Kickstarter Campaign Analysis

Felix Huang

2025-07-14

Contents

ABSTRACT

Abstract: Kickstarter campaigns are notorious for failing. Certain categories find greater success than others, and certain times of the year have greater amounts of campaigns launched. This analysis looks into which categories are most successful and which months have the greatest amount of campaigns launched and succeeded. Findings showed that film&video and music had the greatest amount of successful campaigns. It was also discovered that there was significance in the amount of campaigns launched across each month, March, October and July having the most amount of campaigns launched, with March and October with the greatest amount of success. March and October success rates were compared and found to be proportionate. Strong association was found when running a linear regression model looking at the money goal to backers count for campaigns within March and October.

INTRODUCTION

Introduction: The data used for this analysis is from a Kickstarter dataset provided by Jonathan Leland via ICPSR. The data was obtained from public use Kickstarter Data. The data is not weighted so it may not accurately represent its target population. Important pieces of information provided by this data set are; the types of campaigns launched, when they were launched, how many individuals backed the project, how much money was asked and pledged and if the campaign was successful or not. The data was explored and analysed and statistical testing is used to gather insight.

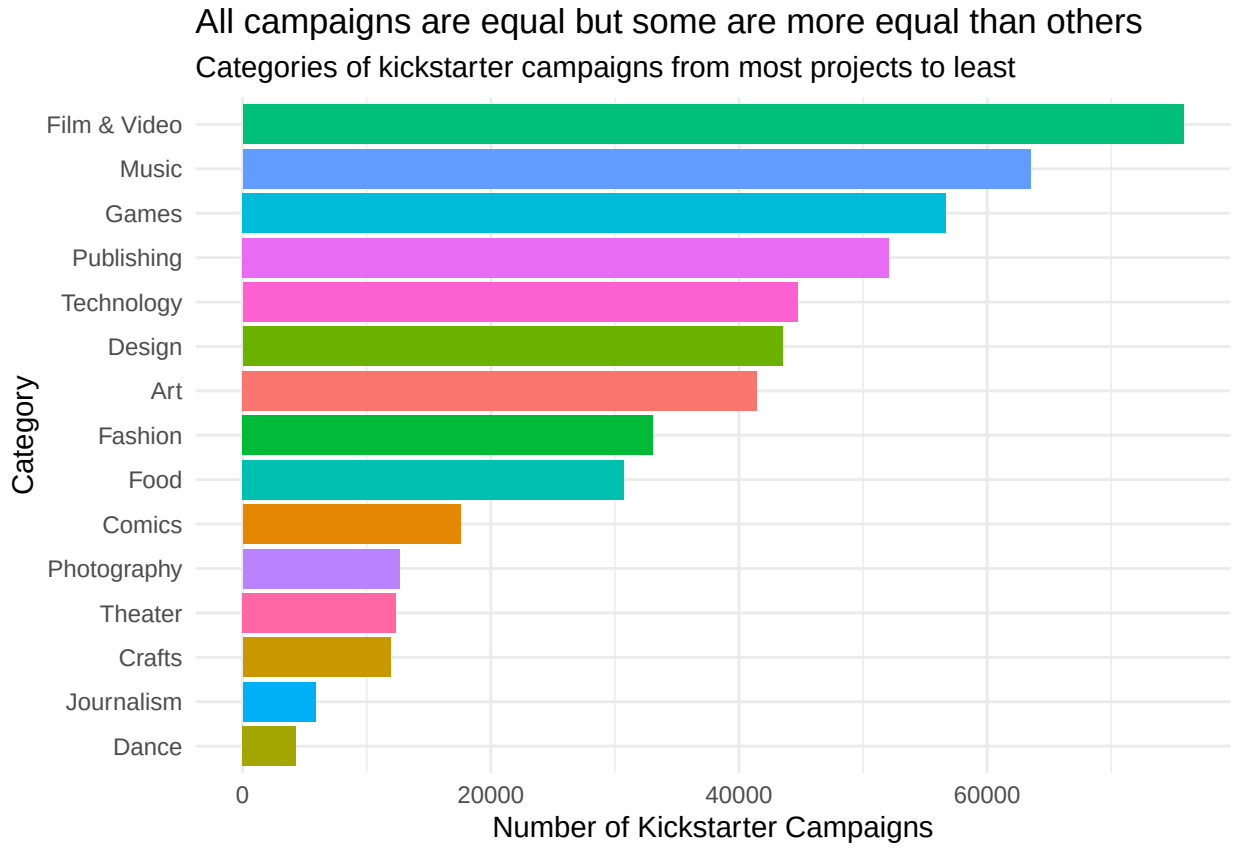
DATA AND METHODOLOGY

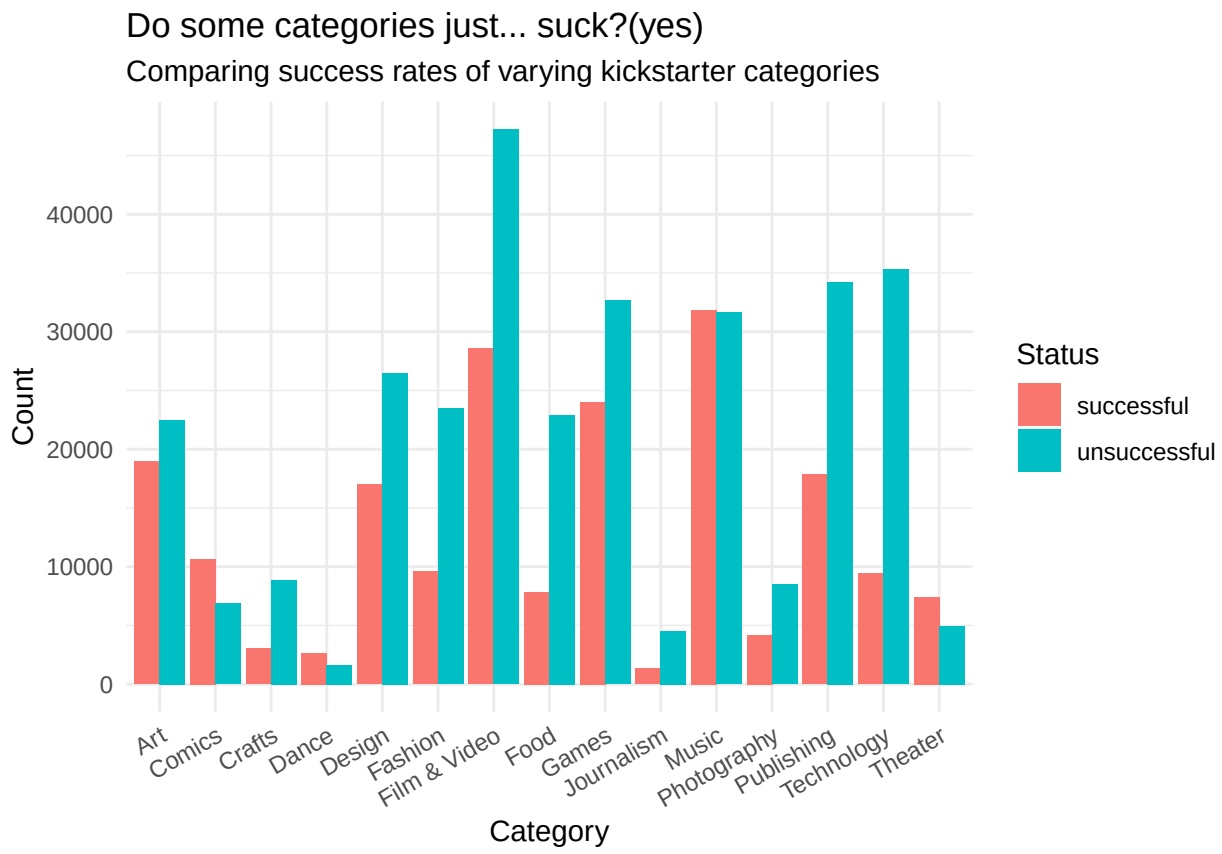
Methodology: Data that were examined were categories with the most campaigns and successess as well as months that had the most campaigns and successes. While analyzing this data there are subgroups such

