TECNOLÓGICO DE MONTERREY

DATA ANALYTICS BOOTCAMP

MODULE 1 – KICKSTARTING WITH EXCEL

---

1. OVERVIEW OF PROJECT

In this module, I helped an up-and-coming playwright, Louise, who wants to start a crowdfunding campaign to help fund her play, **Fever**. She is estimating a budget of over $10,000 USD and is understandably hesitant about jumping into her first fundraising campaign. So Louise decided to contact me, an Excel power user, to help her organize, sort, and analyze several thousand crowdfunding projects with 12 variables to uncover any hidden trends to determine whether there are specific factors that make a project's campaign successful.

PURPOSE: Louise will use these insights to plan her own campaign and set it up for success, mirroring other successful ones in the same category.

2. ANALYSES AND CHALLENGES

After Louise's play **Fever** came close to its fundraising goal in a short amount of time, she wanted to know how different campaigns fared in relation to their launch dates and their funding goals. Using the Kickstarter dataset that I've already combed through, I visualized campaign outcomes based on their launch dates and their funding goals.

2.1. OUTCOMES BASED ON LAUNCH DATE

2.1.1. OVERVIEW

For this analysis I wanted to know if there is a relationship between the month in which a campaign was created, and its outcome. Meaning if there is a higher probability of success according to the month in which Louise launches her campaign. For this, as seen on *Image 2.1.1* "Outcomes Based On Launch Date" Overview, a pivot table was created from the original data frame present in KickStarter sheet using parent category and years as filters, only including observations with "theater" as their parent category. I used "outcome" (I.e., successful, failed, canceled) as columns variables in my pivot table, the month in which the campaign was created and the number of plays (count) that had a certain outcome given a certain month of creation.

*Image 2.1.1* "Outcomes Based On Launch Date" Overview
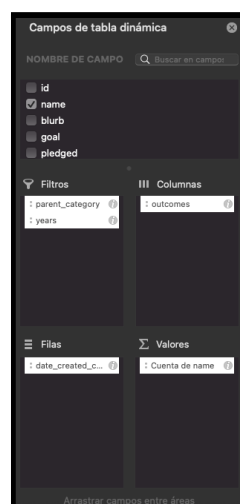
2.1.2. CHALLENGES AND DIFFICULTIES

After creating the pivot table, I wanted my chart to show the frequency of a certain outcome given the month in which the campaign was launched. Nevertheless, the pivot table displayed the information by years and then by quarter and finally by month (as seen on ***Image 2.1.2*** "Outcomes Based On Launch Date" Challenges and Difficulties). What I had to do was to remove the "Años" and "Trimestres" filters that Excel automatically created for me in the rows section of the pivot table field list, and thus, only months would be automatically shown (***Image 2.1.3*** "Outcomes Based On Launch Date" Challenges and Difficulties).

Image 2.1.2 "Outcomes Based On Launch Date" Challenges and Difficulties

Cuenta de name					
parent_category	theater				
years	(Todas)				
Cuenta de name	**Etiquetas de columna**				
Etiquetas de fila	successful		failed	canceled	Total general
2010	3				3
Trim.3	1				1
jul	1				1
Trim.4	2				2
oct	1				1
dic	1				1
2011	4				4
2012	8				8
2013	11				11
2014	241		152	10	403
2015	304		202	12	518
2016	249		130	12	391
2017	19		9	3	31
Total general	839		493	37	1369

Image 2.1.3 "Outcomes Based On Launch Date" Challenges and Difficulties
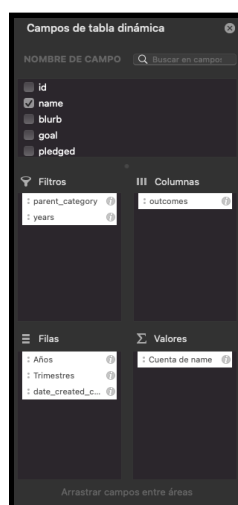
2.2.   OUTCOMES BASED ON GOALS

