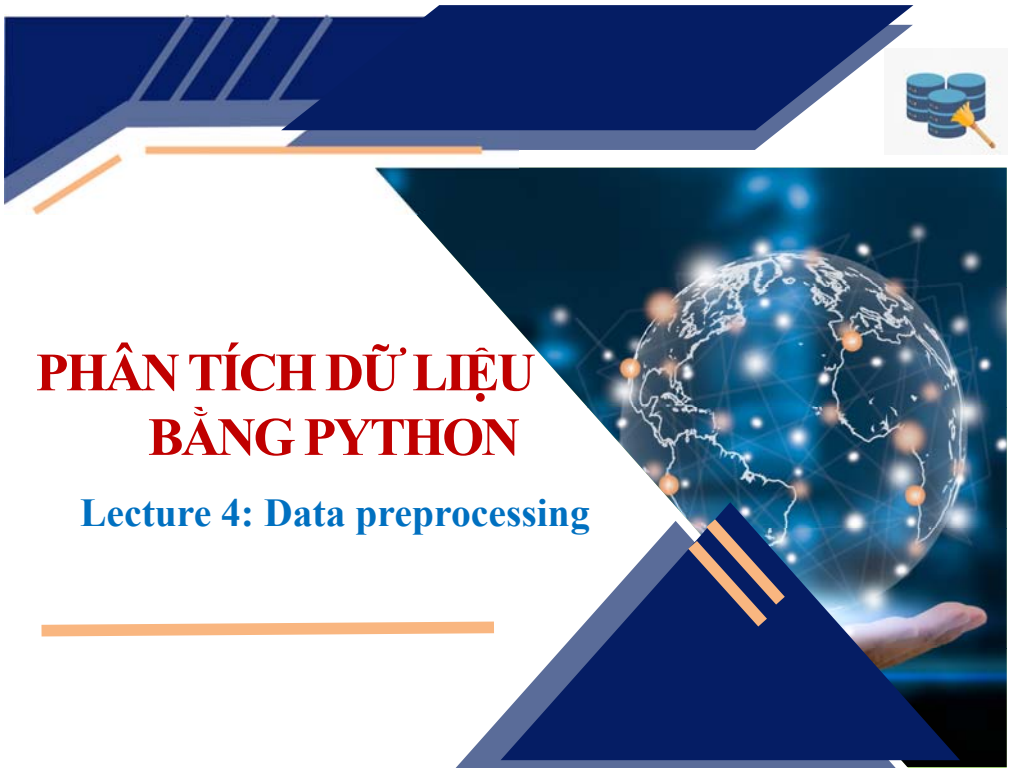


PHÂN TÍCH DỮ LIỆU BẰNG PYTHON

Lecture 4: Data preprocessing

Nội dung

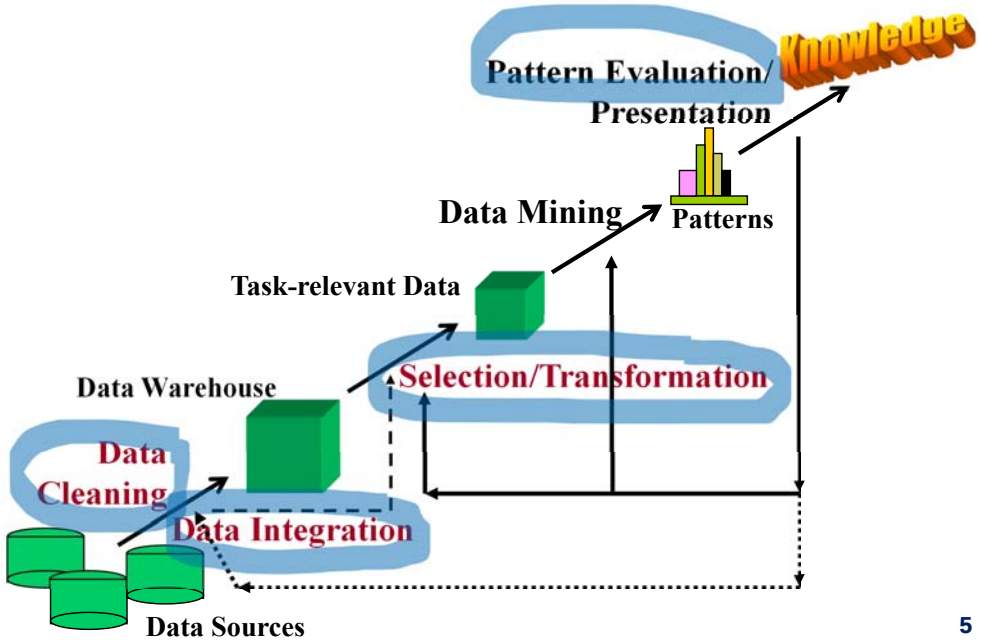
4.1. Tổng quan về giai đoạn tiền xử lý dữ liệu

- Giai đoạn tiền xử lý dữ liệu
  - Quá trình **xử lý dữ liệu thô/gốc** (raw/original data) nhằm cải thiện chất lượng dữ liệu (quality of the data) và do đó, cải thiện chất lượng của kết quả phân tích và khai phá dữ liệu.
    - Dữ liệu thô/gốc
      - Có cấu trúc, bán cấu trúc, phi cấu trúc
      - Các hệ thống cơ sở dữ liệu (database systems)
    - Chất lượng dữ liệu (data quality): tính chính xác, tính hiện hành, tính toàn vẹn, tính nhất quán

