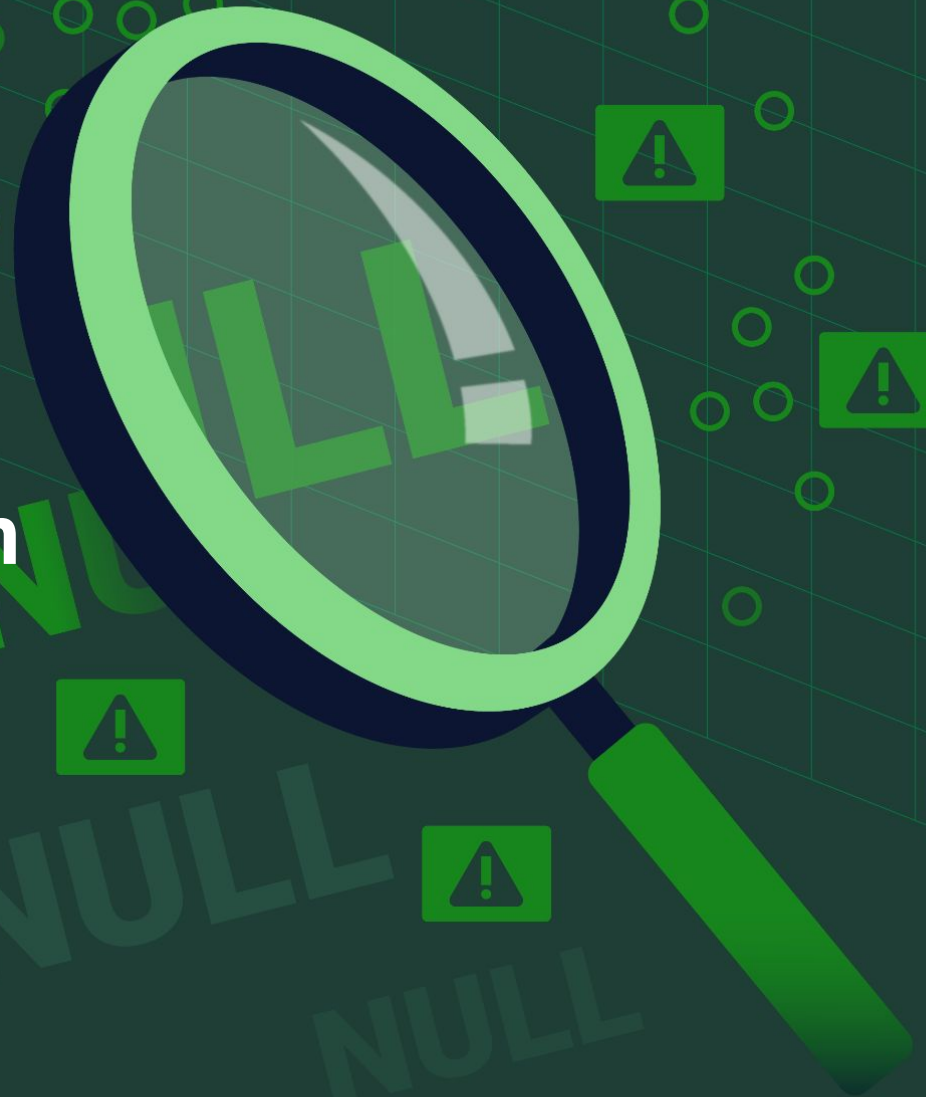


Curso de **Manejo de  
Datos Faltantes:  
Detección y Exploración**

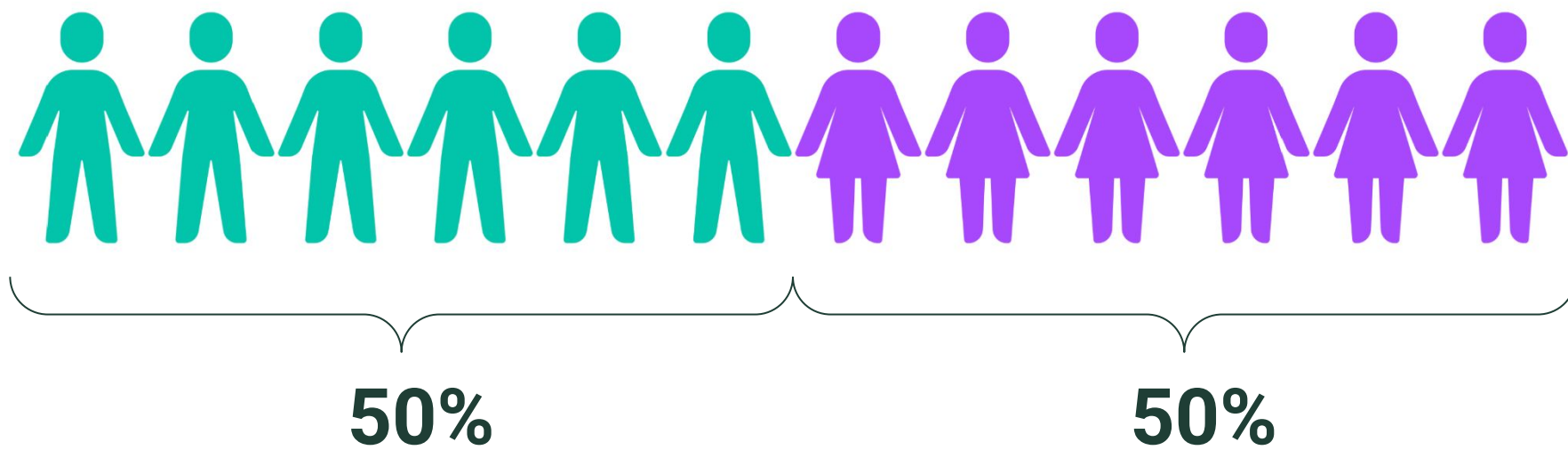
Jesús Vélez | @jvelezmagic

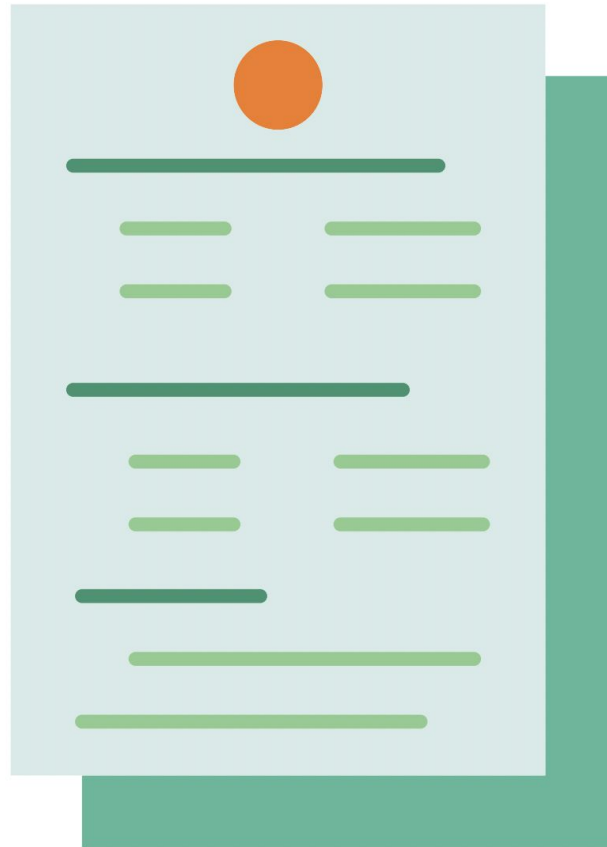


**¿Por qué explorar y  
lidar con los valores  
faltantes?**



**¿Existe una brecha  
salarial entre hombres  
y mujeres?**





survey_id	respondant_name	gender	satisfied	annual_income
1	Arabinda Jagannath	Male	No	70,000
2	Haru Teona	Male	No	60,000
3	Yolonda Hana	Male	Yes	50,000
4	Chalice Libbie	Male	Yes	35,000
5	Pamella Vishnu	Male	Yes	45,000
6	Rajeev Meliton	Male	No	70,000
7	Terrie Bayley	Female	Yes	55,000
8	Sukie Callista	Female	Yes	65,000
9	Denzil Varsha	Female	No	35,000
10	Melvin Dotty	Female	No	40,000
11	Bellamy Suman	Female	Yes	45,000
12	Nikita Mari	Female	No	30,000

survey_id	respondant_name	gender	satisfied	annual_income	Ingreso anual promedio
1	Terrie Bayley	Male	No	70,000	55,000 USD
2	Sukie Callista	Male	No	60,000	
3	Denzil Varsha	Male	Yes	50,000	
4	Melvin Dotty	Male	Yes	35,000	
5	Bellamy Suman	Male	Yes	45,000	
6	Nikita Mari	Male	No	70,000	
7	Arabinda Jagannath	Female	Yes	55,000	45,000 USD
8	Haru Teona	Female	Yes	65,000	
9	Yolonda Hana	Female	No	35,000	
10	Chalice Libbie	Female	No	40,000	
11	Pamella Vishnu	Female	Yes	45,000	
12	Rajeev Meliton	Female	No	30,000	



