**Strategic Policy Recommendation: Caladan's COVID-19 Response**

Team 1: Abigail Merage, Emily Forman, Aidan Clark, Patrick Soong, Naman Nagaria

**Introduction**

As Caladan prepares for a potential surge in COVID-19 cases, Team 1 has been enlisted to assist in devising an effective mitigation plan. We guided our analysis with a key question: What are the least restrictive measures that can be taken to keep the growth rate of deaths below 1% and the growth rate of new cases below 3% based on a 30-day rolling average? Our strategy was to examine the effectiveness of policies from ten representative countries in managing the growth rates of COVID-19 cases and deaths. The goal of this research was to identify which policies have been the most successful and least restrictive in curbing the spread of the virus. This report aims to explore this analysis while delving into the nuances of effectiveness, timing, and external factors that may impact outcomes.

**Data Architecture**

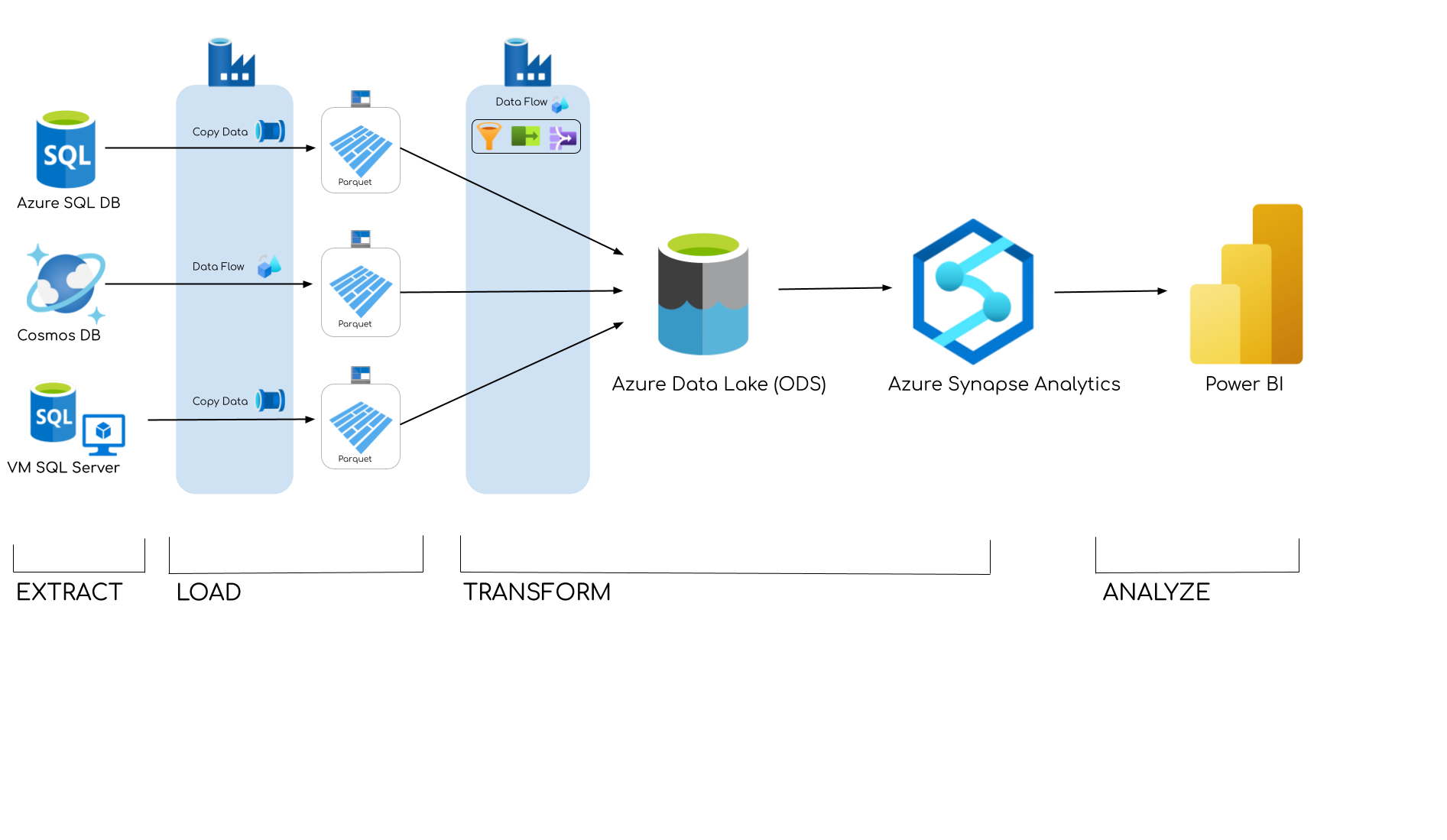


Figure 1: Data Transformation Model

To construct a comprehensive dataset for analysis within Azure Data Factory, Team 1 began by extracting data from three sources: On-Premises SQL Server, Azure CosmosDB (SQL API), and Azure SQL Database. The datasets, originally sourced from various sources including the WHO, CDC, and Public Health Departments, encompass a wide array of variables. The SQL Server and SQL Database include metrics such as confirmed cases of COVID-19, confirmed deaths, and changes in cases from ten representative countries between January 21, 2020 and March 30, 2021. The Cosmos Database provides information about policies implemented by these countries to mitigate COVID-19.

