Project 2 Write-Up

Seven George, Sabina Kamalova, Hannah Lashway

DATA-PT-EAST-APRIL-041524-MTTH
Instructor: Alexander Booth
July 24th, 2024

Table of Contents

Introduction

Crowdfunding is a popular concept often used by small and local businesses, entrepreneurs, and independent projects to raise funds for a variety of reasons whether it be renovations, new equipment, or developing a new product or application. Crowdfunding's main appeal is that it harnesses the power community and social media to let many individuals share and contribute small donations that pool together to meet or exceed a specific monetary goal. There are several popular websites for this, one being GoFundMe, though they are more geared towards helping individuals with unideal circumstances, the other, Kickstarter is much more aligned with the data in our dataset in that they are more focused on raising funds for businesses and entities that with this funding can benefit the community.

Our dataset consisted of two Excel sheets titled Crowdfunding and Contacts, each with 1000 rows of data and was last updated Nov. 11 of 2020. The Crowdfunding file held specific crowdfunding instances and contained relevant column names and hence data such as the name of the company that started the campaign, what they were raising funds for, their goal, how much money had actually been raised, the number of backers, the country they were raising funds from, start and end dates of the campaign, the category of their campaign, and their campaign status (i.e. successful, failed, live). The Contacts file contained the contact information such as first and last name and email address for an individual associated with each crowdfunding instance in our Crowdfunding file.

The main goal of this project was to rearrange the data in these two files into 4 separate csv files so they could be loaded into Postgres where we could perform queries and from there generate subsequent data visualizations that helped us answer questions regarding the popularity of different crowdfunding campaigns from both the perspective of the businesses starting the campaigns and from the perspective of campaign backers.

Database Design Considerations

We began designing our database utilizing QuickDBD to achieve a visualization of our four tables and their respective column names. Seeing that our Campaign table contained column names that were referenced in our Category, Sub-category, and Contacts tables let us know that those column names would be classified as foreign keys in the Campaign table and as primary keys in their respective table. Additionally, we examined our dataset rows to determine the variable classification for each column and whether they would be classified as null or not null. In our case, we classified all our columns at not null.

ETL Code Overview

As noted in our Introduction, our original dataset was made up of two Excel files however we needed to extract the data inside them and transform them into four usable tables that could be loaded into our schema in Postgres for queries and visualizations. To begin we read both our Crowdfunding and Contact files into Juypter Notebooks and converted them into Pandas dataframes.

To create the Category and Subcategory dataframes, we inspected the columns and split them in two using str.split() method with the delimiter '/' and 'expand=True'. Once those were split into separate columns, we used 'unique().tolist()' to create a list of unique values found in each column. We then created an array with numerical ID's using 'numpy.arrange()' and then list comprehensions to add the prefixes 'cat' and subcat' respectively. Last, we created 2 new dataframes, Category and Subcategory, using the new arrays and exported them to csv files.

To create our Campaign dataframe, we began by using a copy of the same Crowdfunding dataframe that we used for our Category and Subcategory dataframes. The first task was renaming the 'blurb', 'launched_at', and 'deadline' columns using the rename() function to the names specified in the project instructions. We had to convert the datatypes of several columns, for the 'goal' and 'pledged' columns we used the 'astype()' method to convert them to float values and the 'to_datetime' method to convert the 'launch_date' and 'end_date' columns to datetime format. Now that all the columns in the Campaign dataframe were assigned their correct datatype and names, we could merge the Campaign dataframe with both the Category and Subcategory dataframes we previously created. We used a left merge so that we could include all the columns and merged on the 'category' and 'subcategory' columns respectively. The last thing we needed to do before exporting our dataframe to a csv file was to clean it up and drop any unnecessary or duplicate columns. In this case we dropped the 'staff_pick' and 'spotlight' because they were unnecessary, the original 'category & sub-category" column and 'category' and 'subcategory' columns because they were repetitive data as we only needed the 'category_id' and 'subcategory_id' columns to reference our Category and Subcategory databases.

