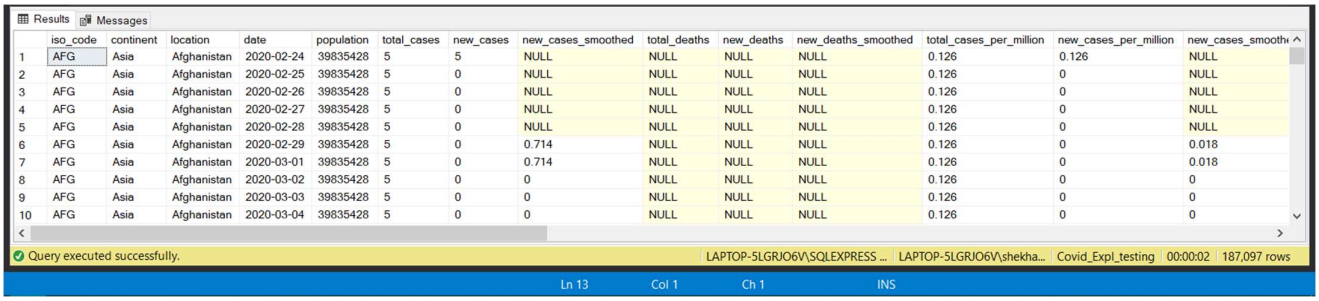
Data Exploration in SQL

Task : Explore the COVID-19 death and vaccination data (in Microsoft SQL Server)

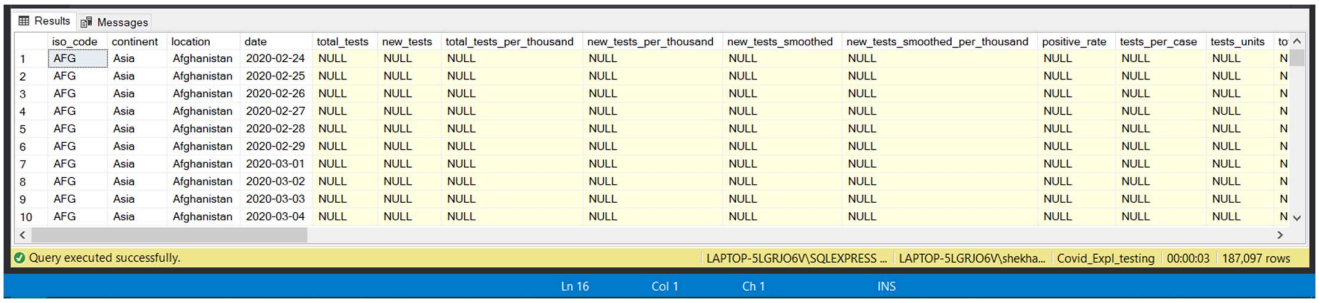
Columns in the “vaccinations” and “deaths” tables:

1) Output after Subtask 1
    a. “deaths” table:



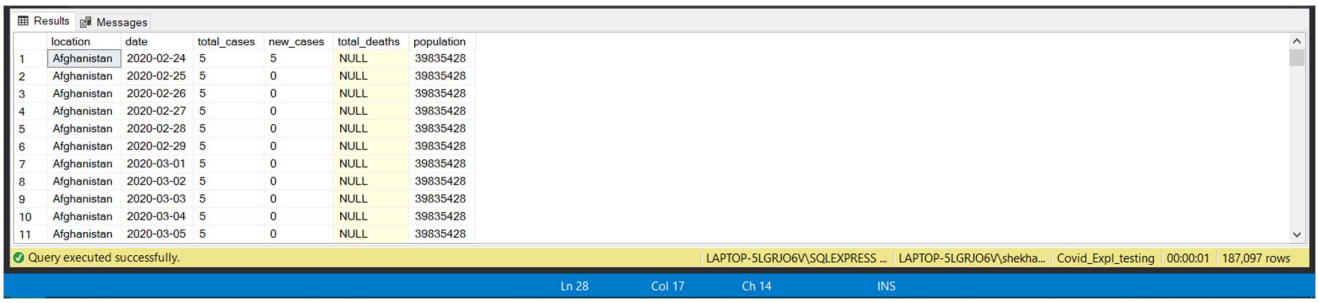
	iso_code	continent	location	date	population	total_cases	new_cases	new_cases_smoothed	total_deaths	new_deaths	new_deaths_smoothed	total_cases_per_million	new_cases_per_million	new_cases_smooth
1	AFG	Asia	Afghanistan	2020-02-24	39835428	5	5	NULL	NULL	NULL	NULL	0.126	0.126	NULL
2	AFG	Asia	Afghanistan	2020-02-25	39835428	5	0	NULL	NULL	NULL	NULL	0.126	0	NULL
3	AFG	Asia	Afghanistan	2020-02-26	39835428	5	0	NULL	NULL	NULL	NULL	0.126	0	NULL
4	AFG	Asia	Afghanistan	2020-02-27	39835428	5	0	NULL	NULL	NULL	NULL	0.126	0	NULL
5	AFG	Asia	Afghanistan	2020-02-28	39835428	5	0	NULL	NULL	NULL	NULL	0.126	0	NULL
6	AFG	Asia	Afghanistan	2020-02-29	39835428	5	0	0.714	NULL	NULL	NULL	0.126	0	0.018
7	AFG	Asia	Afghanistan	2020-03-01	39835428	5	0	0.714	NULL	NULL	NULL	0.126	0	0.018
8	AFG	Asia	Afghanistan	2020-03-02	39835428	5	0	0	NULL	NULL	NULL	0.126	0	0
9	AFG	Asia	Afghanistan	2020-03-03	39835428	5	0	0	NULL	NULL	NULL	0.126	0	0
10	AFG	Asia	Afghanistan	2020-03-04	39835428	5	0	0	NULL	NULL	NULL	0.126	0	0

    b. “vaccinations” table:



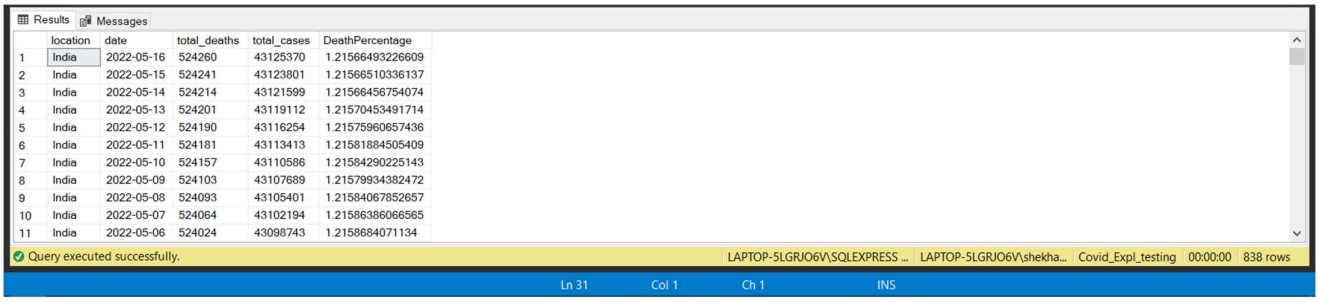
	iso_code	continent	location	date	total_tests	new_tests	total_tests_per_thousand	new_tests_per_thousand	new_tests_smoothed	new_tests_smoothed_per_thousand	positive_rate	tests_per_case	tests_units	to
1	AFG	Asia	Afghanistan	2020-02-24	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	N
2	AFG	Asia	Afghanistan	2020-02-25	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	N
3	AFG	Asia	Afghanistan	2020-02-26	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	N
4	AFG	Asia	Afghanistan	2020-02-27	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	N
5	AFG	Asia	Afghanistan	2020-02-28	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	N
6	AFG	Asia	Afghanistan	2020-02-29	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	N
7	AFG	Asia	Afghanistan	2020-03-01	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	N
8	AFG	Asia	Afghanistan	2020-03-02	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	N
9	AFG	Asia	Afghanistan	2020-03-03	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	N
10	AFG	Asia	Afghanistan	2020-03-04	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	N

