Leveraging ChatGPT for Data Analysis and Visualization

Dr. Benjamin Walter

University of the Virgin Islands
www.uvi.edu

SPECIALIZING IN FUTURES

HISTORICALLY AMERICAN.
UNIQUELY CARIBBEAN.
GLOBALLY INTERACTIVE.

Emerging Caribbean Scientists Seminar

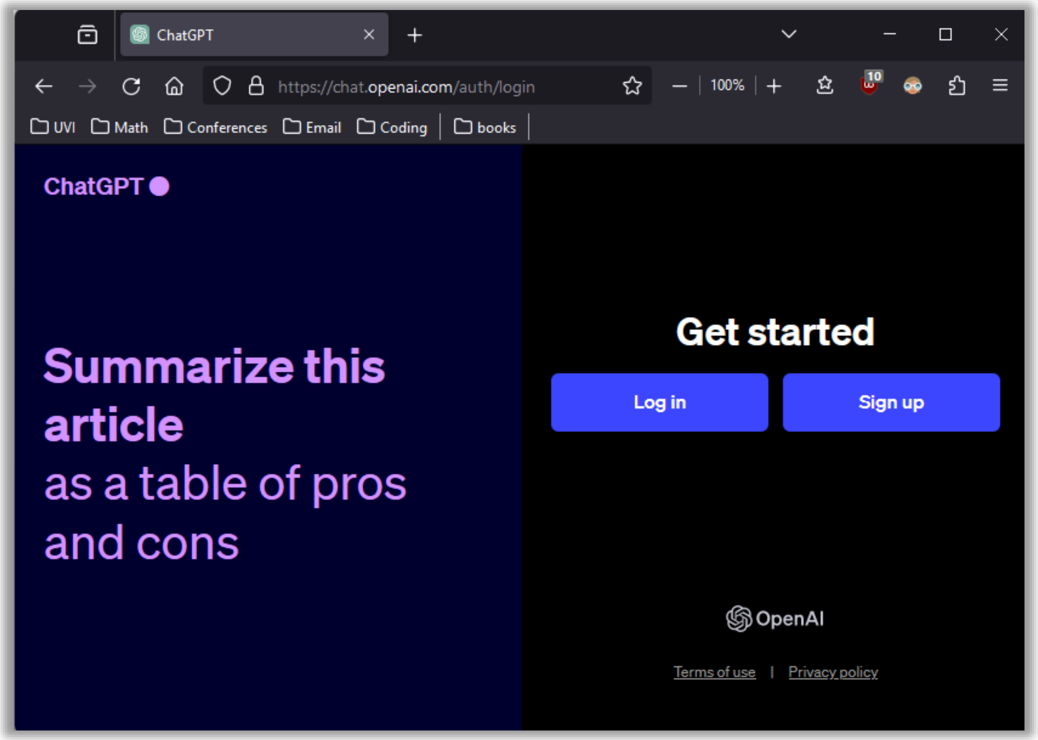
January 31, 2024

Leveraging ChatGPT for Data Analysis and Visualization

ChatGPT (Generative Pre-trained Transformer)

- A type of (predictive, stochastic) Large Language Model (LLM) AI
- Developed by OpenAI, released in Nov 2022 https://chat.openai.com
- GPT-3.5 is free to use
- GPT-4 requires subscription (limit 40 messages / 3 hours)
- Microsoft Copilot is based on GPT-4
  - Office365 Copilot Pro
  - 24H2 "Hudson Valley" Windows update in fall 2024

COPILOT



Summarize this article as a table of pros and cons

> "It saves me 4-5 hours of work per week" says local data scientist.

University of the Virgin Islands
www.uvi.edu
SPECIALIZING IN FUTURES
HISTORICALLY AMERICAN. UNIQUELY CARIBBEAN. GLOBALLY INTERACTIVE.

Dr. Benjamin Walter

ECS **January 31, 2024**

Leveraging ChatGPT for Data Analysis and Visualization

“Cloze probability”
“Semantic priming”

ChatGPT is not a “computer”

- Not computational
- Not deterministic
- Not “garbage in – garbage out”
- Not “does exactly what you tell it”

Note: Song lyrics aren’t exactly deterministic
Example: Love Shack (B-52’s)
“Tin Roof! Rusted!” or “Henry! Busted!”
(this is called a “Mondegreen”)

ChatGPT views the following prompts similarly:

- $2 + 2 =$
- Complete the statement: “2 + …”
- Complete the statement: “I like cats and …”
- What is the next line:

“Hey, I just met you, and this is crazy
But here’s my number …”

So call me, maybe?

University of the Virgin Islands
www.uvi.edu

SPECIALIZING IN FUTURES
HISTORICALLY AMERICAN.
UNIQUELY CARIBBEAN.
GLOBALLY INTERACTIVE.

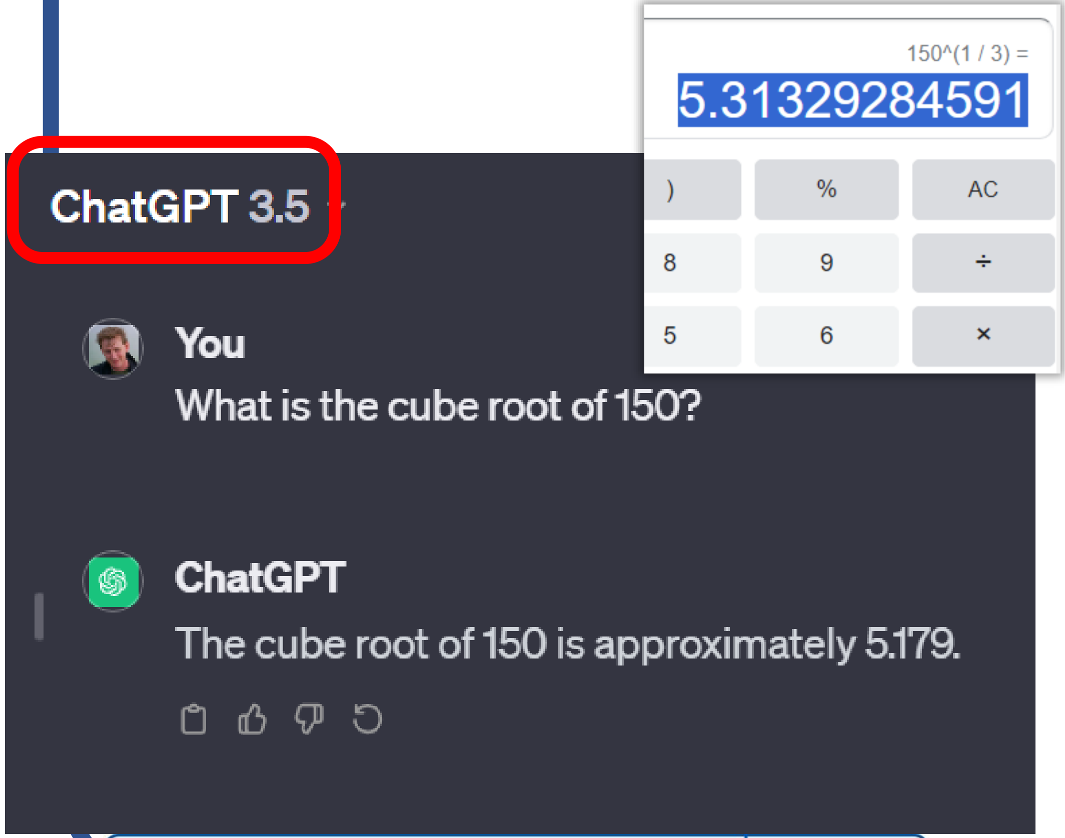
Dr. Benjamin Walter

ECS **January 31, 2024**

Leveraging ChatGPT for Data Analysis and Visualization

ChatGPT is famously bad at computation.

Solution: Give it a calculator!



150^(1 / 3) =

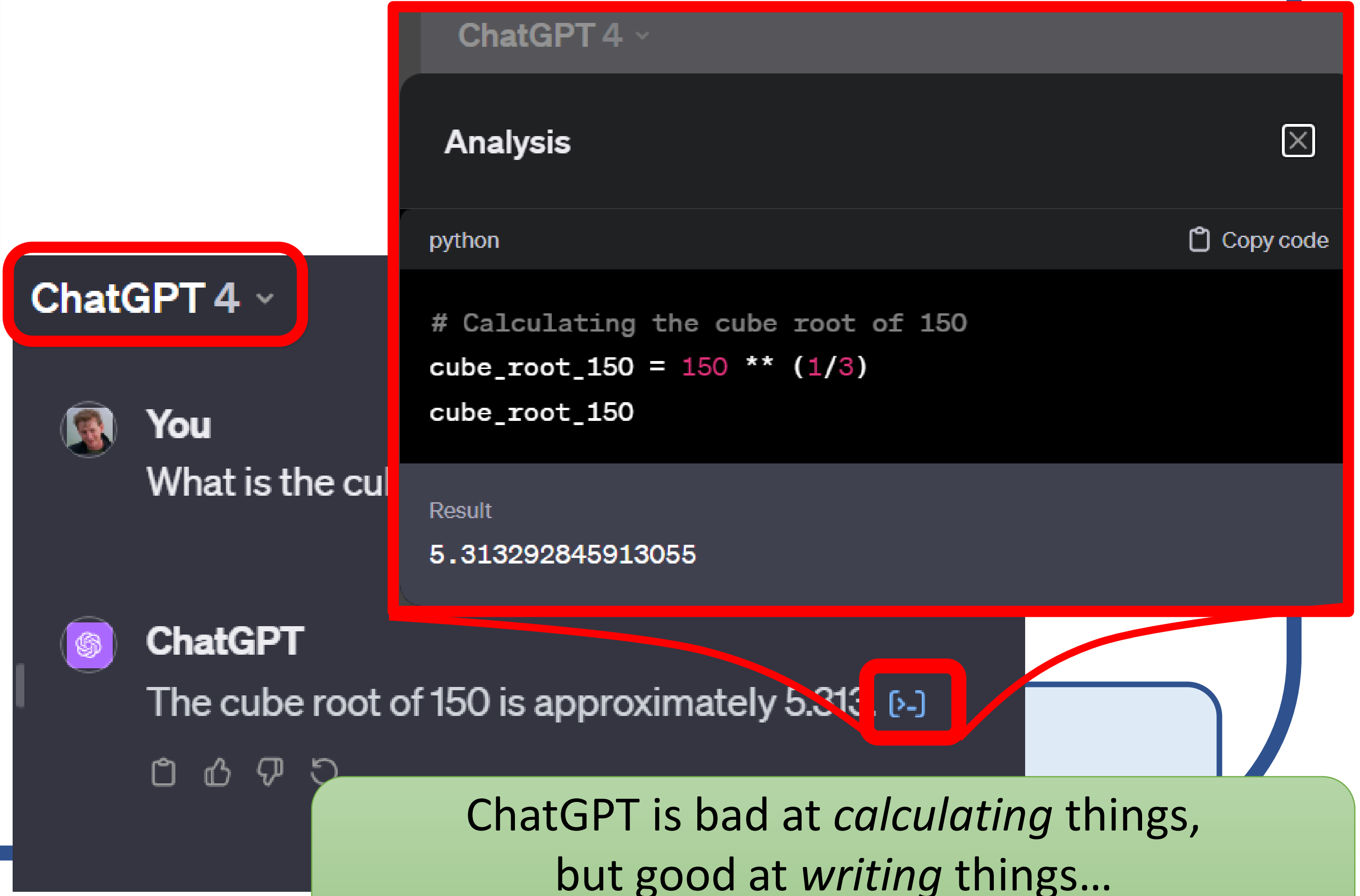
5.31329284591

You

What is the cube root of 150?

ChatGPT

The cube root of 150 is approximately 5.179.



You

What is the cu

```python
# Calculating the cube root of 150
cube_root_150 = 150 ** (1/3)
cube_root_150
```

Result

```
5.313292845913055
```

ChatGPT

The cube root of 150 is approximately 5.313

ChatGPT is bad at *calculating* things,
but good at *writing* things…
including **Python code!!!**

Leveraging ChatGPT for Data Analysis and Visualization

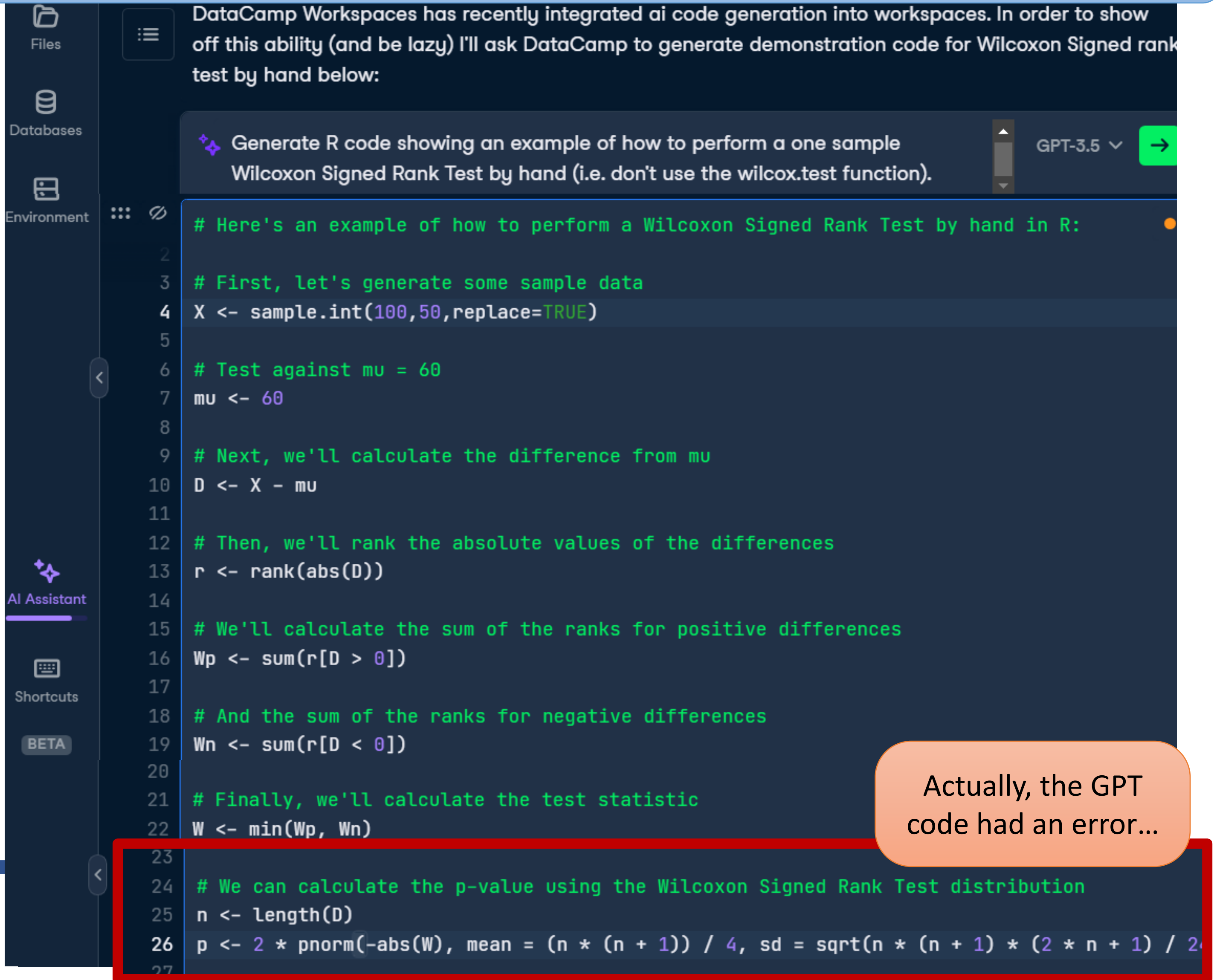
DataCamp Workspace "INE266 7.2 More two sample tests":
https://app.datacamp.com/workspace/w/d75466c5-f4e6-4ea3-9c6c-733561283ef7/

ChatGPT adds data analysis abilities
(interface with python)

Future: Mixing language and code
for data analysis, visualization,
etc.....

Data analysis / computational engines
add ChatGPT / LLM interfaces

University of the Virgin Islands
www.uvi.edu
SPECIALIZING IN FUTURES
HISTORICALLY AMERICAN. UNIQUELY CARIBBEAN. GLOBALLY INTERACTIVE.



DataCamp Workspaces has recently integrated ai code generation into workspaces. In order to show off this ability (and be lazy) I'll ask DataCamp to generate demonstration code for Wilcoxon Signed rank test by hand below:

Generate R code showing an example of how to perform a one sample Wilcoxon Signed Rank Test by hand (i.e. don't use the wilcox.test function).

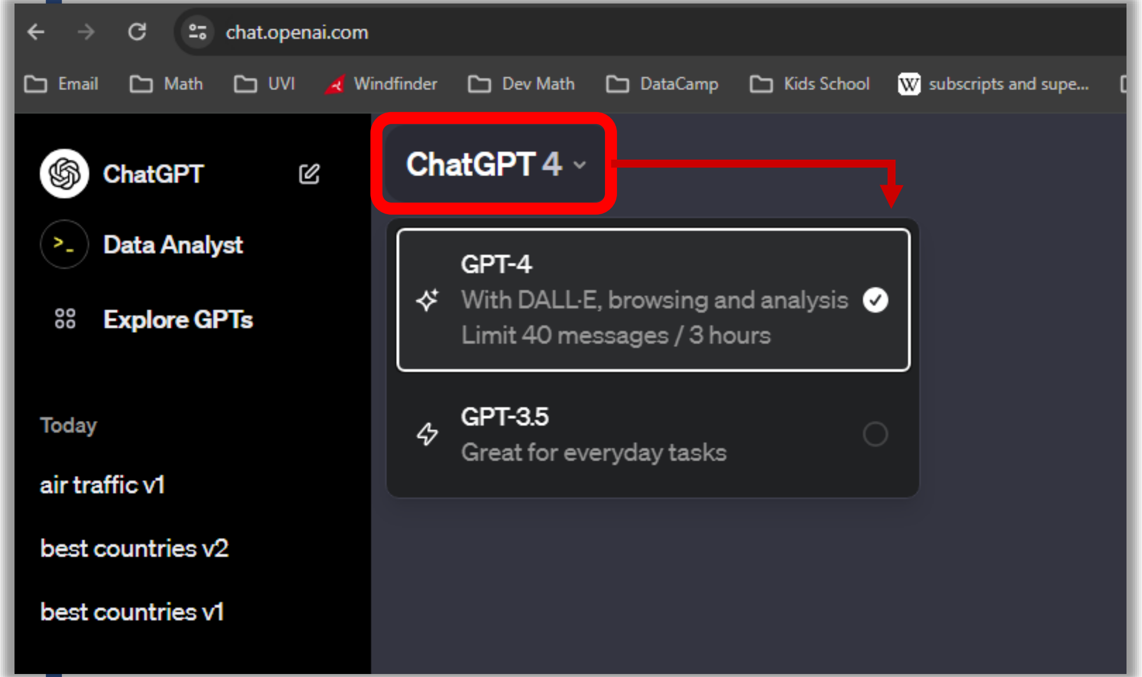
```r
# Here's an example of how to perform a Wilcoxon Signed Rank Test by hand in R:

# First, let's generate some sample data
X <- sample.int(100,50,replace=TRUE)

# Test against mu = 60
mu <- 60

# Next, we'll calculate the difference from mu
D <- X - mu

# Then, we'll rank the absolute values of the differences
r <- rank(abs(D))

# We'll calculate the sum of the ranks for positive differences
Wp <- sum(r[D > 0])

# And the sum of the ranks for negative differences
Wn <- sum(r[D < 0])

# Finally, we'll calculate the test statistic
W <- min(Wp, Wn)

# We can calculate the p-value using the Wilcoxon Signed Rank Test distribution
n <- length(D)
p <- 2 * pnorm(-abs(W), mean = (n * (n + 1)) / 4, sd = sqrt(n * (n + 1) * (2 * n + 1) / 2
```

Actually, the GPT code had an error...

Leveraging ChatGPT for Data Analysis and Visualization

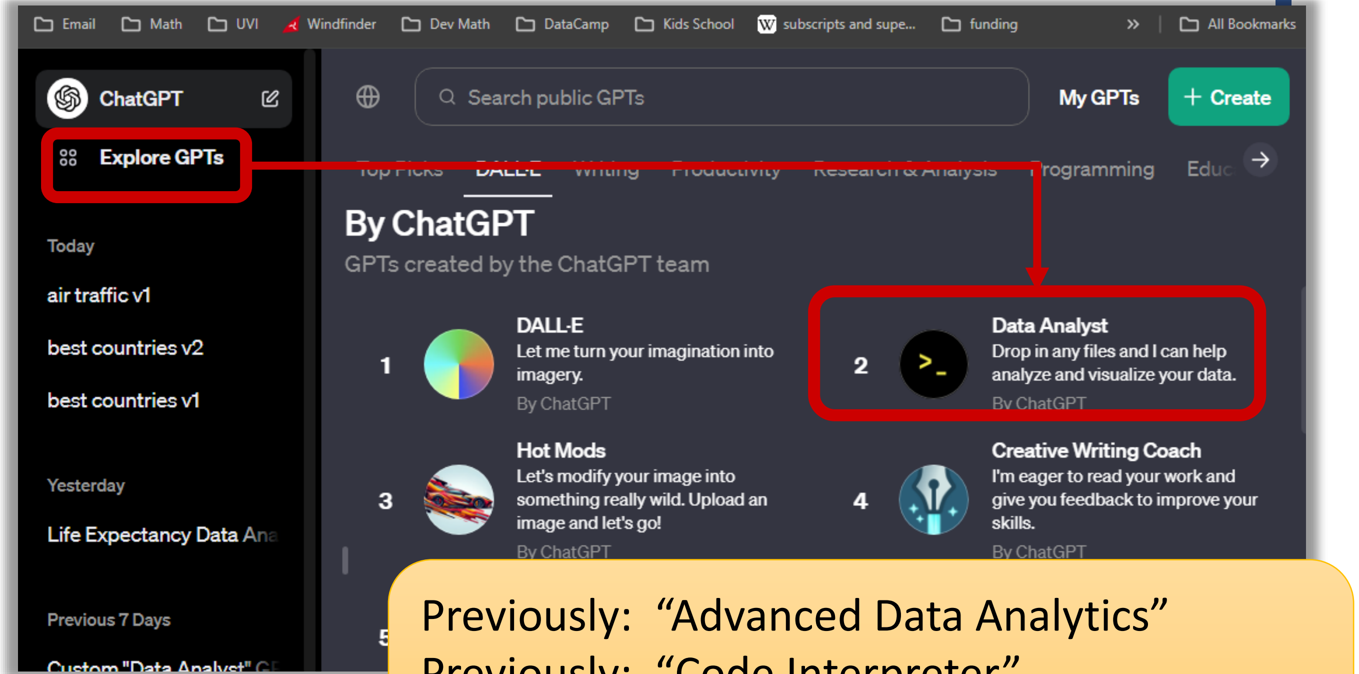
I will use ChatGPT 4

and the Data Analyst custom GPT



GPT-4: With DALL-E, browsing and analysis. Limit 40 messages / 3 hours

GPT-3.5: Great for everyday tasks



By ChatGPT

GPTs created by the ChatGPT team

1 DALL·E: Let me turn your imagination into imagery. By ChatGPT

2 Data Analyst: Drop in any files and I can help analyze and visualize your data. By ChatGPT

3 Hot Mods: Let's modify your image into something really wild. Upload an image and let's go! By ChatGPT

4 Creative Writing Coach: I'm eager to read your work and give you feedback to improve your skills. By ChatGPT

Previously: "Advanced Data Analytics"
Previously: "Code Interpreter"

University of the Virgin Islands
www.uvi.edu
SPECIALIZING IN FUTURES
HISTORICALLY AMERICAN. UNIQUELY CARIBBEAN. GLOBALLY INTERACTIVE.

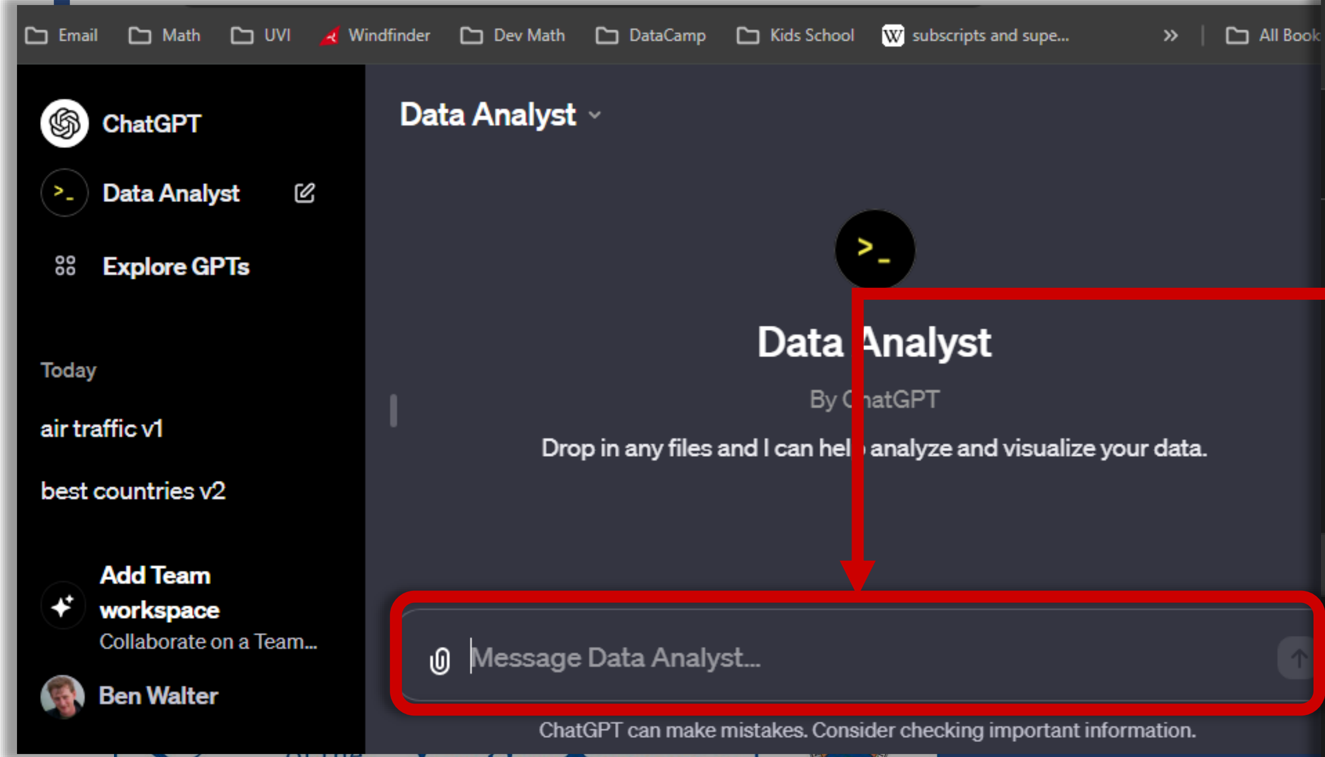
Dr. Benjamin Walter

ECS **January 31, 2024**

Leveraging ChatGPT for Data Analysis and Visualization

Import data:

- Drag and drop into browser window.



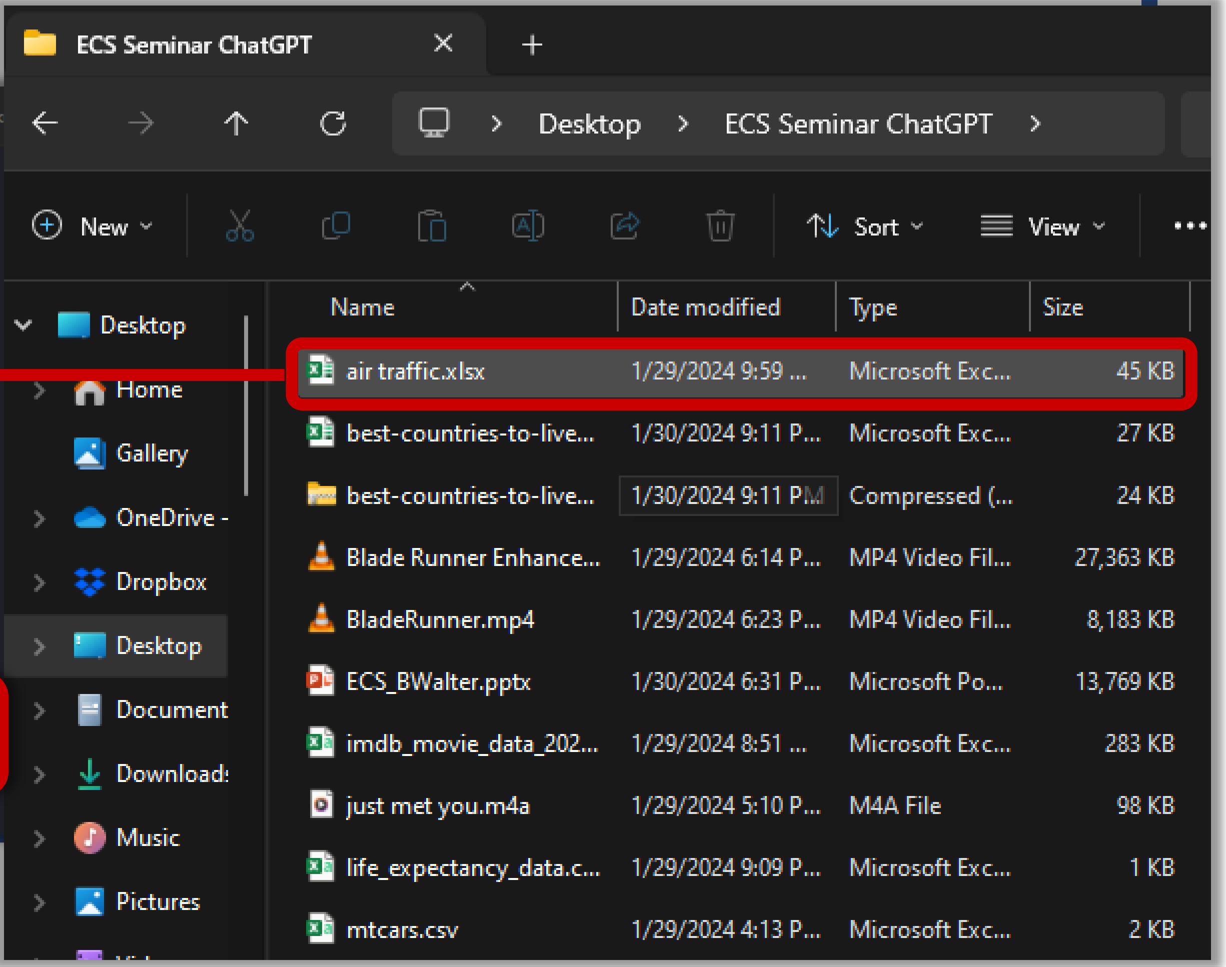
Data Analyst

By ChatGPT

Drop in any files and I can help analyze and visualize your data.

Message Data Analyst...

ChatGPT can make mistakes. Consider checking important information.

