Python (part 2)

KHOA CÔNG NGHỆ THÔNG TIN
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

fit@hcmus

Why Jupyter notebook

- Today
  - More about data science process
  - Demo: using Python to do a data science process

Data science process

Data science process

- ☐ Ask a meaningful question
- ☐ Collect data
- ☐ Explore data
- ☐ Preprocess data
- ☐ Analyze data

  → the answer
- ☐ Communicate results / make decision

What is the purpose of exploring data?

- To **understand more about data**
- From that, we can:
  - Identify problems in data

  If there is a problem, we might need to:
    - Preprocess it (we might need to preprocess it right away in order to continue to explore data)
    - Or even go back to collecting data
  - Refine original questions (if there are) or pose questions which can be answered with this data

What info about (tabular) data do we need to explore?

- **How many** rows and how many columns?
- What is the meaning of each row? Are there **rows having different meaning from the majority**?
- Are there **duplicated rows**?
- What is the meaning of each column?
- What is the current data type of each column? Are there columns having **inappropriate data types**?
- With each numerical column, how are values distributed?
  - What is the percentage of **missing values**?
  - Min? max? ➔ Are they **abnormal**?
- With each categorical column, how are values distributed?
  - What is the percentage of **missing values**?
  - How many different values? Show a few

➔Are they **abnormal**?

What info about (tabular) data do we need to explore?

- Previous slide shows basic info about data we should explore
- In complex cases, we may want to explore additional info about data
- For example, if we want to know more about column distribution, we can compute additional info using descriptive statistics

What is the purpose of preprocessing data?

- During data exploration, when we see a problem about data, we may need to do preprocessing operations to **fix the problem in order to continue to explore data**
- After identifying a specific question (we often reach this point after exploring data and understanding more about data) and the corresponding specific analysis method, we may need to do additional preprocessing operations with the goal: **preparing data which are ready for applying thespecific analysis method**

Demo: using Python to do a data science process

Example from previous lecture

- **Ask a question**: what is the current state of understanding Python of students?
- **Collect data**: let students do quiz about Python in moodle, results can be downloaded as a csv file

Data exploration

- After understanding more about data, we will come back to "**ask a question**" step to refine original questions, or pose new questions

Refined question

- The question “what is the current state of understanding Python of students?” can be refined into 2 more specific questions:
- 1. How are values of the “Grade/16.00” column distributed?
  - *When exploring data, we just knew missing percentage, min, max; here we want to know more ...*
- 2. According to the criterion of being answered correctly by most students, which quiz is in first place, which quiz second, …?

Answer the question

- **Preprocess data** (optional) + **analyze data** to answer question 1
- **Question 1**: How are values of the "Grade/16.00" column distributed?

From data of this column, how to answer this question?

- Option 1: Look at full data and feel …
- Option 2: Summarize data using **descriptive statistics** :

Answer the question

- Different types of data will often have methods different descriptive statistics methods
- Common types of data:
  - Numerical data

    For example: scores, temperature
  - Categorical data
    - Nominal

      For example: color
    - Ordinal

      For example: satisfaction level

Descriptive statistics

- We will use **orange color** to denote descriptive statistics for categorical data, and **green color** for numerical data
- Summarize data level one (center): **mean**, **median**, **mode**
- Summarize data level two (center + range): **mean & standard deviation, lower quartile & median & upper quartile**
- Summarize data level three (full distribution): **histogram**, **bar plot**

Mean

- Give numerical data consisting of elements with values respectively: $v_1, v_2, \ldots, vn$
- $mean = \frac{1}{n}\sum_{i=1}^{n} v_i$

Median

- Median = 50^{th} percentile
- p^{th} percentile ($0 \leq p \leq 100$) of a list of **values** is a value which tells us: there are about p% of values in the list < this **value**
- Example: "75^{th} percentile of the quiz scores = 8" means there are about 75% of students having quiz scores < 8

Median

- There are different ways to compute pth percentile, here is one way:
- 1. Sort values in the list in ascending order
- 2. Find the location corresponding to p% of values in the list:

  location = $p/100 \times$ the number of values in the list

  If it is not an integer, round it up
- 3. pth percentile = value at this location in the sorted list

Median

- Given this list: 1, 5, 3, 4, 2

  Median = 50th percentile = 3
- Given this list: 1, 5, 3, 4, 2, 6

  Median = 50th percentile = 3

Median vs mean

- Given this list: 1, 2, 3, 4, 5
  - Median = 3
  - Mean = 3
- Given this list: 1, 2, 3, 4, **500**
  - Median = 3
  - Mean = **102**
- Median is less affected by **outlier** (a value lying far away from the range of most values) than mean!

