The First Date with Your Data

7 Questions to Ask Before You Model

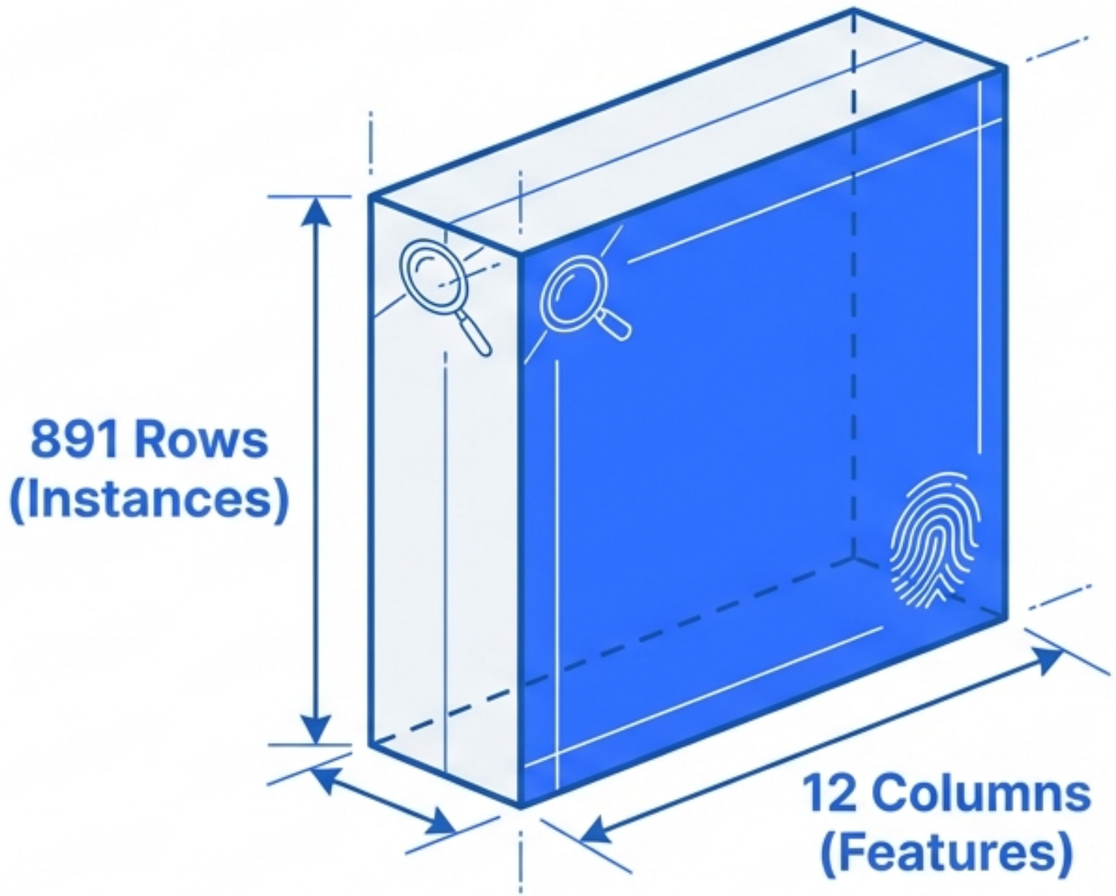
Context: 100 Days of Machine Learning | Day 19

Objective: Establish a baseline truth before
Objective: Establish a baseline truth before
interrogation.

Warning: Jumping straight to modelling is a recipe for failure. Interview the suspect first.

Q1: How big is the beast?

```
df.shape
```



Detective's Note

Output: (891, 12)

Titanic Context:
891 passengers, 12 facts per passenger.

Strategic Insight:

- **Small Data**: Fast iteration, easy processing.
easy processing.
- **Big Data**: Requires memory optimisation.

Q2: What does it look like?

