**Week 1 Excel - Challenge**

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

* The music, parent category, has the highest crowdfunding success rate (77%) and the Theater, parent category, is the most popular project in the crowdfunding project of Kickstarter.
* The projects in the Play sub-category have launched overwhelmingly more than any other projects on Kickstarter.
* Overall, the highest crowdfunding success is in May while December has the lowest crowdfunding success which even lower than failure.

1. What are some limitations of this dataset?

From the current dataset, there is a limit to drive an in-depth conclusion. The project has been launched in 13 different currencies and the goal amount of the crowdfunding varyingly. When it comes to currencies, there could be external factors that impact projects depending on countries such as perception of crowdfunding or recognition of the 'Kickstarter' brand. The default amount of crowdfunding also can be a factor in a project's success and failure as the excessive requirement makes the project unrealistic. Therefore, conclusions made with variables without additional constraints are prone to the error of generalization.

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

Logistic regression graph can find a correlation of a variable affecting the success of project funding, the fundraising period, the amount of fundraising, and the target amount. It is expected that the results of this analysis can cause sufficient interest to project funding developers and sponsors at a time when studies on the success factors of the crowdfunding project are required in various fields.

Bonus Statistical Analysis

|  |  |  |
| --- | --- | --- |
| **Summary Statistics** | | |
|  | **Successful** | **Failed** |
| **Total count** | 2185 | 1530 |
| **Mean** | 194 | 18 |
| **Median** | 62 | 4 |
| **Min** | 1 | 0 |
| **Max** | 26457 | 1293 |
| **Variance** | 712841 | 3773 |
| **STD** | 844 | 61 |

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

The most common method of determining representative values in statistics is the 'mean' value. However, the mean value can be well represented in situations where samples are concentrated near the mean. To compensate for this, we use the median for the value with the most intermediate upper body of the entire sample. There is a huge gap between the mean and median value of each successful and failed outcome, which means the variability of data is very high. Therefore, use the median value to evaluate the data is more meaningful.

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

With the current summary statistics, it does not make sense to compare the variability with the two variables; successful and unsuccessful. Variability describes how far apart data points lie from each other and from the center of a distribution. Along with measures of central tendency, measures of variability give you descriptive statistics that summarize your data. For small samples, the average value is heavily influenced by outliers. When measuring a measure of dispersion (variability), the mean absolute deviation or IQR should be used to eliminate the abnormalities and compare variability.