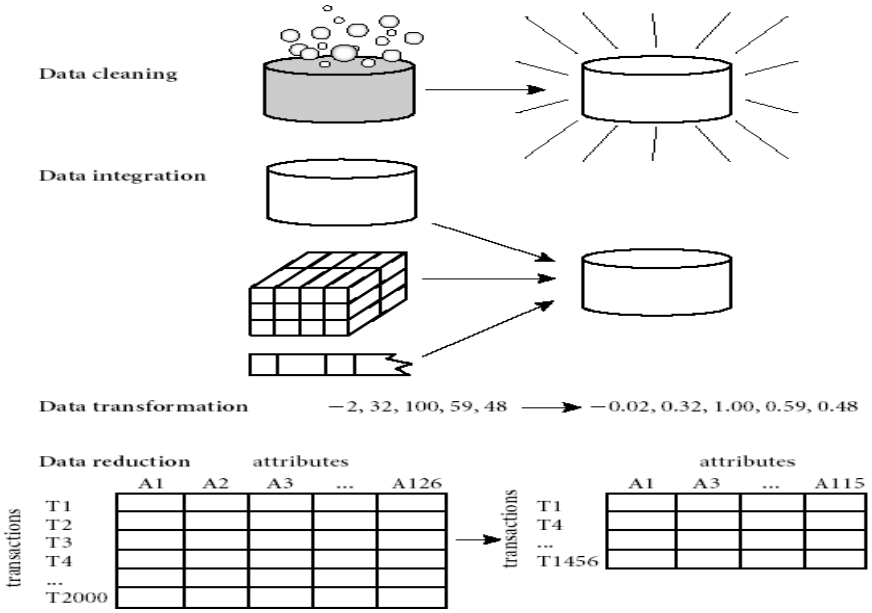
4.1. Tổng quan về giai đoạn tiền xử lý dữ liệu

- Chất lượng dữ liệu (data quality)
  - Tính chính xác (accuracy): giá trị được ghi nhận đúng với giá trị thực.
  - Tính hiện hành (currency/timeliness): giá trị được ghi nhận không bị lỗi thời.
  - Tính toàn vẹn (completeness): tất cả các giá trị dành cho một biến/thuộc tính đều được ghi nhận.
  - Tính nhất quán (consistency): tất cả giá trị dữ liệu đều được biểu diễn như nhau trong tất cả các trường hợp.

4.1. Tổng quan về giai đoạn tiền xử lý dữ liệu



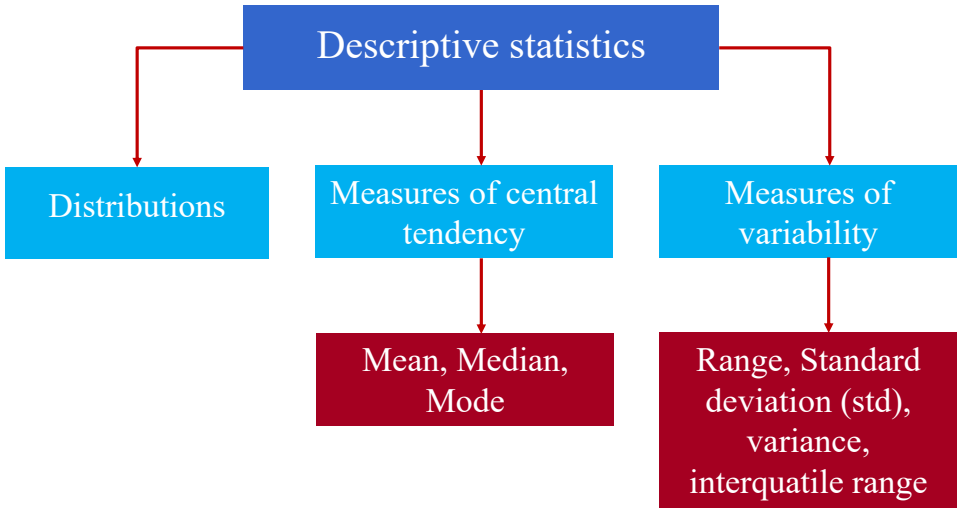
4.1. Tổng quan về giai đoạn tiền xử lý dữ liệu



4.1. Tổng quan về giai đoạn tiền xử lý dữ liệu

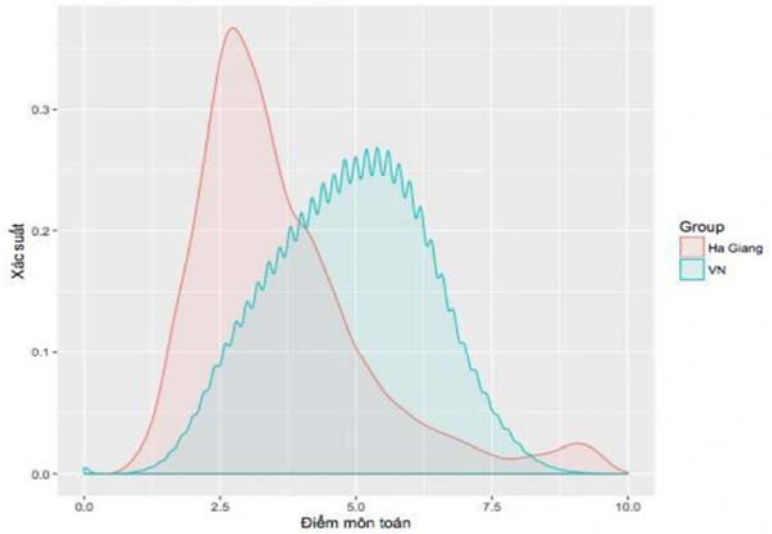
- Các kỹ thuật tiền xử lý dữ liệu
  - Làm sạch dữ liệu (data cleaning/cleansing): loại bỏ nhiễu (remove noise), hiệu chỉnh những phần dữ liệu không nhất quán (correct data inconsistencies)
  - Tích hợp dữ liệu (data integration): trộn dữ liệu (merge data) từ nhiều nguồn khác nhau vào một kho dữ liệu
  - Biến đổi dữ liệu (**data transformation**): chuẩn hoá dữ liệu (**data normalization**)
  - Thu giảm dữ liệu (**data reduction**): thu giảm kích thước dữ liệu (nghĩa là giảm số phần tử) bằng kết hợp dữ liệu (data aggregation), loại bỏ các đặc điểm dư thừa (redundant features) (nghĩa là giảm số chiều/thuộc tính dữ liệu), gom cụm dữ liệu

