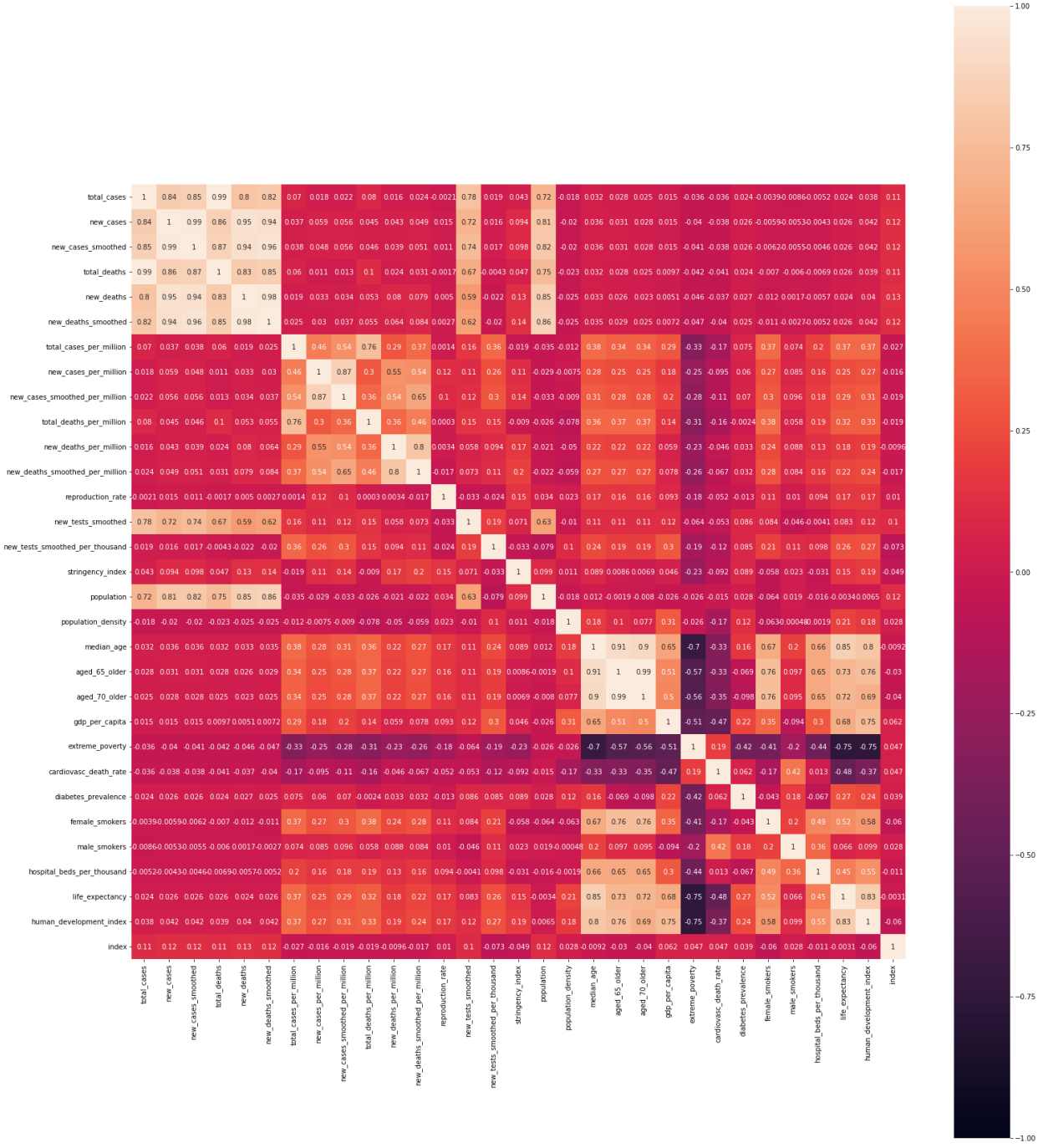
گزارش پروژه دوم علوم داده

محمدرضا صیدگر-97222055

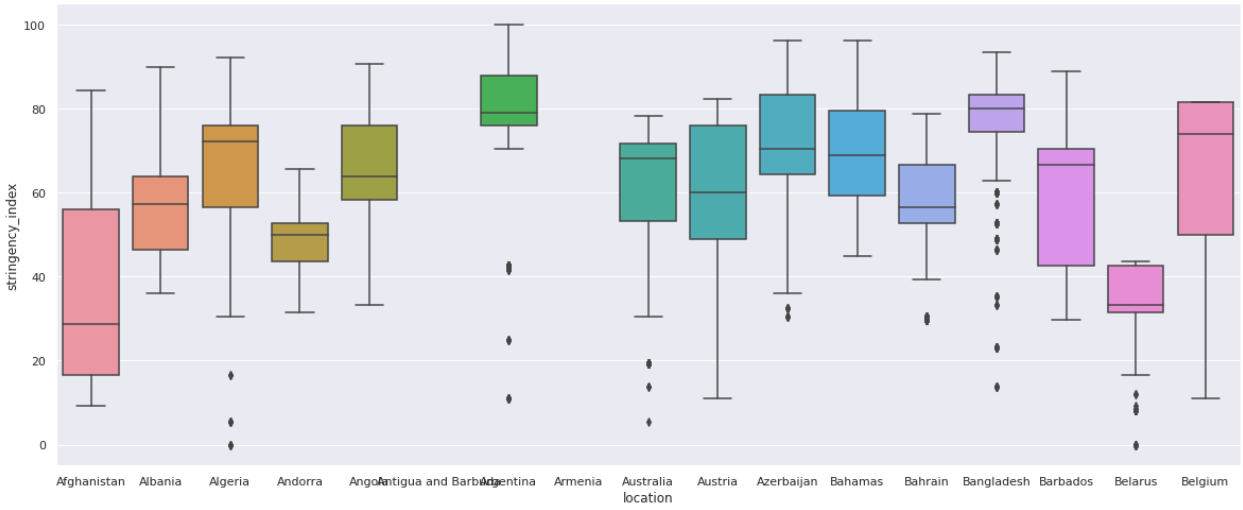
پروژه دوم دارای یک دیتاست از بیماری کوید-19 بود. داده ها را از گوگل درایو خوانده و همان اول یک اطلاعات کلی از آن بدست می آوریم .همه ستون ها بجز ico-code و continent و location و date و test-units داده های عددی هستند.
در مرحله بعدی تعداد داده های null را در دیتاست بررسی کردیم که به شدت دیتاست بد و پر از داده های خالی است که باید با روش های مختلف این قضیه را برطرف که من راهکار های مختلفی را امتحان کردم که جلوتر بهش میپردازم.
تعداد کل ستون ها همان اول 67 ستون است اما چون خیلی از داده ها خالی اند ستون هایی را که بیشتر از 50% آنها خالی است را حذف کردم که تعداد ستون ها با 35 ستون کاهش می یابد.
اما تعداد کل سطر ها یعنی تعداد داده های ما 133596 داده است اما سطر هایی را که بیشتر از 30% آنها خالی است را حذف کردیم تا 111726 داده باقی بماند.
چون جلوتر برخورد کردم به داده هایی که منفی بودند و در این دیتاست داده های منفی مفهومی نداره پس در همینجا داده های منفی را برابر با 0 کردم.
در این مرحله رفتم سراغ داده های کتگوریکال . در ستون continent فقط داده هایی null هستند که location آنها World است پس به همه آنها در continent برچسب All دادم که نشان دهنده همه قاره هاست.
ستون tests_units چون خیلی مورد نیاز نبود این ستون را کلا حذف کردم پس تمام داده های کتگوریکال تا اینجا دیگر null نیستند.
در مرحله بعدی ساغ داده های عددی رفتم.ستون هایی که میانگین آنها نزدیک به صفر است می توانیم به جای داده های خالی آن 0 بگذاریم مثل ستون human_development_index .
ستون new_cases فقط 3 داده خالی دارد پس آن را هم میتوان با 0 پر کرد.
با جدول کورولیشن زیر می توانیم متوجه شویم که با کمک کدام ستون میشود ستونی که داده خالی دارد را پیشبینی کرد.



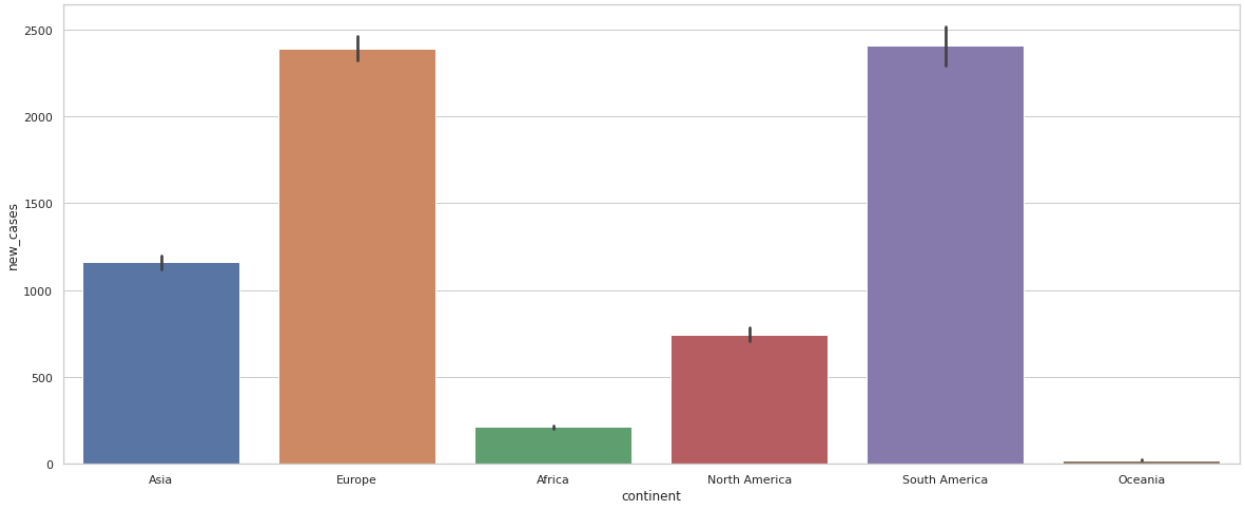
پس برای پر کردن new cases smoothed این ستون همان new_cases را بجای آن قرار دادم چون رابطه بسیار نزدیک باهم دارند.

