Asking Data Analysis Questions with PandasAI

LEVI DOLAN, MLIS

DATA SERVICES LIBRARIAN

INDIANA UNIVERSITY SCHOOL OF MEDICINE

INDIANAPOLIS, IN

What is PandasAI?

Definition

"PandasAI is a Python library that integrates generative artificial intelligence capabilities into Pandas, making dataframes conversational."

-GitHub About

https://github.com/gventuri/pandas-ai

What is PandasAI?

Purpose

PandasAI connects the Pandas library to a large language model so that data analysis tasks can be accomplished through natural language processing.

What is PandasAI?

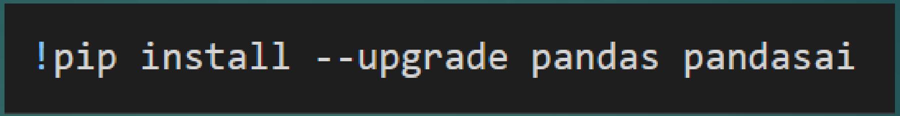
Attribution

There are 40 contributors to this library's GitHub repository. The repository owner is Gabriele Venturi, a software engineer based in Germany.

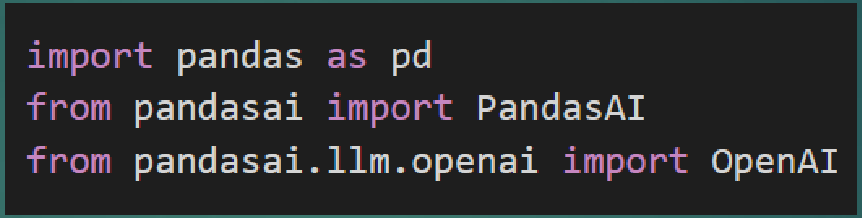
How to Try it Out

in Google Colab

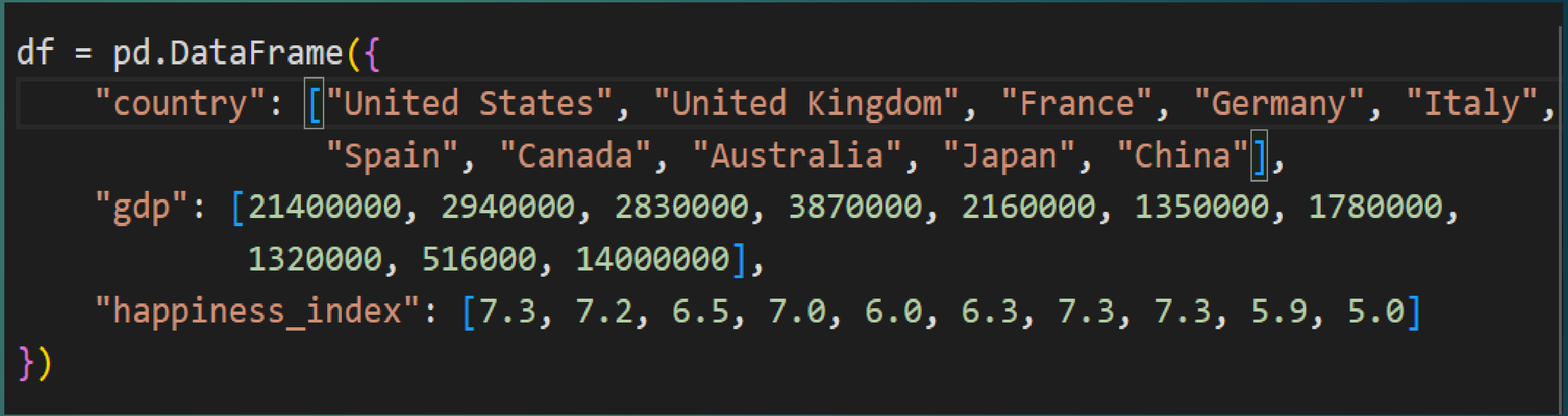
1. Install library
2. Import dependencies
3. Create data
4. Connect to OpenAI
5. Ask questions



