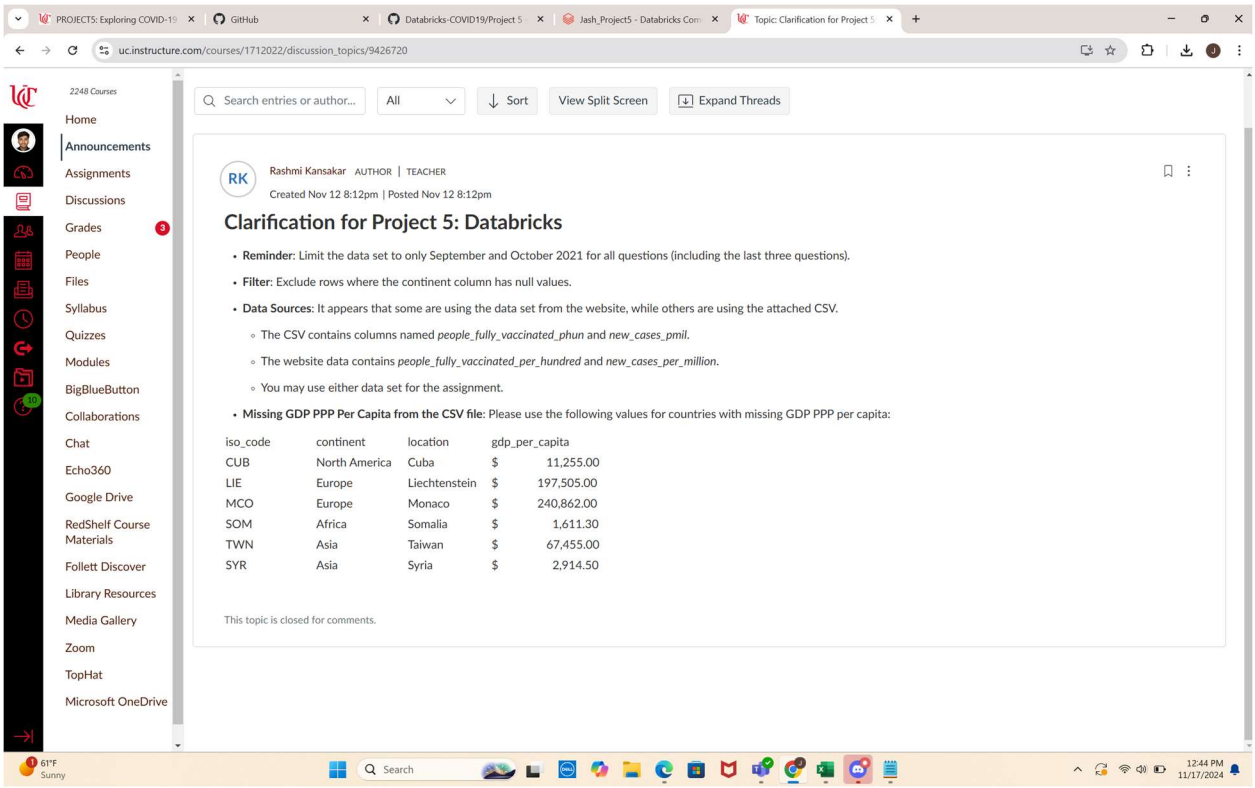
Project 5

Name: Jash Dedhia

M-ID: M15047151 (dedhiaja)

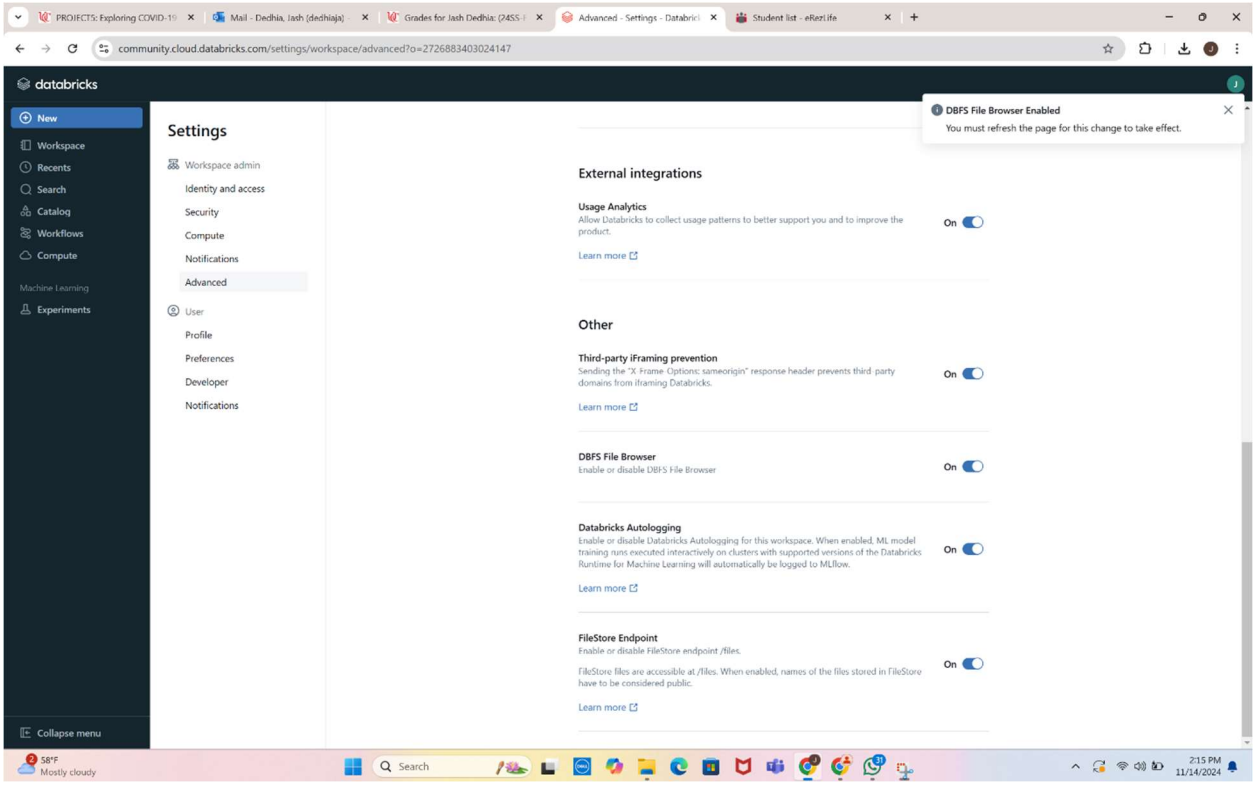
Date: November 17, 2024

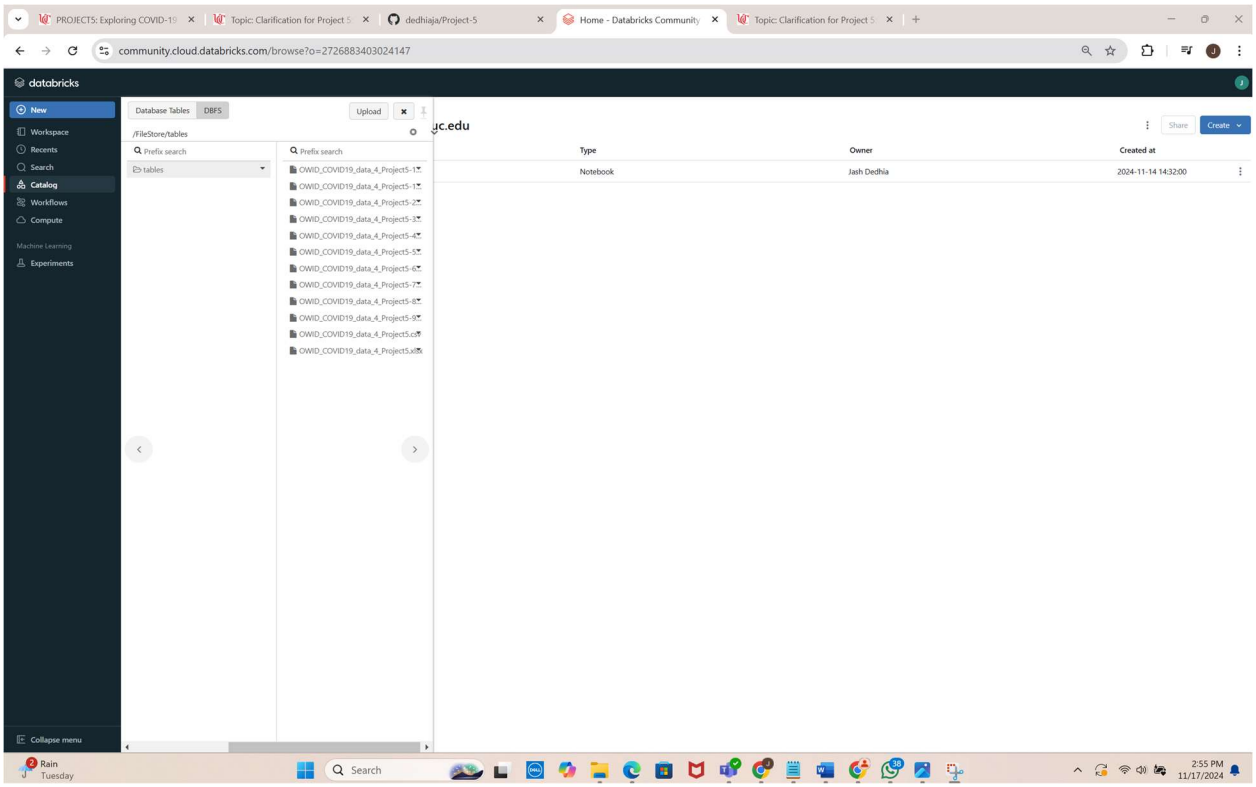
Github Link: https://github.com/dedhiaja/Project-5

