**Part 1**

**Create a report in Microsoft Word, and answer the following questions:**

* **Given the provided data, what are three conclusions that we can draw about crowdfunding campaigns?**
* **What are some limitations of this dataset?**
* **What are some other possible tables and/or graphs that we could create, and what additional value would they provide?**

1. Conclusions

* From this data, it is evident that failed campaigns are frequent. Depending on the category of the campaign, roughly half to three-fourths as many campaigns will fail as succeed. The precent of failed campaigns to successful campaigns appears to be fairly constant for the three categories of campaigns that represent about 70 percent of the total. In this group, 38 percent of the largest category (theater), 34 percent of the second largest category (film and video), and 37 percent of the third largest category (music) failed.

With one exception, that might be attributable to the small sample size, the smaller campaigns are similar in number (between 42 and 96), but represent distinct groups of failure rates, roughly one-half, one-third, or one-quarter of the total. Approximately 48 percent of games campaigns, and 43 percent of food campaigns, fail. Other categories, which might seem to have more cachet, such as technology, are closer to the one-third rate of the most common campaigns. Similarly, 38 percent of publishing campaigns and 29 percent of technology campaigns, fail.

Photography and journalism have the highest success rates. Of the 42 total photography campaigns, 24 percent fail. Journalism is the only category where all of its campaigns succeeded. This could be because of the extremely small sample size of four total campaigns over more than ten years, which is not really large enough to support many conclusions about trends. There also could be some confusion or overlap in categories between photography, journalism, and publishing.

While it might seem natural that if a campaign has raised little to none of its goal, the campaign would be canceled, this dataset shows that vanishingly few campaigns are canceled. Only 57 of the 1000 campaigns were cancelled, and of those, 44 were in the three categories that have far and away the largest numbers of campaigns. It could be that sponsors view the down-to-the-wire period as likely to bring in a large uptick in donations.

* We also can conclude that the type of campaign is key to both the number of campaigns that are run and the percentage of campaigns that are successful. Campaigns involving entertainment, particularly theater, are the most frequent, and, as discussed, also have a higher percentage of successful campaigns to failed campaigns than other categories like food and games. Three categories of campaigns -- theater, film and video, and music -- represent approximately 70% of the total (697 of 1000).

Theater has far and away the largest number of campaigns, with 344, or about one-third of the total. It has approximately fifty percent more campaigns than the next largest. The category of film and video, and the category of music, are the next most frequent, and roughly equivalent in number, at 178 and 175, respectively. It is not clear from this data whether this is because of big-name artists or due to many individuals who want to become that and are seeking support to begin their journeys, but most likely it is the latter.

This also could be because of the nature of the intended audience for the campaign. Crowdfunding tends to be something that is easy to set up at little or no cost, so an organization like a local theater which runs on a shoestring budget and can't raise ticket prices too high would be drawn to this type of fundraising. People who tend to go to plays at a local theater would constitute a built-in group of interested contributors, and would be easy for the theater to reach and to send a link to the fundraising site. Members of the community who wanted to support the theater might be able to obtain contact lists or might know other members, and word of mouth would spread the news about the campaign readily.

It also could be that these sorts of differences are due to sampling error or the size of the dataset, as discussed in part 2.

* In addition, the time of year that the campaign was started, and the particular year of the campaign, are very significant with respect to whether the campaign will be successful. For instance, while it varies slightly from year to year, campaigns that start in late spring or early summer have the highest rate of success, while campaigns that start in August or sometimes September have the lowest. Campaigns that being in the fall tend to produce lower levels of success than those in the spring, with a slight uptick in December. This makes sense intuitively in the United States. In August people likely are on vacation, in early September it is back-to-school time for many, and between Thanksgiving and Christmas there may be a general "holiday" effect of people giving to known charities and spending money on celebrations and gifts. The period around Thanksgiving may be when many large organizations do more direct mail or other direct fundraising, so the fact that this is not the highest period of success also is consistent with the nature of crowdfunding.

