1.3 Data Understanding and Preparation

CRISP-DM

Data Understanding

- Collect Initial Data
- Describe Data
- Explore Data
- Verify Data Quality

Data Preparation

- Data Set
- Select Data
- Clean Data
- Construct Data
- Integrate Data
- Format Data

Data Understanding

Data Summarization

Motivation

- with big data sets it is hard to have an idea of what is going on in the data
- data summaries provide overviews of key properties of the data
- help selecting the most suitable tool for the analysis
- their goal is to describe important properties of the distribution of the values

Types of Summaries

- What is the "most common value"?

- What is the "variability" in the values?

- Are there "strange"/unexpected values in the data set?

- **Data set**: univariate data or multivariate data

- **Variables**: categorical variables or numeric variables

Categorical Variables

- **Mode**: the most frequent value
- **Frequency table**: frequency of each value (absolute or relative)
- **Contingency table**: cross-frequency of values for 2 variables

Numeric Variables

Statistics of location

- **Mean (or sample mean)** - sensitive to extreme values
- **Median** - 50th percentile
- **Mode** - most common value

Statistics of variability of dispersion

- **Range** - $max_x - min_x$
- **Variance** - σ_x^2 - sensitive to extreme values
- **Standard Deviation** - sensitive to extreme values

$$\sigma_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_X)^2}$$

- **Inter-quartile Range (IQR)** - difference between the 3rd (Q_3) and 1st (Q_1) quartiles
  - Q_1 → nr < 25%
  - Q_3 → nr < 75%

Outliers

- For a numeric variable, an outlier can be an extreme value
- In the presence of such values:
  - **median** or **mode** are more robust as a central tendency statistic
  - **interquartile range** is more appropriate as a variability statistic
- **Boxplot Definition** - any value in the interval $[Q_1 - 1.5 * IQR, Q_3 + 1.5 * IQR]$ is an outlier

Multivariate analysis of variability or dispersion

- **Covariance Matrix** - variance between every pair of numeric variables - the value depend on the magnitude of the variable

$$cov(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

- **Correlation Matrix** - correlation between every pair of numeric variables - the influence of the magnitude is removed

$$cor(x, y) = cov(x, y)/\sigma_x\sigma_y$$

- **Pearson Correlation Coefficient (ρ)** - measures the linear correlation between 2 $\in$ [-1, +1]
- **Spearman Rank-Order Correlation Coefficient** - measure the stregth and direction of monotonic association between 2 variables; rank-based and non-parametric version of *Pearson*

Data Visualization

Motivation

- Humans are outstanding at detecting patterns and structures with their eyes
- Data visualization methods try to explore these capabilities
- Help detecting patterns and trends, and also outliers and unusual patterns

Main Types of Graphs

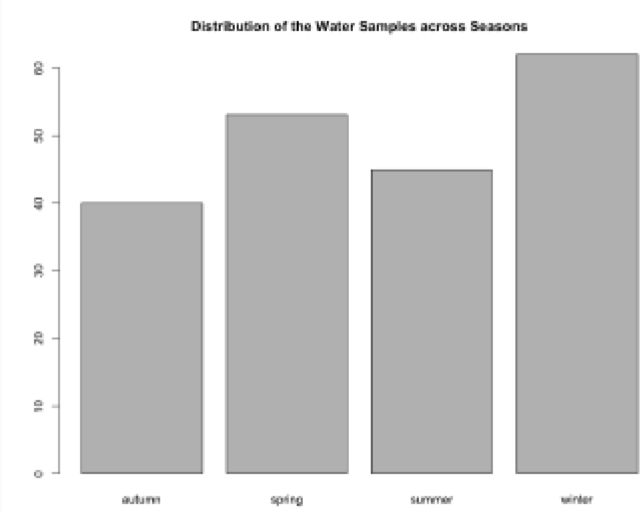
- Univariate Graphs
- Bivariate Graphs
- Multivariate/Conditioned Graphs

Univariate Graphs

- **Categorical Variables**: Barplots, Piecharts, ...
- **Numeric Variables**: Line plots, Histograms, QQ Plots, Boxplots, ...

