

Your NaNs are telling you something.

D8A.ACADEMY/RESOURCES/MISSING-DATA-STRATEGIES

FIRST, LOOK

<code>df.isna().sum()</code>	count per column
<code>df.isna().mean()</code>	share per column
sort by that	worst columns first

THE 3 KINDS

MCAR	truly random, rare
MAR	explained by others
MNAR	← the value itself hides.

DROP

<code>dropna()</code>	rows, if very few
<code>dropna(axis=1)</code>	cols mostly empty
<code>thresh=8</code>	keep partial rows

FILL

<code>fillna(median)</code>	skew-proof default
<code>fillna(mode)</code>	categoricals
<code>ffill()</code>	time series only

SMARTER

<code>KNNImputer</code>	neighbors vote
<code>IterativeImputer</code>	model per column
<code>was_missing = 1</code>	flag it, keep it

GOTCHAS

fit on train only	← or you leak the test.
mean + outliers	fills with nonsense
silent zeros	is 0 really 0?

WORTH MEMORIZING

01 Fill and flag, together

The value was missing: that fact often predicts better than the filled value.

```
# the two-line upgrade over plain fillna
df['age_missing'] = df['age'].isna().astype(int)
df['age'] = df['age'].fillna(df['age'].median())

# leakage-proof version: learn the median on train only
med = train['age'].median()
test['age'] = test['age'].fillna(med)
```

→ Median, not mean: one outlier in the column and the mean fills every hole with nonsense.

02 Trap: not-really-random

People do not skip survey questions at random. Neither do sensors, forms or joins.

LOOKS LIKE	IS OFTEN	SO
random NaNs	MAR: explained by another col	impute using that col
missing income	MNAR: high earners skip it	model it, add a flag
whole block empty	a broken join or export	fix upstream, not fillna

→ Interview phrasing worth memorizing: dropping MNAR rows biases the sample toward people with nothing to hide.