Data and Policy Analysis Executive Summary

Professor Seferlis
Roamah Baray, Matt Frysinger, Sanjeev Pal, Ze Song

DS310: Data Mechanics
Tuesday, May 7, 2024

We have been hired by the country of Caladan to determine what policies—implemented by ten other countries—have been most effective in maintaining a low growth rate of Covid-19 cases and deaths and therefore should be implemented by Caladan. We were able to determine which policies would be the most effective in keeping the growth rate below 3% for the number of new cases and 1% for the number of new deaths. From there we determined which of the policies would not only be the most effective but the least restrictive on the general public.

The first step was gathering data from multiple sources within our Azure environment. The data sources included policy data from CosmosDB, health metrics data from a virtual machine, and country related data from an Azure SQL DB. We used a parquet viewer to examine and understand the initial structure of the data, allowing us to identify the field and data types present in each dataset. After understanding the initial data structure, we outlined the key fields and their data types in order to begin the cleaning process. We conducted basic data cleaning and maintained a consistent format across the datasets. To support effective analysis, we created a galaxy schema. This schema consisted of fact tables for key metrics (created as a separate external table in Azure) and policies, complemented by dimension tables pertaining to the question at hand. To integrate the data, we had to transform and join the datasets based on common keys. Usage of the galaxy schema provided our team additional flexibility in PowerBI.

First, in PowerBI, our team created two new columns for both cases and deaths: a rolling average column and a growth rate column using the rolling average column (see Figures 1 and 2 in the Appendix). The growth rate was used as a key indicator to determine which policies would be the most effective and worthy of implementation. Something our team noticed with the data was inconsistencies among the measurability of the policies; some policies, like “c1_school_closing” are measured on a 0–3 scale, whereas “c3_cancel_public_events” is measured on a 0–2 scale. To gauge a better understanding of “restrictiveness,” each policy column was normalized on a scale of 0–1, with 1 being the most strict. Now, our team has developed a new metric from which to gauge the restrictiveness of combined policies over a period of time in a more accurate manner.

Within PowerBI, the normalized policy sum was plotted over time with lines signifying the growth rate of both new cases and new deaths (Figure 3). The following period, from January 7, 2021 to February 14, 2021, demonstrated low policy

restrictiveness as well as being within the parameters of under 3% and 1% for the growth rate of new cases and deaths, respectively. Additionally, only the countries of Sweden and New Zealand were included in our analysis due to them being the countries with the most similar-sized populations to Caladan. Over this entire period, the normalized policy sum equated to 7.17 — relatively unrestricted when compared to the rest of the available time frame. See Figure 4 for how we arrived at the following final policy conclusions:

- Restrictions on International Travel (policy level = 7)
- Restrictions on Gatherings (4)
- Mask Requirements (3)
- Stay Home From Work (2)
- Stay Home From School (2)
- Restrictions on Public Events (2)
- Restrictions on Public Transportation (1)
- Restrictions on International Movement (1)
- Stay at Home (1)

Our recommendation is to implement those policies listed as they proved to be the least restrictive and most effective policies that meet our thresholds in limiting the growth rate of new cases and deaths.

Appendix

```
Rolling Avg Cases =
CALCULATE(AVERAGE(casesfinal[Confirmed_Change]), FILTER(ALLEXCEPT(casesfinal, casesfinal[Country_Region]),casesfinal[Updated] <= EARLIER(casesfinal[Updated]) &&
casesfinal[Updated] > EARLIER(casesfinal[Updated])-30))
```

Figure 1 *– Rolling average of cases. A nearly identical column was created for deaths.*

```
Growth Rate =
VAR CurrentDate = 'casesfinal'[Updated]
VAR InitialDate = CurrentDate - 1
VAR FilteredTable =
    FILTER(
        'casesfinal',
        'casesfinal'[Updated] >= InitialDate &&
        'casesfinal'[Updated] <= CurrentDate &&
        'casesfinal'[Country_Region] = EARLIER('casesfinal'[Country_Region])
    )
VAR ConfirmedChangeCurrentDate =
    CALCULATE(
        SUM('casesfinal'[Rolling Avg Cases]),
        FilteredTable,
        'casesfinal'[Updated] = CurrentDate
    )
VAR ConfirmedChangeInitialDate =
    CALCULATE(
        SUM('casesfinal'[Rolling Avg Cases]),
        FilteredTable,
        'casesfinal'[Updated] = InitialDate
    )
RETURN
    IF(
        ConfirmedChangeInitialDate < 0,
        BLANK(),
    IF(
        ConfirmedChangeInitialDate = 0,
        0,
        ((ConfirmedChangeCurrentDate - ConfirmedChangeInitialDate) / ConfirmedChangeInitialDate
    )
))
```