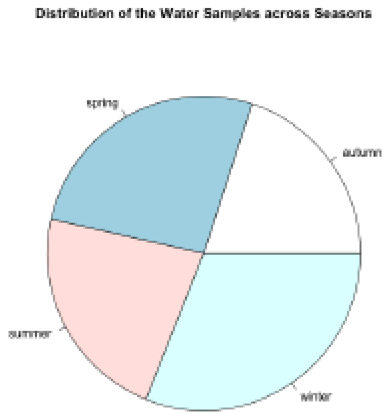
Barplots

- display a set of values as heights of bars
- display frequency of occurrence of different values



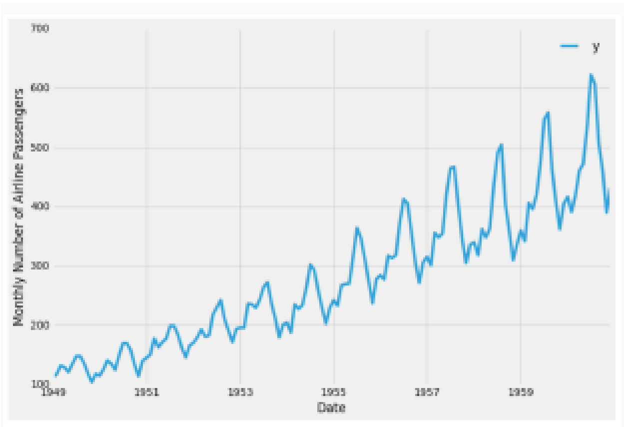
Piecharts

- same purpose as barplots with information in form of a pie
- not so good for comparisons



Line Plots

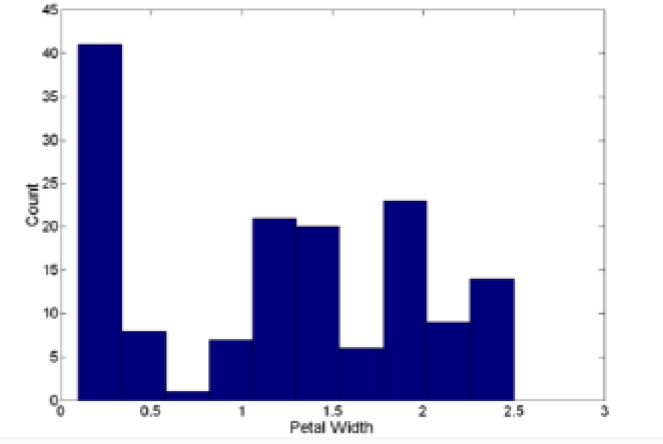
- analyze the evolution of the values of a continuous variable
- x-axis represent a quantitative scale with equal lag between observations
- used to deal with notion of time



Histograms

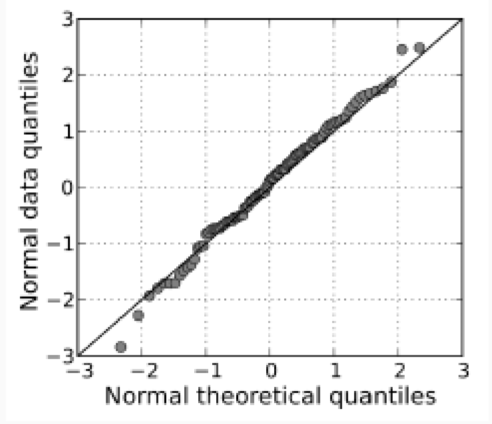
- display how the values of a continuous variable are distributed
- may be misleading in small data sets

- shape depends on the number of bins



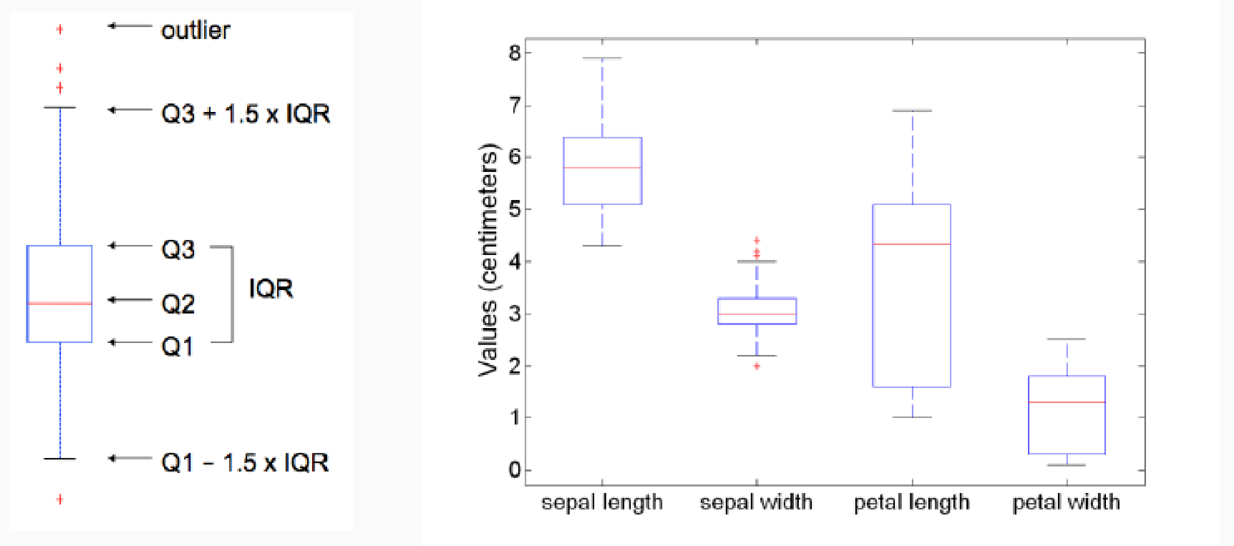
QQ Plots

- how properties such as location, scale and skewness compare in 2 distributions
- visually check the hypothesis that the variable under study follow a normal distribution, comparing the observed distribution against the Normal distribution



