نام و نام خانوادگی: درسا الهوئی

شماره دانشجویی: ۹۷۲۲۲۰۰۵

شماره تمرین: ۲ **تاریخ تحویل: ۲۸ آبان ۱۴۰۰**

عنوان تمرین: EDA & Data Visualization

---

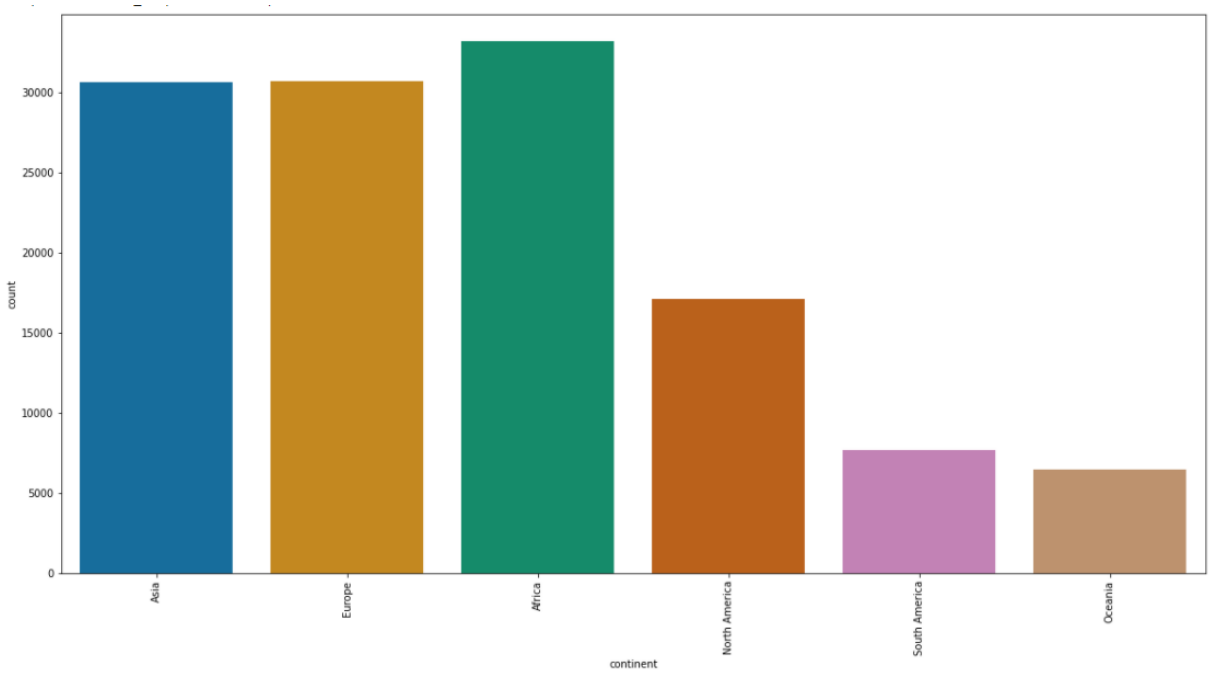
۱. مقدمه

هدف از این تمرین تجزیه و تحلیل مکاشفه ای و بصری سازی داده های مربوط به ویروس کرونا می باشد.

