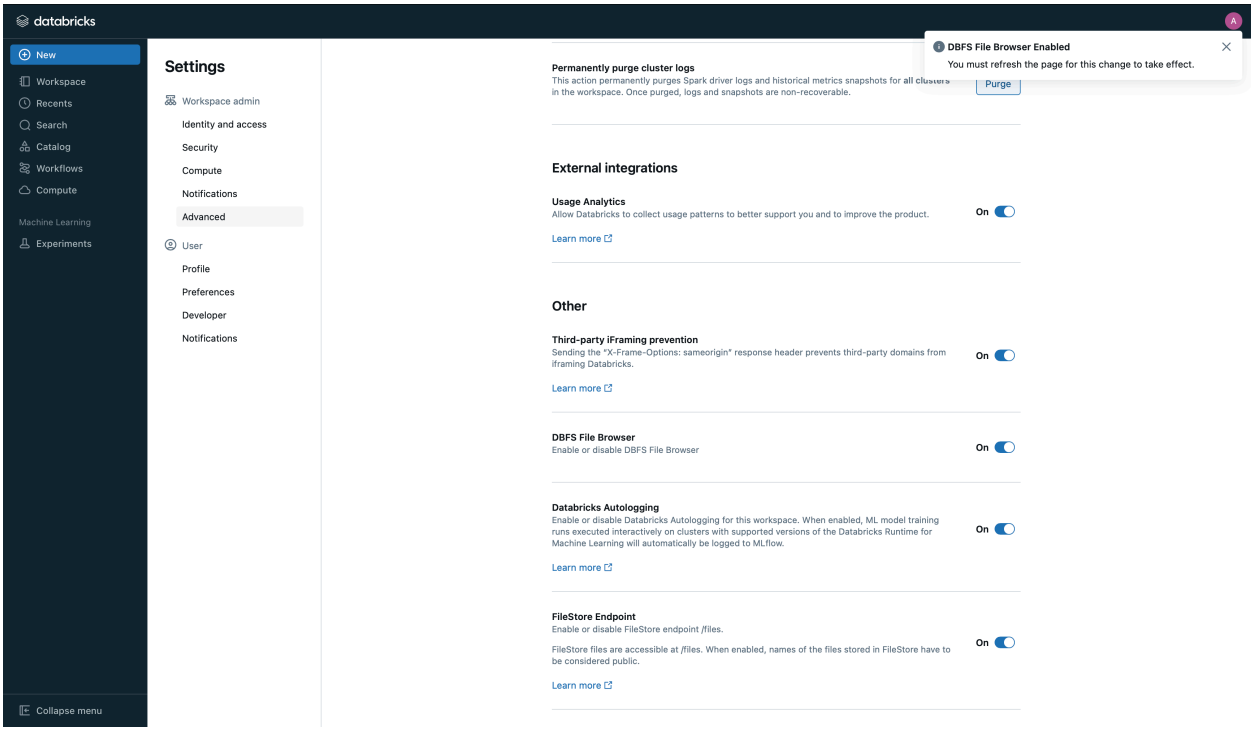
Name: **Anay Abhijit Joshi**

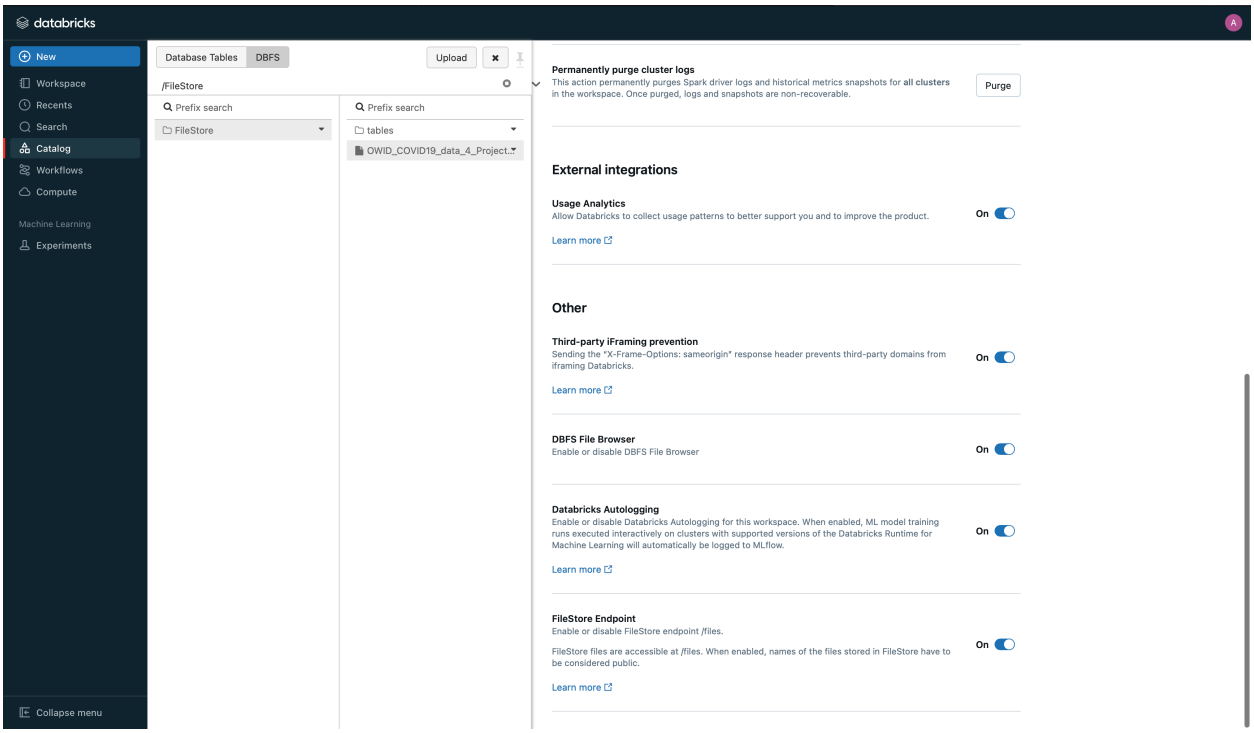
M-ID: **M14391356 (joshi2an)**

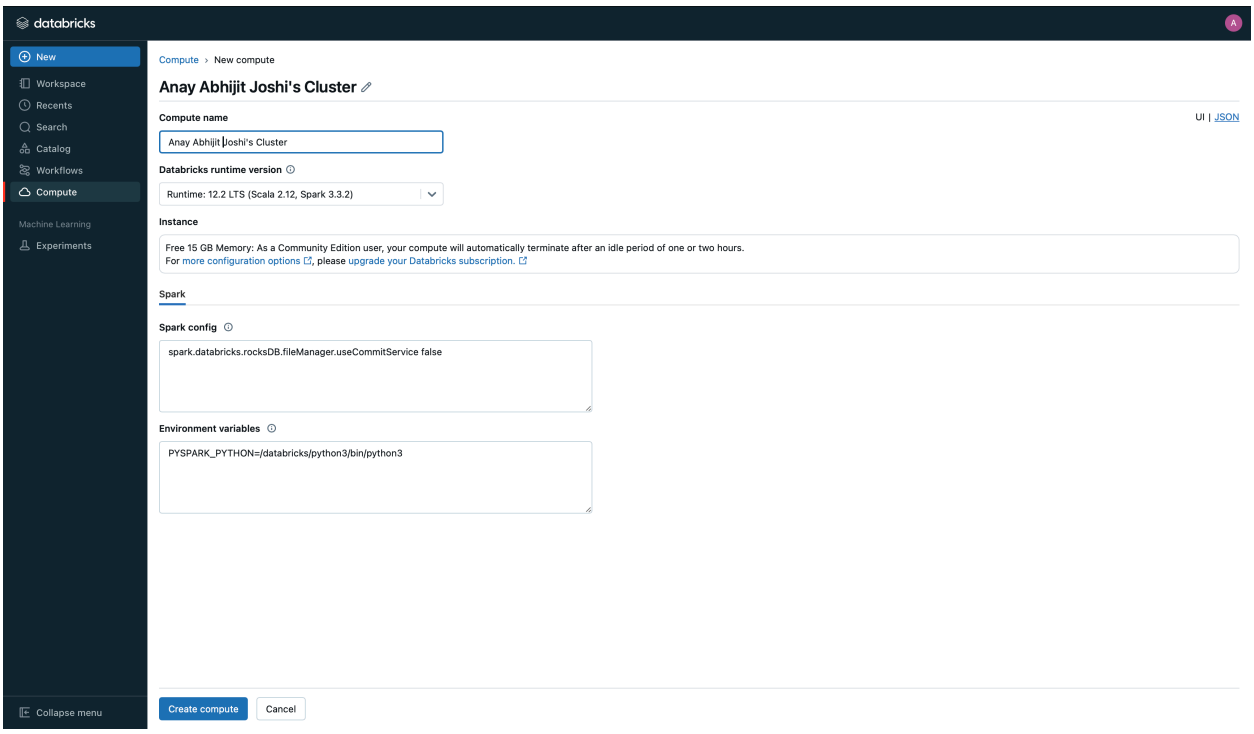
