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 (Figure 1).

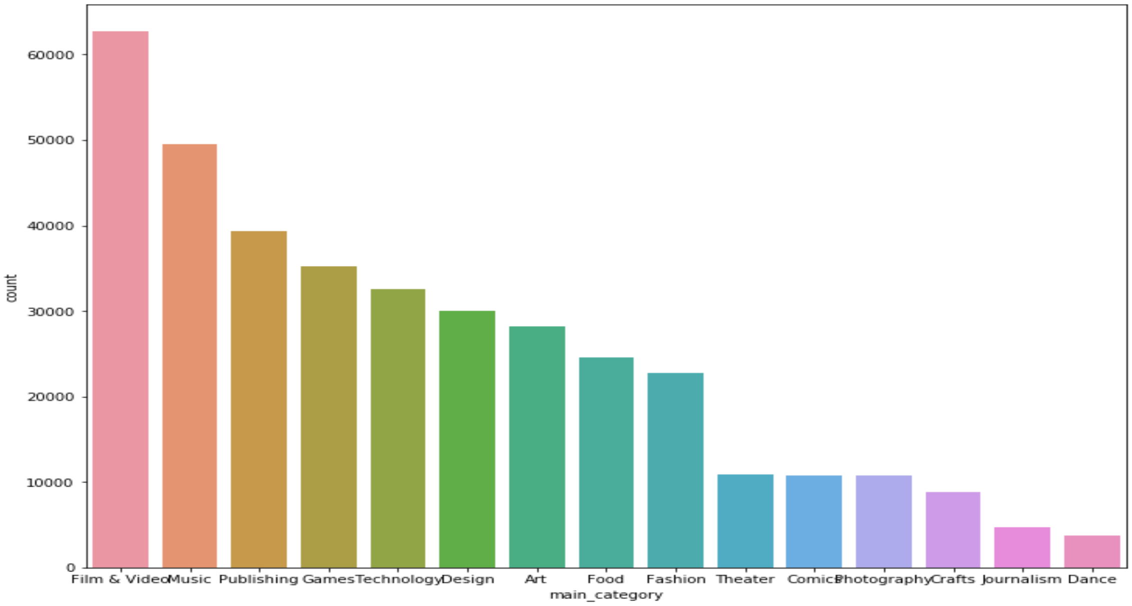


Figure 1. Number of project by main category

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 (Figure 2).

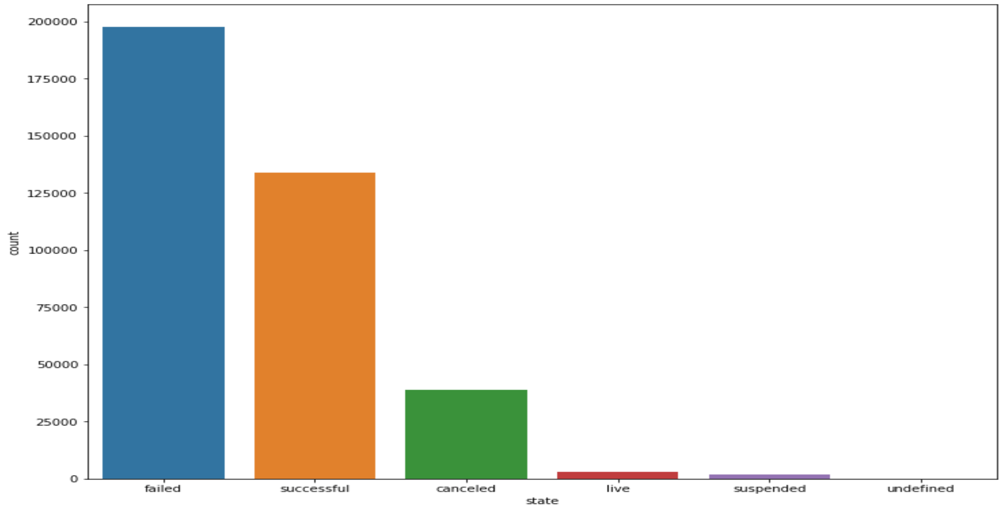


Figure 2. Number of project in each status

As shown below, Film & Video, Music are the two categories that has the most failed projects. They are also the top two categories being successful. There is no surprise due to the fact they are the top categories in terms of number of projects on kickstarter. The number of failed projects in Film & Video is significantly more than succeed ones. Similar to Film & Video, Publishing, Technology also has a higher failure rate than success rate. Only the music category performed slightly better by having a little more succeed project than failed ones.

