# Title: Analysis of COVID-19 tracking data based on the values extracted from different states in the USA

Rohini Vemparala, Sai Tejaswi Bodempudi, Swathi Sri Kilaru, Nikita Firozali Samanani, Proveen Kothagundla, Deepak Patel Chityala.

[rvempara@iu.edu,](mailto:rvempara@iu.edu) [sbodempu@iu.edu,](mailto:sbodempu@iu.edu) [swkilaru@iu.edu,](mailto:swkilaru@iu.edu) [nsamanan@iu.edu,](mailto:nsamanan@iu.edu) [pkothag@iu.edu,](mailto:pkothag@iu.edu) [dchityal@iu.edu,](mailto:dchityal@iu.edu)

Indiana University- Purdue University-Indianapolis (IUPUI) Indianapolis, Indiana, United States of America-46202.

Course Name: SP23: SCIENTIFIC&CLINICAL DATA MGMT: 24860

Instructor: Jeremy Harper

# Abstract

In this paper, we as a team gathered the COVID–19 data and the numbers involving its tracking specific to a country (U.S.A), the data concerning various states. Through our project, we analyzed the available open-source data relating to the global public health pandemic. Our analysis mostly focused on some important columns, like the number of positive cases within the years 2020, and the year 2021. The number of deaths and the number of hospitalized people columns were analyzed, using some of the SQL functions like the SUM, AVG, JOIN BY, GROUP BY, and COUNT functions. Our major research questions were to analyze the pattern of distribution of the data among different states, and to check their form of data distribution by formulating various visualizations. Finally, the ERD diagrams were designed to represent the relationships with each of the specific columns with the available CSV document, using the database management systems. The primary objective of this project is to leverage SQL commands to improve the quality of existing data, enabling more precise and informative research studies on the COVID-19 pandemic. Through this effort, we aim to raise awareness about the detrimental impact of this virus on public health and highlight the need for comprehensive efforts to mitigate its spread. The project seeks to address gaps in existing data by providing a more precise and detailed report that establishes relationships between various factors related to the COVID-19 pandemic. This includes disease severity, hospitalization rates, and mortality rates, among others. The goal is to provide future researchers with more accurate and comprehensive data to facilitate a deeper understanding of the pandemic's impact and inform more effective public health interventions.

# Introduction

Coronavirus disease 2019 is an illness due to the virus SARS-CoV-2 which was first got encountered in Wuhan, China during the year of 2019 in December. The disease is extremely infectious and has speedily been raised and explored all around the globe. Coronavirus-19 most generally affects the respiratory system with symptoms like cold, flu, or pneumonia. Apart from just the respiratory system and the lungs, other parts of the body could also be affected by the disease-causing virus. (Centers for disease control and prevention, 2020).

This virus spread throughout the world over the next few weeks. On January 7, 2020, a Chinese scientific research institute announced that viral pneumonia was caused by a novel coronavirus (SARS-COV-2). According to Worldometer (2020), there were 25,416,807 confirmed cases, 851,102 reported deaths, and 17,724,602 recovered individuals worldwide as of August 31, 2020 (Alsharif & Qurashi, 2021).

# Background

This SARS-COV-2 virus causes numerous diseases from normal headaches/colds or fever including limb and joint pains to more severe SARS (acute respiratory syndrome) and MERS (middle east respiratory syndrome).The spread of the coronavirus is through droplets/aerosols (Centers for disease control and prevention, 2020). As with any other virus, even the SARS-CoV-2 virus also changes its structure constantly making it difficult to find a cure. A genomic surveillance process has been performed to continuously monitor the change of the spike protein virus structure. Due to spike proteins, the crown-like appearance of the coronavirus resulted in the term “Corona”. “The spike proteins are a

crucial part of the SARS-CoV-2 virus, as they are responsible for attaching to human cells, enabling the virus to infect and replicate within the cell, and spread to other cells. Targeting these spike proteins can provide protection against the virus by triggering the production of antibodies that recognize and neutralize them. This is why genomic surveillance is important as it allows for the detection of mutations in the virus's spike proteins, which can affect the efficacy of existing vaccines and therapeutics.” (Centers for disease control and prevention, 2020).

Genetic changes occur over time, by forming the genetic lineages over time, to the covid specific virus. The severity of the illness is determined by how quickly the virus changes its structure and spreads, as well as the effectiveness of antiviral treatments (Centers for disease control and Prevention, 2020). The diagnosis of COVID-19 is a challenging task for most physicians and radiologists as the virus is changing its structure constantly and is like other viral pneumonia when investigated in conventional scans. The integration of artificial intelligence may be helpful in differentiating covid virus from other pneumonia diseases.

There isn't a cure or treatment or vaccine for the coronavirus COVID-19, quarantining people and isolating symptomatic as well as suspected or infected patients is one of the best ways to control the spread of the disease. It is generally considered that the rapid real-time PCR is considered an apt and appropriate tool for making the diagnosis of COVID–19. The low sensitivity of the test, the occurrences of the false positive rate, and the laboratory setting requirements being strict could lead to the correct diagnosis being delayed. At the initial stage of the disease, the only tool which is dependable is computed tomography scans (Alsharif & Qurashi, 2021). Alsharif & Qurashi (2021) stated that “When compared to MERS, the Middle East Respiratory Syndrome, and the severe acute respiratory syndrome, the mortality rate of COVID-19 is very lower.

# COVID-19 data-tracking project:

The topic which we choose is regarding the COVID-19 data tracking and analysis performed for each state’s data, for which we need to perform various methods**.** Through the Data collection method, this project used the data gathered from a variety of websites, including those run by the government, hospitals, news organizations, or social media platforms. Data on confirmed cases, fatalities, recoveries, hospitalizations, testing rates, immunization rates, and other pertinent statistics could be included. To assure the correctness of the data, the technique of data collecting should be dependable and constant. Next, through the process of data cleaning, we made sure that the data is correct and full; the data may need to be cleaned after it has been gathered. Duplicate data should be eliminated, errors should be fixed, and blank fields should be filled up. Imputation techniques, such as mean imputation or regression imputation, could be used to fill in any missing values in the data.

Every data analysis project must begin with data cleaning. Ensuring a dataset is accurate, dependable, and suitable for analysis involves the process of discovering and fixing bugs, irregularities, and incomplete data.

We will go over various COVID-19 data-cleaning techniques:

**Eliminating duplicates and error correction:** Duplicates can appear when the same information is entered more than once or when it appears in several data sources and can be done by deleting outliers and substituting wrong numbers with the proper ones. Inputting missing data can be one by mean imputation, median imputation, mode imputation, and regression imputation.

**Data storage strategy**: To accommodate the volume and complexity of the information, the cleaned data should be kept in a data repository or database. The database should be set up so that data analysis and retrieval are simple. Also, the storage technique should be safe and only available to authenticated users.

**Data analysis techniques:** The COVID-19 data can be analyzed using a wide range of statistical techniques. For instance, time series analysis can be used to monitor the pandemic's development over time. Regression analysis might be utilized to find variables linked to increased rates of COVID-19 cases or deaths. With respect to comparable COVID-19 trends, regions or states can be grouped using cluster analysis. The research questions and the type of data will determine the analysis strategy to use.

**Descriptive statistics:** A technique for summarizing and explaining a dataset's properties is descriptive statistics. It includes statistics like median, variance, mode, mean, ranges, and standard deviation. By condensing the important parameters, such as confirmed cases, deaths, recoveries, hospitalizations, testing rates, and vaccination rates, descriptive statistics can be utilized to interpret COVID-19 data.

**Regression analysis** is a technique for examining the relationship between multiple variables at once. It involves specifying the variables that are both independent and dependent and applying statistical methods to measure the link between them. By identifying variables like gender, and race that are connected to greater rates of COVID-19 cases or fatalities, regression analysis can be used to examine COVID-19 data.

**Cluster analysis:** The COVID-19 data can be grouped and analyzed using cluster analysis, which classifies regions or nations according to common COVID-19 trends like infection rates, mortality rates, or immunization rates.