Boxplots

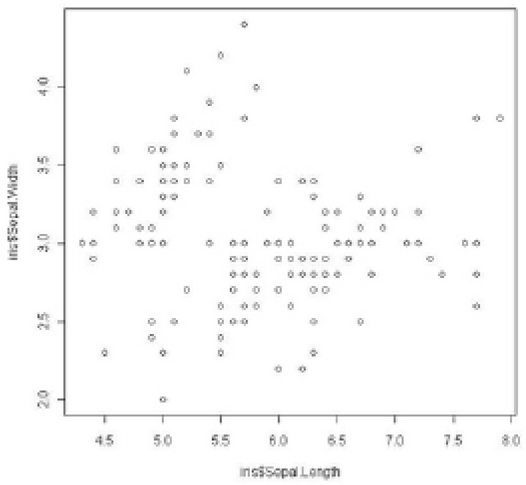
- provide an interesting summary of a variable distribution
- inform us of the interquartile range and of the outliers (if any)



Bivariate Graphs

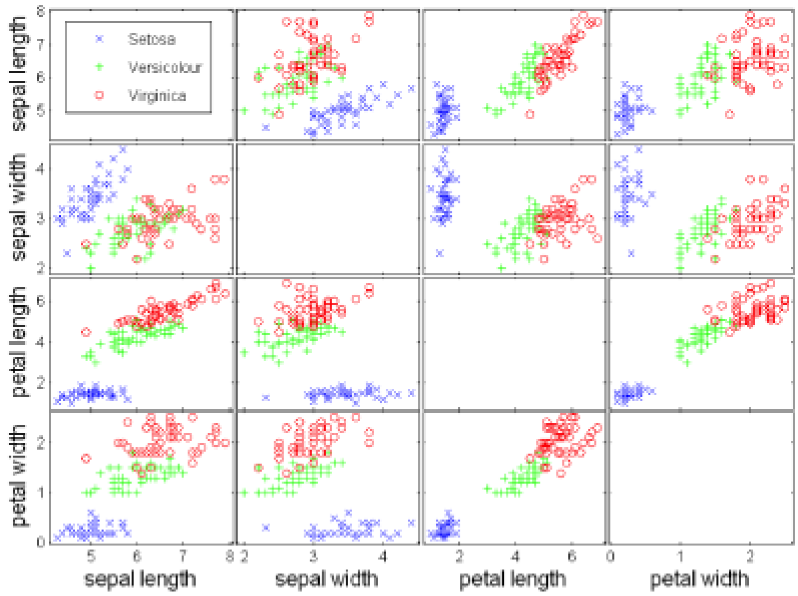
Scatterplots

- show the relationship between 2 numeric variables



Multivariate Graphs

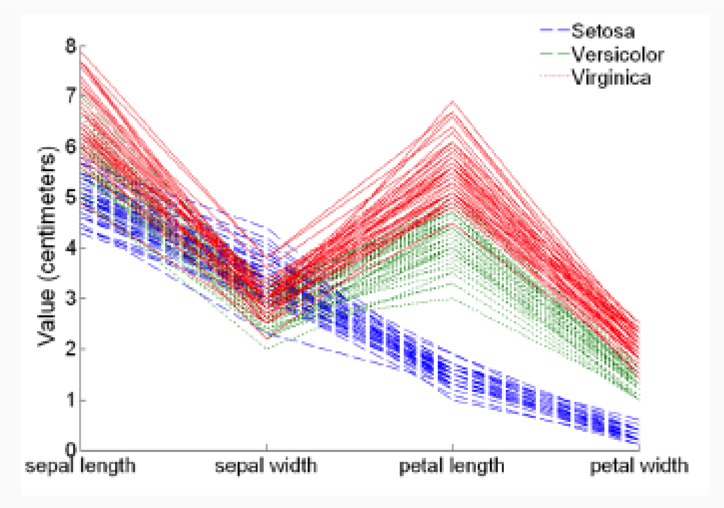
- plot the relationship between every pair of numeric variables and respective groups



