

# Curso de **Manejo de Datos Faltantes: Imputación**

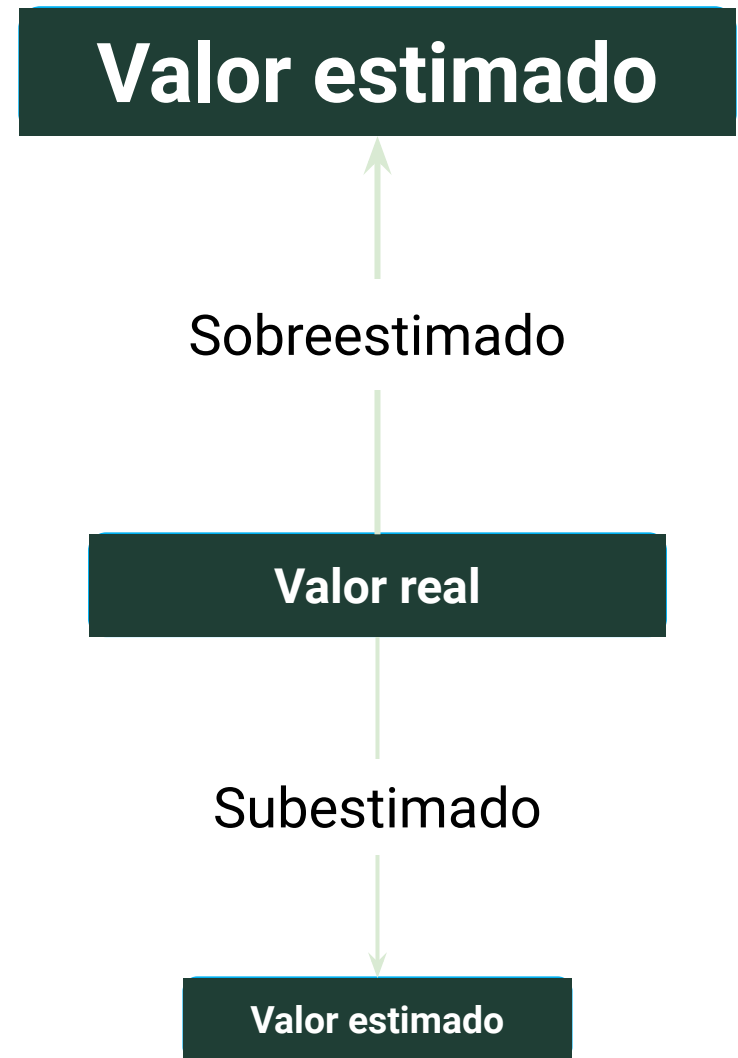
Jesús Vélez | @jvelezmagic

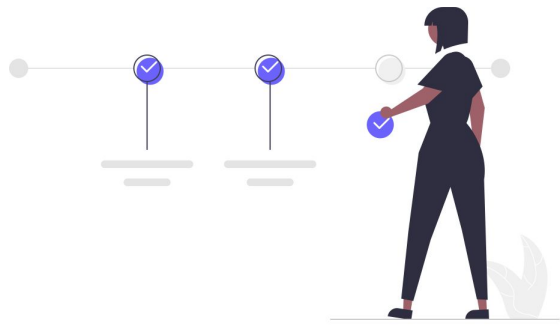


# El problema de trabajar con valores faltantes




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





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



**Estimar los valores  
ausentes con base en  
los valores válidos de  
otras variables y/o casos  
de muestra.**

# Conociendo y limpiando datos antes de imputar

NHANES



# Visualizar y eliminar valores faltantes

NHANES





# Implicaciones de los distintos tipos de valores faltantes

MCAR, MAR, MNAR





**Missing  
Completely  
At Random**

**Missing At  
Random**

**Missing Not  
At Random**



# Missing Completely At Random (MCAR)

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

# Implicaciones - MCAR

## Eliminación de valores faltantes


- Reducción del tamaño de muestra.
- Inferencia limitada.
- No produce sesgos.

## Imputación de valores faltantes

- De hacerlo bien, no introduce sesgos.
- La imputación es recomendada sobre la delección.



# Missing At Random (MAR)



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

# Implicaciones - MAR



## Eliminación de valores faltantes

- Ignorarlos **produce sesgos.**


## Imputación de valores faltantes

- La mayor parte de **métodos** de imputación **asumen MAR.**
- La imputación es **necesaria.**





# Missing Not At Random (MNAR)



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

# Implicaciones - MNAR

## Eliminación de valores faltantes

Ignorarlos **produce sesgos.**

## Imputación de valores faltantes

La imputación es recomendada sobre la delección.

