**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

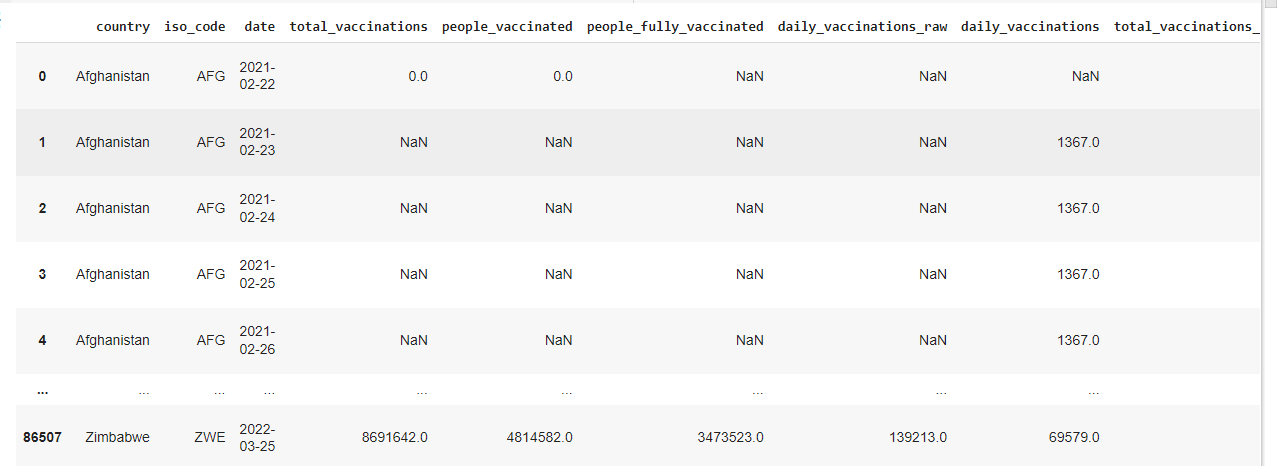
import xgboost as xgb

from sklearn.cluster import KMeans

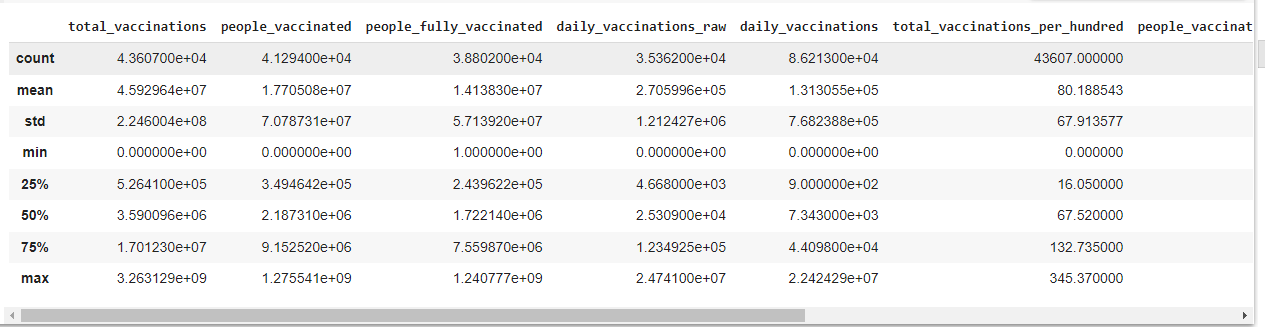
from sklearn.model\_selection import cross\_val\_score, KFold

cov19=pd.read\_csv('/content/drive/MyDrive/dataset/country\_vaccinations.csv')

cov19

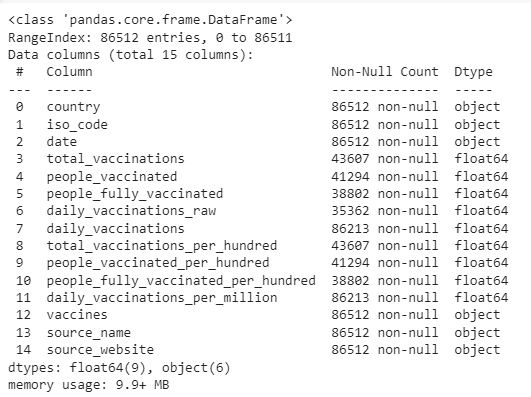


cov19.describe()



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.