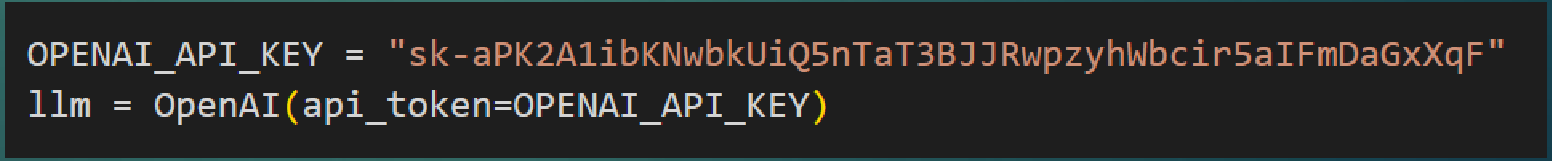
```
!pip install --upgrade pandas pandasai
```



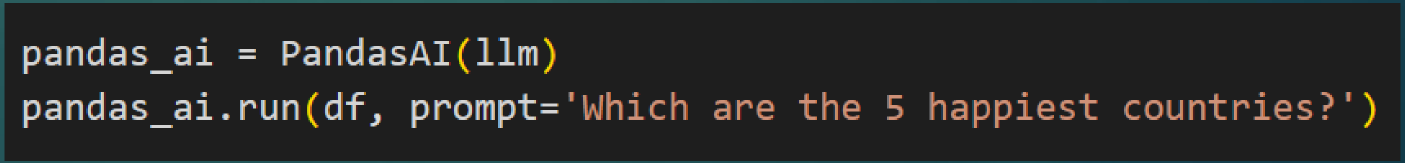
```
import pandas as pd
from pandasai import PandasAI
from pandasai.llm.openai import OpenAI
```



```
df = pd.DataFrame({
    "country": ["United States", "United Kingdom", "France", "Germany", "Italy",
                "Spain", "Canada", "Australia", "Japan", "China"],
    "gdp": [21400000, 2940000, 2830000, 3870000, 2160000, 1350000, 1780000,
            1320000, 516000, 14000000],
    "happiness_index": [7.3, 7.2, 6.5, 7.0, 6.0, 6.3, 7.3, 7.3, 5.9, 5.0]
})
```



```
OPENAI_API_KEY = "sk-aPK2A1ibKNwbkUiQ5nTaT3BJJRwpzyhWbcir5aIFmDaGxXqF"
llm = OpenAI(api_token=OPENAI_API_KEY)
```



```
pandas_ai = PandasAI(llm)
pandas_ai.run(df, prompt='Which are the 5 happiest countries?')
```

Test Questions

- Which country is happiest?
- Which country is richest?
- Which country lost World War 2?