as “Subcategory” and “Launch Date/Deadline Date” that were not utilized. As the data was explored, there was found to be statistical significance among the amount of successful campaigns across months. The observational units were the months from January to December. Sample size was n=73921. A chi-squared test of homogeneity was ran to find significance in the amount of successful campaigns across each month. There was statistical significance and the top to months with similar successful campaigns were chosen for deeper analysis. Data was omitted and filtered out. Campaigns that asked for more than $50000 were omitted as they did not represent the overall data being extremely far from the median and even the 3rd quantile. A 2-prop z test was then run on the two chosen months, March and October. After that a multiple linear regression model was implemented to see if there was a relationship between number of backers, amount of money asked and amount of money pledged.

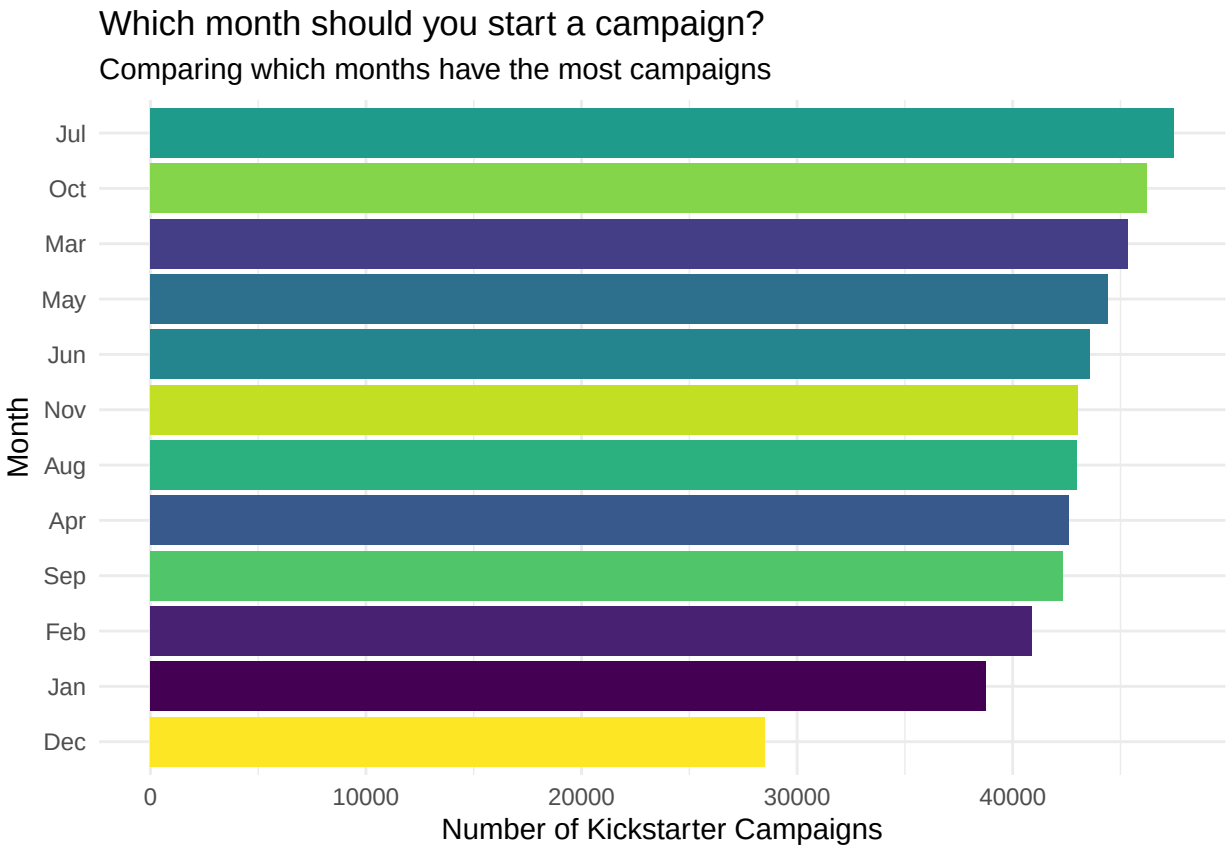
RESULTS

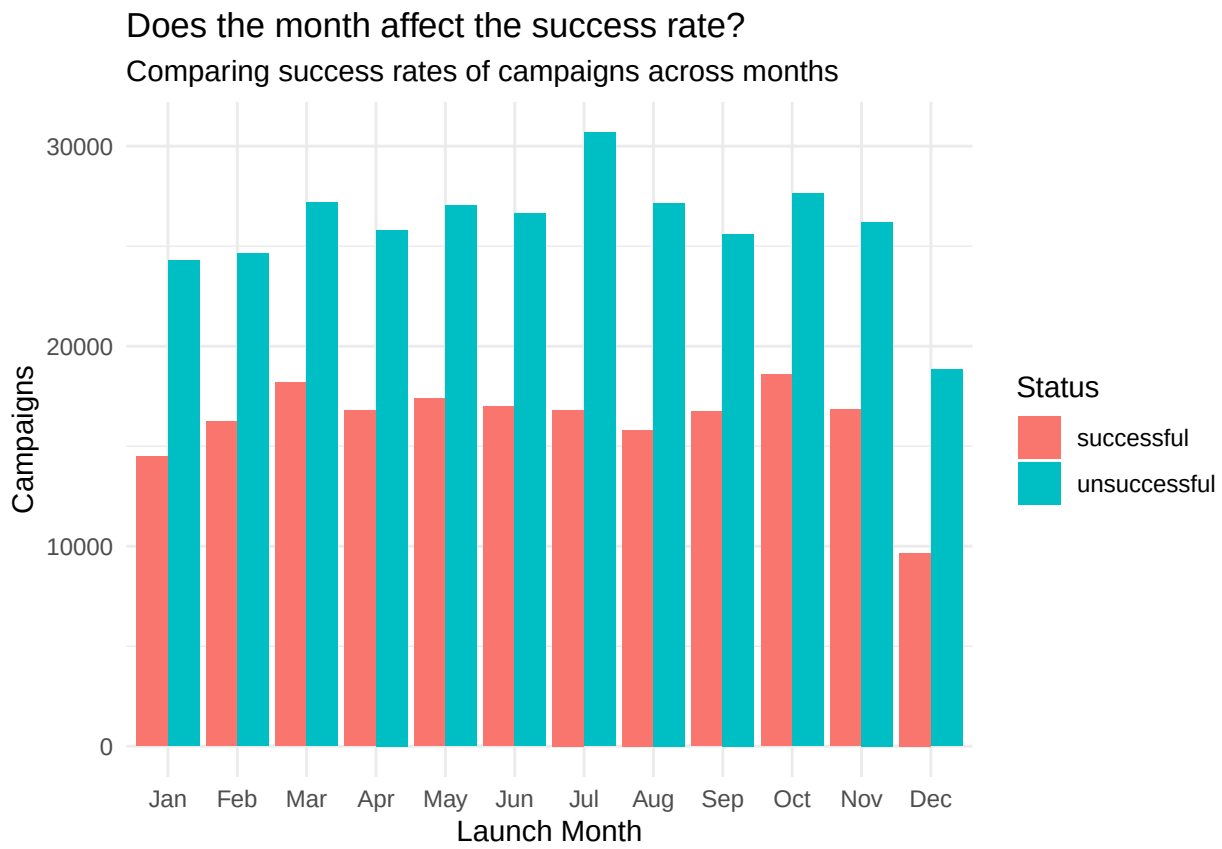
```
## Rows: 506199 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (13): NAME, CATEGORY, SUBCATEGORY, PROJECT_PAGE_LOCATION_NAME, PROJECT_P...
## dbl  (6): CASEID, PID, CATEGORY_ID, SUBCATEGORY_ID, UID, BACKERS_COUNT
## num  (2): GOAL_IN_ORIGINAL_CURRENCY, PLEDGED_IN_ORIGINAL_CURRENCY
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```





Data was explored to answered the questions "which categories has the most amount of campaigns, and which categories were most successful?" It was found that film&video had the largest amount of campaigns but music had the largest amount of success. Music also had the most equal ratio of success to failed campaigns.


Which month should you start a campaign?
Comparing which months have the most campaigns
Jul
Oct
Mar
May
Jun
Nov
Aug
Apr
Sep
Feb
Jan
Dec
Month
0
10000
20000
30000
40000
Number of Kickstarter Campaigns



The next question asked was if the time of the year affected the amount of campaigns that were put out and if there were months that saw greater success. At a glance the number of campaigns were pretty similar and success rates were fairly similar a well.