۲. گزارش

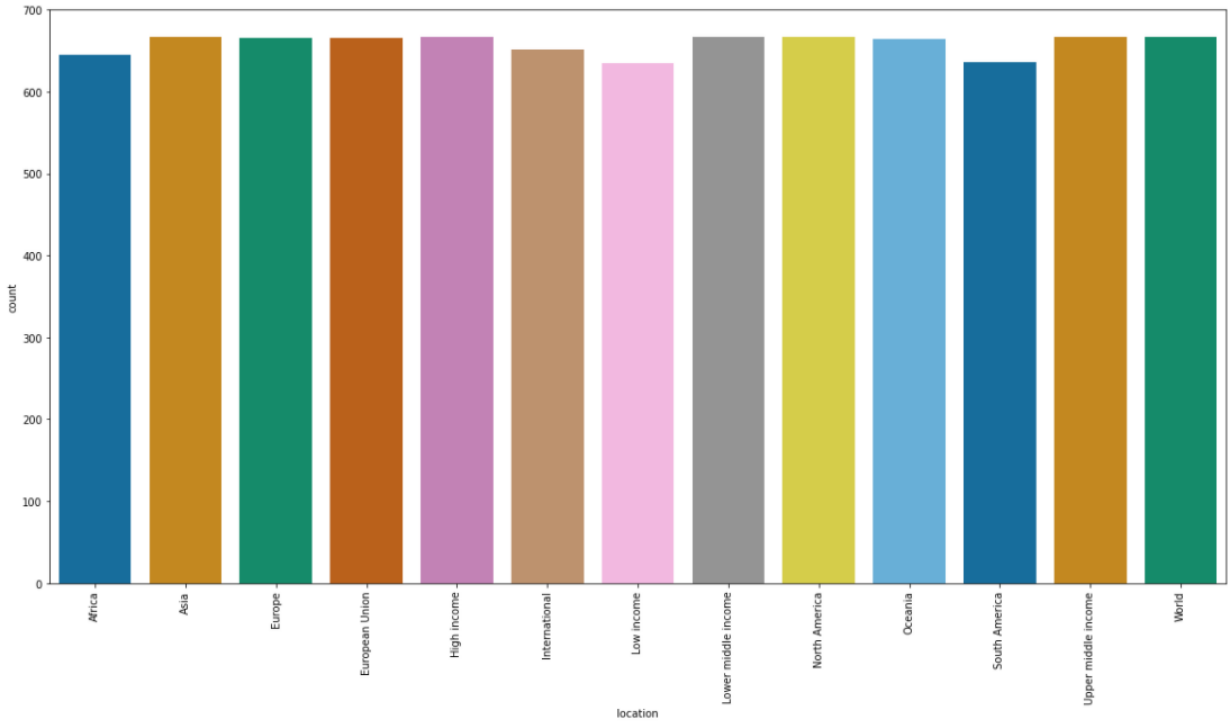
ابتدا داده ها را از جهت وجود سطر تکراری بررسی کردیم. نتیجه بیانگر عدم وجود سطر تکراری بود. در ادامه از ستون date که فرمت string داشت، ستون های سال و ماه و روز را ساختیم و به dataset اولیه اضافه کردیم.

در گام بعد، با استفاده از countplot، حجم داده هایی که از هر قاره می آیند را مشخص کرده ایم.

