Preparing Data for Analysis: Cleaning and Wrangling

Part 1: Office Roleplay Dialogue

Scenario: A Data Analyst, Ayaka, is working with her colleague, Miguel, to collect and clean data from various sources to prepare it for analysis.

Miguel: Hey Ayaka, I've pulled the raw data from different sources, but it looks messy. How should we start the **data cleaning** process?

Ayaka: First, we need to identify any **missing values** and decide whether to fill them in or remove those records.

Miguel: Good point. I also noticed inconsistencies in date formats and text entries. Would that fall under **data wrangling**?

Ayaka: Yes, **data wrangling** helps restructure and standardize the dataset, so everything is consistent before analysis.

Miguel: Got it. Should we use an **ETL (Extract, Transform, Load)** process to automate some of this?

Ayaka: Absolutely. We can extract the raw data, apply transformations like formatting corrections, and then load it into our database.

Miguel: Sounds efficient! I also noticed some numeric fields need **normalization**. Should we scale them to a common range?

Ayaka: Yes, **normalization** ensures values are on a similar scale, which is useful for machine learning models.

Miguel: That makes sense. Once everything is clean, we can start analyzing trends and generating reports.

Ayaka: Exactly! Let's complete the **ETL** process, verify the cleaned data, and then move on to visualization.

Miguel: Perfect. I'll document the changes in case we need to backtrack.

Ayaka: Good idea. Keeping track of our **data cleaning** steps will help with reproducibility.

Part 2: Comprehension Questions

1. Why is Ayaka checking for missing values?
(A) To delete all records from the dataset
(B) To decide whether to fill them in or remove the records
(C) To improve website design
(D) To make the dataset larger

2. What does Miguel mean by data wrangling?
(A) Deleting all numerical values
(B) Converting text into numbers
(C) Changing the database password
(D) Standardizing and restructuring messy data

3. Why is normalization important in data preparation?
(A) It removes duplicate text entries
(B) It reduces database storage usage
(C) It ensures numerical values are on a similar scale
(D) It merges all data columns into one

4. What is the purpose of ETL (Extract, Transform, Load)?
(A) To automate the process of extracting, cleaning, and storing data
(B) To remove unwanted columns
(C) To create charts and graphs
(D) To rename dataset files

Part 3: Key Vocabulary Definitions in Japanese

1. **Data Cleaning (データクレンジング)** – データの誤りや欠損値を特定し、修正または削除して品質を向上させる作業。
2. **Data Wrangling (データ整形)** – データの形式や内容を整え、分析しやすい状態にするプロセス。
3. **ETL (Extract, Transform, Load) (ETL・抽出、変換、ロード)** – データを取得し（抽出）、必要な加工を行い（変換）、最終的に保存する（ロード）プロセス。
4. **Missing Values (欠損値)** – データセット内で特定の項目が欠けている状態。
5. **Normalization (正規化)** – データの値を一定の範囲にスケール調整し、一貫性を持たせる手法。

Part 4: Questions & Correct Answers

1. **Why is Ayaka checking for missing values?**

   ✅ (B) To decide whether to fill them in or remove the records

2. **What does Miguel mean by data wrangling?**

   ✅ (D) Standardizing and restructuring messy data

3. **Why is normalization important in data preparation?**

   ✅ (C) It ensures numerical values are on a similar scale

4. **What is the purpose of ETL (Extract, Transform, Load)?**

   ✅ (A) To automate the process of extracting, cleaning, and storing data