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)

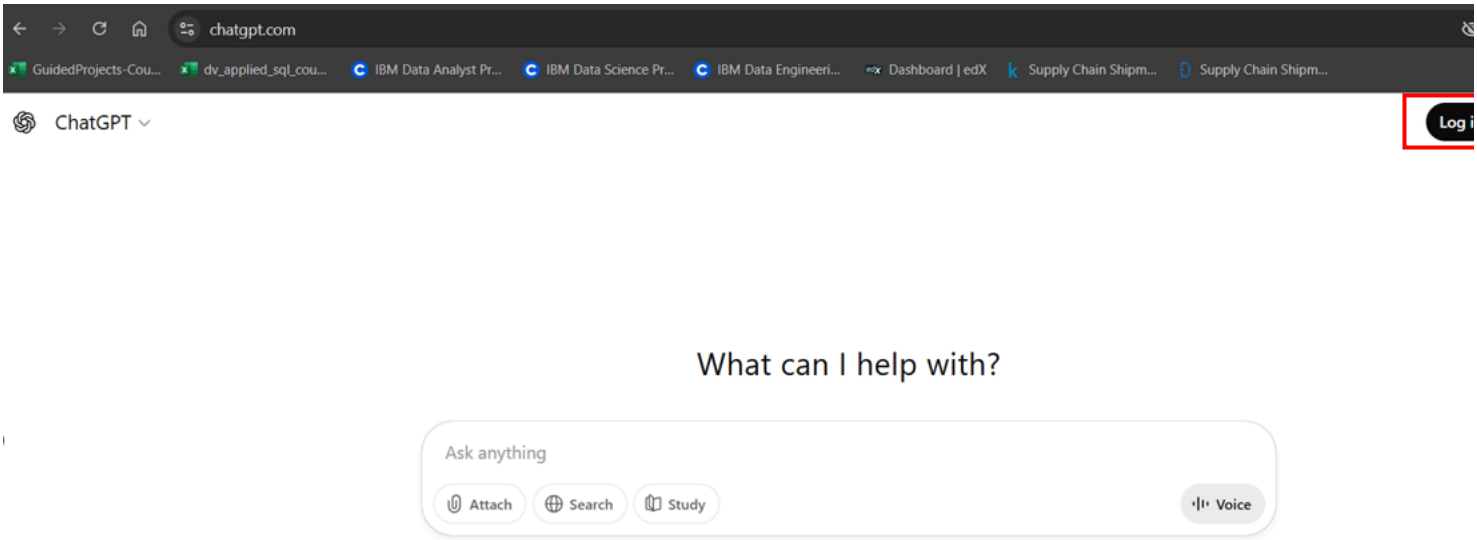
- A free ChatGPT account
- A basic understanding of **Exploratory Data Analysis (EDA)**

Dataset

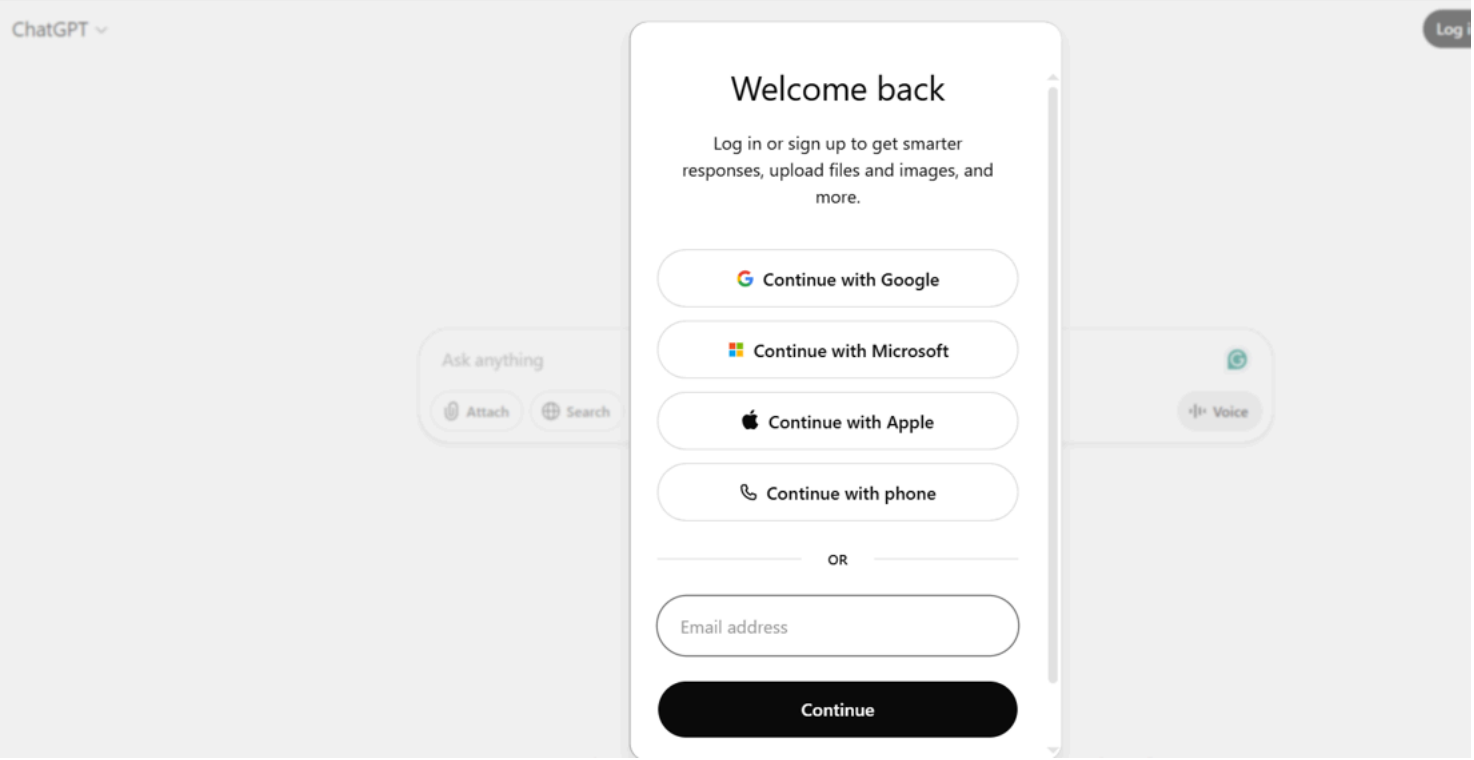
The dataset is a **filtered and modified version** of the Laptop Price Prediction using specifications dataset, available under the **Database Contents License (DbCL) v1.0 on the Kaggle website**.
While holding down the **Ctrl (Windows) or Command (Mac) button**, click here to download the data set.

