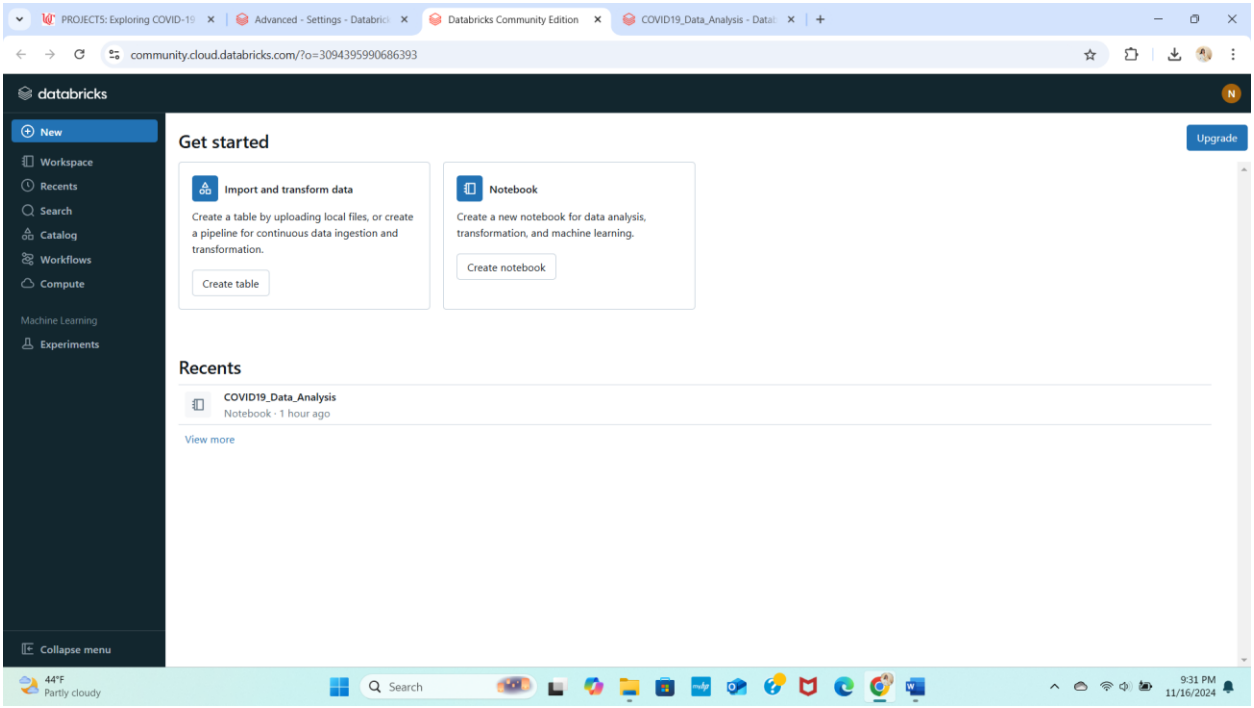
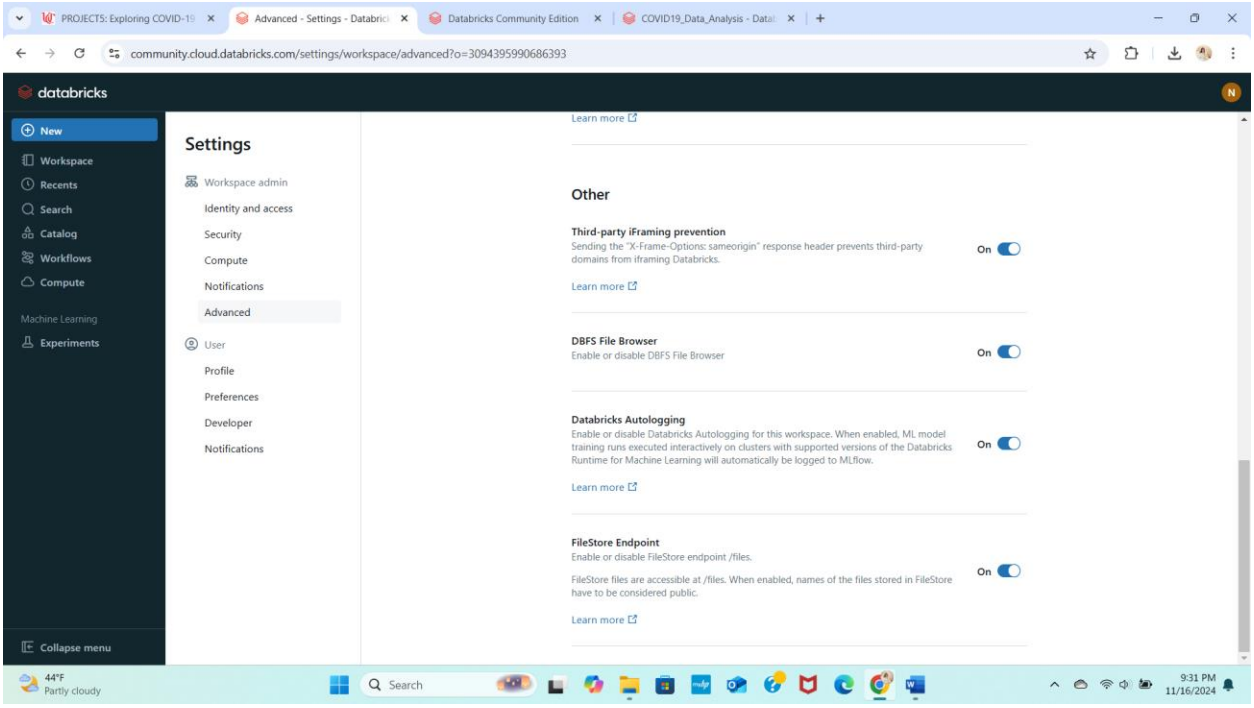
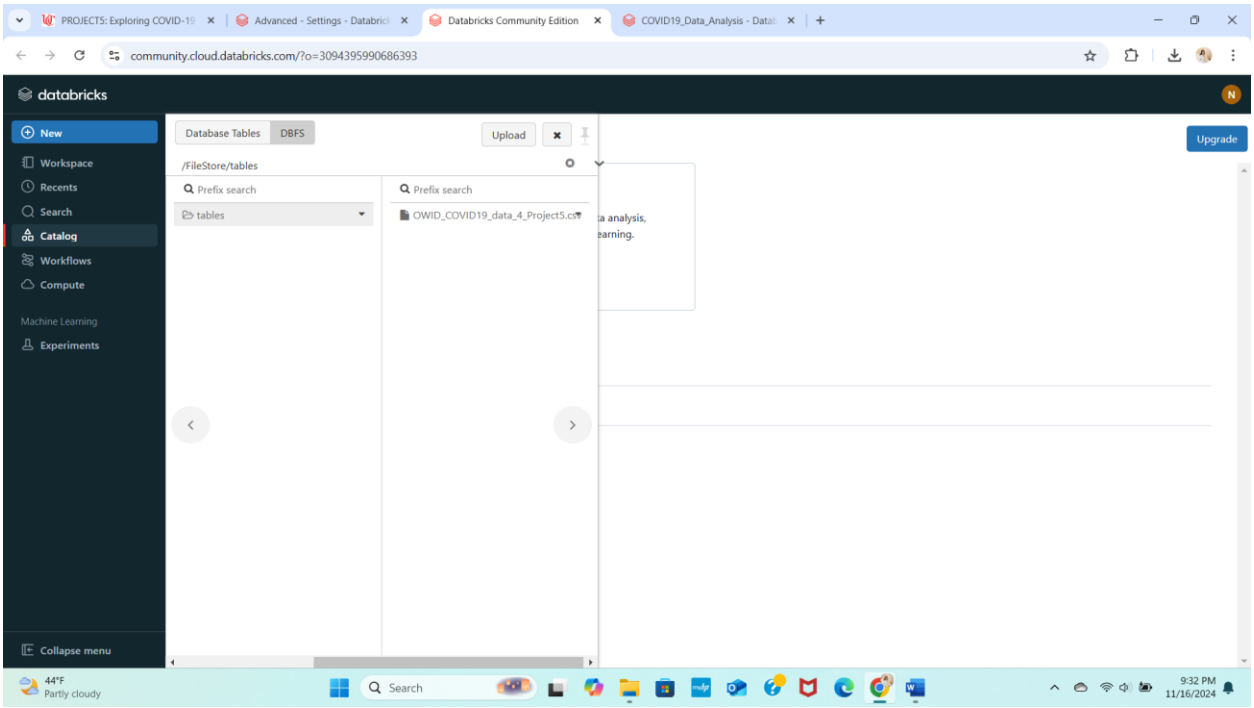
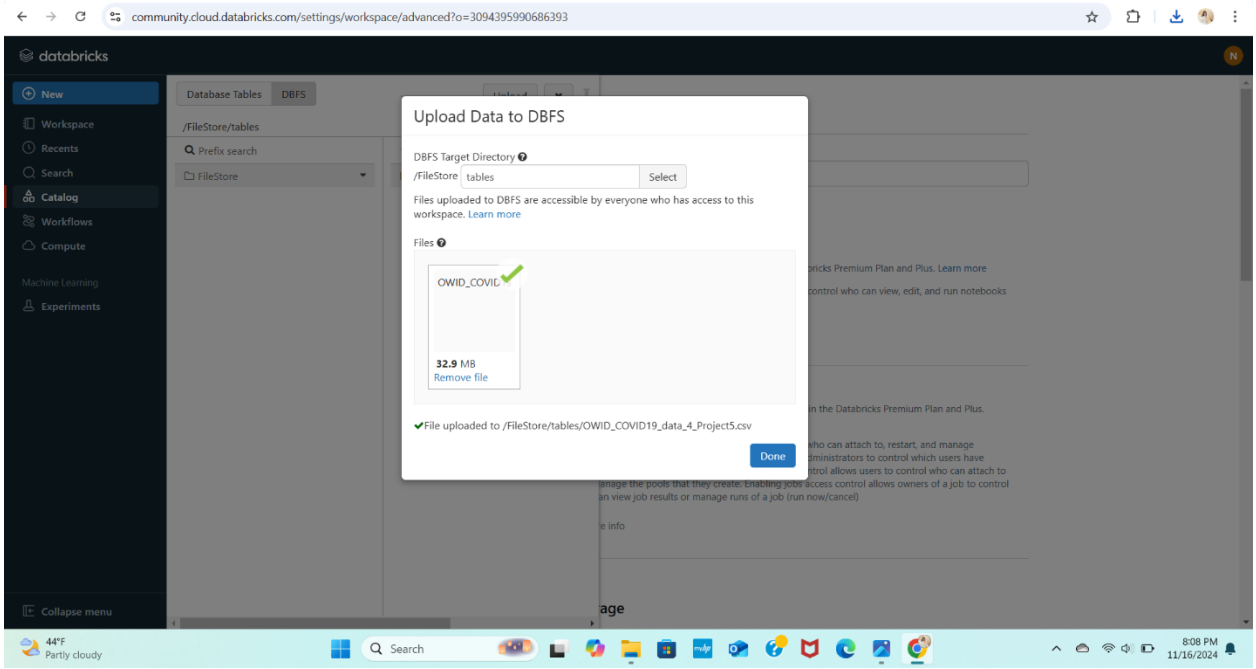
Exploring COVID-19 Data using Databricks

