

COVID ASSIGNMENT BY USING DATABRICKS

Using SQL in Databricks

List out the files

```
%python
```

```
df = spark.read.csv("dbfs:/FileStore/tables/Covid_Filtered_Data.csv")
```

create a tempview

```
%python
```

```
df.createOrReplaceTempView("covid_sql")
```

mouni_covid_sql SQL Free trial ends in 6 days. Continue with a pay-as-you-go subscription by [providing your billing information](#). Schedule Share

Cmd 1

```
1 %python
2 df = spark.read.csv("dbfs:/FileStore/tables/Covid_Filtered_Data.csv")
```

▶ (1) Spark Jobs

df: pyspark.sql.dataframe.DataFrame = [_c0: string, _c1: string ... 65 more fields]

Command took 0.72 seconds -- by training@kpipartners.com at 3/17/2022, 10:47:17 AM on [default]basic-starter-cluster

Cmd 2

```
1 %python
2
3 df.createOrReplaceTempView("covid_sql")
4
```

Command took 0.02 seconds -- by training@kpipartners.com at 3/17/2022, 10:47:37 AM on [default]basic-starter-cluster

Select * from covid_sql

mouni_covid_sql SQL Free trial ends in 6 days. Continue with a pay-as-you-go subscription by [providing your billing information](#). Schedule Share

Cmd 3

```
1 select * from covid_sql
2
```

▶ (1) Spark Jobs

Table Data Profile

	_c0	_c1	_c2	_c3	_c4	_c5	_c6	_c7	_c8	_c9
	iso_code	continent	location	date	total_cases	new_cases	new_cases_smoothed	total_deaths	new_deaths	new_deaths_smoothed
1										
2	AFG	Asia	Afghanistan	2/24/2020	5	5	0	0	0	0
3	AFG	Asia	Afghanistan	2/25/2020	5	0	0	0	0	0
4	AFG	Asia	Afghanistan	2/26/2020	5	0	0	0	0	0
5	AFG	Asia	Afghanistan	2/27/2020	5	0	0	0	0	0
6	AFG	Asia	Afghanistan	2/28/2020	5	0	0	0	0	0
7	AFG	Asia	Afghanistan	2/29/2020	5	0	0.714	0	0	0

Truncated results, showing first 1000 rows.
[Click to re-execute with maximum result limits.](#)

Command took 0.50 seconds -- by training@kpipartners.com at 3/17/2022, 10:49:05 AM on [default]basic-starter-cluster

which continent and location having max of deaths and male_smokers

The screenshot shows a Databricks SQL interface with a query window titled "mouni_covid_sql". The query is: `SELECT _c1,_c2,max(_c7) as max_deaths,max(_c58) as male_smokers from covid_sql group by _c1,_c2 order by max_deaths desc limit 2;`. The result is displayed in a table with 2 rows. The first row shows "continent" and "location" with "total_deaths" and "male_smokers". The second row shows "Europe" and "Austria" with "9997" and "30.9".

	_c1	_c2	max_deaths	male_smokers
1	continent	location	total_deaths	male_smokers
2	Europe	Austria	9997	30.9

display the continent,positivity rate where the stringency index is maximum

The screenshot shows a Databricks SQL interface with a query window titled "mouni_covid_sql". The query is: `SELECT _c1,_c31,max(_c47) from covid_sql group by _c1,_c31,_c47 order by _c47 desc limit 2;`. The result is displayed in a table with 2 rows. The first row shows "continent" and "positive_rate" with "stringency_index". The second row shows "South America" and "0.088" with "98.15".

	_c1	_c31	max(_c47)
1	continent	positive_rate	stringency_index
2	South America	0.088	98.15

which year having the highest number of deaths

mouni_covid_sql SQL Free trial ends in 6 days. Continue with a pay-as-you-go subscription by [providing your billing information](#). Schedule Share

which year having the highest number of deaths

```
1 select _c3,sum(_c7) from covid_sql group by _c3,_c7 order by _c7 desc limit 2;
```

▶ (2) Spark Jobs

Table Data Profile

	_c3	sum(CAST(_c7 AS DOUBLE))
1	date	null
2	4/21/2021	9997

Showing all 2 rows.