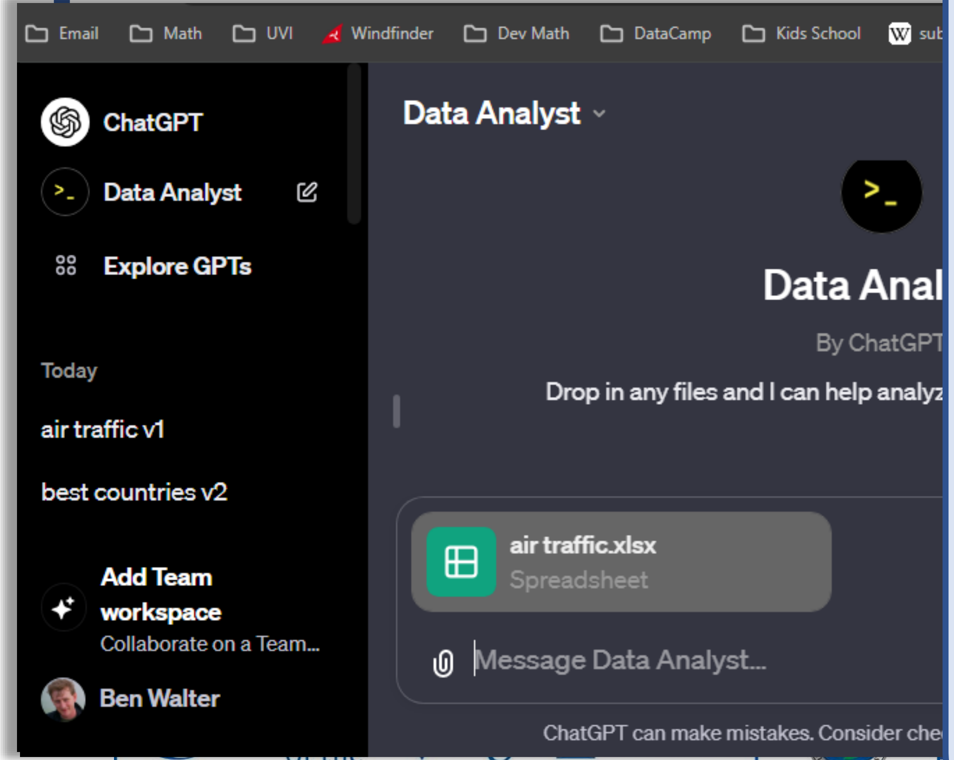


Name	Date modified	Type	Size
air traffic.xlsx	1/29/2024 9:59 ...	Microsoft Exc...	45 KB
best-countries-to-live...	1/30/2024 9:11 P...	Microsoft Exc...	27 KB
best-countries-to-live...	1/30/2024 9:11 PM	Compressed (...	24 KB
Blade Runner Enhance...	1/29/2024 6:14 P...	MP4 Video Fil...	27,363 KB
BladeRunner.mp4	1/29/2024 6:23 P...	MP4 Video Fil...	8,183 KB
ECS_BWalter.pptx	1/30/2024 6:31 P...	Microsoft Po...	13,769 KB
imdb_movie_data_202...	1/29/2024 8:51 ...	Microsoft Exc...	283 KB
just met you.m4a	1/29/2024 5:10 P...	M4A File	98 KB
life_expectancy_data.c...	1/29/2024 9:09 P...	Microsoft Exc...	1 KB
mtcars.csv	1/29/2024 4:13 P...	Microsoft Exc...	2 KB

Leveraging ChatGPT for Data Analysis and Visualization

Import data:

- Drag and drop into browser



Drop in any files and I can help analyz

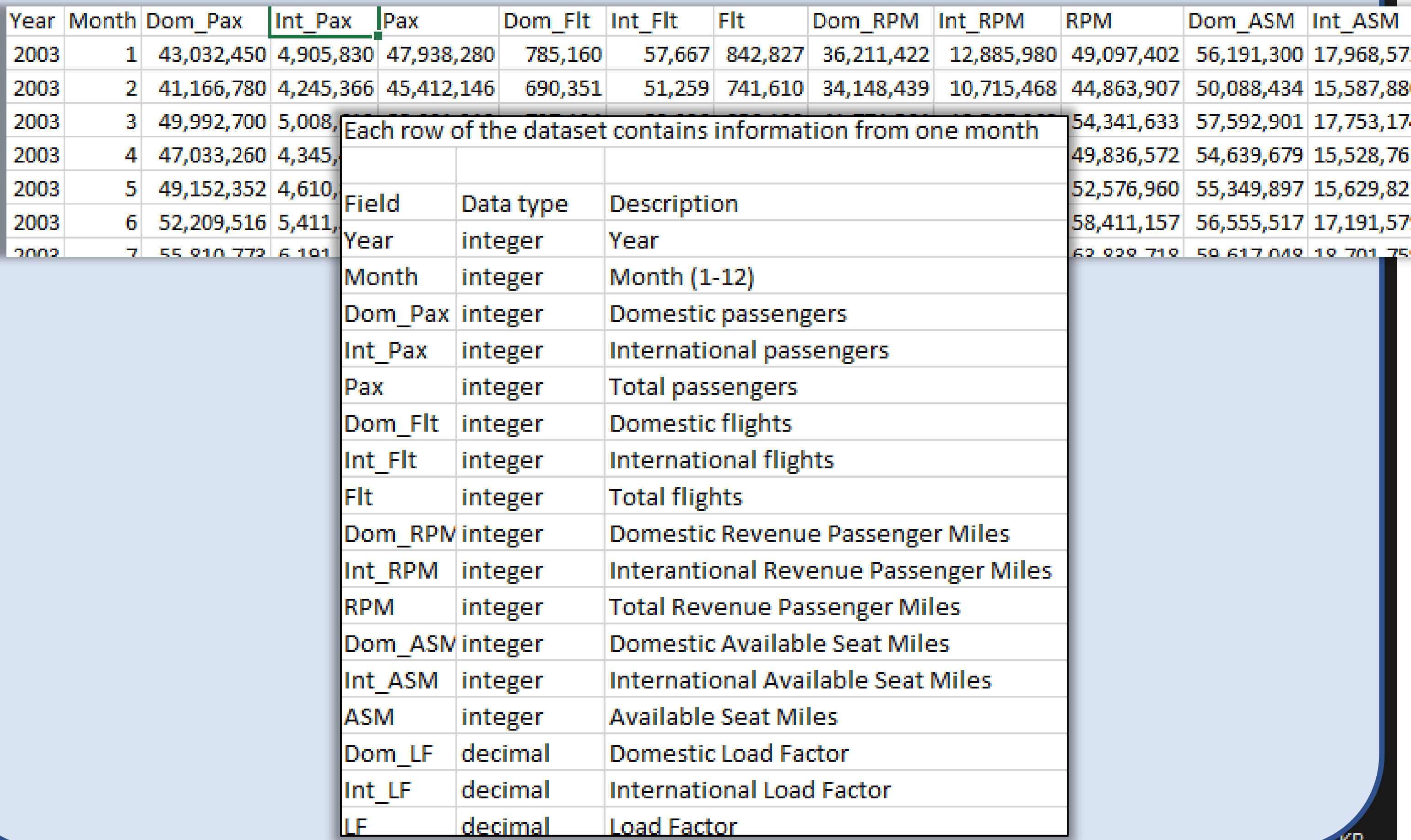
ChatGPT can make mistakes. Consider che

US Airline Traffic Data

Source: Kaggle

https://www.kaggle.com/datasets/yyxian/u-s-airline-traffic-data

Size: 17 columns, 250 rows

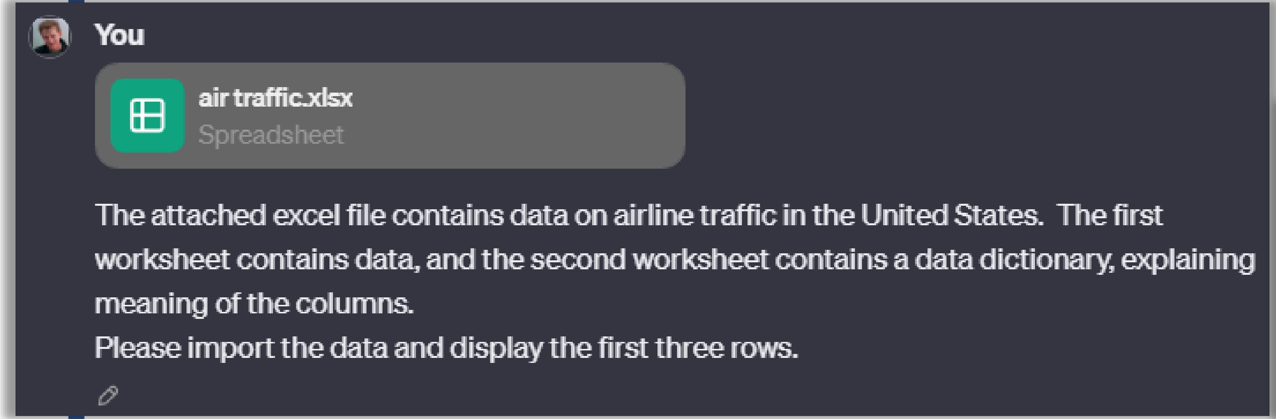


Year	Month	Dom_Pax	Int_Pax	Pax	Dom_Flt	Int_Flt	Flt	Dom_RPM	Int_RPM	RPM	Dom_ASM	Int_ASM
2003	1	43,032,450	4,905,830	47,938,280	785,160	57,667	842,827	36,211,422	12,885,980	49,097,402	56,191,300	17,968,57
2003	2	41,166,780	4,245,366	45,412,146	690,351	51,259	741,610	34,148,439	10,715,468	44,863,907	50,088,434	15,587,88
2003	3	49,992,700	5,008,							54,341,633	57,592,901	17,753,17
2003	4	47,033,260	4,345,							49,836,572	54,639,679	15,528,76
2003	5	49,152,352	4,610,							52,576,960	55,349,897	15,629,82
2003	6	52,209,516	5,411,							58,411,157	56,555,517	17,191,57

Each row of the dataset contains information from one month

Field	Data type	Description
Year	integer	Year
Month	integer	Month (1-12)
Dom_Pax	integer	Domestic passengers
Int_Pax	integer	International passengers
Pax	integer	Total passengers
Dom_Flt	integer	Domestic flights
Int_Flt	integer	International flights
Flt	integer	Total flights
Dom_RPM	integer	Domestic Revenue Passenger Miles
Int_RPM	integer	Interantional Revenue Passenger Miles
RPM	integer	Total Revenue Passenger Miles
Dom_ASM	integer	Domestic Available Seat Miles
Int_ASM	integer	International Available Seat Miles
ASM	integer	Available Seat Miles
Dom_LF	decimal	Domestic Load Factor
Int_LF	decimal	International Load Factor
LF	decimal	Load Factor

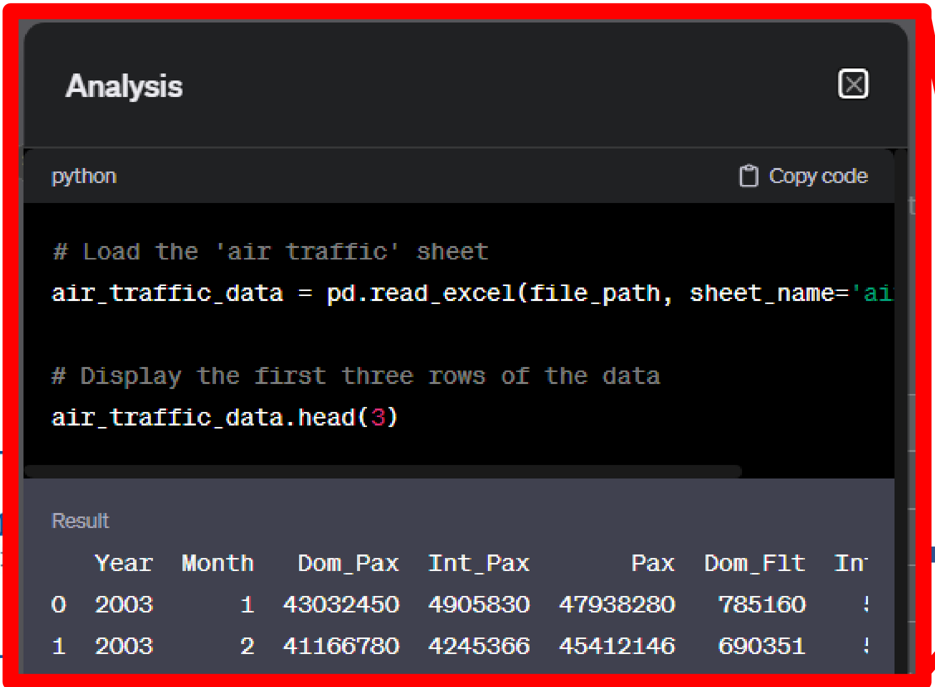
Leveraging ChatGPT for Data Analysis and Visualization



You

air traffic.xlsx
Spreadsheet

The attached excel file contains data on airline traffic in the United States. The first worksheet contains data, and the second worksheet contains a data dictionary, explaining meaning of the columns.
Please import the data and display the first three rows.

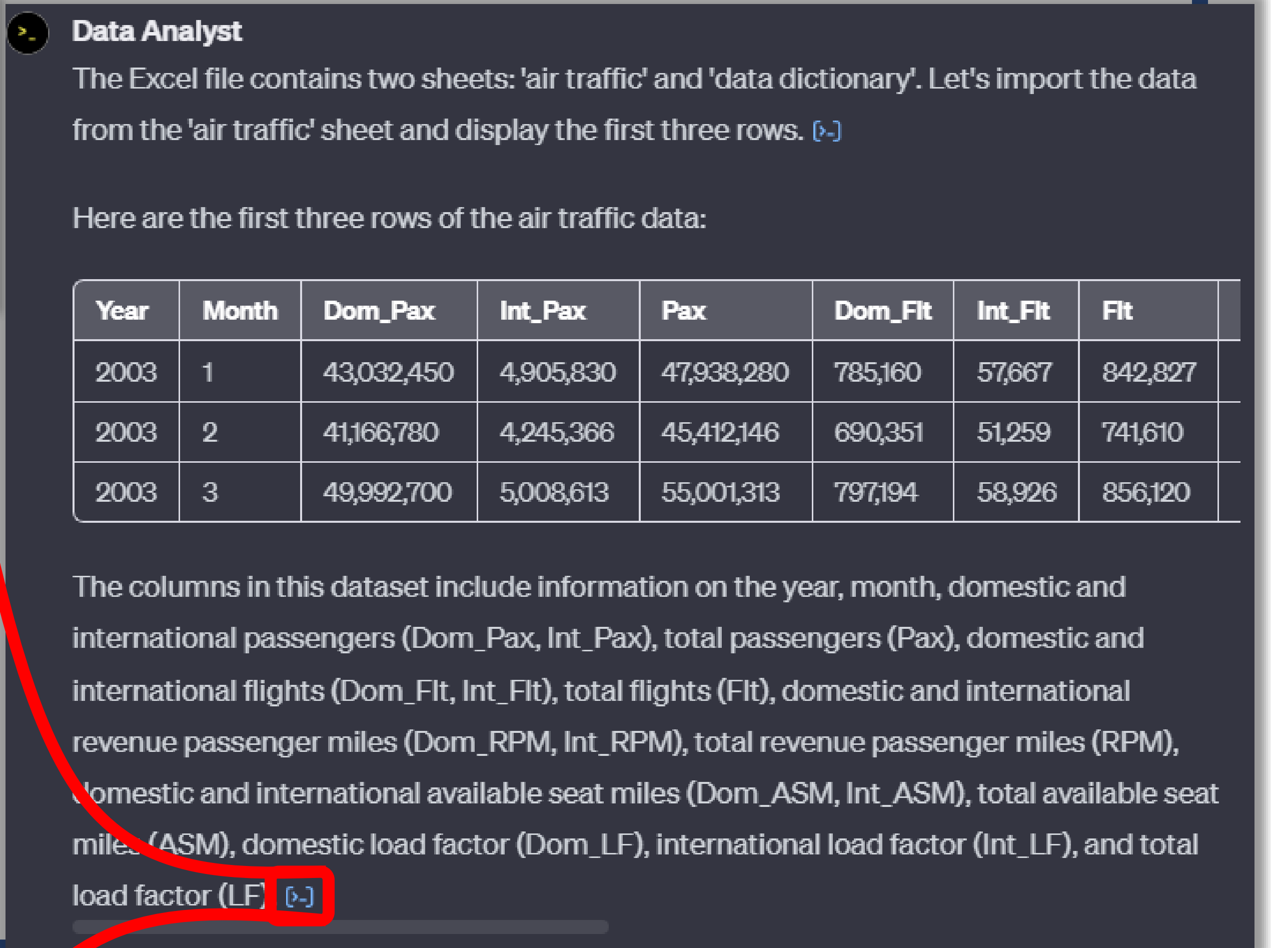


Analysis

```python
# Load the 'air traffic' sheet
air_traffic_data = pd.read_excel(file_path, sheet_name='ai

# Display the first three rows of the data
air_traffic_data.head(3)
```

Result

```
   Year  Month   Dom_Pax  Int_Pax       Pax  Dom_Flt  In
0  2003      1  43032450  4905830  47938280   785160   !
1  2003      2  41166780  4245366  45412146   690351   !
```



Data Analyst

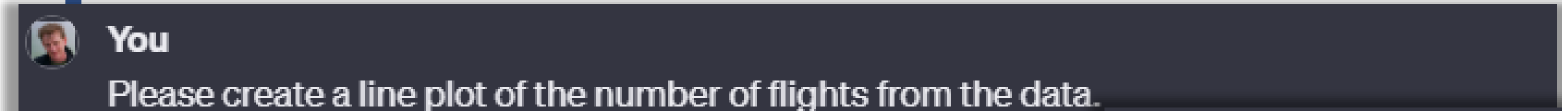
The Excel file contains two sheets: 'air traffic' and 'data dictionary'. Let's import the data from the 'air traffic' sheet and display the first three rows.

Here are the first three rows of the air traffic data:

Year	Month	Dom_Pax	Int_Pax	Pax	Dom_Flt	Int_Flt	Flt
2003	1	43,032,450	4,905,830	47,938,280	785,160	57,667	842,827
2003	2	41,166,780	4,245,366	45,412,146	690,351	51,259	741,610
2003	3	49,992,700	5,008,613	55,001,313	797,194	58,926	856,120

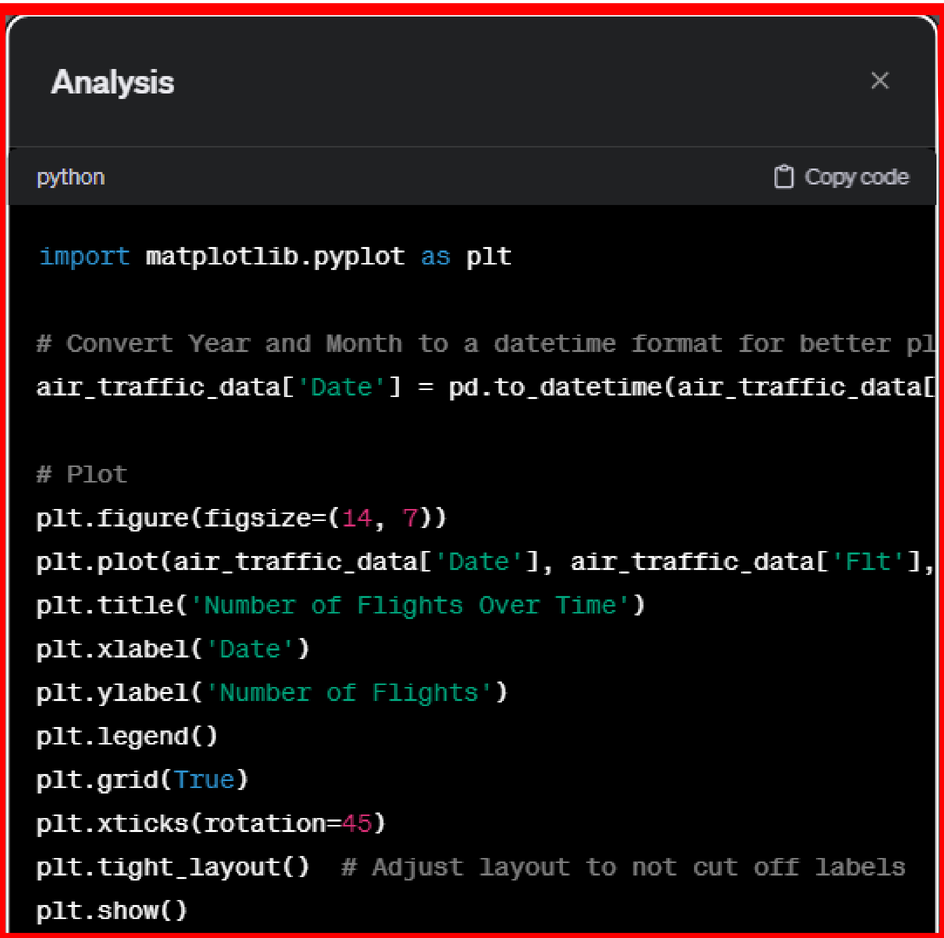
The columns in this dataset include information on the year, month, domestic and international passengers (Dom_Pax, Int_Pax), total passengers (Pax), domestic and international flights (Dom_Flt, Int_Flt), total flights (Flt), domestic and international revenue passenger miles (Dom_RPM, Int_RPM), total revenue passenger miles (RPM), domestic and international available seat miles (Dom_ASM, Int_ASM), total available seat miles (ASM), domestic load factor (Dom_LF), international load factor (Int_LF), and total load factor (LF).

Leveraging ChatGPT for Data Analysis and Visualization

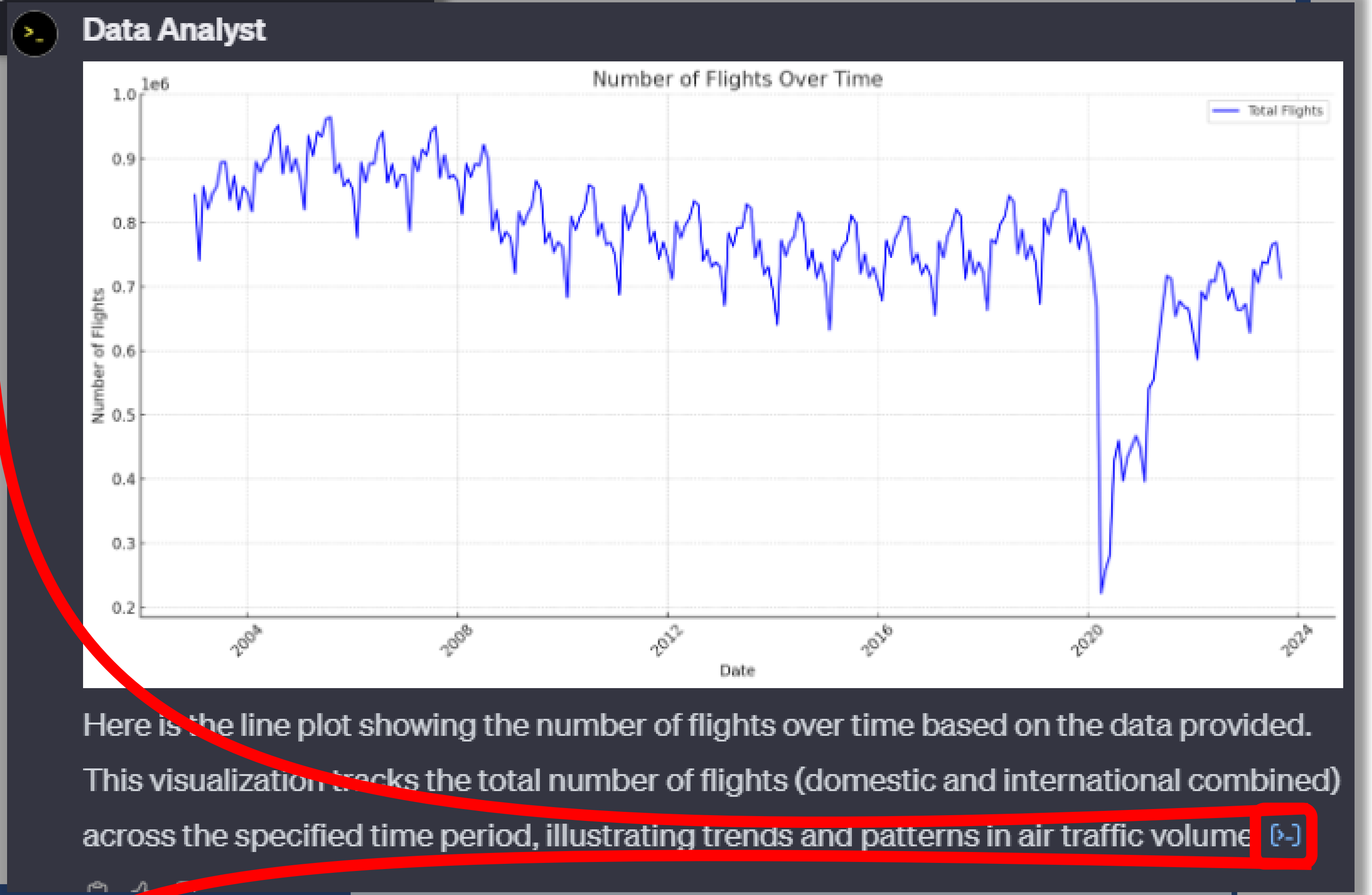


You

Please create a line plot of the number of flights from the data.

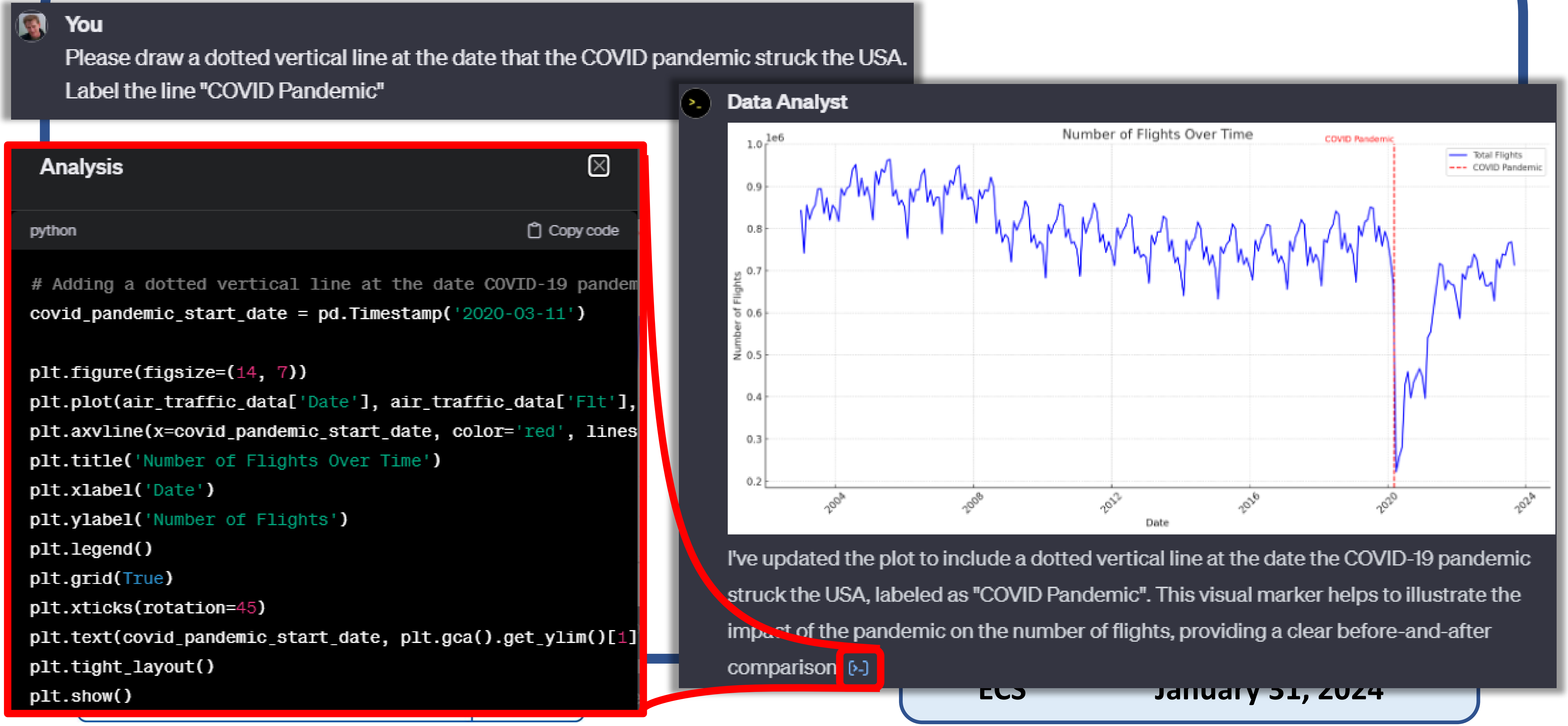


```python
import matplotlib.pyplot as plt

# Convert Year and Month to a datetime format for better pl
air_traffic_data['Date'] = pd.to_datetime(air_traffic_data[

# Plot
plt.figure(figsize=(14, 7))
plt.plot(air_traffic_data['Date'], air_traffic_data['Flt'],
plt.title('Number of Flights Over Time')
plt.xlabel('Date')
plt.ylabel('Number of Flights')
plt.legend()
plt.grid(True)
plt.xticks(rotation=45)
plt.tight_layout()  # Adjust layout to not cut off labels
plt.show()
```



Here is the line plot showing the number of flights over time based on the data provided. This visualization tracks the total number of flights (domestic and international combined) across the specified time period, illustrating trends and patterns in air traffic volume [>_]

Leveraging ChatGPT for Data Analysis and Visualization



You

Please draw a dotted vertical line at the date that the COVID pandemic struck the USA. Label the line "COVID Pandemic"

Analysis

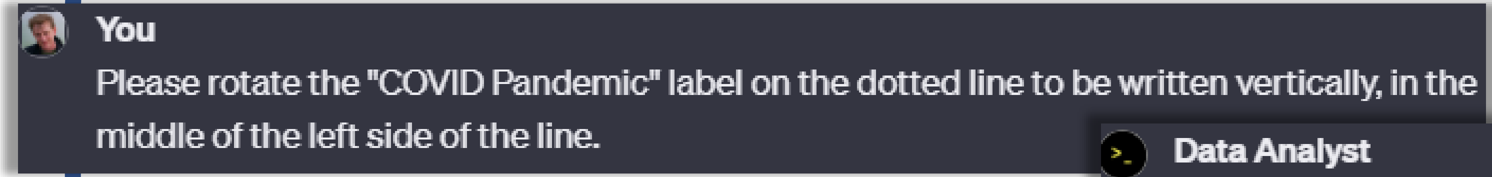
```python
# Adding a dotted vertical line at the date COVID-19 pandem
covid_pandemic_start_date = pd.Timestamp('2020-03-11')

plt.figure(figsize=(14, 7))
plt.plot(air_traffic_data['Date'], air_traffic_data['Flt'],
plt.axvline(x=covid_pandemic_start_date, color='red', lines
plt.title('Number of Flights Over Time')
plt.xlabel('Date')
plt.ylabel('Number of Flights')
plt.legend()
plt.grid(True)
plt.xticks(rotation=45)
plt.text(covid_pandemic_start_date, plt.gca().get_ylim()[1]
plt.tight_layout()
plt.show()
```

Data Analyst

I've updated the plot to include a dotted vertical line at the date the COVID-19 pandemic struck the USA, labeled as "COVID Pandemic". This visual marker helps to illustrate the impact of the pandemic on the number of flights, providing a clear before-and-after comparison

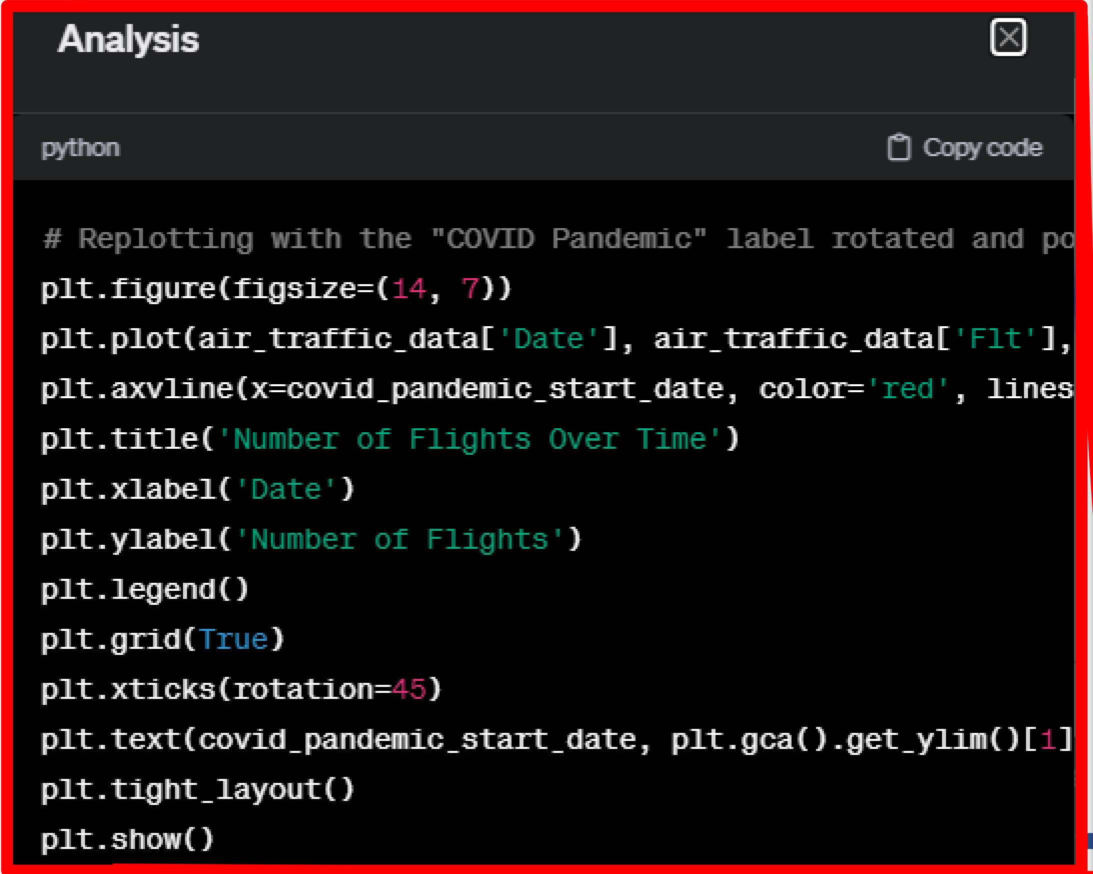
ECS January 31, 2024

Leveraging ChatGPT for Data Analysis and Visualization



You

Please rotate the "COVID Pandemic" label on the dotted line to be written vertically, in the middle of the left side of the line.

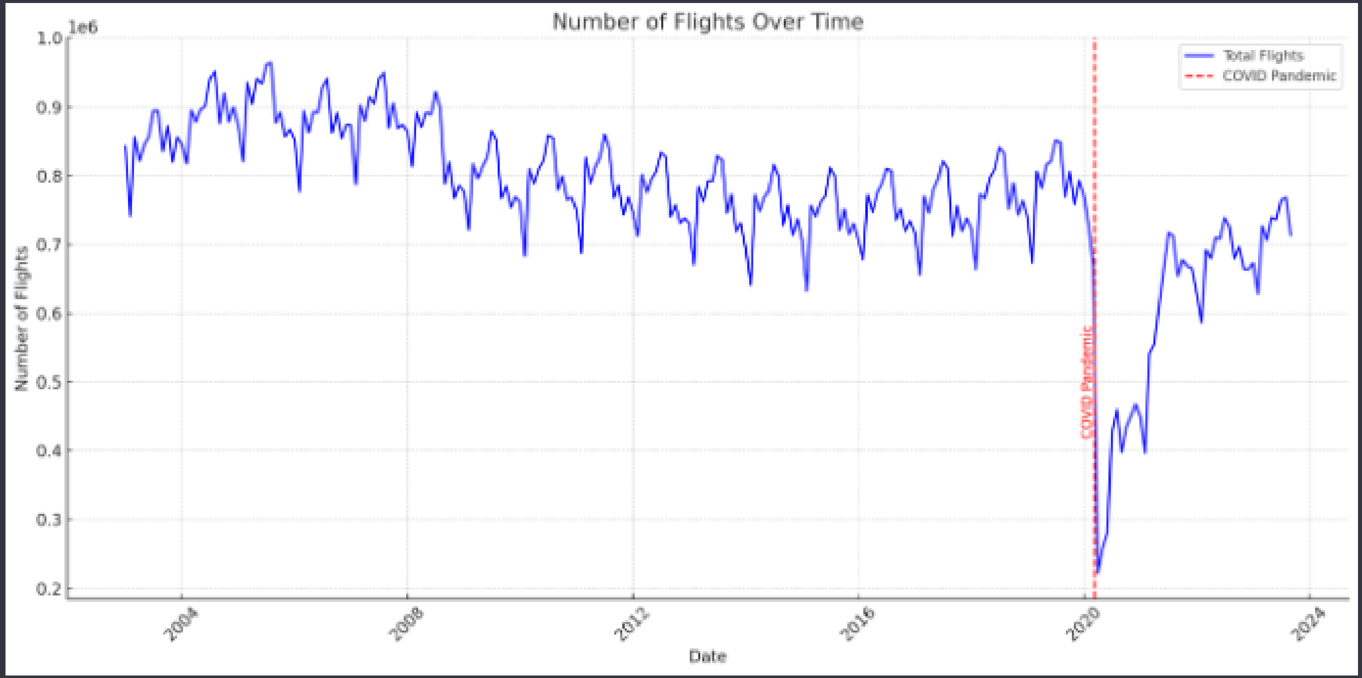


Analysis

python

```python
# Replotting with the "COVID Pandemic" label rotated and po
plt.figure(figsize=(14, 7))
plt.plot(air_traffic_data['Date'], air_traffic_data['Flt'],
plt.axvline(x=covid_pandemic_start_date, color='red', lines
plt.title('Number of Flights Over Time')
plt.xlabel('Date')
plt.ylabel('Number of Flights')
plt.legend()
plt.grid(True)
plt.xticks(rotation=45)
plt.text(covid_pandemic_start_date, plt.gca().get_ylim()[1]
plt.tight_layout()
plt.show()
```

Data Analyst



The plot has been updated with the "COVID Pandemic" label rotated and positioned vertically on the left side of the line, approximately in the middle of the plot's height. This adjustment enhances the visual clarity of the pandemic's impact on the number of flights.

Leveraging ChatGPT for Data Analysis and Visualization

There is more fun analysis that could be done on this dataset, but let’s pause for a bit….

Leveraging ChatGPT for Data Analysis and Visualization

Why the excitement and expectation about AI interfaces?

Bladerunner (1982)

Star Trek IV: The Voyage Home (1986)

University of the Virgin Islands

www.uvi.edu

SPECIALIZING IN FUTURES

HISTORICALLY AMERICAN.
UNIQUELY CARIBBEAN.
GLOBALLY INTERACTIVE.

Dr. Benjamin Walter

ECS **January 31, 2024**

Leveraging ChatGPT for Data Analysis and Visualization

Why the excitement and expectation about AI interfaces?