1) Install Databricks and upload the data

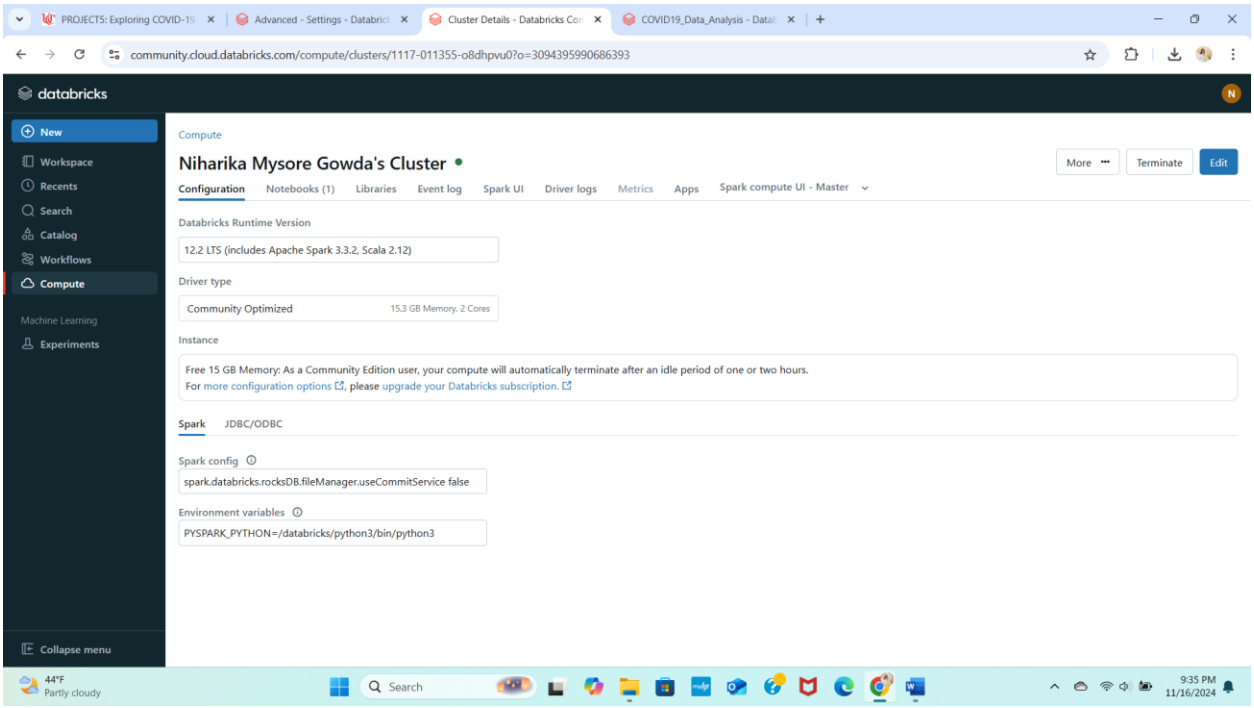


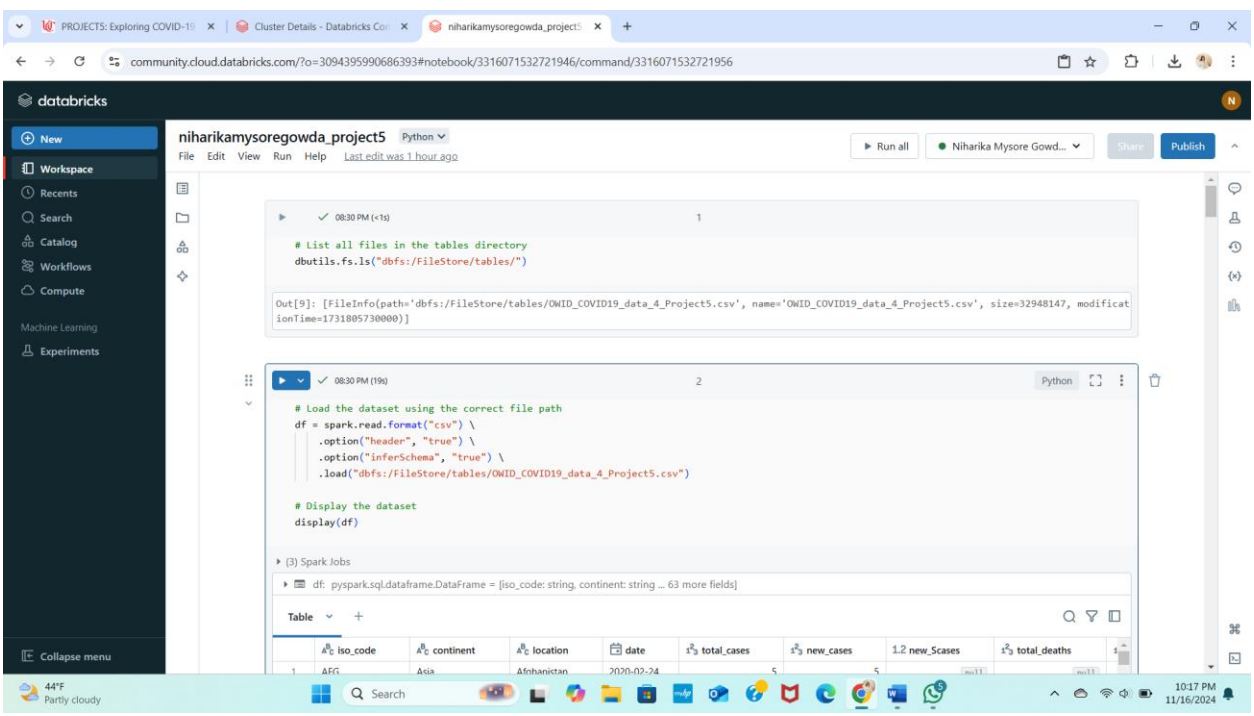

PROJECT5: Exploring COVID-19
Advanced - Settings - Databrick
Databricks Community Edition
COVID19_Data_Analysis - Data
community.cloud.databricks.com/settings/workspace/advanced?o=3094395990686393
databricks
New
Workspace
Recents
Search
Catalog
Workflows
Compute
Machine Learning
Experiments
Collapse menu
Settings
Workspace admin
Identity and access
Security
Compute
Notifications
Advanced
User
Profile
Preferences
Developer
Notifications
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
44°F
Partly cloudy
Search
9:31 PM
11/16/2024