Task 1: Sign in on ChatGPT

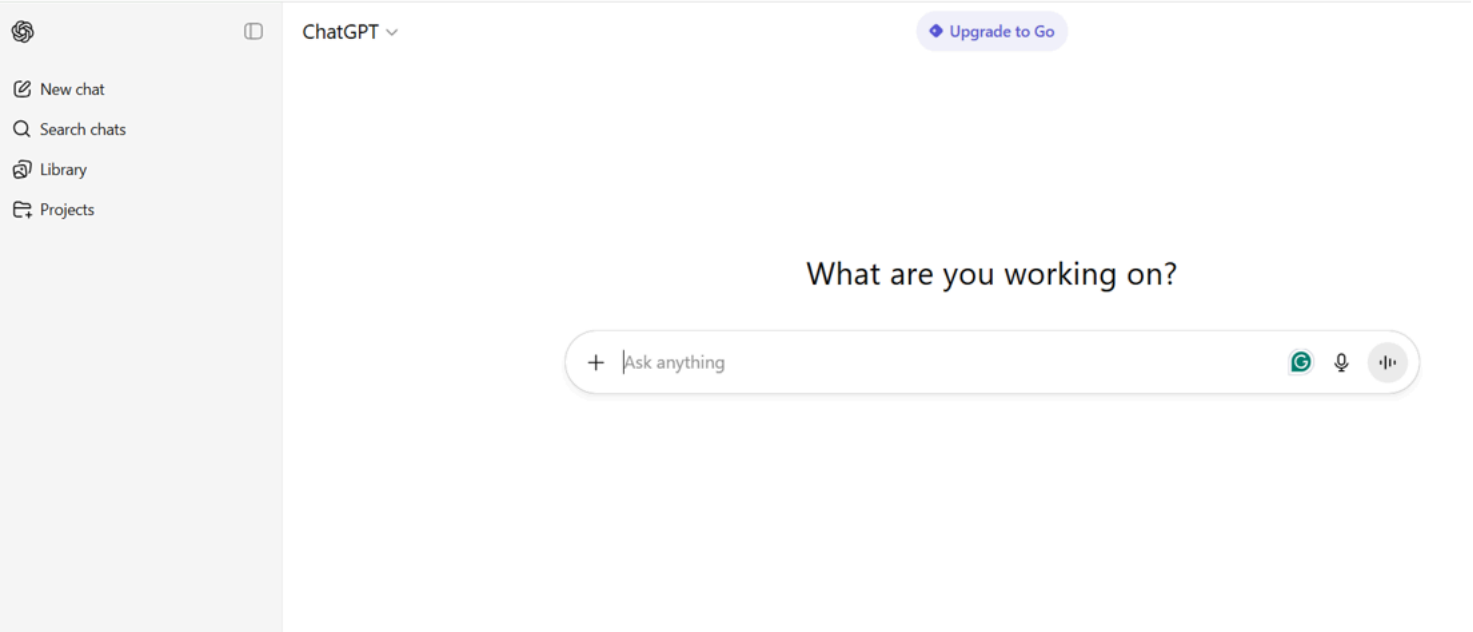
Step 1: If you do not have an account, go to https://chat.openai.com/, and click **Log in or Sign up for free** at the top right corner.



