Which qualities, if any, can help identify a successful crowdfunding campaign?

Is the category or subcategory of a crowdfunding campaign a reliable predictor for the successful outcome of the campaign? To answer this question, we need to look at the spread of success across all categories of campaigns. In Figure 1, we can see that the number of successful and failed campaigns mapped across their respective subcategories. While the '*plays*' subcategory has a significant amount of successful campaigns (187) compared to the amount of failed campaigns (132), this could be the result of being the most popular category of crowdfunding campaign (34% of the total amount of campaigns in the data set). For the remainder of categories, there are no significant differences between the amount of successful and failed campaigns.

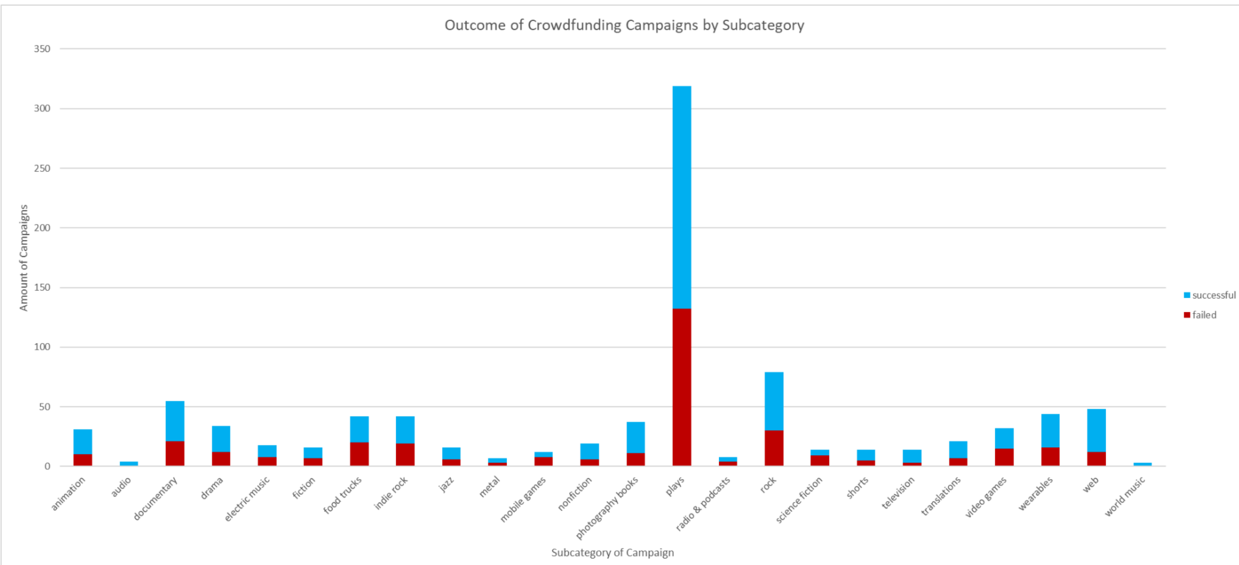

Figure 1: Outcomes of 949 campaigns organized by their subcategory. Successful campaigns are indicated by the blue bars and failed campaigns are indicated by the red bars.

Would the month that a crowdfunding campaign is created indicate whether or not the campaign will be successful? In Figure 2, we see the outcomes of 986 crowdfunding campaigns mapped across the month that they were created. There is a noticeable difference beginning in June with 28 failed campaigns and 55 (difference of 27) successful campaigns which increases through July with 31 failed campaigns and 58 successful campaigns (difference of 27). Another noticeable spike can be found in September with 23 failed campaigns and 45 successful campaigns (difference of 22). Then, there is a noticeable dip in August with 35 failed campaigns and 41 successful campaigns (difference of 6). For the remainder of the months the difference between the number of campaigns remains at an average of 14.8. Within this data set, it would appear that the month of a campaigns' creation is a possible indicator of its' chance to succeed.