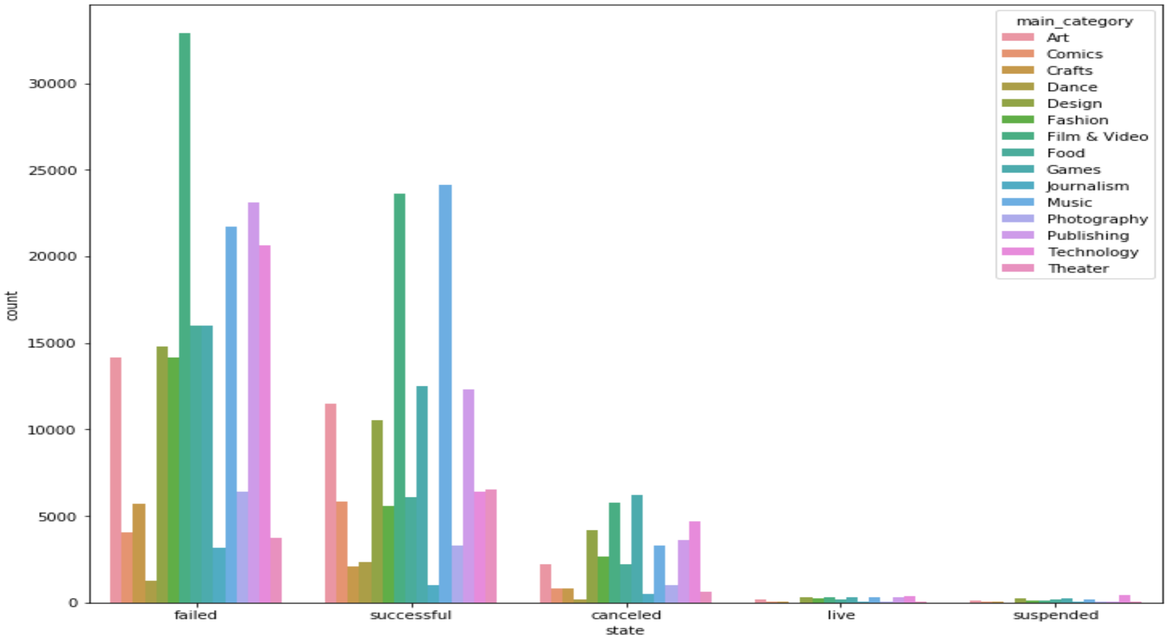


Figure 3 Success number of project by each main category

<u>Categories that attracted most number of backers</u>

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 (Figure 4).