Parallel Coordinates Plot

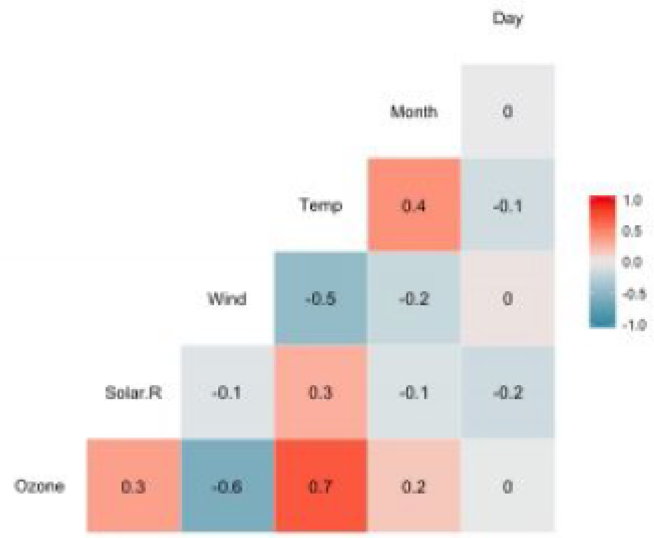
- attributes values for each case (line)

- order might be important to help identify groups



Correlogram

- correlation statistics (e.g. pearson) for each pair of variables



Conditioned Graphs

- allow the simultaneous presentation of subgroup graphs to better allow finding eventual differences between the subgroups

Data Preparation

Set of steps that may be necessary to carry out before any further analysis takes place on the available data

- may face the need to "create" new variables to achieve objectives
- set may be too large
- **Feature Extraction**: extract features from raw data on which analysis can be performed

- **Data Cleaning**: data may be hard to read or require extra parsing efforts
- **Data Transformation**: it may be necessary to change some values of the data
- **Feature Engineering**: to incorporate some domain knowledge
- **Data and Dimensionality Reduction**: to make modeling possible

Feature Extraction

- Very application specific and a very crucial step
  - **Sensor data**: large volume of low-level signals associated with date/time attributes
  - **Image data**: very high-dimensional data that can be represented by pixels, color histograms, etc.
  - **Web logs**: text in a prespecified format with both categorical and numerical attributes
  - **Network traffic**: network packets information
  - **Document data**: raw and unstructured data

Data Cleaning

Handling Missing Values

- **Goal**: make data tidy

Strategies

- Remove all cases in a data set with some unknown value
- Fill-in:
  - the unknowns with the imputation of the most common value
  - with the most common value on the cases that are more "similar" to the one with unknowns
  - with linear interpolation of nearby values in time and/or space
- Explore eventual correlations between variables
- Do nothing

Handling Incorrect Values

- **Inconsistency detection**: data integration techniques within the database field
- **Domain knowledge**: data auditing that use domain knowledge and constraints
- **Data-centric methods**: statistical-based methods to detect outliers

Data Transformation

- Map the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values

Common Strategies

- Normalization
- Binarization/One-Hot Encoding
- Discretization

Normalization

- **Min-Max Scaling (Range-base Normalization)** - not robust for scenarios where there are outliers
$$y_i = \frac{x_i - \min_x}{\max_x - \min_x}$$
- **Standardization (z-score Normalization)**
$$y_i = \frac{x_i - \mu_x}{\sigma_x}$$

Case Dependencies

- In time series, it is common to use different techniques.
- E.g.: adjust mean, variance range; remove unwanted, common signal

Binarization/One-Hot Encoding

- **Binarization**: if the attribute has only 2 possible nominal values, it can be transformed into 1 binary attribute
  - fever: yes/no → fever: 1/0
- **One-Hot Encoding**: if the attribute has k possible nominal values, it can be transformed into k binary attributes
  - eye_color: brown/blue/green → eye_brown: 1/0, eye_blue: 1/0, eye_green 1/0

Discretization

- process of converting continuous attribute into an ordinal attribute of numeric variables
- **Unsupervised discretization**: find breaks in the data values
  - **Equal-width**: divides the original values into equal-width range of values; may be affected by outliers
  - **Equal-frequency**: divides the original values so that the same number of values are assigned to each range; can generate ranges with different amplitudes
- **Supervised discretization**: use class labels to find breaks

Feature Engineering

- Fundamental to the application of machine learning
- The process of using domain knowledge of the data to create features that might help when solving the problem
- New features that can capture the important information in a ta set much more efficiently than the original features

Case Dependencies

- **Case 1**: express known relationships between existing variables
    - create ratios and proportions
- **Case 2**: overcome limitations of some data mining tools regarding cases dependencies
    - create variables that express dependency relationships
- In time series is common to create features that represent relative values instead of absolute values, so to avoid trend effects

  $$y_t = \frac{x_t - x_{t-1}}{x_{t-1}}$$

[< Go back]