An analysis of global Covid-19 data

Author: Rafał Klinowski

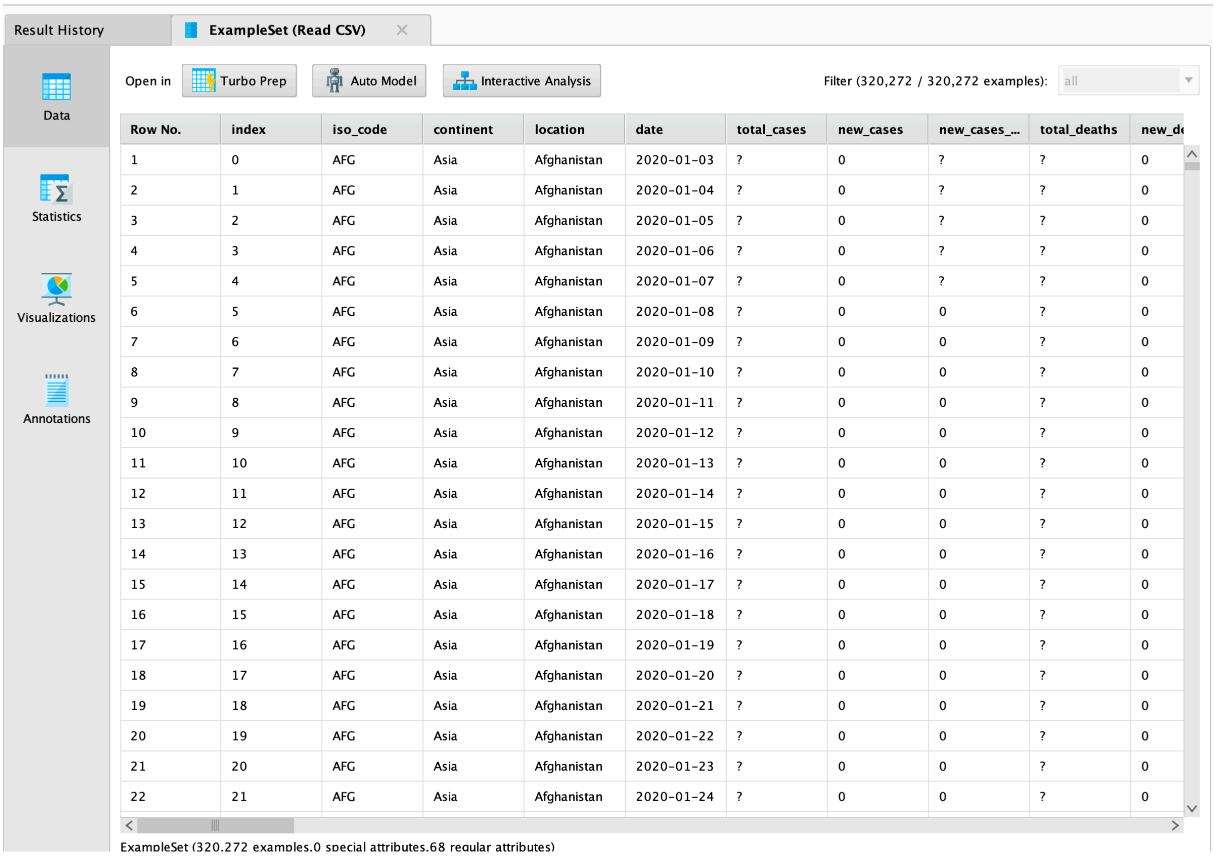
---

1. Used Dataset

The dataset used is the Global Covid-19 Data by Valtteri Kurkela, downloaded from https://www.kaggle.com/datasets/thedevastator/global-covid-19-data on 9.12.2023.

2. Initial data exploration

Firstly, the data was loaded in RapidMiner and stored as a dataset.

