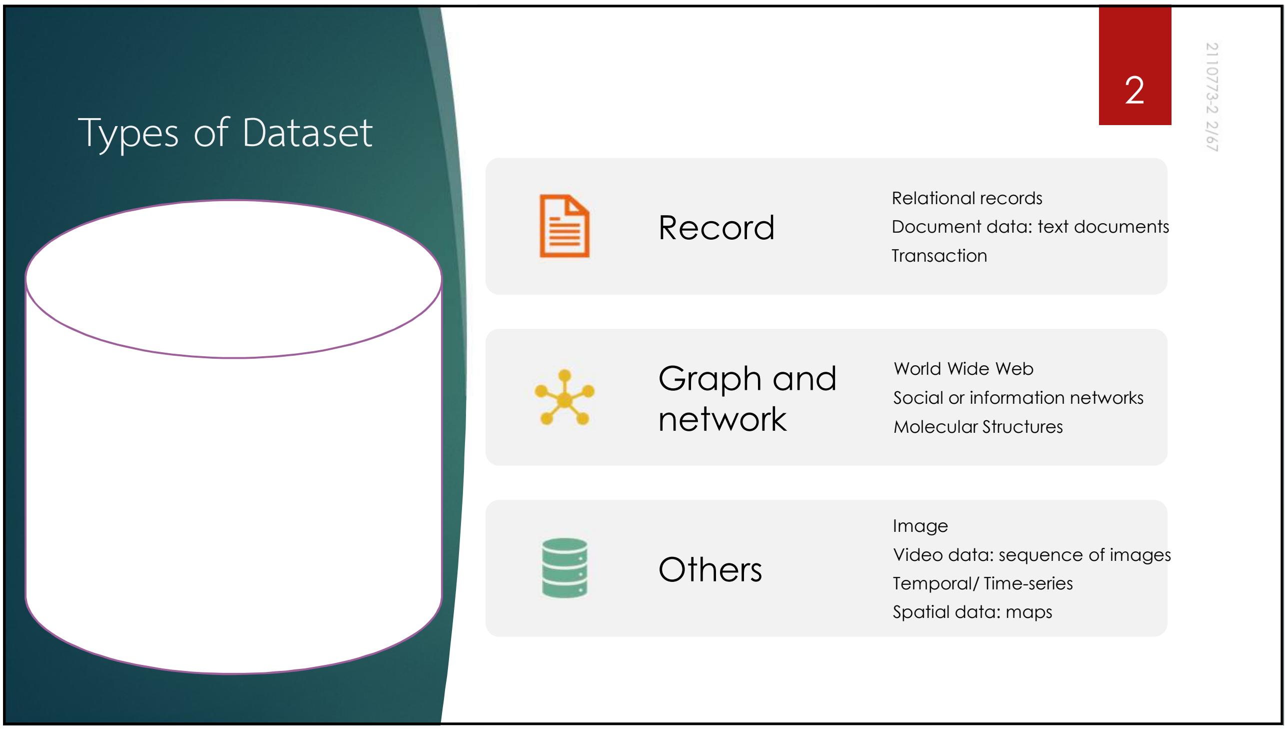
2110773 Data Mining Chapter2: Data Preprocessing

- GARBAGE IN → GARBAGE OUT
- IMPORTANT & TIME-CONSUMING TASK IN KDD
- PRACTICE IS EVERYTHING

- รศ. ดร. ญาใจ ลิ้มปิยะกรณ์



2

2110773-2 2/67

Types of Dataset

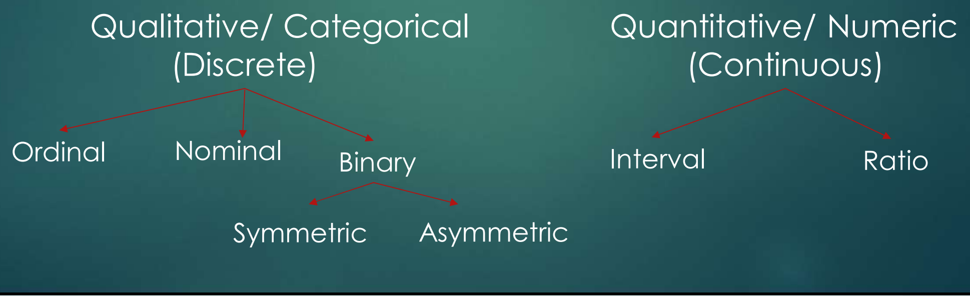
Record	Relational records Document data: text documents Transaction
Graph and network	World Wide Web Social or information networks Molecular Structures
Others	Image Video data: sequence of images Temporal/ Time-series Spatial data: maps