More Test Questions

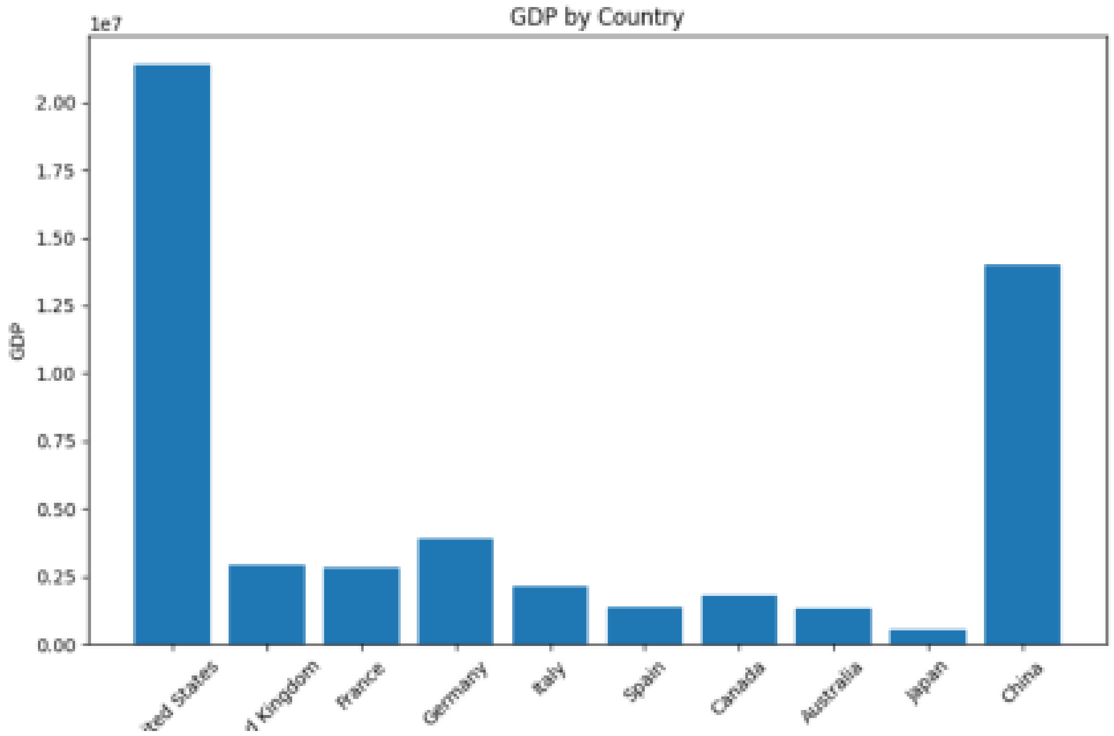
- Which country has the most overweight people?
- Which country will win World War 3?
- Which country is best?

Most recent test

“Unfortunately, I was not able to answer your question, because of the following error:\n\nThe truth value of a DataFrame is ambiguous. Use a.empty, a.bool(), a.item(), a.any() or a.all().\n”

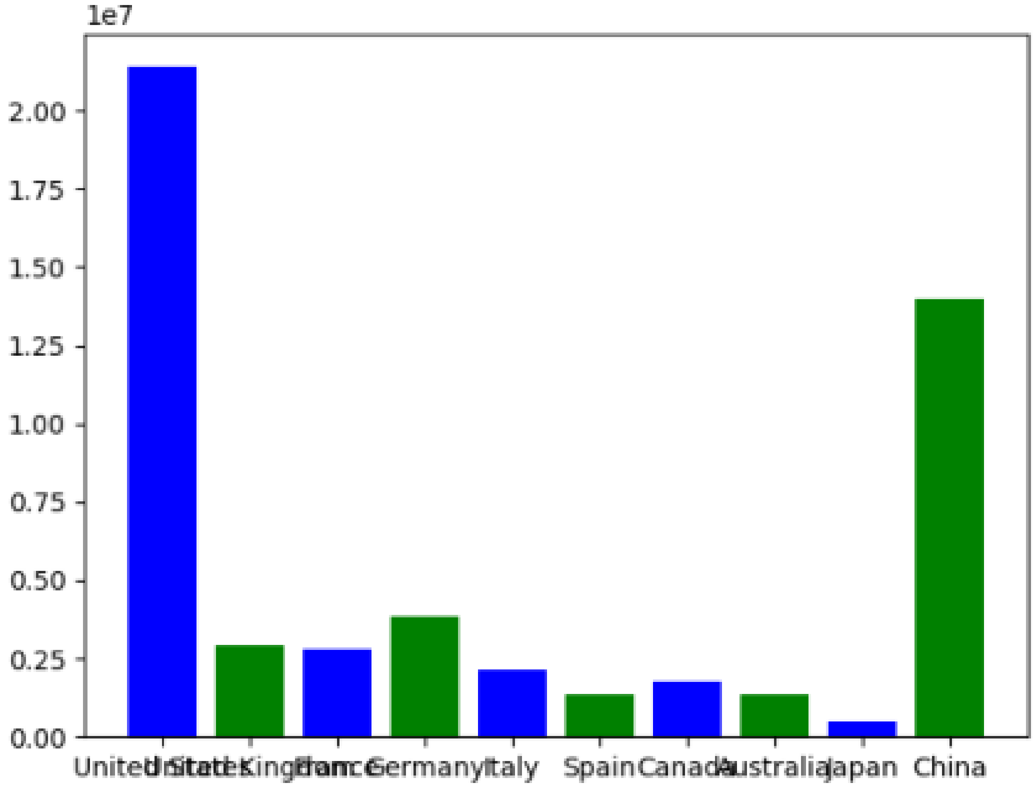
Visualization test

- “plot the histogram of countries showing for each the GDP”



