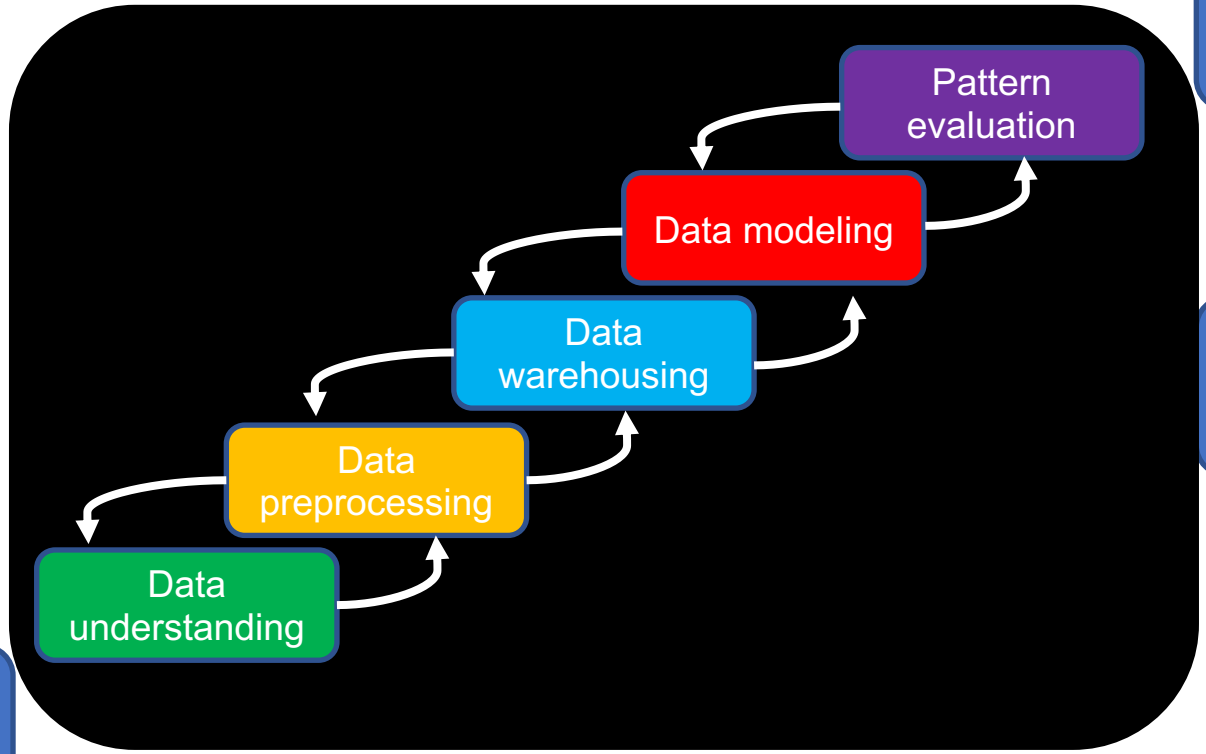
Data Understanding

Data Mining:
Data Mining Pipeline
with **Dr. Qin Lv**

Master of Science in Data Science
UNIVERSITY OF COLORADO **BOULDER**

Learning objective: Describe the key properties of data. Apply techniques to characterize different datasets.

Data Mining Pipeline



Application

Knowledge

Technique

Data

Data Understanding

- Data objects & attributes
- Data statistics
- Data visualization
- Data similarity

Dataset

- A collection of data objects
  - E.g., employee records, product catalog, online posts
- Each described by a number of attributes
  - Also referred to as features, dimensions, variables
  - E.g., employee: name, gender, age, salary, job title
  - E.g., online post: user, time, content, #likes, responses