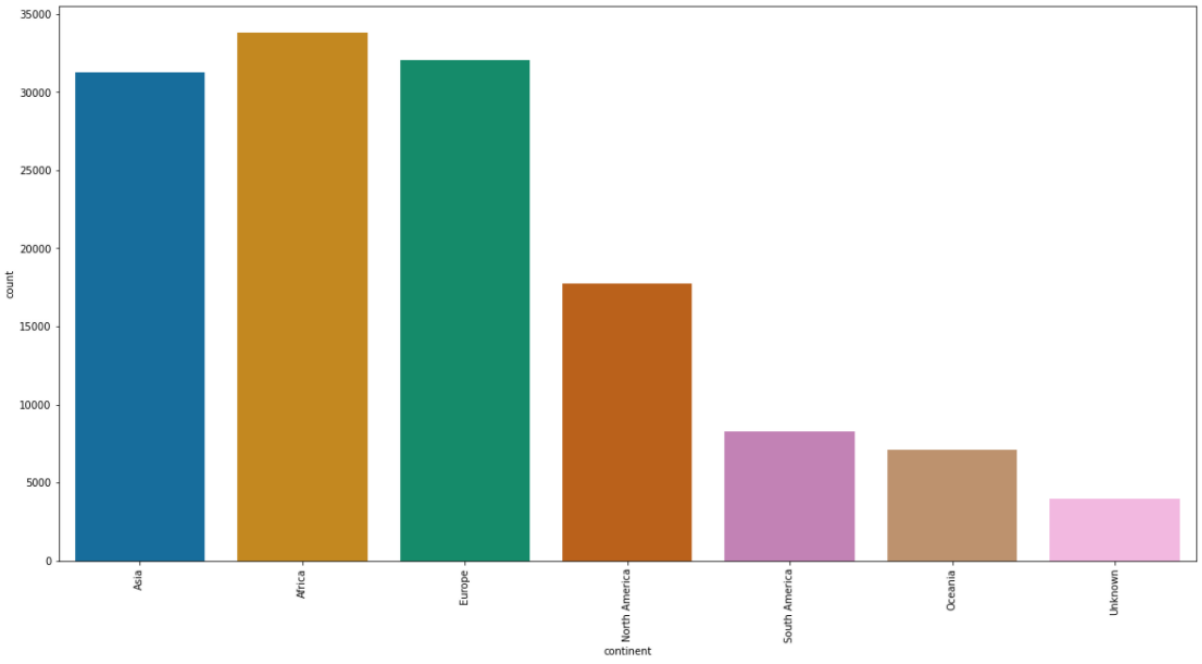


با بررسی داده ها با استفاده متوجه شدیم که در تعدادی از آنها، قاره مشخص نشده است (مقدار null دارد).

برای آنکه ببینیم این مشکل برطرف می شود یا خیر، با استفاده از countplot، location داده هایی که مقدار آنها در ستون قاره برابر null بود را رسم کردیم.



می بینیم که با استفاده از location برخی از داده ها، می توانیم مقدار ستون قاره را در آنها پر کنیم. برای پر کردن این ستون در سایر داده ها نیز از برچسب Unknown استفاده کردیم.

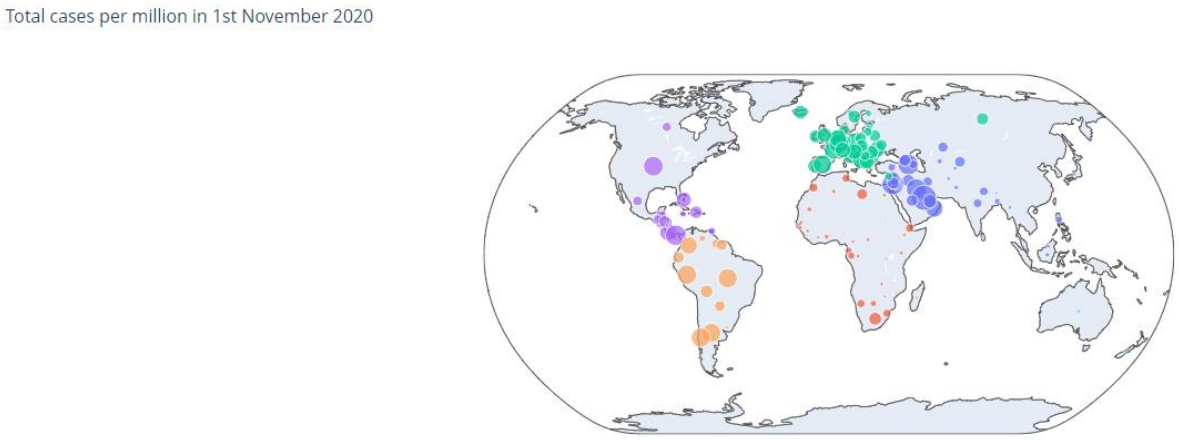


نکته مهمی که در اینجا مطرح است، این است که برای پر کردن مقادیر null در داده ها باید به صورت جزئی تر اقدام کنیم. به این مفهوم که به عنوان مثال در هر کشور جداگانه این عمل را انجام دهیم و دیتا null در یک کشور را با استفاده از سایر دیتایی که از آن کشور داریم، پر کنیم. برای مثال یک نمونه از این روش پر کردن داده های null را می توان در کد مشاهده کرد که مربوط به پر کردن ستون extreme_poverty با استفاده از میانگین این ستون در هر کشور می باشد. (البته در اینجا چون مقادیر null از کشورهایی تشکیل شده است که مقدار این ستون را گزارش نکرده اند، تغییری در تعداد null های دیتا صورت نمی گیرد.)

