Lab 2: What Makes a Successful Kickstarter Campaign?

Bronte Baer, Danny Collins, Amy Ho

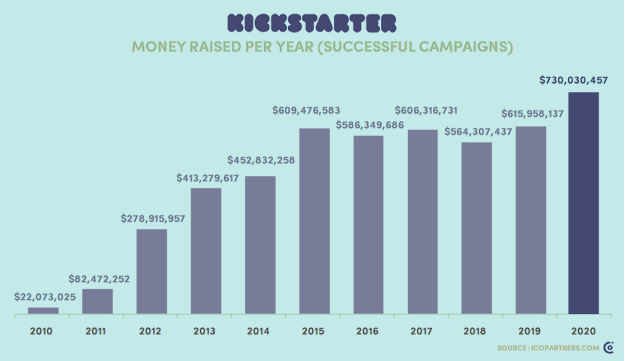
March 14, 2022

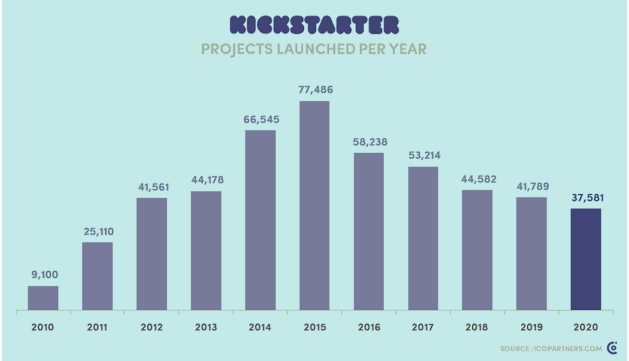
Contents

1 Introduction

1.1 Motivation

Crowdfunding has become an increasingly popular option for entrepreneurs, and while there are many platforms to choose from, Kickstarter is the biggest. The company discloses a wide range of statistics, and its sizeable organic community and press coverage outshine all other platforms.[1] The platform prides itself on allowing creators to "make ideas into reality."[2] The platform focuses on projects that create something tangible, but there are otherwise minimal rules and guidelines for people to follow. Anyone can create a campaign and decide the details for the campaign. While all this freedom is part of the appeal for entrepreneurs to use Kickstarter to fund their ideas, there is a lot of ambiguity regarding what makes a successful campaign. In early 2021, multiple articles were published stating that 2020 holds the record for the most money raised on Kickstarter; however, "the number of funded projects [wa]s actually lower than previous years."[3] Furthermore, according to Kickstarter statistics, less than half of launched campaigns reach their goal.[4] With all this in mind, our research aims to inform someone who is going to start a Kickstarter campaign about the factors that will most likely help their campaign succeed.





1.2 Research Question

Setting up a successful Kickstarter campaign is less straightforward than one might expect. Numerous factors impact a campaign's success, and our research aims to address a few key measurable and tracked options that a campaigner can keep in mind when creating their project.

The main research question our analysis seeks to answer is:

> *Which project details increase the chance of a Kickstarter campaign's success?*
> *More specifically, what magnitude of initial funding goal should a campaigner set?*
> *How does the campaign category change which factors matter for success?* This study explores the relationship between a campaign reaching its funding goal and the funding goal set. Additionally, we examine supplementary variables to further our analysis and help us understand the funding goal amount set. As creators continue to turn to Kickstarter to fund their projects, we hope to provide descriptive statistics to supplement the observational ones Kickstarter supplies. Kickstarter reports the *what*, but we aim to supply insight into the *how*.

[1] V., Narek. "Indiegogo vs Kickstarter: Which One to Choose? (2022 Update)." *The Crowdfunding Formula*, 1 Mar. 2021, https://blog.thecrowdfundingformula.com/indiegogo-vs-kickstarter/.

[2] "About." *Kickstarter*, https://www.kickstarter.com/about.

[3] Bidaux, Thomas. "Kickstarter in 2020 for Games." *Medium*, ICO, 25 Jan. 2021, https://medium.com/icopartners/kickstarter-in-2020-for-games-70d26b5cba73.

[4] "Stats." *Kickstarter*, https://www.kickstarter.com/help/stats.

2 Data and Methodology

2.1 About the Data

Our analysis uses data scraped from Kickstarter by Web Robots, a company that specializes in providing datasets via web crawling. The company provides a new Kickstarter dataset monthly, with datasets dating back to 2014. Each dataset contains one row per campaign ID and category, with some campaigns marked as multiple categories.

The combined dataset contains 39 initial variables (see appendix for full data dictionary). Our particular variables of interest, whether in transformed or raw form, included:

- the `category` that the campaign was listed under
- the `launch` timestamp when the campaign began allowing backers to pledge money
- the `deadline` timestamp by when the campaign needed to hit its goal
- the `goal` amount that the campaign needed to hit to succeed
- the total amount `pledged` to the campaign towards the goal
- whether the campaign was a `staff_pick` or not while it was active
- the `status` of the campaign, i.e. active, successful, canceled, or failed, at the time of data pull

2.1.1 Data Cleaning

Web Robots notes that, as of April 2015,

> "Kickstarter started limiting how many projects user[s] can view in a single category. This limits the amount of historic projects we can get in a single scrape run. But recent and active projects are always included."[5] Given this limitation, we combined and deduplicated four datasets pulled from September 16, 2020, March 18, 2021, September 16, 2021, and March 24, 2022, to ensure we had enough data to answer the questions we are interested in answering. This combined dataset gave us 497,264 rows of data (pre-deduping).

Our ingestion script starts by reading the last CSV in the most recent dataset (3/2022) and then iterates backward through each CSV in the dataset to deduplicate projects in multiple categories and/or in multiple datasets. On each iteration, any duplicate project rows in the CSV are discarded. This deduplicated set is then compared to all projects already captured in the accumulated dataset being built (which on the first iteration is an empty data frame). Any projects not yet seen in the collected dataset are then appended to the accumulated dataset. Once all CSVs in the most recent dataset are exhausted, the iteration continues to the next most recent dataset, and the process repeats. This allows us to capture the most recent "snapshot" of a project before it becomes stale, falls out of pagination, and is no longer accessible by the web scraper. When a project is listed in multiple categories, and both first appear in the same chronological dataset, the row first scanned is chosen, making the category assigned to the project somewhat random. This, and other limitations, are discussed more extensively in our limitations section.

We then built our target dataset from constraints that best facilitate our research question from our raw collection of unique projects. We recognize that the landscape of world commerce was transformed dramatically by the pandemic and that patterns seen before March 2020 may drastically differ from ones seen since. Since our research question revolves around the success of projects moving forward, we limited our dataset to projects launched on April 1, 2020, or later. We further limited our dataset to United States-based projects funded in USD, to avoid complexities from different countries' national economies being in different states since the pandemic and to avoid conversion complexities for currencies.

We then reduced our dataset to projects that had reached finality in either a "successful" or "failed" state. Projects that are still actively raising pledges were discarded because we do not know whether they will be

[5] "Kickstarter Datasets." *Web Robots*, 29 Mar. 2022, https://webrobots.io/kickstarter-datasets/.

successful, and similarly, with canceled projects, we do not know the reason or status as of when they were canceled.

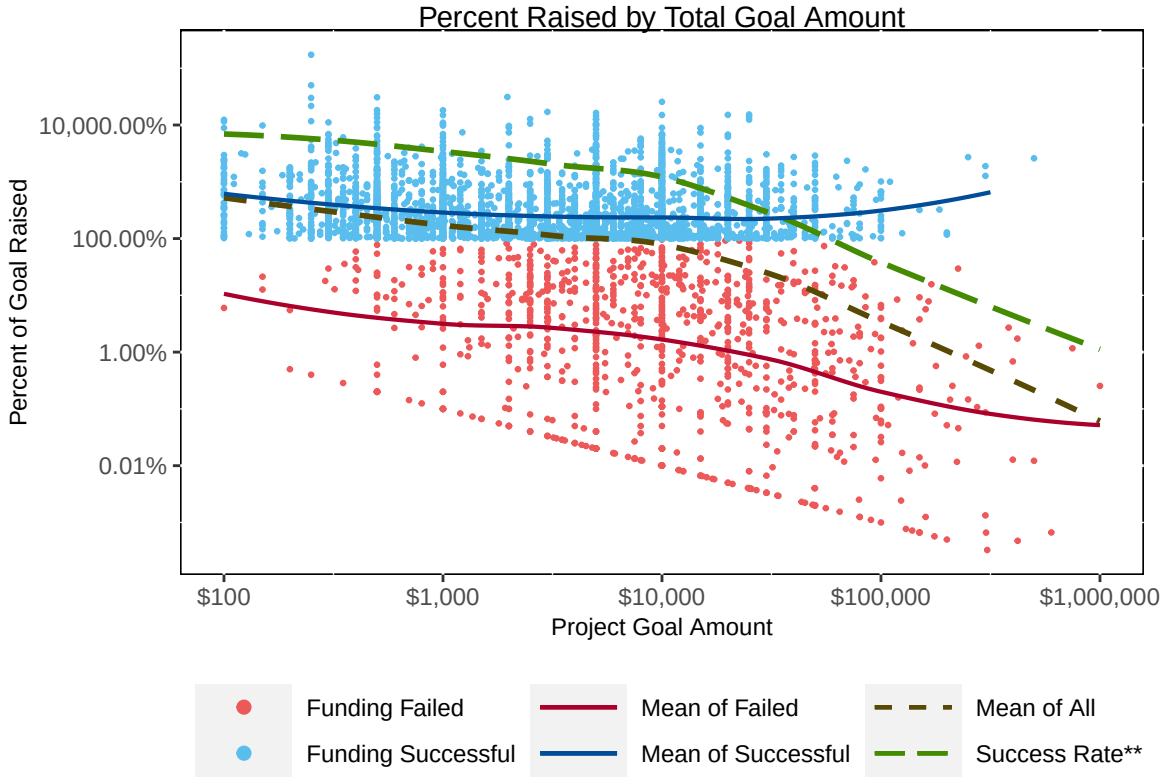
We also removed what we deemed to be extreme outliers that would complicate modeling the dataset at large and are rare enough that there is little value in modeling them. We removed projects that raised $0 or had a goal of less than $100 or greater than $1,000,000. These outliers made up about 1% of the dataset.