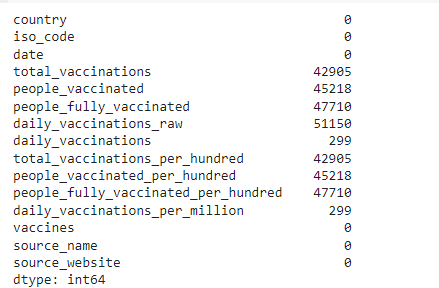
cov19.info()



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

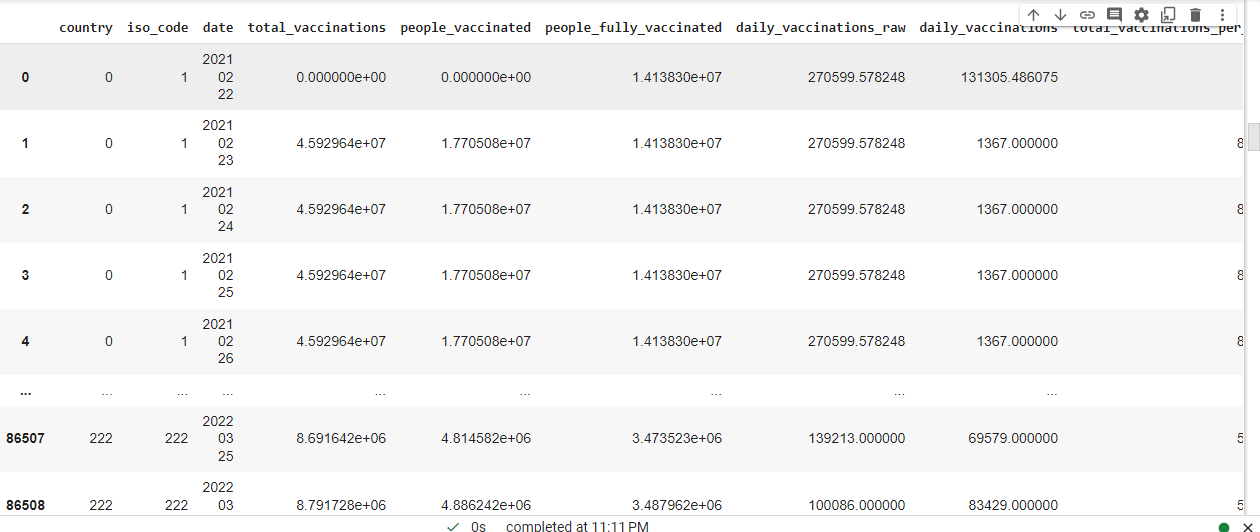
**Data Preprocessing:**

cov19.isnull().sum()



