CROWDFUNDING ETL PROJECT

Kade Rivers

Amar Patil

Data Analytics Bootcamp

Project 2

Professor Booth

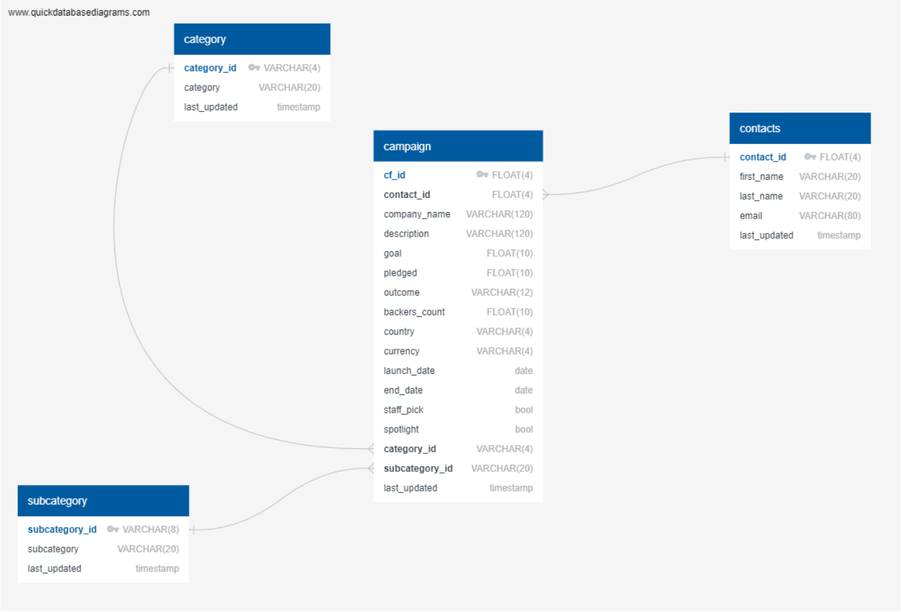
July 21, 2024

Introduction

This project is based on data and information provided by the Data Analytics Bootcamp, and all groups were provided the same information and task for an ETL (Extract, Transform, Load) mini project where partners or small groups were to create an ETL pipeline using Python, Pandas, and dictionary methods or expressions to complete the project. Deliverables include an ERD Diagram, Jupyter Notebooks to show data extraction, transformation, load, and the creation of multiple data frames and visualizations in order to best analyze the given data.

Database Design Considerations

Original data files for this project were given in Excel format where each group extracted the data to form various databases and chose the design for the database. Our group utilized quickdatabasediagrams.com to create the database with primary and foreign keys and Jupyter Notebook was used to load tables.



