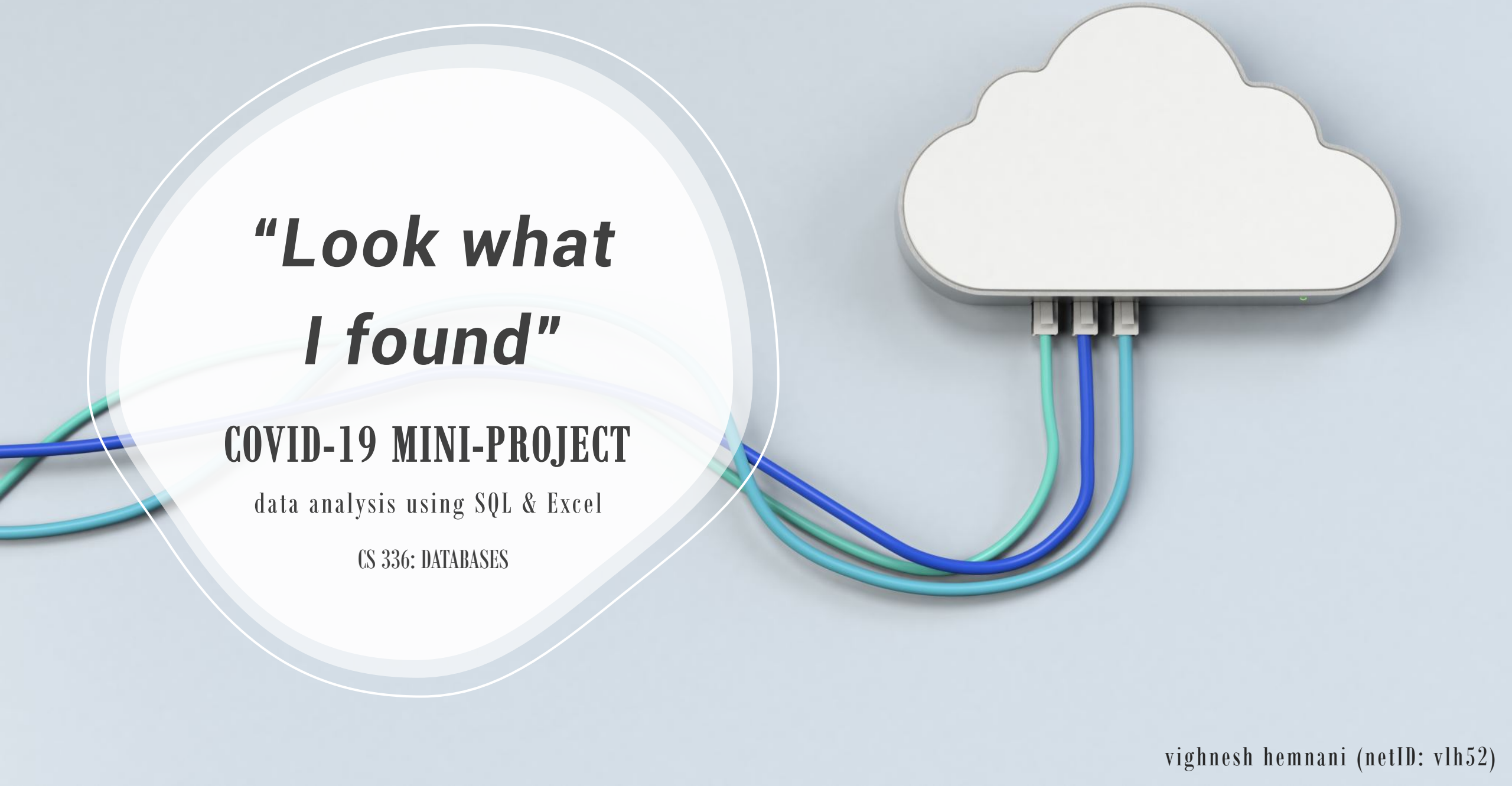


“Look what I found”