cov19\_fillna = cov19

cov19\_fillna

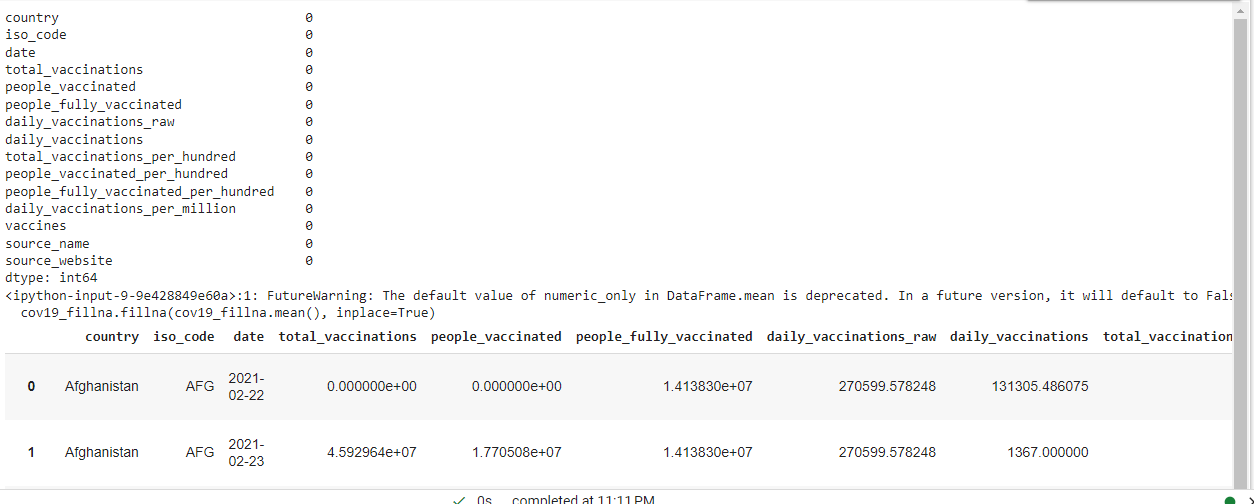


cov19\_fillna.fillna(cov19\_fillna.mean(), inplace=True)

# count the number of NaN values in each column

print(cov19\_fillna.isnull().sum())

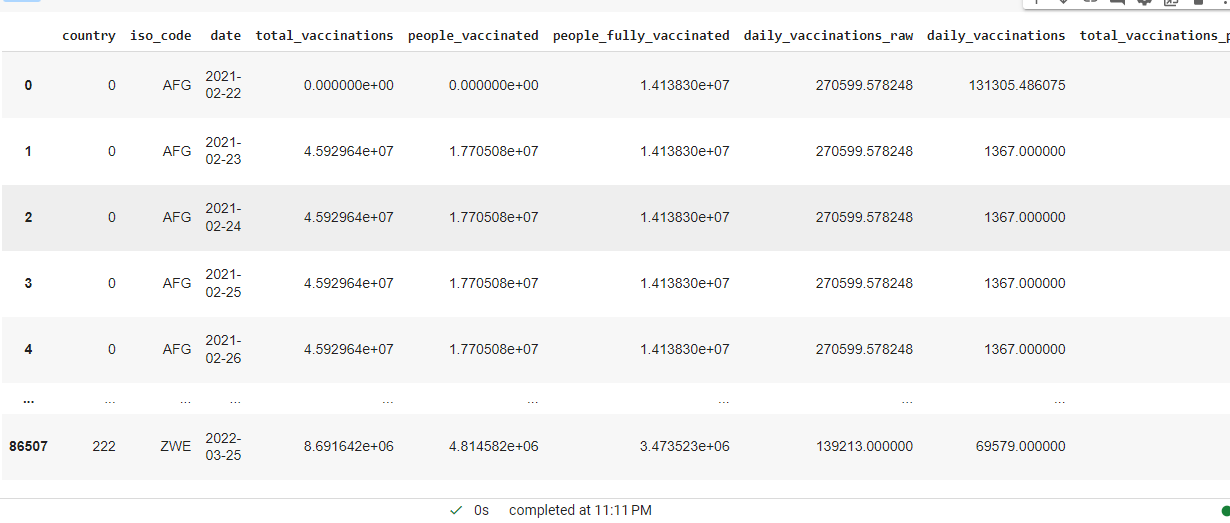
cov19\_fillna



le=LabelEncoder()

cov19['country']=le.fit\_transform(cov19['country'])

cov19



le=LabelEncoder()

