

Final Project – Interactions

DECEMBER 2023

Team

1. Hossein Khoshhal (hkhoshha)
2. Keerthi Jayaram (keerthij)
3. Sathiya Chakravarthy (sathiyac)
4. Sivashakti Komaragiri (skomarag)
5. Tejo Vinay Potti (tpotti)

Exhibits

SA

You

I have a dataframe called `deaths_states` df with these columns:

`Deaths Province_State Date FIPS Lat Long_`

^Date is of format 2021-05-01 in the above case.

FIPS is the code for state starting from 1 till whatever the number is there is in dataset. For single digit add 0 to the FIPS state code.

I have another dataset called `cdc` that has the columns:

`Date FIPS Administered_Dose1_Recip_18Plus Series_Complete_18Plus Metro_status`
and few other columns,.

In this FIPS is of format `xx` where `xx` is the state ID that's similar to the one in `deaths_states` and rest of the other characters is the county.

Now join the two datasets based on two characters of `cdc` from FIPS column and FIPS from `deaths_states` dataframe and also with the date column.

At the end ensure that we have death count for state on each day basis



ChatGPT

To merge the `'deaths_states'` DataFrame with the `'cdc'` DataFrame based on a partial match of the `'FIPS'` codes and the `'Date'` column, and to ensure that the `'Deaths'` count is on a daily basis for each state, you will follow these steps:

1. **Prepare the `'FIPS'` Codes:** In `'deaths_states'`, ensure that the `'FIPS'` codes are two digits (prepend a 0 to single-digit codes). In `'cdc'`, extract the first two characters from the `'FIPS'` column to match the state code.
2. **Merge the DataFrames:** Perform a merge operation based on the prepared `'FIPS'` codes and the `'Date'` column.
3. **Aggregate Deaths:** If necessary, aggregate the deaths count on a daily basis for each state.



You

how do you remove the current date (deaths) columns and then include the recalculated daily death column



ChatGPT

To remove the current cumulative death columns and then include the recalculated daily death column, you'll first calculate the daily deaths as discussed previously, and then drop the original cumulative death columns from your DataFrame. After that, you can add the new daily death column. Here's how you can do it:

1. **Calculate the Daily Deaths:** Using the `'diff'` method on the cumulative death columns.
2. **Drop the Cumulative Death Columns:** Remove the original date columns that represent cumulative deaths.
3. **Add the Daily Deaths Column:** Append the calculated daily deaths to the DataFrame.



You

gettign error: ValueError: invalid literal for int() with base 10: '1001.0'



ChatGPT

The error you're encountering indicates that there are string representations of floating point numbers in your `'FIPS'` column, which cannot be directly converted to integers. This situation can arise if your dataset has some inconsistencies or if it was read into the DataFrame in a way that altered the data types.

To resolve this, you'll need a more robust method to convert these string representations of floats to integers. Here's how you can do it:

Step 1: Handle Non-numeric and NaN Values

Before conversion, ensure that all non-numeric values and NaN values are appropriately handled. This step is crucial to avoid conversion errors.

python

Copy code

```
import pandas as pd
import numpy as np

# Convert FIPS to strings, handle NaN and non-numeric values
covid_deaths_long['FIPS'] = pd.to_numeric(covid_deaths_long['FIPS'], error
```





You

but we r doing number of ppl as opposed to pct becuz of diff popp sizes ri

Creating categorical variables for vaccination rates

```
def categorize_vaccination_rate(rate):
```

```
    if rate < 30:
```

```
        return 'Low'
```

```
    elif 30 <= rate < 60:
```

```
        return 'Moderate'
```

```
    else:
```

```
        return 'High'
```

```
merged_df['Vaccination_Rate_Category'] =
```

```
merged_df['Series_Complete_18PlusPop_Pct'].apply(categorize_vaccinati
```



ChatGPT

You are correct in pointing out that categorizing vaccination rates based on the percentage of the population might be less accurate if these percentages are calculated from populations of different sizes. The approach of using the actual number of people vaccinated, as opposed to percentages, can provide a more consistent and comparable measure across different areas with varying population sizes.



