kickstarter_class_tuncay

April 28, 2019

0.0.1 KICKSTARTER PROJECT FROM KAGGLE

https://www.kaggle.com/kemical/kickstarter-projects/version/7

0.0.2 BY TUNCAY DOGAN

https://github.com/ted2020

0.0.3 FOR ECON 328 - FINAL PROJECT

```
In [1]: Sys.time()
```

```
[1] "2019-04-28 19:31:13 PDT"
```

1 INTRODUCTION

Being able to reach out to clients or investors is one of the many great advantages of technology. Kickstarter enables makers to put forward their ideas and hopefully attain the attention of investors, who are on the lookout for the next big thing. Easiness of the platform, along with its early entrant market advantage, make Kickstarter a unique environment. There are many projects here, relateable to almost every aspect of life. But the changing preferences of people, with their desire of differentiated products can be observed in the data. Also, some sub-categories and categories are the leaders in terms of backers and percentage of fundedness. I will try to break it down as much as possible.

I am, here, looking at the issue from an angle of excess funding, timestamp variables, backers, and title wordcloud.

Kickstarter is a crowdsourcing to encourage creative projects for those who don't have much financial means. It's one of the go-to places for venture capitalists and angel investors. Even individuals can take part in projects with small capitals. The platform includes variety of projects from biotech to painting. If one puts forward a project for funding, that person is called a "creator." Each creator has to go through an immense task of providing a thorough analysis and objectives of the project. Outline of a project should indicate the stage, steps to be taken, possible final outcome, use of it, provide all the links and files, so that the investors can make an informed decision. If an investor decides to pledge an amount, he/she is called a "backer" and

the amount contributed is named as "pledged amount." If the total asked funding by the creator has not been achieved, the pledged amounts of investors are not collected and the project doesn't go through. So, it's an all or none model. Therefore, creators should put strong creativity, research, and sincerity into their projects before they decide to go on to the platform of Kickstarter.

2 SUMMARY

First, I will explore the data. I will create new variables that I am going to use, and see the levels, percentages by the sub and main category. Then I will look at which country has the most Kickstarter projects and which currency is the most used in transacting the business. Additonally, I will find which projects have higher than average funding and how many days the projects stay open.

Second, I will visualize excess pledge by sub and main categories, and what month and day provides the unusual fund pledging.

Third, I will create a wordcloud to analyze the names of projects.

Fourth, I will try to predict whether a project will be successful or fail. (logistic regression and randomforest and wordcloud)

```
In [2]: library(tidyverse)
        library(caret)
        library(janitor) # adorn_percentages
        library(ggplot2)
        library(wordcloud)
        library(tidytext)
        library(stringr)
        library(text2vec)
        library(tm)
        library(NLP)
        library(psych)
        library(randomForest)
        #library(party)
        library(car)
        library(InformationValue)
        library(heuristica)
```

```
Warning message:
"package 'tidyverse' was built under R version 3.5.2"-- Attaching packages -------------------
v ggplot2 3.1.1       v purrr   0.3.2
v tibble  2.1.1       v dplyr   0.8.0.1
v tidyr   0.8.3       v stringr 1.4.0
v readr   1.3.1       v forcats 0.4.0
Warning message:
"package 'ggplot2' was built under R version 3.5.3"Warning message:
```

```
"package 'tibble' was built under R version 3.5.3"Warning message:
"package 'tidyr' was built under R version 3.5.3"Warning message:
"package 'readr' was built under R version 3.5.2"Warning message:
"package 'purrr' was built under R version 3.5.3"Warning message:
"package 'dplyr' was built under R version 3.5.3"Warning message:
"package 'stringr' was built under R version 3.5.2"Warning message:
"package 'forcats' was built under R version 3.5.2"-- Conflicts ------------------------------
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
Warning message:
"package 'caret' was built under R version 3.5.3"Loading required package: lattice
Warning message:
"package 'lattice' was built under R version 3.5.2"
Attaching package: 'caret'

The following object is masked from 'package:purrr':

    lift

Warning message:
"package 'janitor' was built under R version 3.5.3"
Attaching package: 'janitor'

The following objects are masked from 'package:stats':

    chisq.test, fisher.test

Warning message:
"package 'wordcloud' was built under R version 3.5.3"Loading required package: RColorBrewer
Warning message:
"package 'RColorBrewer' was built under R version 3.5.2"Warning message:
"package 'text2vec' was built under R version 3.5.3"Loading required package: NLP

Attaching package: 'NLP'

The following object is masked from 'package:ggplot2':

    annotate

Warning message:
"package 'psych' was built under R version 3.5.2"
Attaching package: 'psych'

The following objects are masked from 'package:ggplot2':

    %+%, alpha

Warning message:
```

```
"package 'randomForest' was built under R version 3.5.2"randomForest 4.6-14
Type rfNews() to see new features/changes/bug fixes.

Attaching package: 'randomForest'

The following object is masked from 'package:psych':

    outlier

The following object is masked from 'package:dplyr':

    combine

The following object is masked from 'package:ggplot2':

    margin

Warning message:
"package 'car' was built under R version 3.5.3"Loading required package: carData
Warning message:
"package 'carData' was built under R version 3.5.2"
Attaching package: 'car'

The following object is masked from 'package:psych':

    logit

The following object is masked from 'package:dplyr':

    recode

The following object is masked from 'package:purrr':

    some

Warning message:
"package 'InformationValue' was built under R version 3.5.3"
Attaching package: 'InformationValue'

The following objects are masked from 'package:caret':

    confusionMatrix, precision, sensitivity, specificity

Warning message:
"package 'heuristica' was built under R version 3.5.3"
```

3 EXPLORATORY

```
In [3]: kickstarter <- read.csv("ks-projects-201801.csv")

In [4]: # additional variables that are used are created here

        kickstarter <- kickstarter %>% mutate(launch_year = as.numeric(as.character(format(as.I
        kickstarter <- kickstarter %>% mutate(launch_month = format(as.Date(launched), "%B"))
        kickstarter <- kickstarter %>% mutate(launch_month_numeric = as.numeric(as.character(fo
        kickstarter <- kickstarter %>% mutate(launch_weekday = format(as.Date(launched), "%A"))
        kickstarter <- kickstarter %>% mutate(launch_weekday_numeric = as.numeric(as.factor(ki
        kickstarter <- kickstarter %>% mutate(excess_p=ifelse(((usd_pledged_real/usd_goal_real)
        kickstarter <- kickstarter %>% mutate(totaldays =  as.numeric(as.Date(deadline) - as.Da
        kickstarter <- kickstarter %>% mutate(category_numeric = as.numeric(as.factor(kickstart
        kickstarter <- kickstarter %>% mutate(main_category_numeric = as.numeric(as.factor(kick
        kickstarter <- kickstarter %>% mutate(state_numeric = as.numeric(as.factor(kickstarter$

        head(kickstarter,1)
```

ID	name	category	main_category	currency	deadline	goal
1000002330	The Songs of Adelaide & Abullah	Poetry	Publishing	GBP	2015-10-09	1000

```
In [5]: str(kickstarter)

'data.frame':        378661 obs. of  25 variables:
 $ ID                    : int  1000002330 1000003930 1000004038 1000007540 1000011046 1000014
 $ name                  : Factor w/ 375765 levels "","\177Not Twins - New EP! \"The View from
 $ category              : Factor w/ 159 levels "3D Printing",..: 109 94 94 91 56 124 59 42 11
 $ 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
 $ deadline              : Factor w/ 3164 levels "2009-05-03","2009-05-16",..: 2288 3042 1333 
 $ 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 804
 $ 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 2
 $ 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 ...
 $ launch_year           : num  2015 2017 2013 2012 2015 ...
 $ launch_month          : chr  "August" "September" "January" "March" ...
 $ launch_month_numeric  : num  8 9 1 3 7 2 12 2 4 7 ...
 $ launch_weekday        : chr  "Tuesday" "Saturday" "Saturday" "Saturday" ...
 $ launch_weekday_numeric: num  6 3 3 3 3 1 2 2 5 1 ...
 $ excess_p              : num  0 0 0 0 0 ...
 $ totaldays             : num  59 60 45 30 56 35 20 45 35 30 ...
 $ category_numeric      : num  109 94 94 91 56 124 59 42 114 40 ...
 $ main_category_numeric : num  13 7 7 11 7 8 8 8 5 7 ...
```

```
 $ state_numeric          : num  2 2 2 2 1 4 4 2 1 1 ...
```

```
In [6]: anyNA(kickstarter)
```

