Building a Successful KickStarter Campaign

Rachel Nelson

8/1/2020

#Section 1

Introduction:

How can I make my KickStarter campaign a success?

Research questions

- Are there certain types/category of campaigns that are more successful?
- How much money should you ask for?
- Is there a time period for the campaign that works better than others?
- What is the average contribution of a backer?
- Is there a better time of year to launch a campaign?

Approach

I will be performing basic data analysis and correlation on the data set provided. I will review things like the mean, median and mode of some of the factors that are of interest.

How your approach addresses (fully or partially) the problem.

By finding out which metrics matter, we can use these elements to ensure your next kickstarter campaign ends in success.

Data

https://www.kaggle.com/kemical/kickstarter-projects

Required Packages

- dplyr
- ggplot2
- plotly
- lm.beta

Plots and Table Needs

* Scatter plots * data tables * correlation tables * box plots

Questions for future steps

- Should I look into neural networks?

Section 2

How to import and clean my data

I am importing the data by connecting the the CSV that was available for download on the Kaggle site. https://www.kaggle.com/kemical/kickstarter-projects

```
# load the data
ks_df <- read.csv("D:/College/DSC520/dsc520/data/ks-projects-201801.csv")
```

I am cleaning the data set to prepare it for analysis. ####Check for missing columns

```
# Check for Missing Columns
names(ks_df)

##  [1] "ID"               "name"             "category"         "main_catego
ry"
##  [5] "currency"         "deadline"         "goal"             "launched"
##  [9] "pledged"          "state"            "backers"          "country"
## [13] "usd.pledged"      "usd_pledged_real" "usd_goal_real"

ks_df$rowid <- paste(ks_df$ID, "-", ks_df$round)
length(unique(ks_df$rowid))

## [1] 378661

length(ks_df$rowid)

## [1] 378661
```

Here I confirmed that all rows have a unique ID. I also reviewed the data to ensure all the data I needed was contained within the data set.

####Check variables names

```
# checks variable names and replace with new name
library(dplyr)

##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
## 
##     filter, lag
```

```
## The following objects are masked from 'package:base':
## 
##     intersect, setdiff, setequal, union
```

```
ks_df <- rename(ks_df, usd_pledged = usd.pledged)
```

Here I renamed the variable usd.pledged to usd_pledged to align the naming conventions of all of my headers, since the rest of the headers uses underscores instead of periods for spaces.

####Check missing observations

```
# checks for missing values in observations
colMeans(is.na(ks_df))
```

```
##               ID             name         category    main_category 
##       0.00000000       0.00000000       0.00000000       0.00000000 
##         currency         deadline             goal         launched 
##       0.00000000       0.00000000       0.00000000       0.00000000 
##          pledged            state          backers          country 
##       0.00000000       0.00000000       0.00000000       0.00000000 
##      usd_pledged usd_pledged_real    usd_goal_real            rowid 
##       0.01002744       0.00000000       0.00000000       0.00000000
```

```
# removes column from data set
ks_df = subset(ks_df, select = -c(usd_pledged) )
```

Here I am looking for missing values. There is a small amount of data in the usd_pledged with missing values. If I wanted to cleanse the data set, I could remove these values, but for now, I want to keep it in mind since there are zero missing values from usd_pledged_real, which is a column giving the same information, but the conversion to USD was done from the fixer.io api instead of done by kickstarter. Instead of removing the rows with the missing data, I am going to remove the column from the data set since it is a duplicate column.

usd_pledged: conversion in US dollars of the pledged column (conversion done by kickstarter). usd pledge real: conversion in US dollars of the pledged column (conversion from Fixer.io API).

