DAC Phase 3:

Problem Statement: COVID Vaccines Analysis

Loading and Pre-processing of data:

```
from google.colab import drive
drive.mount('/content/drive/')
Loading data
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import PolynomialFeatures
from sklearn import metrics
from sklearn.metrics import mean_squared_error
from sklearn.metrics import r2_score
from sklearn.tree import DecisionTreeRegressor
import xgboost as xgb
from sklearn.cluster import KMeans
from sklearn.model_selection import cross_val_score, KFold

cov19=pd.read_csv('/content/drive/MyDrive/dataset/country_vaccinations.csv')
cov19
```

	country	iso_code	date	total_vaccinations	people_vaccinated	people_fully_vaccinated	daily_vaccinations_raw	daily_vaccinations	total_vaccinations_
0	Afghanistan	AFG	2021-02-22	0.0	0.0	NaN	NaN	NaN	
1	Afghanistan	AFG	2021-02-23	NaN	NaN	NaN	NaN	1367.0	
2	Afghanistan	AFG	2021-02-24	NaN	NaN	NaN	NaN	1367.0	
3	Afghanistan	AFG	2021-02-25	NaN	NaN	NaN	NaN	1367.0	
4	Afghanistan	AFG	2021-02-26	NaN	NaN	NaN	NaN	1367.0	
...	...	...	...	...	...	...	...	...	
86507	Zimbabwe	ZWE	2022-03-25	8691642.0	4814582.0	3473523.0	139213.0	69579.0	

cov19.describe()

	total_vaccinations	people_vaccinated	people_fully_vaccinated	daily_vaccinations_raw	daily_vaccinations	total_vaccinations_per_hundred	people_vaccinat
count	4.360700e+04	4.129400e+04	3.880200e+04	3.536200e+04	8.621300e+04	43607.000000	
mean	4.592964e+07	1.770508e+07	1.413830e+07	2.705996e+05	1.313055e+05	80.188543	
std	2.246004e+08	7.078731e+07	5.713920e+07	1.212427e+06	7.682388e+05	67.913577	
min	0.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	0.000000	
25%	5.264100e+05	3.494642e+05	2.439622e+05	4.668000e+03	9.000000e+02	16.050000	
50%	3.590096e+06	2.187310e+06	1.722140e+06	2.530900e+04	7.343000e+03	67.520000	
75%	1.701230e+07	9.152520e+06	7.559870e+06	1.234925e+05	4.409800e+04	132.735000	
max	3.263129e+09	1.275541e+09	1.240777e+09	2.474100e+07	2.242429e+07	345.370000	

This command is used to view the brief summary of the dataset. We can see the mathematical parameters such as percentiles, standard deviation , mean, minimum and maximum values and count of each column.

cov19.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 86512 entries, 0 to 86511
Data columns (total 15 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   country                             86512 non-null  object
 1   iso_code                            86512 non-null  object
 2   date                                86512 non-null  object
 3   total_vaccinations                  43607 non-null  float64
 4   people_vaccinated                   41294 non-null  float64
 5   people_fully_vaccinated             38802 non-null  float64
 6   daily_vaccinations_raw              35362 non-null  float64
 7   daily_vaccinations                  86213 non-null  float64
 8   total_vaccinations_per_hundred      43607 non-null  float64
 9   people_vaccinated_per_hundred       41294 non-null  float64
 10  people_fully_vaccinated_per_hundred 38802 non-null  float64
 11  daily_vaccinations_per_million      86213 non-null  float64
 12  vaccines                            86512 non-null  object
 13  source_name                         86512 non-null  object
 14  source_website                      86512 non-null  object
dtypes: float64(9), object(6)
memory usage: 9.9+ MB
```

Info command is used check the datatype of every column and the count of each column. The difference between the describe() and info() is that describe command will give the mathematical parameters but info command will not give the mathematical parameters such as mean and standard deviation

Data Preprocessing

cov19.isnull().sum()

```
country                                0
iso_code                               0
date                                   0
total_vaccinations                 42905
people_vaccinated                  45218
people_fully_vaccinated            47710
daily_vaccinations_raw             51150
daily_vaccinations                   299
total_vaccinations_per_hundred     42905
people_vaccinated_per_hundred      45218
people_fully_vaccinated_per_hundred    47710
daily_vaccinations_per_million       299
vaccines                               0
source_name                            0
source_website                         0
dtype: int64
```

cov19_fillna = cov19

cov19_fillna

	country	iso_code	date	total_vaccinations	people_vaccinated	people_fully_vaccinated	daily_vaccinations_raw	daily_vaccinations	total_vaccinations_per
0	0	1	2021 02 22	0.000000e+00	0.000000e+00	1.413830e+07	270599.578248	131305.486075	
1	0	1	2021 02 23	4.592964e+07	1.770508e+07	1.413830e+07	270599.578248	1367.000000	8
2	0	1	2021 02 24	4.592964e+07	1.770508e+07	1.413830e+07	270599.578248	1367.000000	8
3	0	1	2021 02 25	4.592964e+07	1.770508e+07	1.413830e+07	270599.578248	1367.000000	8
4	0	1	2021 02 26	4.592964e+07	1.770508e+07	1.413830e+07	270599.578248	1367.000000	8
...	...	...	...	...	...	...	...	...	
