Analysis Of Crowdfunding

Team:

Carlos Ruiz
George Stalker
Joy Weishan

Introduction:

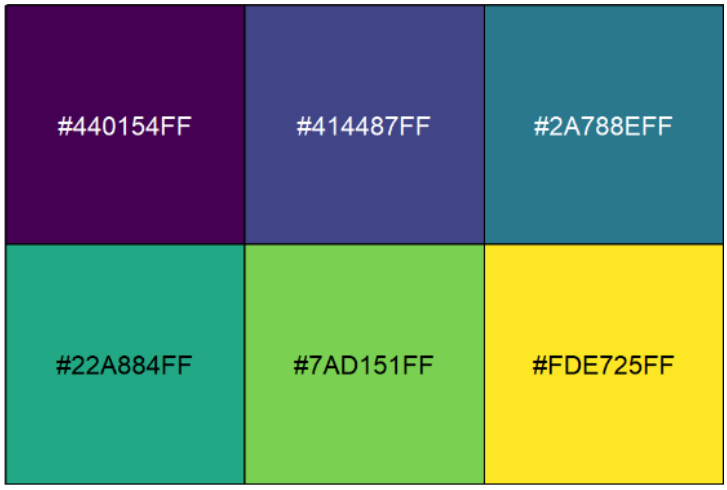
This project is to demonstrate building an ETL pipeline. This will include extracting data from crowdfunding files, data transformation, creating a database and table schema with keys and indexes. Post transformation and database creation, the team will load the database via importing transformed data via an appliation. The team will then extract the data via SQL and/or ORM for data analysis and convert the data into data visualizations as well as include analysis of the data.

Project Scope:

This project will produce a crowdfunding database schema, SQL, database image, 4 transformed data files in CSV format. Post transformation the data will be loaded to the database using an application layer. The project as well shows data analysis using Python and as well demonstration of data retrieval using FLASK with a application layer and SQL layer to separate SQL layer from the application layer. Post data imports, the data was extracted using SQL and/or ORM for analysis and visualizations.

Color Palette:

We used a viridis color pallet for consistency across the graphs.



Dataset Cleaning Steps:

Campaign:

Renamed columns for clarity, such as rename of “blurb” to “description” and “cf_id” to “funding_id”. Renamed columns for consistency across the database. Converted goal and pledged datatypes to floats for better data analysis. Converted date columns to date time for importing to the database as a timestamp and proper data analysis.

Category/Subcategory:

The Category and Subcategory columns were originally in one column and had to be split into two columns and added the prefix “cat” and “subcat” to each so that the join to the campaign table would work. Also, provided column naming consistency across the database. Also insured uniqueness in both the category and subcategory tables. Merged the campaign with category and subcategory to get the ID for further joins on the database during analysis. Dropped category and subcategory columns after the merge as database will be normalized with category subcategory tables.

Contact:

JSON is effective in creating a data frame with the input provided. This project also included the use of regular expressions to find the proper data within a column to extract and assign to the proper columns. Therefore, this project demonstrated JSON and regular expressions in the code. The JSON data was used to load to the database used during analysis. Either file could have been used as the end result is the same formatted data file. The name column was split into first name and last name so that proper data analysis and data retrieval can be done as needed. The full name column was dropped as not needed. The order of the columns was modified to match the database table schema prior to import.

Crowdfunding Analysis and Visualizations:

The top pledged stock pick or spotlight that reached the funding goal was Company Perez Group from the US, it launched on 05/01/2021 and ended 03/11/2021. The category was publishing and the subcategory is translations. If you would like more information the contact for this project is Linares Severino; email: severino@linares@angeli.com