4.2. Thống kê mô tả về dữ liệu



❖ **Measures of variability**

- Observe two hypothetical data sets



❖ **Central measures**

- *Arithmetic mean*: this is the most popular and useful measure of central location

$$\text{Mean} = \frac{\text{Sum of measurements}}{\text{Number of measurements}}$$

Sample mean

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Sample size

Population mean

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

Population size

❖ **Central measures**

Example for sample mean

The mean of the *sample* of six measurements 7, 3, 9, –2, 4, 6 is given by

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{7 + 3 + 9 + (-2) + 4 + 6}{6} = 4.5$$

Example for polulation mean

Suppose the telephone bills represent a *population* of measurements. The population mean is

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{42.19 + 15.30 + \ldots + 53.21}{200} = 43.59$$

❖ **Central measures**

- *Median*: The *median* of a data set is the value that falls in the middle when the measurements are arranged in order of magnitude

Example for odd number

Seven employee salaries were recorded (in 1000s) : 28, 109, 26, 32, 30, 26, 29. Find the median salary.

Solution: First, sort the salaries, then locate the middle value.

26,26,28,29,30,32,109

Odd number of observations

Example for even number

Suppose one salary of $31 000 was added to the group recorded before. Find the median salary.

Solution: Sort the salaries, & locate the two middle values.

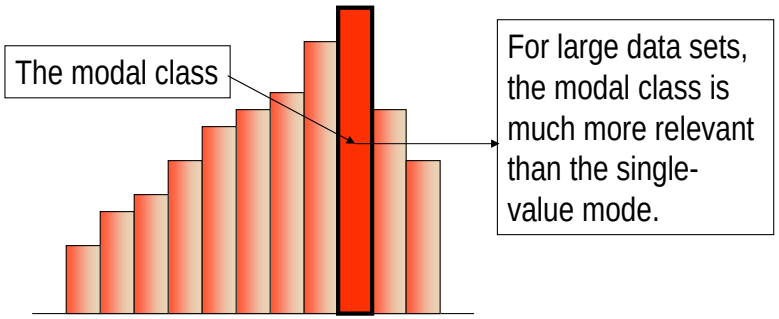
26,26,28,29, 29.5,30,31,32,109

Even number of observations

4.2. Thống kê mô tả về dữ liệu

❖ **Central measures**

- *Mode*: The *mode* of a set of measurements is the value that occurs most frequently. A set of data may have one mode (or modal class), or two or more modes.



4.2. Thống kê mô tả về dữ liệu

❖ **Central measures**

- Example:

  - The manager of a men's store observes the waist size (in centimetres) of trousers sold yesterday:

    77, 85, 90, 85, 82, 70, 85, 75, 85, 80, 77, 100, 85, 70.

  - The mode of this data set is 85 cm.

4.2. Thống kê mô tả về dữ liệu

❖ **Central measures**

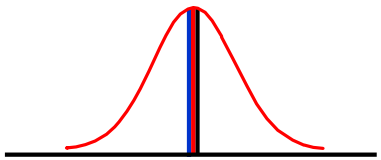
- Example: A tutor wants to report the results of a mid-semester exam taken by 100 students. The results of mean, median and mode are presented as follows.

Marks	
Mean	73.98
Standard Error	2.1502163
Median	81
Mode	84
Standard Deviation	21.502163
Sample Variance	462.34303
Kurtosis	0.3936606
Skewness	-1.073098
Range	89
Minimum	11
Maximum	100
Sum	7398
Count	100

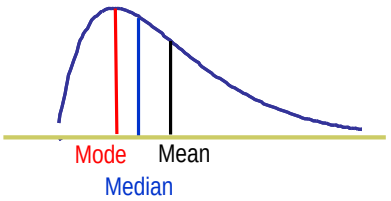
4.2. Thống kê mô tả về dữ liệu

❖ **Central measures: Mean, Median & Mode**

- If a distribution is symmetrical, the mean, median and mode coincide.



A positively skewed distribution ('skewed to the right')

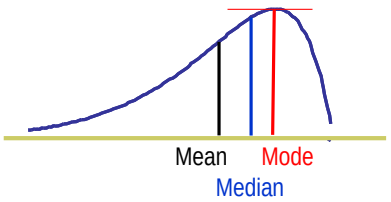


- If a distribution is not symmetrical, & skewed to the left or right, the 3 measures differ.

4.2. Thống kê mô tả về dữ liệu

❖ **Central measures**

A negatively skewed distribution ('skewed to the left')



- If a distribution is not symmetrical, & skewed to the left or right, the three measures differ.

4.2. Thống kê mô tả về dữ liệu

❖ **Measures of variability**

- Observe two hypothetical data sets

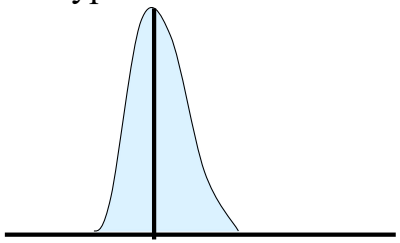


Figure 1

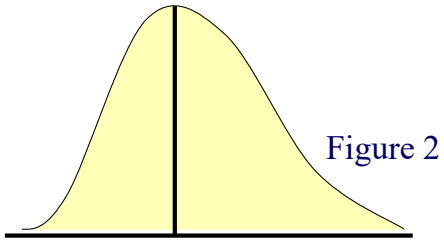


Figure 2

4.2. Thống kê mô tả về dữ liệu

❖ **Measures of variability**

- Range
  - The *range* of a set of measurements is the difference between the largest and smallest measurements
  - Its major advantage is the ease with which it can be computed
  - Its major shortcoming is its failure to provide information on the dispersion of the values between the two end points.

