Ciudad de México, México March 8, 2021

TEC de Monterrey Data Analysis Boot Camp

Inst. Alex Sánchez Vega

Kickstarter Campaigns Statistical Analysis

Executive Report

Juan José Fernández Díaz

[juanjo.fde@gmail.com](mailto:juanjo.fde@gmail.com)

Overview:

Kickstarter is a crowdfunding service launched on April 28, 2009 by Perry Chen and Yancey Strickler. After their initial success on 2012 Kickstarter expands operations outside the U.S. first to the United Kingdon and then moving to a variety of countries, Kickstarter begins operations in Mexico by 2016.

As a result of the coronavirus pandemic Kickstarter was forced to lay-off around 45% of their workers, mostly because the number of projects on the platform decreased by 35%.

Despite this recent events Kickstarter remains as one of the most successful crowdfunding services available but the reduction of active campaigns and a global economic crisis, raise, even more, the importance of finding and analyzing the key to a successful campaign.

Objectives:

* To analyze data from more than 4,000 past campaigns (2009-2015) in order to draw conclusions (at least 3) that help determine whether a campaign is successful or not.
* To identify the potential limitations of the dataset available.
* To suggest further analysis to be made in case the project continues forward.
* To determine the statistical approach best suited for the data provided (mean or median)
* To determine if variability in number of backers for a campaign plays an important role in the outcome of the campaign.

Special considerations:

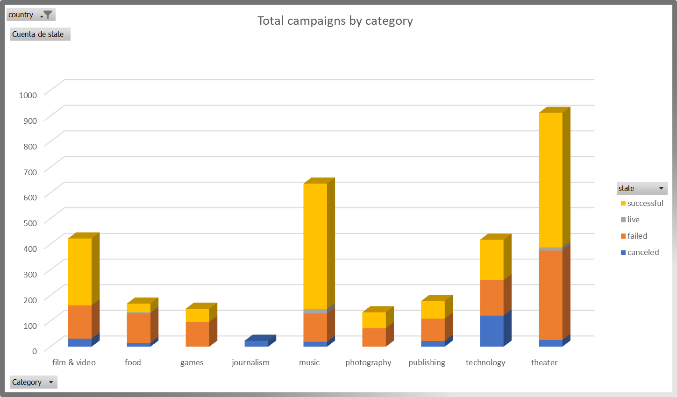
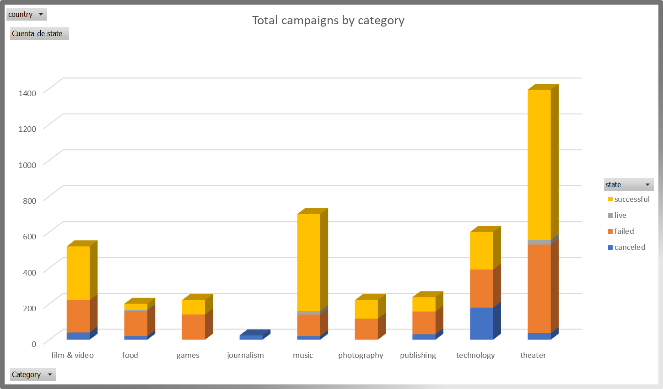
* The workflow of the Data Analysis used in this project can be found in the readme file contained in the repository of this project.
* Data was provided by the project instructor Alex Sánchez Vega

Conclusions:

1. The category of the campaign is important to determine the outcome.
   1. Just by looking at the graph we can already cross out journalism as a viable category to launch a funding campaign, data shows no successful campaigns.
   2. To further determine viable categories, we will consider that the average ratio for a campaign to be successful is around 2:1 and by doing that we can say that the best categories to start a campaign are: *music, theather, film&video and probably photography*
   3. We can clearly determine that category plays an important role determining the success of a campaign, though it is important to point out that this data shows the success of a funding campaign not the success of the project being funded. i.e., journalism is clearly the worst choice for starting a kickstarter campaign, but we all know that journalism is a successful industry.
2. Importance of sub-category in the outcome
3. By doing a deeper analysis in one of our top categories (music) we can clearly see that sub-category is vital determining the success of a campaign, by choosing the right sub-category we can practically guarantee the success of a campaign, let see if this statement holds for other categories.