Most common interaction with computer is via cellphone

- No keyboard or mouse!
- Typing on a cellphone / tablet is not great.
- Coding on a cellphone / table is TERRIBLE!
    - Checking / correcting code isn’t bad….

Leveraging ChatGPT

Example.

- UN World Population Prospects 20
- pdf file: undesa_pd_2022_WPP_s
- 10 Mb; 52 pages (including introdu
- page 19 has data table: Table II.2 Li

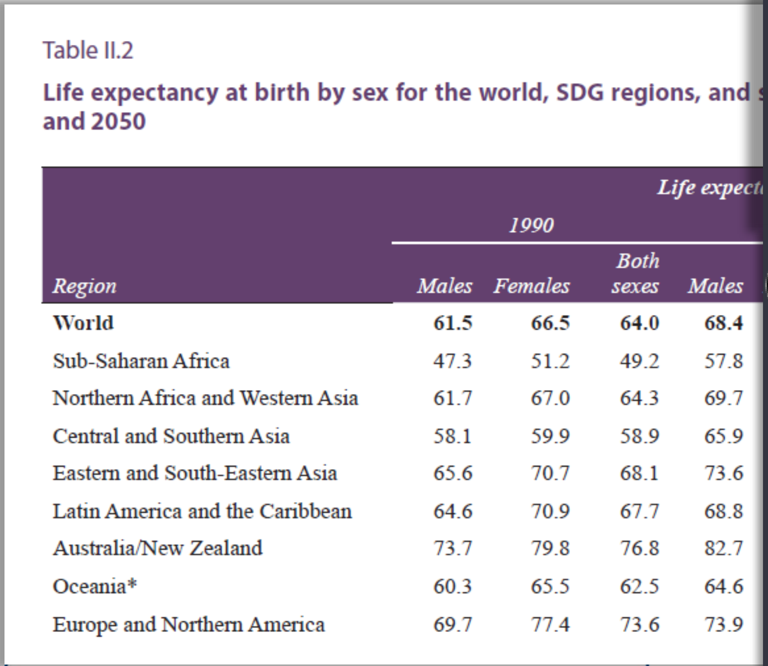
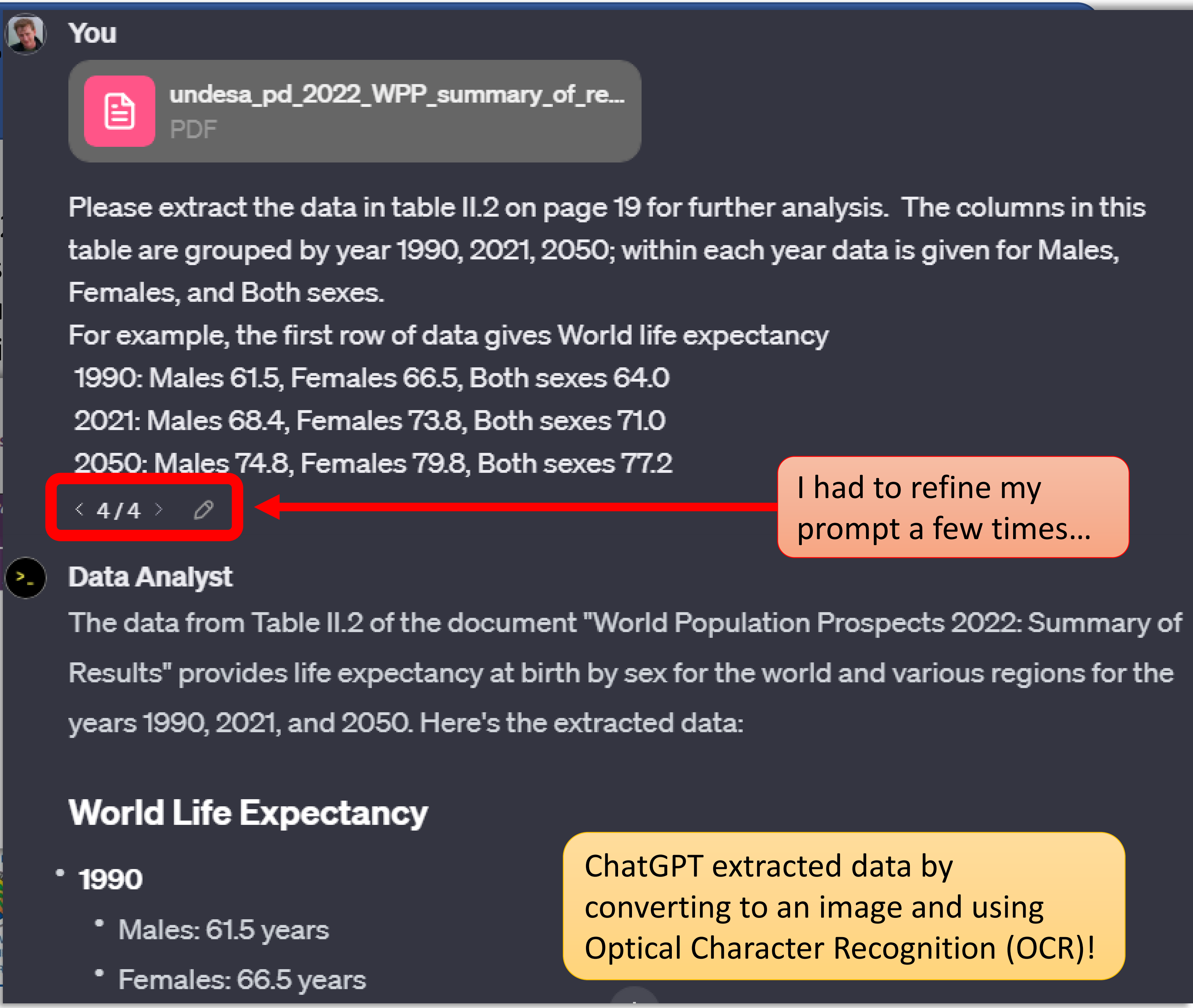


Table II.2

Life expectancy at birth by sex for the world, SDG regions, and s
and 2050

				Life expect
	1990			
Region	Males	Females	Both sexes	Males
World	**61.5**	**66.5**	**64.0**	**68.4**
Sub-Saharan Africa	47.3	51.2	49.2	57.8
Northern Africa and Western Asia	61.7	67.0	64.3	69.7
Central and Southern Asia	58.1	59.9	58.9	65.9
Eastern and South-Eastern Asia	65.6	70.7	68.1	73.6
Latin America and the Caribbean	64.6	70.9	67.7	68.8
Australia/New Zealand	73.7	79.8	76.8	82.7
Oceania*	60.3	65.5	62.5	64.6
Europe and Northern America	69.7	77.4	73.6	73.9



You

undesa_pd_2022_WPP_summary_of_re...
PDF

Please extract the data in table II.2 on page 19 for further analysis. The columns in this table are grouped by year 1990, 2021, 2050; within each year data is given for Males, Females, and Both sexes.
For example, the first row of data gives World life expectancy
1990: Males 61.5, Females 66.5, Both sexes 64.0
2021: Males 68.4, Females 73.8, Both sexes 71.0
2050: Males 74.8, Females 79.8, Both sexes 77.2

Data Analyst

The data from Table II.2 of the document "World Population Prospects 2022: Summary of Results" provides life expectancy at birth by sex for the world and various regions for the years 1990, 2021, and 2050. Here's the extracted data:

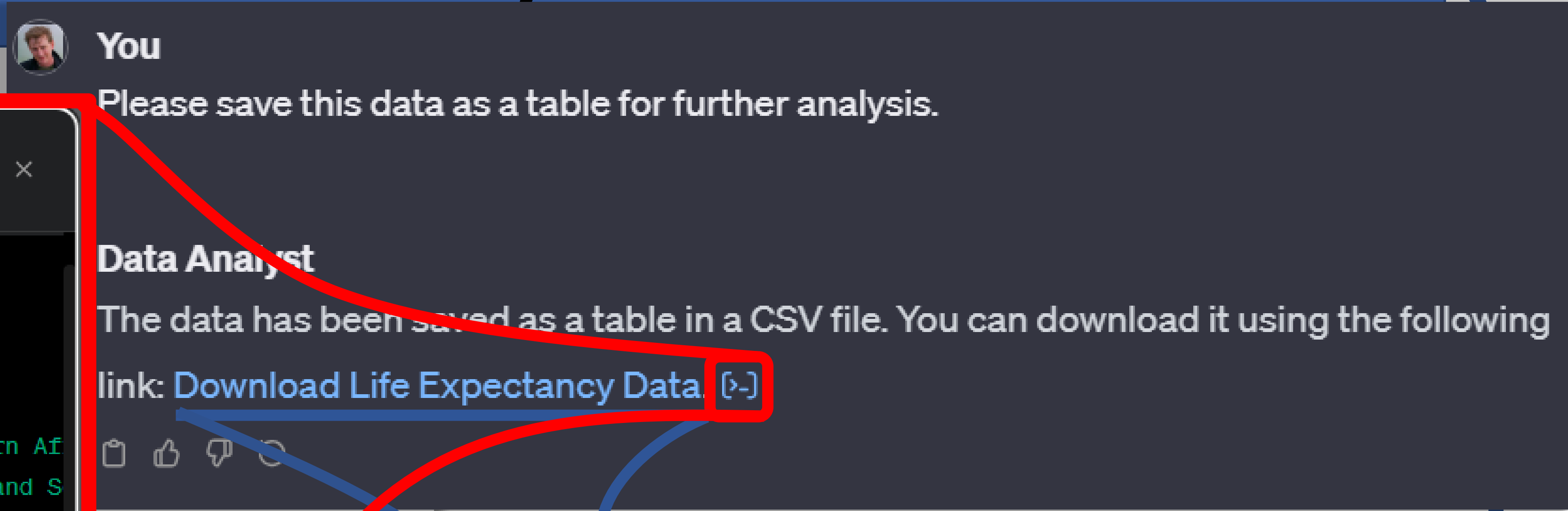
World Life Expectancy

- 1990
  - Males: 61.5 years
  - Females: 66.5 years

I had to refine my prompt a few times…

ChatGPT extracted data by converting to an image and using Optical Character Recognition (OCR)!

Leveraging ChatGPT for Data Analysis and Visualization

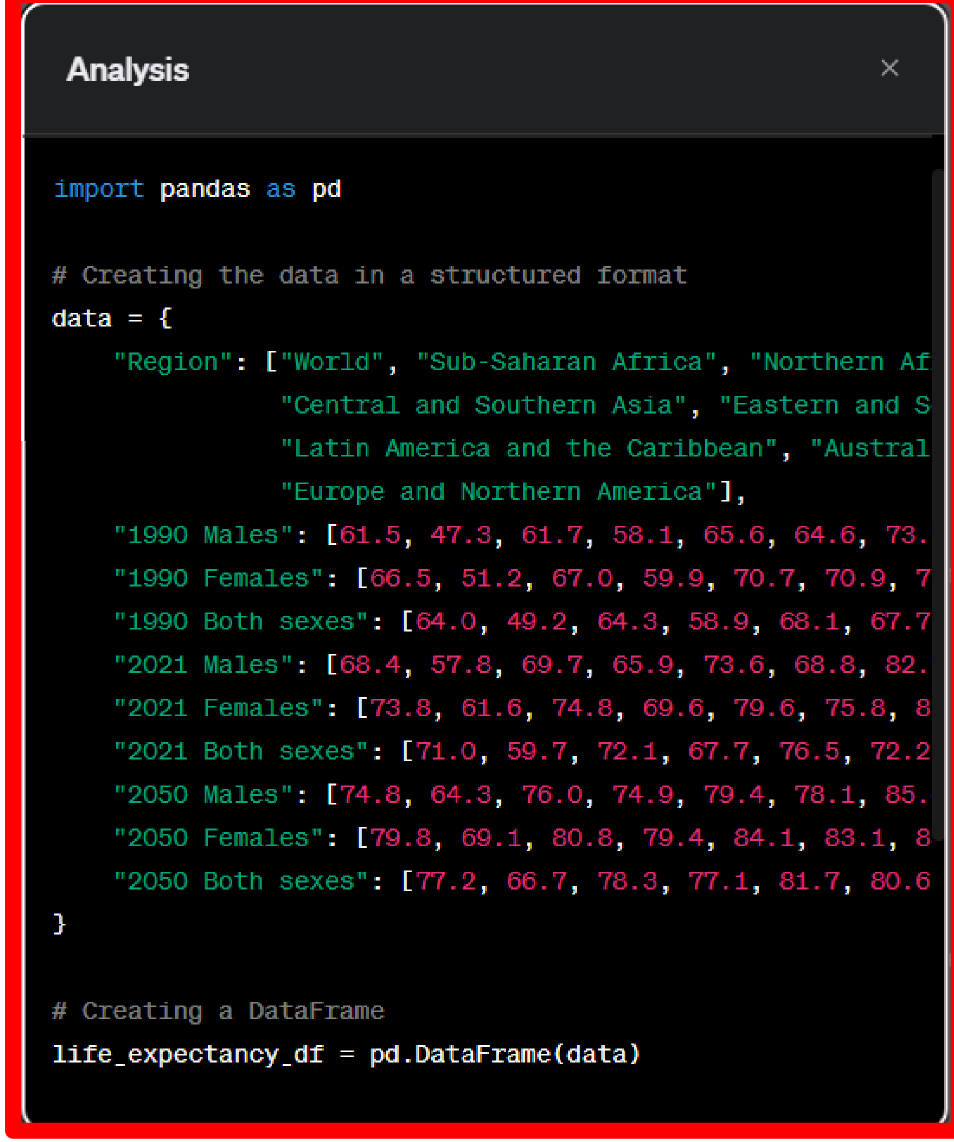


You

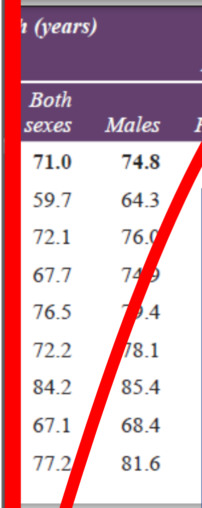
Please save this data as a table for further analysis.

Data Analyst

The data has been saved as a table in a CSV file. You can download it using the following link: Download Life Expectancy Data [>-]

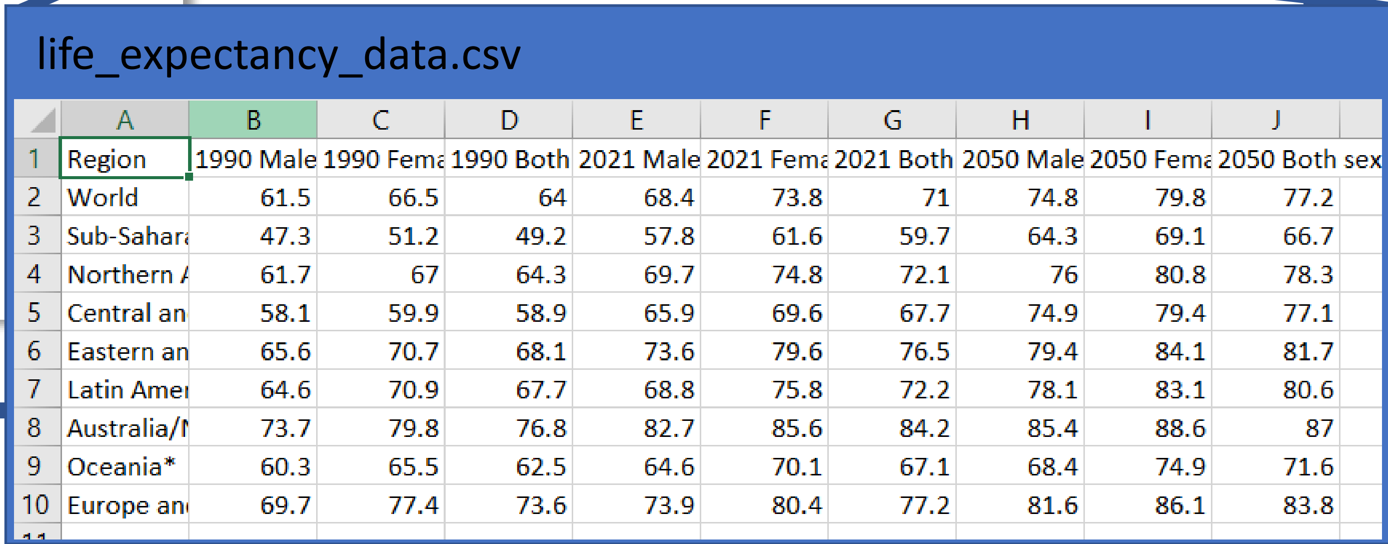


```
import pandas as pd

# Creating the data in a structured format
data = {
    "Region": ["World", "Sub-Saharan Africa", "Northern Af
               "Central and Southern Asia", "Eastern and S
               "Latin America and the Caribbean", "Austral
               "Europe and Northern America"],
    "1990 Males": [61.5, 47.3, 61.7, 58.1, 65.6, 64.6, 73.
    "1990 Females": [66.5, 51.2, 67.0, 59.9, 70.7, 70.9, 7
    "1990 Both sexes": [64.0, 49.2, 64.3, 58.9, 68.1, 67.7
    "2021 Males": [68.4, 57.8, 69.7, 65.9, 73.6, 68.8, 82.
    "2021 Females": [73.8, 61.6, 74.8, 69.6, 79.6, 75.8, 8
    "2021 Both sexes": [71.0, 59.7, 72.1, 67.7, 76.5, 72.2
    "2050 Males": [74.8, 64.3, 76.0, 74.9, 79.4, 78.1, 85.
    "2050 Females": [79.8, 69.1, 80.8, 79.4, 84.1, 83.1, 8
    "2050 Both sexes": [77.2, 66.7, 78.3, 77.1, 81.7, 80.6
}

# Creating a DataFrame
life_expectancy_df = pd.DataFrame(data)
```



h (years)

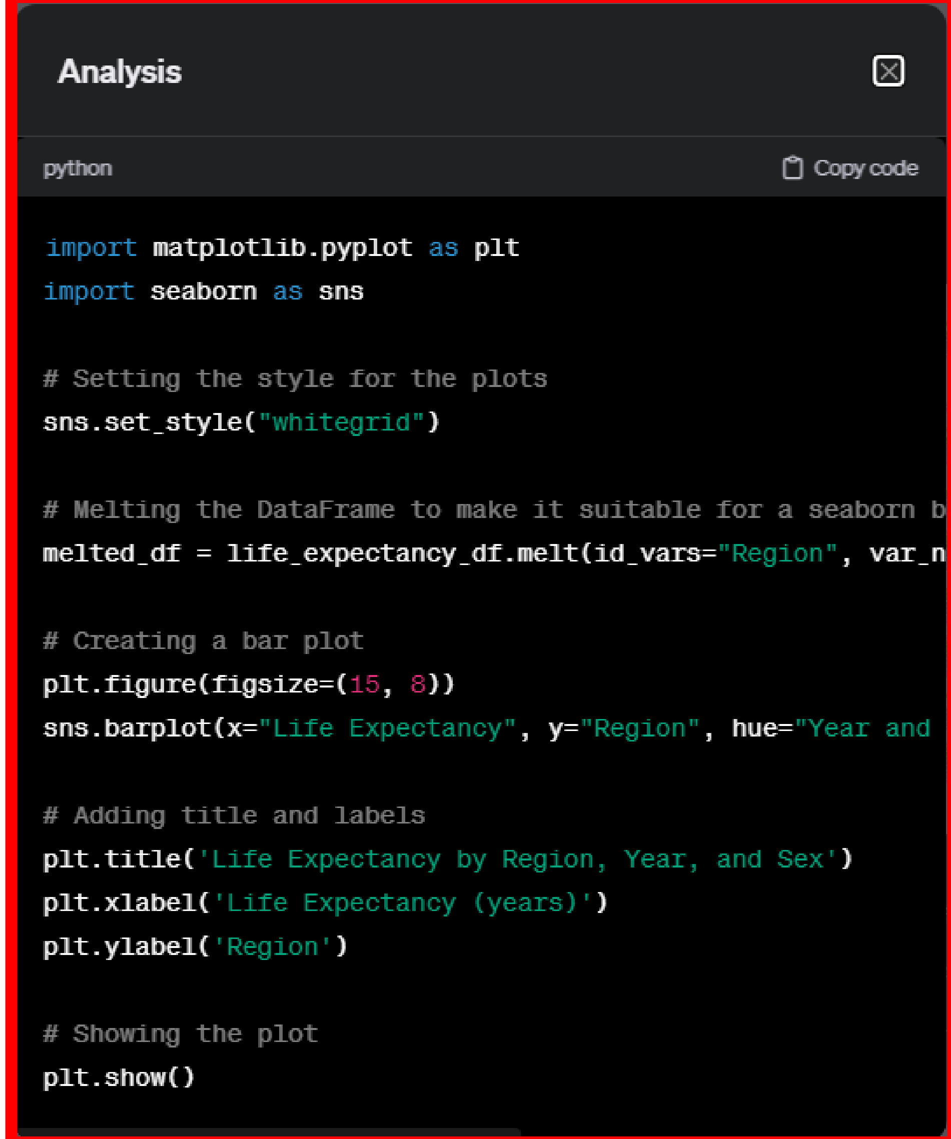
	2050		
Both sexes	Males	Females	Both sexes
71.0	74.8	79.8	77.2
59.7	64.3		
72.1	76.0		
67.7	74.9		
76.5	9.4		
72.2	78.1		
84.2	85.4		
67.1	68.4		
77.2	81.6		



life_expectancy_data.csv

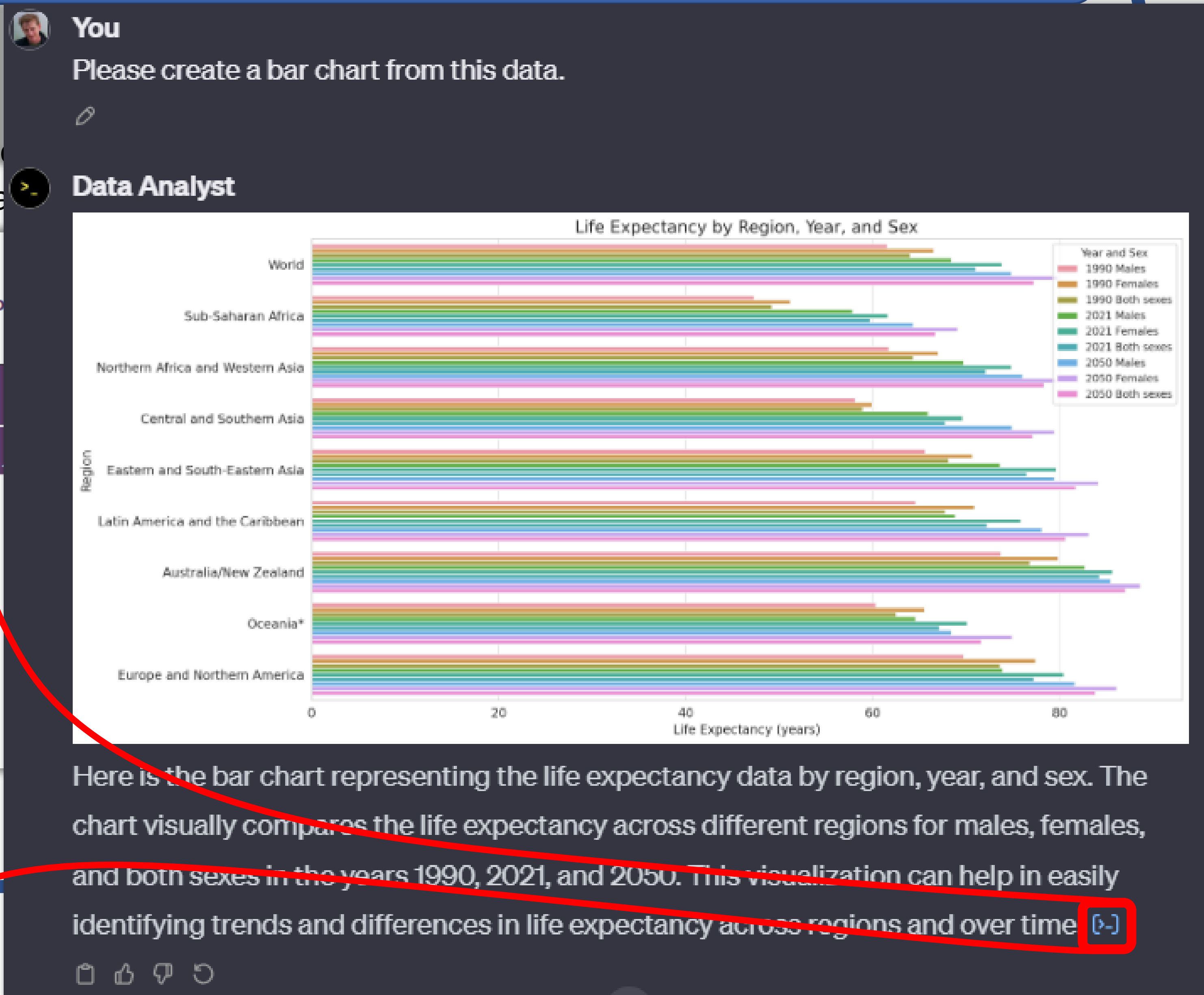
	A	B	C	D	E	F	G	H	I	J
1	Region	1990 Male	1990 Fema	1990 Both	2021 Male	2021 Fema	2021 Both	2050 Male	2050 Fema	2050 Both sex
2	World	61.5	66.5	64	68.4	73.8	71	74.8	79.8	77.2
3	Sub-Sahara	47.3	51.2	49.2	57.8	61.6	59.7	64.3	69.1	66.7
4	Northern A	61.7	67	64.3	69.7	74.8	72.1	76	80.8	78.3
5	Central an	58.1	59.9	58.9	65.9	69.6	67.7	74.9	79.4	77.1
6	Eastern an	65.6	70.7	68.1	73.6	79.6	76.5	79.4	84.1	81.7
7	Latin Amer	64.6	70.9	67.7	68.8	75.8	72.2	78.1	83.1	80.6
8	Australia/N	73.7	79.8	76.8	82.7	85.6	84.2	85.4	88.6	87
9	Oceania*	60.3	65.5	62.5	64.6	70.1	67.1	68.4	74.9	71.6
10	Europe an	69.7	77.4	73.6	73.9	80.4	77.2	81.6	86.1	83.8

Leveraging ChatGPT for Data Analysis and Visualization



```python
import matplotlib.pyplot as plt
import seaborn as sns

# Setting the style for the plots
sns.set_style("whitegrid")

# Melting the DataFrame to make it suitable for a seaborn b
melted_df = life_expectancy_df.melt(id_vars="Region", var_n

# Creating a bar plot
plt.figure(figsize=(15, 8))
sns.barplot(x="Life Expectancy", y="Region", hue="Year and

# Adding title and labels
plt.title('Life Expectancy by Region, Year, and Sex')
plt.xlabel('Life Expectancy (years)')
plt.ylabel('Region')

# Showing the plot
plt.show()
```

nary_
xpecta
d groups o
irth (years)
Both sexes
71.0
59.7
72.1
67.7
76.5
72.2
84.2
67.1
77.2
GLOBALLY INTERACTIVE.



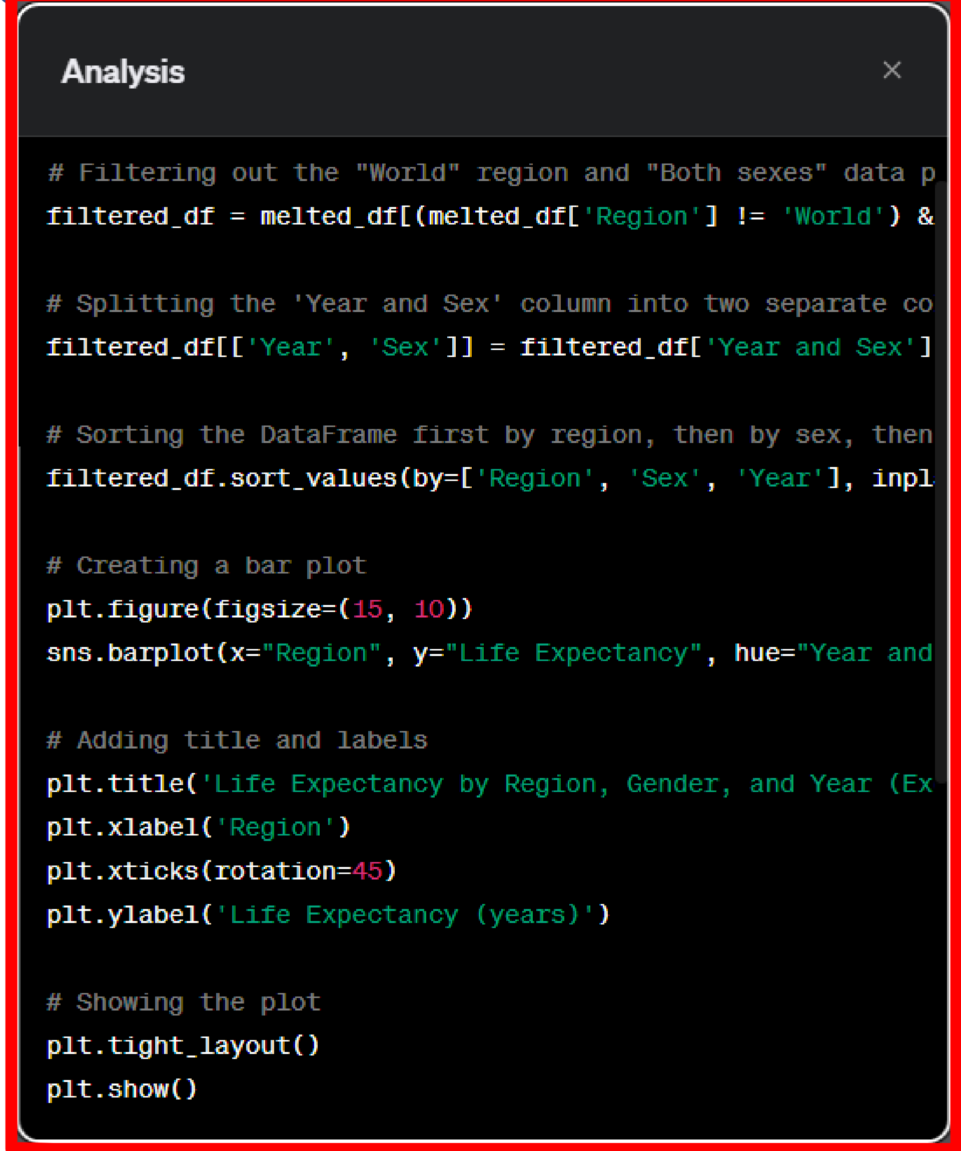
You

Please create a bar chart from this data.