برای پر کردن ستون total_deaths از مدل رگرسیونی برای پیش بینی استفاده کردم و برای این کار از ستون های total_cases و new_cases_smoothed استفاده کردم که روی داده های تست درستی 99 درصد داشت و در نهایت با مدل جاهای خالی را پر کردم.

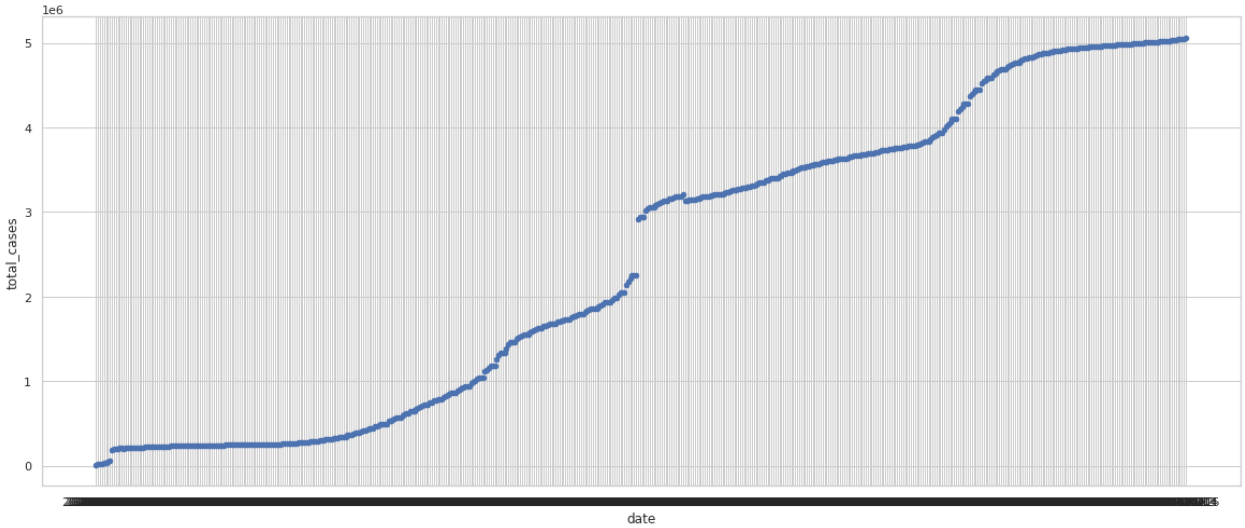
برای پر کردن ستون new_deaths هم از مدل رگرسیونی برای پیش بینی استفاده کردم و برای این کار از ستون های total_cases و new_cases_smoothed استفاده کردم که روی داده های تست درستی 89 درصد داشت و در نهایت با مدل جاهای خالی را پر کردم.
برای پر کردن new_deaths_smoothed هم همان new_deaths را جای داده های خالی گذاشتم.
ستون هایی که کلا تقسیم بر میلیون داشت را حذف کردم چون خیلی به کار نمی آمدند.
ستون reproduction_rate با گرفتن یک توصیف میانگین 1 را داشت با انحراف معیار 0.3 که یعنی 67 درصد داده های ستون تو همین باز است پس داده های خالی را با 1 پر کردم.
برای ستون new_cases_smoothed هم باز از پیشبینی مدل رگرسیونی با درستی 70% استفاده کردم و ستون new_tests_smoothed_per_thousand را هم حذف کردم.



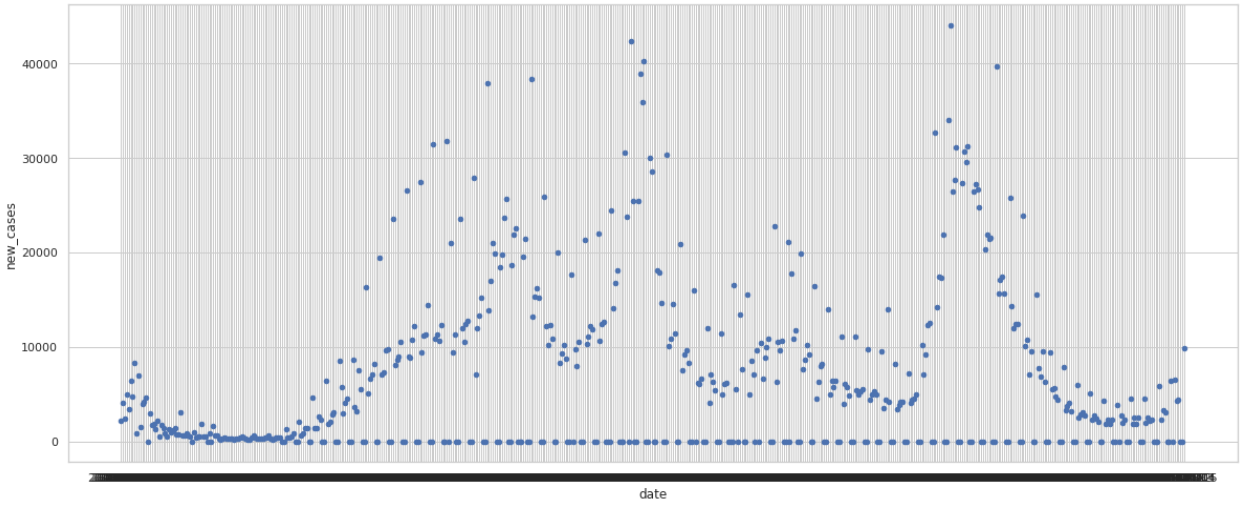
این نمودار پراکندگی شاخص سختگیری هر کشور را نشان میدهد که البته فقط تعداد محدودی کشور قابل مشاهده است. از این نمودار میتوان فهمید هر کشور یک رنج سختگیری خاص خود را دارد پس داده های خالی ستون stringency_index را با توجه به میانگین شاخص هر کشور پر کردم و در ادامه داده هایی از این ستون که همچنان پر نشدند را چاره نیست جز اینکه با 0 پر کرد.
در آخر هم تمام داده های خالی ستون های others را باز با توجه به میانگین هر کشور پر کردم تا تمام داده های null از بین بروند.
برای از بین بردن داده های outlier نمیشد همینطور روی هر ستون داده هایی را که خارج از محدوده قانونی(3 تا انحراف معیار از میانگین) هستند را حذف کرد زیرا اینجا جنس داده های متفاوت است و به طور مثال داده های هر کشور رنج خاص خود را دارد مثلا با انجام کار بالا تقریبا تمام داده های کشور آمریکا حذف میشود چون آمار بسیار بالایی در گرفتن کرونا و مرگ و میر دارد پس حذف کردن داده های outlier برای هر کشور به صورت جدا انجام دادم و داده ها از 111726 به 100799 کاهش پیدا کرد.
اما در مرحله بعد روی داده ها تحلیل و بررسی انجام دادیم:



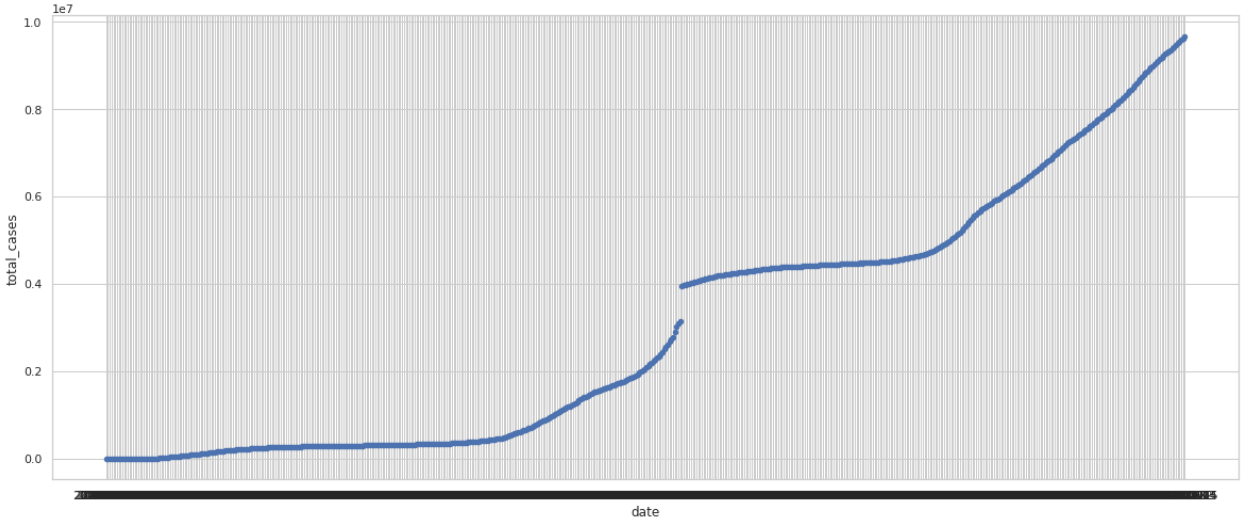
نمودار بالا پراکندگی مورد های جدید کرونا را براساس قاره های مختلف نشان میدهد که بیشتر میزان کرونای جدید در یک روز را اروپا و آمریکای جنوبی داشته اند و بعد آسیا و آمریکای شمالی و در آخر هم آفریقا و اقیانوسیه.

