COVID-19 PANDEMIC

Exploring the Drivers Behind

Presented by
Yu Ye, Yuhan Chen, Ling Yang, Xueting Wang
Data Engineering Project | Dec 10 , 2021

Team Presentation

Xueting Wang

Graduated from the University of Rochester

Full-time Student

Yuhan Chen

Graduated from the University of Rochester

Full-time Student

Ling Yang

Graduated from the Ohio State. University

Full-time Student

Yu Ye

Graduated from Penn State University

Full-time Student

Our Roadmap


Executive Summary
1
Business Use Case
2
Data Preparation & Analysis
3
Database Design & Modeling
4
Reports & Dashboards
5
Future Improvements
6

Executive Summary

Since the beginning of 2020, coronavirus has affected and changed the whole world from many perspectives such as economy, health system, and politics. In this project, we aim to discover the statistical distribution of COVID-19 related data, and how different factors could affect number of cases and mortality rate across the world. The major dataset is a collection of the global COVID-19 data and associated information, including number of cases and deaths, poverty rate and hospital beds, and stringency index, etc. Other minor datasets is a combination of covid data across the United States. Different implementation tools are used for data preparation and analysis. Visualizations are done through Tableau, a powerful tool for driving insights.

Our analysis can provide useful tools for visualization of the pandemic data. It can also provide guidelines for future research on factors associated with the number of cases and mortality rate, as well as the comparisons among continents and countries.

Business Use Case

I. How is the COVID-19 mortality statistics distributed around the world?

II. What are the drivers behind that lead to high or low number of cases and mortality rate across the world? Are there any differences between Asia and North America?

III. Can we build a user-friendly visualization tool for public to explore the pandemic situation across the United States?

1. Data Analysis and Preparation

Implementation Tools

We extracted our data from different platforms: AWS, GitHub, Wikipedia

We used Juptuer Notebook to extract data and perform data cleaning via Python, and used Excel to further prepare the data

We used MySQL Workbench to design our database

We used Tableau to create reports and dashboards to drive insights

Data Source → **Data Processing** → **Database Design & Modeling** → **Data Visualization**




Data Profile

Datasets Description : We extracted 2 major datasets – One from Github, Another from AWS S3 Explorer

Covid_world_stats.csv

Data Size: 34.7 MB
Number of observations: 126,248
Number of Attributes: 67

Covid_us_stats.csv

Data Size: 1.8 MB
Number of observations: 35,173
Number of Attributes: 4

Two minor datasets are scraped from Wikipedia pages and combined with the US data

Covid_us_stats.csv + us_population.csv us_poverty.csv = Covid_us_combined.csv

Data Processing

Problem	Action Taken
Large number of attributes – only need a few desired	Discuss and extract desired attributes
Abnormal data in the column “Continent_name” and “Country_name”	Drop null values
The column "Stringency Index" is not fully up to date	Fill the null values with the data from latest date
Multiple scraped data from Wikipedia	Combine the data with excel vlookup

Code Snippet

```
reduced_df = pd.DataFrame({
    "Continent_Name": combined_df["continent"],
    "Country_Name": combined_df["location"],
    "Date": combined_df["date"],
    "Number_of_cases":combined_df["total_cases"],
    "Number_of_deaths":combined_df["total_deaths"],
    "Stringency_index":combined_df["stringency_index"],
    "Population":combined_df["population"],
```

```
reduced_df = reduced_df.dropna
(subset = ['Continent_Name'])
```

```
pd.set_option('mode.chained_assignment', None)
for i in range(len(split_df)):
    for j in range(len(split_df[i])-1):
        a = split_df[i].iloc[j]["Stringency_index"]
        b = split_df[i].iloc[j+1]["Stringency_index"]
        if np.isnan(b):
            split_df[i]["Stringency_index"].iloc[j+1] = a
```

	date	state	cases	deaths	poverty_rate	population
0	2020/3/13	Alabama	6	0	15.50%	5024279
1	2020/3/14	Alabama	12	0	15.50%	5024279
2	2020/3/15	Alabama	23	0	15.50%	5024279
3	2020/3/16	Alabama	29	0	15.50%	5024279
4	2020/3/17	Alabama	39	0	15.50%	5024279
5	2020/3/18	Alabama	51	0	15.50%	5024279
6	2020/3/19	Alabama	78	0	15.50%	5024279

Sample Result

Continent_Name	Country_Name	Date	Number_of_cases	Number_of_deaths	Stringency_index	P
Asia	Afghanistan	2020-02-24	5.0	NaN	8.33	
Asia	Afghanistan	2020-02-25	5.0	NaN	8.33	
Asia	Afghanistan	2020-02-26	5.0	NaN	8.33	
Asia	Afghanistan	2020-02-27	5.0	NaN	8.33	
Asia	Afghanistan	2020-02-28	5.0	NaN	8.33	

```
reduced_df.isnull().sum()
```

Continent_Name	0
Country_Name	0
Date	0
Number_of_cases	7681

```
merged.isnull().sum()
```

Continent_Name	0
Country_Name	0
Date	0
Number_of_cases	0
Number_of_deaths	0
Stringency_index	0
Population	0
Population_density	0
GDP	0
Human_development	0
Poverty	0
Cardiovasc_deathrate	0
Diabetes	0

Data Overview

After data processing, the major dataset (covid_world_stats.csv) has 17 attributes in total.

