<u>WQD 7009: ALTERNATIVE ASSESSMENT</u>

Question 1

<u>Data capturing</u>

For enterprises, especially in the healthcare industry, capturing data that is clean, comprehensive, accurate, and structured appropriately for usage in multiple systems is a constant battle. From a recent study in an eye clinic, only 23.5 percent records of EHR (Electronic Health Record) data matched with patient-reported data. Patients' reports with three or more eye health symptoms were inconsistent with their EHR data. Ineffective EHR usability, complicated procedures and a lack of understanding of the value of good big data acquisition can all lead to quality problems that will afflict data throughout its course.

<u>Data cleaning</u>

When combining various data sources that may capture clinical or operational variables in different formats, dirty data can easily wreck a big data analytics effort. The process of cleaning data ensures that it is accurate, consistent, correct, relevant and that it is not in any way corrupted.

<u>Data storage</u>

Front-line clinicians seldom consider where their data is kept, but the IT department views this as a crucial cost, security, and performance concern. Some healthcare providers are no longer able to control the costs and effects of on-premise data centres as the volume of healthcare data increases dramatically. Many firms are comfortable with on-premise data storage offering more control over security, access and uptime. However, an on-site server network may be expensive to grow, challenging to manage and prone to creating data silos across various departments.

<u>Data querying</u>

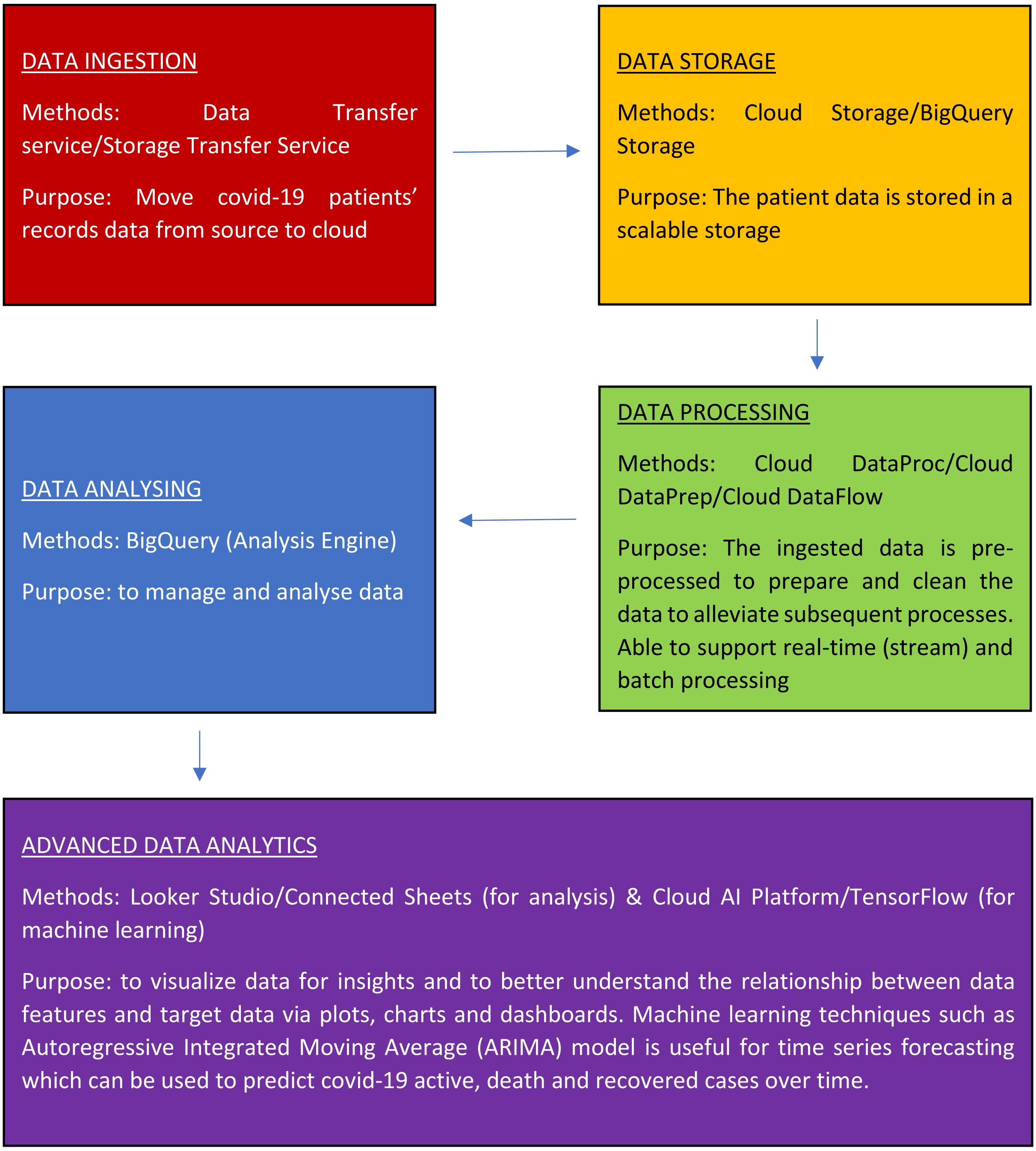
Healthcare organisations typically face a number of obstacles before they can conduct meaningful analysis of their big data assets, such as overcoming data silos and interoperability issues that prevent query tools from accessing the organization's entire repository of data. The ability to query data is fundamental for reporting and analytics in order to obtain the necessary insights and information. It might not be feasible to provide a comprehensive image of the state of an organisation or the health of a specific patient if different components of a dataset are stored in various walled-off systems or in different formats.