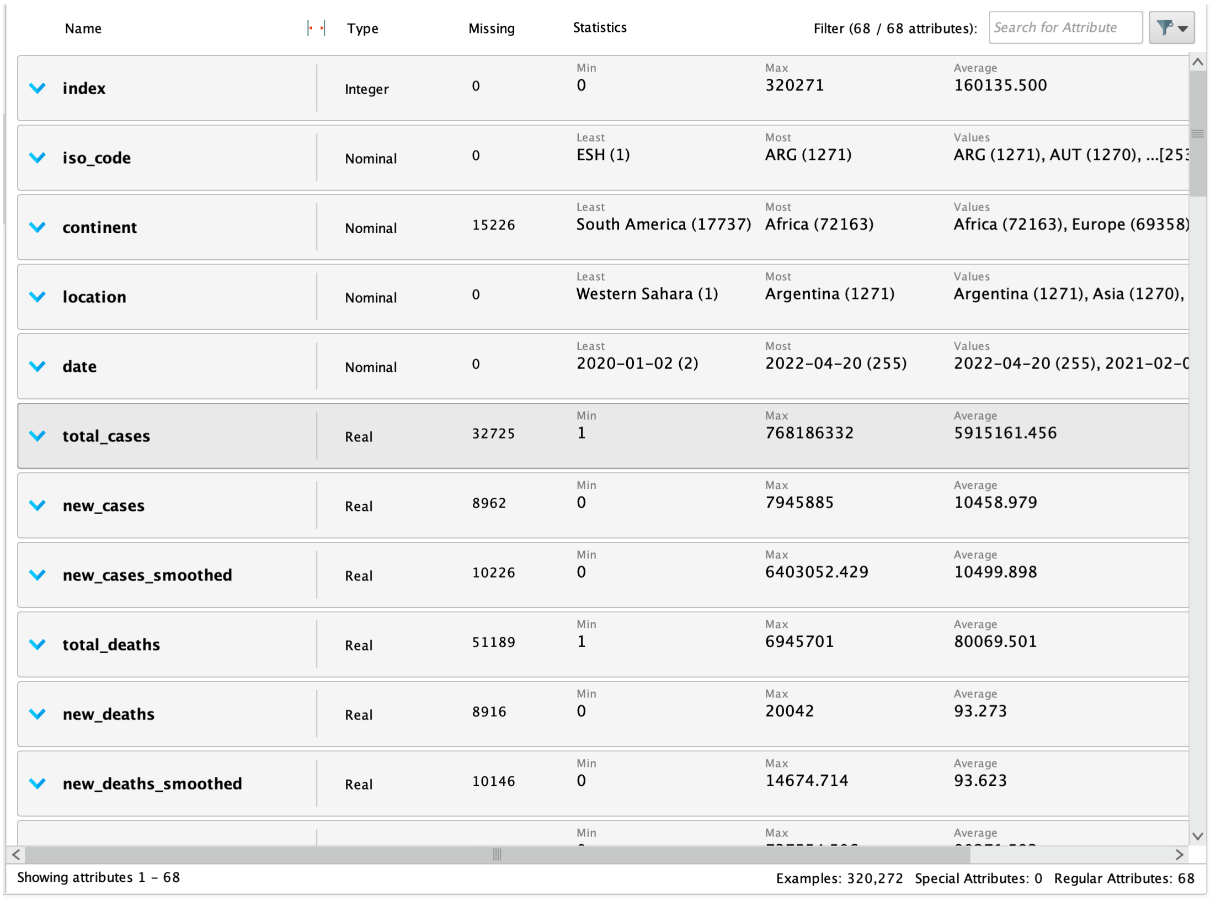


Row No.	index	iso_code	continent	location	date	total_cases	new_cases	new_cases_...	total_deaths	new_de
1	0	AFG	Asia	Afghanistan	2020-01-03	?	0	?	?	0
2	1	AFG	Asia	Afghanistan	2020-01-04	?	0	?	?	0
3	2	AFG	Asia	Afghanistan	2020-01-05	?	0	?	?	0
4	3	AFG	Asia	Afghanistan	2020-01-06	?	0	?	?	0
5	4	AFG	Asia	Afghanistan	2020-01-07	?	0	?	?	0
6	5	AFG	Asia	Afghanistan	2020-01-08	?	0	0	?	0
7	6	AFG	Asia	Afghanistan	2020-01-09	?	0	0	?	0
8	7	AFG	Asia	Afghanistan	2020-01-10	?	0	0	?	0
9	8	AFG	Asia	Afghanistan	2020-01-11	?	0	0	?	0
10	9	AFG	Asia	Afghanistan	2020-01-12	?	0	0	?	0
11	10	AFG	Asia	Afghanistan	2020-01-13	?	0	0	?	0
12	11	AFG	Asia	Afghanistan	2020-01-14	?	0	0	?	0
13	12	AFG	Asia	Afghanistan	2020-01-15	?	0	0	?	0
14	13	AFG	Asia	Afghanistan	2020-01-16	?	0	0	?	0
15	14	AFG	Asia	Afghanistan	2020-01-17	?	0	0	?	0
16	15	AFG	Asia	Afghanistan	2020-01-18	?	0	0	?	0
17	16	AFG	Asia	Afghanistan	2020-01-19	?	0	0	?	0
18	17	AFG	Asia	Afghanistan	2020-01-20	?	0	0	?	0
19	18	AFG	Asia	Afghanistan	2020-01-21	?	0	0	?	0
20	19	AFG	Asia	Afghanistan	2020-01-22	?	0	0	?	0
21	20	AFG	Asia	Afghanistan	2020-01-23	?	0	0	?	0
22	21	AFG	Asia	Afghanistan	2020-01-24	?	0	0	?	0

Figure 1. First few rows of the dataset.

We can see some information about the data, such as:

- There are a total of 320271 data rows with 68 attributes
- The data was collected between 1-01-2020 and 25-06-2023 (the collection was non-uniform – some days have more data than others)
- There are multiple missing values – some columns have as many as 310581 (97%) missing values



Filter (68 / 68 attributes): Search for Attribute

Name	Type	Missing	Statistics		
index	Integer	0	Min 0	Max 320271	Average 160135.500
iso_code	Nominal	0	Least ESH (1)	Most ARG (1271)	Values ARG (1271), AUT (1270), ...[25
continent	Nominal	15226	Least South America (17737)	Most Africa (72163)	Values Africa (72163), Europe (69358)
location	Nominal	0	Least Western Sahara (1)	Most Argentina (1271)	Values Argentina (1271), Asia (1270),
date	Nominal	0	Least 2020-01-02 (2)	Most 2022-04-20 (255)	Values 2022-04-20 (255), 2021-02-0
total_cases	Real	32725	Min 1	Max 768186332	Average 5915161.456
new_cases	Real	8962	Min 0	Max 7945885	Average 10458.979
new_cases_smoothed	Real	10226	Min 0	Max 6403052.429	Average 10499.898
total_deaths	Real	51189	Min 1	Max 6945701	Average 80069.501
new_deaths	Real	8916	Min 0	Max 20042	Average 93.273
new_deaths_smoothed	Real	10146	Min 0	Max 14674.714	Average 93.623