**Geographic mapping**: It is a technique for displaying information on a map. It entails locating data points on a map and representing the data with colors or symbols. By displaying the spatial distribution of COVID-19 cases, fatalities, recoveries, hospitalizations, testing rates, or immunization rates across various regions or states, geographic mapping can be used to study COVID-19 data.

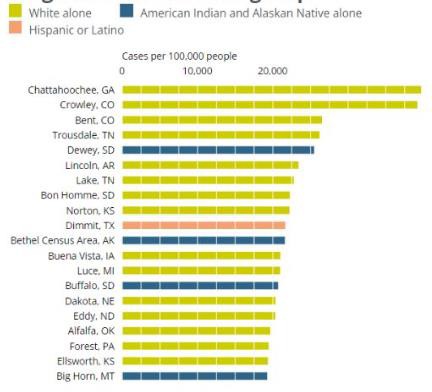
**Data visualization methods**: Data visualization is a crucial component of every data monitoring effort since it makes the data understandable to all stakeholders. Bar charts, line charts, scatterplots, heat maps, and geographical maps are a few examples of visualization techniques. The type of data being examined, and the intended audience will determine the visualization technique to be used.

**Line charts:** Line charts can be used to illustrate how a virus has evolved with time. These graphs typically depict the number of incidents or fatalities over time and aid in highlighting peak times and spotting trends.

**Bar charts:** Bar charts can be used to compare the number of incidents or fatalities in various areas or nations. These graphs aid in determining which states or regions have been afflicted by the pandemic the worst.



**Figure 1:** Representing the effect of COVID-19 on race.



**Figure 2:** Countries representing the top 20, highest death rates within the country, of various states.

**Heat maps:** Heat maps are effective for displaying the concentration of COVID-19 instances in a certain area. These maps make it simpler to locate the hotspots by using colors to show the density of instances in various places.

**Scatter plots:** The link between two variables, such as the number of cases and the number of fatalities, can be demonstrated using scatter plots. With the aid of these graphs, correlations, and patterns could be used to estimate the future trends of the illness.

**Reporting approach:** The outcomes of the data analysis should be presented in a straightforward and productive way. This can include generating a report, a dashboard, or a presentation that highlights the main outcomes of the project. The report should be written specifically for the target audience and should focus on the most significant learnings from the data analysis.

Overall, a COVID-19 data tracking project should include methods for data collection, cleaning, storage, analysis, visualization, and reporting. The methods chosen will depend on the specific objectives of the project and the analyzed data.

The Covid-19 tracking project, as mentioned earlier, collects and analyzes data on the Covid pandemic worldwide. The project compiles a global dataset on Covid-19 using numerous data sources, including government websites, health organizations, and press sources.

O’Hearn et al. (2021) explain the data collection procedure as the method for cleaning, standardizing, and organizing the data. The dataset contains geographic information, as well as

information on the number of confirmed cases, deaths, and death recoveries. The authors also explain how the data is analyzed using statistical methods to spot trends and patterns in the virus's spread (O’Hearn et al., 2021). The methods used in the analysis are time series forecasting, clustering, and correlation analysis. Additionally, the authors used the data from the Covid-19 tracking project and concluded that obesity, hypertension, and diabetes are all closely linked to Covid-19 hospitalizations. This data can identify groups at higher risk of severe outcomes and prioritize prevention and treatment efforts for people suffering from these conditions (O’Hearn et al., 2021).

The article by Xie et al. (2020) describes a time series modeling approach for forecasting the effect of the Covid-19 outbreak on hospitalizations and intensive care unit admissions in the United States. The authors estimate the parameters of their models using freely available statistics on Covid-19 cases, hospitalizations, and deaths. The authors created two models: an SEIR model (susceptible-exposed-infectious-recovered) and a hybrid model that blends the SEIR model with a logistic regression model. The models are used to predict the rate of hospitalizations and ICU admissions over a four-week period. According to the findings of the analysis, the Covid-19 outbreak will result in a significant rise in hospitalizations and ICU admissions in the United States (Xie et al., 2020).

Using an open-access dashboard Hu et al. (2021) describe a statistical technique for determining end dates for Covid-19 outbreaks in various regions. The authors estimate the number of new cases and deaths per day in multiple areas using publicly available data on Covid-19 cases and deaths from the Covid-19 tracking project (Hu et al., 2021).

The authors used a change point analysis technique to determine the inflection point of the regular case and death curves that correlates to the outbreak's end. The technique is based on a Bayesian framework and estimates the change point using a Markov Chain Monte Carlo (MCMC) algorithm (Hu et al., 2021).

The results indicate that the method can accurately predict when the outbreak will end in these areas. The research also investigates the effect of various factors on the outbreak's end date, such as the extent of intervention, the duration of the intervention, and the initial degree of severity of the disease outbreak (Hu et al., 2021). Bardus et al. (2022) investigated the data management and privacy practices of contact-tracing apps. The authors used the data from the Covid-19 tracking project and discussed the importance of contact tracing in controlling the spread of Covid-19, as well as the role of contact-tracing apps.

The study's results are presented in the article, which shows a wide range of approaches to data management and privacy among the various contact-tracing apps. While some apps gather little data and prioritize user privacy, others collect more sensitive information and offer fewer privacy safeguards. The paper emphasizes the importance of clear and transparent data management and privacy policies in contact-tracing apps, as well as user education and consent (Bardus et al., 2022).

# Methods:

First the data is imported to the MySql workbench, to perform and excecute various functions, tables were created to have the data within the specified table, having the approprate table names and associated column.

tructured Query Language) is a powerful tool for data analysis. Some of the SQL methods for data analysis were:

**Filtering data:** For data filtering, SQL's SELECT statement can be used to filter data by specific criteria. For example, you can use the WHERE clause to filter the data for a specific date range or for a specific state.

For instance, the SQL code for showing the numbers of the total positive cases taken within the year 2020 year was presented below.

**Aggregating data:** SQL's aggregate functions, such as COUNT (), SUM (), AVG (), MIN (), and MAX (), can be used to calculate statistics on the COVID-19 data. For example, you can use the SUM () function to calculate the total number of confirmed cases in a state.

**Joining tables**: COVID-19 data is often stored in multiple tables. SQL's JOIN statement can be used to combine data from multiple tables. For example, you can use the JOIN statement to combine COVID-19 data with demographic data to analyze the data gathered from different states.

An inner join returns the records when there is any some common element between both the tables, while a left join returns all records from the left table and the similar type of records sets from the right table, and a right join does the opposite.

**Grouping data:** SQL's GROUP BY statement can be used to group data by specific criteria. For example, you can use the GROUP BY statement to group COVID-19 data by state and calculate the total number of confirmed cases for each state.

**Window functions:** SQL's window functions can be used to perform calculations on a subset of data, such as a rolling average. For example, you can use the AVG () window function to calculate the 7-day rolling average of new COVID-19 cases in a state.

Overall, SQL is a powerful tool for COVID-19 data analysis, and these methods can be used to gain insights into the pandemic and its impact on different regions and populations.

To conduct our COVID-19 tracking project, we utilized two datasets: one containing State data and another indicating the rise in a number of cases. The State data cases dataset contains details about death, positive, negative, and recovered cases data for every state in the United States on a

given date. The number of cases rise dataset contains information on the increase in death and positive cases for each state. To calculate statistics such as case rates and death rates based on the dates and the total rise of cases, we joined these datasets based on the common connection, the state column.

The columns included within the specific tables were as follows, Source Data ERD:

State data Table:

* State
* date
* Death
* Recovered
* positive Number of cases rise Table:
* State
* Date
* Death increase
* Positive increase

Updated Data ERD:

State Data & Number of cases rise Table:

* State
* Date
* Deaths
* positive
* Recovered
* Death increase
* Positive increase
* Total test results
* Total test result increase

After updating the data, the resulting table will contain all the columns from the source data tables, in addition to two new columns: "total test results" and "total test result increase."

The ERD (entity-relationship diagram) for the updated data table will show all the columns mentioned above and will be linked through the "State" column.

Later,

Here are some methods which we followed within our COVID-19 data tracking project:

**Data collection method:** The COVID-19 data tracking project could involve collecting data from various sources, such as official government websites, healthcare organizations, news outlets, or social media platforms. The data could include information on confirmed cases, deaths, recoveries, hospitalizations, testing rates, vaccination rates, and other relevant metrics. The method of data collection should be reliable and consistent to ensure the accuracy of the data. We collected our data from https://covidtracking.com/data/.