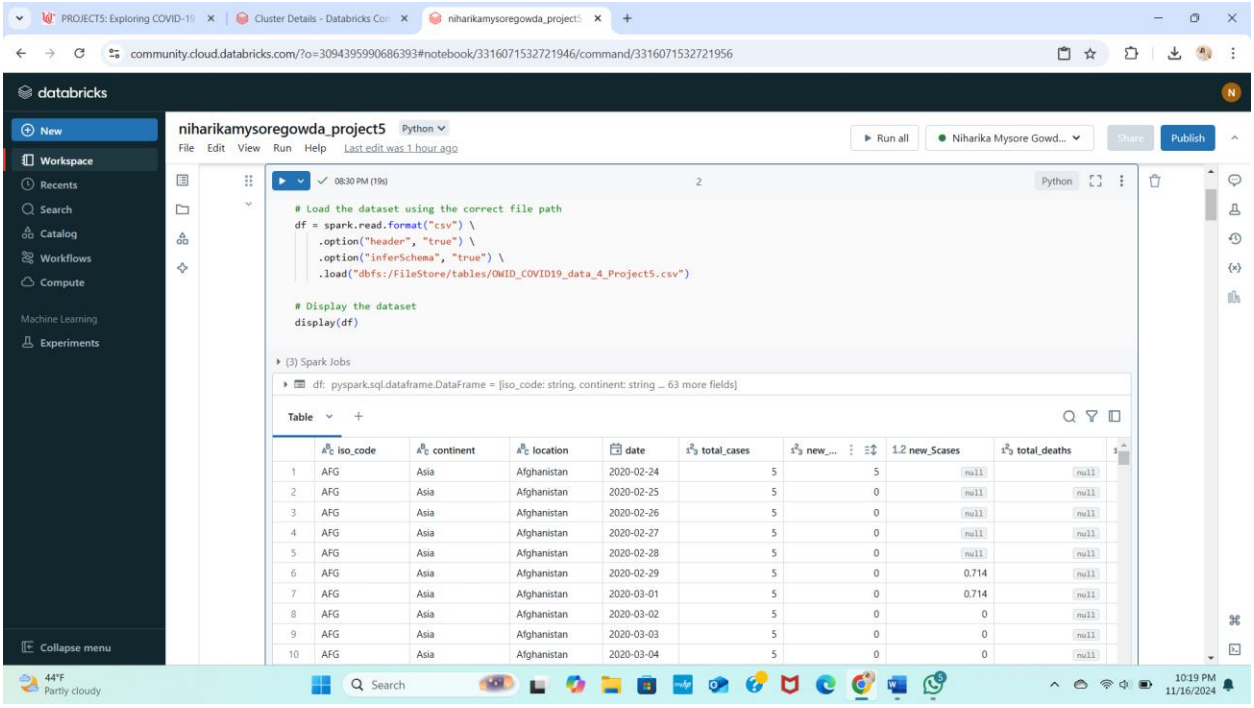

PROJECT5: Exploring COVID-19
Advanced - Settings - Databrick
Databricks Community Edition
COVID19_Data_Analysis - Data
community.cloud.databricks.com/?o=3094395990686393
databricks
New
Workspace
Recents
Search
Catalog
Workflows
Compute
Machine Learning
Experiments
Collapse menu
Database Tables
DBFS
Upload
/FileStore/tables
Prefix search
tables
Prefix search
OWID_COVID19_data_4_Project5.cs
Upgrade
44°F
Partly cloudy
Search
9:32 PM
11/16/2024


community.cloud.databricks.com/settings/workspace/advanced?o=3094395990686393
databricks
New
Workspace
Recents
Search
Catalog
Workflows
Compute
Machine Learning
Experiments
Collapse menu
Database Tables
DBFS
/FileStore/tables
Prefix search
FileStore
Upload Data to DBFS
DBFS Target Directory
/FileStore
tables
Select
Files uploaded to DBFS are accessible by everyone who has access to this workspace. Learn more
Files
OWID_COVID
32.9 MB
Remove file
File uploaded to /FileStore/tables/OWID_COVID19_data_4_Project5.csv
Done
44°F
Partly cloudy
Search
8:08 PM
11/16/2024

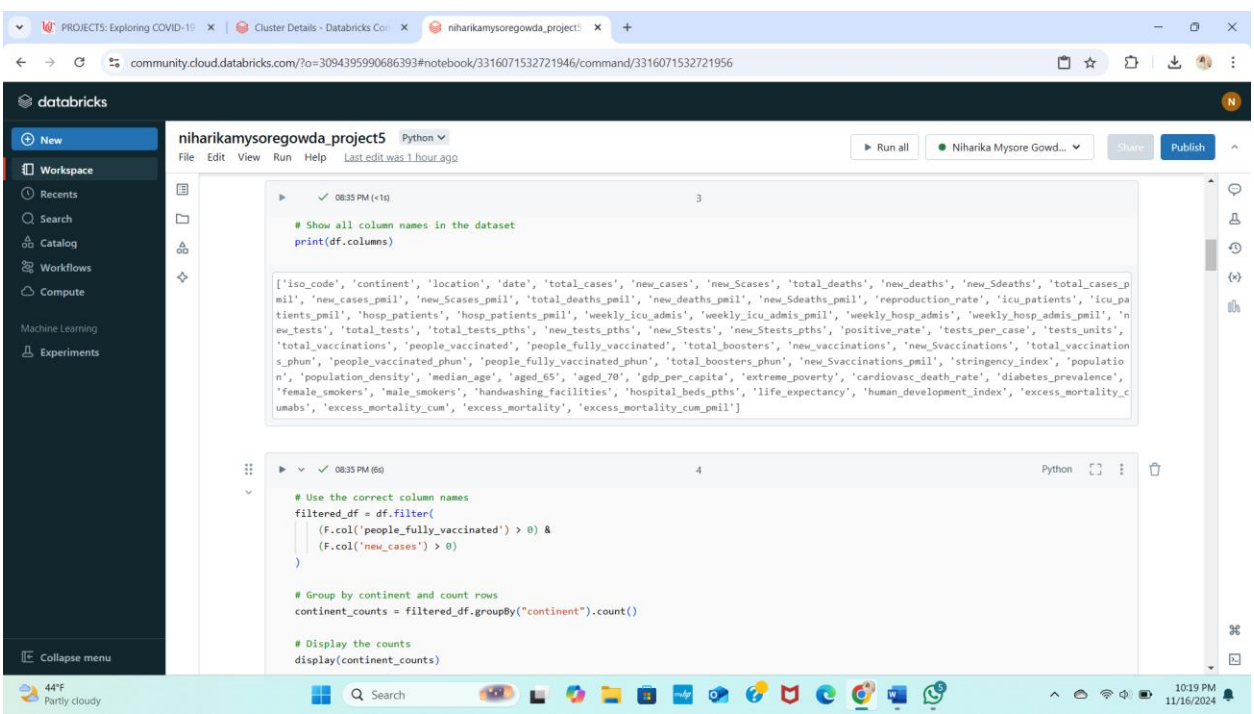
2) Create a Databricks cluster and load the data file using Pandas or PySpark as needed?

