Background

Kickstarter is a crowdfunding website that allows entrepreneurs to obtain fundings for their projects, some of them eventually leads to a startup company that provide either an unique product or service. Currently, the success rate of the projects on Kickstarter is around 50%. As a for profit company, Kickstarter will charge a 5% fee if the project gets launched successfully. The main purpose of this project is to analyze the data of past and current projects to provide some actionable insight for improving the project success rate on Kickstarter as its competitor Indiegogo started to gain more market share. Kickstarter can use the results of this project to either provide better guidance to entrepreneurs or find a better promotion strategies to attract more entrepreneurs who are likely to be successful.

Dataset

The dataset will be used in this project is from Kaggle.com. It is called Kickstarter Projects. It contains data for more than 300K projects on Kickstarter. The date range is from May 2009 until December 2017. The latest version of the data included data up to December 2017. There are 15 variables included whether the project was success or fail. Below is the description of each variable:

ID - Internal Kickstarter ID
Name - Name of project. A project is a finite work with a clear goal that you'd like to bring to life.
Category - Project Category. There are total of 159 categories.
Main Category - Category of campaign. There are 15 main categories.
Currency - Currency used to support, such as USD. There are total 14 different currencies.
Deadline - Deadline for crowdfunding
Goal - Fundraising goal. The funding goal is the amount of money that a creator needs to complete their project.
Launched - Date launched
Pledged - Amount pledged by crowd
State - Current condition the project is in. There are 6 states: Failed, Cancelled, Successful, Live, Suspended, Undefined.
Backers - Number of backers.
Country - Country pledged from.
USD Pledged - Conversion in US dollars of pledged column(Done by Kickstarter)
USD Pledged Real - Conversion in US dollars of pledged column (from Fixer.io API)
USD Goal Real - Conversion in US dollars of goal column(from Fixer.io API)

Importing Data:

As the dataset is a .csv file, used pandas pd.read_csv to import data into a variable called kickstarter_raw with parse_date parameter set to 0. The result is a pandas Dataframe.

Initial Data Assessment:

Used the attribute of file .info() to find the information about the data: there are total 15 columns, 5 columns has float data type, 1 column has int data type, 9 columns has object data type. 13 of 15 columns has 378661 non-null entries, the remaining two columns has less than this optimal number of entries is an indication of missing values in those two columns (**name** and **usd pledged**). Used the attribute of file . head(10) to check the first 10 items in data frame. Noticed that column **deadline** and **launched** should be date and time type. Columns **category**, **main_category**, **currency**, **state** and **country** are all category data type.

Data Cleaning:

Drop duplicates: used drop_duplicate attribute of the file to drop duplicated rows in data frame.

Convert data into their appropriate data format: convert **deadline** and **launched** into datetime data type via pd.to_datetime. Convert **category**, **main_category**, **currency**, **state** and **country** all into category data type for faster processing time. Used .astype() attribute of file to execute the conversion. Rerun the .info() confirmed that the following data type shows up in dataframe: 5 category, 2 datetime, 5 float, 1 int, 2 object, memory reduced to 34.0 MB from 43.3 MB.

Count unique values of country column: used value_counts attribute of country column to show number of projects for each country. The top three countries are US, UK, and Canada. The result table contained an abnormal country name = N0” with 3,797 entries.

Extracting rows with this abnormal country name: from the resulting table, it shows rows with country = N0” all have 0 **backers** even most of them has an pledged amount. In addition to 0 backers, the **state** column of those rows are all undefined. From the first 10 rows in raw data frame, we cannot determine the **state** just from the fact that the pledged amount equal or greater than the goal. Noticed that some project still got cancelled even the pledged amount exceeded the goal. Since our project goal is to find out the factors that impact whether the project is successful or not, then the rows with **state = undefined** should be eliminated. There are 3,797 rows which is only 1% of the data set and should be safe to exclude them.

Drop rows with NaN values: used .dropna() attribute of the file to drop those 3,797 rows and assign the resulting dataframe to a new variable called: kickstarter_clean to indicate it is cleaned dataset. Then apply .isna().sum() to check whether there are more row with missing values. The results confirmed that kickstarter_clean dataframe does not have any missing values in all its columns.

