Campaign Data Analysis

Project 2 Group 6

By: Gorgina Kareem, Kriti Khatri, Shahla Shahnawaz, and Monica Mitry

Introduction/Executive Summary

With this group project we were provided data for crowdfunding sources and contacts. The goal was to clean up the data in order to design a new database. Through the database we can then analyze the data to determine what is the most popular crowdfunding category, what are the outcomes based on goal, which country had the most successful campaigns, How many successful campaigns were there in each subcategory, and which subcategory has the highest number of successful campaigns?

Extract/Transform

Utilizing the provided files *crowdfunding.xlsx* and *contacts.xlsx* we wanted to clean up the data in order to export four CSV files to create our Crowdfunding Database. To begin it was necessary to break out the Category/Subcategory found in the *crowdfunding.xlsx* file. In order to accomplish this that column needed to be split by the backslash found in the cell. After they were split we needed to create separate lists for the unique categories and subcategories. After the lists were created it was necessary to find the number of distinct values for each category and subcategory in order to create a numpy array. To help with the formation of the database in the future we needed to create a list comprehension by adding “cat” and “subcat.” Next it was necessary to create the DataFrame for the two before exporting each into their own CSV file.

Next, a Campaign DataFrame was created by first reading *crowdfunding.xlsx* into pandas. Then, *category* and *sub-category CSV* were also both read into pandas DataFrame. Next, a copy of dataframe named *campaign_df* was created and renamed the following columns: *blurb': 'description, 'launched_at': 'launch_date', 'deadline': 'end_date'*. After renaming the columns, the *goal* and *pledged* columns were converted to a *`float`* data type. Also, *launch_date* and *end_date* columns were formatted to datetime format. Last, Category and Sub-Category columns were split into separate columns and each of them had unique_id assigned.

Lastly, with the *contacts.xlsx* it was necessary to begin with converting each row into a dictionary using json as the data was missing necessary columns to create the pandas DataFrame. With the list of dictionaries it was then possible to transform the data in the DataFrame to create the *contact_id, name*, and *email* columns. After this the name column was split into two columns named *first_name* and *last_name.* We then dropped the *name* column. Before exporting the file into a CSV file the columns were reordered to be in this order: *contact_id, first_name, last_name,* and *email.*

Database Design/Load

Once the clean csv files for Category, Sub-Category, Contact and Campaign data were extracted, we use quickDBD (https://www.quickdatabasediagrams.com) for creating

relational database diagrams. We design the database diagrams and exported the schema as sql file. The sql file was used to create new crowdfunding database in PostgreSQL tools, pgAdmin. We create four tables for “Category”, “Sub-Category”, “Contact” and “Campaign” using the schema. The schema sql file was edited to use default value instead of “Not null” in the campaign before exporting the datasets into the created tables in postgres. Once out tables were created, next step was to import the respective csv dataset in the tables, order wise based on primary and foreign keys assigned in the tables.

We used SQLAlchemy ORM, jupyter notebook to connect the CSV data in pgAdmin. Once we had the csvs imported, we did five queries in the postgres. The queries are discussed in the following data analysis section.