TRUE

```
In [7]: kickstarter[!complete.cases(kickstarter),]
        # looks like missing values are for the usd.pledged column, which i wont use in this a
        # so i can ignore them
```

	ID	name	categor
170	1000694855	STREETFIGHTERZ WHEELIE MURICA	Film &
329	100149523	Duncan Woods - Chameleon EP	Music
633	1003023003	The Making of Ashley Kelley's Debut Album	Music
648	1003130892	Butter Side Down Debut Album	Music
750	1003629045	Chase Goehring debut EP	Music
825	1004013077	Spencer Capier Instrumental Project 2015	Music
845	1004126342	LUKAS LIGETI'S 50TH BIRTHDAY FESTIVAL: ORIGINAL NEW MUSIC!	Music
865	1004255433	The Battle For Breukelen: A Neighborhood Epic	Film &
871	1004298993	"Tamboura Plays Violin" - a collection of Pop & classical!	Music
891	1004402863	Nightingale Noel - An A Cappella Holiday CD	Music
1027	1005185256	Local Music Connection	Music
1037	100522240	DO NOT DUPLICATE - Selected by 2015 Devon Film Commision	Film &
1117	1005653464	SA-4 Studios	Music
1251	1006327667	Letters to the Wild - EP	Music
1685	1008671527	The Fitness Mindset :-)	Publish
1686	1008675685	Inspired & the Sleep EP - Eyelid Kid	Music
1719	1008815806	Into Winter: An Instrumental Christmas Experiment	Music
1785	1009154868	Help advance scientific guitar & bass design	Music
1818	1009312309	Mr. Pickles	Publish
1893	1009774693	Universal Peace	Universal Justice	Music
1962	1010217550	Solarboot-Havel-Tour	Music
2054	1010694326	Surrender	Music
2202	1011508551	EBENEZER Solo Record	Music
2305	1011952280	NHSOB Summer Festival 2015	Music
2327	1012058337	Degobah Max	Music
2396	1012414277	WE ARE WHO WE ARE BUT WHATEVER WE WANT TO BE!!!	Music
2452	1012704793	Life In The Bus Lane	Film &
2458	1012744036	An Oratorio for our Time - Last Stop Cafe	Music
2479	1012807897	Juliet Remembered	Film &
2607	1013389538	Make admission to the London Music Awards ceremony free!	Music
...	...	...	...
375348	982862687	Bydlo - A Novel	Publish
375360	982919439	Faith	Film &
375513	983728691	The Octopodes' 5th Studio Album!	Music
375766	985231591	TWO FISH	Film &
375981	986253191	DEVELOP MAGAZINE FOR THE ARTS, MUSIC, AND DESIGN LOVERS	Publish
376026	986471034	Help Revival finish their new album, Guidance.	Music
376188	98744618	Vinyl for Dancing Pigeons!	Music
376471	988763792	Maud Sings Maud: A Musical Betsy-Tacy Companion	Music
376561	989218116	Keep on Movin' - The Debut Album	Music
376604	989499289	Sojourner	Music
376675	989874415	Love and Soul through Sound	Music
376971	991409513	Chris Copeland's First EP	Music
376985	991483230	Support Vocal Synergy's Second Semester and Tennessee Tour!	Music
377044	991820324	The Blend CD 2015	Music
377159	992433603	Jonny Come Lately	Film &
377473	994080962	Maximum Coppage	Film &
377570	994621918	Star Sailor	Film &
377639	994911312	"Songwriter assistance" service	Music
377641	994933706	Whiskey and Holy Water	Music
377873	99613274	Double EP and Music Video Fundraiser	Music
378012	996908566	The World Needs More Big Thinkers	Publish

```
In [8]: levels(kickstarter$state)
        unique(kickstarter$launch_weekday_numeric)
        unique(kickstarter$launch_weekday)
```

1. 'canceled' 2. 'failed' 3. 'live' 4. 'successful' 5. 'suspended' 6. 'undefined'
1. 6 2. 3 3. 1 4. 2 5. 5 6. 7 7. 4
1. 'Tuesday' 2. 'Saturday' 3. 'Friday' 4. 'Monday' 5. 'Thursday' 6. 'Wednesday' 7. 'Sunday'

```
In [9]: colnames(kickstarter)
```

1. 'ID' 2. 'name' 3. 'category' 4. 'main_category' 5. 'currency' 6. 'deadline' 7. 'goal' 8. 'launched' 9. 'pledged' 10. 'state' 11. 'backers' 12. 'country' 13. 'usd.pledged' 14. 'usd_pledged_real' 15. 'usd_goal_real' 16. 'launch_year' 17. 'launch_month' 18. 'launch_month_numeric' 19. 'launch_weekday' 20. 'launch_weekday_numeric' 21. 'excess_p' 22. 'totaldays' 23. 'category_numeric' 24. 'main_category_numeric' 25. 'state_numeric'

```
In [10]: levels(kickstarter$category)
         n_distinct(kickstarter$category)
```

1. '3D Printing' 2. 'Academic' 3. 'Accessories' 4. 'Action' 5. 'Animals' 6. 'Animation' 7. 'Anthologies' 8. 'Apparel' 9. 'Apps' 10. 'Architecture' 11. 'Art' 12. 'Art Books' 13. 'Audio' 14. 'Bacon' 15. 'Blues' 16. 'Calendars' 17. 'Camera Equipment' 18. 'Candles' 19. 'Ceramics' 20. 'Children\'s Books' 21. 'Childrenswear' 22. 'Chiptune' 23. 'Civic Design' 24. 'Classical Music' 25. 'Comedy' 26. 'Comic Books' 27. 'Comics' 28. 'Community Gardens' 29. 'Conceptual Art' 30. 'Cookbooks' 31. 'Country & Folk' 32. 'Couture' 33. 'Crafts' 34. 'Crochet' 35. 'Dance' 36. 'Design' 37. 'Digital Art' 38. 'DIY' 39. 'DIY Electronics' 40. 'Documentary' 41. 'Drama' 42. 'Drinks' 43. 'Electronic Music' 44. 'Embroidery' 45. 'Events' 46. 'Experimental' 47. 'Fabrication Tools' 48. 'Faith' 49. 'Family' 50. 'Fantasy' 51. 'Farmer\'s Markets' 52. 'Farms' 53. 'Fashion' 54. 'Festivals' 55. 'Fiction' 56. 'Film & Video' 57. 'Fine Art' 58. 'Flight' 59. 'Food' 60. 'Food Trucks' 61. 'Footwear' 62. 'Gadgets' 63. 'Games' 64. 'Gaming Hardware' 65. 'Glass' 66. 'Graphic Design' 67. 'Graphic Novels' 68. 'Hardware' 69. 'Hip-Hop' 70. 'Horror' 71. 'Illustration' 72. 'Immersive' 73. 'Indie Rock' 74. 'Installations' 75. 'Interactive Design' 76. 'Jazz' 77. 'Jewelry' 78. 'Journalism' 79. 'Kids' 80. 'Knitting' 81. 'Latin' 82. 'Letterpress' 83. 'Literary Journals' 84. 'Literary Spaces' 85. 'Live Games' 86. 'Makerspaces' 87. 'Metal' 88. 'Mixed Media' 89. 'Mobile Games' 90. 'Movie Theaters' 91. 'Music' 92. 'Music Videos' 93. 'Musical' 94. 'Narrative Film' 95. 'Nature' 96. 'Nonfiction' 97. 'Painting' 98. 'People' 99. 'Performance Art' 100. 'Performances' 101. 'Periodicals' 102. 'Pet Fashion' 103. 'Photo' 104. 'Photobooks' 105. 'Photography' 106. 'Places' 107. 'Playing Cards' 108. 'Plays' 109. 'Poetry' 110. 'Pop' 111. 'Pottery' 112. 'Print' 113. 'Printing' 114. 'Product Design' 115. 'Public Art' 116. 'Publishing' 117. 'Punk' 118. 'Puzzles' 119. 'Quilts' 120. 'R&B' 121. 'Radio & Podcasts' 122. 'Ready–to-wear' 123. 'Residencies' 124. 'Restaurants' 125. 'Robots' 126. 'Rock' 127. 'Romance' 128. 'Science Fiction' 129. 'Sculpture' 130. 'Shorts' 131. 'Small Batch' 132. 'Software' 133. 'Sound' 134. 'Space Exploration' 135. 'Spaces' 136. 'Stationery' 137. 'Tabletop Games' 138. 'Taxidermy' 139. 'Technology' 140. 'Television' 141. 'Textiles' 142. 'Theater' 143. 'Thrillers' 144. 'Translations' 145. 'Typography' 146. 'Vegan' 147. 'Video' 148. 'Video Art' 149. 'Video Games' 150. 'Wearables' 151. 'Weaving' 152. 'Web' 153. 'Webcomics' 154. 'Webseries' 155. 'Woodworking' 156. 'Workshops' 157. 'World Music' 158. 'Young Adult' 159. 'Zines'

