به نام خدا

مهرانه مقتدائی فر ۹۷۲۲۲۰۸۶

تاریخ: ۱۴۰۰/۸/۲۸

گزارش تمرین سری ۲ **درس: مبانی علوم داده** **استاد درس: دکتر خردپیشه**

در این تمرین به بررسی داده های مربوط به ویروس covid پرداختیم. این دیتا شامل تعداد زیادی داده های NaN بود پس در مرحله اول لازم است تا به پاکسازی داده بپردازیم:

CLEANING DATA :

برای شروع با توجه به info دیتاست، ابتدا ستون هایی که بیش از نصف آنها و مقدار زیادی NaN داشتند را حذف میکنیم. سپس به پر کردن داده های categorical پرداختیم. در اینجا دادهی continent را بررسی کردیم. این داده شامل قاره های Asia, Europe, Africa, Oceania و North and south America است. این قاره ها را با توجه به ستون location == "this continents" پر کردیم همچنین یک LOCATION وجود دارد که برابر با world است، قاره این قسمت را برابر با ALL قرار دادیم که به معنی کل قاره ها و کل جهان است. پس از آن مشاهده کردیم که باز هم حدود ۳هزار داده هنوز Nan هستند. اندیس های این داده ها را پیدا کرده و سپس LOCATION آنها را پیدا کردیم و دیدیم که این location ها اسم کشور نیستند بلکه اطلاعات دیگری در مورد درآمد ها دارند، در اینجا همین سطر ها را از دیتاست حذف کردیم و دیگر با آنها کار نخواهیم کرد.

برای پر کردن داده های عددی متد های مختلفی را پی گرفتیم. ابتدا با بررسی برخی ستون ها متوجه شدیم که جاهایی که total case آنها بسیار کم است(یا کم گزارش شده) یا total death آنها نیز به همین شکل است ، مقادیر new case و new deaths و از این قبلی نیز گزارش نشده و NAN است. پس مقادیر نال آنها را برابر با 0 قرار دادیم.

در مرحله بعدی با بررسی برخی ستون ها مانند total_cases, total_deaths, population, make_smokers, life_expectancyو ... مشاهده میکنیم که برای هر location به طور خاص این مقادیر تقریبا برابر هستن و حتی اگر تغییری داشته باشند(به خصوص داده های مربوط به مرگ و میر و کیس های جدید) خیلی کم است. به همین علت این گونه داده ها را با متد bfill پر کردیم و بدین گونه عمل میکند که هر داده را با داده بعدی آن پر میکند و انقدر ادامه میدهد تا به یک داده غیر nan برسد و سپس تمام قبلی ها را با آن پر میکند.

ستون های دیگری نیز مانند stringency_index, median_age, gpd و ... نیز همینگونه هستند، چون داده های سطر های آخر این مقادیر پر بود به جای متد bfill از متد pad استفاده کردیم که برعکس قبلیست و هر داده را با داده‌ی پیشین پر خواهد کرد.

پس تعداد زیادی از داده های نال ما پر شدند حال تنها داده های مربوط به test و vaccinate باقی مانده‌اند. اینگونه داده ها شاید به طور یکنواخت افزایش نیابند(مخصوصا test) اینگونه داده ها را با استفاده از روش Interpolation پر کردیم. این روش برای آن است که داده های یک ستون با توجه به مابقی داده های در دسترس پر شوند. ساده ترین متد آن یعنی linear استفاده شده که به صورت خطی با توجه به داده های قبل و بعد نوعی درونیابی انجام میدهد تا متوجه شود که بهتر است داده های پوچ را با چه چیزی پر کند.

بدین صورت تمامی داده های ما پر شدند و دیگر داده null نداریم. حال وقت آن است که ببینیم داده های پرت ما به چه صورت هستند.

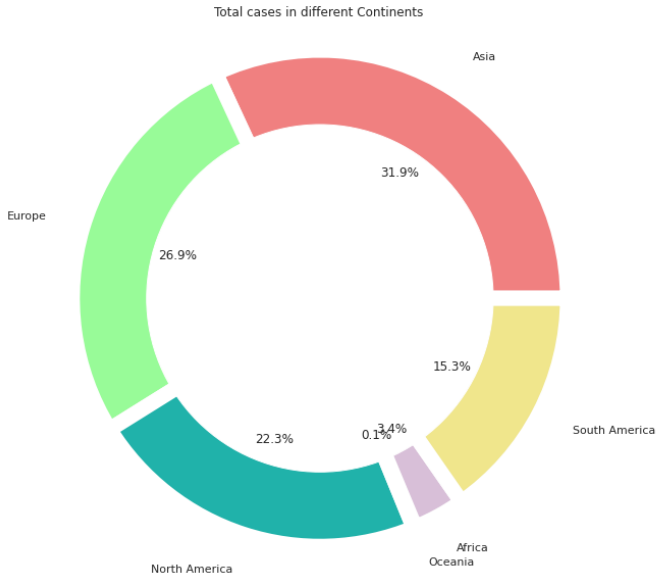
OUTLIERS :

در این بخش به حذف داده های پرت از دیتا ست مورد نظر پرداختیم. از آنجایی که پراکندگی داده های ما بسیار زیاد است نمیتوانیم بر اساس تمامی ستون ها این کار را انجام دهیم زیرا میانگین ستون های مختلف بسیار باهم فاصله دارند و همین باعث خواهد شد تا داده های زیادی به عنوان داده های پرت در نظر گرفته شوند. پس برای اینکار برای هر ستون از دیتاست، این عملیات را انجام میدهیم. بدینگونه که برای هرستون میانگین و std آن را در نظر گرفته و در بازه‌ی $mean \pm 4 * std$ داده ها را نگه داشتیم و هرچه خارج این بازه بود دور ریخته‌ایم.

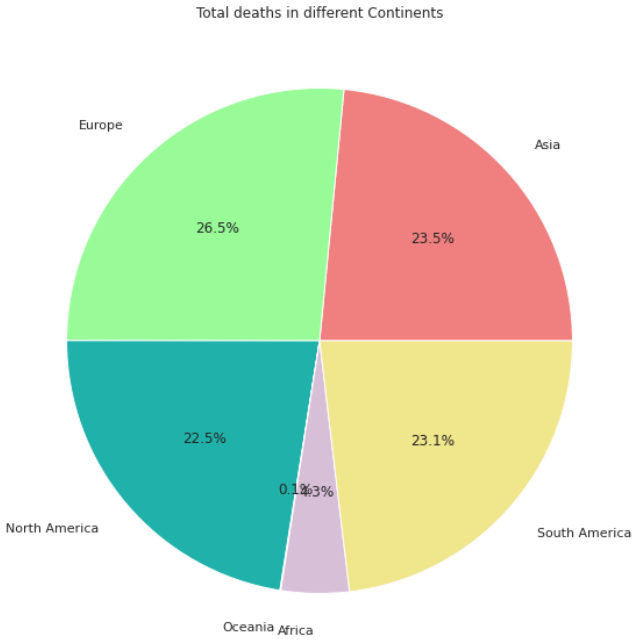
اینگونه حدود ۲۰۰۰ داده از دیتاست موردنظر حذف شد.

