Hung Quoc Tran
CIS - 4400
Prof. Emily Mazo

COVID-19 Situations in Indochina Countries from June - November 2021

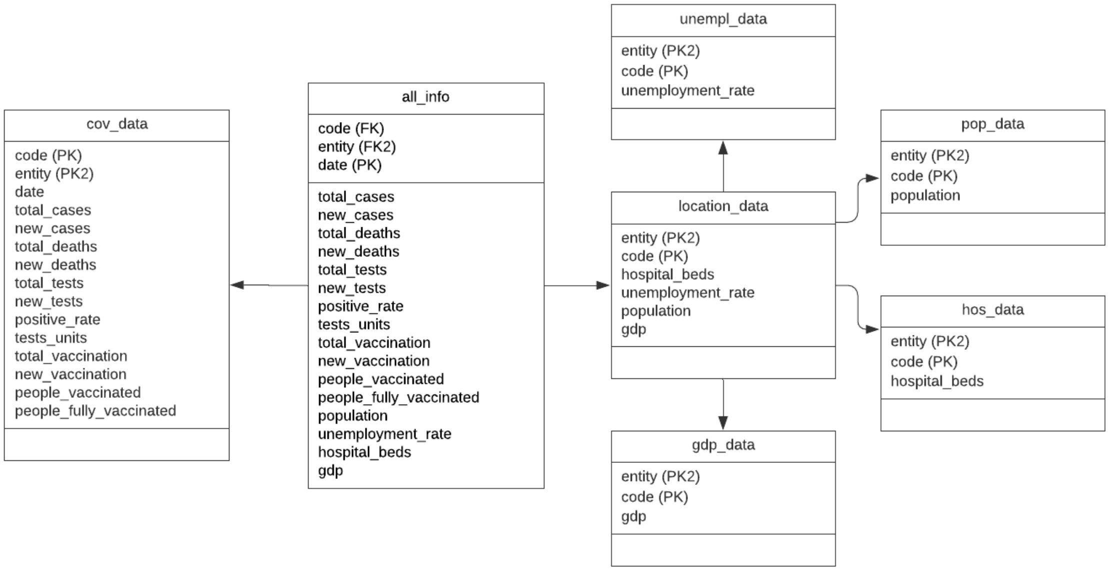
1. Proposal:

In this project, we are focusing on Covid-19 situation in Indochina countries, which are Cambodia, Laos, Myanmar, Thailand, and Vietnam. In detail, we want to understand how factors such as number of people vaccinated, GDP, unemployment rate, hospital beds, population are correlated with the cases that increase daily in these countries. Thus, we decided to segment the data demographically based on geography to get a picture of whether places with more vaccinated people, hospital beds, and wealthier show signs of decreasing daily cases.

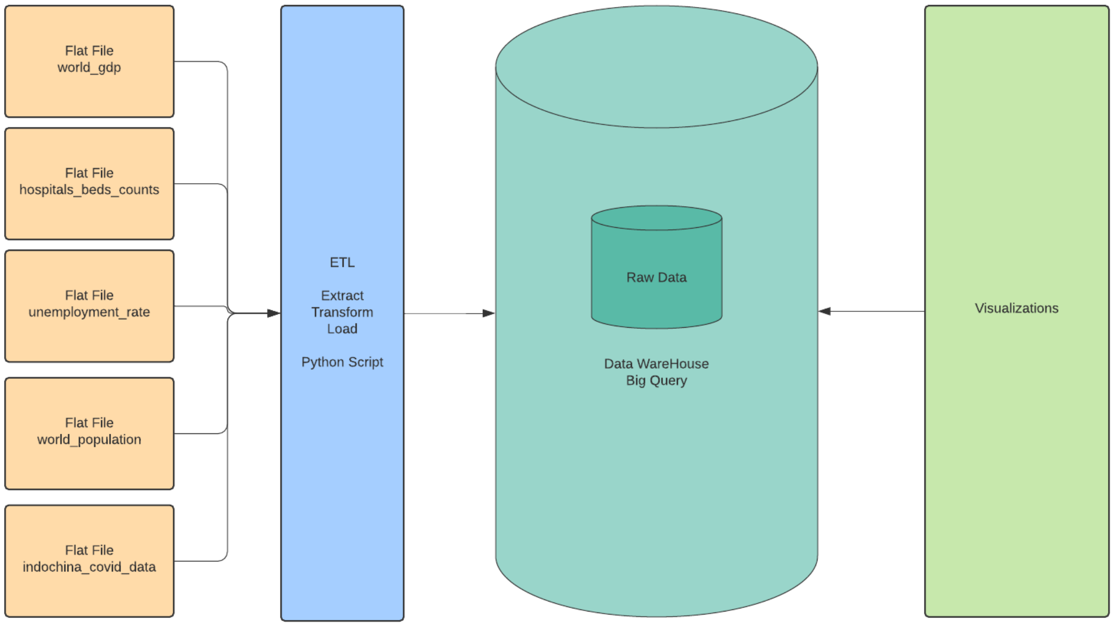
2. Problem:

We are trying to find clear data sets of covid cases in Indochina countries. When we look in-depth in our datasets, we see that there is missing data in different columns and rows. In our data warehouse, we will be able to get the necessary coverage of different countries to do this analysis. Doing further research, we were able to see that many sick people could not come to the hospitals due to shortage of hospital beds or being unable to afford the care due to possibly being unemployed. We believe it is possible that the case numbers that we see in our data would not have been this bad if people had access to vaccines and healthcare despite life situations. Because of this, we are also analyzing the vaccination rate of each country, how many people are vaccinated and fully vaccinated. We will also be tracking the number of cases, deaths, and tests from COVID in each country; moreover we will expand our analysis to whether resources like hospital beds, GDP, and unemployment rate would make a difference in the death rate in different states. Through our database, one will be able to compare and get proper and detailed information about the covid cases in Indochina countries.

Logical Data Model



Architecture Diagram



3. Detailed Design

I. Data Sources

- world_gdp, hospital_beds_counts, unemployment_rate, world_population:

These data sets contain all information on factors such as GDP, unemployment rate, hospital beds, population of the world, which we will extract only from Indochina countries for usage. We got these datasets from **Our World in Data** (ourworldindata.org)

- indochina_covid_data:

The data set contains vaccination information in Indochina countries. In detail, the information includes total/new cases, total/new deaths, total/new vaccination, total/new tests, people vaccinated, people fully vaccinated... We got these datasets from **Our World in Data** (ourworldindata.org)

II. ETL

All the datasets above are csv files.

We chose the datasets as they can best represent the situation of Covid-19 demographic in Indochina Countries daily from June 2021 to November 2021.

We will write a python script performing the ETL process on these flat files locally on my computer. Consequently, we have ourselves a data warehouse loaded onto Big Query. Big Query is free and its data management is efficient for such small size data we have acquired for this class. Moreover, I have experiences working with Big Query beyond individual assignments in this class so Big Query is even more convenient.

III. BI (Business Intelligence) Visualization

We will go with Tableau for visualization. However, if any problems occur, we will use Google Data Studio instead.

IV. Ethical Issues Taken into Consideration:

One of the main “risk-zones” for our project is data privacy and protection. We must ensure that confidential information is secured in accordance with what is being kept, how it is limited in the Data Warehouse, and how it can be retrieved via our business intelligence tools. Enforcing rules at the database level is by far the most effective way to minimize access to the appropriate people. Personal health information should be kept very confidential and secured.

Luckily, our data is general information at countries level being available for public access; thus, there is no sensitive personal health information residing in our data warehouse. However, we still run our data warehouse locally with a 2-step password verification Google account for Big Query; therefore we're still being as cautious as possible with our project.