Mode

- **Mode** = the most frequent value
- A number in a set of numbers that appears the most often.
- For example: if a set of numbers contained the following digits 1, 1, 3, 5, 6, 6, 7, 7, 7, 8, the mode would be 7, as it appears the most out of all the numbers in the set.

Meand & standard deviation

- The **standard deviation** is the average amount of ***variability*** in your dataset.
  - It tells you, on average, how far each value lies from the mean.
- A high standard deviation ➔ values are generally far from the mean
- A low standard deviation ➔ values are clustered close to the mean.
- Given numerical data consisting of elements with values respectively: $v_1, v_2, \ldots, vn$

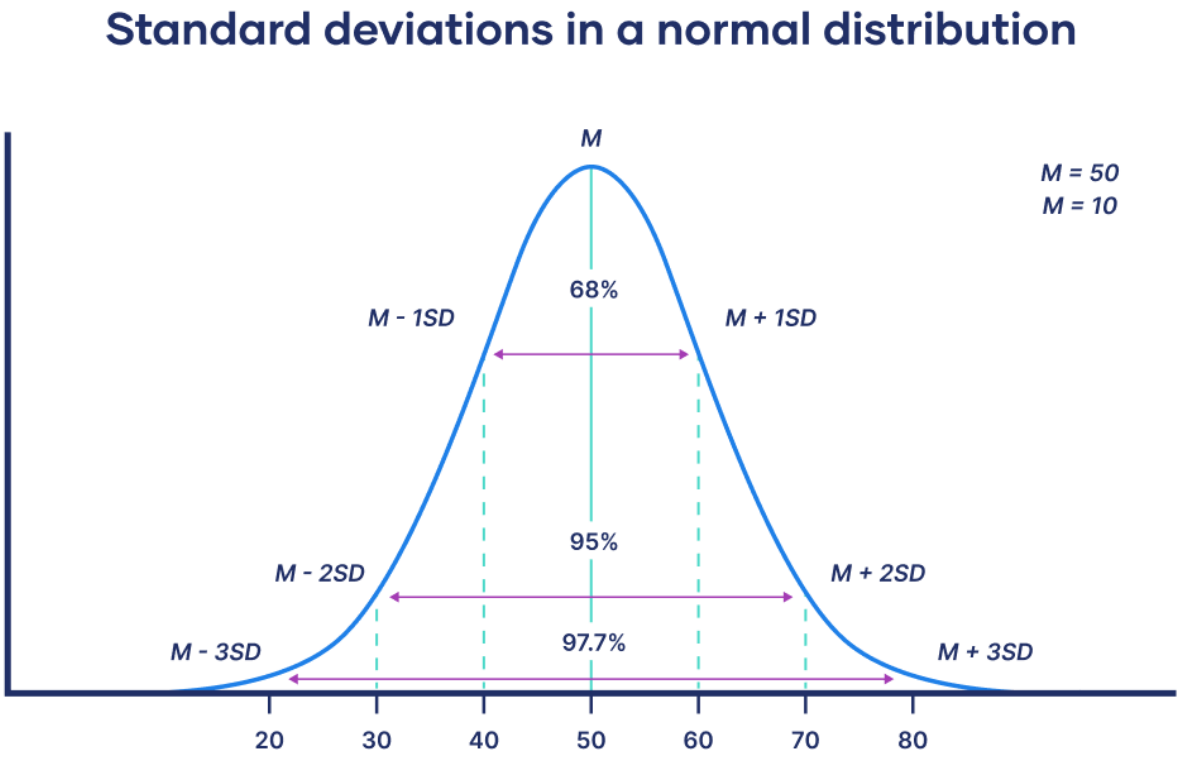
$$SD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(v_i - mean)^2}$$

Standard deviation example

- How data is distributed in a **normal distribution**. The empirical rule, or the 68-95-99.7 rule:
- Around 68% of scores are within 1 SD of the mean,
- Around 95% of scores are within 2 SD of the mean,
- Around 99.7% of scores are within 3 SD of the mean.

