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: Vũ Tuấn Hưng – 2033150
Phạm Hoàng Vĩ – 1937055
Lưu Quốc Bình – 2033009
Nguyễn Thị Kim Khoa – 2049573
Nguyễn Tiến Đăng Khoa – 1832026

Tp. Hồ Chí Minh, Tháng 04/2022

Mục lục

1 Nhóm câu hỏi liên quan đến tổng quát dữ liệu

1) Tập mẫu thể hiện thu thập dữ liệu vào các năm nào.

- A là tập hợp các bản ghi
- $f1 : A \to B$ với $f1(x)$ là hàm lấy ra năm từ ngày của bản ghi x
- Vậy B là tập hợp năm được thống kê

	.
1	2020
2	2021
3	2022

Hình 1: *Dữ liệu thu thập qua các năm*

2) Số lượng đất nước và định danh của mỗi đất nước (hiển thị 10 đất nước đầu tiên).

- A là tập hợp các bản ghi
- $f : A \to C$ với $f2(x)$ là hàm lấy bộ (iso_code, location) từ bản ghi x
- Số đất nước: $x = |C|$
- Tập hợp 10 quốc gia đầu tiên: $D = \{c_i | c_i \in C \land i \in N \land i \geq 1 \land i \leq 10\}$

	iso_code	Country
1	AFG	Afghanistan
2	OWID_AFR	Africa
3	ALB	Albania
4	DZA	Algeria
5	AND	Andorra
6	AGO	Angola
7	AIA	Anguilla
8	ATG	Antigua and Barbuda
9	ARG	Argentina
10	ARM	Armenia

Hình 2: *Hiển thị 10 đất nước đầu tiên theo iso_code và Country*

3) Số lượng châu lục trong tập mẫu

- A là tập hợp các bản ghi
- $f : A \to B$ với f là hàm lấy ra châu lục từ bản ghi của A
- $|B|$: số các châu lục

	continent	so_chau_luc
1	Asia	Châu Á
2	Europe	Châu Âu
3	Africa	Châu Phi
4	North America	Châu Bắc Mĩ
5	South America	Châu Nam Mĩ
6	Oceania	Châu Đại Dương

Hình 3: *Số lượng châu lục trong tập mẫu*

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

- A là tập hợp các bản ghi
- $f : A \to B$ với f là hàm lấy ra châu lục từ bản ghi của A, $b_i \in B$
- $x_i = \sum_1^\infty 1 \forall f(A) = b_i$: số lượng dữ liệu thu thập từng châu lục
- $x = \sum_1^\infty x_i$: tổng số dữ liệu thu thập

	continent	Observations
1	Africa	38647
2	Asia	35528
3	Europe	36375
4	North America	24438
5	Oceania	8993
6	South America	9335
7	Tong	153316

Hình 4: *Dữ liệu thu thập của từng châu lục*

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

- A là tập hợp các bản ghi
- $f : A \to B$ với f là hàm lấy ra đất nước từ bản ghi của A, $b_i \in B$
- $x_i = \sum_1^\infty 1 \forall f(A) = b_i$: số lượng dữ liệu thu thập của quốc gia b_i
- $x = \sum_1^\infty x_i$: tổng số dữ liệu thu thập

	iso_code	Observation
1	VEN	708
2	VGB	694
3	VNM	759
4	VUT	467
5	WLF	489
6	WSM	459
7	YEM	681
8	ZAF	744
9	ZMB	704
10	ZWE	702
11	Tong	163090

Hình 5: *Dữ liệu 10 đất nước cuối cùng trong bản dữ liệu*

6) Cho biết các châu lục nào có lượng dữ liệu thu thập nhỏ nhất và giá trị nhỏ nhất đó?

- A là tập hợp các bản ghi
- $f : A \to B$ với f là hàm lấy ra châu lục từ bản ghi của A, $b_i \in B$
- $x_i = \sum_1^\infty 1 \forall f(A) = b_i$: số lượng dữ liệu thu thập của châu lục b_i
- $x_{min} \in \{x_i | x_{min} \leq x_i \forall x_i \in \{x_i\}\}$: số lượng dữ liệu thu thập tại 1 châu lục thấp nhất
- $b_{min} \in \{b_i | x_i = x_{min}\}$: châu lục có số lượng dữ liệu thu thập thấp nhất

	continent	Observations
1	Oceania	8993

Hình 6: *Châu lục có dữ liệu nhỏ nhất*

7) Cho biết các châu lục nào có lượng dữ liệu thu thập lớn nhất và giá trị lớn nhất đó?

- A là tập hợp các bản ghi
- $f : A \to B$ với f là hàm lấy ra châu lục từ bản ghi của A, $b_i \in B$
- $x_i = \sum_1^\infty 1 \forall f(A) = b_i$: số lượng dữ liệu thu thập của châu lục b_i
- $x_{max} \in \{x_i | x_{max} \leq x_i \forall x_i \in \{x_i\}\}$: số lượng dữ liệu thu thập tại 1 châu lục lớn nhất
- $b_{max} \in \{b_i | x_i = x_{max}\}$: châu lục có số lượng dữ liệu thu thập lớn nhất

	continent	Observations
1	Africa	38647

Hình 7: *Châu lục có dữ liệu lớn nhất*

8) Cho biết các nước nào có lượng dữ liệu thu thập nhỏ nhất và giá trị nhỏ nhất đó?

- A là tập hợp các bản ghi
- $f : A \to B$ với f là hàm lấy ra quốc gia từ bản ghi của A, $b_i \in B$
- $x_i = \sum_1^\infty 1 \forall f(A) = b_i$: số lượng dữ liệu thu thập của quốc gia b_i
- $x_{min} \in \{x_i | x_{min} \leq x_i \forall x_i \in \{x_i\}\}$: số lượng dữ liệu thu thập tại 1 quốc gia thấp nhất
- $b_{min} \in \{b_i | x_i = x_{min}\}$: quốc gia có số lượng dữ liệu thu thập thấp nhất

	location	n
1	Pitcairn	85

Hình 8: *Nước có dữ liệu nhỏ nhất*

9) Cho biết các nước nào có lượng dữ liệu thu thập lớn nhất và giá trị lớn nhất đó?

- A là tập hợp các bản ghi
- $f : A \to B$ với f là hàm lấy ra quốc gia từ bản ghi của A, $b_i \in B$
- $x_i = \sum_1^\infty 1 \forall f(A) = b_i$: số lượng dữ liệu thu thập của quốc gia b_i
- $x_{max} \in \{x_i | x_{max} \leq x_i \forall x_i \in \{x_i\}\}$: số lượng dữ liệu thu thập tại 1 quốc gia cao nhất
- $b_{max} \in \{b_i | x_i = x_{max}\}$: quốc gia có số lượng dữ liệu thu thập cao nhất

	location	n
1	Argentina	781
2	Mexico	781

Hình 9: *Nước có dữ liệu lớn nhất*

10) Cho biết các date nào có lượng dữ liệu thu thập nhỏ nhất và giá trị nhỏ nhất đó?

- A là tập hợp các bản ghi
- $f : A \to B$ với f là hàm lấy ra ngày từ bản ghi của A, $b_i \in B$
- $x_i = \sum_1^\infty 1 \forall f(A) = b_i$: số lượng dữ liệu thu thập của ngày b_i
- $x_{min} \in \{x_i | x_{min} \leq x_i \forall x_i \in \{x_i\}\}$: số lượng dữ liệu thu thập tại 1 ngày thấp nhất
- $b_{min} \in \{b_i | x_i = x_{min}\}$: ngày có số lượng dữ liệu thu thập thấp nhất

	date	n
1	1/1/2020	2
2	1/2/2020	2
3	1/3/2020	2

Hình 10: *Ngày có dữ liệu nhỏ nhất*

11) Cho biết các date nào có lượng dữ liệu thu thập lớn nhất và giá trị lớn nhất đó?

- A là tập hợp các bản ghi
- $f : A \to B$ với f là hàm lấy ra ngày từ bản ghi của A, $b_i \in B$
- $x_i = \sum_1^\infty 1 \forall f(A) = b_i$: số lượng dữ liệu thu thập của ngày b_i
- $x_{max} \in \{x_i | x_{max} \leq x_i \forall x_i \in \{x_i\}\}$: số lượng dữ liệu thu thập tại 1 ngày cao nhất
- $b_{max} \in \{b_i | x_i = x_{max}\}$: ngày có số lượng dữ liệu thu thập cao nhất

	date	n
1	8/22/2021	238
2	8/23/2021	238
3	8/24/2021	238
4	8/25/2021	238
5	8/26/2021	238
6	8/27/2021	238
7	8/28/2021	238
8	8/29/2021	238

Hình 11: *Ngày có dữ liệu lớn nhất*

12) Cho biết số lượng dữ liệu thu thập được theo date và châu lục.

- A là tập hợp các bản ghi
- $f : A \to B$ với f là hàm lấy bộ (date, continent) từ bản ghi, $B = \{(x_i, y_i)\}$
- $z_i = \sum_1^\infty 1 \forall f(A) = (x_i, y_i)$: số lượng dữ liệu thu thập theo date và châu lục

	continent	date	n
1	Africa	1/1/2021	55
2	Africa	1/1/2022	55
3	Africa	1/10/2021	55
4	Africa	1/10/2022	55
5	Africa	1/11/2021	55
6	Africa	1/11/2022	55
7	Africa	1/12/2021	55
8	Africa	1/12/2022	55
9	Africa	1/13/2021	55
10	Africa	1/13/2022	55
11	Africa	1/14/2021	55
12	Africa	1/14/2022	55

Showing 1 to 13 of 4,600 entries, 3 total columns

Hình 12: *Dữ liệu thu thập theo ngày và châu lục*

13) Cho biết số lượng dữ liệu thu thập được là lớn nhất theo date và châu lục.

- A là tập hợp các bản ghi
- $f : A \to B$ với f là hàm lấy bộ (date, continent) từ bản ghi, $B = \{(x_i, y_i)\}$
- $z_i = \sum_1^\infty 1 \forall f(A) = (x_i, y_i)$: số lượng dữ liệu thu thập theo date và châu lục
- $z_{max} \in \{z_i | z_{max} \leq z_i \forall z_i \in \{z_i\}\}$: số lượng dữ liệu thu thập tại 1 ngày ở 1 châu lục cao nhất
- $(x_i, y_i) \in \{(x_i, y_i) | z_i = z_{max}\}$: bộ ngày và châu lục có số lượng dữ liệu thu thập cao nhất

	continent	date	n
1	Africa	1/1/2021	55
2	Africa	1/1/2022	55
3	Africa	1/10/2021	55
4	Africa	1/10/2022	55
5	Africa	1/11/2021	55
6	Africa	1/11/2022	55
7	Africa	1/12/2021	55
8	Africa	1/12/2022	55
9	Africa	1/13/2021	55
10	Africa	1/13/2022	55
11	Africa	1/14/2021	55
12	Africa	1/14/2022	55

Showing 1 to 13 of 531 entries, 3 total columns

Hình 13: *Dữ liệu lớn nhất thu thập theo ngày và châu lục*

14) Cho biết số lượng dữ liệu thu thập được là nhỏ nhất theo date và châu lục.

- A là tập hợp các bản ghi
- $f : A \to B$ với f là hàm lấy bộ (date, continent) từ bản ghi, $B = \{(x_i, y_i)\}$
- $z_i = \sum_1^\infty 1 \forall f(A) = (x_i, y_i)$: số lượng dữ liệu thu thập theo date và châu lục
- $z_{min} \in \{z_i | z_{max} \leq z_i \forall z_i \in \{z_i\}\}$: số lượng dữ liệu thu thập tại 1 ngày ở 1 châu lục nhỏ nhất
- $(x_i, y_i) \in \{(x_i, y_i) | z_i = z_{min}\}$: bộ ngày và châu lục có số lượng dữ liệu thu thập nhỏ nhất

	continent	date	n
1	Asia	1/10/2020	1
2	Asia	1/11/2020	1
3	Asia	1/12/2020	1
4	Asia	1/13/2020	1
5	Asia	1/14/2020	1
6	Asia	1/15/2020	1
7	Asia	1/4/2020	1
8	Asia	1/5/2020	1
9	Asia	1/6/2020	1
10	Asia	1/7/2020	1
11	Asia	1/8/2020	1
12	Asia	1/9/2020	1

Showing 1 to 13 of 93 entries, 3 total columns

Hình 14: *Dữ liệu nhỏ nhất thu thập theo ngày 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.

- A là tập hợp các bản ghi
- $f : A \to B$ với f là hàm lấy bộ (date, continent) từ bản ghi, $B = \{(x_i, y_i)\}$
- $z_i = \sum_1^\infty 1 \forall f(A) = (x_i, y_i)$: số lượng dữ liệu thu thập theo date và châu lục
- $(x_i, y_i) \in \{(x_i, y_i) | (x_i, y_i) = (k, t)\}$: bộ ngày k và châu lục t

	continent	date	n
1	Africa	1/1/2021	55

Hình 15: *Với một date và châu lục cho trước*

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.

- A là tập hợp các bản ghi
- $f : A \to B$ với f là hàm lấy ra quốc gia từ bản ghi của A, $b_i \in B$

- $x_i = \sum_1^\infty 1 \forall f(A) = b_i$: số lượng dữ liệu thu thập của quốc gia b_i, $B = \{x_i\}, z_j \in B$
- $y_j = \sum_1^\infty 1 \forall x_i = z_j$: số lượng quốc gia có cùng số liệu thu thập z_j
- $C = \{b_i | x_i = z_j \wedge y_j \geq 2\}$: tập hợp các quốc gia có chung số liệu thu thập

e × | 2033009_LuuQuocBinh.R × | rs_data ×

Filter

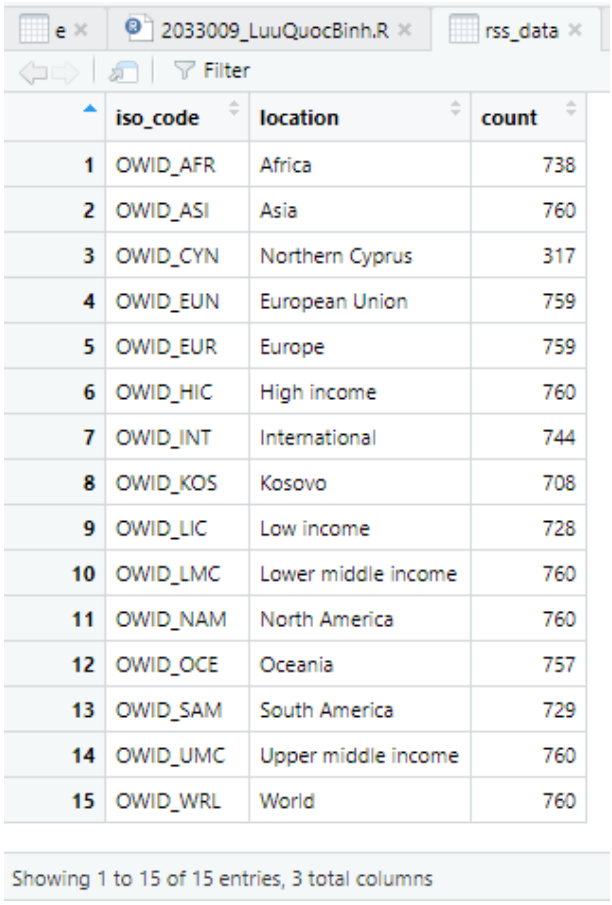
	id	Countries	duplicate_value
1	238	Afghanistan	1
2	238	Africa	1
3	238	Albania	1
4	238	Algeria	1
5	238	Andorra	1
6	238	Angola	1
7	238	Anguilla	1
8	238	Antigua and Barbuda	1
9	238	Argentina	1
10	238	Armenia	1
11	238	Aruba	1
12	238	Asia	1
13	238	Australia	1
14	238	Austria	1
15	238	Azerbaijan	1
16	238	Bahamas	1

Showing 1 to 16 of 238 entries, 3 total columns

Hình 16: *Các đất nước có số lượng dữ liệu bằng nhau*

17) Liệt kê iso_code, tên đất nước mà chiều dài iso_code lớn hơn 3

- A là tập hợp các bản ghi
- $f : A \to B$ với f là hàm lấy ra bộ (iso_code, location) từ bản ghi của A, $(a_i, b_i) \in B$
- $f1(x)$ là hàm lấy độ dài của chuỗi
- $C = \{(a_i, bi) | f1(a_i) > 3\}$: bộ iso_code, tên đất nước mà chiều dài iso_code lớn hơn 3



e | 2033009_LuuQuocBinh.R | rss_data

Filter

	iso_code	location	count
1	OWID_AFR	Africa	738
2	OWID_ASI	Asia	760
3	OWID_CYN	Northern Cyprus	317
4	OWID_EUN	European Union	759
5	OWID_EUR	Europe	759
6	OWID_HIC	High income	760
7	OWID_INT	International	744
8	OWID_KOS	Kosovo	708
9	OWID_LIC	Low income	728
10	OWID_LMC	Lower middle income	760
11	OWID_NAM	North America	760
12	OWID_OCE	Oceania	757
13	OWID_SAM	South America	729
14	OWID_UMC	Upper middle income	760
15	OWID_WRL	World	760

Showing 1 to 15 of 15 entries, 3 total columns

Hình 17: *Danh sách đất nước có iso_ code lớn hơn 3*

2 Nhóm câu hỏi liên quan đến mô tả thống kê cơ bản dữ liệu

1) Tính giá trị nhỏ nhất, lớn nhất.

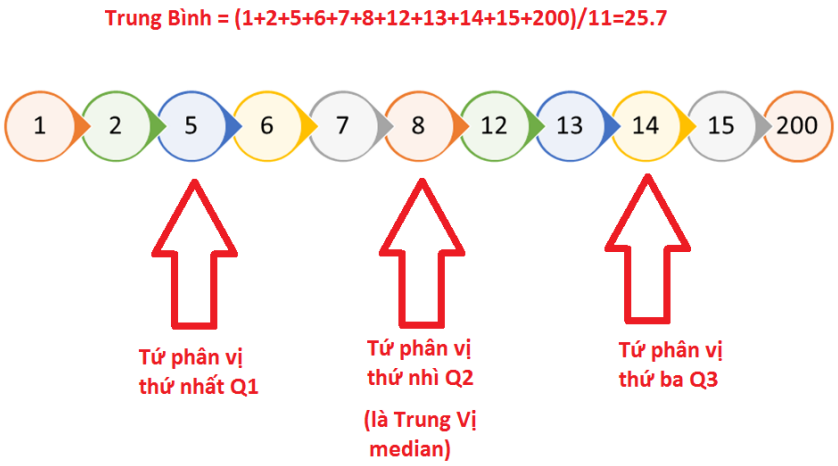
- Công thức biểu diễn.
  - Tìm giá trị lớn nhất: $max_P = maxa_1, a_2, \cdots, a_n$
  - Tìm giá trị nhỏ nhất: $min_P = mina_1, a_2, \cdots, a_n$
    - P = $\{a_1, a_2, \cdots, a_n\}$, là tập đang xét
    - max_P: giá trị lớn nhất trong tập P
    - min_P: giá trị nhỏ nhất trong tập P
    - n: số lượng phần tử trong tập cần tìm
    - $a_1, a_2, \cdots, a_n$: giá trị phần tử thứ $1, 2, \cdots n$ trong tập.
- Sử dụng hàm max(), min() để lấy giá trị lớn nhất, nhỏ nhất cho 2 giá trị new_cases và new_deaths.
- Đưa ra kết quả.

```
> print(max_new_case)
# A tibble: 1 x 6
  iso_code continent location date     new_cases new_deaths
  <chr>    <chr>     <chr>    <chr>        <dbl>      <dbl>
1 JPN      Asia      Japan    2/3/2022    104345         90
> print(min_new_case)
# A tibble: 1 x 6
  iso_code continent location  date     new_cases new_deaths
  <chr>    <chr>     <chr>     <chr>        <dbl>      <dbl>
1 IDN      Asia      Indonesia 3/3/2020         0         NA
```

Hình 18: *Kết quả max, min của new_cases và new_deaths*

2) Tính tứ phân vị thứ nhất(Q1), thứ hai(Q2), thứ ba(Q3)

- Tứ phân vị là đại lượng mô tả sự phân bố và sự phân tán của tập dữ liệu. Tứ phân vị có 3 giá trị, đó là tứ phân vị thứ nhất, thứ nhì, và thứ ba.
- Ba giá trị này chia một tập hợp dữ liệu (đã sắp xếp dữ liệu theo trật từ từ bé đến lớn) thành 4 phần có số lượng quan sát đều nhau.
- Giá trị tứ phân vị thứ hai Q2 chính bằng giá trị trung vị.
- Giá trị tứ phân vị thứ nhất Q1 bằng trung vị phần dưới.
- Giá trị tứ phân vị thứ ba Q3 bằng trung vị phần trên.
- Ví dụ: Tập dữ liệu bao gồm 1,2,5,6,7,8,12,13,14,15,200.
- Tập dữ liệu trên đã được sắp xếp theo thứ tự tăng dần, dễ dàng nhận thấy giá trị trung vị nằm giữa chính là 14.
- Trung vị của tập dữ liệu phần dưới 1,2,5 là 7.
- Và trung vị của tập dữ liệu phần trên 14,15,200 là 34.
- Vậy Q1 = 5, Q2 = 8, Q3 = 14



Hình 19: *Hình minh hoạ Tứ Phân Vị*

- Sử dụng hàm quantile() để lấy tứ phân vị.
- Đưa ra kết quả.
- Source code và kết quả đạt được là

```
> # New death
> print(q_arr_nc)
      0%      25%      50%      75%     100%
     0.0    152.0    687.5   2562.0 104345.0
>
> # New case
> print(q_arr_nd)
  0%  25%  50%  75% 100%
   0    3   11   44 2069
```

Hình 20: *Kết quả tứ phân vị của new_ cases và new_ deaths*

3) Tính giá trị trung bình (Avg)

- Công thức biểu diễn: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\{x_1 + \cdots + x_n\}$
- Sử dụng hàm mean(), để lấy giá trị trung bình.
- Đưa ra kết quả.

```
> # 3) Tính giá trị trung bình (Avg)
> avgnc <- mean(unlist(arr_new_cases))
> avgnd <- mean(unlist(arr_new_deaths))
> # nhiem moi trung binh
> print(avgnc)
[1] 728.8262
>
> # tu vong trung binh
> print(avgnd)
[1] 11.11156
```

Hình 21: *Kết quả giá trị trung bình (Avg) của new_ cases và new_ deaths*

4) Tính giá trị độ lệch chuẩn (Std)

- Công thức tính độ lệch chuẩn: $s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
  - N, n là số phần tử có trong tập hợp/mẫu
  - x_i là phần tử thứ i của quần thể/mẫu
  - $\bar{x}$ là giá trị trung bình của tập
- Sử dụng hàm var(), để lấy độ lệch chuẩn.
- Đưa ra kết quả

```
> # 4) Tính giá trị độ lệch chuẩn (Std)
> variancenc <- var(unlist(arr_new_cases))
> variancend <- var(unlist(arr_new_deaths))
> print(variancenc)
[1] 2273331
> print(variancend)
[1] 373.6357
```

Hình 22: *Kết quả tính độ lệch chuẩn của new_ cases và new_ deaths*

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 $outliers > Q3 + 1.5 * IQR$

- Với $IQR = Q3 - Q1$. Lọc các dòng dữ liệu với giá trị cột $new\_deathsnew\_cases$ thoả điều kiện $Q1 - 1.5 * IQR$ hoặc $outliers > Q3 + 1.5 * IQR$
- Sau đó dùng hàm nrow() để đếm số lượng record
- Đưa ra kết quả.
- Source code và kết quả đạt được là:

```
> # 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
>
> # new case
> q1_nc <- as.numeric(q_arr_nc['25%'])
> q3_nc <- as.numeric(q_arr_nc['75%'])
> qtr_nc <- q3_nc - q1_nc
> data_for_e5nc <- filter(data, !is.na(new_cases))
> data_for_e5nc <- transform(data_for_e5nc, new_cases = as.numeric(new_cases))
> e5nc <- filter(data_for_e5nc, (new_cases < q1_nc - 1.5*qtr_nc) | (new_cases > q3_nc + 1.5*qtr_nc))
> print(nrow(e5nc))
[1] 234
>
> # new death
> q1_nd <- as.numeric(q_arr_nd['25%'])
> q3_nd <- as.numeric(q_arr_nd['75%'])
> qtr_nd <- q3_nd - q1_nd
> data_for_e5nd <- filter(data, !is.na(new_deaths))
> data_for_e5nd <- transform(data_for_e5nc, new_deaths = as.numeric(new_deaths))
> e5nd <- filter(data_for_e5nd, new_deaths < q1_nc - 1.5*qtr_nd | new_deaths > q3_nc - 1.5*qtr_nd)
> print(nrow(e5nd))
```

Hình 23: *Số ca tử vong theo từng quốc gia của tháng 8*

6) Lập bảng mô tả số liệu thống kê cho từng đất nước thuộc về nhóm

- Sử dụng hàm max(), min(), quantile() để tính toán thông số.
- Đưa ra kết quả.
- Source code và kết quả đạt được là:

	Countries	Min	Q1	Q2	Q3	Max	Avg	Std	Outlier
1	Kenya	0	104	271	646	3749	455.10437235543	280263.689655997	709
2	Lesotho	0	0	0	14.25	6925	50.0524691358025	90667.5892056405	648
3	Morocco	0	188.75	577	2241.25	12039	1609.2625	4940957.18134562	720

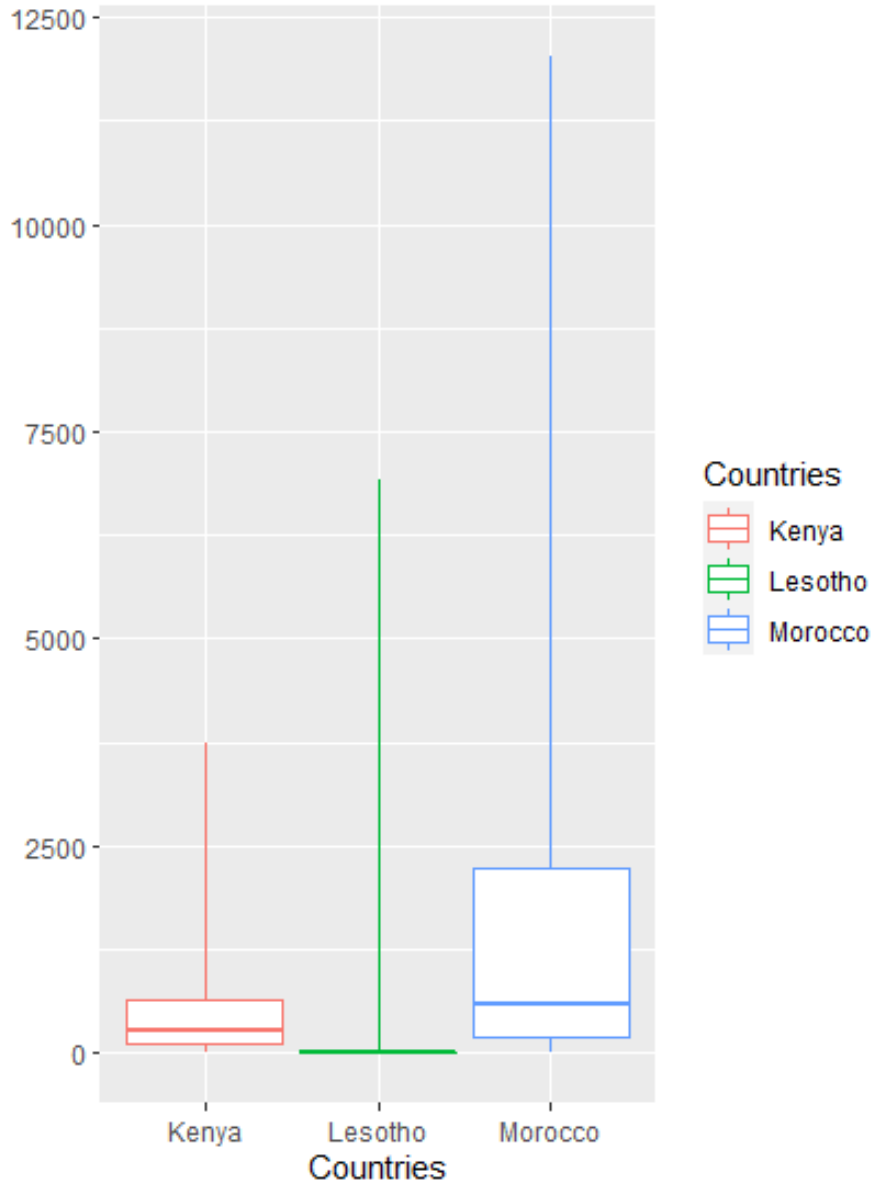
Hình 24: *Câu 6: ket qua cho so truong hợp nhiễm mới (new_ cases).*

	Countries	Min	Q1	Q2	Q3	Max	Avg	Std	Outlier
1	Kenya	0	104	271	646	3749	455.10437235543	280263.689655997	709
2	Lesotho	0	0	0	14.25	6925	50.0524691358025	90667.5892056405	648
3	Morocco	0	188.75	577	2241.25	12039	1609.2625	4940957.18134562	720

Hình 25: *Câu 6: ket qua cho so truong hợp tử vong (new_ deaths).*

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

- Code: tham khảo file R.
- Đưa ra kết quả.



