Data Wrangling Project Report

Jiangwei Wang, Shanshan Liu, Xiaoxi Guo, Xue Yang

Data Sources

Our data are mainly source from three websites as follows:

1. WHO Coronavirus Disease Dataset: https://covid19.who.int/
2. World Bank Open Data: https://data.worldbank.org/
3. OCHA Services Dataset: HUMANITARIAN DATA EXCHANGE: https://data.humdata.org/dataset

The raw data files are CSV and Excel formats

1. WHO-COVID-19-global-data.csv
2. covid_impact_education.csv
3. GlobalEconomicProspectsJune2020GDPgrowthdata.xlsx
4. WPP2019_TotalPopulationBySex.csv
5. HNP_StatsData.csv
6. GEP_csv.zip
7. WDI_csv.zip

Why you chose those data sources

The topic of this project centers on the impact of Covid-19 on economy, education and demographic aspects. Therefore, the data sources should include two aspects: data regarding the Covid –19 pandemic and data regarding the economic and educational indicators before and after Covid –19 pandemic. The three websites that we sourced the data are the most reliable sources for data analysis purposes. Moreover, the three websites are all open data sources, which exempt from possible legal liability for using the data.

What target you chose

The target of choosing the data is in preparation for data analysis on how Covid-19 pandemic affects the economy, education and demographic aspects worldwide.

What difficulties you had to overcome to wrangle the data sources into the target data model

Country codes for data sourced from different websites are not consistent. We needed to match them partially by hand.

File “GlobalEconomicProspectsJune2020GDPgrowthdata.xlsx” contains multiple tabs, and each tab consists of two parts with two themes. Wrangling this file, e.g. splitting and combining, were time consuming.

When assessing the impact of cumulative cases, population density and total cases in percentage of population on status of educational organisitions in October, extra data wrangling was required. Since multiple rows with same country names, we group them together with one country one row with corresponding figures. We can't group population density with simple calculation, we need to calculate corresponding square km first, then calculate density back by the sum of population and sum of square km, also fixing the comma messed up with decimal problem when we load the dataset into dataframe.

What techniques you did use

According to our topic, we classify the raw datasets into two parts for data wrangling.

(a) Economy

Data Wrangling using R

1. Import raw data in R. The raw data include WHO-COVID-19-global-data.csv, WDICountry.csv, WDIData.csv, GEPData.csv.
2. Due to the different sources of the data, we need to match the country codes. We select the country code of WDI Country dataset as the basis. We got 3 digits country codes and 2 digits country codes from WDI Country dataset. Join all country code from each dataset. We find country code in GEP dataset does not match in WDICountry dataset. Create a function to match country code.
3. Import GlobalEconomicProspectsJune2020GDPgrowthdata.xlsx into R basing on each sheet. The raw data includes multiple sheets, and some sheets include 2 datasets. We need to deal with them separately.
4. Remove NA columns and rows and renamed column to names that are easy to read and understand.
5. For a sheet containing two data sets, we use the which function to split the data. According to the key words, we locate specific rows and columns, and use number of rows or columns to split the data.
6. After splitting the data, we obtain data classified by 6 economic regions. Combine 6 datasets into one dataframe.
7. Add one more column as area.
8. We find some of the data contains subscripts and comma, using regular expressions to remove subscripts and comma.
9. Finally, we got dataset called RealGDP_by_areas.csv (see Table 1) and summary of economic by area dataset called Summary_by_areas.csv (see Table 2).

A	B	C	D	E	F	G	H	I
classification	2017	2018	2019e	2020f	2021f	2020fd	2021fd	area
Cambodia	7	7.5	7.1	-1	6	-7.8	-0.8	EAS
China	6.8	6.6	6.1	1	6.9	-4.9	1.1	EAS

Table 1

A	B	C	D	E	F	G	H	I
classification	2017	2018	2019e	2020f	2021f	2020fd	2021fd	area
EMDE EAP G	6.5	6.3	5.9	0.5	6.6	-5.2	1	EAS
GDP per capit	5.8	5.6	5.2	-0.1	6	-5.2	1	EAS

Table 2

10. Select GDP growth (annual %) data from WDIData. We got 2010-2016 GDP data from WDIData dataset and 2017-2021 GDP data from GEPData dataset.
11. Combine two parts of data into one dataframe.
12. We got a dataset includes GDP data from 2010 to 2021, called GDP_History_2010_2021_by_country.csv (see Table 3).