Attributes	Description
Continent_Name	Continent of the geographical location
Country_Name	Country location
Date	Date of observation
Number_of_cases	Total confirmed cases of COVID-19
Number_of_deaths	Total deaths attributed to COVID-19
Stringency_index	Government Response Stringency Index
Population	Continent of the geographical location
Population_density	Number of people divided by land area, measured in square kilometers
GDP	Gross domestic product at purchasing power parity

Attributes	Description
Human_development	A composite index measuring average achievement in three basic dimensions of human development—a long and healthy life, knowledge and a decent standard of living.
Poverty	Share of the population living in extreme poverty
Cardiovasc_deathrate	Death rate from cardiovascular disease in 2017
Diabetes	Diabetes prevalence (% of population aged 20 to 79) in 2017
Female_smokers	Share of women who smoke
Male_smokers	Share of man who smoke
Hospitalbeds	Hospital beds per 1,000 people, most recent year available since 2010
Life_expectancy	Life expectancy at birth in 2019

2. Data Modeling and Design

Design Considerations

Data Integrity

- ✓ Set unique primary key for each table and foreign key relationships
- ✓ Replace NULL value with 0 or average number according to meaning of variables

Data Type

- ✓ Integer -- primary key and numbers
- ✓ String -- continent, country and state names
- ✓ Decimal -- factor numbers (eg. percent of smokers and GDP per capita)
- ✓ Follow standard naming convention for attributes

Database Modeling

- ✓ OLTP normalized physical entity-relationship model
- ✓ Implement best practice of normalization to avoid data redundancy and unnecessary joins across tables

Data Storage

Bucket details

covid_data_dataengineering_team5

Location	Storage class	Public access	Protection
us (multiple regions in United States)	Standard	Not public	None

OBJECTS | CONFIGURATION | PERMISSIONS | PROTECTION | LIFECYCLE

Buckets > covid_data_dataengineering_team5

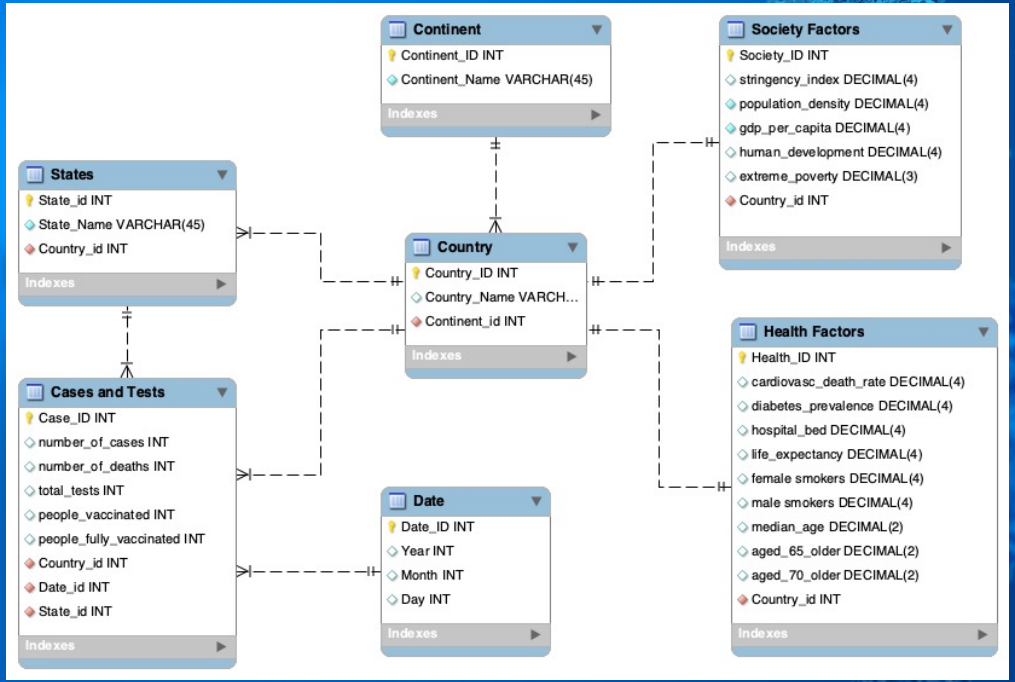
UPLOAD FILES | UPLOAD FOLDER | CREATE FOLDER | MANAGE HOLDS | DOWNLOAD | DELETE

Filter by name prefix only | Filter | Filter objects and folders

Name	Size	Type	Created	Storage class	Last modified	Public access
covid-19-world-cases-deaths-testing.cs	33.1 MB	text/csv	Dec 9, 2021, 6:33:01...	Standard	Dec 9, 2021, 6:33:01...	Not public
covid_us_stats.csv	1.7 MB	text/csv	Dec 9, 2021, 6:32:45...	Standard	Dec 9, 2021, 6:32:45...	Not public
covid_world_stats.csv	12.7 MB	text/csv	Dec 9, 2021, 6:32:53...	Standard	Dec 9, 2021, 6:32:53...	Not public
population.csv	1 KB	text/csv	Dec 9, 2021, 6:32:45...	Standard	Dec 9, 2021, 6:32:45...	Not public
poverty.csv	952 B	text/csv	Dec 9, 2021, 6:32:46...	Standard	Dec 9, 2021, 6:32:46...	Not public

