

Hands-On Lab: Generative AI for Data Preparation

Estimated duration: 30 minutes

Overview

In this lab, you will learn how to use generative AI to prepare data using the tool ChatGPT.

Objective(s)

After completing this lab, you will be able to:

- Sign in on <https://chat.openai.com/>
- Upload a dataset
- Handle missing values
- Perform data standardization
- Perform data normalization

Pre-requisite(s)

- A free ChatGPT account
- A basic understanding of **Exploratory Data Analysis (EDA)**

Dataset

The dataset is a **filtered and modified version** of the [Laptop Price Prediction using specifications dataset](#), available under the **Database Contents License (DbCL) v1.0 on the Kaggle website**.

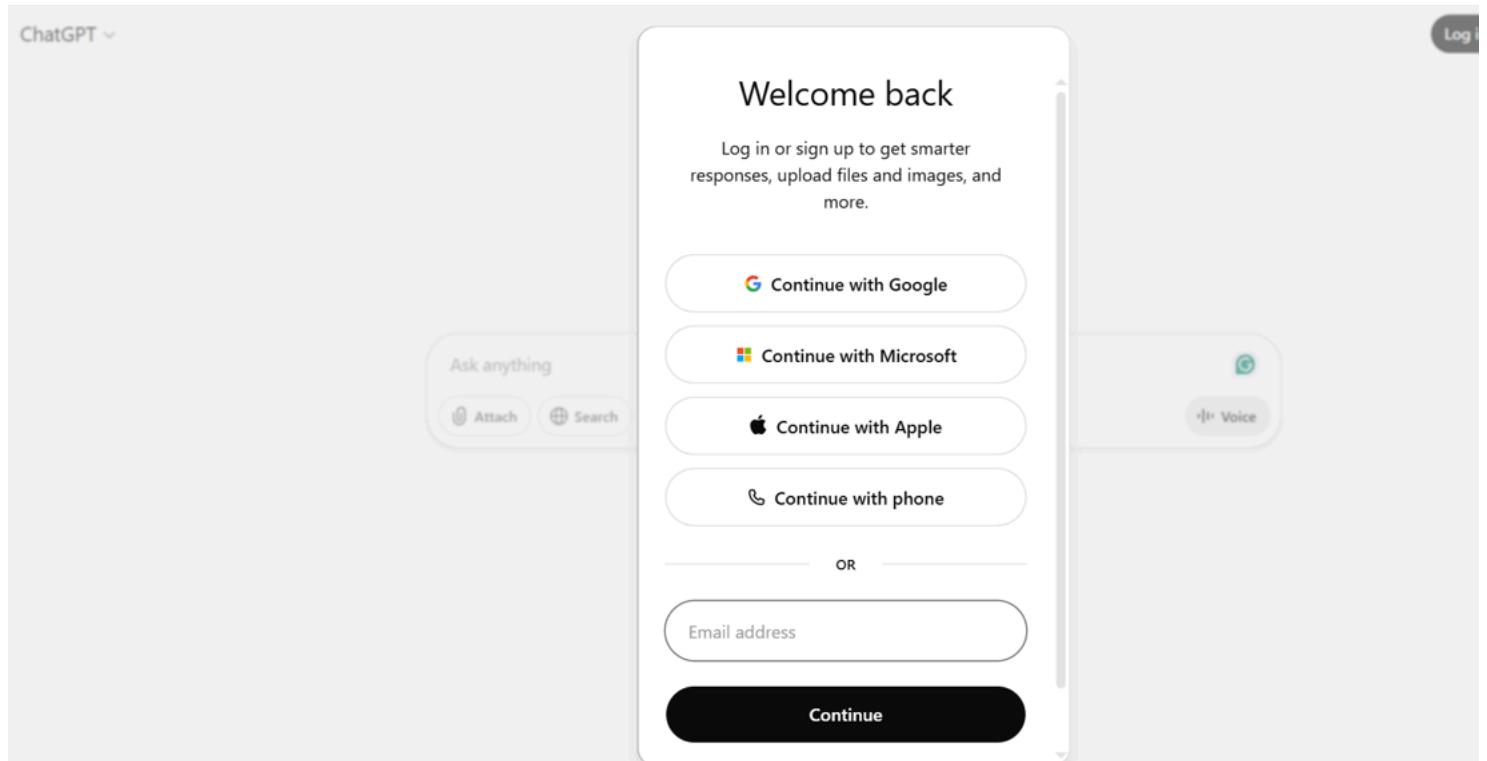
While holding down the **Ctrl (Windows) or Command (Mac)** button, click [here](#) to download the data set.