86507	222	222	2022 03 25	8.691642e+06	4.814582e+06	3.473523e+06	139213.000000	69579.000000	5
86508	222	222	2022 03	8.791728e+06	4.886242e+06	3.487962e+06	100086.000000	83429.000000	5

cov19_fillna.fillna(cov19_fillna.mean(), inplace=True)

count the number of NaN values in each column

print(cov19_fillna.isnull().sum())

cov19_fillna

```
country                              0
iso_code                             0
date                                 0
total_vaccinations                   0
people_vaccinated                    0
people_fully_vaccinated              0
daily_vaccinations_raw               0
daily_vaccinations                   0
total_vaccinations_per_hundred       0
people_vaccinated_per_hundred        0
people_fully_vaccinated_per_hundred  0
daily_vaccinations_per_million       0
vaccines                             0
source_name                          0
source_website                       0
dtype: int64
<ipython-input-9-9e428849e60a>:1: FutureWarning: The default value of numeric_only in DataFrame.mean is deprecated. In a future version, it will default to Fals
  cov19_fillna.fillna(cov19_fillna.mean(), inplace=True)
```

	country	iso_code	date	total_vaccinations	people_vaccinated	people_fully_vaccinated	daily_vaccinations_raw	daily_vaccinations	total_vaccination
0	Afghanistan	AFG	2021-02-22	0.000000e+00	0.000000e+00	1.413830e+07	270599.578248	131305.486075	
1	Afghanistan	AFG	2021-02-23	4.592964e+07	1.770508e+07	1.413830e+07	270599.578248	1367.000000	

✓ 0s completed at 11:11 PM

le=LabelEncoder()

cov19['country']=le.fit_transform(cov19['country'])

cov19

	country	iso_code	date	total_vaccinations	people_vaccinated	people_fully_vaccinated	daily_vaccinations_raw	daily_vaccinations	total_vaccinations_p
0	0	AFG	2021-02-22	0.000000e+00	0.000000e+00	1.413830e+07	270599.578248	131305.486075	
1	0	AFG	2021-02-23	4.592964e+07	1.770508e+07	1.413830e+07	270599.578248	1367.000000	
2	0	AFG	2021-02-24	4.592964e+07	1.770508e+07	1.413830e+07	270599.578248	1367.000000	
3	0	AFG	2021-02-25	4.592964e+07	1.770508e+07	1.413830e+07	270599.578248	1367.000000	
4	0	AFG	2021-02-26	4.592964e+07	1.770508e+07	1.413830e+07	270599.578248	1367.000000	
...	...	...	...	...	...	...	...	...	
