This excel workbook contains 8 sheets with their names shown below.

Data Package

per category

per sub-category

Date Created Conversion

Goal

Backers

Successful\_backers

Failed\_backers

Note: I kept color representations for State: successful, failed, canceled and live consistent across sheets.

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

(1) Based on “per category”, across all countries and all categories, the average successful rate is 53%. Music has the highest successful rate 77%. Campaigns under music launched on Kickstarter had the highest chance of achieving successful.

(2) Based on “per sub-category”, across all countries, under category music, several sub-categories have 100% successful rate, but I would pick rock if I want to launch campaign on Kickstarter since it has the biggest grand total (base).

(3) Based on “Goal”, there is a tendency that the lower the goal amount is, the higher successful percent is and the lower canceled and failed percent are. With goal range increasing, it seems that successful percent is decreasing, canceled and failed percent are increasing. It looks like canceled and failed percent are correlated.

2. What are some limitations of this dataset?

(1) If there are more data provided, I would like to check how closely canceled and failed percent are correlated as indicated in 1(3) above. There are traces of evidence that some turns in two lines happening at same locations.

(2) The whole Kickstarter idea is supported by backers. The dataset only has backers\_count column. If the names of the backers and their contributions could be retrieved in a table, then we can target some potential backers for specific campaigns. This could be processed in SQL.

(3) Data Ended Conversion shows the data are between 2009-2017. It is 2021 now so more recent data set should be provided.

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

(1) Under “Backers”, I provided the histogram chart on backers for successful and failed campaigns correspondingly. The standard deviation of backers for successful is high so its distribution is wide. The standard deviation of backers for failed is relatively low so its distribution is relatively narrow.

(2) Under “Successful\_backers”, I used normal distribution to observe the trend of backers count for successful campaigns. With ideal normal distribution, mean, median and mode are equal. In this case, normal distribution could approximately represent the distribution of backers count.

(3) Under “Failed\_backers”, I used normal distribution to observe the trend of backers count for failed campaigns. With ideal normal distribution, mean, median and mode are equal. In this case, normal distribution could approximately represent the distribution of backers count.

4. 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?

Median summarizes the data more meaningful because the backers\_count is widely distributed with irregularities and uneven spaces between data points. In a [*skewed*](https://statisticsbyjim.com/glossary/skewed-data/)*distribution*, the mean can start to fall outside the central area because [outliers](https://statisticsbyjim.com/glossary/outliers/) have a substantial impact on the mean.

There is more variability with successful campaigns. People are normally generous and opening minded when comes to funding. The next step is to find common causes, which lead to the failed campaigns. Since they are on the smaller data scale side, it may be more efficient to focus on finding rules for failed campaigns so that could avoid later.