category	
category_id	VARCHAR(4)
category	VARCHAR(20)
last_updated	timestamp

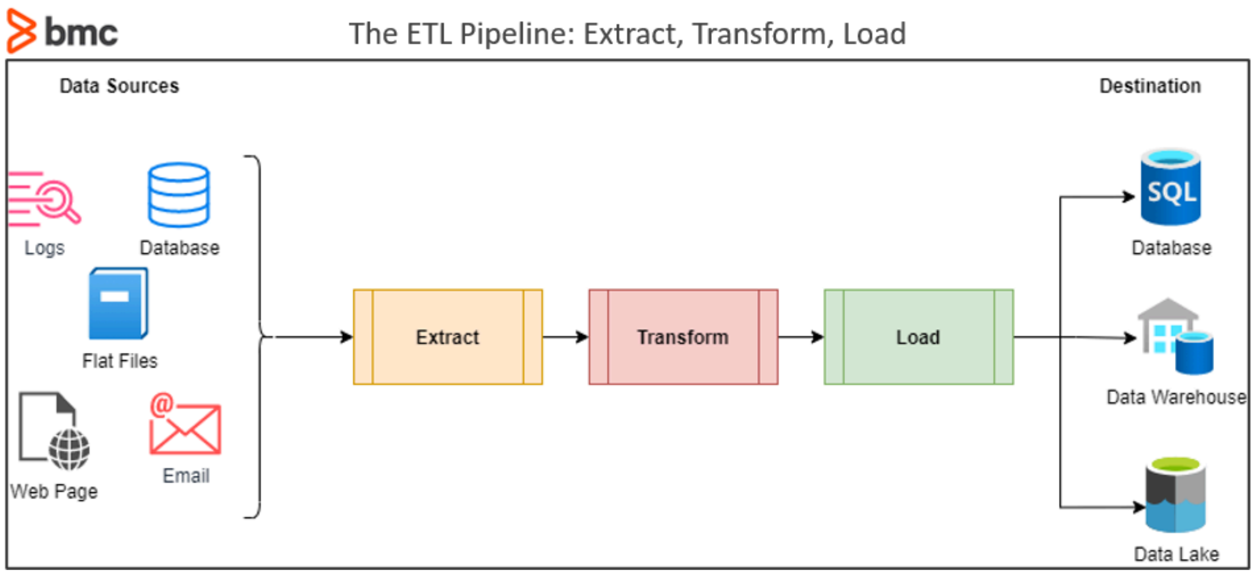
campaign	
cf_id	FLOAT(4)
contact_id	FLOAT(4)
company_name	VARCHAR(120)
description	VARCHAR(120)
goal	FLOAT(10)
pledged	FLOAT(10)
outcome	VARCHAR(12)
backers_count	FLOAT(10)
country	VARCHAR(4)
currency	VARCHAR(4)
launch_date	date
end_date	date
staff_pick	bool
spotlight	bool
category_id	VARCHAR(4)
subcategory_id	VARCHAR(20)
last_updated	timestamp

contacts	
contact_id	FLOAT(4)
first_name	VARCHAR(20)
last_name	VARCHAR(20)
email	VARCHAR(80)
last_updated	timestamp

subcategory	
subcategory_id	VARCHAR(8)
subcategory	VARCHAR(20)
last_updated	timestamp

Extract/Transform/Load Overview

The purpose of an ETL or Extract, Transform, and Load is to be able to streamline the viewing process of a large amount of data from multiple sources. This may include consolidating the data or changing the format of the data being used in the ETL. An overview of ETL is seen here:

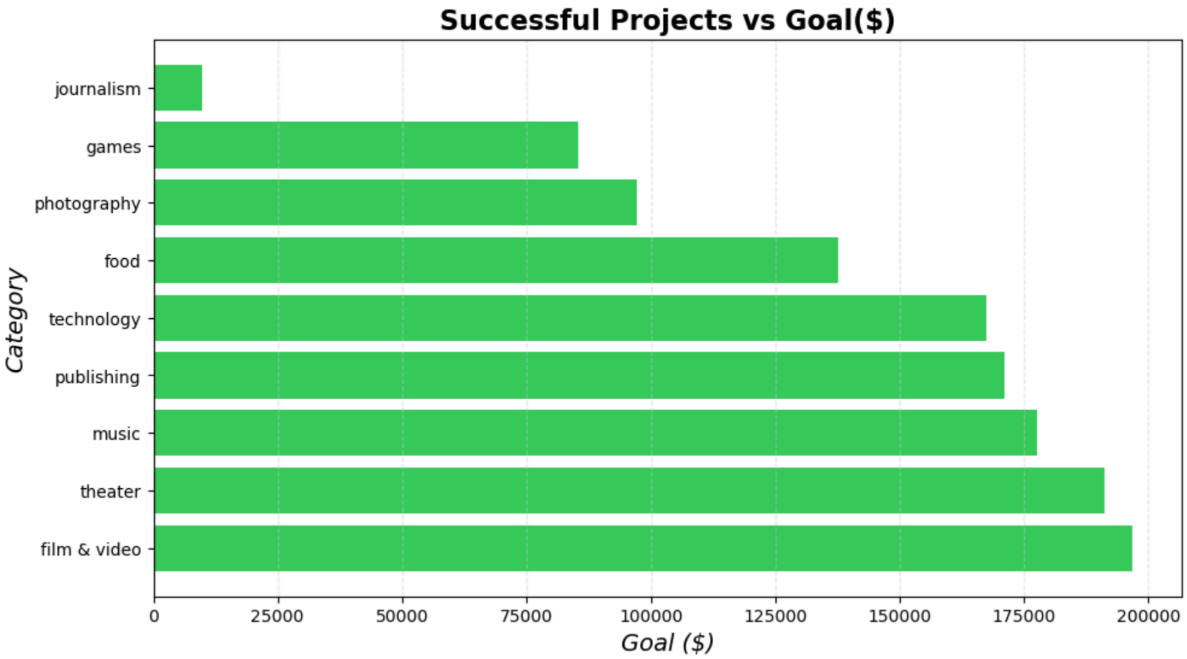


For this group project, we utilized the Excel workbooks, Jupyter Notebook, Postgres, the Quick DBD website and various libraries within the Jupyter Notebooks to be able to extract, transform, and load the given data.

Analysis

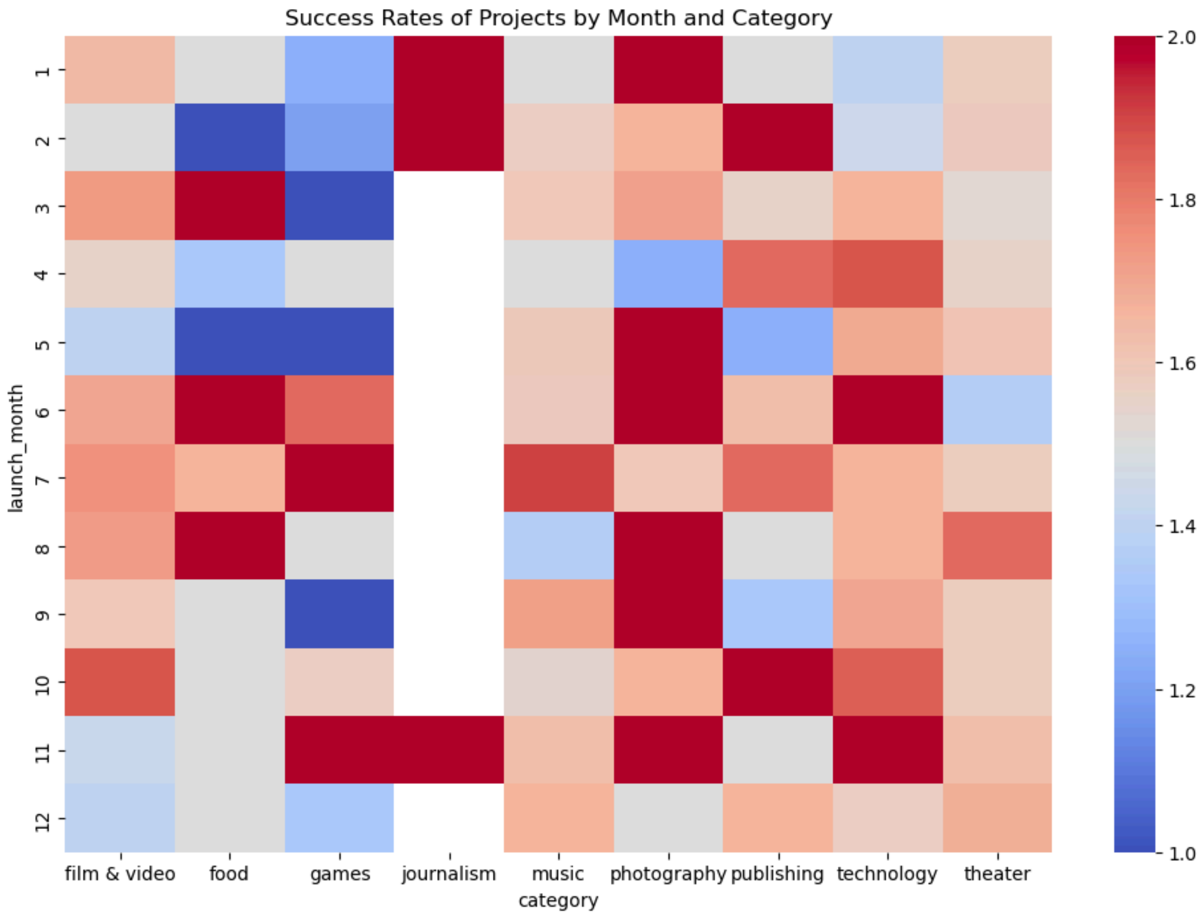
A combination of SQL and ORM queries were used to create dataframes which were used in visualization. The analysis notebook in the git repository provides several visualizations with three of these visualizations offering insight into the crowdfunding successes.

