**Netflix & COVID Data Analysis & Visualization**

**Qussay Al-Qirim**

The case at hand focuses on analysing two different datasets, Netflix and COVID, with the objective of producing visualizations to answer a set of proposed questions in appendix B. As a data analyst, in order to ask insightful questions, a comprehensive understanding of the dataset is essential. This would allow the analyst to realize different attributes, data types and data quality, in turn acting as a starting point for data modelling and piecing attributes together to answer questions.

A data quality overview of the Netflix data set showed inaccuracy in the information where data did not reflect reality, and this could be attributed to poor data collecting. For example, George Lucas did not direct a movie titled “Im standing on a Million Lives”, allowing for similar inaccuracies across the dataset. Whereas the COVID dataset was more accurate and comprehensive.

The created data model will be based on data warehouse principles, composed of fact and dimension tables. Kimball & Ross (2002) define dimension tables as tables with descriptive attributes, characterized by their unique data and are used for slicing and dicing. Whereas fact tables are tables with numerical measurements upon which slicing, and dicing occurs.

The models created in appendix A follow a four-step dimensional design process highlighted by Kimball & Ross (2002) all the while adhering to the principles of data warehousing. The initial step to the process was to identify the end goal for each data set, resulting in a Ranking & Sales Fact table for Netflix in Figure A1 and, similarly, COVID and Vaccination data tables for COVID in Figure A2. Second, understanding what each row in the fact table will describe will define the granularity of the table. For Ranking & Sales in figure A1, each row entails details, reviews & box office for each title. Additionally, the COVID data fact table in figure A2 describes the daily number of cases and deaths for each country, while Vaccination data table details the total number of vaccinations and vaccination rate for each country. After identifying the end goal and granularity for each fact table, the dimension tables that apply to the fact table are created as could be seen in figures A1 and A2. Finally, the numeric attributes are identified in each fact table.

To transform the data, it must be extracted first. For the Netflix data, it is a matter of importing the data into power query through the get data function and selecting Text/CSV as displayed in figure 1.1.1. On the other hand, the COVID data sets are updated daily and therefore recommended to extract data via web URL as displayed in figure 1.1.2 & 1.1.3.

***Graphical user interface, text, application

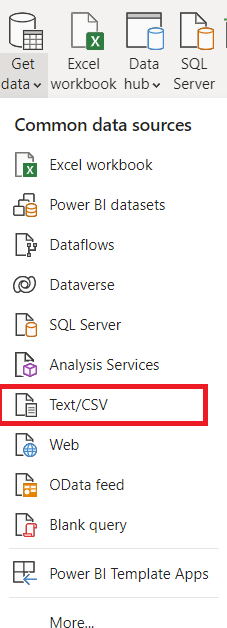
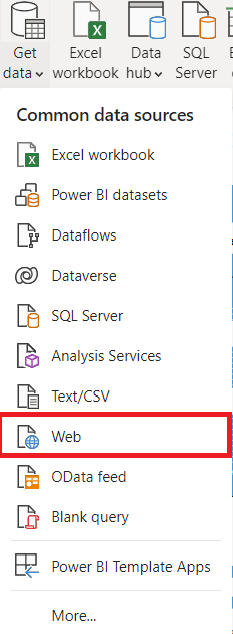
Description automatically generated***

Figure .1.3: COVID URL

Figure 1.1.1: Import COVID Dataset

Figure1.1.2: Import COVID Dataset

After importing the data, transforming the data is essential to develop the models in appendix A. Clicking on Transform Data in figure 1.2 will open the data set in Power Query to perform further transformation processes.

