آریا وارسته‌نژاد
۸۱۰۱۰۰۴۹۸

۱۴۰۰/۰۱/۱۵

# داده‌کاوی
# تمرین اول

a.varaste.n@gmail.com

## مقدمه

- فایل‌های این تمرین عبارت اند از:
  ۱) نوت‌بوک پایتون
  ۲) فایل گزارش
- در فایل نوت‌بوک سعی شده که تمام اصول خوانایی کد رعایت شود و قسمت‌های تمرین به صورت مشخص از هم جدا و قابل تفکیک باشند. هر بخش و زیر بخش مطابق با صورت تمرین و گزارش علامت گزاری شده است.
- نمونه: 2-6
- در گزارش تلاش شده که تا حد امکان کدهای غیر ضروری آورده نشوند.
- در نوت‌بوک بیشتر ران‌های نهایی و خروجی‌های حاصله به صورت ذخیره‌شده موجود هستند.

# بخش اول: پیش‌پردازش

## پیش‌پردازش: ۱

تعداد داده‌های گم شده در هر ویژگی را مشخص کنید. سپس، با ذکر دلیل، رویکرد مورد استفاده خود را برای پر کردن داده‌های گم شده در هر ستون مشخص کرده و اقدام به تکمیل داده‌های گم شده کنید.

ابتدا بررسی می‌کنیم که کدام ستون‌ها داده‌های گم شده دارند.

```
iso_code                                  False
continent                                  True
location                                  False
date                                      False
total_cases                                True
                                           ...
human_development_index                    True
excess_mortality_cumulative_absolute       True
excess_mortality_cumulative                True
excess_mortality                           True
excess_mortality_cumulative_per_million    True
Length: 67, dtype: bool
```

مشخص می‌شود که همه ستون‌ها بجز iso_code، location و date دارای داده‌های گم شده هستند.

برای درک بهتر مجموعه داده بررسی می‌کنیم که کدام ردیف‌ها داده گم شده دارند که مشخص می‌شود همه ردیف‌ها اقلاً در یک ستون داده گم شده دارند.

```
0         True
1         True
2         True
3         True
4         True
          ...
165631    True
165632    True
165633    True
165634    True
165635    True
Length: 165636, d
```

بعد چک می‌کنیم که تقریبی از تعداد موردهای خالی برای هر ردیف را داشته باشیم:

```
0         46
1         46
2         46
3         46
4         46
          ..
165631    14
165632    15
165633    22
165634    23
165635    23
Length: 165636,
```

همانطور که مشاهده می‌شود تعداد زیادی از ۶۷ ستون برای هر سطر خالی هستند.

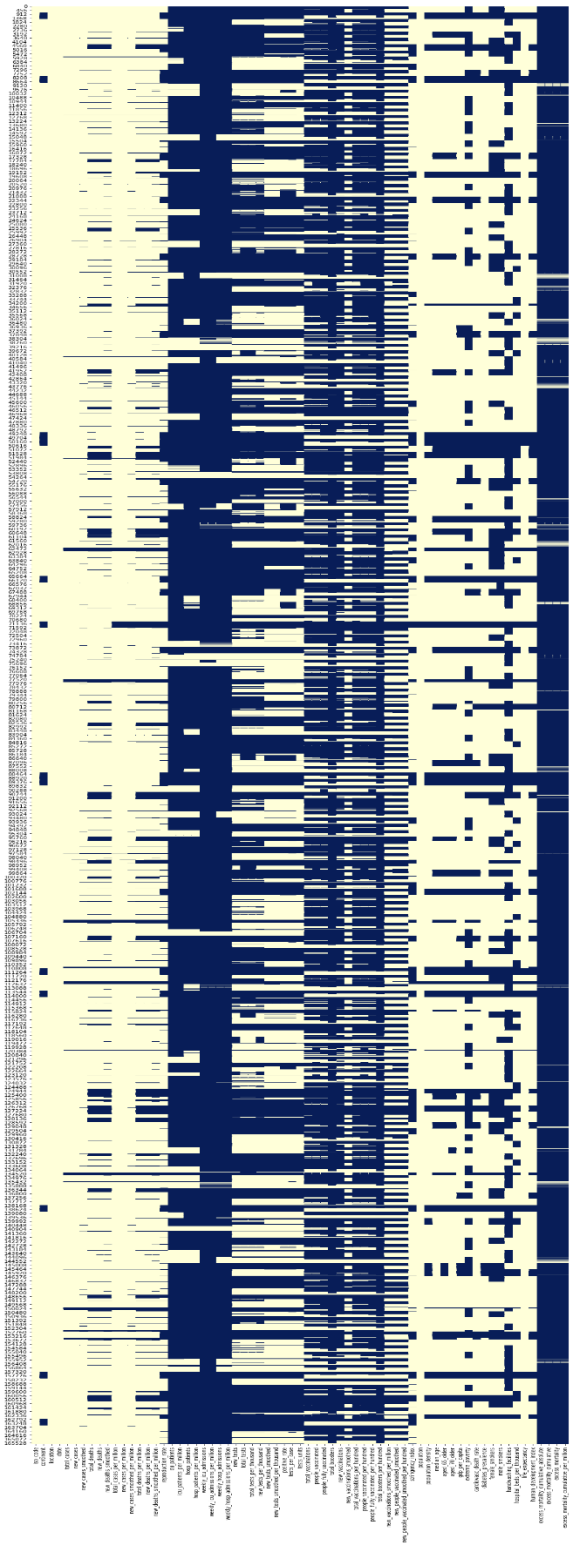
سپس تعداد داده‌های گم شده برای هر ستون را بدست می‌آوریم:

```
iso_code                                        0
continent                                    9917
location                                        0
date                                            0
total_cases                                  3030
new_cases                                    3172
new_cases_smoothed                           5156
total_deaths                                20843
new_deaths                                  20803
new_deaths_smoothed                         22902
total_cases_per_million                      3785
new_cases_per_million                        3927
new_cases_smoothed_per_million               5905
total_deaths_per_million                    21585
new_deaths_per_million                      21545
new_deaths_smoothed_per_million             23638
reproduction_rate                           40569
icu_patients                               142246
icu_patients_per_million                   142246
hosp_patients                              141072
hosp_patients_per_million                  141072
weekly_icu_admissions                      160232
weekly_icu_admissions_per_million          160232
weekly_hosp_admissions                     154759
weekly_hosp_admissions_per_million         154759
new_tests                                   98630
total_tests                                 96692
total_tests_per_thousand                    96692
new_tests_per_thousand                      98630
new_tests_smoothed                          81978
new_tests_smoothed_per_thousand             81978
positive_rate                               87046
tests_per_case                              87609
tests_units                                 79655
total_vaccinations                         120658
people_vaccinated                          122844
people_fully_vaccinated                    125608
total_boosters                             148296
new_vaccinations                           128384
new_vaccinations_smoothed                   81524
total_vaccinations_per_hundred             120658
people_vaccinated_per_hundred              122844
people_fully_vaccinated_per_hundred        125608
total_boosters_per_hundred                 148296
new_vaccinations_smoothed_per_million       81524
new_people_vaccinated_smoothed              82815
new_people_vaccinated_smoothed_per_hundred  82815
stringency_index                            35774
population                                   1072
population_density                          18323
median_age                                  28378
aged_65_older                               29866
aged_70_older                               29114
gdp_per_capita                              27708
extreme_poverty                             74799
cardiovasc_death_rate                       29428
diabetes_prevalence                         22287
female_smokers                              60027
male_smokers                                61476
handwashing_facilities                      97352
hospital_beds_per_thousand                  42485
life_expectancy                             11016
human_development_index                     29953
excess_mortality_cumulative_absolute       159940
excess_mortality_cumulative                159940
excess_mortality                           159940
excess_mortality_cumulative_per_million    159940
dtype: int64
```

در شکل زیر شمای کلی مجموعه داده پیش از تکمیل داده‌های گم شده قابل مشاهده است:

```
In [247]:
1  plt.figure(figsize=(16,67))
2  sns.heatmap(miss_filled_df.isnull(), cbar=False, cmap="YlGnBu")
3  plt.show()
```

executed in 30.5s, finished 13:30:09 2022-04-07



در تصویر بالا قسمت‌هایی که با رنگ سرمه‌ای مشخص شده‌اند داده‌های گم شده و بخش‌هایی که با رنگ زرد روشن رنگ آمیزی شده‌اند دارای داده هستند.

بعد از درک بهتر جوانب و مشخص شدن ستون‌های دارای گم شدگی و تعداد داده‌های گم شده باید برای پرکردن این داده‌ها برای هر ستون تصمیم‌گیری شود. به همین جهت ابتدا روش‌های مختلف موجود برای تکمیل داده‌های گم شده را بررسی و مطالعه کردیم. روش‌های مناسب عبارت بودند از:

| |
|---|
| a single constant value |
| previous row |
| next row |
| median |
| mean |
| Linear Interpolation method |
| most frequent |

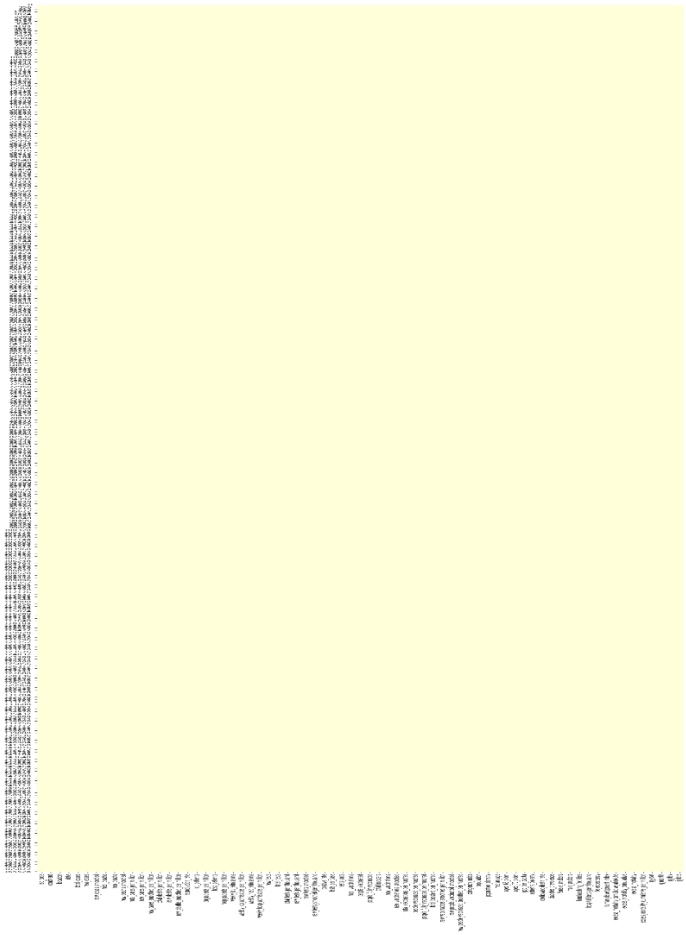
در جدول زیر مورد به مورد مشخص کردیم که داده‌های گم شده هر ستون را با چه روشی تکمیل کنیم:

| Title | Policy |
|---|---|
| iso_code 0 | |
| location 0 | |
| date 0 | |
| population 1072 | دستی |
| total_cases 3030 | Previous Row |
| new_cases 3172 | Linear Interpolation Method |
| total_cases_per_million 3785 | Previous Row |
| new_cases_per_million 3927 | Linear Interpolation Method |
| new_cases_smoothed 5156 | Mean |
| new_cases_smoothed_per_million 5905 | Mean |
| continent 9917 | A Single Constant Value |
| life_expectancy 11016 | Mean |
| population_density 18323 | A Single Constant Value |
| new_deaths 20803 | Linear Interpolation Method |
| total_deaths 20843 | Previous Row |
| new_deaths_per_million 21545 | Linear Interpolation Method |
| total_deaths_per_million 21585 | Previous Row |
| diabetes_prevalence 22287 | Mean |
| new_deaths_smoothed 22902 | Mean |

| | |
|---|---|
| new_deaths_smoothed_per_million 23638 | Mean |
| gdp_per_capita 27708 | Mean |
| median_age 28378 | Mean |
| aged_70_older 29114 | Mean |
| cardiovasc_death_rate 29428 | Mean |
| aged_65_older 29866 | Mean |
| human_development_index 29953 | Mean |
| stringency_index 35774 | Mean |
| reproduction_rate 40569 | Mean |
| hospital_beds_per_thousand 42485 | Mean |
| female_smokers 60027 | Mean |
| male_smokers 61476 | Mean |
| extreme_poverty 74799 | Mean |
| tests_units 79655 | Linear Interpolation Method |
| new_vaccinations_smoothed_per_million 81524 | Linear Interpolation Method |
| new_vaccinations_smoothed 81524 | Linear Interpolation Method |
| new_tests_smoothed 81978 | Linear Interpolation Method |
| new_tests_smoothed_per_thousand 81978 | Linear Interpolation Method |
| new_people_vaccinated_smoothed_per_hundred 82815 | Linear Interpolation Method |
| new_people_vaccinated_smoothed 82815 | Linear Interpolation Method |
| positive_rate 87046 | Linear Interpolation Method |
| tests_per_case 87609 | Linear Interpolation Method |
| total_tests_per_thousand 96692 | Previous Row |
| total_tests 96692 | Previous Row |
| handwashing_facilities 97352 | Mean |
| new_tests 98630 | Linear Interpolation Method |
| new_tests_per_thousand 98630 | Mean |
| total_vaccinations 120658 | Previous Row |
| total_vaccinations_per_hundred 120658 | Previous Row |
| people_vaccinated 122844 | Previous Row |
| people_vaccinated_per_hundred 122844 | Previous Row |
| people_fully_vaccinated 125608 | Previous Row |
| people_fully_vaccinated_per_hundred 125608 | Previous Row |
| new_vaccinations 128384 | Linear Interpolation Method |
| hosp_patients_per_million 141072 | Linear Interpolation Method |

| | |
|---|---|
| hosp_patients 141072 | Linear Interpolation Method |
| icu_patients_per_million 142246 | Linear Interpolation Method |
| icu_patients 142246 | Linear Interpolation Method |
| total_boosters 148296 | Previous Row |
| total_boosters_per_hundred 148296 | Previous Row |
| weekly_hosp_admissions_per_million 154759 | Mean |
| weekly_hosp_admissions 154759 | Mean |
| excess_mortality_cumulative 159940 | A Single Constant Value |
| excess_mortality_cumulative_absolute 159940 | A Single Constant Value |
| excess_mortality 159940 | A Single Constant Value |
| excess_mortality_cumulative_per_million 159940 | A Single Constant Value |
| weekly_icu_admissions_per_million 160232 | Linear Interpolation Method |
| weekly_icu_admissions 160232 | Linear Interpolation Method |

پس از انجام پرکردن داده‌های گم شده شمای کلی مجموعه داده به صورت زیر در می‌آید که نشان دهنده این است که دیگر داده گم شده‌ای وجود ندارد:



## پیش‌پردازش: ۲

دیتافریم دیگری درست نمایید که در آن، تعداد کیس‌های جدید، تعداد واکسینه‌های جدید، تعداد فوتی‌ها و جمعیت برای هر کشور به صورت تجمیع شده محاسبه شده باشد. (محاسبه‌ی جمع داده‌ها از ابتدا تا آخرین تاریخ موجود در مجموعه دادهها برای هر کشور)

دیتافریم را **با جمع کردن مقادیر جدید** به شکل زیر تشکیل دادیم:

In [666]:
```
pd.set_option("display.max_rows", 300, "display.max_columns", 200)
grouped_single = df.groupby('location').agg({'new_cases': ['sum'],
                                             'new_vaccinations': ['sum'],
                                             'new_deaths': ['sum'],
                                             'population': ['max'],
                                             })
grouped_single
```

| | new_cases | new_vaccinations | new_deaths | population |
|---|---|---|---|---|
| | sum | sum | sum | max |
| location | | | | |
| Afghanistan | 174103.0 | 2.974970e+06 | 7645.0 | 3.983543e+07 |
| Africa | 11230524.0 | 5.818190e+08 | 248668.0 | 1.373486e+09 |
| Albania | 271851.0 | 3.002796e+06 | 3489.0 | 2.872934e+06 |
| Algeria | 265079.0 | 8.412034e+06 | 6859.0 | 4.461663e+07 |
| Andorra | 38249.0 | 1.605815e+06 | 171.0 | 7.735400e+04 |
| Angola | 98746.0 | 1.013173e+06 | 1925.0 | 3.393361e+07 |
| Anguilla | 2555.0 | 5.461290e+05 | 552.0 | 1.512500e+04 |
| Antigua and Barbuda | 7457.0 | 3.479066e+06 | 160.0 | 9.872800e+04 |
| Argentina | 8912379.0 | 1.016397e+08 | 126457.0 | 4.560582e+07 |
| Armenia | 420525.0 | 1.908885e+06 | 8551.0 | 2.968128e+06 |
| Aruba | 33684.0 | 1.141134e+06 | 245.0 | 1.071950e+05 |
| Asia | 117811418.0 | 7.544342e+09 | 1351250.0 | 4.678445e+09 |
| Australia | 3389089.0 | 5.549652e+07 | 5357.0 | 2.578822e+07 |
| Austria | 2744023.0 | 1.263144e+07 | 14929.0 | 9.043072e+06 |
| Azerbaijan | 787367.0 | 1.829095e+07 | 9466.0 | 1.022334e+07 |
| Bahamas | 33206.0 | 2.118552e+06 | 787.0 | 3.969140e+05 |
| Bahrain | 519584.0 | 6.483987e+06 | 1534.0 | 1.748295e+06 |
| Bangladesh | 1945123.0 | 3.283306e+08 | 29068.0 | 1.663035e+08 |
| Barbados | 55543.0 | 1.048869e+06 | 335.0 | 2.877080e+05 |
| Belarus | 923432.0 | 2.936000e+04 | 6538.0 | 9.442867e+06 |
| Belgium | 3563842.0 | 2.505387e+07 | 30430.0 | 1.163233e+07 |
| Belize | 56816.0 | 1.386235e+07 | 670.0 | 4.049150e+05 |
| Benin | 26776.0 | 6.162543e+07 | 184.0 | 1.245103e+07 |
| Bermuda | 11561.0 | 6.136759e+07 | 159.0 | 6.209200e+04 |
| Bhutan | 13535.0 | 3.959222e+07 | 314.0 | 7.799000e+05 |
| Bolivia | 893775.0 | 1.331590e+07 | 21461.0 | 1.183294e+07 |
| Bonaire Sint Eustatius and Saba | 7599.0 | 1.888460e+07 | 200.0 | 2.644500e+04 |
| Bosnia and Herzegovina | 371553.0 | 1.815734e+07 | 15604.0 | 3.263459e+06 |
| Botswana | 263955.0 | 5.117875e+06 | 2620.0 | 2.397240e+06 |
| Brazil | 28778780.0 | 3.983332e+08 | 650274.0 | 2.139934e+08 |
| British Virgin Islands | 6085.0 | 5.315700e+05 | 84.0 | 3.042300e+04 |
| Brunei | 71727.0 | 1.110081e+06 | 148.0 | 4.415320e+05 |

## پیش‌پردازش: ۳

ستون جدیدی با اسم تاریخ شمسی ایجاد کنید و برای ایجاد آن، تاریخ میلادی را به شمسی تبدیل نمایید.

