Short Notes — Data Bias, Missing Data Strategies, KPIs vs Metrics

✅ Data Bias

Data bias occurs when the collected data does not fairly represent the real population or business reality, leading to misleading analysis and wrong decisions. Bias usually enters during data collection, sampling, or measurement.
Common types include:

✅ Simple Example — Survey Bias

Suppose you want to analyze **average income of a city**.
You collect survey data only from:

- IT parks
- Corporate offices
- Tech employees

Your dataset result: Average income = ₹80,000/month**Reality:** City average = ₹35,000/month
👉 Your data is biased because low-income and non-tech workers were not included.

✅ Example — ML Model Bias

A company builds a hiring model using past employee data.
Past data mostly contains:

- 90% male employees
- 10% female employees

Model learns pattern → "male candidates are preferred"Now it unfairly rejects qualified female candidates.
👉 Bias came from **historical imbalance in data**.

✅ Example — Product Feedback Bias

A restaurant collects feedback only through a mobile app.Only young customers respond.

Result:

- Rating looks very high
- Older customers' opinions missing

👉 Decision based on incomplete audience = biased insight.

- Sampling bias — data collected from only one group
- Selection bias — excluding important segments
- Measurement bias — wrong or inconsistent measurement method
- Reporting bias — only positive results recorded

🔹 Types of Data Bias

🔹 Sample Bias — Example

Sample bias happens when the data sample is not representative of the whole population, so results become misleading.

✅ Example — College Feedback Survey

A college wants to measure **student satisfaction**.
They collect feedback only from:

- Final year toppers
- Students in honors classes

Results show: **95% satisfaction**
But they did not include:

- Students with backlogs
- Students who rarely attend

👉 The sample is biased because it includes only high-performing students, not the full student population.

2️⃣ Selection Bias

When data is selected in a non-random way that favors certain outcomes.

Example:

A study on "average daily exercise time" uses data from:

- Gym members only

It ignores:

- Non-gym people

Result: Study says average exercise = 1.5 hours/day — unrealistic for the full population.

👉 Dataset was selected from a biased group.

3️⃣ Measurement Bias

When the method of measuring or collecting data is flawed.

A weather station sensor is miscalibrated and always records temperature **+2°C higher** than actual.

- True temperature = 30°C
- Recorded temperature = 32°C

All collected data is systematically wrong.

👉 This is **measurement bias** because the measuring instrument is flawed.

🔹 Reporting Bias — Explained with Example

Reporting bias happens when some results or information are **selectively reported while others are hidden or ignored**, leading to a misleading conclusion.

✅ Example — Medical Study

A drug company runs 10 clinical trials:

- 7 trials show weak or no effect
- 3 trials show strong positive effect

They publish only the 3 successful trials.

👉 Public sees the drug as highly effective — but full data tells a different story.This is **reporting bias**.

✅ Missing Data Strategies

Missing data is common in real datasets due to system errors, optional fields, or integration gaps. The handling strategy depends on column type and business meaning.

Main strategies:

1. Deletion — Remove rows or columnsUsed when missing values are very few and not important. Risk: data loss.
2. Imputation — Fill missing values
   - Numeric → mean or median
   - Categorical → mode or "Unknown"
   - Time series → forward/backward fill
3. Business-rule fillingFill based on domain logic.Example: missing sales value → treat as 0 if it means no recorded sale.
4. Leave as nullUsed when value cannot be safely guessed (like missing dates of events). Analysis then excludes those rows only when needed.

Real example:In a sales dataset, missing discount field may be filled with 0, but missing order_date should not be fabricated.

Best practice:Choose method based on data type and business context — not one rule for all.

☑️ KPIs vs Metrics

Both KPIs and metrics measure performance, but they are not the same.
Metrics are general measurements that track activity or performance. KPIs (Key Performance Indicators) are critical metrics directly tied to business goals and decision-making.
Metrics answer: What are we measuring?KPIs answer: What matters most for success?

Real examples in an online store:
Metrics:

- Website visits
- Page views
- Cart additions
- Email opens

KPIs:

- Conversion rate
- Revenue per customer
- Customer retention rate
- Monthly profit

☑️ Metrics (General Performance Measures)

Metrics are numbers used to **measure activities or performance**, but they may not directly decide business success.
Examples of Metrics:

- Website page views
- Number of app downloads
- Daily active users
- Email open rate

👉 These tell *what is happening*, but not whether business goals are achieved.
Example:A website gets **50,000 page views**, but sales are still low.Page views are a **metric**, not a KPI.

☑️ KPIs (Key Performance Indicators)

KPIs are **critical metrics tied directly to business goals** and decision-making.
Examples of KPIs:

- Conversion rate
- Monthly revenue
- Customer retention rate
- Customer acquisition cost (CAC)

👉 These tell *whether the business is succeeding*.
Example: If the goal is to increase sales, **conversion rate** is a KPI.If conversion rate rises from 2% → 4%, business performance improved.

Aspect	Metrics	KPIs
Purpose	Track activity	Track success
Goal-linked	❌ Not always	☑️ Yes
Decision impact	Low	High