A. Is it ethical to integrate everything into the data warehouse?

The data that we are using is from public sources and they are of countries. Thus no personal health information is being used as no individuals are exposed to a privacy breach looking at our data. Despite such, we always act responsibly and apply the highest standards of research integrity and ethical behaviors. No matter if a formal ethics approval process is involved or not.

B. Is our data accurate and can any organization use it in future?

Yes, our data is accurate as the source is from **Our World in Data** (ourworldindata.org), which is widely known as a credible open source used by data seekers such as data analysts, data engineers, and data scientists. Thus, any organization can use or refer to our data with 100% accuracy. I double checked the numbers with other sources, for example, Vietnam's government Covid data source, and they matched. However, the data is only from June - November 2021, which will be outdated really soon so I'm not sure on how valuable and useful our data is.

C. What have we done to keep our data safe from hackers?

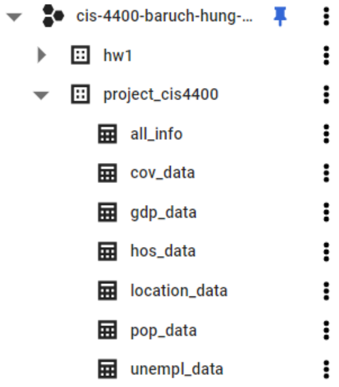
Our data is shared privately among 3 members of the group making this project, who have write-access to it on Big Query. Moreover, there's only one Google account for this data warehouse on Big Query, which has a 2-step verification password so that it's really hard to hack from an anonymous computer. Ultimately our data is from open sources and contains no personal information of any individuals so there's no point for hackers to attack us.

D. Do our outcomes or intentions hurt others? How does our technology help others?

This project is meant to seek a possible suggestion to improve the COVID-19 situation of Indochina countries. There's no potential harm, let alone intention to hurt others. The project will determine the relation between healthcare & wealth and the covid situation of Indochina countries. Our data will display the number of hospital beds available, GDP, unemployment rate, population… in accordance to the cases, deaths, tests, and vaccinated people. Consequently, this will help the countries to see which aspect needs to be improved in order to fight against Covid-19.

4. Technical Documentation

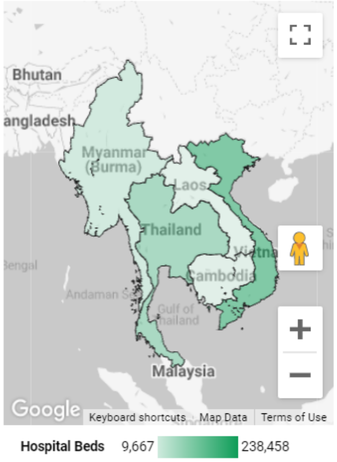
- **ETL Process** (Python Script) with **Documentation** please review this repo on GitHub: https://github.com/erictq96/ETL-Data-Warehouse-Hung-Eric-Tran
  1. Milestone 4 - CIS 4400 - ETL.ipynb - ETL Process
  2. README.md - Documentation
- **Schemas**:

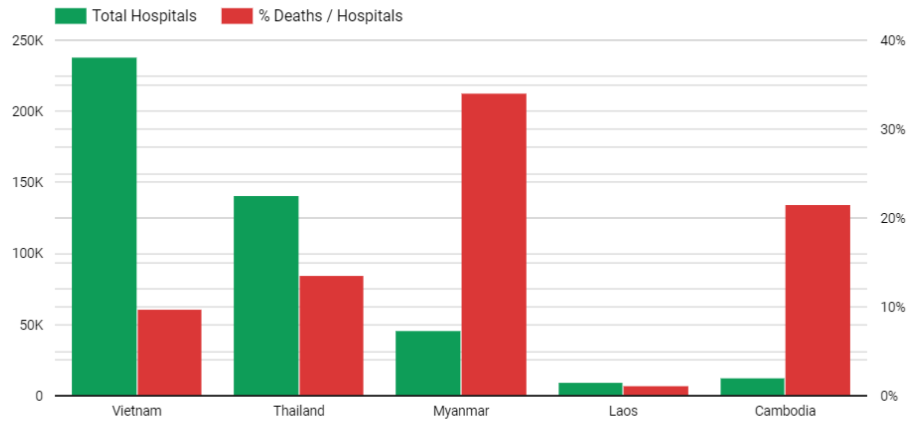


cis-4400-baruch-hung-...
hw1
project_cis4400
all_info
cov_data
gdp_data
hos_data
location_data
pop_data
unempl_data

