**Homework 1 Yukie Kajita**

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

* From ‘Category analysis (‘CategoryStats’ worksheet in Excel), ‘theater’ (Grand Total = 1,393), ‘music’ (700), ‘film & video’ (520), and ‘technology’ (600), were more popular campaigns’ methods compared with other methods including ‘food’ (200), ‘games’ (220), ‘photography’ (220), and ‘publishing’ (237). In addition, ‘theater’ (successful counts = 839), ‘music’ (540), and ‘film & video’ (300) had more successful fundraising counts compared with other campaign methods.
* From “Sub-category” analysis (‘SubcategoryStats’ worksheet in Excel), ‘plays’ was the most popular methods (grand total count = 1,066) and contained more successful campaign counts (694 successful counts) compared with other campaigns. Interestingly, several campaigns, including classical music (grand total count =40), documentary (180), hardware (140), metal (20), nonfiction (60), pop (40), radio & podcasts (20), rock (260), shorts (60), tabletops games (80), television (60), had smaller counts in grand total compared with the ‘play’ counts. Nevertheless, those campaign methods were very successful because all of them met their fundraising goals (100% successful rate).
* From “Launch Date” analysis (‘LaunchDateOutcomes’ worksheet in Excel), I observed highly variable counts in ‘successful’ and ‘failed’ campaigns throughout the years. For example, the campaigns launched at the beginning of the year had more successful campaign counts compared with the campaigns that launched at the end of the years. In addition, a month of May showed the highest successful counts (234), but a month of December showed the lowest successful campaign counts (111). There was a slight variation in ‘canceled’ counts throughout the year.

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

* This dataset provided the counts of successful, failed, canceled; therefore, it is difficult to compare the raw value among categories. I would calculate the percentage of success, failed, and canceled, and then compared those data among categories.
* Each campaign had a different duration for fundraising, and included different styles of the campaign methods. For example, some campaigns might have needed to run a longer period of time compared with other campaigns to reach their goal. Some campaigns might have needed more time or people to prepare to launch the campaign because of the complexity of the campaign method, style, or structure. Therefore, it might be better to obtain the initial costs of the campaign, and include those to evaluate their results and efficiency of the campaign.
* We do not know how each campaign was advertised and launched.
* We do not know backers’ information (e.g., age, sex, race).

**3. What are some other possible tables and/or graphs that we could create?**

* I would convert ‘count’ data to ‘percentage’, and then compare the rate of success, failed, and canceled among categories (‘ExtraDataset’ worksheet in Excel).
* I would sort ‘countries’ to understand if each country shows similar or different patterns in success or failed campaigns among the categories. Also, I would sort ‘years – 2014, 2015, 2016, 2017’ to see whether years show similar or different patterns in the success of campaigns among categories, number of backers, and average donation, etc.
* I would examine a relationship between the number of backers and the average amount of donation per backer. Backers number might be increased if the average amount of donation per backer is low, because people feel more comfortable to donate a small amount of money.

**Bonus Questions**

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

* Our data showed that successful campaigns had higher value of variance (713167.38) and standard deviation (844.49), compared with the unsuccessful campaigns (variance = 3775.69, SD = 61.45) in the number of backers. Because ‘Median’ is the number found at the exact middle point of the dataset, and located at the center of that distribution; therefore, median would be more meaningful to summarize the data than mean in highly variable dataset (i.e., skewed distribution of dataset. See ‘Summary Statistics Table’ in ‘Bonus\_2’ Worksheet in Excel).

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

* As I mentioned above, there was more variability in the number of backers in successful campaigns compared with unsuccessful campaigns (See ‘Summary Statistics Table’ in ‘Bonus\_2’ Worksheet in Excel). That makes sense to me because some successful campaigns might have more backers with donating a small amount of money, and other successful campaigns might have fewer backers with donating a large amount of money. Therefore, the backer number would be highly variable in successful campaigns. On the other hand, in unsuccessful campaigns, backers’ number was very small. Therefore, the backer number was less variable in unsuccessful campaigns.