EER Diagram

- The main table *Case and Tests* is connected to *Date, Country* and *States* table. It represents case related information for country or US state each day.
- *Continent, Social Factors* and *Health Factors* which contains basic information of each country are connected to the *Country* table using Country_id.



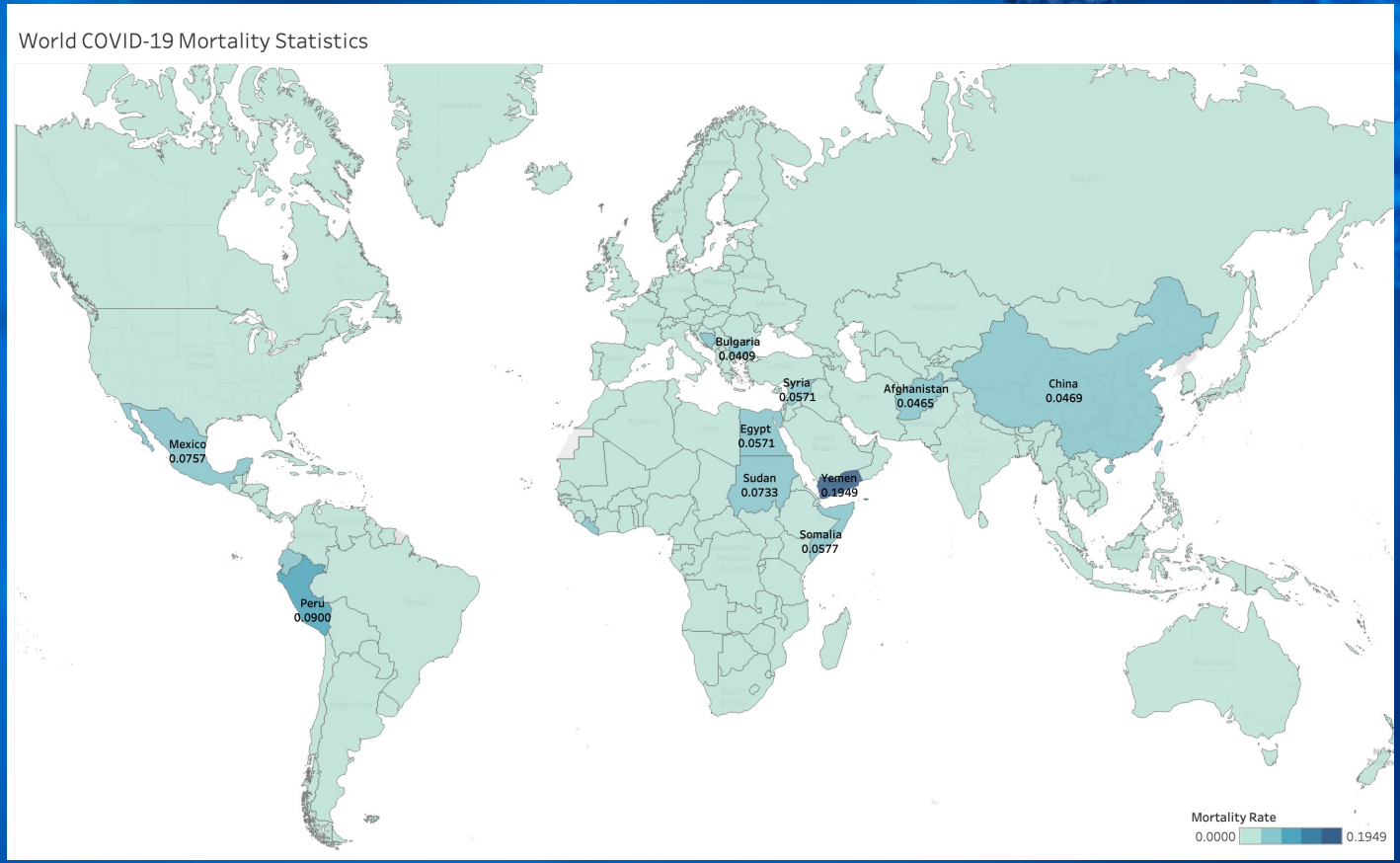
3. Reports and Dashboards

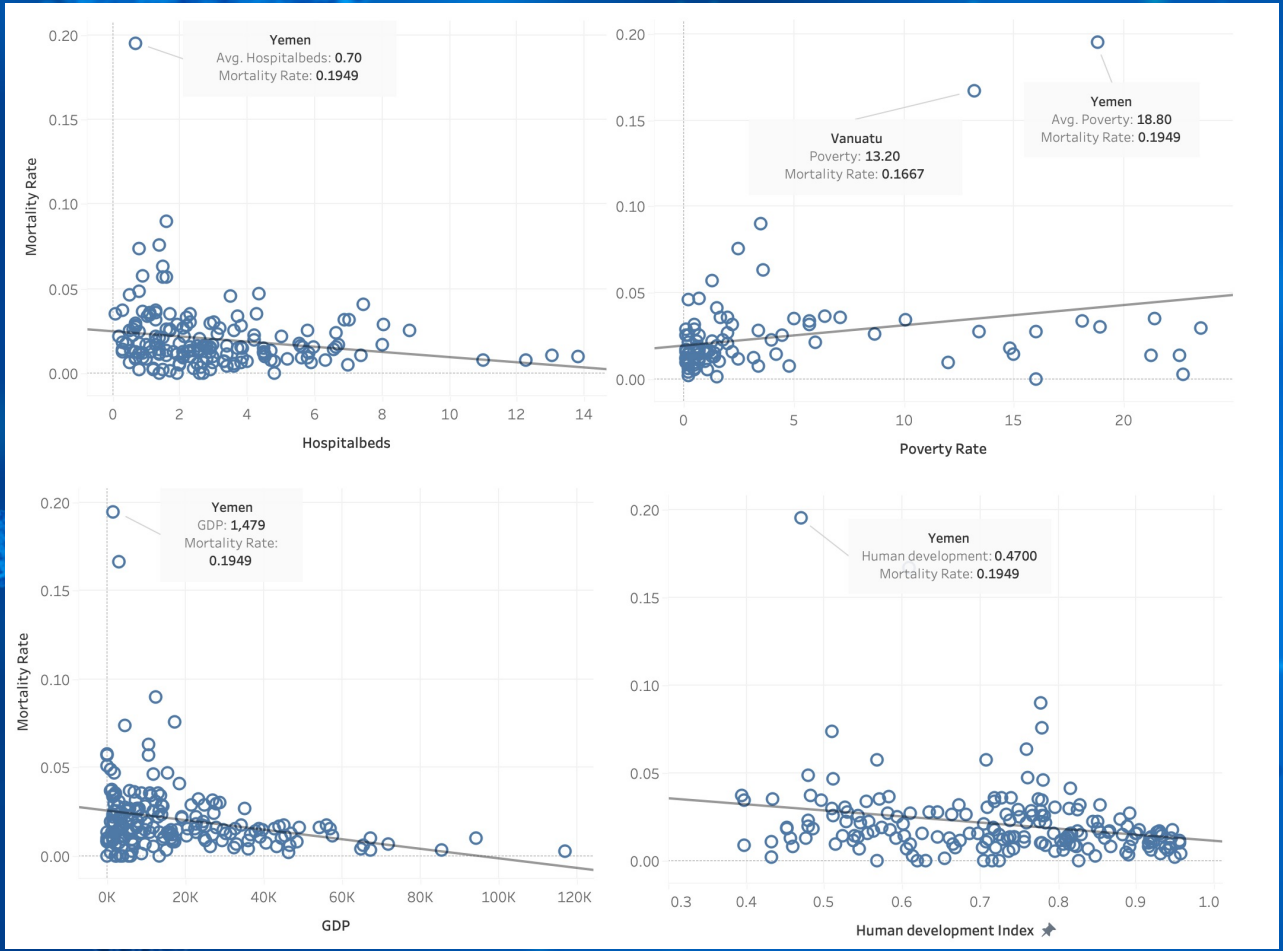
1 From a global perspective...

How is the COVID-19 mortality statistics distributed around the world and what are the drivers?

Visualization #1:
A map showing the world covid-19 mortality statistics:

Yemen has the highest mortality rate across the world. Each continent has a country with an obvious high mortality rate. In South America, Peru has the highest rate, and in North America, Mexico has the highest mortality rate.



