

# Phương pháp nghiên cứu trong khoa học liên ngành

---

**Nguyễn Bích Ngọc**

Khoa các khoa học liên ngành, ĐHQGHN

# Phân tích định lượng

---

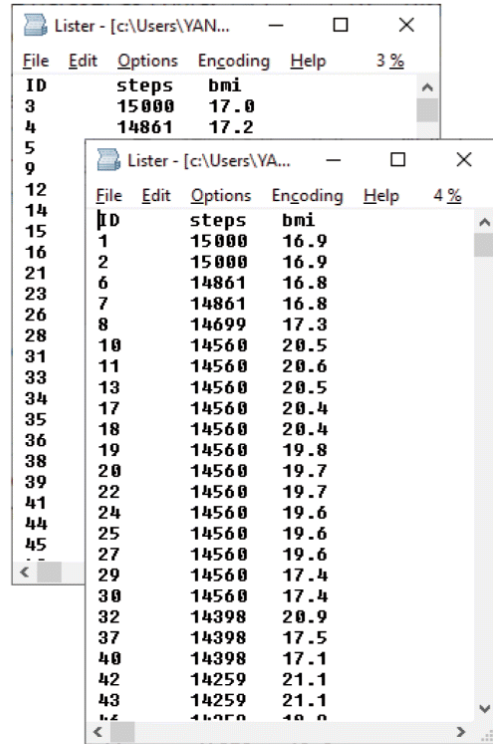
# Các bước phân tích định lượng

- Làm sạch dữ liệu
- Tìm hiểu dữ liệu/Biểu diễn dữ liệu
- Thống kê mô tả
- Thống kê suy luận

# Tìm hiểu dữ liệu/Biểu diễn dữ liệu

- Là bước không thể bỏ qua
- Giúp phát hiện những vấn đề trong dữ liệu
- Giúp có hình dung chung về dữ liệu và các mối tương quan giữa các dữ liệu

a



The image shows a Notepad window titled 'Lister - [c:\Users\YAN...]' with a menu bar (File, Edit, Options, Encoding, Help) and a status bar (3 %). The text content is a list of data points with three columns: ID, steps, and bmi. The data is as follows:

ID	steps	bmi
3	15000	17.0
4	14861	17.2
5		
9		
12		
14		
15	1	15000
16	2	15000
21	6	14861
23	7	14861
26	8	14699
28	10	14560
31	11	14560
33	13	14560
34	17	14560
35	18	14560
36	19	14560
38	20	14560
39	22	14560
41	24	14560
44	25	14560
45	27	14560
	29	14560
	30	14560
	32	14398
	37	14398
	40	14398
	42	14259
	43	14259
	44	14259
	45	14259

Yanai, I., Lercher, M. A hypothesis is a liability. *Genome Biol* **21**, 231 (2020). <https://doi.org/10.1186/s13059-020-02133-w>

a

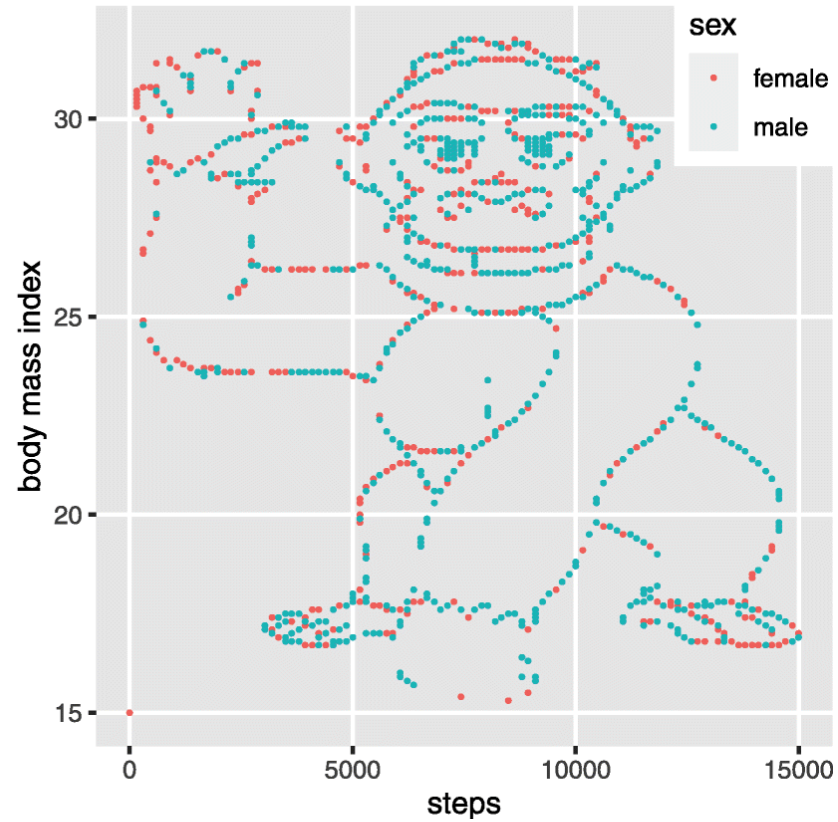
Listner - [c:\Users\YAN...]

ID	steps	bmi
3	15000	17.0
4	14861	17.2

Listner - [c:\Users\YA...]

ID	steps	bmi
1	15000	16.9
2	15000	16.9
6	14861	16.8
7	14861	16.8
8	14699	17.3
10	14560	20.5
11	14560	20.6
13	14560	20.5
17	14560	20.4
18	14560	20.4
19	14560	19.8
20	14560	19.7
22	14560	19.7
24	14560	19.6
25	14560	19.6
27	14560	19.6
29	14560	17.4
30	14560	17.4
32	14398	20.9
37	14398	17.5
40	14398	17.1
42	14259	21.1
43	14259	21.1
44	14259	21.1
45	14259	21.1

b



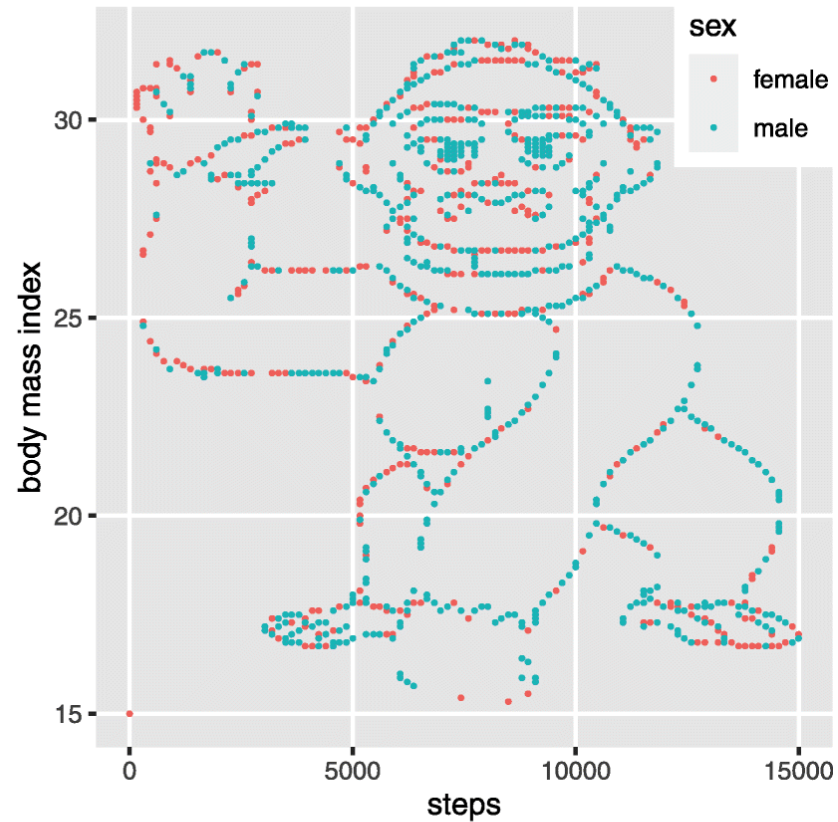
Yanai, I., Lercher, M. A hypothesis is a liability. *Genome Biol* **21**, 231 (2020). <https://doi.org/10.1186/s13059-020-02133-w>

a

Figure a shows two screenshots of a spreadsheet application (likely Excel) displaying data for individuals (ID, steps, bmi). The top screenshot shows a small dataset with 4 rows. The bottom screenshot shows a larger dataset with 45 rows.

ID	steps	bmi
3	15000	17.0
4	14861	17.2
5		
12		
14		
15	1	15000
16	2	15000
21	6	14861
23	7	14861
26	8	14699
28	10	14560
31	11	14560
33	13	14560
34	17	14560
35	18	14560
36	19	14560
38	20	14560
39	22	14560
41	24	14560
44	25	14560
45	27	14560
29	29	14560
30	30	14560
32	32	14398
37	37	14398
40	40	14398
42	42	14259
43	43	14259
44	44	14259
45	45	14259

