**BUAN 6337 Predictive Analytics using SAS**

**Project 1**

**AADITH NARAYAN RAVISHANKAR / AXR180085**

Introduction

The Kickstarter platform when observed has three POVs: -

* Kickstarter
* Project creator
* Backers

The Kickstarter platform serves as a medium for project hosts to host their project creators to host their projects and raise funds to launch their projects. The Kickstarter platform looks to be mainly based on the USA but have their presence in other countries as well mainly in Europe.

The funds are provided by backers who invest in the project creator’s projects based on their viability or feasibility. A project creator can promise a return on investment to the backers if they invest a particular range of money into the project.

The project creators use the Kickstarter platform to host their projects. The main advantage of using Kickstarter for creators is that there is no fee involved in hosting a project. Kickstarter charges a fee of 5% only if the pledged amount is reached.

As with all the datasets, there are some inconsistencies in the dataset. There are certain observations where the goal amount has been reached with 0 backers and the project is deemed successful which completely goes against the idea of how the platform works. Also, for those records the country is not provided which brings me to the understanding that the country is an important parameter to filter out garbage values.

The insights are according to the three POVs of the Kickstarter platform: Kickstarter, Project creator and Backers. The questions which are answered in this report according to the different POVs are as follows:

Kickstarter:

* Which categories of projects are most hosted on the platform?
* Which countries is the platform most utilized on?

Project Creator:

* Is there seasonality when setting deadlines which leads to more successful projects?
* Which categories of projects has most potential to become successful?
* What is the optimal goal amount to set to attract backers for raising funds?

Backers:

* Does the number of backers have an influence on the project becoming successful or not?
* Do backers support high scale projects which have a higher goal amount versus low scale projects? Or is the investment unbiased with respect to the scale of the project?

To get a sense of the state column which deems whether a project is successful or not, the following pie chart is attached. We can see that around 35% of the projects are successful and around 52% of the projects have failed. So failure has a bigger impact on the dataset.

Executive Summary

Kickstarter:

* The category which was most hosted on the platform was that of “Film and Video”. The second highest was “Music”.
* The country which the platform was most utilized on was the United States. The second most active country was Great Britain.

Project Creator:

* We found that the setting of deadlines had seasonality around a year. Deadlines which were set in the second half of the year tended to be more successful than the ones which were set in the first half of the year.
* Within the categories of projects, the most successful one was “Dance”. The second most successful one was “Theatre”
* The optimal goal amount for successful projects is around 3800$

Backers:

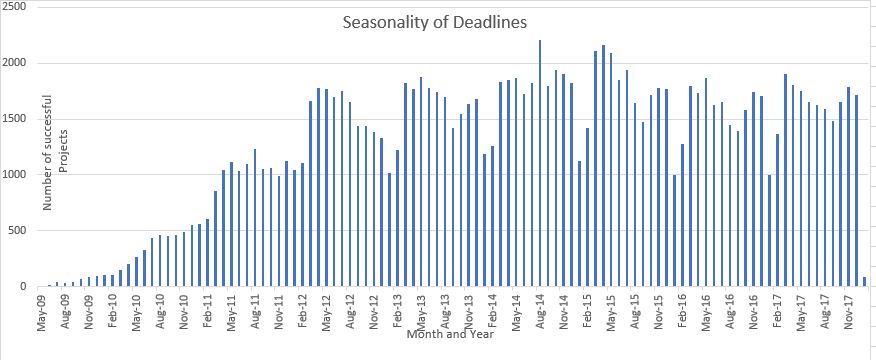
* There is a difference in the number of backers for successful projects and failed ones
* There is a difference in the number of backers with respect to the scale of project according to the goal amount.