¿Qué podría pasar si  
**llegan a faltar valores?**



survey_id	respondant_name	gender	annual_income
1	Terrie Bayley	Male	
2	Sukie Callista	Male	60,000
3	Denzil Varsha	Male	
4	Melvin Dotty	Male	35,000
5	Bellamy Suman	Male	
6	Nikita Mari	Male	70,000
7	Arabinda Jagannath	Female	
8	Haru Teona	Female	65,000
9	Yolonda Hana	Female	
10	Chalice Libbie	Female	40,000
11	Pamella Vishnu	Female	
12	Rajeev Meliton	Female	30,000

Ingreso anual promedio

55,000 USD

45,000 USD

survey_id	respondant_name	gender	annual_income
1	Terrie Bayley	Male	
2	Sukie Callista	Male	60,000
3	Denzil Varsha	Male	
4	Melvin Dotty	Male	35,000
5	Bellamy Suman	Male	
6	Nikita Mari	Male	70,000
7	Arabinda Jagannath	Female	
8	Haru Teona	Female	65,000
9	Yolonda Hana	Female	
10	Chalice Libbie	Female	40,000
11	Pamella Vishnu	Female	
12	Rajeev Meliton	Female	30,000

Ingreso anual promedio

55,000 USD

No cambió nada

45,000 USD



¿Y si hubieran sido  
**otros los valores  
faltantes?**

survey_id	respondant_name	gender	annual_income
1	Terrie Bayley	Male	
2	Sukie Callista	Male	
3	Denzil Varsha	Male	50,000
4	Melvin Dotty	Male	35,000
5	Bellamy Suman	Male	45,000
6	Nikita Mari	Male	
7	Arabinda Jagannath	Female	55,000
8	Haru Teona	Female	65,000
9	Yolonda Hana	Female	
10	Chalice Libbie	Female	
11	Pamella Vishnu	Female	45,000
12	Rajeev Meliton	Female	

Ingreso anual  
promedio

survey_id	respondant_name	gender	annual_income
1	Terrie Bayley	Male	
2	Sukie Callista	Male	
3	Denzil Varsha	Male	50,000
4	Melvin Dotty	Male	35,000
5	Bellamy Suman	Male	45,000
6	Nikita Mari	Male	
7	Arabinda Jagannath	Female	55,000
8	Haru Teona	Female	65,000
9	Yolonda Hana	Female	
10	Chalice Libbie	Female	
11	Pamella Vishnu	Female	45,000
12	Rajeev Meliton	Female	

Ingreso anual promedio

43,333 USD

Se llegó a la conclusión contraria

55,000 USD



¿Se puede explicar  
la ausencia de los  
valores faltantes?

survey_id	respondant_name	gender	satisfied	annual_income
1	Arabinda Jagannath	Male	No	
2	Haru Teona	Male	No	
3	Yolonda Hana	Male	Yes	50,000
4	Chalice Libbie	Male	Yes	35,000
5	Pamella Vishnu	Male	Yes	45,000
6	Rajeev Meliton	Male	No	
7	Terrie Bayley	Female	Yes	55,000
8	Sukie Callista	Female	Yes	65,000
9	Denzil Varsha	Female	No	
10	Melvin Dotty	Female	No	
11	Bellamy Suman	Female	Yes	45,000
12	Nikita Mari	Female	No	

Ingreso anual promedio

43,333 USD

55,000 USD

survey_id	respondant_name	gender	satisfied	annual_income
1	Arabinda Jagannath	Male	No	
2	Haru Teona	Male	No	
3	Yolonda Hana	Male	Yes	50,000
4	Chalice Libbie	Male	Yes	35,000
5	Pamella Vishnu	Male	Yes	45,000
6	Rajeev Meliton	Male	No	
7	Terrie Bayley	Female	Yes	55,000
8	Sukie Callista	Female	Yes	65,000
9	Denzil Varsha	Female	No	
10	Melvin Dotty	Female	No	
11	Bellamy Suman	Female	Yes	45,000
12	Nikita Mari	Female	No	

Ingreso anual promedio

43,333 USD

Se llegó a una conclusión válida dentro de la categoría satisfied = "Yes"

55,000 USD





¿Y si los valores  
faltantes **hubiesen**  
**sido los contrarios?**

survey_id	respondant_name	gender	satisfied	annual_income	Ingreso anual promedio
1	Arabinda Jagannath	Male	No	70,000	
2	Haru Teona	Male	No	60,000	
3	Yolonda Hana	Male	Yes		
4	Chalice Libbie	Male	Yes		
5	Pamella Vishnu	Male	Yes		
6	Rajeev Meliton	Male	No	70,000	
7	Terrie Bayley	Female	Yes		
8	Sukie Callista	Female	Yes		
9	Denzil Varsha	Female	No	35,000	
10	Melvin Dotty	Female	No	40,000	
11	Bellamy Suman	Female	Yes		
12	Nikita Mari	Female	No	30,000	

survey_id	respondant_name	gender	satisfied	annual_income
1	Arabinda Jagannath	Male	No	70,000
2	Haru Teona	Male	No	60,000
3	Yolonda Hana	Male	Yes	
4	Chalice Libbie	Male	Yes	
5	Pamella Vishnu	Male	Yes	
6	Rajeev Meliton	Male	No	70,000
7	Terrie Bayley	Female	Yes	
8	Sukie Callista	Female	Yes	
9	Denzil Varsha	Female	No	35,000
10	Melvin Dotty	Female	No	40,000
11	Bellamy Suman	Female	Yes	
12	Nikita Mari	Female	No	30,000

Ingreso anual promedio

66,666 USD

Se llegó a una conclusión válida dentro de la categoría satisfied = "No"