در گام بعد با location هایی که جمعیت خود را گزارش نکرده اند، پیدا کردیم. دو location بدست آمد: Northern Cyprus و International. به منظور پر کردن ستون جمعیت در قبرس شمالی (Northern Cyprus)، فایل csv جمعیت را از سایت United Nations دانلود کردیم اما این کشور در لیست کشورها نبود! از قرار معلوم تنها کشوری که آن را به رسمیت می شناسد، ترکیه است. به همین جهت به به صورت دستی جمعیت آن را در سال ۲۰۲۰ و ۲۰۲۱ پر کردیم.

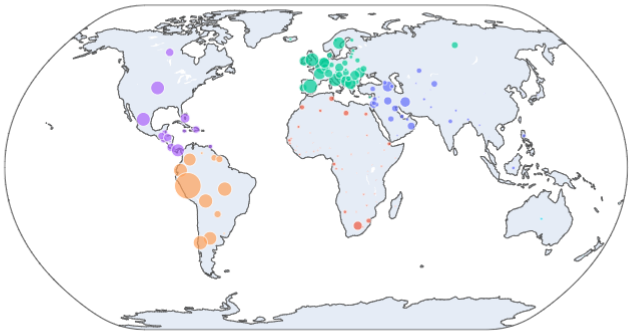
در گام بعد، داده های روز اول ماه ۱۱ میلادی در دو سال ۲۰۲۰ و ۲۰۲۱ را جدا کردیم تا با استفاده از این دو دیتاست جدید، میزان مرگ و میر و تعداد افراد جدید مبتلا شده در مقیاس میلیون را در نقشه رسم کنیم. میزان بزرگی هر یک از این داده ها نیز با استفاده از bubble ها در نقشه نمایش داده می شود. هر قاره نیز به تفکیک رنگ مشخص می باشد.

حجم مورد های جدید مبتلا شده به کرونا در مقیاس میلیون در هر قاره و کشور در تاریخ مذکور در سال ۲۰۲۰.



حجم مورد های فوت شده در مقیاس میلیون در هر قاره و کشور در تاریخ مذکور در سال ۲۰۲۰.

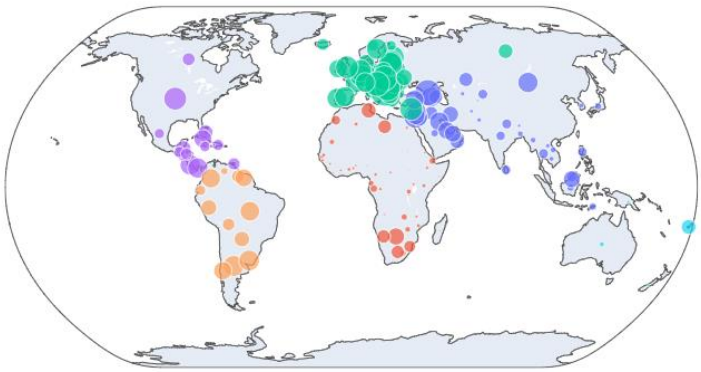
Total deaths per million in 1st November 2020