2) Output after Subtask 2



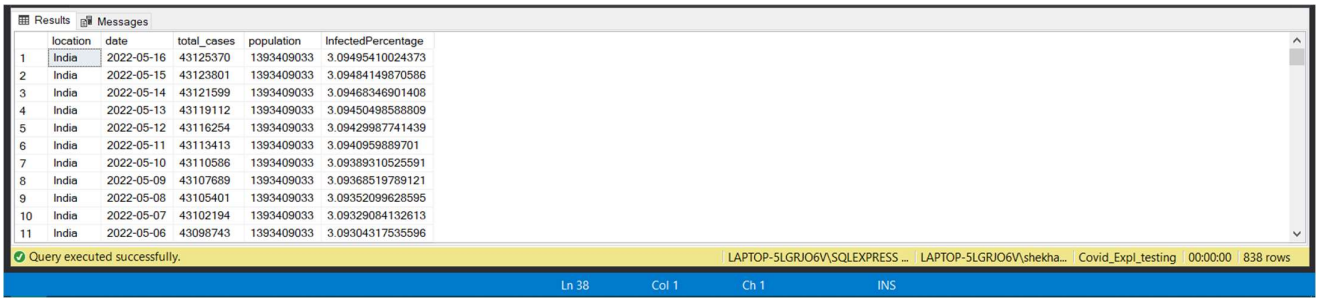
	location	date	total_cases	new_cases	total_deaths	population
1	Afghanistan	2020-02-24	5	5	NULL	39835428
2	Afghanistan	2020-02-25	5	0	NULL	39835428
3	Afghanistan	2020-02-26	5	0	NULL	39835428
4	Afghanistan	2020-02-27	5	0	NULL	39835428
5	Afghanistan	2020-02-28	5	0	NULL	39835428
6	Afghanistan	2020-02-29	5	0	NULL	39835428
7	Afghanistan	2020-03-01	5	0	NULL	39835428
8	Afghanistan	2020-03-02	5	0	NULL	39835428
9	Afghanistan	2020-03-03	5	0	NULL	39835428
10	Afghanistan	2020-03-04	5	0	NULL	39835428
11	Afghanistan	2020-03-05	5	0	NULL	39835428

3) Output after Subtask 3



	location	date	total_deaths	total_cases	DeathPercentage
1	India	2022-05-16	524260	43125370	1.21566493226609
2	India	2022-05-15	524241	43123801	1.21566510336137
3	India	2022-05-14	524214	43121599	1.21566456754074
4	India	2022-05-13	524201	43119112	1.21570453491714
5	India	2022-05-12	524190	43116254	1.21575960657436
6	India	2022-05-11	524181	43113413	1.21581884505409
7	India	2022-05-10	524157	43110586	1.21584290225143
8	India	2022-05-09	524103	43107689	1.21579934382472
9	India	2022-05-08	524093	43105401	1.21584067852657
10	India	2022-05-07	524064	43102194	1.21586386066565
11	India	2022-05-06	524024	43098743	1.2158684071134

