Purpose

This project focuses on our ability to extract, transform and load a given crowdfunding dataset for analysis of trends in an end-user PostgreSQL system. We utilized python code in Jupyter notebook to extract and transform the data into a relational database so queries could be created and run to better understand the factors that contributed to successful campaigns.

Extract

Excel formatted files were supplied to us and extracted into a pandas dateframe using python script in Jupyter notebook. This allowed us to view a brief summary of the imported data and create transformation goals.

Transformation

Within our Jupyter notebook we cleaned, formatted and processed the data for compatibility to load into a postgreSQL end-user system. As a first step we converted data types for “launched_at”, “deadline”, “goal” and “pledged” columns. We then separated “category” and “sub_category” by the “/” split, placed data into new individual columns and deleted the original combined column. Arrays were created for each new individual category and sub-category to ensure all unique categories were retained. A for loop was created to assign categories and sub-categories to their respective ID options contained in the array and a .csv file for category and sub_caterogy was created, category_df.csv and subcategory_df.csv, respectively. Our transformed crowdfunding.csv dataframe was exported as campaign.csv.

A second dataframe was created from a list of dictionaries containing contact information, “contact_id”, “first_name”, “last_name” and “email” data was extracted and respective columns were created. The dataframe column names were re-ordered and the data types were confirmed before being exported to a contacts.csv file.

We used QuickDBD to design our ERD and show the null, data types and primary keys and foreign key relationships (Figure 1). This allowed for a better visualization of how the data is organized and interacts.

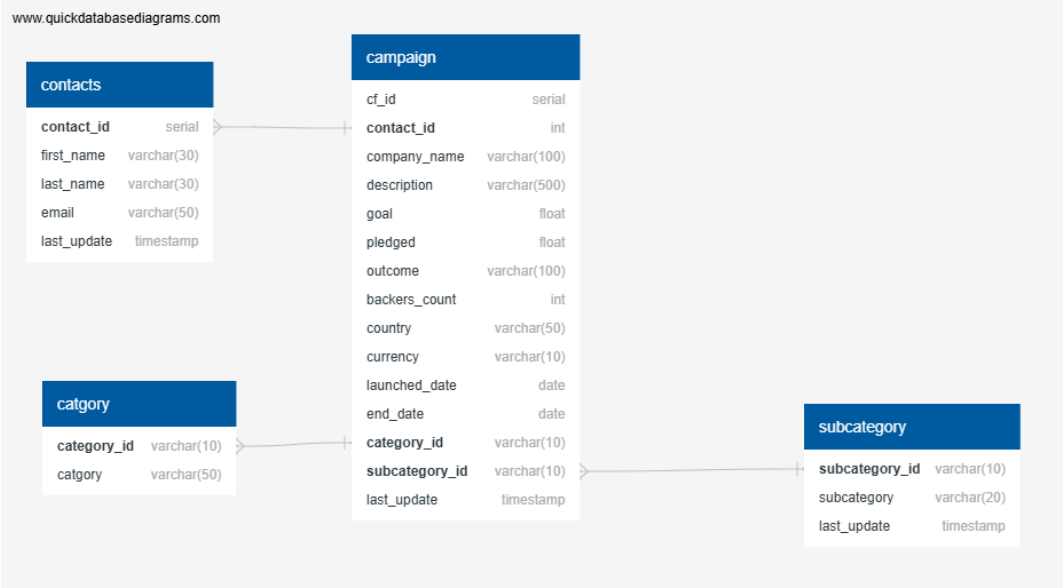

Figure 1. QuickDBD visualization of database relations.

Load

We loaded our transformed data into a new database in pgAdmin, ran a new query containing our schema and manually loaded our transformed .csv files into pgAdmin tables (contacts, campaign, category, and subcategory). We were then able to run queries and visualizations on the data to help us better understand trends for successful campaigns.

Of the nine campaign categories theater, film & video and music were the clear top three categories with the highest number of campaigns. Theater skewed the high end number of campaigns by having almost 40% more campaigns than music and film & video which held similar campaign numbers at just under 250 (Figure 1).

