Final_Summary

Jonathan Henin

November 15, 2018

- I’m interested in using the Kaggle dataset on Kickstarter projects. I’ve always found the notion of crowdfunding interesting and I feel like it would be a fun dataset.
- Having an original idea for a product and bringing it to the market can be a challenge, especially the financing of the product. Kickstarter is a crowd sourcing platform which shifts the control from banks to the people to decide what products are worthy of being funded. However, even on Kickstarter a products success is not guaranteed. If a person would go down the Kickstarter route, they would want to make sure they are doing everything they can to be successful. Vice versa, as a backer you want to know the success rates of projects as well, but you also want to see which products once the goal is reached have the highest chance of being successful.

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

```
library(plyr)
library(dplyr)
library(ggplot2)
library(readr)
library(tidyr)
library(Hmisc)
library(lubridate)
library(scales)
library(tokenizers)
library(stopwords)
library(tidytext)
library(stringr)
library(foreign)
library(caret)
data(stop_words)
options(scipen = 999)
```

- `ID` internal kickstarter id
- `name` name of project - A project is a finite work with a clear goal that you’d like to bring to life. Think albums, books, or films.
- `category` category
- `main_category` category of campaign
- `currency` currency used to support
- `deadline` deadline for crowdfunding
- `goal` fundraising goal - The funding goal is the amount of money that a creator needs to complete their project.
- `launched` date launched
- `pledged` amount pledged by ‘crowd’
- `state` Current condition the project is in
- `backers` number of backers
- `country` country pledged from
- `usd.pledged` amount of money pledged

- `used_pledged_real` amount of money pledged cleaned
- `usd_goal_real` amount in USD

List of 7 research questions I aim to answer.

```
1.  What are the most popular Kickstarter categories, and which have the highest rate of success
/ lowest?
2.  Which Kickstarter campaigns have the most backers and the highest pledges per backer?
3.  Which Kickstarter campaigns goes the most beyond their initial goal (stretch goals)?
4.  What is the correlation between the amount of time given to meet a goal and its success?
5.  Which Kickstarter campaigns have the lowest chance to fail after their goal is met?
6.  Which words have the highest correlation with success and which ones have the lowest?
7.  Can we build a regression model to predict success?
```

```
ks_file <- 'ks-projects-201801.csv'

ks_data <- read.csv(ks_file, header = T)
```

```
str(ks_data)
```

```
## 'data.frame':    378661 obs. of  15 variables:
##  $ ID              : int  1000002330 1000003930 1000004038 1000007540 1000011046 1000014025 1
000023410 1000030581 1000034518 100004195 ...
##  $ name            : Factor w/ 375765 levels "","\177Not Twins - New EP! \"The View from Down
Here\"",..: 332541 135689 365010 344805 77349 206130 293462 69360 284139 290718 ...
##  $ category        : Factor w/ 159 levels "3D Printing",..: 109 94 94 91 56 124 59 42 114 40
...
##  $ main_category   : Factor w/ 15 levels "Art","Comics",..: 13 7 7 11 7 8 8 8 5 7 ...
##  $ currency        : Factor w/ 14 levels "AUD","CAD","CHF",..: 6 14 14 14 14 14 14 14 14 14
...
##  $ deadline        : Factor w/ 3164 levels "2009-05-03","2009-05-16",..: 2288 3042 1333 1017
2247 2463 1996 2448 1790 1863 ...
##  $ goal            : num  1000 30000 45000 5000 19500 50000 1000 25000 125000 65000 ...
##  $ launched        : Factor w/ 378089 levels "1970-01-01 01:00:00",..: 243292 361975 80409 46
557 235943 278600 187500 274014 139367 153766 ...
##  $ pledged         : num  0 2421 220 1 1283 ...
##  $ state           : Factor w/ 6 levels "canceled","failed",..: 2 2 2 2 1 4 4 2 1 1 ...
##  $ backers         : int  0 15 3 1 14 224 16 40 58 43 ...
##  $ country         : Factor w/ 23 levels "AT","AU","BE",..: 10 23 23 23 23 23 23 23 23 23 ...
##  $ usd.pledged     : num  0 100 220 1 1283 ...
##  $ usd_pledged_real: num  0 2421 220 1 1283 ...
##  $ usd_goal_real   : num  1534 30000 45000 5000 19500 ...
```

```
Hmisc::describe(ks_data)
```

```
## ks_data
##
##  15  Variables      378661  Observations
## ---------------------------------------------------------------------------
## ID
##          n    missing   distinct       Info       Mean        Gmd
##     378661          0     378661          1 1074731192  714859359
##        .05        .10        .25        .50        .75        .90
##  108769050  216410277  538263516 1075275634 1610148624 1932082525
##        .95
## 2039733043
##
## lowest :       5971      18520      21109      21371      24380
## highest: 2147455254 2147460119 2147466649 2147472329 2147476221
## ---------------------------------------------------------------------------
## name
##        n  missing distinct
##   378661        0   375765
##
## lowest :                                                    Not Twins - New EP! "The View fro
m Down Here" '' Album''Eyes to Eyes''of Kilimandjaro' '        '' Bone crusher ''
''1985'' Le Spectacle / The Show
## highest: zzz                                               zzz (Canceled)
zZzleepy cat                                       ZzzMask, awesome sleep on a plane.
Zzzymble
## ---------------------------------------------------------------------------
## category
##        n  missing distinct
##   378661        0      159
##
## lowest : 3D Printing Academic    Accessories Action      Animals
## highest: Woodworking Workshops   World Music Young Adult Zines
## ---------------------------------------------------------------------------
## main_category
##        n  missing distinct
##   378661        0       15
##
## Art (28153, 0.074), Comics (10819, 0.029), Crafts (8809, 0.023), Dance
## (3768, 0.010), Design (30070, 0.079), Fashion (22816, 0.060), Film & Video
## (63585, 0.168), Food (24602, 0.065), Games (35231, 0.093), Journalism
## (4755, 0.013), Music (51918, 0.137), Photography (10779, 0.028),
## Publishing (39874, 0.105), Technology (32569, 0.086), Theater (10913,
## 0.029)
## ---------------------------------------------------------------------------
## currency
##        n  missing distinct
##   378661        0       14
##
## Value         AUD    CAD    CHF    DKK    EUR    GBP    HKD    JPY    MXN
## Frequency    7950  14962    768   1129  17405  34132    618     40   1752
## Proportion  0.021  0.040  0.002  0.003  0.046  0.090  0.002  0.000  0.005
##
## Value         NOK    NZD    SEK    SGD    USD
```

```
## Frequency     722   1475   1788    555 295365
## Proportion  0.002  0.004  0.005  0.001  0.780
## ---------------------------------------------------------------------------
## deadline
##        n  missing distinct
##   378661        0     3164
##
## lowest : 2009-05-03 2009-05-16 2009-05-20 2009-05-22 2009-05-26
## highest: 2018-02-27 2018-02-28 2018-03-01 2018-03-02 2018-03-03
## ---------------------------------------------------------------------------
## goal
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##   378661        0     8353    0.999    49081    87272      400      675
##      .25      .50      .75      .90      .95
##     2000     5200    16000    50000    90000
##
## lowest :         0.01         0.15         0.50         1.00         1.85
## highest:  73000000.00  75000000.00  80000000.00  99000000.00 100000000.00
## ---------------------------------------------------------------------------
## launched
##        n  missing distinct
##   378661        0   378089
##
## lowest : 1970-01-01 01:00:00 2009-04-21 21:02:48 2009-04-23 00:07:53 2009-04-24 21:52:03 2009
-04-25 17:36:21
## highest: 2018-01-02 14:13:09 2018-01-02 14:15:38 2018-01-02 14:17:46 2018-01-02 14:38:17 2018
-01-02 15:02:31
## ---------------------------------------------------------------------------
## pledged
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##   378661        0    62130    0.997     9683    17280        0        0
##      .25      .50      .75      .90      .95
##       30      620     4076    14141    29581
##
## lowest :        0.00        1.00        1.01        1.02        1.03
## highest: 10266845.74 12393139.69 12779843.49 13285226.36 20338986.27
## ---------------------------------------------------------------------------
## state
##        n  missing distinct
##   378661        0        6
##
## Value        canceled     failed       live successful  suspended
## Frequency       38779     197719       2799     133956       1846
## Proportion      0.102      0.522      0.007      0.354      0.005
##
## Value       undefined
## Frequency        3562
## Proportion      0.009
## ---------------------------------------------------------------------------
## backers
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##   378661        0     3963    0.996    105.6      182        0        0
##      .25      .50      .75      .90      .95
##        2       12       56      166      334
```