**Data cleaning:** After the data is collected, it may need to be cleaned to ensure that it is accurate and complete. This could involve removing duplicates, correcting errors, or filling in missing data. For example, if there are missing values in the data, they could be filled in using imputation techniques such as mean imputation or regression imputation.

**Data storage:** The cleaned data should be stored in a database or data warehouse making it available to the public through authorized sources, designed in such a way as to accommodate the size and complexity of the data. The database should be organized in a way that makes it easy to query and analyze the data. The storage method should also be secure and accessible to authorized users, for future research purposes. The data which we stored today could be beneficial for upcoming future scientific research cases.

**Data analysis:** Time series analysis could be done for forecasting the data using within the specified periods of time. There are various statistical methods that can be used to analyze COVID- 19 data. Time series analysis, for example, can be used to track the progression of the pandemic

over time. Regression analysis can be used to identify factors that are associated with higher rates of COVID-19 cases or deaths. Cluster analysis can be used to group regions or states based on similar COVID-19 trends. The choice of analysis method will depend on the research questions and the nature of the data.

**Visualization method**: Data visualization is an important part of any data tracking project as it allows the data to be easily interpreted and communicated. Visualization methods could include bar charts, line charts, scatterplots, heat maps, or geographic maps. The choice of visualization method will depend on the type of data being analyzed and the audience for the project.

**Reporting method:** The results of the data analysis should be reported in a clear and concise manner. This could involve creating a report, a dashboard, or a presentation that summarizes the key findings of the project. The report should be tailored to the audience and should highlight the most important insights that were gained from the data analysis.

Overall, a COVID-19 data tracking project should include methods for data collection, cleaning, storage, analysis, visualization, and reporting. The methods chosen will depend on our project’s specific objective and aim, on what factors to focus more on, and what we could neglect by not focusing much on. Identifying and correcting errors, inconsistencies, and missing data in a dataset to ensure that it is accurate, reliable, and ready for analysis.

Also, there were various data cleaning methods that can be used to clean COVID-19 data.

Data cleaning is an essential step in any data analysis project. It involves the process of identifying and correcting errors, inconsistencies, and missing data in a dataset to ensure that it is accurate, reliable, and ready for analysis.

Various data cleaning methods that can be used to clean COVID-19 data are as follows:

**Removing duplicates:** Duplicates can occur when the same data is entered multiple times or when different data sources contain the same information. In general, to remove duplicates, we can use the "drop\_duplicates ()" function in Python or the "Remove Duplicates" feature in Excel. Within SQL we can remove the duplicates by using the “DISTINCT” function.

**Correcting errors:** Errors can occur when data is entered incorrectly or when there is a mistake in the data source. To correct errors, we can use various techniques such as replacing incorrect values with the correct ones, inputting missing values, or removing outliers.

**Imputing missing data:** Missing data can occur when data is not recorded or when data is incomplete. To impute missing data, we can use various techniques such as mean imputation, median imputation, mode imputation, or regression imputation. Mean imputation involves replacing missing values with the mean value of the variable, while regression imputation involves predicting missing values using a regression model.

**Handling outliers:** Outliers are extreme values that are significantly different from other values in the dataset. Outliers can occur due to data entry errors or due to genuine extreme values. To handle outliers, we can either remove them or replace them with a more reasonable value based on the context of the dataset.

**Filtering outliers:** Filtering of the outliers could be done, if we intend to limit the values within a specific range, for example, to limit them between 1 to 100, which that later forwarded to the trimming of data, and replacing the outliers.

**Resolving inconsistencies:** Inconsistencies can occur when data is entered in different formats or when data sources use different units of measurement. To resolve inconsistencies, we can standardize the data by converting it to a common format or unit of measurement.

**Addressing data formatting issues:** Data formatting issues can occur when data is entered in the wrong format or when data is not consistent across the dataset. To address formatting issues, we can use various techniques such as text-to-columns in Excel, regular expressions in Python, or data type conversion. In SQL data formatting could be corrected by changing the case, trimming the whitespace, and concatenation of strings.

**Ensuring data quality:** Data quality is a measure of the accuracy, completeness, and consistency of the data. To ensure data quality, we can perform data validation checks, such as checking for data integrity, accuracy, and completeness.

The database methods we used for our project include MySQL Workbench & Tableau along with its collaboration for better and modified visualizations.

**MySQL Workbench:** MySQL workbench is one of the useful tools which we could use for this paper and in this tracking project. Based on the research questions we had chosen, and performed numerous queries, to showcase and calculate the accurate data. We opted for SQL workbench for our project and analysis because it is easy with the workbench to establish a connection with Tableau for importing the larger datasets and for better visualizations. With the creation of tables

within the workbench, we could also create a faster connection for the ERD diagrams and formulation of database management systems.

**Tableau:** With the use of Tableau, we could use large data of huge amounts with ease and feasibility. The dataset of the COVID-19 tracking project which we choose contains numerical data types. We created separate different worksheets to understand the trend and to analyze the workflow, we created separate worksheets for each scenario and question.

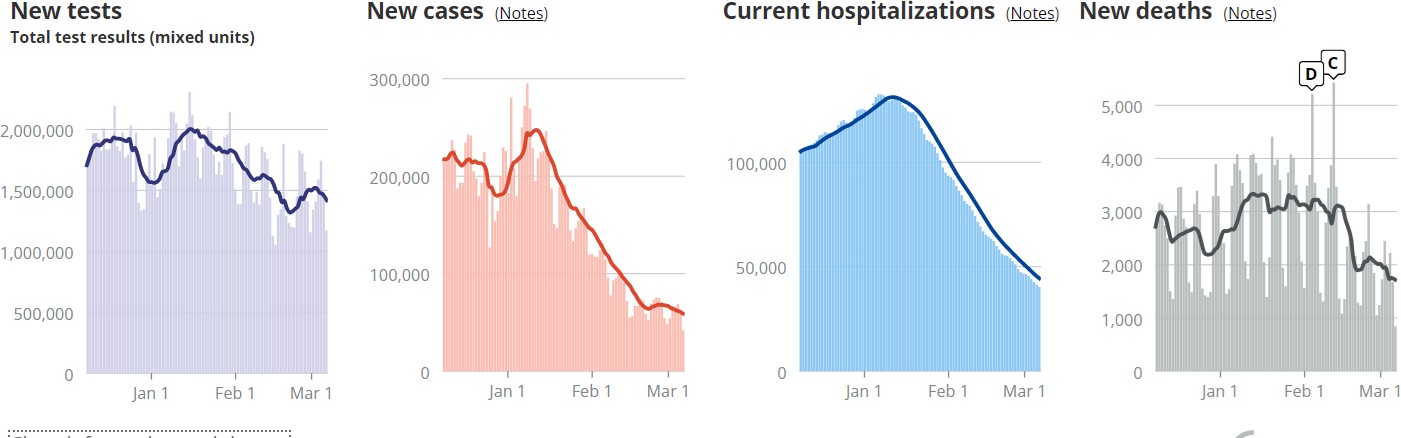
# Results

Importing data into Tableau.

Graphical user interface, application

Description automatically generated

Visualization showing the Overview of COVID-19 data nationally.



0

# 

# 

# Discussions

Special observations –

In general, the COVID-19 tracking project is an important effort to monitor and analyze the pandemic's spread and impact.

Here are some potential special discussions for the project:

**Data accuracy and consistency:** It is critical to ensure that the data collected and analyzed is accurate and consistent across regions and time periods. This can be accomplished by implementing standardized data collection and reporting methods, as well as performing regular data checks and validations.

**Data visualization** is a powerful tool for communicating complex information to various stakeholders. It can aid in the identification of patterns and trends in data and can be used to inform decision-making. Special discussions on the use of effective data visualization techniques can help to improve the COVID-19 tracking project's impact and utility.

**Predictive modeling** can be used to forecast the spread and impact of a pandemic, as well as to inform planning and resource allocation efforts. Special discussions on the use of various predictive modeling techniques, such as machine learning and statistical modeling, can help to improve the COVID-19 tracking project's accuracy and utility.