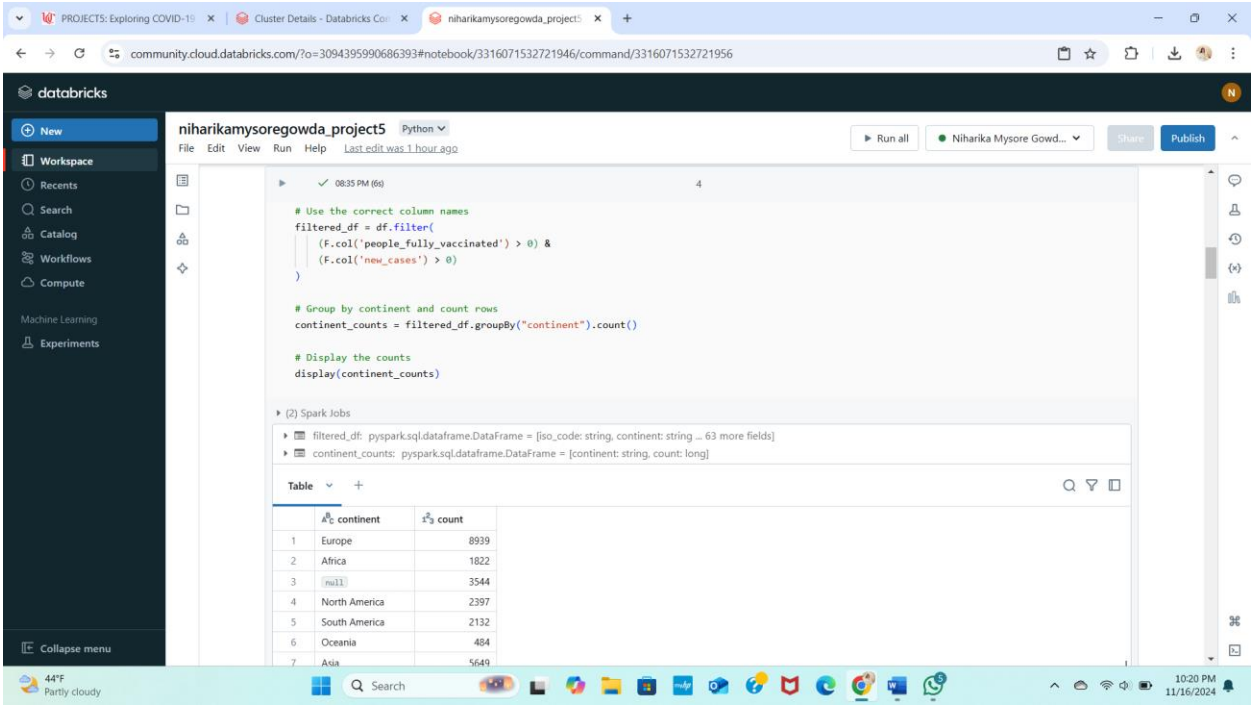


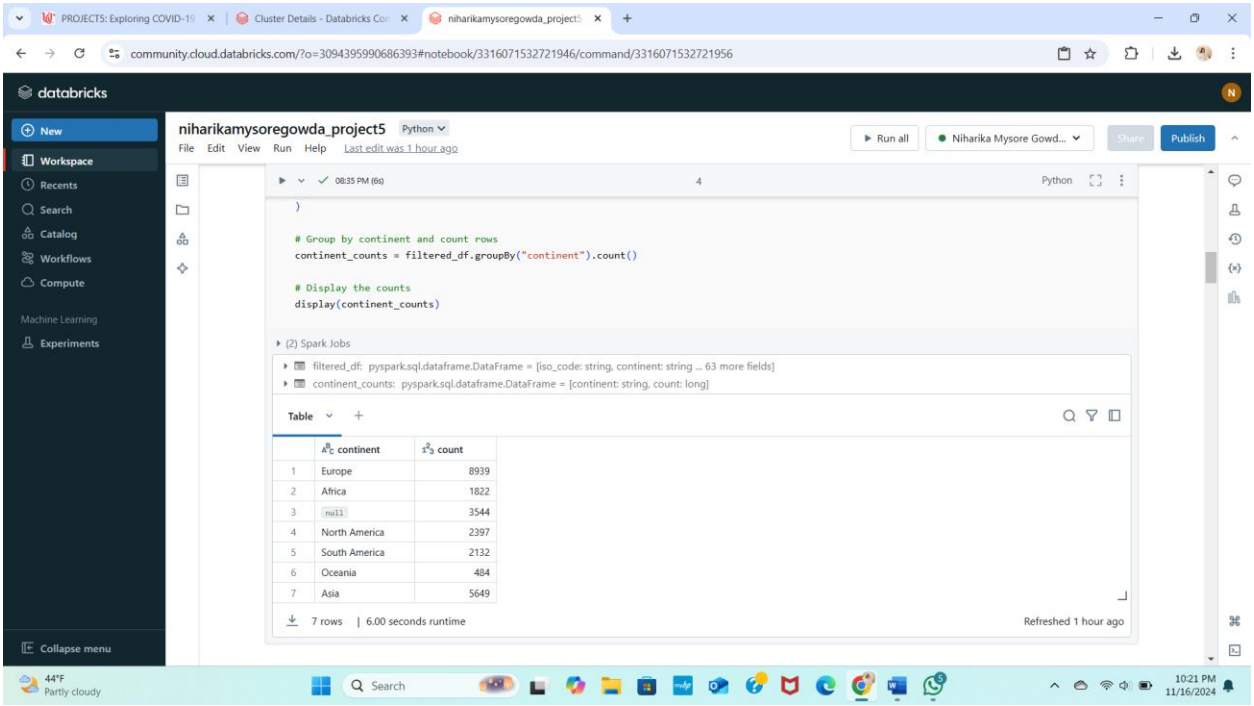




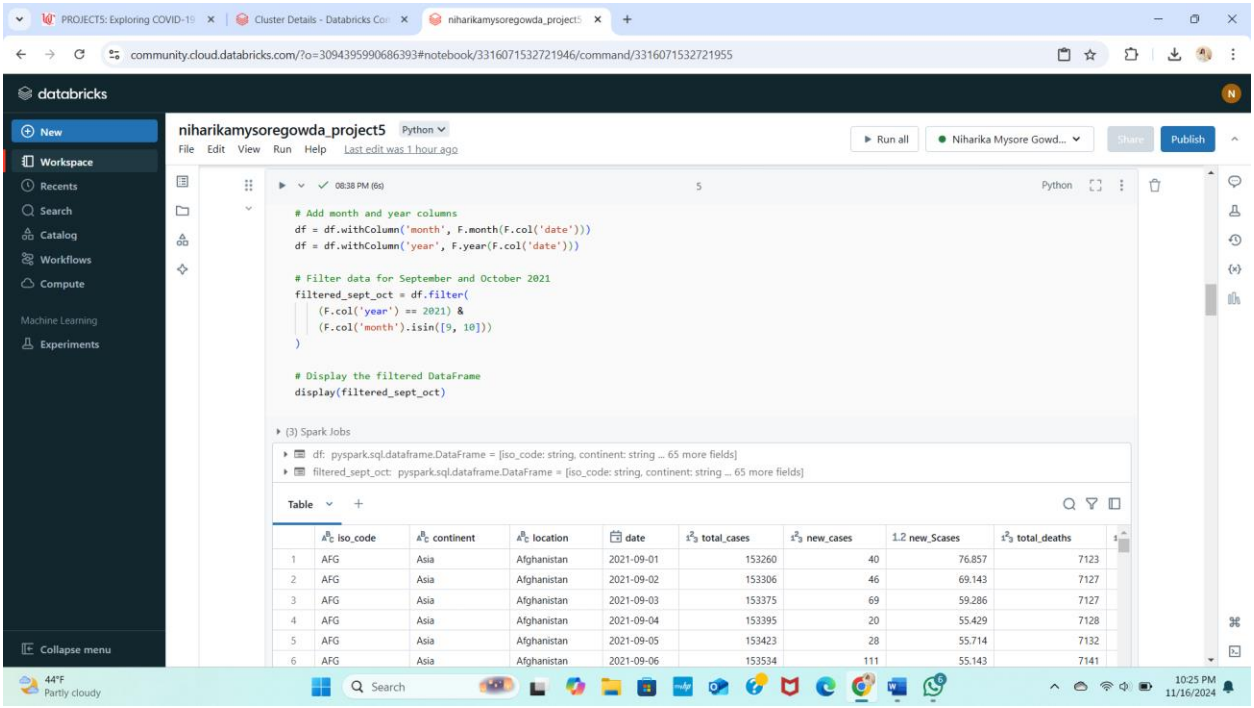
3) Filter Records and display the row count by continent ?