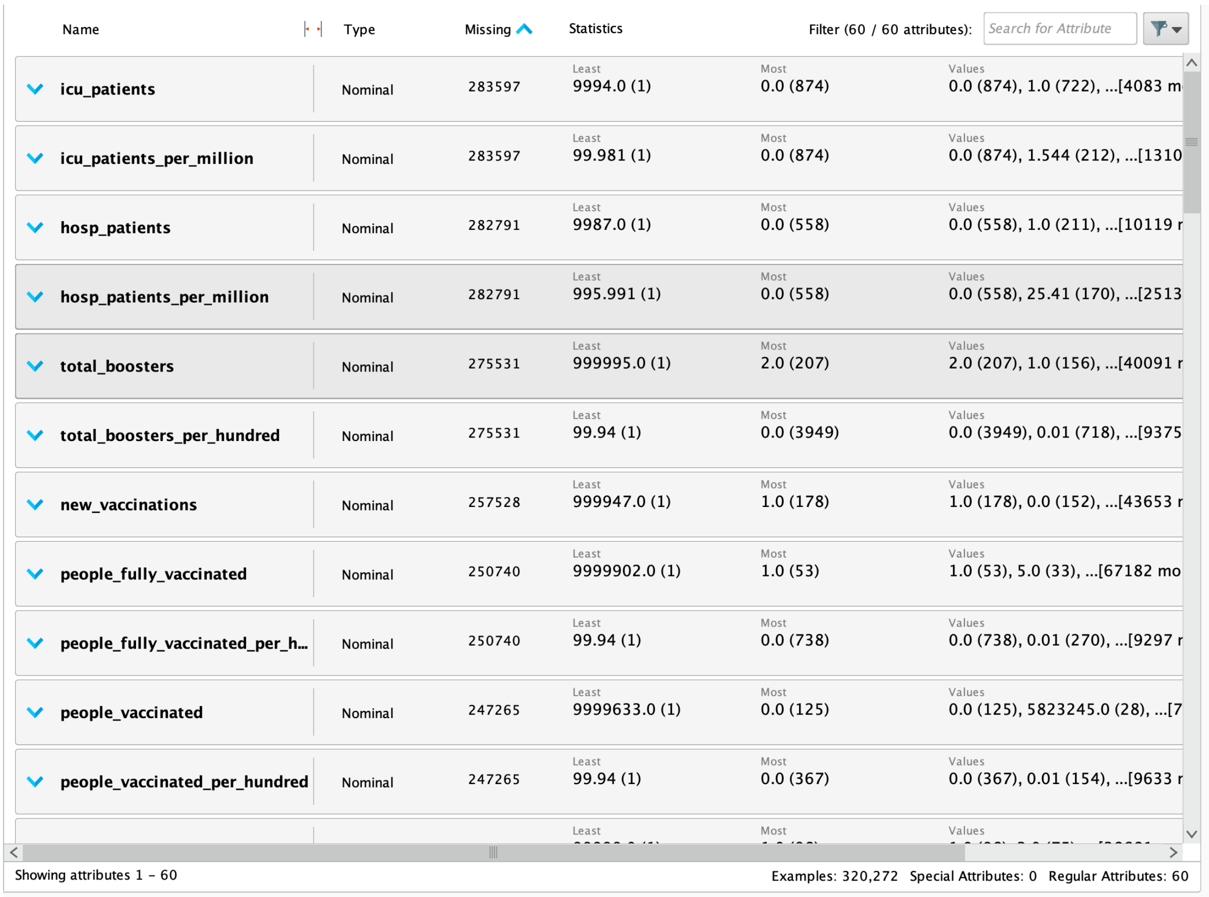
Showing attributes 1 - 68

Examples: 320,272 Special Attributes: 0 Regular Attributes: 68

Figure 2. Some statistics about several attributes from the dataset.

3. Handling missing values

The first step is to remove all attributes with more than 90% missing values. These attributes are not documented well enough to be considered for this project. Some other attributes (such as "icu_patients_per_million") can still be useful if analyzed properly, despite about 88% missing values. Some attributes have missing values likely because the data started to be collected later on (for example the "new_vaccinations" attribute – the vaccine was not yet available when the data started being collected in 2020).



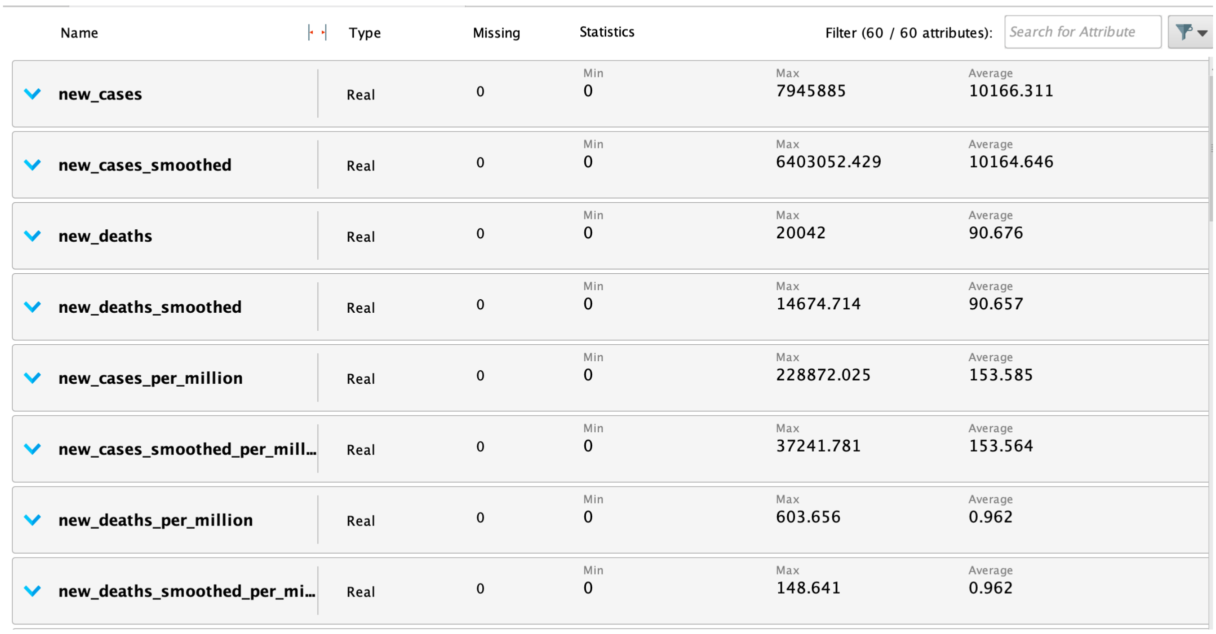
Filter (60 / 60 attributes): Search for Attribute

