CSC343 Project Phase 1

Terry Tu, Owen Zhang

Thursday, September 30, 2021

Project Domain

Recently, a United Nations news report that the COVID crisis may increase global unemployment up to 200 million in 2022. This new makes our interest in how Covid-19 affects the global unemployment rate and different elements to create different inflections in countries [1]. Therefore, we found some datasets connected to these messages from 2019 to 2020, which are pre-pandemic and during the pandemic. These data sets help us explore the unemployment rate from different angles, including the distinct economic entity in different continents, countries in various income levels, and the severity that the country inflect.

Data Sets

1. Country Information Data Set
   (a) Data Set Link: https://www.kaggle.com/andradaolteanu/country-mapping-iso-continent-region
   (b) Information Needed: country code, country name, region, sub-region, region code, sub-region code.
   (c) Relevant Learning: we have leaned countries code and region code' meaning and information about different economies entities around the world.
   (d) Clean Up: We separated the Country Information Data Set into three sub-data relations: country, Region, and Sub-Region
2. 2020 Unemployment Rate Data Set
   (a) Link: https://data.oecd.org/unemp/unemployment-rate.htm
   (b) Information Needed: country code, unemployment rate 2019, unemployment rate 2020
   (c) Relevant Learning: The relationship between the unemployment rate and the economic status of a country, and the general impact of COVID-19 to a country's unemployment rate.
   (d) Clean Up: we only keep the countries' unemployment rate of the year 2019 and 2020 and drop off data in other years.
3. Country Income Level Data Set
   (a) Link: https://data.worldbank.org/indicator/SL.UEM.TOTL.ZS
   (b) Information Needed: country code, income group.
   (c) Relevant Learning: The meaning of high, medium, and low income group, and its relation to the development of a country.
   (d) Clean Up: Remove the Region and SpecialNotes columns. Only keep the country code and the income group the country belongs to.
4. 2020 COVID-19 Cases Data Set
   (a) Link: https://www.kaggle.com/bolkonsky/covid19?select=owid-covid-data.csv
   (b) Information Needed: country code, total cases per million on September 1 2020, total death per million on September 1 2020.
   (c) Relevant Learning: The significance of cases per million and death per million resulted by COVID-19 to a country.
   (d) Clean Up: Only keep the COVID-19 data of a country on September 1, 2020.

Questions to Investigate

1. What is the relation between the quantity of Covid-19 cases and the unemployment rates changes in global countries?
2. How is COVID-19 affecting countries on various continents and sub-regions?
3. Are there different inflections from Covid-19 to countries with varying levels of income?

Schema

Relations

(a) Country(Name, countryCode, regionCode, subRegionCode)

A tuple in this relation represents a Country unit. Name is there formal country name. countryCode is their national code to represent their country. regionCode is the code represent their continent. subRegionCode is the code represent their sub-continent.

(b) Region(regionCode, region)

A tuple in this relation represents the region unit. Region is the continent coresponding to the continent.

(c) SubRegion(subRegionCode, subRegion)

A tuple in this relation represents the sub-region unit. Sub-region is the sub continent corresponding to a sub-region code.

(d) Unemployment(countryCode, 2019UR, 2020UR)

A tuple in this relation represents the unemployment rate per country in different year. countryCode is their national code to represent their country. 2019UR is the unemployed rate for countries in 2019. 2020UR is the unemployed rate for countries in 2020.

(e) IncomeLevel(countryCode, incomeGroup)

A tuple in this relation represents the Income Level for each country in the global status. countryCode is their national code to represent their country. incomeGroup is the income status for countries in the global which separate as high income, upper middle income and lower middle income.

(f) COVID-Data(countryCode, total-case-per-million, total-death-per-million)

A tuple in this relation represents the Covid-19 cases for each country by monthly in 2020. countryCode is their national code to represent their country. total-case-per-million is the data that calculate population by the total case suffer from the Covid-19. total-death-per-million is the data that calculate population by the total case passed away due to the Covid-19.

