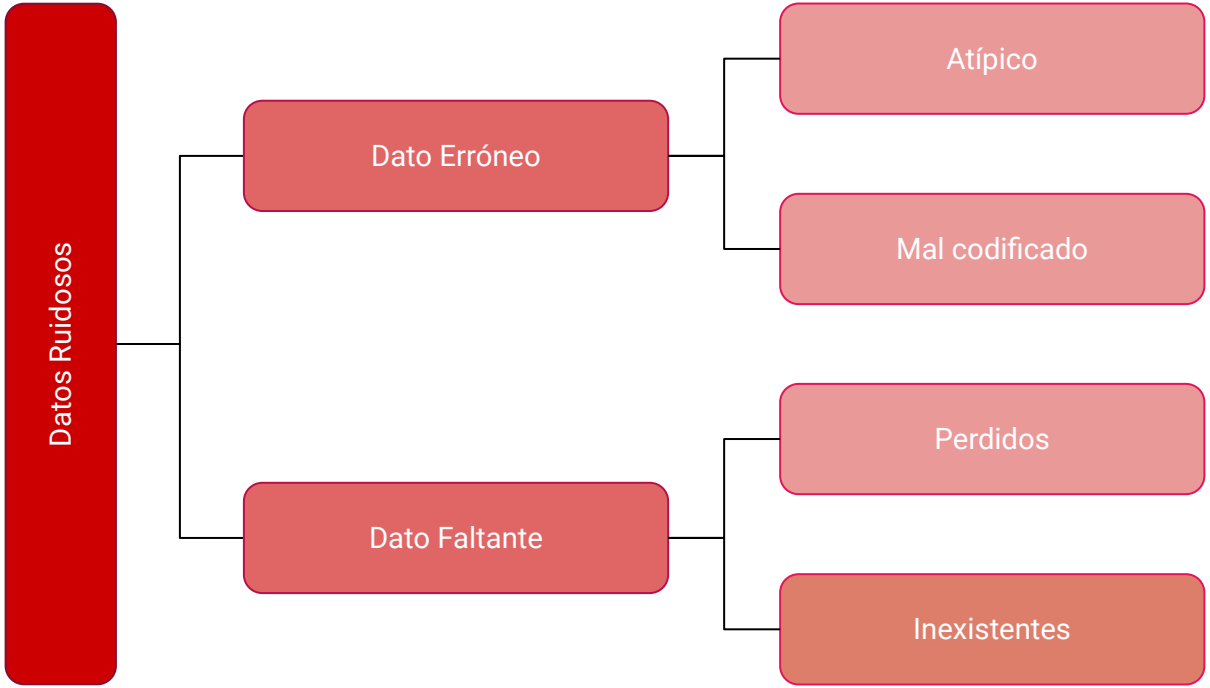
Datos Ruidosos

Indice de temas

Tipos de datos ruidosos



Como trabajamos con datos erróneos

Inspeccionamos los datos

+

Separamos datos atípicos de datos erróneamente codificados

→

Decidimos

- Retirar los datos atípicos
- Retirar los erróneamente codificados
- registrar los problemas y no tomamos acción

