Team: Kowshik Raj Durai Murugan & Sinduja Senthil Kumar Course: BUAN / MKT 6337

SAS Project 1

Executive Summary: -

This project was commissioned to examine the data on past Kickstarter projects and recommend ways to increase the chances of putting up a successful project by project owners, maximize the revenue obtained by the Kickstarter platform and provide insightful suggestions to backers looking to fund projects. Based on various analysis that were done, few key insights are summarized as shown below.

The research draws attention to the fact though the Kickstarter website plays host to more than 250,000 projects and more than 50% of them had failed. In perspective, there were two main reasons for the failure of projects. One major contributing factor for this failure was owing to the unrealistic pledge amount and pledge period. Another factor was that the stipulated commission percentage required for Kickstarter was not achieved. Upon further investigation, it was evident that extending the pledge period for the projects by a single day lead to an increase in revenue of approximately $1400 for the platform.

Developing from this, we further analyze the impact of goal amount and pledge time on successful projects. From the explanatory data analysis performed prior to testing, it was evident that successful projects had lesser goal amounts and shorter pledge periods compared to other categories of projects such as failed and canceled projects. To dive deeper into the analysis, we tested if the difference in pledge time and average goal amount between the successful projects and other project states were significantly different. The test led to a significant result and we concluded that this difference is indeed significant.

Looking into the analysis category wise, “Technology” related projects had certain interesting finds. Technology was among the first on sorting based on categories that pledged more amount from backers than goal amount requested and it stands as the third category with the highest number of backers. But contradictorily, it was seen that despite the keen interest from backers, the success rate of such “Technology” projects was only 20% with nearly 63% of projects failing. This may have been due to the fact that quality of projects was not on par with the quantity of projects. This was then proceeded by testing if the difference in average backers between successful and failed technology projects were significant or not. The test led to a significant result which indicates that certain technology projects are particularly preferred than the others by the backers.

To conclude our insights, we tried to draw relations between the average number of backers and sub-categories. With plots and investigation, it was seen that the category “Chiptune” had the highest number of average backers despite its low count of a mere 35 projects. This was followed by looking into its popularity in the US compared to other countries and it was found that it was largely popular in the US as expected. On an end note, we decided to test if the category “Chiptune” had any significant impact on the “Music” category as a whole. Upon testing however, it resulted in an insignificant result making us to assume that despite it’s high popularity among backers, “Chiptune” does not impact “Music” as a whole.

Introduction: -

Crowdfunding is the practice of funding a project or a venture by raising monetary contributions from many people across the globe. There are several organizations such as DonorsChoose.org, Patreon, Kickstarter which hosts the crowdfunding projects on their platforms.

This project aims at analyzing data from one such organization – the Kickstarter. Using this platform many imaginative and ambitious projects are bought to life. This platform earns revenue through the commissions received from projects that have obtained successful funding and in return connects individuals with innovative ideas with backers who can provide funding for their projects. Kickstarter has hosted more than 250,000 projects on their website with more than $4 Billion collective amount raised up to date.

Business Case:

The essential business use-cases in the crowdfunding scenario can be considered from three different perspectives - from the project owner's perspective, the company's perspective and the backer's perspective

1. From the project owner's perspective, it is highly beneficial to be aware about the key characteristics of a project that greatly influence the success of any project. For instance, it will be interesting to pre-emptively know about following questions:
    - What is an ideal and optimal range of funding goal for my project?
    - On which quarter of the year, I should post the project on Kickstarter to benefit from more backers?
    - What are the apt keywords to use in my project title?
    - What should be the total length of my project description be?

2. From the perspective of companies which hosts the crowdfunding projects, they receive hundreds of thousands of project proposals every year. A large amount of effort is required to screen the project before it is approved to be hosted on the platform. This creates the challenges related to scalability, consistency of project vetting across volunteers, and identification of projects which require special assistance etc., In this scenario it is an essential necessity for these companies to know about:
    - Which category of projects should be promoted more by Kickstarter?
    - What is the success rate of the different categories of projects posted? Etc.,

3. From the backer's perspective they are probably the angel investors to these individual project's owners. Inspite of putting in so much of dollars into these projects, what they receive back is just few rewards and self-satisfaction. This makes it very crucial for them to select projects that not only fits their interests but also that are potential and worthy.

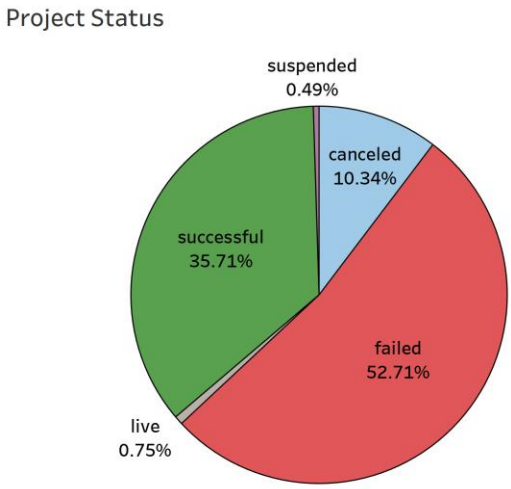
It is due to these perspectives, there is a need to dig deeper and find more intuitive insights related to the project's success. Hence, we begin with Exploratory Data Analysis to understand about the patterns and trends in data and see how these can be converted to valuable insights that benefit one or many of our stake holders.

Exploratory Data Analysis: -

The dataset contains 14 fields with 375,093 observations. The fields contain details about the name of the project, No. of backers supporting each of these projects, the category and the country of the product launch, the project goal amount requested by project owners, launch date of project etc., to name a few.

Though the data set is multi faceted, with various information, the most interesting question would be to know how many projects that are hosted on this Kickstarter platform are successful.

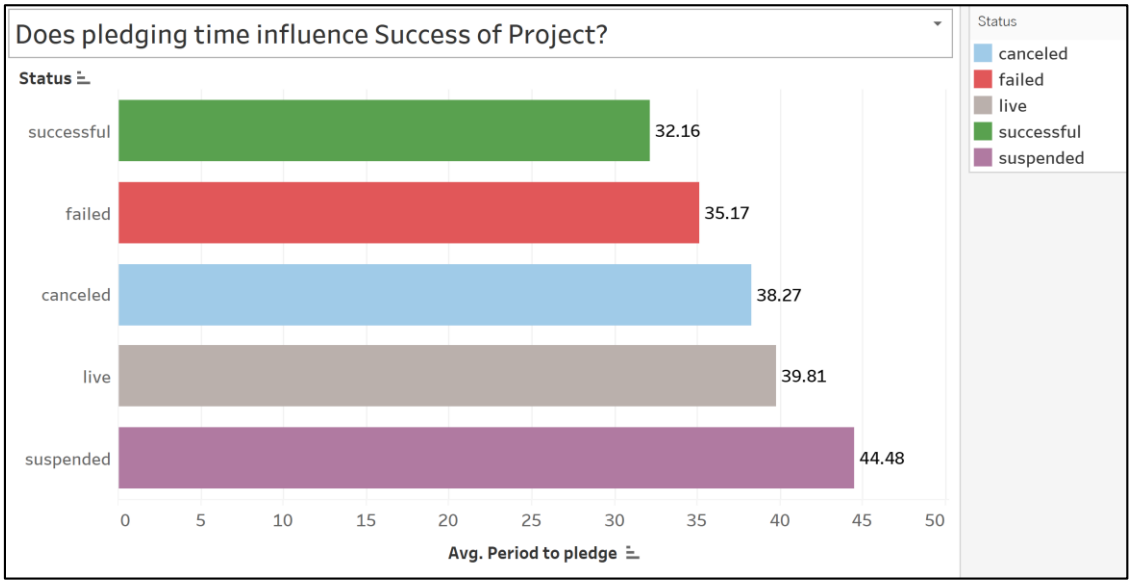
The field "State" from the dataset provides this information, which is as below.



More than 50% of the projects hosted on this platform has failed. And the successful projects account to only 36% of the total projects.

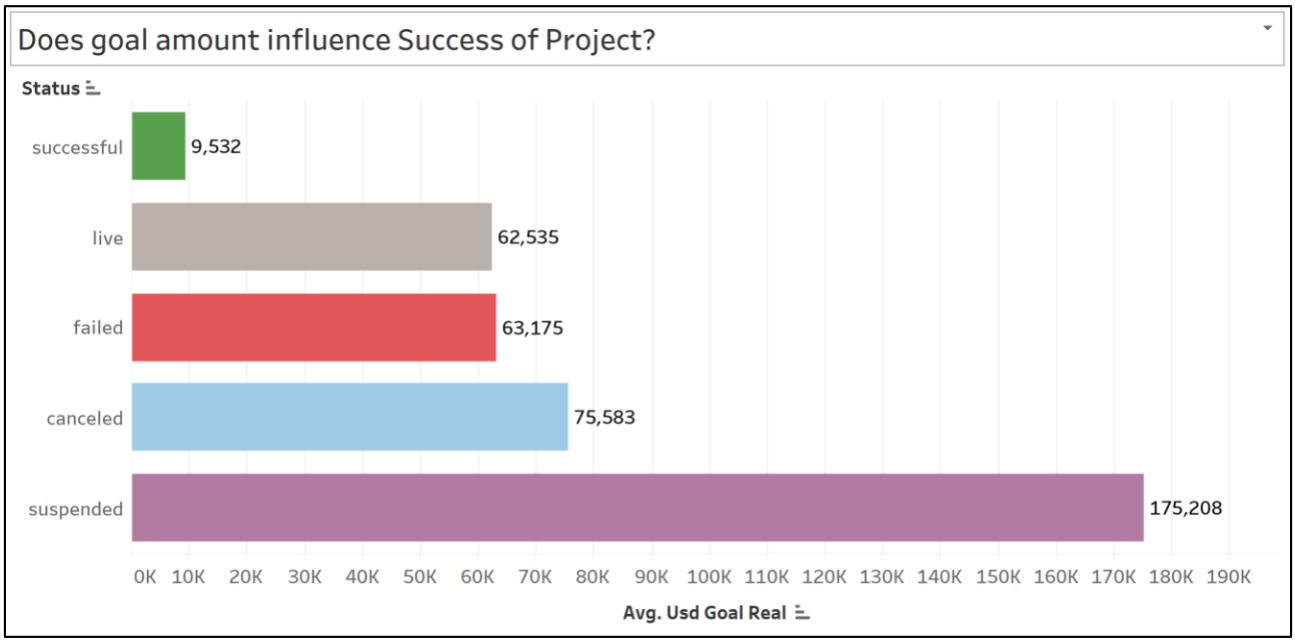
Characteristics of success projects:

To understand the characteristics contributed to each of these project statuses various plots were done. The following bar plot indicates the relationship between project statuses and pledge time. Pledge time is the difference between Launch date and deadline. This is the time duration during which the goal amount as requested by the project owners would be raised.



