Success with Kickstarter

Becca Hicks, Victoria Luu-Hoang, Surya Nandipati, Elena Ojeda, Mark Vieta, and Tucker Zetterberg

05/09/2019

Crowdfunding has existed in various forms for decades. When a person has an idea for a product or service, often that idea requires money. Some require lots of money. It makes sense, then, to ask others to contribute financially in order to achieve success. Traditionally, a person can take out a large loan from a single entity, like a bank, or they can take lots of small loans and donations from a bevy of different entities. The latter is what is formally known as crowdfunding.

The internet has made it possible for millions of people around the world to connect instantly and exchange ideas. It’s no surprise that along with the advent of the internet, crowdfunding has seen a massive surge in popularity over the past 2 decades; it is easier than ever before to find people who will give money to help bring into existence a product or service they support. Today, there are dozens, if not hundreds, of crowdfunding websites, where anyone with an idea can pitch it to the world, and if the world likes what they see, they can help bring it to life. Since anyone can do it, the question becomes who is successful in doing so. What does it take to convince a total stranger that you and your product are worth investing in? What perks garner the most dollars; what legwork produces the best results. In order to tackle these questions, our group did what any responsible citizens would do: analyze the data.

Data Collection

We initially attempted to web-scrape information from several crowdfunding websites, but were unable to successfully get the information we needed succinctly due to the website layouts. Fortunately, we avoided needing to use web-scraping or other time-intensive data collection methods. Our data came from Kaggle, an online data science website that provides free data sets for amateur and professional data scientists to analyze. There was a data set on Kaggle of crowdfunding projects from one website, kickstarter.com. The data was in a .csv file, which made it simple to import into R, and included data on 378,662 projects. The hard part, as per usual, was the cleaning and organizing of the data into something we could use.

The original kickstarter dataset included fifteen variables.

1. A unique identification number for each project.
2. Name of each project.
3. A detailed category for each project.
4. A broad category for each project (15 total categories).
5. Currency abbreviation under which the project was launched.
6. Deadline for the project to raise funds.
7. Funding goal for each project in its native currency.
8. Date and time the project was launched.
9. Funding amount pledged to each project in its native currency.
10. State of the project at the time the data was pulled.
11. Number of backers that pledged funds to each project.
12. Country project was launched from.
13. Dollar amount pledged to the project.
14. Total money pledged to a project in all currencies converted to US dollars.
15. Project goal converted to US dollars.

A lot of our code was simple cleaning and manipulation. We changed dates from character values to date values. For one date variable that included both dates and time, we created two separate variables for each. Furthermore, since projects came from across the world, many project titles were in different languages and contained special characters. A big piece of our cleaning process was dealing with the special characters in names and how to remove them.

Additionally, we found a subset of data where the country abbreviation was missing and was listed as “N0”. Since it mostly showed up when the state of the project was listed as “undefined”, we dropped the undefined projects. We still had a small number of “N0” observations, so to make this data usable, we expanded currency abbreviations into their full names and ran a loop to match a country abbreviation to the currency. This left us with about twenty observations that had euros as the currency and thus could not be properly matched to a country. We decided to drop those observations.

Because Kickstarter was founded in 2009, we removed observations that had launch dates in 1970 since those must have been an error. Finally, we also created three new variables:

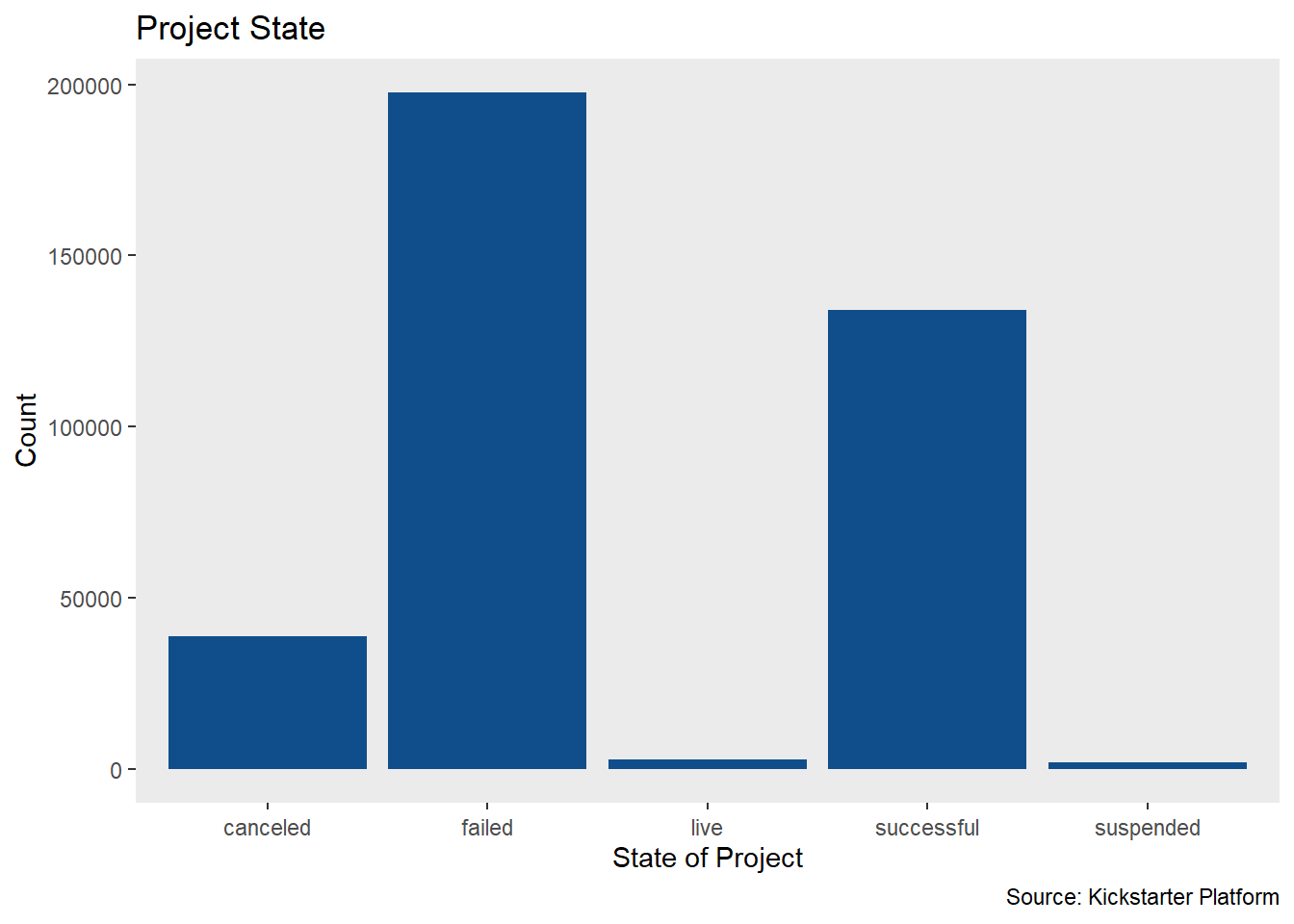
1. The number of days a project was open.
2. The amount of money a project was over/under their goal.
3. A numeric variable with the month the project was launched.

Research Questions and Background

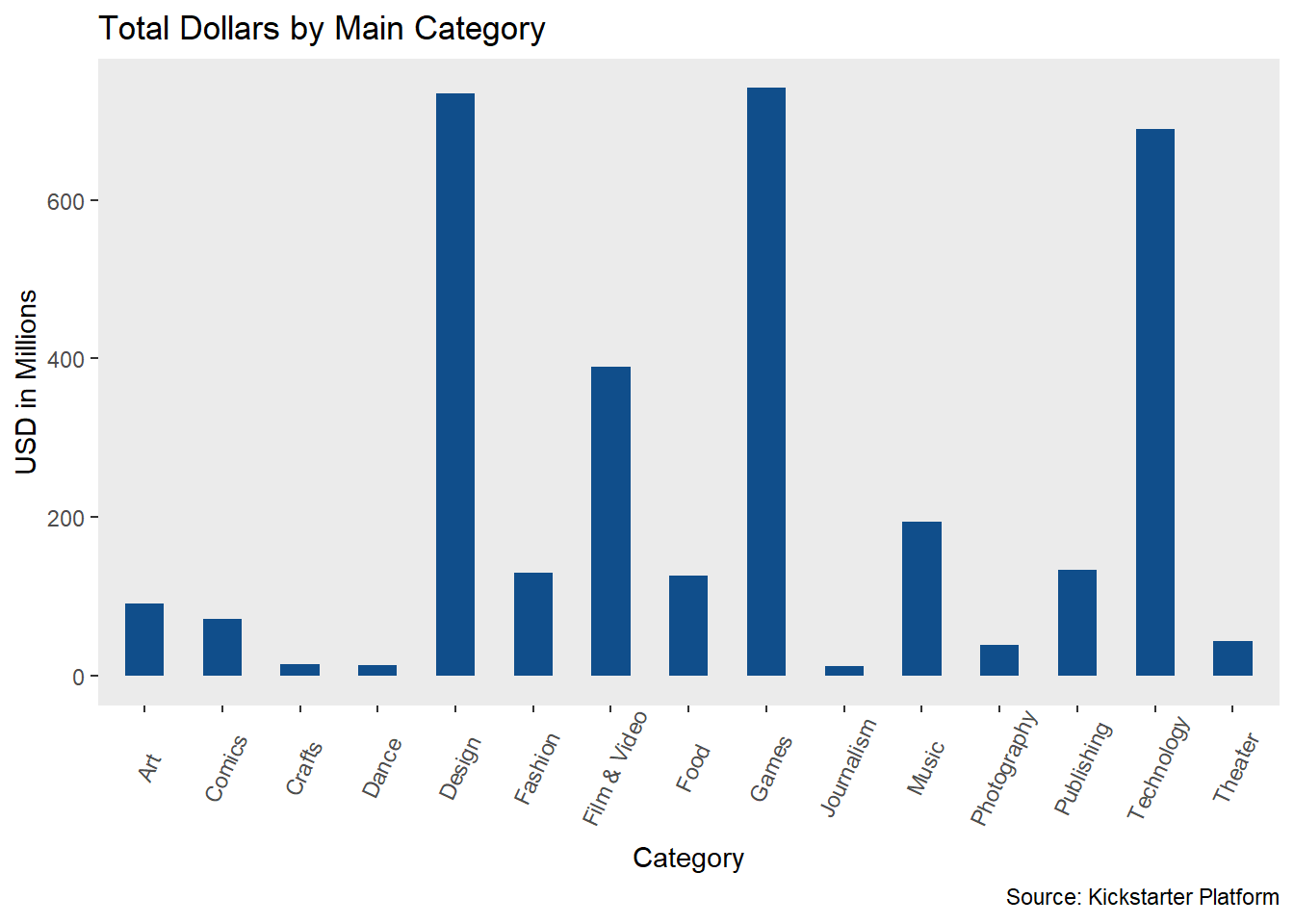
What variables can help us predict if a project will be successful?

While this is a broad question, going into this project, we did not know what data we would collect or what to expect. Nonetheless, we figured that for people wanting to crowdfund their projects it is of interest to know what their probability of reaching their goal is before they launch their Kickstarter campaign. We focused on answering this question with our data. Before proceeding to modeling our data, we visualized it to better understand it.

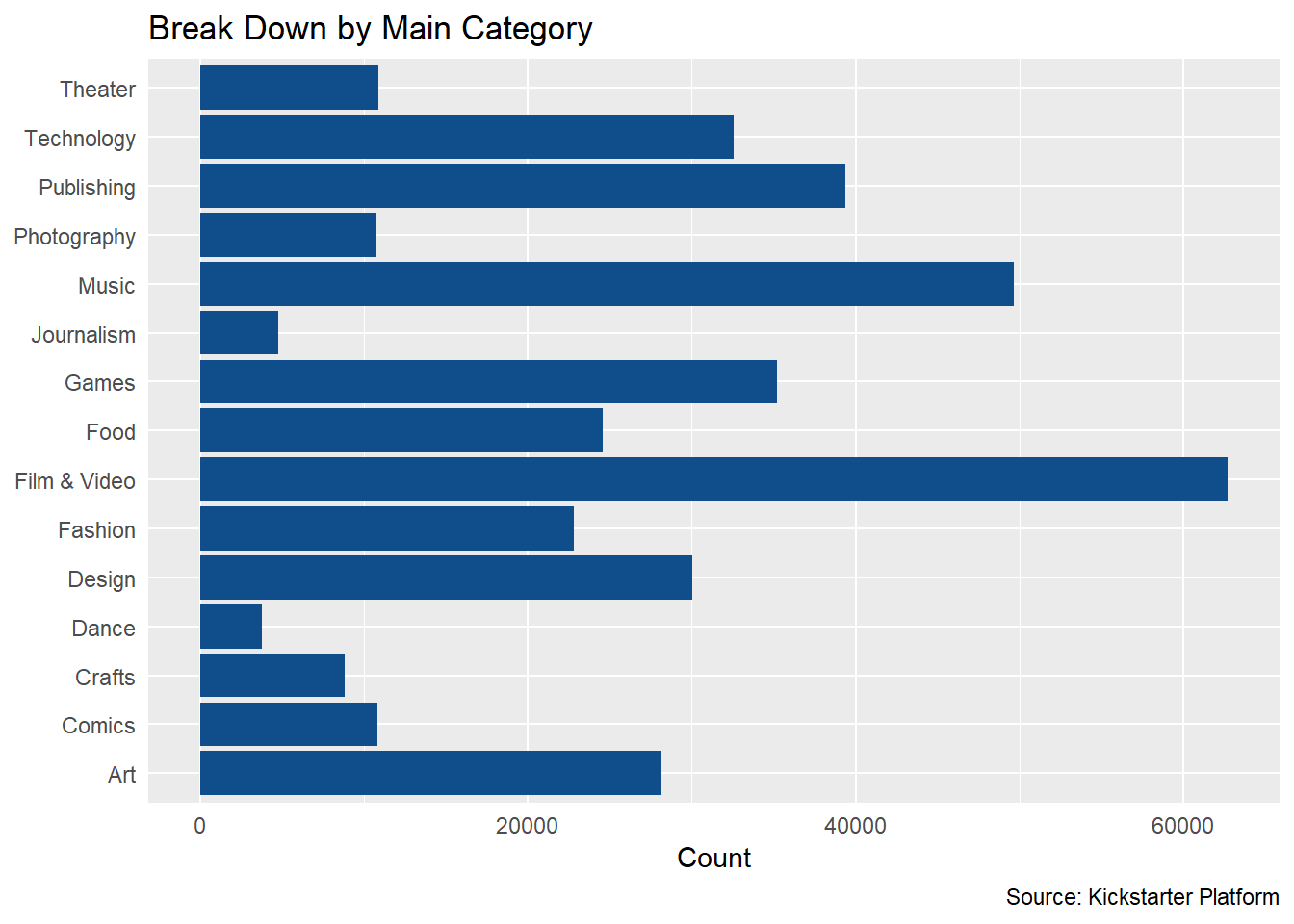
Looking at the success or failure rate of projects was interesting, although not unexpected. The majority of projects failed; 197,702, to be exact. A moderately less number of projects succeeded: 133,952. Perhaps the most interesting insight is the number of cancelled projects, which is more than one might expect, at 38,770.



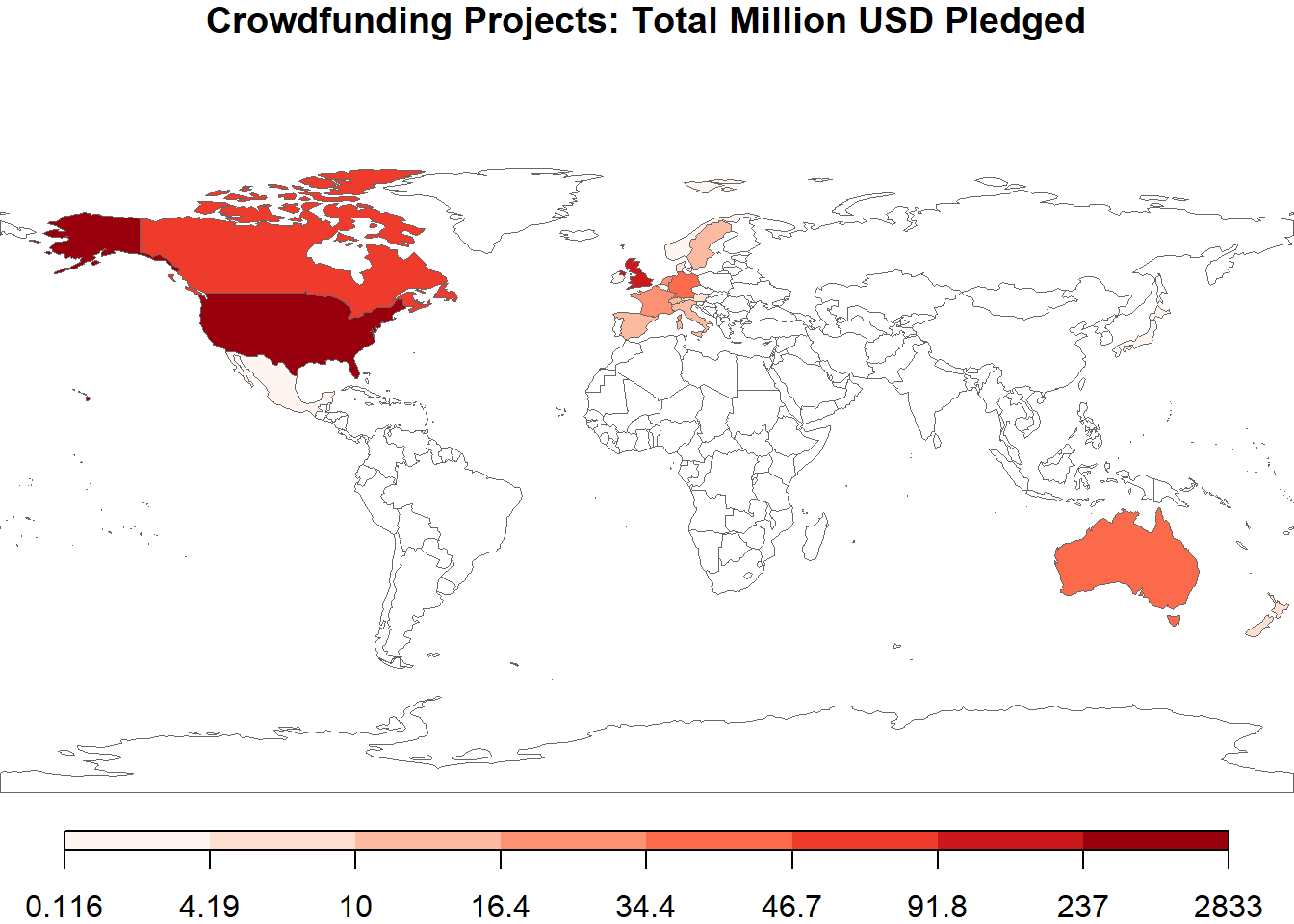
Looking at the projects by category, it is clear that design, games, and technology are leading the charge in sheer dollars. Conducting analysis on this data alone could yield even more interesting insights.



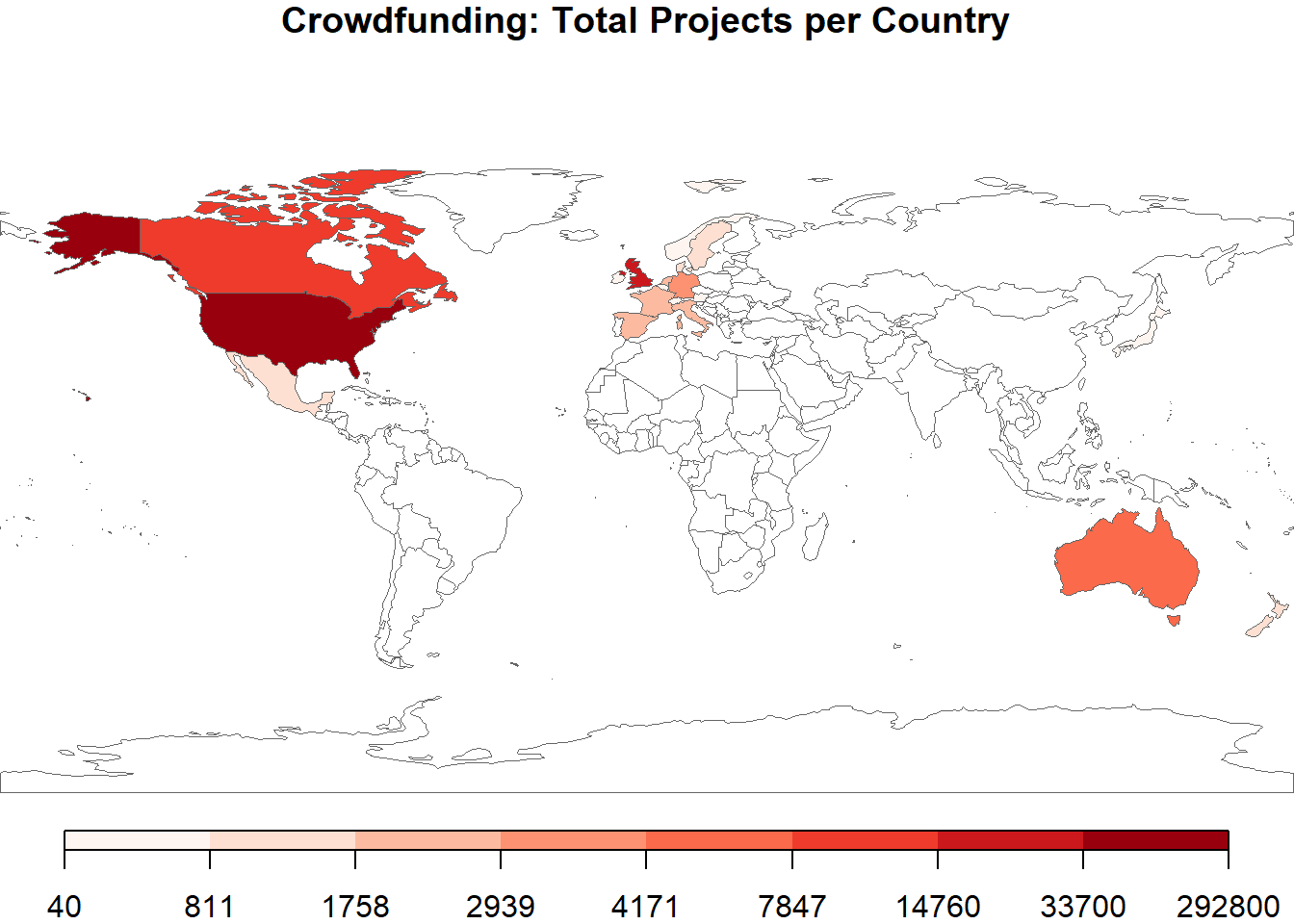
The chart below shows category by count, rather than by dollar amount. This provides another insight into the types of projects being launched. Comparing the two category charts, we can see that the project categories getting the most money are not necessarily the ones which have a larger share of raw counts of projects launched.



There are twenty-two countries represented in the data. We used heat maps to show the total money pledged for projects in their respective countries. The number of projects that were based in the U.S. was astonishing. Almost 300,000 of the 375,000 that we analyzed were in America. Coming in second place was Great Britain at around 33,700. Third was Canada at 14,700, and dramatically falling away from there. This could be due to several factors. Perhaps internet access is not as widespread in other countries; maybe crowdfunding simply isn’t as popular, or people have less money to spend on non-essentials. In any case, it is certainly true that America is leading the charge on crowdfunding, at least on Kickstarter.



Looking at the number of projects by country also shows that the most projects are coming from the U.S., followed by Great Britain.



Methods: Logistic Regression

Since we decided to try to explain what projects would succeed or fail, our main analysis method is a logistic regression model. Logistic regression is a classification method used when the dependent variable is qualitative and dichotomous. This method models the probability that an observation falls into one of two categories. The coefficients of a logistic regression represent the log odds and can be interpreted as follows: for every one unit increase in the independent variable, the log odds increase (decrease) by the coefficient value.

In our case, we are interested in classifying projects by state as either successful or failed based on multiple independent variables. To do this, we focus only on the observations that have a state of “successful” or “failed”. The initial independent variables are the fifteen categories, fourteen currencies, days project was open, month project was launched, and the project goal. We created dummy variables for the categories and currencies and used film & video and U.S. dollars as the baseline, respectively. (We had also fit a model using country dummy variables instead of currency, but the predictive power was almost exactly the same as with currencies. We decided against that model because there were twenty-two countries in our data set and the model had more variables. Our goal was to create the most simple model.)

We checked for collinearity between independent variables using the variable inflation factor (vif). It quantifies how much the variance of the coefficient estimates are inflated by the presence of multicollinearity. Vif values above five suggest multicollinearity, so variables exceeding this threshold are dropped from the model one by one starting with the highest value.

We also perform variable selection to reduce the number of independent variables in the model. We decided to use a backward stepwise selection method because a forward selection method kept a larger number of variables. Backward selection starts with a full model that includes all the independent variables and then eliminates the least significant variables one at a time using the Akaike information criterion (AIC) so that the final model has the lowest AIC.

The logistic regression model is fit using all of our data (331,654 observations). We tried fitting several models using subsets of data (75% of observations, 25% of observations) in an attempt to further reduce the number of variables in the model; one kept the same number of variables, while the other increased them.

Linear Discriminant Analysis

Another classification method we attempted is linear discriminant analysis (LDA). LDA first approximates the distribution of each independent variable for each category of the dependent variable and then uses Bayes’ theorem to turn them into conditional probabilities (James et al., 2017). We fit the LDA model and because the percentage of observations correctly predicted were similar to those of the logistic model, we decided to not keep the LDA model.

Results

The initial logistic regression model we fit with all independent variables. We calculated vif values for the variables and found no indication of multicollinearity. However, the model gave a warning message of separation. Separation happens when a certain value of an independent variable perfectly predicts what category an observation belongs in (Institute for Digital Research & Education). We determined that the variable causing the separation was the variable for the project goal amount. We dropped this variable from the regression model, which resolved the issue, before running the backward selection.

| Final Model Variables | | | | |
| --- | --- | --- | --- | --- |
|  |  |  | **95% Confidence Interval** | |
| **Variable** | **VIF** | **Odds** | **Min** | **Max** |
| Intercept | NA | 1.5631397 | 1.5188800 | 1.6086892 |
| main\_category\_Music | 1.344003 | 1.5606776 | 1.5247542 | 1.5974474 |
| main\_category\_Technology | 1.174519 | 0.4578698 | 0.4436290 | 0.4725677 |
| main\_category\_Publishing | 1.258186 | 0.7238653 | 0.7051848 | 0.7430407 |
| main\_category\_Food | 1.153135 | 0.5198232 | 0.5030008 | 0.5372081 |
| main\_category\_Crafts | 1.058073 | 0.4846287 | 0.4599821 | 0.5105960 |
| main\_category\_Games | 1.236736 | 1.0714942 | 1.0423929 | 1.1014080 |
| main\_category\_Comics | 1.085632 | 1.9970202 | 1.9132403 | 2.0844687 |
| main\_category\_Fashion | 1.142051 | 0.5345742 | 0.5165962 | 0.5531778 |
| main\_category\_Theater | 1.086216 | 2.3905759 | 2.2896306 | 2.4959717 |
| main\_category\_Art | 1.211512 | 1.0853538 | 1.0547692 | 1.1168253 |
| main\_category\_Photography | 1.078961 | 0.7115677 | 0.6804607 | 0.7440967 |
| main\_category\_Dance | 1.029823 | 2.5500166 | 2.3751531 | 2.7377540 |
| main\_category\_Journalism | 1.028549 | 0.4471694 | 0.4157502 | 0.4809629 |
| currency\_GBP | 1.020914 | 0.9430410 | 0.9194597 | 0.9672270 |
| currency\_CAD | 1.010455 | 0.7579509 | 0.7287572 | 0.7883141 |
| currency\_NOK | 1.000523 | 0.5930888 | 0.4924273 | 0.7143275 |
| currency\_SEK | 1.001612 | 0.7484794 | 0.6707777 | 0.8351820 |
| currency\_AUD | 1.007086 | 0.6794184 | 0.6434810 | 0.7173627 |
| currency\_EUR | 1.018528 | 0.6039145 | 0.5809413 | 0.6277963 |
| currency\_MXN | 1.001249 | 0.5746487 | 0.5097325 | 0.6478322 |
| currency\_NZD | 1.001699 | 0.8157337 | 0.7247507 | 0.9181383 |
| currency\_CHF | 1.001175 | 0.6904975 | 0.5799920 | 0.8220575 |
| currency\_HKD | 1.001840 | 1.4128798 | 1.1739008 | 1.7005093 |
| days\_between\_goal | 1.013994 | 0.9793712 | 0.9787870 | 0.9799556 |
| month | 1.002246 | 0.9966681 | 0.9945054 | 0.9988355 |
| *Note:* |  |  |  |  |
| Full currency names shown below. |  |  |  |  |

The final logistic model has 25 variables (shown above). The baseline category and currency categories are: Film & Video and U.S. dollars, respectively.

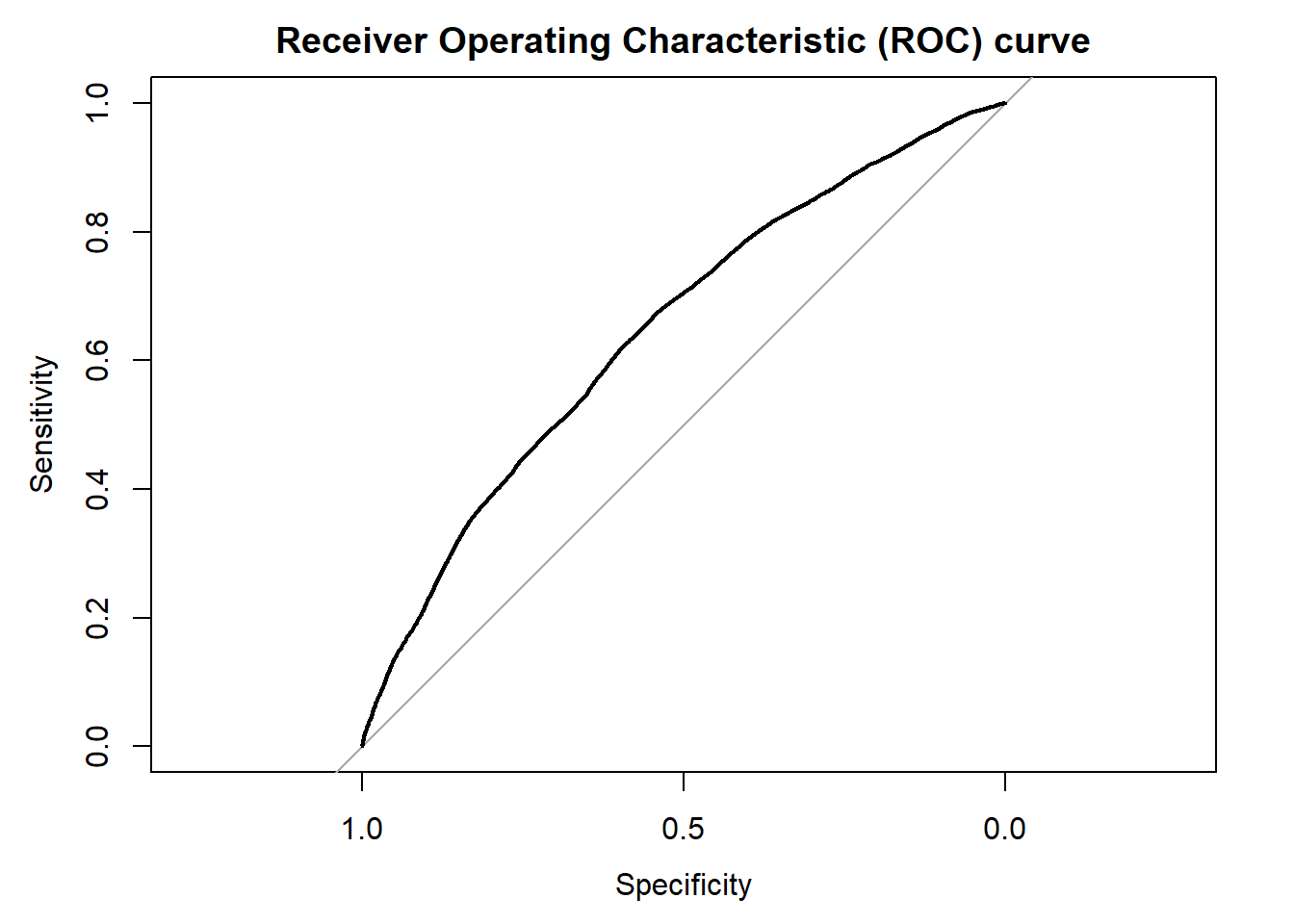
After fitting the final model, we checked for multicollinearity again. All vif values are less than five, which indicates that there is no collinearity between the variables. We also converted the model coefficients from log odds into odds, which allow for better interpretation of the values, where a value of 1 is an expected chance of success. For example, for every project in the music category, the odds of the project being successful are 1.5606776, which is an increase by a factor of 0.5606776. Also, for the quantitative variables, such as days between goal, for every one unit increase in number of days between goal, the odds of the project succeeding decrease by a factor of 0.9793712.

We test the prediction accuracy of our model with a classification table, where an observation in our data was classified as a “success” if the probability is greater than 0.5.

The table below shows that the mean accuracy of our model is 63.5%.

| Classification Table | | |
| --- | --- | --- |
|  | **Actual** | |
| **Predicted** |  |  |
|  | **Fail** | **Success** |
| Fail | 169808 | 93254 |
| Success | 27894 | 40698 |

Another measure of model fit is the receiver operating characteristic (ROC) curve. This plot shows the tradeoff between sensitivity and specificity. Sensitivity is the proportion of actual positives correctly identified by the model. In our case, this is the actual successful projects identified as successful. Specificity is the proportion of actual negatives correctly identified.In our case, this is the actual failed projects identified as failed. To assess model fit, we would like to see a large area underneath the ROC curve. The ROC plot for our model shows a less than ideal area underneath the curve, suggesting that the model is not a great fit.



Conclusion

Our final regression model had a prediction rate of 63.5%, which is decent, but not amazing. With more time, we could improve the prediction rate by dropping some more variables or otherwise refining our model.

Taking a closer look at the chart of odds values, we can see that dance and theater have the two highest values. It seems like dance and theater have a higher chance of success than any other category. However, in terms of dollars raised and number of projects, dance and theater rank among the lowest. Perhaps the odds are so high because the projects are less expensive overall; a smaller goal means a higher success rate. This is one of many insights that can be drawn from our analysis.

We had several problems implementing our regression model. First there was the error message of the separation issue with one of the variables, in which the variables were “overfitting” the model. Also, the regression stepwise selection code was taking excessively long to run on our computers; we attempted to utilize BUDDY to speed up the process. Initially, we had issues getting signed up and logged in to BUDDY; then there was an issue with Windows where BUDDY would not accept the session connection input. We were able to find a workaround. Research Op was used instead, which provides similar functionality.

Ideas for Future Projects

With more time, doing a deep dive into the categorical data could be fascinating. Design, games, and technology massively outearned all other categories; it might be interesting to see exactly what types of projects are being funded under these categories. Since we had more detailed breakdowns for each of the fifteen main categories, future work could focus on modeling the success outcomes in just one category at a time. Another interesting analysis might come from looking at perks offered to donors. We would then be able to answer questions such as: how much incentive is there for people to donate if they receive a free t-shirt compared to a free coffee mug? This last idea would require more complex web-scraping to compile the data.

Currency Legend for Final Model Table

GBP- Great Britain Pound

CAD- Canada Dollar

NOK- Norway Kroner

SEK- Sweden Krona

AUD- Australia Dollar

EUR- Euro

MXN- Mexico Peso

NZD- New Zealand Dollar

CHF- Switzerland Franc

HKD- Hong Kong Dollar

References

1. Currency Acronyms and Abbreviations | Discover. Retrieved from <https://www.easymarkets.com/int/learn-centre/discover-trading/currency-acronyms-and-abbreviations/>
2. Institute for Digital Research & Education. UCLA. FAQ What is complete or quasi-complete separation in logisttic/probit regression and how do we deal with them? Retrieved from <https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faqwhat-is-complete-or-quasi-complete-separation-in-logisticprobit-regression-and-how-do-we-deal-with-them/>
3. James,G., Witten,D., Hastie,T. and Tibshirani,R. (2017) *An Introduction to Statistical Learning*. Retrieved from <http://www-bcf.usc.edu/~gareth/ISL/index.html>
4. Kaggle: Your Home For Data Science. Kickstarter Projects Mickael Mouille. ks-projects-201801 [Data file]. Retrieved from <https://www.kaggle.com/kemical/kickstarter-projects>
5. Wikipedia. Iso 3166-1 Alpha-2. Retrieved from <https://en.wikipedia.org/wiki/ISO_3166-1_alpha-2>