COVID-19 MINI-PROJECT

data analysis using SQL & Excel

CS 336: DATABASES

vighnesh hemnani (netID: vlh52)

introduction

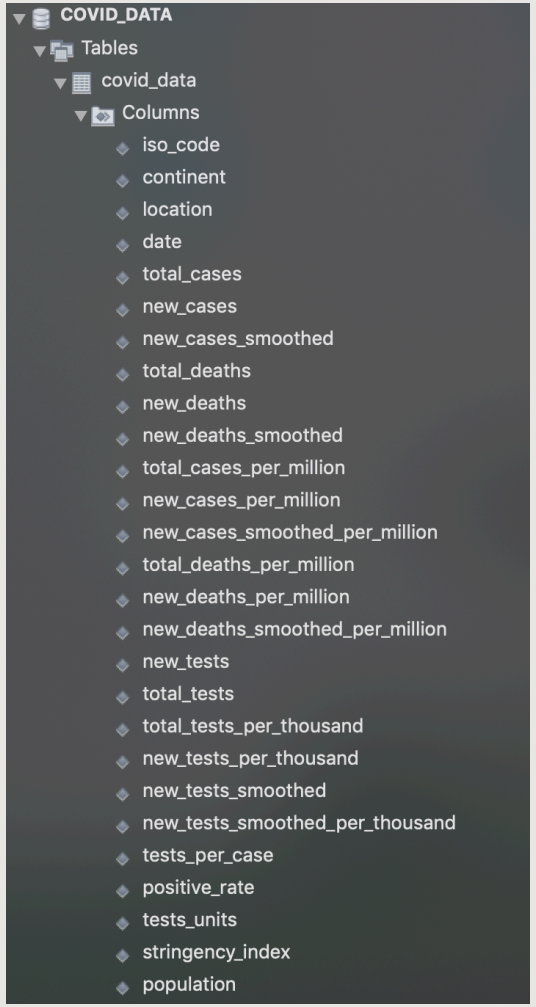
- To begin my investigation of COVID-19 statistics, I wanted to first introduce the two external variables (E) that I investigated:

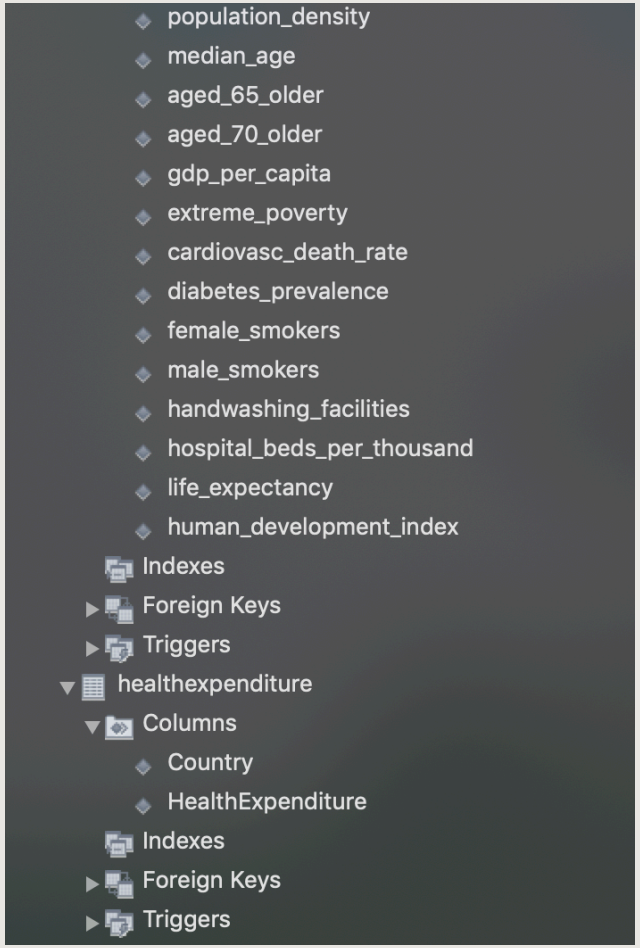
    1. E_1: Health Expenditure (% of GDP per country updated as of 2020-09-16)
    2. E_2: Change in Hotel Stays (Index)

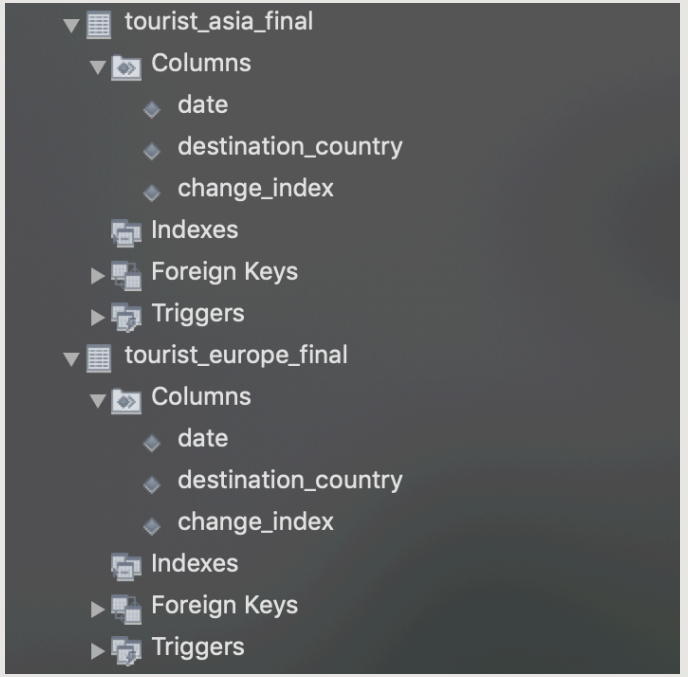
- For each external variable, I hypothesized a prediction in what the trends would look like, wrote several SQL queries that helped make different types of graph (boxplots, aggregated time graphs, histograms, pie charts…) to help me analyze the trends. Afterwards, I analyzed the trends and compared it to my hypothesis.