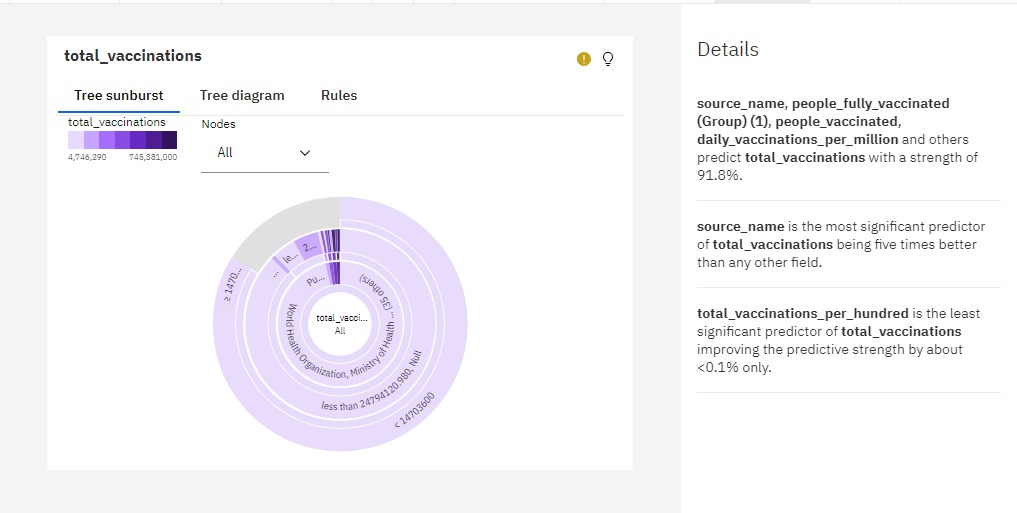
cov19['iso\_code']=le.fit\_transform(cov19['iso\_code'])

cov19



cov19.columns

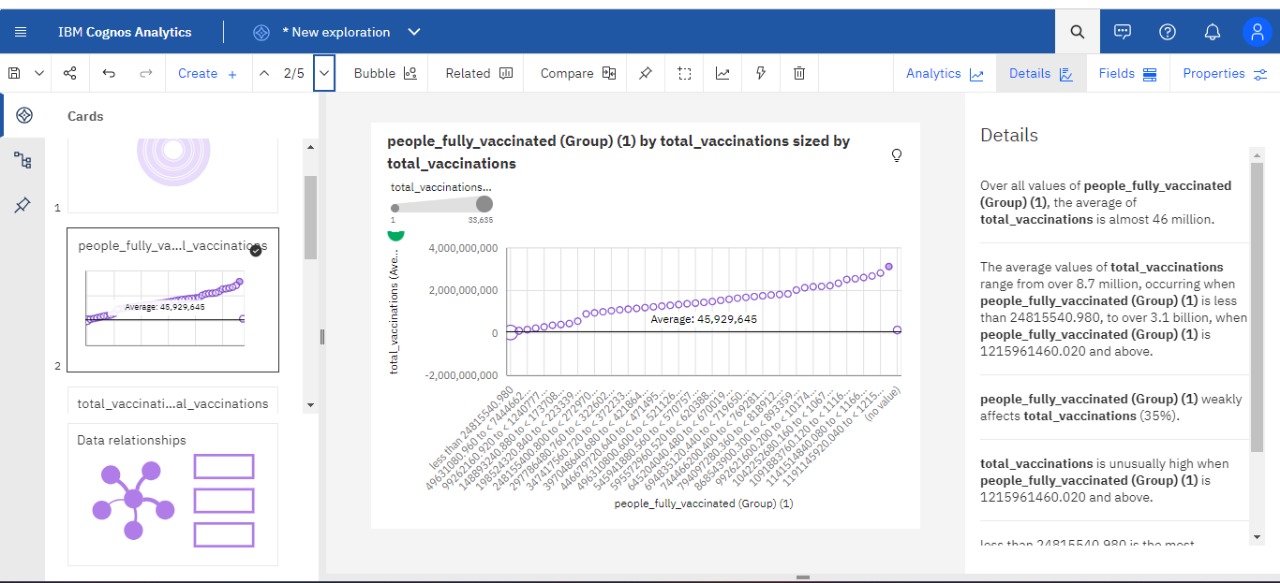
Index(['country', 'iso\_code', 'date', 'total\_vaccinations', 'people\_vaccinated', 'people\_fully\_vaccinated', 'daily\_vaccinations\_raw', 'daily\_vaccinations', 'total\_vaccinations\_per\_hundred', 'people\_vaccinated\_per\_hundred', 'people\_fully\_vaccinated\_per\_hundred', 'daily\_vaccinations\_per\_million', 'vaccines', 'source\_name', 'source\_website'], dtype='object')