<u>Data updating</u>

Healthcare data is not stagnant and the majority of its components will require regular changes to be up to date and useful. Organizations who do not constantly monitor their data assets may find it difficult to comprehend the volatility of big data, or how frequently and how much it changes. In order to update datasets without affecting end users, providers must have a clear understanding of which ones may be automated and which ones require human updating. Organisations should also be careful to avoid producing unneeded duplicate records which may make it difficult for physicians to get the data required for crucial decision-making process for patients.

Question 2

Google cloud is a platform which can help solve all the challenges mentioned in (1) due to its computing capability and big data technologies it possesses. This is made possible via the offered scalability, advanced analysis and analytics techniques and most importantly, performance where the output for each function performed in the cloud platform is provided within seconds. Below is a diagram of a data analytics pipeline in Google Cloud platform to help process, analyse and store patients' records.



Question 3

High Data Storage Capacity (Scalability)

Medical records, medications, and lab results are just a few examples of the digital data that hospitals and other healthcare facilities generate daily. Such data to be stored locally requires procuring expensive storage equipment. However, the cloud provides limitless storage and is effective at handling massive amounts of data. The ability to add extra storage in accordance with your needs is beneficial in the healthcare industry.

Improved interoperability

System interoperability is essential for delivering the best treatment possible. Data exchange from medical apps, devices and systems is made possible through comprehensive cloud interoperability. This guarantees that healthcare professionals and other authorised individuals have access to patient data.

Efficient analysis

Cloud solutions provide improved data monitoring and analysis which is linked to the identification and treatment of various illnesses. It aids healthcare organisations in the development of practitioners, the detection of scan abnormalities, and the forecasting of disease epidemics.

Artificial intelligence and machine learning

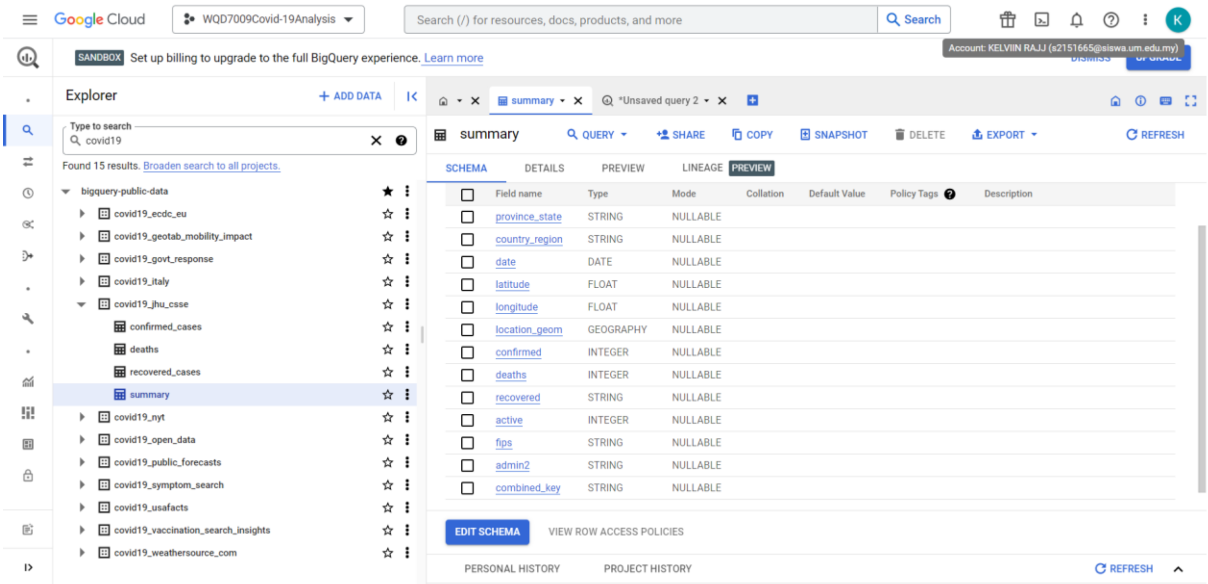
Real-time automated analytics powered by artificial intelligence and machine learning algorithms are provided through cloud solutions. Cloud computing can enable the adoption of artificial intelligence into standard healthcare operations which is vital for supporting massive database administration, clinical decision making, and treatment time reduction.

Improved Data Security

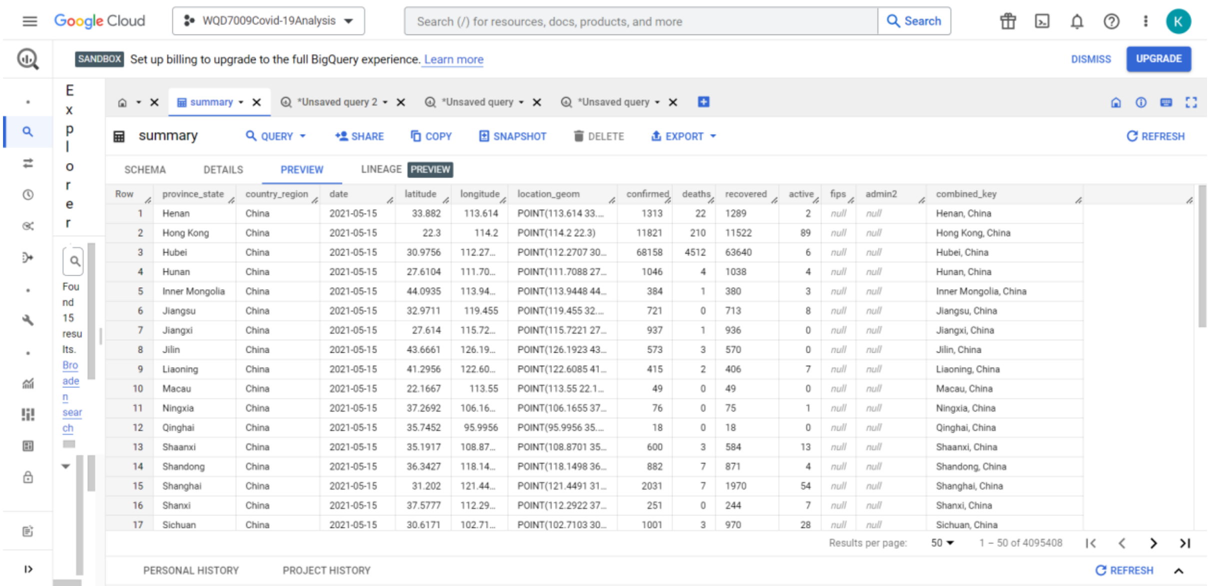
The European General Data Protection Regulation (GDPR), the US Health Insurance Portability and Accountability Act (HIPAA) and the CSF HITRUST Alliance industry standard all support and govern the protection of users' personal information in online services especially cloud platforms. HIPAA prioritises maintaining its security through the use of HIPAA-compliant secure EHR solutions, which is offered by cloud platform providers like Google Cloud who are able to dedicate specialised resources and security mechanisms to assure high level cybersecurity.

