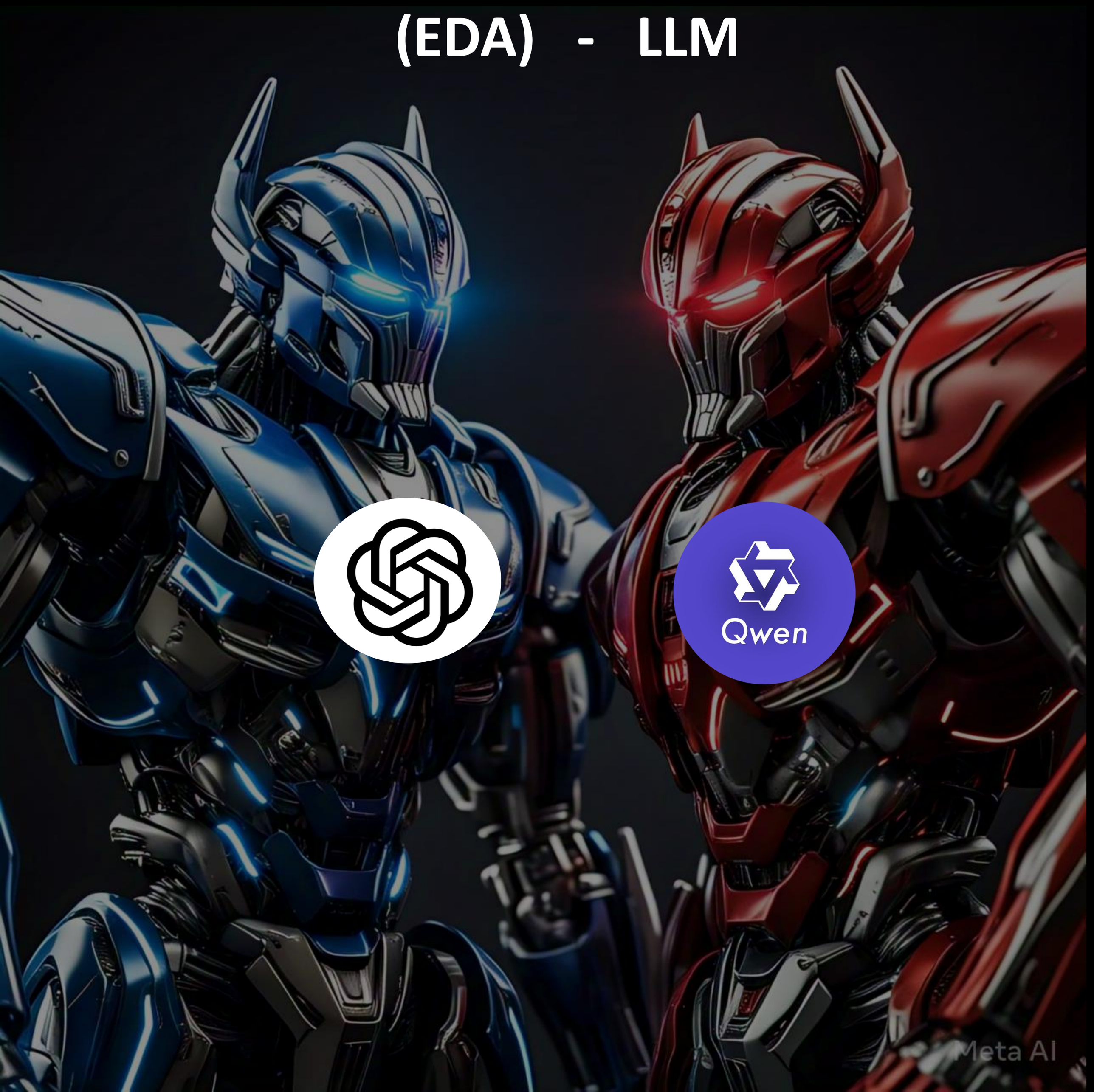
Data Analysis

Exploratory Data Analysis (EDA) - LLM



Problem

Problem description

Context

The original dataset contains 1000 records with 20 categorical attributes prepared by Prof. Hofmann.

In this dataset, each record represents a person taking out a loan from a bank.

Each person is classified as **good** or **bad** credit risk based on the set of attributes.

Link to the original dataset at

UCI Machine Learning

Content

The selected attributes:

- **Age** (numeric)
- **Sex** (text: male, female)
- **Job** (numeric: 0 - unskilled and non-resident, 1 - unskilled and resident, 2 - skilled, 3 - highly skilled)
- **Housing** (text: own, rent, or free)
- **Saving accounts** (text - little, moderate, quite rich, rich)
- **Checking account** (numeric, in DM - Deutsch Mark)
- **Credit amount** (numeric, in DM)
- **Duration** (numeric, in month)
- **Purpose**(text: car, furniture/equipment, radio/TV, domestic appliances, repairs, education, business, vacation/others
- **Risk** (Value target - Good or Bad Risk)