2. Limitations

One major limitation with this data set is its size. There are millions of crowdfunding campaigns annually. Fundera.com, for instance, reports that there were more than 6.55 million campaigns in 2022. This dataset of 1000 is too small to represent all of those campaigns accurately. There are a number of subcategories where there are only two or three datapoints in a year, and no real conclusions can be drawn from that.

In addition, it does not appear that this dataset is representative of the overall population of campaigns, and we have no idea how it was selected from among the entire population, or the level of accuracy in collecting the information. We also do not know the purpose for which it was collected, which could lead to sampling error or bias. Among other things, it appears that all of these campaigns were run or sponsored by professional groups, which is very different than many crowdfunding campaigns would be. This also might contribute to what appears to be a much higher percentage of success than is reported generally for all crowdfunding campaigns; according to fundra.com, the average success rate overall is approximately 22%, while for almost all of the categories in this dataset, the highest percentage of failures is about 50%.

Some types of campaigns, such as raising money for medical care for an individual, also appear to be missing from the dataset, yet are one natural category for crowdfunding. The same is true for fundraising for natural disasters. The category of food has fewer than ten campaigns over the roughly twenty year reporting period, yet restaurant startups and food trucks are other categories where, while not based on empirical information, from personal information I know that numerous others have taken place.

As mentioned in part 1, there also could be some confusion or overlap between certain categories, particularly journalism, publishing, and photography, or music, film, and theater. Journalism and publishing also might have overlap with technology. We don't know how the campaigns were classified, by whom, or whether it was self-reported. Based on the named entities involved, and the "blurb" about the campaign style, it appears that there is not duplication of a single campaign across categories.

The data also is not representative across countries. Only seven countries are included, and, of those, three are Commonwealth countries where English is the primary language and which are closely allied to the United. Of the three where English is not the official language, two are small northern European countries with large international populations and many English speakers, and one is southern European. None of the countries other than the United States have enough data points for reliable predictions. Many regions of the world, including all of central Asia, the Middle East, the Far East, the Pacific Island nations, and all of Africa, are not represented at all. Even among the European countries, the large northern countries such as France and Germany are missing, as are any countries from eastern Europe. It is not clear if this is due to cultural differences in crowdfunding, bias in sampling, or inadequate sample size or other sampling errors.

With the data available, the three most frequent categories of theater, film, and music, followed by technology, exhibit the same relative frequencies as in the United States, except that in Australia there were slightly more technology campaigns than music campaigns. Other regions of the world might show really different results. For example, while it might be the case that crowdfunding is discouraged or illegal in China, from this dataset it is impossible to tell if these numbers represent collection difficulties like the organization gathering the data on the campaign having no one who speaks Chinese, or sampling error, or actually that crowdfunding is not done in China.

Even for those countries that have data, the sample sizes are very small (such as 48 over ten years), and in reality there might be more differences from the United States than are readily apparent from this data. This particular data also shows that no countries outside the United States are fundraising in U.S. dollars, but anecdotally and on statistical reporting sites, that is not the case.

The data is also very inconsistent across years. While there are many reasons that crowdfunding would have been very much reduced during the first year of the pandemic, for example, this data is reporting overall campaigns in the single digits. It is not possible to draw many accurate conclusions about trends based on this, without knowing more about the overall population of crowdfunding for the year 2020. Financial circumstances, elections, and events such as the unrest following the George Floyd killing could contribute enormously to differences in outcomes, without saying anything about the style or goal or outreach for a particular campaign. Similarly, the date before 2010 are very limited.

To interpret this data, more context about the events that were taking place during a particular year and month, or in a particular industry, would be necessary.