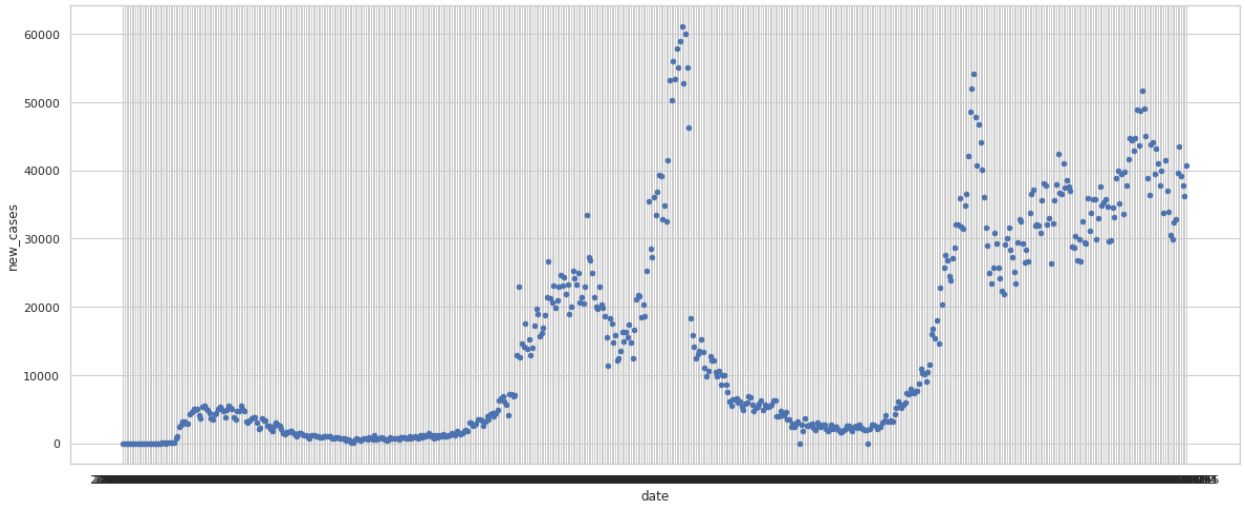


این نمودار محور x آن زمان و محور y آن total case است که روند بالا رفتن تعداد کسانی است که در اسپانیا کرونا گرفته اند.

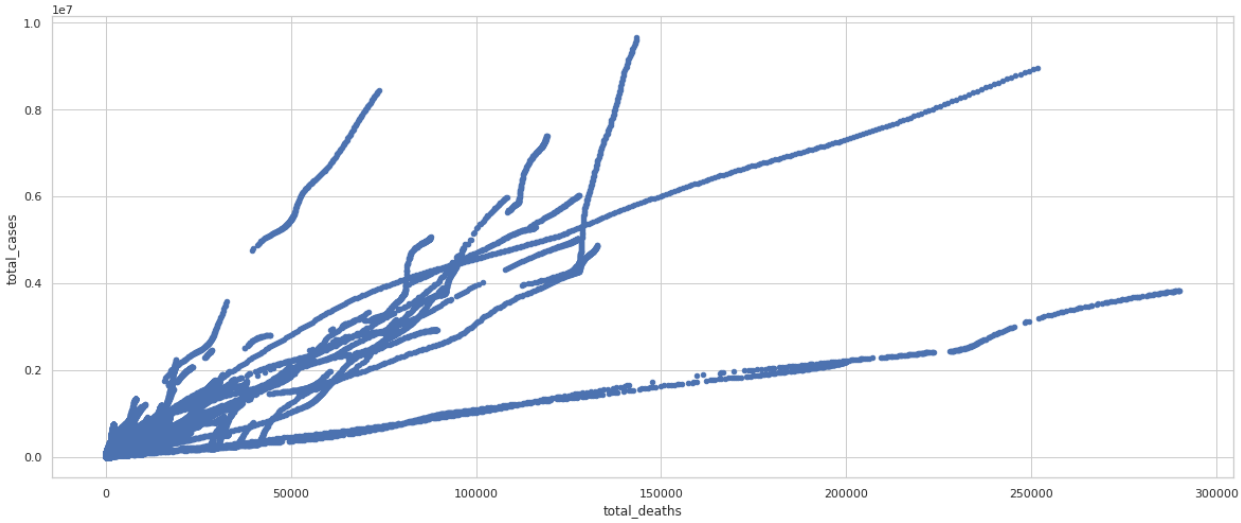


این همان نمودار قبلی است با این تفاوت که y آن new case است یعنی در تعداد افرادی که در روز به عنوان کورونایی جدید ثبت شده اند که نمودار پیک های مختلف کرونا در طول زمان در کشور اسپانیا نشان میدهد.

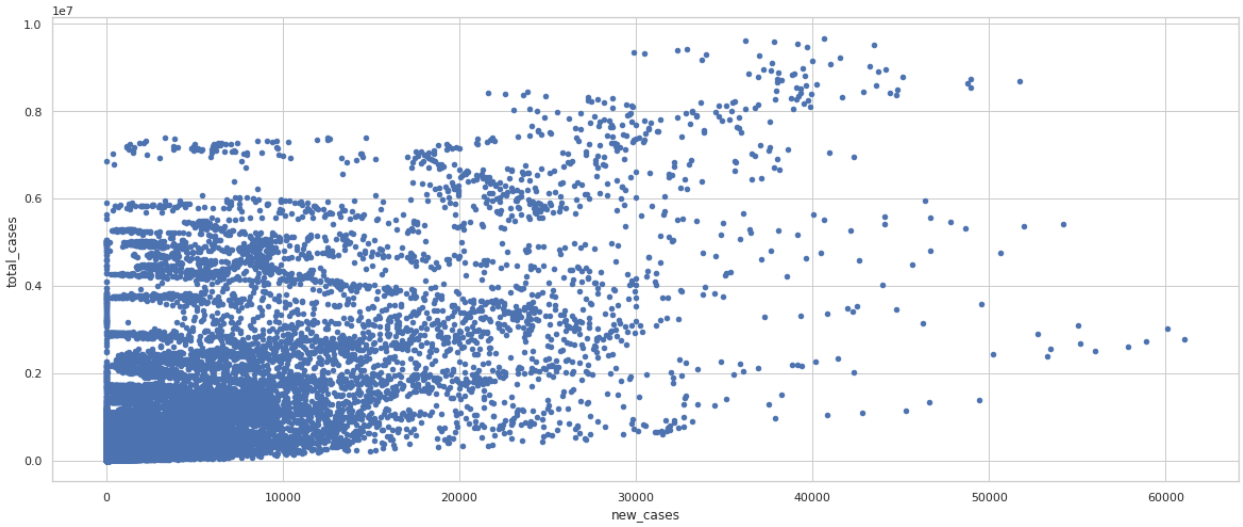




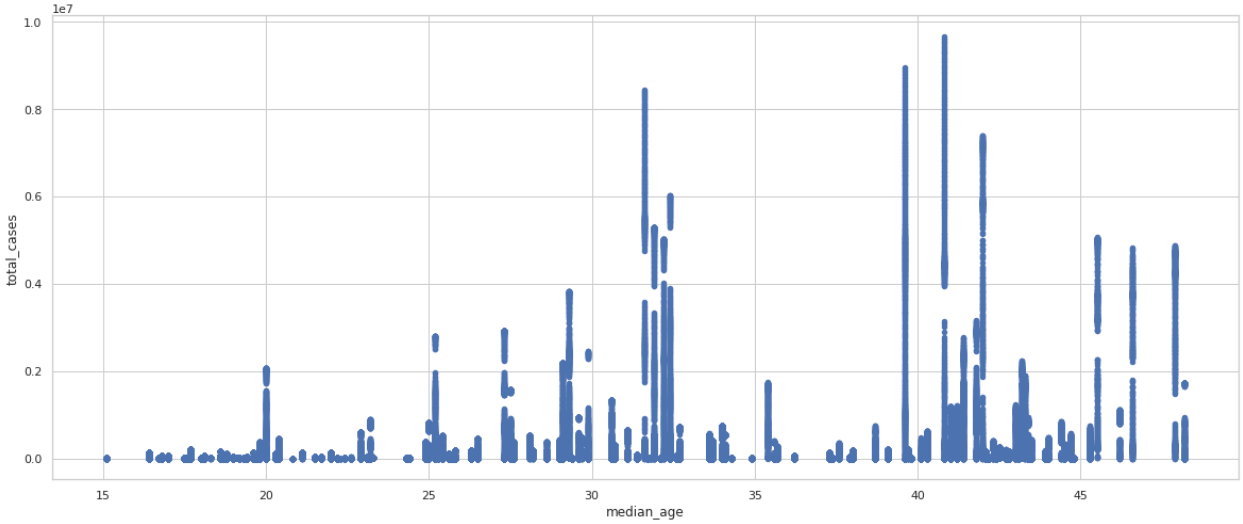
این دو نمودار هم دقیقا همان اطلاعات قبلی است با این تفاوت که این اطلاعات را در کشور انگلستان مشاهده می کنیم و روند کرونا را در این کشور می بینیم که دقیقا آن روزهایی که تعداد مورد های جدید کرونا کاهش میابد شیب نمودار بالایی هم کند می شود.