####Check variable classification

```
# checks attributes of data frame
str(ks_df)
```

```
## 'data.frame':    378661 obs. of  15 variables:
##  $ ID              : int  1000002330 1000003930 1000004038 1000007540 1000
011046 1000014025 1000023410 1000030581 1000034518 100004195 ...
```

```
##  $ name            : chr  "The Songs of Adelaide & Abullah" "Greeting From
Earth: ZGAC Arts Capsule For ET" "Where is Hank?" "ToshiCapital Rekordz Needs
Help to Complete Album" ...
##  $ category        : chr  "Poetry" "Narrative Film" "Narrative Film" "Musi
c" ...
##  $ main_category   : chr  "Publishing" "Film & Video" "Film & Video" "Musi
c" ...
##  $ currency        : chr  "GBP" "USD" "USD" "USD" ...
##  $ deadline        : chr  "2015-10-09" "2017-11-01" "2013-02-26" "2012-04-
16" ...
##  $ goal            : num  1000 30000 45000 5000 19500 50000 1000 25000 125
000 65000 ...
##  $ launched        : chr  "2015-08-11 12:12:28" "2017-09-02 04:43:57" "201
3-01-12 00:20:50" "2012-03-17 03:24:11" ...
##  $ pledged         : num  0 2421 220 1 1283 ...
##  $ state           : chr  "failed" "failed" "failed" "failed" ...
##  $ backers         : int  0 15 3 1 14 224 16 40 58 43 ...
##  $ country         : chr  "GB" "US" "US" "US" ...
##  $ usd_pledged_real: num  0 2421 220 1 1283 ...
##  $ usd_goal_real   : num  1534 30000 45000 5000 19500 ...
##  $ rowid           : chr  "1000002330 - " "1000003930 - " "1000004038 - "
"1000007540 - " ...
```

Checking the variable classification is the step used to make sure the data is the right datatype for analysis.

####Check duplicate rows

```
# Checking if one row is identical to another
distinctdata <- distinct(ks_df)
nrow(ks_df)

## [1] 378661

nrow(distinctdata)

## [1] 378661
```

Checking for duplicate rows within the data. None were found. If duplicate rows are found, the duplicate should be extracted from the dataset.

####Change dates from factors to date

```
ks_df <- transform(ks_df, deadline = as.Date(deadline), launched = as.Date(la
unched), backers = as.numeric(backers))
```

Changes the data type of deadline and launched to date.

What does the final data set look like?

```
head(ks_df)
```

```
##          ID                                                       name
## 1 1000002330                            The Songs of Adelaide & Abullah
## 2 1000003930              Greeting From Earth: ZGAC Arts Capsule For ET
## 3 1000004038                                             Where is Hank?
## 4 1000007540          ToshiCapital Rekordz Needs Help to Complete Album
## 5 1000011046 Community Film Project: The Art of Neighborhood Filmmaking
## 6 1000014025                                       Monarch Espresso Bar
##         category main_category currency   deadline  goal   launched pledge
d
## 1         Poetry    Publishing      GBP 2015-10-09  1000 2015-08-11
0
## 2 Narrative Film  Film & Video      USD 2017-11-01 30000 2017-09-02    242
1
## 3 Narrative Film  Film & Video      USD 2013-02-26 45000 2013-01-12     22
0
## 4          Music         Music      USD 2012-04-16  5000 2012-03-17
1
## 5   Film & Video  Film & Video      USD 2015-08-29 19500 2015-07-04    128
3
## 6    Restaurants          Food      USD 2016-04-01 50000 2016-02-26   5237
5
##        state backers country usd_pledged_real usd_goal_real         rowid
## 1     failed       0      GB                0       1533.95 1000002330 -
## 2     failed      15      US             2421      30000.00 1000003930 -
## 3     failed       3      US              220      45000.00 1000004038 -
## 4     failed       1      US                1       5000.00 1000007540 -
## 5   canceled      14      US             1283      19500.00 1000011046 -
## 6 successful     224      US            52375      50000.00 1000014025 -
```

Questions for future steps

I need to figure out if and how the factor/category data needs to be changed to numerical data. I also had to change dates from factors to date data types.

What information is not self-evident?

I plan to run both correlation and unsupervised learning models on the data to see if I can uncover any new information that is not self-evident.

What are different ways you could look at this data?