To begin, we loaded the Contact excel file containing the contact information. Using the pandas function `pd.read_excel()`, we read the file located at the specified path and loaded it into a DataFrame called `contacts_df`. This step was crucial as it allowed us to work with the data in a tabular format. The data in each cell of the Excel file was in JSON format. To handle this, we iterated over each row of the DataFrame using `contacts_df.iterrows()`. For each row, we used the `json.loads()` function to convert the JSON string into a dictionary. We appended each valid dictionary to a list called `contacts_list`. This step ensured that we could work with the data in a more accessible dictionary format. We also included error handling to skip any rows that did not contain valid JSON, ensuring the robustness of our code. Once we had a list of dictionaries, we extracted the values from each dictionary using list comprehension. We then obtained the column names from the keys of the first dictionary in our list. Using this extracted data, we created a new DataFrame called `new_contacts_df`. This DataFrame contained all the contact information in a structured, tabular format, ready for further processing. To make the contact information more usable, we split the 'name' column into two separate columns: 'first_name' and 'last_name'. First, we ensured the 'name' column was treated as a string. Then, using the `str.split()` function with appropriate parameters, we split the names into first and last names. We assigned these split values to new columns 'first_name' and 'last_name', and subsequently dropped the original 'name' column. This transformation made the contact data more granular and easier to work with. Lastly, we reordered the column names to ensure that they matched with our database in Postgres, this was crucial because without the correct order we couldn't import our csv file. The final step involved saving the cleaned DataFrame to a CSV file. We determined the current working directory using `os.getcwd()` and constructed the path for the output CSV file. Using the `to_csv()` function, we saved the DataFrame to the specified path without including the index. This step ensured that the cleaned and structured data was saved in a portable and accessible format, ready for further use or analysis.

Analysis

Looking at our dataset, and understanding the cultural relevancy and popularity of crowdfunding, we posed two questions: First, what categories of business ventures were most likely to look to a crowdfunding platform to seek funding and secondly, what categories of crowdfunding campaigns were most popular by looking at the quantity of support received and percentage of successful campaigns.

We first created a bar chart that showed the number of successful and failed campaigns for each category. While we didn't need the specific 'failed' and 'successful' information until our second question, our chart also gave us the total counts for each category, giving us insight to answer our first question. Looking at our graph, we can see a significantly higher amount of campaign registrations for the 'theater' category with over 300 registrations. Our next two categories that showed significant registrations were the 'music' and 'film & video' categories with and average of around 160 registrations; every other category saw less than 100 registrations. From this we can draw that the types of business ventures that were most likely to seek funding from a crowdfunding platform could be categorized along the line of the traditional arts as our highest registrations are for campaigns categorized under 'theater', 'music', and 'film & video'. This is not too surprising as historically, the traditional arts have been underfunded from an academic perspective and in terms of careers, they are notorious for being highly competitive and thus difficult to secure financial success hence it makes sense that these categories are the most likely to seek external funding.

To answer our second question, we looked at the first bar graph we created and additionally created a box plot and pie charts for each category for further analysis. Off the bat, by looking at our pie charts, the categories with the highest percentages of successful campaigns were 'journalism' with 100% success, 'photography' with 70.3% success, and 'technology' with 69.6% success. However, it should be noted that both the 'photography' and 'journalism' categories both held less than 50 registrations. Looking to our box plot for further information we observed that categories 'technology' and 'film & video' appeared to be more popular with backers based off the broader spread of backer counts and higher medians. This further proves what our pie charts and bar chart illustrate in terms of successful campaigns, with 'technology' holding a 69.6% success rate and 'film & video' holding a 63% success rate. It should be noted that both these categories have the presence of outliers which suggests that while the campaigns in these categories attract a moderate number of backers, there are standout campaigns that achieve exceptional success in attracting a large number of backers. Overall what we can observe, is that although some categories such as 'journalism' and 'photography' have higher success rates, categories like 'film & video' and 'technology' are more representative of overall category popularity given that they both have high success rates, a high count of median backers, and outliers that are more indicative of higher support and engagement.