4) Output after Subtask 4



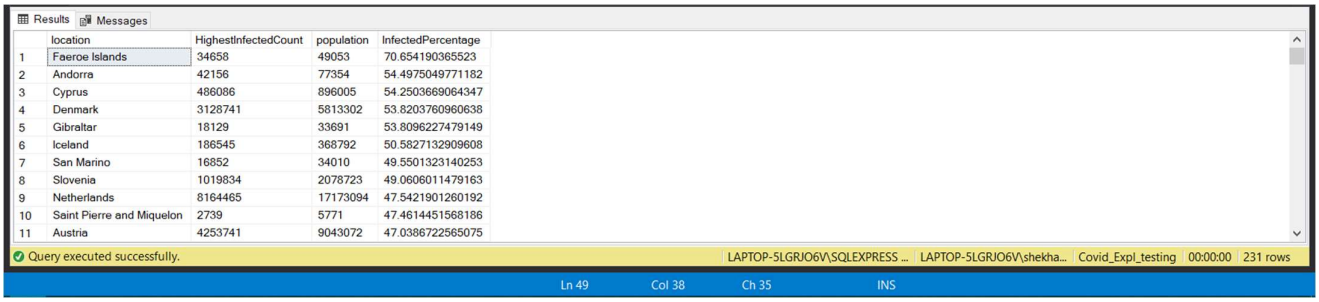
Results Messages

	location	date	total_cases	population	InfectedPercentage
1	India	2022-05-16	43125370	1393409033	3.09495410024373
2	India	2022-05-15	43123801	1393409033	3.09484149870586
3	India	2022-05-14	43121599	1393409033	3.09468346901408
4	India	2022-05-13	43119112	1393409033	3.09450498588809
5	India	2022-05-12	43116254	1393409033	3.09429987741439
6	India	2022-05-11	43113413	1393409033	3.0940959889701
7	India	2022-05-10	43110586	1393409033	3.09389310525591
8	India	2022-05-09	43107689	1393409033	3.09368519789121
9	India	2022-05-08	43105401	1393409033	3.09352099628595
10	India	2022-05-07	43102194	1393409033	3.09329084132613
11	India	2022-05-06	43098743	1393409033	3.09304317535596

Query executed successfully. LAPTOP-5LGRJO6V\SQLEXPRESS ... LAPTOP-5LGRJO6V\shekha... Covid_Expl_testing 00:00:00 838 rows

Ln 38 Col 1 Ch 1 INS

5) Output after Subtask 5



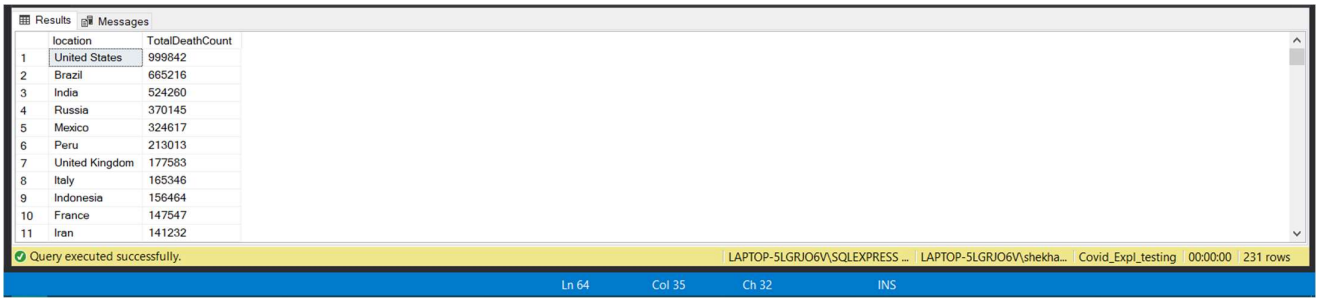
Results Messages

	location	HighestInfectedCount	population	InfectedPercentage
1	Faeroe Islands	34658	49053	70.654190365523
2	Andorra	42156	77354	54.4975049771182
3	Cyprus	486086	896005	54.2503669064347
4	Denmark	3128741	5813302	53.8203760960638
5	Gibraltar	18129	33691	53.8096227479149
6	Iceland	186545	368792	50.5827132909608
7	San Marino	16852	34010	49.5501323140253
8	Slovenia	1019834	2078723	49.0606011479163
9	Netherlands	8164465	17173094	47.5421901260192
10	Saint Pierre and Miquelon	2739	5771	47.4614451568186
11	Austria	4253741	9043072	47.0386722565075

Query executed successfully. LAPTOP-5LGRJO6V\SQLEXPRESS ... LAPTOP-5LGRJO6V\shekha... Covid_Expl_testing 00:00:00 231 rows