**Collaboration across disciplines:** The COVID-19 pandemic is a complex and multifaceted issue that necessitates interdisciplinary cooperation. Special discussions on collaboration among

various stakeholders, such as public health experts, epidemiologists, data analysts, and policymakers, can help to ensure that the project is carried out holistically and effectively.

To conclude, we had some fun understanding the chosen dataset while designing our database. One interesting finding was that many of the variables in the dataset were

numerical. Another intriguing finding in the dataset was that we were able to make columns in the dataset a primary and foreign key.

# Limitations or Challenges:

There were some difficulties even though we worked well as a team. It took a significant amount of time for all team members to conduct research on a project topic at the start of the project. After the project proposal submission, our team faced difficulty with the team members attending the regular meetings. Despite all the challenges faced, we as a team communicated and talked often using varied platforms, like Zoom, and kept informed of the project work progress. Because many of our chosen variables were numerical, selecting the foreign and primary keys while designing the ERD was a challenge. However, we examined the data and discovered the attributes that are relevant to our goal and built a strong model.

Apart from that, our group also faced numerous other challenges after selecting our dataset and that includes:

**Improper standard data:** There is lack of standardized data across different countries as COVID-19 pandemic affected countries differently. This resulted us difficulty

in making comparisons of data across countries and drawing accurate conclusions.

**Incomplete data:** Incomplete data landed us difficulty in making the conclusion because of the pandemic's rapid evolution, the efforts to collect the data have been inconsistent.

**Data privacy:** The data collection and analysis are essential to predict the outcome but due to sharing of health information globally analyzing the data has been a challenging task while maintaining privacy and confidentiality.

**Limited access:** Access to data is restricted to certain organizations and researchers, making it difficult for others to contribute to the analysis and understanding of the pandemic. But in our dataset, we had full access to the data which made our part easy.

These are just a few examples of the COVID-19 project's limitations. It is critical that researchers and analysts are aware of these limitations and take them into account when analyzing and interpreting data.

# Conclusion

Using SQL queries and data visualization methods we were able to analyze the COVID- 19 tracking data based on values taken from different states in the United States. We created an ERD to better understand the relationship between entities and attributes. Based on our research questions, we used various SQL commands, like the GROUP BY, SUM, and AVG functions to formulate the conclusions regarding the data selected, by the creation of tables according to our CSV sheet. The clinical database management system selected is robust and can be used to solve the selected research questions and identify the factors which could potentially have a significant impact on COVID-19 in different states in the USA.

# References

Alsharif, W., & Qurashi, A. (2021, May 27). Effectiveness of COVID-19 diagnosis and management tools: A review. *Radiography (London, England : 1995)*, *27*(2), 682–687. https://doi.org/10.1016/j.radi.2020.09.010

Bardus, M., Al Daccache, M., Maalouf, N., Al Sarih, R., & Elhajj, I. H. (2022). Data Management and privacy policy of COVID-19 contact-tracing apps: Systematic review and content analysis. *JMIR MHealth and UHealth*, *10*(7). https://doi.org/10.2196/35195

Centers for disease control and prevention. (2020, February 11). *COVID-19 and your health*. https:/[/www](http://www.cdc.gov/coronavirus/2019-ncov/your-health/about-covid-19/basics-covid-).[cdc.gov/coronavirus/2019-ncov/your-health/about-covid-19/basics-covid-](http://www.cdc.gov/coronavirus/2019-ncov/your-health/about-covid-19/basics-covid-)

19.html

Hu, S., Xiong, C., Yang, M., Younes, H., Luo, W., & Zhang, L. (2021). A big-data driven approach to analyzing and modeling human mobility trend under non-pharmaceutical interventions during COVID-19 pandemic. *Transportation Research Part C: Emerging Technologies*, *124*, 102955. https://doi.org/10.1016/j.trc.2020.102955

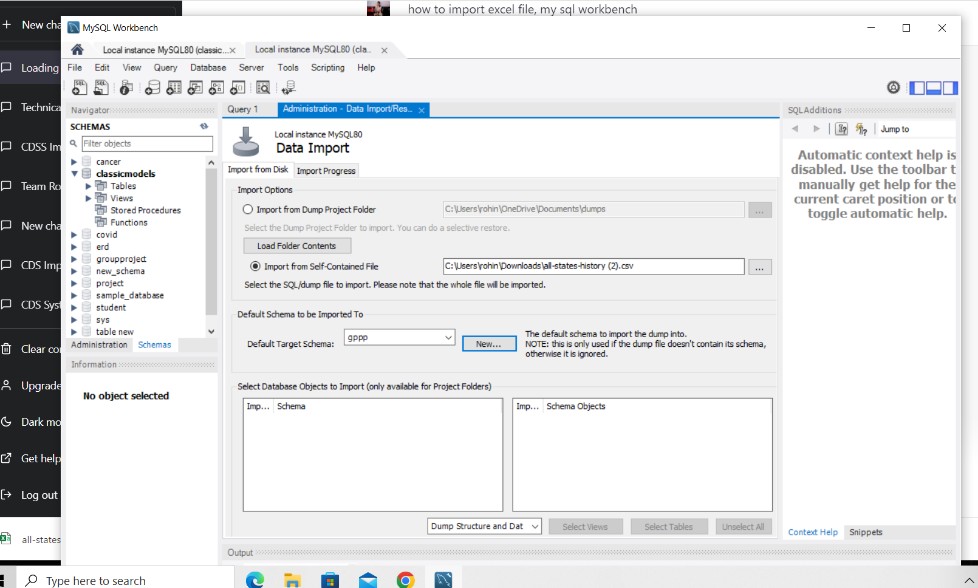
O’Hearn, M., Liu, J., Cudhea, F., Micha, R., & Mozaffarian, D. (2021). Coronavirus disease 2019 hospitalizations attributable to cardiometabolic conditions in the United States: A comparative risk assessment analysis. *Journal of the American Heart Association*,

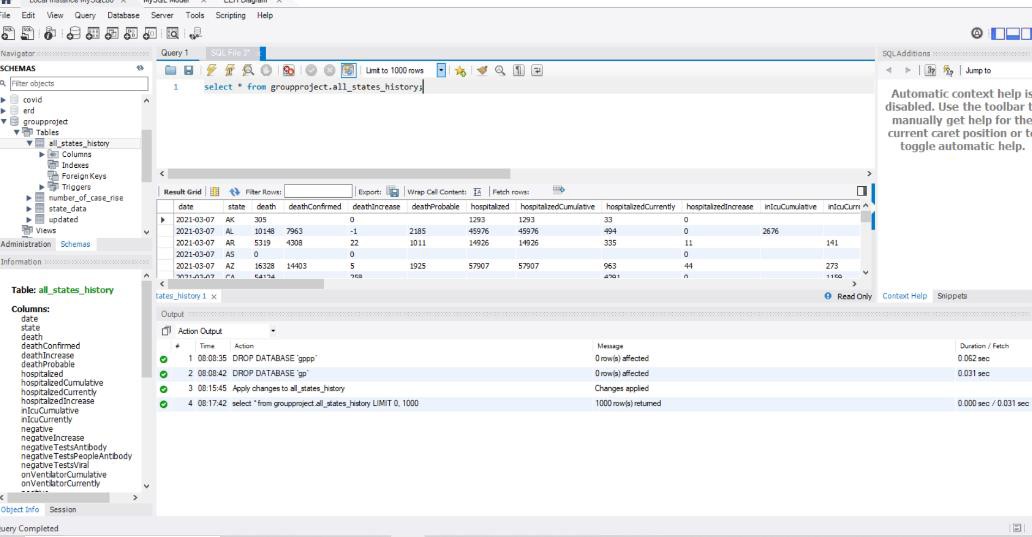
*10*(5). https://doi.org/10.1161/jaha.120.019259

Xie, J., Hungerford, D., Chen, H., Abrams, S. T., Li, S., Wang, G., Wang, Y., Kang, H., Bonnett,

L., Zheng, R., Li, X., Tong, Z., Du, B., Qiu, H., & Toh, C.-H. (2020). Development and external validation of a prognostic multivariable model on admission for hospitalized patients with COVID-19. *MedRxiv*, 2020-03. https://doi.org/10.1101/2020.03.28.

