Data Analytics Bootcamp

Data Analytics Bootcamp

Fundamental Queries

● SELECT	Column
● FROM	Table
● WHERE	Filter
● GROUP BY	Aggregating*
● HAVING	Filtering Aggregates*
● ORDER BY	Arranging
● LIMIT	View

** only if there is an aggregate function*

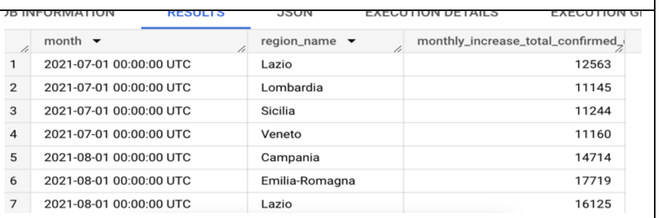
Syntax Guide

```
SELECT
      [ALIAS.column]
      , [ALIAS.column2]
      , AGGREGATE_FUNCTION(ALIAS.column3)
      , …
FROM
      [table ALIAS]
WHERE
      [column] [operator] [condition]
GROUP BY
      [list of columns without functions]
HAVING
     AGGREGATE_FUNCTION(ALIAS.column3) [operator] [condition]
ORDER BY
      [column] [ASC/DESC]
LIMIT
      [number of rows]
```

I. From the previous question, which regions and month registered an increase of more than 10000?

#	*Question/Task*	*Query/Answer*
1	Which table do I get the data **from?**	bigquery-public-data.covid19_italy.data_by_region
2	Which columns do I need to **SELECT**	Region_name, date, new_total_confirmed_cases
3	Do I need to aggregate any of the columns in #2? If so, which ones, and which function?	YES, new_total_confirmed_cases, SUM function
4	Am I getting totals or do I need to aggregate on certain columns? If I'm aggregating on columns, which in #2? (i.e. which columns do not have functions)	SUM(new_total_confirmed_cases) AS monthly_increase_total_confirmed_cases
5	Do I need to filter out any data **BEFORE** aggregating (WHERE)? If #5 is so, what filters do I need?	YES, `DATE(J.date) BETWEEN DATE('2021-07-01') AND DATE('2021-09-30')`
6	Do I need to filter out any data **AFTER** aggregating? (HAVING)	YES, SUM(new_total_confirmed_cases > 10000
7	Do I need to arrange my dataset? Which column? In ascending or descending order?	YES. region_name, date ASC
8	Do I need to limit the results of my dataset? If so, to how many rows?	No

9	Show the query to get the data needed. You can type, copy paste, or paste an image.	`/*From previous question, which regions and month registered an increase of more than 10000*/`
		`SELECT`
		`DATE_TRUNC (J.date,MONTH) AS month`
		`,J.region_name`
		`,SUM (J.new_total_confirmed_cases) AS monthly_increase_total_confirmed_cases`
		`FROM bigquery-public-data.covid19_italy.data_by_region J`

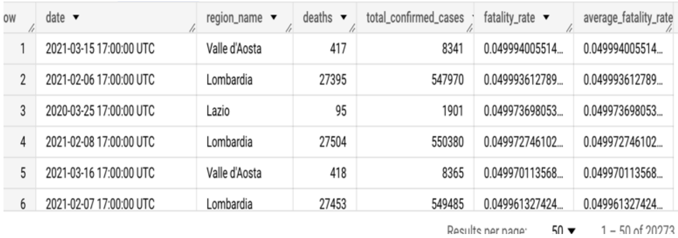
<table>
<tr><td></td><td></td><td><pre>
WHERE
DATE(J.date) BETWEEN DATE('2021-07-01') AND
DATE('2021-09-30')
GROUP BY
1,2
HAVING
SUM(new_total_confirmed_cases)>10000
ORDER BY
1,2 ASC;
</pre></td></tr>
<tr><td>10</td><td>Show a screenshot of the output. No need to show everything, just a sample will do.</td><td>