Insights

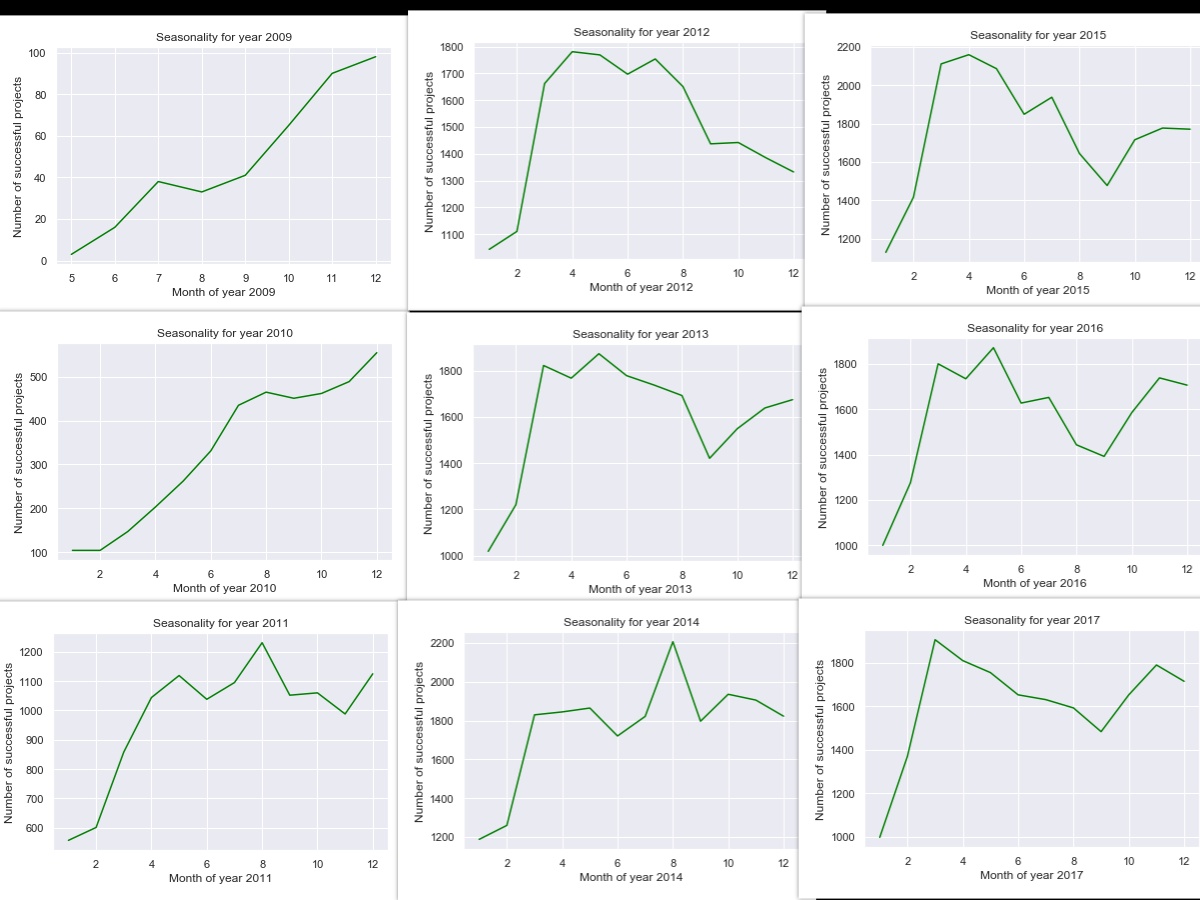
# Project Creator POV:

## Question of Interest: Is there seasonality when setting deadlines which leads to more successful projects?

This intent behind this question is to see if deadlines have any relation to having successful projects. Projects are posted all-round the year but a peak in certain months is common when dealing with annual time patterns. If seasonality exists, it provides a hint to project creators that setting the deadlines in certain months would lead to potential successful projects.



When looking at the above bar graph, it looks like for each year, there is a common pattern which is evident. Therefore, it is worth exploring more in depth to find out the particular months which have a peak in number of successful projects.



These are the list of scatterplots derived for each year by plotting the month versus the number of successful projects. The years 2009 has only last 7 months and the year 2018 only has the month of January considered. So, both those years can be exempted from the analysis.

If we visualize the graphs, in most of the cases, there is a peak in the months of March and April and also there is a peak of successful projects around the month of July which lead to a spike in the number of cases from then on. Therefore, the months of March, April and July can be considered to be important months for setting deadlines for projects to become potentially successful.