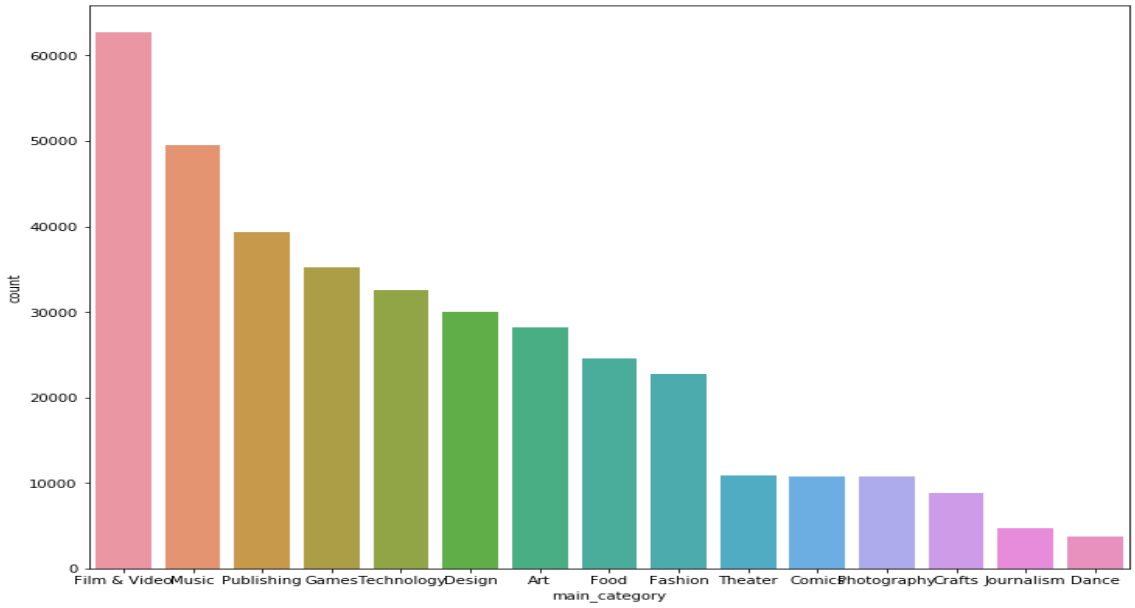
Getting to know more about the cleaned dataset: used value_counts on **main_category** find out the top 3 main categories are: Film & Video, Music and Publishing. Used value_counts on **state** find out the top three states are: failed, successful, and cancelled.

Identify outliers: applied .describe() to kickstarter_clean to get the statistical summary of the numerical data columns. Three main columns are **goal**, **pledged**, and **backers**. Since goal and pledged are the money amount, some of project needs more fund than other projects, it is possible to have a big amount in both columns. Like the money amount, number of backers could also be huge as more people wanted to pledge for certain projects. Or some less popular project only has few number of backers. None of those number should be considered as outliers. Will leave the those number as it is.

Summary of Initial Findings

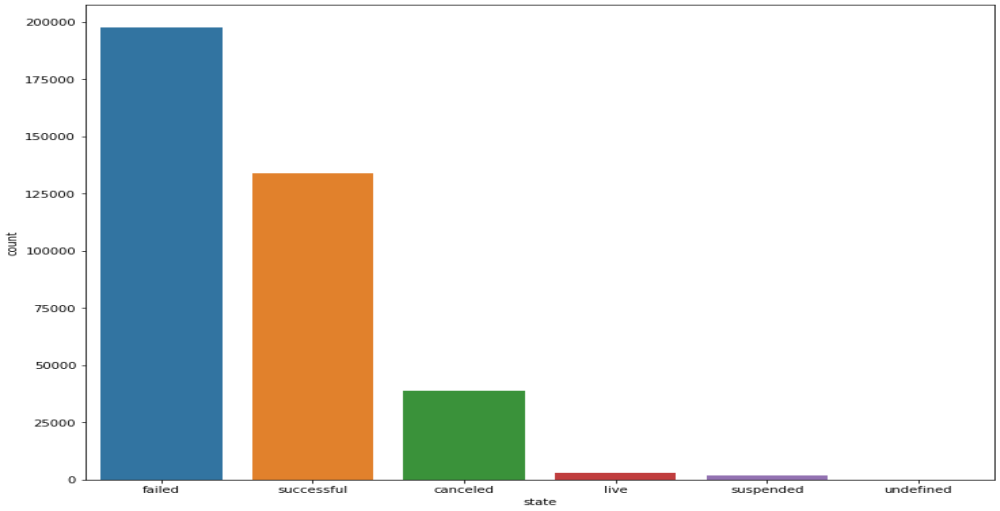
Top main categories of Kickstarter projects

