Анализ данных распространения COVID-19

Источник: https://ourworldindata.org/coronavirus

Подготовила Гришина Екатерина Дмитриевна

Постановка задачи

Найти взаимосвязи между различными характеристиками стран и заболеваемостью/смертностью от COVID-19, а также разделить страны на группы относительно динамики распространения заболевания в них.

Поиск взаимосвязей

Основной задачей был поиск закономерностей между различными социальными, экономическими и медицинскими показателями. Для поиска связей среди данных был применен корреляционный анализ с помощью библиотеки seaborn, и он для некоторых переменных показал неожиданные результаты.

Результаты анализа

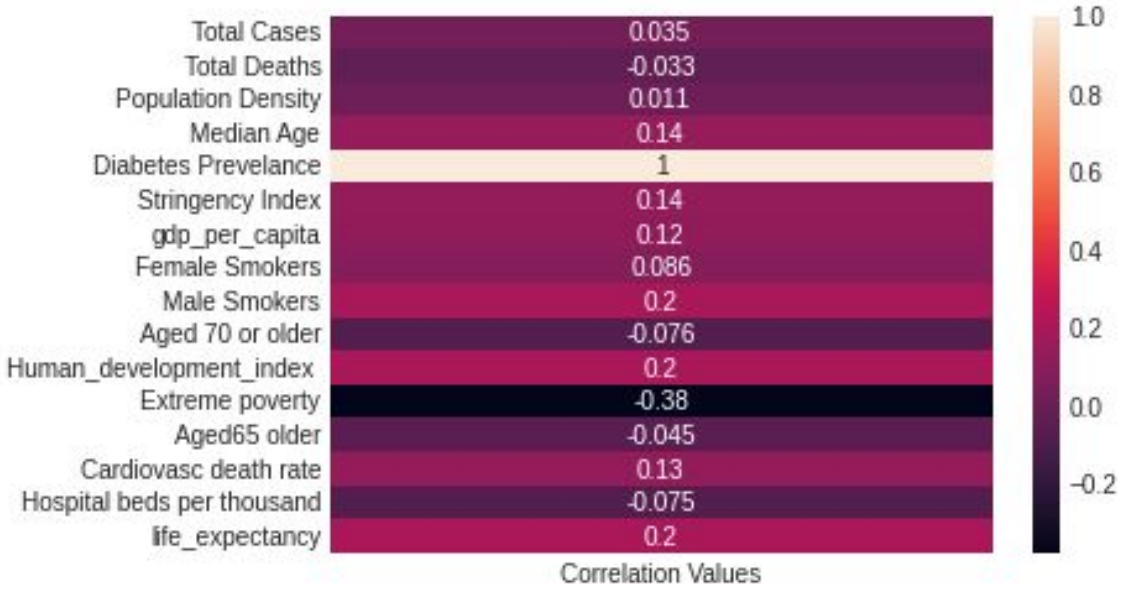
Результаты анализа

Так, оказалось, что относительное число курящих женщин положительно коррелирует с ВВП, продолжительностью жизни, индексом человеческого развития и числом граждан старше 70. Для курящих мужчин таких корреляций нет.

Результаты анализа

Результаты анализа

Еще необычный результат - отрицательная корреляция между отн. числом диабетиков и отн. числом беднейшего населения. Возможно, это связано с тем, что такие слои населения не имеют достаточно средств для покупки лекарств и чаще умирают.



Результаты анализа

Также, чем выше процент бедного населения государства, тем менее остро оно реагирует на распространения заболевания (отрицательная корреляция с индексом строгости).