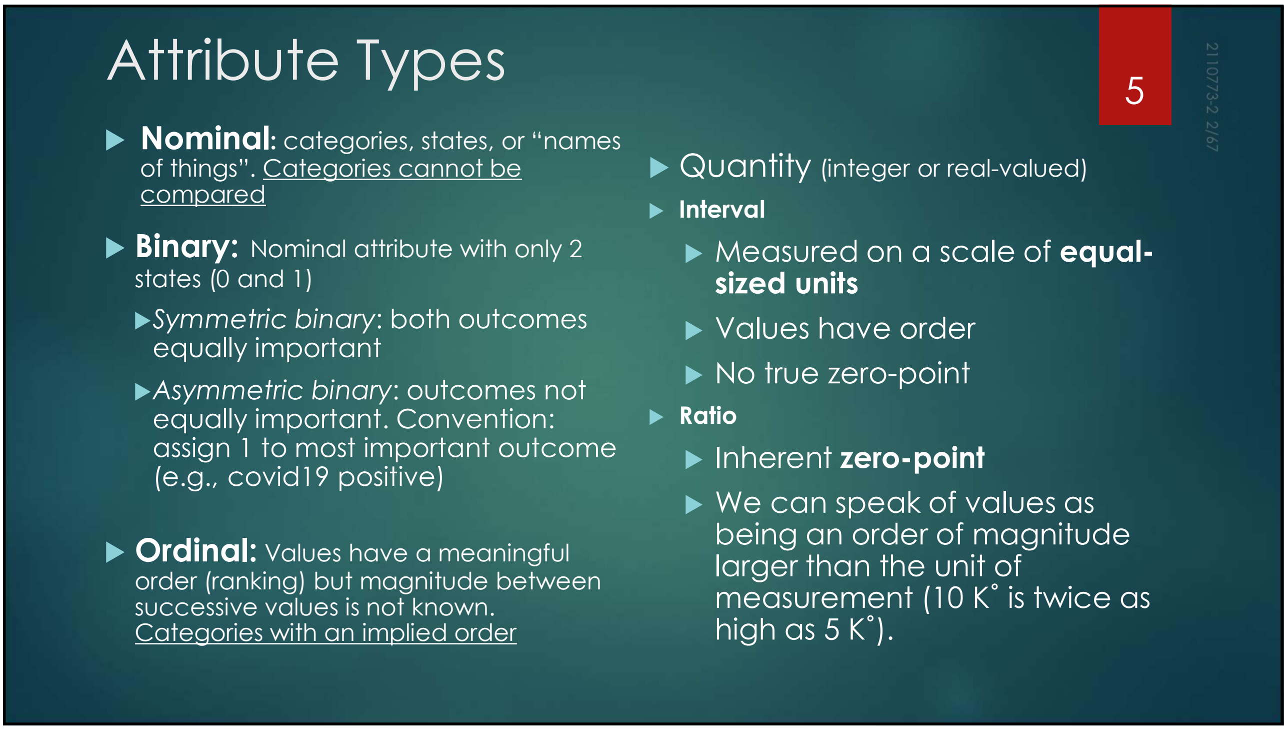
Data Object

- Data sets are made up of data objects.
- A **data object** represents an entity. For examples:
  - medical database: patients, treatments
  - university database: students, professors, courses
- Also called *samples* , *examples, instances, data points, objects, tuples*.
- Data objects are described by **attributes**.
- Database rows -> data objects; columns ->attributes.
- Attribute (or **dimension, feature, variable**): a data field, representing a characteristic or feature of a data object, e.g., *customer _ID, name, address, phone*

Attribute Data Types

1. Qualitative/ Quantitative
2. Categorical/ Numeric
3. Discrete/ Continuous
   - Discrete: Has only a finite or countably infinite set of values. Sometimes, represented as integer variables
   - Continuous: Has real numbers (floating-point) as attribute values. Practically, real values can only be measured.


Qualitative/ Categorical (Discrete)
Ordinal
Nominal
Binary
Symmetric
Asymmetric
Quantitative/ Numeric (Continuous)
Interval
Ratio