# Implicaciones - MNAR

```
graph TD; A[Implicaciones - MNAR] --> B[Eliminación de valores faltantes]; A --> C[Imputación de valores faltantes]; B --> D[Ignorarlos produce sesgos.]; C --> E[La imputación es recomendada sobre la delección.]; D --> F[Mejorar experimentos o realizar análisis de sensibilidad]; E --> F;
```

**Eliminación de  
valores faltantes**

Ignorarlos **produce  
sesgos.**

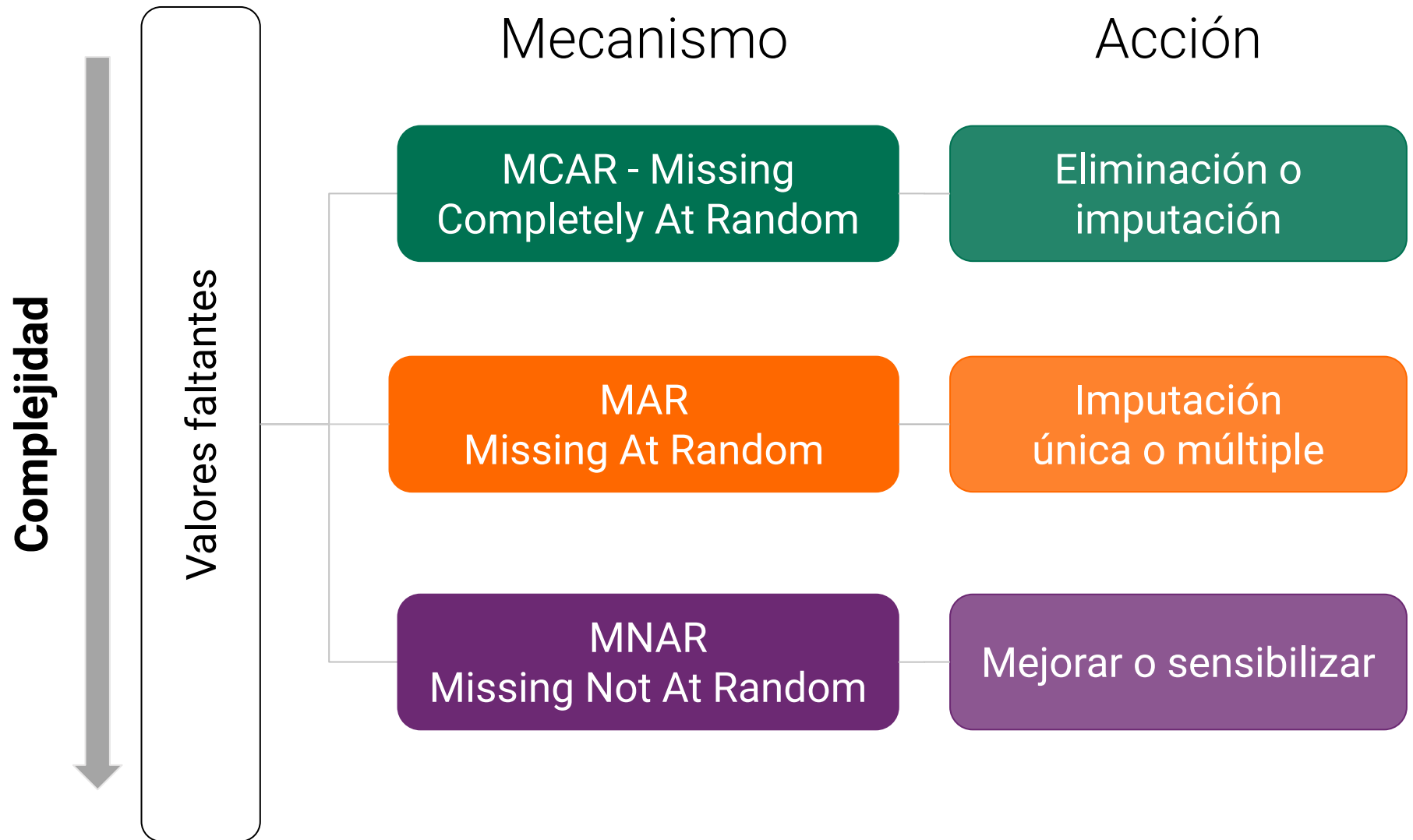
**Imputación de  
valores faltantes**

La imputación es  
recomendada sobre  
la delección.