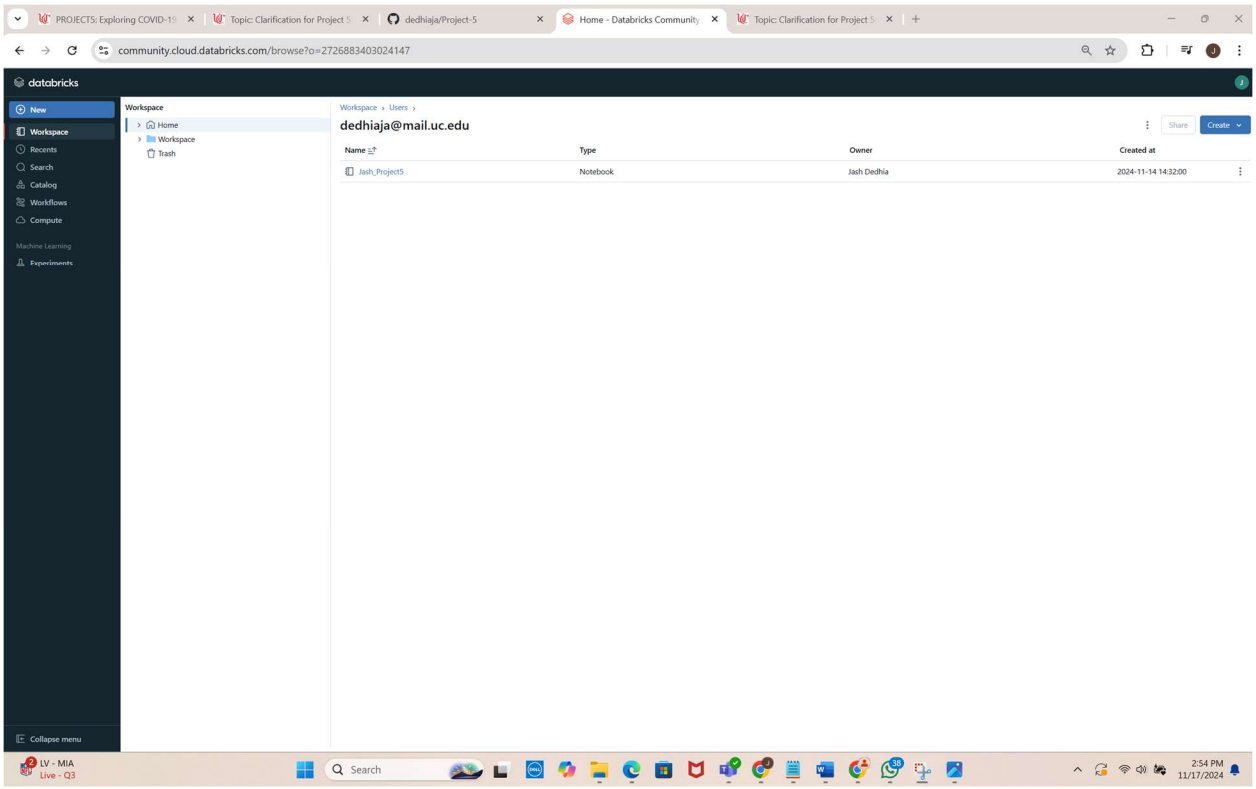


Following professor's instruction that for all the questions the data set has to be limited to September and October 2021. (Screenshot attached above)

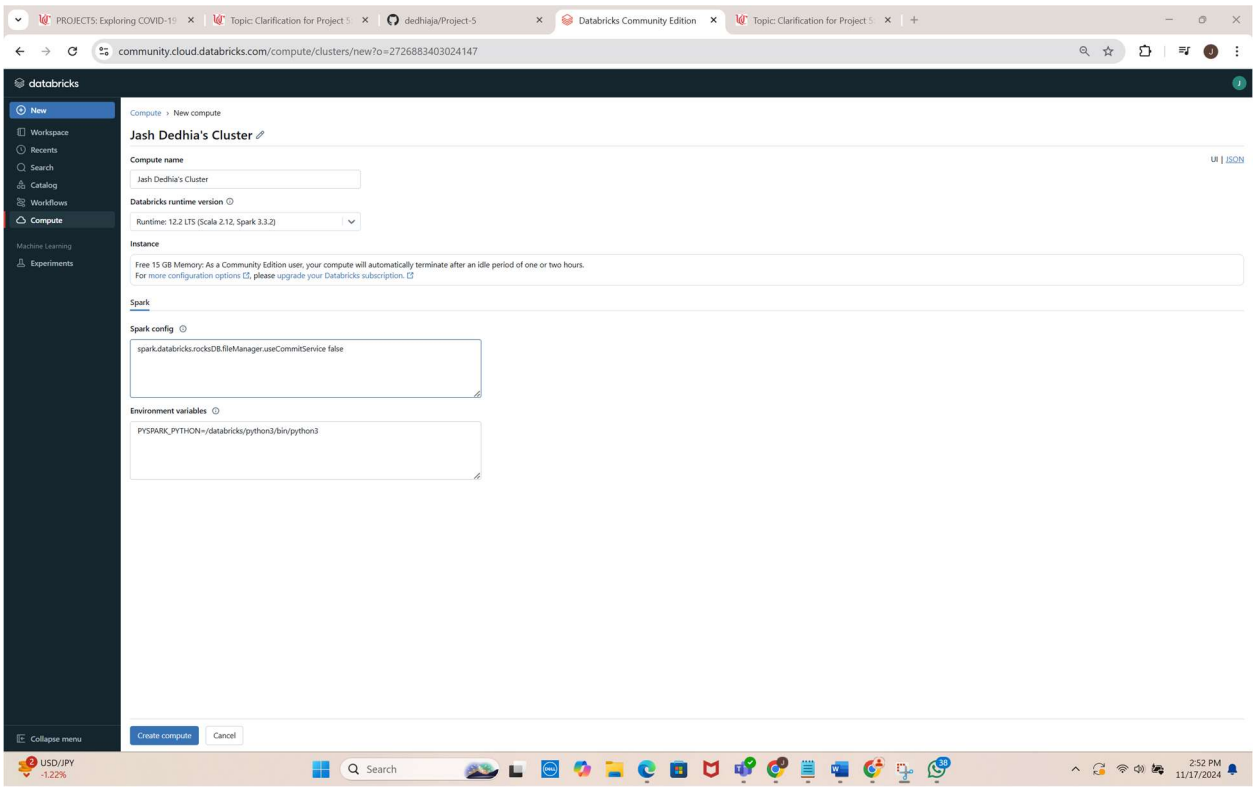
1.