برای انجام این‌کار ابتدا تاریخ میلادی را به سه بخش روز، ماه و سال شکسته و سپس از این سه بخش تاریخ شمسی را تولید می‌کنیم.

In [558]:
```
iran_date_df = df
iran_date_df ["g_year"] = pd.DatetimeIndex(iran_date_df["date"]).year
iran_date_df ["g_month"] = pd.DatetimeIndex(iran_date_df["date"]).month
iran_date_df ["g_day"] = pd.DatetimeIndex(iran_date_df["date"]).day
bb = []
for i in range (165636):
    ax = JalaliDate(datetime.date(iran_date_df ["g_year"][i], iran_date_df ["g_month"][i], iran_date_df
    bb.append(ax)
# print(bb)
iran_date_df ["j_date"] = bb
iran_date_df.head()
```

| | date |
|---|---|
| an | 2020-02-24 |
| an | 2020-02-25 |
| an | 2020-02-26 |
| an | 2020-02-27 |
| an | 2020-02-28 |

| g_year | g_month | g_day | j_date |
|---|---|---|---|
| 2020 | 2 | 24 | 1398-12-05 |
| 2020 | 2 | 25 | 1398-12-06 |
| 2020 | 2 | 26 | 1398-12-07 |
| 2020 | 2 | 27 | 1398-12-08 |
| 2020 | 2 | 28 | 1398-12-09 |

## پیش‌پردازش: ۴

با توجه به تعداد بالای ویژگی‌ها، آیا می‌توان از تعداد ویژگی‌ها کاست؟

بله. ویژگی‌هایی را که بتوان با ویژگی‌های دیگر بدست آورد و یا بازگوکننده یک مفهوم باشند را می‌توان برای کاهش ابعاد حذف کرد.

مثلاً

- `icu_patients`
- `hosp_patients`
- `weekly_icu_admissions`
- `weekly_hosp_admissions`

و

- `total_vaccinations`
- `people_vaccinated`
- `new_vaccinations`

یا مثلاً برخی از ویژگی‌ها از تقسیم یک ویژگی بر جمعیت محاسبه شده‌اند و لذا می‌توان آن‌ها را کنار گذاشت و در صورت نیاز محاسبه کرد.

## پیش‌پردازش: ۵

دیتافریم جدیدی درست نمایید که در آن صرفاً اطلاعات مربوط به کشور ایران قرار داده شده باشد.

```
In [194]:
1 df_Iran = iran_date_df[iran_date_df.location=='Iran']
2 df_Iran
```

executed in 212ms, finished 19:27:28 2022-04-06

| | iso_code | continent | location | date | total_cases | new_cases | new_cases_smoothed | total_deaths | new_deaths |
|---|---|---|---|---|---|---|---|---|---|
| 71639 | IRN | Asia | Iran | 2020-02-19 | 2.0 | 2.0 | NaN | 2.0 | 2.0 |
| 71640 | IRN | Asia | Iran | 2020-02-20 | 5.0 | 3.0 | NaN | 2.0 | 0.0 |
| 71641 | IRN | Asia | Iran | 2020-02-21 | 18.0 | 13.0 | NaN | 4.0 | 2.0 |
| 71642 | IRN | Asia | Iran | 2020-02-22 | 28.0 | 10.0 | NaN | 5.0 | 1.0 |
| 71643 | IRN | Asia | Iran | 2020-02-23 | 43.0 | 15.0 | NaN | 8.0 | 3.0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 72377 | IRN | Asia | Iran | 2022-02-26 | 7030943.0 | 7039.0 | 15065.429 | 136390.0 | 224.0 |
| 72378 | IRN | Asia | Iran | 2022-02-27 | 7040467.0 | 9524.0 | 14002.143 | 136631.0 | 241.0 |
| 72379 | IRN | Asia | Iran | 2022-02-28 | 7051429.0 | 10962.0 | 12838.143 | 136838.0 | 207.0 |
| 72380 | IRN | Asia | Iran | 2022-03-01 | 7060741.0 | 9312.0 | 11015.143 | 137064.0 | 226.0 |
| 72381 | IRN | Asia | Iran | 2022-03-02 | 7066975.0 | 6234.0 | 9714.286 | 137267.0 | 203.0 |

743 rows × 71 columns

## پیش‌پردازش: ۶

در دیتافریم ایران، ستونی ایجاد نمایید که در آن، ماه به عنوان یک ویژگی مستقل در نظر گرفته شده‌است.

در قسمت‌های قبل انجام شده است. (سوال ۳ بخش پیش‌پردازش)

## پیش‌پردازش: ۷

دیتافریم جدیدی ایجاد نمایید که مجموعه‌داده ایران را بر اساس ماه در سال ۲۰۲۱ تجمیع کند.

1-7

In [198]:

```
iran_agg_month = df_Iran
iran_agg_month.groupby("g_month").agg({'new_cases': ['sum'],
                                       'new_vaccinations': ['sum'],
                                       'new_deaths': ['sum'],
                                       'population': ['max'],
                                       })
iran_agg_month
```

executed in 133ms, finished 20:13:38 2022-04-06

| | iso_code | continent | location | date | total_cases | new_cases | new_cases_smoothed | total_deaths | new_deaths |
|---|---|---|---|---|---|---|---|---|---|
| 71639 | IRN | Asia | Iran | 2020-02-19 | 2.0 | 2.0 | NaN | 2.0 | 2.0 |
| 71640 | IRN | Asia | Iran | 2020-02-20 | 5.0 | 3.0 | NaN | 2.0 | 0.0 |
| 71641 | IRN | Asia | Iran | 2020-02-21 | 18.0 | 13.0 | NaN | 4.0 | 2.0 |
| 71642 | IRN | Asia | Iran | 2020-02-22 | 28.0 | 10.0 | NaN | 5.0 | 1.0 |
| 71643 | IRN | Asia | Iran | 2020-02-23 | 43.0 | 15.0 | NaN | 8.0 | 3.0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 72377 | IRN | Asia | Iran | 2022-02-26 | 7030943.0 | 7039.0 | 15065.429 | 136390.0 | 224.0 |
| 72378 | IRN | Asia | Iran | 2022-02-27 | 7040467.0 | 9524.0 | 14002.143 | 136631.0 | 241.0 |
| 72379 | IRN | Asia | Iran | 2022-02-28 | 7051429.0 | 10962.0 | 12838.143 | 136838.0 | 207.0 |
| 72380 | IRN | Asia | Iran | 2022-03-01 | 7060741.0 | 9312.0 | 11015.143 | 137064.0 | 226.0 |
| 72381 | IRN | Asia | Iran | 2022-03-02 | 7066975.0 | 6234.0 | 9714.286 | 137267.0 | 203.0 |

743 rows × 71 columns

# بخش دوم: نمایش دادگان

## نمایش دادگان: ۱

کدام کشورها بهترین و کدام کشورها بدترین عملکرد در مهار ویروس کرونا را داشته‌اند؟ با یک نمودار مناسب این مساله را بررسی نمایید و برداشت خود را از نتایج ذکر نمایید. (منظور از عملکرد، تعداد فوتی نسبت به کل جمعیت است.)

بر اساس نموداری که در صفحه بعد رسم شده است کشورهای:

۱) Niue
۲) Tuvalu
۳) Pitcairn
۴) Nauru
۵) Macao
۶) Vatican
۷) Guernsey
۸) Jersey
۹) Sint Maarten (Dutch part)
۱۰) Cook Islands

بهترین عملکرد را داشته و کشورهای:

۱) Peru
۲) Bulgaria
۳) Bosnia and Herzegovina
۴) Hungary
۵) North Macedonia
۶) Montenegro
۷) Georgia
۸) Croatia
۹) Czechia
۱۰) Slovakia

بدترین عملکرد را داشته اند.

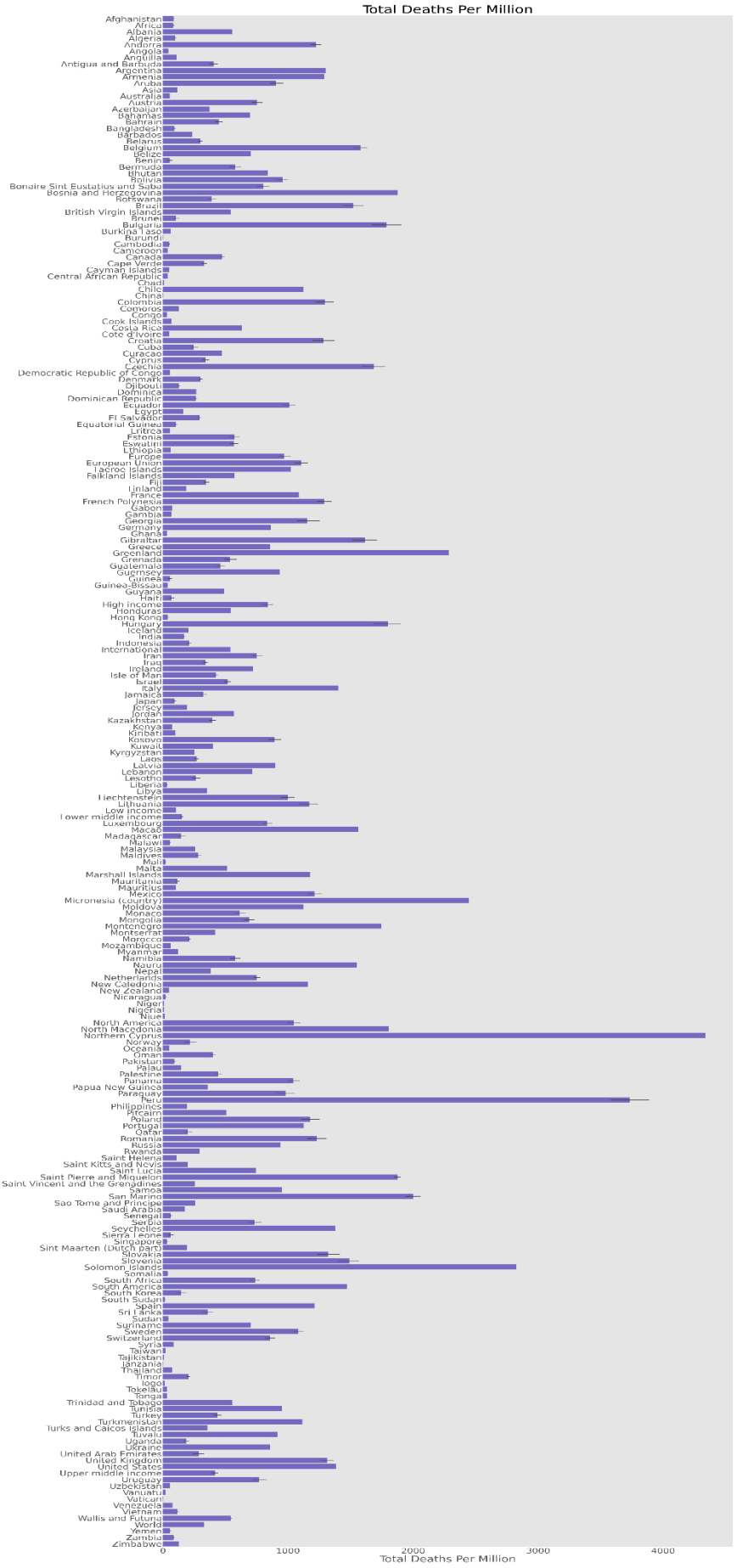
خیلی صادقانه بخوام برداشت خودم رو بگم باید عرض کنم که کشورهایی که در گروه بدترین‌ها هستن، کشورهایی هستن که خیلی وضع خوبی ندارن اما خیلی هم داغون نیستن و از طرف دیگه آمار رو سانسور و مهندسی هم نکردن. همونطور که می‌بینیم بیشتر این کشورها از اروپای شرقی هستن به علاوه پرو از آمریکای جنوبی.

قطعاً خیلی کشورها هستن که آسیب بیشتری از کرونا دیدن و نسبت به این کشورها بدتر عمل کردن که آمار خودشون رو درست ندادن مثل کشورهایی که بهتره اسم نیاریم! یا اصلاً دیگه وضعشون خیلی خراب هست به حدی که توانایی گردآوری آمار نداشتن مثل کشورهای فقیر یا آفریقایی.

In [485]:

```
sorted_df = test_df

fig, ax = plt.subplots(figsize=(60,200))
fig.tight_layout(pad=20)
plt.style.use('ggplot')
# Creating a case-specific function to avoid code repetition
def plot_hor_vs_vert(subplot, x, y, xlabel, ylabel, rotation,
                     tick_bottom, tick_left):
    ax=plt.subplot(1,1,subplot)
    sns.barplot(x, y, data=sorted_df, color='slateblue')
    plt.title('Total Deaths Per Million', fontsize=100)
    plt.xlabel(xlabel, fontsize=80)
    plt.xticks(fontsize=65, rotation=rotation)
    plt.ylabel(ylabel, fontsize=80)
    plt.yticks(fontsize=65)
    sns.despine(bottom=False, left=True)
    ax.grid(False)
    ax.tick_params(bottom=tick_bottom, left=tick_left)
    width = 0.35
    return None

plot_hor_vs_vert(1, x='total_deaths_per_million', y='location',
                 xlabel='Total Deaths Per Million', ylabel=None,
                 rotation=None, tick_bottom=True, tick_left=False)
plt.show()
```

executed in 28.8s, finished 10:29:10 2022-04-08


Total Deaths Per Million
Afghanistan
Africa
Albania
Algeria
Andorra
Angola
Anguilla
Antigua and Barbuda
Argentina
Armenia
Aruba
Asia
Australia
Austria
Azerbaijan
Bahamas
Bahrain
Bangladesh
Barbados
Belarus
Belgium
Belize
Benin
Bermuda
Bhutan
Bolivia
Bonaire Sint Eustatius and Saba
Bosnia and Herzegovina
Botswana
Brazil
British Virgin Islands
Brunei
Bulgaria
Burkina Faso
Burundi
Cambodia
Cameroon
Canada
Cape Verde
Cayman Islands
Central African Republic
Chad
Chile
China
Colombia
Comoros
Congo
Cook Islands
Costa Rica
Cote d'Ivoire
Croatia
Cuba
Curacao
Cyprus
Czechia
Democratic Republic of Congo
Denmark
Djibouti
Dominica
Dominican Republic
Ecuador
Egypt
El Salvador
Equatorial Guinea
Eritrea
Estonia
Eswatini
Ethiopia
Europe
European Union
Faeroe Islands
Falkland Islands
Fiji
Finland
France
French Polynesia
Gabon
Gambia
Georgia
Germany
Ghana
Gibraltar
Greece
Greenland
Grenada
Guatemala
Guernsey
Guinea
Guinea-Bissau
Guyana
Haiti
High income
Honduras
Hong Kong
Hungary
Iceland
India
Indonesia
International
Iran
Iraq
Ireland
Isle of Man
Israel
Italy
Jamaica
Japan
Jersey
Jordan
Kazakhstan
Kenya
Kiribati
Kosovo
Kuwait
Kyrgyzstan
Laos
Latvia
Lebanon
Lesotho
Liberia
Libya
Liechtenstein
Lithuania
Low income
Lower middle income
Luxembourg
Macao
Madagascar
Malawi
Malaysia
Maldives
Mali
Malta
Marshall Islands
Mauritania
Mauritius
Mexico
Micronesia (country)
Moldova
Monaco
Mongolia
Montenegro
Montserrat
Morocco
Mozambique
Myanmar
Namibia
Nauru
Nepal
Netherlands
New Caledonia
New Zealand
Nicaragua
Niger
Nigeria
Niue
North America
North Macedonia
Northern Cyprus
Norway
Oceania
Oman
Pakistan
Palau
Palestine
Panama
Papua New Guinea
Paraguay
Peru
Philippines
Pitcairn
Poland
Portugal
Qatar
Romania
Russia
Rwanda
Saint Helena
Saint Kitts and Nevis
Saint Lucia
Saint Pierre and Miquelon
Saint Vincent and the Grenadines
Samoa
San Marino
Sao Tome and Principe
Saudi Arabia
Senegal
Serbia
Seychelles
Sierra Leone
Singapore
Sint Maarten (Dutch part)
Slovakia
Slovenia
Solomon Islands
Somalia
South Africa
South America
South Korea
South Sudan
Spain
Sri Lanka
Sudan
Suriname
Sweden
Switzerland
Syria
Taiwan
Tajikistan
Tanzania
Thailand
Timor
Togo
Tokelau
Tonga
Trinidad and Tobago
Tunisia
Turkey
Turkmenistan
Turks and Caicos Islands
Tuvalu
Uganda
Ukraine
United Arab Emirates
United Kingdom
United States
Upper middle income
Uruguay
Uzbekistan
Vanuatu
Vatican
Venezuela
Vietnam
Wallis and Futuna
World
Yemen
Zambia
Zimbabwe
0
1000
2000
3000
4000
Total Deaths Per Million

## نمایش دادگان: ۲

می‌خواهیم تاثیر واکسیناسیون بر تعداد فوتی‌ها را بررسی کنیم. برای این کار فرض کنید الزام است که اطلاعات ۵ کشور را بررسی کنیم. شما کدام کشورها را برای مقایسه انتخاب می‌کنید؟ با یک نمودار مناسب این مساله را بررسی نمایید و برداشت خود را از نتایج ذکر نمایید.