✪ Note: to make easier to understand how my SQL queries work, I have shown my Database Schema with all the tables and its respective columns

database (schema) structure







exploring E_1

E_1: health expenditure (% of gdp per country)

plan + hypothesis of E_1 trends

- Despite the way COVID-19 attacks human is the same, countries react differently to the pandemic. So I decided to look into the Health Expenditure governments spend as a % of their GDP via a box-plot and a pie-chart that looks at the health expenditure spread across all countries and a bar chart that compares the average health expenditure for each continent and the respective deaths and hospital beds/thousands.
- It would be logical to assume that countries with ***higher health expenditure would have lower COVID-19 related deaths and higher hospital beds/thousands*** as they would have all necessary equipment and funding too. Also, because of the wide disparity between GDPs and sizes of economies across the world, I am expecting a ***skewed box-plot and pie chart***.
- To do this, I exported the outputs of my SQL queries that explored the given COVID Database as well as an external attribute: health expenditure, and then proceeded to graph my results.

✪ Note: The reason I used deaths as a marker for COVID-19 is because cases have an ambiguity and can be over-reported/inflated for more government funding; however, deaths are official and must be reported and is a reliable marker.

exploring E_1

0. query + boxplot

- This query outputs the Health Expenditure for the respective country, to help me plot a boxplot showing the min, Q_1, median, Q_3, and max, using basic Excel functions and its graphing tools.

```sql
SELECT
    covid_data.location as Country,
    avg(healthexpenditure.HealthExpenditure) as 'Health Expenditure'
FROM covid_data
INNER JOIN healthexpenditure
ON covid_data.location = healthexpenditure.Country
GROUP BY covid_data.location
```

figure 1: SQL query for boxplot

Minimum	1.77
Lower Quartile (Q_1)	4.63
Median (Q_2)	6.45
Upper Quartile (Q_3)	8.27
Maximum	17.06