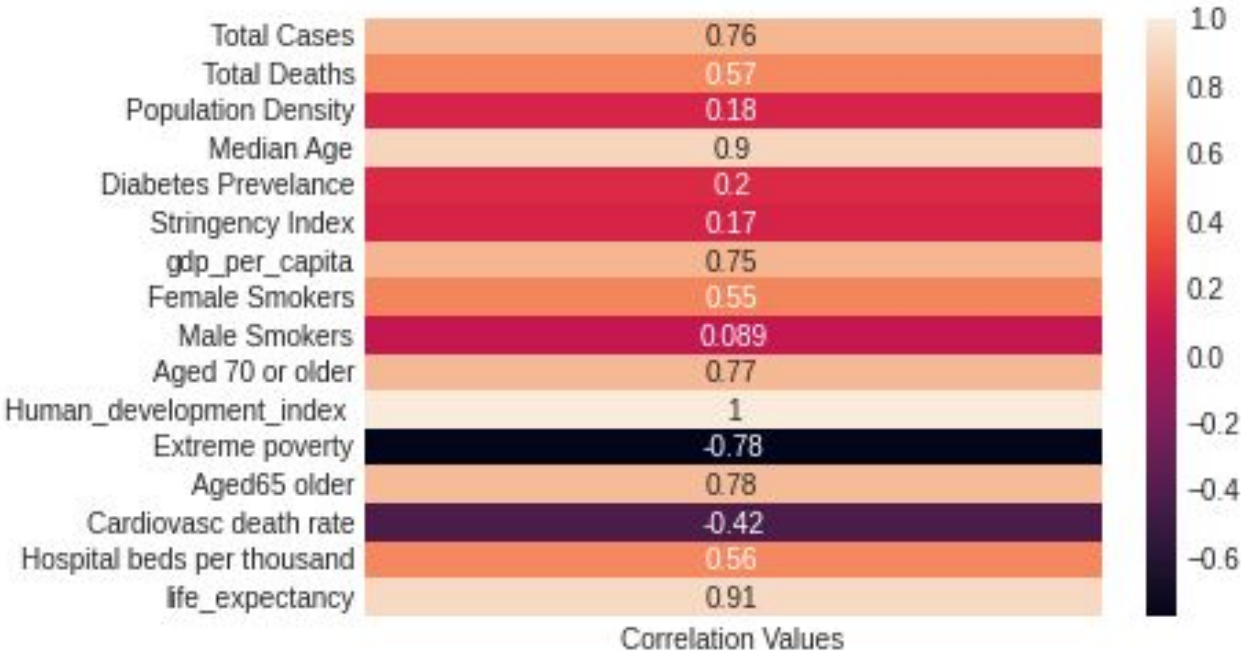
Результаты анализа

Интересна положительная корреляция между ВВП на душу населения и числом заболевших. Причем ВВП также положительно с плотностью населения, а она не коррелирует с числом заболеваний (хотя так кажется чисто интуитивно).

Результаты анализа

Наконец, была обнаружена интересная взаимосвязь между ИЧР и числом заболевших и умерших от коронавируса. Как видно, эти величины положительно коррелируют, несмотря на положительную корреляцию между числом коек и ИЧР (в случае числа смертей).

Результаты анализа



Кластерный анализ

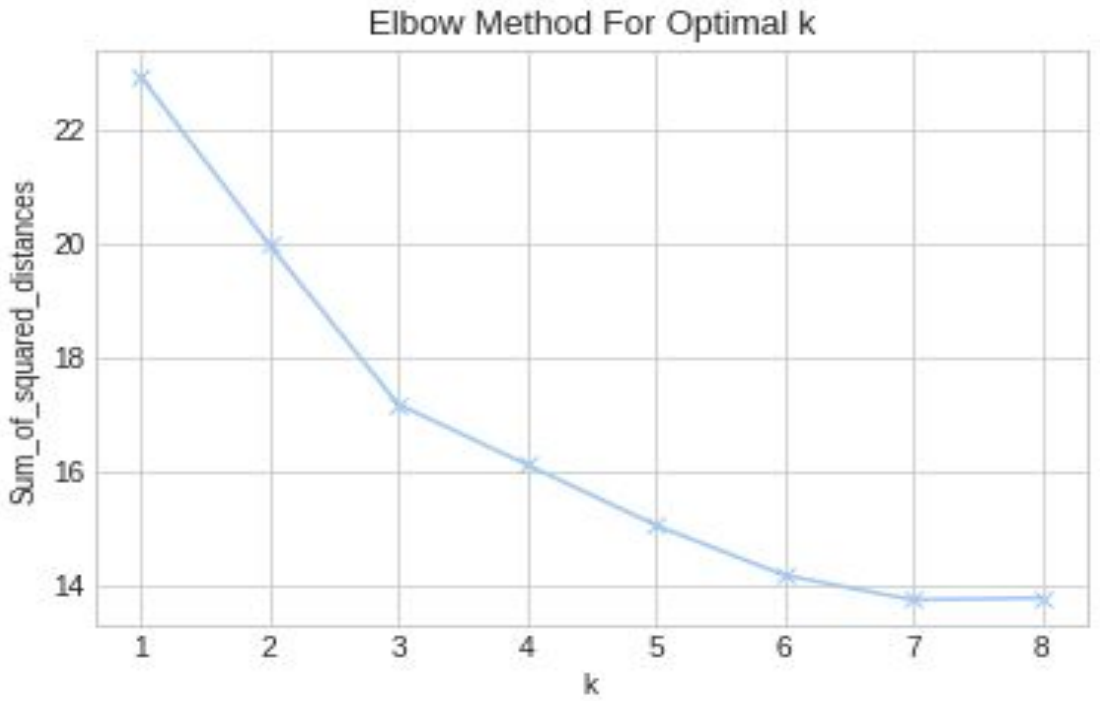
Мне было интересно проанализировать динамику заболеваемости в различных странах и по этому поведению разделить их на группы. Исследовалась величина, соответствующая числу заболевших на 100 человек.

