Aayush Keshari

CS 5165

PROJECT 5: Exploring COVID-19 Data using Databricks

Links

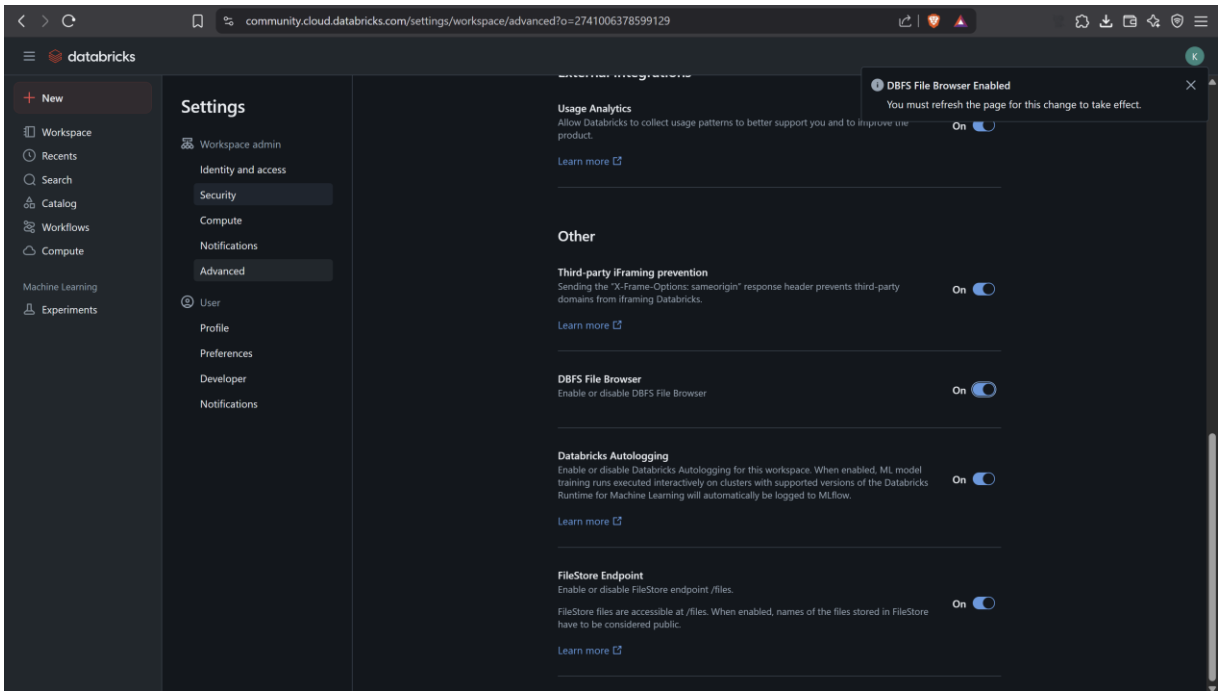
Databricks Notebook Link:

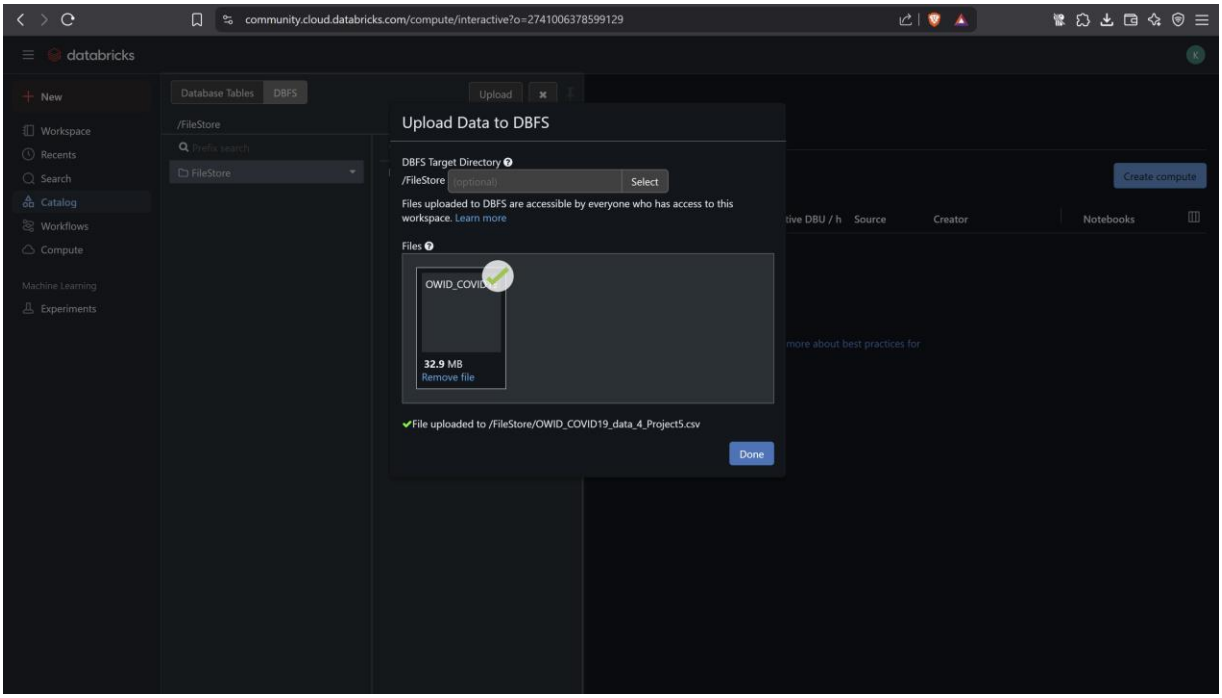
https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/2741006378599129/3779699637522574/1037910492077287/latest.html

