Name ________________________________

Would You Eat Off This Data Table?

Before an AI can learn, its data has to be **clean.** Messy data - with missing labels, duplicates, or inconsistent entries - can totally confuse an AI model. Think of it like cooking: if your ingredients are spoiled or mislabeled, the final dish won't taste right. Your job is to examine the messy dataset below, find what's wrong, and list the steps you'd take to clean and organize it before labeling.

The Messy Dataset - Below is part of the "raw data" for an AI that's supposed to analyze **customer snack reviews**. Read carefully - there are duplicates, mistakes, and inconsistencies all over.

Raw Snack Review Data

Snack: "Choco Bites" | Review: "Loved it!" | Label: Positive

Snack: "choco bites" | Review: "loved it" | Label: POS

Snack: "Salties" | Review: "Too salty salty salty" | Label: negative

Snack: "?" | Review: "Yum." | Label:

Snack: "Fruit Crunch" | Review: "Too sour, but maybe okay" | Label: neutral

Snack: "Salties" | Review: "TOO SALTY SALTY SALTY" | Label: neg

Snack: "Fruit Crunch" | Review: "To sour but maybe ok" | Label: N/A

Snack: "Trail Mix" | Review: "fine, I guess" | Label: neutral

Snack: "Fruit Crunch" | Review: "Too sour but maybe okay" | Label: Neutral

Snack: "Choco Bites" | Review: "Loved it!" | Label: Positive

Snack: "Trail mix" | Review: "Fine I guess" | Label: NUETRAL

Snack: "Salties" | Review: "Tooo salty!!" | Label: Negative?

Snack: "Salties" | Review: "too salty salty salty" | Label: neg.

1. Spot the Problems: What kinds of errors or inconsistencies do you see in this dataset? (List at least three.)

2. Duplicates: Which entries look like duplicates or near-duplicates that might confuse the AI?

3. Missing or Incorrect Labels: What reviews don't have a clear or correct label? How would you handle them?

4. Standardizing the Data: What rules would you create to make sure all data entries use the same format (spelling, capitalization, labeling style, etc.)?

© GetWorksheets.com