Objective

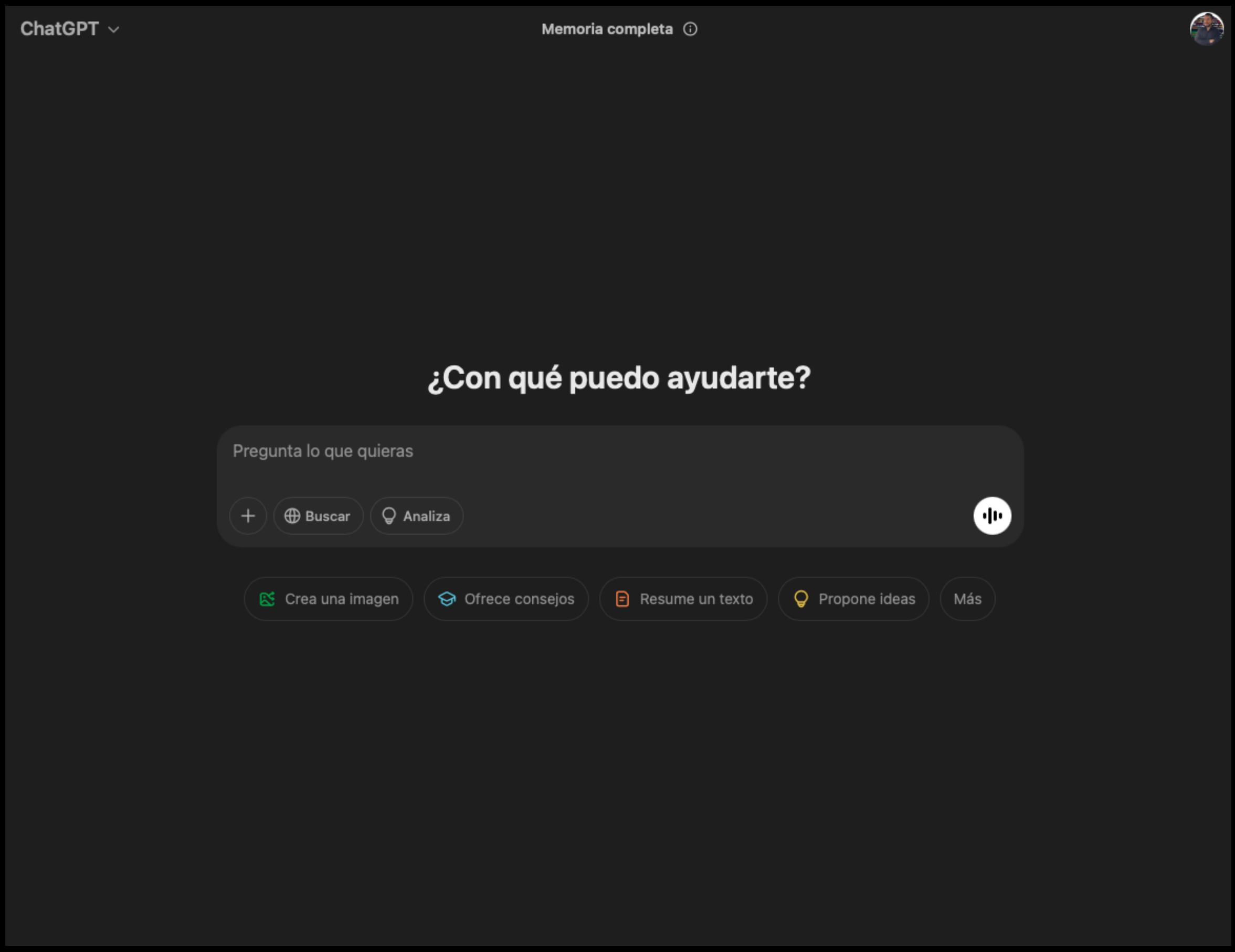
Train a model to predict from new data whether a person applying for a loan represents a good or bad risk

PROMPT

Develop a Python program that performs the following exploratory data analysis (EDA) tasks on a CSV file provided by user german_credit_data.csv:
Reading CSV file: Load the CSV file into a DataFrame using the pandas library.
Initial display: Displays the first 5 rows of the dataset.
Dataset overview: Provides an informative summary of the dataset, including the number of rows and columns, the data types of each column, and the number of non-null values.
Statistical description: Generates a statistical summary of the dataset, including metrics such as mean, median, standard deviation, minimum, maximum, and quartiles for numeric columns.
Correlation matrix: Calculates the correlation matrix for numeric variables in the dataset.
Heatmap: Visualize the correlation matrix using a heatmap with the seaborn or matplotlib library, making sure that the correlation values are readable and the plot is well labeled.
The program should be modular, easy to understand, and well-commented. It should also include error handling for cases such as a missing CSV file, non-numeric columns in the correlation matrix, or formatting issues in the file.

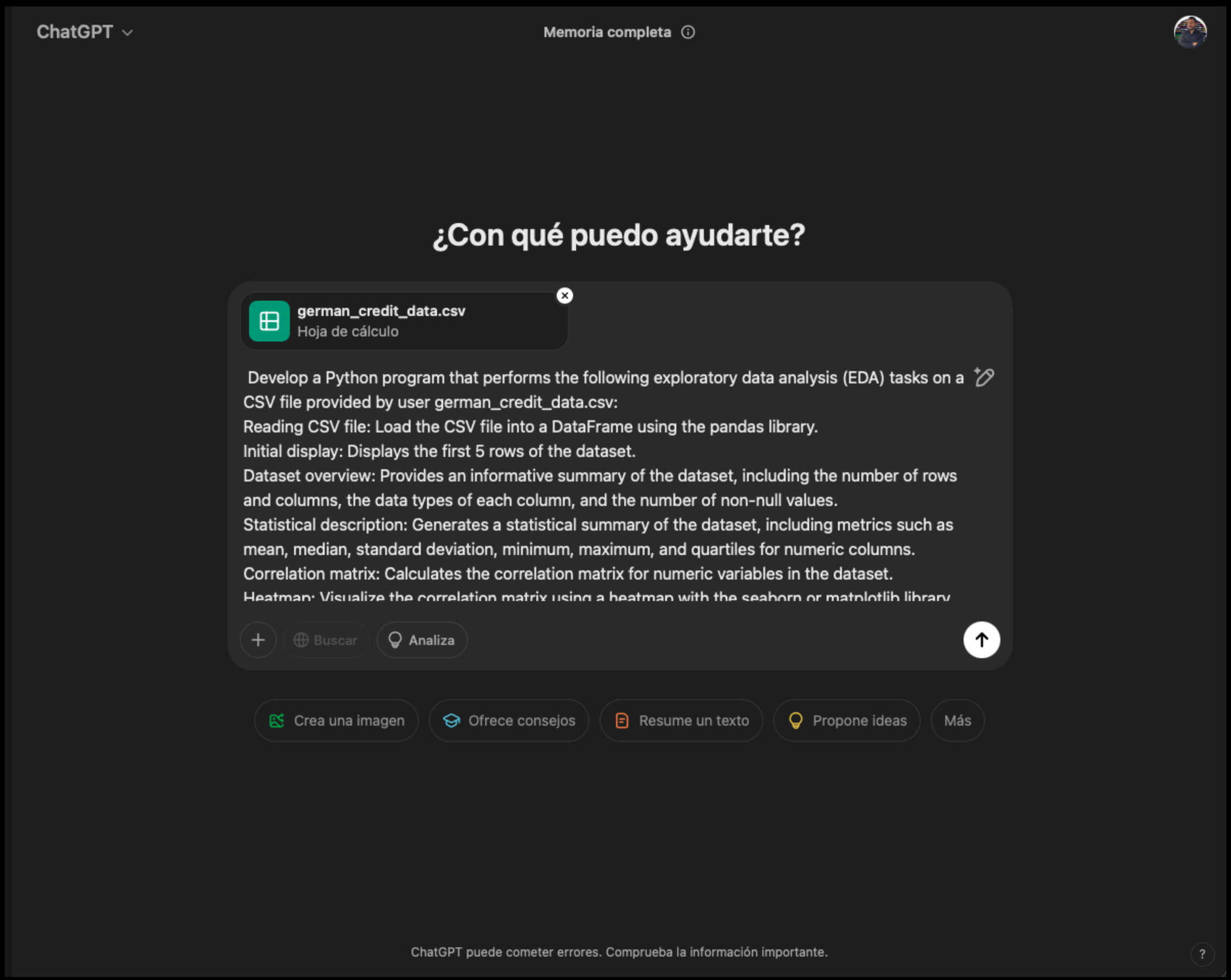
ChatGPT

Go to ChatGPT website



ChatGPT

Introduce a prompt to create EDA in python



ChatGPT

See result of LLM

ChatGPT Memoria completa Compartir

columns.
Correlation matrix: Calculates the correlation matrix for numeric variables in the dataset.
Heatmap: Visualize the correlation matrix using a heatmap with the seaborn or matplotlib library, making sure that the correlation values are readable and the plot is well labeled.
The program should be modular, easy to understand, and well-commented. It should also include error handling for cases such as a missing CSV file, non-numeric columns in the correlation matrix, or formatting issues in the file.

I'll develop a Python program that performs the requested Exploratory Data Analysis (EDA) tasks on the uploaded `german_credit_data.csv` file. The program will include modular functions, error handling, and clear comments. Here's the plan:

1. **Read the CSV file** and load it into a Pandas DataFrame.
2. **Display the first 5 rows** of the dataset.
3. **Provide a dataset overview** including shape, data types, and non-null values.
4. **Generate a statistical summary** for numeric columns.
5. **Calculate the correlation matrix** for numeric columns.
6. **Visualize the correlation matrix** using a heatmap.

I'll now write and execute the Python script.

It looks like there was an issue accessing the file. I'll retry loading the CSV file to see if we can proceed with the analysis.

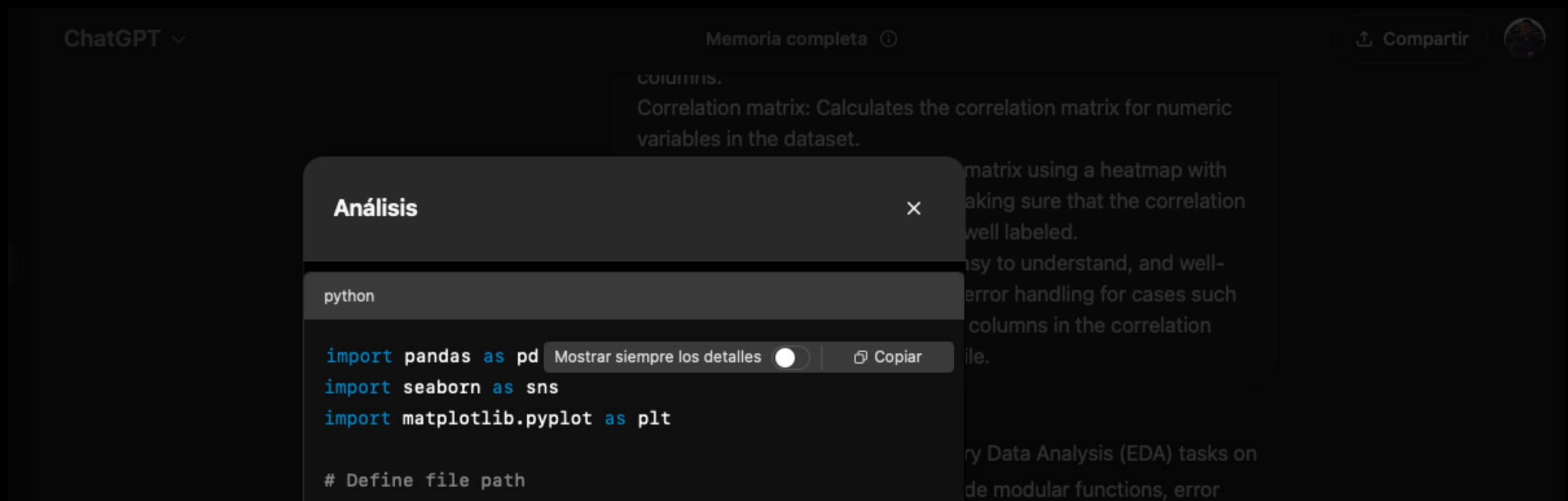
Pregunta lo que quieras

Buscar Analiza

ChatGPT puede cometer errores. Comprueba la información importante.