Name	Type	Missing	Least	Most	Values
icu_patients	Nominal	283597	9994.0 (1)	0.0 (874)	0.0 (874), 1.0 (722), ...[4083 m
icu_patients_per_million	Nominal	283597	99.981 (1)	0.0 (874)	0.0 (874), 1.544 (212), ...[1310
hosp_patients	Nominal	282791	9987.0 (1)	0.0 (558)	0.0 (558), 1.0 (211), ...[10119 r
hosp_patients_per_million	Nominal	282791	995.991 (1)	0.0 (558)	0.0 (558), 25.41 (170), ...[2513
total_boosters	Nominal	275531	999995.0 (1)	2.0 (207)	2.0 (207), 1.0 (156), ...[40091 r
total_boosters_per_hundred	Nominal	275531	99.94 (1)	0.0 (3949)	0.0 (3949), 0.01 (718), ...[9375
new_vaccinations	Nominal	257528	999947.0 (1)	1.0 (178)	1.0 (178), 0.0 (152), ...[43653 r
people_fully_vaccinated	Nominal	250740	9999902.0 (1)	1.0 (53)	1.0 (53), 5.0 (33), ...[67182 mo
people_fully_vaccinated_per_h...	Nominal	250740	99.94 (1)	0.0 (738)	0.0 (738), 0.01 (270), ...[9297 r
people_vaccinated	Nominal	247265	9999633.0 (1)	0.0 (125)	0.0 (125), 5823245.0 (28), ...[7
people_vaccinated_per_hundred	Nominal	247265	99.94 (1)	0.0 (367)	0.0 (367), 0.01 (154), ...[9633 r

Showing attributes 1 – 60 Examples: 320,272 Special Attributes: 0 Regular Attributes: 60

Figure 3. The results after the first step of handling missing values.

For the statistics attributes such as “new deaths” or “new cases”, we will assume they are equal to 0 on a given date if there is no given value.



Filter (60 / 60 attributes): Search for Attribute

Name	Type	Missing	Min	Max	Average
new_cases	Real	0	0	7945885	10166.311
new_cases_smoothed	Real	0	0	6403052.429	10164.646
new_deaths	Real	0	0	20042	90.676
new_deaths_smoothed	Real	0	0	14674.714	90.657
new_cases_per_million	Real	0	0	228872.025	153.585
new_cases_smoothed_per_mill...	Real	0	0	37241.781	153.564
new_deaths_per_million	Real	0	0	603.656	0.962
new_deaths_smoothed_per_mi...	Real	0	0	148.641	0.962

Figure 4. Several attributes after replacing their missing values with zeroes (where it made sense).

An interesting note is that there are missing values for the Continent attribute, but no missing values for the Location or ISO_code attributes. Therefore, the missing values for Continents could be deducted from the country names.

To fix this, we use the Countries-Continents dataset from https://github.com/dbouquin/IS_608/blob/master/NanosatDB_munging/Countries-Continents.csv that matches each country name to the continent it's located in. The idea is to first merge the two datasets, then fill in the missing Continent values using the values from the "Countries-Continents" set.