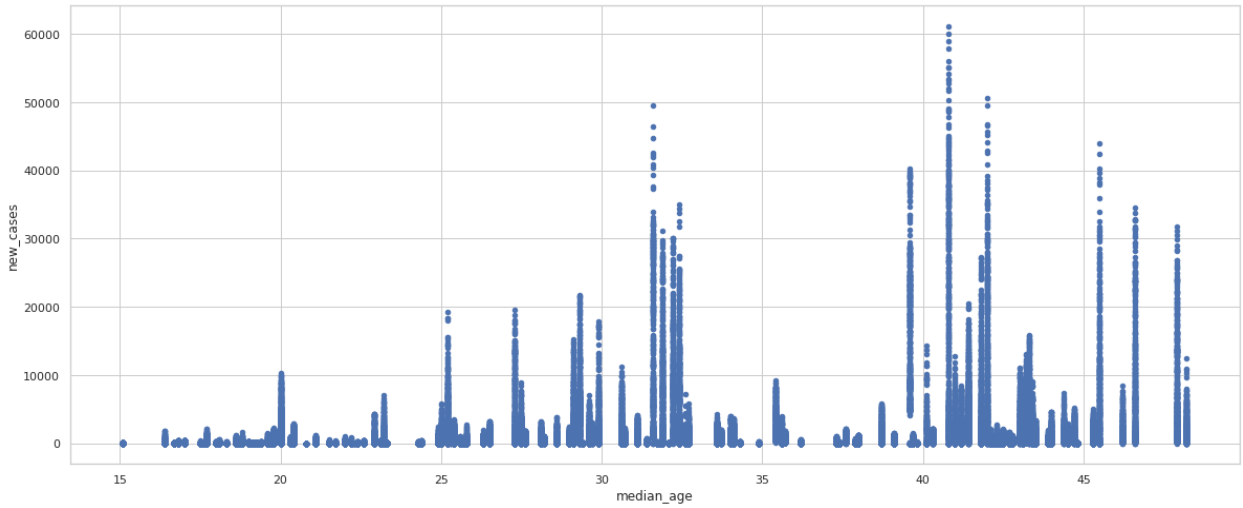
این نمودار رابطه بین دو متغیر total death و total cases را نشان میدهد که وقتی تعداد کل موارد کورونایی زیاد میشود تقریبا تعداد مرگ و میر کرونا هم بیشتر میشود یه جورایی با هم در ارتباط اند.



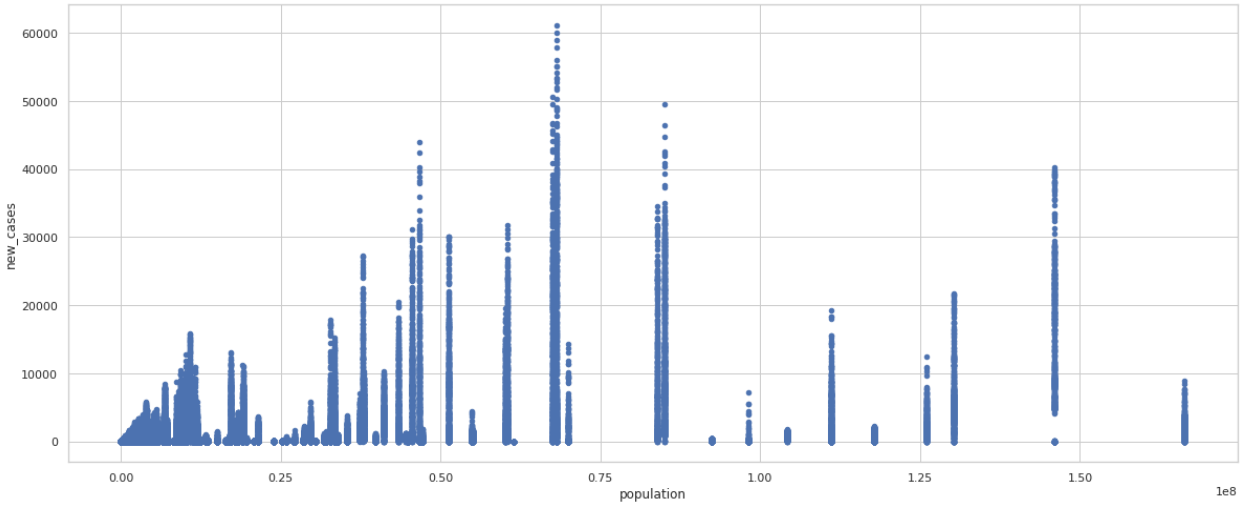
این نمودار رابطه بین دو متغیر new cases و total cases را نشان میدهد که رابطه خاصی با هم ندارند و نقاط همه جای صفحه پراکنده اند بجز قسمت های پایین سمت راست چون با تعداد مورد های زیاد جدید نمیشود که تعداد مورد های کلی کم باشد.



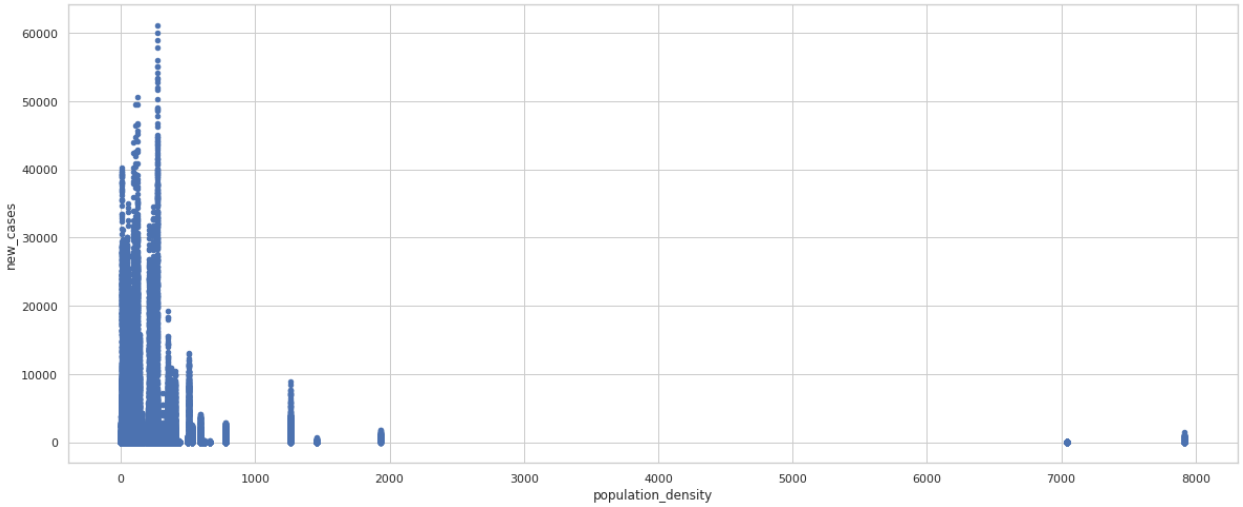
این نمودار نشان میدهد که وقتی تعداد مورد های کلی کرونا در کشورها زیاد میشود میانگین سنی هم بالا میرود و نشان میدهد که کرونا اغلب انسان های با سن بالا تر را نسبت به جوان ها درگیر میکند.



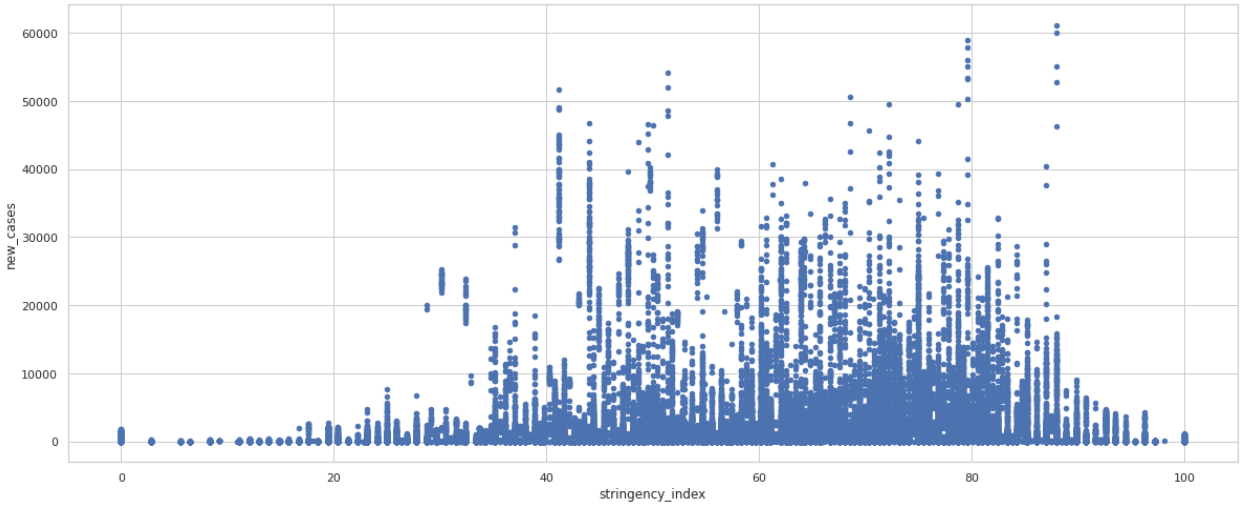
این هم همان نمودار قبلی است با این تفاوت که مورد های جدید کرونا در یک روز را نشان میدهد که باز هم نتایج قبلی تصدیق میشوند که یعنی وقتی مورد های جدید زیادی در یک روز میبینیم میانگین سنی بالا میرود.



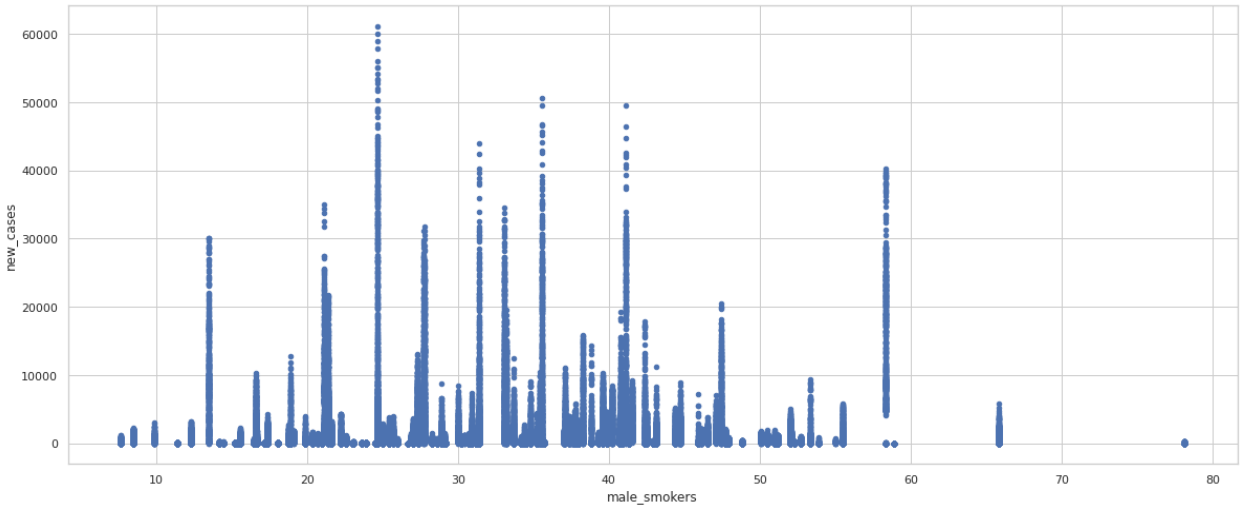
این نمودار تعداد مورد جدید کرونایی را در روز بر اساس جمعیت آن کشور ارائه میدهد که نشان میدهد کشور با جمعیت کمتر در یک روز تعداد کورونایی کمتری دارد ولی در جمعیت زیاد خیلی دقیق نمی توان پاسخ داد و بالا پایین دارد ولی کلا از کشور های با جمعیت خیل کم بیشتر مورد کورونایی دارند.



