FACULTAD DE INGENIERÍA ECONÓMICA, ESTADÍSTICA Y CIENCIAS SOCIALES

Data Advanced Analytics

Programas de Especialización en

Business Inteligence and Data Analytics

Evelyn Gutierrez

(egutierreza@pucp.edu.pe)

Tratamiento de datos perdidos

Evelyn Gutierrez (egutierreza@pucp.edu.pe)

Agenda: Datos Perdidos

- Exploración
  - Tipos de patrones.
  - Análisis de patrones.
- Tratamiento.
  - Eliminación de casos.
  - Imputación
    - Univariada/Multivariada.
    - Simple/Múltiple.

Datos Perdidos

- En R, los datos perdidos se denotan como NA (*Not Available*)

```
##    nombre nota
## 1   Jesus   12
## 2   Carla   15
## 3 Rodrigo   13
## 4  Javier   NA
```

¿Por qué tenemos datos perdidos?

- Errores en la recolección de datos.
- No respuesta a preguntas sensibles (Ejemplo: ingresos).
  - Esto suele ser común en ciencias sociales.

Clasificación de los datos perdidos:

- Antes de lanzarnos a aplicar un tratamiento es importante considerar qué tipo de patrón encontramos en los datos perdidos.

MCAR

Missing completely at random

MAR

Missing at random

MNAR

Missing Not At Random

Rubin. 1976. “Inference and Missing Data.” *Biometrika* 63 (3): 581–90.

Clasificación de los datos perdidos:

- **MCAR -** ***Missing completely at random***

- Cada observación tuvo la misma posibilidad de ser un valor perdido.
- Ejemplo: Una balanza deja de funcionar aleatoriamente.

-> Los valores que hemos perdido tendrán la misma distribución que los valores completos.

No observable	Datos
	Peso
2	2
3	3
4	4
3	NA
3	3
2	NA
4	4
3	3
3	NA
3	3
2	2
7	NA
8	8
9	NA
7	7
6	7

Clasificación de los datos perdidos:

- **MAR -** ***Missing at random***

- Los valores perdidos dependen de alguna(s) de la(s) variable(s) observada(s).
- Ejemplo:
  - Una balanza deja de funcionar más frecuentemente cuando la superficie es blanda.

    Observar:
    - Dentro de cada categoría, los NA están presentes aleatoriamente (MCAR).
    - El comportamiento de estos NA será similar al de los datos no perdidos del mismo grupo.

Datos		No observable
Superficie	Peso	
Sólida	4	4
Sólida	5	5
Sólida	6	6
Sólida	NA	4
Sólida	5	5
Sólida	4	4
Sólida	4	4
Sólida	6	6
Sólida	NA	4
Sólida	5	5
Blanda	NA	2
Blanda	3	3
Blanda	2	2
Blanda	NA	3
Blanda	1	1
Blanda	NA	1
Blanda	2	2
Blanda	NA	2
Blanda	NA	2
Blanda	3	3
Blanda	NA	2
Blanda	1	1
Blanda	3	3

Clasificación de los datos perdidos:

- **MNAR (*Missing Not At Random*)**

- Los valores perdidos dependen de la variable que estamos analizando.
- Ejemplo:
  - Una balanza deja de funcionar más frecuentemente cuando los pesos son altos.

Observar:

- Imputar datos sería difícil en este caso porque no sabemos cómo se comportan los datos perdidos (Ya no son una muestra aleatoria de ningún grupo).

Peso no observable	Data
	Peso
2	2
3	3
4	4
3	3
3	3
2	2
4	4
3	3
5	5
8	NA
7	7
6	NA
8	NA
9	NA
7	NA
6	NA

Clasificación de los datos perdidos:

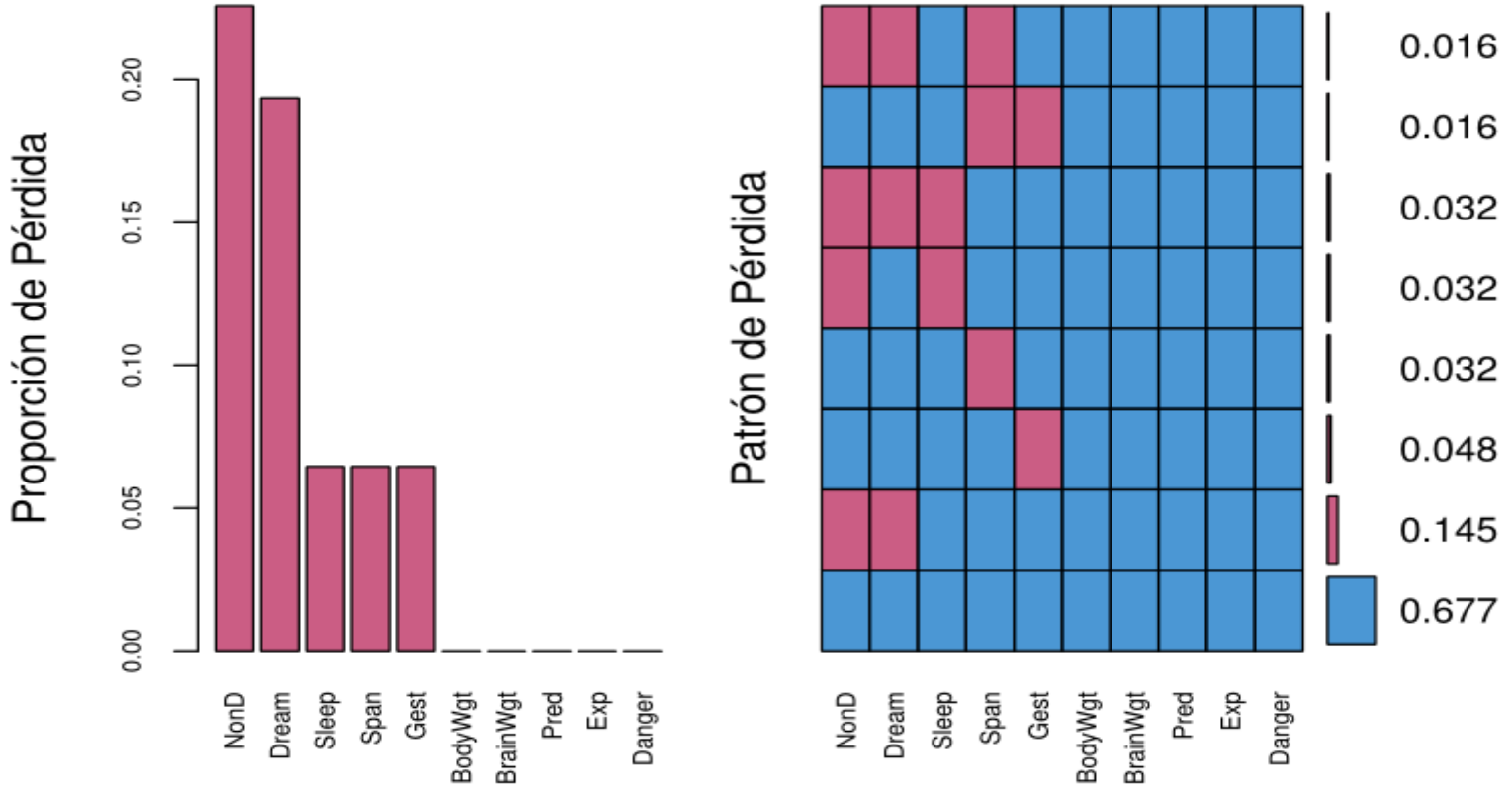
1. **MCAR (Missing Completely At Random) -** Datos perdidos completamente aleatorios.
   - No existe relación entre las observaciones perdidas y las covariables observadas
   - Si nos encontramos en este caso, se puede emplear análisis de casos completos, imputación múltiple o cualquier otro método de imputación.
   - Eliminar casos con datos faltantes no sesga las inferencias.
2. **MAR (Missing at Random)** - Datos perdidos aleatorios.
   - La probabilidad de que una variable tenga valores perdidos depende únicamente de la información disponible (variables observadas).
   - En esta situación, es válido emplear imputación múltiple. El método resulta ser sesgado, pero con un sesgo que es considerado despreciable.
3. **MNAR (Missing Not At Random) -** Datos perdidos no aleatorios
   - La información faltante depende de los valores de los datos perdidos.
   - Los llamados datos censurados, o 'censored data', pertenecen a esta categoría.
   - Se debe considerar la información perdida explícitamente, modelando conjuntamente los valores observados y los valores perdidos. Existen metodologías ya conocidas cuando se trata por ejemplo de un Análisis de Supervivencia. Caso contrario, se debe aceptar que habrá algún sesgo en las inferencias.

Explorando a los datos perdidos

Evelyn Gutierrez (egutierreza@pucp.edu.pe)

- ¿Cuánta data perdida tengo?

Exploración de valores perdidos.