Avoiding the Bias Trap

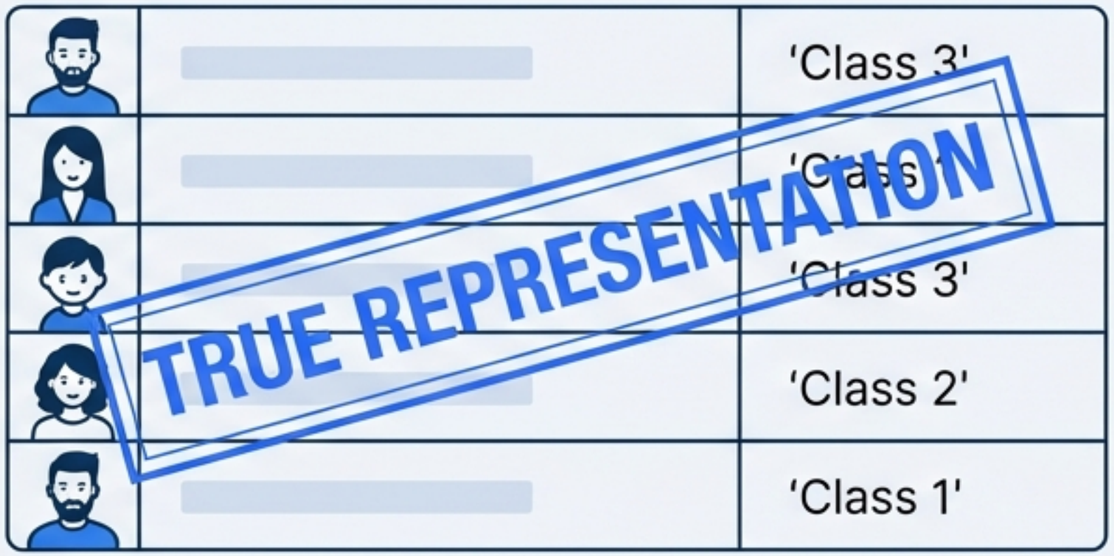
The Novice Approach

```
df.head()
```

Danger: Data is often sorted by time or ID. Viewing the top 5 rows creates a false impression of homogeneity.

The Expert Approach

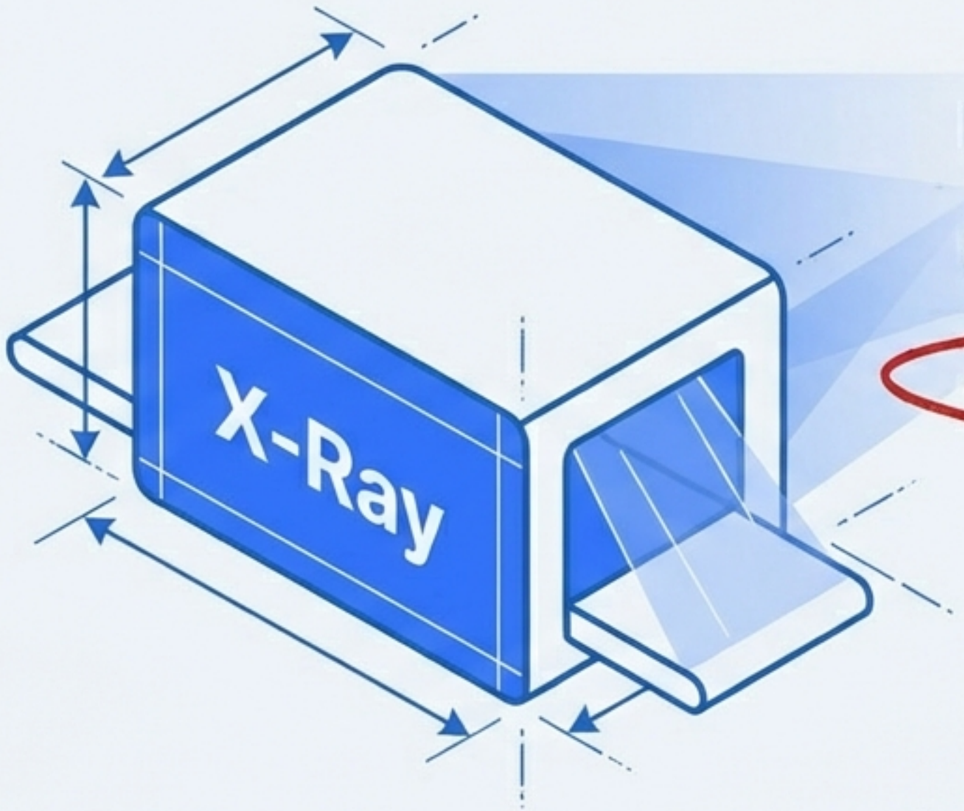
```
df.sample(5)
```



Solution: Pick random rows to see the variety. Don't assume the whole dataset is male just because the first 10 rows are.

