Project 2 Group 5
Bootcamp: DATA-PT-EAST-APRIL-041524-MTTH
July 17, 2024

Presented By:
Alexandra Reid
Jorge Arroyo
Lina Blanco Restrepo

Introduction

In this project we analyzed crowdfunding campaigns data. The dataset contains various attributes related to each campaign, such as the campaign's outcome, launch date, end date, funding goal, amount pledged, number of backers, and other relevant details. The purpose of this project was to use a more practical and hands on approach in the creation of an ETL pipeline. Close to what we would face as a data analyst in our work environment.

Database Design Considerations

In this project, various Python libraries and functions were used to extract, transform and load data from the crowdfunding.xlsx and contacts.xlsx files. Here are some of the key functions and techniques we used:

- Pandas library for data manipulation, analysis, and exporting data to CSV files using.
- Creating DataFrames and manipulating data using Python dictionary methods by converting rows to dictionaries by using the function to_dict().
- Iterating through DataFrame rows using the iterrows() function.
- Data cleaning techniques such as converting data types, renaming columns, handling missing values, and dropping columns.
- Regular Expressions (Regex) for extracting specific patterns, like emails, from text data.
- Split columns using the .str and .split('/', expand=True) function.
- This data manipulation py library, allows to split the values of a column into multiple columns based on a specified delimiter. The .str accessor is used to access string methods for pandas Series (columns), and the .split method is used to split the string by a specified delimiter. The expand=True argument is used to return the split strings as a DataFrame, which allows each split part to be placed in a separate column.
- Creating relational database tables using SQL and defining primary keys, foreign keys, and constraints.
- Using QuickDBD to sketch an Entity-Relationship Diagram (ERD) for the database schema.
- Creating a Postgres database and tables based on the defined schema.

