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

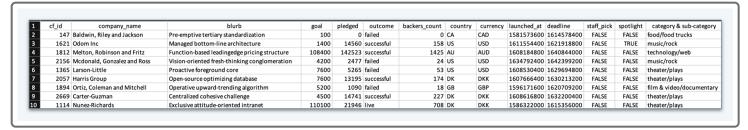


(https://2u-data-curriculum-team.s3.amazonaws.com/dataviz-online/v2/module\_8/crowdfunding.xlsx), named crowdfunding.xlsx), to your class folder.

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

Notice that the crowdfunding Excel file has two worksheets, named <a href="mailto:crowdfunding\_info">crowdfunding\_info</a> and <a href="mailto:contact\_info">contact\_info</a>.

The <u>crowdfunding\_info</u> 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."



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 lannotti", "email": "jeanette.iannotti@yahoo.com"}
9	{"contact_id": 2199, "name": "Samuel Sorgatz", "email": "samuel.sorgatz@gmail.com"}
LO	{"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.

© 2020 - 2022 Trilogy Education Services, a 2U, Inc. brand. All Rights Reserved.