Hình 26: *Câu 7: biểu đồ boxplot hay còn được gọi là box-and-whisker cho nhiễm coronavirus.*

3 Nhóm câu hỏi liên quan đến dữ liệu thể hiện thu thập dữ liệu

1) Có bao nhiêu ngày có số lần dữ liệu không được báo cáo mới.

	location	new_cases
1	Kenya	15
2	Lesotho	366
3	Morocco	30

Hình 27: *Số ngày new case không được báo cáo mới*

	location	new_deaths
1	Kenya	114
2	Lesotho	527
3	Morocco	82

Hình 28: *Số ngày new death không được báo cáo mới*

2) Có bao nhiêu ngày có số ca nhiễm/ tử vong là thấp nhất được báo cáo mới.

	location	new_cases	n
1	Kenya	1	2
2	Lesotho	1	17
3	Morocco	1	6

Hình 29: *Số ngày new case có số lần thu thập dữ liệu thấp nhất được báo cáo mới*

	location	new_deaths	n
1	Kenya	1	59
2	Lesotho	1	49
3	Morocco	1	50

Hình 30: *Số ngày new death có số lần thu thập dữ liệu thấp nhất được báo cáo mới*

3) Có bao nhiêu ngày có số ca nhiễm/ tử vong là cao nhất được báo cáo mới

	location	new_cases	max_new_cases
1	Kenya	3749	1
2	Lesotho	6925	1
3	Morocco	12039	1

Hình 31: *Số ngày new case có số lần thu thập dữ liệu cao nhất được báo cáo mới*

	location	new_deaths	max_new_deaths
1	Kenya	41	1
2	Lesotho	230	1
3	Morocco	127	1

Hình 32: *Số ngày new death có số lần thu thập dữ liệu cao nhất được báo cáo mới*

4) Thể hiện bảng số liệu như sau không được báo cáo mới và báo cáo mới

	location	new_cases	new_deaths
1	Kenya	15	114
2	Lesotho	366	527
3	Morocco	30	82

Hình 33: *Số ngày new case và new death không được báo cáo mới*

	location	new_cases	max_new_cases	new_deaths	max_new_deaths
1	Kenya	3749	1	41	1
2	Lesotho	6925	1	230	1
3	Morocco	12039	1	127	1

Hình 34: *Số dữ liệu max min new case được báo cáo mới*

	location	min_death	max_death
1	Kenya	1	41
2	Lesotho	1	230
3	Morocco	1	127

Hình 35: *Số dữ liệu max min new death được báo cáo mới*

5) Cho biết số ngày ngắn nhất liên tiếp mà không có dữ liệu được báo cáo

- Giải pháp: duyệt từ row đầu tiên đến row cuối cùng, đếm chuỗi các chuỗi ngày mà "liên tiếp mà không có dữ liệu được báo cáo", sinh ra 1 table, tìm max/min từ table đó.
- Code: tham khảo trong file R.
- Đưa ra kết quả.

```
> print(minVal)
  iso_code continent location continuity_new_case start_date  end_date
1      KEN    Africa    Kenya                   6   3/6/2020 3/12/2020
```

Hình 36: *Câu 6: ket qua cho 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

- Dùng chung ý tưởng, và function với câu 5.
- Đưa ra kết quả.

```
> print(maxVal)
  iso_code continent location continuity_new_case start_date end_date
2      MAR    Africa  Morocco                  24   2/7/2020 3/1/2020
```

Hình 37: *Kết quả cho cho số ngày ngắn 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

- Giải pháp: duyệt từ row đầu tiên đến row cuối cùng, đếm chuỗi các chuỗi ngày mà "liên tiếp mà không có người nhiễm bệnh mới", sinh ra 1 table, tìm max/min từ table đó.
- Code: tham khảo trong file R.
- Đưa ra kết quả.

```
> print(minVal)                                                          :e
  iso_code continent location continuity_new_case start_date  end_date '1
1      KEN    Africa    Kenya                   1  3/14/2020 3/14/2020
```

Hình 38: *Kết quả cho 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

- Dùng chung ý tưởng, và function với câu 7.
- Đưa ra kết quả.

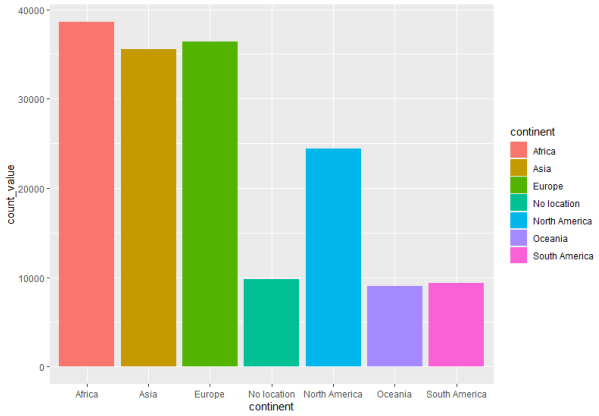
```
> print(maxVal)
    iso_code continent location continuity_new_case start_date  end_date
117      LSO    Africa  Lesotho                  35  8/27/2021 9/30/2021
```

Hình 39: *Kết quả cho số ngày dài nhất liên tiếp mà không có người nhiễm bệnh mới.*

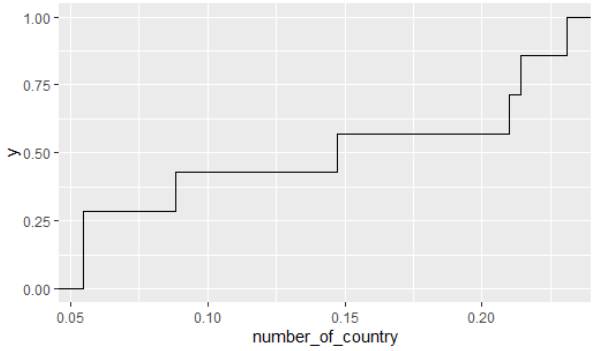
4 Nhóm câu hỏi liên quan đến trực quan dữ liệu

1 Vẽ biểu đồ tần số tích lũy quốc gia cho các châu lục

- Cách giải.
  - Bước 1: Group by *location* và *continent*, ta sẽ đếm được so quoc gia. Sau đó lưu kết quả.
  - Bước 2: Từ kết quả đạt được, group by *continent* sẽ đếm. Sẽ ra được số lượng quốc gia theo từng châu lục.
  - Bước 3: Dùng hàm *stat_ecdf*(empirical cumulative distribution function), để tính tần số tích luỹ.
  - Bước 4: Sau đó từ kết quả thu dược dùng ggplot, để vẽ biểu đồ tần số tích tuỹ quốc gia theo châu lục.
- Code: tham khảo file R.
- Đưa ra kết quả.



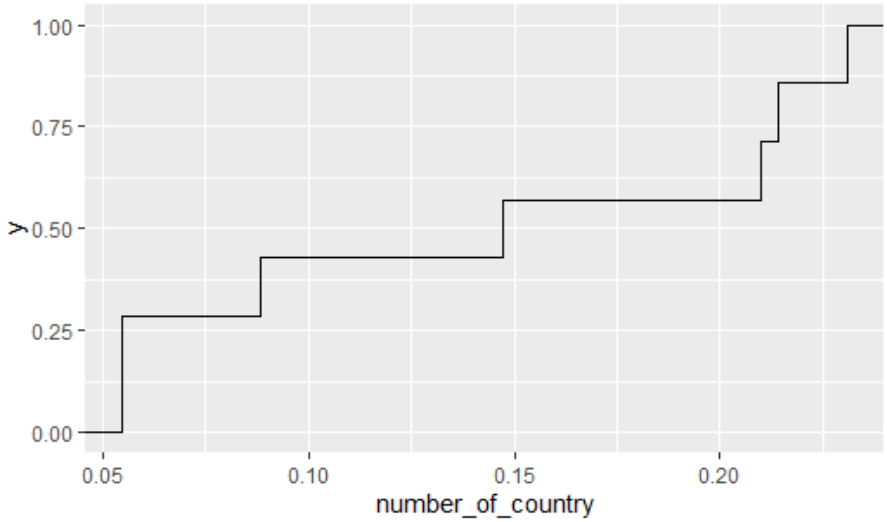
Hình 40: *Câu 6: ket qua cho so truong hợp nhiễm mới (new_cases).*



Hình 41: *Câu 6: ket qua cho so truong hợp nhiễm mới (new_cases).*

2 Vẽ biểu đồ tần số tương đối quốc gia cho các châu lục

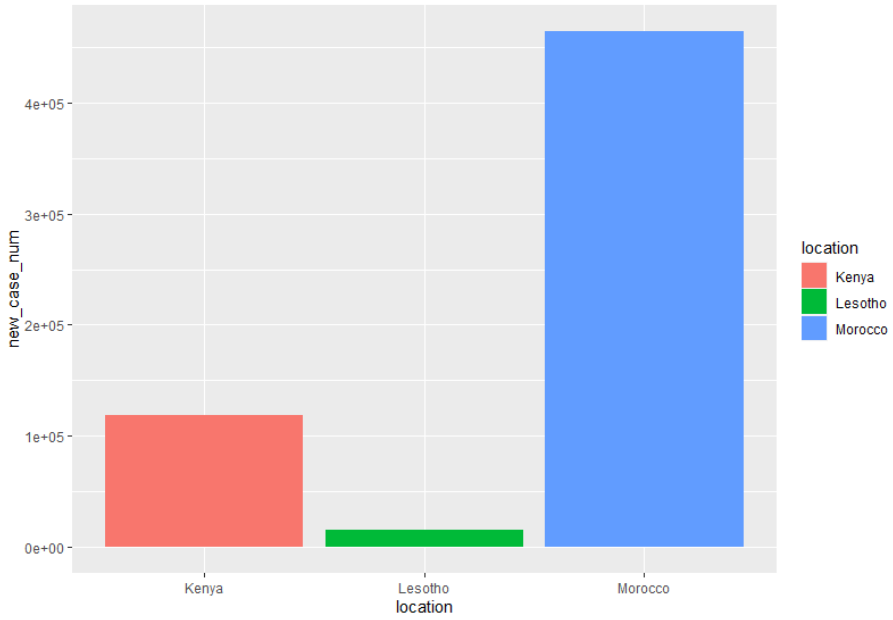
- Cách giải tương tự câu 1, tuy nhiên có thêm bước 2, 5, chia *slngqucgiatheochulc* cho *tngttcqucgia*
  - Bước 1: Group by *location* và *continent*, ta sẽ đếm được so quoc gia. Sau đó lưu kết quả.
  - Bước 2: Từ kết quả đạt được, group by *continent* sẽ đếm. Sẽ ra được số lượng quốc gia theo từng châu lục.
  - Bước 2, 5: chia *slngqucgiatheochulc* cho *tngttcqucgia*
  - Bước 3: Dùng hàm *stat_ecdf*(empirical cumulative distribution function), để tính tần số tích luỹ.
  - Bước 4: Sau đó từ kết quả thu dược dùng ggplot, để vẽ biểu đồ tần số tích tuỹ quốc gia theo châu lục.
- Code: tham khảo file R.
- Đưa ra kết quả.



Hình 42: *Câu 6: ket qua cho so truong hợp nhiễm mới (new_cases).*

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

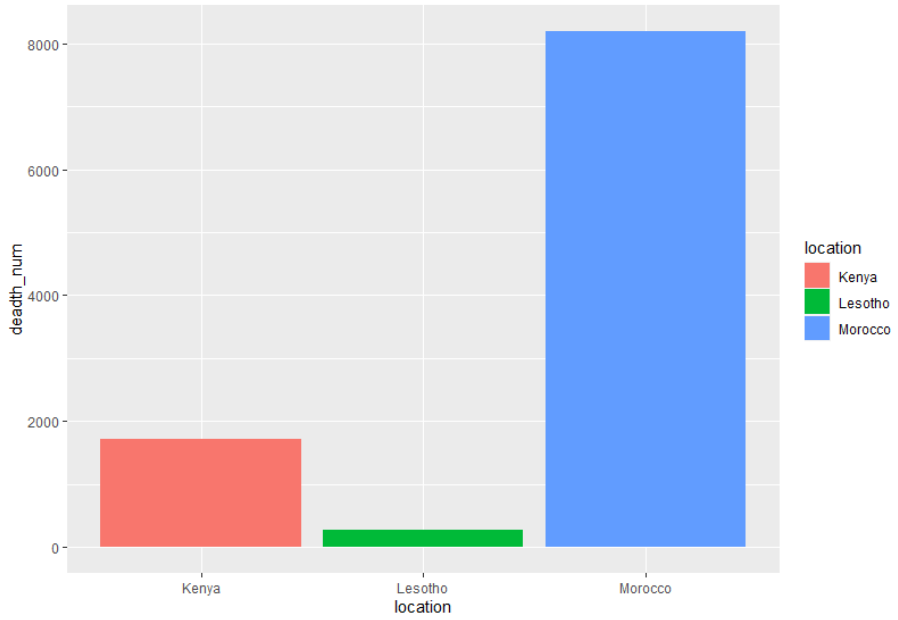
- Cách giải.
  - Bước 1: tính ngày cuối cùng.
  - Bước 2: lấy ngày cuối cùng trừ đi 7.
  - Bước 3: khởi tạo data, filter N/A cho field new_cases
  - Bước 4: filter data có ngày > ngày vừa tìm được ở bước 2, và in kết quả.
- Code: tham khảo file R.
- Đưa ra kết quả.



Hình 43: *Câu 6: ket qua cho so truong hợp nhiễm mới (new_ cases).*

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

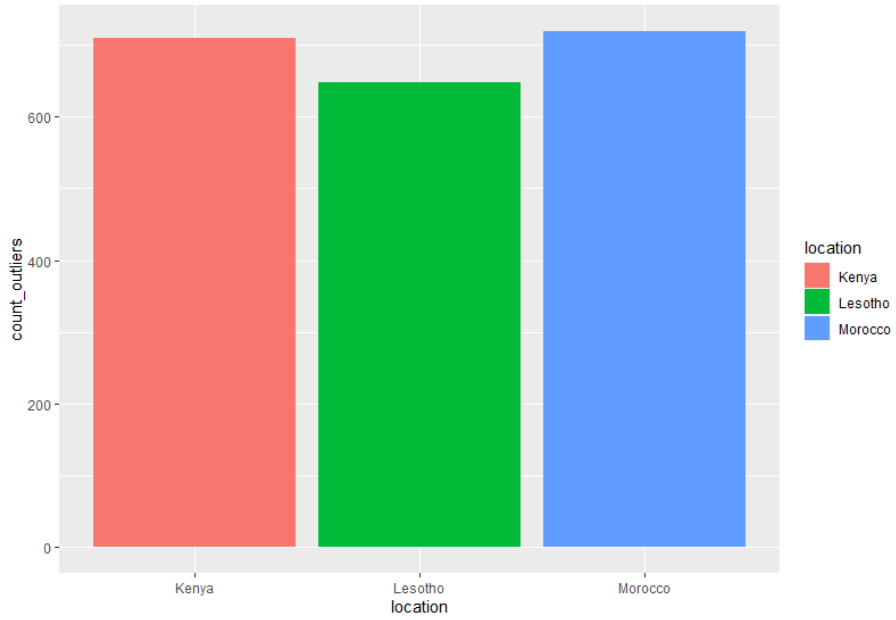
- Cách giải.
  - Bước 1: tính ngày cuối cùng.
  - Bước 2: lấy ngày cuối cùng trừ đi 7.
  - Bước 3: khởi tạo data, filter N/A cho field new_deaths
  - Bước 4: filter data có ngày > ngày vừa tìm được ở bước 2, và in kết quả.
- Code: tham khảo file R.
- Đưa ra kết quả.



Hình 44: *Câu 6: ket qua cho so truong hợp nhiễm mới (new_ cases).*

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

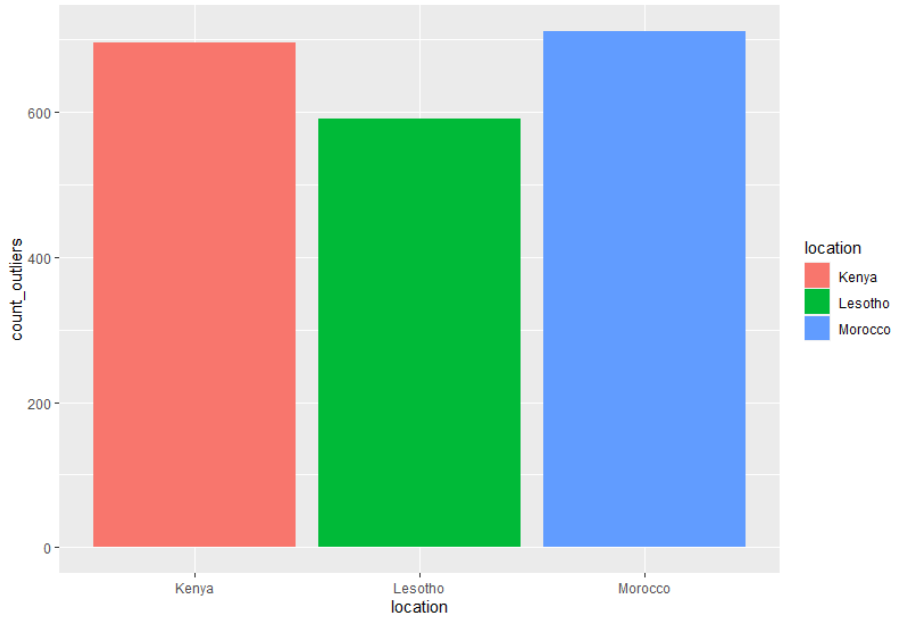
- Cách giải.
  - Bước 1: tính ngày cuối cùng.
  - Bước 2: lấy ngày cuối cùng trừ đi 7.
  - Bước 3: khởi tạo data, tính điều kiện dùng để lọc outliers. (đã để cập ở phần *ii*5)
    $IQR = Q3 - Q1$
    $outliers < Q1 - 1.5 * IQR$ hoặc $outliers > Q3 + 1.5 * IQR$
  - Bước 4: filter data thoả $new\_cases < Q1 - 1.5 * IQR$ hoặc $new\_cases > Q3 + 1.5 * IQR$.
- Code: tham khảo file R.
- Đưa ra kết quả.



Hình 45: *Câu 6: ket qua cho so truong hợp nhiễm mới (new_cases).*

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

- Cách giải.
  - Bước 1: tính ngày cuối cùng.
  - Bước 2: lấy ngày cuối cùng trừ đi 7.
  - Bước 3: khởi tạo data, tính điều kiện dùng để lọc outliers. (đã để cập ở phần *ii*5)
    $IQR = Q3 - Q1$
    $outliers < Q1 - 1.5 * IQR$ hoặc $outliers > Q3 + 1.5 * IQR$
  - Bước 4: filter data thoả $new\_deaths < Q1 - 1.5 * IQR$ hoặc $new\_deaths > Q3 + 1.5 * IQR$.
- Code: tham khảo file R.
- Đưa ra kết quả.



Hình 46: *Câu 6: ket qua cho so truong hợp nhiễm mới (new_cases).*

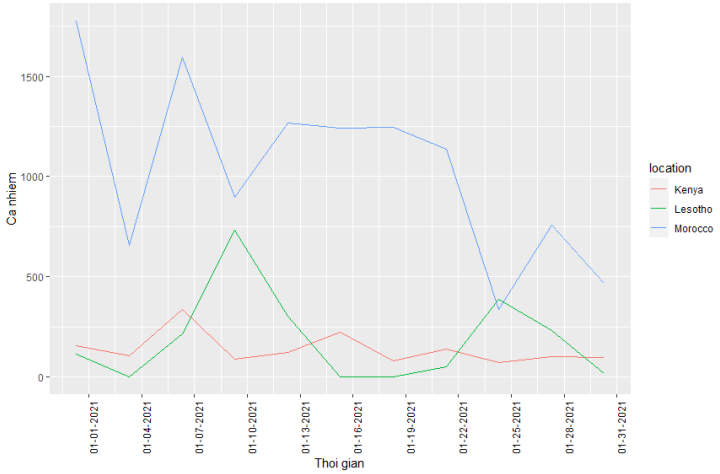
5 Nhóm câu hỏi liên quan đến trực quan dữ liệu theo thời gian là tháng

Cách giải chung

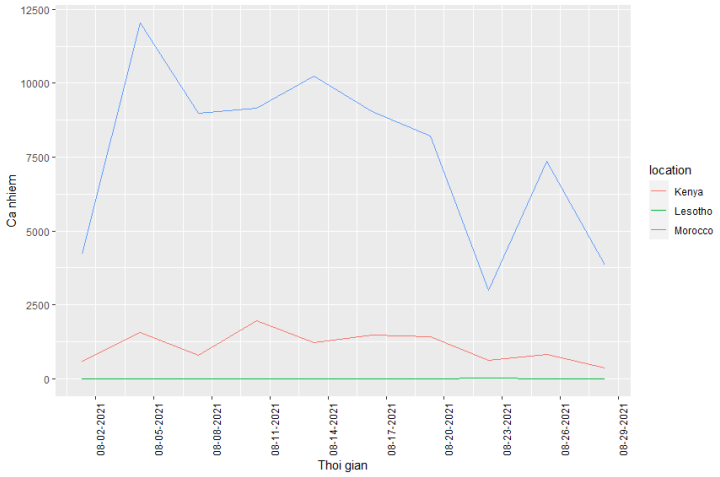
- $A = \{d_i\}$: tập hợp dữ liệu của tất cả các quốc gia
- $P = \{Kenya, Lesotho, Morocco\}$: tập hợp các quốc gia cần thống kê
- $M = \{1, 8, 4, 5\}$: các cần tháng thống kê
- x_i: số ca nhiễm bệnh của ngày d_i

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

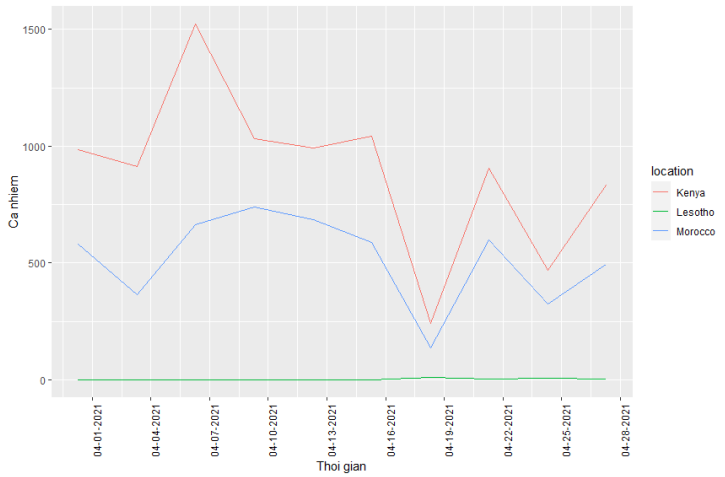
- Thêm cột tháng và năm, định dạng cột đó thành dạng number.
- Tạo 1 vector ngày cần xử lý là 1-8-4-5.
- Vẽ biểu đồ thu thập số ca nhiễm theo từng quốc gia của mỗi tháng



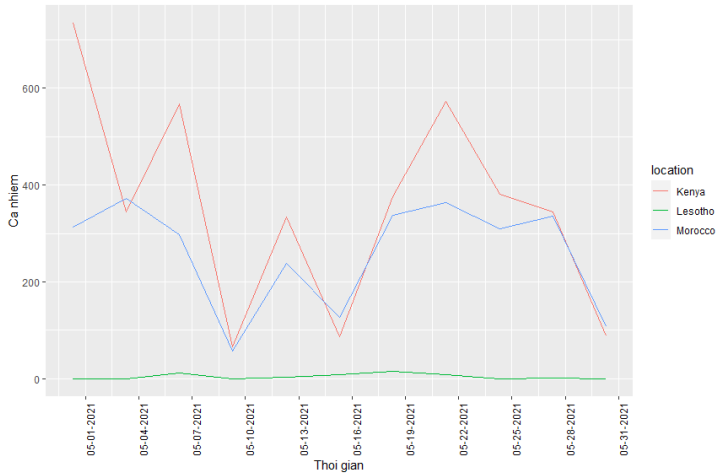
Hình 47: *Biểu đồ thu thập ca nhiễm của cả ba quốc gia trong 4 tháng 1-8-5-4*



Hình 48: *Biểu đồ thu thập ca nhiễm của cả ba quốc gia trong 4 tháng 1-8-5-4*



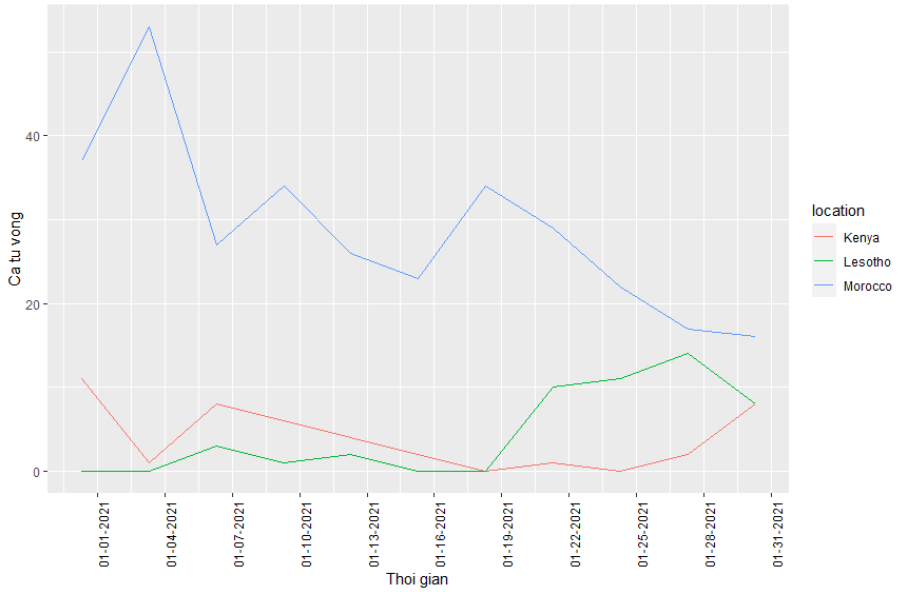
Hình 49: *Biểu đồ thu thập ca nhiễm của cả ba quốc gia trong 4 tháng 1-8-5-4*



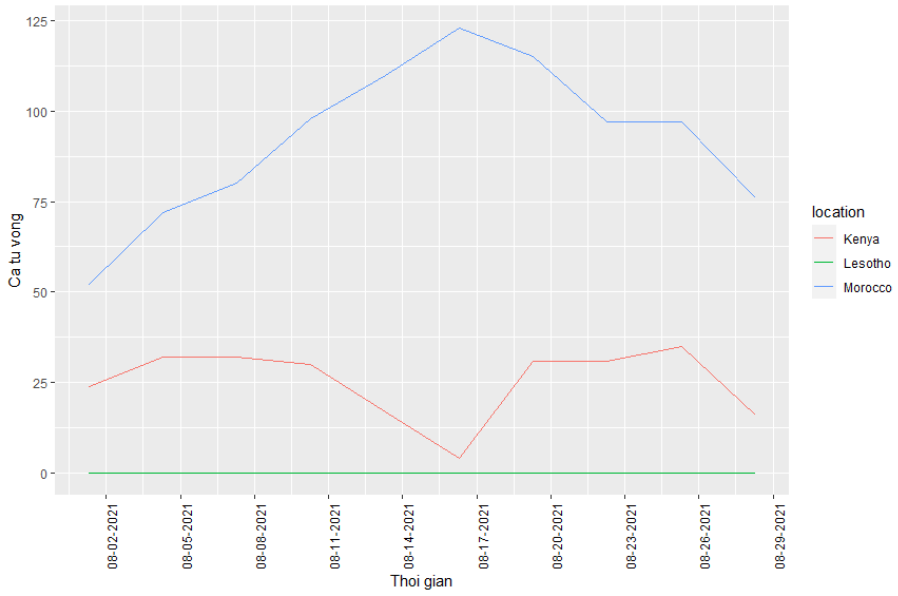
Hình 50: *Biểu đồ thu thập ca nhiễm của cả ba quốc gia trong 4 tháng 1-8-5-4*