The average pledge time (in days) is small for successful projects when compared to others.

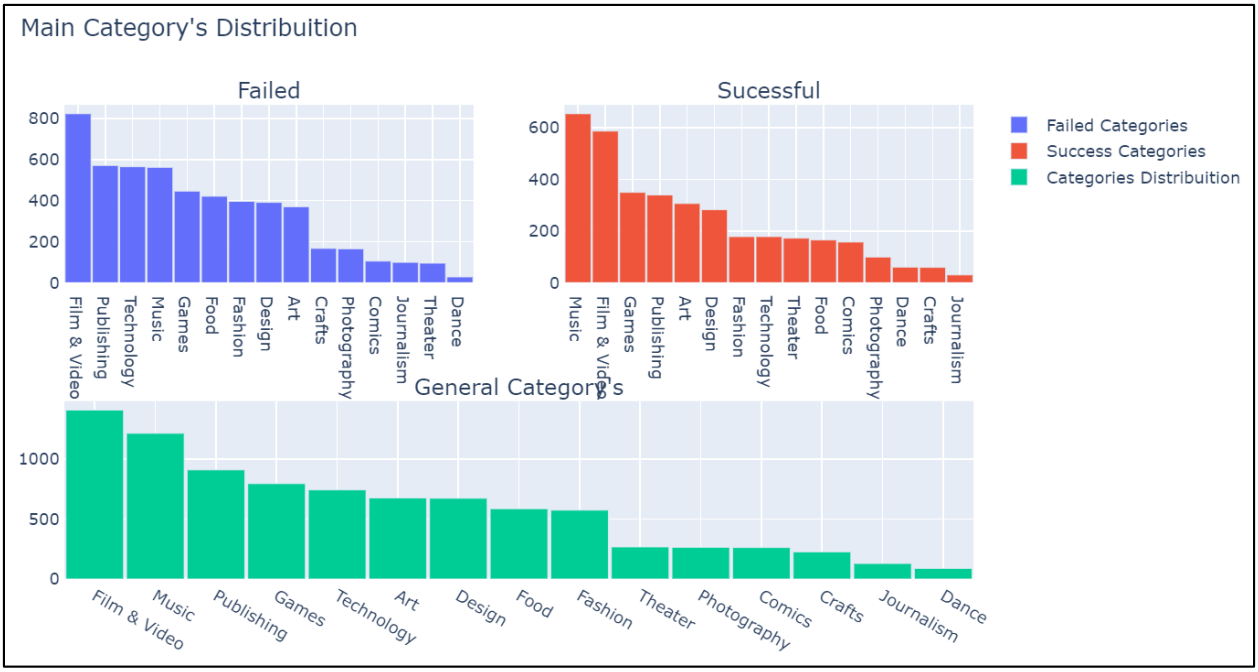
To understand the relationship between goal amount set by the project owners and the project statuses, the following bar chart is plotted.



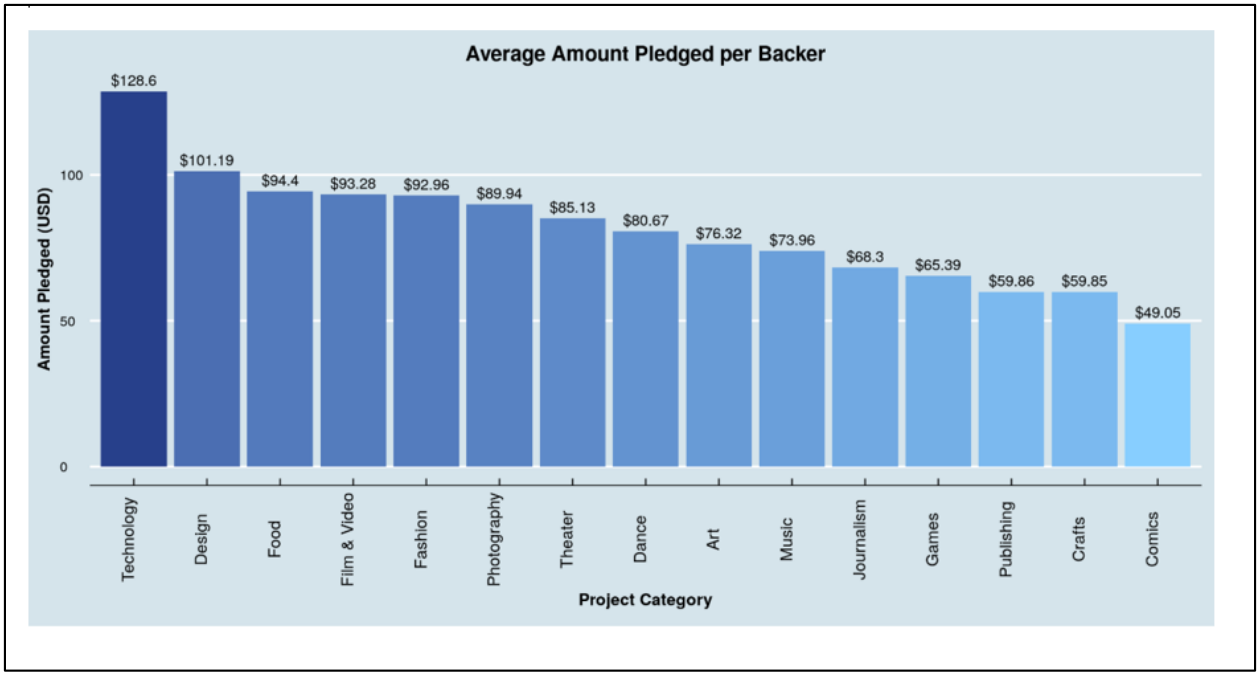
The average goal amount (in USD) is small for successful projects when compared to others.

Main Categories Vs project statuses:

Moving on to the "main categories" in which the projects where launched, projects that belong to certain main categories are more successful than the others and it is illustrated by the following plot.



We can see that the top 5 main categories that had the highest count of projects were Film & video, Music, Publishing, games and Technology. However, the top 5 successful projects include the same categories of projects as before except for Technology. Also, from the below plot we witness the backer's interest in various main categories in which the projects were launched, and that technology tops it.

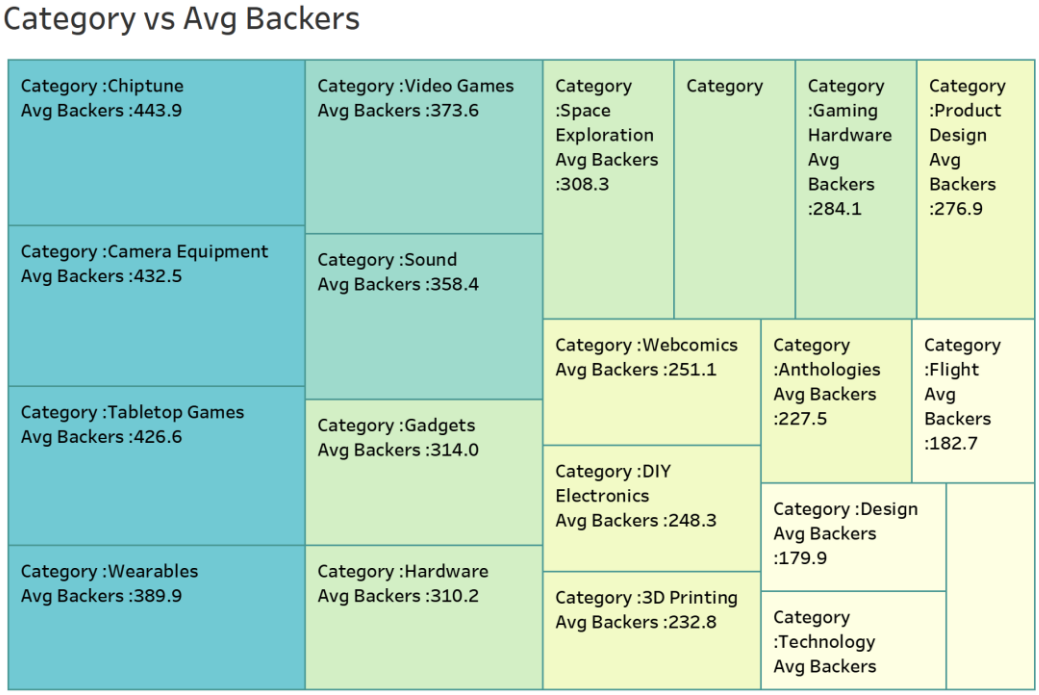


The highest average amount pledged is highest for Main Category - Technology

Further we move on to the subcategories and explore further.

Categories Vs Backers:

We also investigate the average number of backers' category wise to get an idea of which category of projects gain the greatest number of backer's support. Shown below, is a treemap displaying the top 20 categories with the highest average number of backers.



As one would expect, certain categories have a greater number of average backers compared to other categories.

In a similar fashion, many other explorations were done with plots and tables. With the preliminary analysis been done, further analysis needs to be performed keeping in mind the possible questions that might be of interests to our three stake holders the Kick starter Platform, the Individual Project owners and the Project Backers. Some of the possible goals for these stake holders are as tabulated and the objectives are set in order to satisfy one or many of these goals.