Предварительно данные прошли очистку: был выбран период времени, в течении которого минимальное значение пропусков в данных (с 31 марта 2020 до 29 августа 2022). Также выбирались страны без пропусков (всего 109 стран).

Кластерный анализ

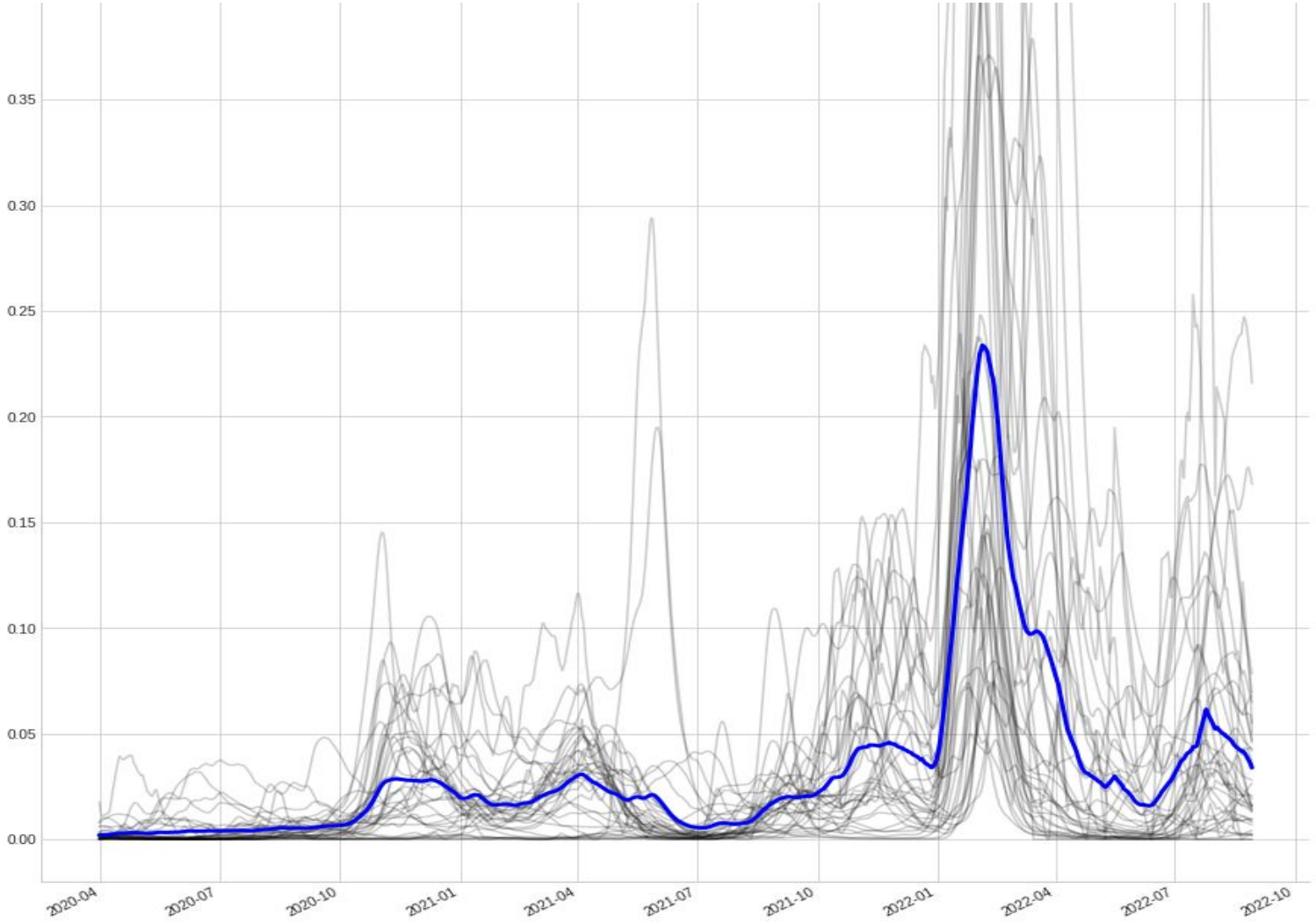
Затем данные были нормализованы. Для кластеризации был выбран алгоритм TimeSeriesKMeans из модуля tslearn для анализа временных рядов. Оптимальное число кластеров подбиралось локтевым методом. Причем задача была разделить кластеры глобально, излишняя точность была не нужна, а так как данные даже после очистки достаточно неоднородны (в частности, из-за шумов или, например, различной политикой фиксирования случаев заболеваний), то оптимальной оценкой послужил первый “изгиб” локтя - число 3.

