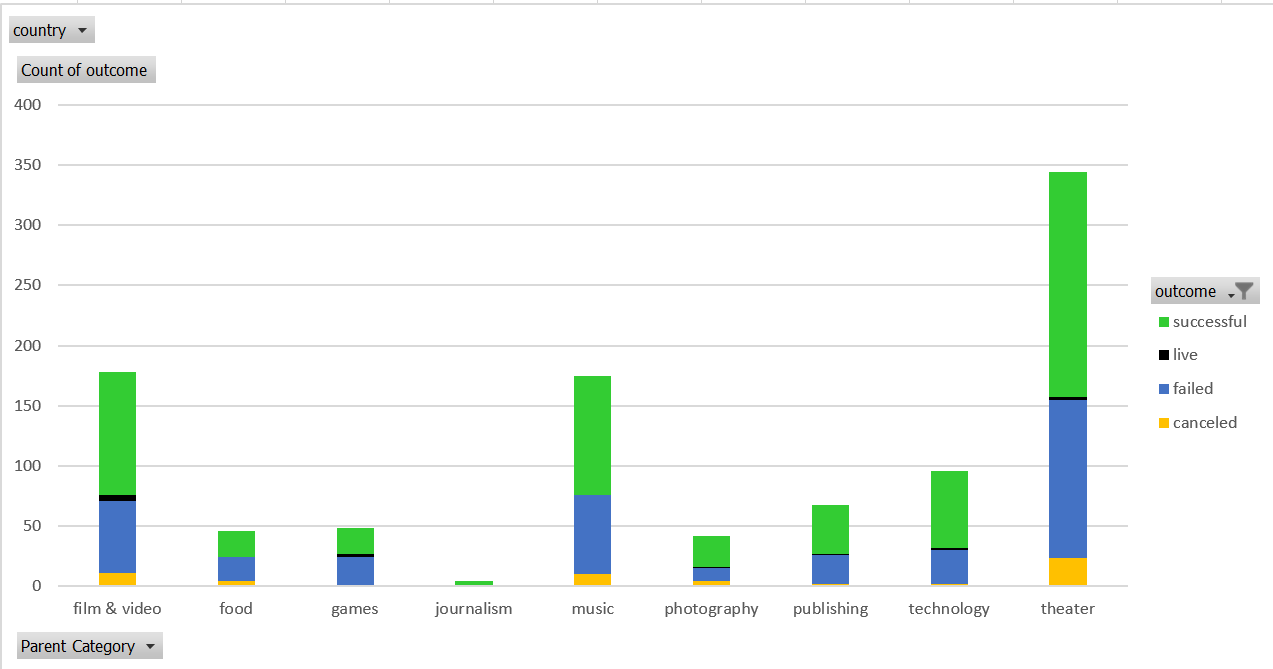
**Module 1 Challenge - MSExcel**

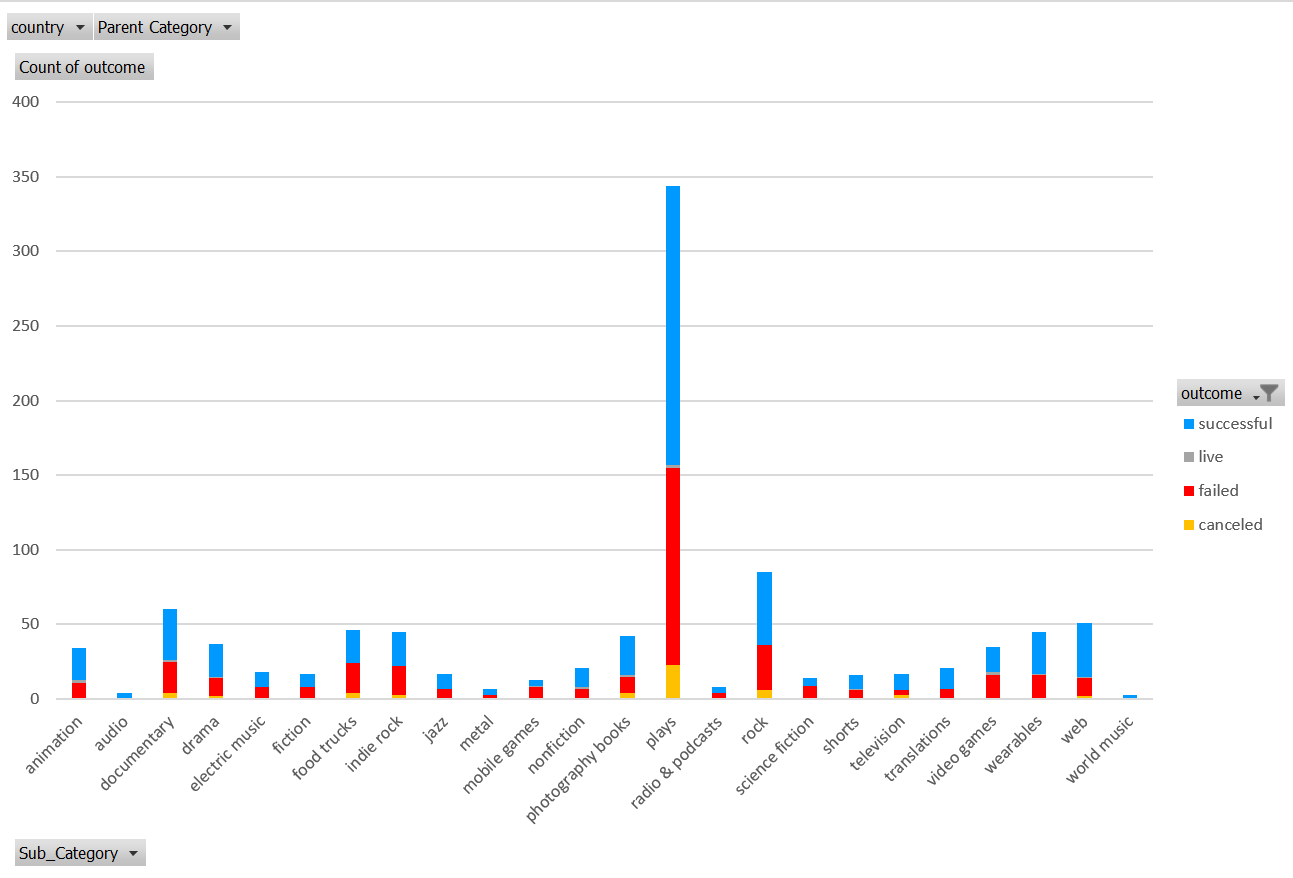
Create a report in Microsoft Word, and answer the following questions:

**Given the provided data, what are three conclusions that we can draw about crowdfunding campaigns?**

1. Theater, Film & video and music are the three most popular crowdfunding parent category in descending order (Figure 1) while Plays (Figure 2) is the most popular sub-category in crowdfunding (refer to Sheet1 & 2).