**100% Done ✅**

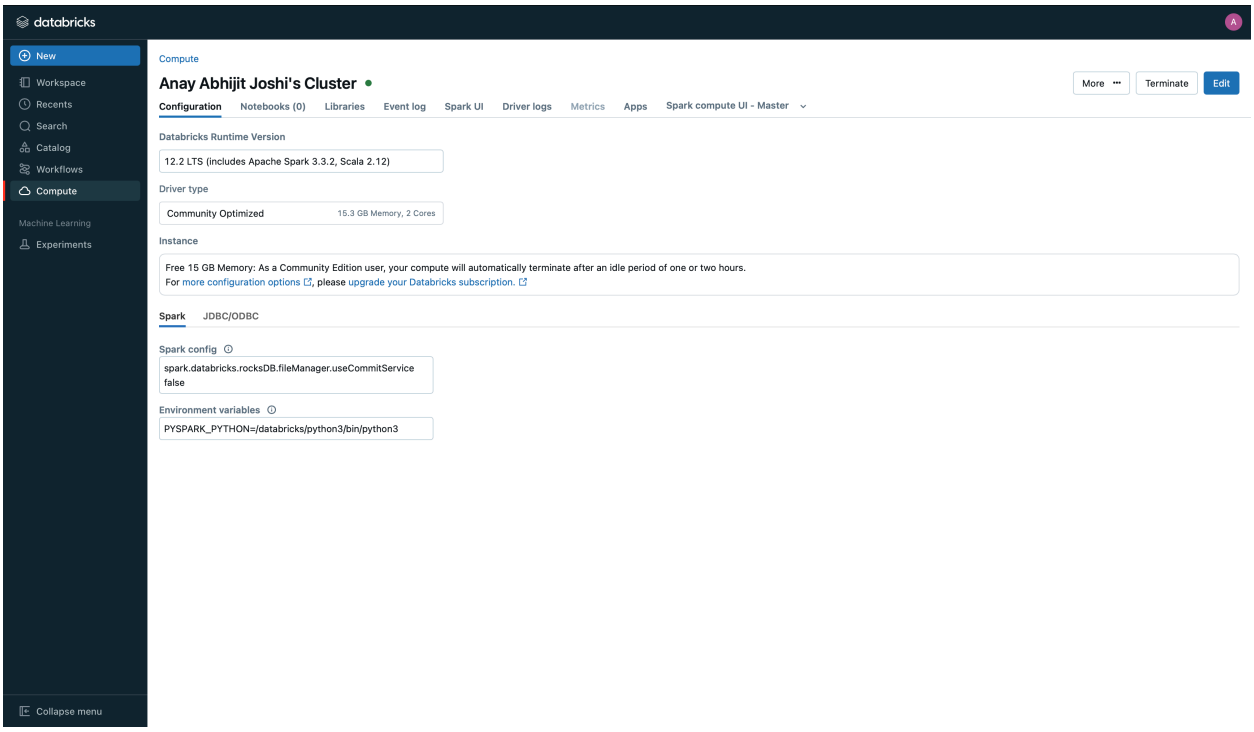
**Code is attached in a Zip Folder and is also available on GitHub:**
**https://github.com/anay-a-joshi/Databricks-COVID19**

