UAX Universidad Alfonso X el Sabio

CIENCIA DE DATOS

GRADO EN BIOMEDICINA

DR. ARIEL CARIAGA-MARTÍNEZ

GOBERNANZA DEL DATO.

- Políticas y procedimientos: no podemos hacer ”lo que queremos como queremos”. Documentación legal vinculante.
- Roles y responsabilidades: *data owners, data stewards y data users* (entre otros).
- Calidad del dato → HOY
- Seguridad/privacidad, accesiblidad/disponibilidad.
- Trazabilidad y auditoría. → Auditorías → RESPONSABLES

EN DATA SCIENCE TODO POR ESCRITO

GOBERNANZA DEL DATO: CONTROL DE CALIDAD DEL DATO

UAX

- Exactitud: verificar que los datos reflejan la realidad.
- Integridad: comprobar que no falten datos importantes.
- Consistencia: mantener uniformidad en múltiples sistemas.
- Validez: asegurarse de que los datos cumplen con los formatos correctos.
- Puntualidad: mantener los datos actualizados.
- Verificabilidad: asegurar que las transformaciones sean trazables.

IMPORTANCIA DEL “WRANGLING”

- CALIDAD DE LOS DATOS → VERIFICAR.
- PRESENCIA DE DATOS INCOHERENTES, INCOMPLETOS O INCOSISTENTES.
- “FORMATEO” DE LOS DATOS → CODIFICACIONES NO CONSISTENTES (PRESENCIA DE PUNTOS Y COMAS COMO DECIMALES, PRESENCIA DE ESPACIOS EN BLANCO, ETC.).
- MANEJO DE FECHAS.
- ¿OTROS “PROBLEMAS”?

Valores faltantes: ¿por qué faltan los datos?

Valores faltantes completamente al azar (MCAR - Missing Completely at Random)

- Los valores faltantes no están relacionados ni con las variables observadas ni con las no observadas. Es decir, la ausencia de datos es completamente aleatoria. MCAR es la situación más favorable, ya que los análisis que ignoren los valores faltantes todavía podrían ser válidos si no hay un sesgo introducido por la falta de datos.
- Por ejemplo, si en una encuesta, algunas personas omiten preguntas sin ninguna razón aparente, esos valores podrían considerarse MCAR.

Valores faltantes al azar (MAR - Missing at Random)

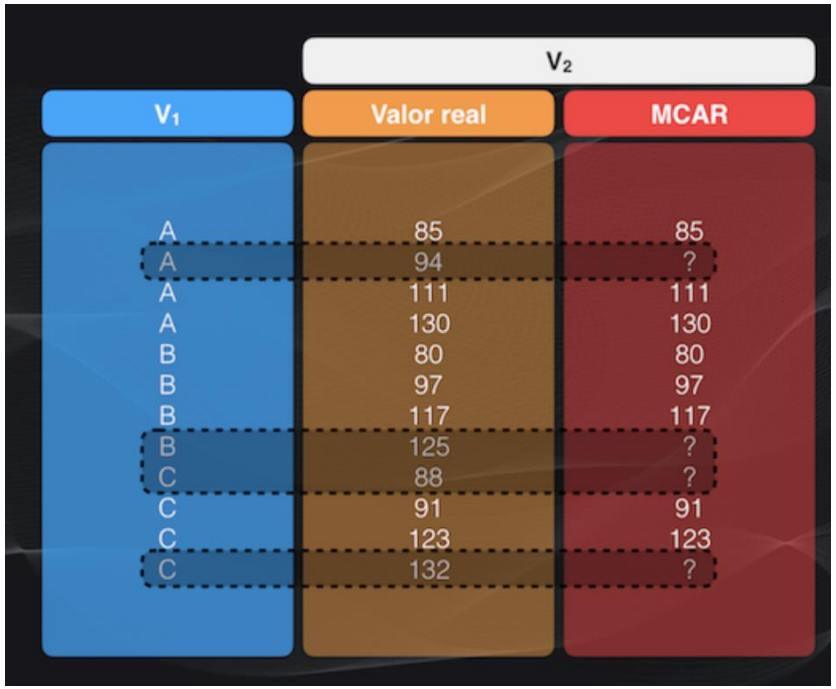
- La probabilidad de que un valor esté faltante depende de las variables observadas, pero no de los valores no observados. En este caso, si conocemos algo sobre las características observadas, podríamos inferir información sobre los datos faltantes.
- Por ejemplo, en un estudio médico, las personas mayores podrían ser menos propensas a responder a preguntas relacionadas con tecnología, pero si tenemos su edad, podemos usar esa información para imputar los valores faltantes. **TAMBIÉN DESEABILIDAD SOCIAL.**

