**Data Preprocessing and Storing Datasets on GitHub**

**🎯 Objective**

By the end of this lab, you will:

* Load and preprocess a sample dataset using Python
* Use **GitHub Copilot** to assist with writing preprocessing code
* Create a GitHub repository and push your cleaned dataset and script

**🧰 Prerequisites**

* GitHub account with Copilot activated
* Visual Studio Code or Jupyter Notebook with GitHub Copilot installed
* Basic knowledge of Python, pandas, and Git
* Git installed on your machine and configured

**🪜 Step-by-Step Instructions**

**🟦 Step 1: Set Up Environment with bash**

1. Create a folder on your local system:

mkdir data-preprocessing-lab

cd data-preprocessing-lab

1. Create a virtual environment (optional but recommended):

python -m venv venv

source venv/bin/activate # On Windows: venv\Scripts\activate

1. Install required libraries:

pip install pandas numpy matplotlib jupyter

1. Launch VS Code or Jupyter Notebook in this directory.

**🟨 Step 2: Create and Load Dataset**

1. Create a new Python file: preprocess\_data.py
2. Use GitHub Copilot to write the code to load a sample dataset: Start typing:

import pandas as pd

# Load the Titanic dataset from a URL

👉 GitHub Copilot will suggest:

df = pd.read\_csv("https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv")

1. Run the script and confirm the dataset is loaded:

print(df.head())

**🟩 Step 3: Data Cleaning and Preprocessing with AI**

Now let’s guide Copilot to help preprocess the data.

**Example Prompts to Type:**

* # Drop rows with missing values
* # Encode the 'Sex' column as binary values
* # Fill missing 'Age' values with the median

Copilot suggestions:

df.dropna(inplace=True)

df['Sex'] = df['Sex'].map({'male': 0, 'female': 1})

df['Age'].fillna(df['Age'].median(), inplace=True)

**🟧 Step 4: Save Cleaned Dataset**

df.to\_csv("cleaned\_titanic.csv", index=False)

print("Cleaned dataset saved.")

**🟥 Step 5: Push to GitHub**

1. Initialize a Git repo:

git init

git add .

git commit -m "Initial commit with preprocessing code and dataset"

1. Create a new repo on GitHub (manually or via GitHub CLI):

bash

CopyEdit

gh repo create data-preprocessing-lab --public --source=. --remote=origin --push

or manually:

* + Go to GitHub → New Repo → name it data-preprocessing-lab
  + Follow the instructions to push code:

bash

CopyEdit

git remote add origin https://github.com/your-username/data-preprocessing-lab.git

git branch -M main

git push -u origin main

**Deliverables**

By the end of this lab, your GitHub repo should contain:

data-preprocessing-lab/

├── cleaned\_titanic.csv

├── preprocess\_data.py

└── README.md (optional)