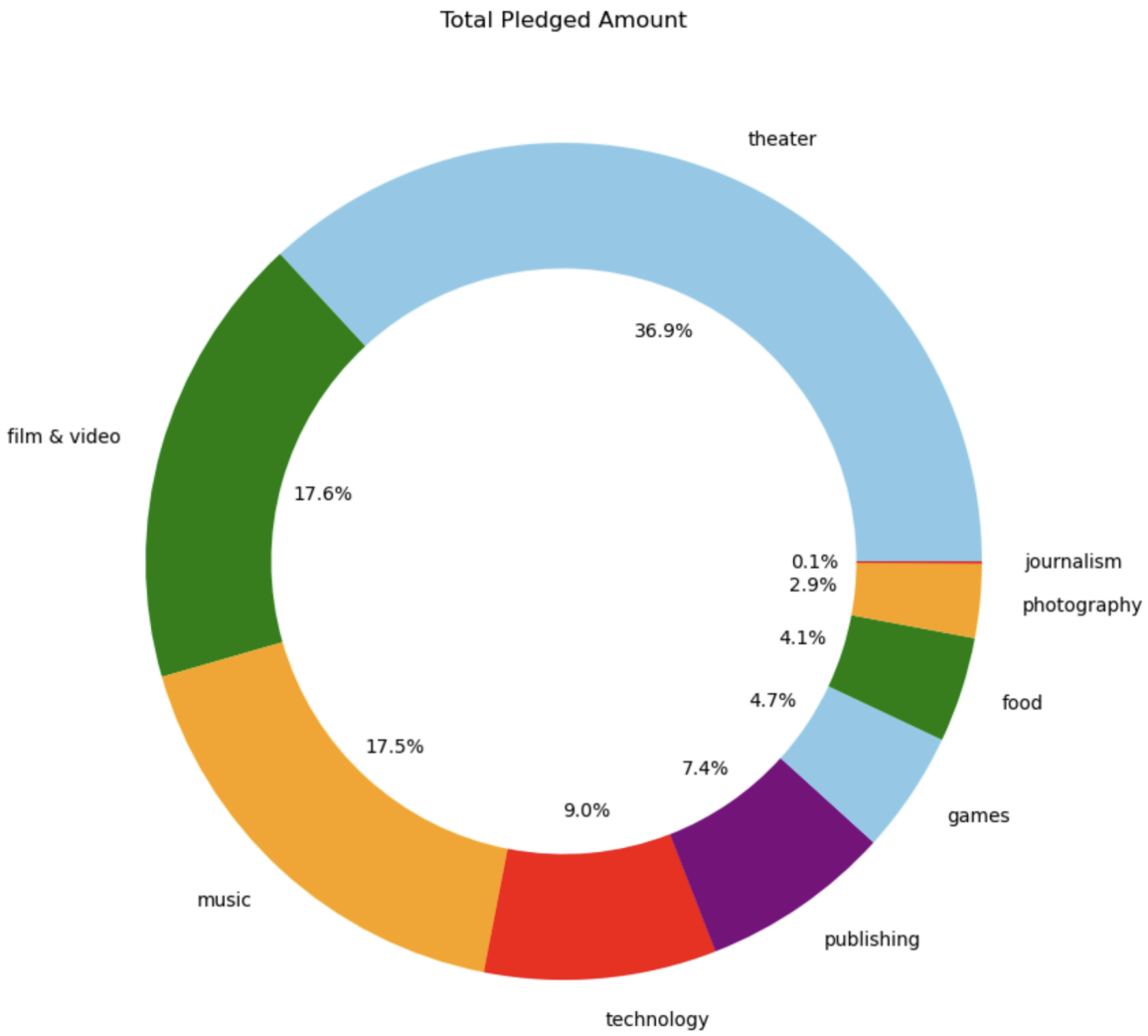
Analysis

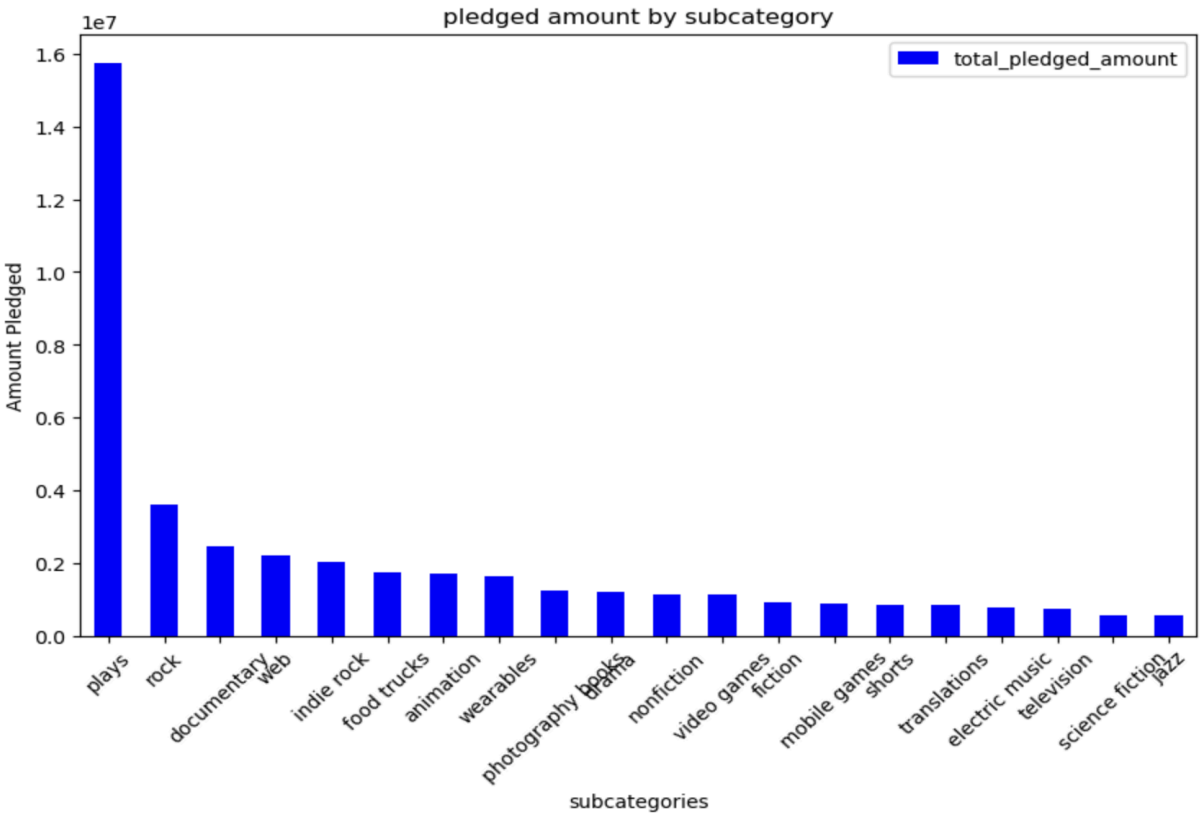
Questions that were asked after loading the data and querying from postgres were as follows:

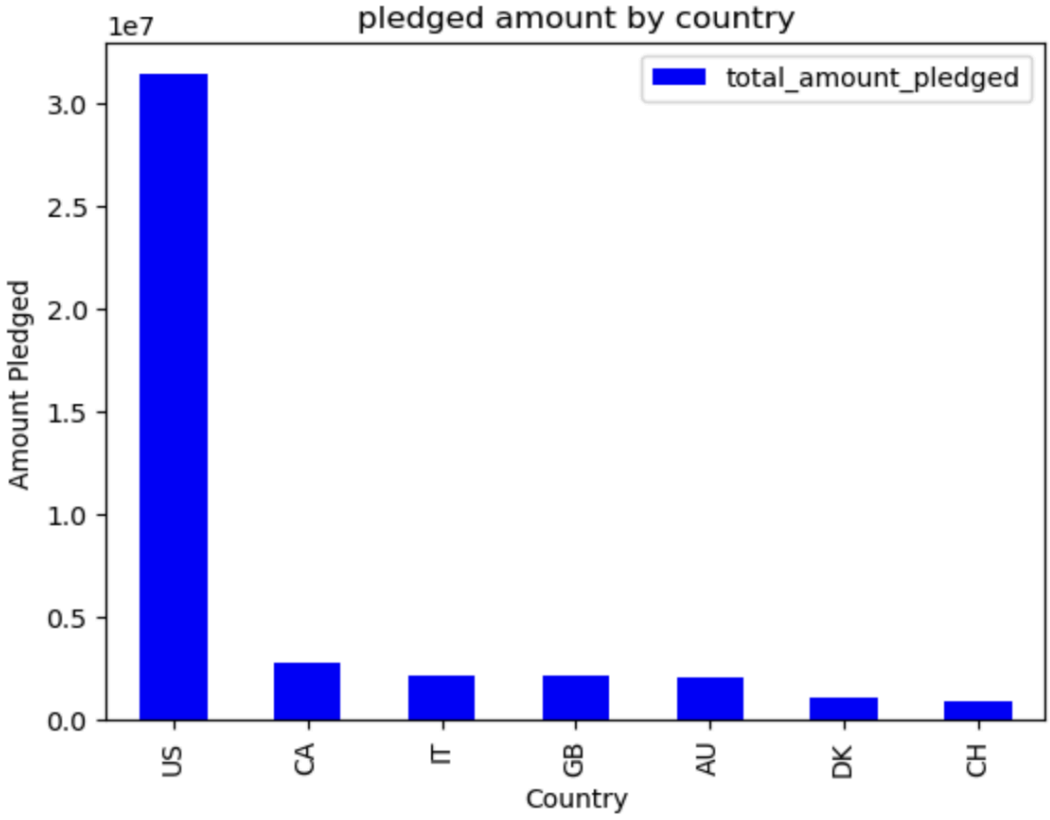
- What was the amount of money pledged by category?
- What was the amount of money pledged by subcategory?
- What was the amount of money pledged by country?
- What was the relationship between outcome and pledged money?