159

```
In [11]: levels(kickstarter$main_category)
         n_distinct(kickstarter$main_category)
```

1. 'Art' 2. 'Comics' 3. 'Crafts' 4. 'Dance' 5. 'Design' 6. 'Fashion' 7. 'Film & Video' 8. 'Food' 9. 'Games' 10. 'Journalism' 11. 'Music' 12. 'Photography' 13. 'Publishing' 14. 'Technology' 15. 'Theater'

15

```
In [12]: levels(kickstarter$currency)
         n_distinct(kickstarter$currency)
```

1. 'AUD' 2. 'CAD' 3. 'CHF' 4. 'DKK' 5. 'EUR' 6. 'GBP' 7. 'HKD' 8. 'JPY' 9. 'MXN' 10. 'NOK' 11. 'NZD' 12. 'SEK' 13. 'SGD' 14. 'USD'

14

```
In [13]: df1 <- (kickstarter %>% group_by(main_category,f_s=state_numeric==4) %>%
         summarize(successful_count = n()))

         df2 <- (kickstarter %>% group_by(main_category,f_s=state_numeric==2) %>%
         summarize(failed_count = n()))

         df3 <- (kickstarter %>% group_by(main_category,f_s=state_numeric==1) %>%
         summarize(canceled_count = n()))

         df4 <- (kickstarter %>% group_by(main_category,f_s=state_numeric==3) %>%
         summarize(live_count = n()))

         df5 <- (kickstarter %>% group_by(main_category,f_s=state_numeric==5) %>%
         summarize(suspended_count = n()))

         df6 <- (kickstarter %>% group_by(main_category,f_s=state_numeric==6) %>%
         summarize(undefined_count = n()))
```

```
In [14]: merged <- Reduce(function(x, y) left_join(x, y, by=c("main_category","f_s"), all=TRUE
         merged <- merged %>% mutate_if(is.integer, ~replace(., is.na(.), 0))
         merged   <- merged[seq(2,nrow(merged),2),]
         merged
```

```
`mutate_if()` ignored the following grouping variables:
Column `main_category`
```

main_category	f_s	successful_count	failed_count	canceled_count	live_count	suspended_coun
Art	TRUE	11510	14131	2222	194	96
Comics	TRUE	5842	4036	842	76	23
Crafts	TRUE	2115	5703	843	76	72
Dance	TRUE	2338	1235	163	18	13
Design	TRUE	10550	14814	4152	305	247
Fashion	TRUE	5593	14182	2650	250	138
Film & Video	TRUE	23623	32904	5755	332	117
Food	TRUE	6085	15969	2211	184	153
Games	TRUE	12518	16003	6202	287	220
Journalism	TRUE	1012	3137	523	31	52
Music	TRUE	24197	21752	3305	281	149
Photography	TRUE	3305	6384	986	48	55
Publishing	TRUE	12300	23145	3602	299	66
Technology	TRUE	6434	20616	4715	377	424
Theater	TRUE	6534	3708	608	41	21

In [15]:
```
# sum(merged$successful_count[seq(2,30,2)]) # sums only TRUE values
# sum(merged$successful_count[seq(1,30,2)])
#sum(merged[,3:8])
#sum(merged$successful_count)
# sum(merged$failed_count)+
# sum(merged$canceled_count)+
# sum(merged$live_count)+
# sum(merged$suspended_count)+
# sum(merged$undefined_count)
# dim(kickstarter)
# sum(merged[1,3:8])
# sum(merged[2,3:8])
# sum(kickstarter$fail_success==4)
```

In [16]:
```
merged_p <- merged %>% adorn_percentages()
merged_p %>% arrange(desc(successful_count))
```

main_category	f_s	successful_count	failed_count	canceled_count	live_count	suspended_cou
Dance	TRUE	0.6204883	0.3277601	0.04325902	0.004777070	0.003450106
Theater	TRUE	0.5987355	0.3397782	0.05571337	0.003756987	0.001924310
Comics	TRUE	0.5399760	0.3730474	0.07782605	0.007024679	0.002125890
Music	TRUE	0.4660619	0.4189684	0.06365808	0.005412381	0.002869910
Art	TRUE	0.4088374	0.5019359	0.07892587	0.006890917	0.003409939
Film & Video	TRUE	0.3715184	0.5174805	0.09050877	0.005221357	0.001840057
Games	TRUE	0.3553121	0.4542306	0.17603815	0.008146235	0.006244501
Design	TRUE	0.3508480	0.4926505	0.13807782	0.010143000	0.008214167
Publishing	TRUE	0.3084717	0.5804534	0.09033455	0.007498621	0.001655214
Photography	TRUE	0.3066147	0.5922627	0.09147416	0.004453103	0.005102514
Food	TRUE	0.2473376	0.6490936	0.08987074	0.007479067	0.006219007
Fashion	TRUE	0.2451350	0.6215813	0.11614656	0.010957223	0.006048387
Crafts	TRUE	0.2400954	0.6474061	0.09569758	0.008627540	0.008173459
Journalism	TRUE	0.2128286	0.6597266	0.10998948	0.006519453	0.010935857
Technology	TRUE	0.1975498	0.6329946	0.14476957	0.011575424	0.013018515

```
In [17]: merged_p <- merged %>% adorn_percentages()
         merged_p %>% arrange(desc(failed_count))
```

main_category	f_s	successful_count	failed_count	canceled_count	live_count	suspended_cou
Journalism	TRUE	0.2128286	0.6597266	0.10998948	0.006519453	0.010935857
Food	TRUE	0.2473376	0.6490936	0.08987074	0.007479067	0.006219007
Crafts	TRUE	0.2400954	0.6474061	0.09569758	0.008627540	0.008173459
Technology	TRUE	0.1975498	0.6329946	0.14476957	0.011575424	0.013018515
Fashion	TRUE	0.2451350	0.6215813	0.11614656	0.010957223	0.006048387
Photography	TRUE	0.3066147	0.5922627	0.09147416	0.004453103	0.005102514
Publishing	TRUE	0.3084717	0.5804534	0.09033455	0.007498621	0.001655214
Film & Video	TRUE	0.3715184	0.5174805	0.09050877	0.005221357	0.001840057
Art	TRUE	0.4088374	0.5019359	0.07892587	0.006890917	0.003409939
Design	TRUE	0.3508480	0.4926505	0.13807782	0.010143000	0.008214167
Games	TRUE	0.3553121	0.4542306	0.17603815	0.008146235	0.006244501
Music	TRUE	0.4660619	0.4189684	0.06365808	0.005412381	0.002869910
Comics	TRUE	0.5399760	0.3730474	0.07782605	0.007024679	0.002125890
Theater	TRUE	0.5987355	0.3397782	0.05571337	0.003756987	0.001924310
Dance	TRUE	0.6204883	0.3277601	0.04325902	0.004777070	0.003450106

```
In [18]: df7 <- (kickstarter %>% group_by(category,f_s=state_numeric==4) %>%
         summarize(successful_count = n()))

         df8 <- (kickstarter %>% group_by(category,f_s=state_numeric==2) %>%
         summarize(failed_count = n()))

         df9 <- (kickstarter %>% group_by(category,f_s=state_numeric==1) %>%
         summarize(canceled_count = n()))

         df10 <- (kickstarter %>% group_by(category,f_s=state_numeric==3) %>%
         summarize(live_count = n()))
```

```
df11 <- (kickstarter %>% group_by(category,f_s=state_numeric==5) %>%
summarize(suspended_count = n()))

df12 <- (kickstarter %>% group_by(category,f_s=state_numeric==6) %>%
summarize(undefined_count = n()))
```

