**Bài tập tuần 7**

**Bài 1:** Dữ liệu (dulieu2.csv) đã cho là thông tin về doanh số bán hàng của một công ty qua các năm.

Thực hiện các thao tác sau với phần mềm R.

a. Nhập dữ liệu (“dulieu2.csv”) vào R và loại bỏ dữ liệu trống.

b. Trích ra một dữ liệu có bộ phận (Segment) là “Government”, đặt nó là data1 và “Midmarket” đặt tên nó là data2.

c. Hãy sử dụng các câu lệnh if/for để tính tổng doanh thu (Gross.sales) của từng loại mặt hàng (Product). Hãy cho biết sản phẩm nào có doanh thu lớn nhất?

d. Kiểm tra sản phẩm “VTT” được bán tại “Mexico” vào tháng và năm nào?

e. Hãy cho biết sản phẩm nào đã bán được nhiều nhất?

f. Hãy so sánh tổng doanh thu của hai bộ phận Government và midmarket.

**Bài 2:**

Từ dữ liệu “WHO1”, ta có mức thu nhập quốc dân (GNI) của 195 quốc gia. Với những quốc gia có GNI lớn hơn 20 nghìn ta xét vào nhóm quốc gia phát triển, những quốc gia có GNI nằm trong khoảng từ 10 – 20 nghìn được xếp vào nhóm quốc gia đang phát triển, các nước có GNI nhỏ hơn 10000 được xếp vào nhóm các quốc gia chưa phát triển, còn lại là các quốc gia khác. Sử dụng R và làm các câu hỏi sau:

a. Nhập dữ liệu “WHO1” vào R và đặt tên nó là who

b. Loại bỏ các dữ liệu trống trong và đặt tên nó là who1.

c. Thực hiện bổ sung dữ liệu trống trong dữ liệu who bằng cách thêm giá trị trung bình cộng của giá trị lớn nhất và giá trị nhỏ nhất.

d. Hãy tính thu nhập quốc dân trung bình của hai dữ liệu who và who1, và đưa ra nhận xét.

e. Từ dữ liệu trích ra dữ liệu của hai nước “Bahrain” và “Mexico” và so sánh tuổi thọ của hai nước này.

f. Hãy cho biết tỷ lệ sinh sản của nước “Canada”.

g. Viết thêm cột phân nhóm các quốc gia vào dữ liệu. Hãy cho biết nước “China” thuộc nhóm nào?

h. Hãy tìm nước có dân số cao nhất.

**Bài 3:** Thực hiện lần lượt các thao tác sau với R

1. Sinh ra một bộ dữ liệu về thông tin của 300 sinh viên: bao gồm ID, giới tính, tuổi (từ 18-24), chiều cao, cân nặng, điểm thường xuyên, điểm giữa kỳ, điểm cuối kỳ.
2. Lấy từ dữ liệu trên danh sách sinh viên của điểm thi cuối kỳ lớn hơn hoặc bằng 9.
3. Bạn sinh viên có ID=5 có điểm chuyên cần bị sai, phải sửa lại là 10. Thực hiện thao tác đó với R
4. Thêm một cột tên là “tổng điểm” được tính theo công thức tổng= điểm thường xuyên\*0.2+điểm giữa kỳ\*0.2+điểm cuối kỳ\*0.6.
5. Nếu tổng điểm lớn hơn hoặc bằng 9.0 được xếp điểm A+, 8.5<= điểm tổng < 9 được xếp điểm A, 8.0<= điểm tổng < 8.5 được xếp điểm B+, 7.0<= điểm tổng <8.0 được xếp điểm B, điểm tổng < 7.0 được xếp điểm C. Hãy kiểm tra xem bạn có ID=10 được điểm chữ là bao nhiêu?
6. Hãy đưa ra điểm chữ cho các bạn sinh viên có điểm thi cuối kỳ lớn hơn hoặc bằng 9.0.
7. Có bao nhiêu bạn sinh viên có điểm giữa kì là 0 nhưng lại qua môn?
8. Hãy tính tỉ lệ phần trăm của các bạn được điểm “B”
9. Tỉ lệ này sẽ thay đổi như nào nếu ta cộng thêm mỗi sinh viên 2đ giữa kỳ. (Bạn nào quá 10 điểm thì chỉ đạt 10đ). Câu hỏi này tương tự với ý g)

**Bài 4:** Thực hiện các mô phỏng sinh ngẫu nhiên dưới đây với phần mềm R. Một nhà đầu tư muốn mô tả thị trường nhà ở tại Hà nội, ông ta yêu cầu R sinh ra dữ liệu với các thông tin sau:

+ **crim** - tỷ lệ tội phạm bình quân đầu người theo thị trấn: là một số liên tục có phân phối chuẩn với trung bình là 3.613 và độ lệch chuẩn là 8.6

+ **indus** - tỷ lệ diện tích đất kinh doanh phi bán lẻ trên mỗi thị trấn: là một số liên tục có phân phối chuẩn với trung bình là 11.14 và độ lệch chuẩn là 6.86

+ **rm** - số lượng phòng trung bình trên một ngôi nhà: là một số bất kì từ 2-6.