Biases + Limitations

In any data analysis project, it is essential to acknowledge the inherent biases and limitations that could impact the findings and interpretations - this project is no exception. Our dataset is derived from two Excel sheets, each with 1000 rows, and was last updated in November 2020. This relatively small and potentially outdated dataset may not fully capture recent trends or the broader scope of crowdfunding activities. Additionally, the dataset focuses primarily on campaigns similar to those on Kickstarter, potentially excluding data from other popular crowdfunding platforms like GoFundMe or Indiegogo, which may have different user bases and campaign characteristics.

Looking first at the categories and subcategories in our dataset, we understand that they may not represent the full spectrum of crowdfunding projects. Some categories might be underrepresented or missing, leading to biased insights. Certain categories, such as "Journalism" and "Photography," have fewer entries, which may have skewed their success rate calculations because they had objectively less competition on a categorical level that other categories like 'theater' or 'technology'.

Our analysis focused on the percentages of successful and unsuccessful campaigns as well as the number of backers as the primary success metrics for each category. While these are important, they do not encompass all factors contributing to a campaign's success, such as social impact, long-term viability, or community engagement. In addition, the definition of success itself is binary in that a campaign either was or was not successful and does not take into account the degree of success for campaigns that may have far surpassed their original goal or the potential for near-successful campaigns to achieve their goals through other means. Additionally outliers, particularly in categories like "Technology" and "Film & Video," suggest exceptional success but can disproportionately influence averages and medians. These outliers may distort the perceived popularity and effectiveness of certain campaign types.

Looking at our dataset itself, we also run into other potential biases. The process of transforming the dataset into separate CSV files and merging them into a relational database introduces potential errors. Any inaccuracies in these steps could affect the final analysis. The splitting of the "category & sub-category" column into separate fields relies on consistent formatting, which may not always be the case in a larger, more diverse dataset. Speaking of our dataset, it was last updated in November 2020, almost four years ago and thus the findings may not reflect the current state of crowdfunding, which can change rapidly due to socio-economic factors, technological advancements, and shifts in public interest.

In summary, while the project provides valuable insights into crowdfunding trends, it is important to consider these biases and limitations when interpreting the results. Future analyses could benefit from a more extensive and updated dataset, inclusion of additional success metrics, and careful handling of outliers to provide a more comprehensive understanding of crowdfunding dynamics.

Conclusion

Our project analyzed a dataset consisting of two exel files containing data from Kickstarter-like campaigns focusing on campaign details, funding goals, amounts raised, backer numbers, and campaign status, alongside contact information for campaign representatives. The primary goal was to restructure this data into a format suitable for PostgreSQL to perform queries and generate visualizations in Jupyter Notebooks to reveal insights into the popularity and success of various crowdfunding campaign categories.

Our findings indicated that categories like 'theater,' 'music,' and 'film & video' were the most likely to look to crowdfunding as a viable form for funding, displaying a trend towards artistic endeavors. While categories like 'journalism' and 'photography' had high success rates, they had fewer entries, which we took into consideration when addressing dataset biases. Categories like ‘technology', and 'film & video' saw relatively high success rates and we feel provided a broader representation of crowdfunding dynamics due to their higher median backer counts and significant outliers indicative of exceptional success.

Our dataset had various limitations, such as its relatively small size, potential outdatedness at nearly four years old, and not being inclusive of data that is more reflective of other popular crowdfunding websites which could have had an influence in determining what are truly the most successful types of campaigns. In the future, a more extensive and updated dataset would be ideal, and if we continued to work with this dataset, performing queries such as the average campaign goal for all campaigns, average campaign goal for successful campaigns, and further breaking down the success rates on subcategories could provide us with additional success metrics for a deeper understanding of crowdfunding trends.