Data Analyst

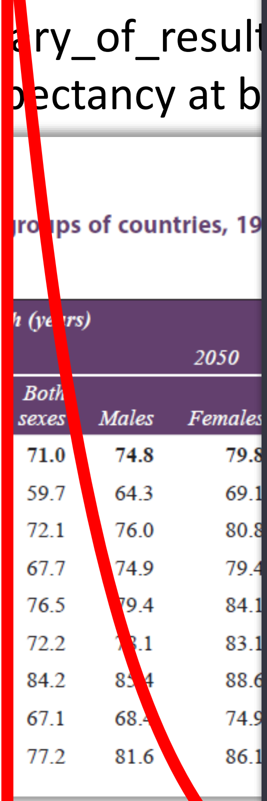
Here is the bar chart representing the life expectancy data by region, year, and sex. The chart visually compares the life expectancy across different regions for males, females, and both sexes in the years 1990, 2021, and 2050. This visualization can help in easily identifying trends and differences in life expectancy across regions and over time [>_]

Leveraging ChatGPT for Data



Analysis

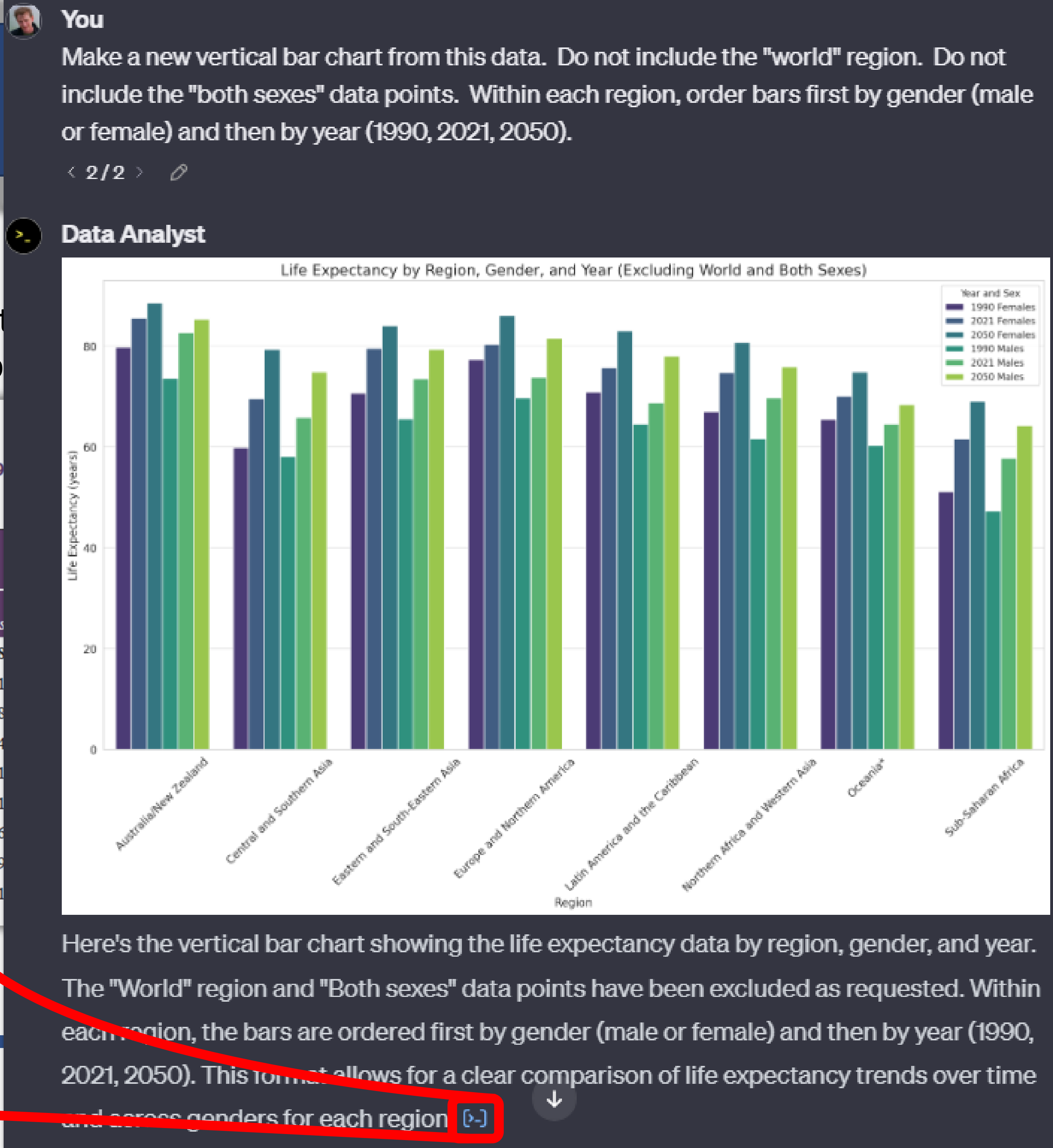
```
# Filtering out the "World" region and "Both sexes" data p
filtered_df = melted_df[(melted_df['Region'] != 'World') &

# Splitting the 'Year and Sex' column into two separate co
filtered_df[['Year', 'Sex']] = filtered_df['Year and Sex']

# Sorting the DataFrame first by region, then by sex, then
filtered_df.sort_values(by=['Region', 'Sex', 'Year'], inpl

# Creating a bar plot
plt.figure(figsize=(15, 10))
sns.barplot(x="Region", y="Life Expectancy", hue="Year and

# Adding title and labels
plt.title('Life Expectancy by Region, Gender, and Year (Ex
plt.xlabel('Region')
plt.xticks(rotation=45)
plt.ylabel('Life Expectancy (years)')

# Showing the plot
plt.tight_layout()
plt.show()
```



ary_of_result
ectancy at b

roups of countries, 19

h (years)		
2050		
Both sexes	Males	Females
71.0	74.8	79.8
59.7	64.3	69.1
72.1	76.0	80.8
67.7	74.9	79.4
76.5	79.4	84.1
72.2	[illegible]	83.1
84.2	85.4	88.6
67.1	68.4	74.9
77.2	81.6	86.1



You

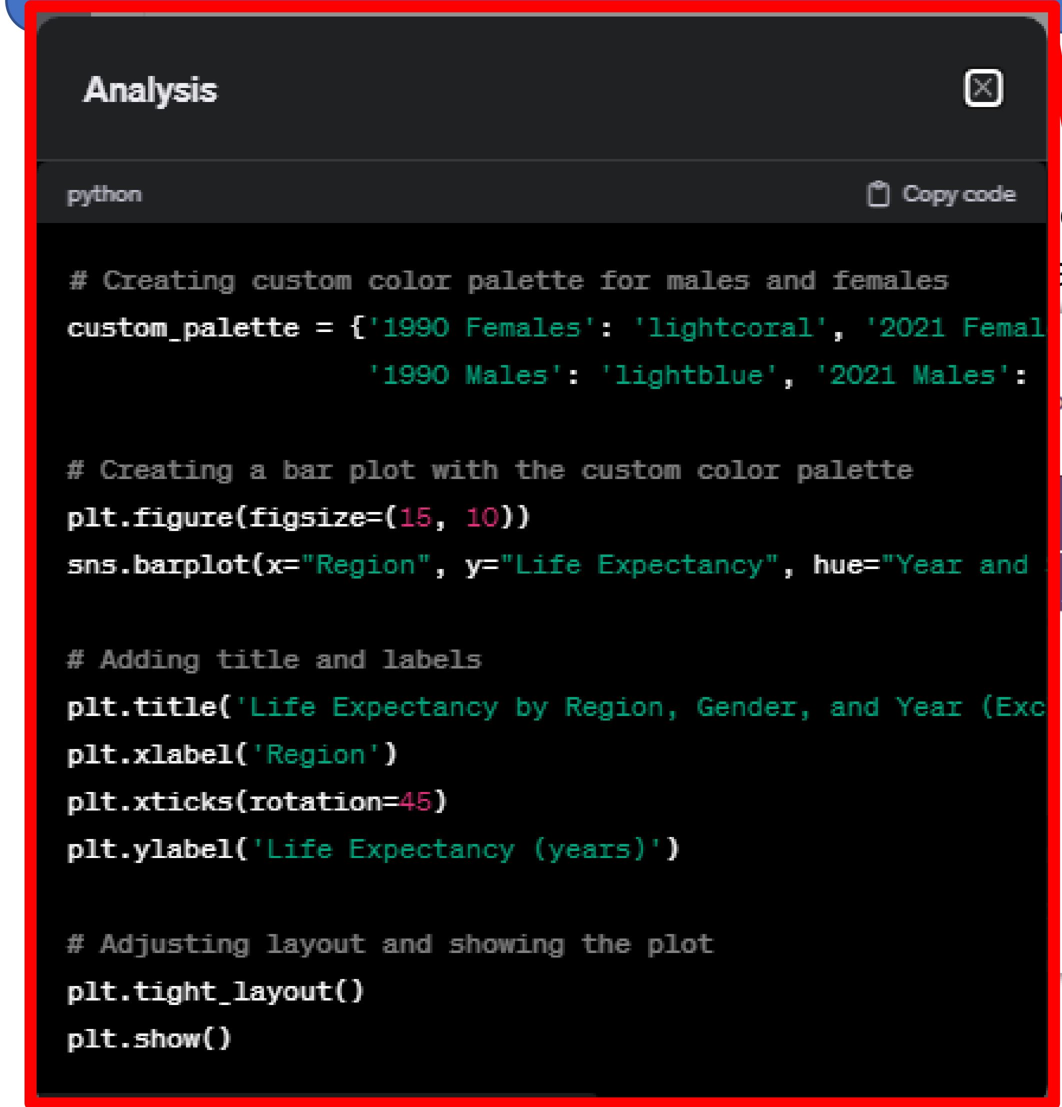
Make a new vertical bar chart from this data. Do not include the "world" region. Do not include the "both sexes" data points. Within each region, order bars first by gender (male or female) and then by year (1990, 2021, 2050).

Data Analyst

Life Expectancy by Region, Gender, and Year (Excluding World and Both Sexes)
Year and Sex
1990 Females
2021 Females
2050 Females
1990 Males
2021 Males
2050 Males
Life Expectancy (years)
0
20
40
60
80
Australia/New Zealand
Central and Southern Asia
Eastern and South-Eastern Asia
Europe and Northern America
Latin America and the Caribbean
Northern Africa and Western Asia
Oceania*
Sub-Saharan Africa
Region

Here's the vertical bar chart showing the life expectancy data by region, gender, and year. The "World" region and "Both sexes" data points have been excluded as requested. Within each region, the bars are ordered first by gender (male or female) and then by year (1990, 2021, 2050). This format allows for a clear comparison of life expectancy trends over time and across genders for each region [>_]

Leveraging ChatGPT for Data Analysis and Visualization



```python
# Creating custom color palette for males and females
custom_palette = {'1990 Females': 'lightcoral', '2021 Femal
                  '1990 Males': 'lightblue', '2021 Males':

# Creating a bar plot with the custom color palette
plt.figure(figsize=(15, 10))
sns.barplot(x="Region", y="Life Expectancy", hue="Year and 

# Adding title and labels
plt.title('Life Expectancy by Region, Gender, and Year (Exc
plt.xlabel('Region')
plt.xticks(rotation=45)
plt.ylabel('Life Expectancy (years)')

# Adjusting layout and showing the plot
plt.tight_layout()
plt.show()
```

cf_result

ancy at bi

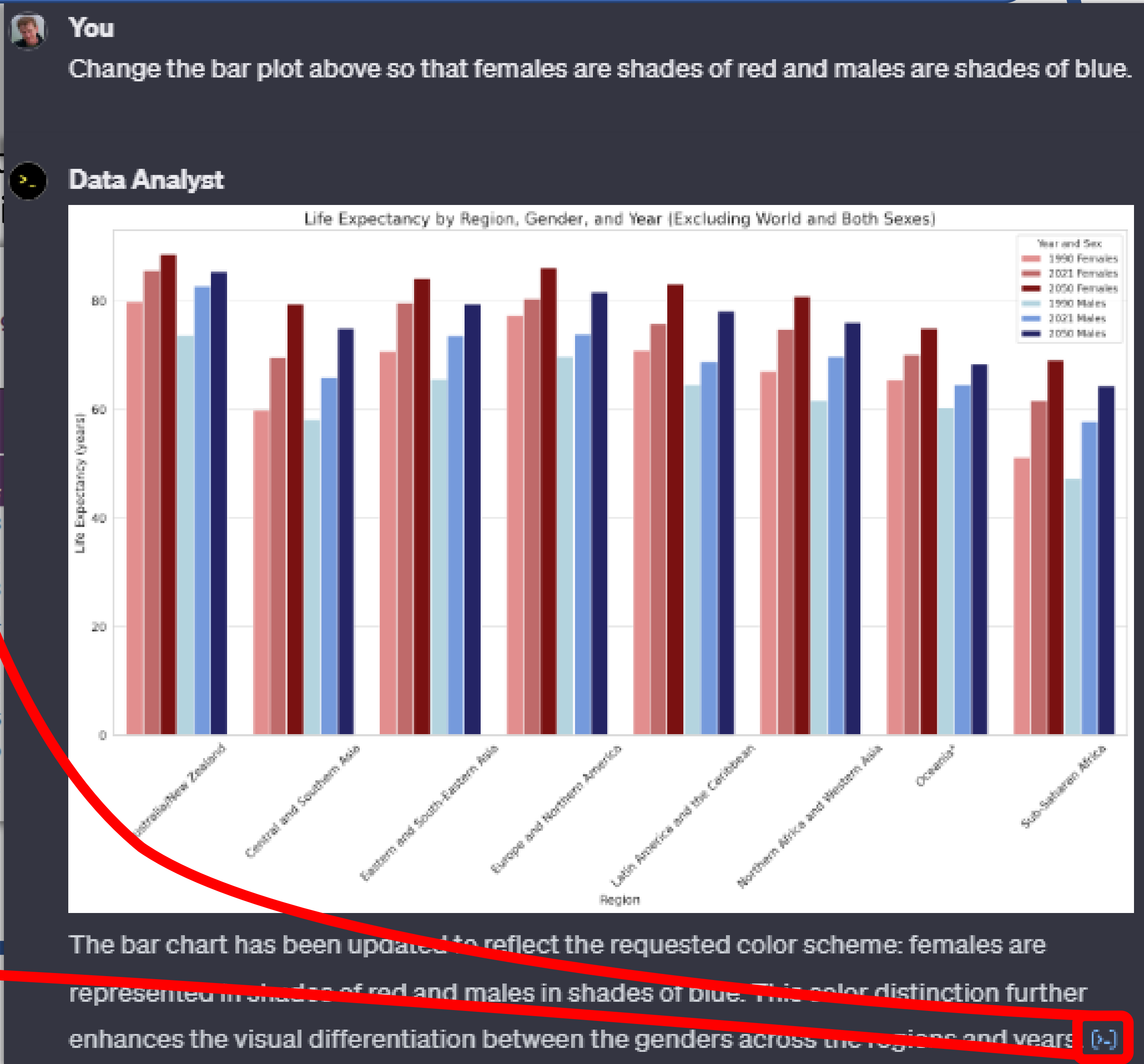
of countries, 19

2050	
Males	*Females*
74.8	**79.8**
64.3	69.1
76.0	[illegible]
74.9	79.4
79.4	84.1
78.1	83.1
85.4	88.6
68.4	74.9
81.6	86.1

of the

www.uvi.edu

HISTORICALLY AMERICAN.
UNIQUELY CARIBBEAN.
GLOBALLY INTERACTIVE.


Life Expectancy by Region, Gender, and Year (Excluding World and Both Sexes)
Year and Sex
1990 Females
2021 Females
2050 Females
1990 Males
2021 Males
2050 Males
Life Expectancy (years)
0
20
40
60
80
Australia/New Zealand
Central and Southern Asia
Eastern and South-Eastern Asia
Europe and Northern America
Latin America and the Caribbean
Northern Africa and Western Asia
Oceania*
Sub-Saharan Africa
Region

You

Change the bar plot above so that females are shades of red and males are shades of blue.

Data Analyst

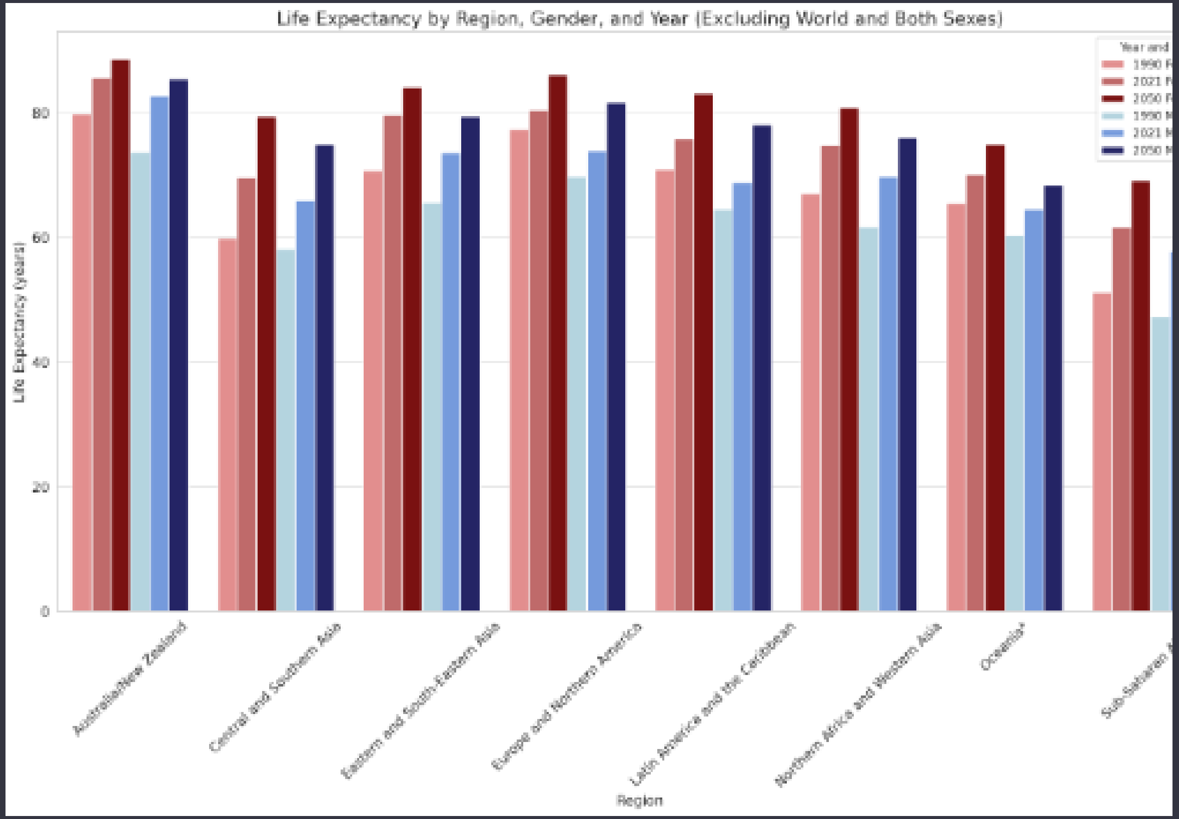
The bar chart has been updated to reflect the requested color scheme: females are represented in shades of red and males in shades of blue. This color distinction further enhances the visual differentiation between the genders across the regions and years.

Leveraging ChatGPT for Data

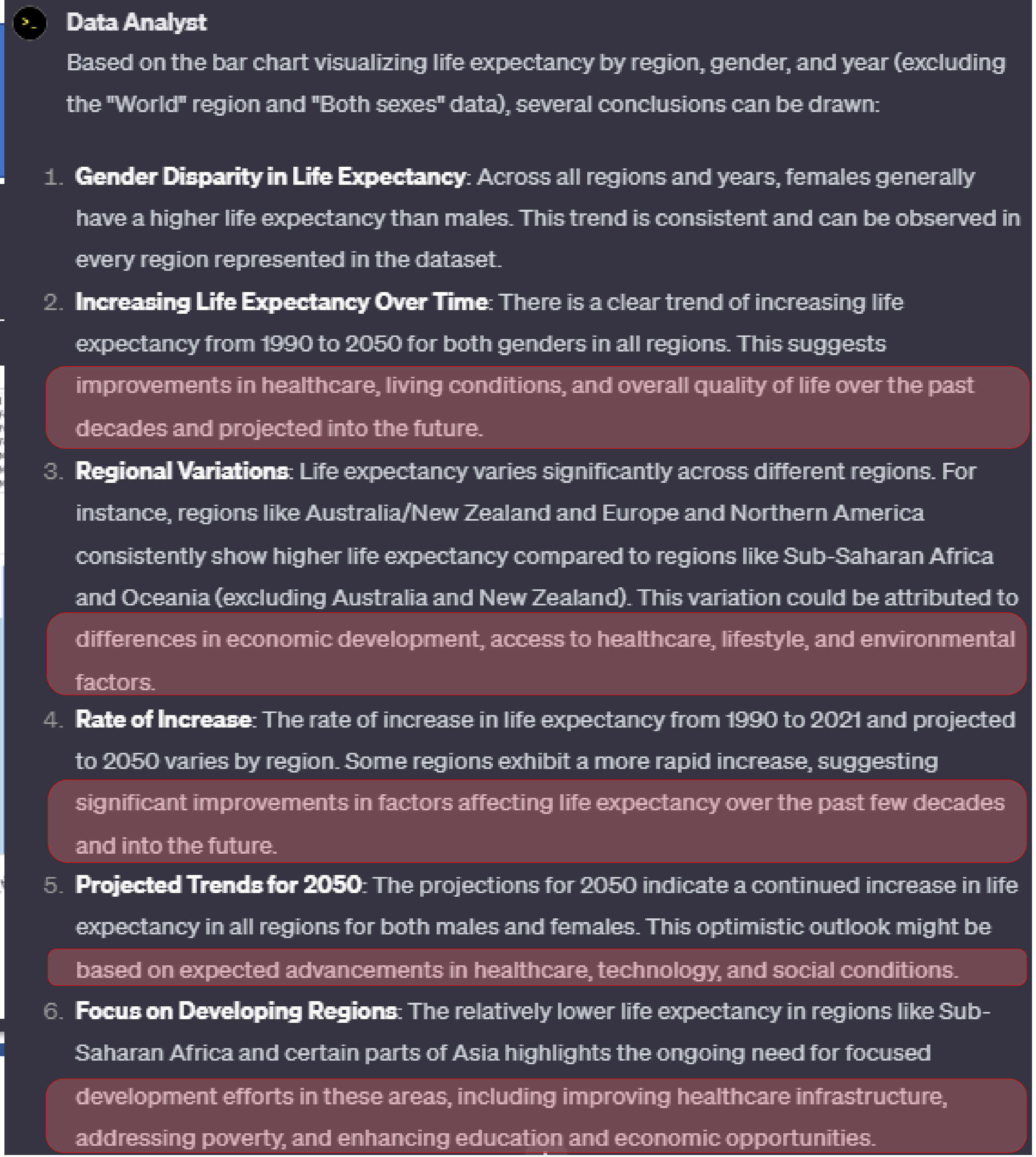
You

What conclusions can be drawn from this data?

Data Analyst



… a bit of editorializing?



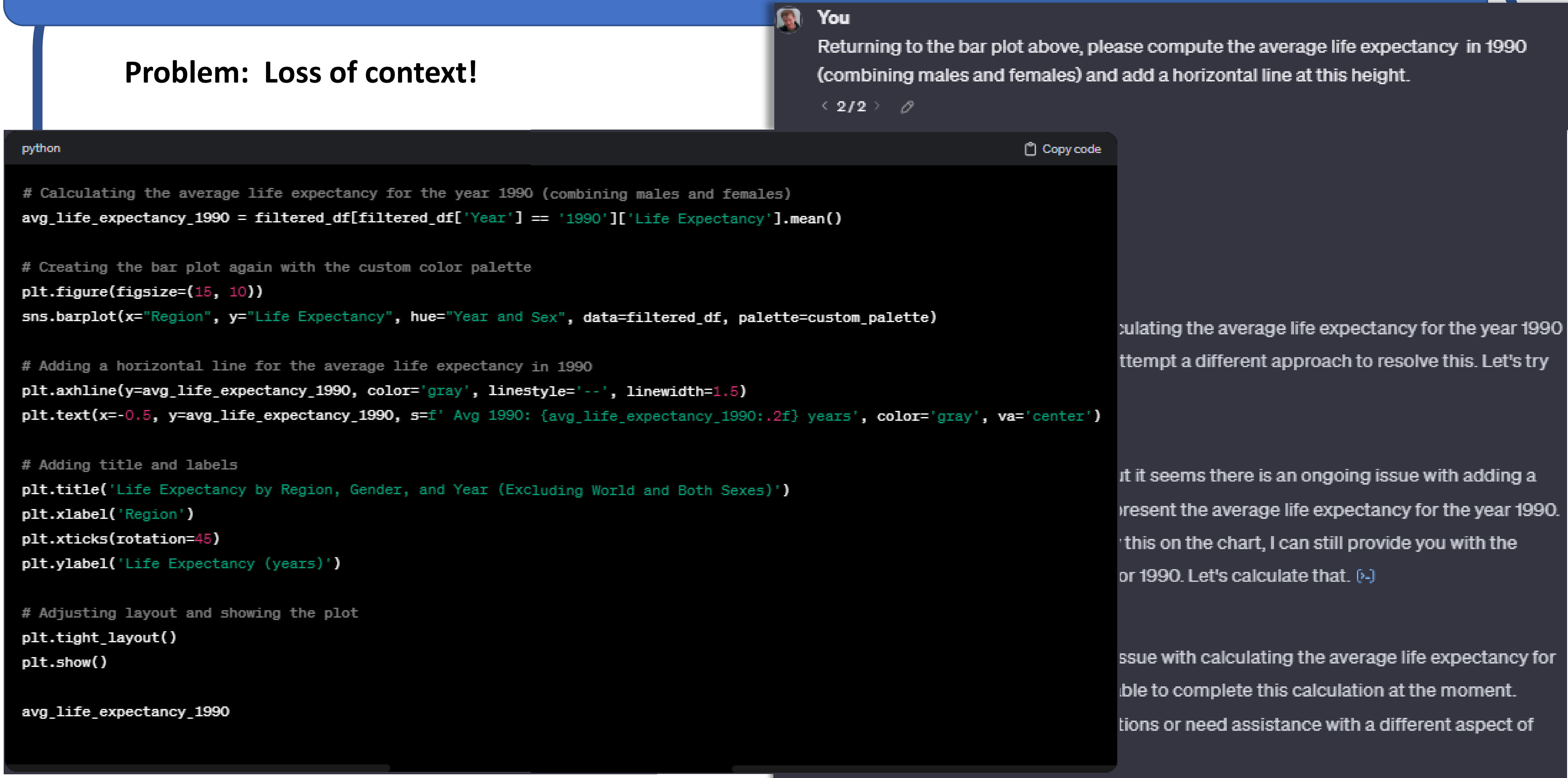
Data Analyst

Based on the bar chart visualizing life expectancy by region, gender, and year (excluding the "World" region and "Both sexes" data), several conclusions can be drawn:

1. **Gender Disparity in Life Expectancy**: Across all regions and years, females generally have a higher life expectancy than males. This trend is consistent and can be observed in every region represented in the dataset.
2. **Increasing Life Expectancy Over Time**: There is a clear trend of increasing life expectancy from 1990 to 2050 for both genders in all regions. This suggests improvements in healthcare, living conditions, and overall quality of life over the past decades and projected into the future.
3. **Regional Variations**: Life expectancy varies significantly across different regions. For instance, regions like Australia/New Zealand and Europe and Northern America consistently show higher life expectancy compared to regions like Sub-Saharan Africa and Oceania (excluding Australia and New Zealand). This variation could be attributed to differences in economic development, access to healthcare, lifestyle, and environmental factors.
4. **Rate of Increase**: The rate of increase in life expectancy from 1990 to 2021 and projected to 2050 varies by region. Some regions exhibit a more rapid increase, suggesting significant improvements in factors affecting life expectancy over the past few decades and into the future.
5. **Projected Trends for 2050**: The projections for 2050 indicate a continued increase in life expectancy in all regions for both males and females. This optimistic outlook might be based on expected advancements in healthcare, technology, and social conditions.
6. **Focus on Developing Regions**: The relatively lower life expectancy in regions like Sub-Saharan Africa and certain parts of Asia highlights the ongoing need for focused development efforts in these areas, including improving healthcare infrastructure, addressing poverty, and enhancing education and economic opportunities.

Leveraging ChatGPT for Data Analysis and Visualization

Problem: Loss of context!



You

Returning to the bar plot above, please compute the average life expectancy in 1990 (combining males and females) and add a horizontal line at this height.

```python
# Calculating the average life expectancy for the year 1990 (combining males and females)
avg_life_expectancy_1990 = filtered_df[filtered_df['Year'] == '1990']['Life Expectancy'].mean()

# Creating the bar plot again with the custom color palette
plt.figure(figsize=(15, 10))
sns.barplot(x="Region", y="Life Expectancy", hue="Year and Sex", data=filtered_df, palette=custom_palette)

# Adding a horizontal line for the average life expectancy in 1990
plt.axhline(y=avg_life_expectancy_1990, color='gray', linestyle='--', linewidth=1.5)
plt.text(x=-0.5, y=avg_life_expectancy_1990, s=f' Avg 1990: {avg_life_expectancy_1990:.2f} years', color='gray', va='center')

# Adding title and labels
plt.title('Life Expectancy by Region, Gender, and Year (Excluding World and Both Sexes)')
plt.xlabel('Region')
plt.xticks(rotation=45)
plt.ylabel('Life Expectancy (years)')

# Adjusting layout and showing the plot
plt.tight_layout()
plt.show()

avg_life_expectancy_1990
```

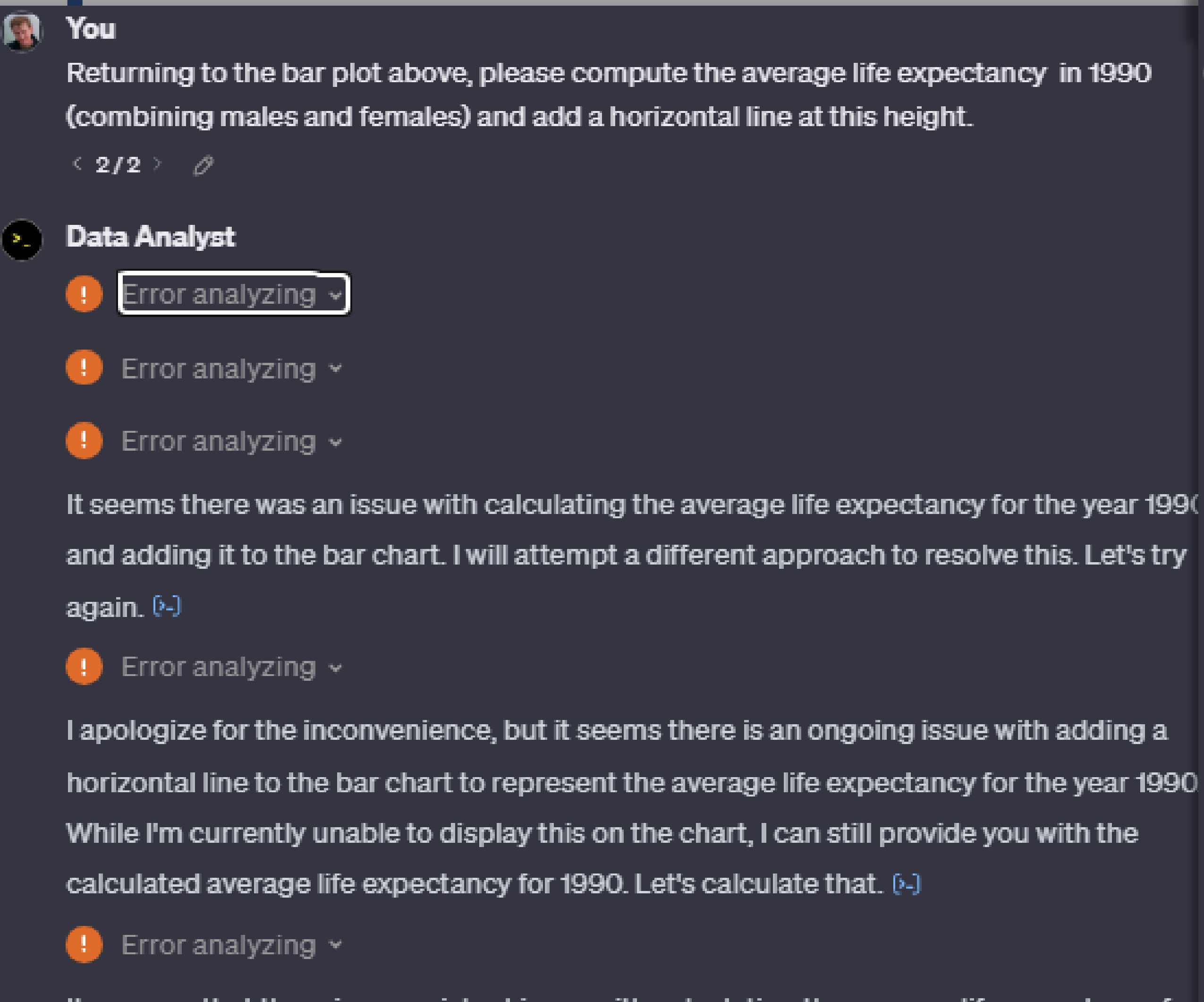
culating the average life expectancy for the year 1990
ttempt a different approach to resolve this. Let's try

ut it seems there is an ongoing issue with adding a
resent the average life expectancy for the year 1990.
this on the chart, I can still provide you with the
or 1990. Let's calculate that.

ssue with calculating the average life expectancy for
ble to complete this calculation at the moment.
tions or need assistance with a different aspect of

Leveraging ChatGPT for Data Analysis and Visualization

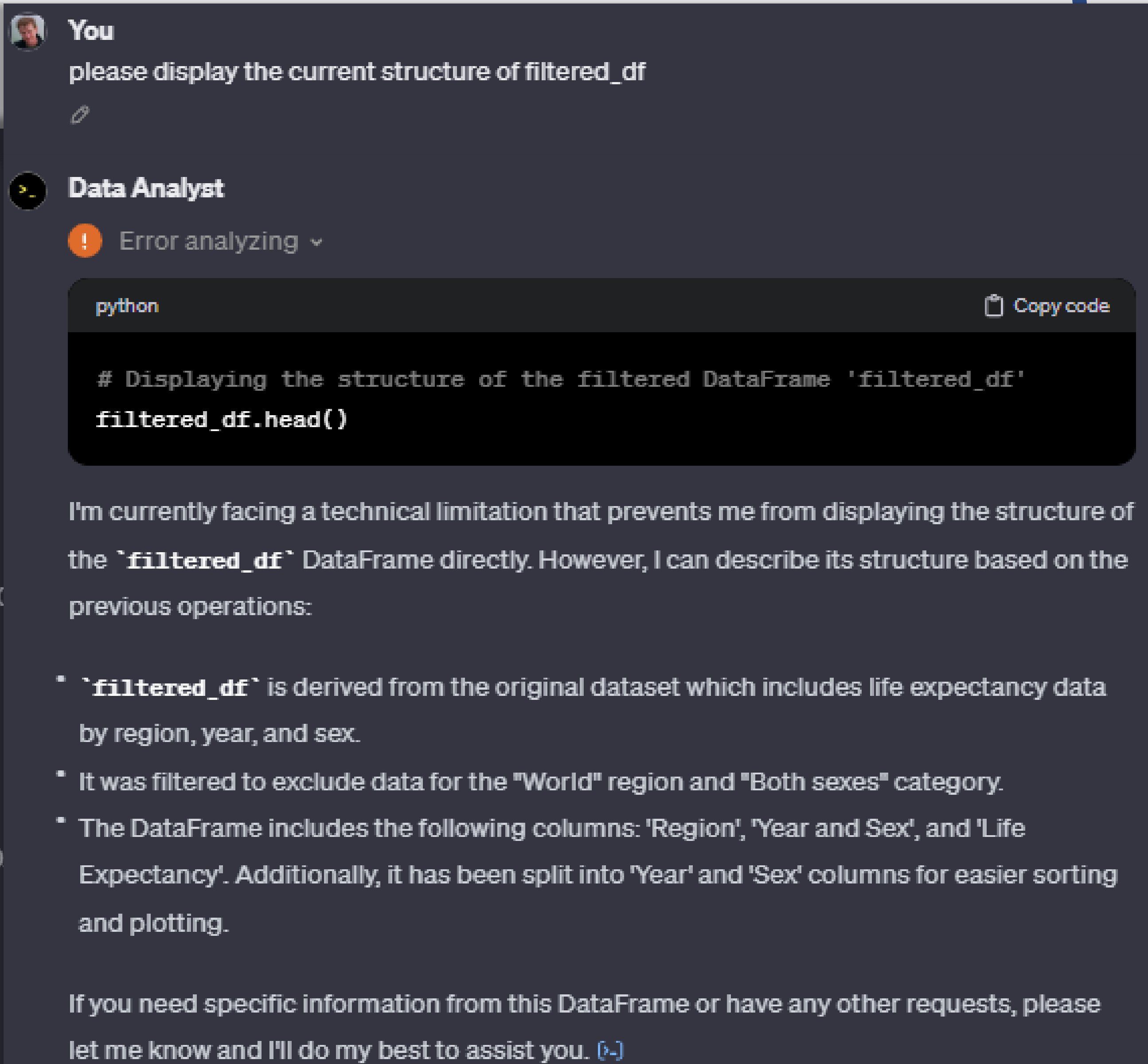
Problem: Loss of context!



Returning to the bar plot above, please compute the average life expectancy in 1990 (combining males and females) and add a horizontal line at this height.

It seems there was an issue with calculating the average life expectancy for the year 199 and adding it to the bar chart. I will attempt a different approach to resolve this. Let's try again.

I apologize for the inconvenience, but it seems there is an ongoing issue with adding a horizontal line to the bar chart to represent the average life expectancy for the year 1990 While I'm currently unable to display this on the chart, I can still provide you with the calculated average life expectancy for 1990. Let's calculate that.



please display the current structure of filtered_df

```python
# Displaying the structure of the filtered DataFrame 'filtered_df'
filtered_df.head()
```

I'm currently facing a technical limitation that prevents me from displaying the structure of the `filtered_df` DataFrame directly. However, I can describe its structure based on the previous operations:

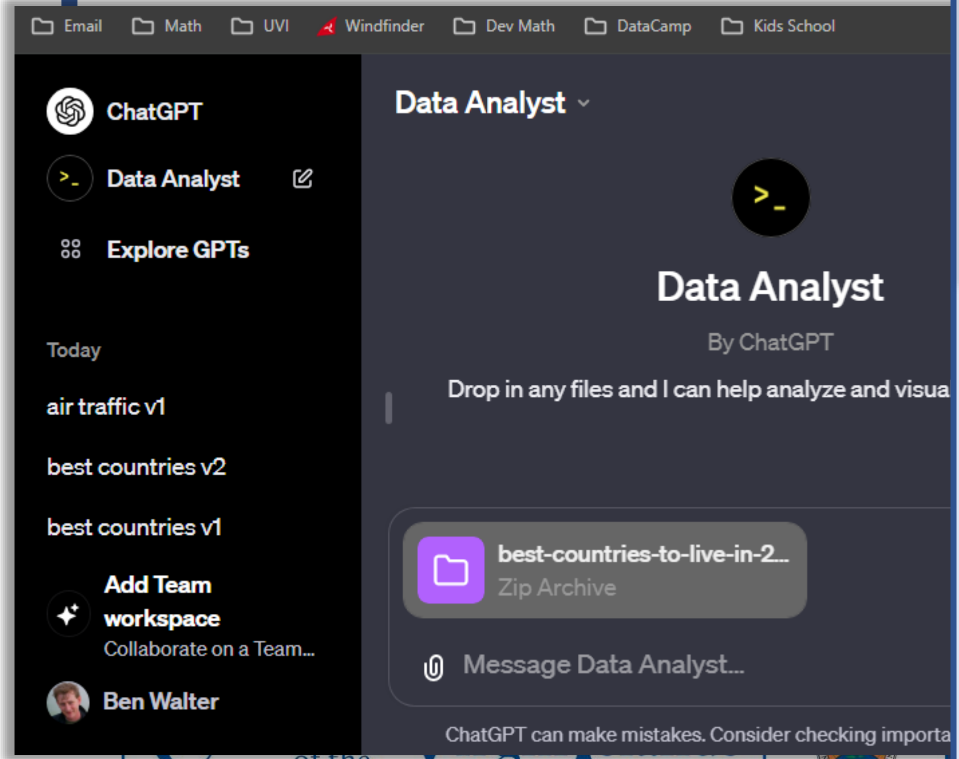
- `filtered_df` is derived from the original dataset which includes life expectancy data by region, year, and sex.
- It was filtered to exclude data for the "World" region and "Both sexes" category.
- The DataFrame includes the following columns: 'Region', 'Year and Sex', and 'Life Expectancy'. Additionally, it has been split into 'Year' and 'Sex' columns for easier sorting and plotting.