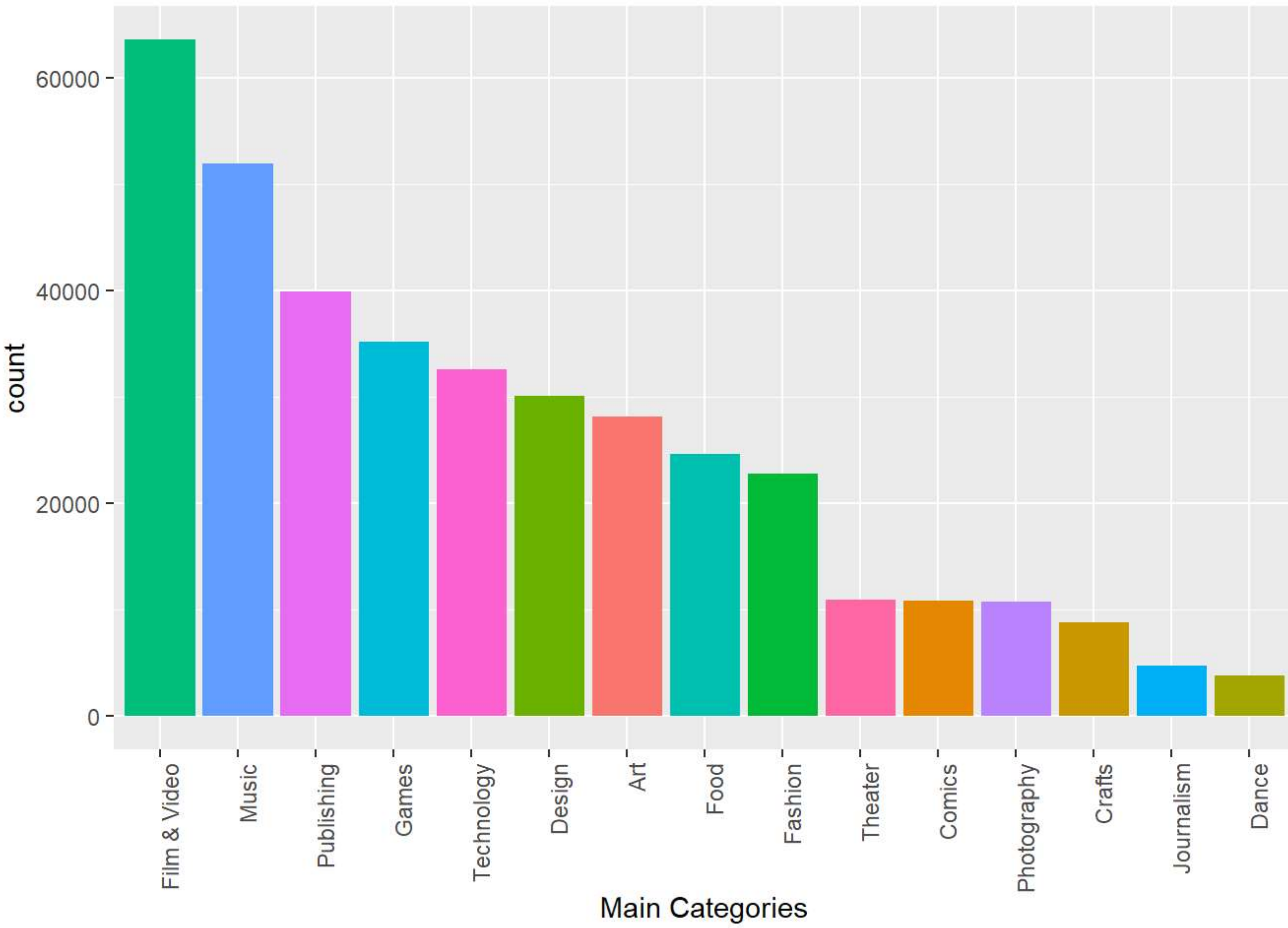
```
##
## lowest :      0      1      2      3      4, highest:  87142  91585 105857 154926 219382
## ---------------------------------------------------------------------------
## country
##        n  missing distinct
##   378661        0       23
##
## lowest : AT AU BE CA CH, highest: NO NZ SE SG US
## ---------------------------------------------------------------------------
## usd.pledged
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##   374864     3797    95455    0.994     7037    12556     0.00     0.00
##      .25      .50      .75      .90      .95
##    16.98   394.72  3034.09 10859.70 22432.85
##
## lowest :        0.00        0.47        0.48        0.51        0.52
## highest:  9192055.66 10266845.74 12779843.49 13285226.36 20338986.27
## ---------------------------------------------------------------------------
## usd_pledged_real
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##   378661        0   106065    0.997     9059    16072      0.0      0.0
##      .25      .50      .75      .90      .95
##     31.0    624.3   4050.0  13671.0  28090.0
##
## lowest :        0.00        0.45        0.47        0.48        0.49
## highest: 10266845.74 12393139.69 12779843.49 13285226.36 20338986.27
## ---------------------------------------------------------------------------
## usd_goal_real
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##   378661        0    50339    0.999    45454    80280      400      700
##      .25      .50      .75      .90      .95
##     2000     5500    15500    45000    80000
##
## lowest :         0.01         0.15         0.49         0.50         0.55
## highest: 104057189.83 107369867.72 110169771.62 151395869.92 166361390.71
## ---------------------------------------------------------------------------
```

In general this looks good, a few things of note.
* There are 3797 missing values in usd.pledged but according to the data source usd_pledged_real is already a cleaned up version of that column. * `launched` and `deadline` need to be changed to dates. * Add `days_to_goal` column calculating the difference between launched and deadline * There are some strange values in launched with years in 1970, these are probably dummy values and I'll remove those records. * Make name a character instead of a factor

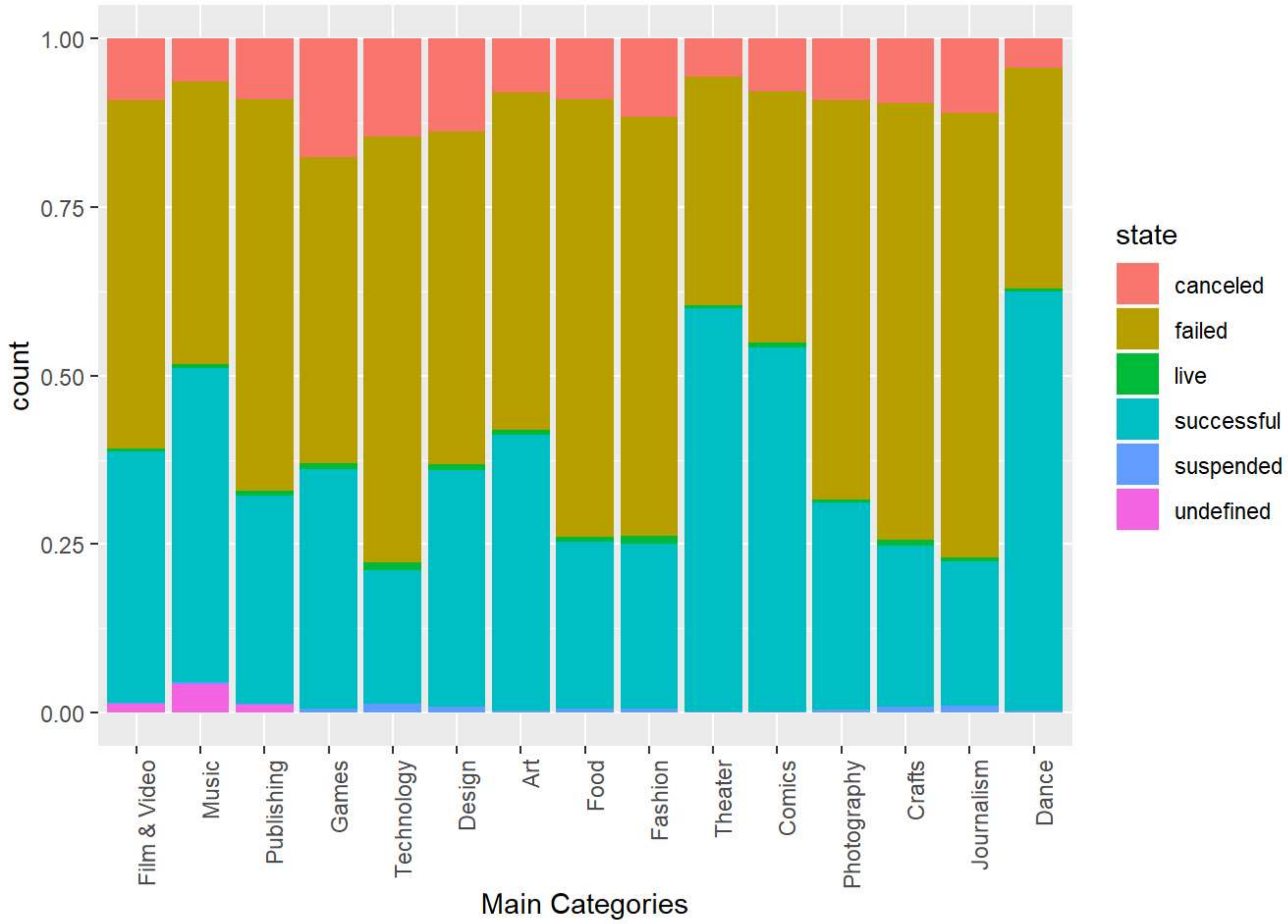
```
ks_data_cleaned <- ks_data

ks_data_cleaned$launched <- ymd_hms(as.character(ks_data_cleaned$launched))
ks_data_cleaned$deadline <- ymd(as.character(ks_data_cleaned$deadline))
ks_data_cleaned$days_to_goal <- interval(ks_data_cleaned$launched, ks_data_cleaned$deadline) %/%
days(1)
ks_data_cleaned <- ks_data_cleaned[(ks_data_cleaned$launched >= '2000-01-01'),]
ks_data_cleaned$name <- as.character(ks_data_cleaned$name)
```