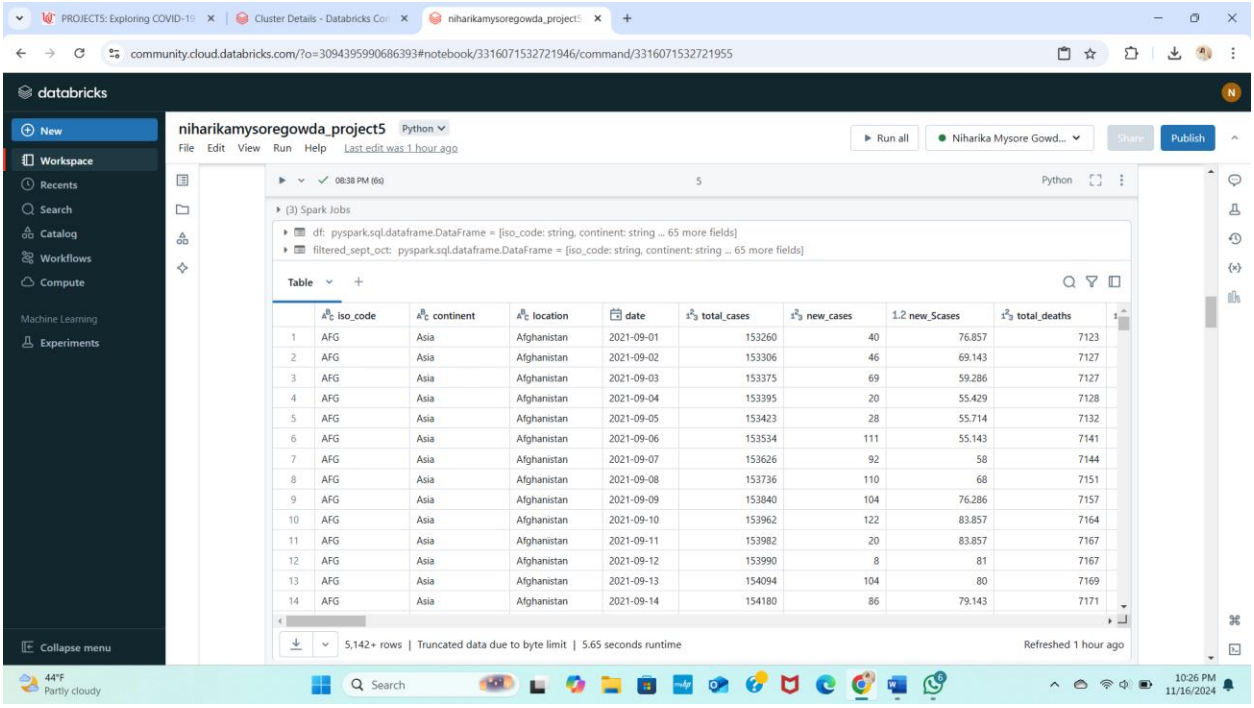



PROJECT5: Exploring COVID-19
Cluster Details - Databricks Co
niharikamysoregowda_project5
community.cloud.databricks.com/?o=3094395990686393#notebook/3316071532721946/command/3316071532721956
databricks
New
Workspace
Recents
Search
Catalog
Workflows
Compute
Machine Learning
Experiments
Collapse menu
niharikamysoregowda_project5
Python
File Edit View Run Help
Last edit was 1 hour ago
Run all
Niharika Mysore Gowd...
Share
Publish
08:35 PM (6s)
4
# Use the correct column names
filtered_df = df.filter(
(F.col('people_fully_vaccinated') > 0) &
(F.col('new_cases') > 0)
)
# Group by continent and count rows
continent_counts = filtered_df.groupBy("continent").count()
# Display the counts
display(continent_counts)
(2) Spark Jobs
filtered_df: pyspark.sql.dataframe.DataFrame = [iso_code: string, continent: string ... 63 more fields]
continent_counts: pyspark.sql.dataframe.DataFrame = [continent: string, count: long]
Table
continent
count
1 Europe 8939
2 Africa 1822
3 null 3544
4 North America 2397
5 South America 2132
6 Oceania 484
7 Asia 5649
44°F
Partly cloudy
Search
10:20 PM
11/16/2024


PROJECT5: Exploring COVID-19
Cluster Details - Databricks Co
niharikamysoregowda_project5
community.cloud.databricks.com/?o=3094395990686393#notebook/3316071532721946/command/3316071532721956
databricks
New
Workspace
Recents
Search
Catalog
Workflows
Compute
Machine Learning
Experiments
Collapse menu
niharikamysoregowda_project5
Python
File Edit View Run Help
Last edit was 1 hour ago
Run all
Niharika Mysore Gowd...
Share
Publish
08:35 PM (6s)
4
Python
)
# Group by continent and count rows
continent_counts = filtered_df.groupBy("continent").count()
# Display the counts
display(continent_counts)
(2) Spark Jobs
filtered_df: pyspark.sql.dataframe.DataFrame = [iso_code: string, continent: string ... 63 more fields]
continent_counts: pyspark.sql.dataframe.DataFrame = [continent: string, count: long]
Table
continent
count
1 Europe 8939
2 Africa 1822
3 null 3544
4 North America 2397
5 South America 2132
6 Oceania 484
7 Asia 5649
7 rows | 6.00 seconds runtime
Refreshed 1 hour ago
44°F
Partly cloudy
Search
10:21 PM
11/16/2024

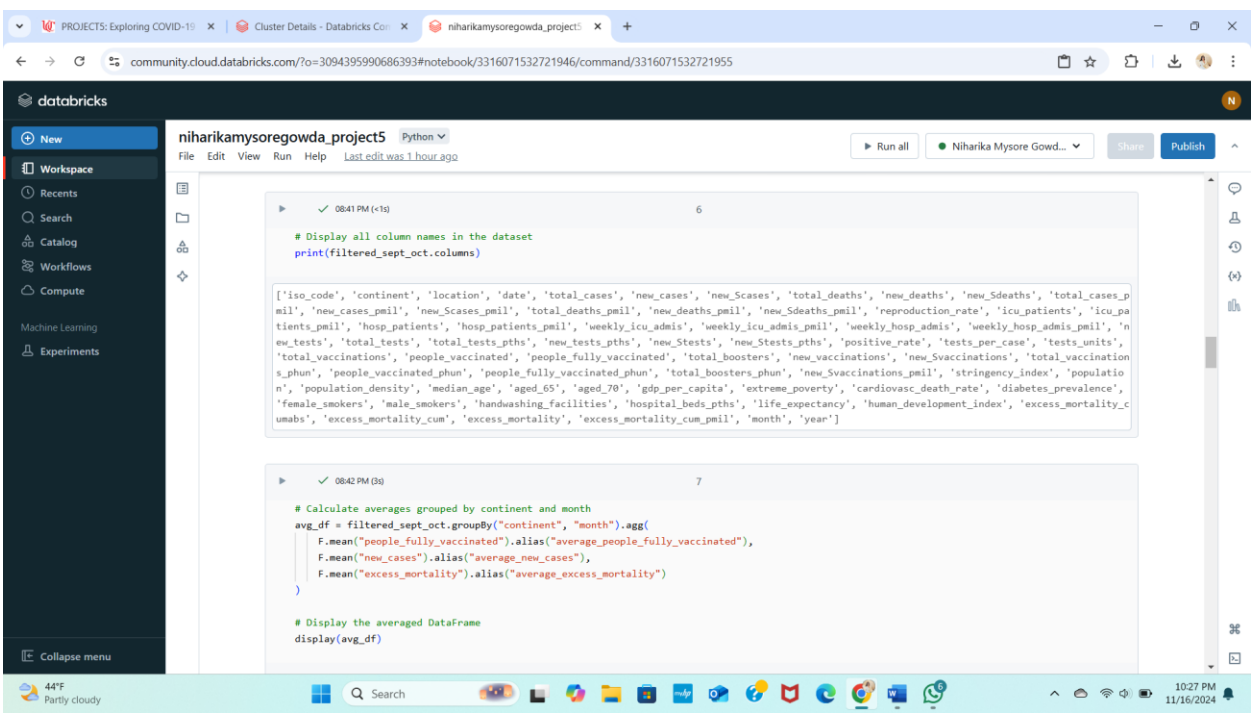
4) Create Month and Year Columns and display the total record count ?





	iso_code	continent	location	date	total_cases	new_cases	new_Scases	total_deaths
1	AFG	Asia	Afghanistan	2021-09-01	153260	40	76.857	7123
2	AFG	Asia	Afghanistan	2021-09-02	153306	46	69.143	7127
3	AFG	Asia	Afghanistan	2021-09-03	153375	69	59.286	7127
4	AFG	Asia	Afghanistan	2021-09-04	153395	20	55.429	7128
5	AFG	Asia	Afghanistan	2021-09-05	153423	28	55.714	7132
6	AFG	Asia	Afghanistan	2021-09-06	153534	111	55.143	7141
7	AFG	Asia	Afghanistan	2021-09-07	153626	92	58	7144
8	AFG	Asia	Afghanistan	2021-09-08	153736	110	68	7151
9	AFG	Asia	Afghanistan	2021-09-09	153840	104	76.286	7157
10	AFG	Asia	Afghanistan	2021-09-10	153962	122	83.857	7164
11	AFG	Asia	Afghanistan	2021-09-11	153982	20	83.857	7167
12	AFG	Asia	Afghanistan	2021-09-12	153990	8	81	7167
13	AFG	Asia	Afghanistan	2021-09-13	154094	104	80	7169
14	AFG	Asia	Afghanistan	2021-09-14	154180	86	79.143	7171

5) Calculate Averages and display the row count by **continent** by **month**



PROJECT5: Exploring COVID-19 | Cluster Details - Databricks Co | niharikamysoregowda_project5

community.cloud.databricks.com/?o=3094395990686393#notebook/3316071532721946/command/3316071532721955

databricks

niharikamysoregowda_project5 Python

File Edit View Run Help Last edit was 1 hour ago

Run all | Niharika Mysore Gowd... | Share | Publish

08:42 PM (3s) 7 Python

```python
# Calculate averages grouped by continent and month
avg_df = filtered_sept_oct.groupBy("continent", "month").agg(
    F.mean("people_fully_vaccinated").alias("average_people_fully_vaccinated"),
    F.mean("new_cases").alias("average_new_cases"),
    F.mean("excess_mortality").alias("average_excess_mortality")
)

# Display the averaged DataFrame
display(avg_df)
```

(2) Spark Jobs

avg_df: pyspark.sql.dataframe.DataFrame = [continent: string, month: integer ... 3 more fields]

Table

	continent	month	average_people_fully_vaccinated	average_new_cases	average_excess_mortality
1	Asia	9	22096526.44235294	4053.748936170213	45.19416666666667
2	Asia	10	27797579.723353293	2434.627316403569	46.093333333333334
3	Africa	9	1883448.3812316717	306.1043209876543	25.88
4	Oceania	9	3583451.804347826	180.3121212121212	-6.7
5	null	9	628643291.5722222	125350.9282051282	null
6	South America	9	15159482.979338843	2053.3454038997215	9.0475
7	Africa	10	1975453.0733695652	127.30227001194743	9.0975
8	null	10	736008439.7580645	102029.5012406948	null
9	Europe	10	10071081.166177908	3875.915147265077	12.14