35,000 USD

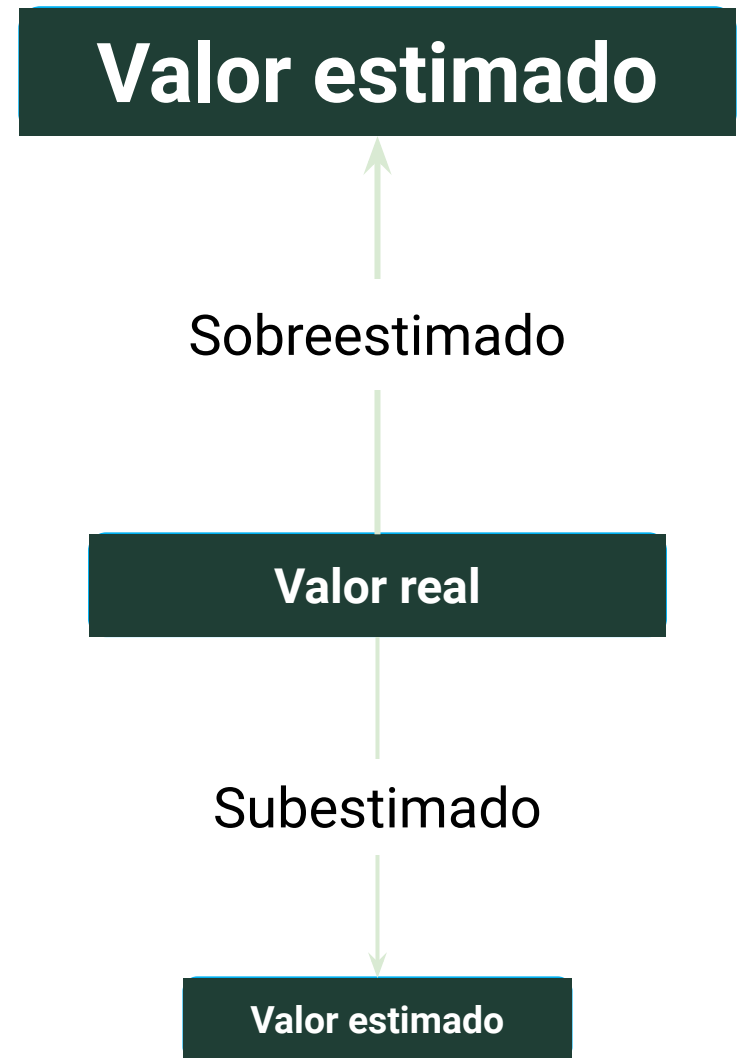


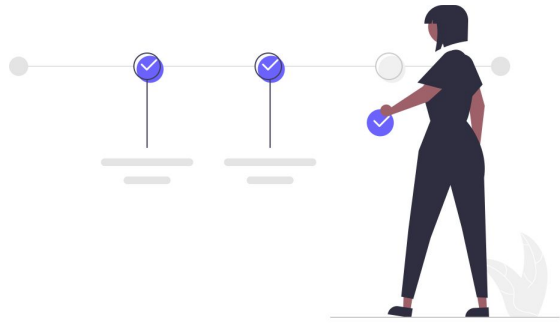
¿Y si los valores faltantes  
tuvieran un **comportamiento**  
**más complejo?**



¿Por qué deberías  
**explorar y lidiar con  
los valores faltantes?**

Ignorar a los valores faltantes puede introducir **sesgos** en tus **análisis** y **modelos**.



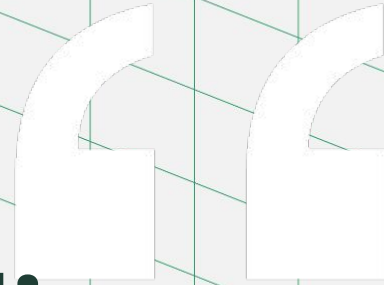


Múltiples de  
los **algoritmos**  
disponibles **fallarán.**

**...los colegios de cartógrafos  
levantaron un mapa del  
Imperio, que tenía el tamaño  
del Imperio y coincidía  
puntualmente con él.**


*Jorge Luis Borges - Del Rigor en la Ciencia*





**Entendieron que ese  
dilatado mapa era inútil...**

*Jorge Luis Borges - Del Rigor en la Ciencia*



**Obviamente, la mejor  
manera de tratar los datos  
que faltan es no tenerlos.**

*Woodbury, M. A. (1970). A missing information  
principle: theory and applications. Duke University  
Medical Center Durham United States.*

# Operaciones con valores faltantes



$$2 + 2 = 4$$

$$2 + "" = ?$$

$$2 + "" = ?$$



```
python

import pandas as pd

s = pd.Series([1, 2, 3, 4, 5, None])

s.mean()
#> 3.0
```

Diferentes herramientas, diferentes resultados por defecto.

```
R

s <- c(1, 2, 3, 4, 5, NA)

mean(s)
#> NA
```

**Aprender otro idioma  
no es solo aprender  
diferentes palabras para  
las mismas cosas, sino  
aprender otra forma de  
pensar sobre las cosas.**

*Flora Lewis*



# Conociendo nuestros datasets



# Extendiendo la API de Pandas

Añade tu propio sabor



# pandas ecosystem

Increasingly, packages are being built on top of pandas to address specific needs in data preparation, analysis and visualization. This is encouraging because it means pandas is not only helping users to handle their data tasks but also that it provides a better starting point for developers to build powerful and more focused data tools. The creation of libraries that complement pandas' functionality also allows pandas development to remain focused around its original requirements.

This is an inexhaustive list of projects that build on pandas in order to provide tools in the PyData space. For a list of projects that depend on pandas, see the [Github network dependents for pandas](#) or [search pypi for pandas](#).

We'd like to make it easier for users to find these projects, if you know of other substantial projects that you feel should be on this list, please let us know.



Utilizando Pandas  
normalmente.



Personalizando  
Pandas para  
analizar tus datos.

# Tabulación de valores faltantes





**Tabular es expresar  
valores, magnitudes  
u otros datos  
por medio de tablas.**



survey_id	respondant_name	gender	satisfied	annual_income
1	Arabinda Jagannath	Male	No	
2	Haru Teona			60,000
3	Yolonda Hana	Male	Yes	
4	Chalice Libbie	Male	Yes	35,000
5	Pamella Vishnu	Male	Yes	
6	Rajeev Meliton	Male	No	70,000
7	Terrie Bayley	Female	Yes	
8	Sukie Callista	Female		65,000
9	Denzil Varsha	Female	No	
10	Melvin Dotty	Female	No	40,000
11	Bellamy Suman	Female	Yes	
12	Nikita Mari	Female	No	30,000

# Empezar con resúmenes simples, como números.

¿Cuántos valores deberían existir en el conjunto de datos?



¿Cuántos **valores faltantes** existen en el conjunto de datos?

¿Cuántos **valores completos** existen en el conjunto de datos?





# Empezar con resúmenes simples, como números.

¿Cuántos valores deberían existir en el conjunto de datos?

¿Cuántos **valores faltantes** existen en el conjunto de datos?

¿Cuántos **valores completos** existen en el conjunto de datos?



# Construir resúmenes por variables y observaciones

¿Cuántos **valores faltantes** existen por cada **variable**?

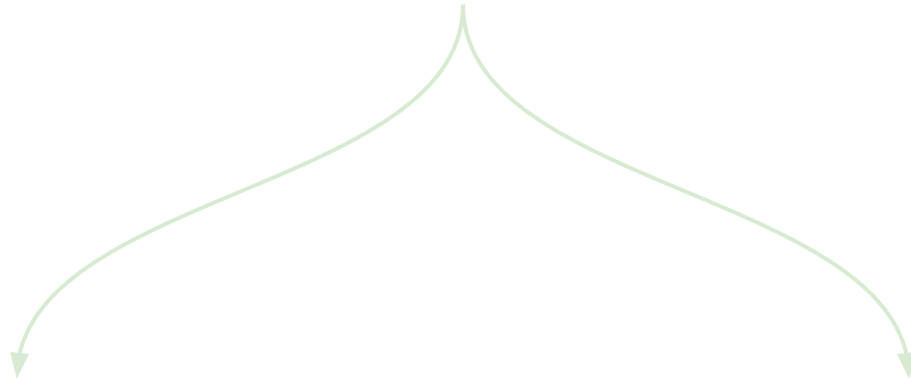
¿Cuántos **valores faltantes** existen por cada **observación**?

¿Cuántas **variables** tiene **X** número de valores faltantes?

¿Cuántas **observaciones** tienen **X** número de valores faltantes?



# Salir de la caja y hacer más preguntas



¿Cuántos **valores faltantes** tengo en una **variable** cada **X pasos**?