4.2. Thống kê mô tả về dữ liệu

❖ **Measures of variability**

- Variance
  - This measure of dispersion reflects the values of *all* the measurements
  - The variance of *a population* of N measurements x_1, x_2, …, x_N having a mean m is defined as

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

  - The variance of a *sample* of n measurements x_1, x_2, …, x_n having a mean $\bar{x}$ is defined

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

4.2. Thống kê mô tả về dữ liệu

❖ **Measures of variability**

- Standard deviation
  - The standard deviation of a set of measurements is the square root of the variance.

- Sample standard deviation: $s = \sqrt{s^2}$

- Population standard deviation : $\sigma = \sqrt{\sigma^2}$

4.2. Thống kê mô tả về dữ liệu

❖ **Measures of variability**

- Coefficient of variation
  - The *coefficient of variation* of a set of measurements is the standard deviation divided by the mean value.

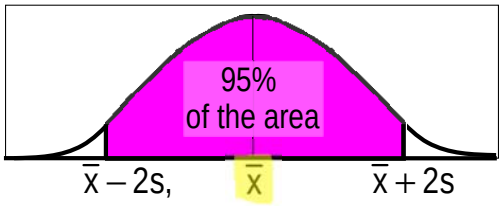
- Sample coefficient of variation: $CV = \frac{S}{\bar{x}}$

- Population coefficient of variation: $CV = \frac{\sigma}{\mu}$

4.2. Thống kê mô tả về dữ liệu

❖ **Measures of variability**

- Interpreting standard deviation
  - The empirical rule: if a sample of measurements has a mound-shaped distribution, the following intervals apply:



($\bar{x} - s, \bar{x} + s$) contains approximately 68% of the measurements

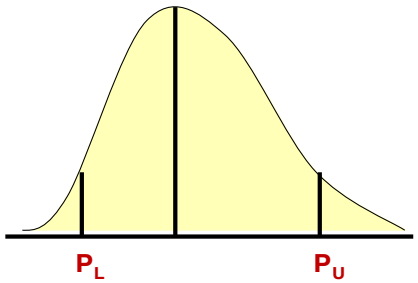
($\bar{x} - 2s, \bar{x} + 2s$) contains approximately 95% of the measurements

($\bar{x} - 3s, \bar{x} + 3s$) contains virtually all of the measurements

4.2. Thống kê mô tả về dữ liệu

❖ **Measures of relative standing and box plots**

- Percentiles: The p^{th} percentile of a set of measurements is the value for which
  - at most, p% of the measurements are less than that value
  - at most, 100 (1 – p)% of all the measurements are greater than that value.



4.2. Thống kê mô tả về dữ liệu

❖ **Measures of relative standing and box plots**

- Percentiles: Commonly used percentiles
  - First (**lower**) decile, $\mathbf{p_{10}}$ = 10th percentile
  - First (lower) **quartile**, $\mathbf{Q_1}$= 25th percentile
  - Second (middle) **quartile**, $\mathbf{Q_2}$ = 50th percentile
  - Third **quartile**, $\mathbf{Q_3}$ = 75th percentile
  - Ninth (**upper**) decile, $\mathbf{p_{90}}$ = 90th percentile

4.2. Thống kê mô tả về dữ liệu

❖ **Measures of relative standing and box plots**

- Find the location of any percentile using the formula:

$$L_P = (n+1)\frac{P}{100}$$

where L_P is the location of the P^{th} percentile

- Example: dataset={30, 33, 30, 35, 30, 31, 34, 33, 27, 29, 28, 35, 37}. Calculate the values of P_{25} =Q_1, P_{50} =Q_2, P_{75} =Q_3

4.2. Thống kê mô tả về dữ liệu

❖ **Measures of relative standing and box plots**

- Example: dataset={30, 33, 30, 35, 30, 31, 34, 33, 27, 29, 28, 35, 37}. Calculate the values of P_{25}, P_{50}, P_{75}

Ans:

- Sort the dataset: 27, 28, 29, 30, 30, 30, 31, 33, 33, 34, 35, 35, 37
- Location of P_{25} : Q_1=L_{25} =(n+1)*(P/100)=(13+1)*25/100=3.5
-> L_{25} =(29+30)/2=29.5
- Location of Q_2=P_{50} : L_{50} =(n+1)*(P/100)=(13+1)*50/100=7
-> L_{50} =31
- Location of P_{75} : Q_3=L_{75} =(n+1)*(P/100)=(13+1)*75/100=10.5
-> L_{75} =(34+35)/2=34.5

4.2. Thống kê mô tả về dữ liệu

Data from list, array,...

```
from statistics import quantiles
List=[30,33,30,35,30,31,34,33,27,29,28,35,37]
Q1=quantiles(List, n=100)[24]
Q2=quantiles(List, n=100)[49]
Q3=quantiles(List, n=100)[74]
print("25th percentile of arr : ",Q1)
print("50th percentile of arr : ",Q2)
print("75th percentile of arr : ",Q3)
```

```
25th percentile of arr :  29.5
50th percentile of arr :  31.0
75th percentile of arr :  34.5
```

4.2. Thống kê mô tả về dữ liệu

Data from Dataframe

```
# importing pandas as pd
import pandas as pd
from statistics import quantiles

# Creating the dataframe
df = pd.DataFrame({"A":[1, 5, 3, 4, 2],
                   "B":[3, 2, 4, 3, 4],
                   "C":[2, 2, 7, 3, 4],
                   "D":[4, 3, 6, 12, 7]})
Q1=df.A.quantile(0.5)
print(Q1)
```

```
3.0
```

