Data Preprocessing

Process of converting raw data into a clean and usable format

- Improves data quality, accuracy, and reliability
- **Data Cleaning**
  - Handle missing values
  - Remove duplicates and outliers
  - Fix inconsistent data
- **Data Integration**
  - Combine data from multiple sources
  - Resolve data conflicts
- **Data Transformation**
  - Normalization and standardization
  - Encoding categorical variables
  - Feature construction
- **Data Reduction**
  - Feature selection
  - Dimensionality reduction (PCA, sampling)
- **Data Discretization**
  - Convert continuous data into intervals
- Essential step before data analysis and machine learning

Dimensionality Reduction

- **Min–Max Normalization/ Column normalization**
- **Definition**: Feature scaling technique that rescales data to a fixed range, usually **[0, 1]**
- Used to bring all numerical features to a **common scale**
- **Formula**:
- $$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$
- **Key Characteristics**:
  - Preserves original data distribution
  - Maintains relative distances between values
- **Advantages**:
  - Easy to understand and implement
  - Useful for distance-based algorithms
- **Limitations**:
  - Highly sensitive to outliers
  - New data outside range may distort scaling

Column Standardization

- **Column Standardization (Z-Score Standardization)**
- **Definition**: Feature scaling technique that transforms data to have **mean = 0** and **standard deviation = 1**
- Used when features have **different units or scales**
- **Formula**: $$X' = \frac{X - \mu}{\sigma}$$
- **Key Characteristics**:
  - Centers data around zero
  - Reduces effect of scale differences
- **Advantages**:
  - Works well for normally distributed data
  - Less sensitive to outliers than Min–Max normalization
- **Limitations**:
  - Does not bound values to a fixed range
  - Assumes meaningful mean and variance

Covariance Matrix

- **Definition**: A square matrix that shows **covariance between pairs of variables** in a dataset
- Describes how variables **vary together**

$$\begin{bmatrix} \mathrm{Var}(x_1) & \cdots & \mathrm{Cov}(x_n, x_1) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(x_n, x_1) & \cdots & \mathrm{Var}(x_n) \end{bmatrix}$$

- ***Sample Variance:*** *var(x_1) =* $\frac{\sum_1^n (x_i - \overline{x})^2}{n-1}$
- ***Sample Covariance:*** *cov(x_1, y_1) =* $\frac{\sum_1^n (x_i - \overline{x})(y_i - \overline{y})}{n-1}$
- ***Population Variance:*** *var(x_n) =* $\frac{\sum_1^n (x_i - \mu)^2}{n}$
- ***Population Covariance:*** *cov(x_n, y_n) =* $\frac{\sum_1^n (x_i - \mu_x)(y_i - \mu_y)}{n}$

- **Key Properties**:
- Diagonal elements → variances
- Off-diagonal elements → covariances
- Symmetric matrix
- **Interpretation**:
- Positive covariance → variables increase together
- Negative covariance → one increases, other decreases
- Zero covariance → no linear relationship

How to find covariance matrix?

Student	Psychology(X)	History(Y)
Anna	80	70
Caroline	63	20
Laura	100	50

Step 1: *Find the mean of variable X. Sum up all the observations in variable X and divide the sum obtained with the number of terms. Thus, (80 + 63 + 100)/3 = 81.*

Step 2: *Subtract the mean from all observations. (80 - 81), (63 - 81), (100 - 81).*

Step 3: *Take the squares of the differences obtained above and then add them up. Thus,* $(80 - 81)^2 + (63 - 81)^2 + (100 - 81)^2$.

Step 4: *Find the variance of X by dividing the value obtained in Step 3 by 1 less than the total number of observations.* $var(X) = [(80 - 81)^2 + (63 - 81)^2 + (100 - 81)^2] / (3 - 1) = 343$.

Step 5: *Similarly, repeat steps 1 to 4 to calculate the variance of Y. Var(Y) = 633.333*

Step 6: *Choose a pair of variables.*

Step 7: *Subtract the mean of the first variable (X) from all observations; (80 - 81), (63 - 81), (100 - 81).*

Step 8: *Repeat the same for variable Y; (70 - 47), (20 - 47), (50 - 47).*

Step 9: *Multiply the corresponding terms: (80 - 81)(70 - 47), (63 - 81)(20 - 47), (100 - 81)(50 - 47).*

Step 10: *Find the covariance by adding these values and dividing them by (n - 1).* $Cov(X, Y) = [(80 - 81)(70 - 47) + (63 - 81)(20 - 47) + (100 - 81)(50 - 47)]/(3-1) = 260$.

Step 11: *Use the general formula for the covariance matrix to arrange the terms.*

The matrix becomes: $\begin{bmatrix} 343 & 260 \\ 260 & 633.333 \end{bmatrix}$