Chris Lunger
Deeksha Yennam
Reece Wooten
Yue Wang

Pattern Analysis of Kickstarter Projects

1 Overview of the project

We investigated the crowdfunding industry to get a basic understanding of its attributes. More importantly, we wanted to know the key attributes to successful crowdfunding projects. We used data from Kickstarter, a famous crowdfunding website launched in 2009 with a key focus on creativity. Kickstarter helps artists, musicians, filmmakers, and other creators find funding to support their projects. Creators provide information about their projects on the platform, and then set a timeframe for the funding process, the minimum fund they need, and how they will reward their backers. Then, backers can choose which projects to fund and how much they are willing to pledge. If creators cannot meet their minimum funding requirement by the deadline, then they don't receive any funds. If their funding process is successful, Kickstarter will apply around an 8% fee as commission.

We retrieved the Kickstarter data from Webrobots. This site has a publicly available Kickstarter dataset that is updated with information about all current and previous projects on Kickstarter every month. We used the most recent dataset available on the website, which was built on July 15th, 2017. As part of data processing, we extracted the project category and the location details from the raw dataset using regular expressions. In addition, to better perform exploratory analysis, we also combined U.S. population statistics for each state to our main dataset. Our final dataset had 49813 observations after randomly subsetting the original 250,000+ data set.

2 Importance of the problem

The crowdfunding industry has become increasingly popular in recent times, and is now a very important medium for creators to fund their projects. This makes it essential to answer questions about what the crowdfunding industry looks like, what category of projects are more likely to be funded and what factors can drive the success of a project.

There are four major distinct types of crowdfunding, reward, equity, donation and lending, and each type has its own platform. In this project, we only focus on the data from Kickstarter, a reward-based platform that emphasizes creativity. But we can still use our results to draw inferences about the whole industry.

3 Exploratory analysis

3.1 Countries & Locations

We organized our analysis to first look at the macro environment of Kickstarter crowd funding, and increasingly got more granular. First, we analyzed Kickstarter projects at the country level. Kickstarter's main service is in the U.S, although it has expanded to other countries, such as Canada, and Australia. Fig.3.1.1 shows that a large majority of projects are U.S. projects, 38066 out of 49813. The second large country is U.K with 4236 projects, and the third one is Canada, having 1934 projects. Therefore, the dataset is unbalanced in terms of countries included, and we decided to focus only on the U.S. projects

We then investigated the relationship between different states and projects as shown in Fig.3.1.2. Consistent with conventional knowledge, California, New York, and Texas have the largest number of projects, as these three states are highly developed. However, we decided to consider population, as these states are heavily populated as well. Fig.3.1.3 shows the per capita number of projects, and the graph is smoother. The top three states by per capita are Washington D.C., Oregon, and Vermont.

Next, we examined the project success rate by location. The overall success rate for the U.S. is 44.27%, far above other countries. For example, the success rates for the U.K, Canada, and Australia are 14.21%, 38.42%, and 31.11% respectively. Then, in terms of different states, Vermont has the largest success rate, 60.18%, followed by Rhode Island and Massachusetts, with 57.86% and 56.95% respectively(Fig.3.1.4). Based on this analysis, we concluded that state could be a potential predictor to predict the success rate of a project.

3.2 Categories

We looked for patterns in the data to see how differences in the category of the project affects various factors like number of backers, goal of the project and amount of funding received.

All the projects in the dataset are divided into 15 categories such as art, technology, music etc. Fig.3.2.1 shows the number of projects in each category, each bar also marked in orange by the proportion of projects which were successful. We observe that music, film, publishing, technology and art come up as the top five categories. Next, we looked at the average length per project in each of the categories. As shown in Fig.3.2.2, irrespective of the category, the average length of a project seems to remain fairly constant between 30-35 days.

Fig.3.2.3 and Fig.3.2.4 show the total goal and average goal per project by category. Film, technology and art come up as top categories in terms of the cumulative goal but journalism and theater are top categories if we go by mean project goals. However, we notice that journalism and theater have very few projects. Hence, average goal does not give a fair idea of which categories are most popular.

Fig.3.2.5 shows the average number of backers per project by category. Games and design followed by technology have the highest

number of backers per project, indicating that these categories are more popular among investors. Fig.3.2.6 shows that tabletop games have the highest success rate among all projects.