	A	B	C	D
0	1	3	2	4
1	5	2	2	3
2	3	4	7	6
3	4	3	3	12
4	2	4	4	7

4.2. Thống kê mô tả về dữ liệu

❖ **Measures of relative standing and box plots**

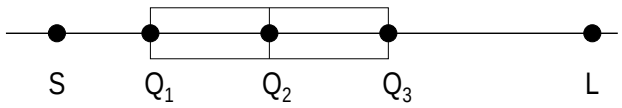
- Interquartile range (IQR):
  - This is a measure of the spread of the middle 50% of the observations.
  - A large value indicates a large spread of observations.

$$\mathbf{IQR = Q_3 - Q_1}$$

4.2. Thống kê mô tả về dữ liệu

❖ **Box plots**

- These are pictorial displays that provide the main descriptive measures of the measurement set:
  - L - The largest measurement
  - Q_3 - The upper quartile
  - Q_2 - The median
  - Q_1 - The lower quartile
  - S - The smallest measurement



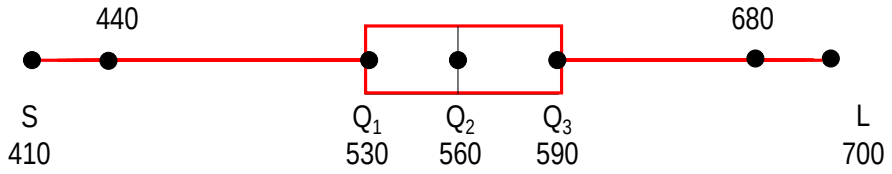
4.2. Thống kê mô tả về dữ liệu

❖ **Box plots**

- Example: create a box plot for the data regarding the number of customers who purchased petrol in an independent petrol station each day in the last 200 days.
- Solution: The following are the relevant summary statistics for the data:
  - smallest number = 410
  - Q_1 = 530
  - Q_2 = 560
  - Q_3 = 590
  - largest number = 700

4.2. Thống kê mô tả về dữ liệu

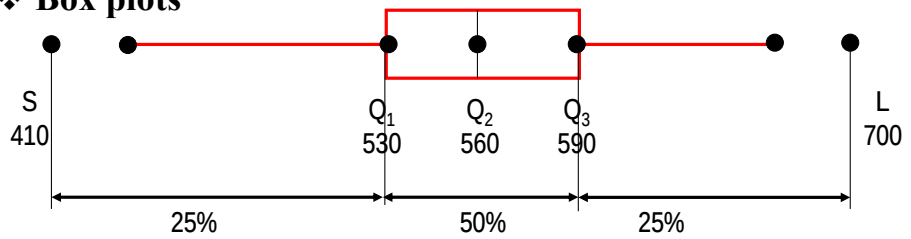
❖ **Box plots**



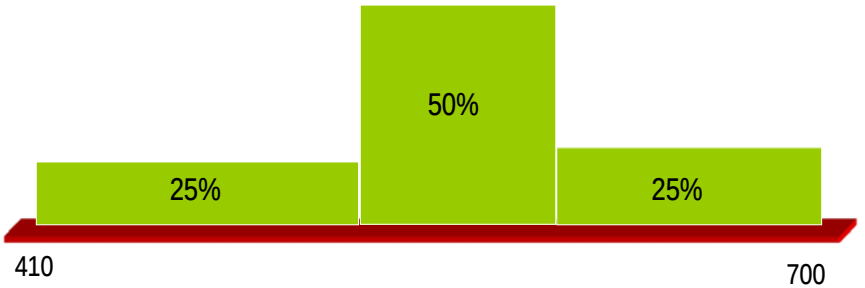
$IQR = Q_3 - Q_1 = 590 - 530 = 60$

4.2. Thống kê mô tả về dữ liệu

❖ **Box plots**



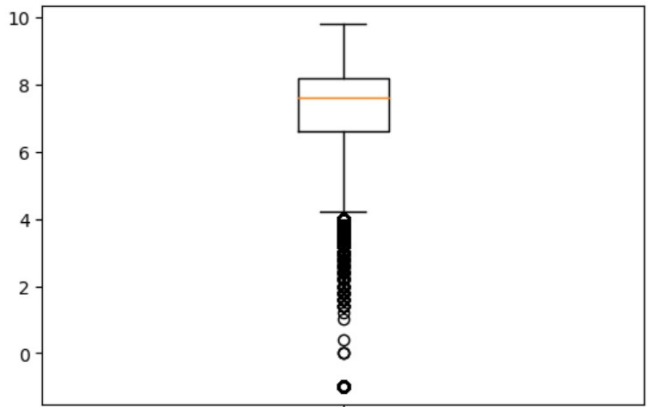
The distribution is very symmetrical.



4.2. Thống kê mô tả về dữ liệu

❖ **Box plots**

```
import matplotlib
from matplotlib import pyplot as plt
matplotlib.rcParams['figure.figsize']=(6,4)
plt.boxplot(df.toan);
```

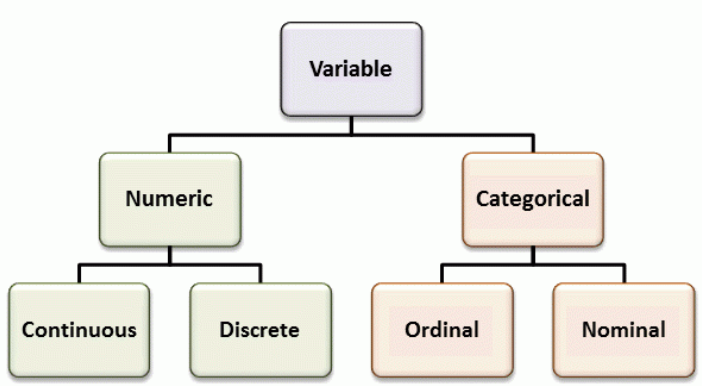


4.3. Làm sạch dữ liệu

- Xử lý dữ liệu bị thiếu (missing data)
- Nhận diện phần tử biên (outliers) và giảm thiểu nhiễu (noisy data)
- Xử lý dữ liệu không nhất quán (inconsistent data)