**APPENDIX:**

Importing data to MySql workbench with table names and columns.



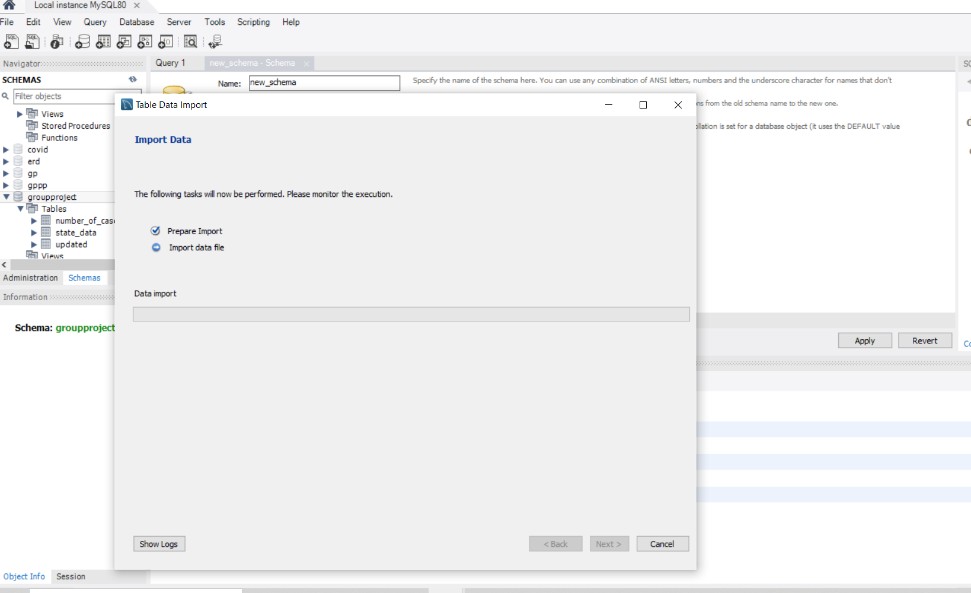
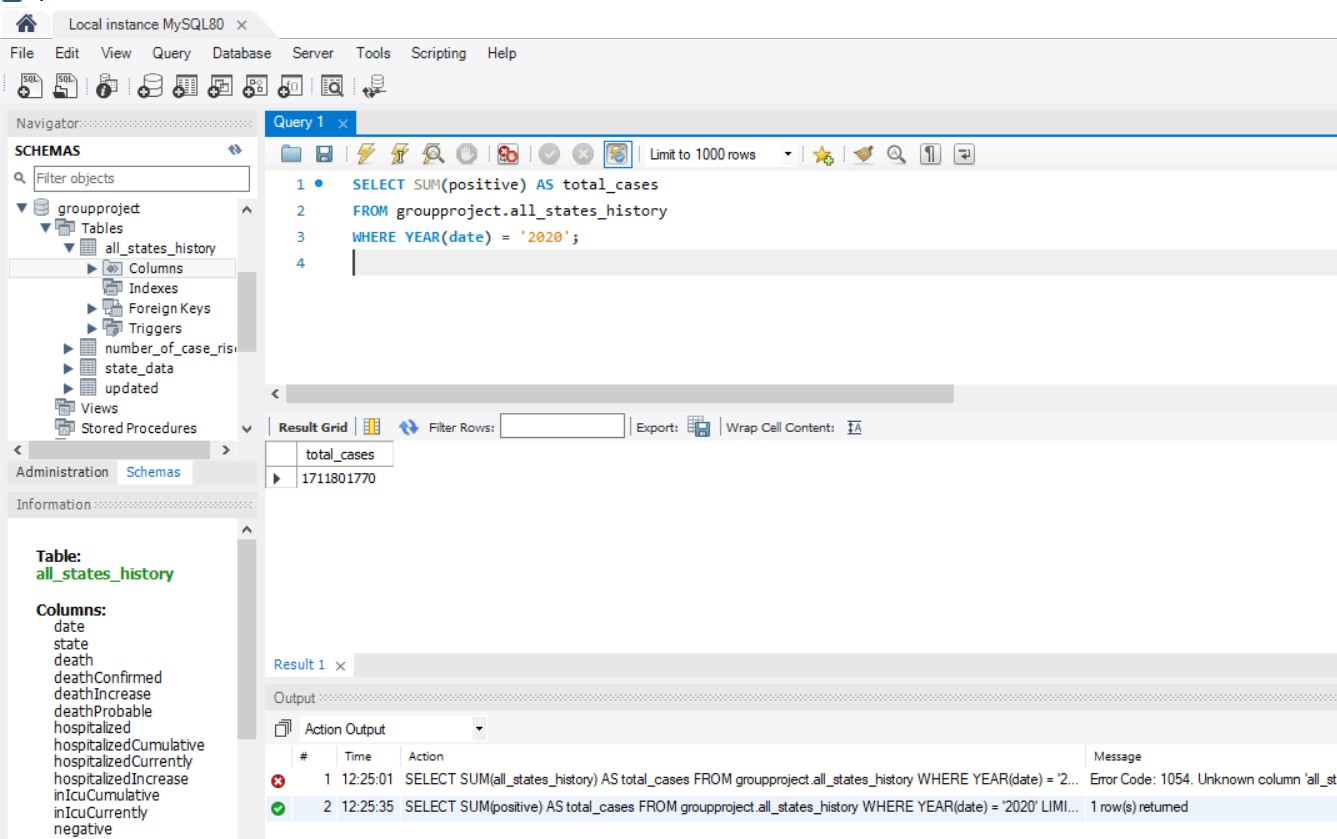
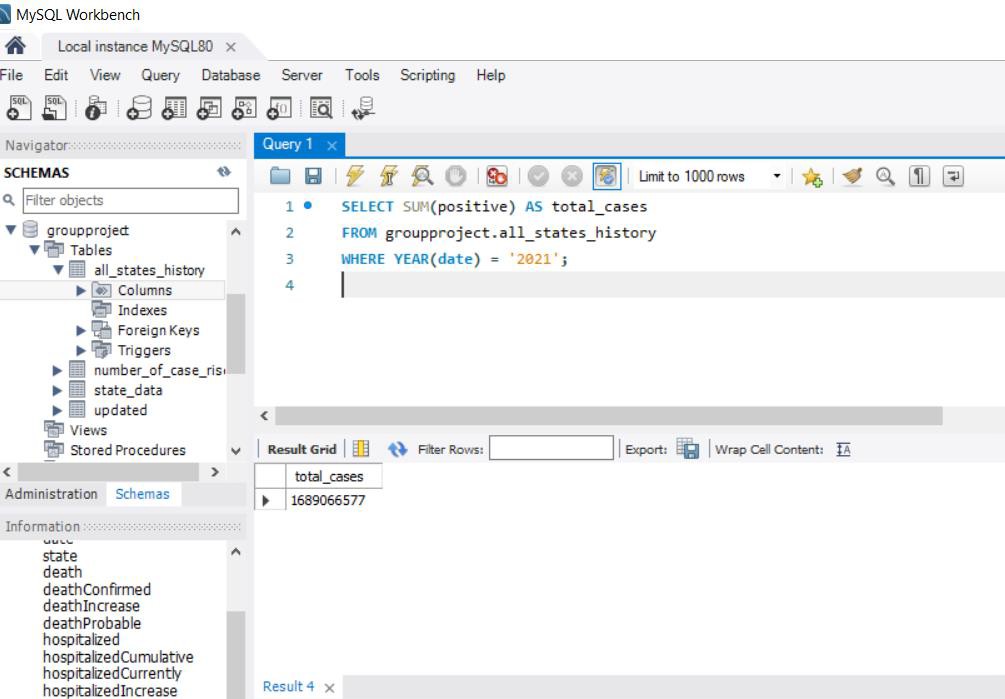


