به نام خدا

محمد زیاری – (97222047)

تاریخ: 28/ 8/ 1400

گزارش تمرین سری 2 **درس: مبانی علوم داده** **استاد درس: دکتر خردپیشه**

در این دیتاست ما با داده هایی مرتبط با ویروس کرونا از کشورهای متفاوت مواجه هستیم. پس از خواندن داده ها ، برای از بین بردن داده های null باید روش های متفاوتی را در هر ستون پیاده سازی کنیم.
طبق document میدانیم بعضی از کشورها در داده هایی که منتشر می کنند ، یک سری از آمارها را منتشر نمی کنند و از آنجا که ما داده های null نمیخواهیم باید ترتیبی اتخاذ شود تا روی ستونها تصمیم گیری درستی انجام شود. نتیجه گیری هایی که یادداشت می شود از بررسی چند ستون و سطر به صورت رندم و روند آنها است که در کد وجود ندارد.

ستون هایی که مربوط به Confirmed cases و Confirmed deaths بودند و مقدار null داشتند و همچنین برچسب new داشتند را با توجه به بررسی که روی داده ها داشتم با 0 پر میکنم. چون بیشتر آن کشورها به منظور دادن مواردی غیر مرتبط با این 2 مورد ، آماری را فرستادند و یا این موارد مربوط به روزهای اولیه ی دادن اطلاعات بوده است. به همین منظور فرض می شود موارد دیگر 0 بوده است. اما ستون هایی که total داشتند در بعضی روزهای وسط پر نشده اند و مقادیر null دارند. از این جهت نمیتوان آنها را با صفر پر کرد چون مطمئنا پس از 5 ماه total_case یک کشور null نیست. من برای پر کردن این داده ها از متد bfill استفاده کردم و آنها را با داده های سطر های بعدیشان پر کردم.

در ارتباط با ستون های stringency_index و reproduction_rate ، اولی را بیشتر کشورهایی که پیشرفته نبودند گزارش نکردند یا گزارش هایی که مربوط به روزهای اولیه کرونا می شد پس قاعدتا باید سختگیری کمتری در آنها صورت می گرفت پس null ها را با 0 پر میکنیم. مواردی که reproduction_rate را گزارش نکرده بودند نیز بیشتر به روزهایی که مرگ و میر کمی داشت مربوط بود یا مکان های غیر مشخصی بودند که در آخر کار از جدول حذف میشوند. پس منطقی است که آنها را نیز با 0 پر کنیم.

موارد مورد گزارش درباره Hospital & ICU بسیار کم بود ، به همین علت پر کردن آن به صورت ساده با میانگین و میانه و ... منطقی به نظر نمی رسد. از این جهت این ستون ها از جدول پاک شده است .

ستون های مربوط به Excess mortality نیز در کشور های بسیار کمی آمارش گزارش میشود و به مانند بخش قبلی از جدول کنار گذاشته میشود و در آخر به صورت کوتاهی به تاثیر آن پرداخته می شود.

در بخش Tests & positivity نیز با بررسی به نظر می آمد آنهایی null بودند که تستی نگرفته بودند. پس در ستون های عددی آنها که برچسب new داشتند 0 گذاشتیم و بقیه موارد که total یا نسبتی بودند (مثلا test_per_case) را بر اساس سطرهای بعدیش پر کرده ایم. در test_units نیز No Where را قرار دادیم تا با بقیه متمایز باشد. تنها یک مورد در اینجا مشکل دارد ، آن هم آنکه برخی داده ها که در test_smoothed مقداری غیر null دارند و در بقیه جاها null که بنابر document به علت نوع گزارششان است ، هم در بقیه ستونهایش صفر قرار داده شد که اگر وقت بیشتر بود یا کد حرفه ای تر بود میشد بقیه ستون های این موارد را نیز به اقتضا (مثلا با همان مقدار smoothed یا به هر شکل دیگری جز 0) پر کرد.

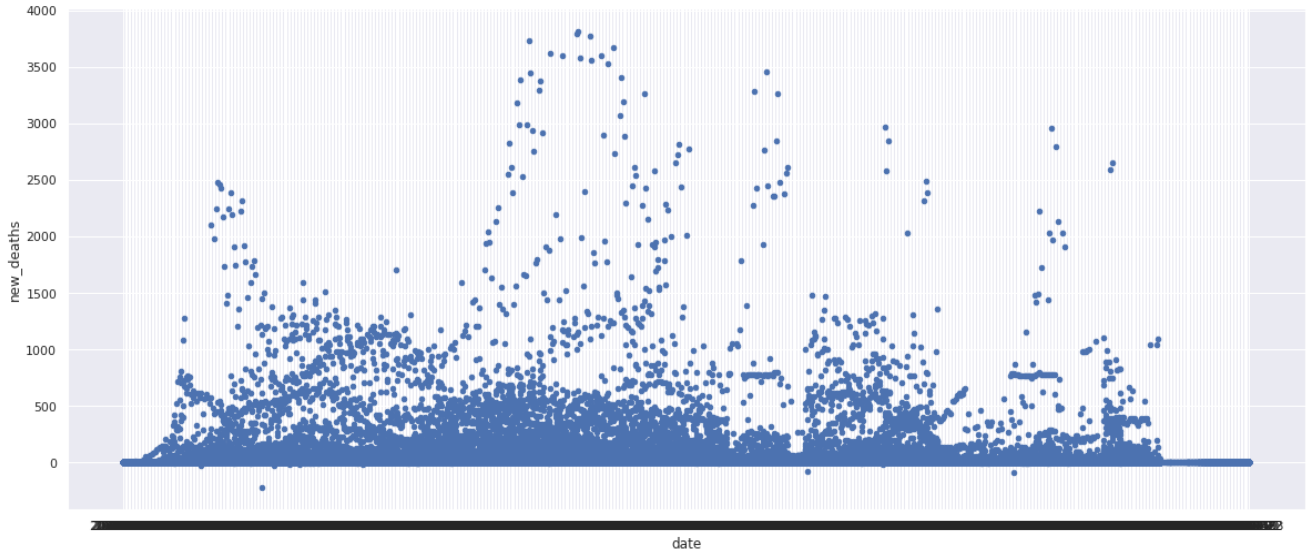
در بخش Vaccinations برخی داده هایی که ارسال می شدند تنها در بخش smoothed موارد واکسیناسیون شان را مطرح میکردند. با بررسی روی داده ها (به طور مثال کشور افغانستان) مشخص شد بعضی کشورها total و یا new را در برخی موارد آپدیت نمی کردند و null می گذاشتند. از این جهت راهکار درست از نظر من آن بود که این سطرهای با داده های سطرهای بعدی پر شوند که از متد bfill استفاده میشود. اما پس از این کار این طور به نظر می رسید که بعضی ها از هفته اول شیوع ، واکسینه شده اند که مطمئنا اشتباه است. از این جهت با 0 پر می شوند . هرچند یک ایده درست آن است که برای هر کشور اولی ها با صفر پر شوند و پس از اولین واکسیناسیون با bfill پر شوند تا total منطقی بماند اما من چنین روشی را پیاده سازی نکردم.
اما booster ها را خیلی از کشورها استفاده نکرده اند یا گزارش نکرده اند. از این جهت این ستونها ، null هایشان با 0 پر شد.

در بخش others اول از همه آن سطرهایی که continent شان null بود را از دیتاست drop کردیم . ستون iso_code از دیتاست drop شده است . در ستون های gdp_per_capita , extreme_poverty بر

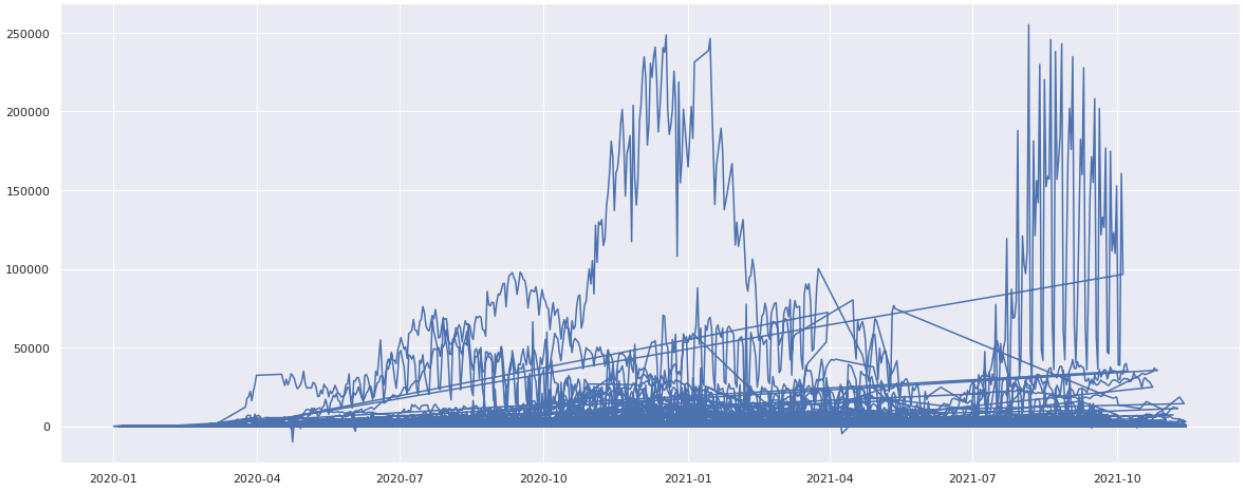
اساس مشاهدات روی جدول ، بیشتر کشورهای فقیرتر آن را گزارش نکرده بودند و پر کردن آن با mean شاید منطقی به نظر نمی رسید از این جهت اولی را با mean*2 و دومی را با mean/2 پر کردیم. مطمئنا مشکلاتی در این نوع پر کردن وجود دارد اما راه حل منطقی تری به نظر می رسید. در بقیه ستون ها هم از میانگین استفاده شده است.

برای حذف کردن outlier داده ها بر اساس هر ستون ، اگر آن داده از میانگین هر ستون به اندازه 3 برابر std فاصله داشت آن سطر را از جدول جدا میکنیم. مشکل در این نوع جداسازی آن است که برای مثال طبق بررسی من فقط 7 سطر از کشور آمریکا باقی ماند و باقی موارد به علت اینکه از میانگین دور بودند حذف شدند. حالا چه به علت مرگ و میر بالا چه به علت واکسیناسیون بیشتر و ... . پس من در این مورد کد پاک کردن outlier ها را کامنت کردم و از آنها استفاده نکردم ، چون بعضی از همین داده های پرت به درد ما میخورند. ذکر این نکته ضروری است که قاعدتا case و death ها واریانس بالایی دارند و از این رو به نظرم نمیشود داده هایی که مقدار بالا یا بسیار پایینی را دارند حذف کرد.
اما ایده جدیدی که به ذهنم رسید آن بود که بر اساس کشورها ، داده های پرت را مشخص کنیم. یعنی کشور به کشور جلو برویم و داده های outlier را حذف کنیم . در نتیجه این کار نزدیک به 38 هزار سطر از داده ها از جدول پاک شدند.