برای انجام مقایسه از بین کشورهای جهان، کشورهای:

۱) ایران
۲) کانادا
۳) ایتالیا
۴) آمریکا
۵) آلمان

را انتخاب کردیم. یکی از علت‌های این انتخاب‌ها این است که برای موردهای ۲ الی ۵ در این کشورها مطمئن هستیم که داده‌های موجود از سایر کشورها دقیق‌تر بوده و هم داده‌ای خالی ندارند و هم نویز کمی دارند. این خواص از این جهت مهم هستند که زمانی که می‌خواهیم تاثیرات پارامترهای مختلف برهم را در کنار یک‌دیگر ارزیابی کنیم نویزها و داده‌هایی که ناقص هستند می‌توانند باعث نادرست شدن برداشت‌های ما شوند.

برای انجام این بررسی ابتدا برای هر کشور یک دیتافریم می‌سازیم:

```
In [385]:
df_22_iran = test_df[iran_date_df.location=='Iran']
df_22_canada = test_df[iran_date_df.location=='Canada']
df_22_italy = test_df[iran_date_df.location=='Italy']
df_22_united_states = test_df[iran_date_df.location=='United States']
df_22_germany = test_df[iran_date_df.location=='Germany']
```

| | iso_code | continent | location | date | total_cases | new_cases | new_cases_smoothed | total_deaths | new_deaths |
|---|---|---|---|---|---|---|---|---|---|
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 156571 | USA | North America | United States | 2020-01-22 | 1.0 | 11486.032247 | 11492.440366 | 161848.0 | 1.0 |
| 156572 | USA | North America | United States | 2020-01-23 | 1.0 | 0.000000 | 11492.440366 | 161848.0 | 1.0 |
| 156573 | USA | North America | United States | 2020-01-24 | 2.0 | 1.000000 | 11492.440366 | 161848.0 | 1.0 |
| 156574 | USA | North America | United States | 2020-01-25 | 2.0 | 0.000000 | 11492.440366 | 161848.0 | 1.0 |
| 156575 | USA | North America | United States | 2020-01-26 | 5.0 | 3.000000 | 11492.440366 | 161848.0 | 1.0 |
| 157337 | USA | North America | United States | 2022-02-26 | 78929797.0 | 48638.000000 | 67034.286000 | 948457.0 | 798.0 |
| 157338 | USA | North America | United States | 2022-02-27 | 78947866.0 | 18069.000000 | 66453.857000 | 948639.0 | 182.0 |
| 157339 | USA | North America | United States | 2022-02-28 | 79044330.0 | 96464.000000 | 68640.857000 | 950732.0 | 2093.0 |
| 157340 | USA | North America | United States | 2022-03-01 | 79091361.0 | 47031.000000 | 62331.429000 | 952423.0 | 1691.0 |
| 157341 | USA | North America | United States | 2022-03-02 | 79143716.0 | 52355.000000 | 57824.857000 | 954518.0 | 2095.0 |

771 rows × 71 columns

سپس شاخص‌های مد نظر را برای هر کشور از مجموعه داده استخراج می‌کنیم:

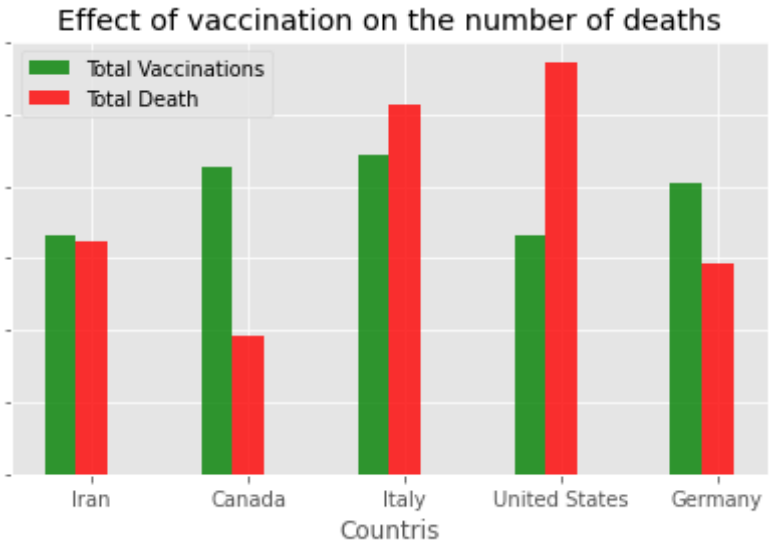
In [696]:

```
con_list = ['Iran', 'Canada', 'Italy', 'United States', 'Germany']

df_2222 = test_df[test_df.location == 'Germany']
df_2222

df_2222 = df_2222.groupby('location').agg({#'total_vaccinations_per_hundred': ['max'],
                                           #'people_vaccinated_per_hundred': ['max'],
                                           #'people_fully_vaccinated_per_hundred': ['max'],
                                           #'total_boosters_per_hundred': ['max'],
                                           #'new_deaths_per_million': ['max'],
                                           'total_deaths_per_million': ['max'],
                                           'total_vaccinations_per_hundred': ['max'],
                                            })


df_2222
```

executed in 41ms, finished 17:03:48 2022-04-08

| | total_deaths_per_million | total_vaccinations_per_hundred |
|---|---|---|
| | max | max |
| location | | |
| Germany | 1469.73 | 202.83 |

نهایتاً نمودار زیر را برای بررسی اثر تعداد دوز واکسن تزریقی به نسبت جمعیت بر تعداد فوتی‌ها نسبت به جمعیت رسم کردیم.




به عنوان تحلیل کشور به کشور می‌توان گفت هرچه که میزان واکسیناسیون (ارتفاع ستون سبز) بیشتر باشد انتظار می‌رود که ارتفاع ستون قرمز رنگ کمتر باشد. برای نمونه کشور کانادا مثال دقیقی از این تحیلی است. اما مشاهده می‌شود که ایتالیا علی‌رغم این که در واکسیناسیون عملکردی نزدیک به کانادا و حتی بهتر از آن داشته است، نسبت فوتی به جمعیت بالاتری از کانادا دارد و این شبهه را ایجاد می‌کند که انتظاری که در چند خط قبل گفتم نادرست باشد! اما در پاسخ می‌توان گفت که بیشتر فوتی‌های ایتالیا برای زمان قبل از عرضه واکسن‌ها بوده. (این مسئله را در نمودارهایی که برای تحلیل دقیق‌تر رسم کرده‌ام که در ادامه گزارش قرار دارند می‌توان تحقیق کرد.)

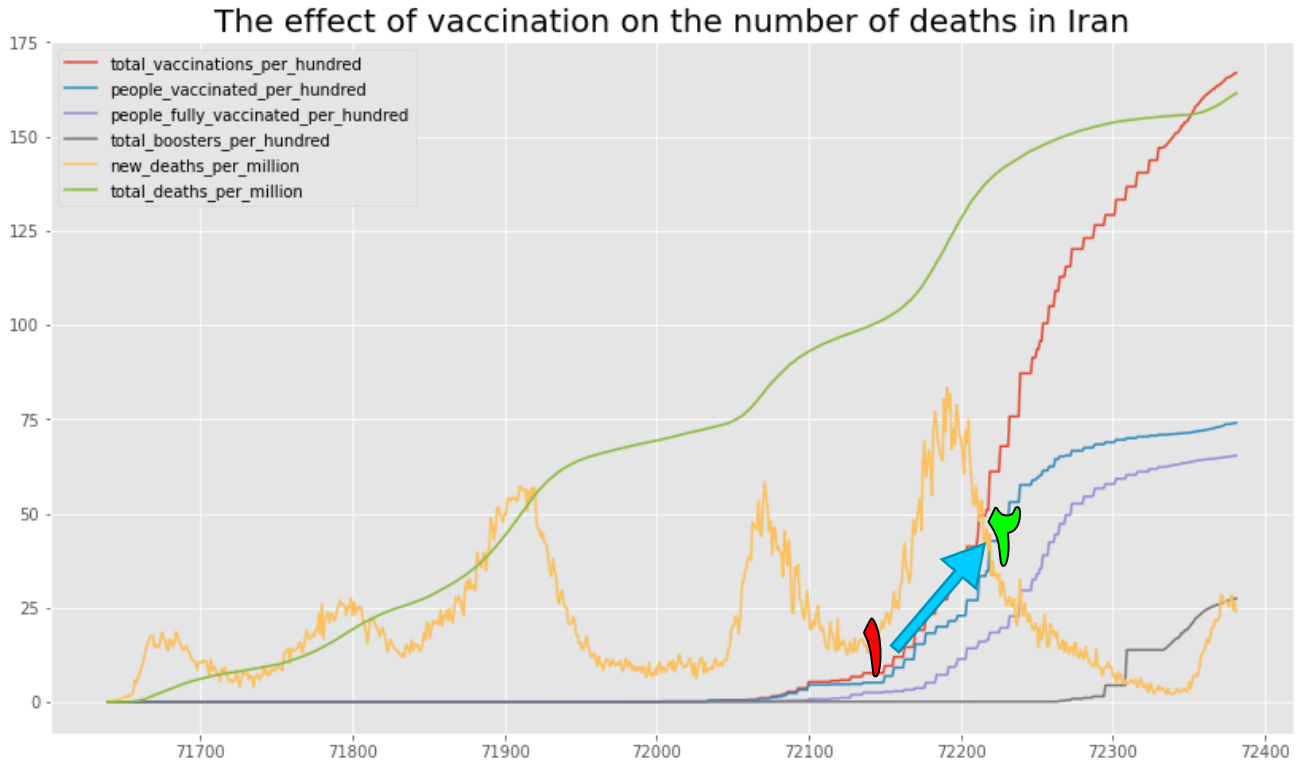
در مورد ایران، خیلی نکته عجیبی وجود دارد و آن این است که با این که تعداد واکسیناسیون در ایران نسبت به جمعیت از تعداد واکسیناسیون در کشورهای آلمان و کانادا به نسبت جمعیتشان کمتر است اما کشته‌های ایران بیشتر هستند!!!! چرا؟!؟ مضافاً باید این راهم در نظر گرفت که همین تعداد واکسن‌ها برای ایران خیلی دیرتر از دیگر کشورهای حاضر در این نمودار به دست مردم رسیده است. (ان موضوع با کمک نمودارهایی که در ادامه و در بخش بررسی سرعت واکسیناسیون در کشورهای مختلف رسم کرده ام قابل اثبات است.)

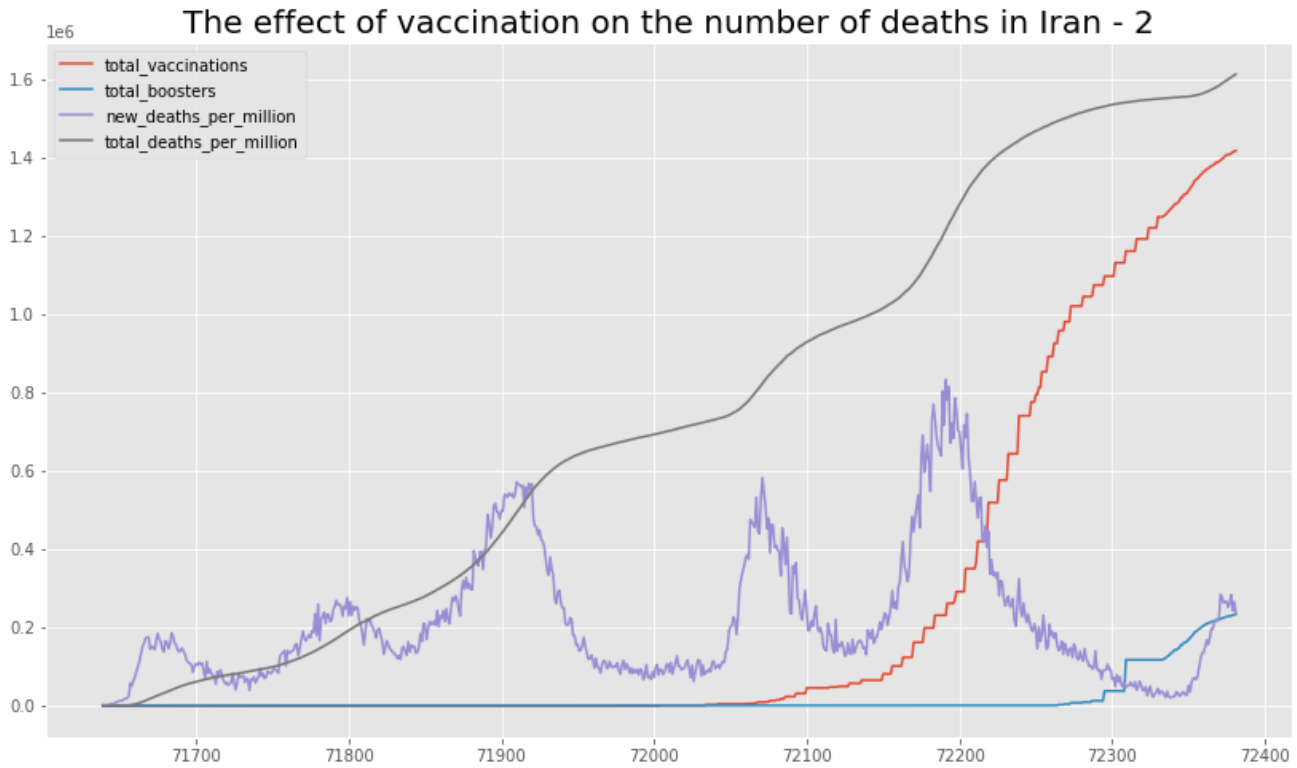
علت زیاد بودن فوتی‌ها در آمریکا نیز رعایت نکردن شیوه‌های بهداشتی از سوی عده قابل ملاحظه‌ای از مردم است. (ماجراهای اعتراض به محدودیت‌ها که یک عده می‌گفتند آزادی ما دارد محدود می‌شود! -خب بشود بهتر از این است که بمیرید!-)

آلمان و کانادا مورد خاصی ندارند و به نظر نرمال می‌آیند.