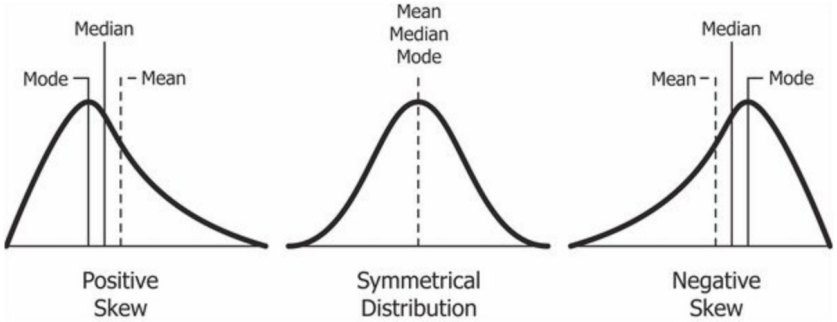
Attribute Types

- Categorical
  - Nominal, binary, ordinal
  - E.g., major, CS major, academic ranks
- Numeric: discrete or continuous
  - Interval-scale or ratio-scaled (true zero)
  - E.g., year 2000, number of users, annual income

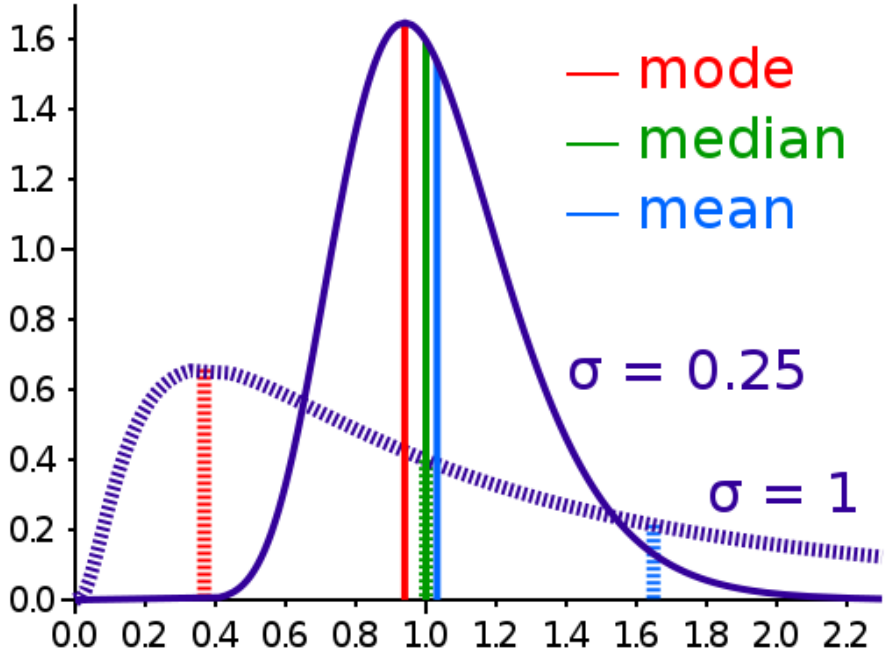
Data Statistics

- #objects, #attributes
- Distribution of each attribute's values
  - Categorical: % of each value
  - Numeric: central tendency, dispersion
- Comparison across attributes & datasets

Central Tendency



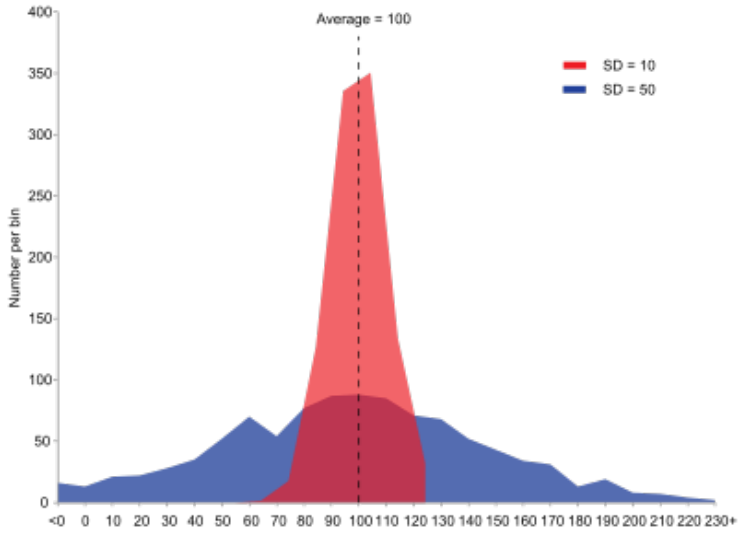
- Mean
- Median
- Mode
- Midrange
  - (Max – Min)/2