Como trabajamos con datos faltantes

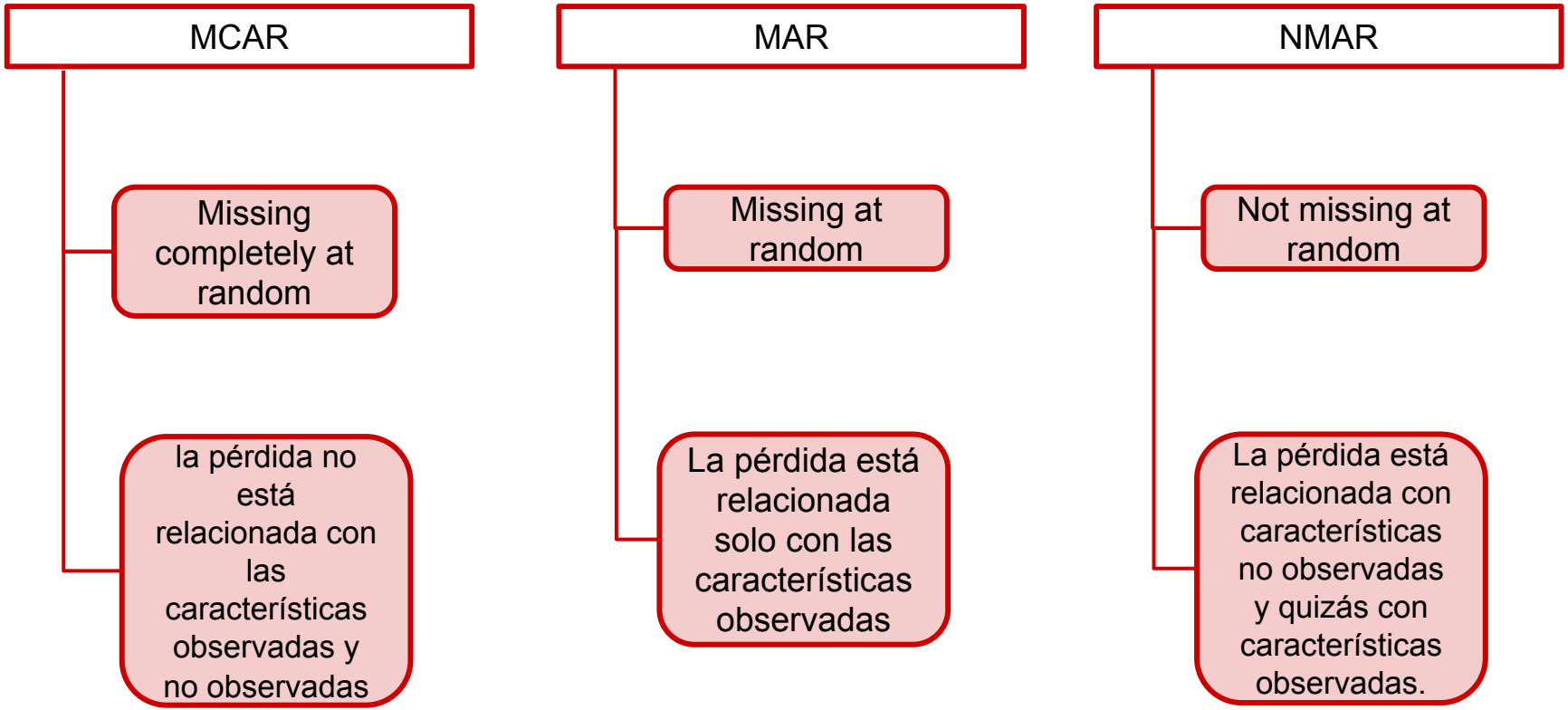
En estadística,

- **predecir** es otorgar valor a un dato que todavía no ha sido muestreado,

- **imputar** es estimar un valor que puede haber sido muestreado pero no se lo conoce.

Si uno logra realizar un modelo de predicción con los datos que no tienen problemas…
imputar es **predecir** esos datos.