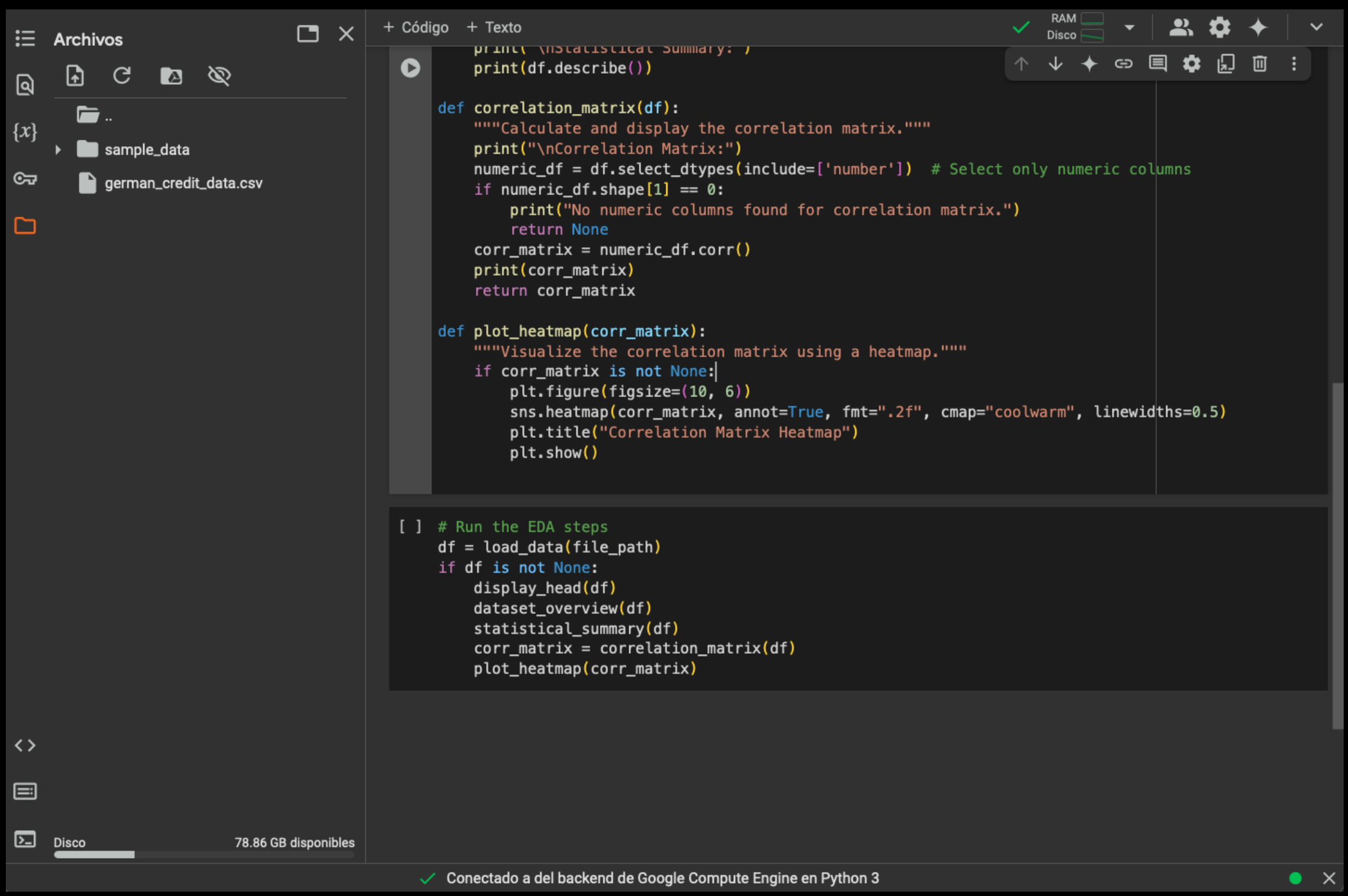
ChatGPT

Copy Python code



ChatGPT+ Colab

Go to Colab, paste code & check results



ChatGPT+ Colab

Go to Colab, paste code & check results

EDA_ChatGPT.ipynb

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda

Compartir

Archivos

+ Código + Texto

RAM Disco

```
# Run the EDA steps
df = load_data(file_path)
if df is not None:
    display_head(df)
    dataset_overview(df)
    statistical_summary(df)
    corr_matrix = correlation_matrix(df)
    plot_heatmap(corr_matrix)
```

```
File successfully loaded!

First 5 rows of the dataset:
   Unnamed: 0  age     sex  job housing saving_accounts checking account  \
0           0   67    male    2     own             NaN           little
1           1   22  female    2     own          little         moderate
2           2   49    male    1     own          little              NaN
3           3   45    male    2    free          little           little
4           4   53    male    2    free          little           little

   credit amount  duration              purpose  risk
0           1169         6             radio/TV  good
1           5951        48             radio/TV   bad
2           2096        12            education  good
3           7882        42  furniture/equipment  good
4           4870        24                  car   bad

Dataset Overview:
Number of Rows: 1000, Number of Columns: 11

Column Data Types and Non-null Counts:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Unnamed: 0       1000 non-null   int64
 1   age              1000 non-null   int64
 2   sex              1000 non-null   object
```

sample_data

german_credit_data.csv

Disco 78.59 GB disponibles

0 s se ejecutó 8:57 a.m.