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.

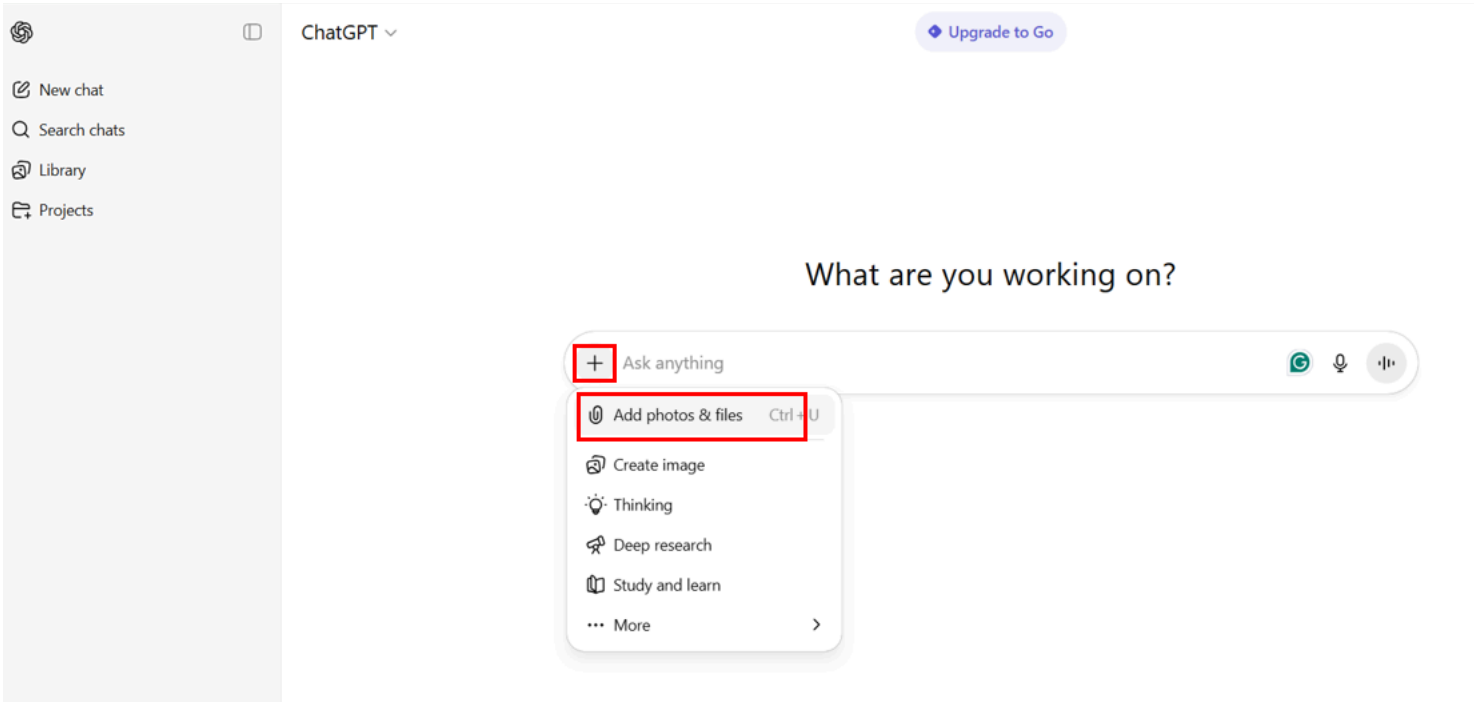


Step 3: After creating your account, log in with your credentials, and the ChatGPT interface will be displayed.