Standard deviation example

- The data follows a normal distribution with a mean score of 50 and a standard deviation of 10.
- Around 68% of scores are between 40 and 60.
- Around 95% of scores are between 30 and 70.
- Around 99.7% of scores are between 20 and 80.



Mean & standard deviation

- **Chebychev** discovered: with ***any*** list of values, the range mean ± zSD collects ***at least***

$1 - \frac{1}{z^2} \times 100\%$ of values in the list

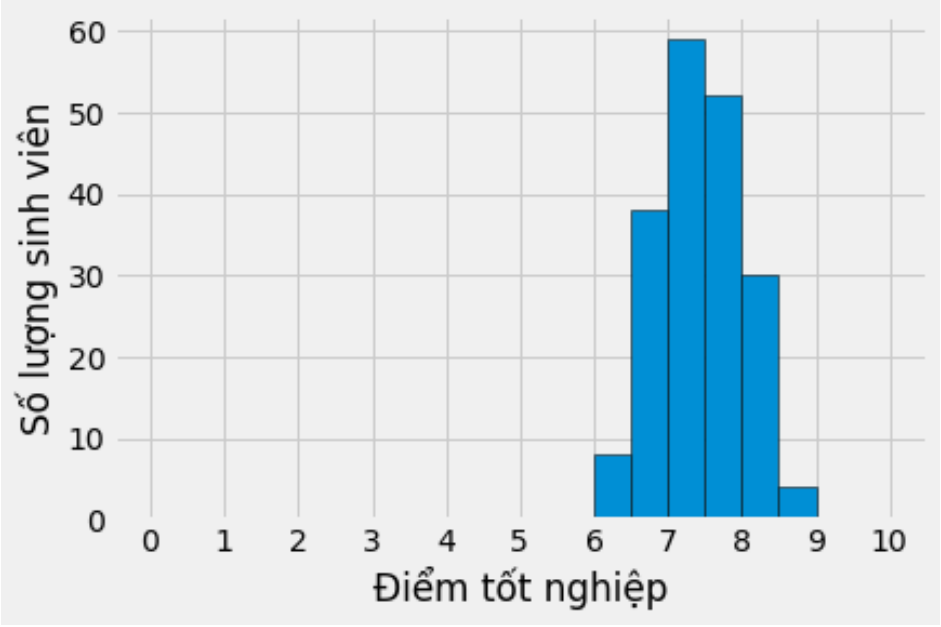
- Mean ± 1SD collects at least 0% of values
- Mean ± 2SD collects at least 75% of values
- Mean ± 3SD collects at least 88.9% of values