2.2.1.   OVERVIEW

For this analysis I wanted to know if there is a relationship between the fundraising goal and its outcome. For this, as seen on *Image 2.2.1* "Outcomes Based On Goals" Overview, a summary table was created from the original data frame present in KickStarter sheet, using the columns: Goal, Number Successful, Number Failed, Number Canceled, Total Projects, Percentage Successful, Percentage Failed, and Percentage Canceled. This meant that I wanted to calculate the number of observations with a given outcome according to its original fundraising goal. For this, I had to segment fundraising goals into 12 buckets, starting from less than $1,000 to more than $50,000, mainly with $4,999 increments (except for the lower buckets). Using the =COUNTIFS() function I was able to count the number of observations with more than one filter parameter as input, using both the outcome column and the goal column from the original data frame present in the KickStarter sheet. Finally, I summed the number of observations for a given goal and calculated the percentage that each outcome represented given a certain goal.

Image 2.2.1 "Outcomes Based On Goals" Overview

Goal	Number	Number	Number	Total Pro	Percenta	Percenta	Percenta
Less Than 1,	322	113	18	453	71%	25%	4%
1,000 to 4,9	931	420	60	1,411	66%	30%	4%
5,000 to 9,9	380	282	51	713	53%	40%	7%
10,000 to 14	168	144	40	352	48%	41%	11%
15,000 to 19	94	90	17	201	47%	45%	8%
20,000 to 24	62	72	14	148	42%	49%	9%
25,000 to 29	55	64	18	137	40%	47%	13%
30,000 to 34	32	37	13	82	39%	45%	16%
35,000 to 39	26	22	7	55	47%	40%	13%
40,000 to 44	21	16	6	43	49%	37%	14%
45,000 to 49	6	11	4	21	29%	52%	19%
Greater tha	86	258	100	444	19%	58%	23%

2.2.2.   CHALLENGES AND DIFFICULTIES

While working on this analysis I encountered four main challenges:

i.   I did not want to write down by hand the names of all of the 12 different buckets, so I created two columns with the lower and upper bound of the bucket and used the =CONCAT() function using " to " as a separator.

ii.   As it was my first time working with the =COUNTIFS() formula, I had trouble working with it. I read the documentation and noticed that the applied criterias must be applied to different columns in question.

iii. As not all buckets were separated by a linear difference, I had to fill in the filters (numbers, i.e., lower and upper bound of the bucket) by hand. Also, some criterias included greater or EQUAL than and others just greater than, which made the process manual.

iv.   While running the formula for the rest of the columns I used the =UPPER() and the =REPLACE() formulas so the outcome could be extracted from the name of the column so I wouldn't have to fill in manually the outcome for every cell.

EXAMPLE:

=CONTAR.SI.CONJUNTO('KickStarter   '!$F:$F,MINUSC(SUSTITUIR(B1,"NUMBER   ", "")),'KickStarter   '!$D:$D,">=5000",'KickStarter   '!$D:$D,"<9999",'KickStarter   '!$H:$H,"USD", 'KickStarter '!$R:$R, "plays")
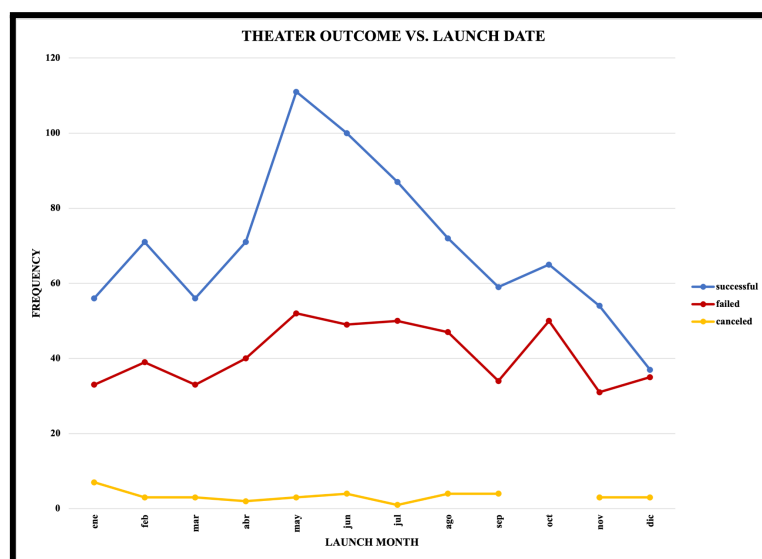
3.   RESULTS

3.1.   RESULTS BASED ON LAUNCH DATE

What are two conclusions you can draw about the Outcomes based on Launch Date?