44°F Partly cloudy | Search | 10:27 PM 11/16/2024

PROJECT5: Exploring COVID-19 | Cluster Details - Databricks Co | niharikamysoregowda_project5

community.cloud.databricks.com/?o=3094395990686393#notebook/3316071532721946/command/3316071532721956

databricks

niharikamysoregowda_project5 Python

File Edit View Run Help Last edit was 1 hour ago

Run all | Terminated | Share | Publish

08:42 PM (3s) 7 Python

(2) Spark Jobs

avg_df: pyspark.sql.dataframe.DataFrame = [continent: string, month: integer ... 3 more fields]

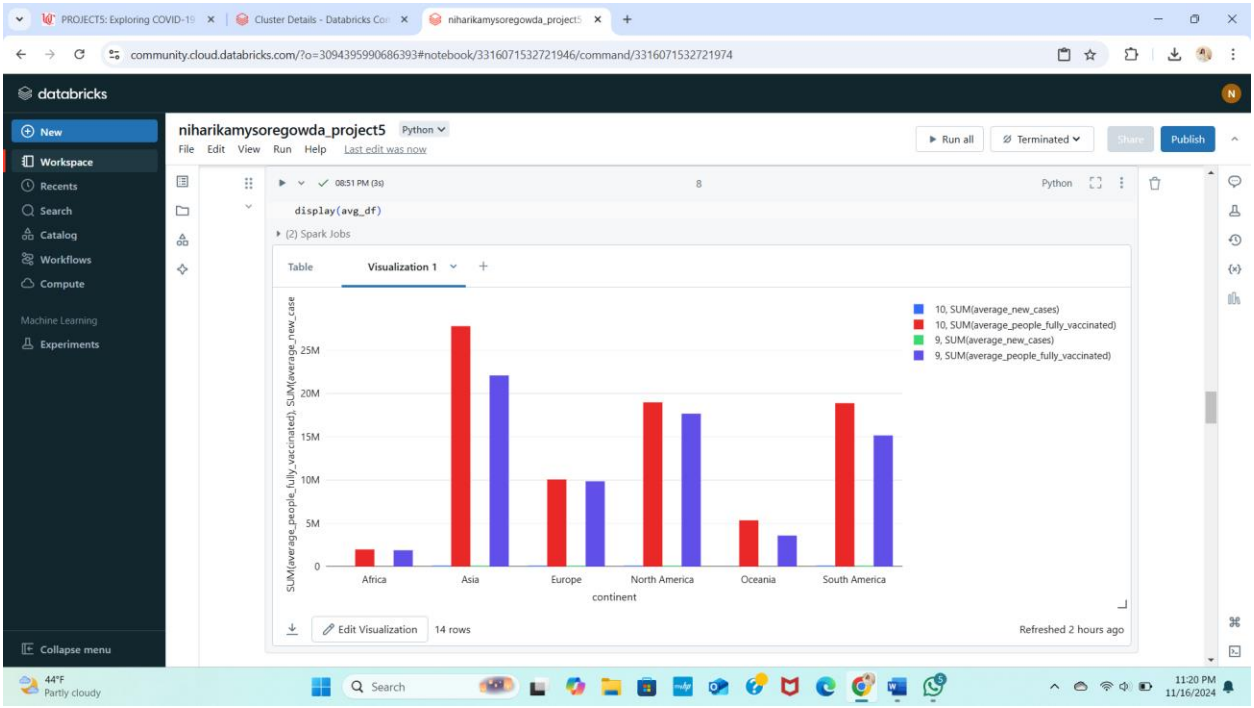
Table

	continent	month	average_people_fully_vaccinated	average_new_cases	average_excess_mortality
1	Asia	9	22096526.44235294	4053.748936170213	45.19416666666667
2	Asia	10	27797579.723353293	2434.627316403569	46.093333333333334
3	Africa	9	1883448.3812316717	306.1043209876543	25.88
4	Oceania	9	3583451.804347826	180.3121212121212	-6.7
5	null	9	628643291.5722222	125350.9282051282	null
6	South America	9	15159482.979338843	2053.3454038997215	9.0475
7	Africa	10	1975453.0733695652	127.30227001194743	9.0975
8	null	10	736008439.7580645	102029.5012406948	null
9	Europe	10	10071081.166177908	3875.915147265077	12.14
10	South America	10	18892681.772925764	1621.3279569892472	7.786666666666668
11	North America	9	17677649.330337077	7338.973913043478	74.689
12	Oceania	10	5345090.62244898	226.07558139534885	-13
13	Europe	9	9860297.254154447	2693.226811594203	12.174639175257733
14	North America	10	18987487.34304933	4142.352033660589	null

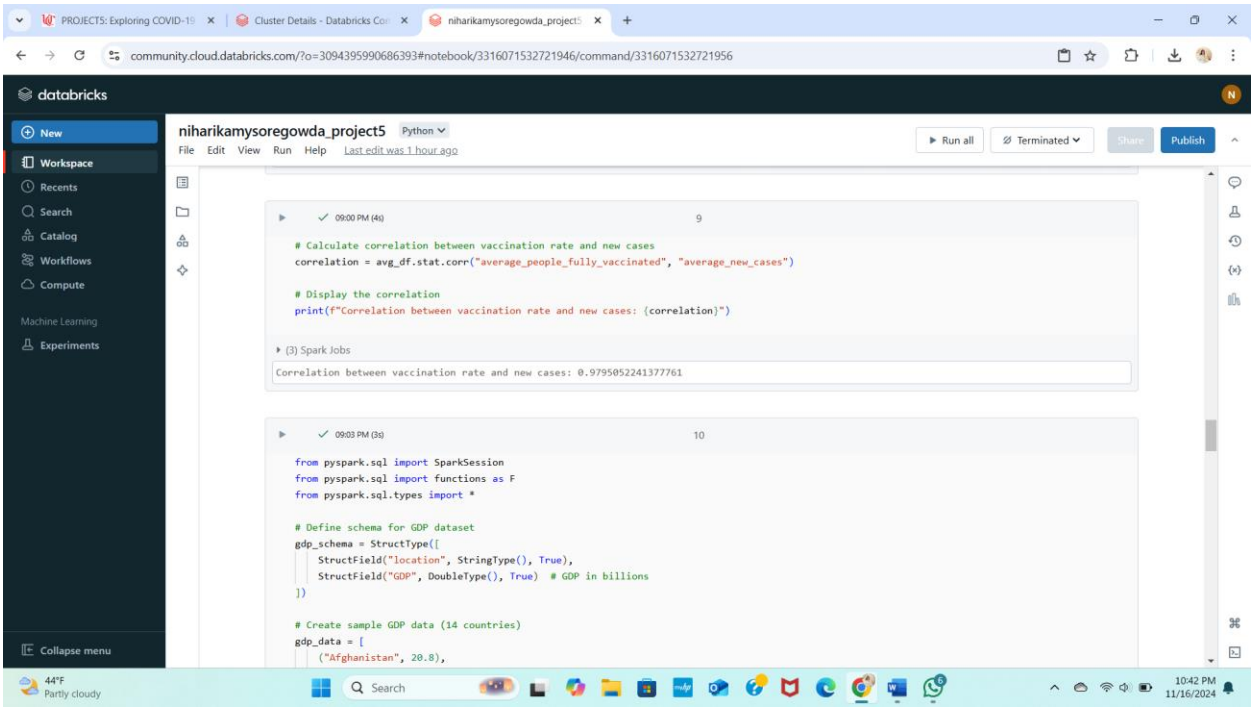
14 rows | 3.24 seconds runtime Refreshed 2 hours ago