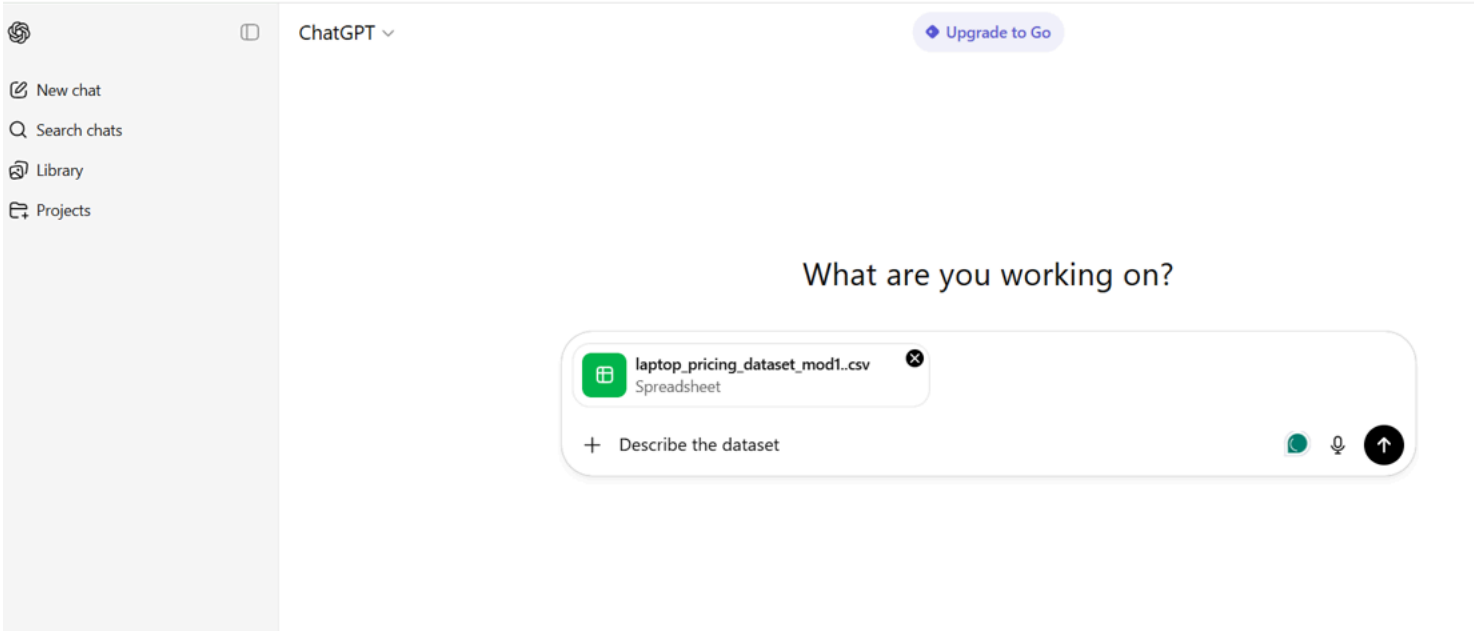
Task 2: Upload Dataset

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.

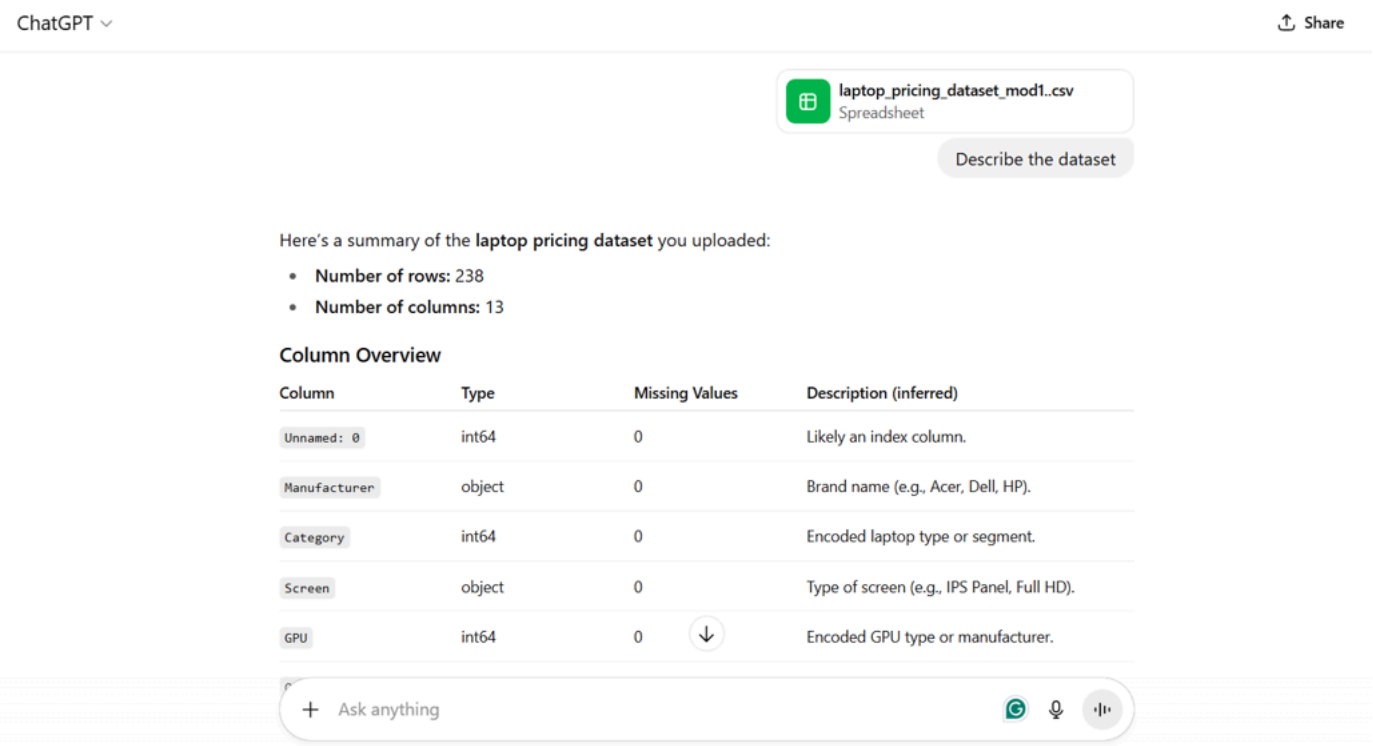


Step 2: Once the file is uploaded, write a prompt **Describe the dataset** to get the statistical description of the dataset. Copy the prompt below, paste it into the ChatGPT interface, and press Enter.

```
Describe the dataset
```



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



Task 3: Handle missing values

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

```
**Identify the attributes with missing data**
```

and press **Enter**.

