Kickstarter Analysis

11/20/2020

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

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 launching in 2009, 19 million people have pledged to back up 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:

- Does the amount of money a creator asks for influence it's chance of success?
- Does the category of project influence it's chance of success?

Our hypotheses include:

- The more money the project asks for, the less successful it will be in terms of getting funding.
- The project category is associated with the success rate.

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. Modelling 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 cancelled), 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.

Explorative Data Analysis

Overview of project state

Distribution of project states
More projects failed than those that succeeded
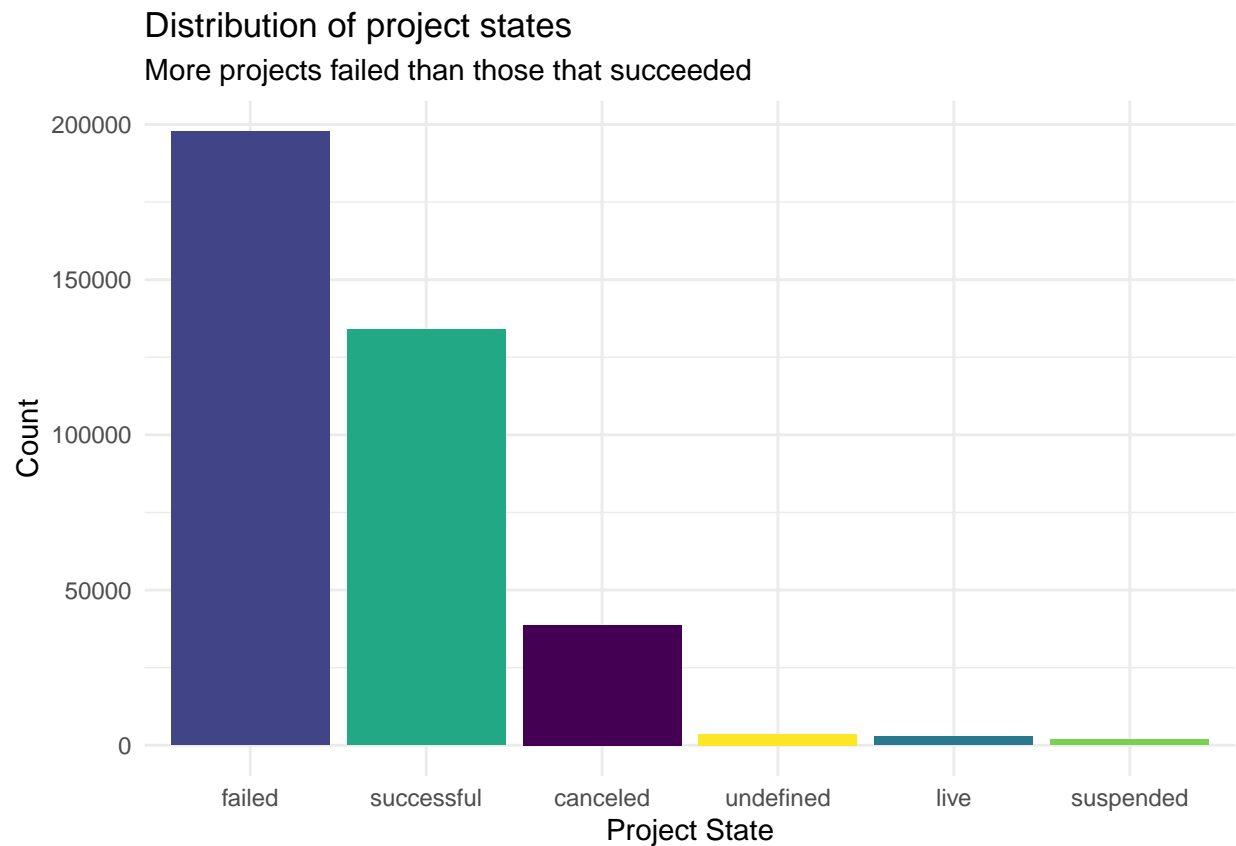

```
## # A tibble: 6 x 2
## # Groups:   state [6]
##   state         n
##   <chr>     <int>
## 1 failed    197719
## 2 successful 133956
## 3 canceled   38779
## 4 undefined   3562
## 5 live        2799
## 6 suspended   1846

## # A tibble: 1 x 1
##   'project success rate'
##                   <dbl>
## 1                 0.354
```
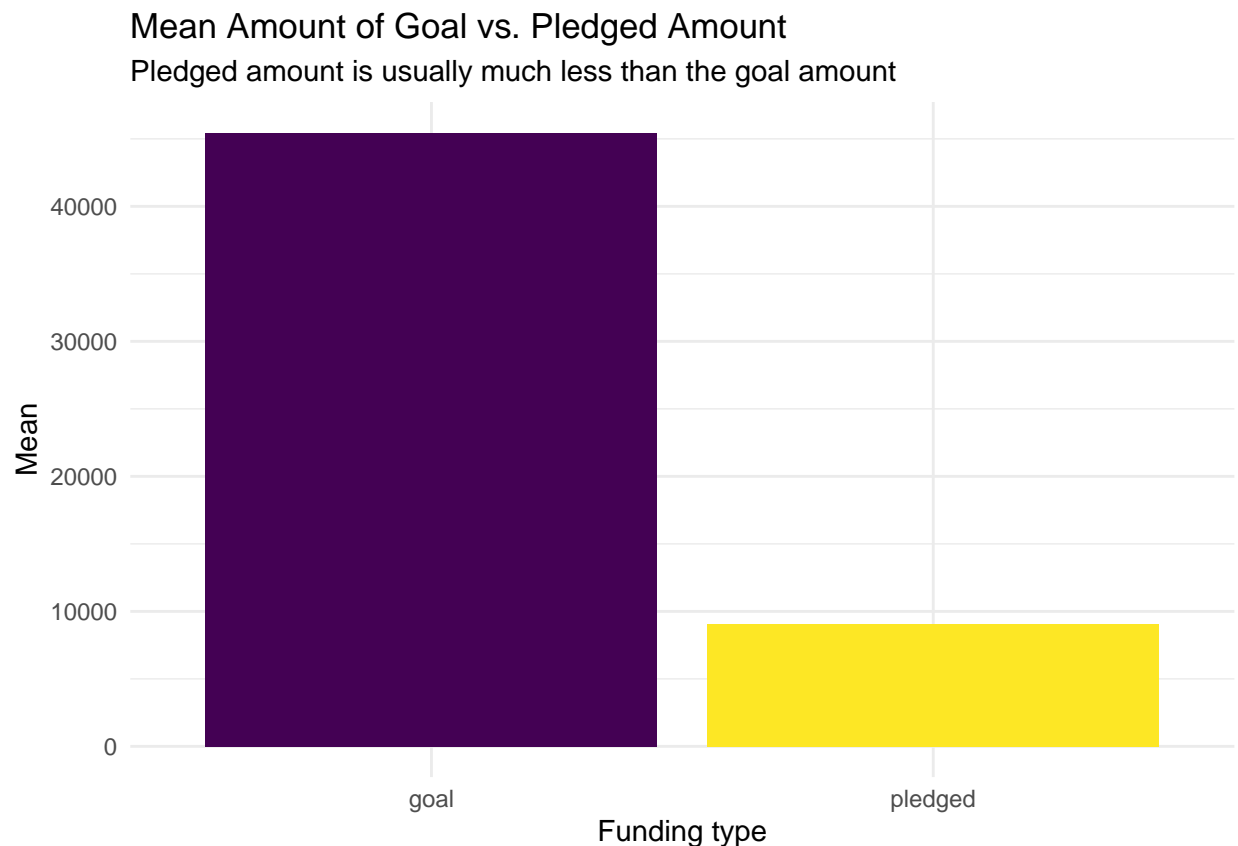
More projects failed than those that succeeded, with an average success rate of 35.4%.