Ln 49 Col 38 Ch 35 INS

6) Output after Subtask 7



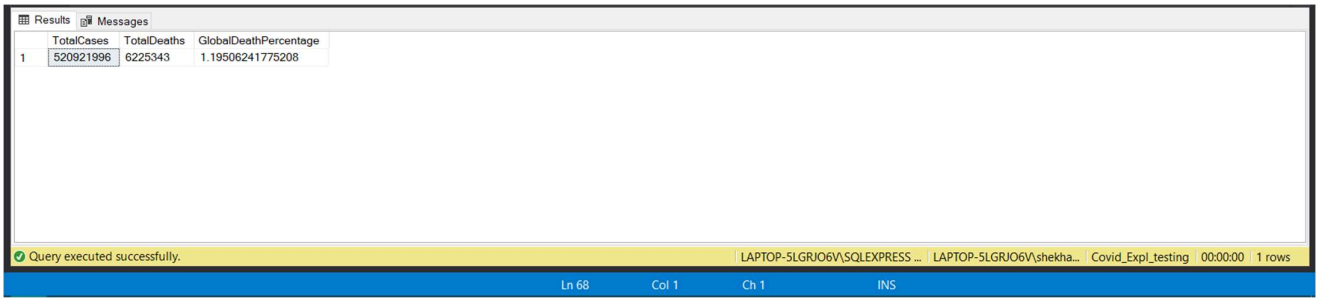
Results Messages

	location	TotalDeathCount
1	United States	999842
2	Brazil	665216
3	India	524260
4	Russia	370145
5	Mexico	324617
6	Peru	213013
7	United Kingdom	177583
8	Italy	165346
9	Indonesia	156464
10	France	147547
11	Iran	141232

Query executed successfully. LAPTOP-5LGRJO6V\SQLEXPRESS ... LAPTOP-5LGRJO6V\shekha... Covid_Expl_testing 00:00:00 231 rows

Ln 64 Col 35 Ch 32 INS

7) Output after Subtask 8



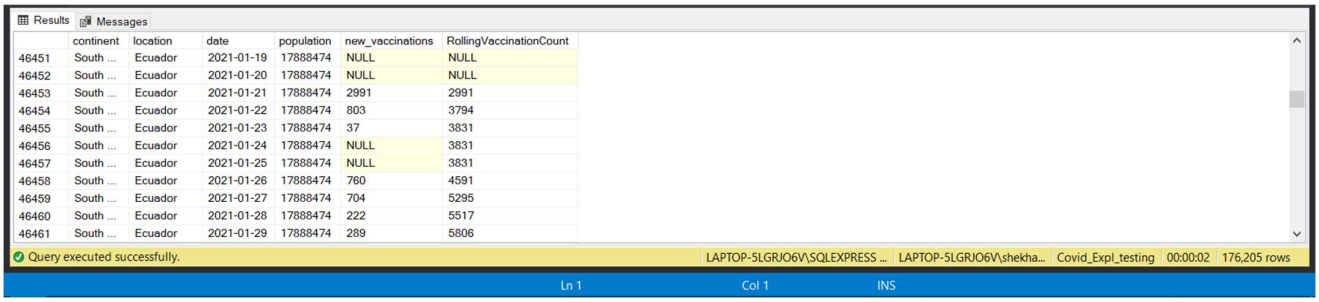
Results Messages

	TotalCases	TotalDeaths	GlobalDeathPercentage
1	520921996	6225343	1.19506241775208

Query executed successfully. LAPTOP-5LGRJO6V\SQLEXPRESS ... LAPTOP-5LGRJO6V\shekha... Covid_Expl_testing 00:00:00 1 rows