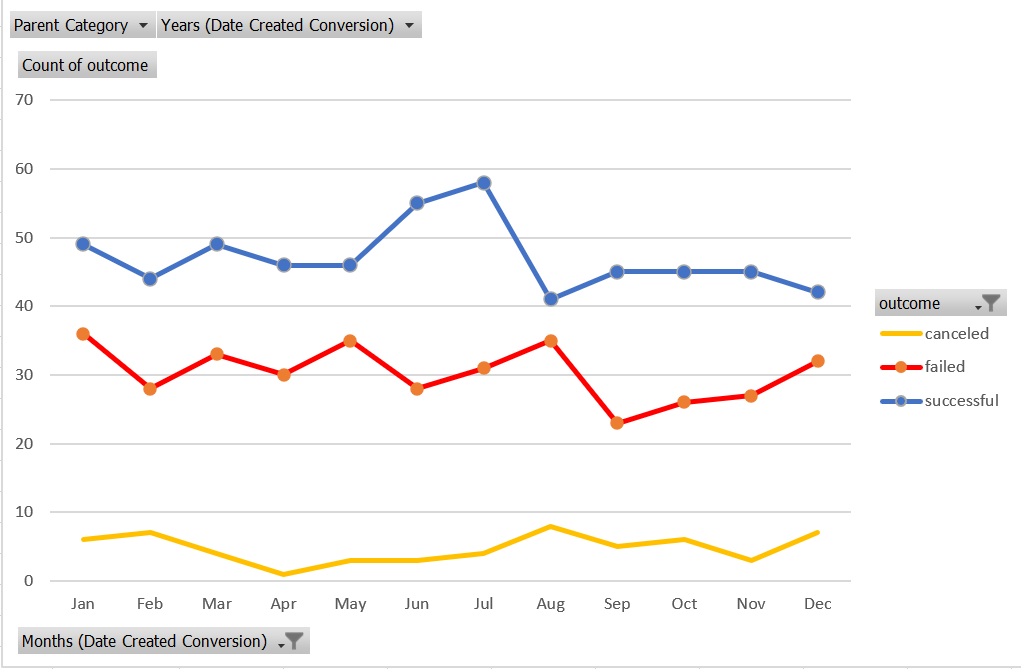


**Figure 1**: Theater, film & video and music are the most popular crowdfunding project by parent category.



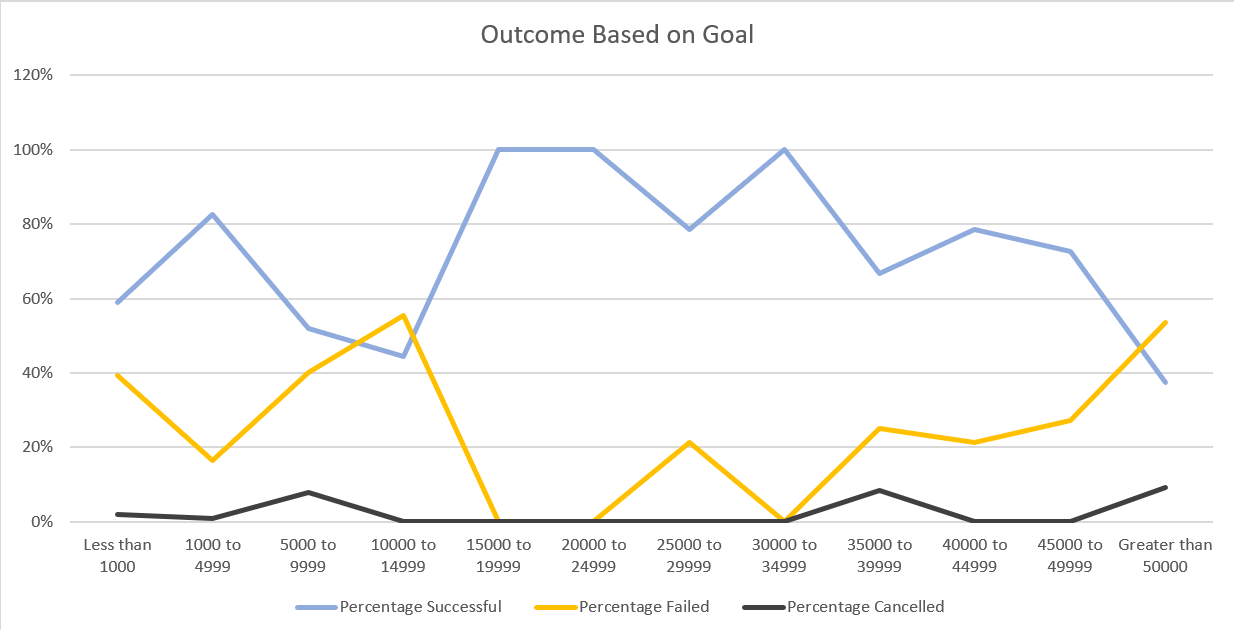
**Figure 2**: Plays is the most popular sub-category.

1. Month of Jun and Jul are the most successful months to crowdfund while August and December are the worst month to crowdfund (Figure 3 and refer to Sheet3.)



**Figure 3**: Month of Jun and Jul are most successful months while August and December are the worst month to crowdfund.

1. Crowdfunding for 15000 to 24999 and 30000-39999 target have 100% success rate while 10000-14999 and greater than 50000 target have highest failure rate (Figure 4 and refer to Sheet4).



**Figure 4**: Showing 100% success for 15000-24999 and 30000-39999 crowdfunding target while 10000-14999 and >50000 are the least successful.