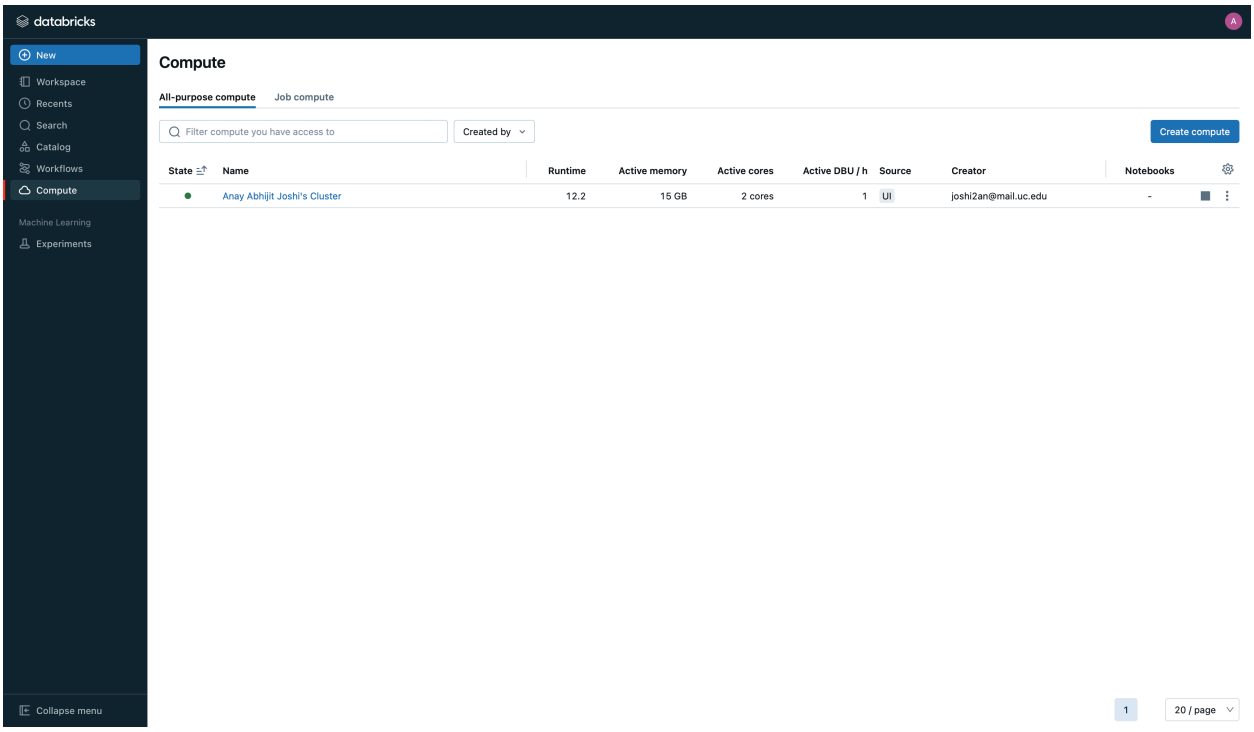
**All the required screenshots are attached below !**