figure 2: important data for boxplot

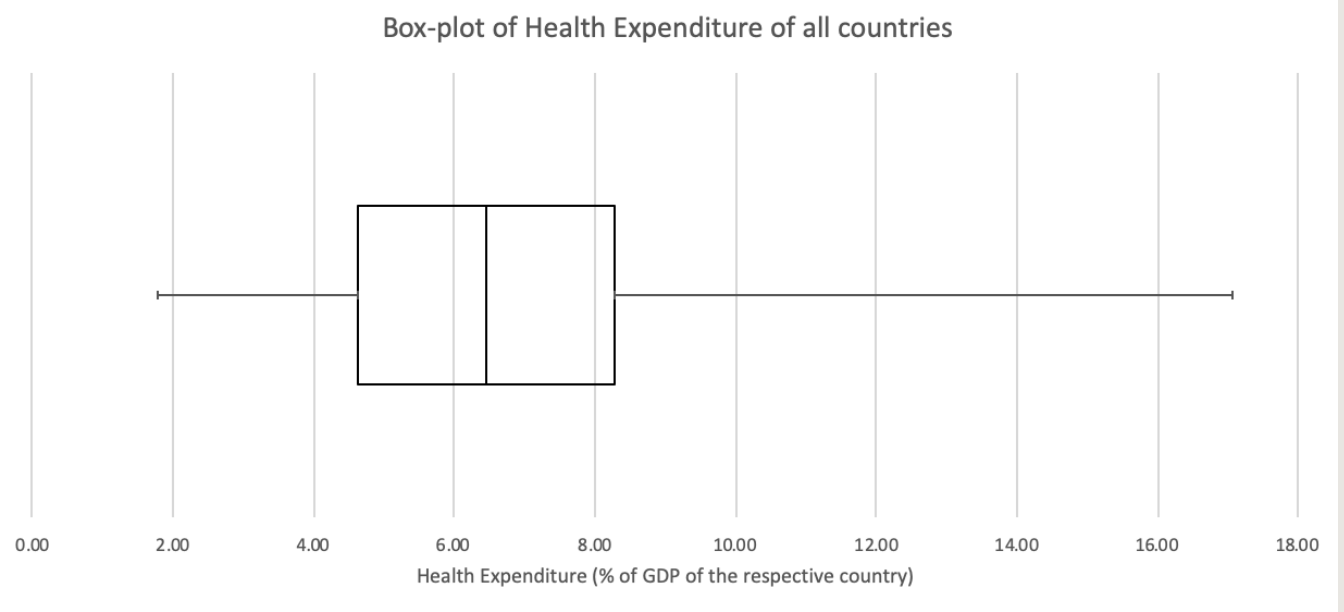


figure 3: boxplot of health expenditure (% of GDP)

exploring E_1

1. query + pie chart

- Using the same query as the last slide, I will use the output to plot a pie-chart that looks at the % of countries that have low, around average and high Health Expenditure, where the mean was 6.65%.

```sql
SELECT
    covid_data.location as Country,
    avg(healthexpenditure.HealthExpenditure) as 'Health Expenditure'
FROM covid_data
INNER JOIN healthexpenditure
ON covid_data.location = healthexpenditure.Country
GROUP BY covid_data.location
```

figure 4: SQL query for pie chart

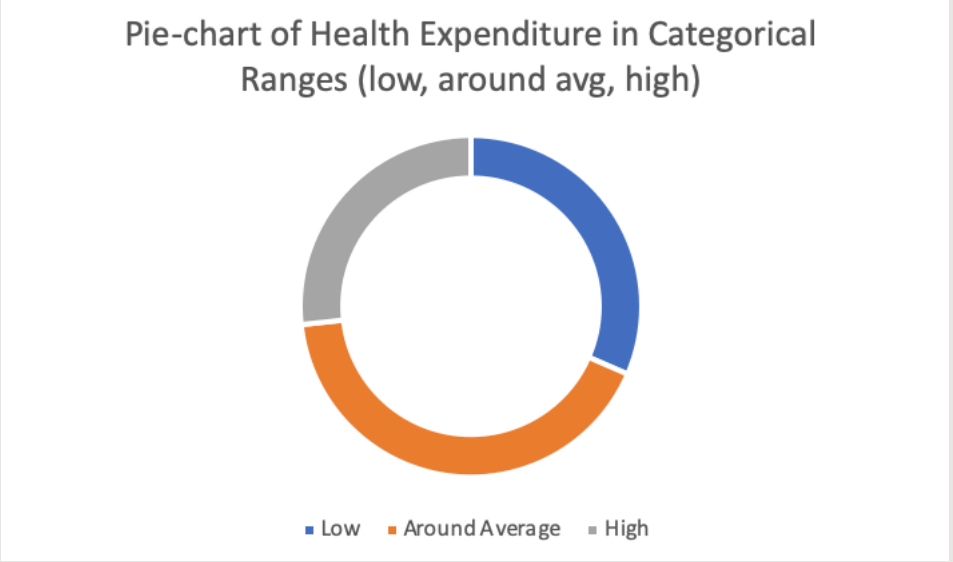


figure 5: pie chart of health expenditure in categorical ranges

2. query + bar chart

- The output of this new query is the total deaths (as a sum of all new deaths till date), average Health Exp. and average hospital beds (per thousands) for the respective continent. Using this output, we will plot a histogram that shows all this information with the respective continents.

```
SELECT
    covid_data.continent as Continent,
    CEILING(SUM(covid_data.new_deaths)) as Deaths,
    avg(healthexpenditure.HealthExpenditure) as 'Health Expenditure',
    avg(covid_data.hospital_beds_per_thousand) as 'Hospital Beds'
FROM covid_data
INNER JOIN healthexpenditure
ON covid_data.location = healthexpenditure.Country
GROUP BY covid_data.continent
Order by Deaths desc
```