In [19]:
```
merged2 <- Reduce(function(x, y) left_join(x, y, by=c("category","f_s"), all=TRUE), l
merged2 <- merged2 %>% mutate_if(is.integer, ~replace(., is.na(.), 0))
merged2   <- merged2[seq(2,nrow(merged2),2),]
head(merged2)
```

```
`mutate_if()` ignored the following grouping variables:
Column `category`
```

category	f_s	successful_count	failed_count	canceled_count	live_count	suspended_count
3D Printing	TRUE	242	326	91	8	15
Academic	TRUE	188	589	115	11	13
Accessories	TRUE	1073	1667	340	53	29
Action	TRUE	107	514	109	7	3
Animals	TRUE	63	166	18	3	5
Animation	TRUE	682	1531	306	16	6

In [20]:
```
merged2_p <- merged2 %>% adorn_percentages()
head(merged2_p %>% arrange(desc(successful_count)),10)
```

category	f_s	successful_count	failed_count	canceled_count	live_count	suspended_cc
Chiptune	TRUE	0.7714286	0.1714286	0.05714286	0.000000000	0.0000000000
Residencies	TRUE	0.7246377	0.2608696	0.01449275	0.000000000	0.0000000000
Anthologies	TRUE	0.6645408	0.2755102	0.04719388	0.011479592	0.0012755102
Dance	TRUE	0.6640827	0.2911283	0.04134367	0.002153316	0.0008613264
Indie Rock	TRUE	0.6395616	0.3024571	0.05479936	0.002651582	0.0005303164
Letterpress	TRUE	0.6326531	0.3061224	0.02040816	0.020408163	0.0204081633
Country & Folk	TRUE	0.6317681	0.3147607	0.04830375	0.004268704	0.0008986745
Classical Music	TRUE	0.6303100	0.3034826	0.06008419	0.004209721	0.0019135094
Theater	TRUE	0.6242029	0.3229418	0.04987955	0.001133626	0.0017004393
Performances	TRUE	0.6159921	0.3326752	0.03849951	0.006910168	0.0059230010

In [21]:
```
merged2_p <- merged2 %>% adorn_percentages()
head(merged2_p %>% arrange(desc(failed_count)),10)
```

category	f_s	successful_count	failed_count	canceled_count	live_count	suspended_
Video	TRUE	0.11915888	0.7803738	0.07943925	0.004672897	0.016355140
Food Trucks	TRUE	0.12385845	0.7745434	0.08675799	0.011415525	0.003424658
Apps	TRUE	0.05957447	0.7736801	0.15003940	0.012450749	0.004255319
Candles	TRUE	0.12820513	0.7529138	0.11421911	0.004662005	0.000000000
Web	TRUE	0.08596934	0.7502426	0.14477004	0.010867456	0.008150592
Farmer's Markets	TRUE	0.16509434	0.7405660	0.08018868	0.009433962	0.004716981
Restaurants	TRUE	0.16211422	0.7314651	0.09116708	0.010996807	0.004256829
Hip-Hop	TRUE	0.15388548	0.7303170	0.10071575	0.007924335	0.006901840
Software	TRUE	0.12171916	0.7224409	0.14140420	0.009842520	0.003937008
Mobile Games	TRUE	0.08552264	0.7210732	0.17272219	0.011738401	0.008943544

In [22]:
```
# countries that do kickstarter projects the most to the least
# united states, great britain, canada, and australia lead the way in crowd funding.
# weirdly, these are all english spoken countries.
(kickstarter %>% group_by(country) %>%
summarize(n=n())) %>% arrange(desc(n))
```

country	n
US	292627
GB	33672
CA	14756
AU	7839
DE	4171
N,0"	3797
FR	2939
IT	2878
NL	2868
ES	2276
SE	1757
MX	1752
NZ	1447
DK	1113
IE	811
CH	761
NO	708
HK	618
BE	617
AT	597
SG	555
LU	62
JP	40

In [23]:
```
# most preferred currencies
# dollar, euro, and british pound are the most used currencies in crowd funding.
(kickstarter %>% group_by(currency) %>%
summarize(n=n())) %>% arrange(desc(n))
```

currency	n
USD	295365
GBP	34132
EUR	17405
CAD	14962
AUD	7950
SEK	1788
MXN	1752
NZD	1475
DKK	1129
CHF	768
NOK	722
HKD	618
SGD	555
JPY	40

In [24]:
```
by_excess_p <- (kickstarter %>% group_by(main_category,excess_p>0)) %>%
summarize(by_group_excess_p=sum(excess_p/n()))

by_excess_p    <- by_excess_p[seq(2,nrow(by_excess_p),2),]
by_excess_p
```

main_category	excess_p > 0	by_group_excess_p
Art	TRUE	5.1131469
Comics	TRUE	10.8190684
Crafts	TRUE	9.0844371
Dance	TRUE	0.2500935
Design	TRUE	4.5595111
Fashion	TRUE	2.8380101
Film & Video	TRUE	2.9961448
Food	TRUE	2.1918508
Games	TRUE	19.5581550
Journalism	TRUE	1.2200308
Music	TRUE	15.3790834
Photography	TRUE	0.8393530
Publishing	TRUE	4.8649677
Technology	TRUE	13.2552497
Theater	TRUE	0.6508495

In [25]:
```
(kickstarter %>% group_by(main_category,f_s=state=="successful")) %>%
summarize(avgdays=sum(totaldays/n()))
```

main_category	f_s	avgdays
Art	FALSE	35.07018
Art	TRUE	30.34005
Comics	FALSE	36.79586
Comics	TRUE	31.90859
Crafts	FALSE	32.39991
Crafts	TRUE	29.02695
Dance	FALSE	35.20699
Dance	TRUE	31.81480
Design	FALSE	35.80072
Design	TRUE	33.46777
Fashion	FALSE	33.35905
Fashion	TRUE	31.30914
Film & Video	FALSE	37.78149
Film & Video	TRUE	32.31004
Food	FALSE	34.92191
Food	TRUE	31.51159
Games	FALSE	34.00916
Games	TRUE	29.91277
Journalism	FALSE	35.26262
Journalism	TRUE	32.03656
Music	FALSE	37.28632
Music	TRUE	33.87759
Photography	FALSE	34.61413
Photography	TRUE	32.13737
Publishing	FALSE	35.47976
Publishing	TRUE	32.01220
Technology	FALSE	35.78217
Technology	TRUE	34.16211
Theater	FALSE	39.95661
Theater	TRUE	31.61570

```
In [26]: head((kickstarter %>% group_by(category, f_s=state=="successful")) %>%
         summarize(avgdays=sum(totaldays/n())))
```

category	f_s	avgdays
3D Printing	FALSE	34.43764
3D Printing	TRUE	32.50826
Academic	FALSE	35.66071
Academic	TRUE	31.76064
Accessories	FALSE	32.06549
Accessories	TRUE	30.31873

almost exclusively, average days the project is open for funding differ for succesful and failed projects. Failed projects tend to stay open longer than the successful ones.

it looks like the projects owners, to fill the pledge gap, keep their project open for investment longer than their counterparts which are pledged the required amount.