Dispersion

- How much a distribution is stretched or squeezed

- Range: max – min
- Quartiles: Q1 (25%), Q3 (75%)
- IQR (interquartile range): Q3 – Q1
- Variance
- Standard deviation



Data Visualization

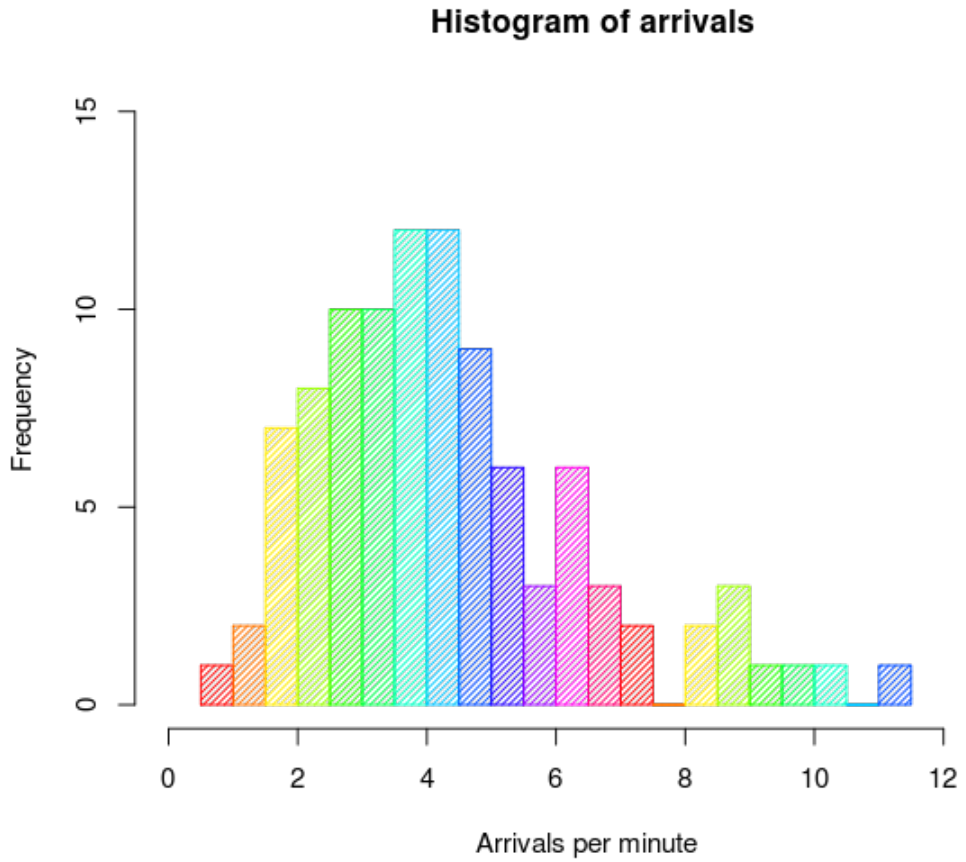
- Boxplot
- Histogram
- Quantile plot
- Q-Q plot
- Scatter plot
- …