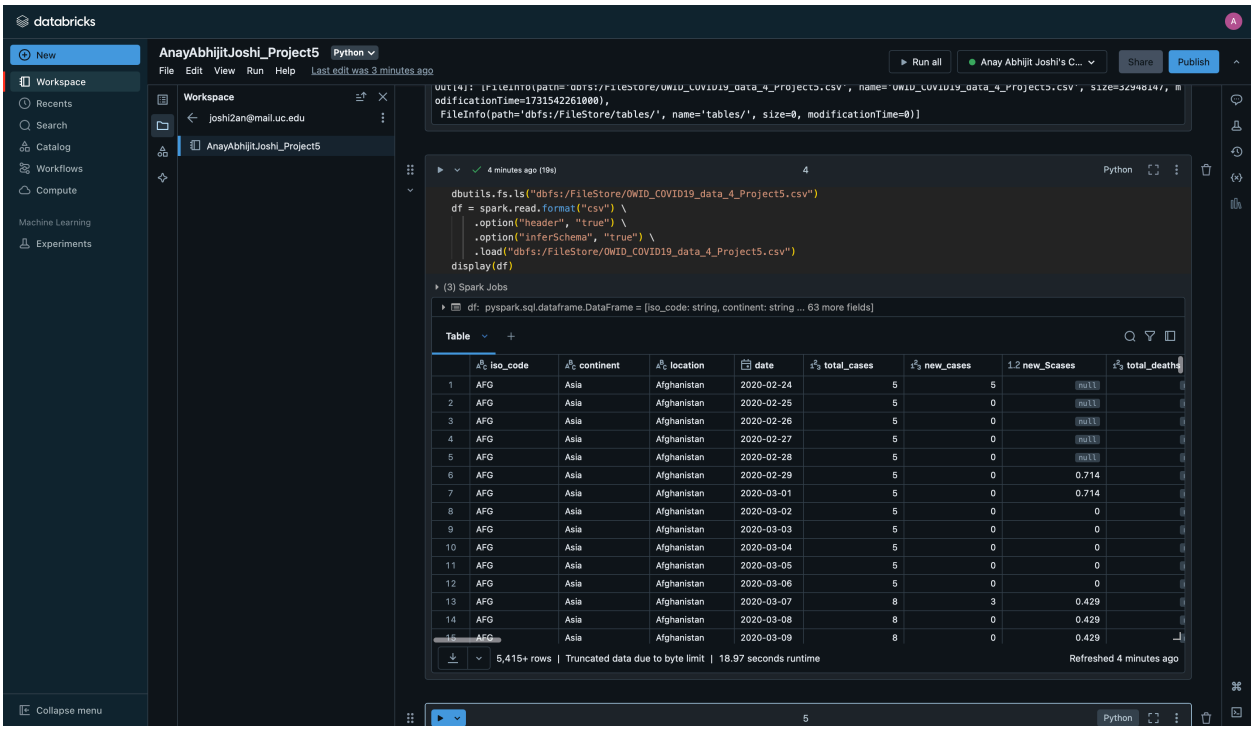


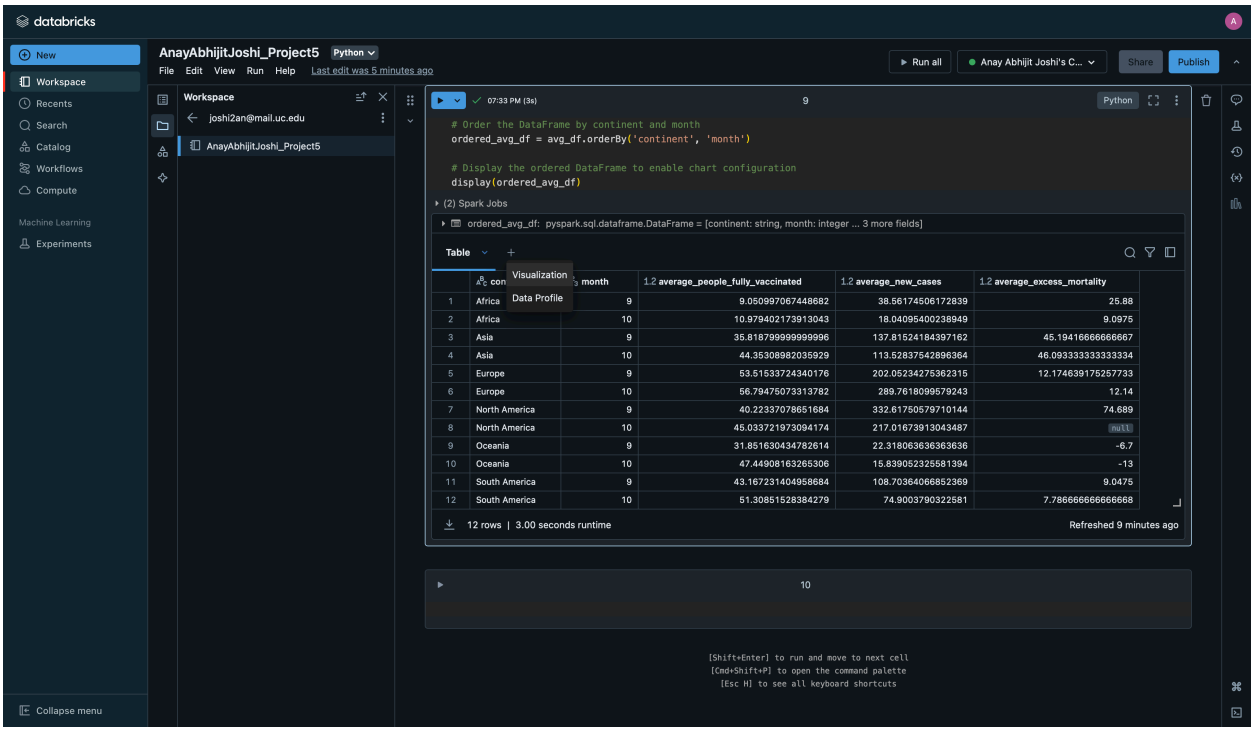

databricks
New
Workspace
Recents
Search
Catalog
Workflows
Compute
Machine Learning
Experiments
Collapse menu
Database Tables
DBFS
Upload
/FileStore
Prefix search
FileStore
tables
OWID_COVID19_data_4_Project...
Permanently purge cluster logs
This action permanently purges Spark driver logs and historical metrics snapshots for all clusters in the workspace. Once purged, logs and snapshots are non-recoverable.
Purge
External integrations
Usage Analytics
Allow Databricks to collect usage patterns to better support you and to improve the product.
On
Learn more
Other
Third-party iFraming prevention
Sending the "X-Frame-Options: sameorigin" response header prevents third-party domains from iframing Databricks.
On
Learn more
DBFS File Browser
Enable or disable DBFS File Browser
On
Databricks Autologging
Enable or disable Databricks Autologging for this workspace. When enabled, ML model training runs executed interactively on clusters with supported versions of the Databricks Runtime for Machine Learning will automatically be logged to MLflow.
On
Learn more
FileStore Endpoint
Enable or disable FileStore endpoint /files.
FileStore files are accessible at /files. When enabled, names of the files stored in FileStore have to be considered public.
On
Learn more