4.3. Làm sạch dữ liệu

❖ Xử lý dữ liệu bị thiếu (missing data): dùng giá trị thay thế như giá trị trung bình, số trung vị, số mode,…



4.3. Làm sạch dữ liệu

❖ Nhận diện phần tử biên (outliers) và giảm thiểu nhiễu (noisy data)

- Định nghĩa
  - Outliers: những dữ liệu không tuân theo đặc tính/hành vi chung của tập dữ liệu.
  - Noisy data: để giảm thiểu nhiễu (noisy data) thì outliers cần bị loại bỏ ra khỏi tập dữ liệu phân tích.

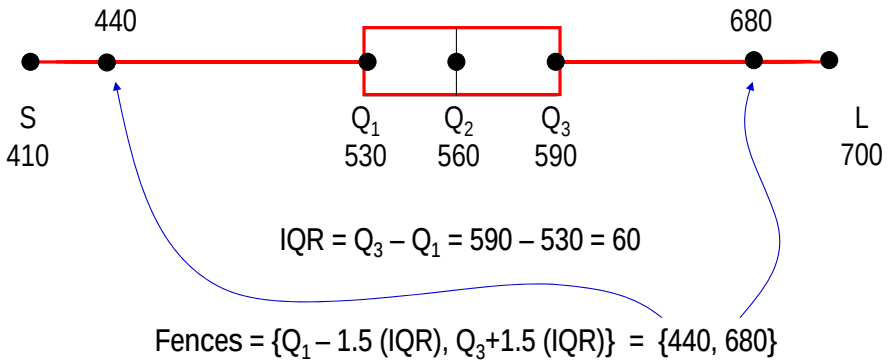
4.3. Làm sạch dữ liệu

❖ Nhận diện phần tử biên (outliers) và giảm thiểu nhiễu (noisy data)

- Giải pháp nhận diện phần tử biên dựa vào:
  - Phân bố thống kê (statistical distributions)
  - Khoảng cách (distance)
  - Mật độ (density)
  - Độ lệch (deviation)
- Giải pháp giảm thiểu nhiễu dựa vào
  - IQR
  - Độ lệch chuẩn hoặc Z-scores
  - …

4.3. Làm sạch dữ liệu

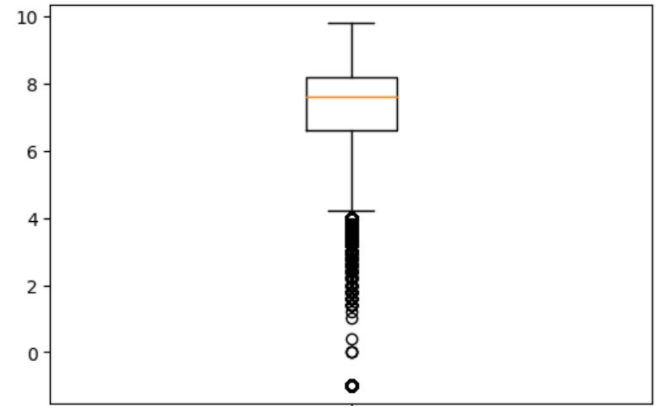
❖ Giảm thiểu nhiễu (noisy data) dựa vào IQR



4.2. Thống kê mô tả về dữ liệu

❖ **Box plots**

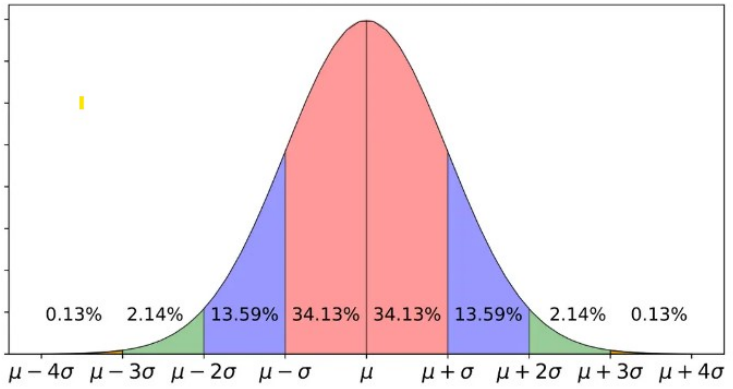
```
import matplotlib
from matplotlib import pyplot as plt
matplotlib.rcParams['figure.figsize']=(6,4)
plt.boxplot(df.toan);
```



4.3. Làm sạch dữ liệu

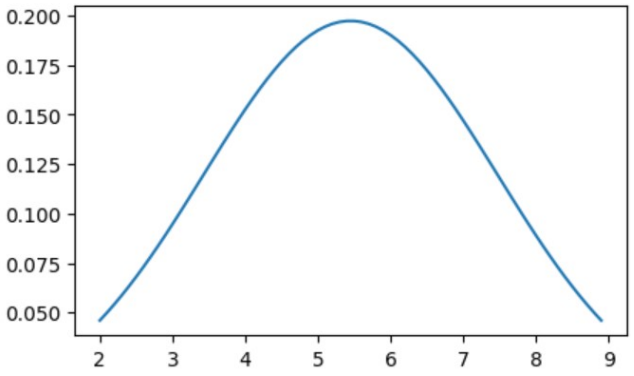
Mật đô phân phối xác suất theo độ lệch tiêu chuẩn (μ, σ)