In [27]:
```
describe(kickstarter$backers)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew
X1	1	378661	105.6175	907.185	12	28.83679	17.7912	0	219382	219382	86.76232

In [28]:
```
# 1 mad away from mean backers distribution by main_category
# mad = median absolute deviation
backed <- (kickstarter %>% group_by(backers,main_category) %>%
summarize(n=n()))

head((backed %>% arrange(desc(backers>(mean(kickstarter$backers)+mad(kickstarter$backe
```

backers	main_category	n
124	Art	22
124	Comics	18
124	Crafts	2
124	Dance	5
124	Design	37
124	Fashion	13
124	Film & Video	58
124	Food	17
124	Games	34
124	Journalism	4

In [29]:
```
# 1 mad away from mean backers distribution by sub_category
backed <- (kickstarter %>% group_by(backers,category) %>%
summarize(n=n()))

head((backed %>% arrange(desc(backers>(mean(kickstarter$backers)+mad(kickstarter$backe
```

backers	category	n
124	Action	1
124	Animation	2
124	Anthologies	1
124	Apparel	4
124	Art	10
124	Art Books	7
124	Camera Equipment	1
124	Ceramics	1
124	Children's Books	10
124	Childrenswear	2

by looking backers distribution and its 1 mad (median absolute deviation) from mean, i can say that art, dance, design, books are leading projects.

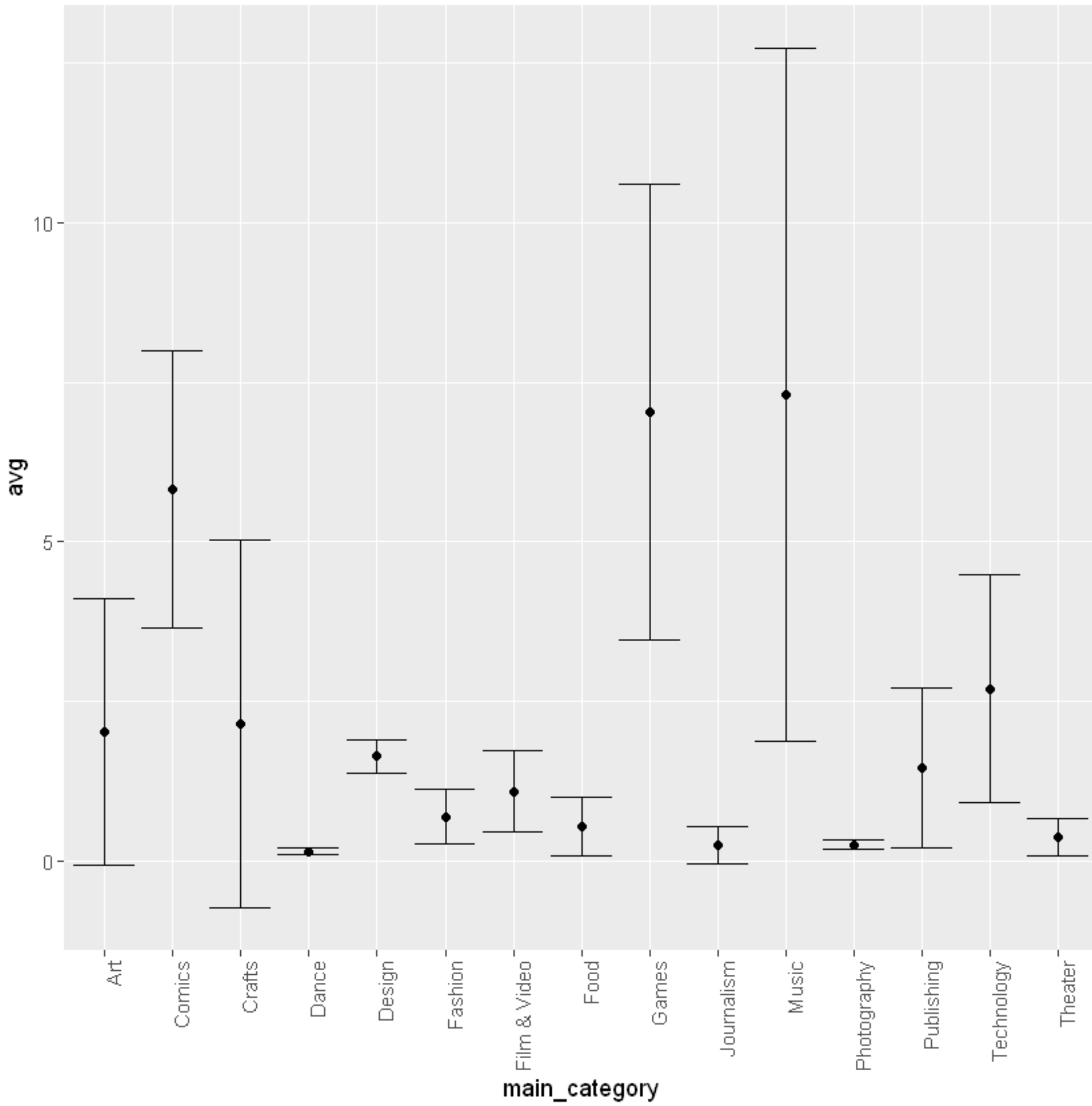
this is no suprise since their main and sub categories are found to be providing higher than average excess funding.

higher than average excess funding requires higher than average number of backers, and this proves it.

4 VISUALIZATION

```
In [30]: kickstarter %>% group_by(main_category) %>%
                 summarize(n = n(), avg = mean(excess_p), se = sd(excess_p)/sqrt(n())) %>%
                 filter(n > 1) %>%
                 mutate(reorder(main_category, avg)) %>%
                 ggplot(aes(x = main_category, y = avg, ymin = avg - 2*se, ymax = avg + 2*se))
                 geom_point() +
                 geom_errorbar() +
                 theme(axis.text.x = element_text(angle = 90, hjust = 1))
                 ggtitle("excess pledge by main category")
```

```
$title
[1] "excess pledge by main category"

attr(,"class")
[1] "labels"
```



games, music, and technology, and comics main categories seem to be the ones with the highest excess funding.

excess funding refers to the funds donated more than the project requires.

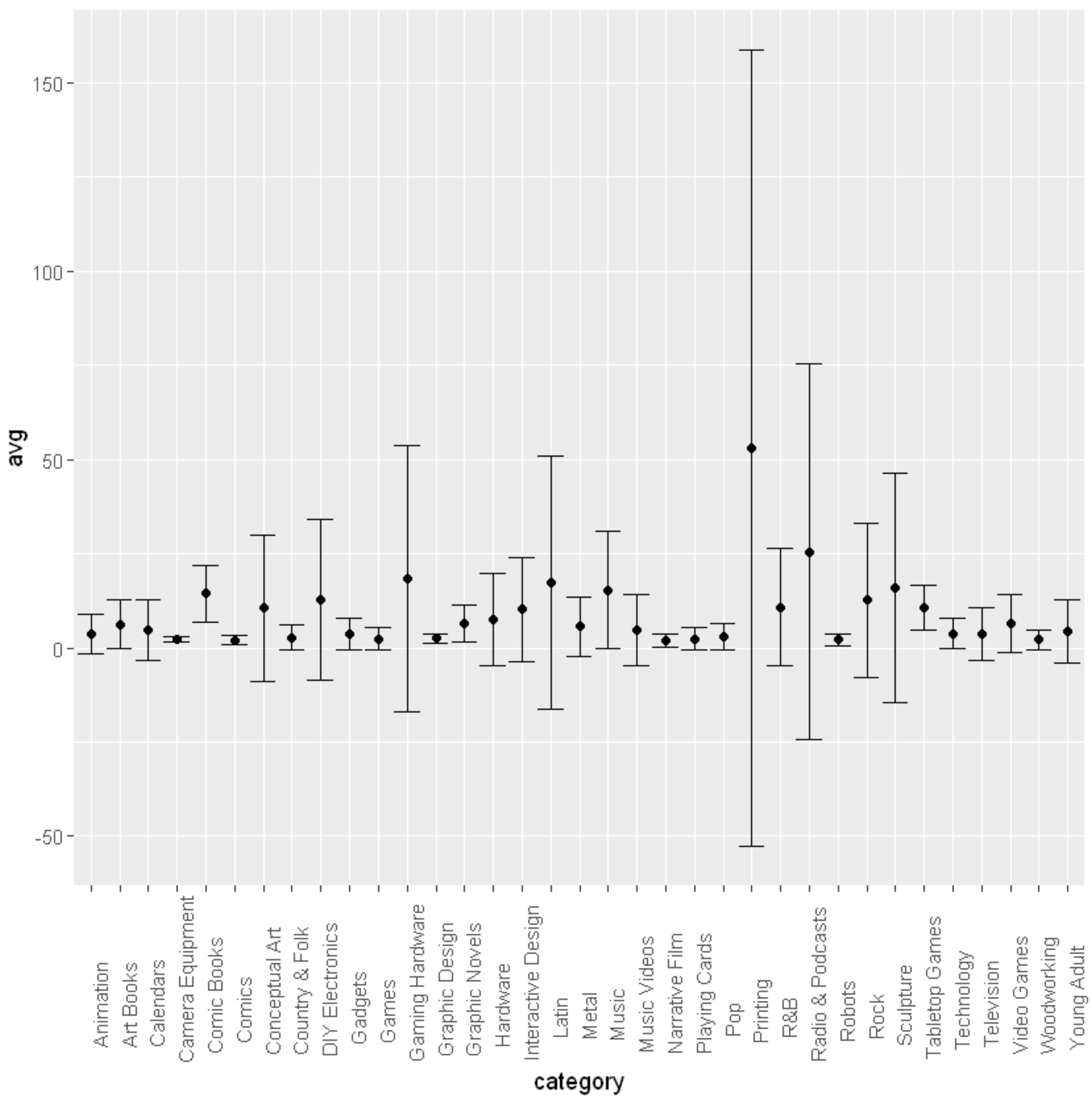
for example, if you launch a gaming category kickstarter project, funds being donated, on average, across many projects, is likely to be 20 times more than asked for.