Question 4

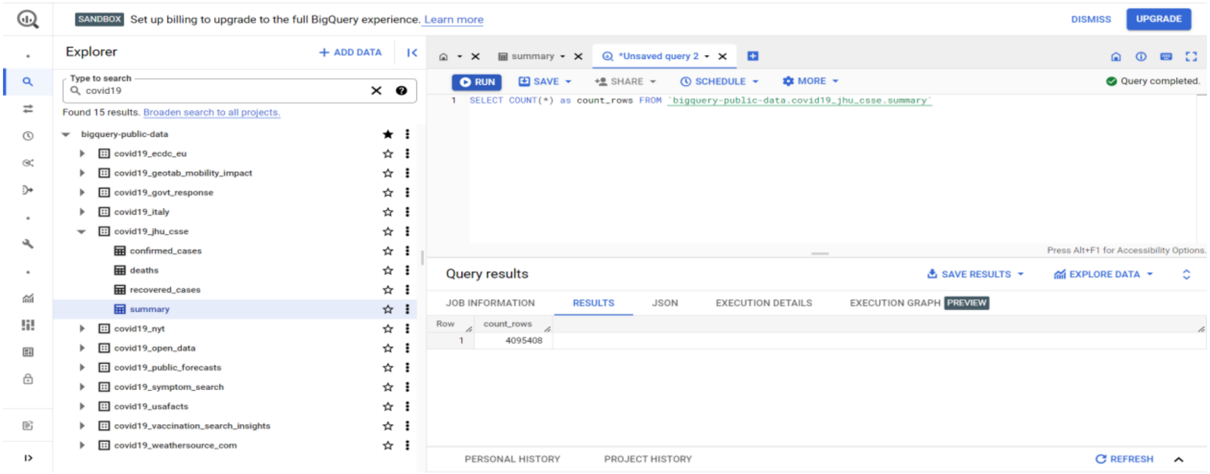
For the implementation section, the public dataset available in Google Cloud platform is selected for querying and visualization. As can be seen in the diagram below, there are 13 features (columns) in this dataset that can be analysed.



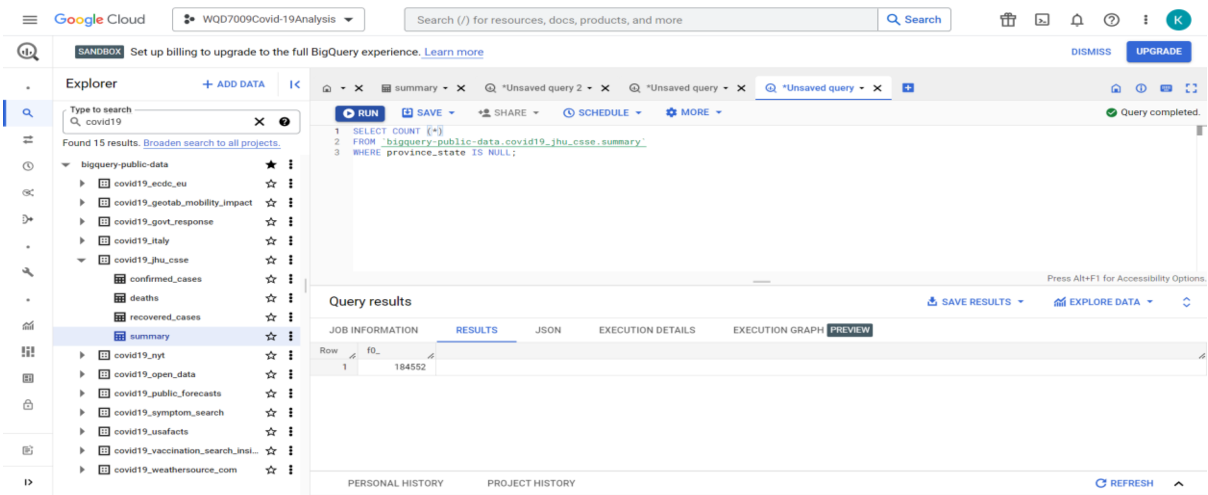
Below is an overview of the data in each column. This analysis will revolve on the countries with **confirmed, deaths, recovered** and **active** cases to determine which country has the highest cases in each category.