**Mejorar experimentos  
o realizar análisis de sensibilidad**



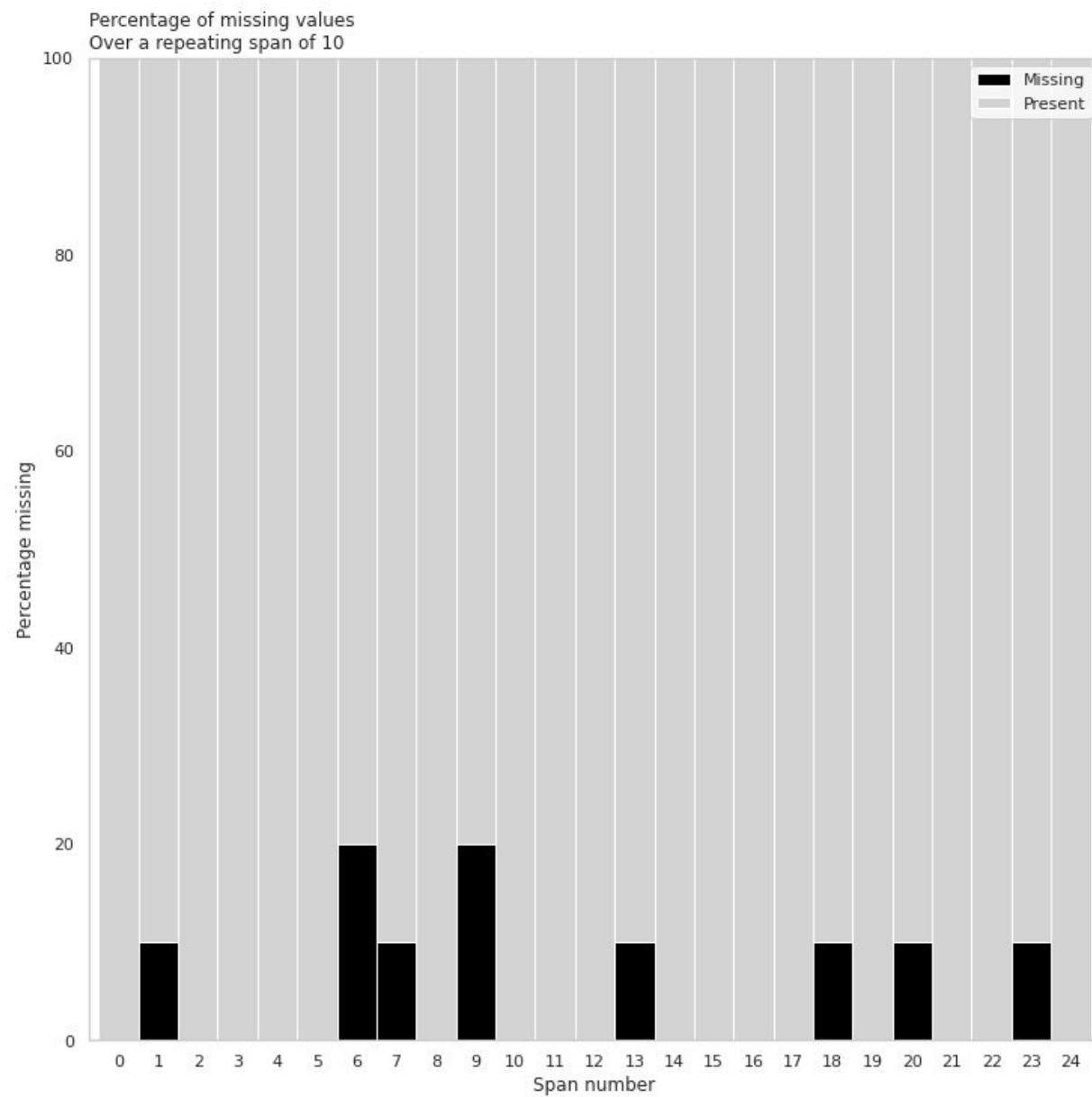
# En resumen

