Predicting the Success of Kickstarter Campaigns

Peter (Haochen) Zhou

STAT 157 Final Project

Professor Aldous

December 5th, 2017

1. Introduction

Over the past decade, we have witnessed an increasing popularity in crowdfunding. Crowdfunding is the practice of funding a project or venture by raising many small amounts of money from a large number of people, typically via the Internet. Due to its easy accessibility and broad exposure to the public, entrepreneurs and aspiring creators seek to utilize the platform to raise money for their projects. According to crowdfunding research firm Massolution, crowdfunding grew 167 percent in 2014. To put that into dollars, crowdfunding platforms raised $16.2 billion in 2014, escalating to $34.4 billion in 2015. For 2016, crowdfunding trends are predicted to pass VC funding for the first time.

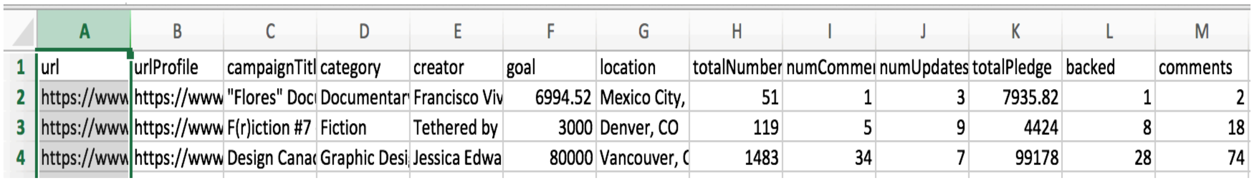
Kickstarter, launched in 2009, is one of the major crowdfunding platforms along with Indiegogo, RocketHub and GoFundMe. With a mission to help bring creative projects to life, Kickstarter has attracted over $13 million backers to successfully fund 134, 767 projects with a total amount exceeding $3.4 billion to date. In Kickstarter, projects are listed into 15 categories: Art, Comics, Crafts, Dance, Design, Fashion, Film and Video, Food, Games, Journalism, Music, Photography, Publishing, Technology and Theater. In Kickstarter, “Creators” are people behind the projects who are seeking funding. “Backers” are people who pledge money for projects they believe in and “pledges” are monetary contributions towards the projects.

Kickstarter has a unique “All-or-Nothing” model, meaning unless a project reaches its funding goal, no backer will be charged any pledge towards a project. On the one hand, it is less risky for backers because projects that are not fully funded have less likelihood of completion. On the other hand, it adds more incentives for creators to connect with the public and to finish their projects. Another unique feature of Kickstarter is that backers will only be rewarded in experience or creative products instead of equity. In addition, Kickstarter claims no equity in creators’ projects and creators maintain full ownership of their work.

Although numerous projects enjoyed tremendous success on Kickstarter, the aggregated average success rate is declining over the years. Data have shown that the aggregated average success rate of Kickstarter campaigns was 43.70% in 2011, 43.56% in 2013 and 35.82% at May 2017. In this paper, I seek to explore factors that affect the success of Kickstarter campaigns. I believe that success rate differs across different categories and prior researches have shown that Film & Video and Music are the largest categories and have raised the most amount of money. Film & Video, Music and Games combined account for over half of the money raised. My project attempts to predict the success of Kickstarter campaigns by analyzing the significance of factors identified in the data. I hypothesize that among all factors listed, the creator's prior success, goal, numbers of perks, creator's Facebook friends, total word count in the description are more significant than other factors.

2. Data and Methods

I collected my data from Kaggle.com, the open online database. The data contain 4000 most-backed projects and 4000 live projects with limited information such as the pledged amount, category, goal, location, blurb and number of backers. Another dataset contains 3652 Kickstarter projects in 2017 with comprehensive information such as a project's creator, goal, main category, duration, number of comments, number of updates and creator's Facebook friends. There are 51 unique characteristics for each project and I intend to use this as my main source of data. Here is a snapshot of my dataset:



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	url	urlProfile	campaignTitl	category	creator	goal	location	totalNumber	numComme	numUpdates	totalPledge	backed	comments
2	https://www	https://www	"Flores" Doc	Documentar	Francisco Viv	6994.52	Mexico City,	51	1	3	7935.82	1	2
3	https://www	https://www	F(r)iction #7	Fiction	Tethered by	3000	Denver, CO	119	5	9	4424	8	18
4	https://www	https://www	Design Cana	Graphic Desi	Jessica Edwa	80000	Vancouver, C	1483	34	7	99178	28	74

Since the data is pre-processed, I will first clean the data in R to convert them into a "ready-to-use" format. Secondly, I will perform linear regression in Stata to explore the significance of key variables to test my original hypothesis that the creator's prior success, goal, numbers of perks, creator's Facebook friends and total word count in the description are more significant than other factors. In addition to linear regression, I will generate plots to visualize the data and further explore the relationship between success rate and key factors. Due to Kickstarter's "All-or-Nothing" model, I will limit the scope of my project and only analyze projects with success=1 ("success" is a dummy variable and takes the value of 1 when the project meets its funding goal and 0 otherwise).

Given the declining aggregated average success rate of Kickstarter campaigns, I hope this paper can uncover the myths behind the success and serve as a prediction of success for creators and backers.

3. Literature Review

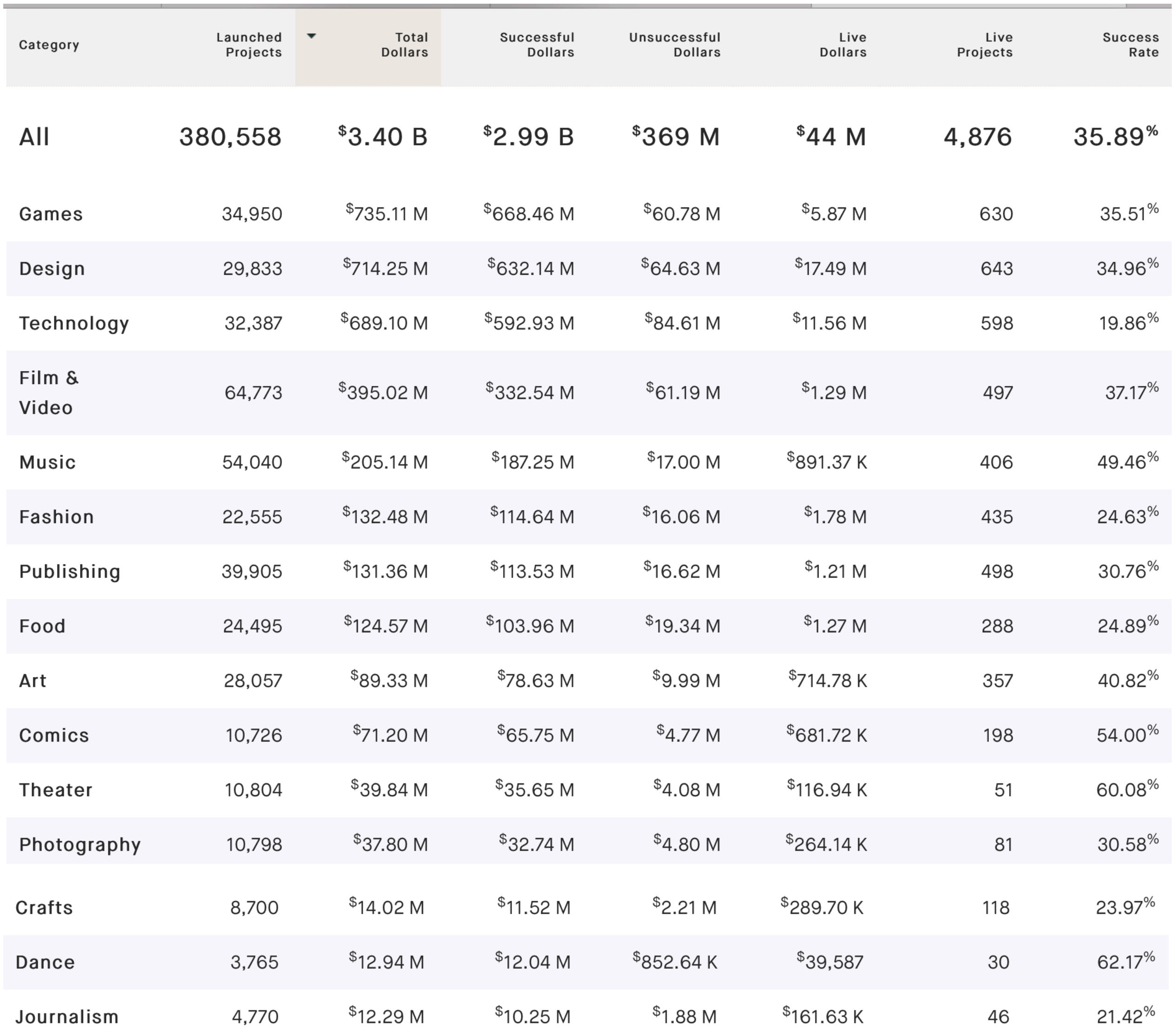
Prior researchers have studied similar topics and provided insightful findings to the topic. There is even a website (http://sidekick.epfl.ch) that shows real-time prediction of the success of Kickstarter campaigns. In this section, I will conduct literature review of two academic papers "Predicting the Success of Kickstarter Campaigns" (Hussain, Kamel & Radhakrishna) and "Launch Hard or Go Home" (Etter, Grossglauser & Thiran) in order to gain more insights for my own analysis.