b

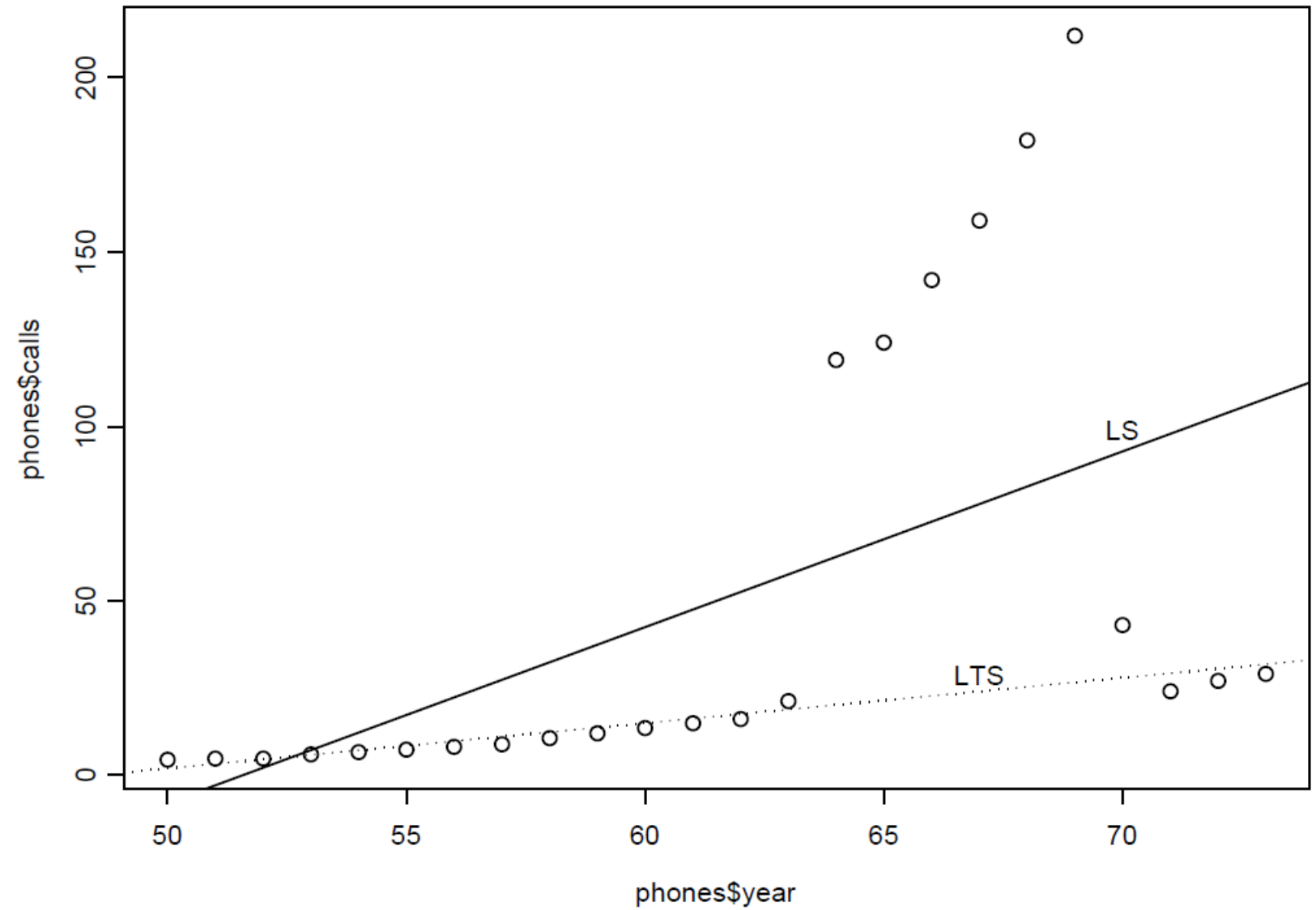


c

	Gorilla <u>not</u> discovered	Gorilla discovered
Hypothesis-focused	14	5
Hypothesis-free	5	9

Yanai, I., Lercher, M. A hypothesis is a liability. *Genome Biol* **21**, 231 (2020). <https://doi.org/10.1186/s13059-020-02133-w>

- Dữ liệu điện thoại
- Cuộc gọi (triệu) ra nước ngoài từ Bỉ từ 1950-1973.





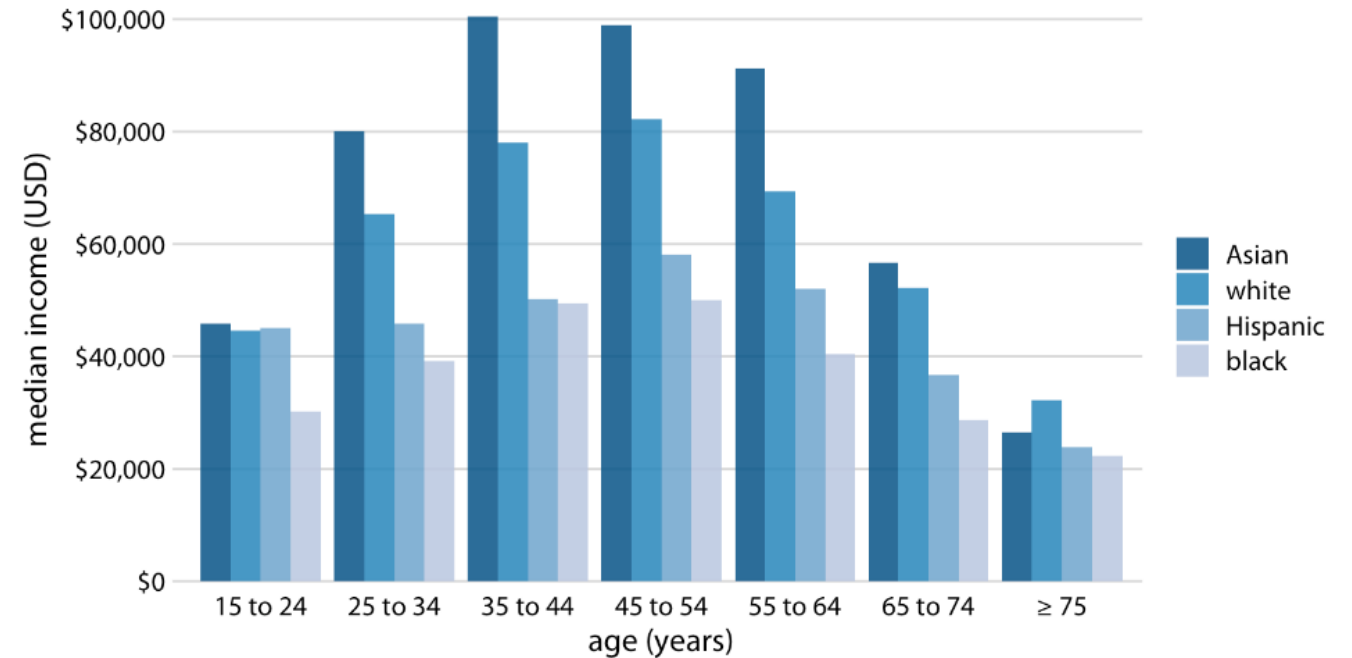
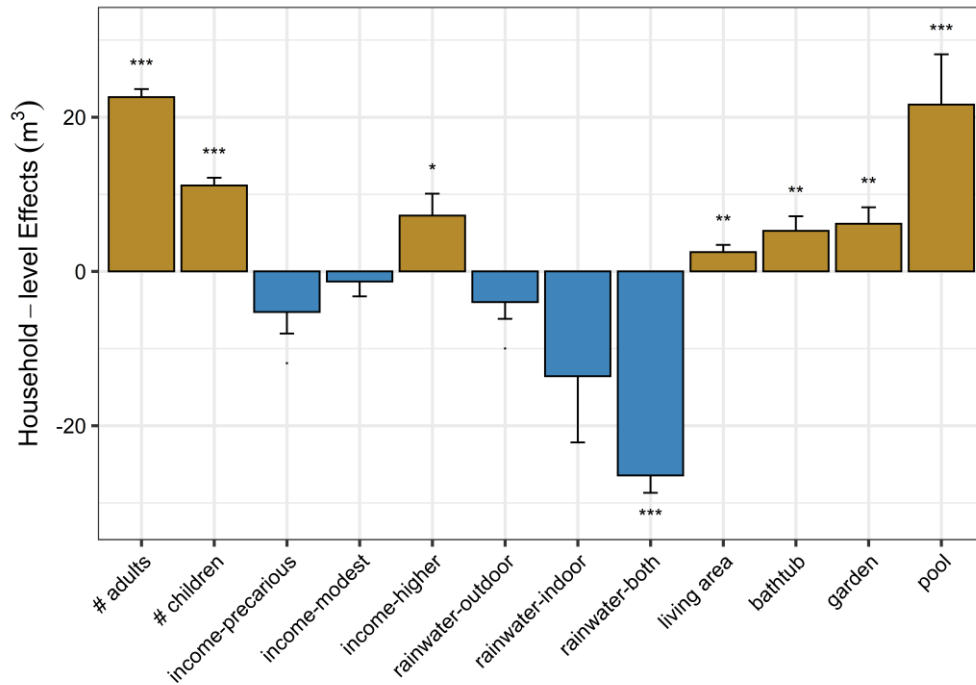
# Đồ thị

- Rõ ràng
- Chính xác
- Hiệu quả
- Tối đa thông tin, tối thiểu mực in

[https://www.ted.com/talks/hans\\_rosling\\_the\\_best\\_stats\\_you\\_ve\\_ever\\_seen](https://www.ted.com/talks/hans_rosling_the_best_stats_you_ve_ever_seen)