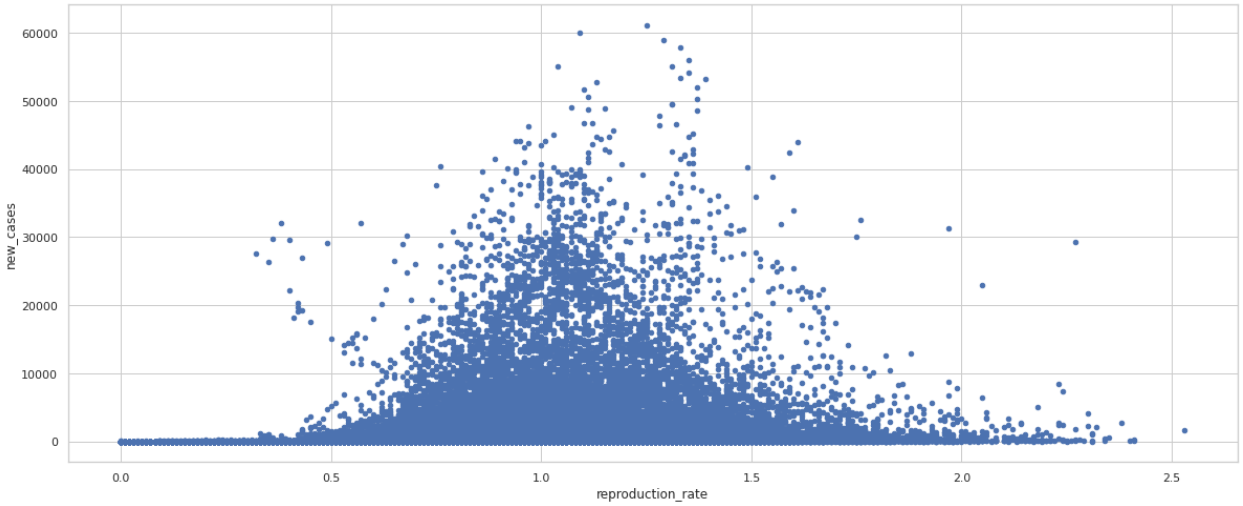
این نمودار تعداد مورد جدید کرونایی را در روز بر اساس تراکم جمعیت آن کشور ارائه میدهد که نشان میدهد که خیلی دید خوبی را نمی دهد شاید بخاطر داده های خالی ای که در این ستون باشد و یا شاید هم کورونا ربطی به تراکم جمعیت نداشته باشد.



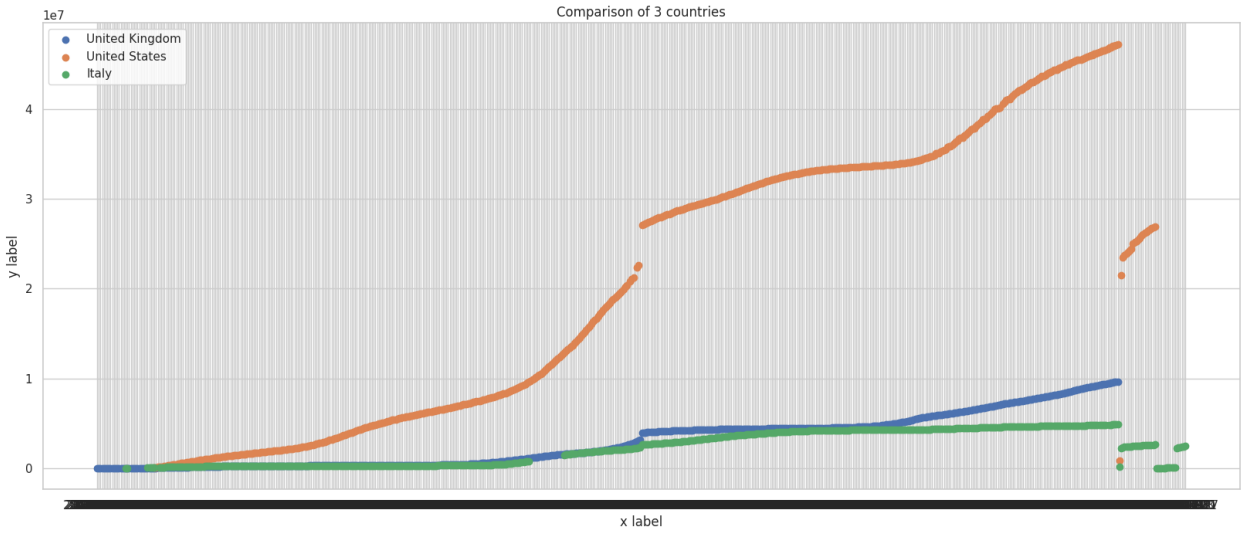
این نمودار یه جورایی نوع برخورد دولت ها با میزان مورد کورونایی های جدیدشان در روز را نشان میدهد که یعنی وقتی مورد کورونایی کم است دیگر سخت گیری ندارند ولی وقتی موارد کرونایی در روز زیاد میشود سختگیری را بیشتر میکنند تا جایی که خیلی سختگیری زیاد میشود و دوباره موارد کورونایی کم میشود.



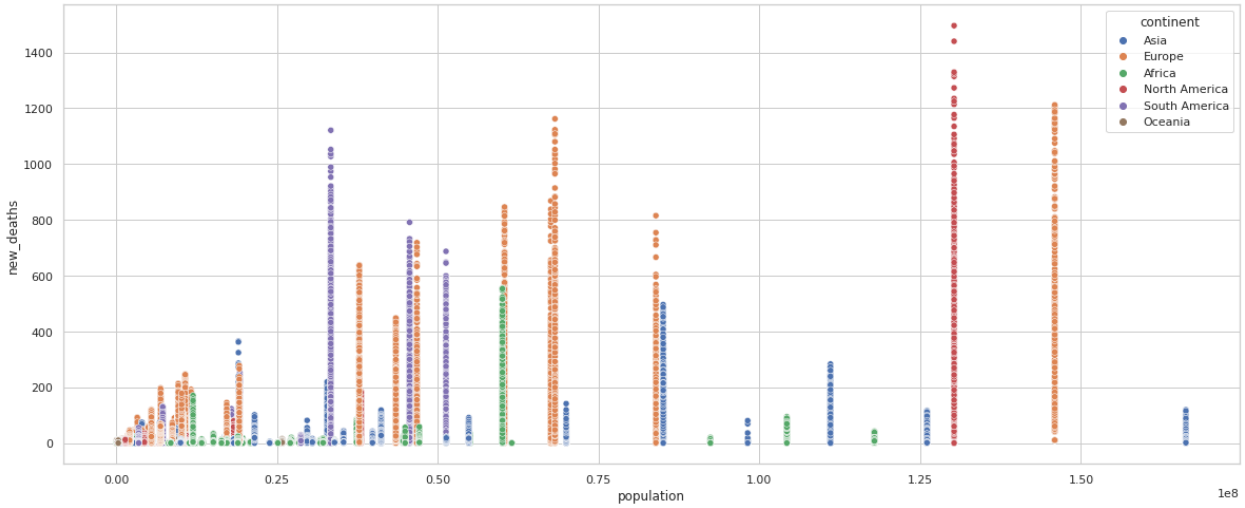
این نمودار موارد کورونایی جدید در روز را براساس مصرف سیگار مرد ها نشان می دهد که نشان میدهد سیگار کشیدن زیاد یا کم فرقی در بیماری کرونا ندارد.



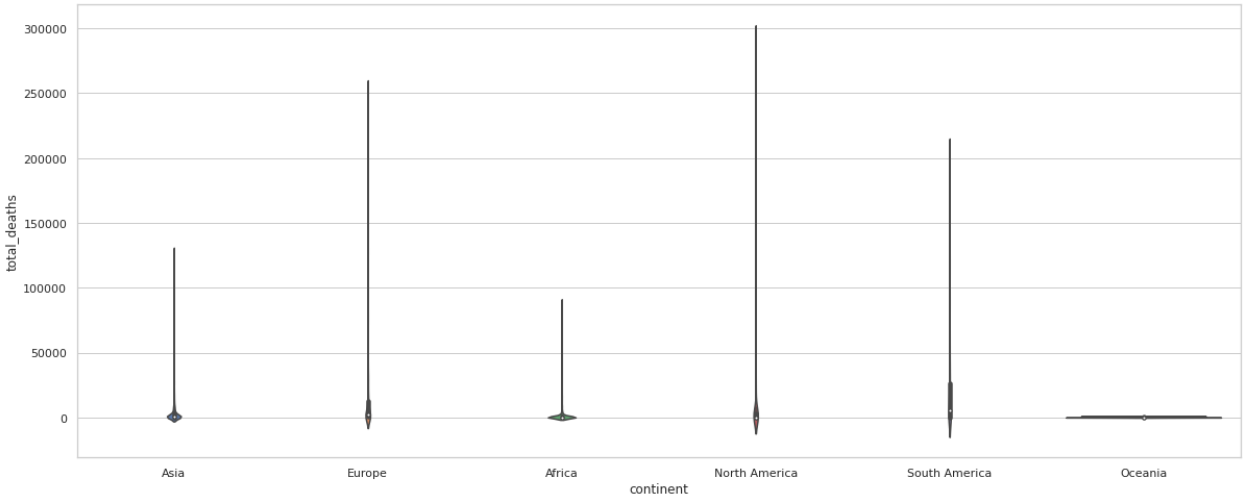
این نمودار هم موارد کورونایی جدید را براساس نرخ رشد بیماری نشان میدهد که شاید این هم باز واکنش دولت را نشان میدهد که وقتی نرخ رشد زیاد میشود سختگیری ها زیاد و موارد کورونایی هم کم میشود.



