Questions:

1. Given the provided data, what are three conclusions we can draw about Kickstarter campaigns?
2. What are some limitations of this dataset?
3. What are some other possible tables and graphs that we could create?

From the data provided regarding Kickstarter campaigns, there are three main conclusions:

1. According to the analysis of the relationship between the parent categories and the campaign states, campaigns about Theater, Music, and Film & Video rank the top three favorite categories within all campaigns. These categories take 77.80% (1679 out of 2158) of all successful campaigns, with 63.51% (2613 out of 4114) of total campaigns. Also, they are the only categories that pass 50% of the success rate.
2. All successful campaigns within the Film & Video category come from the documentary, shorts, and television. In the Music category, faith, jazz, and world music sub-categories appear no attraction to potential funding. Under the Theater category, funding is mostly generated from projects related to plays (694 successful campaigns out of 1393 total projects of Theater). Still, the other sub-categories fail to pass 50% of the success rate, which illustrates funding prefers only plays. Besides three parent categories, sub-categories like small batch (Food), tabletop games (Games), photobooks (Photography), nonfiction (Publishing), radio & podcasts (Publishing), hardware (Technology), and space exploration (Technology) all have the potential of fundraising.
3. There is no evidence showing the success rate and failure rate are related to seasonality. From the chart, spring and summer have (February, March, April, May, June, and July) apparent discrepancies between success and failure. This indicates that campaigns created in spring and summer are more willing to succeed.

However, the dataset has a lot of limitations. Firstly, it is tough to consider the category of campaigns. For example, the TV show about food or games should be categorized as Film & Video, Food, or Games. Secondly, it is hard to test if the success rate is related to the target amount (goal). For example, the campaign which has a goal of 5,000 but failed could have been successful if the target is 4,500. Thirdly, the state of failure and cancellation are not clear enough to analyze the attraction of different categories of the campaign. Campaigns’ failure could be due to lack of attraction, short period, over-optimistic. None of these details are clear. The dataset also does not provide the reason for the cancellation of campaigns. Last but not least, the population sizes of some countries are not big enough to conclude the potential to launch campaigns there.

There are some other possible investigations that I feel useful.

1. The relationship between goal amounts and time durations should be investigated, different categories and sub-categories could have intervals of goal amount/time duration among which the success rate is higher.
2. The failed campaigns need deeper investigations to help them raise more money. For example, lower the goal amounts, extend the time for campaigns, as well as launch in different countries might have made positive results.

Bonus Questions:

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

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

Answer

For successful campaigns, the median summarizes the dataset more meaningfully. According to the range of the numbers, most of them stay within 100. However, extreme numbers drive the average to more than three times of median. In this case, the median should be a better choice to summarize.

For the failed campaigns, since there are extreme numbers and multiple zeros in the dataset, the median is a better representative of the dataset.

The variances between the successful and unsuccessful campaigns illustrate that successful campaigns are more variability. A high volume of zeros in the unsuccessful data lowers the average to a number close to zero. Also, the number of extremes of unsuccessful campaigns is less than the number of successful campaigns. Both reasons result in the lower variability of unsuccessful campaigns.