Command took 1.05 seconds -- by training@kpipartners.com at 3/17/2022, 10:56:46 AM on [default]basic-starter-cluster

Cmd 7

Using scala:

```
%scala
```

```
display(dbutils.fs.ls("/FileStore/tables"))
```

mouni_covid_scala SQL Free trial ends in 6 days. Continue with a pay-as-you-go subscription by [providing your billing information](#). Schedule Share

```
1 %scala
2 display(dbutils.fs.ls("/FileStore/tables"))
```

Table Data Profile

	path	name	size
1	dbfs:/FileStore/tables/Covid_Filtered_Data.csv	Covid_Filtered_Data.csv	44987720
2	dbfs:/FileStore/tables/Crops_price-1.csv	Crops_price-1.csv	1813
3	dbfs:/FileStore/tables/Crops_price-2.csv	Crops_price-2.csv	1813
4	dbfs:/FileStore/tables/Crops_price.csv	Crops_price.csv	1835
5	dbfs:/FileStore/tables/bamboo.csv	bamboo.csv	664
6	dbfs:/FileStore/tables/bamboo_plantation.txt	bamboo_plantation.txt	666
7	dbfs:/FileStore/tables/bamboo_plantation 15 16.txt	bamboo_plantation 15 16.txt	666

Showing all 31 rows.

Command took 2.55 seconds -- by training@kpipartners.com at 3/17/2022, 11:07:18 AM on [default]basic-starter-cluster

Cmd 2

```
%scala
```

```
df1.createOrReplaceTempView("temp_covid")
```

```
1 %scala
2 val df1 = spark.read.csv("dbfs:/FileStore/tables/Covid_Filtered_Data.csv")

(1) Spark Jobs
  df1: org.apache.spark.sql.DataFrame = [_c0: string, _c1: string ... 65 more fields]
df1: org.apache.spark.sql.DataFrame = [_c0: string, _c1: string ... 65 more fields]
Command took 0.70 seconds -- by training@kpipartners.com at 3/17/2022, 11:07:28 AM on [default]basic-starter-cluster

Cmd 3
1 %scala
2 df1.createOrReplaceTempView("temp_covid")

Command took 0.14 seconds -- by training@kpipartners.com at 3/17/2022, 11:07:51 AM on [default]basic-starter-cluster
```

which location having lowest recovery rate for each continent

```
6. which location having lowest recovery rate for each continent

1 %scala
2 spark.sql("select _c1,max(recovery_rate) from(select _c1,_c2,(_c5 - _c7) as recovery_rate from temp_covid group by _c1,_c2,_c5,_c7 order by recovery_rate asc) group by _c1 limit 6").show()

(3) Spark Jobs
+-----+
|_c1|max(recovery_rate)|
+-----+
|Europe|372950.0|
|Africa|24257.0|
|North America|526880.0|
|South America|21952.0|
|Oceania|172749.0|
|Asia|234933.0|
+-----+

Command took 1.96 seconds -- by training@kpipartners.com at 3/17/2022, 11:08:07 AM on [default]basic-starter-cluster
```

Countries with max cardiovasc_death_rate and diabetes_prevalence greater than 10

mouni_covid_scala SQL Free trial ends in 6 days. Continue with a pay-as-you-go subscription by providing your billing information. Schedule Share

7. Countries with max cardiovasc_death_rate and diabetes_prevalence greater than

```

1 %scala
2 spark.sql("select _c2,max(_c55),max(_c56) from temp_covid where _c55 >10 and _c56 >10 group by _c2").show()

```

(2) Spark Jobs

_c2	max(_c55)	max(_c56)
Antigua and Barbuda	191.511	13.17
Bahamas	235.954	13.17
Bahrain	151.689	16.52
Barbados	170.05	13.57
Belize	176.957	17.11
Bermuda	139.547	13
Brunei	201.285	12.79
Comoros	261.516	11.88
Dominica	227.376	11.62
Egypt	525.432	17.31
Fiji	412.82	14.49
Guyana	373.159	11.62
Jamaica	206.537	11.28
Jordan	208.257	11.75
Kiribati	434.657	22.66
Kuwait	132.235	15.84
Lebanon	266.591	12.71

which location has highest number of vaccinated people in each continent

mouni_covid_scala SQL Free trial ends in 6 days. Continue with a pay-as-you-go subscription by providing your billing information. Schedule Share

8. which location has highest number of vaccinated people in each continent

```

1 %scala
2 spark.sql("select _c1,max(vaccinated_people) from (select distinct _c1,_c2,max(_c34) as vaccinated_people from temp_covid group by _c1,_c2 order by vaccinated_people desc) group by _c1 limit 7").show()