Stakeholder	Goals
Kickstarter Platform	• To generate more revenue through commissions • To increase their face value by increasing the percentage of success projects • To attract more Individual Project owners and backers to their platform to withstand competition
Individual Project owners	• To gain enough support from backers to ensure success in their projects • To imbibe trust and maintain good relationship with kick starter platform with clear project goals
Project Backers	• To identify projects that maybe of interest owing to personal interest. • To invest in projects that hold benefits in the future.

Based on the understanding that is obtained through the EDA, the following objectives are set to aid in identifying insights.

Objectives:

1. To explore ways to convert failure projects to successful ones.
2. To test and identify whether factors such as pledge time and goal amount influences the success of a project.
3. To test and identify whether there is a significant difference in number of backers for successful and failed projects in the Main category "Technology".
4. To identify the relation between the average number of backers and categories to draw insightful conclusions.

Analysis and Insights: -

Objective 1: To explore ways to convert failure projects to successful ones-

On viewing the Kickstarter data and through exploratory analysis, the predominant fact about the project status is that nearly 53% of the projects that were released using Kickstarter platform had been a failure. On analyzing further about the projects that failed, the following were observed. The projects that failed can be classified into two groups:

a. Type A: Projects that have raised funds more than or equal to the goal amount but still failed since the commission % required by kick platform was not reached
b. Type B: Projects that did not pledge any money at all or that pledged a lesser amount than the goal amount required within the fundraising deadline

Descriptive Statistics:

	Count
Total Failed Projects	199,718
Type A	6
Type B	199,712

Assumption: The Type A projects are very low in number and hence might be an outlier. Although, the reason behind the failure of these projects is not known for sure, it might have failed because of their insufficient pledge amount that did not accommodate the Kickstarter platform commission.

Analysis:

If the number of failed projects reduce and successful projects increase, it is beneficial to both the Kick starter platform and individual project owners. But how could they do this? The failed projects can become successful if the number of backers supporting the projects increase but this is outside their circle of influence, however what is within their control is deciding and extending the number of pledge days.

Since the kick starter platform commission for projects varies around 3-5% (Information taken from kick starter website) of goal amount, we assume a constant average of 4% commission for all projects. With this scenario, the average fund raised per day, total days required to pledge goal amount + 4% commission is calculated, thus leading to the number of extended days required for funding. Thus, by extending the fundraising period, the success rate of these projects can be increased as per the summarized table below.

EXTENDED FUND RAISING AND BENEFITS TABLE:

S.No.	No. of days extended for fund raising	No. of new successful projects	Increase in revenue for Kick starter (in USD)
1.	1 day	7	1393.04
2.	2 days	20	7017.34
3.	3 days	37	15275.66
4.	4 days	57	24743.85
5.	5 days	76	28461.44
6.	6 days	124	49406.83
7.	7 days	151	69170.79

Results and Benefits:

Kick starter Platform: By extending the fund-raising period to 1 extra day, their commission can extend up to $1393 and so on. Thus, kick starter platform gains more commission as well as good fame by facilitating more successful projects in their platform.

Individual Project owners: Also, for projects owners extending the pledge days based on the average pledge amount they receive each day would help reach the required goal amount and make the project go live.

Project Backers: Increasing the pledge days a little might inspire more capable backers to sponsor, by making them realize that goal amount is about to be reached which means the project is going live very soon and also since many backers have already sponsored they might understand that this is a potential project which might spur their interest.

Thus, by increasing the number of fund-raising days based on the goal amount for specific project, all the stake holders can benefit.

Insight 1: Failed projects can be converted to successful projects, thus increasing commission for the kick starter platform by increasing the pledge time.

Objective 2: To test and identify whether factors such as pledge time and goal amount influences the success of a project-

Every stakeholder is interested in successful projects. There might be various factors which might influence the success of a project such as the how many backers sponsor the project; which country is the project going to be launched and so on. However, our question now is does pledge time and goal amount influence the success of the projects.

Analysis 1: Pledge time Vs Successful projects

The average pledge time is calculated based on the launch date and deadline and was compared to the project status. From the plots in EDA it is evident that the average pledge time (in days) is small for successful projects when compared to others. However, in order to verify whether this difference in average pledge time between projects is statistically different we do ANOVA and DUNNETT tests.

Hypothesis Testing:

H0: μ1 = μ2 = μ3 = μ4 = μ5 = 0
H1: At least one of the means is different,
Where,
μ1 = Mean value of 'Pledge time' for level "Canceled" in factor variable 'State'
μ2 = Mean value of 'Pledge time' for level "failed" in factor variable 'State' and so on….

Here we test whether the average pledge time for the various status of projects are similar or different.

ANOVA TEST RESULTS:

```
                            Df    Sum Sq Mean Sq F value Pr(>F)
Pledge_Time.dt$state         4 1.651e+06  412771   94.23 <2e-16 ***
Residuals               375088 1.643e+09    4381
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the ANOVA test result, we conclude that the average pledge time for the various projects states are different.

DUNNETT TEST RESULTS:

Dunnett test is used to test how similar or different the average pledge time for the various project states are when compared to the reference state – "Successful"

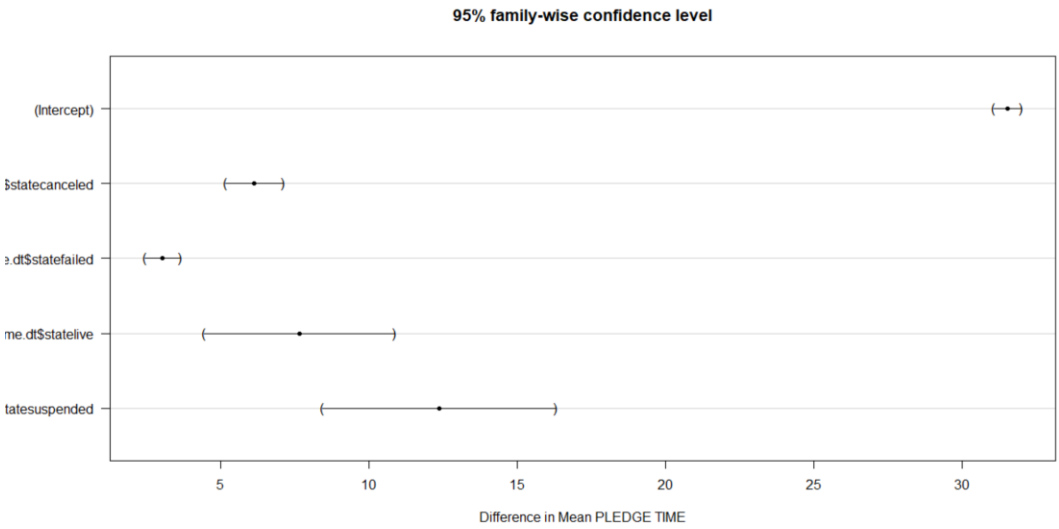
```
        Simultaneous Tests for General Linear Hypotheses

Fit: aov(formula = Pledge_Time.dt$Period_to_pledge ~ Pledge_Time.dt$state)

Linear Hypotheses:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept) == 0                     31.5194     0.1808 174.296   <1e-08 ***
Pledge_Time.dt$statecanceled == 0     6.1266     0.3817  16.052   <1e-08 ***
Pledge_Time.dt$statefailed == 0       3.0422     0.2342  12.989   <1e-08 ***
Pledge_Time.dt$statelive == 0         7.6507     1.2640   6.053   <1e-08 ***
Pledge_Time.dt$statesuspended == 0   12.3582     1.5510   7.968   <1e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

From the results we see that the average pledge time for projects that are successful is significantly different from all other groups.

DUNNETT TEST PLOT:



From the plot also we see that the mean values of all groups are significantly different from Projects that are successful as the confidence intervals of none of these values passes through zero.

Analysis 2: Goal amount (in USD) Vs Successful projects

The average goal amount requested by the individual project owners was compared to the project status. From the plot in EDA it is evident that the average goal amount (in USD) is small for successful projects when compared to others. However, in order to verify whether this difference in average goal amount between status groups is significantly different we do ANOVA and DUNNETT tests.

Hypothesis Testing:

H0: μ1 = μ2 = μ3 = μ4 = μ5 = 0

H1: At least one of the means is different,

Where,

μ1 = Mean value of 'USD goal real' for level "Canceled" in factor variable 'State'

μ2 = Mean value of 'USD goal real' for level "failed" in factor variable 'State' and so on….

Here we test whether the average goal amount for the various states of projects are similar or different.

ANOVA TEST RESULTS:

```
                          Df    Sum Sq   Mean Sq F value Pr(>F)
GoalAmt_USD.dt$state       4 3.02e+14 7.549e+13   56.29 <2e-16 ***
Residuals            375088 5.03e+17 1.341e+12
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the results we see that the average goal amount for projects that are successful is significantly different from all other groups.

DUNNETT TEST RESULTS:

Dunnett test is used to test how similar or different the average goal amount for the various project states are, when compared to the reference state – "Successful"

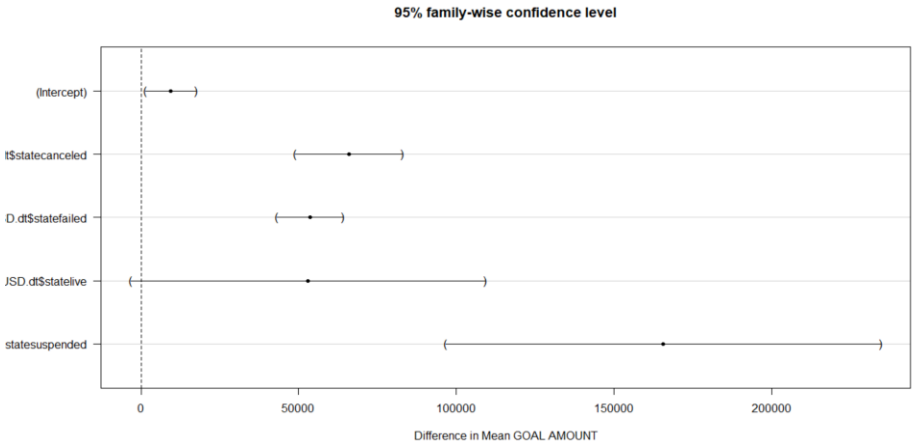
```
         Simultaneous Tests for General Linear Hypotheses

Fit: aov(formula = GoalAmt_USD.dt$usd_goal_real ~ GoalAmt_USD.dt$state)

Linear Hypotheses:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept) == 0                      9532       3164   3.013   0.0121 *
GoalAmt_USD.dt$statecanceled == 0     66051      6678   9.891   <0.001 ***
GoalAmt_USD.dt$statefailed == 0       53643      4098  13.089   <0.001 ***
GoalAmt_USD.dt$statelive == 0         53003     22117   2.396   0.0735 .
GoalAmt_USD.dt$statesuspended == 0   165676     27139   6.105   <0.001 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

From the results we see that the average goal amount for projects that are successful is significantly different from all other groups.

