Module 1 Challenge – Written Report

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

Conclusion 1 – Technology over Everything?
When determining the success of crowdfunded projects, Technology campaigns are your safest bet to fund, games should be avoided if possible. Using the Category_Pivot figure, we see that while theater categories have the largest successful number of campaigns. But given that theater categories significantly outnumber other campaigns, (nearly 33% of **all** crowdfunded projects are for theater), we should look at the percentage of successful to failed campaigns. Through this framework, we see that the best performing category is technology, with 67% of campaigns being successful. The worst performing group is games, with only 44% of campaigns being successful. Games also is the only category that has a higher failure rate than success rate.

Conclusion 2 – Shoot for the Stars or Just Try to Get to the Moon?
Can we use the **size** of a campaign to determine if it will be successful or not? Yes, strangely. Looking at the Outcomes Based on Goal graph, we see what percentage of campaigns are successful based on their goal size. The sweet spot is the mid-range, with campaigns that are $15,000 – $35,000 perform significantly better than campaigns that are both larger and smaller. This tracks if we think about it: too little of a goal and people can't be bothered to care, while too large of a goal might be a daunting task that pessimists would dismiss as Sisyphean.

Conclusion 3 – When Should We Launch?
Looking at the Date_Pivot chart, we gain some insight as to **when** we should launch a campaign. Filtering the data to see if the month of when a campaign launched can identify success, we see July as the highest chance of success and August as the highest chance of failure. While this might initially answer a question, it's more important to look at the month that has the largest delta between successful and failed campaigns. This allows for backers as a function of success to be limited and makes performance relative to the two criteria we are judging the campaigns on. Looking at the data through this framework, both June and July are the best months to start a campaign while August is the worst month to start a campaign.

What are some limitations of this dataset?

The dataset is limited in the fact that we have defined success as just being funded. There currently is no data on whether a funded project achieved its goals or if it was funded and then fizzled out. In this sense, it's identical if a backer gets behind a failed campaign and if they get behind a successful one that is unable to achieve the campaign's stated goals.

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

To account for the effects of the global economy, a chart showing successful and failed campaigns as a function of the year they were active (time between start and stop). This allows us a means of

evaluating how many projects may have suffered due to these forces. Conversely, we can also see which campaigns were benefiting from a strong economy. A rising tide may in fact lift all boats.

We could also see if there is a relationship to success and the words used in their blurb. We could do this by making every word a unique variable and counting the instances. Time is a luxury and if your elevator pitch is lacking, it may explain your lack of success. This shouldn't be shocking to anyone, as SEO and LLM have dominated the news recently. When trying to stand out, beating the algorithm goes a long way.

Statistical Analysis – Mean and Median Data

We want to see if the numbers of backers to a campaign can predict success. To perform the analysis, only campaigns that were successful or failed were considered. We start by creating our summary statistics, which can be found below:

	Successful Campaigns	Failed Campaigns
Mean	851	586
Median	201	115
Minimum	16	0
Maximum	7295	6080
Variance	1.60E+06	9.22E+05
StDev	1266	960
Count	565	364
Mode	85	1
25% Q	127.5	38
75% Q	1288.5	789.5
IQR (25%-75%)	1161	751.5

After calculating our statistics, we need to determine if the dataset should be characterized by the mean, median, or some other statistic. First, we need to determine if we have a statistically significant dataset. If we were drawing from five data points, we might as well just flip a coin and save ourselves the hassle. But this dataset has 1000 points, which is well within the ranges of statistical significance. Let's break these down:

- We see that the average successful campaign has 45% more backers than the failed campaigns. So far, so good, average backer count may be a good proxy. Should we stop here?
    - Lol, no, let's keep going
- The median can be used to determine where the midpoint in the dataset lies and compare it to the average. In a normally distributed dataset, we would expect the mean and median to be similar. Looking at our data, both the median values of the two datasets are significantly smaller than their respective average values. This means we should not be using mean to summarize the values but should be using median to characterize the dataset.

Let's talk about variability! Some of the best campaigns / companies have been supported by a relatively small number of backers (looking at you, Cargill family). Using only the mean and median provides an

incomplete portrait of the dataset. Let's look at the variance and standard deviation to see the spread of datapoints. Successful campaigns have a larger variance and standard deviation from the mean than failed campaigns. This makes sense due to the nature of failure and success. While there are many ways to move forward and succeed, there appears to be relatively few ways of staying still and failing. The more backers you have – statistically at least – the higher chance you have of success. That being said, the data spread is so large that using backer statistics is flimsy at best when trying to predict the outcome of a campaign.