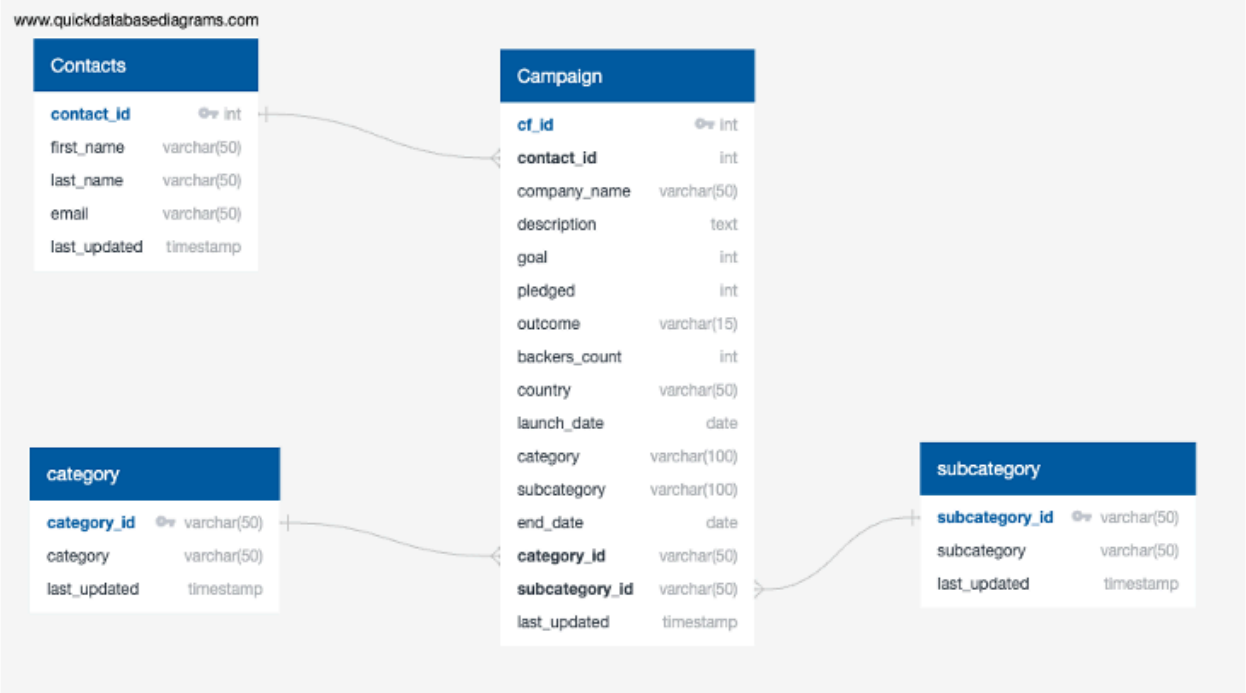

Figure 1: Image of the quickDBD diagram.

Data Analysis:

1. **What is the total number of backers per category and per outcome?** The first two sql queries aggregates the total number of backers found within the Campaign table. One query focuses on determining the total number of backers based on the outcome of the campaigns. The results showcase that successful campaigns has the highest amount of backers with a total of 480,898, followed by failed campaigns with a total of 213,164.

The second query showcases the total number of backers grouped by campaign categories. The data expresses that the theater category has the highest amount of backers with a total of 264,269. Immediately following that is the music, film & video, technology, and publishing with the least amount of backers with a total of 52,619.

2. **Which country had the most successful campaigns?** This query retrieves the number of successful campaigns for each country by grouping the data by county and filtering campaigns with a successful outcome. The results were organized in descending order based on the count of successful campaigns. With a significant lead, the US had the most number of successful campaigns with 436 while the other countries remain behind by over 400 successful campaigns. Among the countries Great Britain leads with 28 campaigns, then Italy, Australia, and lastly Canada.