```
In [31]: by_excess_p2 <- (kickstarter %>% group_by(category,excess_p>0)) %>%
         summarize(by_group_excess_p2=sum(excess_p/n()))

         by_excess_p2   <- by_excess_p2[seq(2,nrow(by_excess_p2),2),]
         head(by_excess_p2 %>% arrange(desc(by_group_excess_p2)))
```

category	excess_p > 0	by_group_excess_p2
Printing	TRUE	252.39184
Gaming Hardware	TRUE	76.16636
Latin	TRUE	65.79265
Radio & Podcasts	TRUE	63.16044
Interactive Design	TRUE	51.87512
R&B	TRUE	46.96202

```
In [32]: kickstarter %>% group_by(category) %>%
             summarize(n = n(), avg = mean(excess_p), se = sd(excess_p)/sqrt(n())) %>%
             filter(avg > 2) %>%
             mutate(reorder(category, avg)) %>%
             ggplot(aes(x = category, y = avg, ymin = avg - 2*se, ymax = avg + 2*se)) +
             geom_point() +
             geom_errorbar() +
             theme(axis.text.x = element_text(angle = 90, hjust = 0.2))
             ggtitle("excess pledge by sub-category, filtered by average excess 2-fold")
```

```
$title
[1] "excess pledge by sub-category, filtered by average excess 2-fold"

attr(,"class")
[1] "labels"
```



in sub-category, i see that printing, gaming, latin, design, and music leads the way in excess funding.

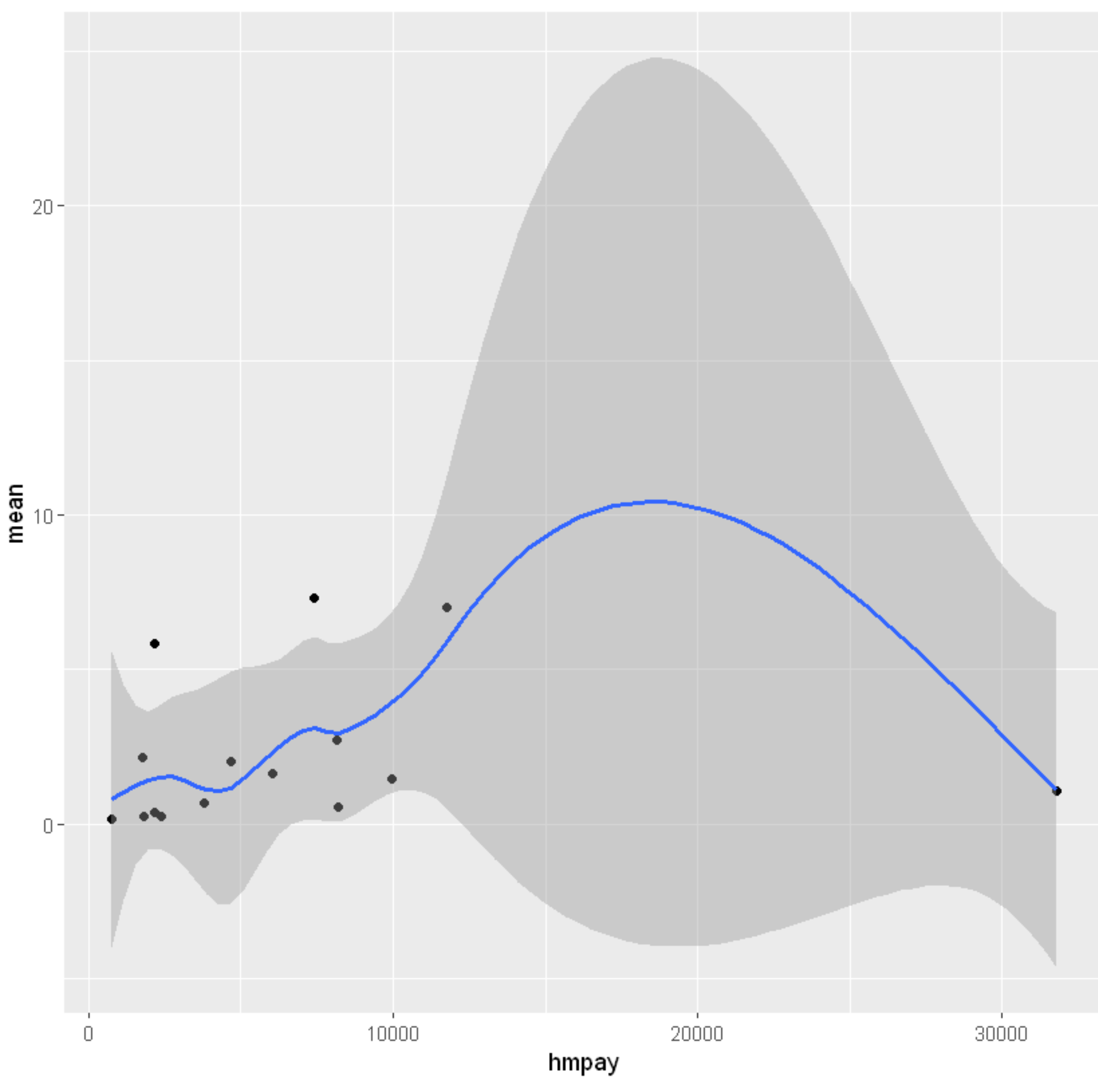
i am assuming that this list includes printing due to high demand for 3D printing material and its vast applicability from bio-tech and arms industry.

https://www.researchnester.com/reports/us-3d-printing-market-analysis-opportunity-outlook-2024/88

another interesting excess funding here is from radio&podcasts. I looked it up and most of them require small amounts. That may be the reason why it's one of the highly excessly funded project topics

