غزل رفیعی | شماره دانشجویی: ۹۷۲۲۲۰۴۴ | تاریخ: ۲۸/ ۸/ ۱۴۰۰
گزارش تمرین سری ۲ | درس: مبانی علوم داده | استاد درس: دکتر خردپیشه

پیش‌پردازش:

نام دادگان: آمار مبتلایان و مرگ و میر ناشی از ویروس کرونا از روز ۲ فوریه ۲۰۲۰ تا ۶ نوامبر ۲۰۲۱ (۶۲۱ روز)
تعداد ردیف: ۱۳۱.۵۰۰
تعداد ستون:۶۵
نام ستون‌ها:

iso_code
continent
location
date
total_cases
new_cases
new_cases_smoothed
total_deaths
new_deaths
new_deaths_smoothed
total_cases_per_million
new_cases_per_million
new_cases_smoothed_per_million
total_deaths_per_million
new_deaths_per_million
new_deaths_smoothed_per_million
reproduction_rate
icu_patients
icu_patients_per_million
hosp_patients
hosp_patients_per_million
weekly_icu_admissions
weekly_icu_admissions_per_million
weekly_hosp_admissions
weekly_hosp_admissions_per_million

new_tests
total_tests
total_tests_per_thousand
new_tests_per_thousand
new_tests_smoothed
new_tests_smoothed_per_thousand
positive_rate
tests_per_case
tests_units
total_vaccinations
people_vaccinated
people_fully_vaccinated
total_boosters
new_vaccinations
new_vaccinations_smoothed
total_vaccinations_per_hundred
people_vaccinated_per_hundred
people_fully_vaccinated_per_hundred
total_boosters_per_hundred
new_vaccinations_smoothed_per_million
stringency_index
population
population_density
median_age
aged__older

aged__older
gdp_per_capita
extreme_poverty
cardiovasc_death_rate
diabetes_prevalence
female_smokers
male_smokers
handwashing_facilities
hospital_beds_per_thousand
life_expectancy
human_development_index
excess_mortality_cumulative_absolute
excess_mortality_cumulative
excess_mortality
excess_mortality_cumulative_per_million

به طور خلاصه اگر به جدول نگاهی بیندازیم، برخی از آماره‌های آن به صورت زیر خواهد بود.

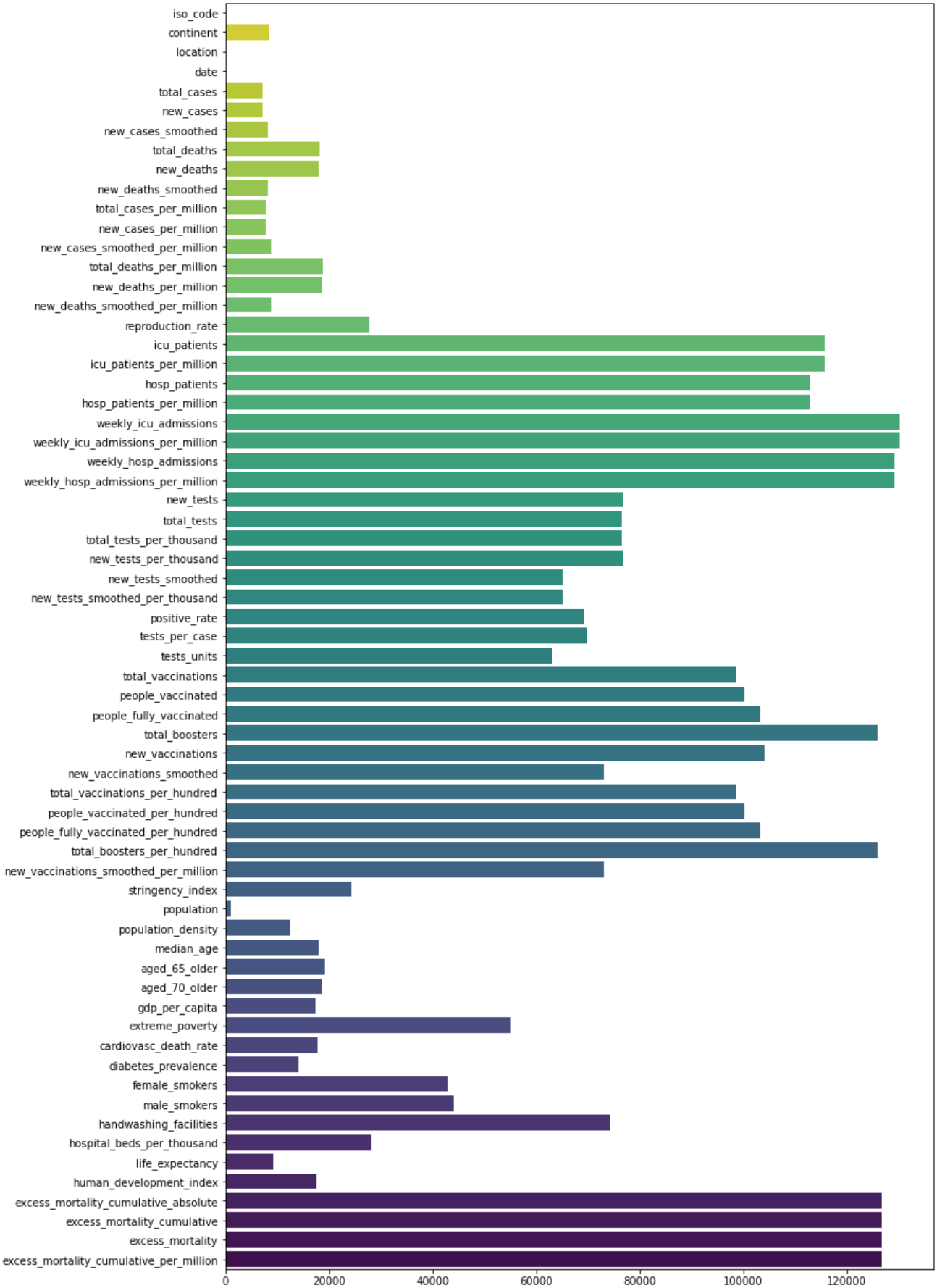
	total_cases	new_cases	new_cases_smoothed	total_deaths	new_deaths	new_deaths_smoothed	total_cases_per_million	new_cases_per_million	new_cases_smoothed_per_million
count	1.243750e+05	124373.000000	123330.000000	1.132980e+05	113494.000000	123330.000000	123736.00000	123734.000000	122696.000000
mean	2.013939e+06	8338.897799	8359.638338	4.996560e+04	184.311972	168.840030	19375.14109	85.686714	85.575258
std	1.150869e+07	43516.048476	43023.843904	2.565859e+05	873.118464	819.030964	32256.71325	197.748216	167.062820
min	1.000000e+00	-74347.000000	-6223.000000	1.000000e+00	-1918.000000	-232.143000	0.00100	-3125.829000	-272.971000
25%	2.347000e+03	3.000000	10.286000	7.900000e+01	0.000000	0.143000	401.88400	0.335000	1.652000
50%	2.652900e+04	104.000000	129.714000	7.270000e+02	2.000000	2.000000	3102.23600	11.294000	15.808500
75%	2.598825e+05	1073.000000	1128.821250	6.416000e+03	22.000000	18.571000	24380.29875	82.823750	92.410000
max	2.495419e+08	907963.000000	826457.571000	5.044839e+06	18007.000000	14703.286000	235692.64300	8620.690000	3385.473000