# Đồ thị thông thường cho biến liên tục

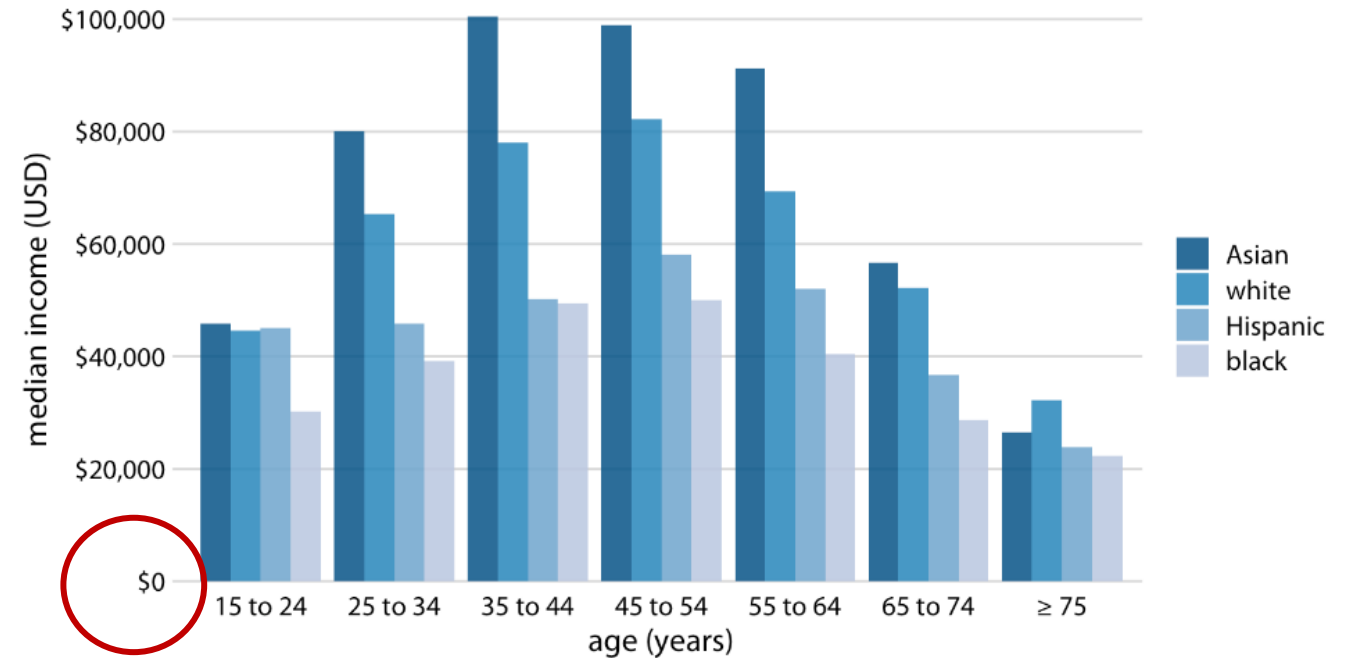
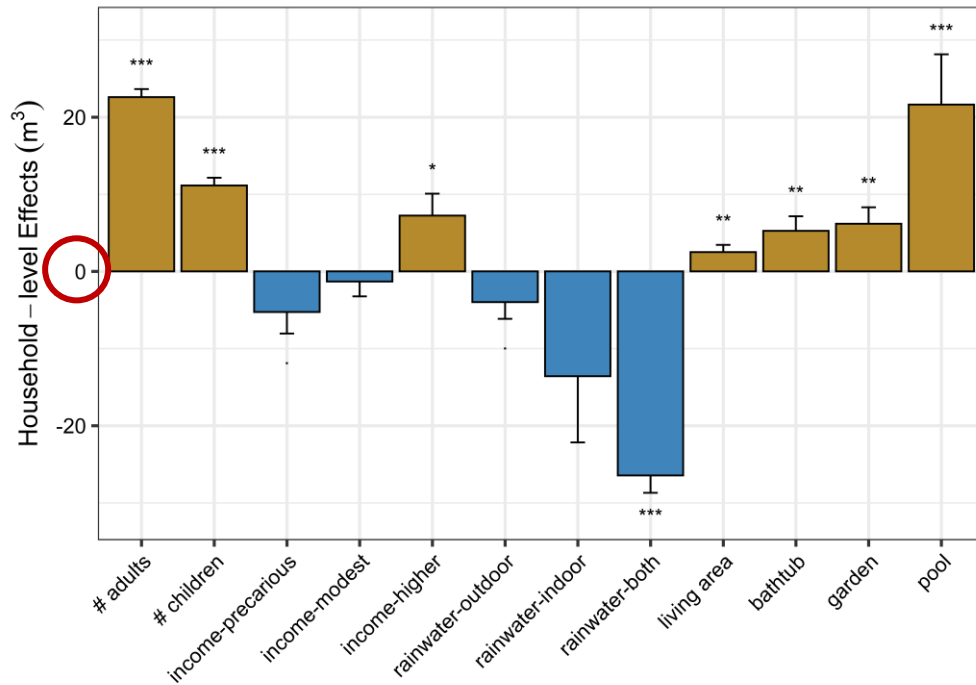
## Đồ thị cột – bar chart



Wilke (2018)

# Đồ thị thông thường cho biến liên tục

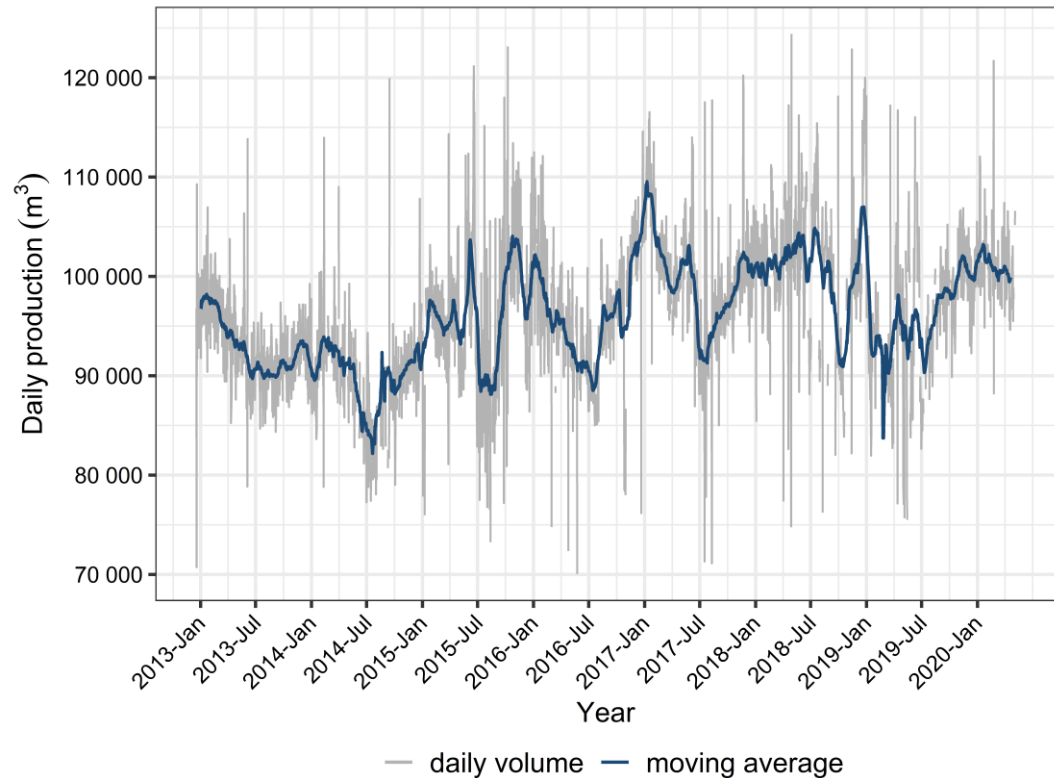
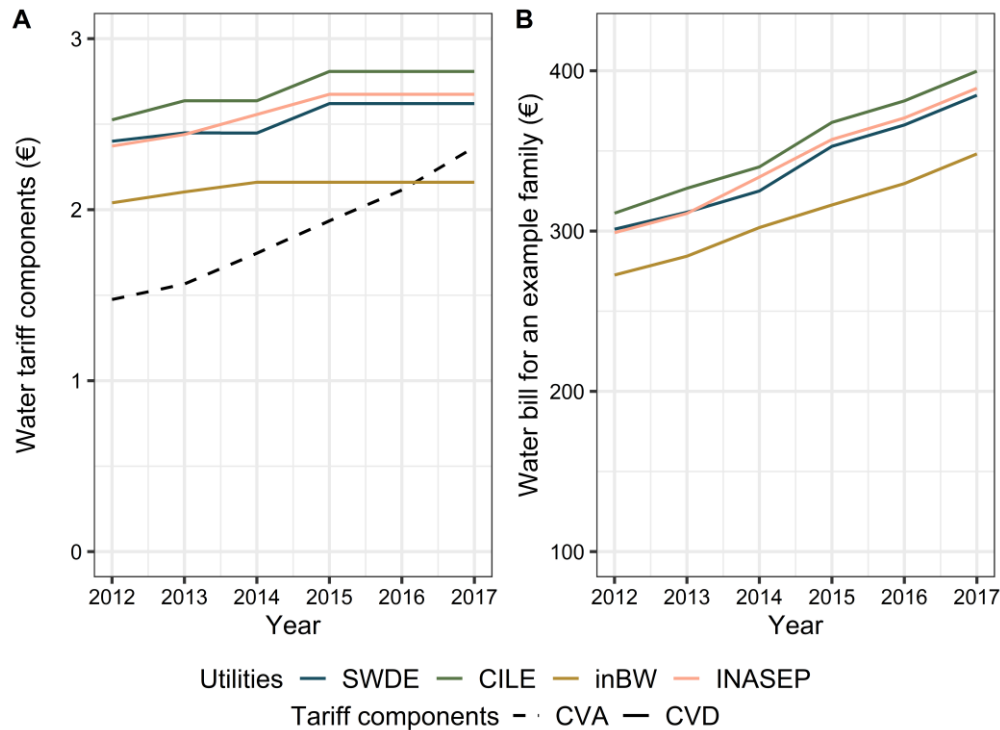
## Đồ thị cột – bar chart



Source: Wilke (2018)