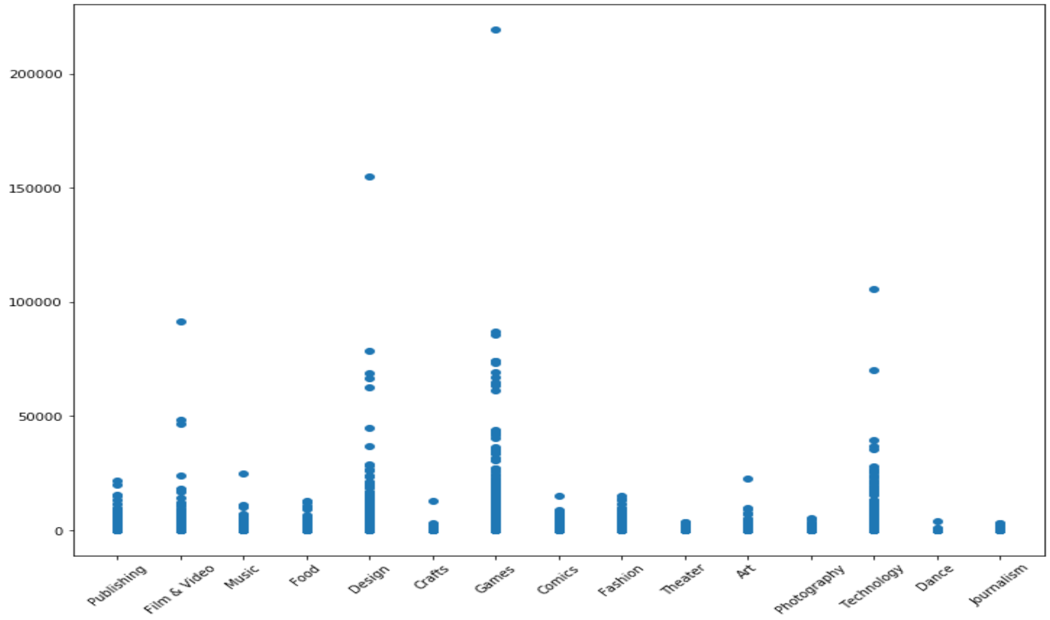


Figure 4. Number of backers by main category

Total pledged amount and number of backers varies monthly

Total pledged amount and number of backers shows an obvious seasonality. Both of them drops to the lowest point towards the end of every year since 2012. Then people started to invest again during the beginning of new year. As time progresses, there is typically another drop in the middle of the year possibly due to summer vacation. 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. Figure 4 and Figure 5