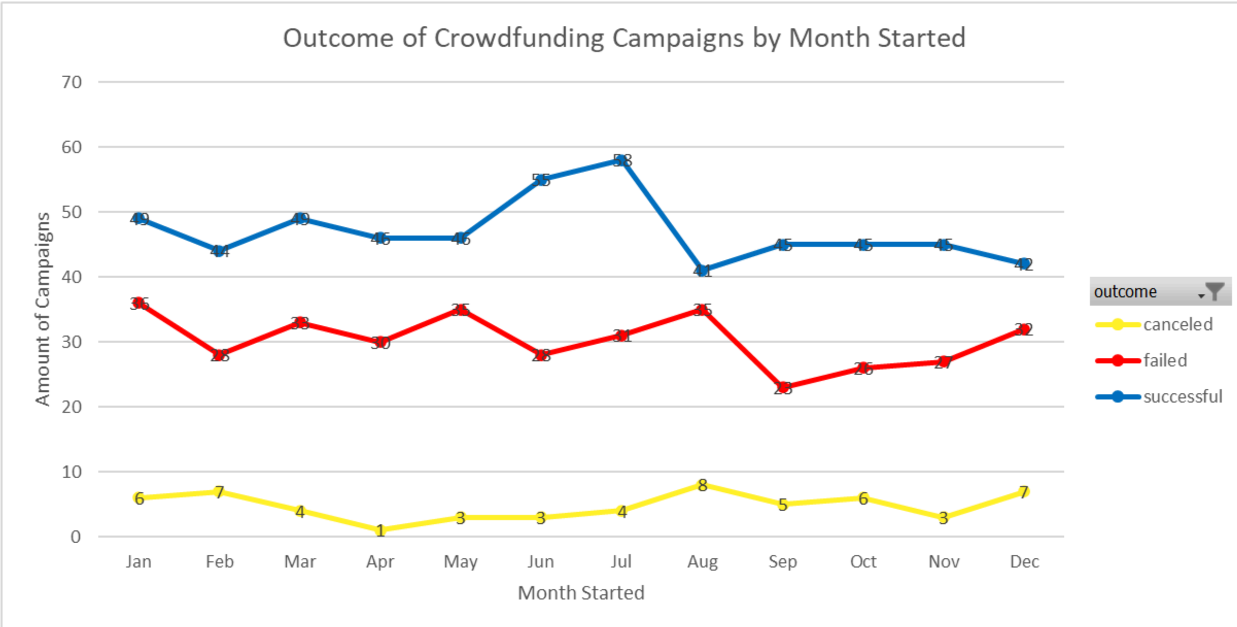

Figure 2: Outcomes of 986 campaigns charted over the month of their creation. Canceled campaigns are shown as a yellow trace, failed campaigns are shown as a red trace, and successful campaigns are shown as a blue trace.

Can the success of a crowdfunding campaign be predicted with its' funding goal? To get an answer to this question, we must look at the outcomes of campaigns in relation to their target funding goals. Figure 3 shows us the rate of success, failure, and cancellation for 961 crowdfunding campaigns mapped across their target funding goals. There are significant differences at ranges of: '1,000 to 4,999' with a success rate of 83%, '25,000 to 29,999' with a success rate of 79%, '35,000 to 39,000' with a success rate of 67%, '40,000 to 44,999' with a success rate of 79%, '45,000 to 49,999' with a success rate of 73%. In addition to these ranges, there is a success rate of 100% at ranges of '15,000 to 19,999' and '20,000 to 24,999' and '30,000 to 34,999.' Within this data set, there exists a comfortable success rate for smaller projects which wanes until the project has a funding goal of at least 15,000. Ranges after 15,000 see a great overall success rate which persists until the goal exceeds 50,000; at which point the rate begins to decline beneath 50%.

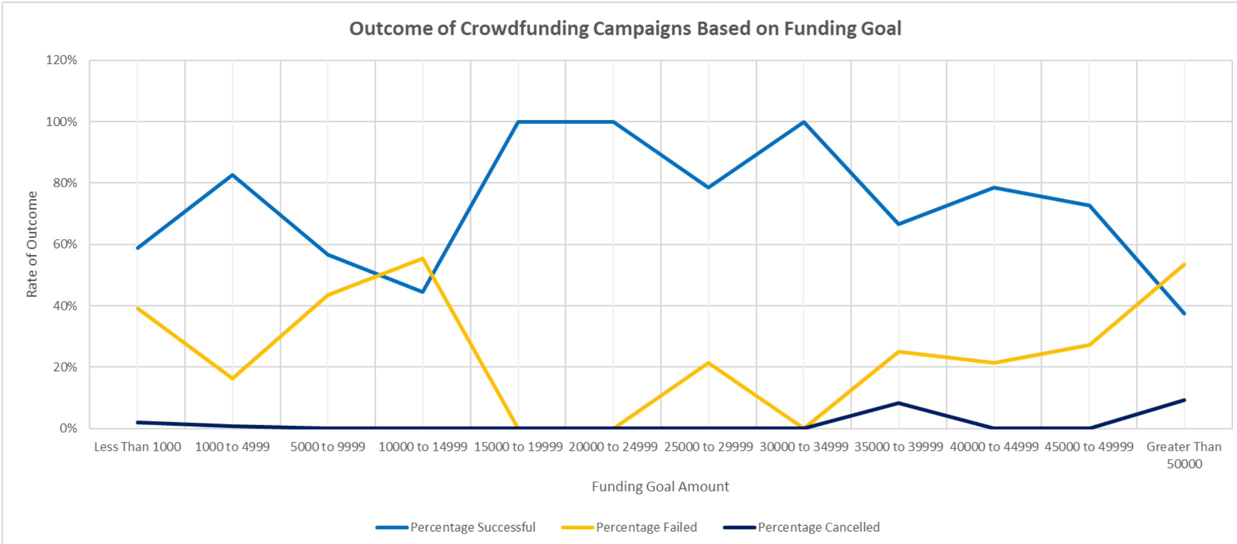

Figure 3: Rates of failure, success, and cancellation of … campaigns are charted over their target goal ranges. These ranges begin at "less than 1000," "1000 to 4999", then increase by 5000 until reaching a maximum of "greater than 50,000." Success rate is shown with a blue trace, failure rate is shown with a yellow trace, and cancellation rate is shown with a dark blue trace.

