5.2.1

Import and Inspect CSV Files

The moment has finally come! You're ready to download the CSV files that are worth their weight in gold: two datasets containing four months of rideshare data, just waiting for you to unlock their secrets.

Import the Data

The first step is to import the data. To do that, follow these steps:

1. Click the following links to download the city_data.csv and ride_data.csv files into your Resources folder.

[Download city_data.csv (https://2u-data-curriculum-team.s3.amazonaws.com/dataviz-online/v2/module_5/city_data.csv)]

[Download ride_data.csv (https://2u-data-curriculum-team.s3.amazonaws.com/dataviz-online/v2/module_5/ride_data.csv)]

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

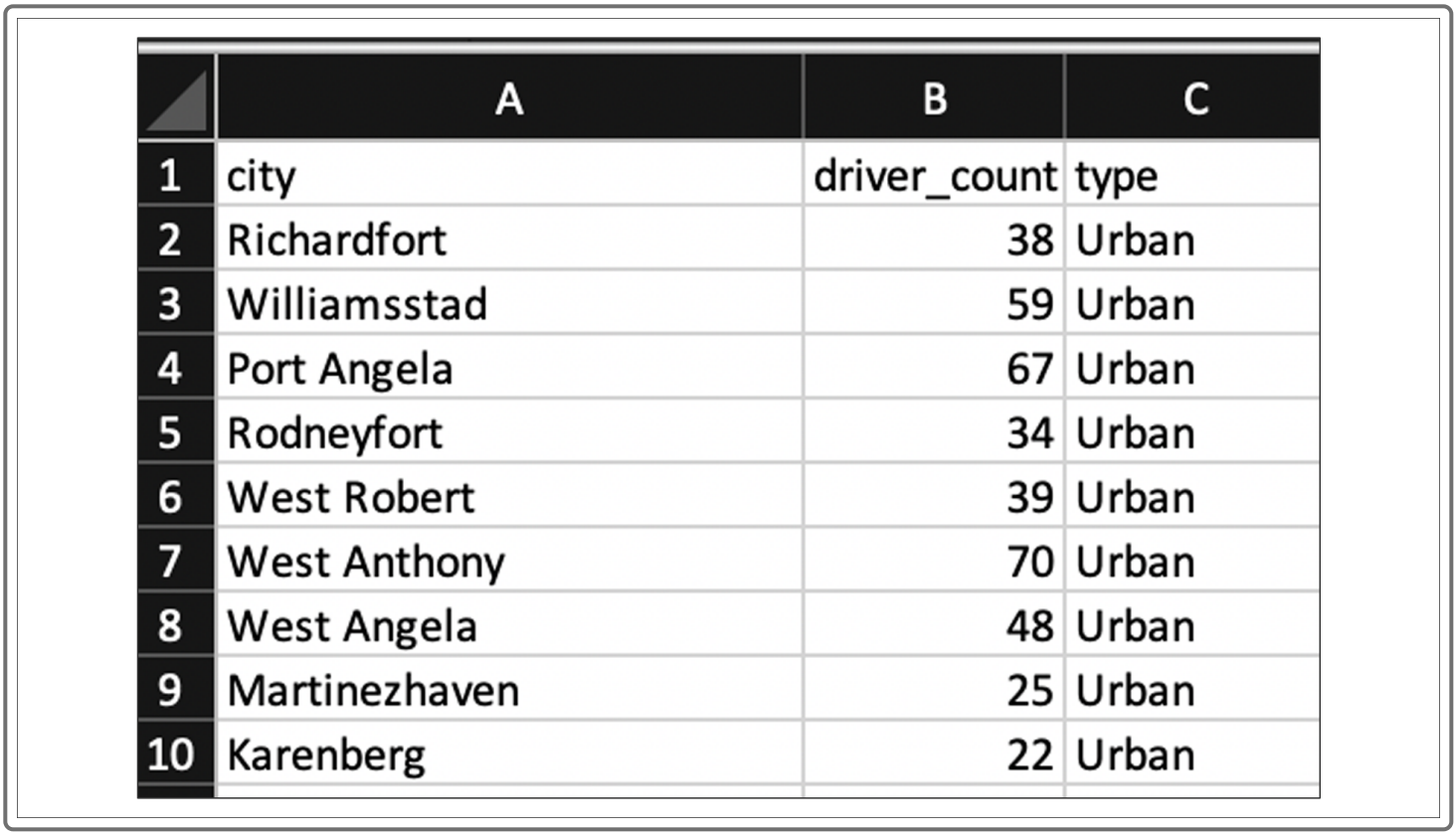
- `city_data.csv`
- `ride_data.csv`

Inspect the Data

Before we do any analysis, we will inspect the data and answer 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?

If you open the `city_data.csv` file, you will see three columns: city, driver_count, and type. Here's a snapshot of the first ten rows of data in this file:



	A	B	C
1	city	driver_count	type
2	Richardfort	38	Urban
3	Williamsstad	59	Urban
4	Port Angela	67	Urban
5	Rodneyfort	34	Urban
6	West Robert	39	Urban
7	West Anthony	70	Urban
8	West Angela	48	Urban
9	Martinezhaven	25	Urban
10	Karenberg	22	Urban

Let's see how much data is in this file. Can you remember how to jump to the end of a column in Excel?

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 + down arrow on a Mac, or CTRL+ down arrow on Windows.

Here's a snapshot of the last ten rows of the `city_data.csv` file:

112	Lake Jamie	4	Rural
113	Lake Latoyabury	2	Rural
114	North Jaime	1	Rural
115	South Marychester	1	Rural
116	Garzaport	7	Rural
117	Bradshawfurt	7	Rural
118	New Ryantown	2	Rural
119	Randallchester	9	Rural