PROJECT5: Exploring COVID-19
Mail - Dedhia, Jash (dedhiaja)
Grades for Jash Dedhia: (24SS-I
Advanced - Settings - Databrick
Student list - eRezLife
community.cloud.databricks.com/settings/workspace/advanced?o=2726883403024147
databricks
New
Workspace
Recents
Search
Catalog
Workflows
Compute
Machine Learning
Experiments
Collapse menu
Settings
Workspace admin
Identity and access
Security
Compute
Notifications
Advanced
User
Profile
Preferences
Developer
Notifications
DBFS File Browser Enabled
You must refresh the page for this change to take effect.
External integrations
Usage Analytics
Allow Databricks to collect usage patterns to better support you and to improve the product.
Learn more
On
Other
Third-party iFraming prevention
Sending the "X-Frame-Options: sameorigin" response header prevents third-party domains from iframing Databricks.
Learn more
On
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.
Learn more
On
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.
Learn more
On
58°F
Mostly cloudy
Search
2:15 PM
11/14/2024

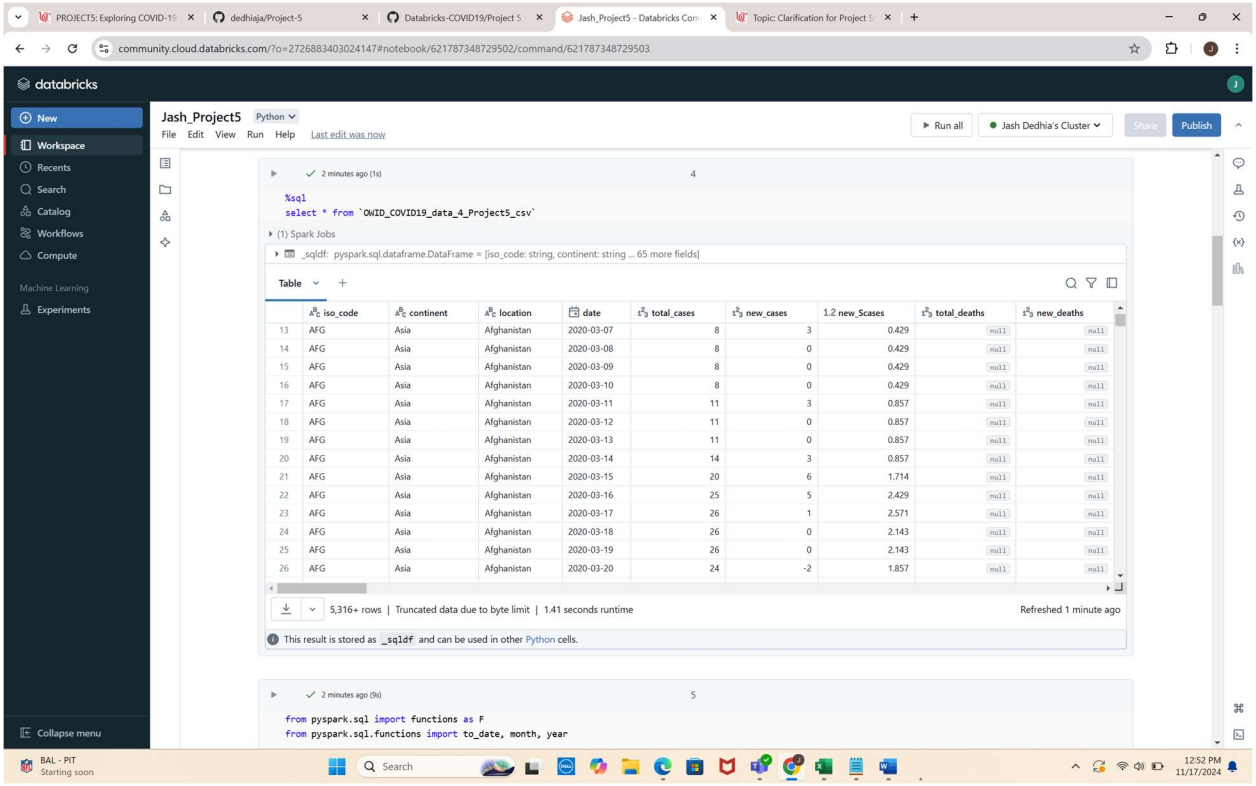

PROJECT5: Exploring COVID-19
Topic: Clarification for Project 5
dedhiaja/Project-5
Home - Databricks Community
Topic: Clarification for Project 5
community.cloud.databricks.com/browse?o=2726883403024147
databricks
New
Workspace
Recents
Search
Catalog
Workflows
Compute
Machine Learning
Experiments
Collapse menu
Database Tables
DBFS
Upload
/FileStore/tables
Prefix search
tables
OWID_COVID19_data_4_Project5-1
OWID_COVID19_data_4_Project5-1
OWID_COVID19_data_4_Project5-2
OWID_COVID19_data_4_Project5-3
OWID_COVID19_data_4_Project5-4
OWID_COVID19_data_4_Project5-5
OWID_COVID19_data_4_Project5-6
OWID_COVID19_data_4_Project5-7
OWID_COVID19_data_4_Project5-8
OWID_COVID19_data_4_Project5-9
OWID_COVID19_data_4_Project5.cs
OWID_COVID19_data_4_Project5.xls
yc.edu
Share
Create
Type
Owner
Created at
Notebook
Jash Dedhia
2024-11-14 14:32:00
Rain
Tuesday
Search
2:55 PM
11/17/2024


PROJECT5: Exploring COVID-19
Topic: Clarification for Project 5
dedhiaja/Project-5
Home - Databricks Community
Topic: Clarification for Project 5
community.cloud.databricks.com/browse?o=2726883403024147
databricks
New
Workspace
Recents
Search
Catalog
Workflows
Compute
Machine Learning
Experiments
Collapse menu
Workspace
Home
Workspace
Trash
Workspace › Users ›
dedhiaja@mail.uc.edu
Share
Create
Name
Type
Owner
Created at
Jash_Project5
Notebook
Jash Dedhia
2024-11-14 14:32:00
LV - MIA
Live - Q3
Search
2:54 PM
11/17/2024

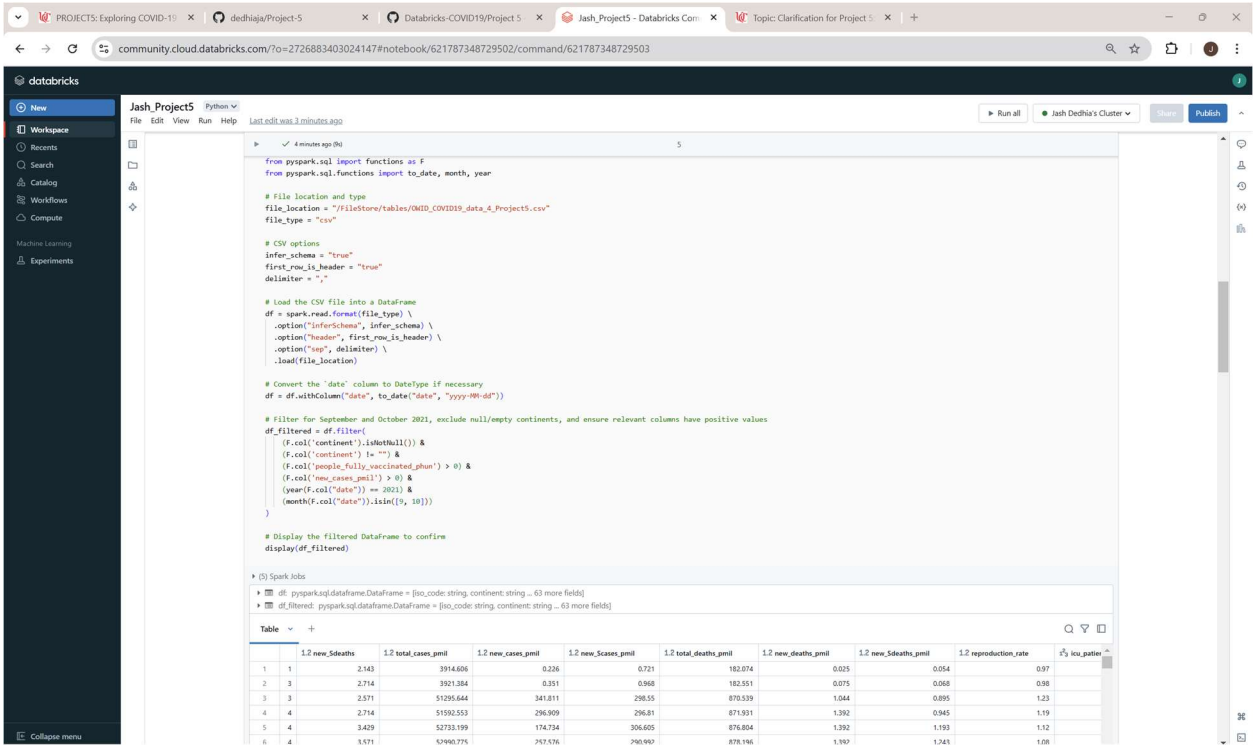
2.