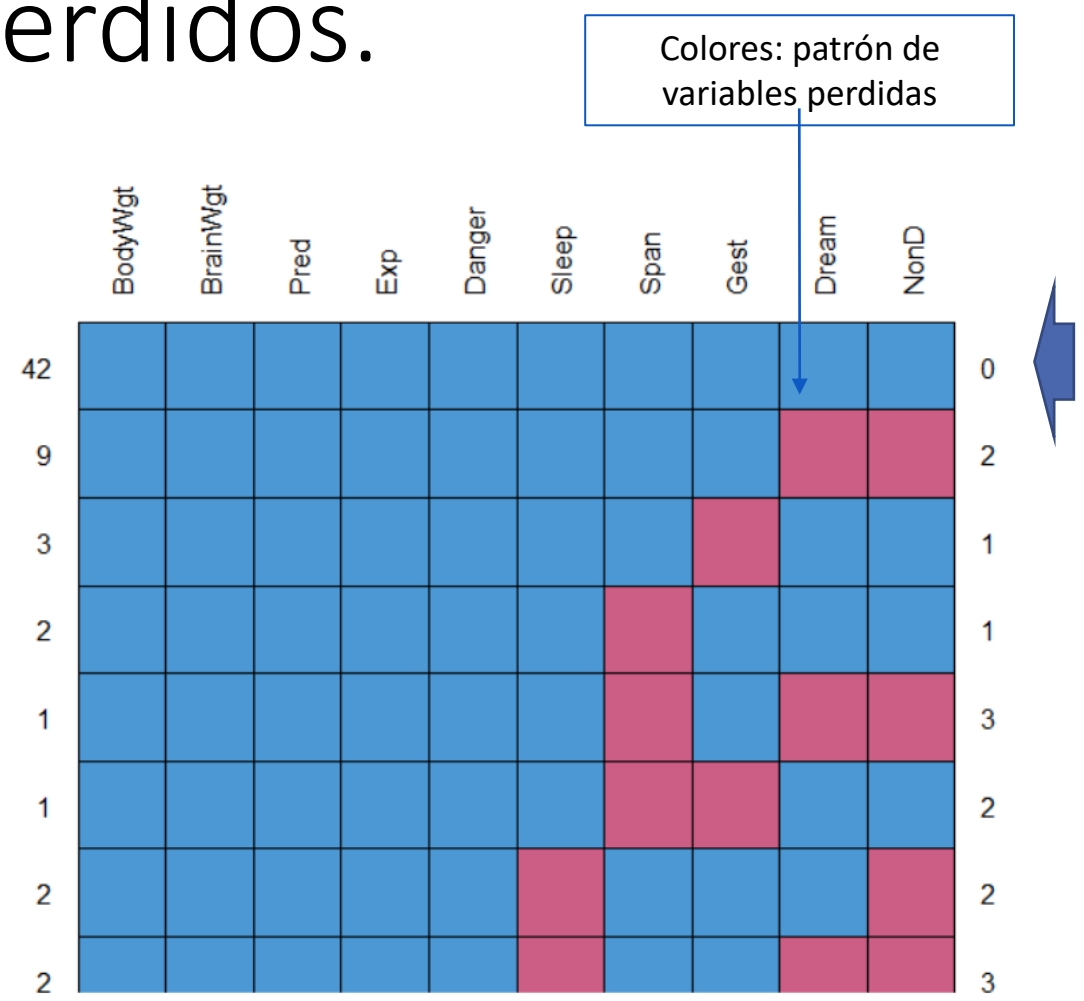
- ¿Cuánta data perdida tengo?
- ¿Cuánta data perdida tengo por grupos de variables?

VIM::aggr()

Exploración de valores perdidos.