Результаты анализа



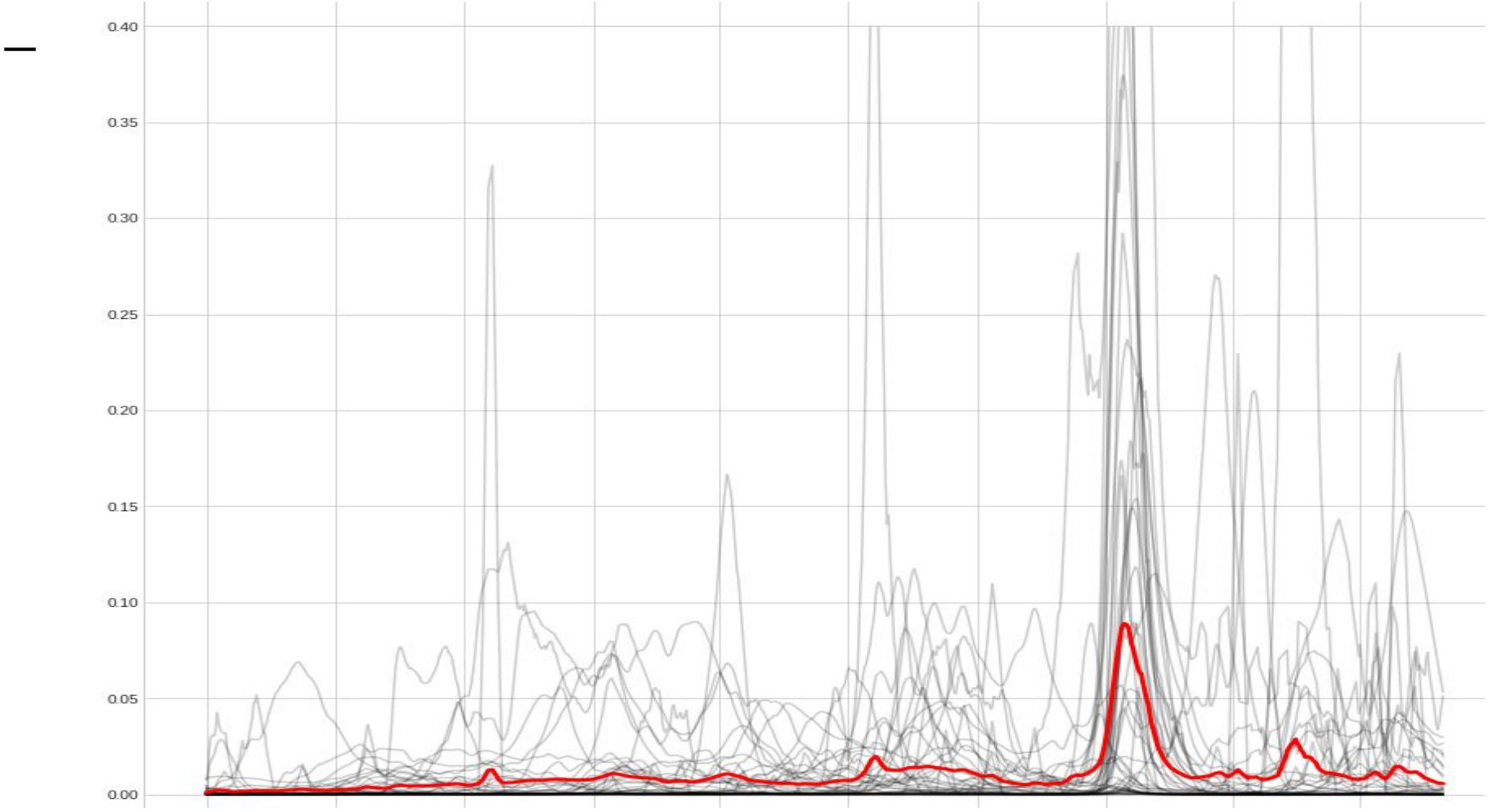
Результаты анализа

Страны разделились на 3 кластера - по 37, 49 и 25 государства. К первому кластеру можно отнести страны, которые плохо справляются с распространением инфекции - в них отчетливо видны высокие бугры, соответствующие различным волнам коронавируса. К этому же кластеру принадлежат страны с максимальными значениями новых случаев.


