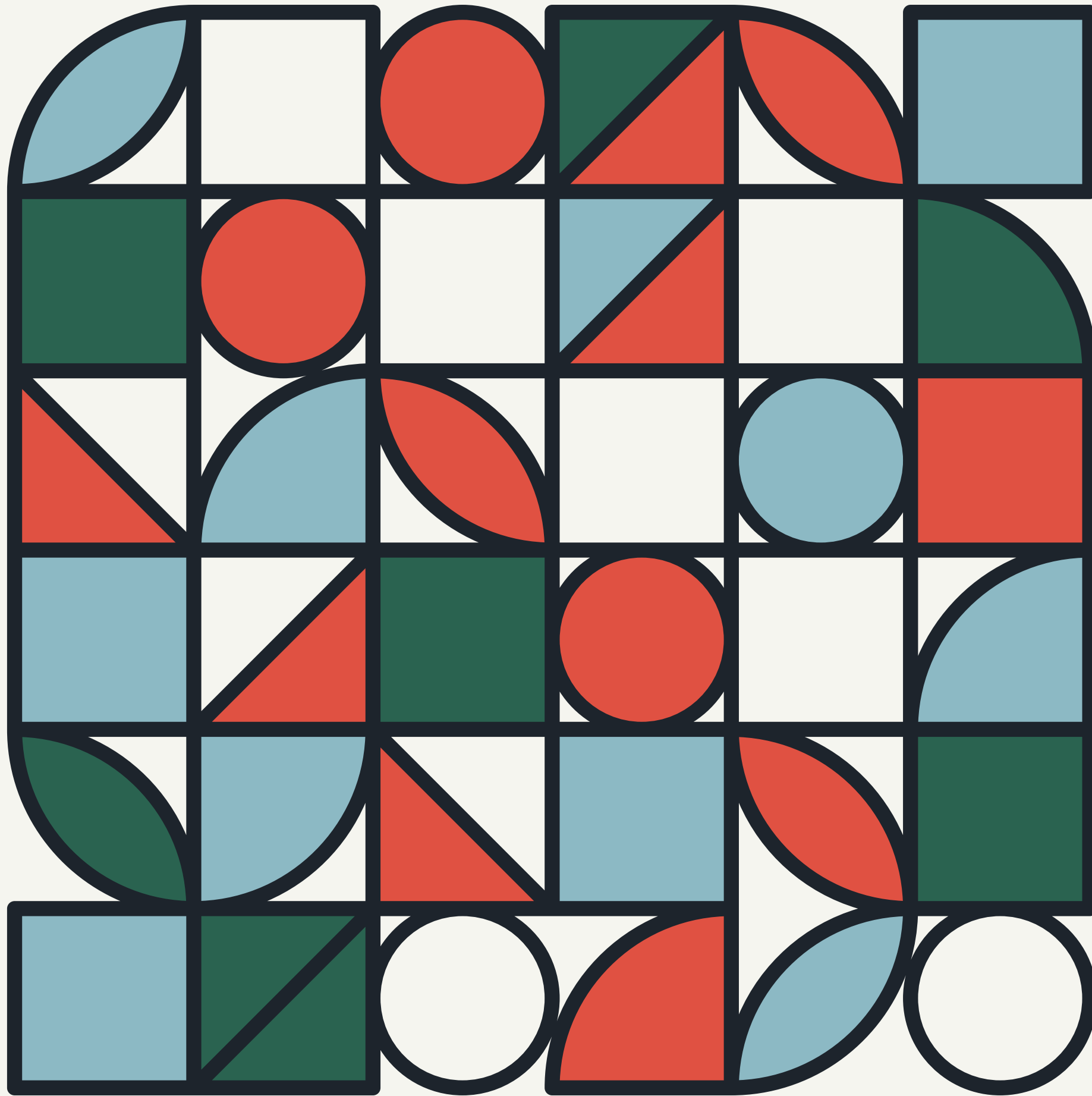
Puntos fuera de los "bigotes del gráfico".

PUNTUACIÓN Z

También llamada estandarización, transforma los datos para que tengan media 0 y desviación típica 1.

