Homework #1 Excel – Report on Kickstarter campaigns

1. Given the provided data, what are three conclusions we can draw about Kickstarter campaigns?
   1. Kickstarter campaigns in the categories of film & video, music, and theater are likely to be more successful in getting funding.
      1. Given the provided data on 4,000 past projects, the overall success rate (# successful campaigns to the total # campaigns) during the funding process is 53% (n=2185/4114).
      2. The categories of music (77%; n=540/700), theater (60%; n=839/1393), and film & video (58%; n=300/520) have the highest category success rate, with all 3 above the overall success rate of 53.1%.
   2. Kickstarter campaigns in the categories of food, games, photography, and publishing are likely to fail in obtaining their funding goals.
      1. The overall failure rate during the funding process is 37% (n=349/4114). In this dataset, the categories of food (70%; n=140/200), games (64%; n=140/220), publishing (54%; n=127/237), and photography (53%; n=117/220) had higher failure rates than the overall failure rate.
      2. Also noteworthy is all journalism campaigns were cancelled (n=24). The
   3. Campaigns in some sub-categories were 100% successful in reaching their funding goal.
      1. Interestingly there are several sub-categories for which all campaigns initiated were successful during the funding process. Of the 12 sub-categories with 100% successful campaigns, 8 were from high performing parent categories (music and film & video).
      2. While the parent categories of technology, publishing, and games performed below the average success rate of 53.1%, some sub-categories had 100% successful campaigns.
      3. Sub-categories with 100% success rate were classical music (n=40; music), documentary (n=180; film & video), electronic music (n=40; music), hardware (n=140; technology), metal (n=20; music), nonfiction (n=60; publishing), pop (n=40; music), radio & podcasts (n=20; publishing), rock (n=260; music), shorts (n=60; film & video), tabletop games (n=80; games), and television (n=60; film & video).
   4. Some sub-categories of campaigns completely failed to reach funding goals despite being in a parent category that is more likely to succeed.
      1. Campaigns in sub-categories of animation and drama had 100% failure even though the film & video category has a high success rate.
      2. Likewise, the sub-category jazz had 100% failure rate though the music category performed above average.
      3. The sub-categories with a 100% rate of failure were animation (n=100; film & video), children’s books (n=40; publishing), drama (n=80; film & video), fiction (n=40; publishing), gadgets (n=20; technology), jazz (n=60; music), mobile games (n=40; games), nature (n=20; photography), people (n=20; photography), places (n=20; photography), restaurants (n=20; food), video games (n=100; games).
   5. Campaigns launched in the spring, specifically the months of April and May, were most likely to reach funding goals. Those launched in the winter, specifically in December and January, were likely to fail to reach funding goals.
      1. The month of May had the largest number of campaigns launched with a successful outcome in the funding process and the highest proportion of successful outcomes (61%; n=234/386). The month of April has a similar performance, with a success rate of 60%.
      2. The number of campaigns successful steadily declines through September. There is an increase in the number of successful campaigns in the autumn (October and November) with a steep decline in December.
      3. The month of December had more campaigns fail (46.8%; n=118/252) than succeed (44.0%; n=111/252). January follows a similar trend.
2. What are some limitations of this dataset?
   1. The main limitation of this dataset is its size of just 4,000 past projects. The dataset indicates 53.1% (n=2185/4114) of campaigns have a successful outcome whereas the background information states only a third of the 300,000+ Kickstarter projects have made it through the funding process with a successful (i.e. positive) outcome. It is likely a larger dataset may yield a funding success rate more representative of the total population of over 300,000 campaigns.
   2. Another limitation of this dataset is the time period it spans. While the data set includes data from 2009, the year Kickstarter began, the last year it includes is 2017. It may be helpful to have the more recent data from 2018 onward.
   3. It is unclear how this sample of 4,000 Kickstarter campaigns were selected from the population of campaigns. Perhaps the parameters to select the sample could be adjusted to be more representative of the population.
3. What are some other possible tables and/or graphs that we could create?
   1. Statistics on the funding amounts for successful and failed campaigns by category. For example, the mean, median, variance and standard deviation of funding goals and donation size by category.
   2. Which months have the greatest number of donor and highest amounts of funds donated. More analysis could be done on the launch timing and funding received.
   3. What categories and sub-categories have the highest rates of cancelling.

Bonus Statistical Analysis

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

* In this dataset, the median is the better measure of central tendency.
  + When trying to decide whether the mean or median summarizes a dataset better, one must consider if the dataset is skewed with outliers, has a high degree of variability, or have values that extend the tails in either or both directions. The mean is the arithmetic average, calculated by summing all the values and then dividing by the number of values in the dataset. Because the mean is calculated with every value in the dataset it is impacted by extremely high and low values. This pulls the mean away from the “center” of the dataset and pulled closer to the extremes of the outlier values.
  + The median is the middle value in a dataset ordered from small to large. This measure of central tendency essentially splits the dataset in half, literally as it is the middle value in the ordered dataset. Because of the way the median is calculated, it is less impacted by outliers and skewed data. If the smallest or largest values in the dataset become more extreme (e.g. get even bigger or smaller), the median value does not change. However, the mean value would change.
  + There is a large degree of variability in this dataset. The minimum and maximum values of both successful and unsuccessful campaigns show a wide range of values and a large range. The standard deviation is much larger than both the mean and median, which indicates a high variability in both datasets.
* The mean and median of both successful and unsuccessful campaigns are on the low end of the range. The standard deviation is much larger than the mean. These observations indicate there are more projects successful with a small number of backers and just a few projects with an exceptionally large number of backers. This skewing of data points to the median as the better measure of central tendency compared to the mean.

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

* There is more variability with successful campaigns than unsuccessful campaigns.
  + The variance and standard deviation, which are measures of dispersion and spread in a dataset, are greater for successful campaigns. Variance is calculated by taking the difference between each data point and the mean, squaring, and then averaging it. Because the units of variance are squared and no longer the same units as the dataset, variance in it of itself is not particularly valuable. However, variance is needed for the calculation of standard deviation, which does have the same units as the dataset.
  + The standard deviation for successful campaigns is 844 (mean=194) compared to the standard deviation for failed campaigns at 61 (mean=18).
  + Other measures of variability, like range and interquartile range, are also greater for successful campaigns. Successful campaigns range is 26456 and IQR 112 whereas failed campaigns range is 1293 and IQR is 11.
* It makes sense there is more variability for successful campaigns than unsuccessful campaigns.
  + A campaign is more likely to be successful with a larger number of backers. Thus, successful campaigns are more likely to have outliers of large number of backers and have more variability compared to failed campaigns.
  + Even with campaigns varying in their funding goal, with some goals smaller and some larger, higher numbers of backers will be more common to successful campaigns. Extreme outliers can form when a campaign goes viral and backers keep joining the campaign past the funding goal. People may view campaigns that are close to its funding goal and with moderate number of backers as a “lower risk” investment than a Kickstarter project that has just begun its campaign.
  + The summary statistics indicate successful campaigns have around 60-200 backers per campaign on average and a huge variability of number of backers with a standard deviation of 845. The data has a long, skewed tail extending to the outliers such as the maximum number of backers of 26,457 for a single campaign.
  + Failed campaigns are more likely to have smaller numbers of backers per project and not likely to have extreme outliers of large number of backers compared to successful campaigns. The failed campaigns will not have quite as skewed of a tail towards the maximum value of 1293 backers for a single campaign. With less outliers on large number of backers, the variability of failed campaigns is lower.