Project 2 Group 8 Write Up

By: Gavin Bozan, Misha Mambully Muralidharan, Arya Maredia, Brian Marowsky

Introduction

The project aims at building an ETL (Extract,Transform,Load) pipeline to process and then analyse the crowdfunding campaign data. We have used Pandas and postgresSQL.

Extract Data and Transform

In the first part we extracted data from the excel files and then transformed the data with pandas. We created four DataFrames, category, subcategory, campaign and contacts from the crowdfunding and contacts excel sheets.

- For category dataframe we created a category_id column with uniques ids and the category column contains the names of these values

- For the subcategory dataframe, a subcategory_id column was created with unique ids and the subcategory column containing the values.

- The campaign dataframe is a subset of the crowdfunding dataframe and merged with category and subcategory dataframe.The datatype of pledged and goal column were changed to float and the launched_at and deadline were changed to datetime.the launched_at and deadline column names were changed to launch_date and end_date respectively

- The contacts dataframe was extracted from the contacts dataframe.The contacts dataframe had contacts_info

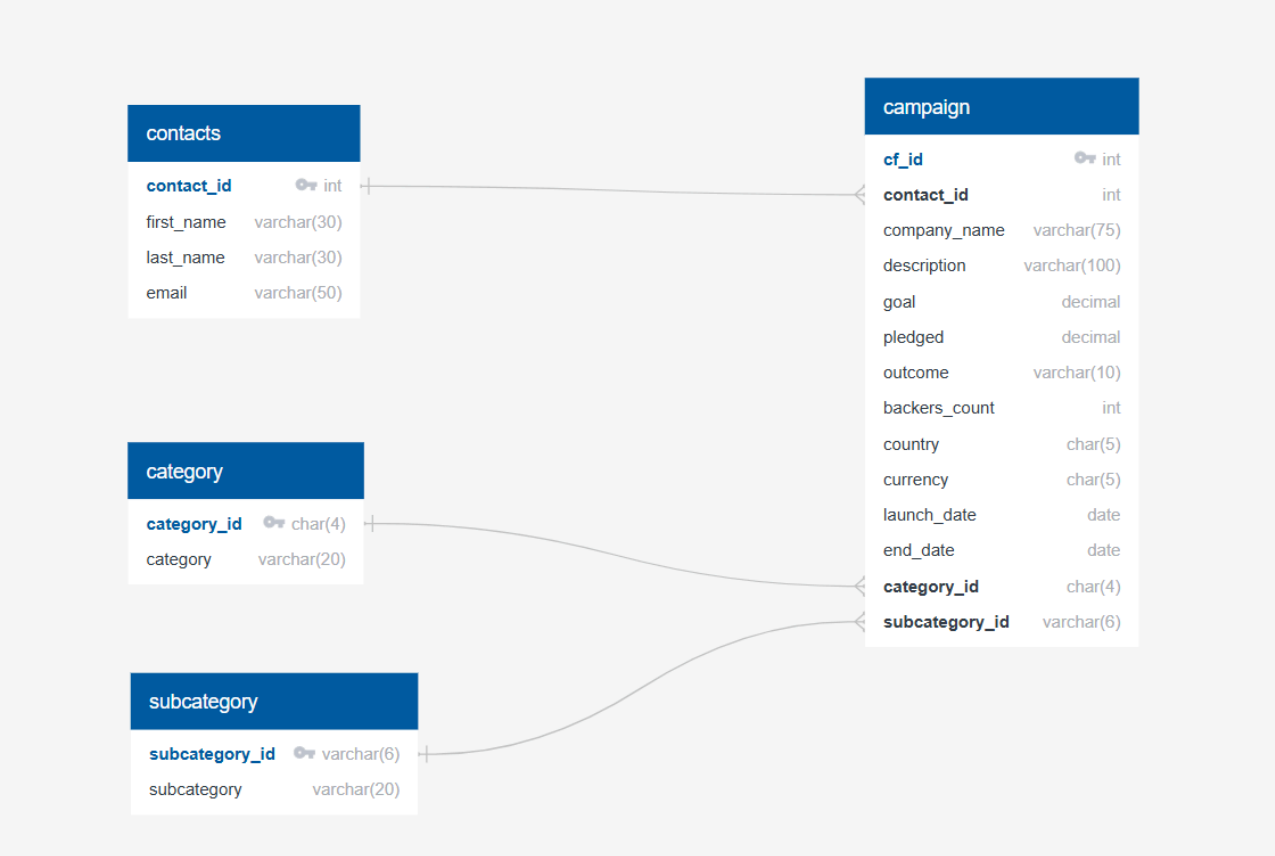
(contact_id,name,email) in a string format.We had to transform it into a python dictionary using the json.load() method in python.

All the four dataframes were exported as a csv file.