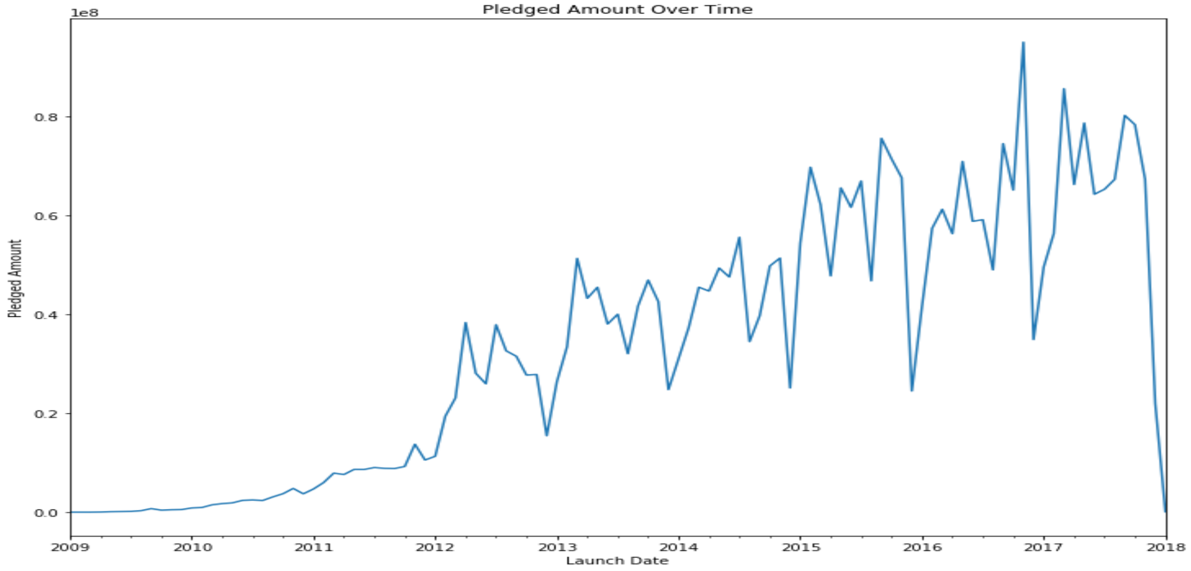


Figure 5. Pledged Amount Over Time

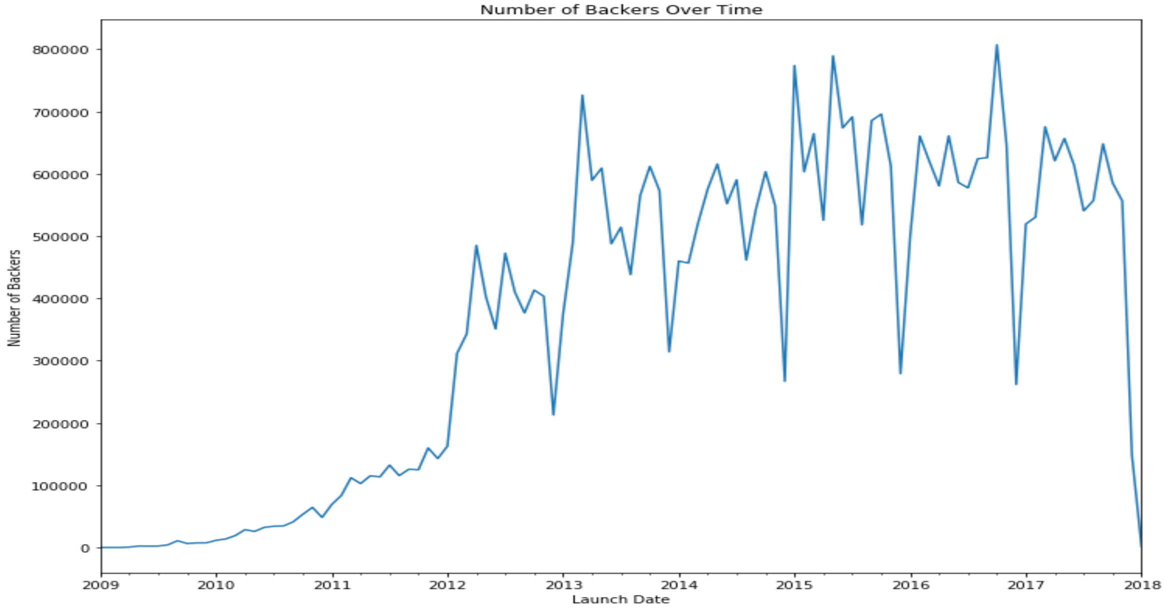


Figure 6. Number Of Backers Over Time

Correlation among independent variables

Backers and pledged

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. This strong correlation can also be observed 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. Since there is a correlation between Number of Backers and Amount Pledged, we should only include either one feature in our model to avoid the impact of correlations. (Figure 7).

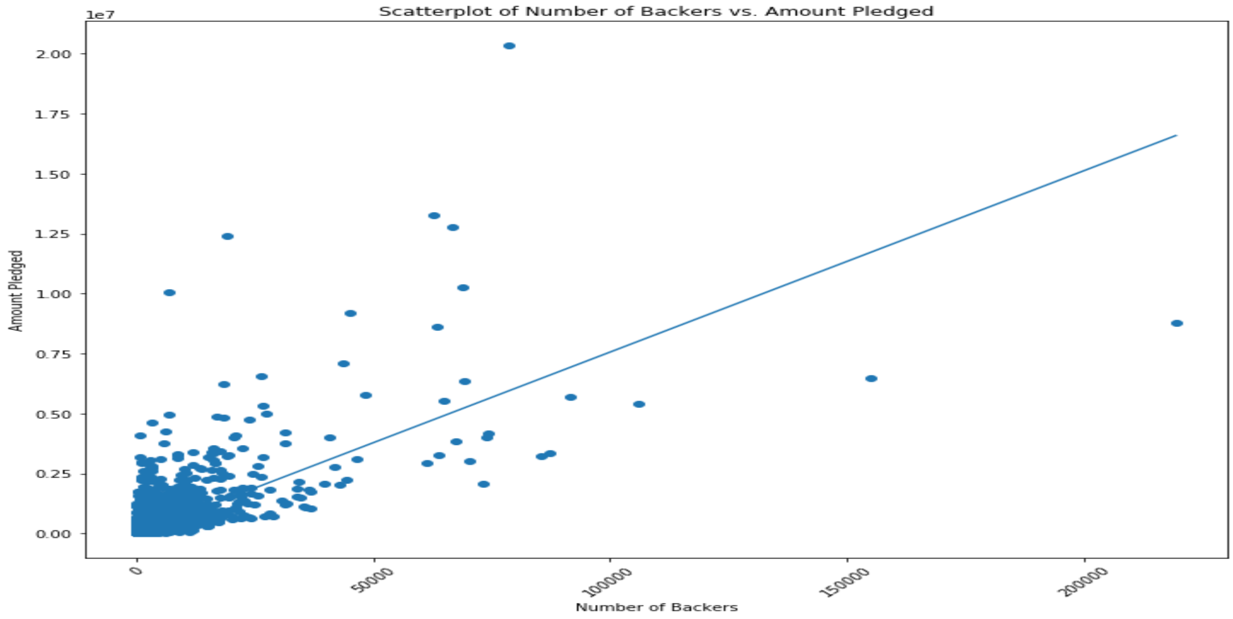


Figure 7. Scatterplot of Number of Backers vs. Amount Pledged.

