0.0-data-wrangling

October 20, 2020

1 Goal

My goal is to visualize various aspect of the `COVID-19` pandemic. In this notebook I describe how the data is acquired and processed.

2 Data sources

Link	Source
https://github.com/CSSEGISandData/COVID-19	JHU CSSE
GDP per capita PPP	The World Bank
Population	The World Bank
Urban Population	The World Bank
Population living in slums	The World Bank
Rural population	The World Bank
Life expectancy at birth	The World Bank
Current healthcare expenditure	The World Bank
https://datahub.io/JohnSnowLabs/country-and-continent-codes-list	Datahub

The process of obtaining the data has been automated. See the `src/data` directory.

3 Data wrangling

3.1 COVID-19

3.1.1 Original data

This dataset is downloaded from a `repository` on `github`. The data about `COVID-19` cases is in `.csv` files where each region has a seperate row. We group the data by country and store each country in a different column. Cases that happened on boats are removed from the data.

See the script `src/features/make_cases.py` for details.

3.1.2 Derived data

From the original data about `COVID-19` cases we calculate what follows:

- `mortality rate = dead / confirmed`
- `active cases = confirmed - recovered - dead`.

We also extract a list of countries and apply the differencing operator to `confirmed` to extract the `daily change in cases` for each country.

3.2 World Bank data

The data from the World Bank is downloaded using the `wbdata` library. The data includes is `Life expectancy` and `GDP per capita` to name a few. We extract the last known value of an indicator for a given county.

See the script `src/features/make_world_bank.py` for details.

3.3 Continents

In order to analyse the data by continent, we download a list of countries with continents and a list of countries with their respective 3 letter codes.

See the script `src/features/make_continent.py` for details.

4 Summary

After preparing, cleaning and joining the downloaded datasets we store newly created `.csv` files in `data/processed` directory for further use. Here is table with a brief description of the contents of each file.

Name	Description
active_cases.csv	Calculation: `confirmed - recovered - dead`
confirmed_cases.csv	Time series of confirmed cases from JHU CSSE.
confirmed_cases_daily_change.csv	Daily change in confirmed cases, derived from JHU CSSE.
confirmed_cases_since_t0.csv	Reindexed time series of confirmed cases.
continents.csv	Countries mapped to continents.
coordinates.csv	Country coordinates.
country_stats.csv	Newest available case data by county.
country_to_continent.csv	A mapping of countries to continents.
dead_cases.csv	Time series of fatalities from JHU CSSE.
mortality_rate.csv	Calculation: `dead / confirmed`, derived from JHU CSSE.
recovered_cases.csv	Time series of recovered cases from JHU CSSE.
world_bank.csv	Socioeconomic from the World Bank merged with data about covid.
world_bank_codes.csv	3 letter country codes from the World Bank.