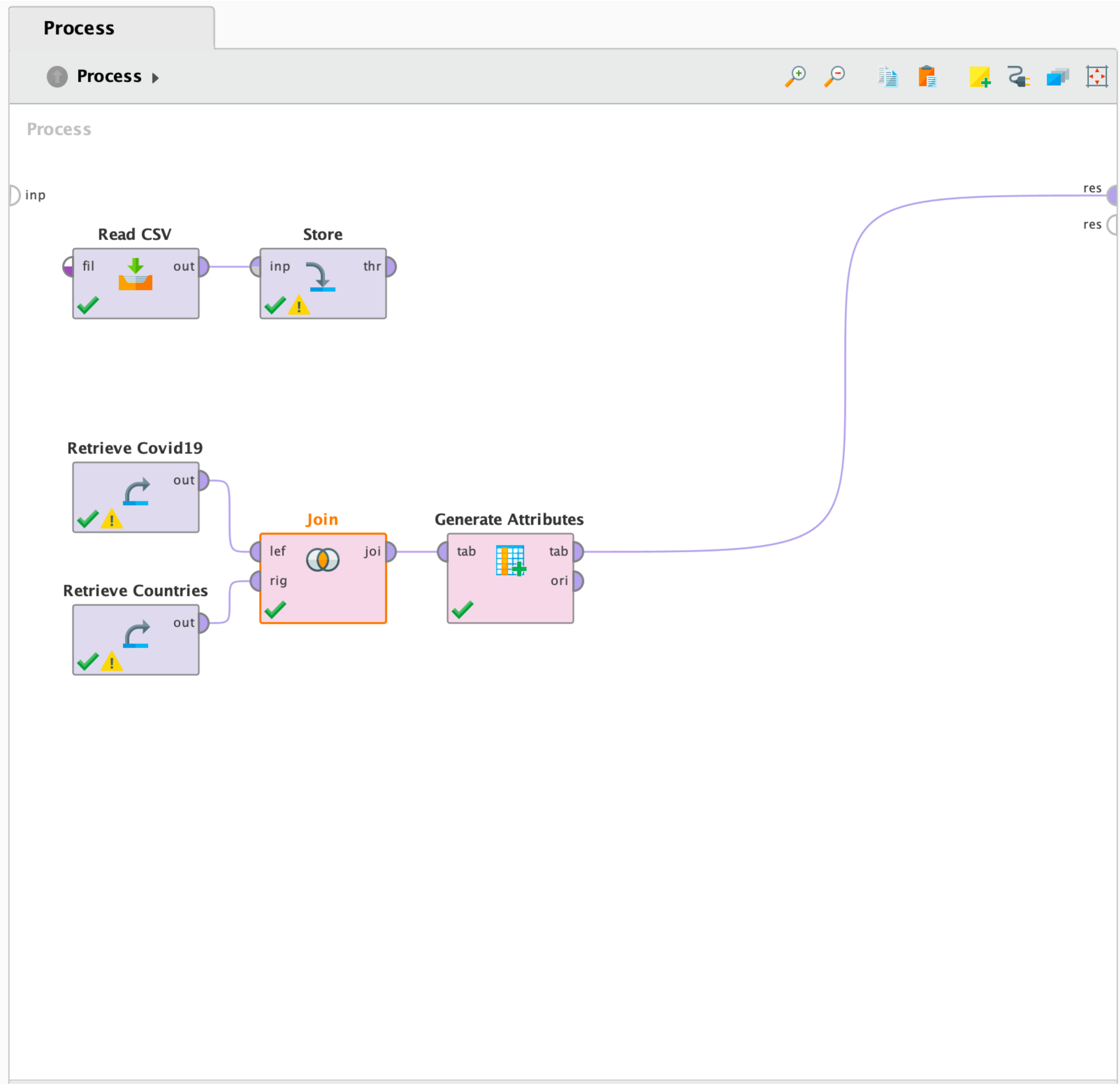


Figure 5. Replacing missing Continent values based on the secondary dataset.

4. Creating some visualizations

Let's take a look at some interesting statistics.

Firstly, let's display the data for Poland.