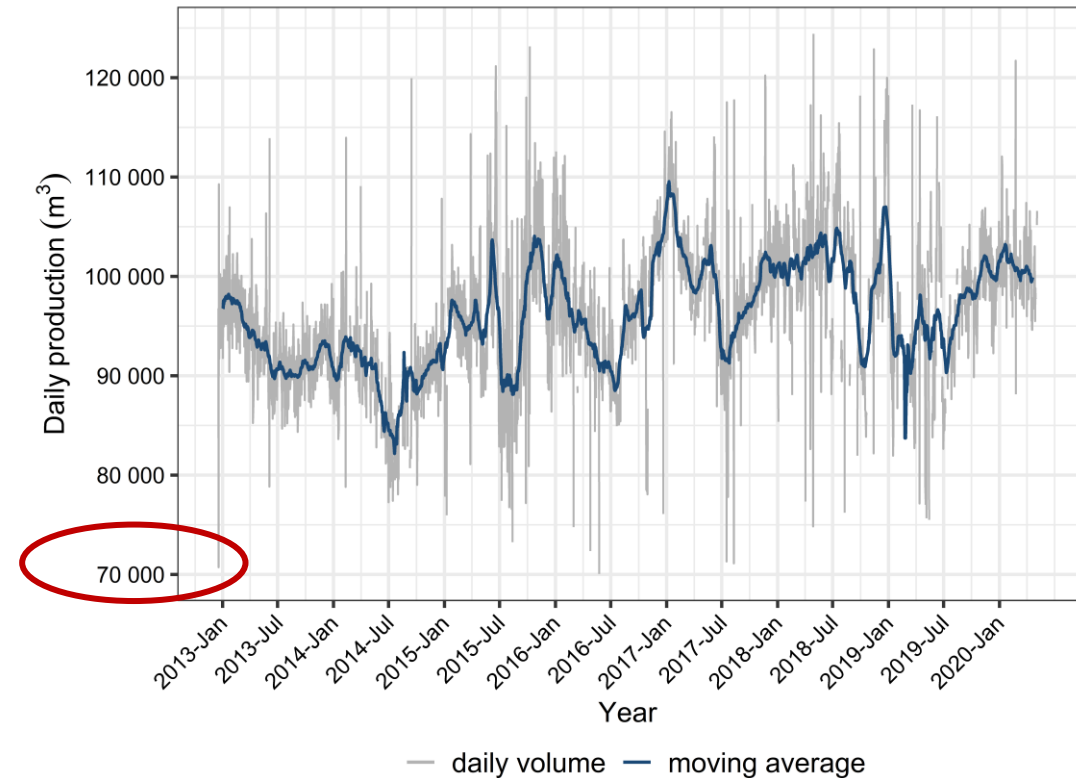
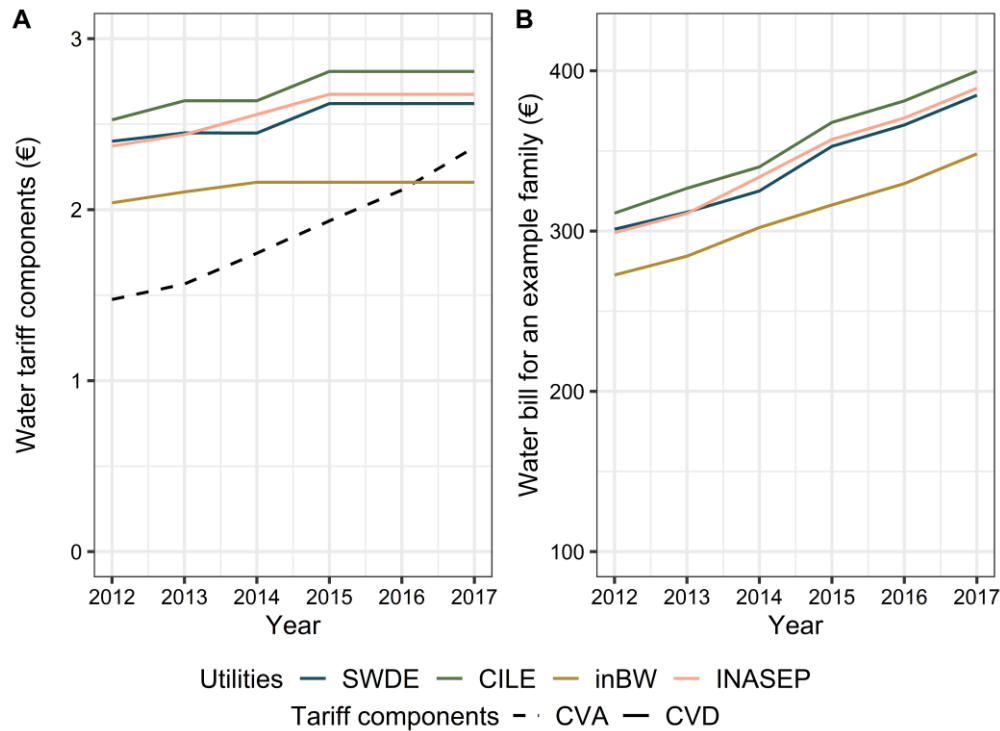
# Đồ thị thông thường cho biến liên tục

## Đồ thị đường – line chart



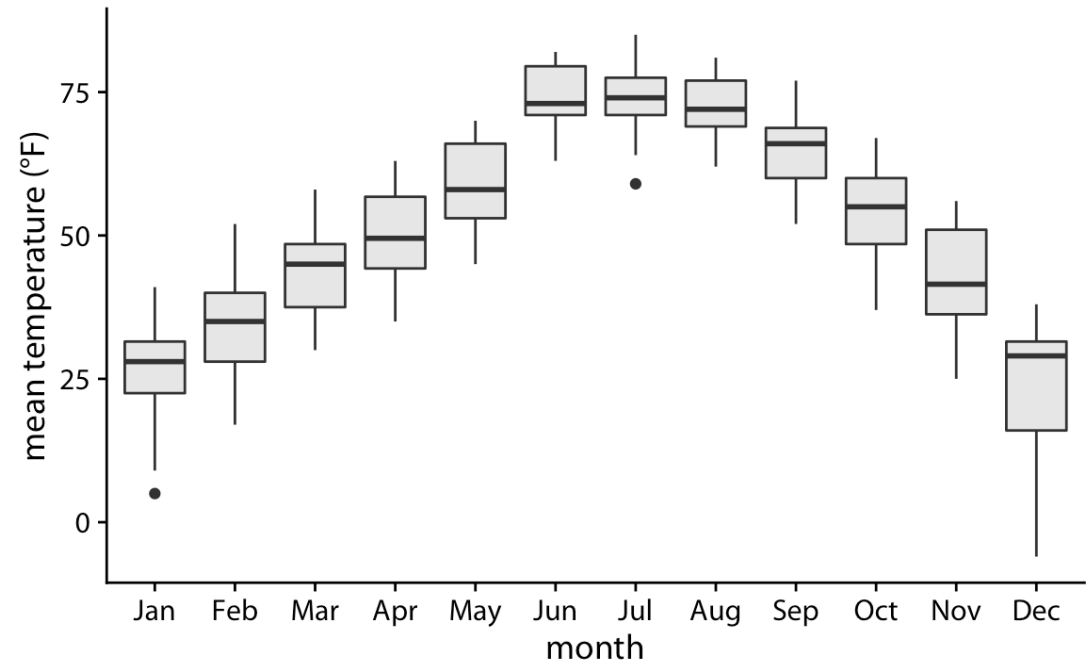
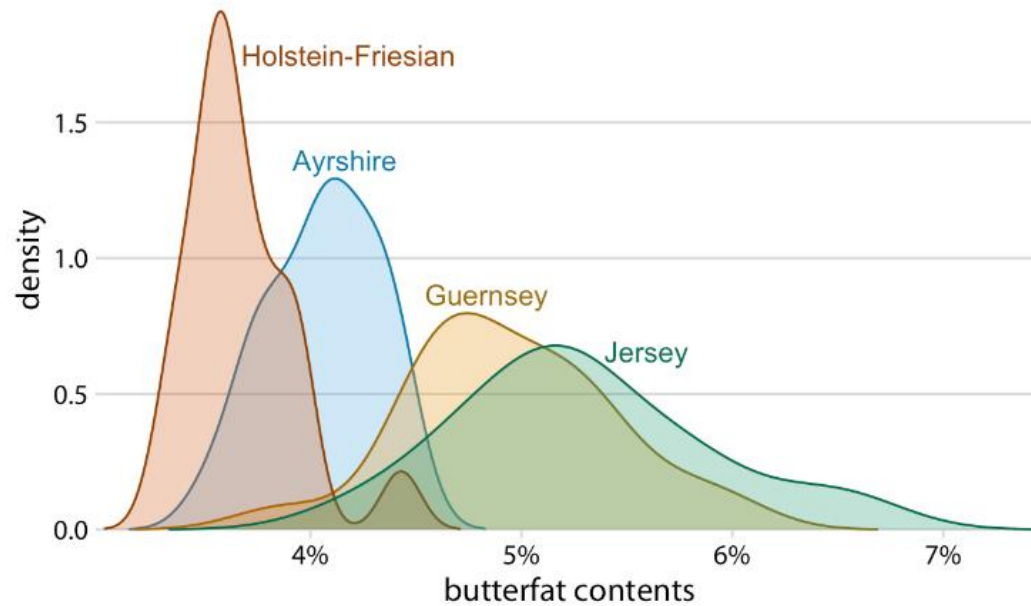
# Đồ thị thông thường cho biến liên tục

