COLUMBIA UNIVERSITY | MAILMAN SCHOOL of PUBLIC HEALTH
BIOSTATISTICS

In-Class Assignment 16

Applying CTEs, Temporary Tables, and Window Functions to SQL Queries

Read the scenario below, then follow the instructions and answer the questions that follow. Submit your solutions in a MySQL script named **Assignment16.sql** to Canvas by the deadline listed above. Save your file frequently to avoid losing work!

Scenario:

You are part of a biostatistics team and have been asked to provide insights about confirmed COVID-19 cases throughout New York State. You've been given a dataset from Johns Hopkins University containing all confirmed cases for all counties in New York State from 1/22/20 through 11/29/2021. Write SQL queries using the techniques covered in lecture this week to answer the questions that follow.

Adding Confirmed Cases Data to MySQL:

Download the MySQL script called **covid_confirmed_ny.sql** from Canvas and run the entire script in Workbench to create a schema called 'covid' and populate a table called 'confirmed.'

The three key columns you will work with are:

admin2: represents the County name → *please always us an appropriate alias when referencing this variable!*

confirm_date: date on which the number of cases was reported

confirmed_cases: cumulative number of confirmed cases to that date

Questions:

Please answer the questions below in /*block comments*/ within your SQL script, along with your query code.

1. How many counties are reported in the dataset? In your query, display the count as **county_count.**

2. List the names of counties reporting less than 1000 confirmed cases as of 11/29/21. Give the number of confirmed cases reported.

3. List the counties from fewest to greatest number of cases as of 11/29/21 and add a dense rank to show the relative rank of each county. Display the rank as **county_rank** in your query. (No need to answer anything in a comment for this question.)

4. Using your query from question 3 as a CTE called **bottom_ten,** limit the list to any counties falling within the bottom 10 rank. Show the county name, confirmed cases, date, rank, and add the average number of confirmed cases in the bottom 10. Call the average **bottom_ten_avg**. What is the average, and what are the bottom 10 counties for confirmed cases?

In-Class Assignment 16

5. Create a temporary table called **top_counties** that (dense) ranks the NY State counties from greatest to fewest number of cases as of 11/29/21. Show the county name, date, number of confirmed cases, and the rank. Again, display the rank as **county_rank.** Use **ORDER BY confirmed_cases DESC** in the appropriate window function.

6. As with question 4, produce a list of the counties with the 10 highest counts of confirmed cases as of 11/29/21. Show the county name, date, number of cases, rank, and the overall average confirmed cases for the top 10. Call the average **top_ten_avg.** What is the average and what are the top 10 counties for confirmed cases?

7. What is the total number of cases that have been recorded in New York State as of 11/29/21? Display the total as **current_confirmed_cases.**

8. Create a temporary table called **nyc_confirmed**, which contains county name, date, and confirmed cases for all 2021 reporting dates in New York County. Include a column called **confirm_month**, which extracts the month from the confirm_date and saves it in a separate column. Also add a field that shows the day-to-day increase in confirmed cases and call the column **daily_increase**. (No need to answer any questions in a comment for this question.)

9. Display a result set that adds the average of daily_increase for each month to every record in nyc_confirmed. Call the daily average **avg_daily_increase.** What were the average day-to-day increases for each of the months reported in 2021?

10. Create a CTE called **new_case_rank**, based on nyc_confirmed, that ranks (with RANK()) days from highest to lowest daily increase in confirmed cases. Name the rank column **day_rank.** Then, retrieve the record(s) for the day(s) with the highest daily increase. Which day in 2021 so far had the highest day-to-day increase in confirmed cases in NYC? How many cases were confirmed that day?