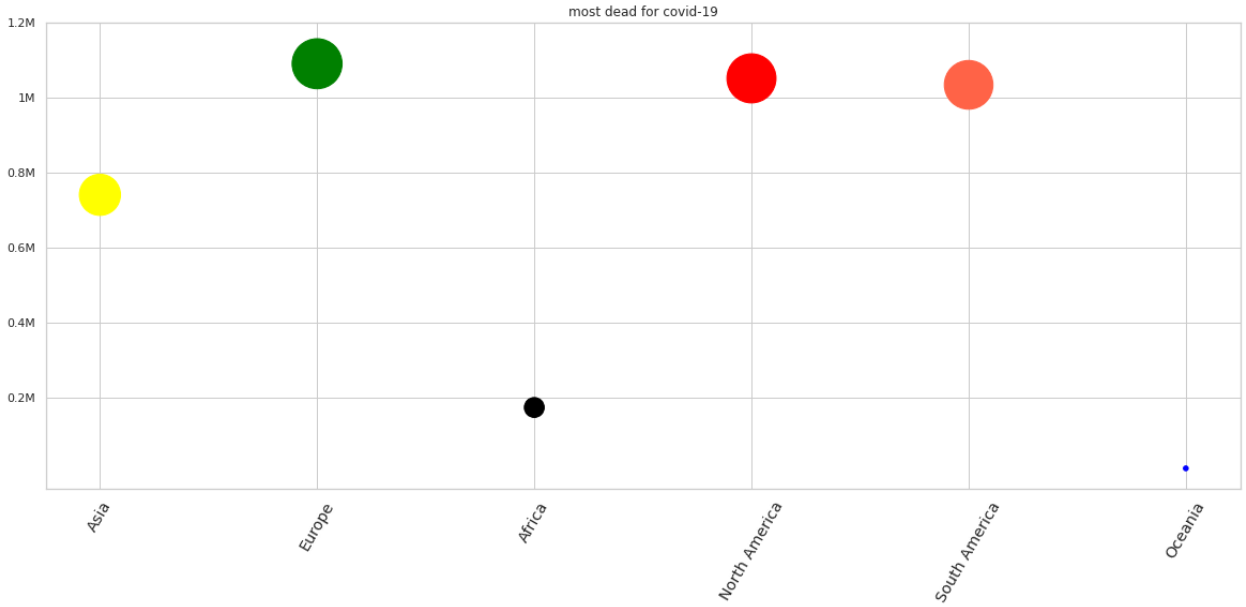
این نمودار زیاد شدن موارد کلی کرونا را در 3 کشور انگلستان و آمریکا و ایتالیا نشان میدهد که مشخص است وضعیت کشور آمریکا بسیار بدتر از کشور انگلستان و ایتالیا بوده است.



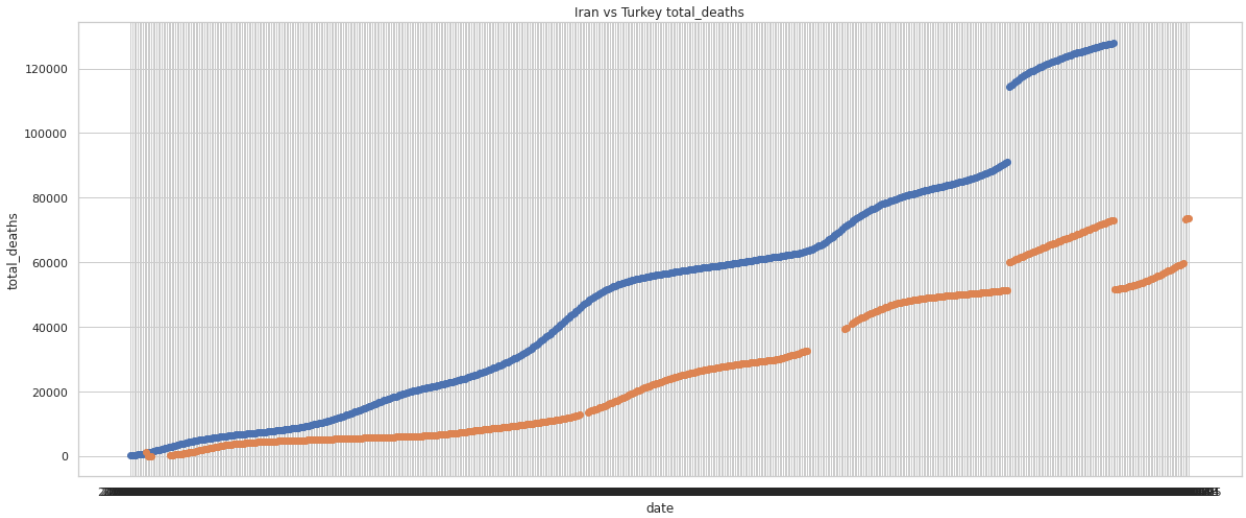
این نمودار تعداد مرگ های جدید در روز را بر اساس جمعیت کشور ها نشان میدهد که هر قاره با رنگ متفاوت مشخص است.نشان میدهد به طور مثال در آمریکای شمالی که تعداد مرگ و میر زیاد بوده است جمعیت هم زیاد است و این از دلایل آن است.



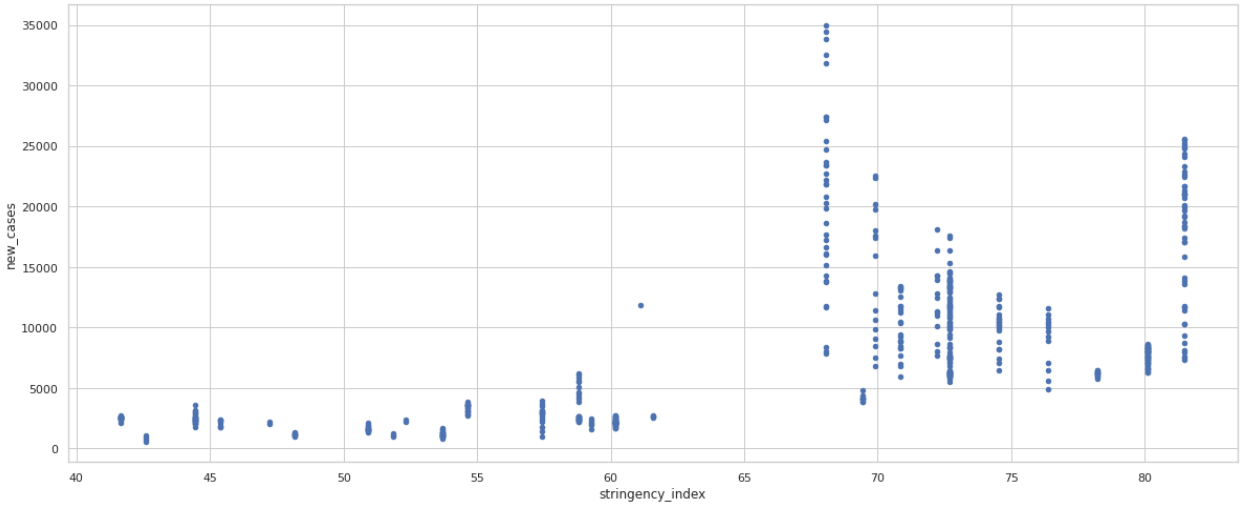
این نمودار پراکندگی مرگ و میر کلی را بر اساس هر قاره نشان میدهد که کلا آمریکای شمالی و اروپا بیشترین مرگ و میر کلی را داشته اند.



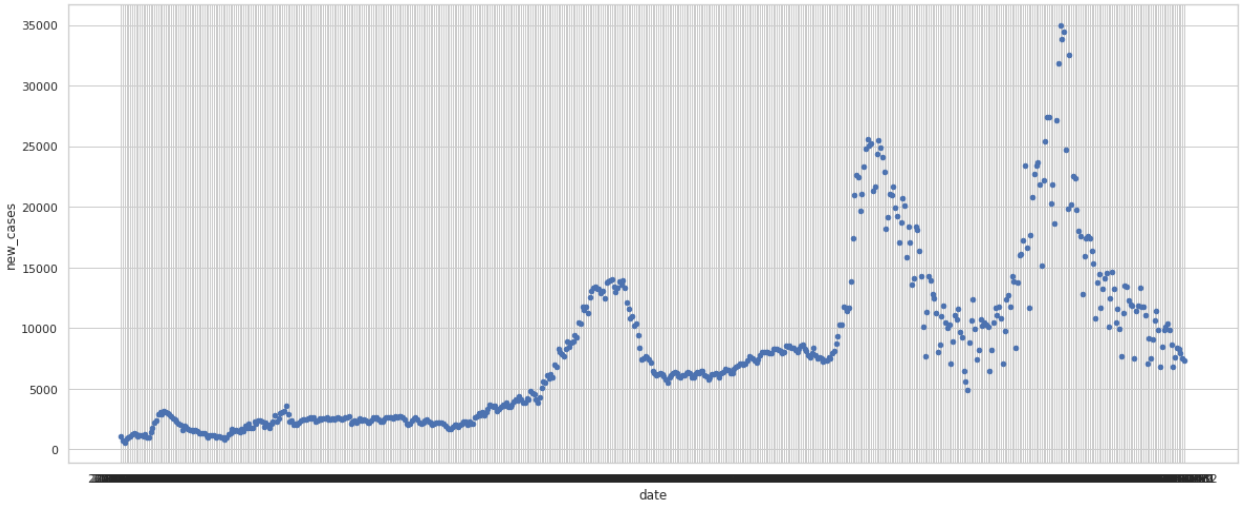
این هم همان اطلاعات قبلی را نمایش میدهد با شکلی بهتر. آمریکای شمالی و اروپا و آمریکای جنوبی هرکدام حدود 1 تا 1.2 میلیون مرگ و میر داشتند و بعد آسیا با 0.8 میلیون مرگ و آفریقا با 0.2 میلیون مرگ و اقیانوسیه که بسیار ناچیز بوده است.



