

# Gợi ý làm bài tập thực hành

Ngày 01 – Câu hỏi 3

Lê Ngọc Thạch

# Câu hỏi 3

Trong USB kèm theo có file tên là "**test excel file 2.xls**". Đây là một nghiên cứu mà đối tượng (thể hiện qua biến id) được theo dõi theo thời gian, tiếp theo câu hỏi 2. Các bạn hãy đọc file này vào R và gọi đối tượng là "data". Làm các thao tác sau đây:

- Có bao nhiêu bệnh nhân trong file này?
- Mỗi bệnh nhân được theo dõi theo thời gian, và số lần thăm clinic thể hiện qua cột "visit". Đếm xem số bệnh nhân đã thăm clinic 1, 2, 3, 4, 5, 6 lần. Nói cách khác, các bạn hãy điền vào bảng số liệu sau đây:

Số lần visit	Số bệnh nhân	Phần trăm
1		
2		
3		
...		

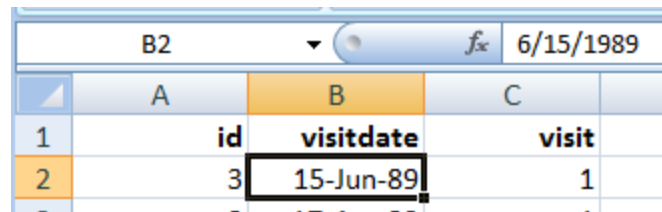
# Tải file “test excel file 2.xls”

- <https://github.com/phantichdulieu/phantichdulieukhoahoc/blob/master/huong-dan-thuc-hanh/DataFiles/test%20excel%20file%202.xls>

# Hướng dẫn

**Bước 1:** Định dạng lại dữ liệu thời gian (ngày tháng năm) trong file Excel trước khi lưu thành dạng CSV.

- Bấm chuột vào ô B2 bạn sẽ thấy giá trị “visitdate” là “15-Jun-89” được hiển thị trên thanh Address “6/15/1989”

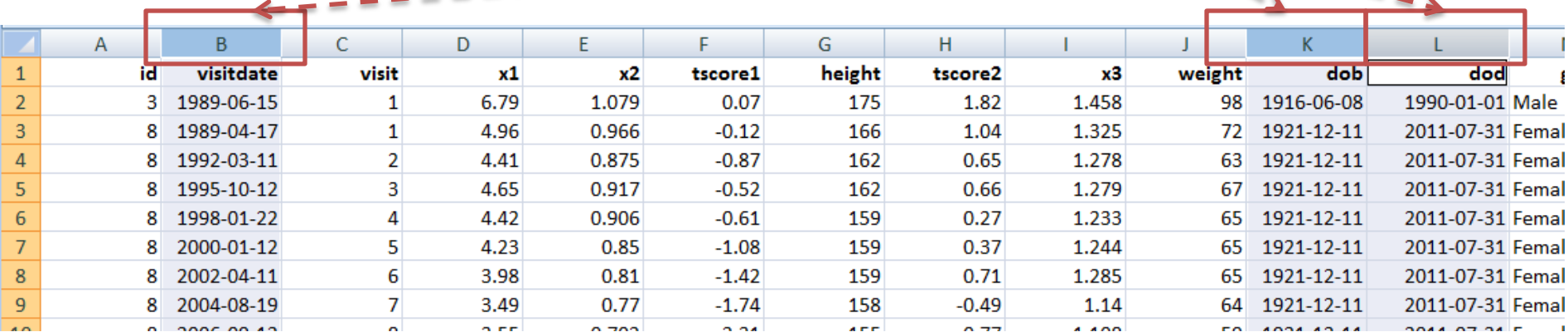


	A	B	C
1	id	visitdate	visit
2	3	15-Jun-89	1

Bạn cần chuyển tất cả các cột ngày (trong bài này là 3 cột “visitdate”, “dob” và “dod”) thành định dạng “**yyyy-mm-dd**”