در این جدول دادگانی که کمترین مقدار آن‌ها منفی‌ست، خطا هستند و می‌بایست به ۰ تبدیل شوند.

برای مرحله‌ی اول تنها ستون های رنگی را انتخاب می‌کنیم و تحلیل باقی ستون‌ها را به مرحله بعد واگذار می‌کنیم.
یکی از دلایل این کار این است که طبق نمودار زیر، اکثر مقادیر ستون‌هایی که انتخاب نکرده‌ایم، پوچ هستند و کاملا قابل اطمینان نیستند یا می‌توان مقدار آن ها را با محاسبات کمی از روی ستون‌های انتخاب شده به دست آورد.
خلاصه مقادیر ستون‌های جدول به دست آمده در جدول زیر آمده است.

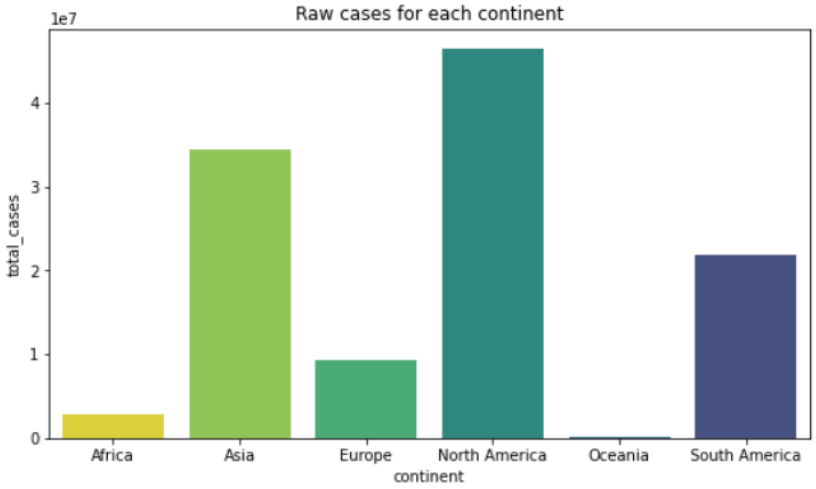
	total_cases	new_cases	new_deaths	population	new_tests	median_age	total_vaccinations
count	1.243750e+05	131500.000000	131500.000000	1.305650e+05	1.315000e+05	113490.000000	1.315000e+05
mean	2.013939e+06	7888.850935	159.139430	1.584128e+08	2.363510e+04	30.501934	1.056160e+08
std	1.150869e+07	42361.064675	813.555158	7.314010e+08	1.291820e+05	9.116176	5.944725e+08
min	1.000000e+00	0.000000	0.000000	4.700000e+01	0.000000e+00	15.100000	0.000000e+00
25%	2.347000e+03	1.000000	0.000000	2.078723e+06	0.000000e+00	22.200000	1.404180e+05
50%	2.652900e+04	77.000000	1.000000	9.749625e+06	0.000000e+00	29.700000	1.204244e+06
75%	2.598825e+05	949.000000	15.000000	3.734479e+07	4.624000e+03	39.100000	1.069756e+07
max	2.495419e+08	907963.000000	18007.000000	7.874966e+09	3.740296e+06	48.200000	7.247579e+09

در این ستون‌ها داده‌های پوچ ۵ ستون اول را با ۰ ، ستون ششم را با میانه و ستون آخر را با داده‌ی قبلی غیر پوچ پر می‌کنیم.