PROJECT5: Exploring COVID-19
Topic: Clarification for Project 5
dedhiaja/Project-5
Databricks Community Edition
Topic: Clarification for Project 5
community.cloud.databricks.com/compute/clusters/new?o=2726883403024147
databricks
New
Workspace
Recents
Search
Catalog
Workflows
Compute
Machine Learning
Experiments
Collapse menu
Compute › New compute
Jash Dedhia's Cluster
UI | JSON
Compute name
Jash Dedhia's Cluster
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
USD/JPY
-1.22%
Search
2:52 PM
11/17/2024

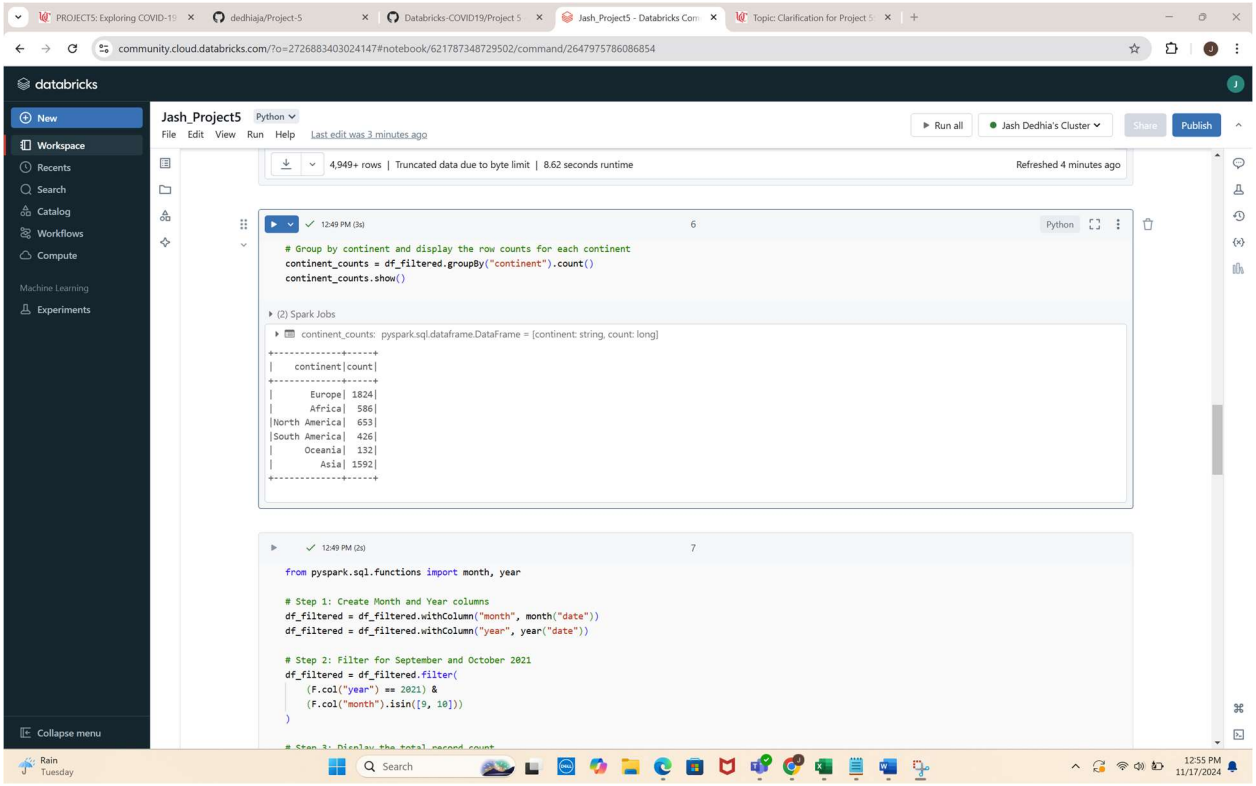
This shows that the entire file is loaded:



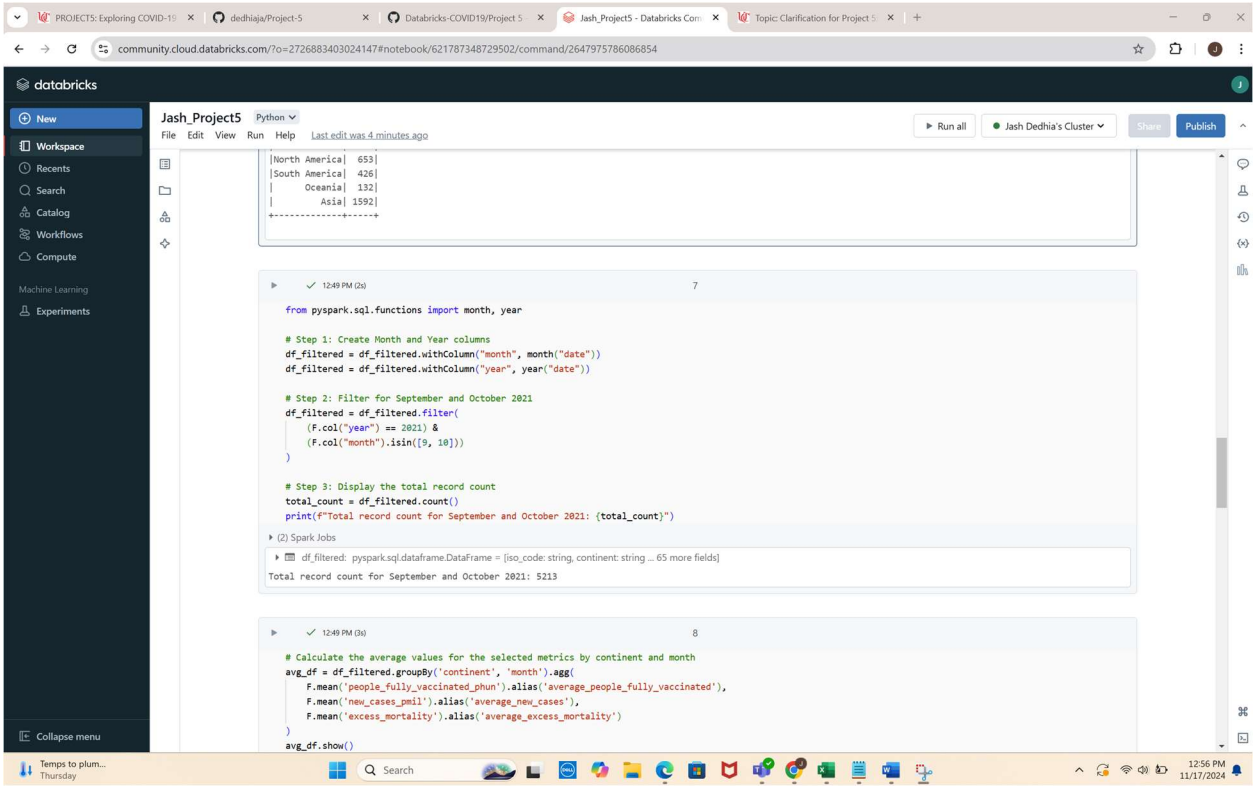
This shows that the records are now just narrowed down for September and October 2021:



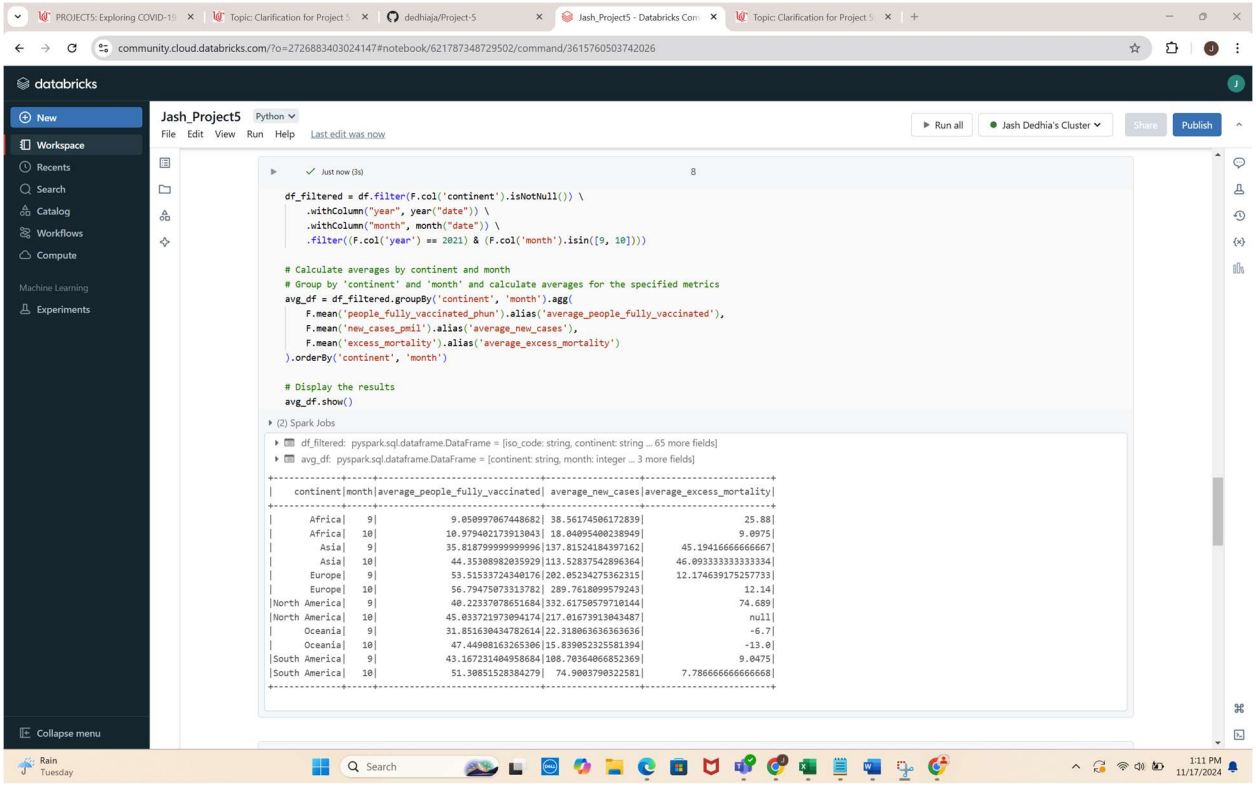
3.