Task 1: Sign in on ChatGPT

Step 1: If you do not have an account, go to <https://chat.openai.com/>, and click **Log in or Sign up for free** at the top right corner.

The screenshot shows a web browser window with the URL chat.openai.com in the address bar. The page header includes the ChatGPT logo and a dropdown menu labeled "ChatGPT". On the right side of the header, there is a red-bordered "Log in" button. The main content area features a large input field with the placeholder text "What can I help with?". Below this is a rounded rectangular button with the text "Ask anything" and three smaller buttons for "Attach", "Search", and "Study". In the bottom right corner of the input field, there is a "Voice" button with a microphone icon.

Step 2: Log in or sign up using any of the options — Google, Microsoft, Apple, GitHub, or Email — and follow the prompts to sign in.

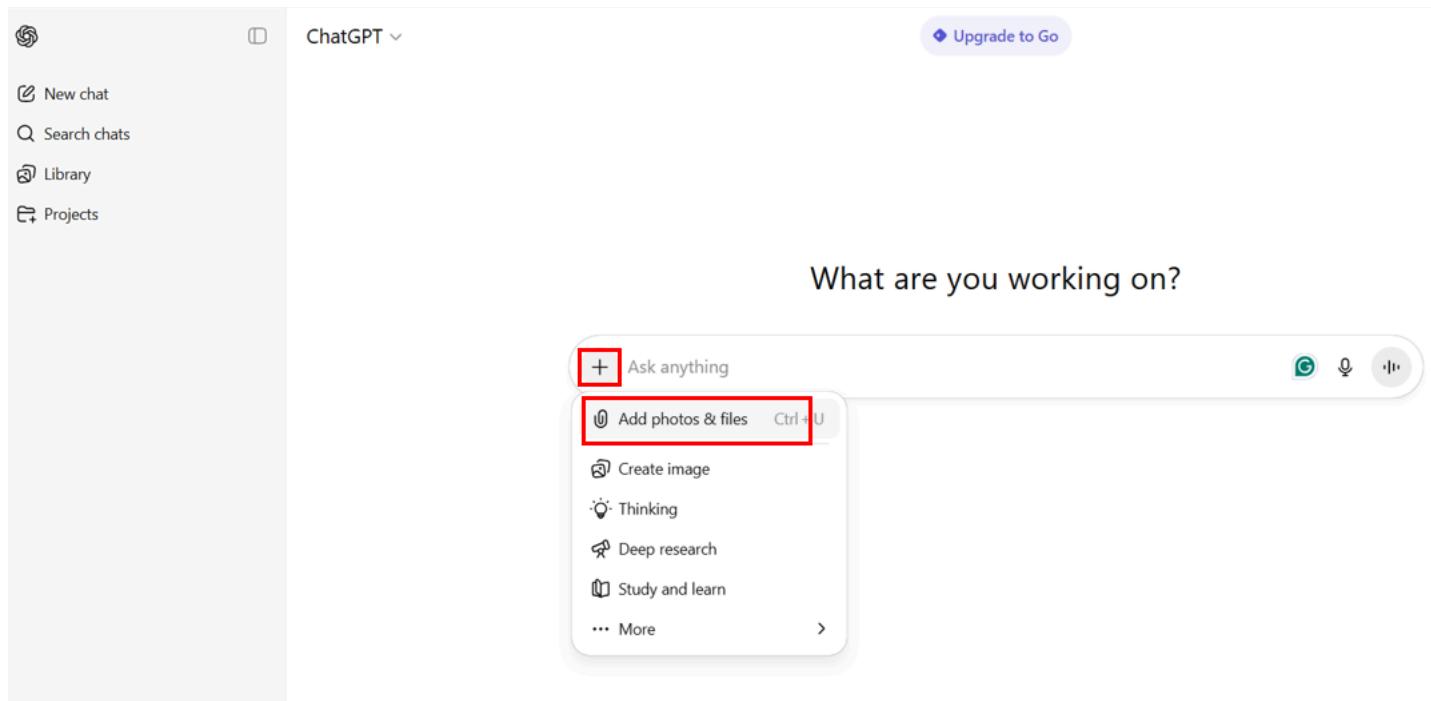


Step 3: After creating your account, log in with your credentials, and the ChatGPT interface will be displayed.