	first_name	last_name	email	company_name	description	goal	pledged	outcome	backers_count	country	currency	launched_date	enc
0	Severino	Linares	severino.linares@angeli.com	Perez Group	Cross-platform tertiary hub	63400.0	197728.0	successful	2038	US	USD	2021-05-01 05:00:00	202 05
1	Roberto	Guyot	roberto.guyot@bennett.com	Hicks, Wall and Webb	Managed discrete framework	174500.0	197018.0	successful	2526	US	USD	2020-04-08 05:00:00	202 05
2	Modesto	Wright	modesto.wright@pareto.com	Baker, Collins and Smith	Reactive real-time software	114400.0	196779.0	successful	4799	US	USD	2021-09-13 05:00:00	202 06
3	Darren	Bernardi	darren.bernardi@brooks-martin.com	Santana-George	Re-engineered client-driven knowledge user	71500.0	194912.0	successful	2320	US	USD	2021-03-16 05:00:00	202 05
4	Gebhard	Thanel	gebhard.thanel@gmail.com	Clarke, Anderson and Lee	Robust explicit hardware	191200.0	191222.0	successful	1821	US	USD	2021-05-22 05:00:00	202 05

When looking at the two categories with the most campaigns; Theater and Film&Video the most successful campaign is film&video with a 57% success rate over theater with a 54% success rate. When we look at top categories with similar numbers of campaigns; film & video and music, film & video is still more successful than music with 56%..

	category	canceled	failed	live	successful	grand_total
0	theater	23	132	2	187	344
1	film & video	11	60	5	102	178
2	music	10	66	0	99	175
3	technology	2	28	2	64	96
4	publishing	2	24	1	40	67

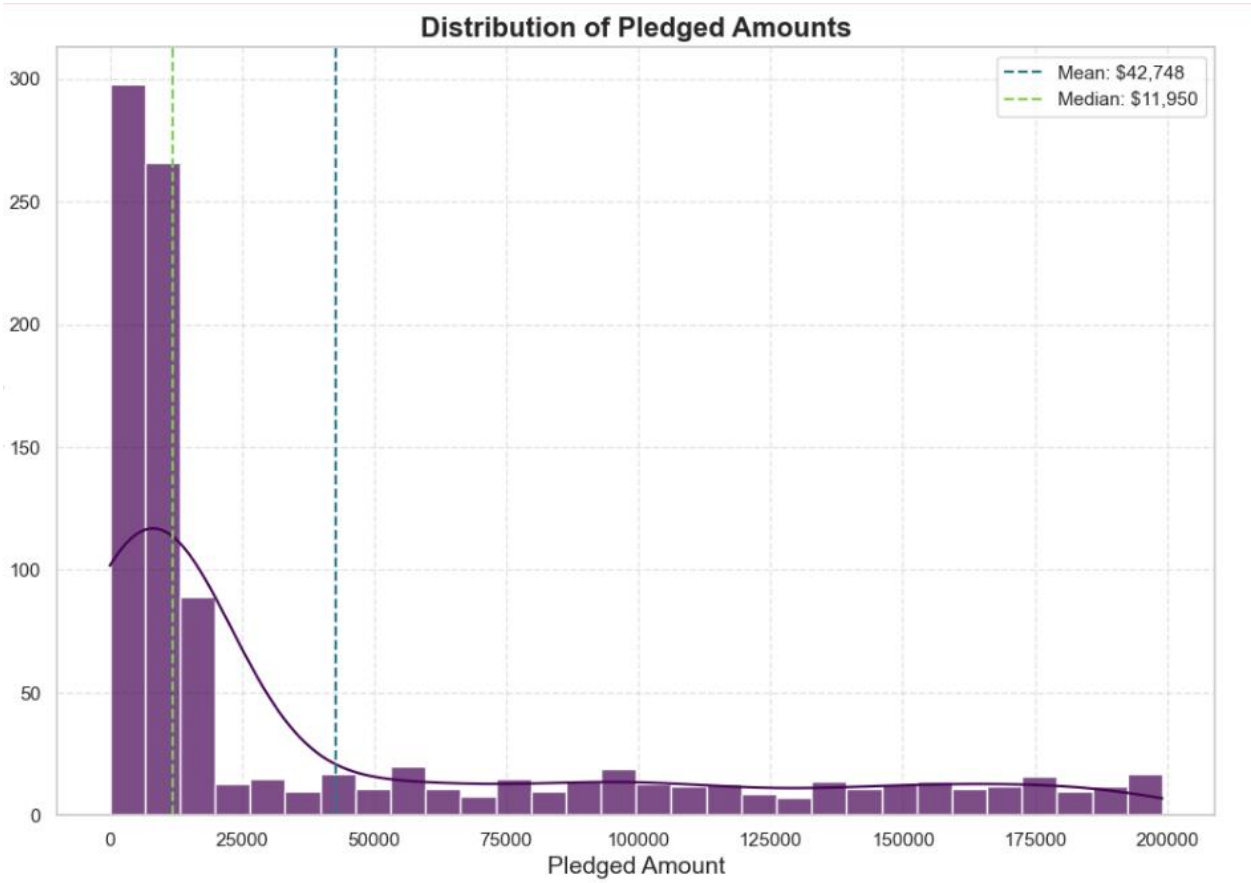
When looking at the two subcategories with the most campaigns; plays and rock the most successful campaign is rock with a 58% success rate over plays with a 54% success rate. When we look at top categories with more similar numbers of campaigns; rock and documentary, rock is still more successful than documentary with 57%

