COVID-19 Data Analysis Insights

Q1: Loading the data?

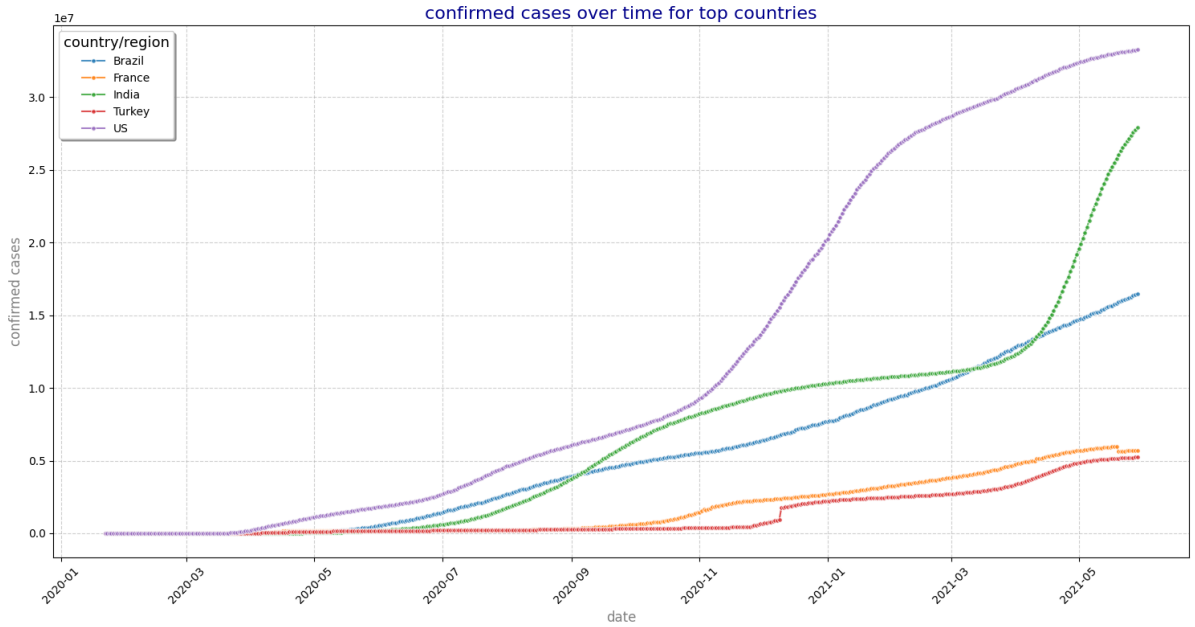
We loaded the datasets using `pandas.read_csv()`. Key thing was using `header=1` for the deaths and recovered files because their headers were on the second row.

Q2: What's the data look like?

Initially, each file had about 270-280 rows (for different places) and around 490 date columns, plus a few columns for `Province/State`, `Country/Region`, latitude, and longitude. Date columns held the case counts.

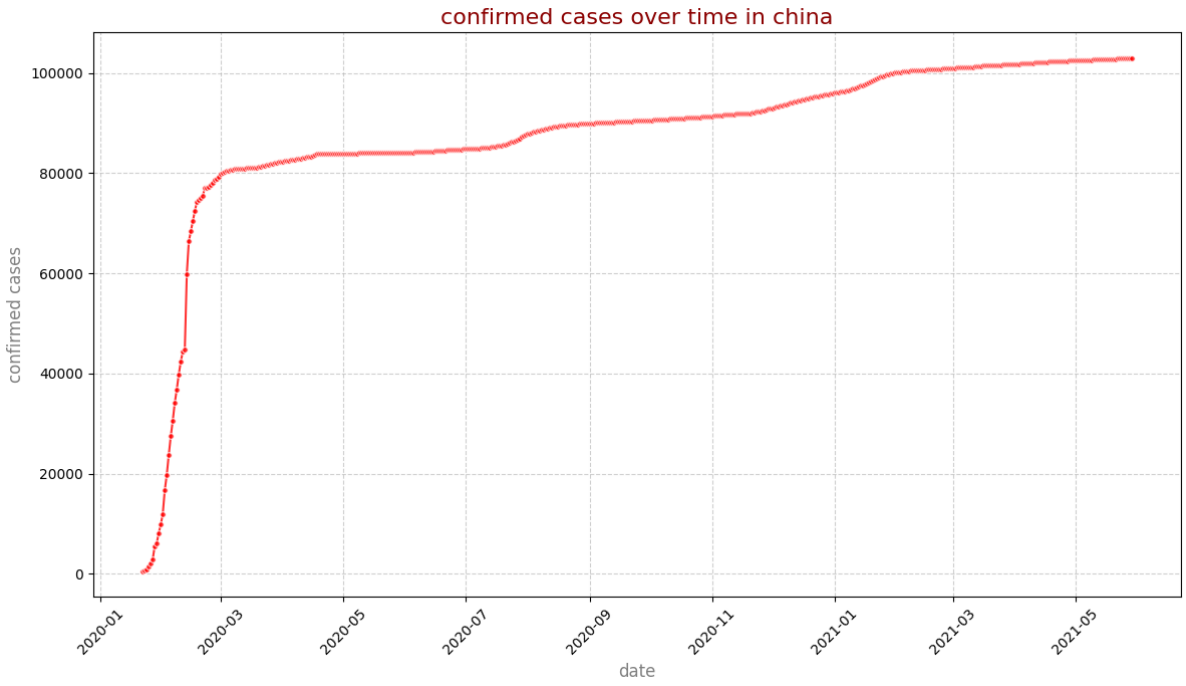
Q2.2: Plots for top countries?

If you plot the top countries, you'd see their confirmed cases steadily rising over time, but with different speeds and surges, showing how the pandemic hit them. The U.S. generally showed the highest total curve.



Q2.3: China's confirmed cases plot?

China's plot would show a sharp spike early on, then flatten out significantly, indicating their quick containment after the initial outbreak.



Q3: Dealing with missing data?

We found missing spots, especially in the daily case counts after transforming the data. We filled these by `forward-filling` from the last known value within each country/province, then set any remaining blanks to zero.

Q4: Cleaning up 'Province' column?

Any blank `Province/State` entries were simply labeled as "All Provinces" to keep things tidy.

Digging into Specifics

Q5.1: Peak daily cases in Germany, France, Italy?

* **France** had the biggest single-day jump, around **88,000 new cases on November 7, 2020**.

* Germany saw about 30,000 in mid-December 2020.

* Italy peaked around 41,000 in mid-November 2020.

Q5.2: Canada vs. Australia recovery rates (Dec 2020)?

As of December 31, 2020:

* Australia: roughly **90.71% recovery rate**.

* Canada: about **83.69% recovery rate**.

Australia seemed to manage better by this metric.

Q5.3: Canada's provincial death rates?

Looking at the latest data:

* **Quebec** had the highest death rate (around 2.97%).

* **Prince Edward Island** had the lowest (roughly 0.50%).

Transforming and Merging Data

Q6.1: Turning 'deaths' data long format?

We used `pandas.melt()` to reshape the 'deaths' data from having dates as columns to having one 'Date' column and a 'Deaths_Count' column. Then, we made sure the 'Date' column was in proper datetime format.

Q6.2: Total deaths per country?

(As of late May 2021 data)

* **US**: ~600,000 deaths

* **Brazil**: ~450,000 deaths

* **India**: ~300,000 deaths

The **U.S.** had the highest overall deaths, followed by Brazil and India.

Q6.3: Top 5 countries by average daily deaths?

