Hult International Business School

Business Intelligence – DAT-8564 – BOS2

A1: Individual Post-Assessment

M Raafi Jahangir

August 6, 2021

Table of Contents

[Introduction 3](#_Toc79161019)

[Methodology 3](#_Toc79161020)

[Data Selection 3](#_Toc79161021)

[Data Cleaning 3](#_Toc79161022)

[Relationships 3](#_Toc79161023)

[Measurements 3](#_Toc79161024)

[Pitfalls 3](#_Toc79161025)

[Dashboard 3](#_Toc79161026)

[Screenshot 3](#_Toc79161027)

[Design 3](#_Toc79161028)

[Final Remarks 3](#_Toc79161029)

[Key Insights 3](#_Toc79161030)

[Recommendations 4](#_Toc79161031)

[Conclusion 4](#_Toc79161032)

# Introduction

Coronavirus (COVID-19) has taken the world by a surprise. It has almost been two years since COVID-19 had its first case in December of 2019. The new normal now became staying indoors, wearing masks, and lockdowns. The world quickly adapted and pivoted to moving completely virtual, including businesses, schools, and other forms of services.

The aim of this dashboard is to show the pandemic in numbers, including cases that are confirmed, active, recovered and deaths. The Enterprise DNA challenge only used these values, which have been updated last in January 2021. Since vaccines quickly became prominent in the later part of the year, I decided to combine and include vaccination information to create a more relevant and interesting dashboard.

# Methodology

## Data Selection

The COVID-19 cases and vaccination data were obtained from the John Hopkins University (JHU) reports that are updated daily. These were publicly available on JHU’s GitHub repositories and were imported on PowerBI directly from the website URL. There were three different datasets for global confirmed cases, global recovered cases, global deaths, and vaccination information.

## Data Cleaning

The COVID-19 cases datasets are structured in a wide time-series which included columns of Province/State, Country/Region, Latitude, Longitude, and 280 more individual date columns. The dates were first unpivoted, so they are available in a long format with date as a and values as a column, with values respect to each specific country’s data.

Some columns were renamed, and data types were changed to their respective format, specifically date. A column was added at the very end to identify each dataset’s status. For example, the ‘Confirmed’ dataset had an added ‘status’ column that identified all values in that dataset as ‘Confirmed’.

The same changes were done to the ‘Recovered’ and ‘Deaths’ dataset so they all follow uniformed steps. This is important because all of them is appended onto a merged fact table with data values from the three datasets called ‘Global Cases Fact’. The three separate datasets were then disabled load since we have a merged fact table.

The ‘Confirmed’ table was then duplicated to create a country dimensions table called ‘DimCountry’. Duplicates were removed so each area were unique. An index column was also added to signify the dimensions. The ‘area\_id’ was merged onto the ‘Global Cases Fact’ table to associate each area.

The COVID-19 vaccination dataset was cleaned by removing null ‘country’ values and removing columns that weren’t needed. The ‘area\_id’ was also merged onto this table to continue the uniqueness of data. The table was renamed to ‘Global Vax Fact’.

## Additional Data

Two other datasets were imported onto excel as CSV files that included supplement information. The first table was ‘country\_population’ that was taken from the World Bank dataset. This dataset included information on the population, region, sub-region and whether it was a developing/developed or developing country. This information was merged onto the ‘DimCountry’ table which stored unique values. ‘Country\_population’ was then disabled load.

The second dataset imported was the ‘country\_ISO’ table that included the specific ISO-3 letters that are associated with all identified countries. The specific country\_ISO was merged onto ‘DimCountry’ to keep its specific country code. ‘country\_ISO’ was then disabled load.

There was a final table that was added, which was the dates table named ‘Calendar’. The code was copy pasted and added onto modifying a table. It was then set to dates table.

## Relationships

The relationships shared are ‘area\_id’ from the ‘Global Cases Fact’ to ‘DimCountry’ tables as well as the ‘Global\_Vax\_Fact’ to ‘DimCountry’. Both Fact tables were also connected with the ‘Calendar’ table with ‘Date’.

## Measurements

Multiple measurements were created throughout the process of creating this dashboard. This mainly included ratios and mathematical expressions. The main ones were individual ‘Confirmed’, ‘Recovered’, ‘Deaths’, ‘Partially\_vax’, ‘Fully\_vax’ totals by using SUM. These measurements were also divided by ‘Country\_population’ to get the % per population.

## Pitfalls

There have been multiple challenges when cleaning these data sets. One of the most frustration one was the discrepancies in country name that didn’t allow a smooth transition of country code to connect with each other. This was fixed manually by replacing the values of specific countries to match all tables.

Some of the values in the ‘Recovered’ table yielded a 0 value on the latest date import, which did not help the visualization since it must be filtered out for the last date that had values.

All values were in the form of cumulative so additional calculations had to be created to find out daily totals. The SUM function did not work if we wanted to find the final total because it would SUM the cumulative values. Instead of using SUM, MAX was used integrated with LastDate to find out the actual final values.

# Dashboard

## Screenshot

A screenshot of a video game

Description automatically generated

## Design

The design was chosen for its bright colors that identified the urgencies of specific attributes. The background was dark, so all visuals and colors pop out. This dashboard serves as a one-pager that includes global numbers of cases and vaccinations over time. All the data can be filtered out from the three slicers provided, including the date rage, country selection and region selection.

The map included a heatmap that showed the highest number of cases in each specific area in the world. The final matrix serves as a very important visualization because it is all relative to the country’s population. The numbers are very relative so finding out the specific percentages per population speaks a lot for the countries.

# Final Remarks

## Recommendations

1. Although countries may have a high confirmed percentage per population, the targets should be to fully vaccinate the nationals of the countries to mitigate the risk of spreading the virus.
2. Measuring the difference between recovered percentage and death percentage will show the biggest difference in the countries’ performance on how they are dealing with the pandemic.
3. The rate of vaccinating is exceeding the rate of confirmed cases, which show that the world is starting to keep up with getting everyone vaccinated. There are some countries with over 50% of the population fully vaccinated, which is great news. Having majority of the population at least getting partially vaccinated is the best way to go about slowing down the pace of the virus.

## Conclusion

Under developing countries are the ones struggling with vaccination and struggling with cases. Once developed countries are in a state where they can mitigate the virus spread, they should offer to continue helping to develop and under developing countries