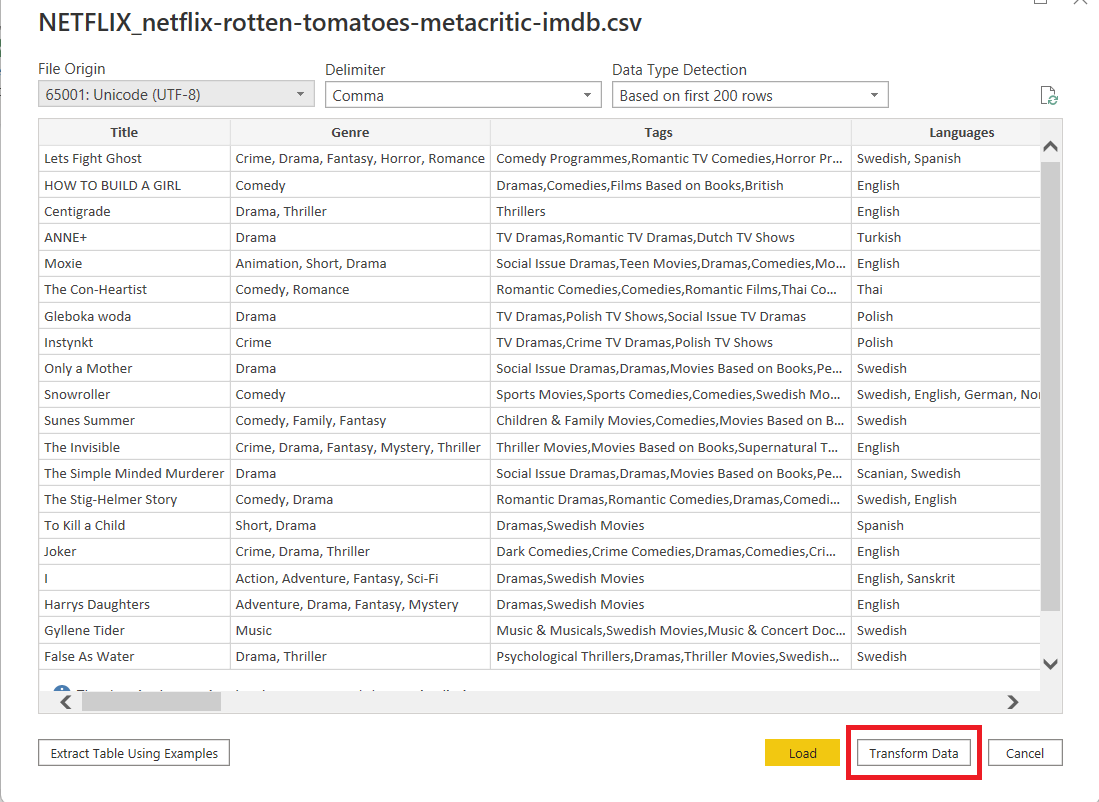


Figure 1.2: Transform Data

Dimension and Fact Tables were created by duplicating the Main data set after which cleaning and transforming the data was performed. To create unique dimensional tables, duplicate values were removed. As such, data was cleaned using the following format functions: Uppercase, Trim & Clean. After which empty and duplicate rows were removed.

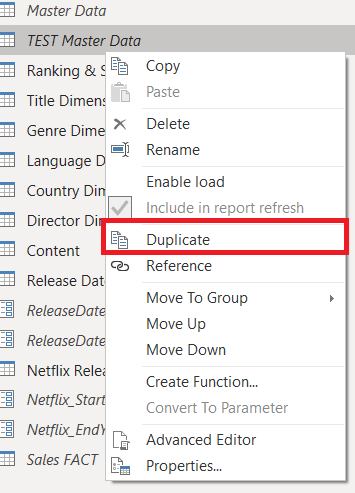


Figure 1.3.3: Clearing Rows

Figure 1.3.1: Duplicates

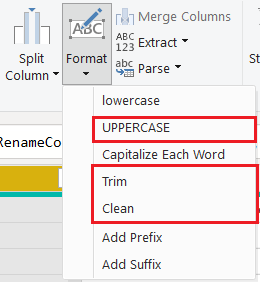
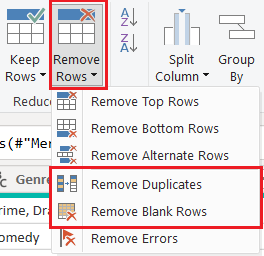


Figure 1.3.2: Format Functions

Each dimension table must contain a surrogate key that is unique per row for that dimension table. This achieved by creating an index column as shown in figure 1.4.1 & 1.4.2.

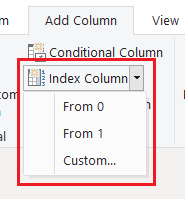


Figure 1.4.1: Index Column



Figure 1.4.2: Genre Dimension Example

The dimension table was then connected to the fact table via a common column through the merge query shown in figure 1.5.1. For instance, Genre dimension is merged to Fact table based on the Genre Column, displayed in figure 1.5.2.