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

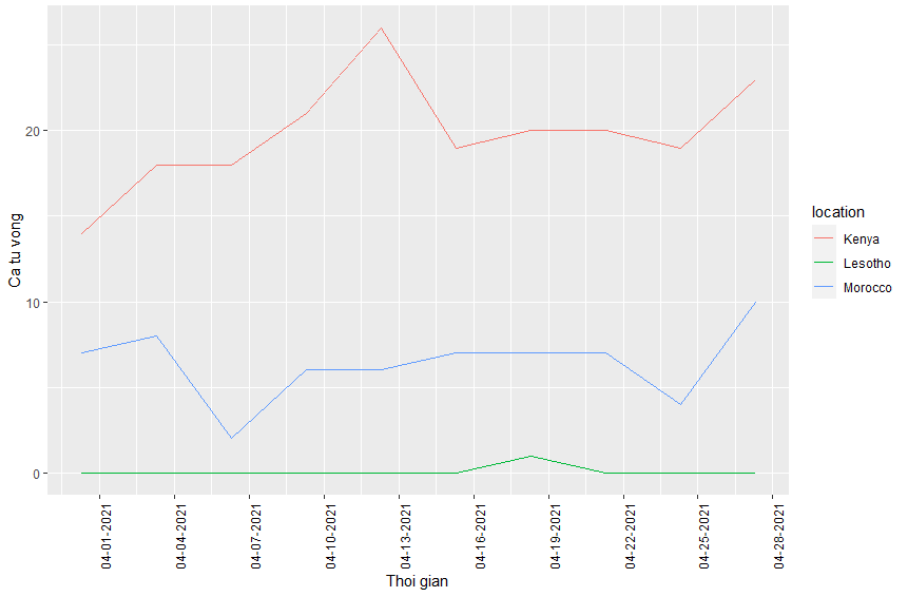
- Thêm cột tháng và năm, định dạng cột đó thành dạng number.
- Tạo 1 vector ngày cần xử lý là 1-8-4-5.
- Vẽ biểu đồ thu thập số ca tử vong theo từng quốc gia của mỗi tháng



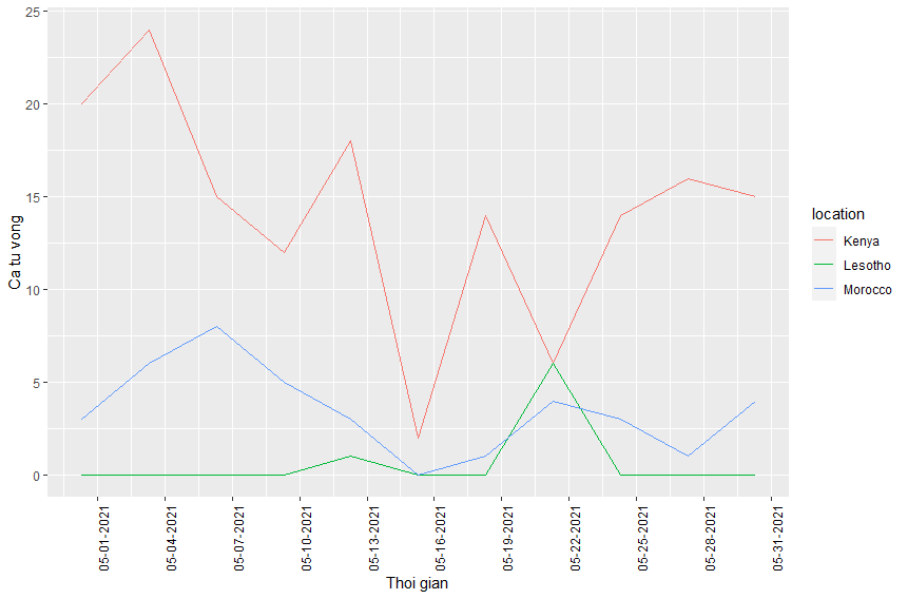
Hình 51: *Biểu đồ thu thập ca tử vong của cả ba quốc gia trong 4 tháng 1-8-5-4.*



Hình 52: *Biểu đồ thu thập ca tử vong của cả ba quốc gia trong 4 tháng 1-8-5-4.*



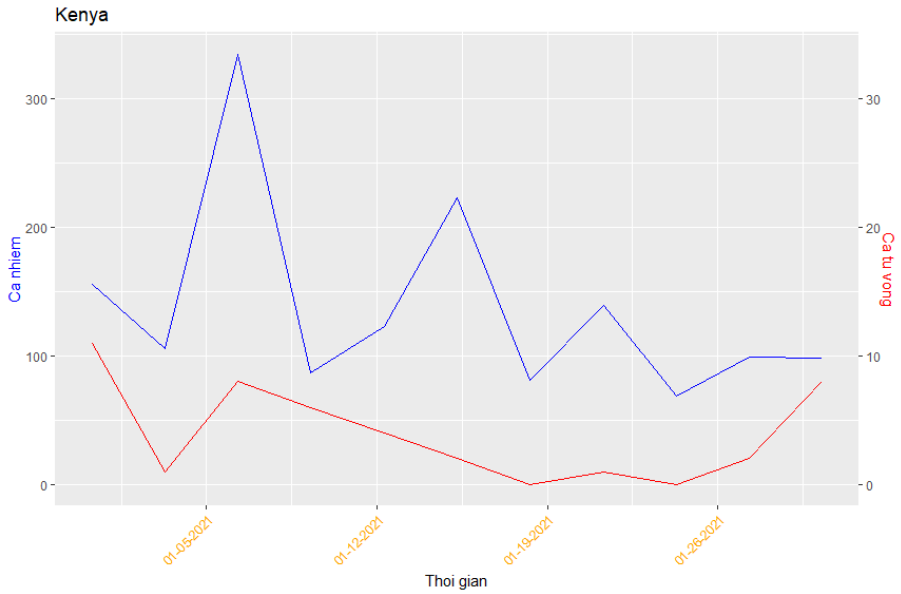
Hình 53: *Biểu đồ thu thập ca tử vong của cả ba quốc gia trong 4 tháng 1-8-5-4.*



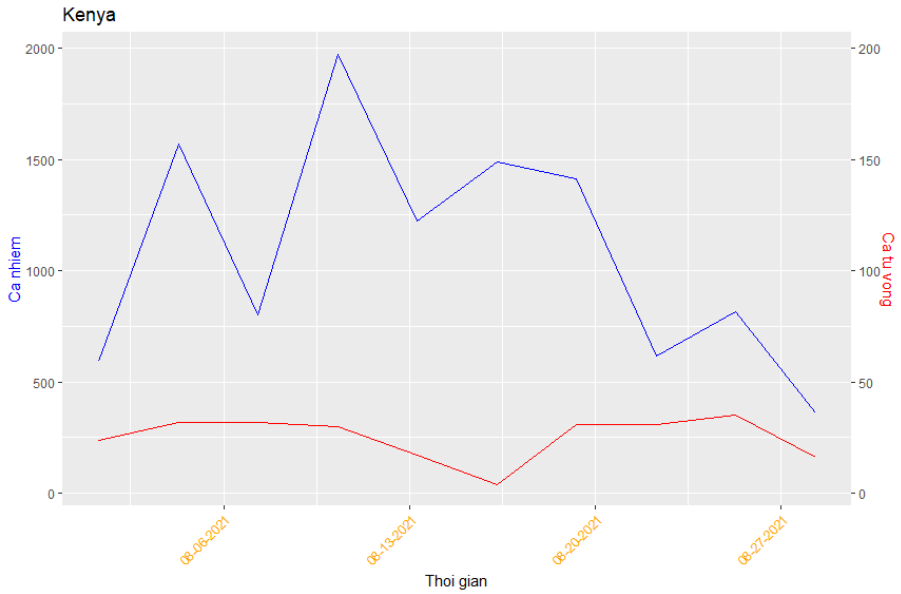
Hình 54: *Biểu đồ thu thập ca tử vong của cả ba quốc gia trong 4 tháng 1-8-5-4.*

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

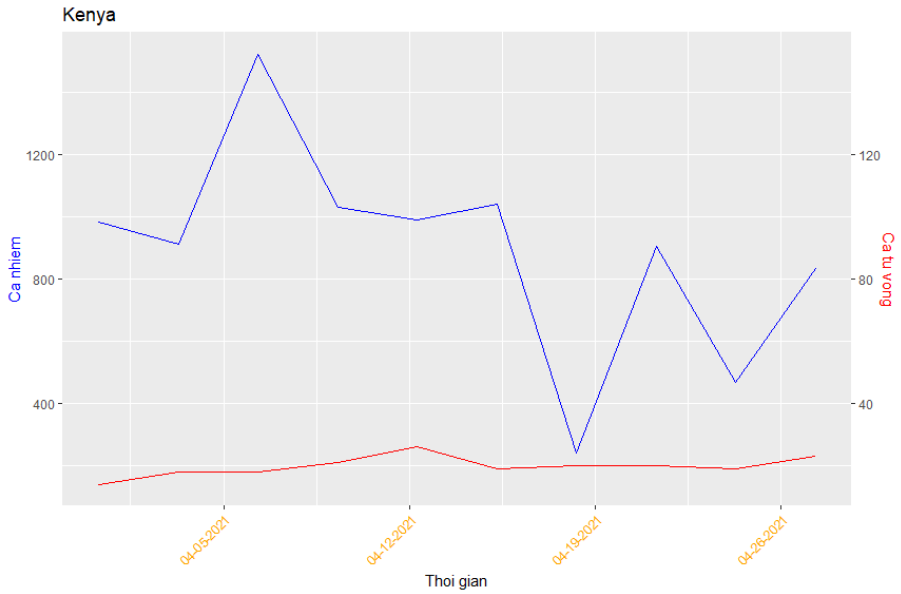
- x_i: số ca nhiễm bệnh của ngày d_i
- y_i: số ca tử vong của ngày d_i
- $a_j = x_{ij} \forall i \in M | j \in P$: dữ liệu nhiễm bệnh mỗi tháng của từng quốc gia .
- $a_j = x_{ij} \forall i \in M | j \in P$: dữ liệu tử vong mỗi của 3 từng gia .



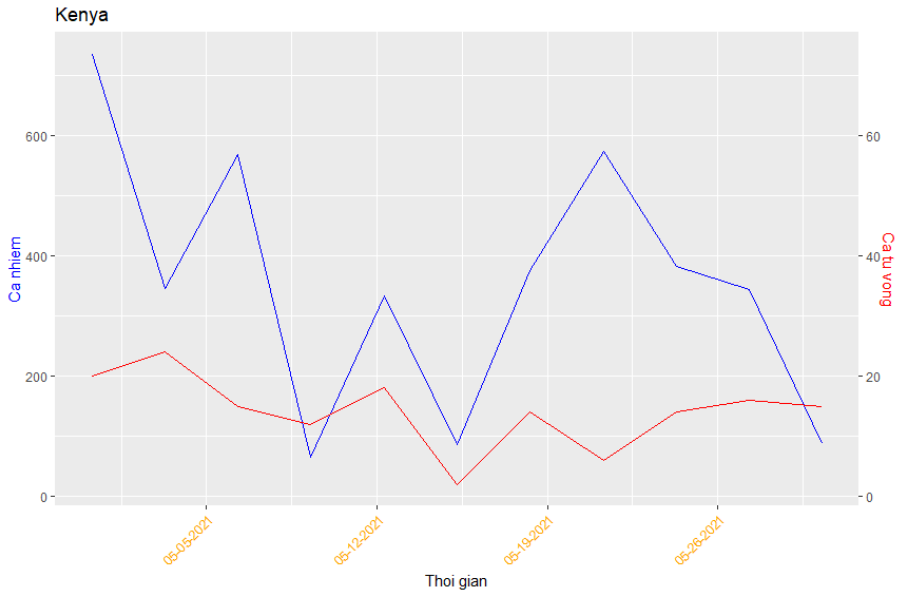
Hình 55: *ữ liệu thu thập ca nhiễm và tử vong từng tháng*



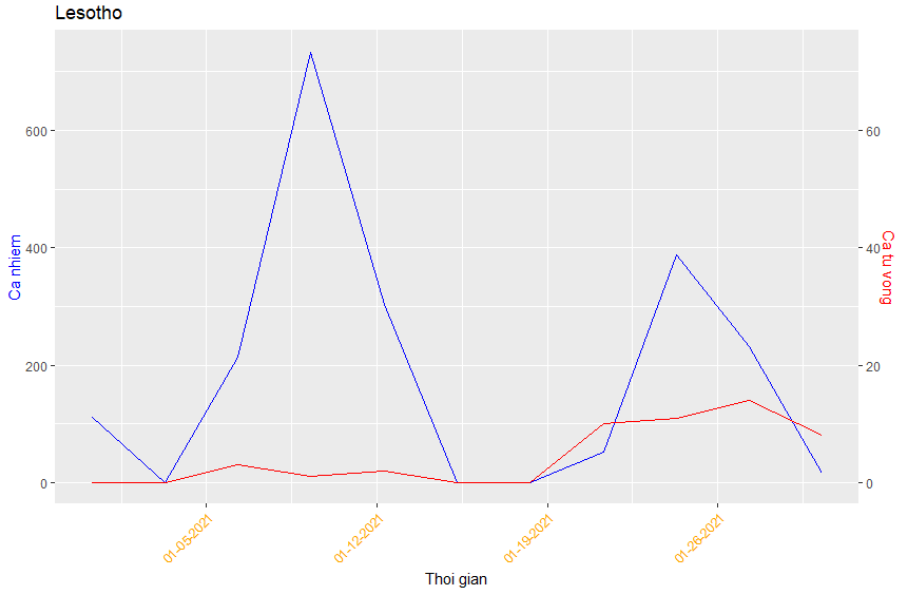
Hình 56: *ữ liệu thu thập ca nhiễm và tử vong từng tháng*



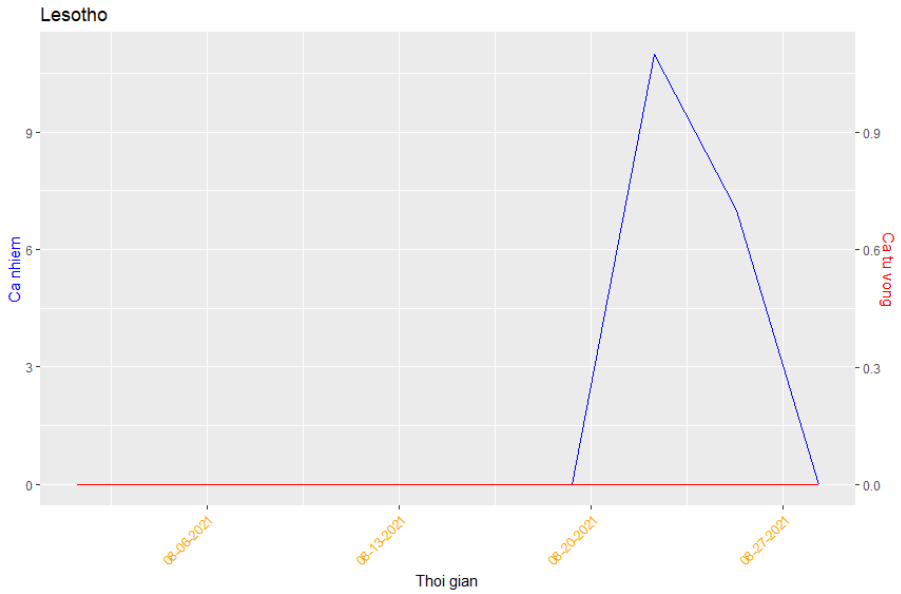
Hình 57: *ữ liệu thu thập ca nhiễm và tử vong từng tháng*



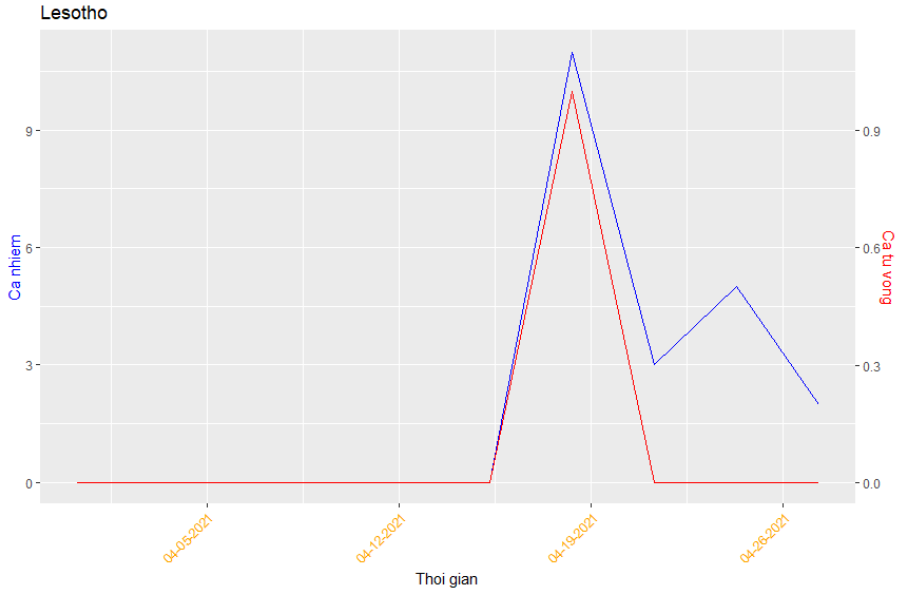
Hình 58: *ữ liệu thu thập ca nhiễm và tử vong từng tháng*



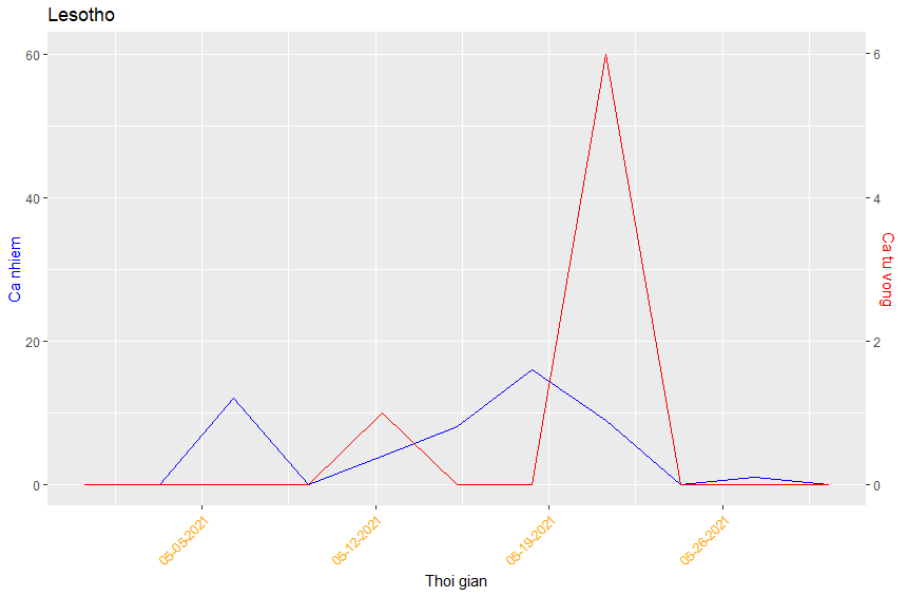
Hình 59: *ữ liệu thu thập ca nhiễm và tử vong từng tháng*



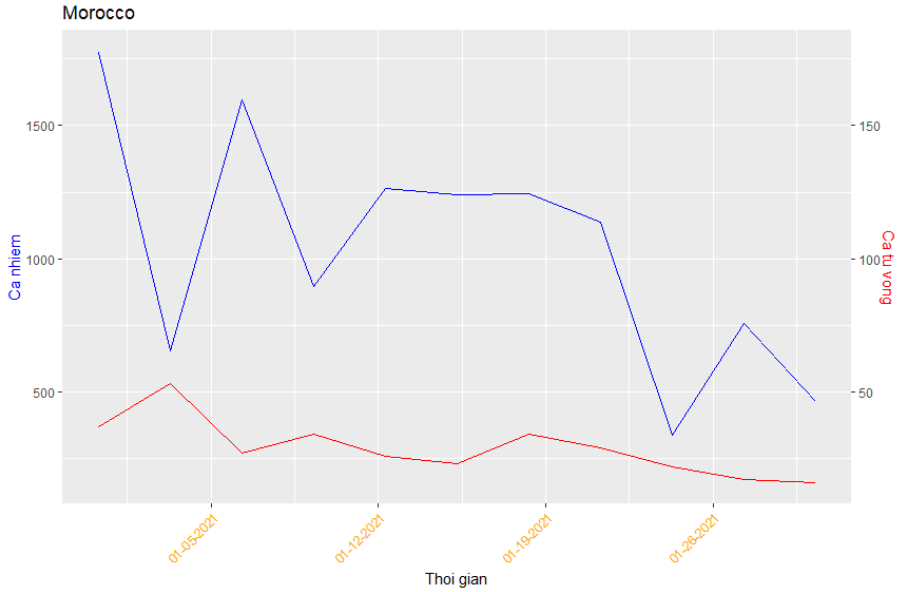
Hình 60: *ữ liệu thu thập ca nhiễm và tử vong từng tháng*



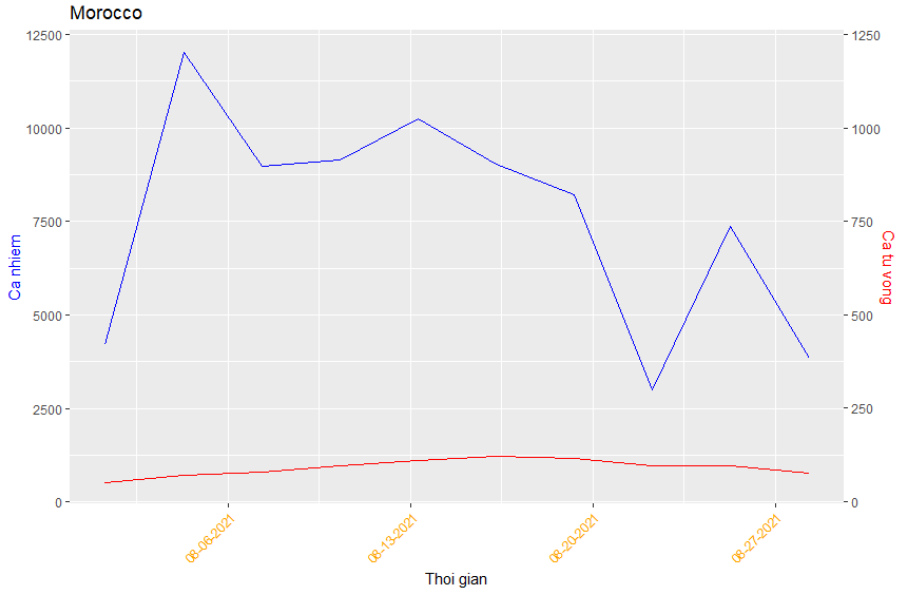
Hình 61: *ữ liệu thu thập ca nhiễm và tử vong từng tháng*



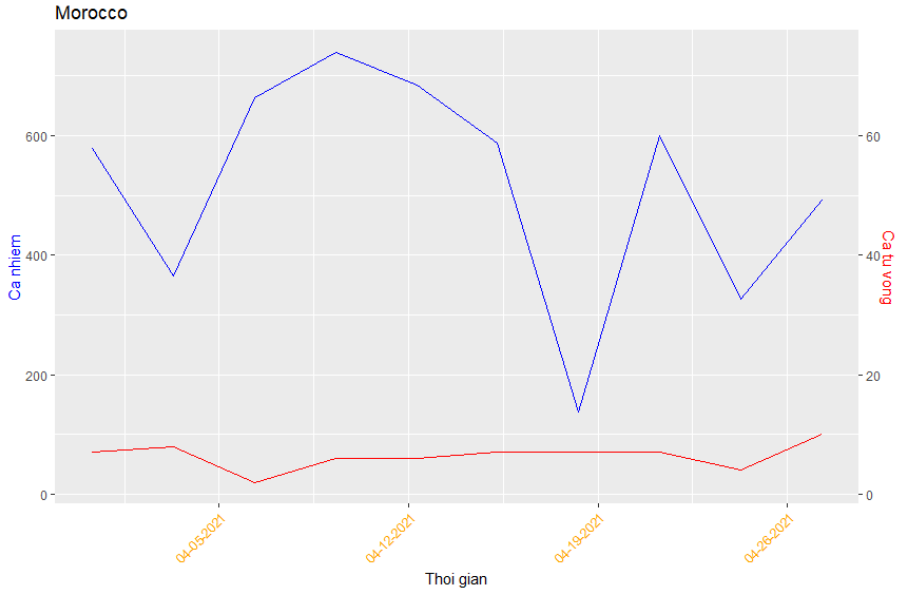
Hình 62: *ữ liệu thu thập ca nhiễm và tử vong từng tháng*



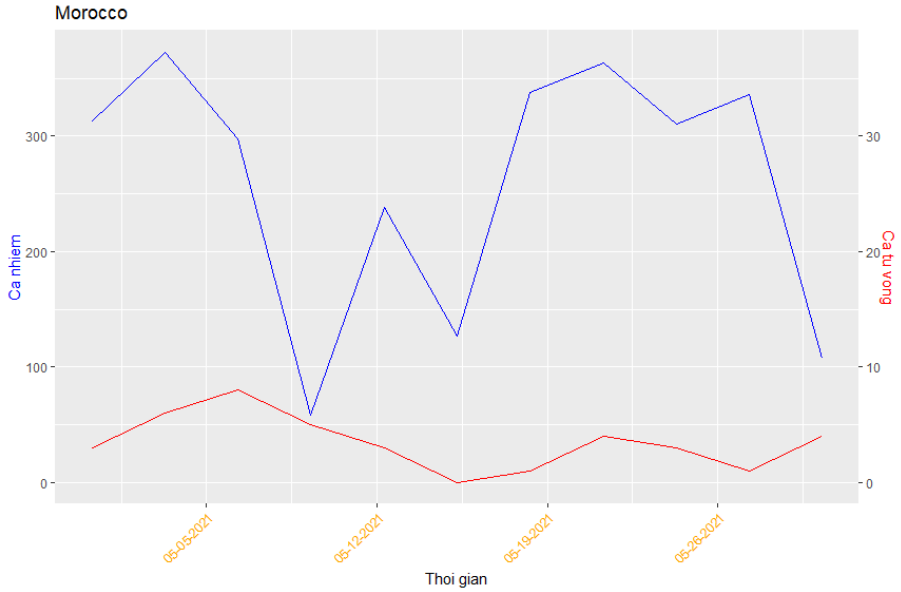
Hình 63: *ữ liệu thu thập ca nhiễm và tử vong từng tháng*



Hình 64: *ữ liệu thu thập ca nhiễm và tử vong từng tháng*



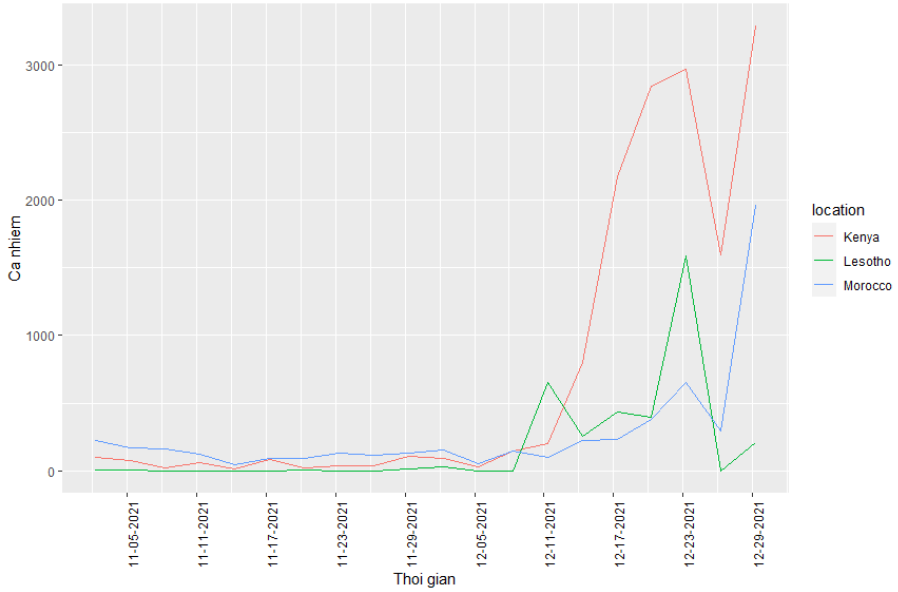
Hình 65: *ữ liệu thu thập ca nhiễm và tử vong từng tháng*



Hình 66: *ữ liệu thu thập ca nhiễm và tử vong từng tháng*

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

- $M = \{11, 12\}$: các tháng thống kê
- x_i: số ca tử vong của ngày d_i
- $a_j = x_i \forall i = M$: dữ liệu nhiễm bệnh theo 2 tháng cuôi năm của 3 quốc gia.