Ln 68 Col 1 Ch 1 INS

8) Output after Subtask 9



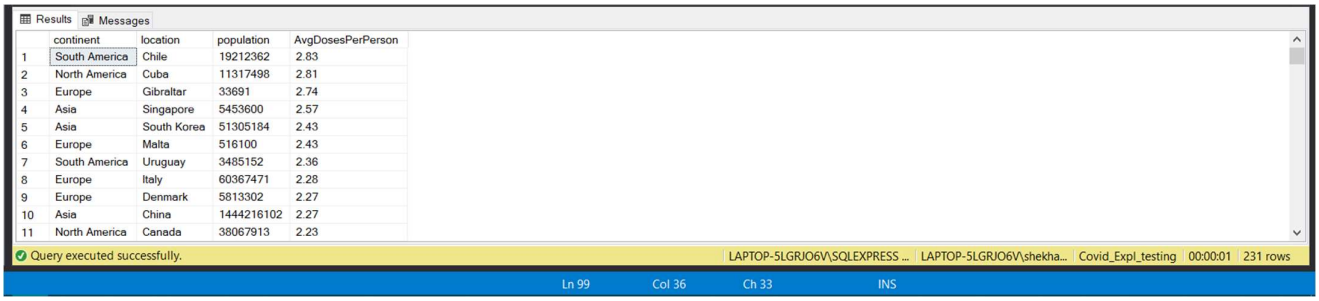
Results Messages

	continent	location	date	population	new_vaccinations	RollingVaccinationCount
46451	South ...	Ecuador	2021-01-19	17888474	NULL	NULL
46452	South ...	Ecuador	2021-01-20	17888474	NULL	NULL
46453	South ...	Ecuador	2021-01-21	17888474	2991	2991
46454	South ...	Ecuador	2021-01-22	17888474	803	3794
46455	South ...	Ecuador	2021-01-23	17888474	37	3831
46456	South ...	Ecuador	2021-01-24	17888474	NULL	3831
46457	South ...	Ecuador	2021-01-25	17888474	NULL	3831
46458	South ...	Ecuador	2021-01-26	17888474	760	4591
46459	South ...	Ecuador	2021-01-27	17888474	704	5295
46460	South ...	Ecuador	2021-01-28	17888474	222	5517
46461	South ...	Ecuador	2021-01-29	17888474	289	5806

Query executed successfully. LAPTOP-5LGRJO6V\SQLEXPRESS ... LAPTOP-5LGRJO6V\shekha... Covid_Expl_testing 00:00:02 176,205 rows

Ln 1 Col 1 INS

9) Output after Subtask 10



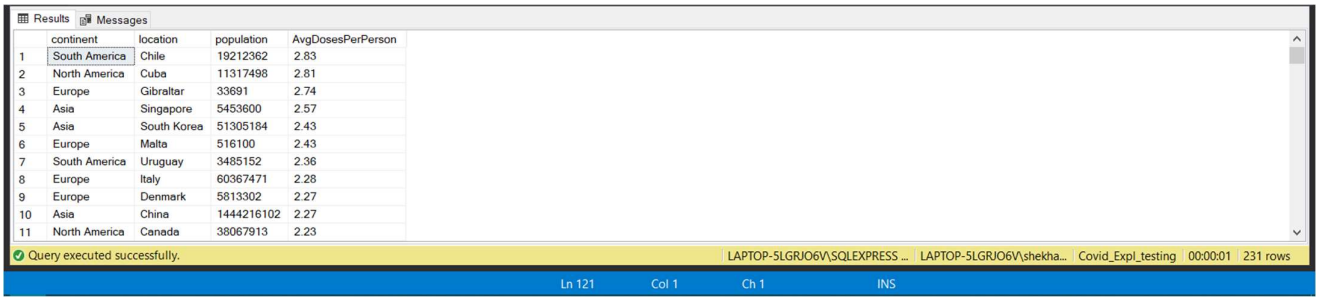
Results Messages

	continent	location	population	AvgDosesPerPerson
1	South America	Chile	19212362	2.83
2	North America	Cuba	11317498	2.81
3	Europe	Gibraltar	33691	2.74
4	Asia	Singapore	5453600	2.57
5	Asia	South Korea	51305184	2.43
6	Europe	Malta	516100	2.43
7	South America	Uruguay	3485152	2.36
8	Europe	Italy	60367471	2.28
9	Europe	Denmark	5813302	2.27
10	Asia	China	1444216102	2.27
11	North America	Canada	38067913	2.23