Normal distribution



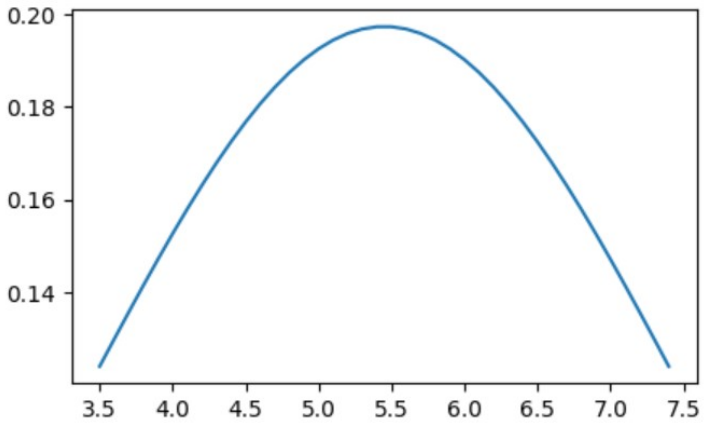
4.3. Làm sạch dữ liệu

```
import matplotlib
from matplotlib import pyplot as plt
matplotlib.rcParams['figure.figsize']=(5,3)
from scipy.stats import norm
data=np.arange(2,9,0.1)
print(data.shape)
plt.plot(data, norm.pdf(data,data.mean(),data.std()))
```



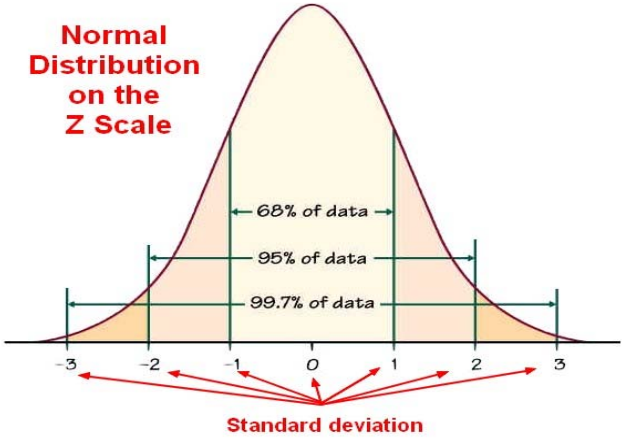
4.3. Làm sạch dữ liệu

```
uper=data.mean()+data.std()
lower=data.mean()-data.std()
indexes=np.where((data>lower) & (data<uper))
data_new=data[indexes]
print(data_new.shape)
plt.plot(data_new, norm.pdf(data_new,data.mean(),data.std()))
```



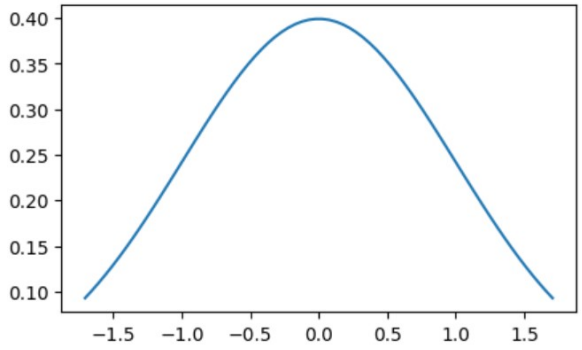
4.3. Làm sạch dữ liệu

❖ Mật đô phân phối xác suất theo Z-score (0,1)



4.3. Làm sạch dữ liệu

```
import matplotlib
from matplotlib import pyplot as plt
matplotlib.rcParams['figure.figsize']=(5,3)
from scipy.stats import norm
from scipy import stats # method of zscore
data=np.arange(2,9,0.1)
data_zscore=stats.zscore(data)
plt.plot(data_zscore, norm.pdf(data_zscore,data_zscore.mean(),data_zscore.std()))
```



4.3. Làm sạch dữ liệu

❖ Xử lý dữ liệu không nhất quán

- Định nghĩa của **dữ liệu không nhất quán**
  - Dữ liệu được ghi nhận khác nhau cho cùng một đối tượng/thực thể
    - Ràng buộc khóa ngoại

4.3. Làm sạch dữ liệu

FIGURE 6.4 RELATIONAL DATABASE TABLES

SUPPLIER

Columns (Attributes, Fields)

Supplier_Number	Supplier_Name	Supplier_Street	Supplier_City	Supplier_State	Supplier_Zip
8259	CBM Inc.	74 5th Avenue	Dayton	OH	45220
8261	B. R. Molds	1277 Gandolly Street	Cleveland	OH	49345
8263	Jackson Composites	8233 Micklin Street	Lexington	KY	56723
8444	Bryant Corporation	4315 Mill Drive	Rochester	NY	11344

Rows (Records, Tuples)

Key Field (Primary Key)

PART

Part_Number	Part_Name	Unit_Price	Supplier_Number
137	Door latch	22.00	8259
145	Side mirror	12.00	8444
150	Door molding	6.00	8263
152	Door lock	31.00	8259
155	Compressor	54.00	8261
178	Door handle	10.00	8259

Primary Key

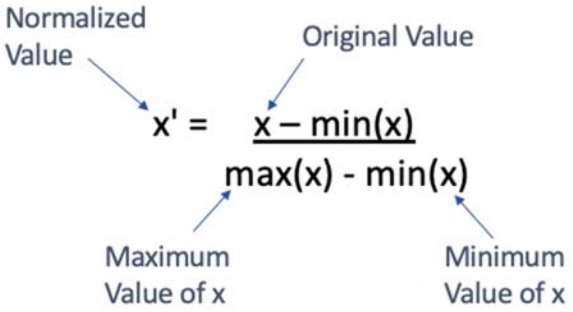
Foreign Key

4.4. Biến đổi dữ liệu

- Biến đổi dữ liệu: quá trình **biến đổi hay kết hợp dữ liệu** vào những dạng thích hợp cho quá trình phân tích và khai phá dữ liệu
  - Làm trơn dữ liệu (rời rạc hóa dữ liệu, loại outliers,…)
  - Chuẩn hoá (normalization)
  - Xây dựng thêm thuộc tính