86507	222	ZWE	2022-03-25	8.691642e+06	4.814582e+06	3.473523e+06	139213.000000	69579.000000	

✓ 0s completed at 11:11 PM

le=LabelEncoder()

cov19['iso_code']=le.fit_transform(cov19['iso_code'])

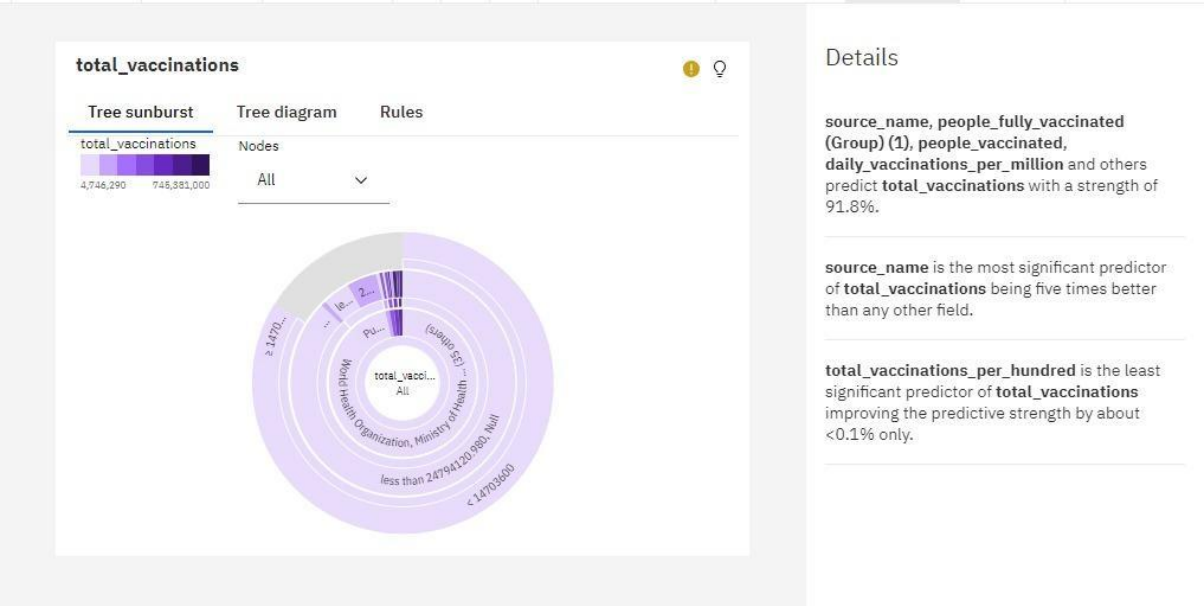
cov19

	country	iso_code	date	total_vaccinations	people_vaccinated	people_fully_vaccinated	daily_vaccinations_raw	daily_vaccinations	total_vaccinations_per
0	0	1	2021-02-22	0.000000e+00	0.000000e+00	1.413830e+07	270599.578248	131305.486075	
1	0	1	2021-02-23	4.592964e+07	1.770508e+07	1.413830e+07	270599.578248	1367.000000	
2	0	1	2021-02-24	4.592964e+07	1.770508e+07	1.413830e+07	270599.578248	1367.000000	
3	0	1	2021-02-25	4.592964e+07	1.770508e+07	1.413830e+07	270599.578248	1367.000000	
4	0	1	2021-02-26	4.592964e+07	1.770508e+07	1.413830e+07	270599.578248	1367.000000	
...	...	...	...	...	...	...	...	...	
86507	222	222	2022-03-25	8.691642e+06	4.814582e+06	3.473523e+06	139213.000000	69579.000000	

0s completed at 11:11 PM

cov19.columns

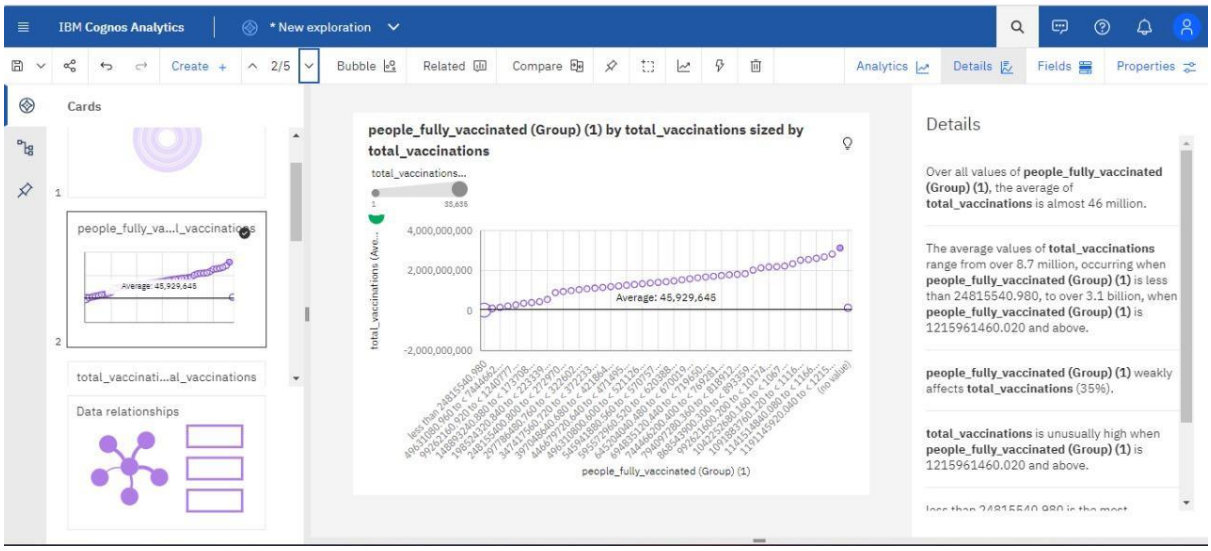
Index(['country', 'iso_code', 'date', 'total_vaccinations', 'people_vaccinated', 'people_fully_vaccinated', 'daily_vaccinations_raw', 'daily_vaccinations', 'total_vaccinations_per_hundred', 'people_vaccinated_per_hundred', 'people_fully_vaccinated_per_hundred', 'daily_vaccinations_per_million', 'vaccines', 'source_name', 'source_website'], dtype='object')

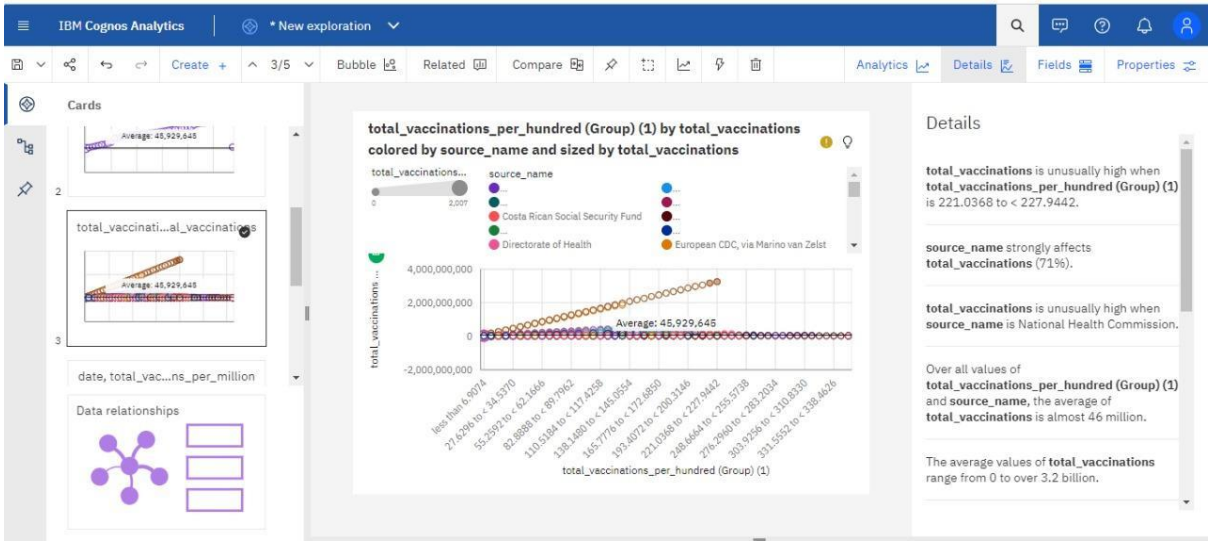


source_name, **people_fully_vaccinated (Group) (1)**, **people_vaccinated**, **daily_vaccinations_per_million** and others predict **total_vaccinations** with a strength of 91.8%.

