Kickstarter Analysis

11/20/2020

The Laureates: Farzeen Najam, Olivia Olsher, Vincent Liu, Emile Therrien

Introduction and Data

Introduction

Have you ever had an ambitious idea but did not have resources to pursue it? Kickstarter is a crowdfunding platform for creators who need support for their projects. Since the platform launched in 2009, 19 million people have pledged to back up various projects, and nearly 190,000 projects have been successfully funded.

However, many more projects failed to reach their goals, and our group is interested in analyzing what made projects succeed and fail. The dataset Kickstarter Projects comes from Kaggle, which contains Kickstarter projects up until January 2018.

We will be exploring which variables influence the success of a Kickstarter project by observing which types of projects are more likely to be funded.

Our questions include:

1. Is the project goal (in USD) associated with its chance of success?
2. Is the project category associated with its chance of success?
3. Which category of project is more likely to succeed?

Our hypotheses include:

1. The more money a project asks for, the less likely it will be successful raising funding.
2. The project category is associated with the success rate, and certain categories perform better than others.

The goal of the project is to give future Kickstarter creators insight into which projects failed and succeeded. This will give them the tools to perform better against their competition, by giving estimates for what has and has not worked based on this historical dataset. Modeling which categories will be most successful best will give creators insight into predicted category success, assuming that there exists a relationship between project categories and success rates.

Data Description

```
## Rows: 378,661
## Columns: 15
## $ ID               <dbl> 1000002330, 1000003930, 1000004038, 1000007540, 10...
## $ name             <chr> "The Songs of Adelaide & Abullah", "Greeting From ...
## $ category         <chr> "Poetry", "Narrative Film", "Narrative Film", "Mus...
## $ main_category    <chr> "Publishing", "Film & Video", "Film & Video", "Mus...
```

```
## $ currency         <chr> "GBP", "USD", "USD", "USD", "USD", "USD", "USD", "...
## $ deadline         <date> 2015-10-09, 2017-11-01, 2013-02-26, 2012-04-16, 2...
## $ goal             <dbl> 1000, 30000, 45000, 5000, 19500, 50000, 1000, 2500...
## $ launched         <dttm> 2015-08-11 12:12:28, 2017-09-02 04:43:57, 2013-01...
## $ pledged          <dbl> 0.00, 2421.00, 220.00, 1.00, 1283.00, 52375.00, 12...
## $ state            <chr> "failed", "failed", "failed", "failed", "canceled"...
## $ backers          <dbl> 0, 15, 3, 1, 14, 224, 16, 40, 58, 43, 0, 100, 0, 0...
## $ country          <chr> "GB", "US", "US", "US", "US", "US", "US", "US", "U...
## $ `usd pledged`    <dbl> 0.00, 100.00, 220.00, 1.00, 1283.00, 52375.00, 120...
## $ usd_pledged_real <dbl> 0.00, 2421.00, 220.00, 1.00, 1283.00, 52375.00, 12...
## $ usd_goal_real    <dbl> 1533.95, 30000.00, 45000.00, 5000.00, 19500.00, 50...
```

This data set contains 15 variables and 378,661 observations, where each observation is one kickstarter project.

The categorical variables include the name of each project, the category of each project (music, narrative film, restaurant, etc.), a broader category of each (food, film, publishing, etc.), the crowdsourcing currency, the state of each project (failed, successful, or canceled), and the country of origin for each project.

The numerical data are the project's ID number, the monetary goal for each, the money pledged to each project, how many backers of each project; there are three numerical variables that are not self-explanatory, *usd_pledged*: conversion in US dollars of the pledged column (conversion done by kickstarter), *usd_pledge_real*: conversion in US dollars of the pledged column (conversion from Fixer.io API), and *usd_goal_real*: conversion in US dollars of the goal column (conversion from Fixer.io API).

There are also two date columns, one for the project launch and the other for the crowdsourcing deadline.