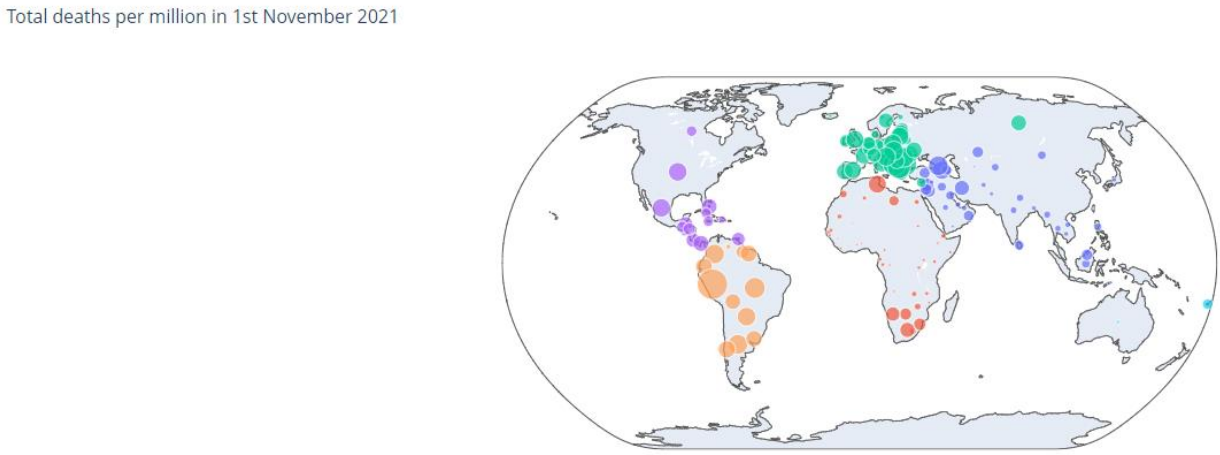
با مقایسه دو شکل برای مثال به صورت شهودی می توانیم ببینیم که در این روز خاص، حجم افراد مبتلا شده ی جدید در سه کشور ایران و عراق و عربستان و کشورهای غرب اروپا زیاد و میزان مرگ و میر در این سه کشور نسبت به کشورهای غرب اروپا، کمتر است.

حجم مورد های جدید مبتلا شده به کرونا در مقیاس میلیون در هر قاره و کشور در تاریخ مذکور در سال ۲۰۲۱.

Total cases per million in 1st November 2021



حجم مورد های فوت شده در مقیاس میلیون در هر قاره و کشور در تاریخ مذکور در سال ۲۰۲۰.



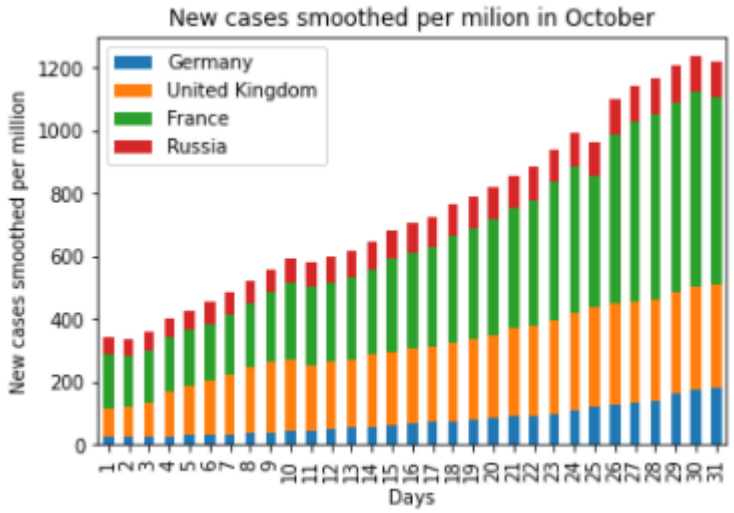
با مقایسه نقشه میزان مرگ و میر در مقیاس میلیون در یک روز خاص و در دو سال متوالی می توان دید که حجم این مرگ و میرها در سال ۲۰۲۱ به وضوح افزایش پیدا کرده است.

