Project 2 - Group 12
Josh Ehlke
Leonardo Rodrigues
Kerry Oostdyk

Introduction

We began our project with a large Excel file containing information about crowdfunding startups. The dataset included details on various companies' projects, their funding goals, and the start and end dates for each project. Additionally, it categorized projects and indicated whether they were fully funded. Another table provided contact information for each project. Our goal was to clean, simplify, and analyze this data using SQL.

Extract

First, we imported the data from the `crowdfunding.xlsx` file. Our objective was to extract the data and transform it into several simpler CSV files. We started by splitting the combined "category and subcategory" column into two separate columns, enabling us to create distinct category and subcategory tables.

This process allowed us to assign unique IDs to the categories and subcategories, reducing the amount of data stored. Once the category and subcategory dataframes were created, we exported them as individual CSV files. These dataframes are shown below.

Project 2 - Group 12
Josh Ehlke
Leonardo Rodrigues
Kerry Oostdyk

[11]:

	cat_ids	category
0	1cat	food
1	2cat	music
2	3cat	technology
3	4cat	theater
4	5cat	film & video

	sub_ids	subcategory
0	1subcat	food trucks
1	2subcat	rock
2	3subcat	web
3	4subcat	plays
4	5subcat	documentary

Transform

Next, we made a copy of the crowdfunding table and named it `campaign_df`. We examined the data types and renamed several columns for clarity as follows:

- `'blurb'` → `'description'`
- `'launched_at'` → `'launch_date'`
- `'deadline'` → `'end_date'`

We converted the "goal" and "pledged" columns into floats and transformed the "launch_date" and "end_date" columns into datetime objects.

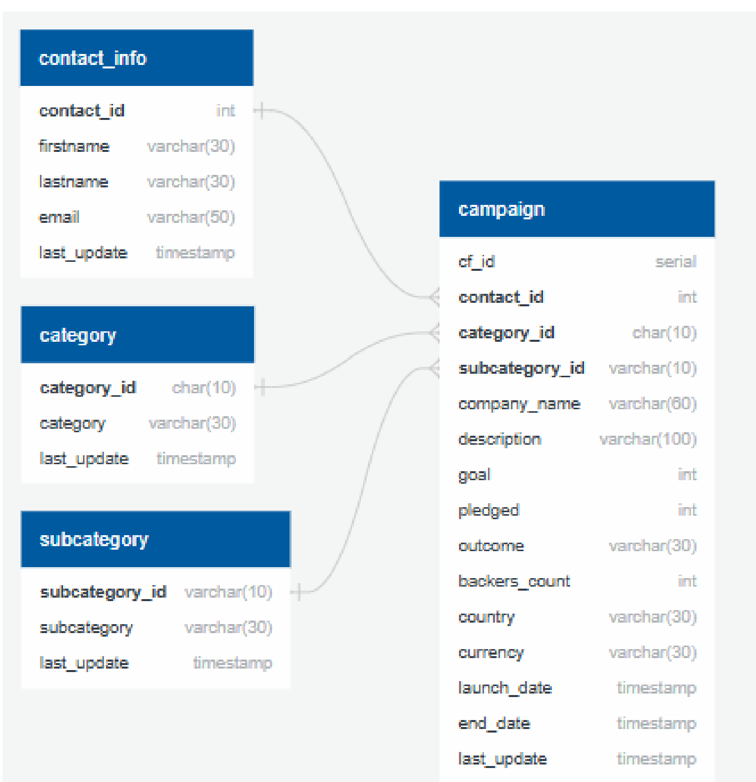
The next major step was merging our dataframes on the "category" and "subcategory" columns. After merging, we removed the original "category," "subcategory," and "category & subcategory" columns, as well as the "staff_pick" and "spotlight" columns, to streamline the dataset and reduce storage requirements. Finally, we exported the merged `campaign_df` as a CSV file.

Project 2 - Group 12
Josh Ehlke
Leonardo Rodrigues
Kerry Oostdyk

For the contact information, we created a separate contacts dataframe. To achieve this, we parsed the `contacts.csv` file into a dictionary using a for loop. This process allowed us to split rows into separate columns, as they were originally delimited by commas. We split the "first name" and "last name" columns at the comma and reordered the dataframe columns for better organization.

Load

To integrate the data, we designed an Entity-Relationship Diagram (ERD) connecting the four CSV files created in the earlier steps. The ERD diagram is shown below. Primary keys such as `contact_id`, `category_id`, and `subcategory_id` were established, which served as foreign keys in the campaign table.