3.3 Kickstarter's View

Kickstarter stays in business by having projects being successfully funded. If a project is fully funded in America, 5% of the total amount of funds raised plus 3% per contribution is taken as a fee for Kickstarter. Fig.3.3.1 shows that the top 5 categories where Kickstarter receives the most money is from technology, games, design, film, and music. When adjusted by the number of backers per category,as in Fig.3.3.2, we see that the top 5 categories are technology, design, fashion, dance, and film. This information can point Kickstarter to the projects in their respective categories which give Kickstarter the most money. These categories, perhaps are the ones which Kickstarter should try and advertise more on their staff pick page so that they draw more attention, and therefore increase their cut per backer. When investigating if Kickstarter already chooses their staff picks based upon their most lucrative categories, as in Fig.3.3.3, we see that the top five categories that get the most staff picks are dance, comics, games, design, and publishing. We see that only two categories overlap between the most lucrative categories and staff picks, dance, and design. While the posted reason for staff picks are 'cool projects picked by our staff', it might be more advantageous for Kickstarter to pick from the categories which receive the most Kickstarter cut per backer.

3.4 Time Series

868 of the projects we observed were actively occurring. Our goal was to develop an analysis of how Kickstarters function over time, using the mean of each category over time to function as a proxy for characteristics of a category at different points.Our thoughts were that different categories may be funded in different ways: some may get funded quickly and then drop off, or that some may get funded late or more steadily.

The main issue we faced is that 868 is too few data points to gain any sort of real insight as to the function of 15 groups over 10 different points in their funding cycle. Fewer than 10 points would have drawn a figure that was too crude to be of any use.

What we could find from Fig.3.4.1 was that most of the funding of projects occurred during the first half of their project length and plateaued around 45%. The results dip at the 60-70% mark because these are all different projects and not just following the same projects over different points in time. Were we to have more projects or follow the same projects over time, we believe this would follow a more upward shape.

4 Predictive Model and Insights

After exploring the data, and finding patterns between the various variables of the data set, we ran a series of predictive models to see if we could predict whether a project would be successful based upon what the project creators see. This means we picked the predictors which only the project creator knows, *Project length in days*, *Goal*, *Category*, *State*, and *Sub-Category.* We decided not to use *staff-pick*, *number of backers,* or any other variables which the project creator won't see before beginning his/her project.

For all of our predictive models we compared our test accuracy, to the baseline accuracy rate which is the distribution of failed projects in the test data set, 57.5%. The baseline accuracy rate is the best guess of whether or not a project will fail without any predictive modeling. We first started with a logistic regression, which did relatively poorly compared to the baseline accuracy, but gave us a good indication of which variables are important. The top five variables which gave more weight towards a successful project was: *Tabletop games*, *product design*, *music*, *design*, and *games*, all of which are either a category, or subcategory. The variables which predicted the most strongly for an unsuccessful project was: *Project length in days*, *technology*, *food*, *fashion*, *Florida*. Most of the unsuccessful predictors were categories except for Florida, which would be interesting for further research.

The next predictive models we conducted dealt with trees, after testing a tree model, random forest, and boosting against a training and test set, the best model ended up being the boosting model with 500 trees, a learning rate of .01, and a depth of 4. This indicates that while each tree needs to be relatively complex, only small amounts of those complex trees are included in the final model. Our largest accuracy rate in the test set was 72.2% which is 14.7 percentage points above baseline accuracy. We believe that this model can be used by both Kickstarter and the project owners to determine whether or not a project will succeed on their platform. This can also be an indicator of the types of projects the crowdfunding industry funds. More research should be conducted in the other crowdfunding websites like Indiegogo to see if there are similar patterns.

5 Conclusion

We believe this analysis can be valuable to both Kickstarter and the project creators. Kickstarter can use this report to better market the categories which provide them the most return. Also, the project owners can use this report to understand which categories and states are most favorable for their projects and focus their advertising accordingly. These insights can be used to better optimize the crowdfunding process by connecting backers and project owners more efficiently.