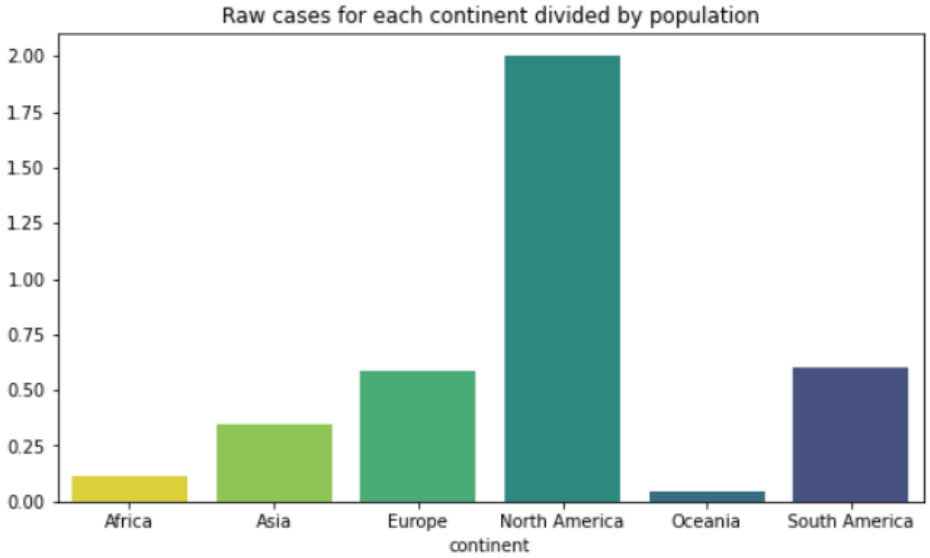

iso_code
continent
location
date
total_cases
new_cases
new_cases_smoothed
total_deaths
new_deaths
new_deaths_smoothed
total_cases_per_million
new_cases_per_million
new_cases_smoothed_per_million
total_deaths_per_million
new_deaths_per_million
new_deaths_smoothed_per_million
reproduction_rate
icu_patients
icu_patients_per_million
hosp_patients
hosp_patients_per_million
weekly_icu_admissions
weekly_icu_admissions_per_million
weekly_hosp_admissions
weekly_hosp_admissions_per_million
new_tests
total_tests
total_tests_per_thousand
new_tests_per_thousand
new_tests_smoothed
new_tests_smoothed_per_thousand
positive_rate
tests_per_case
tests_units
total_vaccinations
people_vaccinated
people_fully_vaccinated
total_boosters
new_vaccinations
new_vaccinations_smoothed
total_vaccinations_per_hundred
people_vaccinated_per_hundred
people_fully_vaccinated_per_hundred
total_boosters_per_hundred
new_vaccinations_smoothed_per_million
stringency_index
population
population_density
median_age
aged_65_older
aged_70_older
gdp_per_capita
extreme_poverty
cardiovasc_death_rate
diabetes_prevalence
female_smokers
male_smokers
handwashing_facilities
hospital_beds_per_thousand
life_expectancy
human_development_index
excess_mortality_cumulative_absolute
excess_mortality_cumulative
excess_mortality
excess_mortality_cumulative_per_million
0
20000
40000
60000
80000
100000
120000

تحلیل داده:

به طور کلی در هر قاره و کشور، چه تعداد مبتلا وجود دارد؟

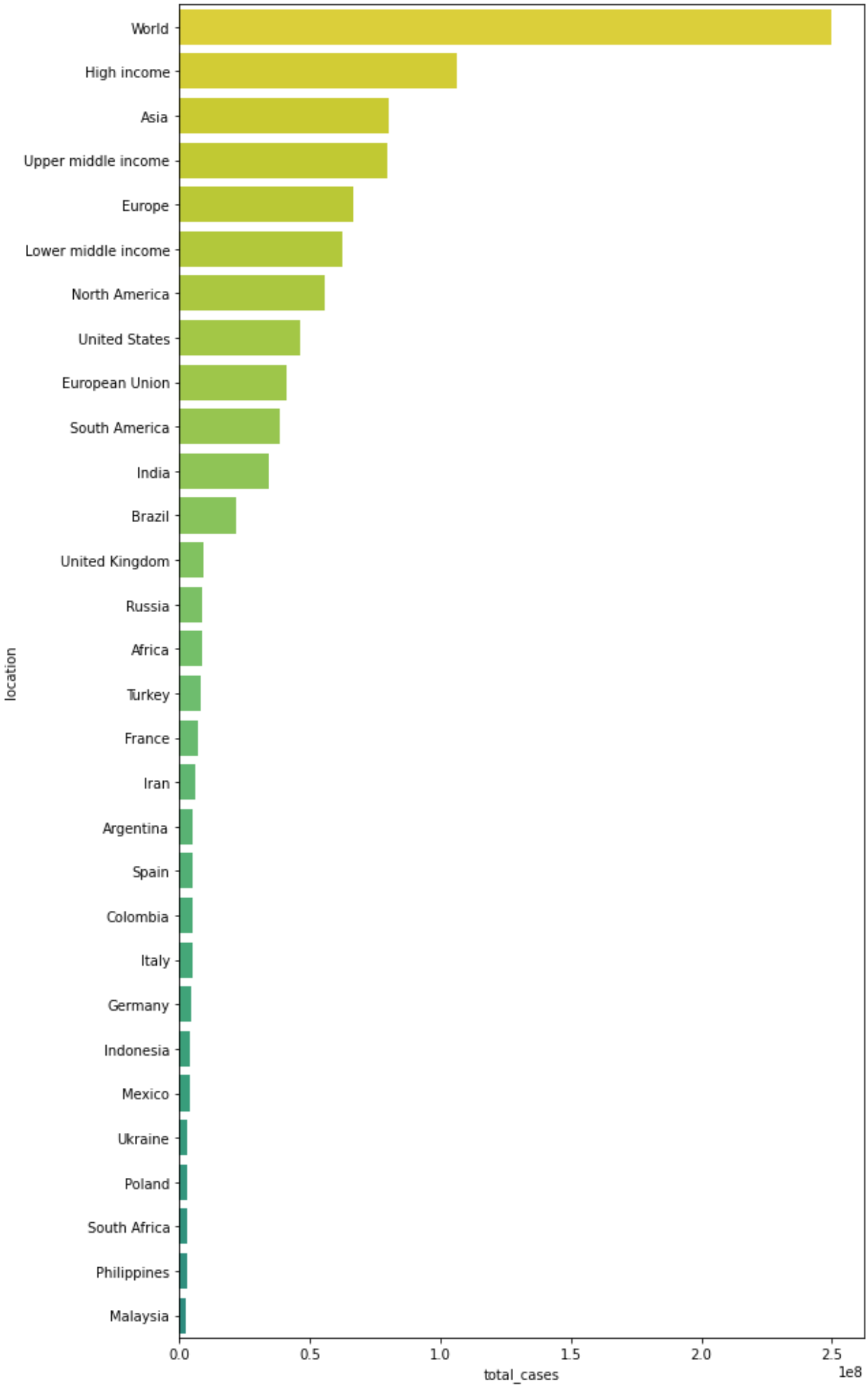


اما این جدول به ما اطلاعات دقیقی نمی‌دهد چون تعداد مبتلایان بر حسب جمعیت یک منطقه است که معنا دارد. در این نمودار، می‌قختوانیم درصد مبتلایان در هر قاره را مشاهده کنیم.



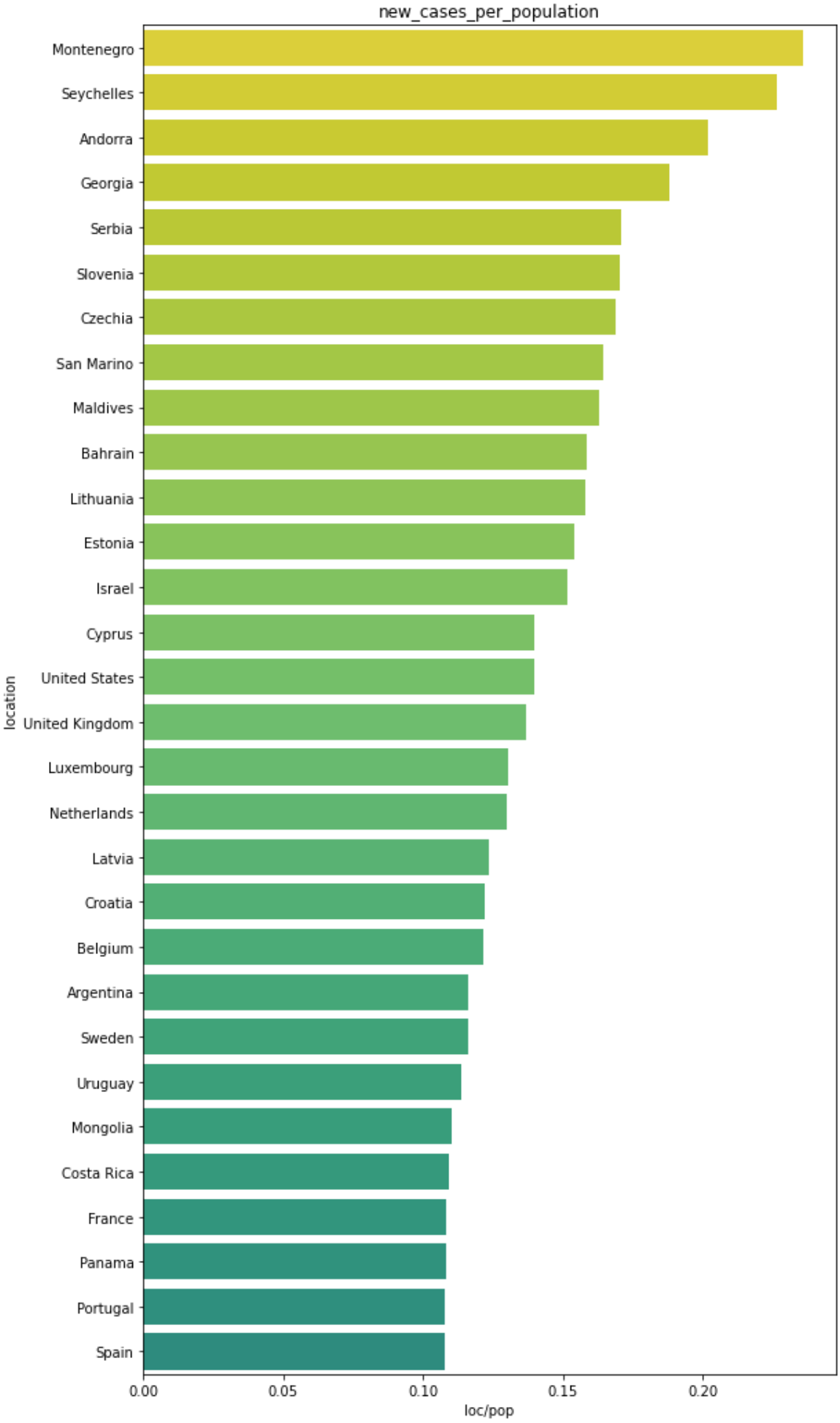
بنابراین به ترتیب آمریکای شمالی، آمریکای جنوبی، اروپا، آسیا، آفریقا و در نهایت قاره استرالیا بیشترین درصد ابتلا را داشته‌اند.