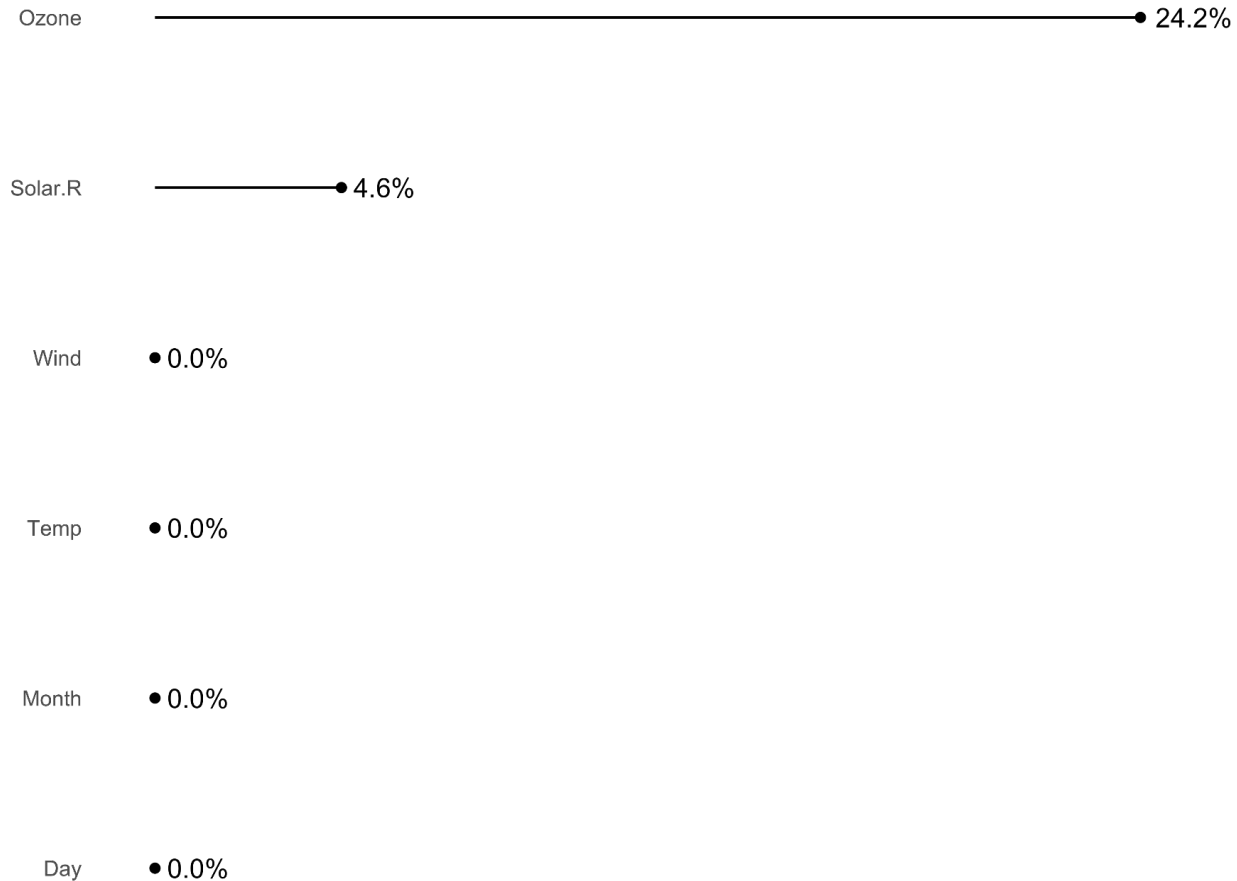
¿Cuál es mi **racha** de valores completos y faltantes en una **variable**?