پس از انجام این کار ها نوبت به بررسی رو ستون های جدول و روابط آن هاست. برای این کار اول از همه از date شروع کرده ایم. نمودار زیر scatter plot ای از new_death ها طی روند زمانی را نشان می دهد که مشخصا در اواسط پیدایش کرونا این آمار بیشترین مقدار ها را دارد ، اما پس از مدتی و پس از واکسیناسیون دوباره نزولی تر شده است :

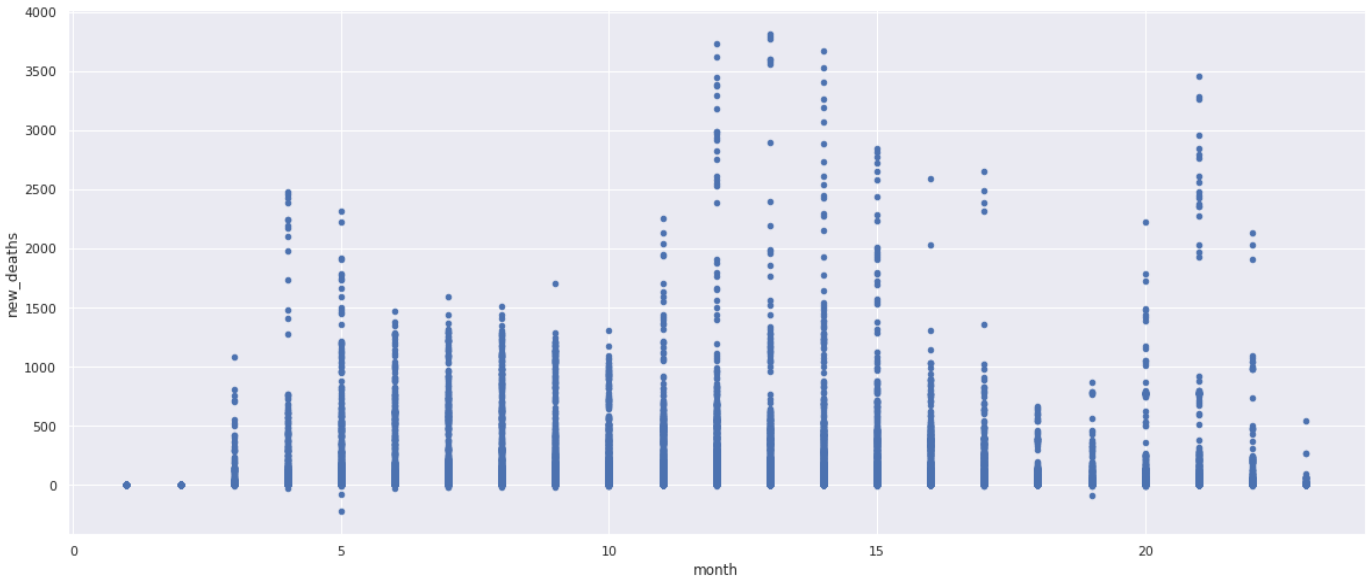


اما نمودار زیر new-cases را در طول زمان نشان می دهد که مشخص است روند مشخص و خیلی جالبی ندارد. هرچند در 2 ناحیه از جدول پیشرفت زیادی در میزان new-cases شاهد هستیم.



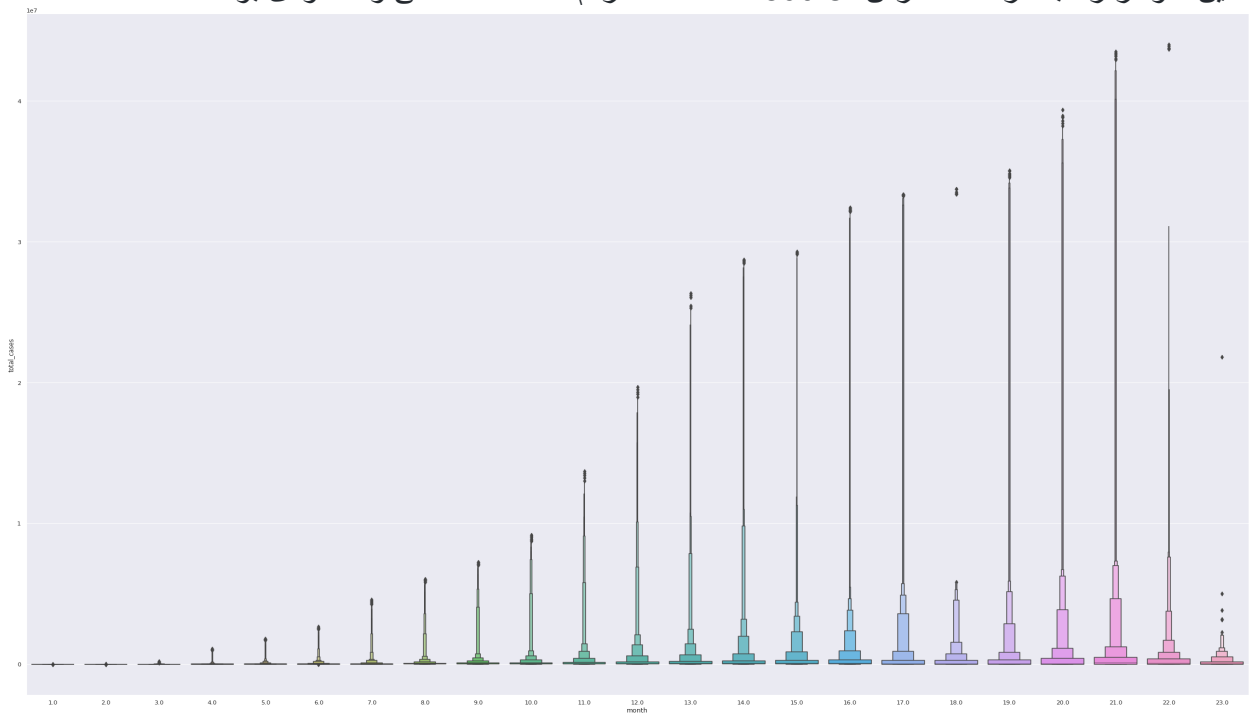
ایده ای که پیاده سازی کردم تا نمودارها منطقی تر به نظر بیاید آن بود که ستون جدیدی تحت عنوان month به جدول اضافه کردم و می خواهیم موارد مختلف را طی ماه های مختلف با هم بسنجیم.

پس از ساختن ستون month نمودار new_death را بر اساس ماه ها در شکل زیر مشاهده می کنیم که مشخص تر است در اواسط کرونا میزان بیشتری داشته است.(ماه های ابتدایی سال 2021)

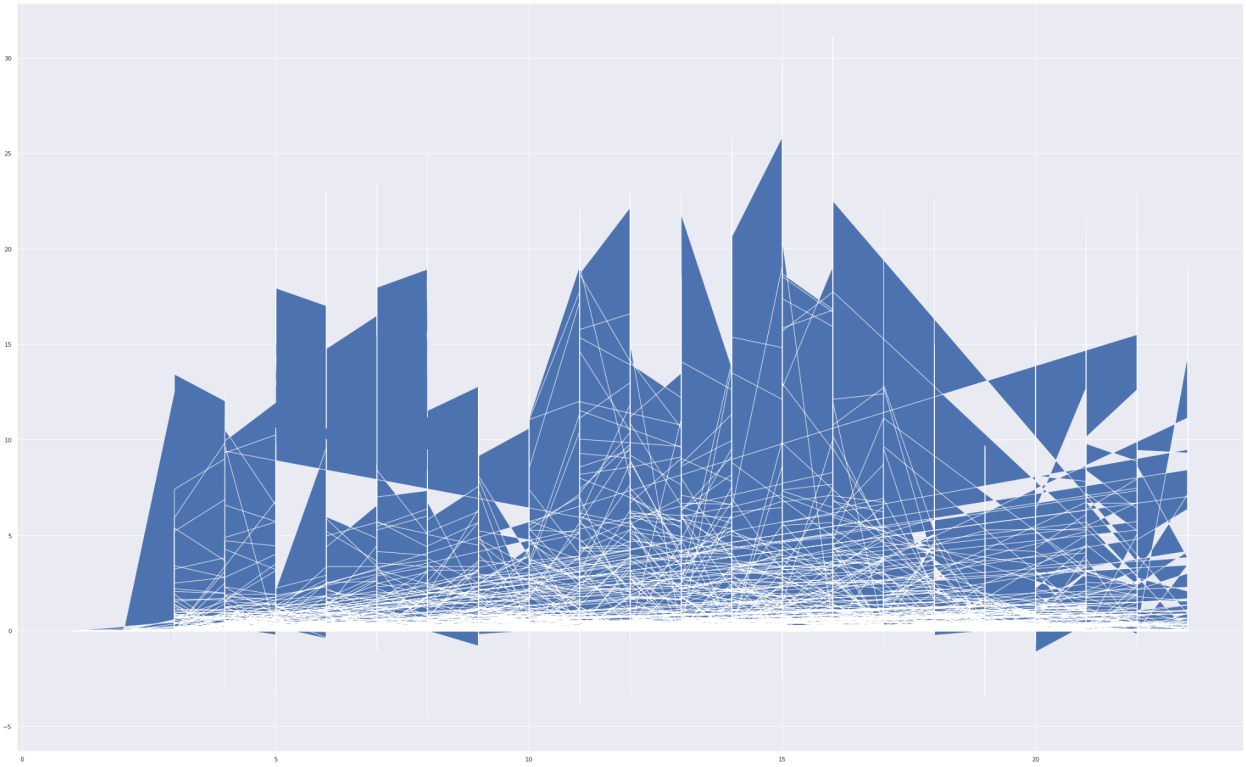


نمودار زیر total_cases را در گذر ماه ها نمایش می دهد. مشخص است طی گذر ماه ها total_cases باید افزایشی باشد اما نکته مهمی که در بالاتر هم به آن اشاره کردم آن است که با حذف outlier ها

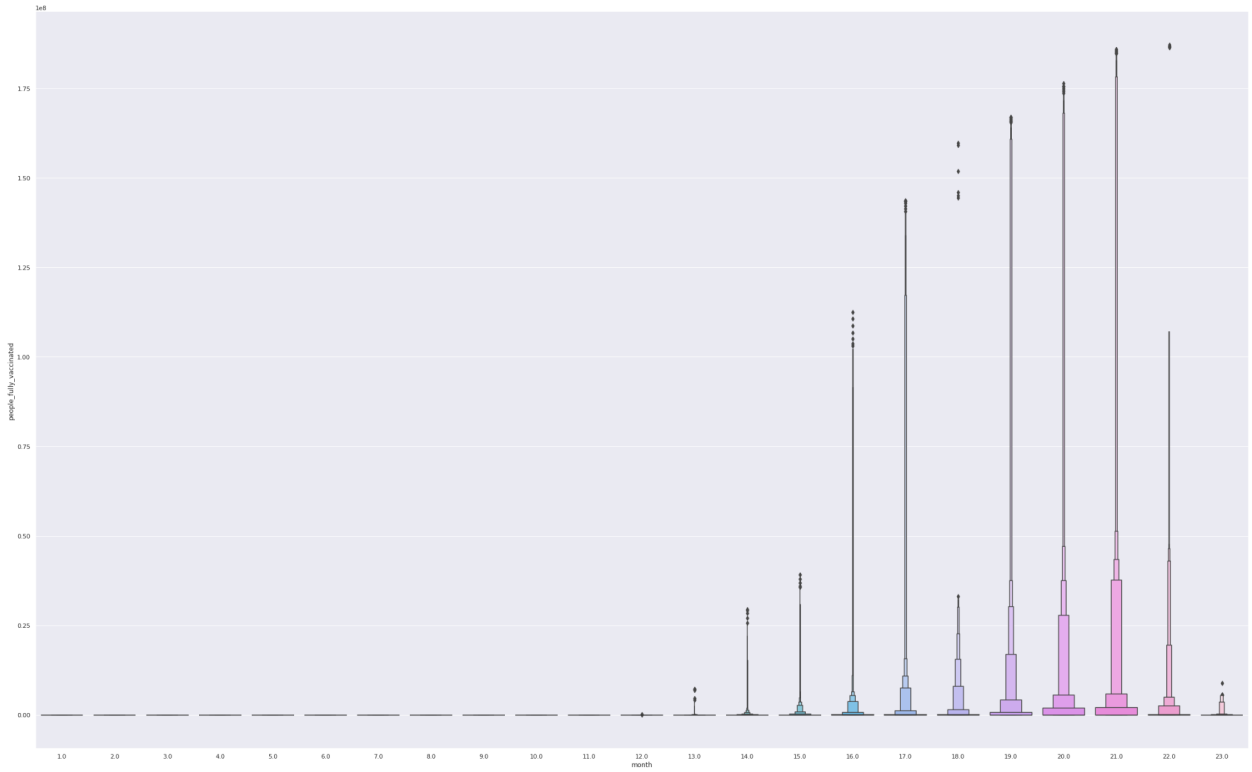
total_cases هایی که مقادیر بالاتری داشتند و به روزهای پایانی مربوط بودند از جدول حذف شدند. چون من این نمودار را قبل از حذف کردن outlier ها مشاهده کردم که کاملا منطقی و صعودی بود.