Goal and pledged

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 Goal 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: 0.007327433128476939

The p-value is: 7.246195827024917e-06

The z-test p value is very close to 0 which is less than alpha = 0.05, therefore, we can reject the null hypothesis.
The Pearson Correlation Coefficient showed a very weak positive correlation between Goal and Amount Pledged. Although the correlation test shows a very weak correlation between those two features, however, Amount of Pledged cannot be determined until the deadline passed. Therefore, Amount of Pledged should not be used to determine the success rate of the project.

Goal and backers

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 Goal 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: 0.003968769170630538
The p-value is: 0.01510246424061905

The z-test p value is around 0.015 which is less than alpha = 0.05, therefore, we can reject the null hypothesis.
The Pearson Correlation Coefficient showed a very weak positive correlation between Goal and backers. It is possible to use both goal and backers in the model building since the correlation between those two features is not very strong.

State and main_category

Since both state and main_category are categorical variable, a chi-square test will be used to evaluate the correlation between two variables.
Null Hypothesis: state and main_category are independent.
Alternative Hypothesis: state and main_category are not independent.
The Test Results are shown below:
Chi-Square Test value: 21526.469859728186
p-value: 0.0
Degree of Freedom: 56

State and country

Similar to State and main_category, state and country are two categorical variable, chi-square test will be used here.
Null Hypothesis: state and country are independent.
Alternative Hypothesis: state and country are not independent.
The Test Results are shown below:

Chi-Square Test value: 5089.318159226757
p-value: 0.0
Degree of Freedom: 84

State and currency

Similar to State and country, state and currency are two categorical variable, chi-square test will be used here.
Null Hypothesis: state and currency are independent.
Alternative Hypothesis: state and currency are not independent.
The Test Results are shown below:
Chi-Square Test value: 4849.285036287392
p-value: 0.0
Degree of Freedom: 52

Currency and country

Similar to State and country, currency and country are two categorical variable, chi-square test will be used here.
Null Hypothesis: country and currency are independent.
Alternative Hypothesis: country and currency are not independent.
The Test Results are shown below:
Chi-Square Test value: 4873180.0
p-value: 0.0
Degree of Freedom: 273

The p-value for chi-square is 0 which is < 0.05 between all three categorical variables and project state, therefore, we can reject the null hypothesis. State and Main_category, State and Country, and State and Currency are not independent. Main_category, Country, Currency could be considered in analysis.
At the same time , the p-value for chi-square between Currency and Country is 0 which is also < 0.05, therefore, we can reject the null hypothesis. Currency and country are not independent. Either one of them should be included in depth analysis.

Based on above, we should consider following combinations of categorical variables in depth analysis:

Option 1: Main_category, country
Option 2: Main_category, currency.

In-Depth Analysis

Since the goal of this project is to determine whether the project will be successful or not, this is a classification problem. Three models have been explored: Logistic Regression, KNN, and Bagging (ensemble).

Pre-processing Data

Since there are only three numerical features in the dataset (goal, pledged, backer), and the pledged value cannot be determined until the end of the project and is correlated to the goal features (from our pearson correlation analysis) assuming the project will be successful when the pledged amount is equal or greater than the goal amount. Therefore, the pledged feature will not be included in the analysis.

The currency feature will not be included in depth analysis due to its high correlation with country. Usually the currency type is close related to the country, therefore, it will not be a good feature to add to the analysis dataset. Therefore, we will use Option 1 of the categorical variable selection as our categorical features for the model (Main_category, country). In order to incorporate the categorical features into the analysis, one-hot encoding scheme has been applied to features main_cateogry and country. The process created binary column for each category and returned a sparse matrix and dense array.

Training and Testing Dataset

In order to evaluate how good is the model, the cleaned dataset has been randomly splitted into training set, test set, and validation set. The percentage split is 60%, 20%, and 20% respectively. By utilizing train_test_split package from sklearn.model_selection twice, the desired data set were obtained.