In addition, certain data, such as the "name" column, are not well-defined and it might be that incorrect conclusions could be drawn based on a misinterpretation of what the column represents. This discussion, for example, is treating "name" as the organization that organized the campaign, because the names do not seem to have any relationship to the category of the campaign, and all sound like law firms, accounting firms, or advertising firms. That might not be correct. The "blurb" also appears to be aimed at advertising firms or as developed by advertisers to get corporate sponsorship. It is also not clear what "staff pick" and "spotlight" are intended to describe. Assuming that either of these refer to interest within the organization that is to be the beneficiary of the campaign, these two fields might be expected to be predictors of success because they would tend to draw more interest and energy by the organization. Notably, the identity of the entity that is going to benefit from the campaign is at best unclear, or likely not named. Maybe this is important to protect privacy for some purposes, but it makes it more difficult to determine whether something like name recognition, or advertising budget, or size and type of outreach, such as internet, TV, or radio ads, had any effect on the outcomes.

One reason to suspect sampling error or having extracted a small section from a dataset without thought to its being representative of the population, is that is that the "name" and "blurb" columns appear to be unique. If this field is the advertising entity that is running the campaign, one might expect that more than one successful campaign had been run by top agencies, or one campaign style would be more effective, but there are not enough data to compare these items.

While, on this dataset, the size of the goal appears to have a very predictable effect, with three goal sizes having 100% success, these goals are represented by no more than 14 projects, whereas goals for amounts less than $10,000, or $50,000 or more, have hundreds of campaigns each. Again, a larger sample might make this information much more useful for predictions.

3. Other tables or graphs

Some other comparisons that might be useful to attempt to predict success would be as mentioned above, looking at "spotlight" or "staff pick" compared to outcome. Another set would be looking at the type of the campaign itself, which the blurb suggests, or the entity that is running the campaign (or the beneficiary) compared to outcome. I ran a number of informal comparisons like this, and was surprised to see that these did not appear to be very predictive of success or failure. I did not graph any of them, which might show such trends more readily.

Another set of graphs that might be very helpful is to look at "failed" campaigns by percentage of goal reached, considering the category or subcategory of the campaign. For instance, for a campaign raising money for a children's hospital, having achieved 99.87% of the goal, and raised hundreds of thousands of dollars, would not be a "failure" in the same way as a community theater that needs to meet a minimum goal to pay rent in order to keep the theater open, and misses the goal by 2%. Maybe in terms of an advertising agency, and psychologically, not reaching a goal could be seen as a failure, but some of the "failed" campaigns in this dataset did raise hundreds of thousands of dollars, far more money than many of the "successful" campaigns, and did reach over 99% of their goals.

It also would be interesting to look at total dollars raised by number of campaigns, filterable by category and country and subcategory and country, as well as dollars raised compared to number of backers, and average donation by category and subcategory. The definitions of "success" and "failed" do not depend on the amount raised or the average donation, but understanding the distribution of those contributed to reaching the goal could help to design better and more targeted outreach for future campaigns, to attempt to reduce the number of campaigns that do not reach their target goal, and to increase donations overall.

Another factor that might be predictive would be looking at length of campaign by outcome. I did this a little bit informally, and did not see the differences that I expected between a two-day campaign, a ten-day, and a thirty-day, but that was by eyeballing the data in the full table, while a chart and a filter box would be much more informative. If people get very excited during the first few days of a campaign, or for something like a weekend pledge walk, keeping the campaign going longer might result in much more work for diminishing returns. On the other hand, if information about the campaign is being spread through word of mouth or individual social media pages, or reaching older people who generally use computers much less frequently and generally don't spend as much time on their phones as teenagers and those in their twenties, a longer campaign might be much more beneficial and more likely to succeed. In something like journalism or technology, people may need longer to investigate the merits and trustworthiness of the beneficiary, read articles about the proposed product, or attempt to fact check journal articles before they are willing to give money to a particular campaign.

Finally, examining differences between countries might be highly predictive and helpful in understanding outcomes, particularly as crowdfunding can and does have a global reach. Looking at differences in outcome by currency and by date could add to the information available through comparing categories and subcategories. Although this dataset indicates that the only campaigns to use US dollars are based in the United States, other outside sources suggest that many campaigns that are not begun in the United States use US dollars, or offer a choice of currency. Some of the data here indicate that the campaign is using Euros, while others are using their own country's currency. It would be interesting to compare outcomes for each of those, although the numbers of campaigns to examine are likely too small to be informative. In some countries, such as in Argentina, the value of the exchange rate might be so volatile that campaigns in US dollars would be more successful, whereas in a country like Russia, it might be difficult to access US currency, and Euros or rubles would be more effective. Exchange rates and the stability of a particular currency might also dictate the choice of one or a few specific currencies other than a country's own.