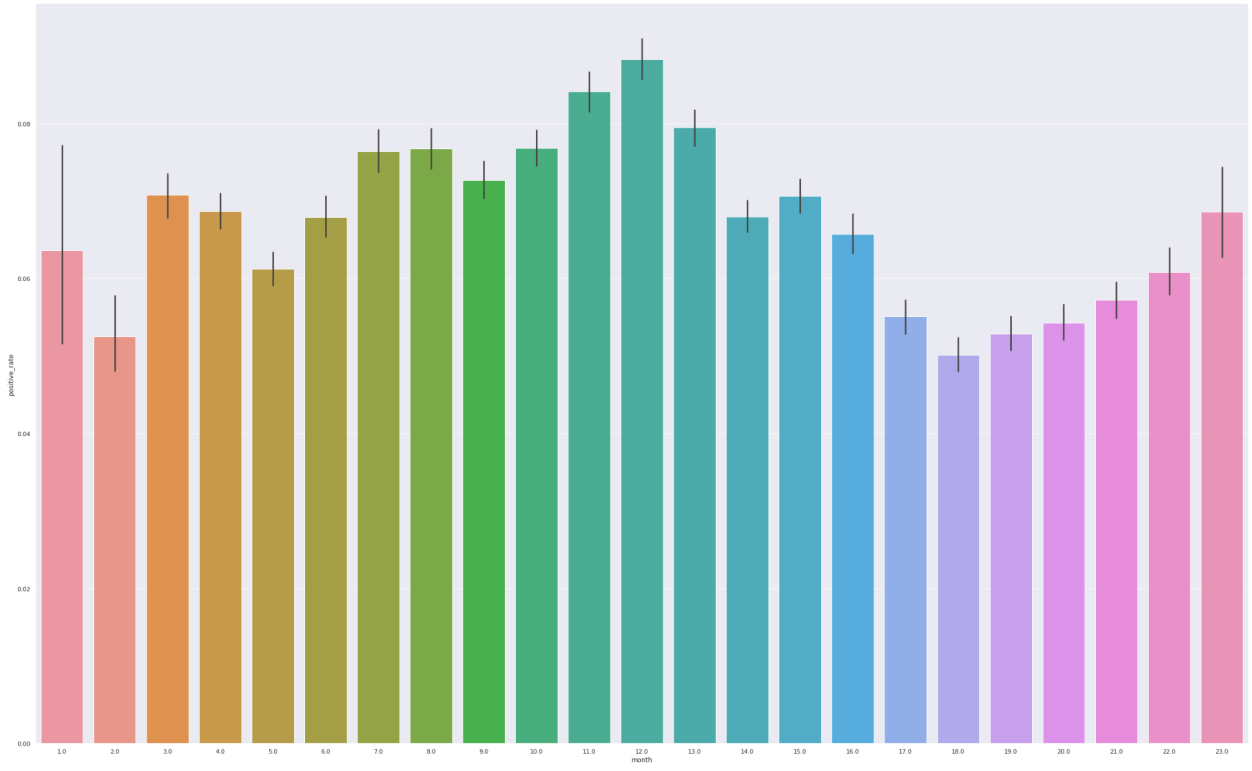
نمودار زیر نیز new_deaths_per_million در گذر ماه است که باز هم در اوایل سال 2021 میزان بالاتری دارد و در آخر رو به کاهش می رود



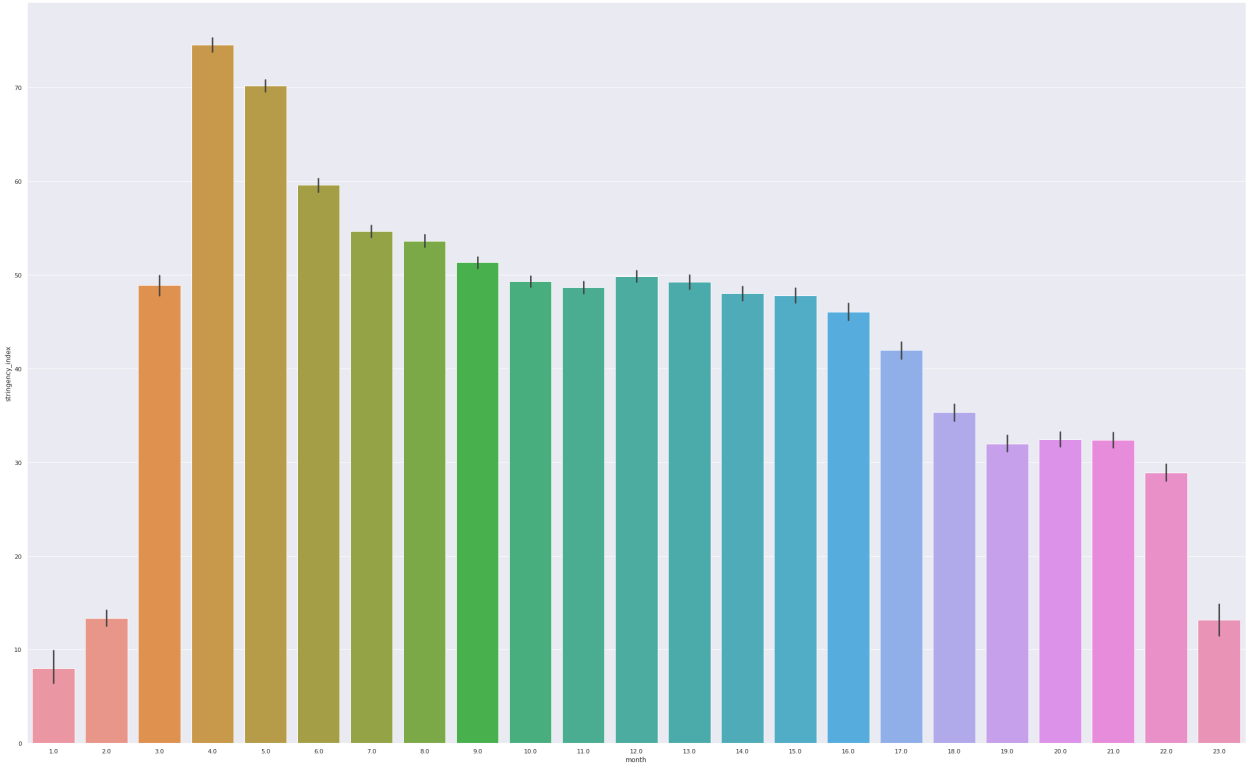
نمودار زیر people_fully_vaccinated را بر اساس ماه ها نشان می دهد که روند افزایشی دارد به غیر از ماه های آخر که باز هم به علت outlier بودن از جدول پاک شده است.



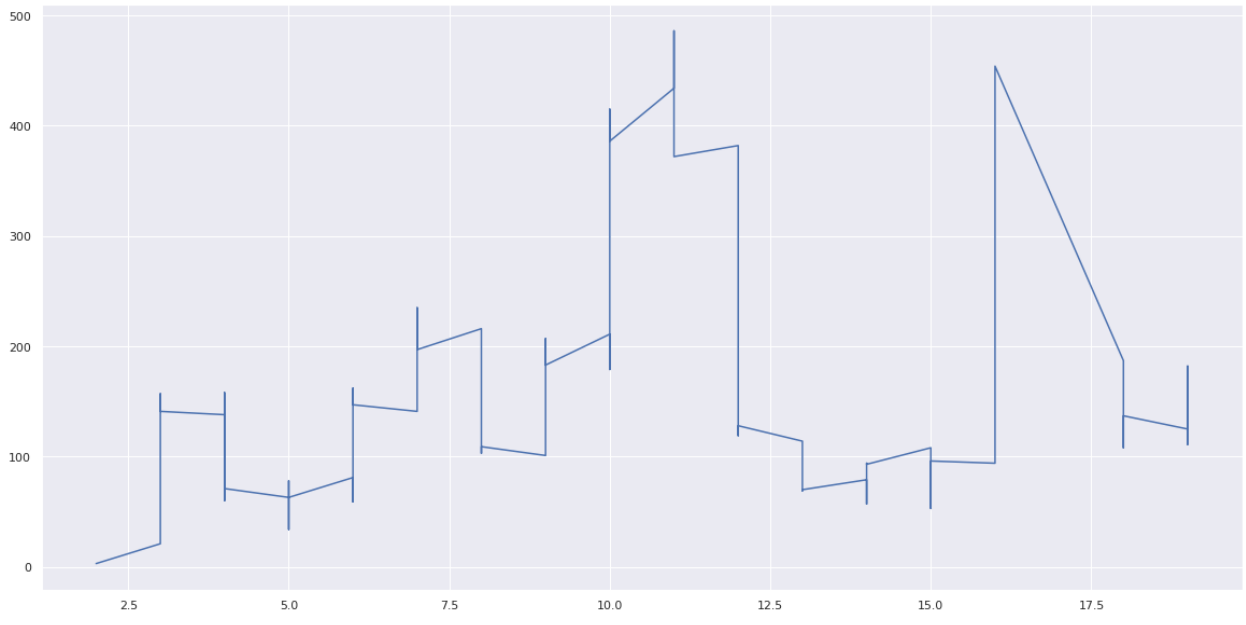
نمودار زیر نیز positive_rate را در گذر ماه ها نشان داده که تا اوایل سال 2021 افزایش داشته اما پس از آن کاهش داشته. البته در اواسط سال 2021 باری دیگر افزایش داشته است.