If you need specific information from this DataFrame or have any other requests, please let me know and I'll do my best to assist you.

Leveraging ChatGPT for Data Analysis and Visualization

Extra time???

Another example!!!!



Drop in any files and I can help analyze and visua

ChatGPT can make mistakes. Consider checking importa

www.uvi.edu

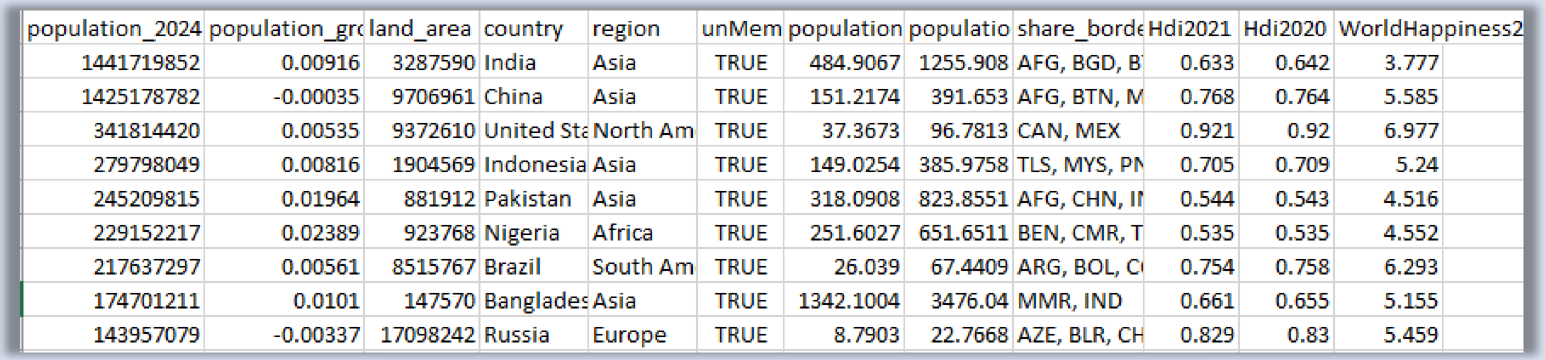
HISTORICALLY AMERICAN.
UNIQUELY CARIBBEAN.
GLOBALLY INTERACTIVE.

Best Country to Live In 2024

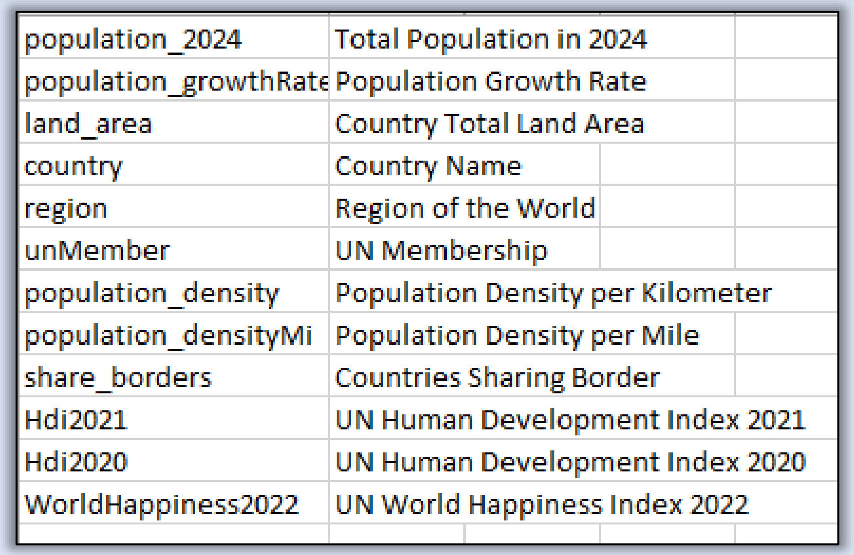
Source: Kaggle

https://www.kaggle.com/datasets/rafsunahmad/best-country-to-live-in-2024/

Size: 12 columns, 141 rows

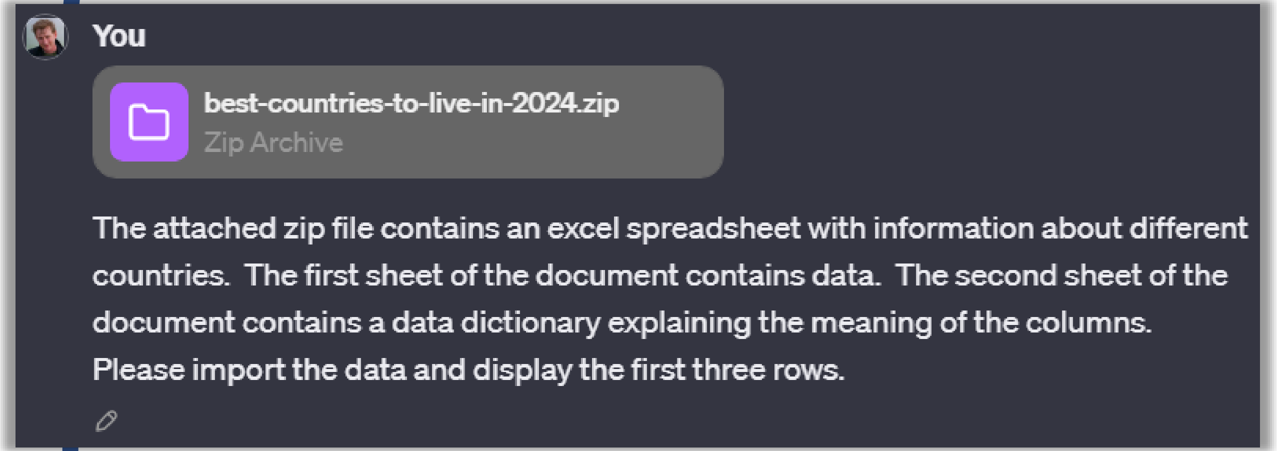


population_2024	population_gr	land_area	country	region	unMem	population	populatio	share_borde	Hdi2021	Hdi2020	WorldHappiness2
1441719852	0.00916	3287590	India	Asia	TRUE	484.9067	1255.908	AFG, BGD, B	0.633	0.642	3.777
1425178782	-0.00035	9706961	China	Asia	TRUE	151.2174	391.653	AFG, BTN, M	0.768	0.764	5.585
341814420	0.00535	9372610	United Sta	North Am	TRUE	37.3673	96.7813	CAN, MEX	0.921	0.92	6.977
279798049	0.00816	1904569	Indonesia	Asia	TRUE	149.0254	385.9758	TLS, MYS, PN	0.705	0.709	5.24
245209815	0.01964	881912	Pakistan	Asia	TRUE	318.0908	823.8551	AFG, CHN, IN	0.544	0.543	4.516
229152217	0.02389	923768	Nigeria	Africa	TRUE	251.6027	651.6511	BEN, CMR, T	0.535	0.535	4.552
217637297	0.00561	8515767	Brazil	South Am	TRUE	26.039	67.4409	ARG, BOL, C	0.754	0.758	6.293
174701211	0.0101	147570	Banglades	Asia	TRUE	1342.1004	3476.04	MMR, IND	0.661	0.655	5.155
143957079	-0.00337	17098242	Russia	Europe	TRUE	8.7903	22.7668	AZE, BLR, CH	0.829	0.83	5.459



population_2024	Total Population in 2024
population_growthRate	Population Growth Rate
land_area	Country Total Land Area
country	Country Name
region	Region of the World
unMember	UN Membership
population_density	Population Density per Kilometer
population_densityMi	Population Density per Mile
share_borders	Countries Sharing Border
Hdi2021	UN Human Development Index 2021
Hdi2020	UN Human Development Index 2020
WorldHappiness2022	UN World Happiness Index 2022

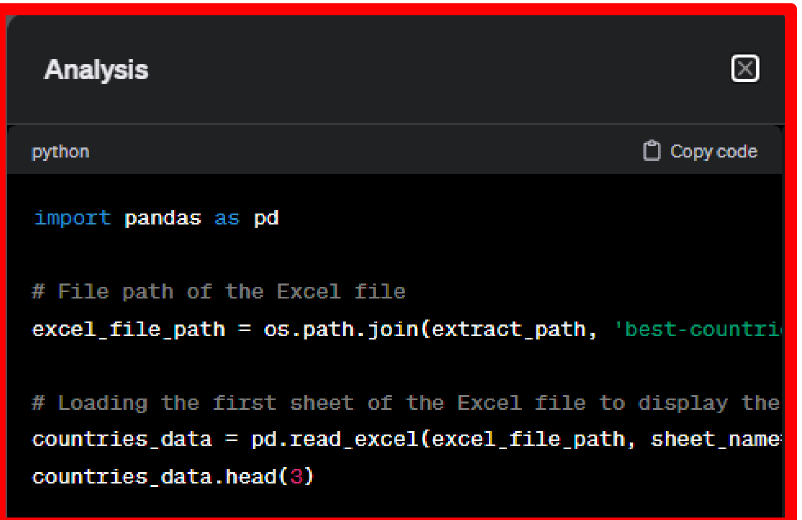
Leveraging ChatGPT for Data Analysis and Visualization



You

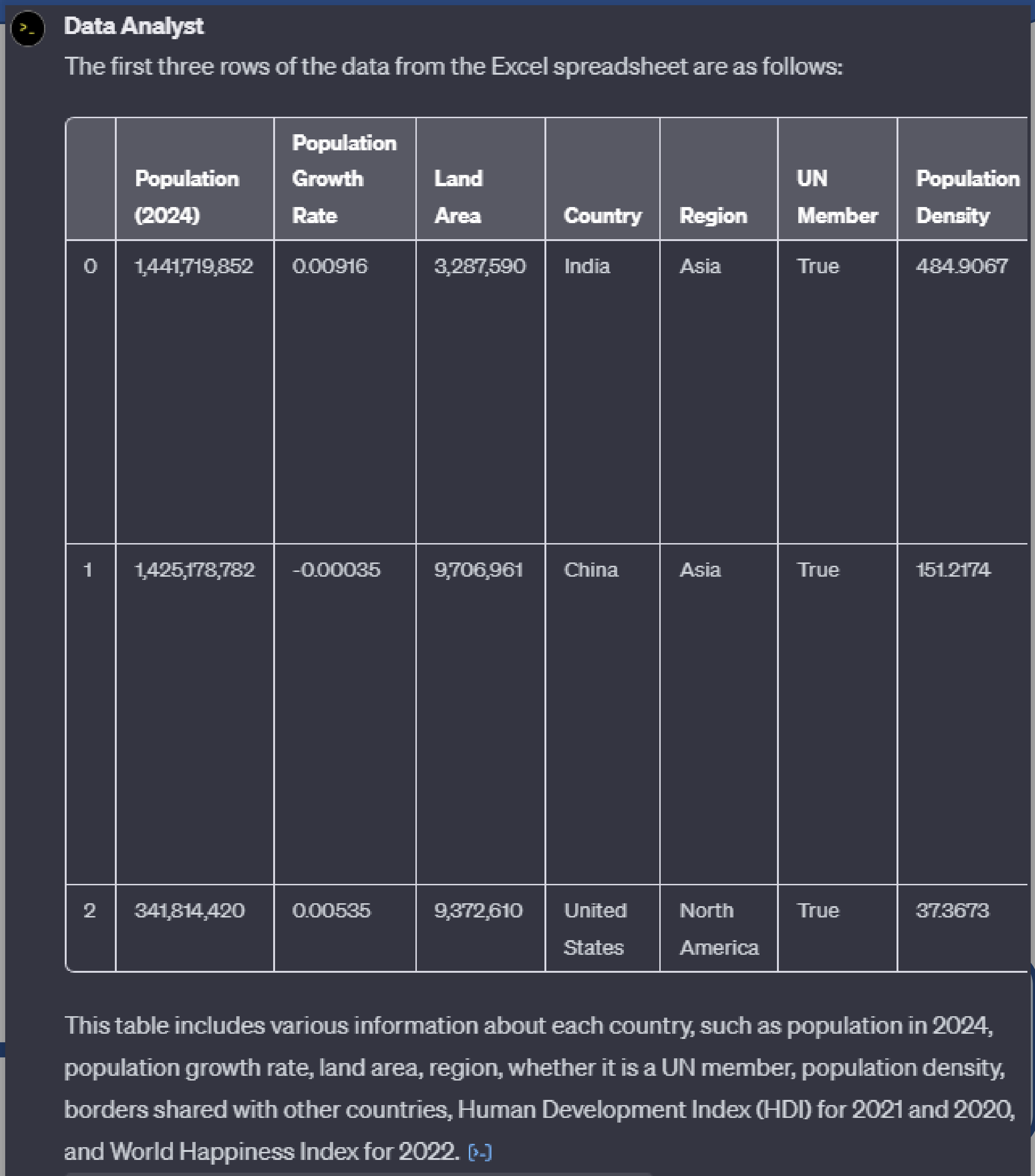
best-countries-to-live-in-2024.zip
Zip Archive

The attached zip file contains an excel spreadsheet with information about different countries. The first sheet of the document contains data. The second sheet of the document contains a data dictionary explaining the meaning of the columns. Please import the data and display the first three rows.



```python
import pandas as pd

# File path of the Excel file
excel_file_path = os.path.join(extract_path, 'best-countri

# Loading the first sheet of the Excel file to display the
countries_data = pd.read_excel(excel_file_path, sheet_name
countries_data.head(3)
```

University of the Virgin Islands
www.uvi.edu
SPECIALIZING IN FUTURES
HISTORICALLY AMERICAN. UNIQUELY CARIBBEAN. GLOBALLY INTERACTIVE.



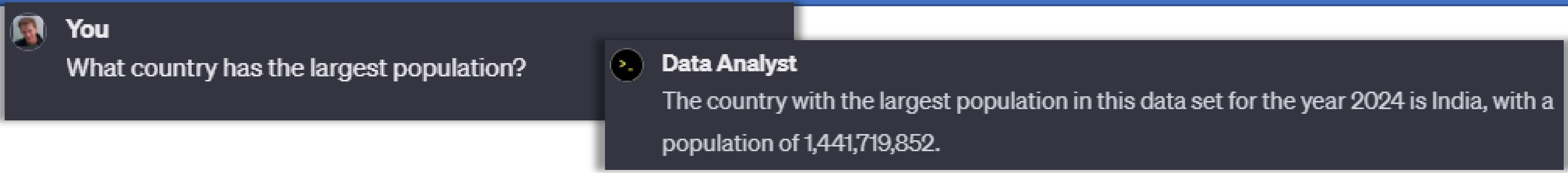
Data Analyst

The first three rows of the data from the Excel spreadsheet are as follows:

	Population (2024)	Population Growth Rate	Land Area	Country	Region	UN Member	Population Density
0	1,441,719,852	0.00916	3,287,590	India	Asia	True	484.9067
1	1,425,178,782	-0.00035	9,706,961	China	Asia	True	151.2174
2	341,814,420	0.00535	9,372,610	United States	North America	True	37.3673

This table includes various information about each country, such as population in 2024, population growth rate, land area, region, whether it is a UN member, population density, borders shared with other countries, Human Development Index (HDI) for 2021 and 2020, and World Happiness Index for 2022.

Leveraging ChatGPT for Data Analysis and Visualization

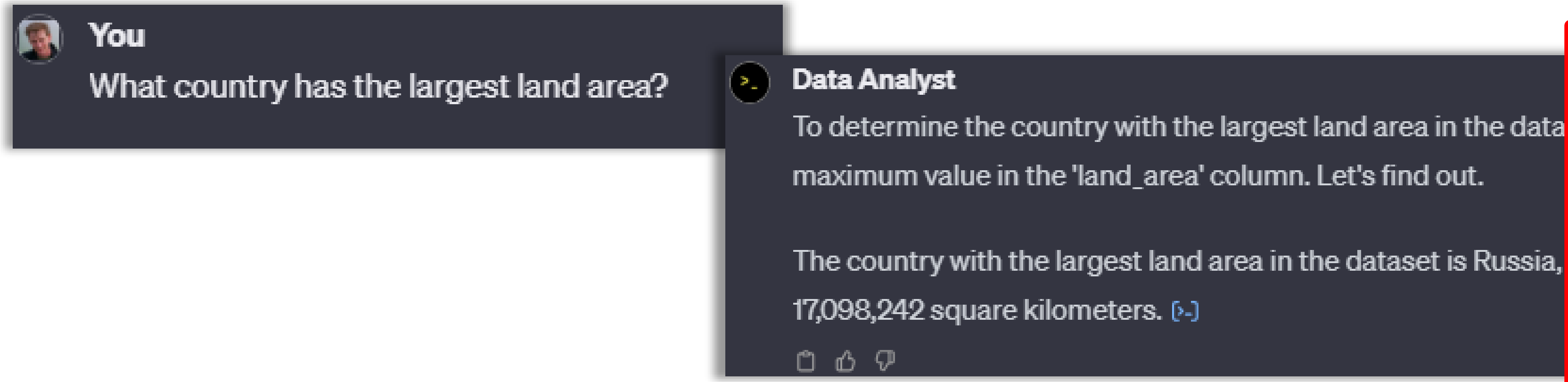


You

What country has the largest population?

Data Analyst

The country with the largest population in this data set for the year 2024 is India, with a population of 1,441,719,852.



You

What country has the largest land area?

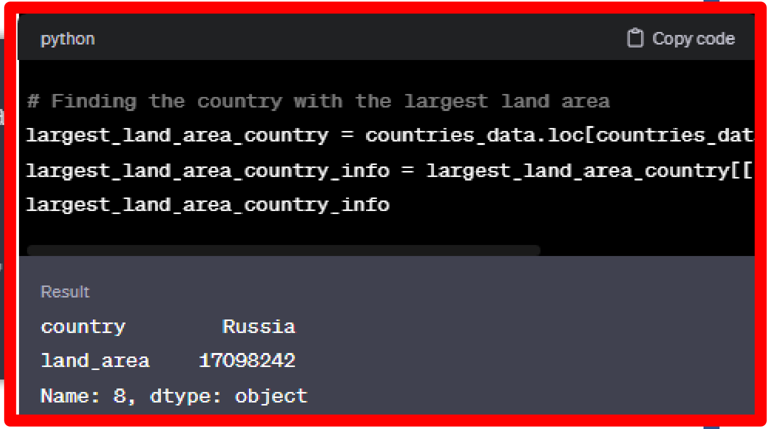
Data Analyst

To determine the country with the largest land area in the data

maximum value in the 'land_area' column. Let's find out.

The country with the largest land area in the dataset is Russia,

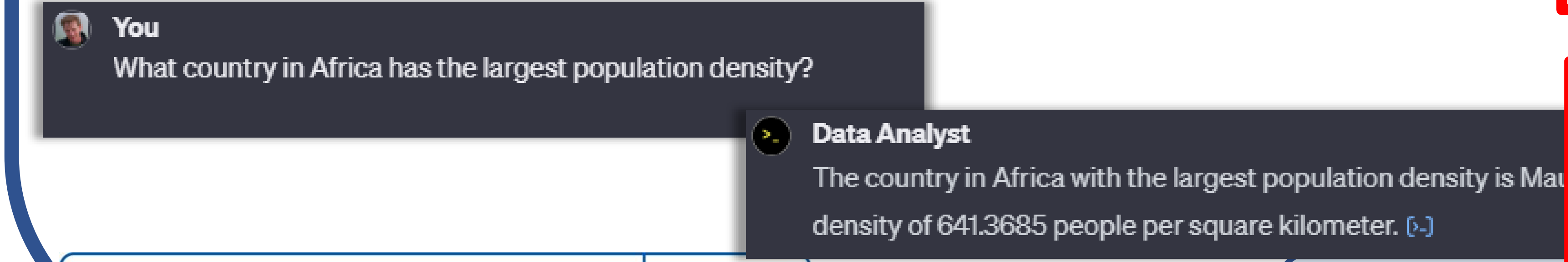
17,098,242 square kilometers.



```python
# Finding the country with the largest land area
largest_land_area_country = countries_data.loc[countries_dat
largest_land_area_country_info = largest_land_area_country[[
largest_land_area_country_info
```

Result

```
country          Russia
land_area      17098242
Name: 8, dtype: object
```



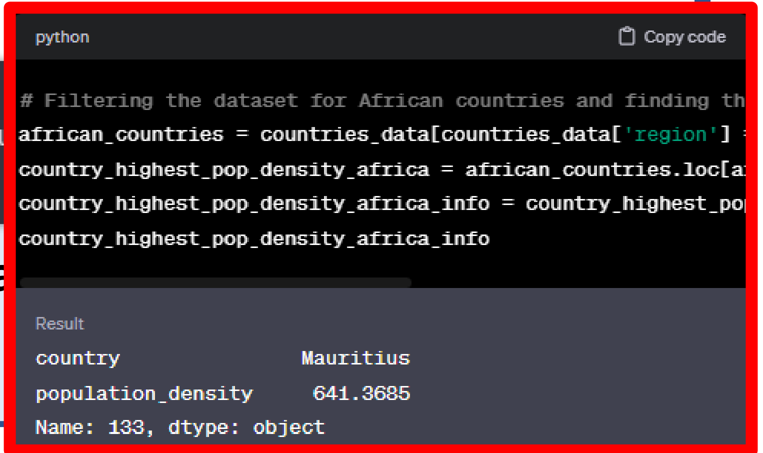
You

What country in Africa has the largest population density?

Data Analyst

The country in Africa with the largest population density is Mau

density of 641.3685 people per square kilometer.



```python
# Filtering the dataset for African countries and finding th
african_countries = countries_data[countries_data['region']
country_highest_pop_density_africa = african_countries.loc[a
country_highest_pop_density_africa_info = country_highest_po
country_highest_pop_density_africa_info
```

Result

```
country                  Mauritius
population_density        641.3685
Name: 133, dtype: object
```

University of the Virgin Islands

www.uvi.edu

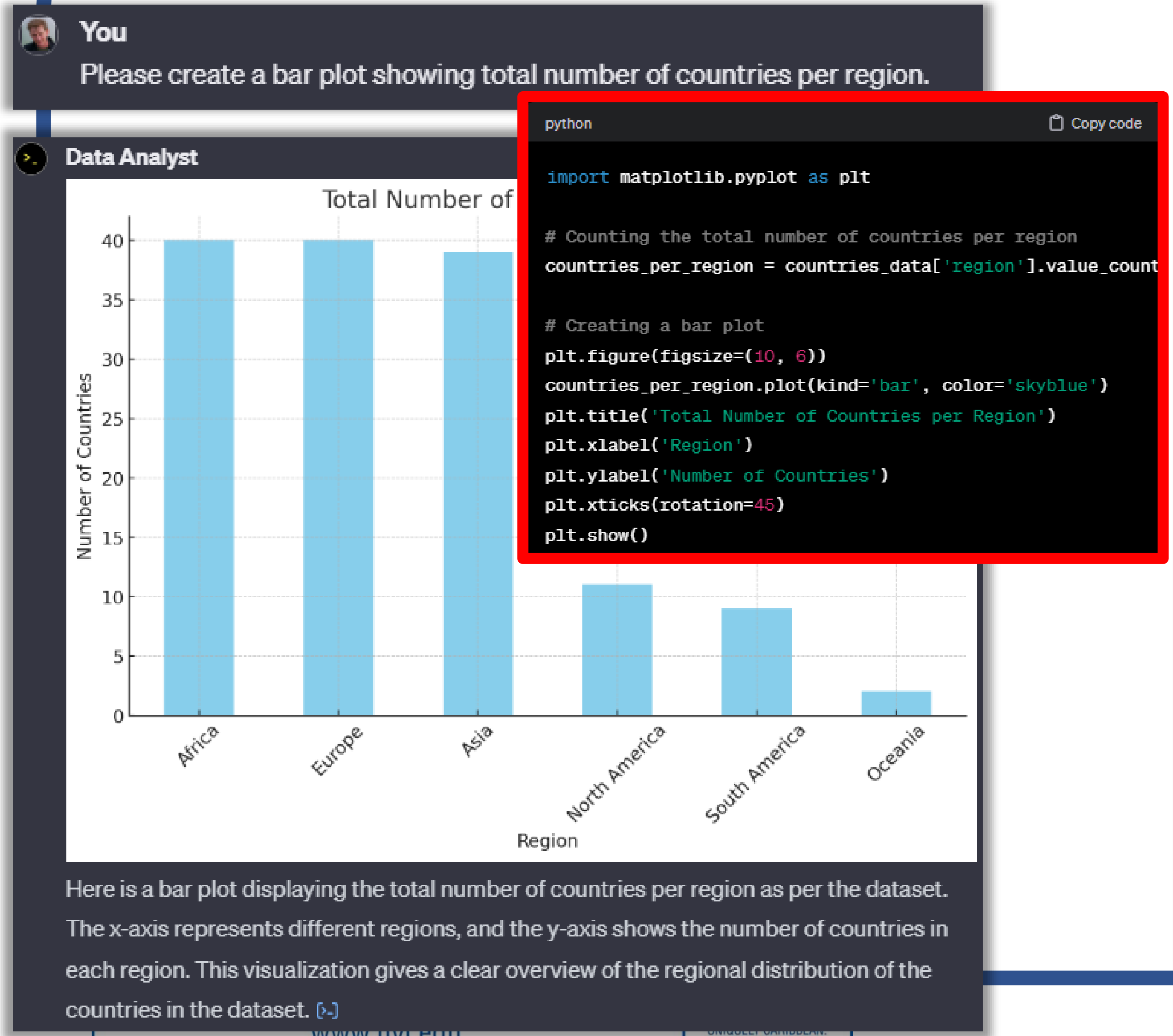
SPECIALIZING IN FUTURES

HISTORICALLY AMERICAN. UNIQUELY CARIBBEAN. GLOBALLY INTERACTIVE.

Dr. Benja

ECS

Leveraging ChatGPT for Data Analysis and Visualization



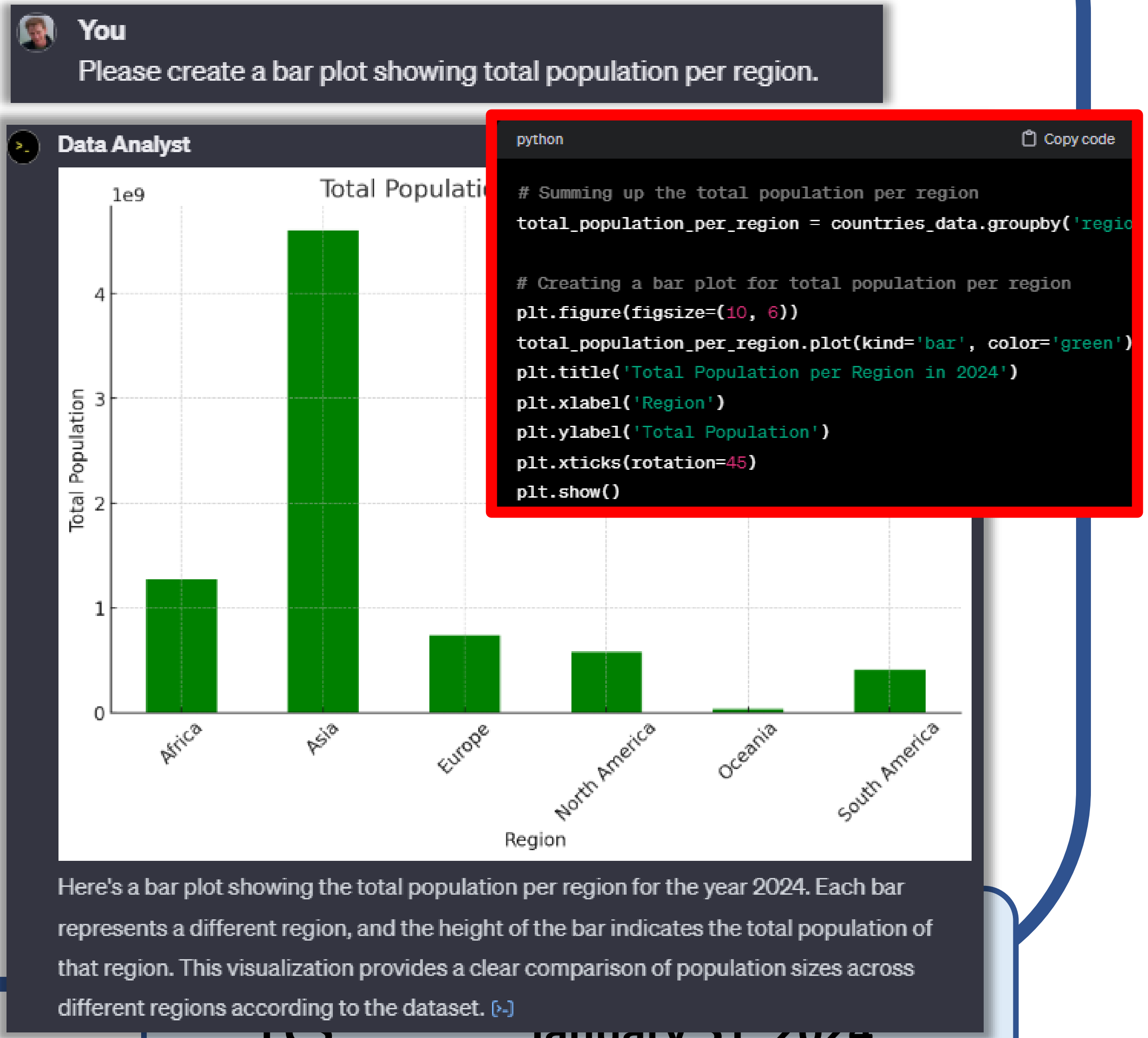
You

Please create a bar plot showing total number of countries per region.

Data Analyst

```python
import matplotlib.pyplot as plt

# Counting the total number of countries per region
countries_per_region = countries_data['region'].value_count

# Creating a bar plot
plt.figure(figsize=(10, 6))
countries_per_region.plot(kind='bar', color='skyblue')
plt.title('Total Number of Countries per Region')
plt.xlabel('Region')
plt.ylabel('Number of Countries')
plt.xticks(rotation=45)
plt.show()
```

Here is a bar plot displaying the total number of countries per region as per the dataset. The x-axis represents different regions, and the y-axis shows the number of countries in each region. This visualization gives a clear overview of the regional distribution of the countries in the dataset.



You

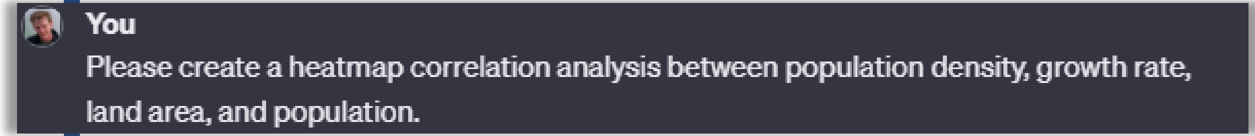
Please create a bar plot showing total population per region.

Data Analyst

```python
# Summing up the total population per region
total_population_per_region = countries_data.groupby('regio

# Creating a bar plot for total population per region
plt.figure(figsize=(10, 6))
total_population_per_region.plot(kind='bar', color='green')
plt.title('Total Population per Region in 2024')
plt.xlabel('Region')
plt.ylabel('Total Population')
plt.xticks(rotation=45)
plt.show()
```

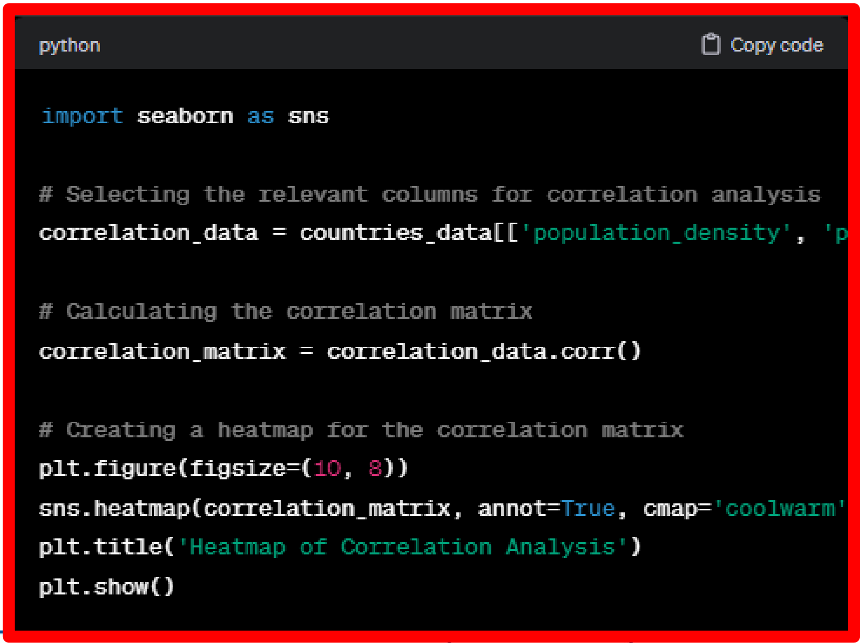
Here's a bar plot showing the total population per region for the year 2024. Each bar represents a different region, and the height of the bar indicates the total population of that region. This visualization provides a clear comparison of population sizes across different regions according to the dataset.