Q3: What are the Data Types?

```
df.info()
```



Name	→	object (String)
Survived	→	int64 (Integer)
Age	→	float64 (Decimal)

⚠ **Optimisation Opportunity:** Why store simple numbers as decimals? Converting floats to integers reduces memory footprint and speeds up training.

Goal: Identify type mismatches and memory waste before they crash your pipeline.

Q4: Are there holes in the story?

Missing Value Investigation

```
df.isnull().sum()
```

Forensic Strategy

- High Missing (Cabin): Likely drop the column.
- Medium Missing (Age): 🧮 Imputation (Fill with Mean/Median).
- Low Missing (Embarked): ❎ Drop the rows.

Q5: How does it look mathematically?

```
df.describe()
```

Anomaly Detected: Data contains infants (0.42 years). Not just integers.

	Count	Mean	Std	Min	25%	50%	75%	Max
Age	714	29.7	14.5	**0.42**	20.1	28.0	38.0	80
Fare	891	**32.2**	49.7	0.0	7.9	14.5	31.0	512.3

Baseline Established: Average ticket price is £32. Helps identify pricing outliers.

Q6: Is there noise?

Duplicate Check

```
df.duplicated().sum()
```

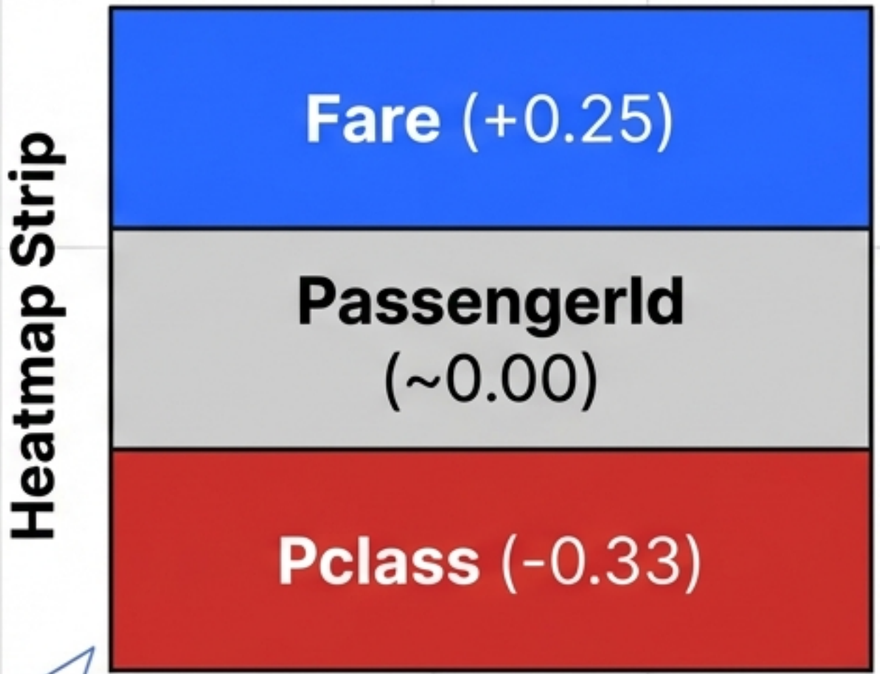
Concept: Redundancy creates bias.

Why it matters: If a specific data point appears twice, the model artificially gives it double importance. It “over-learns” that specific example.

The Fix: If result > 0, execute “df.drop_duplicates()”.