## Question of Interest: Which categories of projects have more potential to become successful?

This insight gives a new project creator to set his direction to get his/her project funded and to launch it using the Kickstarter platform.

The following table gives the summary of the categories and the number of projects under that which were successful.

|  |  |  |  |
| --- | --- | --- | --- |
| Main\_category | count success | count main | Conversion rate |
| Art | 11510 | 28153 | 40.88374241 |
| Comics | 5842 | 10819 | 53.99759682 |
| Crafts | 2115 | 8809 | 24.0095357 |
| Dance | 2338 | 3767 | 62.06530396 |
| Design | 10550 | 30068 | 35.08713583 |
| Fashion | 5593 | 22812 | 24.51779765 |
| Film & Video | 23622 | 62730 | 37.65662363 |
| Food | 6085 | 24602 | 24.73376148 |
| Games | 12518 | 35230 | 35.53221686 |
| Journalism | 1012 | 4755 | 21.28286015 |
| Music | 24195 | 49680 | 48.70169082 |
| Photography | 3305 | 10778 | 30.6643162 |
| Publishing | 12300 | 39412 | 31.2087689 |
| Technology | 6434 | 32566 | 19.75680157 |
| Theater | 6534 | 10912 | 59.87903226 |

Main\_category- Category of the project

Count success- Number of successful projects

Count main- Number of projects

Conversion rate- The probability of projects under that category to become successful.

We see that the category which has the greatest number of successful projects is “Film & Video” but we can see that its conversion rate is 37% which is pretty low. So we see that the number of successful projects as a factor is deceiving. The category with the greatest conversion rate is Dance and hence is considered the go-to project category for new project creators. Second in line is Theatre.

## Question of Interest: What is the optimal goal amount to set to attract backers for raising funds?

This insight is important because it gives new project creators to decide on the scale of their project. Whenever a project creator is hosting a project, he has to make sure that either he has a decent amount of backers to invest in his project and also for the goal amount to be reasonable for launch.

The following descriptive statistics gives us a decent picture of the range of goal amount for successful projects:

Descriptive statistics of “usd\_goal\_real” column:

|  |  |
| --- | --- |
| count | 1.339530e+05 |
| mean | 9.532308e+03 |
| std | 2.796064e+04 |
| min | 1.000000e-02 |
| 25% | 1.303670e+03 |
| 50% | 3.837890e+03 |
| 75% | 1.000000e+04 |
| max | 2.015609e+06 |

We consider the usd\_goal\_real column because the amounts are standardized to USD.

The lower limit range of the goal amount is from 0.01$ to 1303$. We see that the median amount for successful projects is around 3800$. On the other hand the upper limit of goal amount is between 10000$ to 2015609$. Anything above this upper limit range might be considered an outlier and would not be advised to set.

The optimal amount can be set as around 3800$ for a project to be successful.

# Kickstarter POV:

## Question of Interest: Which categories of projects are most hosted on the platform?

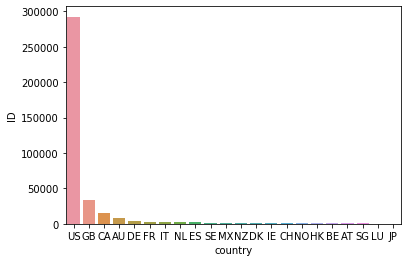
This insight is important to the platform to base its marketing campaign on. In this world of targeted ads where advertisements are focused on people who have similar interests, this insight provides a highly effective focus to target the right audience.

|  |  |
| --- | --- |
| Main Category | Number of Projects |
| Film & Video | 62730 |
| Music | 49680 |
| Publishing | 39412 |
| Games | 35230 |
| Technology | 32566 |
| Design | 30068 |
| Art | 28153 |
| Food | 24602 |
| Fashion | 22812 |
| Theater | 10912 |
| Comics | 10819 |
| Photography | 10778 |
| Crafts | 8809 |
| Journalism | 4755 |
| Dance | 3767 |

As we can see, the greatest number of projects hosted on the platform is of Film&Video. The second highest is Music. Therefore, focusing their marketing campaigns on Movie and music enthusiasts would increase the amount of activity and reception of the Kickstarter platform.