Yes, the questions I want to answer can be viewed though looking at bar charts, frequency plots and statistical models. * Are there certain types/category of campaigns that are more successful? * How much money should you ask for? * Is there a time period for the campaign that works better than others? * What is the average contribution of a backer? * Is there a better time of year to launch a campaign?

How do you plan to slice and dice the data?.

Created a new variable for % successful by taking the pledged and dividing it by the goal. I also slided out the month for both deadline and launch dates.

```r
# Adding new rows to slide and dice the data later
ks_df <-
  ks_df %>%
  mutate(
    pledged_to_goal = usd_pledged_real/usd_goal_real,
    count = 1,
    deadline_month = format(deadline,"%m"),
    launched_month = format(launched,"%m"),
    backers_per_pledge = usd_pledged_real/backers
    )
```

How could you summarize your data to answer key questions?

This ties into the different ways I can look at the data set. Charts and visualizations are a great way to summarize the data and answer key questions.

What types of plots and tables will help you to illustrate the findings to your questions?

Bar charts, box plots and scatter charts will help illustrate findings to my questions.

Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

Yes, I plan to see if there are any supervised (like decision tree or random forest) models and unsupervised (clustering) that can help make sense of what is funded verses unfunded.

Questions for future steps

This still ties in to question #3, where I need to figure out if the factor/categyory data needs to be changed to numerical data and if so, how I go about doing that.

Section 3

Introduction

Kickstarter campaigns is a way to crowdsource funding to support projects, people or situations. It's a way to raise money. In this analysis, I will be finding out if there are controllable factors which can lead to a successful campaign.

The problem statement you addressed

Is there a way to design a kickstarter campaign to increase it's likelihood to be successful?

How you addressed this problem statement

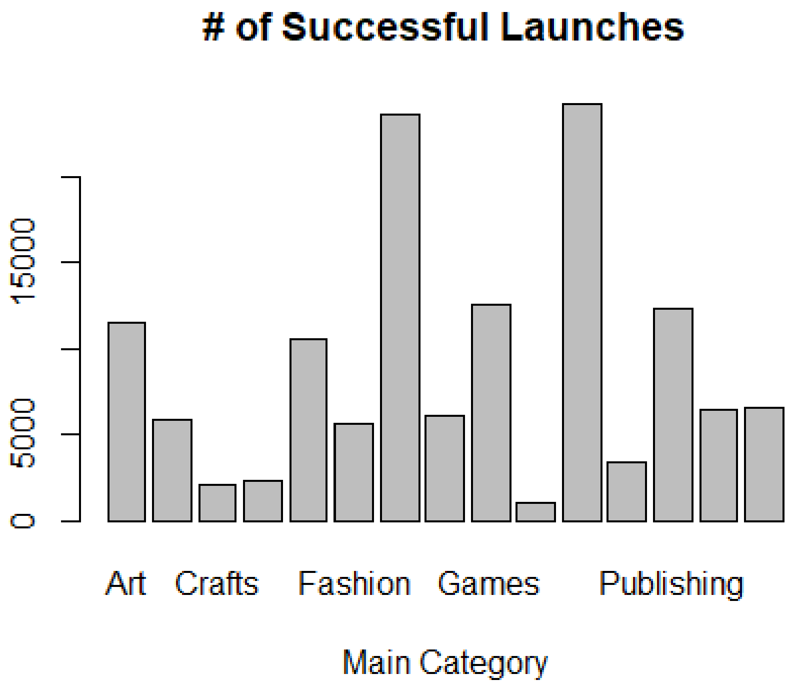
I addressed this problem statement by looking into answering 5 questions: * Are there certain types/category of campaigns that are more successful? * How much money should you ask for? * Is there a time period for the campaign that works better than others? * What is the average contribution of a backer? * Is there a better time of year to launch a campaign?

I also performed correlation and applied machine learning techniques to see if there are ways to increase the likelihood of building successful campaigns.

Analysis

Are there certain types/category of campaigns that are more successful?