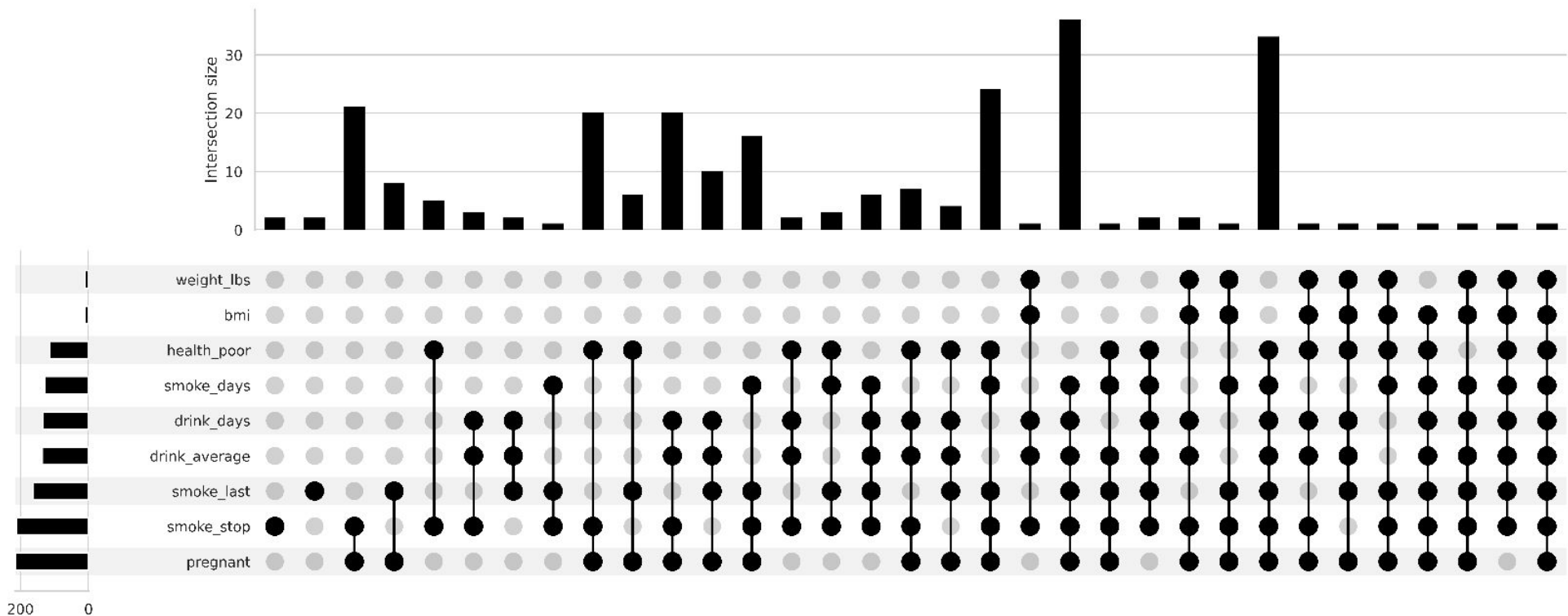
# Visualización de valores faltantes



# Porcentaje de valores faltantes por cada variable



# Apariciones conjuntas de valores faltantes



# Codificación de valores faltantes





**NA**



-99

**N/A**


**None**

**NA**

**Missing**

**-1**

Not  
Available



**Asumir que los valores  
faltantes siempre  
vendrán en un único  
formato es un error.**

# Conversión de valores faltantes implícitos en explícitos



-99

**N/A**

**None**

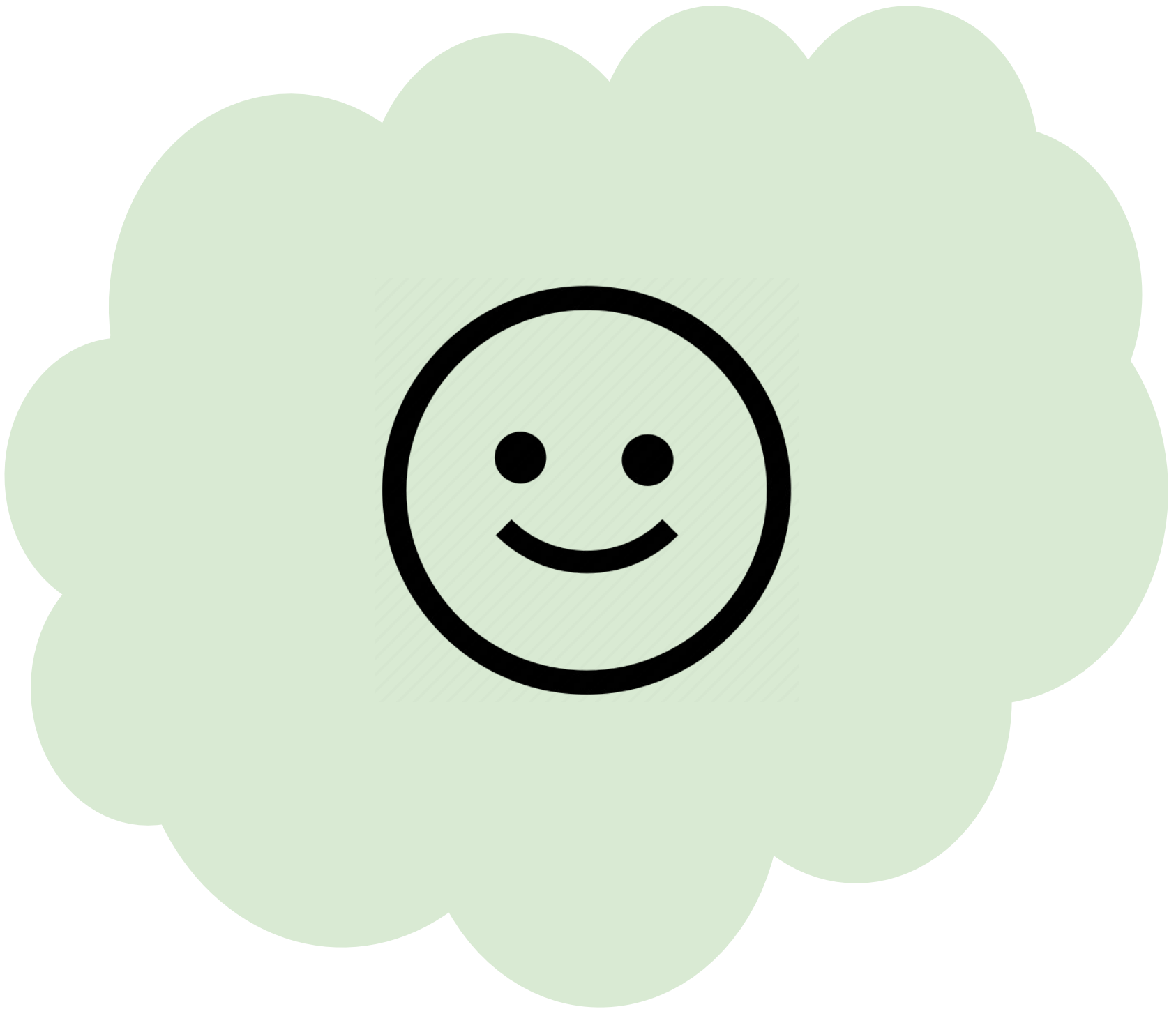
**NA**

**Missing**

**-1**

Not  
Available







# ¿Cuántos valores faltantes existen en la tabla?

name	time	value
lynn	morning	350
lynn	afternoon	310
lynn	night	150
zelda	morning	320





# ¿Cuántos valores faltantes existen en la tabla?

name	time	value
lynn	morning	350
lynn	afternoon	310
lynn	night	150
zelda	morning	320
zelda	afternoon	
zelda	night	



# Exponer filas faltantes implícitas en explícitas



# Tipos de valores faltantes



**Missing  
Completely  
at Random**

**Missing at  
Random**

**Missing not  
at Random**



# Missing Completely at Random (MCAR)

*Faltan completamente al azar*





En algunas ocasiones las herramientas dejan de funcionar **sin ninguna razón** por detrás.

**La localización de los valores faltantes en el conjunto de datos ocurre completamente al azar, estos no dependen de ningún otro dato.**






# Missing at Random (MAR)

*Faltan al azar*





Las herramientas necesitan  
mantenimiento periódico para asegurar  
su **funcionamiento constante.**



**La localización de los  
valores faltantes en el  
conjunto de datos  
depende de otros  
valores observados.**



# Missing not at Random (MNAR)

*Faltan no al azar*








Las herramientas tienen límites. Al tratar de hacer seguimiento en zonas fuera de su rango de medición, **se generan valores faltantes.**





**La localización de los  
valores faltantes en el  
conjunto de datos  
dependen de los valores  
faltantes en sí mismos.**