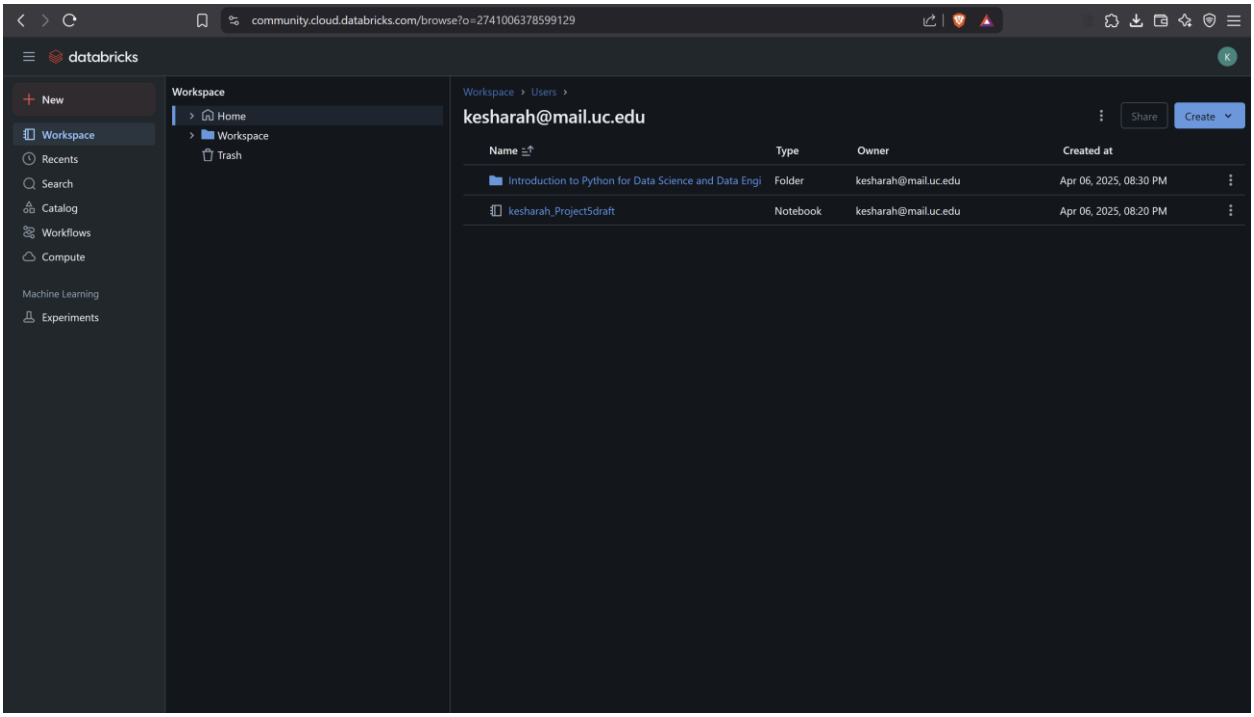
GitHub Repository Link:

https://github.com/aayushkeshari/Project-5-Databricks

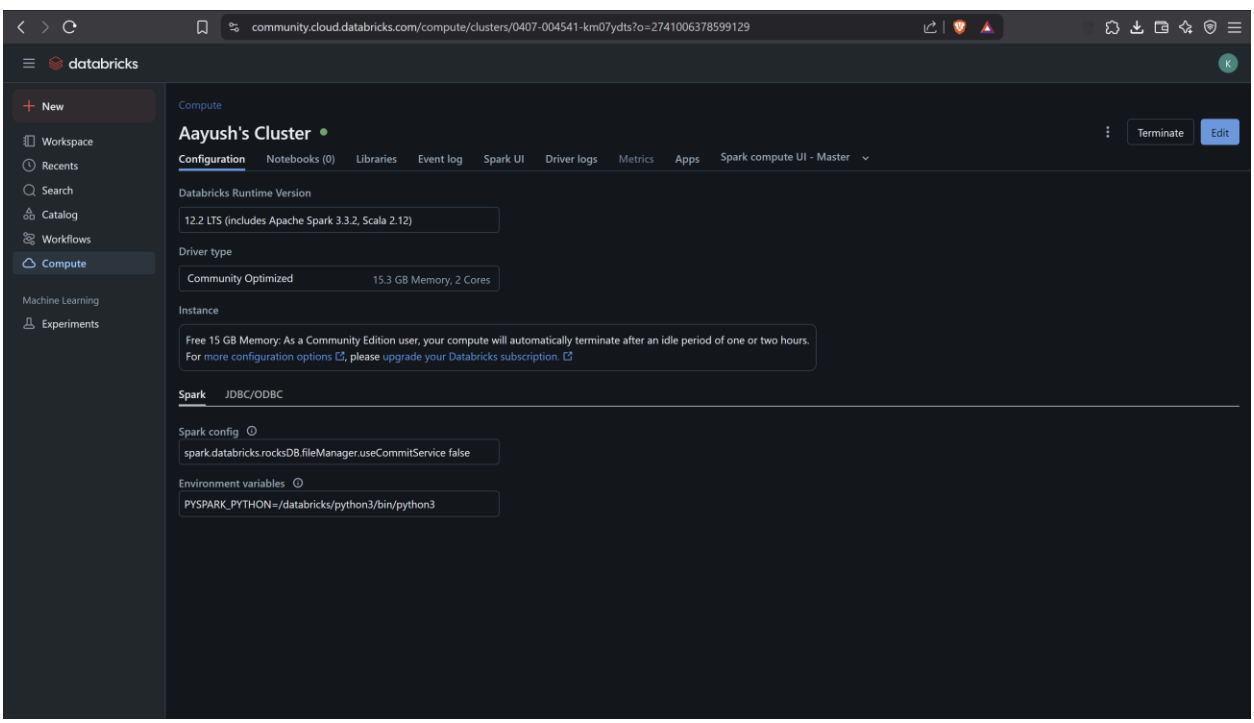
1. Install Databricks and upload the data.