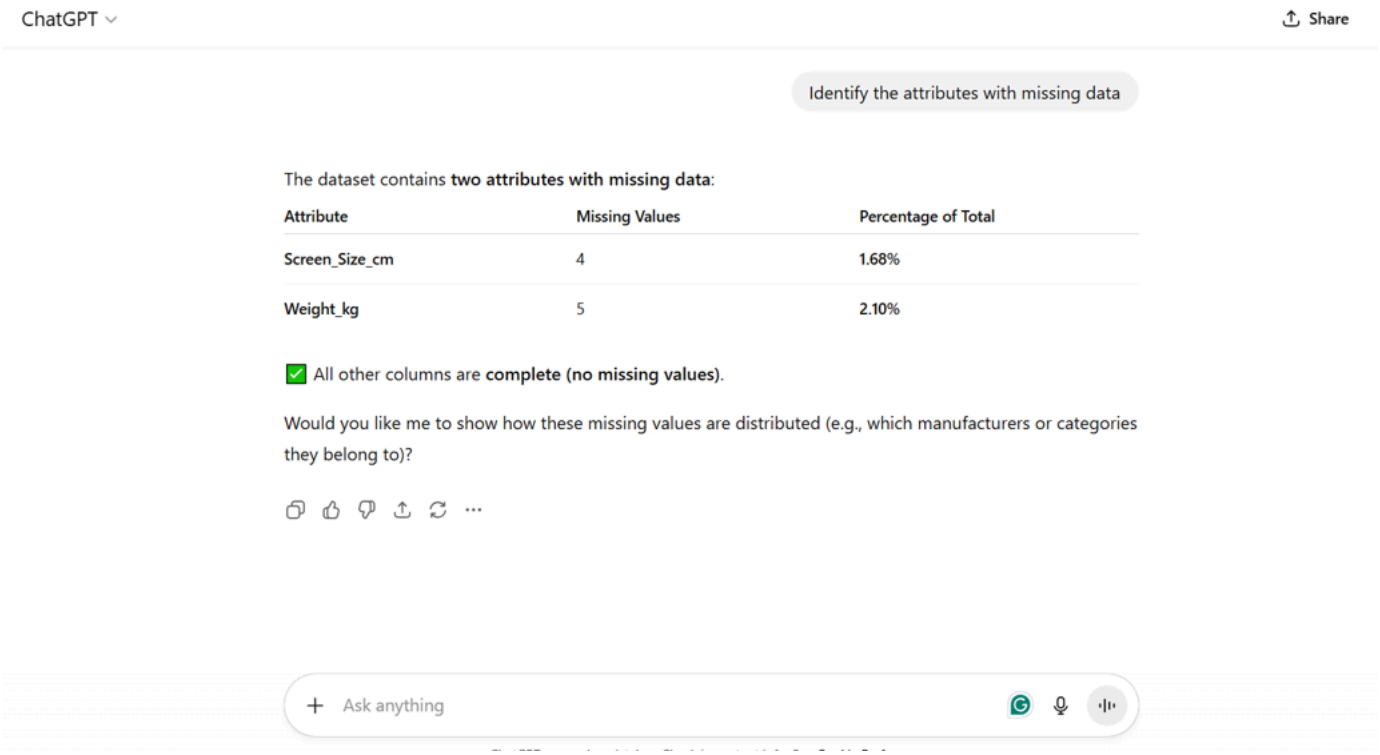
The response will display two attributes with missing values as:

- **"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.