Query executed successfully. LAPTOP-5LGRJO6V\SQLEXPRESS ... LAPTOP-5LGRJO6V\shekha... Covid_Expl_testing 00:00:01 231 rows

Ln 99 Col 36 Ch 33 INS

10) Output after Subtask 12



Results Messages

	continent	location	population	AvgDosesPerPerson
1	South America	Chile	19212362	2.83
2	North America	Cuba	11317498	2.81
3	Europe	Gibraltar	33691	2.74
4	Asia	Singapore	5453600	2.57
5	Asia	South Korea	51305184	2.43
6	Europe	Malta	516100	2.43
7	South America	Uruguay	3485152	2.36
8	Europe	Italy	60367471	2.28
9	Europe	Denmark	5813302	2.27
10	Asia	China	1444216102	2.27
11	North America	Canada	38067913	2.23

Query executed successfully. LAPTOP-5LGRJO6V\SQLEXPRESS ... LAPTOP-5LGRJO6V\shekha... Covid_Expl_testing 00:00:01 231 rows

Ln 121 Col 1 Ch 1 INS

Script:

```
---- Project: DATA EXPLORATION IN SQL ----------------
------------------------------------------------
---- Description: In this project we explore the Covid-
19 death and vaccination data in SQL Server. ----
------------------------------------------------------
------------------------------------------------
---- Note: The data and observations are valid as of 16
May 2022. ------------------------------------
---- Note: First move the "population" column to
position 5 to get owid-covid-data.xlsx file. ----------
---- Note: To get deaths.xlsx, remove columns AA to BO
from owid-covid-data.xlsx file ------------------
---- Note: To get vaccinations.xlsx, remove columns E
to Z from owid-covid-data.xlsx file --------------
---- Note: The functions and commands are consistent
with Microsoft SQL Server. ------------------------
------------------------------------------------------
------------------------------------------------

-- 1) Let's look at both the "deaths" and
"vaccinations" tables.
        SELECT * FROM deaths
        ORDER BY location, date

        SELECT * FROM vaccinations
        ORDER BY location, date

-- 2) Let's change the data type of "date" column in
both tables to DATE, and look at the important columns.
        ALTER TABLE deaths
        ALTER COLUMN date DATE

        ALTER TABLE vaccinations
        ALTER COLUMN date DATE

        SELECT location, date, total_cases, new_cases,
total_deaths, population
        FROM deaths
        ORDER BY 1,2

-- 3) Let's look at Total Cases vs Total Deaths in
India.
        SELECT location, date, total_deaths,
total_cases, total_deaths/total_cases*100 as
DeathPercentage
        FROM deaths
        WHERE location LIKE '%india%'
        ORDER BY 1, 2 DESC
---  Observation: Death percentage in India is 1.22%.

-- 4) Let's look at Total Cases vs Population in India.
        SELECT location, date, total_cases, population,
total_cases/population*100 as InfectedPercentage
        FROM deaths
        WHERE location LIKE '%india%'
        ORDER BY 1, 2 DESC
---  Observation: Case percentage in India is 3.09%.

-- 5) Let's look at the countries with Highest
Infection Rate compared to the population.
        SELECT location, MAX(total_cases) AS
HighestInfectedCount, population,
MAX(total_cases)/population*100 as InfectedPercentage
        FROM deaths
        WHERE continent IS NOT NULL
        GROUP BY location, population
        ORDER BY InfectedPercentage DESC
---  Observation: Faroe Islands has had 70.65%
population infected over the period.

-- 6) Let's convert the type of columns "total_deaths"
and "new_deaths" in "deaths" table from NVARCHAR(255)
to INT.
        ALTER TABLE deaths
        ALTER COLUMN total_deaths INT

        ALTER TABLE deaths
        ALTER COLUMN new_deaths INT

-- 7) Now let's look at the countries with Highest
number of Deaths.
