CS 5: What Students Know :

Part 1: Select a threshold concept and write the 'what students know', 'what they dont know', 'principles':

List me all the threshold concepts for the course Data mining. Make a list only.

Here's a list of key threshold concepts for a **Data Mining** course:

- Data preprocessing and cleaning
- Data transformation and normalization
- Feature selection and dimensionality reduction
- Similarity and distance measures
- Classification concepts and decision boundaries
- Regression analysis fundamentals
- Clustering principles
- Association rule mining
- Evaluation metrics (precision, recall, F1-score, ROC)
- Overfitting and underfitting
- Bias–variance tradeoff
- Cross-validation techniques
- Ensemble learning fundamentals
- Scalability and big data considerations
- Interpreting and visualizing mined patterns

List me all the principles for Data preprocessing and cleaning Data transformation and normalization. Just make a list.

Principles of Data Preprocessing and Cleaning

- Identify and handle missing data
- Detect and remove duplicates
- Handle noisy data (smoothing, binning)
- Resolve data inconsistencies
- Outlier detection and treatment
- Data type validation
- Standardize formats (dates, text, units)
- Handle imbalanced datasets
- Ensure data integrity and accuracy

Principles of Data Transformation and Normalization

- Attribute scaling (min–max scaling, z-score standardization)
- Logarithmic or power transformations
- Encoding categorical variables (one-hot, label encoding)
- Aggregation of data
- Discretization (binning continuous variables)
- Feature construction (creating new features)
- Normalizing data to a specific range
- Handling skewed data distributions

- Dimensionality reduction before modeling

Design me a case study for the concept Data preprocessing and cleaning Data transformation and normalization . My students already know various methods of data cleaning and preprocessing. My students don't know when to apply which method. Use the principles: Handle noisy data, Detect and remove duplicates, Outlier detection and treatment, Ensure data integrity and accuracy. Design the case study which is relevant to the current generation students. Don't use any technical terms in the case study. Use Finks Taxonomy.

Case Study Title:

“The Campus Fest Playlist Challenge”

Scenario

Your university is hosting a **Campus Fest** next month, and you’ve been asked to create the **official music playlist** for the event.
You collect song suggestions from:

- A **Google Form** circulated among students
- **Social media polls** (Instagram, X/Twitter)
- **WhatsApp group chats** from different clubs

When you put all the collected songs into one spreadsheet, you notice a problem:

- Some songs appear **multiple times** in slightly different spellings.
- Some song entries have **typos or incomplete names**.
- A few suggestions include **weird symbols or random text** instead of a song name.
- Some students added songs that **don’t exist on Spotify or YouTube**.
- The loudness/energy level of songs varies wildly — some are so soft they’ll get lost in the crowd noise, while others are way too intense for a campus gathering.

Learning Objectives (Fink’s Taxonomy)

1. **Foundational Knowledge**
   - Recall the methods you know for cleaning and preparing messy information.
2. **Application**
   - Apply the right cleaning method to each problem in the playlist scenario.
3. **Integration**
   - Connect the choices you make to the quality of the final playlist experience.
4. **Human Dimension**
   - Realize how your cleaning decisions can affect audience mood and event success.
5. **Caring**
   - Appreciate the importance of organizing information so everyone enjoys the music.

6. **Learning How to Learn**
    - Develop the habit of asking: *"What exactly is the problem here, and which method fits best?"* instead of blindly applying every method.

Student Task

1. **Identify** each type of problem in the playlist (duplicates, errors, irrelevant entries, unbalanced loudness).
2. **Decide** which cleaning or transformation approach fits each case.
3. **Justify** why you chose that method over others.
4. **Explain** how your decisions will make the playlist better for the audience.

Hints for Students (without revealing terms)

- If something appears more than once but in different disguises, think "merge."
- If something is broken, think "fix or remove."
- If something doesn't belong, think "filter out."
- If something is too soft or too loud, think "balance."

Rebuild the case study using SOLO taxonomy.

Case Study Title:

"The Campus Fest Playlist Challenge"

Scenario

You're in charge of creating the **official music playlist** for next month's **Campus Fest**. Song suggestions have come from:

- Google Form responses
- Instagram & Twitter polls
- WhatsApp group chats from different student clubs

When you put them together:

- Some songs appear **multiple times** in slightly different spellings.
- Some entries have **typos or incomplete titles**.
- A few entries have **symbols or random words** instead of song names.
- Some songs **don't exist on streaming platforms**.

- Loudness levels vary a lot — some songs are too quiet, others are overly loud for a friendly gathering.

Your goal: Prepare a playlist that is **fun, clean, and balanced** so the event runs smoothly.

SOLO Taxonomy Stages

1. Prestructural

Students recognize there's a problem with the playlist but cannot yet decide how to handle each issue.
Prompt: "What problems do you notice in the song list?"

---

2. Unistructural

Students can identify **one** type of problem and think of a single method to fix it.
Prompt: "Pick one problem from the playlist and suggest a way to fix it."

---

3. Multistructural

Students identify **multiple** problems but treat each one separately without connecting them.
Prompt: "List all the different problems you see and a matching way to fix each one."

---

4. Relational

Students connect the problems to see how fixing one can affect the others, and how all cleaning steps fit together for the final playlist quality.
Prompt: "How do your fixes work together to make the playlist more enjoyable for the audience?"

---

5. Extended Abstract

Students go beyond the current playlist — they generalize the method selection process so they can apply it in **other messy-data situations** in life or work.