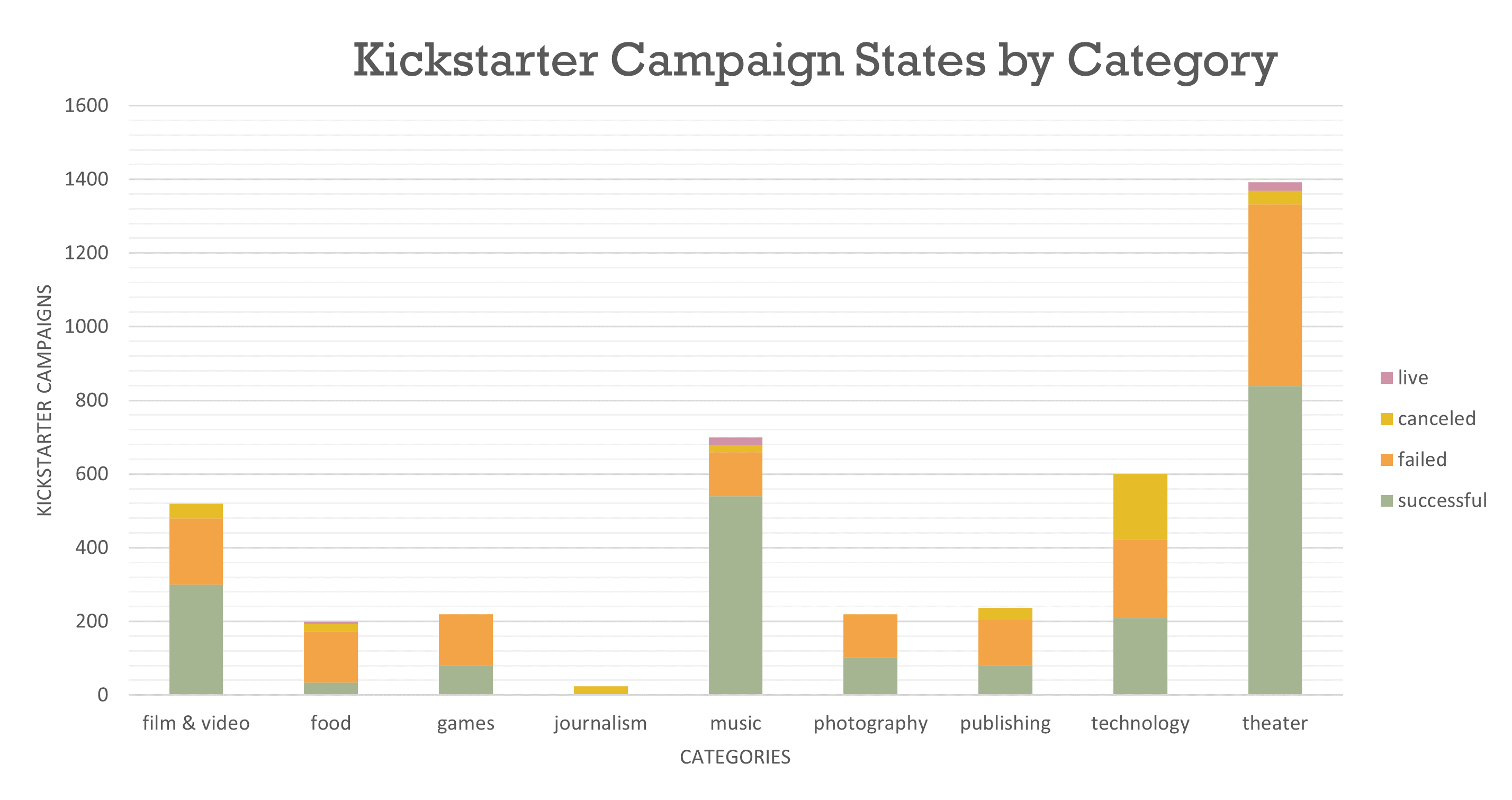
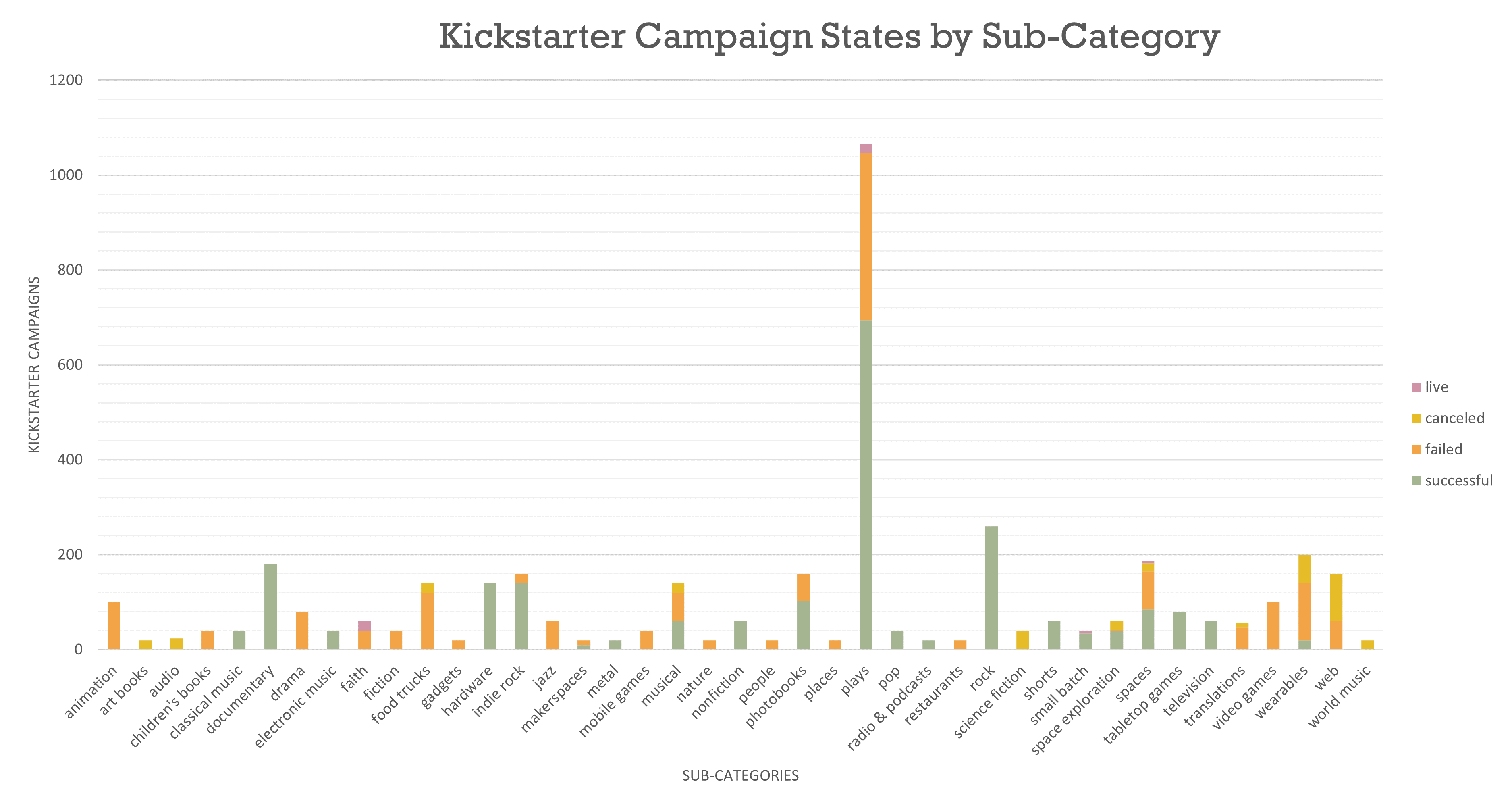
# Excel Homework: Kickstart My Chart