```

(3) Spark Jobs

_c1	max(vaccinated_people)
Africa	9993402
Asia	99963895
Europe	9997608
North America	999990
Oceania	996214
South America	999929
continent	total_vaccinations

Using pyspark

List out the file

mouni_covid_cleandata Python Free trial ends in 6 days. Continue with a pay-as-you-go subscription by providing your billing information. Schedule Share

[default]basic-start...

```
1 display(dbutils.fs.ls("/FileStore/tables"))
```

(2) Spark Jobs

Table Data Profile

	path	name	size
1	dbfs:/FileStore/tables/Covid_Filtered_Data.csv	Covid_Filtered_Data.csv	44987720
2	dbfs:/FileStore/tables/Crops_price-1.csv	Crops_price-1.csv	1813
3	dbfs:/FileStore/tables/Crops_price-2.csv	Crops_price-2.csv	1813
4	dbfs:/FileStore/tables/Crops_price.csv	Crops_price.csv	1835
5	dbfs:/FileStore/tables/bamboo.csv	bamboo.csv	664
6	dbfs:/FileStore/tables/bamboo_plantation.txt	bamboo_plantation.txt	666
7	dbfs:/FileStore/tables/bamboo_plantation 15 16.txt	bamboo_plantation 15 16.txt	666

Showing all 31 rows.

Command took 1.00 second -- by training@kpipartners.com at 3/16/2022, 6:03:12 PM on [default]basic-starter-cluster

mouni_covid_cleandata Python Free trial ends in 6 days. Continue with a pay-as-you-go subscription by providing your billing information. Schedule Share

[default]basic-start...

```
7 dbfs:/FileStore/tables/bamboo_plantation 15 16.txt bamboo_plantation 15 16.txt 666
```

Showing all 31 rows.

Command took 1.00 second -- by training@kpipartners.com at 3/16/2022, 6:03:12 PM on [default]basic-starter-cluster

Cmd 2

```
1 import pandas as pd
2 df = pd.read_csv("/dbfs/FileStore/tables/owid_covid_data.csv")
```

Command took 1.12 seconds -- by training@kpipartners.com at 3/16/2022, 6:03:51 PM on [default]basic-starter-cluster

1. remove the records where the continent column having blanks

mouni_covid_cleandata Python Free trial ends in 6 days. Continue with a pay-as-you-go subscription by providing your billing information. Schedule Share

[default]basic-start...

```
1.. remove the records where the continent column having blanks
```

```
1 for i in range(len(df)):
2     if (df.at[i,'continent'])==0 : # (pd.isnull(df.at[i,'continent']) or
3         df.drop([i], axis=0, inplace=True)
4
```

Command took 0.91 seconds -- by training@kpipartners.com at 3/16/2022, 6:06:07 PM on [default]basic-starter-cluster

2.replace all the NEGATIVE and NULL valued records with 0

mouni_covid_cleandata Python Free trial ends in 6 days. Continue with a pay-as-you-go subscription by providing your billing information. Schedule Share

[default]basic-start... Command took 0.51 seconds -- by training@kpiipartners.com at 3/16/2022, 6:06:07 PM on [default]basic-starter-cluster

Cmd 4

2.replace all the NEGATIVE and NULL valued records with 0

```
1 df=df.fillna(0)
2 if (df.at[i,'new_cases'])<0:
3     df.at[i,'new_cases']=0
4
5 if (df.at[i,'new_cases_smoothed'])<0:
6     df.at[i,'new_cases_smoothed']=0
7
8 if (df.at[i,'new_deaths'])<0:
9     df.at[i,'new_deaths']=0
10
11 if (df.at[i,'new_deaths_smoothed'])<0:
12     df.at[i,'new_deaths_smoothed']=0
13
14 if (df.at[i,'new_cases_per_million'])<0:
15     df.at[i,'new_cases_per_million']=0
16
17 if (df.at[i,'new_cases_smoothed_per_million'])<0:
18     df.at[i,'new_cases_smoothed_per_million']=0
19
20 if (df.at[i,'new_deaths_per_million'])<0:
21     df.at[i,'new_deaths_per_million']=0
22
```

mouni_covid_cleandata Python Free trial ends in 6 days. Continue with a pay-as-you-go subscription by providing your billing information. Schedule Share

[default]basic-start... Command took 0.63 seconds -- by training@kpiipartners.com at 3/16/2022, 8:09:17 PM on [default]basic-starter-cluster