Leveraging ChatGPT for Data Analysis and Visualization

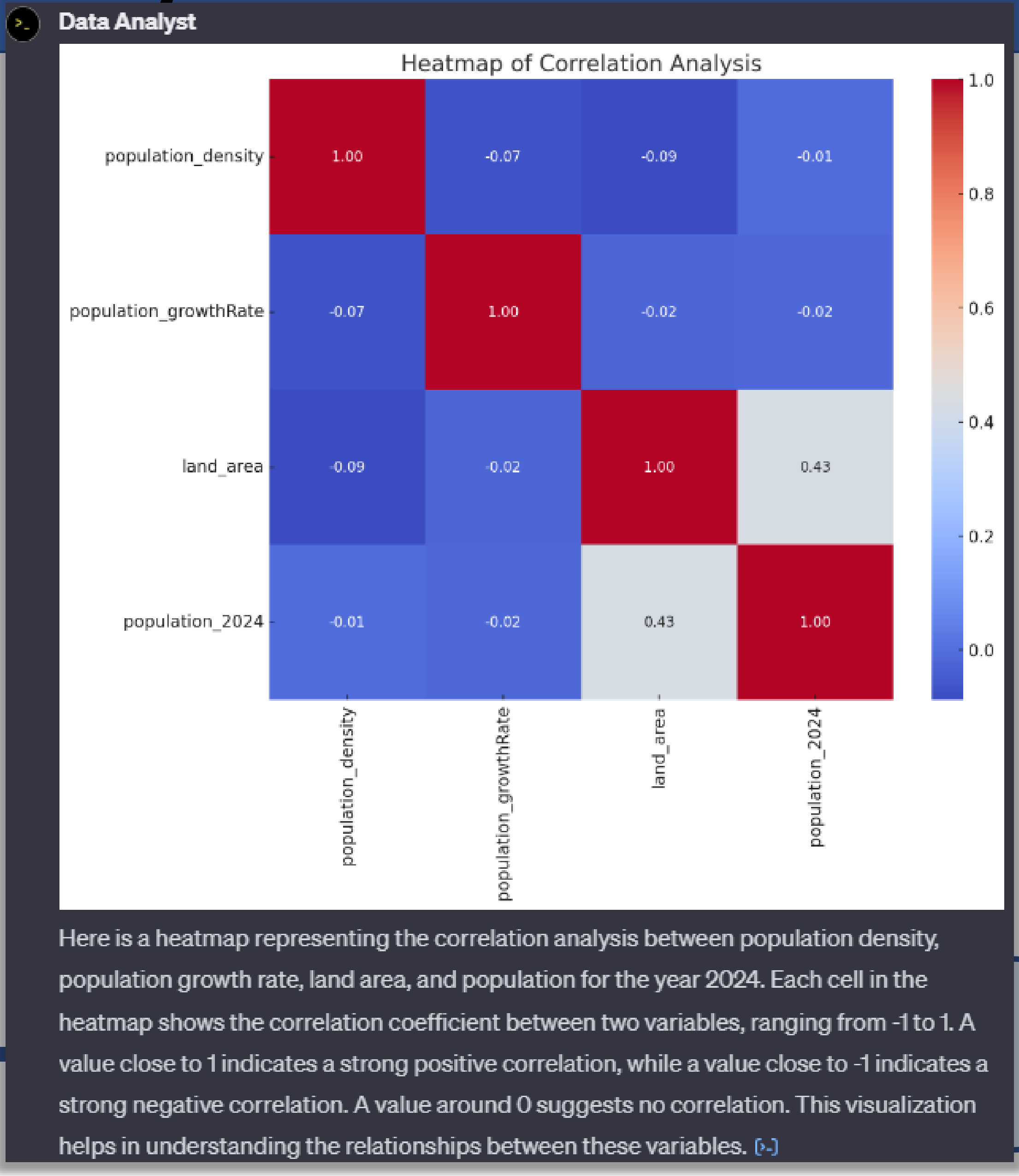


You

Please create a heatmap correlation analysis between population density, growth rate, land area, and population.



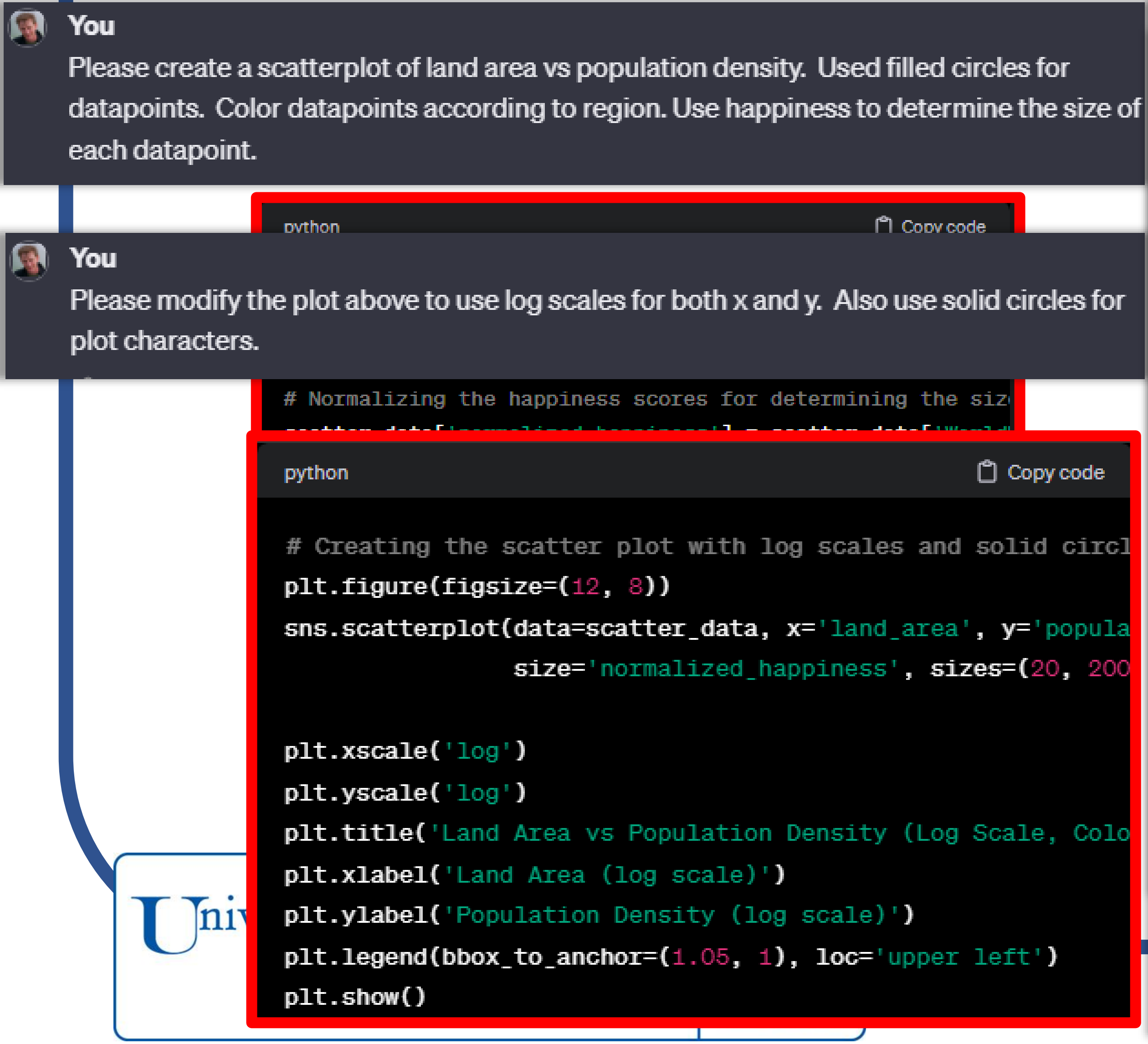
```python
import seaborn as sns

# Selecting the relevant columns for correlation analysis
correlation_data = countries_data[['population_density', 'p

# Calculating the correlation matrix
correlation_matrix = correlation_data.corr()

# Creating a heatmap for the correlation matrix
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm'
plt.title('Heatmap of Correlation Analysis')
plt.show()
```



Here is a heatmap representing the correlation analysis between population density, population growth rate, land area, and population for the year 2024. Each cell in the heatmap shows the correlation coefficient between two variables, ranging from -1 to 1. A value close to 1 indicates a strong positive correlation, while a value close to -1 indicates a strong negative correlation. A value around 0 suggests no correlation. This visualization helps in understanding the relationships between these variables.

University of the Virgin Islands
www.uvi.edu
SPECIALIZING IN FUTURES
HISTORICALLY AMERICAN.
UNIQUELY CARIBBEAN.
GLOBALLY INTERACTIVE.

Leveraging ChatGPT for Data Analysis and Visualization



You

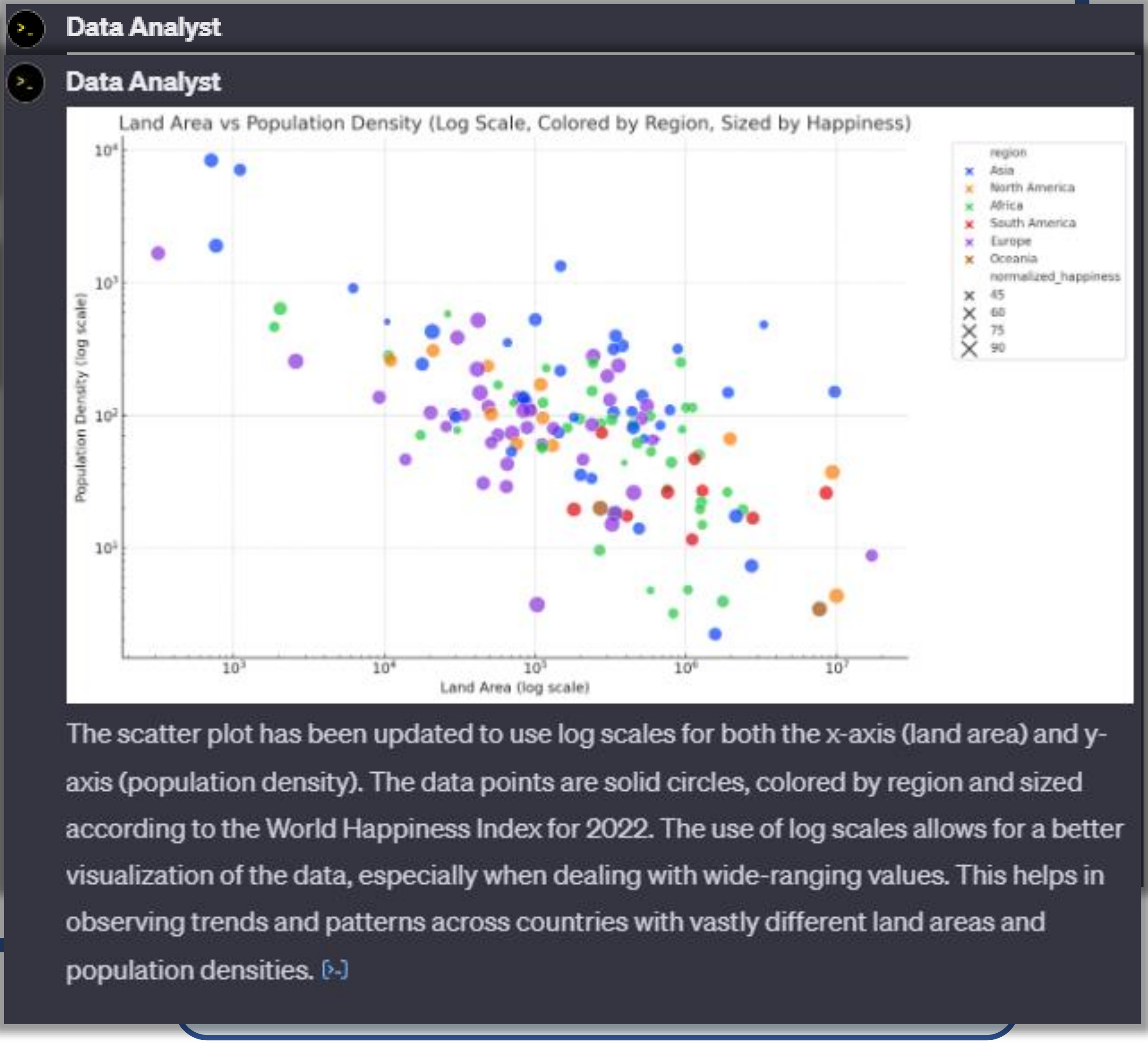
Please create a scatterplot of land area vs population density. Used filled circles for datapoints. Color datapoints according to region. Use happiness to determine the size of each datapoint.

```
# Normalizing the happiness scores for determining the siz
```

You

Please modify the plot above to use log scales for both x and y. Also use solid circles for plot characters.

```
# Creating the scatter plot with log scales and solid circl
plt.figure(figsize=(12, 8))
sns.scatterplot(data=scatter_data, x='land_area', y='popula
                size='normalized_happiness', sizes=(20, 200

plt.xscale('log')
plt.yscale('log')
plt.title('Land Area vs Population Density (Log Scale, Colo
plt.xlabel('Land Area (log scale)')
plt.ylabel('Population Density (log scale)')
plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left')
plt.show()
```

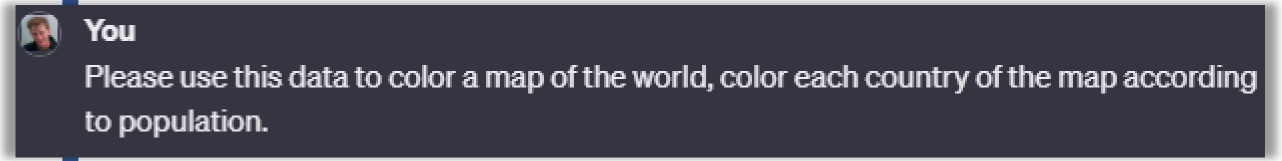


Data Analyst

Data Analyst

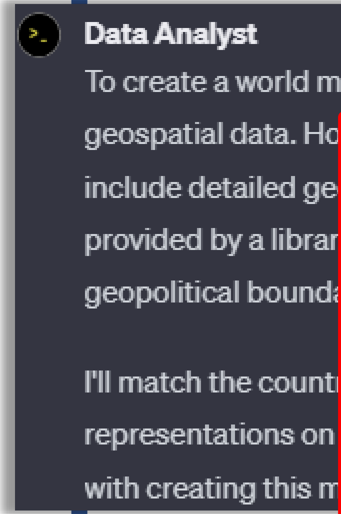
The scatter plot has been updated to use log scales for both the x-axis (land area) and y-axis (population density). The data points are solid circles, colored by region and sized according to the World Happiness Index for 2022. The use of log scales allows for a better visualization of the data, especially when dealing with wide-ranging values. This helps in observing trends and patterns across countries with vastly different land areas and population densities. [>-]

Leveraging ChatGPT for Data Analysis and Visualization



You

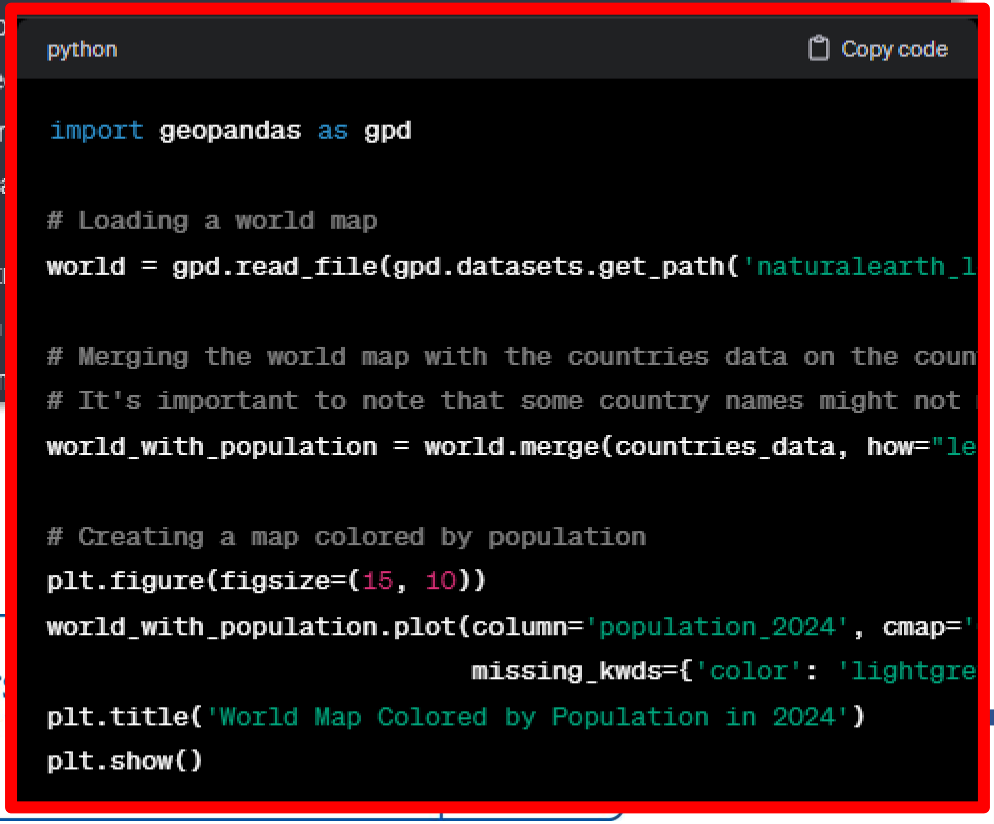
Please use this data to color a map of the world, color each country of the map according to population.



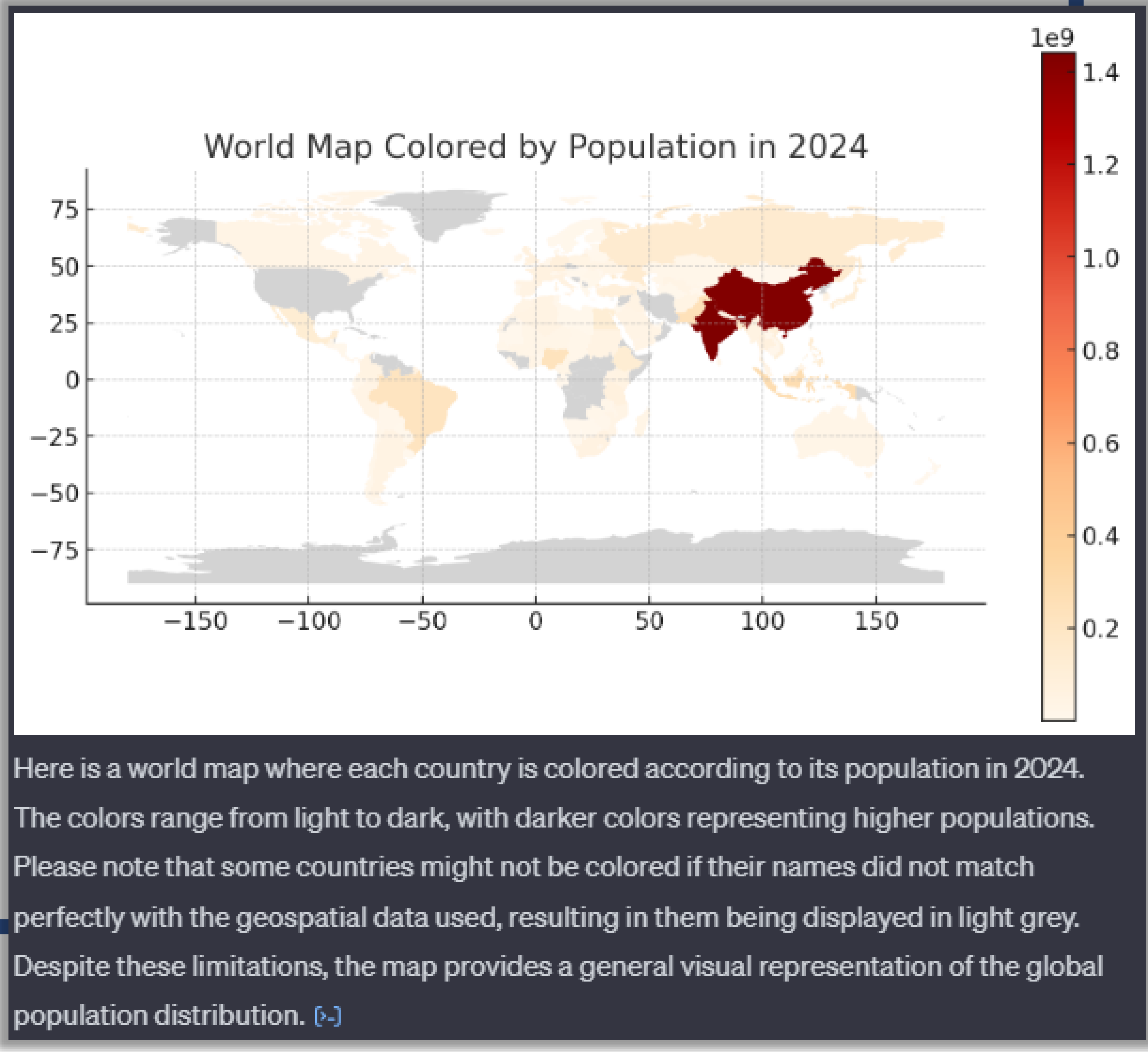
Data Analyst

To create a world map colored according to the population of each country, I will use
geospatial data. Ho
include detailed ge
provided by a librar
geopolitical bound

I'll match the countr
representations on
with creating this m

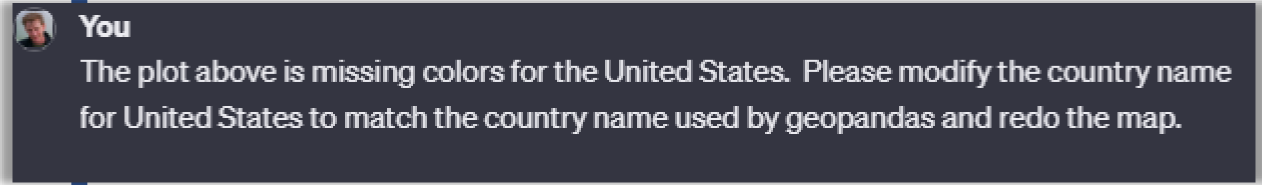


```python
import geopandas as gpd

# Loading a world map
world = gpd.read_file(gpd.datasets.get_path('naturalearth_l

# Merging the world map with the countries data on the coun
# It's important to note that some country names might not 
world_with_population = world.merge(countries_data, how="le

# Creating a map colored by population
plt.figure(figsize=(15, 10))
world_with_population.plot(column='population_2024', cmap='
                           missing_kwds={'color': 'lightgre
plt.title('World Map Colored by Population in 2024')
plt.show()
```



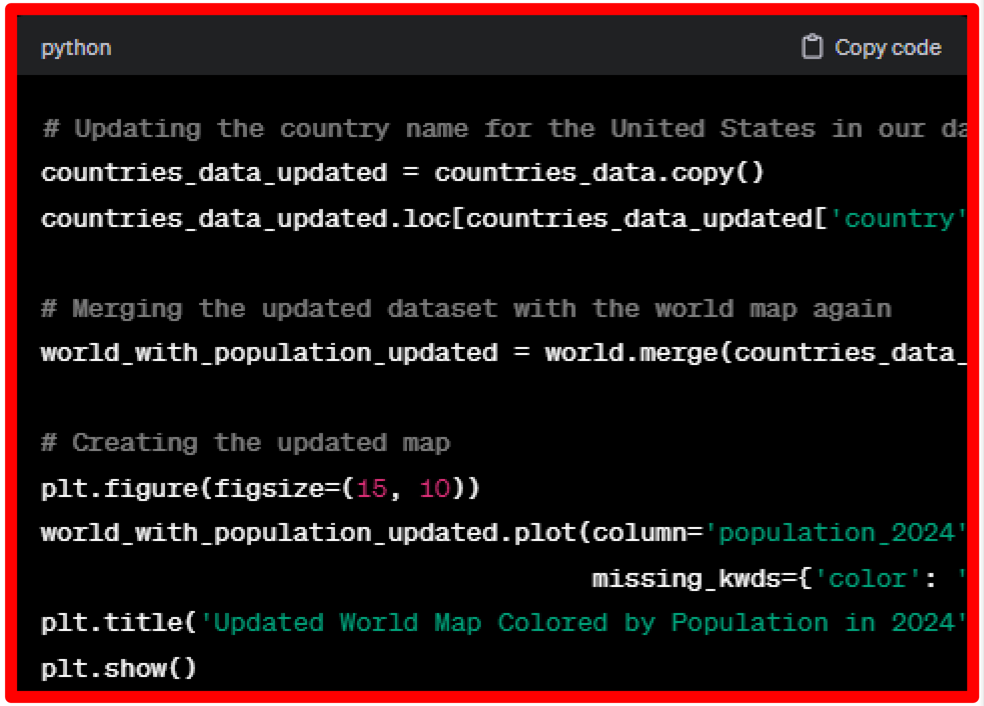
Here is a world map where each country is colored according to its population in 2024. The colors range from light to dark, with darker colors representing higher populations. Please note that some countries might not be colored if their names did not match perfectly with the geospatial data used, resulting in them being displayed in light grey. Despite these limitations, the map provides a general visual representation of the global population distribution.

Leveraging ChatGPT for Data Analysis and Visualization

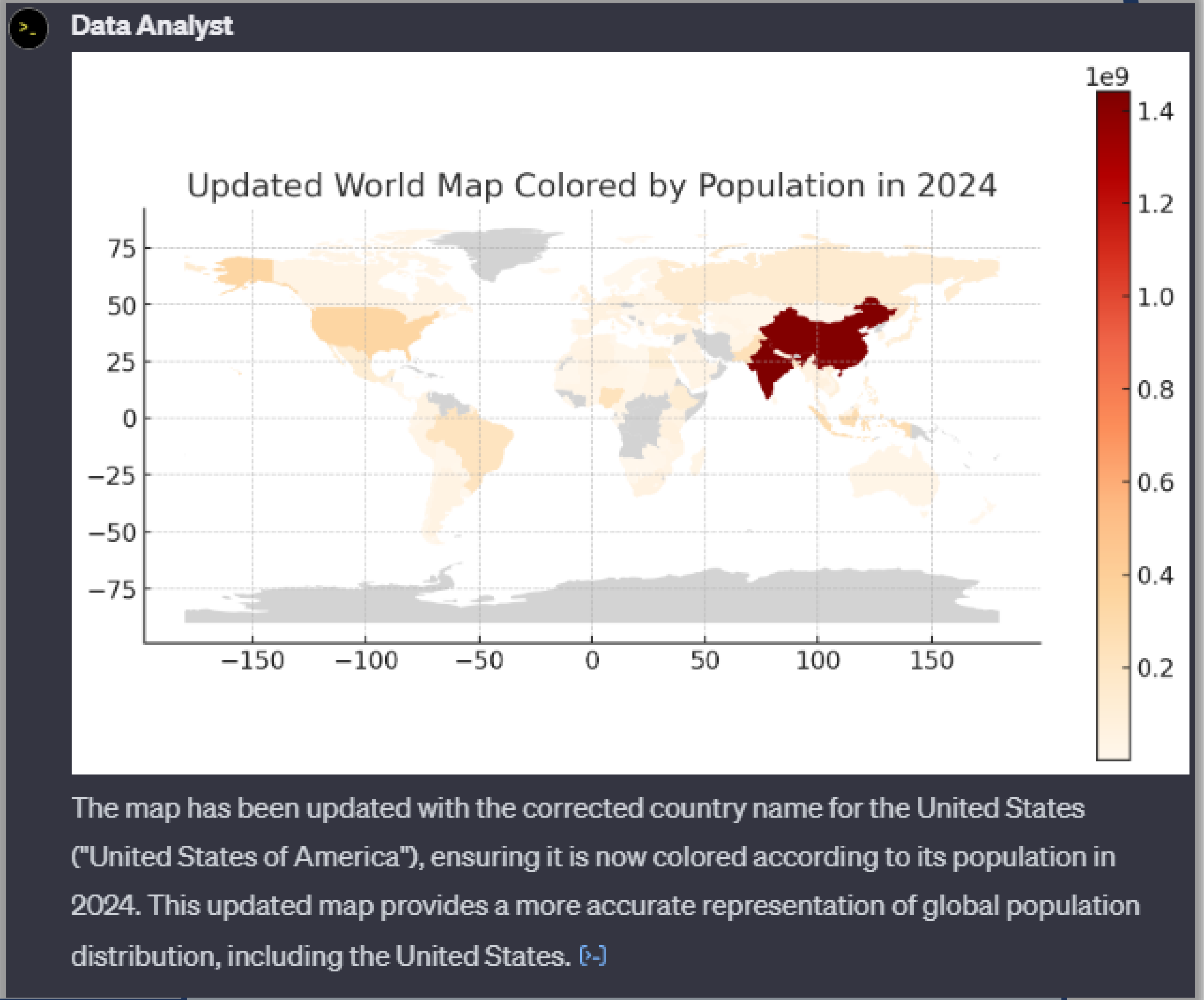


You

The plot above is missing colors for the United States. Please modify the country name for United States to match the country name used by geopandas and redo the map.

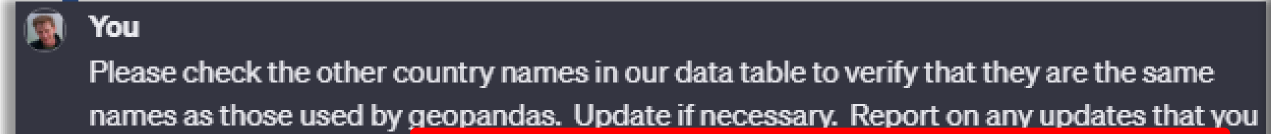


```
# Updating the country name for the United States in our da
countries_data_updated = countries_data.copy()
countries_data_updated.loc[countries_data_updated['country'

# Merging the updated dataset with the world map again
world_with_population_updated = world.merge(countries_data_

# Creating the updated map
plt.figure(figsize=(15, 10))
world_with_population_updated.plot(column='population_2024'
                                   missing_kwds={'color': '
plt.title('Updated World Map Colored by Population in 2024'
plt.show()
```



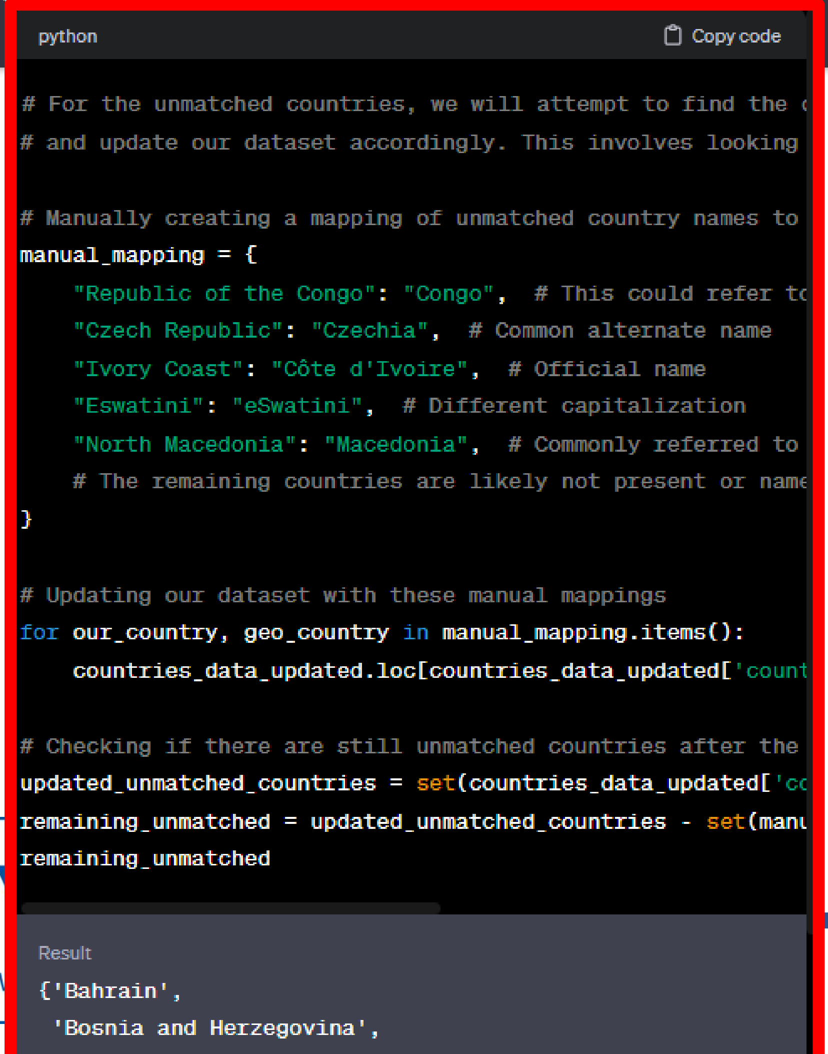
The map has been updated with the corrected country name for the United States ("United States of America"), ensuring it is now colored according to its population in 2024. This updated map provides a more accurate representation of global population distribution, including the United States.