Valores faltantes no al azar (MNAR - Missing Not at Random)

- Los valores faltantes dependen de la información no observada, es decir, de los mismos valores faltantes → **REALMENTE HAY UN SESGO SISTEMÁTICO**. Los datos MNAR son los más difíciles de manejar, ya que no se puede simplemente imputar o ignorar los valores sin introducir sesgo.
- Un ejemplo sería si las personas con ingresos más bajos son menos propensas a reportar sus ingresos en una encuesta, lo que genera una falta de datos directamente relacionada con el valor faltante en sí mismo. Alcohólicos/jugadores preguntados por sus hábitos al respecto.

Ejemplo

SESGO DE DESEABILIDAD / MALA TOMA DE DATOS



V_1	V_2 Valor real	V_2 MCAR
A	85	85
A	94	?
A	111	111
A	130	130
B	80	80
B	97	97
B	117	117
B	125	?
C	88	?
C	91	91
C	123	123
C	132	?

No importa lo que suceda en V1, se "pierden datos" en V2.

V_1	V_2 Valor real	V_2 MAR
A	85	85
A	94	94
A	111	111
A	130	130
B	80	?
B	97	?
B	117	?
B	125	?
C	88	88
C	91	91
C	123	123
C	132	132

Según lo que suceda en V1 se "pierden los mismos niveles de datos" en V2.

V_1	V_2 Valor real	V_2 MNAR
A	85	?
A	94	?
A	111	111
A	130	130
B	80	?
B	97	?
B	117	117
B	125	125
C	88	?
C	91	?
C	123	123
C	132	132

Según lo que suceda en V1 se "pierden ciertos niveles de datos" en V2 (por ejemplo, menores a 100)

https://www.codificandobits.com/blog/manejo-datos-faltantes/

TRATAMIENTO E IMPUTACIONES

UAX

Valores faltantes completamente al azar (MCAR - Missing Completely at Random)

- En estudios **con datos MCAR**, eliminar o ignorar los valores faltantes podría ser una solución adecuada.

Valores faltantes al azar (MAR - Missing at Random)

- En situaciones **con datos MAR**, la imputación basada en otras variables observadas suele ser una buena solución, como el uso de modelos de regresión o kNN.

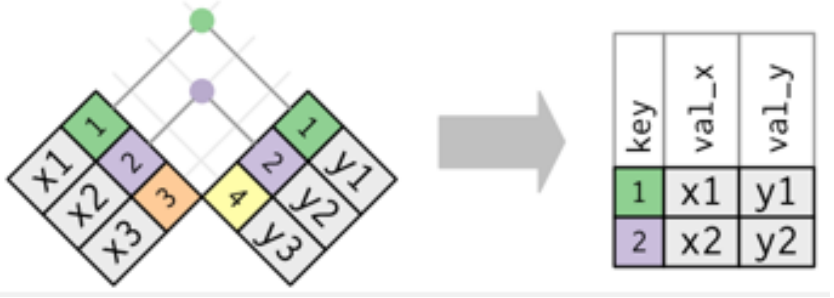
Valores faltantes no al azar (MNAR - Missing Not at Random)

- Para los **datos MNAR**, se recomienda realizar un análisis cuidadoso para tratar de entender el patrón de falta de datos y, cuando sea posible, recolectar más información para explicar por qué faltan los datos. En ocasiones, puede ser necesario un rediseño del estudio o la aplicación de técnicas avanzadas de imputación, como el modelo EM (Expectation-Maximization) o técnicas bayesianas.