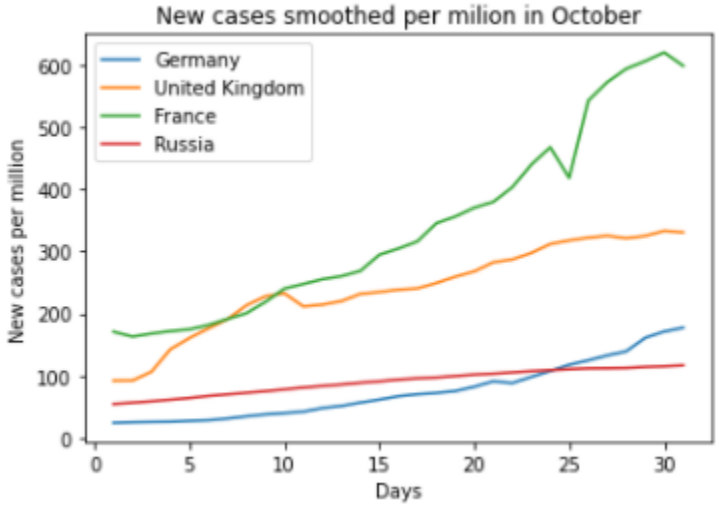
دو انیمیشن از تغییر حجم افراد مبتلا شده جدید در مقیاس میلیون در ماه های دو سال ۲۰۲۰ و ۲۰۲۱ (هر سال به صورت جداگانه) در کد قابل مشاهده می باشند. در این دو انیمیشن، هر bubble حجم افراد مبتلا شده جدید در مقیاس میلیون در اولین روز از هر ماه را نمایش می دهد. میتوانیم تغییرات حجم را با توجه به تغییر اندازه هر bubble مشاهده کنیم. (لازم به ذکر است که شروع دریافت داده های مربوط به ویروس کرونا، از ماه سوم سال ۲۰۲۰ میلادی می باشد.)

در گام بعد، داده های چهار کشور آلمان، انگلیس، فرانسه و روسیه را در ماه ۱۱ سال ۲۰۲۰ به دلخواه جدا کردیم. ابتدا با استفاده از stacked barplot و سپس با استفاده از timeseries تعداد افراد مبتلا شده جدید smooth شده در مقیاس میلیون را در این ماه رسم کرده ایم.

