# **General observation:**

Our database (4,113 projects) would be not representative of the population (300,000 projects), particularly because the background stablishes that *“only a third have made it through the founding process with a positive outcome”*. Nevertheless, our worksheet suggests a different rate of successful: 53% of total projects. Making questionable each inference based on this dataset.

Also, at first glad, you can see several big outliers.

However, let’s suppose that this database is representative and we are able to make inferences regarding to all projects founding on Kickstarter.

In particular and as we can see below, we are able to create a strategy considering the visualizations of this activity, following these points:

* The relative participation of every project by category.
* Probability of success for each category and sub-category.
* Geographic location of the successful projects by volume.

1. **Given the provided data, what are three conclusions we can draw about Kickstarter campaigns?**

At first glance and based on the visualizations created during this activity, the main conclusion:

* 1. **The Pivot table 1 (Categories)** shows several patterns, in particular the proportion of successful projects are concentrated in categories music (70% of success), theater (60.2%), film&video (57.7%), photography (46.8%) and games (36.4%). In contrast, journalism, food, games and publishing categories seem to be the worst projects in terms of funding achieved.
  2. **The Pivot table 2 (Sub-categories)** let us see another interesting fact about the successful categories: music, theater, film&video, photography and games:
* Music projects are particularly successful in the following subcategories Rock, Classic Music, Electronic Music, Pop and Metal at 100% (it means that the probability of success is 1), and indie rock at least the 87.5% of the projects get funds required as the goal (.87% of probability of success).
* Theater projects has less probability of success, their categories show only 0.66, and less of .5 are successful in Musical sub-category, as well as Spaces.
* Film&video in sub-categories documentary, shorts and television are completely successful even with several projects over financed.
* Regarding to Photography, the sub-categories are not 100% successful, the unique category is photobooks with (0.6 of probability).
* Games is the other category with more successful projects. Only one of three sub-categories has a probability of 1 to be successful: Tabletop games.

Having in mind the trends below, we are able to consider a suggestion to the client to prioritize categories and subcategories of projects according to the relevant number of projects successful based on a classic definition of probability in order to promote more of them instead to the other projects with less probability of success:

|  |  |  |
| --- | --- | --- |
| **Categories / Sub-categories** | **Share** | **Probability of success** |
| **Music** | 17% | 0.77 all category |
| Rock, Classic Music, Electronic Music, Pop and Metal | 1 each sub-category |
| Indie Rock | 0.87 |
| **Film&Video** | 13% | 0.57 all category |
| Documentary, shorts and television | 1 each sub category |
| Theater | 34% | 0.6 all category |
| Musical and Spaces | Less than 0.5 |
| **Technology** | 15% | 0.35 all category |
| Hardware | 1 in this sub-category |
| **Publishing** | 6% | 0.34 all category |
| Nonfiction and radio&podcast | 1 in each sub category |
| **Games** | 5% | 0.36 in all category |
| Tabletop games | 1 in this sub-category |

As we can see, Kickstarter should be focused on projects with high probability of successful (particularly if they have a probability of success of 1 most of the cases). Champions in terms of success such as:

* Music in sub-categories Rock, Classic, Electronic, Prop and Metal.
* Film&Video in sub-categories Documentary, Shorts and Television
* Technology in sub-category hardware
* Publishing in sub-category nonfiction and radio&podcast
* Games in sub-category tabletop games

This pattern make feasible to stablish a prioritization criterion, however it could be more specific and robust as we see in the answer of question 3.

* 1. Based on the **Pivot table 3 (Sub-categories)**, the company also should put attention in the time. Successful years where from 2010 to 2013 when each month the proportion of projects with a pledge bigger than the goal funding are 80% or so per month. The successful projects where almost the same proportion of the failed projects each year until March 2017 when the number of projects fall considerably.

Those fluctuations could be explained by an economic crisis or other drivers such as regulatory reforms with limits in the amount allowed for every backer, and so on[[1]](#footnote-1).

The other conclusion we are able to get from this visualization is about the countries with more successful projects:

* United Kingdom (GB) with 366 successful projects (61% of the total projects)
* United States (US) with 1651 successful projects (55% of the total projects)

Only this measure suggest that Kickstarter should have focused his efforts on some locations, based on the rate of successful applied to the overall projects. This is another prioritization criterion that we should take in to account so as to build a strategy for maximize successful projects.