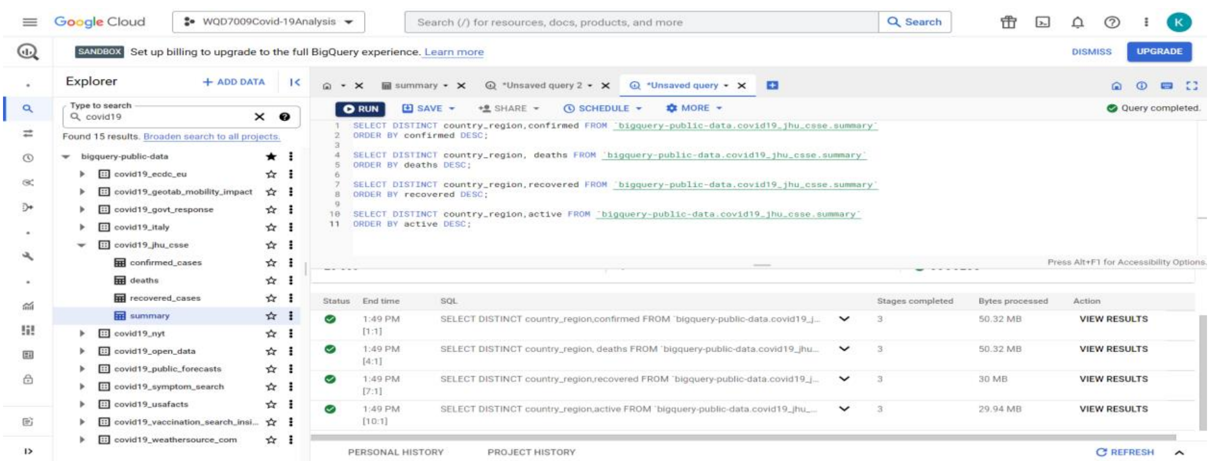
The dataset contains 4,095,408 rows of records.



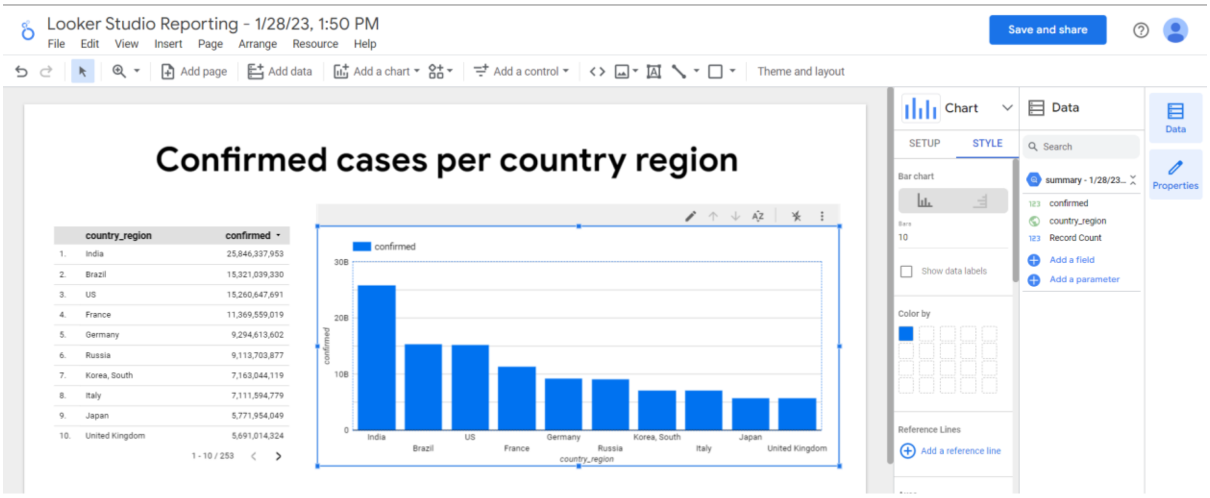
When checked for missing data, the **province_state** feature has significant number of null values (184,552 rows of records contain no data). Hence, the alternate feature (**country_region**) will be used for this analysis.