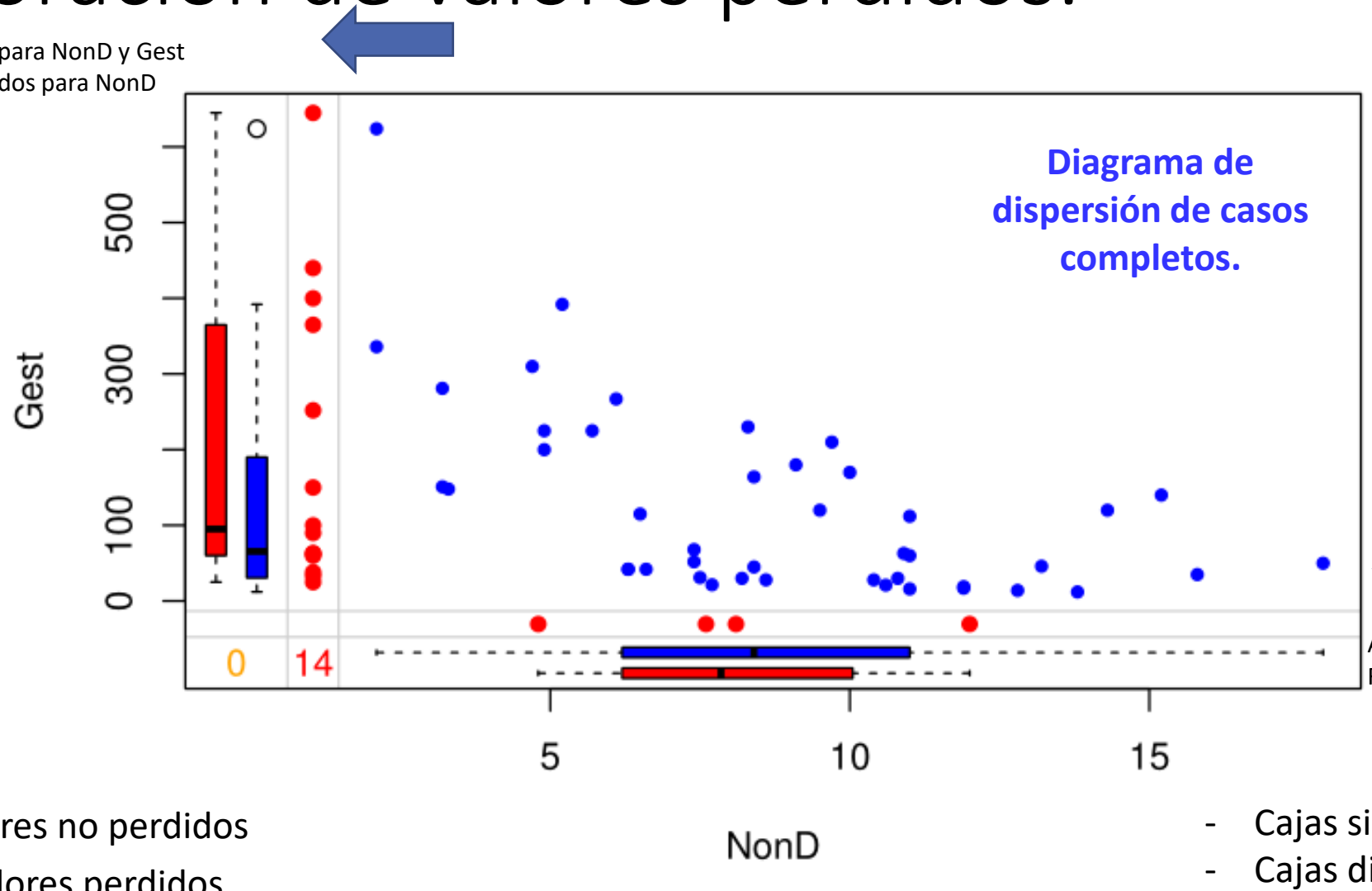
md.pattern()

	BodyWgt	BrainWgt	Pred	Exp	Danger	Sleep	Span	Gest	Dream	NonD	
42	1	1	1	1	1	1	1	1	1	1	0
9	1	1	1	1	1	1	1	1	0	0	2
3	1	1	1	1	1	1	1	0	1	1	1
2	1	1	1	1	1	1	0	1	1	1	1
1	1	1	1	1	1	1	0	1	0	0	3
1	1	1	1	1	1	1	0	0	1	1	2
2	1	1	1	1	1	0	1	1	1	0	2
2	1	1	1	1	1	0	1	1	0	0	3
	0	0	0	0	0	4	4	4	12	14	38



Exploración de valores perdidos.

Azul: Completo para NonD y Gest
Rojo: Solo perdidos para NonD



Azul: Completo para NonD y Gest
Rojo: Solo perdidos para Gest

Boxplot de valores no perdidos

Boxplot de valores perdidos

- Cajas similares -> MCAR
- Cajas diferentes -> MAR

Estrategias para el tratamiento de datos perdidos.

1. Eliminación de casos.
2. Imputación
    a. Imputación simple.
    b. Imputación múltiple.

1. Eliminación de casos.

- Lo más fácil es eliminarlos.
- Sin embargo, esto puede tener algunas consecuencias:
  - La precisión en las estimaciones.
  - Limitados modelos pueden usar pocos datos.
  - Modelos más complejos no podrán ser utilizados.
- Recomendado: Si los incompletos son menos del 5% de los resgistros, podemos considerarlo como opción.

Registros completos	Registros disponibles.
Solo se usan los datos SIN ningún valor perdido.	Según el análisis se utilizan los datos disponibles según análisis.