In "Predicting the Success of Kickstarter Campaigns", the authors used a mixture of simple and complex features to perform their predictive task. Simple features include: time features, location features, category features, text features, goal, staff picked and number of reward levels. The complex features include: previous projects by creator, average money per reward level, preparedness and distance to mean blurb length. After identifying the features, the authors used k-nearest neighbors, logistic regression, support vector classification and random forest to carry out the model. K-nearest Neighbors produced the result that only 4 of the simple features have predictive value: the goal amount, length of the project description, the number of reward levels and the average dollar amount per reward level. In addition, Random Forest is proved to be the best classifier yielding the accuracy of 80.37%. It authors proposed a better model to predict the success and failure of Kickstarter campaigns.

In “Launch Hard or Go Home”, the authors studied the prediction of success on two features: the time-series of money pledged and social attributes. They showed that predictors that use time-series features reach a high prediction accuracy of 85%. The social predictors reach a lower accuracy but the two features combined yield high prediction accuracy. The two papers are highly significant and laid the foundation for future researches on similar topics.

4. Analysis

4.1 Analysis on Kickstarter project statistics



Category	Launched Projects	Total Dollars	Successful Dollars	Unsuccessful Dollars	Live Dollars	Live Projects	Success Rate
All	380,558	$3.40 B	$2.99 B	$369 M	$44 M	4,876	35.89%
Games	34,950	$735.11 M	$668.46 M	$60.78 M	$5.87 M	630	35.51%
Design	29,833	$714.25 M	$632.14 M	$64.63 M	$17.49 M	643	34.96%
Technology	32,387	$689.10 M	$592.93 M	$84.61 M	$11.56 M	598	19.86%
Film & Video	64,773	$395.02 M	$332.54 M	$61.19 M	$1.29 M	497	37.17%
Music	54,040	$205.14 M	$187.25 M	$17.00 M	$891.37 K	406	49.46%
Fashion	22,555	$132.48 M	$114.64 M	$16.06 M	$1.78 M	435	24.63%
Publishing	39,905	$131.36 M	$113.53 M	$16.62 M	$1.21 M	498	30.76%
Food	24,495	$124.57 M	$103.96 M	$19.34 M	$1.27 M	288	24.89%
Art	28,057	$89.33 M	$78.63 M	$9.99 M	$714.78 K	357	40.82%
Comics	10,726	$71.20 M	$65.75 M	$4.77 M	$681.72 K	198	54.00%
Theater	10,804	$39.84 M	$35.65 M	$4.08 M	$116.94 K	51	60.08%
Photography	10,798	$37.80 M	$32.74 M	$4.80 M	$264.14 K	81	30.58%
Crafts	8,700	$14.02 M	$11.52 M	$2.21 M	$289.70 K	118	23.97%
Dance	3,765	$12.94 M	$12.04 M	$852.64 K	$39,587	30	62.17%
Journalism	4,770	$12.29 M	$10.25 M	$1.88 M	$161.63 K	46	21.42%

From the data on the Kickstarter website, we observe that across 15 categories, the 5 categories that have the highest success rates are: Dance, Theater, Comics, Music and Art. The 5 categories that have the lowest success rates are: Technology, Journalism, Crafts, Fashion and Publishing. It is intuitive that success rate varies across different categories due to the unique nature of each category.

However, the data itself is not convincing enough because the number of launch projects differs significantly across categories. “Dance” only has 3,765 launch projects whereas “Film & Video” has 64,773 projects. The small project size may not be representative of the whole picture and thus we need to conduct more in-depth analysis.

4.2 Linear Regression on key variables

Next, I will conduct linear regression to explore the significance of key variables. In my regression, the dependent variable is natural log of total pledges: *ln(totalPledge)* and I will add important independent variables one at a time to explore the best fitted model. Intuitively, I hypothesize that creator’s initial goal, number of comments under the projects, number of Facebook friends the creator has, creator’s past success rate, number of competitors, number of perks the creator provides for backers, total word count in the blurb, number of images in the descriptions and number of updates are important factors for success.

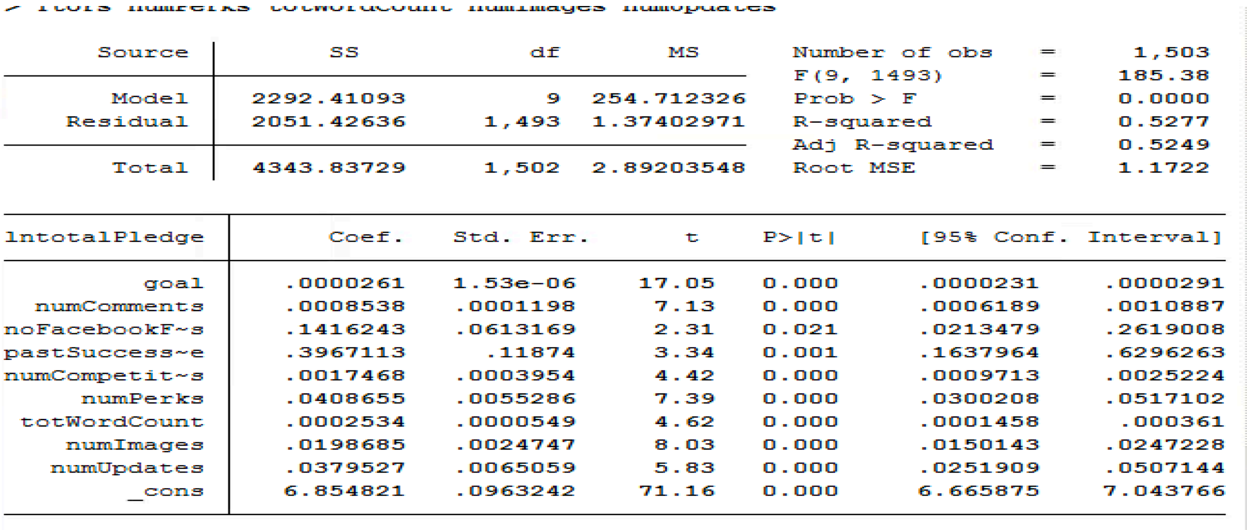
Here attached a regression table with different independent variables. In the regression output, adjusted R-squared measures how well our independent variables jointly explain the variation in dependent variable. We notice that the adjusted R-squared has a big jump (0.3938 to 0.4538) from regression (5) to (6). It shows that past success rate is an important independent variable in our model.

regression	(1)	(2)	(3)	(4)	(5)	(6)	(7)
dependent	lnpledge	lnpledge	lnpledge	lnpledge	lnpledge	lnpledge	lnpledge
intercept	8.2597 (0.0403) p=0.000	8.2309 (0.0389) p=0.000	8.1947 (0.0497) p=0.000	7.8686 (0.0927) p=0.000	7.6155 (0.9492) p=0.000	7.0257 (0.1011) p=0.000	6.8416 (0.0994) p=0.000
goal	0.000043 (1.66e-06) p=0.000	0.00004 (1.63e-06) p=0.000	0.00004 (1.64e-06) p=0.000	0.00004 (1.63e-06) p=0.000	0.00004 (1.60e-06) p=0.000	0.00003 (1.58e-06) p=0.000	0.00003 (1.56e-06) p=0.000
comments	--	0.001426 (0.0001) p=0.000	0.00142 (0.0001) p=0.000	0.00139 (0.0001) p=0.000	0.00129 (0.0001) p=0.000	0.00134 (0.0001) p=0.000	0.00124 (0.0001) p=0.000
friends	--	--	0.08287 (0.0709) p=0.243	0.0906 (0.0706) p=0.199	0.0869 (0.0689) p=0.207	0.1301 (0.0655) p=0.047	0.1320 (0.0633) p=0.037
pastsuccess	--	--	--	0.5689 (0.1367) p=0.000	0.5329 (0.1335) p=0.000	0.4205 (0.1270) p=0.001	0.3509 (0.1230) p=0.004
competitors	--	--	--	--	0.0037 (0.0004) p=0.000	0.0037 (0.0004) p=0.000	0.0031 (0.0004) p=0.000
perks	--	--	--	--	--	0.0709 (0.0055) p=0.000	0.0558 (0.0055) p=0.000
wordcount	--	--	--	--	--	--	0.0005 (0.0001) p=0.000
images	--	--	--	--	--	--	--
updates	--	--	--	--	--	--	--