```
##
##  Chi-squared test for given probabilities
##
## data:  tab
## X-squared = 3635.8, df = 11, p-value < 2.2e-16
```

To answer the question of if there was statistical significance when comparing the amount of campaigns launched for each month, a chi-squared test of homogeneity was performed. The finding produced a p-value of 2,2e-16 which is extremely small. This leads us to reject our null hypothesis that every month had a non-significant identical amount of campaigns.

```
##   Launch_Month           CATEGORY      GOAL_IN_USD    PLEDGED_IN_USD
##  Oct    :41974   Film & Video:11964   Min.   :    1   Min.   :      0
##  Mar    :41347   Music       :11035   1st Qu.: 1632   1st Qu.:     70
##  Jan    :    0   Games       : 9745   Median : 5000   Median :    953
##  Feb    :    0   Publishing  : 9194   Mean   : 8509   Mean   :   7896
##  Apr    :    0   Design      : 7084   3rd Qu.:10979   3rd Qu.:   4984
##  May    :    0   Art         : 7061   Max.   :49999   Max.   :4167720
##  (Other):    0   (Other)     :27238
##            STATE       BACKERS_COUNT
##  successful  :35627   Min.   :  0.00
##  unsuccessful:47694   1st Qu.:  3.00
```

```
##                       Median : 17.00
##                       Mean   : 63.83
##                       3rd Qu.: 66.00
##                       Max.   :999.00
##                       NA's   :1383
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(March_success, October_success) out of c(March_total, October_total)
## X-squared = 0.094519, df = 1, p-value = 0.7585
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.007820493  0.005664719
## sample estimates:
##    prop 1    prop 2
## 0.4270443 0.4281222
```

The top 2 months with the greatest amount of successful campaigns were chosen; March and October. A 2-prop z test was then run to see if there was statistical significance between the success rate of both groups. The prop for both March and October were about equal, and the p-value was a staggering 0.758. This leads us to fail to reject the null hypothesis that both Months had similar proportions of success rates.