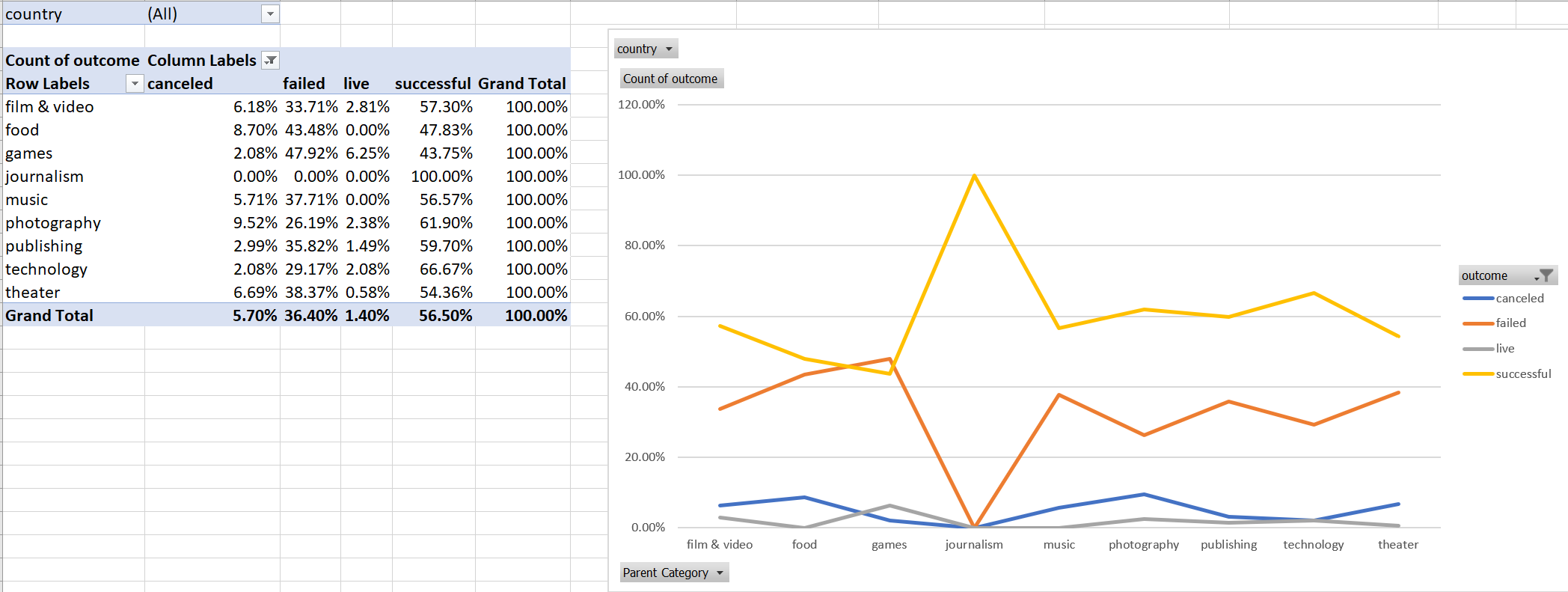
**What are some limitations of this dataset?**

The following are the limitation to this dataset.