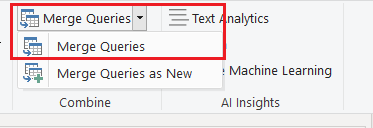


Figure 1.5.1: Merge Query Button

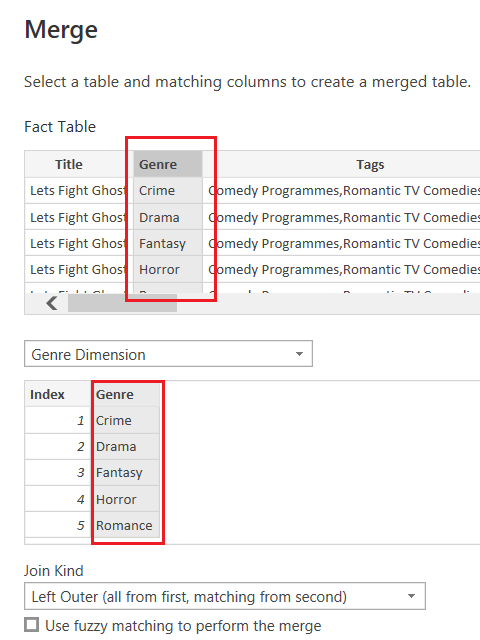
After completing the transformation process and creating the dimension and fact tables, Close & Apply was selected to load the transformed data. Under the relationship tab, the relationships created based on the primary keys along with their cardinalities were visible.