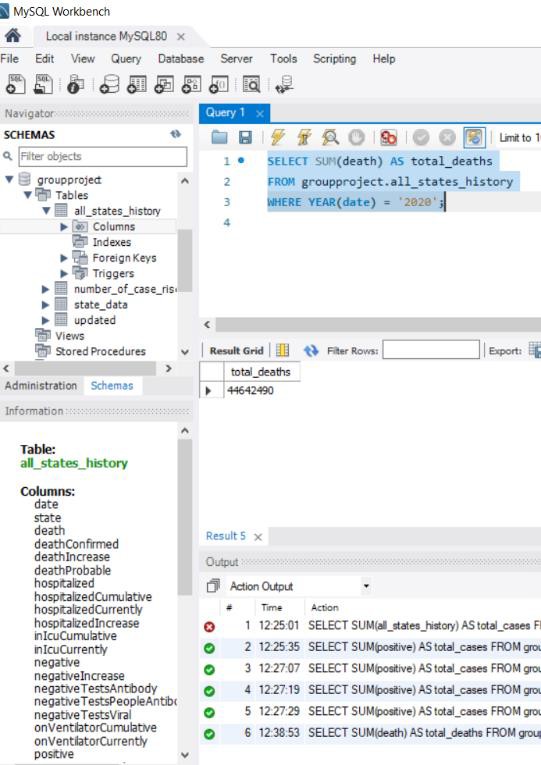
Figure 3 : Image showing the import completion process



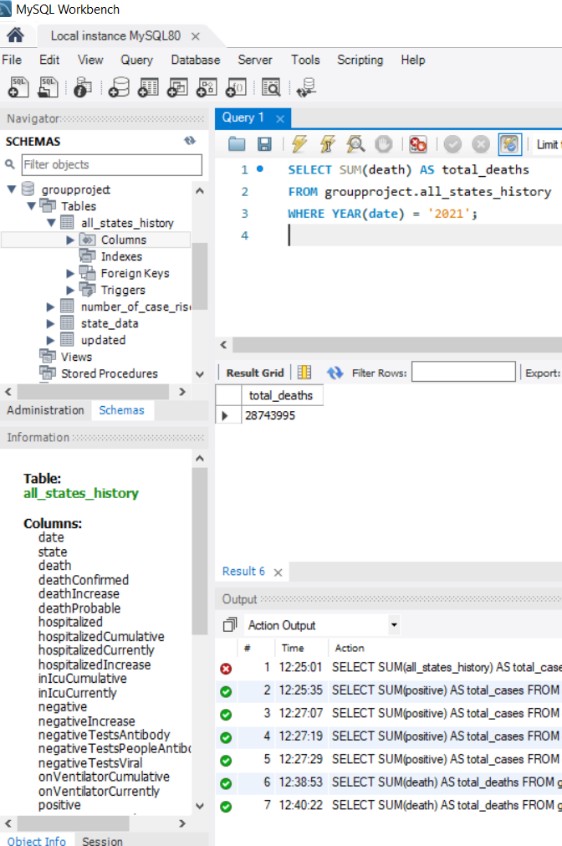
**Figure 4:** Code executed to get the total number of positive cases from 2020.



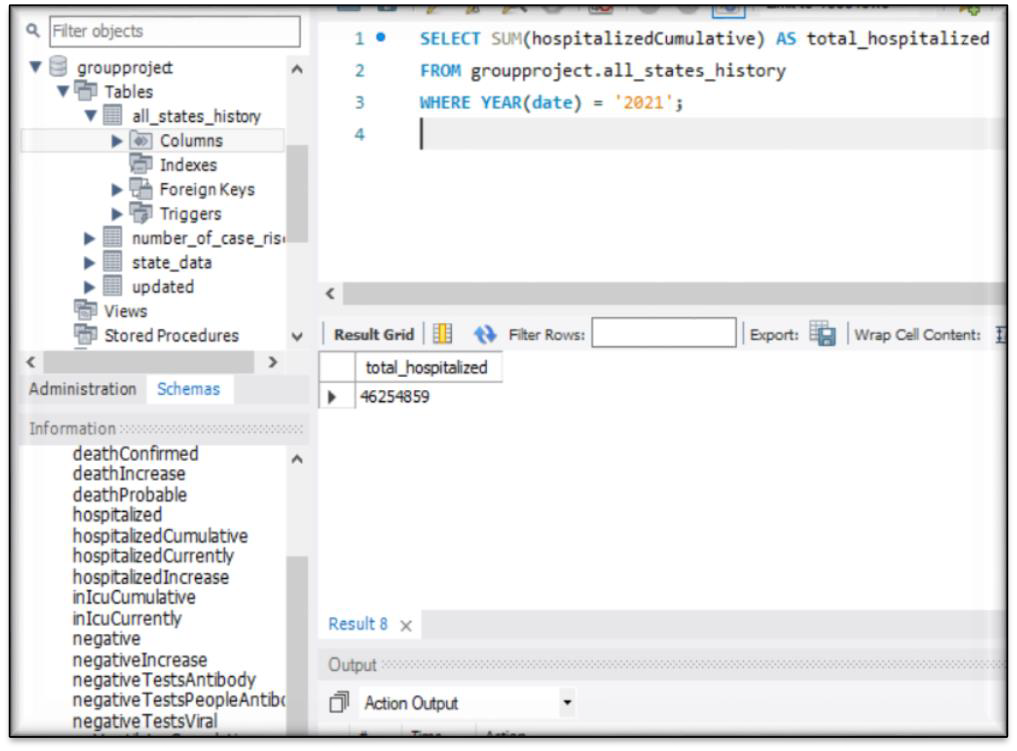
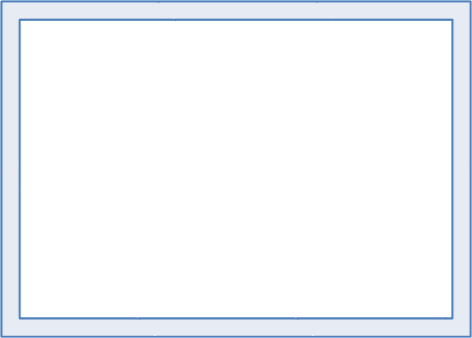
**Figure 5:** Code executed to show the number of positive cases from the year 2021.



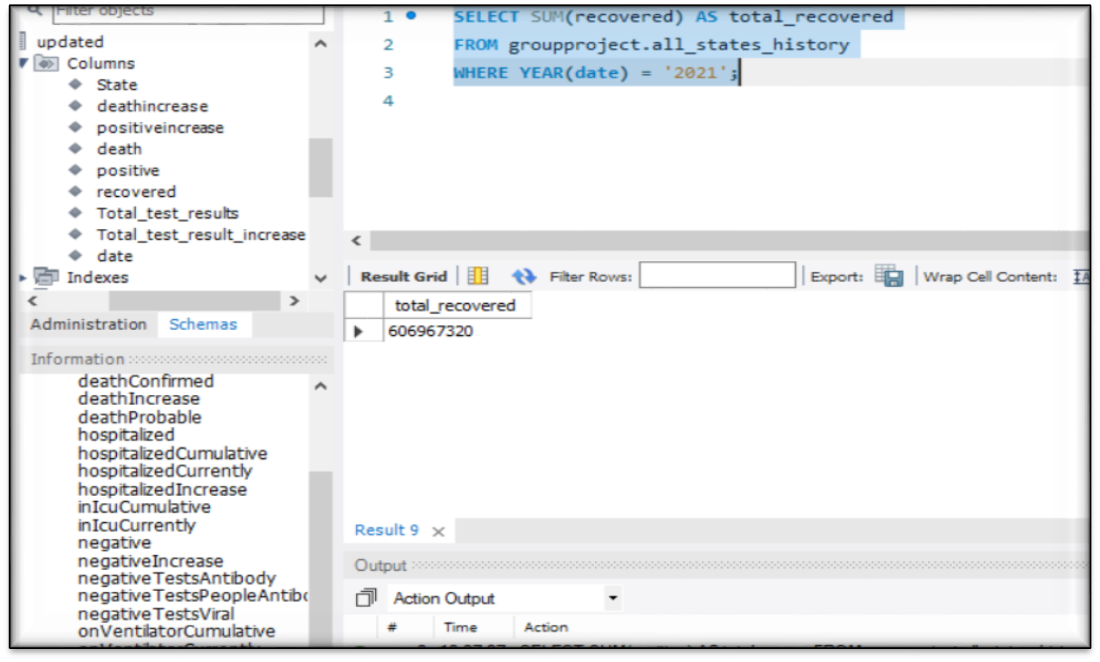
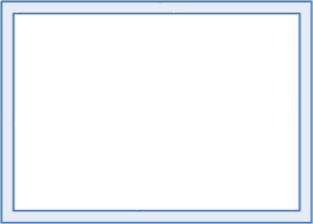
**Figure 6:** Code showing the total number of deaths within the specific year 2020.