ChatGPT+ Colab

Go to Colab, paste code & check results

EDA_ChatGPT.ipynb

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda

Compartir

+ Código + Texto

RAM Disco

Archivos

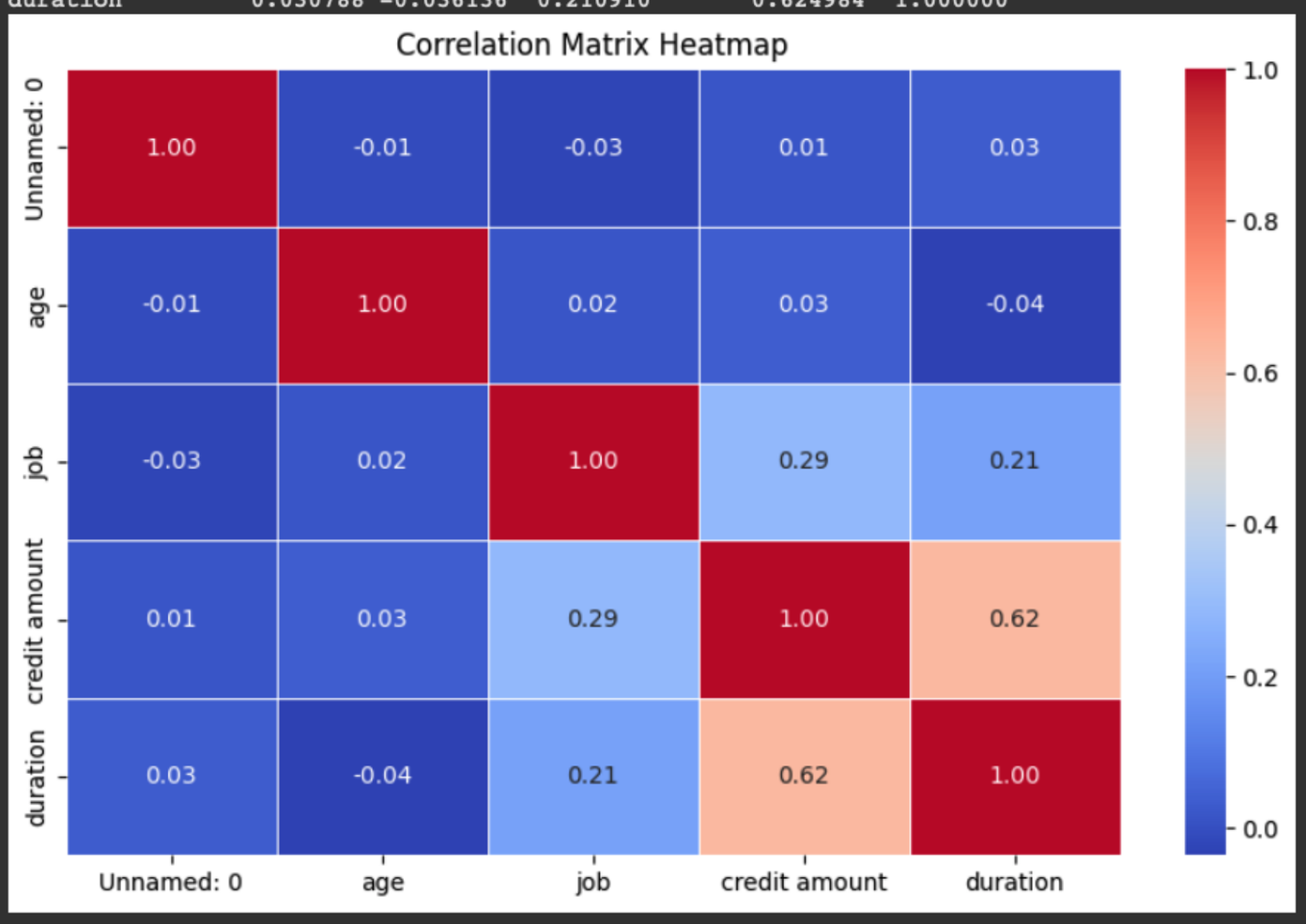
..

sample_data

german_credit_data.csv

```
75%      749.250000    42.000000     2.000000    3972.250000    24.000000
max      999.000000    75.000000     3.000000   18424.000000    72.000000

Correlation Matrix:
               Unnamed: 0       age       job  credit amount  duration
Unnamed: 0       1.000000 -0.010096 -0.027345       0.013488  0.030788
age             -0.010096  1.000000  0.015673       0.032716 -0.036136
job             -0.027345  0.015673  1.000000       0.285385  0.210910
credit amount    0.013488  0.032716  0.285385       1.000000  0.624984
duration         0.030788 -0.036136  0.210910       0.624984  1.000000
```