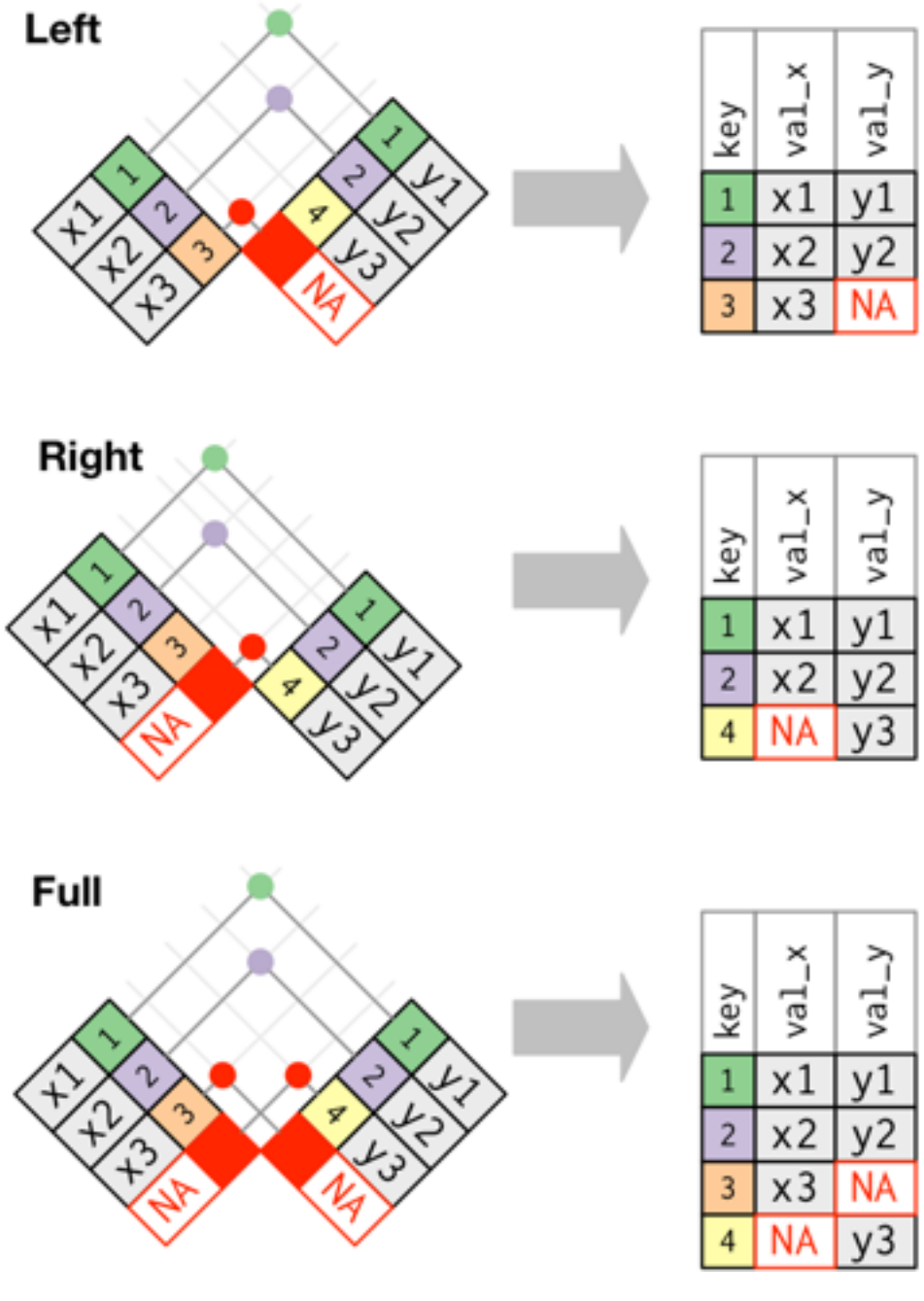
Estrategias: PROS/CONS

Technique	Advantage	Disadvantage
Omit the values	Easy to perform	You lose the information from an observation
Set value to `null`	Easy to perform	Not every modeling technique and/or implementation can handle `null` values
Impute a static value such as 0 or the mean	Easy to perform You don't lose information from the other variables in the observation	Can lead to false estimations from a model
Impute a value from an estimated or theoretical distribution	Does not disturb the model as much	Harder to execute You make data assumptions
Modeling the value (nondependent)	Does not disturb the model too much	Can lead to too much confidence in the model Can artificially raise dependence among the variables Harder to execute You make data assumptions