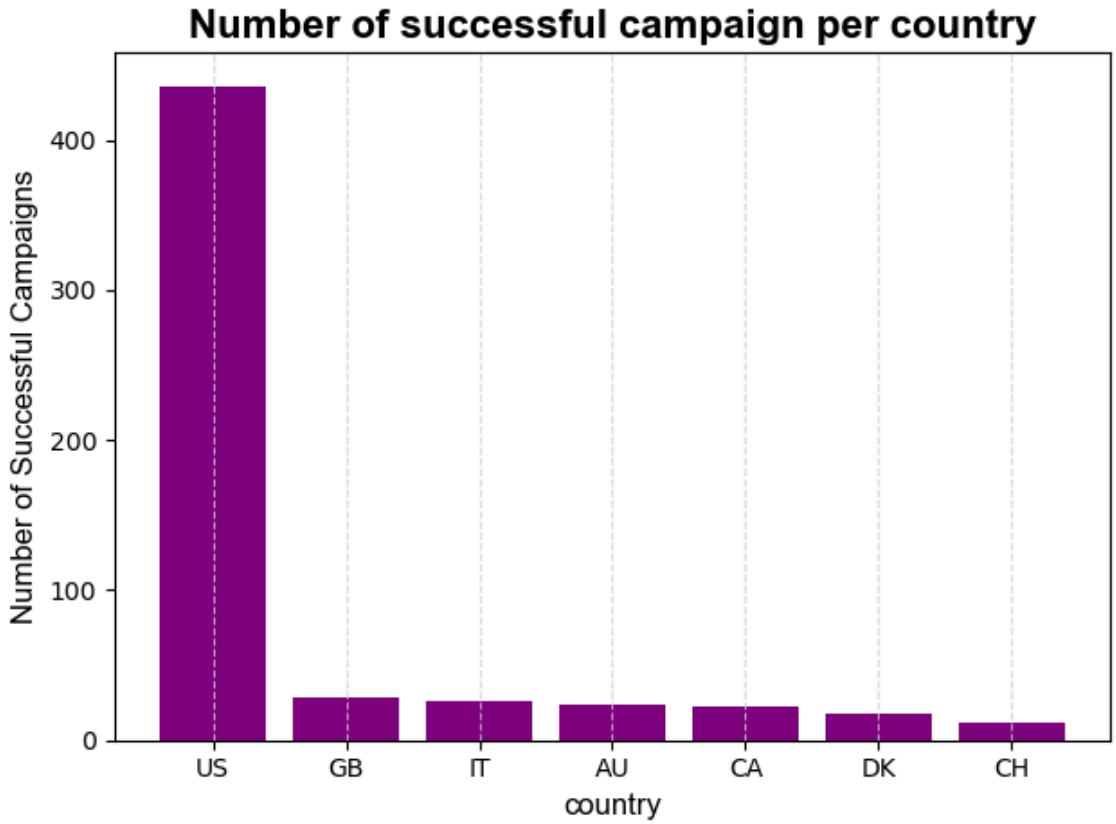

Figure 1: Bar graph depicting the total number of successful campaigns per country.

3. **What is the total amount pledged per subcategory?** The third query calculates the total pledged amount for the campaign subcategories by totaling the pledged values. The data shows that plays was the highest pledged total with $15,763,227. With a large difference the rock subcategory follows with

$3,603,659, while documentary, web, and indie rock round out the top five subcategories.

4. **What was the total number of success campaigns by subcategory?** In the final query it counts the total of successful campaigns grouped by their subcategories. Form the output, the subcategory with the highest number of successful campaigns at 187 is plays. Immediately follows it rock with 49, web with 36, documentary with 34, and wearables with 28. The plays significantly outperformed the other subcategories by over 130 campaigns.