A screenshot of the ChatGPT interface. On the left is a sidebar with icons for 'New chat', 'Search chats', 'Library', and 'Projects'. The main area has a header with 'ChatGPT' and an 'Upgrade to Go' button. The central part asks 'What are you working on?'. Below it is a search bar with a plus sign and the placeholder 'Ask anything'. On the far right are three circular icons: a green one with a 'G', a blue one with a question mark, and a grey one with a refresh symbol.

Task 2: Upload Dataset

Step 1: On the ChatGPT interface, click the Add (+) icon and choose Add photos & files to upload the Laptop Pricing Data that you have downloaded.



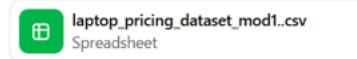
Step 2: Once the file is uploaded, write a prompt **Describe the dataset** to get the statistical description of the dataset. Copy the prompt below, paste it into the ChatGPT interface, and press Enter.

Describe the dataset

A screenshot of the ChatGPT interface showing the result of a previous step. The sidebar and top navigation are identical to the first screenshot. The main area shows a message card with a green icon of a spreadsheet and the text 'laptop_pricing_dataset_mod1..csv Spreadsheet'. Below this, the 'What are you working on?' input field contains the prompt 'Describe the dataset'. To the right of the input field are the same three circular icons: green 'G', blue microphone, and grey document.

Step 3: Scroll down through the dataset output to view the details presented by ChatGPT.

ChatGPT ▾



Here's a summary of the **laptop pricing dataset** you uploaded:

- Number of rows: 238
- Number of columns: 13

Column Overview

Column	Type	Missing Values	Description (inferred)
Unnamed: 0	int64	0	Likely an index column.
Manufacturer	object	0	Brand name (e.g., Acer, Dell, HP).
Category	int64	0	Encoded laptop type or segment.
Screen	object	0	Type of screen (e.g., IPS Panel, Full HD).
GPU	int64	0	Encoded GPU type or manufacturer.

+ Ask anything



Task 3: Handle missing values

Step 1: To identify the missing values, write a prompt:

Identify the attributes with missing data

and press **Enter**.

The response will display two attributes with missing values as:

- "Screen_Size_cm" - categorical variable
- "Weight_kg" - continuous variable.

To replace the missing values, copy the prompt below, paste it into the ChatGPT interface, and press Enter.

- Missing entries in columns containing categorical values need to be replaced with the most frequent entries.
- Missing entries in columns with continuous data need to be replaced with the mean value of the column. If a value is missing in the target column,

The prompt response will be shown as below:

ChatGPT ▾

Share

Identify the attributes with missing data

The dataset contains **two** attributes with missing data:

Attribute	Missing Values	Percentage of Total
Screen_Size_cm	4	1.68%
Weight_kg	5	2.10%

All other columns are **complete (no missing values)**.

Would you like me to show how these missing values are distributed (e.g., which manufacturers or categories they belong to)?



+ Ask anything

G 0 ⓘ

Step 2: Write a prompt:

Replace the missing values with the mean value in the respective column and save the updated dataset.

ChatGPT ▾

Share

Replace the missing values with the mean value in the respective column and save the updated dataset.

The missing values in `Screen_Size_cm` and `Weight_kg` have been replaced with their respective mean values.

You can download the cleaned dataset using this link:

[laptop_pricing_dataset_cleaned.csv](#) ⓘ



+ Ask anything

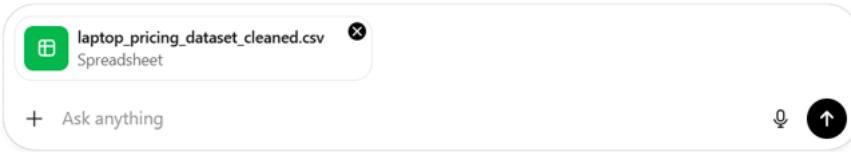
G 0 ⓘ

ChatGPT will provide a **download link** for the cleaned dataset. You can **download the CSV** and open it in Excel if needed.

