ETL Mini Project

Sierra Sarkis, Kelsey Mersinas and Judith Landa

Introduction

This project presented two Excel files containing a large amount of raw data on crowdfunding campaigns between them, with one containing significantly more data than the other. Given the challenge of reading and understanding differently formatted data and transforming them into usable tables to be read into our database in pgAdmin, we set off to extract, transform, and load our data into a usable format appropriate for visualizations and analysis using Pandas and Postgres.

ETL Code

Beginning with our extracted data, we went to transform the files from two excel files to four csv files that could be used to create our database. This required both cleaning and structuring the data into the correct data types. Initially, we started with mostly strings and integers but inspected this to make sure we understood what the data type was for each column.

Our crowdfunding dataframe was in the first normal form, with each row filled with the values, most of this data was specific to each campaign and not generalizable in any way. However, with our category and subcategory columns, we had the opportunity to remove redundancies and create a table of categories and subcategories that our crowdfunding table could reference. We assigned an ID we could use as the primary keys for both our category and subcategory tables and saved these to new csv files for use in our upcoming database. Using these new tables we could merge them with our crowdfunding data frame and remove the category values themselves, leaving the IDs that we would use as the foreign keys once in pgAdmin.

The remainder of the campaign database would need adjustments to some data types and a rearrangement of the columns in a way that prioritized the data that provided qualitative data for us to analyze. Saving this off as a third csv file we had our crowdfunding data transformed to the desired state.

Our contacts were given to us in a series of dictionaries that we needed to format into a dataframe. This could be done with a for loop to iterate through each row and return JSON data, which could be separated and sorted into the desired dataframe. The contact id column would be the only values shared by our campaign data. Meaning that other than checking and converting our data types we were ready to save our desired format to a csv file and move to our database.

Database Design

With our four csv files we had to define and understand the relationships between our tables to make an effective schema. Our contact, category and subcategory tables were relatively small and had a district primary key that we knew would always be unique to the table and would need to be included. Moving to our larger campaign table, we identified our foreign keys that would be referring to our other three tables and set our primary key. This organized our csv files in a simplistic and effective format.

Creating a comprehensive schema includes assigning the correct data types to our values, which were identified in our ETL notebook when we reviewed our values and dataframes. Once we had our schema we could read our csv files in with Pandas to Postgres. Our tables were now loaded and we could select \* to view the data within.

Analysis

This data provided a wide range of values but the outcome column was a clear descriptor of success or failure. We asked how we could compare the successful campaigns to the failed ones and ran a query to return the amount of crowdfunding campaigns per each outcome listed. From this, we could make a donut chart that displayed a clear picture of the overall data results and were able to show the breakdown of the final outcomes in these crowdfunding campaigns.

Delving further into our successful campaigns, we looked to breakdown the amount of succeeded campaigns per category to uncover any trends. To do this we needed to run a query for both our campaign table for a dataframe of the successful campaigns and their category\_id, and additionally, we needed a query for the category names themselves, which are now in the category table. These results could be turned into a dataframe and merged to deliver the data we needed for our plot. We now had two visualizations on our successful campaigns that we could use for analysis.

This data shows that the majority of these campaigns concluded successfully. Of the successful campaigns our top three categories making up our successful campaigns were theater, film & video and music. This points to shows and entertainment as being good predictors of success in a crowdfunding campaign.

Limitations

When looking at the data being evaluated we can see that there are a couple primary ways in which the data is limited. First, the dataset is only capturing data from 2020 - 2023. To gain a more solid understanding of trends over time it would be useful to have data spanning over several years. While 4 years seems like a long period of time, there are many economic factors that can change spending trends. For example, natural disasters, elections and pandemics could all have large impacts on the shape of these trends. It would be interesting to know how the goals and pledges may have varied without the economic impact of Covid.

Second, there are only 7 countries included in the dataset, expanding to include additional countries could provide a better understanding of the global picture when it comes to these campaigns. Overall the dataset appears to be small, but if data collection continues we would be able to make a more sound analysis with each passing year.

Future Work

There are many ways in which this work could be improved upon, including technical adjustments, data collection, and data expansion. Technical improvements could look like assessing options to improve the design and efficiency of our pipeline, simplifying code to be more reader friendly, or even devising different ways of capturing the data instead of reading it from an excel file. Data collection as referenced above would be simply continuing to collect this data over time to ensure a robust dataset. This would help us to understand how outcomes change year over year, and indicate optimal timeframes for each outcome to be implemented. Finally, data expansion could include new companies, countries, or categories to be captured in the dataset.

Conclusions

Through this project, we successfully transformed raw crowdfunding data from Excel files into structured CSV files for database integration and analysis in pgAdmin.Our process involved thorough data cleaning, transformation, and organization to make sure the data was user-friendly. We set up a clear database schema and established relationships between tables, which allowed us to load and search the data effectively, resulting in insightful visualizations that highlighted crowdfunding campaign outcomes and trends. However, the dataset had some limitations due to its short timeframe and narrow geographic focus. Looking ahead, we plan to enhance the data pipeline, keep collecting data for more in-depth analysis, and expand the dataset to include more countries and categories. These improvements would provide a more comprehensive understanding of crowdfunding trends and outcomes, leading to more accurate and impactful insights.