Más radical

Más flexible.

Puede crear confusión.

2. Imputación

Variables utilizadas

Proceso de imputación

	Univariada	Multivariada
Simple		
Múltiple		

Evelyn Gutierrez (egutierreza@pucp.edu.pe)

2. Imputación

- Univariada:

- Imputar utilizando la misma variable
- Es una forma simple de imputación.

Desventaja:

- No hay variabilidad en los valores imputados.
  - (Se subestimará la varianza)
- Puede contaminar la relación entre variables.
- Si no hay MCAR, los estimaciones pueden ser sesgadas.

¿Cuándo usarlo?

- Cuando se requiere algo rápido y
- Solo hay unos cuantos valores perdidos por imputar.

Por la media

Utilizamos la media para reemplazar.
No añade mayor información

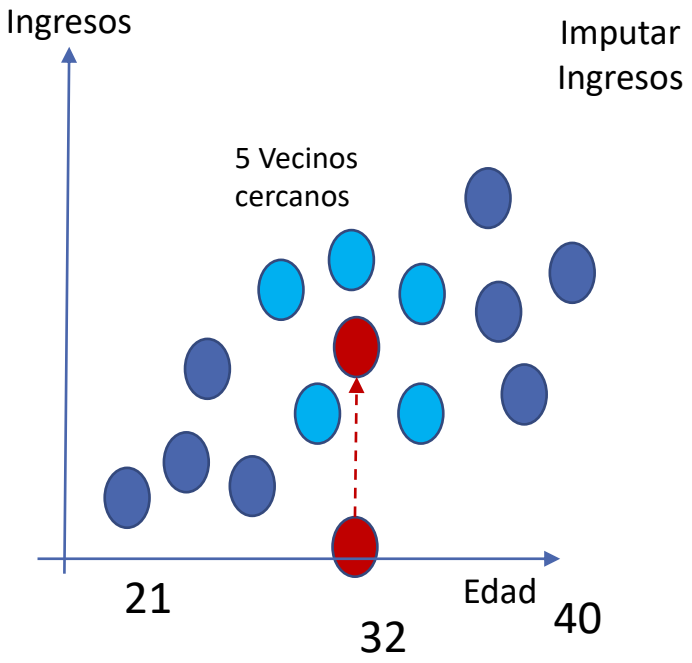
2. Imputación

- Multivariada Utilizando información de otras variables

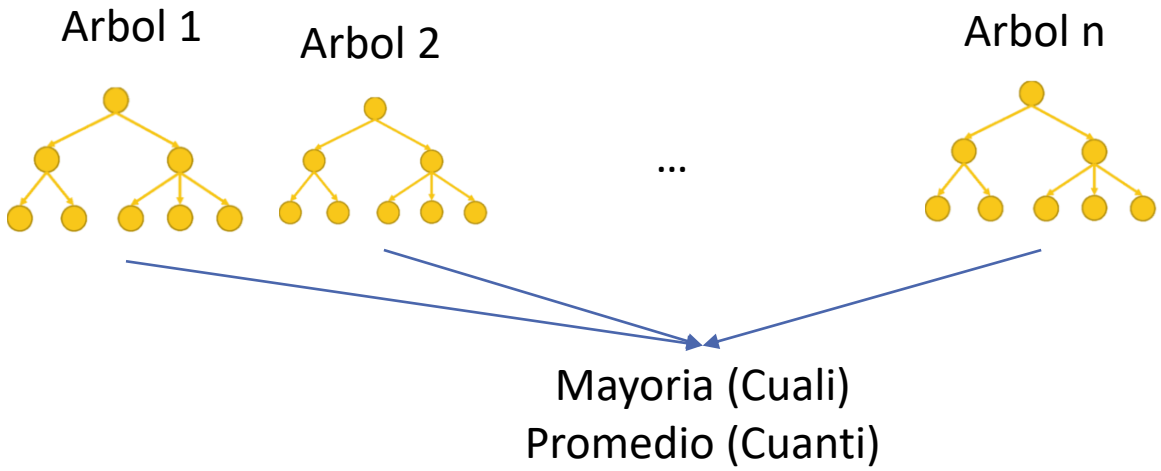
Por regresión	Por vecinos más cercanos	Por bosques aleatorios	MICE: MULTIVARIATE IMPUTATION BY CHAINED EQUATIONS
Se reemplaza los valores perdidos por el valor más probable **basados en las otras variables** disponibles	Usa **KNN** para estimar el valor perdido. (Cuanti-vars)	**Random Forest** para predecir missings. (Cuali/Cuanti vars)	Se utiliza **una cadena de regresiones** para obtener imputaciones,

- La variabilidad general de los datos se subestima. (Aunque en menor medida que con la univariada)
- Las relaciones entre variables pueden sobreestimarse.