A	B	C	D	E	F	G	H	I	J	K	L	M	N
Country_Code	Country_Name	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
EAS	East Asia & Pac	7.069893	4.610478	4.694137	4.76982	4.189865	4.199749	4.091404	6.5	6.3	5.9	0.5	6.6

Table 3

(b) Education

Data wrangling using R

1. Load required libraries and import raw data in R.
2. Rename the Date column to a shorter name.
3. Overview the dataframe structure and count how many dates and countries are recorded in the dataset.
4. Create a function to translate WHO region names.
5. Reshape the covid-19 dataframe to a long table with all the numeric variables in one column for plotting.
6. Plot the average worldwide daily Covid-19 status over months.
7. Filter out cumulative cases value only and ready to plot top 10 countries.
8. Create average daily ranking column and two more columns and filter out top 10 countries.
9. Load population dataset, we only need the latest figures and one row from duplicates, rename with a suitable name.
10. Find out country name with “Other” and assign “Other” to the “Other” country.

11. Create a list with irrelevant countries and filter out the irrelevant rows by matching the area list.
12. Join the covid-19 dataframe with population dataframe with a full join by country names.
13. Find out the unmatched countries from covid-19 dataframe.
14. Create a function to match them manually.
15. Due to multiple rows with same country names, we group them together with one country one row.
16. We calculate corresponding square km, then calculate density by the sum of population and sum of square km, also fixing the comma messed up with decimal problem.
17. Remove missing values and add a column with cumulative Covid-19 cases on population percentage.
18. Find out how many days and countries are covered in the dataframe.
19. Join the Covid-19 dataframe with education open status dataframe with a left join on data and country.
20. Remove rows with missing open status values.
21. We got a dataset include Covid-19, education and population data (Table 4).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
		Date	Country_code	Country	WHO_region	New_case	Cumulative_cases	New_deaths	Cumulative_deaths	Pop_total	PopDensity	CumCasePop_percent	ISO	Status
1		######	AF	Afghanista	Eastern Medi	0	0	0	0	38928341	59627	0	NA	NA
2		######	AF	Afghanista	Eastern Medi	0	0	0	0	38928341	59627	0	NA	NA

Table 4

Data wrangling using Julia

1. Read HNP csv file into a dataframe.
2. Select the useful rows and columns we need.
3. Find out which row number the actual country name starts, not continent, area, organization or any other irrelevant rows.
4. Filter out the rows are only countries, not continent, area, organization or any other irrelevant rows.
5. Remove the missing rows and rename some columns to more sensible names.
6. Read Covid-19 .csv file and rename the columns' names.
7. We only need the WHO regions' information with corresponding countries from this dataset.
8. Create a function to translate WHO region names.
9. Outer join WHO region information with unemployment dataframe.
10. Find out the unmatched countries and list the countries still available to be joined.
11. Create a function to amend the country names to get join again.
12. We use left join to make a join again.
13. Fill the missing WHO region names to "Other".
14. Reshape a wide dataframe to a long dataframe.
15. Convert year column back to string data type.
16. Create a new column to calculate the unemployment rate change from year 2011 to 2020.
17. Rearrange the columns order.
18. Convert all columns to correct data type.

19. Finally, we got unemployment.csv file (see Table 5).

A	B	C	D	E	F	G	H	I	J	K	L	M	N
Country	ISO	WHO_region	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	19-20Diff
Afghanista	AFG	Eastern Medi	11.488	11.508	11.534	11.448	11.387	11.313	11.184	11.057	11.118	11.164	0.046
Albania	ALB	European	13.481	13.376	15.866	17.49	17.08	15.22	13.75	12.34	12.331	12.813	0.481999