There are total 15 main categories that Kickstarter identified. The Top 5 main categories are: Film & Video with 63K projects, Music with 50K projects, Publishing with 40K projects, Games with 35K projects, Technology with 32K project. The top 5 main categories accounted for almost 60% of the total projects.



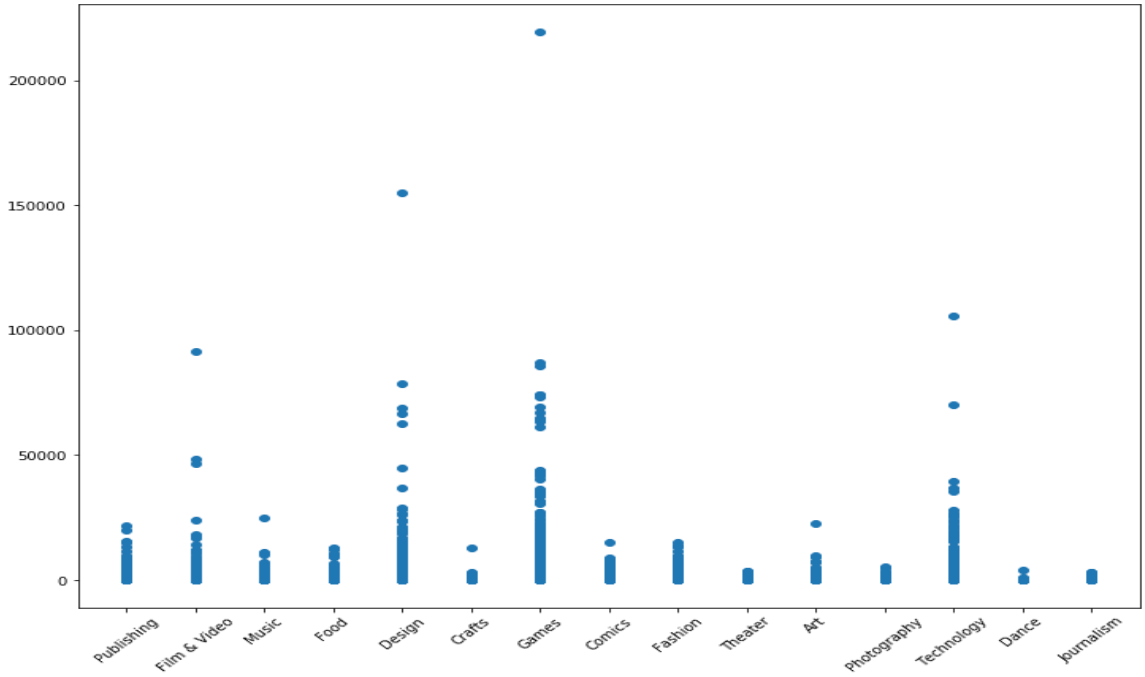
Percentage of project successful

There are about 200K project failed, 133K projects succeed, 39K projects cancelled. There are around 2.8K project still live by the time the data is collected. Lastly, there are 1.8K project were suspended. Roughly 63.5% of the projects were either failed, cancelled or suspended. Only 35.7% project were successful.