**¿Puedo tener seguridad  
sobre qué mecanismo de  
valores faltantes es  
correcto para mis datos?**

**NO**

**NO**, pero a través de  
análisis y conocimiento  
del tema puedes hacer  
**suposiciones**  
**razonables.**

# MCAR, MAR, MNAR en Python

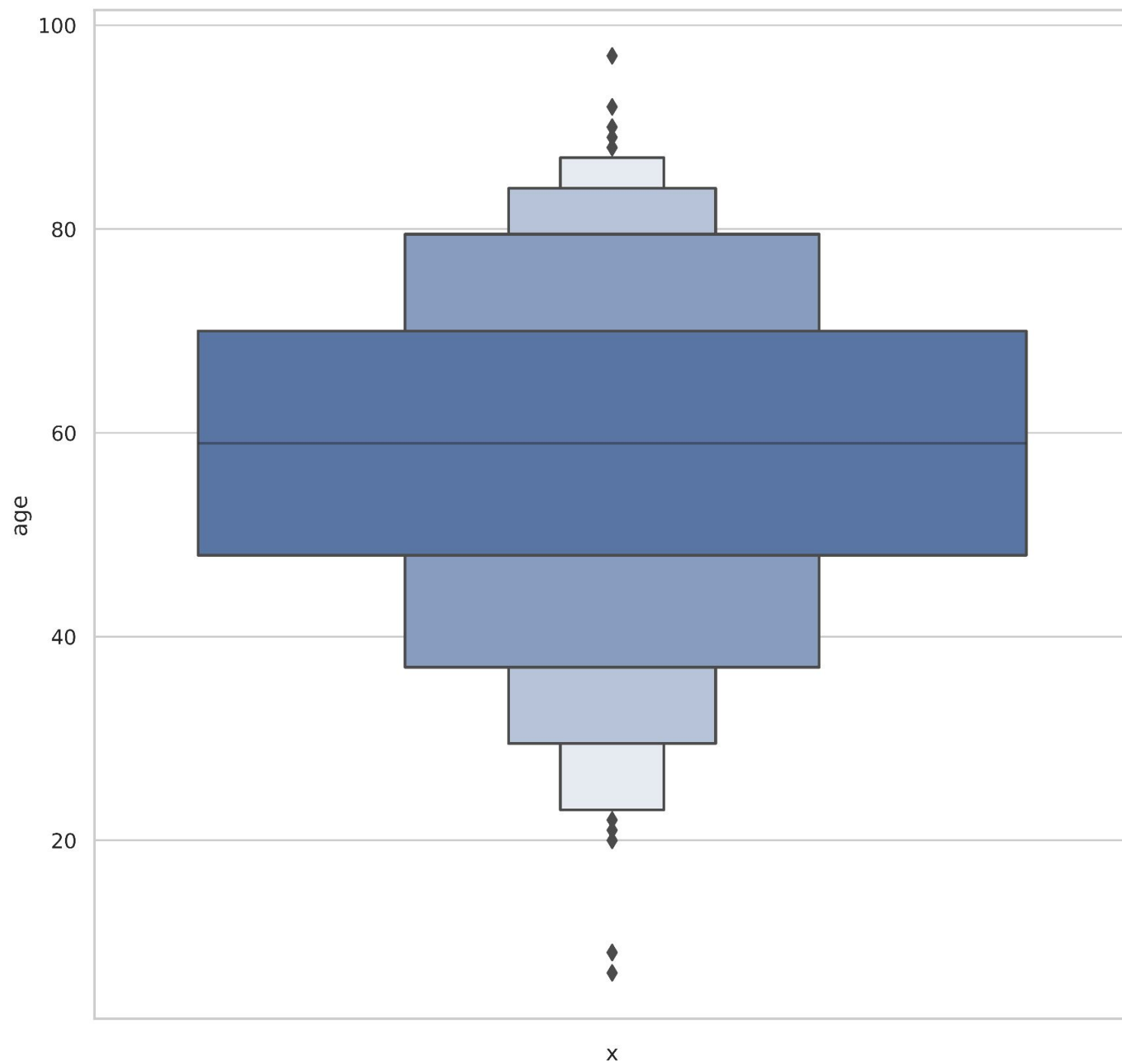
Tipos de valores faltantes

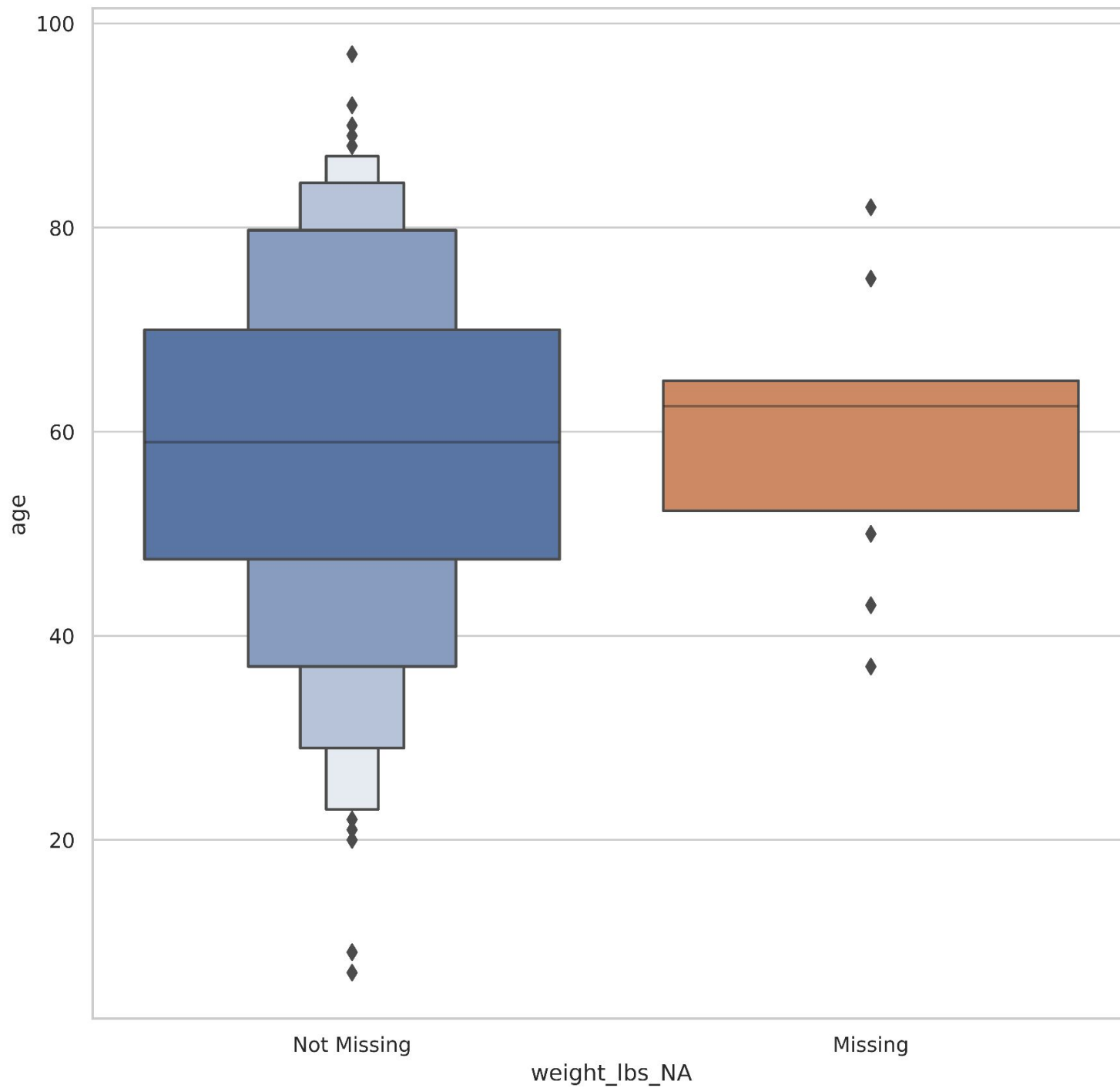


# Matriz de sombras

Shadow Matrix









**La matriz de sombras**

**=**

**The shadow matrix**

1. Identifica todos tus valores faltantes.

name	time	value
lynn	NA	350
NA	afternoon	310
lynn	night	150
zelda	morning	NA

2. Reemplaza los valores faltantes con **True** (1) y el resto con **False** (0).

name	time	value
False	True	False
True	False	False
False	False	False
False	False	True

3. Reemplaza los **True** y **False** por algo que te sea más informativo. Añade un sufijo a los nombres de tus variables.

name_NA	time_NA	value_NA
!NA	NA	!NA
NA	!NA	!NA
!NA	!NA	!NA
!NA	!NA	NA

Características  
de la matriz  
de sombras



Nombres  
coordinados.



Valores  
explícitos.

## Tabla original

name	time	value
lynn	NA	350
NA	afternoon	310
lynn	night	150
zelda	morning	NA



## Matriz de sombras

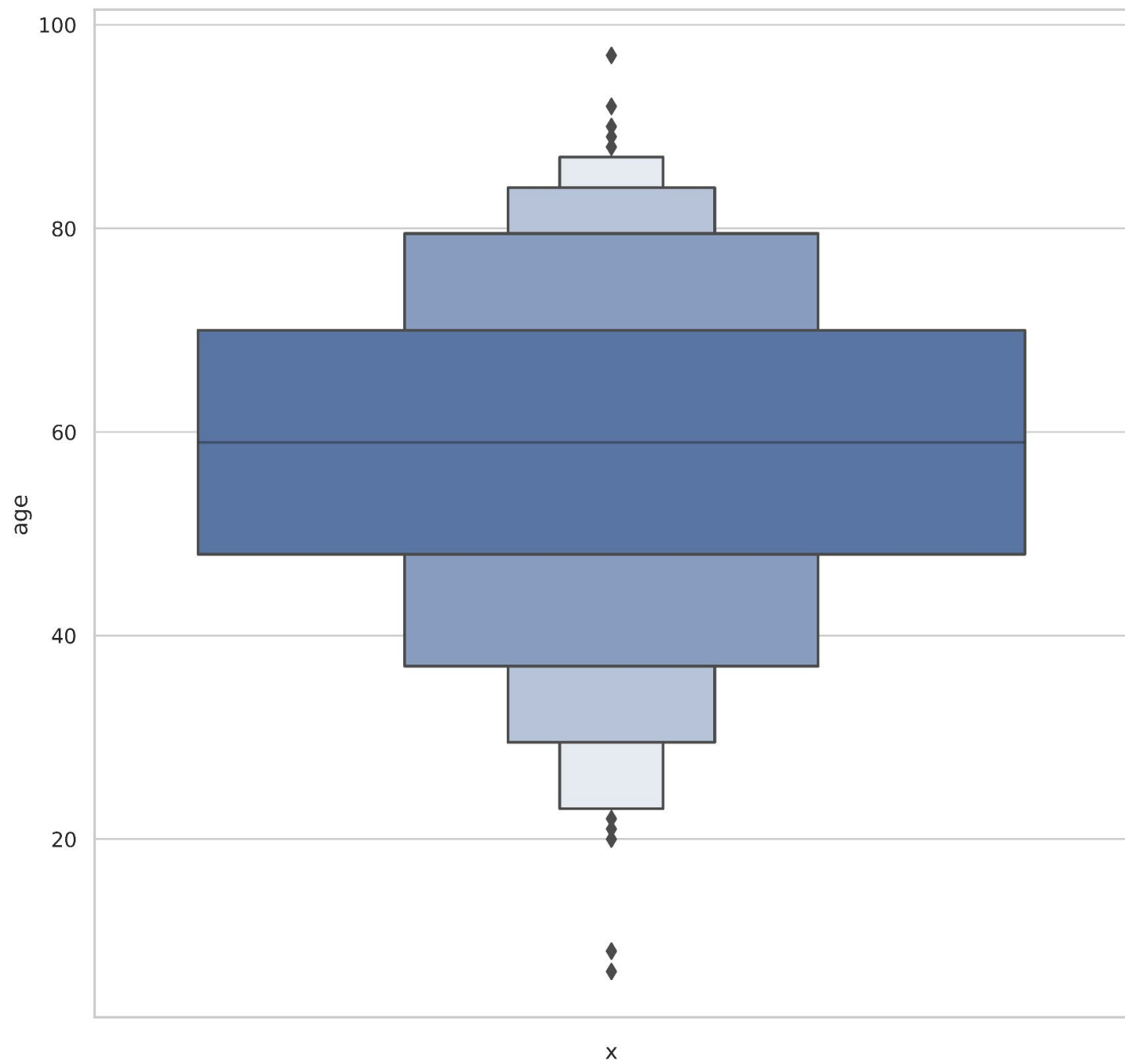
name_NA	time_NA	value_NA
!NA	NA	!NA
NA	!NA	!NA
!NA	!NA	!NA
!NA	!NA	NA

