HIV Dataset

Context:

The Publication regions include the India, Indonesia, Vietnam, Afghanistan, Bangladesh - 13 Websites. The data was scraped using BeautifulSoup, Selenium and Newspaper3K. Initially, I've stored each website article into separate CSVs here and At Final I concatenated all CSVs into one file. Data sources that were missing a label were simply assigned a label of "Nan".

The dataset is consisting of 10 columns

* Headlines
* Description
* Authors
* Published Dates
* Publication
* Category
* Article
* Keywords
* Summary
* URL

and represents 2295 HIV posts in total. The data primarily falls between the years of 2000 and 2019, there is possibly insignificant number of articles from 2000.

Content:

I have chopped the dataset into two CSVs file by concatenating where the initial dataset was (2295,10) after cleaning the dataset by removing Nan values. The resultant dataset was (580, 10)

Final\_HIV.csv – 2295 news articles (Articles 1- 2295)

Final\_HIV\_without\_Nan.csv – 580 news articles (Articles 1-580)