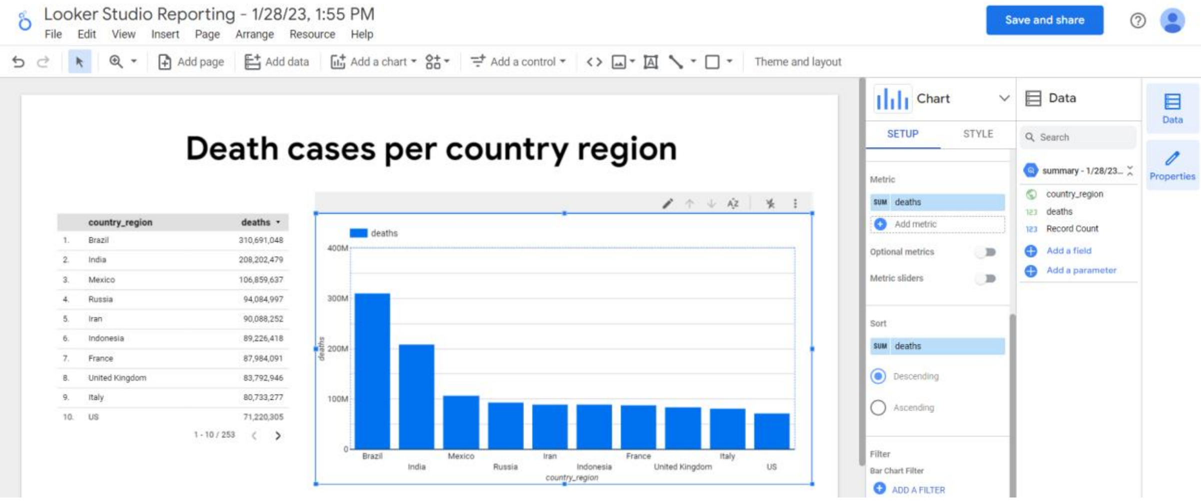


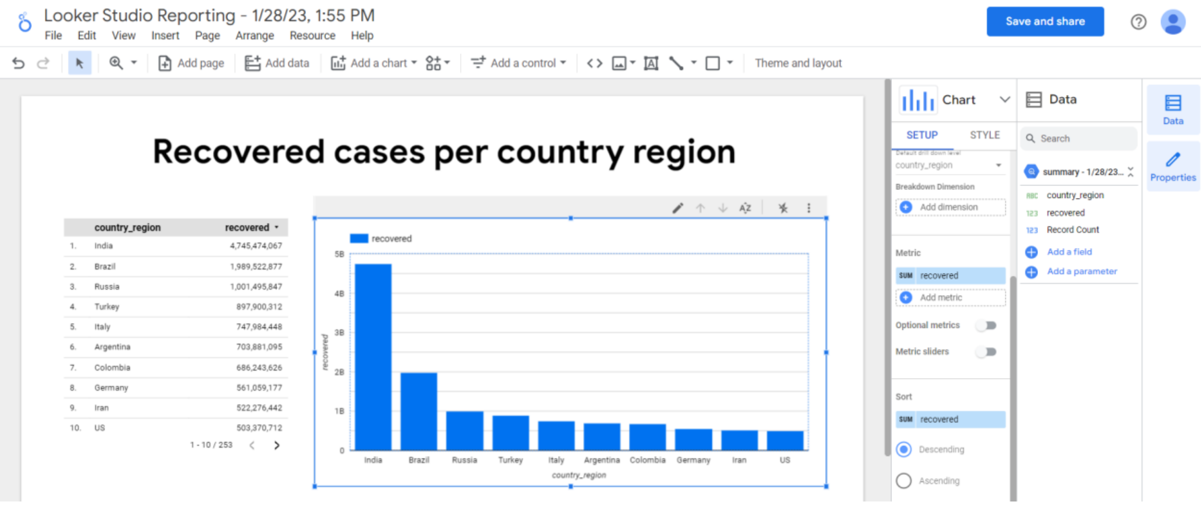
The query is run for **country_region** feature on each category of cases. The query was completed successfully as can be seen below.