Leveraging ChatGPT for Data Analysis and Visualization



You

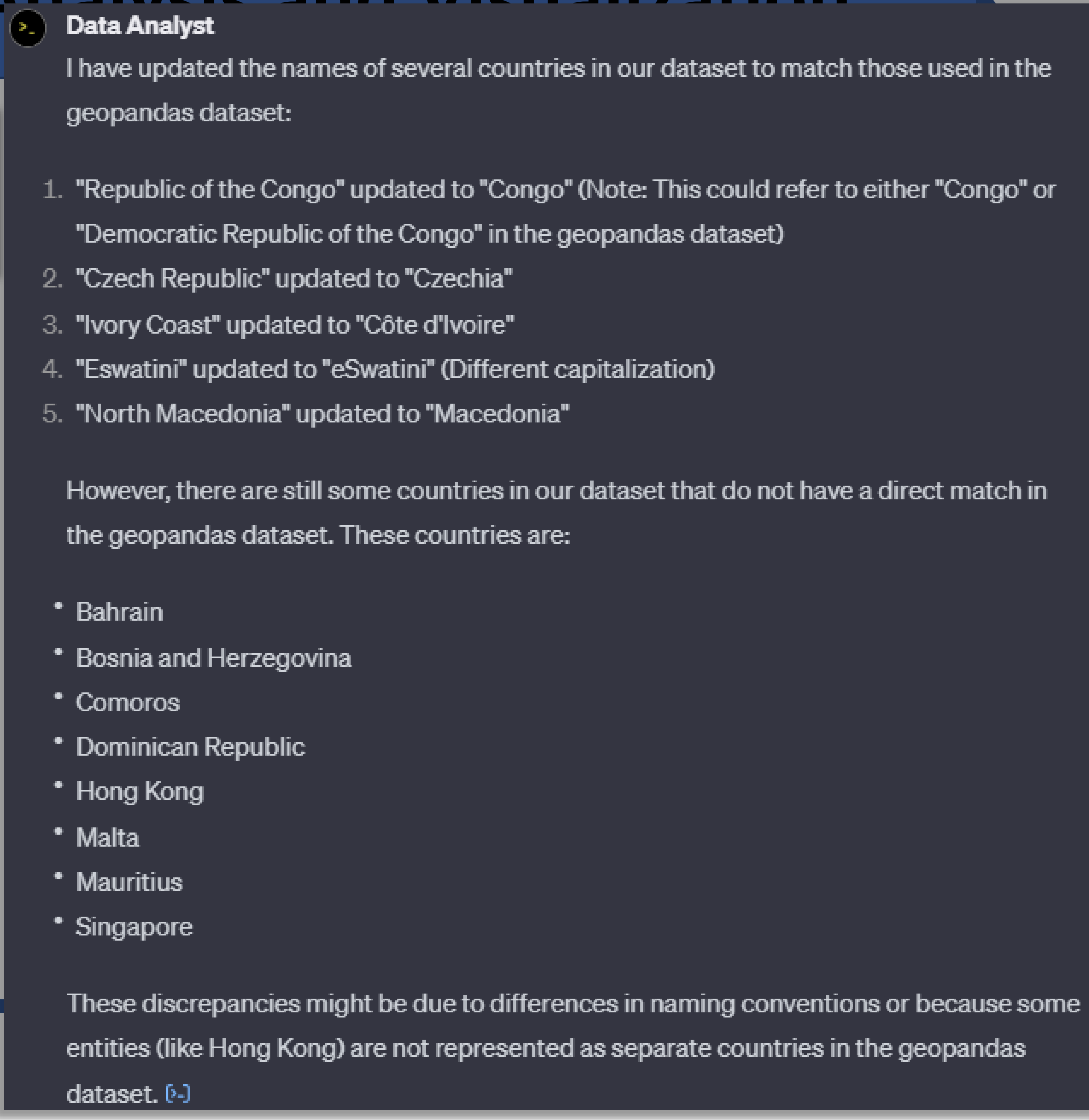
Please check the other country names in our data table to verify that they are the same names as those used by geopandas. Update if necessary. Report on any updates that you do.



```python
# For the unmatched countries, we will attempt to find the c
# and update our dataset accordingly. This involves looking

# Manually creating a mapping of unmatched country names to
manual_mapping = {
    "Republic of the Congo": "Congo",  # This could refer to
    "Czech Republic": "Czechia",  # Common alternate name
    "Ivory Coast": "Côte d'Ivoire",  # Official name
    "Eswatini": "eSwatini",  # Different capitalization
    "North Macedonia": "Macedonia",  # Commonly referred to
    # The remaining countries are likely not present or name
}

# Updating our dataset with these manual mappings
for our_country, geo_country in manual_mapping.items():
    countries_data_updated.loc[countries_data_updated['count

# Checking if there are still unmatched countries after the
updated_unmatched_countries = set(countries_data_updated['co
remaining_unmatched = updated_unmatched_countries - set(manu
remaining_unmatched
```

Result

```
{'Bahrain',
 'Bosnia and Herzegovina',
```



Data Analyst

I have updated the names of several countries in our dataset to match those used in the geopandas dataset:

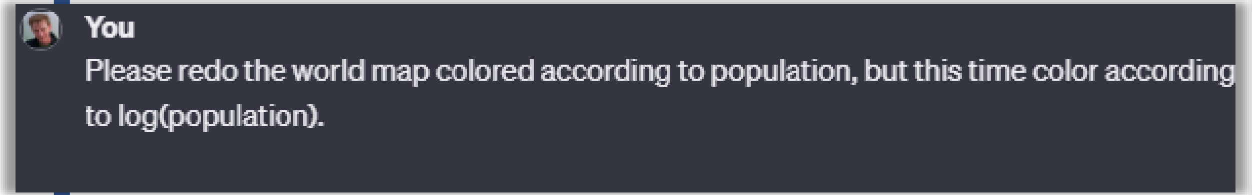
1. "Republic of the Congo" updated to "Congo" (Note: This could refer to either "Congo" or "Democratic Republic of the Congo" in the geopandas dataset)
2. "Czech Republic" updated to "Czechia"
3. "Ivory Coast" updated to "Côte d'Ivoire"
4. "Eswatini" updated to "eSwatini" (Different capitalization)
5. "North Macedonia" updated to "Macedonia"

However, there are still some countries in our dataset that do not have a direct match in the geopandas dataset. These countries are:

- Bahrain
- Bosnia and Herzegovina
- Comoros
- Dominican Republic
- Hong Kong
- Malta
- Mauritius
- Singapore

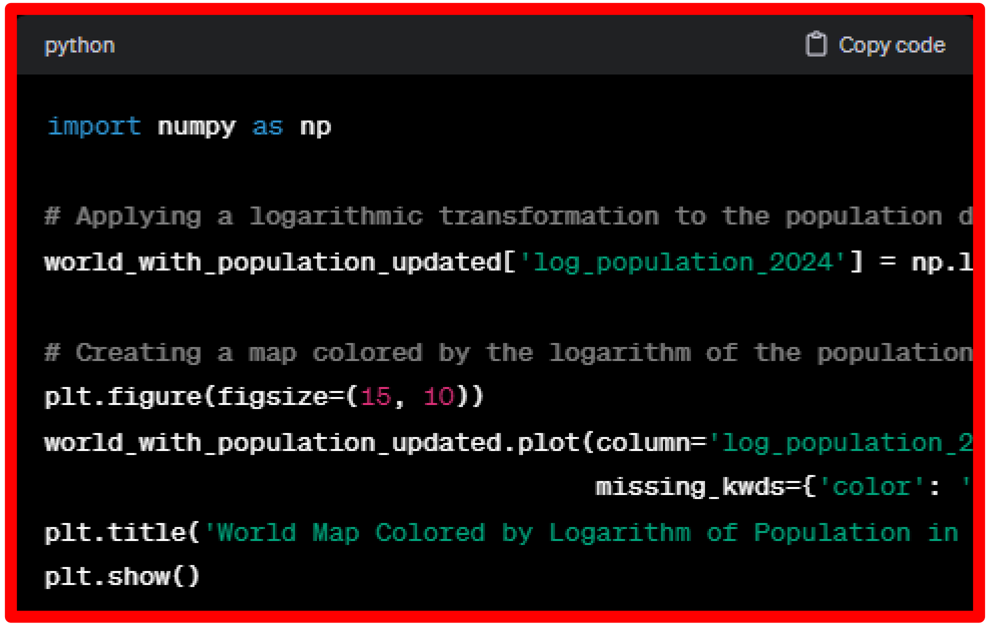
These discrepancies might be due to differences in naming conventions or because some entities (like Hong Kong) are not represented as separate countries in the geopandas dataset.

Leveraging ChatGPT for Data Analysis and Visualization

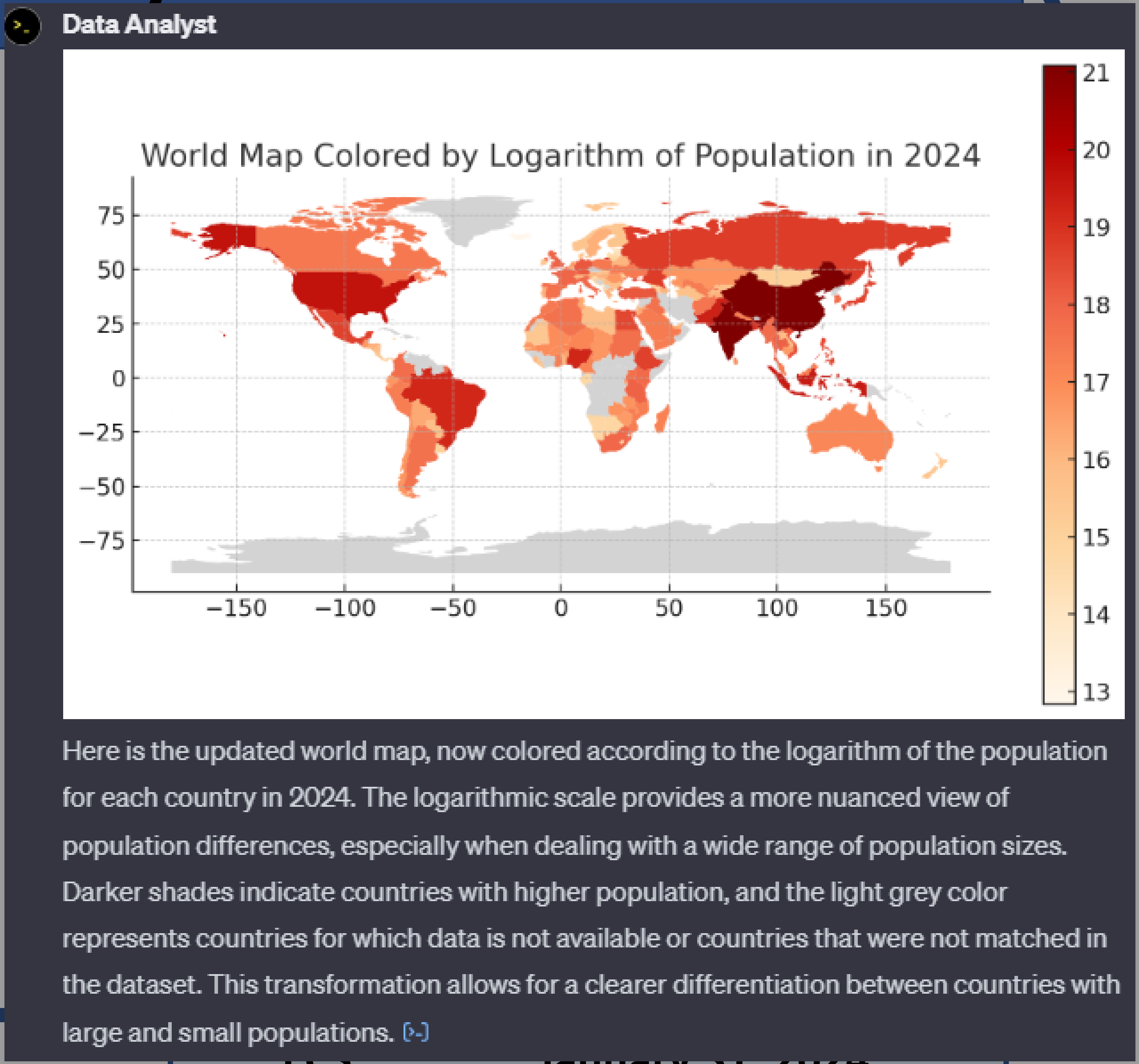


You

Please redo the world map colored according to population, but this time color according to log(population).



```python
import numpy as np

# Applying a logarithmic transformation to the population d
world_with_population_updated['log_population_2024'] = np.l

# Creating a map colored by the logarithm of the population
plt.figure(figsize=(15, 10))
world_with_population_updated.plot(column='log_population_2
                                   missing_kwds={'color': '
plt.title('World Map Colored by Logarithm of Population in
plt.show()
```



Data Analyst

Here is the updated world map, now colored according to the logarithm of the population for each country in 2024. The logarithmic scale provides a more nuanced view of population differences, especially when dealing with a wide range of population sizes. Darker shades indicate countries with higher population, and the light grey color represents countries for which data is not available or countries that were not matched in the dataset. This transformation allows for a clearer differentiation between countries with large and small populations. [>_]

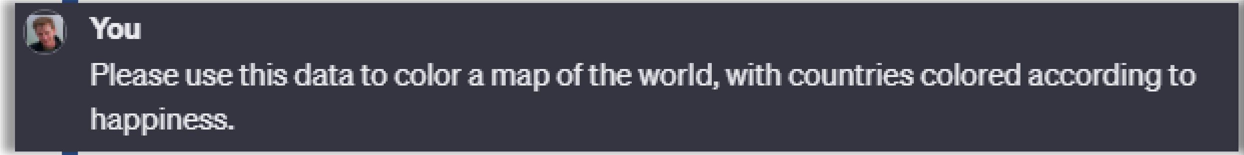
University of the Virgin Islands

www.uvi.edu

SPECIALIZING IN FUTURES

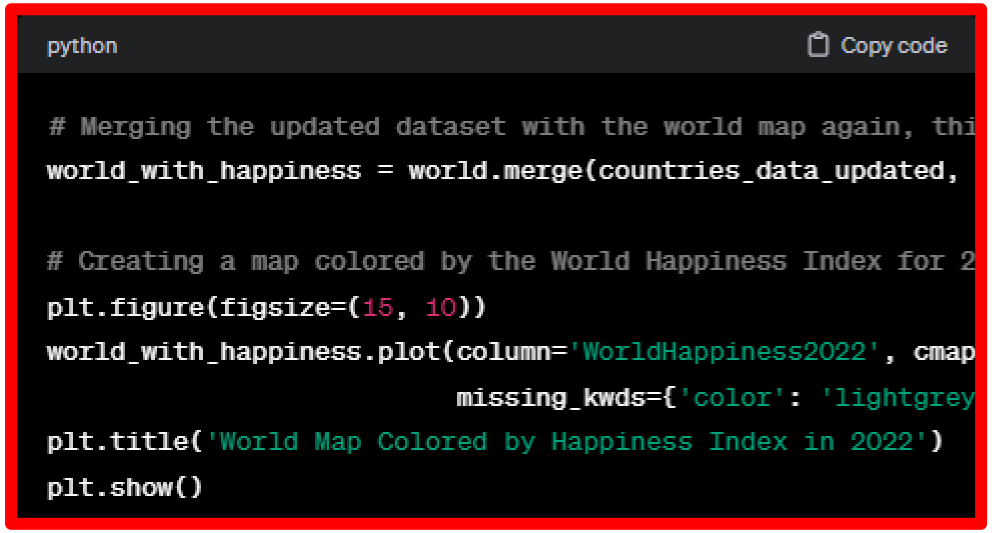
HISTORICALLY AMERICAN.
UNIQUELY CARIBBEAN.
GLOBALLY INTERACTIVE.

Leveraging ChatGPT for Data Analysis and Visualization

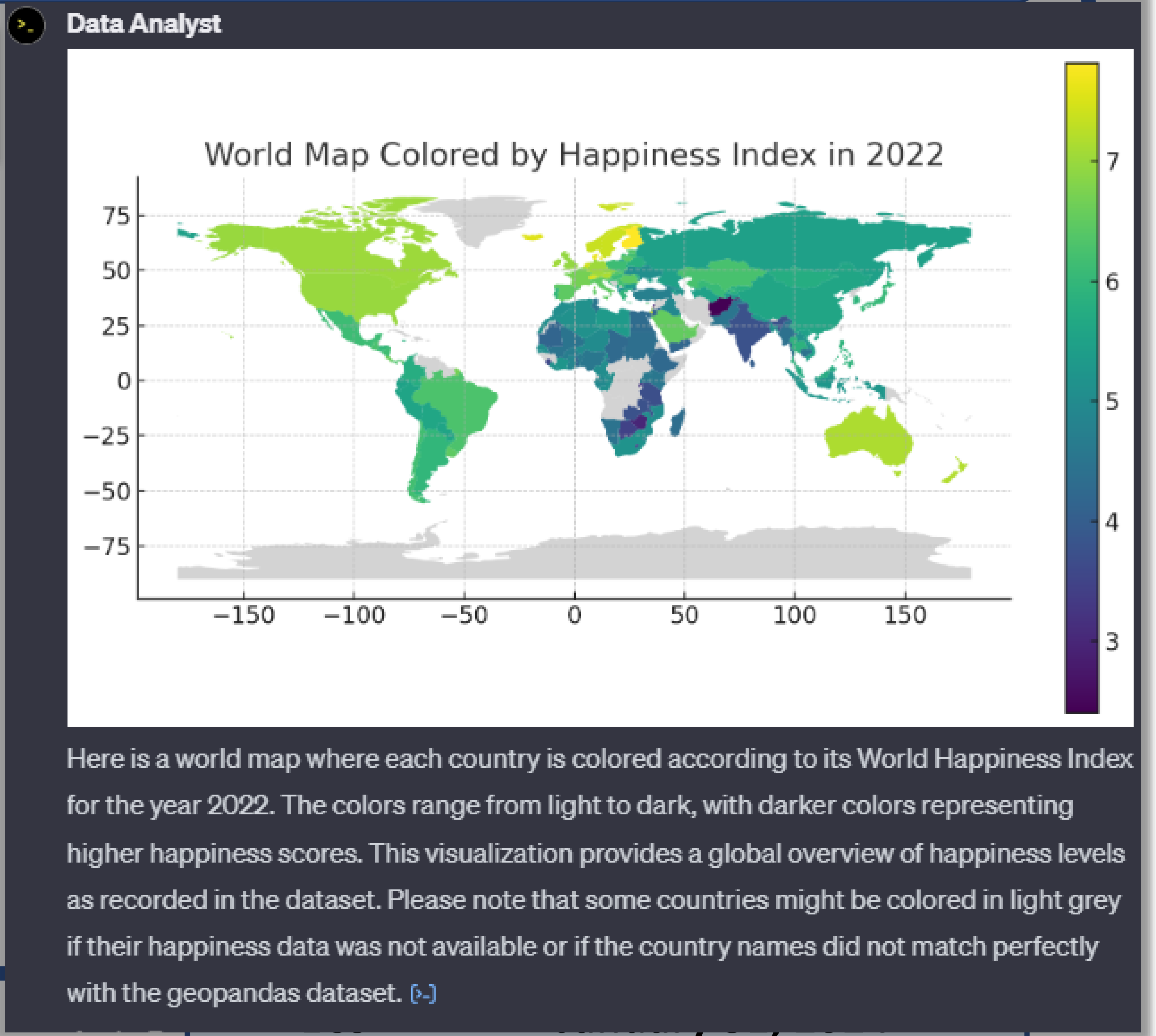


You

Please use this data to color a map of the world, with countries colored according to happiness.



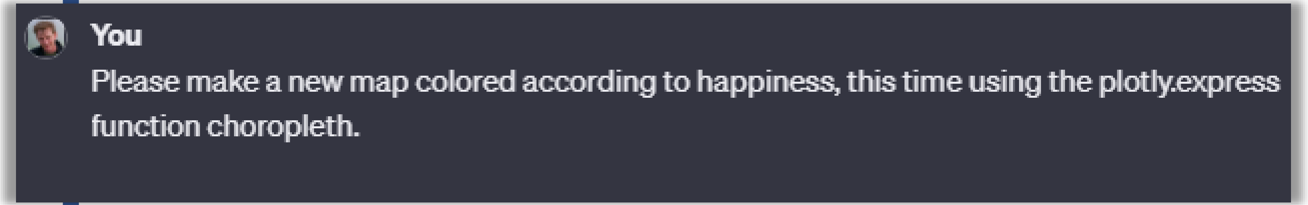
```python
# Merging the updated dataset with the world map again, thi
world_with_happiness = world.merge(countries_data_updated,

# Creating a map colored by the World Happiness Index for 2
plt.figure(figsize=(15, 10))
world_with_happiness.plot(column='WorldHappiness2022', cmap
                          missing_kwds={'color': 'lightgrey
plt.title('World Map Colored by Happiness Index in 2022')
plt.show()
```



Data Analyst

Here is a world map where each country is colored according to its World Happiness Index for the year 2022. The colors range from light to dark, with darker colors representing higher happiness scores. This visualization provides a global overview of happiness levels as recorded in the dataset. Please note that some countries might be colored in light grey if their happiness data was not available or if the country names did not match perfectly with the geopandas dataset.

Leveraging ChatGPT for Data Ar

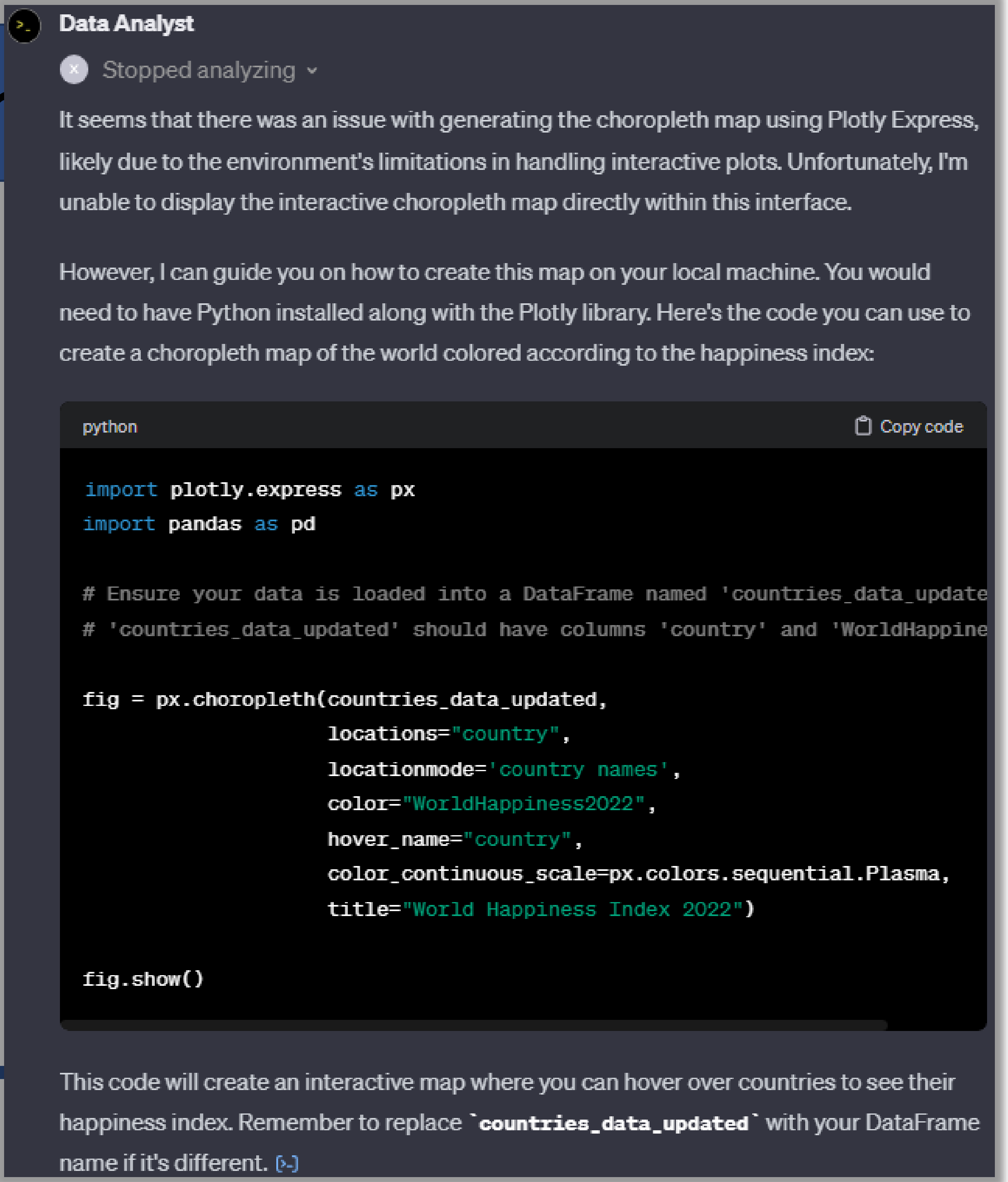


You

Please make a new map colored according to happiness, this time using the plotly.express function choropleth.

Chloropleth creates a dynamic figure, where moving your mouse across different areas will show extra information as a tooltip.

ChatGPT is disabled from creating dynamic figures….



Data Analyst

It seems that there was an issue with generating the choropleth map using Plotly Express, likely due to the environment's limitations in handling interactive plots. Unfortunately, I'm unable to display the interactive choropleth map directly within this interface.

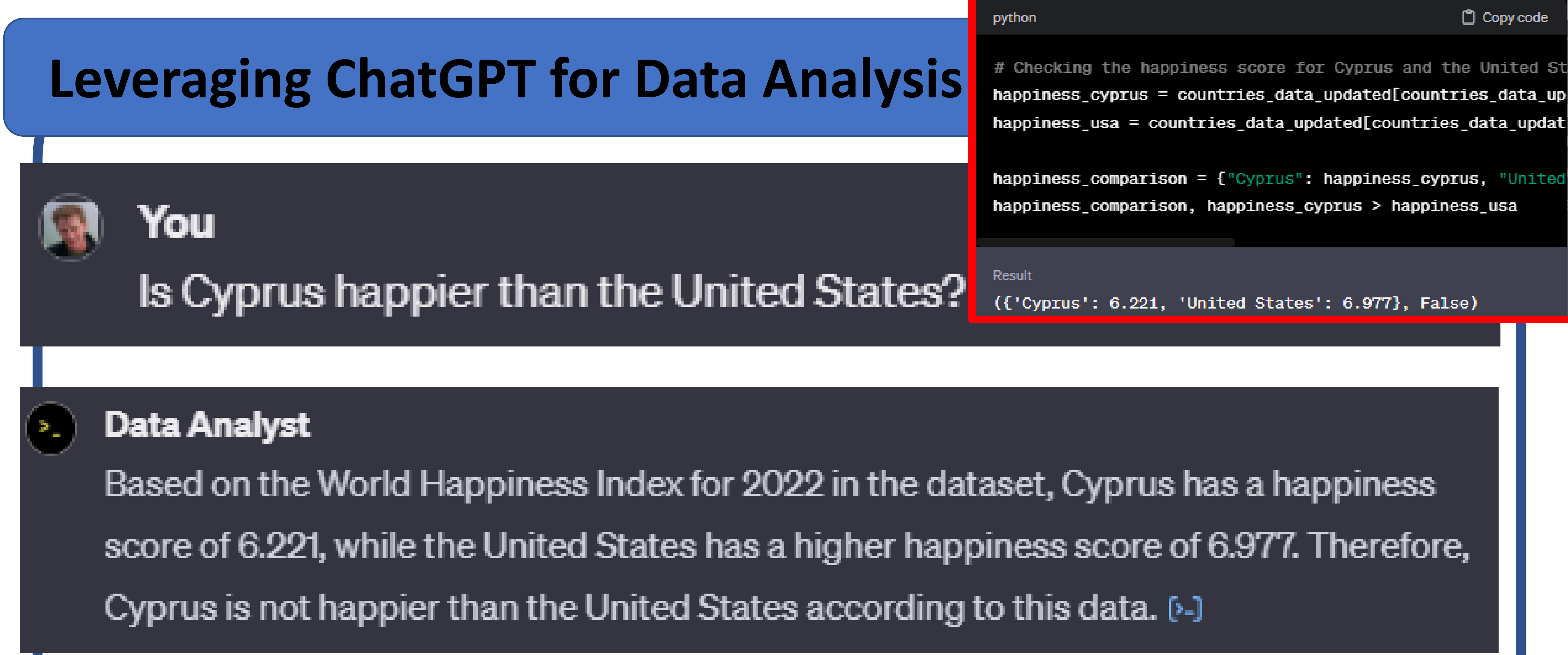
However, I can guide you on how to create this map on your local machine. You would need to have Python installed along with the Plotly library. Here's the code you can use to create a choropleth map of the world colored according to the happiness index:

```python
import plotly.express as px
import pandas as pd

# Ensure your data is loaded into a DataFrame named 'countries_data_update
# 'countries_data_updated' should have columns 'country' and 'WorldHappine

fig = px.choropleth(countries_data_updated,
                    locations="country",
                    locationmode='country names',
                    color="WorldHappiness2022",
                    hover_name="country",
                    color_continuous_scale=px.colors.sequential.Plasma,
                    title="World Happiness Index 2022")

fig.show()
```

This code will create an interactive map where you can hover over countries to see their happiness index. Remember to replace `countries_data_updated` with your DataFrame name if it's different.

University of the Virgin Islands
www.uvi.edu
SPECIALIZING IN FUTURES
HISTORICALLY AMERICAN.
UNIQUELY CARIBBEAN.
GLOBALLY INTERACTIVE.

Leveraging ChatGPT for Data Analysis



You

Is Cyprus happier than the United States?

Data Analyst

Based on the World Happiness Index for 2022 in the dataset, Cyprus has a happiness score of 6.221, while the United States has a higher happiness score of 6.977. Therefore, Cyprus is not happier than the United States according to this data.

```python
# Checking the happiness score for Cyprus and the United St
happiness_cyprus = countries_data_updated[countries_data_up
happiness_usa = countries_data_updated[countries_data_updat

happiness_comparison = {"Cyprus": happiness_cyprus, "United
happiness_comparison, happiness_cyprus > happiness_usa
```

Result

```
({'Cyprus': 6.221, 'United States': 6.977}, False)
```

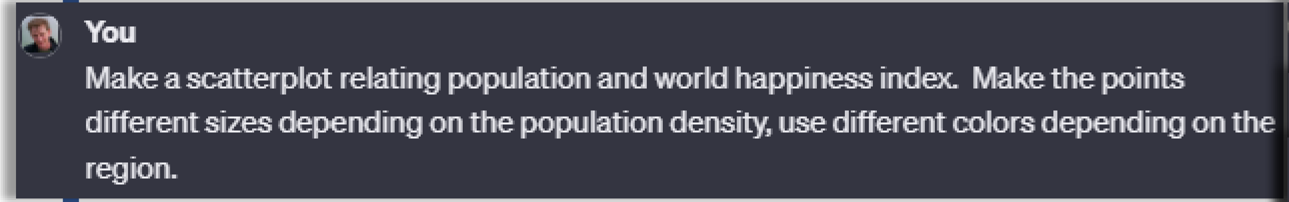
University of the Virgin Islands
www.uvi.edu
SPECIALIZING IN FUTURES
HISTORICALLY AMERICAN. UNIQUELY CARIBBEAN. GLOBALLY INTERACTIVE.

Dr. Benjamin Walter

ECS January 31, 2024

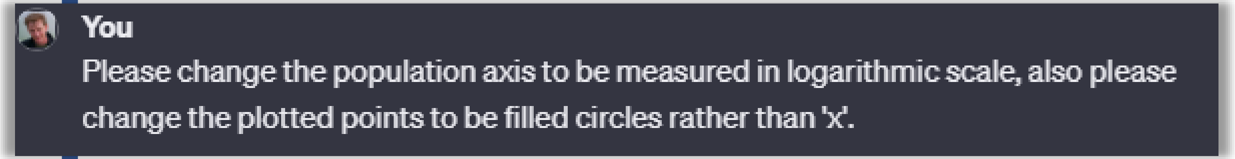
Leveraging ChatGPT for Data Analysis and Visualization

Let’s look at happiness…..



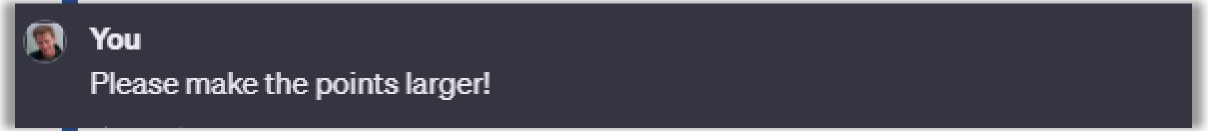
You

Make a scatterplot relating population and world happiness index. Make the points different sizes depending on the population density, use different colors depending on the region.



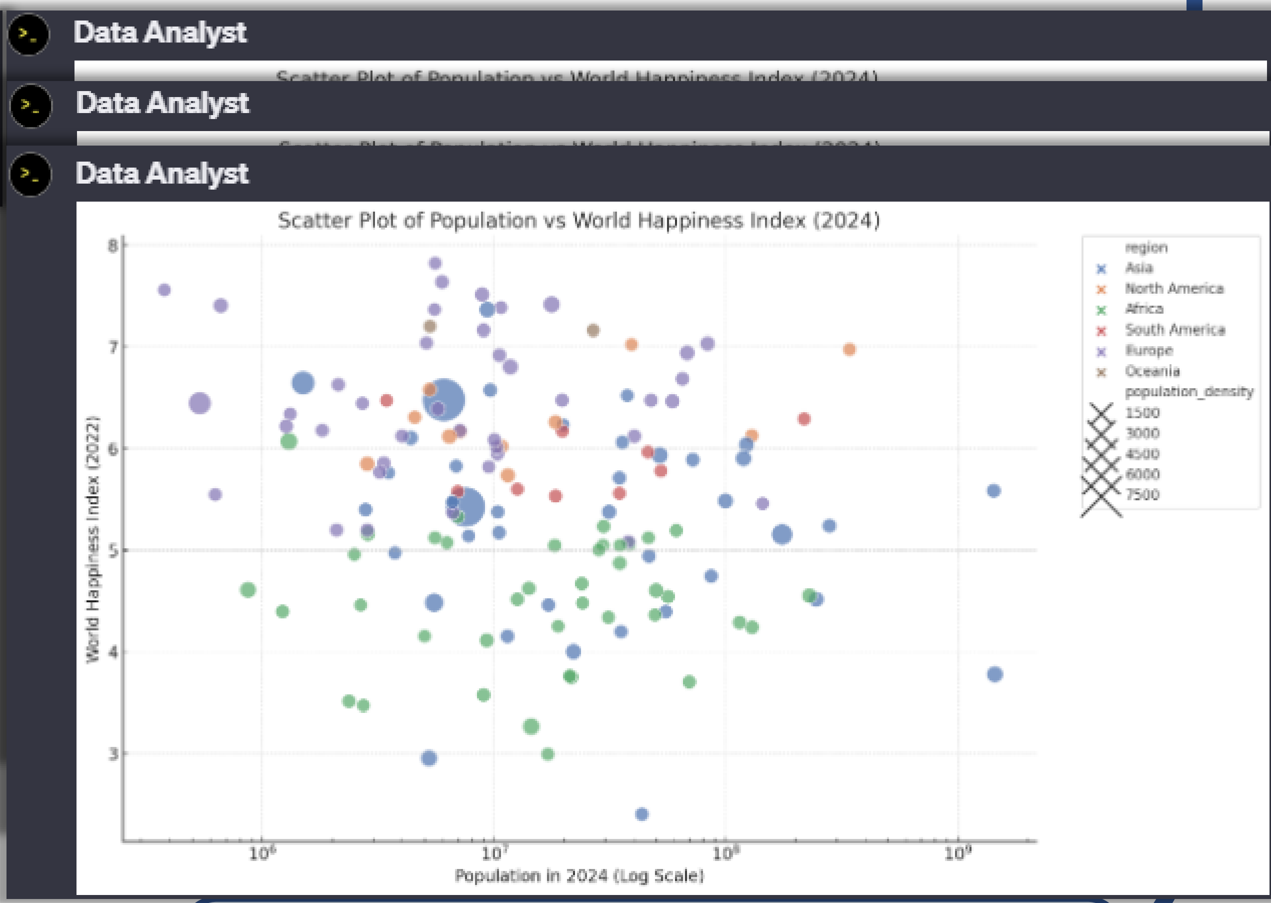
You

Please change the population axis to be measured in logarithmic scale, also please change the plotted points to be filled circles rather than 'x'.



You

Please make the points larger!



Data Analyst

Data Analyst

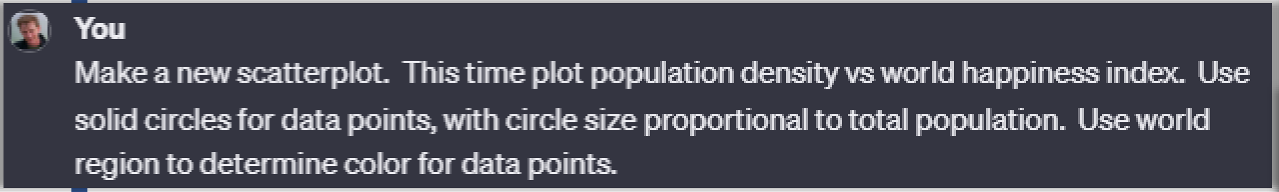
Data Analyst

University of the Virgin Islands
www.uvi.edu
SPECIALIZING IN FUTURES
HISTORICALLY AMERICAN. UNIQUELY CARIBBEAN. GLOBALLY INTERACTIVE.

Dr. Benjamin Walter

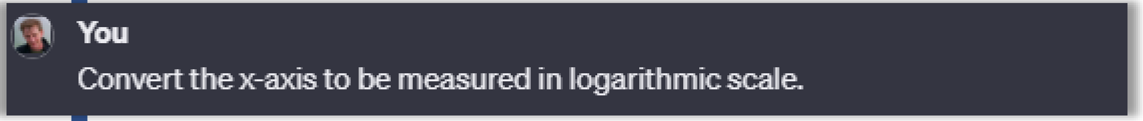
ECS **January 31, 2024**

Leveraging ChatGPT for Data Analysis and Visualization



You

Make a new scatterplot. This time plot population density vs world happiness index. Use solid circles for data points, with circle size proportional to total population. Use world region to determine color for data points.

