WILLIAM VANN

HOMEWORK #1

**Narrative Report**

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

First conclusion: 3 parent categories - across all countries and also the U.S. alone – are of special interest because they lead the dataset dramatically in **most successful projects** and **most total projects**. These top 3 parent categories are **theater, film & video,** and **music**; these are all **arts-related**.

Thus, I conclude that **theater, film & video,** and **music** are responsible for more overall projects, and more successful projects, than any other parent categories. Further investigation of *why* this is so is warranted.

(Source: “**Outcome Count by Parent Category**” chart, “Parent Category” sheet)

Theater, Film & Video, Music

Second conclusion: The sub-category with the **greatest total crowdfunding projects** and the **greatest number of successful crowdfunding projects** - across all countries and also the U.S. alone - is **plays** (a sub-category under the parent category of **theater**). Again, *why* plays have been more represented and more successful deserves to be explored further.

(Source: “**Outcome Count by Sub-Category”** chart, “Sub-Category” sheet)

Plays

Third conclusion: A close examination of the state (**canceled, failed, successful**) of the projects over time reveals a substantial uptick in successful projects **between May and July**. This trend held true for all parent categories (successful: May = 46, June = 55, July = 58), as well as for the three aforementioned parent categories of interest: **theater, film & video,** and **music** (successful: May = 27, June = 37, July = 41). All of the countries represented in the dataset except for Australia (AU) are in the Northern Hemisphere, hence end of spring to mid-summer may therefore be the most fortuitous time of the year for Northern Hemisphere crowdfunding campaigns.

(Source: “**State over Time”** chart, “State ” sheet)

May – June – July

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

One limitation of the dataset is that it may not be a randomly selected, representative sample of crowdfunding projects worldwide. For starters, its relatively small size (1000) is troubling, as a quick look at <https://www.kickstarter.com/help/stats> reveals total launched projects at the time of this writing to be 561,534. I was unable to discover a similar figure for IndieGoGo, but the point remains that 1000 may be too small of a sample. Moreover, the countries represented in the dataset may skew results. Of the countries included (AU-Australia, CA-Canada, CH-Switzerland, DK-Denmark, GB-Great Britain, IT-Italy, US-United States), 1 is in Asia, 2 are in North America, and 4 are in Europe. It is curious why China and India are not included, while much less populous countries like Switzerland and Denmark are included, for example. Finally, 763 out of the 1000 projects are from the U.S., which would seem to *overrepresent* the U.S. vis-à-vis other countries.

A second limitation is that some parent categories only have 1 sub-category (e.g., “journalism/audio”), while other parent categories have several sub-categories (e.g., “music/electric music”, “music/indie rock”, “music/jazz”, etc.). It’s unclear why parent categories like music have been sub-categorized much more than those like journalism. (The kickstarter website shows additional sub-categories for “journalism” such as “print”, “photo”, “web”.) This mismatch can lead to distortions in the visualizations based on parent category and sub-categories.

Third, the dataset contains columns with little usefulness to the analyst, at least without further explication or granularity. The fields in question are: **blurb, staff\_pick,** and **spotlight**. Blurb is puzzling and cryptic, and bears no obvious relation to the parent category and sub-category for many projects (e.g., id=943, “food/food trucks”, blurb=”Synchronized fault-tolerant algorithm”). Perhaps a more verbose campaign description could be useful with some natural language processing algorithms. **staff\_pick** and **spotlight** are intriguing, but we are left guessing how their boolean values were arrived at. Perhaps a numerical rating on a scale for staff\_pick would be more illuminating (e.g., “4 out of 5 stars”), as well as some criteria for the staff picks. More info on what “spotlight” refers to would be helpful as well. (I assume it’s related to showcasing the campaign on the crowdfunding website, but this is a speculation. Maybe some deeper website analytics related to such showcasing – like “click throughs” or “time spent on the project page” – would lead to deeper insights.

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

Examining the relationship between **spotlight** and **outcome** would be of interest to me. The additional value would be to see if there is a significant connection between being “showcased” on the crowdfunding website and how the campaign turns out. For instance, if it were discovered that spotlight drives outcome, then perhaps a savvy crowdfunder would pay the extra charge for a “spotlight” on the website.

Another table or graph that drills down on the “Outcomes Based on Goal” graph from the “Bonus” sheet would be interesting. There appears to be a dramatic uptick in successful outcomes from “15000 to 19999” through “30000 to 34999”. Does this trend hold for all parent categories and sub-categories, or just some? It would be fascinating to see which categories and sub-categories accounted for *failures* during this funding goal “sweet spot”, if that is in fact what it is.

**Statistical Analysis**

1. *Use your data to determine whether the mean or the median summarizes the data more meaningfully.*

I attempted to create 2 histograms with bins, one for successful backers\_count and one for failed backers\_count. Since neither distribution appears to be normal (Gaussian), I conclude that the mean isn’t a good measure of central tendency for these datasets. In other words, the distributions are heavily skewed toward the leftmost (lowest value) bin. The median for each seems to capture the center of the datasets better, as the value of the median is within this leftmost (lowest value) bin.

1. *Use your data to determine if there is more variability with successful or unsuccessful campaigns. Does this make sense? Why or why not?*

If we use variance, and hence standard deviation, of the **backers\_count** as our measure of variability, then the successful campaigns are more variable than the failed campaigns (**stdev** is 1266 compared to 960, respectively). I think this makes sense because there are many more rows of successful backers\_count than rows of failed backers\_count (565 vs. 364, respectively). Simply put, there are more values to be dispersed in the successful backers\_count, and hence more values to populate more bins.