Valores $\geq |3|$





DATOS PERDIDOS

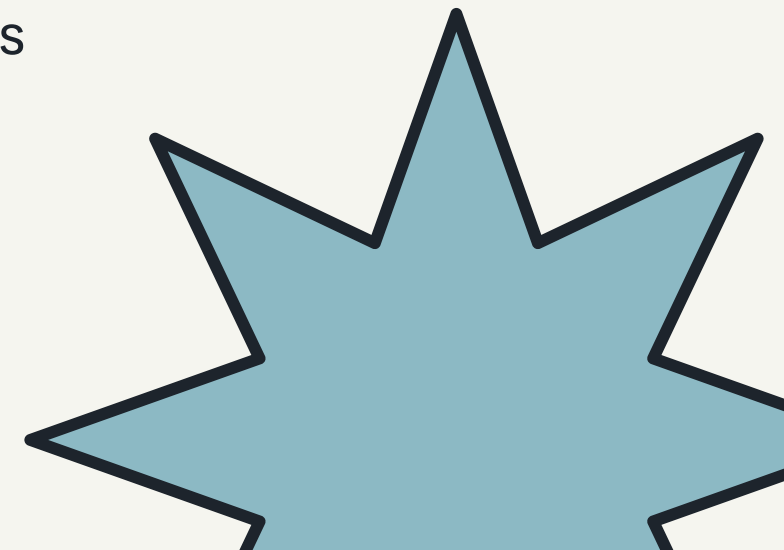
Datos que no están presentes en la variable. (Null u otros)

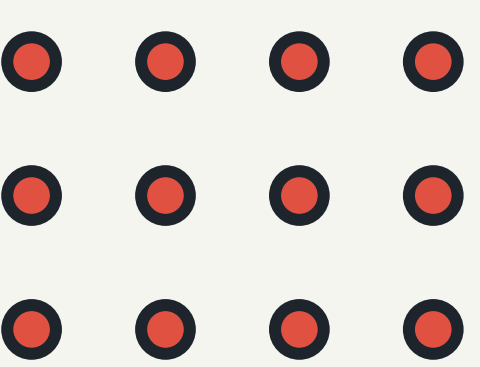
Riesgos que generan los datos perdidos en nuestros análisis:

- Ineficiencia
- Falta de consistencia / sesgo
- Falta de coincidencia entre error tipo I empírico y nominal
- Pérdida de potencia estadística
- Degradación de intervalos de confianza
- Errores típicos sesgados

Como lidiar con ellos:

- Eliminar cuando son muy pocos
- Sustitución de los valores:
 - Media, Moda, Mediana.
 - ML y Imputación Múltiple





CORRELACIONES

Esta medida de dependencia lineal nos indica el grado de relación entre dos variables.

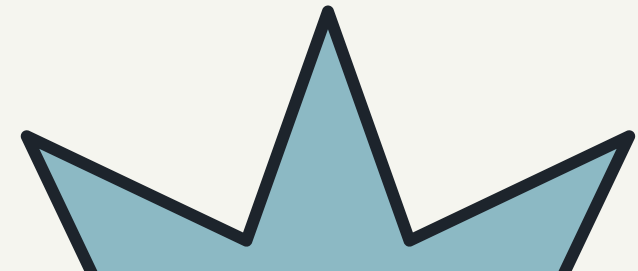
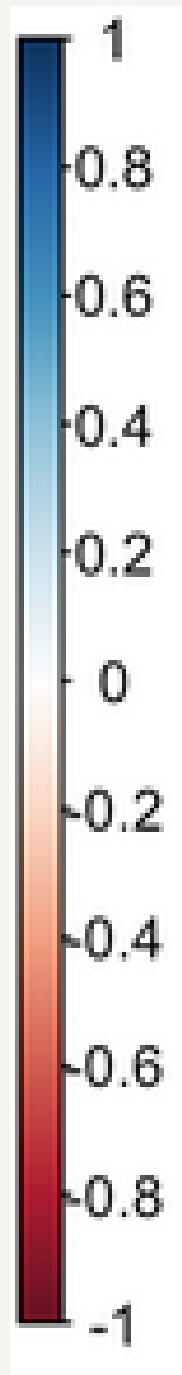
Diferentes estadísticos en función de los tipos de las variables a comparar (Pearson, Spearman, Biserial...)