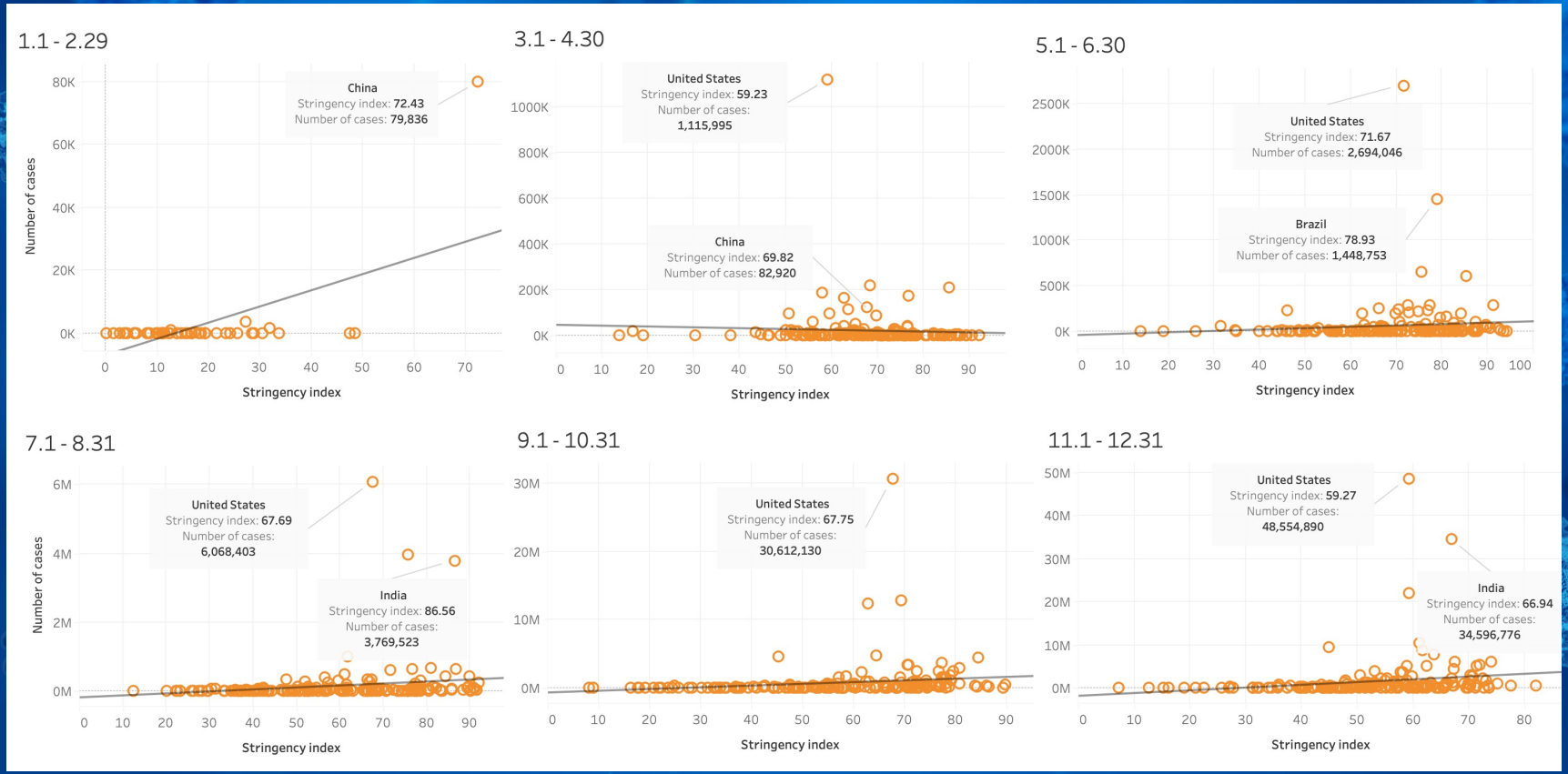


Visualization #2: Correlations of four factors with the mortality rate in the whole world (hospital beds, poverty rate, GDP, and homan development index).

Overall, hospitalbeds, GDP, and human development index have negative correlations with the mortality rate. And poverty rate has a positive correlation with the mortality rate. Yemen stands out in every correlation.

Visualization #3: Correlation between stringency index and number of cases in the world over time in 2020

The higher the number of cases, the higher the stringency index. In Janurary and Febauray, the correlation is positive and strong. However, in March and April, there is a weak negative correlation.