## Đồ thị đường – line chart



# Đồ thị thông thường cho biến liên tục

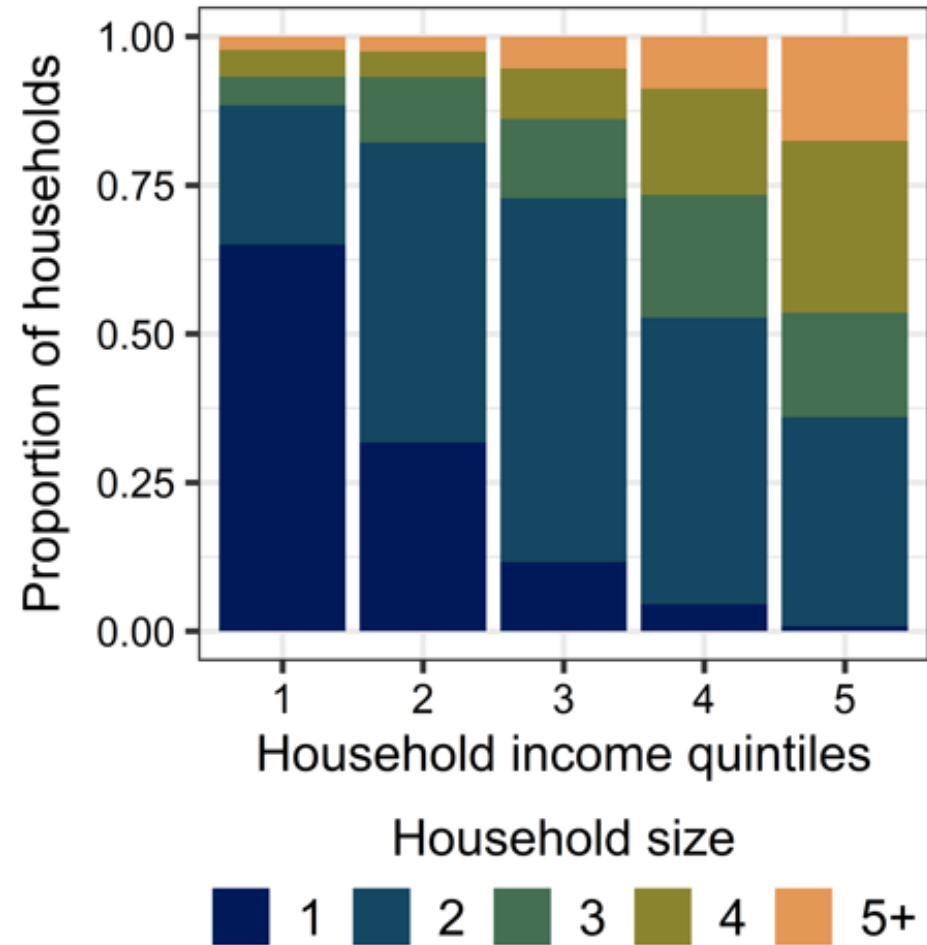
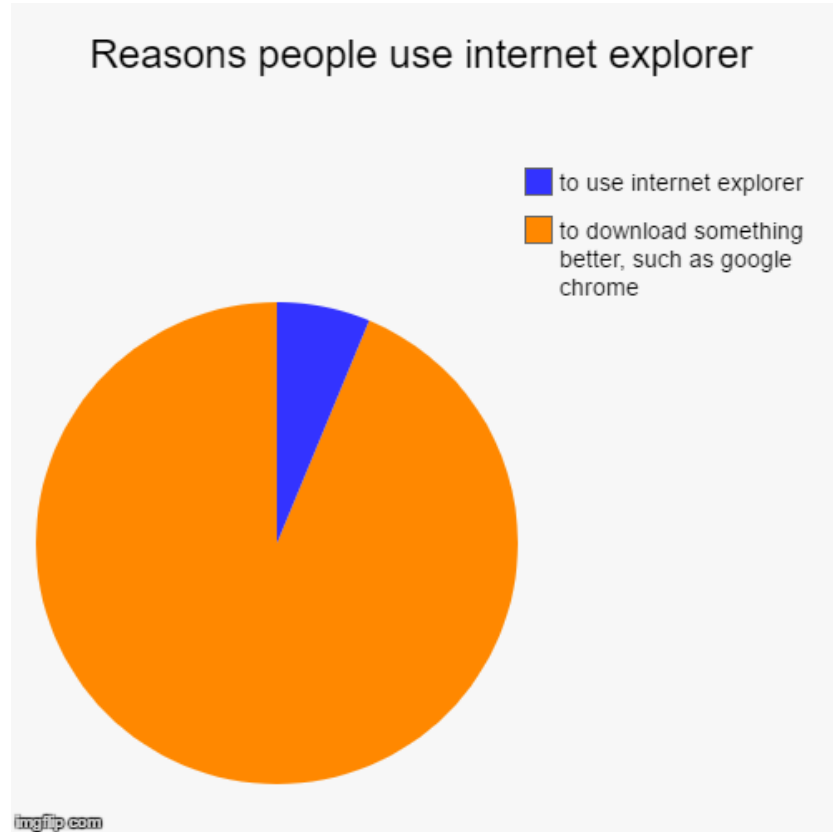
## Biểu đồ tần suất - Histogram



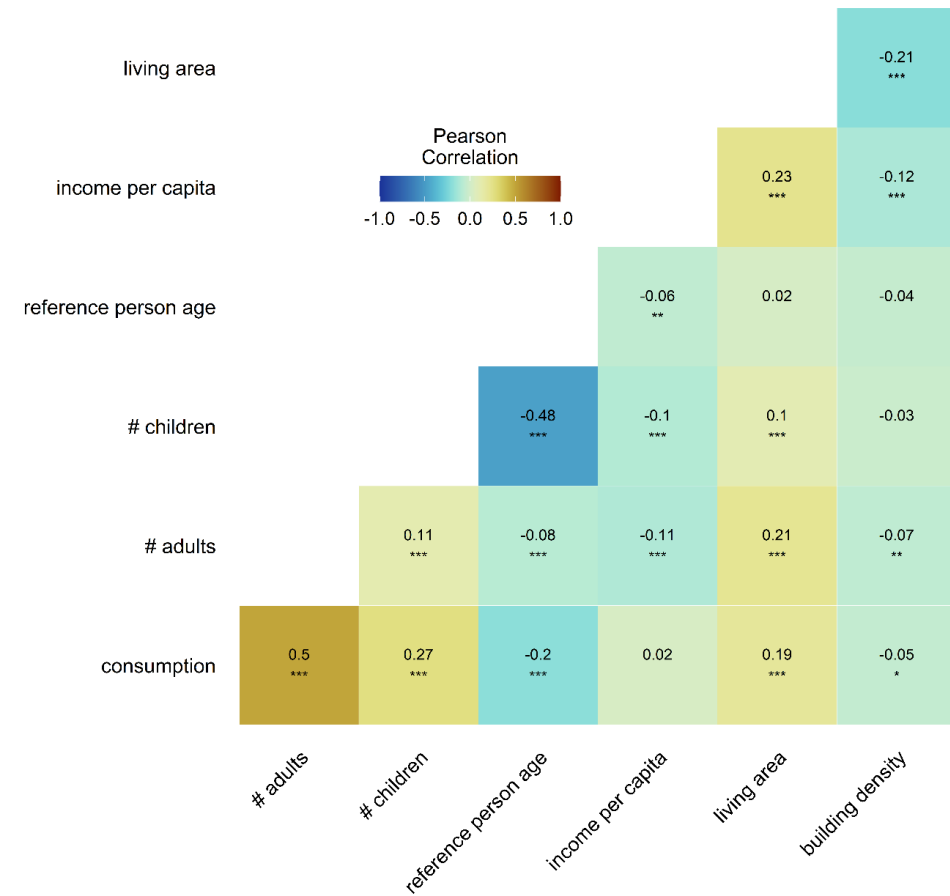
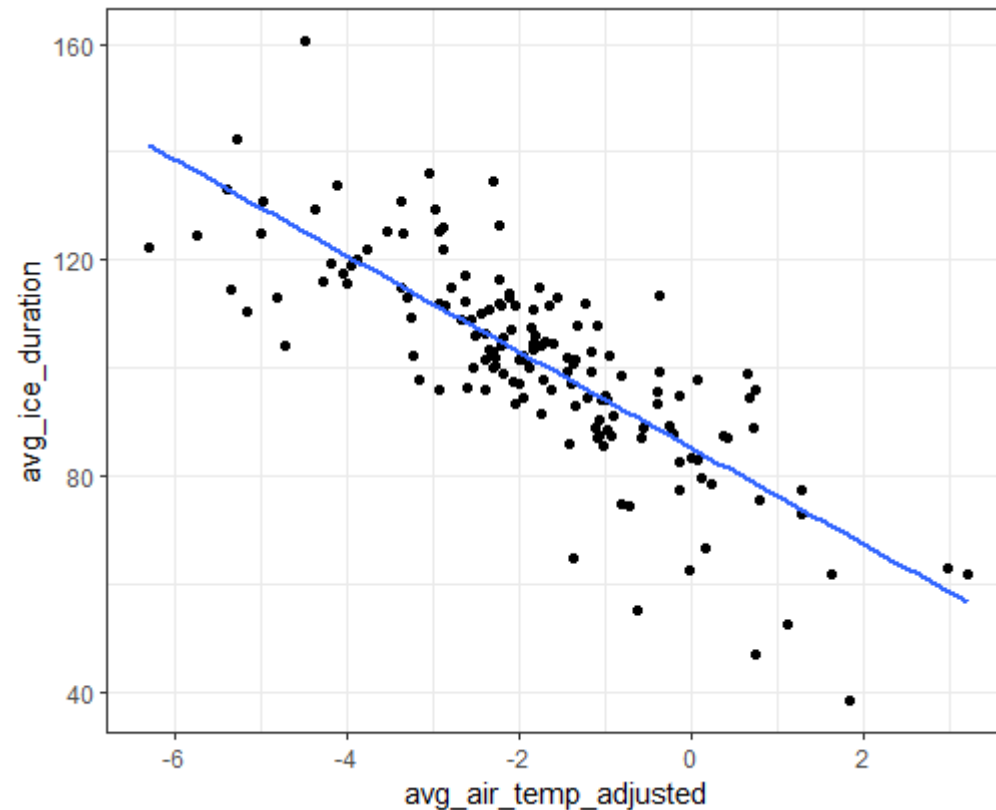
## Đồ thị hộp – Box plots

Source Wilke (2018)