حال به طور جزئی‌تر در مورد کشورهای مختلف این آمار را بررسی می‌کنیم.


World
High income
Asia
Upper middle income
Europe
Lower middle income
North America
United States
European Union
South America
India
Brazil
United Kingdom
Russia
Africa
Turkey
France
Iran
Argentina
Spain
Colombia
Italy
Germany
Indonesia
Mexico
Ukraine
Poland
South Africa
Philippines
Malaysia
location
0.0
0.5
1.0
1.5
2.0
2.5
1e8
total_cases

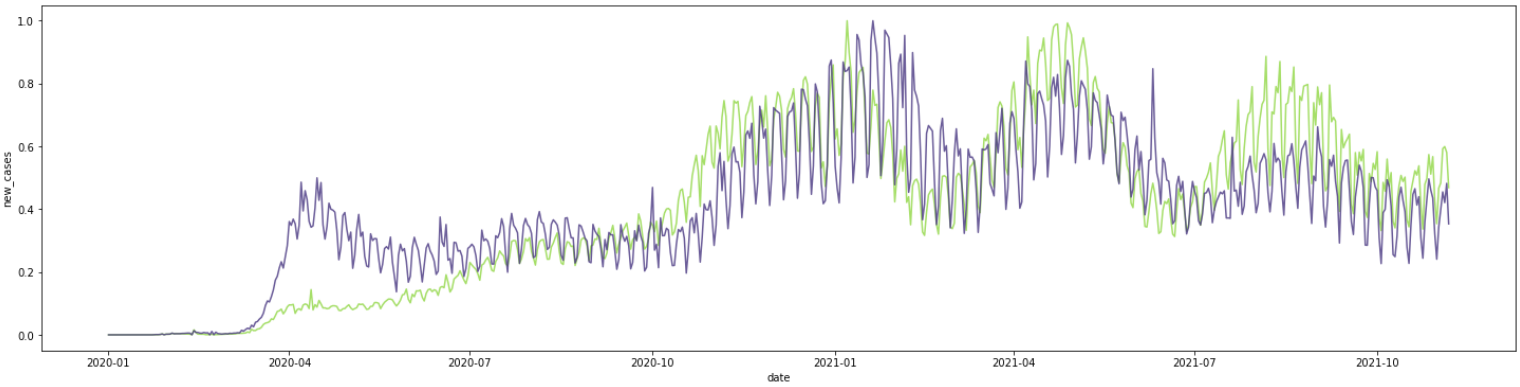
به دلیل تعداد بالای کشورها، تنها برای ۳۰ کشور اول رسم شده است.
توجه داشته باشید در این نمودار علاوه بر کشورها، آمار کلی جهان، قاره‌ها، و افراد با حقوق بالا، متوسط و کم هم در نظر گرفته شده است.
در ادامه‌ی تحلیل این ردیف‌ها را از جدول حذف نمی‌کنیم زیرا در نمودارها تاثیری ندارند. در واقع هر آمار یک بار در کشور، یک بار در قاره، در جهان، و یک بار در ردیف حقوق‌ها آورده شده است و جمعا هر نفر در ۴ بار آورده شده است.
در مقایسه‌ی نسبت میزان مبتلایان به میزان حقوق دریافتی، می‌توان حدس زد که این اتفاق به این دلیل است که افراد دارای حقوق بالاتر، دسترسی بیشتری به امکانات بهداشتی و تست کرونا و بیمارستان‌ها دارند یا در کشورهای پیشرفته‌تری زندگی می‌کنند، به همین دلیل تست بیشتری از آن ها گرفته شده. ممکن است افرادی که حقوق کمتری دارند و به امکانات دسترسی ندارند هم بیمار شده باشند، اما از آن ها تست گرفته نشده باشد یا علت مرگ و میر آن‌ها به نام کرونا ثبت نشده باشد.

حال این نمودار را برای درصد ابتلا، و نه تعداد خالص رسم می‌کنیم.