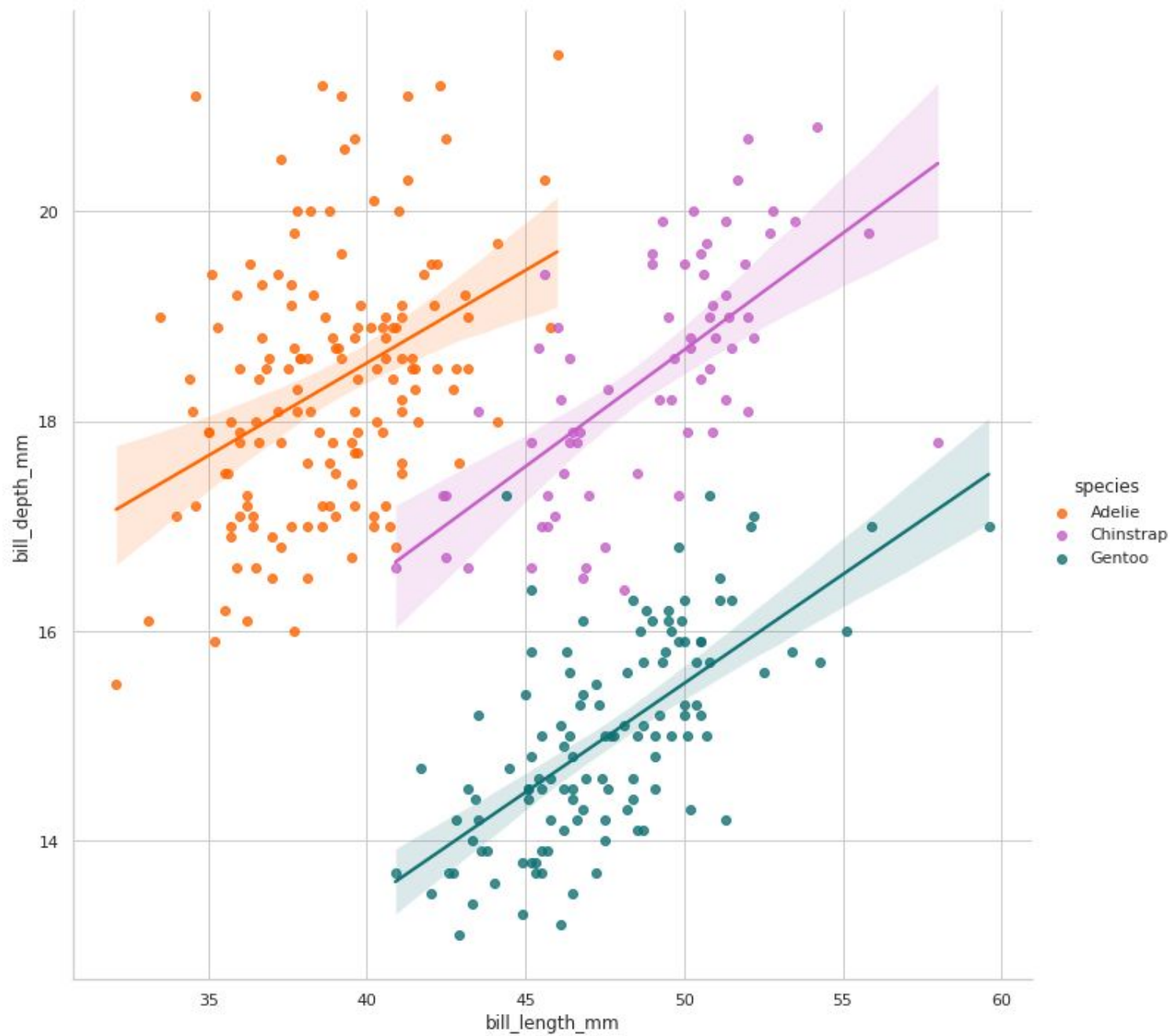


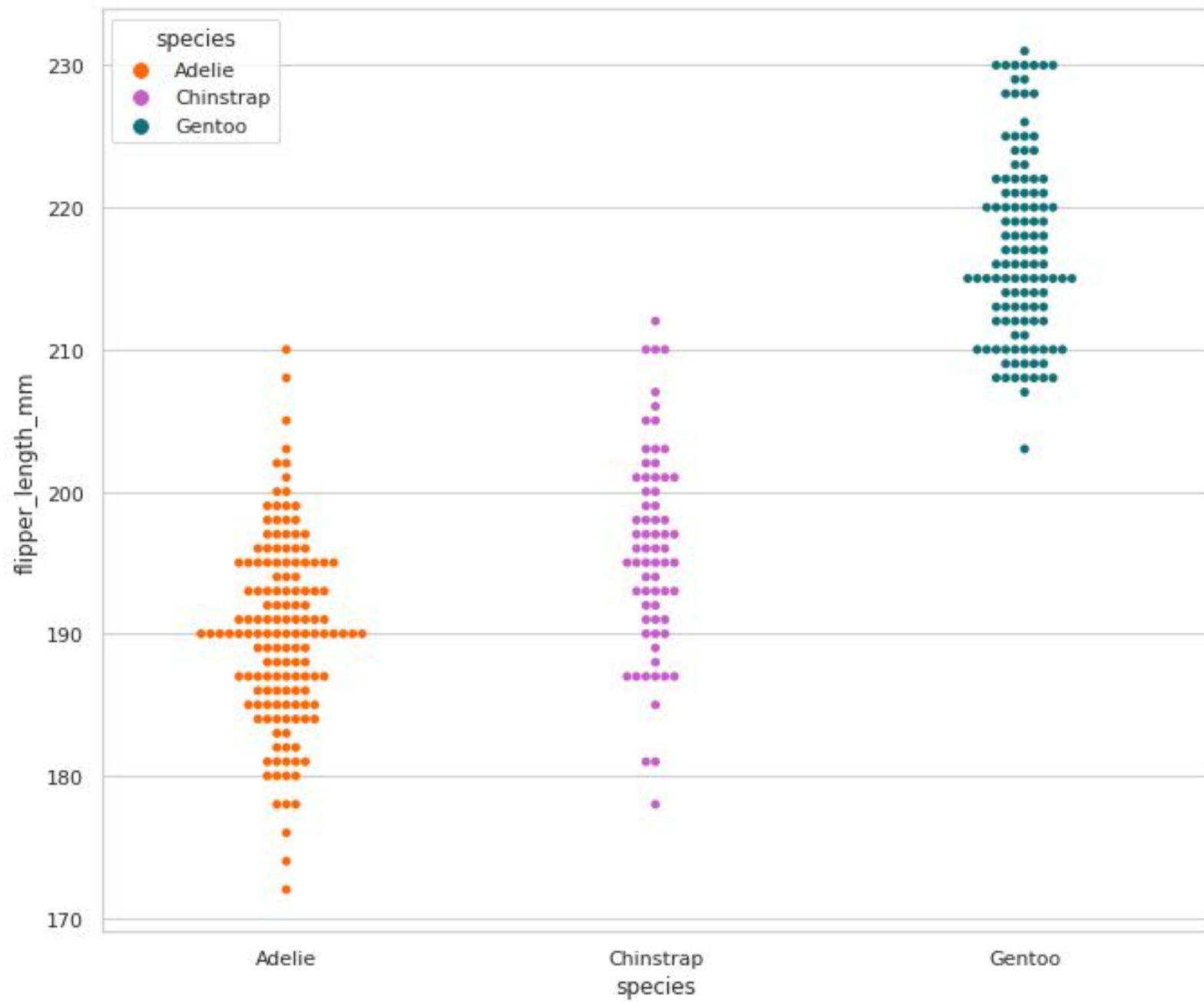
**Amplía tu conjunto  
de herramientas  
para explorar  
valores faltantes**