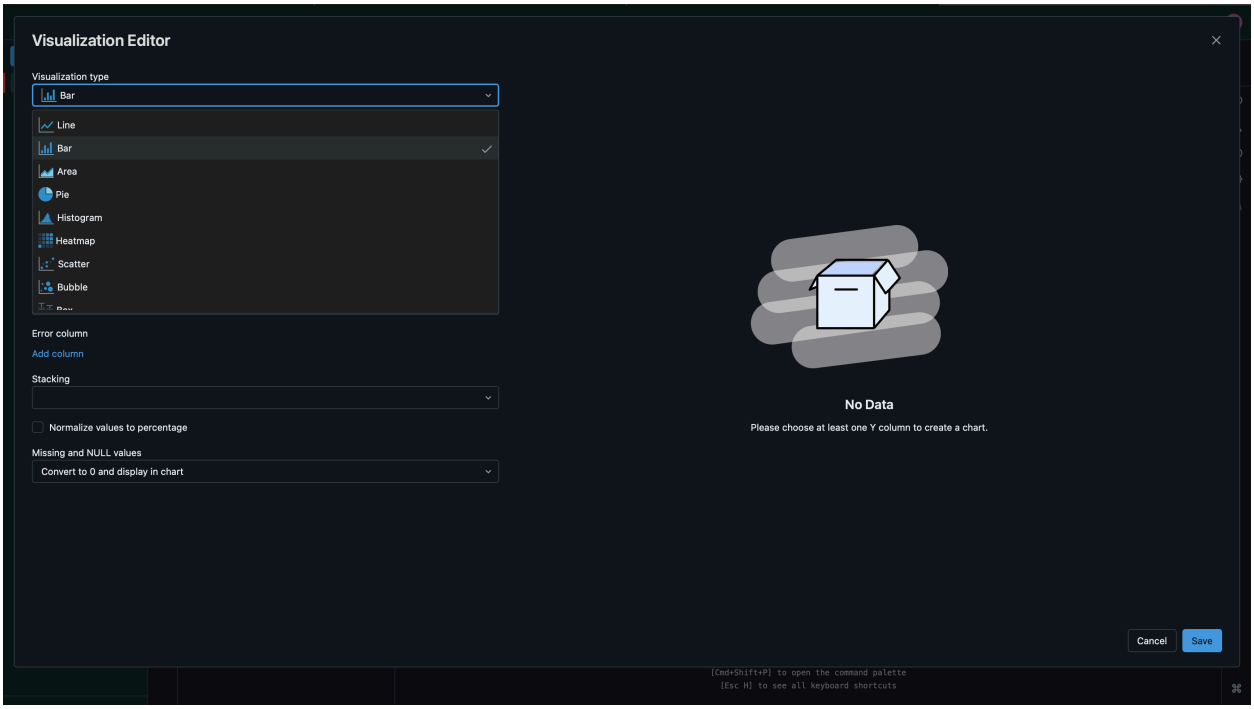

databricks
New
Workspace
Recents
Search
Catalog
Workflows
Compute
Machine Learning
Experiments
Collapse menu
Compute › New compute
Anay Abhijit Joshi's Cluster
Compute name
Anay Abhijit Joshi's Cluster
UI | JSON
Databricks runtime version
Runtime: 12.2 LTS (Scala 2.12, Spark 3.3.2)
Instance
Free 15 GB Memory: As a Community Edition user, your compute will automatically terminate after an idle period of one or two hours.
For more configuration options, please upgrade your Databricks subscription.
Spark
Spark config
spark.databricks.rocksDB.fileManager.useCommitService false
Environment variables
PYSPARK_PYTHON=/databricks/python3/bin/python3
Create compute
Cancel