Jash_Project5
Python
File Edit View Run Help
Last edit was 3 minutes ago
Run all
Jash Dedhia's Cluster
Share
Publish
4,949+ rows | Truncated data due to byte limit | 8.62 seconds runtime
Refreshed 4 minutes ago
12:49 PM (3s)
6
# Group by continent and display the row counts for each continent
continent_counts = df_filtered.groupBy("continent").count()
continent_counts.show()
(2) Spark Jobs
continent_counts: pyspark.sql.dataframe.DataFrame = [continent: string, count: long]
+-------------+-----+
| continent|count|
+-------------+-----+
| Europe| 1824|
| Africa| 586|
|North America| 653|
|South America| 426|
| Oceania| 132|
| Asia| 1592|
+-------------+-----+
12:49 PM (2s)
7
from pyspark.sql.functions import month, year
# Step 1: Create Month and Year columns
df_filtered = df_filtered.withColumn("month", month("date"))
df_filtered = df_filtered.withColumn("year", year("date"))
# Step 2: Filter for September and October 2021
df_filtered = df_filtered.filter(
(F.col("year") == 2021) &
(F.col("month").isin([9, 10]))
)
Collapse menu

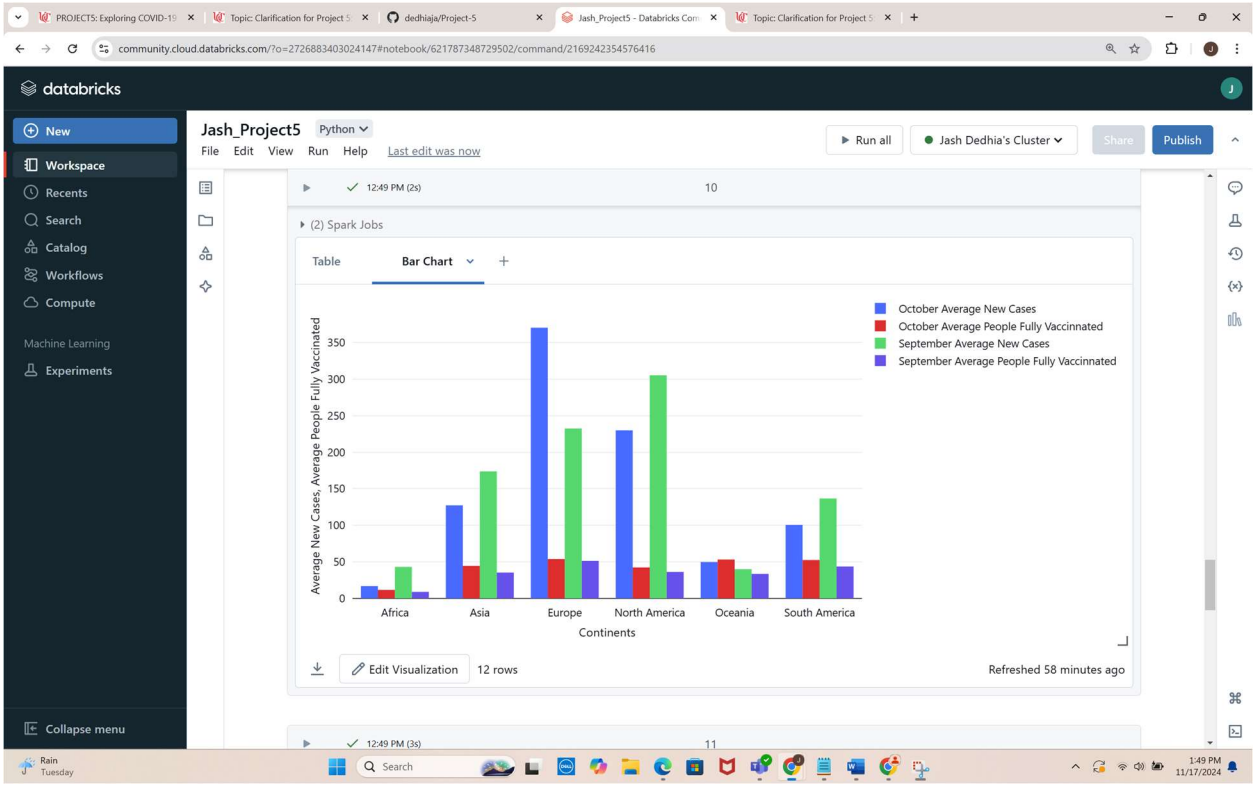
4.


Jash_Project5
Python
File Edit View Run Help
Last edit was 4 minutes ago
Run all
Jash Dedhia's Cluster
Share
Publish
|North America| 653|
|South America| 426|
| Oceania| 132|
| Asia| 1592|
+-------------+-----+
12:49 PM (2s)
7
from pyspark.sql.functions import month, year
# Step 1: Create Month and Year columns
df_filtered = df_filtered.withColumn("month", month("date"))
df_filtered = df_filtered.withColumn("year", year("date"))
# Step 2: Filter for September and October 2021
df_filtered = df_filtered.filter(
(F.col("year") == 2021) &
(F.col("month").isin([9, 10]))
)
# Step 3: Display the total record count
total_count = df_filtered.count()
print(f"Total record count for September and October 2021: {total_count}")
(2) Spark Jobs
df_filtered: pyspark.sql.dataframe.DataFrame = [iso_code: string, continent: string ... 65 more fields]
Total record count for September and October 2021: 5213
12:49 PM (3s)
8
# Calculate the average values for the selected metrics by continent and month
avg_df = df_filtered.groupBy('continent', 'month').agg(
F.mean('people_fully_vaccinated_phun').alias('average_people_fully_vaccinated'),
F.mean('new_cases_pmil').alias('average_new_cases'),
F.mean('excess_mortality').alias('average_excess_mortality')
)
avg_df.show()
Collapse menu