1. **What are some limitations of this dataset?**

If we propose a broader definition of successful project, not only the successful rate we have used previously as the main criteria (goal/pledges), we need a more robust dataset.

For instance, we should think in the strategy used by each project to connect potential backers considering data about **strategies of communication**: social networks, videos, blogs, etc. Patterns around this information could complement our strategy to stablish a communication standard for every project in order to increase probability of success.

Another key word that we should measure is the **project’s quality**, an indirect way to measure this dimension is based on the time spent to achieve the funding goal. This could give us a valuable pattern that let us define other prioritization criteria.

We also need information to **build backers profiles** in order to focalize the campaign, based on age, location, interests, and so on.

The time series is short, we only can know what happen from 2009 to 1Q2017.

1. **What are some other possible tables and/or graphs that we could create?**

I would start with a data exploratory analysis using basic statistics, such as: proportions regarding with important variables, measures of central tendency, standar deviation, in order to get information about data distribution and outliers.

Proportions regarding with state categories



Histograms related with percent funded, average donation, pledge, backers count, total founding, *etc*.

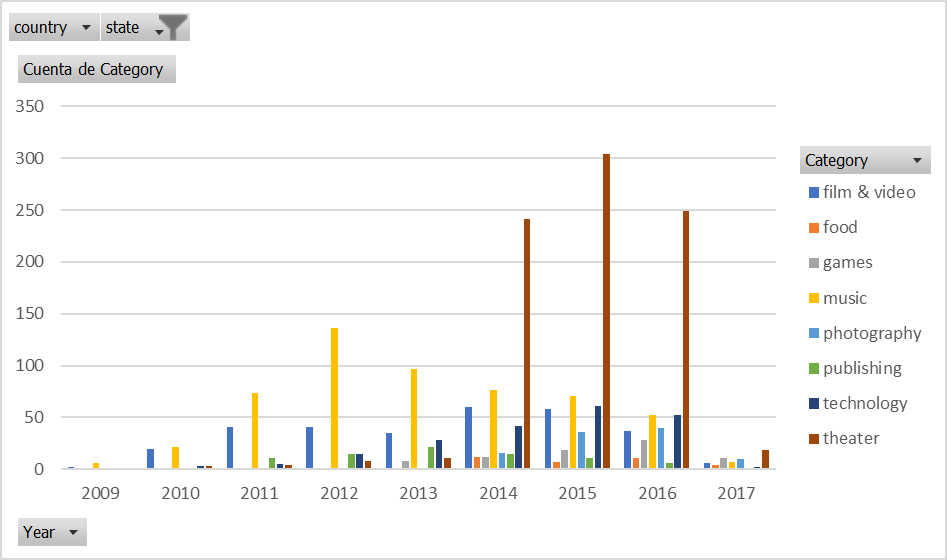
|  |  |
| --- | --- |
|  |  |
|  |  |

Another sort of visualizations are related with the identification of outliers, for instance box plots and definition of quantiles. Unfortunately I didn’t control outliers in the dataset and they must be an issue if we want an inference exercise:

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

This visualizations should be complementary and has to consider basic statistics such as average, sd, max and minimum value, mode, median, etc.

Several visualizations could be for instance:



Where we can see theater is coming the most important category of successful projects since 2014 until 2016. What changes in 2017 is that we only have three months.

Music had his best year in 2012 and is decreasing slow, as well as Film&video, photography, and publishing were very constant until 2017.

Also, we can find several interesting correlations. For instance, there is a negative statistical relationship between the average of number of days of campaign in the projects and the pledge achieved:

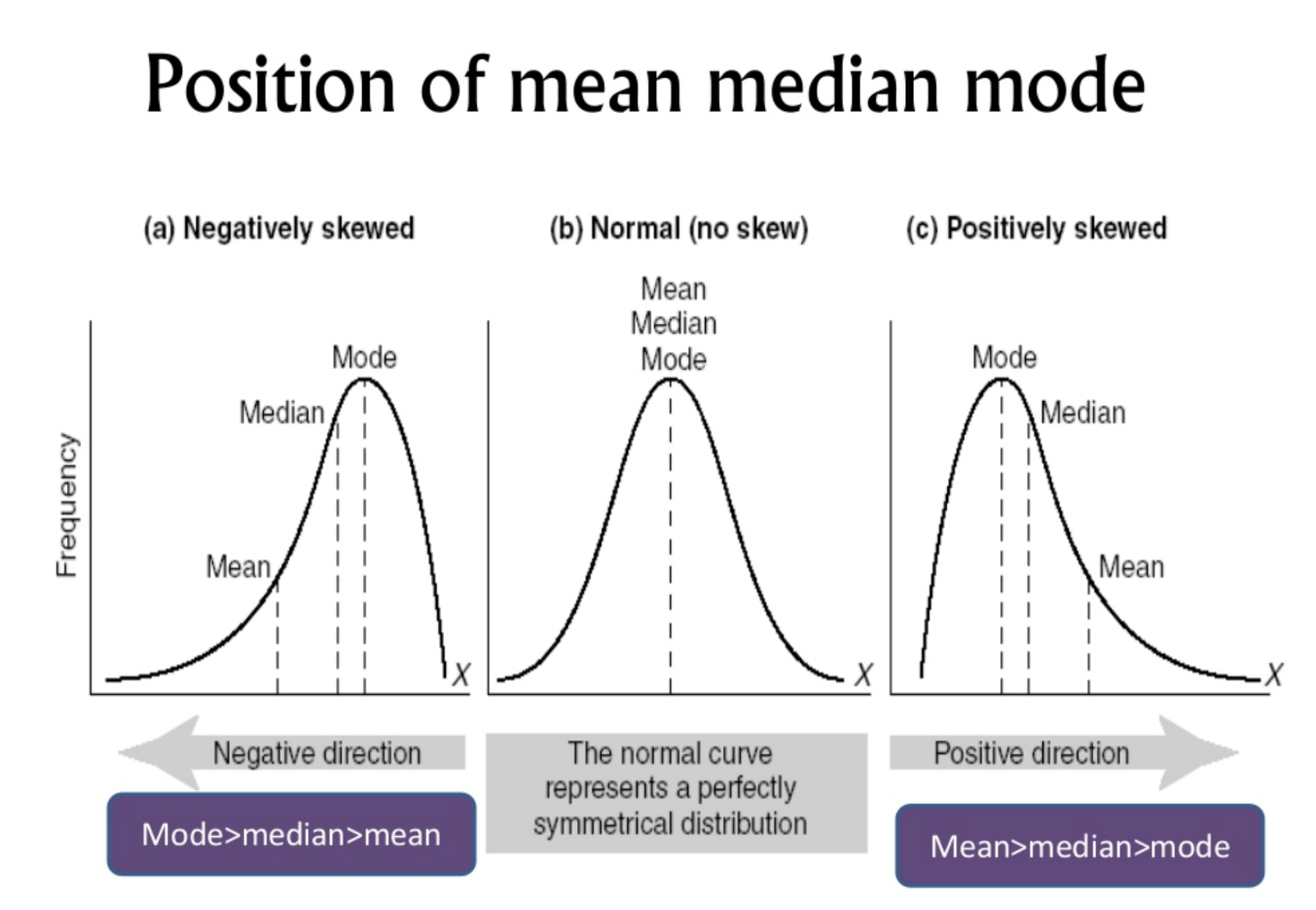
Also, this relationship is seen with the number of backers:

# **BONUS**

# **BONUS statistical analysis**

|  |  |  |
| --- | --- | --- |
| Summary | | |
| **Statistics** | **successful** | **failed** |
| Mean of number of backers | 194.4 | 17.7 |
| Median number of backers | 62.0 | 4.0 |
| Mode | 27.0 | 0.0 |
| Minimum number of backers | 1.0 | 0.0 |
| Maximum | 26457.0 | 1293.0 |
| Variance | 713167.4 | 3775.7 |
| Standard deviation | 844.5 | 61.4 |

As we see during the first week in the boot camp, when mean is greater than median and median is greater than mode we have a positively skewed (thanks Héctor for the following image).



This is confirmed also to all population such as in the following box plot (Thanks JM for that):

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

Because of that, median summarizes the data more meaningfully.

Variability is greater in successful () than in failed projects (). The number of outliers could explain this, because the maximum value in successful cases is 426 times the median and in the case of failed projects, the maximum is 323 times the median. Almost 18% of the projects do not get any backer, which is the mode in the distribution.

1. For instance: <https://www.fca.org.uk/publication/thematic-reviews/crowdfunding-review.pdf> [↑](#footnote-ref-1)