Database design and loading data

We used QuickDBD to design our ERD diagram.We used category, subcategory, contacts and campaign tables. The columns were allocated the correct datatypes.The Primary Key and the Foreign Key were properly implemented. We generated the schema and then in our database crowdfundingdata created the four tables. Using SQLAlchemy and Pandas the data was loaded successfully into these four tables.

ERD Diagram



Data Analysis and Visualization

We queried the database with the following queries.

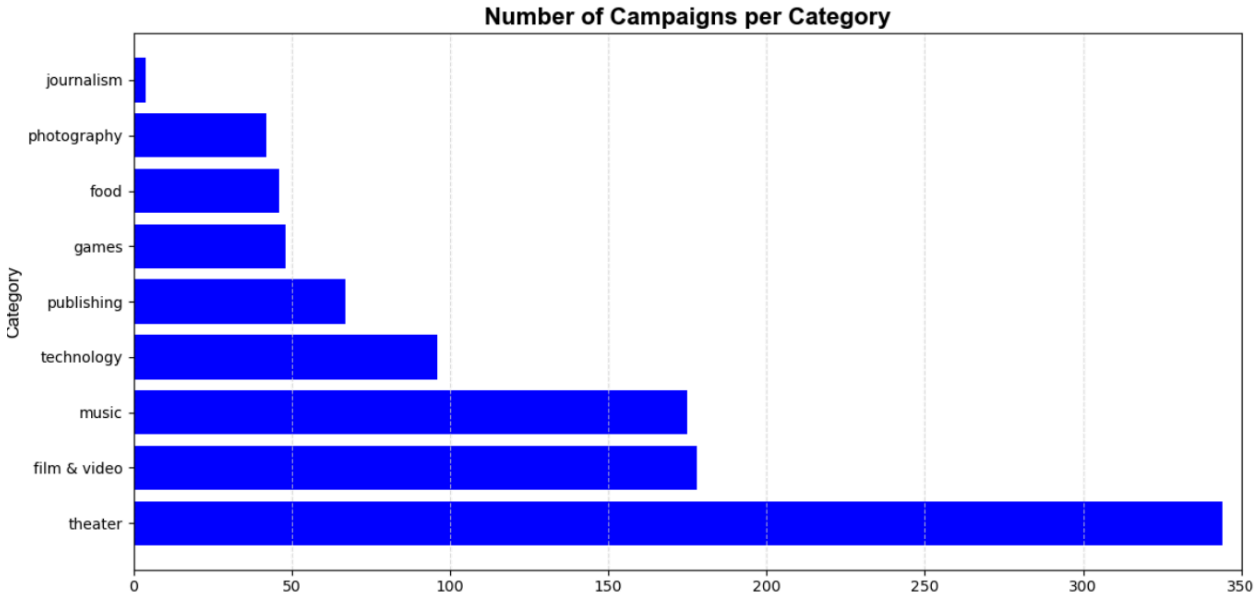
- First we analyzed the total number of campaigns by grouping it by category. The SQL query counted the number of campaigns for each category and then sorted the values in descending order.From this we were able to find out the category with the highest number of campaigns.
- The query we used was:

  select c.category, count(cam.cf_id) as number_of_campaigns

  from category c join campaign cam on c.category_id = cam.category_idgroup by c.category

  order by number_of_campaigns desc

The resulting data was then visualized using a bar graph. The bar graph provided a quick overview of which categories dominated the crowdfunding area.



- Then we analyzed the distribution of campaign outcomes.we queried the outcome column in the campaign table by grouping them by outcomes and counted how many campaigns had each outcome.
- The query used was :

  select cam.outcome, count(cam.outcome) as outcome_count

  from campaign cam

  group by cam.outcome;

A pie chart was created to represent this distribution. The pie chart helped to visually represent the percentage of successful, failed, live and cancelled campaigns.