Boxplot

- Box
  - Q1, Q2, Q3, IQR
- Whiskers
  - Min, max,
  - 1.5 x IQR
- Outliers

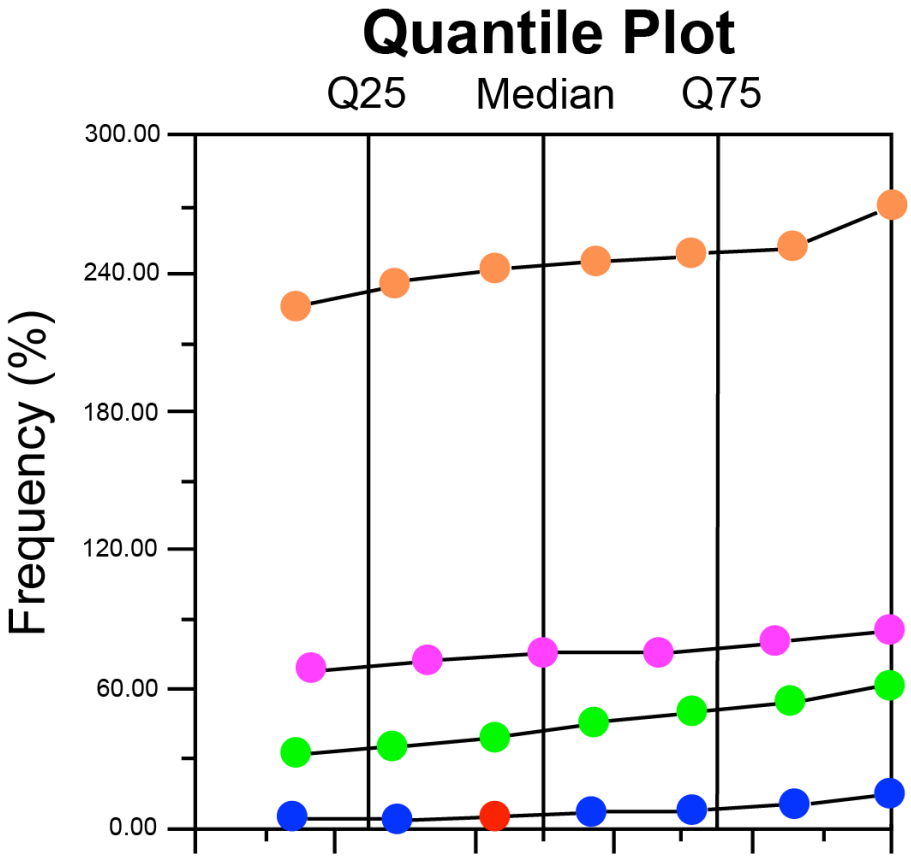
Histogram

- Bars of different height
- X: sub-range (bin grouping)
- Y: frequency (bar height)