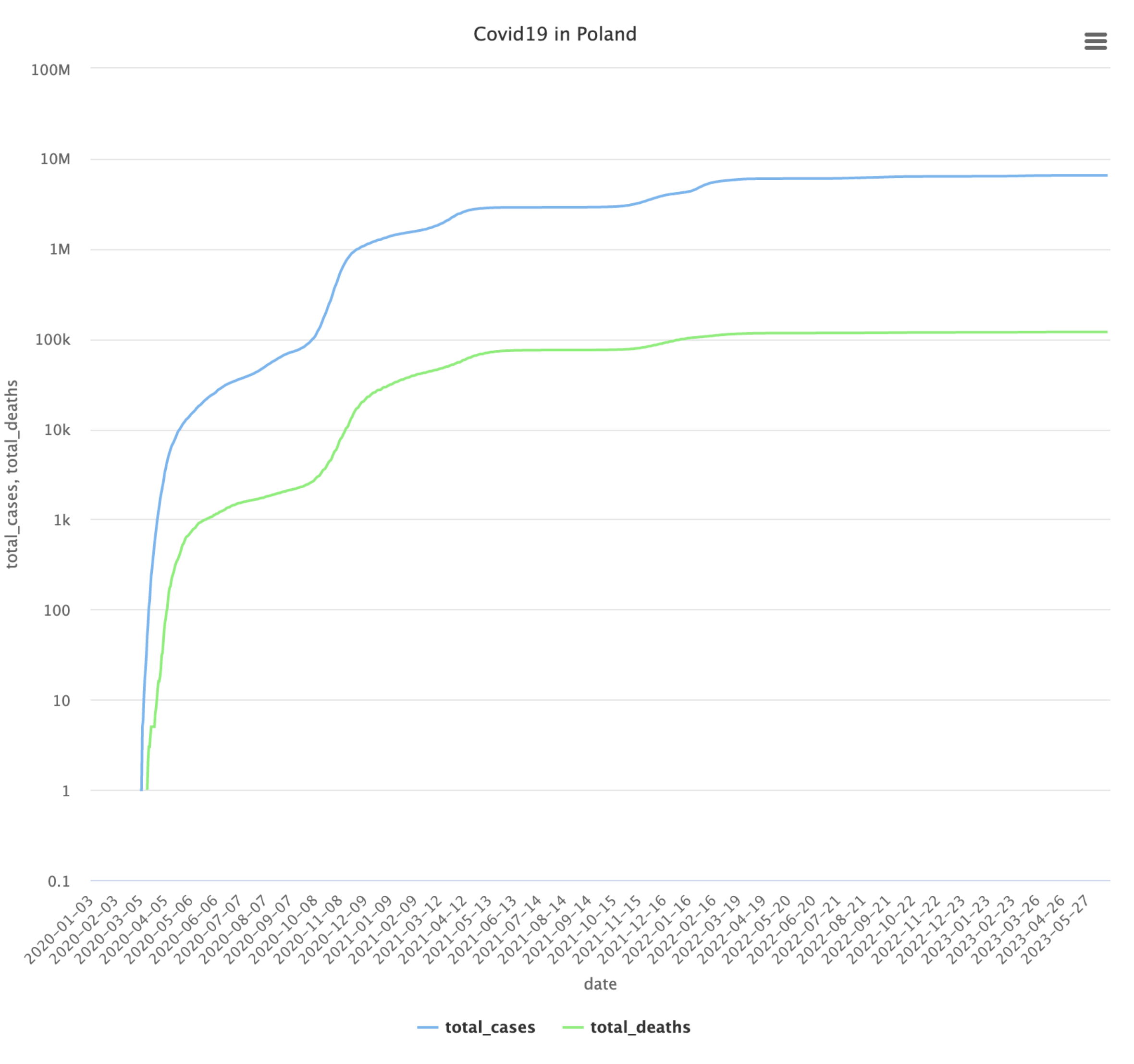


Figure 6. Covid19 statistics for the whole population in Poland. Logarithmic scale.

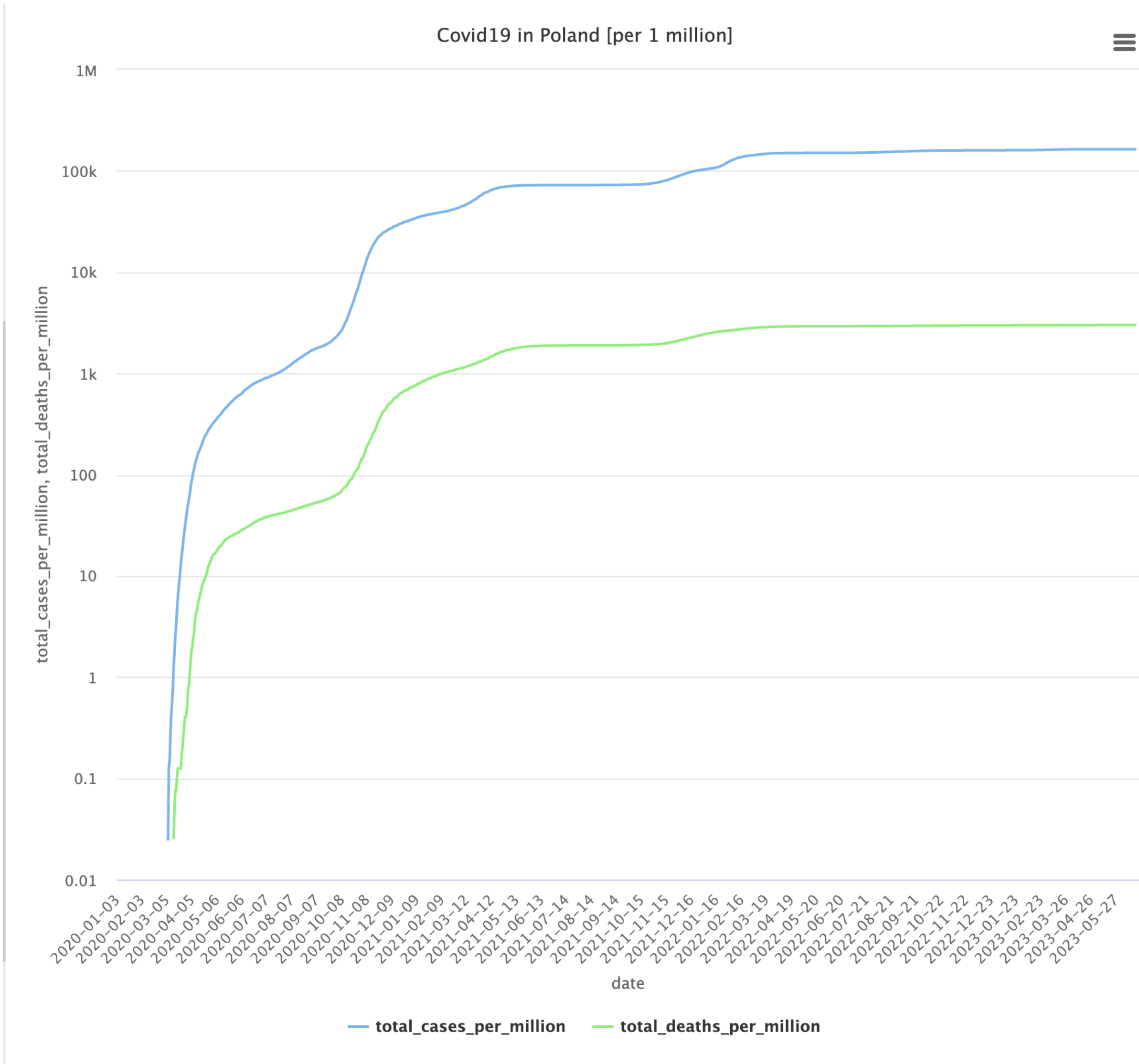


Figure 7. COVID-19 statistics per 1 million population in Poland. Logarithmic scale.