Table 5

Data Visualization

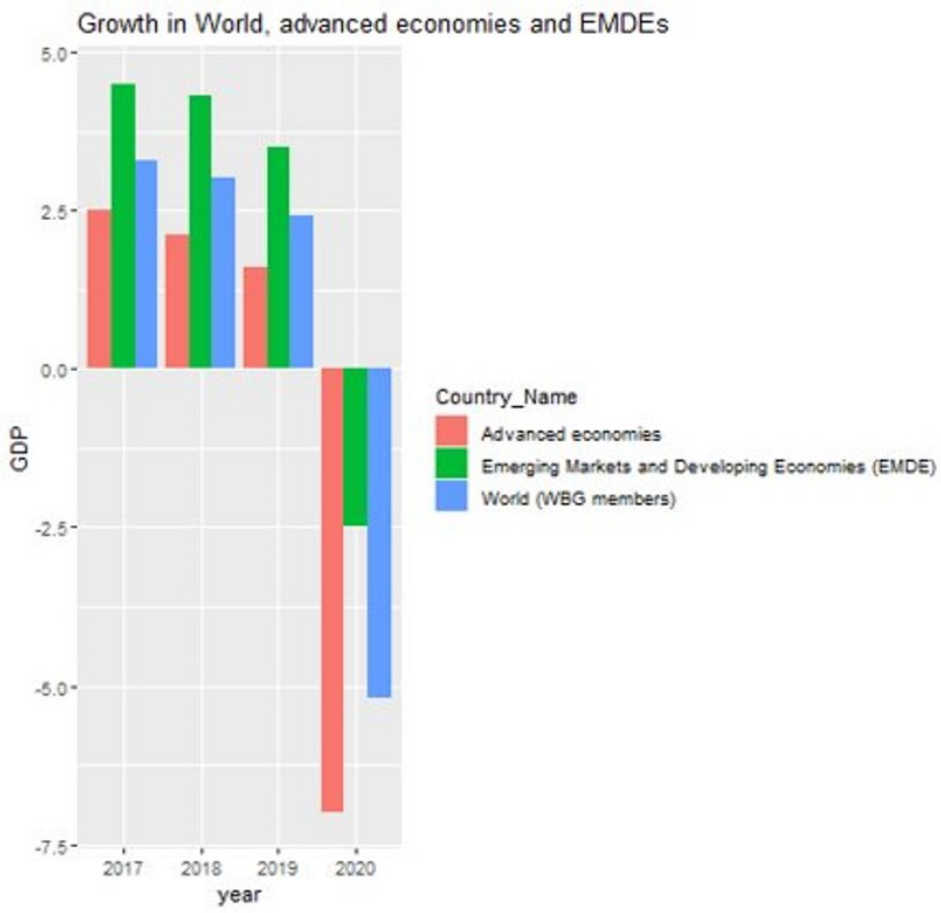

Figure 1

The histogram in Figure 1 reflects the GDP growth of the world, advanced economies and emerging market and developing economies over the last four years, and it is clear to see that GDP growth is negative in 2020, especially in advanced economies.