	subcategory	canceled	failed	live	successful	grand_total
0	plays	23	132	2	187	344
1	rock	6	30	0	49	85
2	documentary	4	21	1	34	60
3	web	2	12	1	36	51
4	food trucks	4	20	0	22	46

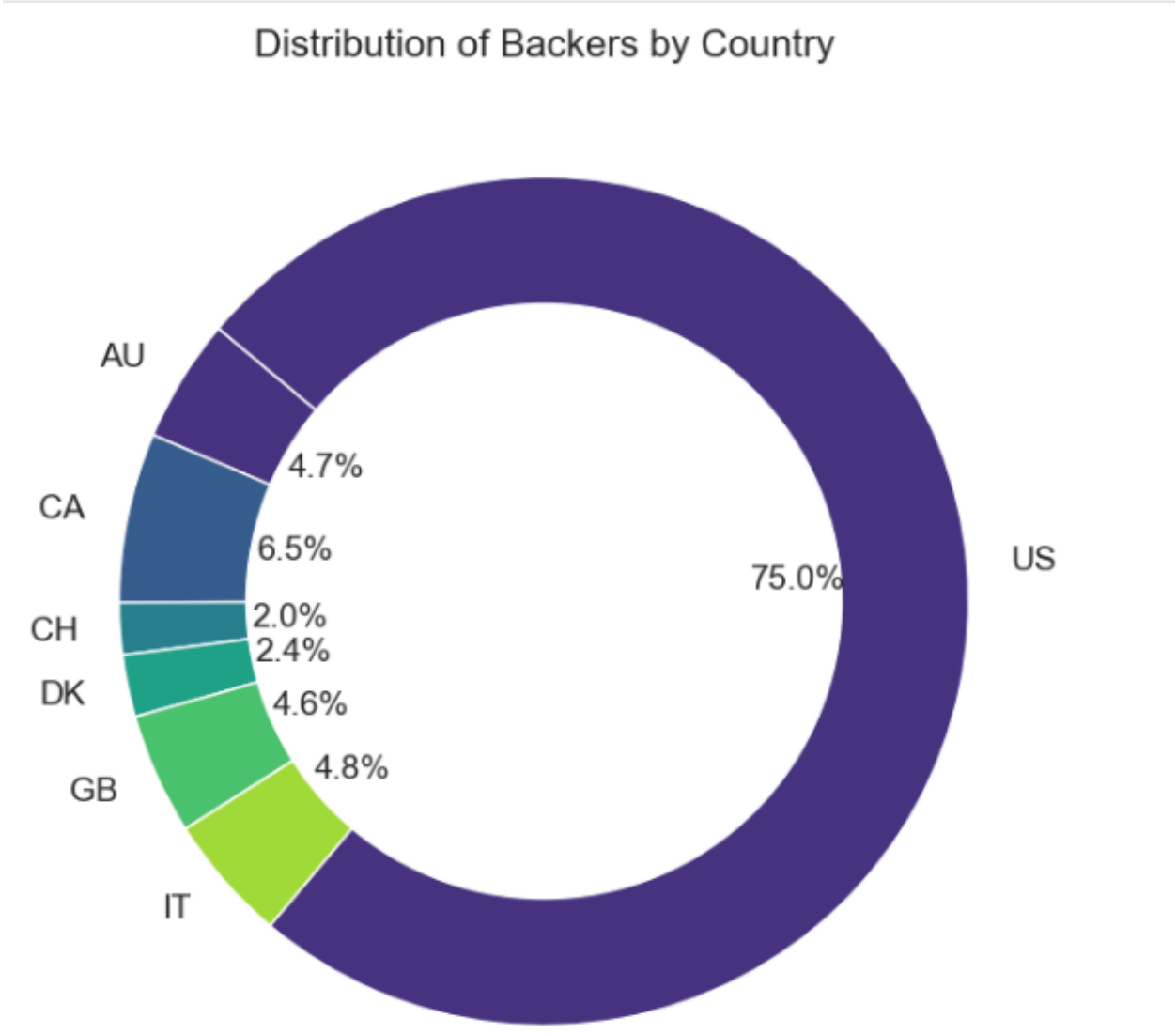
Reviewing percentage of goal by Country it's noted the US has the greatest number of Campaigns. Therefore, to avoid biases it's difficult to compare the data to other countries. Comparing Italy and Great Britain with similar total campaigns those who reached 50% of goal and above were as well similar with Great Britain slightly higher. However, Italy was higher with 70 and 80 percent of goal where Great Britain having a few in 50% of goal. Therefore, those two countries have very similar success of goal to pledge ratios.