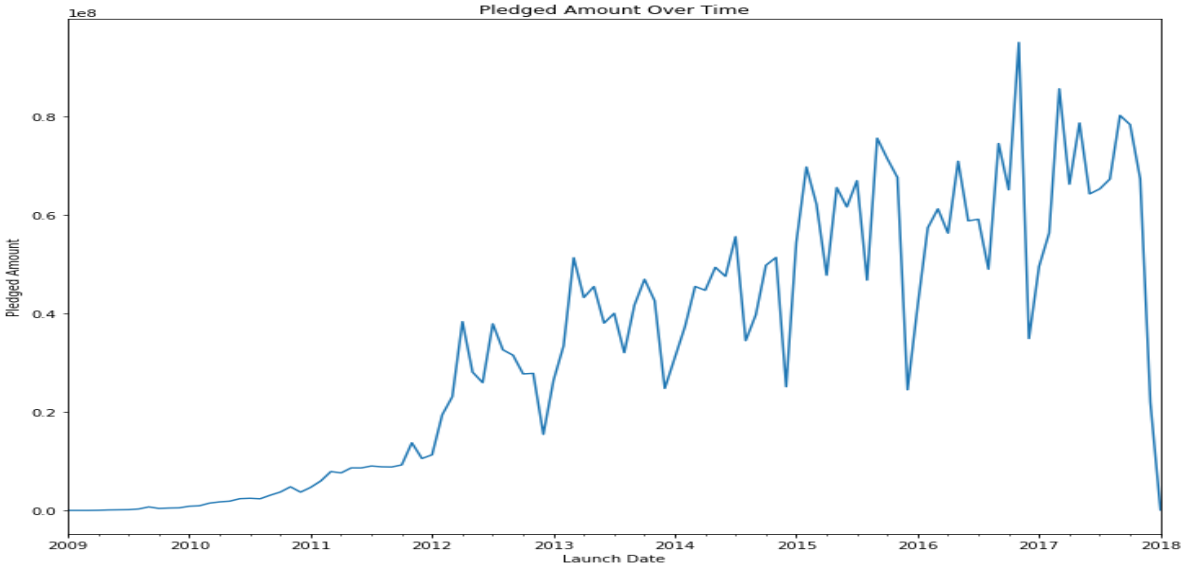
Categories that attracted most number of backers

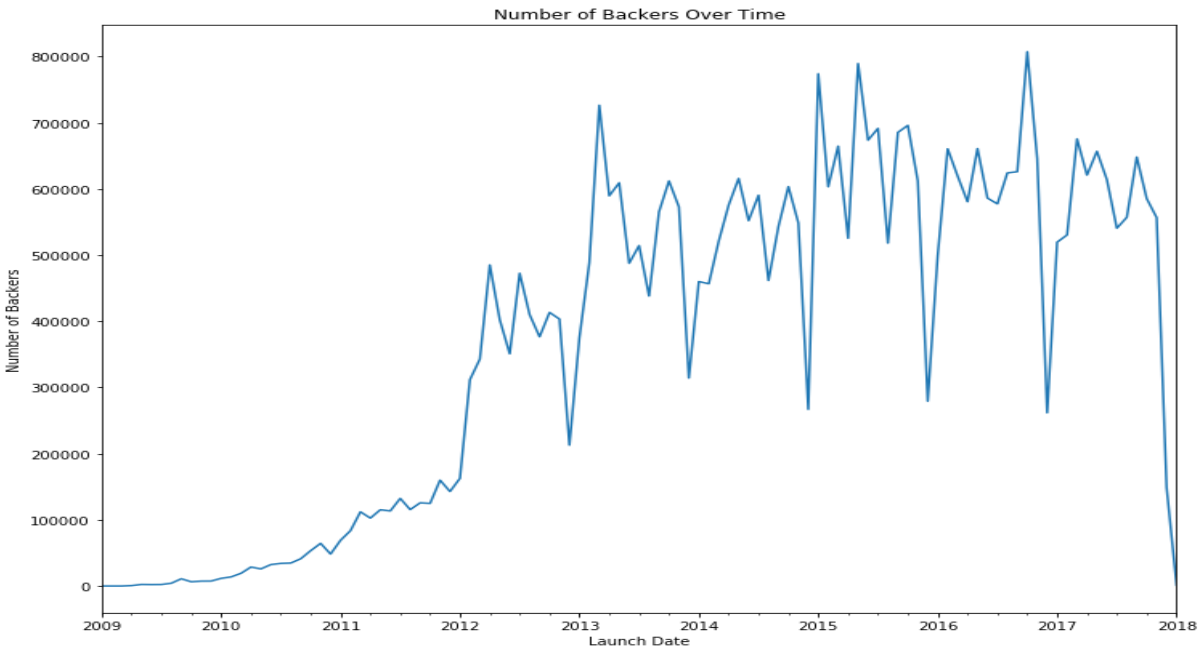
Game and Design category contains several projects that has extremely large number of backers while majority of the projects in other categories has less than 25K backers. Film & Video has the most number of projects, however, only few project gets large number of backers to support. Technology category had several projects that attracted a significant amount of backers. People are generally interested in Game, Design and Technology projects.