Integrity Constraints

(a) Unemployment[countryCode] $\subseteq$ Country[countryCode]

(b) IncomeLevel[countryCode] $\subseteq$ Country[countryCode]

(c) COVID-Data[countryCode] $\subseteq$ Country[countryCode]

(d) Country[regionCode] $\subseteq$ Region[regionCode]

(e) Country[subRegionCode] $\subseteq$ SubRegion[subRegionCode]

(f) IncomeLevel[incomeGroup] $\subseteq$ {"High income", "Upper middle income", "Lower middle income", "Low income"}

Data Dictionaries

Country Dictionary

Attribute	Description	Type	Required
Name	The name of the country that we used	TEXT	Yes
countryCode	The 3-letters national code to represent their country.	TEXT	Yes
regionCode	The code represent countries continent.	TEXT	Yes
subRegionCode	The code represent countries sub-continent.	TEXT	Yes

Figure 1: Data dictionary for Country relation

Region Dictionary

Attribute	Description	Type	Required
regionCode	The code represent countries continent.	TEXT	Yes
region	The regions in the world.	TEXT	Yes

Figure 2: Data dictionary for Region relation

SubRegion Dictionary

Attribute	Description	Type	Required
subRegionCode	The code represent countries sub continent	TEXT	Yes
subRegion	The sub regions in the world.	TEXT	Yes

Figure 3: Data dictionary for SubRegion relation

Unemployment Dictionary

Attribute	Description	Type	Required
countryCode	The national code to represent their country.	TEXT	Yes
2019UR	The unemployed rate for countries in 2019.	INT	Yes
2020UR	The unemployed rate for countries in 2020.	INT	Yes

Figure 4: Data dictionary for Unemployment relation

IncomeLevel Dictionary

Attribute	Description	Type	Required
countryCode	The national code to represent their country.	TEXT	Yes
incomeGroup	The income status for countries in the global.	TEXT	Yes

Figure 5: Data dictionary for IncomeLevel relation

COVID-Data Dictionary

Attribute	Description	Type	Required
countryCode	The national code to represent their country.	TEXT	Yes
total-case-per-million	The total case suffer from the Covid-19 until 2020/09/01.	INT	Yes
total-death-per-million	The total case passed away due to the Covid-19 until 2020/09/01.	INT	Yes

Figure 6: Data dictionary for COVID-Data relation

Justification of Design

In our project, We combined economic knowledge and geography knowledge to make our design more professional. We explore the unemployment rate from different angles, including the distinct economic entity in different continents, countries in various income levels, and the severity that the country inflects during the pandemic.

We make the country code the key through datasets since the national country code is the same, which could help us connect each dataset accuracy by country. In the dataset of the Covid-19 Data, we keep the data of total cases for people suffering from and death by the Covid-19, which can help us make the data more comprehensively.

We also separated the Country Information Data Set into three sub-data relations: country, Region, and Sub-Region. Since most of our question only requires the country code and country name. It is a good idea to separate the region and sub-region information from the Country table. When needed to combine the country and the country's continent, we can use the regionCode and subRegionCode to integrate with the Region and SubRegion table.

We only keep the income level column and country code in the income countries data set since the original data set containing a suit separate income group connected to our project questions.

For the unemployment data set, we will be using the country code and the country's unemployment rate for the years 2019 and 2020. We chose these two particular years since the year 2019 represents the unemployment rate of pre-pandemic, and 2020 represents the unemployment rate during the pandemic. Comparing these two years will give us the information we need to analyze the impact of COVID-19 on a country's unemployment rate.

References

[1] "COVID crisis to push global unemployment over 200 million mark in 2022 — — UN NEWS," United Nations. [Online]. Available: https://news.un.org/en/story/2021/06/1093182. [Accessed: 30-Sep-2021].