Hình 67: *Dữ liệu thu thập ca nhiễm bệnh 2 tháng cuối năm*

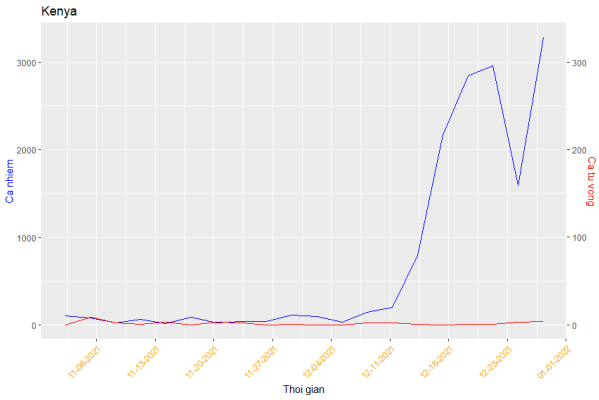
5 Biểu đồ thể hiện thu thập dữ liệu tử vong gồm 2 tháng cuối của năm

- $M = \{11, 12\}$: các tháng thống kê
- x_i: số ca tử vong của ngày d_i
- $b_j = x_i \forall i = M$: dữ liệu tử vong theo 2 tháng cuôi năm của 3 quốc gia.

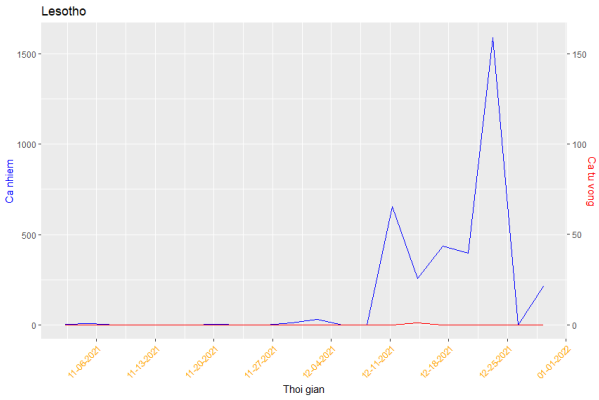
Hình 68: *Dữ liệu thu thập ca tử vong 2 tháng cuối năm*

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

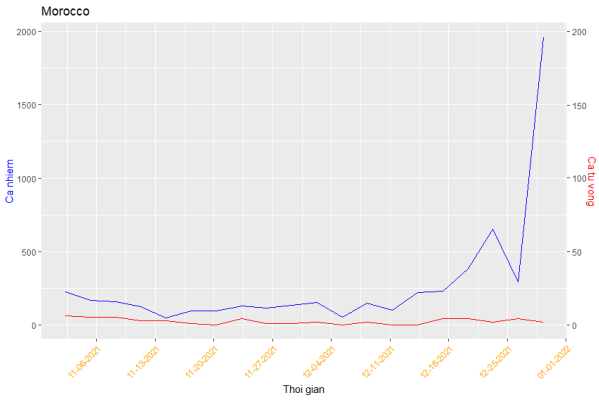
- $M = \{11, 12\}$: các tháng thống kê
- x_i, y_i: số ca nhiễm bệnh, tử vong của ngày d_i
- $a_j = x_{ij} \forall i = M | j \in P$: dữ liệu nhiễm bệnh mỗi tháng của từng quốc gia .
- $a_j = x_{ij} \forall i = M | j \in P$: dữ liệu tử vong mỗi của 3 từng gia .



Hình 69: *Biểu đồ ca nhiễm bệnh và tử vong 2 tháng cuối năm*



Hình 70: *Biểu đồ ca nhiễm bệnh và tử vong 2 tháng cuối năm*



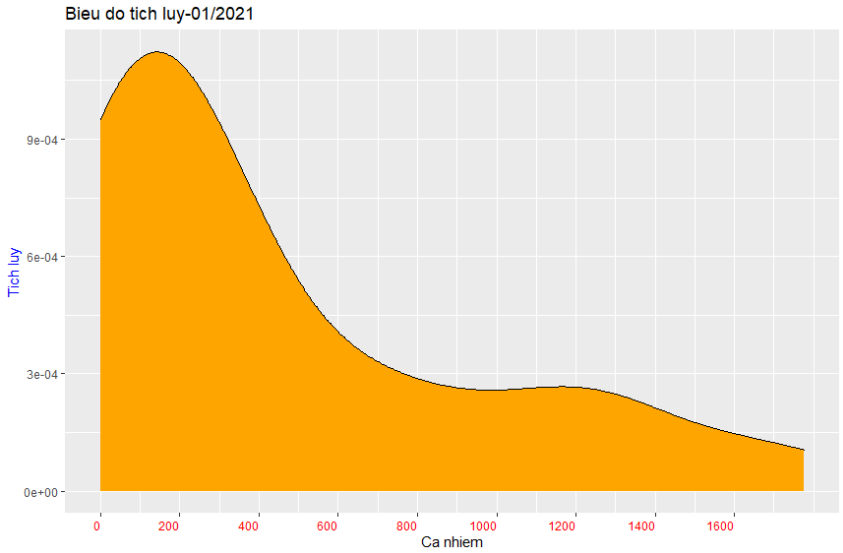
Hình 71: *Biểu đồ ca nhiễm bệnh và tử vong 2 tháng cuối năm*

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

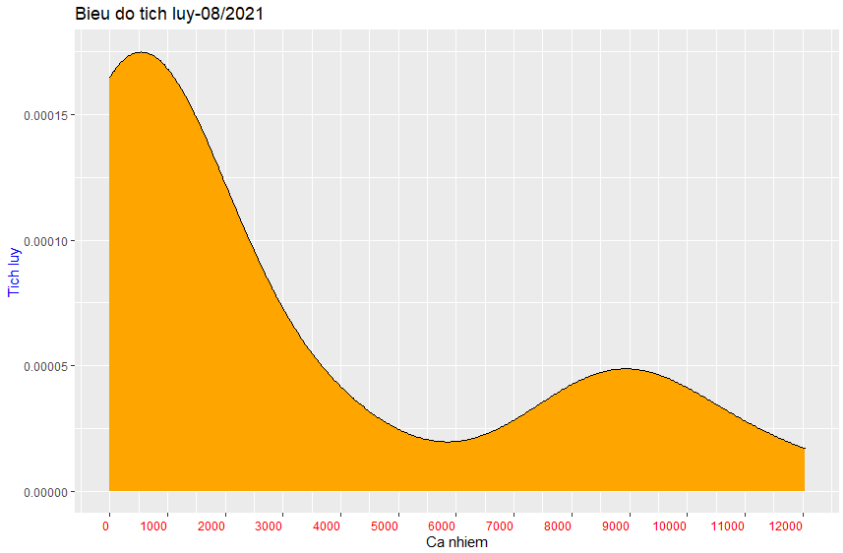
- Tìm dữ liệu nhiễm bệnh tích lũy cho từng tháng

$$A = \frac{\sum_{i=1}^{n} ai*i}{\sum_{i=1}^{n} n}$$

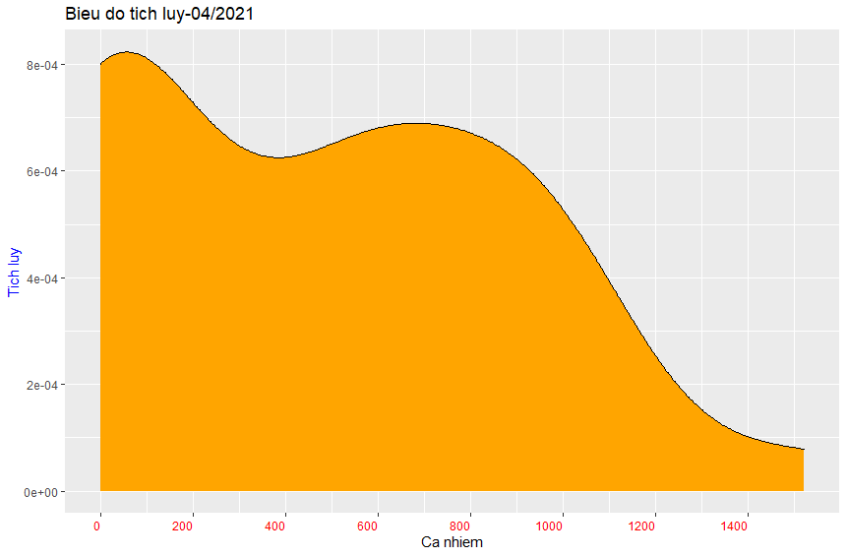
- A:giá trị tích lũy của dữ liệu
- i: ngày thứ i của tháng
- ai: dữ liệu ca nhiễm thu thập được của ngày thứ i
- n: tổng số ca nhiễm trong tháng



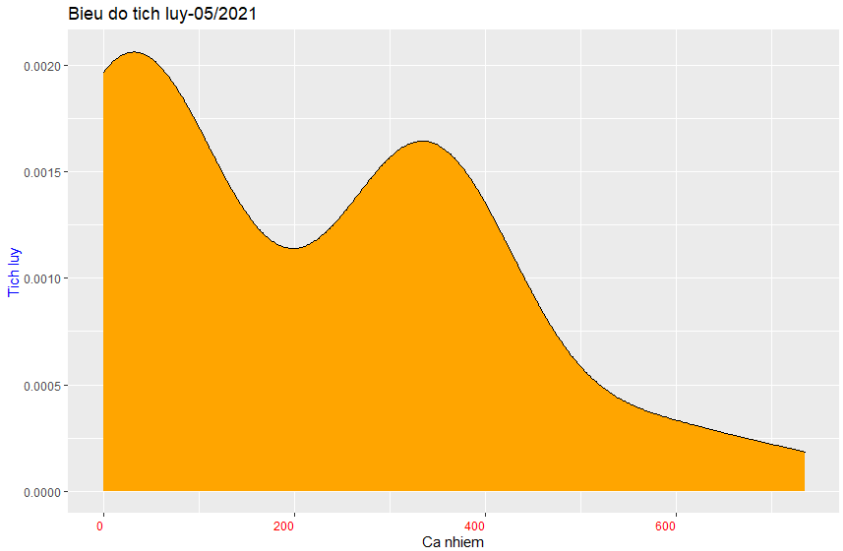
Hình 72: *Biểu đồ thu thập dữ liểu ca nhiễm tích lũy trong mỗi tháng của 3 quốc gia*



Hình 73: *Biểu đồ thu thập dữ liểu ca nhiễm tích lũy trong mỗi tháng của 3 quốc gia*



Hình 74: *Biểu đồ thu thập dữ liểu ca nhiễm tích lũy trong mỗi tháng của 3 quốc gia*



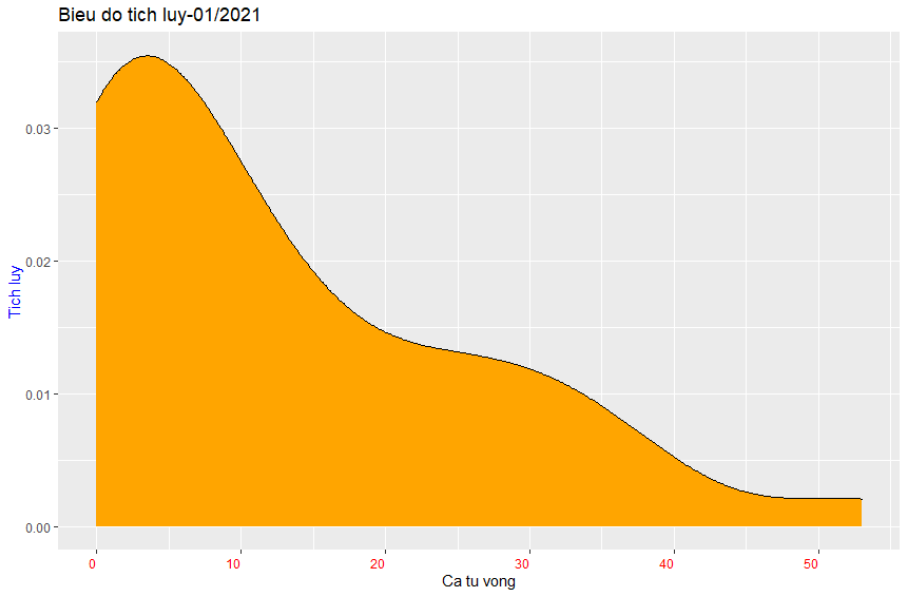
Hình 75: *Biểu đồ thu thập dữ liểu ca nhiễm tích lũy trong mỗi tháng của 3 quốc gia*

8 Biểu đồ thu thập tử vong tích lũy cho từng tháng

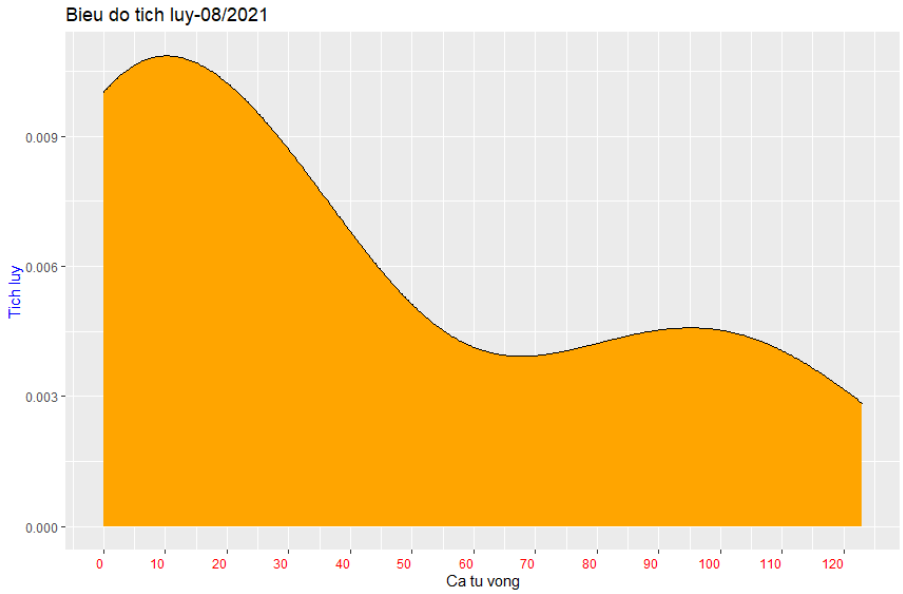
- Tìm dữ liệu nhiễm bệnh tích lũy cho từng tháng

$$A = \frac{\sum_{i=1}^{n} ai*i}{\sum_{i=1}^{n} n}$$

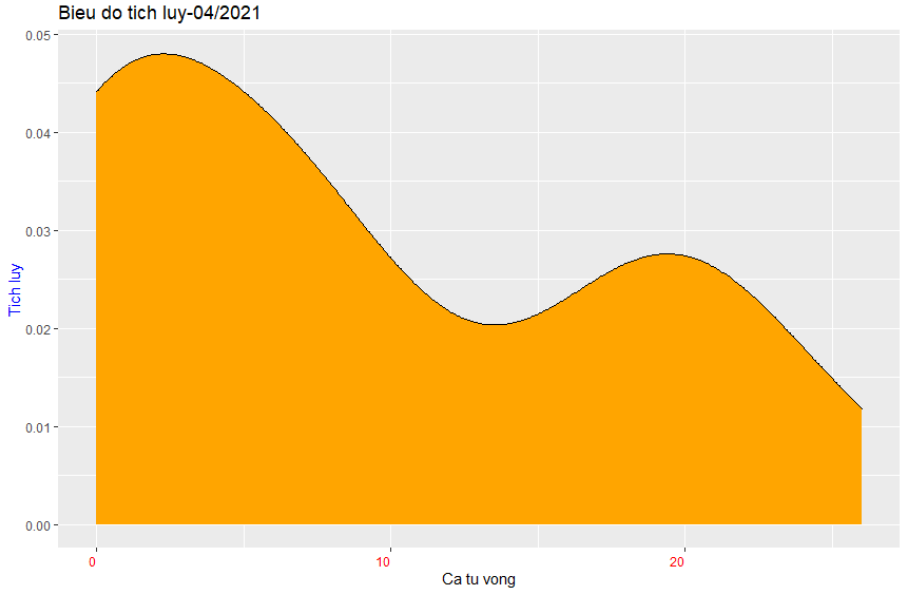
- A:giá trị tích lũy của dữ liệu
- i: ngày thứ i của tháng
- ai: dữ liệu ca nhiễm thu thập được của ngày thứ i
- n: tổng số ca tử vong trong tháng



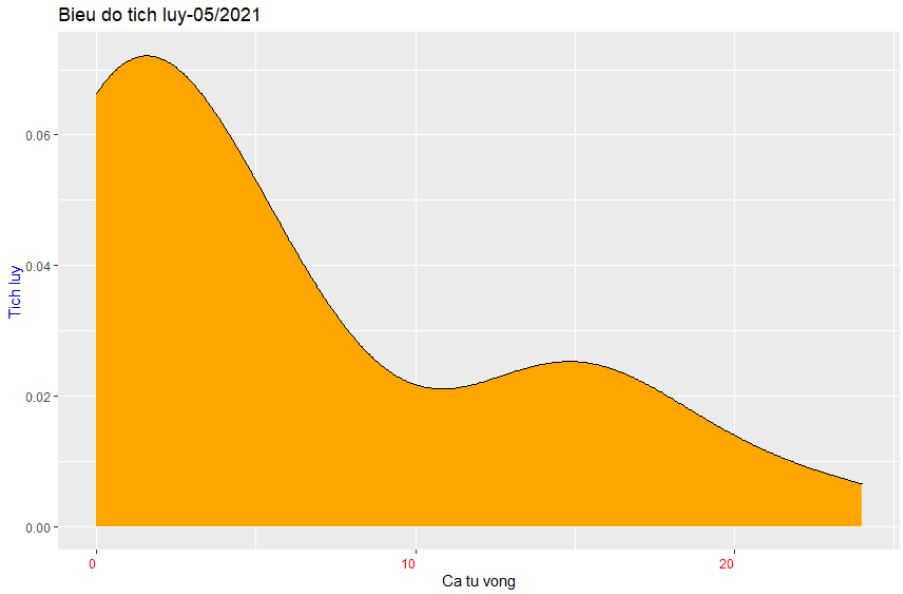
Hình 76: *Biểu đồ thu thập dữ liểu ca tử vong tích lũy trong mỗi tháng của 3 quốc gia*



Hình 77: *Biểu đồ thu thập dữ liểu ca tử vong tích lũy trong mỗi tháng của 3 quốc gia*



Hình 78: *Biểu đồ thu thập dữ liểu ca tử vong tích lũy trong mỗi tháng của 3 quốc gia*



Hình 79: *Biểu đồ thu thập dữ liểu ca tử vong tích lũy trong mỗi tháng của 3 quốc gia*

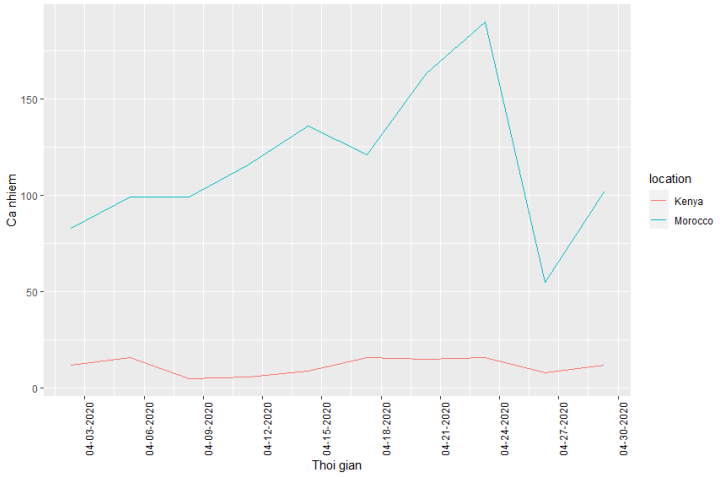
6 Nhóm câu hỏi liên quan đến trực quan dữ liệu theo trung bình 7 ngày gần nhất

Cách giải chung

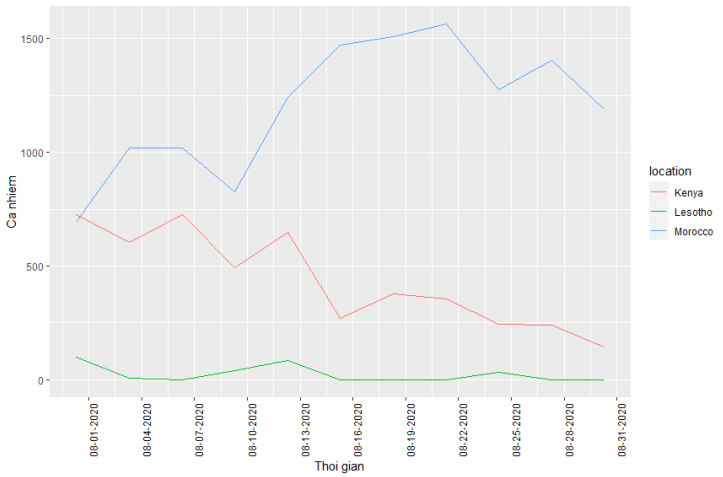
- $A = \{d_i\}$: tập hợp dữ liệu của tất cả các quốc gia
- $P = \{Kenya, Lesotho, Morocco\}$: tập hợp các quốc gia cần thống kê
- $M = \{1, 8, 4, 5\}$: các cần tháng thống kê
- x_i: số ca nhiễm bệnh/tử vong của ngày d_i
- $avg = \frac{\sum_{i=1}^{7} xi}{7}$: số ca nhiễm bệnh/tử vong trung bình trong 7 ngày gần nhất

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

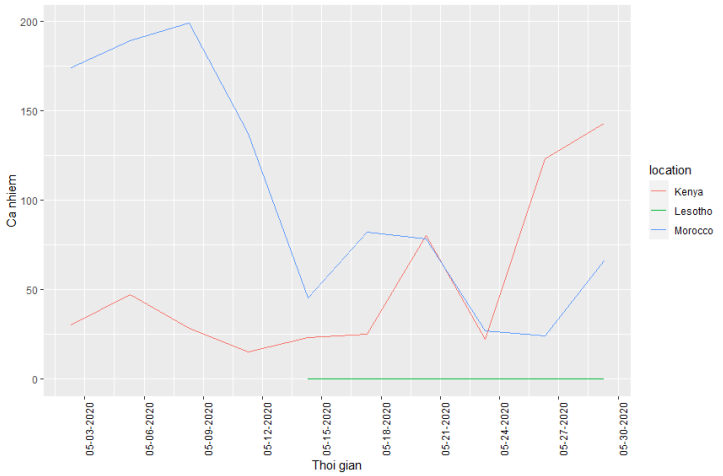
- x_i: số ca nhiễm bệnh của ngày d_i
- $a_j = x_i \forall i \in M$: dữ liệu nhiễm bệnh mỗi tháng của 3 quốc gia.
- $a_i = avg$: Thay thế những báo cáo không thường xuyên bằng giá trị trung bình của 7 ngày gần nhất.
- Vì các số liệu ở các tháng 1-4-5-8/2021 đã được cập nhật liên tục nên biểu đồ sẽ tương tự với câu v.2.



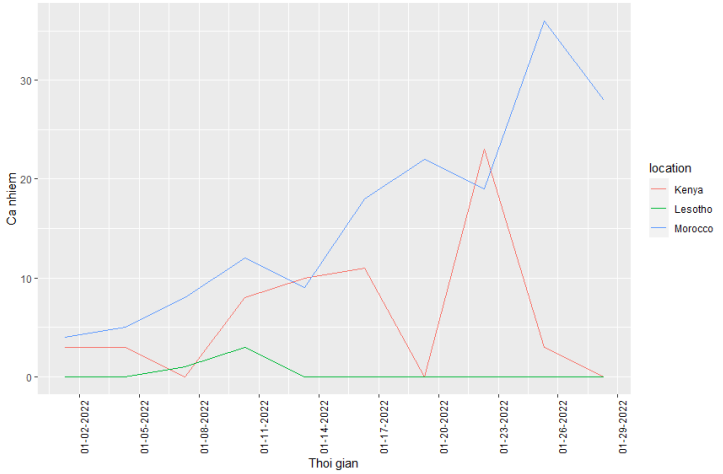
Hình 80: *Biểu đồ thu thập ca nhiễm của cả ba quốc gia trong 4 tháng 1-8-5-4*



Hình 81: *Biểu đồ thu thập ca nhiễm của cả ba quốc gia trong 4 tháng 1-8-5-4*



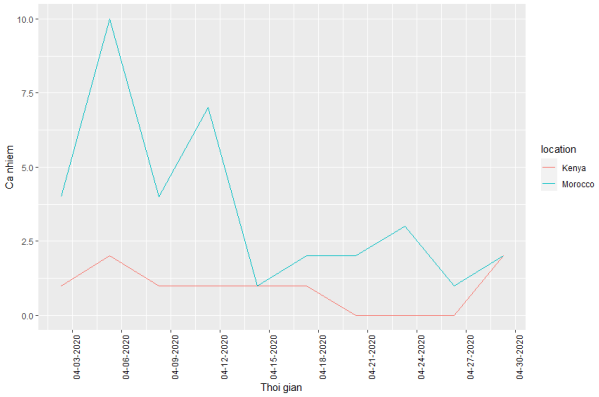
Hình 82: *Biểu đồ thu thập ca nhiễm của cả ba quốc gia trong 4 tháng 1-8-5-4*



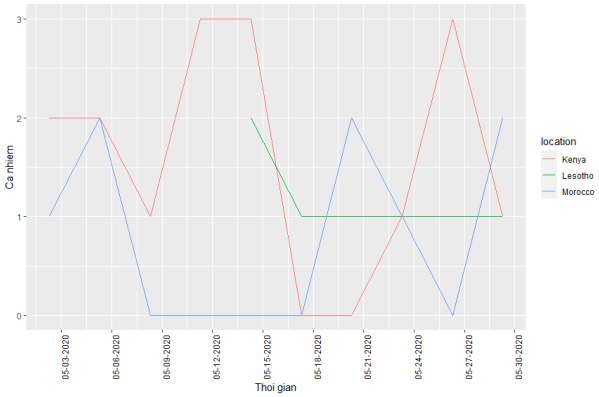
Hình 83: *Biểu đồ thu thập ca nhiễm của cả ba quốc gia trong 4 tháng 1-8-5-4*

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

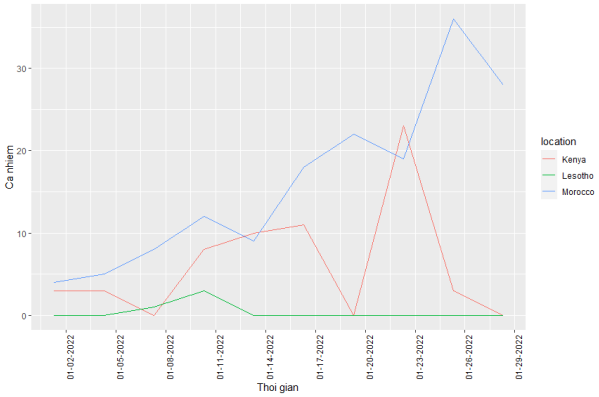
- x_i: số ca nhiễm bệnh của ngày d_i
- $a_j = x_i \forall i \in M$: dữ liệu tử vong mỗi tháng của 3 quốc gia.
- $a_i = avg$: Thay thế những báo cáo không thường xuyên bằng giá trị trung bình của 7 ngày gần nhất.
- Vì các số liệu ở các tháng 1-4-5-8/2021 đã được cập nhật liên tục nên biểu đồ sẽ tương tự với câu v.3.