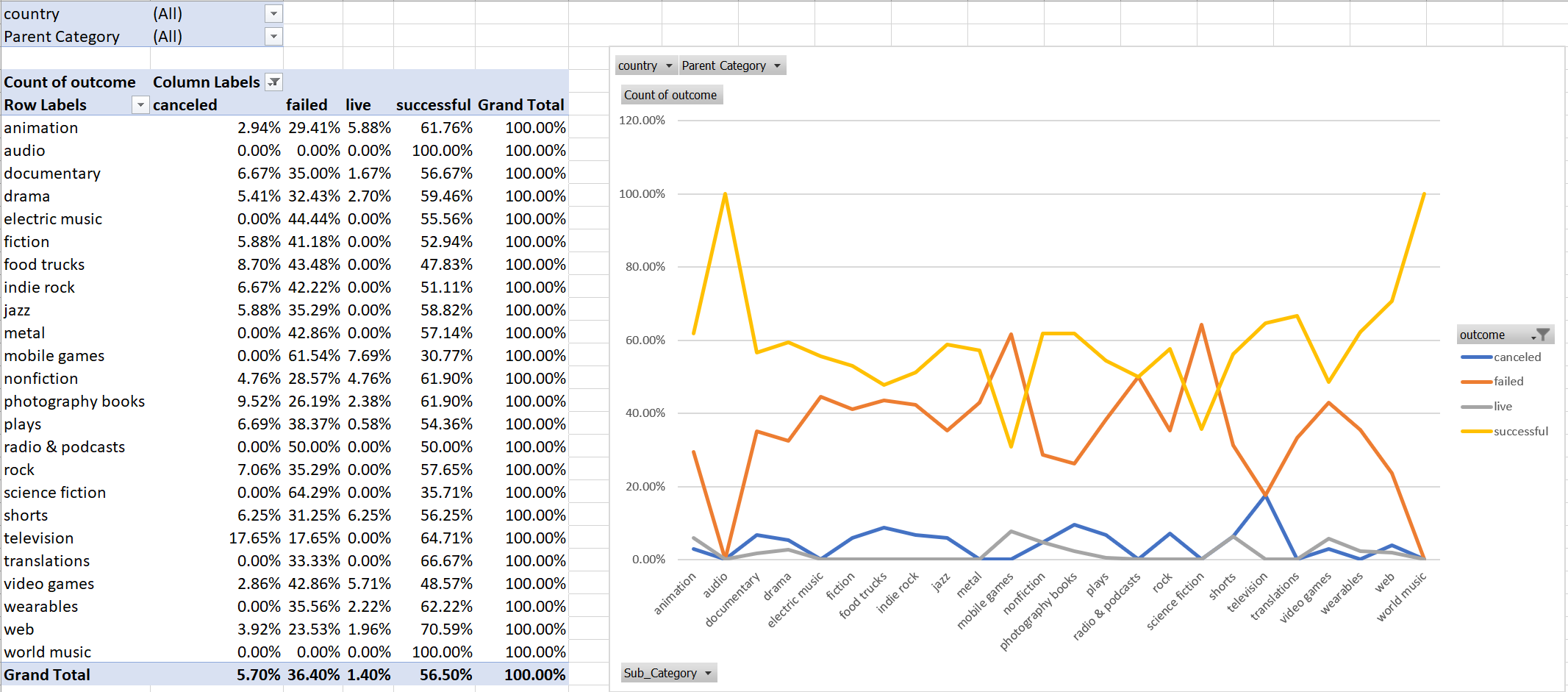
1. Dataset might not have been randomly sampled to be good representative of the whole population. Hence, our conclusion could be biased.
2. Dataset only has 1000 sample and we don’t know the actual total number of the whole population and hence not able to make a judgement on how small or big this dataset represent.
3. We are not told how the dataset was derived at. What are the source and method of collection and whether any preprocessing or filters have been applied to the dataset. It is always good to start from raw dataset and applies filter ourselves.

**What are some other possible tables and/or graphs that we could create, and what additional value would they provide?**

1. Create a pivot table to calculate the percentage of canceled, failed, live and successful per Parent-Category and make a pivot line chart at the same time. We will be able to see which Parent-category would be the most successful from this (Figure 5 and refer to Sheet1\_Copy).

**Figure 5**: This is a variation of the first exercise but instead of showing actual number, this one is showing percentage per Column. This clearly show that journalism parent category has 100% success while the highest failure is games.

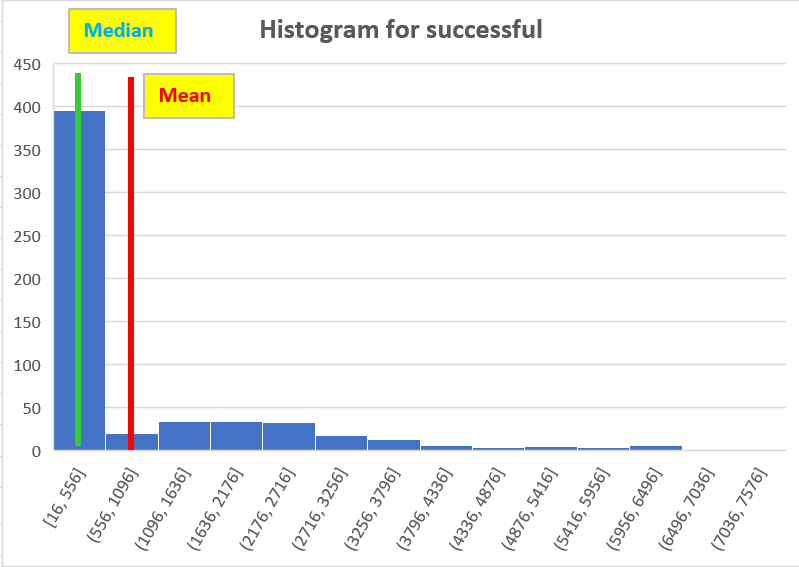
1. Create a pivot table to calculate the percentage of canceled, failed, live and successful per Sub-Category and make a pivot line chart at the same time. We will be able to see which sub-category would be the most successful from this (Figure 6 and refer to Sheet2\_Copy).