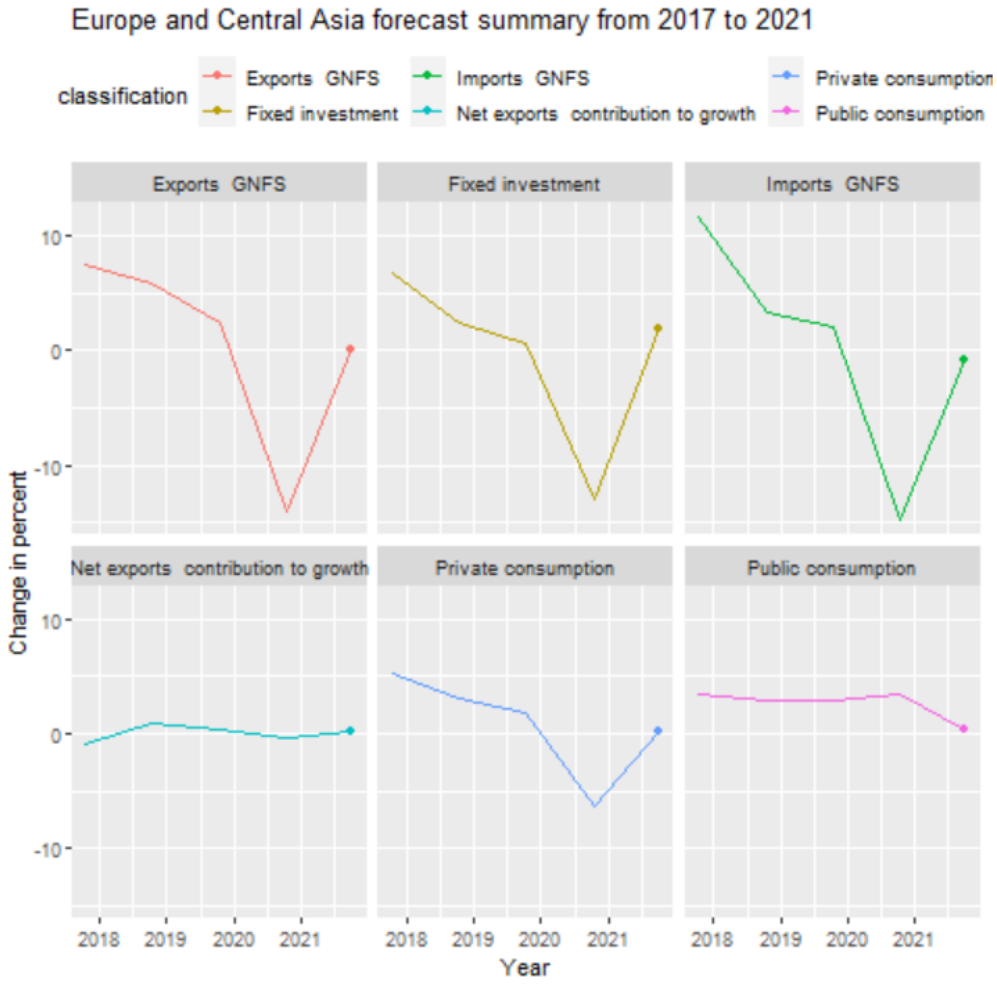

Figure 2

Figure 2 (originally a gif) contains percentage changes in six economic indicators for Europe and Central Asia between 2017 and 2021, with four of the six indicators declining significantly in 2020, while all four are projected to rebound in 2021.

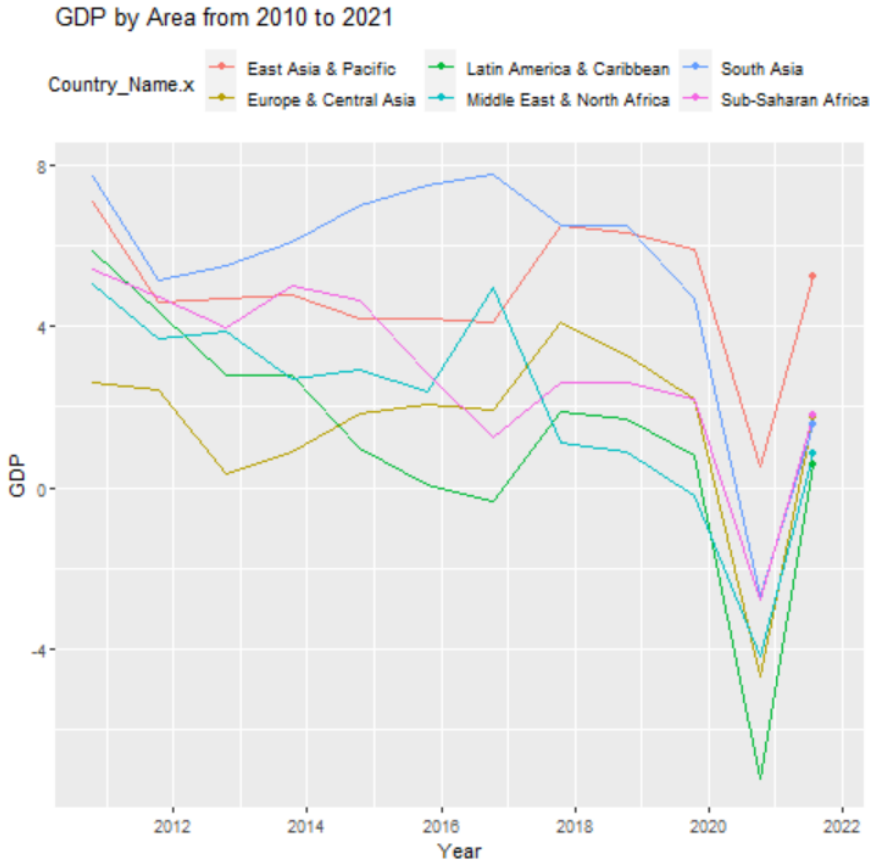

Figure 3

Figure 3 (originally a gif) reflects the percentage change in GDP from 2010 to 2021 for the six world regions, which shows that all six regions experienced significant declines and negative growth in 2020 and are expected to pick up significantly after 2021.