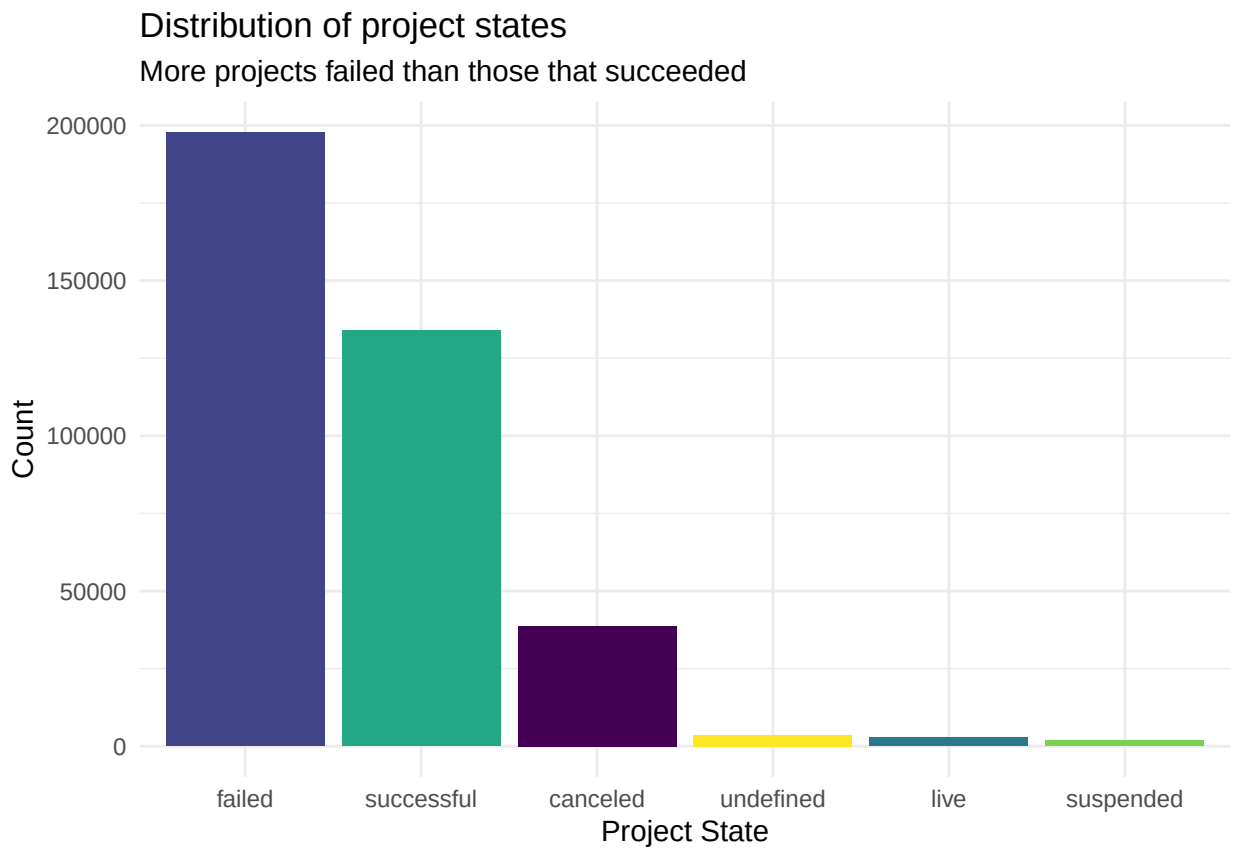
The data were collected from Kickstarter Platform likely using web scraping methods on their own site, to be used by data scientists to model whether or not a project will be successful or not when it is launched.

Exploratory Data Analysis

Overview of project state

Table 1: Count of each Kickstarter Project State

state	n
failed	197719
successful	133956
canceled	38779
undefined	3562
live	2799
suspended	1846



More projects failed than those that succeeded, with an average success percentage of 35.4%.

