Group 04:

Write UP

The main goal of this project was to enhance knowledge in ETL. This mini project involves extracting data from multiple sources, cleaning and transforming the data using Jupyter Notebook with pandas, numpy, and loading the cleaned data into a relational database using pgAdmin.

For this project we created four CSV files, and then used the CSV files to create an ERD and a table schema. As a final stage, we loaded the CSV files into a PostgreSQL database.

During this project we created DataFrames :

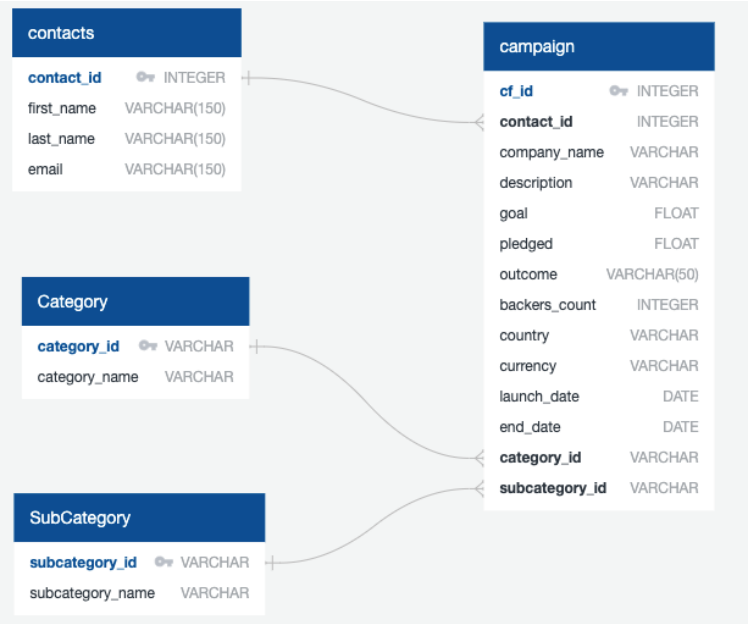
- Category. Where "category_id" column that has entries going sequentially from "cat1" to "catn", and n is the number of unique categories

- Subcategory. Where subcategory_id" column that has entries going sequentially from "subcat1" to "subcatn", where n is the number of unique subcategories

- Campaign

- Contacts. Two create this dataframe we used two options : Python dictionary methods and used regular expressions.

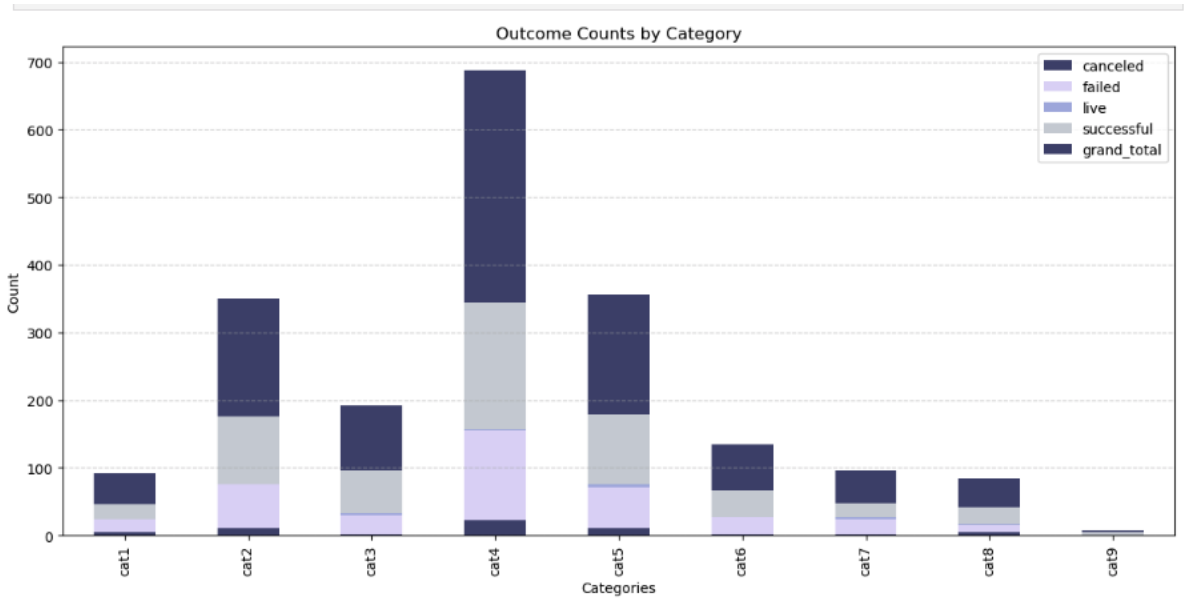
And Crowdfunding Database. We used information from ERD create a table schema for each CSV file. By using ERD we easily identify connections (FK and PK) between our existing tables