MANEJO BÁSICO DE DATOS TABULARES: JOINS



INNER: MANTIENE LAS FILAS QUE COINCIDEN EN AMBOS DATASETS



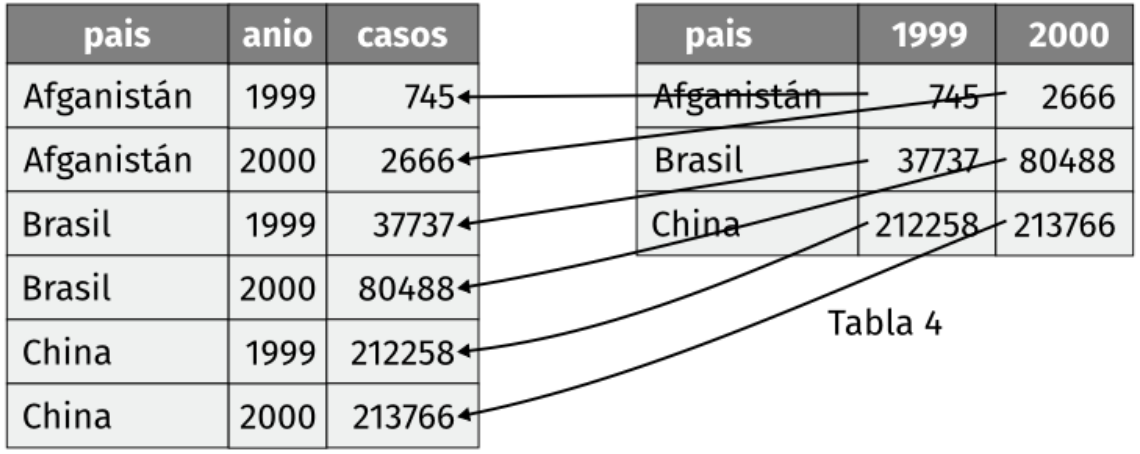
LEFT: MANTIENE TODAS LAS FILAS DEL PRIMER DATASET Y UNE LAS COLUMNAS DEL SEGUNDO DATASET.

RIGHT: MANTIENE TODAS LAS FILAS DEL SEGUNDO DATASET Y UNE LAS COLUMNAS DEL PRIMER DATASET

FULL: MANTIENE TODAS LAS FILAS DE AMBOS DATASETS

MANEJO BÁSICO DE DATOS TABULARES: PIVOTs

- Cada variable en su propia columna.
- Cada observación en su propia fila
- Cada valor en su propia celda.

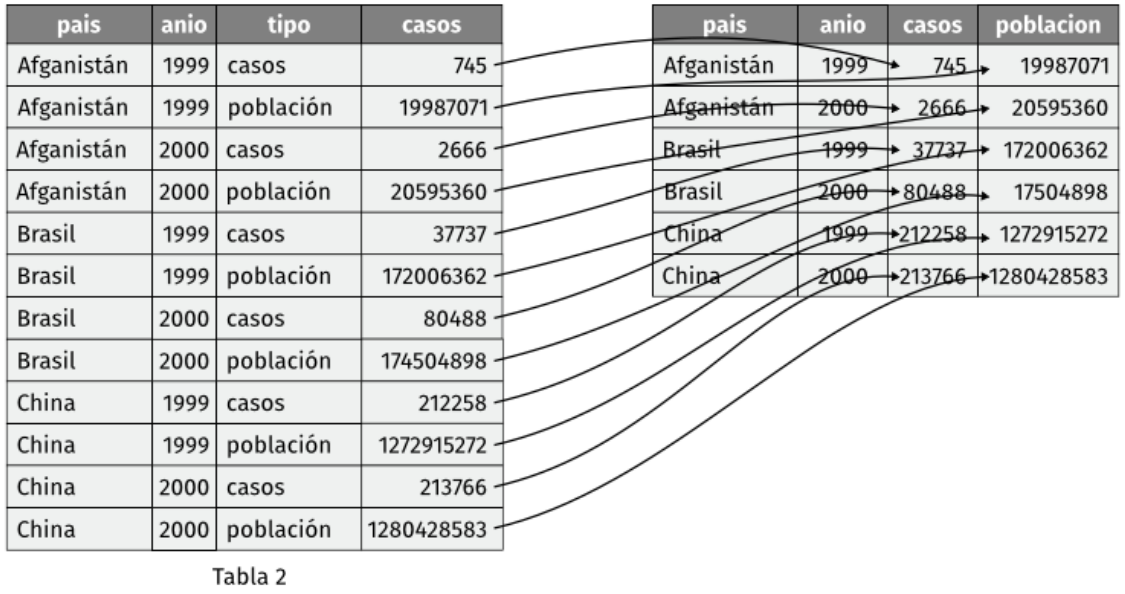


pais	anio	casos
Afganistán	1999	745
Afganistán	2000	2666
Brasil	1999	37737
Brasil	2000	80488
China	1999	212258
China	2000	213766

pais	1999	2000
Afganistán	745	2666
Brasil	37737	80488
China	212258	213766

Tabla 4

Pivot longer: datos “alargados”



pais	anio	tipo	casos
Afganistán	1999	casos	745
Afganistán	1999	población	19987071
Afganistán	2000	casos	2666
Afganistán	2000	población	20595360
Brasil	1999	casos	37737
Brasil	1999	población	172006362
Brasil	2000	casos	80488
Brasil	2000	población	174504898
China	1999	casos	212258
China	1999	población	1272915272
China	2000	casos	213766
China	2000	población	1280428583