Figure 1.5.2: Merging Tables

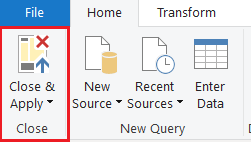
******

Figure 1.6.1: Close & Apply Button

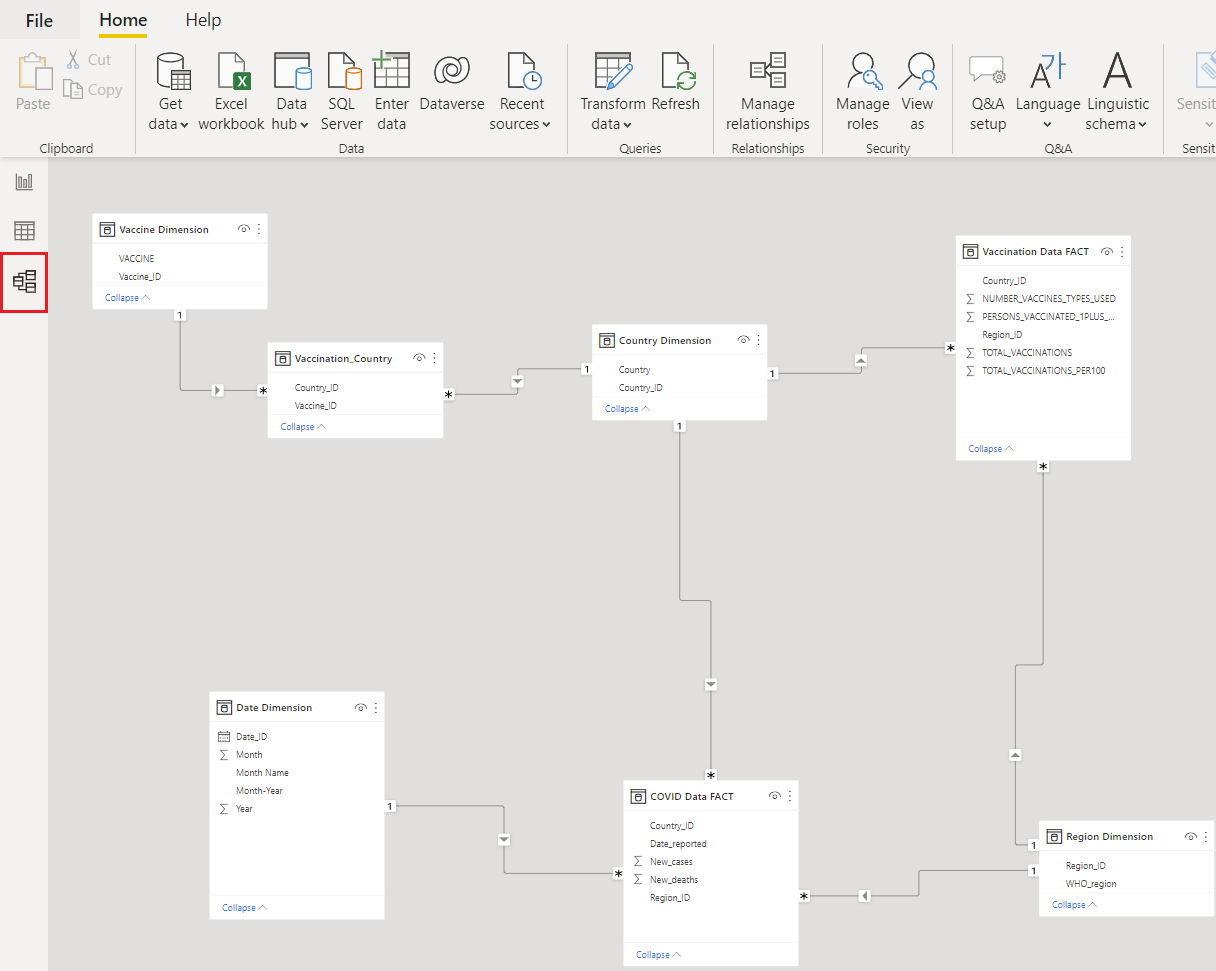
******

Figure 1.6.2: COVID Data Model

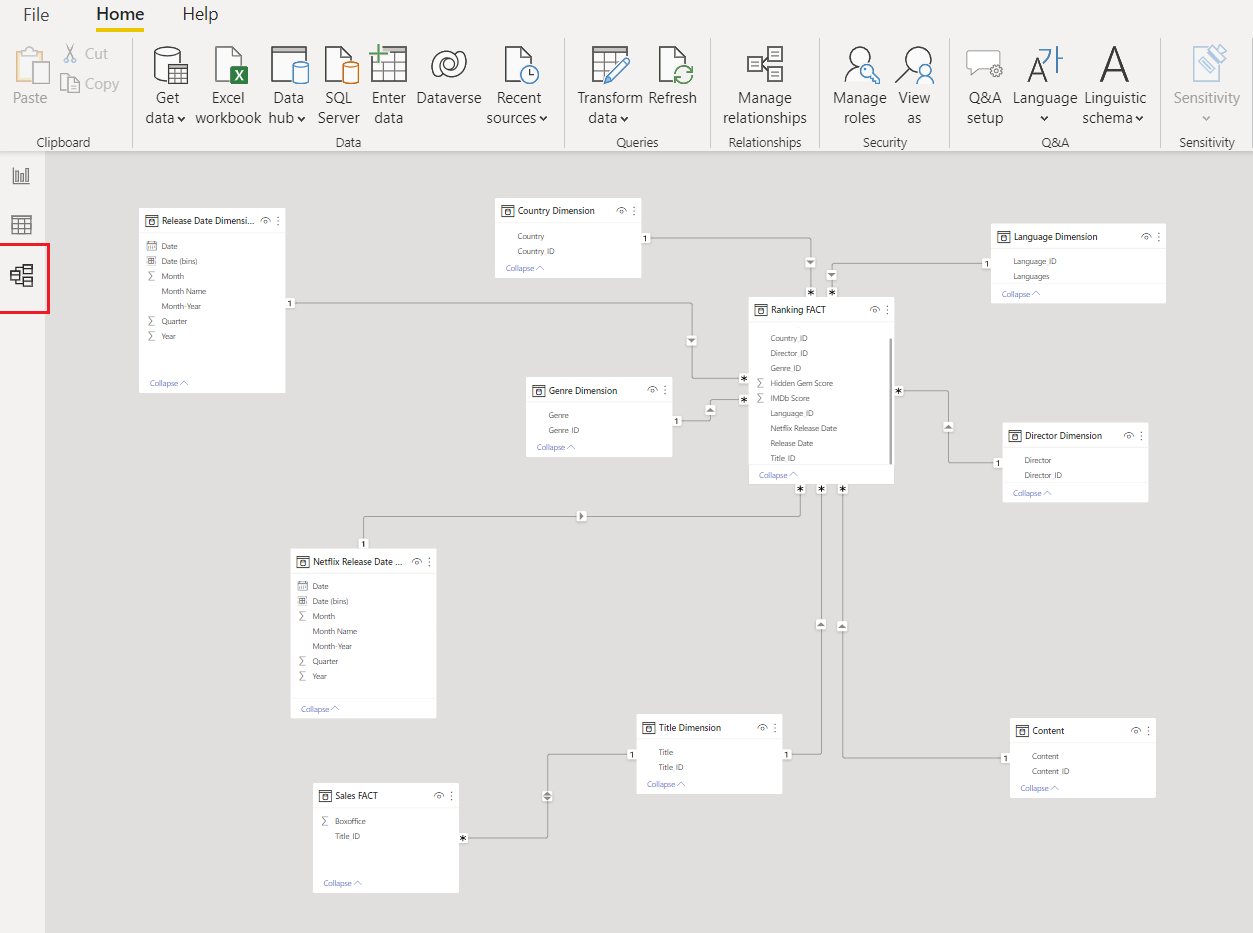
****

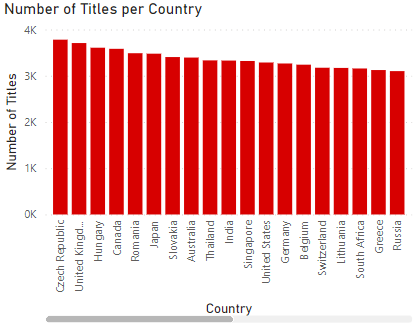
Figure 1.6.3: Netflix Data Model

Sequentially, after completing the previous processes, producing visualizations to answer the proposed questions was possible. Below are the answers for the questions along with the choice of chart selection.

**Netflix Dataset:**

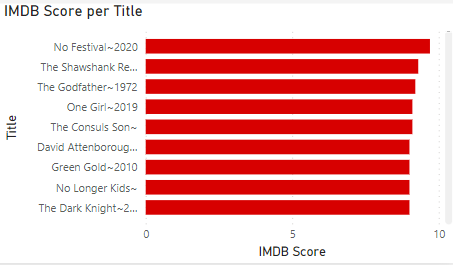
1. Which Country has the highest amount of movies?

A Clustered Column Chart was used to answer this question and from the visual Czech Republic is the country with the highest number of movies.



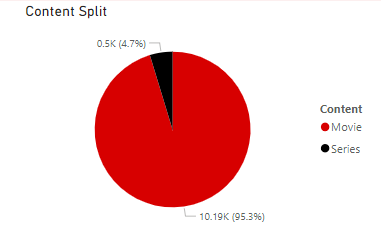
1. Which Movie has the highest IMDB Score?

A Clustered Bar chart was used to answer this question and the movie No Festival released in 2020 was the highest rated.



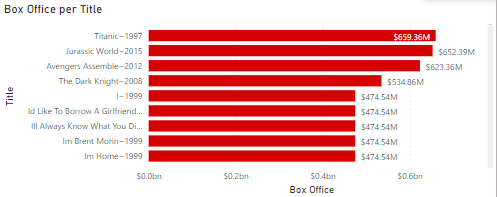
1. Content split between series and Movies

Given 2 categories, a pie chart was selected to compare amount of content between series and movies. The analysis shows a split of 95.3% for Movies and 4.7% for Series.



1. Which Movies has the highest box office?

A Clustered Bar Chart shows that Titanic released in 1997 had the highest amount of Box Office sales at 659 million dollars.



1. Which Series has the highest IMDB Score?

A clustered bar chart filtered on Series displays the title Im Sorry released in 2021 was rated the highest IMDB score at 9.2.

Chart, bar chart

Description automatically generated

**COVID Dataset:**

1. **Which month had the greatest number of COVID Deaths?**

On the Line chart displayed below, May 2021 had the highest amount of deaths totalling 383,122.

Chart, line chart

Description automatically generated

1. **Which region had the greatest number of cases?**

According to the clustered bar chart, Europe had the greatest number of cases.

Timeline

Description automatically generated with medium confidence

1. **Which region had the greatest number of deaths?**

The clustered bar chart below shows the Eastern Mediterranean Region with the greatest number of deaths.

Chart, bar chart

Description automatically generated

1. **What are the total number of administered doses?**

A card visual was used to demonstrate the total number of administered vaccines, numbering at 12, 782,955,639.

Graphical user interface

Description automatically generated with medium confidence

1. **What are the total number of people with at least one dose?**

Similarly, another card visual was used and the number of people with at least one dose is 5,386,127,571.