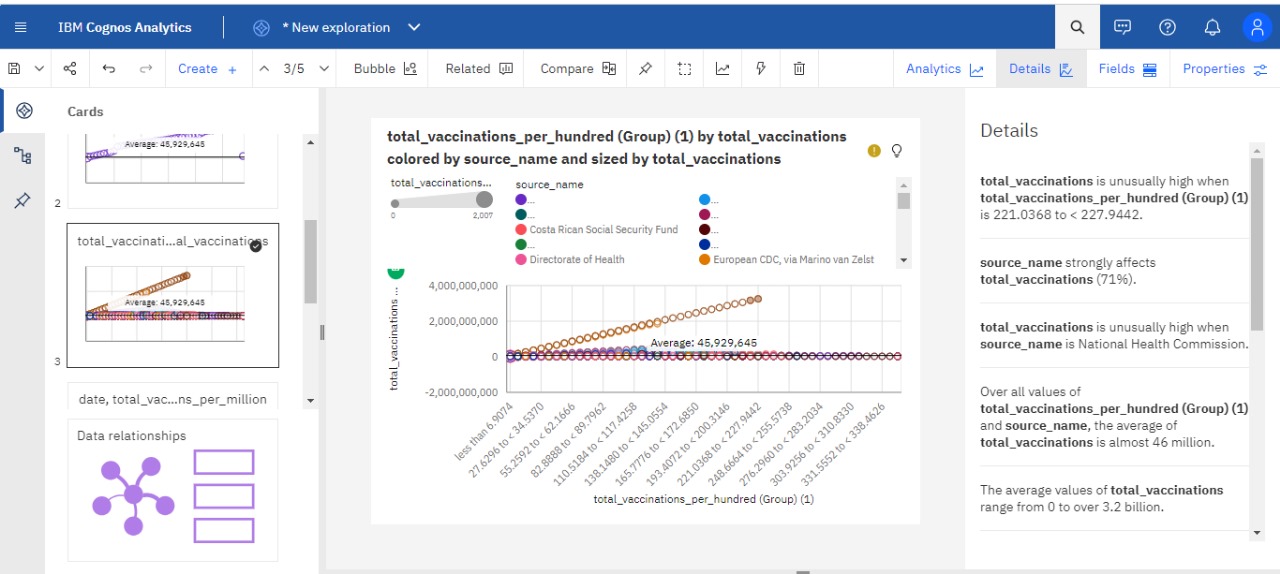


**source\_name**, **people\_fully\_vaccinated (Group) (1)**, **people\_vaccinated**, **daily\_vaccinations\_per\_million** and others predict **total\_vaccinations** with a strength of 91.8%.