What are the limitations of this data set?

One of the limitations of our data set is the sample size relative to the amount of crowdfunding campaigns. This data set has a total of 1000 entries; however, according to Kickstarter (a well-known crowdfunding platform) their website has hosted a total of 624,775 campaigns (as of April 20 2024). With that context, the sample size of this data represents only 0.16% of all campaigns on Kickstarter alone. Kickstarter is not the only platform hosting crowdfunding projects, either, and the representation would decline further if it were to include other sites.

This sample is not representative of the world-wide community of crowdfunding. All of the data contains campaigns that originated in the United States, Canada, United Kingdom, Australia, Switzerland, Denmark, or Italy. These 7 countries cannot represent the entire world of crowdfunding; it is possible that with the inclusion of data from more countries, trends could change and different observations could be made regarding success of crowdfunding campaigns.

Another limitation of this data set is the gap that exists between the last year of collected campaign data and the current year. In the data set, there are no entries for campaigns that were created beyond the year 2020. In addition, there are a total of 2 entries for the year 2020. With this in mind, observations drawn from this data might not give reliable insights regarding outcomes of campaigns started in the current year. Trends can change fast in the current age and more up-to-date data would be beneficial for this data set.

Table 1: (right) Shows the amount of campaigns created in each year for the data set.

Year Created	Amount
2010	108
2011	103
2012	84
2013	88
2014	102
2015	105
2016	98
2017	101
2018	102
2019	107
2020	2

What additional kinds of tables or graphs could we create? What insights would they provide?

An additional table that I created is the table pictured above (Table 1, above). This table sorts the entries in the data set by the year that they were created. This table provides insight regarding the span of time the data covers. There is a significant gap between the final year of the data set (2020) and the current year. This gap shows us one of the data's limitations: there is a lot of missing time between the data set and our current time, and the data that is missing could provide us with better insights regarding the crowdfunding campaign landscape today.

Another resource that I created was a line chart using the data from Figure 1 (Outcome of Crowdfunding Campaign by Subcategory). This line chart assisted me in seeing the slope of the two series of data which I found difficult with the bar graph. With this insight, I was able to identify the similarities between the data regarding the outcomes mapped across the subcategories of the campaigns.

Given that the campaigns originate from different countries, a useful resource would be a standard conversion of the goal amount and amount pledged. Whether or not this conversion has been done is not stated within the data set; there is a portion of the data set named 'currency'. However, it is not clear if that is the currency used for that entries' goal and amount pledged. If there is no conversion in the data set, then it would be beneficial to create three new columns named 'goal adjusted,' 'pledged adjusted,' and 'average donation adjusted' to allow for a clearer comparison of all of the campaigns.

In our data set, it appears that the median is better suited to summarize the data. Both the *failed* and *successful* outcomes have large maximum values relative to their mean which would indicate that there are high outliers. For both of these series of data, the mean number of backers is a large margin higher than the median: the *failed* backer mean is 115 with a mean of 586, and the *successful* backer mean is 851 with a median of 201. These discrepancies lead us to believe that the mean has been skewed higher given the outliers that exist near the upper range of the set. So, it would follow that the median would provide us with a better representation of the data set.

Between the two outcomes *failed* and *successful* in our data set, the variance is different when counting the *backer count*: for the *failed* campaigns, the variance is 921,575, and for the *successful* campaigns, the variance is 1,603,374. From looking at the data, this makes sense: the range of the *failed* campaigns is 0 to 6080 backers where the range of the *successful* campaigns is 16 to 7295 backers. This increase of 1215 in range with an increase in the median of only 86 indicates that the outliers are more prevalent in the *successful* data series. Given this information, it would make sense that the *successful* entries have a higher variance than the *failed* entries.

In terms of whether or not this makes sense from a logical standpoint: there are a few things to consider about the data set. First, for the *failed* campaigns there is a trend of campaigns not getting a lot of backers and failing. A large number of the entries for *failed* campaigns have backer counts that are low and this makes sense: a campaign that does not have backers is most likely going to fail. These *failed* campaigns mimic each other more closely than the *successful* campaigns. A *successful* campaign is seen to have multiple ranges of goals where their success is possible; i.e., a smaller campaign can see success with less backers while larger campaigns can also see success with a larger amount of backers. Compared to the *failed* campaigns, which have a lower chance of failing when the backer count gets higher, the *successful* campaigns are going to have data with a higher spread given their chances of success with high or low amounts of backers.

References

Kickstarter. (2009, April 28). Stats. Retrieved April 21, 2020.
https://www.kickstarter.com/help/stats?ref=hello

Note: the resource below was used as a reference for formatting this report.
Scott, James G. (2016, February 1st). Data Exploration And Simple Models. James Scott.
https://jgscott.github.io/teaching/writeups/files/example_writeup1.pdf