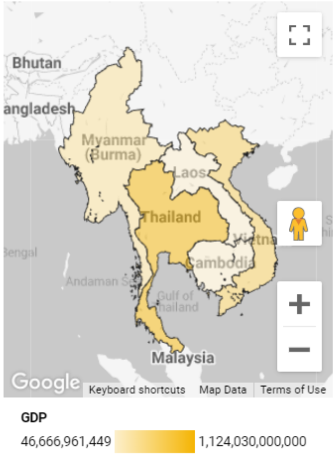
5. Analysis

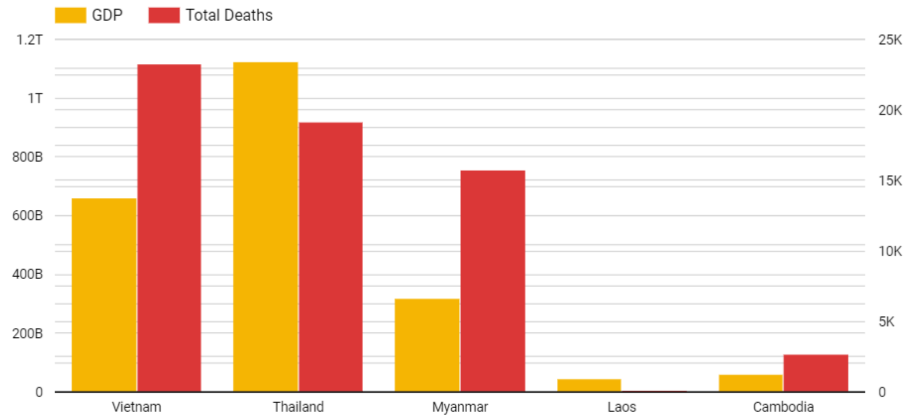
- Number Hospital Beds:
  Comparing total number of deaths & total number of hospital beds, we can see on the graph that as a country has more hospital beds the death rate is lower (Vietnam & Thailand), and vice versa (Myanmar & Cambodia).



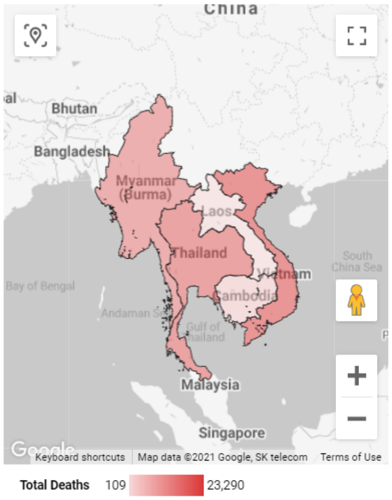


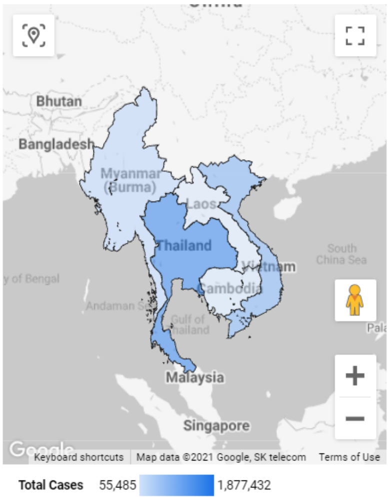
- GDP:
  As a country has a high GDP, the number of deaths is low. In other words, the wealthier the country, the less deaths it has, and vice versa.