figure 6: SQL query for bar chart

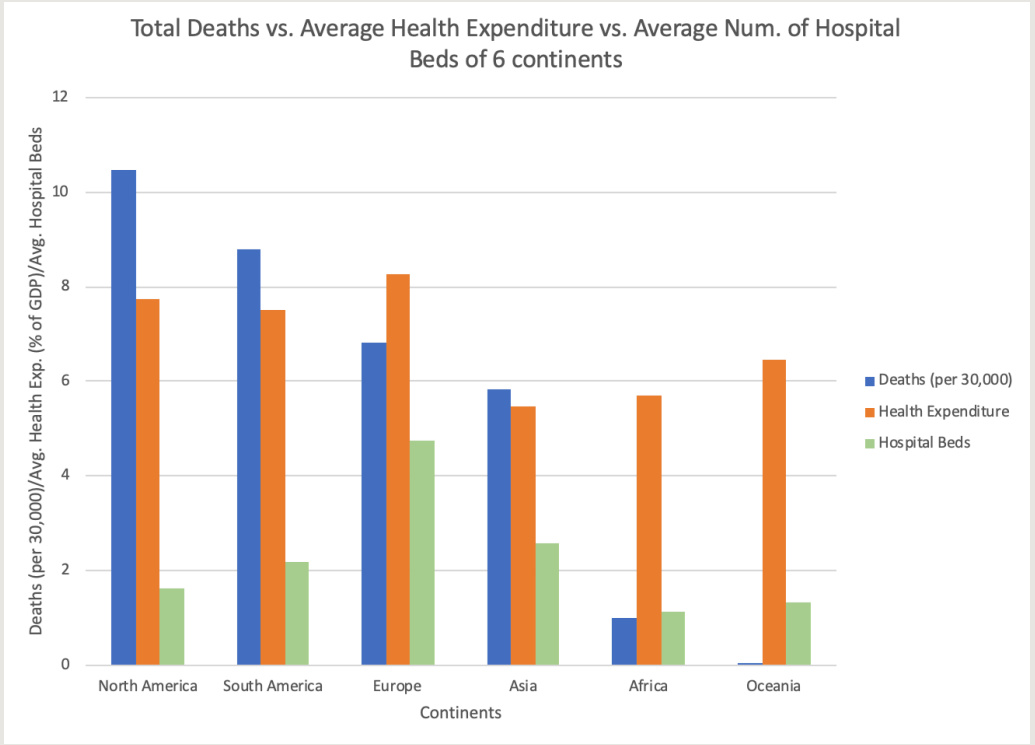


figure 7: bar chart of 6 continents with average health exp., average hospital beds and total deaths

discussion of E_1 trends

- As predicted, the box plot of the health expenditure and the pie-chart was indeed skewed, however the bar chart was very interesting. Although the high total deaths correlated with a high number of hospital beds like predicted, the health expenditure displayed an ***opposite result: the higher the health expenditure, the higher total deaths.***
- As seen, an inverse relation may exist, but another way to analyse this is that perhaps health expenditure was not a factor in total COVID-related deaths and didn't affect it at all. Curiouser and curiouser...
- So what does affect or even cause a higher total deaths/COVID-19?

exploring E_2

E_2: change in hotel stays (index)

plan + hypothesis of E_2 trends

- Since the health expenditure showed that there was no correlation or reason for the data, I decided to look into ways that could increase interaction among people and therefore COVID-19 deaths: population movement, specifically, travel and hotel lodging. In relation to this another important variable, government stringency index: an index developed by Oxford that uses 9 metrics such as school, workplace, transport closures, stay-at-home requirements and domestic/international travel control.
- After some browsing online, there was some world-wide data of the changes in hotel stays (as an index). To look into how correlated deaths, stringency index and change in hotel stays index are, I made a set of 3 aggregated time graphs that look at how two countries (Portugal and Japan) in Europe and Asia responded as short/mini case-studies.
- Based on how COVID-19 is spread and public interaction is necessary, so therefore, I would only assume that a ***higher stringency index would cause a decrease in total deaths and hotel stays index***.

3. query + aggregated time/line graph

- The output of this query is the total deaths (as a sum of all new deaths till date), stringency index and the change in hotel stays index for the respective dates in Portugal. Using this output, we will plot a set of 3 aggregated time graph that shows each variable against another over time.

```
SELECT
    covid_data.date as Date,
    covid_data.location as Country,
    covid_data.total_deaths as Deaths,
    covid_data.stringency_index as 'Strigency Index',
    tourist_europe_final.change_index as 'Change in Hotel Stays Index'
FROM covid_data
INNER JOIN tourist_europe_final
ON covid_data.date = tourist_europe_final.date
WHERE covid_data.location = 'Portugal' AND tourist_europe_final.destination_country = 'Portugal'
```

figure 8: SQL query for Portugal aggregated time graph

3. query + aggregated time line graph

- ✪ Here are the set of 3 graphs that show each variable against each other:

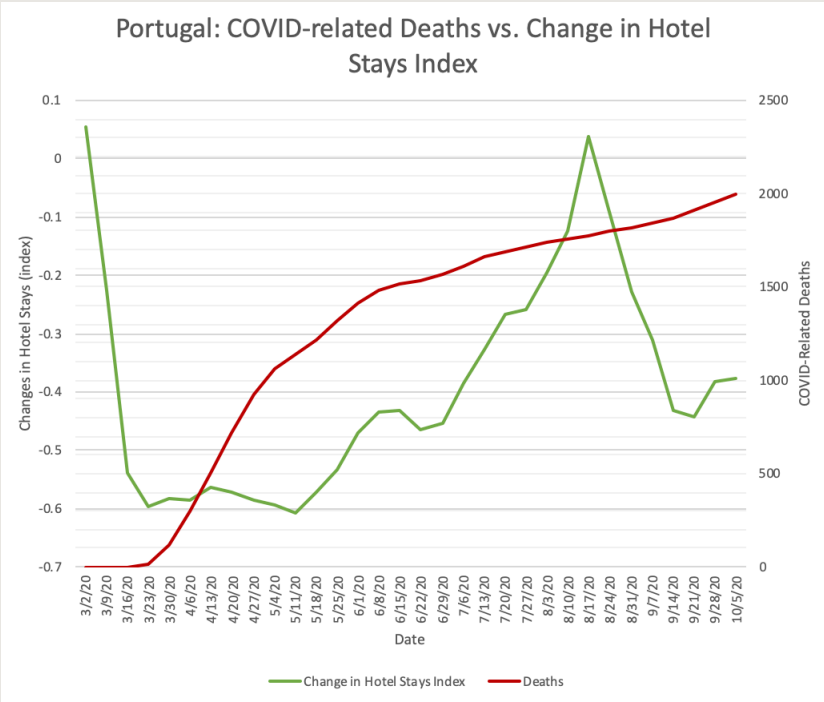


figure 9: Portugal aggregated time graph of total deaths vs. change in hotel stays index

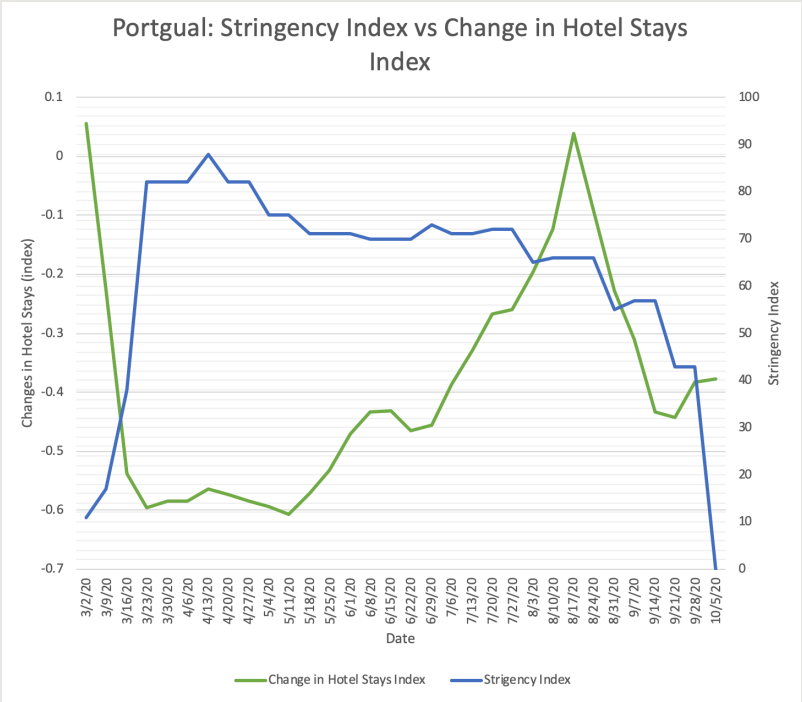


figure 10: Portugal aggregated time graph of stringency index vs. change in hotel stays index