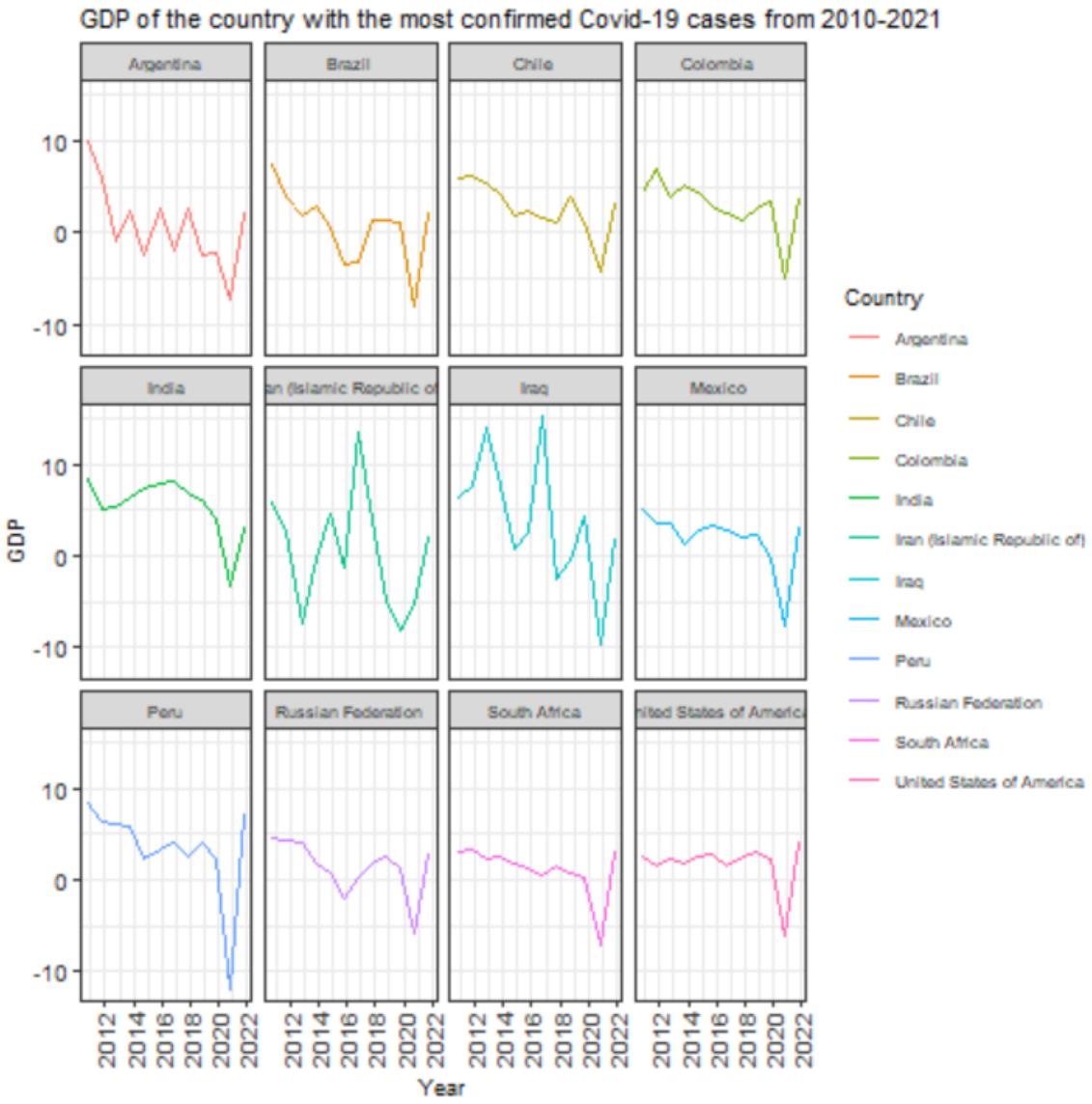

Figure 4

Figure 4 shows GDP growth over the past 10 years for the 12 countries with the highest number of confirmed Covid-19 cases, as well as projections for GDP growth in 2021 for these countries, all of which have experienced sharp declines and negative growth in 2020.