```
In [33]: kickstarter %>%
         group_by(main_category) %>%
         summarize(n=n(),howold=2019-first(launch_year),mean=mean(excess_p)) %>%
         mutate (hmpay=n/howold)%>% #hmray=how many projects a year
         ggplot(aes(hmpay, mean)) +
         geom_point() +
         geom_smooth() +
         ggtitle("")
```

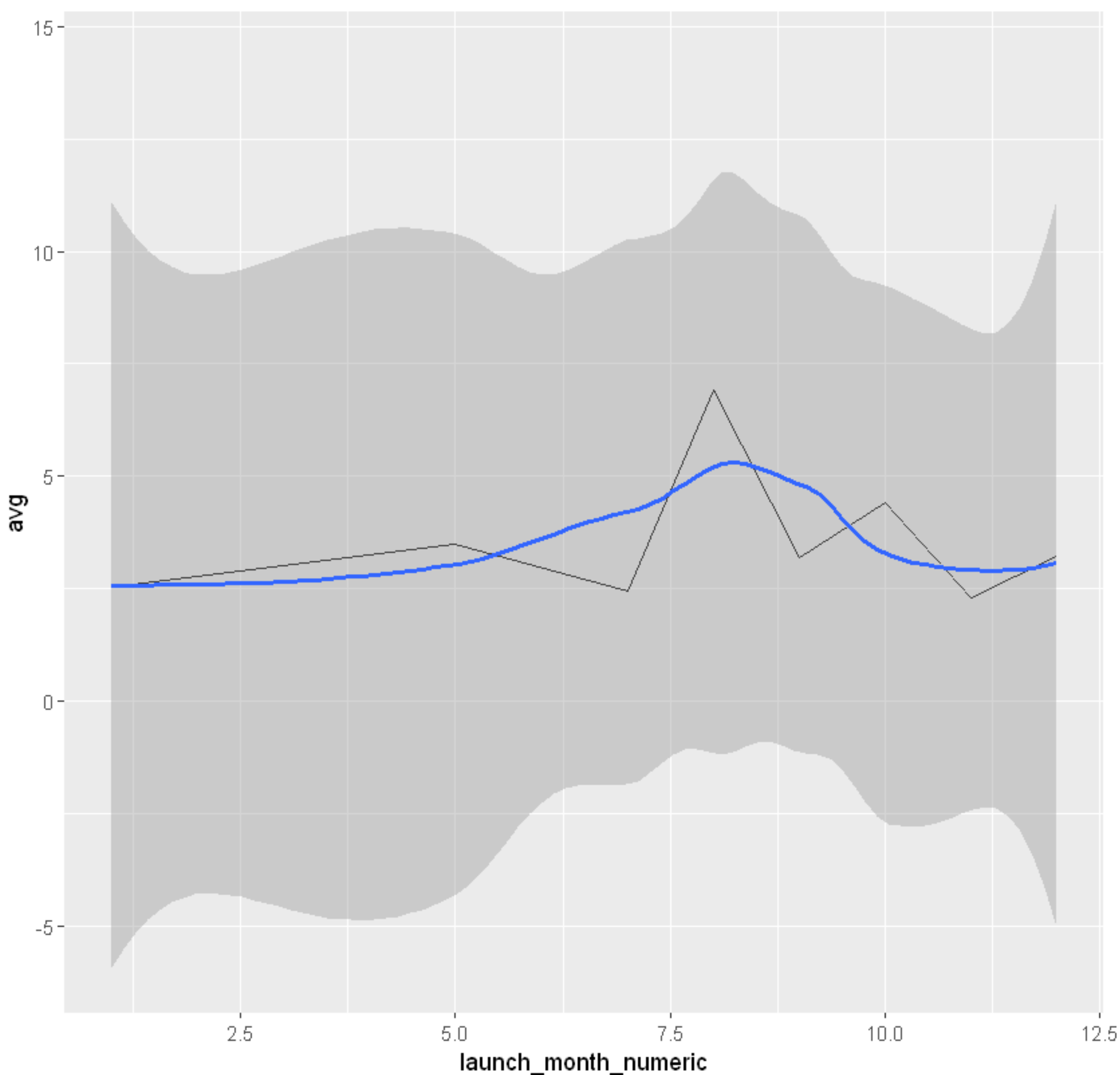
```
`geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
In [34]: kickstarter %>% group_by(launch_month_numeric) %>%
                 summarize(n = n(), avg = mean(excess_p), se = sd(excess_p)/sqrt(n())) %>%
```

```
                filter(avg > 2) %>%
                mutate(reorder(launch_month_numeric, avg)) %>%
                ggplot(aes(x = launch_month_numeric, y = avg, ymin = avg - 2*se, ymax = avg +
                geom_line() +
                geom_smooth() +
                ggtitle("")
```

```
`geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

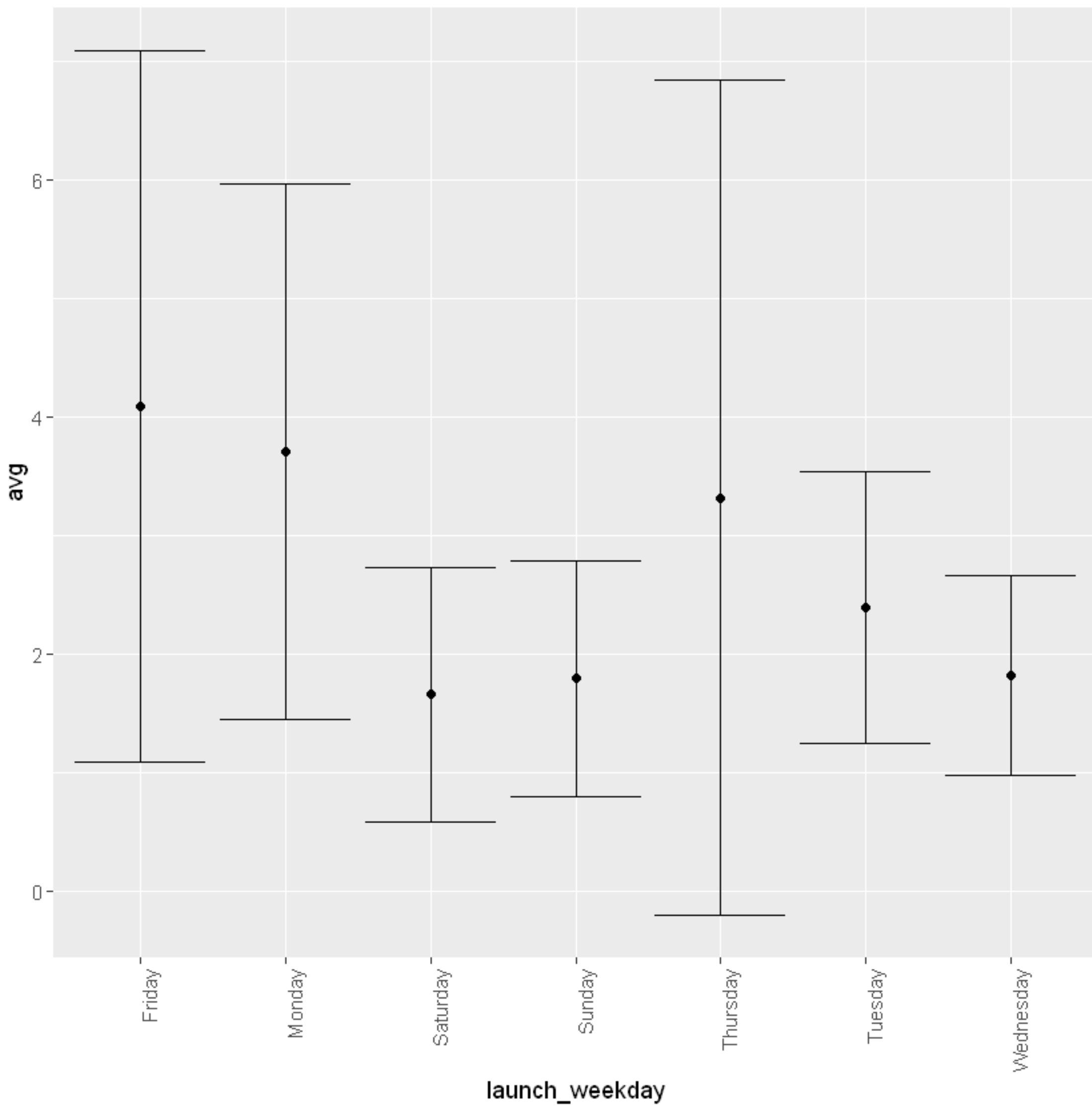


it looks like there is a little bump in funding in august and september.

```
In [35]: kickstarter %>% group_by(launch_weekday) %>%
                summarize(n = n(), avg = mean(excess_p), se = sd(excess_p)/sqrt(n())) %>%
```

```
filter(avg > 0) %>%
mutate(reorder(launch_weekday, avg)) %>%
ggplot(aes(x = launch_weekday, y = avg, ymin = avg - 2*se, ymax = avg + 2*se)
geom_point() +
geom_errorbar() +
theme(axis.text.x = element_text(angle = 90, hjust = 1))
ggtitle("")
```

```
$title
[1] ""

attr(,"class")
[1] "labels"
```



friday, monday, and thurday looks like the day with the highest excess funding return

of course, these day and months with high observed returns are based on launch timestamp. There is going to be a lag between the launch and when it gets noticed by the investors