6 Appendix

Exploratory Analysis

3.1 Countries and locations

Fig.3.1.1 : Country Vs. No. of Projects (Excluding outliers)

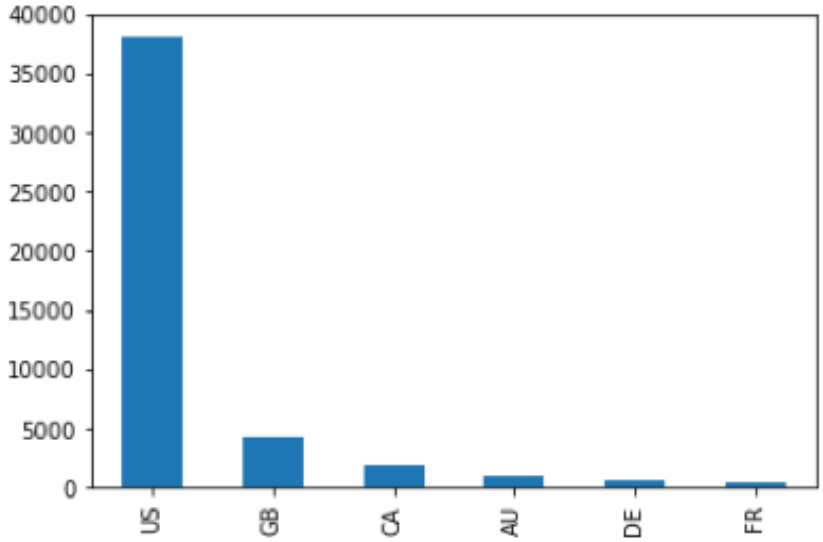


Fig.3.1.2 : No. of projects by states in the US

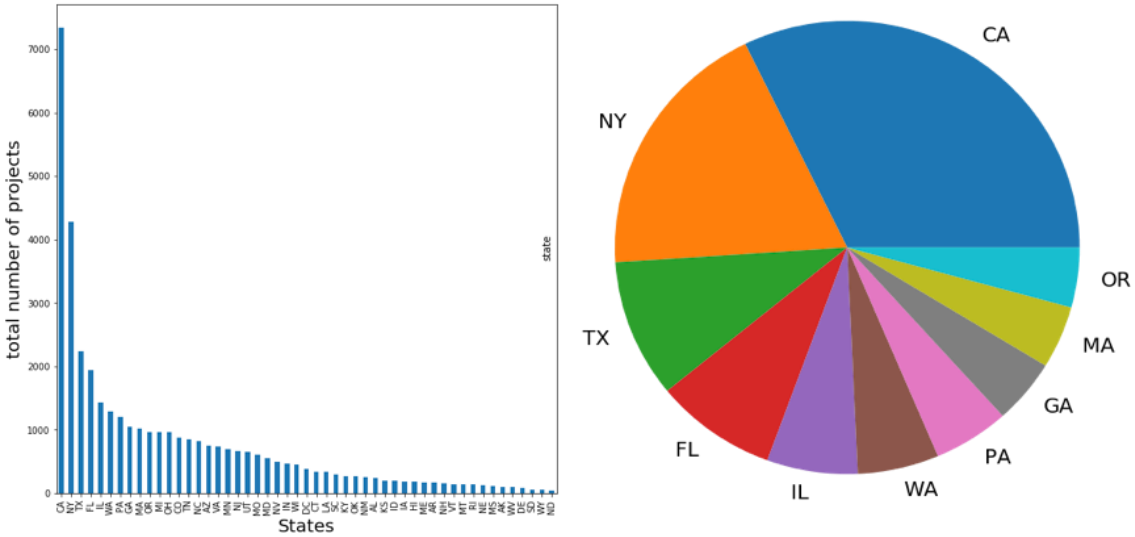


Fig.3.1.3: Per capita number of projects by states in the US

