# 1 Dữ liệu và một số thao tác tiền xử lí dữ liệu

**Bài 1.** Giả sử bạn theo dõi số tiền trong hóa đơn diện thoại hàng tháng theo thứ tự từ tháng 1 đến tháng 12 trong năm vừa qua như sau (đơn vị nghìn đồng)

198 185 223 221 207 203 180 195 222 177 214 216

**a.** Nhập dữ liệu thành một vec tơ có tên **TienDT**.

**b.** Tính tổng số tiền bạn phải trả cho phí điện thoại trong năm đó.

**c.** Cho biết tháng nào có số tiền ít (nhiều) nhất, và số tiền là bao nhiêu?

**d.** Cho biết những tháng bạn phải trả hơn 200 nghìn tiền điện thoại. Có bao nhiêu tháng như thế?

**e.** Tính xem có bao nhiêu tháng mà tiền điện thoại không quá 190 nghìn.

**f.** Tính xem có bao nhiêu tháng mà tiền điện thoại dao động trong khoảng [190, 210] nghìn.

**g.** Tính số tiền điện thoại trung bình một tháng (dùng hàm `mean`)..

**Bài 2.** Dùng dữ liệu **TienDT** trên để thực hiện tiếp những yêu cầu sau

**a.** Lưu file dữ liệu này dưới dạng **.rda**

**b.** Sửa tiền điện thoại tháng 2 thành 175.

**c.** Nhập thêm vào dãy trên tiền điện thoại của 3 tháng tiếp theo nhưng bạn quên mất số tiền tháng thứ 13, biết tháng thứ 14, 15 số tiền lần lượt là 201, 185. Sau đó tính lại số tiền trung bình bạn phải trả mỗi tháng.

**Bài 3.** Cho ba tập dữ liệu dạng véc tơ

$$x = c(1, 3, 5, 7, 9), \quad y = c(1, 2, 8, 6, 4, 5, 7), \quad z = c(2, 8, 1, 0, 3)$$

Hãy thực hiện các thao tác sau:

**a.** $z - x$, $x + z$, $x \star z$, $z/x$

**b.** $x + 1$, $y^\star 2$, length(x), length(y), $x + y$

**c.** sum$(x > 5)$ và sum(x$[x > 5]$)

**Bài 4.** Dưới đây là thông tin về 8 sinh viên mới ra trường của một khóa học

| Thứ tự | Lương khởi điểm | Giới tính | Xếp loại tốt nghiệp | Tuổi |
|---|---|---|---|---|
| 1 | 6.0 | Nam | K | 22 |
| 2 | 5.0 | Nu | K | 25 |
| 3 | 4.5 | Nam | TB | 23 |
| 4 | 3.8 | Nu | K | 22 |
| 5 | 8.0 | Nu | G | 22 |
| 6 | 12.0 | Nam | G | 23 |
| 7 | 4.0 | Nam | TB | 22 |
| 8 | 5.0 | Nu | TB | 24 |

**a.** Nhập dữ liệu vào một data frame (đặt tên là **SinhVien**) gồm các cột TT, Luong, GioiTinh, TotNghiep, Tuoi.

**b.** Đưa ra dữ liệu về những sinh viên nữ.

**c.** Đưa ra dữ liệu về những sinh viên nam.

**d.** Đưa ra danh sách lương khởi điểm của nhóm sinh viên nữ.

**e.** Đưa ra danh sách tuổi của nhóm sinh viên nam.

**f.** Đưa ra danh sách những sinh viên có lương khởi điểm trên 6 triệu/tháng.

**g** Cho biết những thông tin về những người có lương cao nhất trong danh sách.

**h.** Thêm vào danh sách một sinh viên nam tốt nghiệp xếp loại giỏi, lương khởi điểm 7.5 triệu nhưng không có thông tin về tuổi.

**i.** Thêm vào data frame trên một cột điểm khóa luận tốt nghiệp của các sinh viên trên theo thứ tự như sau: 8, 7.5, 7, 7, 9, 9.5, 8, 8, 9.

**j.** Cho biết dữ liệu trong từng cột được đo bằng thang đo nào?

**k.** Loại đi số liệu trống không trong data frame. Nhận xét.

**Bài 5.** File dữ liệu **HocSinh.rda** cho ta thông tin của một nhóm học sinh về giới tính (GioiTinh), lớp (Lop), tuổi (Tuoi), nơi ở (NoiO), học lực (HocLuc), mức độ yêu thích thể thao (TheThao) và đánh giá hình thức (HinhThuc). Hãy lấy file dữ liệu và thực hiện các yêu cầu sau:

**a.** Cho biết dữ liệu trong từng cột được đo bằng thang đo nào?

**b.** Cho biết tập dữ liệu này có bao nhiêu cột, bao nhiêu dòng?

**c.** Lấy ra dữ liệu ở cột thứ 3 (Tuoi).

**d.** Lấy ra toàn bộ dữ liệu ở dòng thứ 10.

**e.** Tính số học sinh nữ, nam.

**f.** Tính tỷ lệ học sinh có mức độ yêu thích thể thao là 4. Tính tỷ lệ này trong nhóm học sinh nam, trong nhóm học sinh nữ.

**Bài 6.** Trong file **HocSinh.rda**, hãy chọn ngẫu nhiên 50 học sinh và lấy ra tất cả thông tin về 50 người đó. Lưu dữ liệu dưới dạng file **.rda**.

**Bài 7. a.** Tạo dãy số từ 1 đến 100.

**b.** Tạo dãy số chẵn từ 0 đến 100.

**c.** Tạo dãy số trong đó 3 lặp 4 lần, 5 lặp 10 lần, 16 lặp 7 lần.

**d.** Tạo dãy số trong đó có các giá trị 1, 2, 3, 4 lần lượt lặp lại 10 lần.

**e.** Tạo dãy số mà cả cụm 1, 2, 3 lặp lại 8 lần.

**f.** Tạo biến thứ bậc gồm 3 bậc, mỗi bậc lặp 4 lần.

**g.** Tạo biến thứ bậc gồm 4 bậc, mỗi bậc lặp 4 lần, chiều dài biến bằng 15.

**h.** Tạo biến thứ bậc gồm 3 bậc, số lần lặp lại tương ứng là 2, 5, 8 với ký hiệu a, b, c.

**Bài 8.** Xác định loại thang đo trong các trường hợp sau

**a.** Thời gian chờ thang máy của một người tại một khu chung cư.

**b.** Số khối nước một gia đình sử dụng trong một tháng.

**c.** Xếp hạng 5 chiếc máy trong nhà máy theo đánh giá: rất tốt, tốt, trung bình, kém.

**d.** Mã vùng điện thoại của các địa phương.

**e.** Tuổi của các nhân viên trong công ty.

**f.** Doanh thu (VN đồng) của một cửa hàng bán báo trong một tháng

**g.** Mã sinh viên trong một trường đại học.

**h.** Điểm thi một môn của sinh viên một lớp.

**i.** Chiều cao của một người.

**Bài 9.** Chọn ngẫu nhiên 5 người từ danh sách gồm 40 người.

**Bài 10.** Tung một đồng xu 50 lần. Mô phỏng phép thử và đếm số mặt sấp.

**Bài 11.** Tung một con xúc xắc 100 lần. Mô phỏng phép thử và đếm số lần xuất hiện mặt 6 chấm.

**Bài 12.** Chọn ngẫu nhiên năm cây bài từ bộ bài tú lơ khơ. Mô phỏng phép thử và kiểm tra xem có bộ đôi nào trong mẫu không? Hãy lặp lại cho đến khi có được một đôi trong 5 cây bài.

# 2 Tóm tắt dữ liệu

**Bài 13.** Trong file dữ liệu có tên là **SoLieu.csv** chứa một số thông tin cá nhân của 100 người về giới tính (GioiTinh), tuổi (Tuoi), khu vực sống (KhuVuc) và tổng thu nhập (đơn vị triệu VND) trong năm 2008 (ThuNhap). Hãy lấy file dữ liệu và thực hiện các yêu cầu sau:

**a.** Trong nhóm được điều tra có bao nhiêu nam và có bao nhiêu người sống ở thành phố.

**b.** Tính số nam sống ở hải đảo và nữ sống ở nông thôn trong nhóm những người được điều tra.

**c.** Trong số nữ được điều tra, hãy tính tỉ lệ nữ sống ở thành phố và miền núi.

**d.** Tiến hành phân tổ cột dữ liệu về tuổi thành các tổ với các điểm chia là $20, 30, 40, 50, 60, 70, 80$ và tính tỉ lệ những người được điều tra có độ tuổi không vượt quá 50.

**e.** Tiến hành phân tổ cột dữ liệu về thu nhập thành các tổ với các điểm chia là $20, 40, 60, 80, 100$ và tính:

  **i.** tỉ lệ những người phải đóng thuế thu nhập nếu biết một người phải đóng thuế thu nhập nếu tổng thu nhập trong năm của người đó vượt quá 60 triệu VND.