KNN: K Nearest Neighbors
K Vecinos más cercanos



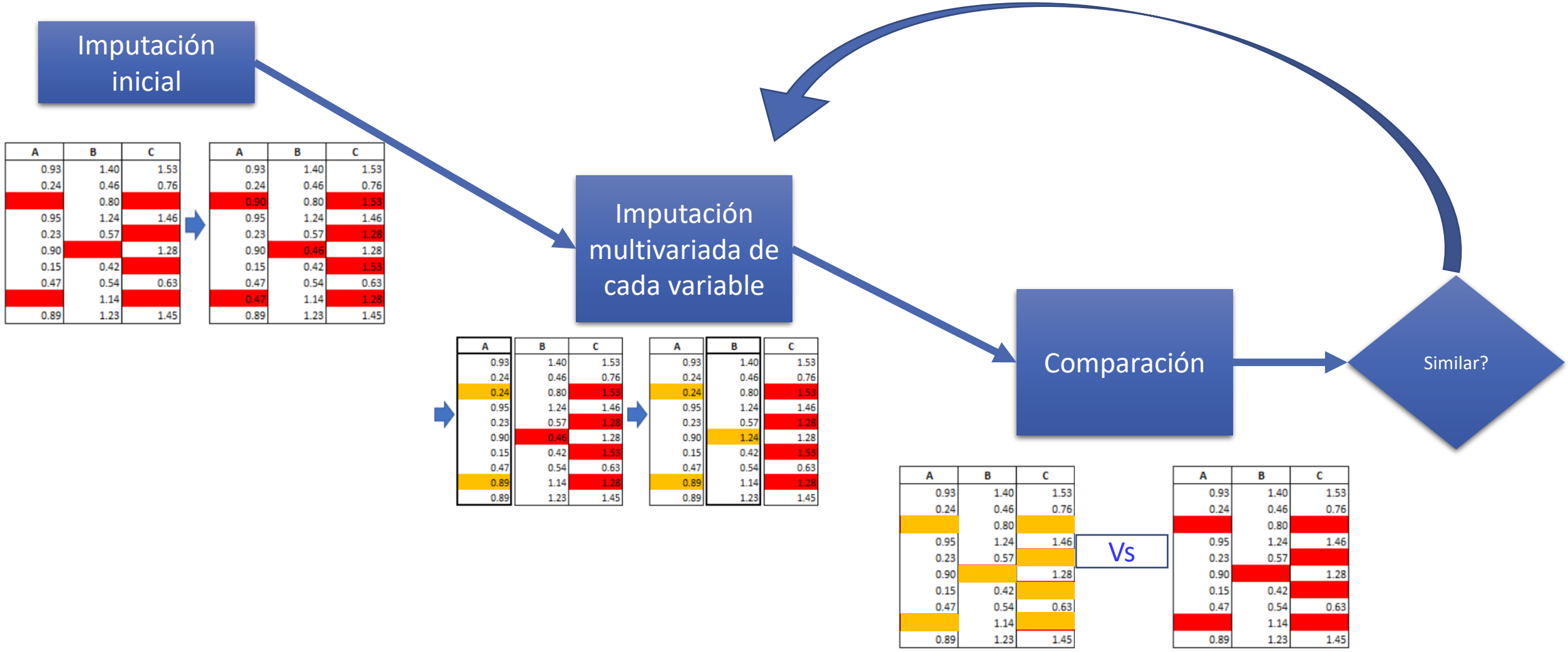
RF: Random Forest
Bosque aleatorio



Evelyn Gutierrez (egutierreza@pucp.edu.pe)

MICE

- Una imputación iterativa



Evelyn Gutierrez (egutierreza@pucp.edu.pe)

Imputación inicial

A	B	C
0.93	1.40	1.53
0.24	0.46	0.76
	0.80	
0.95	1.24	1.46
0.23	0.57	
0.90		1.28
0.15	0.42	
0.47	0.54	0.63
	1.14	
0.89	1.23	1.45

A	B	C
0.93	1.40	1.53
0.24	0.46	0.76
0.90	0.80	1.53
0.95	1.24	1.46
0.23	0.57	1.28
0.90	0.46	1.28
0.15	0.42	1.53
0.47	0.54	0.63
0.47	1.14	1.28
0.89	1.23	1.45

