جامـــــعة المـــــستقبل
AL MUSTAQBAL UNIVERSITY

كلية العلـوم

قســــم الأنظمة الطبية الذكية

المحاضرة الرابعة

Data Cleaning and Transformation

المادة: Simulation and Modeling

المرحلة: الرابعة

اسم الاستاذ: م.م هادي صلاح هادي

Introduction to Data Cleaning

- Data Cleaning is the process of identifying and correcting errors, inconsistencies, or missing parts in the data to ensure quality and reliability.
- In Medical Simulation:
- Real-world data often contains errors due to sensor faults or manual entry.
- Cleaning ensures simulation accuracy and model safety.
- Without cleaning, simulation outcomes may be misleading or unsafe.
- Key Idea: Clean data = reliable simulation results.

Why Cleaning is Essential

➨ 1. Improves Accuracy:

- Unclean data leads to wrong predictions and simulation errors.

➨ 2. Reduces Bias:

- Cleaning removes abnormal or duplicated entries.

➨ 3. Ensures Consistency:

- Standardizing units (e.g., °C, mg/dL, bpm) avoids confusion.

➨ 4. Builds Trust:

- Reliable data helps healthcare professionals depend on model results.

Example (Conceptual):

➨ If ECG readings have noise or missing signals → the heart rhythm simulation will be inaccurate.

Common Data Problems in Medical Datasets

Problem Type	Description	Example
Missing Values	Data not recorded due to device/sensor error	Glucose = NaN
Outliers	Unusual data far from the normal range	Blood pressure = 500 mmHg
Duplicate Records	Same patient data repeated	ID = 102 twice
Incorrect Format	Data stored as text instead of number	“Ninety” instead of 90
Noise	Random fluctuations in signals	ECG waveform distortion

Handling Missing Values

- Missing values occur when no data value is stored for a variable in a dataset. In medical datasets, they often appear due to:
- Sensor malfunction (e.g., failed ECG reading)
- Human error in data entry
- Data transmission loss
- Patient absence from measurement

Methods to Handle Missing Data

- In medical datasets, missing data must be handled carefully using mathematical techniques to avoid bias and maintain the accuracy of simulations.

1. Deletion Methods

2. Imputation Methods

3. Regression Imputation

1. Deletion Methods

a. Listwise Deletion: Removes entire records (rows) with any missing values.

Mathematical Model: $D' = D - \{x_i \mid \exists j,\, x_{ij} = NaN\}$

Advantages: Simple to implement.

Disadvantages: May reduce dataset size significantly.

b. Pairwise Deletion: Uses all available data for each pair of variables.

Formula: $Cov(X,\, Y) = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n_{xY}}$

Advantages: Keeps more data.

Disadvantages: Inconsistent sample sizes across variables.

2. Imputation Methods

a. **Mean Imputation:** Replace missing values with the mean of observed data.

Formula: $x_{ij} * \; = \left(\frac{1}{n_j}\right) \Sigma\, x_{ij}$

- Advantages: Preserves sample size.
- Disadvantages: Reduces variability and may bias correlations.

b. **Median Imputation:** Replace missing values with the median.

Formula: $x_{ij} * \; = \; Median(x_j)$

- Advantages: More robust to outliers.
- Disadvantages: Still assumes missing values are random.

c. **Mode Imputation:** Replace missing values with the mode (most frequent value).

$$\text{Formula: } x_{\mathrm{ij}} * \ = Mode(x_{\mathrm{j}})$$

- Advantages: Best for categorical variables (e.g., gender, diagnosis).
- Disadvantages: May oversimplify category diversity

3. Regression Imputation

Predict missing values based on relationships between variables:

$$\text{Formula: } x_{\mathrm{ij}} * \ = \beta_0 + \beta_1 x_{\mathrm{i1}} + \beta_2 x_{\mathrm{i2}} + \cdots + \beta_{\mathrm{k}} x_{\mathrm{ik}}$$

- Advantages: More accurate for correlated medical features.
- Disadvantages: Assumes linearity and may introduce model bias.