$\bar{R}^2$	0.3057	0.3560	0.3562	0.3631	0.3938	0.4538	0.4890

regression	(8)	(9)
dependent	lnpledge	lnpledge
intercept	6.8957 (0.0971) p=0.000	6.8548 (0.0963) p=0.000
goal	0.00003 (1.55e-06) p=0.000	0.00003 (1.53e-06) p=0.000
comments	0.0010 (0.0001) p=0.000	0.0009 (0.0001) p=0.000
Friends	0.1110 (0.6718) p=0.072	0.1416 (0.0613) p=0.021
pastsuccess	0.3920 (0.1200) p=0.001	0.3967 (0.1187) p=0001
competitors	0.0021 (0.0004) p=0.000	0.0017 (0.0004) p=0.000
perks	0.0472 (0.0055) p=0.000	0.0409 (0.0055) p=0.000
wordcount	0.0003 (0.0001) p=0.000	0.00025 (0.00005) p=0.000
images	0.0220	0.01987

	(0.0025) p=0.000	(0.0025) p=0.000
updates	--	0.0379 (0.0065) p=0.000
$\bar{R}^2$	0.5144	0.5249



```
> itors numPerks totWordCount numImages numUpdates

      Source |       SS           df       MS      Number of obs   =     1,503
-------------+----------------------------------   F(9, 1493)      =    185.38
       Model |  2292.41093         9  254.712326   Prob > F        =    0.0000
    Residual |  2051.42636     1,493  1.37402971   R-squared       =    0.5277
-------------+----------------------------------   Adj R-squared   =    0.5249
       Total |  4343.83729     1,502  2.89203548   Root MSE        =    1.1722

------------------------------------------------------------------------------
lntotalPledge|      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        goal |   .0000261   1.53e-06    17.05   0.000     .0000231    .0000291
 numComments |   .0008538   .0001198     7.13   0.000     .0006189    .0010887
noFacebookF~s|   .1416243   .0613169     2.31   0.021     .0213479    .2619008
pastSuccess~e|   .3967113     .11874     3.34   0.001     .1637964    .6296263
numCompetit~s|   .0017468   .0003954     4.42   0.000     .0009713    .0025224
    numPerks |   .0408655   .0055286     7.39   0.000     .0300208    .0517102
totWordCount |   .0002534   .0000549     4.62   0.000     .0001458     .000361
   numImages |   .0198685   .0024747     8.03   0.000     .0150143    .0247228
  numUpdates |   .0379527   .0065059     5.83   0.000     .0251909    .0507144
       _cons |   6.854821   .0963242    71.16   0.000     6.665875    7.043766
------------------------------------------------------------------------------
```

Analyzing the regression output, regression (9) gives us the best fit with the adjusted R-squared of 0.5249. In regression (9), the dependent variable is *ln(totalPledge)* and the independent variables are: *goal, numComments, noFacebookFriends, pastSuccessRate, numCompetitors, numPerks, totWordCount, numImages and numUpdates.*

Thus, my initial model can be written as:

$$\ln(totalPledge) = \beta_0 + \beta_1 goal + \beta_2 comments + \beta_3 friends + \beta_4 pastsuccess + \beta_5 competitors + \beta_6 perks + \beta_7 totalword + \beta_8 images + \beta_9 update$$

Plugging in the values from the regression output, my model is:

$$\begin{aligned}\ln(totalPledge) &= \beta_0 + 0.0000261 goal + 0.0008538 comments + 0.1416243 friends \\ &+ 0.3967113 pastsuccess + 0.0017468 competitors + 0.0408655 perks \\ &+ 0.0002534 totalword + 0.0198685 images + 0.0379527 update\end{aligned}$$

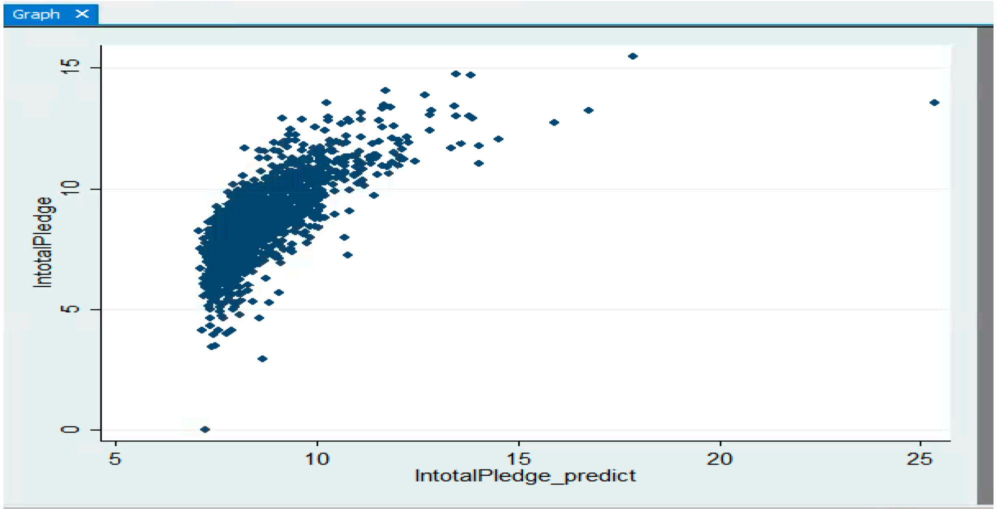
I use log-lin model for the regression in order to correct skewed data and provide more intuitive interpretations of the coefficients. From the regression result, we can observe that $\beta_1, \beta_2, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8$ *and* β_9 are significant because their respective p-value is less than 0.01 and it means that they are significant at 1% significance level. However, we should note that β_3 has a p-value of 0.021, which is not significant at 1% significance level but is significant at 5% significance level.

Firstly, looking at the regression result, we can observe that adjusted R-squared is 0.5249. It shows that 52.49% of change in ln(totalpledge) can be explained by the independent variables in the model. 52.49% is a decent fit to begin with and the result validates my choice of independent variables.

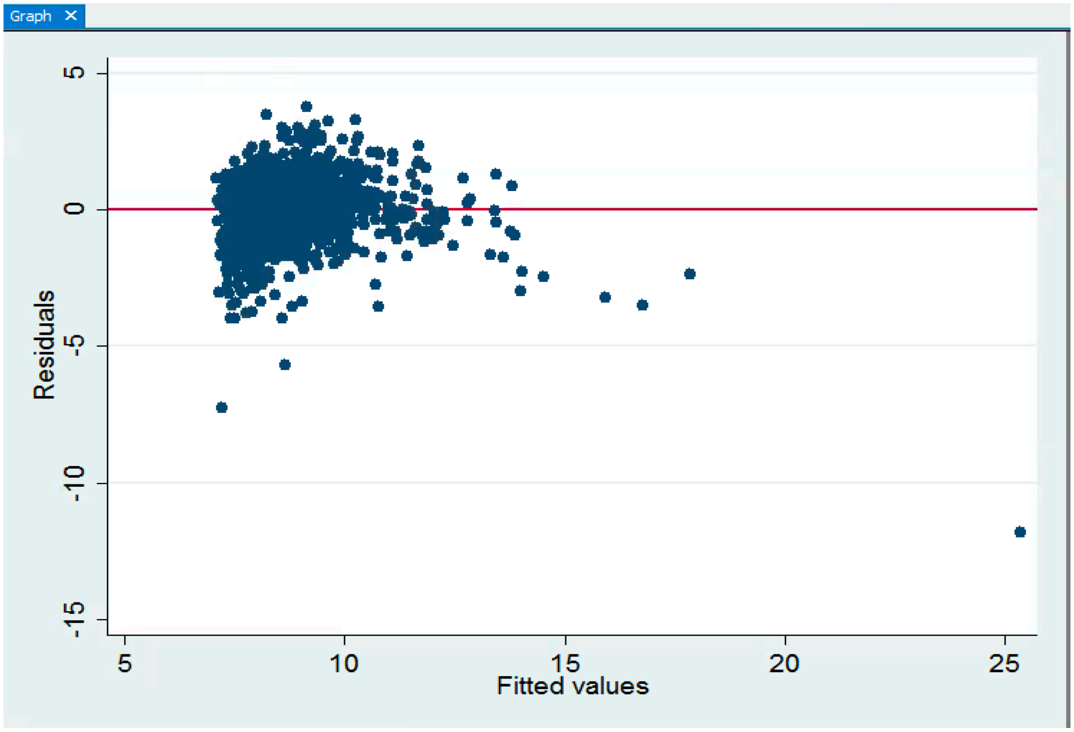
Secondly, from the regression result: *numbers of Facebook friends*, *numbers of perks*, *numbers of updates*, *numbers of images* and *past success rate* influence ln(totalPledge) the most and are thus the most significant independent variables in regression (9).