The next part of the project was done in Python.

```
data = pd.read_csv("../data/covid19-cleaned.csv")
# Display the relation between number of total vaccinations and number of
covid cases in Poland
data_poland = data[data["location"] == "Poland"]
data_poland = data_poland[["date", "people_vaccinated", "new_cases"]]

print(data_poland.info())

# Drop all rows where there is no data about number of vaccinations
data_poland = data_poland.dropna(subset=["people_vaccinated"])

print(data_poland.head())

# Group data by month
data_poland["date"] = pd.to_datetime(data_poland["date"])
# Sum only the "new_cases" column, for the "people_vaccinated" take the
last value for each month
data_poland = data_poland.groupby(pd.Grouper(key="date",
freq="M")).agg({"new_cases": "sum", "people_vaccinated":
"last"}).reset_index()

print(data_poland.head())

plt.plot(data_poland["date"], data_poland["people_vaccinated"],
label="people_vaccinated")
plt.plot(data_poland["date"], data_poland["new_cases"], label="new_cases")
plt.legend()
plt.title("Relation between number of total vaccinations and number of
monthly covid cases in Poland")
# Logarithmic scale
plt.yscale("log")
# Tilt the x-axis labels
plt.xticks(rotation=45)
plt.show()
```

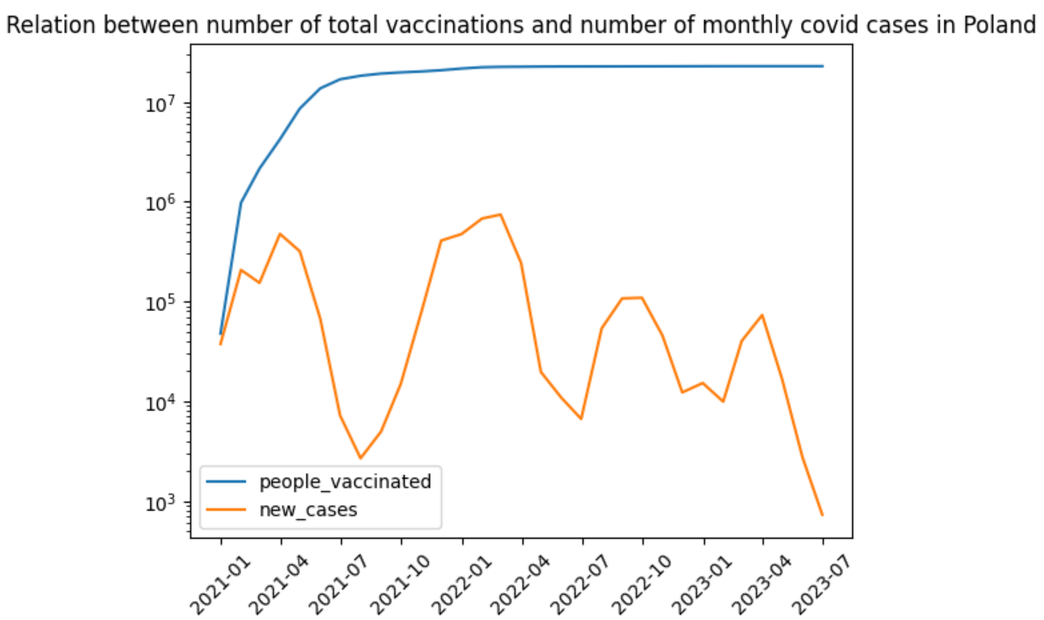


Figure 8. The relation between number of cases and the total amount of vaccinated people in Poland, grouped by month.

The correlation between “people_vaccinated” and “new_cases” was calculated to be equal to: -0.13536198, which means there is a weak negative correlation between the two attributes, and as one increases, the other attribute value decreases. This however does not mean causation.

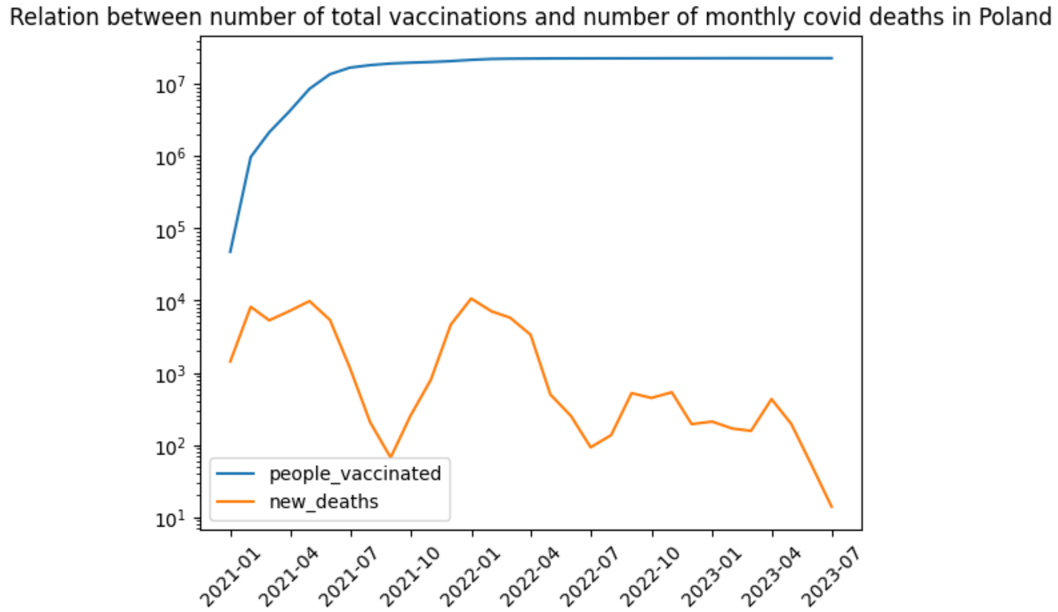


Figure 9. The relations between Covid deaths and the total number of vaccinated people in Poland, grouped by month.