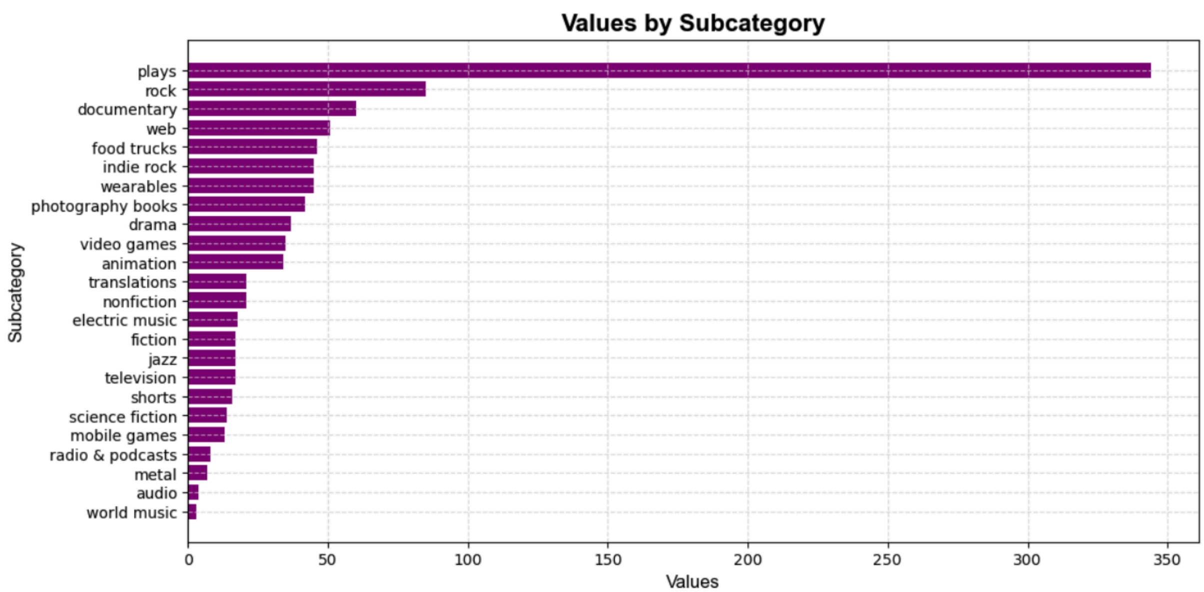
Project 2 - Group 12
Josh Ehlke
Leonardo Rodrigues
Kerry Oostdyk

We then loaded the schema into pgAdmin and added the four tables to our database.

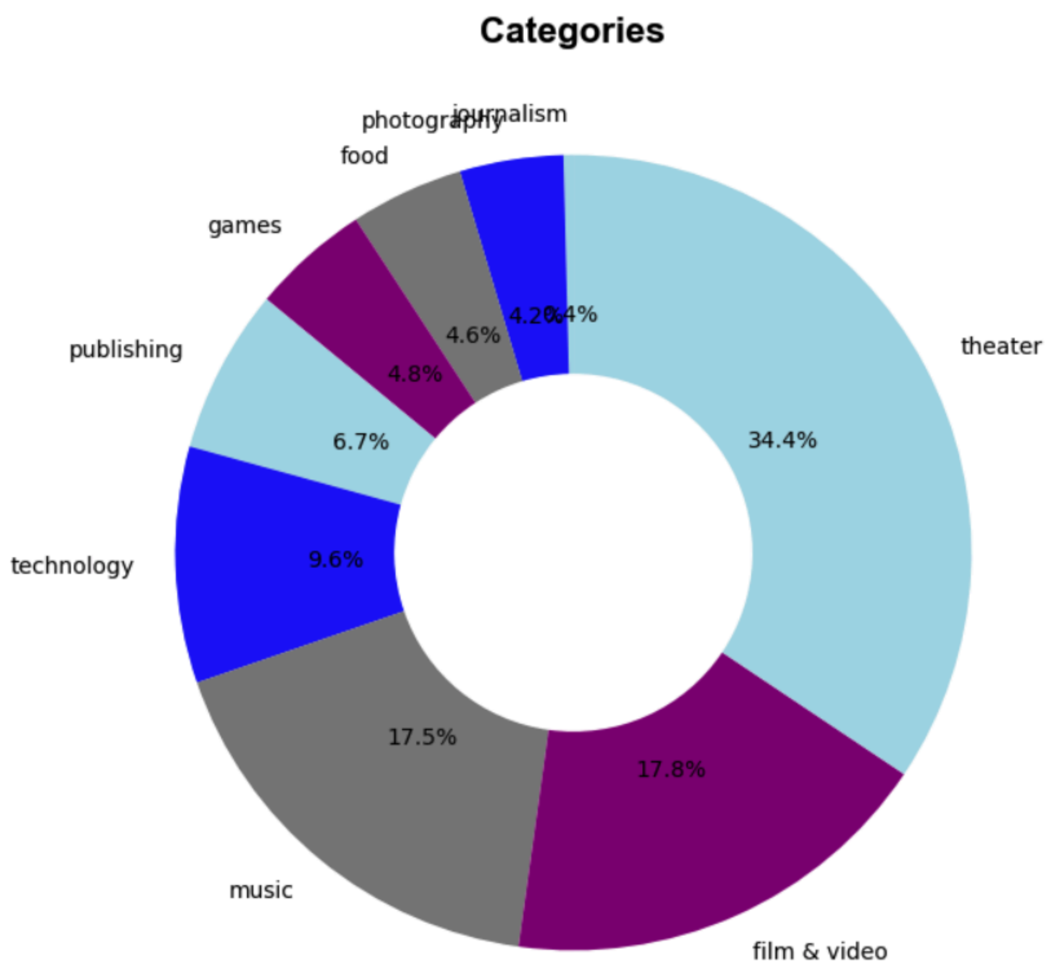
After that, we loaded the 4 CSVs into pgAdmin and ran some exploratory queries on them.