During the loading phase, we employed Azure Data Factory pipelines to standardize the three datasets into parquet format. Subsequently, we integrated data from both SQL Server and SQL Database, combining country metrics, while employing a filter pipeline to eliminate redundant country data. Following this, the data was aggregated into an Operational Data Store, comprising of six parquet files: Dates, Geographies, Deaths, Cases, Recoveries, and Policies. The structured data was then loaded into external tables within Azure Synapse to construct Fact and Dimension tables using a linked service. Team 1 chose to establish a data schema in the next phase of the project, Power BI analytics, to allow for more flexibility in data model creation.

**Data Exploration**

Following the integration of our data into Power BI, Team 1 created a model that would allow us to connect the data by common entities for best utilization when displaying our findings. We created a snowflake schema, in which there is one main fact table with a primary key and additional dimension tables. We established our Policy data as the fact table because it contained the most amount of common features and eliminated many-to-many relationships. A factor in our decision was that Policy data includes the feature “Dates,” which connects to our dimension table of dates without having to utilize countries, which then appears in the remaining tables. Additionally, we created a concatenated unique identifier of countries and dates to reduce redundancy amongst all of the remaining tables. Our external population data snowflaked from the geography dimension table and served to visualize the demographics of our countries.

**Evaluation**

In the analysis, Team 1 began by examining the growth rates of COVID-19 cases and deaths on a 30-day rolling average, focusing on time periods where growth rates of cases remained under 3% and deaths under 1% per country. By identifying timeframes where both deaths and cases fell below our desired threshold,we narrowed down the date range for further investigation. Next, we compared these dates to a heat map depicting policy implementations to gauge how various policies fared against each other and understand each country's policy