DUNNETT TEST PLOT:



Results and Benefits:

Thus, from the statistically significant test results it is evident that the average pledge time and average goal amount (in USD) of successful projects are significantly different from other groups.

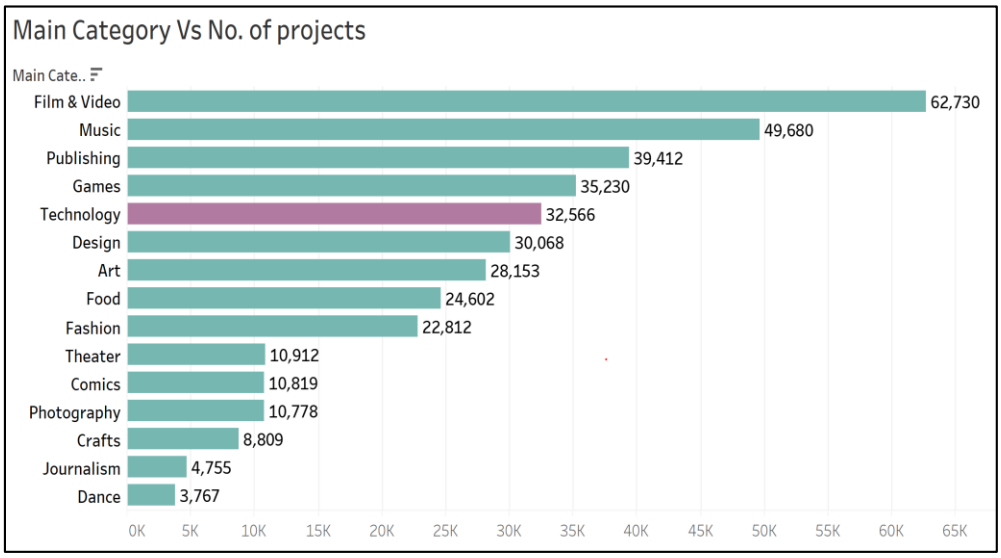
Kick starter Platform: With pledge time and goal amount influencing project success, Kick starter platform can use this as one of the criteria to choose the right projects to be launched in their platform, thus reducing more failure projects

Individual Project owners: From the statistical tests it is evident that pledge time and goal amount influence the project success. Also, since these are decided by the individual project's owners themselves, they can play it to their advantage to ensure their projects are successful

Insight 2: Pledge time and goal amount influences the success of a project

Objective 3: To test and identify whether there is a significant difference in number of backers for successful and failed projects in the Main category "Technology"

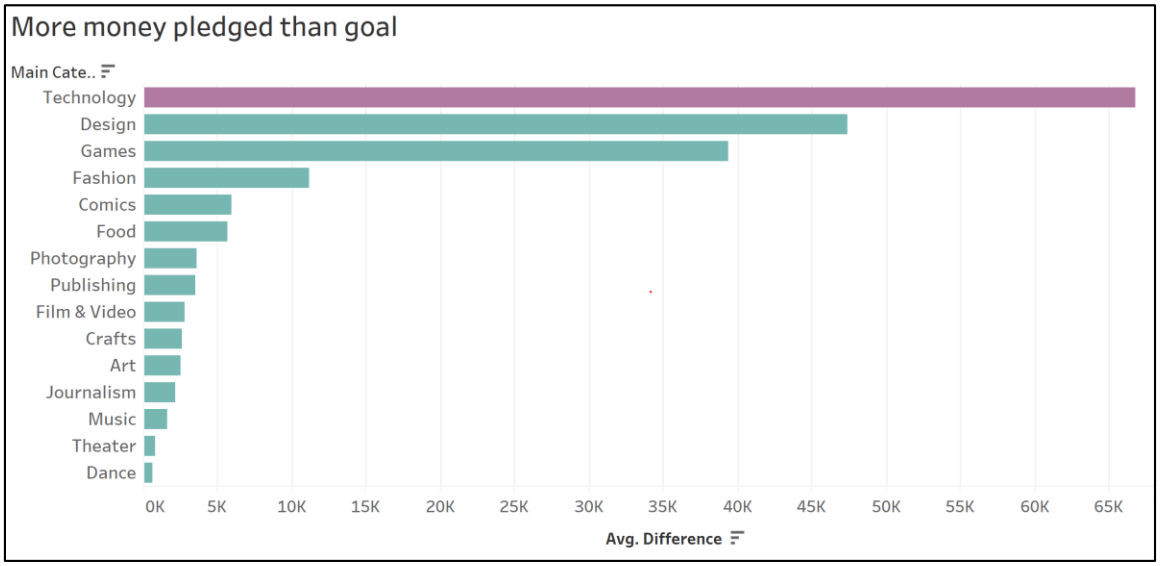
Projects that belong to certain Main categories are more successful than the others. Out of the 15 Main Categories, Technology is one of them. Technology has become an important aspect in our everyday life, and it is no surprise that Technology is fifth Main category with largest number of projects. Also, Technology had the third highest numbers of backers for the goal amount pledging.



Main Category Vs backers

Main Cat..	Avg. Backers
Film & Video	67
Music	55
Publishing	57
Games	322
Technology	164
Design	241
Art	42
Food	54
Fashion	61
Theater	47
Comics	135
Photography	40
Crafts	27
Journalism	38
Dance	43

And the difference between average amount pledged and goal amount is highest for projects in this Main Category- Technology



In spite of having higher number of backers, more projects and being the Main Categories that has pledged more money than the goal amount, yet Technology has the lowest success rate.

Main category Vs Success rate

	Status				
Main Category	successf..	failed	live	canceled	suspend..
Dance	62.07%	32.78%	0.48%	4.33%	0.35%
Theater	59.88%	33.98%	0.38%	5.57%	0.19%
Comics	54.00%	37.30%	0.70%	7.78%	0.21%
Music	48.70%	43.78%	0.57%	6.65%	0.30%
Art	40.88%	50.19%	0.69%	7.89%	0.34%
Film & Video	37.66%	52.45%	0.53%	9.17%	0.19%
Games	35.53%	45.42%	0.81%	17.60%	0.62%
Design	35.09%	49.27%	1.01%	13.81%	0.82%
Publishing	31.21%	58.73%	0.76%	9.14%	0.17%
Photography	30.66%	59.23%	0.45%	9.15%	0.51%
Food	24.73%	64.91%	0.75%	8.99%	0.62%
Fashion	24.52%	62.16%	1.10%	11.62%	0.60%
Crafts	24.01%	64.74%	0.86%	9.57%	0.82%
Journalism	21.28%	65.97%	0.65%	11.00%	1.09%
Technology	19.76%	63.31%	1.16%	14.48%	1.30%

This might be because though backers were interested in technology products, they might have not been interested in sponsoring all technology products. This can be verified using a T test.

Analysis:

Hypothesis Testing:

H0: There is no significant difference in mean values of backers for successful and failed projects in Main Category - Technology

H1: There is no significant difference in mean values of backers for successful and failed projects in Main Category - Technology

T TEST RESULTS:

```
        Welch Two Sample t-test

data:  TechProj.df$backers by TechProj.df$state
t = -22.923, df = 6436.6, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -757.0577 -637.7749
sample estimates:
    mean in group failed mean in group successful
                20.26242                717.67874
```

Here p value is <0.05, hence we reject null. Thus, there is significant difference in number of backer's for successful and failed projects in Main Category – Technology

Results and Benefits:

Kick starter Platform: Despite its high popularity, the platform should try to enforce stricter guidelines when projects related to "Technology" are put up on the website. This will ensure a higher success rate, more backers which will in turn result in better revenue.

Individual Project owners: Project owners should focus on putting up more projects that are actually beneficial for the future and not focus on personal gain. This way the chances of their projects becoming successful increases and attracts more backers.