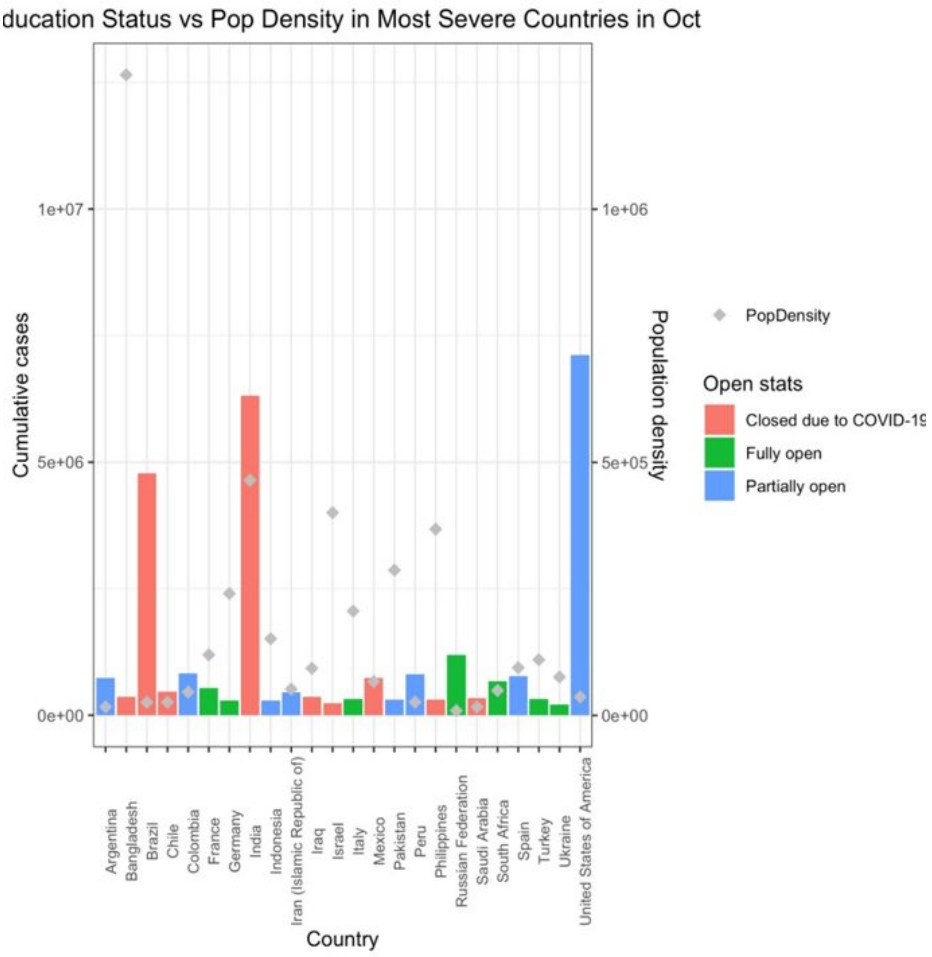

Figure 5

The bar chart in Figure 5 demonstrates how the status of educational organizations (Fully open/ Closed/Partially open) was impacted by the cumulative confirmed cases and population density in the most severe countries worldwide in October 2020.

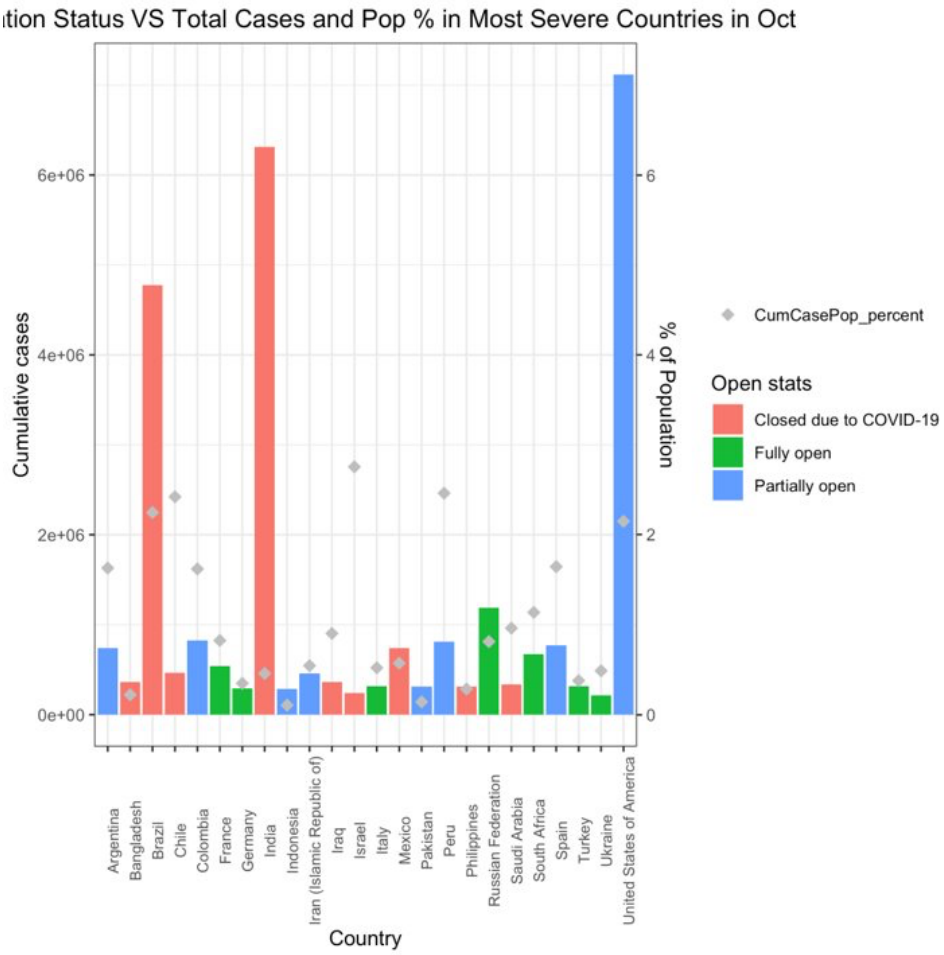

Figure 6

The bar chart in Figure 6 displays how the status of educational organizations (Fully open/ Closed/Partially open) was impacted by the cumulative confirmed cases and the percentage of the cumulative confirmed cases taking up to the total population in the most severe countries worldwide in October 2020.