Modelo de pérdida de datos

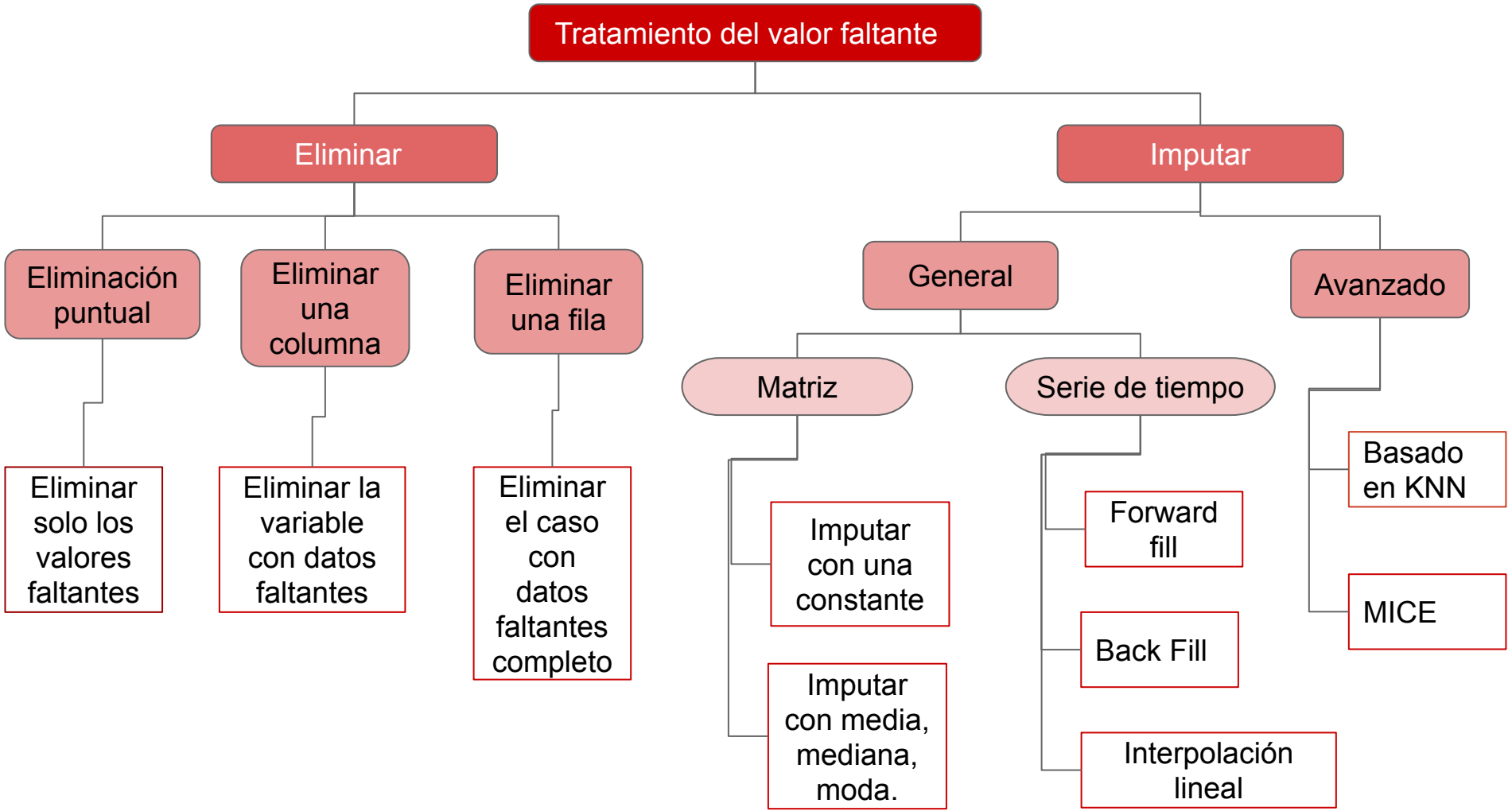


Missingno: librería para explorar datos faltantes

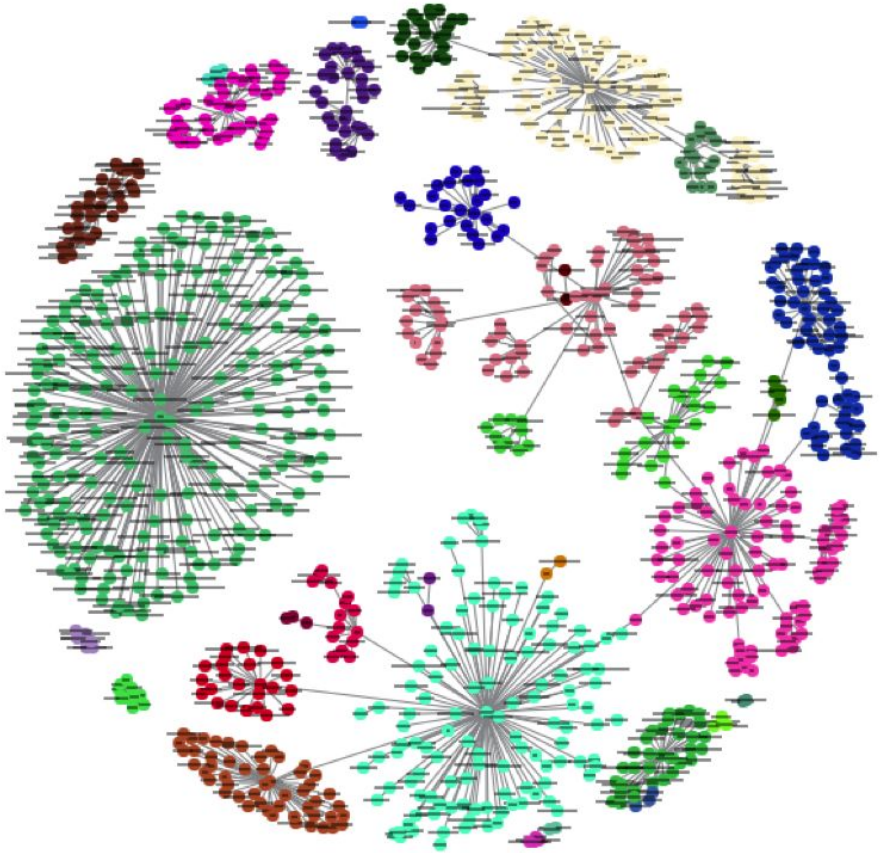
- `pip install missingno`

- **Bar Chart :**
  - Este gráfico de barras le dá una idea de cuántos valores faltantes hay en cada columna.
- **Matrix :**
  - Con este gráfico de barras especial se puede encontrar muy rápidamente el patrón de pérdidas en el conjunto de datos.
- **Heatmap :**
  - Este mapa visualiza la correlación de la pérdida entre dos columnas con un heatmap.