1. What are the most popular Kickstarter categories, and which have the highest rate of success / lowest?

```
ggplot(ks_data_cleaned, aes(x = reorder(main_category, main_category, function(x)-length(x)), fi
ll=main_category)) +
  geom_bar() +
  labs(x = 'Main Categories') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1), legend.position='none')
```



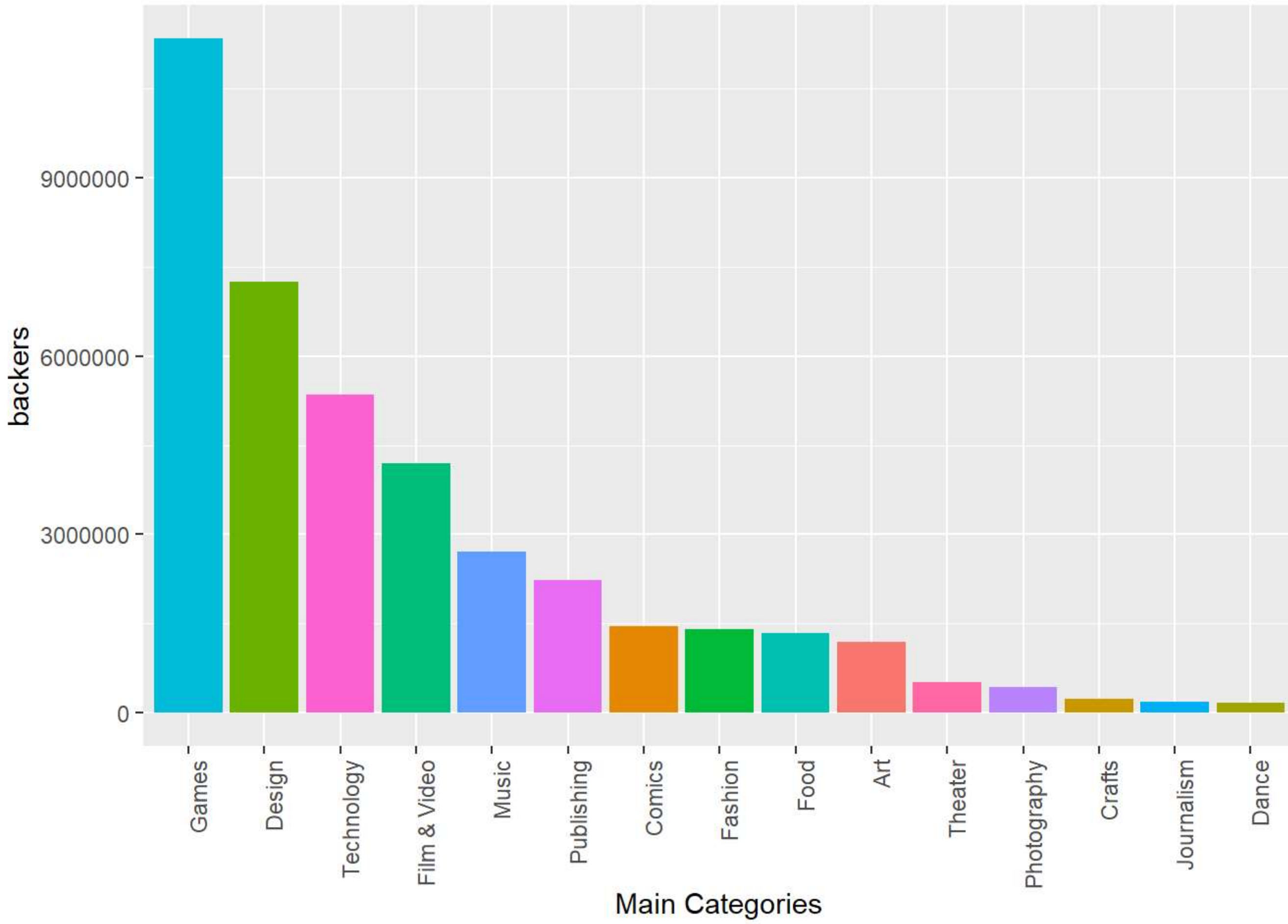

```
ggplot(ks_data_cleaned, aes(fill=state, x = reorder(main_category, main_category, function(x)-le
ngth(x)))) +
  geom_bar(position='fill') +
  labs(x = 'Main Categories') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



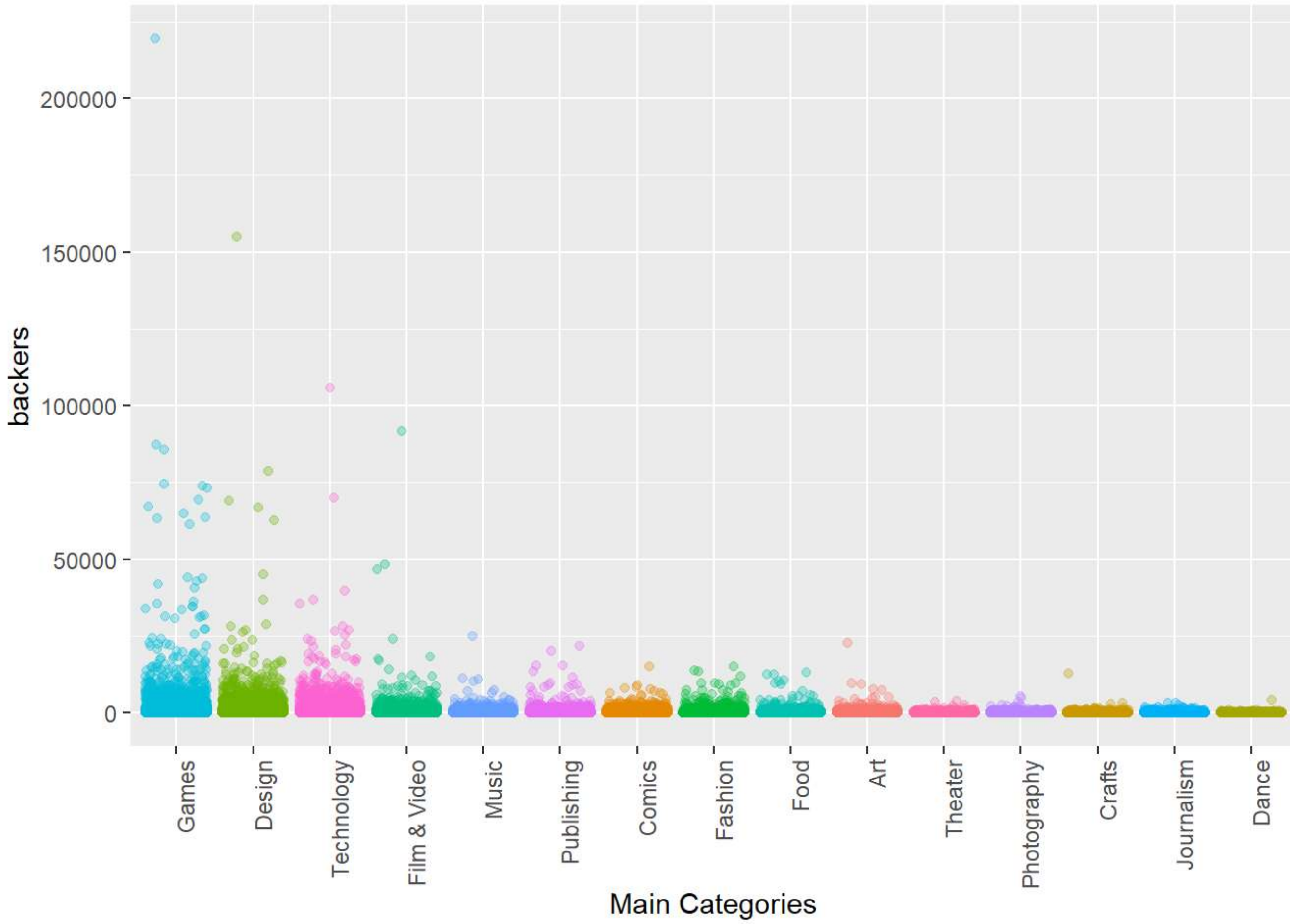
We can see that Film & Video, Music, and Publishing are the top 3 categories in terms of number of kickstarter projects. However, neither of those three are in the top 3 for highest chance of success. That honor goes to Dance, Theater and Comics.
The worst 3 categories in terms of chance of success seem to be Journalism, Technology, and Crafts.

2. Which Kickstarter campaigns have the most backers and the highest pledges per backer?

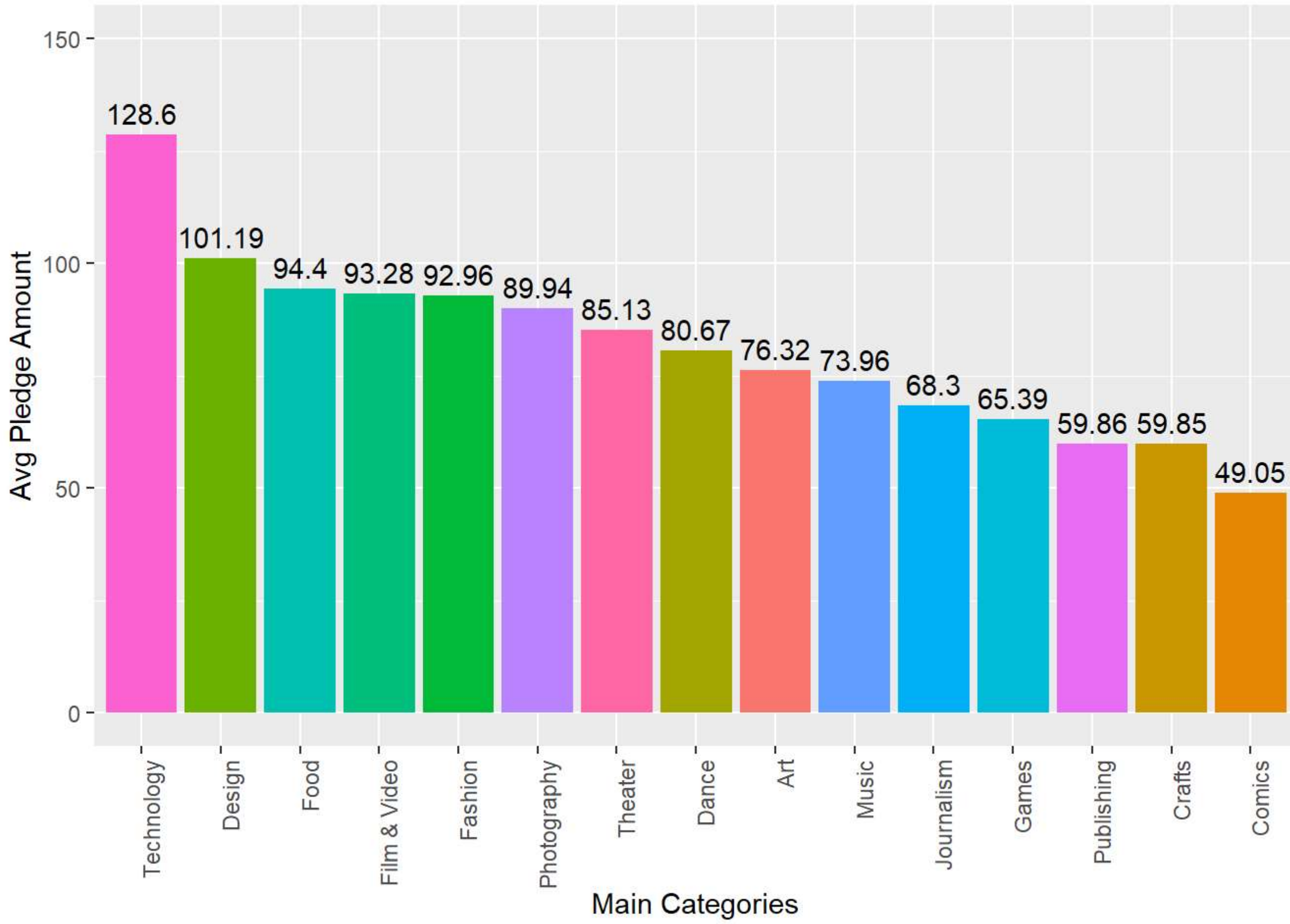
```
ggplot(ks_data_cleaned, aes(x = reorder(main_category, -backers, sum), y = backers, fill=main_ca
tegory)) +
  geom_col() +
  labs(x = 'Main Categories') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1), legend.position='none')
```