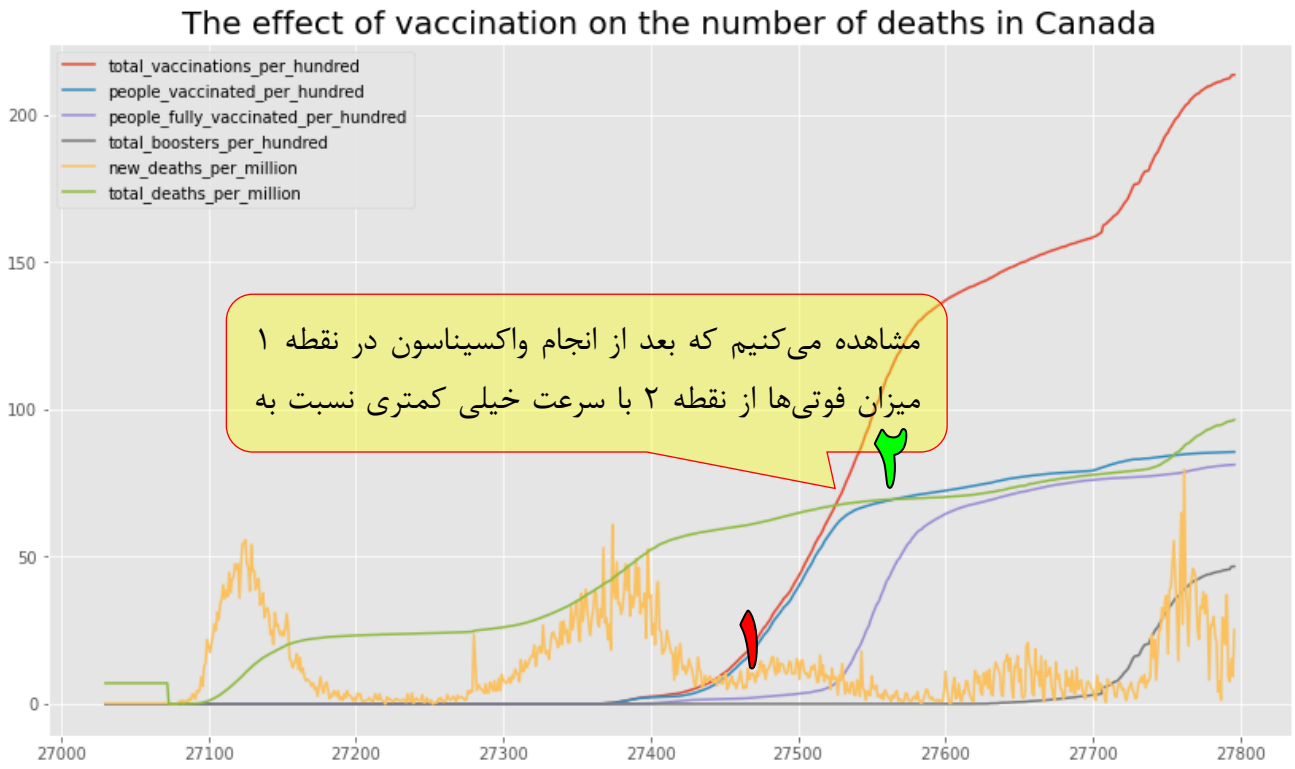
در ادامه برای دقیق‌تر بودن تحلیل‌ها برای هر کشور به صورت جداگانه دو نمودار رسم کرده‌ایم. در این نمودارها می‌توان دید زمانی که خطهای قرمز، آبی، بنفش و خاکستری از نمودارهای اول هر کشور با صعود می‌کنند، یعنی میزان واکسیناسیون زیاد می‌شود، خط طلایی رنگ که تعداد فوتی‌های جدید است در حالی که رو به صعود داشته و از روند قبلی بر می‌آید که بعد از هر صعود، صعود بعدی قوی‌تر است، روند کاهشی شروع می‌شود.

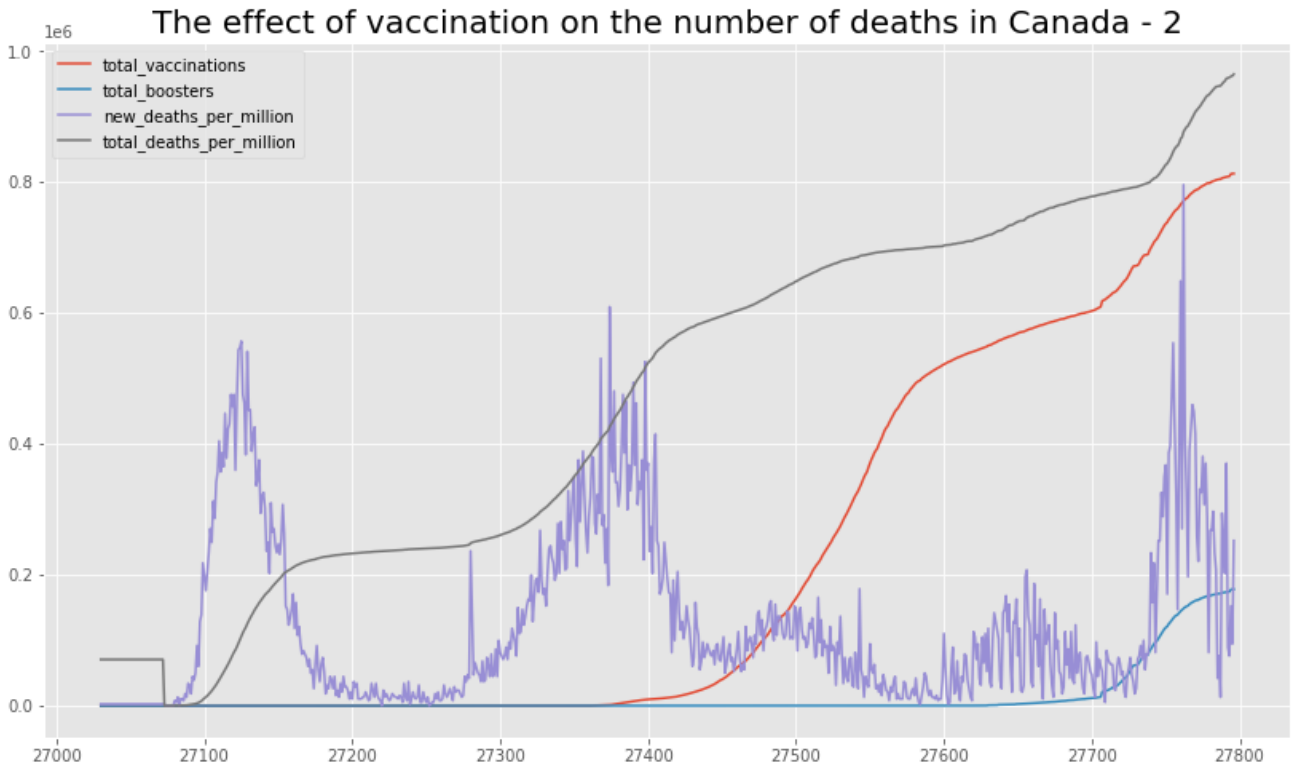
## ایران


The effect of vaccination on the number of deaths in Iran
total_vaccinations_per_hundred
people_vaccinated_per_hundred
people_fully_vaccinated_per_hundred
total_boosters_per_hundred
new_deaths_per_million
total_deaths_per_million
175
150
125
100
75
50
25
0
71700
71800
71900
72000
72100
72200
72300
72400
۱
۲

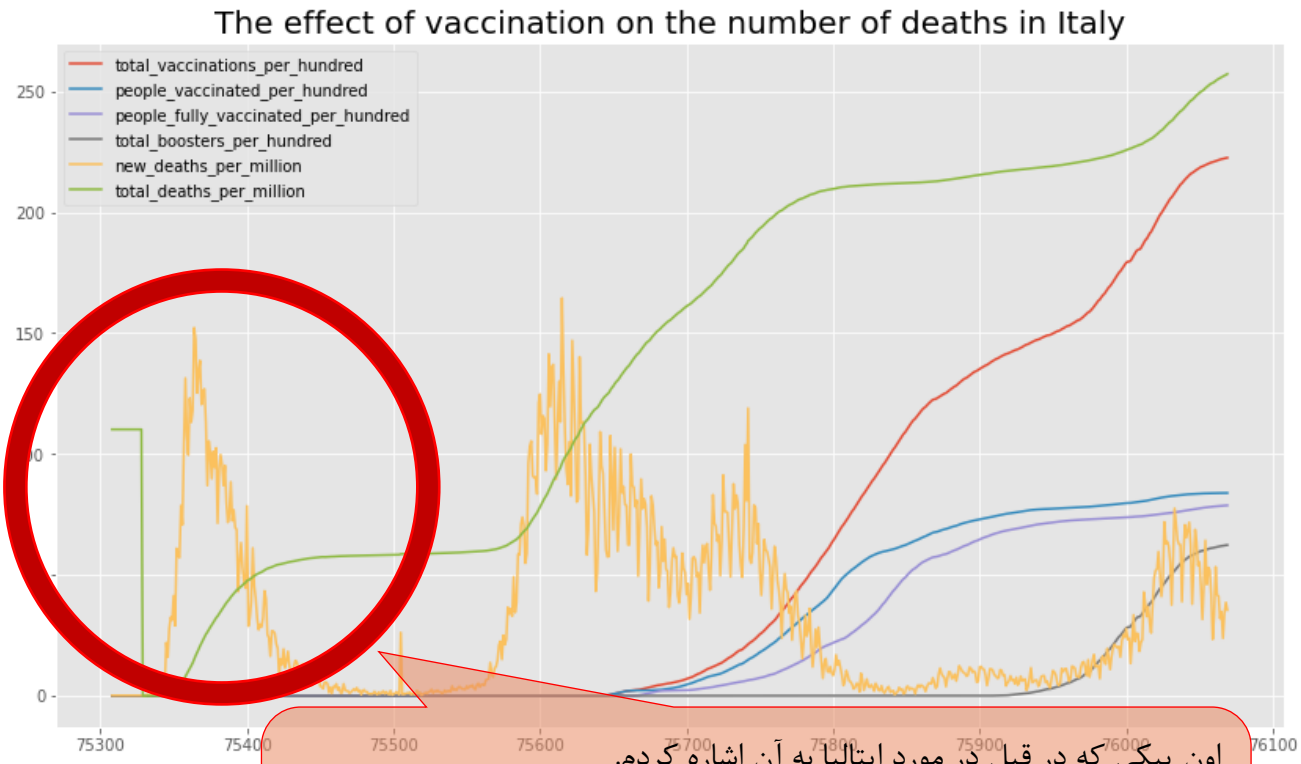

The effect of vaccination on the number of deaths in Iran - 2
1e6
total_vaccinations
total_boosters
new_deaths_per_million
total_deaths_per_million
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
71700
71800
71900
72000
72100
72200
72300
72400

## کانادا


The effect of vaccination on the number of deaths in Canada
total_vaccinations_per_hundred
people_vaccinated_per_hundred
people_fully_vaccinated_per_hundred
total_boosters_per_hundred
new_deaths_per_million
total_deaths_per_million
مشاهده می‌کنیم که بعد از انجام واکسیناسون در نقطه ۱
میزان فوتی‌ها از نقطه ۲ با سرعت خیلی کمتری نسبت به
۲
۱
200
150
100
50
0
27000
27100
27200
27300
27400
27500
27600
27700
27800

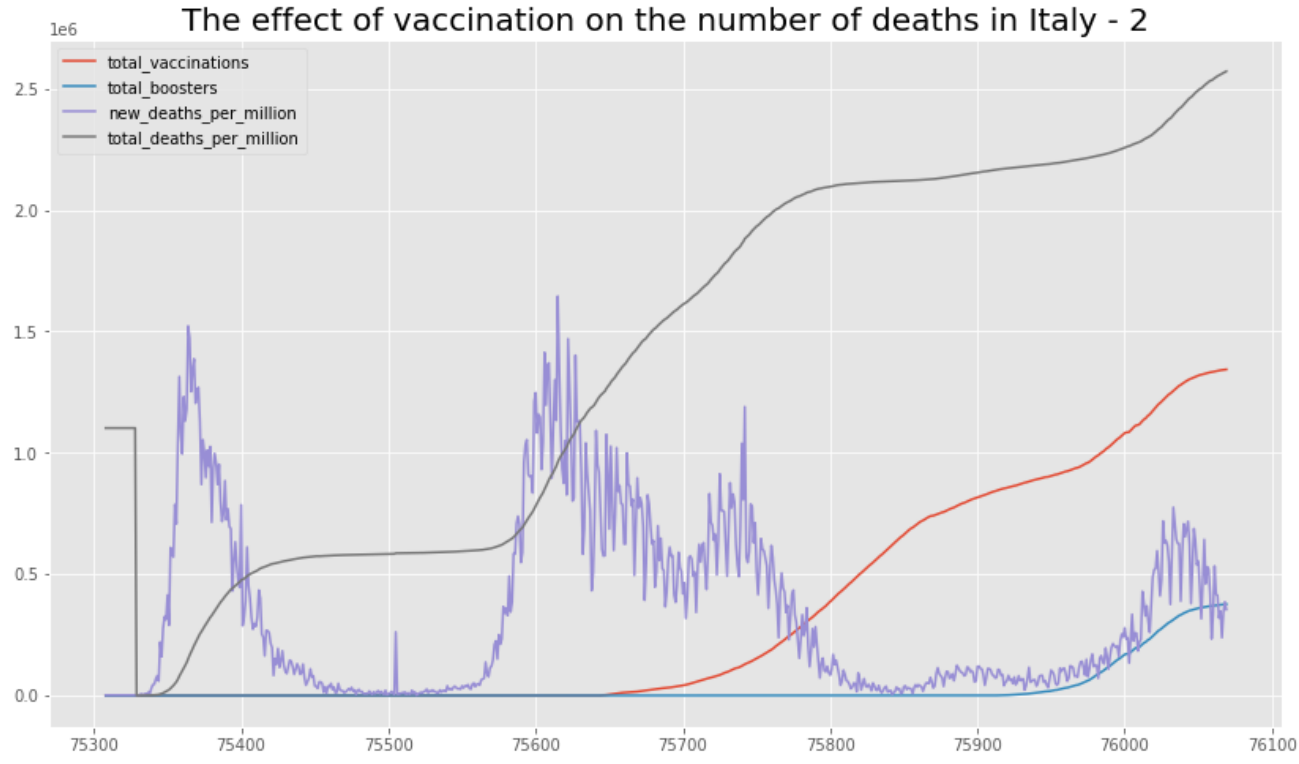

The effect of vaccination on the number of deaths in Canada - 2
1e6
total_vaccinations
total_boosters
new_deaths_per_million
total_deaths_per_million
1.0
0.8
0.6
0.4
0.2
0.0
27000
27100
27200
27300
27400
27500
27600
27700
27800

