**Crowdfunding\_ETL**

Project 2 Data Analytics Challenge

For the ETL mini project, we have worked in a team to practice building an ETL pipeline using Python, Pandas, and either Python dictionary methods or regular expressions to extract and transform the data. After transforming the data, we have created four CSV files and used the CSV file data to create an ERD and a table schema. Lastly, we have uploaded the CSV file data into a Postgres database.​Before starting the work on the project:​

1. Before we got started, we downloaded the starter code and files.
2. One memeber of the team created a new repository named Crowdfunding\_ETL.
3. Each member has cloned the new repository to their computer.
4. One person renamed the ETL\_Mini\_Project\_starter\_code.ipynb file with the first name initial and last name of each member of the group, to ETL\_Mini\_Project\_SPribojac\_PLehel\_IPirogan.
5. We have added this Jupyter notebook file and the Resources folder containing the crowdfunding.xlsx and the contacts.xlsx files to the new repository.
6. We pushed the changes to GitHub.
7. Each team member have pulled the changes, so both all of us have the same notebook available on our computers.

**Instructions:**

The instructions for this mini project are divided into the following subsections:

* Create the Category and Subcategory DataFrames
* Create the Campaign DataFrame
* Create the Contacts DataFrame
* Create the Crowdfunding Database​

**Create the Category and Subcategory DataFrames:**

1. Extract and transform the crowdfunding.xlsx Excel data to create a category DataFrame that has the following columns:
   * A "category\_id" column that has entries going sequentially from "cat1" to "catn", where n is the number of unique categories
   * A "category" column that contains only the category titles
2. Export the category DataFrame as category.csv and save it to your GitHub repository.​
3. Extract and transform the crowdfunding.xlsx Excel data to create a subcategory DataFrame that has the following columns:
   * A "subcategory\_id" column that has entries going sequentially from "subcat1" to "subcatn", where n is the number of unique subcategories
   * A "subcategory" column that contains only the subcategory titles
4. Export the subcategory DataFrame as subcategory.csv and save it to your GitHub repository.

**Create the Campaign DataFrame:**

​

1. Extract and transform the crowdfunding.xlsx Excel data to create a campaign DataFrame has the following columns:

* The "cf\_id" column
* The "contact\_id" column
* The "company\_name" column
* The "blurb" column, renamed to "description"
* The "goal" column, converted to the float data type
* The "pledged" column, converted to the float data type
* The "outcome" column
* The "backers\_count" column
* The "country" column
* The "currency" column
* The "launched\_at" column, renamed to "launch\_date" and with the UTC times converted to the datetime format
* The "deadline" column, renamed to "end\_date" and with the UTC times converted to the datetime format
* The "category\_id" column, with unique identification numbers matching those in the "category\_id" column of the category DataFrame
* The "subcategory\_id" column, with the unique identification numbers matching those in the "subcategory\_id" column of the subcategory DataFrame​

1. Export the campaign DataFrame as campaign.csv and save it to your GitHub repository.​

**Create the Contacts DataFrame:**

​1.Use Python dictionary methods for extracting and transforming the data from the contacts.xlsx Excel data:​

* Import the contacts.xlsx file into a DataFrame.
* Iterate through the DataFrame, converting each row to a dictionary.
* Iterate through each dictionary, doing the following:
* Extract the dictionary values from the keys by using a Python list comprehension.
* Add the values for each row to a new list.
* Create a new DataFrame that contains the extracted data.
* Split each "name" column value into a first and last name, and place each in a new column.
* Clean and export the DataFrame as contacts.csv and save it to your GitHub repository.​

**Create the Crowdfunding Database:**

​

1. Inspect the four CSV files, and then sketch an ERD of the tables by using QuickDBDLinks <http://www.quickdatabasediagrams.com/>​
2. Use the information from the ERD to create a table schema for each CSV file. Remember to specify the data types, primary keys, foreign keys, and other constraints.​
3. Save the database schema as a Postgres file named crowdfunding\_db\_schema.sql, and save it to your GitHub repository.​
4. Create a new Postgres database, named crowdfunding\_db.​
5. Using the database schema, create the tables in the correct order to handle the foreign keys.​
6. Verify the table creation by running a SELECT statement for each table.​
7. Import each CSV file into its corresponding SQL table.​
8. Verify that each table has the correct data by running a SELECT statement for each.​​----------------------------------------------------------------------------

**Credits**

​---------------------------------------------------------------------------- We have used the resources provided in class, such as class material, homework and extra resources to support our understanding and completion of the project.​Credits to Sophie Pribojac, Peter Lehel, Iana Pirogan.​