        SELECT location, MAX(total_deaths) AS
TotalDeathCount
        FROM deaths
        WHERE continent IS NOT NULL
        GROUP BY location
        ORDER BY TotalDeathCount DESC
---  Observation: United State has highest number of
deaths at 999842.

-- 8) Let's focus on the data for the world.
        SELECT SUM(new_cases) AS TotalCases,
SUM(new_deaths) AS TotalDeaths,
                SUM(new_deaths)/SUM(new_cases)*100 AS
GlobalDeathPercentage
        FROM deaths
        WHERE continent IS NOT NULL
---  Observation: Globally, the death percentage has
been 1.20%.

-- 9) Let's get the Total number of Vaccinations on
each day for each country using a rolling count.
        SELECT dea.continent, dea.location, dea.date,
dea.population, vac.new_vaccinations,
                SUM(CONVERT(bigint,
vac.new_vaccinations)) OVER (PARTITION BY dea.location
ORDER BY dea.location, dea.date) AS
RollingVaccinationCount
        FROM deaths as dea
        JOIN vaccinations as vac
                ON dea.location = vac.location
                AND dea.date = vac.date
        WHERE dea.continent IS NOT NULL
        ORDER BY 2,3

-- 10) Let's find the Average of number of doses
received by a person in a country.
        WITH AvgDoses (continent, location, date,
population, new_vaccinations, RollingVaccinationCount)
AS
        (
        SELECT dea.continent, dea.location, dea.date,
dea.population, vac.new_vaccinations,
                SUM(CONVERT(bigint,
vac.new_vaccinations)) OVER (PARTITION BY dea.location
ORDER BY dea.location, dea.date) AS
RollingVaccinationCount
        FROM deaths as dea
        JOIN vaccinations as vac
                ON dea.location = vac.location
                AND dea.date = vac.date
        WHERE dea.continent IS NOT NULL
        )
        SELECT continent, location, population,
ROUND(MAX(RollingVaccinationCount)/population,2) AS
AvgDosesPerPerson
        FROM AvgDoses
        WHERE continent IS NOT NULL
        GROUP BY continent, location, population
        ORDER BY AvgDosesPerPerson DESC
---  Observation: Chile has given an average of 2.83
doses per person (highest), while India has given 1.32.

-- 11) Let's store the results from the previous query
into a view for later use.
        CREATE VIEW ViewAvgDoses AS
```

```
        WITH AvgDoses (continent, location, date,
population, new_vaccinations, RollingVaccinationCount)
AS
        (
        SELECT dea.continent, dea.location, dea.date,
dea.population, vac.new_vaccinations,
                SUM(CONVERT(bigint,
vac.new_vaccinations)) OVER (PARTITION BY dea.location
ORDER BY dea.location, dea.date) AS
RollingVaccinationCount
        FROM deaths as dea
        JOIN vaccinations as vac
                ON dea.location = vac.location
                AND dea.date = vac.date
        WHERE dea.continent IS NOT NULL
        )
        SELECT continent, location, population,
ROUND(MAX(RollingVaccinationCount)/population,2) AS
AvgDosesPerPerson
        FROM AvgDoses
        WHERE continent IS NOT NULL
        GROUP BY continent, location, population
        --ORDER BY AvgDosesPerPerson DESC (commented
out because orderby is invalid in creating views)

-- 12) Let's see the created view.
        SELECT * FROM ViewAvgDoses
        ORDER BY AvgDosesPerPerson DESC
```

Project Owner: More Shekhar Sanjay

Project Repository: https://github.com/MoreShekharSanjay/project-data-exploration-in-sql

Email me at: moreshekharsanjay@gmail.com

My LinkedIn: https://www.linkedin.com/in/moreshekharsanjay/

Dataset source: https://ourworldindata.org/covid-deaths

Software and languages used:
SQL, Microsoft SQL Server Management Studio 18, Microsoft Excel

---