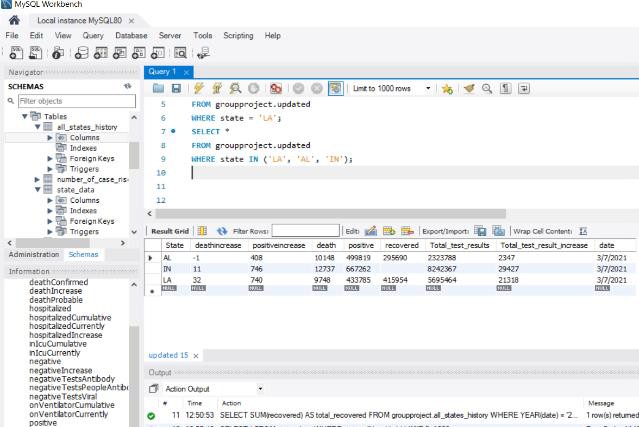
**Figure 7:** Code showing the total deaths within the specific year of 2021.



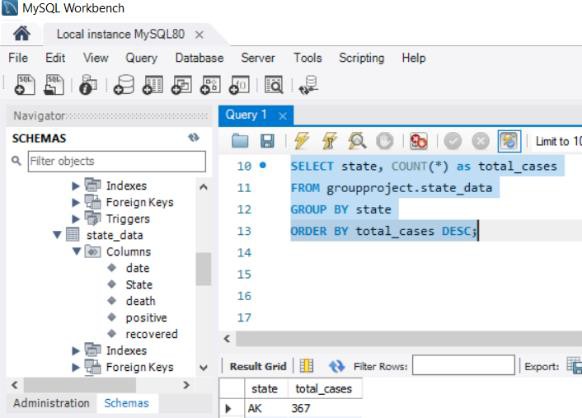
**Figure 8:** Code showing the hospitalized cases.



**Figure 9:** Code showing the recovered counts.

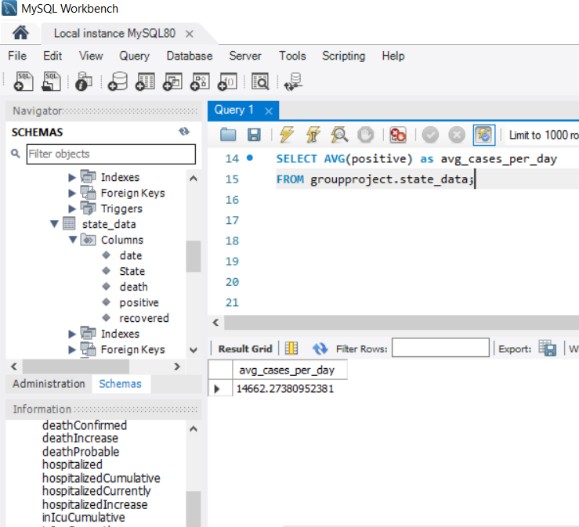


**Figure 10:** Filtering of data performed to extract the specific data from different states.

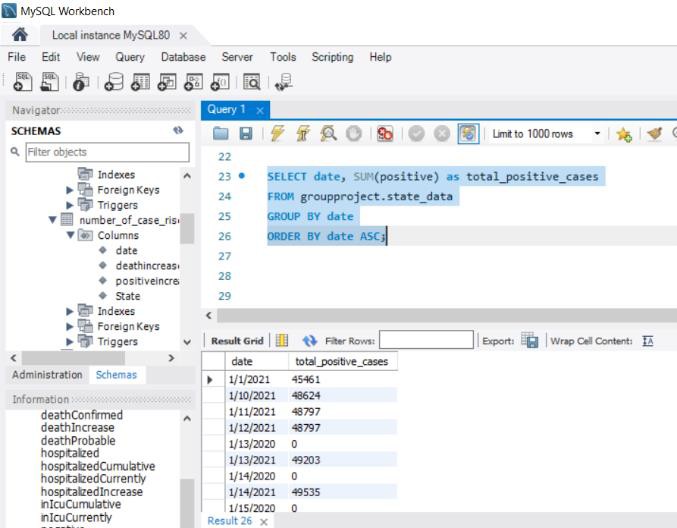


**Figure 11:** Data aggregation with the use of COUNT function.

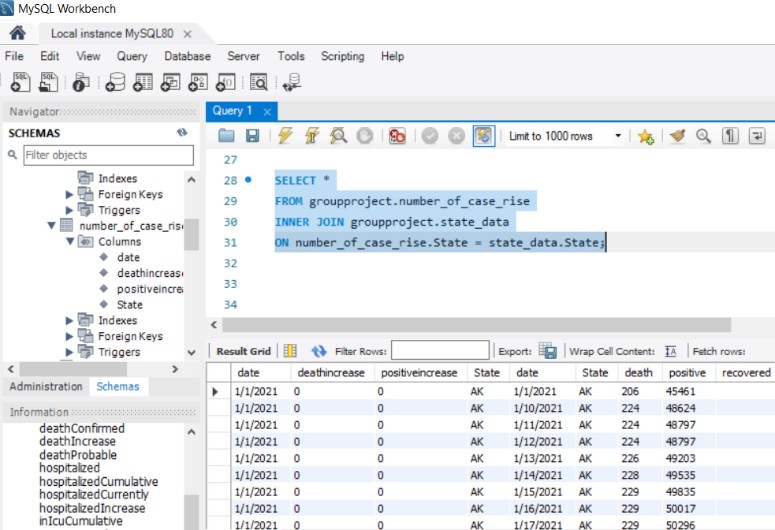
**Figure 11:** Data aggregation with the use of COUNT function.



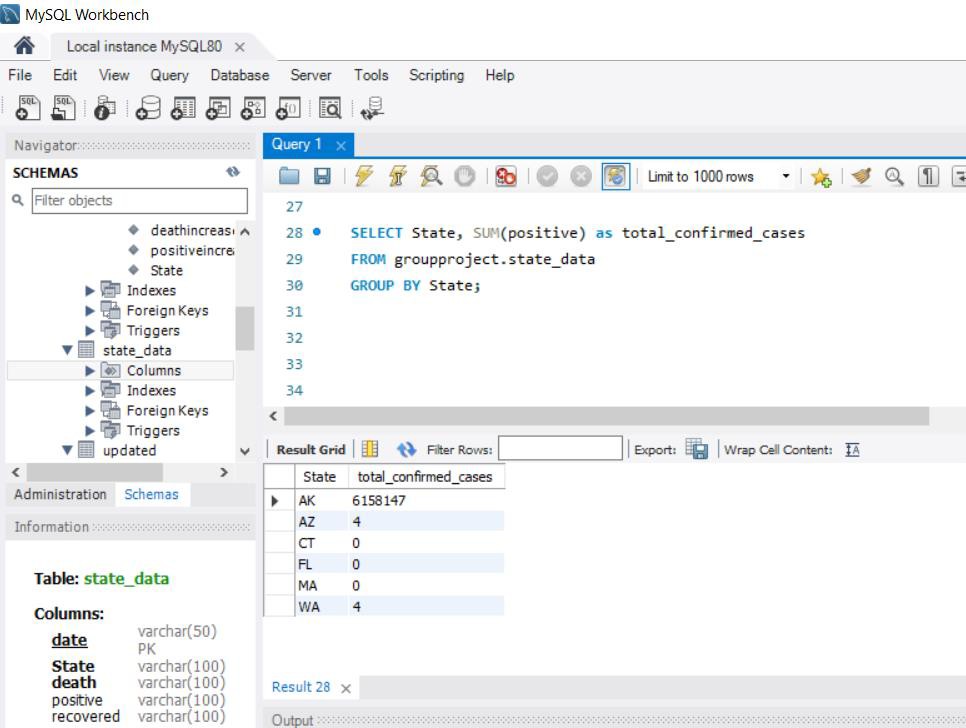
**Figure 12:** Data aggregation, with the use of AVG function.



