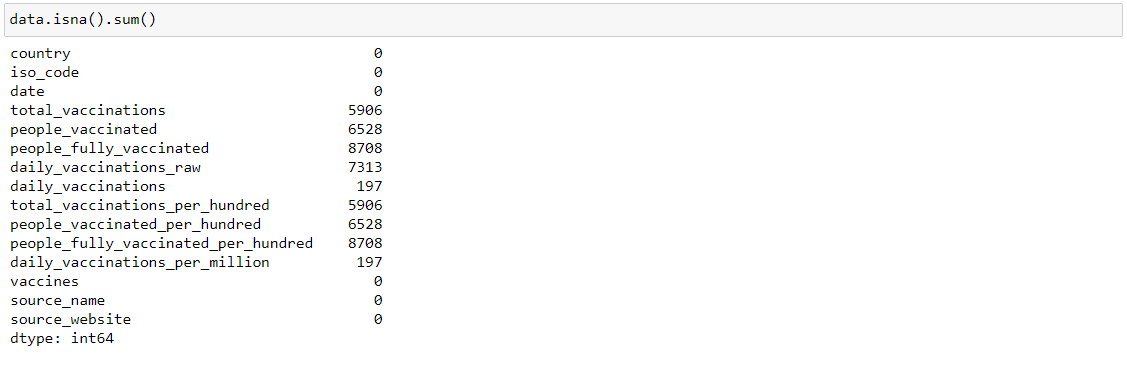
# Dataset

The dataset used for this shows the progress of the COVID-19 vaccine in different countries. The features of the dataset are as follows:

* Country: name of the country
* Iso\_code: the internationally accepted standard of referring to a country
* Date: the date on which the data was captured
* Total\_vaccinations: total number of vaccines in the country
* People\_vaccinated: number of people vaccinated in the country
* People\_fully\_vaccinated: number of people vaccinated with all their doses complete
* Daily\_vaccinations\_raw: The raw number of daily vaccinations happening in the country
* Daily\_vaccinations: Number of daily vaccinations happening in the country
* Total\_vaccinations\_per\_hundred: total number of vaccines grouped by 100
* People\_vaccinations\_per\_hundred: number of people vaccinated in the country grouped by 100
* People\_fully\_vaccinated\_per\_hundred: number of people vaccinated with all their doses complete grouped by 100
* Daily\_vaccinations\_per\_million: Number of daily vaccinations happening in the country grouped by million

# Machine Learning

The aim of this project code is to tackle the anomaly of missing data instances in each feature of the dataset. To this extent, heuristic and statistical approaches like the null-hypothesis tests will be applied.

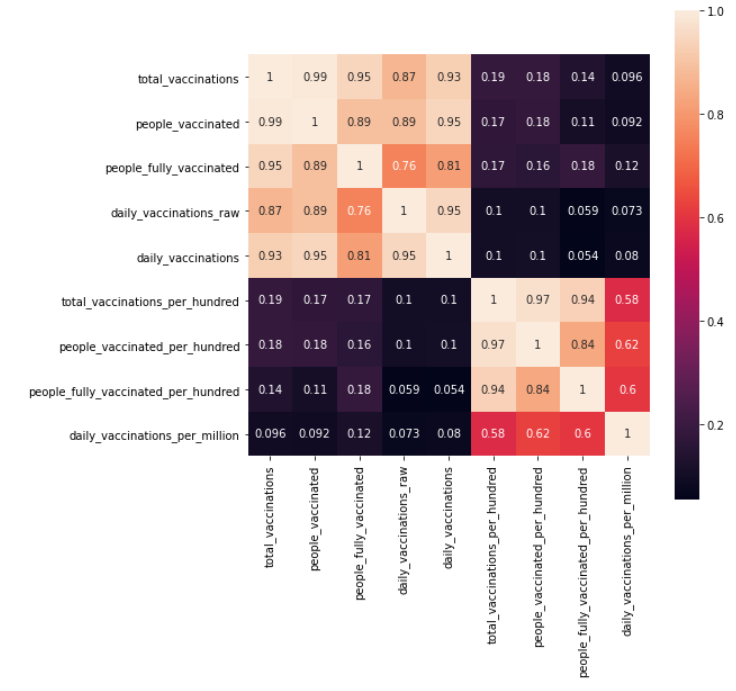


The above figure shows the initial number of null instances in the dataset.