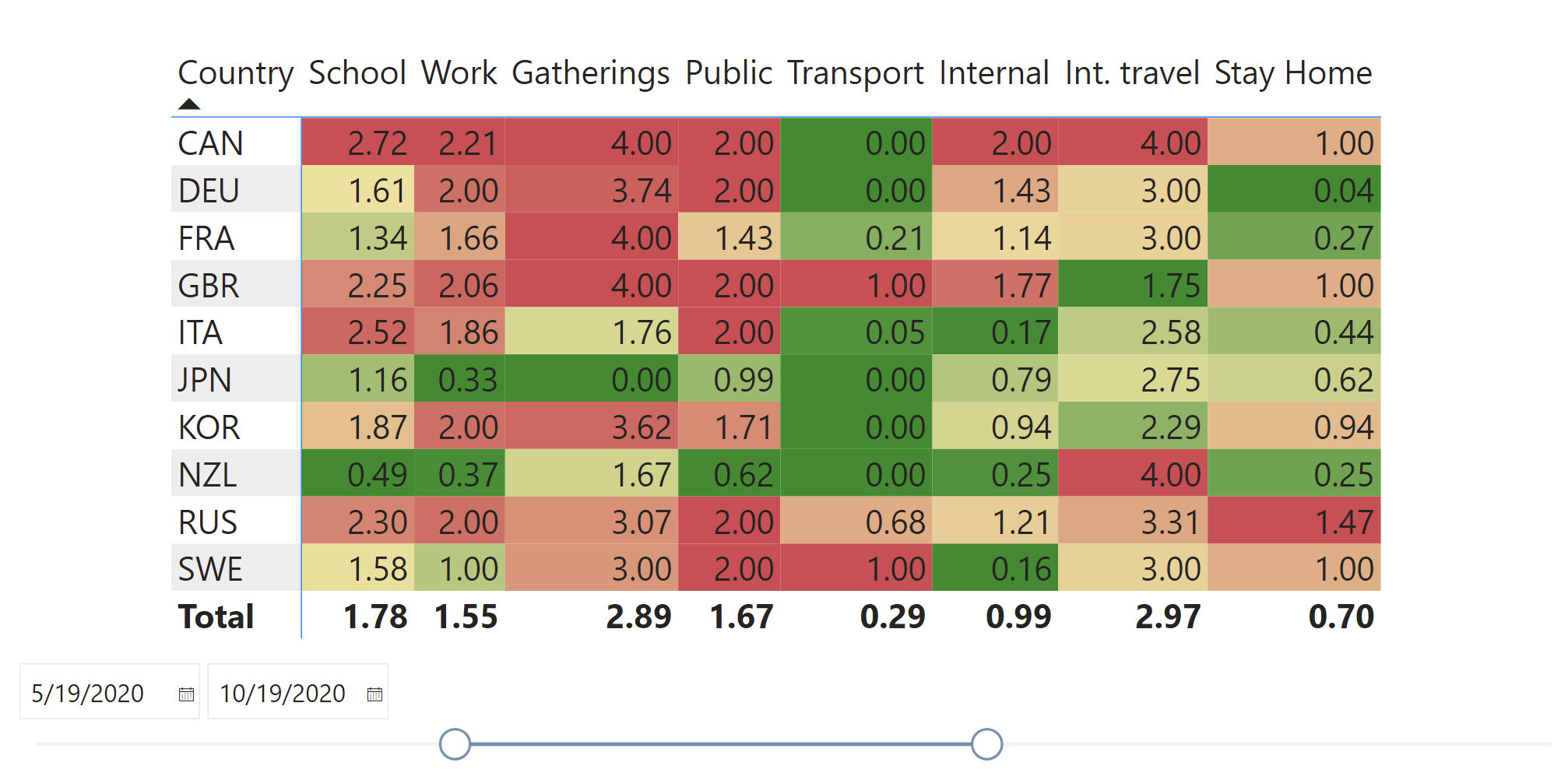
landscape simultaneously.