Task 4: Perform the data standardization

Note that the parameter 'Screen_Size_cm' represents the screen size in **centimeters** instead of the standard unit (inches). Similarly, the parameter 'Weight_kg' needs to be expressed as **pounds**.

Step 1: Upload the updated dataset by clicking the '+' icon and attaching the new dataset you saved in the previous step.



Step 2: Write the following prompts to change the attributes into standard units:

- Data under Screen_Size_cm; is assumed to be in centimeters. Convert this data into inches. Modify the name of the attribute to Screen_Size_inch.
- Data under Weight_kg is assumed to be in kilograms. Convert this data into pounds. Modify the name of the attribute to Weight_pounds.
- Provide the updated dataset as a downloadable CSV.

ChatGPT ▾

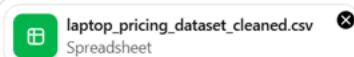
Share

The missing values in `Screen_Size_cm` and `Weight_kg` have been replaced with their respective mean values.

You can download the cleaned dataset using this link:

[CSV](#) [laptop_pricing_dataset_cleaned.csv](#)

Reply Like Report Upvote ...



Convert the 'Screen_Size_cm' values from centimeters to inches and rename the column to 'Screen_Size_inch'.

Convert the 'Weight_kg' values from kilograms to pounds and rename the column to 'Weight_pounds'. Provide the updated dataset for download as a CSV.

+

0



ChatGPT will generate and **provide the new dataset link**. You can **download the CSV** and open it in Excel if needed.

Share

ChatGPT ▾

laptop_pricing_dataset_cleaned.csv

laptop_pricing_dataset_cleaned.csv
Spreadsheet

Convert the 'Screen_Size_cm' values from centimeters to inches and rename the column to 'Screen_Size_inch'.
 Convert the 'Weight_kg' values from kilograms to pounds and rename the column to 'Weight_pounds'.
 Provide the updated dataset for download as a CSV.

The conversions and renaming have been completed successfully.

You can download the updated dataset here:

[laptop_pricing_dataset_updated.csv](#)



+ Ask anything

0



Task 5: Perform the data normalization

Some attributes might require normalization. Since there are multiple normalization methods, you need to specify your exact requirements and tasks in the prompt.

You can either:

- Save the normalized data as a new attribute, or
- Overwrite the original attribute with the normalized values.

Example: Assume the data under 'CPU_frequency' needs to be normalized using the maximum value of the attribute. You want the normalized values to replace the original attribute instead of creating a new one.

Make sure to provide clear instructions in your prompt so ChatGPT can process the normalization correctly.

Step 1: Upload the new updated dataset from the previous step by clicking the '+' icon and attach the new dataset.

laptop_pricing_dataset_updated.csv

Spreadsheet

+ Ask anything



Step 2: Write the following prompt:

- Normalize the values under the column 'CPU_frequency' with respect to its maximum value.
- Do not create a new attribute; replace the existing values.
- Provide the updated dataset as downloadable CSV.

ChatGPT ▾

◆ Upgrade to Go ×

Share

laptop_pricing_dataset_updated.csv
Spreadsheet

Normalize the values under the column 'CPU_frequency' with respect to its maximum value.
Do not create a new attribute; replace the existing values.
Provide the updated dataset as downloadable CSV.

Here's your updated dataset with the 'CPU_frequency' column normalized (values replaced in place):

[Download laptop_pricing_dataset_normalized.csv](#)

🔗 🗑️ 🙌 🔍 ⏵ ⏷ ⏸ ⏹

+ Ask anything

G Q ⓘ

Practice problems

- Create a prompt to generate a Python code that converts the values under Price from USD to Euros.
- Modify the normalization prompt to perform min-max normalization on the CPU_frequency parameter.

Conclusion

In this lab, you have learned to handle missing values in your dataset, and performed data standardization and data normalization.

Author(s)

Pratiksha Verma

© IBM Corporation. All rights reserved.



skills Network