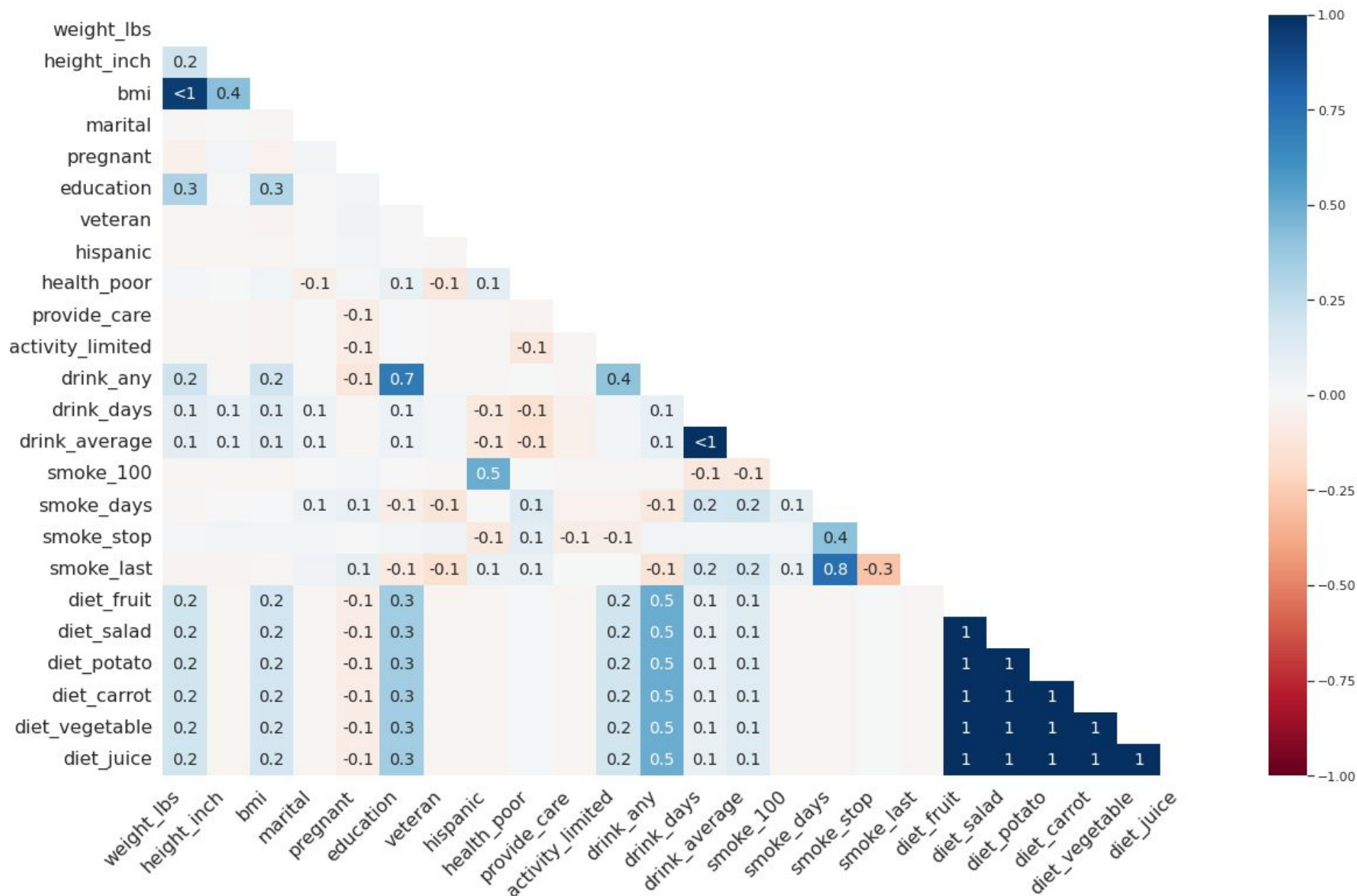














# Tratamiento de variables categóricas para imputación

Codificación ordinal



La mayor parte de las variables categóricas son cadenas de texto.



Realizar operaciones con cadenas de texto  
**NO es trivial.**



Surge la necesidad de **convertir o codificar** las cadenas de texto a números.