```
ggplot(ks_data_cleaned, aes(x = reorder(main_category, -backers, sum), y = backers, color = main
_category)) +
  geom_jitter(alpha = .3) +
  labs(x = 'Main Categories') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1), legend.position='none')
```



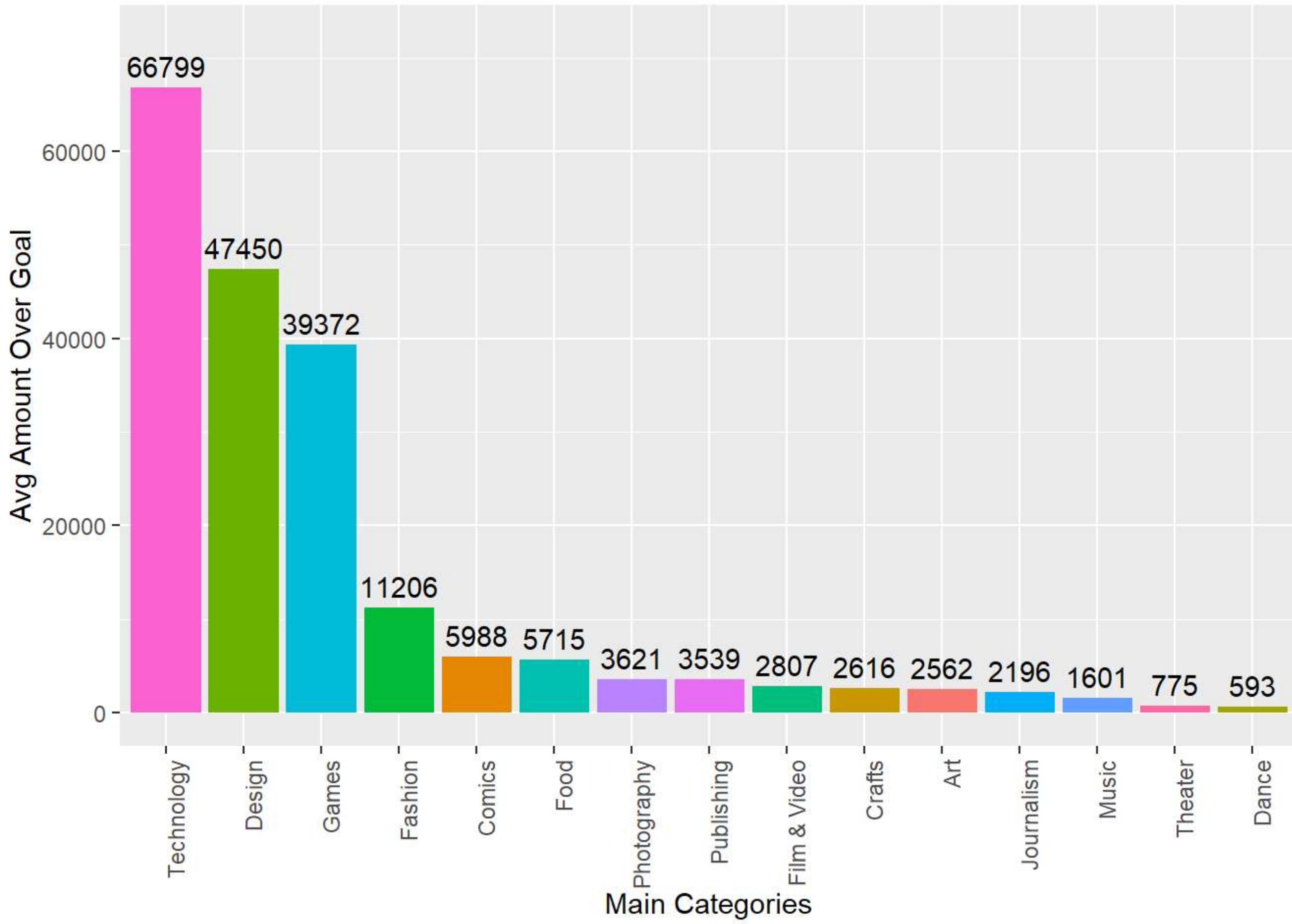