Overview of pledged and goal amount in USD:

```
## # A tibble: 1 x 4
```

```
##    mean median     sd type
##   <dbl>  <dbl>  <dbl> <chr>
## 1 9059.   624. 90973. pledged
```

```
## # A tibble: 1 x 4
##    mean median       sd type
##   <dbl>  <dbl>    <dbl> <chr>
## 1 45454.   5500 1152950. goal
```

Mean Amount of Goal vs. Pledged Amount
Pledged amount is usually much less than the goal amount

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.

In this analysis, we conducted two central limit thereom (CLT)-based tests. Due to the sheer size of this dataset, using a simulation based method for analysis is not appropriate. We removed all projects that were "live" and currently asking for funding, so that any discrepencies 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. To analyze success, rather than using the given variable "success", we created our own. The original variable "success" contained cancelled, suspended, and undefined states along with successful and failed states. We had no indication if those projects that were cancelled, suspended, or undefined met their goals and cancelled prematurely or if they cancelled due to no funding at all. To get around this, we created a variable for success that was a ratio of the project's funding raised over their original funding goal. If the funding raised was greater than the funding asked for, the project was successful; if the project goal was greater than the raised funds, the Kickstarter funding was unsuccessful.

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 CLT-based test, placing the lower and higher groups in the same category when running a t-test. It would make intuitive sense if the projects that require less funding 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.