## ایتالیا

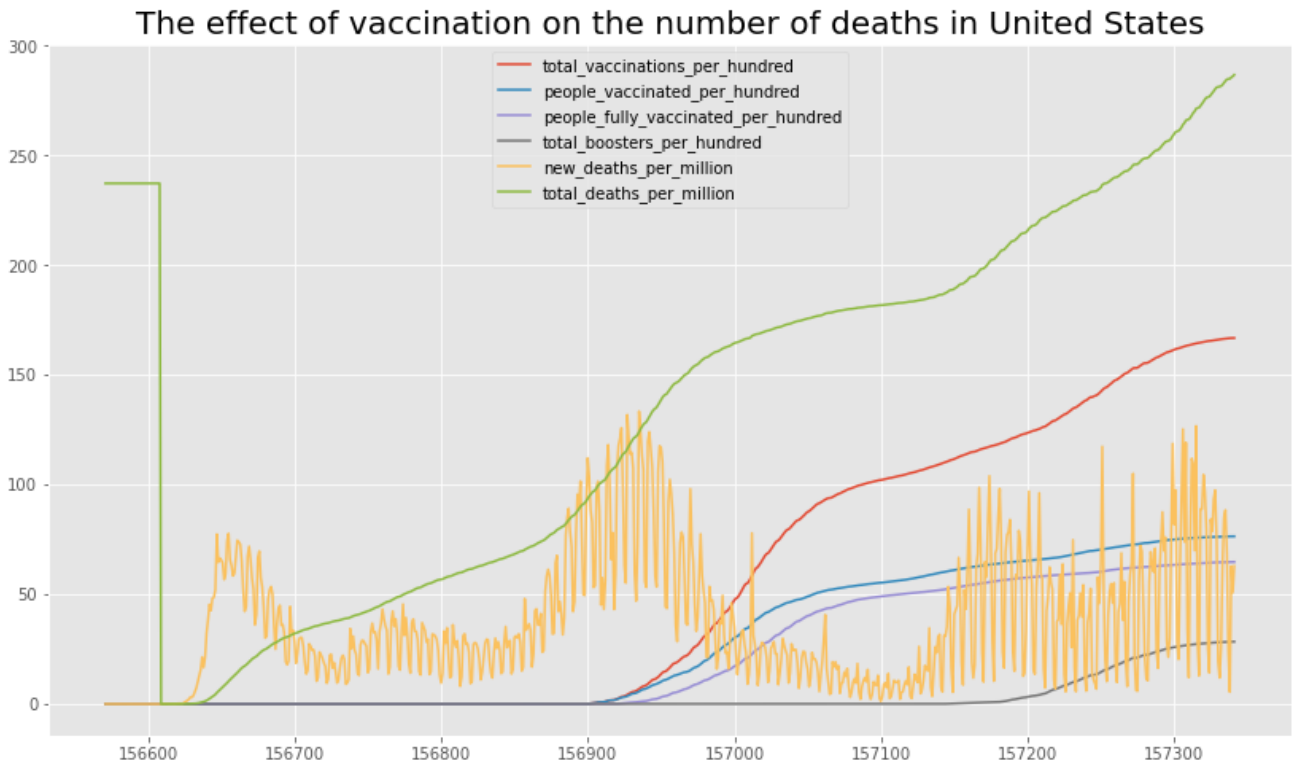


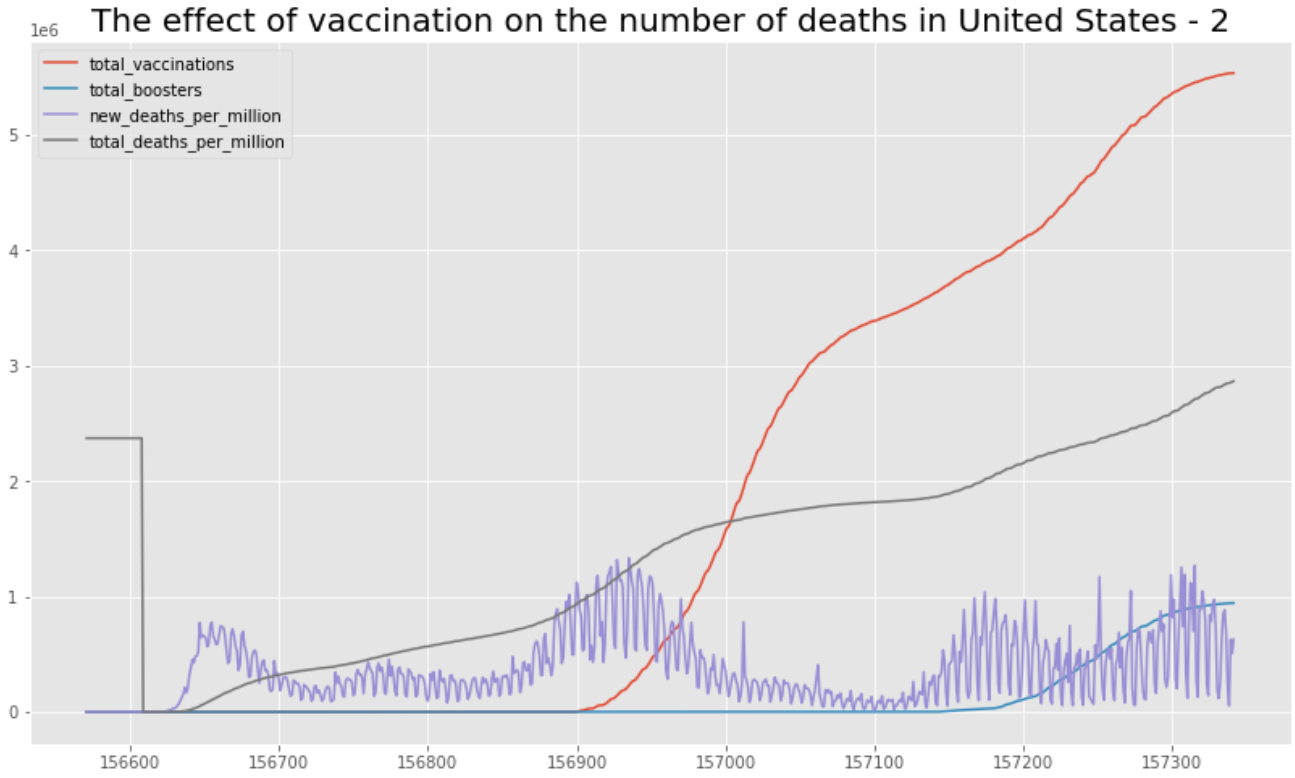

اون پیکی که در قبل در مورد ایتالیا به آن اشاره کردم.

همانطور که ملاحظه می‌کنید در بین کشورهایی که بررسی کردیم ایتالیا قبل از شروع عرضه واکسن وضع وخیم تری داشته

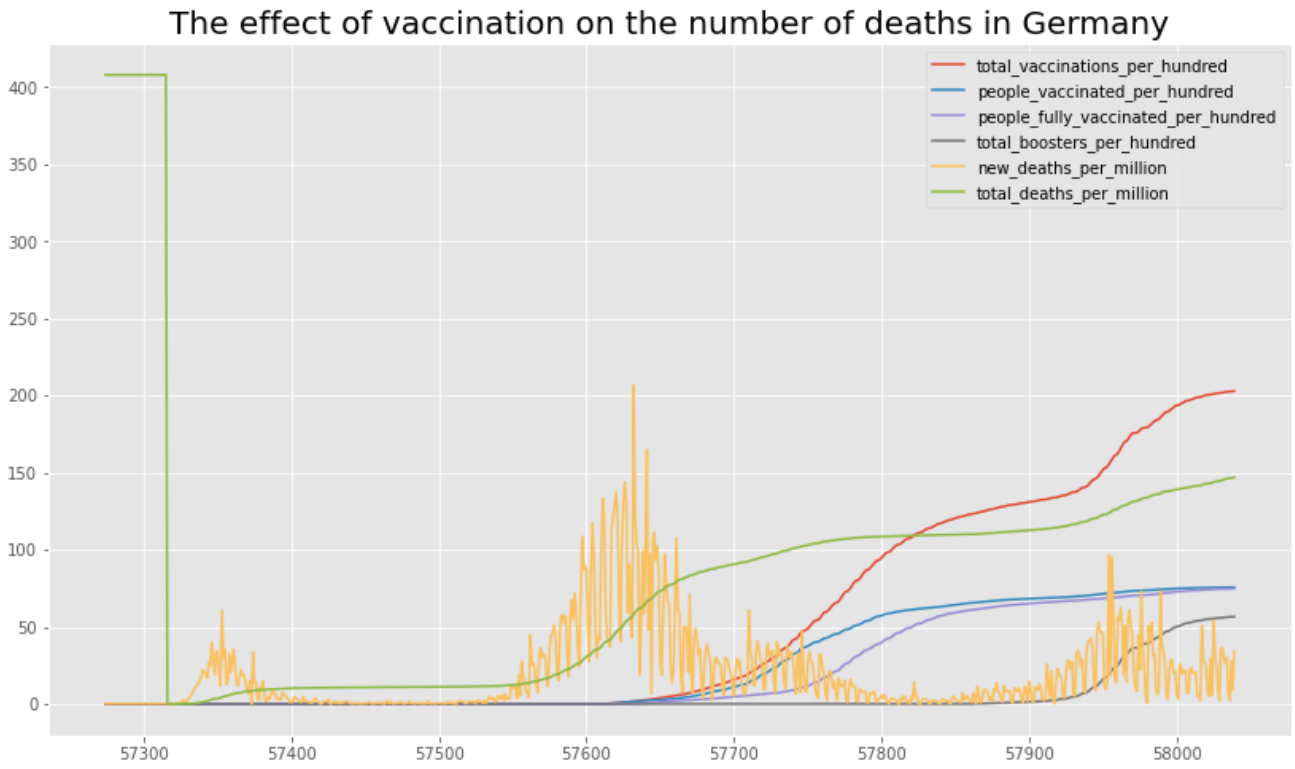


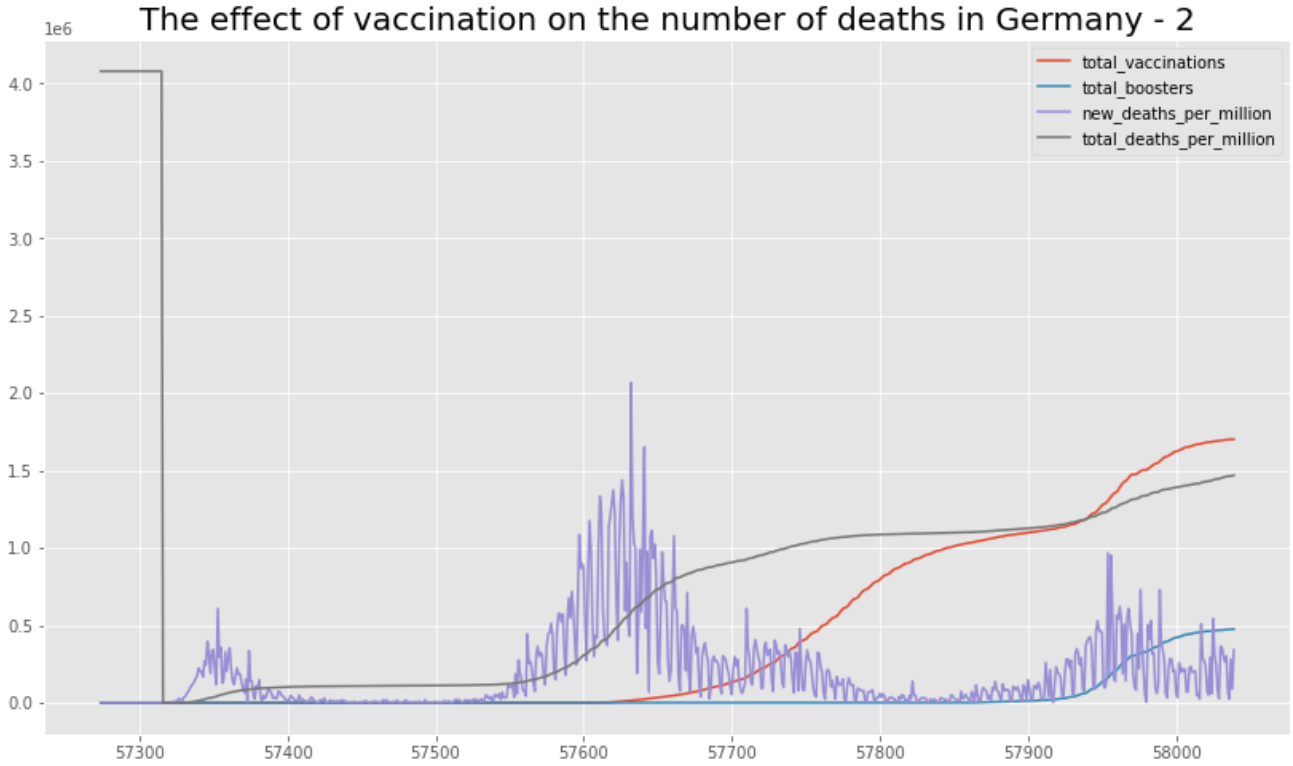
## آمریکا


The effect of vaccination on the number of deaths in United States
total_vaccinations_per_hundred
people_vaccinated_per_hundred
people_fully_vaccinated_per_hundred
total_boosters_per_hundred
new_deaths_per_million
total_deaths_per_million
300
250
200
150
100
50
0
156600
156700
156800
156900
157000
157100
157200
157300


The effect of vaccination on the number of deaths in United States - 2
1e6
total_vaccinations
total_boosters
new_deaths_per_million
total_deaths_per_million
5
4
3
2
1
0
156600
156700
156800
156900
157000
157100
157200
157300

## آلمان


The effect of vaccination on the number of deaths in Germany
total_vaccinations_per_hundred
people_vaccinated_per_hundred
people_fully_vaccinated_per_hundred
total_boosters_per_hundred
new_deaths_per_million
total_deaths_per_million
400
350
300
250
200
150
100
50
0
57300
57400
57500
57600
57700
57800
57900
58000


The effect of vaccination on the number of deaths in Germany - 2
1e6
total_vaccinations
total_boosters
new_deaths_per_million
total_deaths_per_million
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
57300
57400
57500
57600
57700
57800
57900
58000

## نمایش دادگان: ۳

قصد داریم سرعت واکسیناسیون در کشورهای مختلف را بررسی کنیم. برای این کار فرض کنید الزام است که اطلاعات پنج کشور را ارزیابی کنیم. شما کدام کشورها را برای مقایسه انتخاب می‌کنید؟ با یک نمودار مناسب این مساله را بررسی نمایید و برداشت خود را از نتایج ذکر نمایید.

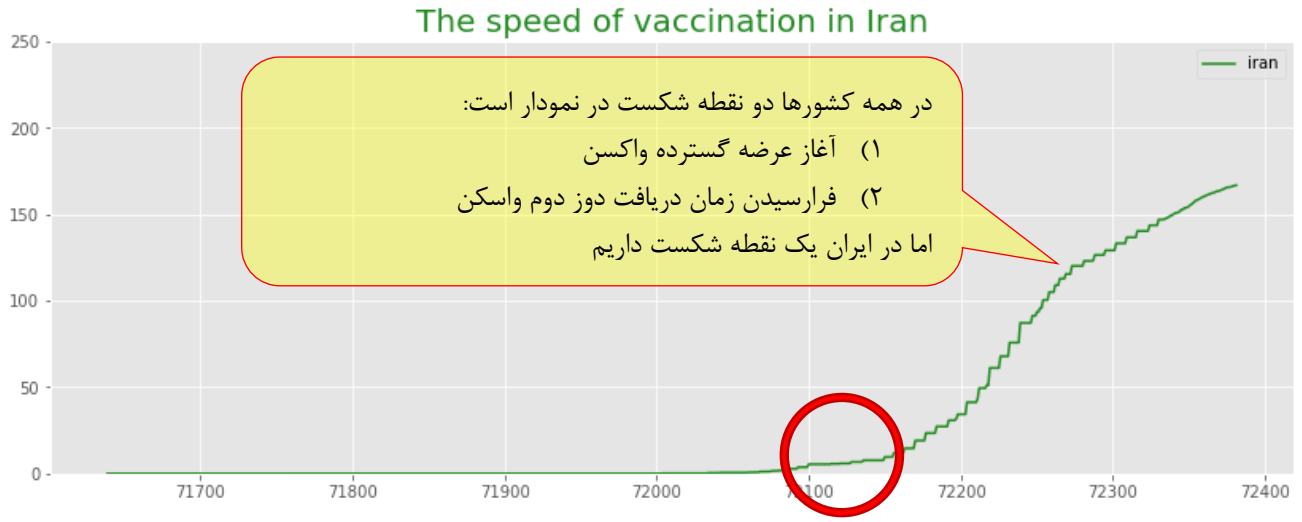
برای انجام مقایسه از بین کشورهای جهان، کشورهای:

۱) ایران
۲) کانادا
۳) ایتالیا
۴) آمریکا
۵) آلمان

را انتخاب کردیم. یکی از علت‌های این انتخاب‌ها این است که برای موردهای ۲ الی ۵ در این کشورها مطمئن هستیم که داده‌های موجود از سایر کشورها دقیق‌تر بوده و هم داده‌ای خالی ندارند و هم نویز کمی دارند. این خواص از این جهت مهم هستند که زمانی که می‌خواهیم تاثیرات پارامترهای مختلف برهم را در کنار یک‌دیگر ارزیابی کنیم نویزها و داده‌هایی که ناقص هستند می‌توانند باعث نادرست شدن برداشت‌های ما شوند.

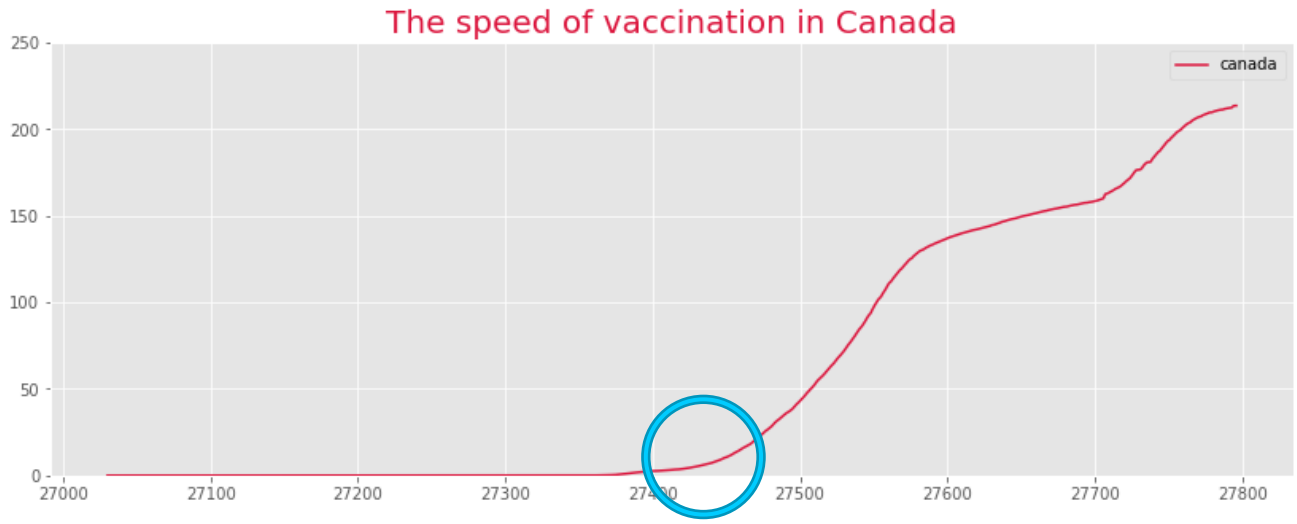
یک نمونه از این نویزهایی که در این سوال و سوال قبل بدان‌ها اشاره کردم در شکل دندانه دندانه‌ای نمودار ایران قابل مشاهده است. همانطور که در شکل سایر چهار کشور ملاحظه می‌شود خط خیلی نرم‌تر بوده و شکستگی‌های کمتری دارد.