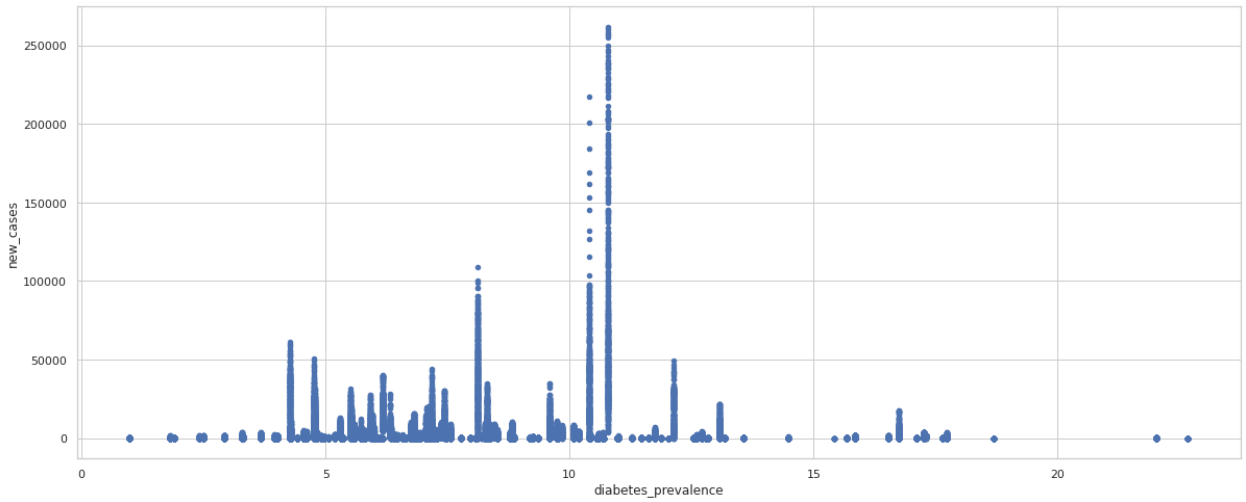
این نمودار آمار مرگ و میر کلی کرونا را در طول زمان در ایران و ترکیه بررسی میکند که ایران نمودار آبی و ترکیه نمودار نارنجی است که یعنی ایران وضعیت بدتری نسبت به کشور همسایه خود داشته است .



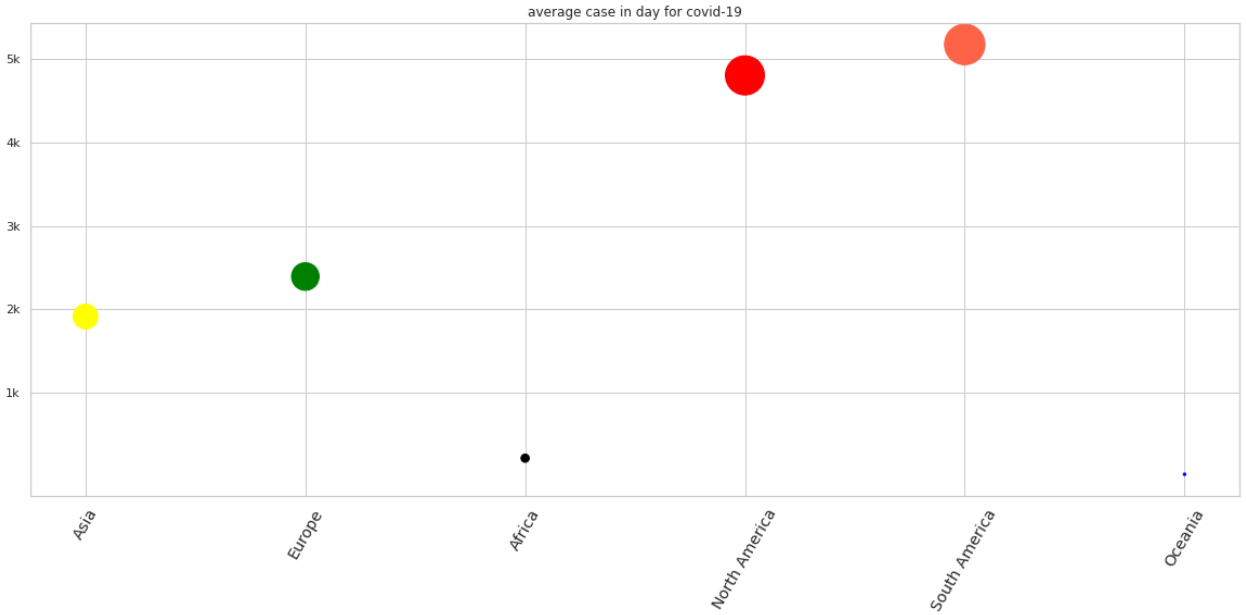
این هم واکنش دولت ایران نسبت به موارد کورونایی جدید را نشان میدهد که اکثرا مواقع در حالت 65 تا 75 است و مواردی هم بسیار سختگیر بوده اند.



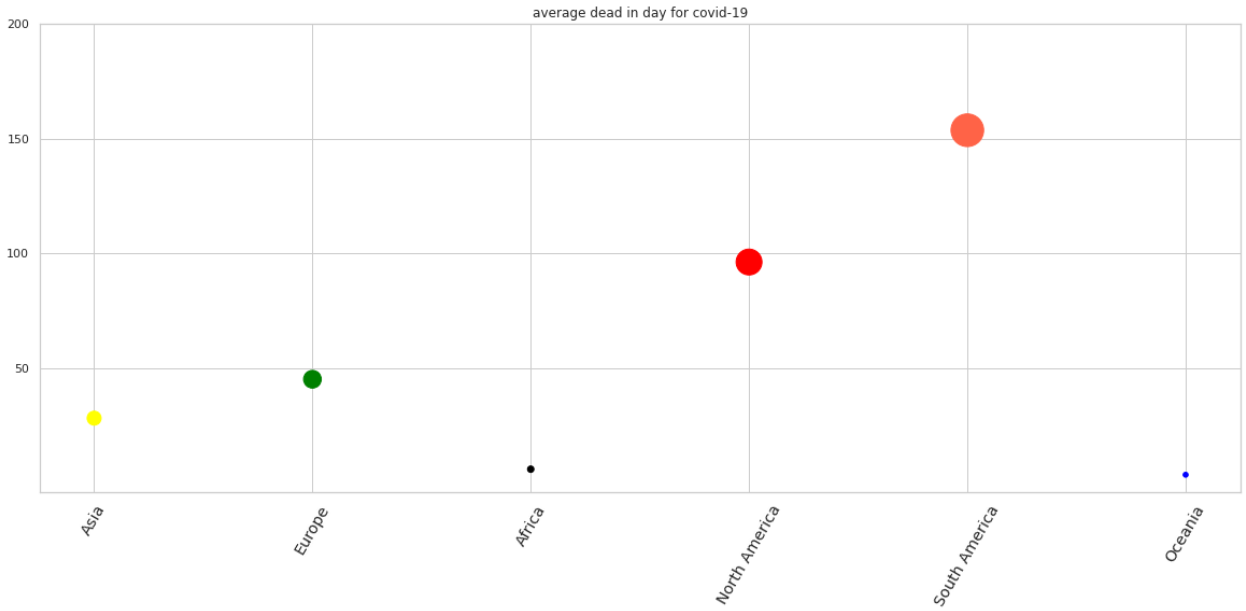
این هم موارد کورونایی های جدید را در طول زمان در ایران نشان میدهد که دقیقا نشان دهنده پیک هایی است که در طول این 2 سال شاهدش بودیم.