	country	grand_total	goalreached	ninetyperc_goal	eightyperc_goal	seventyperc_goal	sixtyperc_goal	fiftyperc_goal	lessfiftyperc_goal
0	US	763	436	26	33	27	41	34	166
1	IT	48	26	2	4	2	2	0	12
2	GB	48	28	3	2	1	0	2	12
3	CA	44	22	3	2	2	3	3	9
4	AU	43	24	1	2	1	2	7	6

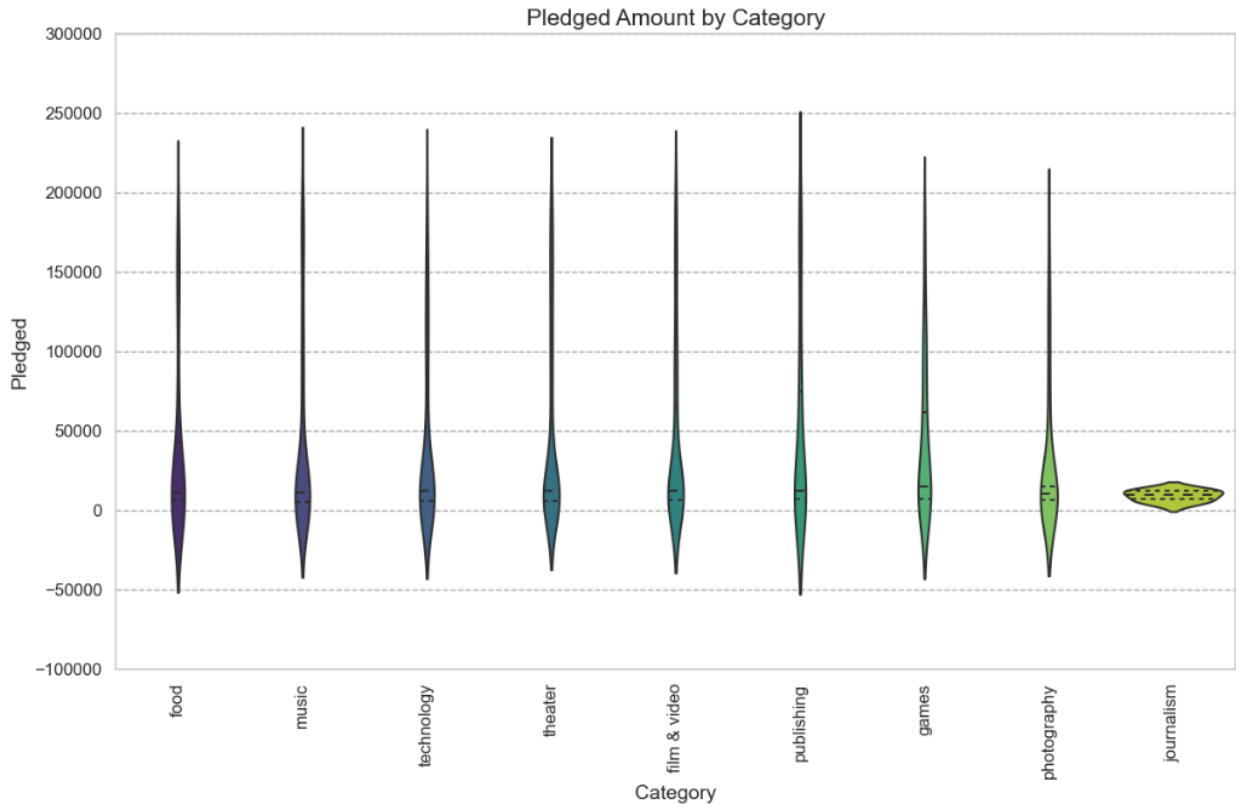
It appears pledge amounts are left skewed with the Median quite a bit lower than the Mean. We do see those outliers at the higher pledge amounts. With this data, leaving the outliers and showing the skew shows that bigger picture. The data does not appear to be bi-modal. We see that one peak on the graph.