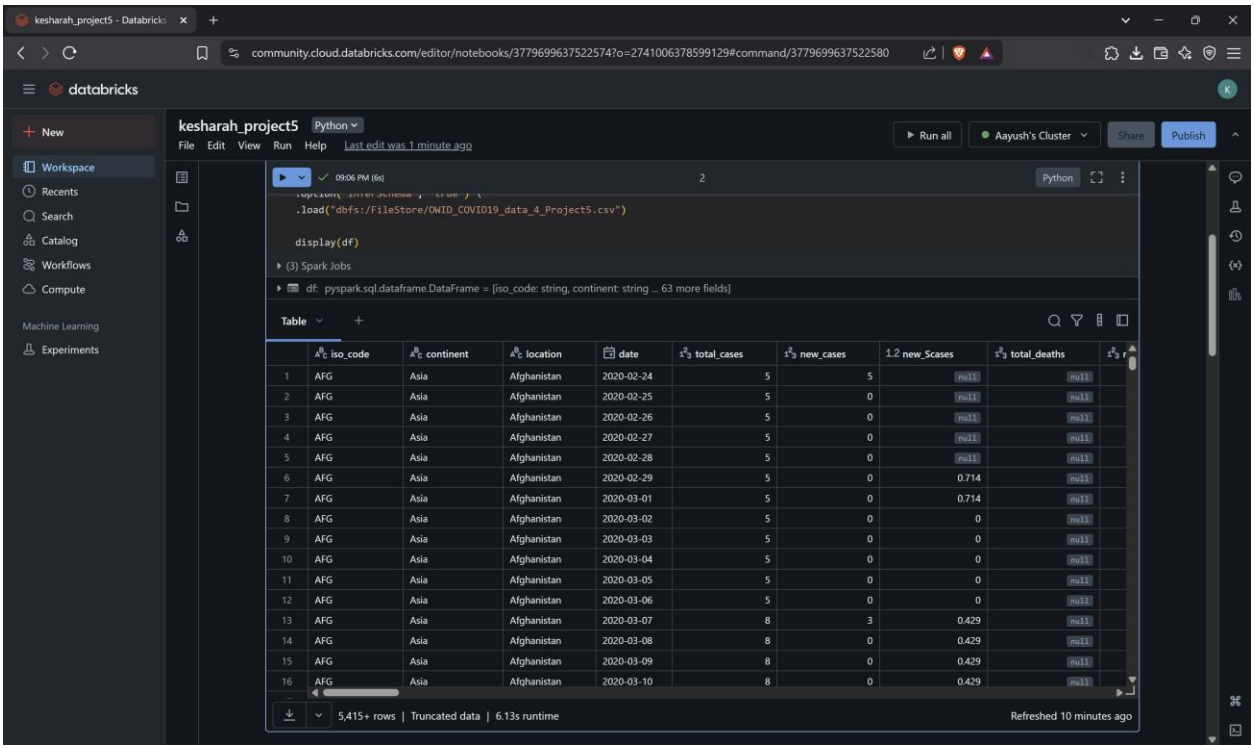




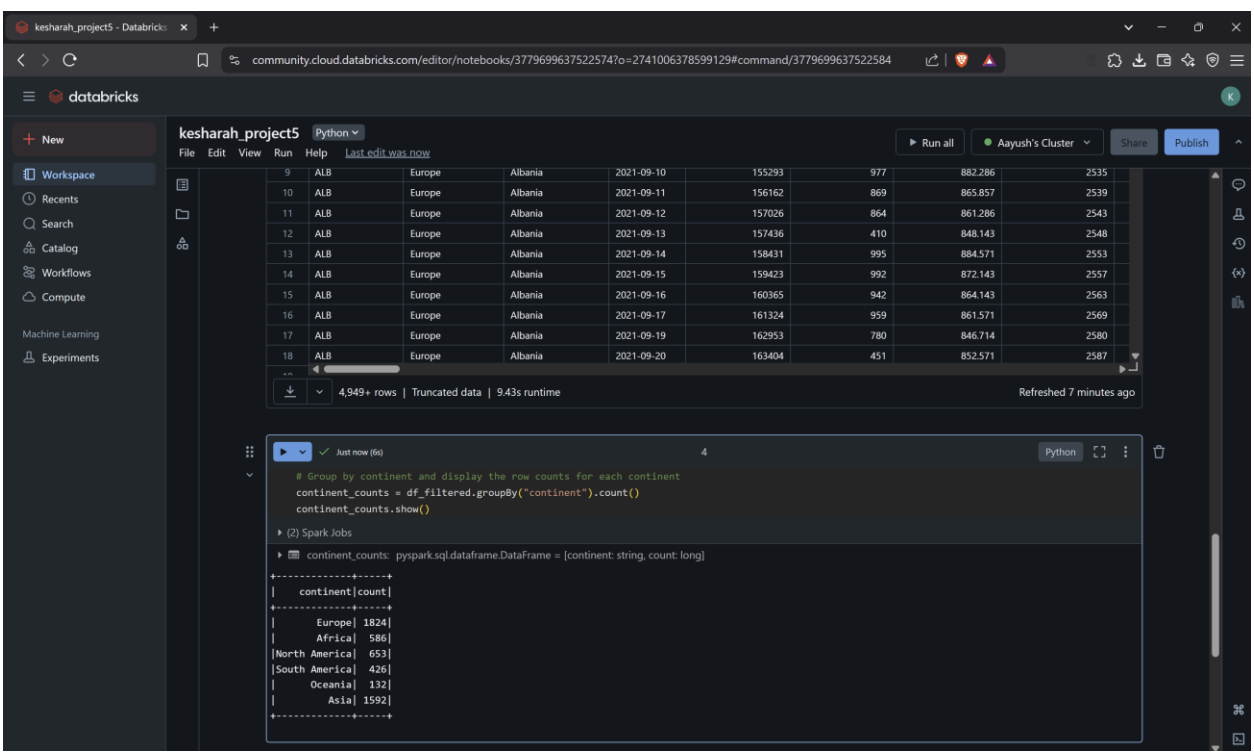
2. Create a Databricks cluster and load the data file using Pandas or PySpark as needed.