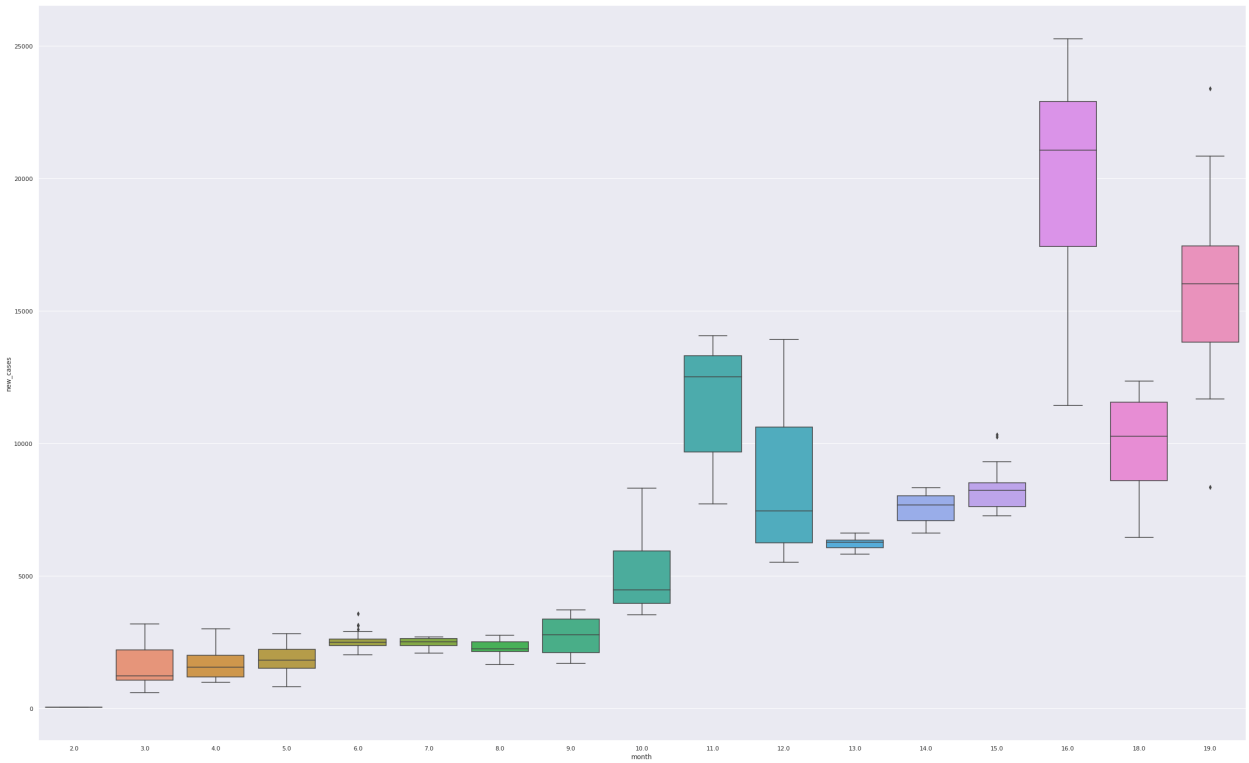
نمودار پایین stringency را طی ماه ها نمایش می دهد که پس از 2 ، 3 ماه از شیوع کرونا که این ویروس همه گیر تر شده بود ، سخت گیری بسیار بیشتر بود اما پس از مدتی این میزان رو به کاهش داشته است.



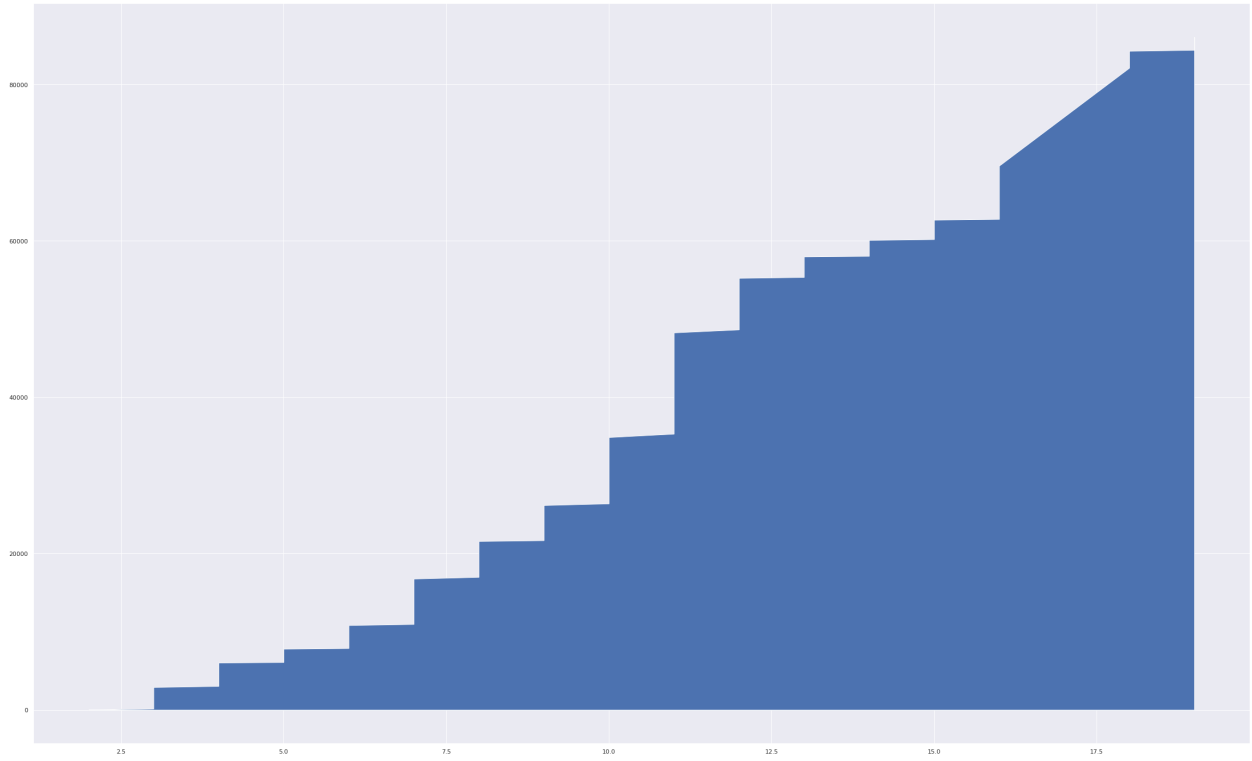
حالا تصمیم گرفتم ، چندتا از این موارد را بر روی داده های کشور ایران امتحان کنم تا بررسی کنم طی ماه ها ، چه اتفاقاتی در ایران افتاده است. نمودار زیر نشان دهنده روند new deaths در طول ماه ها در ایران است که به نظر در ماه های اواخر 2021 و اواسط ان میزان بیشتری داشته است ولی نکته آن است که به نظر داده های ماه های آخر ایران در هنگام پاک کردن outliers پاک شده است.



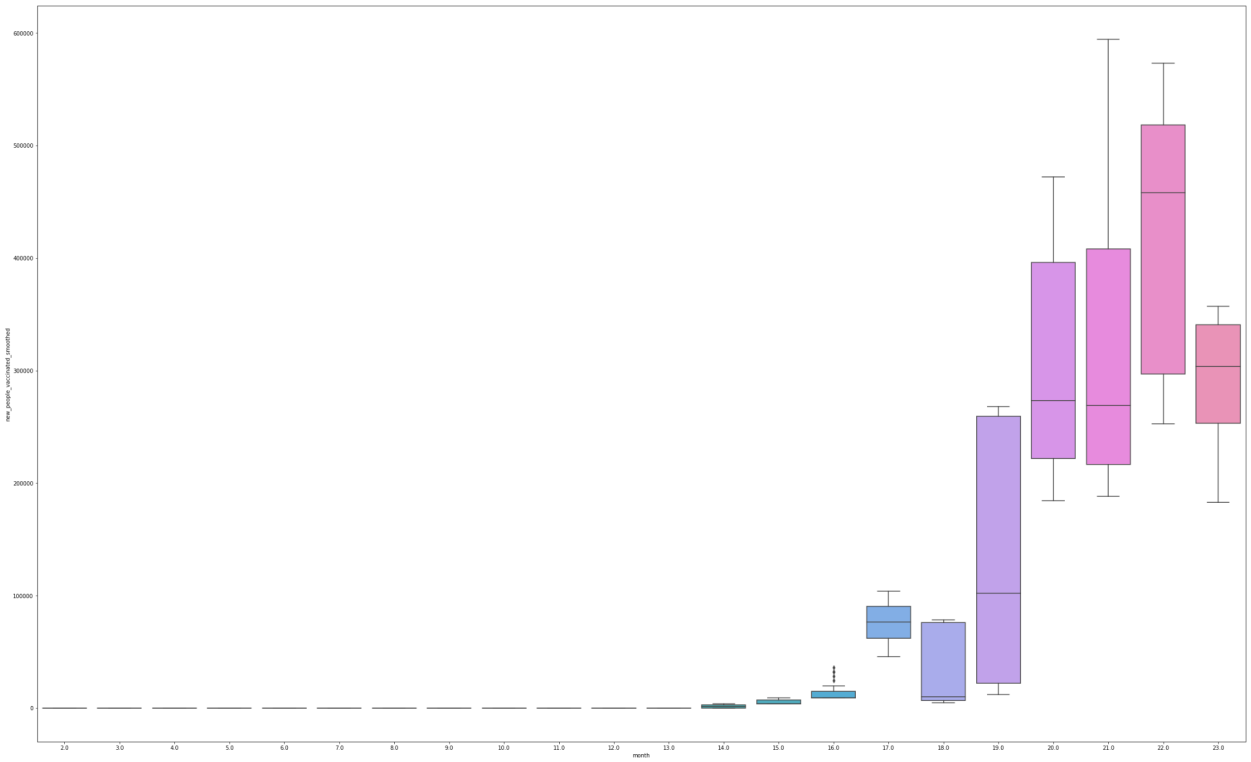
نمودار زیر box_plot از new_cases های ایران در هر ماه را نشان میدهد. بیشترین میزان بیماری نیز در که همانطور که مشخص است در اواخر بوده است.

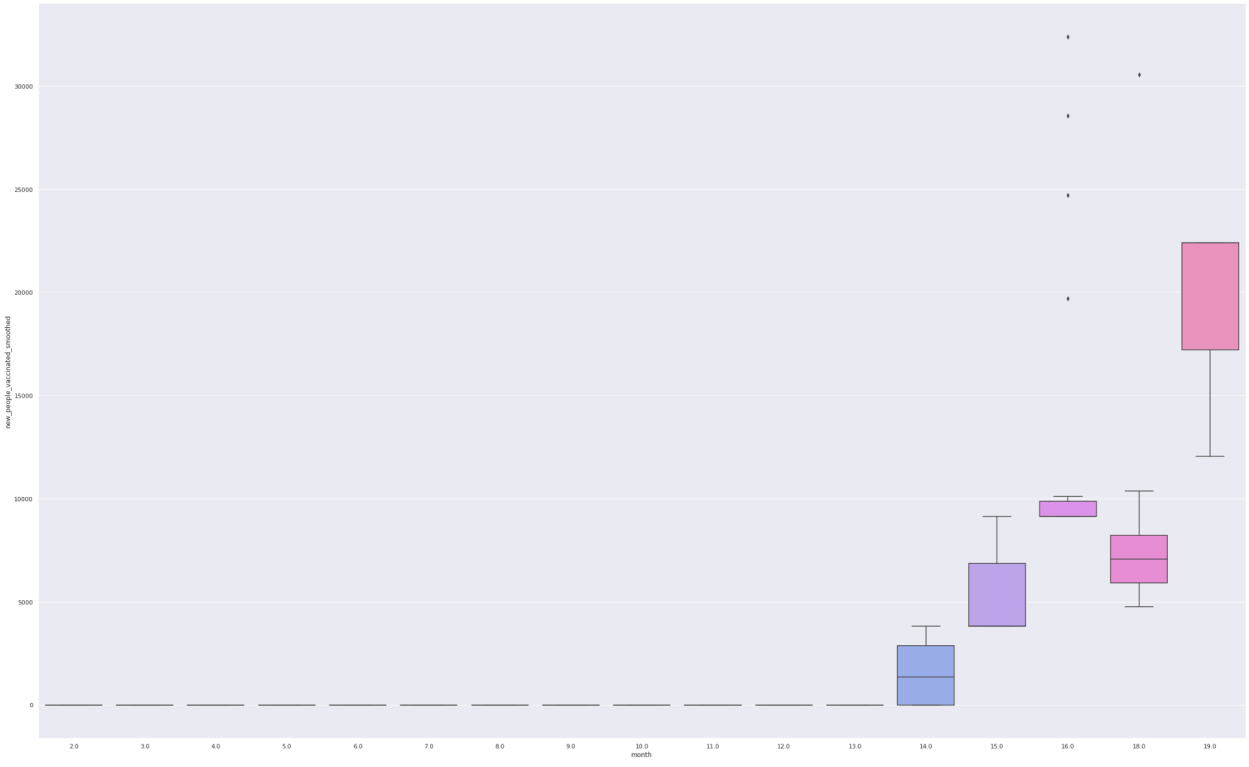


نمودار زیر total_death در گذر زمان است که مشخصا رو به صعود است.