source_name is the most significant predictor of **total_vaccinations** being five times better than any other field.

total_vaccinations_per_hundred is the least significant predictor of **total_vaccinations** improving the predictive strength by about <0.1% only.





total_vaccinations is unusually high when **total_vaccinations_per_hundred (Group) (1)** is 221.0368 to < 227.9442.

source_name strongly affects **total_vaccinations** (71%).

total_vaccinations is unusually high when **source_name** is National Health Commission.

Over all values of **total_vaccinations_per_hundred (Group) (1)** and **source_name**, the average of **total_vaccinations** is almost 46 million.

The average values of **total_vaccinations** range from 0 to over 3.2 billion.

total_vaccinations_per_hundred (Group) (1) and **source_name** strongly affect **total_vaccinations** (100%).

total_vaccinations is unusually high when the combinations of **total_vaccinations_per_hundred (Group) (1)** and **source_name** are 221.0368 to < 227.9442 and National Health Commission and 214.1294 to < 221.0368 and National Health Commission.

less than 6.9074 is the most frequently occurring category of **total_vaccinations_per_hundred (Group) (1)** with a count of 7505 items with **total_vaccinations** values (17.2 % of the total).

Ministry of Health is the most frequently occurring category of **source_name** with a count of 9981 items with **total_vaccinations** values (22.9 % of the total).