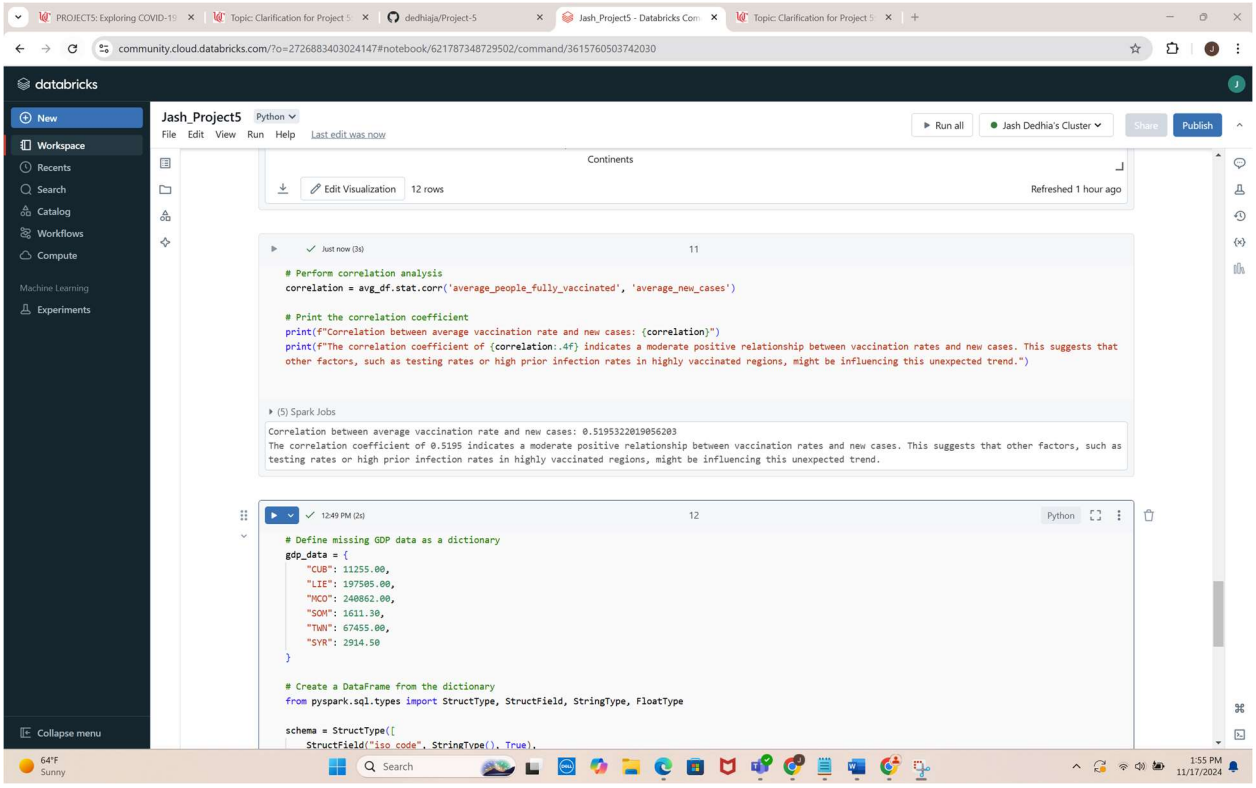
5.


PROJECT5: Exploring COVID-19
Topic: Clarification for Project 5
dedhiaja/Project-5
Jash_Project5 - Databricks Com
Topic: Clarification for Project 5
community.cloud.databricks.com/?o=2726883403024147#notebook/621787348729502/command/3615760503742026
databricks
New
Workspace
Recents
Search
Catalog
Workflows
Compute
Machine Learning
Experiments
Collapse menu
Jash_Project5
Python
File Edit View Run Help
Last edit was now
Run all
Jash Dedhia's Cluster
Share
Publish
Just now (3s)
8
df_filtered = df.filter(F.col('continent').isNotNull()) \
.withColumn("year", year("date")) \
.withColumn("month", month("date")) \
.filter((F.col('year') == 2021) & (F.col('month').isin([9, 10])))
# Calculate averages by continent and month
# Group by 'continent' and 'month' and calculate averages for the specified metrics
avg_df = df_filtered.groupBy('continent', 'month').agg(
F.mean('people_fully_vaccinated_phun').alias('average_people_fully_vaccinated'),
F.mean('new_cases_pmil').alias('average_new_cases'),
F.mean('excess_mortality').alias('average_excess_mortality')
).orderBy('continent', 'month')
# Display the results
avg_df.show()
(2) Spark Jobs
df_filtered: pyspark.sql.dataframe.DataFrame = [iso_code: string, continent: string ... 65 more fields]
avg_df: pyspark.sql.dataframe.DataFrame = [continent: string, month: integer ... 3 more fields]
continent month average_people_fully_vaccinated average_new_cases average_excess_mortality
Africa 9 9.050997067448682 38.56174506172839 25.88
Africa 10 10.979402173913043 18.04095400238949 9.0975
Asia 9 35.81879999999996 137.81524184397162 45.19416666666667
Asia 10 44.35308982035929 113.52837542896364 46.09333333333334
Europe 9 53.51533724340176 202.05234275362315 12.174639175257733
Europe 10 56.79475073313782 289.7618099579243 12.14
North America 9 40.22337078651684 332.61750579710144 74.689
North America 10 45.03372197309417 217.01673913043487 null
Oceania 9 31.851630434782614 22.318063636363636 -6.7
Oceania 10 47.44908163265306 15.839052325581394 -13.0
South America 9 43.16723140495868 108.70364066852369 9.0475
South America 10 51.30851528384279 74.90037903225 81 7.786666666666668

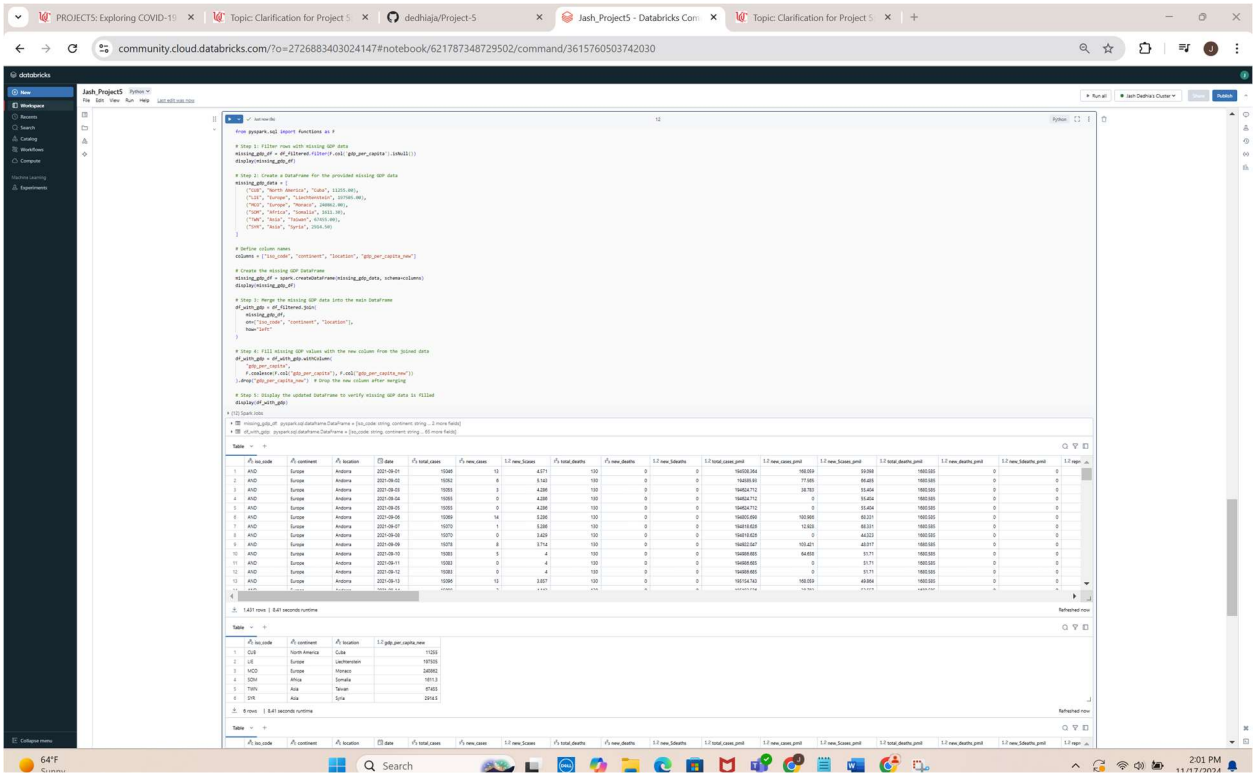
6.