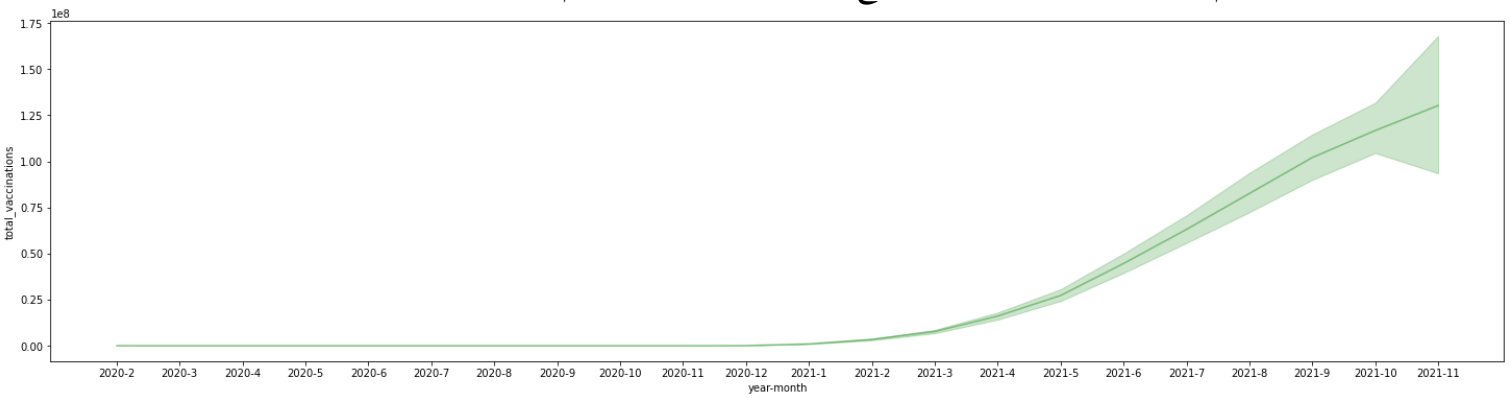

new_cases_per_population
Montenegro
Seychelles
Andorra
Georgia
Serbia
Slovenia
Czechia
San Marino
Maldives
Bahrain
Lithuania
Estonia
Israel
Cyprus
United States
United Kingdom
Luxembourg
Netherlands
Latvia
Croatia
Belgium
Argentina
Sweden
Uruguay
Mongolia
Costa Rica
France
Panama
Portugal
Spain
location
0.00
0.05
0.10
0.15
0.20
loc/pop

کشورهایی که در این نمودار دارای بیشترین درصد مبتلایان هستند، در واقع کشورهای کوچکتر و با تراکم بیشتری هستند.

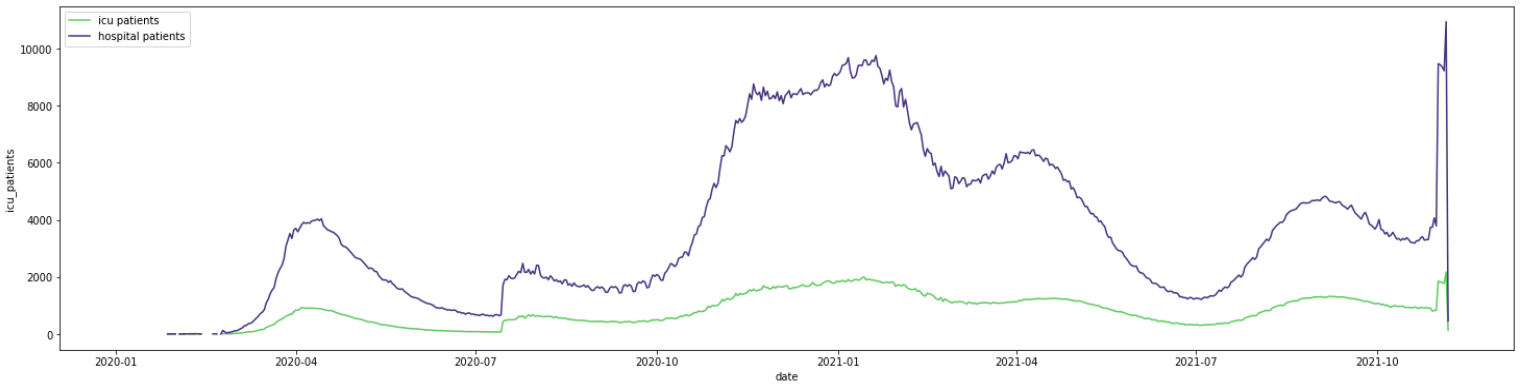
در نمودار زیر می‌توانیم تعداد کیس‌های جدید و در مقابل آن تعداد مرگ و میر جدید را در طول این ۲ سال مشاهده کنیم.



رنگ بنفش مربوط به مرگ و میر، و رنگ سبز مربوط به تعداد تست‌های جدید است.
توجه کنید برای اینکه بتوان این دو داده را مقایسه کرد، هر دو نرمال شده اند.
در ابتدا مشاهده می‌کنیم که مرگ و میر شاهد قله‌ای بوده که در نمودار سبز رنگ وجود ندارد. علت این پدیده می‌تواند این باشد که در ابتدا این ویروس ناشناخته بوده و کشورها هنوز قرنطینه را اعمال نکرده‌اند. همچنین تست‌های کرونا به اندازه کافی در دسترس همه نبوده.
سپس از ۲۰۲۱-۰۱ تا ۲۰۲۱-۱۰ شاهد سه قله هستیم که هر دو نمودار سبز و بنفش تجربه کرده‌اند و مطابق انتظار است. اما در قله‌ی سوم مشاهده می‌کنیم که قله‌ی مرگ و میر کوتاهتر از تعداد مبتلایان است و علت این قضیه می‌تواند این باشد که در این مدت تعداد واکسن زیادی به افراد تزریق شده و ایمنی نسبی به وجود آمده.
در زیر می‌توانیم تعداد افراد واکسینه‌شده تا تاریخ‌های مشخصه شده را ببینیم(به طور تجمعی):

