

## **Project 2 Crowdfunding\_ETL: Writeup**

For this project, our team came together to organize and process data from two main resources: the “crowdfunding.xlsx” and “contacts.xlsx” files. We started by creating two tables, one for categories and another for subcategories. Each unique category and subcategory was assigned a sequential ID, which helped us create structured and easy to use DataFrames. These were then exported as CSV files to be used in later steps.

Next, we worked on the crowdfunding campaign data, transforming it into a detailed and organized campaign DataFrame. This included key details like campaign ID, contact ID, company name, description, funding goal, pledged amount, outcome, number of backers, country, and currency. We also converted the launch and end dates into an easy to read format and renamed them as "launch\_date" and "end\_date" for better clarity. To ensure everything was linked, we matched each campaign to the correct category and subcategory using unique IDs. Once cleaned, this DataFrame was also exported as a CSV file.

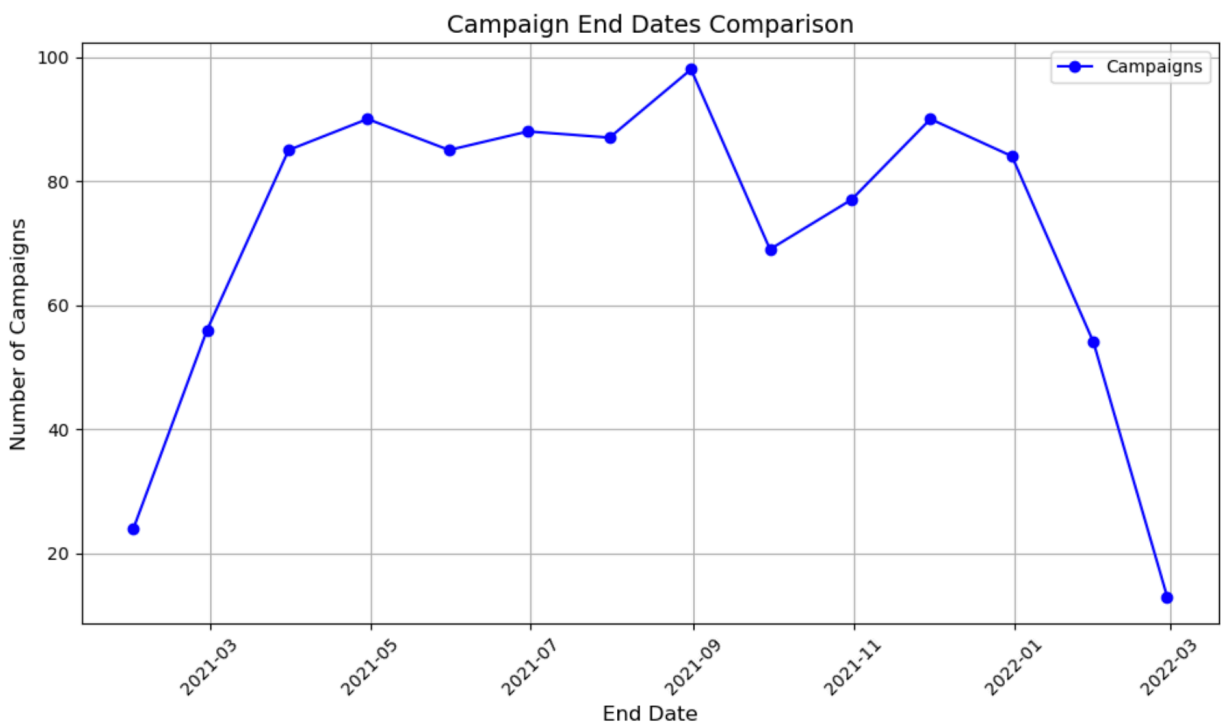
For the contacts data, we took the time to extract and clean the information carefully. We split full names into "first\_name" and "last\_name" columns and kept email addresses intact. The result was a clear and well structured contacts data frame, which we exported as another CSV file.

To bring all the pieces together, we designed an entity-relationship diagram (ERD) to show how the different data tables were connected. Using this as a guide, we created a PostgreSQL database schema with primary and foreign key constraints to ensure everything was properly linked. After setting up the crowdfunding\_db database, we imported the CSV files into their respective tables and double checked that everything was accurate using SELECT queries.

With this database in place, we now have a solid data to analyze and uncover meaningful insights in the future. For example, we can compare pledged amounts to funding goals to identify the types of campaigns that perform best. We can also look at popular categories and subcategories to see which niches attract the most backers or pass their funding goals. Time based patterns, like the best times of the year to launch campaigns, and geographical trends, such as which countries

have the highest success rates, are other areas to explore. These analyses can help us understand what makes a crowdfunding campaign successful.

To wrap things up, we created a Jupyter notebook where we ran queries and visualized the data, generating charts and graphs to highlight key trends in campaign activity and funding distribution. The visualizations revealed that campaign end dates peaked in mid 2021, suggesting a seasonal surge, followed by a steep decline in early 2022, possibly due to external factors reducing campaign completions. In terms of funding distribution, theater campaigns dominated, accounting for 34.3% of campaigns pledging over 1500, highlighting their strong appeal or funding needs. Other notable categories include film & video (18.1%), music (17.0%), and technology (9.9%), showcasing a diverse range of interests among backers, with creative and technological projects being particularly popular. Smaller shares were observed for publishing, games, and photography, indicating niche appeal.



Distribution of Categories for Campaigns that Pledged Over 1500