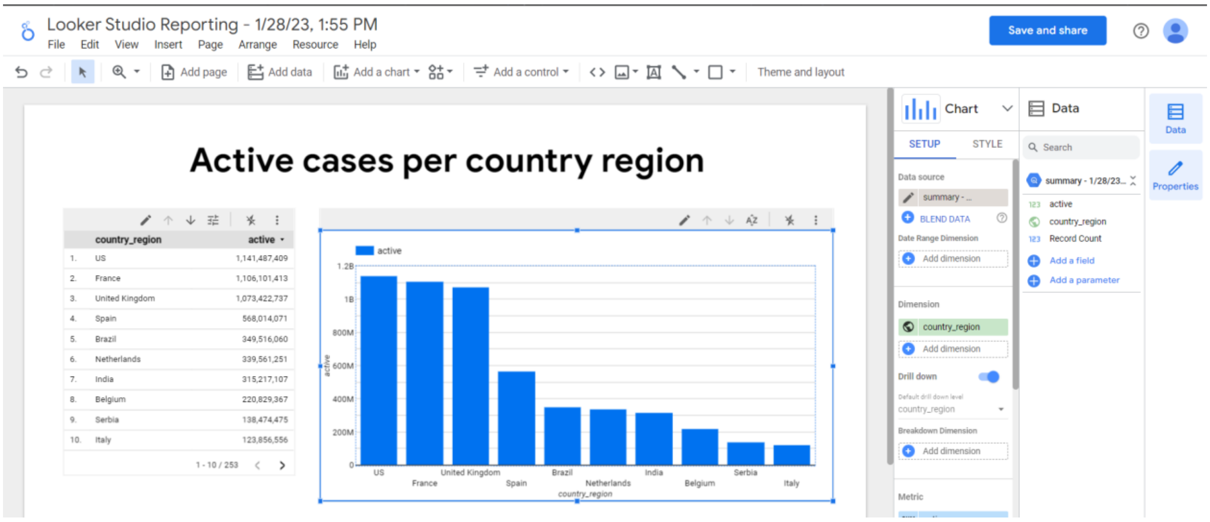


Below are visualization graphs generated in Looker Studio for each query performed for each case category (**confirmed, deaths, recovered** and **active**).









Based on the visualization obtained from Looker Studio, below is the summarized rank order from highest to lowest for each case category using the top 5 **country_region** features:

Confirmed:

1. India (25.8 billion)
2. Brazil (15.3 billion)
3. United States (15.2 billion)
4. France (11.4 billion)
5. Germany (9.3 billion)

Death:

1. Brazil (310 million)
2. India (208 million)
3. Mexico (106 million)
4. Russia (94 million)
5. Iran (90 million)

Recovered:

1. India (4.7 billion)
2. Brazil (1.9 billion)
3. Russia (1.0 billion)
4. Turkey (897 million)
5. Italy (747 million)

Active:

1. United States (1.14 billion)
2. France (1.10 billion)
3. United Kingdom (1.07 billion)
4. Spain (568 million)
5. Brazil (349 million)