8.5.1

Identify the Data Relationships

Now that you and Britta have transformed the data, you need to move to the next phase in the ETL process: load. Specifically, you'll load the data into a database. You have the data—the four CSV files. But, you don't have a database or a table for each file.

Before you and Britta can create the database and the tables, you need to review each CSV file to find out how the CSV files are connected so you can create a map of the database and table schemas.

Try to find the four CSV files that you exported during the transform phase. If you can't find them, that's OK. You can download the following CSV files to your computer and add them to your Crowdfunding-ETL repository:

- category.csv (https://2u-data-curriculum-team.s3.amazonaws.com/dataviz-online/v2/module_8/category.csv)
- subcategory.csv (https://2u-data-curriculum-team.s3.amazonaws.com/dataviz-online/v2/module_8/subcategory.csv)
- campaign.csv (https://2u-data-curriculum-team.s3.amazonaws.com/dataviz-online/v2/module_8/campaign.csv)
- contacts.csv (https://2u-data-curriculum-team.s3.amazonaws.com/dataviz-online/v2/module_8/contacts.csv)

Now that you have the four CSV files, open the `category.csv` dataset, and then review the data.

	category_id	category
2	cat01	food
3	cat02	music
4	cat03	technology
5	cat04	theater
6	cat05	film & video
7	cat06	publishing
8	cat07	games
9	cat08	photography
10	cat09	journalism

This table doesn't contain lots of data—only 2 columns and 10 rows.

Next, review the `subcatecory.csv` dataset.

	subcategory_id	subcategory
2	scat01	food trucks
3	scat02	rock
4	scat03	web
5	scat04	plays
6	scat05	documentary
7	scat06	electric music
8	scat07	drama
9	scat08	indie rock
10	scat09	wearables
11	scat010	nonfiction
12	scat011	animation
13	scat012	video games
14	scat013	shorts
15	scat014	fiction
16	scat015	photography books
17	scat016	radio & podcasts
18	scat017	metal
19	scat018	jazz
20	scat019	translations
21	scat020	television
22	scat021	mobile games
23	scat022	world music
24	scat023	science fiction
25	scat024	audio

Like with the `category.csv` dataset, this table doesn't contain lots of data—only 2 columns and 25 rows.

Before moving on, check your knowledge with the following assessment:

Practice identifying the primary keys of these datasets yourself in the following Skill Drill:

SKILL DRILL

Can you identify the primary keys in the `category.csv` and `subcategory.csv` datasets? Recall that a primary key is a unique identifier for each row in a table.

The primary key in the `category.csv` dataset is the "category_id" column. The primary key in the `subcategory.csv` dataset is the "subcategory_id" column.

Next, review the `campaign.csv` dataset.

	cf_id	contact_id	company_name	description	goal	pledged	outcome	backers_count	country	currency	launched_date	end_date	category_id	subcategory_id
2	147	4661	Baldwin, Riley an	Pre-emptive tertia	100	0	failed	0	CA	CAD	2/13/20	3/1/21	cat01	scat01
3	1621	3765	Odom Inc	Managed bottom-	1400	14560	successful	158	US	USD	1/25/21	5/25/21	cat02	scat02
4	1812	4187	Melton, Robinson	Function-based le	108400	142523	successful	1425	AU	AUD	12/17/20	12/30/21	cat03	scat03
5	2156	4941	Mcdonald, Gonzal	Vision-oriented fre	4200	2477	failed	24	US	USD	10/21/21	1/17/22	cat02	scat02
6	1365	2199	Larson-Little	Proactive foregrou	7600	5265	failed	53	US	USD	12/21/20	8/23/21	cat04	scat04
7	2057	5650	Harris Group	Open-source optin	7600	13195	successful	174	DK	DKK	12/11/20	8/29/21	cat04	scat04
8	1894	5889	Ortiz, Coleman an	Operative upward	5200	1090	failed	18	GB	GBP	7/31/20	5/11/21	cat05	scat05
9	2669	4842	Carter-Guzman	Centralized cohesi	4500	14741	successful	227	DK	DKK	12/22/20	9/21/21	cat04	scat04
10	1114	3280	Nunez-Richards	Exclusive attitude-	110100	21946	live	708	DK	DKK	4/8/20	3/10/21	cat04	scat04
11	970	5468	Rangel, Holt and J	Open-source fresh	6200	3208	failed	44	US	USD	8/13/21	8/31/21	cat02	scat06
12	2340	3064	Green Ltd	Monitored empow	5200	13838	successful	220	US	USD	7/11/20	8/2/21	cat05	scat07
13	601	4904	Perez, Johnson an	Grass-roots zero a	6300	3030	failed	27	US	USD	8/11/20	6/26/21	cat04	scat04
14	1950	1299	Kim Ltd	Assimilated hybric	6300	5629	failed	55	US	USD	11/14/20	4/9/21	cat05	scat07
15	671	5602	Walker, Taylor an	Multi-tiered direct	4200	10295	successful	98	US	USD	11/11/20	11/6/21	cat02	scat08

Before moving on, check your knowledge in the following assessment:

Next, review the `contacts.csv` dataset.

1	contact_id	first_name	last_name	email
2	4661	Cecilia	Velasco	cecilia.velasco@rodrigues.fr
3	3765	Mariana	Ellis	mariana.ellis@rossi.org
4	4187	Sofie	Woods	sofie.woods@riviere.com
5	4941	Jeanette	Iannotti	jeanette.iannotti@yahoo.com
6	2199	Samuel	Sorgatz	samuel.sorgatz@gmail.com
7	5650	Socorro	Luna	socorro.luna@hotmail.com
8	5889	Carolina	Murray	carolina.murray@knight.com
9	4842	Kayla	Moon	kayla.moon@yahoo.de
10	3280	Ariadna	Geisel	ariadna.geisel@rangel.com

Did you notice that both the `contacts.csv` and `campaign.csv` datasets contain the "contact_id" column, which has unique contact ID numbers?

Let's figure out which of these datasets has the "contact_id" column as the primary key. We know that "cf_id" is the primary key in the `campaign.csv` dataset. This means that the "contact_id" column in the `contacts.csv` dataset is the primary key.

SKILL DRILL

Can you identify all the foreign keys in the `campaign.csv` dataset? Recall that a foreign key references the primary key of another dataset.

The foreign keys in the `campaign.csv` dataset are the "contact_id," "category_id," and "subcategory_id" columns.

Terrific job identifying the primary and the foreign keys in the datasets! Next, we'll create an entity relationship diagram (ERD).