4.4.1 Import and Inspect CSV Files

Now that you have learned about the Pandas library and what it can do, Maria is going to share the datasets with you, which are in CSV format. After you import these CSV files, you will need to inspect them. Then you'll be given a set of tasks to complete for the analysis.

Import the Data

Create a folder named "Resources" in your "School_District_Analysis" folder. Click the following links to download the datasets into the Resources folder.

Download schools_complete.csv (https://2u-data-curriculum-team.s3.amazonaws.com/dataviz-online/module_4/schools_complete.csv)

Download students_complete.csv (https://2u-data-curriculum-team.s3.amazonaws.com/dataviz-online/module_4/students_complete.csv)

You should now have two CSV files in your Resources folder:

- `schools_complete.csv`
- `students_complete.csv`

Inspect the Data

Now that the CSV files downloaded and imported into the correct folder, let's inspect the data in the files.

REWIND

Remember, when inspecting the data, focus on the following questions:

- How many columns and rows are there?
- What types of data are present?
- Is the data readable, or does it need to be converted in some way?

When we open `schools_complete.csv`, we can see that there is a top header row followed by 15 rows of data, labeled with the names "School ID," "school_name," "type," "size," and "budget." See the following image for reference.

	A	B	C	D	E
1	School ID	school_name	type	size	budget
2	0	Huang High School	District	2917	1910635
3	1	Figueroa High School	District	2949	1884411
4	2	Shelton High School	Charter	1761	1056600
5	3	Hernandez High School	District	4635	3022020
6	4	Griffin High School	Charter	1468	917500
7	5	Wilson High School	Charter	2283	1319574
8	6	Cabrera High School	Charter	1858	1081356
9	7	Bailey High School	District	4976	3124928
10	8	Holden High School	Charter	427	248087
11	9	Pena High School	Charter	962	585858
12	10	Wright High School	Charter	1800	1049400
13	11	Rodriguez High School	District	3999	2547363
14	12	Johnson High School	District	4761	3094650
15	13	Ford High School	District	2739	1763916
16	14	Thomas High School	Charter	1635	1043130

Take a few moments to read through the data. There are no misspellings, changes of case, duplicates, or special characters; the dataset looks clean! This means that this dataset will not require any cleaning once loaded into the DataFrame, which is convenient.

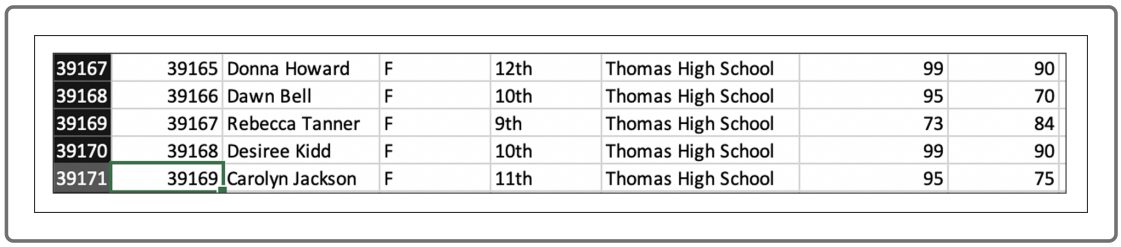
Next, open `students_complete.csv`. This is a large file! Here are the first 10 rows:

	A	B	C	D	E	F	G
1	Student ID	student_name	gender	grade	school_name	reading_score	math_score
2	0	Paul Bradley	M	9th	Huang High School	66	79
3	1	Victor Smith	M	12th	Huang High School	94	61
4	2	Kevin Rodriguez	M	12th	Huang High School	90	60
5	3	Dr. Richard Scott	M	12th	Huang High School	67	58
6	4	Bonnie Ray	F	9th	Huang High School	97	84
7	5	Bryan Miranda	M	9th	Huang High School	94	94
8	6	Sheena Carter	F	11th	Huang High School	82	80
9	7	Nicole Baker	F	12th	Huang High School	96	69
10	8	Michael Roth	M	10th	Huang High School	95	87

Now go to the last row of the file.

REWIND

To get to the last row of an Excel file, place the cursor in a column that doesn't have any empty cells and press Command + the down-arrow key (on a Mac) or CTRL + the down-arrow key (on Windows).



39167	39165	Donna Howard	F	12th	Thomas High School	99	90
39168	39166	Dawn Bell	F	10th	Thomas High School	95	70
39169	39167	Rebecca Tanner	F	9th	Thomas High School	73	84
39170	39168	Desiree Kidd	F	10th	Thomas High School	99	90
39171	39169	Carolyn Jackson	F	11th	Thomas High School	95	75

When we get to the last row, we see there are 39,170 rows (the first row is the header row, so it's not counted). Also, each row contains a student ID that is associated with a student's name, gender, grade in school, the name of the school they attend, and their reading and math scores. Students are grouped by school. Like `schools_complete.csv`, this file also includes the "school_name" in the header.

Retake