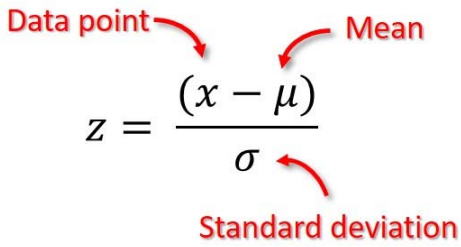
4.4. Biến đổi dữ liệu

- Chuẩn hóa (normalization)
  - min-max normalization



4.4. Biến đổi dữ liệu

- Chuẩn hóa (normalization)
  - z-score normalization



4.4. Biến đổi dữ liệu

- Xây dựng thuộc tính/đặc tính (attribute/feature construction)
  - Các thuộc tính mới được xây dựng và thêm vào từ tập các thuộc tính sẵn có.
  - Hỗ trợ kiểm tra tính chính xác và giúp hiểu cấu trúc của **dữ liệu nhiều chiều.**
  - Hỗ trợ phát hiện thông tin thiếu sót về các mối quan hệ giữa các thuộc tính dữ liệu.

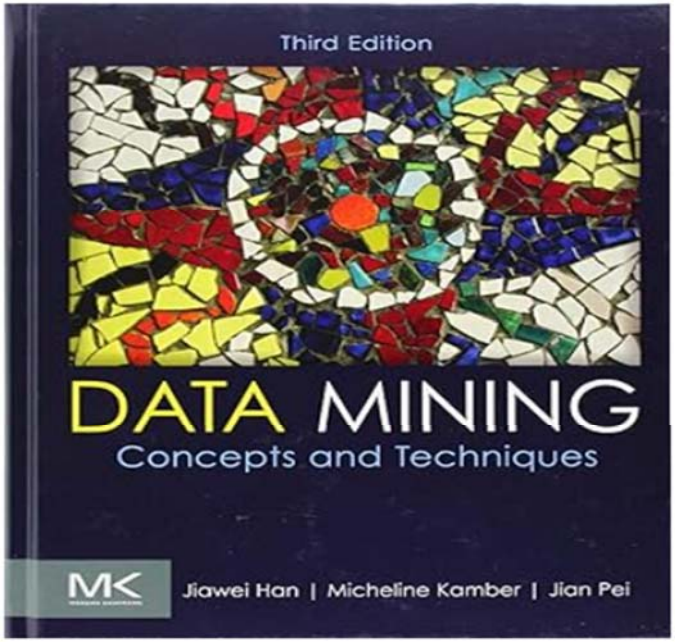
4.5. Rời rạc hóa dữ liệu

- Giảm số lượng giá trị của một thuộc tính liên tục (continuous attribute) bằng các chia miền trị thuộc tính thành các khoảng (**intervals**)
- Các nhãn (**labels**) được gán cho các khoảng (intervals) này và được dùng thay giá trị thực của thuộc tính
- Các trị thuộc tính có thể được phân hoạch theo một phân cấp (**hierarchical**) hay ở nhiều mức phân giải khác nhau (multiresolution)

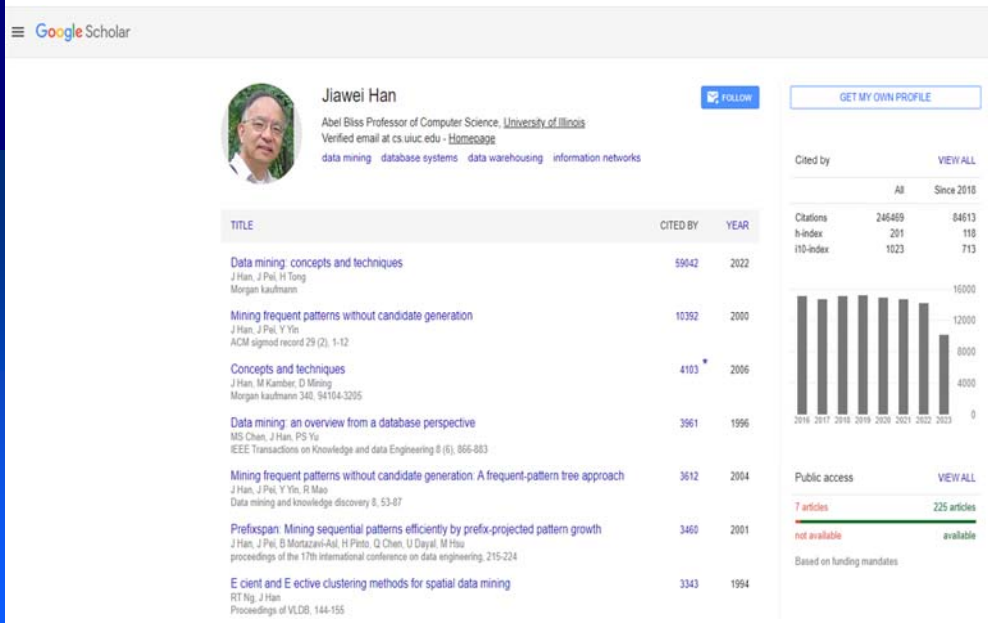
4.5. Rời rạc hóa dữ liệu

- Các phương pháp rời rạc hóa dữ liệu cho các thuộc tính số
  - Binning
  - Cluster analysis
  - …

Further discussion



Further discussion



Hỏi & Đáp

Tham khảo slides từ Textbook "Data mining: Concepts and Techniques" và bài giảng của Trường Đại học bách khoa Tp. HCM