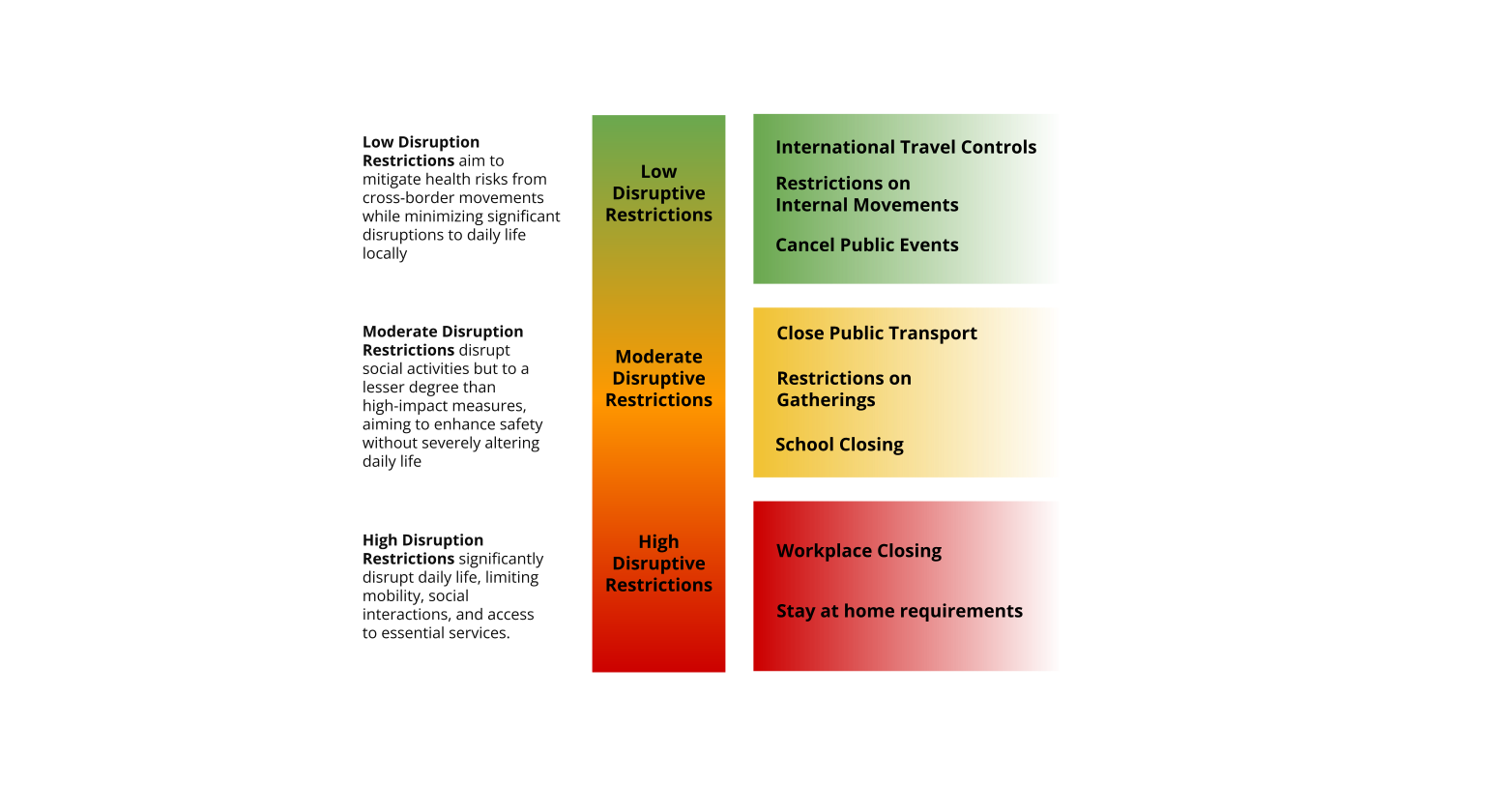
These eight policies were chosen for their preventative nature. Our team chose to focus on measures that would curb the growth of cases and deaths rather than measures that address the aftermath of COVID spread. An analysis of public gatherings over a five-month period revealed differences in policy implementation among countries such as Canada, Germany, and the UK compared to Japan, New Zealand, and Italy, providing insights into active policy enforcement.

We cross-referenced the heat map with growth rate trends to identify which policies were in effect during the studied time period, determining our focus areas accordingly.

Policies were then ranked based on their level of restrictiveness, with less disruptive policies prioritized (Figure 3). Using area charts, we compared growth rates in cases and deaths against specific policies, allowing us to discern trends in growth rates concerning policy implementation. We concluded that effective policies should have a lower restrictiveness while impacting growth rates optimally. For instance, analyzing the impact of stay-at-home policies revealed spikes in policy implementation correlating with a reduction in COVID cases, suggesting effectiveness.