Detecting Outliers Using Statistical Methods

- Outliers are data points that deviate significantly from the majority of the dataset.
- In medical data, outliers may represent measurement errors or rare clinical conditions**.**

Why Detect Outliers?

- To improve model accuracy and simulation stability.
- To remove data noise caused by faulty sensors.
- To detect potential anomalies (e.g., abnormal heart rate or glucose spike).

1. **Z-Score Method (Standard Score)**

The Z-score measures how far a value is from the mean in terms of standard deviations.

$$\text{Formula: } Z_{\mathrm{i}} = \frac{(X_{\mathrm{i}} - \mu)}{\sigma}$$

Where:

X_{i} = individual data point

μ = mean of the data

σ = standard deviation

Rule of Thumb: If $|Z_{\text{i}}| > 3$, the point is considered an outlier.

2. IQR Method (Interquartile Range)

The IQR method identifies outliers based on the spread of the middle 50% of the data.

Formula: Where:$IQR = Q_3 - Q_1$

Q_1 = First quartile (25th percentile)

Q_3 = Third quartile (75th percentile)

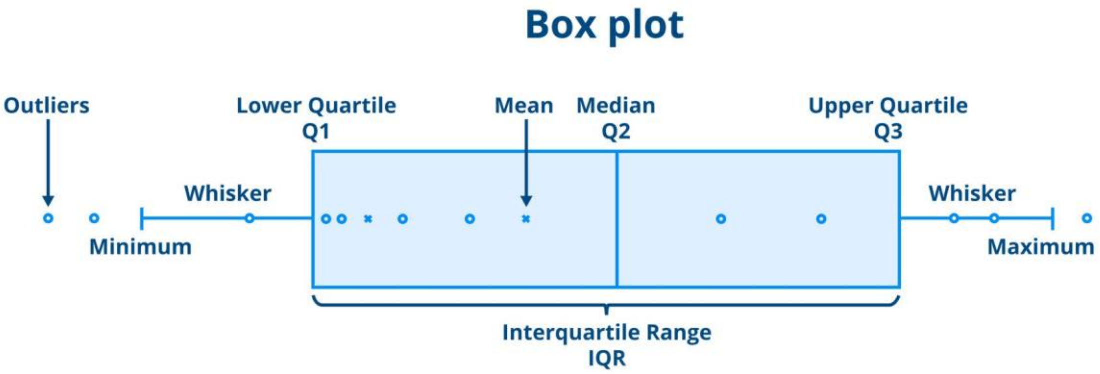
A data point X_{i} is considered an outlier if:

$$X_{\text{i}} < Q_1 - 1.5 \times IQR \ or \ X_{\text{i}} > Q_3 + 1.5 \times IQR$$

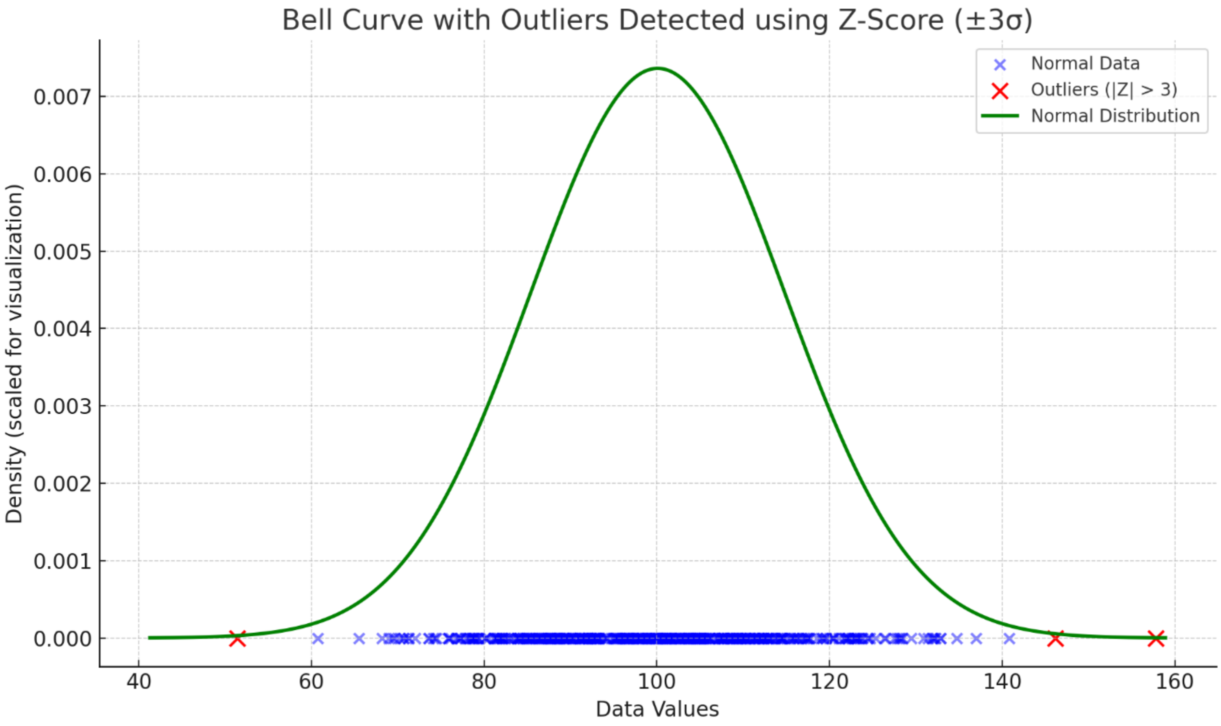
Comparison Table:

Method	**Basis**	**Formula**	**Best Used For**
Z-Score	**Mean & Standard Deviation**	$(\boldsymbol{Z} = \frac{(\boldsymbol{X} - \boldsymbol{\mu})}{\boldsymbol{\sigma}})$	**Normally distributed data**
IQR	**Quartiles (25%, 75%)**	$(\boldsymbol{X} < \boldsymbol{Q\_1} - \mathbf{1.5} \times \boldsymbol{IQR}) \ \boldsymbol{or} \ (\boldsymbol{X} > \boldsymbol{Q\_3} + \mathbf{1.5} \times \boldsymbol{IQR})$	**Skewed or non-normal data**

Boxplot



Bell Curve with Outliers Detected using Z-Score (±3σ)



Data Transformation

Data Transformation is the process of converting raw or cleaned data into a suitable format or scale for analysis, modeling, and simulation. It includes scaling, normalizing, encoding, and applying mathematical transformations such as logarithmic or standardization functions

Purpose in Medical Simulation

Goal	Explanation
Consistency	Ensures all variables are in compatible scales (e.g., heart rate vs. glucose level).
Improved Accuracy	Prevents bias caused by large-valued features dominating smaller ones.
Better Model Training	Helps algorithms converge faster and perform more accurately.
Smooth Simulation Behavior	Reduces instability in medical system models.

Mean and Standard Deviation

population is known

$$\mu = \frac{\sum_i^N xi}{N}$$

$$\sigma = \sqrt{\sum \frac{(xi - \mu)^2}{N}}$$

population is unknown

$$\bar{x} = \frac{\sum_i^N xi}{N-1}$$

$$\sigma = \sqrt{\sum \frac{(xi - \bar{x})^2}{N-1}}$$

- **HW:** Write a short paragraph (5–6 lines) explaining why the denominator is (N−1) for a sampleand give a small numeric example in your answer.

Covariance and Variance

- Covariance quantifies the extent to which two random variables exhibit simultaneous variation.
- If two variables have a tendency to rise or decrease in a similar manner, their covariance is positive. Conversely, if one variable increases while the other decreases, their covariance is negative.