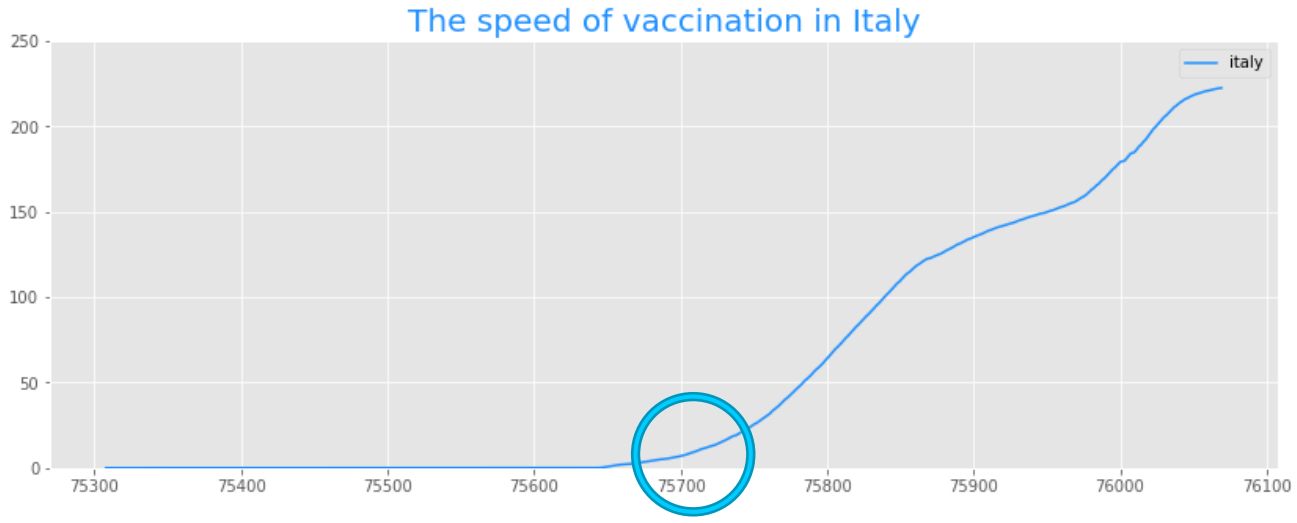
با نوشتن کدهای زیر برای هر کشور نمودار سرعت واکسیناسیون آن‌ها را مقایسه می‌کنیم:

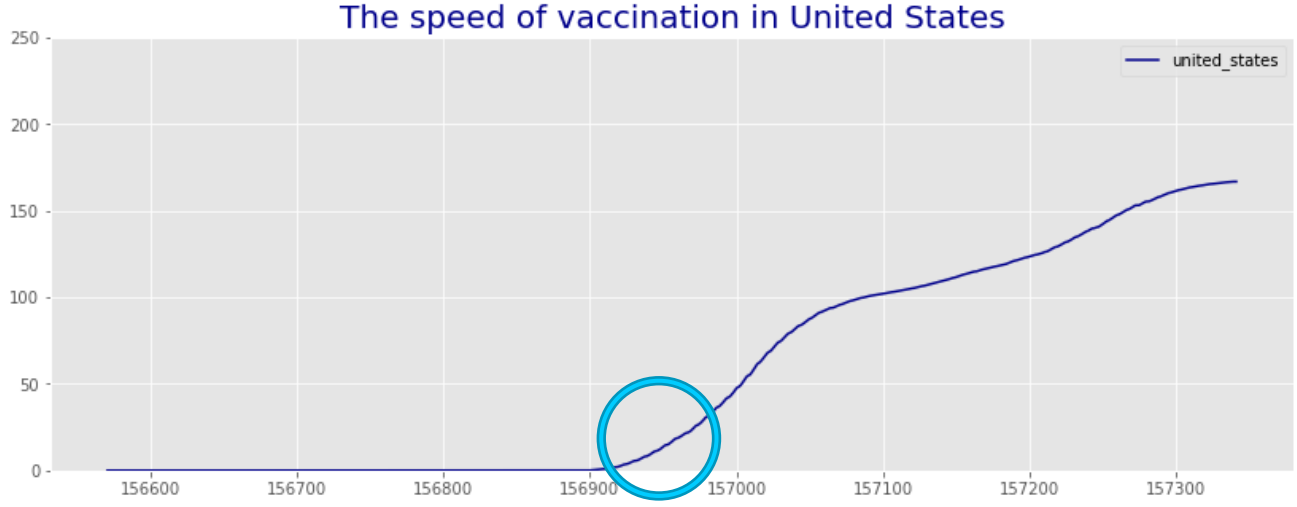


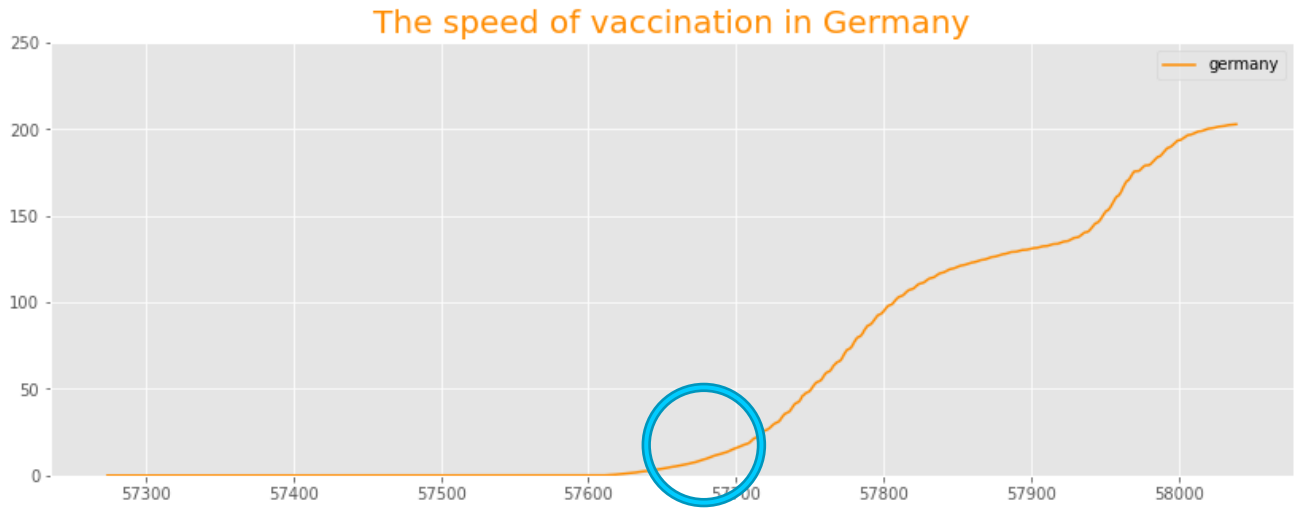

محل شروع دریافت واکسن در نمودارها با دایره‌های توخالی مشخص شده اند.

همانطور که ملاحظه می‌کنید ایران حدود ۱۰۰ روز دیرتر از کشورهای توسعه یافته واکسیناسیون سراسری را شروع کرده است. (مقایسه محل دایره‌های آبی و دایره قرمز)



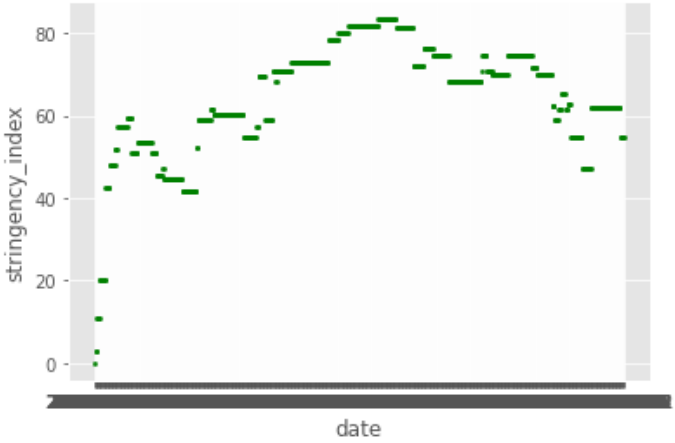





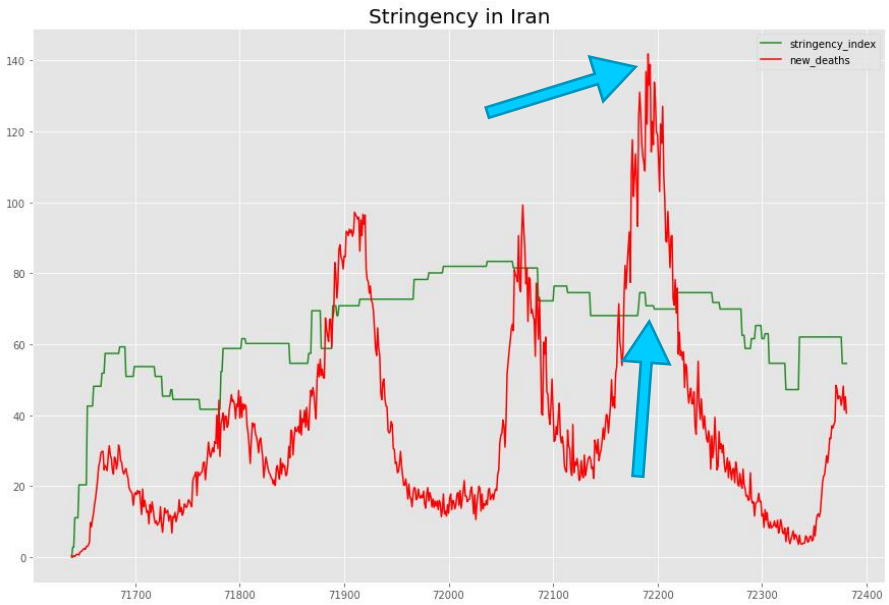

## نمایش دادگان: ۴

روند سختگیری در حوزه‌ی کرونا در ایران را در طول زمان بررسی کنید، توجه نمایید برای پاسخگویی به این سوال براساس تحلیل خود می‌توانید از ویژگی یا ویژگی‌های دلخواه استفاده نمایید، تحلیل خود را بیان نمایید.

برای بررسی روند سختگیری در حوزه کرونا در ایران از شاخص `stringency_index` که در مجموعه‌داده به همین منظور ثبت شده استفاده کردیم. در این راستا نمودار زیر میزان سختگیری‌های اعمال شده در ایران بر اساس این معیار را رسم کرده ایم:



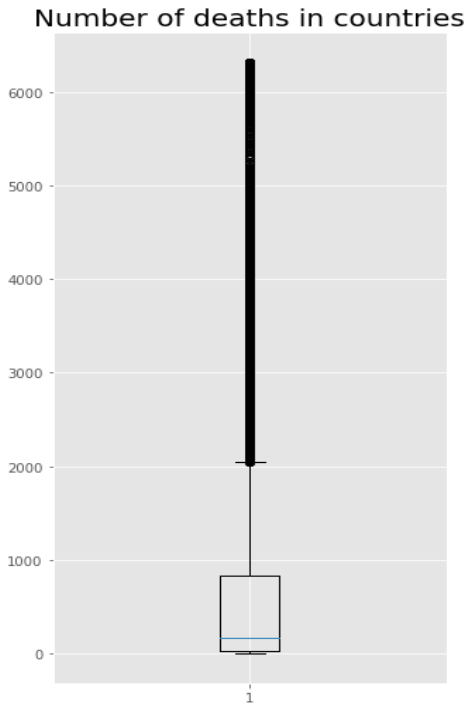

در ادامه برای تحلیل بهتر از میزان موفقیت و کارا بودن این سختگیری‌ها آن‌را در کنار میزان مرگ و میر با نمودار ارزیابی می‌کنیم:




برداشت و تحلیلی که من از این نمودار دارم این است که هر زمان میزان مرگ و میر زیاد شده محدودیت‌ها هم با کمی تاخیر زمانی زیاد شده اند و زیاد شدن محدودیت‌ها در ادامه موجب کمتر شدن فوتی‌های روزانه شده است. اما در محدوده ۷۲۲۰۰ مشاهده می‌کنیم که یک پیک درحال شکل گیری سریع و شارپ است اما محدودیت‌ها زیاد نشده اند و نتیجتاً بد ترین پیک رخ می‌دهد. (-فکر کنم دوره پیک دلتا باشه)

## نمایش دادگان: ۵

با استفاده از دیتافریم تجمیع شدهای که ایجاد کردید، برای ویژگی تعداد فوتی‌های هر کشور نمودار BoxPlot رسم کنید. با توجه به مقدار میانه و میانگین، چولگی نمودار به کدام سمت می‌باشد؟



با توجه به نمودار چولگی به سمت داده های کوچک است.

## نمایش دادگان: ۶

تاثیر ویژگی‌های تراکم جمعیت، میانگین سنی، وجود امکانات بهداشتی، تعداد تخت بیمارستان‌ها و شاخص پیشرفت انسانی را بر تعداد فوتی‌ها و تعداد کیس‌های جدید با رسم نمودار مناسب بررسی کنید.

ابتدا هریک از این شاخص‌ها را به صورت تجمیع شده برای هر کشور محاسبه کرده و یک دیتافریم برایشان شکل می‌دهیم و سپس برای شاخص‌ها نمودار اسکترپلات رسم می‌کنیم.

2-6

In [569]:

```
df_26 = test_df.groupby('location').agg({'total_cases_per_million': ['max'],
                                         'total_deaths_per_million': ['max'],
#                                          'population_density': ['max'],
                                         'median_age': ['max'],
                                         'handwashing_facilities': ['max'],
                                         'hospital_beds_per_thousand': ['max'],
                                         'human_development_index': ['max'],
                                         'gdp_per_capita': ['max'],
                                         'extreme_poverty': ['max'],
                                         'life_expectancy': ['max'],
                                            })
# df_26 = df_26[df_26['population_density'] != 'Not Cal']

df_26

```

executed in 253ms, finished 13:55:56 2022-04-08

| | total_cases_per_million | total_deaths_per_million | median_age | handwashing_facilities | hospital_beds_per_thousand |
|---|---|---|---|---|---|
| | max | max | max | max | max |
| location | | | | | |
| Peru | 105479.098 | 6317.377 | 29.100000 | 50.790872 | 1.600000 |
| Bulgaria | 158945.750 | 5175.842 | 44.700000 | 50.790872 | 7.454000 |
| Bosnia and Herzegovina | 113852.510 | 4744.659 | 42.500000 | 97.164000 | 3.500000 |
| Hungary | 186121.014 | 4580.990 | 43.400000 | 50.790872 | 7.020000 |
| North Macedonia | 143179.807 | 4338.680 | 39.100000 | 50.790872 | 4.280000 |
| Montenegro | 367027.518 | 4270.354 | 39.100000 | 50.790872 | 3.861000 |
| Georgia | 406093.262 | 4078.373 | 38.700000 | 50.790872 | 2.600000 |
| Croatia | 259319.438 | 3704.868 | 44.000000 | 50.790872 | 5.540000 |
| Czechia | 335634.874 | 3612.365 | 43.300000 | 50.790872 | 6.630000 |
| Slovakia | 394670.479 | 3407.245 | 41.200000 | 50.790872 | 5.820000 |
| Romania | 143706.073 | 3328.563 | 43.000000 | 50.790872 | 6.892000 |
| San Marino | 424433.990 | 3293.149 | 30.568558 | 50.790872 | 3.800000 |

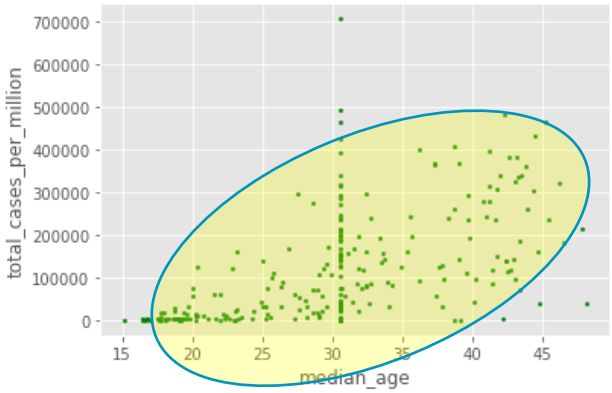
** در ابتدا یک توضیح مهم در مورد اسکترپلات‌ها

در بسیاری از اسکترپلات‌های رسم شده در این سوال و سوال بعدی شاهد وجود الگوهایی به شکل زیر هستیم:

وجود این نقاط روی یک محور به این دلیل است که چون مجموعه‌داده‌ای که در اختیار ما قرار گرفته است تعداد زیادی مقادیر گم شده داشته و ماهم برای پر کردن تعدادی از ویژگی‌ها از روش میانگین‌گیری استفاده کردیم، تعداد زیادی داده وجود دارند که مقدار ویژگی آن‌ها برابر با میانگین است که این مسئله خود را در نمودار ما به این شکل بروز می‌دهد.