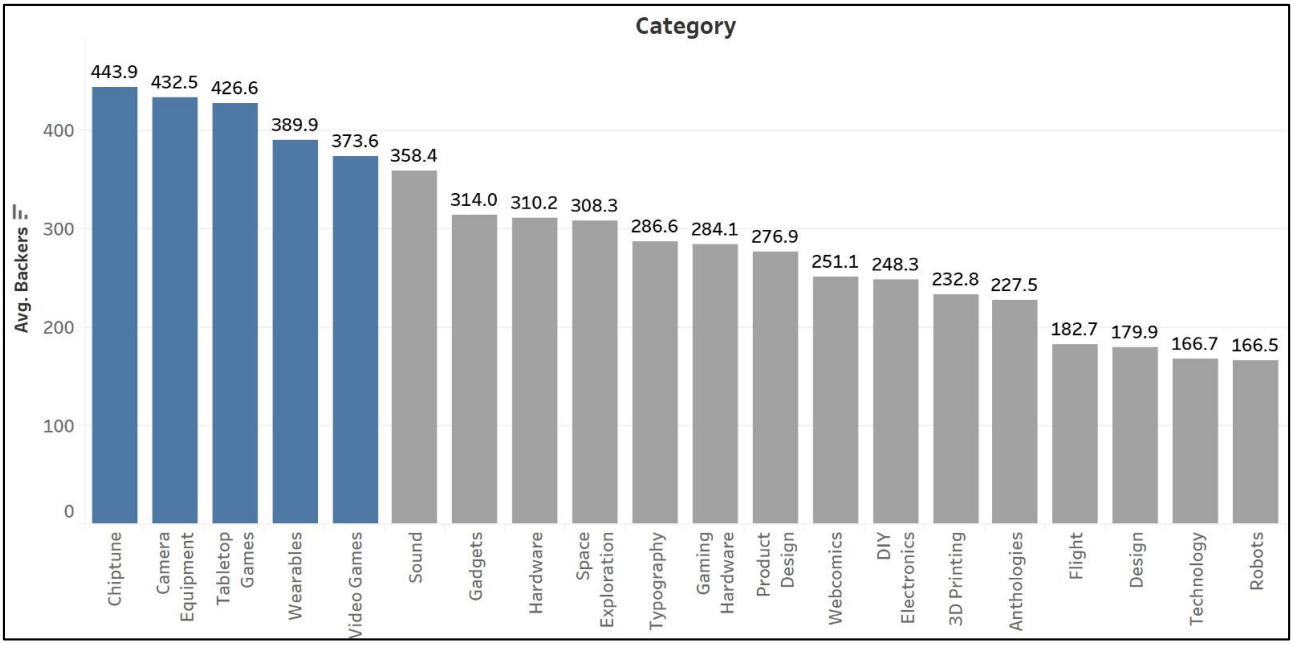
Project Backers: Backers while looking to invest in "Technology" projects should check the project description to see if it is clear, have a sense if the goal amount posted is reasonable to the project objective and fund such projects with caution.

Insight 3: Though technology projects have received the highest pledge amount from backers and also the third Main Category with highest number of backers, not all technology projects are of interests to backers.

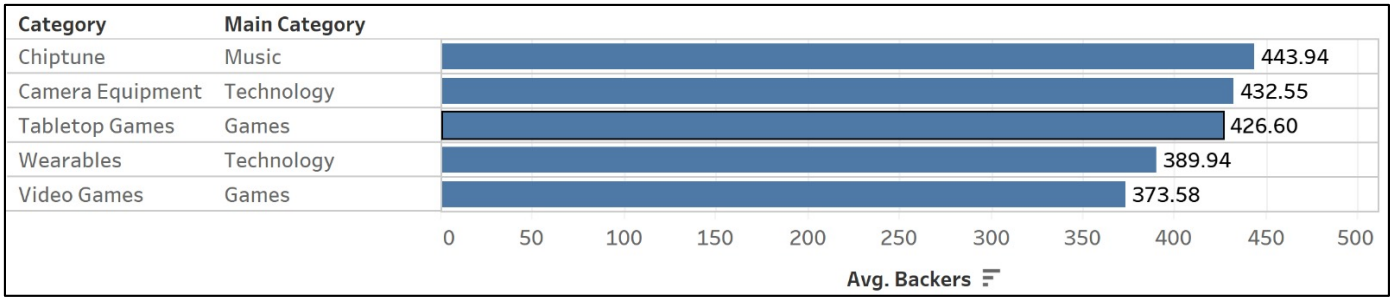
Objective 4: To identify the relation between the average number of backers and categories to draw insightful conclusions.

Analysis 1: Overall identification of project categories with highest number of average backers

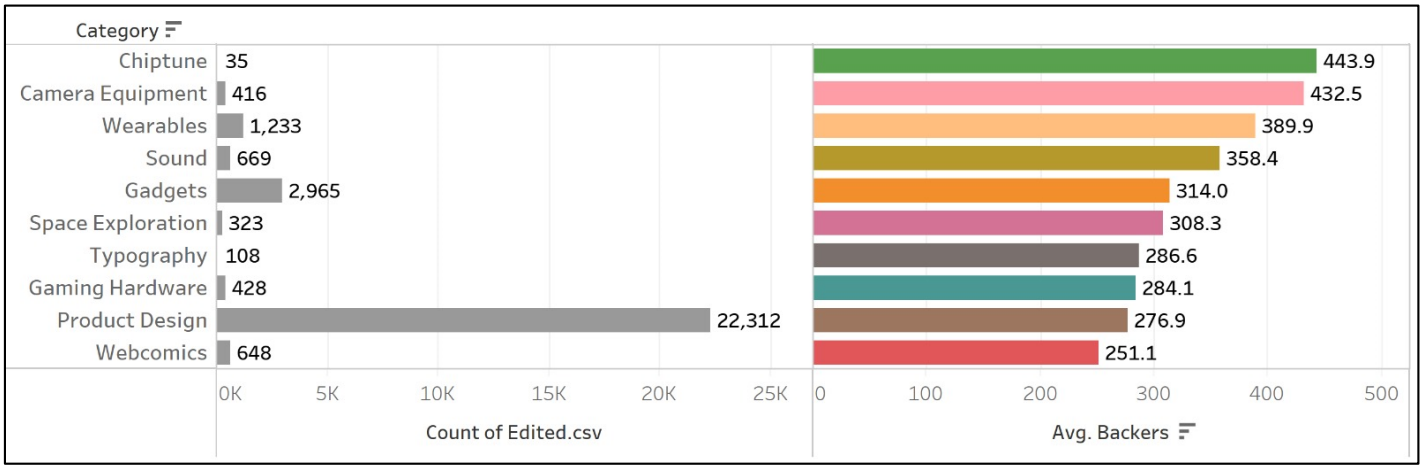
In the EDA performed prior to the objectives, we identified the top 20 categories with the greatest number of average backers. To streamline our analysis further, we consider the top 5 categories namely, Chiptune, Camera Equipment, Tabletop Games, Wearables and Video Games.



To understand why these particular categories, have the number shown above, we identify the main categories they belong to as shown below,



From the above plot, we see that the main categories are those that are mostly in development phase with numerous potential improvements in the future. This can be one of the main reasons as to why projects in such categories have a greater number of average backers. However, we have to check if this number correlates with the number of projects being released.

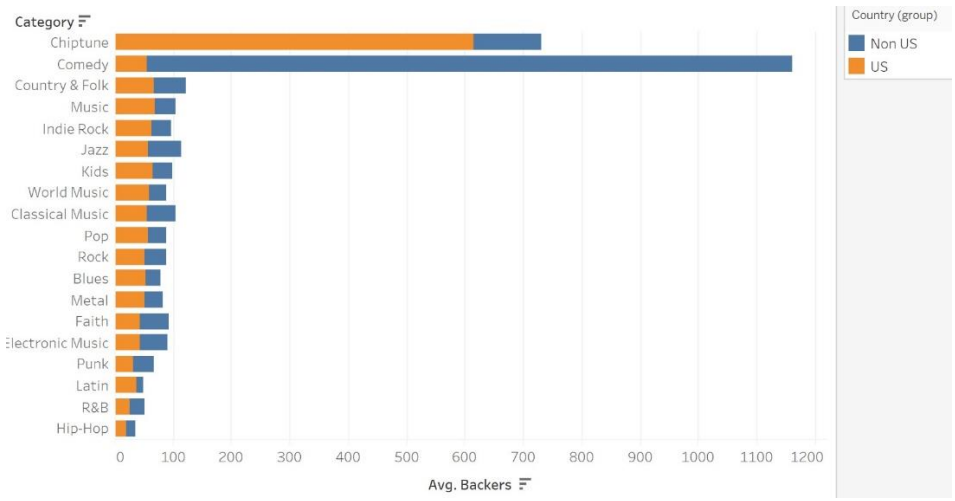


From the above plot, we observe that the results deviate from what we expected.

For example, though Chiptune has the greatest number of average backers, the number of Chiptune projects are significantly less in number. This means that such projects require more promotion in the form of online ads, digital marketing campaigns etc.
On the contrary, though the category "Product Design" has good number of projects in relation to its high amount of average backers, it becomes vital for the kick starter platform to analyze and approve quality projects to keep up the motivation of backers in these categories.

Analysis 2: Further analysis of the music category "Chiptune"

From the previous analysis, we observed that "Chiptune" had the highest number of average backers. We will now further inspect this category to check for some better insights. The below plot shows how influential "Chiptune" is in the US compared to other categories.



From the above plot, it is evident that the "Chiptune" category is immensely popular in the US compared to the other countries. Next, we check how popular is the Music category among the backers with and without Chiptune to see if Chiptune plays a significant role.

Hypothesis Testing:

Before conducting the hypothesis, we have two groups. One has all the observations in the music category with Chiptune and the second group has all the observations excluding Chiptune.

H0: There is no significant difference in the average number of backers between the two groups.
H1: There is a significant difference in the average number of backers between the two groups.

Here, we conduct the Welch two sample t-test to check if "Chiptune" is significant for the music category.

```
        Welch Two Sample t-test

data:  Chiptune.df$backers by Chiptune.df$Groups
t = 0.21566, df = 99298, p-value = 0.8293
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.220594  2.769688
sample estimates:
   mean in group With Chiptune mean in group Without Chiptune
                       54.51765                       54.24311
```

From the above results, we see that there isn't much difference in the averages between the two groups. In addition to this, the p-value is not significant too. Thus, we can conclude that though Chiptune has the greatest number of average backers, it does not have a significant impact on the Music category as expected.

Results and Benefits:

Kick starter Platform: Observing the average number of backers for each category, the Kickstarter platform should try to promote more of these types of projects which are preferred by backers in order to generate better revenue.

Project Backers: By putting up more projects from such categories, the backers' interest is increased. Thus, they will tend to sponsor such projects which have a tendency to be fruitful in the future.

Insight 4: ***Though projects in the categories such as Chiptune and Camera Equipment have received the highest average number of backers, they should be promoted more to generate higher profits.***

Appendix:

The exploratory analysis was done using analytical tools such as Python, Excel and Tableau. The statistical tests were executed using R and the code for the same is as in Appendix A. The Python code for the EDA plots is as in Appendix B. The dataset modification was done in Excel as in Appendix C (attached separately). Plots and other visualizations were done in Tableau as in Appendix D (attached separately).

For the analysis done with Excel and Tableau, the steps are as below.