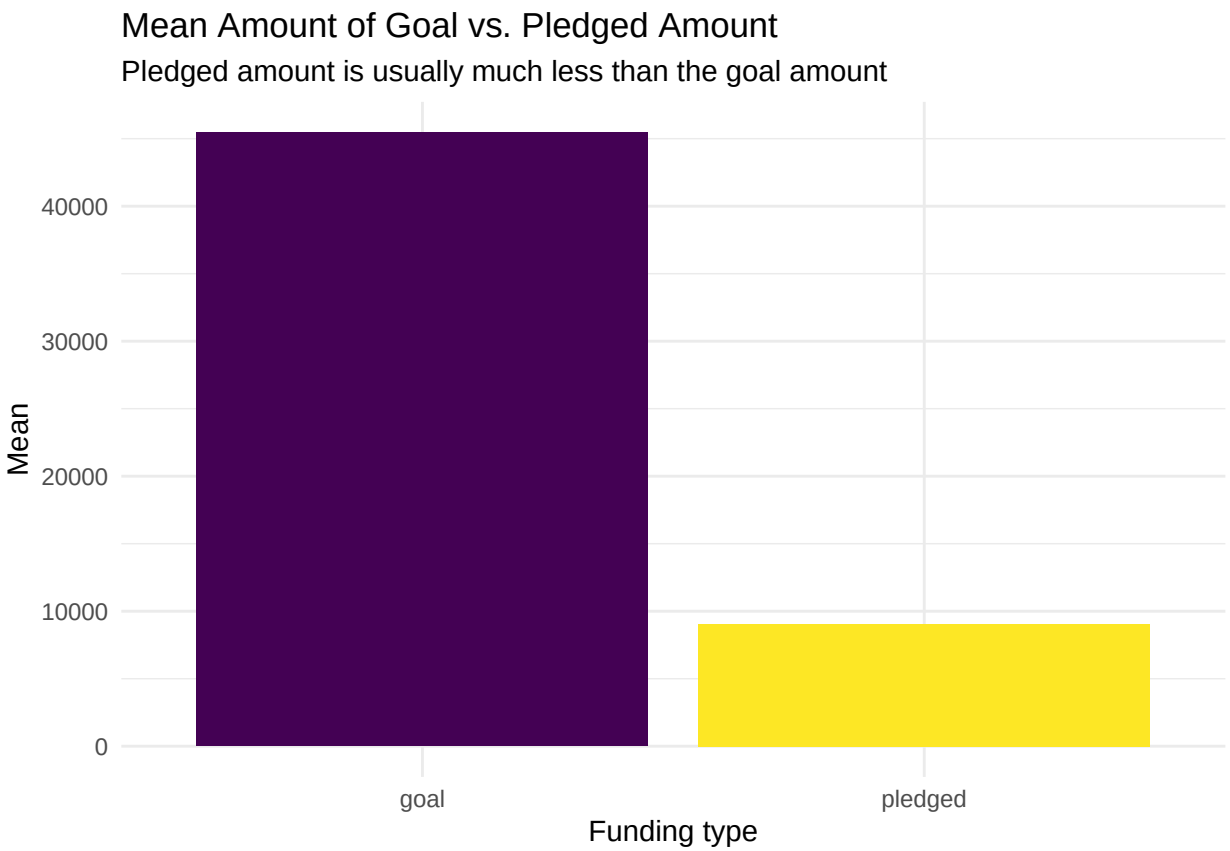
Overview of pledged and goal amount in USD:

Table 2: Overview of Pledged Amount

mean	median	sd	type
9058.924	624.33	90973.34	pledged

Table 3: Overview of Goal Amount

mean	median	sd	type
45454.4	5500	1152950	goal



The average funding pledged was 9058.92, with a standard deviation of 90,973.34. In comparison, the average funding goal was 45,454 with a standard deviation of 1,152,950. The observed differences between both groups is tremendous.

Methodology

First of all, we removed all projects that were *live* and currently asking for funding, so that any discrepancies seen would be negated. If we had included *live* projects, each project category may not have been representative of the population and project goal amounts may be skewed. Therefore, in the first section of the analysis, we overwrote the Kickstarter data with observations that are not *live*. To analyze success, rather than using the given variable "success", we created our own. The original variable "success" contained canceled, suspended, and undefined states along with successful and failed states. We had no indication if those projects that were canceled, suspended, or undefined met their goals and canceled prematurely or if they canceled due to no funding at all, which may skew our data. To analyze our data, we created an indicator for success: 1 being successful and 0 being unsuccessful (failed, canceled, suspended, or undefined).

Due to the sheer size of this dataset, using a simulation based method for analysis is not appropriate. To begin our analysis after removing *live* projects, we assessed whether there is a relationship between the amount of money a creator asks for and its success. We categorized projects by tiers, on a scale of 1-7, with Tier 1 asking for the least amount of funding and Tier 7 the most. We grouped Tiers 1-4 and 5-7 together when we ran our Central Limit Theorem (CLT)-based test, placing the lower and higher groups in the same category when running this test. We had to first determine a relationship between tiers and success, so we used a χ^2 Test.

After determining this relationship, we used a Logistic Regression Model to show the differences in success between Tiers. It would make intuitive sense if the projects that require less funding then they will be more

successful. These projects that ask for less funding should require a lower volume of money funded and meet their goal and on average meet more of their goals before project funding deadlines. Furthermore, we believed that these projects would require less backers donating money, assuming each backer donates an equal amount, and thus be dependent on a lower amount of people for funding.