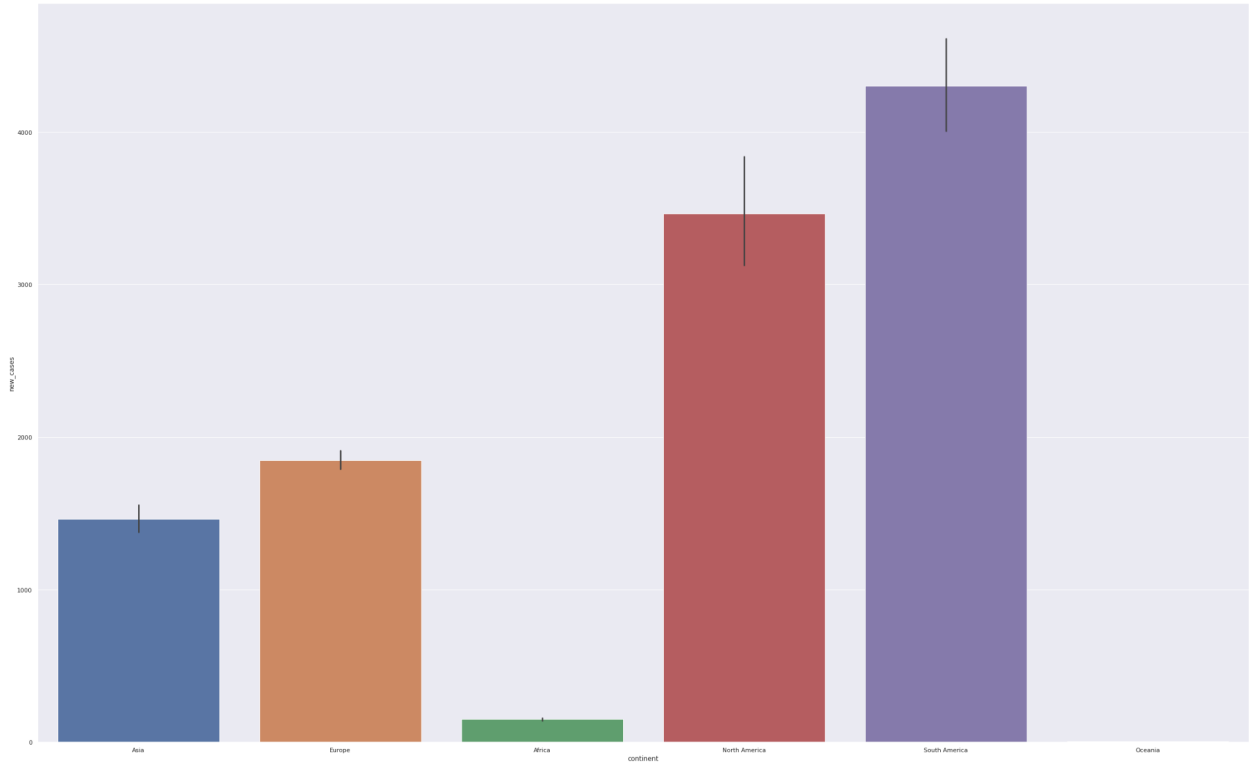


دو نمودار زیر به خوبی نشان دهنده تاثیر پاک کردن outliers خواهند بود.نمودارها new_people_vaccinated_smoothed برحسب ماه هاست که نمودار اول مربوط به قبل از پاک کردن outliers می باشد که به خوبی نشان می دهد که واکسیناسیون در ماه های انتهایی رو به افزایش است و بیشترین میزان آن در ماه مهر بوده است ولی نمودار دوم به کلی چندین ماه را حذف کرده و طبق چیزی که نشان می دهد واکسیناسیون حتی در ماه های پایانی بسیار کند پیش می رود که این موضوع صحیح نیست.

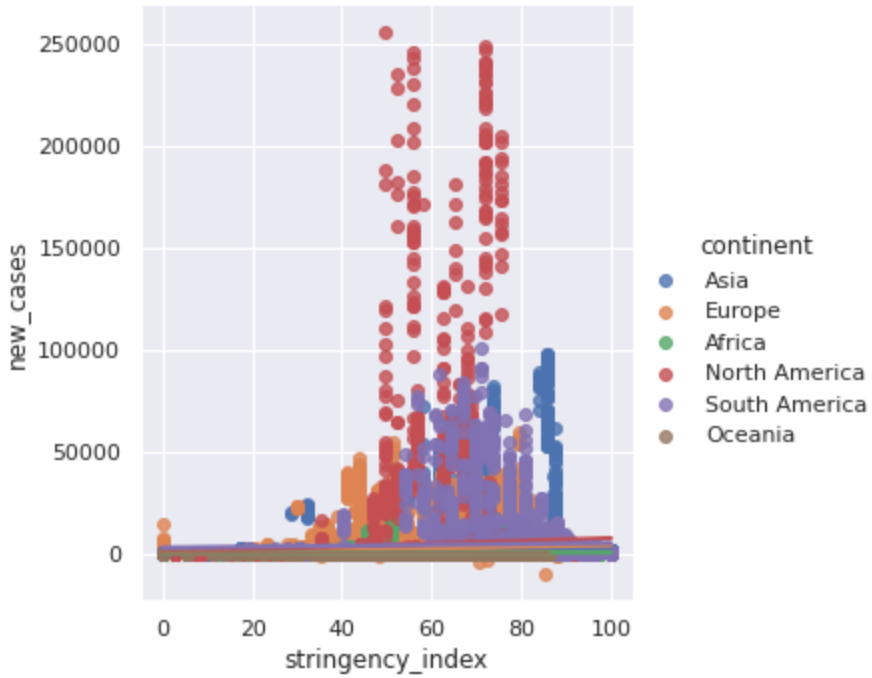




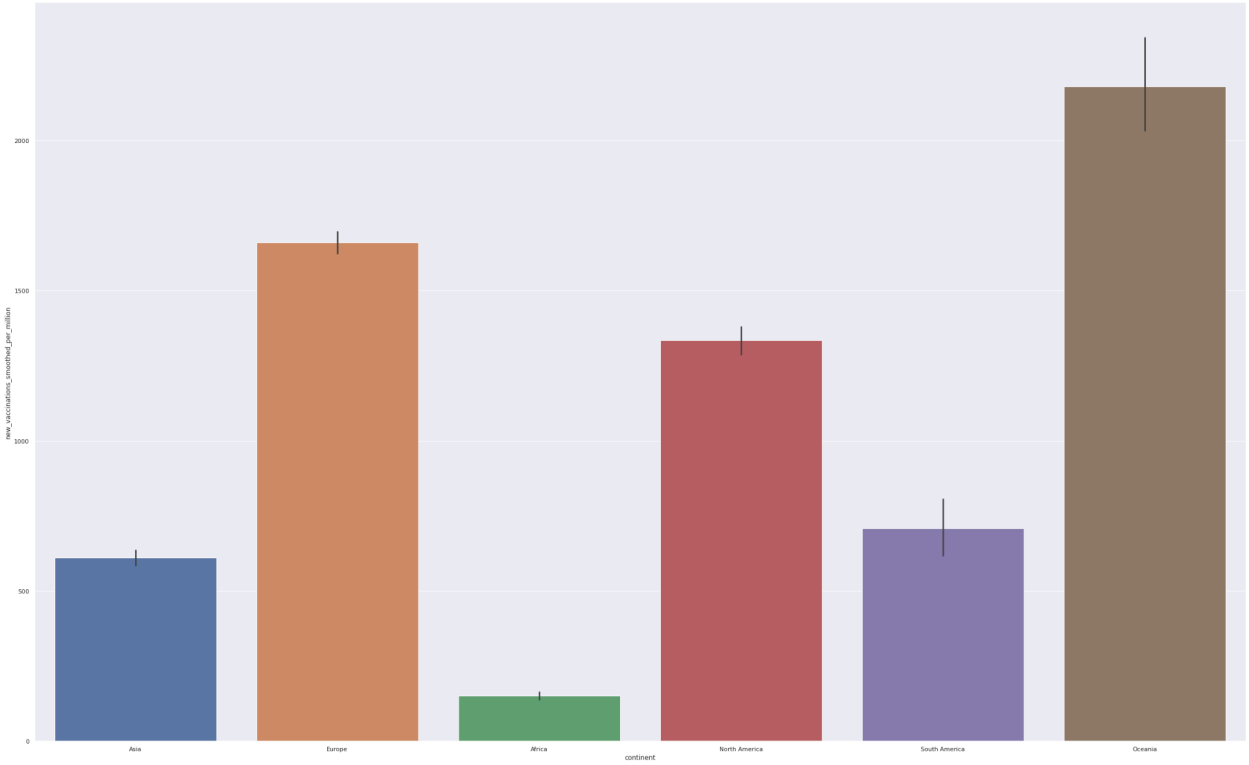
پس از آن به بررسی قاره های مختلف می پردازیم. نمودار زیر new_cases را در قاره های مختلف نشان می دهد که این مورد در قاره آمریکا شمالی و جنوبی میزان بیشتری را داراست. به نظر داده های اقیانوسیه آنقدر کم هستند که انگار وجود ندارند.



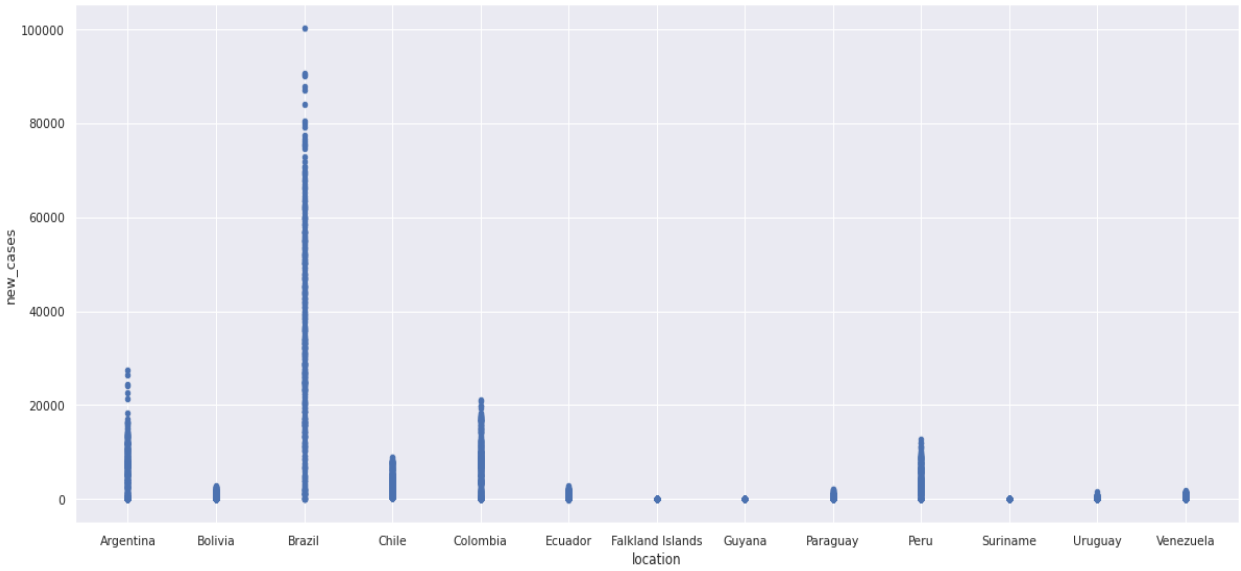
نمودار زیر stringency را با new_cases را برای هر قاره بررسی کرده است و سعی کرده خطی را با مدل رگرسیون بر آن فیت کند که خط صافی است و نشان می دهد این مقادیر را نمی توان به این شکل رابطه شان را بررسی کرد.