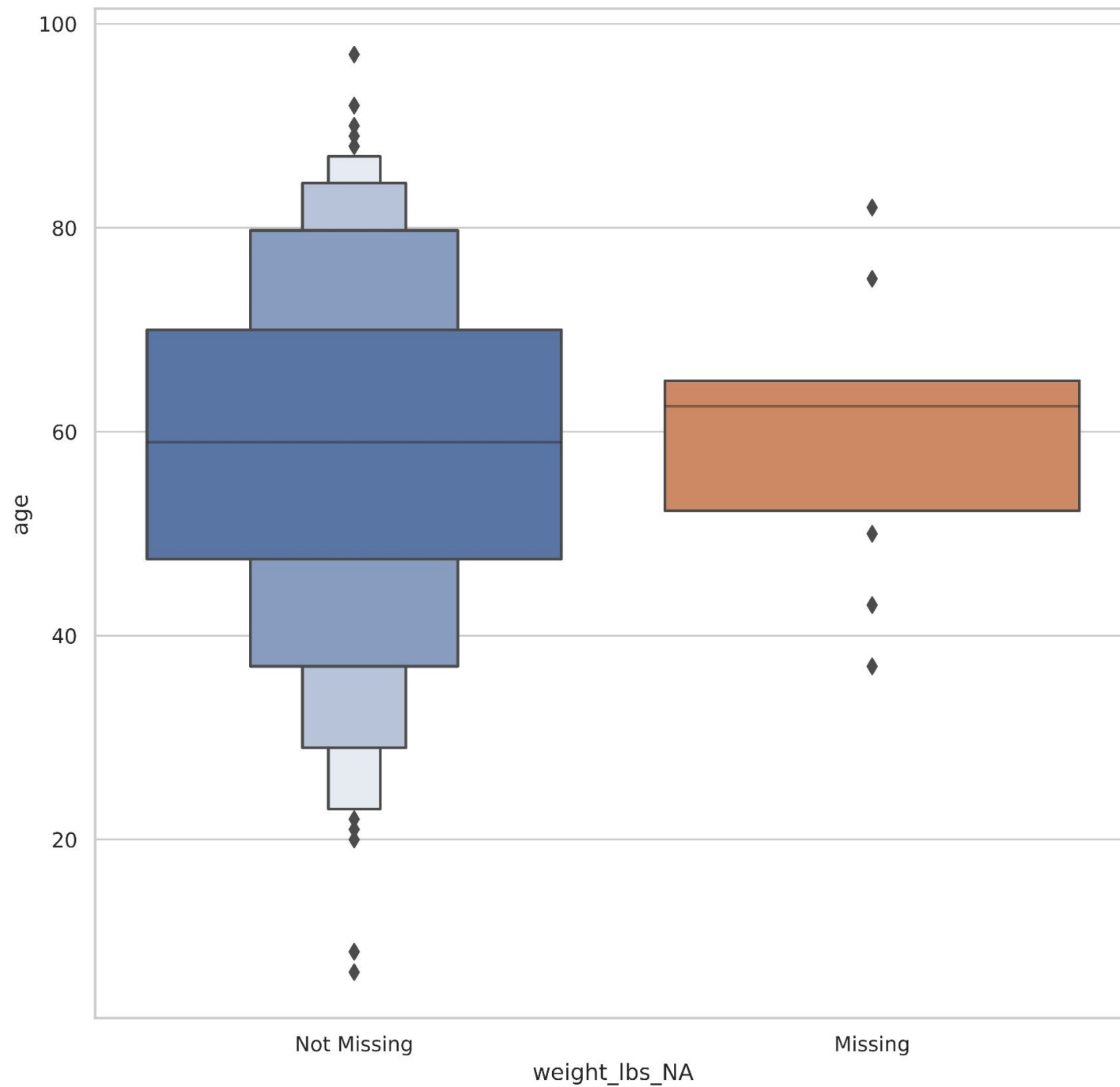
= Nabular

name	time	value	name_NA	time_NA	value_NA
lynn	NA	350	!NA	NA	!NA
NA	afternoon	310	NA	!NA	!NA
lynn	night	150	!NA	!NA	!NA
zelda	morning	NA	!NA	!NA	NA