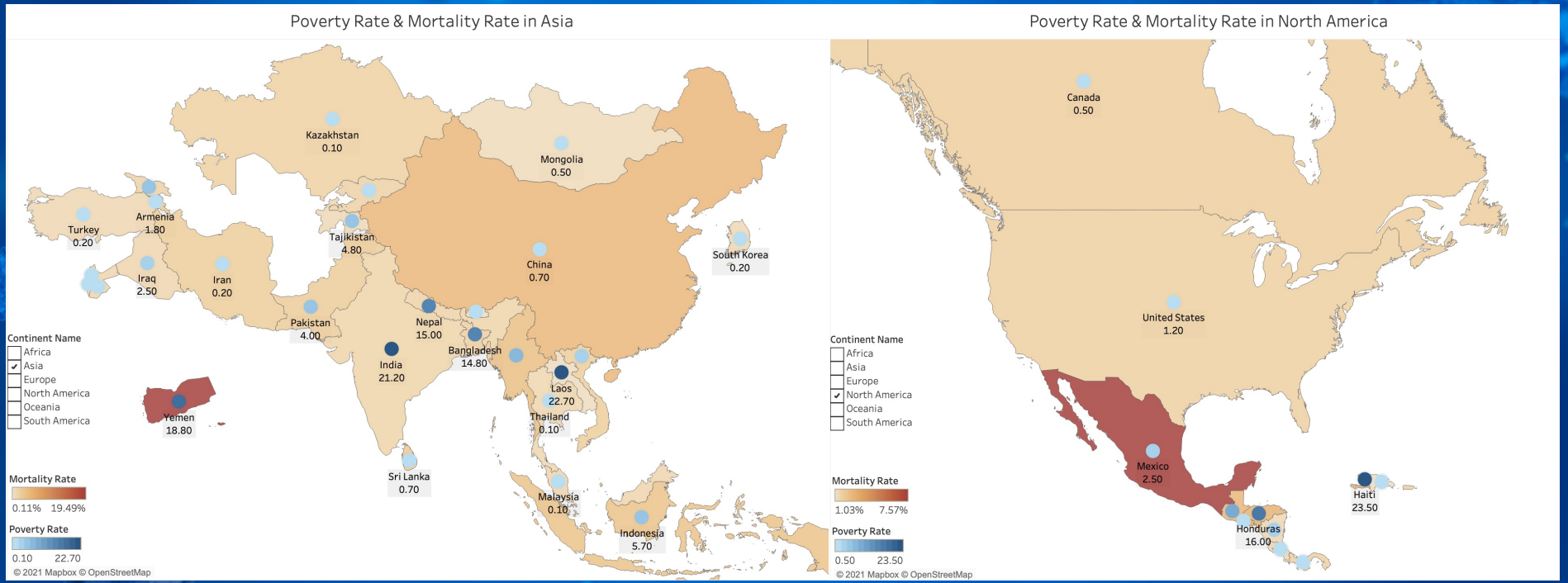


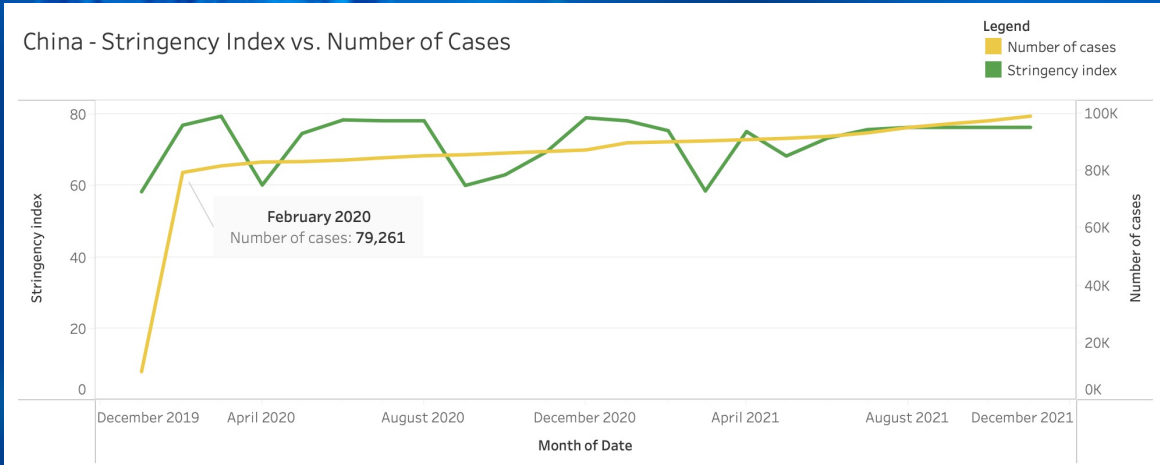
2 From a regional perspective...

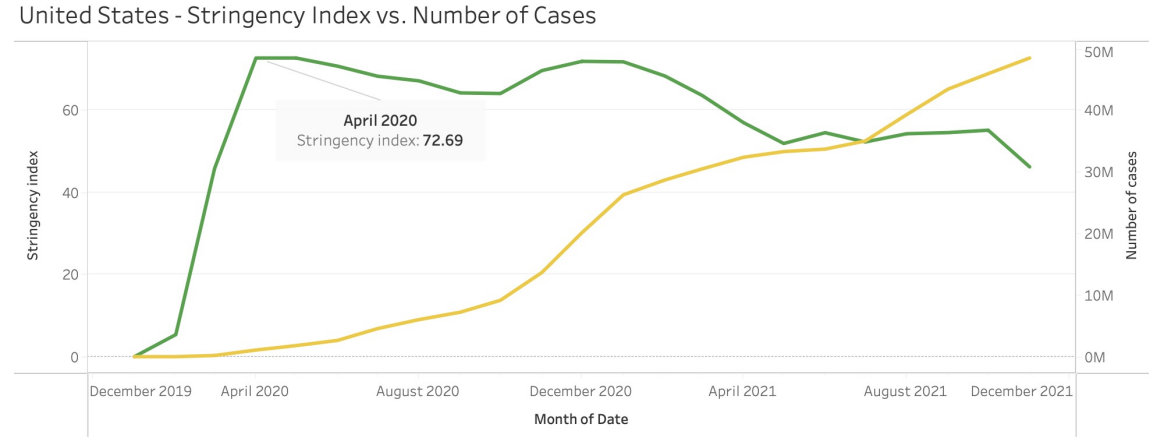
Asia vs. North America and China vs. United States, what are the differences?

