</talentlabs>

Chapter 5 Assignment - SQL Insight Generation

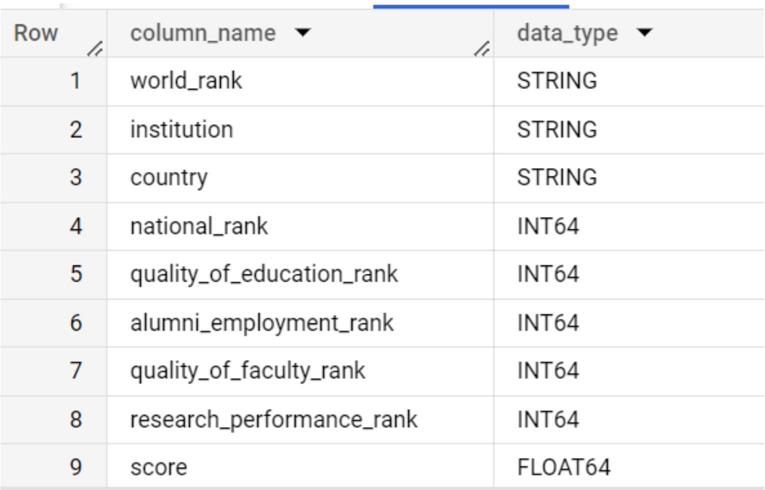
Instructions

1. You can take help from the lecture notes to revise the concepts that we have covered
2. Choose the best suitable answer and submit the word document
3. You have been provided a csv file named “Top 2000 Universities of the World.csv”, this is your dataset for this assignment.
4. For these questions, you need to work on Google BigQuery and answer the questions in this document.
5. To get started with the assignment, you need to create a database and dataset in Google BigQuery using the csv file provided to you as a dataset. (You can take help from the Hands On exercise video from lectures)
6. For each question, apart from answering the questions, please also paste a screenshot of the SQL with the SQL output, as a proof of your work.
7. Please submit the assignment through TalentLabs Learning System. You will need to submit this word document.

Question 1 (3 point):
If you recall the SQL hands-on analysis video, you should remember that we need to create a table schema when we set up the data table before working on BigQuery.

What are the data types in your schema?

Types of all columns (2 points for the answers)



Row	column_name	data_type
1	world_rank	STRING
2	institution	STRING
3	country	STRING
4	national_rank	INT64
5	quality_of_education_rank	INT64
6	alumni_employment_rank	INT64
7	quality_of_faculty_rank	INT64
8	research_performance_rank	INT64
9	score	FLOAT64

SQL Query (1 point for the SQL query):

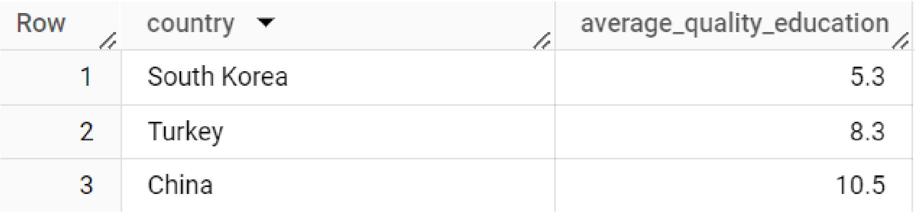
```
SELECT column_name, data_type FROM `striking-arbor-
398510.university`.INFORMATION_SCHEMA.COLUMNS;
```

Question 2 (6 points):
Data aggregation can help us in understanding how different groups of data compare to each other. In this exercise, we would like to understand which country is having the best ranking in terms of quality education.

To achieve this, we can calculate the average”quality education” ranking of each country, and see which country is having the best or worst average ranking.

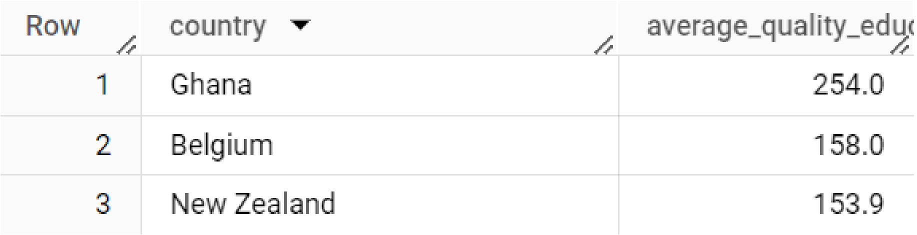
(Hint: you can do a group by, and you should ignore countries with no data on quality education ranking)