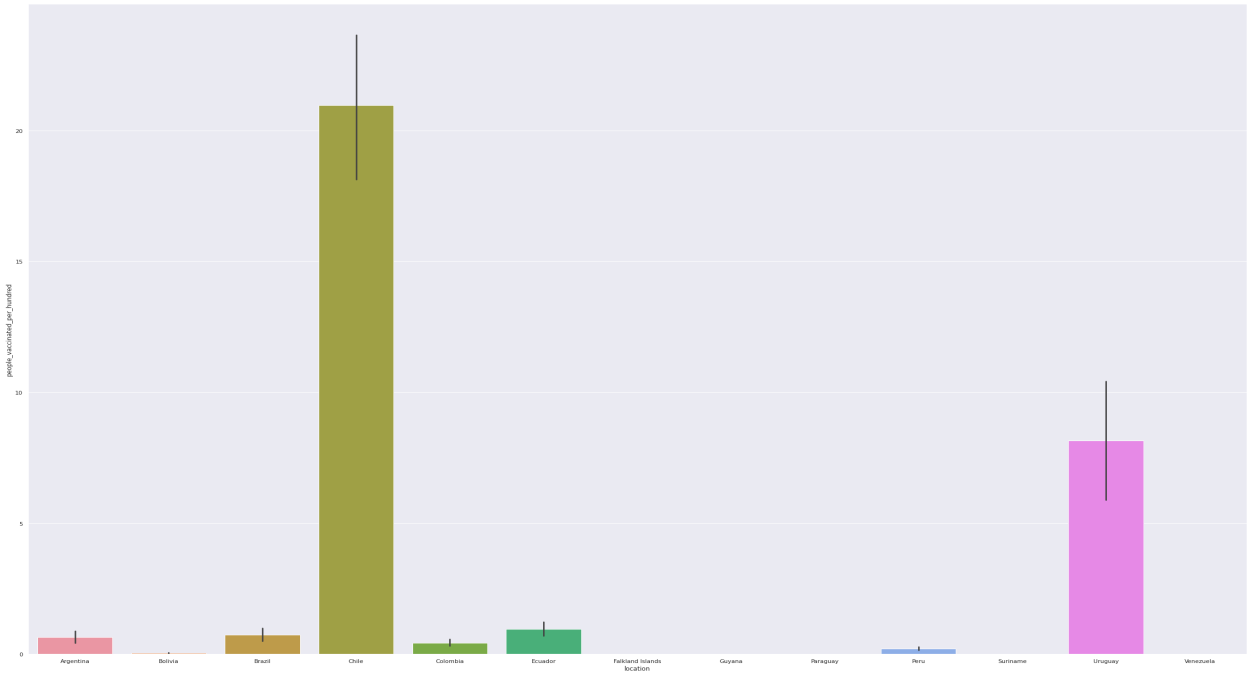
نمودار پایین new_vaccinated_smoothed در قاره های مختلف است که مشخصا در اقیانوسیه بیشترین میزان و در آفریقا به علت فقر کمترین میزان را داراست.



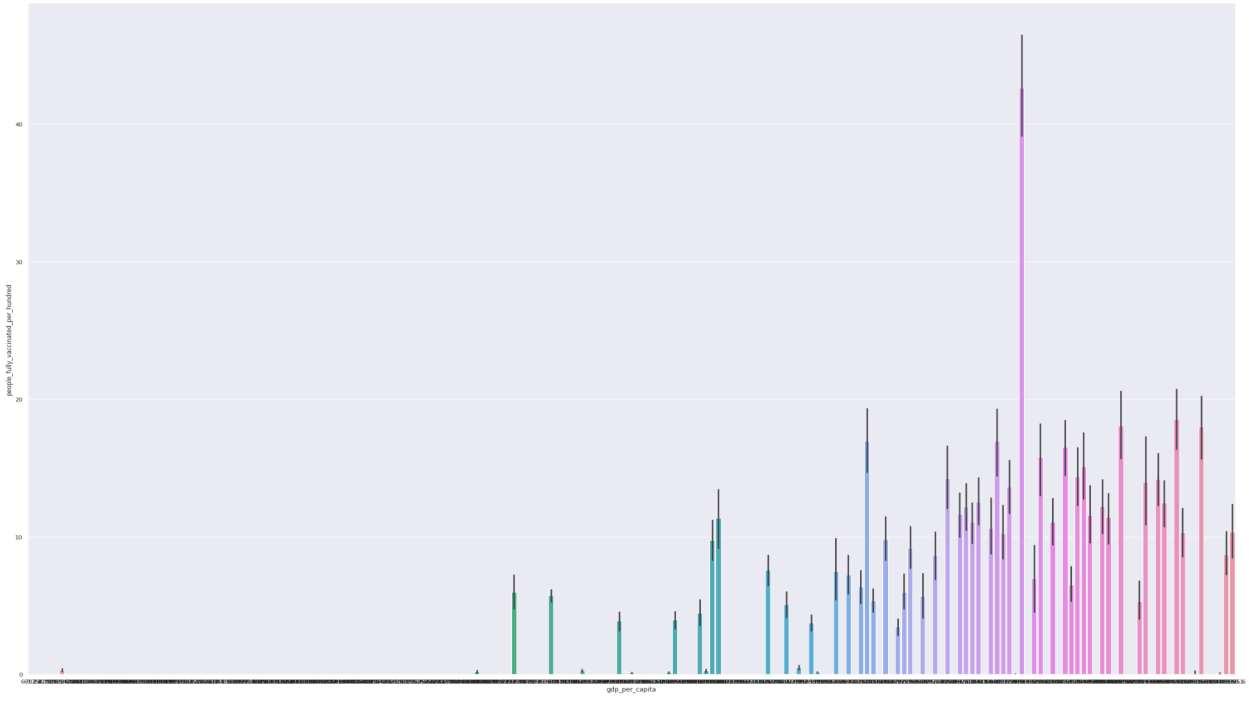
سپس خواستم کشورهای یک قاره را به صورت خاص مورد بررسی قرار دهم که کشورهای اروپا زیاد بودند از این رو این کار را روی آمریکای جنوبی انجام دادم که نمودار زیر new-cases را در کشورهای مختلف نشان می دهد که مشخصا در برزیل این آمار با اختلاف زیادی از کشورهای دیگر بیشتر است.



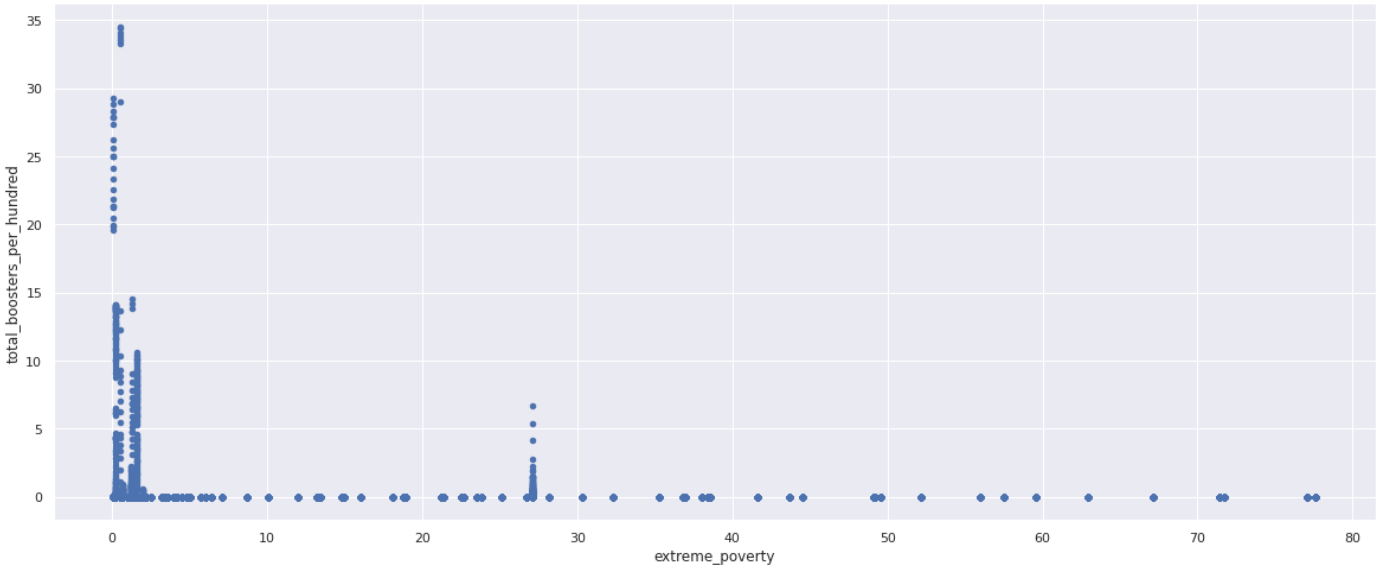
نمودار زیر را نیز new_vaccinated_per_hundred در کشورهای مختلف آمریکای جنوبی است که اینطور به نظر می آید که شیلی با اختلاف درصد بیشتری از مردمش را واکسینه کرده است.



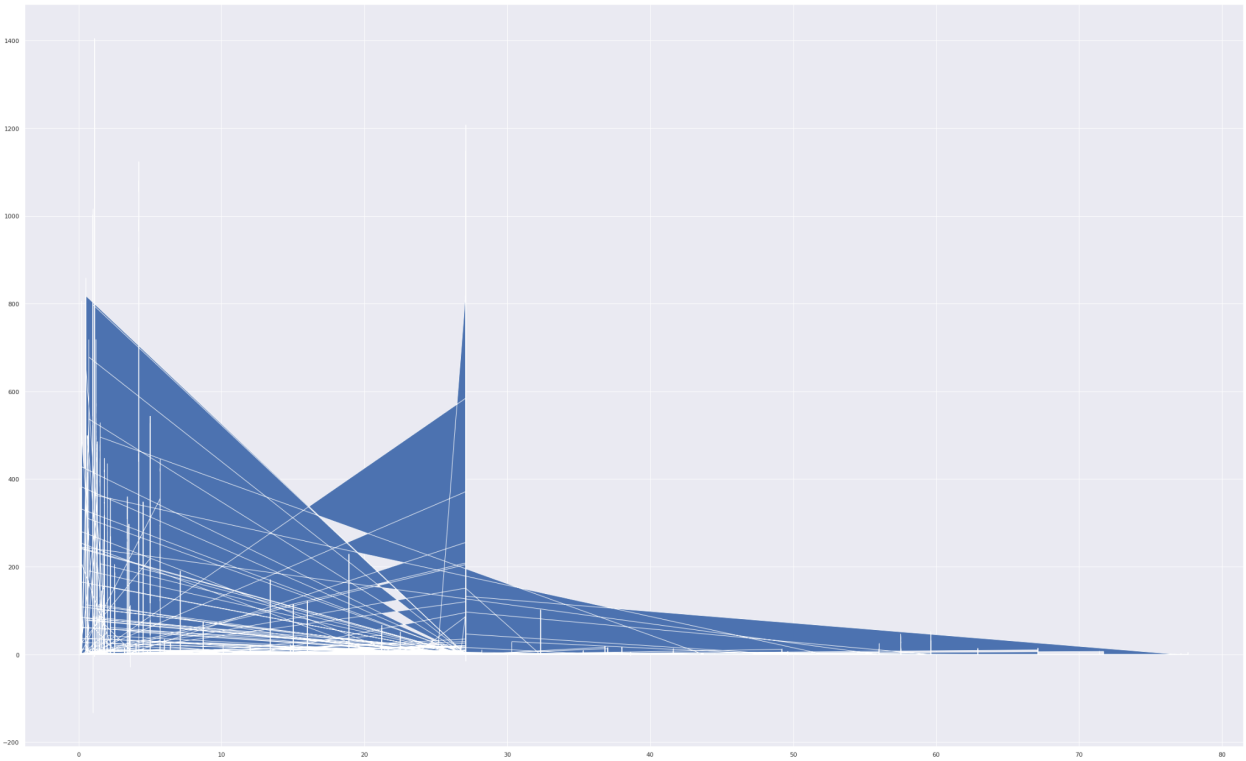
نمودارهای بیشتری در کد وجود دارد اما به سراغ بخش بعدی می رویم که بررسی gdp_per_capita و poverty می باشد. در ابتدا بررسی می کنیم که این موارد چه تاثیری بر واکسیناسیون داشته اند. نمودار زیر people-fully_vaccinated را بر اساس gdp_per_capita نشان میدهد. تقریبا هر چه قدر این میزان بالاتر باشد مردم بیشتری به طور کامل واکسینه شده اند که این به نظر به توان تولیدی و اقتصادی آن کشور بر میگردد .