```
ks_data_pledged <- ks_data_cleaned %>%
  group_by(main_category) %>%
  dplyr::summarise(pledged = sum(usd_pledged_real), backers = sum(backers))

ggplot(ks_data_pledged, aes(x = reorder(main_category, -(pledged / backers), sum), y = (pledged
/ backers), fill=main_category)) +
  geom_col() +
  labs(x = 'Main Categories', y = 'Avg Pledge Amount') +
  geom_text(aes(label = round((pledged / backers),2)), vjust = -0.5) +
  ylim(0, 150) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1), legend.position='none')
```



Games, Design and Technology have the most number of backers, but that doesn't necessarily mean they are willing to pay out more. As we can see, Games backers generally pledge a lot less than Design and Technology. With Technology being the highest. This is interesting considering technology has one of the lowest chances of success. This might because the pledge categories are higher for technology vs games but unfortunately we don't have the level of detail.

3. Which Kickstarter campaign goes the most beyond their initial goal (stretch goals)?

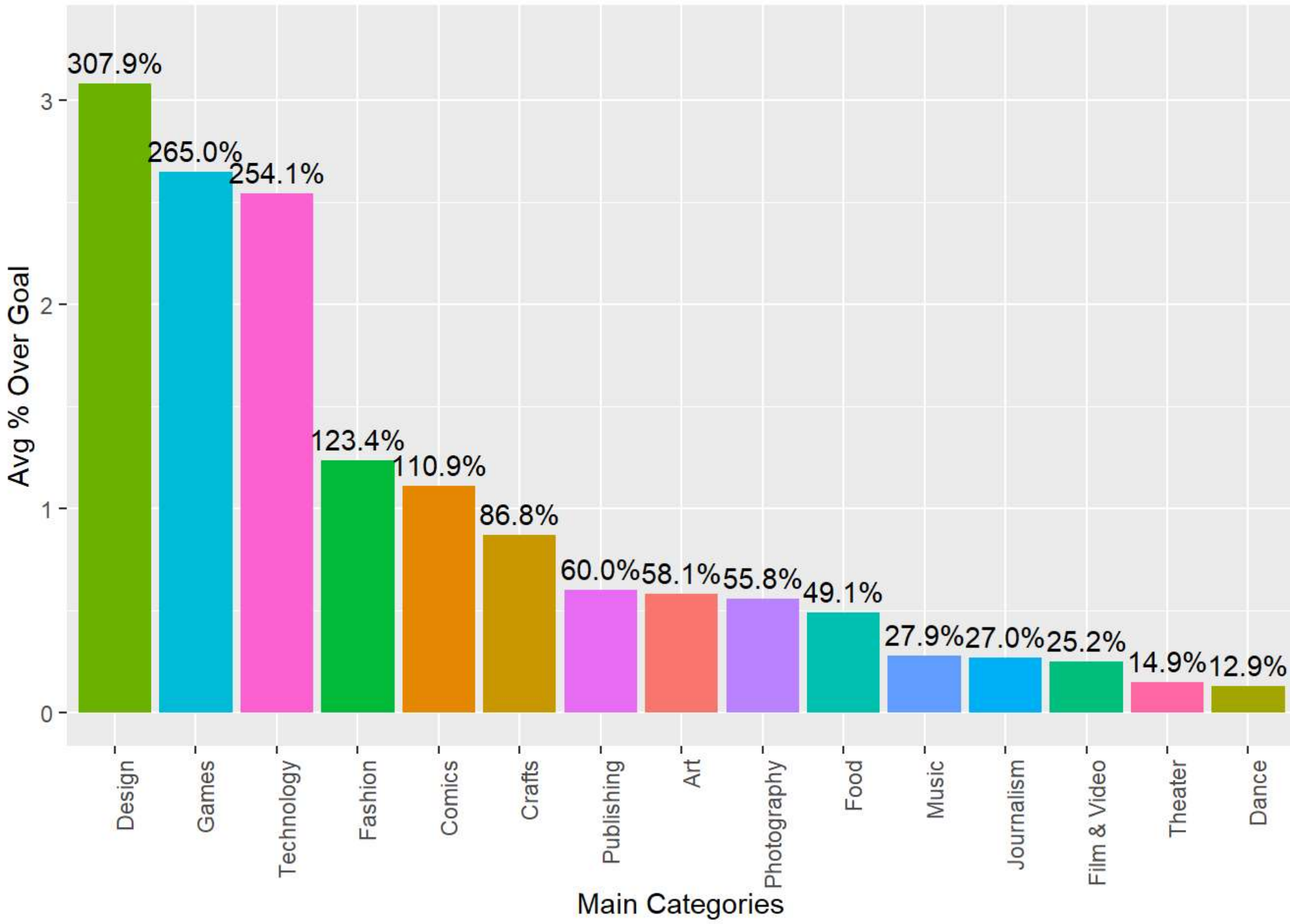
```
beyond_goal <- ks_data_cleaned %>%
  filter(state %in% c('successful')) %>%
  group_by(main_category) %>%
  dplyr::summarise(count=n(), pledged = sum(usd_pledged_real), goal = sum(usd_goal_real)) %>%
  mutate(avgover=(pledged-goal)/count)

ggplot(beyond_goal, aes(x = reorder(main_category, -avgover, sum), y = avgover, fill=main_catego
ry)) +
  geom_col() +
  labs(x = 'Main Categories', y = 'Avg Amount Over Goal') +
  ylim(0, 72000) +
  geom_text(aes(label = round(avgover, 0), vjust = -0.5)) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1), legend.position='none')
```



Right in line with campaigns that have the most backers, it makes sense Technology, Design and Games raise a lot more money past their goals. But what if we normalized it to look at percent over the goal.

```
percent1 <- function(x, digits = 1, format = "f", ...) {
  paste0(formatC(100 * x, format = format, digits = digits, ...), "%")
}
```