44°F Partly cloudy | Search | 10:37 PM 11/16/2024

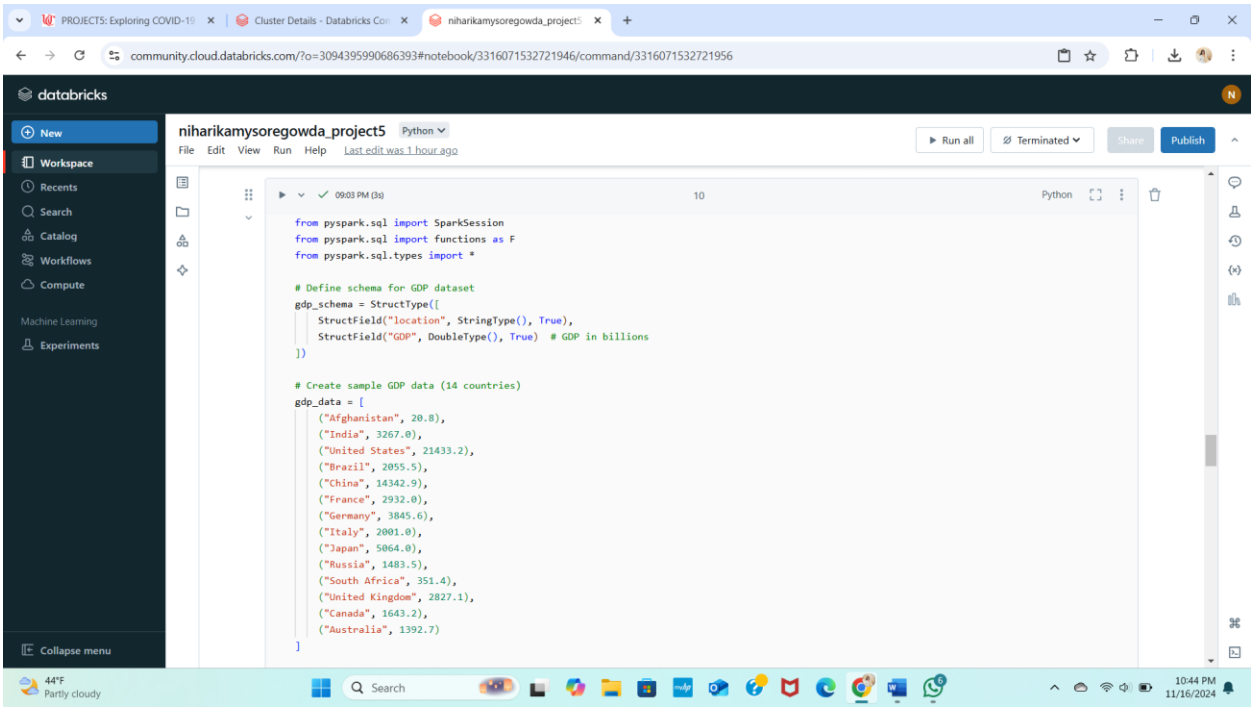
5) Plotting a Bar Chart?

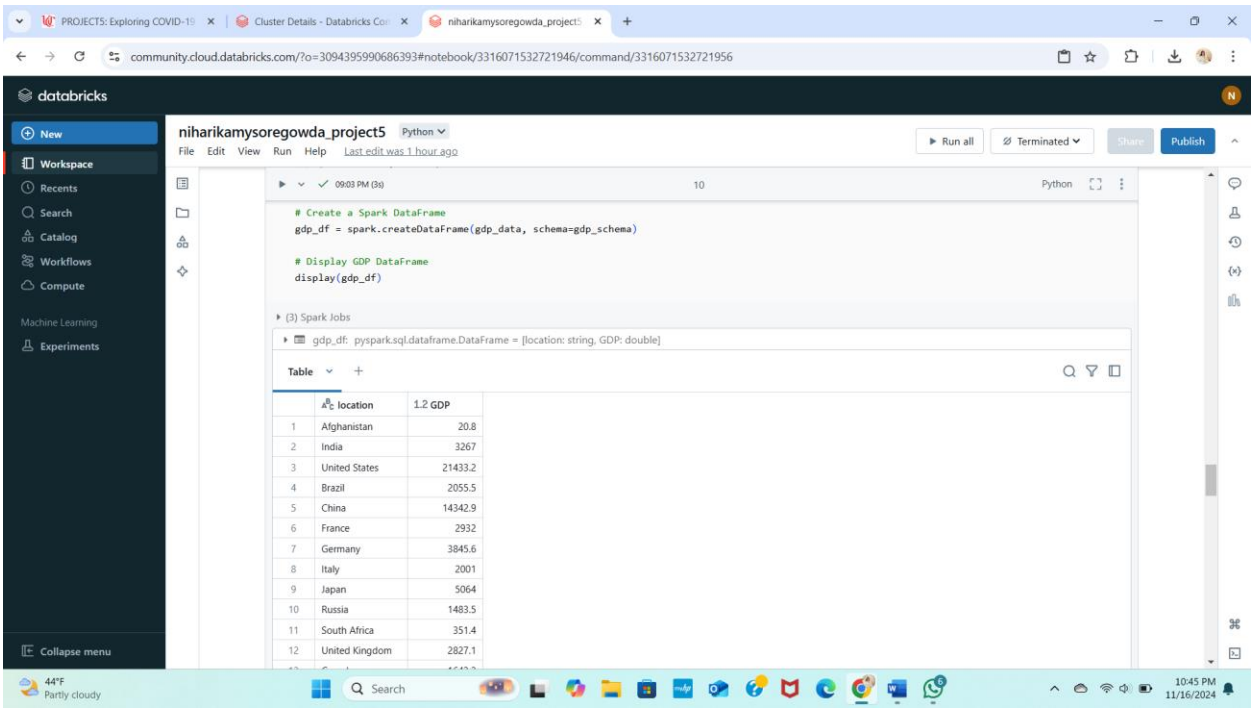


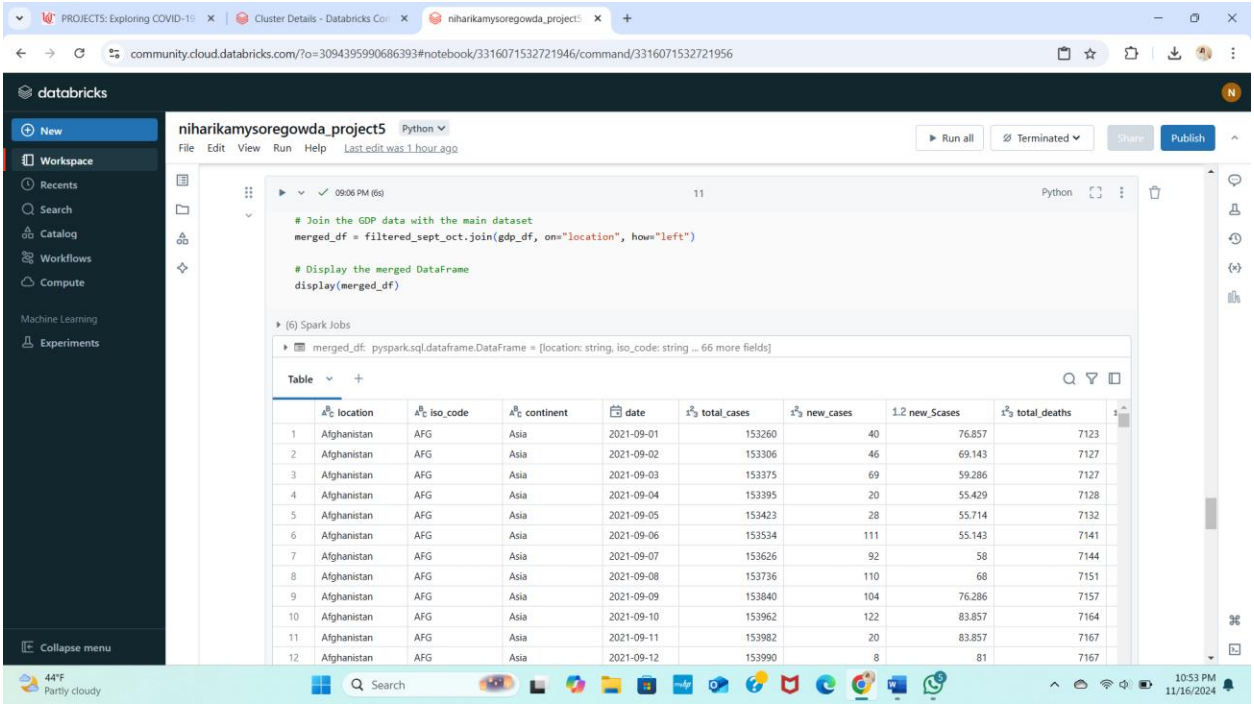
7) Run Correlation Analysis ?