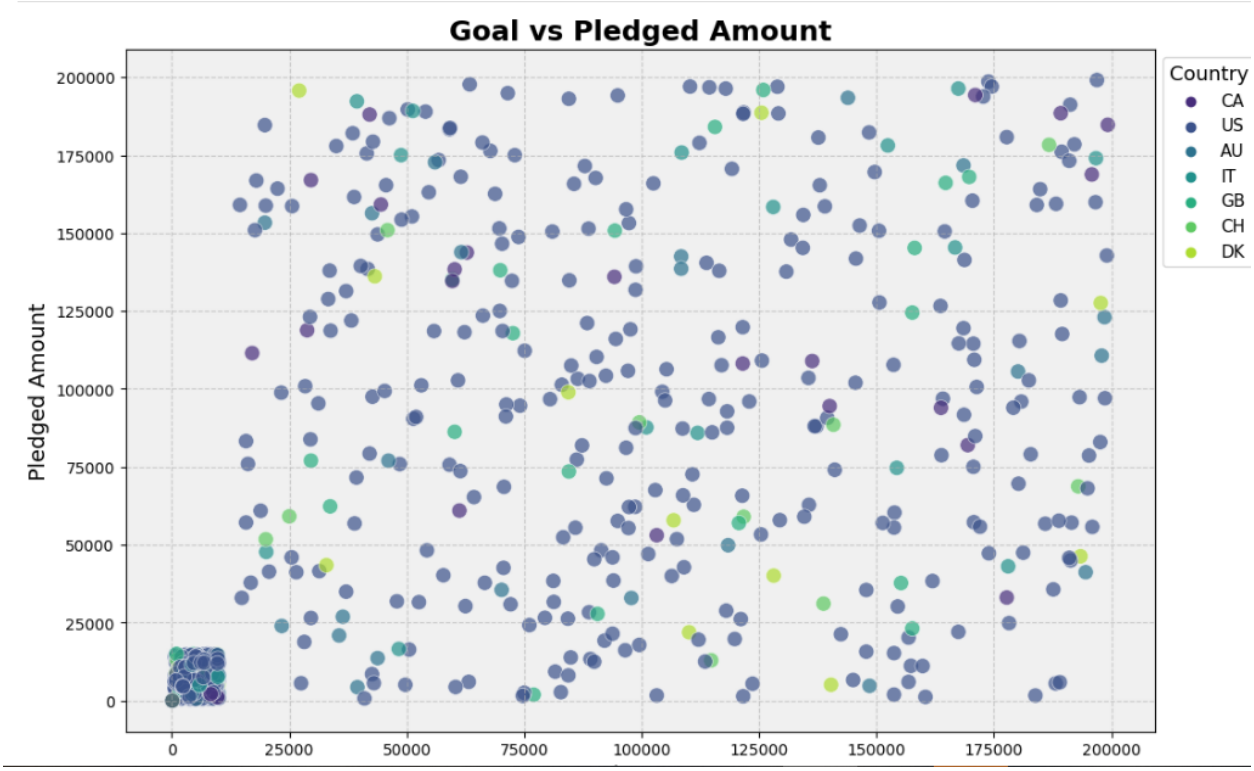
This analysis shows backers by Country. Of course, we see the US with the majority as the number of campaigns in the US was much higher than any other country. Italy and Great Britain had similar amount of Campaigns as we analyzed above. This chart does show Italy had slightly higher backing for their campaigns.
Canada is more interesting as they had more backers and less campaigns than Great Britain and Italy. More analysis would be needed to understand this percentage difference for Canada. Did Canada have bigger Goal amounts and therefore perhaps more backers? As this project doesn't include full data analysis, this could be future work.