Imputación multivariada de cada variable

A	B	C
0.93	1.40	1.53
0.24	0.46	0.76
0.24	0.80	1.53
0.95	1.24	1.46
0.23	0.57	1.28
0.90	0.46	1.28
0.15	0.42	1.53
0.47	0.54	0.63
0.89	1.14	1.28
0.89	1.23	1.45

A	B	C
0.93	1.40	1.53
0.24	0.46	0.76
0.24	0.80	1.53
0.95	1.24	1.46
0.23	0.57	1.28
0.90	1.24	1.28
0.15	0.42	1.53
0.47	0.54	0.63
0.89	1.14	1.28
0.89	1.23	1.45

A	B	C
0.93	1.40	1.53
0.24	0.46	0.76
0.90	0.80	1.53
0.95	1.24	1.46
0.23	0.57	1.28
0.90	0.46	1.28
0.15	0.42	1.53
0.47	0.54	0.63
0.47	1.14	1.28
0.89	1.23	1.45

Vs

A	B	C
0.93	1.40	1.53
0.24	0.46	0.76
0.24	0.80	1.20
0.95	1.24	1.46
0.23	0.57	0.89
0.9	1.24	1.28
0.15	0.42	1.05
0.47	0.54	0.63
0.89	1.14	0.70
0.89	1.23	1.45

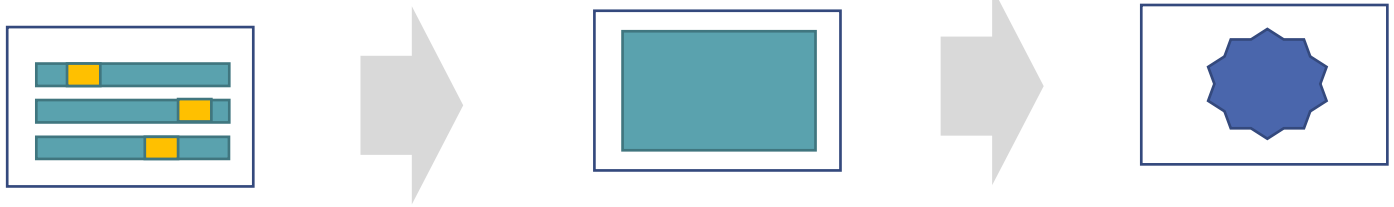
Evelyn Gutierrez (egutierreza@pucp.edu.pe)

Imputación Simple y Múltiple.

Simple: Imputar un solo dataset

Multiple: Imputar varios dataset

Explorar datos perdidos → Imputar → Realizar análisis



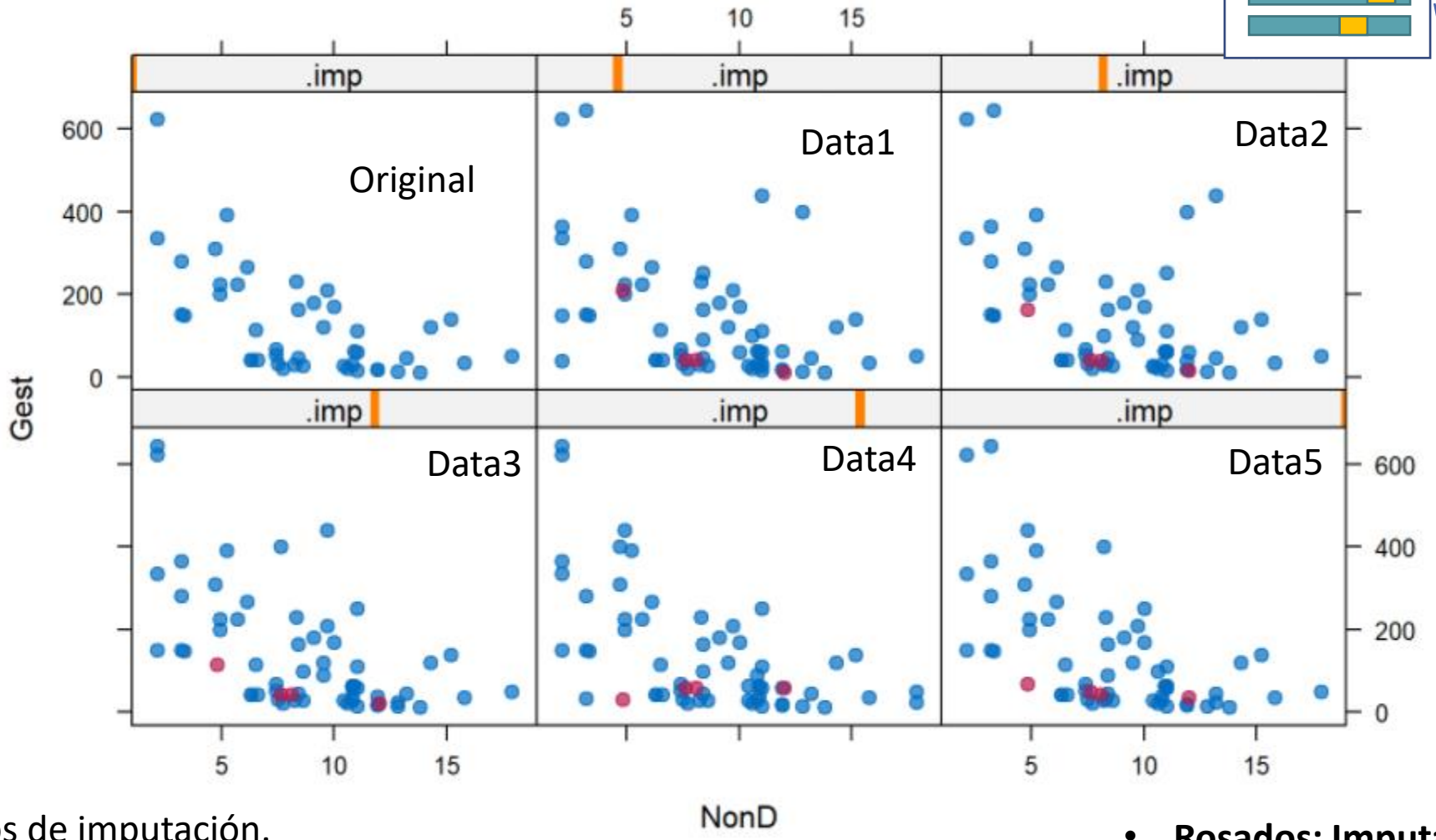
Evelyn Gutierrez (egutierreza@pucp.edu.pe)

