Covid-19 Data Analysis and Visualization - Final Report

-Suyash Pandit

Dataset link:
https://data.who.int/dashboards/covid19/data
https://www.kaggle.com/datasets/andradaolteanu/country-mapping-iso-continent-region

Goals:
Data Collection, Data Cleaning, Data Visualization

Problem statement:
1. Datasets with lots of null values
2. Missing columns in datasets
3. Country codes mentioned in rows of country column
4. Different criterias of data submission by different countries
5. Small island nations and independent territories not having WHO ISO codes

Milestone 1:
Data Collection and Cleaning was done and a python notebook with the first milestone goals was created.
Some of the methods used were:
- Dropping columns for lack of significance and to ensure clarity while building dashboards
- Renaming columns for better visualization when building dashborads
- Replacing null values with 0
- Changing data types
- Reordering columns
- Merging queries using join functions to obtain new column as needed in the dataset

Milestone 2:
Exploratory Data Analysis and Insight Documentation was done and a python notebook with the second milestone goals was created.
Some of the actions done were:
- Creating measures to obtaining descriptive statistics
- Clear explanation of why some statistic was obtained while others were not
- Table visuals with the created statistical values
- KPI cards presented for some metrics
- New measure created for total deaths by country using “removefilters” and “max” functions in the DAX query to get a trend analysis

Milestone 3:

Insight documentation:

A clear data visualization with presentable insight was made using PowerBI. Included below are details about each visualization and insights. The entire PowerBI interface has been pushed to github for direct viewing of the insight.

1. Global median being close to zero for both cases and deaths clearly showed the spikes that was happening during the pandemic.
2. Due to the close to zero median in global deaths and cases, I created a new table with only 11 countries with varied deaths and cases which clearly showed that:
    a) Median values for individual countries which had large outbreaks, like US and India are high compared to the countries which had low population and also had low reporting
    b) There are many countries with zero cases or deaths especially during the early days of the pandemic as seen in the dataset.
    c) Median is the middle value. This presents the concept of data being skewed (ups and downs in values really quick).
3. Line chart depicting the gradual increase in global new cases upto early 2022 and the decrease after that shows that COVID-19 cases decreased after 2022. Deaths had already started decreasing as per the line chart in the same visualization which shows that people had starting adopting vaccine and masking mandates. Vaccine acceptance and more preventative measures being affected is reflected in this line chart as the case and death both decrease.
4. Uptick in the total cases from China in 2023 which is clearly due to under-reporting by China during the pandemic. WHO probably had to adjust the previous values in this quarterly report, which showed that China had this sudden uptick. This can be accessed interactively in the PowerBI interface file that is uploaded.
5. The fifth visualization has two bar charts. China always had less data coming out of the country formally, there is very less death reported despite China being the origin and having many death reports in the media. Objectively looking, China is not even in the top 10 countries with death.
6. China is in the first place for vaccine coverage for atleast one dose. This could be due to China having the first vaccine Sinovac during the pandemic (information I obtained while working as a healthcare professional). This was insightful because immediately follows India in second place who also started its own production of the Astrageneca vaccine from the UK. United States having less vaccine coverage could be due to vaccine hesitancy in the country.
7. There is no clear indication of increased vaccination and lower death rates based on the limited data from WHO with many flawed datasets. The vaccine accessibilty, poverty level and demographic and cultural factors cannot be ignored.