**BMEG3105 Data Analytics for Personalized Genomics and Precision Medicine**
**Lecture 5: Data Exploration**

17 September, 2025

Lecturer: Yu LI
Scriber: Lan Hoi Lee(1155191196)

# Data cleaning

## A. Recap

- A data matrix is a collection of records. Each record is made up of a fixed set of attributes.
- It can be shown as an n-row (rows = n) × m-column (columns = m) matrix. Each row stands for one data object, and each column stands for one attribute.
- If you swap all the elements of one entire column with another, or swap all the elements of one entire row with another, the actual data content will not change.

**1. Noise**

- Noise is a change to the original data value. It affects what the original data should be.

**2. Outliers**

- Outliers are data points that are very different from other data points.
- Sometimes they're just random or mistakes—they aren't useful.
- Some outliers give important information.

**3. Missing values**

Missing values mean some information wasn't collected or doesn't apply.
Solutions:

- Delete the whole data point that has missing values. Risk: You might delete a lot of data.
- Guess the missing values. Use assumptions (e.g., if two people have similar height, they might have similar weight).
- Ignore the missing values.
- Replace missing values with all possible answers. Use how likely each answer is (weighted by probability).

**4. Duplicate data**

- Duplicate data happens when you combine two datasets—you might get the same data more than once.

**5. Unnormalized data**

Unnormalized data can't be compared easily.

- We need comparable data to calculate things like norms or Euclidean distance.
- The different scales of unnormalized data will make one parameter more influential during calculations.
  Example: In gene expression studies—this fixes technical differences. Like if one sample is "sequenced too deep" (makes more copies of all reads for that cell/system).

Methods to fix:

- Min-max normalization: Makes data range from 0 to 1.

$$v' = \frac{v - v_{\min}}{v_{\max} - v_{\min}}$$

- Z-score normalization: Assumes data follows a gaussian/normal distribution.

$$v' = \frac{v - Mean(v)}{Std(v)}$$

- One-hot encoding: For example, split "gender" into two attributes, and use 0 or 1 for each.

Note: The order of data cleaning affects the final results.

**6. Categorical data**

- Categorical data is data that groups things (like "gender" or "color") instead of using numbers you can add/subtract.

# Data Exploration

## B. Summary Statistics

Summary statistics are numbers that describe the properties of data. They include frequency, location, spread, and measures like mean and standard deviation (SD).

**1. Measures of location**

- The mean is the most common way to show the "center" of a set of data points.
- It is sensitive to outliers.
- Alternatives to the mean: median or trimmed mean.

$$mean(x) = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$median(x) = \begin{cases} x_{(r+1)} & if\ m\ is\ odd \\ \frac{1}{2}\left(x_{(r)} + x_{(r+1)}\right) & if\ m\ is\ even \end{cases}$$

**2. Measures of spread**

- Range: the difference between the maximum and minimum values in the data.
- Variance or SD are the most common ways to show how spread out the data is.
- Variance and SD are sensitive to outliers.
- Alternatives to variance and SD: median absolute deviation (MAD) or interquartile range.

$$variance(x) = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - mean(x))^2$$

$$median(|x_1 - mean(x)|, \ldots, |x_m - mean(x)|) \qquad x_{75\%} - x_{25\%}$$

**3. Percentiles**

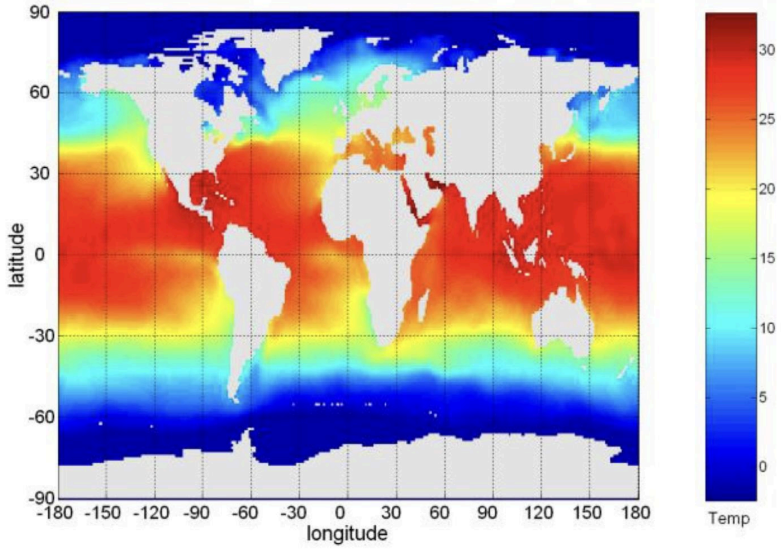
- Used for ordinal or continuous attributes (e.g., attribute x).
- The value of p (percentile) is between 0 and 100.
- The p-th percentile of x is a value (xp) where p% of the observed x values are less than xp.
- Example: For the data set [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], the 30th percentile is 4 (since 30% of the values, which are [1, 2, 3], are less than 4).