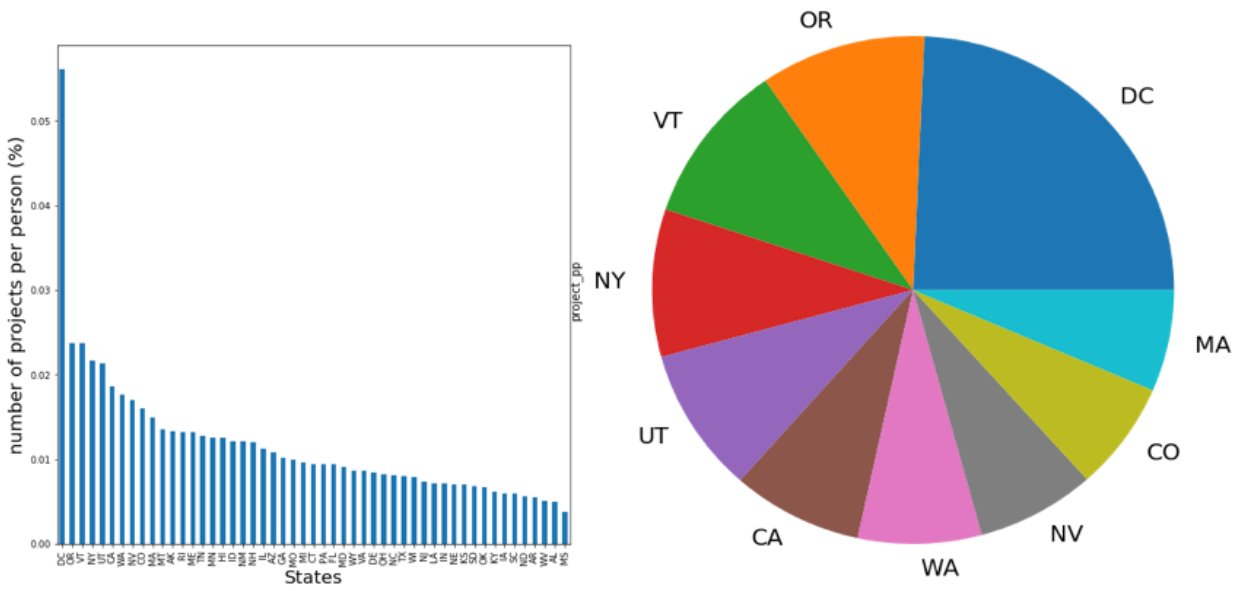
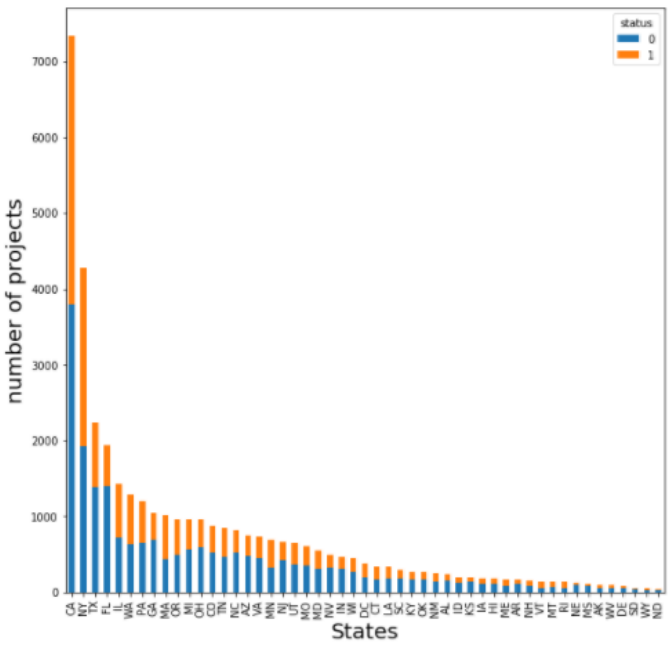


Fig.3.1.4: Success Rate of projects by State

state	Success Rate	No. of Projects
VT	0.6081	148
RI	0.5786	140
MA	0.5694	1015
NY	0.5500	4278
MN	0.5318	692
MT	0.5177	141
WA	0.5070	288
IL	0.5003	437
CT	0.4918	970
CA	0.4818	7335