Visualization 1:



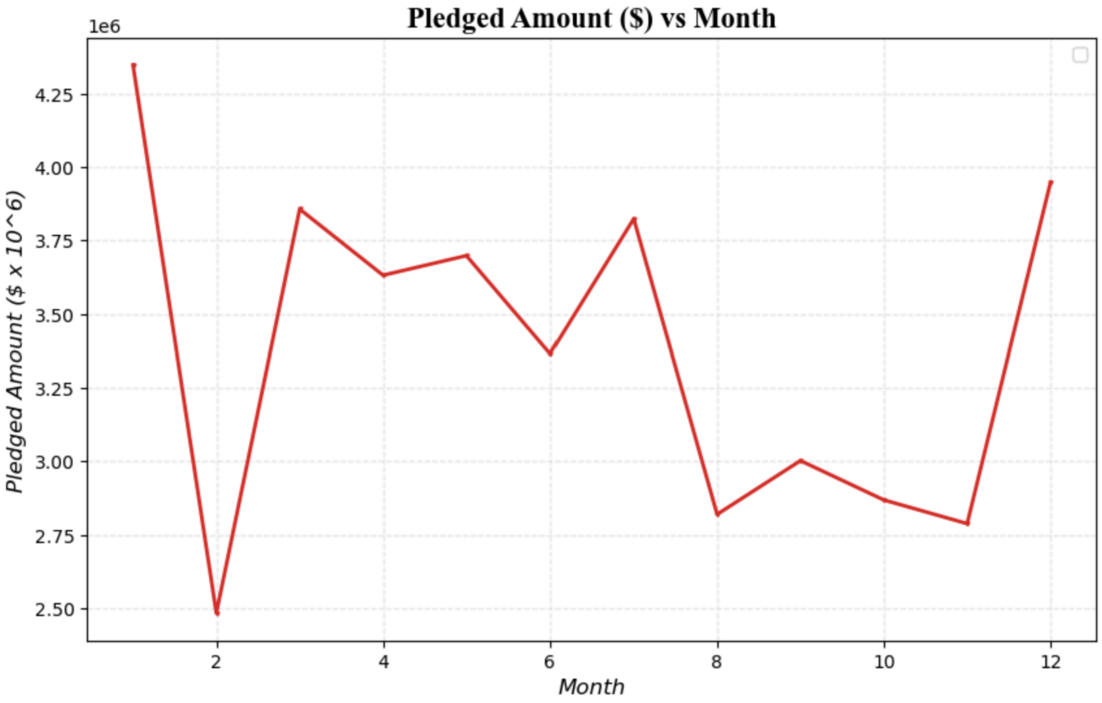
Visualization 1 shows that the Film and Video category of crowdfunded projects have the highest goal dollar amount successfully funded and that the category of Journalism has the lowest. A **limitation** to this visualization is that it does not provide the overall number of projects from each category that were successful, so it cannot be determined if the film and video also had the most success overall or if it just had the highest goal with a successful funding amount.