The outlier reduction resulted in 15993 projects, and of these, we dedicated 30% (4797 projects) for early data analysis and discovery. We reserved the remaining 70% (11196 projects) for analyzing the performance of our designed estimators.

2.2 Defining Success

Kickstarter allows creators to set their own funding goals and follows an all-or-nothing model. If the total amount of money pledged is at least equal to the funding goal by the deadline, then the campaign is marked as a success, and the funds raised are disbursed. If the goal is not met, then regardless of how much the campaign was short, it is considered a failed campaign, and funds are not taken from backers.

We considered there might be value in using the rate-of-goal pledged (i.e., how close the campaign got to reaching 100% or exceeding 100%) instead of discarding this fidelity to look at whether 100% of the goal was met or exceeded as a binary variable. However, on the contrary, we observed that successful and failed campaigns follow distinct trendlines that become increasingly divergent as campaign goals increase.



We found successful campaigns succeed by relatively consistent margins regardless of the goal, even ticking upward for huge goal campaigns (blue line). However, unsuccessful campaigns failed with exponential spectacularity as the goal amount increased. Complications from heteroskedasticity aside, this presents an unavoidable downward bias as the average rate of large goals dips below the theoretical minimum possible for lower goals (red line).

Conversely, it becomes increasingly improbable that large-goal campaigns would exceed their goals by such extreme orders of magnitude. This asymmetry would coerce a predictor (e.g. brown short-dashed line) to

travel well below the 100% threshold, even if more campaigns succeeded than failed at a given goal amount (green long-dashed line). This would result in underpredicting the chance of success for large-goal campaigns.

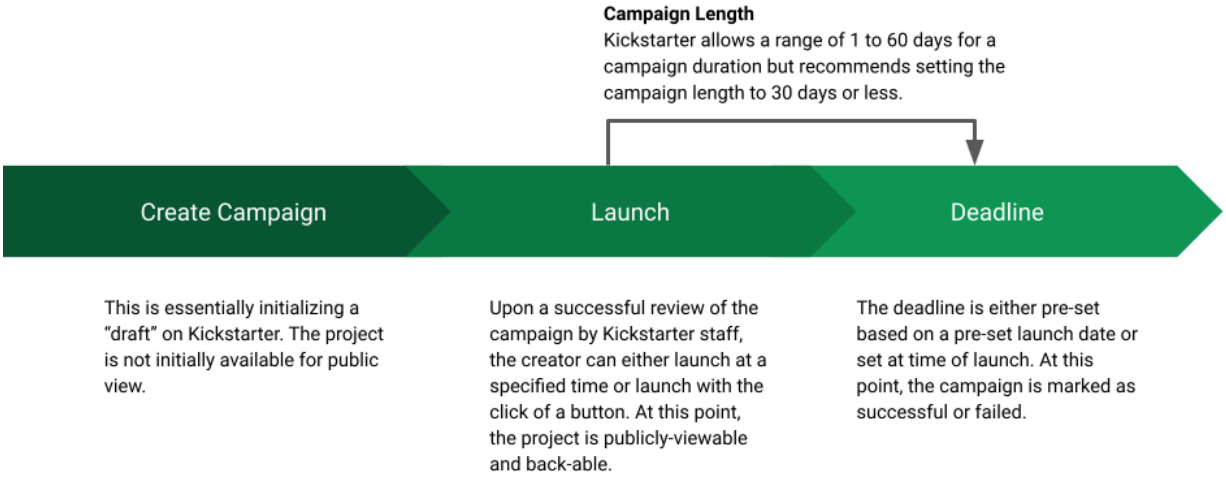
**The success rate line does not map directly to the y-axis values and is illustrative to demonstrate the skew mentioned above, with the blue-red boundary serving as a 50/50 success rate.

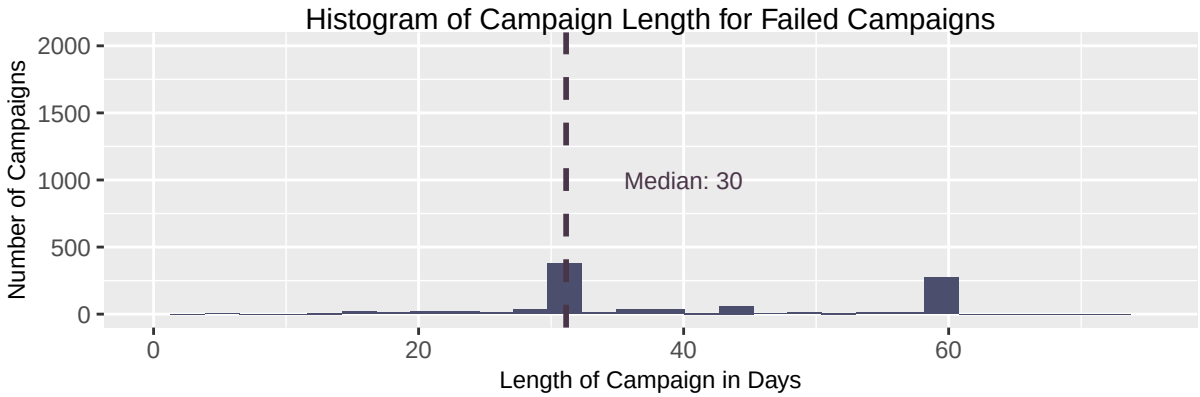
Due to this discovery, we decided to use the campaign `status` in our dataset as our outcome variable, categorizing the success of a campaign as a binary success or failure rather than a metric variable measuring how much success a campaign saw.

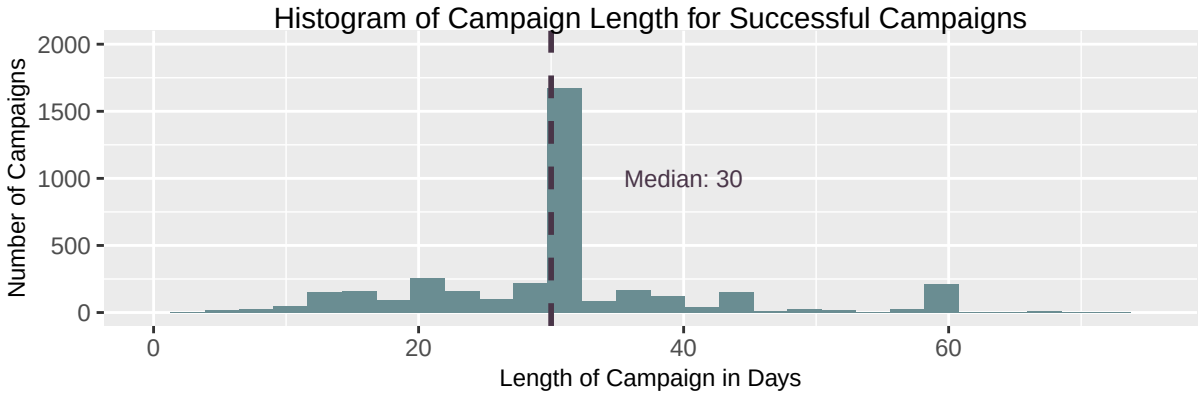
2.3 Campaign Characteristics

For our research design, we chose to analyze variables that the campaign creator had complete control of, such as campaign length and details about the campaign deadline, and variables that the creator had less control over, such as staff pick status and campaign category.

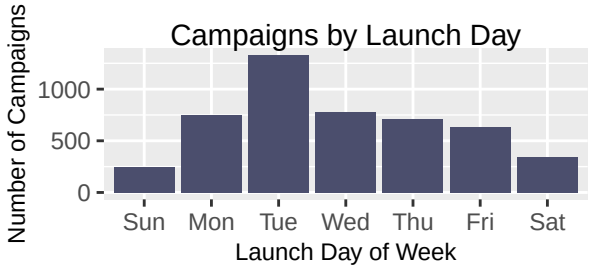
There are three key dates for a Kickstarter campaign: a `created_at` date, `launched_at` date, and `deadline` date.