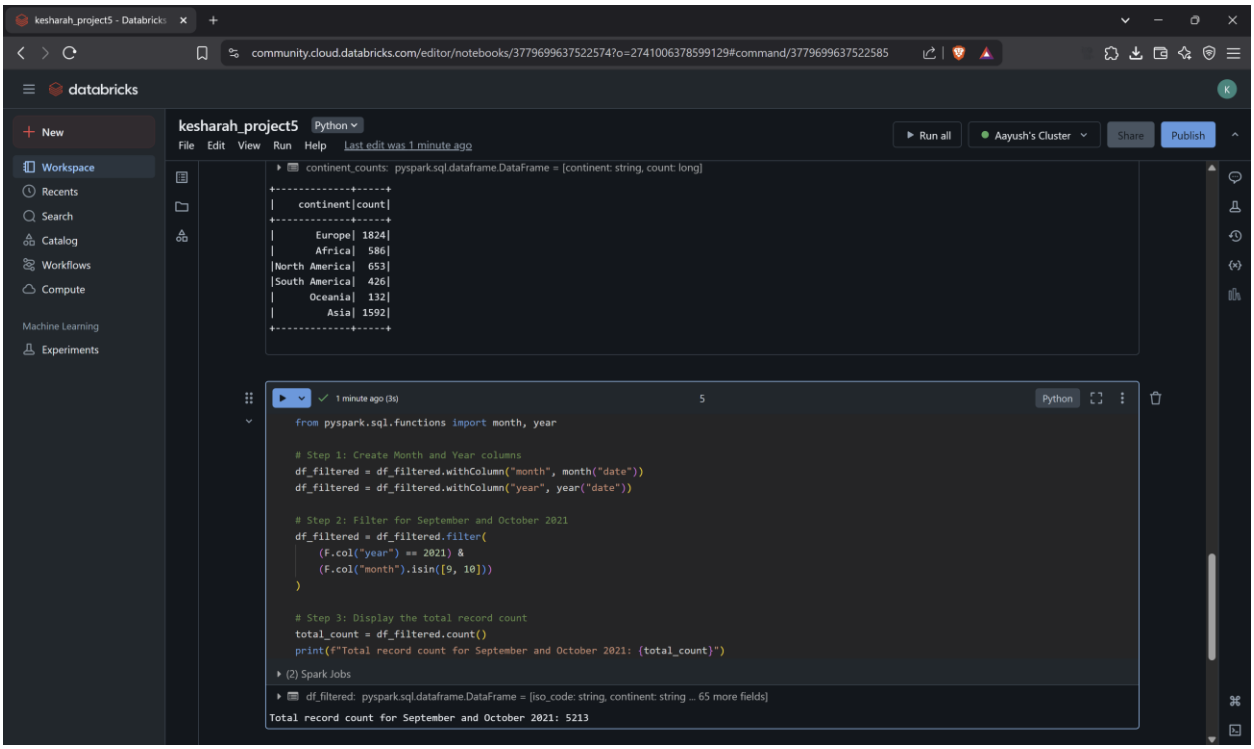
You can see the entire file uploaded here.



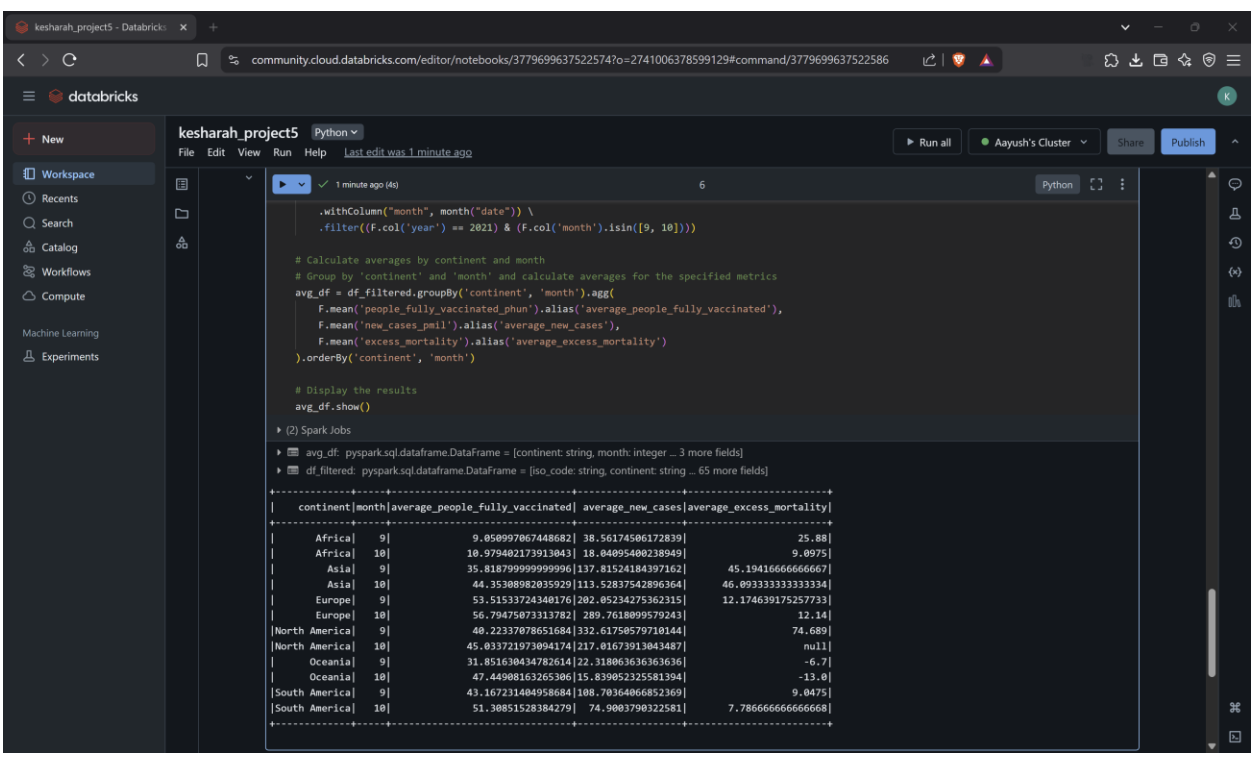
3. Filter Records and display the row count by continent.