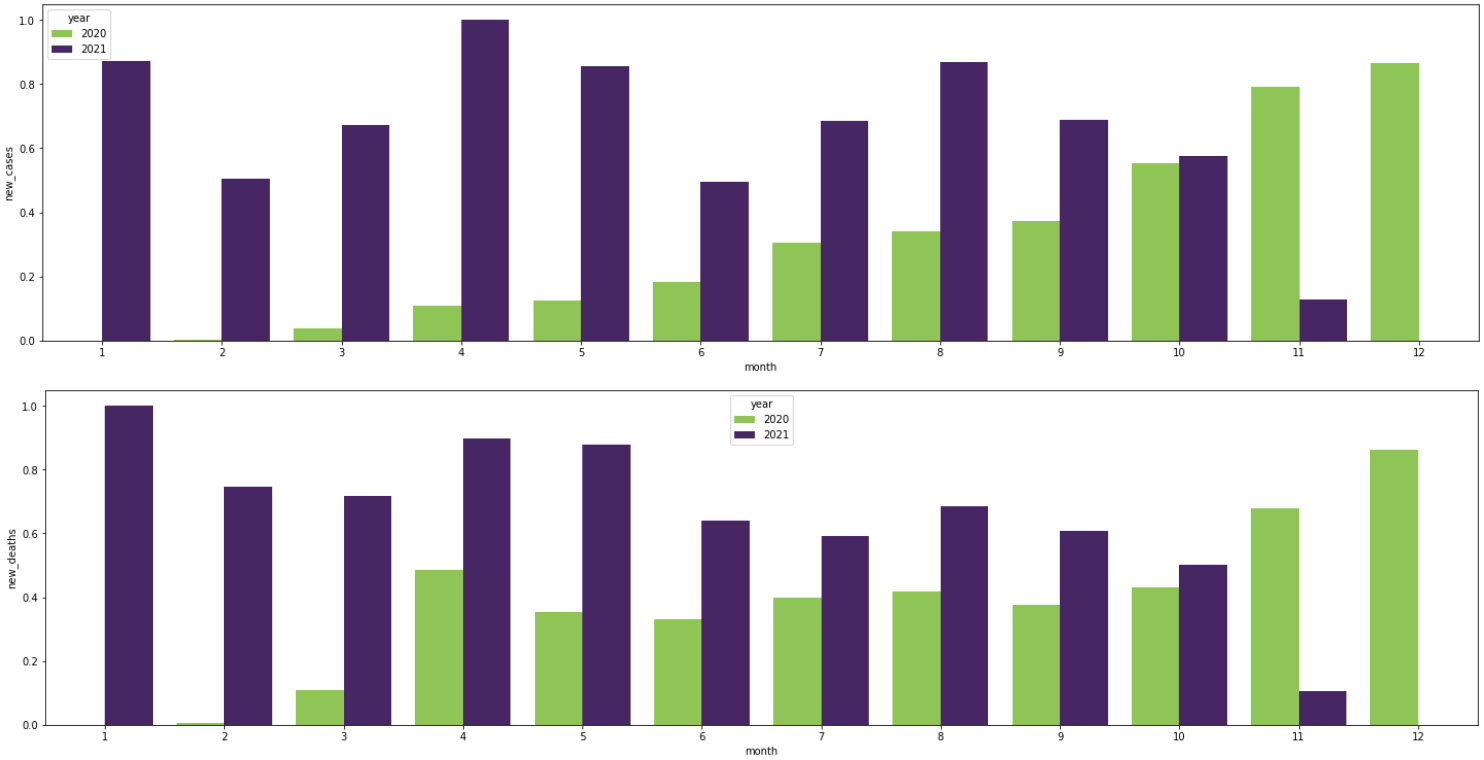


همچنین نمودار تعداد کسانی که در بیمارستان بستری شده‌اند، در مقابل کسانی که به بخش icu رفته‌اند در زیر آمده.



مقایسه‌ی آمار ۲۰۲۱ و ۲۰۲۰

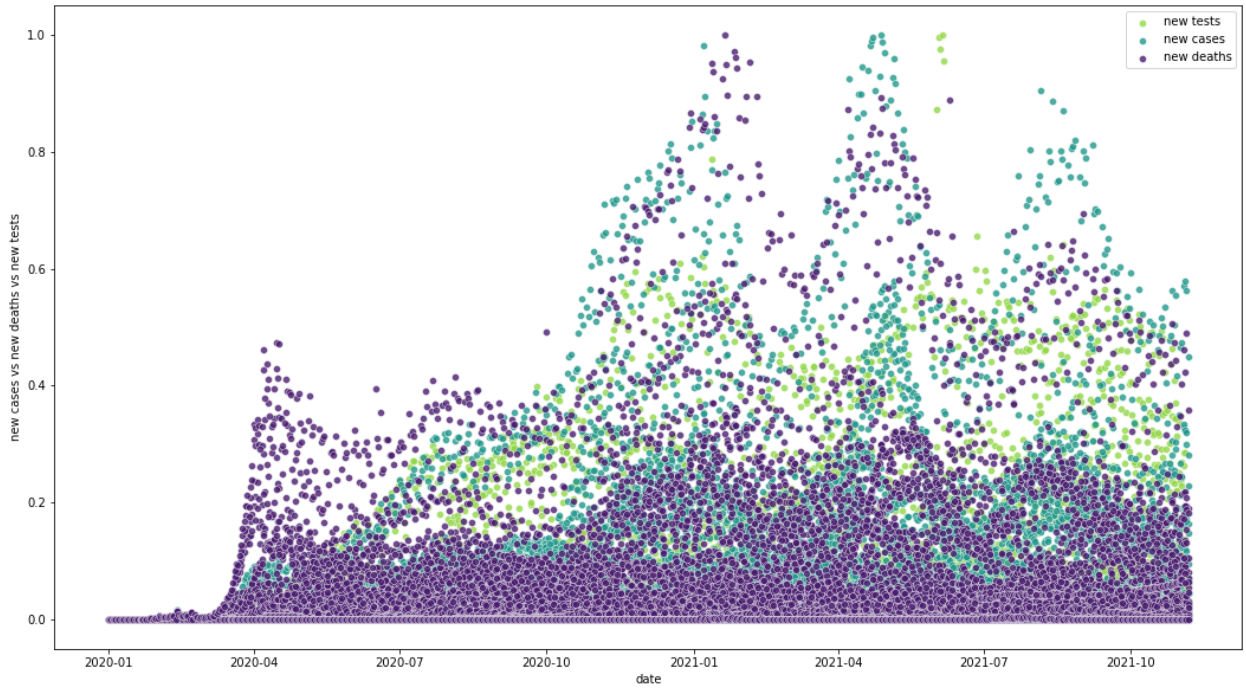
تعداد افراد مبتلا به کرونا در سال ۲۰۲۱ و ۲۰۲۰
رنگ سبز مربوط به ۲۰۲۰ و بنفش مربوط به ۲۰۲۱ است.
برای مقایسه‌ی صحیح هر دو آمار نرمال شده اند.



