Overview preprocessing media dataset

As the dataset is huge, especially in terms of the amount of columns, I decided to inspect the columns and found there to be a big problem that needed to be addressed. There were four event type columns. Event type 1 was always filled, and the rest of these columns were only filled when the report had multiple event types. Same goes for the coordinates of the respective cities, a report could have up to four cities but this was rare so a lot of empty values here as well. I decided to split the rows per event type AND per city. This means we now have a row for every event type and for every city per report. This also means multiple rows per report are frequent. This way the data becomes way more processable, as well as easier to visualize. Furthermore, we greatly increased our amount of rows meaning there is more to show. The table below shows information per column. I would like to know your opinions on possible further operations and what you want to see visualized.

The full dataset contains 4244 rows and 28 columns. In my opinion this is a huge improvement over the initial dataset that contained 563 rows and 103 columns.

| **Column name** | **Amount of Unique Values** | **Amount of Empty Values** | **Short description** |
| --- | --- | --- | --- |
| Article ID | 566 (13.34%) | 0 | Contains the unique identifier of the article/report. |
| Event Date | 208 (4.9%) | 890 (20.97%) | The date of the respective report. |
| Event Type | 6 (0.14%) | 1 (0.02%) | A type of the report. A report can have multiple types, in that case there are more rows of this report with the same Article ID. |
| Event Type Precision | 29 (0.68%) | 722 (17.01%) | Further elaboration on the event type, for example “Departures” is more specific than just “Movements”. |
| Number of infected migrants | 12 (0.28%) | 4026 (94.86%) |  |
| Number of migrant in quarantine | 34 (0.8%) | 3816 (89.92%) |  |
| Number of deaths at sea | 35 (0.82%) | 3690 (86.95%) |  |
| Number of deaths when crossing borders | 7 (0.16%) | 4183 (98.56%) |  |
| Number of rescued people when crossing borders | 60 (1.41%) | 3417 (80.51%) |  |
| Number of arrived people by sea | 85 (2%) | 3199 (75.38%) |  |
| Number of arrived people through boarders | 20 (0.47%) | 3988 (93.97%) |  |
| Number of migrants deaths by COVID -19 | 0 | 4243 (100%) |  |
| Number of migrants healed of COVID-19 | 0 | 4243 (100%) |  |
| Number of people deported on sea | 20 (0.47%) | 4042 (95.24%) |  |
| Number of people deported through borders | 28 (0.66%) | 3917 (92.3%) |  |
| City | 59 (1.39%) | 219 (5.16%) |  |
| Latitude-city1\_degrees | 21 (0.49%) | 225 (5.30%) |  |
| Latitude-city1\_minutes | 56 (1.32%) | 239 (5.63%) |  |
| Longitude-city1\_seconds | 51 (1.2%) | 646 (15.22%) |  |
| Longitude-city1\_direction | 3 (0.07%) | 234 (5.51%) |  |
| Source-Country | 29 (0.8%) | 5 (0.12%) |  |
| Source-Type | 6 (0.14%) | 1 (0.02%) |  |
| Language | 7 (0.16%) | 1 (0.02%) |  |