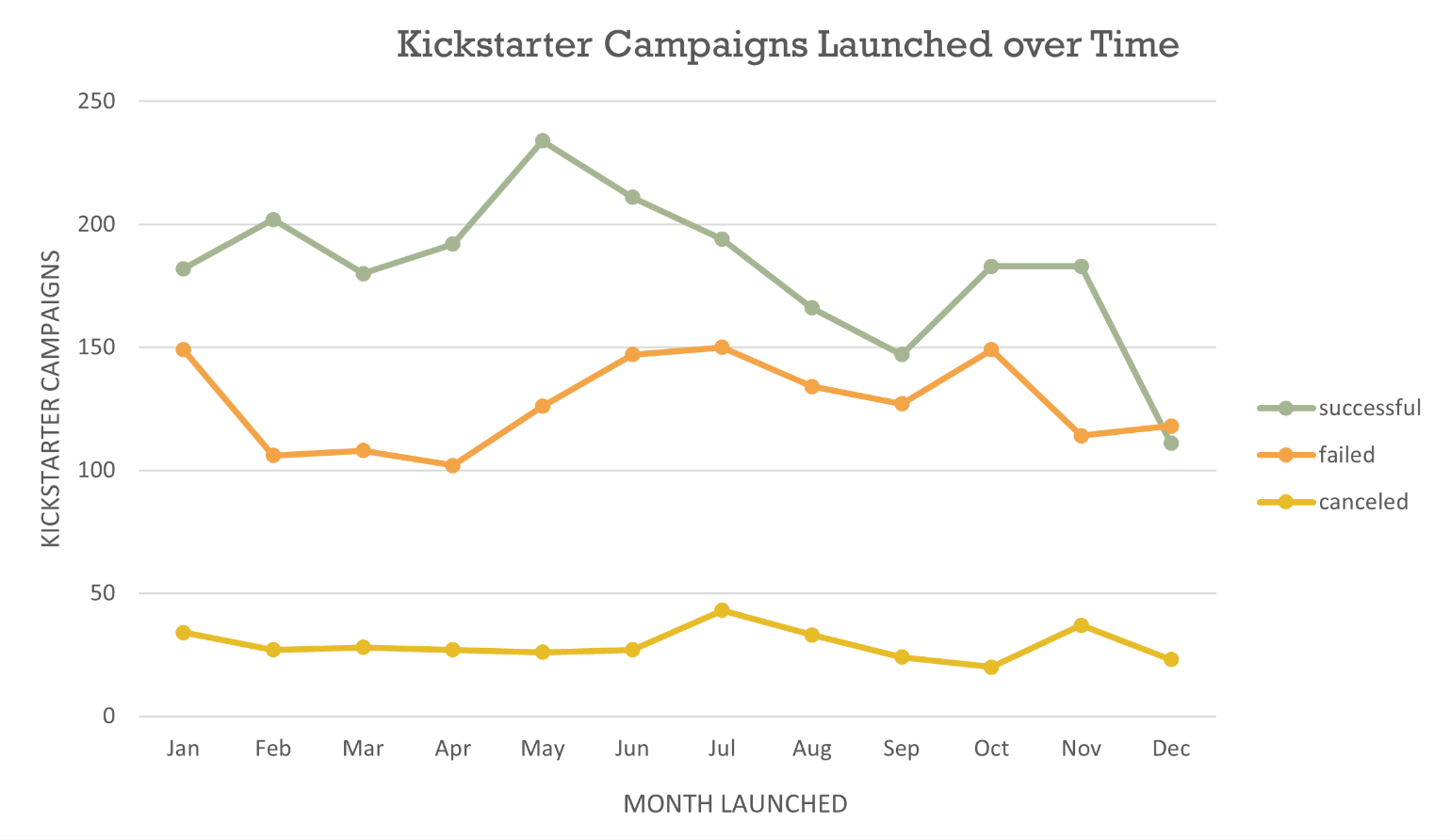
Student: Welan Chu | Monash University Data Analytics Bootcamp | Nov. 2020 PT | Week 01