# Visualización de valores faltantes en una variable



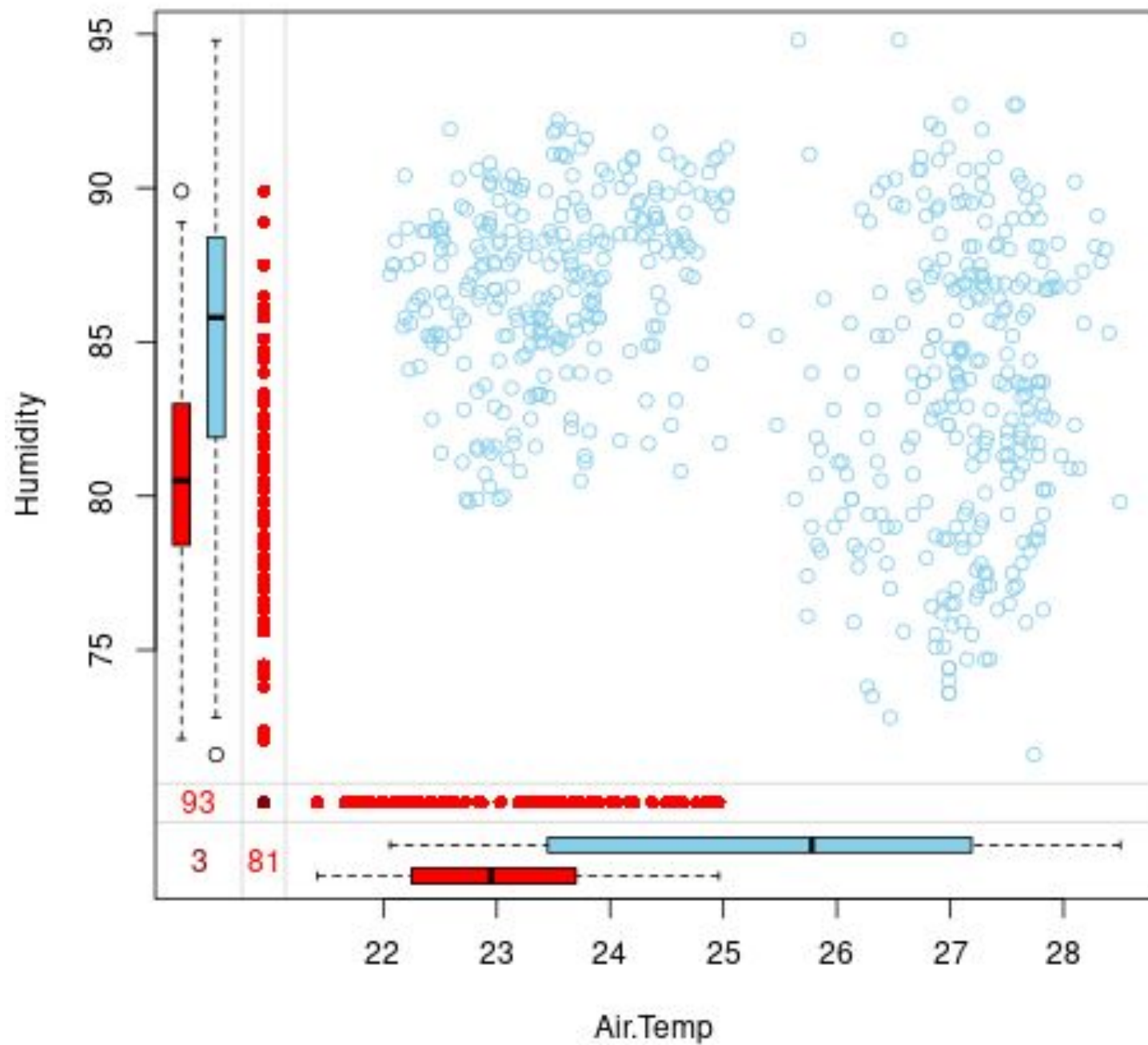




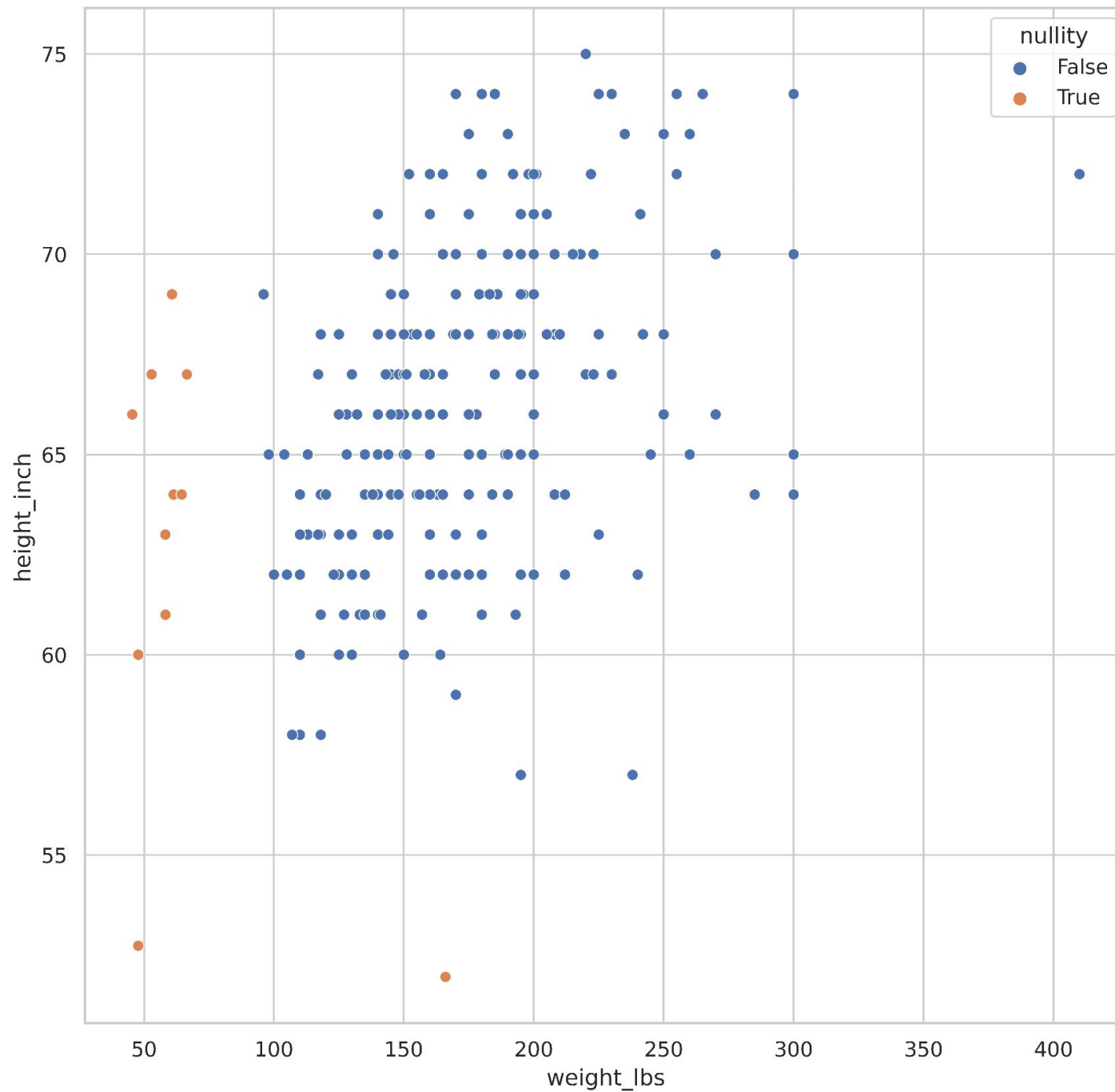
# Visualización de valores faltantes en dos variables

Cómo explorar el concepto con múltiples variables









# Scatterplot con valores faltantes

Con dos variables  
con datos faltantes



# Correlación de nulidad





**¿Existen valores faltantes  
que aparecen conjuntamente  
con otras variables en el  
conjunto de datos?**

# Eliminación de valores faltantes





**NA**







The background of the slide is a light gray isometric grid. The grid lines are thin and green, forming a pattern of rhombi that recede into the distance, creating a three-dimensional effect.

**¿Cómo lograrlo?**

**Eliminación de  
valores faltantes**

**Imputación de  
valores faltantes**

**Eliminación de  
valores faltantes**

**Imputación de  
valores faltantes**

# Imputación básica de datos



**Eliminación de  
valores faltantes**

**Imputación a  
valores faltantes**

**Eliminación de  
valores faltantes**

**Imputación a  
valores faltantes**

# Bonus: visualización múltiple de imputaciones



# Continúa aprendiendo sobre el manejo de valores faltantes







**¿Qué aprendiste  
en este curso?**

# Conclusiones

- Aprendiste sobre las operaciones con valores faltantes.
- Fuiste capaz de detectar valores faltantes con sus distintas codificaciones.
- Lograste encontrar los valores faltantes implícitos y transformarlos en explícitos.



# Conclusiones

- Aprendiste sobre los diferentes tipos de valores faltantes (MCAR, MAR y MNAR).
- Creaste visualizaciones que ayudan a la exploración de patrones de valores faltantes complejos.
- Conociste las bases de la eliminación e imputación de valores faltantes.

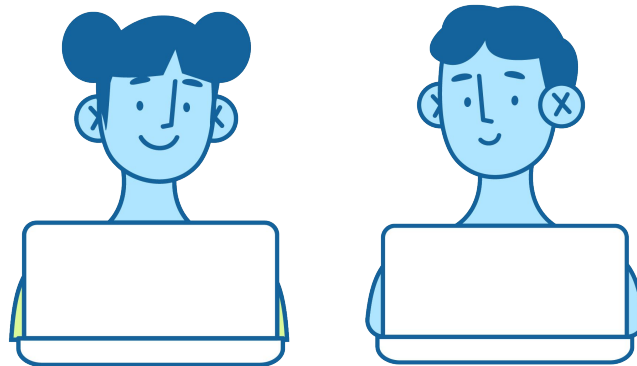




**¿Cómo continuar  
practicando?**

# Continúa explorando realizando un proyecto

- Explora y detecta datos faltantes en otro dataset disponible en el curso.



# Continúa con los cursos

- Curso de Manejo de Datos Faltantes: Imputación.
- Curso de Configuración Profesional de Entorno de Trabajo para Ciencia de Datos.



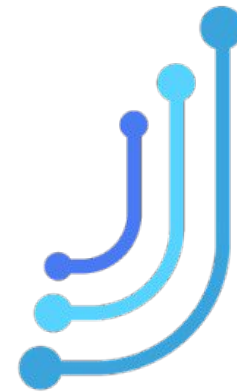
**¡Felicidades!**



# ¡Felicidades!



**@jvelezmagic**



**jvelezmagic.com**