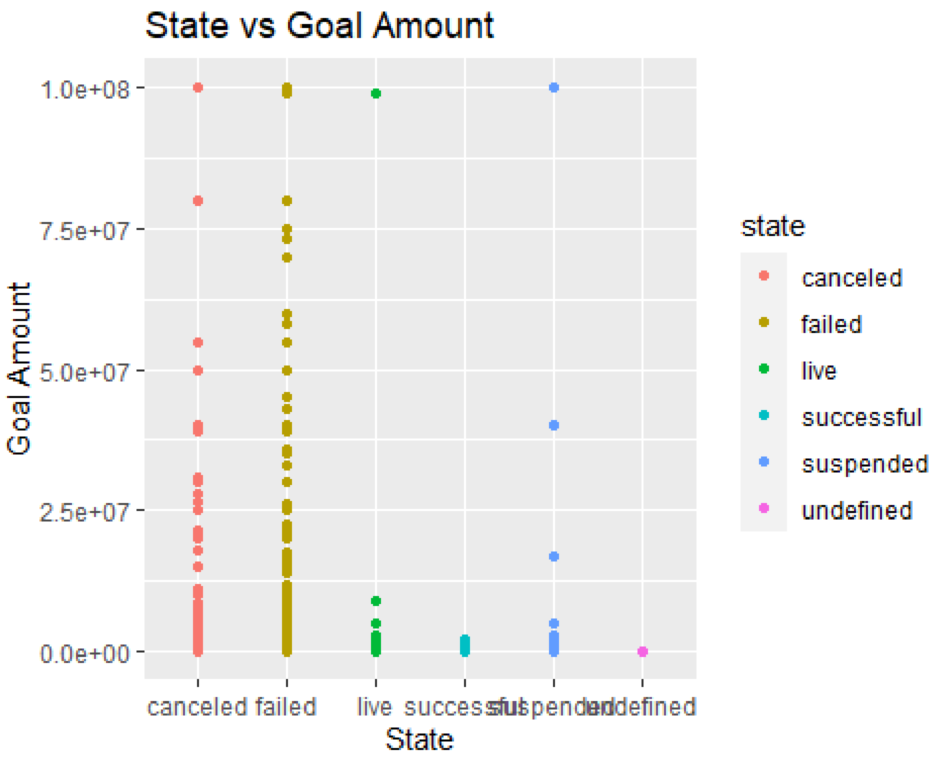
```
# Filtering by one criterion
ks_dff <- filter(ks_df, state == "successful")

## Successful Launches based on Deadline Month
counts <- table(ks_dff$main_category)
barplot(counts, main="# of Successful Launches",
   xlab="Main Category")
```



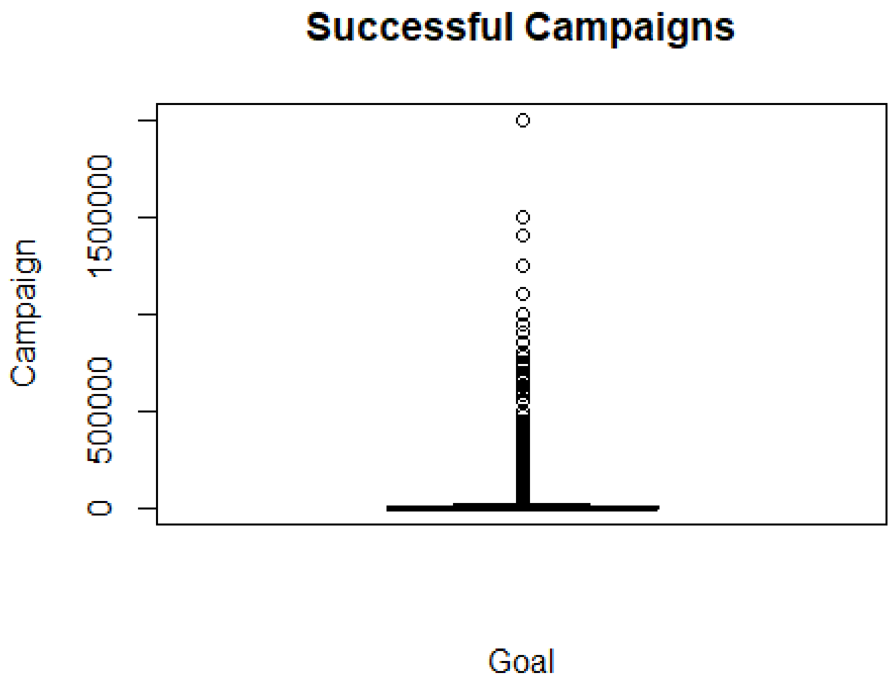
The top 5 categories with successful campaigns are: 1. Music 2. Film & Video 3. Games 4. Publishing 5. Art