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Theater by sub-category | Film & Video by subcategory | Photography by subcategory |

1. Analyzing our other top categories film & video shows the same behavior as music but about theater and photography even though we cannot eliminate the risk of failing a campaign we can greatly increase the probability of success by choosing the appropriate sub-category.
2. Analyzing our tough category, photography, the analysis show that it would be a viable category but only in the photobook sub-category.
3. Does the country in which the campaign was created affects its outcome?
   1. In order to compare if country affects the outcome of a campaign, I compared the total campaigns with the country that has the greatest number of campaigns available (US)



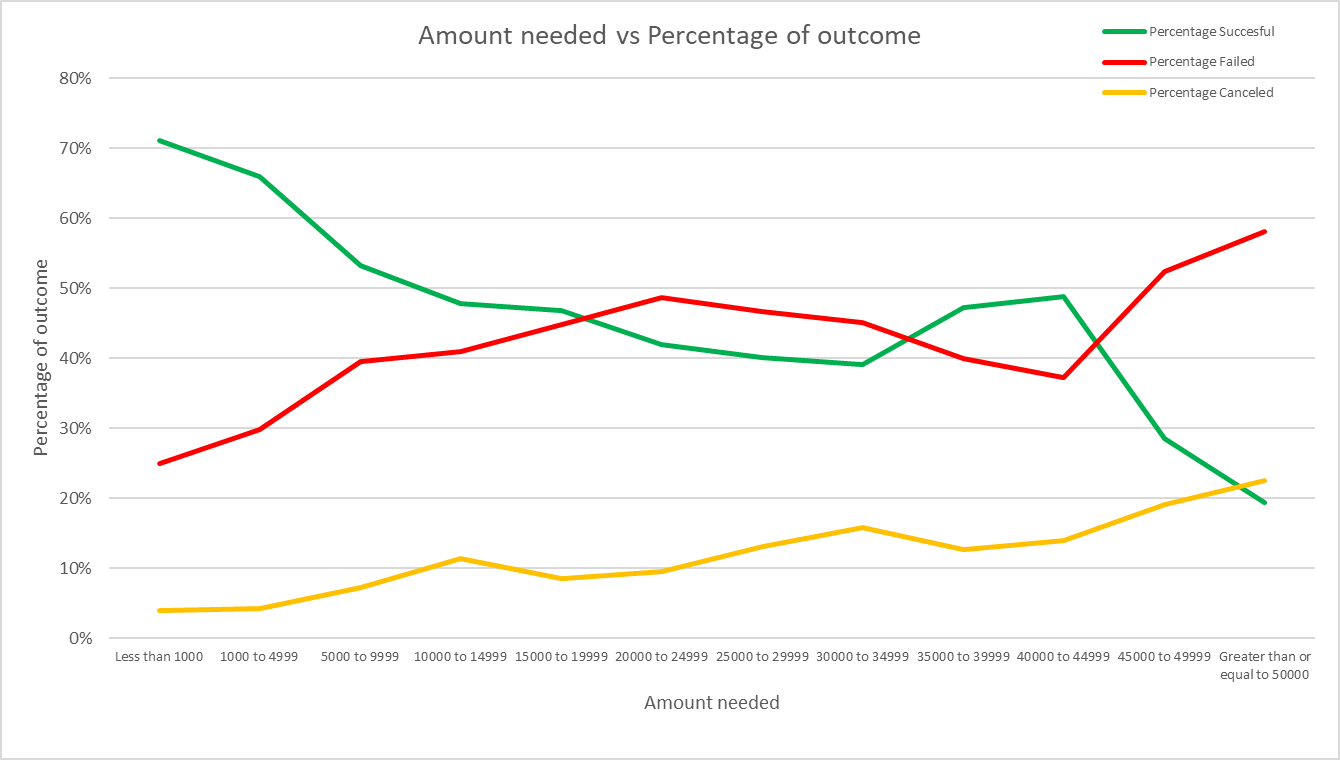
* 1. Total number of campaigns worldwide is 520 and in the United States there are 422, with the United States hoarding more than 80% of the data there is no statistical way to determine if country is a relevant variable, the sample size is too small for the rest of the countries.