3.2 Categories

Fig.3.2.1 : No. of projects by category

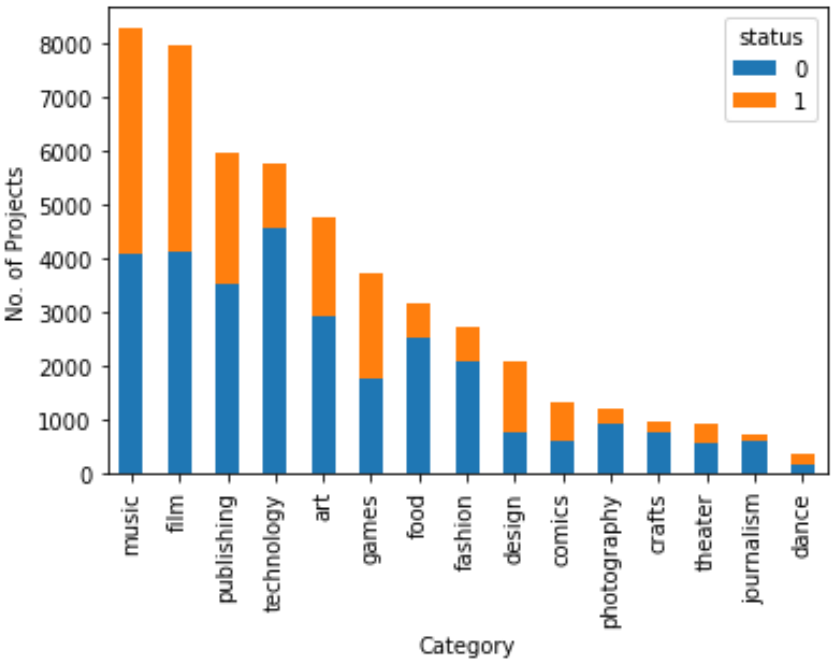


Fig.3.2.2 : Average project length by category

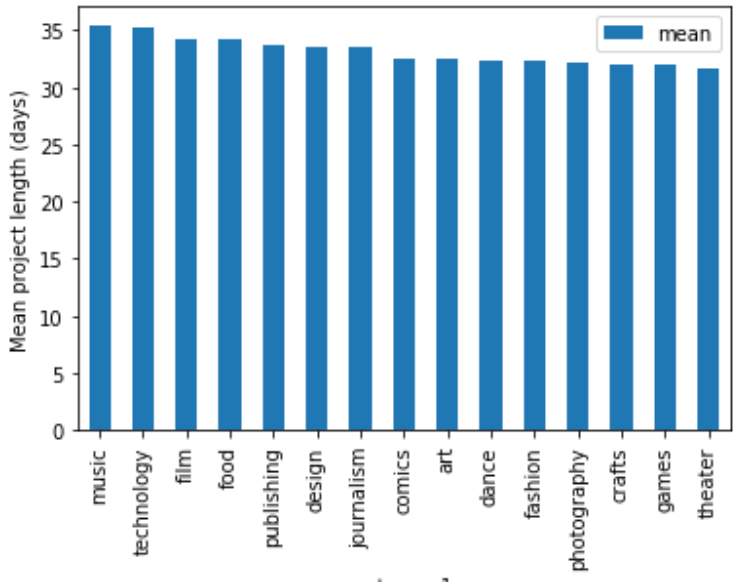


Fig.3.2.3 : Cumulative goal in $B by category