To test whether this model is a good fit, I will perform two tests: observed vs. predicted values test and homoscedasticity test.

To test for observed vs. predicted value, I plot ln(totalPledge) against ln(totalPledge)_predict in order to explore the linear relationship. We should expect a 45-degree pattern in the data. From the plot, if we ignore outliers, the linear relationship is relatively close to 45-degree. It shows that our model is doing a good job in predicting *lntotalPledge*.



To test for homoscedasticity, I use residual vs fitted value plot to study if there are any patterns of the residuals plotted against fitted values. If there are no patterns, it indicates that the model does not violate the Multiple Linear Regression assumptions and our model is well-fitted. Ignoring the outliers, we cannot observe any patterns of residuals vs fitted values and the errors appear to be homoscedastic. Thus, we can conclude that our linear model is a good fit.



I understand that there might be omitted variables in the model. However, given the model is a good fit proved by two tests above and our independent variables can explain up to 52.49% of the change in ln(totalPledge), I will use this model to carry out my analyses. One potential improvement of this paper is to identify and study other independent variables that are significant in explaining ln(totalPledge).

4.3 Hypotheses Formulation

Given the result of my regression, I will modify my hypotheses. Before conducting any research, my original hypotheses were: Among all factors listed, *the creator's prior success, goal, numbers of perks, number of Facebook friends, total word count in the description* are more significant than other factors. My modified hypotheses are: *numbers of Facebook friends*, *numbers of perks*, *numbers of updates*, *numbers of images* and *past success rate* are the most important factors for a campaign's success.

For the numbers of Facebook friends, I hypothesize that there is a positive relationship. The more Facebook friends the creator has, the more support he is likely to receive. When the creator launches a project, their Facebook friends are more likely to support them initially compared to strangers. In addition, in the current era, social media is more powerful than ever and Facebook is an ideal platform to connect with supporters and garner support for creative projects.

For numbers of perks, I hypothesize that perks are attractive to backers. Since backers cannot claim any equity from the project, perks are the only rewards they can receive from creators. The perks in Kickstarter vary in forms and monetary values and serve as incentives for backers.

The relationship between the number of updates and the success of campaigns is not obvious at the beginning. One may argue that each update conveys more information of the project and provides more valuable information to help backers make their decisions.

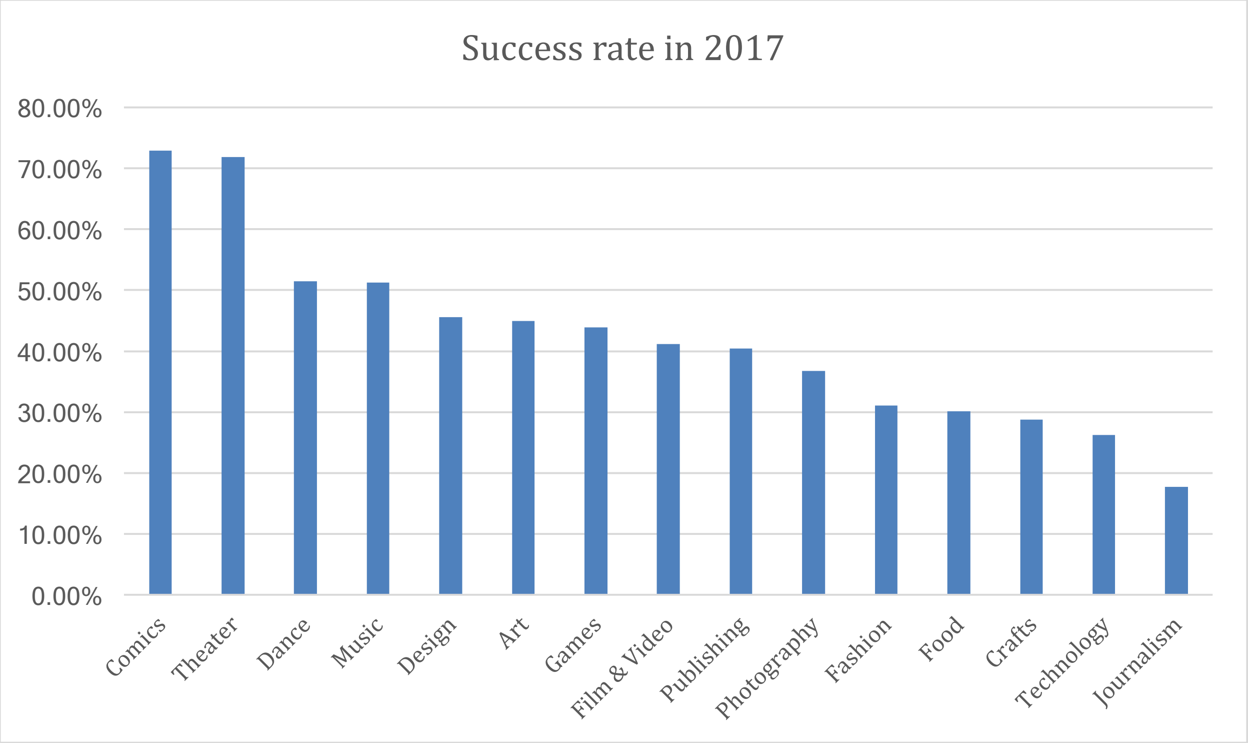
However, it is difficult to quantify this relationship and more research needs to be conducted to further explore this relationship.

The number of images increases the probability of success of Kickstarter campaigns. Backers are more attracted by visual images and the number of images helps them understand the projects better. Thus, it is important for creators to use images to communicate with the Kickstarter community and attract backers.

Creator's past success rate is an important indicator of the success of their current projects. Past success rate indicates the creator's experience and connections to the Kickstarter community. Creators are likely to create multiple projects in the same category and if they have connected with backers for their previous projects, it is likely that the backers will continue to support their current endeavors. However, in the dataset, it is difficult to distinguish new creators and creators with all prior failures since the prior success rates are both 0.

4.4 Data Visualization and Summary Statistics

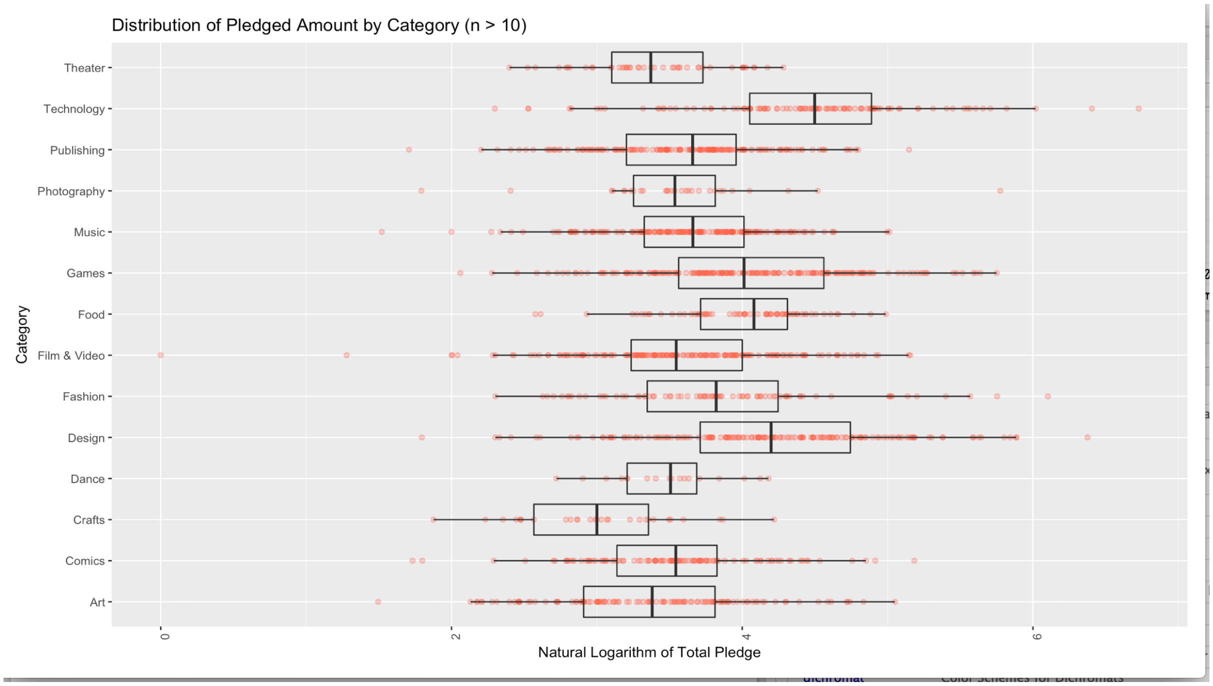
My dataset contains 3652 Kickstarter projects in 2017 only. Of the total 3652 projects, 1503 projects met their funding goals and are considered "successful". For this paper, I will analyze the attributes of 1503 successful projects. As mentioned above, Kickstarter has a unique "All-or-Nothing" model and even projects with huge amounts of total pledges will fail if they do not meet their funding goals. Another potential improvement of this paper is to analyze the attributes of "failed" projects such as funding goals. First, I will visualize the success rate in each of the 15 categories in 2017:



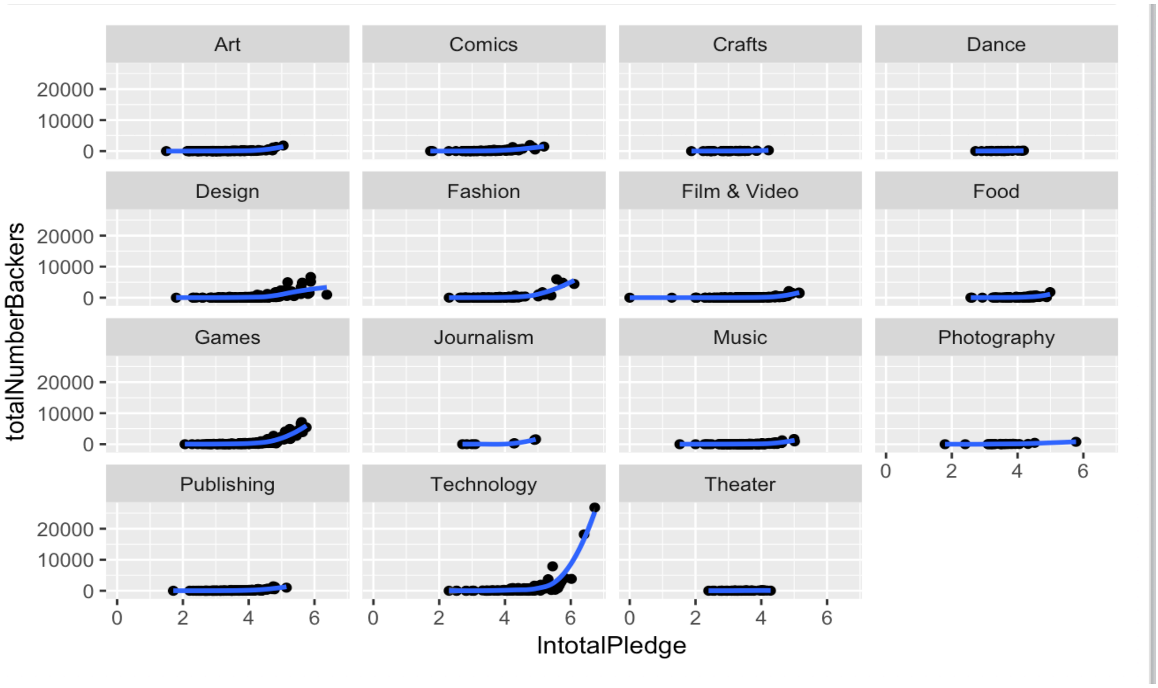
For projects in 2017, the three categories that have the highest success rates are: Comics (72.92%), Theater (71.88%) and Dance (51.43%). The three categories that have the lowest success rates are: Crafts (28.71%), Technology (26.25%) and Journalism (17.78%). The result is consistent with the all-time data. Since the launch of Kickstarter, the 5 categories that have the highest success rates are: Dance, Theater, Comics, Music and Art. Originally, I expected that the success rate by category varies over year due to the shift of trend. The result is surprising and shows that the success rate by category hardly changes over time.

In Kickstarter, a project is considered "successful" once it meets its initial funding goal. However, success rate on Kickstarter alone is not indicative of the project's success in general. There are many nuances behind the success rate. After observing the success rate by category in 2017, the next step is to study the relationship between pledge amount and category. It is natural to think that total pledge amount is related to the project size, which differs by category. Now we use data to further explore this relationship.

Since there are only 8 successful projects in the category "Journalism", I omit that category from the analysis. As we can see from the graph above, across 14 categories, Technology, Design, Food, Games and Fashion have the highest mean natural log of total pledges. Crafts, Theater, Art, Dance and Comics have the lowest mean natural log of total pledges. It verifies our previous assumption that pledge amount is related to project size, which differs across categories. Projects in Technology, Games and Food might require more Research and Development costs and more capital investments. Thus, creators require more funding in order to complete the projects in these categories. In addition, due to the large size of the projects in these categories, the creators are more likely to raise high amounts of pledges.

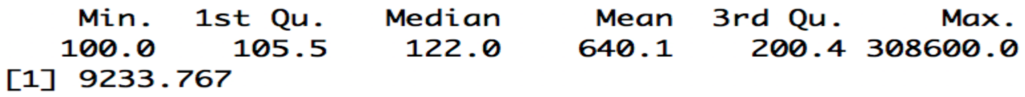


Naturally, we think that total amount of pledges is related to the number of backers. However, backers with large amounts of pledges might affect the accuracy of our analysis. Now, I will explore the relationship between natural log of total pledges and the total number of backers by category to study which categories are more prone to outliers. I expect the larger the total number of backers, the more pledges there are.



From the result above, we can conclude that: Technology, Games, Design and Fashion are more prone to outliers than other categories. It means that these four categories are more likely to attract backers with huge amounts of pledges. Previously, we found out that Technology, Games, Design and Fashion are four out of the five categories that receive the most pledges in 2017. We can thus conclude that they are indeed the most popular categories in 2017. In addition, the categories with more outliers correspond to the categories that receive higher average amount of pledges. We can conclude that the larger the size of the project, the more likely the creator will receive large amounts of pledges.

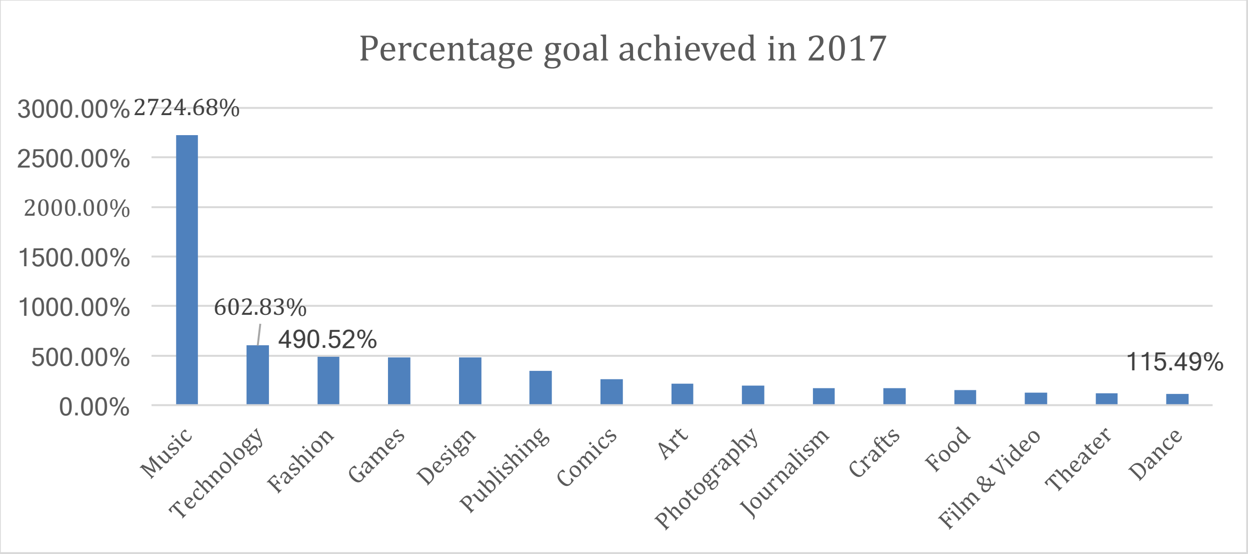
In Kickstarter's "All-or-Nothing" model, creators need to receive the full amount of their goals within a certain time period, otherwise, they will receive nothing. While nearly half of the projects fail, there are still a large number projects that are significantly overfunded. Now we study the average percentage of overfunded by category. In our analysis, overfunding status is calculated by the total pledge divided by creator's goal and it gives us a percentage amount.



```
    Min.  1st Qu.   Median     Mean  3rd Qu.      Max.
   100.0    105.5    122.0    640.1    200.4  308600.0
[1] 9233.767
```