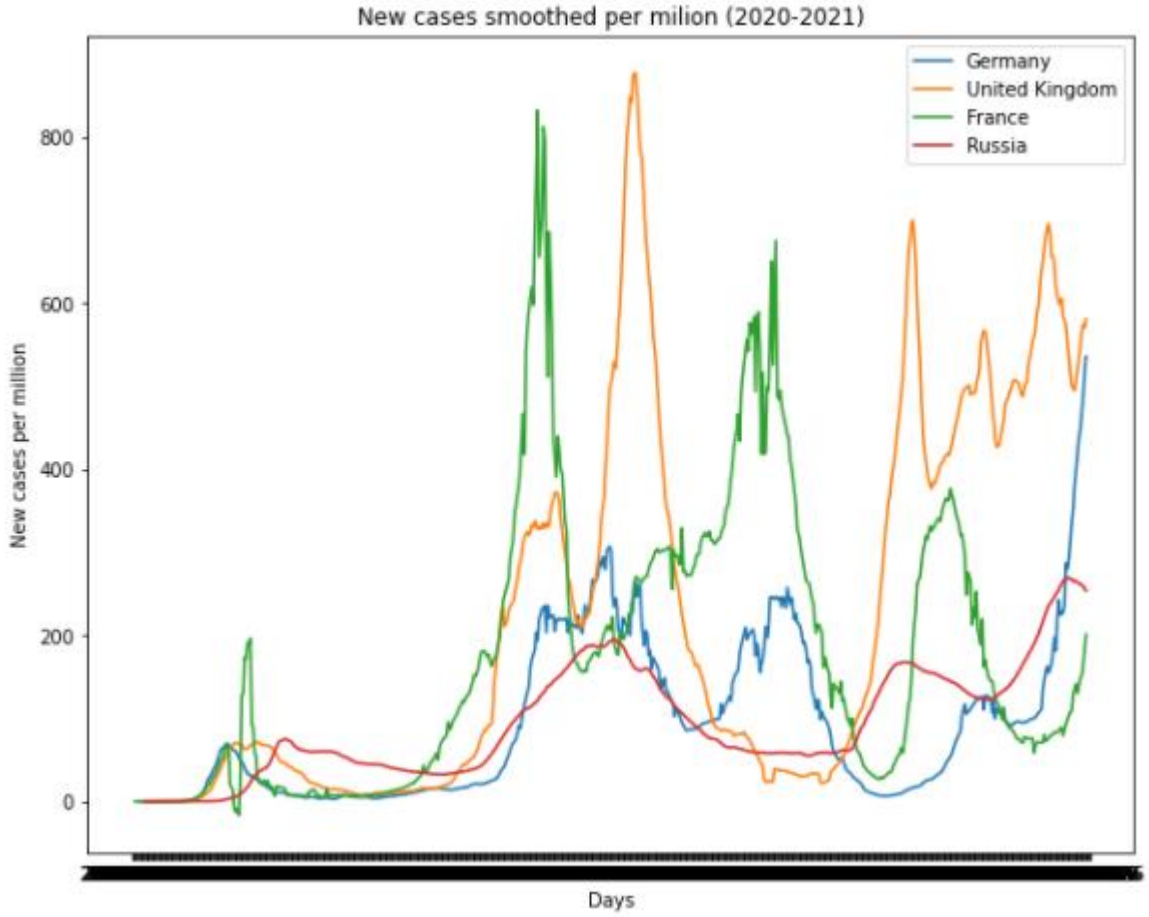
در نمودار stacked barplot، اندازه هر قسمت رنگی هر bar، حجم افراد مذکور (در هر کشور) در آن روز خاص را مشخص می کند. اما در timeseries، حجم این افراد (در هر کشور) در هر روز با یک نقطه مشخص می شود و در نهایت این نقطه ها به یکدیگر متصل می شوند و نمودار را تشکیل می دهند.

در نمودار دوم بهتر مشاهده می شود که شیب تعداد افراد مبتلا شده جدید smooth شده (در مقیاس میلیون) در این ماه در فرانسه بیشتر از سه کشور دیگر است.


New cases smoothed per milion in October
Germany
United Kingdom
France
Russia
New cases smoothed per million
Days