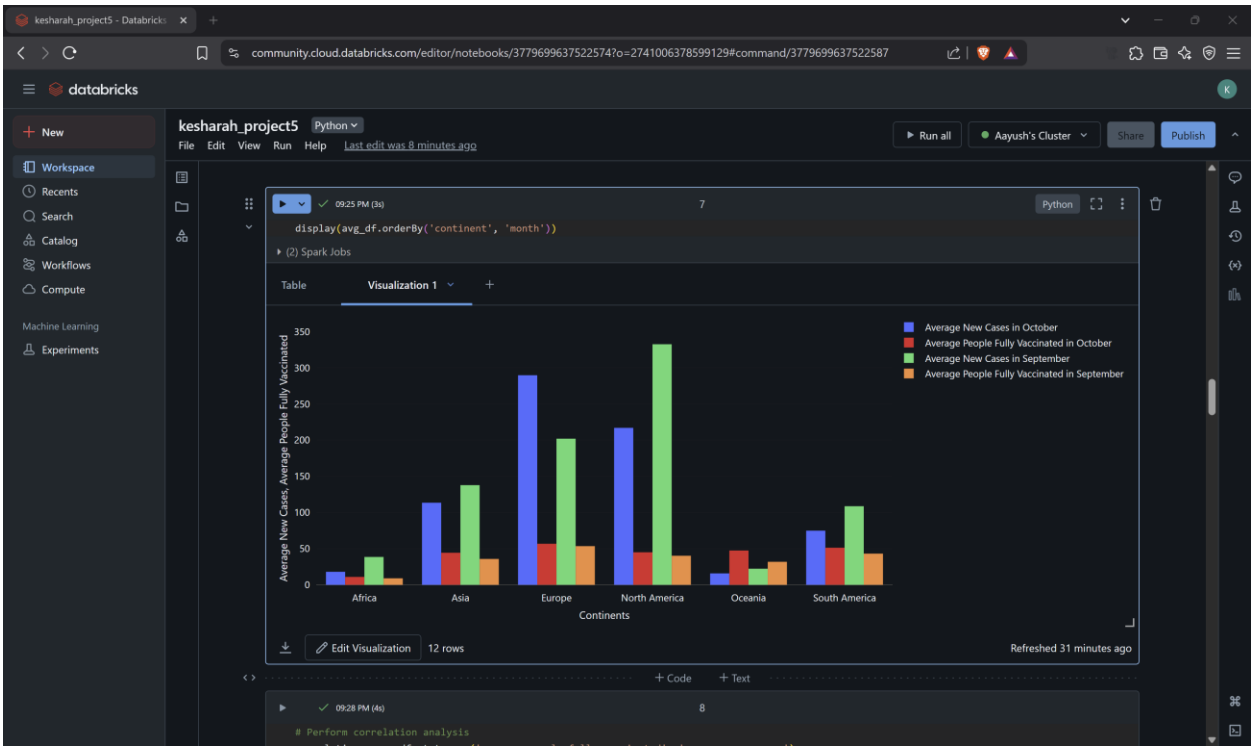
4. Create Month and Year Columns and display the total record count.



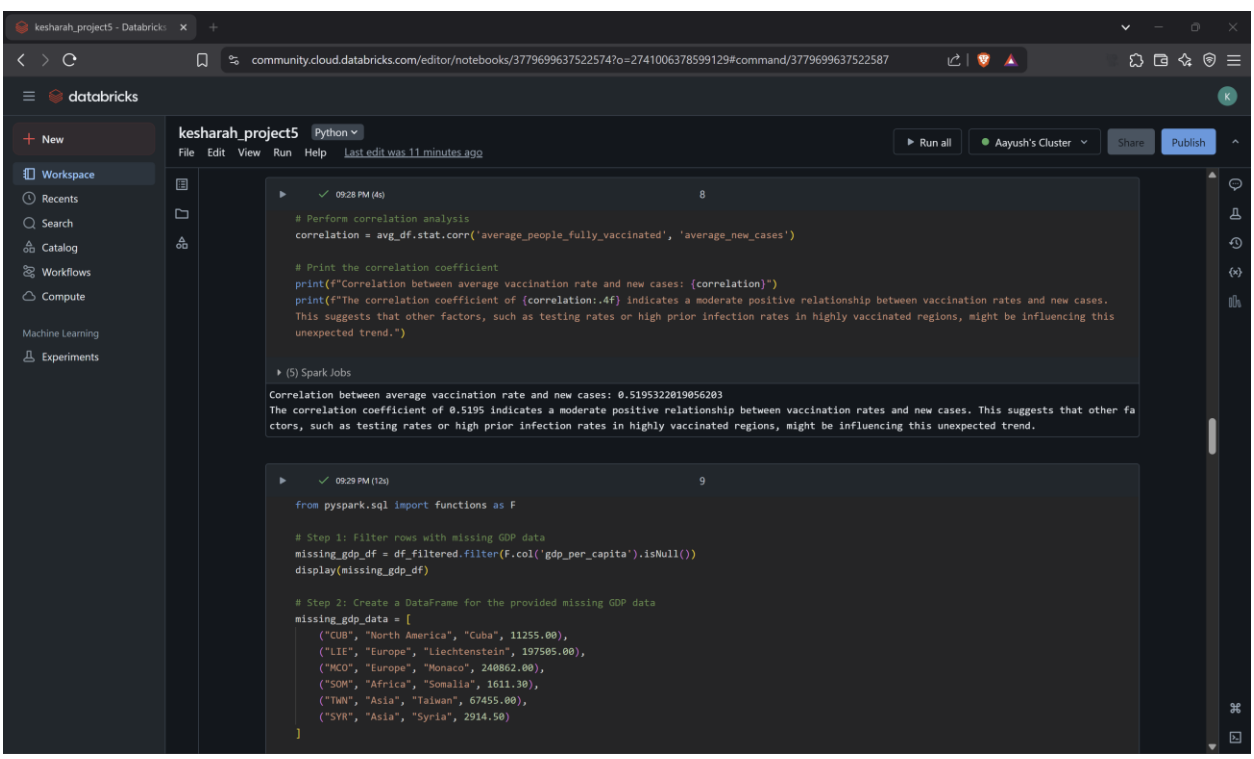
5. Calculate Averages and display the row count by continent by month.