How much money should you ask for?

```
library(ggplot2)
## Create a scatterplot of all states
ggplot(ks_df, aes(x=state, y=goal, col=state)) + ggtitle("State vs Goal Amoun
t") + xlab("State") + ylab("Goal Amount") + geom_point(aes(colour = state))
```



```
# Boxplot of only successful campaigns
boxplot(goal~count,data=ks_dff, main="Successful Campaigns",
   xlab="Goal", ylab="Campaign")
```

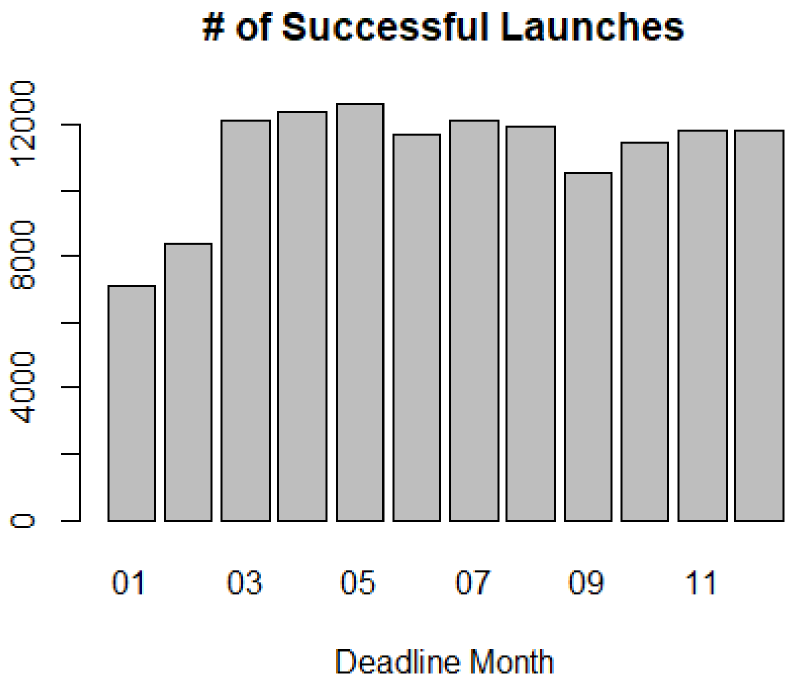


```
summary(ks_dff)

##        ID                name             category         main_category
##  Min.   :2.111e+04   Length:133956      Length:133956      Length:133956
##  1st Qu.:5.354e+08   Class :character   Class :character   Class :characte
r
##  Median :1.077e+09   Mode  :character   Mode  :character   Mode  :characte
r
##  Mean   :1.074e+09
##  3rd Qu.:1.608e+09
##  Max.   :2.147e+09
##    currency             deadline                goal            launched
##  Length:133956      Min.   :2009-05-03   Min.   :      0   Min.   :2009-04
-24
##  Class :character   1st Qu.:2012-12-13   1st Qu.:   1250   1st Qu.:2012-11
-13
##  Mode  :character   Median :2014-08-29   Median :   3923   Median :2014-07
-29
##                     Mean   :2014-07-31   Mean   :  10163   Mean   :2014-06
-29
##                     3rd Qu.:2016-04-13   3rd Qu.:  10000   3rd Qu.:2016-03
-12
##                     Max.   :2018-01-02   Max.   :2000000   Max.   :2017-12
-29
##     pledged             state              backers            country
##  Min.   :      1   Length:133956      Min.   :     0.0   Length:133956
##  1st Qu.:   1978   Class :character   1st Qu.:    33.0   Class :character
##  Median :    5117   Mode  :character   Median :    71.0   Mode  :character
##  Mean   :   24100                      Mean   :   263.9
##  3rd Qu.:   13440                      3rd Qu.:   167.0
##  Max.   :20338986                      Max.   :219382.0
##  usd_pledged_real   usd_goal_real        rowid            pledged_to_goal
##  Min.   :       1   Min.   :      0   Length:133956      Min.   :     0.85
##  1st Qu.:    2000   1st Qu.:   1302   Class :character   1st Qu.:     1.05
##  Median :    5107   Median :   3838   Mode  :character   Median :     1.17
##  Mean   :   22671   Mean   :   9533                      Mean   :     8.56
##  3rd Qu.:   13232   3rd Qu.:  10000                      3rd Qu.:     1.63
##  Max.   :20338986   Max.   :2015609                      Max.   :104277.89
##      count    deadline_month     launched_month     backers_per_pledge
##  Min.   :1   Length:133956      Length:133956      Min.   :  0.7835
##  1st Qu.:1   Class :character   Class :character   1st Qu.: 41.1972
##  Median :1   Mode  :character   Mode  :character   Median : 63.3473
##  Mean   :1                                         Mean   :      Inf
##  3rd Qu.:1                                         3rd Qu.:102.3367
##  Max.   :1                                         Max.   :      Inf
```