Hình 84: *Biểu đồ thu thập ca tử vong của cả ba quốc gia trong 4 tháng 1-8-5-4*



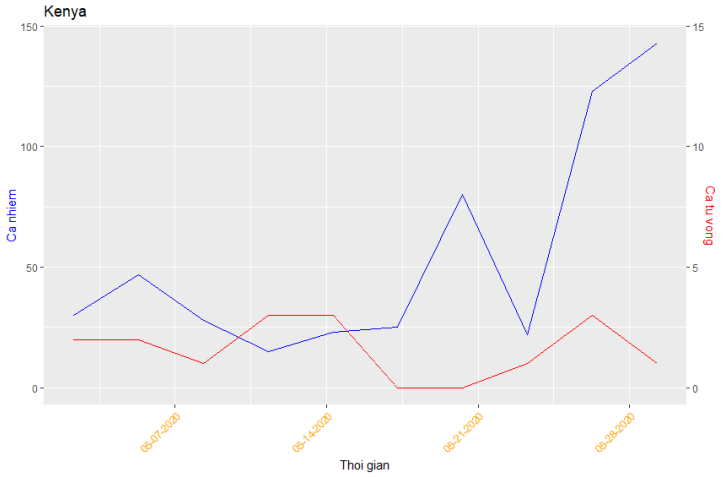
Hình 85: *Biểu đồ thu thập ca tử vong của cả ba quốc gia trong 4 tháng 1-8-5-4*



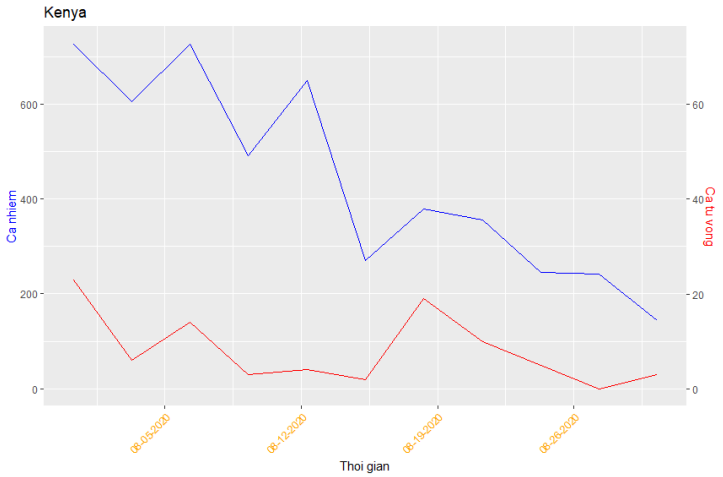
Hình 86: *Biểu đồ thu thập ca tử vong của cả ba quốc gia trong 4 tháng 1-8-5-4*

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

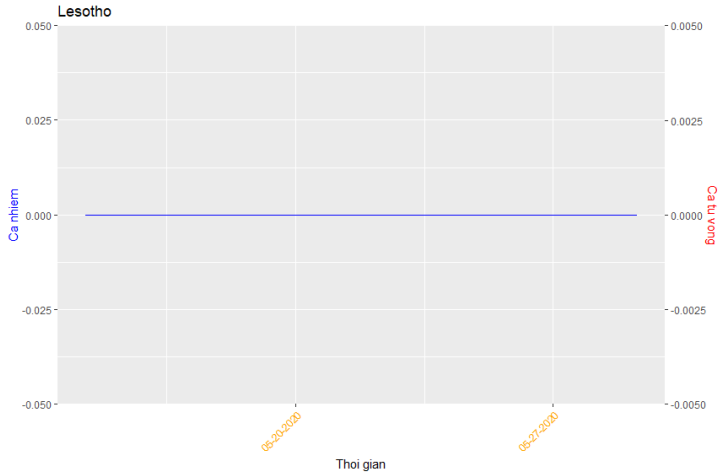
- x_i: số ca nhiễm bệnh của ngày d_i
- y_i: số ca tử vong của ngày d_i
- $a_j = x_{ij} \forall i \in M | j \in P$: dữ liệu nhiễm bệnh mỗi tháng của từng quốc gia .
- $a_j = x_{ij} \forall i \in M | j \in P$: dữ liệu tử vong mỗi của 3 từng gia .
- $a_i / b_i = avg$: Thay thế những báo cáo không thường xuyên bằng giá trị trung bình của 7 ngày gần nhất.



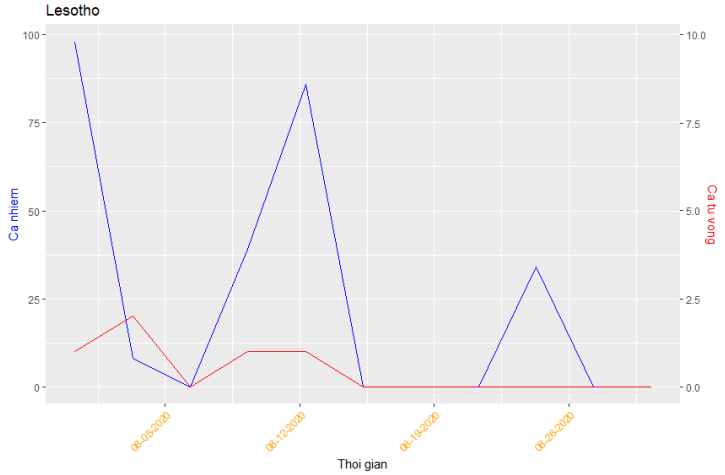
Hình 87: *Biểu đồ dữ liệu thu thập ca nhiễm tử vong của từng quốc gia theo từng tháng theo trung bình 7 ngày gần nhất*



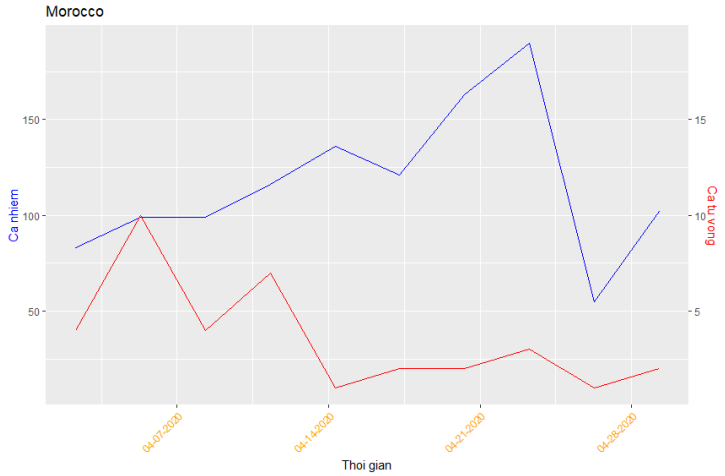
Hình 88: *Biểu đồ dữ liệu thu thập ca nhiễm tử vong của từng quốc gia theo từng tháng theo trung bình 7 ngày gần nhất*



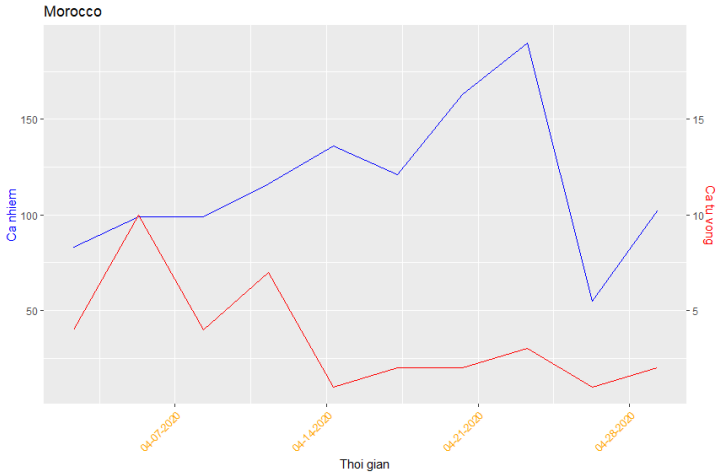
Hình 89: *Biểu đồ dữ liệu thu thập ca nhiễm tử vong của từng quốc gia theo từng tháng theo trung bình 7 ngày gần nhất*



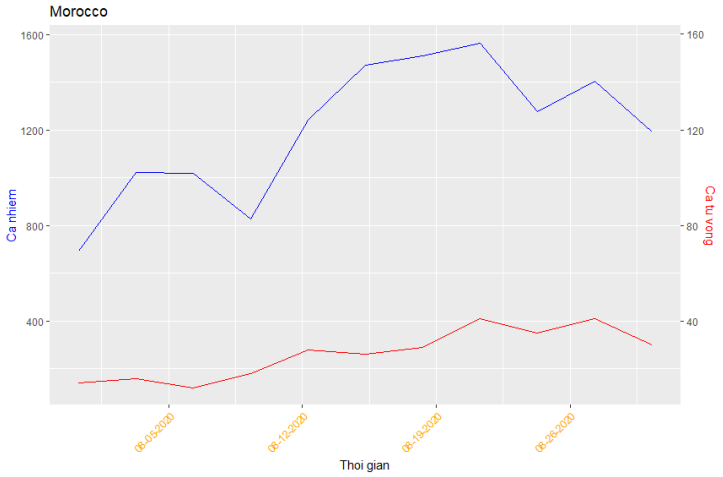
Hình 90: *Biểu đồ dữ liệu thu thập ca nhiễm tử vong của từng quốc gia theo từng tháng theo trung bình 7 ngày gần nhất*



Hình 91: *Biểu đồ dữ liệu thu thập ca nhiễm tử vong của từng quốc gia theo từng tháng theo trung bình 7 ngày gần nhất*



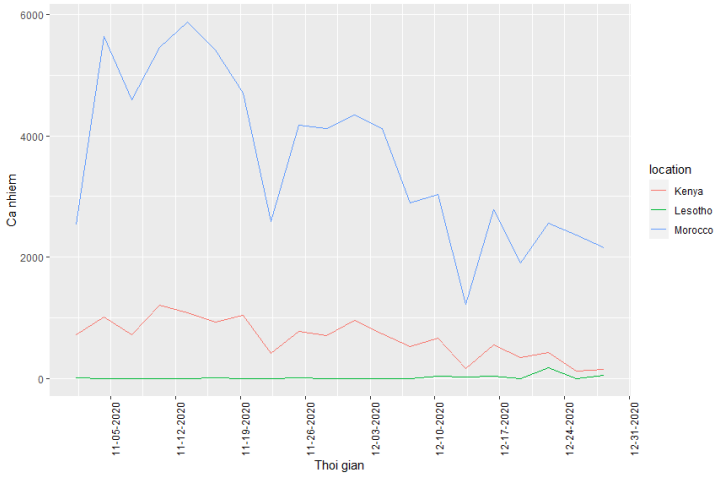
Hình 92: *Biểu đồ dữ liệu thu thập ca nhiễm tử vong của từng quốc gia theo từng tháng theo trung bình 7 ngày gần nhất*



Hình 93: *Biểu đồ dữ liệu thu thập ca nhiễm tử vong của từng quốc gia theo từng tháng theo trung bình 7 ngày gần nhất*

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

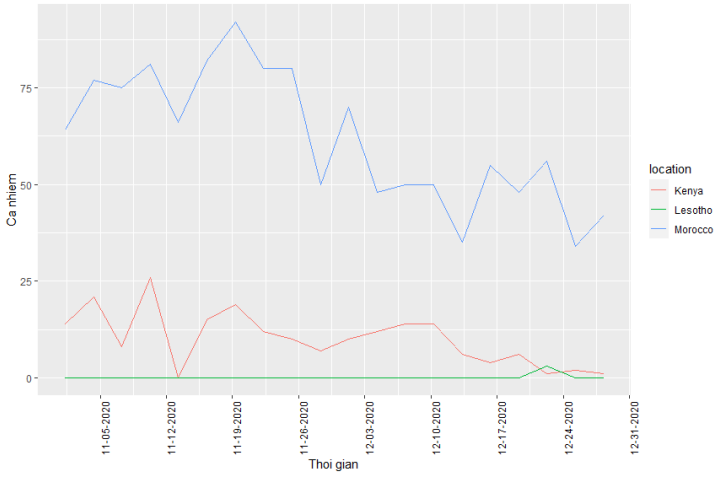
- $M = \{11, 12\}$: các tháng thống kê
- x_i: số ca tử vong của ngày d_i
- $a_j = x_i \forall i = M$: dữ liệu nhiễm bệnh theo 2 tháng cuôi năm của 3 quốc gia.
- $a_i = avg$: Thay thế những báo cáo không thường xuyên bằng giá trị trung bình của 7 ngày gần nhất.
- Vì các số liệu ở các tháng 1-4-5-8/2021 đã được cập nhật liên tục nên biểu đồ sẽ tương tự với câu v.4.



Hình 94: *Dữ liệu thu thập ca nhiễm bệnh 2 tháng cuối năm*

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

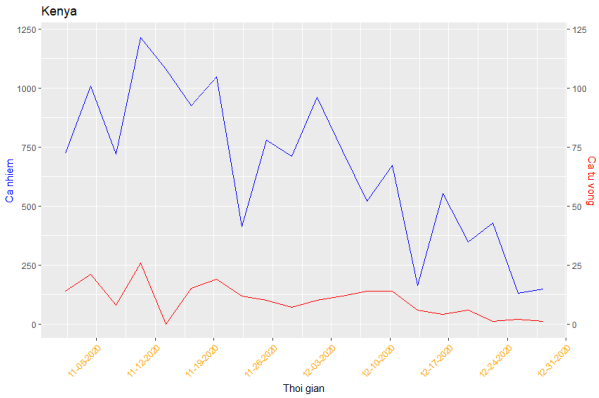
- $M = \{11, 12\}$: các tháng thống kê
- x_i: số ca tử vong của ngày d_i
- $b_j = x_i \forall i = M$: dữ liệu tử vong theo 2 tháng cuôi năm của 3 quốc gia.
- $b_i = avg$: Thay thế những báo cáo không thường xuyên bằng giá trị trung bình của 7 ngày gần nhất.
- Vì các số liệu ở các tháng 1-4-5-8/2021 đã được cập nhật liên tục nên biểu đồ sẽ tương tự với câu v.5.



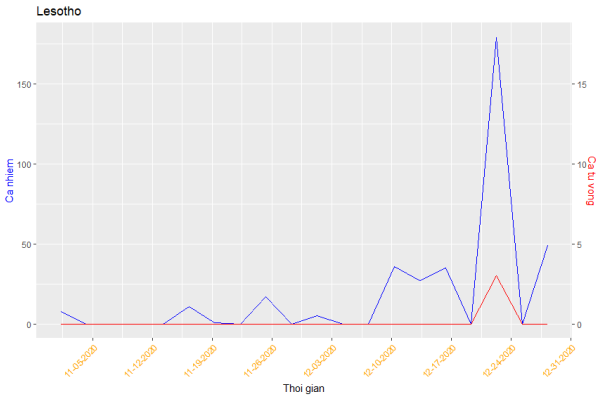
Hình 95: *Dữ liệu thu thập ca tử vong 2 tháng cuối năm*

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

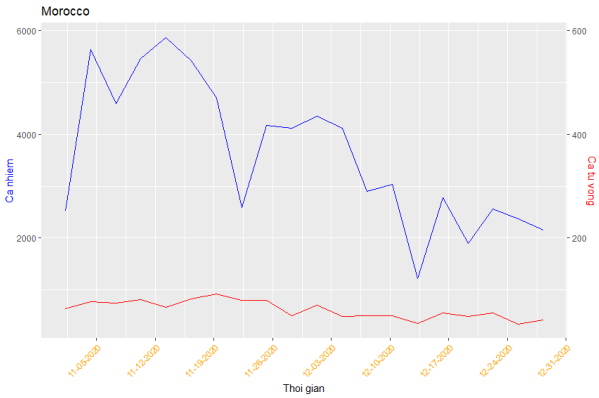
- $M = \{11, 12\}$: các tháng thống kê
- x_i: số ca nhiễm bệnh của ngày d_i
- y_i: số ca tử vong của ngày d_i
- $a_j = x_{ij} \forall i = M | j \in P$: dữ liệu nhiễm bệnh mỗi tháng của từng quốc gia .
- $b_j = y_{ij} \forall i = M | j \in P$: dữ liệu tử vong mỗi của 3 từng gia .
- Vì các số liệu ở các tháng 1-4-5-8/2021 đã được cập nhật liên tục nên biểu đồ sẽ tương tự với câu v.6.



Hình 96: *Biểu đồ ca nhiễm bệnh và tử vong 2 tháng cuối năm*



Hình 97: *Biểu đồ ca nhiễm bệnh và tử vong 2 tháng cuối năm*



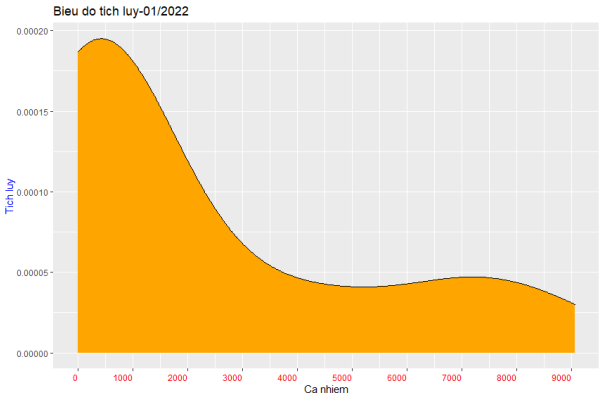
Hình 98: *Biểu đồ ca nhiễm bệnh và tử vong 2 tháng cuối năm*

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

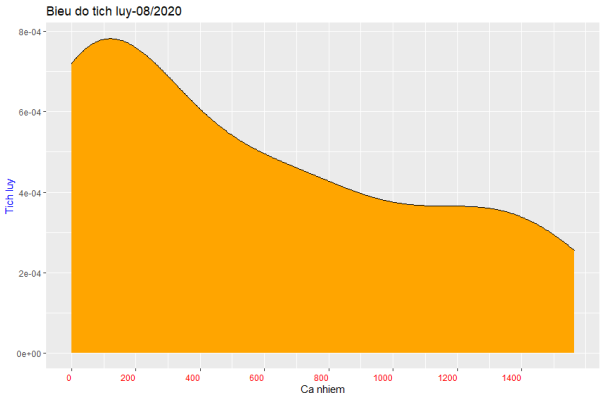
- Tìm dữ liệu nhiễm bệnh tích lũy cho từng tháng

$$A = \frac{\sum_{i=1}^{n} ai*i}{\sum_{i=1}^{n} n}$$

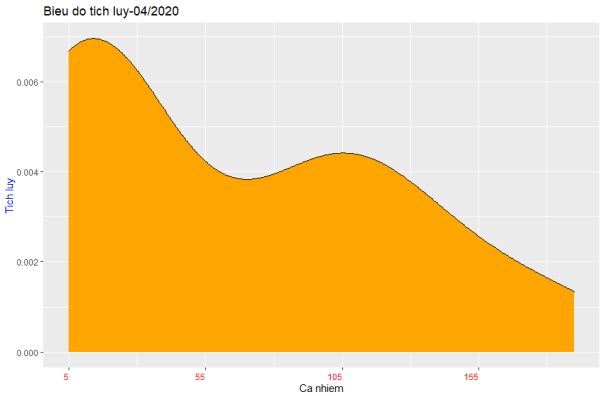
- A:giá trị tích lũy của dữ liệu
- i: ngày thứ i của tháng
- ai: dữ liệu ca nhiễm thu thập được của ngày thứ i
- $a_i = avg$: thay thế giá trị không thường xuyên bằng trung bình 7 ngày gần nhất
- n: tổng số ca nhiễm trong tháng
- Vì các số liệu ở các tháng 1-4-5-8/2021 đã được cập nhật liên tục nên biểu đồ sẽ tương tự với câu v.7.



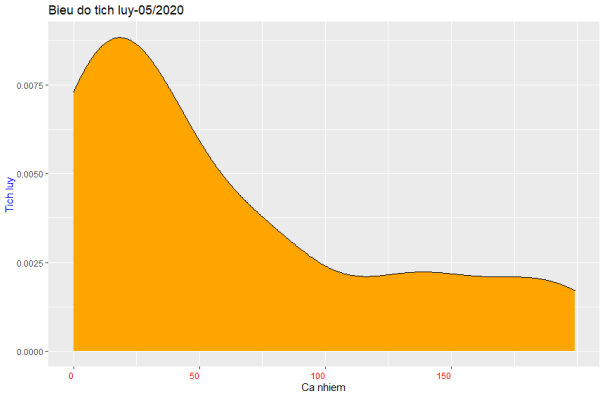
Hình 99: *Biểu đồ thu thập dữ liểu ca nhiễm tích lũy trong mỗi tháng của 3 quốc gia*



Hình 100: *Biểu đồ thu thập dữ liểu ca nhiễm tích lũy trong mỗi tháng của 3 quốc gia*



Hình 101: *Biểu đồ thu thập dữ liểu ca nhiễm tích lũy trong mỗi tháng của 3 quốc gia*



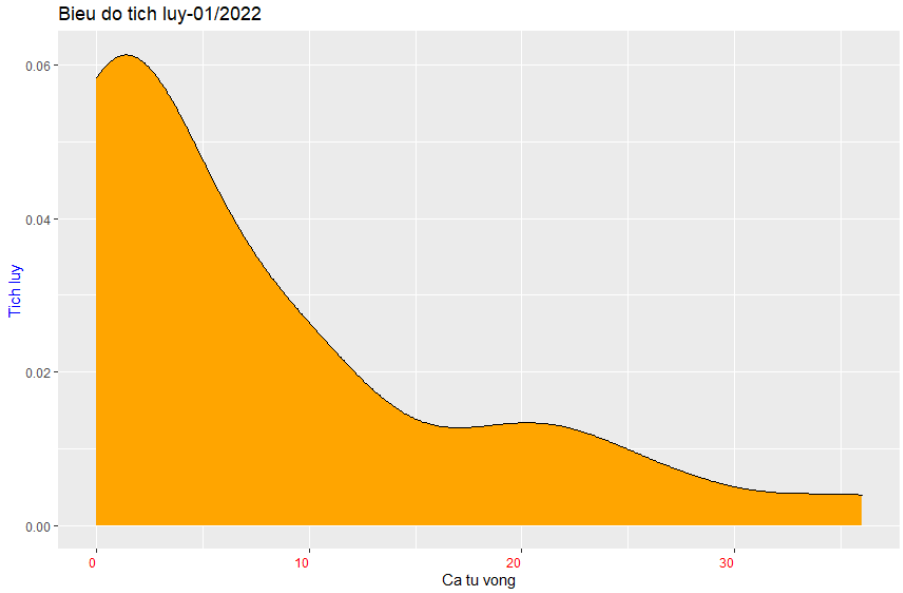
Hình 102: *Biểu đồ thu thập dữ liểu ca nhiễm tích lũy trong mỗi tháng của 3 quốc gia*

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

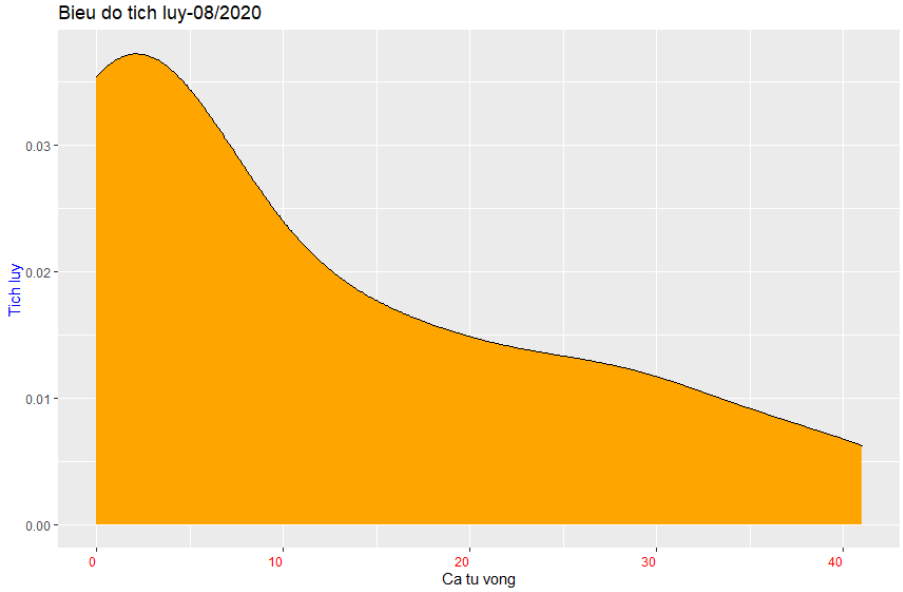
- Tìm dữ liệu nhiễm bệnh tích lũy cho từng tháng

$$A = \frac{\sum_{i=1}^{n} bi*i}{\sum_{i=1}^{n} n}$$

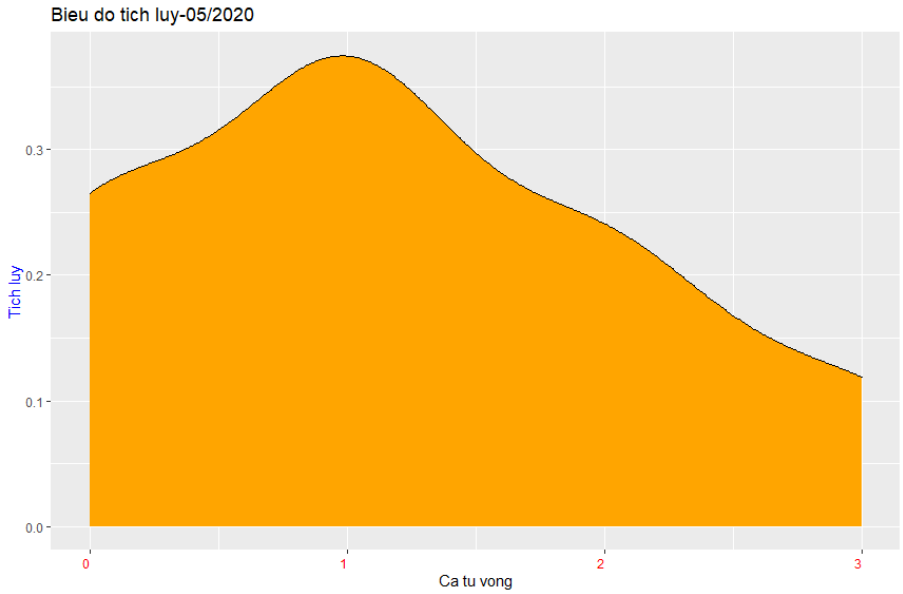
- A:giá trị tích lũy của dữ liệu
- i: ngày thứ i của tháng
- bi: dữ liệu ca nhiễm thu thập được của ngày thứ i
- $b_i = avg$: thay thế giá trị không thường xuyên bằng trung bình 7 ngày gần nhất
- n: tổng số ca tử vong trong tháng
- Vì số liệu ở các tháng 1-4-5-8/2021 đã được cập nhật liên tục nên biểu đồ tương tự với câu v.8.