Visualization #4: Two maps showing the comparison of poverty vs. mortality rate between Asia and North America

In Asia, a higher poverty rate leads to a higher mortality rate, such as Yemen. In North America, Although Mexico has a low poverty rate comparing to most countries in Asia, its mortality rate is the highest.



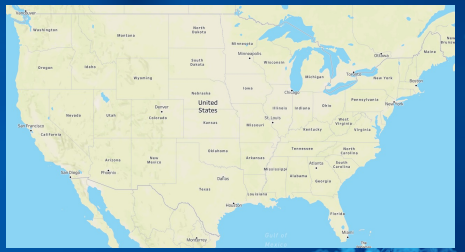




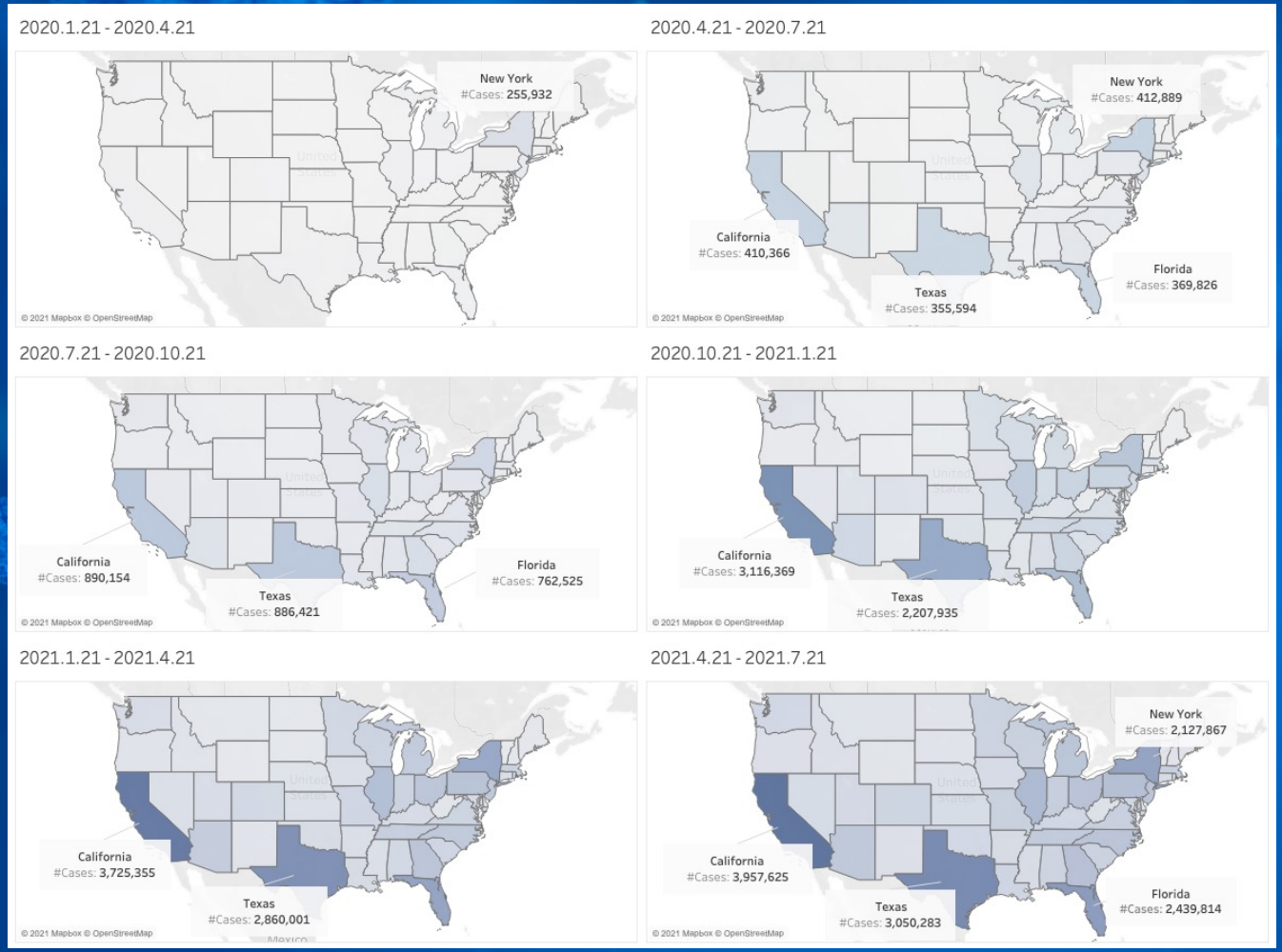
Visualization #5: China vs. United States the average stringenxy index per month and sum of number of cases per month.

China's stringency index remains steady overall with little fluctuation, with the number of cases reached a break point in February, 2020. America's average stringency index was the highest in April, 2020, with an abrupt increase from February. And the number of cases are constantly increasing at a rapid speed.