- In excel the column "launched" was converted to short date to get the date without time. Then by subtracting this date from "deadline" we obtain the column "period to pledge" in days.
- This updated excel is saved as a CSV file and was used to create charts in Tableau.

Data preparation for Insight 1:

1. The dataset was filtered to have observations with projects that failed.
2. The failed projects were categorized into type A and type B projects depending on whether the pledge amount was greater than the requested goal amount or not. If the pledge amount is greater than it is named as Type A projects else Type B.
3. A column with 4% commission based on the goal amount and a column with total required amount (goal amount + 4% commission) was calculated and inserted.
4. By dividing the column "amount pledged" by the column "period to pledge" we obtain average amount pledged per day.
5. Then the column "total amount required with commission" is dividing by "pledge amt per day" to obtain "Total time required for pledging". Then by subtracting this column from "period to pledge" we would arrive at "More days required" for the project to become successful.
6. Then by using this column the revenue that would be generated for the Kickstarter platform by increasing the pledging time was calculated. A sample of the table is as below:

	ID	name	category	main_category	Period_to_pledge	state	backers	country	usd_pledged_real	usd_goal_real	Project Type	Amt required with 4% commision	4% commision	Pledge amt per day	Total Time reqd to pledge	More day reqd	Amt/ backer	total backers reqd	Addition backers needed
1																			
3	1000003930	Greeting From Earth: ZGAC Arts (	Narrative Film	Film & Video	60	failed	15	US	2421	30000	Type B	31200.00	1200.00	40.35	773.23	713.23	161.40	193.31	178.31
4	1000004038	Where is Hank?	Narrative Film	Film & Video	45	failed	3	US	220	45000	Type B	46800.00	1800.00	4.89	9572.73	9527.73	73.33	638.18	635.18
5	1000007540	ToshiCapital Rekordz Needs Help	Music	Music	30	failed	1	US	1	5000	Type B	5200.00	200.00	0.03	156000.00	155970.00	1.00	5200.00	5199.00
6	1000030581	Chaser Strips. Our Strips make Sh	Drinks	Food	45	failed	40	US	453	25000	Type B	26000.00	1000.00	10.07	2582.78	2537.78	11.33	2295.81	2255.81
10	1000064368	Survival Rings	Design	Design	30	failed	11	US	664	2500	Type B	2600.00	100.00	22.13	117.47	87.47	60.36	43.07	32.07
11	1000064918	The Beard	Comic Books	Comics	30	failed	16	US	395	1500	Type B	1560.00	60.00	13.17	118.48	88.48	24.69	63.19	47.19

Appendix A

Appendix- SAS Project Tests and Analysis

Sinduja, Kowshik

7/5/2020

a. Load the packages:

```
if(!require("pacman")) install.packages("pacman")
```

```
## Loading required package: pacman
```

```
pacman::p_load(data.table, ggplot2, dplyr, tidyverse, forecast, gplots, GGall
y, mosaic,multcomp,scales, mosaic, mapproj, mlbench)
search()
```

```
##  [1] ".GlobalEnv"         "package:mlbench"    "package:mapproj"
##  [4] "package:maps"       "package:scales"     "package:multcomp"
##  [7] "package:TH.data"    "package:MASS"       "package:survival"
## [10] "package:mvtnorm"    "package:mosaic"     "package:Matrix"
## [13] "package:mosaicData" "package:ggformula"  "package:ggstance"
## [16] "package:lattice"    "package:GGally"     "package:gplots"
## [19] "package:forecast"   "package:forcats"    "package:stringr"
## [22] "package:purrr"      "package:readr"      "package:tidyr"
## [25] "package:tibble"     "package:tidyverse"  "package:dplyr"
## [28] "package:ggplot2"    "package:data.table" "package:pacman"
## [31] "package:stats"      "package:graphics"   "package:grDevices"
## [34] "package:utils"      "package:datasets"   "package:methods"
## [37] "Autoloads"          "package:base"
```

b. Read in the data from “Proj Pledging Time.csv”:

```
Pledge_Time.df <- read.csv("Proj Pledging Time.csv")
```

c. Setting to dataframe:

```
Pledge_Time.dt <- setDT(Pledge_Time.df)
```

d. Structure of data:

```
dim(Pledge_Time.dt)
```

```
## [1] 375093      3
```

```
str(Pledge_Time.dt)
```

```
## Classes 'data.table' and 'data.frame':   375093 obs. of  3 variables:
##  $ name            : Factor w/ 372252 levels "","\177Not Twins - New EP! \
"The View from Down Here\"",..: 204143 290690 184347 340808 200593 73982 3032
97 196262 232763 251165 ...
##  $ state           : Factor w/ 5 levels "canceled","failed",..: 4 4 4 4 4
4 4 4 4 4 ...
```

```
##  $ Period_to_pledge: int  34 19 30 27 14 35 19 29 29 29 ...
##  - attr(*, ".internal.selfref")=<externalptr>

summary(Pledge_Time.dt)

##                          name              state        Period_to_pledge
##  #NAME?                   :    43   canceled  : 38777   Min.   :    0.00
##  New EP/Music Development:    15   failed    :197718   1st Qu.:   29.00
##  Canceled (Canceled)      :    13   live      :  2799   Median :   30.00
##  Music Video              :    11   successful:133953   Mean   :   33.87
##  N/A (Canceled)           :    11   suspended :  1846   3rd Qu.:   37.00
##  Cancelled (Canceled)     :    10                       Max.   :16739.00
##  (Other)                  :374990
```

Anova 1 - comparison of mean values of Pledging time within levels of "Status":

```
levels(Pledge_Time.dt$state)

## [1] "canceled"   "failed"     "live"       "successful" "suspended"

Pledge_Time.dt$state <- factor(Pledge_Time.dt$state, levels=c("successful","canceled","failed","live","suspended"))
Anova_analysis1 <- aov(Pledge_Time.dt$Period_to_pledge ~ Pledge_Time.dt$state)
summary(Anova_analysis1)

##                           Df    Sum Sq Mean Sq F value Pr(>F)
## Pledge_Time.dt$state       4 1.651e+06  412771   94.23 <2e-16 ***
## Residuals            375088 1.643e+09    4381
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

##Dunnett Test 2 - COMPARE MEAN VALUE OF PLEDGE TIME TO STATUS TYPE- "SUCCESSFUL" PROJECTS

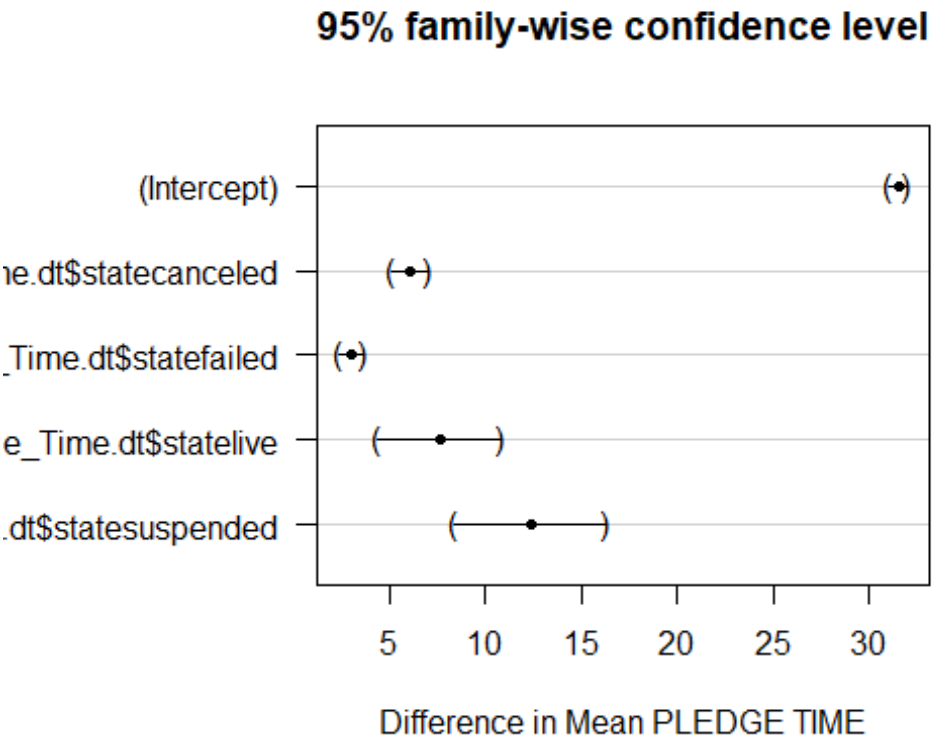
```
Dunnet <- glht(Anova_analysis1, clinfct =mcp(Type="Dunnett"))
summary(Dunnet)

##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: aov(formula = Pledge_Time.dt$Period_to_pledge ~ Pledge_Time.dt$state)
##
## Linear Hypotheses:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept) == 0                      31.5194     0.1808 174.296   <1e-08 **
*
## Pledge_Time.dt$statecanceled == 0      6.1266     0.3817  16.052   <1e-08 **
*
## Pledge_Time.dt$statefailed == 0        3.0422     0.2342  12.989   <1e-08 **
*
## Pledge_Time.dt$statelive == 0          7.6507     1.2640   6.053   <1e-08 **
```

```
*
## Pledge_Time.dt$statesuspended == 0  12.3582     1.5510   7.968   <1e-08 **
*
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

##Dunnett Test PLOT 1

```
if(require(multcomp)){     # Use the multcomp library
  myDunnett  <- glht(Anova_analysis1, clinfct = mcp(Type = "Dunnett"))
  summary(myDunnett)
  confint(myDunnett,level=.95)
  opar <- par(no.readonly=TRUE)  # Save current graphics parameter settings
  par(mar=c(4.1,8.1,4.1,1.1)) # Change margins
  plot(myDunnett,
       xlab="Difference in Mean PLEDGE TIME")
  par(opar)  # Restore original graphics parameter settings
}
```