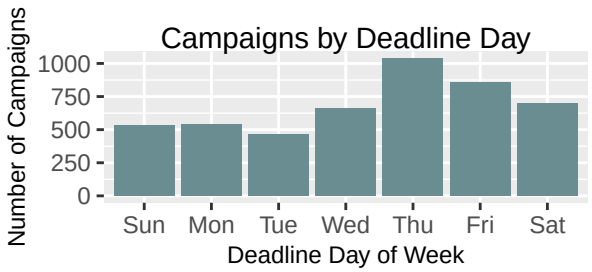


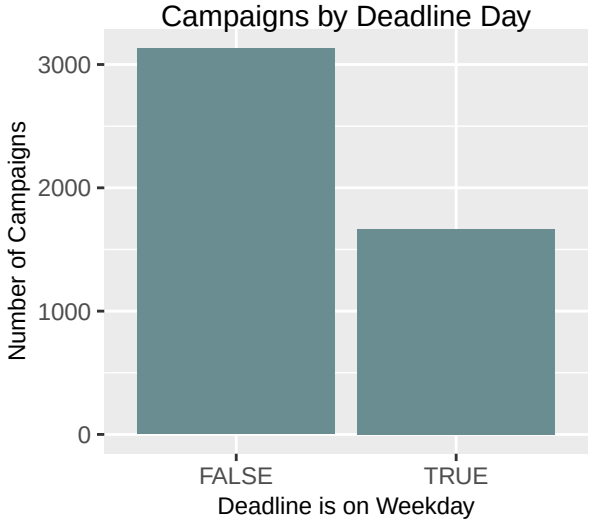


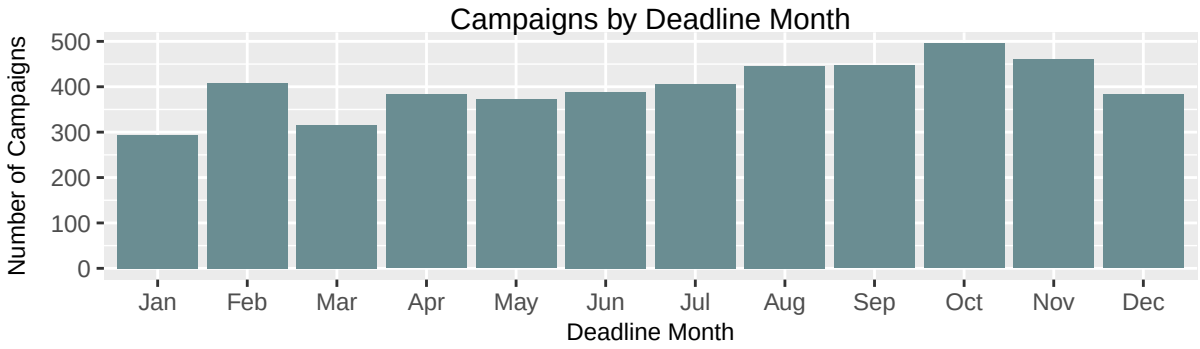


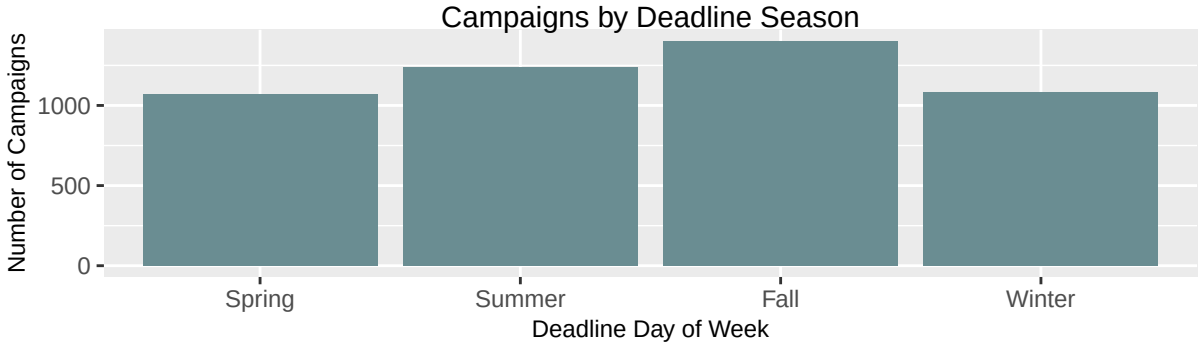
Campaign deadline details:
We believe there is a seasonal logic to when potential backers are considering crowdfunding. For example, people may be less likely to check crowdfunding sites on busier days like weekends or holidays, but may be more curious about crowdfunding during months like October, when people are starting to consider holiday gifts.











While we could have applied the same consideration to the campaign launch date, we ultimately focused on the campaign deadline date because we wanted to limit the number of features, considering our sample size. We chose `deadline` over `launched_at` details because Kickstarter sends out reminder emails to users for bookmarked projects 48 hours and 8 hours before the deadline.[6]

Staff pick

Kickstarter staff can curate projects through `staff_pick` status.[7] We consider the `staff_pick` variable noteworthy because it improves the campaign's visibility in a sea of thousands of active campaigns. Also, it likely reflects some qualitative aspects of the campaign, such as the worthiness of the cause, coolness factor, or creator trustworthiness, that we cannot capture in quantitative variables.

Campaign category

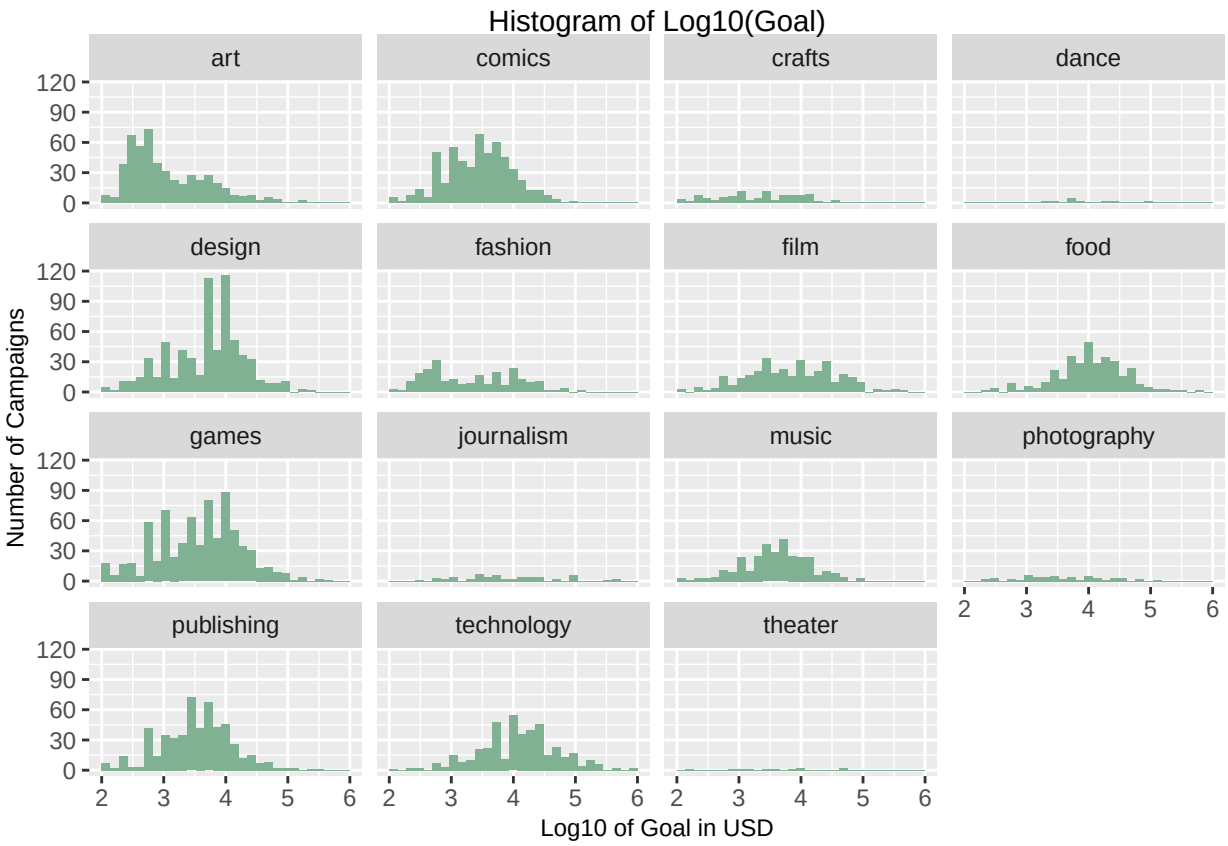
While the category is technically something the creator has control over, in the sense that the creator decides what category their campaign falls under, it is usually not very flexible. For instance, a food truck does not make sense for the "technology" category. We consider this variable to be noteworthy because funding goals can vary greatly by category.

[6] "What Does the 'Remind Me' Button Do?" *Kickstarter*, https://help.kickstarter.com/hc/en-us/articles/115005126574-What-does-the-Remind-Me-button-do-.

[7] "Introducing Badges for Projects We Love." *Kickstarter*, 2 Feb. 2016, https://www.kickstarter.com/blog/introducing-projects-we-love-badges.