## Question of Interest: Which country is the platform most utilized on?

This question might seem simple to analyze and answer but the impact of the answer to this is huge in a marketing perspective. The platform was started in United States but the growth seems to be worldwide. So it is essential to know where to focus marketing efforts to attract more projects and hence more revenue to keep the platform running for the long term.



|  |  |
| --- | --- |
| country | Count of projects |
| US | 292621 |
| GB | 33672 |
| CA | 14756 |
| AU | 7839 |
| DE | 4171 |
| FR | 2939 |
| IT | 2878 |
| NL | 2868 |
| ES | 2276 |
| SE | 1757 |
| MX | 1752 |
| NZ | 1447 |
| DK | 1113 |
| IE | 811 |
| CH | 761 |
| NO | 708 |
| HK | 618 |
| BE | 617 |
| AT | 597 |
| SG | 555 |
| LU | 62 |
| JP | 40 |

As we can see, the USA has the greatest number of projects which are hosted on the platform. Next in line is Great Britain. Both the countries make up for 87% of projects which are hosted. Therefore, marketing can be focused on these countries to promote the platform.

# Backers POV:

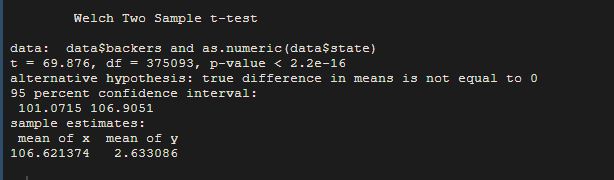
## Question of Interest: Does the number of backers have an influence on the project becoming successful or not?

This question at first glance might be too simple to answer. Let us consider this situation. Two projects A and B have goal amounts 5000$ and 10000$. Both have the same number of backers as 50. Project A reached its goal while project B could not. So it is not just the number of backers which is at play here but the amount of money invested by each backer and the goal amount. So this analysis tells us whether there is a bias in the number of backers with respect to a project’s success or failure or not. Here we considered successful and live states to be successful and the rest to be failures.

In this analysis a two sample T-test is conducted with unequal variances (Welch T-test) to see if the differences between the mean of backers for successful and failed is 0 or not i.e if the means of both the groups are equal. The hypothesis is as follows:

Null Hypothesis H0: The difference between the means are equal to 0

Alternative Hypothesis Ha: The differences between the means are not equal to 0



The Welch two sample T-test gives a p-value of less than 0.01, thus rejecting the null hypothesis. Therefore, the differences between the means are not equal to 0. This implies that the number of backers for successful and failed projects differ significantly.

## Question of Interest: Do backers support high scale projects which have a higher goal amount versus low scale projects? Or is the investment unbiased with respect to the scale of the project?

The intent behind this insight is to see if backers invest in high scale projects vs low scale projects. Investing in low scale projects involves minimum risk and also low return on investment. On the other hand, large scale investments have maximum risk but high ROI. Therefore, we would like to find out the backer behavior when investing in projects. If the backers have a bias for high scale projects or low scale projects over the other.

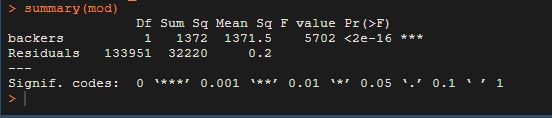
For this analysis, I am binning the usd\_goal\_real column into 4 bins as follows:

1. 0-5000 as 1
2. 5000-100000 as 2
3. 100000-500000 as 3
4. 500000-3000000 as 4

Then im running an ANOVA test to check if there is a difference of means within these bins:

Null Hypothesis H0: µ1 = µ2 = µ3 = µ4 = µ5 = 0

Alternative Hypothesis Ha: At least one of the means is different



We get a p-value of less than 0.01 which rejects the null hypothesis. Therefore, there is a difference in means within these groups which suggests that there is some bias involved with respect to the investments based on the goal amounts.

