8.2.1

Open and Inspect the Crowdfunding Dataset

The crowdfunding dataset has lots of data about the crowdfunding projects. For each project, it includes an identification number, the company name, a blurb about the project, the goal amount, the amount pledged so far, the project outcome, the number of backers, a category-and-subcategory combination, and more. Before beginning the ETL process, you need to inspect the dataset.

To get started inspecting the dataset, complete the following steps:

1. Create a new GitHub repository named Crowdfunding-ETL.
2. Navigate to your class folder, and then clone your new repo within the folder.
3. Download the crowdfunding Excel file

(https://2u-data-curriculum-team.s3.amazonaws.com/dataviz-online/v2/module_8/crowdfunding.xlsx) , named `crowdfunding.xlsx` , to your class folder.

4. Open the file, and then inspect the data that it contains.

Notice that the crowdfunding Excel file has two worksheets, named `crowdfunding_info` and `contact_info.`

The `crowdfunding_info` worksheet has a header row followed by 10 rows of data. The header row consists of the following labels: "cf_id," "company_name," "blurb," "goal," "pledged," "outcome," "backers_count," "country," "currency," "launched_at," "deadline," "staff_pick," "spotlight,", and "category & sub-category."

	cf_id	company_name	blurb	goal	pledged	outcome	backers_count	country	currency	launched_at	deadline	staff_pick	spotlight	category & sub-category
2	147	Baldwin, Riley and Jackson	Pre-emptive tertiary standardization	100	0	failed	0	CA	CAD	1581573600	1614578400	FALSE	FALSE	food/food trucks
3	1621	Odom Inc	Managed bottom-line architecture	1400	14560	successful	158	US	USD	1611554400	1621918800	FALSE	TRUE	music/rock
4	1812	Melton, Robinson and Fritz	Function-based leadingedge pricing structure	108400	142523	successful	1425	AU	AUD	1608184800	1640844000	FALSE	FALSE	technology/web
5	2156	Mcdonald, Gonzalez and Ross	Vision-oriented fresh-thinking conglomeration	4200	2477	failed	24	US	USD	1634792400	1642399200	FALSE	FALSE	music/rock
6	1365	Larson-Little	Proactive foreground core	7600	5265	failed	53	US	USD	1608530400	1629694800	FALSE	FALSE	theater/plays
7	2057	Harris Group	Open-source optimizing database	7600	13195	successful	174	DK	DKK	1607666400	1630213200	FALSE	FALSE	theater/plays
8	1894	Ortiz, Coleman and Mitchell	Operative upward-trending algorithm	5200	1090	failed	18	GB	GBP	1596171600	1620709200	FALSE	FALSE	film & video/documentary
9	2669	Carter-Guzman	Centralized cohesive challenge	4500	14741	successful	227	DK	DKK	1608616800	1632200400	FALSE	FALSE	theater/plays
10	1114	Nunez-Richards	Exclusive attitude-oriented intranet	110100	21946	live	708	DK	DKK	1586322000	1615356000	FALSE	FALSE	theater/plays

Take a moment to review the data. The "cf_id" column contains the crowdfunding identification number for each project. The "goal" and "pledged" columns contain dollar amounts. The "launched_at" and "deadline" columns contain the time of day, in seconds. The "staff_pick" and "spotlight" columns contain Boolean values. And, the remaining columns contain text. We'll further explore the data later in this lesson.

REWIND

A time of day in seconds is a database timestamp, which people also refer to as epoch time.

Moving on to the `contact_info` worksheet, notice that the first two rows have information about the data on the sheet. Then come two blank rows, a header row that's labeled "contact_info," and six rows of data.

1	This list of contacts was updated on 11/10/2020.
2	Note: The contact information needs to be separated into the following columns: contact_id, first name, last name, and email.
3	
4	contact_info
5	{"contact_id": 4661, "name": "Cecilia Velasco", "email": "cecilia.velasco@rodrigues.fr"}
6	{"contact_id": 3765, "name": "Mariana Ellis", "email": "mariana.ellis@rossi.org"}
7	{"contact_id": 4187, "name": "Sofie Woods", "email": "sofie.woods@riviere.com"}
8	{"contact_id": 4941, "name": "Jeanette Iannotti", "email": "jeanette.iannotti@yahoo.com"}
9	{"contact_id": 2199, "name": "Samuel Sorgatz", "email": "samuel.sorgatz@gmail.com"}
10	{"contact_id": 5650, "name": "Socorro Luna", "email": "socorro.luna@hotmail.com"}

Take a moment to review the data. Each of the six rows of data has a dictionary which, in turn, contains "contact_id", "name", and "email" keys along with a value for each key. Later, when we load the worksheet into a DataFrame, we'll have to determine the data type of the rows and assess whether these are Python dictionary key-value pairs or text.

Now that we've inspected the crowdfunding dataset, it's time to prepare for the ETL project. Let's do that next.