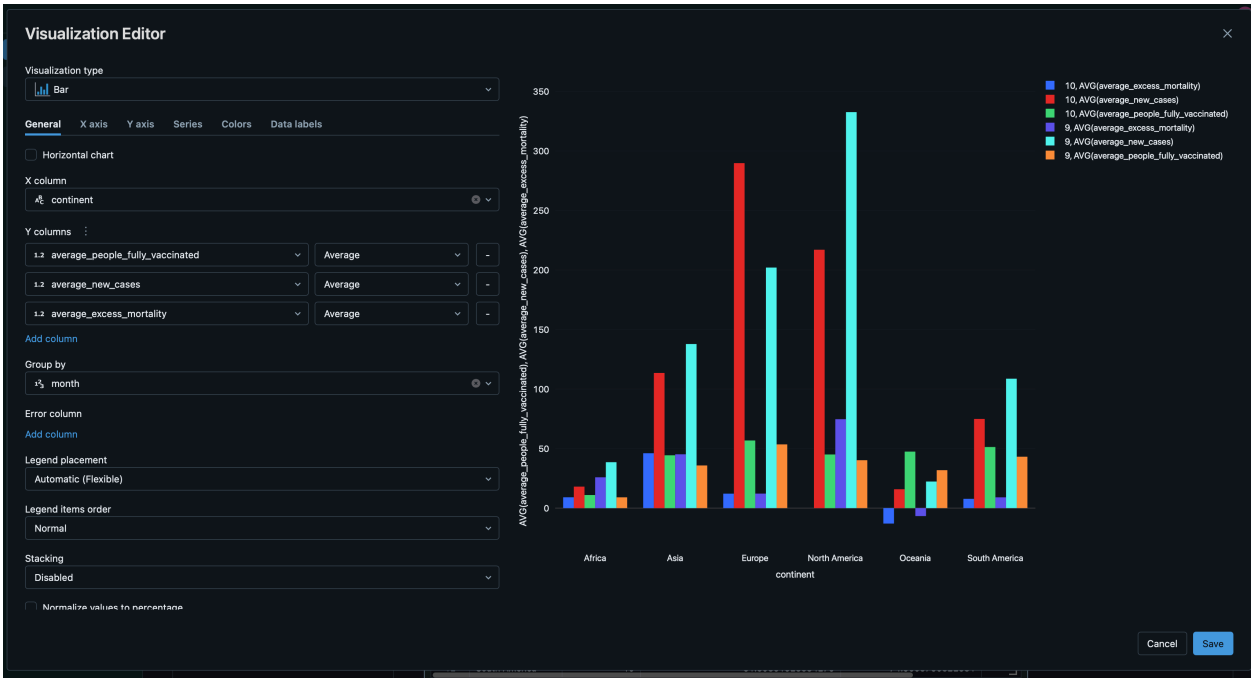

databricks
New
Workspace
Recents
Search
Catalog
Workflows
Compute
Machine Learning
Experiments
Collapse menu
Compute
Anay Abhijit Joshi's Cluster
More
Terminate
Edit
Configuration
Notebooks (0)
Libraries
Event log
Spark UI
Driver logs
Metrics
Apps
Spark compute UI - Master
Databricks Runtime Version
12.2 LTS (includes Apache Spark 3.3.2, Scala 2.12)
Driver type
Community Optimized
15.3 GB Memory, 2 Cores
Instance
Free 15 GB Memory: As a Community Edition user, your compute will automatically terminate after an idle period of one or two hours.
For more configuration options, please upgrade your Databricks subscription.
Spark
JDBC/ODBC
Spark config
spark.databricks.rocksDB.fileManager.useCommitService
false
Environment variables
PYSPARK_PYTHON=/databricks/python3/bin/python3


databricks
New
Workspace
Recents
Search
Catalog
Workflows
Compute
Machine Learning
Experiments
Collapse menu
Compute
All-purpose compute
Job compute
Filter compute you have access to
Created by
Create compute
State
Name
Runtime
Active memory
Active cores
Active DBU / h
Source
Creator
Notebooks
Anay Abhijit Joshi's Cluster
12.2
15 GB
2 cores
1
UI
joshi2an@mail.uc.edu
-
1
20 / page


databricks
AnayAbhijitJoshi_Project5
Python
File Edit View Run Help Last edit was 3 minutes ago
Run all
Anay Abhijit Joshi's C...
Share
Publish
New
Workspace
Recents
Search
Catalog
Workflows
Compute
Machine Learning
Experiments
Collapse menu
joshi2an@mail.uc.edu
AnayAbhijitJoshi_Project5
FileInfo(path='dbfs:/FileStore/tables/', name='tables/', size=0, modificationTime=0)]
4 minutes ago (19s)
4
Python
dbutils.fs.ls("dbfs:/FileStore/OWID_COVID19_data_4_Project5.csv")
df = spark.read.format("csv") \
.option("header", "true") \
.option("inferSchema", "true") \
.load("dbfs:/FileStore/OWID_COVID19_data_4_Project5.csv")
display(df)
(3) Spark Jobs
df: pyspark.sql.dataframe.DataFrame = [iso_code: string, continent: string ... 63 more fields]
Table
iso_code continent location date total_cases new_cases new_Scases total_deaths
1 AFG Asia Afghanistan 2020-02-24 5 5 null
2 AFG Asia Afghanistan 2020-02-25 5 0 null
3 AFG Asia Afghanistan 2020-02-26 5 0 null
4 AFG Asia Afghanistan 2020-02-27 5 0 null
5 AFG Asia Afghanistan 2020-02-28 5 0 null
6 AFG Asia Afghanistan 2020-02-29 5 0 0.714
7 AFG Asia Afghanistan 2020-03-01 5 0 0.714
8 AFG Asia Afghanistan 2020-03-02 5 0 0
9 AFG Asia Afghanistan 2020-03-03 5 0 0
10 AFG Asia Afghanistan 2020-03-04 5 0 0
11 AFG Asia Afghanistan 2020-03-05 5 0 0
12 AFG Asia Afghanistan 2020-03-06 5 0 0
13 AFG Asia Afghanistan 2020-03-07 8 3 0.429
14 AFG Asia Afghanistan 2020-03-08 8 0 0.429
15 AFG Asia Afghanistan 2020-03-09 8 0 0.429
5,415+ rows | Truncated data due to byte limit | 18.97 seconds runtime
Refreshed 4 minutes ago
5
Python