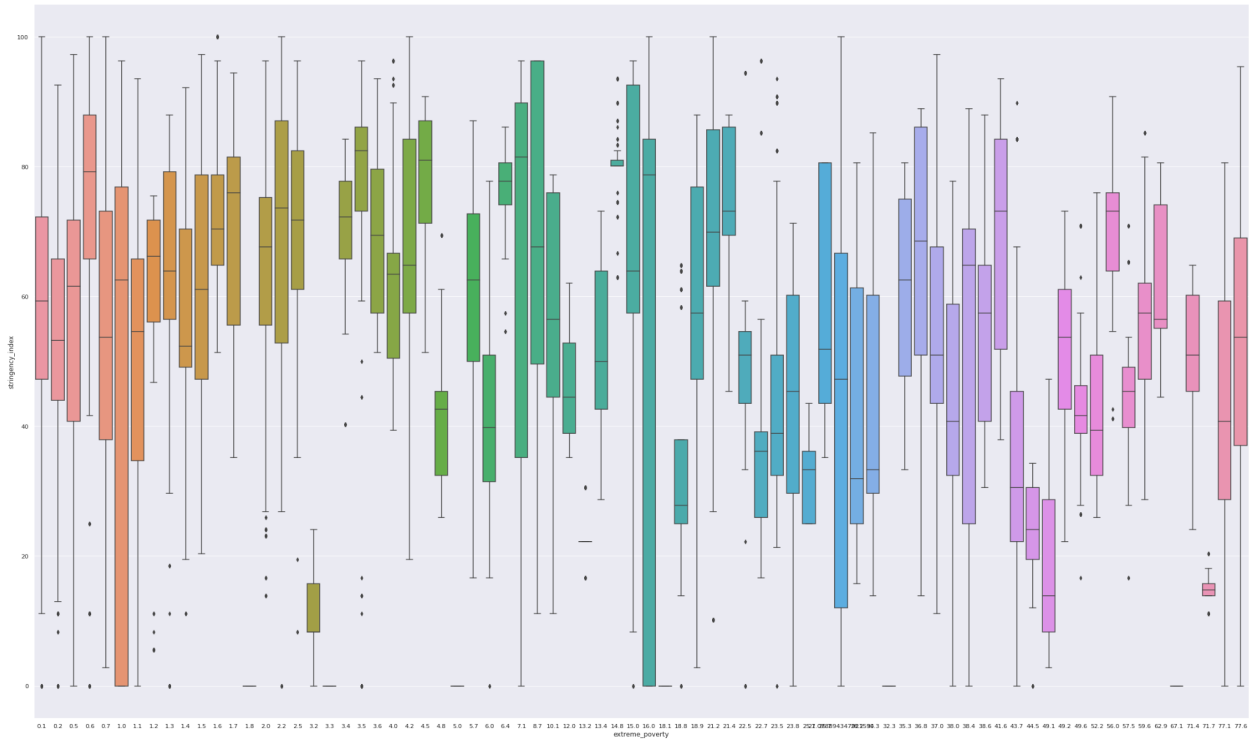
در نمودار زیر نیز که total_boosters_per_hundred را بر اساس extreme_poverty نمایش داده است ، تنها کشورهایی توانسته اند boosters استفاده کنند که فقر پایین تری دارند و قاعدتا توان اقتصادی بیشتری دارند.



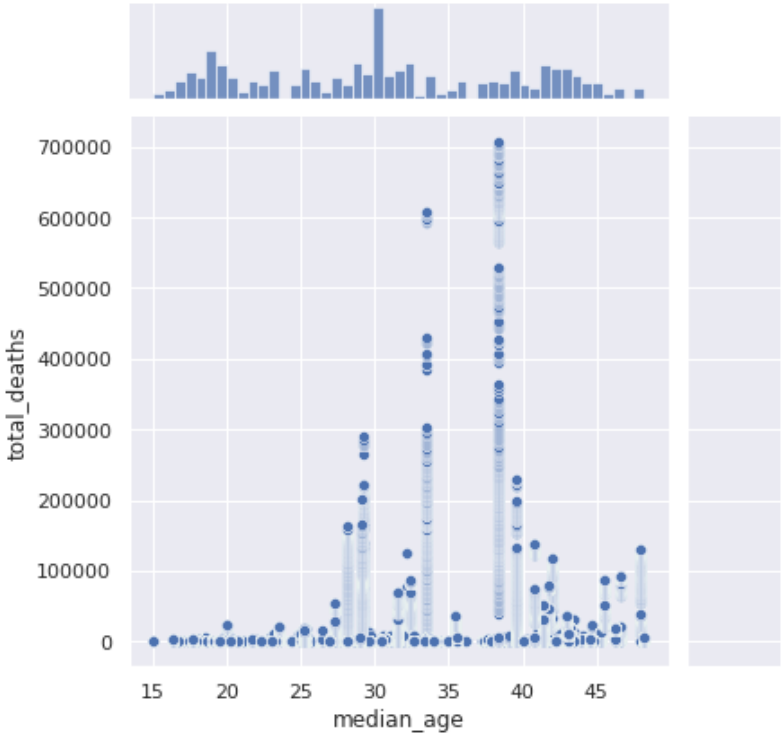
نمودار زیر new_cases_smoothed_per_million بر حسب extreme_poverty است که نشان می دهد بیشتر کشورهایی درگیر کرونا که وضع اقتصادی بهتری دارند. گرچه به نظر من یکی از این دلایل این است که کشورهای فقیرتر آمار درست و تست درستی انجام نمی دهند.



در آخر خواستیم ببینیم که آیا stringency با فقر و یا تولید رابطه ای دارد که با نمودارهایی که در کد وجود دارد مشخص است که نمی توان درد این مورد اظهار نظری کرد. نمودار زیر که stringency را بر حسب poverty نشان میدهد نمایانگر همین مسئله است.



در بخش بعدی median-age را بررسی می کنیم تا ببینیم سن افراد یک جامعه بر میزان تاثیر کرونا بر آنها اثر دارد یا خیر. نمودار زیر که total-death را بر اساس median-age نشان می دهد نمایانگر ان است که جوامعی که میانه سن پایین تری دارند میزان مرگ و میر کمتری داشتند. بیشتر میزان مرگ بین سنین 35 تا 40 بوده است که در نمودار زیر مشاهده می کنیم.



نمودار های بیشتری در کد وجود دارد که به همین یک مورد قناعت میکنم و به سراغ بخش بعدی می روم. در این بخش می خواهیم correlation نمودارهای مختلف را با heatmap بررسی کنیم. اولین نمودار که به شکل زیر است که همانطور که مشخص است بیشترین ارتباط را ستون های total cases per million و total deaths per million دارند که مشخص است هرجایی که case بیماری بیشتری داشته باشد مرگ و میر بیشتری نیز دارد.

رابطه دیگری که بالاست بین life expectancy با gdp و hand washing facilities است که تا حدودی به کرونا بی ربط است ولی هر چقدر قدرت تولید بیشتر باشد امید به زندگی نیز بیشتر خواهد بود و همچنین کشورهایی که امید به زندگی بالاتری دارند بیشتر handwashing را رعایت می کنند.

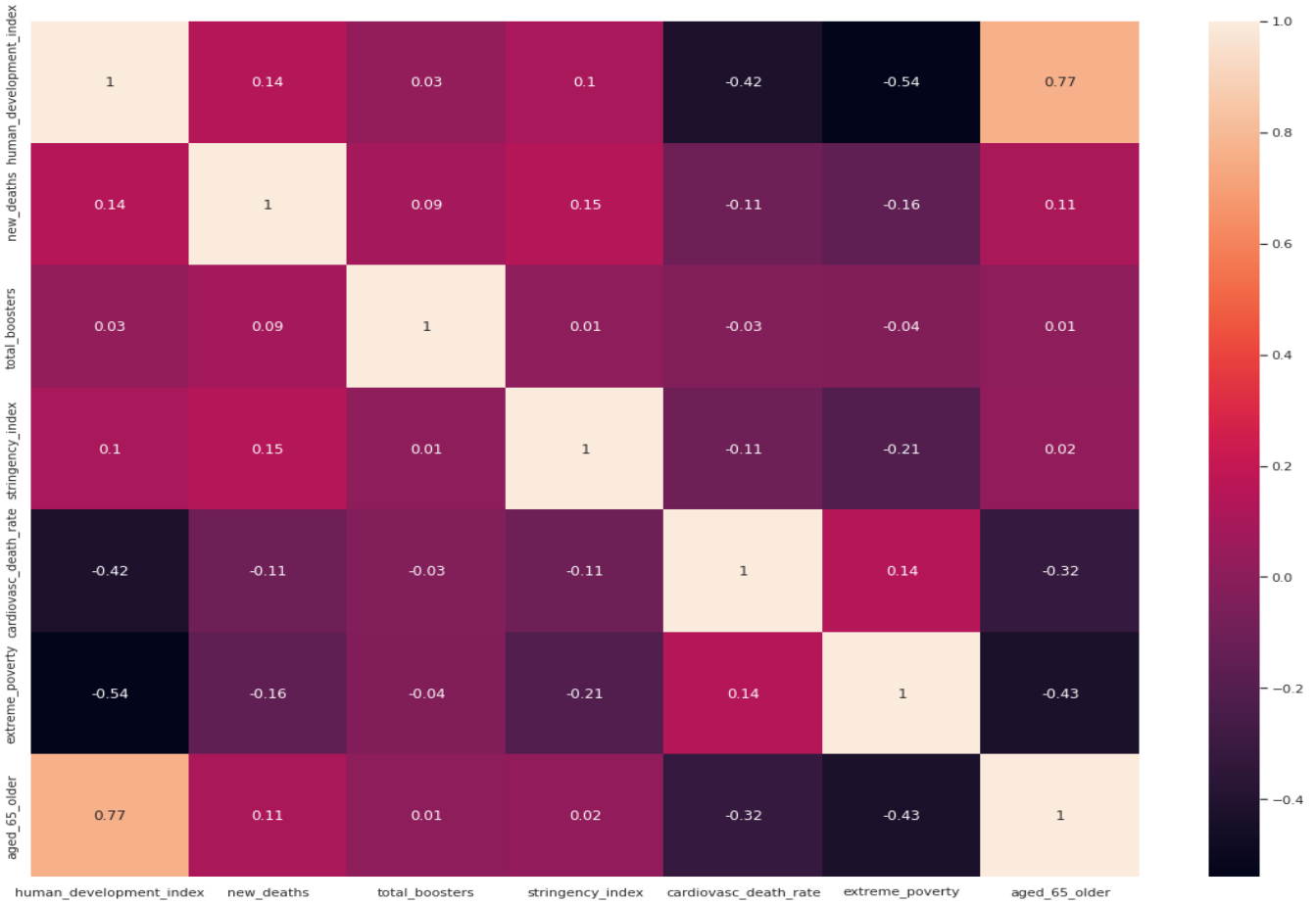
اما اخرین رابطه ای که به چشم می آید آن است که life expectancy با death و case ها رابطه مستقیم دارد و در جوامعی که امید به زندگی بیشتر بوده کرونا بیشتر اختلال ایجاد کرده است که این مورد را تا حدودی در نمودار های قبلی بررسی کردیم که کشورهای پیشرفته تر آمار مرگ و میر بیشتری داشتند.