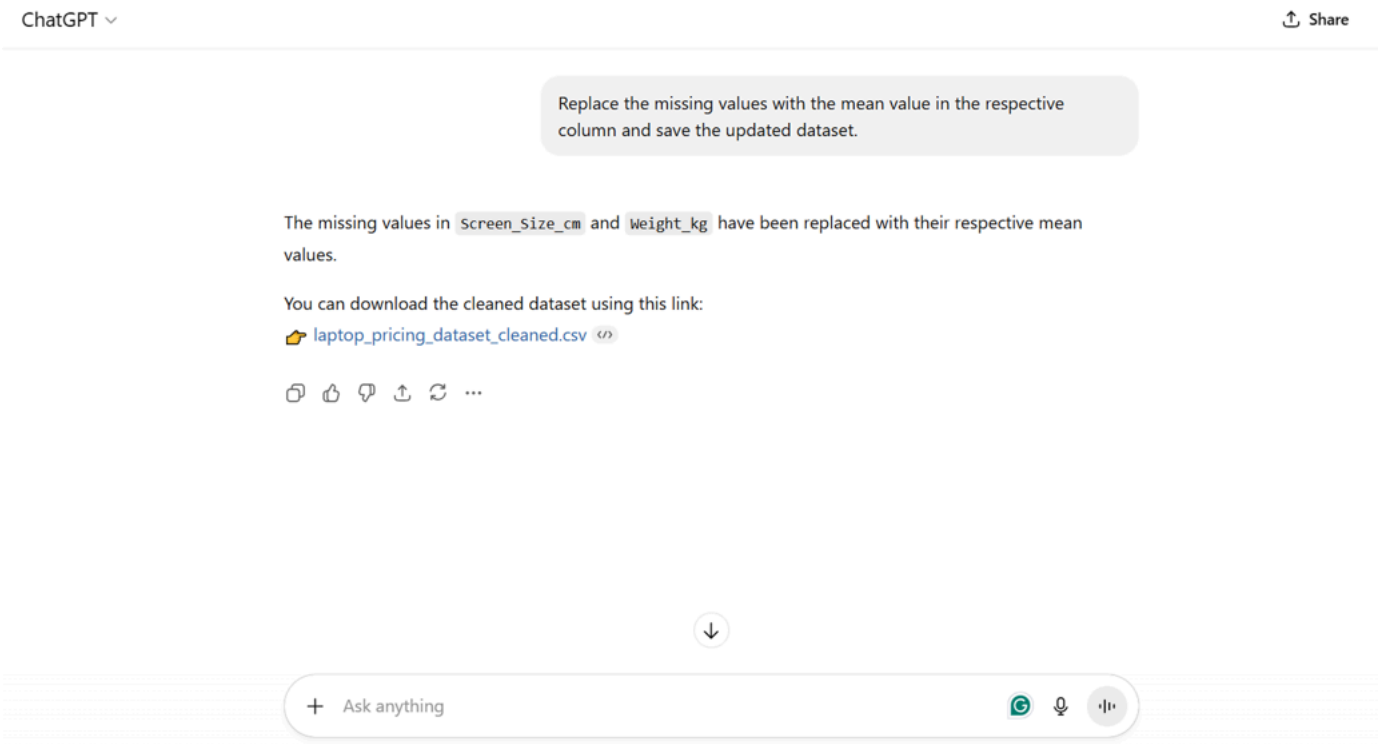
```
- Missing entries in columns containing categorical values need to be replaced with the most frequent entries.
- Missing entries in columns with continuous data need to be replaced with the mean value of the column. If a value is missing in the target column,
```

The prompt response will be shown as below:



Step 2: Write a prompt:

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

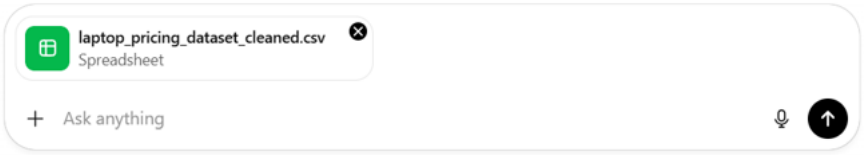


ChatGPT will provide a **download link** for the cleaned dataset. You can **download the CSV** and open it in Excel if needed.