Appendix

The platforms used to carry out analysis for this project were:

* Python- Jupyter Notebook
* Microsoft Excel
* R- R Studio

The dataset was loaded into Excel and the N,0” characters had to be changed to NULL.

Then the modified dataset was loaded to Python as follows:

df=pd.read\_csv(“DATA.csv”)

# Project Creator POV:

## Question of Interest: Is there seasonality when setting deadlines which leads to more successful projects?

* The overall trend has to be analyzed as to how the successful projects are distributed across the deadline time.
* For this a separate column deadline\_month\_year is created which has the month and year of the deadline dates extracted.
* Separate columns deadline\_month and deadline\_year are created for month and year respectively.
* The dataset is filtered out such that only the records which have successful as the state exists
* The project ID is determined as the primary key for the dataset. Therefore, the count of project ID is segregated against the deadline\_month\_year.
* Those values are plotted as a bar graph in Excel to observe seasonality across all the years.
* Once the overall graph is analyzed, line graphs are plotted for each month of a particular year using the deadline\_month column.

*#To get the bar graph of successful projects for the whole timeline*

*df['deadline\_year']=df['deadline'].transform(lambda x: x[-4:])*

*df['deadline\_month']=df['deadline'].transform(lambda x: x[:2])*

*df['deadline\_month']=df['deadline\_month'].transform(lambda x: x.replace('/',''))*

*df['deadline\_month']=df['deadline\_month'].transform(lambda x: x.strip())*

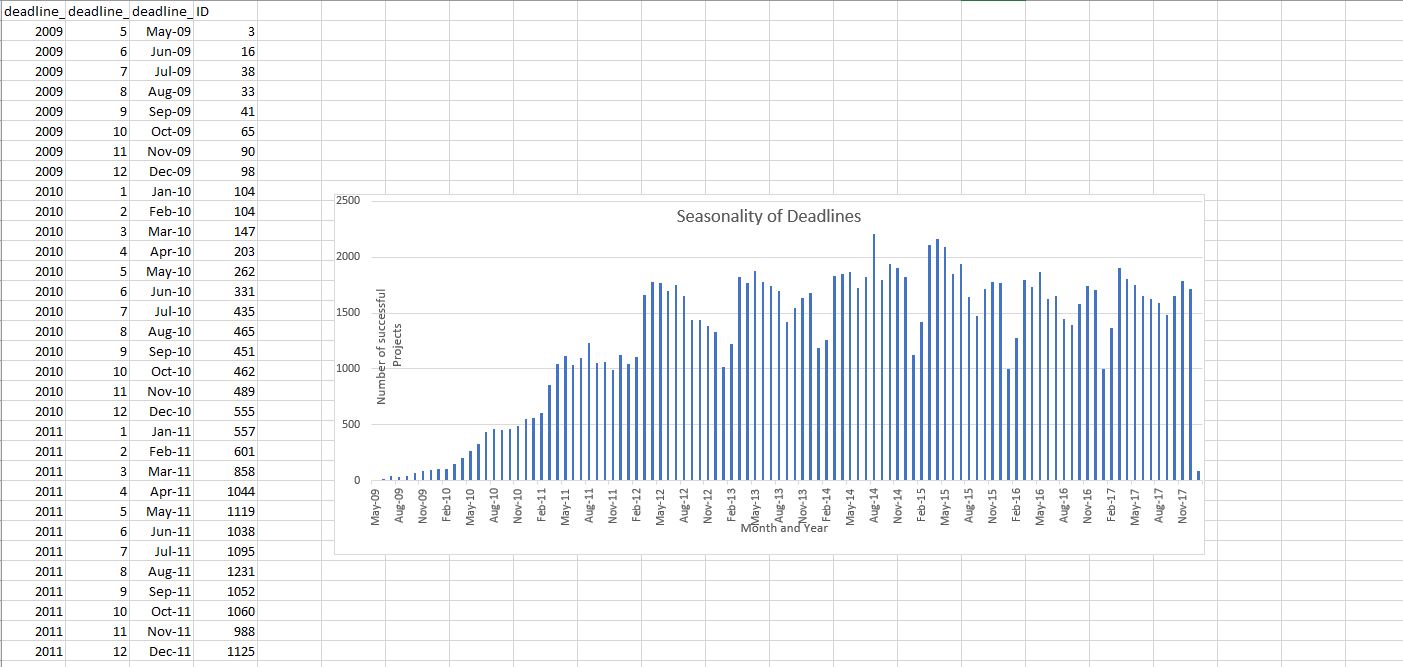
*df['deadline\_month\_year'] = df['deadline\_month'] +"/"+ df['deadline\_year']*

