Data exploration

Types of data:

- Sequential data – data where order of elements is important
  ex) DNA sequence

- Data matrix – data is organized into rows and columns
  ex) n x m matrix

- Spatial data – data that includes information about physical location or geometry of objects.
  ex) images

- Temporal data – data that involves time-related information
  ex) temperature, humidity over certain period

- Graph or network- data that represents relationship between entities

- Text – data in the form of written words

- Multi-modality data – data that combines multiple types of data
  ex) video, electronic health records

Data cleaning

Essential steps in the data preprocessing so that the data is ready for analysis.

Examples of data quality problems

1. Noise
2. Outlier
3. Missing values – solution: remove, estimate (mean)
4. Duplicate data – solution: remove
5. Unnormalized data – solution: Min-max or Z-score normalization
6. Categorical data – solution: one-hot encoding

Steps of data cleaning (order is important)

1. Denoise data (if applicable)
2. Remove outliers
3. Handle missing data
4. Remove duplicates
5. Categorical data encoding
6. Data normalization

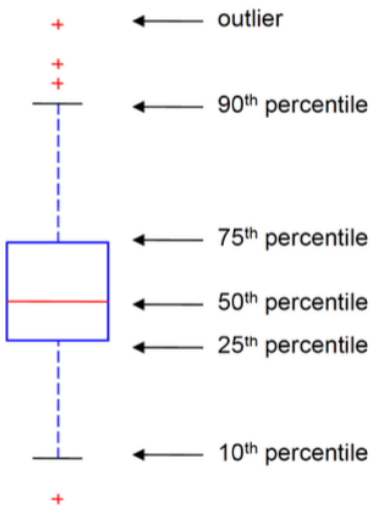
Summary statistics

Measure of location

- Mean - $\frac{1}{m}\sum_{i=1}^{m} x_i$ *sensitive to outliers
- Median - $\begin{cases} x_{(r+1)} & if\ m\ is\ odd, i.e.,\ \ m = 2r+1 \\ \frac{1}{2}\left(x_{(r)} + x_{(r+1)}\right) & if\ m\ is\ even,\ \ i.e.,\ \ m = 2r \end{cases}$

Measure of spread

- Range - Difference between Max and Min
- Variance / standard deviation - $\frac{1}{m-1}\sum_{i=1}^{m}\left(x_i - mean(x)\right)^2$
- Percentiles



  - Interquartile range – $x_{75\%} - x_{25\%}$

Frequency – the percentage of time the value occurs in data set

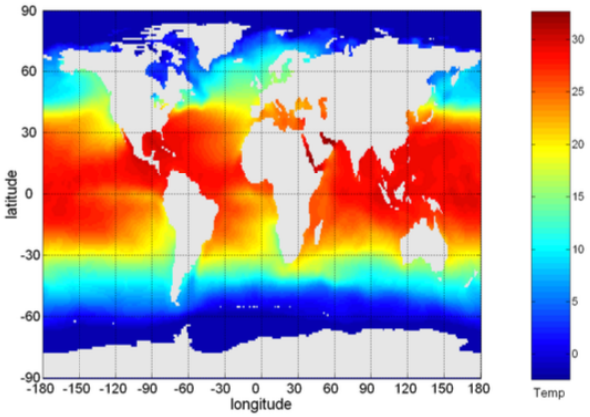
Mode – the most frequent attribute value

*Usually used in categorical data

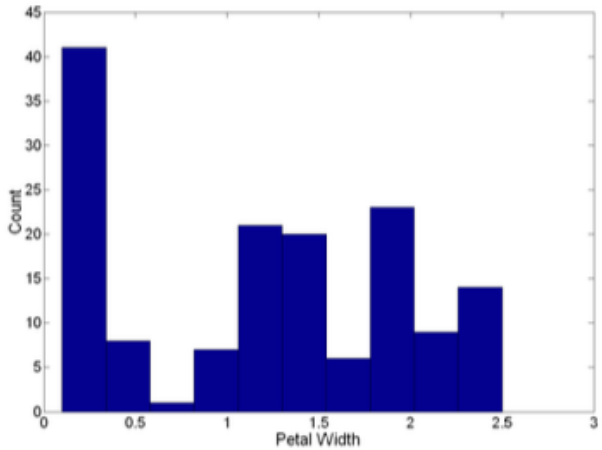
Visualization of data

- ◊ Can detect general patterns and trends
- ◊ Can detect outliers and unusual patterns

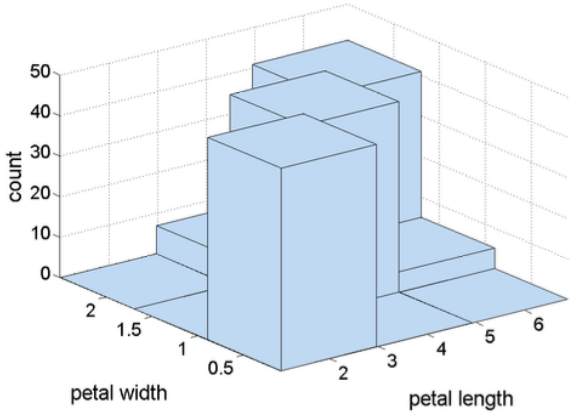
1. Sea Surface Temperature (SST)



2. Histograms



3. 2D - Histograms



4. Box Plots