Logistic Regression

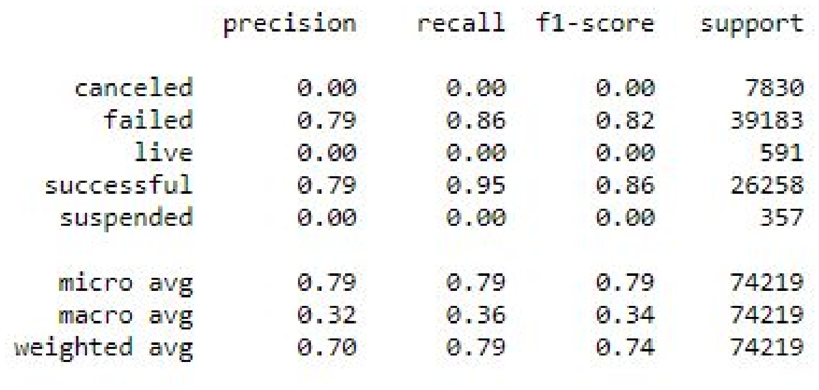
One of the most popular classification models. In order to optimize the performance of the model, the GridSearchCV module has been used to find the best model parameter C(regularization strength) with cross validation. The penalty has been set as default value of l2 and number of fold set to 10 and 8. Below is a table that summarize the model accuracy score with varies C values (Table 1).

Logistic Regression	(penalty = l2, cv = 10)	**Logistic Regression**	(penalty = l2, cv = 8)
C value	Accuracy	C value	Accuracy
0.1	0.7919	0.1	0.7938
1	0.792	1	0.7938

10	0.792	10	0.7938
50	0.7919	50	0.7937
100	0.7919	100	0.7937
500	0.7919	500	0.7937
1000	0.792	1000	0.7937
3000	0.7919	3000	0.7928
5000	0.792	5000	0.7927

Table 1. Summary of C value and Test Accuracy in Logistic Regression

As seen from the table, the accuracy with different C values does not vary significantly as long as penalty method is l2. The number of folds cv also did not make big impact on the accuracy of the model output with varies C values. Therefore, C = 10, penalty = l2 , cv =8 will be our model parameters. The model is then trained with training set, and used to predict y value based on the test dataset. Below table 1 shows the Classification Report for test set. Last, the remaining validation set was used to check the validity of the model. Table 2 shows the Classification Report for validation set.

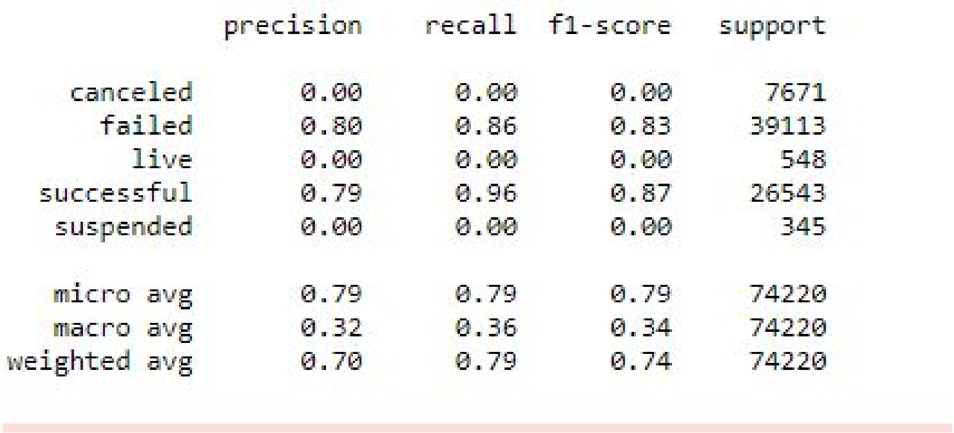


```
              precision    recall  f1-score   support

    canceled       0.00      0.00      0.00      7830
      failed       0.79      0.86      0.82     39183
        live       0.00      0.00      0.00       591
  successful       0.79      0.95      0.86     26258
   suspended       0.00      0.00      0.00       357

   micro avg       0.79      0.79      0.79     74219
   macro avg       0.32      0.36      0.34     74219
weighted avg       0.70      0.79      0.74     74219
```

Table 2. Classification Report for Test Data



```
              precision    recall  f1-score   support

    canceled       0.00      0.00      0.00      7671
      failed       0.80      0.86      0.83     39113
        live       0.00      0.00      0.00       548
  successful       0.79      0.96      0.87     26543
   suspended       0.00      0.00      0.00       345

   micro avg       0.79      0.79      0.79     74220
   macro avg       0.32      0.36      0.34     74220
weighted avg       0.70      0.79      0.74     74220
```