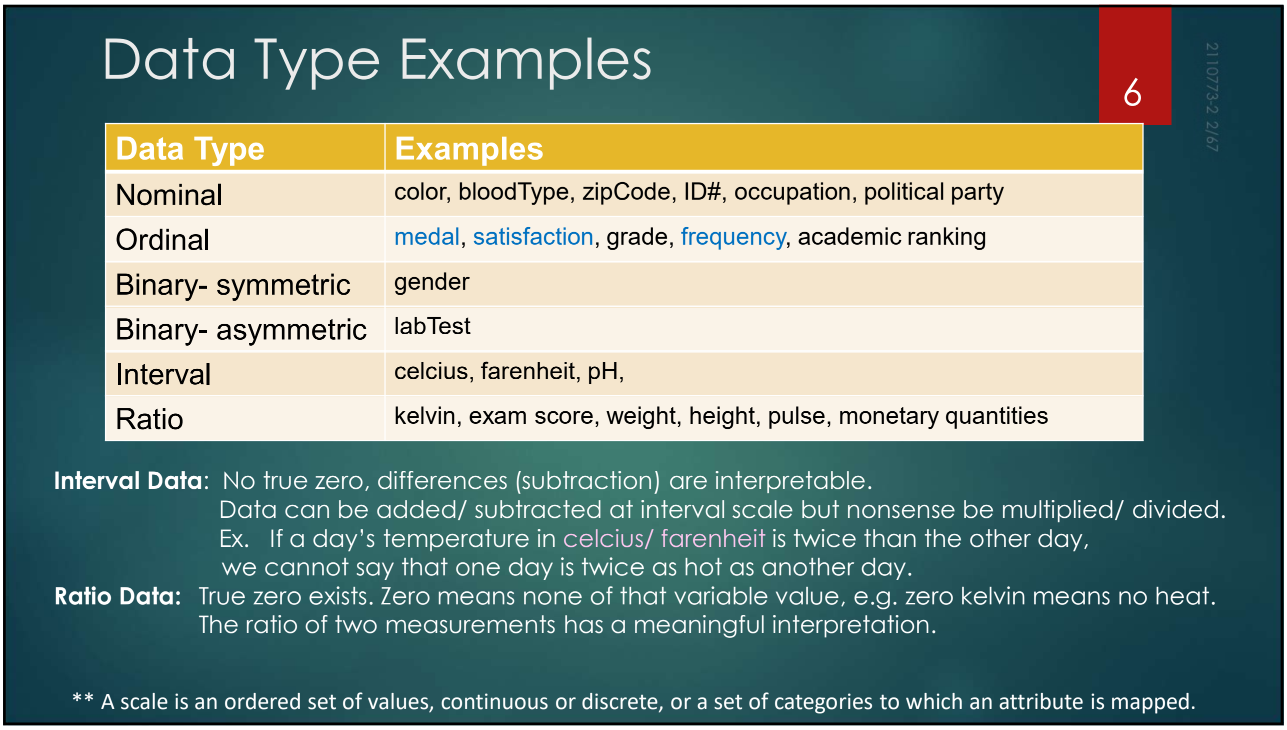
Attribute Types

5

2110773-2 2/67

- **Nominal:** categories, states, or "names of things". Categories cannot be compared
- **Binary:** Nominal attribute with only 2 states (0 and 1)
  - *Symmetric binary*: both outcomes equally important
  - *Asymmetric binary*: outcomes not equally important. Convention: assign 1 to most important outcome (e.g., covid19 positive)
- **Ordinal:** Values have a meaningful order (ranking) but magnitude between successive values is not known. Categories with an implied order

- Quantity (integer or real-valued)
- **Interval**
  - Measured on a scale of **equal-sized units**
  - Values have order
  - No true zero-point
- **Ratio**
  - Inherent **zero-point**
  - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K˚ is twice as high as 5 K˚).



Data Type Examples

6

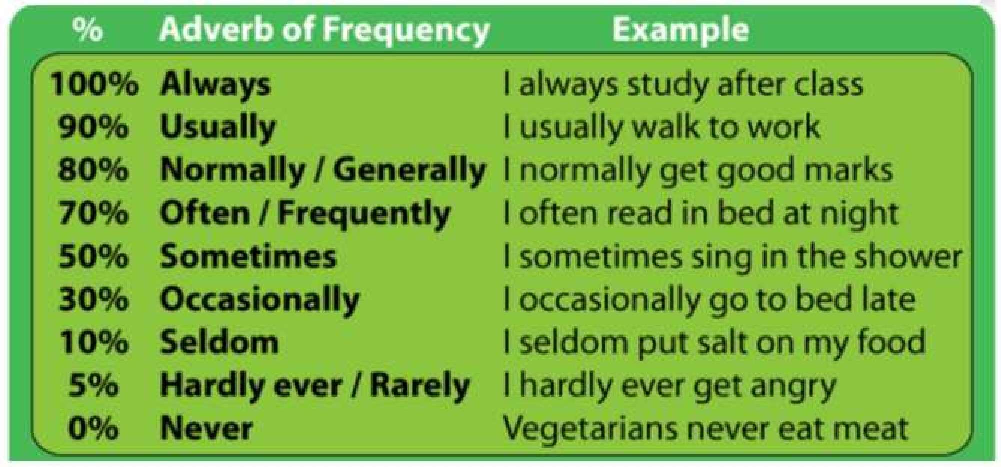
2110773-2 2/67

Data Type	Examples
Nominal	color, bloodType, zipCode, ID#, occupation, political party
Ordinal	medal, satisfaction, grade, frequency, academic ranking
Binary- symmetric	gender
Binary- asymmetric	labTest
Interval	celcius, farenheit, pH,
Ratio	kelvin, exam score, weight, height, pulse, monetary quantities

Interval Data: No true zero, differences (subtraction) are interpretable.
Data can be added/ subtracted at interval scale but nonsense be multiplied/ divided.
Ex. If a day's temperature in celcius/ farenheit is twice than the other day,
we cannot say that one day is twice as hot as another day.

Ratio Data: True zero exists. Zero means none of that variable value, e.g. zero kelvin means no heat.
The ratio of two measurements has a meaningful interpretation.

** A scale is an ordered set of values, continuous or discrete, or a set of categories to which an attribute is mapped.



%	Adverb of Frequency	Example
100%	Always	I always study after class
90%	Usually	I usually walk to work
80%	Normally / Generally	I normally get good marks
70%	Often / Frequently	I often read in bed at night
50%	Sometimes	I sometimes sing in the shower
30%	Occasionally	I occasionally go to bed late
10%	Seldom	I seldom put salt on my food
5%	Hardly ever / Rarely	I hardly ever get angry
0%	Never	Vegetarians never eat meat

Scales of Measurement

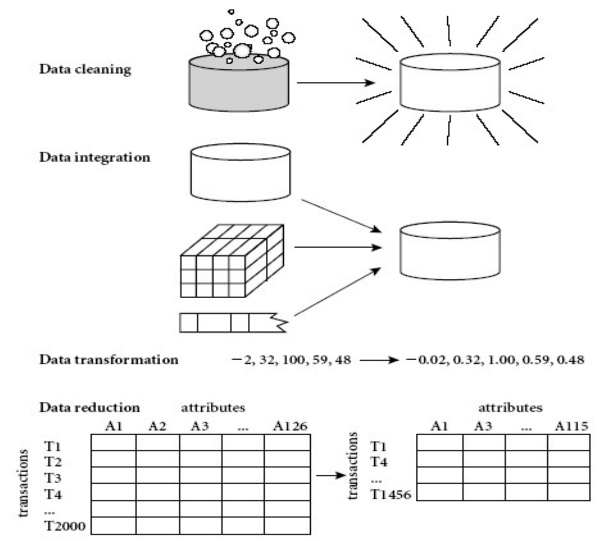
Data	Nominal	Ordinal	Interval	Ratio
Labeled				
Order				
Measurable Difference				
True Zero Starting Point				

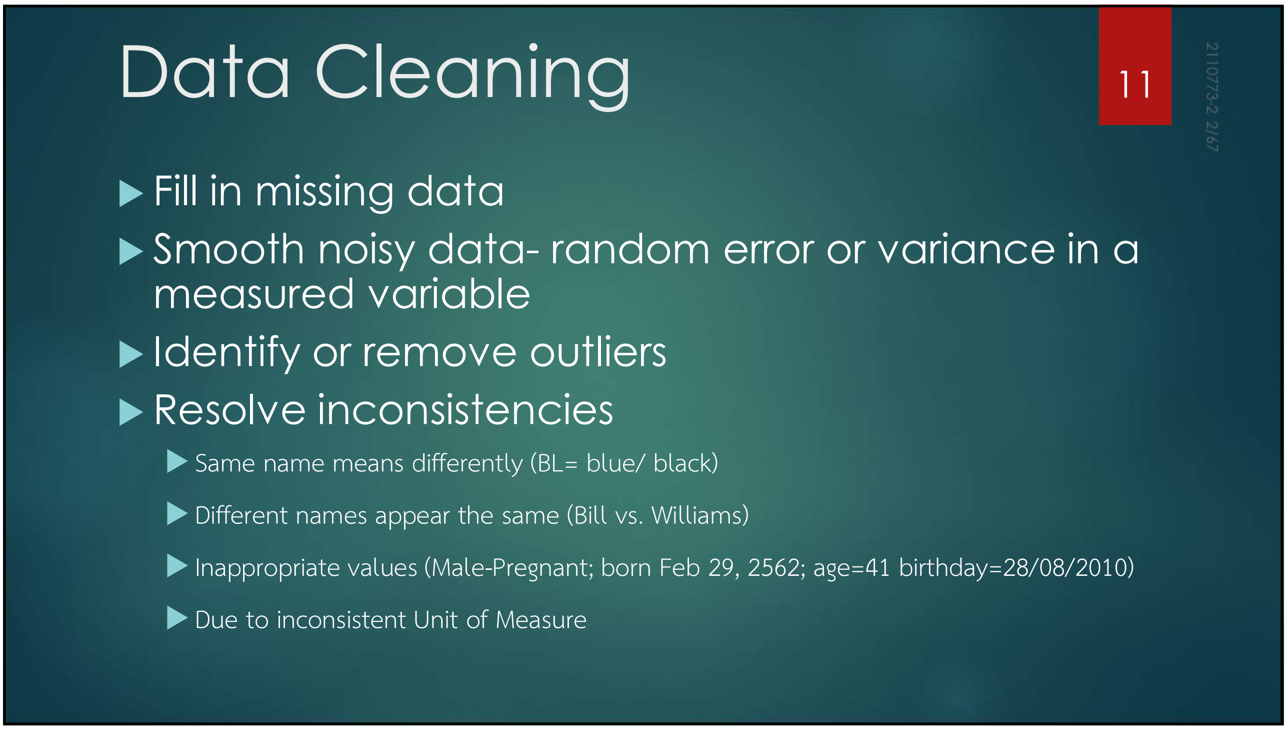
Survey

1. How old are you? ___________years
2. Are you: Male Female
3. How much do you spend on groceries each week? _______Baht
4. How many cups of coffee do you buy in a week? ________
5. Which type of coffee do you like most?
   Latte Espresso Cappuccino Americano
6. How likely are you to buy more than a cup of coffee per day?
   Very Likely Likely Not Likely Very Unlikely

Data Preprocessing

- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction



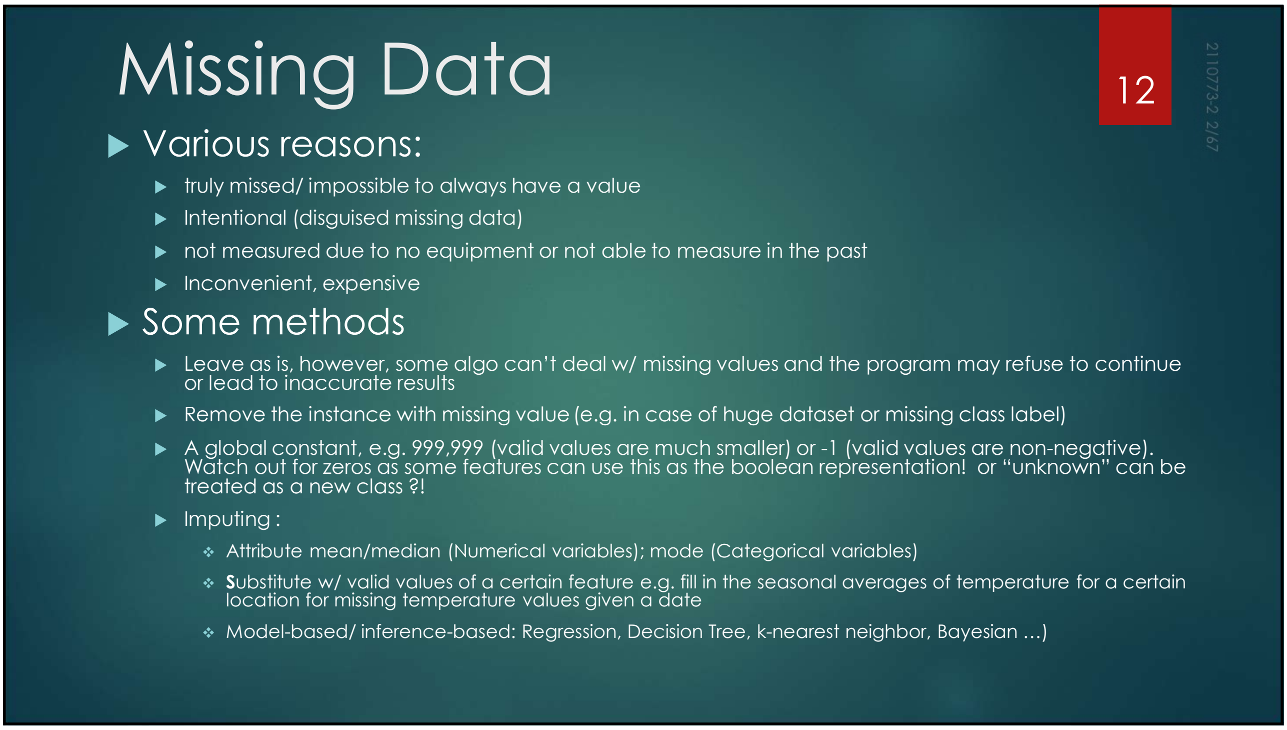


Data Cleaning

11

2110773-2 2/67

- Fill in missing data
- Smooth noisy data- random error or variance in a measured variable
- Identify or remove outliers
- Resolve inconsistencies
  - Same name means differently (BL= blue/ black)
  - Different names appear the same (Bill vs. Williams)
  - Inappropriate values (Male-Pregnant; born Feb 29, 2562; age=41 birthday=28/08/2010)
  - Due to inconsistent Unit of Measure



Missing Data

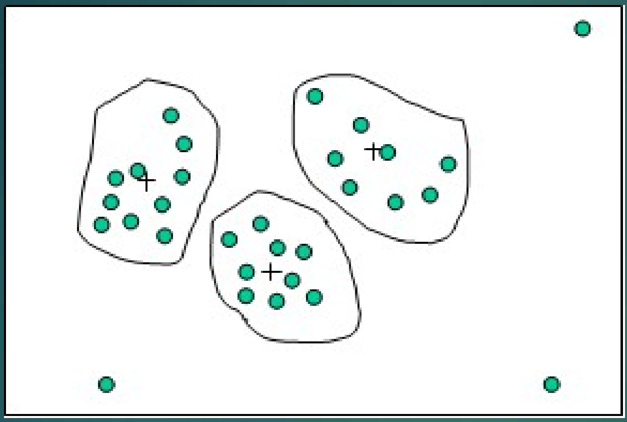
12

2110773-2 2/67

- Various reasons:
  - truly missed/ impossible to always have a value
  - Intentional (disguised missing data)
  - not measured due to no equipment or not able to measure in the past
  - Inconvenient, expensive
- Some methods
  - Leave as is, however, some algo can't deal w/ missing values and the program may refuse to continue or lead to inaccurate results
  - Remove the instance with missing value (e.g. in case of huge dataset or missing class label)
  - A global constant, e.g. 999,999 (valid values are much smaller) or -1 (valid values are non-negative). Watch out for zeros as some features can use this as the boolean representation! or "unknown" can be treated as a new class ?!
  - Imputing :
    - Attribute mean/median (Numerical variables); mode (Categorical variables)
    - **S**ubstitute w/ valid values of a certain feature e.g. fill in the seasonal averages of temperature for a certain location for missing temperature values given a date
    - Model-based/ inference-based: Regression, Decision Tree, k-nearest neighbor, Bayesian …)

Noisy Data

- Random error or variance in a measured variable
  - Regression- smooth by fitting the data into regression functions
- Outliers are noisy data or data points inconsistent with the majority of data, e.g. one's age = 200 year, height=3 metre, widely deviated points
  - Detect and remove outliers- Clustering
  - Truncate outliers- Bell curve, Box plots

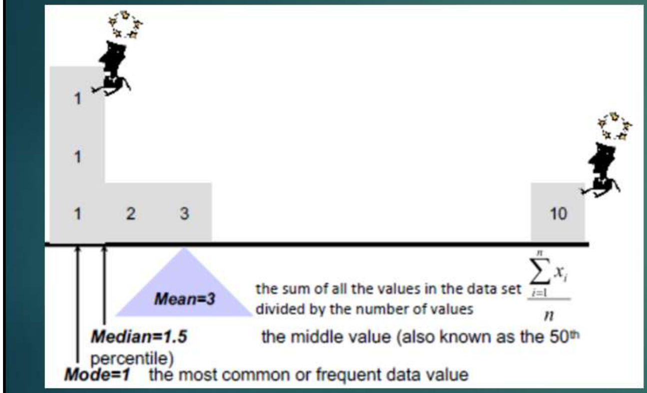


Clustering

Data Distribution

1. Central Tendency/ Center

2. Spread/ Dispersion

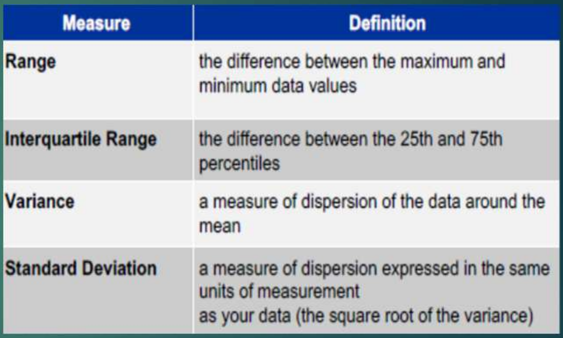


the sum of all the values in the data set divided by the number of values $\frac{\sum_{i=1}^{n} x_i}{n}$

Median=1.5 the middle value (also known as the 50th percentile)

Mode=1 the most common or frequent data value

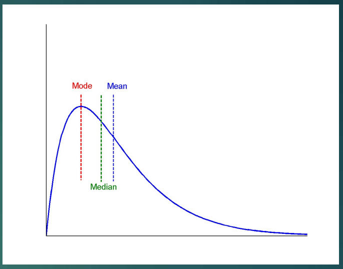
Measure of Central Tendency (Representative value): Mean, Median, Mode



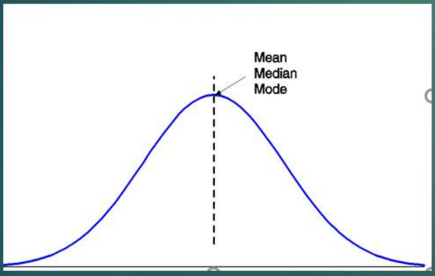
Measure	Definition
Range	the difference between the maximum and minimum data values
Interquartile Range	the difference between the 25th and 75th percentiles
Variance	a measure of dispersion of the data around the mean
Standard Deviation	a measure of dispersion expressed in the same units of measurement as your data (the square root of the variance)

Symmetric vs. Skewed Data

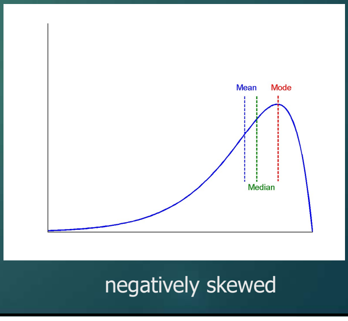
- Median, mean and mode of symmetric, positively and negatively skewed data



positively skewed



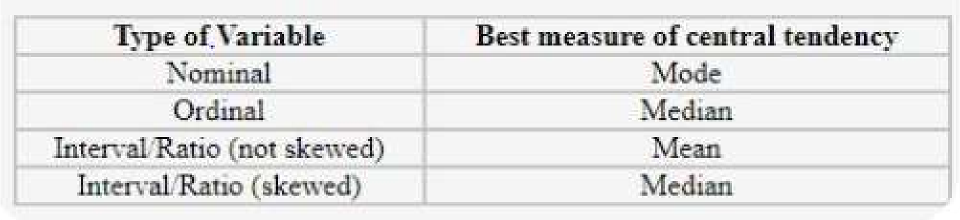
symmetric



negatively skewed

17

2110773-2 2/67



Type of Variable	Best measure of central tendency
Nominal	Mode
Ordinal	Median
Interval/Ratio (not skewed)	Mean
Interval/Ratio (skewed)	Median

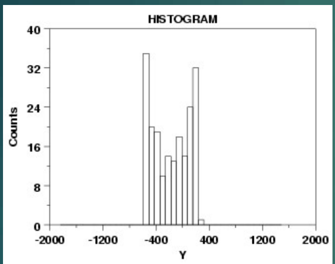
When to use Mean, Median, Mode

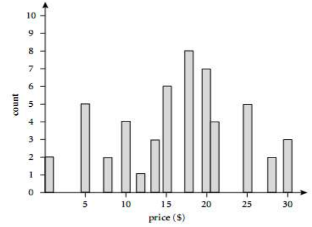
Graphical Displays of Distribution

18

2110773-2 2/67

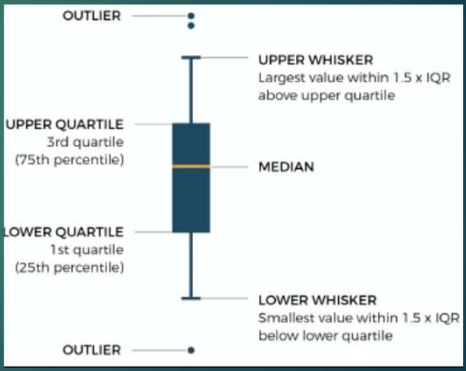
- Histogram Graph display of frequencies, shown as bars with numeric values on X axis

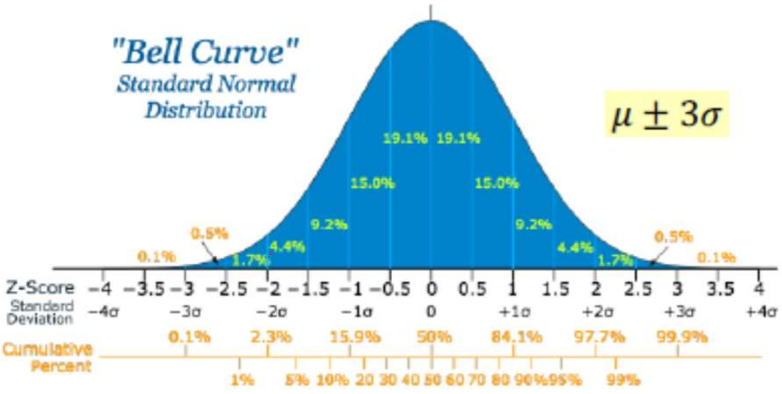




Singleton Histogram

- Box plots





$\mu \pm 3\sigma$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i-\mu)^2 = \frac{1}{N}\sum_{i=1}^{n}x_i^2 - \mu^2$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n}x_i^2 - \frac{1}{n}(\sum_{i=1}^{n}x_i)^2]$$

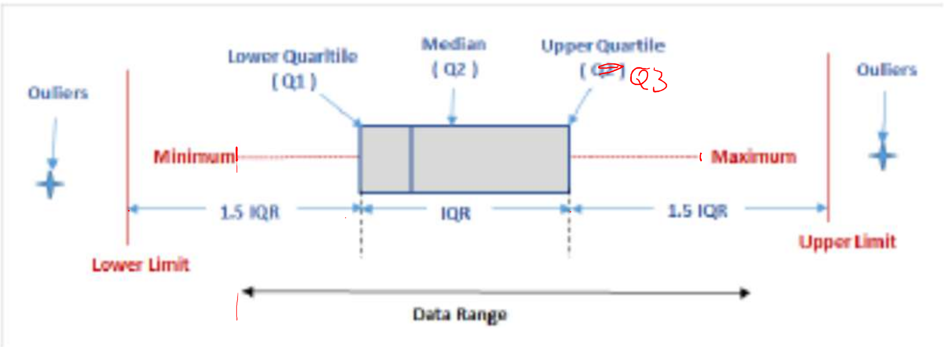
19

2110773-2 2/67

Truncate Outliers: Bell Curve

Variance and standard deviation (sample: s, population: σ)

Standard deviation is the square root of variance



20

2110773-2 2/67

Truncate Outliers: Box Plots

Interquartile Range

- **IQR is a measure of spread indicating where the bulk of the values lie.**
  - **Quartiles**: Q_1 (25th percentile), Q_3 (75th percentile)
  - **Inter-quartile range**: IQR = $Q_3 – Q_1$
  - **Five number summary**: min, Q_1, median, Q_3, max
  - **Boxplot**: ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually
  - **Outlier**: usually, a value higher/lower than 1.5 x IQR

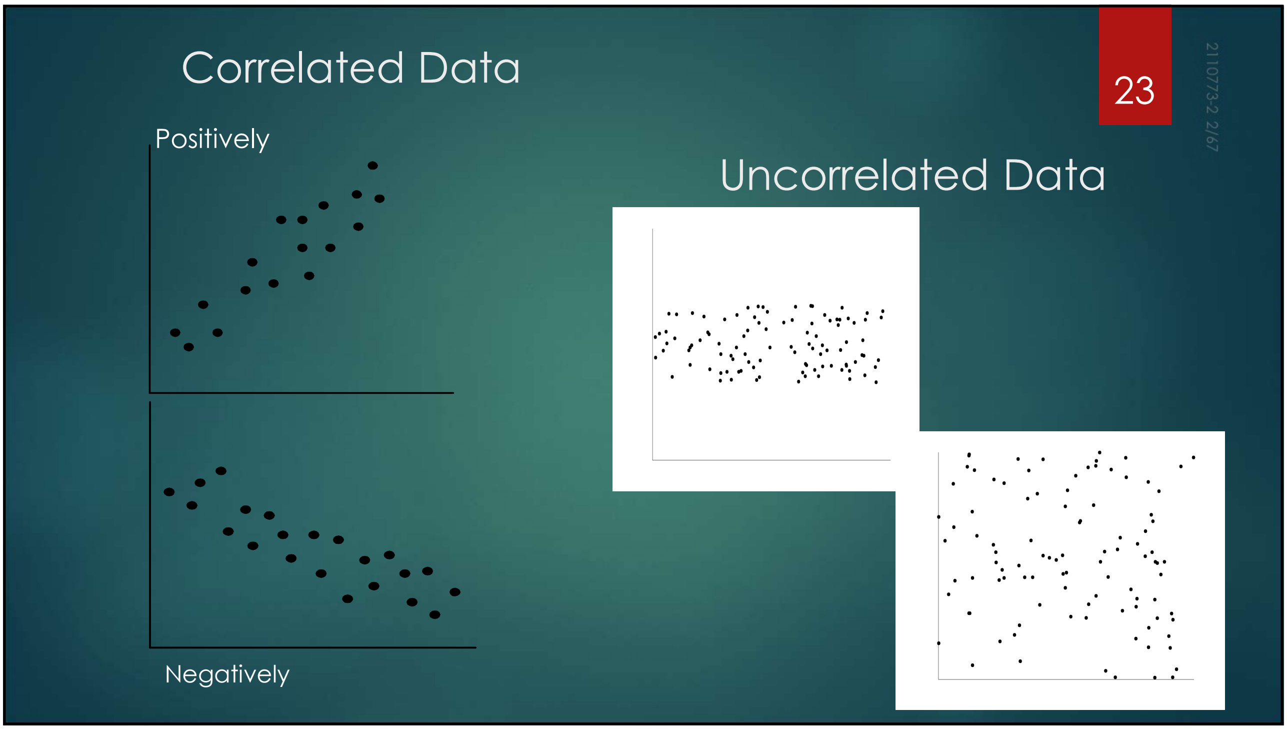
IQR Calculation

Odd set of numbers

- Step 1: **Put the numbers in order.**
  1, 2, 5, 6, 7, 9, 12, 15, 18, 19, 27.
- Step 2: **Find the median.**
  1, 2, 5, 6, 7, **9**, 12, 15, 18, 19, 27.
- Step 3: **Place parentheses around the numbers above and below the median.** Not necessary **statistically**, but it makes Q1 and Q3 easier to spot.
  (1, 2, 5, 6, 7), 9, (12, 15, 18, 19, 27).
- Step 4: **Find Q1 and Q3**
  Think of Q1 as a median in the lower half of the data and think of Q3 as a median for the upper half of data.
  (1, 2, **5**, 6, 7), **9**, (12, 15, **18**, 19, 27). Q1 = 5 and Q3 = 18.
- Step 5: **Subtract Q1 from Q3 to find the interquartile range**.
  18 – 5 = 13.

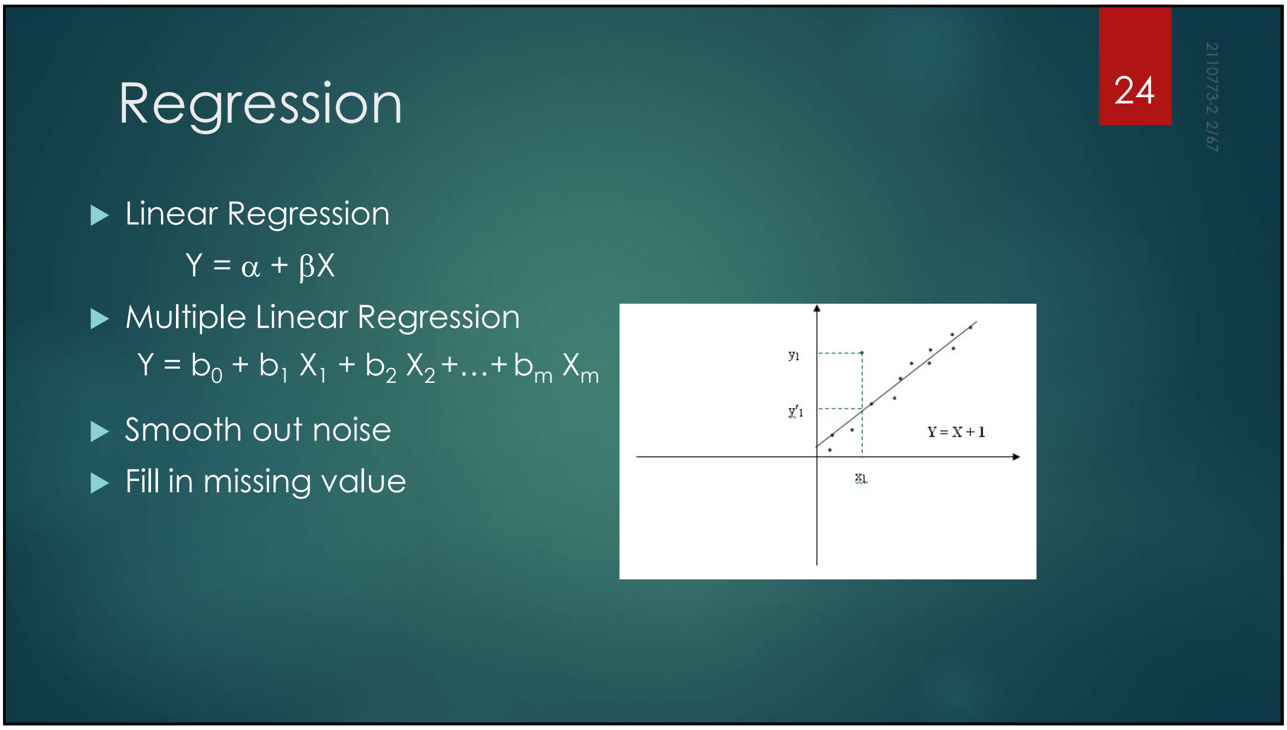
Even set of numbers

- Step 1: **Put the numbers in order**.
  3, 5, 7, 8, 9, 11, 15, 16, 20, 21.
- Step 2: **Make a mark in the center of the data**:
  3, 5, 7, 8, 9, | 11, 15, 16, 20, 21.
- Step 3: **Place parentheses around the numbers above and below the mark you made in Step 2–it makes Q1 and Q3 easier to spot.**
  (3, 5, 7, 8, 9), | (11, 15, 16, 20, 21).
- Step 4: **Find Q1 and Q3**
  Q1 is the median (the middle) of the lower half of the data, and Q3 is the median (the middle) of the upper half of the data.
  (3, 5, **7**, 8, 9), | (11, 15, **16,** 20, 21). Q1 = 7 and Q3 = 16.
- Step 5: **Subtract Q1 from Q3**.
  16 – 7 = 9.



Correlated Data

Uncorrelated Data



Regression

- Linear Regression

  $Y = \alpha + \beta X$

- Multiple Linear Regression

  $Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_m X_m$

- Smooth out noise
- Fill in missing value