Methodology

In this analysis, we conducted two central limit thereom (CLT)-based tests. Due to the sheer size of this dataset, using a simulation based method for analysis is not appropriate. We removed all projects that were "Live" and currently asking for funding, so that any discrepencies 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. Thereforein the first section of the analysis, we overwrote the Kickstarter dataframe 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 cancelled, suspended, and undefined states along with successful and failed states. We had no indication if those projects that were cancelled, suspended, or undefined met their goals and cancelled prematurely or if they cancelled 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, cancelled, suspended, or undefined).

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 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 linear 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 intial funding goal and modelling 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 neglible amount of projects, if it existed at all, and thus continued with "main_category" over "category" for analysis.

Following these analyses, we modelled 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. Technology was used as the reference level for our model, so 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

As explained in the Methodology section, we first overwrited the initial kickstarter dataframe with projects that are not "Live".
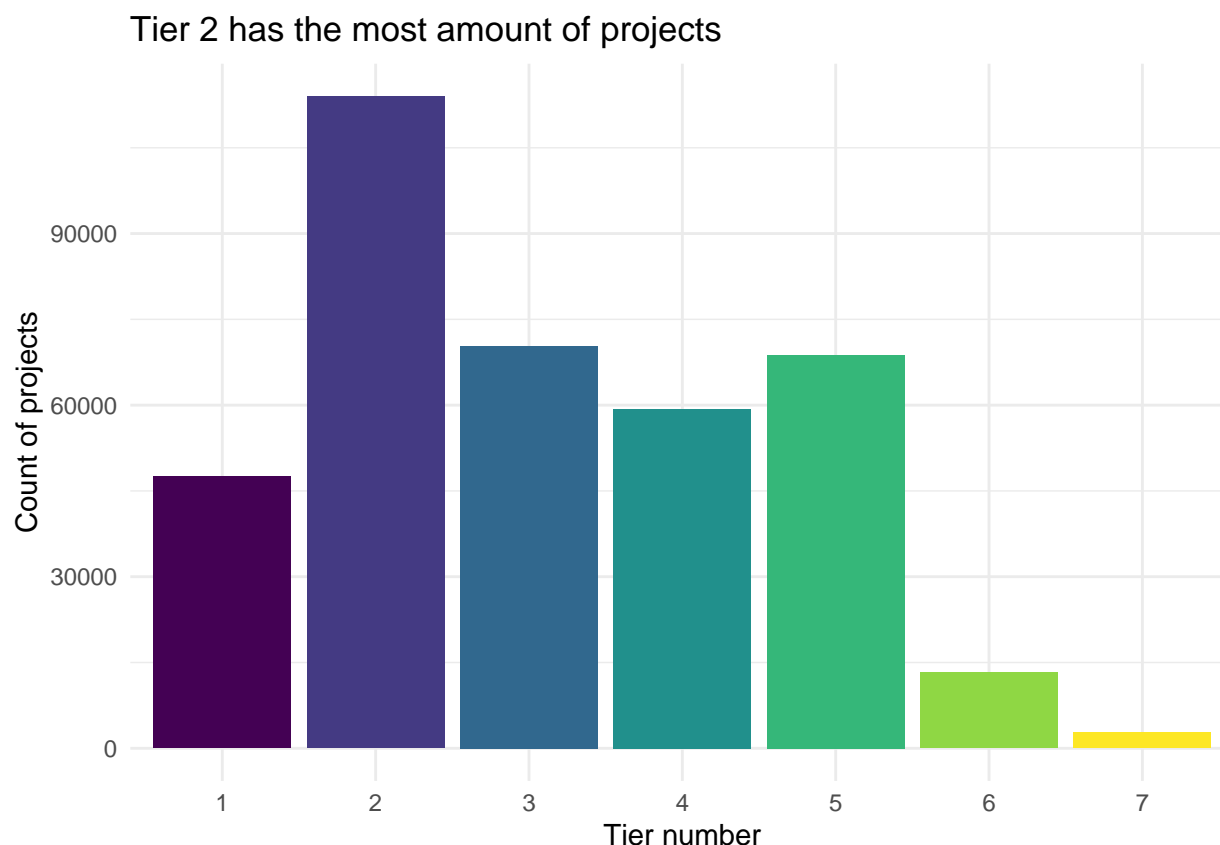
Project Goal Amount and Success

In the Explorative 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 tiering 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 $7 \geq 500,000$.

```
## # A tibble: 7 x 2
##   usd_goal_real_tier numbers
##                <dbl>   <int>
## 1                  1   47494
## 2                  2  113993
## 3                  3   70338
## 4                  4   59285
## 5                  5   68727
## 6                  6   13231
## 7                  7    2794
```