Total pledged amount and number of backers varies monthly

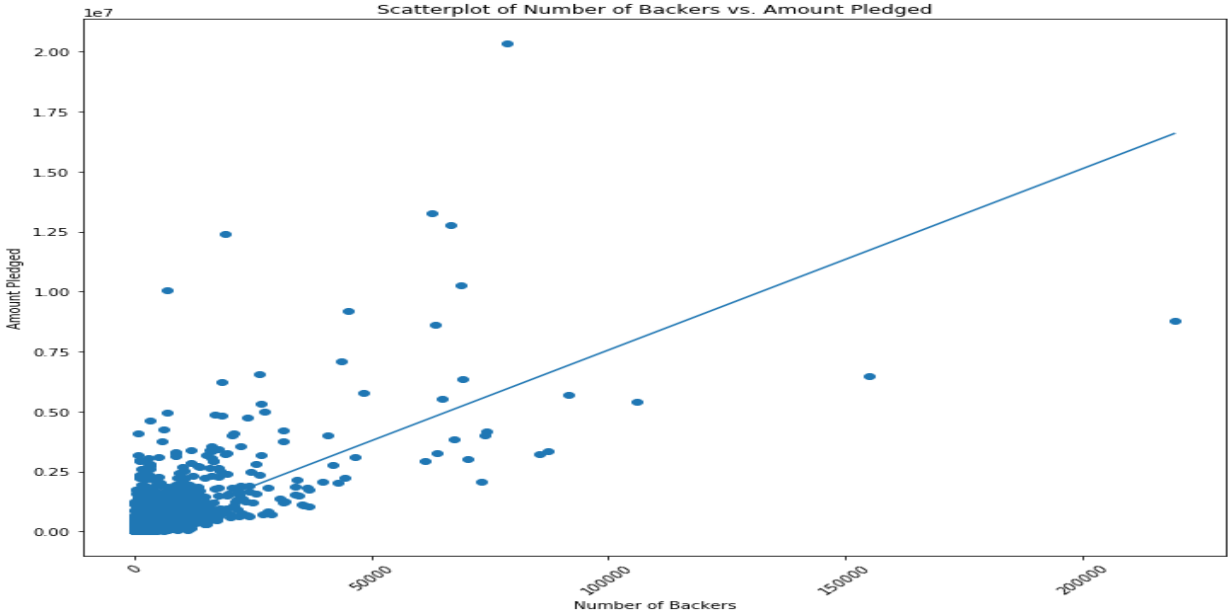
The total pledged amount drops to the lowest point towards the end of every year since 2012. Usually end of year is the holiday season where people will spend most of their money on buying gifts for family members and friends. They are less likely to putting money during that time frame. Then when new year comes, people start to invest in those projects again as many of those people probably get their annual bonus. As time progresses, there is typically another drop in the middle of the year where families spend money on the summer vacation for kids. The pledged amount usually reaches the peak right after new year or second half of the year before the holiday starts. This observation can also be seen from the Number of Backers over time plot. It shows very similar pattern as the pledged amount. A logical follow up question will be : Is the seasonality play a big factor in people's pledge behavior?





Correlation among independent variables

There is a correlation between number of backers and pledged from initial scatter plot with the linear regression line. Generally, as the number of backers goes up, the amount pledged also goes up. This make senses as more backers typically means more fundings.



Pearson Correlation Coefficient has been calculated and a two-sample z-test was run for those two variables.

Null Hypothesis: There is no correlation between Number of Backers and Amount of Pledged

Alternative Hypothesis: There is a positive correlation between Number of Backers and Amount of Pledged

The Test Results are shown below:

The Pearson Correlation Coefficient is: 0.7178584514187824

The p-value is: 0.0

The z-test p value is 0 which is less than alpha = 0.05, therefore, we can reject the null hypothesis.

The Pearson Correlation Coefficient showed a strong positive correlation between Number of Backers and Amount Pledged.