Chart A

date - Top 10 by daily_vaccinations_per_million

date, total_vaccinations and daily_vaccinations_per_million

5

date

total_vaccinations

daily_vaccinations_per_million

6/22/2021

2,699,790,526

965,713

6/23/2021

2,788,620,339

954,815

6/26/2021

2,877,147,766

954,034

6/28/2021

2,996,944,602

951,522

Chart B

daily_vaccinations and total_vaccinations by country colored by country

10,562,357

2 of 200 items

Summary	Select Chart A : total_vaccinationsChart A : daily_vaccinations_per_million	Select Chart B : daily_vaccinationsChart B : total_vaccinations	Combined
Chart percent of data set	1.72%	100%	-
Average	3,434,983,805.7	50,763,407.49	-
Chart total	34,349,838,057	11,320,239,871	-

people_fully_vaccinated (Group) (1) by total_vaccinations sized by total_vaccinations

less than 24815540.98049631080.960 to < 7444662...99262160.920 to < 1240777...148893240.880 to < 173708...198524320.840 to < 223339...248155400.800 to < 272970...297786480.760 to < 322602...347417560.720 to < 372233...397048640.680 to < 421864...446679720.640 to < 471495...496310800.600 to < 521126...545941880.560 to < 570757...595572960.520 to < 620388...645204040.480 to < 670019...694835120.440 to < 719650...744466200.400 to < 769281...794097280.360 to < 818912...868543900.300 to < 893359...992621600.200 to < 10174...1042252680.160 to < 1067...1091883760.120 to < 1116...1141514840.080 to < 1166...1191145920.040 to < 1215...(no value)people_fully_vaccinated (Group) (1)-2,000,000,00002,000,000,0004,000,000,000total_vaccinations (Ave…5Average: 45,929,645

total_vaccinations (Count)

133,635

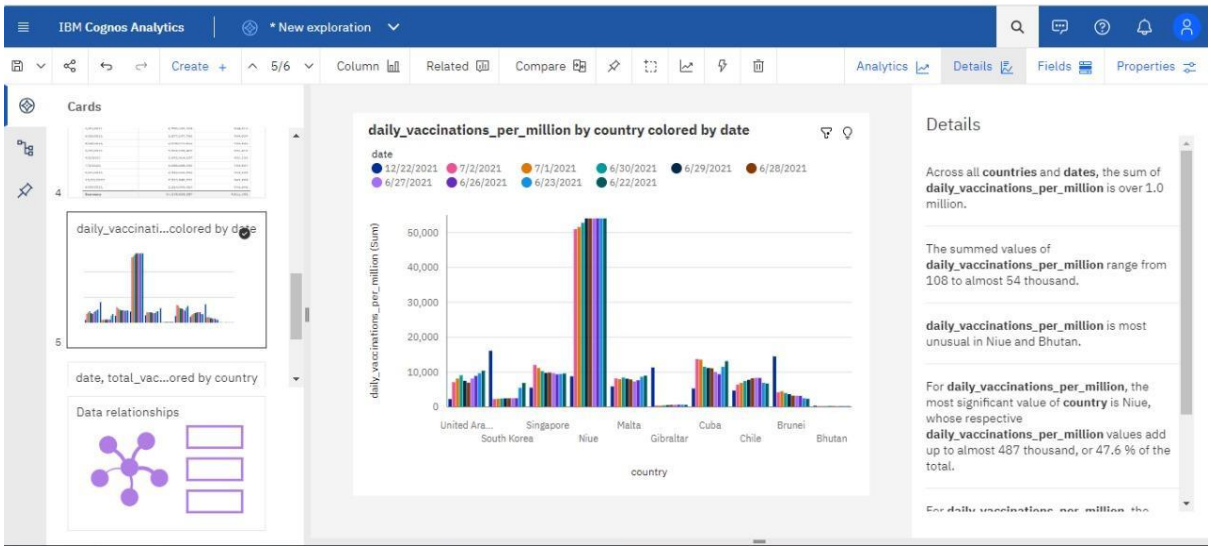
daily_vaccinations_per_million by country colored by date

United Arab...South KoreaSingaporeNiueMaltaGibraltarCubaChileBruneiBhutancountry010,00020,00030,00040,00050,000daily_vaccinations_per_million (Sum)

date

- 12/22/2021
- 7/2/2021
- 7/1/2021
- 6/30/2021
- 6/29/2021
- 6/28/2021
- 6/27/2021

- 6/26/2021
- 6/23/2021
- 6/22/2021



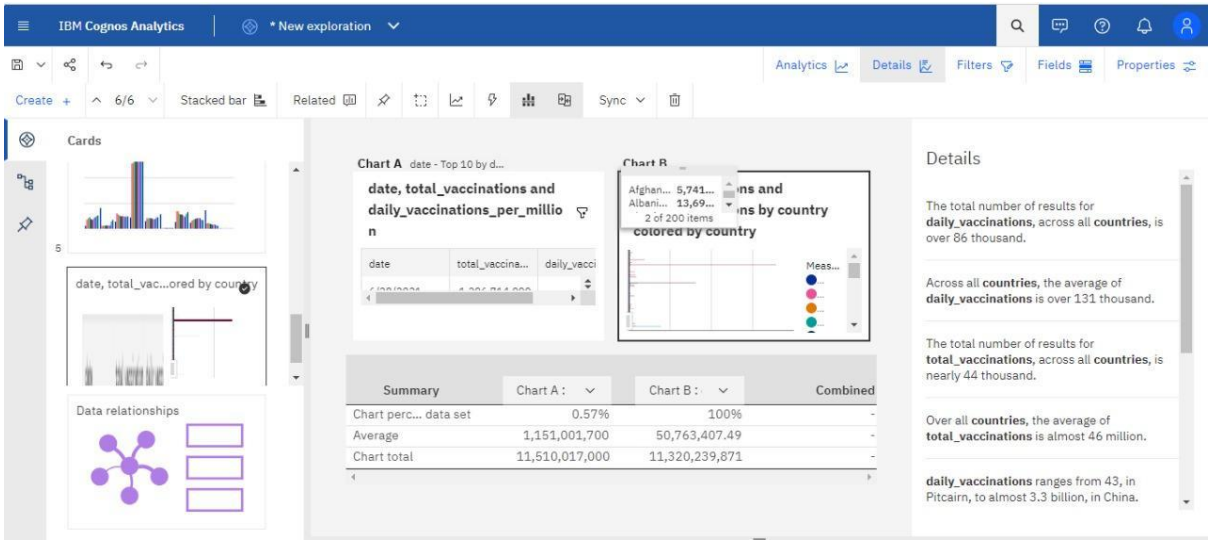