Top 3 countries in terms of quality education (2 points)



Row	country	average_quality_education
1	South Korea	5.3
2	Turkey	8.3
3	China	10.5

Last 3 countries in terms of quality education (2 points)



Row	country	average_quality_edu
1	Ghana	254.0
2	Belgium	158.0
3	New Zealand	153.9

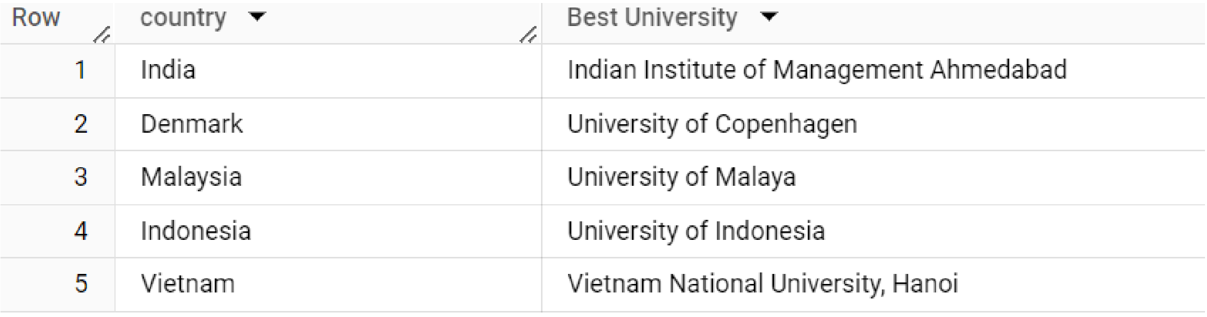
SQL Query (2 points)

```
    (i)    Top 3 countries in terms of quality education:

SELECT country, ROUND(AVG(quality_of_education_rank),1) AS
average_quality_education
FROM ` striking-arbor-398510.university.student`
GROUP BY country
HAVING average_quality_education > 0
ORDER BY average_quality_education ASC
LIMIT 3;

    (ii)   Last 3 countries in terms of quality education:

SELECT country, ROUND(AVG(quality_of_education_rank),1) AS
average_quality_education
FROM `striking-arbor-398510.university.student`
GROUP BY country
HAVING average_quality_education > 0
ORDER BY average_quality_education DESC
LIMIT 3;
```

Question 3 (7 points):

In this dataset, there is a column named “National Ranking”, which shows the ranking of the universities within their own country. This can help us in identifying the best university in each of the countries.

Let’s try to find out the top universities of the countries listed below:



Row	country	Best University
1	India	Indian Institute of Management Ahmedabad
2	Denmark	University of Copenhagen
3	Malaysia	University of Malaya
4	Indonesia	University of Indonesia
5	Vietnam	Vietnam National University, Hanoi

SQL Query (2 points):

```
SELECT country, Institution AS `Best University` FROM
 `striking-arbor-398510.university.student`
WHERE National_Rank=1
AND country IN ("India","Denmark","Malaysia","Indonesia", "Vietnam");
```

Question 4 (5 points):
Data Summaries like mean, mode and media are great ways of summarising large datasets and generating insights.

In this question, we would like to do some analysis for universities in the UK. In order to do that,

1. you need to make a sub table for United Kingdom Institutions
2. summarise the column of interest using measures of location which should include Mean, Mode and Median.

Let’s try to answer the following:
In terms of UK universities research performance ranking:

Mean (1 point)
775.9

Median (1 point)
707

Although the UK got really good university (e.g. Oxford University and Cambridge University), why is the mean ranking in research performance still bad? (1 point)

Mean (775.9) > Median (707). This suggests that there are some universities with very high research performance rankings that are pushing the mean ranking up.

SQL Query: (2 points)

- Create sub table for United Kingdom institution

```
CREATE TABLE `striking-arbor-398510.university.student_uk` AS (
 SELECT * FROM `striking-arbor-398510.university.student`
 WHERE country = "United Kingdom");
```

- Mean:

```
SELECT
  ROUND(AVG(research_performance_rank),1) as mean_score,
FROM `striking-arbor-398510.university.student_uk`
```

- Median:

```
SELECT
  PERCENTILE_CONT(research_performance_rank, 0.5) OVER(PARTITION BY country) as
median_score,
FROM `striking-arbor-398510.university.student_uk`
LIMIT 1;
```