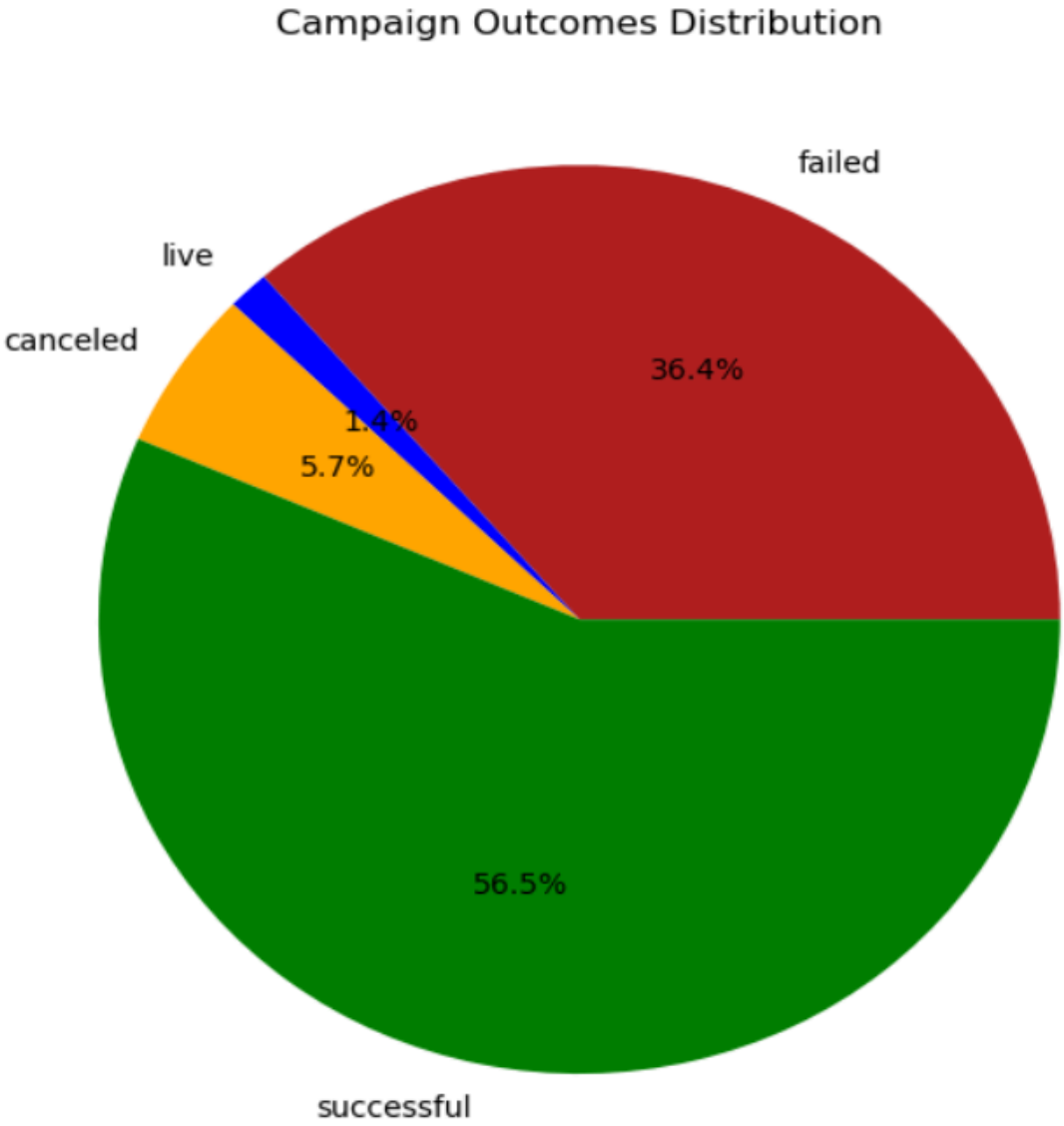
- We wanted to explore the distribution of campaigns across different countries. The SQL query grouped campaigns by country and counted the number of campaigns for each.
- The query used was:

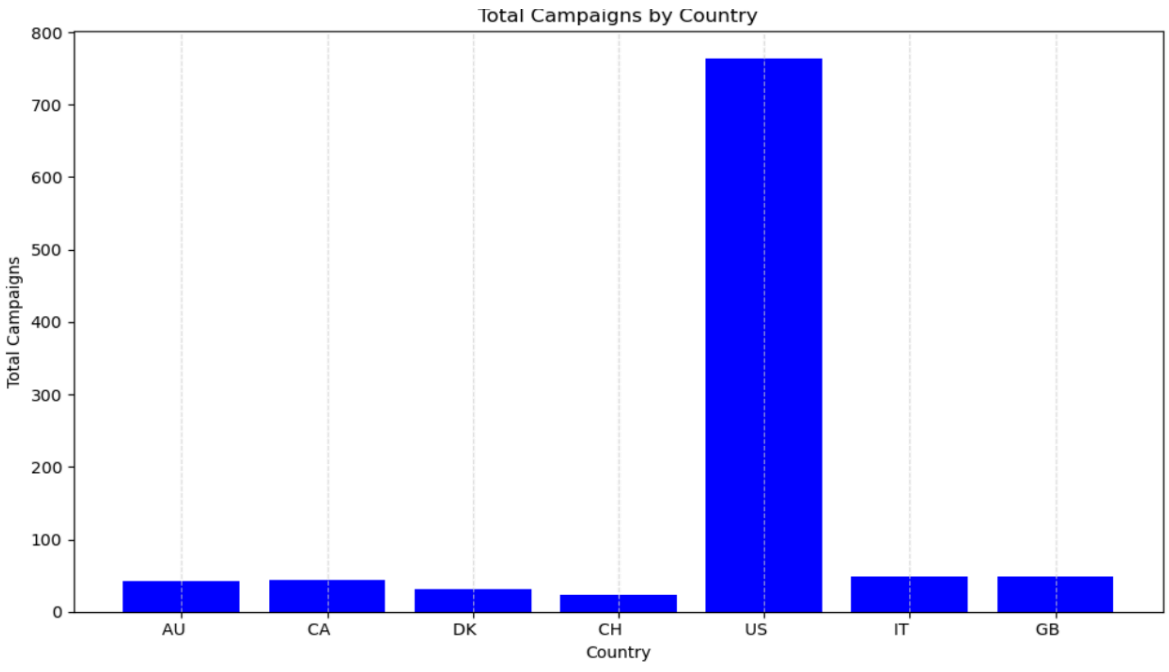
  select cam.country, count(cam.cf_id) as total_campaigns