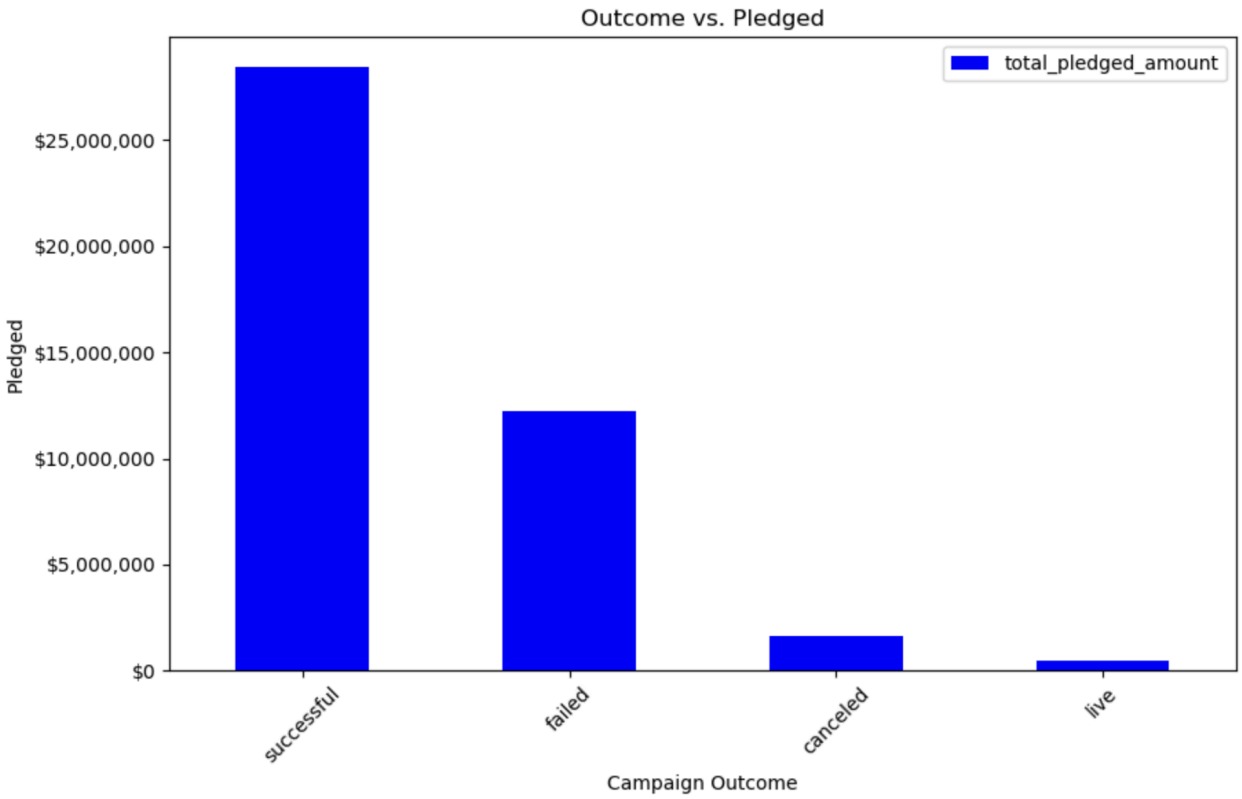
The subcategory bar chart and category pie chart had strong correlations with one another since they both were borrowed from the money pledged columns and the subcategory column derived from the category column. Theater had the most money pledged behind it and when we looked into the subcategory dataframe, we could see what specific activities played into theater

and music being the top pledged sections in their respective columns. Next we looked at how much money was pledged based on the country. The US had the most money pledged among all the countries by a significant margin. This could have been due to the US having a higher participation rate within the data compared to other countries. Finally, we looked at the relationship between outcome and pledged money. Successful campaigns had close to double the amount of pledged money compared to the failed campaigns.




pledged amount by subcategory
1e7
total_pledged_amount
Amount Pledged
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
plays
rock
documentary
web
indie rock
food trucks
animation
wearables
photography
drama
nonfiction
video games
fiction
mobile games
shorts
translations
electric music
television
science fiction
jazz
subcategories


pledged amount by country
1e7
total_amount_pledged
Amount Pledged
3.0
2.5
2.0
1.5
1.0
0.5
0.0
US
CA
IT
GB
AU
DK
CH
Country



Bias/Limitations

While we didn't have many issues with the data there are some common limitations that can occur when doing an ETL pipeline. Some of those limitations include data quality issues, data loss, and processing time. When creating the pipeline it's important to make sure that you are using quality data. Data that is inaccurate, incomplete, or inconsistent can lead to errors in the transform stage and can cause issues with the analysis and results. Data loss is another very important limitation. There is risk of corruption or losing data at any point in the ETL process. To help mitigate data loss issues in our project we made a few copies of dataframes to ensure that we had back ups just in case something was deleted or changed. Processing time is another limitation that could cause issues when creating an ETL pipeline. If you are trying to process large amounts of data it could slow down processing times and can impact the efficiency of the project. This could then impact how long it takes to make data-driven decisions with the pipeline.

Conclusions/Reflection

It’s a wrap! As a data analyst, transforming and manipulating data can provide a better understanding of the factors that contribute to the success or failure of a project in this case the crowdfunding data provided just that. The opportunity to derive insights that can help future campaigners optimize their chances of success. This project was a great test of our skills in creating an ETL pipeline and how we may be able to use it in the future.

References

Xpert by edX Boot Camps LLC and is intended for educational purposes only.

Panda web inquiry: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html

Classroom exercises from Boot Camp Consortium East Coast DATA-PT-EAST-APRIL-041524

Postgresql Inquiry:
https://www.postgresql.org/docs/

Regex Inquiry
https://www.rexegg.com/regex-quickstart.php