

*Worksheet 3 - Sep 26, 2025*

**Data Cleaning with Pandas**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Goal:** Learn how to clean and prepare real-world data for analysis.

**Step 1 – Download the Dataset**

1. Go to the class GitHub repo:  
   👉 [Intro-to-Data-Scinece](https://github.com/Sammyjoon/Intro-to-Data-Scinece)
2. Open the **Week3 folder.**
3. Click on **mini\_weather.csv.**
4. Click the **“Download raw file”** button.
5. Save it in your **Downloads folder.**

**Step 2 – Upload the Dataset to Colab**

1. In Colab, click the **📂 folder icon**on the left sidebar.
2. Click the **⬆️ Upload** button.
3. Select mini\_weather.csv from your Downloads folder.

**Step 3 –** Load the Data in Python

A screenshot of a computer program

Description automatically generated

**Exercises**

**Exercise 1 – First Look**

* Print the first 5 rows using .head().
* How many rows and columns are in the dataset?

### 🧪 ****Exercise 2 – Get to Know the Columns****

#### 👩‍🏫 What We’re Learning:

When we look at a dataset, it helps to know:

* What each column represents
* What kind of data is inside (numbers, words, dates)
* How computers “see” that data (they use something called **data types**)

We’ll use some code to explore this together!

#### 🧩 Part A – What Are the Columns?

Type this in your Colab notebook:

data.columns

📌 **Question:**

1. What are the names of the columns in this dataset?
2. Which one looks like a date?
3. Which ones do you think are numbers?
4. Which ones are words or descriptions?

📝 Write your guesses in a text box below this code cell.

#### 🧪 Part B – What Kind of Data Is In Each Column?

Type this next:

data.dtypes

This tells us the **data type** of each column. Don’t worry if this looks weird — we’ll break it down:

* **object** = a mix of text, symbols, or anything that’s not a number (like "Cloudy" or "Rain")
* **float64** = a number that can have decimals (like 21.3, 0.8, etc.)
* **int64** = a whole number (like 0, 1, 20 — although you might not see this in this dataset)

📌 **Question:**  
Match the columns with their data types. For each one, write:

* Column Name
* What kind of data it has (numbers, words, date?)
* Why do you think the computer classified it that way?

Here’s a template to help:

A screenshot of a phone

Description automatically generated

**🔍 Exercise 3 – Explore PrecipType and Summary**

Now look into what values appear in important columns:

A close-up of a computer code

Description automatically generated

* What types of precipitation are in the data?
* Which weather summary appears the most?

### Exercise 4 – ****Are there any missing values?****

Introduce the idea of missing data without fixing it yet.

Use .isnull().sum() to find how many missing values are in each column.

* Which columns have missing values?

### 🧹 ****Exercise 5 – Fix the Missing Data****

Now they're ready to do something:

* Fill missing Precip Type with "unknown"
* Fill missing Temperature with average



### 🚫 ****Exercise 6 – Remove Duplicates****

Introduce .duplicated() and .drop\_duplicates()

A close-up of a computer code

Description automatically generated

* What happens if we keep the same row twice?
* Why is it important to remove duplicates?

**Exercise 7 – Rename Columns**

* Rename "Temperature (C)" to Temp\_C.
* Rename "Precip Type" to PrecipType.