# Đồ thị thông thường cho biến gián đoạn

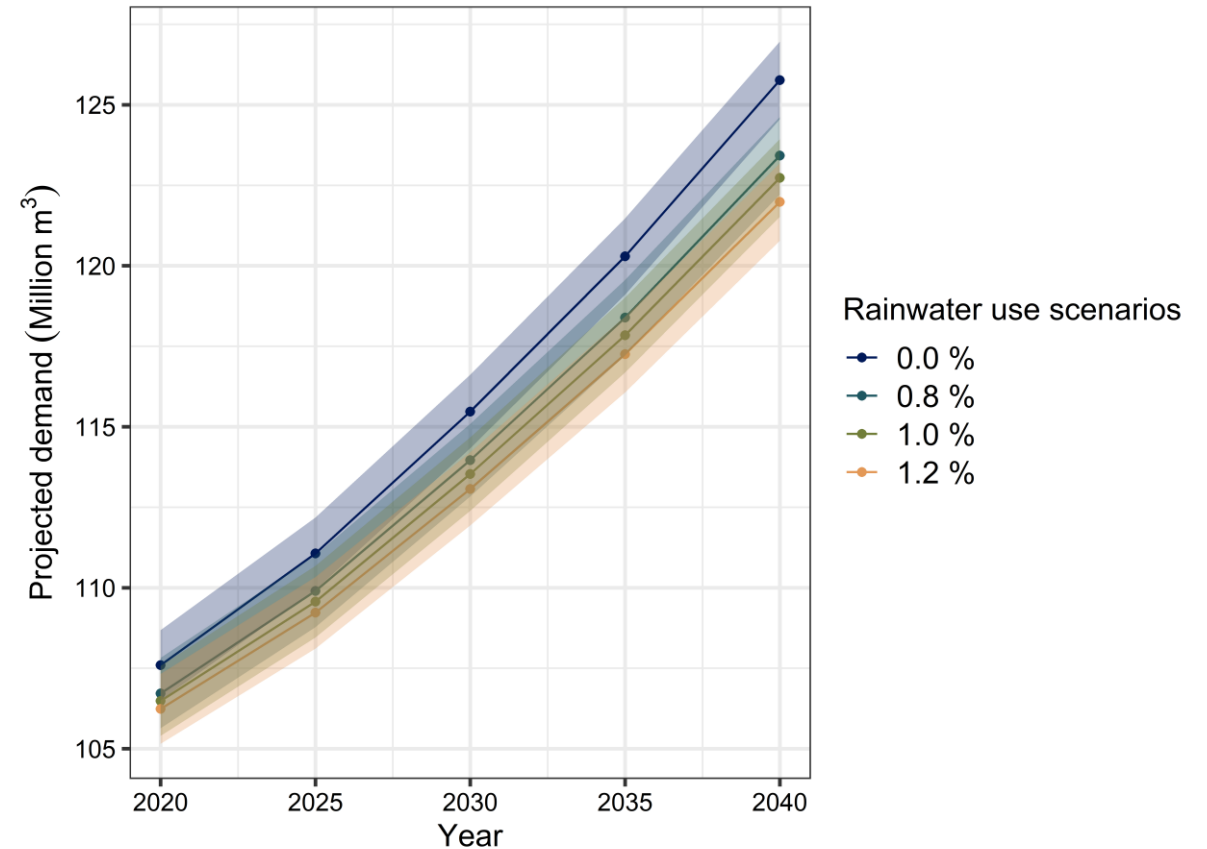
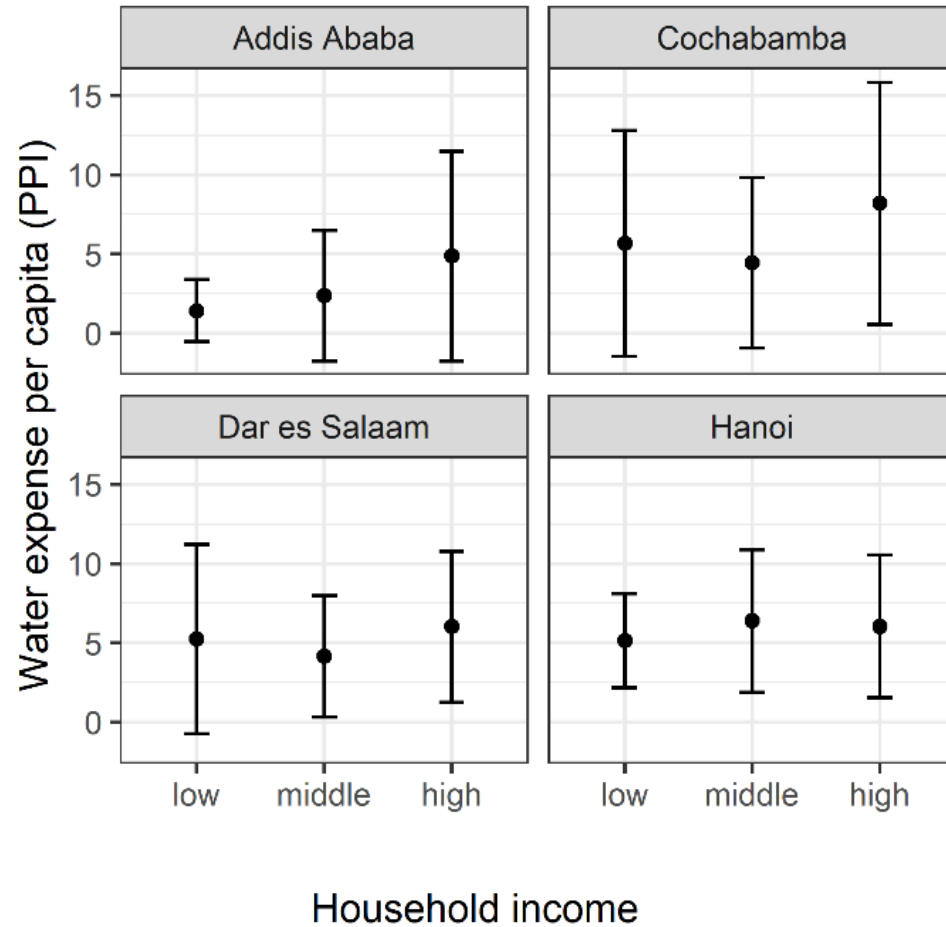