Across all **countries** and **dates**, the sum of **daily_vaccinations_per_million** is over 1.0 million.

The summed values of **daily_vaccinations_per_million** range from 108 to almost 54 thousand.

daily_vaccinations_per_million is most unusual in Niue and Bhutan.

For **daily_vaccinations_per_million**, the most significant value of **country** is Niue, whose respective **daily_vaccinations_per_million** values add up to almost 487 thousand, or 47.6 % of the total.

For **daily_vaccinations_per_million**, the most significant values of **date** are 2021-06-22, 2021-06-23, 2021-07-01, 2021-06-30, and 2021-07-02, whose respective **daily_vaccinations_per_million** values add up to over 535 thousand, or 52.3 % of the total.



The total number of results for **daily_vaccinations**, across all **countries**, is over 86 thousand.

Across all **countries**, the average of **daily_vaccinations** is over 131 thousand.

The total number of results for **total_vaccinations**, across all **countries**, is nearly 44 thousand.

Over all **countries**, the average of **total_vaccinations** is almost 46 million.

daily_vaccinations ranges from 43, in Pitcairn, to almost 3.3 billion, in China.

total_vaccinations ranges from 348, in Pitcairn, to approximately 709 billion, in China.

Norway (0.6 %), Latvia (0.6 %), and Denmark (0.6 %) are the most frequently occurring categories of **country** with a combined count of 1435 items with **daily_vaccinations** values (1.7 % of the total).

Norway is the most frequently occurring category of **country** with a count of 482 items with **total_vaccinations** values (1.1 % of the total).

date, total_vaccinations and daily_vaccinations_per_million 5

date	total_vaccinations	daily_vaccinations_per_million
6/22/2021	2,699,790,526	965,713
6/23/2021	2,788,620,339	954,815
6/26/2021	2,877,147,766	954,034
6/28/2021	2,996,944,602	951,522
6/30/2021	3,062,159,402	951,412
7/2/2021	3,072,014,637	951,132
7/1/2021	3,085,188,933	950,829
6/27/2021	2,942,024,392	944,228
12/22/2021	7,810,948,031	943,909
6/29/2021	3,014,999,429	943,898
Summary	34,349,838,057	9,511,492