Project 2 ETL Write Up

Matthew Adent, Tyler Beringer, Victor Rincon

Purpose

The purpose of this project was to successfully design and implement an ETL pipeline for our given crowdfunding data. The reason that we want to do this is because by doing so, we can get some useful insights about statistics such as campaign success rates, high-performing categories and sub-categories, and backer behavior within our dataset. Our plan to achieve this was to take in the raw data from excel spreadsheets in .xlsx format, clean it and plan its database structure out via an ERD, load it into a relational database, then make it ready for analysis.

ETL Pipeline

For the extraction step, we imported the raw data from two excel files: crowdfunding.xlsx and contacts.xlsx. The crowdfunding.xlsx file contained data about campaigns including the company name, a blurb about the campaign, the money goal and pledge amount, the outcome and backer count, the category and subcategory, and the launch and deadline date, along with a few other columns. Contacts.xlsx contained rows of contacts in the format of contact_id, name, and email.

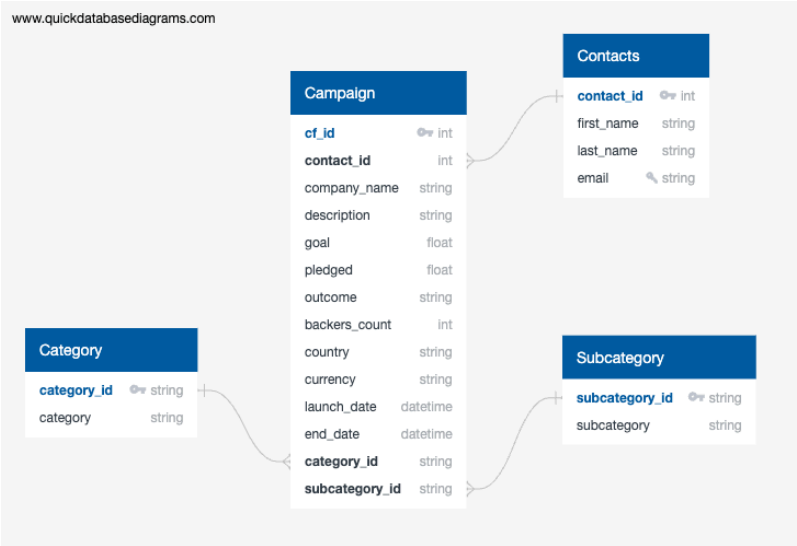
During the transformation process, we cleaned and restructured our data to make it more usable for analysis and put it into our relational database. One of the key transformations we made was splitting the category and subcategory column in crowdfunding.xlsx in order to have them be separate attributes, and we assigned them unique IDs to fit the structure of our database. As for

contacts.xlsx, we wanted to make contact_id, first and last name, and email their own attributes for each row, so we split the unorganized rows that contained the data using a lambda function and some string splitting. The screenshots below showed the form of our data after the transformation process was complete.

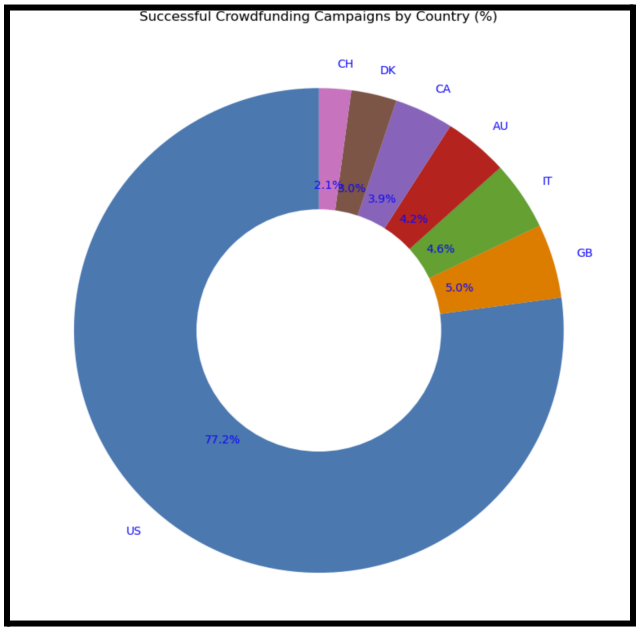
	cf_id	contact_id	company_name	description	goal	pledged	outcome	backers_count	country	currency	launched_date	end_date	category_id	subcategory_id
0	147	4661	Baldwin, Riley and Jackson	Pre-emptive tertiary standardization	100.0	0.0	failed	0	CA	CAD	2020-02-13	2021-03-01	category1	subcategory1
1	1621	3765	Odom Inc	Managed bottom-line architecture	1400.0	14560.0	successful	158	US	USD	2021-01-25	2021-05-25	category2	subcategory2
2	1812	4187	Melton, Robinson and Fritz	Function-based leadingedge pricing structure	108400.0	142523.0	successful	1425	AU	AUD	2020-12-17	2021-12-30	category3	subcategory3
3	2156	4941	Mcdonald, Gonzalez and Ross	Vision-oriented fresh-thinking conglomeration	4200.0	2477.0	failed	24	US	USD	2021-10-21	2022-01-17	category2	subcategory2
4	1365	2199	Larson-Little	Proactive foreground core	7600.0	5265.0	failed	53	US	USD	2020-12-21	2021-08-23	category4	subcategory4