```
ggplot(beyond_goal, aes(x = reorder(main_category, -(avgover / (goal / count)), sum), y = (avgov
er / (goal / count)), fill=main_category)) +
  geom_col() +
  labs(x = 'Main Categories', y = 'Avg % Over Goal') +
  ylim(0, 3.3) +
  geom_text(aes(label = percent1((avgover / (goal / count))), vjust = -0.5)) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1), legend.position='none')
```

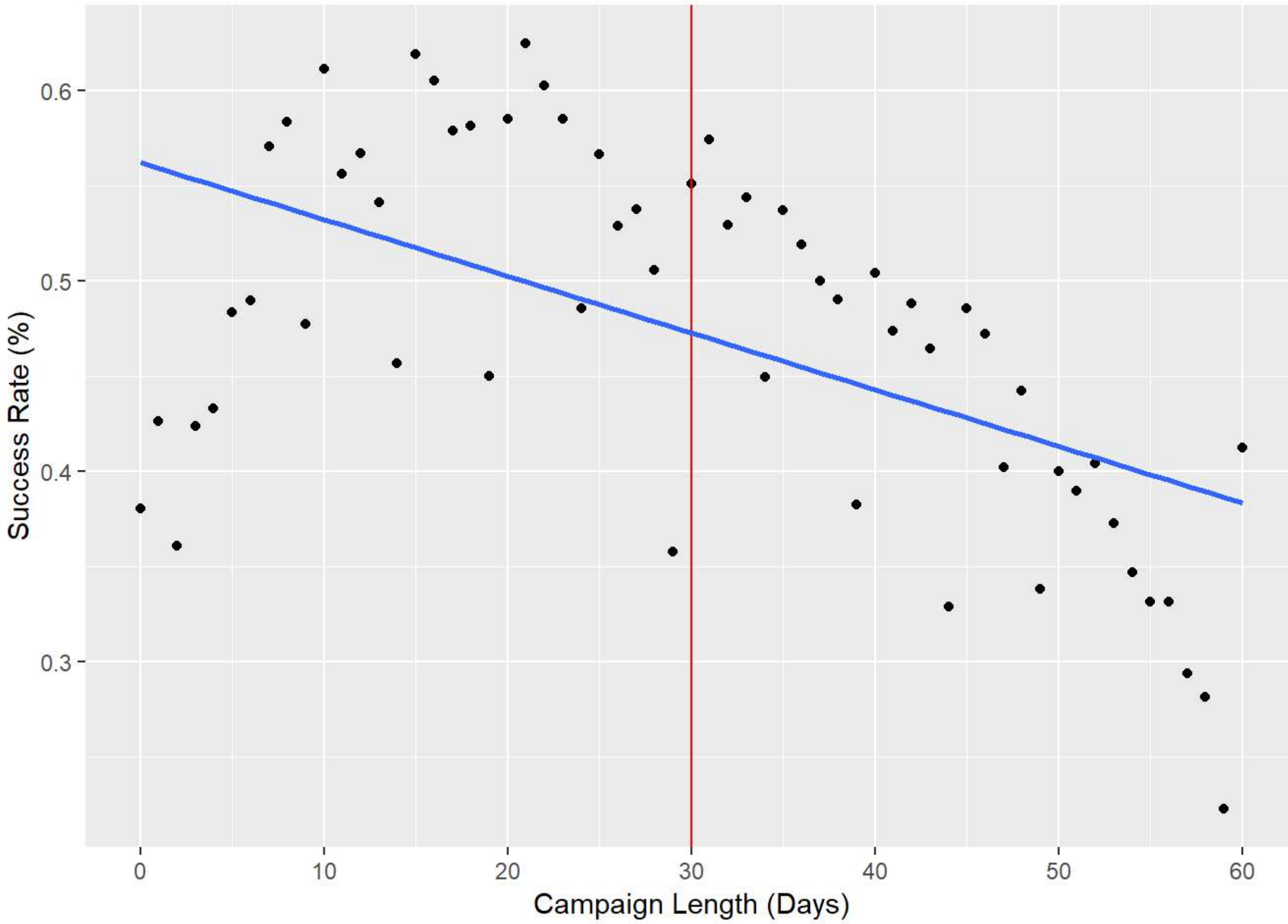


We can see now that Design and Games overtake Technology as raising the most past their initial goal in terms of percentage over.

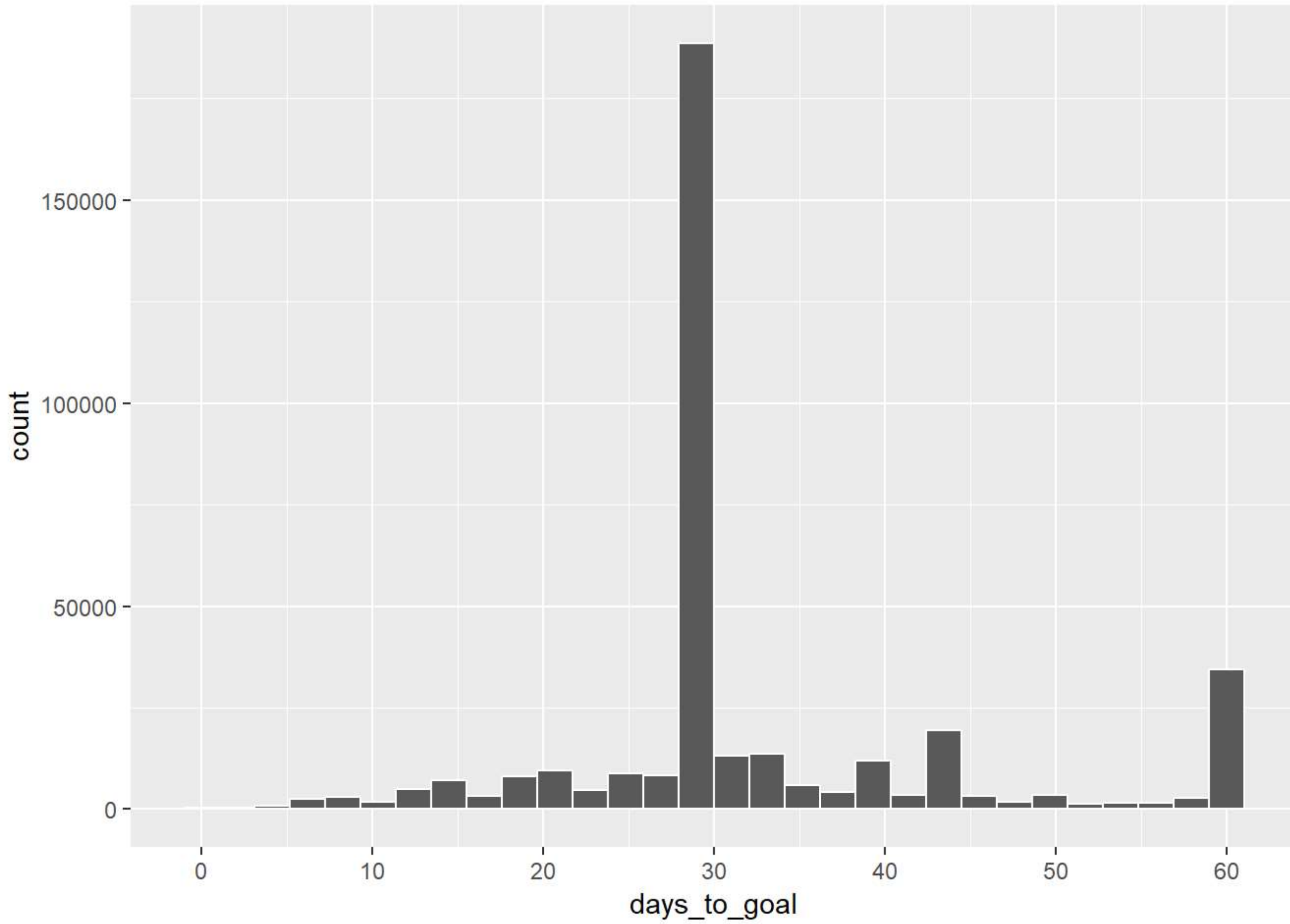
4. What is the correlation between the amount of time given to meet a goal and its success?

Just a note that kickstarter sets a maximum amount of time for a goal to 60 days, and recommends a little less than 30 days, let's see if the data supports that recommendation.

```
ks_days <- ks_data_cleaned %>%
  filter(state %in% c('successful', 'failed'), days_to_goal <= 60) %>%
  group_by(days_to_goal, state) %>%
  dplyr::summarise(count=n()) %>%
  mutate(pct=count/sum(count))