*df\_success=df[df['state']=='successful']*

*df\_time=df\_success.groupby(["deadline\_year","deadline\_month","deadline\_month\_year"],as\_index=False)["ID"].count()*

*df\_time[["deadline\_year","deadline\_month"]]= df\_time[["deadline\_year","deadline\_month"]].apply(pd.to\_numeric)*

*df\_time=df\_time.sort\_values(["deadline\_year","deadline\_month"], ascending=[True,True])*



*#To get the line graph of successful projects for each month*

*df\_2009=df\_time[df\_time["deadline\_year"]==2009]*

*plt.plot(df\_2009["deadline\_month"],df\_2009["ID"], color='green')*

*plt.ylabel('Number of successful projects')*

*plt.xlabel('Month of year 2009')*

*plt.title("Seasonality for year 2009")*

*plt.show()*

## Question of Interest: Which categories of projects have more potential to become successful?

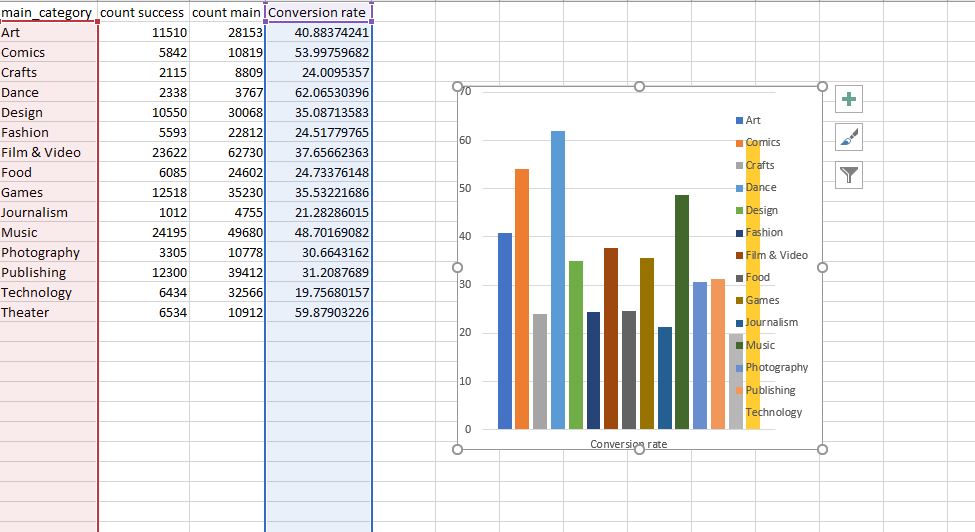
* The dataset with only successful project records is used to group by the main\_category column and also with respect to the count of ID which shows the number of successful projects for each category.

*df\_success.groupby("main\_category",as\_index=False)["ID"].count()*

* Also the same procedure is done to the original dataset which shows the total number of projects for each category.

*df.groupby("main\_category",as\_index=False)["ID"].count()*

* The data is transferred to Excel where a new column Conversion Rate is created as the percentage difference between the successful projects and total number of projects.
* Then a bar graph is plotted in Excel consisting of category vs conversion rate.