Hình 103: *Biểu đồ thu thập dữ liểu tử vong tích lũy trong mỗi tháng của 3 quốc gia*



Hình 104: *Biểu đồ thu thập dữ liểu tử vong tích lũy trong mỗi tháng của 3 quốc gia*

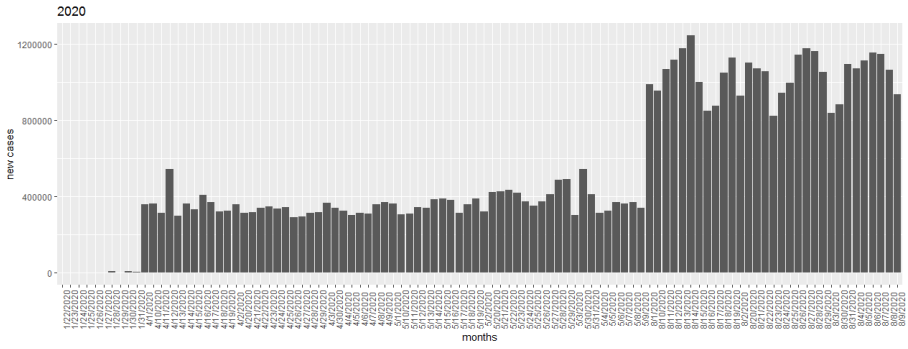


Hình 105: *Biểu đồ thu thập dữ liểu tử vong tích lũy trong mỗi tháng của 3 quốc gia*

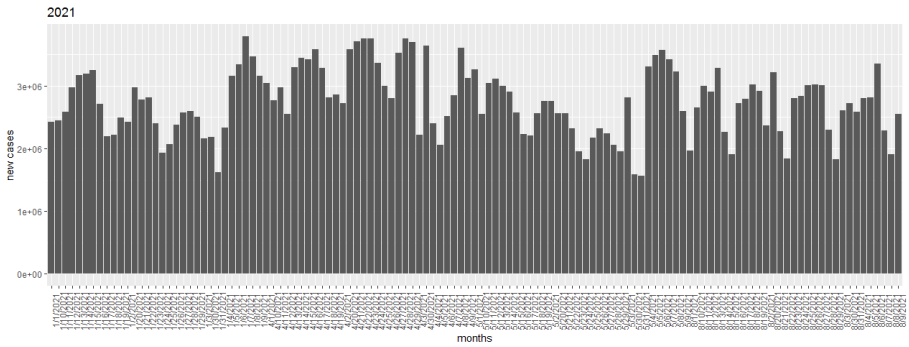
7 Nhóm câu hỏi liên quan đến tất cả quốc gia theo thời gian là tháng

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

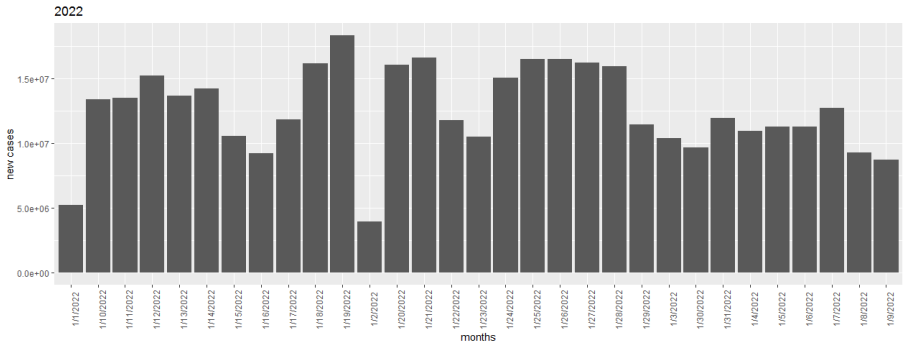
- $D = \{d_i\}$: tập hợp các ngày thu thập dữ liệu của tất cả quốc gia
- $f : D \to Z$: $f(x)$ là hàm tìm tháng từ ngày x
- $A = \{1, 8, 4, 5\}$: các tháng thống kê
- x_i: số ca nhiễm bệnh của ngày d_i
- $a_j = x_i \forall z_j \in A$: dữ liệu nhiễm bệnh tháng 1, 8, 4, 5 của tất cả quốc gia



Hình 106: *Dữ liệu nhiễm bệnh tất cả quốc gia năm 2020*



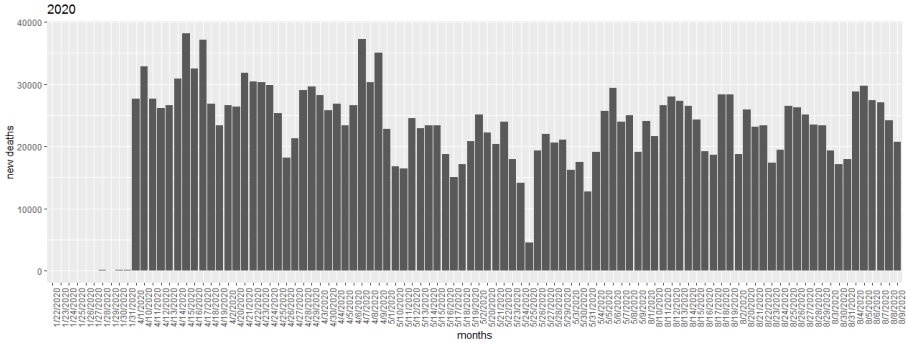
Hình 107: *Dữ liệu nhiễm bệnh tất cả quốc gia năm 2021*



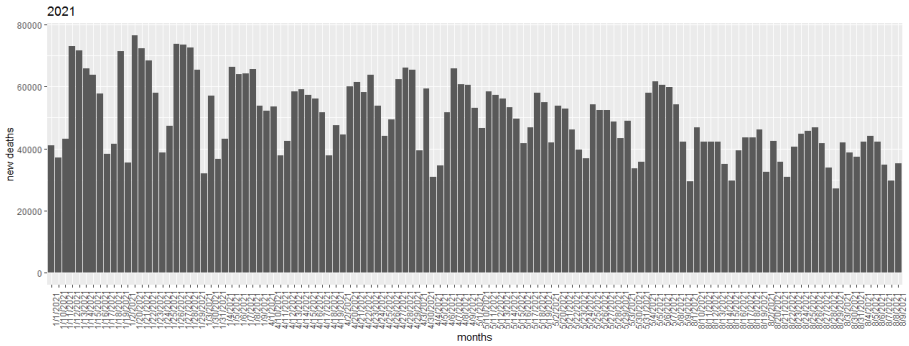
Hình 108: *Dữ liệu nhiễm bệnh tất cả quốc gia năm 2022*

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

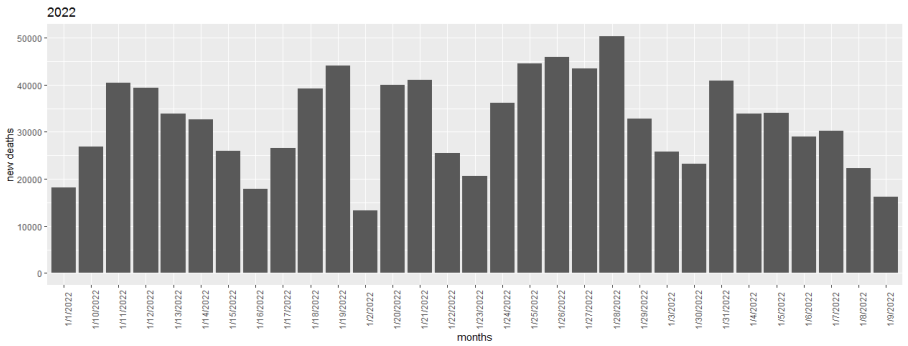
- $D = \{d_i\}$: tập hợp các ngày thu thập dữ liệu của tất cả quốc gia
- $f : D \to Z$: $f(x)$ là hàm tìm tháng từ ngày x
- $A = \{1, 8, 4, 5\}$: các tháng thống kê
- x_i: số ca tử vong của ngày d_i
- $a_j = x_i \forall z_j \in A$: dữ liệu tử vong tháng 1, 8, 4, 5 của tất cả quốc gia



Hình 109: *Dữ liệu tử vong tất cả quốc gia năm 2020*



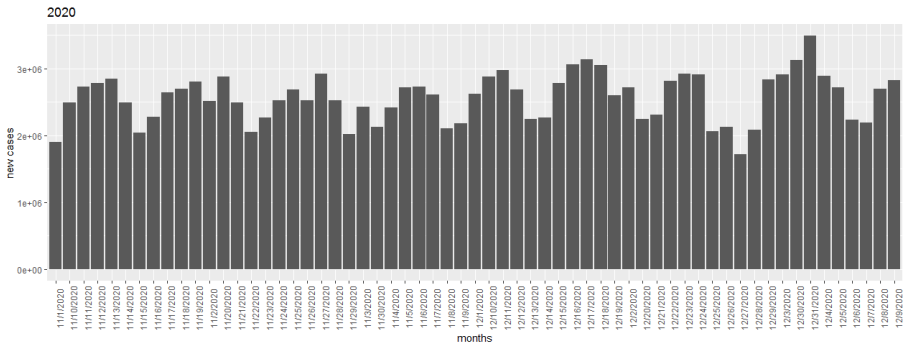
Hình 110: *Dữ liệu tử vong tất cả quốc gia năm 2021*



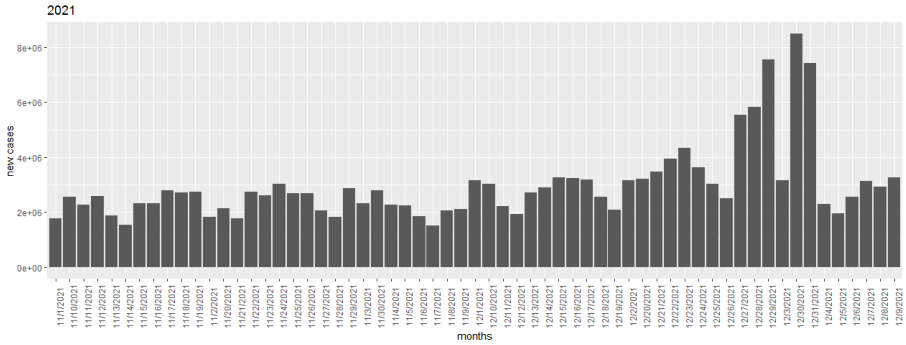
Hình 111: *Dữ liệu tử vong tất cả quốc gia năm 2022*

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

- $D = \{d_i\}$: tập hợp các ngày thu thập dữ liệu của tất cả quốc gia
- $f : D \to Z$: $f(x)$ là hàm tìm tháng từ ngày x
- $A = \{11, 12\}$: các tháng thống kê
- x_i: số ca nhiễm bệnh của ngày d_i
- $a_j = x_i \forall z_j \in A$: dữ liệu nhiễm bệnh tháng 11, 12 của tất cả quốc gia



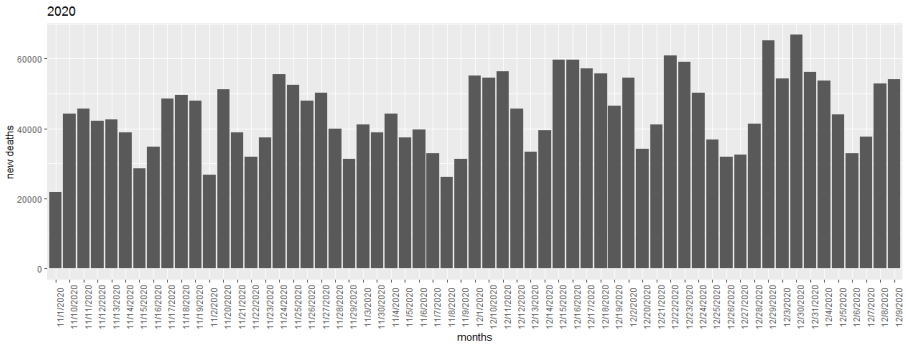
Hình 112: *Dữ liệu nhiễm bệnh 2 tháng cuối năm 2020 của tất cả quốc gia*



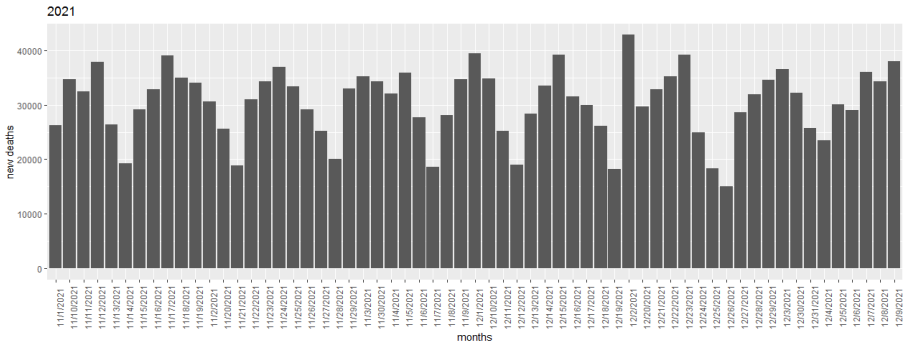
Hình 113: *Dữ liệu nhiễm bệnh 2 tháng cuối năm 2021 của tất cả quốc gia*

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

- $D = \{d_i\}$: tập hợp các ngày thu thập dữ liệu của tất cả quốc gia
- $f : D \to Z$: $f(x)$ là hàm tìm tháng từ ngày x
- $A = \{11, 12\}$: các tháng thống kê
- x_i: số ca tử vong của ngày d_i
- $a_j = x_i \forall z_j \in A$: dữ liệu tử vong tháng 11, 12 của tất cả quốc gia



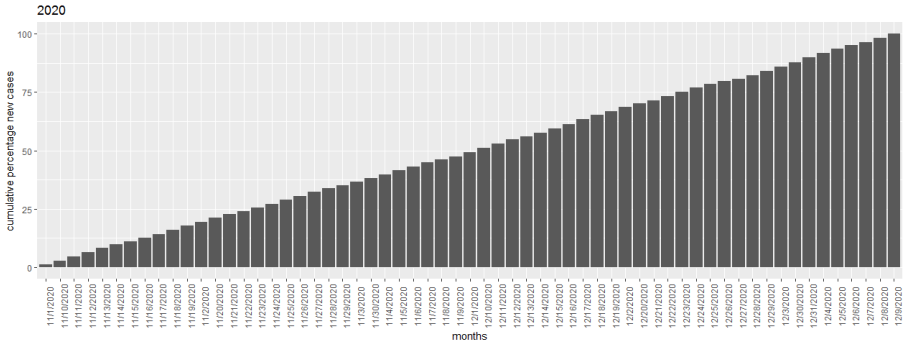
Hình 114: *Dữ liệu tử vong 2 tháng cuối năm 2020 của tất cả quốc gia*



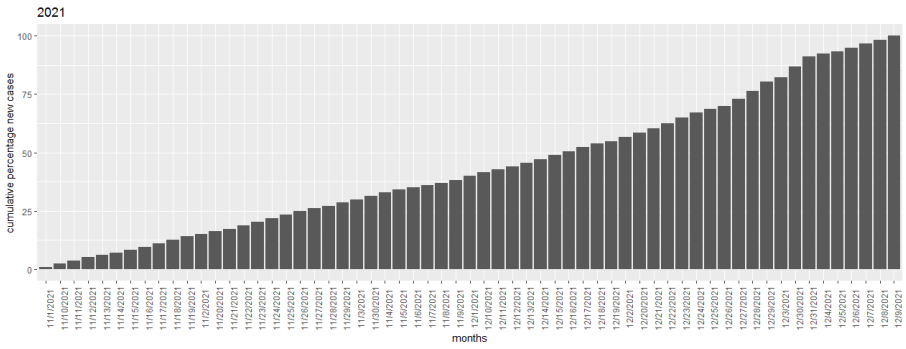
Hình 115: *Dữ liệu tử vong 2 tháng cuối năm 2021 của tất cả quốc gia*

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 gia

- $D = \{d_i\}$: tập hợp các ngày thu thập dữ liệu của tất cả quốc gia
- $f : D \to Z$: $f(x)$ là hàm tìm tháng từ ngày x
- $A = \{11, 12\}$: các tháng thống kê
- x_i: số ca nhiễm bệnh của ngày d_i
- $d_{max} \in \{d_{max} | d_{max} \in A \wedge d_{max} \geq d_i \forall d_i \in A\}$: ngày cuối cùng thu thập dữ liệu
- $d_{min} \in \{d_{min} | d_{min} \in A \wedge d_{min} \leq d_i \forall d_i \in A\}$: ngày đầu thu thập dữ liệu
- $a_j = \frac{\sum_{n=d_{min}}^{d_j} x_j}{\sum_{n=d_{min}}^{d_{max}} x_j} * 100 \forall z_j \in A$: dữ liệu nhiễm bệnh tương đối tích lũy của ngày a_j



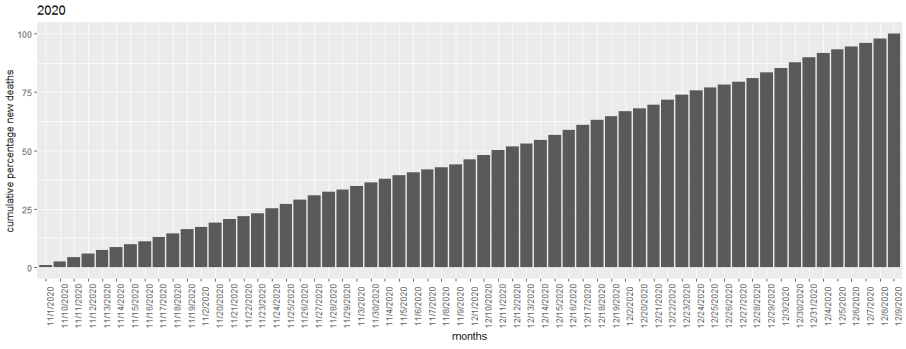
Hình 116: *Dữ liệu nhiễm bệnh tương đối tích lũy 2 tháng cuối năm 2020 của tất cả quốc gia*



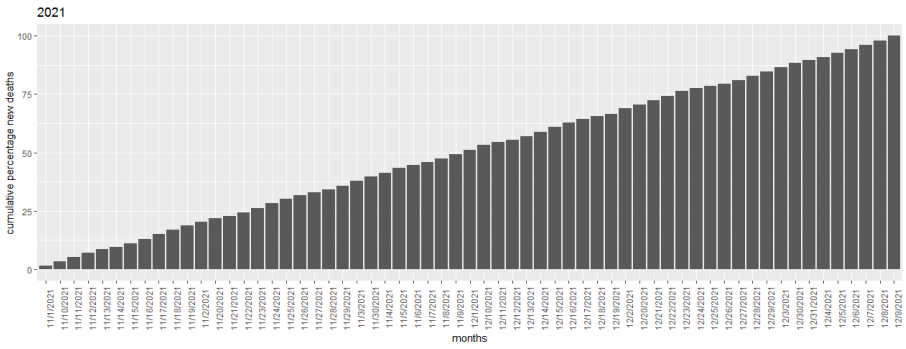
Hình 117: *Dữ liệu nhiễm bệnh tương đối tích lũy 2 tháng cuối năm 2021 của tất cả quốc gia*

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

- $D = \{d_i\}$: tập hợp các ngày thu thập dữ liệu của tất cả quốc gia
- $f : D \to Z$: $f(x)$ là hàm tìm tháng từ ngày x
- $A = \{11, 12\}$: các tháng thống kê
- x_i: số ca tử vong của ngày d_i
- $d_{max} \in \{d_{max} | d_{max} \in A \wedge d_{max} \geq d_i \forall d_i \in A\}$: ngày cuối cùng thu thập dữ liệu
- $d_{min} \in \{d_{min} | d_{min} \in A \wedge d_{min} \leq d_i \forall d_i \in A\}$: ngày đầu thu thập dữ liệu
- $a_j = \frac{\sum_{n=d_{min}}^{d_j} x_i}{\sum_{n=d_{min}}^{d_{max}} x_i} * 100 \forall z_j \in A$: dữ liệu tử vong tương đối tích lũy của ngày a_j



Hình 118: *Dữ liệu tử vong tương đối tích lũy 2 tháng cuối năm 2020 của tất cả quốc gia*

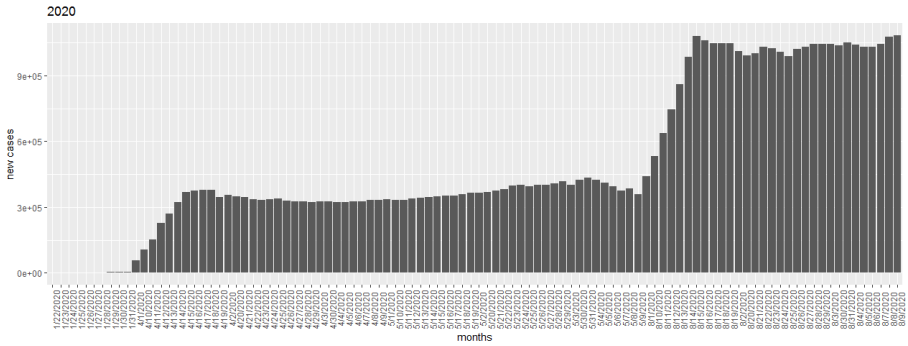


Hình 119: *Dữ liệu tử vong tương đối tích lũy 2 tháng cuối năm 2021 của tất cả quốc gia*

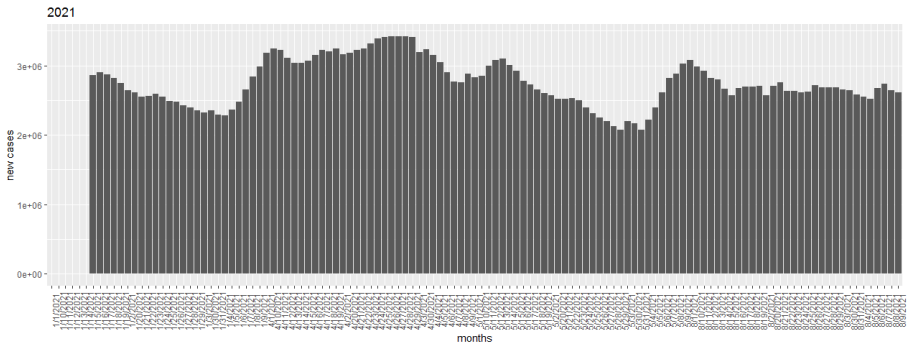
8 Nhóm câu hỏi liên quan đến tất cả quốc gia theo trung bình 7 ngày gần nhất

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

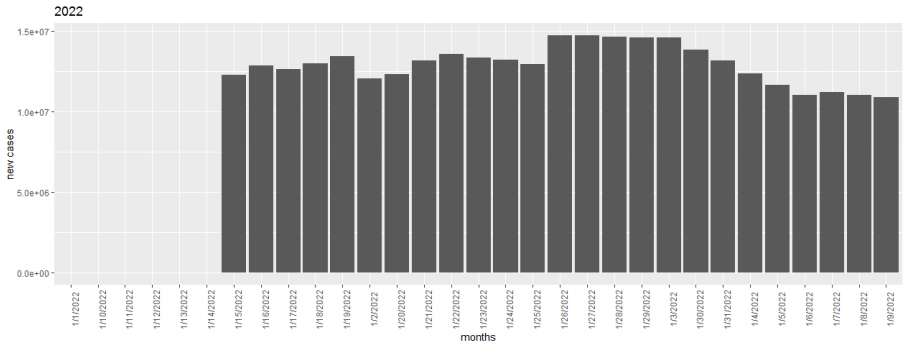
- $D = \{d_i\}$: tập hợp các ngày thu thập dữ liệu của tất cả quốc gia
- $f : D \to Z$: $f(x)$ là hàm tìm tháng từ ngày x
- $A = \{1, 8, 4, 5\}$: các tháng thống kê
- x_i: số ca nhiễm bệnh của ngày d_i
- $a_j = \frac{\sum_{n=i-6}^{i} x_i}{\sum_{n=i-6}^{i} 1} \forall z_j \in A$: dữ liệu nhiễm bệnh theo trung bình 7 ngày gần nhất của ngày a_j



Hình 120: *Dữ liệu nhiễm bệnh theo thời gian là tháng của tất cả quốc gia theo trung bình 7 ngày gần nhất năm 2020*



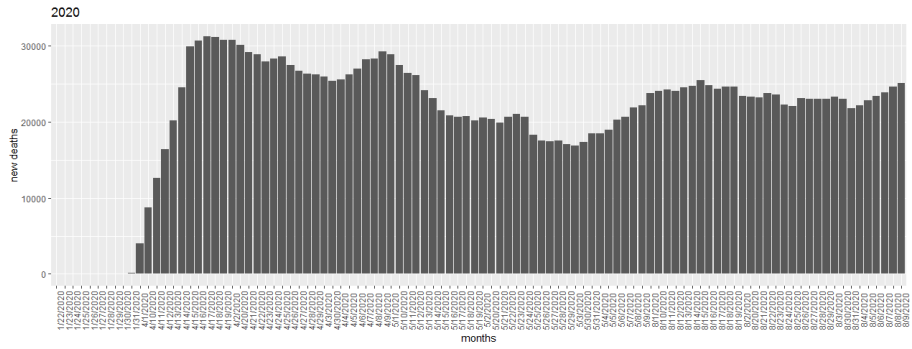
Hình 121: *Dữ liệu nhiễm bệnh theo thời gian là tháng của tất cả quốc gia theo trung bình 7 ngày gần nhất năm 2021*



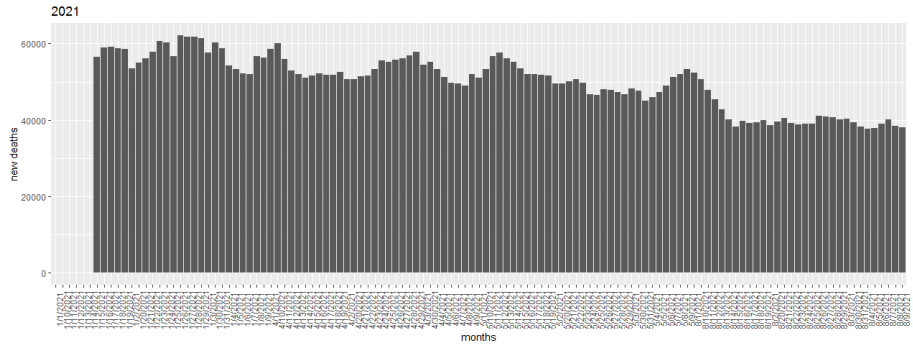
Hình 122: *Dữ liệu nhiễm bệnh theo thời gian là tháng của tất cả quốc gia theo trung bình 7 ngày gần nhất năm 2022*

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

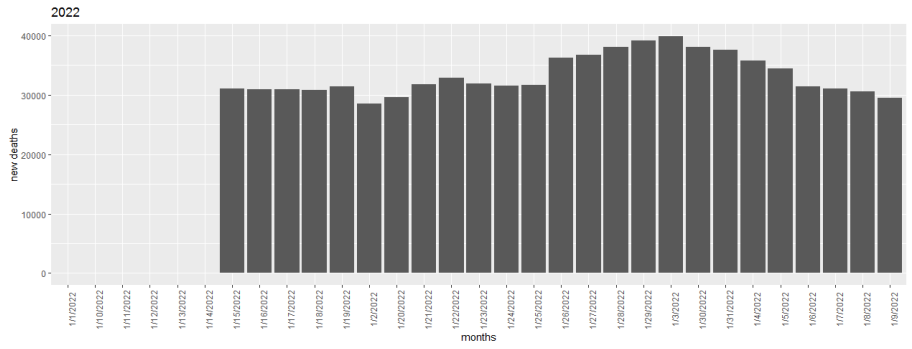
- $D = \{d_i\}$: tập hợp các ngày thu thập dữ liệu của tất cả quốc gia
- $f : D \to Z$: $f(x)$ là hàm tìm tháng từ ngày x
- $A = \{1, 8, 4, 5\}$: các tháng thống kê
- x_i: số ca tử vong của ngày d_i
- $a_j = \dfrac{\sum_{n=i-6}^{i} x_i}{\sum_{n=i-6}^{i} 1} \forall z_j \in A$: dữ liệu tử vong trung bình 7 ngày gần nhất của ngày a_j



Hình 123: *Dữ liệu tử vong theo thời gian là tháng của tất cả quốc gia theo trung bình 7 ngày gần nhất năm 2020*



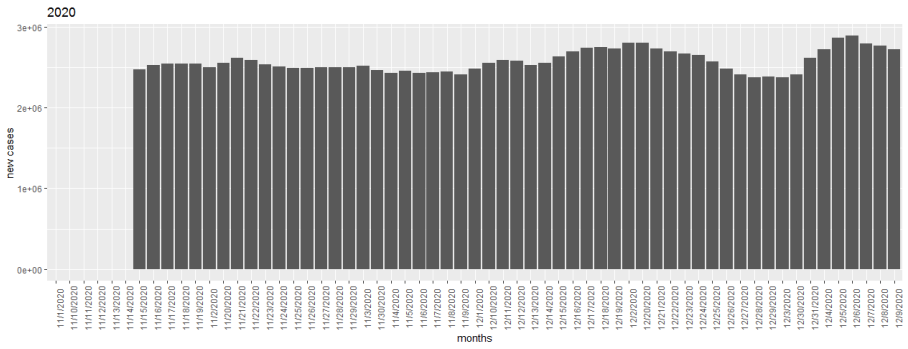
Hình 124: *Dữ liệu tử vong theo thời gian là tháng của tất cả quốc gia theo trung bình 7 ngày gần nhất năm 2021*



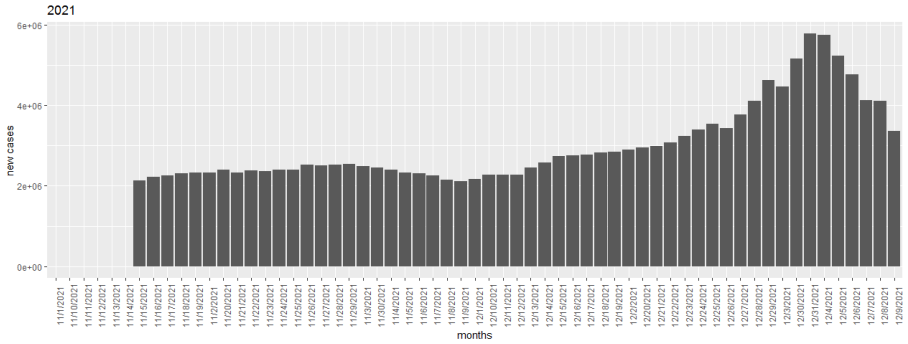
Hình 125: *Dữ liệu tử vong theo thời gian là tháng của tất cả quốc gia theo trung bình 7 ngày gần nhất năm 2022*