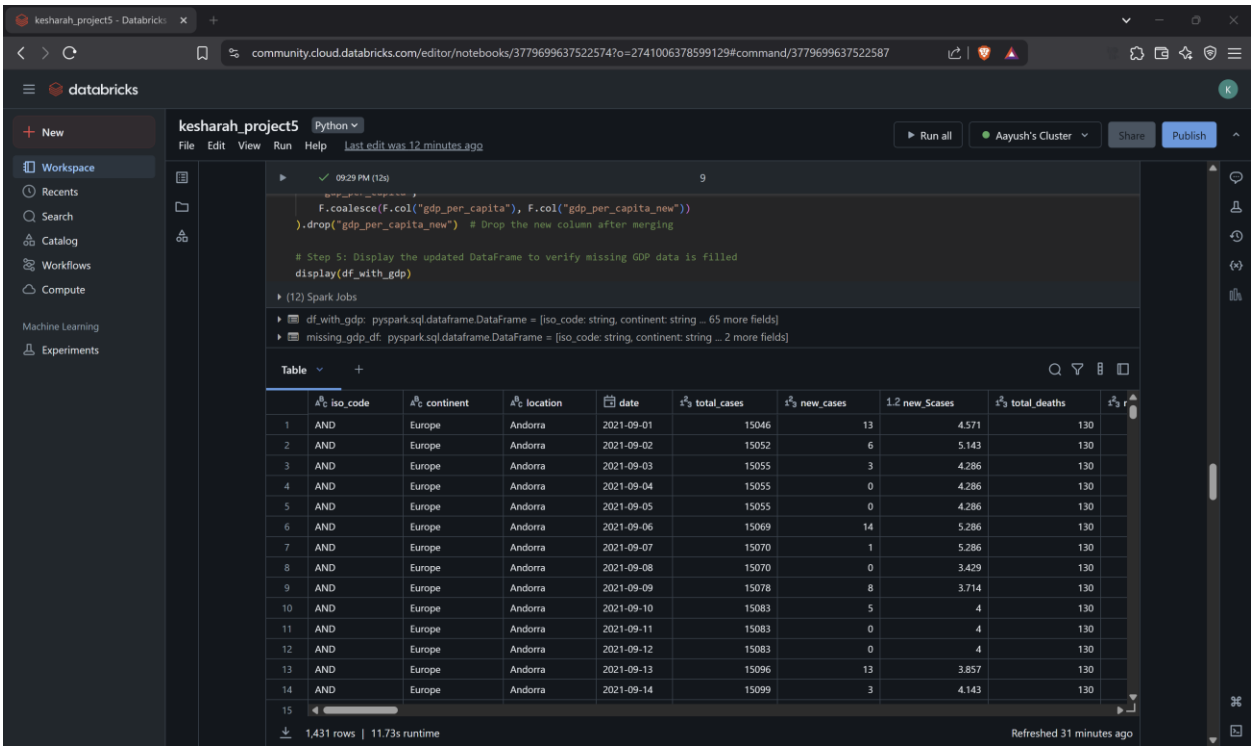
6. Plotting a Bar Chart.



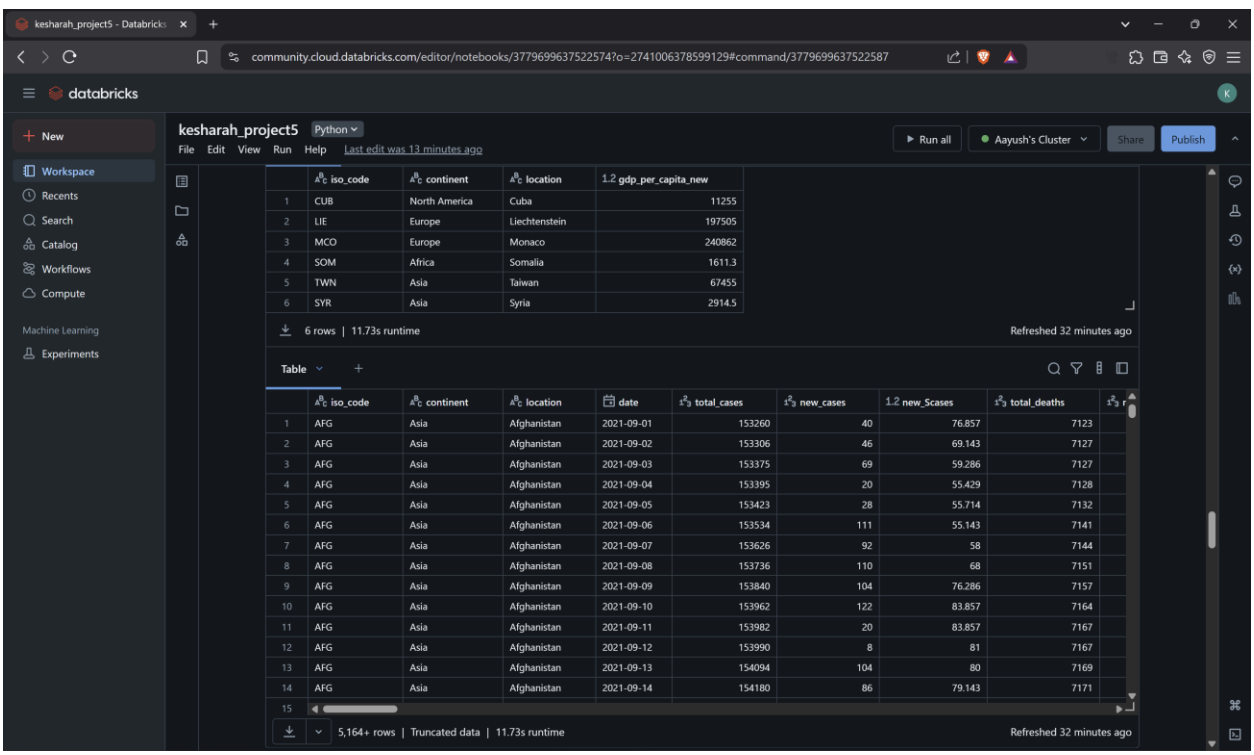
7. Run Correlation Analysis.