1. Does the time of year determine the success of a campaign?
   1. To determine seasonality in the data we need to compare every year to each other.
   2. 2009 was not a full year and I took it out of the comparison.
   3. 2010 only has 65 campaigns, besides that, there are months with no occurrences, hence, it is irrelevant for the analysis.
   4. 2011 even though it does have months with no occurrences it totals 171 campaigns which I will consider the first stable year and we will begin our analysis from there.

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 2011 | 2012 | 2013 |
|  |  |  |
| 2014 | 2015 | 2016 |

* 1. Looking at the previous graphs it seems there is no seasonality in the data but there’s a consistent increase in successful campaigns in April every year, this is consistent if we plot a graph considering every year available.
  2. 2017 is not a full year and does not include April either, so I am taking it out as well from the comparison.
  3. I conclude that time of the year should not be considered as an important factor for launching a funding campaign but, if choosing the time of the year does not affect anything in the project (costs, amount needed to rise, opportunity, etc.) it will slightly increase the chances of success launching around April.

1. Does the amount needed affects the chances of success?
   1. Plotting amount needed vs percentage of outcome.



* 1. The graph clearly shows that as the amount needed for funding the campaign increases, the chances of the campaign being successful decreases, which can seem like an obvious conclusion.
  2. What is interesting about this analysis is the points in which the tendencies cross: between $15,000 and $40,000 the chances of failing a campaign are significantly larger than succeeding, but between $40,000 and $50,000 the plots cross over again and the success percentage increases, giving us a far better chance at succeeding. Beyond $50,000 the success rate takes a plunge and at that point I would probably consider alternative forms of funding.

Dataset analysis

The dataset provided proved to be consistent and stable since I was able to find some tendencies in the data, the sample size was enough since it contained every record available during the timestamp of the data, that being said, it would have been interesting to have more data about cancelled campaigns, specifically the reasons behind it, having this information could lead to decisions that increase the chances of success.

As limitations to the data provided, I would have considered trimming down the first two years of Kickstarter, mostly because it can be misleading since it was a startup, and the data was not stable during those first two years.

Another limitation to the data is age, the most recent record goes back three years, which is a long time specially for internet-based companies, besides that, looking a bit about the story about Kickstarter 2020 was their first tough year and it would have been interesting to see how the data behaved and if the tendencies continued during that time.

Lastly, even though it is not a limitation of the data, it would have been interesting to compare the results of the analysis with the same analysis done on the data of one of kickstarter closest competitors, GoFundMe, this way we could have determined if the tendencies are industry wide or just particular to Kickstarter.

Further Analysis Suggested

As complimentary analysis I would like to focus on the pledger. Do previous campaigns by the same pledger influence the outcome of a new one? His/her presence in complimentary social media influence the number of backers behind a project?

Even though the analysis done to Kickstarter data provides very valuable insight I would also like to move forward in more complex calculations aimed to determine the correlation each variable has in determining a campaign success.

Mean/median, variance?

Usually speaking median is a better measure for discrete data, which is the case for this dataset, we cannot have 0.5 backers, this gives me the first clue that median will be a better statistic for this dataset. Now, if we look at our statistics table for the number of backers per project.



We can clearly see that mean and median differ by a lot, which is a clear indicator that the distribution of our data is skewed, another indicator, this time a very strong one, that median will be a better statistic, now let us look at the distribution of our data:

|  |  |
| --- | --- |
|  |  |

This just confirms in a visual form, that we do have a skewed distribution of values, proving, now for sure, that median will be a better statistic for this dataset.

Variance, being a statistic that relies heavily on the mean value of a distribution will not be a valuable statistic either, although, the number of backers does apparently influence the success of a campaign. To evaluate that, we will need to take out the outliers in the data in order to have more reliable measures of central tendency.





Looking at the revised data we can now safely assume that any campaign that manages to get at least 6 backers will greatly increase their chance of success since there are not failed campaigns (except outliers) above 6 backers.

Bibliography:

https://en.wikipedia.org/wiki/Kickstarter