# Codificación ordinal

# ¿Cuál es tu animal preferido?

<b>animal</b>
perro
gato
perro
gato
gato
beluga
perezoso
gato
Turritopsis nutricula



# ¿Cuál es tu animal preferido?

animal
perro
gato
perro
gato
gato
beluga
perezoso
gato
Turritopsis nutricula

animal	valor
perro	0
gato	1
perro	0
gato	1
gato	1
beluga	2
perezoso	3
gato	1
Turritopsis nutricula	4

Ordinal Encoder  
o Codificación  
Ordinal

animal	valor
perro	0
gato	1
beluga	2
perezoso	3
Turritopsis nutricula	4



# Tratamiento de variables categóricas para imputación

One - Hot Encoding



# ¿Cuál es tu animal preferido?

One - Hot Encoding

animal
Perro
Gato
Perro
Gato
Gato
Beluga
Perezoso
Gato
Turritopsis nutricula

animal	Perro	Gato	Beluga	Perezoso	Turritopsis nutricula
Perro	1	0	0	0	0
Gato	0	1	0	0	0
Perro	1	0	0	0	0
Gato	0	1	0	0	0
Gato	0	1	0	0	0
Beluga	0	0	1	0	0
Perezoso	0	0	0	1	0
Gato	0	1	0	0	0
Turritopsis nutricula	0	0	0	0	1



# Métodos de imputación de valores faltantes



# Tratamiento de valores faltantes

```
graph TD; A[Tratamiento de valores faltantes] --> B[Deleciones / Eliminaciones]; A --> C[Imputaciones]
```

Deleciones /  
Eliminaciones

Imputaciones



## Deleciones / Eliminaciones

```
graph TD; A[Deleciones / Eliminaciones] --> B["Pairwise deletion  
(Eliminación por parejas)"]; A --> C["Listwise deletion  
(Eliminación por filas)"]; A --> D["Eliminación completa  
de columnas"]; B --> E["Elimina únicamente  
los valores faltantes"]; C --> F["Elimina las filas con  
valores faltantes"]; D --> G["Elimina las columnas  
con valores faltantes"];
```

*Pairwise deletion*  
(Eliminación por  
parejas)

Elimina únicamente  
los valores faltantes

*Listwise deletion*  
(Eliminación por filas)

Elimina las filas con  
valores faltantes

Eliminación completa  
de columnas

Elimina las columnas  
con valores faltantes

# Imputaciones

General

Avanzada

Datos que no son series  
de tiempo

Datos que son series  
de tiempo

Imputar con una  
constante

Imputar con media,  
mediana o moda

Llenado hacia  
atrás

Llenado hacia  
adelante

Interpolación

KNN

MICE

NN

SVM

Otros  
modelos

# Imputaciones

General

Avanzada

Datos que no son series de tiempo

Datos que son series de tiempo

Imputar con una constante



Imputar con media, mediana o moda



Llenado hacia atrás



Llenado hacia adelante



Interpolación

KNN



MICE



NN



SVM



Otros modelos





**¿Qué son las  
imputaciones con  
base en el donante?**

**Completa los valores que faltan para una unidad dada copiando los valores observados de otra unidad, el donante.**

*Donor imputation (theme). (2013, October 22).  
CROS - European Commission.*



**¿Qué son las  
imputaciones con  
base en modelos?**

**El objetivo de la imputación basada en modelos es encontrar un modelo predictivo para cada variable objetivo en el conjunto de datos que contiene valores faltantes.**

*Model-Based imputation (theme). (2013, October 22). CROS - European Commission.*

# Imputaciones

General

Avanzada

Datos que no son series de tiempo

Datos que son series de tiempo

Imputar con una constante



Imputar con media, mediana o moda



Llenado hacia atrás



Llenado hacia adelante



Interpolación

KNN



MICE



NN



SVM



Otros modelos





# Imputación de media, mediana y moda





Puede sesgar los resultados, dado que modifica la distribución por debajo (curtosis).

1

1

Rápido y fácil.

Pierde correlaciones entre variables. No es muy preciso.

2

2

La media puede ser útil en presencia de outliers.

No puede usar variables categóricas (a excepción de la moda).

3

3

No afectará el estadístico en cuestión ni el tamaño de muestra.

# Imputación por llenado hacia atrás y hacia adelante





Relaciones multivariadas pueden ser distorsionadas.

1

1

Rápido y fácil.

2

Los datos imputados no son constantes.

3

Existen trucos para evitar romper las relaciones entre variables.

# Imputación por interpolación



# Imputaciones

General

Avanzada

Datos que no son series de tiempo

Datos que son series de tiempo

Imputar con una constante



Imputar con media, mediana o moda



Llenado hacia atrás



Llenado hacia adelante



Interpolación

KNN



MICE



NN



SVM



Otros modelos





Puede romper relaciones entre variables.

1

1

Sencillo de implementar.

Puede introducir valores fuera de rango.

2

2

Útil para series de tiempo.

3

Variabilidad de opciones al alcance.

# Imputación por algoritmo de K-vecinos más cercanos

KNN





# Pasos para imputación por **k-Nearest-Neighbors**

**Para cada observación con valores faltantes:**

1. Encuentra otras **K** observaciones (donadores, vecinos) que sean más similares a esa observación.
2. Reemplaza los valores faltantes con los valores agregados de los **K** vecinos.