  from campaign cam

  group by cam.country;

A bar graph was used to visualize the total number of campaigns by country. This helped us to identify the regions where crowdfunding efforts were most concentrated.


Campaign Outcomes Distribution
failed
live
canceled
36.4%
1.4%
5.7%
56.5%
successful

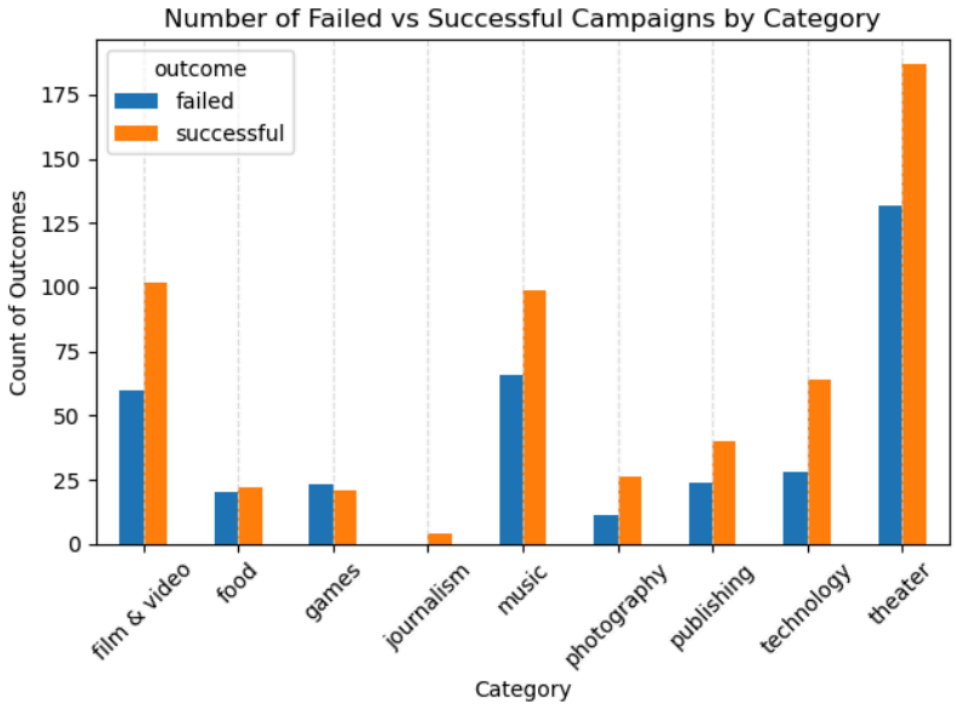

Total Campaigns by Country
Total Campaigns
800
700
600
500
400
300
200
100
0
AU
CA
DK
CH
US
IT
GB
Country

- We then wrote a query to find the total number of failed and successful campaigns and grouped them according to the category.

  The query used was:

- **select c.category,cam.outcome, count(cam.outcome) as outcome_count from category c join campaign cam on c.category_id = cam.category_id where upper(cam.outcome) in ('FAILED','SUCCESSFUL') group by cam.outcome,c.category order by outcome_count desc;**

We used a grouped bar chart for side-by-side comparisons between the two outcomes for each category. This gave us an overview of the number of successful and failed campaigns for each category .



Conclusion

Our analysis on the crowdfunding data provided valuable insights on campaign distributions, outcomes and trends. Transforming and organizing

the data, helped us query and visualize key patterns. Categories with the highest numbers of campaigns are shown, presenting an understanding of areas that are dominating the crowdfunding aspect. The analysis of campaign outcomes similarly highlighted the success rates and failure patterns across categories, offering insights for future planning.

Exploring the campaign distributions by country, showcases regional hotspots for crowdfunding efforts. Taking a closer look at successful and failed campaigns in each category provided data to understand audience preference and market demands. Combining SQL queries, data visualization, and database management allows the project to successfully demonstrate structured analysis that uncovers trends and helps in strategic decision making in the crowdfunding domain.