1. My hypothesis was that the behavior of successful and failed theater plays would be opposite, meaning that in the months where there are more successful theater plays, there would be a drop in failed theater plays. Nevertheless, this is not the case, with successful and failed outcomes following the same trend no matter the month, successful theater plays having a more visible trend, suggesting that the database might be biased towards theater successful plays, i.e. The database has more observations with successful outcomes.

2. Because of the conclusion stated in the previous point, launching a campaign in May does not necessarily mean that there is a higher probability of success in May than in the rest of the months, as failed plays also reach a maximum during this month. Nevertheless, my recommendation would be to launch the campaign in May as the difference between successful and failed is greater, but having the previous point under consideration.

*Image 3.1* "Outcomes Based On Launch Date" Results

3.2. RESULTS BASED ON LAUNCH DATE

What can you conclude about the Outcomes based on Goals?

1. I was not able to reproduce the exact same graph as the one presented on the Module 1 Challenge explanation. I believe the statement is not well described as it does not clearly state to filter out observations whose goals are not in USD. I did filter out everything besides USD and used "plays" as a sub category. It is also extremely weird that there are no canceled observations using the filters I used for the =COUNTIFS() formula.

2. Despite not having the exact same graph, the general tendency remains the same. Percentage failed and percentage successful are mutually exclusive, meaning they are inversely related. The general tendency says that as the monetary goal in USD increases, your probability of success decreases (and thus, your probability of failure increases). There is a point in the $25,000 USD to $29,000 USD where Louise has the same probability of succeeding and failing, every goal before this bucket is more probable for Louise to succeed and vice versa.

Image 3.2 "Outcomes Based On Goals" Results
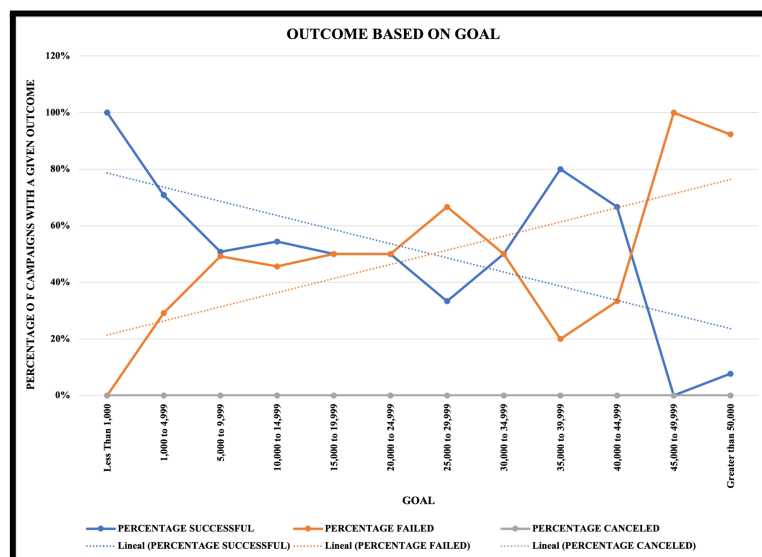
3.3. LIMITATIONS AND OTHER VISUALIZATIONS

For the two analyses I did, there is one opportunity area per analysis that I would like to assess:

1. I would like to analyze a base that has 50% successful, 50% failed or 33.33% successful, 33.33% failed, 33.33% canceled, this way we can really know with sufficient information the temporal tendency of failed and canceled theater plays.

2. I would like to further segment the base having a conversion of all goals to USD, this way we can compare pears with pears and apples with apples.