3) Biểu đồ thể hiện thu thập dữ liệu nhiễm bệnh theo thời gian là 2 tháng cuối năm của tất cả quốc gia theo trung bình 7 ngày gần nhất

- $D = \{d_i\}$: tập hợp các ngày thu thập dữ liệu của tất cả quốc gia
- $f : D \to Z$: $f(x)$ là hàm tìm tháng từ ngày x
- $A = \{11, 12\}$: các tháng thống kê
- x_i: số ca nhiễm bệnh của ngày d_i
- $a_j = \dfrac{\sum_{n=i-6}^{i} x_i}{\sum_{n=i-6}^{i} 1} \forall z_j \in A$: dữ liệu nhiễm bệnh theo trung bình 7 ngày gần nhất của ngày a_j



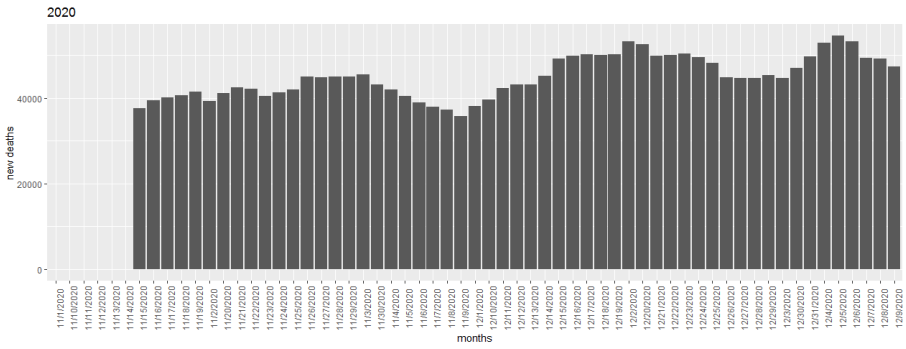
Hình 126: *Dữ liệu nhiễm bệnh theo thời gian 2 tháng cuối năm 2020 của tất cả quốc gia theo trung bình 7 ngày gần nhất*



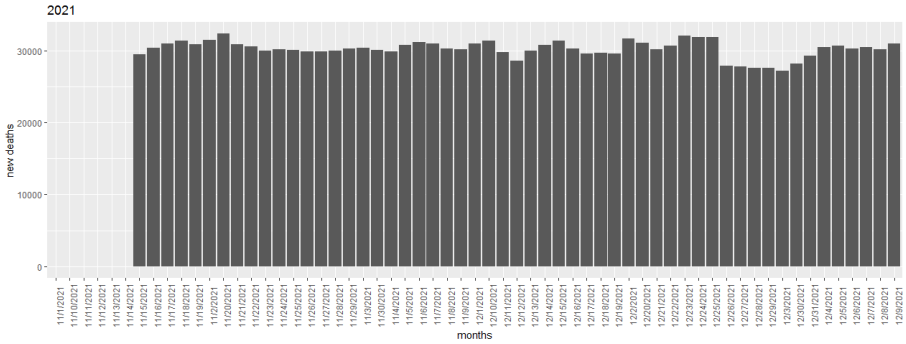
Hình 127: *Dữ liệu nhiễm bệnh theo thời gian 2 tháng cuối năm 2021 của tất cả quốc gia theo trung bình 7 ngày gần nhất*

4) Biểu đồ thể hiện thu thập dữ liệu tử vong theo thời gian là 2 tháng cuối năm của tất cả quốc gia theo trung bình 7 ngày gần nhất

- $D = \{d_i\}$: tập hợp các ngày thu thập dữ liệu của tất cả quốc gia
- $f : D \to Z$: $f(x)$ là hàm tìm tháng từ ngày x
- $A = \{11, 12\}$: các tháng thống kê
- x_i: số ca tử vong của ngày d_i
- $a_j = \dfrac{\sum_{n=i-6}^{i} x_i}{\sum_{n=i-6}^{i} 1} \forall z_j \in A$: dữ liệu tử vong theo trung bình 7 ngày gần nhất của ngày a_j



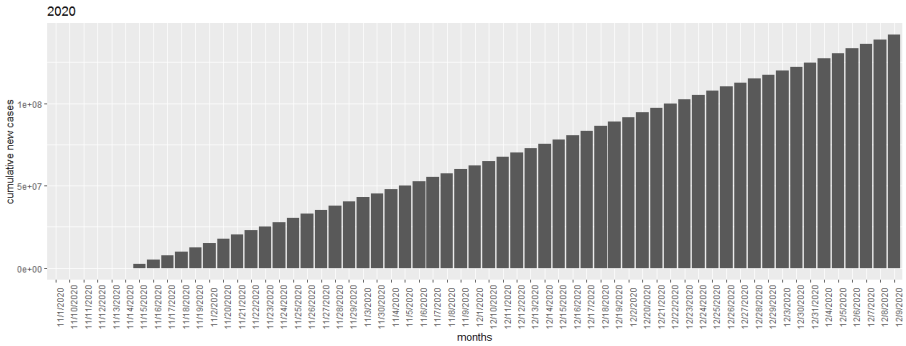
Hình 128: *Dữ liệu tử vong theo thời gian 2 tháng cuối năm 2020 của tất cả quốc gia theo trung bình 7 ngày gần nhất*



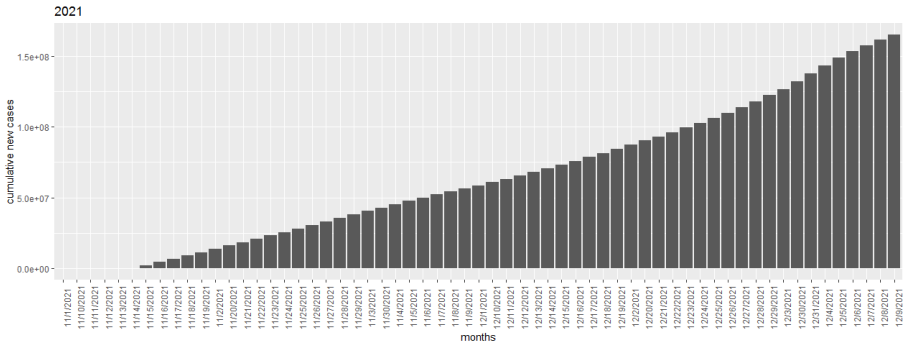
Hình 129: *Dữ liệu tử vong theo thời gian 2 tháng cuối năm 2021 của tất cả quốc gia theo trung bình 7 ngày gần nhất*

5) Biểu đồ thể hiện thu thập dữ liệu nhiễm bệnh tích lũy theo thời gian là 2 tháng cuối năm của tất cả quốc giaị theo trung bình 7 ngày gần nhất

- $D = \{d_i\}$: tập hợp các ngày thu thập dữ liệu của tất cả quốc gia
- $f : D \to Z$: $f(x)$ là hàm tìm tháng từ ngày x
- $A = \{11, 12\}$: các tháng thống kê
- x_i: số ca nhiễm bệnh của ngày d_i
- $a_j = \sum_{n=d_{min}}^{d_j} x_j \forall z_j \in A$: dữ liệu nhiễm bệnh tích lũy của ngày a_j



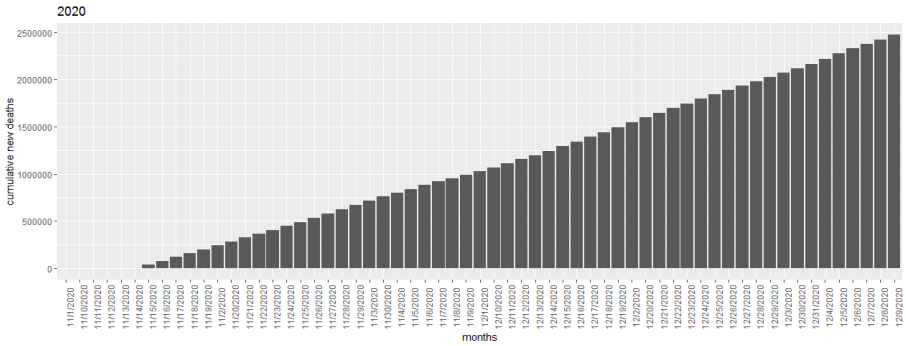
Hình 130: *Dữ liệu nhiễm bệnh theo thời gian 2 tháng cuối năm 2020 của tất cả quốc gia theo trung bình 7 ngày gần nhất*



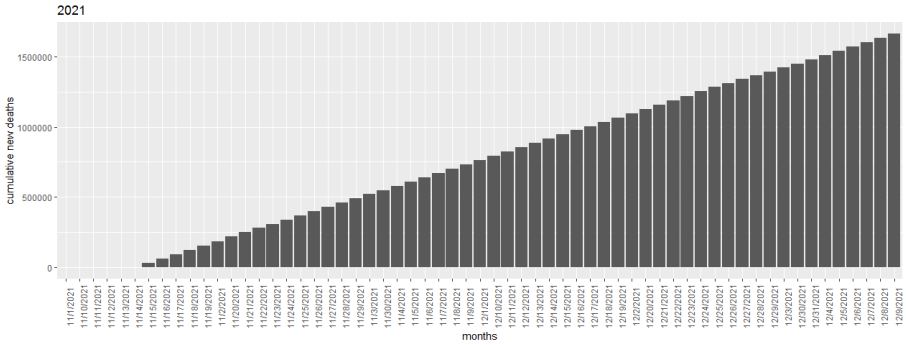
Hình 131: *Dữ liệu nhiễm bệnh theo thời gian 2 tháng cuối năm 2021 của tất cả quốc gia theo trung bình 7 ngày gần nhất*

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

- $D = \{d_i\}$: tập hợp các ngày thu thập dữ liệu của tất cả quốc gia
- $f : D \to Z$: $f(x)$ là hàm tìm tháng từ ngày x
- $A = \{11, 12\}$: các tháng thống kê
- x_i: số ca tử vong của ngày d_i
- $a_j = \sum_{n=d_{min}}^{d_j} x_j \forall z_j \in A$: dữ liệu tử vong tích lũy của ngày a_j



Hình 132: *Dữ liệu tử vong theo thời gian 2 tháng cuối năm 2020 của tất cả quốc gia theo trung bình 7 ngày gần nhất*



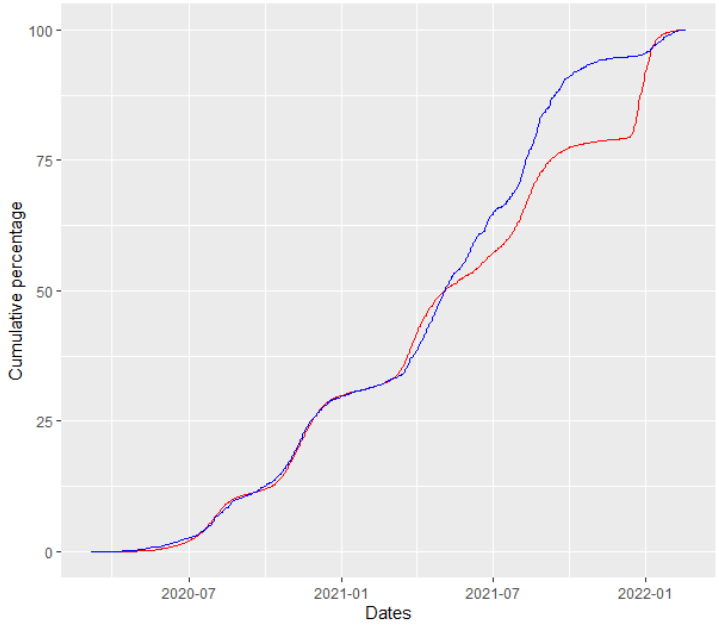
Hình 133: *Dữ liệu tử vong theo thời gian 2 tháng cuối năm 2021 của tất cả quốc gia theo trung bình 7 ngày gần nhất*

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

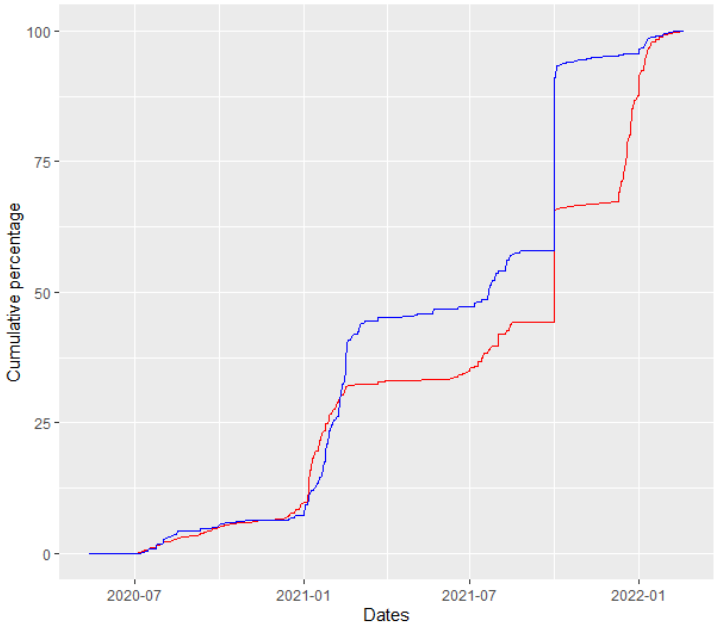
1) Vẽ biểu đồ thể hiện phần trăm giữa nhiễm bệnh tích lũy trên tổng nhiễm bệnh và phần trăm tử vong tích lũy trên tổng số tử vong cho từng quốc gia theo thời gian. Vẽ 2 đường trên cùng biểu đồ
Cách giải

- P: tập hợp các ngày thu thập dữ liệu
- n_i: ngày thứ i
- x_i: số ca nhiễm bệnh của ngày thứ i
- y_i: số ca tử vong của ngày thứ i
- $n_{max} \in \{n_{max} | n_{max} \in P, n_{max} \geq n_i \forall n_i \in P\}$: ngày cuối cùng thu thập dữ liệu
- $n_{min} \in \{n_{min} | n_{min} \in P, n_{min} \leq n_i \forall n_i \in P\}$: ngày đầu thu thập dữ liệu
- $a_i = \sum_{n=n_{min}}^{n_i} x_i$: Số ca nhiễm bệnh tích lũy ngày i
- $b_i = \sum_{n=n_{min}}^{n_i} y_i$: Số ca tử vong tích lũy ngày i
- $c_i = \frac{a_i}{\sum_{n=n_{min}}^{n_{max}} x_i}$: Tỷ lệ ca nhiễm bệnh tích lũy ngày i trên tổng ca nhiễm
- $d_i = \frac{b_i}{\sum_{n=n_{min}}^{n_{max}} y_i}$: Tỷ lệ ca tử vong tích lũy ngày i trên tổng ca nhiễm

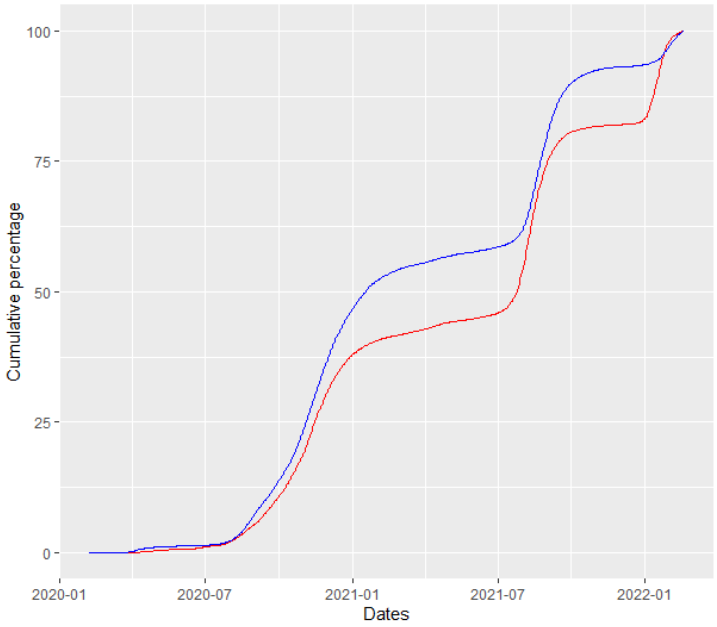
Kết quả



Hình 134: *Phần trăm nhiễm bệnh tích lũy trên tổng nhiễm bệnh và phần trăm tử vong tích lũy trên tổng số tử vong tại Kenya*



Hình 135: *Phần trăm nhiễm bệnh tích lũy trên tổng nhiễm bệnh và phần trăm tử vong tích lũy trên tổng số tử vong tại Lesotho*



Hình 136: *Phần trăm nhiễm bệnh tích lũy trên tổng nhiễm bệnh và phần trăm tử vong tích lũy trên tổng số tử vong tại Morocco*

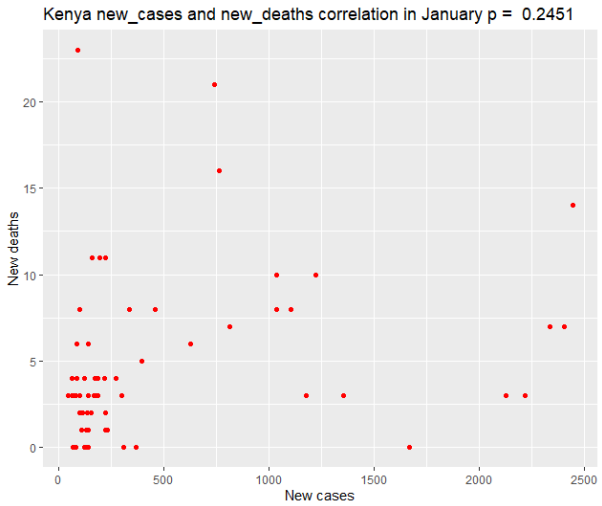
Trên từng quốc gia riêng của nhóm hãy vẽ biểu đồ thể hiện trục Ox là nhiễm bệnh, trục Oy là tử vong. Hãy lấy 4 tháng theo 4 ký số mã đề thể hiện. Nếu ký số là 0 thì lấy tháng là 10.

2) Xét tương quan trong mỗi tháng
Cách giải

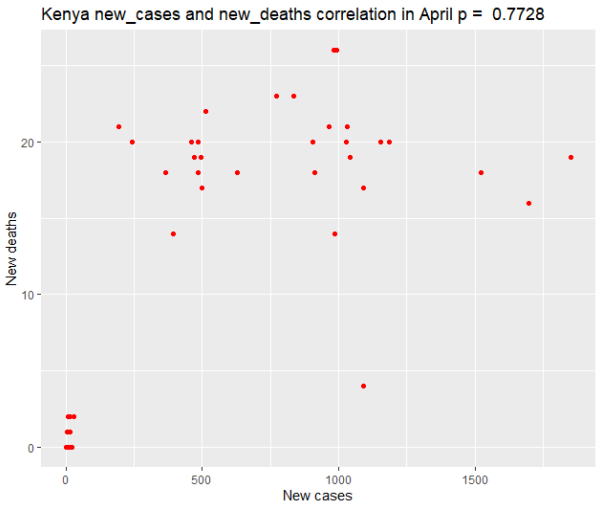
- $D = \{d_i\}$: tập hợp các ngày thu thập dữ liệu của 1 quốc gia
- $f : D \to Z$: $f(x)$ là hàm tìm tháng từ ngày x
- $A = \{1, 8, 4, 5\}$: các tháng thống kê
- x_i: số ca nhiễm bệnh của ngày d_i
- y_i: số ca tử vong của ngày d_i
- $d_{max} \in \{d_{max} | d_{max} \in P \wedge d_{max} \geq d_i \forall d_i \in D\}$: ngày cuối cùng thu thập dữ liệu
- $d_{min} \in \{d_{min} | d_{min} \in P \wedge d_{min} \leq d_i \forall d_i \in D\}$: ngày đầu thu thập dữ liệu
- $\bar{x_j} = \frac{\sum_{n=d_{min}}^{d_i} x_i}{\sum_{n=d_{min}}^{d_i} 1} \forall z_j \in A$: trung bình số nhiễm bệnh của từng tháng 1, 8, 4, 5
- $\bar{y_j} = \frac{\sum_{n=d_{min}}^{d_i} y_i}{\sum_{n=d_{min}}^{d_i} 1} \forall z_j \in A$: trung bình số tử vong của từng tháng 1, 8, 4, 5
- p_j: hệ số tương quan giữa nhiễm bệnh và tử vong trong tháng z_j

$$p_j = \frac{\sum_{n=d_{min}}^{d_{max}} (x_i - \bar{x_j})(y_i - \bar{y_j})}{\sqrt{\sum_{n=d_{min}}^{d_{max}} (x_i - \bar{x_j})^2 \sum_{n=d_{min}}^{d_{max}} (y_i - \bar{y_j})^2}}$$

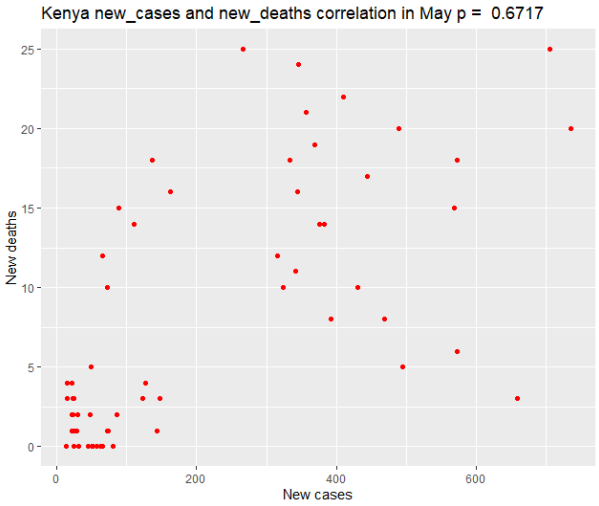
Kết quả



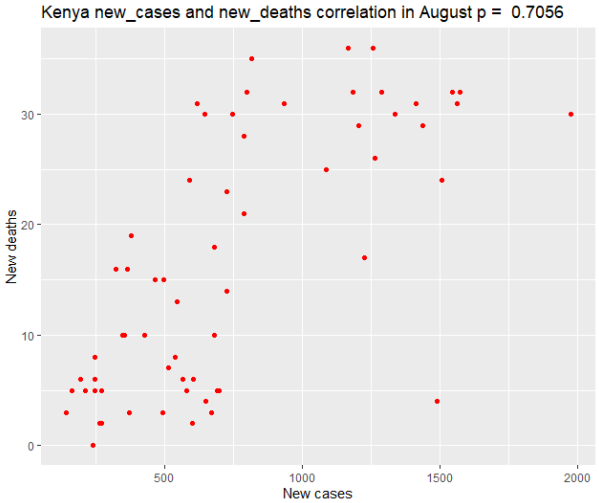
Hình 137: *Tương quan nhiễm bệnh và tử vong tháng 1 tại Kenya*



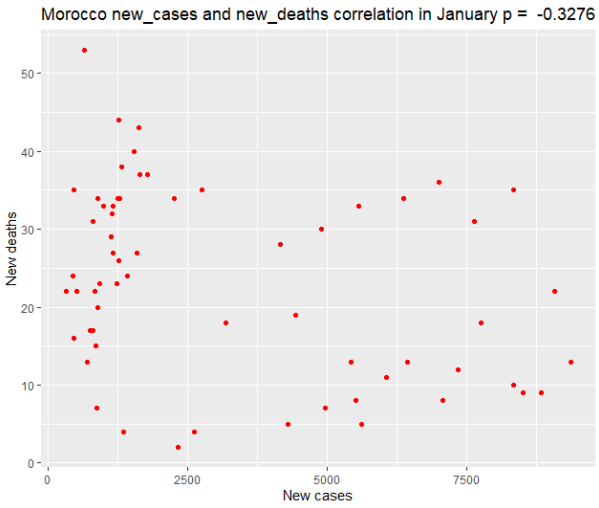
Hình 138: *Tương quan nhiễm bệnh và tử vong tháng 4 tại Kenya*



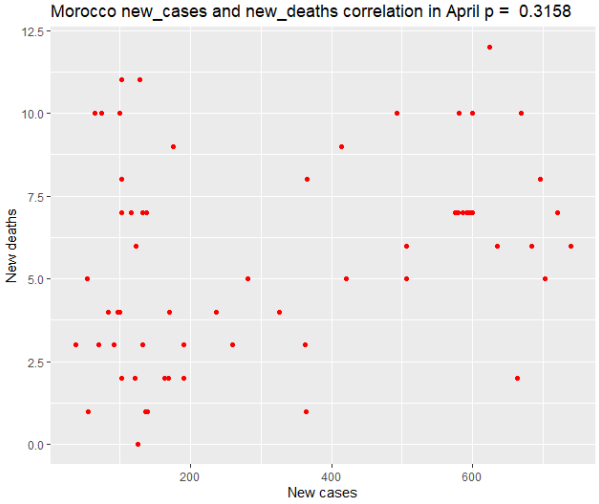
Hình 139: *Tương quan nhiễm bệnh và tử vong tháng 5 tại Kenya*



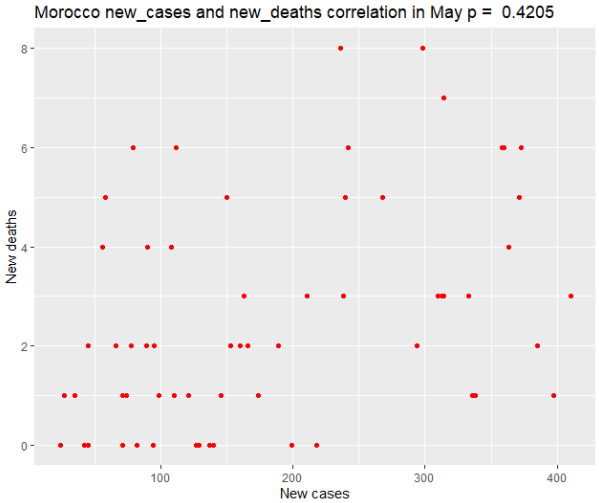
Hình 140: *Tương quan nhiễm bệnh và tử vong tháng 8 tại Kenya*



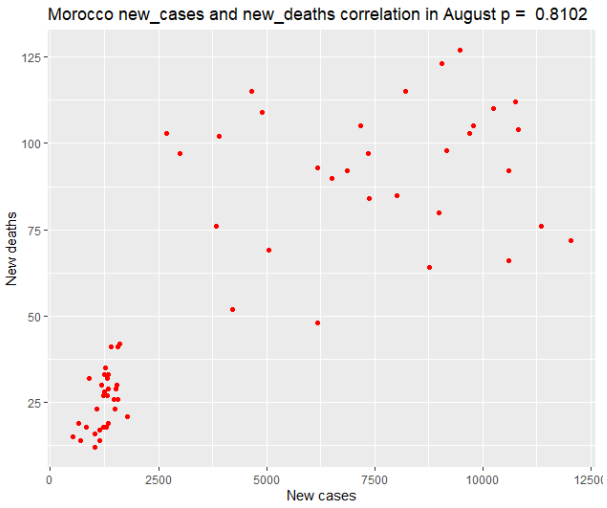
Hình 141: *Tương quan nhiễm bệnh và tử vong tháng 1 tại Morocco*



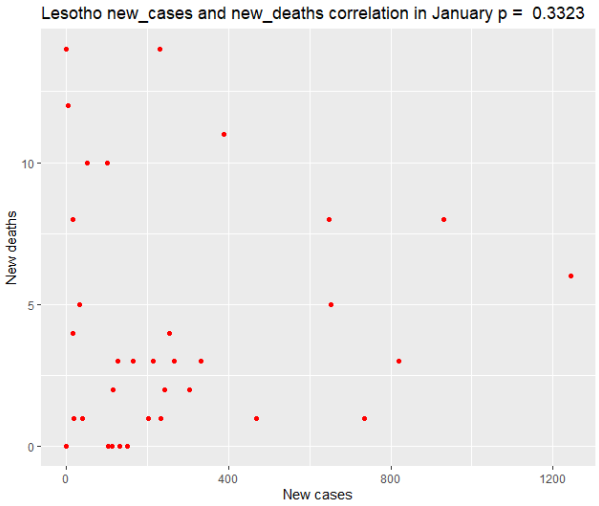
Hình 142: *Tương quan nhiễm bệnh và tử vong tháng 4 tại Morocco*