However, our team aimed to find less restrictive policies that still led to a reduction in growth rates. We excluded the top two most restrictive policies from immediate consideration, recommending them as last-resort measures rather than premeditated actions (Figure 3). Initially, school closings emerged as an unfavorable choice due to its comparatively higher restrictiveness than the other policies. However, it was determined that proper preparation, such as investing in infrastructure that supports remote learning, could allow school closings to be far less restrictive while significantly reducing the growth rate of cases. Our analysis culminated in the identification of three recommended policies for implementation: internal movements, public transport regulations, and school closings.

Figure 2: The heat map illustrates the level of implementation for eight policies of interest, with a color corresponding to level of implementation.

Figure 3: Scale of policy restrictiveness based on disruption to everyday life.

**Recommendation**

Using our evaluation process, we ultimately decided on three key policies from which Caladan could benefit without being unnecessarily disruptive to the general public. From least restrictive to most, we recommend that Caladan put restrictions on internal movements, public transport, and school closings. We were able to come up with this conclusion based on several factors including our exploratory and confirmatory data analysis.

As we explored our data through cross validation with the heatmap and growth rates of cases and deaths, we found date ranges where the growth rate rolling average of deaths and cases were below the thresholds that we were interested in. Using these date ranges and the heatmap, we were able to conclude that these three policies were the least restrictive and reduced the growth rates for both cases and deaths.

To further confirm this, we delved into two methods of statistical analysis during our confirmatory analysis stage. Utilizing a key influencers graph in PowerBI and logistic regression in Python, we found significant results further supporting our initial findings. Using the key influencers graph with our cases growth rate rolling average, we found that these three policies mentioned earlier were effective in reducing the growth rate of cases by -.94% while the overall average was at 3.03. Creating a key influencer graph with the rolling average of the growth rate of deaths, we found similar results, showcasing an average growth rate of -1.26% with school closing and restrictions on internal movements against an average growth rate of 1.99%. On top of using a key influencers graph, we utilized a logistic regression in Python with the scikit-learn library. We set the thresholds for the growth of the 30-day moving average for death at 1% and for cases at 3%. We found that for the rolling average of the growth of cases, all of our three policies were within the top 5 policies that had the highest coefficient. For deaths, public transport and internal movement were the top 2 and school closing also had a significant coefficient. We excluded contact tracing because it is not a “restrictive” policy. Both of these regressions included variables that were either not included in the data description or a policy that we deemed to be too restrictive such as stay-at-home requirements.

Thus, through our exploratory and confirmatory data analysis, we highly recommend that Caladan implement policies for internal movements, public transport, and school closing before moving to more restrictive changes.

**Machine Learning**

For the machine learning aspect of the project, Team 1 decided to use XGBoost Regressors to predict the confirmed change and deaths change as a proportion to the population, using the three policies we found had the most influence. We found the best levels of these policies for each month and trained a model, excluding Sweden (to prevent leakage) and Great Britain (negative deaths outlier, reduced model accuracy). Below is a plot of our predictions of Sweden if they had the levels of policy we found were the best for each month (Figure 4).