Tabla 2

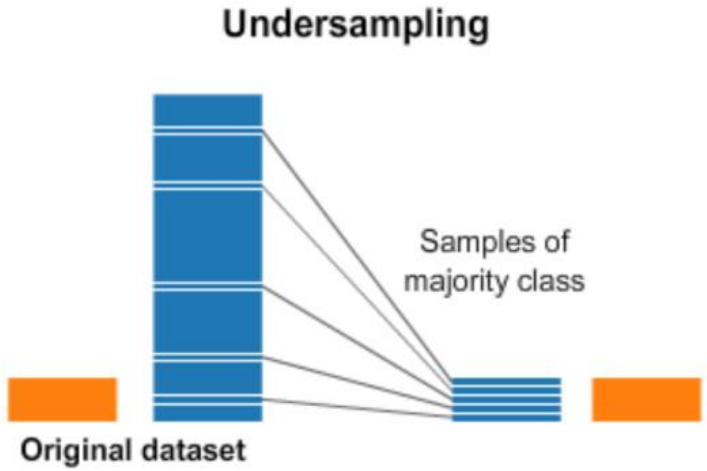
pais	anio	casos	poblacion
Afganistán	1999	745	19987071
Afganistán	2000	2666	20595360
Brasil	1999	37737	172006362
Brasil	2000	80488	17504898
China	1999	212258	1272915272
China	2000	213766	1280428583

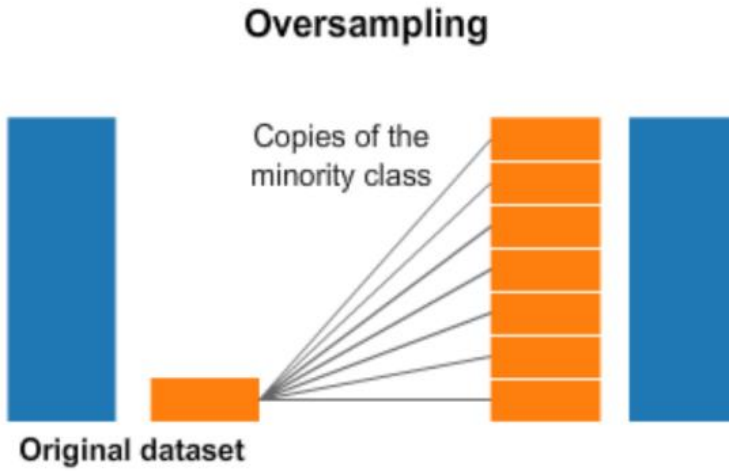
Pivot wider: datos “anchos”

Manejo del desbalance

- **Problema (típicamente de clasificación) → ¿Son datos atípicos, en muy baja frecuencia, “reales”? 99%vs1%.**
- **Implican un sesgo hacia la clase mayoritaria (problema en los modelos de predicción). → clasificar algo porque es lo más común.**
- **Existen estrategias en el modelado para lidiar con esta situación (mejora de métricas, ponderación de pesos en los modelos etc.).**
- **Es un punto que tenemos que observar en el EDA y tratar de darle una solución (si ello es requerido).**

Manejo del desbalance: MUESTREO (sobre/submuestreo)





SMOTE: Synthetic Minority Over-sampling Technique (kNN)

UAX Universidad Alfonso X el Sabio

GRACIAS

DR. ARIEL CARIAGA-MARTINEZ

CIENCIA DE DATOS

ACARIMAR@UAX.ES

UAX Universidad Alfonso X el Sabio

TRABAJOS

GRADO EN BIOMEDICINA

DR. ARIEL CARIAGA-MARTÍNEZ

UNA LUCHA…

- Formar equipos (2-3 personas).
- Realizar un EDA rápido del dataset disponible.
- **¿Con qué problemas te encuentras?**
- GENERA UNA SOLUCIÓN A LOS PROBLEMAS DE CALIDAD DEL DATO OBSERVADOS.
- EL ENTREGABLE ES UN DOCUMENTO TIPO (IMRD (INTRODUCCIÓN/MATERIALES-MÉTODOS/RESULTADOS/DISCUSIÓN): IDEALMENTE RMARKDOWN CON EXPLICACIÓN Y CÓDIGO. NO MÁS DE 2 FOLIOS CENTRADOS EN CALIDAD DEL DATO.
- RETO: ¿PUEDES JUSTIFICAR QUÉ TIPO DE DATO FALTANTE CORRESPONDE EN CADA COLUMNA (MNAR, MAR, MCAR)?T

https://vmblog.com/archive/2022/06/07/understanding-how-to-approach-data-wrangling.aspx