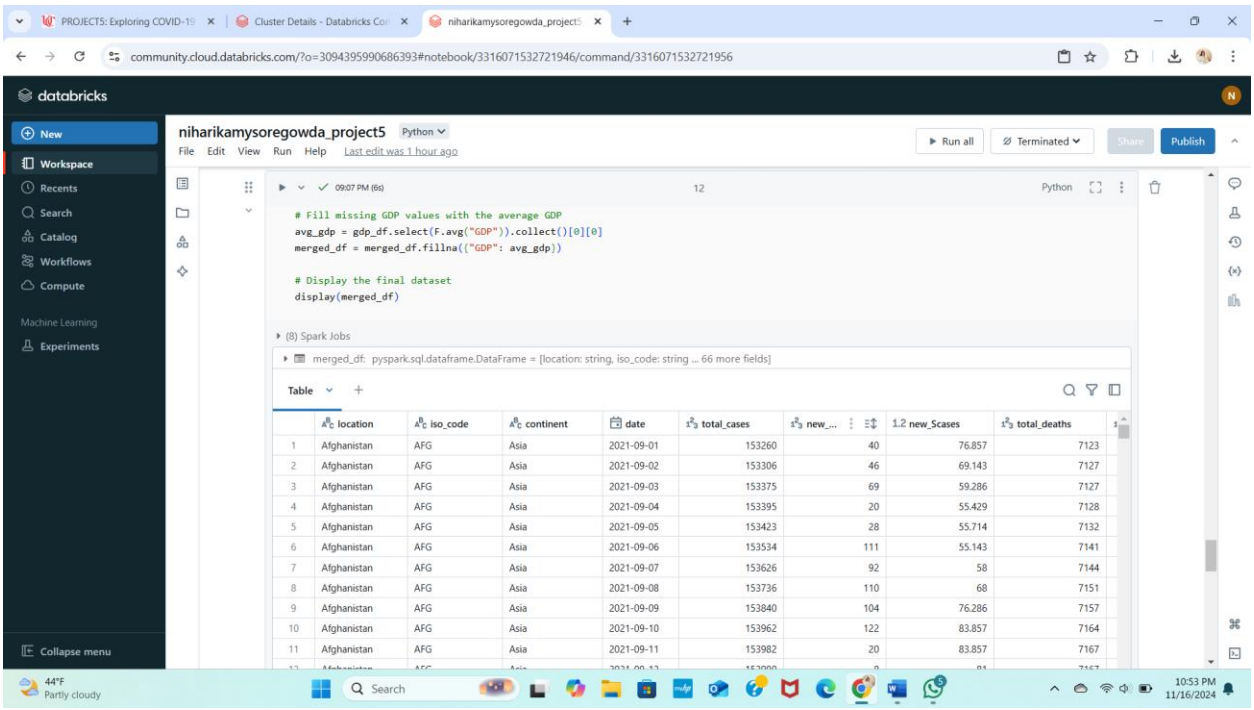


8) Fill missing GDP data

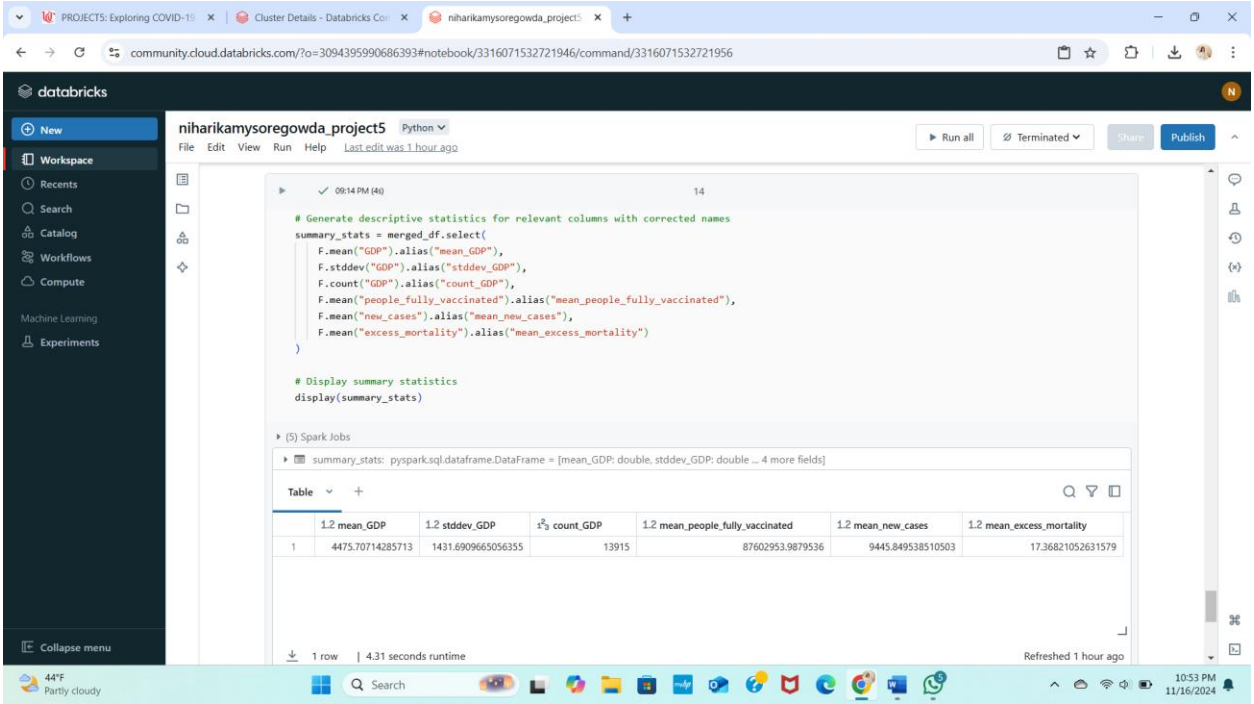





PROJECT5: Exploring COVID-19
Cluster Details - Databricks Co
niharikamysoregowda_project5
community.cloud.databricks.com/?o=3094395990686393#notebook/3316071532721946/command/3316071532721956
databricks
New
Workspace
Recents
Search
Catalog
Workflows
Compute
Machine Learning
Experiments
Collapse menu
niharikamysoregowda_project5
Python
File Edit View Run Help Last edit was 1 hour ago
Run all
Terminated
Share
Publish
09:06 PM (6s)
11
Python
# Join the GDP data with the main dataset
merged_df = filtered_sept_oct.join(gdp_df, on="location", how="left")
# Display the merged DataFrame
display(merged_df)
(6) Spark Jobs
merged_df: pyspark.sql.dataframe.DataFrame = [location: string, iso_code: string ... 66 more fields]
Table
location iso_code continent date total_cases new_cases new_Scases total_deaths
1 Afghanistan AFG Asia 2021-09-01 153260 40 76.857 7123
2 Afghanistan AFG Asia 2021-09-02 153306 46 69.143 7127
3 Afghanistan AFG Asia 2021-09-03 153375 69 59.286 7127
4 Afghanistan AFG Asia 2021-09-04 153395 20 55.429 7128
5 Afghanistan AFG Asia 2021-09-05 153423 28 55.714 7132
6 Afghanistan AFG Asia 2021-09-06 153534 111 55.143 7141
7 Afghanistan AFG Asia 2021-09-07 153626 92 58 7144
8 Afghanistan AFG Asia 2021-09-08 153736 110 68 7151
9 Afghanistan AFG Asia 2021-09-09 153840 104 76.286 7157
10 Afghanistan AFG Asia 2021-09-10 153962 122 83.857 7164
11 Afghanistan AFG Asia 2021-09-11 153982 20 83.857 7167
12 Afghanistan AFG Asia 2021-09-12 153990 8 81 7167
44°F
Partly cloudy
Search
10:53 PM
11/16/2024


PROJECT5: Exploring COVID-19
Cluster Details - Databricks Co
niharikamysoregowda_project5
community.cloud.databricks.com/?o=3094395990686393#notebook/3316071532721946/command/3316071532721956
databricks
New
Workspace
Recents
Search
Catalog
Workflows
Compute
Machine Learning
Experiments
Collapse menu
niharikamysoregowda_project5
Python
File Edit View Run Help Last edit was 1 hour ago
Run all
Terminated
Share
Publish
09:07 PM (6s)
12
Python
# Fill missing GDP values with the average GDP
avg_gdp = gdp_df.select(F.avg("GDP")).collect()[0][0]
merged_df = merged_df.fillna({"GDP": avg_gdp})
# Display the final dataset
display(merged_df)
(8) Spark Jobs
merged_df: pyspark.sql.dataframe.DataFrame = [location: string, iso_code: string ... 66 more fields]
Table
location iso_code continent date total_cases new_... new_Scases total_deaths
1 Afghanistan AFG Asia 2021-09-01 153260 40 76.857 7123
2 Afghanistan AFG Asia 2021-09-02 153306 46 69.143 7127
3 Afghanistan AFG Asia 2021-09-03 153375 69 59.286 7127
4 Afghanistan AFG Asia 2021-09-04 153395 20 55.429 7128
5 Afghanistan AFG Asia 2021-09-05 153423 28 55.714 7132
6 Afghanistan AFG Asia 2021-09-06 153534 111 55.143 7141
7 Afghanistan AFG Asia 2021-09-07 153626 92 58 7144
8 Afghanistan AFG Asia 2021-09-08 153736 110 68 7151
9 Afghanistan AFG Asia 2021-09-09 153840 104 76.286 7157
10 Afghanistan AFG Asia 2021-09-10 153962 122 83.857 7164
11 Afghanistan AFG Asia 2021-09-11 153982 20 83.857 7167
44°F
Partly cloudy
Search
10:53 PM
11/16/2024

9) Create Summary/Descriptive Statistics Table



10) Reporting Results - Higher COVID-19 immunization rates are strongly linked to fewer cases and lower excess mortality, according to the findings. With a correlation coefficient at -0.98, correlation analysis showed a strong inverse link, meaning that areas with higher vaccination rates reported noticeably fewer new cases. This tendency is further supported by visual evidence from bar graphs, which demonstrate that continents like Europe and North America that have extensive vaccination campaigns have continuously had lower case counts and fatality rates than those with less vaccination coverage. Regression analysis was used to quantify the impact even more, emphasizing the importance of vaccination in halting the virus's spread and saving lives. These results highlight how important immunization campaigns are to containing the pandemic and lessening its effects on world health.