**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)

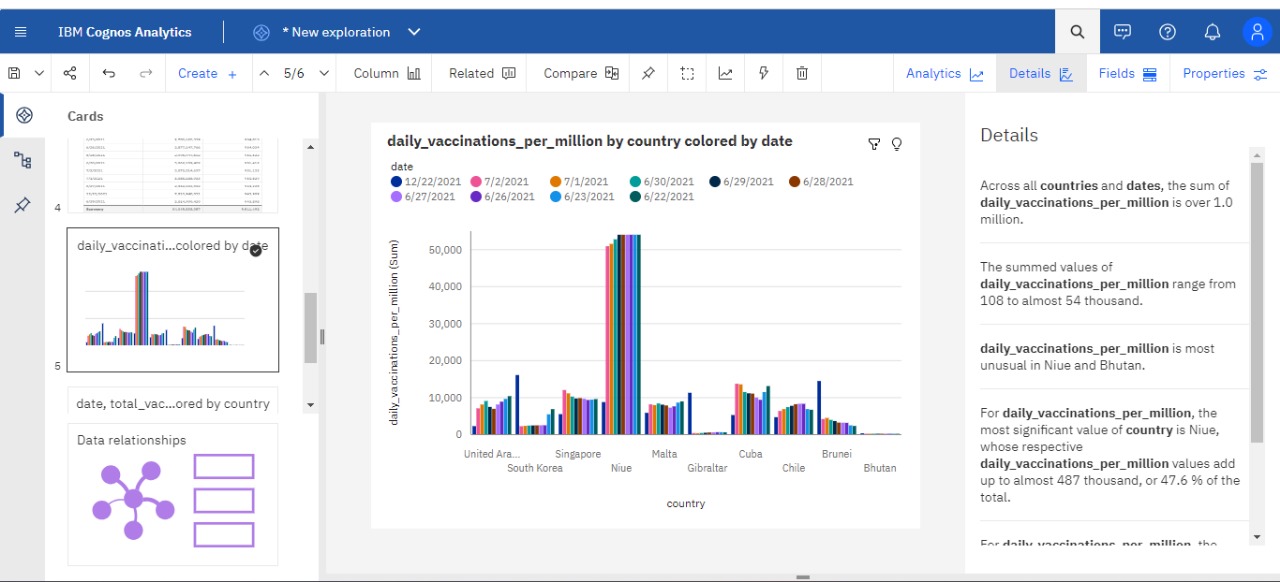
133,635

**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



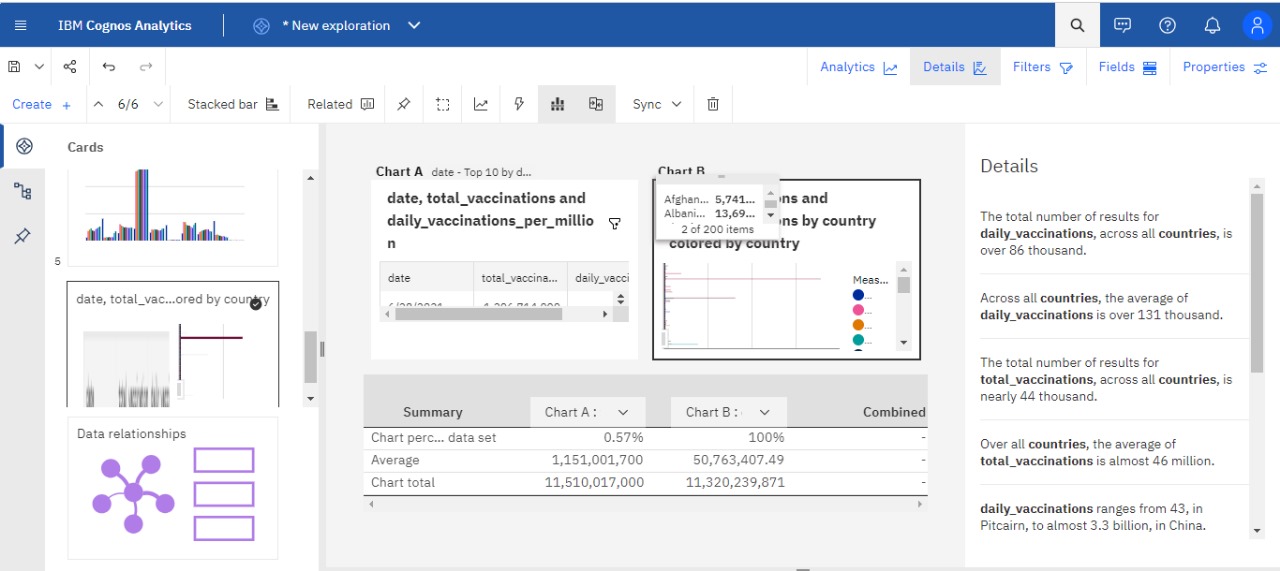
Across all **countries** and **dates**, the sum of **daily\_vaccinations\_per\_million** is over 1.0 million.

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).