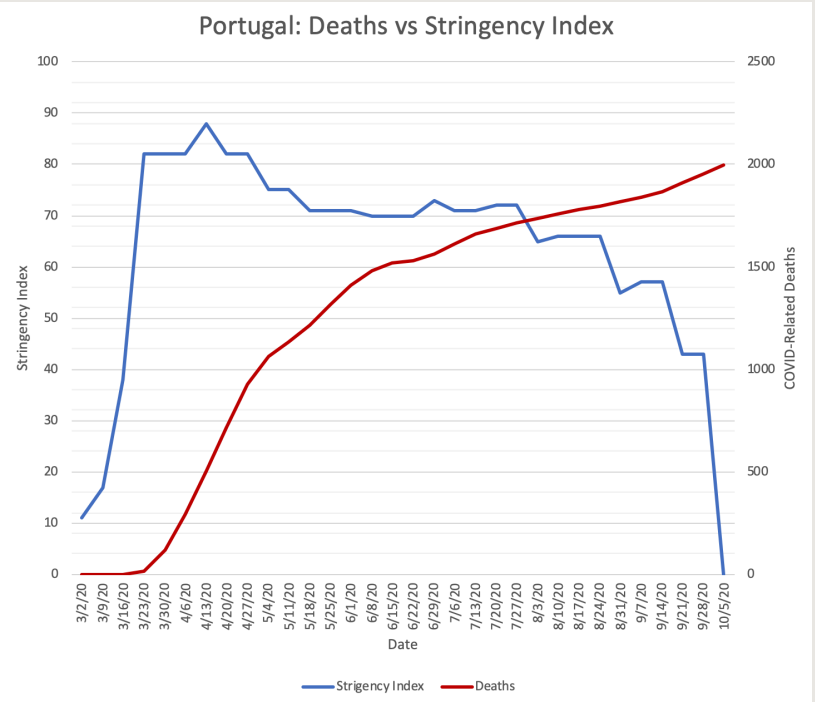


figure 11: Portugal aggregated time graph of stringency index vs. total deaths

exploring E_2

4. query + aggregated time/line graph

- The output of this query is the total deaths (as a sum of all new deaths till date), stringency index and the change in hotel stays index for the respective dates in Japan. Using this output, we will plot a set of 3 aggregated time graph that shows each variable against another over time.

```
SELECT
    covid_data.date as Date,
    covid_data.location as Country,
    covid_data.total_deaths as Deaths,
    covid_data.stringency_index as 'Strigency Index',
    tourist_asia_final.change_index as 'Change in Hotel Stays Index'
FROM covid_data
INNER JOIN tourist_asia_final
ON covid_data.date = tourist_asia_final.date
WHERE covid_data.location = 'Japan' AND tourist_asia_final.destination_country = 'Japan'
```

figure 8: SQL query for Japan aggregated time graph

4. query + aggregated time line graph

- Here are the set of 3 graphs that show each variable against each other:

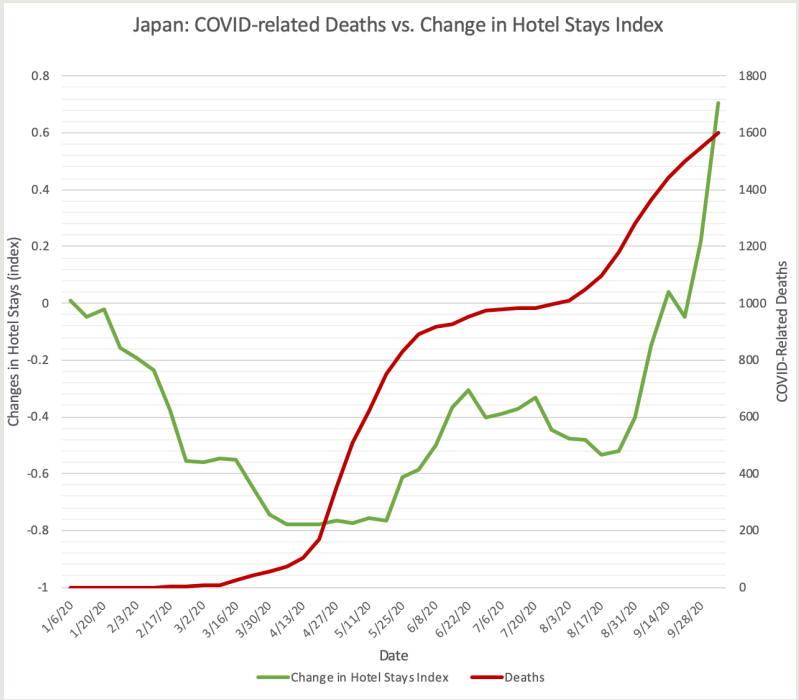


figure 13: Japan aggregated time graph of total deaths vs. change in hotel stays index