ggplot(ks_days[ks_days$state=='successful',], aes(days_to_goal, pct)) +
  geom_point() +
  labs(x='Campaign Length (Days)', y='Success Rate (%)') +
  scale_x_continuous(breaks=c(0,10,20,30,40,50,60)) +
  geom_vline(xintercept=30, col='red') +
  geom_smooth(method = 'lm', se = FALSE)
```



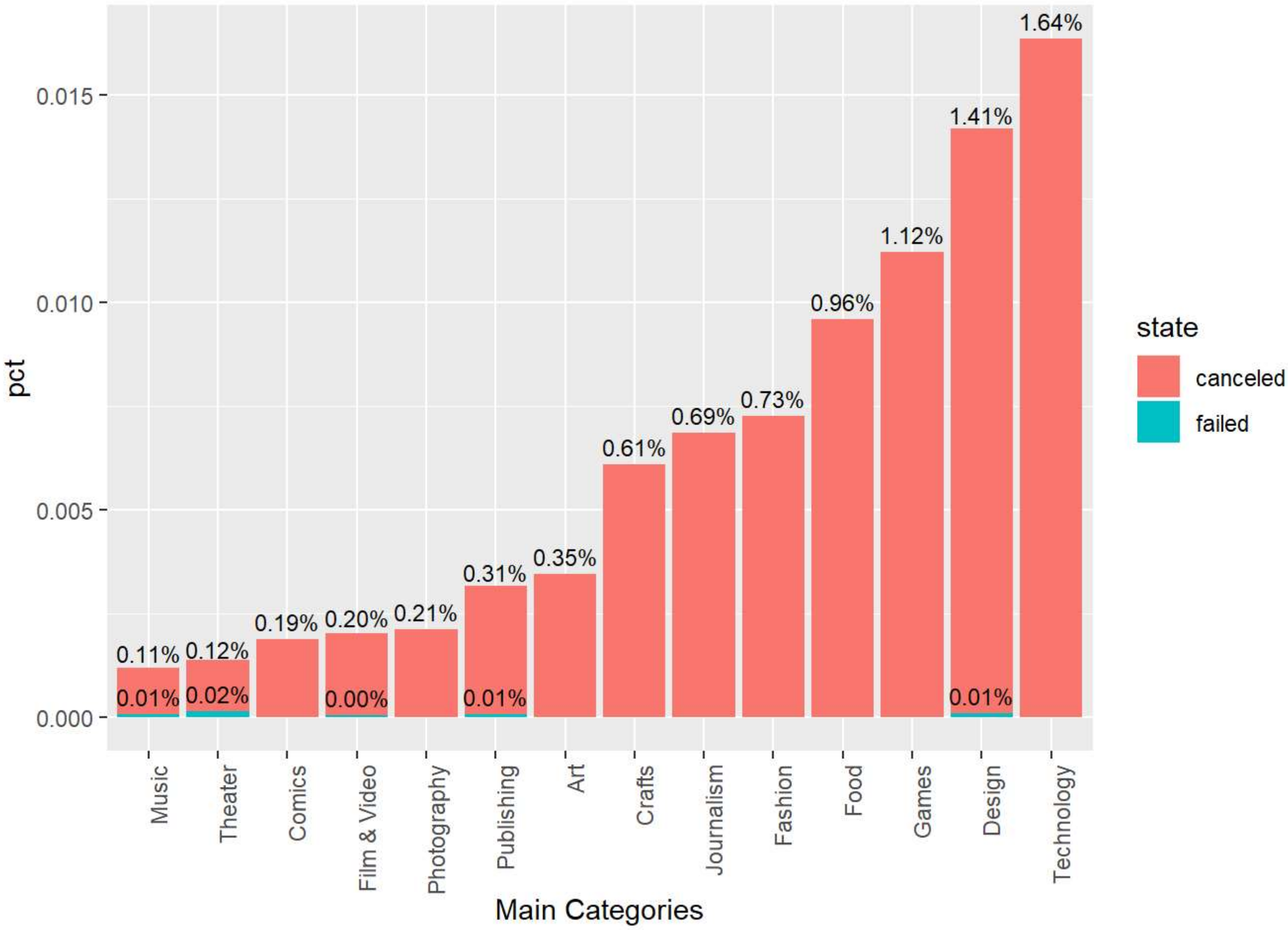

```
ggplot(ks_data_cleaned[ks_data_cleaned$days_to_goal <= 60,], aes(x=days_to_goal)) +
  geom_histogram(col = 'white') +
  scale_x_continuous(breaks=c(0,10,20,30,40,50,60))
```



It seems that the recommendation by kickstarter to have a duration less than 30 days is accurate as we can those to the left 30 days have a greater chance than those to the left and this is further reinforced by the trend line. It also seems that a lot of people have listened to kickstarter and set their time to 29 days (most picked time), but the popularity of this number has brought this amount of to lower than any other day less than 30. The optimal amount seems to be between 7 and 25 days.

5. Which Kickstarter campaigns have the lowest chance to fail after their goal is met?

```
ks_success <- ks_data_cleaned %>%
  filter(state %in% c('successful', 'failed', 'canceled'), pledged >= goal) %>%
  group_by(main_category, state) %>%
  dplyr::summarise(count=n()) %>%
  mutate(pct=count/sum(count)) %>%
  arrange(desc(state), pct)

ggplot(ks_success[ks_success$state != 'successful',], aes(x = reorder(main_category, pct, sum),
 y = pct, fill = state)) +
  geom_col() +
  labs(x = 'Main Categories') +
  geom_text(aes(label = percent(pct), vjust = -0.5), size = 3) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

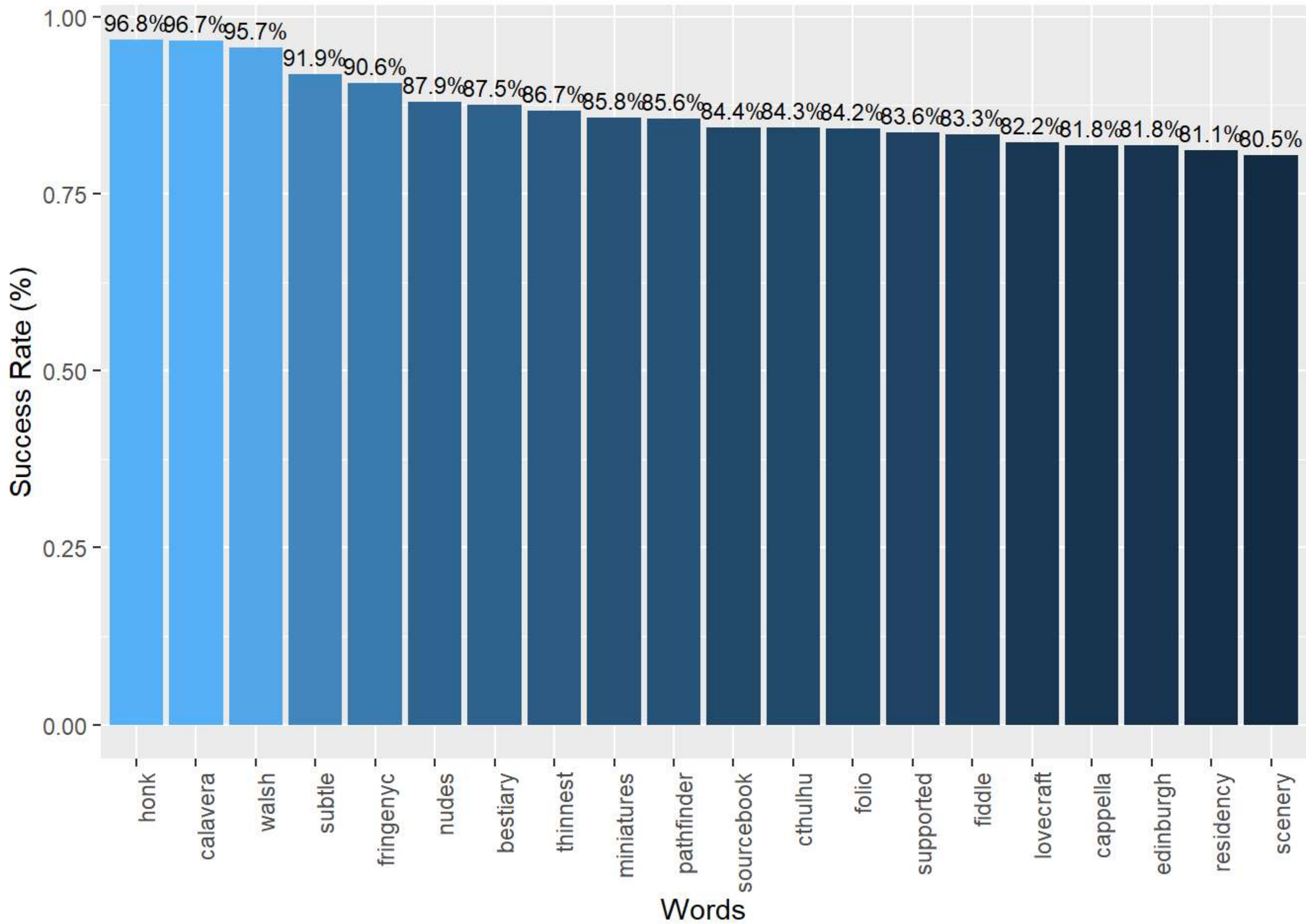


Of the projects that made their initial goal we can see that Technology, Design and Games lead in the highest chance to fail. Frankly these percents seem a little low, so I wonder if Kickstarter accurately tracks projects that do not deliver what they promised. However, at face value, Music, Theater and Comics have the best chance of success once they meet their goals.