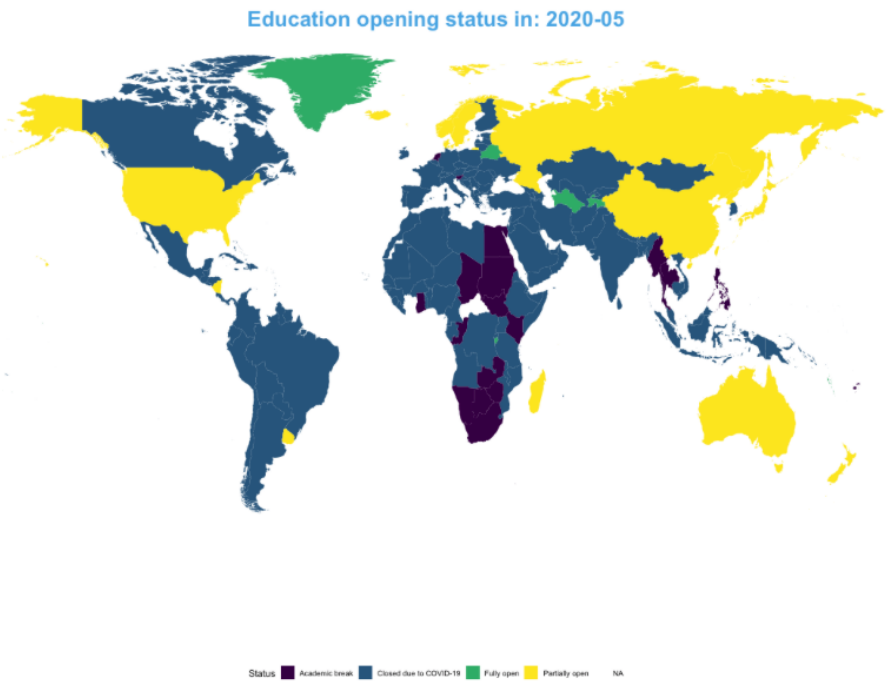

Figure 7

Figure 7 (it is a gif originally) shows the monthly changes of status of educational institutes worldwide from 2020 February to 2020 October.

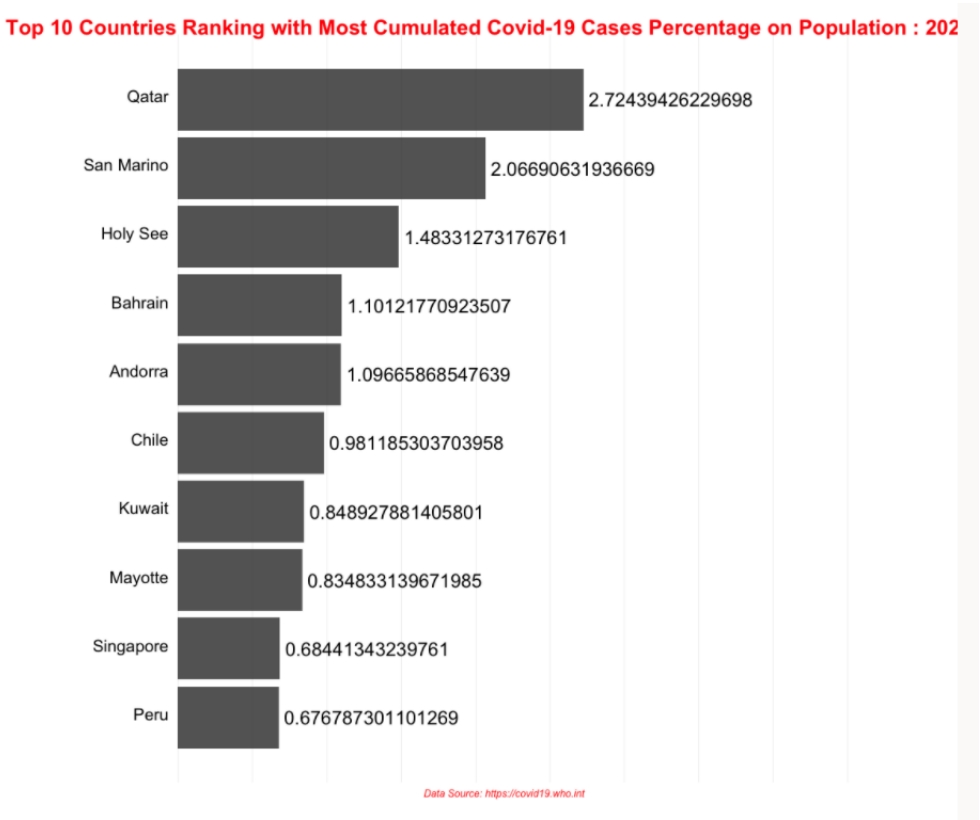

Figure 8

Figure 8 (originally a gif) demonstrates the top 10 countries with the most cumulative confirmed cases by percentage from 2020 Jan to present.