Table 1: Count of Campaigns by Primary Category

Category	Campaigns
games	747
design	667
comics	543
publishing	533
art	497
technology	418
food	323
film	319
music	302
fashion	229
crafts	90
journalism	56
photography	52
theater	11
dance	10



2.4 Research Design

Our research aims to understand the factors that increase a Kickstarter campaign's success, which we defined as the campaigner reaching their funding goal. We focus on the funding goal amount as our main dependent variable of interest, with the hypothesis that it is more difficult to crowdfund at higher orders of magnitude of funding goal.

We use large-sample linear models for our analysis, and as with all models, this option has limitations, which we detail in the Limitations section below. Our models include one outcome variable, which we defined as binary.

To enhance our research, we included variables that help explain project visibility and potential backer interest, such as when the campaign is running, how long it runs for, whether it is a staff pick, and what type of project it is. We chose to test these additional dependent variables because we believe that the funding goal amount chosen by the campaign creator is not comprehensive enough to explain campaign success.

Before testing variable correlation and collinearity, we hypothesized the effect variables might have on the model.

- Since Kickstarter's analysis of previously successful projects suggests that campaigns longer than 30 days struggle, we hypothesized that campaigns that run longer than 30 days would reach their funding goals less frequently than the shorter campaigns.
- We hypothesized that campaigns that ended on weekdays would reach their funding goals more often than campaigns that ended on weekends. Additionally, we guessed campaigns in the fall months, September, October, and November, would reach their funding goals more often than campaigns in the other seasons.
- We expect the 'staff_pick' variable to greatly increase the likelihood of a campaign reaching its funding goal because staff picks have more visibility on the platform, and there are likely success factors that are unavailable to us that play a role in staff picks being chosen.

It is important to note that since our outcome variable is binary, with only two levels, our residuals will not be normal if we have a large omitted variable bias, which we discuss in further detail in the Limitations section. We accept this limitation because we do not need normality for a large-sample linear model.

3 Model Building Process

We use the funding goal amount in our analysis to help creators understand whether their minimum viable product is the right level of ambition to be a successful Kickstarter campaign. We also consider factors both in and out of creators' control that may help their campaigns succeed among the many active campaigns.

Specifically, we decided to include campaign length, whether the deadline is on a weekday, what season the deadline is in, and staff pick in additional models. These explanatory variables are only moderately correlated (see the Limitations - Statistical section).

In particular, goal amount and campaign length are positively correlated because campaigners with larger goals tend to give themselves more time to hit those goals. When we included these variables in our models, we remembered that the inferred effect sizes of the goal amount and campaign length might be impacted by each other. We also considered the campaign category as a categorical variable by itself, and we made it interact with the log of goal amount to understand whether the effect size of a campaign's goal amount differed by category. Lastly, we subset the data by campaign category to assess whether the significance of our dependent variables changed by category.

We will use robust standard errors throughout our models to follow best practice.

3.1 Model 1: Base Model

In the first modeling iteration, we used only the primary explanatory variable of interest: funding goal amount in USD. Due to data skewness, as depicted in the exploratory data analysis section, we log-transformed the funding goal amount. Kickstarter's platform requires the funding goal amount to be at least $1, so we were not concerned about negative values.

$$success\_metric = \beta_0 + \beta_1 \log_{10}\text{goal}$$

3.2 Model 2: Creator-Controlled Variables

The second model included additional explanatory variables that a campaigner could reasonably control when launching their Kickstarter project.

- `normalized_campaign_length`: the duration of the campaign on the platform (the number of days from campaign launched date to campaign end date).
  - We noted that most campaigns had a campaign length of Kickstarter's default, 30 days. To better understand the effect of campaign length, we transformed this variable by subtracting 30 from campaign length, giving us negative values in what was previously a positive-value-only field. This allowed us to interpret a negative coefficient to indicate that shorter campaigns are more likely to succeed. In contrast, a positive coefficient would mean that longer campaigns are more likely to succeed.
- `deadline_on_weekend` (categorical variable): shows if the deadline falls on a week or weekend (True, False).
  - Our causal theory suggests that whether the campaign ends on a weekday matters, but it does not consider the specific day of the week meaningful. Therefore, we created a binary variable indicating whether a campaign deadline was on a weekday.
- `deadline_season` (categorical variable): the season of the year the campaign ends.
  - We also wanted to consider seasonality non-ordinal (e.g. fall comes before winter, but fall does not necessarily "rank above" winter for our interest). Based both on our causal theory and to reduce the number of features, we chose not to break campaign deadline into 12-month categories but rather into 4-month groupings to indicate season:
    - Summer: June, July, August
    - Fall: September, October, November
    - Winter: December, January, February
    - Spring: March, April, May

$$\begin{aligned} success\_metric = \beta_0 &+ \beta_1 \log_{10}\text{goal} \\ &+ \beta_2 \text{normalized\_campaign\_length} + \beta_3 \text{deadline\_on\_weekday} \\ &+ \beta_4 \text{deadline\_season(Spring)} + \beta_5 \text{deadline\_season(Summer)} + \beta_6 \text{deadline\_season(Winter)} \end{aligned}$$

3.3 Model 3: Non-Controllable Variables

In the third model, we included everything from Model 2 and added variables the campaigner cannot control.

- `staff_pick` (categorical variable): indicates whether the campaign was marked as a "Kickstarter staff pick" while the campaign was live; the campaign does not have to be successful to be featured as a staff pick

- `clean_category`: transformed variable that only contains the main category of the campaign (e.g. music, art, food etc)

$$
\begin{aligned}
success\_metric = \beta_0 &+ \beta_1 \log_{10}\text{goal} \\
&+ \beta_2 \text{normalized\_campaign\_length} + \beta_3 \text{deadline\_on\_weekday} \\
&+ \beta_4 \text{deadline\_season(Spring)} + \beta_5 \text{deadline\_season(Summer)} + \beta_6 \text{deadline\_season(Winter)} \\
&+ \beta_7 \text{clean\_category(comics)} + \beta_8 \text{clean\_category(crafts)} + \beta_9 \text{clean\_category(dance)} \\
&+ \beta_{10} \text{clean\_category(design)} + \beta_{11} \text{clean\_category(fashion)} + \beta_{12} \text{clean\_category(film)} \\
&+ \beta_{13} \text{clean\_category(food)} + \beta_{14} \text{clean\_category(games)} + \beta_{15} \text{clean\_category(journalism)} \\
&+ \beta_{16} \text{clean\_category(music)} + \beta_{17} \text{clean\_category(photography)} + \beta_{18} \text{clean\_category(publishing)} \\
&+ \beta_{19} \text{clean\_category(technology)} + \beta_{20} \text{clean\_category(theater)} + \beta_{21} \text{staff\_pick}
\end{aligned}
$$

3.4 Model 4: Category and Goal Interactions

In the fourth model, we wanted to further our investigation of how categories affect campaign success. In the third model, we observed the initial effect of categories; however, we did not observe how categories impacted the success rate as the project's goal increased.

To accomplish this, we wanted to observe an interaction variable between `category` and `log10(goal)`. We believed that adding this to Model 3 would add too many degrees of freedom to get valid results, so we returned to Model 1 as a basis to define our model:

$$
\begin{aligned}
success\_metric = \beta_0 &+ \beta_1 \log_{10}\text{goal} \\
&+ \beta_2 \text{clean\_category(comics)} + \beta_3 \text{clean\_category(comics)}{:}\log_{10}\text{goal} \\
&+ \beta_4 \text{clean\_category(crafts)} + \beta_5 \text{clean\_category(crafts)}{:}\log_{10}\text{goal} \\
&+ \beta_6 \text{clean\_category(dance)} + \beta_7 \text{clean\_category(dance)}{:}\log_{10}\text{goal} \\
&+ \beta_8 \text{clean\_category(design)} + \beta_9 \text{clean\_category(design)}{:}\log_{10}\text{goal} \\
&+ \beta_{10} \text{clean\_category(fashion)} + \beta_{11} \text{clean\_category(fashion)}{:}\log_{10}\text{goal} \\
&+ \beta_{12} \text{clean\_category(film)} + \beta_{13} \text{clean\_category(film)}{:}\log_{10}\text{goal} \\
&+ \beta_{14} \text{clean\_category(food)} + \beta_{15} \text{clean\_category(food)}{:}\log_{10}\text{goal} \\
&+ \beta_{16} \text{clean\_category(games)} + \beta_{17} \text{clean\_category(games)}{:}\log_{10}\text{goal} \\
&+ \beta_{18} \text{clean\_category(journalism)} + \beta_{19} \text{clean\_category(journalism)}{:}\log_{10}\text{goal} \\
&+ \beta_{20} \text{clean\_category(music)} + \beta_{21} \text{clean\_category(music)}{:}\log_{10}\text{goal} \\
&+ \beta_{22} \text{clean\_category(photography)} + \beta_{23} \text{clean\_category(photography)}{:}\log_{10}\text{goal} \\
&+ \beta_{24} \text{clean\_category(publishing)} + \beta_{25} \text{clean\_category(publishing)}{:}\log_{10}\text{goal} \\
&+ \beta_{26} \text{clean\_category(technology)} + \beta_{27} \text{clean\_category(technology)}{:}\log_{10}\text{goal} \\
&+ \beta_{28} \text{clean\_category(theater)} + \beta_{29} \text{clean\_category(theater)}{:}\log_{10}\text{goal}
\end{aligned}
$$

4 Results

Table 2: Model Results

	Dependent variable:		
	Pledged Amount Reached or Exceeded Goal Before Deadline		
	(1)	(2)	(3)
Log10 of Goal Amount in USD	-0.177^{***} (0.005)	-0.135^{***} (0.006)	-0.137^{***} (0.006)
Normalized Campaign Length		-0.008^{***} (0.0003)	-0.006^{***} (0.0003)
Deadline was on a Weekday		-0.030^{***} (0.008)	-0.022^{***} (0.007)
Deadline was in Spring		-0.020^{**} (0.010)	-0.027^{***} (0.009)
Deadline was in Summer		-0.011 (0.010)	-0.004 (0.009)
Deadline was in Winter		-0.021^{**} (0.010)	-0.018^{**} (0.009)
Comics Category			0.065^{***} (0.013)
Crafts Category			-0.269^{***} (0.031)
Dance Category			-0.119 (0.104)
Design Category			0.187^{***} (0.012)
Fashion Category			-0.004 (0.018)
Film Category			-0.132^{***} (0.019)
Food Category			-0.090^{***} (0.020)
Game Category			0.119^{***} (0.012)
Journalism Category			-0.340^{***} (0.037)
Music Category			0.016 (0.018)
Photography Category			-0.118^{***} (0.039)
Publishing Category			0.020 (0.014)
Technology Category			-0.264^{***} (0.019)
Theater Category			-0.113^{*} (0.058)
Was Staff Pick			0.247^{***} (0.007)
Constant	1.416^{***} (0.018)	1.307^{***} (0.021)	1.258^{***} (0.020)
Observations	11,196	11,196	11,196
R^2	0.088	0.151	0.310
Adjusted R^2	0.088	0.151	0.309
Residual Std. Error	0.395 (df = 11194)	0.381 (df = 11189)	0.344 (df = 11174)
F Statistic	$1{,}081.877^{***}$ (df = 1; 11194)	332.441^{***} (df = 6; 11189)	239.028^{***} (df = 21; 11174)
Note:			*p<0.1; **p<0.05; ***p<0.01

4.1 Model 1

In our base model, which does not include any covariates, the log-transformed funding goal amount variable was statistically significant, with a p-value < 0.001. We rejected the null hypothesis that the funding goal amount does not impact a Kickstarter campaign's success. Although we had a statistically significant p-value and rejected the null hypothesis, it was important to note that we had many observations. Thus, we considered additional statistics to explain our base model, like the coefficient of determination and the effect size. The log-transformed funding goal amount variable explained 8.8% of the variance in whether a campaign was successful or failed (adj. $R^2 = 0.088$). Since we limited our dataset to campaigns with a goal of a minimum of \$100, when we plugged the lowest value into our model, it said that campaigns with a goal of \$100 have $\sim$ 106% chance of success. Furthermore, for every increase in the order of magnitude (e.g. $\log_{10}(\$1000)$), the chance of success decreases by -17.7 $\pm$ 0.5%.

Variable	Categories for which the variable significantly...	
	Increases chance of success (p-value < 0.1)	Decreases chance of success (p-value < 0.1)
deadline_on_weekday		Design, Technology
deadline_season(spring)		Games, Music, Technology
deadline_season(summer)	Art, Other	Games, Fashion
deadline_season(winter)	Other	Music

Note:
The following categories did not find any of the deadline detail variables to be significant: Comics, Film, Food, and Publishing

4.2 Model 2

Model 2 included multiple categorical independent variables, `deadline_on_weekday` and `deadline_season` variables, and an independent metric variable, `normalized_campaign_length`. For the `deadline_on_weekday` variable, `False` was the baseline, and the model showed that if the campaign ends on a weekday, there is a -3 $\pm$ 0.8% chance of the campaign successfully reaching its goal. This negative effect was proven statistically significant with a p-value < 0.001. In terms of the `deadline_season` variable, `fall` (September through November) was the reference, and the model returned a statistically significant negative coefficient for both `spring` and `winter`, exhibiting there is a -2.1 $\pm$ 1% chance of a campaign reaching its funding goal. The coefficient to explain how `summer` impacts a campaign's success is not statistically significant. Lastly, the normalized campaign length coefficient was statistically significant and showed that the chance of campaign success goes down by -0.8% for each day past thirty days. Overall, Model 2 explains 15.1% of the variance in our dependent variable, which is almost double Model 1's R^2 value.

4.2.1 Model 2 Category Subsets

In addition to testing `category` as a variable, we tested Model 2 for each category. By doing this, we were able to determine if the various factors we analyzed for impact on campaign success differ between categories. Our analysis revealed a few noteworthy differences.

First, the log-transformed `goal` variable was statistically significant for every category. Second, the `normalized_campaign_length` was statistically significant for every category, and for every day past thirty days the campaign ran, its chance of successfully reaching its funding goal declined by around 1%. The deadline seasonality and deadline day of the week classification demonstrated varying effects across categories, with some being statistically significant and others not.

Table 3: Model 2 Results by Category: Set 1

	Dependent variable:		
	success_metric		
	art	comics	design
	(1)	(2)	(3)
Log10 of Goal Amount in USD	-0.126^{***} (0.019)	-0.041^{***} (0.014)	-0.056^{***} (0.013)
Normalized Campaign Length	-0.006^{***} (0.001)	-0.009^{***} (0.001)	-0.001 (0.001)
Deadline was on a Weekday	0.003 (0.020)	-0.022 (0.015)	-0.040^{***} (0.015)
Deadline was in Spring	0.009 (0.030)	0.019 (0.019)	-0.002 (0.017)
Deadline was in Summer	0.046^{*} (0.025)	0.021 (0.018)	0.020 (0.015)
Deadline was in Winter	0.040 (0.027)	-0.022 (0.022)	-0.018 (0.019)
Constant	1.198^{***} (0.059)	1.068^{***} (0.051)	1.161^{***} (0.049)
Observations	1,228	1,277	1,430
R^2	0.123	0.129	0.032
Adjusted R^2	0.119	0.125	0.027
Residual Std. Error	0.342 (df = 1221)	0.254 (df = 1270)	0.228 (df = 1423)
F Statistic	28.604^{***} (df = 6; 1221)	31.350^{***} (df = 6; 1270)	7.716^{***} (df = 6; 1423)
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 4: Model 2 Results by Category: Set 2

	Dependent variable:		
	success_metric		
	fashion	film	food
	(1)	(2)	(3)
Log10 of Goal Amount in USD	-0.213^{***} (0.027)	-0.132^{***} (0.023)	-0.162^{***} (0.027)
Normalized Campaign Length	-0.009^{***} (0.002)	-0.008^{***} (0.001)	-0.008^{***} (0.001)
Deadline was on a Weekday	-0.010 (0.030)	-0.051 (0.036)	0.004 (0.035)
Deadline was in Spring	-0.073 (0.045)	0.020 (0.047)	0.005 (0.047)
Deadline was in Summer	-0.099^{***} (0.038)	-0.003 (0.046)	0.010 (0.047)
Deadline was in Winter	0.043 (0.041)	-0.024 (0.048)	-0.058 (0.048)
Constant	1.540^{***} (0.090)	1.177^{***} (0.093)	1.325^{***} (0.109)
Observations	533	758	748
R^2	0.285	0.110	0.112
Adjusted R^2	0.277	0.103	0.105
Residual Std. Error	0.343 (df = 526)	0.466 (df = 751)	0.460 (df = 741)
F Statistic	34.992^{***} (df = 6; 526)	15.539^{***} (df = 6; 751)	15.571^{***} (df = 6; 741)
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 5: Model 2 Results by Category: Set 3

	Dependent variable:		
	success_metric		
	games	music	publishing
	(1)	(2)	(3)
Log10 of Goal Amount in USD	-0.032^{***} (0.010)	-0.028 (0.029)	-0.072^{***} (0.020)
Normalized Campaign Length	-0.006^{***} (0.001)	-0.009^{***} (0.001)	-0.005^{***} (0.001)
Deadline was on a Weekday	-0.006 (0.013)	-0.026 (0.032)	-0.031 (0.021)
Deadline was in Spring	-0.062^{***} (0.018)	-0.078^{*} (0.042)	-0.018 (0.029)
Deadline was in Summer	-0.071^{***} (0.018)	-0.048 (0.039)	0.030 (0.026)
Deadline was in Winter	-0.013 (0.013)	-0.086^{**} (0.041)	-0.010 (0.027)
Constant	1.052^{***} (0.036)	0.967^{***} (0.112)	1.126^{***} (0.070)
Observations	1,746	729	1,158
R^2	0.089	0.098	0.055
Adjusted R^2	0.086	0.090	0.051
Residual Std. Error	0.254 (df = 1739)	0.404 (df = 722)	0.341 (df = 1151)
F Statistic	28.296^{***} (df = 6; 1739)	13.017^{***} (df = 6; 722)	11.264^{***} (df = 6; 1151)
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 6: Model 2 Results by Category: Set 4

	Dependent variable:	
	success_metric	
	technology	other categories
	(1)	(2)
Log10 of Goal Amount in USD	-0.233^{***} (0.019)	-0.165^{***} (0.028)
Normalized Campaign Length	-0.005^{***} (0.001)	-0.009^{***} (0.001)
Deadline was on a Weekday	-0.094^{***} (0.029)	0.007 (0.040)
Deadline was in Spring	-0.098^{**} (0.041)	0.070 (0.056)
Deadline was in Summer	-0.049 (0.037)	0.129^{**} (0.053)
Deadline was in Winter	-0.041 (0.038)	0.152^{***} (0.052)
Constant	1.455^{***} (0.084)	1.074^{***} (0.107)
Observations	1,024	565
R^2	0.155	0.174
Adjusted R^2	0.150	0.165
Residual Std. Error	0.448 (df = 1017)	0.457 (df = 558)
F Statistic	31.019^{***} (df = 6; 1017)	19.530^{***} (df = 6; 558)
Note:		*p<0.1; **p<0.05; ***p<0.01

4.3 Model 3

Model 3 included all the variables in Model 2 with the addition of two categorical variables that are out of a creator's control, `category` and `staff_pick`. The inclusion of these variables increased the R^2 value to 30.9%. For the 'staff_pick' variable, the default was `False`, and the model showed that if a campaign was marked as a staff pick its chance for success increased by 25%. The model did not show much change in the log-transformed `goal` coefficient from Model 2, and it was still statistically significant. There was a slight change in the `deadline_on_weekday` effect from -3 ± 0.8% to -2.2 ± 0.7%. The `deadline_season` variables' impacts changed slightly from a campaign with a `spring` deadline being -2.7 ± 0.9% less likely to reach its funding goal, compared to `fall`, and `winter` decreasing the chance of success by -1.8 ± 0.9%. The `normalized_campaign_length` variable's negative impact on campaign success decreased because it went from -0.8% less likely to reach the funding goal to -0.6% less likely.

- Categories that lower success chance (significantly), compared to art:
  - crafts (-26.9 ± 2.5%)
  - film (-13.2 ± 1.7%)
  - food (-9 ± 1.7%)
  - journalism (-34 ± 3.1%)
  - technology (-26.4 ± 1.6%)
  - theater (-11.3 ± 4.9%)
- Categories that increase success chance (significantly), compared to art:
  - comics (6.5 ± 1.4%)
  - design (18.7 ± 1.4%)
  - games (11.9 ± 1.3%)
  - photography (-11.8 ± 3.3%)
- Categories with no (significant) impact on success chance, compared to art:
  - dance
  - fashion
  - music
  - publishing
  - theater (has only one * of significance)

The five categories that did not present statistical significance in Model 3 compared to `art` make sense because they all fall into a kind of "arts" category themselves. Therefore, we would not expect to find any statistically significant difference when comparing `art` to "arts".

4.4 Model 4

Model 4 includes interaction terms between the log-transformed `goal` and category variables. We wanted to test whether the effect of the log-transformed `goal` on success differs among categories.

Art, our default category, has an intercept of 1.34, at goal=0, with an adjusted intercept of 1.01, at goal=$100, and a slope of -0.17. This means that there is ~101% probability that a $100 project is successful for art and that the projected success rate decreases by -17% per order of magnitude of the target goal.

The chart below shows the relative change in these coefficients for other categories and the resulting take-aways. For example, a Comics campaign for $100 has an expected 101% chance of success, and the projected success rate decreases by -6% per order of magnitude (O.O.M.) of the target goal.

	Delta v. Art		Effective	
Category	Delta Base % @$100	Delta %/O.O.M.	Effective Base % @$100	Effective %/O.O.M.
comics	0%	10%	101%	-6%
crafts	-30%†	-15%	71%†	-32%
design	2%	11%	104%	-6%
fashion	12%	-10%	114%	-27%
games	-1%	11%	101%	-5%
journalism	-52%§	N/S	50%§	N/S
music	-16%	11%	86%	-6%
publishing	-3%	8%	98%	-8%
technology	-18%†	-9%	84%†	-25%

Note:

N/S = Not Significant

† = Delta intercept @$0 N/S, recalculated using Art Intercept @$0 with significant delta %/O.O.M.

§ = Delta slope N/S, recalculated using Art %/O.O.M. with significant delta intercept @$0

(dance, film, food, photography, theater had no significant changes)

Table 7: Model 4 Results

	Dependent variable:
	success_metric
factor(clean_category)comics	-0.204^{***} (0.073)
factor(clean_category)crafts	0.212 (0.149)
factor(clean_category)dance	-0.870 (0.895)
factor(clean_category)design	-0.195^{***} (0.071)
factor(clean_category)fashion	0.323^{***} (0.091)
factor(clean_category)film	-0.123 (0.103)
factor(clean_category)food	0.060 (0.119)
factor(clean_category)games	-0.230^{***} (0.062)
factor(clean_category)journalism	-0.515^{**} (0.228)
factor(clean_category)music	-0.376^{***} (0.127)
factor(clean_category)photography	-0.045 (0.225)
factor(clean_category)publishing	-0.197^{**} (0.085)
factor(clean_category)technology	0.084 (0.097)
factor(clean_category)theater	0.212 (0.287)
log10(goal)	-0.165^{***} (0.019)
factor(clean_category)comics:log10(goal)	0.101^{***} (0.024)
factor(clean_category)crafts:log10(goal)	-0.150^{***} (0.046)
factor(clean_category)dance:log10(goal)	0.217 (0.239)
factor(clean_category)design:log10(goal)	0.109^{***} (0.023)
factor(clean_category)fashion:log10(goal)	-0.101^{***} (0.031)
factor(clean_category)film:log10(goal)	0.003 (0.029)
factor(clean_category)food:log10(goal)	-0.033 (0.033)
factor(clean_category)games:log10(goal)	0.111^{***} (0.021)
factor(clean_category)journalism:log10(goal)	0.047 (0.059)
factor(clean_category)music:log10(goal)	0.109^{***} (0.037)
factor(clean_category)photography:log10(goal)	-0.010 (0.067)
factor(clean_category)publishing:log10(goal)	0.081^{***} (0.027)
factor(clean_category)technology:log10(goal)	-0.088^{***} (0.026)
factor(clean_category)theater:log10(goal)	-0.089 (0.078)
Constant	1.345^{***} (0.054)
Observations	11,196
R^2	0.243
Adjusted R^2	0.241
Residual Std. Error	0.360 (df = 11166)
F Statistic	123.786^{***} (df = 29; 11166)
Note:	*p<0.1; **p<0.05; ***p<0.01

5 Limitations

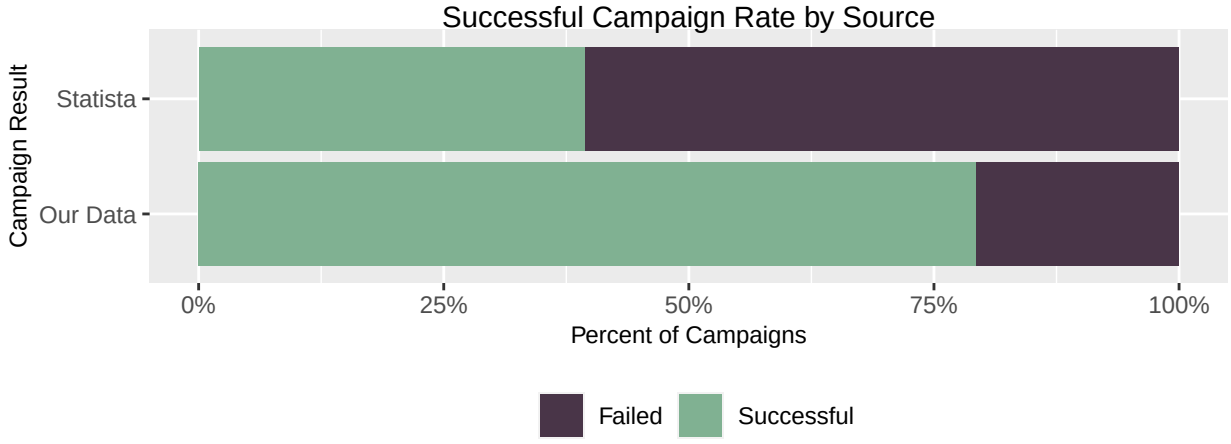
5.1 Statistical

Assumption 1: Independent and identically distributed (I.I.D.) samples
Due to the timeline of this research project and technical problems with DataHub, we did not fully resolve all the issues present in our dataset, which has implications for assuming the sample is I.I.D. An example of one of these issues is that the web crawler pages through the history of projects for each category; however, Kickstarter only provides a finite number of pages to traverse.

We observed projects launched back in 2014 in the latest 3/2022 dataset, yet successful projects launched in 2021 were found in earlier datasets (e.g., the 9/2021 dataset included projects that were not in the 3/2022 dataset). Therefore, we can safely confirm that projects are not pushed off of Kickstarter's categorial pagination through chronology only. Kickstarter's algorithms used to determine whether projects appear in history are a blackbox and almost certainly violate I.I.D. requirements in the strictest sense. It is improbable that the projects that make it into pagination are through random selection. Sampling additional datasets not used in our final dataset, e.g. 10/2021, confirms that there are still post-pandemic projects that are not captured in any of our four chosen datasets.

A solution to minimize the number of missing projects and allow us to better converge on I.I.D. would be to sample the web crawler results more frequently than every six months. A preliminary discovery found, for example, that sampling every three months instead of every six months would have yielded 10% more post-pandemic projects. However, due to R's poor memory management and DataHub's low maximum memory allocation, our efforts to get data every six months reached diminishing returns. This prevented us from loading additional datasets without DataHub's Virtual Machine crashing. With proper resources and time, we could have scanned every monthly dataset published; however, it is worth noting that even this would almost certainly not have acquired all projects.

Some external findings confirm this. Statista compiled a report that only ~39% of Kickstarter campaigns are successful[8], compared to 79% of our data's campaigns.

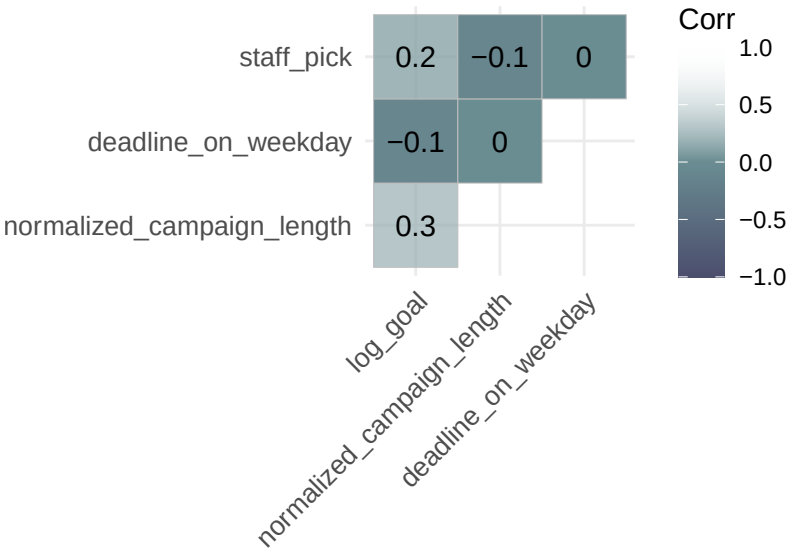


When deduplicating projects with multiple categories, we pick the category of the first occurrence seen in the CSVs for a given dataset. For proper I.I.D., we should be picking one of the categories randomly, especially given that categories are likely crawled in the same order each time. By selecting the first one seen, specific categories are likely overrepresented. Unfortunately, because of DataHub's limited resources, we had to perform deduplication on each chunk of data as we loaded it. We would have preferred loading all duplicates and then performing deduplication randomly on the full dataset.

Assumption 2: A unique best linear predictor (BLP) exists
A unique BLP exists so long as the variables in our models are not perfectly collinear.

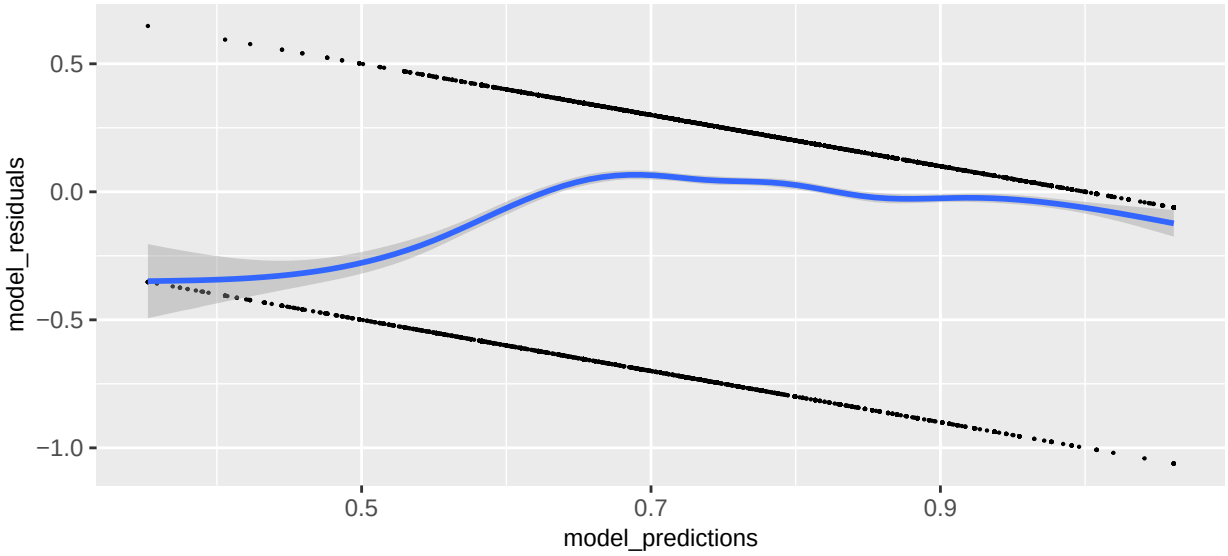
[8] "Kickstarter: Project Funding Success Rate 2021." *Statista*, 13 Dec. 2021, https://www.statista.com/statistics/235405/kickstarter-project-funding-success-rate/.

The correlation coefficients in the correlation matrix below show no perfect collinearity between our chosen metric variables, `log(goal)`, `normalized_campaign_length`, `deadline_on_weekday`, and `staff_pick`. We used R's `factor()` function for our categorical variables, `deadline_season` and `clean_category`, which automatically drops one factor for each variable to prevent perfect collinearity.



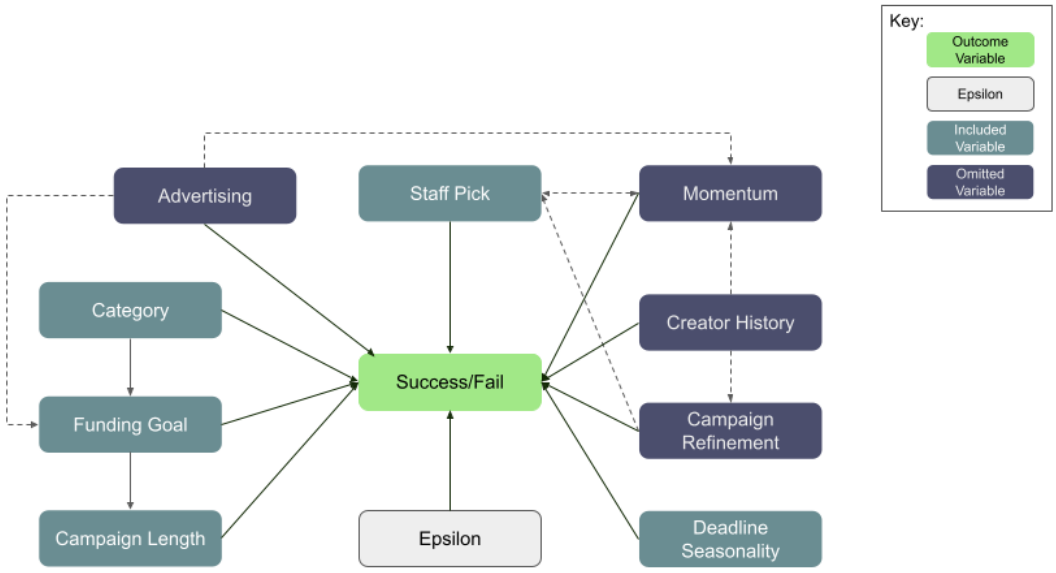
Assumption 3: Conditional Linear Expectation

The goals of our sample projects were mainly clustered between $1,000 and $50,000. Due to this, the model is disproportionately fitted to projects within this range, causing accuracy degradation for lower and extremely high goals. The graph below shows that campaign success is severely overpredicted for large-goal campaigns on the left side (which already have the lowest predicted success rate). In the rightmost octile, you can see a sudden rigid linear downward overprediction that represents where the model predicts over 100% success for low-goal campaigns, which we know to be impossible.



5.2 Structural

The success of creative endeavors depends greatly on qualitative factors, luck, and factors localized to the particular type of endeavor. Thus, we recognize that the causal theory applied here is greatly simplified. Below is a slightly less simplified causal diagram to highlight some omitted variables.



Intentionally omitted variables
We intentionally chose to omit the following variables:

- location
- state
- name
- blurb

Despite limiting our dataset to campaigns located only in the United States, the 'location' variable contained many unique categorical variables. Given our sample size, we could not effectively model with all unique locations included. Due to time and hardware constraints, we chose to forego a cluster analysis to reduce the number of variables.

Kickstarter campaign names and blurbs are unique to each campaign and frequently contain non-ASCII characters, such as emojis and other characters outside of the English alphabet. Therefore, due to time and constraints, we chose to refrain from conducting string analyses of the name or campaign blurb.

Unintentionally omitted variables
There are several omitted variables we would have liked to include in our model if the data were available. These are listed below and followed by their possible implications.

- momentum
- advertising
- creator experience
- campaign refinement

Momentum
Campaign "momentum" generally describes how much early support a project receives, such as the number of backers or money pledged in the first 48 hours of a campaign's launch. This has two effects: 1) it increases the chances of "bandwagoning", with the idea that someone else has already determined that this project is something worth backing, and 2) it contributes to visibility, as Kickstarter has a page for "trending" or

"popular" projects. We expect momentum to be positively correlated with success. In contrast, we think momentum would be negatively correlated with the funding goal amount. Hence, the omitted variable campaign momentum has an unknown effect on the coefficients.

Advertising
In recent years, Kickstarter campaign creators have started advertising through social media ads on sites such as Instagram and YouTube. We expect this to be positively correlated with success since project visibility is increased. We would also anticipate the funding goal amount to increase because the creator might include advertising costs in their campaign budget. This causes the "true" coefficient value of the funding goal amount to be higher than our fitted coefficient value. Since the coefficient in front of the funding goal amount is negative, this omitted variable bias is towards zero.

Creator Experience
Since Kickstarter has been crowdfunding projects for over a decade, some creators have had time to come back and run multiple campaigns. Running previous campaigns gives creators access to previous backers' emails, which can act as a base of support for new campaigns. Furthermore, their experience might allow them to create more polished campaigns (see next section). We expect creator history to be positively correlated with campaign success. Assuming that smaller goal amounts reflect projects that take less effort, we expect creator history to be negatively correlated with the funding goal amount. The more projects a creator has launched in the past, the more likely their campaigns are smaller-scale campaigns. This causes the "true" coefficient value of the funding goal amount to be lower than our fitted coefficient value, meaning that this omitted variable bias is away from zero.

Measures of Campaign Refinement
Under the assumption that people want to crowdfund projects they believe will succeed while also understanding what they are paying for, a campaign with detailed information is more likely to attract backers than a poorly worded, poorly visualized campaign. The length of a campaign's description and the number of images included could be considered measurements of the level of details provided. Furthermore, it has become increasingly popular for creators to include "stretch goals" in their campaign descriptions. The stretch goals are often shared in illustration form and entice backers to pledge additional money for additional rewards. With all this in mind, we would expect the inclusion of the description's character length and/or the number of images to be positively correlated with a campaign's success. It is not easy to say how an increased level of detail in a campaign would impact the funding goal amount; therefore, we need to test this variable before knowing its effect on the coefficients.

6 Conclusion

6.1 Overall Conclusion

In this study, we sought to provide Kickstarter campaign creators with an analysis to help them understand how to create a successful campaign. After thorough data exploration, we realized limitations with the data Kickstarter allows to be available to the public. With these limitations in mind, we continued our research and analysis and provided clear explanations for the challenges we faced and their impact on our findings.

In summary, we combined four datasets from Web Robots scraping the Kickstarter platform and wrote up a detailed outline of the data cleansing, transformation, and pipeline building we performed. We also listed the limitations we faced while working with the data, including large-sample model assumptions and omitted variable biases. We tested multiple models and provided model descriptions and an explanation of our results.

Our analysis discovered that the log-transformed funding goal amount a creator sets for their Kickstarter campaign significantly impacts their likelihood of success. We looked into the effect of additional variables and detailed their impact results. To highlight a few, we found that having a campaign deadline on a weekday, compared to a weekend, and in the spring or winter months, compared to the fall months, decreases a campaign's chance of successfully reaching its funding goal. Additionally, for each day past thirty days a

creator runs their campaign, they diminish their chance of reaching their funding goal by a little less than 1%. Furthermore, we confirmed our theory that staff-picked campaigns would be more successful because when `staff_pick` = True, a campaign is 25% more likely to reach its funding goal.

In conclusion, and based on our sample data, our analysis shed light on a few factors creators can consider to increase their chances of successfully reaching their funding goal amount when creating a Kickstarter campaign. Due to the various limitations we described, we strongly suggest further studying these factors' effects.

6.2 Further Study

We hope that our study results will help inspire Kickstarter creators in bringing their campaigns to life. While our models have relatively low R^2 values, we expected this to be the case. We would be disappointed if we found that metadata details and aspects out of creators' control, such as `staff_pick`, more strongly determined whether a project was successful than the unique aspects of each campaign. With that in mind, we still hope this guides creators regarding those metadata details that are only incidental to their core project but may help get them to the 100% pledged mark to reach success.

Besides qualitative, unique details, we also highlighted some omitted variables, such as advertising money and minimum reward tier, that we believe quantitative and highly important but not available in our existing dataset. These omitted variables could form the basis for future research.

In the interest of reproducibility, we recommend re-running this study with more samples. Due to hardware limitations, we could not include all of the data that we wanted to.

7 Appendix

7.1 Kickstarter Data Dictionary

- `backers_count`: integer value of the total backers that this campaign had at the time of data pull
- `blurb`: varchar of the subtitle of the campaign
- `category`: JSON blob with values such as category ID, category name, and parent category of the campaign (e.g. music, specifically pop music)
- `converted_pledged_amount`: integer value of amount pledged to the campaign at the time of data pull, converted from whatever currency the funding goal is in to the currency_currency
- `country`: two letter ISO code for the country that the campaign was launched in
- `country_displayable_name`: varchar name of the country that the campaign was launched in
- `created_at`: when the campaign was first created in Kickstarter (think of this as entering "draft" inside of Kickstarter)
  - In the original dataset, this is in epoch time in seconds
  - We transform this to be POSIXct type, in format YYYY-MM-DD HH:MM:SS
- `creator`: JSON blob identifying who created the campaign with values such as creator ID, creator name, and the image link for the creator's profile picture
- `currency`: three letter currency code identifying the currency that the funding goal for this campaign is in
- `currency_symbol`: symbol for the currency, such as $ for USD
- `currency_trailing_code`: boolean value for whether the currency has a trailing code
- `current_currency`: the default currency for the account that is looking at the campaign; in this case, the webcrawler always defaults to USD
- `deadline`: the deadline for the campaign in Kickstarter, at which point no more pledges will be accepted. If the funding goal is reached by this deadline, then funds are taken from backers and released to the creator, and the campaign is marked as successful. If the funding goal is not reached by this deadline, no funds are taken from backers, and the campaign is marked as a failure.
  - In the original dataset, this is in epoch time in seconds
  - We transform this to be POSIXct type, in format YYYY-MM-DD HH:MM:SS
- `disable_communication`: boolean value for whether the campaign has disabled communications
- `friends`: whether or not the user currently looking at the campaign is friends (e.g. through social media account linking) with the campaign creator
- `fx_rate`: a float value specifying the foreign exchange rate at time of data pull between the original campaign goal currency and current_currency
- `goal`: an integer value specifying the goal amount for this campaign, automatically converted from the original goal amount in its original currency to an amount displayed in current_currency
- `id`: the unique identifier for this Kickstarter campaign
- `is_backing`: boolean value for whether the account that is looking at the campaign is a campaign backer at the time of data pull; in this case, the webcrawler always defaults to null or false
- `is_starrable`: boolean value for whether or not this campaign can be starred by users at the time of data pull; false for inactive campaigns
- `is_starred`: boolean value for whether the account that is looking at the campaign has starred this campaign or not at the time of data pull; in this case, the webcrawler always defaults to null or false
- `launched_at`: the time at which the Kickstarter campaign launched, as in was open for backers to begin making pledges
  - In the original dataset, this is in epoch time in seconds
  - We transform this to be POSIXct type, in format YYYY-MM-DD HH:MM:SS
- `location`: JSON blob identifying the location of the campaign, as selected by the creator and as displayed on the campaign page, with values such as country, state, and localized name of the location
- `name`: varchar value that is the name/title of the campaign

- `permissions`: multiple users can be added as creators of a campaign; this indicates what permissions the account that is looking at the campaign has; in this case, the webcrawler always has the same value
- `photo`: JSON blob containing IDs of and links to various versions of the primary photo (or video thumbnail) of the Kickstarter campaign
- `pledged`: total amount pledged to the campaign at time of data pull, in the original currency of the funding goal
- `profile`: JSON blob with details about the campaign's profile including visual settings
- `slug`: name of the campaign with hyphens instead of spaces, appended after the "kickstarter.com" hostname to resolve to this campaign's page
- `source_url`: where the web crawler found the campaign from; in this case, based on how the web crawler works, this is always the campaign category
- `spotlight`: boolean value that indicates whether the campaign made use of the "spotlight" feature, available only to successful campaigns, which essentially preserves the campaign's journey and allows creators to communicate with backers easily in one webpage
- `staff_pick`: boolean value that indicates whether the campaign was marked as a "Kickstarter staff pick" while the campaign was live; the campaign does not have to be successful to be featured as a staff pick
- `state`: campaign state at the time of data pull; options are successful or failed, as detailed previously, as well as canceled, meaning that the campaign creator canceled the campaign before the campaign deadline, or "active" for campaigns whose deadlines have not been reached and are not canceled as of time of data pull
- `state_changed_at`: when the campaign status was changed; this is the same as the deadline for successful and failed campaigns
  - In the original dataset, this is in epoch time in seconds
  - We transform this to be POSIXct type, in format YYYY-MM-DD HH:MM:SS
- `static_usd_rate`: the fixed conversion rate between the original campaign goal currency and USD for the first transaction
- `urls`: JSON blob containing URL links to various pieces of the campaign
- `usd_exchange_rate`: the exchange rate between the original campaign goal currency and USD at the time the funds were disbursed
- `usd_pledged`: pledged amount converted to USD, using the static_usd_rate, and therefore the actual USD value of the amount pledged that was disbursed
- `usd_type`: either "domestic" or "international"; it is unclear what this value means