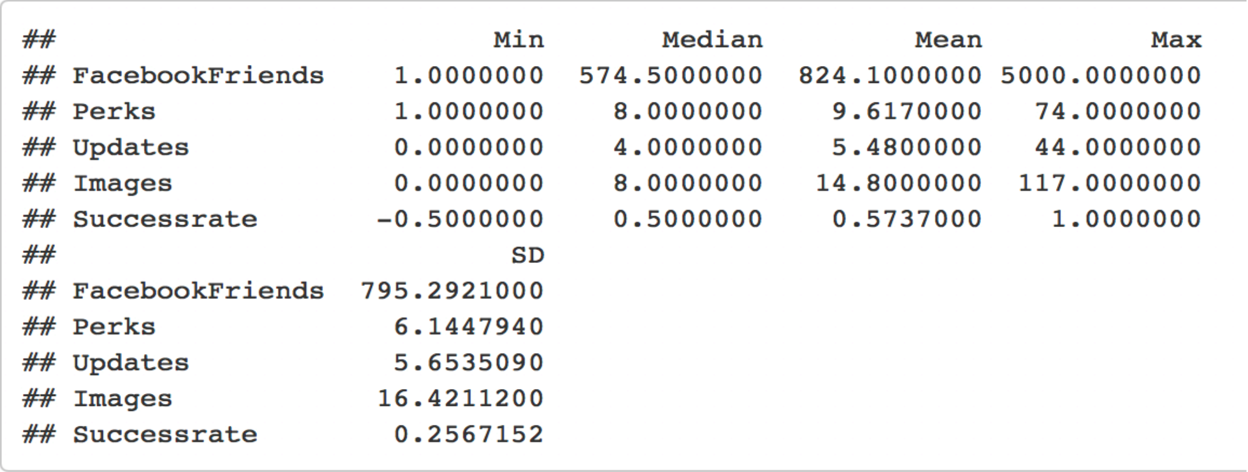
When we look at the summary statistics of overfunded projects, we find that the mean is 640.1. It indicates that the mean amount of total pledges for successful projects is 6 times the original goal. However, the standard deviation is extremely large at 9233.767 and it shows that there are a lot of variations in the overfunding projects.

From the calculation of mean percentage goal, we find that Music has the highest percentage goal of 2724.68. It shows that on average, projects in the Music category can achieve over 27 times of their funding goals. Dance has the lowest percentage goal of 115.4907 and it shows that successful projects in the Dance category barely meet their funding goals on average. The ranking of all categories according to their overfunding status is: Music, Technology, Fashion, Games, Design, Publishing, Comics, Art, Photography, Journalism, Crafts, Food, Film & Video, Theater and Dance. The ranking provides creators an indication on how well their projects will perform based on their categories according to the data in 2017.



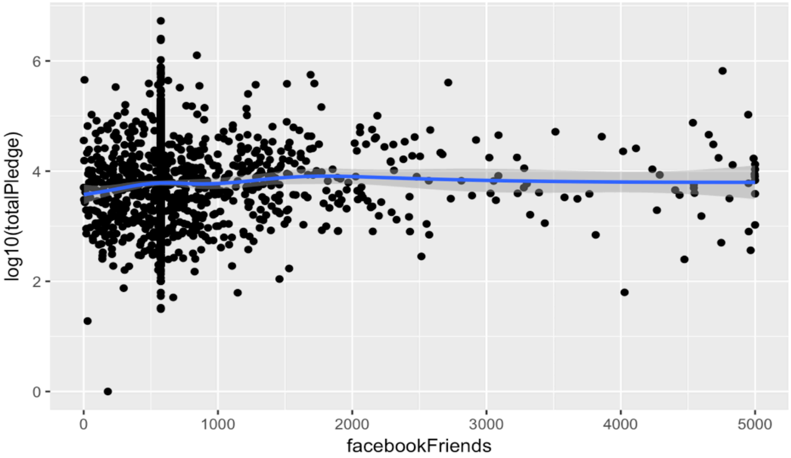
The data only cover projects in 2017 and do not represent the overall overfunding status of projects since the inception of Kickstarter. However, I do not expect a big difference as we found out previously that the success rate across categories is consistent throughout the years.

After exploring different attributes of successful projects by category, I will now use a table to summarize the statistics of the 5 independent variables in my hypotheses:



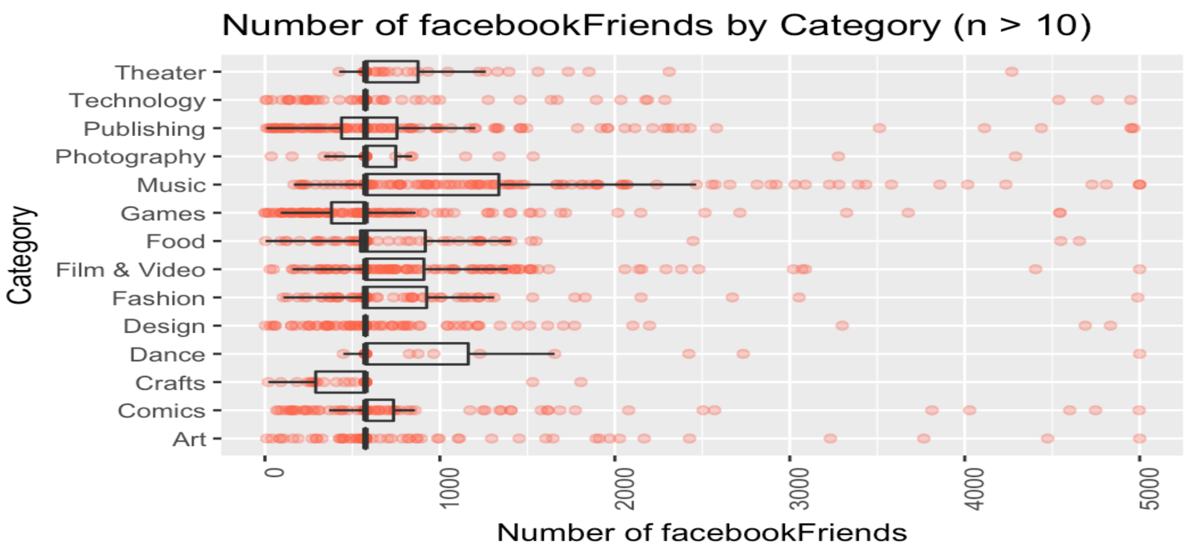
```
##                          Min        Median         Mean           Max
## FacebookFriends    1.0000000   574.5000000  824.1000000  5000.0000000
## Perks              1.0000000     8.0000000    9.6170000    74.0000000
## Updates            0.0000000     4.0000000    5.4800000    44.0000000
## Images             0.0000000     8.0000000   14.8000000   117.0000000
## Successrate       -0.5000000     0.5000000    0.5737000     1.0000000
##                           SD
## FacebookFriends  795.2921000
## Perks              6.1447940
## Updates            5.6535090
## Images            16.4211200
## Successrate        0.2567152
```

From the summary, we can observe the min, median, mean, max and standard deviation of each independent variable. The number of Facebook friends has the widest range (1 to 5000) and largest standard deviation of 795. The success rate has the lowest range (0 to 1) and smallest standard deviation of 0.26. These independent variables are proven by the linear model identified before and they collectively explain up to 52.49% of variations in predicted natural log of pledge amounts.



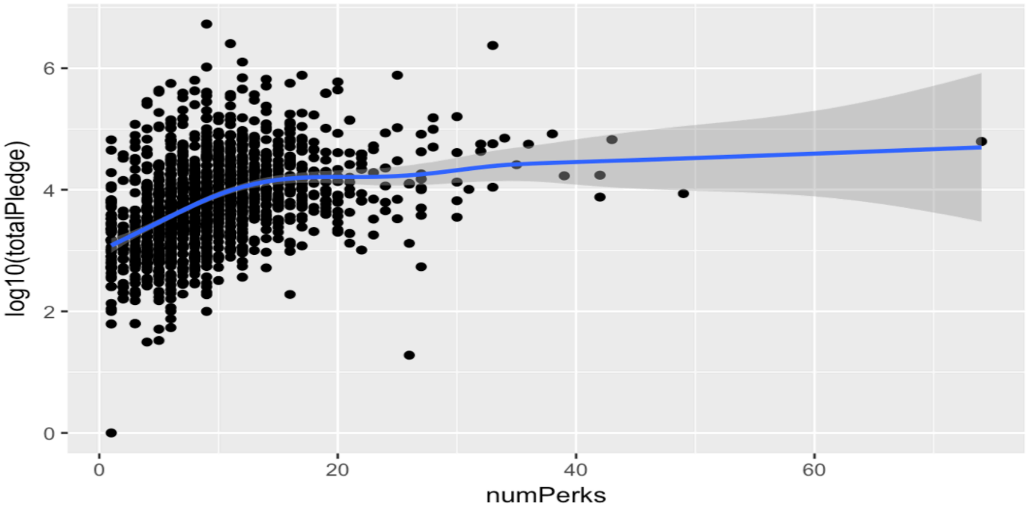
The number of Facebook friends is a crucial factor for the success of projects since the more Facebook friends a creator has, the more connections he has with the community. However, the number of Facebook friends is also the independent variable that has the largest standard error. Some people may argue that the more Facebook friends a creator has, the better. I would argue that although a creator should have a certain number of Facebook friends, the quality of their Facebook friends is also important. From the plot above, we observe that the relationship between the number of Facebook friends and the natural log of total pledge amount is extremely weak. In the age of social media, people tend to befriend with strangers on social media platforms. In addition, some people may even purchase "friends" and "followers" on Facebook or Twitter to appear to be more popular than they actually are. Although it may seem that the creator has a wide connection on social media with a large number of followers, the truth is that the "fake friends" will not have much influence on the success of projects. In general, while we recognize the significance of the number of Facebook friends on the success of projects, we should not ignore the nuances behind the number of Facebook friends.