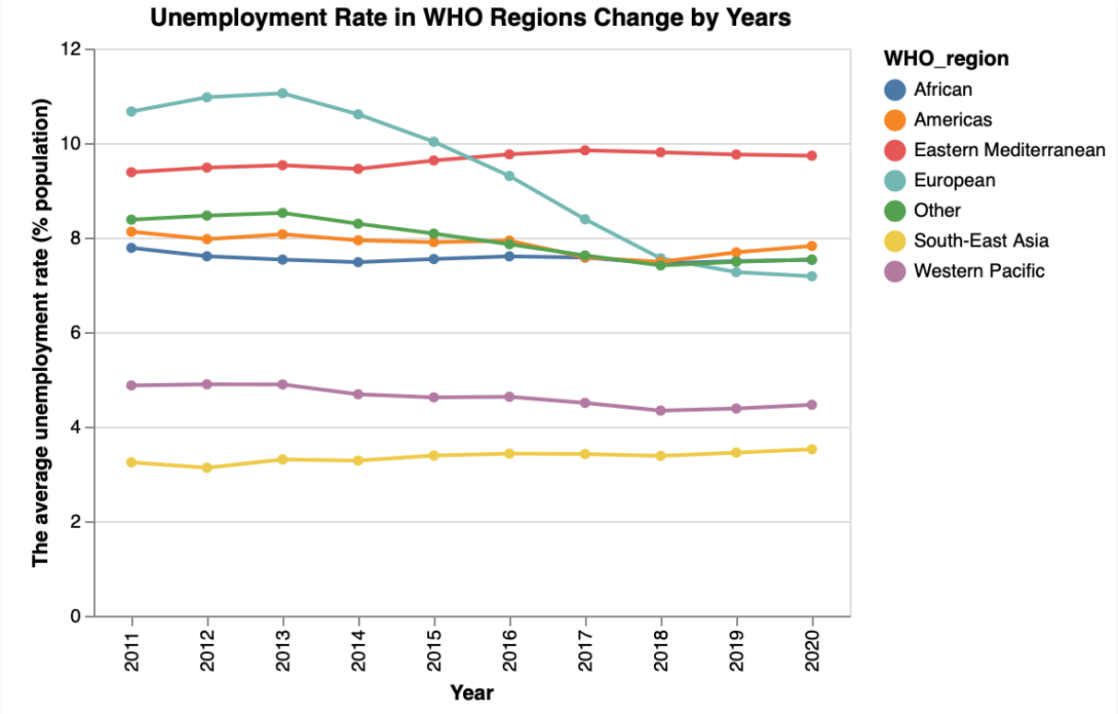

Figure 9

The line graph (Figure 7) shows the yearly unemployment rate during the last decade for the 6 economic regions defined by WHO. The result is out of our expectation, which might because of the data source has not been updated correctly.

What you managed to achieve and what you failed to do

We planned to scrap data from Twitter using Twitter API. However, we could not perform this plan finally for three reasons. First of all, we were advised by Dr Giulio Dalla Riva that scraping data regarding Covid-19 pandemic may be at the risk of too much noise in the raw data. Secondly, the workload of wrangling data with too much noise is beyond the scope of this project of this paper. Thirdly, it was difficult to acquire the permission of Twitter API from Twitter.

Legality and Ethics

In accordance with the terms and conditions of use of the datasets on the websites of the World Health Organization and the World Bank, both organizations encourage public access to and use of the data they collect and publish on their websites. Our use of these datasets in this project does not go beyond their permitted use, nor does it comply with their prohibited terms of use. Accordingly, this project does not violate any legal requirements or terms of use.

The data for this project was obtained from public platforms and is not involved with private or commercially confidential. Current research results may reflect the lack of control over Covid-19 in some countries or regions; however, negative assessments of governmental or public institutions' incompetence are within the realm of reasonable discussion and do not break ethical boundaries. Therefore, this project is not unethical, nor does it cause harm to any individual or institution.

Intended use

Due to the very broad scope of the project, the current results can be used as a basis for many different research directions. For example, the next step could be to develop a time series model to predict future changes in Covid-19 and its impact on the economy and education.