120	Jessicaport	1	Rural
121	South Saramouth	7	Rural

We can see that there are 121 rows. Inspecting this file further, we notice that each column has a header. Each row contains a city that has a driver_count and the type of city: Urban, Suburban, or Rural.

Scrolling through this CSV file, we see that there are no empty rows. Once we add this CSV file into a Pandas DataFrame, we'll be able to determine the data type for each column.

If we open the `ride_data.csv` file, we can see four columns: city, date, fare, and ride_id. Here are the first 10 rows:

1	city	date	fare	ride_id
2	Lake Jonathanshire	1/14/19 10:14	13.83	5.74E+12
3	South Michelleport	3/4/19 18:24	30.24	2.34E+12
4	Port Samanthamouth	2/24/19 4:29	33.44	2.01E+12
5	Rodneyfort	2/10/19 23:22	23.44	5.15E+12
6	South Jack	3/6/19 4:28	34.58	3.91E+12
7	South Latoya	3/11/19 12:26	9.52	2.00E+12
8	New Paulville	2/27/19 11:17	43.25	7.93E+11
9	Simpsonburgh	4/26/19 0:43	35.98	1.12E+11
10	South Karenland	1/8/19 3:28	35.09	8.00E+12
11	North Jasmine	3/9/19 6:26	42.81	5.33E+12

And if we go to the end of the file, we can see that there are 2,376 rows:

2370	Lake Jamie	4/29/19 1:58	54.22	2.49E+12
2371	Bradshawfurt	1/30/19 10:55	51.39	1.33E+12
2372	Michaelberg	4/29/19 17:04	13.38	8.55E+12
2373	Lake Latoyabury	1/30/19 0:05	20.76	9.02E+12
2374	North Jaime	2/10/19 21:03	11.11	2.78E+12
2375	West Heather	5/7/19 19:22	44.94	4.26E+12
2376	Newtonview	4/25/19 10:20	55.84	9.99E+12

Looking more closely at this file, we notice that each column has a header. Each row contains a city that has a date when the ride was taken, the fare for the ride, and a 12-to-13 digit ride identification number.

There's no way to scroll through the `ride_data.csv` file efficiently, so we'll have to use Pandas to determine if there are empty rows and the data type for each column.