8. Fill missing GDP (PPP) per capita.



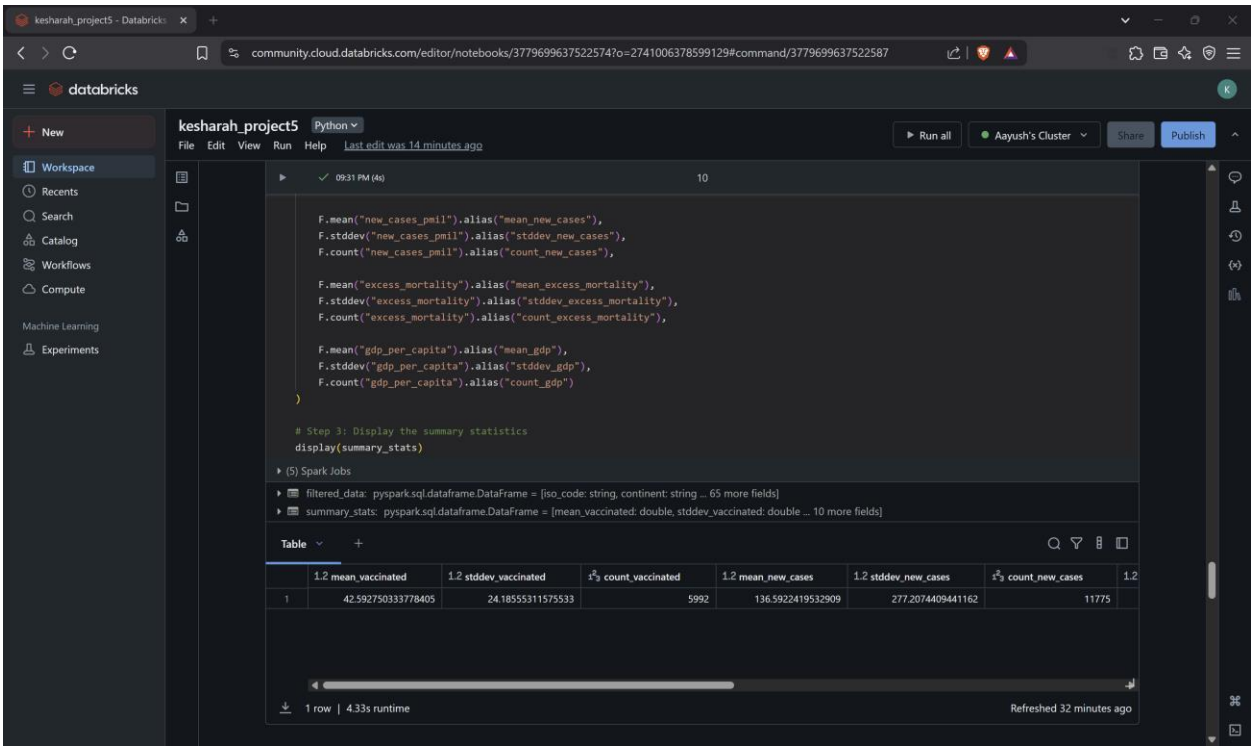
	iso_code	continent	location	date	total_cases	new_cases	new_Scases	total_deaths
1	AND	Europe	Andorra	2021-09-01	15046	13	4.571	130
2	AND	Europe	Andorra	2021-09-02	15052	6	5.143	130
3	AND	Europe	Andorra	2021-09-03	15055	3	4.286	130
4	AND	Europe	Andorra	2021-09-04	15055	0	4.286	130
5	AND	Europe	Andorra	2021-09-05	15055	0	4.286	130
6	AND	Europe	Andorra	2021-09-06	15069	14	5.286	130
7	AND	Europe	Andorra	2021-09-07	15070	1	5.286	130
8	AND	Europe	Andorra	2021-09-08	15070	0	3.429	130
9	AND	Europe	Andorra	2021-09-09	15078	8	3.714	130
10	AND	Europe	Andorra	2021-09-10	15083	5	4	130
11	AND	Europe	Andorra	2021-09-11	15083	0	4	130
12	AND	Europe	Andorra	2021-09-12	15083	0	4	130
13	AND	Europe	Andorra	2021-09-13	15096	13	3.857	130
14	AND	Europe	Andorra	2021-09-14	15099	3	4.143	130



	iso_code	continent	location	gdp_per_capita_new
1	CUB	North America	Cuba	11255
2	LIE	Europe	Liechtenstein	197505
3	MCO	Europe	Monaco	240862
4	SOM	Africa	Somalia	1611.3
5	TWN	Asia	Taiwan	67455
6	SYR	Asia	Syria	2914.5