Q7: How are thinings connected?

```
df.corr()['Survived']
```



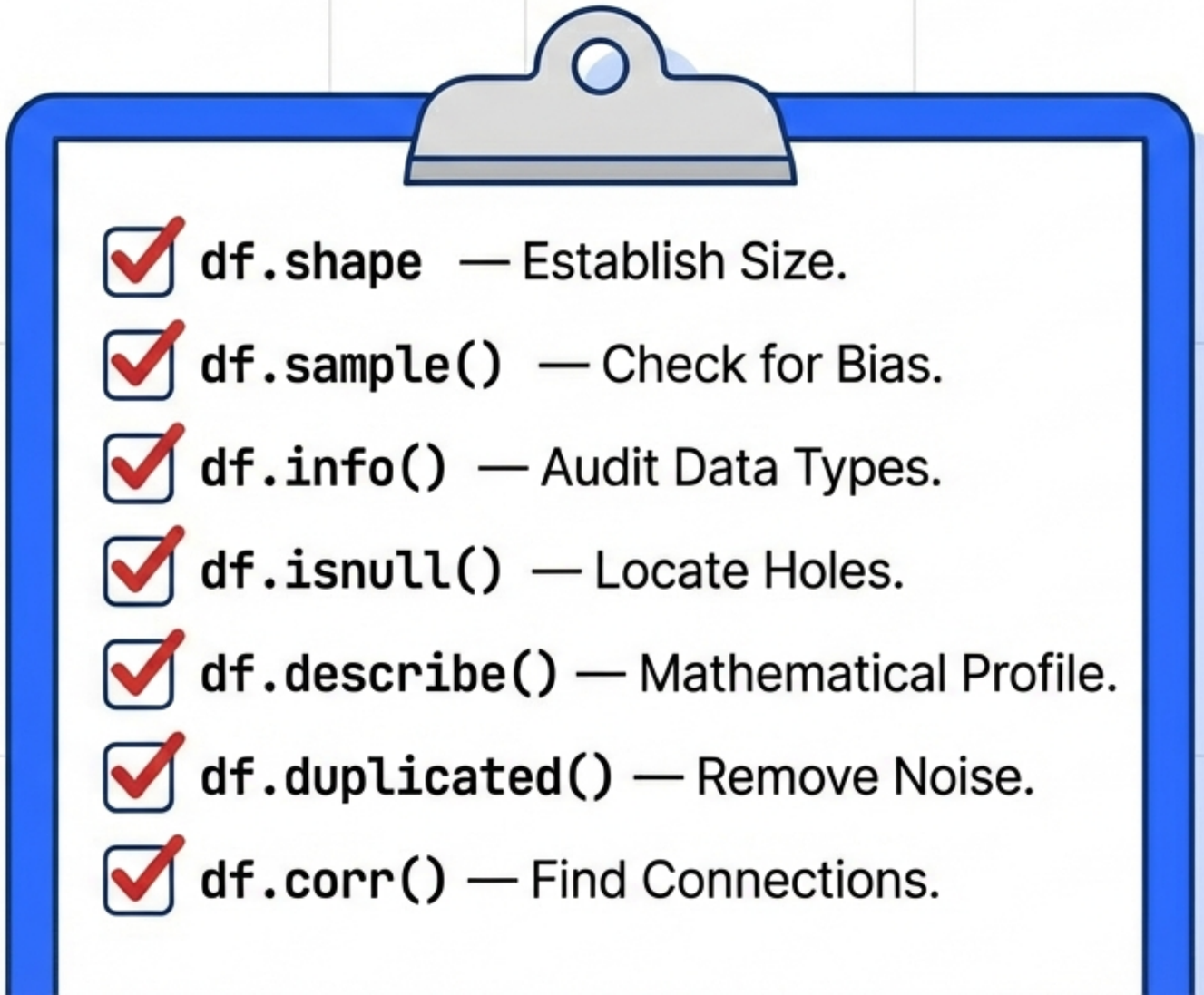
Positive Correlation: Wealthier passengers saved first.

Zero Correlation: Random noise. No predictive power.

Negative Correlation: As class number goes up (3rd Class), survival chance goes down.

This step separates the evidence (Signal) from the distraction (Noise).

The Interrogation Roadmap



S.O.P.

These 7 questions are your Standard Operating Procedure for any new dataset.
Memorise them.

Looking Ahead

From Interrogation to Forensics

Basic Questions (Understanding)

NEXT STEP

Exploratory Data Analysis (EDA)

Univariate & Multivariate Analysis

Pro Tip: Expert Insight: Automated Interrogation 

Once you master the manual process, tools like `'Pandas Profiling'` can perform all 7 **checks** with a single line of code. But you must understand the logic first.