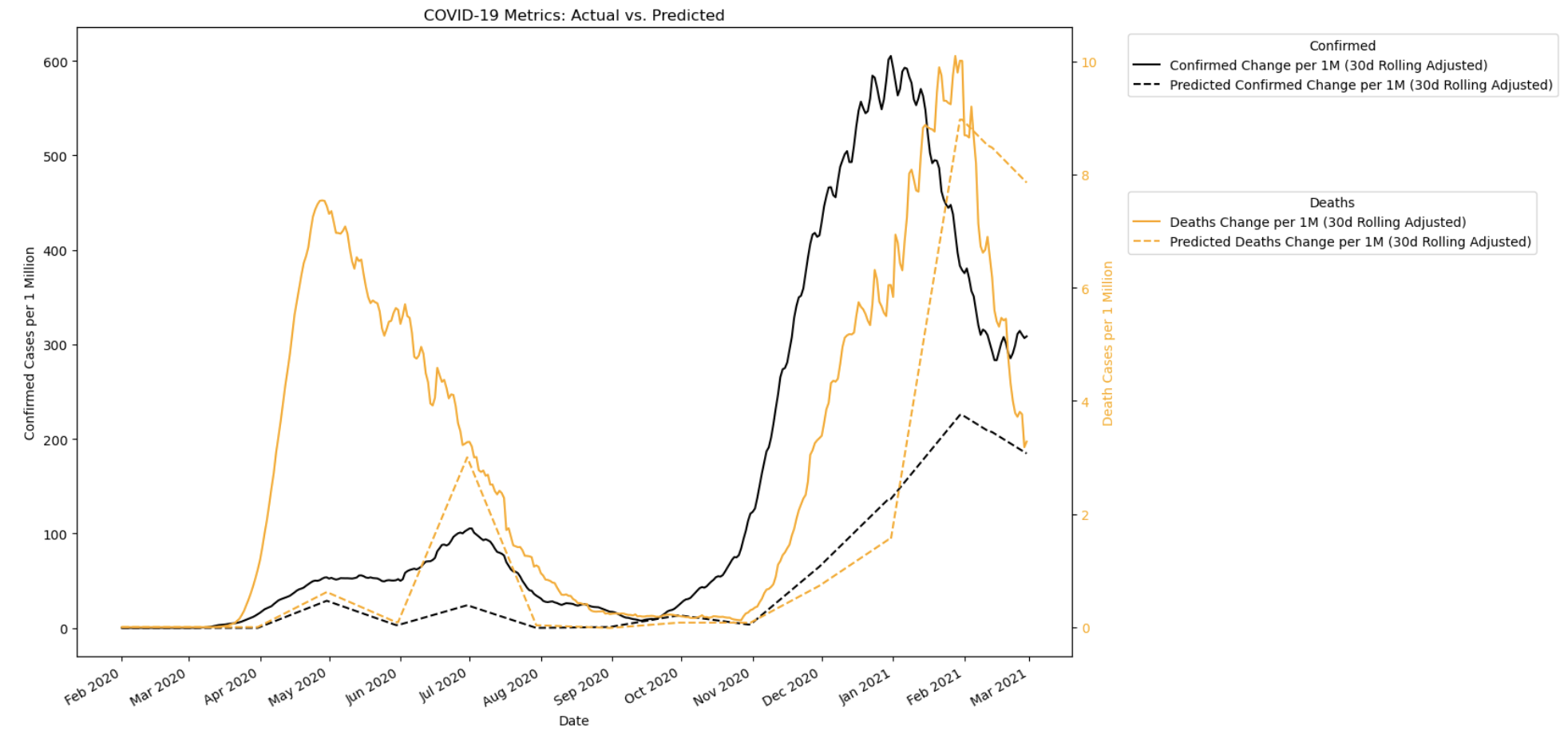


Figure 4: Predictive Machine Learning

We also built a seasonal ARIMA model for forecasting the rolling average of cases for the next two months of sweden. Below are the plots for the results from this model (Figure 5). The entire process is explained in depth in the attached document for Machine Learning and Forecasting.