## Question of Interest: What is the optimal goal amount to set to attract backers for raising funds?

* The descriptive statistics for the column “usd\_goal\_real” is pulled up using the following code:

*df\_success['usd\_goal\_real'].describe()*

*df\_success['usd\_goal\_real'].median()*

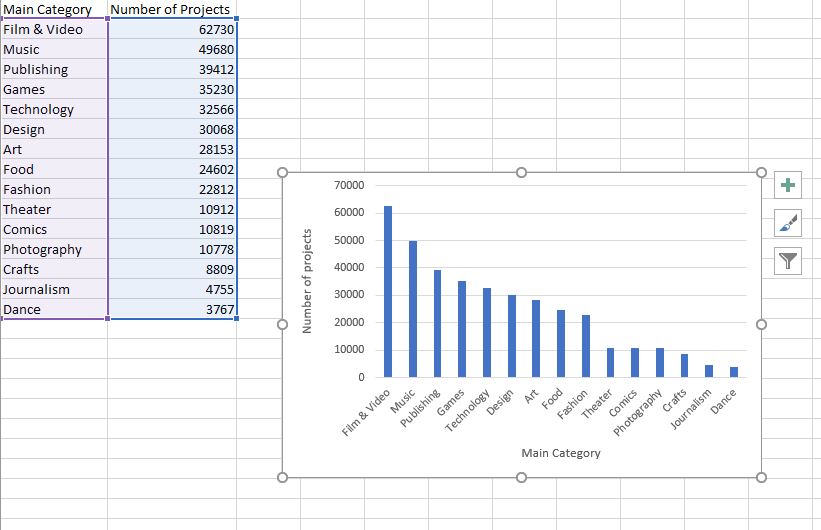
# Kickstarter POV:

## Question of Interest: Which categories of projects are most hosted on the platform?

* Derivation of the second question in the Project Creator POV. The dataset is used to group by the main\_category column and also with respect to the count of ID which shows the number of projects for each category.

*df.groupby("main\_category",as\_index=False)["ID"].count()*

* Then the table is transferred to excel and a bar graph is plotted

**

## Question of Interest: Which country is the platform most utilized on?

* The count of projects are grouped against the countries. Then a bar plot is plotted with the two columns derived using seaborn.

*df\_country\_ID=df.groupby("country",as\_index=False)["ID"].count()*

*df\_country\_ID=df\_country\_ID.sort\_values(["ID"], ascending=[False])*

*df\_country\_ID*

*#Plotting the bar plot*

*sns.barplot(x=df\_country\_ID["country"],y=df\_country\_ID["ID"])*

# Backers POV:

## Question of Interest: Does the number of backers have an influence on the project becoming successful or not?

* Necessary binnings are performed in Python for the “State” Column. The Successful and live states are binned as successful. The failed,canceled,suspended states are considered as failed.
* The dataset is exported as a csv to be loaded in R
* In R, the dataset is loaded and a two sample T-test is executed between the backers and binned state column.

*#Python code*

*df\_r=df.copy()*

*df\_r=df\_r.replace("canceled","failed")*

*df\_r=df\_r.replace("suspended","failed")*

*df\_r=df\_r.replace("live","successful")*

*df.to\_csv(r'E:\UTD\BUAN 6337\Project 1\Sample datasets\\data\_r.csv', index = False)*

*# R code*

*library(data.table)*

*setwd("E:/UTD/BUAN 6337/Project 1/Sample datasets")*

*data<-read.csv("data\_r.csv")*

*data<-data.frame(data)*

*t.test(data$backers,as.numeric(data$state))*

## Question of Interest: Do backers support high scale projects which have a higher goal amount versus low scale projects? Or is the investment unbiased with respect to the scale of the project?

* The usd\_goal\_real column is binned as follows:

1. 0-5000 as 1
2. 5000-100000 as 2
3. 100000-500000 as 3
4. 500000-3000000 as 4

* Then the dataset is exported as a csv to be loaded in R
* In R, the dataset is loaded and an ANOVA test is executed

*#Python code*

*bins = [0, 5000, 100000, 500000, 3000000]*

*labels=[1,2,3,4]*

*df\_success['goal\_bin'] = pd.cut(df\_success['usd\_goal\_real'], bins,labels=labels)*

*df\_success*

*# R code*

*data<-read.csv("data\_success.csv")*

*print(data$goal\_bin)*

*mod<-aov(goal\_bin~backers,data=data)*

*summary(mod)*