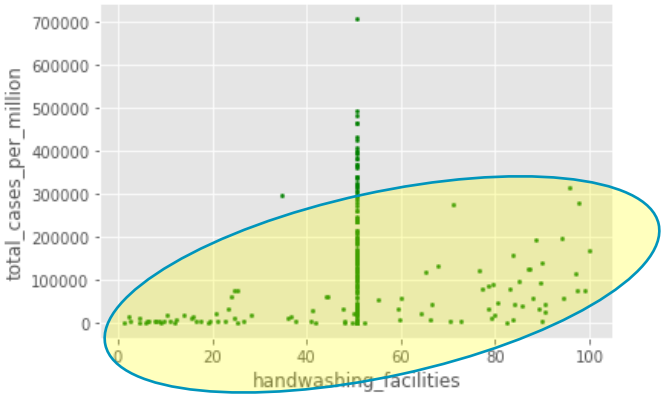
اگر به جای این مقادیر صفر میگذاشتیم یا سیاست دیگری انتخاب می‌کردیم شاید این خطوط شکل نمی‌گرفتند اما چون در بسیاری از موارد میانگین کل این ویژگی‌ها برایمان مورد استفاده داشت و با پرکردن آن ویژگی‌ها با حالتی غیر از میانگین‌گیری، مقدار میانگین خراب می‌شد، روش پرکردن داده‌های گم شده برای این ویژگی‌ها را تغییر ندادیم.

**میانه سن به نرخ ابتلا نسبت به جمعیت**



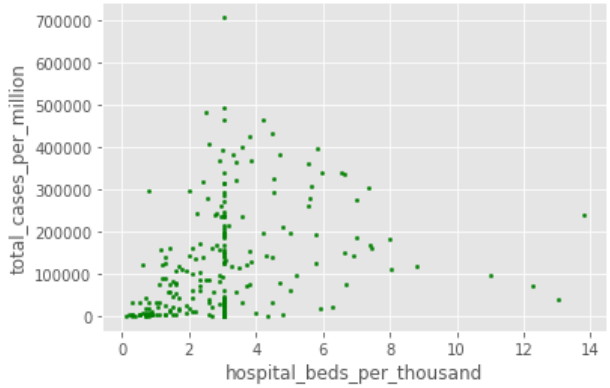

این نمودار به وضوح مشخص می‌کند که هرچه میانه سن جمعیت یک کشور بالاتر باشد میزان شیوع کرونا هم بیشتر است.

### زیر ساخت‌های شستوشوی دست به نرخ ابتلا نسبت به جمعیت



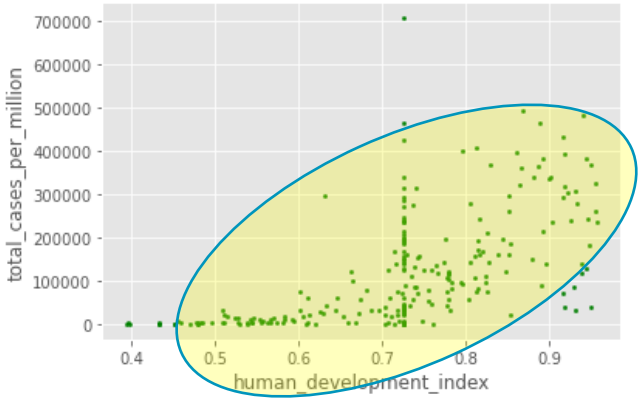

با کمال تعجب شاهد هستیم که هرچه امکانات براش شستوشوی دست بیشتر باشد بیماری هم بیشتر است. شاید اگر محاسبات ما درست باشد بتوان گفت که این اصل که ویروس از تماس فیزیکی هم منتقل می‌شود درست نیست!.

### تعداد تخت‌های بیمارستانی نسبت به جمعیت به نرخ ابتلا نسبت به جمعیت



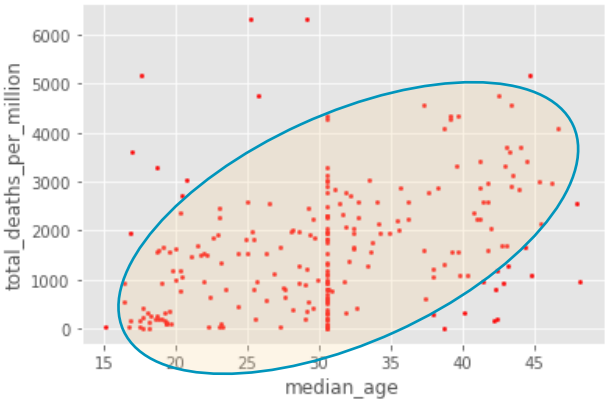

با بررسی این نمودار رابطه خاصی را بین این دو مشخصه درک نکردم.

### شاخص توسعه فردی به نرخ ابتلا نسبت به جمعیت



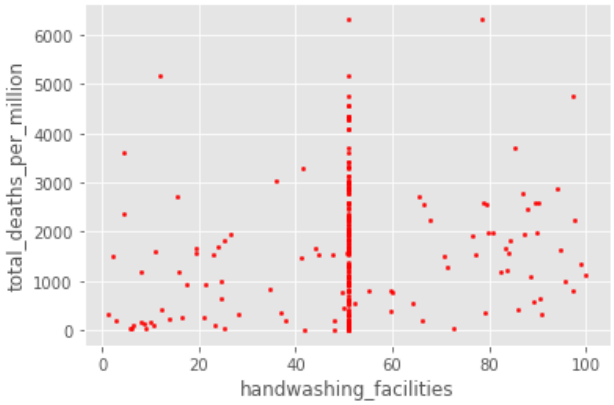
شکل بالا به خوبی نشان می‌دهد که هرچه شاخص توسعه فردی بیشتر باشد ابتلا کاهش پیدا می‌کند. (خوش به حال مردم کشورهایی که این شاخص‌شون خوبه)

### میانه سن به نرخ فوتی نسبت به جمعیت



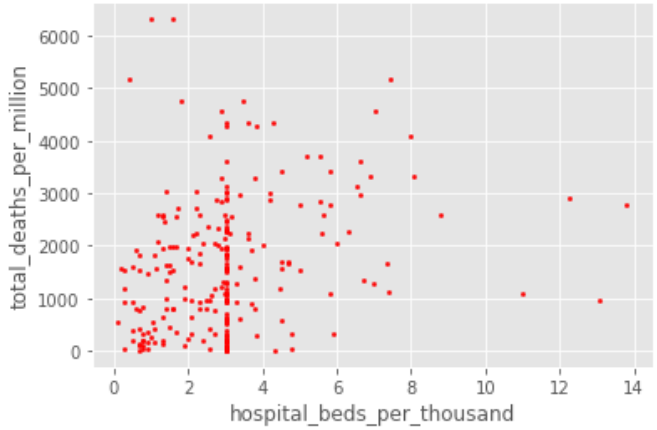

همانند میزان ابتلا، میانه سن بیشتر برای کشورها فوتی بیشتری برای آن‌ها در همه‌گیری کرونا را در پی داشته است.

### زیر ساخت‌های شستوشوی دست به نرخ فوتی نسبت به جمعیت



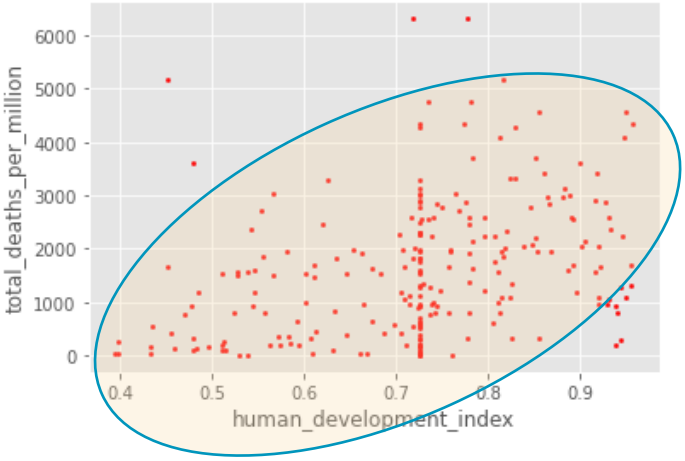

از این شکل نیز برداشت خاصی ندارم!

### تعداد تخت‌های بیمارستانی نسبت به جمعیت به نرخ فوتی نسبت به جمعیت




تحلیلی برای این شکل ندارم.

### شاخص توسعه فردی به نرخ فوتی نسبت به جمعیت




کشورهای با توسعه فردی بالاتر، مرگ و میر به نسبت جمعیت کمتری نسبت به کشورهای با توسعه فردی کمتر دارند.

## نمایش دادگان: ۷

رابطه بین وضعیت اقتصادی کشورها و تعداد افراد واکسینه شده را بررسی کنید و تحلیل خود را بیان نمایید.

برای ارزیابی وضعیت اقتصادی کشورها از بین معیارهای موجود در مجموعه‌داده و یا قابل محاسبه از آن از چهار ویژگی:

۱) human_development_index[1]
۲) gdp_per_capita[2]
۳) extreme_poverty[3]

---

[1] شاخص توسعه فردی
[2] تولید ناخالص ملی
[3] فقر مطلق

۴) life_expectancy[4]

استفاده کرده و با کمک آن‌ها صفت total_economy را به مجموعه‌داده اضافه می‌کنیم. نحوه محاسبه total_economy به صورت زیر است:

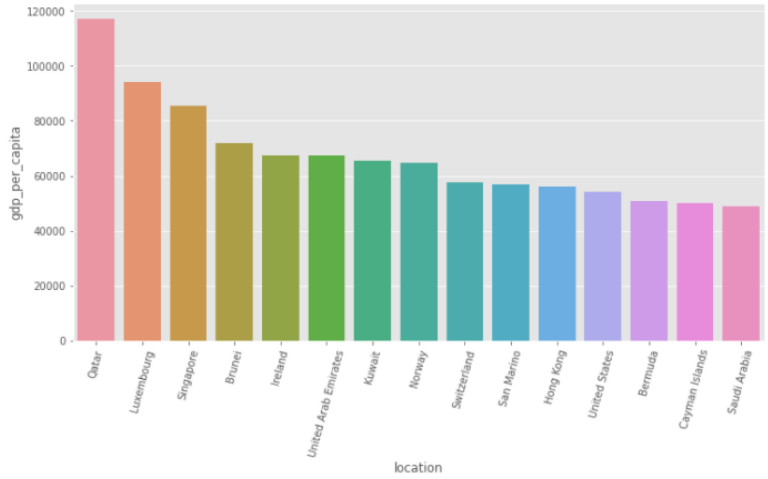
```
div_val = (4)*(52800)
df_27["total_economy"] = ((df_27["extreme_poverty"] * 0.1) *
                          (df_27["gdp_per_capita"] * 0.0001) *
                          (df_27["life_expectancy"] * 3) *
                          (df_27["human_development_index"] * 176))/div_val
```

ابتدا هریک از این شاخص‌ها را به صورت تجمیع شده برای هر کشور محاسبه کرده و یک دیتافریم برایشان شکل می‌دهیم و سپس برای شاخص‌ها نمودار اسکترپلات رسم می‌کنیم.

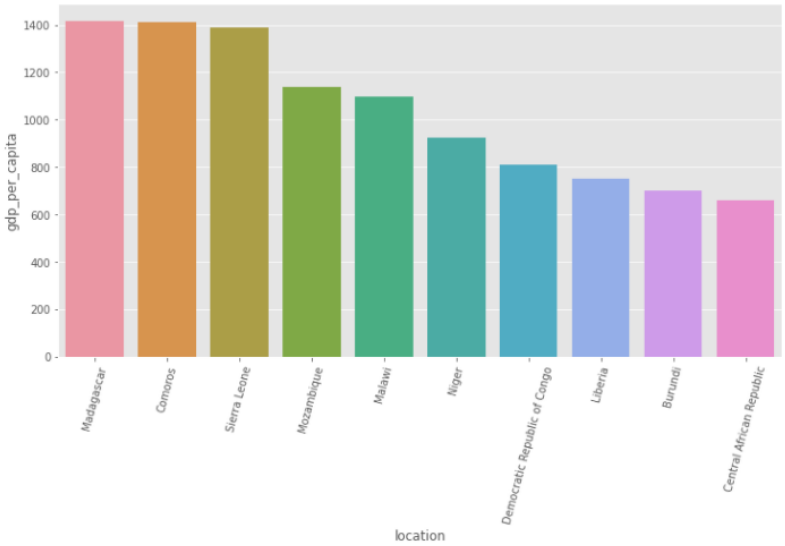
| | people_vaccinated_per_hundred | people_fully_vaccinated_per_hundred | human_development_index | gdp_per_capita |
|---|---|---|---|---|
| | max | max | max | max |
| location | | | | |
| Gibraltar | 124.57 | 121.45 | 0.725595 | 19645.934476 |
| Pitcairn | 100.00 | 100.00 | 0.725595 | 19645.934476 |
| United Arab Emirates | 98.99 | 95.13 | 0.890000 | 67293.483000 |
| Portugal | 95.02 | 92.51 | 0.864000 | 27936.896000 |
| Brunei | 92.31 | 91.57 | 0.838000 | 71809.251000 |
| Singapore | 91.48 | 90.25 | 0.938000 | 85535.383000 |
| Cayman Islands | 92.49 | 89.82 | 0.725595 | 49903.029000 |
| Malta | 91.22 | 89.77 | 0.895000 | 36513.323000 |
| Chile | 92.53 | 89.58 | 0.851000 | 22767.037000 |
| Cuba | 93.74 | 87.32 | 0.783000 | 19645.934476 |
| South Korea | 87.44 | 86.49 | 0.916000 | 35938.374000 |
| China | 87.89 | 85.48 | 0.761000 | 15308.712000 |
| Spain | 87.87 | 83.56 | 0.904000 | 34272.360000 |
| Faeroe Islands | 85.04 | 83.37 | 0.725595 | 19645.934476 |
| Cambodia | 86.41 | 81.77 | 0.594000 | 3645.070000 |
| Denmark | 83.33 | 81.58 | 0.940000 | 46682.515000 |
| Canada | 85.59 | 81.24 | 0.929000 | 44017.591000 |
| Guernsey | 85.42 | 81.01 | 0.725595 | 19645.934476 |
| Seychelles | 85.00 | 80.70 | 0.796000 | 26382.287000 |
| Niue | 83.02 | 79.80 | 0.725595 | 19645.934476 |
| Ireland | 81.35 | 79.73 | 0.955000 | 67335.293000 |
| Japan | 80.70 | 79.53 | 0.919000 | 39002.223000 |
| Australia | 85.58 | 79.48 | 0.944000 | 44648.710000 |

۱۰ کشور با بالاترین نرخ تولید ناخالص ملی:

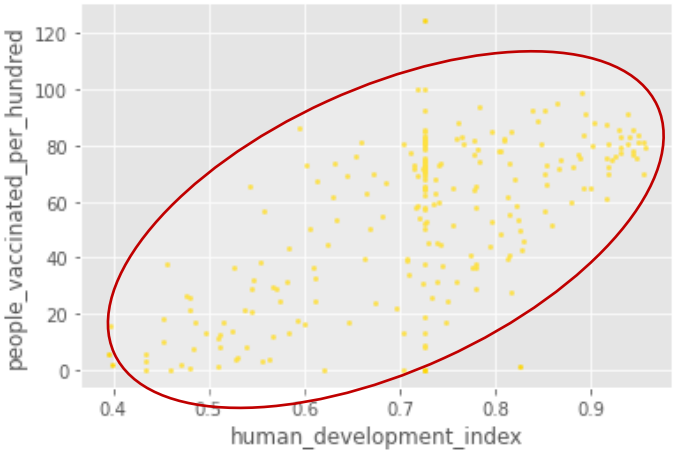
[4] امید به زندگی




۱۰ کشور با کمترین تولید نرخ ناخالص ملی

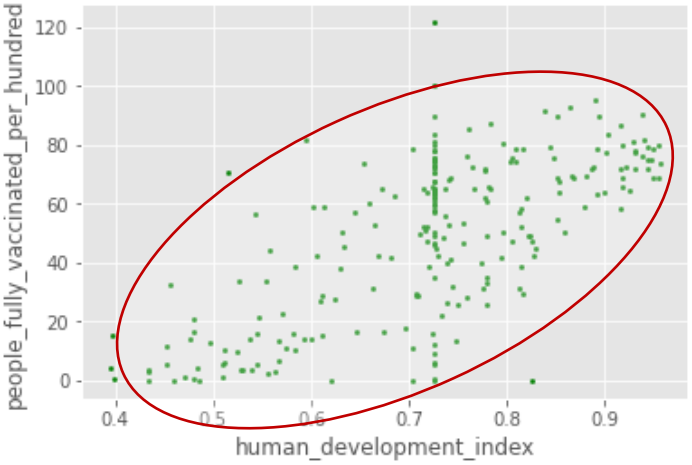



**شاخص توسعه فردی کشور به تعداد افراد واکسینه شده کشور**



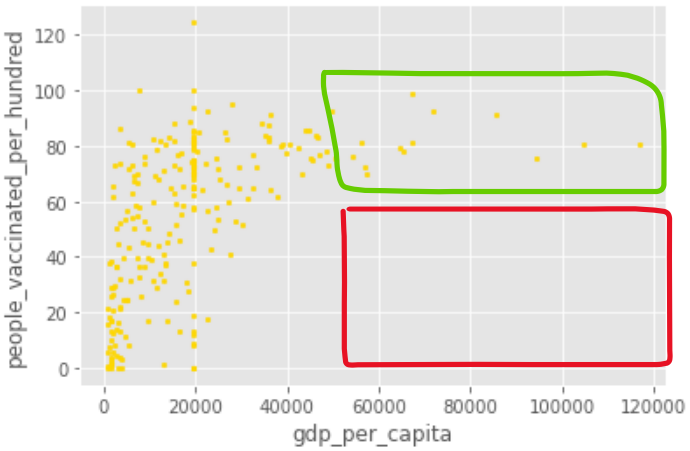

در کشورهایی با توسعه فردی شاهد هستیم که تعداد افرادی که واکسن دریافت کرده اند هم به نسبت جمعیت بیشتر از کشورهای ضعیف‌تر از نظر این مشخصه است.