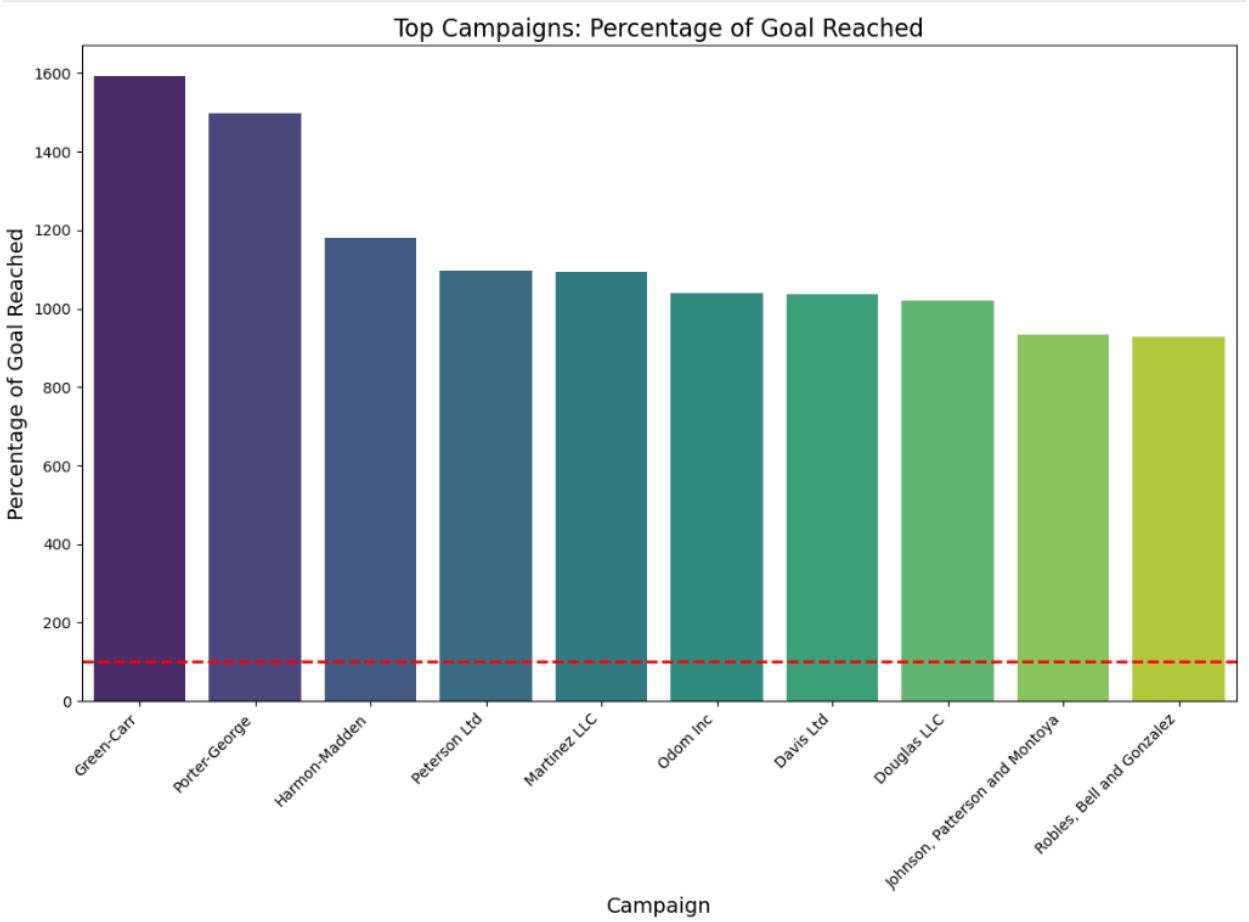
Below shows analysis of pledge amount by category. It's interesting to see journalism very flat. With further analysis we see few campaigns for journalism.