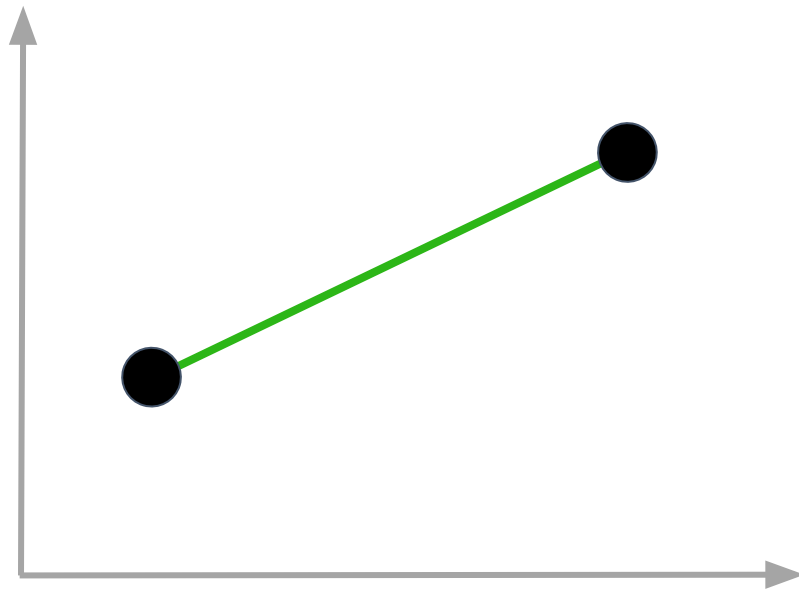




**¿Cómo determinar  
cuáles son los vecinos  
más similares?**

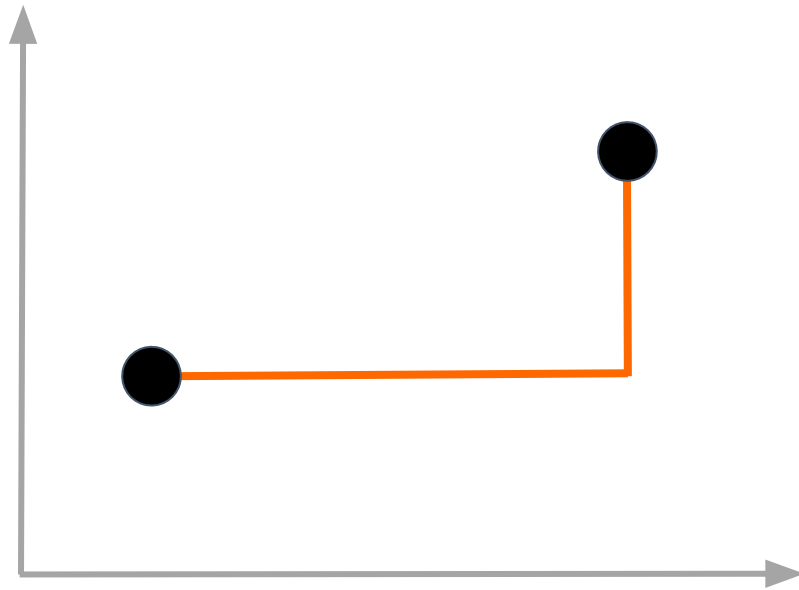
# Cuantificación de distancia: **distancia euclidiana**

Útil para variables numéricas.



# Cuantificación de distancia: **distancia Manhattan**

Útil para variables tipo factor.



# Cuantificación de distancia: **distancia de Hamming**

Útil para variables categóricas.

Hola mundo

Hoja masdo



# Cuantificación de distancia: **distancia de Gower**

Útil para conjuntos de datos con variables mixtas.

Variables numéricas					Variables tipo factor			Variables Categóricas	
Distancia Euclidiana					Distancia Manhattan			Distancia Hamming	

**Distancia de Gower**





Su escalabilidad puede ser comprometedora.

1

1

Sencillo de implementar.

Requiere transformaciones especiales para las variables categóricas.

2

2

Buen rendimiento con conjuntos de datos pequeños.

Posee sensibilidad a valores atípicos.

3

3

Excelente para datos numéricos, pero también funciona para datos mixtos.

# Imputación basada en modelos







Puede subestimar la varianza.

1

1

Mejora sobre la imputación basada en donante sencilla.

Los modelos funcionan mal si las variables observadas y faltantes son independientes.

2

2

Gran variedad de opciones para imputar.

Más complicado que la imputación basada en donantes.

3

3

Preservación de relaciones entre variables.