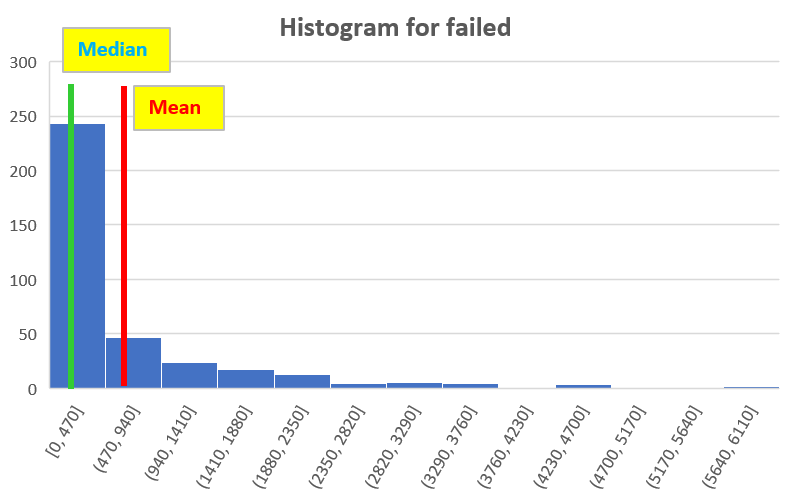
**Figure 6**: This is a variation of the second exercise but instead of showing actual number, this one is showing percentage per Column. This clearly show that world music and audio sub-categories have 100% success while the highest failure is science fiction.

**Use your data to determine whether the mean or the median better summarises the data.**

Since both histogram for successful and failed showed positive skewness, the Median is better than Mean in showing the central tendencies as shown in the Figure 7 & 8 below. Mean is not plotted in the middle of the distribution while Median is.



**Figure 7**: Histogram for Successful showing skewness. **Green line** drawn for **Median** and **red line** for **Mean**.

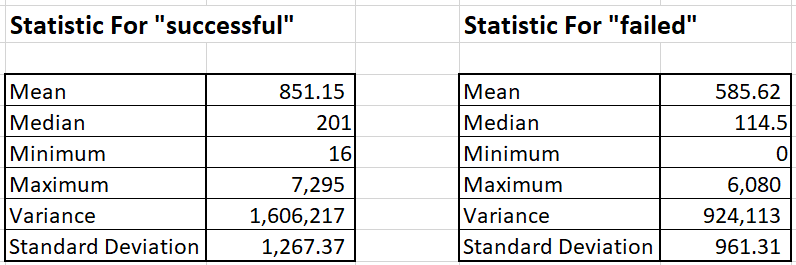


**Figure 8**: Histogram for Failed showing skewness. **Green line** drawn for **Median** and **red line** for **Mean**.

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

Variability of the data can be seen from the variance and standard deviation number. The higher the variance, the more spread out the data. Standard deviation is a measure of how spread out the data is from the mean. Since our dataset is skewed, mean is not really good to be used as central tendency and hence standard deviation is not so meaningful too. Therefore, judging from the variance data alone in Figure 9, we conclude that successful campaign is more spread out than unsuccessful campaign.

This makes sense since in Figure 7 & 8, we see that successful campaign has longer tail than failed campaign. Furthermore, the range is more in successful campaign (7295-16=7279) than in unsuccessful campaign (6080-0=6080).



**Figure 9**: Showing the basic statistic for successful and failed campaign.