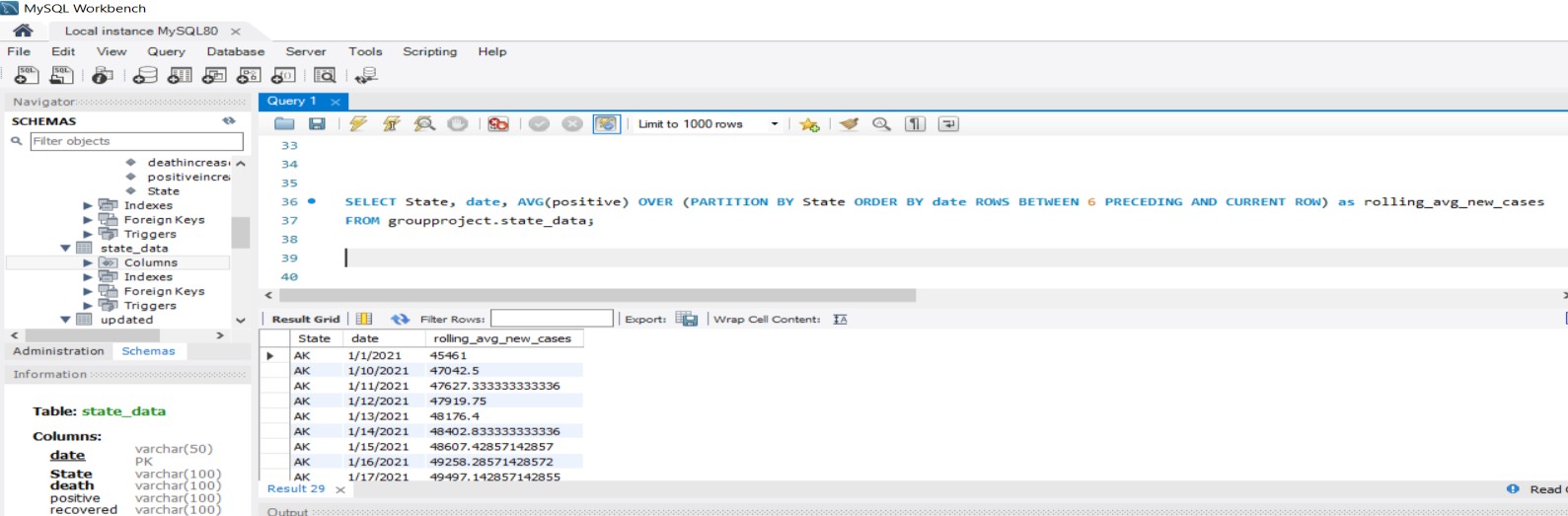
**Figure 13:** Data aggregation using the function SUM.



**Figure 14:** Joining of tables performed using the inner join.

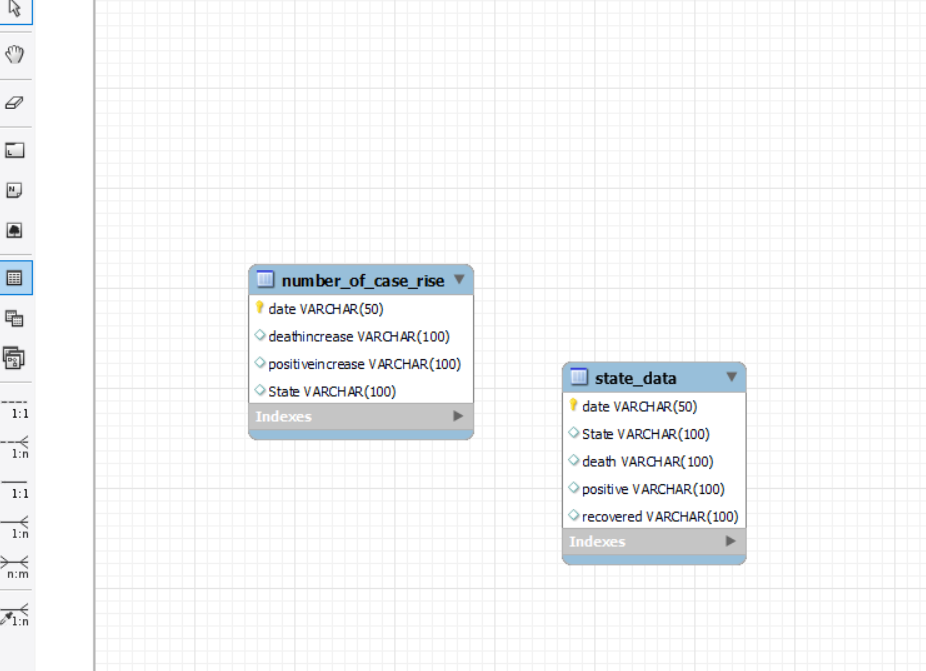


**Figure 15:** Using GROUP BY function to calculate the total confirmed cases based on the respective information gathered from the respective state.

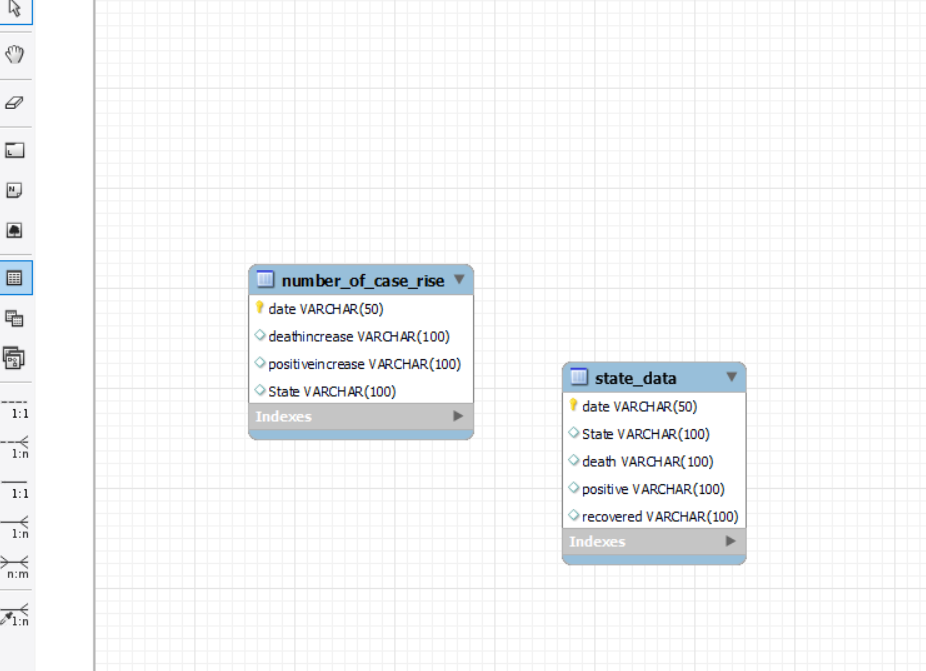


**Figure 16:** Using windows function of AVG to calculate the average of all the positive cases from the specific state.

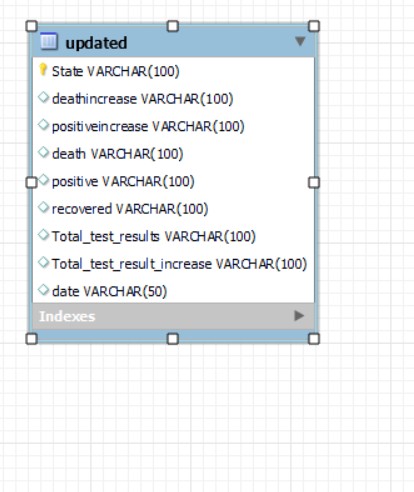
Finally, the ERD Diagrams drawn for our model were, as follows:



The comparison between the source data and the updated data: ERD for the source data could be represented as follows,



ERD for the updated data you've designed:



**Setting up the connections:**