PROJECT5: Exploring COVID-19
Topic: Clarification for Project 5
dedhiaja/Project-5
Jash_Project5 - Databricks Com
Topic: Clarification for Project 5
community.cloud.databricks.com/?o=2726883403024147#notebook/621787348729502/command/2169242354576416
databricks
New
Workspace
Recents
Search
Catalog
Workflows
Compute
Machine Learning
Experiments
Collapse menu
Jash_Project5
Python
File Edit View Run Help
Last edit was now
Run all
Jash Dedhia's Cluster
Share
Publish
12:49 PM (2s)
10
(2) Spark Jobs
Table
Bar Chart
October Average New Cases
October Average People Fully Vaccinnated
September Average New Cases
September Average People Fully Vaccinnated
Average New Cases, Average People Fully Vaccinated
0
50
100
150
200
250
300
350
Africa
Asia
Europe
North America
Oceania
South America
Continents
Edit Visualization
12 rows
Refreshed 58 minutes ago
12:49 PM (3s)
11

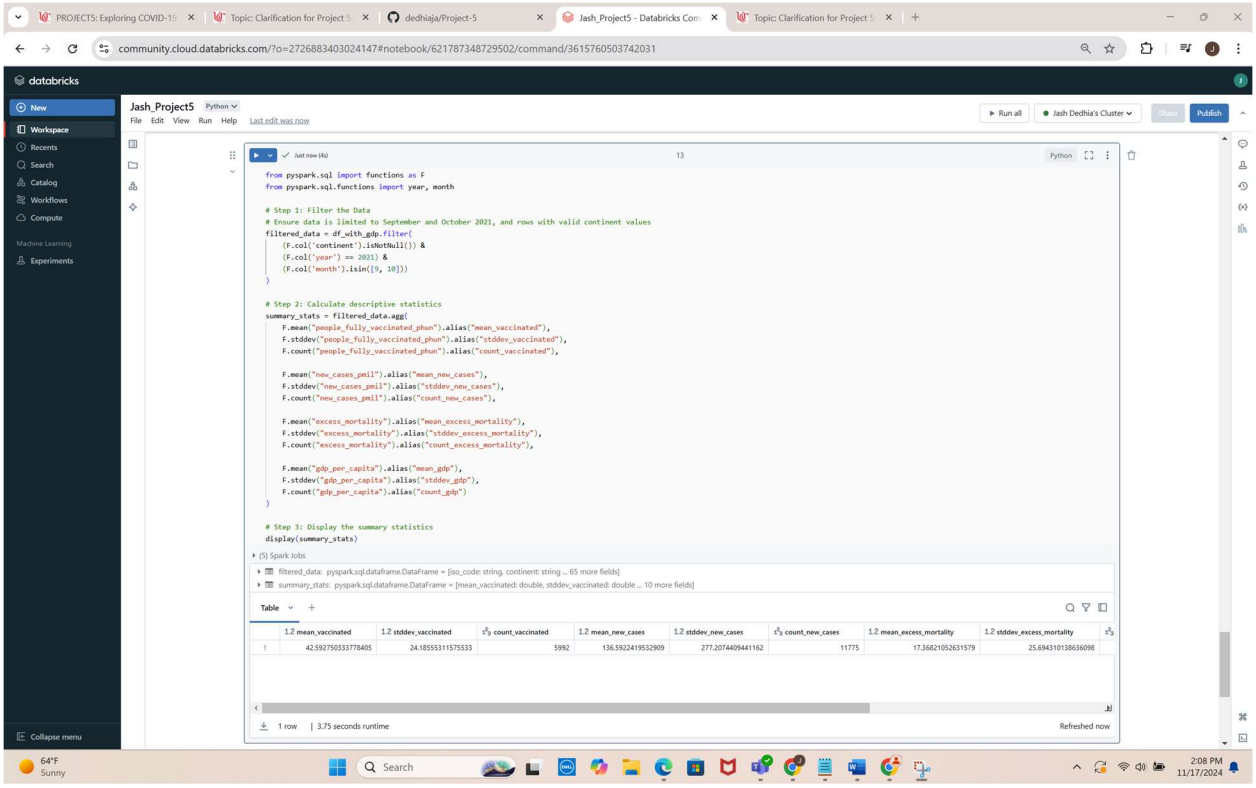
7.


Jash_Project5
Continents
Edit Visualization
12 rows
Refreshed 1 hour ago
Just now (3s)
11
# Perform correlation analysis
correlation = avg_df.stat.corr('average_people_fully_vaccinated', 'average_new_cases')
# Print the correlation coefficient
print(f"Correlation between average vaccination rate and new cases: {correlation}")
print(f"The correlation coefficient of {correlation:.4f} indicates a moderate positive relationship between vaccination rates and new cases. This suggests that other factors, such as testing rates or high prior infection rates in highly vaccinated regions, might be influencing this unexpected trend.")
(5) Spark Jobs
Correlation between average vaccination rate and new cases: 0.5195322019056203
The correlation coefficient of 0.5195 indicates a moderate positive relationship between vaccination rates and new cases. This suggests that other factors, such as testing rates or high prior infection rates in highly vaccinated regions, might be influencing this unexpected trend.
12:49 PM (2s)
12
# Define missing GDP data as a dictionary
gdp_data = {
"CUB": 11255.00,
"LIE": 197505.00,
"MCO": 240862.00,
"SOM": 1611.30,
"TWN": 67455.00,
"SYR": 2914.50
}
# Create a DataFrame from the dictionary
from pyspark.sql.types import StructType, StructField, StringType, FloatType
schema = StructType([
StructField("iso_code", StringType(), True),

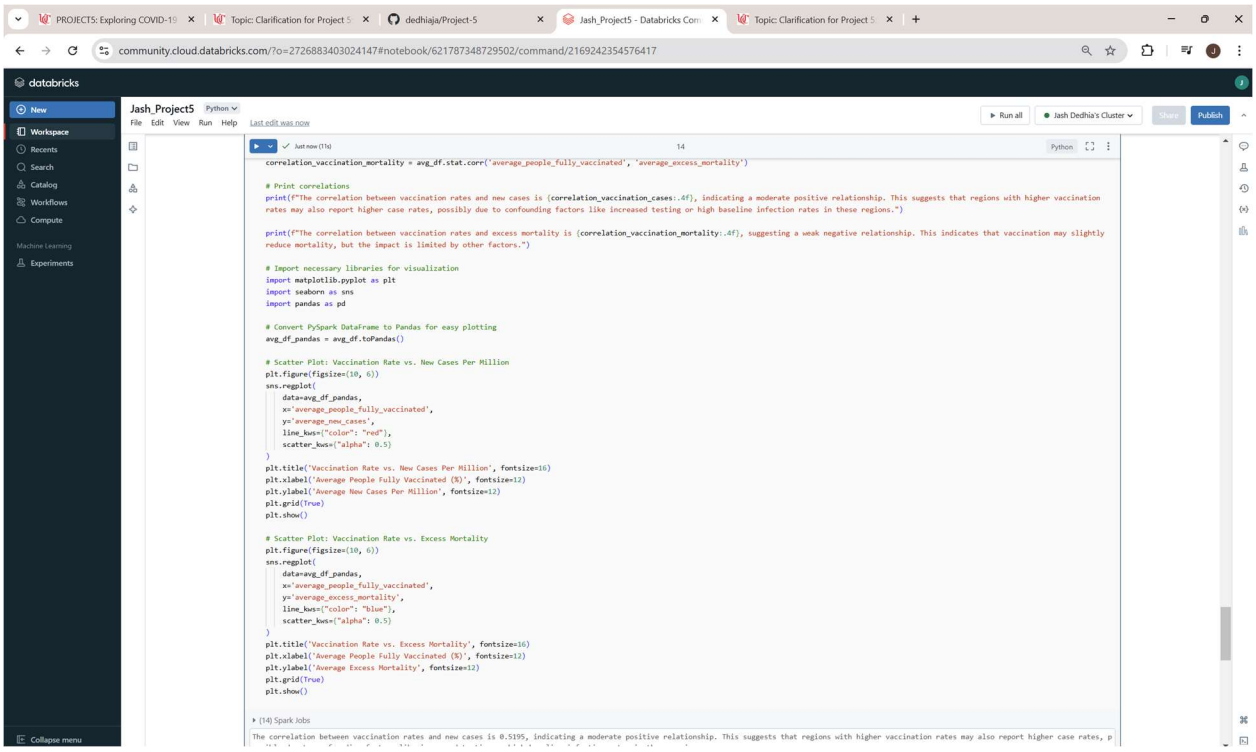
8.

