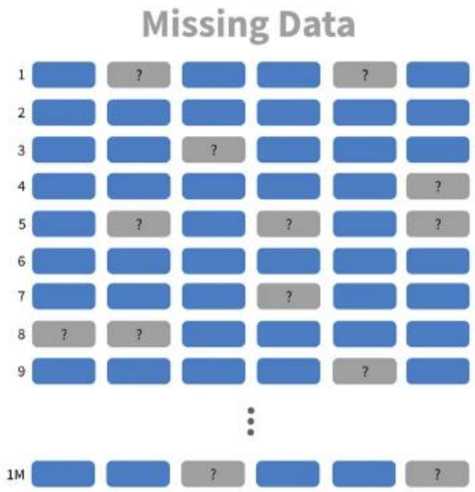
What is missing data?



Imagine we're keeping track of customer purchases at our men's sports apparel store.

What is Missing Data?

- Sometimes, we might have incomplete records.
- For example, we might know a customer bought a t-shirt, but we forgot to write down the exact price. Or maybe our computer system glitched and didn't record the delivery time.
- These missing pieces of information are what we call "missing data."

Why Do We Care About Missing Data?

- If we try to analyze our sales data with missing information, our results might be wrong.
- For example, if we're trying to find the average purchase amount and some purchase amounts are missing, our average won't be accurate.
- Missing data can also cause problems with our prediction models.

Why does missing data occur?

Missing data is a common problem in real-world datasets and can arise for various reasons, including:

- **Data Entry Errors:** Mistakes during manual data input can lead to blanks or incorrect entries that are later treated as missing.

- **Non-Response:** In surveys or questionnaires, respondents may choose not to answer certain questions.
- **System Errors:** Software glitches, database issues, or sensor malfunctions can prevent data from being recorded or lead to data loss.
- **Data Collection Issues:** Problems with the data collection process itself, such as faulty equipment or incomplete procedures.
- **Data Merging:** When combining datasets from different sources, some variables might not be available in all sources.
- **Privacy Concerns:** Individuals might opt out of providing certain information due to privacy reasons.
- **Data Corruption:** Errors during data transmission or storage can lead to data becoming unreadable or being interpreted as missing.

Why is missing data a problem?

Missing data can significantly impact data analysis and modeling:

- **Reduced Statistical Power:** Fewer complete observations can lead to less precise estimates and lower statistical power in hypothesis testing.
- **Biased Estimates:** If the missing data is not MCAR, the remaining complete observations might not be representative of the overall population, leading to biased results.
- **Complicated Analysis:** Many statistical and machine learning algorithms cannot directly handle missing values and may produce errors or require specific handling techniques.
- **Loss of Information:** Ignoring or simply removing missing data can lead to a loss of valuable information contained in the other variables of those observations.
- **Inefficient Modeling:** Models trained on data with missing values might be less accurate and less reliable.

Therefore, identifying, understanding the patterns of, and appropriately handling missing data is a critical step in the data cleaning and preparation process in data science.