Successful campaigns have a smaller range then non-successful campaigns. The average successful campaign has a goal of around 10,000 with a median of around 4,000.

Is there a time period for the campaign that works better than others?

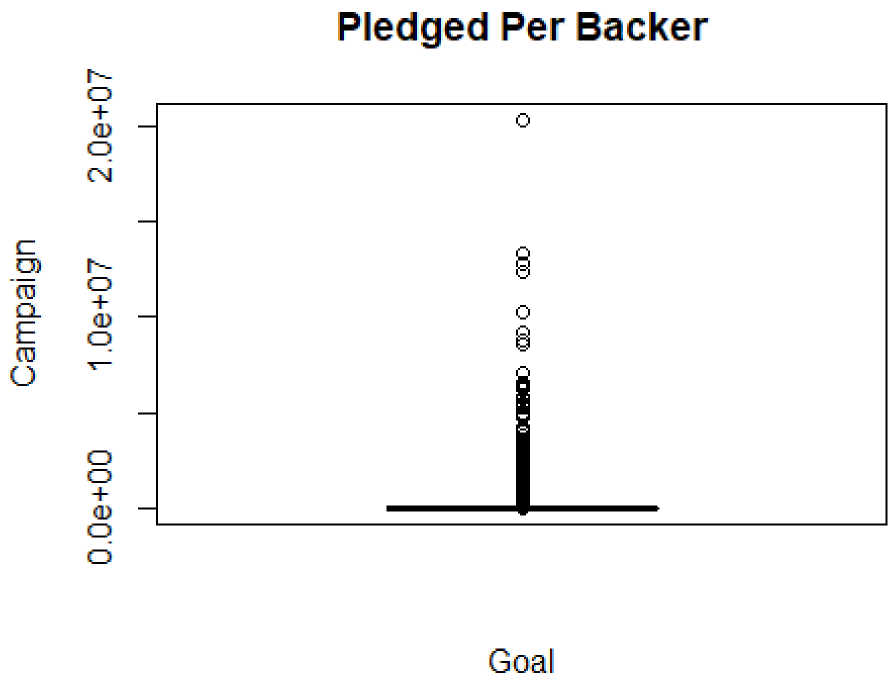
```
## Successful Launches based on Deadline Month
counts <- table(ks_dff$deadline_month)
barplot(counts, main="# of Successful Launches",
   xlab="Deadline Month")
```



May has the highest number of campaigns that are successful.