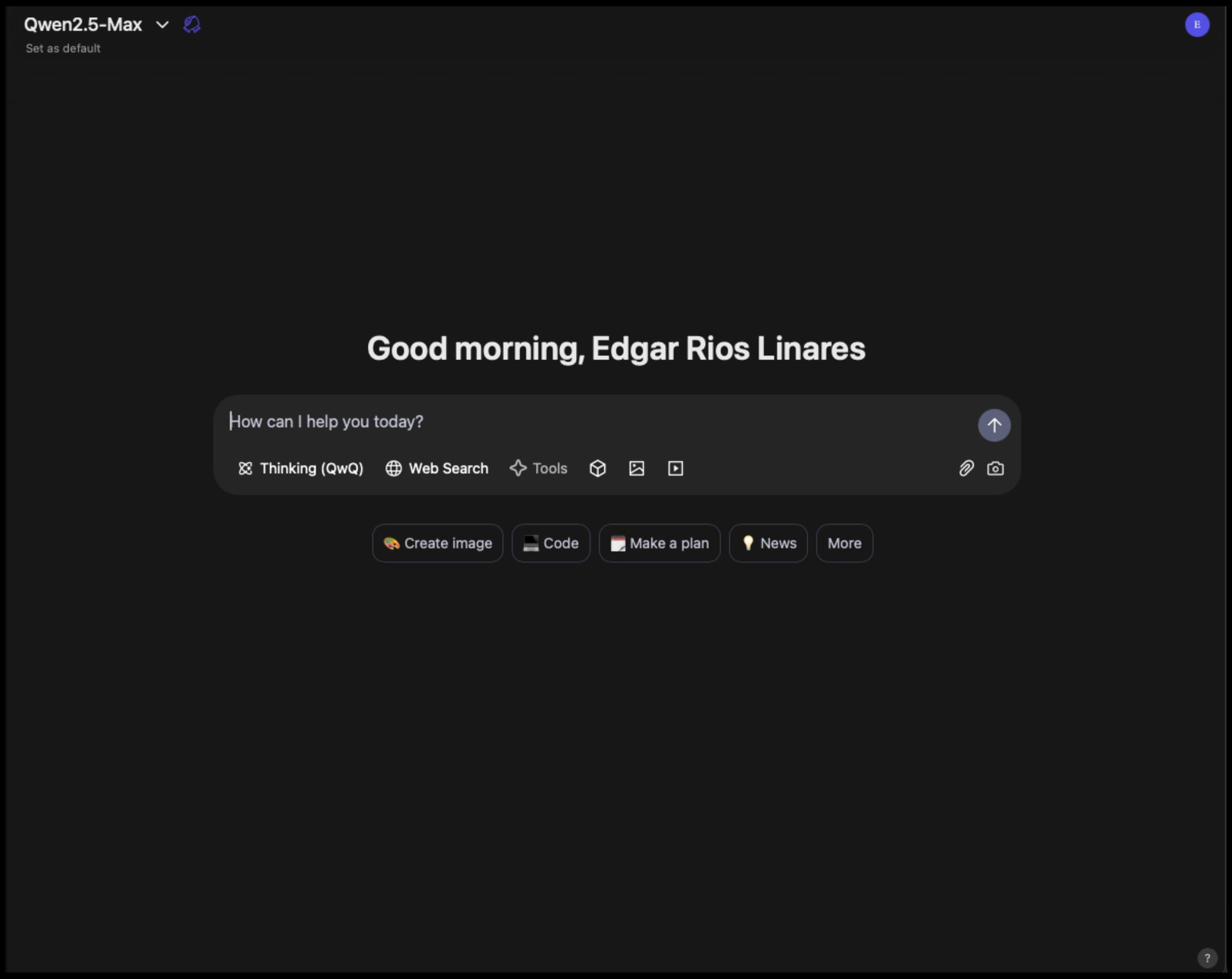


Disco 78.59 GB disponibles

0 s se ejecutó 8:57 a.m.