Tier 2 has the most amount of projects

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 alpha 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?

```
## # A tibble: 30 x 5
## # Groups:   main_category [15]
##    main_category success_state     n  prop plot_success
##    <chr>                 <dbl> <int> <dbl> <chr>
## 1 Art                       0 16449 0.588 Others
## 2 Art                       1 11510 0.412 Success
## 3 Comics                    0  4901 0.456 Others
## 4 Comics                    1  5842 0.544 Success
## 5 Crafts                    0  6618 0.758 Others
```
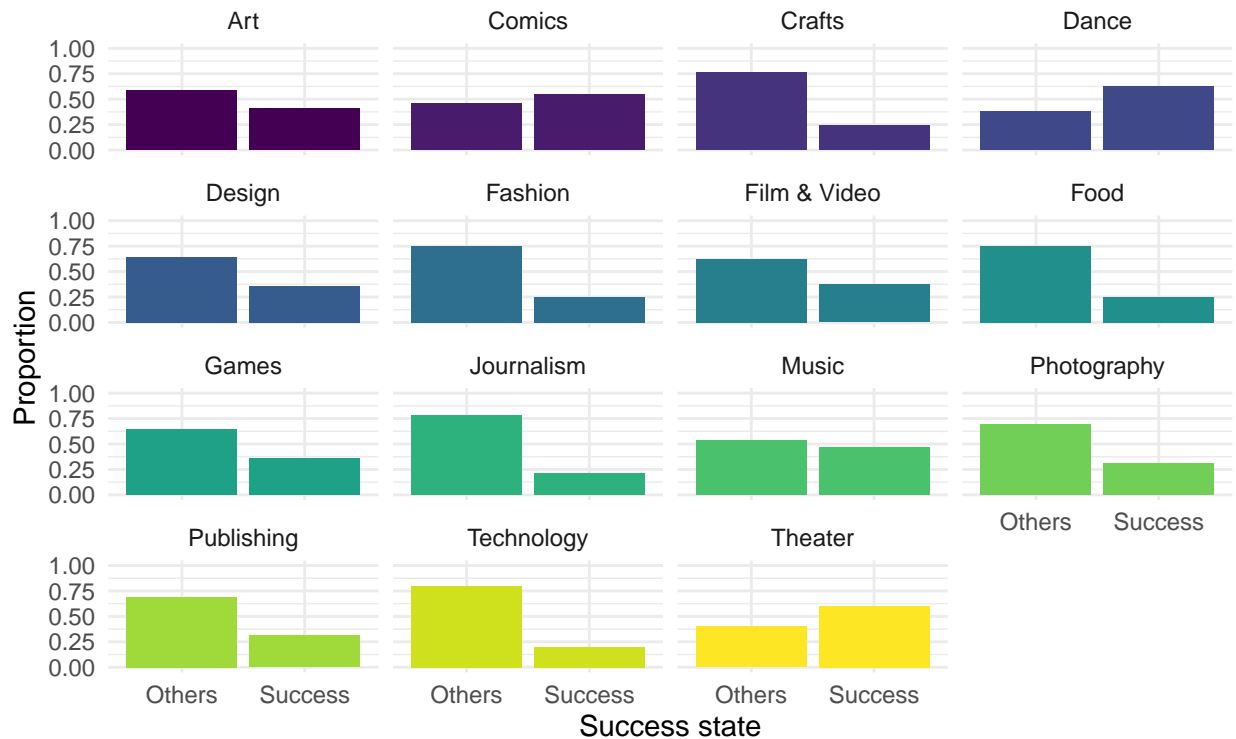
```
##  6 Crafts                    1  2115 0.242 Success
##  7 Dance                     0  1412 0.377 Others
##  8 Dance                     1  2338 0.623 Success
##  9 Design                    0 19215 0.646 Others
## 10 Design                    1 10550 0.354 Success
## # ... with 20 more rows
```

Successful vs. Other Projects, faceted by category
Success ratio of Comics, Dance, Music, and Theater categories > 1

We created another variable called "plot_success" which enabled us to plot the results while excluding the 1's and 0's. We plotted bar charts to show the proportion of successful to unsuccessful ("other") projects, faceted by category. By using a ratio of successes to others, we were able to visualize the relative success of each category.

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 successess 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 CLT simulation 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:

The Chi-squared test compares observed vs. the expected counts that we would expect if the null hypothesis were true.

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.

From our Chi-squared test, we calculate the p-value to be < 2.2e-16.

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.

Emile: A paragraph on what you would do differently if you were able to start over with the project or what you would do next if you were going to continue work on the project should also be included.