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.

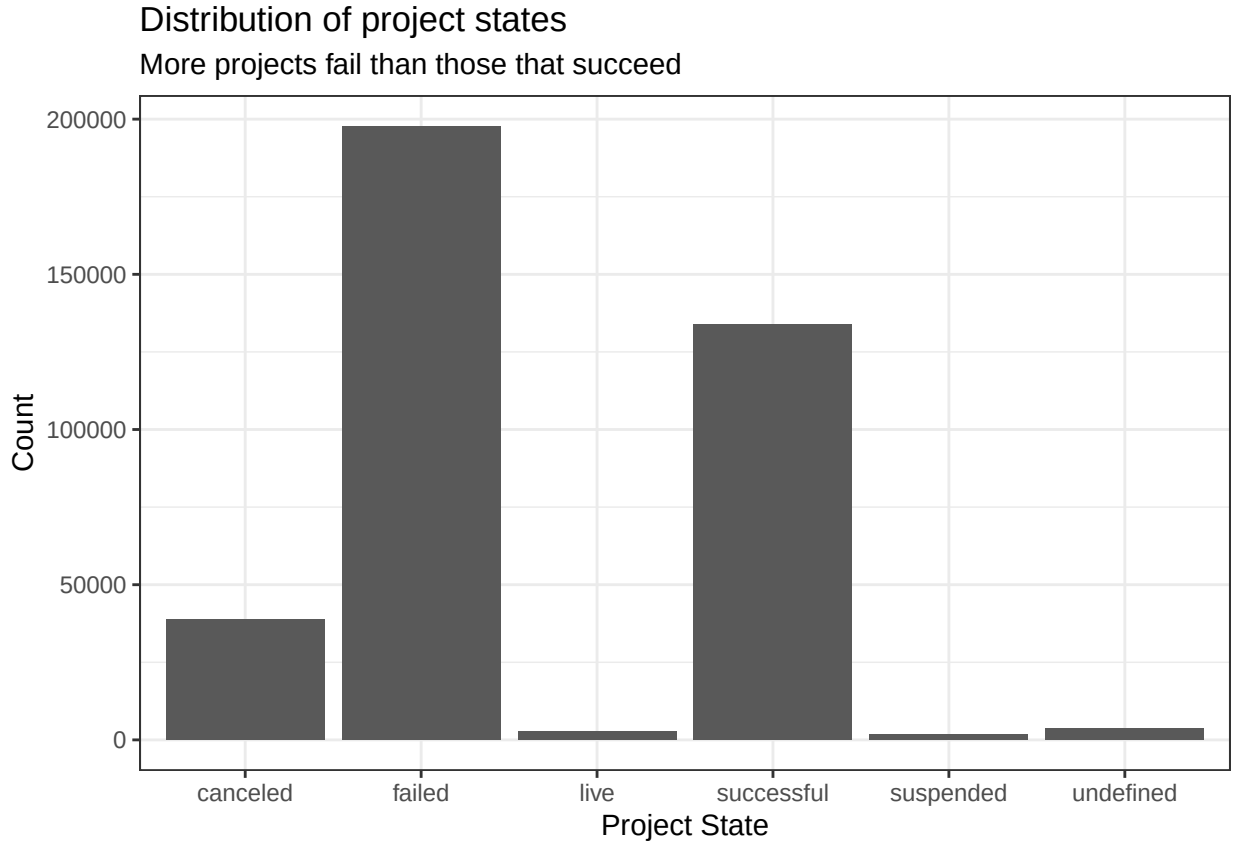
The numerical data are the project's ID number, the monetary goal for each, the money pledged to each project, how many backers of each project; there are three numerical variables that are not self-explanatory, usd_pledged: conversion in US dollars of the pledged column (conversion done by kickstarter),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




```
## # 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
```

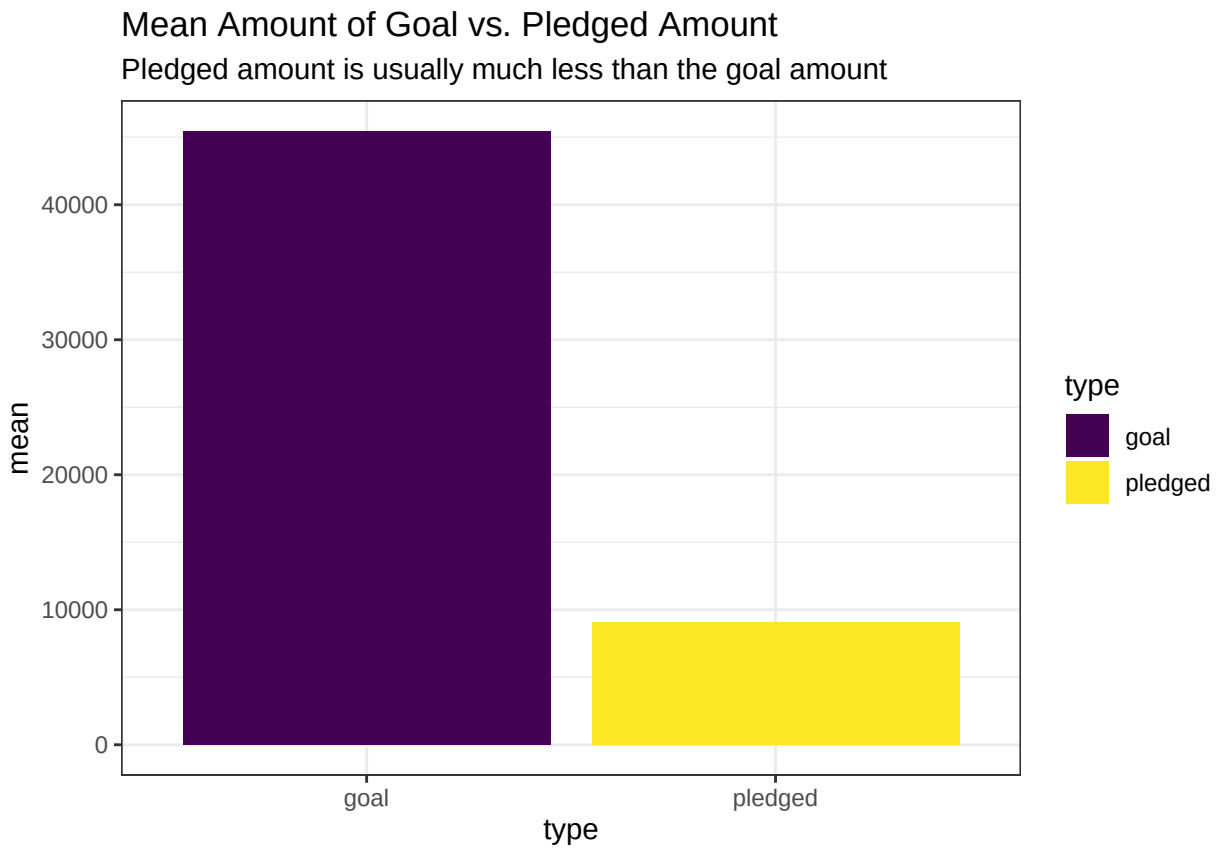
```
## # A tibble: 1 x 1
##   `project success rate`
##                    <dbl>
## 1                  0.354
```

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
```



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. 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.

After establishing a relationship between project success and the intial funding goal, we used a CLT-based 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 analyss 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.

Data Analysis

Project Goal Amount and Success

Claim: The more money the project asks for (project goal), the less successful it will be in terms of getting funding.

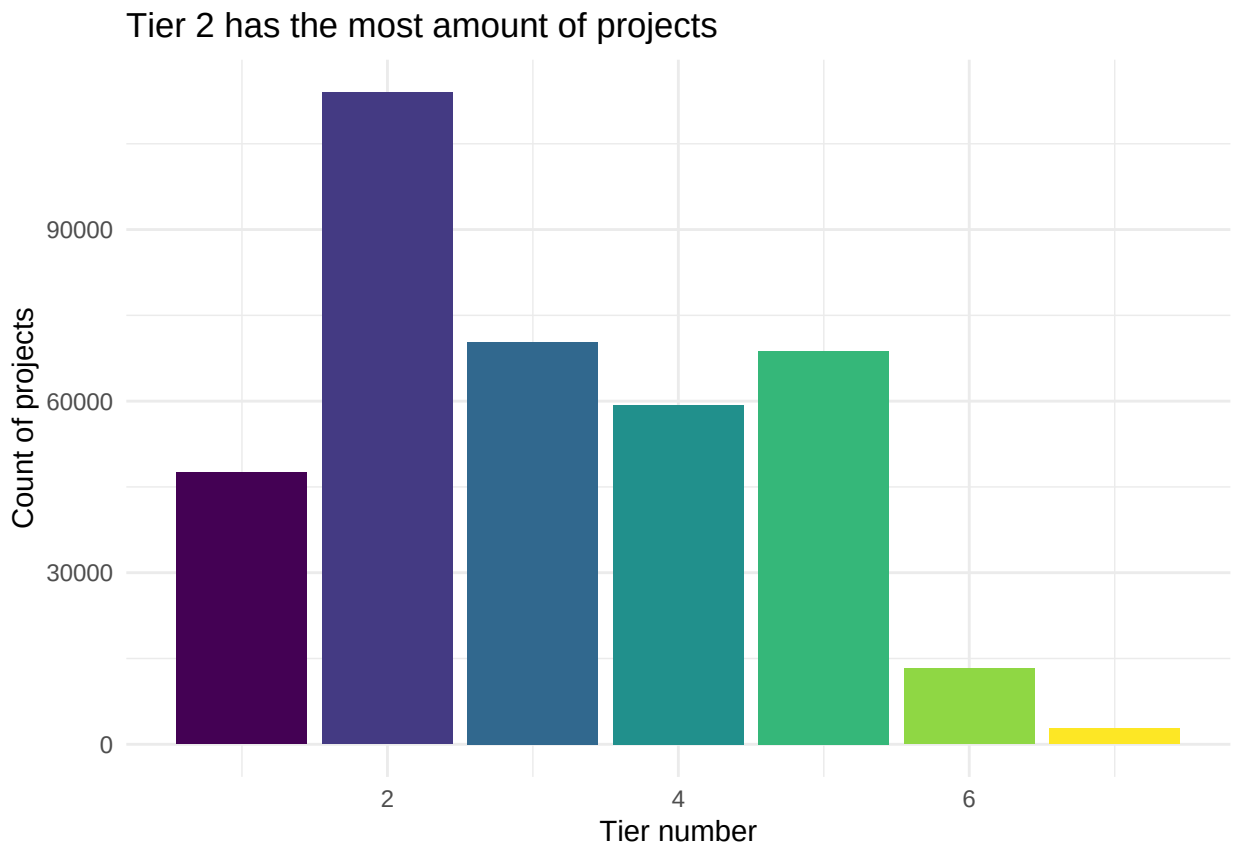
TODO:

- add some context here why we examine this relationship