# Imputaciones múltiples por ecuaciones encadenadas

MICE

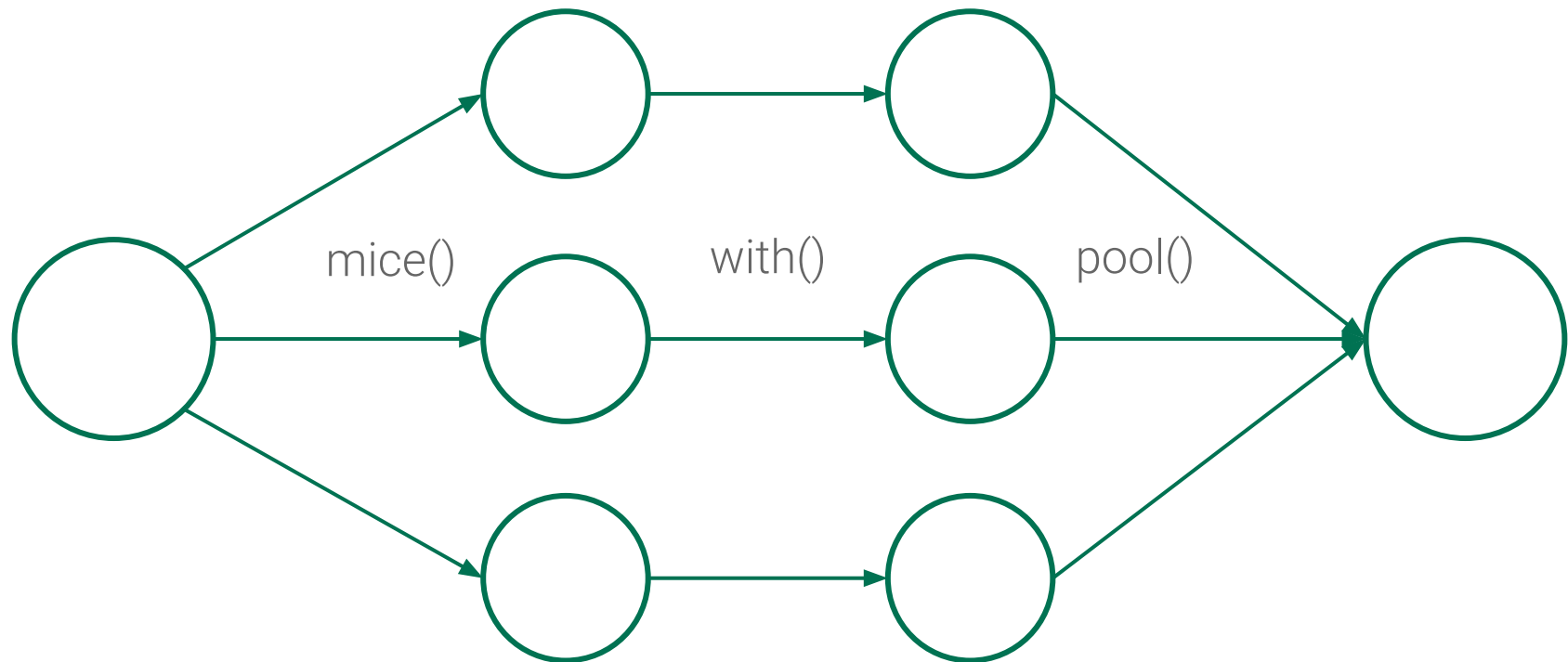


**Datos  
incompletos**

**Datos  
imputados**

**Análisis de  
resultados**

**Resultados  
agrupados**





Para funcionar bien, necesitas pensar en el modelo de imputación y el modelo de análisis.

Solo funciona como los métodos de imputación seleccionados.

1

2

1

2

3

Mantiene la distribución relativa similar antes y después de la imputación.

Puede ser utilizada en cualquier tipo de análisis.

Múltiples variables son imputadas.

# Transformación inversa de los datos



# ¿Cómo continuar practicando?





**¿Qué aprendiste  
en este curso?**

# Conclusiones

- Aprendiste que trabajar con valores faltantes representa un trabajo con tratamiento especial.
- Conociste las consideraciones al trabajar con los distintos tipos de valores faltantes (MCAR, MAR y MNAR).
- Lograste explorar los tipos de valores faltantes a través de visualizaciones y pruebas estadísticas nuevas.





# Conclusiones

- Aprendiste a tratar valores categóricos al momento de realizar imputaciones.
- Lograste identificar los distintos tipos de imputación de valores faltantes: con base en donantes y modelos.
- Realizaste múltiples imputaciones a través de distintos algoritmos.
- Entendiste las ventajas y desventajas de cada herramienta de imputación.





**¿Cómo continuar  
aprendiendo sobre  
valores faltantes?**

# Realiza un proyecto: National Health and Nutrition Examination Survey

8  
variables



**197  
variables**



# Realiza un proyecto:

## National Health and Nutrition Examination Survey

```
nhanes_raw_df.select_columns("*activi*")
```

- vigorous\_work\_activity
- moderate\_work\_activity
- vigorous\_recreational\_activities
- moderate\_recreational\_activities
- minutes\_sedentary\_activity



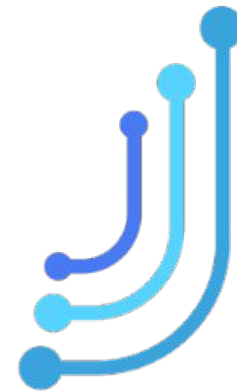
# ¡Felicidades!



# ¡Felicidades!



**@jvelezmagic**



**jvelezmagic.com**

