For this exam, you'll need to use the department_of_education database.

Query prompts

Below, you'll find 9 numbered prompts. Each prompt will require you to write a SQL query. These prompts are split up into 2 distinct sections focusing on data exploration and data analysis.

Data exploration

You'll begin your analysis with the naep table. It's always a good idea to get a better understanding of your data BEFORE doing any analysis. This allows you to gather key insights before you jump into any complex operations. You'll want to know what columns are reported in your data, what the data types are for each column, and what the first few observations look like.

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

```
--1
SELECT column_name, data_type
FROM information_schema.columns
WHERE table_name = 'naep';
```

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

```
--2
SELECT *
FROM naep
LIMIT 50;
```

Another good way to understand your data is to calculate various summary statistics. Summary statistics can give you very useful information, such as where your data is centered and how spread out it is. These summary statistics include **count**, **average**, **min**, and **max** values.

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

```
--3
SELECT state, count(*) AS record_count,
        ROUND(AVG(avg_math_4_score),2) AS avg_avg_math_4_score,
        MIN(avg_math_4_score) AS min_avg_math_4_score,
        MAX(avg_math_4_score) AS max_avg_math_4_score
FROM naep
GROUP BY state
ORDER BY state;
```

When a state has a large gap between the max and min values for a score, that's a good indicator that there may be problems with the education system in that state. You decide that for avg_math_4_score, a gap of more than 30 between max and min values is probably a bad sign.

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.

```
--4
WITH stat_avg_math_4 AS
(
		SELECT state, count(*) AS record_count,
			ROUND(AVG(avg_math_4_score),2) AS avg_avg_math_4_score,
			MIN(avg_math_4_score) AS min_avg_math_4_score,
			MAX(avg_math_4_score) AS max_avg_math_4_score
		FROM naep
		GROUP BY state
		ORDER BY state
)
SELECT *
FROM stat_avg_math_4
WHERE (max_avg_math_4_score-min_avg_math_4_score) > 30;
```

Analyzing your data

Now that you've gathered key insights about your data, you're ready to do some analysis! You want to report the bottom 10 performing states for avg_math_4_score in the year 2000. You also want to report the states that scored below the average avg_math_4_score over all states in the year 2000.

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.

```
--5
SELECT state AS bottom_10_states
FROM naep
WHERE year = 2000
ORDER BY avg_math_4_score ASC
LIMIT 10;
```

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.

```
--6
SELECT ROUND(AVG(avg_math_4_score),2) AS avg_avg_math_4_score
FROM naep
WHERE year = 2000;
```

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.

```
--7
SELECT state AS below_average_states_y2000
FROM naep
WHERE avg_math_4_score < ALL
(       SELECT ROUND(AVG(avg_math_4_score),2) AS avg_avg_math_4_score
        FROM naep
        WHERE year = 2000)
AND year = 2000;
```

To answer the question list in 5 (lowest score states) is present in list 7 (below average score states) . The two lists should overlap despite null values in average score for 10 states as the average will be calculated for the remaining 41 states.

Take a look at your results. Do your above lists overlap? Should they overlap? It's important to remember that if missing values are not handled properly, you may end up with inaccurate calculations and incorrect conclusions. In the lists you've created, you would expect some of the states that showed up in the bottom 10 to also show up as scoring below the average over all states.

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.

```
--8
SELECT state AS scores_missing_y2000
FROM naep
WHERE year = 2000 AND avg_math_4_score IS NULL;
```

After finding out that some states have missing values for avg_math_4_score in the year 2000, you may decide to alter how you report on the states in the bottom 10. To be clear: we're not asking you to do this for the exam. But in a real-world scenario, you might do this!

Proceeding with your analysis, you suspect that there may be a correlation between avg_math_4_score and total_expenditure for the year 2000. You hypothesize that where less money is spent, scores will be lower. Rigorously proving something like this requires some basic statistics knowledge that we haven't covered yet. Nevertheless, you can write a query that should allow you to "eyeball" this correlation.

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.

At first glance, you should see that there seems to be a correlation.

```
--9
SELECT naep.state, ROUND(naep.avg_math_4_score,2) as round_avg_math_4_score,
finance.total_expenditure
FROM naep LEFT OUTER JOIN finance
```

```
ON naep.id = finance.id
WHERE naep.avg_math_4_Score IS NOT NULL AND
        naep.year = 2000
ORDER BY finance.total_expenditure DESC;
```

Looking at the average scores and total expenditure , there is no strong correlation between the two. For example the lowest score state District of Columbia has third lowest expenditure at the same time California has the third lowest score with highest expenditure. Also Texas has third highest expenditure and the fifth highest average score.