Analysis

Our first investigation led us to look at the categories and subcategories of the campaigns. The most common subcategory was “Plays” with 344 total campaigns. This makes sense when compared to the categories number 1 pick, theater, holding majority with 34.4% of total campaigns. This is shown in the figures below.



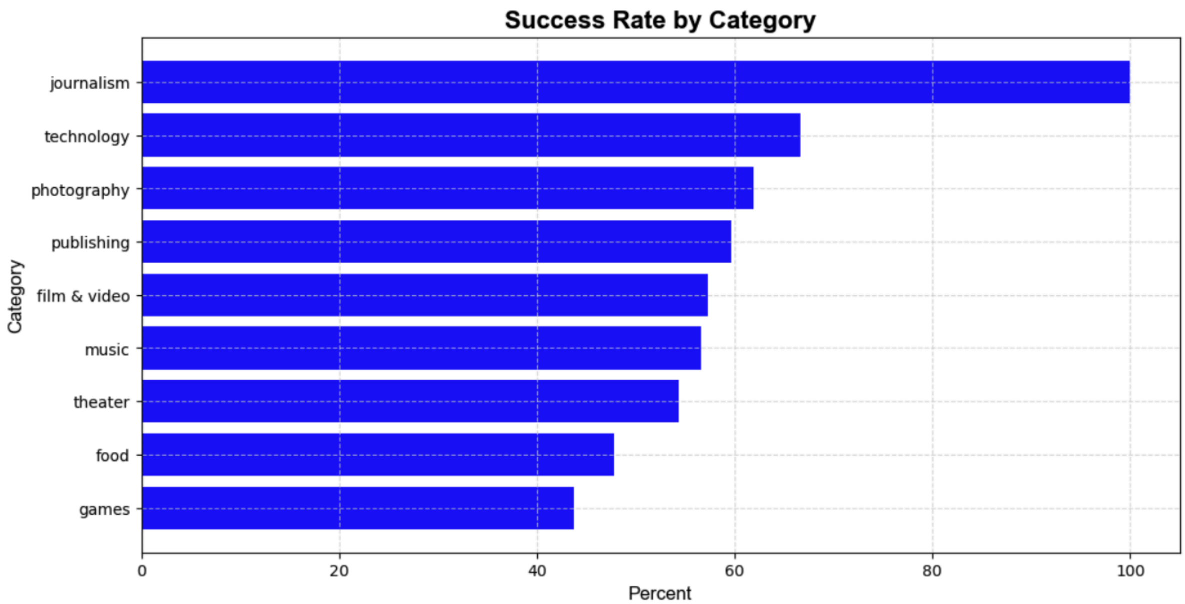
Project 2 - Group 12
Josh Ehlke
Leonardo Rodrigues
Kerry Oostdyk



After querying the total number of campaigns by category, we wanted to evaluate the success rate for each category. The most successful was journalism, but when looking closer we noted there were only 4 journalism related campaigns. This can be seen in our donut chart as well. Interestingly, the second highest was technology at a success rate of

Project 2 - Group 12
Josh Ehlke
Leonardo Rodrigues
Kerry Oostdyk

67%.



Conclusion/Future Work

In future work, we would look at the success rate with subcategories as we did with teh categories. We would also like to look at the success rate depending on the length of the campaign. Although we discovered a lot about the categories of these campaigns there is still further work to do.