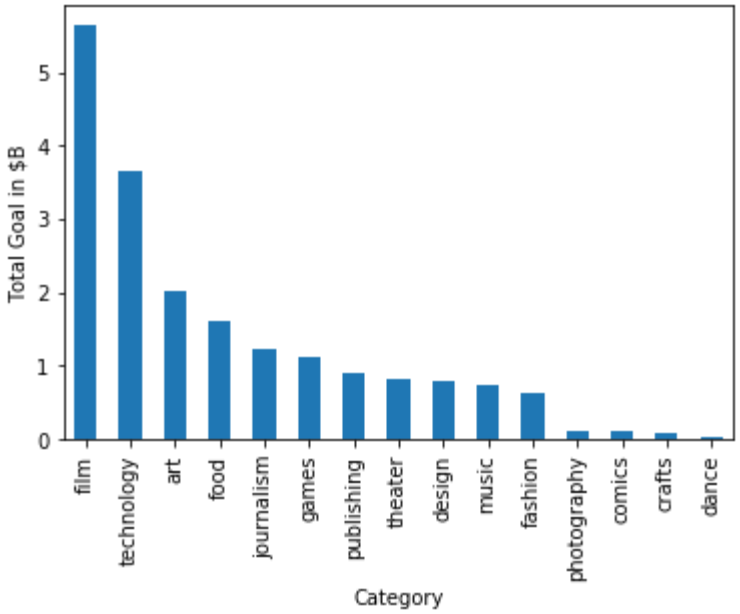


Fig.3.2.4 : Mean goal per project in $ by category

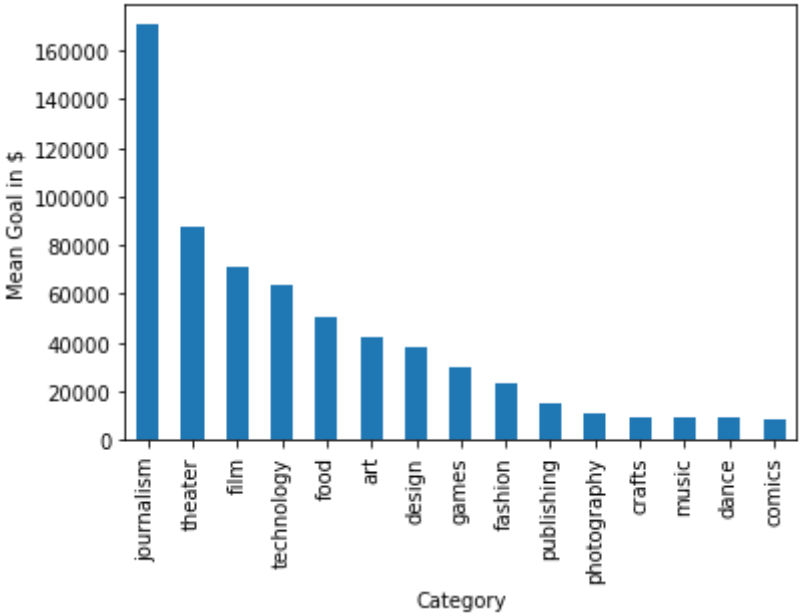


Fig.3.2.5 : Avg. no. of backers per project by category

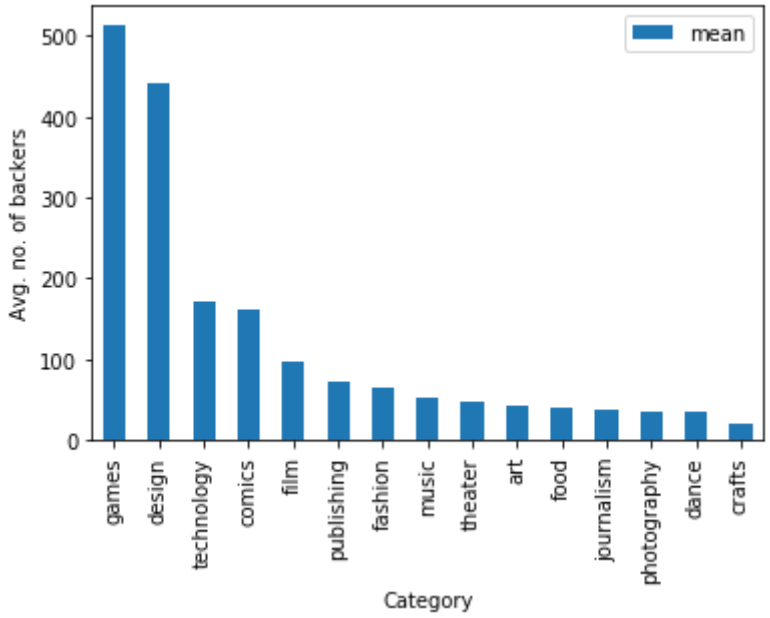
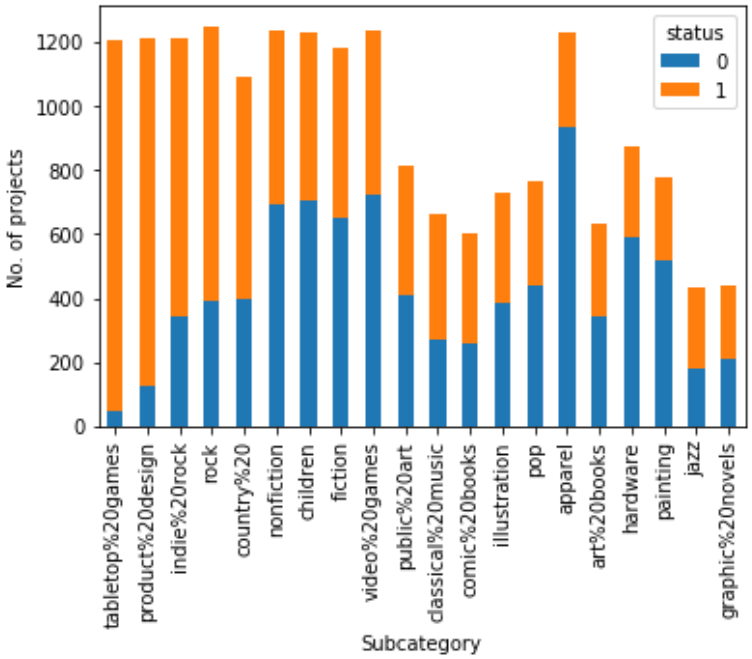