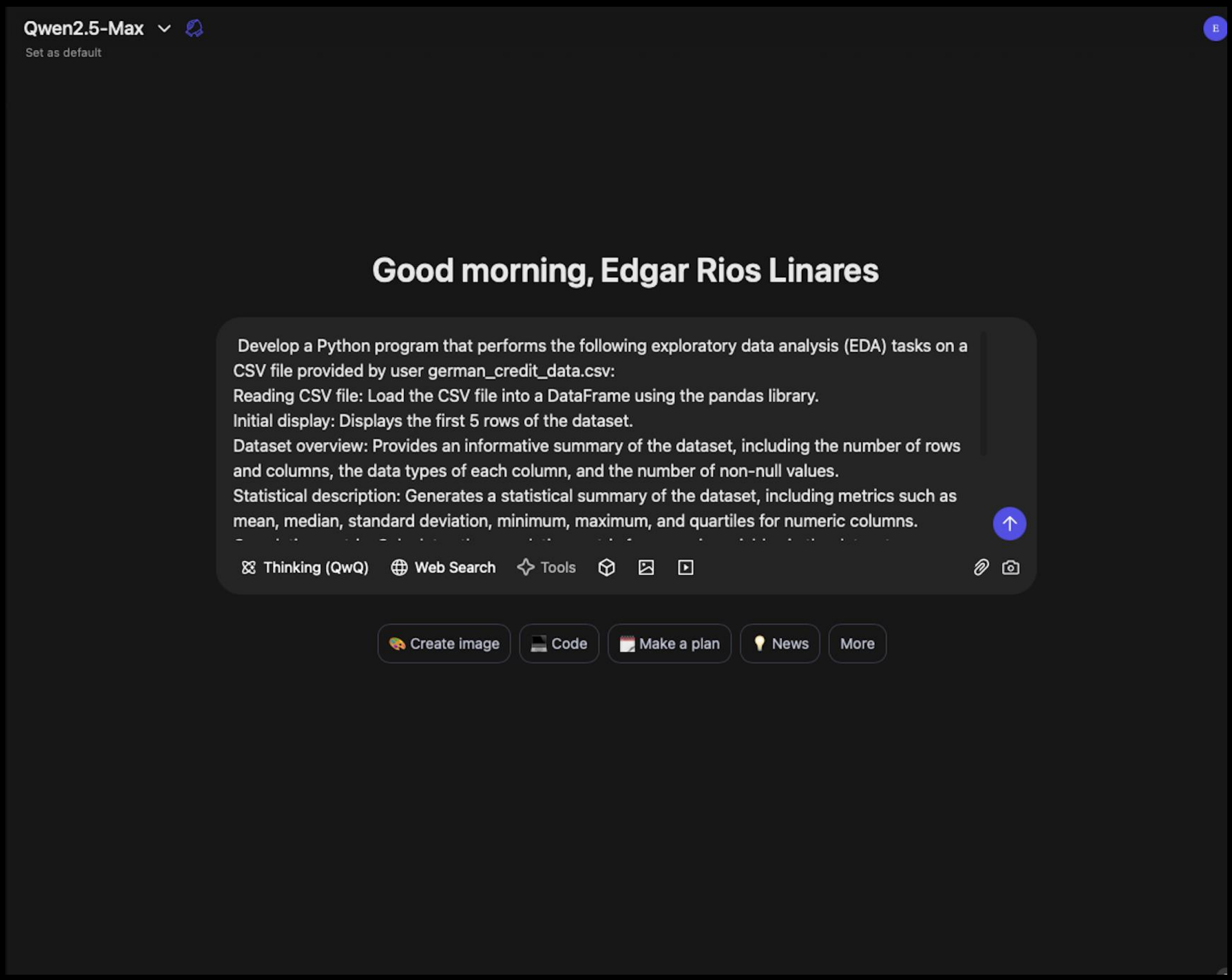
Qwen

Go to Qwen website



Qwen

Introduce a prompt to create EDA in python



Qwen

See result of LLM

Qwen2.5-Max

Qwen2.5-Max 8:59 AM

Below is a Python program that performs the requested exploratory data analysis (EDA) tasks on the `german_credit_data.csv` file. The program is modular, well-commented, and includes error handling for common issues.

python Copy

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

def load_csv(file_path):
    """
    Load a CSV file into a pandas DataFrame.

    Parameters:
        file_path (str): Path to the CSV file.

    Returns:
        pd.DataFrame: Loaded DataFrame.
    """
    try:
        df = pd.read_csv(file_path)
        print("CSV file successfully loaded.")
        return df
    except FileNotFoundError:
        print(f"Error: The file '{file_path}' was not found. Please check the file path.")
        return None
    except pd.errors.EmptyDataError:
        print("Error: The file is empty or contains no data.")
        return None
    except pd.errors.ParserError:
        print("Error: The file could not be parsed. Check the file format.")
        return None

def display_initial_rows(df):
```

Send a Message

Thinking (QwQ) Web Search Tools

Thinking (QwQ) is a preview model and is still being updated.

Qwen

See result of LLM

Qwen2.5-Max

Explanation of the Code:

1. **Modular Design** :
   - Each task is encapsulated in its own function (`load_csv` , `display_initial_rows` , `dataset_overview` , etc.), making the code reusable and easy to understand.
2. **Error Handling** :
   - The `load_csv` function handles errors such as missing files, empty files, and parsing issues.
   - The `calculate_correlation_matrix` function checks if there are numeric columns before attempting to calculate the correlation matrix.
3. **Dataset Overview** :
   - The `dataset_overview` function provides a detailed summary of the dataset, including the number of rows, columns, data types, and descriptive statistics.
4. **Correlation Matrix and Heatmap** :
   - The correlation matrix is calculated only for numeric columns using `select_dtypes` .
   - The heatmap is visualized using `seaborn.heatmap` , with annotations for better readability.
5. **User Interaction** :
   - The program prompts the user to input the file path, making it flexible for different datasets.
6. **Visualization** :
   - The heatmap is styled with a color map (`coolwarm`) and annotated with correlation values rounded to two decimal places.

How to Run the Program:

1. Save the code in a Python file, e.g., `eda_script.py` .
2. Place the `german_credit_data.csv` file in the same directory or provide the full path to the file

Send a Message

Thinking (QwQ) Web Search Tools

Thinking (QwQ) is a preview model and is still being updated.

Qwen

Copy Python code

Qwen2.5-Max

Qwen2.5-Max 8:59 AM

Below is a Python program that performs the requested exploratory data analysis (EDA) tasks on the `german_credit_data.csv` file. The program is modular, well-commented, and includes error handling for common issues.

python Copy

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

def load_csv(file_path):
    """
    Load a CSV file into a pandas DataFrame.

    Parameters:
        file_path (str): Path to the CSV file.

    Returns:
        pd.DataFrame: Loaded DataFrame.
    """
    try:
        df = pd.read_csv(file_path)
        print("CSV file successfully loaded.")
        return df
    except FileNotFoundError:
        print(f"Error: The file '{file_path}' was not found. Please check the file path.")
        return None
    except pd.errors.EmptyDataError:
        print("Error: The file is empty or contains no data.")
        return None
    except pd.errors.ParserError:
        print("Error: The file could not be parsed. Check the file format.")
        return None

def display_initial_rows(df):
```

Send a Message

Thinking (QwQ) Web Search Tools

Thinking (QwQ) is a preview model and is still being updated.