Analysis :

During our analysis we created the following graphs:

The first graph shows the breakdown of campaigns by category. Category 4 has the highest number of successful campaigns with 187, followed by cat5 with 102 successful campaigns. Category 4 also has the highest number of failed campaigns with 132, followed by cat2 with 66 failed campaigns. Category 4 has the highest number of canceled campaigns with 23, followed by cat5 with 11 canceled campaigns.

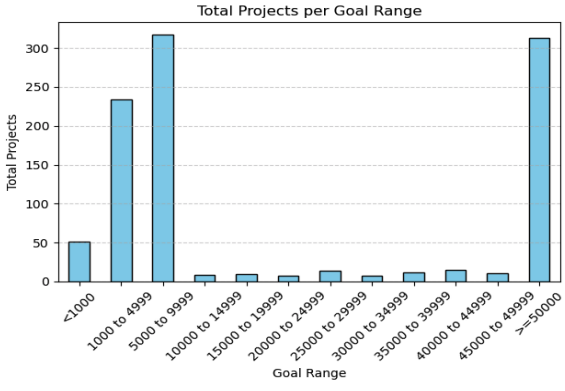


	row_labels	canceled	failed	live	successful	grand_total
0	cat1	4	20	0	22	46
1	cat2	10	66	0	99	175
2	cat3	2	28	2	64	96
3	cat4	23	132	2	187	344
4	cat5	11	60	5	102	178
5	cat6	2	24	1	40	67
6	cat7	1	23	3	21	48
7	cat8	4	11	1	26	42
8	cat9	0	0	0	4	4

From the second graph and table it is clear that the

- success rates vary significantly across different goal ranges. For example, campaigns with goals between 15,000 and 24,999 have a 100% success rate, while campaigns with goals between 10,000 and 14,999 have a success rate of only 44%.
- The majority of projects fall within the goal range of 1,000 to 4,999, with a total of 234 projects in this range. This indicates that most campaigns on the platform have relatively modest funding goals
- Campaigns with lower funding goals tend to have higher success rates. For instance, campaigns with goals between 1,000 and 4,999 have a success rate of 82%, while campaigns with goals over 50,000 have a success rate of only 36%.
- Most goal ranges have low cancellation rates, with the highest cancellation rate (8%) observed in the 5,000 to 9,999 goal range. This indicates that campaigns in this range may face challenges leading to cancellations.

	goal_range	number_successful	number_failed	number_canceled	total_projects	percentage_successful	percentage_failed
0	<1000	30	20	1	51	59%	39%
1	1000 to 4999	191	38	2	234	82%	16%
2	5000 to 9999	164	126	25	317	52%	40%
3	10000 to 14999	4	5	0	9	44%	56%
4	15000 to 19999	10	0	0	10	100%	0%
5	20000 to 24999	7	0	0	7	100%	0%
6	25000 to 29999	11	3	0	14	79%	21%
7	30000 to 34999	7	0	0	7	100%	0%
8	35000 to 39999	8	3	1	12	67%	25%
9	40000 to 44999	11	3	0	15	73%	20%
10	45000 to 49999	8	3	0	11	73%	27%
11	>=50000	114	163	28	313	36%	52%



In this project, we illustrate the practical application of ETL processes in real-world scenarios involving data manipulation and storage.