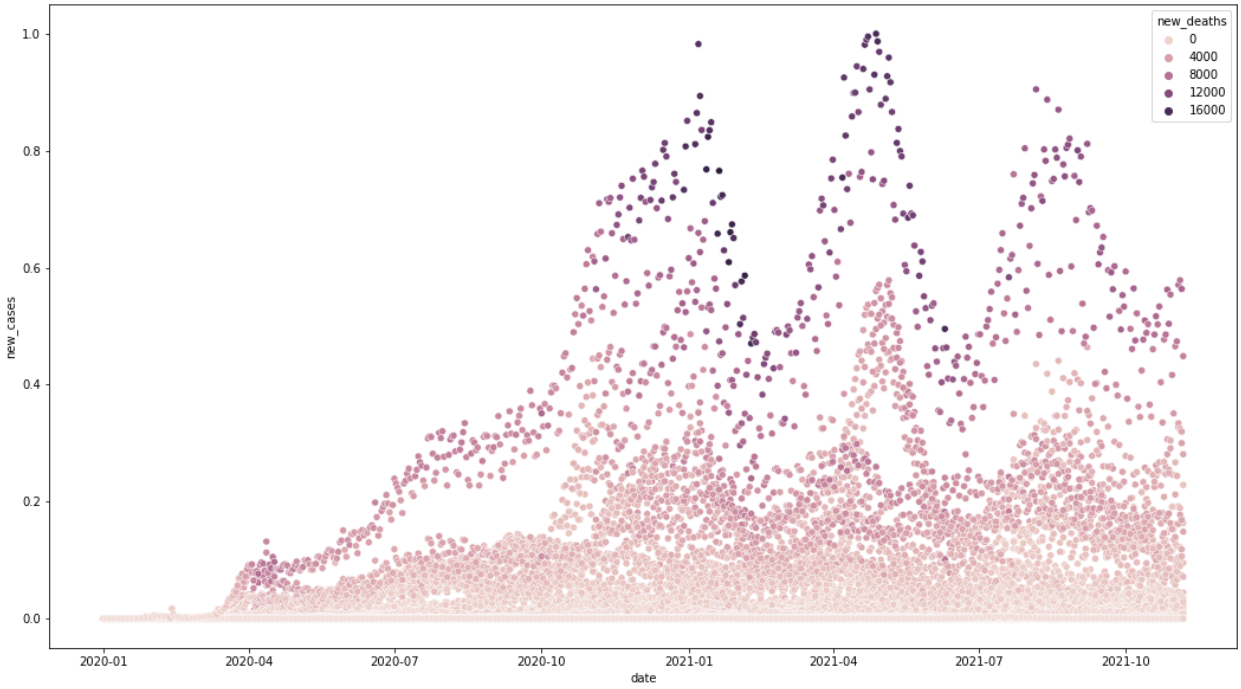
همانطور که مشاهده می‌کنید هر دو آمار مبتلایان و مرگ و میر در سال ۲۰۲۱ به طور کلی از ۲۰۲۰ بیشتر بوده، اما در ماه‌های ۱۱ و ۱۲ در سال ۲۰۲۰ آمار بیشتری داشته ایم که بعد از آن آمار کمی افت داشته است.

بررسی چند همبستگی

می‌خواهیم بررسی کنیم آیا افزایش و کاهش تعداد تست‌ها، مرگ و میر، و مبتلایان جدید به طور مشابه پیش می‌رود یا خیر. یعنی به عبارت دیگر ممکن است این شبهه به وجود بیاید که چون تعداد تست‌های بیشتری گرفته شده، در نتیجه تعداد مبتلایان هم بیشتر شده. در این نمودارها برای مقایسه نرمال شده‌اند و محور y نشان‌دهنده درصد است.
اگر به رنگ سبز کمرنگ در شکل نگاه کنیم، متوجه می‌شویم که مرگ و میر و مبتلایان جدید، دارای قله‌ها و دره‌های مشابه هستند اما تعداد تست‌های جدید از آن‌ها پیروی نمی‌کند. بنابراین در واقعیت هم تعداد مبتلایان بیشتر شده.



در نمودار زیر مشاهده می‌کنیم در روزهایی که تعداد کیس‌های جدید بیشتر بوده، به مناسب آن تعداد مرگ و میر هم افزایش یافته(یعنی در پایین نمودار رنگ روشن و در بالا رنگ تیره‌تر داریم).
هم‌چنین اگر به چندین روز اخیر نگاه کنیم، متوجه می‌شویم مرگ و میر امروزه، در مقایسه با روزهایی که تعداد کیس جدید یکسان داشته‌ایم، کاهش یافته(رنگ روشن‌تر). درنتیجه می‌توان نتیجه گرفت از کشنده بودن این ویروس کم شده، یا افراد زیادی در مقابل این ویروس ایمن شده‌اند (یا به دلیل ابتلای قبلی به این بیماری، یا تزریق واکسن).



در این شکل نیز به طور دقیق‌تر می‌بینیم در روزهایی که تعداد کیس‌ها بیشتر بوده، لزوما تعداد تست بیشتری از افراد گرفته نشده.