

The scenario

You are a data analyst for your state's department of education. You're given a database containing 2 tables: naep and finance. NAEP is the National Assessment of Educational Progress for states. The naep table contains each state's average NAEP scores in math and reading for students in grades 4 and 8 for various years between 1992 and 2017. The finance table contains each state's total K-12 education revenue and expenditures for the years 1992 through 2016. You are tasked with assessing the quality of this data. You must also find useful ways to analyze it.

Data Exploration

1. Write a query that allows you to inspect the schema of the naep table.

```
SELECT *  
FROM  
    information_schema.COLUMNS  
WHERE  
    Table_name = 'naep';
```

2. Write a query that returns the first 50 records of the naep table.

```
SELECT *  
FROM  
    naep  
LIMIT 50;
```

3. Write a query that returns summary statistics for 'avg_math_4_score' by state. Make sure to sort alphabetically by state name.

```
SELECT  
    state,  
    COUNT(avg_math_4_score) AS count,  
    ROUND(AVG(avg_math_4_score),3) AS average,  
    MIN(avg_math_4_score) AS minimum,  
    MAX(avg_math_4_score) AS maximum  
FROM  
    naep  
GROUP BY state  
ORDER BY state ASC;
```

4. Write a query that alters the previous query so that it returns only the summary statistics for avg_math_4_score by state with differences in max and min values that are greater than 30.

```
SELECT
    state,
    COUNT(avg_math_4_score) AS count,
    ROUND(AVG(avg_math_4_score),3) AS average,
    MIN(avg_math_4_score) AS minimum,
    MAX(avg_math_4_score) AS maximum
FROM
    naep
GROUP BY
    state
HAVING (MAX(avg_math_4_score) - MIN(avg_math_4_score)) > 30
ORDER BY
    state ASC;
```

Analyzing the Data

5. Write a query that returns a field called 'bottom_10_states' that lists the states in the bottom 10 for 'avg_math_4_score' in the year 2000.

```
SELECT
    state AS bottom_10_states
FROM
    naep
WHERE
    year = 2000 AND
    avg_math_4_score IS NOT null (# test for non-null values)
ORDER BY
    avg_math_4_score ASC
LIMIT 10;
```

6. Write a query that calculates the average 'avg_math_4_score' rounded to the nearest 2 decimal places over all states in the year 2000.

```
SELECT
    state,
    ROUND(AVG(avg_math_4_score),2) AS average
FROM
    naep
WHERE
    year = 2000
GROUP BY
    state
ORDER BY
    state ASC;
```

7. Write a query that returns a field called 'below_average_states_y2000' that lists all states with an 'avg_math_4_score' less than the average over all states in the year 2000.

```
WITH AVR_STATES AS
    (SELECT AVG(avg_math_4_score) as AVG_STATES
     FROM naep
     WHERE year = '2000'
     GROUP BY state)
SELECT
    avg_math_4_score AS below_average_states_y2000, state
FROM
    naep, AVR_STATES
WHERE
    avg_math_4_score < AVG_STATES;
```

8. Write a query that returns a field called 'scores_missing_y2000' that lists any states with missing values in the 'avg_math_4_score' column of the naep data table for the year 2000.

```
SELECT
    state AS scores_missing_y2000
FROM
    naep
WHERE
    year = '2000' AND
    avg_math_4_score IS null;
```

9. Write a query that returns for the year 2000 the state, 'avg_math_4_score', and 'total_expenditure' from the naep table left outer joined with the finance table, using id as the key and ordered by 'total_expenditure' greatest to least. Be sure to round 'avg_math_4_score' to the nearest 2 decimal places, and then filter out NULL 'avg_math_4_scores' in order to see any correlation more clearly.

```
SELECT
    naep.state,
    ROUND(naep.avg_math_4_score, 2) as avg_math_4_score, finance.total_expenditure
FROM
    naep
LEFT OUTER JOIN
    finance ON naep.id = finance.id
WHERE
    naep.year = '2000' AND avg_math_4_score IS NOT NULL
ORDER BY
    finance.total_expenditure DESC;
```