databricks
AnayAbhijitJoshi_Project5
Python
File Edit View Run Help Last edit was 5 minutes ago
Run all
Anay Abhijit Joshi's C...
Share
Publish
New
Workspace
Recents
Search
Catalog
Workflows
Compute
Machine Learning
Experiments
Collapse menu
joshi2an@mail.uc.edu
AnayAbhijitJoshi_Project5
07:33 PM (3s)
9
Python
# Order the DataFrame by continent and month
ordered_avg_df = avg_df.orderBy('continent', 'month')
# Display the ordered DataFrame to enable chart configuration
display(ordered_avg_df)
(2) Spark Jobs
ordered_avg_df: pyspark.sql.dataframe.DataFrame = [continent: string, month: integer ... 3 more fields]
Table
Visualization
Data Profile
con month 1.2 average_people_fully_vaccinated 1.2 average_new_cases 1.2 average_excess_mortality
1 Africa 9 9.050997067448682 38.56174506172839 25.88
2 Africa 10 10.979402173913043 18.04095400238949 9.0975
3 Asia 9 35.818799999999996 137.81524184397162 45.19416666666667
4 Asia 10 44.35308982035929 113.52837542896364 46.093333333333334
5 Europe 9 53.51533724340176 202.05234275362315 12.174639175257733
6 Europe 10 56.79475073313782 289.7618099579243 12.14
7 North America 9 40.22337078651684 332.61750579710144 74.689
8 North America 10 45.033721973094174 217.01673913043487 null
9 Oceania 9 31.851630434782614 22.318063636363636 -6.7
10 Oceania 10 47.44908163265306 15.839052325581394 -13
11 South America 9 43.167231404958684 108.70364066852369 9.0475
12 South America 10 51.30851528384279 74.9003790322581 7.786666666666668
12 rows | 3.00 seconds runtime
Refreshed 9 minutes ago
10
[Shift+Enter] to run and move to next cell
[Cmd+Shift+P] to open the command palette
[Esc H] to see all keyboard shortcuts


Visualization Editor
Visualization type
Bar
Line
Bar
Area
Pie
Histogram
Heatmap
Scatter
Bubble
Error column
Add column
Stacking
Normalize values to percentage
Missing and NULL values
Convert to 0 and display in chart
No Data
Please choose at least one Y column to create a chart.
Cancel
Save
[Cmd+Shift+P] to open the command palette
[Esc H] to see all keyboard shortcuts


Visualization Editor
Visualization type
Bar
General
X axis
Y axis
Series
Colors
Data labels
Horizontal chart
X column
continent
Y columns
average_people_fully_vaccinated
Average
average_new_cases
Average
average_excess_mortality
Average
Add column
Group by
month
Error column
Add column
Legend placement
Automatic (Flexible)
Legend items order
Normal
Stacking
Disabled
Normalize values to percentage
AVG(average_people_fully_vaccinated), AVG(average_new_cases), AVG(average_excess_mortality)
350
300
250
200
150
100
50
0
Africa
Asia
Europe
North America
Oceania
South America
continent
10, AVG(average_excess_mortality)
10, AVG(average_new_cases)
10, AVG(average_people_fully_vaccinated)
9, AVG(average_excess_mortality)
9, AVG(average_new_cases)
9, AVG(average_people_fully_vaccinated)
Cancel
Save

# STEP 10

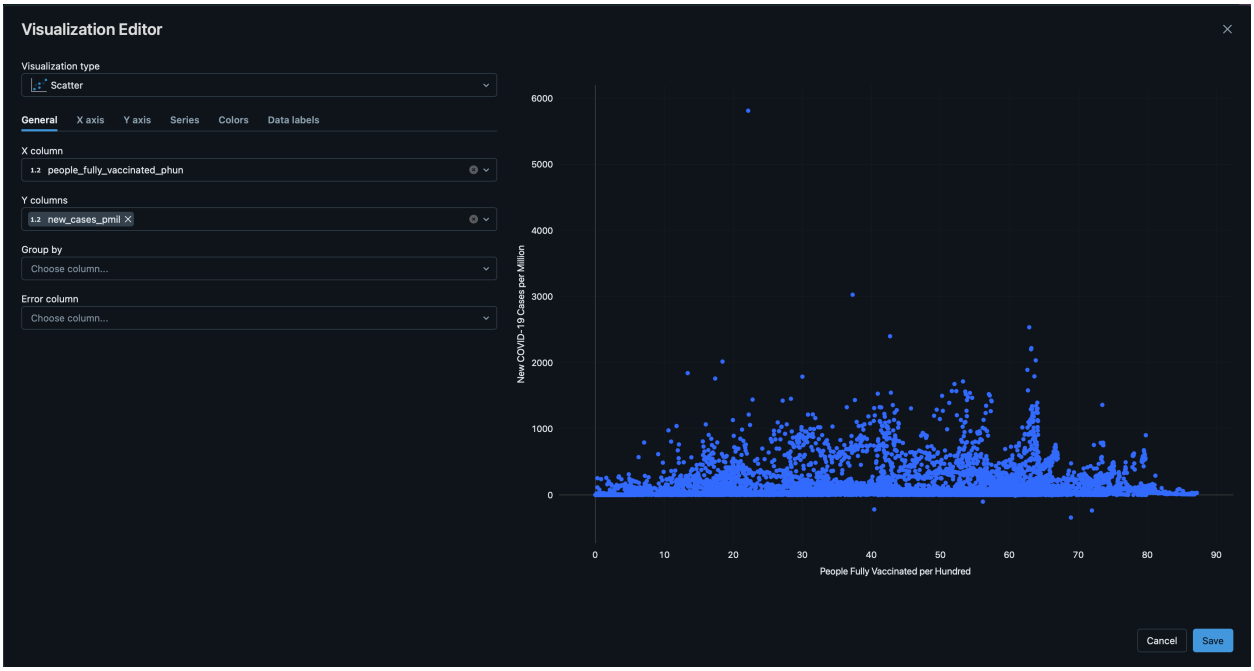
Based on our analysis, there appears to be a moderate positive association between COVID-19 vaccination rates and new cases per million, with a correlation coefficient of approximately **0.52**. This finding is somewhat surprising, as we might intuitively expect higher vaccination rates to correlate with lower case rates. However, this positive correlation could be influenced by several factors, such as differing rates of variant transmission, testing availability, or case reporting practices across regions.