Cmd 5

Queries by using pyspark

```
1 display(dbutils.fs.ls("/FileStore/tables"))
```

▶ (2) Spark Jobs

Table Data Profile

	path	name	size
1	dbfs:/FileStore/tables/Covid_Filtered_Data.csv	Covid_Filtered_Data.csv	44987720
2	dbfs:/FileStore/tables/Crops_price-1.csv	Crops_price-1.csv	1813
3	dbfs:/FileStore/tables/Crops_price-2.csv	Crops_price-2.csv	1813
4	dbfs:/FileStore/tables/Crops_price.csv	Crops_price.csv	1835
5	dbfs:/FileStore/tables/bamboo.csv	bamboo.csv	664
6	dbfs:/FileStore/tables/bamboo_plantation.txt	bamboo_plantation.txt	666
7	dbfs:/FileStore/tables/bamboo_plantation 15 16.txt	bamboo_plantation 15 16.txt	666

Showing all 31 rows.

Cmd 6

mouni_covid_cleandata Python Free trial ends in 6 days. Continue with a pay-as-you-go subscription by providing your billing information. Schedule Share

Command took 0.63 seconds -- by training@kpipartners.com at 3/16/2022, 8:09:17 PM on [default]basic-starter-cluster

Cmd 6

```
1 df1 = spark.read.csv("dbfs://FileStore/tables/Covid_Filtered_Data.csv")
2
```

(1) Spark Jobs

- df1: pyspark.sql.dataframe.DataFrame = [_c0: string, _c1: string ... 65 more fields]

Command took 0.67 seconds -- by training@kpipartners.com at 3/16/2022, 8:09:58 PM on [default]basic-starter-cluster

Cmd 7

```
1 df1.createOrReplaceTempView("covid1")
```

Command took 0.03 seconds -- by training@kpipartners.com at 3/16/2022, 10:20:21 PM on [default]basic-starter-cluster

3. display the continent,positivity rate where the strigency index is maximum

mouni_covid_cleandata Python Free trial ends in 6 days. Continue with a pay-as-you-go subscription by providing your billing information. Schedule Share

Command took 0.03 seconds -- by training@kpipartners.com at 3/16/2022, 10:20:21 PM on [default]basic-starter-cluster

Cmd 8

3.display the continent,positivity rate where the strigency index is maximum

```
1 spark.sql(" SELECT _c1,_c31,max(_c47) from covid1 group by _c1,_c31,_c47 order by _c47 desc limit 2").show()
```

(2) Spark Jobs

_c1	_c31	max(_c47)
continent	positive_rate	stringency_index
South America	0.088	98.15

Command took 1.52 seconds -- by training@kpipartners.com at 3/16/2022, 10:26:26 PM on [default]basic-starter-cluster

4.which continent and location having max of deaths and male_smokers

mouni_covid_cleandata Python Free trial ends in 6 days. Continue with a pay-as-you-go subscription by providing your billing information. Schedule Share

Command took 1.52 seconds -- by training@kpipartners.com at 3/16/2022, 10:26:26 PM on [default]basic-starter-cluster

Cmd 9

4. which continent and location having max of deaths and male_smokers

```
1 spark.sql(" SELECT _c1,_c2,max(_c7) as max_deaths,max(_c58) as male_smokers from covid1 group by _c1,_c2 order by max_deaths desc limit 2").show()
```

(2) Spark Jobs

_c1	_c2	max_deaths	male_smokers
continent	location	total_deaths	male_smokers
Europe	Austria	9997	30.9

Command took 0.92 seconds -- by training@kpipartners.com at 3/16/2022, 10:27:57 PM on [default]basic-starter-cluster

5. which year having the highest number of deaths

mouni_covid_cleandata Python Free trial ends in 6 days. Continue with a pay-as-you-go subscription by providing your billing information. Schedule Share

[default]basic-start...

Cmd 10

5. which year having the highest number of deaths

```
1 spark.sql("select _c3,sum(_c7) from covid1 group by _c3,_c7 order by _c7 desc limit 2 ").show()
```

► (2) Spark Jobs

_c3 sum(CAST(_c7 AS DOUBLE))	
date	null
4/21/2021	9997.0

Command took 1.63 seconds -- by training@kpipartners.com at 3/16/2022, 10:31:30 PM on [default]basic-starter-cluster

Cmd 11