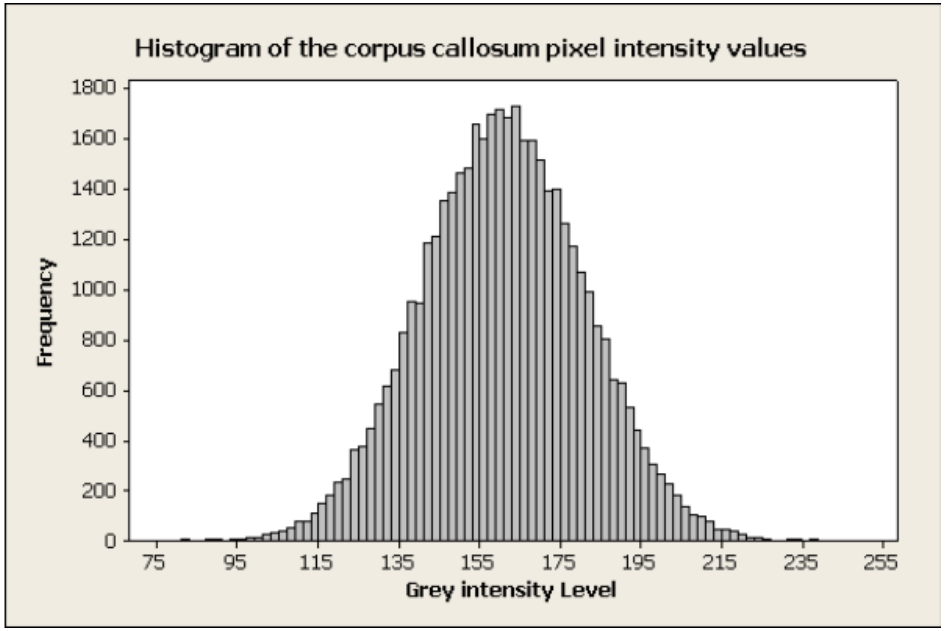
Lower quartile & median & upper quartile

- Lower quartile = 25th percentile
- Median = 50th percentile
- Upper quartile = 75th percentile

Histogram

- Histogram is a graphical representation of the distribution of data
- A frequency distribution shows how often each different value in a set of data occurs.
- A histogram is the most commonly used graph to show frequency distributions.