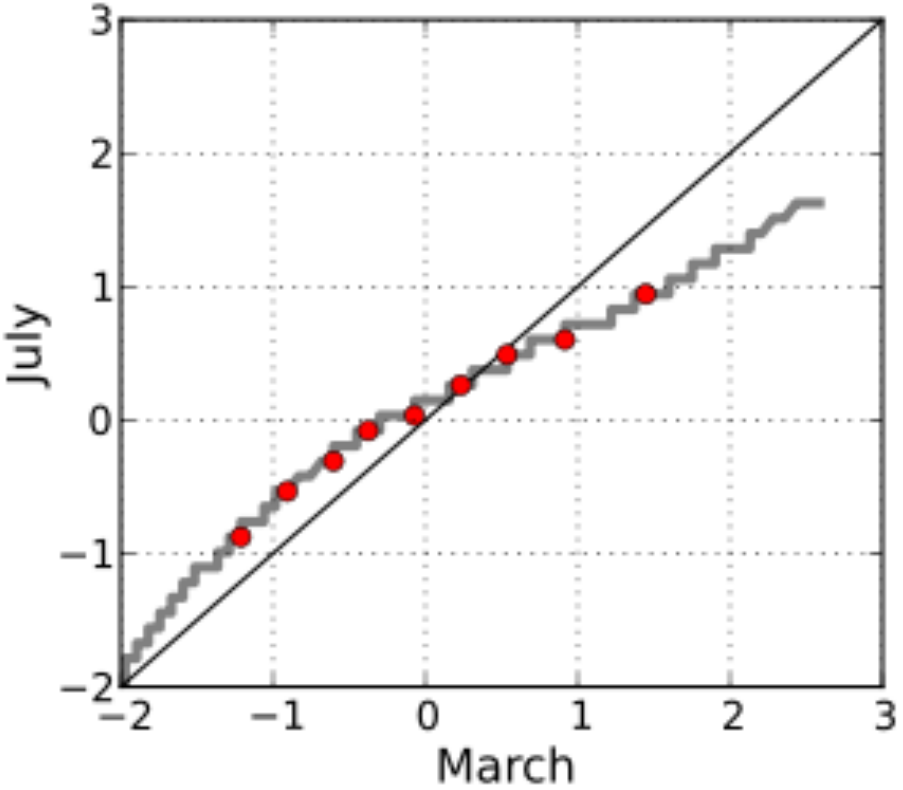
Quantile Plot

- Quantile: percent of points below the given value
- X: percent
- Y: quantile



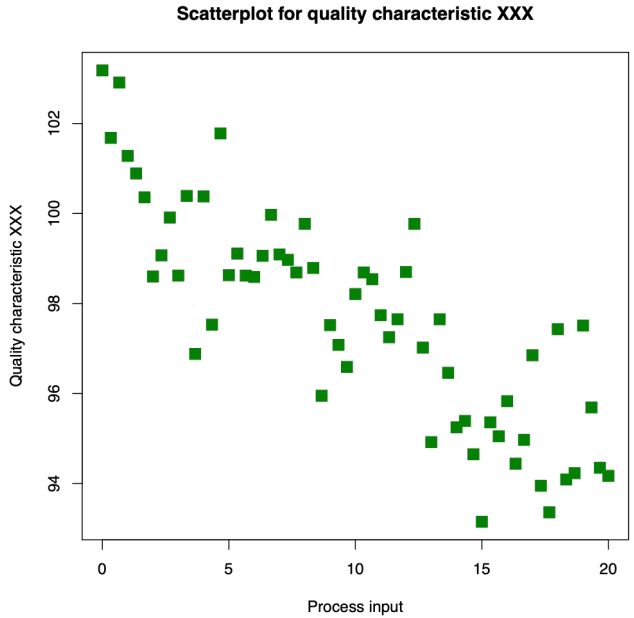
Quantile-Quantile (Q-Q) Plot

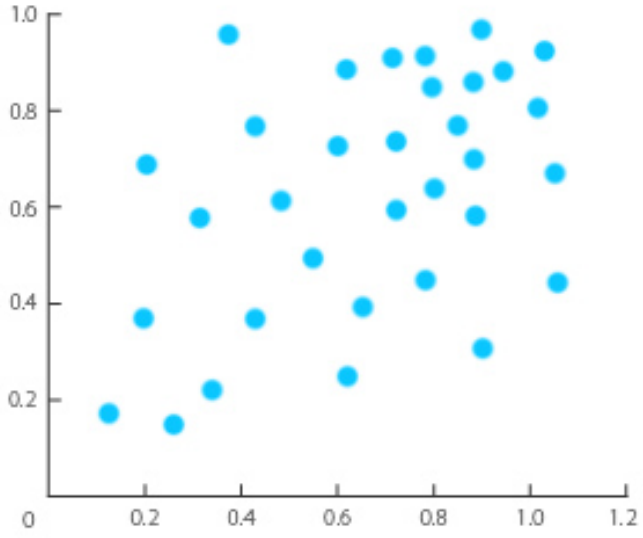
- Comparison of two quantiles
- 45-degree reference line