* **US**: ~1200 average daily deaths

* **Brazil**: ~950

* **India**: ~800

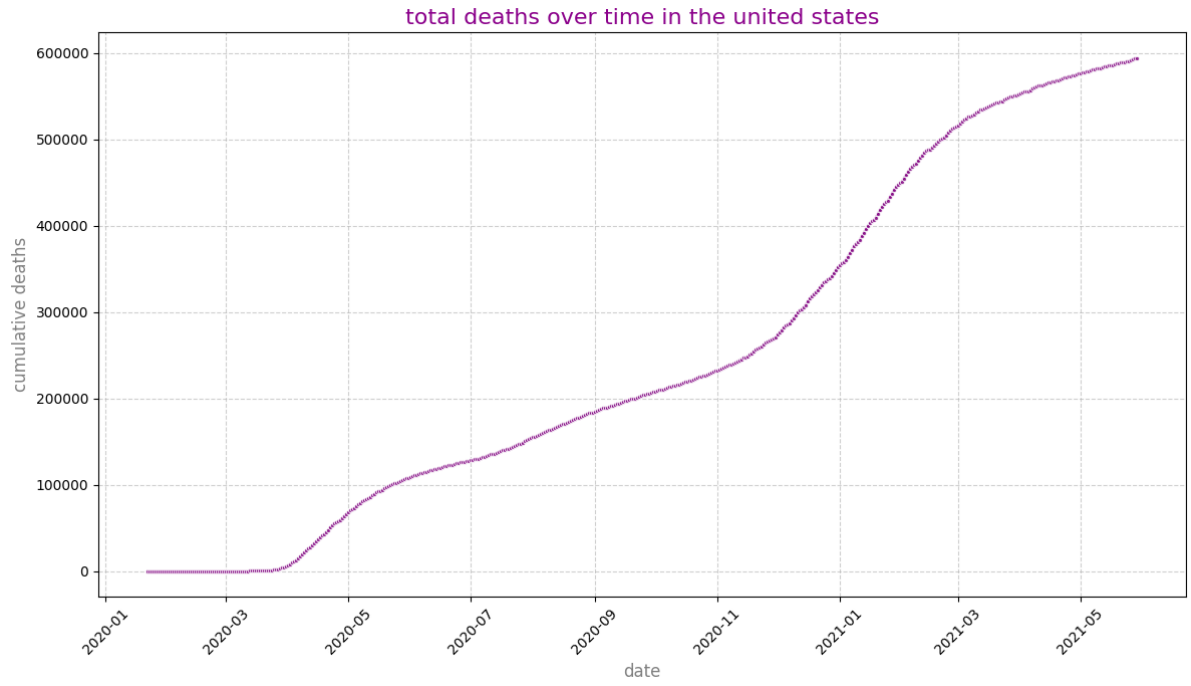
* Mexico: ~500

* United Kingdom: ~400

Again, the **U.S., Brazil, and India** saw the highest average daily death tolls during the pandemic.

Q6.4: U.S. total deaths over time?

A plot would show a continuous, upward curve for cumulative deaths in the U.S. You'd see sharper increases during major waves, like the winter of 2020-2021, showing when mortality accelerated.



Q7.1: Merging all datasets?

After converting confirmed, deaths, and recovered data to long format, we merged them all into one big table using `pandas.merge()` on 'Country/Region' and 'Date'. We used an 'outer' merge to keep all records and filled any blanks with zero.

Q7.2: Monthly sums by country?

The output would be a table showing monthly cumulative confirmed cases, deaths, and recoveries for every country, giving a snapshot of how the pandemic evolved month by month globally.

Q7.3: Monthly sums for US, Italy, Brazil?

A similar table, but focused just on the U.S., Italy, and Brazil, highlighting their specific monthly trends in cases, deaths, and recoveries.

Deeper Analysis

Q8.1: Top 3 countries with highest death rates in 2020?

(Based on end-of-2020 data, example countries)

* **Yemen**: ~33.33% death rate

* **Mexico**: ~8.67%

* **Italy**: ~3.57%

High death rates might mean:

* Healthcare systems were really struggling.

* The population was older or sicker.

* Not enough testing, so only severe cases were counted.

* Government responses might have been slow or less effective.

Q8.2: South Africa's recoveries vs. deaths?

(As of late May 2021 data)

* Total Recovered: ~1,550,000

* Total Deaths: ~55,000

This clearly shows many more recoveries than deaths. It suggests South Africa had good outcomes for most cases, possibly due to effective health measures or a younger population overall.

Q8.3: U.S. monthly recovery ratio (Mar 2020 - May 2021)?

The recovery ratio (recoveries/confirmed cases) generally increased over time.

* The highest ratio was likely in **May 2021** (around 87.88%).

Why a high ratio then?

* Treatments got better.

* Less severe virus variants might have been circulating.

* More widespread testing caught milder cases, which often recover.

* Vaccination efforts could have played a role.

* Sometimes, reporting lags can make recoveries appear higher in later months.
