🎬 Video 22 – Understanding and Handling Duplicate Data

🔁 What Is Duplicate Data?

- **Definition**: A record that repeats the information of another record, either partially or exactly.
- **Types**:
  - **Partial Duplicate**: Some fields match, others differ (e.g., missing phone number).
  - **Exact Duplicate**: All fields match completely.

⚠️ Why Duplicates Are a Problem

- **Data Integrity**: Duplicates distort analysis and reporting.
- **Skewed Metrics**: Aggregations (e.g., averages) can be inflated by repeated values.
- **Wasted Resources**: Duplicate entries can lead to:
  - Redundant marketing outreach
  - Increased storage costs
  - Inefficient decision-making

📊 Real-World Examples

- **Sales Data**: A shirt sold for $20 entered three times skews average price.
- **Marketing Campaigns**: Duplicate addresses cause multiple ads to be sent to the same person, wasting budget.

🧹 Deduplication Techniques

- **Manual Deduplication**:
  - Good for small datasets
  - Involves row-by-row comparison
- **Automated Deduplication Tools**:
  - Use algorithms to detect duplicates

- Efficient for large datasets
- Can identify both exact and partial duplicates

🧠 Key Takeaway

Deduplication is essential for maintaining data quality and is a critical part of any data transformation plan.