Fig.3.2.6 : No. of projects by subcategory



3.3 : Kickstarter’s cut

Fig.3.3.1 Kickstarter’s Cut vs. Category

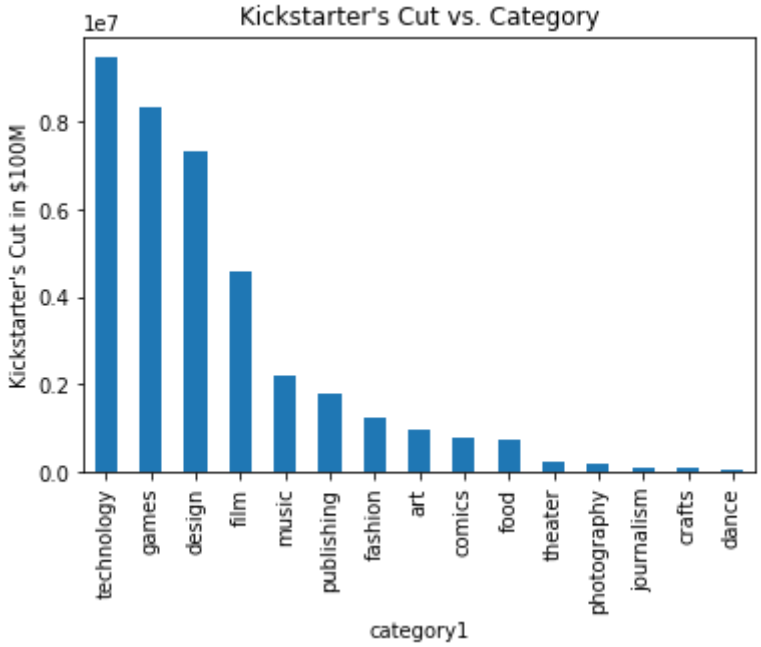


Fig.3.3.2 Kickstarter’s Cut Per Backer vs. Category

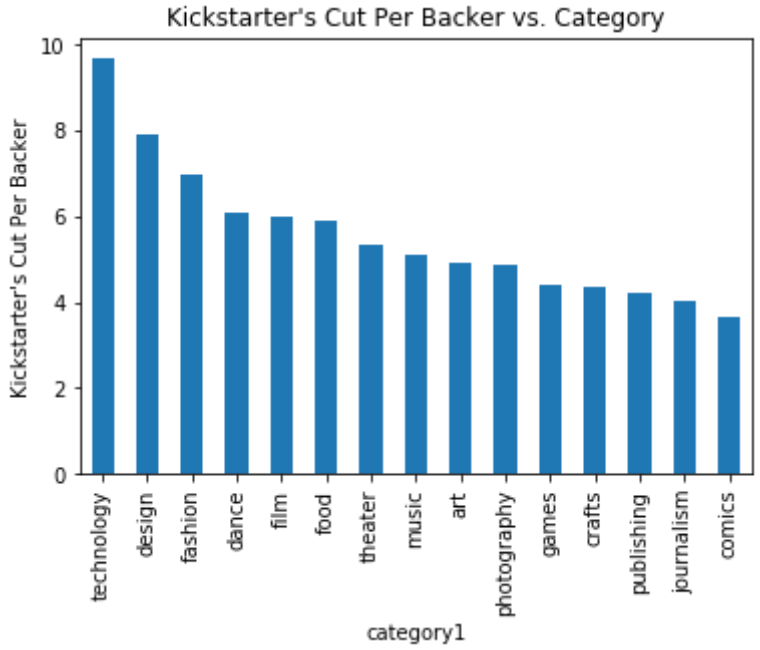