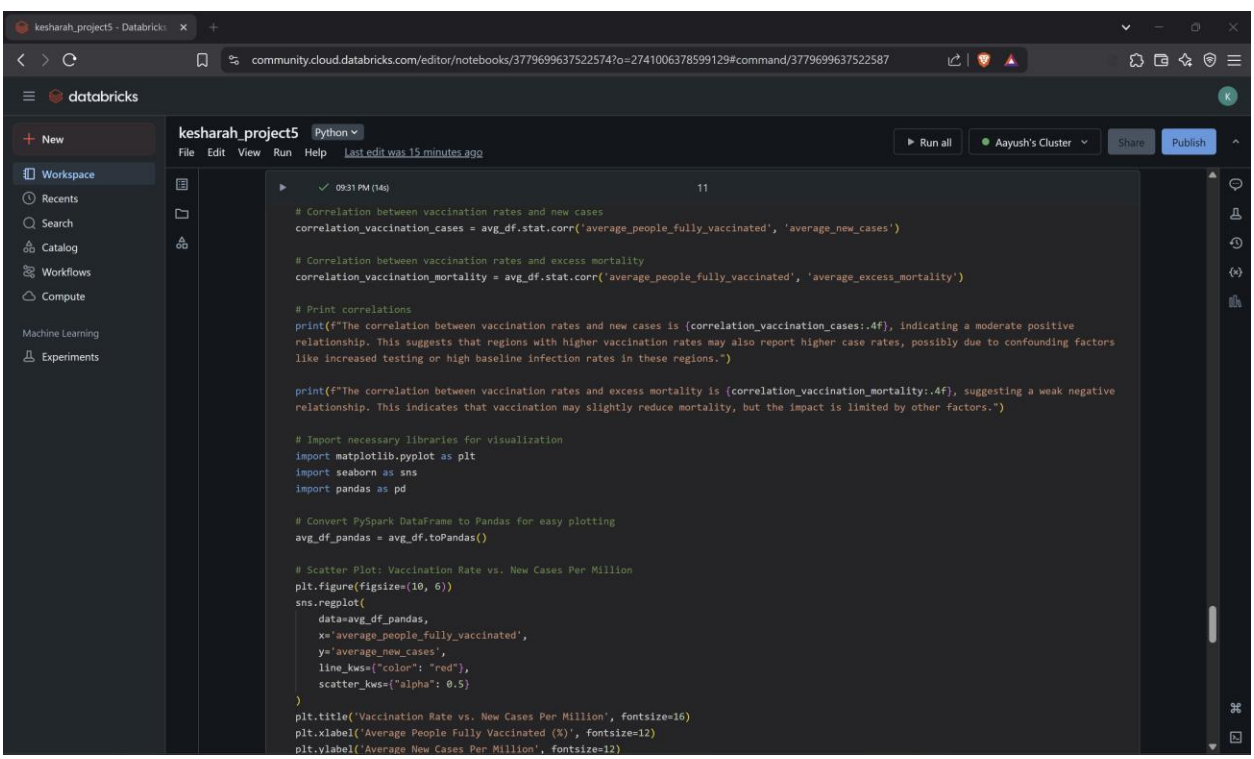
	iso_code	continent	location	date	total_cases	new_cases	new_Scases	total_deaths
1	AFG	Asia	Afghanistan	2021-09-01	153260	40	76.857	7123
2	AFG	Asia	Afghanistan	2021-09-02	153306	46	69.143	7127
3	AFG	Asia	Afghanistan	2021-09-03	153375	69	59.286	7127
4	AFG	Asia	Afghanistan	2021-09-04	153395	20	55.429	7128
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

9. Create Summary/Descriptive Statistics Table.



	mean_vaccinated	stddev_vaccinated	count_vaccinated	mean_new_cases	stddev_new_cases	count_new_cases
1	42.592750333778405	24.18555311575533	5992	136.5922419532909	277.2074409441162	11775

10. Reporting Results.



kesharah_project5 - Databricks

community.cloud.databricks.com/editor/notebooks/3779699637522574?o=2741006378599129#command/3779699637522587

kesharah_project5 Python

File Edit View Run Help Last edit was 16 minutes ago

Run all | Aayush's Cluster | Share | Publish

09:31 PM (14s) 11

```python
    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 highe
r vaccination rates may also report higher case rates, possibly due to confounding factors like increased testing or high baseline infection rates in
these regions.
The correlation between vaccination rates and excess mortality is -0.0940, suggesting a weak negative relationship. This indicates that vaccination m
ay slightly reduce mortality, but the impact is limited by other factors.
```

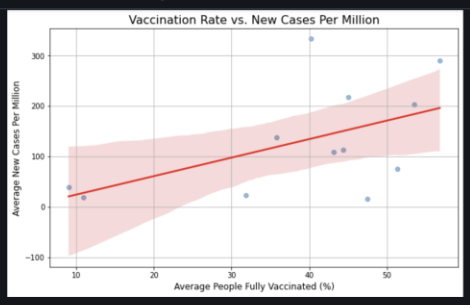
kesharah_project5 - Databricks

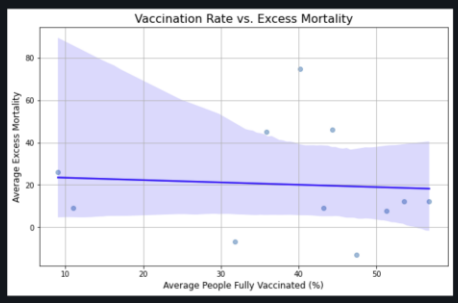
community.cloud.databricks.com/editor/notebooks/3779699637522574?o=2741006378599129#command/3779699637522591

kesharah_project5 Python

File Edit View Run Help Last edit was 18 minutes ago

Run all | Aayush's Cluster | Share | Publish





Findings

The analysis indicates a **moderate positive correlation (r = 0.5195)** between COVID-19 vaccination rates and the number of new cases per million during the months of September and October 2021. This observation suggests that regions with higher vaccination coverage may simultaneously report increased case counts. While this relationship appears counterintuitive, it is likely influenced by confounding variables such as elevated testing rates, higher population densities, or historically elevated baseline infection levels in areas with robust vaccination initiatives. Moreover, vaccination may induce behavioral adaptations such as decreased adherence to masking and social distancing protocols particularly in populations that perceive themselves as protected, further complicating the observed association.

Conversely, the **weak negative correlation (r = -0.0940)** between vaccination rates and excess mortality is consistent with the anticipated protective effect of vaccines against severe disease outcomes and death. Although the magnitude of this association is relatively modest, it provides evidence that vaccination contributes to a reduction in mortality. It is important to acknowledge, however, that other factors such as the quality of healthcare systems, demographic characteristics, and the virulence of circulating variants also play significant roles in shaping mortality trends. These findings underscore the importance of a multifaceted public health strategy, wherein vaccination serves as a cornerstone for mitigating disease severity but must be integrated with additional interventions to effectively control case incidence.