What is the average contribution of a backer?

```
# Boxplot of only successful campaigns
boxplot(usd_pledged_real~count,data=ks_dff, main="Pledged Per Backer",
   xlab="Goal", ylab="Campaign")
```

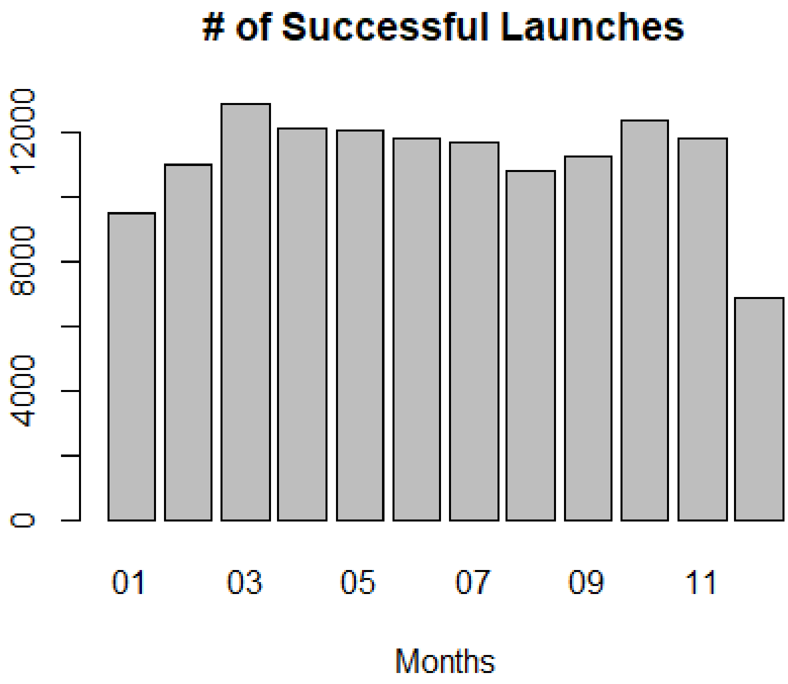


```
summary(ks_dff)

##        ID                  name             category          main_category
##  Min.   :2.111e+04   Length:133956      Length:133956      Length:133956
##  1st Qu.:5.354e+08   Class :character   Class :character   Class :characte
r
##  Median :1.077e+09   Mode  :character   Mode  :character   Mode  :characte
r
##  Mean   :1.074e+09
##  3rd Qu.:1.608e+09
##  Max.   :2.147e+09
##    currency             deadline               goal             launched
##  Length:133956      Min.   :2009-05-03   Min.   :      0   Min.   :2009-04
-24
##  Class :character   1st Qu.:2012-12-13   1st Qu.:   1250   1st Qu.:2012-11
-13
##  Mode  :character   Median :2014-08-29   Median :   3923   Median :2014-07
-29
##                     Mean   :2014-07-31   Mean   :  10163   Mean   :2014-06
-29
##                     3rd Qu.:2016-04-13   3rd Qu.:  10000   3rd Qu.:2016-03
-12
##                     Max.   :2018-01-02   Max.   :2000000   Max.   :2017-12
-29
##     pledged              state              backers            country
##  Min.   :        1   Length:133956      Min.   :     0.0   Length:133956
##  1st Qu.:     1978   Class :character   1st Qu.:    33.0   Class :character
##  Median :    5117   Mode  :character   Median :    71.0   Mode  :character
##  Mean   :   24100                      Mean   :   263.9
##  3rd Qu.:   13440                      3rd Qu.:   167.0
##  Max.   :20338986                      Max.   :219382.0
##  usd_pledged_real   usd_goal_real        rowid           pledged_to_goal
##  Min.   :       1   Min.   :      0   Length:133956      Min.   :     0.85
##  1st Qu.:    2000   1st Qu.:   1302   Class :character   1st Qu.:     1.05
##  Median :    5107   Median :   3838   Mode  :character   Median :     1.17
##  Mean   :   22671   Mean   :   9533                      Mean   :     8.56
##  3rd Qu.:   13232   3rd Qu.:  10000                      3rd Qu.:     1.63
##  Max.   :20338986   Max.   :2015609                      Max.   :104277.89
##      count    deadline_month     launched_month     backers_per_pledge
##  Min.   :1   Length:133956      Length:133956      Min.   :  0.7835
##  1st Qu.:1   Class :character   Class :character   1st Qu.: 41.1972
##  Median :1   Mode  :character   Mode  :character   Median : 63.3473
##  Mean   :1                                         Mean   :     Inf
##  3rd Qu.:1                                         3rd Qu.:102.3367
##  Max.   :1                                         Max.   :     Inf
```