Text

Description automatically generated with low confidence

1. **Which countries used the most type of vaccines?**

According the clustered bar chart, Iran used the most type of vaccines totalling at 12 different types of vaccines.

Graphical user interface

Description automatically generated with low confidence

References:

Kimball, R., & Ross, M. (2002). *The data warehouse toolkit*. New York: Wiley.

**Appendix A**

Netflix and COVID Data Entity Relationship Diagrams

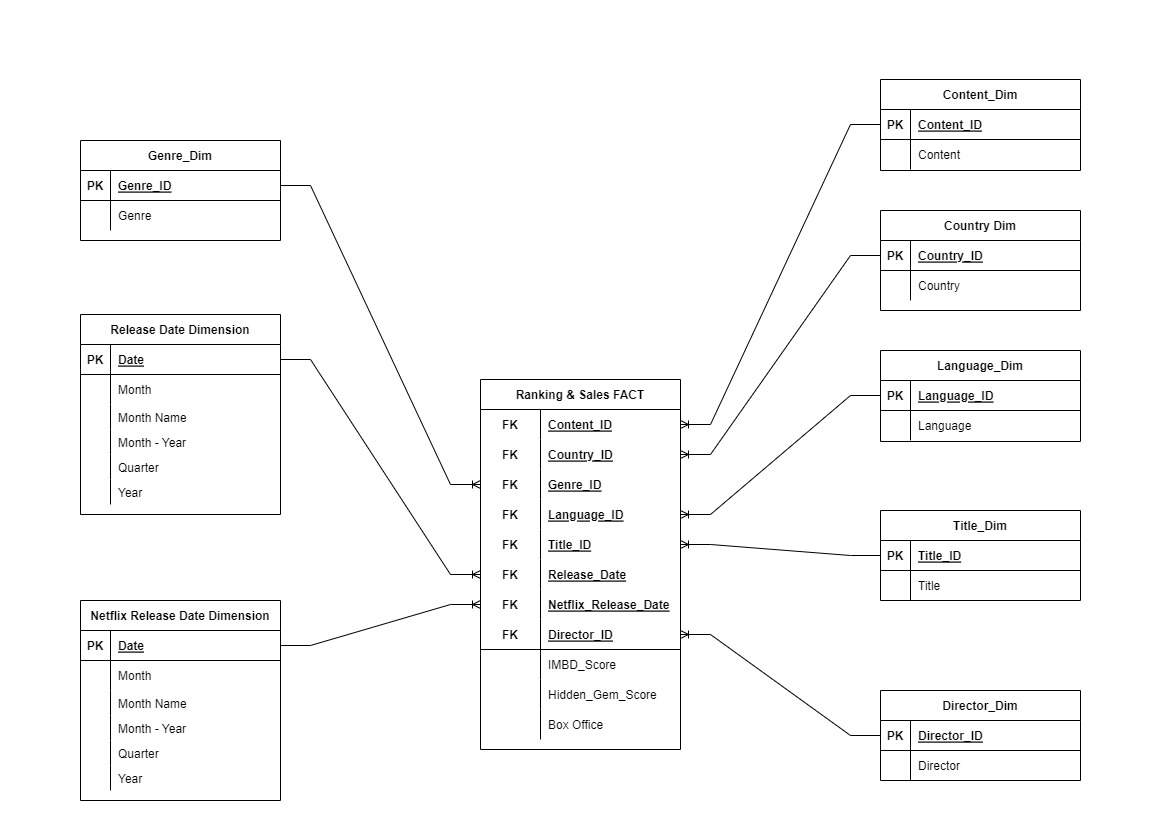


Figure A1: Netflix Entity Relationship Diagram

Diagram

Description automatically generated

Figure A2: COVID Entity Relationship Diagram

**Appendix B**

Proposed Analytical Questions

**Netflix Data:**

1. Which Country has the highest amount of Movies?
2. Which Movie has the highest IMDB Score?
3. Content split between series and Movies
4. Which Movies has the highest box office?
5. Which Series has the highest IMDB Score?

**COVID Data:**

1. Which month had the greatest number of COVID Deaths?
2. Which region had the greatest number of cases?
3. which region had the greatest number of deaths?
4. What are the total number of administered doses?
5. What are the total number of people with at least one dose?
6. Which countries used the most type of vaccines?