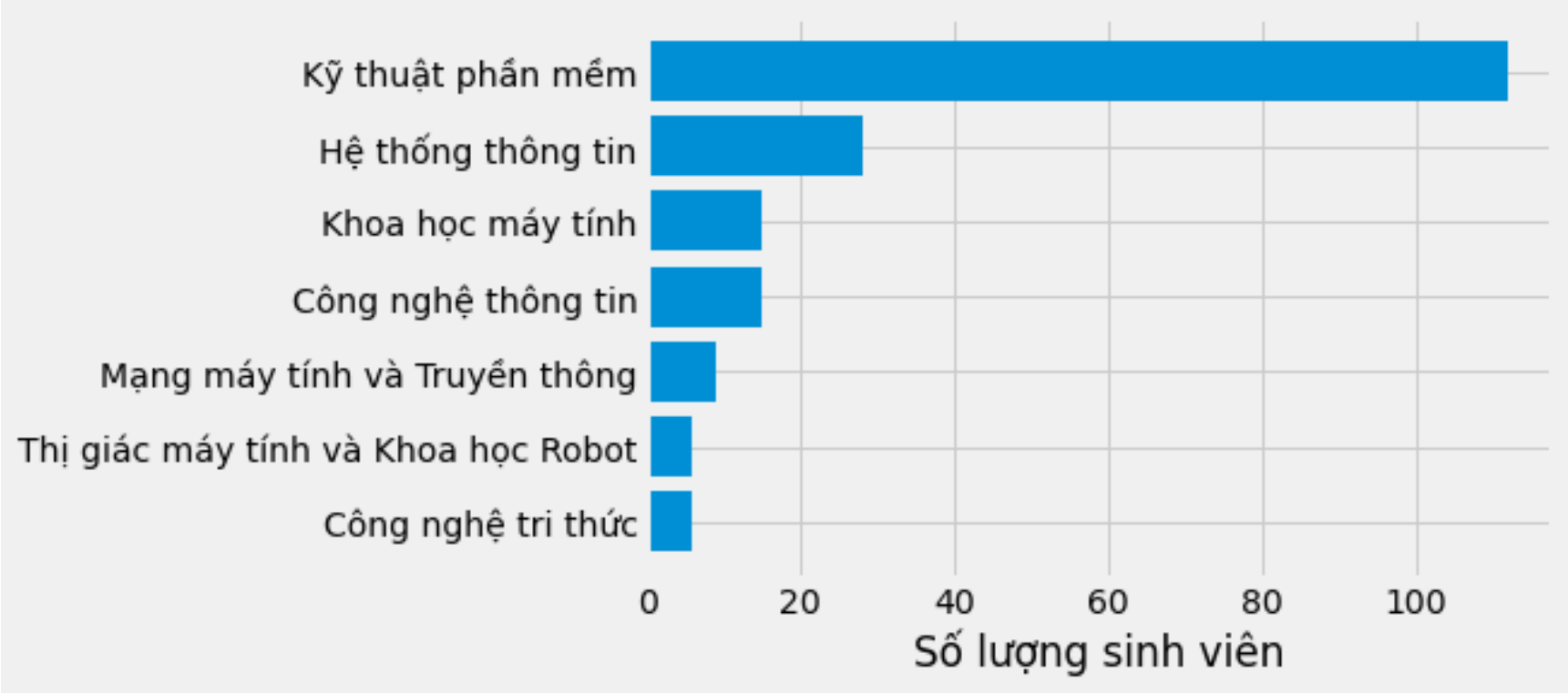




Bar chart

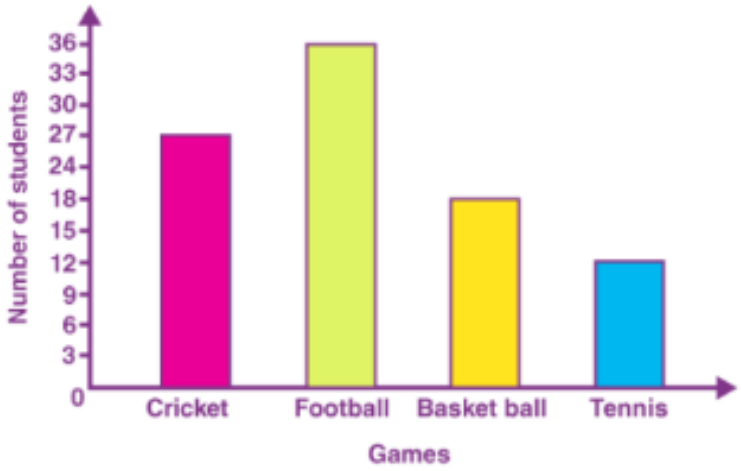
- **Bar chart** is a way of showing the distribution of data with a ***categorical data*** set.
- Bar charts can be represented horizontally or vertically, they can also represent more than one set of data.
- One axis is labelled with the category/group and the other labelled with the frequency of the category/group.

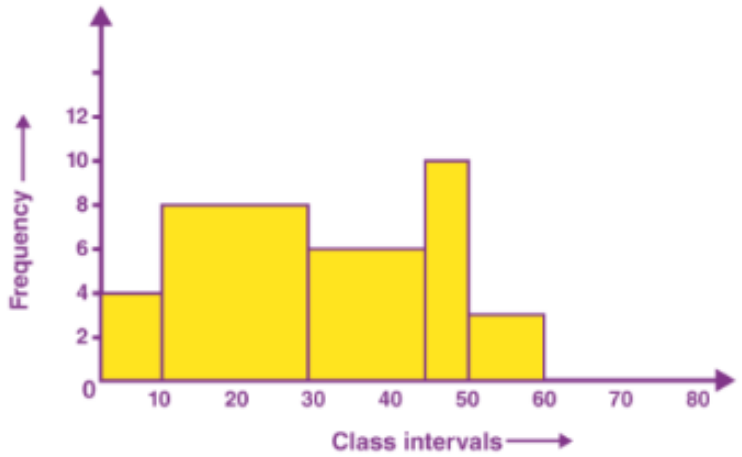
Bar chart


Kỹ thuật phần mềm
Hệ thống thông tin
Khoa học máy tính
Công nghệ thông tin
Mạng máy tính và Truyền thông
Thị giác máy tính và Khoa học Robot
Công nghệ tri thức
0
20
40
60
80
100
Số lượng sinh viên

Bar Chart vs. Histogram

Bar chart	Histogram
Graphical representation of **categorical** data	Graphical representation of **numberal** data
There is **equal space** between each pair of consecutive bars	There is **no space** between the consecutive bars
The **height** of the bars shows the frequency, and the width of the bars are **same**	The **area** of rectangular bars shows the frequency and the width of the bars **need not** to be same





Reference