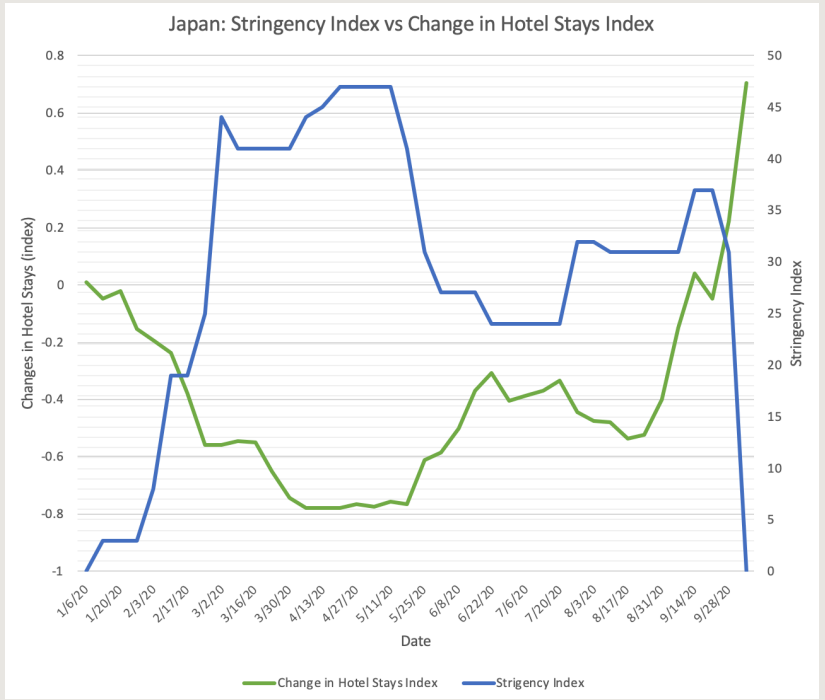


figure 14: Japan aggregated time graph of stringency index vs. change in hotel stays index

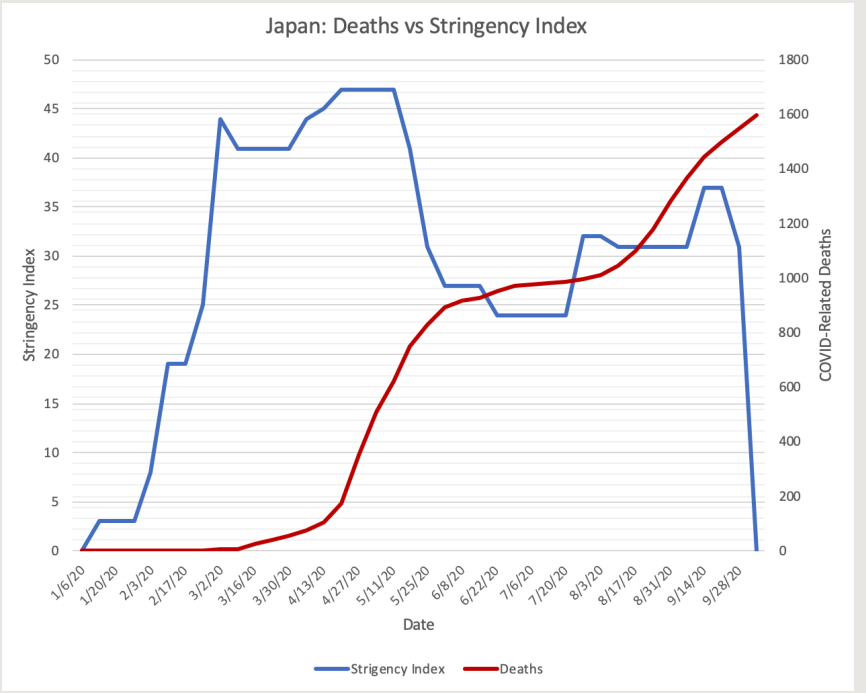


figure 15: Japan aggregated time graph of stringency index vs. total deaths

exploring E_2

discussion of E_2 trends

- Very interestingly, we see almost a *direct* correlation which is pretty *amazing* – for a ***high stringency index or a low change in hotel stays, it showed a lower total deaths for that time period***.
- For both countries, Portugal and Japan, it was seen that the total deaths either decreased or at least slowed down whenever there was a high stringency index or a low change in hotel stays index.

references

- https://datastudio.google.com/u/0/reporting/1V-6CANQZRbqjd_P3jJeBdVgYkgOY2Uup/page/ODqSB
- https://ourworldindata.org/coronavirus-data-explorer?zoomToSelection=true&time=earliest..latest&country=~PRT&region=World&deathsMetric=true&interval=smoothed&perCapita=true&smoothing=7&pickerMetric=total_deaths&pickerSort=desc

Thank you!

COVID-19 MINI-PROJECT

vighnesh hemnani (netID: vlh52)