Anova 2 - comparison of mean values of Goal amount within levels of "Status":

```
GoalAmt_USD.df  <- read.csv("Proj goal.csv")
GoalAmt_USD.dt <- setDT(GoalAmt_USD.df)
levels(GoalAmt_USD.dt$state)
```

```
## [1] "canceled"   "failed"     "live"       "successful" "suspended"
```

```
GoalAmt_USD.dt$state <- factor(GoalAmt_USD.dt$state, levels=c("successful","c
anceled","failed","live","suspended"))
Anova_analysis2 <- aov(GoalAmt_USD.dt$usd_goal_real ~ GoalAmt_USD.dt$state)
summary(Anova_analysis2)
```

```
##                           Df   Sum Sq   Mean Sq F value Pr(>F)
## GoalAmt_USD.dt$state       4 3.02e+14 7.549e+13   56.29 <2e-16 ***
## Residuals             375088 5.03e+17 1.341e+12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

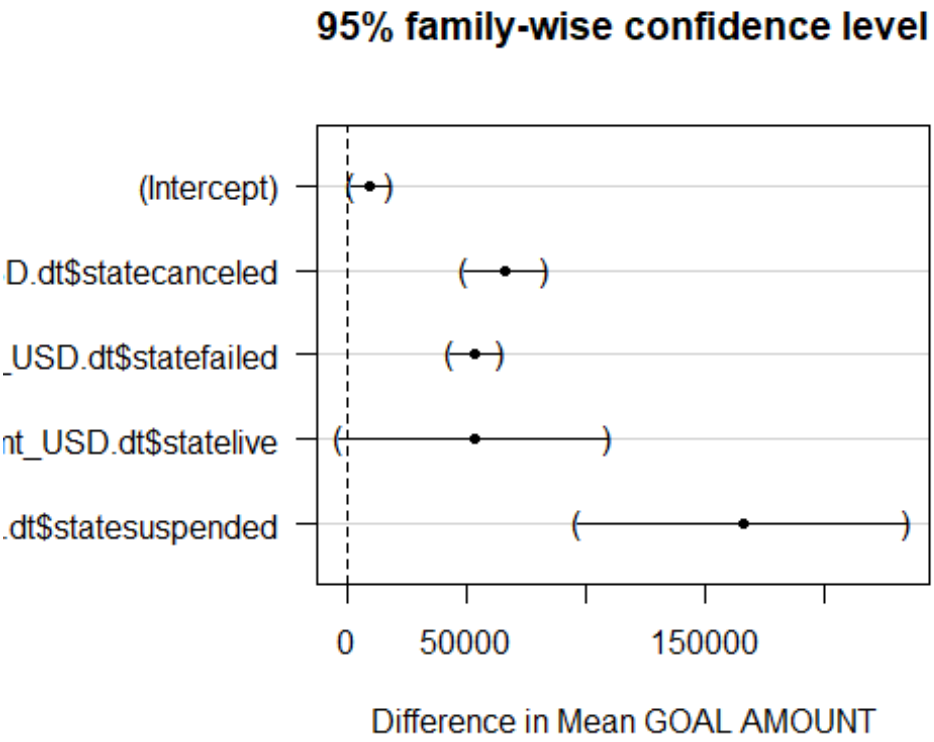
##Dunnett Test 2 - COMPARE MEAN VALUE OF GOAL AMOUNT STATUS TYPE- "SUCCESSFUL" PROJECTS

```
Dunnet <- glht(Anova_analysis2, clinfct =mcp(Type="Dunnett"))
summary(Dunnet)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: aov(formula = GoalAmt_USD.dt$usd_goal_real ~ GoalAmt_USD.dt$state)
##
## Linear Hypotheses:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept) == 0                        9532       3164   3.013   0.0122 *
## GoalAmt_USD.dt$statecanceled == 0      66051       6678   9.891   <0.001 **
*
## GoalAmt_USD.dt$statefailed == 0        53643       4098  13.089   <0.001 **
*
## GoalAmt_USD.dt$statelive == 0          53003      22117   2.396   0.0733 .
## GoalAmt_USD.dt$statesuspended == 0    165676      27139   6.105   <0.001 **
*
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

##Dunnett Test PLOT 2

```
if(require(multcomp)){     # Use the multcomp library
  myDunnett  <- glht(Anova_analysis2, clinfct = mcp(Type = "Dunnett"))
  summary(myDunnett)
  confint(myDunnett,level=.95)
  opar <- par(no.readonly=TRUE)  # Save current graphics parameter settings
  par(mar=c(4.1,8.1,4.1,1.1)) # Change margins
  plot(myDunnett,
       xlab="Difference in Mean GOAL AMOUNT")
  par(opar)  # Restore original graphics parameter settings
}
```



T Test 1 - comparison of mean values of number of backers for success and failre projects for Main Category Technology:

```
TechProj.df  <- read.csv("Tech Proj.csv")
TechProj.dt <- setDT(TechProj.df)
t.test(TechProj.df$backers ~ TechProj.df$state, alternative = "two.sided")

##
##  Welch Two Sample t-test
##
## data:  TechProj.df$backers by TechProj.df$state
## t = -22.923, df = 6436.6, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -757.0577 -637.7749
## sample estimates:
##     mean in group failed mean in group successful
##                 20.26242                717.67874
```

T Test 2 - comparison of mean values of number of backers for Music projects with and without Chiptune category

```
Chiptune.df  <- read.csv("Chiptune.csv")
Chiptune.dt <- setDT(Chiptune.df)
t.test(Chiptune.df$backers ~ Chiptune.df$Groups, alternative = "two.sided")

##
##  Welch Two Sample t-test
```

```
##
## data:  Chiptune.df$backers by Chiptune.df$Groups
## t = 0.21566, df = 99298, p-value = 0.8293
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.220594  2.769688
## sample estimates:
##    mean in group With Chiptune mean in group Without Chiptune
##                       54.51765                       54.24311
```

In [16]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from scipy import stats
from scipy.stats import chi2

import plotly.tools as tls
import plotly
import plotly.offline as py
from plotly.offline import init_notebook_mode, iplot, plot
import plotly.graph_objs as go
init_notebook_mode(connected=True)
import plotly.graph_objs as go
import plotly.figure_factory as ff
from plotly import tools
py.init_notebook_mode(connected=True)

pd.options.mode.chained_assignment = None
pd.options.display.max_columns = 999
pd.options.display.float_format = '{:.2f}'.format

import warnings
from collections import Counter
```

In [2]:
```
df_kick = pd.read_csv("C:/Users/Kowshik Raj/OneDrive/Desktop/UTD/Predictive Analytics using SAS/Projects/Pro
df_kick = df_kick.sample(10000, random_state=42).reset_index().drop('index', axis=1)
```

In [3]:
```
df_kick.head()
```

Out[3]:

	ID	name	category	main_category	currency	deadline	goal	launched	pledged	state	backers	c
0	1963584816	Lilly & The Flight	Shorts	Film & Video	USD	10/27/2017	2500.0	9/28/2017 16:01	2702.0	successful	17	
1	702213029	Barbie Doomsday Shelters (Canceled)	Games	Games	USD	6/2/2014	100000.0	5/3/2014 3:07	0.0	canceled	0	
2	1144326745	Apple - Chore Commissioner	Technology	Technology	USD	1/5/2014	900.0	12/16/2013 7:20	34.0	failed	6	
3	461710720	Crime and Consequences	Webseries	Film & Video	USD	9/11/2012	10500.0	7/22/2012 6:40	11170.0	successful	82	
4	619431407	a story for "class artists" everywhere (re-lau...	Children's Books	Publishing	USD	4/26/2013	2100.0	2/25/2013 6:04	2125.0	successful	27	

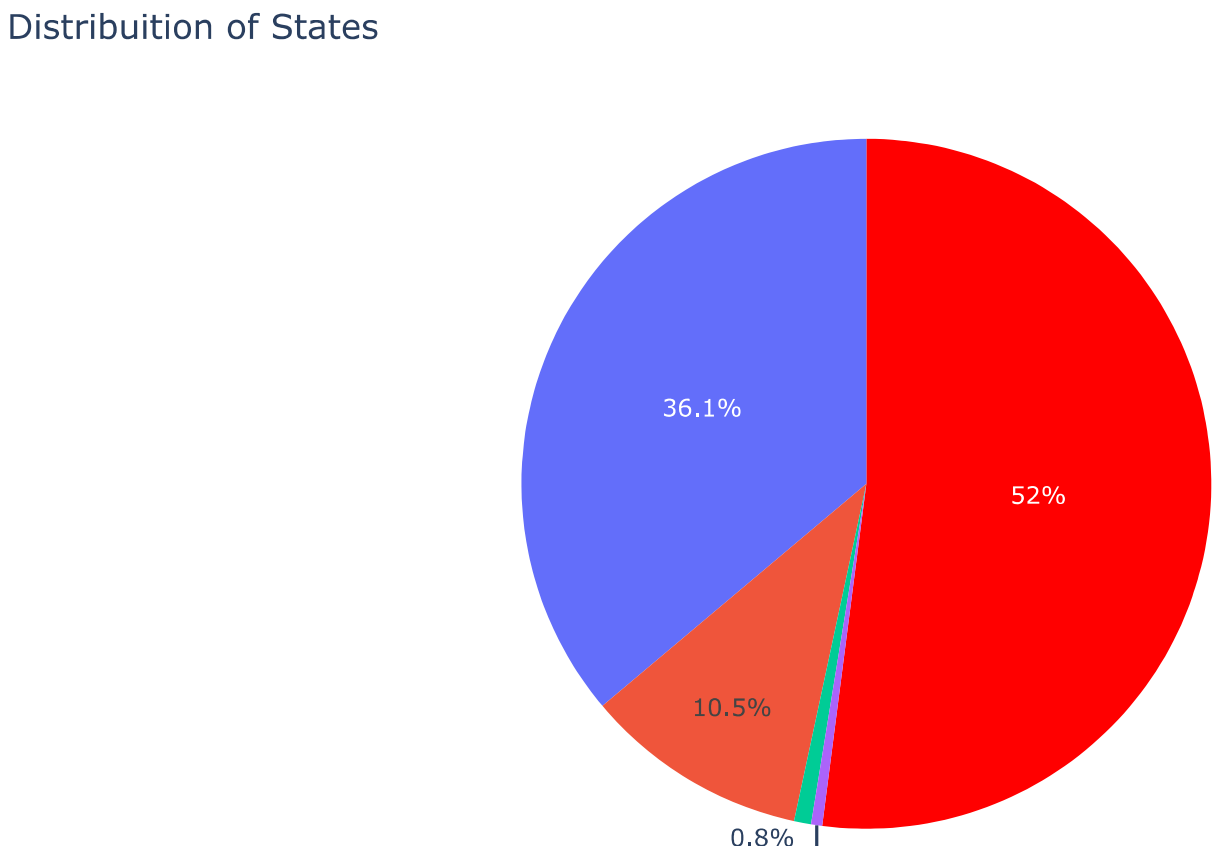
In [4]:
```
state = round(df_kick["state"].value_counts() / len(df_kick["state"]) * 100,2)

labels = list(state.index)
values = list(state.values)

trace1 = go.Pie(labels=labels, values=values, marker=dict(colors=['red']))

layout = go.Layout(title='Distribuition of States', legend=dict(orientation="h"));

fig = go.Figure(data=[trace1], layout=layout)
iplot(fig)
```



In [5]:
```
df_kick = df_kick.loc[df_kick['state'].isin(['failed','successful'])]
```

In [6]:

```
df_failed = df_kick[df_kick["state"] == "failed"].sample(10000, replace=True)
df_sucess = df_kick[df_kick["state"] == "successful"].sample(10000, replace=True)

#First plot
trace0 = go.Histogram(
    x= np.log(df_kick.usd_goal_real + 1),
    histnorm='probability', showlegend=False,
    xbins=dict(
        start=-5.0,
        end=19.0,
        size=1),
    autobiny=True)

#Second plot
trace1 = go.Histogram(
    x = np.log(df_kick.usd_pledged_real + 1),
    histnorm='probability', showlegend=False,
    xbins=dict(
        start=-1.0,
        end=17.0,
        size=1))

# Add histogram data
failed = np.log(df_failed['usd_goal_real']+1)
success = np.log(df_sucess["usd_goal_real"]+1)

trace3 = go.Histogram(
    x=failed,
    opacity=0.60, nbinsx=30, name='Goals Failed', histnorm='probability'
)
trace4 = go.Histogram(
    x=success,
    opacity=0.60, nbinsx=30, name='Goals Sucessful', histnorm='probability'
)


data = [trace0, trace1, trace3, trace4]
layout = go.Layout(barmode='overlay')

#Creating the grid
fig = plotly.tools.make_subplots(rows=2, cols=2, specs=[ [{'colspan': 2}, None], [{}, {}]],
                                 subplot_titles=('Failed and Sucessful Projects',
```

```
                                        'Goal','Pledged'))

#setting the figs
fig.append_trace(trace0, 2, 1)
fig.append_trace(trace1, 2, 2)
fig.append_trace(trace3, 1, 1)
fig.append_trace(trace4, 1, 1)

fig['layout'].update(title="Distribuitions",
                     height=500, width=900, barmode='overlay')
iplot(fig)
```

```
C:\Users\Kowshik Raj\anaconda3\lib\site-packages\plotly\tools.py:465: DeprecationWarning:

plotly.tools.make_subplots is deprecated, please use plotly.subplots.make_subplots instead
```

In [7]:
```
stat, p = stats.shapiro(np.log(df_kick['usd_goal_real']+1).sample(500, random_state=42))

print("Shapiro stat:", stat)
print("P-value: ", p)
if p >= .01:
    print('Normal Distribution')
else:
    print("Non-Normal Distribution")
```

```
Shapiro stat: 0.9965996146202087
P-value:  0.37331530451774597
Normal Distribution
```

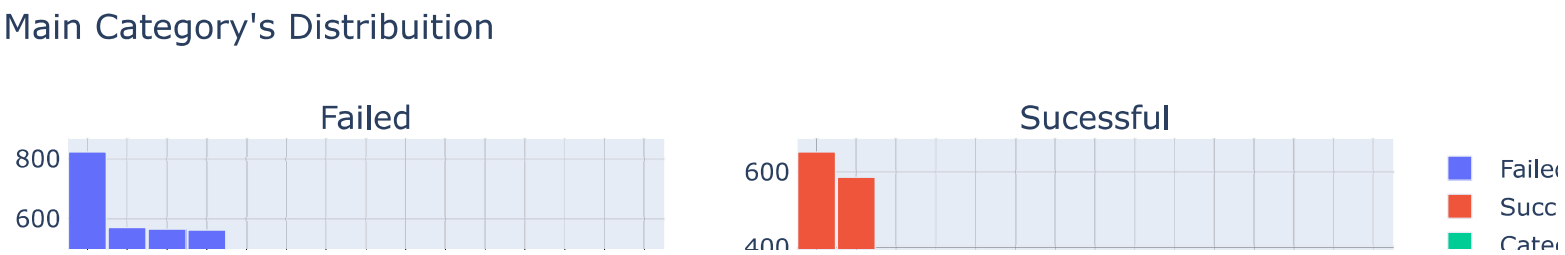
In [9]:
```
main_cats = df_kick["main_category"].value_counts()
main_cats_failed = df_kick[df_kick["state"] == "failed"]["main_category"].value_counts()
main_cats_sucess = df_kick[df_kick["state"] == "successful"]["main_category"].value_counts()
```

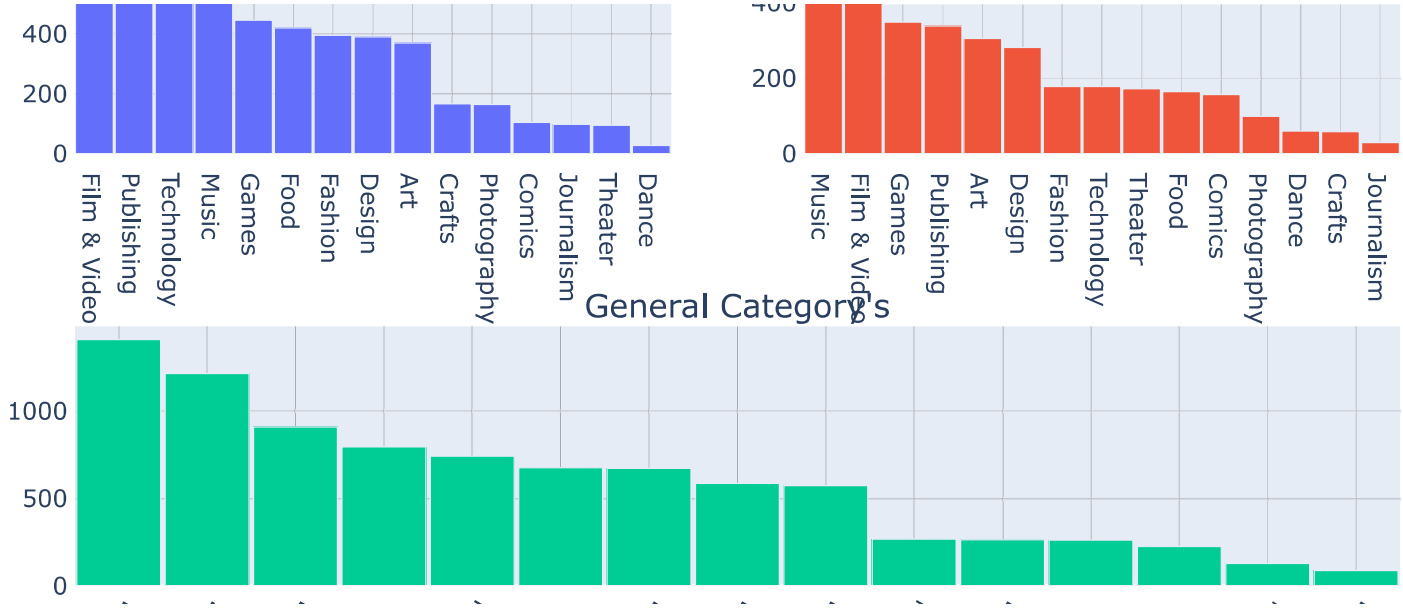
In [10]:
```
#First plot
trace0 = go.Bar(
    x=main_cats_failed.index,
    y=main_cats_failed.values,
    name="Failed Categories"
)
#Second plot
trace1 = go.Bar(
    x=main_cats_sucess.index,
    y=main_cats_sucess.values,
    name="Success Categories"
)
#Third plot
trace2 = go.Bar(
    x=main_cats.index,
    y=main_cats.values,
    name="Categories Distribuition"
)

#Creating the grid
fig = tls.make_subplots(rows=2, cols=2, specs=[[{}, {}], [{'colspan': 2}, None]],
                          subplot_titles=('Failed','Sucessful', "General Category's"))

#setting the figs
fig.append_trace(trace0, 1, 1)
fig.append_trace(trace1, 1, 2)
fig.append_trace(trace2, 2, 1)

fig['layout'].update(showlegend=True,
                     title="Main Category's Distribuition",
                     bargap=0.05)
iplot(fig)
```





In [14]:

```
def chi2_test(col, prob=.95):
    stat, p, dof, expected = stats.chi2_contingency((pd.crosstab(df_kick[col[0]],
                                                                 df_kick[col[1]]
                                                                )))
    print("CHI-SQUARED TEST: ")
    # calculating the value to compare with chi2 statistic
    critical = stats.chi2.ppf(prob, dof)
    print(f'dof={dof}, probability={round(prob,3)}, critical={round(critical,5)}, stat={round(stat,5)}')
    print("Accept or Reject H0: ")
    # interpret test statistic
    if abs(stat) >= critical:
        print('Dependent (reject H0)')
    else:
        print('Independent (fail to reject H0)')
```

In [15]:
```
chi2_test(['state','main_category'])
```

```
CHI-SQUARED TEST:
dof=14, probability=0.95, critical=23.68479, stat=413.12072
Accept or Reject H0:
Dependent (reject H0)
```