The first step in removing the null instances in the dataset is to use the in-built function of dropping null values in pandas to drop all instances from the dataset where ‘total\_vaccinations’ has a null value. The results are as follows:



There is a significant reduction in the null values of the dataset features across the dataset. The next step is to plot a correlation matrix to see if any of the features in the dataset are degrading the quality of the dataset.



The correlation matrix shows a high-quality dataset with high correlation values across the dataset. However, from the correlation matrix, it can be seen from the trend of the correlation values that there might be features with highly similar data values in the dataset due to unusually high correlation values. This can be used to determine whether to discard the null instance or refill it with a different method.

We will be using the Mann-Whitney U test which is a test for null hypotheses to determine the similarity of the suspect pairs of features in the dataset suspected of having similar values. The first suspect pair is the ‘total\_vaccinations\_per\_hundred’ and ‘people\_vaccinated\_per\_hundred’. Upon performing the Mann-whitney U test, based on the p-value, this hypothesis cannot be ignored. In this case, we will use the following algorithm for refilling the values:

* Calculating the mean for all the values in both the features
* Taking the difference of both mean values
* Subtracting the difference from the values in ‘total\_vaccinations\_per\_hundred’ in indexes where ‘people\_vaccinations\_per\_hundred’ is null and replacing the null values with the subtraction results. The vice-versa is also performed.

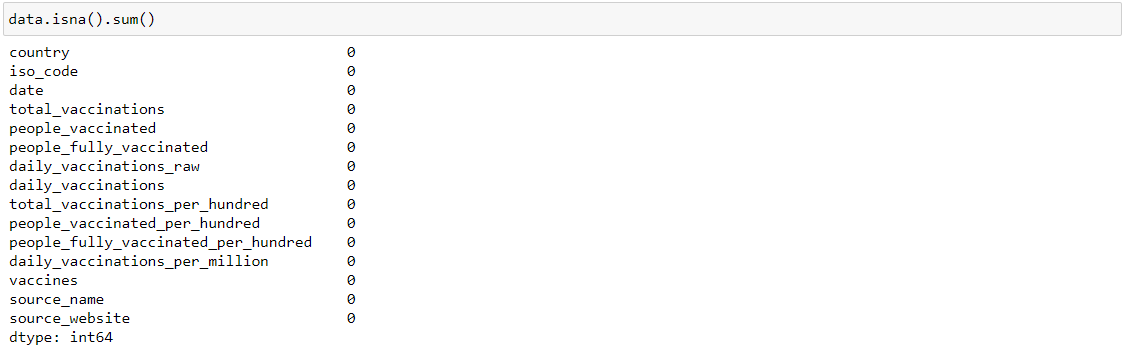
Suspect two pair of features are the ‘total\_vaccinations’ and ‘people\_vaccinated’. Based on the p-value, the hypothesis cannot be ignored. Therefore, the same algorithm is applied to replace the null values in both the features.

The next suspect pair is the pair of ‘people\_vaccinated’ and ‘daily\_vaccinations’. Based on the p-value, the hypothesis can be ignored. Therefore, the null instances will just be dropped. The same goes for the suspect pair ‘people\_vaccinated\_per\_hundred’ and ‘daily\_vaccinations\_per\_million’.

The next suspect pair is the ‘people\_fully\_vaccinated’ and ‘total\_vaccinations’. Based on the p-value, the hypothesis can be ignored, Therefore, the null instances will just be dropped. The same goes for ‘people\_fully\_vaccinated\_per\_hundred’ and ‘total\_vaccinations\_per\_hundred’.

The next and the final suspect pair is the ‘daily\_vaccinations\_raw’ and the ‘daily\_vaccinations’. Again, based on the p-value, the hypothesis can be ignored, Therefore, the null instances will just be dropped.

The final result is as follows:



As it can be seen in the figure above, there are no more null values in any of the features in the dataset.