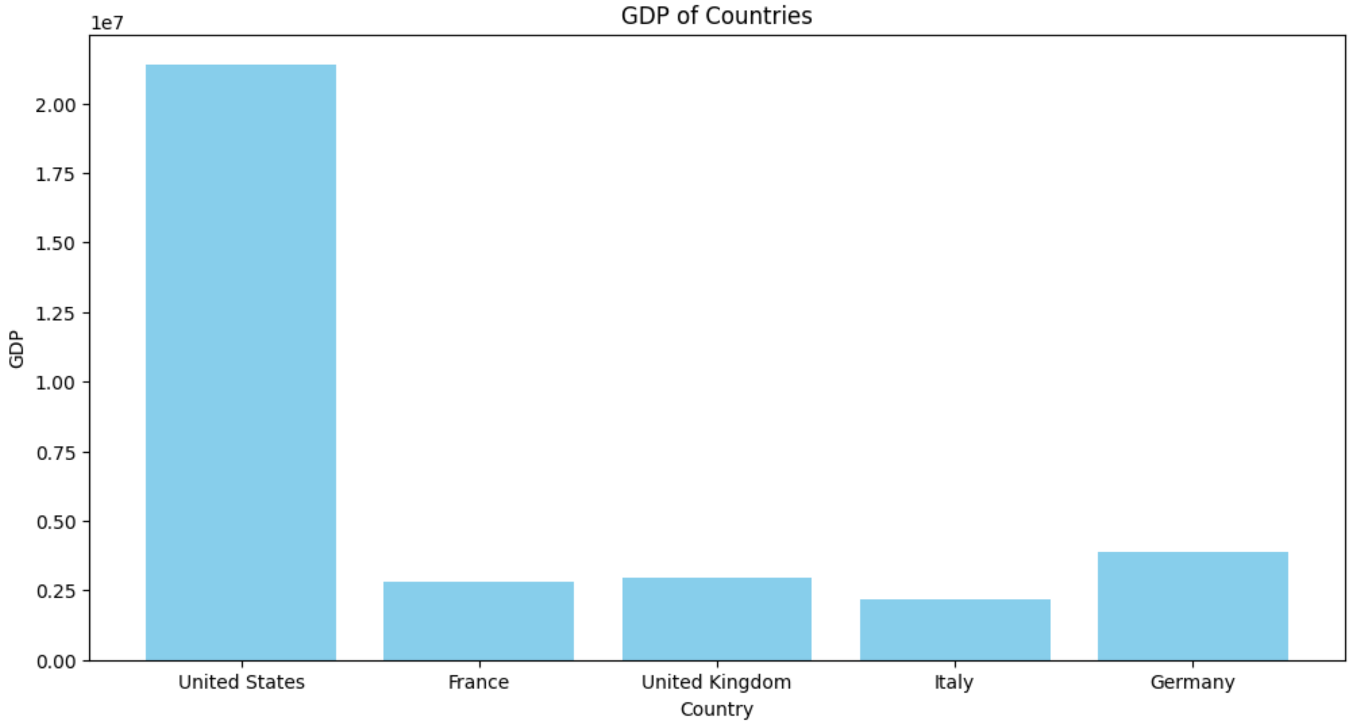
Visualization test

- “plot the histogram of countries showing for each the GDP, alternating blue and green for each bar”