Fig.3.3.3 Staff Pick Percentage vs. Category

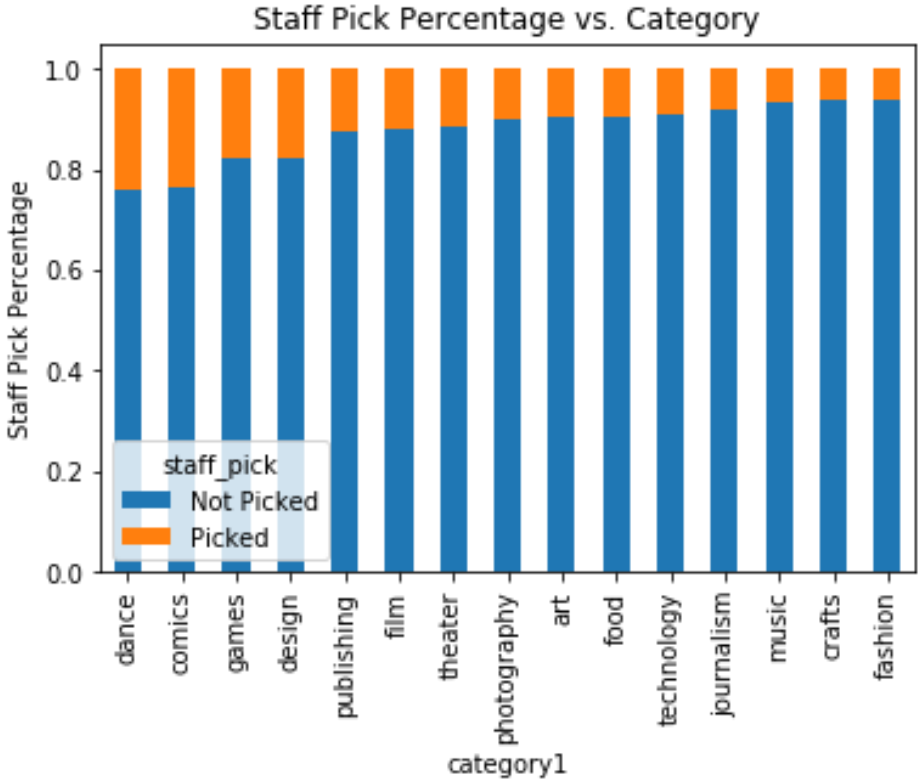


Fig.3.4.1 : Time Vs. Average Percent Funded

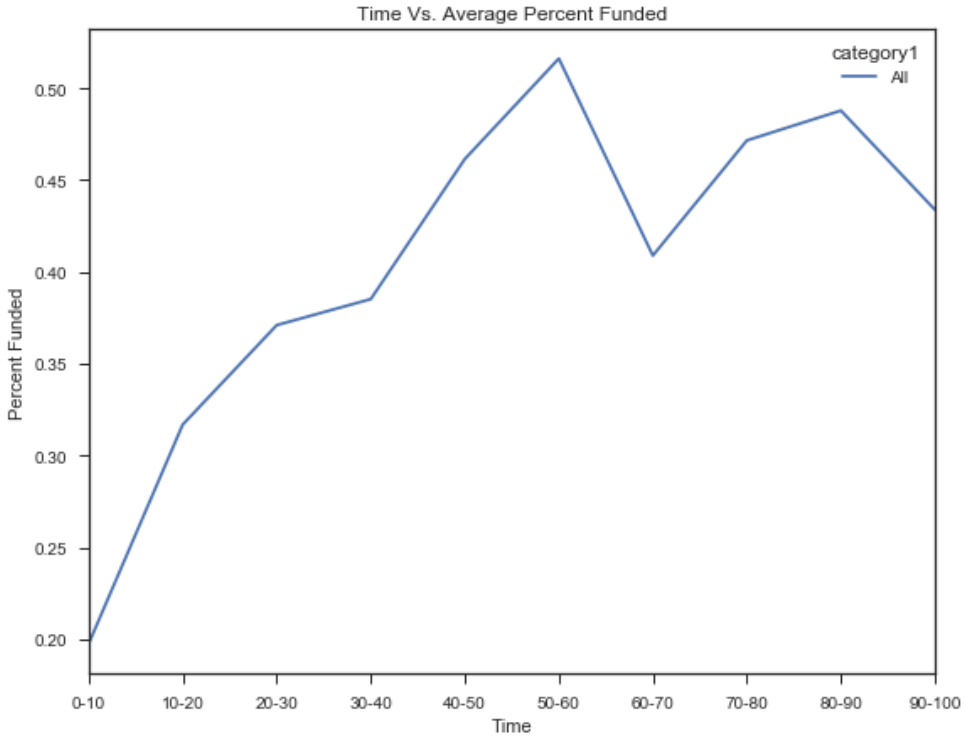


Fig.4.0.1 Confusion Matrix

		Predicted	
Actual		Not Funded	Funded
	Not Funded	4744	812
	Funded	1966	2133