3 Take a closer look at the United States

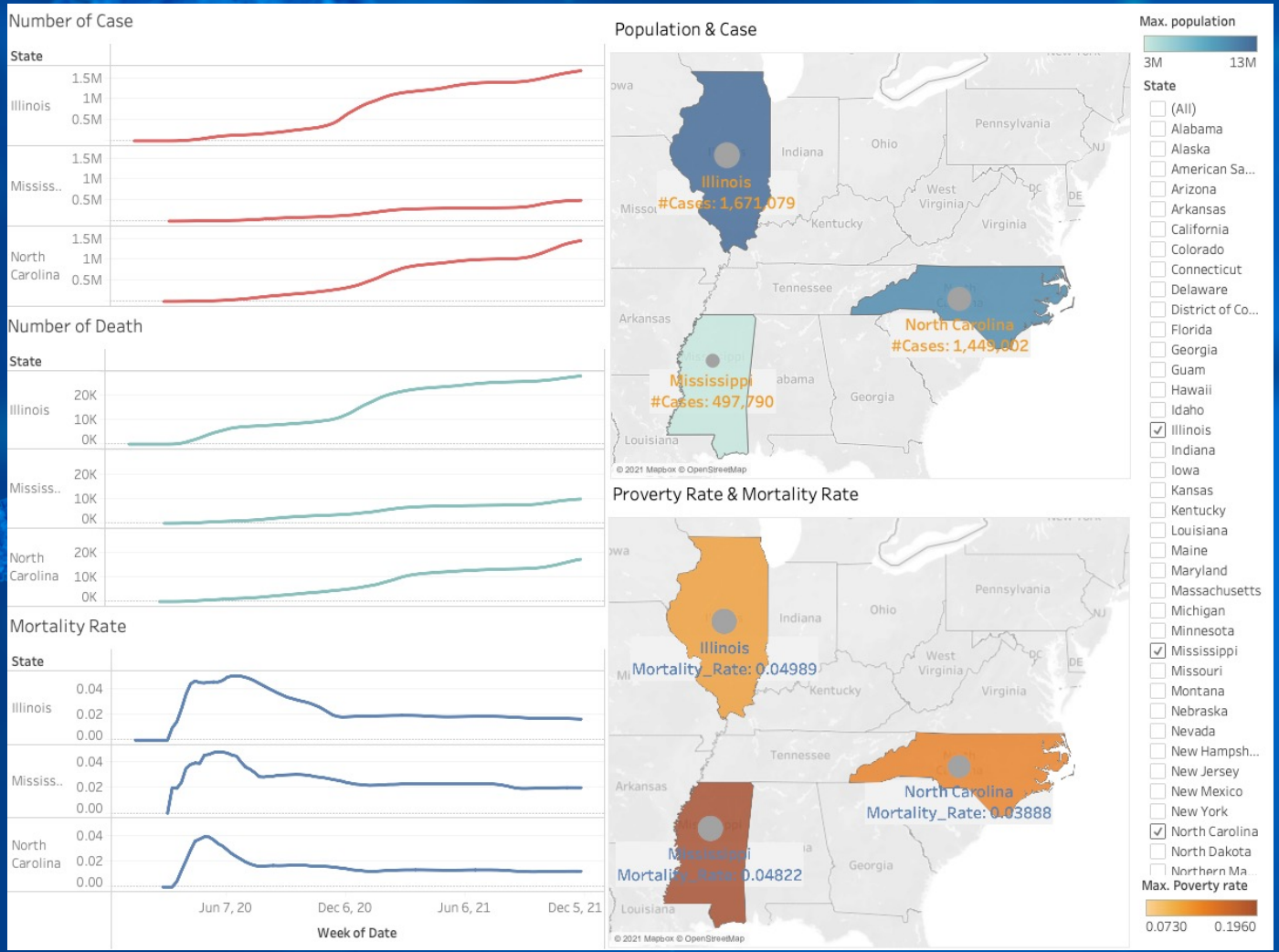


How is COVID-19 statistics distributed across the United States?



Visualization #6: Number of Cases across the United States over time

In the first three months, New York was the state affected the most by the pandemic. Gradually, coastal states, such as California, Texas, and Florida became more and more serious. Starting from October 2020, California has become the most affected state in the United States.



Visualization #7: A tool for visualizing COVID-19 situations across the states

This is a visualization tool for governments and publics to visualize and compare number of cases and deaths, population, mortality rate and poverty rate in each state.

Users can select different states to see not only the basic statistics, but also different correlations in a map setting .

Lessons Learned & Improvements

Data Preparation

- ✓ Data accuracy and integrity is important
- ✓ Scraping data from other sources requires lots of data cleaning
- ✓ Try to incorporate more data types for better visualization

Database Consideration & Design

- ✓ Careful considerations are required before designing the EER diagram
- ✓ Loading tables from csv files into GCP requires trouble shooting to ensure consistent data types and labeling
- ✓ Learn how to load large datasets into MongoDB

Data Visualization

- ✓ Constructing useful dashboards as a tool for monitoring and analysis
- ✓ Clear business case is needed for reports and dashboards

References

https://github.com/owid/covid-19-data/tree/master/public/data

https://dj2taa9i652rf.cloudfront.net

https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_poverty_rate

https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_population