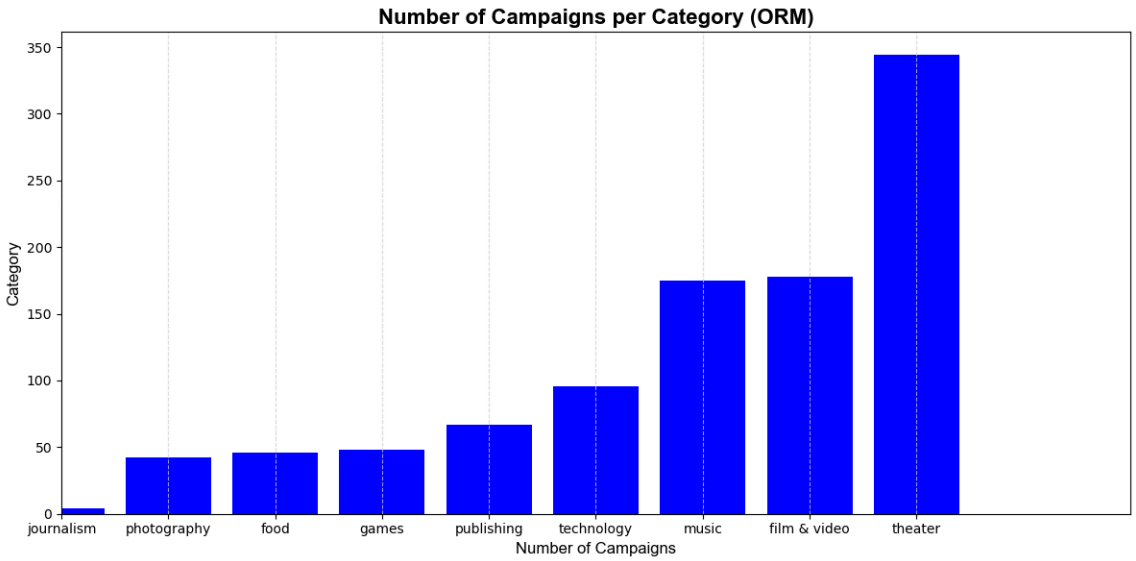

Figure 1: Total number of campaigns for all categories.

We utilized a heatmap visualization to better tell the story of success formusic campaigns. The below heat map shows that there is a strong relationship between the pledged and goal amounts for a campaigns success, those campaigns who met their pledged goals were more likely to be successful.

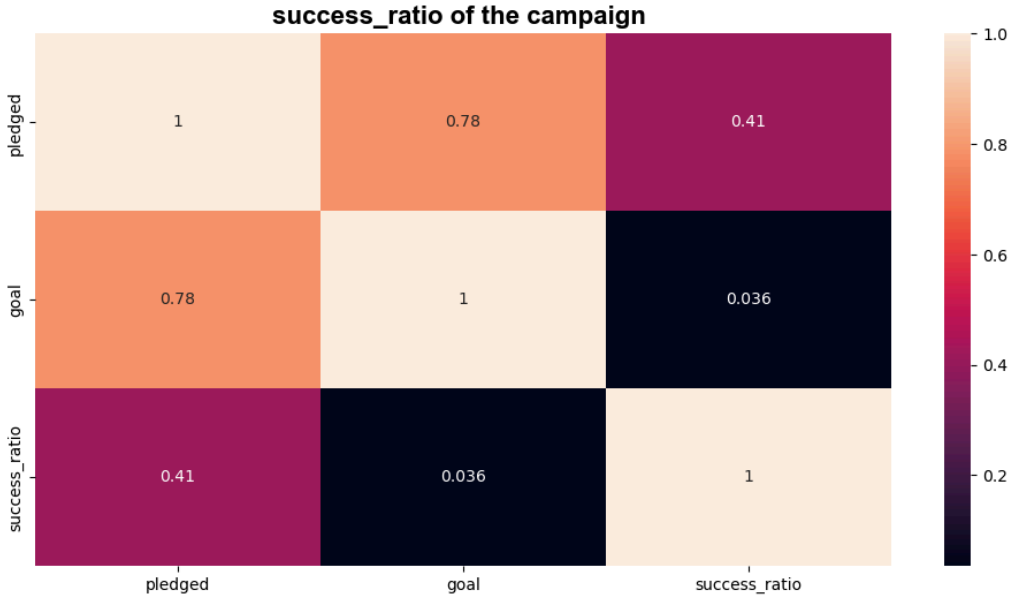

Figure 2: Heatmap visualization of amount pledged and goal metrics for successful music campaigns.

Our final visualization was a linear regression showing the relationship between success ratio of a campaign and the desired goal amount. There was no correlation (y=0.0x + 0.62) between a set goal amount and the successful outcome of a campaign. A campaign's success was marked more by their ability to meet their funding goal, regardless of the goal being a set number.