You

so the number is coming in decimal, how can no. of ppl be in decimal?




ChatGPT

You're correct that the number of people cannot be a decimal. The issue arises due to the representation of percentages and population figures. When the percentage of vaccinated people (which is in percentage form) is multiplied by the total population (an integer), the result should ideally be rounded to the nearest whole number, as it represents a count of individuals.

Here's how you can modify the code to round the result to the nearest whole number:

python

 Copy code

```
df_covid19['Number_OneDose'] = ((df_covid19['Administered_Dose'] * df_covid19['Total_Population']) / 100).round()
```





You

i see that adult pop values are diff for diff counties. i wanna add a column to find number of people (1dosepercentage of the population)



ChatGPT

To add a column that calculates the number of people who have received at least one dose of the vaccine in each county, you'll need to use the percentage of the adult population that has received at least one dose and the total population of the county. Assuming you have these two pieces of information in your dataset, you can calculate this number as follows:

1. Identify the column representing the percentage of the adult population that has received at least one dose. Let's call this column **`Administered_Dose1_Recip_18PlusPop_Pct`**.
2. Identify the column representing the total population of the county. Let's call this column **`Population`**.
3. Create a new column, for example, **`Number_OneDose`**, which is the product of **`Administered_Dose1_Recip_18PlusPop_Pct`** (as a

Perform interesting visualizations (somewhat complex) and NOT ordinary bar plot or pie chart. DO with Plotly and Seaborn and analyse the trends. Do this in a way such that we can understand what's going on and how data behaves. DO some interactive interesting plots too. The columns I got in the df are : ['Date', 'State_FIPS', 'Administered_Doses_65Plus_Daily', 'Series_Complete_65Plus_Daily', 'Booster_Doses_18Plus_Daily', 'Series_Complete_18Plus_Daily', 'Administered_Doses_18Plus_Daily', 'Census2019_18PlusPop', 'Deaths', 'Confirmed', 'Province_State', 'FIPS', 'Lat', 'Long_']



Data Analysis

The dataset contains information about COVID-19 vaccination and associated deaths for each state in the USA, spanning from May 2021 to September 2022. The columns include daily administered doses, series completion, booster doses for different age groups, population data, and COVID-19 related deaths and confirmed cases, along with geographical information.

To analyze and visualize these data, I will create a series of plots using Seaborn and Plotly. These visualizations will help us understand trends and correlations in the data, focusing on more complex and informative visualizations rather than basic bar or pie charts.

1. **Time Series Analysis:** Plotting vaccination and death rates over time to observe trends.
2. **Heatmaps:** Showing correlations between different variables.
3. **Interactive Plots:** Using Plotly for interactive visualizations, such as geographic



You

How do I integrate MLflow into my COVID-19 data analysis project for tracking experiments if I'm coding on Google Colab? Give me steps and keep it direct.

< 3 / 3 >



ChatGPT

Integrating MLflow into your COVID-19 data analysis project on Google Colab involves several steps. MLflow is a platform for managing the machine learning lifecycle, including experimentation, reproducibility, and deployment. Here's a direct, step-by-step guide:

Step 1: Install MLflow

In a cell in your Google Colab notebook, install MLflow:

python

Copy code


```
!pip install mlflow
```

Step 2: Start an MLflow Tracking Server (Optional)

2. **Start MLflow Tracking Server:** MLflow tracks experiments to log parameters, code versions, metrics, and output files. In Colab, you can start a local MLflow tracking server or connect to a remote server.

To start a local server, use:

python

 Copy code


```
get_ipython().system_raw("mlflow ui --port 5000 &")
```

This will start the MLflow tracking UI in the background on port 5000.

3. **Access MLflow UI:** Google Colab does not allow direct access to local ports, so you need to use a tunneling service like ``ngrok`` to access the MLflow UI.

First, install ``ngrok``:

python

 Copy code

```
!wget -q https://bin.equinox.io/c/4VmDzA7iaHb/ngrok-stable-  
!unzip -qq ngrok-stable-linux-amd64.zip
```

Then, start an ``ngrok`` tunnel:

