Bellevue University

Winter 2020

What Might Determine A Successful Kickstarter Campaign

January 30, 2020

Kevin Angotti

Executive Summary

When becoming your own boss and getting ideas off the ground, companies like Kickstarter have become a tool to help provide this means—looking into campaigns that might provide helpful insights to aid future entrepreneurs to determine if a Kickstarter campaign is either the right direction for their ideas or which category might be the best approach for success. This project will look at the main categories of a Kickstarter campaign to determine the success rate based on different criteria. Looking at the data, my goal was to see which variables could help predict if a campaign would become successful.

Diving into the data, there are a bunch of promising variables that could be used in prediction. Some key variables I looked at to run against my models were the main category, pledged, pledged in the United States (usd_pledged), backers and goal. By using these variables as insights into the data, I hope to be able to accurately find the best campaign options for anyone looking to start or grow their idea.

When you venture to the Kickstarter site, there is a nice layout of beginning a campaign and providing tips for starting and growing your funding base. However, they do not currently provide any insights into what types of campaigns or which categories will provide the best outcomes. This could help focus campaigns or even if a Kickstarter campaign is a right direction for their idea/company.

Preparation

This project will utilize various predictive analytic methods to understand relationships within the data to identify factors that could allow a person to determine the best route to take when starting a Kickstarter campaign. Key factors examined during this project were main_category, currency, goal, pledged, state, backers, country, and usd_pledged.

Initial Findings

From this project, the initial findings yielded several interesting returns; more on this regarding findings can be found below. Two of my hypotheses guiding this project were put through my modeling process, and the outcome points to predicting a campaign type for their success or failure.

Abstract

When a person is looking to build their idea or business, one place they can turn is Kickstarter.com. This site allows for one to open a campaign and see investors for their idea. The process is straight forward, and you select the category your idea or product might fall into, provide some background to what the product is or what you are creating, then choose your location and answer some simple information to set up an account.

After you are up and running with an account and fully set up your idea/product, you can start receiving pledges. This step will depend on the timeframe to get your funding; if you set a large pledge goal, you may be waiting a while to achieve the goal, thus obtaining funding. People all over the globe look at Kickstarter every day for investment opportunities. However, the people at Kickstarter suggest reaching out to friends and family to get the campaign moving. Connecting Kickstarter to all your social platforms can also provide added exposure.

The Kickstarter community explains that the beginning stage of a campaign is the busiest time, and you can expect the first few days to be jam-packed as you will need to spread the word

on your campaign, answer questions from potential backers and rally all your social media networks in order to reach your fundraising goal [2].

When you first start, it is suggested to look through the helpful guide for setting up a campaign. There are many options for how you fundraise. All or nothing campaigns allow for higher success in case you do not meet your fundraising goal. The site provides an excellent example of how to set goals. However, it is suggested to research your idea for set up cost and manufacturing to determine the best goals to set. The goal you set depends on the cost it will take to get the idea off the ground.

This project will help determine the best outcome one might have when creating goals and what factors might help campaigns become more successful.

Keywords: Predictive Analysis, Kickstarter, K-Modes Algorithm, Neural Network, EDA

Introduction

Since April 2009, Kickstarter has become a place millions of people come to expand and get their ideas, package, and sell them to the world. There have been many high-profile products that started from a Kickstarter campaign. Campaigns have won Grammys and Oscars and can be found everywhere you look. Many factors can go into a successful campaign, and understanding the process and ways to success are key when creating one.

One of the most important factors during setup is to ensure the details of a campaign are explained entirely—backers tend to choose campaigns based on how feasible they are. Backers also go into campaigns knowing their funding toward one can lead to not getting their investment back. Kickstarter Campaigns can fail, and quite easily, just as starting a company or growing an idea tends to be. Keys to success can all depend on how the campaign owners sell the idea. This is the main factor for this project, what can provide the best outcome when creating a campaign. Are there things that variables in this data that can point to for a more successful campaign? What modeling will work best in making these kinds of predictions? These are just a few ideas I brought to this project, and from the project outcome, I would like to expand the next steps to Deep Learning or AI.

Background Information

Methods

Data Understanding

When approaching the data for this project, I began by examining variables within the dataset. This took time due to the vast number of rows in the dataset. I used a combination of the R programming language, Python programming language, and Power BI. To begin the project, I turned to R; I conducted summary statistics of the datasets to see initial trends and what each variable might have to tell me. I followed this up with a quick linear regression model to see what these trends might mean. The R programing language provides a good plug-and-play model for different LR models. This is useful to get an idea of what your variables might have to tell you in a more invasive model setting. I took the initial LR model from R and my summary statistics. I used that to create my initial visuals in Power BI. Doing this at the beginning of the project allowed me to see what the data looks like before running any type of model in Python. Initial visualizations provided evidence that the main_category variable may be a possible target for the models. Other areas of interest were to review the state of a campaign, usd_pledged,

pledged (world, and backers. I then used Python to create all of the data models that would be needed for analysis and predictions. The first model I created was a K-Modes clustering model; I then followed that with a linear regression model that scores on the mean squared error MSE, finally, a neural network. These models provide insight into the questions, allowing for future analysis, and eventually, Deep Learning and AI.

Data Preparation

Extensive data preparation was performed on the dataset to transform key categorical predictors into a numeric form. This was done to prepare the data for modeling and further analysis, though EDA and basic visuals were completed beforehand. During the data cleaning process, certain variables were transformed into numeric values from categorical ones. This would allow for the summary statistics, linear models, and the neural network to run and provide insights into the data. The variables were transformed in Python on the data set by the use of the formula, (df['columnName'] = le.fit_transform(df.columnName.values). A new dataset created with the transformed variables was saved and then loaded into the Rstudio to run the summary stats and the initial liner model. The main variables that were transformed were state, main category, country, and currency.

The next process I conducted was to create each of the highlighted variables visuals to get a better look at the data. As you will see in Appendix A, Figures 1 through 4 provide a good sense of what the data is currently suggesting. These visuals were used to compare the data before it was transformed for modeling. Each of these four variables were normalized for plotting using the function train[variable].value_counts(normalize=True)*100.

Modeling

The extensive data preparation allowed for linear regression models to be performed. R's initial linear model was for the backers variable as a function of the state, usd_pledged, pledged, and man_category. These variables returned some good insights for significance, and all but three categories returned a p-value under the 0.05 threshold. The overall P-value of the model yielded < 2.2e-16 or (0.00000000000000022) with an F-statistic of 2.249e+04 or 10.1134 and an adjusted R-squared of 0.5455.

I then shifted to Python to run another linear model. I utilized a clustering algorithm that could better understand some of the categorical variables that might be important. For that, I chose to deploy a K-Modes clustering model to better deal with the complex variables contained within the dataset. This model allows for matching clusters based on the number of its matching categories between each data point. This clustering model works well with data with a large amount of categorical mixed in numeric data or large datasets with mixed numeric and categorical values.

Finally, I run the data through a Neural Network model that would take in the cleaned and transformed dataset using the Keras built-in library for Python. The model run two different ways with an epoch or number of passes on the dataset with a batch size or hyperparameter that tells the model how many samples to work through. Running multiple model types provides a good range into what the data could tell us, thus allowing for sound analysis.

Results

From the preliminary results, things look promising. While further analysis is needed to draw definitive conclusions, the initial model results in R are shown below.

Linear Model

```
lm(formula = backers ~ state + usd_pledged + pledged + man_category,
    data = ks_projects_201801)
Residuals:
   Min     1Q Median     3Q    Max
-80367    -54      3     19 150616
Coefficients:
                                   Estimate      Std. Error       t value          Pr(>|t|)
(Intercept)                    -31.53275549   4.77307911      -6.606       3.94e-11 ***
state[T.failed]                 11.79552182   3.43123396       3.438       0.000587 ***
state[T.live]                   25.09029304  12.03567625       2.085       0.037101 *
state[T.successful]            105.77598921   3.59886821      29.391       < 2e-16 ***
state[T.suspended]              27.94592893  14.66021333       1.906       0.056619 .
usd.pledged                      0.00358098   0.00002481     144.341       < 2e-16 ***
pledged                          0.00422223   0.00002038     207.127       < 2e-16 ***
main_category[T.Comics]         56.28421891   6.95874674       8.088       6.07e-16 ***
main_category[T.Crafts]         12.60341164   7.51286804       1.678       0.093431 .
main_category[T.Dance]         -21.58605072  10.67342135      -2.022       0.043135 *
main_category[T.Design]         53.93863182   5.10961957      10.556       < 2e-16 ***
main_category[T.Fashion]        16.30116175   5.48748864       2.971       0.002972 **
main_category[T.Film&Video]      5.33015204   4.41067605       1.208       0.226869
main_category[T.Food]           13.90263175   5.37634689       2.586       0.009713 **
main_category[T.Games]         159.31938462   4.92912596      32.322       < 2e-16 ***
main_category[T.Journalism]     17.75709699   9.64766272       1.841       0.065687 .
main_category[T.Music]           0.23800840   4.59081633       0.052       0.958653
main_category[T.Photography]     6.07508382   6.96588901       0.872       0.383144
main_category[T.Publishing]     23.02229817   4.80192001       4.794       1.63e-06 ***
main_category[T.Technology]     16.63935174   5.03193213       3.307       0.000944 ***
main_category[T.Theater]       -18.92767473   6.94341564      -2.726       0.006411 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 614.7 on 374843 degrees of freedom
  (3797 observations deleted due to missingness)
Multiple R-squared:  0.5455,  Adjusted R-squared:  0.5455
F-statistic: 2.249e+04 on 20 and 374843 DF,  p-value: < 2.2e-16
```

Table 1. Linear Regression Model

After running this linear regression model, which used different summary statistics to determine significance, such as an F-Statistic, P-Value, and Multiple R-Squared, these measures only provided some of the information needed to determine if the data could predict on a model. I also created visuals in Power BI (Appendix A) from these statistics to get a peek at what a model might conclude before I started building my models.

Shifting to Python, I started with some exploratory data analysis (EDA), where I could get some insights into some of the key variables and run some other linear regression and random forest models on the data. Appendix A will show -some of these results and be further discussed in the discussion section of this paper. After the EDA was completed, I started to code the main models for the analysis. I first created a form of the K-Means clustering model called K-Modes. This type of model handles categorical data very well. A K-Modes model takes in an entire dataset and returns its prediction on each cluster's most common outcome. Each time the models ran, the results returned the measured number of clusters; in this case, my last run returned a failed campaign in 3 of the 5 clusters. Of the 5 clusters, the 3 failed campaigns were film and video, Food, and publishing. The two that were successful were music and a film and video. In all five clusters, the United States provided the most funding. What I found surprising and should be pointed out was one of the successful campaigns was the music category, but a game was the product, so that could be insightful for future entrepreneurs looking for successful campaigns. Maybe make sure to put your product/idea into a category that is known to succeed.

When building my next model(s) in Python, the linear regression model took the man_category as the target variable and utilized a 60/40 test, train, split. The result, which is shown in figure 5, returned predictions for each of the category types. The model, which was scored with MSE, returned 14.7 against the entire dataset and an MSE of 15.3 for just the man_category vs. usd_pledged. The thought process here was from the EDA analysis the United States was shown to provide the most funding to campaigns so, running a linear model just on the United States pledged amounts was a necessary step. The error did go up by a full point; however, it is still a strong model since this change was very little. I was only able to provide an accuracy of .26 to 0.31 percent.

Looking forward to the next steps, I would like to see some form of Deep Learning or AI.

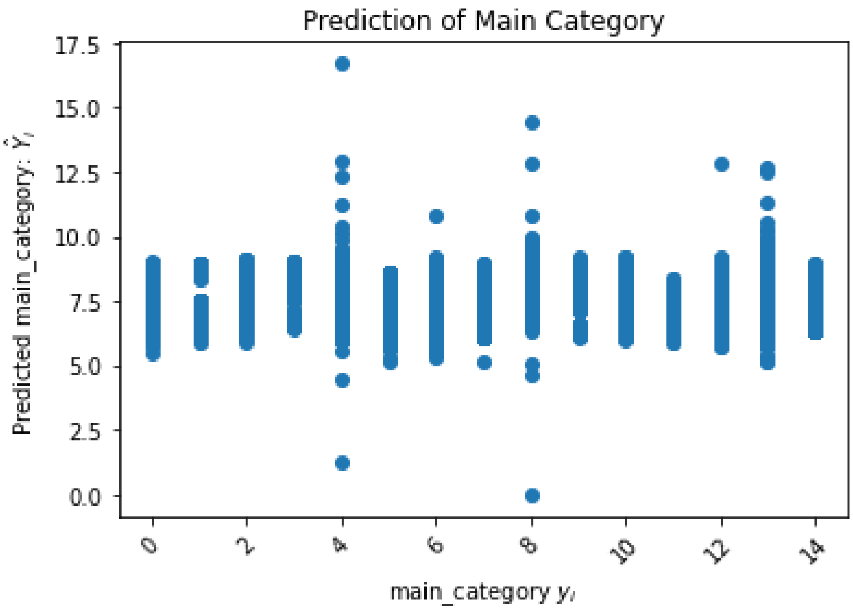

Figure 5: Linear Regress by MSE

Rerunning the model with state as the target returns an MSE of 49.95, and the variables state vs. usd_pledged returned an MSE of 35.53, meaning that state as a whole against the entire dataset does not provide a significant return as it does with man_category, but the change in MSE from state to usd_pledged is much more significant. Between the three variables; state, man_category, and usd_pledged, I provide the best analysis for predictions on future runs.

The last model I created was a Neural Network. This was my first time running a successful NN on a dataset. The results were good and bad. There were two Neural network setups; the first one had a dense or number of nodes as follows input layer 11, output layer 1, two hidden layers of 32 and 32. This NN model provided an accuracy of only 11%. The second NN model had a better outcome, the input and output layers were also 11 and output 1 but had three hidden layers of 1000, and this model returned 52-60% on runs. They were both run on 25 passes or epochs with a batch size of 32, which was how many samples the model took.

I believe the adjusting of the layers and input/outputs could help obtain better results. The main issue I ran into was CPU and memory. The models I was able to run were not the inputs I would have like to test on. I had to reduce the epochs from 100-150 to 25 just to run the analysis. Running this setup on a better-matched PC could allow for better results. Appendix A provides a visual of the validation loss on just the 25 epochs; given more increase in passes, the outcome could tell a different story.

Discussion

Insights from the summary statistics and linear models returned some helpful information that one can use to create a Kickstarter campaign. It can be quickly noticed that most of the funding stems from right here in the United States, followed by Great Britten and other European nations. As you will see in Figure 1 (Appendix A), campaigns fail 52% of the time. Knowing the categories that provide the best results could provide someone looking to get their products funded with an added advantage. Figure 2 (Appendix A) shows that the top categories chosen are Film and Video, Music, Publishing, Games, and Technology. However, are these the best categories to get funded? We will get to that in just a bit. Figures 3 & 4 show that without a doubt, the United States provides the most funding, with Great Britten and Canada drastically behind but well ahead of other countries. This could provide useful insights to focus your product. If the company is based in the US, you may want to reach out more to Canada or GB for funds, or since most funding comes from the US, keep the focus here in the US.

Just looking at funding from the world to that of just the United States, Figures 6 & 7 provide a visual of the difference; though they look similar, the metrics are what you should focus on. Here both visuals show the density of funding, and the United States is higher across the board ranging between 2-4, while the rest of the world is under 1. The box and whisker plot that accompanies both density plots show that, on average, the world pledges around $25,000 while the United States $40,000.

Figure 8 may be the most helpful because it shows the man_category and the most likely state that it returns. Looking at the figure, you can see that the best outcome is from the Music category and that more than half the time, their campaigns are successful. Technology has the worst outcome with less than 25% of successful campaigns.

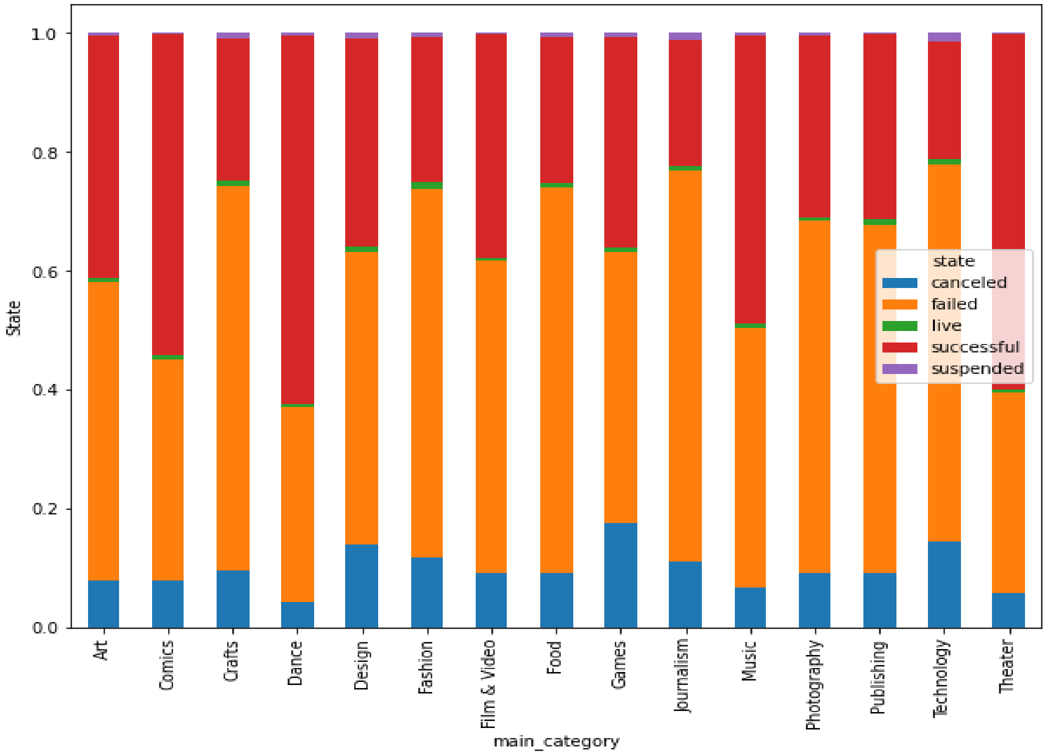

Figure 8. Stacked Bar Graph Main_Category vs. State.

I ended up running two other regression models during the EDA process to see what the data could determine. The first Logistic Regression returned an accuracy of 55% on a 64/40 train, test, split. The other was a Random Forest, which returned 53% accuracy with a depth of

10 and n_estimators of 50. Both of these models returning percentages in the mid-'50s suggest that the variables are significant. However, is that enough to make predictions?

The results I achieved from the traditional linear model and the Neural Network, I believe, provide a good model for aiding a new business or entrepreneur in making decisions on their Kickstarter campaigns. If nothing else, it provides some guidance into where to look and focus attention. Figures 4 and 8 can help determine that the Design and Game categories are ones to stay away from if possible as they tend to fail more often than not.

Lastly, in Power BI, I created visuals based on the modeling to see if they could provide a better visual aspect of what I saw from the analysis. Figures 9 through 14 in Appendix A show the outcomes. I will note that from Figure 10, you can see that Film and Video, Music and Art are your categories that prove to be most successful. Figure 13 shows this in better detail and even shows each state type and their outcomes to the category's overall percentage.

Conclusion

Looking at Kickstarter data and trying to find insights became a real possibility and the analysis I was able to perform is just a small picture of the whole aspect of the questions I looked to answer. Digging deeper into which category and state were successful and determining the best category can be possible. This can help future entrepreneurs make better decisions on where to start or which category to place their products. As was determined, it can be possible to place a product into a category that might not be exactly the best description but may provide a better result.

If there are other similar datasets out there that could be added to this one for further analysis, this project can provide a good model for those seeking to start campaign options before they began.

The next step, I feel, would be adding more variable data to see how the models fair; there may need to be some tweaking of each model to account for added data. However, the outcome would provide better results. In addition, the models created in this project call for high computing power to be more accurate in their results.

Acknowledgments

I want to thank Professor Catie Williams for her continued guidance and support throughout the project. I would also like to thank Data World and Kickstarter for making the project data publicly available.

References

1. Mouille, M. (2018). Kickstarter Projects. Datasets. Kaggle. Retrieved from https://www.kaggle.com/kemical/kickstarter-projects

Information on what and how the data is to be examined. The author also provides a good amount of information on how the data was compiled and what he would like out of any projects.

2. https://www.kickstarter.com/

This site s the main company page and provide adequate background on how a campaign gets started, and provides the mission goals and company outline.

3. Peters, D. (2013). 10 Tips I Wish I Knew Before I Launched My Kickstarter Campaign. Entrepreneur. Retrieved from https://www.entrepreneur.com/article/229782

This article gives a layout of other Kickstarter campaigns that people would have liked to know before they started out. This can help provide insights to results we ay see in the data.

4. Stimmel, G. (2019). 11 Proven Tips For Launching A Successful Kickstarter Project. Product hype. Retrieved from https://producthype.co/kickstarter-tips-2019/

This is another site that can be used to help make sense of the results and may even provide useful information in determining a target for the models.

5. Cood Backer. (2019). HOW TO TELL IF A KICKSTARTER CAMPAIGN IS LEGITIMATE. Cool Backer. Retrieved from https://coolbacker.com/how-to-tell-if-a-kickstarter-campaign-is-legitimate/

Looks into what could be a scam or legit Kickstarter campaign.

6. Albright, D. (2016). 3 Things to Consider Before Backing a Kickstarter Project. Make Use Of. Retrieved from https://www.makeuseof.com/tag/3-things-consider-backing-kickstarter-project/

Looks at what might help a Kickstarter become a project people will back.

7. Prondle, D. (2019). Don't get burned! How to back crowdfunding projects the smart way. Digital Trends. Retrieved from https://www.digitaltrends.com/cool-tech/crowdfunding-tips-avoid-scams-kickstarter-indiegogo/

Provides another aspect that could be used as helpful insights that might contribute to success in a campaign. This might be helpful in how people might pick a Kickstarter.

8. Benvides, N. (2017). What Makes a Successful Kickstarter Campaign? Towards Data Science. Medium. Retrieved from https://towardsdatascience.com/what-makes-a-successful-kickstarter-campaign-ad36fb3eaf69

Further information on what it takes for a Kickstarter to become successful. These types of articles can help provide reasons a campaign will be successful or fail.