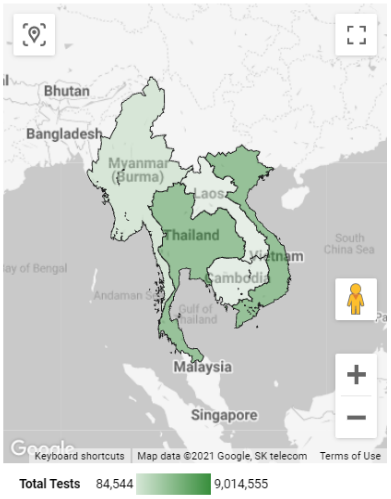




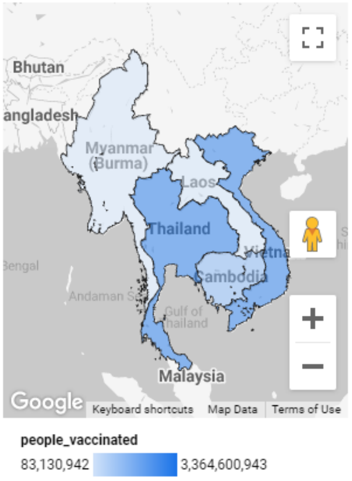
- Infographic on total cases, total deaths, and total tests:

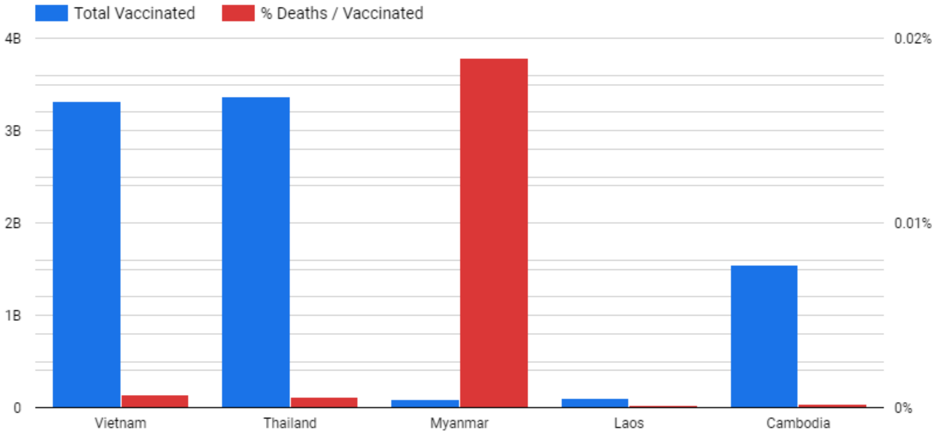






- Infographic on vaccinated vs deaths:
  Countries with a high number of vaccinated people (at least one dose) have a significantly low death rate and vice versa.





6. Conclusion

Having finished this data warehousing project, I find myself to gain much more knowledge about how to build relational models, ETL pipelines, and visualizations. The most challenging part for me is finding the data I needed because before finalizing this dataset, we were working with Covid-19 Data of India and there was so much missing data that we couldn't move forward with the project. Moreover, building a relational model is quite hard as well since I have not fully understood about the knowledge. Thus, I'm not sure if the one we built above is correct or not, though it looks fine to me. On the other hand, the most enjoyable part is to write a python script doing the ETL process, especially because I love working with python. However, the process of using python to load our data warehouse into Big Query is completely new to me and I found it very valuable as I also work with Big Query. The visualization is quite difficult to me. I was originally going to use Tableau but failed due to being unable to download the software. Thus, I chose Google Data Studio instead, and there I was facing the problems of visualizing the data that I have. Consequently, the visualization got questioned by the professor and other students since it was potentially "biased" and "misleading", which I totally agree with.
The ethical part of managing data of this project is quite easy since our data doesn't breach any privacy issue due to having general information of the countries rather than personal ones, thus not having any potential harms to anyone. If I have to redo this project all over again, I would definitely begin with choosing another set of data with a deeper level of information that sparks my interests more, then go with Tableau for visualization rather than Google Data Studio.