Professor Seferlis
CDS DS310
5/6/2024

Data Engineering Project Executive Summary:
Analysis of COVID-19 Data and Policy Recommendations for the Commonwealth of Caladan

Objective

This project aims to construct a Modern Data Pipeline using COVID-19 data for Caladan, a fictitious midsize country with a population of 3.2 million. Given the concerns about a potential new wave of the virus as spring approaches, our task is to assist government leaders in developing a policy plan. The objective is to mitigate the impact of the virus while maintaining minimal restrictions. To do this, we wanted to identify policies that can keep the growth rate of deaths below 1% and the growth rate of new cases below 3% over a 30-day rolling average.

Data Pipeline

This section details the data exploration steps undertaken in the design of a Modern Data Pipeline for Caladan. The primary task involves extracting, cleaning, loading, and analyzing COVID-19 data from multiple international sources stored in different formats within an Azure environment. The data is sourced from three key storage solutions within Azure: Cosmos DB, a Virtual Machine (VM), and Azure SQL DB. This diversified storage approach simulates real-world data collection environments, providing a comprehensive dataset that includes policy data, as well as case, recovery, and death metrics. After creating a new storage account in our Azure resource group, we connected to the COVID-19 VM to simulate the extraction of on-premises data. The policy data in JSON format was extracted from Cosmos DB and converted into Parquet format for better performance and cost-effective storage in the data lake. Additionally, data from the Azure SQL DB, including information from the Country, Covid19_Metrics, and Dates tables, was extracted and stored as separate Parquet files to maintain data structure and ensure query efficiency. Data from the VM SQL Server, representing half of the international COVID-19 data metrics, was also converted and stored in Parquet format. Data from all sources was loaded into a structured directory within the Azure Data Lake. This structured format ensures that each data type is stored in its respective container, facilitating efficient data retrieval for analysis. The data was then processed and analyzed using Azure Data Factory and Synapse Analytics. By methodically acquiring, cleaning, integrating, and storing COVID-19-related data, we have prepared a robust dataset that will allow us to effectively assess the impact of health policies in Caladan.

Exploratory Data Analysis

Following the collection, processing, and exploration of our Covid-19 data, we performed our analysis. We primarily used PowerBI reports and machine learning techniques to reach our final conclusions. First, we visually examined the effect of each policy by country on the growth rate of new cases. Based on our initial graph analysis, we chose ten policies that we believe to be the most effective in keeping the growth rate below 3%. Our initial hypothesis considered school and workplace closing, stay at home requirements, internal movement restrictions, and international travel bans as effective closing policies. For effective health policies, we considered protection of the elderly, investment in healthcare and vaccines, public health campaigns, and facial coverings.

To assess restrictiveness, we explored the level of restriction implemented for each policy by country during the time period that met our growth rate threshold. Keeping our objective in mind, we prioritized the effectiveness of our policies first, and chose the least restrictive out of those initial findings. We ruled out international travel bans, internal movement restrictions, and school closings, as they were the maximum level of restrictiveness for the majority of our timeline. Additionally, we considered these policies to be inherently restrictive and we would not recommend enforcing them permanently. Our final selection was two closing policies (selected workplace closings and stay at home requirements) and two health policies (protection of elderly people and Covid-19 testing availability).

Confirmatory Data Analysis

To perform our CDA, we first programmed a correlation matrix and heatmap using Python, showcasing the correlation between a number of policies and the confirmed growth rate. We found a strong negative correlation between both protection of elderly and testing policy with the confirmed growth rate. Additionally, we observed a negative correlation with our two closing policies and confirmed growth rate.

To further our analysis, we used Python to train a gradient boosting regression algorithm, which assessed the feature importance of Covid-19 policies on the confirmed growth rate of cases and deaths. Among the policies, workplace closing, stay-at-home requirements, and protection of elderly people emerged as standout features, with feature importance values of 20.21%, 34.04%, and 23.38%, respectively.

To determine the optimal restriction level of these policies, we visualized the average growth rate of new cases by each of the restriction levels, for each policy. It was evident that, with the exception of stay-at-home requirements, level 2 had the lowest average growth rate of both cases and deaths across our chosen policies. We recommend implementing a level 1 of stay at home requirements.

Conclusion

To reiterate our solution, we are recommending that Caladan implements workplace closing, protection of elderly people, and testing policy at a level 2, and stay at home requirements at a level 1 to keep the growth rate of new cases and deaths below the threshold. We believe that by promoting remote work for certain employees and protecting those most vulnerable, Caladan will avoid spikes in confirmed cases and deaths due to Covid-19. If we were given more time for further research, we would like to explore more metrics and policies around this disease. To advance our statistical research, we could use more machine learning functions such as decision trees to support our findings. Lastly, we could use languages like R in order to find new statistics. Overall, we learned a lot from this project and enjoyed problem solving among our teammates.

Appendix

Responsibilities:

Miya Stauss: Project Manager

- Facilitated weekly meetings with advisor and group
- Helped with completing each challenge
- Created base of Data Flow
- Made base snowflake schema
- Created snowflake, galaxy, and data flow visualizations
- Helped with creating measures for growth and death rates
- Creating presentation

Maddie: Data Engineer I

- Extracted data from Azure SQL Database, Azure Cosmos DB, and an on-premise SQL Server using a virtual machine into data lake storage
- Processed data in Azure Data Factory to ensure consistent data types and formats across all columns and tables
- Created external tables in Azure Synapse
- Imported trade data for galaxy schema
- Created snowflake schema in PowerBI
- Designed final report in PowerBI

Perry: Data Architect

- Extracted and cleaned outside/extra data for use in a galaxy schema
- Creation of the galaxy schema
- Ensuring the Population by Year table was in usable format for the other measures such as the Gradient Boost
- Created a way to visualize population by day, using new cases per million, with Population table
- Helped create statistic on the growth rate of deaths per policy and growth rate of cases per policy

Lance: Data Engineer II

- Extracted data from Azure SQL Database, Azure Cosmos DB, and an on-premise SQL Server using a virtual machine into data lake storage processed data to ensure consistent data types and formats across all columns and tables
- Brought in one of the new datasets into our Azure Synapse to be ready for PowerBI analysis
- Standardized the confirmed cases by population from the second fact table in our Galaxy Schema and incorporated a Gradient Boosting Regressor to evaluate the model accuracy using MSE and evaluating the feature importance to make a bar chart