9. James, T. (2021). 7 Keys for Successful Kickstarter Campaigns. Cover Kit. Retrieved from https://convertkit.com/successful-kickstarter-campaigns

More help t determine success.

10. Crockett, Z. (2019). What are your chances of successfully raising money on Kickstarter? The Hustle. Retrieved from https://thehustle.co/crowdfunding-success-rate

Appendix A

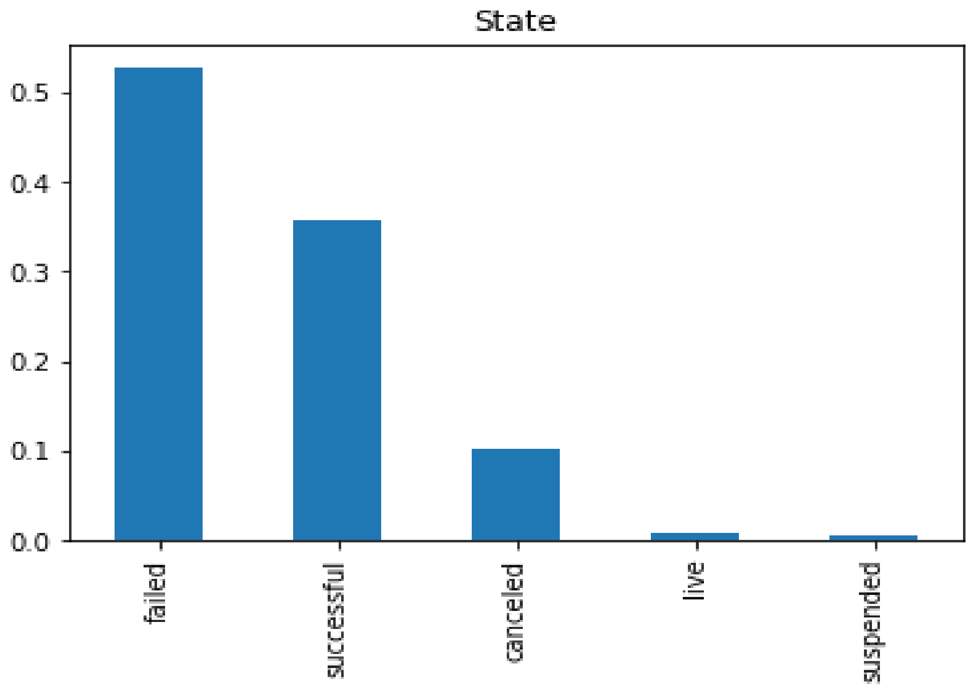

Figure 1: State

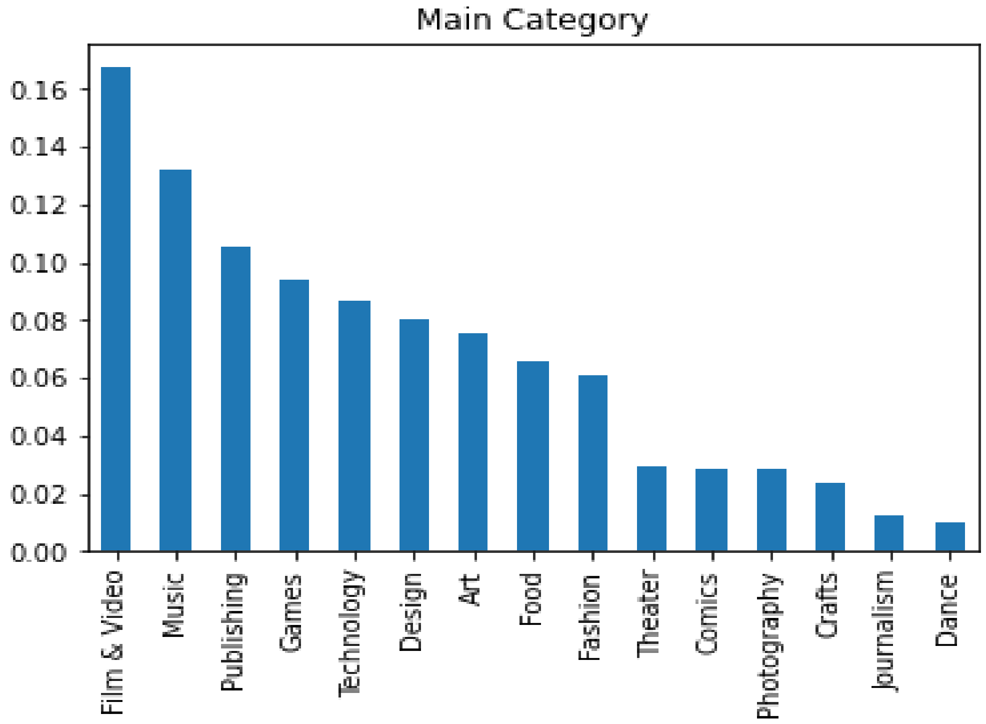

Figure 2: Main Category

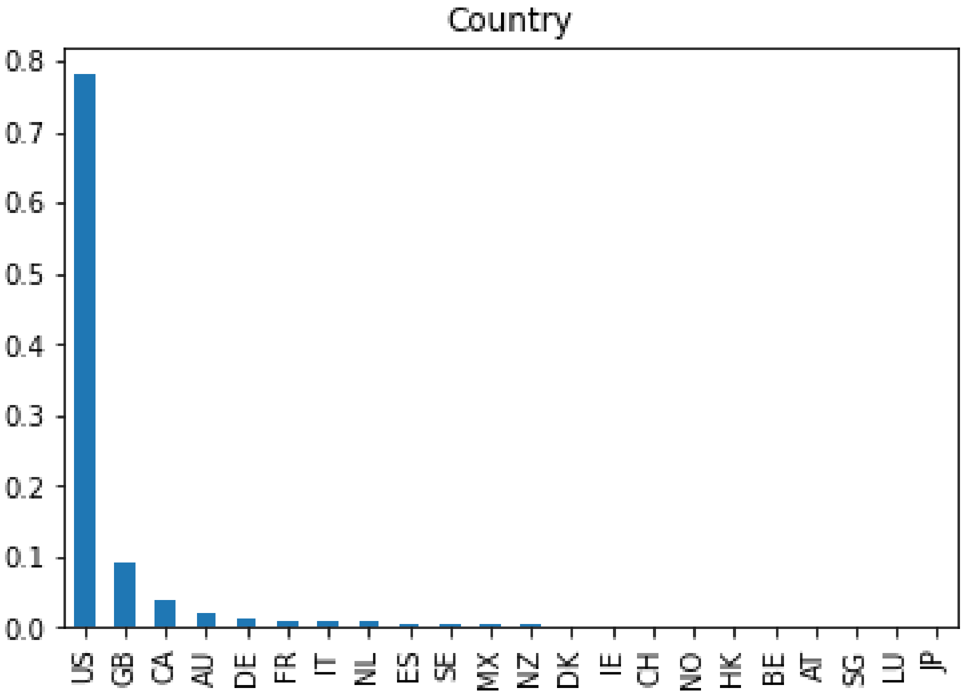

Figure 3: Country

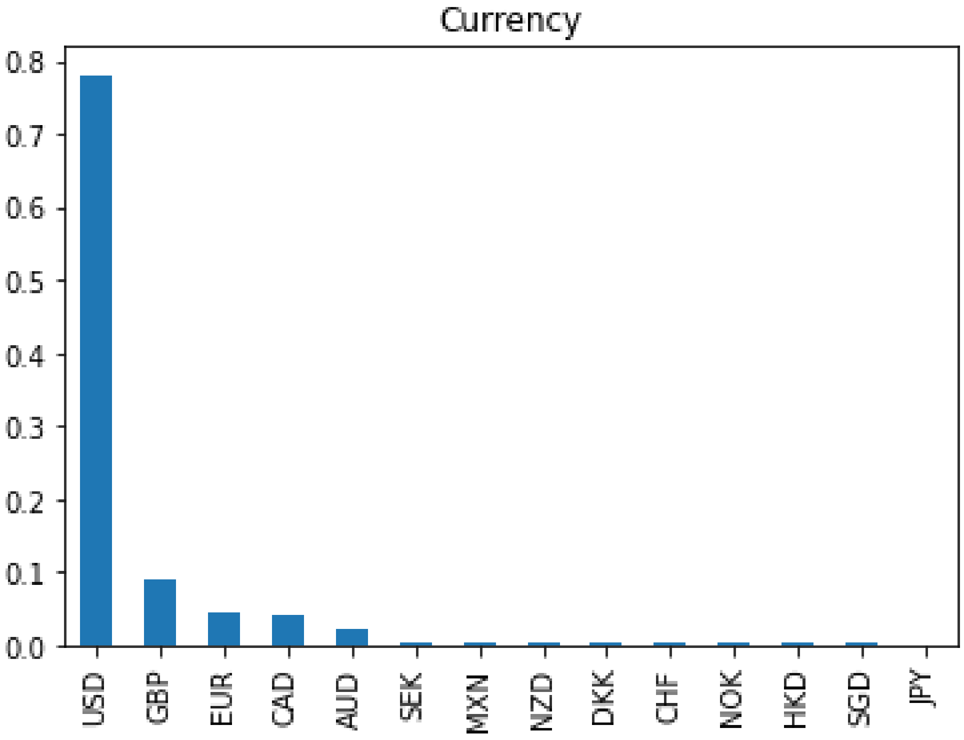

Figure 4: Currency

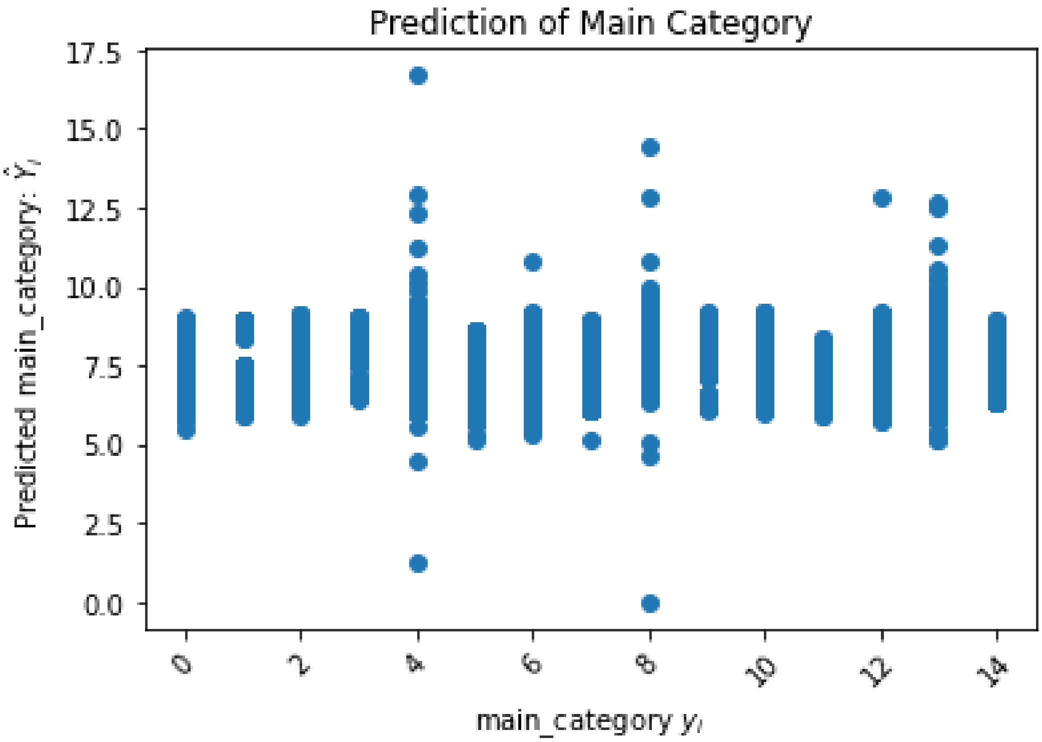