حال به بررسی ستون ها و مقاسیه آنها با یکدیگر و رسم نمودار های مختلف میپردازیم:

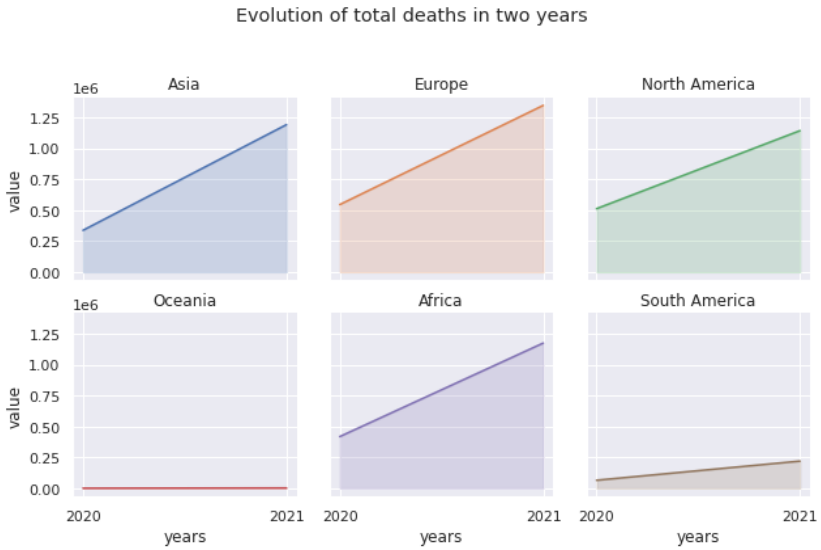
در ابتدا با توجه به location برای قاره های مختلف و همچنین کل جهان(world) دیتا فریم جداگانه ساخته‌ایم و سطر آخر total_cases که برابر است با تعداد کل مبتلایان قاره ها را در نظر گرفتیم. این مقدار را بر مقدار کل جهان تقسیم کردیم تا بتوانیم به صورت درصد، مقایسه کنیم. نمودار دونات زیر نشان میدهد که در قاره Asia از بقیه مکان ها بیشتر و در قاره Oceania از همه کمتر بوده است.



همچنین برای آمار مرگ و میر نیز همینکار را با نمودار Pie chart انجام دادیم. آمار مرگ و میر برای اکثر قاره ها به جز اقیانوسیه و آفریقا، تقریبا نزدیک به هم بوده است اما همانطور که مشخص است مرگ و میر اروپا از همه بیشتر بوده است.

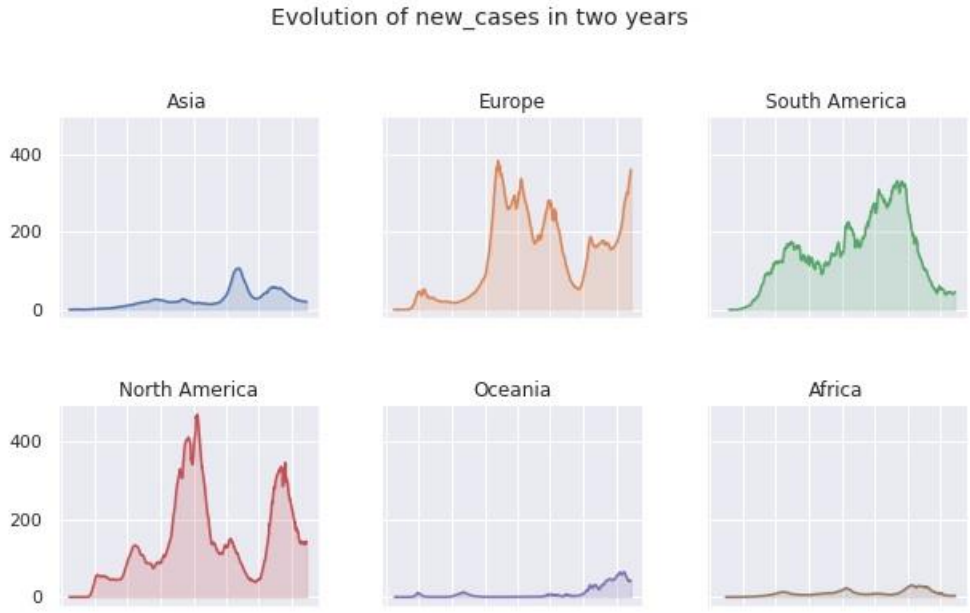


در ابتدای کد با توجه به ستون date یک ستون جدید year برای دیتاست خود درست کردیم که در این ستون دو سال ۲۰۲۰ و ۲۰۲۱ قرار دارد. حال با استفاده از این دو سال روند افزایش مرگ و میر در قاره های مختلف را باهم میبینیم.



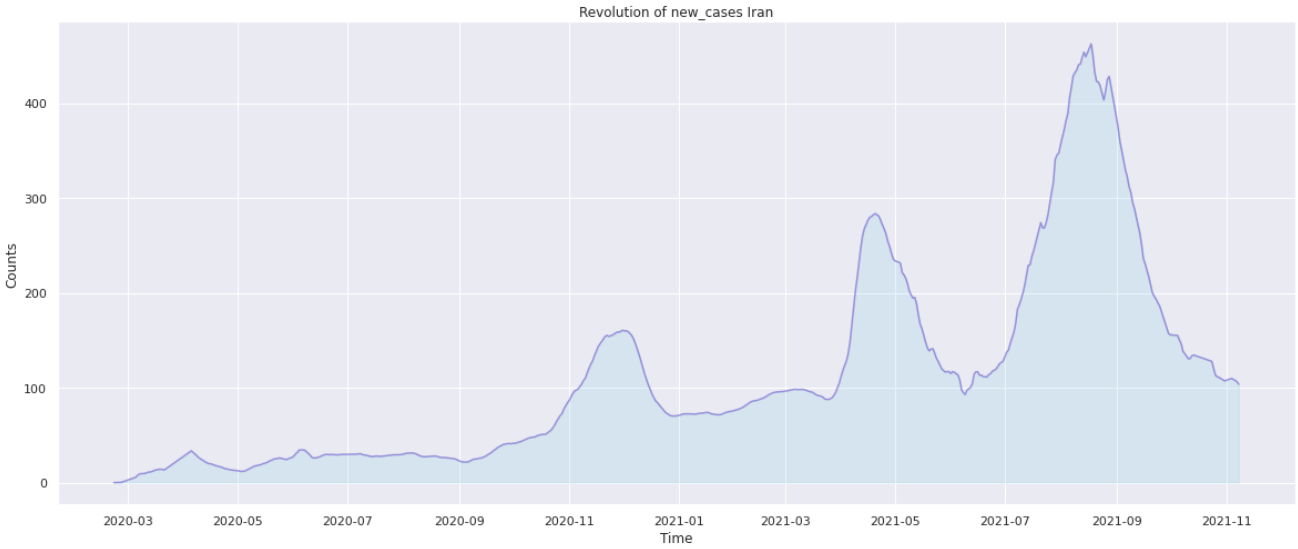
این نمودار روند صعودی مرگ و میر را نشان میدهد و همانطور که از نمودار پیشین نیز نتیجه گرفته بودیم این روند در اروپا از همه بیشتر بوده است و در اقیانوسیه تغییر بسزایی نداشته است.

حال نمودار بعدی که باهم میبینیم، بررسی روند افزایش(یا کاهش)میزان مبتلایان و new case ها هست.

