ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC & KỸ THUẬT MÁY TÍNH

CẤU TRÚC RỜI RẠC CHO KHMT (CO1007)

Thống kê khảo sát kết quả Covid-19 môn Cấu trúc rời rạc

GVHD: Huỳnh Tường Nguyên
Nguyễn Ngọc Lễ
SV thực hiện: Nguyễn Văn A – 22102134
Trần Văn B – 88471475
Lê Thị C – 36811334
Phạm Ngọc D – 97501334
Kiều Thị E – 12341334

Tp. Hồ Chí Minh, Tháng 09/2021

Mục lục

1 Động cơ nghiên cứu

Bệnh Corona do virus gây ra còn gọi là COVID-19 đã tạo ra những tác động tiêu cực đến nền đời sống của cư dân trên thể giới. Các đợt bùng phát của COVID-19 hay những biến thể virus đã mang đến những thách thức chưa từng có và được dự báo sẽ có tác động đáng kể đến sự phát triển kinh tế. Nhiều thông tin, tin tức về tình hình dịch bệnh cũng như dữ liệu về COVID-19 được phổ biến rộng rải trong đời sống hay trên internet để giúp cho mọi người quan sát, phân tích, nghiên cứu đươc cập nhật hàng ngày.

Phân tích & thống kê dữ liệu về COVID-19 giúp cho ta thấy được số ca nhiễm bệnh, tử vong của một quốc gia, so sánh tình trạng của các quốc gia trong khu vực hay diễn biến dịch trên thế giới. Từ số liệu được báo cáo mơi chúng ta muốn biết các ca nhiễm bệnh có xu hướng tăng lên hay giảm xuống quy mô các đợt bùng phát ở mỗi quốc gia. Dữ liệu dùng cho bài tập lớn có tham khào từ nguồn có thể xử lý trước với một vài thống kê cơ bản trước khi nó được truyền đi để khai thác dữ liệu thông minh sâu hơn.

2 Mục tiêu

Trong bài tập lớn này, các sinh viên sẽ bắt đầu với các bài toán thống kê đơn giản từ những dữ liệu được cung cấp. Qua đó, các em sẽ tìm ra những con số thú vị, có ý nghĩa đối với các dữ liệu thực tế từ tình hình dịch corona. Những kết quả mà các em tìm ra sẽ là bước khởi đầu cho việc khai phá nguồn dữ liệu của hệ thống sau này, nhằm đạt tới mục tiêu nâng cao kỹ năng lập trình, kỹ năng giải quyết vấn đề cho người học, kỹ năng làm việc nhóm cũng như hướng tới mục tiêu cao hơn là đam mê trong làm việc, học tập và nghiên cứu.

3 Mô tả dữ liệu

Dữ liệu gồm các thuộc tính chính **"iso_code, continent, location, date, new_cases, new_deaths"** được lưu tron file **csv**.

1. *iso_code*: Định danh đất nước
2. *continent* Tên châu lục
3. *location*: Tên quốc gia
4. *date*: Ngày quan sát với định dạng Month-Day-Year
5. *new_cases*: Số trường hợp COVID-19 mới được xác nhận
6. *new_deaths*: Số tử vong mới do COVID-19

4 Nhiệm vụ

Gọi MD là mã đề riêng cho mỗi nhóm (gồm 4 ký số) không trùng nhau, nhóm sinh viên sẽ thực hiện các yêu cầu dưới đây với các giá trị xác định như sau:

- Mỗi nhóm sẽ dùng R để thao tác trên số file dữ liệu khác nhau được chọn theo cột "STT" theo cách tính $kq = (kytu1 + kytu2 + kytu3 + kytu4)\%6$:
  - Nếu $kq = 0$ thì làm các stt là 1,2,3
  - Nếu $kq = 1$ thì làm các stt là 4,5,6
  - Nếu $kq = 2$ thì làm các stt là 7,8,9
  - Nếu $kq = 3$ thì làm các stt là 10,11,12
  - Nếu $kq = 4$ thì làm các stt là 13,14,15
  - Nếu $kq = 5$ thì làm các stt là 16,17,18

STT	đất nước	STT	đất nước
1	Kenya	10	Canada
2	Lesotho	11	Greenland
3	Morocco	12	United States
4	Indonesia	13	Australia
5	Japan	14	New Caledonia
6	Vietnam	15	New Zealand
7	Andorra	16	Brazil
8	Slovenia	17	Chile
9	United Kingdom	18	Venezuela

Hoàn thành các bài tập:

- Đối với các bài tập chung gồm các phần $\{i, ..., ix\}$, tất cả các nhóm đều phải làm
- Mỗi nhóm sẽ thực hiện 4 câu riêng của mình trong phần x bằng các lấy 4 ký số trong mã đề của mình là chỉ số câu hỏi tương ứng. Nếu ký số là 0 thì làm câu 10.

i) Nhóm câu hỏi liên quan đến tổng quát dữ liệu
Dùng tập dữ liệu để trả lời các câu hỏi và trình bày theo đinh dạng

1) Tập mẫu thể hiện thu thập dữ liệu vào các năm nào
2) Số lượng đất nước và định danh của mỗi đất nước (hiển thị 10 đất nước đầu tiên).

iso_code:	Country
AFG	Afghanistan
OWID_AFR	Africa
ALB	Albania
Count	Số đất nước

3) Số lượng châu lục trong tập mẫu

Continent :	Số châu lục
Africa:	Châu phi
Asia:	Châu Á

4) Số lượng dữ liệu thể hiện thu thập dữ liệu được trong từng từng châu lục và tổng số

Continent:	Observations
Africa	value1
Asia	value2
Tổng:	giá trị tổng

5) Số lượng dữ liệu thể hiện thu thập dữ liệu được trong từng từng đất nước (hiển thị 10 dất nước cuối cùng) và tổng số

iso_code	Observations
AFG	value1
OWID_AFR	value2
ALB	value3
Tổng:	giá trị tổng

6) Cho biết các châu lục nào có lượng dữ liệu thu thập nhỏ nhất và giá trị nhó nhất đó?
7) Cho biết các châu lục nào có lượng dữ liệu thu thập lớn nhất và giá trị lớn nhất đó?
8) Cho biết các nước nào có lượng dữ liệu thu thập nhỏ nhất và giá trị nhó nhất đó?
9) Cho biết các nước nào có lượng dữ liệu thu thập lớn nhất và giá trị lớn nhất đó?
10) Cho biết các date nào có lượng dữ liệu thu thập nhỏ nhất và giá trị nhó nhất đó?
11) Cho biết các date nào có lượng dữ liệu thu thập lớn nhất và giá trị lớn nhất đó?
12) Cho biết số lượng dữ liệu thu thập được theo date và châu lục.

13) Cho biết số lượng dữ liệu thu thập được là lớn nhất theo date và châu lục.
14) Cho biết số lượng dữ liệu thu thập được là nhỏ nhất theo date và châu lục.
15) Với một date là k và châu lục t cho trước, hãy cho biết số lượng dữ liệu thể hiện thu thập dữ liệu được.
16) Có đất nước nào mà số lượng dữ liệu thu thập được là bằng nhau không? Hãy cho biết các iso_code của đất nước đó.
17) Liệt kê iso_code, tên đất nước mà chiều dài iso_code lớn hơn 3.

ii) Nhóm câu hỏi liên quan đến mô tả thống kê cơ bản dữ liệu
Với mỗi quốc gia mà thuộc về nhóm cần tính số liệu thống kê lần lượt cho nhiễm và tử vong do coronavirus được báo cáo mới:

1) Tính giá trị nhỏ nhất, lớn nhất
2) Tính tứ phân vị thứ nhất(Q1), thứ hai(Q2), thứ ba(Q3)
3) Tính giá trị trung bình (Avg)
4) Tính giá trị độ lệch chuẩn (Std)
5) Đếm xem có bao nhiêu outliers, một quan sát mà giá trị của nó nằm trong khoảng sau:
$IQR = Q3 - Q1$
$outliers < Q1 - 1.5 * IQR$ hoặc $outliers > Q3 + 1.5 * IQR$
6) Lập bảng mô tả số liệu thống kê cho từng đất nước thuộc về nhóm:

Countries	Min	Q1	Q2	Q3	Max	Avg	Std	Outlier
ctr_i	?	?	?	?	?	?	?	?

7) Vẽ biểu đồ boxplot hay còn được gọi là box-and-whisker cho nhiễm coronavirus

iii) Nhóm câu hỏi liên quan đến dữ liệu thể hiện thu thập dữ liệu
Với mỗi quốc gia mà thuộc về nhóm cần tính số liệu thống kê lần lượt cho nhiễm và tử vong do coronavirus:

1) Có bao nhiêu ngày có số lần dữ liệu không được báo cáo mới.
2) Có bao nhiêu ngày có số ca nhiễm/ tử vong là thấp nhất được báo cáo mới.
3) Có bao nhiêu ngày có số ca nhiễm/ tử vong là cao nhất được báo cáo mới
4) Thể hiện bảng số liệu như sau:
Không được báo cáo mới:

Countries	Infections	Deaths
ctr_i	value	value

Báo cáo mới:

Countries	Infections	Deaths
ctr_i	value	value

5) Cho biết số ngày ngắn nhất liên tiếp mà không có dữ liệu được báo cáo
6) Cho biết số ngày dài nhất liên tiếp mà không có dữ liệu được báo cáo
7) Cho biết số ngày ngắn nhất liên tiếp mà không có người nhiễm bệnh mới
8) Cho biết số ngày dài nhất liên tiếp mà không có người nhiễm bệnh mới

iv) Nhóm câu hỏi liên quan đến trực quan dữ liệu

1 Vẽ biểu đồ tần số tích lũy quốc gia cho các châu lục
2 Vẽ biểu đồ tần số tương đối quốc gia cho các châu lục
3 Vẽ biểu đồ thể hiện nhiễm bệnh đã báo cáo của các quốc gia mà thuộc về nhóm trong 7 ngày cuối của năm cuối cùng

4 Vẽ biểu đồ thể hiện tử vong đã báo cáo của các quốc gia mà thuộc về nhóm trong 7 ngày cuối của năm cuối cùng

5 Vẽ biểu đồ phổ đất nước xuất hiện outliers cho nhiễm bệnh

6 Vẽ biểu đồ phổ đất nước xuất hiện outliers cho tử vong

v) Nhóm câu hỏi liên quan đến trực quan dữ liệu theo thời gian là tháng
Với mỗi quốc gia mà thuộc về nhóm, trên từng năm hãy vẽ biểu đồ thể hiện trục Ox là thời gian, trục Oy là nhiễm bệnh/tử vong. Hãy dùng 4 ký số của mã đề để vẽ 4 tháng tương ứng theo ký số đó. Nếu ký số là 0 thì lấy tháng là 10.

1 Biểu đồ thể hiện thu thập dữ liệu nhiễm bệnh cho từng tháng

2 Biểu đồ thể hiện thu thập dữ liệu tử vong cho từng tháng

3 Biểu đồ thể hiện thu thập dữ liệu gồm nhiễm bệnh và tử vong cho từng tháng

4 Biểu đồ thể hiện thu thập dữ liệu nhiễm bệnh gồm 2 tháng cuối của năm

5 Biểu đồ thể hiện thu thập dữ liệu tử vong gồm 2 tháng cuối của năm

6 Biểu đồ thể hiện thu thập dữ liệu gồm nhiễm bệnh và tử vong gồm 2 tháng cuối của năm

7 Biểu đồ thể hiện thu thập dữ liệu nhiễm bệnh tích lũy cho từng tháng

8 Biểu đồ thể hiện thu thập dữ liệu tử vong tích lũy cho từng tháng

vi) Nhóm câu hỏi liên quan đến trực quan dữ liệu theo trung bình 7 ngày gần nhất:
- Với mỗi quốc gia mà thuộc về nhóm, trên từng năm hãy vẽ biểu đồ thể hiện trục Ox là thời gian, trục Oy là nhiễm bệnh/tử vong. Hãy dùng 4 ký số của mã đề để vẽ 4 tháng tương ứng theo ký số đó. Nếu ký số là 0 thì lấy tháng là 10.

- Dùng trung bình của các ca nhiễm bệnh và tử vong được báo cáo trong 7 ngày gần nhất để loại trừ một số báo cáo không thường xuyên và đưa chúng ta đến gần hơn với con số hàng ngày.

1) Biểu đồ thể hiện thu thập dữ liệu nhiễm bệnh cho từng tháng

2) Biểu đồ thể hiện thu thập dữ liệu tử vong cho từng tháng

3) Biểu đồ thể hiện thu thập dữ liệu gồm nhiễm bệnh và tử vong cho từng tháng

4) Biểu đồ thể hiện thu thập dữ liệu nhiễm bệnh gồm 2 tháng cuối của năm

5) Biểu đồ thể hiện thu thập dữ liệu tử vong gồm 2 tháng cuối của năm

6) Biểu đồ thể hiện thu thập dữ liệu gồm nhiễm bệnh và tử vong gồm 2 tháng cuối của năm

7) Biểu đồ thể hiện thu thập dữ liệu nhiễm bệnh tích lũy cho từng tháng

8) Biểu đồ thể hiện thu thập dữ liệu tử vong tích lũy cho từng tháng

vii) Nhóm câu hỏi liên quan đến tất cả quốc gia theo thời gian là tháng

- Trên từng năm hãy vẽ biểu đồ thể hiện trục Ox là thời gian, trục Oy là nhiễm bệnh/tử vong. Hãy dùng 4 ký số của mã đề để vẽ 4 tháng tương ứng theo ký số đó. Nếu ký số là 0 thì lấy tháng là 10.

1 Biểu đồ thể hiện thu thập dữ liệu nhiễm bệnh theo thời gian là tháng của tất cả quốc gia

2 Biểu đồ thể hiện thu thập dữ liệu tử vong theo thời gian là tháng của tất cả quốc gia

3 Biểu đồ thể hiện thu thập dữ liệu nhiễm bệnh theo thời gian là 2 tháng cuối của năm của tất cả quốc giaị

4 Biểu đồ thể hiện thu thập dữ liệu tử vong theo thời gian là 2 tháng cuối của năm của tất cả quốc gia

5 Biểu đồ thể hiện thu thập dữ liệu nhiễm bệnh tương đối tích lũy theo thời gian là 2 tháng cuối của năm của tất cả quốc giaị

6 Biểu đồ thể hiện thu thập dữ liệu tử vong tương đối tích lũy theo thời gian là 2 tháng cuối của năm của tất cả quốc gia

viii) Nhóm câu hỏi liên quan đến tất cả quốc gia theo trung bình 7 ngày gần nhất Trên từng năm hãy vẽ biểu đồ thể hiện trục Ox là thời gian, trục Oy là nhiễm bệnh/tử vong. Hãy dùng 4 ký số của mã đề để vẽ 4 tháng tương ứng theo ký số đó. Nếu ký số là 0 thì lấy tháng là 10.

1) Biểu đồ thể hiện thu thập dữ liệu nhiễm bệnh theo thời gian là tháng của tất cả quốc gia theo trung bình 7 ngày gần nhất
2) Biểu đồ thể hiện thu thập dữ liệu tử vong theo thời gian là tháng của tất cả quốc gia theo trung bình 7 ngày gần nhất
3) Biểu đồ thể hiện thu thập dữ liệu nhiễm bệnh theo thời gian là 2 tháng của năm của tất cả quốc giai theo trung bình 7 ngày gần nhất
4) Biểu đồ thể hiện thu thập dữ liệu tử vong theo thời gian là 2 tháng của năm của tất cả quốc gia theo trung bình 7 ngày gần nhất
5) Biểu đồ thể hiện thu thập dữ liệu nhiễm bệnh tích lũy theo thời gian là 2 tháng của năm của tất cả quốc giai theo trung bình 7 ngày gần nhất
6) Biểu đồ thể hiện thu thập dữ liệu tử vong tích lũy theo thời gian là 2 tháng của năm của tất cả quốc gia theo trung bình 7 ngày gần nhất

ix) Nhóm câu hỏi liên quan đến sự tương quan giữa nhiễm bệnh và tử vong

1) Vẽ biểu đồ thể hiện phần trăm giữa nhiễm bệnh tích lũy trên tổng nhiễm bệnh và phần trăm tử vong tích lũy trên tổng số tử vong cho từng quốc gia theo thời gian. Vẽ 2 đường trên cùng biểu đồ

   Trên từng quốc gia riêng của nhóm hãy vẽ biểu đồ thể hiện trục Ox là nhiễm bệnh, trục Oy là tử vong. Hãy lấy 4 tháng theo 4 ký số mã đề thể hiện. Nếu ký số là 0 thì lấy tháng là 10.
2) Xét tương quan trong mỗi tháng,
3) Xét tương quan trong mỗi tháng theo trung bình 7 ngày gần nhất

x) Nhóm câu hỏi riêng

1) So sánh tình trạng nhiễm bệnh của các quốc gia trong 7 ngày cuối của năm cuối cùng
2) So sánh tình trạng tử vong của các quốc gia trong 7 ngày cuối của năm cuối cùng
3) Cho biết các khoảng thời gian nào mà tỉ lệ tử vong tích lũy giảm mạnh nhưng tỉ lệ nhiễm bệnh tích lũy tăng mạnh hoặc ngược lại cho các quốc gia.
4) Với k là mốc bùng phát dịch, hãy xác định k và cho biết các khoảng thời gian bùng phát
5) Với k là mốc bùng tử vong, hãy xác định k và cho biết các khoảng thời gian bùng phát
6) Khoảng thời gian bùng phát nhiễm bệnh lớn nhất giữa các quốc gia có chồng lên nhau không, Cho biết khoảng thời gian giao nhau đó?
7) Khoảng thời gian bùng phát tử vong lớn nhất giữa các quốc gia có chồng lên nhau không, Cho biết khoảng thời gian giao nhau đó?
8) Thử dự đoán thời gian nào dịch sẽ giảm tối thiểu hay kết thúc ở các quốc gia nhóm đã phân tích, đưa ra giải thích của nhóm
9) Cho nhận xét của các bạn về tình hình dịch theo các quốc mà nhóm đã phân tích
10) Hãy mô tả mối quan hệ tuyến tính giữa nhiễm bệnh và tử vong bằng cách đo độ kết hợp của mối quan hệ dùng correlation r (correlation coefficient) và hướng kết hợp.

5 Hướng dẫn và yêu cầu

5.1 Hướng dẫn

- Cài đặt đồng thời cả R và Rstudio.
- Đọc kĩ và xử lý lại tất cả những thí dụ đã có trong file mẫu.
- Tìm hiểu kĩ cách soạn thảo văn bản bằng LaTeX và cách sử dụng phần mềm R trong các file hướng dẫn và tìm hiểu thêm trong các tài liệu khác.
- Tạo một folder chung chứa mọi thứ cần thiết để share giữa các thành viên trong nhóm trên các cloud services như Google Drive hay Dropbox,...
- Dùng Doodle để lên kế hoạch họp nhóm.
- Dùng Trello để quản lý project.

5.2 Yêu cầu

Mỗi nhóm, từ 3 đến 6 sinh viên, đề xuất giải pháp. Nhóm cần nộp báo cáo trình bày về lời giải cho các câu hỏi và kết quả thực nghiệm. Đồng thời, nhóm cũng cần nộp source code, và trình bày các kết quả của nhóm trong khoảng 5 minutes.

Báo cáo và slide trình bày cần được viết dưới dạng LaTeX.

- Thời gian làm bài: Từ ngày 21/02/2022 – 18g00 ngày 20/03/2022.

  Đối với mỗi bài toán, yêu cầu sinh viên trình bày lời giải theo lối truyền thống, sử dụng các công thức, kết quả lý thuyết trong phần kiến thức chuẩn bị. Đồng thời, sau đó trình bày kết quả tính toán và biểu đồ minh họa bằng R.

- Trình bày cả code R và kết quả tính toán trong R giống như file mẫu.

- Viết báo cáo theo đúng **bố cục như trong file mẫu** bằng LaTeX.

- Mỗi nhóm khi nộp bài **cần phải nộp theo file log (nhật ký)** ghi rõ: tiến độ công việc, phân công nhiệm vụ, trao đổi của các thành viên,...

5.3 Nộp bài

- SV chỉ nộp bài qua hệ thống BKEL: nén tất cả các file cần thiết (file .tex, file .R, ...) thành một file tên là "*LOP-NHOM-MADE.zip*": 1-3456.zip và nộp trong mục Assignment.

- Lưu ý: mỗi nhóm **chỉ cần một thành viên là nhóm trưởng nộp bài.**

6 Cách đánh giá và xử lý gian lận

6.1 Đánh giá

Mỗi bài làm sẽ được đánh giá như sau.

Nội dung	Tỉ lệ điểm (%)
Giải đúng các bài toán bằng công thức và lập luận	30%
Các lệnh (hàm) R được sử dụng đúng đắn và hợp lý	30%
Trình bày kiến thức chuẩn bị rõ ràng, phù hợp	20%
Trình bày văn bản đẹp, đúng chuẩn	20%

6.2 Xử lý gian lận

Bài tập lớn phải được sinh viên (nhóm) TỰ LÀM. Sinh viên (nhóm) sẽ bị coi là gian lận nếu:

- Có sự giống nhau bất thường giữa các bài thu hoạch (nhất là phần kiến thức chuẩn bị). Trong trường hợp này, TẤT CẢ các bài nộp có sự giống nhau đều bị coi là gian lận. Do vậy sinh viên (nhóm) phải bảo vệ bài làm của mình.

- Sinh viên (nhóm) không hiểu bài làm do chính mình viết. Sinh viên (nhóm) có thể tham khảo từ bất kỳ nguồn tài liệu nào, tuy nhiên phải đảm bảo rằng mình hiểu rõ ý nghĩa của tất cả những gì mình viết.

Bài bị phát hiện gian lận thì sinh viên sẽ bị xử ý theo quy định của nhà trường.

Tài liệu

[Dal] Dalgaard, P. *Introductory Statistics with R.* Springer 2008.

[K-Z] Kenett, R. S. and Zacks, S. *Modern Industrial Statistics: with applications in R, MINITAB and JMP,* 2nd ed., John Wiley and Sons, 2014.

[Ker] Kerns, G. J. *Introduction to Probability and Statistics Using R,* 2nd ed., CRC 2015.