The correlation for this one was calculated to be equal to:
-0.5063519, which means there is a strong negative correlation between the two attributes.

5. Analysis of Covid based on seasons

Let’s first take a look at whether the season of the year is correlated with the amount of Covid cases and deaths in Europe and the US.

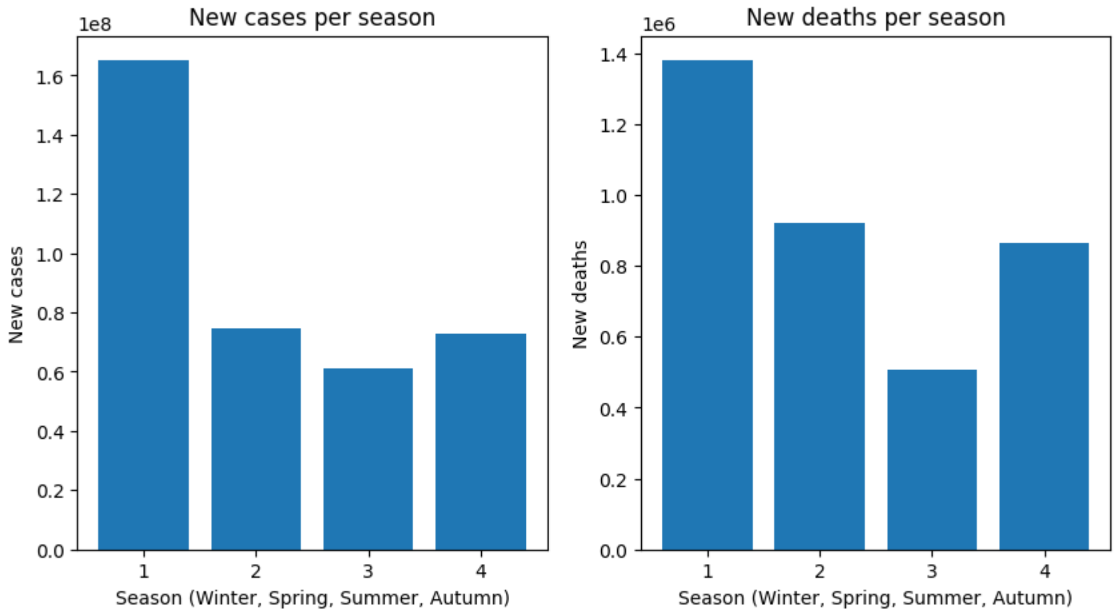


Figure 10. The amount of COVID cases and deaths recorded in Europe and North America, grouped by season (1 – Winter, 2 – Spring, 3 – Summer, 4 – Autumn).

Since Covid began in early 2020, let's exclude the first months up until December 2020.

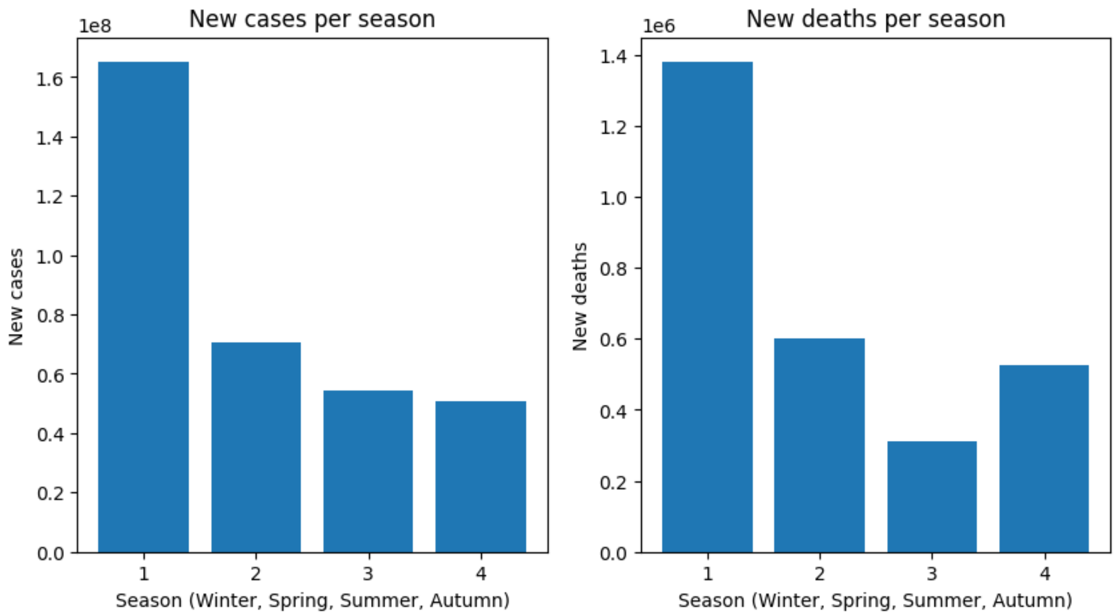


Figure 11. The amount of COVID cases and deaths recorded in Europe and North America, grouped by season (1 – Winter, 2 – Spring, 3 – Summer, 4 – Autumn). Data from 1.12.2020 – 25.06.2023.

We can see that Winter comes with a much-increased amount of both Covid cases and Covid deaths, even if we exclude the initial few months of the pandemic.