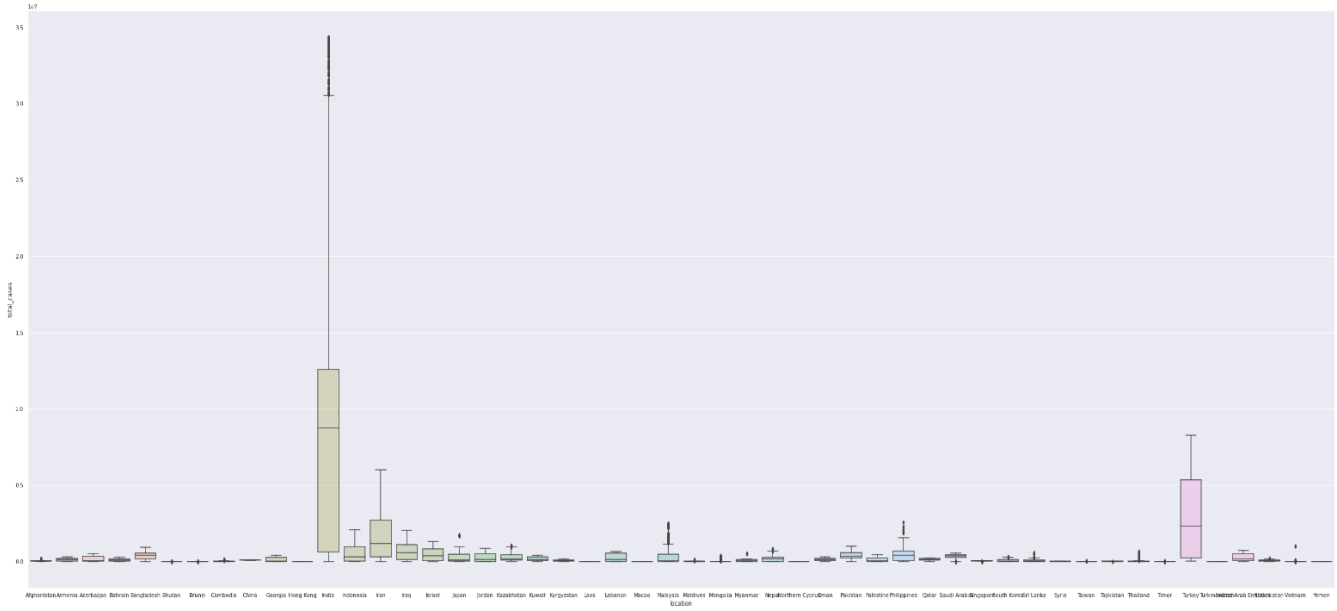


در این نمودار ها کاملا مشخص است که در چه زمانی تعداد مبتلایان جدید افزایش داشته و در چه زمانی کاهش یافته است.

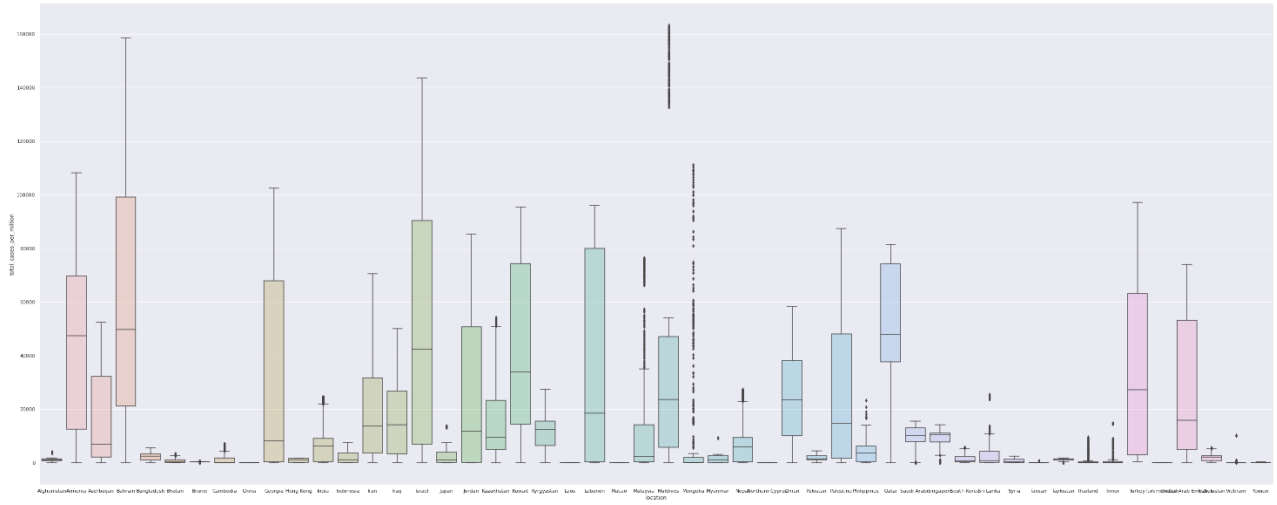
در شکل زیر هم به طور خاص این روند را در ایران مشاهده میکنیم:



در بخش بعدی به بررسی کشور هایی که به طور مخصوص در قاره آسیا قرار دارند را بررسی میکنیم. در بخش قبل دیدم که total cases در قاره آسیا از مابقی جاها بیشتر بود است حال ببینیم در کدام کشور:



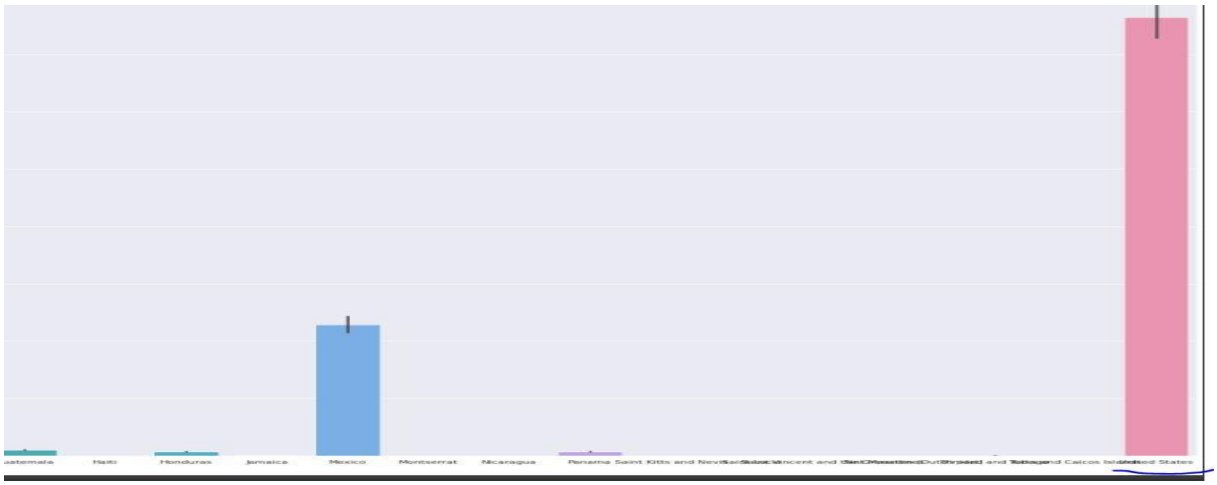
همانطور که در نمودار بالا مشاهده میکنیم، کشور هند با اختلاف زیادی از بقیه کشور ها قرار دارد و total case آن بیشتر بوده است. البته اینکه جمعیت هند نیز از بقیه کشور ها بسیار بیشتر است نیز بی تاثیر نیست به همین علت per_million آن را نیز بررسی کردیم:



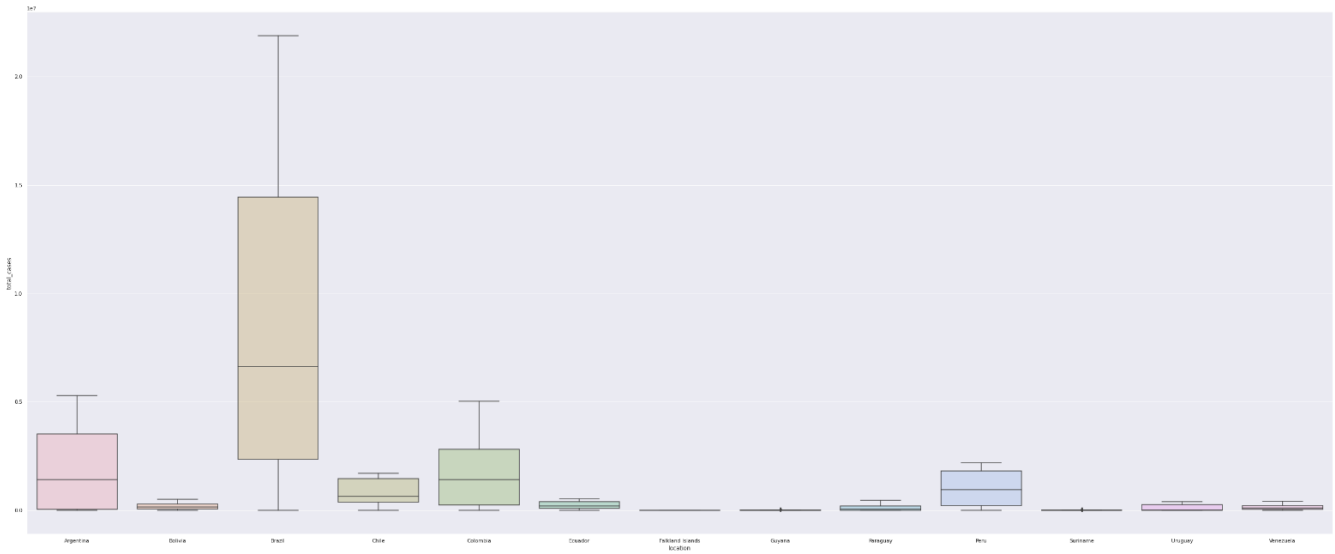
همانطور که میبینیم، در این نمودار آمار بحرین از مابقی بیشتر است(نسبت به جمعیت آن)

در اروپا نیز این آمار برای فرانسه بیشترین بوده است.

در نمودار بعدی نشان داده شده که در آمریکاری شمالی، ایالت متحده آمریکا آمار مرگ و میر و همچنین new_case آن از مابقی بیشتر بوده است.

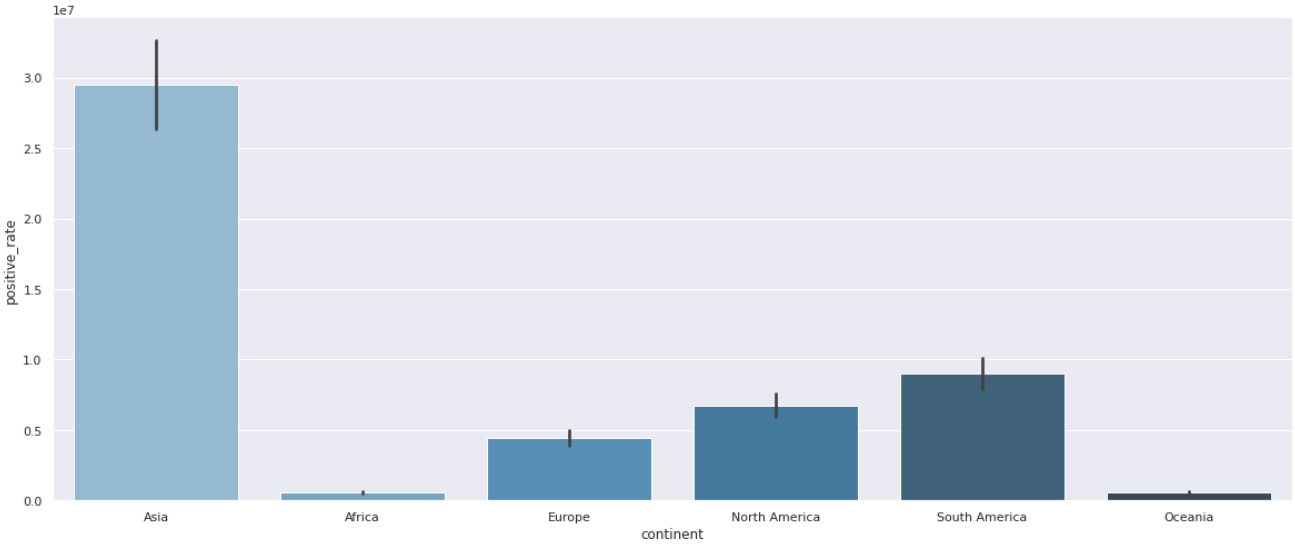


برای south America نیز میبینیم که آمار new case برای کشور برزیل از بقیه بیشتر است:

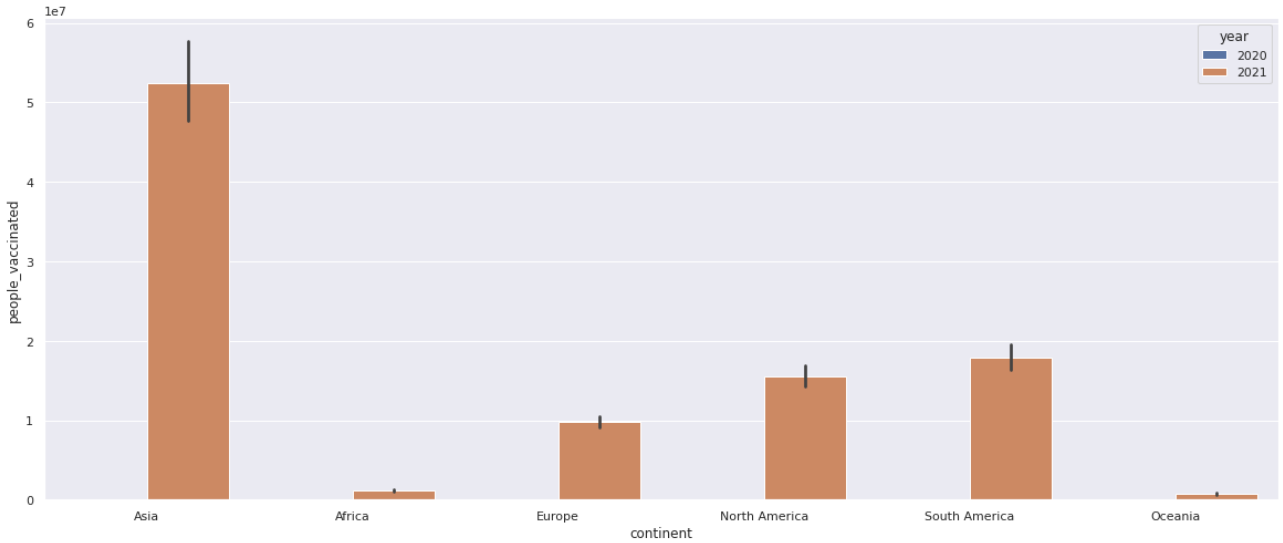


حال در ادامه چندین فاکتور مختلف را برای تمامی قاره ها باهم بررسی میکنیم:

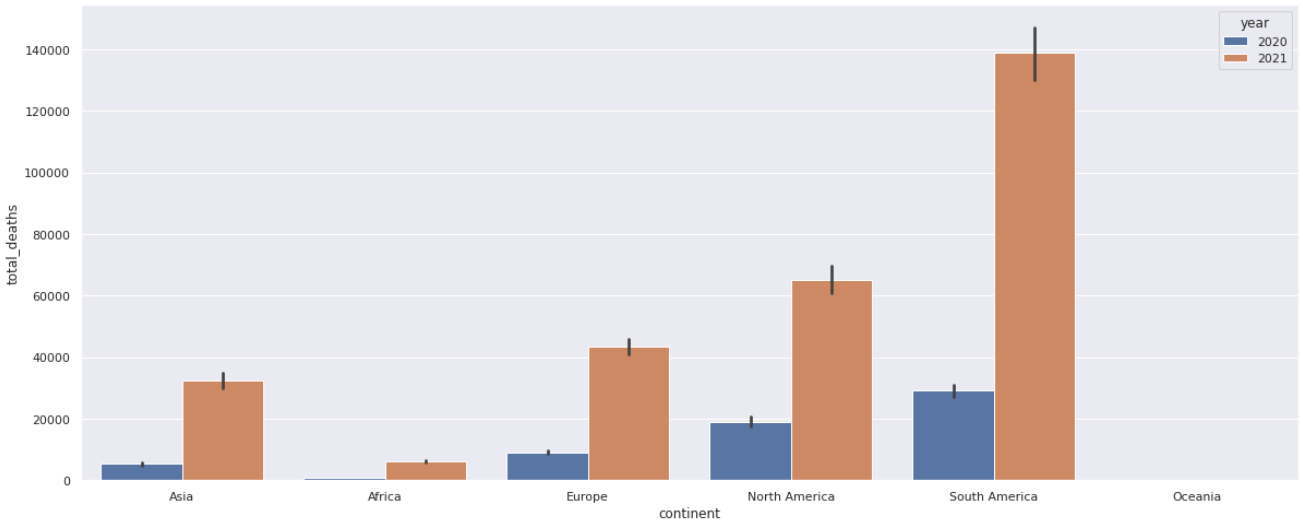
میزان مثبت شدن تست ها:



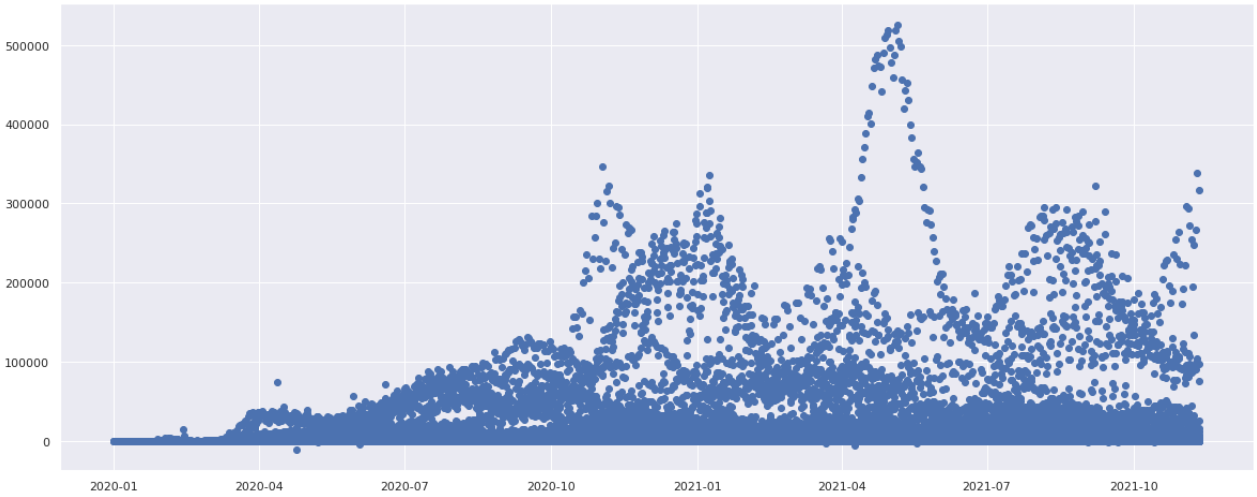
تعداد افرادی که در دوسال ۲۰۲۰ و ۲۰۲۱ واکسن زدند:



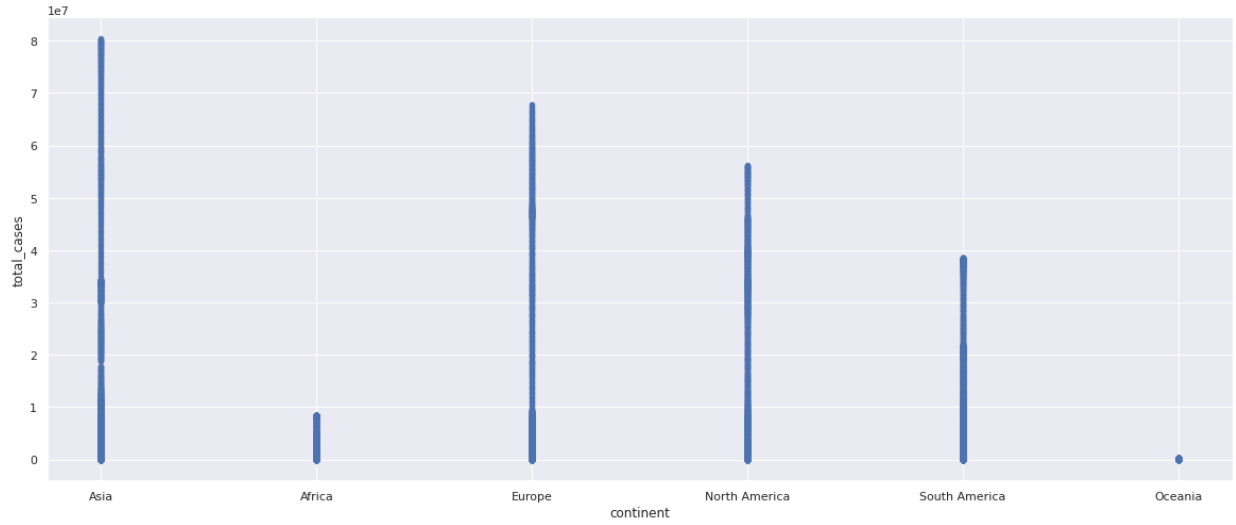
میزان مرگ و میر در دو سال متوالی ۲۰۲۰ و ۲۰۲۱:



نمودار زیر نشان میدهد که در کل دیتاست و در تمامی زمان ها تعداد کیس های جدید به چه صورت بوده است:

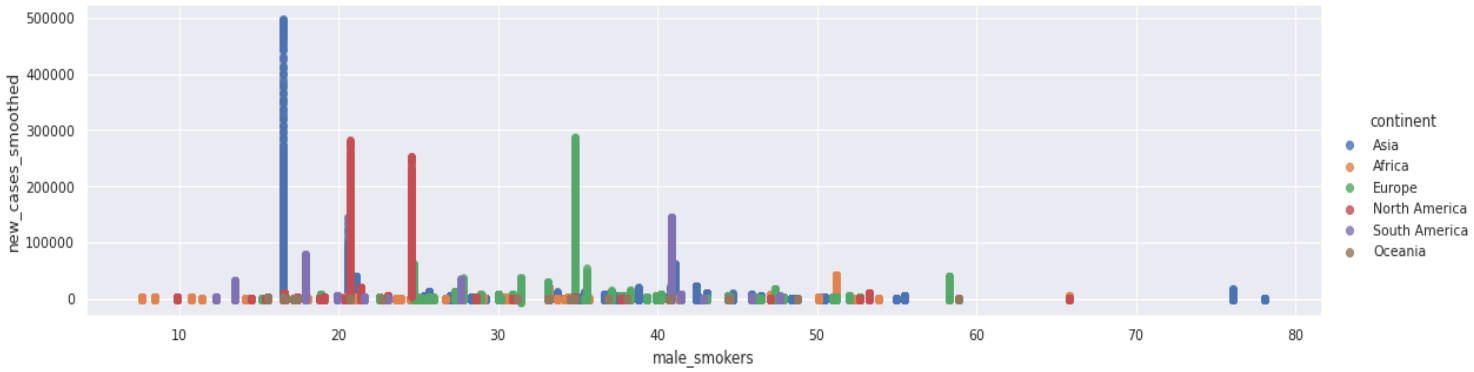


میزان total_cases برای قاره ها:



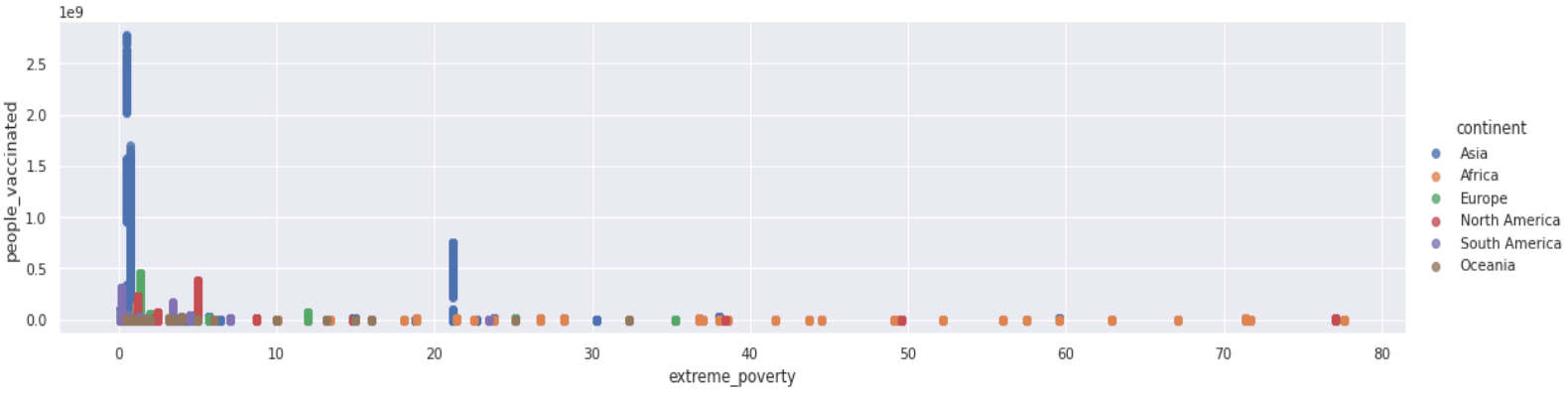
حال بر اساس قاره ها چند فاکتور مختلف را نیز باهم بررسی میکنیم:

در قاره های مختلف میزان male_smokers و میزان مبتلایان را بررسی کردیم:

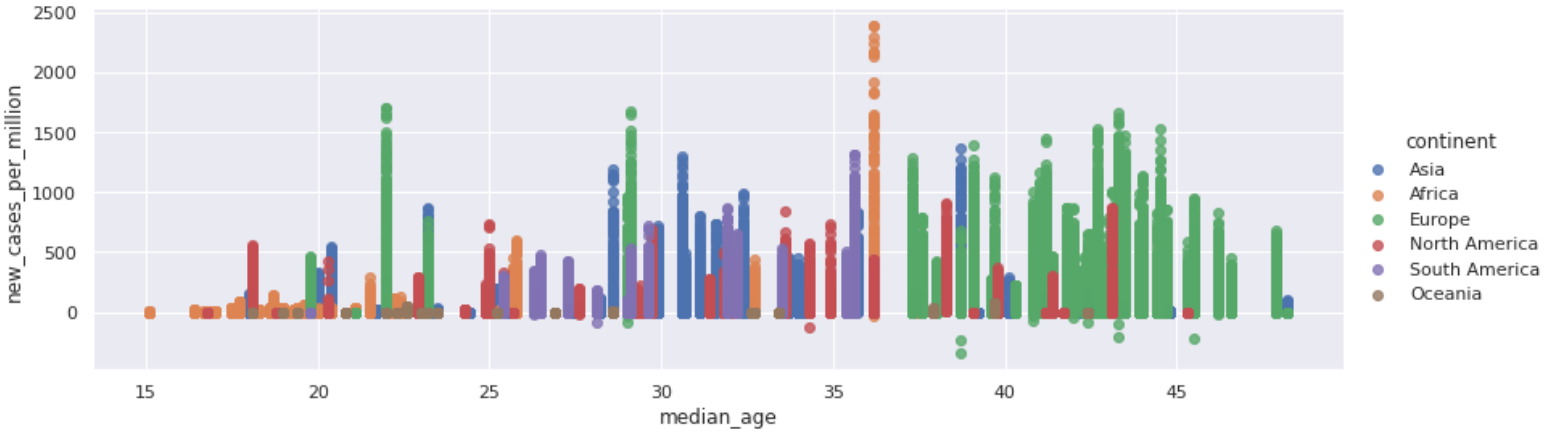


با توجه به نمودار بالا مشاهده میکنیم که با افزایش نرخ مرد های سیگاری، میزان ابتلای آنها کاهش یافته است.

در نمودار زیر با توجه به میزان فقیر بودن، تعداد افرادی که واکسن زده‌اند را بررسی میکنیم:



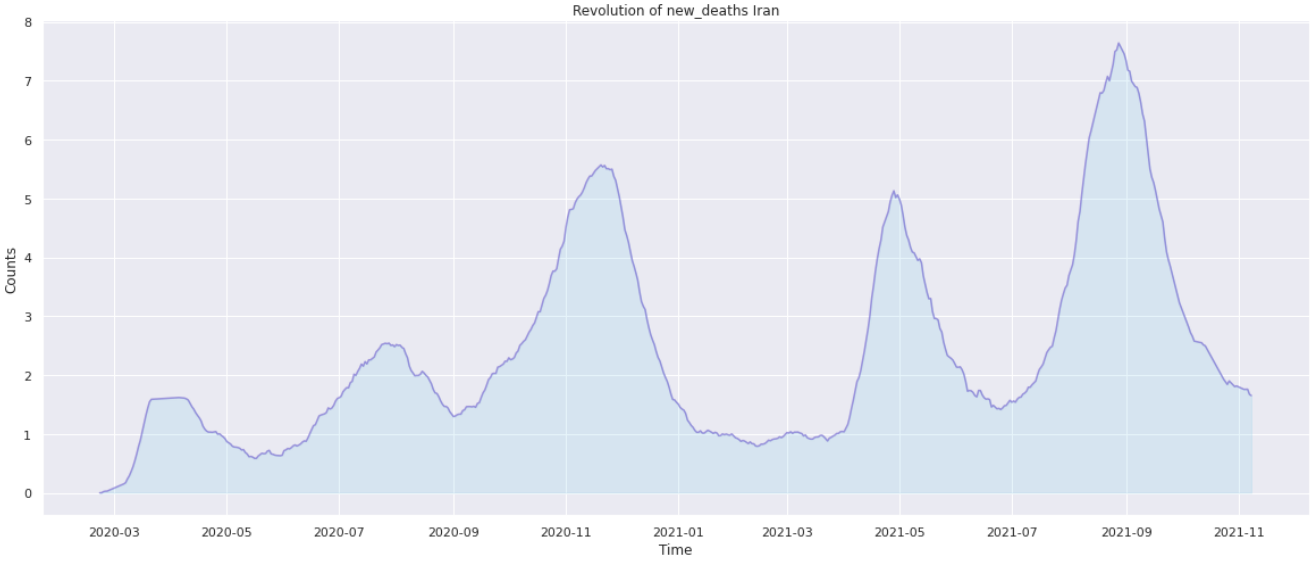
در نمودار بعد با توجه به سن افراد آمار مبتلایان را بررسی کردیم:



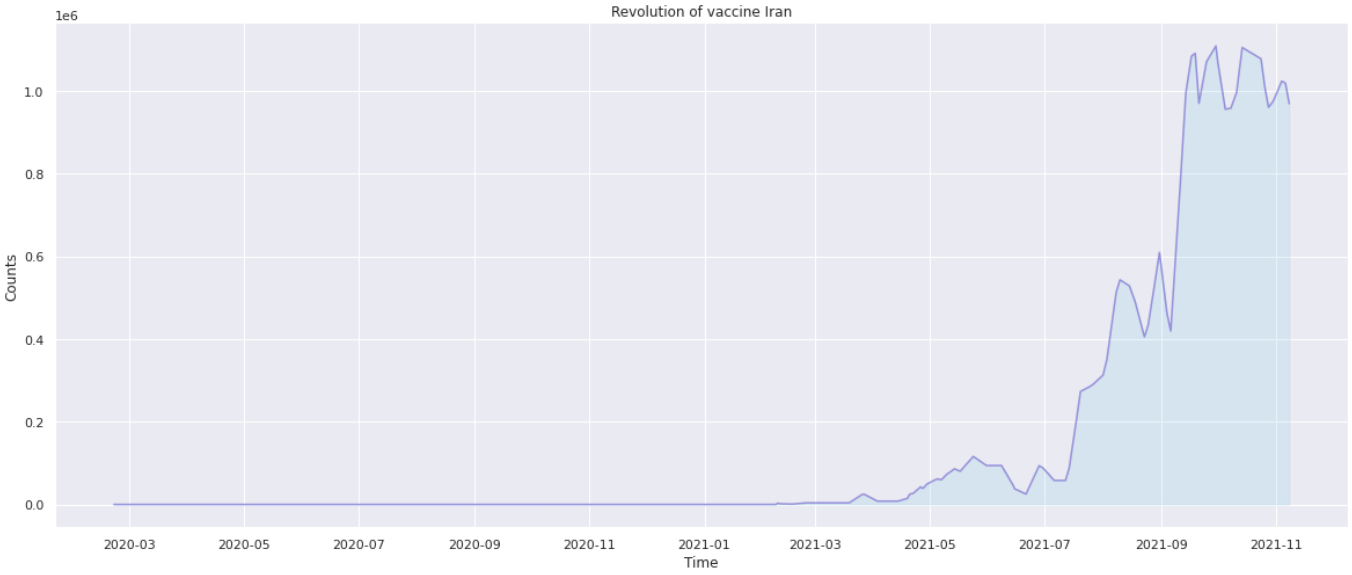
در قاره اروپا بیشتر مبتلایان بین سن ۴۰ الی ۵۰ هستند. در آمریکای جنوبی بیشتر بین ۳۰ الی ۳۵ و در آسیا نیز بیشتر بین ۲۵ الی ۳۵ میباشند.

سپس برخی نتایج را بر روی کشور ایران بررسی کردیم:

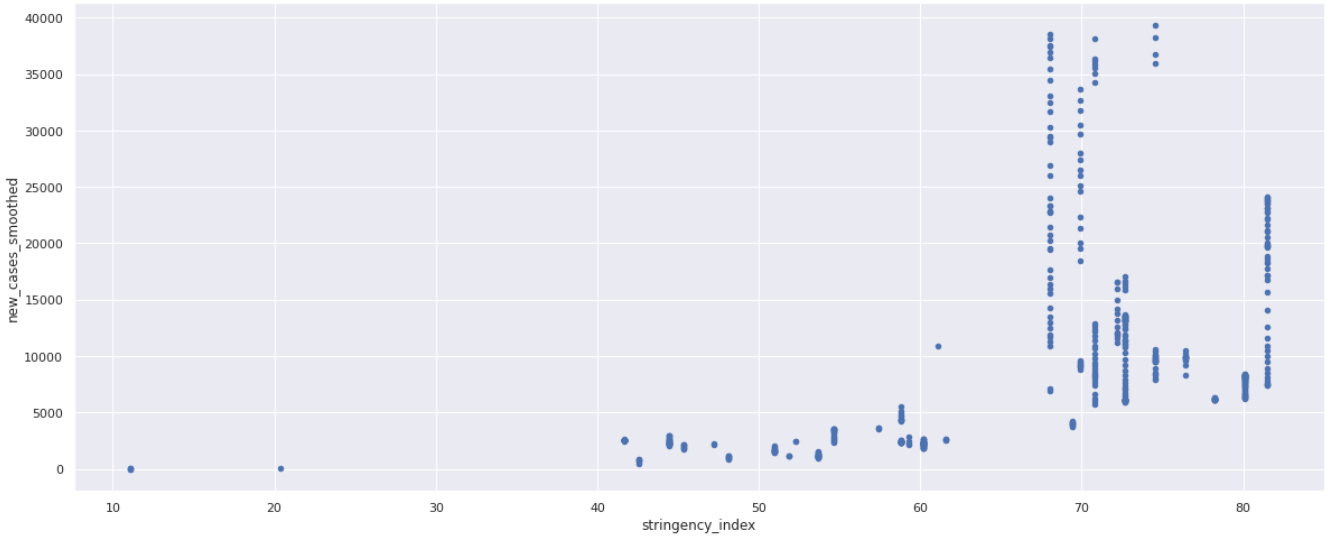
نرخ مرگ و میر بر اساس زمان



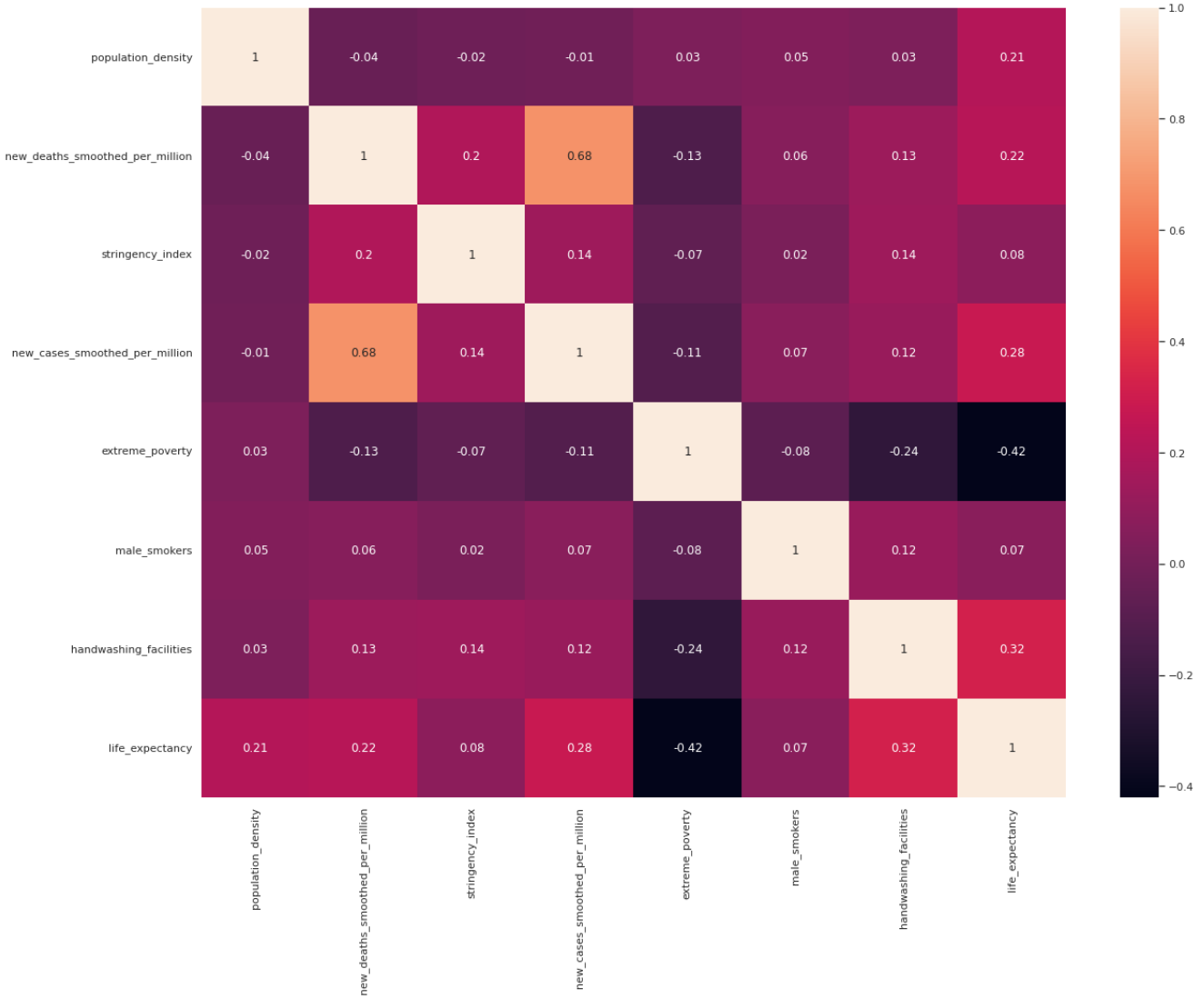
آمار واکسیناسیون بر اساس زمان:



آمار مبتلایان با توجه به نرخ سخت گیری دولت:



در انتها نیز با توجه به برخی ستون ها ماتریس کورولیشن را رسم کردیم تا ارتباط ستون ها با یکدیگر را دقیق تر مشاهده کنیم:

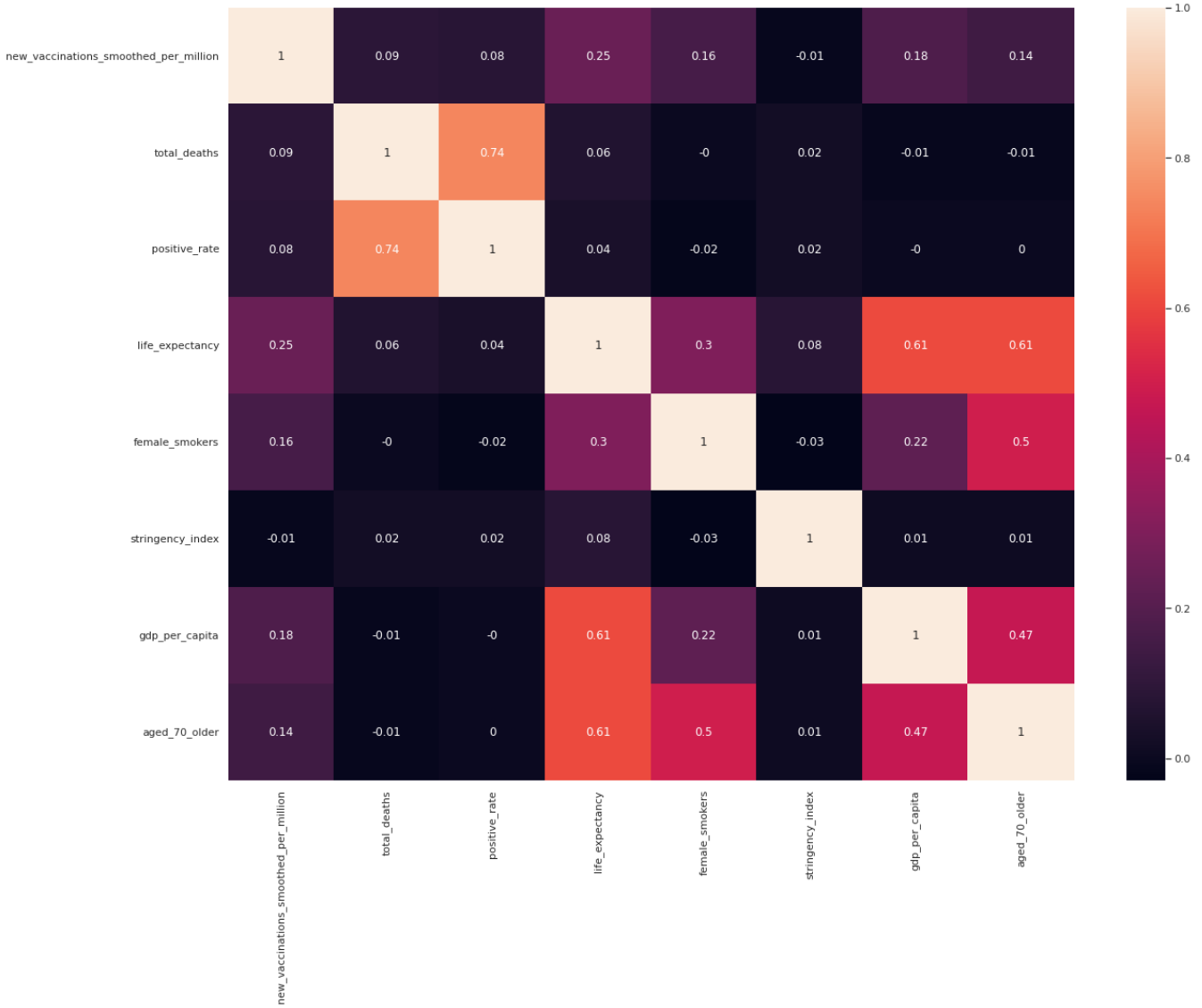


در این نمودار اگه مقادیر بیشتر از 0.2 برای رابطه دو ستون باشد یعنی این دو ستون باهم رابطه مستقیم دارند و اگر کمتر از 0.2- باشد نیز یعنی رابطه عکس دارند.

برای مثال extreme_poverty و life_expectancy مقدار 0.42- را دارند که یعنی هرچه میزان فقر بیشتر باشد امید به زندگی نیز کمتر است.

دو ستون new_death و new_cases باهم رابطه مستقیم بامقدار 0.68 دارند.

همچنین new_death و life expectancy نیز رابطه مستقیم دارند.



یک نمودار دیگر را نیز در بالا تماشا میکنیم. همانطور که میبینیم و منطقی نیز هست، با افزایش نتایج تست های مثبت(positive_rate) مرگ و میر نیز بیشتر شده است و این دو رابطه مستقیم با مقدار 0.74 دارند.

همچنین مشاهده میشود که دو ستون female_smokers و aged_70_older نیز باهم رابطه مستفیم دارند و این به این معناست که خانم های بالا ۷۰ سال بیشتر سیگار میکشند.