Task 4: Perform the data standardization

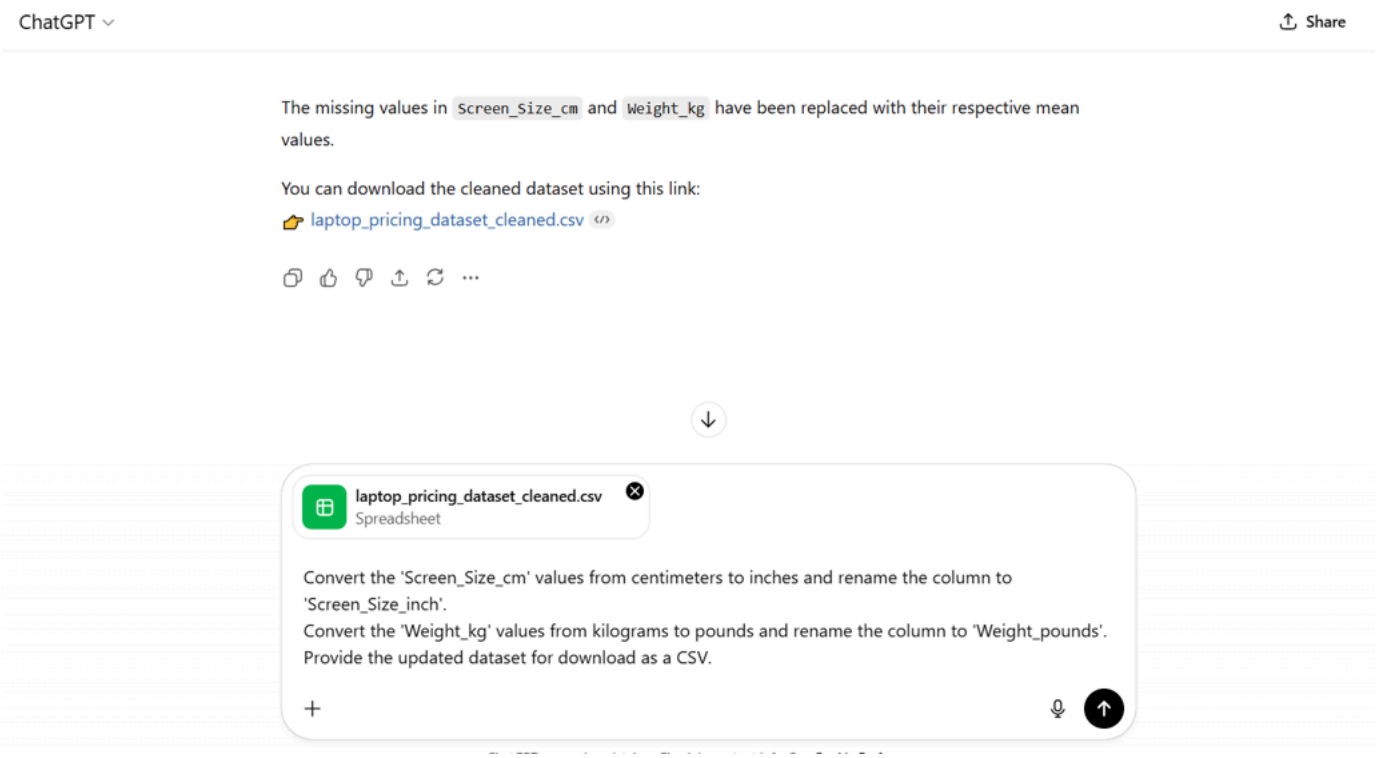
> Note that the parameter **'Screen_Size_cm'** represents the screen size in **centimeters** instead of the standard unit (inches). Similarly, the parameter **'Weight_kg'** needs to be expressed as **pounds**.