```
##
## Call:
## lm(formula = BACKERS_COUNT ~ GOAL_IN_USD + PLEDGED_IN_USD, data = Oct_Mar)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3411.5   -29.1   -23.1    -0.2   852.4
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.160e+01  4.121e-01   76.70  < 2e-16 ***
## GOAL_IN_USD    -1.996e-04  3.343e-05   -5.97 2.38e-09 ***
## PLEDGED_IN_USD  7.048e-03  2.634e-05  267.57  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 88.43 on 81935 degrees of freedom
##   (1383 observations deleted due to missingness)
## Multiple R-squared:  0.4838, Adjusted R-squared:  0.4838
## F-statistic: 3.839e+04 on 2 and 81935 DF,  p-value: < 2.2e-16
```

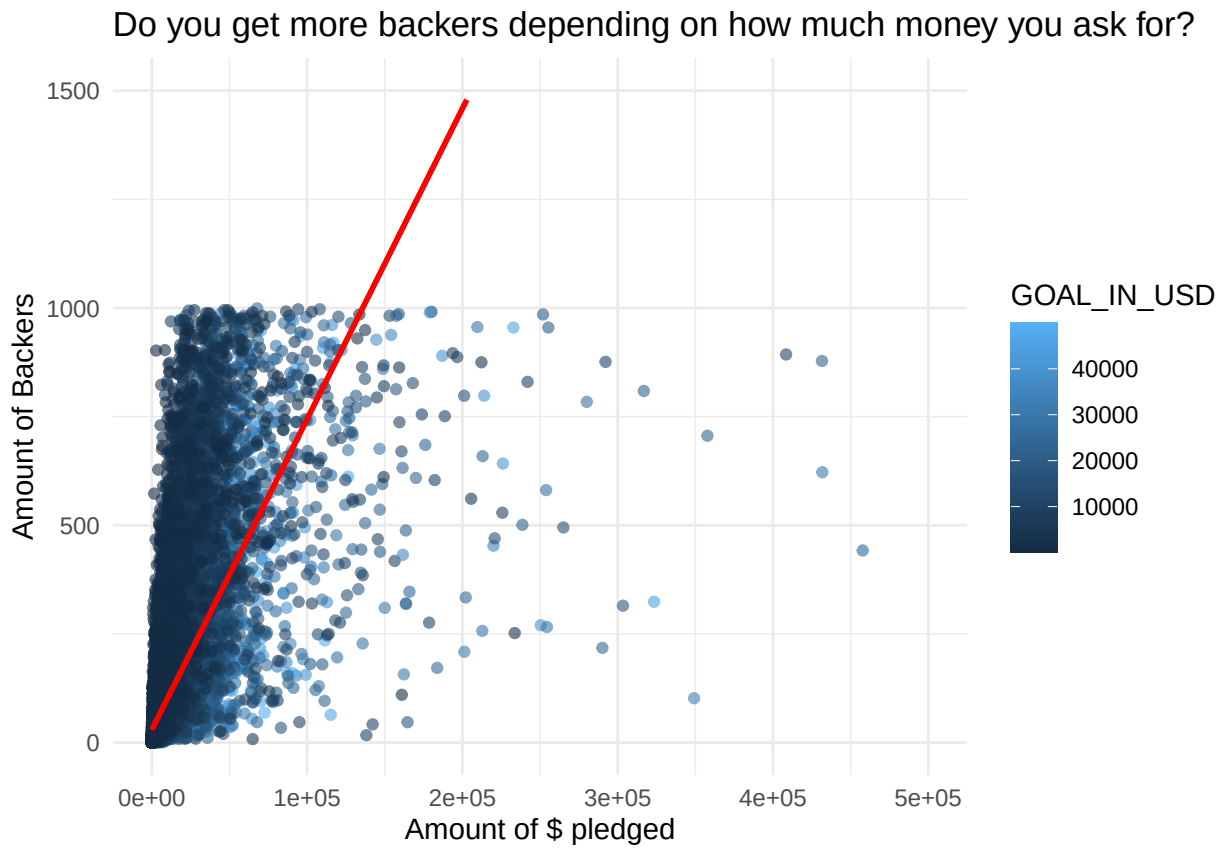
```
##                    Df    Sum Sq   Mean Sq F value Pr(>F)
## GOAL_IN_USD         1  40559713  40559713    5186 <2e-16 ***
## PLEDGED_IN_USD      1 559914737 559914737   71596 <2e-16 ***
## Residuals       81935 640774831      7821
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 1383 observations deleted due to missingness
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 1384 rows containing non-finite outside the scale range
## (‘stat_smooth()‘).

## Warning: Removed 1384 rows containing missing values or values outside the scale range
## (‘geom_point()‘).

## Warning: Removed 44 rows containing missing values or values outside the scale range
## (‘geom_smooth()‘).
```



A multiple linear regression model was run to see if there was an association between the amount of money a campaign asked for and amount of money pledged to the amount of backers. The P value is extremely low in both situations, sitting at <2e-16 and 2.38e-09. This leads us to believe that there is a strong association between backers and money asked as well as money pledged. The linear regression model is not particularly great for this, as the R^2 value is at 0.483. Over 50% of the variance is not explained with this model. There were quite a few outliers that skewed the data. It is also important to note that the date set caps amount of backers to 999. This is not realistic and does not represent the true population.

CONCLUSION

In conclusion, it was found that film&video had the most campaigns launched however music had the largest amount of successful campaigns. Findings also showed that there was a statistically significant difference in the amount of campaigns launched across each month. It was found that March and October had the largest amount of campaigns that succeeded and that there was no statistical significance when comparing the proportions of success rates between the two. A multiple linear regression model then showed that were was strong association between number of backers, money asked for a campaign and money pledged for a

campaign. The linear regression model is not accurate with an r^2 value of .475. For future analysis, the relationship between success and the proportion of amount pledged to amount asked for would be looked at. The different categories and their relationships with success rates would also be analyzed.

APPENDIX

REFERENCES