The median backer pledges 63 USD to projects.

Is there a better time of year to launch a campaign?

```
## Successful Launches
counts <- table(ks_dff$launched_month)
barplot(counts, main="# of Successful Launches",
   xlab="Months")
```



March and October has the most for count of successful launches. December has the least.

What are the factors that contribute to sucessful campaigns?

```
# Prepping the data for modelling:

# Adding new rows to indicate successful campaigns
ks_dff <-
  ks_dff %>%
  mutate(
    successful = 1
    )

# Filtering by one criterion where campaigns not successful
ks_dfn <- filter(ks_df, state != "successful")


# Adding new rows to indicate unsuccessful campaigns
ks_dfn <-
  ks_dfn %>%
  mutate(
    successful = 0
    )

#combines successful and unsuccessful campaigns
df_union1<-merge(ks_dff,ks_dfn,all=TRUE)
```

```
df_union1 <- transform(df_union1, deadline_month = as.integer(deadline_month)
, launched_month = as.integer(launched_month))

model_1 <- lm(successful ~ backers+usd_pledged_real,usd_goal_real,pledged_to_
goal+deadline_month+launched_month, data = df_union1)
summary(model_1)

##
## Call:
## lm(formula = successful ~ backers + usd_pledged_real, data = df_union1,
##     subset = usd_goal_real, weights = pledged_to_goal + deadline_month +
##         launched_month)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -46.295  -1.627  -1.223   2.021  94.783
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4.044e-01  8.040e-04 502.957   <2e-16 ***
## backers           1.562e-04  8.386e-07 186.275   <2e-16 ***
## usd_pledged_real -6.924e-09  1.088e-08  -0.637    0.524
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.939 on 375212 degrees of freedom
##   (3408 observations deleted due to missingness)
## Multiple R-squared:  0.1211, Adjusted R-squared:  0.1211
## F-statistic: 2.584e+04 on 2 and 375212 DF,  p-value: < 2.2e-16

library(lm.beta)
model_1.beta <- lm.beta(model_1)
coef(model_1.beta)

##      (Intercept)          backers usd_pledged_real
##     0.0000000000     0.1543442876    -0.0006805745

# linear regression on backers
linearMod <- lm(successful ~ backers, data=df_union1)
print(linearMod)

##
## Call:
## lm(formula = successful ~ backers, data = df_union1)
##
## Coefficients:
## (Intercept)      backers
##   3.466e-01    6.805e-05

summary(linearMod)
```

```
##
## Call:
## lm(formula = successful ~ backers, data = df_union1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2750  -0.3470  -0.3466   0.6439   0.6534
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.466e-01  7.757e-04  446.78   <2e-16 ***
## backers     6.805e-05  8.493e-07   80.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4741 on 378659 degrees of freedom
## Multiple R-squared:  0.01667,    Adjusted R-squared:  0.01667
## F-statistic:  6419 on 1 and 378659 DF,  p-value: < 2.2e-16
```

The number of backers is a significant factor when predicting if the kickstarter will be a success.

Conclusion

The best way to have a successful campaign is to increase the number of backers for that campaign. Would not recommend campaigning during the holiday season.