# Biểu diễn quan hệ giữa các biến liên tục

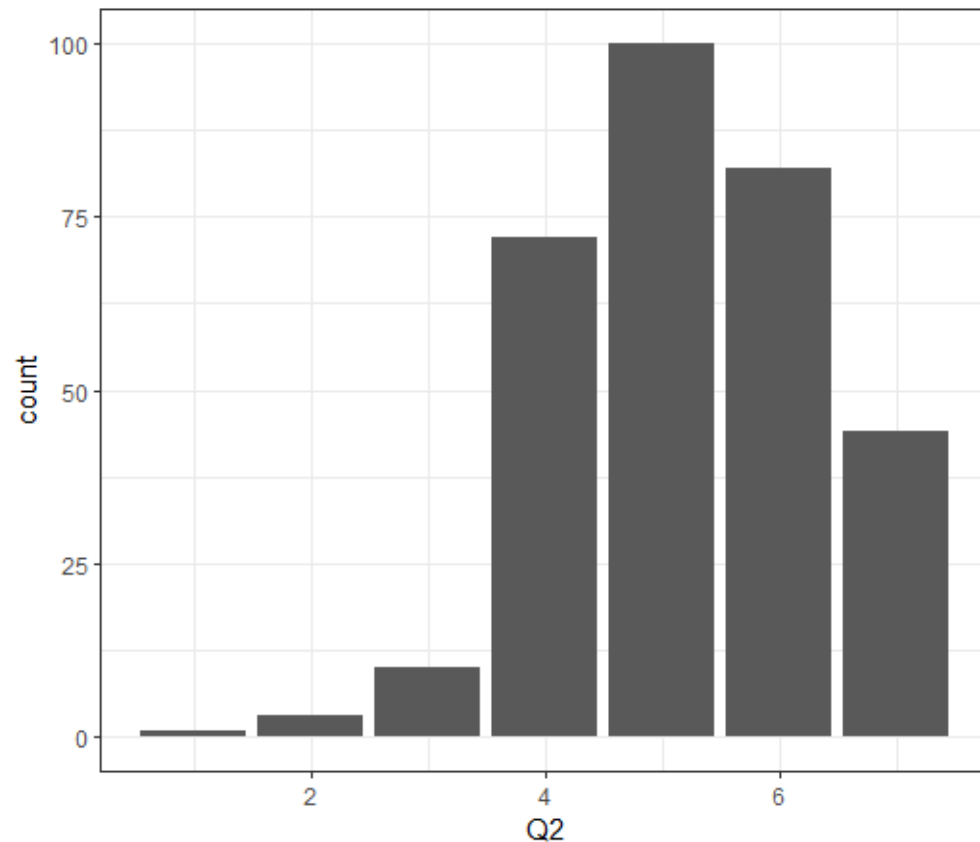




# Biểu diễn quan hệ giữa biến liên tục và gián đoạn



# Thống kê mô tả - biến gián đoạn

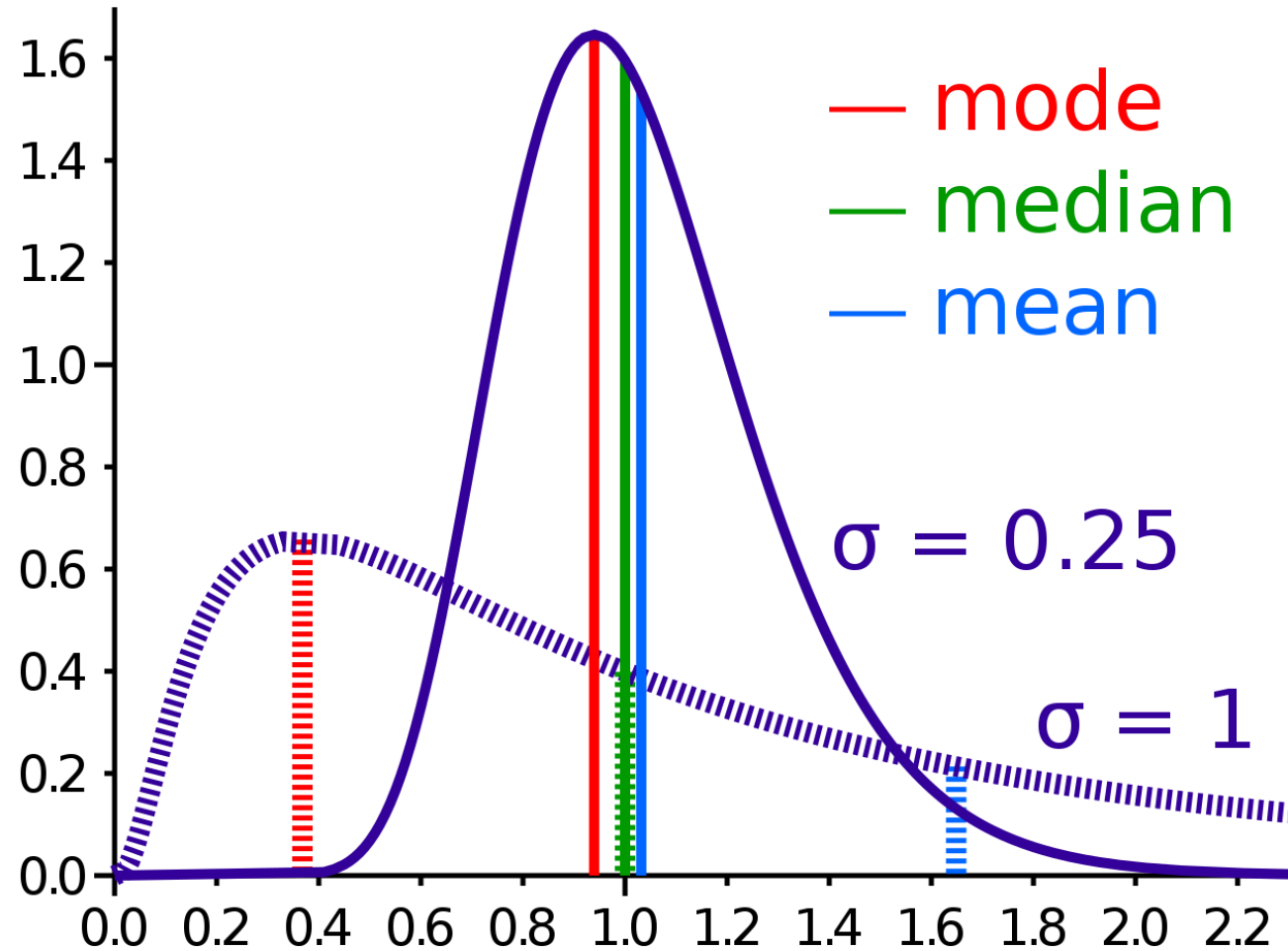


	Q2	count	percent
1	1	1	0.3205128
2	2	3	0.9615385
3	3	10	3.2051282
4	4	72	23.0769231
5	5	100	32.0512821
6	6	82	26.2820513
7	7	44	14.1025641

# Thống kê mô tả - Biến liên tục

Đo độ tập trung

Trung bình, trung vị, mode

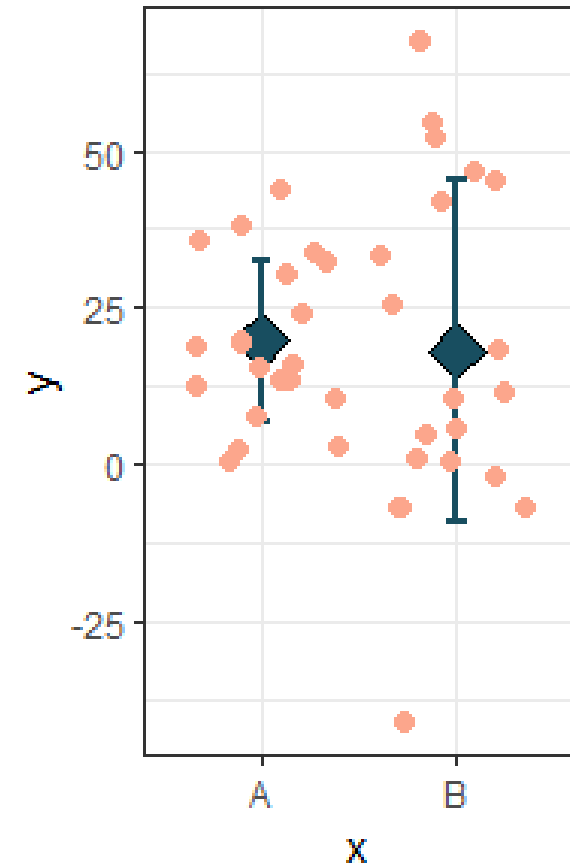
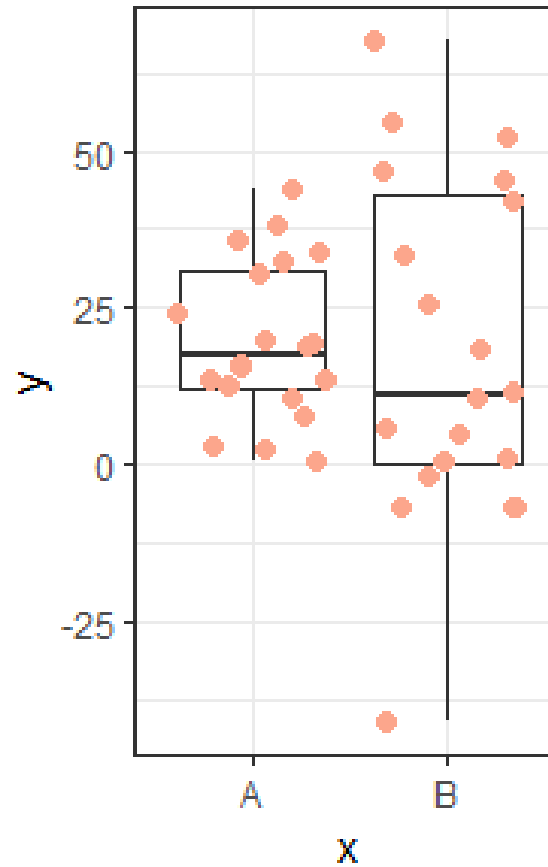


[This Photo](#) by Unknown Author is licensed under [CC BY-SA](#)

# Thống kê mô tả - Biến liên tục

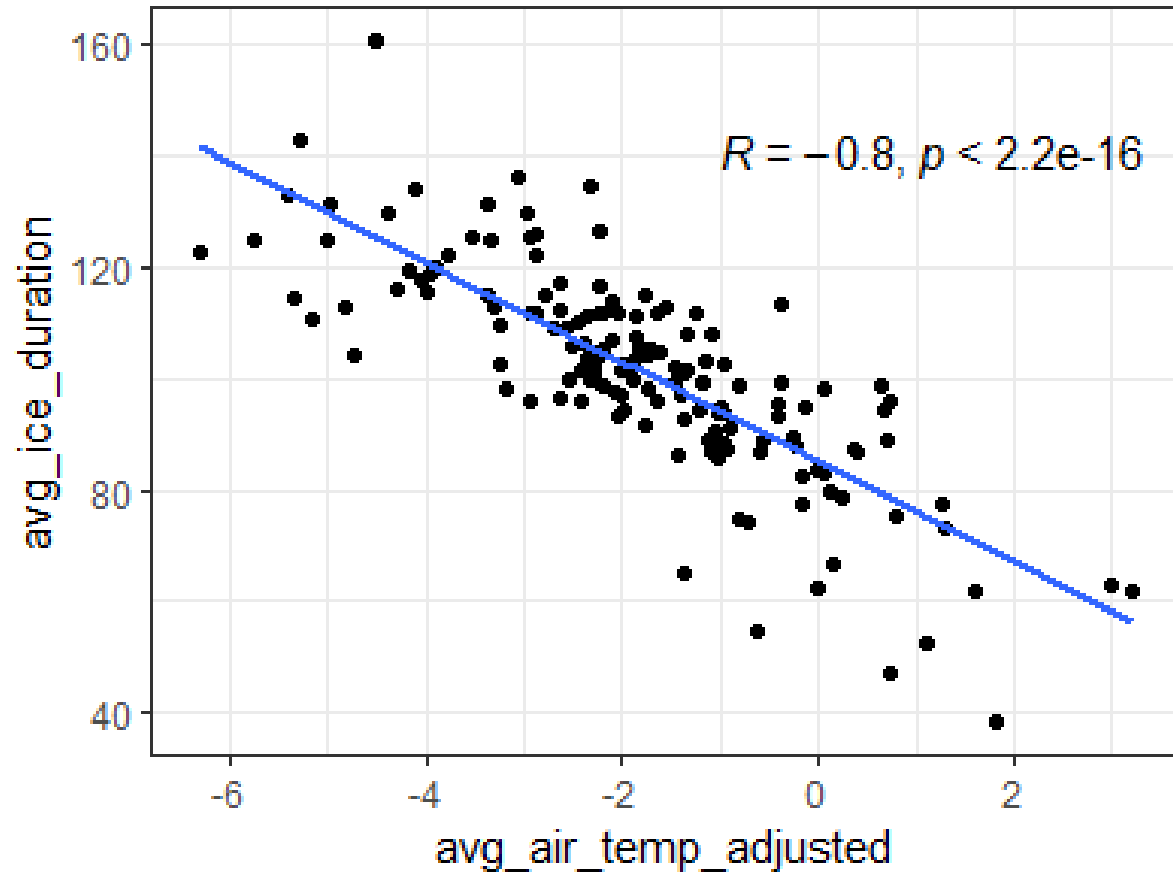
## Đo độ phân tán

- Phương sai (Variance)
- Độ lệch chuẩn (Standard deviation)
- Phân vị (Quantile)
- Điểm tứ phân vị (Quartile)
- Giá trị tối thiểu/tối đa (min/max)