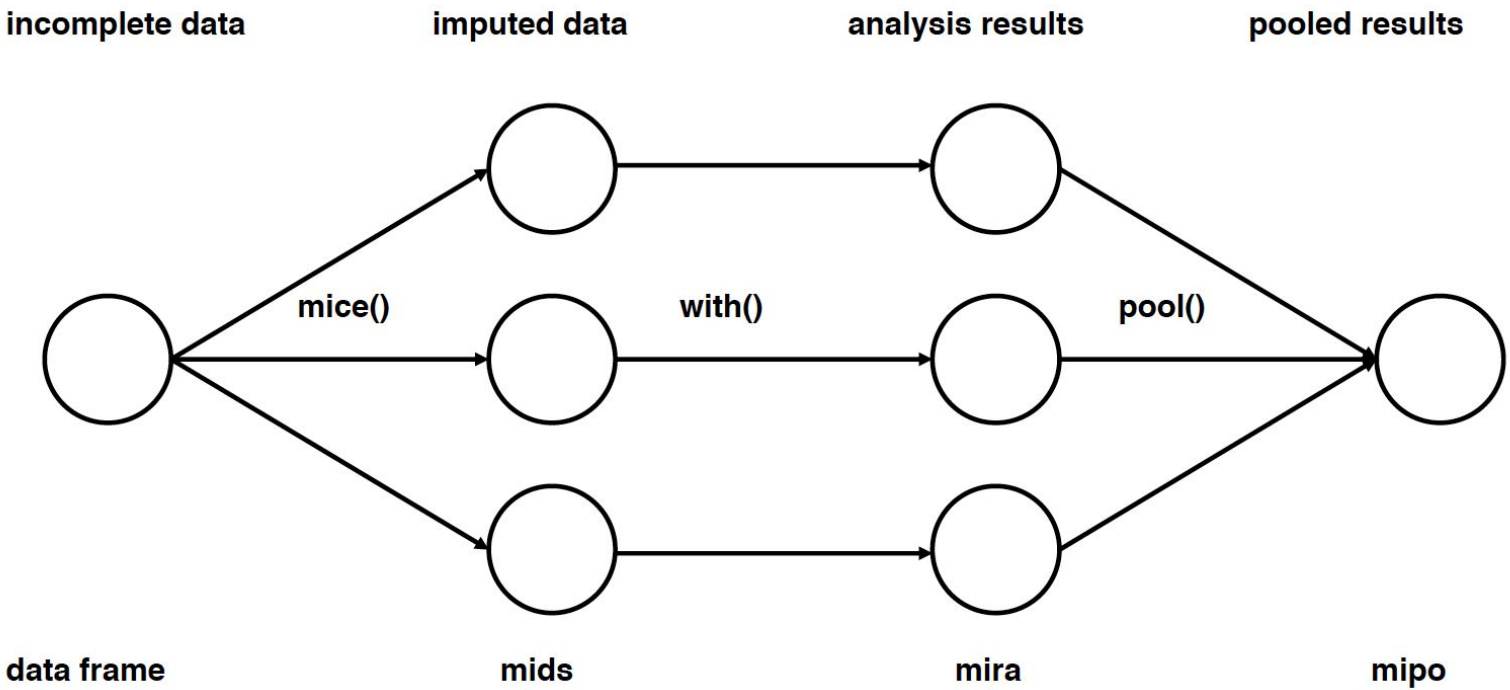
Que vemos en estos gráficos?


Tratamiento del valor faltante
Eliminar
Imputar
Eliminación puntual
Eliminar una columna
Eliminar una fila
General
Avanzado
Matriz
Serie de tiempo
Eliminar solo los valores faltantes
Eliminar la variable con datos faltantes
Eliminar el caso con datos faltantes completo
Imputar con una constante
Imputar con media, mediana, moda.
Forward fill
Back Fill
Interpolación lineal
Basado en KNN
MICE

KNN imputation



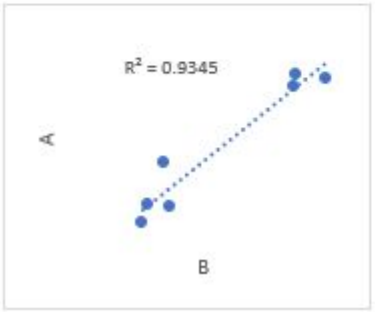
MICE: Multivariate Feature Imputation



MICE Forest

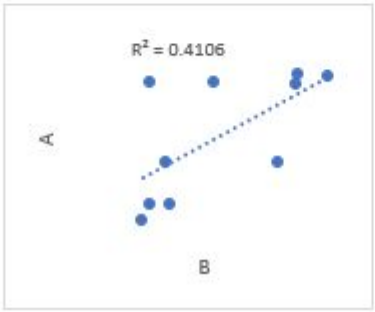
Missing data is in red. There is a strong correlation between A and B, so let's try to impute A using B and C.

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



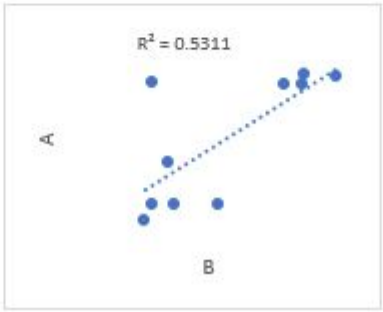
Missing data is filled in randomly. This dillutes the correlations, but allows us to impute using all available data.

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



A random forest is used to predict A with B and C. Notice the correlation between A and B improved.

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



After Imputing B using A and C, we have achieved a correlation between A and B much closer to the original data.

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