Module 1 Report: Startup

By Karoly Burgyan

Given the provided data, what are three conclusions that we can draw about crowdfunding campaigns?

1. Theaters and plays are the most common categories and sub-categories respectively for startups.
2. As the dataset and graphs currently are, they appear to indicate that the most successful category and sub-category that has been attempted beyond just a handful of times (i.e. e.g. excluding journalism, audio, and world music) is technology and web (and wearables) respectively especially when they start in June and any of the countries evaluated with the exception of Denmark (DK) (as shown in the **Outcome By Category** graph, **Outcome By Sub-Category** graph and the **Outcome By Month** graph).
3. Setting a goal between $15000 and $25000 may have some special qualities that it produces for success but those qualities, if they exist, can’t be determined by the dataset and graphs as they are.
4. Extra: By looking at the **Outcome In Percentage By Goal** graph (and we assume “live” is not a final outcome) cancelation is the least likely final outcome and success is the most likely outcome for startups.

What are some limitations of this dataset?

1. As the dataset currently stands, it doesn’t show **explicitly** which category has the greatest percentage of success compared to other categories (unless one is going to claim that “journalism” is the most successful category on such a small number of attempts (4 attempts and 4 successes)).
2. As the dataset currently stands, it doesn’t show **explicitly** which sub-category has the greatest percentage of success compared to other sub-categories (unless one is going to claim that “audio” is the most successful sub-category on such a small number of attempts (4 attempts and 4 successes)).
3. As the dataset currently stands, it doesn’t show **explicitly** which month has the greatest percentage of success compared to other months (“July”, “September”, and “June” look like they have the highest chance of success from the graph but it’s not clear without further processing).
4. As the dataset currently stands, we can’t separate noise from signal which prevents us from claiming things like “journalism” or “audio” produce the most successful startups.

What are some other possible tables and/or graphs that we could create, and what additional value would they provide?

1. Add a column that calculates percent success per final-outcome-attempt for all categories (i.e. number of success/(number of success + number of fails + number of cancelations)). Show these percentages per category in a pie-chart. This would allow you to see the most successful category (excluding the categories that were rarely attempted such as “journalism”).
2. Add a column that calculates percent success per final-outcome-attempt for all sub-categories (i.e. number of success/(number of success + number of fails + number of cancelations)). Show these percentages per sub-category in a pie-chart. This would allow you to see the most successful sub-category (excluding the sub-categories that were rarely attempted such as “audio”).
3. Calculate p-values for the categories, sub-categories, and months so that you can determine what differences between categories, sub-categories, or months is “highly significant”, “significant”, and “not significant” (as a statistician might describe in accordance with p-value).