Visualization test

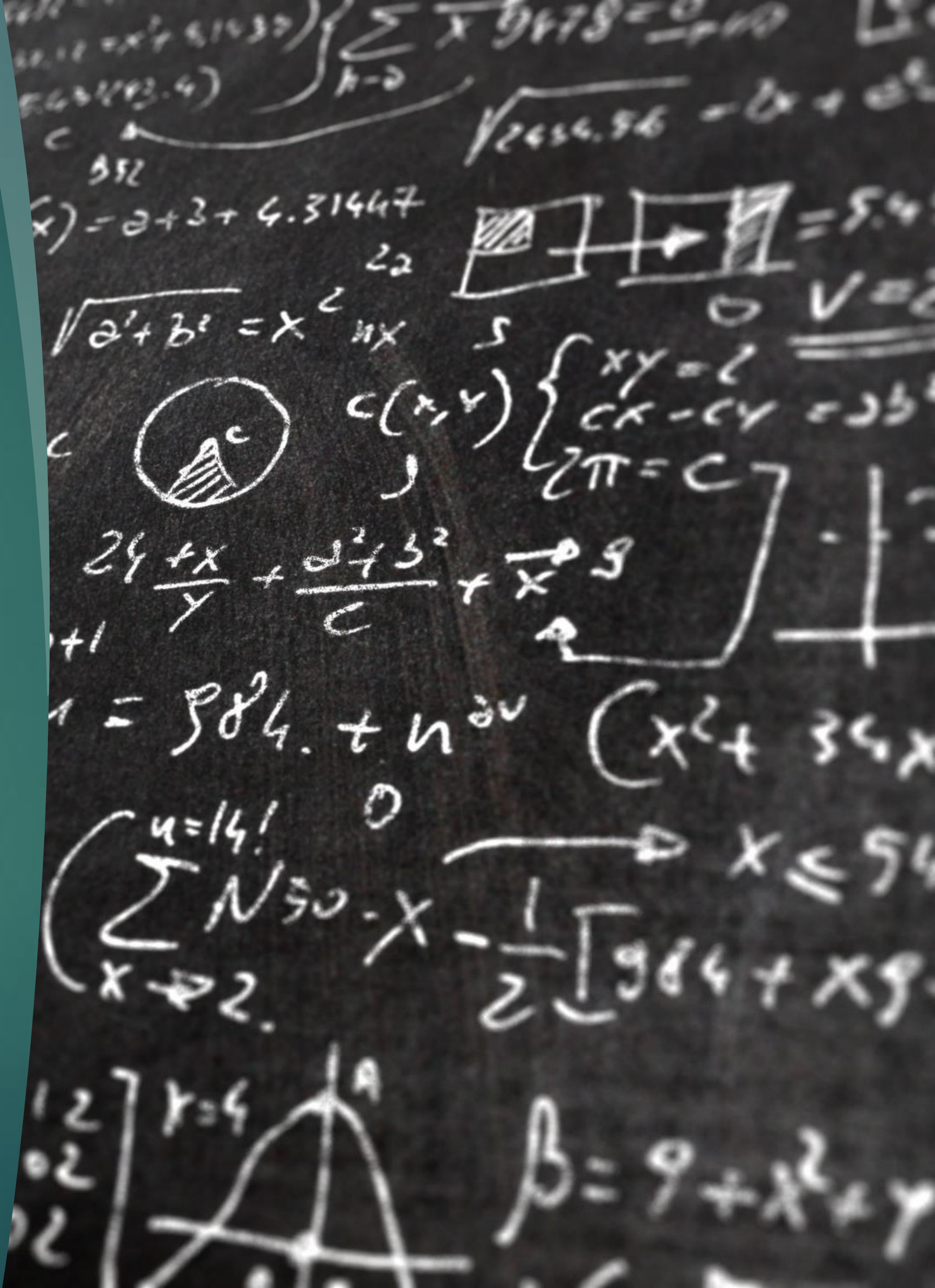
- "plot a large histogram of countries showing for each the GDP"



Conclusions

PandasAI is:

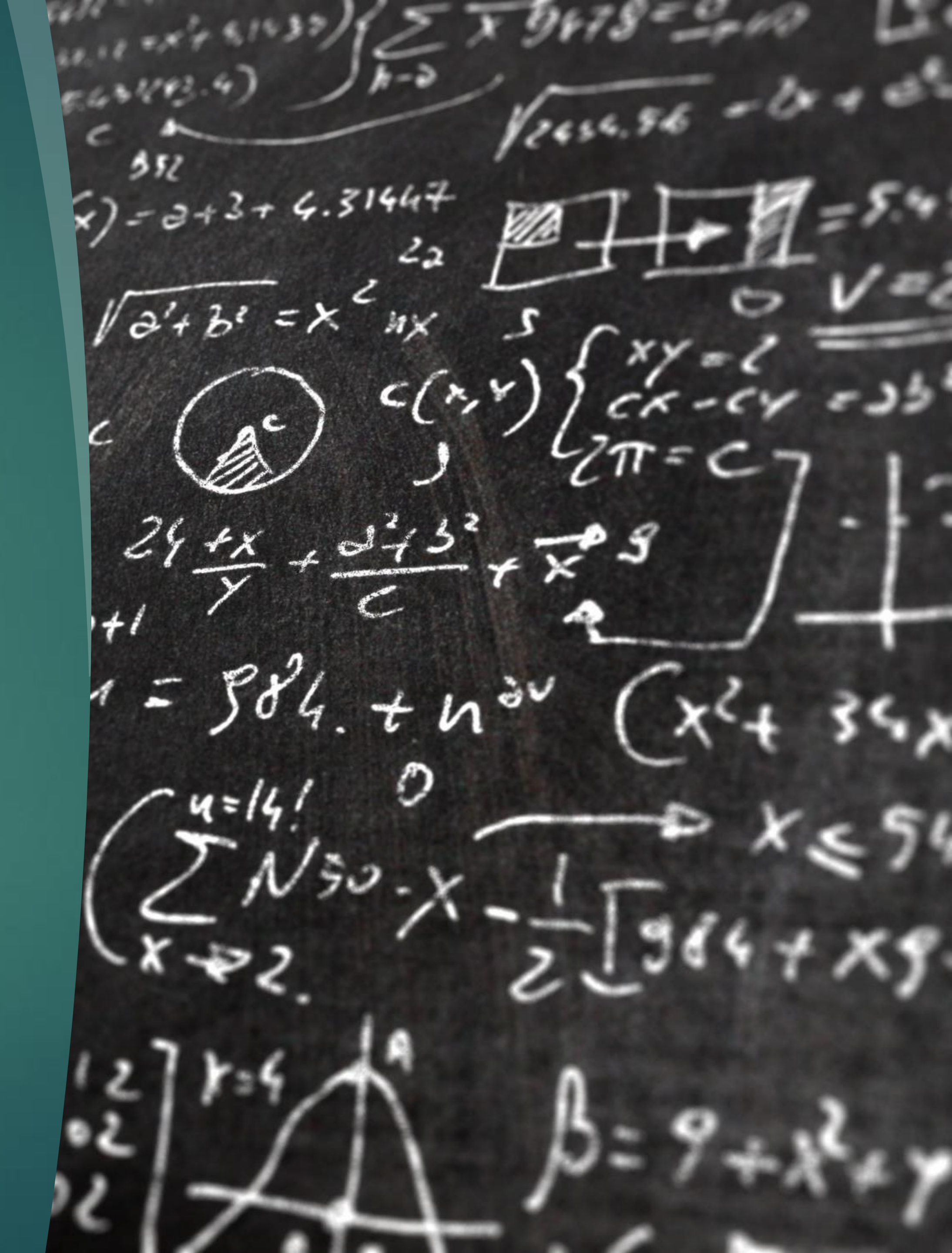
- A tool that can streamline exploration of a dataset's scope.
- A tool that can fill the gap between coding each instruction and app presets.
- A tool that is faster to learn than a GUI with extensive parameter-setting.



Conclusions

PandasAI is:

- NOT an onramp to understanding qualitative information.
- NOT a solution that supports validation.
- NOT a way to make your research reproducible.
- NOT completely free.



Thank you!

Levi Dolan

Data Services Librarian

Indiana University School of Medicine

Indianapolis, IN

dolanl@iu.edu

Google Colab Notebook:

https://colab.research.google.com/drive/1A-KyksGZ5eOaaEmeNJTGd_CDZCBKrr68?usp=sharing