Table 3. Classification Report for Validation Data

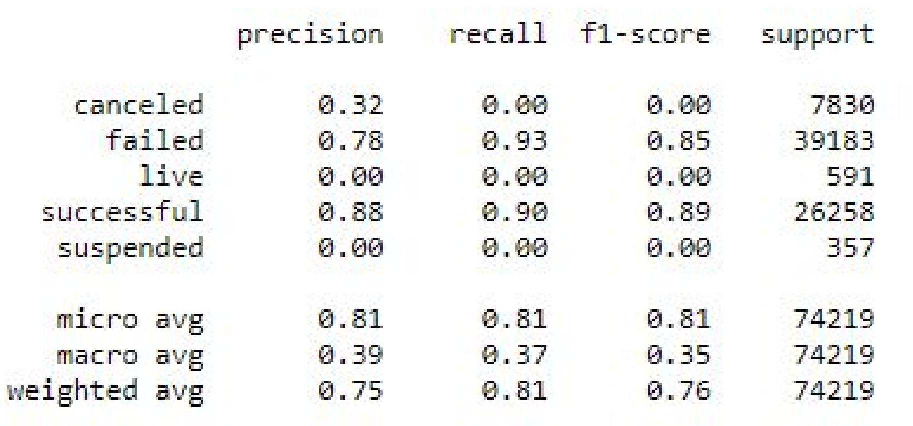
Knn

For knn, the most important parameter is the number of neighbors. Several numbers has been tried to evaluate the performance of the model. The result is shown in below table (Table 4).

KNN	
n_neighbor	Accuracy
3	0.7617
8	0.7952
10	0.8015
15	0.8086
20	0.8102
25	0.8117
30	0.8117
40	0.8114

Table 4. Summary of n_neighbor and Test Accuracy

When n_neighbor = 25 or 30, it will generate a test accuracy of 0.8117, therefore, we will choose n_neighbor = 30 for our model. The accuracy of the validation set is 0.8117. Table 5 and Table 6 shows classification report for Test and Validation data.

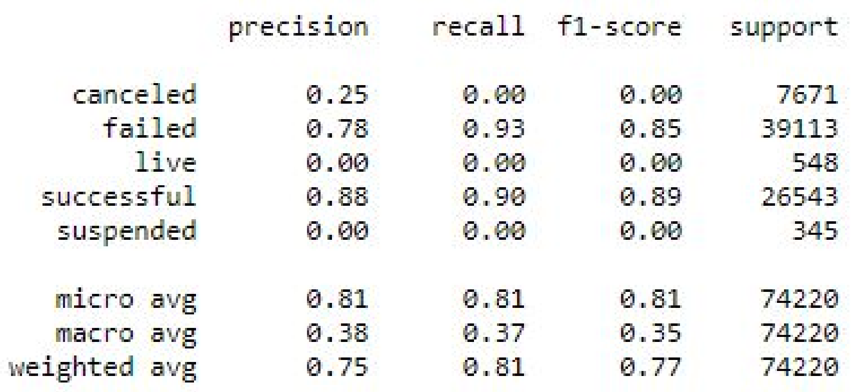


```
              precision    recall  f1-score   support

    canceled       0.32      0.00      0.00      7830
      failed       0.78      0.93      0.85     39183
        live       0.00      0.00      0.00       591
  successful       0.88      0.90      0.89     26258
   suspended       0.00      0.00      0.00       357

   micro avg       0.81      0.81      0.81     74219
   macro avg       0.39      0.37      0.35     74219
weighted avg       0.75      0.81      0.76     74219
```

Table 5. Classification Report for Test Data



```
              precision    recall  f1-score   support

    canceled       0.25      0.00      0.00      7671
      failed       0.78      0.93      0.85     39113
        live       0.00      0.00      0.00       548
  successful       0.88      0.90      0.89     26543
   suspended       0.00      0.00      0.00       345

   micro avg       0.81      0.81      0.81     74220
   macro avg       0.38      0.37      0.35     74220
weighted avg       0.75      0.81      0.77     74220
```