# Cách định dạng cho cả cột trong Excel

- Một tay nhấn giữ phím Ctrl, một tay bấm chuột vào thanh header của cột B, K, L

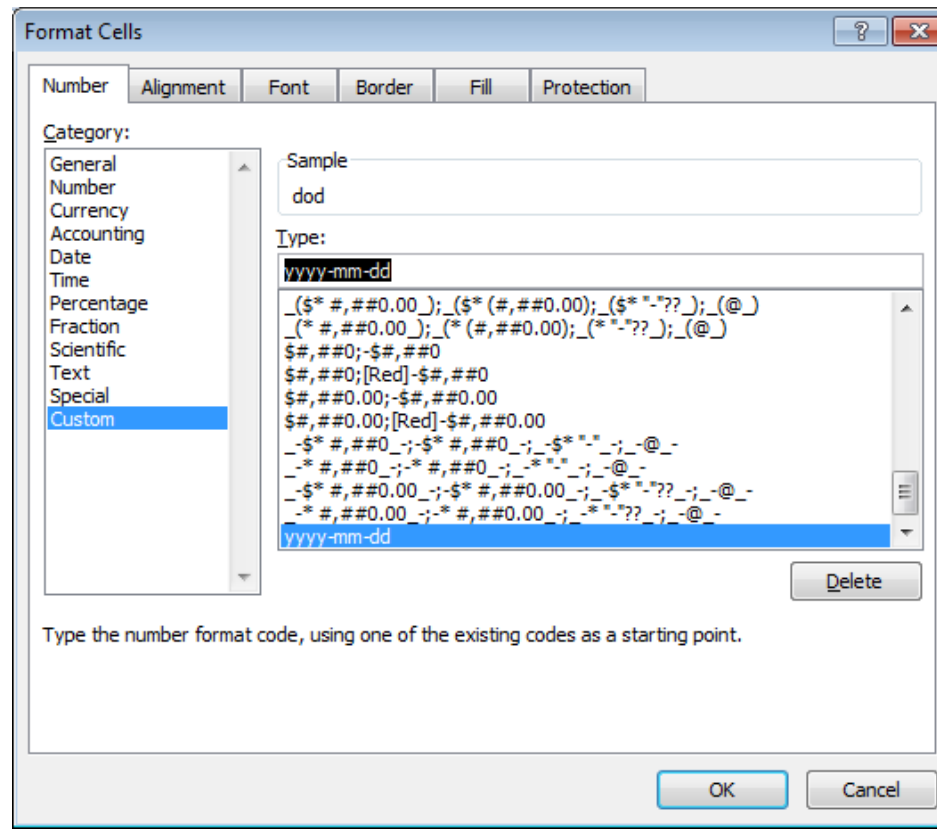


	A	B	C	D	E	F	G	H	I	J	K	L	M
1		id	visitdate	visit	x1	x2	tscore1	height	tscore2	x3	weight	dob	dod
2	3	1989-06-15	1	6.79	1.079	0.07	175	1.82	1.458	98	1916-06-08	1990-01-01	Male
3	8	1989-04-17	1	4.96	0.966	-0.12	166	1.04	1.325	72	1921-12-11	2011-07-31	Femal
4	8	1992-03-11	2	4.41	0.875	-0.87	162	0.65	1.278	63	1921-12-11	2011-07-31	Femal
5	8	1995-10-12	3	4.65	0.917	-0.52	162	0.66	1.279	67	1921-12-11	2011-07-31	Femal
6	8	1998-01-22	4	4.42	0.906	-0.61	159	0.27	1.233	65	1921-12-11	2011-07-31	Femal
7	8	2000-01-12	5	4.23	0.85	-1.08	159	0.37	1.244	65	1921-12-11	2011-07-31	Femal
8	8	2002-04-11	6	3.98	0.81	-1.42	159	0.71	1.285	65	1921-12-11	2011-07-31	Femal
9	8	2004-08-19	7	3.49	0.77	-1.74	158	-0.49	1.14	64	1921-12-11	2011-07-31	Femal

- Bạn thấy 3 cột đổi màu tức là đã chọn thành công. Lúc này có thể thả phím Ctrl và thả chuột.
- Tiếp theo nhấn tổ hợp phím Ctrl+1

# Định dạng yyyy-mm-dd cho ô thời gian

- Trong hộp thoại “Format Cells”, bấm chọn mục “Custom” trong “Category” của tab “Number”.
- Gõ giá trị vào ô “Type”: **yyyy-mm-dd**



# Định dạng yyyy-mm-dd cho ô thời gian

- Sau khi định dạng xong bạn sẽ thấy 3 cột B, K, L có giá trị như bên dưới:

	A	B		K	L
1	id	visitdate		dob	dod
2	3	1989-06-15		1916-06-08	1990-01-01 M
3	8	1989-04-17		1921-12-11	2011-07-31 F
4	8	1992-03-11		1921-12-11	2011-07-31 F
5	8	1995-10-12		1921-12-11	2011-07-31 F
6	8	1998-01-22		1921-12-11	2011-07-31 F

# Hướng dẫn [tt]

## Bước 2: Lưu file dữ liệu dạng Excel thành file CSV.

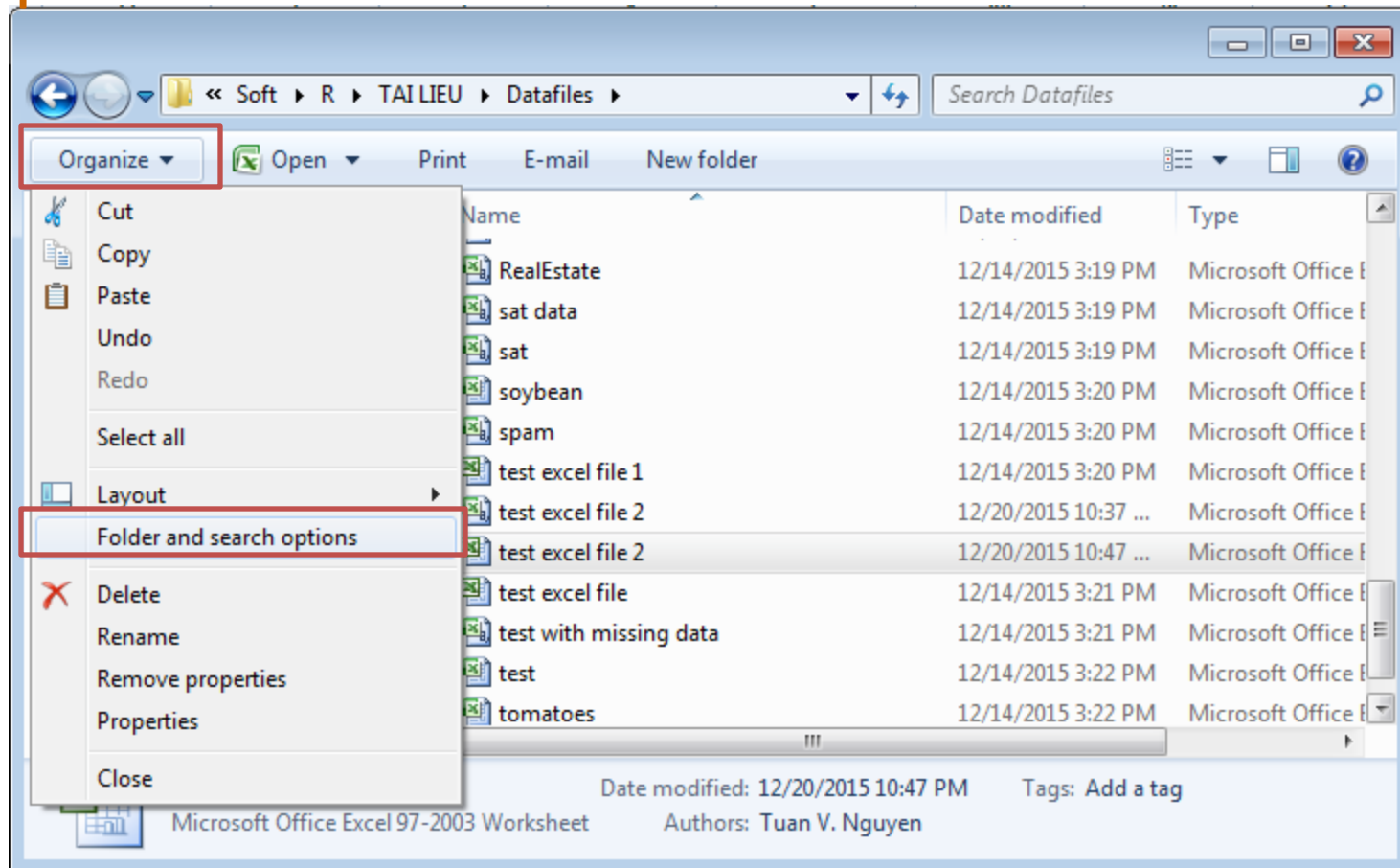
- Mặc định trong cửa sổ Explorer của Windows bạn sẽ không thấy phần mở rộng của file nên rất dễ nhầm lẫn file .csv và .xls nếu không để ý kỹ biểu tượng.
- Tốt nhất là bạn cấu hình để Windows hiển thị tên file đầy đủ bằng cách sau:

Trong cửa sổ Windows Explorer của Windows 7:



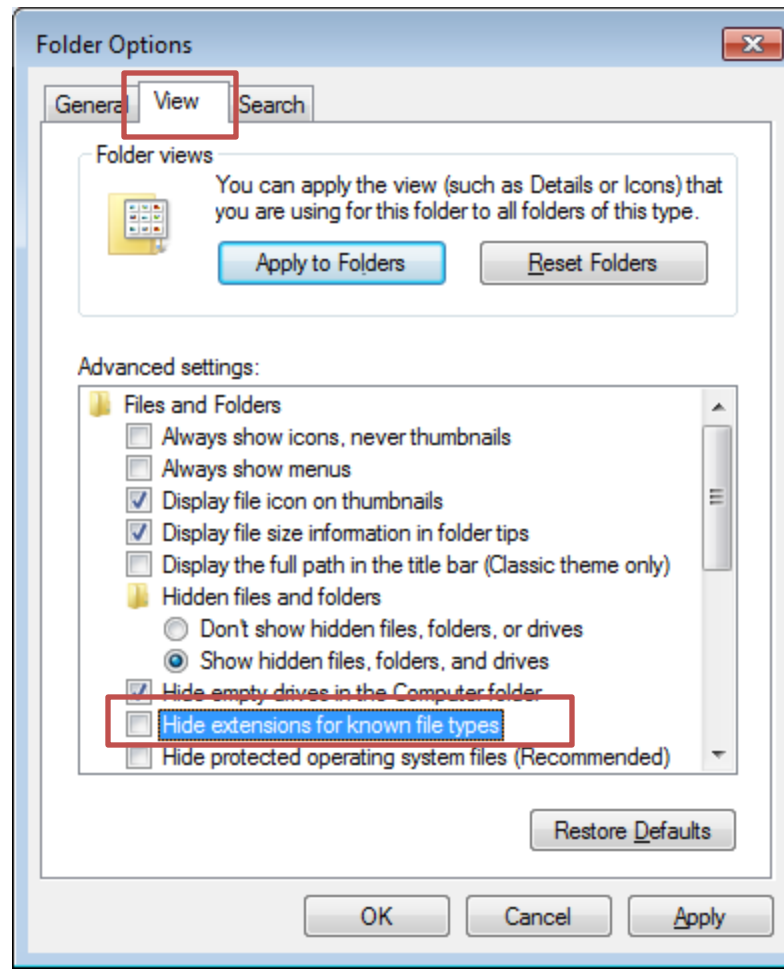
# Cấu hình Windows Explorer để hiển thị tên file đầy đủ

- Bấm chọn mũi tên nhỏ xíu bên phải nút “Organize”. Bấm tiếp mục “Folder and search options”



# Cấu hình Windows Explorer để hiển thị tên file đầy đủ

- Trong tab View, bỏ chọn “Hide extensions for known file types”. Sau đó nhấn nút “OK” để hoàn tất.



# Đọc dữ liệu

## Bước 3: Đọc file CSV vào R:

- Dùng 2 lệnh sau để chọn file và đọc vào đối tượng “data”:

```
> f = file.choose()
```

```
> data = read.csv(f, header = T)
```

- Xem lại tên cột và vài dòng dữ liệu:

```
> head(data)
```

- Xem số dòng và số cột có trong “data”:

```
> dim(data)
```

# Tổng hợp dữ liệu

- Tính số bệnh nhân theo id
  - > nBenhNhan = dim(table(data\$id))
  - > nBenhNhan #3666 bệnh nhân
- Tính tổng số bệnh nhân theo số visit:
  - > sumVisit = table(data\$visit)
- Tính phần trăm số bệnh nhân theo số visit:
  - > percentSumVisit = table(data\$visit) / nBenhNhan \* 100
- Xem số dòng và số cột có trong “data”:
  - > dim(data)
- Chuẩn bị dữ liệu cho bảng kết quả:
  - > nVisit = 1:6
  - > soBenhNhan = sumVisit[c(1:6)]
  - > phanTram = percentSumVisit[c(1:6)]
  - > resultTable = data.frame(nVisit, soBenhNhan, phanTram)