+ **tax** - thuế suất tài sản toàn giá trị trên 10,000$: là một số liên tục có phân phối chuẩn với trung bình là 408 và độ lệch chuẩn là 167

+ **nox** - nồng độ nitric oxide (phần trên 10 triệu - ppm): là một số liên tục từ 0.38 đến 0.87 và có phân phối chuẩn với trung bình là 0.55 và độ lệch chuẩn là 0.116. Hint: Nếu giá trị sinh ra lớn hơn 0.87 hoặc nhỏ hơn 0.8 thì yêu cầu máy tính sinh lại những dữ liệu đó.

+ **chas** - biến giần sông Charles: 1 nếu là gần sông với xác suất là 1/5; 0 nếu ngược lại.

+ **age** - tỷ lệ các căn hộ có chủ sở hữu được xây dựng trước năm 1940: là một số liên tục có phân phối chuẩn với trung bình là 68 và độ lệch chuẩn là 28. Lưu ý: tỷ lệ này chỉ nằm trong khoảng 0-100.

+ **medv** - giá trị của những ngôi nhà có chủ sở hữu tính bằng 1,000$: là một số liên tục có phân phối chuẩn với trung bình là 22.53 và độ lệch chuẩn là 9.2.

a. Thực hiện sinh ra dữ liệu và đặt tên nó là hanoi.

b. Một mô hình thống kê cho thấy:

medv = - 0.138108\*crim - 0.068990\*indus - 4.716034\*nox + 7.677527\*rm - 0.007923\*tax - 18.599942 .

Hãy tạo ra biến medv1 bằng công thức trên, sau đó hãy tính độ lệch tiêu chuẩn cho hiệu số giữa medv1 và medv trong Boston.

c. Hãy so sánh giá trị trung bình của những ngôi nhà gần sông và những ngôi nhà không gần sông.

**Bài 5:** Bộ dữ liệu "COVID19.csv" chứa thông tin liên quan đến các bệnh nhân COVID-19. Dữ liệu gồm 7 biến sau: Số thức tự (ID), bệnh nhân (Patient), tuổi (Age), giới tính (Sex), địa chỉ (Location), tình trang bệnh (Status) và quốc tịch (Nationality). Thực hiện các thao tác sau với phần mềm R.

a) Nhập dữ liệu “COVID19.csv” vào R và viết hàm tính tuổi trung bình của bệnh nhân theo giới tính.

b) Viết một hàm để tính tỷ lệ bệnh nhân đã khỏi bệnh trên tổng số bệnh nhân mắc bệnh ở các địa điểm. Từ đó, giữa hai thành phố Hà Nội và Hồ Chí Minh thì thành phố nào có tỷ lệ bệnh nhân được chữa khỏi bệnh cao hơn.

c) Có bao nhiêu bệnh nhân quốc tịch nước ngoài đã được chữa khỏi bệnh.

d) Tính toán tỷ lệ bệnh nhân khỏi bệnh, đang điều trị, và đã tử vong trong từng nhóm tuổi khác nhau (ví dụ: dưới 18, 18-40, 41-60, trên 60). Nhóm tuổi nào có tỷ lệ hồi phục cao nhất.

**Bài 6:** Cho dữ liệu về thông tin bán hàng của một công ty. Dữ liệu gồm 8 biến sau: ngày đặt hàng (OrderDate), khu vực (Region), thành phố (City), loại sản phẩm (Category), tên sản phẩm (Product), số lượng sản phẩm bán ra (Quantity), đơn giá (UnitPrice) và doanh thu (TotalPrice). Dữ liệu được cho trong file Excel **Product**.

a) Nhập dữ liệu vào R và đặt tên là data. Loại bỏ các bản ghi có giá trị trống.

b) Xác định tổng số sản phẩm “Chocolate” được bán ra tại thành phố San Diego.

c) Trích ra một bộ dữ liệu có loại sản phẩm là “Bars”, đặt tên là data1 và một bộ dữ liệu có loại sản phẩm là “Crackers”, đặt tên là data2. So sánh doanh thu trung bình trong ngày của hai loại sản phẩm trên. Nếu doanh thu trung bình trong ngày của “Bars” lớn hơn/ nhỏ hơn/ bằng doanh thu trung bình trong ngày của “Crackers” thì in ra “Doanh thu trung bình trong ngày của “Bars” lớn hơn/ nhỏ hơn/ bằng doanh thu trung bình trong ngày của “Crackers”.

d) Viết một hàm trong R để kiểm tra tổng số lượng bán ra của một sản phẩm tại thành phố New York. Áp dụng hàm để kiểm tra với sản phẩm “Carrot”.

e) Viết một hàm để đưa ra tên thành phố có tổng số lượng sản phẩm thuộc một loại nhất định bán ra nhiều nhất. Áp dụng hàm với loại sản phẩm “Cookies”.

f) Có một sản phẩm mà nhân viên đã nhập sai giá. Tìm thông tin của sản phẩm đó bằng cách so sánh tổng giá được tính theo công thức:

TotalPrice=Quantity×UnitPrice

với tổng giá có trong dữ liệu. Hãy sửa lại giá đúng, biết rằng mỗi sản phẩm có đơn giá (UnitPrice) khác nhau.