Data Quality Issues

Lesson

AI Academy

Data Quality Considerations

- What kinds of data quality problems might we have?
- How can we detect them?
- What can we do about them?

Examples of data quality problems:

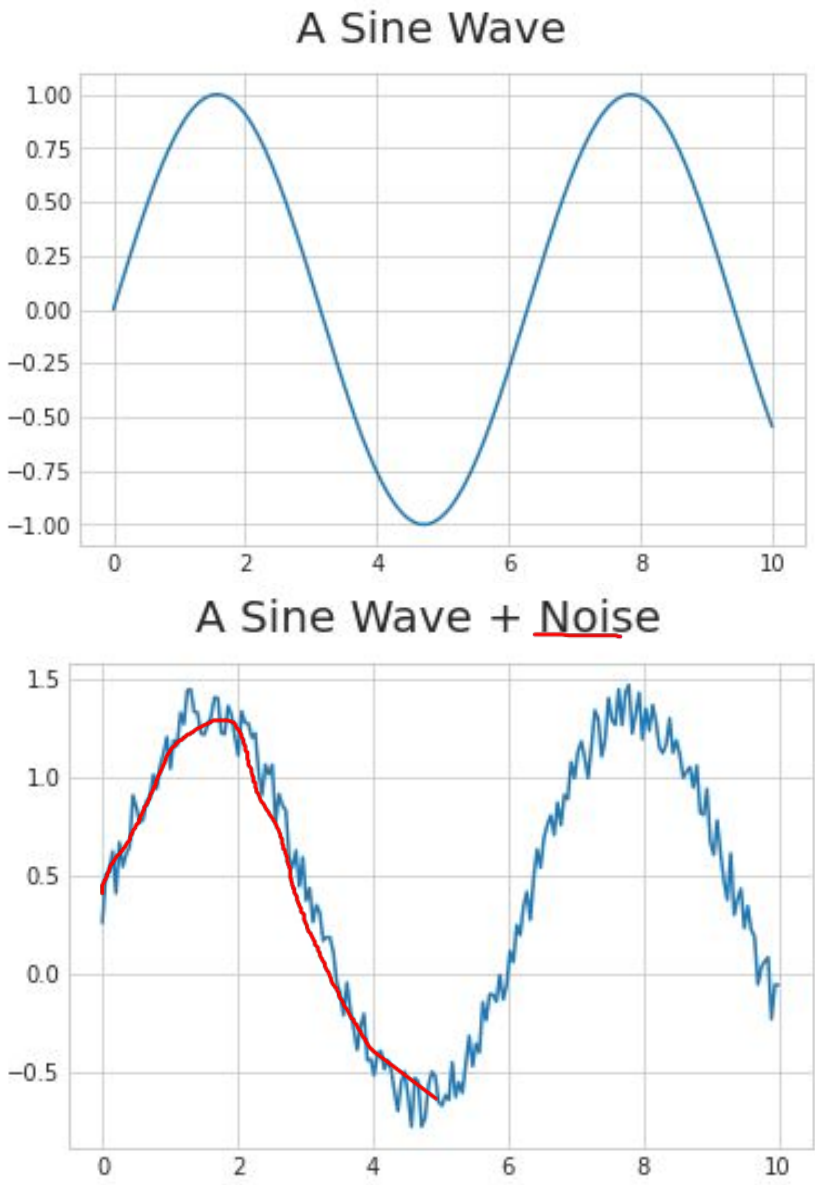
- Noise
- Outliers
- Missing values
- Duplicate data

Noise

Noise refers to a modification of original attribute values

Examples:

- instrument error
- distortion of audio
- unexplainable variation



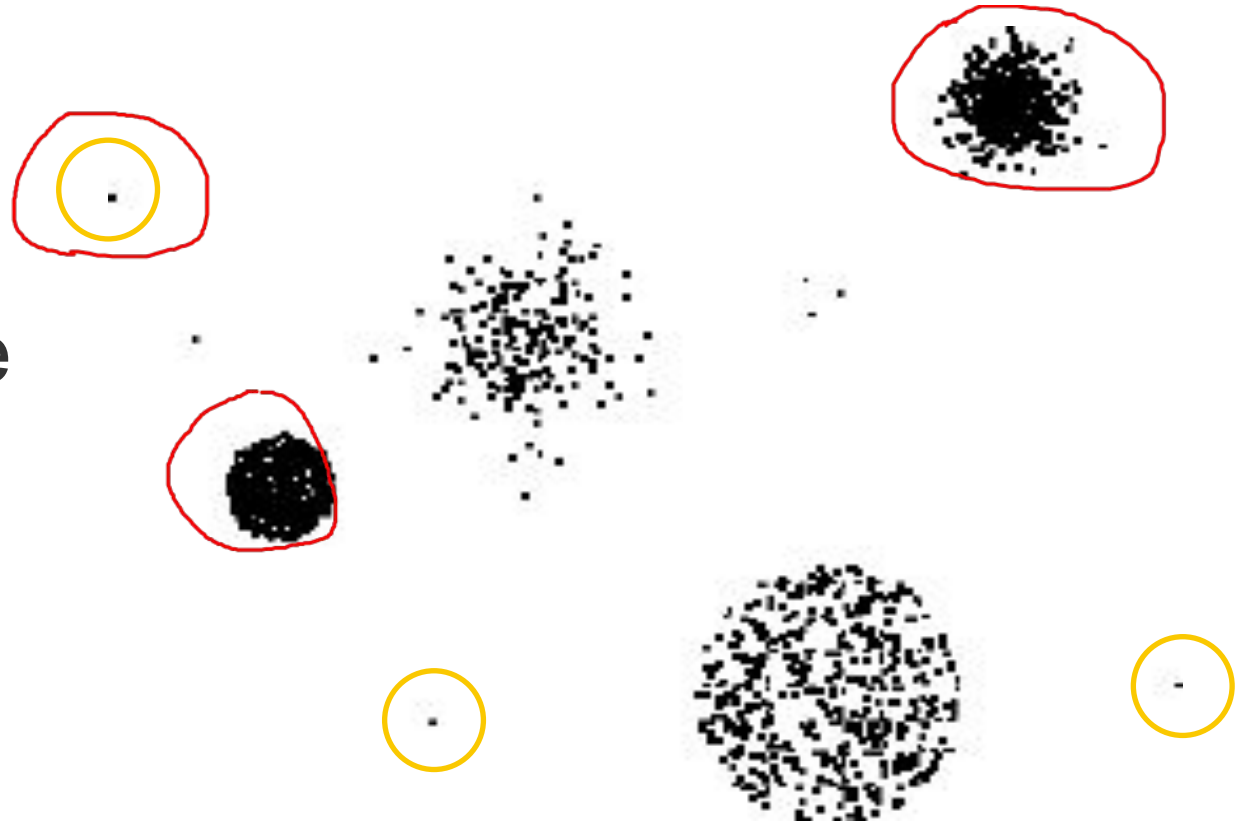
Addressing Noise

Almost all data has noise – what can you do?

- **Identify** noise, e.g. using visualization.
- **Remove attributes** with too much noise.
- Ensure models do not **overfit** to noisy data points.

Outliers

Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set



Deviation/Anomaly Detection

- Detect significant deviations from normal behavior
- Applications
  - Credit Card Fraud Detection
  - Network Intrusion Detection
- *Typical network traffic at a University may reach over 100 million connections per day*

Addressing Outliers

- **Identify** outliers using visualization or summary statistics.
- **Remove outliers** if they do not inform your analysis.
- Use summary statistics that are **robust** to outliers (e.g. median).

Missing Values

Reasons for missing values

- Information is not collected
  e.g. people may decline to give their weight
- Attributes may not be applicable to all cases
  e.g. children don't have annual income

Ways to handle missing values

- Eliminate the entire object
- Estimate missing values
- Ignore missing values during analysis
- Replace with all possible values (weighted by their probabilities)

Duplicate Data

Data may include data objects that are duplicates, or almost duplicates of one another

- Major issue when merging data from diverse sources

- **Example**: Same person with multiple email addresses

Data Cleaning: The process of addressing data issues (duplicates, missing values, etc.)

Learning Objectives: Data Quality Issues

You now should be able to:

Apply strategies to address **data quality issues**

- E.g., Explain the significance of noise and outliers

AI Academy
NC STATE

AI Academy

Data Quality Issues

Exercises

AI Academy

Practice Question: Operations

You have a **weight** attribute that was measured with a scale that has an error of +/- 2 grams. This is most likely to introduce _____ into your data?

A. Noise
B. Outliers
C. Missing Data
D. Duplicate Data

AI Academy

go.ncsu.edu/aiacademy

NC STATE UNIVERSITY