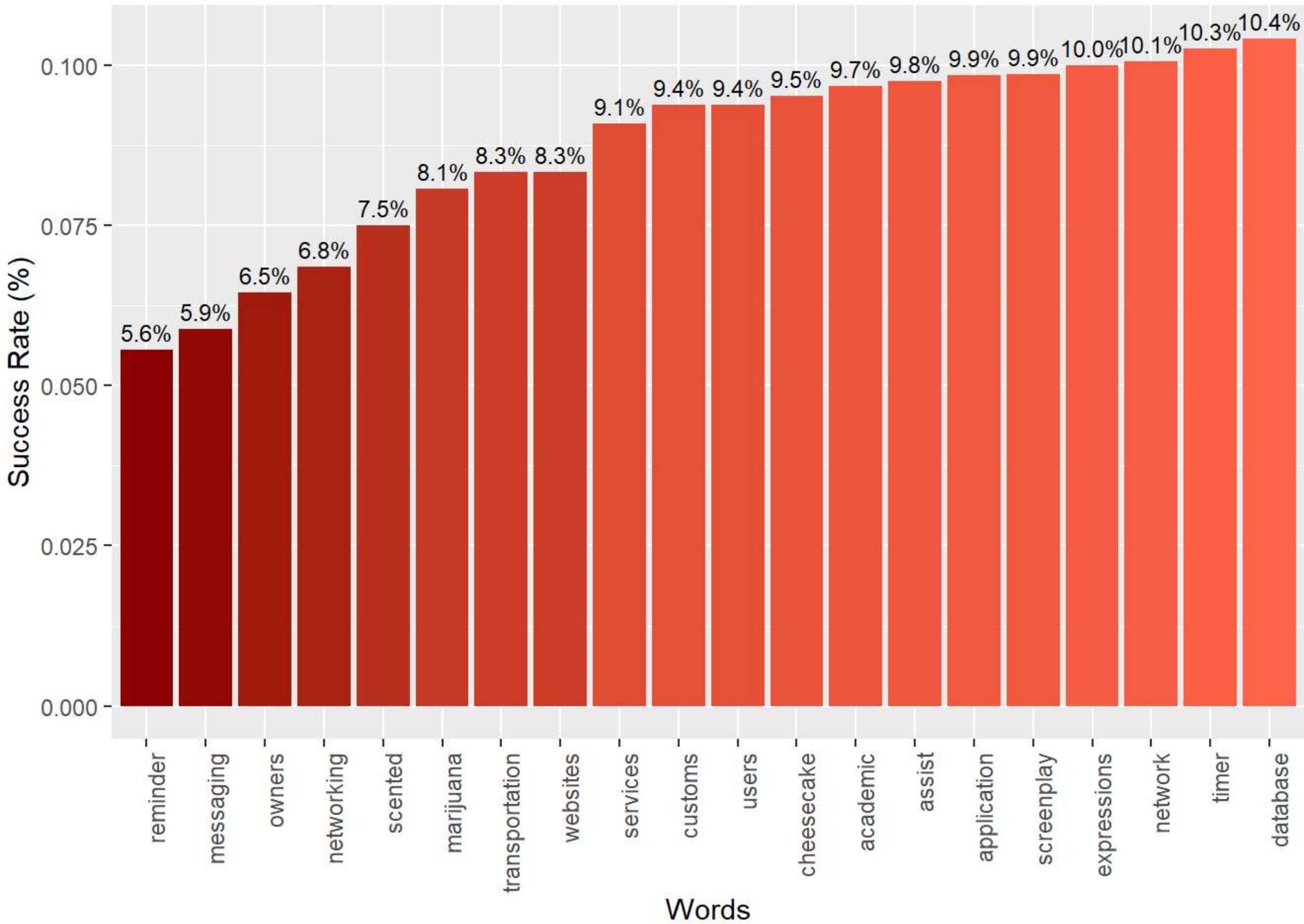
6. Which words have the highest correlation with success and which ones have the lowest?

```r
ks_tokens <- ks_data_cleaned %>%
  filter(state %in% c('successful', 'failed')) %>%
  select(state, main_category, name) %>%
  unnest_tokens(word, name) %>%
  anti_join(stop_words)

ks_tokens_success <- ks_tokens %>%
  filter(state %in% c('successful')) %>%
  dplyr::count(word, sort = TRUE)

colnames(ks_tokens_success)[2] <- 'n_success'

ks_tokens_failed <- ks_tokens %>%
  filter(state %in% c('failed')) %>%
  dplyr::count(word, sort = TRUE)

colnames(ks_tokens_failed)[2] <- 'n_failed'

freq <- ks_tokens_success %>%
  full_join(ks_tokens_failed) %>%
  mutate(word = str_extract(word, "[a-z']+"),
         n_total = n_success + n_failed,
         n_success_pct = n_success / n_total,
         n_success_wgt = n_success_pct * n_success,
         n_lean = n_success - n_failed) %>%
  filter(nchar(word) > 3) %>%
  na.omit()
```

```r
n_appear = 30 # Minimum number of times a word must show up to be counted
n_num = 20 # Number of words on the graph

top_n(freq[freq$n_total >= n_appear,], n=n_num, n_success_pct) %>%
  ggplot(., aes(x = reorder(word, -n_success_pct, sum), y = n_success_pct, fill = n_success_pc
t)) +
  geom_col() +
  labs(x = 'Words', y = 'Success Rate (%)') +
  geom_text(aes(label = percent1(n_success_pct), vjust = -0.5), size = 3) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1), legend.position='none')
```




```
top_n(freq[freq$n_total >= n_appear,], n=-n_num, n_success_pct) %>%
  ggplot(., aes(x = reorder(word, n_success_pct, sum), y = n_success_pct, fill = n_success_pct))
+
  geom_col() +
  labs(x = 'Words', y = 'Success Rate (%)') +
  scale_fill_gradient(low="darkred",high="tomato") +
  geom_text(aes(label = percent1(n_success_pct), vjust = -0.5), size = 3) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1), legend.position='none')
```



For the word frequency analysis, I tokenized all the names of kickstarter campaigns, unnested them and filtered out stop words. However I also removed characters that were not letters and this left some words that were really short so I also decided the word had to be greater than 3 letters to count. There were a lot of really high and low percentages for both categories that had very low usage rates so I made an arbitrary decision to filter the list to at least 30 total appearances. Feel free to play around with this number, it produces some interesting results.

7. Binomial Logistic Regression Model

```
set.seed(25)

ks_binary <- ks_data_cleaned %>%
  filter(state %in% c('successful', 'failed')) %>%
  mutate(state_binary = as.numeric(as.character(revalue(state, c('successful'=1,'failed'=0)))),
         pledge = usd_pledged_real,
         goal = usd_goal_real) %>%
  select(state_binary, goal, days_to_goal)


train_index = createDataPartition(ks_binary$state_binary, p = .8, list = F)
train = ks_binary[train_index,]
test = ks_binary[-train_index,]

model <- glm(state_binary ~., family = "binomial", data = train)
summary(model)
```

```
##
## Call:
## glm(formula = state_binary ~ ., family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3319  -1.0731  -0.8332   1.2437   8.4904
##
## Coefficients:
##                   Estimate    Std. Error z value            Pr(>|z|)
## (Intercept)   0.3562923181  0.0115582039   30.83 <0.0000000000000002 ***
## goal         -0.0000160198  0.0000002276  -70.39 <0.0000000000000002 ***
## days_to_goal -0.0159594443  0.0003348440  -47.66 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 358066  on 265339  degrees of freedom
## Residual deviance: 344823  on 265337  degrees of freedom
## AIC: 344829
##
## Number of Fisher Scoring iterations: 9
```

```
pred <- predict(model, newdata = test, type = "response")
pred.fit <- ifelse(pred > 0.5, 1, 0)
misClasificError <- mean(pred.fit != test$state_binary)
print(paste('Accuracy', percent(1 - misClasificError)))
```

```
## [1] "Accuracy 61.0%"
```

Conclusion

I wanted to explore the data from two different perspectives, a backer and a creator, and see if we could pull out meaningful analysis for both.

- From a backer perspective I want to see if I'm going to invest my money, which projects are the safest and it seems like Dance, Theater and Comics are the safest bet.
- Not only did they have the highest success rates, but they also have the lowest cancellation rates. While Technology, Design and Journalism seem to the be the riskiest.

- From a creator standpoint we have a few things to help.
- First we see that Games, Design and Technology get the most backers and are also the most likely to go over the initial goal so stretch goals are very important
- Though Games backers don't pay out as much as the other two categories so goals should be lower.
- Even though those categories get a lot more backers, we've already seen that those are categories that don't see as much success.
- The number of days to set our goal to see the best chance of success would be 10, 15 or 21 days.
- We've also seen words that have done really well, such as Cthulhu and Calaveras. We've also seen words that haven't such as reminder, messaging and networking

- Finally, we have a binomial logistic regression model which is showing 61% accuracy on predicting success just by using `goal` amount and `days_to_goal` , which is certainly better than chance.

Limitations

- Finally I want to end with some limitations to the analysis. Some of the analysis were broken out by category and some weren't, however there is a finer level of detail and that is the sub_category group.
- A fully flushed out EDA would explore all these nuanced differences because there might be a lot of variability within each category. This is especially true of the most successful words as many of them are probably only successful in certain categories.
- A next step analysis would include dialing in what is the most successful goal amount per category as this is one of the controllable variables for success.