- you could explain/justify the specific variables you chose to include such as "used_goal_real_tier"– what does this variable mean and how does the classification into tiers with your mutated variable relate to the larger analysis on goal v. success rate? This will make it easier for the reader to follow along and allow you to have methods with enough detail such that they can be reproducible. Explain how tiers were defined. rationale.

Create a new data frame with projects that are no longer live:

Create a new variable named usd_goal_real_tier to classify usd_goal_real into tiers:

```
## # 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
```



Create a new variable for success state

Hypotheses:

H_0: $p(less - money) \leq p(more - money)$ H_1: $p(less - money) > p(more - money)$

TODO:

- "proportion of people who successful with low money goal." was initially difficult to follow/understand exactly what you are trying to test.
- in the analysis paragraphs following the tests, including the actual p-value found and the alpha level it was tested against is information that is necessary. This was present in some tests, but was missing in the project goal amount and success section. As an example of how this specifically could be improved, in the project goal amount and success section, you may wish to support your claim of "our p-value is less than our alpha value" with the actual p-value you calculate and the alpha value you are testing in. Additionally, when there is sufficient evidence to reject the null hypothesis, make a statement about how this means there is sufficient evidence to claim that the alternative is true. This sentence will help bridge the two conclusions made. In the analysis for the project category and success section, it may be clearer to use a word other than "main category" in the hypothesis test conclusion (i.e. say project of category or something more specific in the context of the research than simply what the variable is named in the dataset).
- Each bootstrap analysis is only performed with the number of simulations being set to 500. However, the more simulations there are, the more accurate the prediction of the relationship in the data will be. Therefore, any simulation-based methods used to evaluate hypotheses of interest should be set to at least 1000 simulations, instead.
- explain why you chose certain variables and how you then incorporated those into your hypo test

- express the hypotheses in notation and then explain the notation's meaning.

H_0:proportion of people who successful with low money goal - proportion of people who successful with high money <= 0 H_1: proportion of people who successful with low money goal - proportion of people who successful with high money > 0

$\alpha = 0.05$

```
##
##  One Sample t-test
##
## data:  kick1$usd_goal_real_tier
## t = 3111.8, df = 84751, p-value < 2.2e-16
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  5.219288      Inf
## sample estimates:
## mean of x
##  5.222048
```

TODO:

- State the p-value and alpha value explicitly here.
- Currently the statement is repeatitive because you said twice "we reject our null hypothesis that".

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.

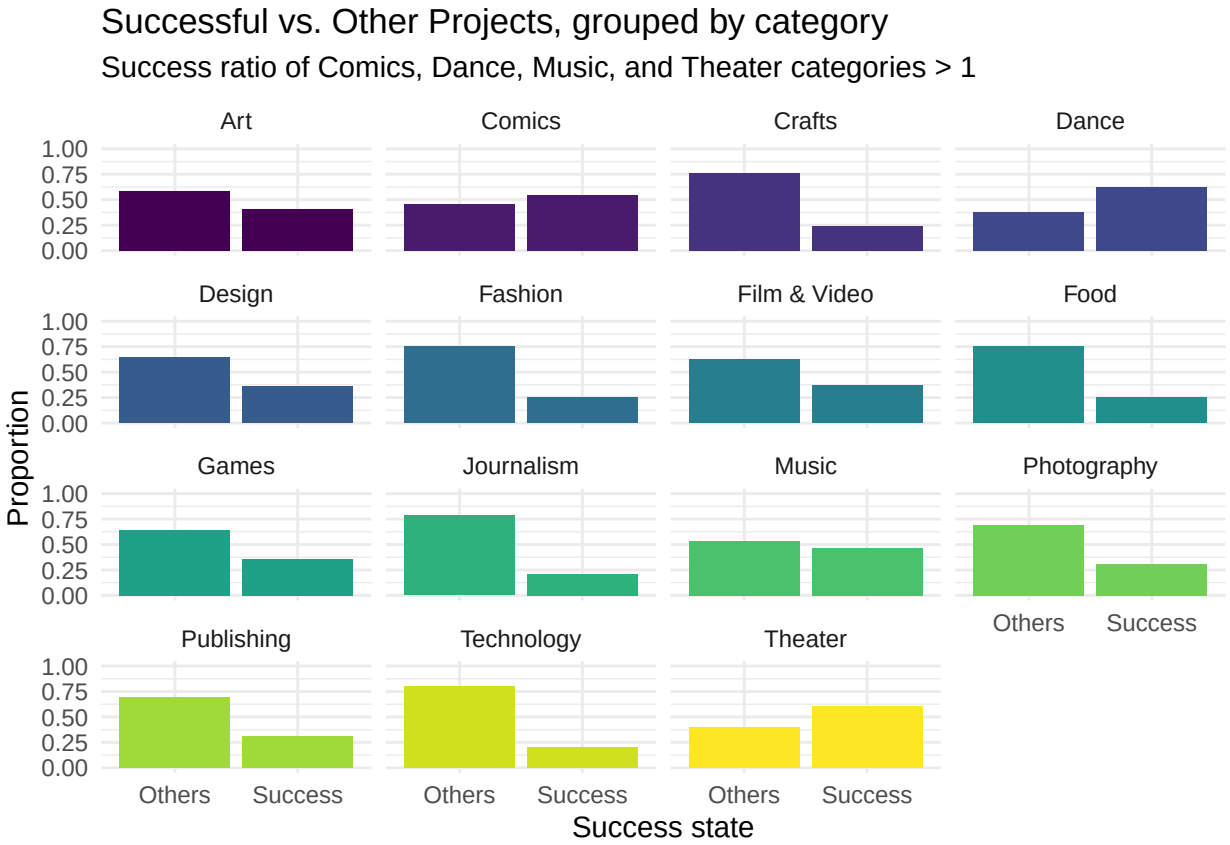
Project Category and Success

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

TODO

- limitations: possible crossover in the main_categories that we cannot screen for

```
## # 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
```



main_category vs. success

H_0: There is no relationship between main_category and success.

H_1: There is a relationship between main_category and success.

Where the $alpha/ = 0.05$

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

TODO:

- state the p-value here
- describe and justify why you ran a chi-squared test and what information this test will provide you in context of supporting your larger question/hypotheses. There could also be greater explanation of the individual steps you took before completing the overall chi-squared test
- when there is sufficient evidence to reject the null hypothesis, make a statement about how this means there is sufficient evidence to claim that the alternative is true. This sentence will help bridge the two conclusions made. In the analysis for the project category and success section, it may be clearer to use a word other than "main category" in the hypothesis test conclusion (i.e. say project of category or something more specific in the context of the research than simply what the variable is named in the dataset).

- Interpretation of chi-square test could follow a format like this: "Our test statistic was 19624, which has a chi-square distribution with 6 degree of freedom under H_0. This corresponds to a p-value less than 2.2e-16. So our decision is to reject H_0, and there is sufficient evidence that there is a relationship between a project's monetary goal and a project's state of success."

Given that the p-value is less than $alpha/ = 0.05$, we can reject the null hypothesis. Thus, there is sufficient evidence to suggest that there is a relationship between main category and success.

Project Goal Amount and Pledge Amount

Here we used a linear model to model success based on main category. 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 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

TODO: Furthermore, the entire discussion section is missing, as there is no overall summary of what all of the hypothesis tests have shown in the context of the research question, along with the specific p-values that support these conclusions. To make your results and conclusions stronger, you should also critique your own methods and provide suggestions for improving your analysis, as showing possible faults in reliability and validity of your data and the appropriateness of the statistical analyses helps support your positions as researchers and knowledge of the data. To add onto conclusions that you write, you should also discuss what you would do next if you were going to continue work on the project to show where your analyses could have gone farther.