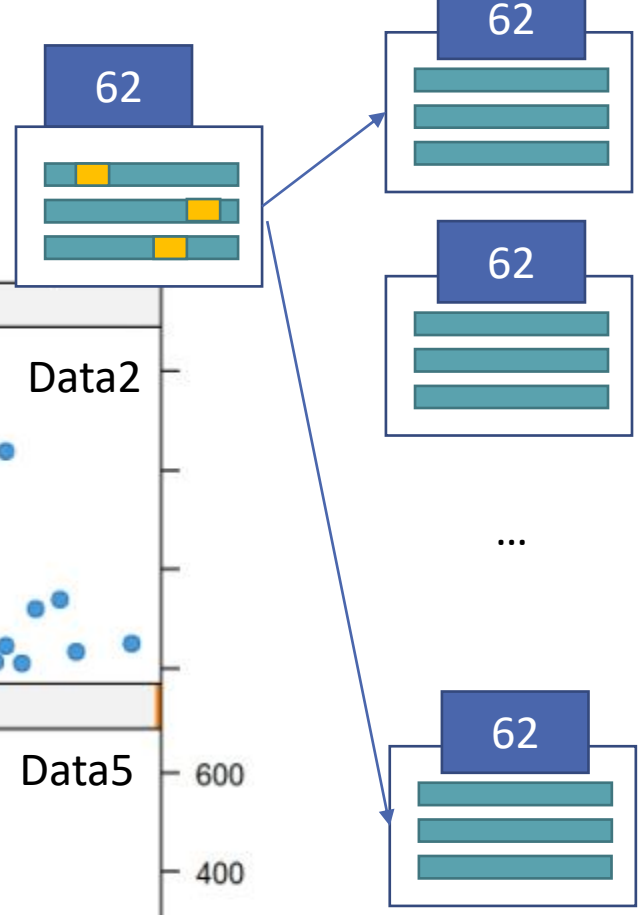
Paquetes en R utilizados:

Hmisc -> install.packages(“Hmisc”)
VIM -> install.packages(“VIM”)
Mice -> install.packages(“mice”)
DMwR -> remotes::install_github("cran/DMwR")

Install.packages(“performance”)

Análisis de imputaciones para **Gest**





- 5 escenarios de imputación.

- **Rosados: Imputados para Gest**
- **Azul: "Completo" para Gest-NonD**

Preguntas

Data Advanced Analytics

Programas de Especialización en

Business Inteligence and Data Analytics

Evelyn Gutierrez (egutierreza@pucp.edu.pe)

Referencias:

Discusión:

- MICE vs Forest
- MICE vs KNN
- Imputación estocástica.

Libros:

- Flexible imputation of missing data
- Applied Missing Data Analysis