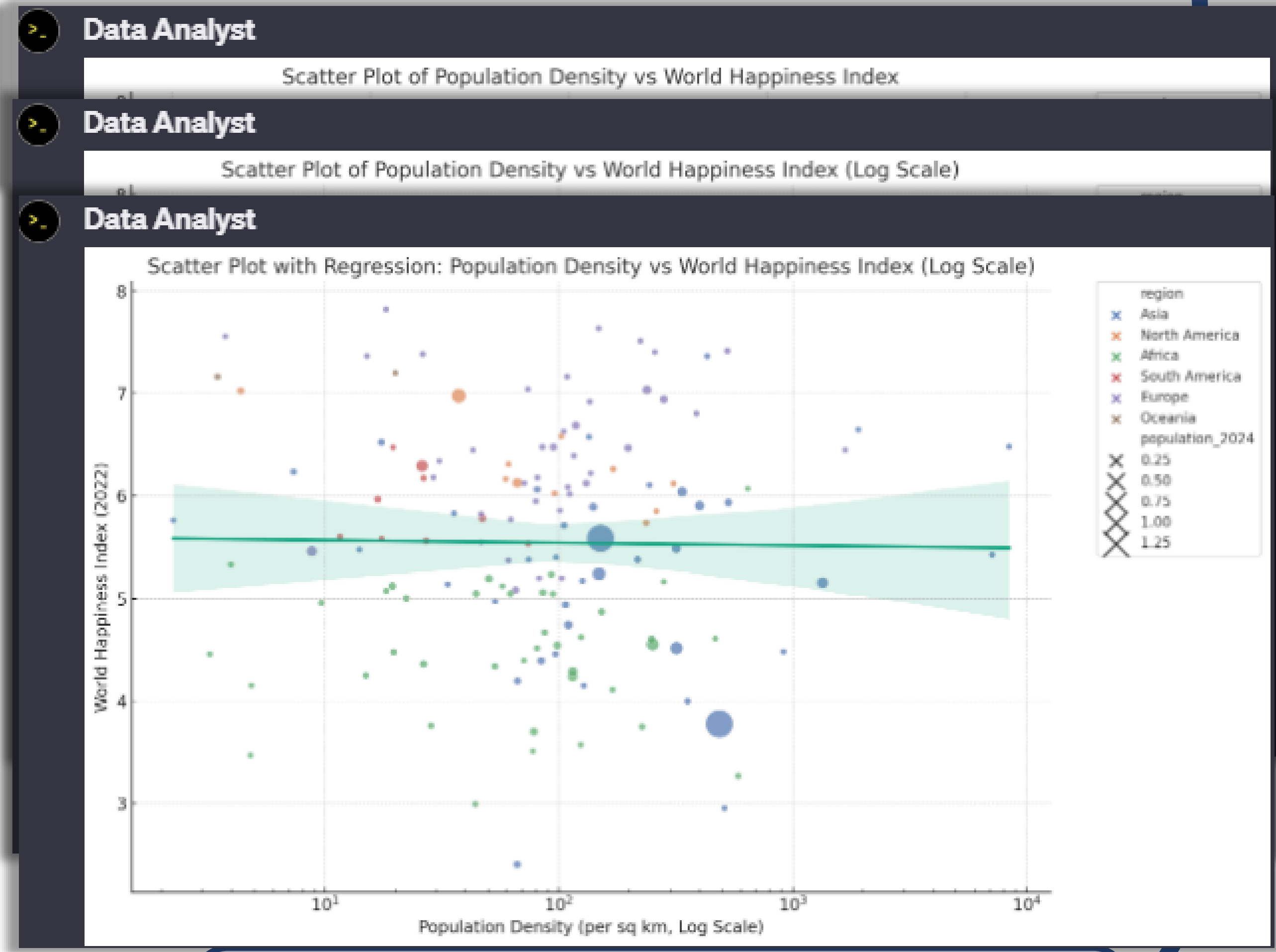


You

Convert the x-axis to be measured in logarithmic scale.

You

Include a regression line in the plot above.



Data Analyst

Scatter Plot of Population Density vs World Happiness Index

Data Analyst

Scatter Plot of Population Density vs World Happiness Index (Log Scale)

Data Analyst

Scatter Plot with Regression: Population Density vs World Happiness Index (Log Scale)

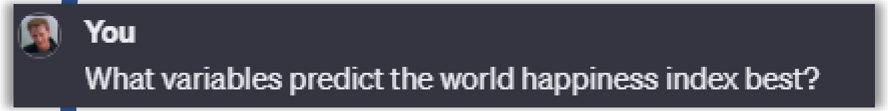
University of the Virgin Islands
www.uvi.edu
SPECIALIZING IN FUTURES
HISTORICALLY AMERICAN.
UNIQUELY CARIBBEAN.
GLOBALLY INTERACTIVE.

Dr. Benjamin Walter

ECS **January 31, 2024**

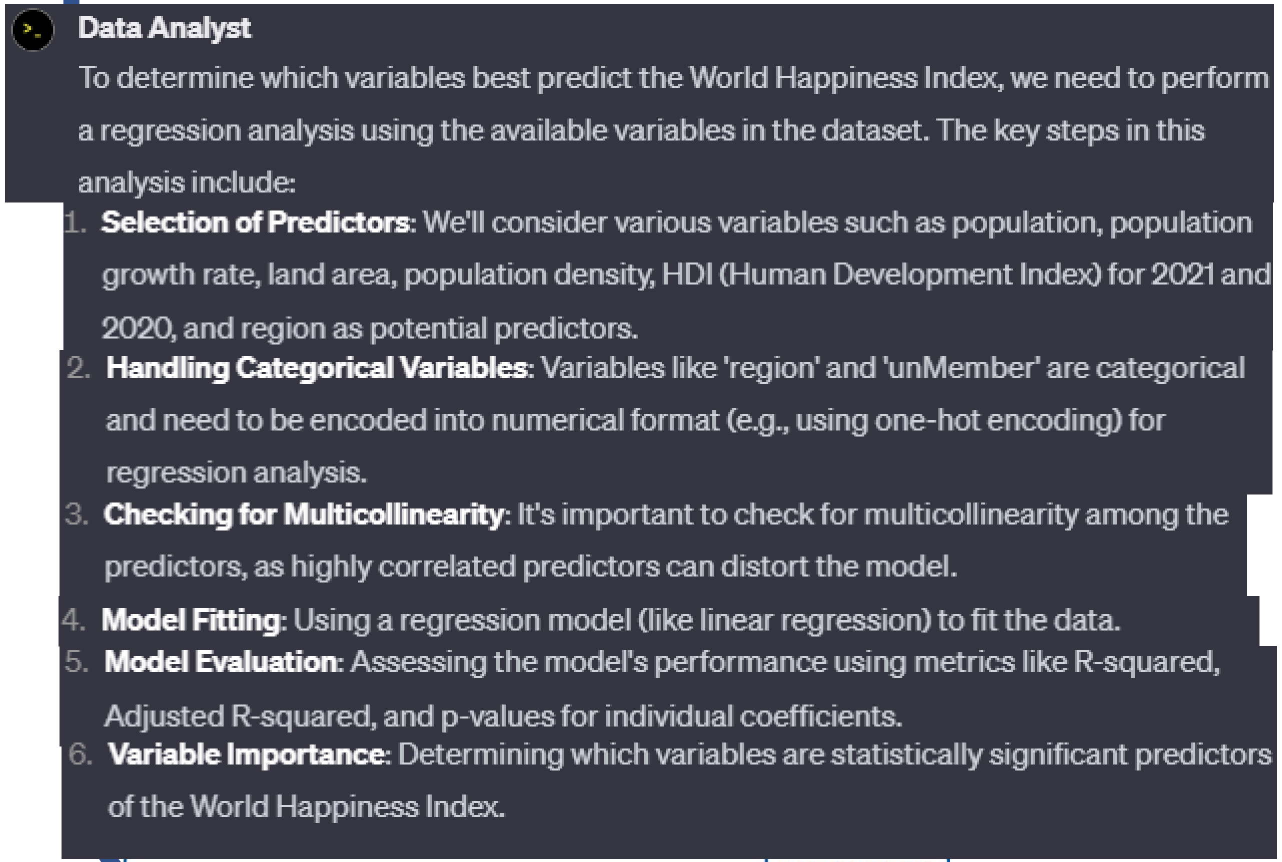
Leveraging ChatGPT for Data Analysis and Visualization

ChatGPT concluded that I am an idiot and needed to be taught basic statistics and experimental design….



You

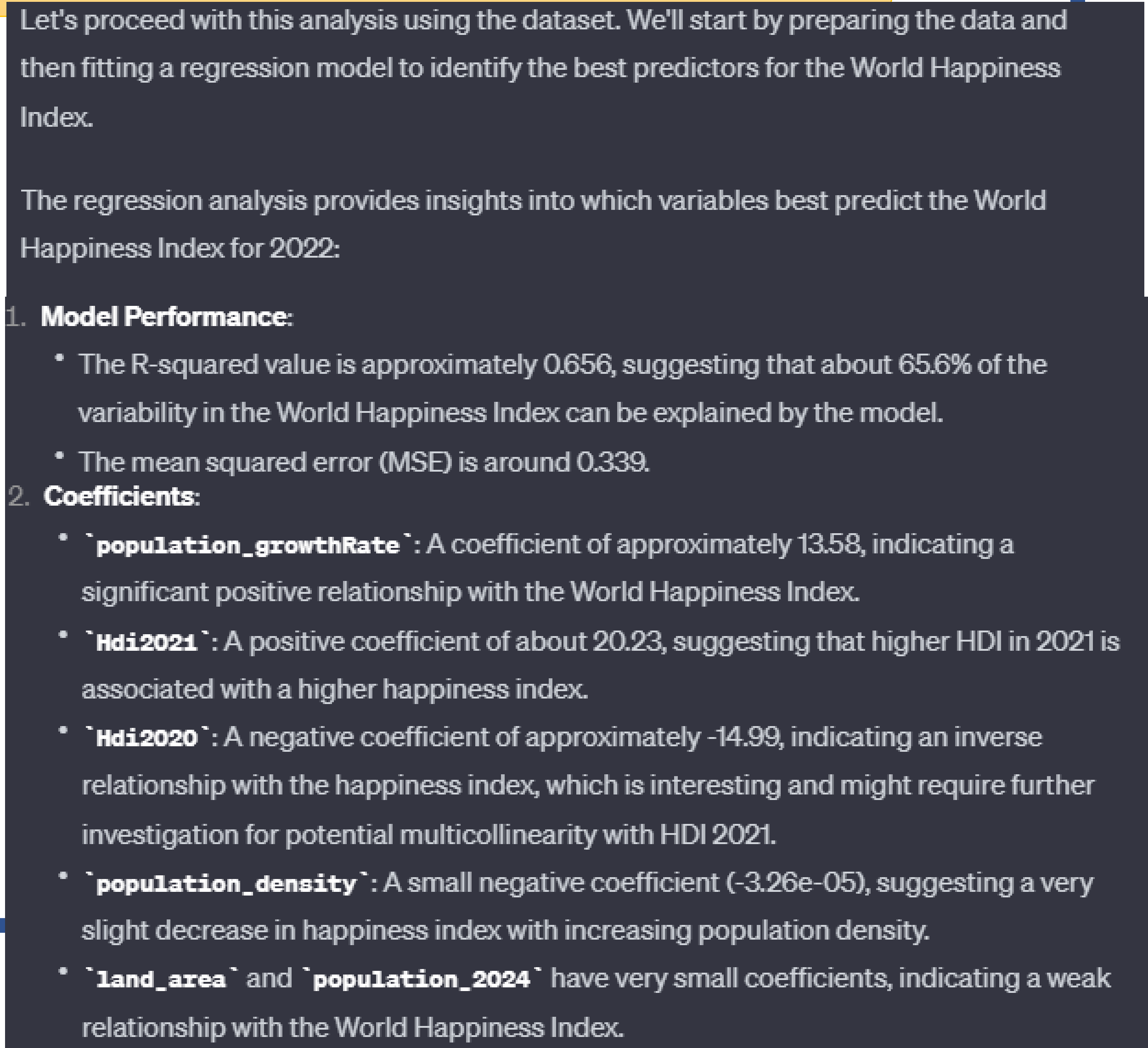
What variables predict the world happiness index best?



Data Analyst

To determine which variables best predict the World Happiness Index, we need to perform a regression analysis using the available variables in the dataset. The key steps in this analysis include:

1. **Selection of Predictors**: We'll consider various variables such as population, population growth rate, land area, population density, HDI (Human Development Index) for 2021 and 2020, and region as potential predictors.
2. **Handling Categorical Variables**: Variables like 'region' and 'unMember' are categorical and need to be encoded into numerical format (e.g., using one-hot encoding) for regression analysis.
3. **Checking for Multicollinearity**: It's important to check for multicollinearity among the predictors, as highly correlated predictors can distort the model.
4. **Model Fitting**: Using a regression model (like linear regression) to fit the data.
5. **Model Evaluation**: Assessing the model's performance using metrics like R-squared, Adjusted R-squared, and p-values for individual coefficients.
6. **Variable Importance**: Determining which variables are statistically significant predictors of the World Happiness Index.



Let's proceed with this analysis using the dataset. We'll start by preparing the data and then fitting a regression model to identify the best predictors for the World Happiness Index.

The regression analysis provides insights into which variables best predict the World Happiness Index for 2022:

1. **Model Performance**:
   - The R-squared value is approximately 0.656, suggesting that about 65.6% of the variability in the World Happiness Index can be explained by the model.
   - The mean squared error (MSE) is around 0.339.
2. **Coefficients**:
   - `population_growthRate`: A coefficient of approximately 13.58, indicating a significant positive relationship with the World Happiness Index.
   - `Hdi2021`: A positive coefficient of about 20.23, suggesting that higher HDI in 2021 is associated with a higher happiness index.
   - `Hdi2020`: A negative coefficient of approximately -14.99, indicating an inverse relationship with the happiness index, which is interesting and might require further investigation for potential multicollinearity with HDI 2021.
   - `population_density`: A small negative coefficient (-3.26e-05), suggesting a very slight decrease in happiness index with increasing population density.
   - `land_area` and `population_2024` have very small coefficients, indicating a weak relationship with the World Happiness Index.

University of the Virgin Islands

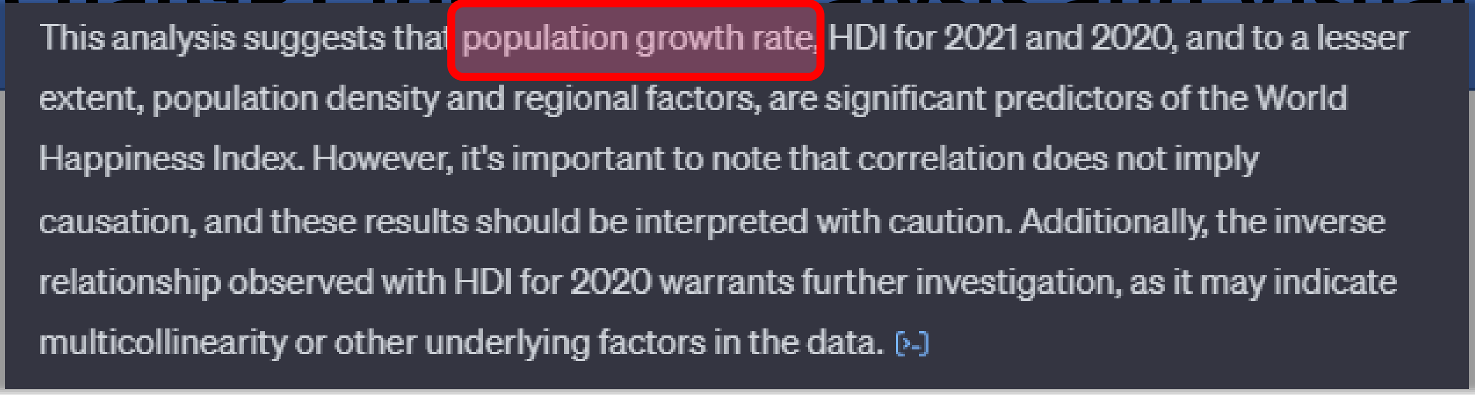
www.uvi.edu

SPECIALIZING IN FUTURES

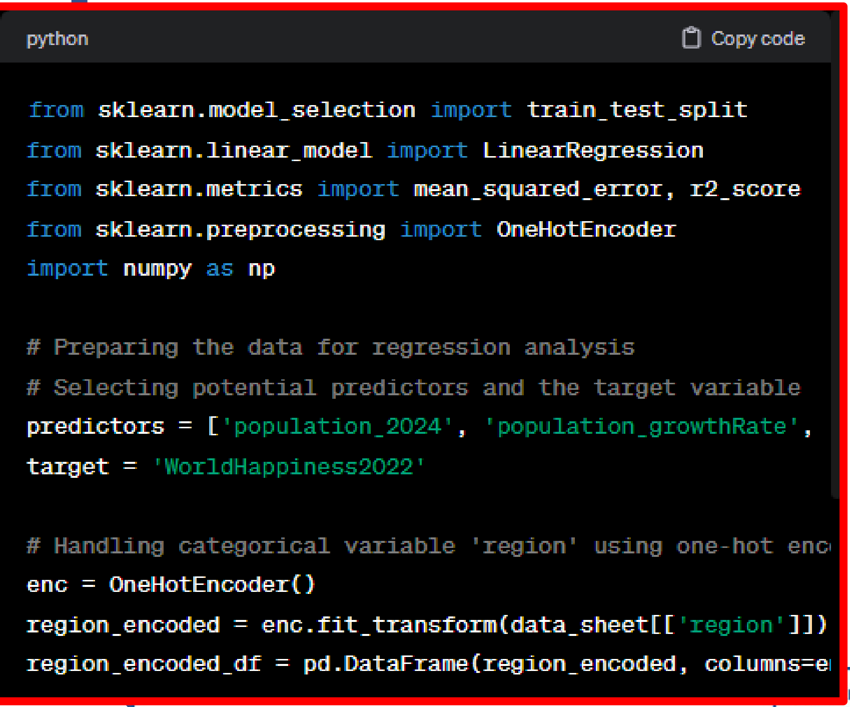
HISTORICALLY AMERICAN. UNIQUELY CARIBBEAN. GLOBALLY INTERACTIVE.

Leveraging ChatGPT for Data Analysis and Visualization

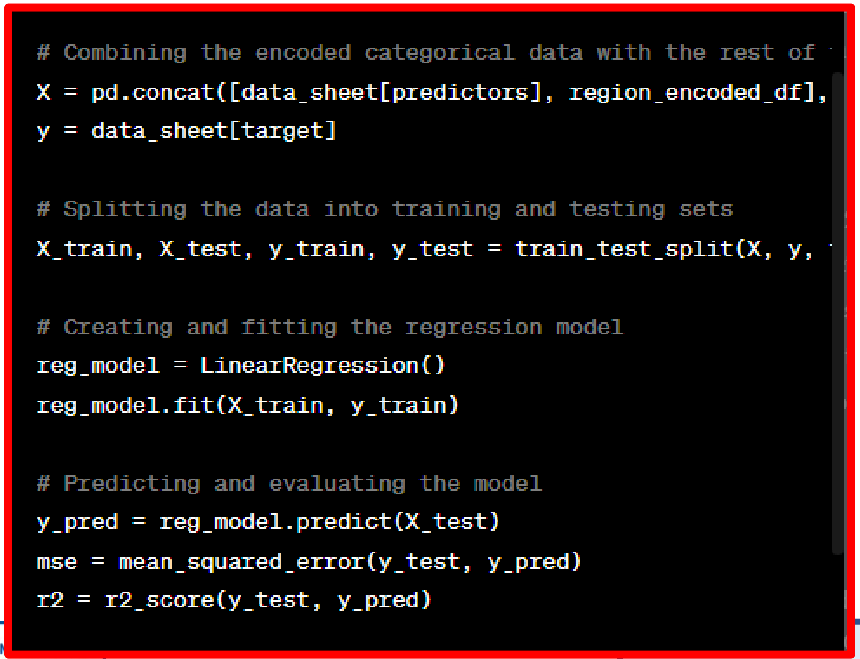
Blah blah blah



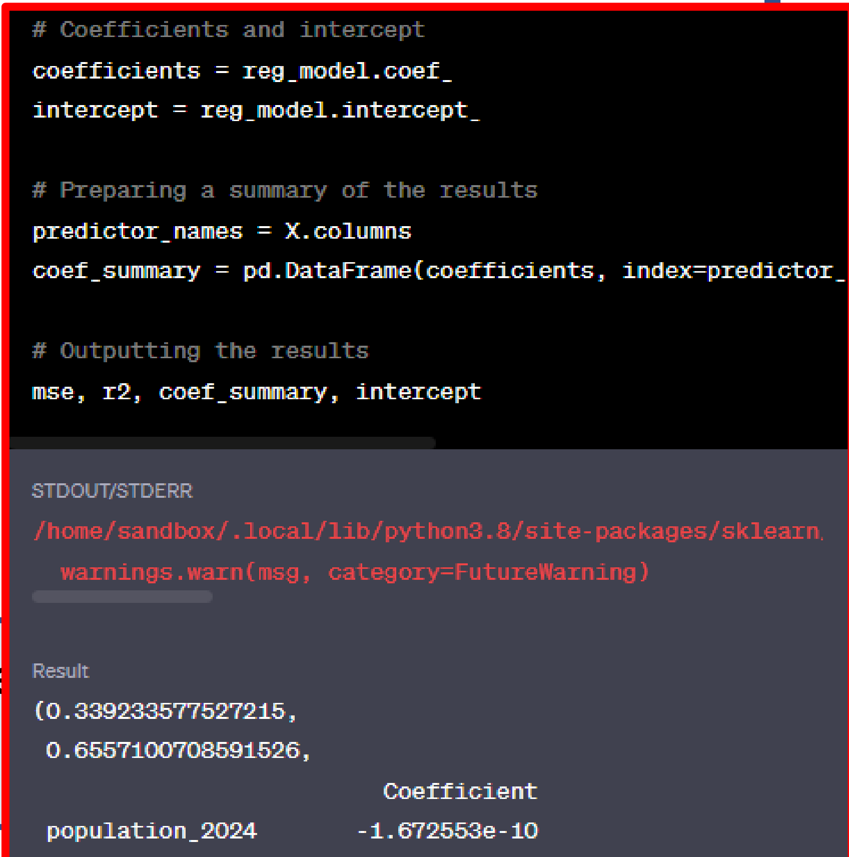
This analysis suggests that population growth rate, HDI for 2021 and 2020, and to a lesser extent, population density and regional factors, are significant predictors of the World Happiness Index. However, it's important to note that correlation does not imply causation, and these results should be interpreted with caution. Additionally, the inverse relationship observed with HDI for 2020 warrants further investigation, as it may indicate multicollinearity or other underlying factors in the data.



```python
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.preprocessing import OneHotEncoder
import numpy as np

# Preparing the data for regression analysis
# Selecting potential predictors and the target variable
predictors = ['population_2024', 'population_growthRate',
target = 'WorldHappiness2022'

# Handling categorical variable 'region' using one-hot enc
enc = OneHotEncoder()
region_encoded = enc.fit_transform(data_sheet[['region']])
region_encoded_df = pd.DataFrame(region_encoded, columns=e
```



```python
# Combining the encoded categorical data with the rest of
X = pd.concat([data_sheet[predictors], region_encoded_df],
y = data_sheet[target]

# Splitting the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y,

# Creating and fitting the regression model
reg_model = LinearRegression()
reg_model.fit(X_train, y_train)

# Predicting and evaluating the model
y_pred = reg_model.predict(X_test)
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
```



```python
# Coefficients and intercept
coefficients = reg_model.coef_
intercept = reg_model.intercept_

# Preparing a summary of the results
predictor_names = X.columns
coef_summary = pd.DataFrame(coefficients, index=predictor_

# Outputting the results
mse, r2, coef_summary, intercept
```

STDOUT/STDERR

```
/home/sandbox/.local/lib/python3.8/site-packages/sklearn
  warnings.warn(msg, category=FutureWarning)
```

Result

```
(0.339233577527215,
 0.6557100708591526,
                        Coefficient
 population_2024       -1.672553e-10
```

University of the Virgin Islands

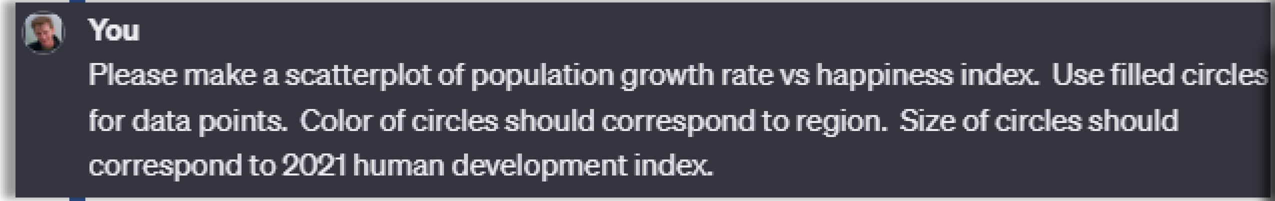
www.uvi.edu

HISTORICALLY AMERICAN.
UNIQUELY CARIBBEAN.
GLOBALLY INTERACTIVE.

Dr. Be

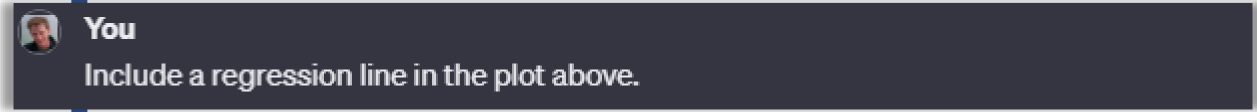
ECS

Leveraging ChatGPT for Data Analysis and Visualization



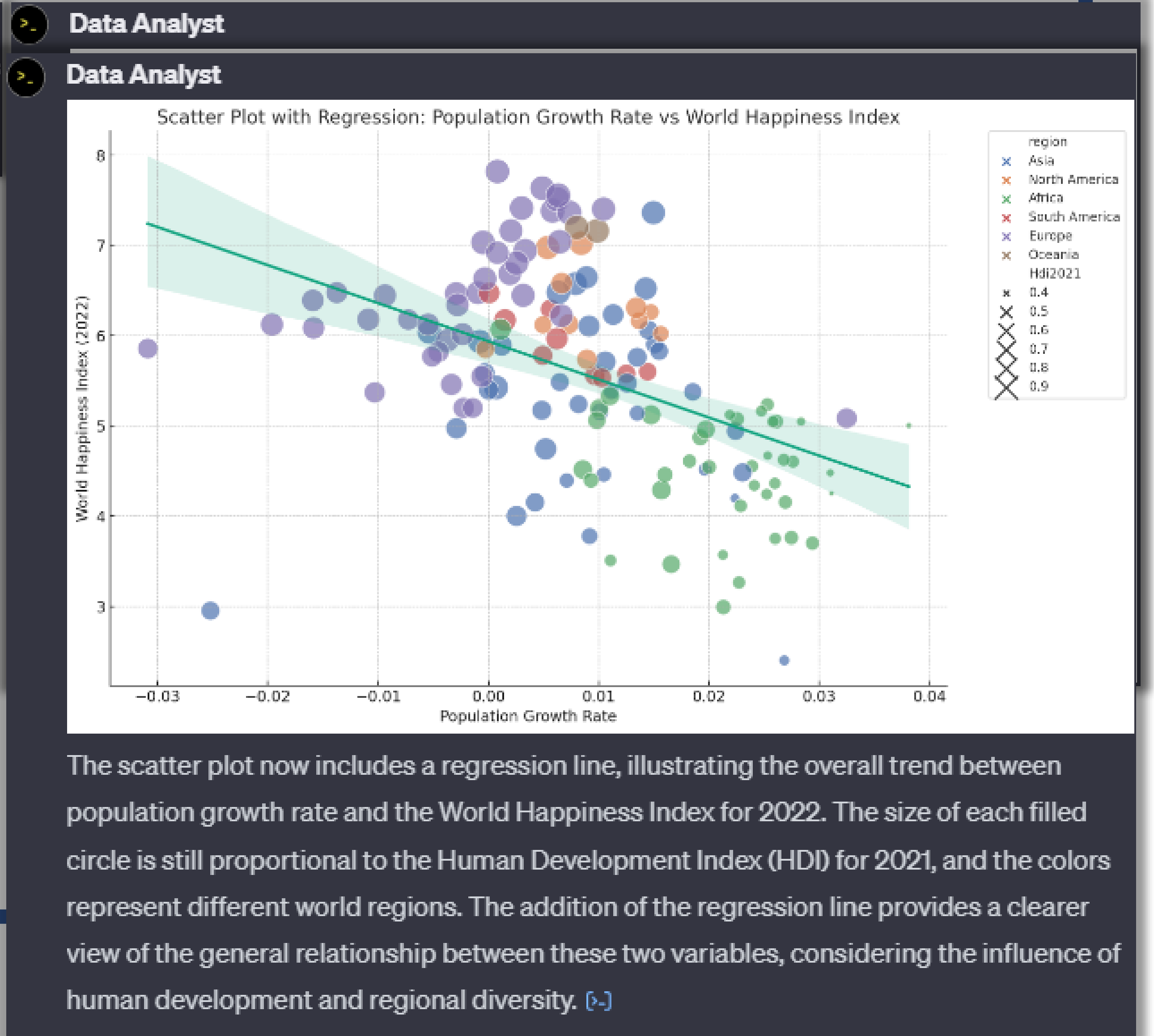
You

Please make a scatterplot of population growth rate vs happiness index. Use filled circles for data points. Color of circles should correspond to region. Size of circles should correspond to 2021 human development index.



You

Include a regression line in the plot above.



Data Analyst

Data Analyst

The scatter plot now includes a regression line, illustrating the overall trend between population growth rate and the World Happiness Index for 2022. The size of each filled circle is still proportional to the Human Development Index (HDI) for 2021, and the colors represent different world regions. The addition of the regression line provides a clearer view of the general relationship between these two variables, considering the influence of human development and regional diversity. [>_]

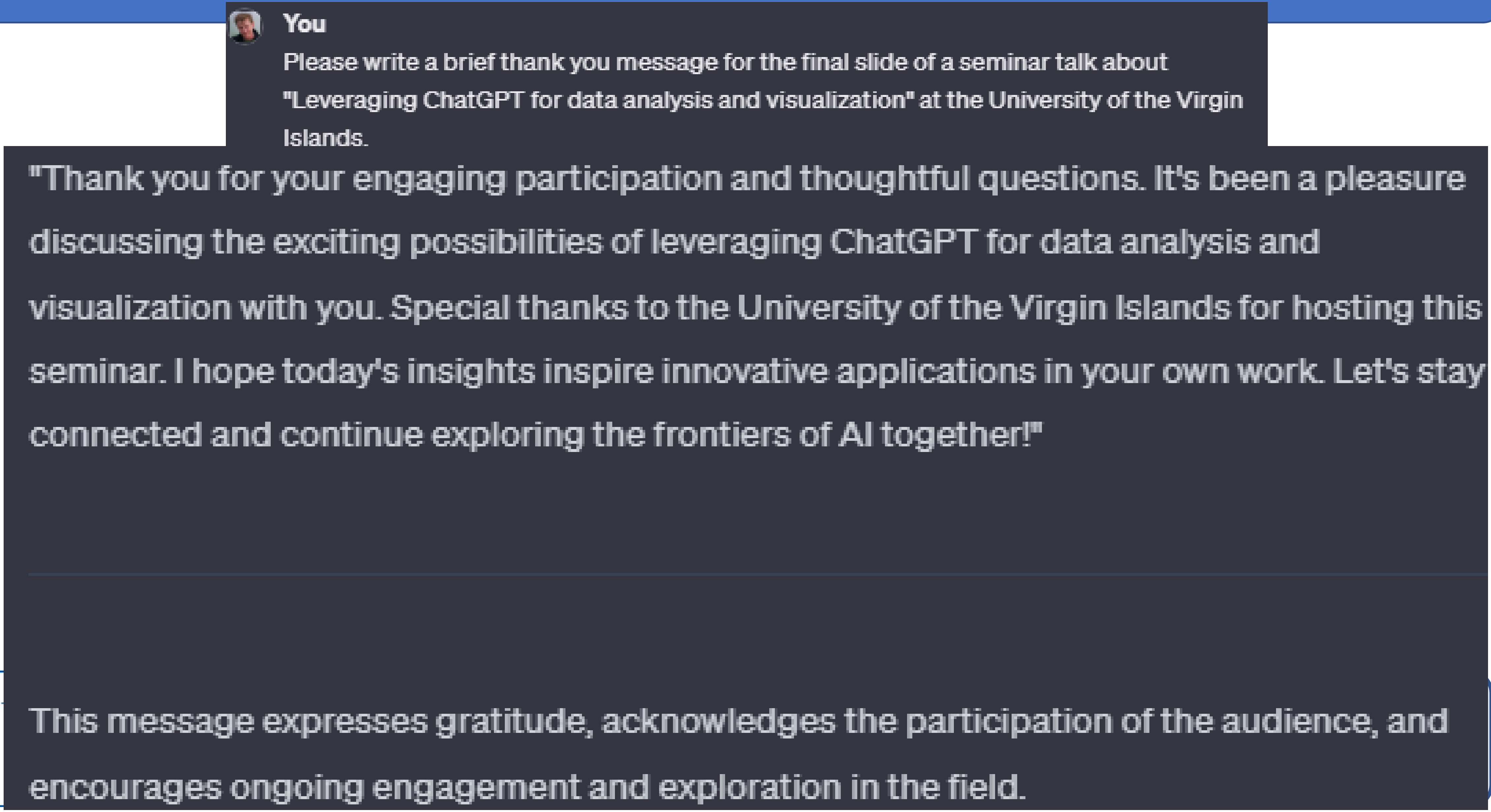
University of the Virgin Islands

www.uvi.edu

SPECIALIZING IN FUTURES

HISTORICALLY AMERICAN.
UNIQUELY CARIBBEAN.
GLOBALLY INTERACTIVE.

Leveraging ChatGPT for Data Analysis and Visualization



You

Please write a brief thank you message for the final slide of a seminar talk about "Leveraging ChatGPT for data analysis and visualization" at the University of the Virgin Islands.

"Thank you for your engaging participation and thoughtful questions. It's been a pleasure discussing the exciting possibilities of leveraging ChatGPT for data analysis and visualization with you. Special thanks to the University of the Virgin Islands for hosting this seminar. I hope today's insights inspire innovative applications in your own work. Let's stay connected and continue exploring the frontiers of AI together!"

This message expresses gratitude, acknowledges the participation of the audience, and encourages ongoing engagement and exploration in the field.

Leveraging ChatGPT for Data Analysis and Visualization

University of the Virgin Islands
www.uvi.edu

SPECIALIZING IN FUTURES
HISTORICALLY AMERICAN.
UNIQUELY CARIBBEAN.
GLOBALLY INTERACTIVE.

Dr. Benjamin Walter

ECS **January 31, 2024**