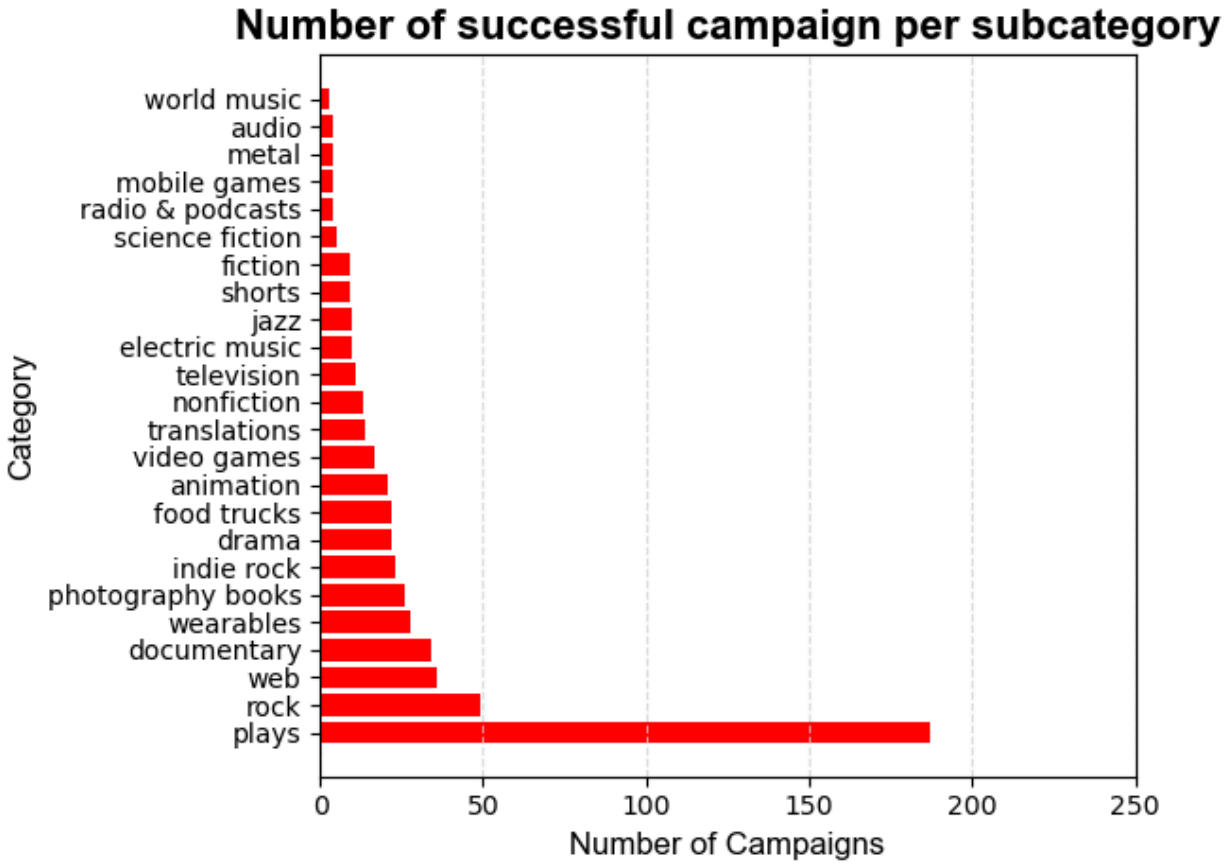

Figure 2: Bar graph depicting the total number of successful campaigns per subcategory.

Bias/Limitations

1. **Selection Bias:** If the data is not representative of the entire crowdfunding platform, the results could be biased.
2. **Reporting Bias:** Some campaigns may not report their outcomes accurately, or they may not provide data on goals and success rates, leading to incomplete or inaccurate insights.

3. **Success Criteria Bias**: The criteria used to define success might not align with how the stakeholders define success
4. Subcategory Classification: The classification of subcategories could be arbitrary or inconsistent, and different platforms might classify campaigns differently, limiting the ability to compare across datasets and platforms.

Conclusions

1. **Most Popular Crowdfunding Category:** Through the analysis of the category data, we identified the most popular crowdfunding categories based on the frequency of campaigns within each. This helped to highlight which sectors or causes are most likely to attract backers and funding.
2. **Outcomes Based on Goal:** By analyzing the relationship between campaign goals and outcomes (such as pledged amounts), we could determine whether campaigns with higher goals had a higher chance of success. This analysis provides valuable insights into how ambitious campaign goals impact the likelihood of success.
3. **Country with the Most Successful Campaigns:** By aggregating campaign success data by country, we were able to determine which country had the most successful campaigns. This can give us a better understanding of the geographic areas where crowdfunding is most successful, and how regional factors may play a role in the success of campaigns.
4. **Successful Campaigns by Subcategory:** We examined the number of successful campaigns within each subcategory, providing insights into which specific niches or types of projects tend to be more successful. Identifying subcategories with the highest number of successful campaigns allows us to understand what types of crowdfunding efforts resonate more with backers.

Reflections

1. **Data Cleaning and Transformation:** The process of cleaning and transforming the data, such as splitting the category and subcategory columns and dealing with missing values, was crucial to ensuring that the analysis was based on accurate and well-structured data. Without this step, the analysis could have been skewed by incomplete or inconsistent information.
2. **Challenges with Data Completeness:** A major challenge encountered during the project was the incomplete or inconsistent data in both the crowdfunding and contacts files. For instance, some campaigns lacked detailed outcome information or had missing launch and end dates. Overcoming this by creating workarounds and using assumptions when necessary helped to move the project forward but also emphasized the importance of clean, complete data for accurate insights.

3. **Database Design Considerations:** The process of designing the database using QuickDBD and creating the corresponding tables in PostgreSQL highlighted the importance of a well-structured schema for efficient data storage and retrieval. The database design provided a foundation for future analysis, ensuring scalability and ease of access for large datasets.
4. **Insights for Future Campaigns:** The insights gathered from the analysis, particularly on what factors lead to campaign success (such as goal size and subcategory focus), provide valuable guidance for individuals or organizations looking to run crowdfunding campaigns. These conclusions could help optimize strategies to increase the likelihood of success.
5. **Limitations and Future Work:** While the analysis offered valuable insights, there are limitations to the current dataset, such as missing data or the inability to account for external factors influencing campaign outcomes. Future work could involve gathering more granular data (e.g., campaign updates, social media activity) to provide a more complete picture of the factors driving crowdfunding success.
6. **Reflection on Database Performance:** The choice of PostgreSQL as the database solution proved efficient for handling the dataset's size and structure. The normalization of the data into separate tables helped ensure faster query performance during analysis. However, in future projects with larger datasets, optimizing queries and indexing could further improve database performance.

Final Thoughts

This project reinforced the significance of careful data cleaning and transformation in ensuring the quality and accuracy of analysis. By applying these best practices, we gained valuable insights into crowdfunding success factors and how various elements like category, goal size, and geography influence outcomes. Moving forward, further data exploration and validation could improve the overall results and broaden the scope of analysis, leading to even more actionable conclusions for campaign organizers and stakeholders.