**4. Frequency and mode**

- Frequency: the percentage of time a specific value appears in the data set.
- Mode: the attribute value that appears most often in the data set.
- Both frequency and mode are usually used with categorical data.

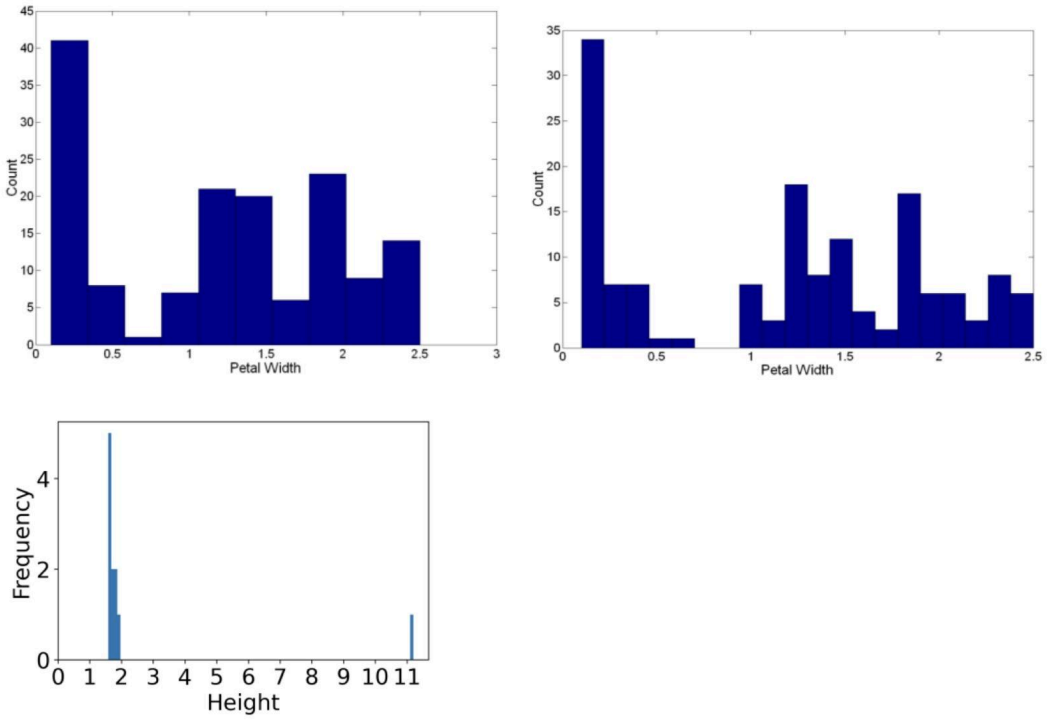
## C. Exploratory visualization

- It is used to analyze or report on two key things: the characteristics of the data, and the relationships between different items or attributes.
- With visualization, you can detect general patterns, trends, outliers, and unusual patterns in the data.
- Example: Using visualization to represent sea surface temperature (e.g., through color-coded maps or line graphs showing temperature changes over time).



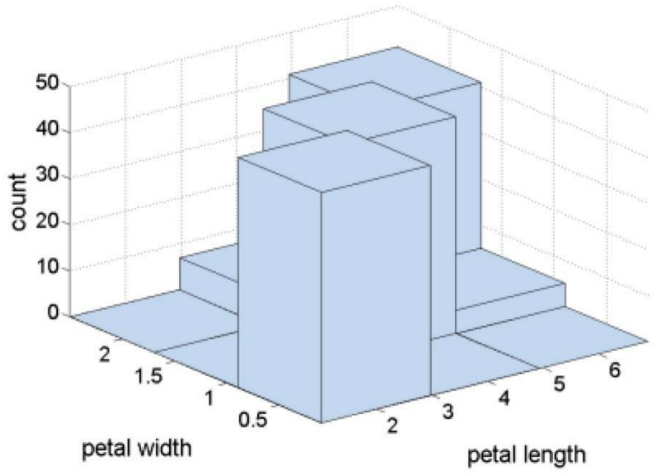

**1. Histograms**

- Histograms are used to visualize the distribution of values for a single variable.
- They work by dividing the variable's values into groups called "bins," then showing a bar plot where each bar represents the number of data objects in its corresponding bin.
- The height of each bar equals the number of objects in that bin.
- The overall shape of a histogram depends on how many bins are used (e.g., fewer bins make the shape broader, more bins show more detailed variations).




**2. 2D histograms**

- 2D histograms visualize the joint distribution of values from two attributes (i.e., how the two attributes' values occur together).
- Example: Using a 2D histogram to show the relationship between petal width and petal length (e.g., grouping petal width into one set of bins and petal length into another, then using color or height to represent how many data points fall into each combined bin of width and length).




### 3. Box plots

- Box plots are used for displaying the distribution of a single dataset and comparing the distributions of multiple datasets.
- They show key statistical measures (like median, quartiles, and potential outliers) in a compact format, making it easy to see differences in spread, central tendency, and range across different groups or variables.