Para considerar una correlación como relevante debe tener significación estadística, no nos vale con que el valor sea alto. Es más, es erróneo pensar que el valor del coeficiente de correlación sea indicativo de nada por sí solo.

Otro aspecto **MUY IMPORTANTE** respecto a la correlación es que **CORRELACIÓN NO IMPLICA CAUSALIDAD**.

No podemos asumir que porque dos variables presenten una correlación muy alta entre ellas presenten algún tipo de relación causa-efecto, ya que puede ser que no tengan una relación directa, lo que se conoce como una correlación espuria. Como ejemplo, número de cigüeñas en una ciudad y número de nacimientos correlaciona.

A partir de los resultados de las correlaciones podemos empezar a sacar conclusiones sobre nuestros datos y orientar los siguientes pasos en el análisis. Incluso, como veréis más adelante, es útil para una herramienta muy importante en ML, feature engineering, pero eso ya lo veréis en su momento.



CONCLUSIONES Y FUTUROS PASOS

NOTEBOOK

En el notebook del repositorio podéis ampliar los conceptos vistos en la clase y hacer pruebas con código.

TEORÍA

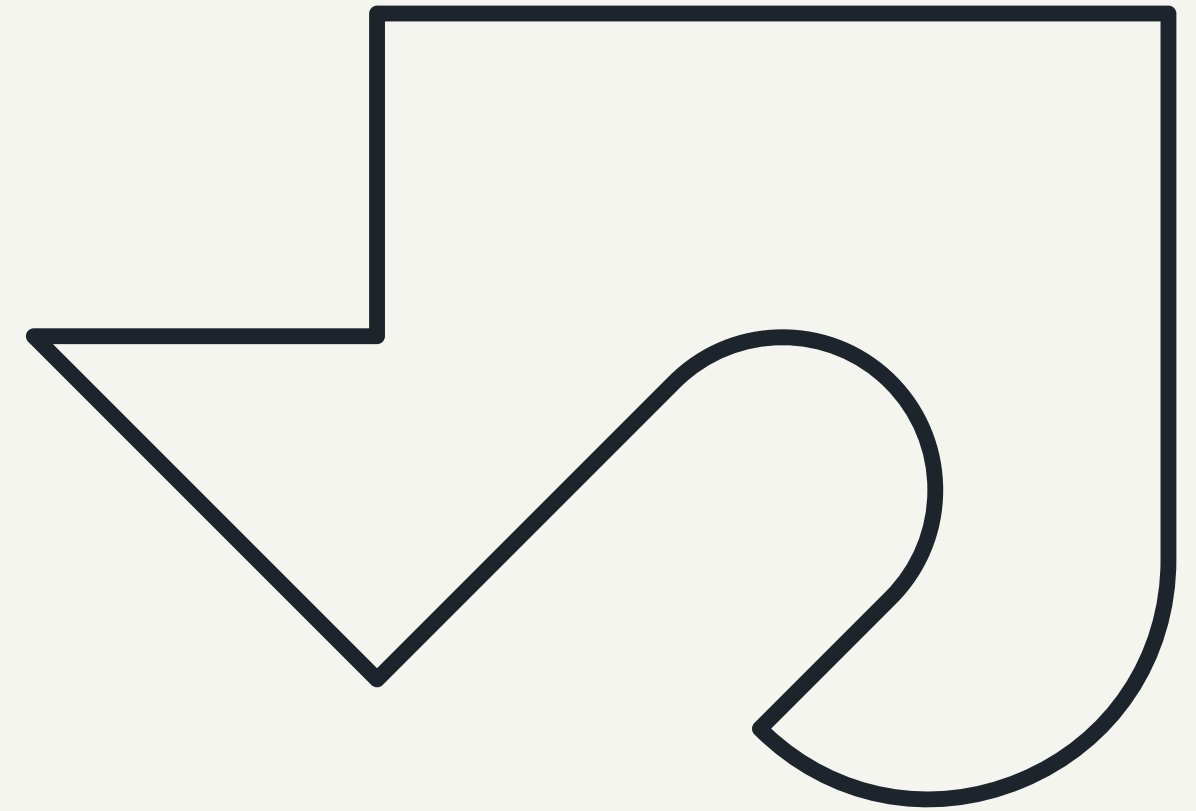
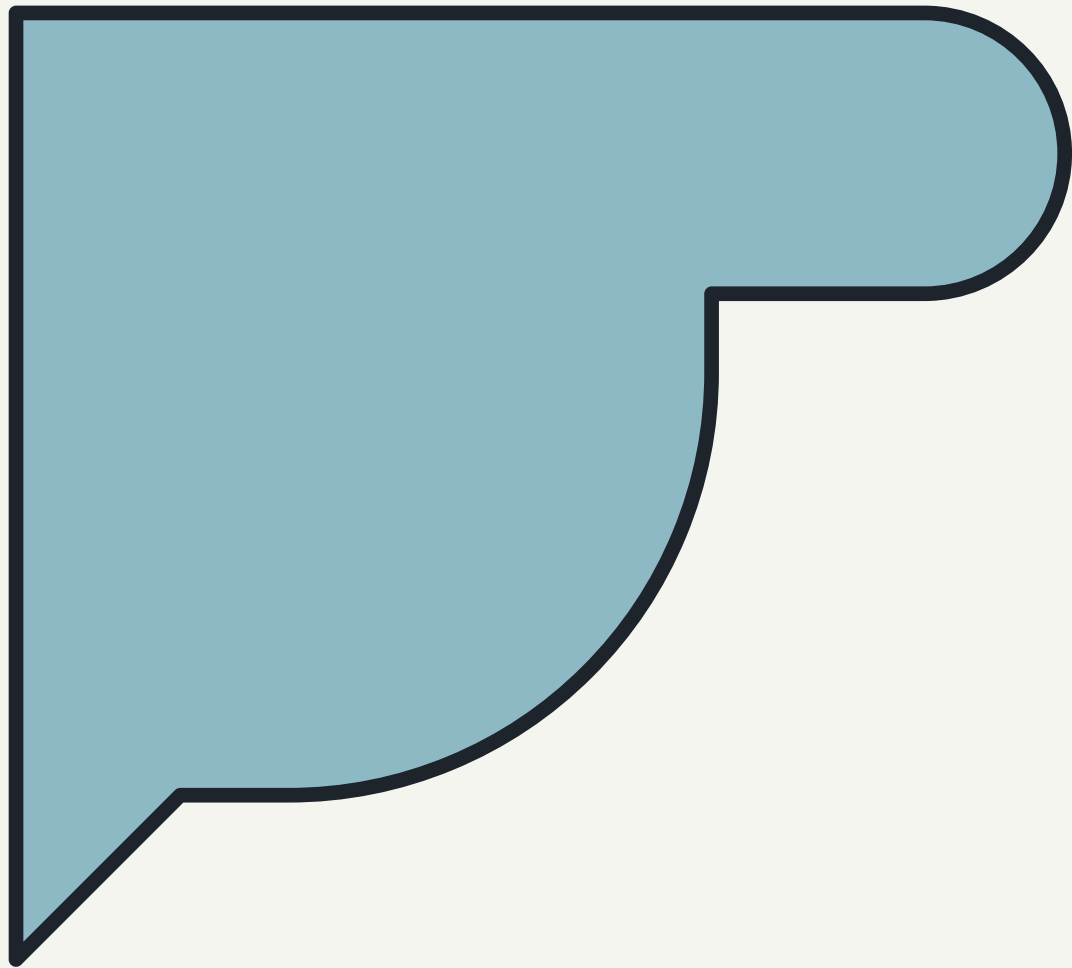
Relevante intentar empaparse de la teoría que subyace a muchos de estos procesos, ya que nos hará mejores profesionales.

PRACTICAR

Tenéis un csv y una invitación a realizar vuestro propio EDA para ir cogiendo ritmo.