**شاخص توسعه فردی کشور به تعداد افراد کاملاً واکسینه شده کشور (اقلاً دو دوز یا بیشتر دریافت کرده‌اند)**



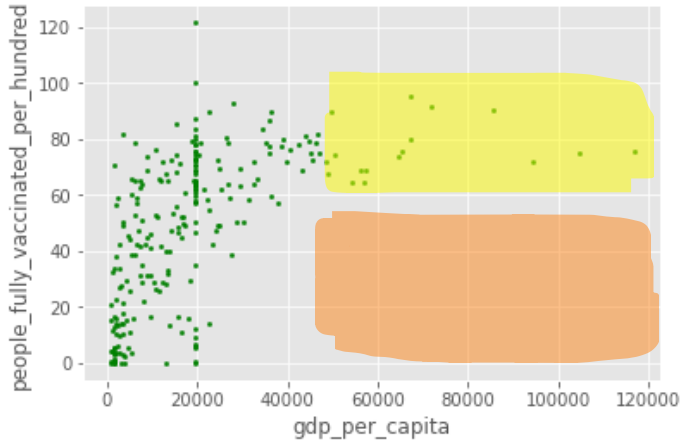

شاخص توسعه فردی با واکسینه شدن کامل افراد هم رابطه مستقیم دارد.

**تولید ناخالص ملی کشور به تعداد افراد واکسینه شده کشور**



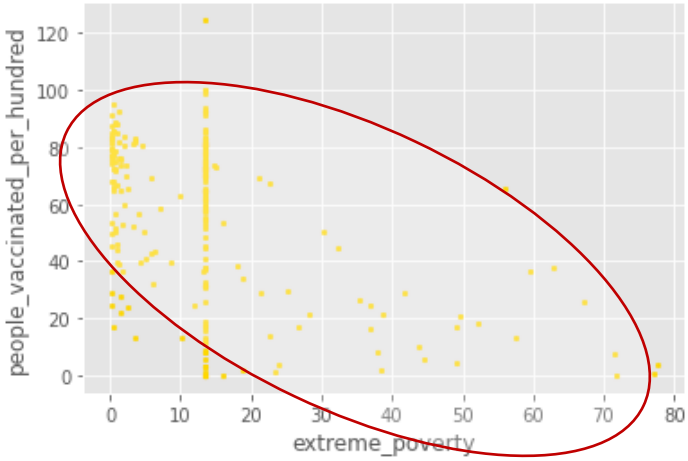

از این نمودار برداشتی که می‌توان کرد این است که داشتن شاخص تولید ناخالص ملی باعث کمتر گرفتن مشهود واکسن نیست اما کشورهایی که تولید ناخاص ملی بالایی دارند همه خوب واکسن زده‌اند.

**تولید ناخالص ملی کشور به تعداد افراد کاملاً واکسینه شده کشور (اقلاً دو دوز یا بیشتر دریافت کرده‌اند)**



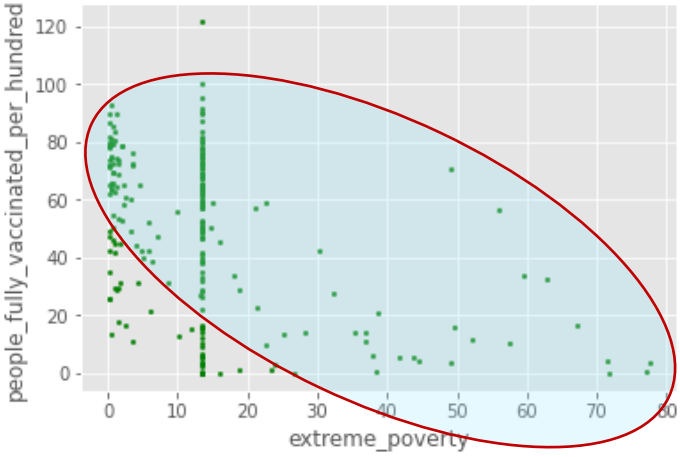
تحلیل این نمودار هم مثل نمودار قبلی است.

**فقر مطلق در کشور به تعداد افراد واکسینه شده کشور**



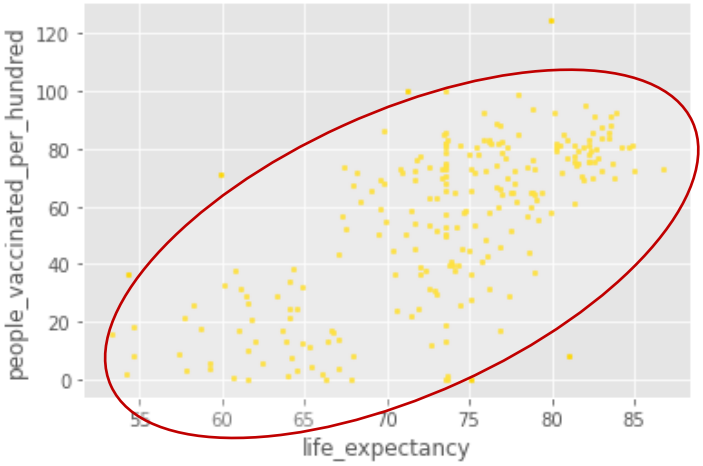

این شکل متاسفانه نشان می‌دهد که کشورهایی که فقیرتر هستند واکسن کمتری دریافت کرده‌اند.

**فقر مطلق به تعداد افراد کاملاً واکسینه شده کشور (اقلاً دو دوز یا بیشتر دریافت کرده‌اند)**



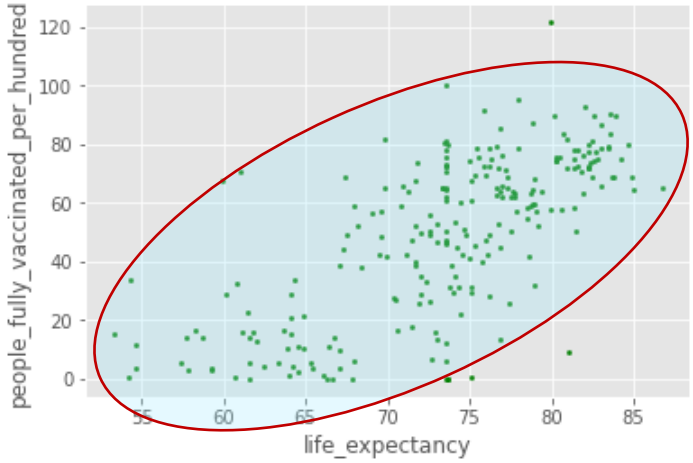

کشورهایی که واکسن کمتری به نسبت جمعیت زده اند فقیرتر هم هستند.

**امید به زندگی در کشور به تعداد افراد واکسینه شده کشور**




کشورهایی که امید به زندگی بیشتری دارند که اغلب مرفه‌تر و ثروتمندتر هم هستند واکسیناسیون بیشتری انجام داده‌اند.

**امید به زندگی در کشور به تعداد افراد کاملاً واکسینه شده کشور (اقلاً دو دوز یا بیشتر دریافت کرده‌اند)**




مردم در کشورهایی با امید به زندگی بیشتر دسترسی بیشتری به واکسن داشته‌اند.

## نمایش دادگان: ۸

در سال ۲۰۲۱ توزیع تعداد مبتلایان به تفکیک ماه را بررسی نمایید و تحلیل خود را ذکر نمایید.

ابتدا داده‌های سال ۲۰۲۱ را از مجموعه داده جدا کرده و یک دیتافریم جدید شکل می‌دهیم:

## 2-8

In [562]:

```
df_28 = test_df
df_28 = df_28[df_28['g_year'] == 2021]
df_28
```

executed in 220ms, finished 13:39:51 2022-04-08

| | iso_code | continent | location | date | total_cases | new_cases | new_cases_smoothed | total_deaths | new_deaths |
|---|---|---|---|---|---|---|---|---|---|
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 312 | AFG | Asia | Afghanistan | 2021-01-01 | 52513.0 | 183.0 | 131.143 | 2201.0 | 12.0 |
| 313 | AFG | Asia | Afghanistan | 2021-01-02 | 52586.0 | 73.0 | 117.429 | 2211.0 | 10.0 |
| 314 | AFG | Asia | Afghanistan | 2021-01-03 | 52709.0 | 123.0 | 123.000 | 2221.0 | 10.0 |
| 315 | AFG | Asia | Afghanistan | 2021-01-04 | 52909.0 | 200.0 | 128.857 | 2230.0 | 9.0 |
| 316 | AFG | Asia | Afghanistan | 2021-01-05 | 53011.0 | 102.0 | 123.429 | 2237.0 | 7.0 |
| 165570 | ZWE | Africa | Zimbabwe | 2021-12-27 | 205449.0 | 1098.0 | 1481.429 | 4908.0 | 17.0 |
| 165571 | ZWE | Africa | Zimbabwe | 2021-12-28 | 207548.0 | 2099.0 | 1397.143 | 4940.0 | 32.0 |
| 165572 | ZWE | Africa | Zimbabwe | 2021-12-29 | 207548.0 | 0.0 | 1163.429 | 4940.0 | 0.0 |
| 165573 | ZWE | Africa | Zimbabwe | 2021-12-30 | 211728.0 | 4180.0 | 1483.429 | 4997.0 | 57.0 |
| 165574 | ZWE | Africa | Zimbabwe | 2021-12-31 | 213258.0 | 1530.0 | 1503.143 | 5004.0 | 7.0 |

84758 rows × 71 columns

سپس این دیتافریم را بر اساس ماه برای تعداد موارد ابتلای جدید تجمیع می‌کنیم:

```
In [565]:
df_28 = df_28.groupby('g_month').agg({'new_cases': ['sum']})
df_28
```

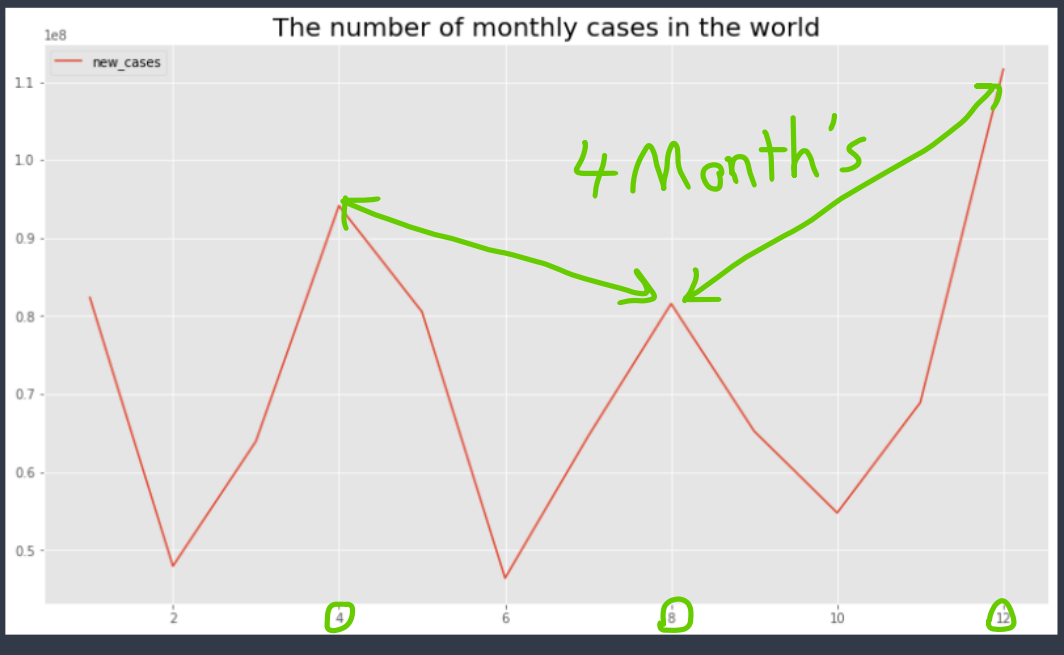
executed in 28ms, finished 13:42:27 2022-04-08

| | new_cases |
|---|---|
| | sum |
| g_month | |
| 1 | 82370562.0 |
| 2 | 47970396.0 |
| 3 | 63963191.0 |
| 4 | 94132024.0 |
| 5 | 80552750.0 |
| 6 | 46436759.0 |
| 7 | 64590165.0 |
| 8 | 81587128.0 |
| 9 | 65264545.0 |
| 10 | 54795747.0 |
| 11 | 68912535.0 |
| 12 | 111619153.0 |

در آخر نمودار ابتلای جهانی به صورت ماهیانه برای سال ۲۰۲۱ را رسم می‌کنیم.

```
In [567]:
plt.figure(num = 3, figsize=(14, 8))
plt.plot(df_28["new_cases"], label="new_cases")
plt.title('The number of monthly cases in the world', fontsize=20)
plt.legend()
plt.show()
```

executed in 287ms, finished 13:43:35 2022-04-08



بر اساس این نمودار شاهد هستیم که هر ۴ ماه یک بار تقریباً یک پیک از نظر شیوع بیماری داریم.

## بخش امتیازی

تعداد فوتی‌های سه ماه اخیر کشورهای مختلف را به نسبت جمعیت آن‌ها بر روی نقشه نمایش دهید.

برای این کار از کتابخانه plotly استفاده کردیم. ابتدا دیتافریم مخصوص را ایجاد کرده:

```
df_extra = test_df.groupby('location').agg({'total_deaths_per_million': ['max'],
                                            'iso_code': ['max'],
                                            'location': ['max']
                                               })
df_extra
```

executed in 780ms, finished 23:19:22 2022-04-08

| | total_cases_per_million | total_deaths_per_million | iso_code | location |
|---|---|---|---|---|
| | max | max | max | max |
| location | | | | |
| Afghanistan | 4369.804 | 191.212 | AFG | Afghanistan |
| Albania | 94615.818 | 1209.217 | ALB | Albania |
| Algeria | 5941.261 | 153.373 | DZA | Algeria |
| Andorra | 494466.996 | 1952.065 | AND | Andorra |
| Angola | 2909.976 | 55.992 | AGO | Angola |
| Anguilla | 168925.620 | 595.041 | AIA | Anguilla |
| Antigua and Barbuda | 75449.720 | 1367.393 | ATG | Antigua and Barbuda |
| Argentina | 195420.594 | 2771.357 | ARG | Argentina |
| Armenia | 141671.114 | 2861.400 | ARM | Armenia |
| Aruba | 314231.074 | 1968.375 | ABW | Aruba |
| Australia | 127840.711 | 206.141 | AUS | Australia |
| Austria | 303439.252 | 1646.343 | AUT | Austria |
| Azerbaijan | 77016.581 | 924.746 | AZE | Azerbaijan |
| Bahamas | 83519.352 | 1942.486 | BHS | Bahamas |
| Bahrain | 297194.695 | 856.835 | BHR | Bahrain |
| Bangladesh | 11696.134 | 174.699 | BGD | Bangladesh |
| Barbados | 193053.374 | 1098.336 | BRB | Barbados |

سپس با کتابخانه ذکر شده نقشه را می‌سازیم:

In [753]:

```
choropleth_map = go.Figure(
    data = {
        'type':'choropleth',
        'locations':df_extra['location'],
        'locationmode':'country names',
#         'colorscale':'Portland',
        'z':df_extra['total_deaths_per_million'],
        'colorbar':{'title':'total deaths per million'},
        'marker': {
            'line': {
                'color':'rgb(255,255,255)',
                'width':2
            }
        }
    },
    layout = {
      'geo':{
          'scope':'world',
      }
    })
choropleth_map
```

متاسفانه در آخرین ران کردن‌ها که می‌خواستیم ظاهر و رنگ‌بندی نقشه را بهتر کنیم، دیگر رنگ‌ها ظاهر نشدند. متوجه نشدم که چه چیزی کم یا زیاد شده و تغییر کرده. نهایتاً الآن نقشه به صورت زیر است.