Visualization 2:



Visualization 2 shows the overall successes and failures for categories by month start date with the utilization of a heat map. This heat map indicates that overall the Theater and Photography have successful projects started throughout the year and that Music, Publishing, and Technology are fairly close behind. It also indicates that there are far less Journalism projects submitted throughout the year as there were some months without data to utilize for this heatmap. Food and Games both show a variety of overall successful vs failure months with several neutral colors throughout the heatmap indicating the presence of both success and failures for starting projects in those months. The analysis of this heatmap also indicates that overall, starting any project in the months of June, July, or August may impact the success rate to be higher as well as projects started in December through May for categories such as Film and Video, Food, and Games will have a lowered chance for a success rate given the outcomes on the heatmap. This heatmap visualization can be useful in predicting the categories with the highest chance of success as well as which months may be best to launch the project for an even higher chance of success. Visualization two could also be utilized to determine which month to launch a tougher category such as Film and Video, Food, or Games if those are the desired categories to be launched. The **limitation** of this visualization is not having the exact numbers of successful and projects per category listed per month as the visualization would lose impact with so much additional text.

Visualization 3:



Visualization 3 is a visualization with months and the amount pledged per month. This visualization works well with the heatmap to determine which months overall have the highest amount of money actually pledged for all projects. While this visualization does not include categories, it can be shown to indicate that January is the month with the highest amount of money pledged for all projects followed by December, March, then July. The lowest amounts of money pledged fell in February, November, and August. The **limitations** to this visualization are that it does not indicate the pledge amount by category, or the month the projects started or ended based on the pledge. However, used in conjunction with the heatmap, both can provide insight into determining category, start date and an idea of a realistic goal if further analysis is completed as characteristics for a starting project are narrowed down.

Bias and Further Limitations

No analysis was done by project country which could lead to a different outcome of success vs failure and should be taken into consideration if this analysis will be utilized for future decisions regarding new projects. Exact numbers of successes, failures, and cancellations were not included in the visualizations, however, they were included in the Jupyter Notebooks and can be utilized for further analysis to gain a better understanding of success and failure rates averages for the amount of projects completed regardless of success or failure.

Conclusions

Project two increased our teamwork skills when working on shared notebooks and working together in place of simply delegating tasks and completing them individually. Using the skills to create the ETL and then create visualizations based on the data we extracted provided a unique challenge of determining the most important data to visualize from a variety of information that can all affect the outcome of a chosen project, and many other visualizations were discussed, and some created, to determine which visualizations showed the most appropriate amount of information that could be useful for an analysis and recommendation for someone looking to begin a new project.

While the visualizations are a great starting point, this analysis and project could be taken in a variety of directions for a future project individual or company who may already have a category in mind, who may want to see if Staff Pick or Spotlight make a difference in success rates, or if the goal amount vs the rates of success and failure can help determine the best amount to set for a future goal.

Additional time on the project would also allow data analysts to create further in-depth analysis and a wider variety of visualizations related to the given data.

Reflections

This project had less group members and was more hands-on with using our various skills obtained throughout the Data Analytics Bootcamp. While the data was pre-selected for all groups, the visualizations were left to each group to determine which visualizations best showcased the data, and this allowed us all to gain a better understanding of how to look at the data and determine which visualizations would make the most sense for the data story we wanted to tell. We believe that our visualizations would prompt further discussions and decisions as well as further digging into the data to take it further once some general guidelines were established based on our original findings.

Additional Sources

Websites:

https://app.quickdatabasediagrams.com/#/

https://www.bmc.com/blogs/what-is-etl-extract-transform-load-etl-explained/