To study whether the number of Facebook friends varies by category, I generate the following graph:

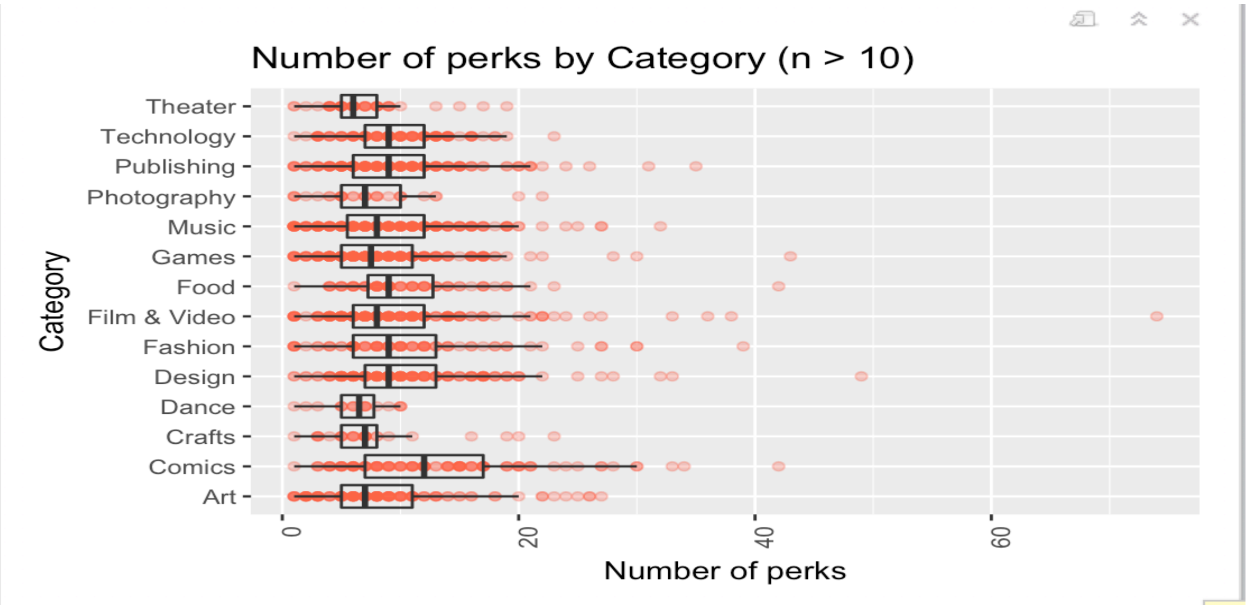


As we can observe from the plot above, there does not seem to be a strong relationship of the average number of Facebook friends across different categories. However, Music and Dance categories stand out since creators have more Facebook friends on average compared to other categories. There could be various reasons behind this finding. For example, creators in Music and Dance categories may tend to be more artistic and spend more time and energy on social media. However, the reason is not valid enough without further research and I will exclude this detailed part of analysis from my paper.

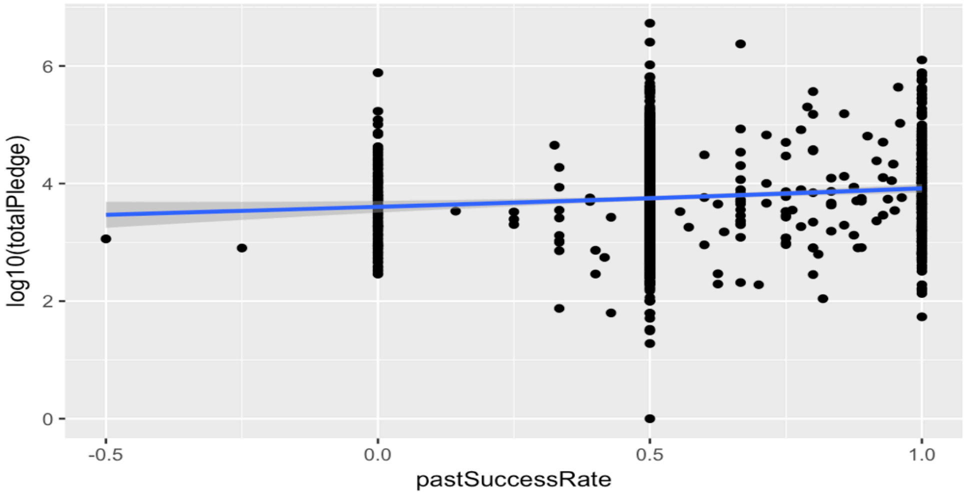
As for the numbers of perks, on average creators provide backers with around 9 perks and the range is from 1 to 74. For the project, I am using the number of the perks in my analysis instead of the values of the perks. The numbers of perks provide an incentive for backers when they consider about backing the projects since they cannot claim any equity from the projects. Here is the visualization of the number of perks and the natural log of total pledges. As we can observe from the plot below, there is a positive relationship between the number of perks and the natural log of total pledge. It shows that perks are attractive to backers and can influence their decisions on Kickstarter.



To study whether the number of perks varies by category, I generate the following graph. From the graph, we can observe that the average number of perks are similar across categories except for Comics. In our previous analysis, we found out that projects in the Comics category have the highest success rate in 2017. The high success rate could be due to the large number of perks creators provide in Comics category. Perks are indeed important for the success of Kickstarter campaigns.

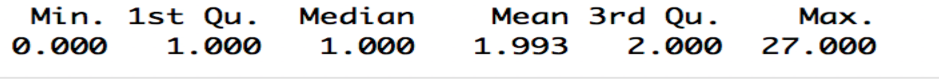


Another important independent variable in my model is past success rate. The mean past success rate is 0.5737 and the max is 1. The data visualization is attached below:



As we can observe from the plot, the majority of projects have a past success rate between 0.5 and 1 and only 2 projects have past success rate below 0. We would expect creators who have a high past success rate to be more experienced in Kickstarter and have more supporters in the Kickstarter community. As a result, they will have a higher probability of success for their new projects. Creators who have a high past success rate may also have more motivations to start new projects. In addition, past success rate can serve as the “resume” for creators and may influence the choice of backers. Backers will likely to have more faith in creators with high past success rates.

As far as the number of successful campaigns, creators who have a successful project in 2017 have 1.993 successful projects on average in the past with the standard deviation of 2.89. It further shows that not all creators with current successful projects have extensive experiences on Kickstarter. However, more experiences definitely help creators in launching their future campaigns.



```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   1.000   1.000   1.993   2.000  27.000 
```

5. Conclusion and Future Work

In my project, I explored key factors that affect the success of Kickstarter campaigns. Due to the declining success rate over the years, I believe that my project is relevant and I hope to uncover the myths behind the success of Kickstarter campaigns.

I identified a linear regression model that explains up to 52.49% of the variations in natural log of total pledge. In my model, the dependent variable is the natural log of total pledges and the independent variables are: *goal, numComments, noFacebookFriends, pasSuccessRate, numCompetitors, numPerks, totWordCount*, *numImages and numUpdates.* Among the nine independent variables, the five variables that have the largest coefficient are: Facebook friends, number of perks, number of updates, number of images and past success rate.

In my graphic analysis, I first find out that the success rate of categories stays constant through time. In addition, projects in Technology, Game and Design have the largest average number of total money raised. Coincidently, projects in Technology, Game and Design have more outliers than other categories. It shows that projects in these categories have larger sizes on average and require more funding. Moreover, for projects in 2017, the top three overfunded categories are: Music, Technology and Fashion.