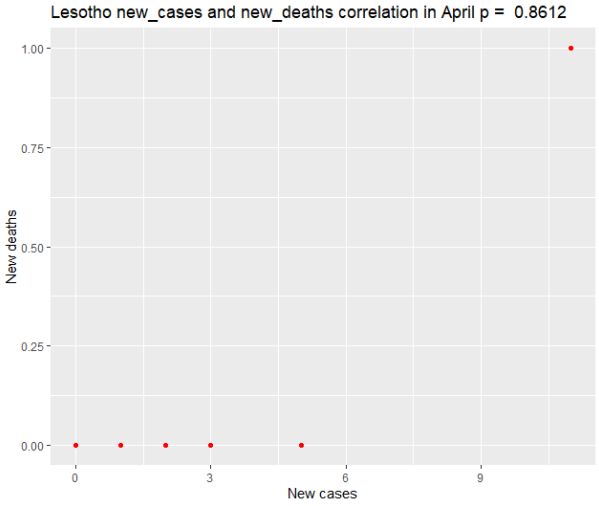
Hình 143: *Tương quan nhiễm bệnh và tử vong tháng 5 tại Morocco*



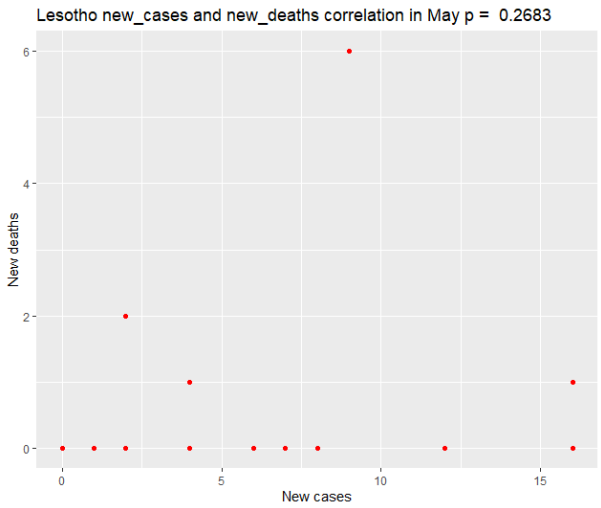
Hình 144: *Tương quan nhiễm bệnh và tử vong tháng 8 tại Morocco*



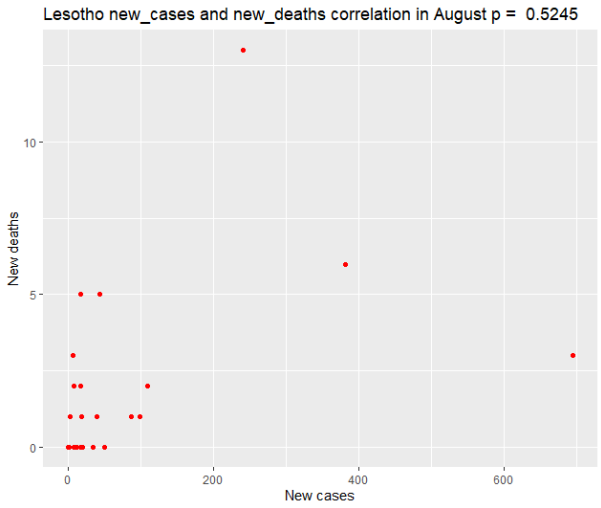
Hình 145: *Tương quan nhiễm bệnh và tử vong tháng 1 tại Lesotho*



Hình 146: *Tương quan nhiễm bệnh và tử vong tháng 4 tại Lesotho*



Hình 147: *Tương quan nhiễm bệnh và tử vong tháng 5 tại Lesotho*



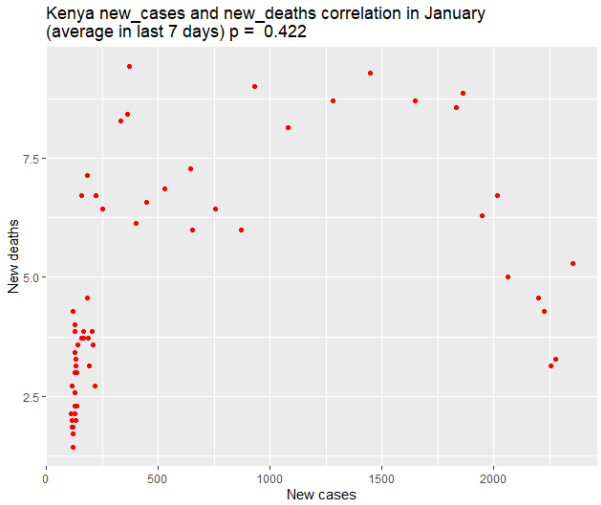
Hình 148: *Tương quan nhiễm bệnh và tử vong tháng 8 tại Lesotho*

3) Xét tương quan trong mỗi tháng theo trung bình 7 ngày gần nhất
Cách giải

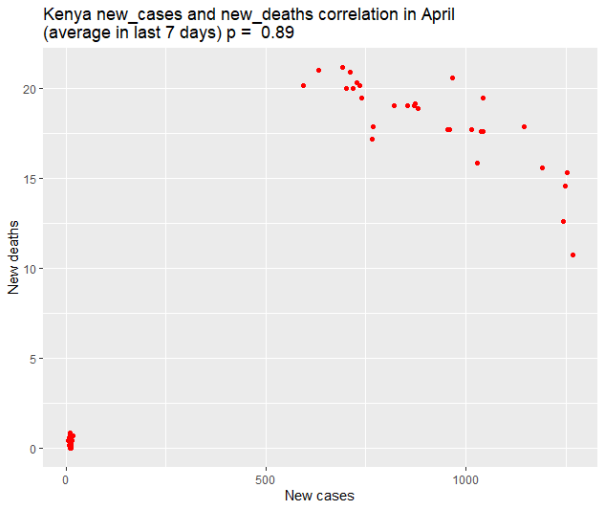
- $D = \{d_i\}$: tập hợp các ngày thu thập dữ liệu của 1 quốc gia
- $f : D \to Z$: $f(x)$ là hàm tìm tháng từ ngày x
- $A = \{1, 8, 4, 5\}$: các tháng thống kê
- x_i: số ca nhiễm bệnh của ngày d_i
- $a_i = \frac{\sum_{n=i-6}^{i} x_i}{\sum_{n=i-6}^{i} 1}$: trung bình ca nhiễm bệnh trong 7 ngày gần nhất
- y_i: số ca tử vong của ngày d_i
- $b_i = \frac{\sum_{n=i-6}^{i} y_i}{\sum_{n=i-6}^{i} 1}$: trung bình ca tử vong trong 7 ngày gần nhất
- $d_{max} \in \{d_{max} | d_{max} \in P \wedge d_{max} \geq d_i \forall d_i \in D\}$: ngày cuối cùng thu thập dữ liệu
- $d_{min} \in \{d_{min} | d_{min} \in P \wedge d_{min} \leq d_i \forall d_i \in D\}$: ngày đầu thu thập dữ liệu
- $\bar{a_j} = \frac{\sum_{n=d_{min}}^{d_i} a_i}{\sum_{n=d_{min}}^{d_i} 1} \forall z_j \in A$: trung bình của trung bình số nhiễm bệnh trong 7 ngày gần nhất của từng tháng 1, 8, 4, 5
- $\bar{b_j} = \frac{\sum_{n=d_{min}}^{d_i} b_i}{\sum_{n=d_{min}}^{d_i} 1} \forall z_j \in A$: trung bình số tử vong của từng tháng 1, 8, 4, 5
- p_j: hệ số tương quan giữa nhiễm bệnh và tử vong (trung bình 7 ngày gần nhất) trong tháng z_j

$$p_j = \frac{\sum_{n=d_{min}}^{d_{max}} (a_i - \bar{a_j})(b_i - \bar{b_j})}{\sqrt{\sum_{n=d_{min}}^{d_{max}} (a_i - \bar{a_j})^2 \sum_{n=d_{min}}^{d_{max}} (b_i - \bar{b_j})^2}}$$

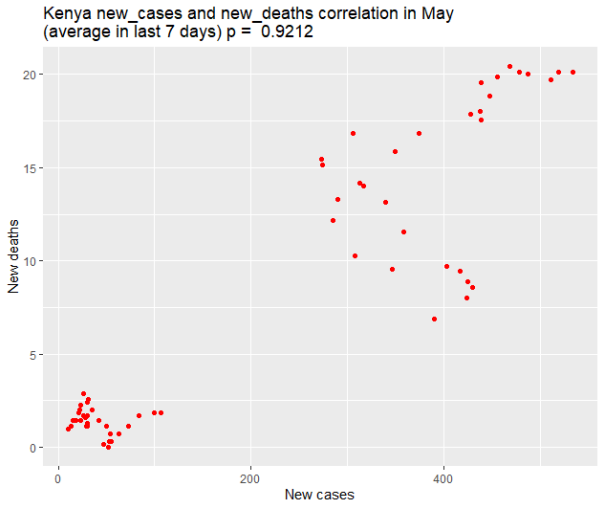
Kết quả



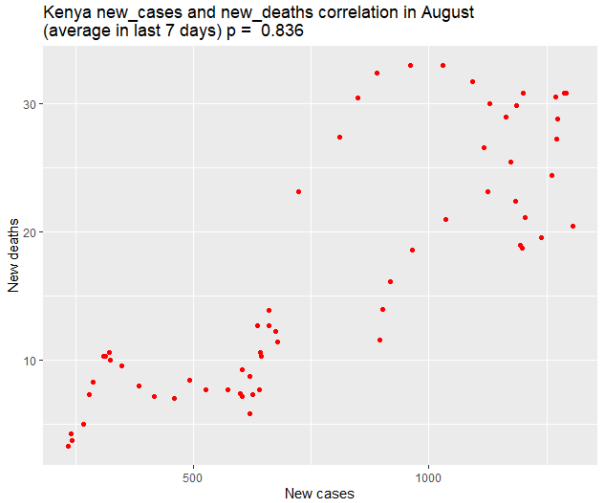
Hình 149: *Tương quan nhiễm bệnh và tử vong (trung bình 7 ngày) tháng 1 tại Kenya*



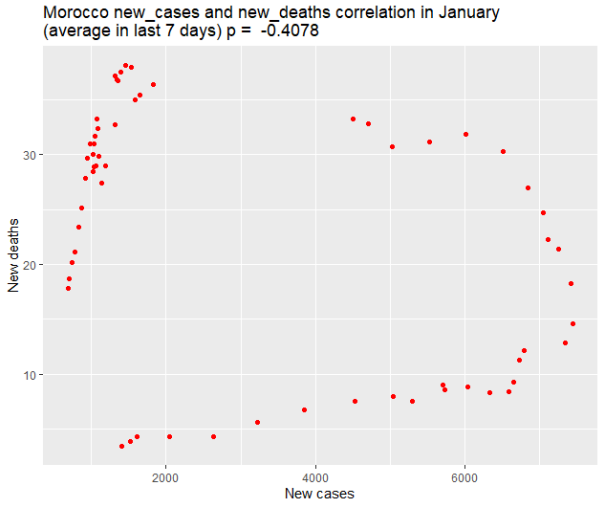
Hình 150: *Tương quan nhiễm bệnh và tử vong (trung bình 7 ngày) tháng 4 tại Kenya*



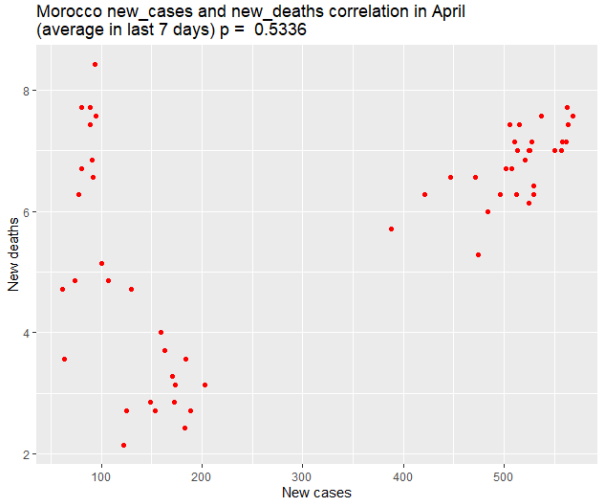
Hình 151: *Tương quan nhiễm bệnh và tử vong (trung bình 7 ngày) tháng 5 tại Kenya*



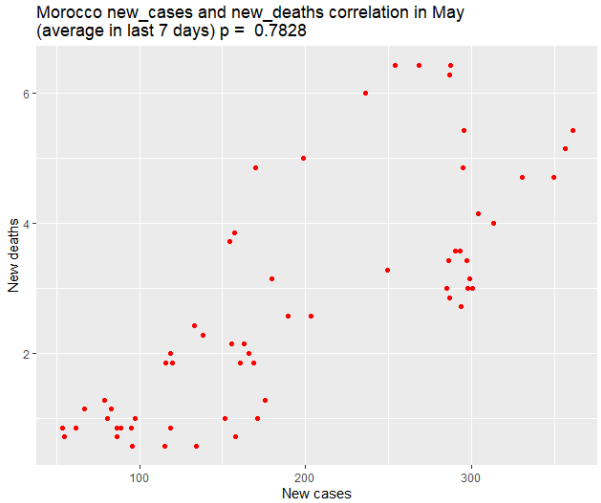
Hình 152: *Tương quan nhiễm bệnh và tử vong (trung bình 7 ngày) tháng 8 tại Kenya*



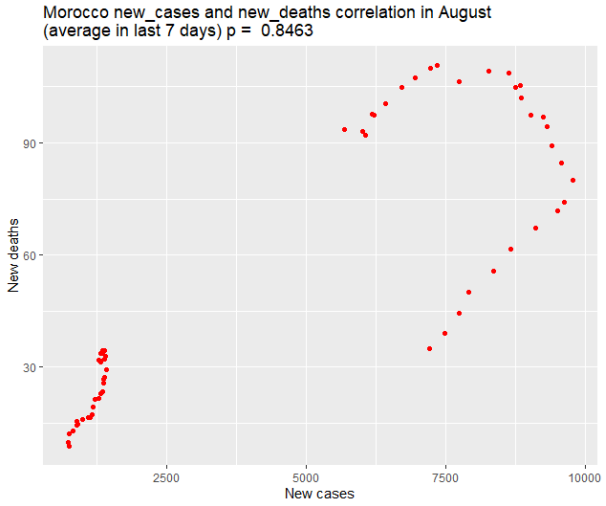
Hình 153: *Tương quan nhiễm bệnh và tử vong (trung bình 7 ngày) tháng 1 tại Morocco*



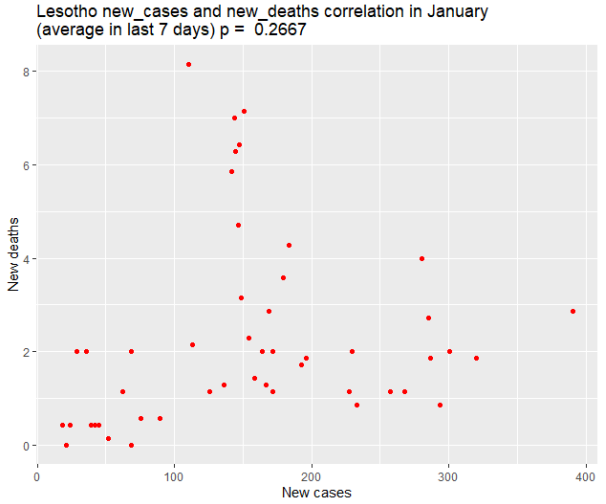
Hình 154: *Tương quan nhiễm bệnh và tử vong (trung bình 7 ngày) tháng 4 tại Morocco*



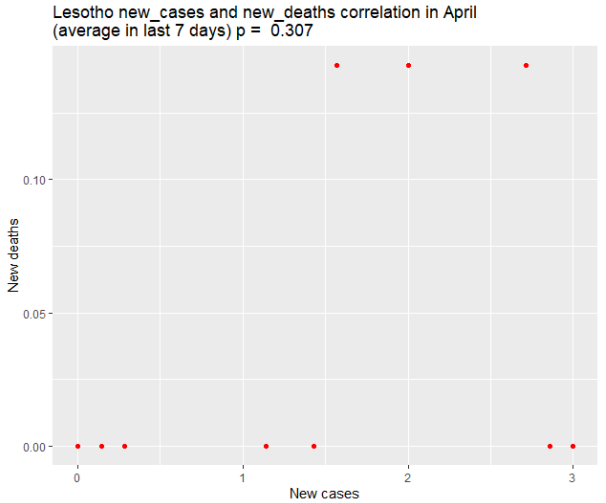
Hình 155: *Tương quan nhiễm bệnh và tử vong (trung bình 7 ngày) tháng 5 tại Morocco*



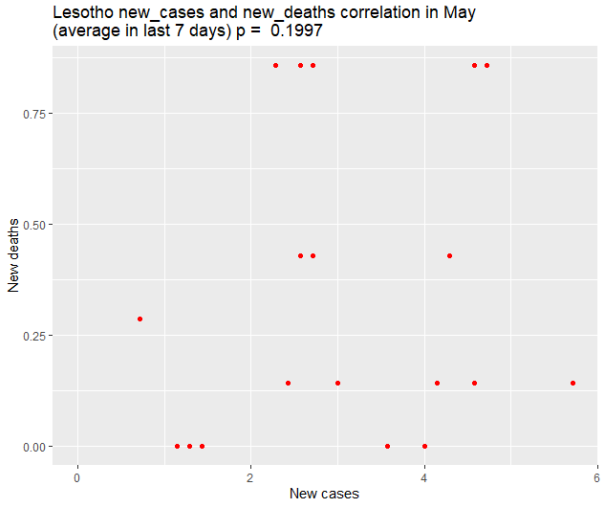
Hình 156: *Tương quan nhiễm bệnh và tử vong (trung bình 7 ngày) tháng 8 tại Morocco*



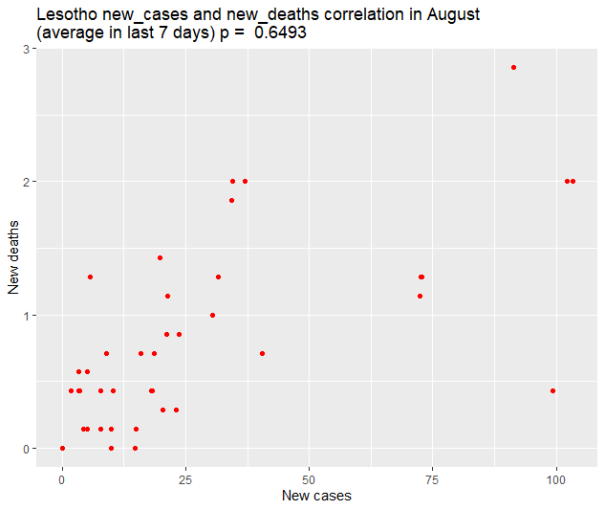
Hình 157: *Tương quan nhiễm bệnh và tử vong (trung bình 7 ngày) tháng 1 tại Lesotho*



Hình 158: *Tương quan nhiễm bệnh và tử vong (trung bình 7 ngày) tháng 4 tại Lesotho*



Hình 159: *Tương quan nhiễm bệnh và tử vong (trung bình 7 ngày) tháng 5 tại Lesotho*



Hình 160: *Tương quan nhiễm bệnh và tử vong (trung bình 7 ngày) tháng 8 tại Lesotho*

10 Nhóm câu hỏi riêng

1) So sánh tình trạng nhiễm bệnh của các quốc gia trong 7 ngày cuối của năm cuối cùng
Cách giải

- A: tập hợp các bản ghi
- d_i: ngày của bản ghi thứ i của A
- $f1 : A \to B$ với $f1$ là hàm lấy ra năm từ ngày của bản ghi của A
- $y_{max} \in \{y_{max} | y_{max} \in B \wedge y_{max} \geq y_i \forall y_i \in B\}$: năm cuối cùng
- $d_{max} \in \{d_{max} | d_{max} \in A \wedge d_{max} \geq d_i \forall d_i \in A\}$: ngày cuối cùng thu thập dữ liệu
- $G = \{A_i | f1(A_i) = y_{max} \wedge d_i \geq d_{max} - 6\}$: tập hợp các bản ghi trong 7 ngày cuối của năm cuối cùng
- g_j: số ca nhiễm mới của bản ghi thứ j của G
- $f2 : G \to C$ với $f2$ là hàm lấy ra quốc gia từ bản ghi của G
- $b_{c_k} = \sum_{n=d_{max}-6}^{d_{max}} g_i \forall f2(G) = c_k$: tổng số ca nhiễm mới của quốc gia c_k trong 7 ngày cuối của năm cuối cùng
- $e_{c_k} = \frac{\sum_{n=d_{max}-6}^{d_{max}} g_i}{\sum_{n=d_{max}-6}^{d_{max}} 1} \forall f2(G) = c_k$: trung bình số ca nhiễm mới của quốc gia c_k trong 7 ngày cuối của năm cuối cùng

Kết quả

	location	Total_new_cases	Avg_new_cases
1	Russia	1453180	181647.50
2	Germany	1330792	166349.00
3	Brazil	878031	109753.88
4	United States	812284	101535.50
5	France	724273	90534.12
6	Turkey	685789	85723.62
7	South Korea	668632	83579.00
8	Japan	645090	80636.25
9	Netherlands	442205	55275.62
10	Italy	436664	54583.00

Hình 161: *Danh sách 10 nước có tổng số ca nhiễm bệnh trong 7 ngày cuối của năm cuối cùng cao nhất*

2) Với k là mốc bùng phát dịch, hãy xác định k và cho biết các khoảng thời gian bùng phát

Cách giải

Chọn mốc $k = 3000000$ ứng với tổng số ca nhiễm mới trong 7 ngày gần nhất.

- A: tập hợp các bản ghi, d_i: ngày của bản ghi thứ i của A
- x_i: số ca nhiễm bệnh của bản ghi thứ i của A
- $a_i = \sum_{d_{i-6}}^{d_i} x_i$: tổng số ca nhiễm bệnh của 7 ngày gần nhất ở ngày d_i
- $b_i = \sum_{d_{i-7}}^{d_{i-1}} x_i$: tổng số ca nhiễm bệnh của 7 ngày gần nhất ở ngày d_{i-1}
- $c_i = \sum_{d_{i-5}}^{d_{i+1}} x_i$: tổng số ca nhiễm bệnh của 7 ngày gần nhất ở ngày d_{i+1}
- o_i: số lần bùng phát dịch đã từng xảy ra tính tới ngày d_i

  $o_i = \sum_1^i f1(a_i, b_i, c_i)$ với $f1(a_i, b_i, c_i) = \begin{cases} 1, & a_i > k \wedge c_i > k \wedge b_i < k \\ 0, & \neg(a_i > k \wedge c_i > k \wedge b_i < k) \end{cases}$
- e_i: ngày d_i thuộc đợt bùng dịch thứ mấy, nếu 0 tức là đang không thuộc đợt bùng dịch nào cả

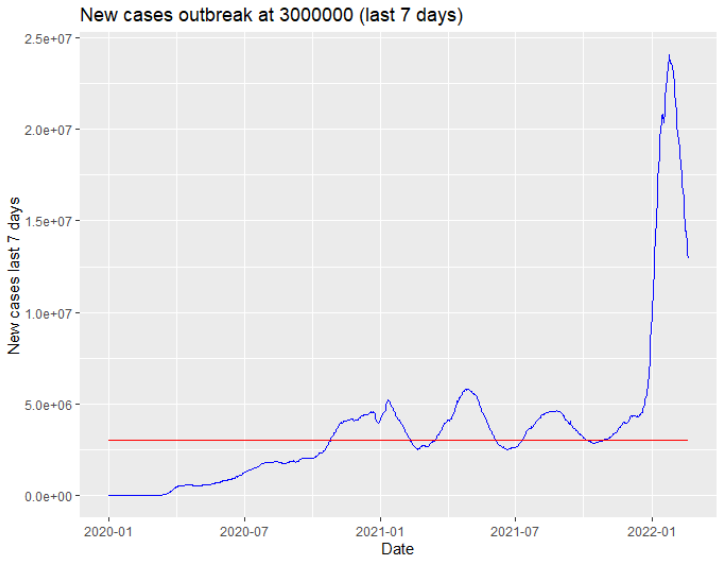
  $e_i = \begin{cases} o_i, & a_i > k \\ 0, & a_i \leq k \end{cases}$
- $B = \{e_i\}, g_j \in B$: B là tập hợp các đợt bùng dịch, g_j là số thứ tự của đợt bùng dịch
- $h_j \in \{h_j | h_j \in \{d_i\} \wedge h_j \leq d_i \forall e_i = g_j\}$: ngày bắt đầu đợt bùng phát dịch g_j
- $m_j \in \{m_j | m_j \in \{d_i\} \wedge m_j \geq d_i \forall e_i = g_j\}$: ngày kết thúc đợt bùng phát dịch g_j

Kết quả

Có 4 khoảng thời gian bùng phát dịch như hình sau:

outbreak_no	Start_date	End_date
1	2020-10-26	2021-02-09
2	2021-03-15	2021-06-05
3	2021-07-10	2021-10-07
4	2021-10-27	2022-02-19

Hình 162: *Danh sách khoảng thời gian bùng phát dịch*



Hình 163: *Biểu đồ số ca nhiễm bệnh trong 7 ngày gần nhất so với mốc bùng phát dịch* $k = 3000000$

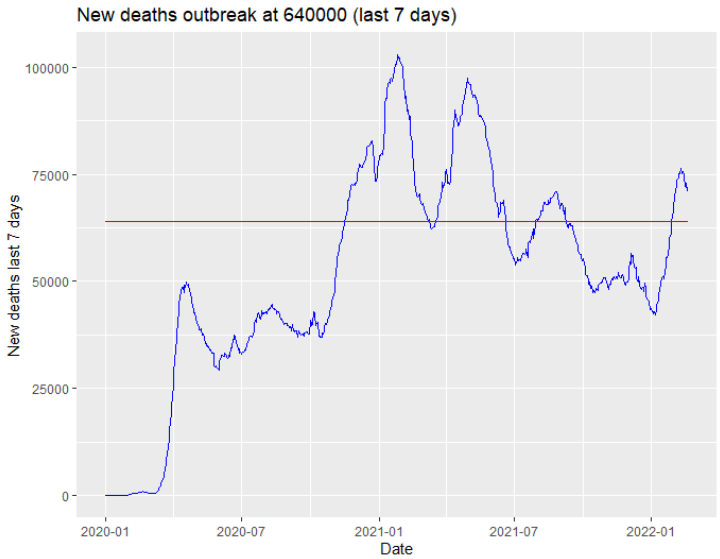
3) Với k là mốc bùng tử vong, hãy xác định k và cho biết các khoảng thời gian bùng phát
Cách giải
Chọn mốc $k = 64000$ ứng với tổng số ca tử vong trong 7 ngày gần nhất.

- A: tập hợp các bản ghi, d_i: ngày của bản ghi thứ i của A
- x_i: số ca tử vong của bản ghi thứ i của A
- $a_i = \sum_{d_{i-6}}^{d_i} x_i$: tổng số ca tử vong của 7 ngày gần nhất ở ngày d_i
- $b_i = \sum_{d_{i-7}}^{d_{i-1}} x_i$: tổng số ca tử vong của 7 ngày gần nhất ở ngày d_{i-1}
- $c_i = \sum_{d_{i-5}}^{d_{i+1}} x_i$: tổng số ca tử vong của 7 ngày gần nhất ở ngày d_{i+1}
- o_i: số lần bùng tử vong đã từng xảy ra tính tới ngày d_i
  $o_i = \sum_1^i f1(a_i, b_i, c_i)$ với $f1(a_i, b_i, c_i) = \begin{cases} 1, & a_i > k \wedge c_i > k \wedge b_i < k \\ 0, & \neg(a_i > k \wedge c_i > k \wedge b_i < k) \end{cases}$
- e_i: ngày d_i thuộc đợt bùng tử vong thứ mấy, nếu 0 tức là đang không thuộc đợt bùng tử vong nào cả
  $e_i = \begin{cases} o_i, & a_i > k \\ 0, & a_i \leq k \end{cases}$
- $B = \{e_i\}, g_j \in B$: B là tập hợp các đợt bùng tử vong, g_j là số thứ tự của đợt bùng tử vong
- $h_j \in \{h_j | h_j \in \{d_i\} \wedge h_j \leq d_i \forall e_i = g_j\}$: ngày bắt đầu đợt bùng tử vong g_j
- $m_j \in \{m_j | m_j \in \{d_i\} \wedge m_j \geq d_i \forall e_i = g_j\}$: ngày kết thúc đợt bùng tử vong g_j

Có 4 khoảng thời gian bùng tử vong như hình sau:

outbreak_no	Start_date	End_date
1	2020-11-17	2021-03-09
2	2021-03-19	2021-06-19
3	2021-07-30	2021-09-09
4	2022-01-29	2022-02-19

Hình 164: *Danh sách khoảng thời gian bùng tử vong*



Hình 165: *Biểu đồ số ca tử vong trong 7 ngày gần nhất so với mốc bùng tử vong $k = 64000$*