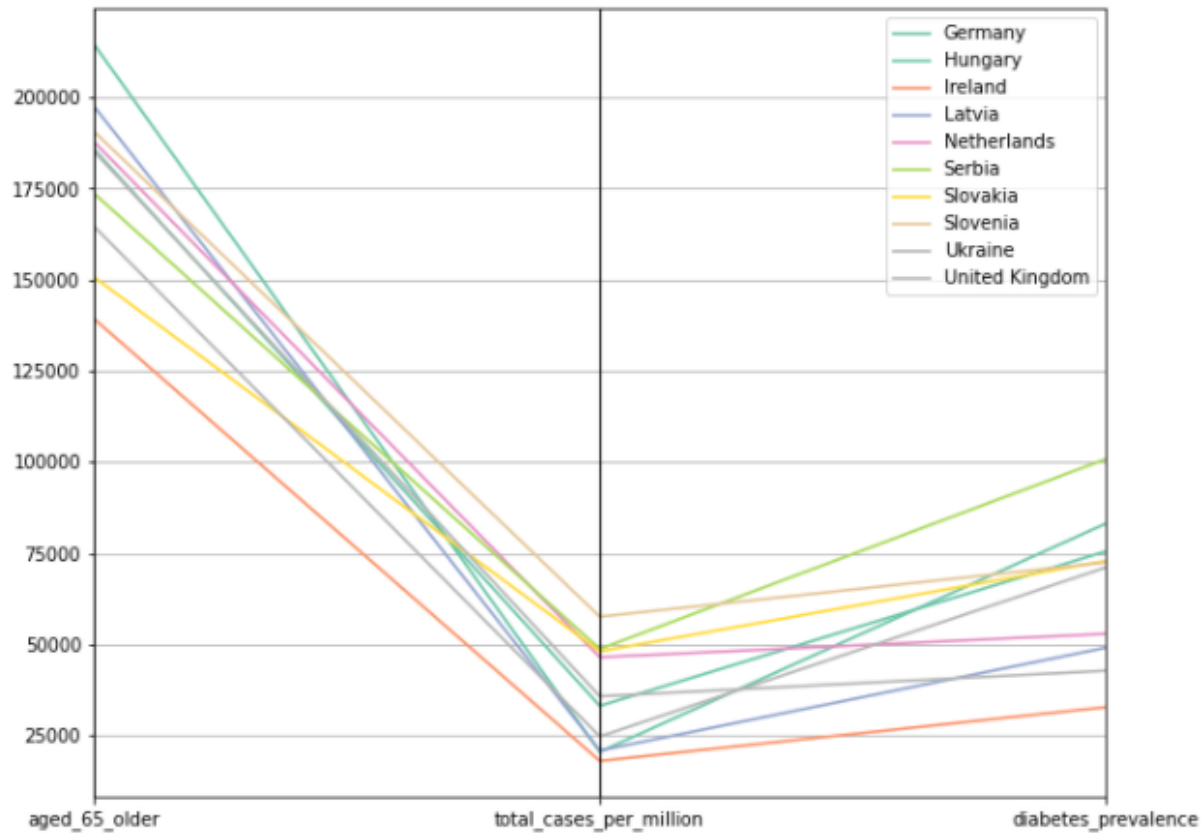

New cases smoothed per milion in October
Germany
United Kingdom
France
Russia
New cases per million
Days

در نمودار بعدی، تعداد افراد مبتلا شده جدید smooth شده (در مقیاس میلیون) در این چهار کشور را در دو سال ۲۰۲۰ و ۲۰۲۱ بررسی می کنیم. می بینیم که پیک های تشکیل شده در نمودار دو کشور انگلیس و فرانسه نسبت به دو کشور دیگر، شدید تر می باشند.



در گام بعد، ۱۰ کشور را به صورت تصادفی از قاره اروپا انتخاب می کنیم و از آنها، داده های مربوط آخرین ماه سال ۲۰۲۰ را انتخاب می کنیم. با استفاده از نمودار parallel_coordinates، ستون های aged_65_older و total_cases_per_million و diabetes_prevalence این ۱۰ کشور در تاریخ مذکور را روی نمودار می آوریم.

به طور کلی مشاهده می شود که هر چه تعداد کل افراد مبتلا شده در مقیاس میلیون بیشتر باشد، تعداد افراد مبتلا شده با سن ۶۵ سال یا بیشتر نیز بیشتر است. با بررسی طرف دیگر نمودار نیز به طور کلی می بینیم هر چه تعداد کل افراد مبتلا شده در مقیاس میلیون بیشتر است، حجم افراد دیابتی نیز در آن جامعه بیشتر است.



در مرحله بعد ۱۰ کشور جدید از قاره اروپا انتخاب می کنیم. میانگین positive_rate (میانگین هر ۷ روز می باشد.) را در هر یک از این کشورها محاسبه می کنیم. تراکم جمعیت مربوط به هر یک از این کشور ها نیز مقدار یکتایی است. با استفاده از scatterplot ی خواهیم رابطه بین تراکم جمعیت و نرخ تست ها مثبت کرونا را بررسی کنیم. طبق نمودار می بینیم که هر چه تراکم جمعیت در یک ناحیه بیشتر است، به صورت کلی positive_rate در آن ناحیه نیز بیشتر است.