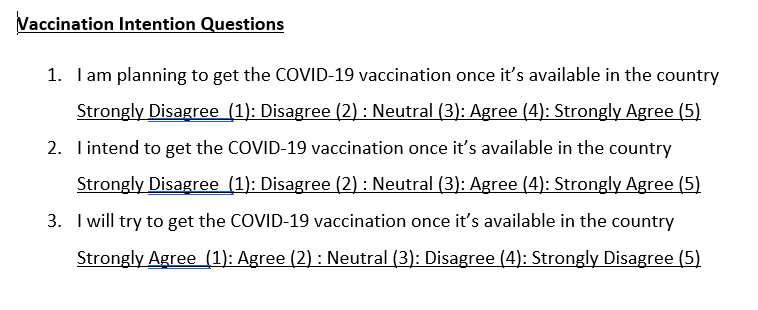
**Label Creation of Dataset**

**Missing Label**

In the paper “COVID-19 vaccine hesitancy: Vaccination intention and attitudes of community health

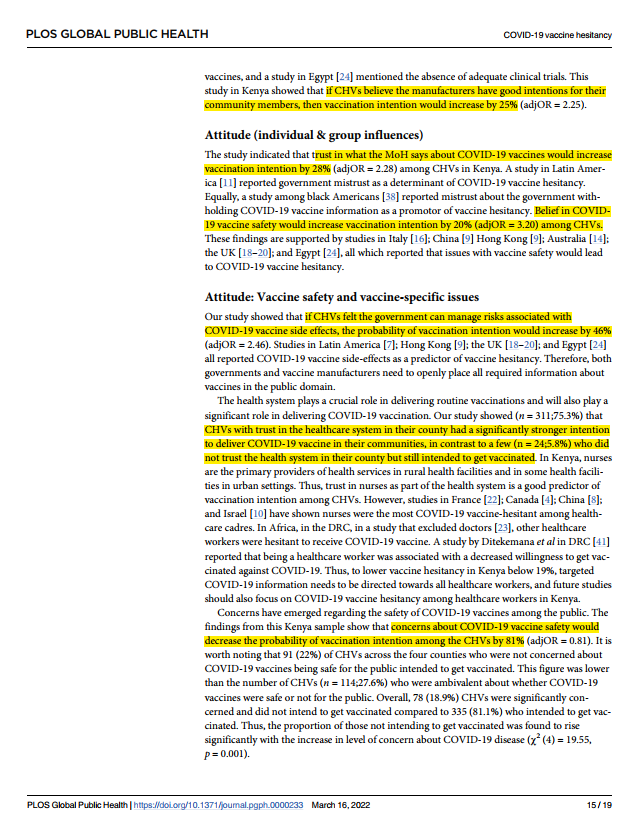
volunteers in Kenya”, they mentioned that the data collected for the research was based on a “WHO SAGE vaccine hesitancy matrix” form which is in 3 parts and this form can be easily downloaded using the hyperlink of the pdf paper. In the paper they mention statistics of COVID-19 vaccination intention among the volunteers, and this information can be seen in the third part of the questions that was asked to understand vaccine hesitancy, but when inspecting the dataset, we couldn’t find this data in the dataset. So, the labels are missing.

Missing values from dataset that could’ve been used as labels:



**Useful Data from the Dataset**

In the paper they mentioned a few information that corresponds to few columns/features in the dataset, that can influence vaccination intention significantly.



Columns in dataset with the questions asked for the data in the column:

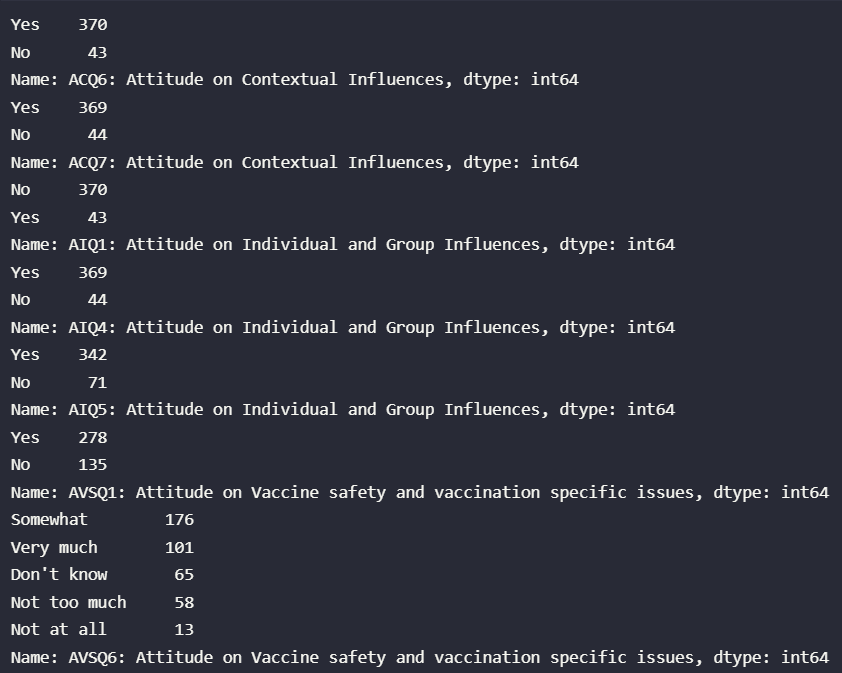
* acq6 - In your view is the MoH making the right decisions on COVID-19 vaccination?
* acq7 - Do you think vaccine manufacturers have good intentions for you and people in your community?
* aiq1 - Are you aware of any bad reactions in people who have had COVID-19 vaccination?
* aiq4 - Do you trust what the MoH says about COVID-19 vaccination?
* aiq5 - In your view is the COVID-19 vaccine safe enough for people to be injected?
* avq1 - Do you feel our country can manage risks associated with COVID-19 vaccine side effects?
* avq6 - In general, how safe do you think COVID-19 vaccine is for the general population?

These columns have few distinct values/answers, making it easy to encode and use.

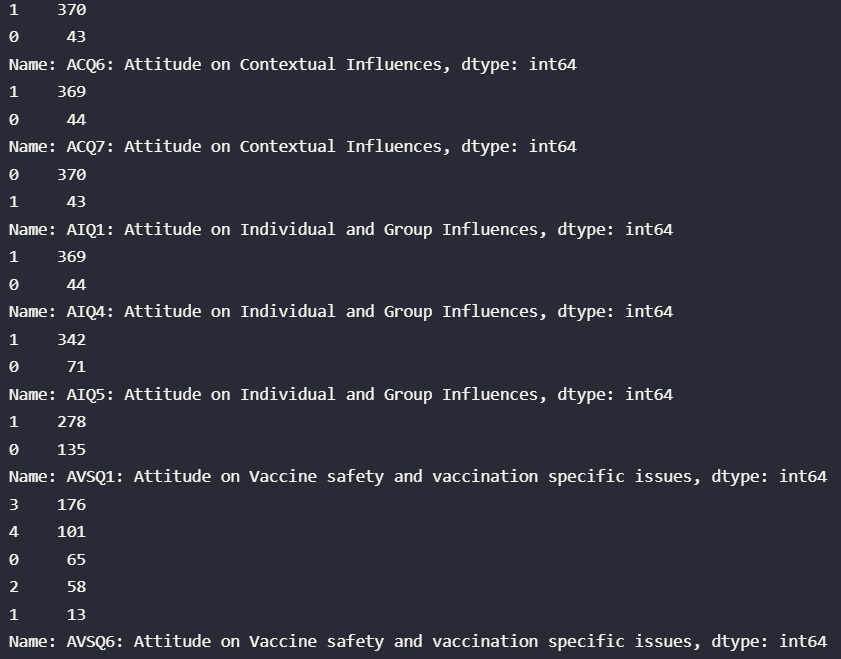
**Encoding**

To use the values from the columns we need to encode them to numerical data. So we use sklearn’s LabelEncoder.

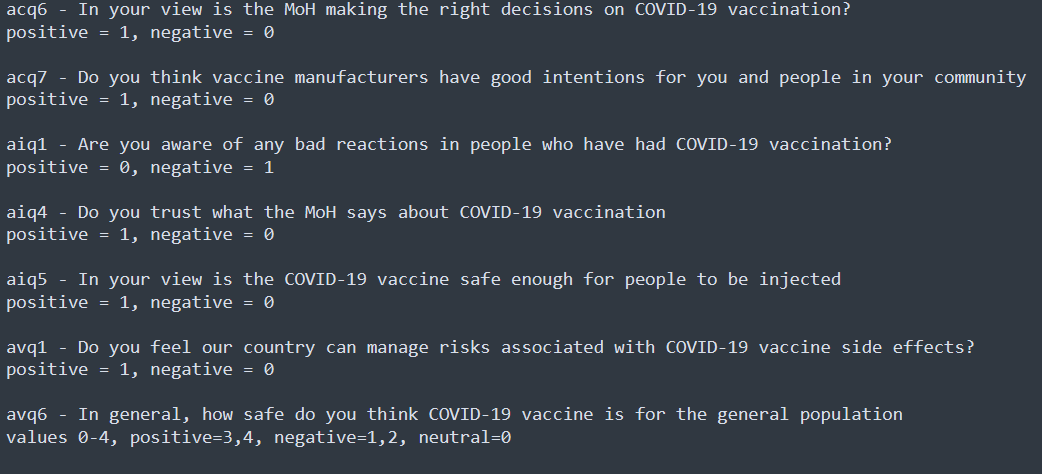
Inspecting values in the selected columns:



Inspecting the values after encoding:

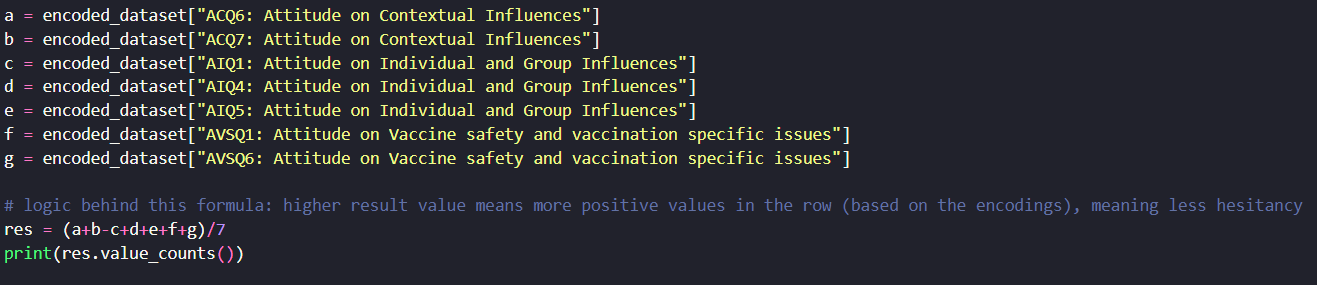


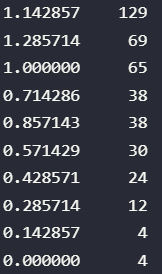
After inspection we can match the encoded values to be either responding positively to vaccination or negatively (this is decided based on response to the questions recorded in the dataset)



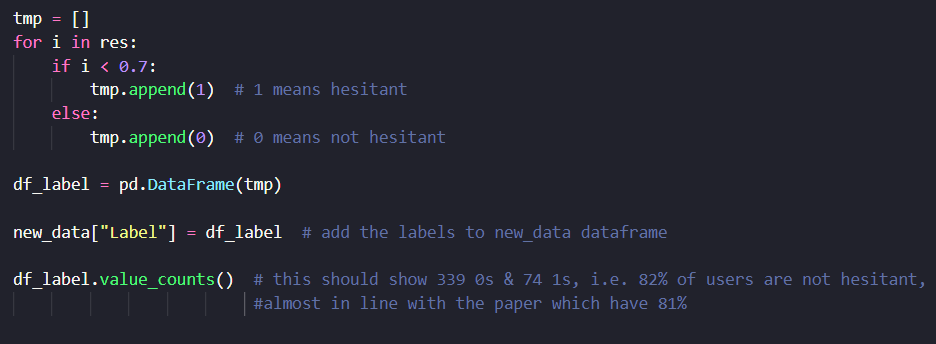
**Label Creation**

From previous inspection we can see in 6/7 cases positive reaction towards vaccination is encoded as a higher value, and negative reaction is a lower value, so we can add these when finding average. The one value where positive is lower value and negative is higher value, we can subtract it when finding average. So, if we add them and create an average, we can understand the overall vaccine hesitation from the volunteers.





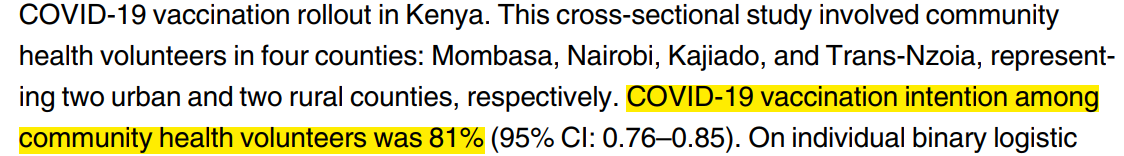
When doing a value count, we have a few distinct values, so now we need to find the threshold value that can best represent the percentage of volunteers who intent to get vaccinated.





By using 0.7 as the threshold from the average, we can split the values into a binary label to detect vaccine hesitation among the volunteers.

So now if we do (339/(339+74)), we get 0.8208, about 82% which is close to the value reported in the paper.



This way we can say the label generated was close to the actual labels that was removed before publishing the dataset (for licensing or whatever reason it may be).