Scatter Plot

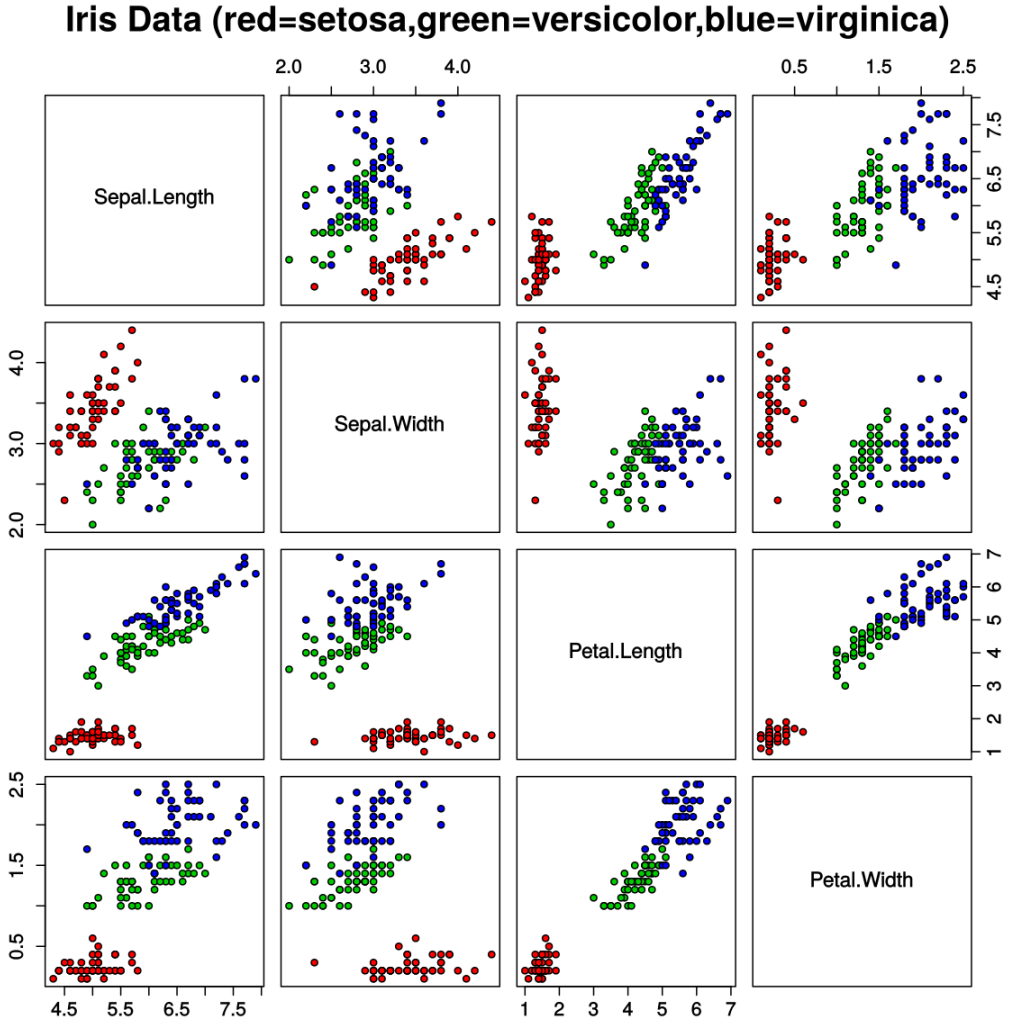
- Comparison of two attributes: X vs. Y





Scatter Plot

- Pairwise comparison across multiple attributes



Data Visualization

Data Visualization Methods

- Visualizing complex data & relations

- Chart type: line, pie, (stacked) bar, bubble, area, heatmap, word cloud, network, …
- Color, size, layout, hierarchy, ...

- Exploration vs. explanation
- Automation, interaction, efficiency & effectiveness