<table>
<tr><th></th><th>month</th><th>region_name</th><th>monthly_increase_total_confirmed_</th></tr>
<tr><td>1</td><td>2021-07-01 00:00:00 UTC</td><td>Lazio</td><td>12563</td></tr>
<tr><td>2</td><td>2021-07-01 00:00:00 UTC</td><td>Lombardia</td><td>11145</td></tr>
<tr><td>3</td><td>2021-07-01 00:00:00 UTC</td><td>Sicilia</td><td>11244</td></tr>
<tr><td>4</td><td>2021-07-01 00:00:00 UTC</td><td>Veneto</td><td>11160</td></tr>
<tr><td>5</td><td>2021-08-01 00:00:00 UTC</td><td>Campania</td><td>14714</td></tr>
<tr><td>6</td><td>2021-08-01 00:00:00 UTC</td><td>Emilia-Romagna</td><td>17719</td></tr>
<tr><td>7</td><td>2021-08-01 00:00:00 UTC</td><td>Lazio</td><td>16125</td></tr>
</table></td></tr>
</table>

II. Which regions have an average fatality rate of less than 5%? Consider only days where total cases > 0, and sort results from highest fatality rate to lowest.

#	*Question/Task*	*Query/Answer*
1	Which table do I get the data **from?**	bigquery-public-data.covid19_italy.data_by_region
2	Which columns do I need to **SELECT**	Date, region_name, deaths, total_confirmed_cases
3	Do I need to aggregate any of the columns in #2? If so, which ones, and which function?	YES. deaths, total_confirmed_cases. AVG
4	Am I getting totals or do I need to aggregate on certain columns? If I'm aggregating on columns, which in #2? (i.e. which columns do not have functions)	Deaths/total_confirmed_cases AS fatality_rate, AVG(deaths/total_confirmed_cases) AS average_fatality_rate

5	Do I need to filter out any data **BEFORE** aggregating (WHERE)? If #5 is so, what filters do I need?	total_confirmed_cases > 0
6	Do I need to filter out any data **AFTER** aggregating? (HAVING)	YES. AVG(A.deaths/A.total_confirmed_cases) < 0.05
7	Do I need to arrange my dataset? Which column? In ascending or descending order?	Fatality_rate DESC
8	Do I need to limit the results of my dataset? If so, to how many rows?	No
9	Show the query to get the data needed. You can type, copy paste, or paste an image.	/*Show regions having an average fatality rate of less than 5%. Consider only days where total cases > 0, and sort results from highest fatality rate to lowest*/ SELECT A.date ,A.region_name ,A.deaths ,A.total_confirmed_cases ,A.deaths/A.total_confirmed_cases AS fatality_rate ,AVG(A.deaths/A.total_confirmed_cases) AS average_fatality_rate FROM bigquery-public-data.covid19_italy.data_by_region A WHERE A.total_confirmed_cases > 0 GROUP BY A.date ,A.region_name ,A.deaths ,A.total_confirmed_cases HAVING AVG(A.deaths/A.total_confirmed_cases) < 0.05 ORDER BY fatality_rate DESC;

10	Show a screenshot of the output. No need to show everything, just a sample will do.	

ow	date ▾	region_name ▾	deaths ▾	total_confirmed_cases	fatality_rate ▾	average_fatality_rate
1	2021-03-15 17:00:00 UTC	Valle d'Aosta	417	8341	0.049994005514...	0.049994005514...
2	2021-02-06 17:00:00 UTC	Lombardia	27395	547970	0.049993612789...	0.049993612789...
3	2020-03-25 17:00:00 UTC	Lazio	95	1901	0.049973698053...	0.049973698053...
4	2021-02-08 17:00:00 UTC	Lombardia	27504	550380	0.049972746102...	0.049972746102...
5	2021-03-16 17:00:00 UTC	Valle d'Aosta	418	8365	0.049970113568...	0.049970113568...
6	2021-02-07 17:00:00 UTC	Lombardia	27453	549485	0.049961327424...	0.049961327424...

Results per page: 50 ▾ 1 – 50 of 20273