After establishing a relationship between project success and the initial funding goal and modeling the predicted success based on each Tier, we used another χ^2 Test to determine if there was a relationship between project categories and their success. We used the variable *main_category* instead of *category* because the latter was far too specific for our purposes. *main_category* was composed of 15 distinct categories, each for a unique industry. We felt that 15 categories allowed our analysis to be broader and therefore each could encompass many more projects as not to pigeon hole a creator when using our analysis for their purposes. There may be possible crossover between main categories that we were unable to screen for, however, but we assumed this to be a negligible amount of projects, if it existed at all, and thus continued with *main_category* over *category* for analysis.

Following these analyses, we modeled each project's log-odds of success based on *main_category*. Success here was a boolean value, with 1 representing successful funding and 0 representing unsuccessful funding. This model enables future creators to think about their project in the larger scheme of a category and base their opinions off these values. We ran a Logistic Regression Model, where *Technology* was used as the reference level, and each value is based off success relative to the technology category. We used a proportionality level of 0.50 to determine if a project category was worth pursuing. A level greater than 0.50 meant that the category was predicted to have more successful projects than unsuccessful ones.

Results

Project Goal Amount and Success

In the Exploratory Data Analysis section, the first visualization gave an overview of project states, and the second revealed the large difference between the amount of money asked for and raised. However, from those two plots, we could neither see the goal and pledge amount of each successful and failed projects, nor could we know if whether or not there is a relationship between the goal and pledge amount. In lieu of this, we decided to examine the association between projects' success and their goal amount. To do so, we first came up with a claim that states our assumed association.

Claim: The amount of money a project asks for is related to its success in getting enough funding.

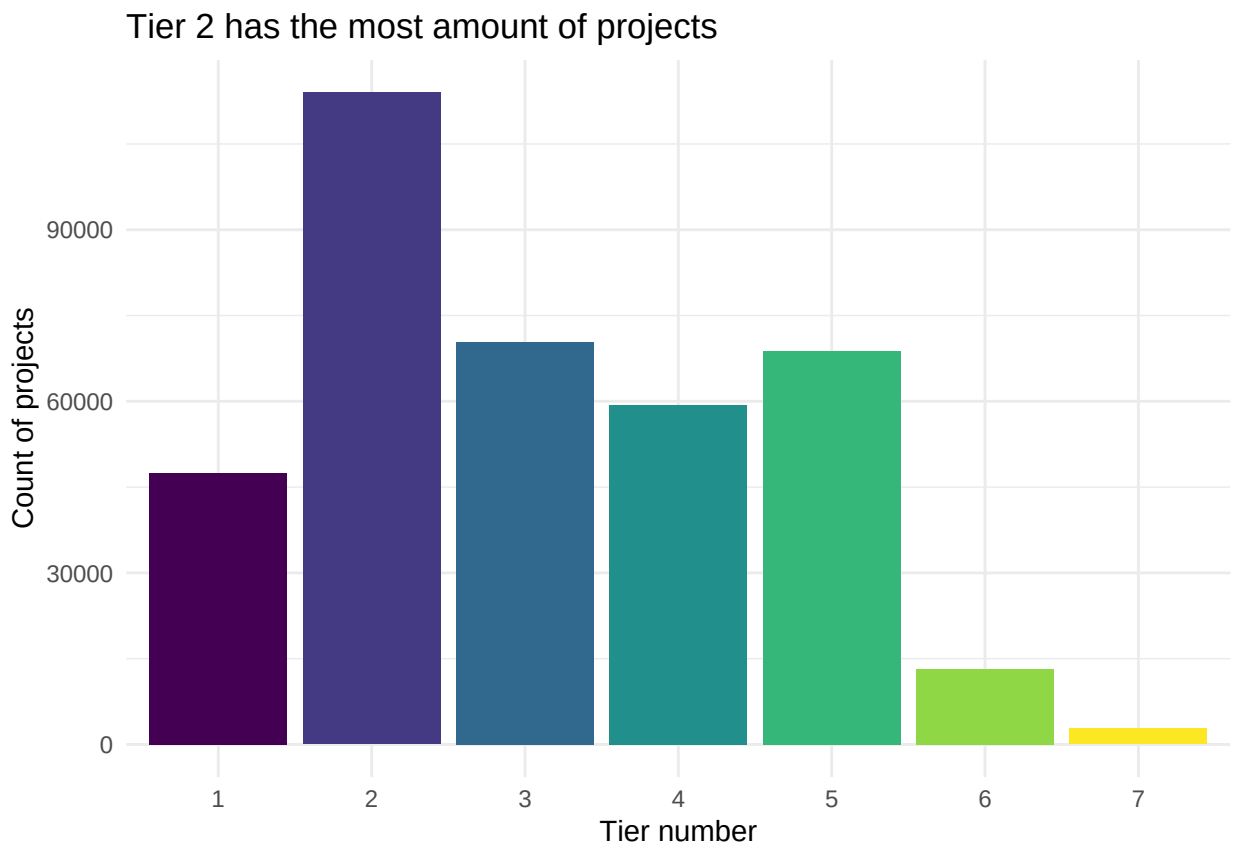
In order to perform a hypothesis testing on this claim, we need to quantify the goal amount into only a few levels rather than using the original discrete data (amount in USD). Otherwise, sample tests wouldn't work well because of too many data points. We decided that using a tier system would better suit our purposes of determining a relationship between success and how much money a project asks for. Furthermore, visualizing the project tier can give insight to how many projects fall in each range. Therefore, we decided to categorize the goal amount using the following metric and to create new variable named *usd_goal_real_tier* to classify *usd_goal_real* into tiers.

The tiers are as follows: Tier 1 $<$ 1,000 in goal USD, Tier 2 $\geq$ 1,000 and $<$ 5,000, Tier 3 $\geq$ 5,000 and $<$ 10,000, Tier 4 $\geq$ 10,000 and $<$ 20,000, Tier 5 $\geq$ 20,000 and $<$ 100,000, Tier 6 $\geq$ 100,000 and $<$ 500,000, and Tier 7 $\geq$ 500,000.

Table 4: Overview of Tiers

usd_goal_real_tier	numbers
1	47494
2	113993
3	70338
4	59285

usd_goal_real_tier	numbers
5	68727
6	13231
7	2794



Then, we created a new binary variable named *success_state* to represent whether a project was successful or not. If a project is not successful ("failed", "undefined", "suspended", or "canceled"), then it could carry a value of 0. Creating this binary variable gets rid of unnecessary project states so that we could focus only on successful projects.

Now we test our first claim using a CLT-based approach:

Hypotheses:

At the $\alpha = 0.05$ level:

- H_0: $project_{tiers}$ and $project_{success}$ have no relationship
- H_1: There is a relationship between $project_{tiers}$ and $project_{success}$, where $project_{tiers}$ is the variable *usd_goal_real_tiers* and $project_{success}$ the variable *success_state*.

Our following CLT approach is testing whether or not there is a relationship between how much money a project attempts to raise and its success.

```
##
##  Pearson's Chi-squared test
##
```

```
## data:  table(kickstarter$usd_goal_real_tier, kickstarter$success_state)
## X-squared = 19624, df = 6, p-value < 2.2e-16
```

Analysis of Results:

At the previously stated α level of 0.05, our χ^2 value is 19624 with 6 degrees of freedom and a p-value of less than 2.2e-16. Since our p-value is less than our α value, we have enough evidence to reject our null hypothesis that the Tier a project is independent of its success. There is enough evidence to suggest that a project's Tier is related to its success rate.

Since our p-value is less than our α value, we reject our null hypothesis that the difference in the proportion of projects with less money as goal and proportion of projects with more money as goal is less than or equal to 0. This means we reject the null hypothesis that success rate for projects with bigger money is equal or higher than projects with lower money.

Logistic Model for Success based on Project Tiers

Here we used a logistic regression model to predict project success based on the project's tier.

We used Tier 1 as our reference level.

```
## # A tibble: 7 x 2
##   term                estimate
##   <chr>                  <dbl>
## 1 (Intercept)           0.505
## 2 usd_goal_real_tier2  -0.0675
## 3 usd_goal_real_tier3  -0.147
## 4 usd_goal_real_tier4  -0.188
## 5 usd_goal_real_tier5  -0.289
## 6 usd_goal_real_tier6  -0.416
## 7 usd_goal_real_tier7  -0.479
```

Relative to tier 1, the most likely funded project category tier is Tier 1. The odds of success for tier 2, holding everything constant, is $e^{-0.067468} = 0.9347$ times the odds of success for the Tier 1 projects.

Furthermore, all else being equal, the estimated probability of success for the Tier 1 projects are 0.62, whereas for Tier 2 the probability of success is 0.61. There is sufficient evidence based on our model to suggest that Tier 1 funding may be the most readily successful project Tier, and that the probability of success is over 0.50 for Tiers 1, 2, 3 (p = 0.59), 4 (p = 0.58), 5 (p = 0.55), 6 (p = 0.52), and 7 (p = 0.51).

Project Category and Success

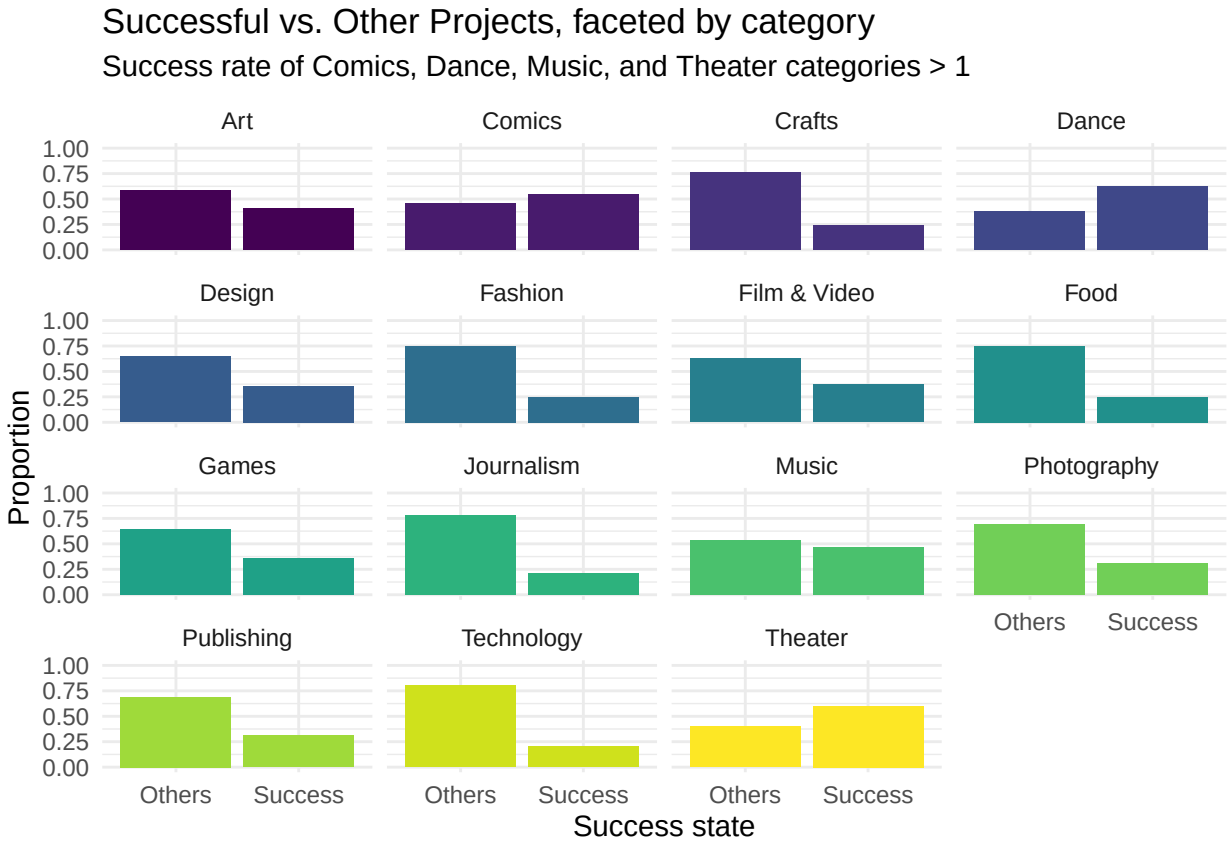
We are studying the question of whether the main category of a project influences its rate of success. We decided to use main category rather than category, because the latter was too broad for our project. *main_category* is composed of 15 distinct categories, each for a unique industry.

The accuracy of or analysis for this question is limited by the fact there could be possible crossovers in the main_categories that we cannot screen for. We assumed this to be a negligible amount of projects, if it existed at all, and thus continued with *main_category* over *category* for analysis.

Question: Does the main category of project influence it's chance of success?

Table 5: Success Rate by Category

main_category	prop
Art	0.412
Comics	0.544
Crafts	0.242
Dance	0.623
Design	0.354
Fashion	0.248
Film & Video	0.373
Food	0.249
Games	0.358
Journalism	0.214
Music	0.469
Photography	0.308
Publishing	0.311
Technology	0.200
Theater	0.601



We created another variable called *plot_success* which enabled us to plot the whether the project was successful or not, excluding the 1's and 0's that our previous indicator used. We plotted bar charts to show the proportion of successful to all the other projects, faceted by category. By using a ratio of successes to others, we were able to visualize the relative success of each category on the same scale.

Interestingly, the visualization shows that more categories failed in raising enough money to meet their goals than those that achieved their monetary goal. Only the categories of Comics, Dance, and Theatre had more

successes than failures according to our visualizations.

Hypotheses:

- H_0: There is no relationship between *main_category* and *success*.
- H_1: There is a relationship between *main_category* and *success*.

We will perform a Chi-squared test at the $\alpha = 0.05$.

```
##
##  Pearson's Chi-squared test
##
## data:  table(kickstarter$main_category, kickstarter$success_state)
## X-squared = 16137, df = 14, p-value < 2.2e-16
```

Analysis of Results:

We used a Chi-squared test because we want to see if the variables main_category and success_rate are independent of one another in this data set. In other words, running a Chi-squared test helps us evaluate our hypothesis; that there is an association between project category and project success rate.

Our test statistic was 16137, which has a Chi-square distribution with 14 degrees of freedom under the null hypothesis. This corresponds to a p-value less than 2.2e-16. Thus, our decision is to reject the null hypothesis. Moreover, there is sufficient evidence to claim that the alternative hypothesis, that there is an an association between *main_category* and *success*, is true.

Logistic Regression Model to Predict Category Success

Here we used a logistic regression model to predict category success. Specifically, we wanted to see how the project's main category leads to differences in the odds of success.

We used Technology as our reference level.

```
## # A tibble: 15 x 2
##    term                      estimate
##    <chr>                        <dbl>
##  1 (Intercept)                -1.39
##  2 main_categoryArt            1.03
##  3 main_categoryComics         1.56
##  4 main_categoryCrafts         0.246
##  5 main_categoryDance          1.89
##  6 main_categoryDesign         0.788
##  7 main_categoryFashion        0.277
##  8 main_categoryFilm & Video   0.870
##  9 main_categoryFood           0.284
## 10 main_categoryGames          0.804
## 11 main_categoryJournalism     0.0875
## 12 main_categoryMusic          1.26
## 13 main_categoryPhotography    0.578
## 14 main_categoryPublishing     0.591
## 15 main_categoryTheater        1.80
```

Relative to Technology, the most likely funded project category is Dance. The odds of success for Dance are $e^{-1.38714883} = 6.374285$ times the odds of success for Technology. Furthermore, all else being equal, the estimated probability of success for the Dance category is 0.62, whereas for Technology the probability of success is 0.21. There is sufficient evidence based on our model to suggest that Dance may be the most readily successful project type and is likely worth spending time looking into this category for project creators.

Discussion

Farzeen: This will require a summary of what you have learned about your research question along with statistical arguments supporting your conclusions.

Olivia: Also, critique your own methods and provide suggestions for improving your analysis. Issues pertaining to the reliability and validity of your data and appropriateness of the statistical analysis should also be discussed here.

The Kickstarter dataset we have is a rigorous collection of data to give project creators insight into which projects have historically succeeded. The dataset is limited, however. We have been able to give creators the statistical odds of success based on monetary and categorical variables, yet have neither looked into how long each project was on Kickstarter for funding, nor what time of the year donations may spike. People may be more charitable during the holiday period, and that may lead to more projects being funded. Not then only would creators have the knowledge of which projects secured funding based off the funding goal and category, but they would also have tools now on when to launch their Kickstarter funding. Furthermore, a dataset on how each project managed their marketing campaign would be more than helpful to creators. Sentiment analysis of the campaign, along with its length, how often it changed, the target audience, and where it was marketed (socials, print, word of mouth, etc) would be the start of this dataset. All told, this analysis we have done is the start of a much larger project. We have shown that there are influencers on a project's funding success, and finding those with the largest influence should be the next priority to complete this project.