Population mean is known

$$cov(x, y) = \frac{\sum_i^n (x_i - \mu) \cdot (y_i - \mu)}{N}$$

Population mean is unknown

$$cov(x) = \frac{\sum_i^n (x_i - \overline{x}) \cdot (y_i - \overline{y})}{N-1}$$

- Variance quantifies the extent to which a single random variable deviate from its average value.
- population is know: $\sigma^2 = \sum \frac{(xi-\mu)^2}{N}$, population is unknow :$\sigma^2 = \sum \frac{(xi-\bar{x})^2}{N-1}$

Mathematical Concept

- Let X be a raw variable, and X' the transformed value: $X' = f(X)$
- Where f is a transformation function, such as:
- Standardization: $X' = (X - \mu) / \sigma$
- Normalization: $X' = (X - X_{\min}) / (X_{\max} - X_{\min})$
- Log Transformation: $X' = log\ (X)$

Simple Medical Example

Raw glucose levels (mg/dL) range from 70 to 300. After normalization, values are scaled between 0 and 1. This makes the model treat glucose level like any other variable, such as heart rate**.**

Example: Standardization of Heart Rate Data

- Let’s consider three heart rate readings (in beats per minute): $X = [60,\ 80,\ 100]$

Step 1: Calculate the Mean (μ)

$$\mu = \frac{(60\ +\ 80\ +\ 100)}{3} = 80$$

Step 2: Calculate the Standard Deviation (σ)

$$\sigma = \sqrt{\left[\frac{(\,(60-80)^2 +\ (80-80)^2 +\ (100-80)^2)}{3}\right]}$$

$$\sigma = \sqrt{\left[\frac{(400 + 0 + 400)}{3}\right]} = \sqrt{266.67} \approx 16.33$$

Step 3: Apply the Standardization Formula

$$X' = \frac{(X - \mu)}{\sigma}$$

Original Value (X)	Calculation	Standardized (X')
60	$(60 - 80) / 16.33$	-1.22
80	$(80 - 80) / 16.33$	0.00
100	$(100 - 80) / 16.33$	$+1.22$

Final Result

$$X' = [-1.22,\ 0.00,\ +1.22]$$

- Interpretation

- Value 0 represents the mean (normal reading).
- Negative value (-1.22) indicates below-average heart rate.
- Positive value $(+1.22)$ indicates above-average heart rate.
- After standardization, all data are represented on the same Z-score scale, making them easier to compare and analyze in medical models.

Example: Normalization of Blood Pressure Data

- Let’s consider systolic blood pressure readings (mmHg):

$$X = [90, 110, 130, 150]$$

- Step 1: Identify Minimum and Maximum Values

$$X_{\min} = 90, \quad X_{\max} = 150$$

- Step 2: Apply Normalization Formula

$$X' = \frac{(X - X_{\min})}{(X_{\max} - X_{\min})}$$

- Step 3: Compute for Each Value

Original (X)	Calculation	Normalized (X')
90	$(90 - 90) / (150 - 90)$	0.00
110	$(110 - 90) / (150 - 90)$	0.33
130	$(130 - 90) / (150 - 90)$	0.67
150	$(150 - 90) / (150 - 90)$	1.00

- **Final Result**

$$X' = [0.00, 0.33, 0.67, 1.00]$$

Interpretation

- The lowest value (90) becomes 0, and the highest (150) becomes 1.
- All other values are scaled between 0 and 1 according to their relative distance.
- This makes the data comparable across features like glucose or heart rate, ensuring fair weighting in medical models and simulations.

Example: Log Transformation

- Let's consider the white blood cell count (WBC) in units of $10^3/\mu L$:

$$X = [4,\ 10,\ 20,\ 40]$$

- Step 1: Apply the Log Transformation

$$\text{Formula: } X' = log_{10}(X)$$

- Step 2: Compute for Each Value

Original (X)	Calculation	Transformed (X′)
4	$log_{10}(4)$	0.60
10	$log_{10}(10)$	1.00
20	$log_{10}(20)$	1.30
40	$log_{10}(40)$	1.60

Final Result

$$\boldsymbol{X' = [0.60, 1.00, 1.30, 1.60]}$$

Interpretation

- The log transformation compresses large values, reducing data skewness.
- This is particularly useful in medical data where some lab results vary widely.
- After transformation, all values become closer in scale, allowing fairer analysis in models.

Before the log transformation, the original data values were:

$$X = [4,\ 10,\ 20,\ 40]$$

- The value 40 is almost ten times greater than 4, which causes it to dominate the dataset. This means that statistical models or simulations might focus too much on large values, distorting the overall analysis.

After applying the logarithmic transformation:

$$X' = [0.60,\ 1.00,\ 1.30,\ 1.60]$$