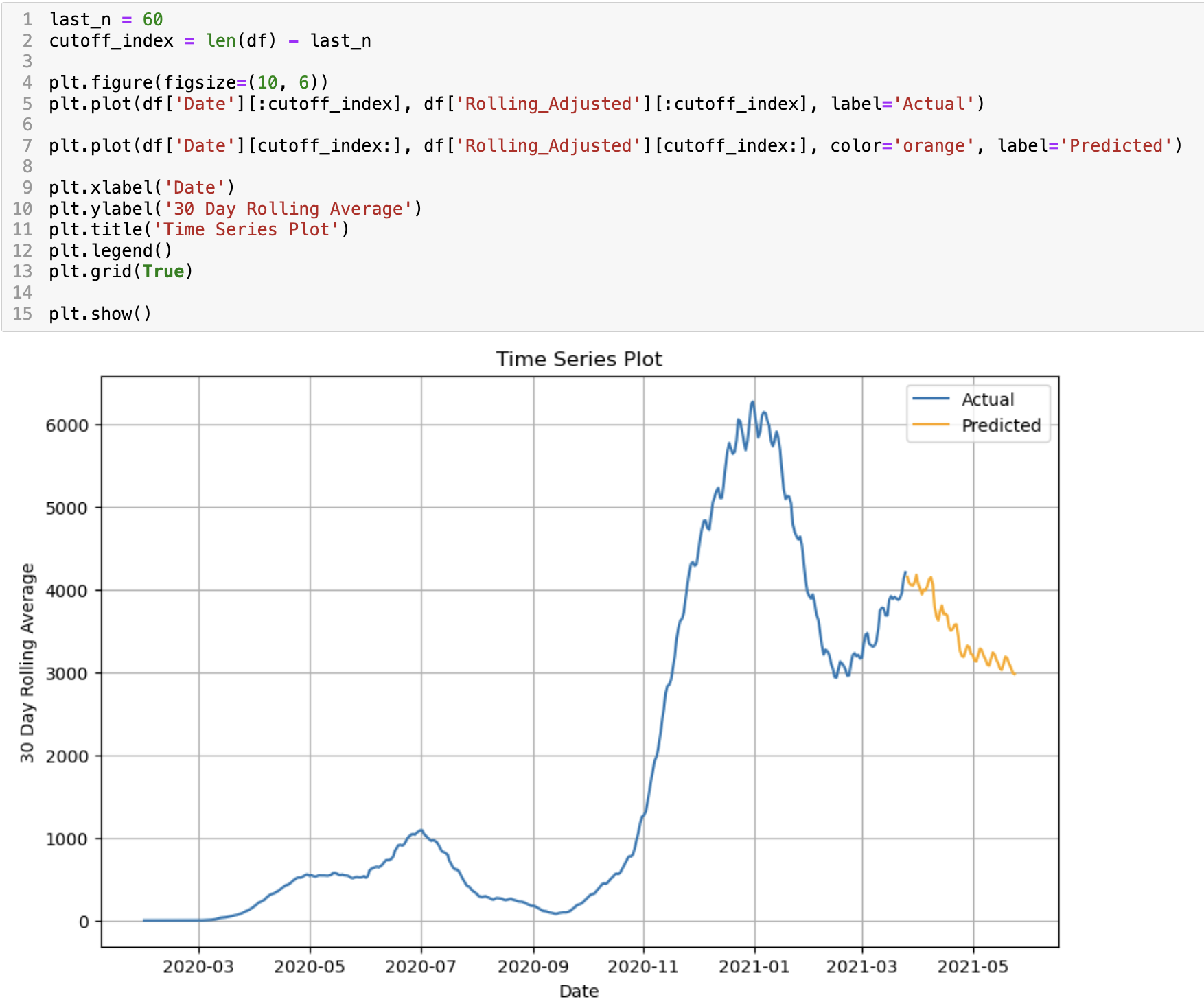
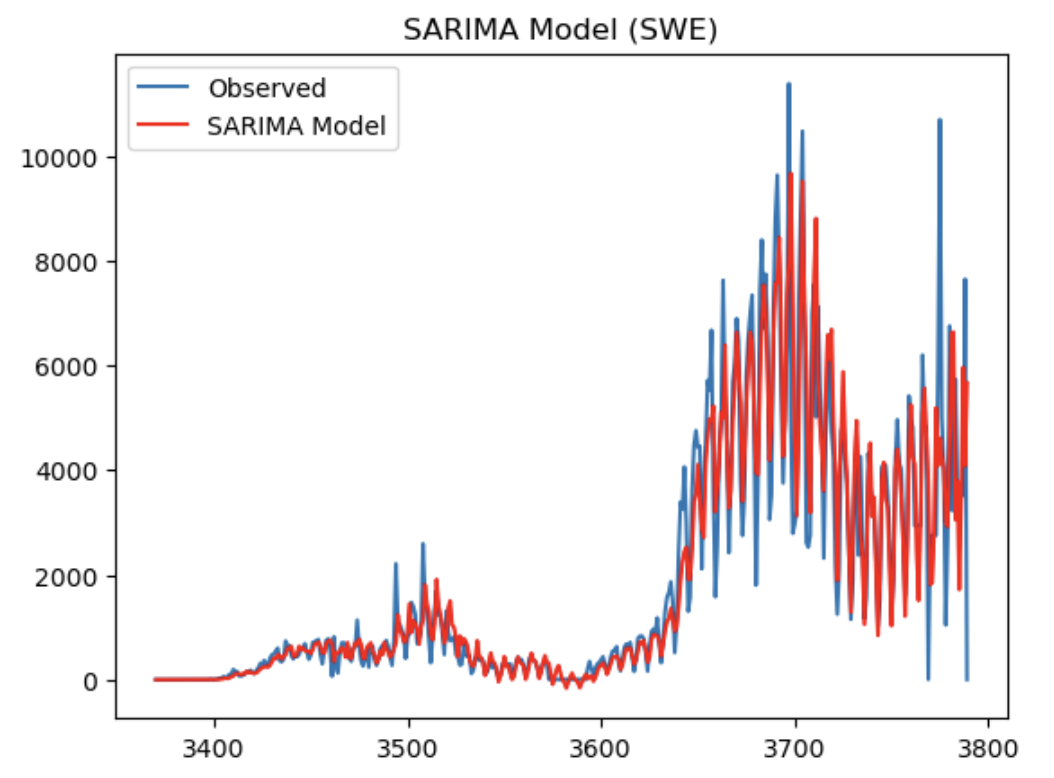