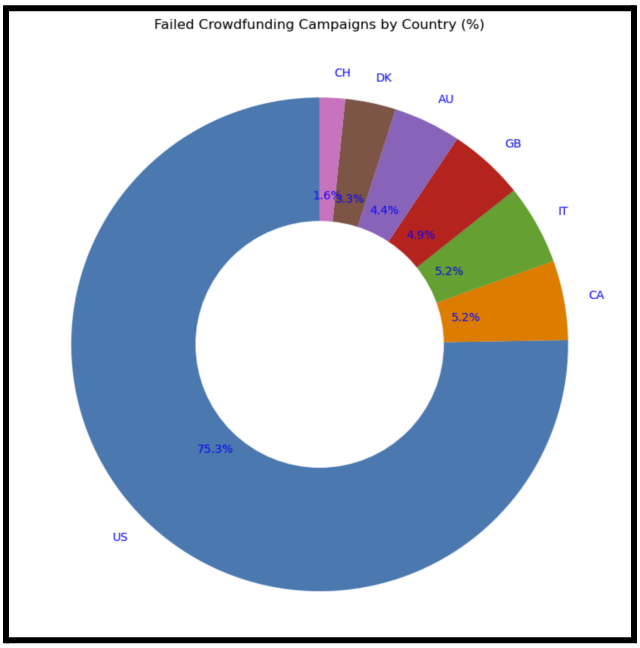
	contact_id	first_name	last_name	email
0	4661	cecilia	velasco	cecilia.velasco@rodrigues.fr
1	3765	mariana	ellis	mariana.ellis@rossi.org
2	4187	sofie	woods	sofie.woods@riviere.com
3	4941	jeanette	iannotti	jeanette.iannotti@yahoo.com
4	2199	samuel	sorgatz	samuel.sorgatz@gmail.com

We then loaded this data into our PostgreSQL relational database following our ERD. We had four tables: Category, Subcategory, Campaign, and COntacts. Below is our ERD.

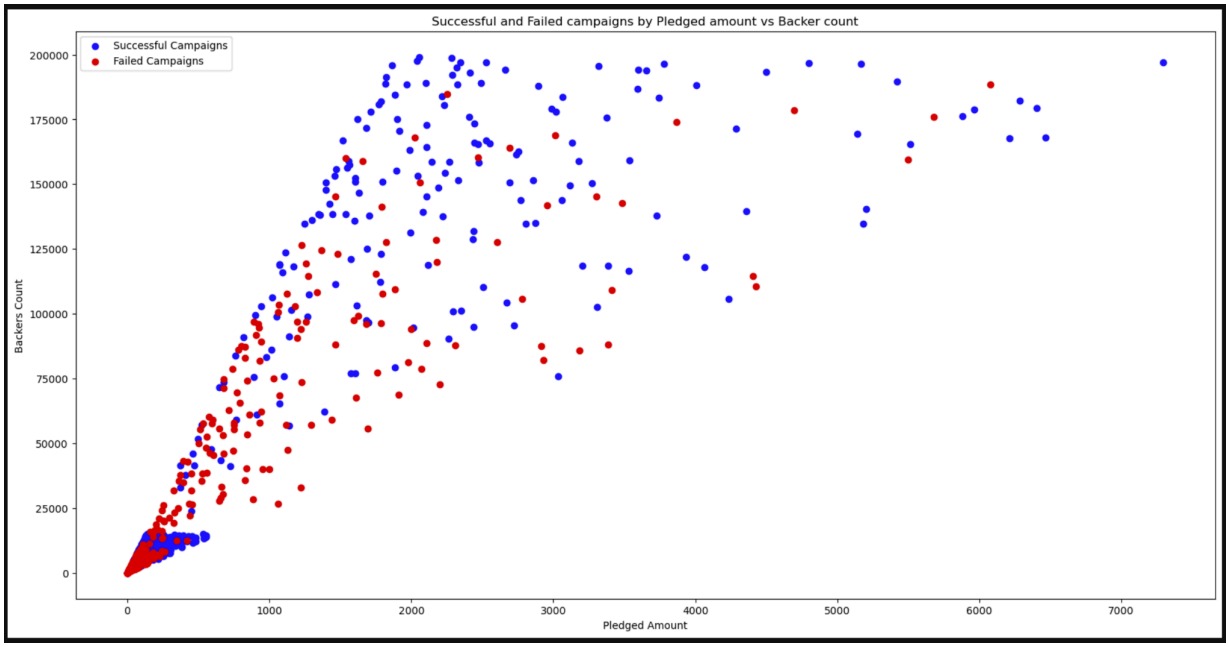


Results





In this dataset, the U.S. had the most campaigns. This is evident because while they made up 77.2% of successful campaigns, they also made up a whopping 75.3% of failed campaigns. The U.S. and China succeeded more than they failed, while the rest pictured did not.



This shows a positive correlation between backers and pledged amounts. We can also see a lot more blue than red as pledge amount and backer count increase together, suggesting that these variables together contribute to success. However, we can see some campaigns with high pledges and backers can fail, so it's not a definitive fact.

Conclusions and Future Work

This project gave us valuable insights into campaign performance, such as country campaign success rate and the correlation between a campaign's success and its backers count and pledged amount based on historical data. In the future, this pipeline could be productionalized by automating the ingestion of new crowdfunding data and using scheduled scripts to perform ETL seamlessly. From there, we'd need to regularly perform data validation and monitor its processes and outputs to ensure it remains accurate.