In addition, examining the start month and year, and trends by year, by country, likely would be informative, and would be predicted to be quite different than in the United States, at least by month. Trends by year likely would share some similarities with the United States, such as the COVID-19 pandemic, and the war in Ukraine, whereas other things, like elections, sports events, or political unrest, likely would be more country specific. Certain sports events, such as the World Cup, could have a huge impact on when to start crowdfunding campaigns in other countries, but very little impact in the United States. Monthly differences likely would be significant. For example, school starts in late October in many European countries, and there is a six- or eight-week vacation in August and September. School also frequently continues into July for a summer term. Australia's summer and winter months are the opposite of all of the other countries in this dataset, and potentially could have a large impact on outcomes depending on when the campaign started. It's not clear if there are enough campaigns to examine these patterns by country, but it would be worth an attempt.

**Part 2:**

** Use Excel to evaluate the following values for successful campaigns, and then do the same for unsuccessful campaigns:**

* **The mean number of backers**
* **The median number of backers**
* **The minimum number of backers**
* **The maximum number of backers**
* **The variance of the number of backers**
* **The standard deviation of the number of backers**

** Use your data to determine whether the mean or the median better summarizes the data.**

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

Mean or Median. The median better describes this dataset than does the mean. The median for successful campaigns is 201, and for failed campaigns it is 114. The mean for successful campaigns is 851, and for failed campaigns it is 505.

The number of backers per campaign is enormously variable, and ranges from zero to 7,295. Sixteen campaigns (less than 2% of the total) had at least 5,000 backers, and 552 (approximately 55%) had 201 or fewer. Indeed, 292, or approximately 30%, had 100 or fewer backers. Although neither statistic presents a very clear picture, the median better conveys this large variability, whereas the mean masks it by making all campaigns appear to have hundreds more backers each than they actually have. While 55% have 201 or fewer backers, the mean for successful campaigns is more than four times larger, and for failed campaigns it is approximately three times larger.

Although it was not one of the statistics to calculate for the assignment, I calculated the mode for both sets of data. It was 85 for successful campaigns and 1 for failed campaigns. This also is not very descriptive of this dataset, because of the huge variance in the number of backers per campaign, but it comes closer to describing the ordinary case here than does the mean. On the other hand, it does not take into account the outlier backers in the multiple thousands for some campaigns, in some cases close to 10,000, for the campaigns with the largest numbers of backers. These large campaigns include 3 failed campaigns which each had more than 5,000 backers.

Variability. There is greater variability in the successful campaigns than in the failed campaigns, by a significant amount. The minimum number of backers for successful campaigns is 16, as compared to zero for failed campaigns. While there is little difference between these numbers at the law end, on the upper end, the difference in the number of backers between successful and failed campaigns is more than 1,500. The maximum for successful campaigns is 7,295, and for failed campaigns it is 6080. Thus, the standard deviation for each is enormous. It makes sense that this would be so. With respect to the most successful campaigns in terms of backers, publicity, and presumably also funds obtained, logically more backers would be supporting them, whereas the hundreds of middle-of-the-road campaigns, with fewer than 200 backers, are frequent in both groups, with the lowest numbers more common in failed campaigns.

It would be interesting to see if the total funds raised has a linear or exponential relationship to the number of backers, or if one or a few backers are donating millions and therefore skewing the mean donation far from the center of the normal curve, similarly to the mean and the numbers of backers, or if, where there are thousands of backers, each is giving $1. It might be much easier for a sponsor or a researcher to understand the numbers of backers and the amounts of their donations by seeing the data presented in a stacked bar graph rather than in table form.