## Kickstarter Data Analysis

The charts below represent a set of Kickstarter campaign data and their states ranging from successfully backed, failed, cancelled and live.

* Chart 01.*

*Chart 02.*



*Chart 03.*

A number of conclusions may be drawn from the charts, including:

* Performance-related arts and entertainment categories appeared to have the most campaign launches and better success than other categories, evidenced by *theatre*, *music* and *film & television* categories containing high campaign numbers (around 1,400, 700 and 500 respectively), compared with other categories, including other arts and entertainment categories such as *games*, *photography* and *publishing*. The most commonly launched (and most successful, on aggregate) campaign appears to be *theatre*, by a significant margin, with the least common (and least successful, on aggregate) being *journalism*. Almost all categories contain successes and failures, except journalism, where all such campaigns in the dataset were cancelled.
* Out of the successful sub-categories, *plays* were substantially more common than any other sub-category, approximately 3 times more campaigns launched than the next most common sub-category (*rock*). While plays have the most successful campaigns on aggregate, rock is the most successful sub-category without failures or cancellations. Conversely, *animation* is the most failed sub-category without any successful or cancelled campaigns. This is particularly interesting given that film & television is the third-most common campaign category, indicating animation’s unpopularity or otherwise unsuccessful nature on the Kickstarter platform.
* Campaign launches result in success more often than failure in every month except December. This could be due to several factors that are not immediately evident in the data, such as the quality and type/category of campaigns launched in December and the behaviour and possible influence of Christmas holidays on both backers and creators, including time, money and level of engagement, among other factors. The most successful months appear to be May and June, with spikes in February and October and November, while the most unsuccessful months are January, July and October.