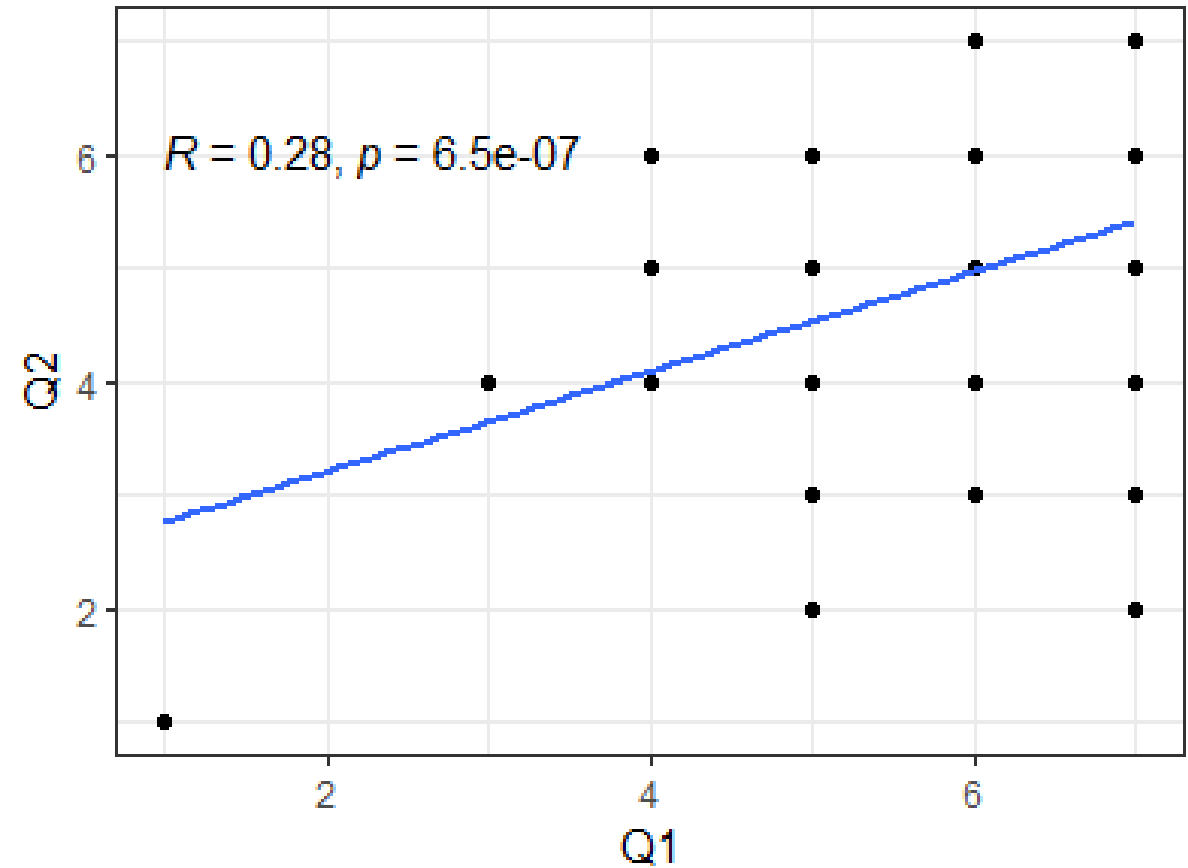


# Thống kê mô tả - tương quan

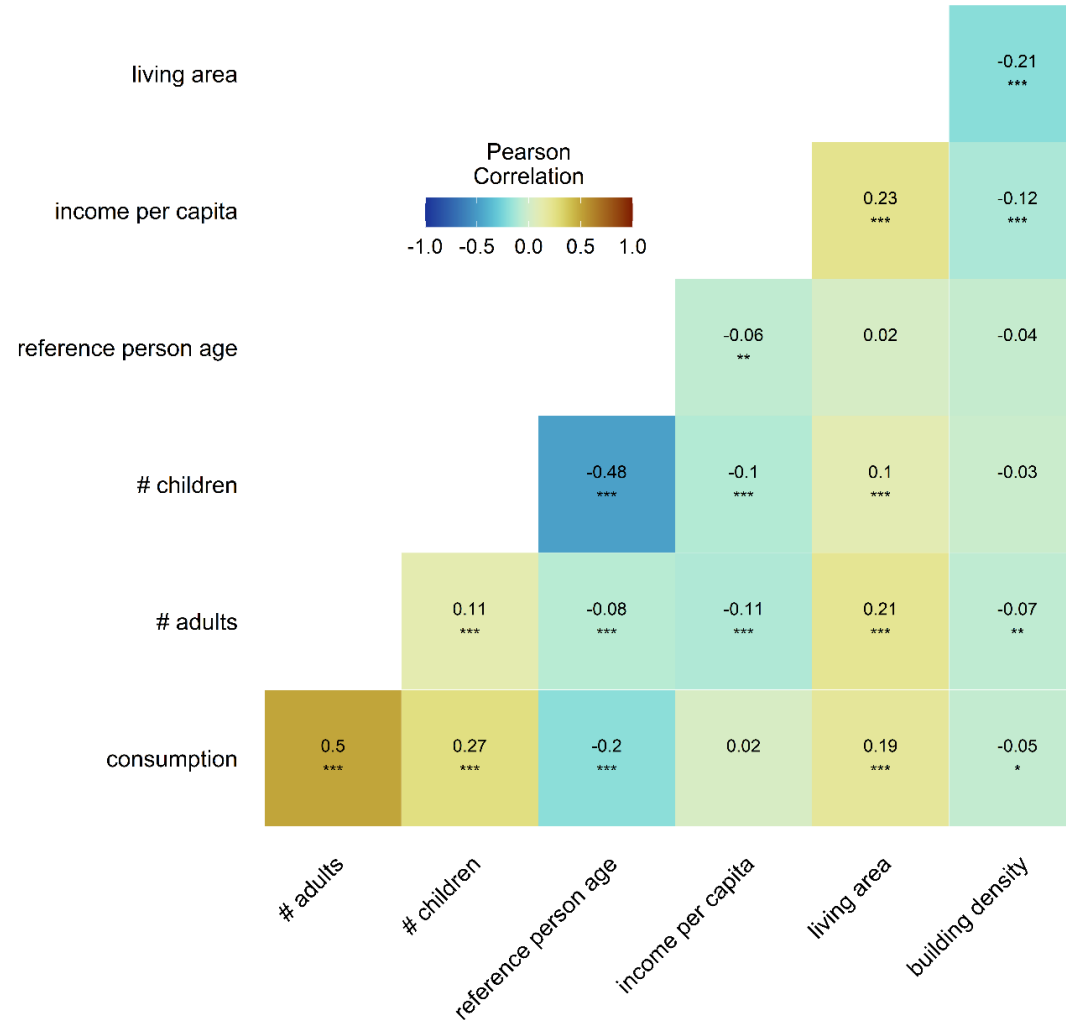
Pearson



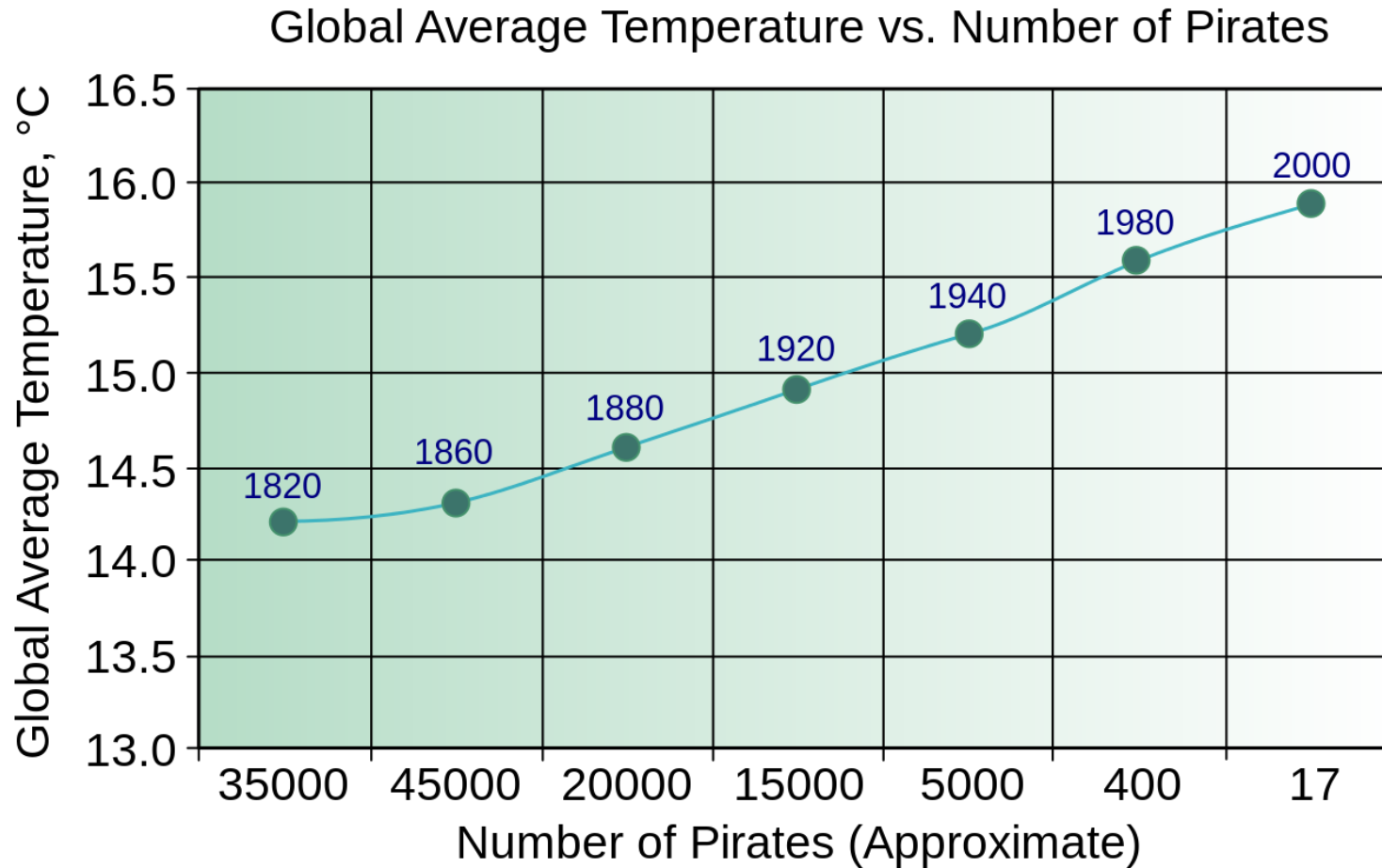
Spearman



# Thống kê mô tả - Tương quan



# Tương quan và quan hệ nhân quả



# Bài tập

- Tìm hiểu và biểu diễn dữ liệu nước sạch tiêu thụ
- Tính toán một vài đại lượng thống kê miêu tả của dữ liệu nước sạch tiêu thụ