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

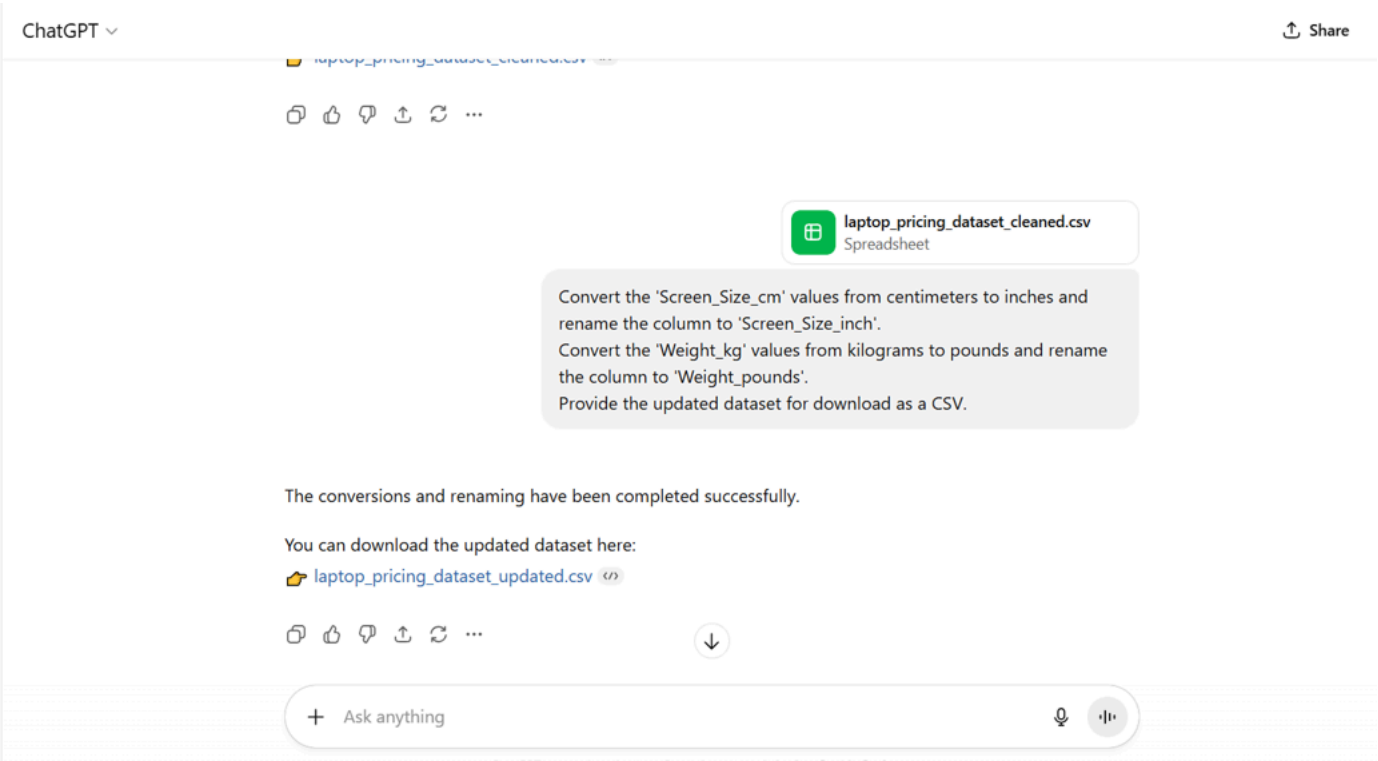


Step 2: Write the following prompts to change the attributes into standard units:

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



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.

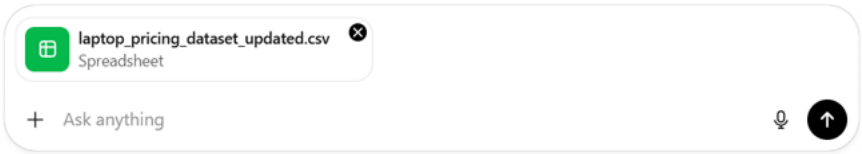
You can either:

- 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.

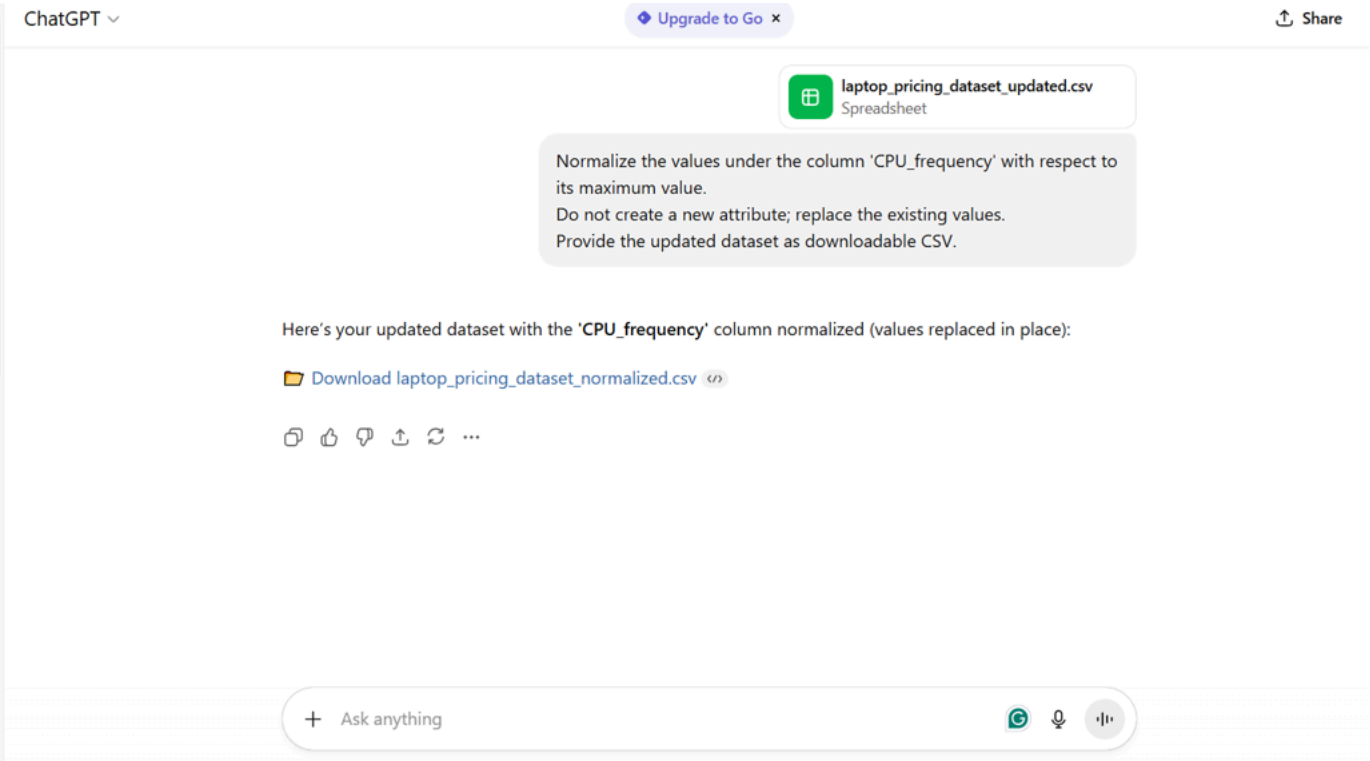
> Make sure to provide clear instructions in your prompt so ChatGPT can process the normalization correctly.

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



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.
```



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