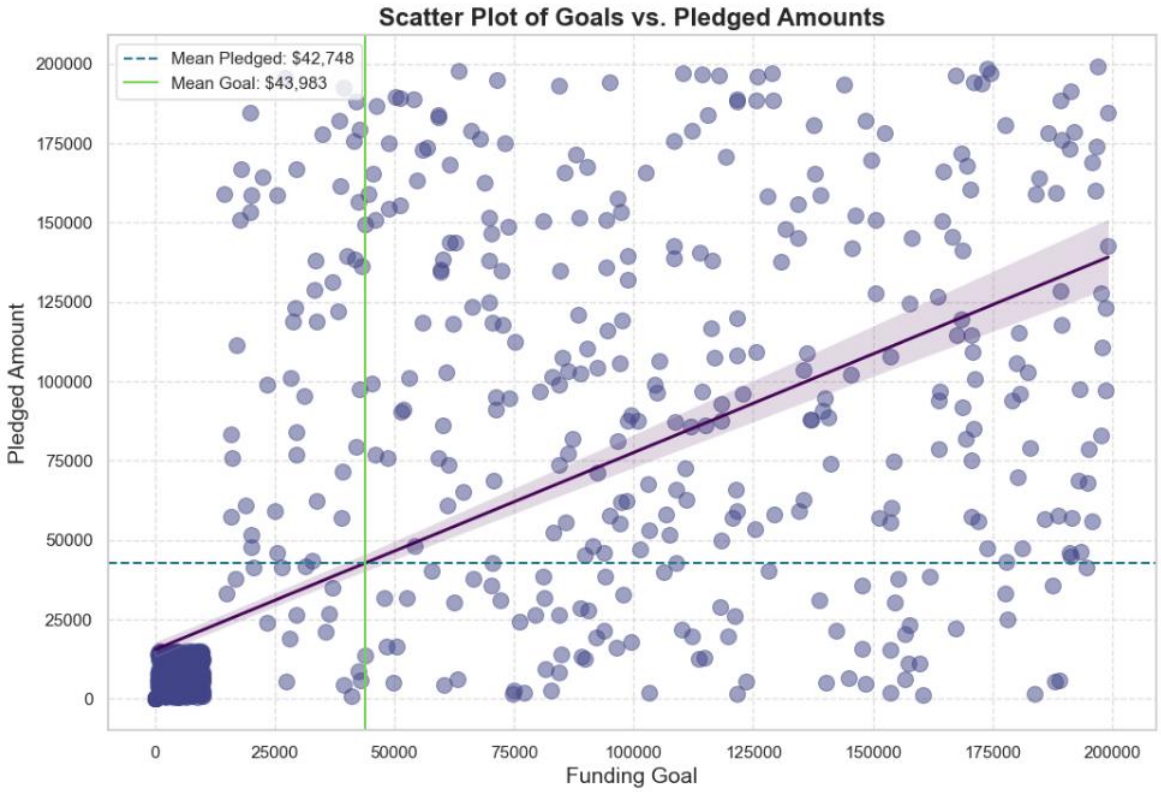
This Scatter plot shows Goal vs Pledge by Country. It's very noticeable there is a cluster of smaller pledges which is shown in the bottom left corner of the chart. Otherwise, as expected pledged amounts will vary based on Goal as well as the US as expected shows more pledges. It's also nice to see the higher Denmark pledges vs goal which could be analyzed further in the actual data.



This Bar plot shows Top Campaigns based on Percentage of Goal Reached. Green-Carr had the highest pledges to goal.



The linear relationship between the funding goal and the pledged amount are displayed below as well as the average pledged and average goal. As funding goal increases, pledge amount increases which is expected. It's a nice steady line which shows the relationship between funding goal and pledge amount appears to be there. However, it's not a real steep line showing it's not an extremely strong correlation. The relationship between funding goal and pledged amount does appear to be present.



Sources:

ChatGPT for some sorting of grand_totals fields in the ORM code
https://chatgpt.com/