Figure 5: SARIMA Time Series Forecast

**Challenges**

Our team faced challenges in accomplishing our project goals. We had to overcome learning curves in working with Azure Data Factory, Azure Synapse, and PowerBI as a group. A specific example of this challenge occurred early on in the project, where pipelines were overwritten due to multiple users creating and running pipelines at the same time. As a result, our team compromised by setting up a weekly schedule where we would work on the project together. This prevented any overlaps in work and ensured our group members were providing equal contributions, even if the roles differed.

A major limitation that our team noticed in the initial extraction was the vast differences between the policy data and the rest of the metrics (cases, deaths, recoveries), being the country focus. In the policy data, there were numerous countries from around the world with a much larger variety in economic and geopolitical standing. However, the rest of the metrics featured ten countries all considered highly developed on the global stage. As a result, this presented a major limitation to the scope of the data since Caladan’s geopolitical, geographical, and economic status were left up to interpretation, presenting a difficulty in determining effective policies based on other countries, like New Zealand.

An alternative way of performing this project could have been to use more external data, in the form of policies. Using subjectivity to weigh the policies based on restriction presented a challenge to our team by leaving doubt about each of our choices in terms of determining the best “unrestrictive” policies to implement. Including more policies and weighing them in a way that represents Caladan’s priorities would provide a foundation for our team to work on, rather than place the engineers in charge of determining a significant element of the business requirements.

As a whole, this project allowed our team to learn from our own mistakes and develop the necessary skills in all elements of working with data. In a time where data is ubiquitous in all aspects of people’s lives, this understanding will push us forward in an industry and time where data engineers and business analysts are most needed.

**Group Roles**

Emily Forman - Project Manager

Abigail Merage - Data Engineer

Naman Nagaria - Data Architect

Patrick Soong - Project Manager

Aidan Clark - Data/BI Analyst

Despite our initial roles, we found that the vast majority of our efforts were done when we all could meet at once. That way, we got the chance to complete tasks together to gather a full understanding throughout the process.