Figure 2 *– Growth rate for cases column. A nearly identical column was created for deaths.*

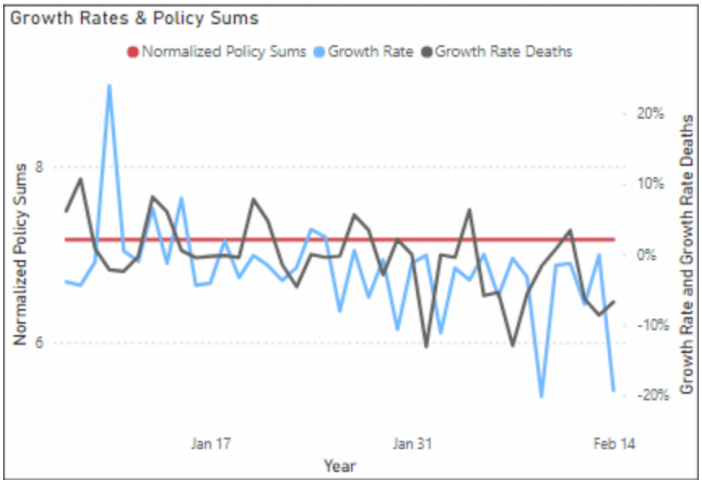

Figure 3 *– Growth rate of cases, deaths as well as a line marking the normalized policy sum (y=7.17)*

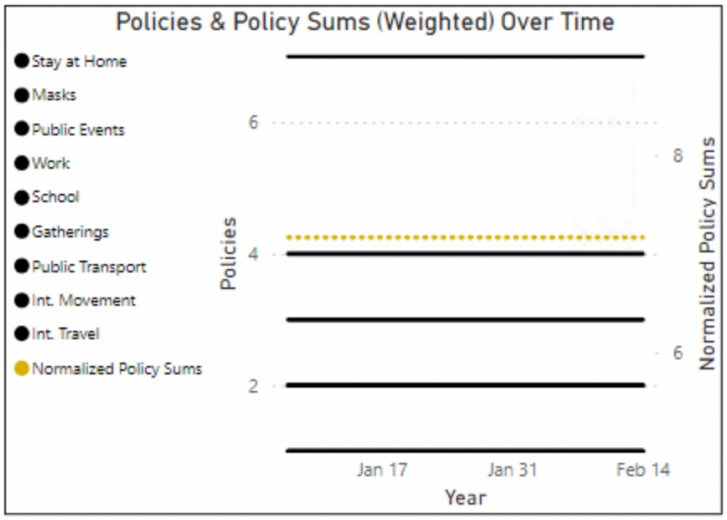

Figure 4 – Policies in place, their levels and weighted policy sum over desired timeframe

---

Challenge Responsibilities

Matt Frysinger

- Challenge 1 – creating ADLS Gen 1, configuring the RDP port, creating the on-premise SHIR, VM configuration, grabbing the data from CosmosDB and Azure SQL DB and converting to .parquet format
- Challenge 2 – creating ODS, creating “Cases” and “Policies” datasets using Azure Data Factory
- Challenge 3 – creating galaxy schema, creating external tables for “Cases” and “Policies”
- Challenge 4 – loading data into PowerBI, setting up relationships, creating additional columns using DAX, visualizations, presentation of the BIreport
- Executive Summary – editing and figures
- Presentation – “Analytical Justification” slide

Roamah Baray

- Challenge 1 - assisted in grabbing the “recoveries” data
- Challenge 2 - created “Recoveries” dataset using Azure Data Factory
- Challenge 3 - assisted in the creation of galaxy schema by finding common metric
- Challenge 4 - thinking of relationships to consider and visualizations that would be helpful for the PowerBI
- Executive Summary - wrote the summary
- Presentation - “Findings” slide
- Set up the Github repository

Sanjeev Pal

- Challenge 1 - assisted in grabbing the “deaths” data
- Challenge 2 - created “deaths” dataset using Azure Data Factory
- Challenge 3 - brainstorm relationships in galaxy schema by finding common metric
- Challenge 4 - brainstorm relationships to consider and visualizations that would be helpful for the PowerBI
- Presentation - “Architectural Overview” and “Schema” slide with the assistance of draw.io

Ze Song

- Challenge 1 - assisted in configuring the SHIR, grabbing data via SQL server, helped with converting data into parquet from Azure Cosmos DB and SQL database.
- Challenge 2 - creating ‘Geography’ datasets using Azure Data Factory, join two tables for the ‘Date’ dataset through Azure Data Factory
- Challenge 3 - assisted in the design of the galaxy schema, identifying the appropriate primary keys and common metrics
- Challenge 4 - assisted in brainstorming the approaches for visualizing results in PowerBI
- Presentation - Creating the slides and the introduction for the project