0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
2020-04
2020-07
2020-10
2021-01
2021-04
2021-07
2021-10
2022-01
2022-04
2022-07
2022-10

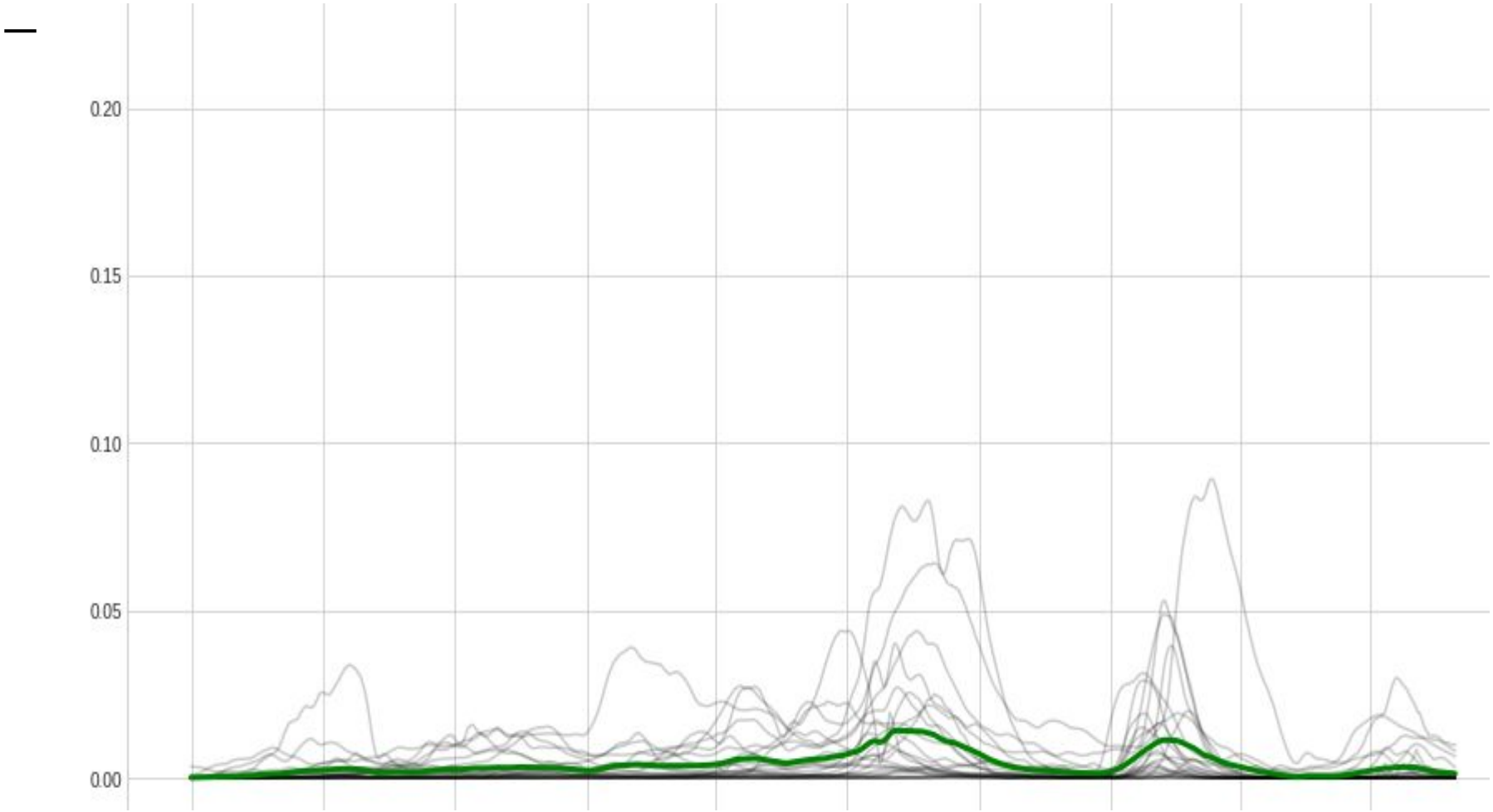
Результаты анализа

Волны коронавируса прослеживаются во всех трех кластерах. Во второй группе находятся страны, которые в основном справились с распространением вируса, за исключением пика заболеваемости в промежутке между январем и апрелем 2022 года.


0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00

Результаты анализа

Наконец, в третьей группе находятся страны, которым удалось сдержать распространение вируса - в этой группе нет высоких пиков, а средняя линия имеет довольно пологие бугры.


0.20
0.15
0.10
0.05
0.00