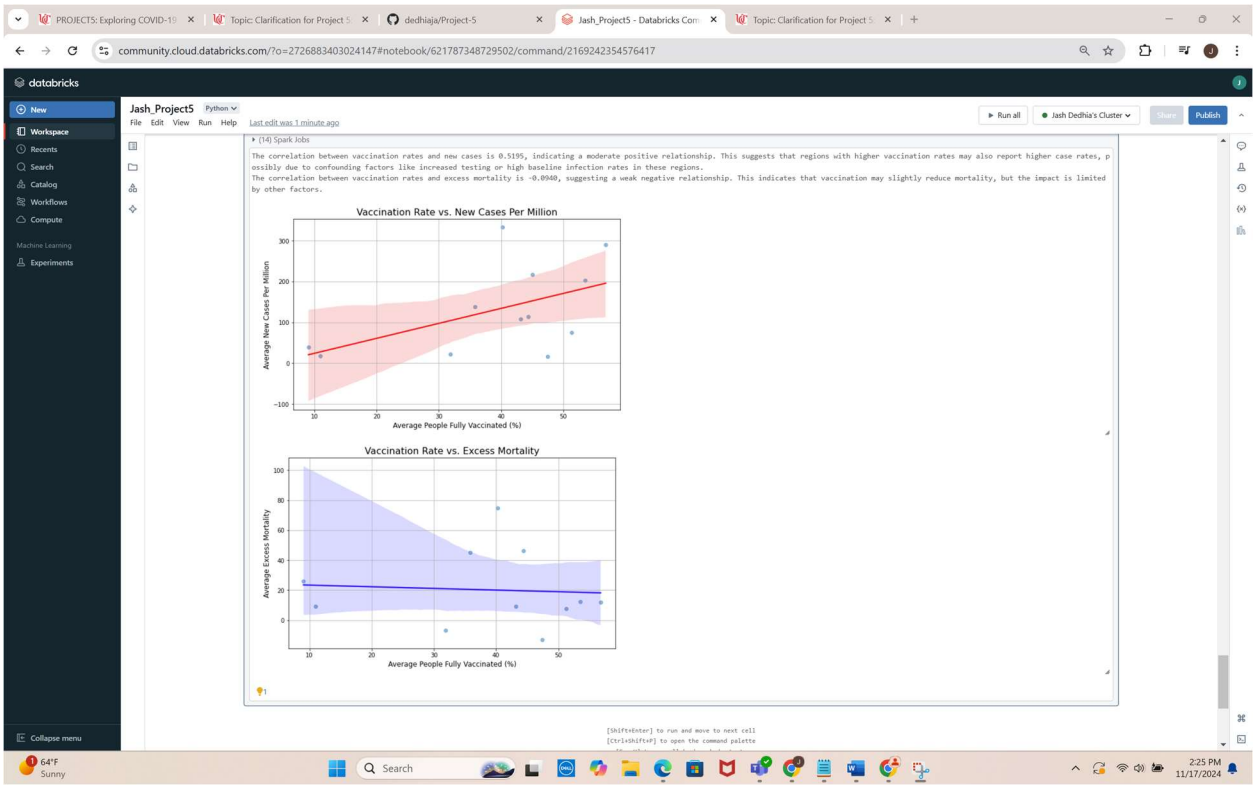


9.


PROJECT5: Exploring COVID-19
Topic: Clarification for Project 5
dedhiaja/Project-5
Jash_Project5 - Databricks Com
Topic: Clarification for Project 5
community.cloud.databricks.com/?o=2726883403024147#notebook/621787348729502/command/3615760503742031
databricks
New
Workspace
Recents
Search
Catalog
Workflows
Compute
Machine Learning
Experiments
Collapse menu
Jash_Project5
Python
File Edit View Run Help
Last edit was now
Run all
Jash Dedhia's Cluster
Share
Publish
Just now (4s)
13
Python
from pyspark.sql import functions as F
from pyspark.sql.functions import year, month
# Step 1: Filter the Data
# Ensure data is limited to September and October 2021, and rows with valid continent values
filtered_data = df_with_gdp.filter(
(F.col('continent').isNotNull()) &
(F.col('year') == 2021) &
(F.col('month').isin([9, 10]))
)
# Step 2: Calculate descriptive statistics
summary_stats = filtered_data.agg(
F.mean("people_fully_vaccinated_phun").alias("mean_vaccinated"),
F.stddev("people_fully_vaccinated_phun").alias("stddev_vaccinated"),
F.count("people_fully_vaccinated_phun").alias("count_vaccinated"),
F.mean("new_cases_pmil").alias("mean_new_cases"),
F.stddev("new_cases_pmil").alias("stddev_new_cases"),
F.count("new_cases_pmil").alias("count_new_cases"),
F.mean("excess_mortality").alias("mean_excess_mortality"),
F.stddev("excess_mortality").alias("stddev_excess_mortality"),
F.count("excess_mortality").alias("count_excess_mortality"),
F.mean("gdp_per_capita").alias("mean_gdp"),
F.stddev("gdp_per_capita").alias("stddev_gdp"),
F.count("gdp_per_capita").alias("count_gdp")
)
# Step 3: Display the summary statistics
display(summary_stats)
(5) Spark Jobs
filtered_data: pyspark.sql.dataframe.DataFrame = [iso_code: string, continent: string ... 65 more fields]
summary_stats: pyspark.sql.dataframe.DataFrame = [mean_vaccinated: double, stddev_vaccinated: double ... 10 more fields]
Table
1.2 mean_vaccinated | 1.2 stddev_vaccinated | count_vaccinated | 1.2 mean_new_cases | 1.2 stddev_new_cases | count_new_cases | 1.2 mean_excess_mortality | 1.2 stddev_excess_mortality
1 | 42.59275033378405 | 24.18555311575533 | 5992 | 136.5922419532909 | 277.2074409441162 | 11775 | 17.36821052631579 | 25.69431013863609 8
1 row | 3.75 seconds runtime
Refreshed now
64°F
Sunny
Search
2:08 PM
11/17/2024

10.


PROJECT5: Exploring COVID-19
Topic: Clarification for Project 5
dedhiaja/Project-5
Jash_Project5 - Databricks Com
Topic: Clarification for Project 5
community.cloud.databricks.com/?o=2726883403024147#notebook/621787348729502/command/2169242354576417
databricks
New
Workspace
Recents
Search
Catalog
Workflows
Compute
Machine Learning
Experiments
Collapse menu
Jash_Project5
Python
File Edit View Run Help
Last edit was now
Run all
Jash Dedhia's Cluster
Share
Publish
Just now (11s)
14
Python
correlation_vaccination_mortality = avg_df.stat.corr('average_people_fully_vaccinated', 'average_excess_mortality')
# Print correlations
print(f"The correlation between vaccination rates and new cases is {correlation_vaccination_cases:.4f}, indicating a moderate positive relationship. This suggests that regions with higher vaccination rates may also report higher case rates, possibly due to confounding factors like increased testing or high baseline infection rates in these regions.")
print(f"The correlation between vaccination rates and excess mortality is {correlation_vaccination_mortality:.4f}, suggesting a weak negative relationship. This indicates that vaccination may slightly reduce mortality, but the impact is limited by other factors.")
# Import necessary libraries for visualization
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
# Convert PySpark DataFrame to Pandas for easy plotting
avg_df_pandas = avg_df.toPandas()
# Scatter Plot: Vaccination Rate vs. New Cases Per Million
plt.figure(figsize=(10, 6))
sns.regplot(
data=avg_df_pandas,
x='average_people_fully_vaccinated',
y='average_new_cases',
line_kws={"color": "red"},
scatter_kws={"alpha": 0.5}
)
plt.title('Vaccination Rate vs. New Cases Per Million', fontsize=16)
plt.xlabel('Average People Fully Vaccinated (%)', fontsize=12)
plt.ylabel('Average New Cases Per Million', fontsize=12)
plt.grid(True)
plt.show()
# Scatter Plot: Vaccination Rate vs. Excess Mortality
plt.figure(figsize=(10, 6))
sns.regplot(
data=avg_df_pandas,
x='average_people_fully_vaccinated',
y='average_excess_mortality',
line_kws={"color": "blue"},
scatter_kws={"alpha": 0.5}
)
plt.title('Vaccination Rate vs. Excess Mortality', fontsize=16)
plt.xlabel('Average People Fully Vaccinated (%)', fontsize=12)
plt.ylabel('Average Excess Mortality', fontsize=12)
plt.grid(True)
plt.show()
(14) Spark Jobs
The correlation between vaccination rates and new cases is 0.5195, indicating a moderate positive relationship. This suggests that regions with higher vaccination rates may also report higher case rates, p



The analysis reveals a **moderate positive correlation (0.5195)** between vaccination rates and new COVID-19 cases per million. This suggests that regions with higher vaccination rates may also report more cases during September and October 2021. However, this counterintuitive trend can be attributed to external factors, such as increased testing rates, higher population densities, or previously high baseline infection levels in regions with strong vaccination campaigns. Additionally, vaccination may encourage behavioral changes like reduced mask usage or social distancing, particularly in regions where people perceive themselves as protected. These factors likely confound the direct relationship between vaccination rates and case counts.

In contrast, the **weak negative correlation (-0.0940)** between vaccination rates and excess mortality aligns with expectations that vaccines reduce severe outcomes and deaths. While the effect appears small, it indicates that vaccination may help mitigate mortality, even if the direct impact on case numbers is less clear. Factors such as healthcare quality, demographic differences, and variant severity likely also influence mortality rates, underscoring the importance of comprehensive pandemic responses beyond vaccination campaigns alone. These findings suggest that vaccination plays a critical role in reducing the severity of COVID-19 but must be complemented by other measures to manage overall case rates.