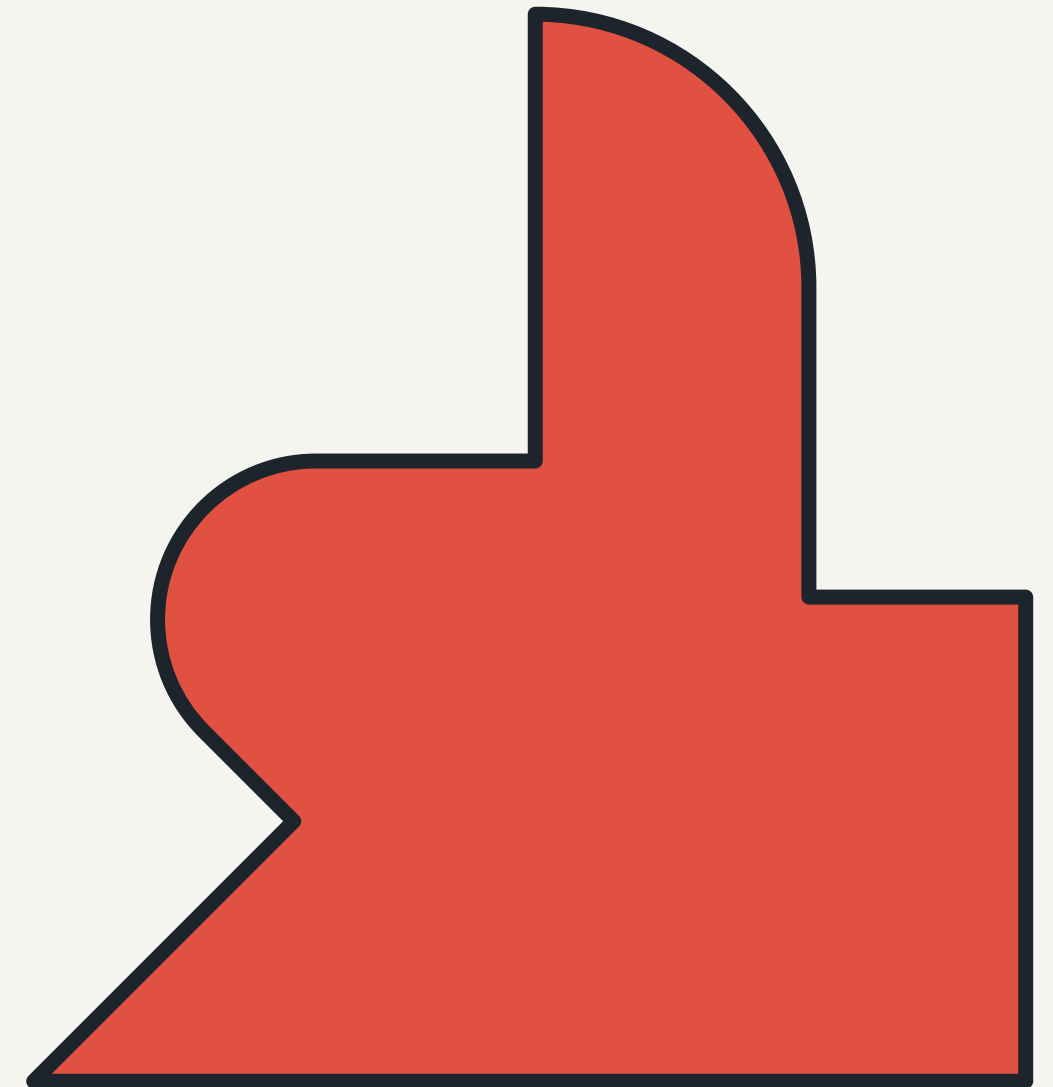
PASO CLAVE


Este proceso es clave para poder trabajar eficientemente en cualquier análisis o modelado que vayáis a hacer posteriormente.



PERO ANTES DE IRNOS

VAMOS A COMPROBAR QUE REALMENTE HABÉIS ESTADO
ATENTOS.



The background features a central square frame with rounded corners, outlined in dark blue. The frame is divided into four quadrants by a horizontal and vertical line, each containing a semi-circular shape: top-left is light blue, top-right is red, bottom-left is red, and bottom-right is light blue. The corners of the frame are filled with dark green semi-circles. Surrounding the frame are decorative elements: two orange four-pointed stars (one in the top-right and one in the bottom-left) and two hexagonal molecular-like structures (one in the top-left and one in the bottom-right). These structures consist of dark blue lines forming hexagons with colored circles (white, red, green, and light blue) at the vertices.

**MUCHAS
GRACIAS POR
AGUANTAR EL
CHAPARRÓN**