Figure 5: Predicted Main Category

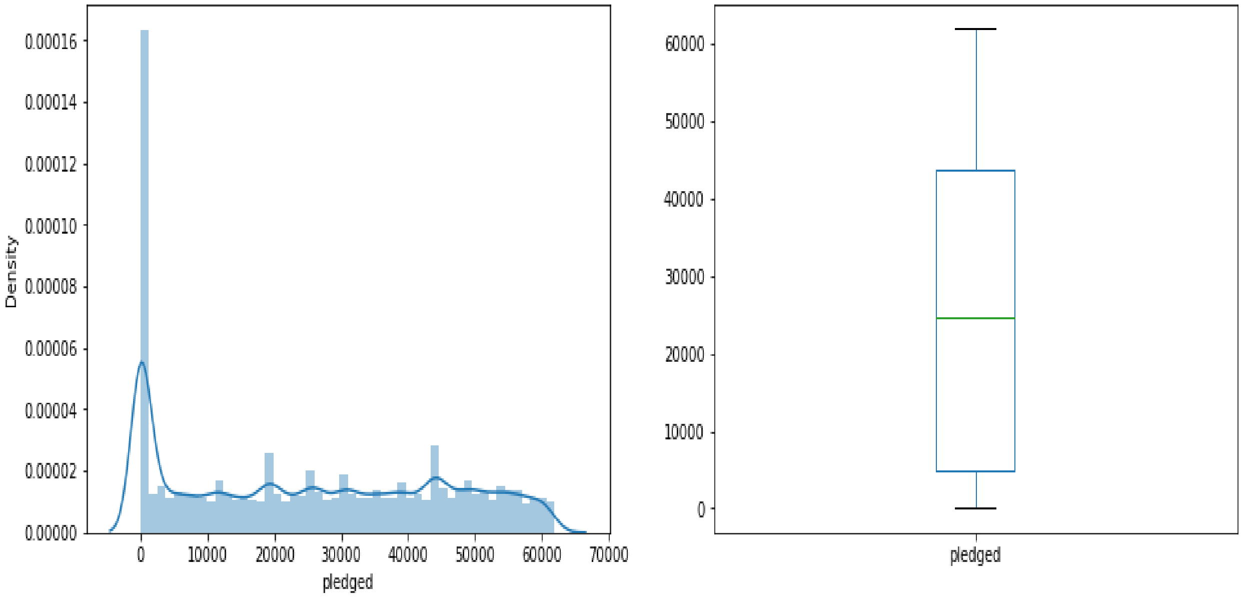

Figure 6: Worldwide Pledges

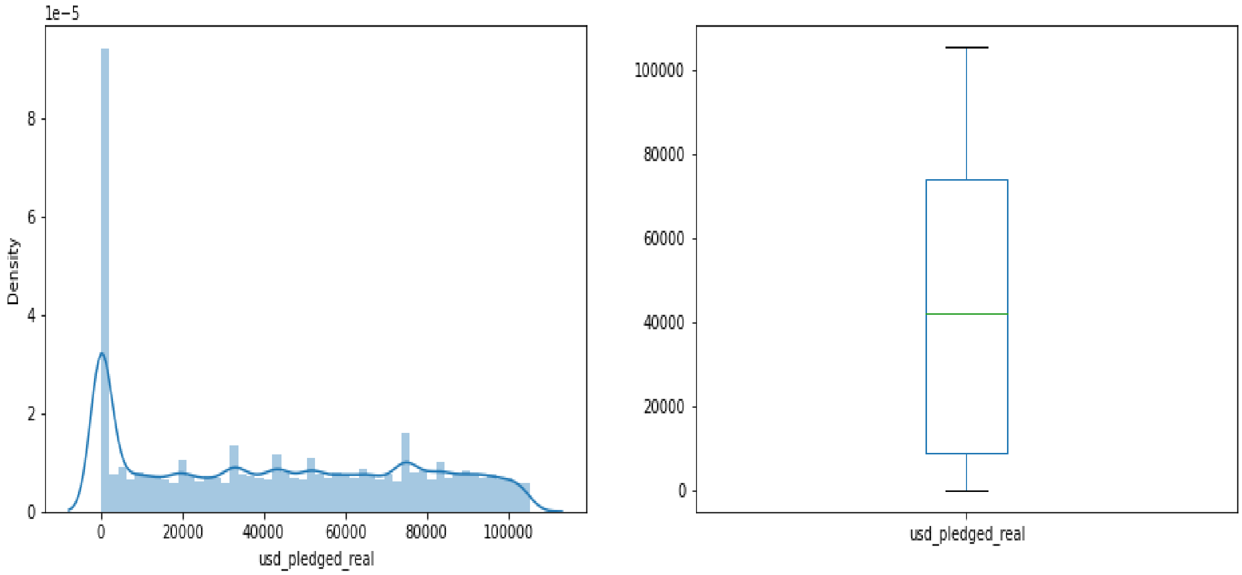

Figure 7: United States Pledges

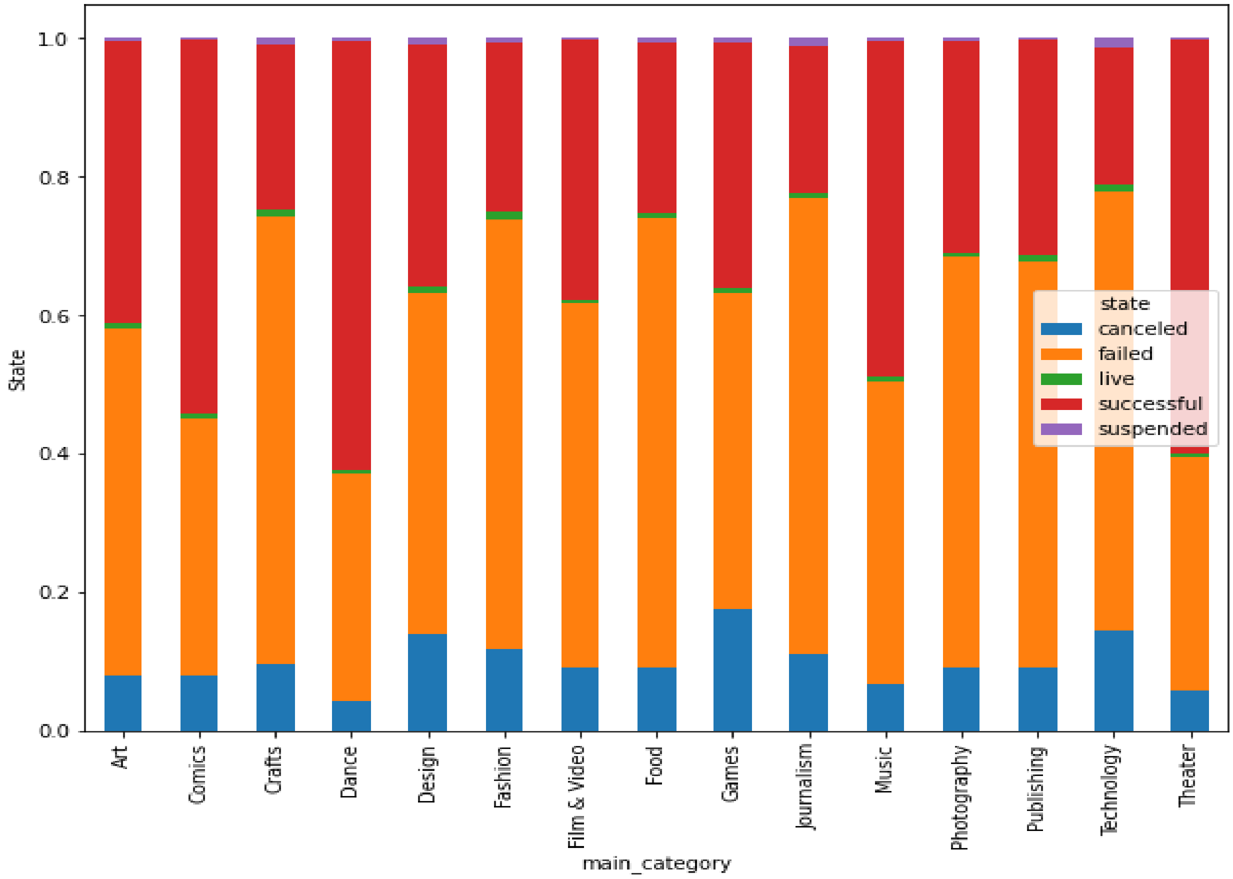

Figure 8: Main Category vs. State

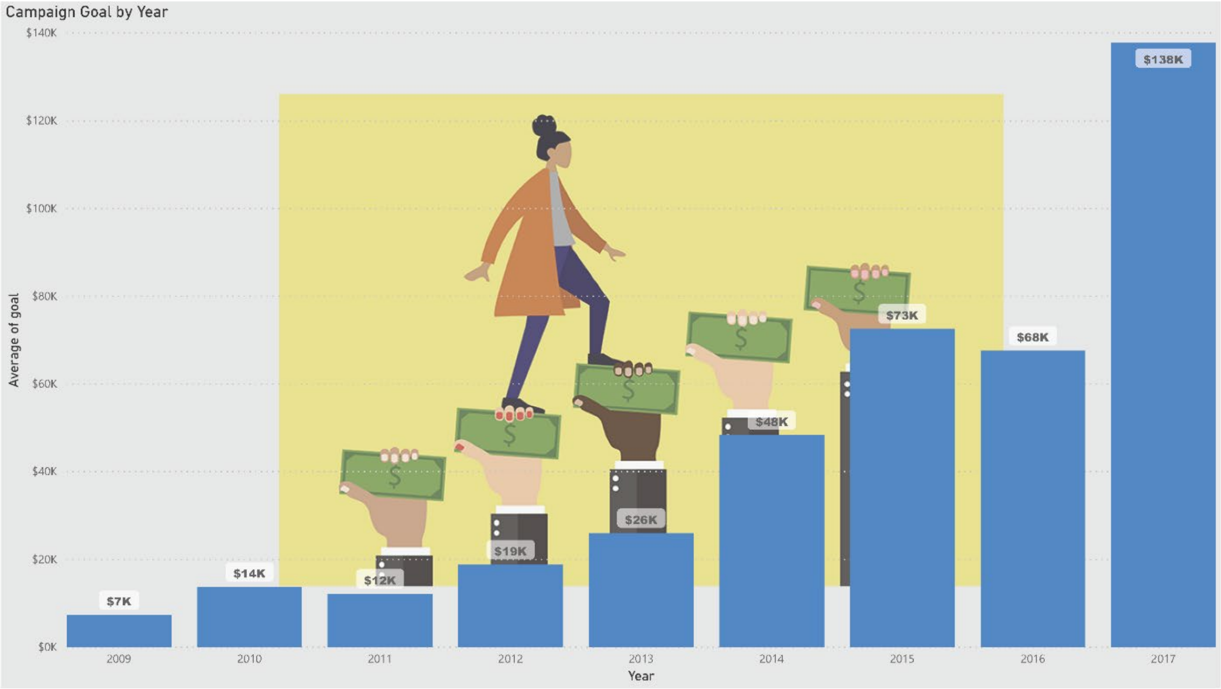

Figure 9:

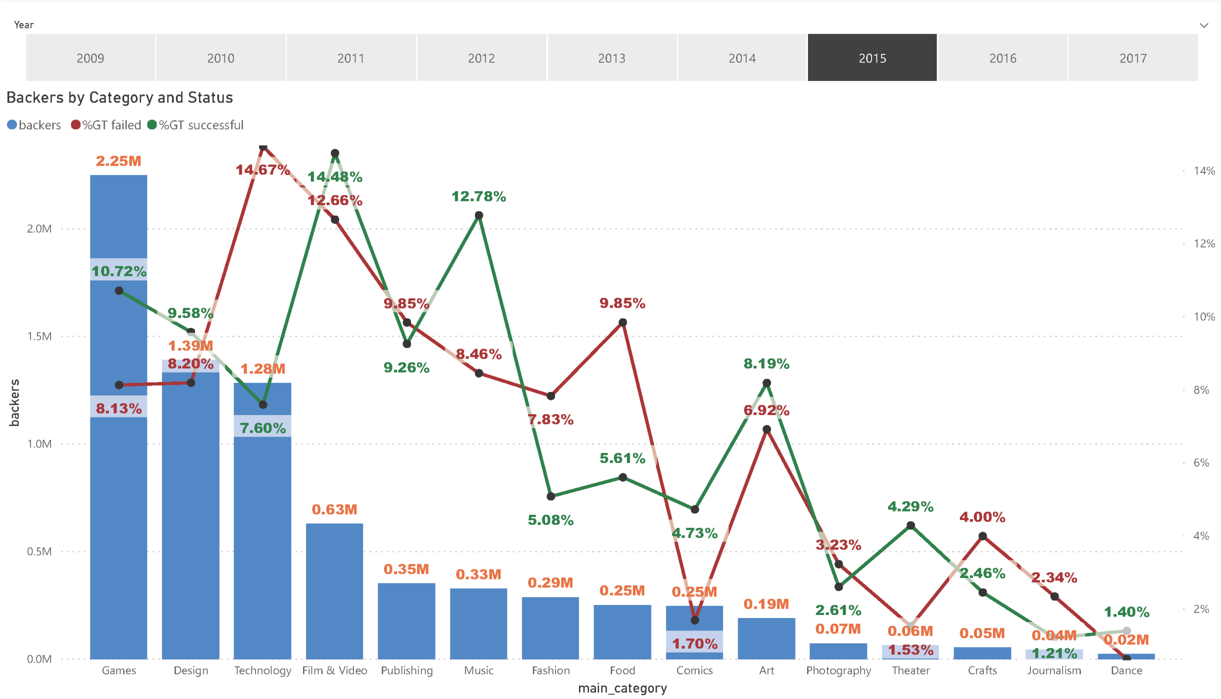

Figure 10:

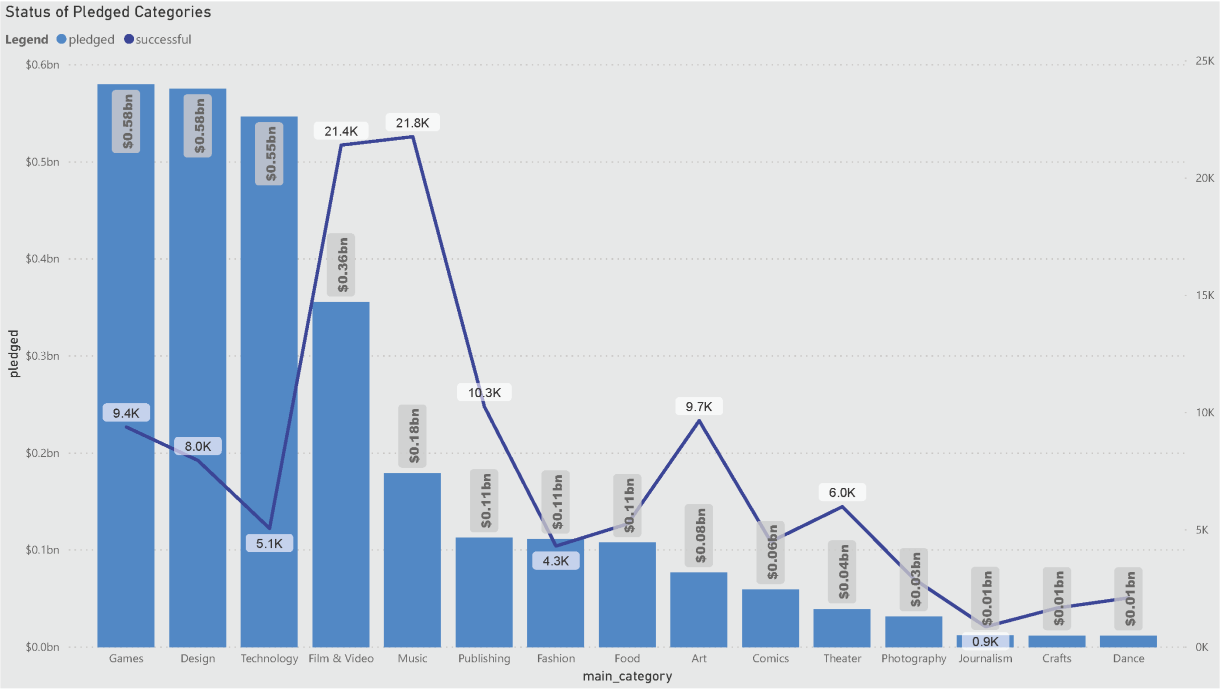

Figure 11:

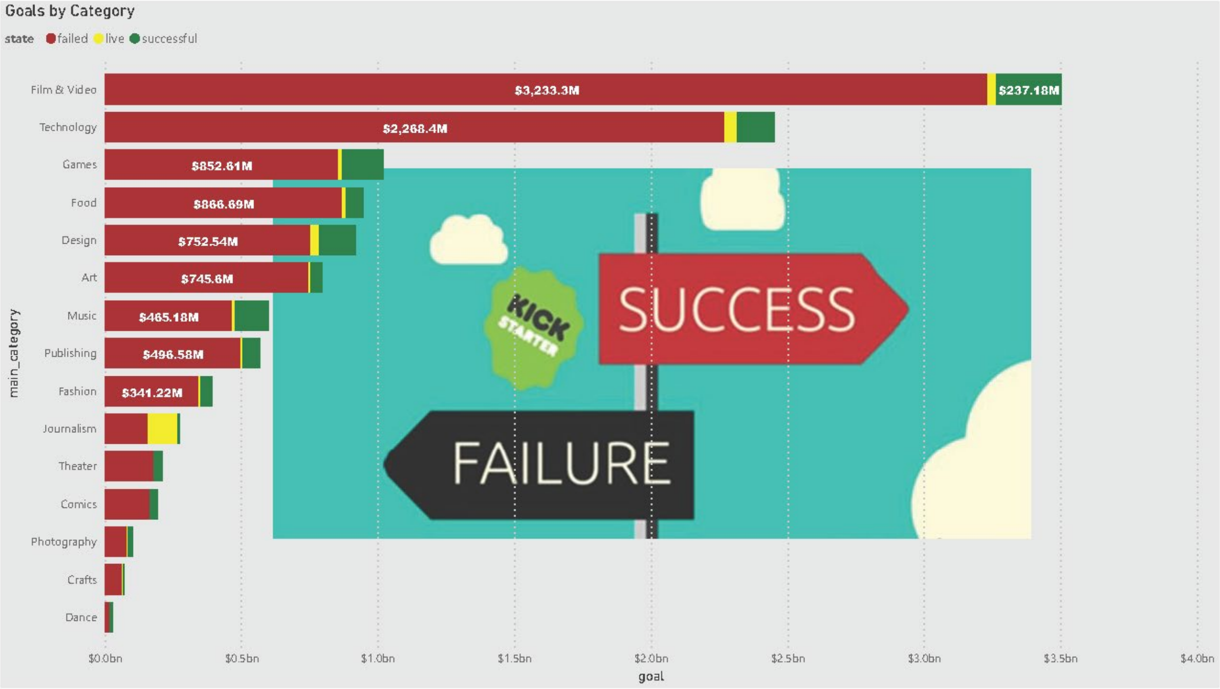

Figure 12:

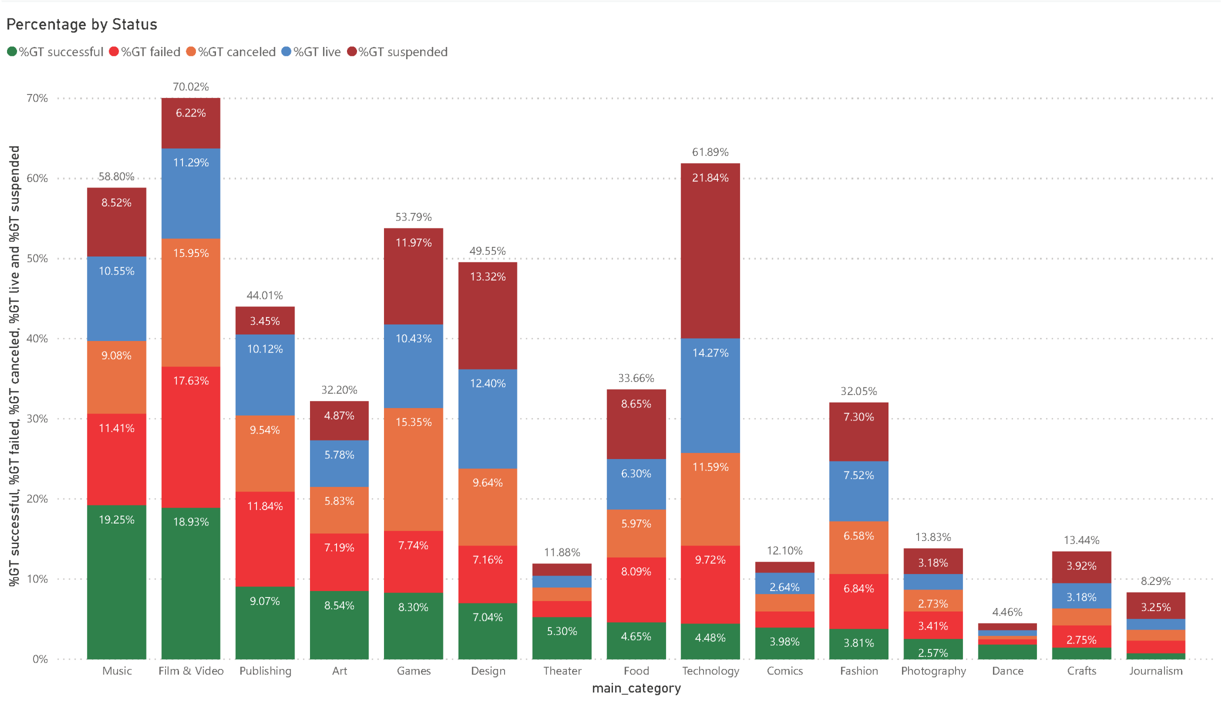

Figure 13:

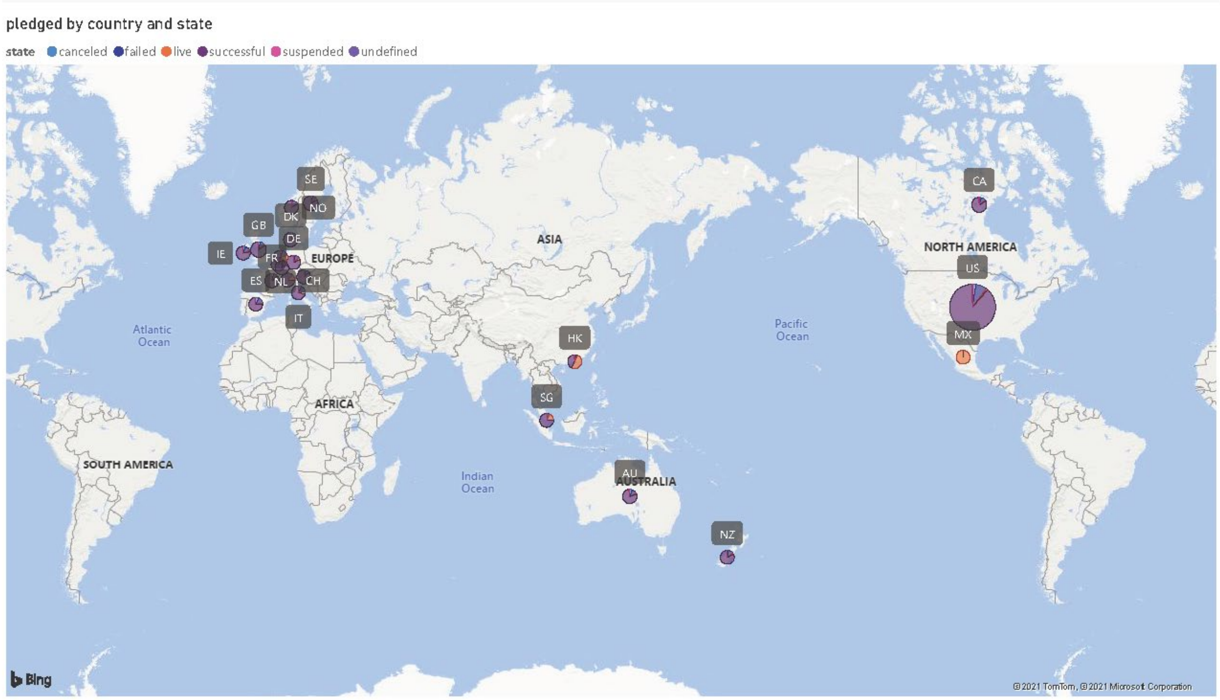
pledged by country and state
state ● canceled ● failed ● live ● successful ● suspended ● undefined
SE
NO
DK
GB
DE
IE
FR
EUROPE
ES
NL
CH
IT
ASIA
CA
NORTH AMERICA
US
MX
HK
SG
AFRICA
Atlantic Ocean
Pacific Ocean
Indian Ocean
SOUTH AMERICA
AU
AUSTRALIA
NZ
Bing
© 2021 TomTom, © 2021 Microsoft Corporation

Figure 14: