**SENG8081 - Case Studies**

**Big Data Solution Architecture**

**Section 1**

**Project Title**

**Pipeline to Insights: Data Engineering and Visualization of Bitcoin Market Trends**

**Team 3**

**Team Members:**

[Cherukupalli Krishna Vamsi](javascript://)

Lohith Reddy Danda

Paras Rupani

**Git Hub Repository**

[**https://github.com/SENG8081/SENG8081-S25-Team3**](https://github.com/SENG8081/SENG8081-S25-Team3)

**Data Source : Yahoo Finance API**

**A computer screen shot of a program code

AI-generated content may be incorrect.**

**Dataset Cleaning Process**

This step focuses on cleaning, organizing, and storing the raw Bitcoin and Ethereum market data so that it's ready for meaningful analysis in the next stage (Data Research).

We ensured that the data is:

* Clean (free from duplicates, missing values, and errors)
* Accurate
* Well-organized (saved in a structured CSV format)

**Tools Used:**

* Python 3.x
* Pandas’ library
* VS Code
* Git + GitHub for version control and collaboration

**Cleaning Process – Step-by-Step:**

**1. Loaded Raw Datasets**

We used the pd.read\_csv() function to load:

* btc\_daily\_2014\_to\_yesterday.csv
* eth\_daily\_2014\_to\_yesterday.csv  
  These contain daily historical data for Bitcoin and Ethereum from 2014 to the present.

**2. Removed Duplicate Rows**

To avoid repeated entries, we used. drop duplicates() to eliminate any rows that were exactly the same.

**3. Removed Invalid (Negative) Prices**

Prices of cryptocurrencies should never be negative.  
We checked the **Open**, **High**, **Low**, and **Close** columns and removed any rows where the price was < 0.

**4. Converted Date Column**

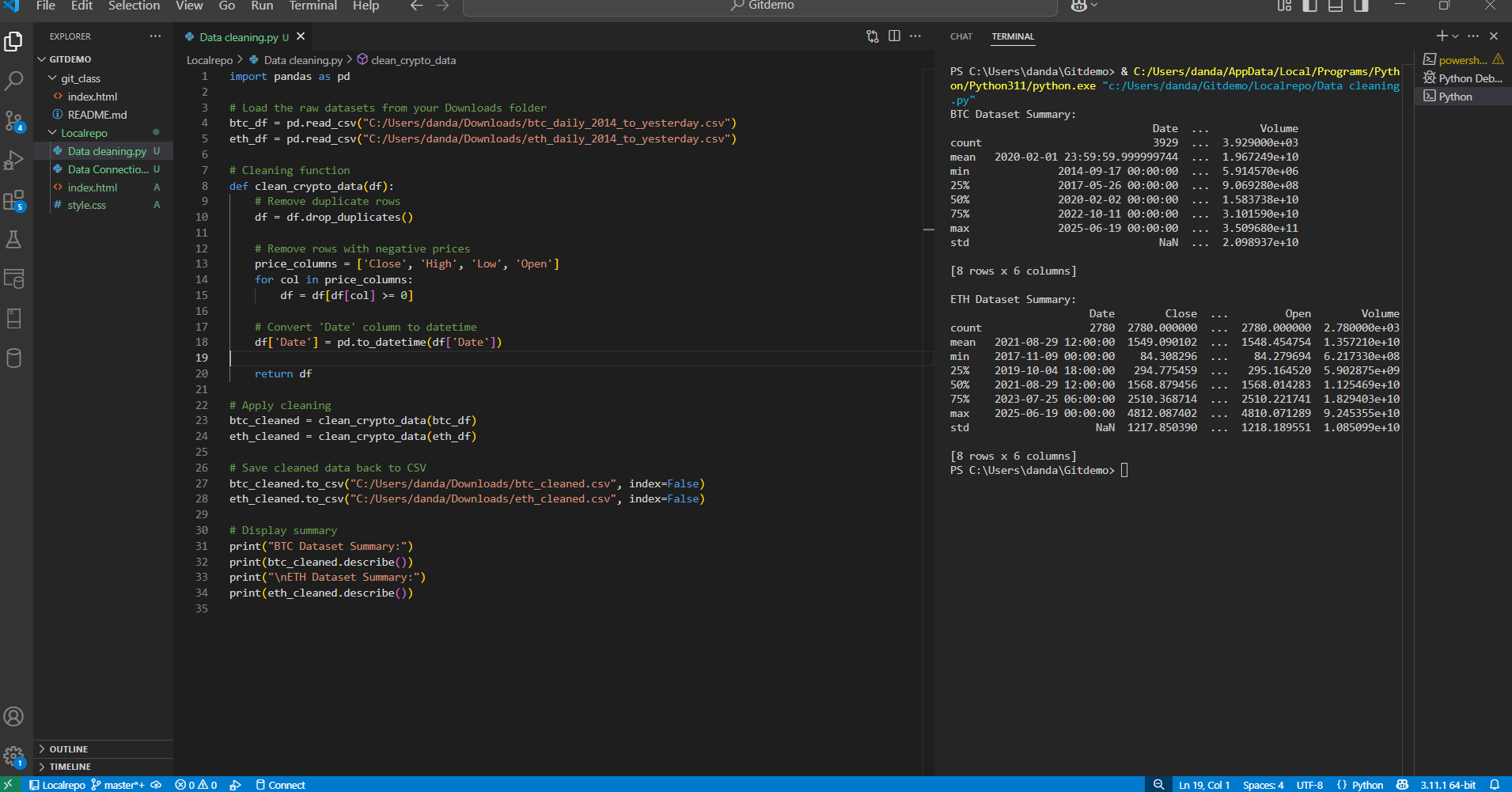
The original date column was just plain text.  
We used pd.to\_datetime() to convert it into proper datetime format so we can sort and analyze by time.

**5. Saved Cleaned Files**

We saved the cleaned datasets as:

* btc\_cleaned.csv
* eth\_cleaned.csv

These cleaned files were added to our GitHub repository under BTC\_ETH\_USD\_DATA/.



**Create the Database and Tables**

**A computer screen with a white screen

AI-generated content may be incorrect.**

**Used Python to Upload the Data**

**A screenshot of a computer

AI-generated content may be incorrect.**

We used Python to open the cleaned crypto CSV files like reading a spreadsheet. Then, connected Python to SQL Server and gently drops each row of data into the right table like moving rows from Excel into a digital filing cabinet. This way, we make data safely stored and ready for analysis.

**The Uploaded Data**

**A screenshot of a computer

AI-generated content may be incorrect.**

A computer screen with a white screen

AI-generated content may be incorrect.

Data is uploaded into SQL Server; you can view it using SQL queries in SQL Server Management Studio (SSMS). Think of it like opening a folder to check that everything you filed is there. By running a command like SELECT \* FROM Bitcoin Data and SELECT\*FROM EthereumData, you can instantly see the rows you've inserted just like scrolling through a spreadsheet and confirm that your data has been successfully stored in the database.