Qwen + Colab

Go to Colab, paste code & check results

EDA_QWen.ipynb ☆

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda

Compartir

Archivos

..

sample_data

german_credit_data.csv

Disco 78.59 GB disponibles

+ Código + Texto

RAM Disco

```python
def main():
    """
    Main function to perform EDA on the provided CSV file.
    """
    # Prompt user for the file path
    file_path = "german_credit_data.csv"

    # Step 1: Load the CSV file
    df = load_csv(file_path)
    if df is None:
        return

    # Step 2: Display the first 5 rows
    display_initial_rows(df)

    # Step 3: Provide an overview of the dataset
    dataset_overview(df)

    # Step 4: Calculate the correlation matrix
    corr_matrix = calculate_correlation_matrix(df)

    # Step 5: Plot the heatmap
    plot_heatmap(corr_matrix)

if __name__ == "__main__":
    main()
```

```
CSV file successfully loaded.

First 5 rows of the dataset:
   Unnamed: 0  age     sex  job housing saving_accounts checking account  \
0           0   67    male    2     own             NaN           little
1           1   22  female    2     own          little         moderate
2           2   49    male    1     own          little              NaN
3           3   45    male    2    free          little           little
4           4   53    male    2    free          little           little

   credit amount  duration              purpose  risk
```

1 s se ejecutó 9:05 a.m.

Qwen + Colab

Go to Colab, paste code & check results

EDA_QWen.ipynb

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda

Compartir

Archivos

..

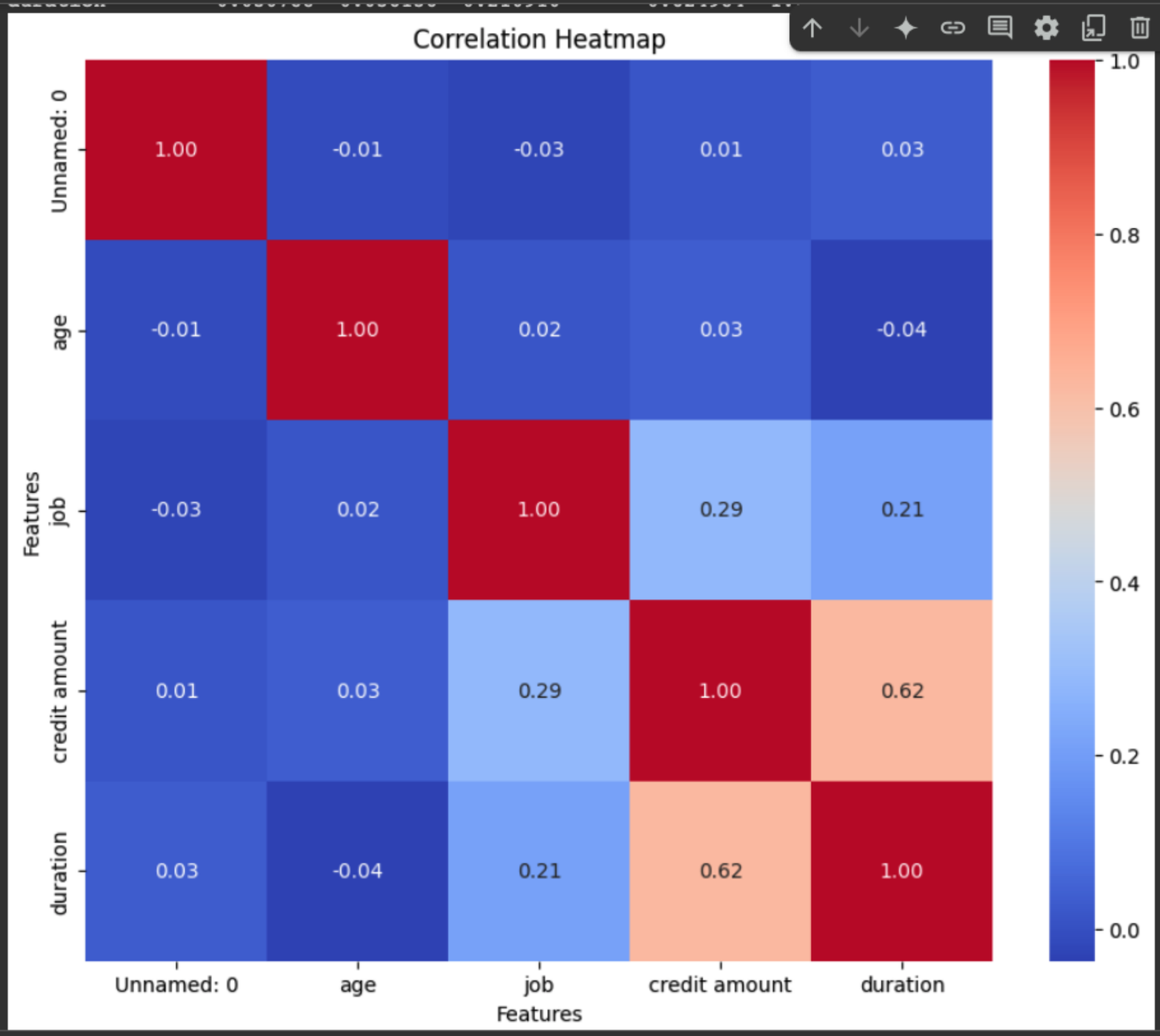
sample_data

german_credit_data.csv

Disco 78.59 GB disponibles

+ Código + Texto

RAM Disco



1 s se ejecutó 9:05 a.m.

The work is done, great job!"

You have a basic EDA quickly

Educator in AI

Artificial Intelligence

Data Engineering

Machine Learning

Data Science

📍Linkedin —> https://www.linkedin.com/in/erlinares/

👋 Follow us on 𝕏: https://x.com/erlinares

💻 GitHub: https://github.com/erlinares/365_AI_Journey/

💬 Discord: https://discord.gg/5fFM2zh8

Edgar Rios Linares