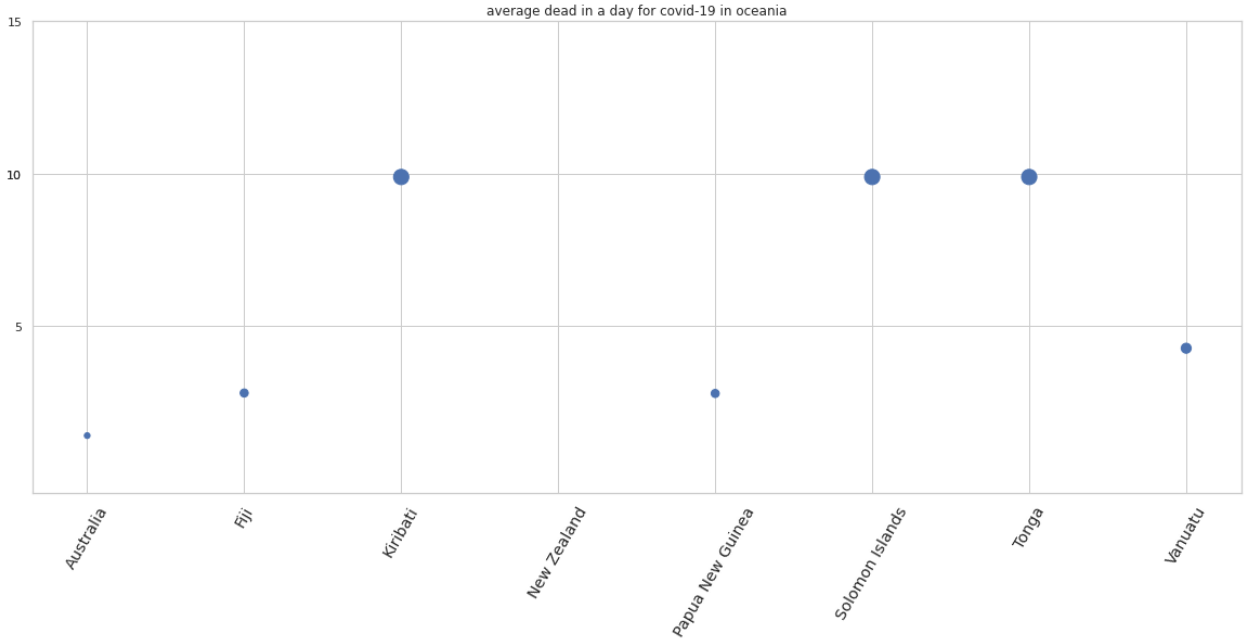
این نمودار موارد جدید کورنایی را بر اساس میزان دیابت نشان میدهد که نشان میدهد که این بیماری ارتباط چندانی با دیابت ندارد.



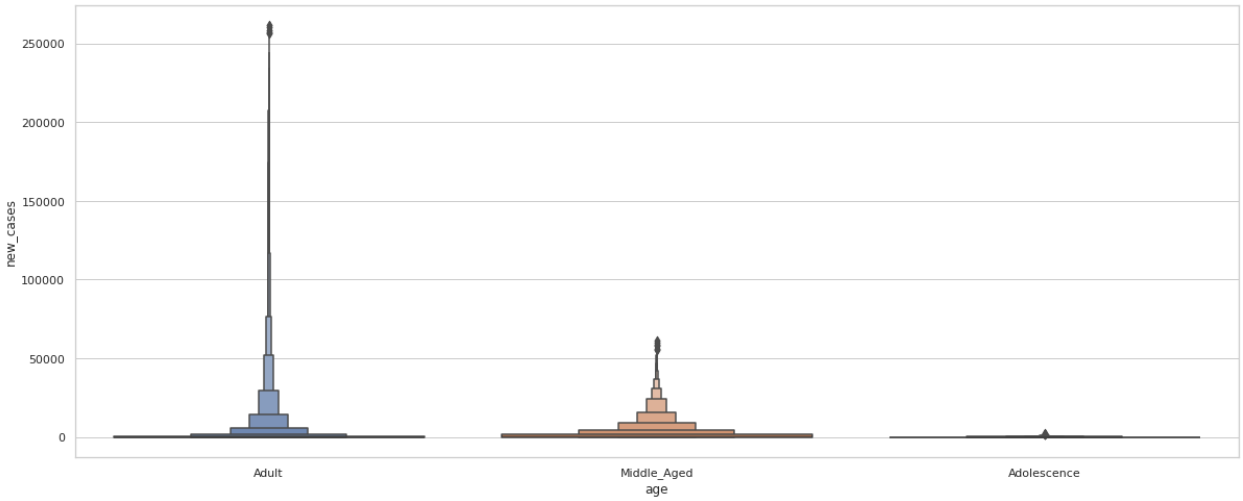
این نمودار میانگین تعداد افراد کورونایی در هر روز را در هر قاره نشان میدهد که به طور متوسط آمریکای جنوبی آمار بالاتری دارد و بعد آمریکای شمالی و سپس اروپا و آسیا و در نهایت آفریقا و اقیانوسیه.



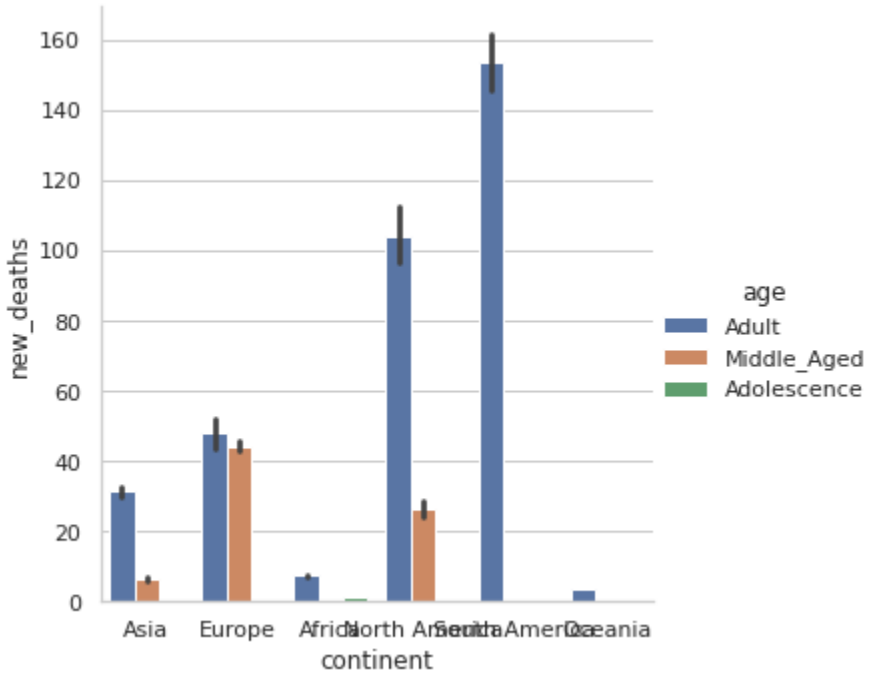
این نمودار میانگین تعداد افرادی که بخاطر کرونا در هر روز مردند را در هر قاره نشان میدهد که به طور متوسط آمریکای جنوبی آمار بالاتری دارد و بعد آمریکای شمالی و سپس اروپا و آسیا و در نهایت آفریقا و اقیانوسیه.



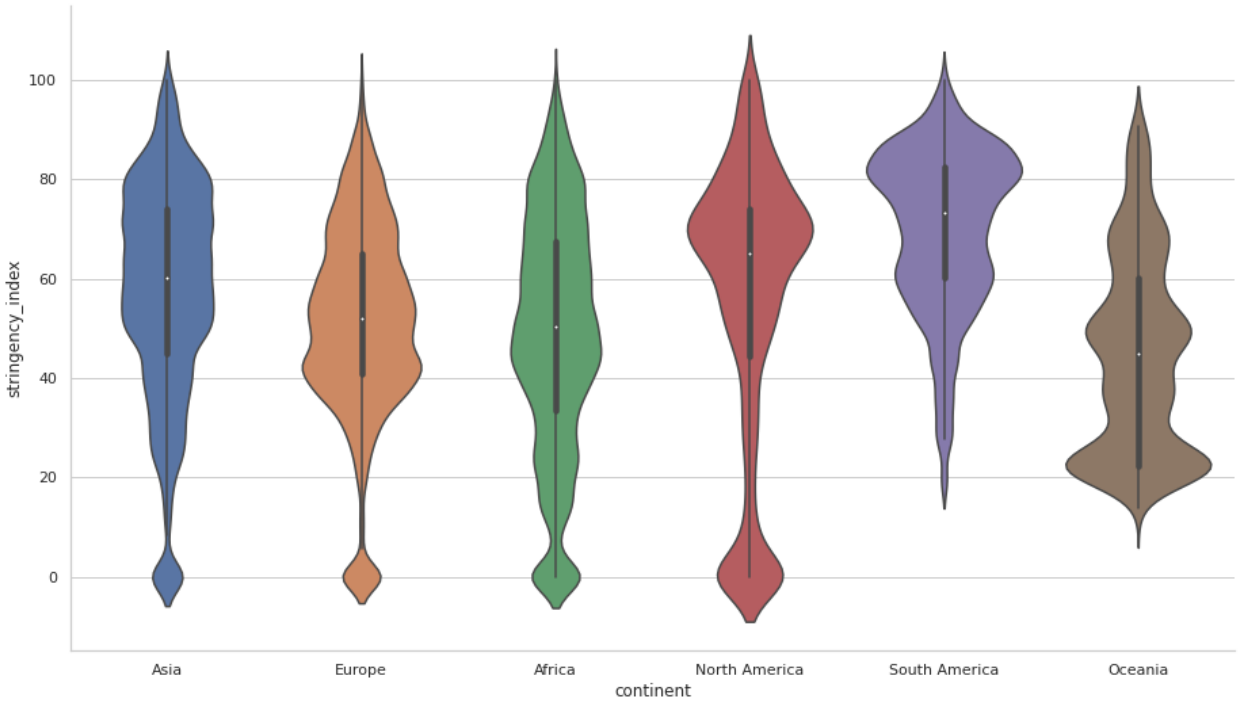
چون کلا در آمار ها اقیانوسیه پایین بود این هم آمار مرگ از کرونا در کشور های اقیانوسیه که نشان میدهد کیریباتی و جزیره سولومون و تونگا آمار بالاتری داشتند.



در آخر هم آمار موارد جدید کورونایی را بر اساس رده سنی مختلف میدهد که همانطور که گفتیم کمترین موارد کرونا از افراد نوجوان بوده است.



این نمودار هم مرگ های جدید را در قاره های مختلف بر اساس رده سنی نشان میدهد که در اروپا افراد میانسال در اندازه افراد بالغ مرده اند ولی در آمریکای جنوبی اکثرا افراد بالغ بر اثر کرونا مرده اند.



این نمودار نشان دهنده سختگیری کشور ها است بر اساس هر قاره که نشان میدهد اقیانوسیه و آمریکای جنوبی سختگیری های بیشتری داشته اند هرچند آمریکای جنوبی آمار های بالایی در کرونا داشته است.