To study deeply about the number of Facebook friends, I conclude that the average number of Facebook friends stays relatively constant across categories. However, creators tend to have more Facebook friends on average in Music and Dance categories. As for the number of perks, there is a positive relationship between the number of perks and the natural log of total pledge. And the more past success a creator has, the more likely he will succeed in the current project.

I believe that there is still much room for improvements for my project. To begin with, my data only cover projects in 2017. My conclusion will be more accurate and predictive if I can obtain the all-time data of Kickstarter. In addition, there are other characteristics of the projects that might be interesting to explore such as the duration, location etc. Last but not least, I hope to gain more knowledge about machine learning and build a model with more accuracy in predicting the success of Kickstarter campaigns.

This project definitely broadens my horizon on statistical methods and analysis. I would like to thank Professor Aldous for his guidance and support on the project. I hope to keep learning in the field of Statistics and equip myself with more knowledge to conduct more in-depth analysis in the future.

6. Appendix

Stata command:

```
clear all
import excel "\\Client\H$\Desktop\STAT 157\Kickstarter_success.xlsx", sheet("Sheet1")
firstrow
gen lntotalPledge = ln(totalPledge)
reg lntotalPledge goal
reg lntotalPledge goal numComments
reg lntotalPledge goal numComments noFacebookFriends
reg lntotalPledge goal numComments noFacebookFriends pastSuccessRate
reg lntotalPledge goal numComments noFacebookFriends pastSuccessRate
numCompetitors
reg lntotalPledge goal numComments noFacebookFriends pastSuccessRate
numCompetitors numPerks
reg lntotalPledge goal numComments noFacebookFriends pastSuccessRate
numCompetitors numPerks totWordCount
reg lntotalPledge goal numComments noFacebookFriends pastSuccessRate
numCompetitors numPerks totWordCount numImages
reg lntotalPledge goal numComments noFacebookFriends pastSuccessRate
numCompetitors numPerks totWordCount numImages numUpdates
predict lntotalPledge_predict
label variable lntotalPledge_predict "lntotalPledge_predict"
scatter lntotalPledge lntotalPledge_predict
rvfplot, yline(0)
```

R code:

~~~
```{r}
library(readxl)
Kickstarter_success <- read_excel("~/Desktop/STAT 157/Kickstarter_success.xlsx")
head(Kickstarter_success)
~~~

~~~
```
```{r}
a<-Kickstarter_success[Kickstarter_success$mainCategory=='Art',]
b<-Kickstarter_success[Kickstarter_success$mainCategory=='Food',]
c<-Kickstarter_success[Kickstarter_success$mainCategory=='Music',]
d<-Kickstarter_success[Kickstarter_success$mainCategory=='Games',]
e<-Kickstarter_success[Kickstarter_success$mainCategory=='Publishing',]
f<-Kickstarter_success[Kickstarter_success$mainCategory=='Photography',]
g<-Kickstarter_success[Kickstarter_success$mainCategory=='Design',]
h<-Kickstarter_success[Kickstarter_success$mainCategory=='Comics',]
i<-Kickstarter_success[Kickstarter_success$mainCategory=='Technology',]
j<-Kickstarter_success[Kickstarter_success$mainCategory=='Fashion',]
k<-Kickstarter_success[Kickstarter_success$mainCategory=='Theater',]
l<-Kickstarter_success[Kickstarter_success$mainCategory=='Journalism',]
m<-Kickstarter_success[Kickstarter_success$mainCategory=='Dance',]
n<-Kickstarter_success[Kickstarter_success$mainCategory=='Film & Video',]
o<-Kickstarter_success[Kickstarter_success$mainCategory=='Crafts',]
```
```{r}
A <- c(121/269,70/232, 191/373, 196/447, 170/420, 29/79, 158/347, 105/144, 110/419, 79/254, 45/64, 8/45, 18/35, 174/423, 29/101)
labels <- c("Art","Food","Music","Games","Publishing","Photography","Design","Comics","Technology","Fashion","Theater","Journal","Dance","Film & Video","Crafts")
pie(A, labels=labels,radius=1, cex=0.8)
```
```{r}
summary(Kickstarter_success$facebookFriends)
summary(Kickstarter_success$numPerks)
summary(Kickstarter_success$numUpdates)
summary(Kickstarter_success$numImages)
~~~

~~~
summary(Kickstarter_success$pastSuccessRate)
```
```{r}
sd(Kickstarter_success$facebookFriends)
sd(Kickstarter_success$numPerks)
sd(Kickstarter_success$numUpdates)
sd(Kickstarter_success$numImages)
sd(Kickstarter_success$pastSuccessRate)
```
```{r}
statistics <- matrix(c(1, 574.5, 824.1, 5000, 795.2921,1, 8, 9.617, 74, 6.144794, 0, 4, 5.48,44,
5.653509, 0, 8, 14.8, 117, 16.42112, -0.5, 0.5, 0.5737, 1, 0.2567152), ncol=5, byrow=TRUE)
colnames(statistics) <- c("Min","Median","Mean","Max","SD")
rownames(statistics) <- c("FacebookFriends","Perks","Updates","Images","Successrate")
statistics <-as.table(statistics)
statistics
```
```{r}
library(ggplot2)
p1 <-ggplot(Kickstarter_success, aes(x=facebookFriends))
p1+geom_histogram()+ggtitle("Number of Facebook friends in successful projects")

p2 <-ggplot(Kickstarter_success, aes(x=numPerks))
p2+geom_histogram()+ggtitle("Number of perks in successful projects")
p3 <-ggplot(Kickstarter_success, aes(x=numUpdates))
p3+geom_histogram()+ggtitle("Number of updates in successful projects")
p4 <-ggplot(Kickstarter_success, aes(x=numImages))
p4+geom_histogram()+ggtitle("Number of images in successful projects")
p5 <-ggplot(Kickstarter_success, aes(x=pastSuccessRate))
p5+geom_histogram()+ggtitle("Past success rate in successful projects")
```
~~~

~~~~
```{r}
g6 <-ggplot(Kickstarter_success, aes(x=facebookFriends, y=log10(totalPledge)))
g6+geom_point()+geom_smooth()

g7 <-ggplot(Kickstarter_success, aes(x=numPerks, y=log10(totalPledge)))
g7+geom_point()+geom_smooth()

g8 <-ggplot(Kickstarter_success, aes(x=numUpdates, y=log10(totalPledge)))
g8+geom_point()+geom_smooth()

g9 <-ggplot(Kickstarter_success, aes(x=numImages, y=log10(totalPledge)))
g9+geom_point()+geom_smooth()

g10 <-ggplot(Kickstarter_success, aes(x=pastSuccessRate, y=log10(totalPledge)))
g10+geom_point()+geom_smooth()
```
```{r}
g11 <-ggplot(c, aes(x=facebookFriends, y=log10(totalPledge)))
g11+geom_point()+geom_smooth()
```
```{r}
Kickstarter_success %>%
   group_by(mainCategory) %>%
   filter(n() > 10) %>%
   ggplot() + geom_point(aes(x=mainCategory, y=numPerks), alpha=0.3, color="tomato") +
   geom_boxplot(aes(x=mainCategory, y=numPerks), alpha=0) +
   ggtitle("Number of perks by Category (n > 10)") + xlab("Category") +
   theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
   ylab("Number of perks") + coord_flip()
```
```{r}
~~~~

~~~~
Kickstarter_success %>%
  group_by(mainCategory) %>%
  filter(n() > 10) %>%
  ggplot() + geom_point(aes(x=mainCategory, y=facebookFriends), alpha=0.3, color="tomato") +
  geom_boxplot(aes(x=mainCategory, y=facebookFriends), alpha=0) +
  ggtitle("Number of facebookFriends by Category (n > 10)") + xlab("Category") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  ylab("Number of facebookFriends") + coord_flip()
```
ggplot(Kickstarter_success,aes(totalNumberBackers,lntotalPledge))+geom_point()+geom_smooth()+facet_wrap(~mainCategory)
```{r}
summary(Kickstarter_success$percentageGoal)
sd(Kickstarter_success$percentageGoal)
summary(Kickstarter_success$numSuccessfulCampaigns)
sd(Kickstarter_success$numSuccessfulCampaigns)
```
~~~~

7. References

Chen, K., Jones, B., Kim, I., & Schlamp, B. (n.d.). KickPredict : Predicting Kickstarter Success.

CrowdExpert. (n.d.). Total crowdfunding volume worldwide from 2012 to 2015 (in billion U.S. dollars).

Etter, V., Grossglauser, M., & Thiran, P. (2013). Launch Hard or Go Home! Predicting the Success of Kickstarter Campaigns.

Hussain, N., Kamel, K., & Radhakrishna, A. (n.d.). Predicting the success of Kickstarter campaigns.

Kickstarter. (2017). Kickstarter Stats.