Limitations

Some limitations of the dataset include:

* No clear demographics of backers: there is no way of knowing who the campaigns appeal to, why certain campaigns are successful and others are failures and who in the community does the data represent and describe. While the data may suggest that its backers predominantly like and support theatre campaigns, it is unclear who these people are, their age, gender, location, disposable income and appetite for risk, amongst other characteristics.
* Difficult to determine the riskiness of the campaigns based solely on the available data. Some categories may have more ambitious or riskier campaigns than others, and backers respond according to level of risk, thus people who may have an interest in several categories may not ultimately invest in every category of their interest. Some categories may have lower goals than other campaign categories, or more creators and campaigns, therefore comparing a smaller data set of campaigns with similar goals in different categories may reveal a different perspective.
* No indication of how many campaigns a single backer has backed: some categories may have inflated numbers due to repeat backers or a dedicated community supporting one another, while others may have none or rely entirely on individual, one-off backers.
* Difficult to determine the quality of the campaigns between categories based only on the dataset: some categories may have better production and marketing that drives the competition within the category, while other categories may enjoy a high success rate (or suffer a high failure rate) regardless of the quality of the pitch. The inability to determine quality makes it difficult to learn from the results and determine whether a high-quality campaign makes a difference, and even what the difference is, objectively, between a high-quality and low-quality campaign pitch.

Other tables and/or graphs to represent the data

* A scatter graph could be used to plot each campaign individually and compare it to the others. The data could be filtered in numerous ways depending on the question, such as comparisons within categories, or comparisons between each category. This visualisation can help determine whether there are any positive or negative relationships between two variables and whether there are any trends and the nature of that trend (e.g. linear, exponential, logarithmic et cetera). It can also expose outliers, either where campaigns received no funding, or where the pledges exceeded the goals and by what magnitude.
* Given the data includes geographic information regarding the campaign creator’s location, a map chart could be used to represent the data, such as illustrating the number of campaigns in locations throughout the world and their geographic spread. This would allow for visual analysis of active and inactive locations, and could be filtered by category to show whether any location has a greater propensity of launching one type of campaign over another, among other possible analyses.

## Bonus

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| successful |  |  | failed |  |
| mean | 194.4252 |  | mean | 17.7098 |
| median | 62 |  | median | 4 |
| minimum | 1 |  | minimum | 0 |
| maximum | 26457 |  | maximum | 1293 |
| variance | 713167.4 |  | variance | 3775.689 |
| standard deviation | 844.2991 |  | standard deviation | 61.42656 |

Based on the above table, I think the median summarises the data more meaningfully in both successful and failed cases. In both cases, but particularly for successful campaigns, the high maximum skews the mean to be far beyond what the median is, resulting in less than half of all successful and failed campaigns being equal or close to the mean.

Also, the variance and standard deviation being very high in successful campaigns, and higher than the variance and standard deviation in failed campaigns, makes sense for a number of reasons. The substantial difference between the mean and the median in the successful campaigns, and high maximum, indicates that the variance is likely to be high. Since variance is a measure of how far values in the dataset are from the mean and that more than half the numbers are significantly lower than the mean (62 median compared to ~194 mean), it makes sense for the variance to be a high value. Contextually, this shows that successful campaigns can have a wide-ranging number of backers, from one to tens of thousands, depending on various factors such as the popularity of the campaign and the goal amount. Also, given that there is no upper limit to the success of a campaign (i.e., a campaign can be funded +200% its goal), but there is inherently an upper limit for a campaign to fail (that is, it must be less than 100% of the campaign’s goal, otherwise it would be a successful campaign), it supports the evidence in the data that successful campaigns can have larger numbers of backers than failed campaigns overall, particularly the extremely successful campaigns. Having no upper limit allows for successful campaigns to have ever-increasing variance and standard deviations, while failed campaigns are a sign of insufficient support, meaning the variance and standard deviation will always be smaller than successful campaigns.