یک heatmap دیگر از روابط ستون ها را در پایین با هم بررسی میکنیم.

اولین نکته ای که به چشمم آمد آن است که کشورهایی که افراد بالای 65 سال بیشتری دارند فقر و بیماری قلبی عروقی کمتری نیز دارند و همچنین human_development_index بالاتری دارند که فکر میکنم مربوط به کشورهای پیشرفته تر باشد و در این صورت این امار تا حدی قابل تصور است.

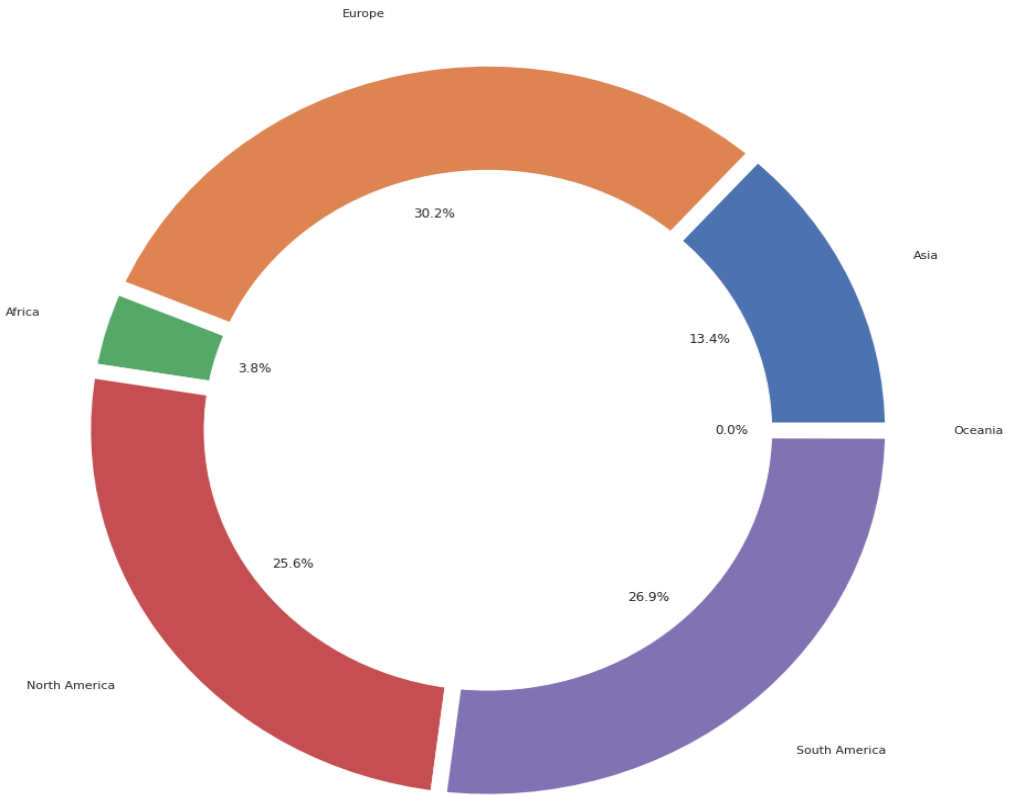
حالا چند مورد از عوامل موثر بر آمار کرونا را بررسی میکنیم. کشورهایی که آمار مرگ و میر بیشتری داشتند stringency بالاتری نیز داشتند و سختگیری بیشتری داشتند که قابل درک است. همچنین کشورهای فقیر تر stringency کمتری داشتند که به علت عدم اگاهیشان است.

بقیه مواردی که مقادیر بالاتری دارند به آمار کرونا بی ربط است که در زیر مشاهده می کنیم.

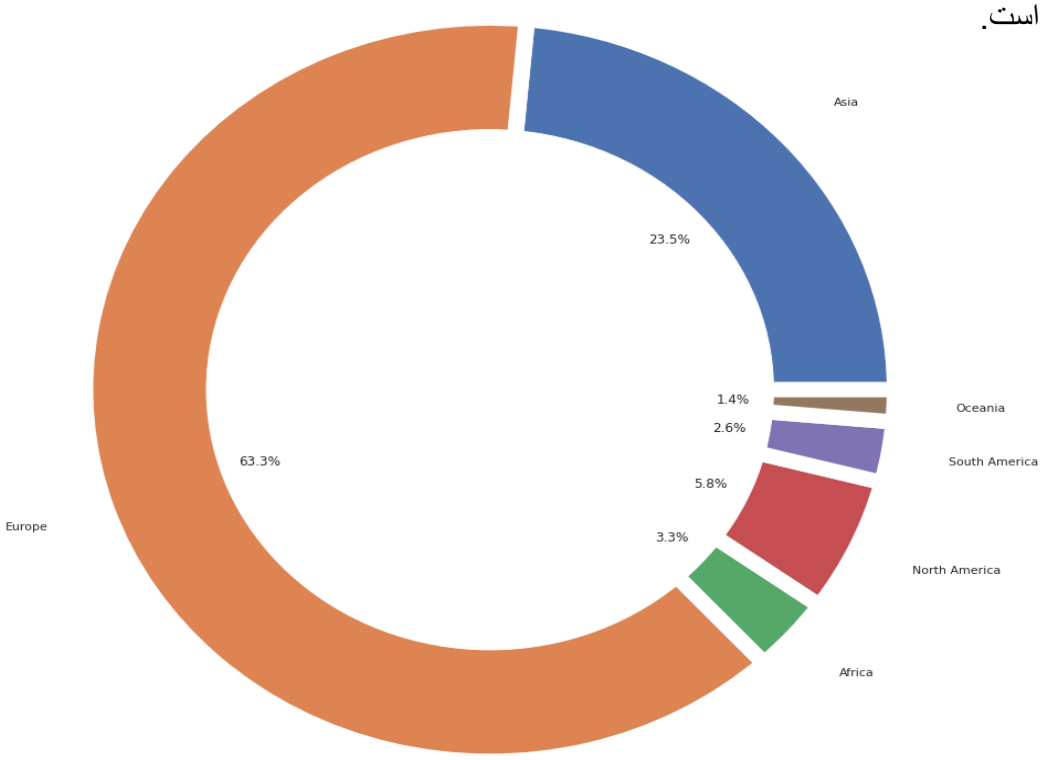


در بخش بعدی برای اینکه نمودار هایمان تا حدی متفاوت تر باشد تلاش کردیم تا نمودار دونات را برای قاره ها پیاده سازی کنیم. نمودار اول که total deaths را در قاره های مختلف نشان می دهد نشان دهنده آن است که مرگ و میر در اروپا از همه قاره ها بیشتر بوده است. همچنین مرگ و میر در اقیانوسیه آنقدر درصد پایینی را به خود اختصاص داده است که عدد 0 را برای آن مشاهده میکنیم.

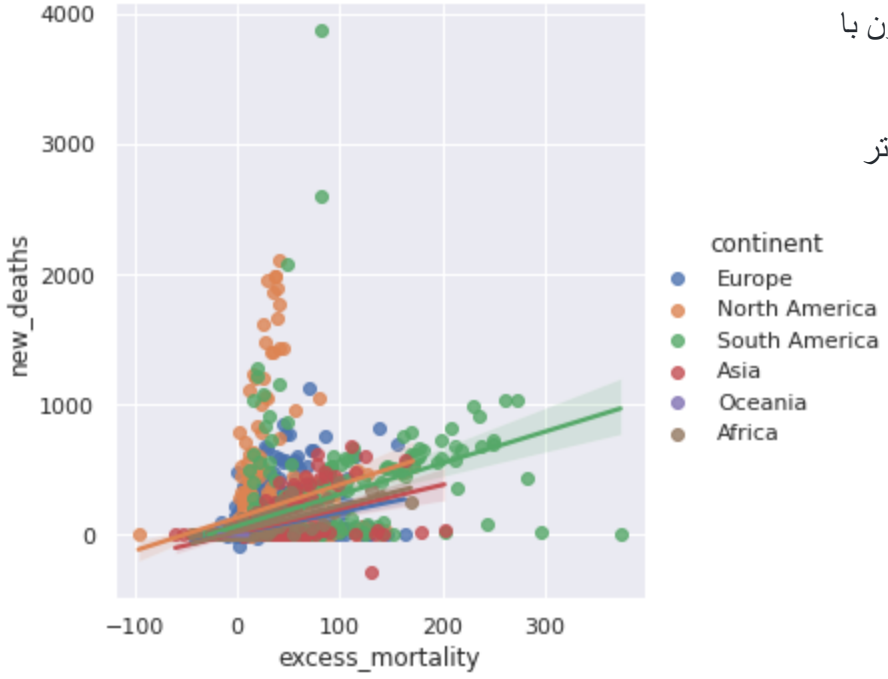
ذکر این نکته لازم است که برای به دست آوردن مجموع مرگ ها اخرین total death هر کشور در هر قاره مختلف را با هم جمع کردیم و مقایسه کردیم.



نمودار زیر total_tests_per_thousand در قاره های مختلف است که این مقدار در اروپا به شکل قابل توجه ای بالاتر است. زیرا در اروپا تست های بیشتری گرفته می شود. آسیا نیز 23 درصد از تست ها را به خود اختصاص داده است.



در آخرین بخش خواستیم تاثیرات ستون های excess_mortality که از جدول حذف شده بودند را بررسی کنیم. برای این کار دیتا فریم جدیدی تنها با داده هایی که excess_mortality شان null نیست درست کرده ایم و پس از آن چند نمودار در ارتباط این ستون با بقیه ستونها میبینیم که نمودار زیر نمایانگر آن است که هر چقدر excess_mortality بالاتر بوده new death هم مقدار بالاتری داشته است زیرا اگر به خطی که نمودار زیر برای هر قاره سعی دارد تا فیت کند توجه کنیم می بینیم که شیب مثبت و صعودی دارد.



اما در آخر heatmap آنها را مشاهده میکنیم که کاملا واضح است که ستونهای مربوط به excess_mortality همگی با ستون های مربوط به مرگ و میر رابطه مستقیم دارند. یعنی بیشتر شدن نرخ مرگ و میر احتمالا تا حدودی تحت تاثیر این ویروس قرار دارد. بقیه موارد را در جدول مشاهده میکنیم.