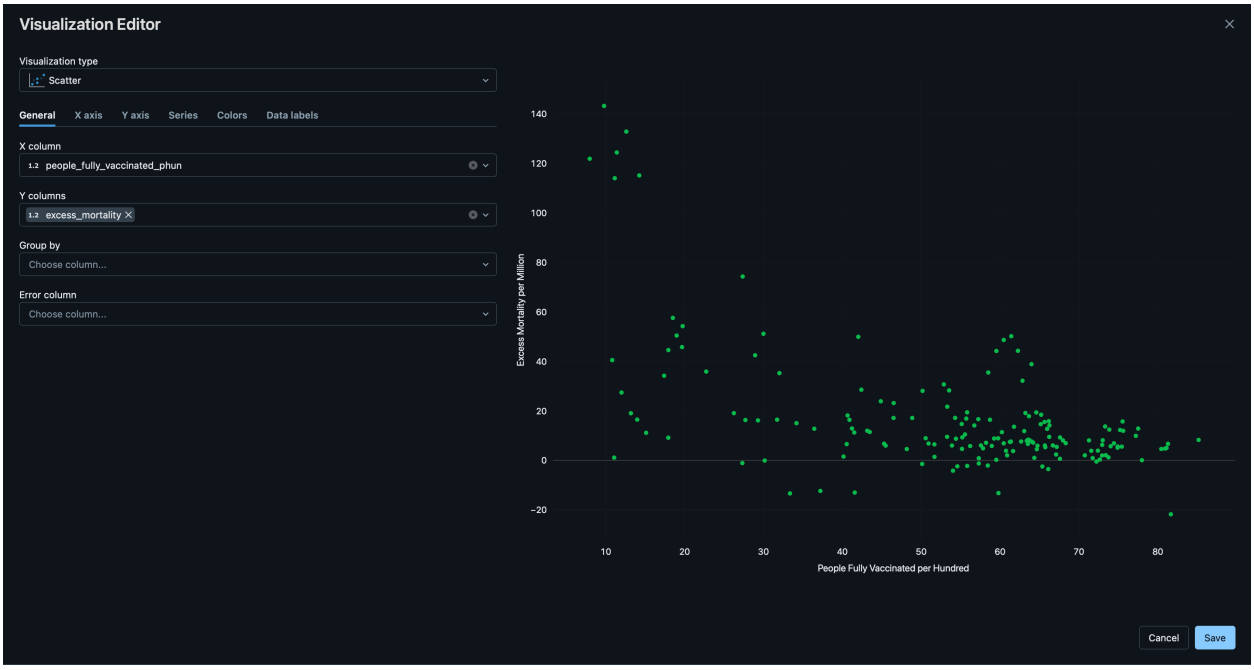
Looking at the summary statistics, we found that the average vaccination rate for September and October 2021 was **42.59%** fully vaccinated individuals per hundred, with a standard deviation of **24.19**. The mean new cases per million was **136.59**, indicating continued viral transmission even in regions with relatively high vaccination coverage. Meanwhile, the mean excess mortality was **17.37** per million, with a large standard deviation of **25.69**, suggesting that mortality outcomes varied widely across locations.

Given these insights, while vaccination efforts have likely contributed to reducing severe outcomes and mortality overall, the data suggests that vaccination alone might not be enough to reduce case numbers uniformly. Other factors, including public health interventions and population density, likely play significant roles in influencing case rates and mortality outcomes.

## *EXTRA*

In my analysis, I observed that COVID-19 vaccination rates did not exhibit a straightforward negative correlation with new cases per million, as I might initially have expected. To support the obtained results and my analysis, I used regression and plotted its results with scatter plots, representing the relationships between vaccination rates (in people per hundred) and new cases (per million) as well as excess mortality (per million). The regression results, while showing some degree of association, suggest that higher vaccination rates alone do not necessarily correlate with significantly lower case numbers or uniformly reduced excess mortality. My correlation and scatter plots display a moderate positive association, which, upon further inspection, likely reflects the complex interplay of other factors, such as variant spread, testing availability, and localized health policies. Though vaccines are crucial in reducing severe illness and mortality, my data analysis reinforces that vaccination may need to be complemented by other public health interventions to effectively curb case rates and excess mortality on a broader scale.


Visualization Editor
Visualization type
Scatter
General
X axis
Y axis
Series
Colors
Data labels
X column
1.2 people_fully_vaccinated_phun
Y columns
1.2 new_cases_pmil
Group by
Choose column...
Error column
Choose column...
New COVID-19 Cases per Million
6000
5000
4000
3000
2000
1000
0
0
10
20
30
40
50
60
70
80
90
People Fully Vaccinated per Hundred
Cancel
Save


Visualization Editor
Visualization type
Scatter
General
X axis
Y axis
Series
Colors
Data labels
X column
1.2 people_fully_vaccinated_phun
Y columns
1.2 excess_mortality
Group by
Choose column...
Error column
Choose column...
Excess Mortality per Million
140
120
100
80
60
40
20
0
−20
10
20
30
40
50
60
70
80
People Fully Vaccinated per Hundred
Cancel
Save