# Policy Guidelines for Caladan to Prevent Spread of COVID-19.

Aryan Jain, Jaden Cho, Samyuktaa Jayakrishnan, Naveah Zhan

*The Commonwealth of Caladan hereby instills the following 4 policies*

1. **Closing Schools**: All schools K-12 will be shut down until further notice
2. **Stay at Home Requirements**: All citizens of Caladan must stay at home with exceptions for daily exercise, grocery shopping, and 'essential' trips.
3. **International Travel Controls**: The Commonwealth of Caladan will ban international travel arrivals from some regions.
4. **Workplace Closing:** The Commonwealth of Caladan will require closing (or work from home) for some sectors or categories of workers.

Our team of 4 data scientists came together to analyze data and support this plan.

In the first challenge, the main objective was to collect the raw data necessary for analysis. We extracted our Policy data from Cosmos DB, and other metrics data and dates from Azure Data Factory Source and Azure SQL Database. The tasks required setting up Azure Data Factory, configuring connections using the Self Hosted Integration Runtime for SQL Server data extraction, and ensuring the data was clean and organized for future use. The output format for all data was specified to be in Parquet format to ensure compatibility and performance within the Azure Data Lake.

In the second challenge, we focused on transforming the raw data stored in the Azure Data Lake into a more usable format by creating an Operational Data Store. For example, we created a unified dataset that included metric and policy data.

In the third challenge, the transformed data from ODS was loaded into Azure Synapse Studio for further processing and analysis. We first created a data warehouse using Azure Synapse to facilitate the efficient querying and analysis of large datasets. We loaded those external tables into PowerBI. In PowerBI, we first implemented a snowflake schema to organize the data into fact and dimensional tables.

We implemented a Snowflake Schema for this project, which is essential for our analysis involving various datasets related to COVID-19 policy effectiveness.

The central fact table contains data about different policies implemented to reduce the growth and deaths rate. This includes information on school closings, workplace closings, cancellations of public events, restrictions on gatherings, and other related measures.

In addition, there are 6 dimensional tables, including Geography, Data, Cases, Recoveries, and Deaths. First, Geography contains geographical data which is vital for regional analysis and impact of policies on specific countries. Second, Dates are crucial for tracking policy changes and their effects over time. Third, Cases contains confirmed cases and changes, and growth rates. This table allows for direct correlation between policy implementation and case outcomes. Fourth, Recoveries are essential for assessing the effectiveness of health policies and measures over time. Lastly, Deaths contains mortality data associated with COVID-19. This table helps analyze the severity and mortal impact of the virus under various policy conditions. We also added two external datasets: Population and GDP to the Geography table.

To get an idea of the data we first made a simple visualization for the 14 months to see cases and death growth rates. We identified a time period where the rates were under the benchmarks using horizontal measures (0.01 for deaths and 0.03 for cases). We found that the time period was May 2nd to July 8.

To figure out which policies were most effective, we created visualizations for each policy between April 15 to July 8th. We chose April 15, more than 2 weeks before May 2nd, so that we could take into account the time required for the policies to take effect. Using the visual markers at the cutoff points of .01 for deaths and .03 for cases, we analyzed the graphs and identified the top 4 policies for both growth rates.

To confirm our findings, we performed a decision tree regression for all 14 months. First, In this section we regressed the policies and growth rates for cases and deaths. We created a new column that checks when both cases and deaths growth rates are below 3% and 1% respectively and regresses over that column. The features that are most important for that regression helped us determine the most important policies. Next, we used the information we concluded from our Power BI visualization,  and regressed to see the most important policies during the time period of May 2nd to July 8th. We tried a model with our external datasets but they were increasing our MSE values and we hypothesized that it was due to a faulty dataset, so we decided not to include it in our EDA or CDA.

Once our CDA confirmed our findings, we used the visualization to identify the specific levels at which the policies seemed to be more effective. Using this information, we created our top four policies and their specific regulations for Caladan.

Work Done:

Jin Kyu (Jaden) Cho: Challenge 1,2, and 3, Making Schema, Report: Data Flows and Schema, Draw.io Data Flows, PowerBI Visualization, Uploading files in GitHub, Presentation Slides and Presentation Recording.

Aryan Jain: Challenge 1, 3, Making Schema, PowerBI Visualization, Report: Regression Analysis, Presentation Slides and Presentation Recording, Machine Learning: Decision Tree Regressor with Feature Importances

Samyuktaa Jayakrishna: Challenge 1, 3, PowerBI Visualization, Report:EDA, Presentation Slides and Presentation Recording

Neveah Zhan: Draw.io Schema, Presentation