Table 6. Classification Report for Validation Data

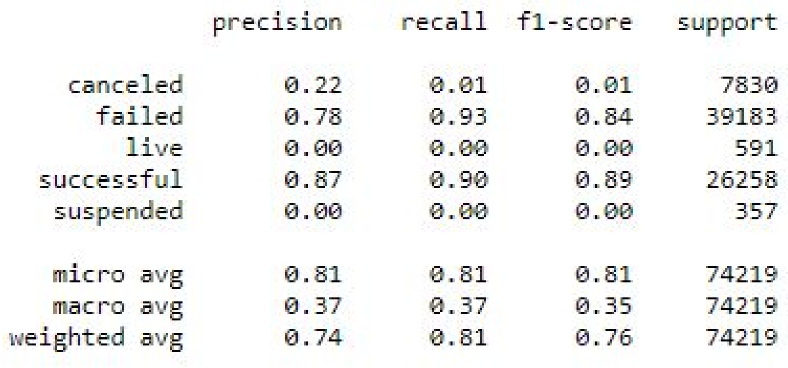
Ensemble

For ensemble model, we have tried several different combinations of parameter max_sample and max_features to find best test accuracy score. Below table contains the results:

Ensemable		
max_samples	max_features	Test Accuracy
0.3	0.3	0.7516
0.3	0.5	0.7825
0.3	0.8	0.8051
0.5	0.3	0.7744
0.5	0.5	0.8021
0.5	0.8	0.809

Table 7.

The test accuracy varies with different combination of max_samples and max_features. When max_sample = 0.5, max_features = 0.8, it generated the best accuracy score for test set = 0.809. The accuracy of the validation set is 0.8116.

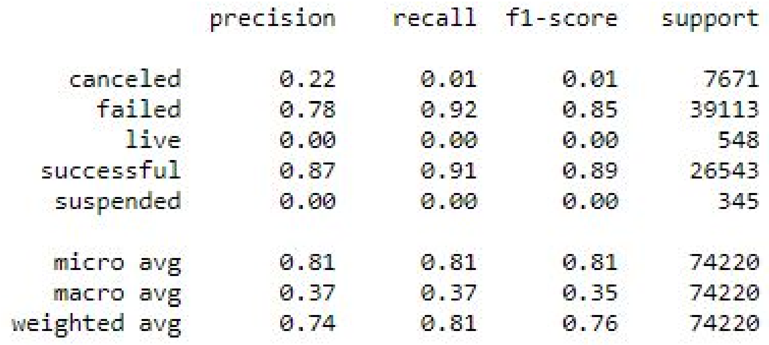


```
              precision    recall  f1-score   support

    canceled       0.22      0.01      0.01      7830
      failed       0.78      0.93      0.84     39183
        live       0.00      0.00      0.00       591
  successful       0.87      0.90      0.89     26258
   suspended       0.00      0.00      0.00       357

   micro avg       0.81      0.81      0.81     74219
   macro avg       0.37      0.37      0.35     74219
weighted avg       0.74      0.81      0.76     74219
```

Table 8. Classification report for test data



```
              precision    recall  f1-score   support

    canceled       0.22      0.01      0.01      7671
      failed       0.78      0.92      0.85     39113
        live       0.00      0.00      0.00       548
  successful       0.87      0.91      0.89     26543
   suspended       0.00      0.00      0.00       345

   micro avg       0.81      0.81      0.81     74220
   macro avg       0.37      0.37      0.35     74220
weighted avg       0.74      0.81      0.76     74220
```

Table 9. Classification Report for Validation Data

Results

Out of the three models, knn and ensemble generated similar test data accuracy score. However, when applying validation data set to the model, only KNN performed best. Therefore, KNN with n_neighbor = 30 will be a reasonable model used to predict the outcome of the kickstarter projects.

Model	**Test Accuracy Score**	**alidation Accuracy Score**
Logistic	0.7938	0.7935
KNN	0.8117	0.8144
Ensemble	0.809	0.8116

Conclusion

The top 5 categories of kickstarter projects accounts for more than 50% of the total projects. Projects with a lot of creativity, such as music, video and movie are the top 2 most popular categories. This aligns with Kickstarter's main focus: maintain a global crowdfunding platform that focused on creativity and merchandising. As the current data showed, the success rate of kickstarter project is only 37.5% which is less than 50%. The top 5 categories of projects are also the most popular target for interested backers. There are seasonality trend on when and how much people will fund the project. A KNN model provided the best prediction outcome based on both numerical and categorical features in the dataset and will be a good model to predict the success chance when a new project is established on Kickstarter. Kickstarter could use this model to monitor the projects on the website, however, further analysis will be needed to figure out how to improve the success rate.