5 WORDCLOUD

In [36]:
```
names_of_projects <- (kickstarter$name)
names_of_projects <- tolower(names_of_projects)
names_of_projects <- gsub("[^[A-Za-z]"," ", names_of_projects)
names_of_projects <- gsub("\\b\\w{1,2}\\b", "", names_of_projects)
names_of_projects <- gsub("canceled", "", names_of_projects)
names_of_projects <- gsub("project", "", names_of_projects)

names_of_projects <- str_squish(names_of_projects)
head(names_of_projects)
```

1. 'the songs adelaide abullah' 2. 'greeting from earth zgac arts capsule for' 3. 'where hank' 4. 'toshicapital rekordz needs help complete album' 5. 'community film the art neighborhood filmmaking' 6. 'monarch espresso bar'

In [37]:
```
data("stop_words")
```

In [38]:
```
head(stop_words)
which(stop_words$lexicon == 'snowball')[1:10]
```

word	lexicon
a	SMART
a's	SMART
able	SMART
about	SMART
above	SMART
according	SMART

1. 572 2. 573 3. 574 4. 575 5. 576 6. 577 7. 578 8. 579 9. 580 10. 581

In [39]:
```
stop_words_for_removal <- as.vector((stop_words %>% filter(lexicon=="snowball"))[1])
as.character(stop_words_for_removal)
```

'c("i", "me", "my", "myself", "we", "our", "ours", "ourselves", "you", "your", "yours", "yourself", "yourselves", "he", "him", "his", "himself", "she", "her", "hers", "herself", "it", "its", "itself", "they", "them", "their", "theirs", "themselves", "what", "which", "who", "whom", "this", "that", "these", "those", "am", "is", "are", "was", "were", "be", "been", "being", "have", "has", "had", "having", "do", "does", "did", "doing", "would", "should", "could", "ought", "i\'m", "you\'re", "he\'s", "she\'s", \n"it\'s", "we\'re", "they\'re", "i\'ve", "you\'ve", "we\'ve", "they\'ve", "i\'d", "you\'d", "he\'d", "she\'d", "we\'d", "they\'d", "i\'ll", "you\'ll", "he\'ll", "she\'ll", "we\'ll", "they\'ll", "isn\'t", "aren\'t", "wasn\'t", "weren\'t", "hasn\'t", "haven\'t", "hadn\'t", "doesn\'t", "don\'t", "didn\'t", "won\'t", "wouldn\'t", "shan\'t", "shouldn\'t", "can\'t", "cannot", "couldn\'t", "mustn\'t", "let\'s", "that\'s", "who\'s", "what\'s", "here\'s", "there\'s", "when\'s", "where\'s", "why\'s", "how\'s", "a", "an", "the", "and", "but", "if", "or", \n"because", "as", "until", "while", "of", "at", "by", "for", "with", "about", "against", "between", "into", "through", "during", "before", "after", "above", "below", "to", "from", "up", "down", "in", "out", "on", "off", "over", "under", "again", "further", "then", "once", "here", "there", "when", "where", "why", "how", "all", "any", "both", "each", "few", "more", "most", "other", "some", "such", "no", "nor", "not", "only", "own", "same", "so", "than", "too", "very")'

```
In [40]: names_of_projects_tm <- VCorpus(VectorSource(names_of_projects))
         names_of_projects_tm <- tm_map(names_of_projects_tm,removeWords,stop_words_for_removal
         names_of_projects_tm <- tm_map(names_of_projects_tm,stripWhitespace)
         names_of_projects_tm <- tm_map(names_of_projects_tm,removePunctuation)
         inspect(names_of_projects_tm[[10]])
```

```
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 35

studio sky documentary feature film
```

```
In [41]: NGramTokenizer <- function(x) {
           unlist(lapply(ngrams(words(x), GRAMS), paste, collapse = " "),
           use.names = FALSE)
         }
```

```
In [42]: GRAMS <- 1
         NGramTokenizer(names_of_projects_tm[[1]])
```

1. 'songs' 2. 'adelaide' 3. 'abullah'

```
In [43]: GRAMS <- 2
         NGramTokenizer(names_of_projects_tm[[1]])
```

1. 'songs adelaide' 2. 'adelaide abullah'

```
In [44]: GRAMS <- 1
         names_of_projects_dtm_1 <- DocumentTermMatrix(names_of_projects_tm,control = list(toke
         names_of_projects_dtm_1 <- removeSparseTerms(names_of_projects_dtm_1,0.99)
```

```
In [45]: head(names_of_projects_dtm_1$dimnames$Terms)
         tail(names_of_projects_dtm_1$dimnames$Terms)
```

1. 'album' 2. 'art' 3. 'book' 4. 'debut' 5. 'documentary' 6. 'film'
1. 'new' 2. 'one' 3. 'series' 4. 'short' 5. 'video' 6. 'world'

```
In [46]: GRAMS <- 2
         names_of_projects_dtm_2 <- DocumentTermMatrix(names_of_projects_tm,control = list(toke
         names_of_projects_dtm_2 <- removeSparseTerms(names_of_projects_dtm_2,0.997)
```

```
In [47]: head(names_of_projects_dtm_2$dimnames$Terms)
         tail(names_of_projects_dtm_2$dimnames$Terms)
```

1. 'card game' 2. 'debut album' 3. 'feature film' 4. 'full length' 5. 'music video' 6. 'new album'
1. 'feature film' 2. 'full length' 3. 'music video' 4. 'new album' 5. 'playing cards' 6. 'short film'

```
In [48]: names_of_projects_dtm_freq_1 <- colSums(as.matrix(names_of_projects_dtm_1), na.rm = T)
         names_of_projects_dtm_freq_1 <- sort(names_of_projects_dtm_freq_1,decreasing = T)
         names_of_projects_dtm_freq_1[1:10]
         #barplot(inst2_dtm_freq_1[1:20])
```

album 14239 **new** 13971 **film** 11762 **book** 10048 **game** 9261 **art** 8329 **music** 7685 **first** 6691 **help** 6590 **world** 6553

```
In [49]: names_of_projects_dtm_freq_2 <- colSums(as.matrix(names_of_projects_dtm_2), na.rm = T)
         names_of_projects_dtm_freq_2 <- sort(names_of_projects_dtm_freq_2,decreasing = T)
         names_of_projects_dtm_freq_2[1:10]
         #barplot(inst2_dtm_freq_1[1:20])
```

short film 4103 **debut album** 2641 **new album** 2412 **music video** 1892 **card game** 1748 **feature film** 1634 **playing cards** 1584 **full length** 1452 **9** <NA> **10** <NA>

```
In [50]: colorlist = c("red","blue","green","red","pink","orange","grey","black","brown","navy'
         wordcloud(names(names_of_projects_dtm_freq_1),names_of_projects_dtm_freq_1, random.ord
         wordcloud(names(names_of_projects_dtm_freq_2),names_of_projects_dtm_freq_2, random.ord
```


debut
series
new
world
book
documentary
love
album
one
first
short
film
help
game
art
music
life
video

it's not a coincidence to observe the words "album" and "film" in the wordcloud, since the most succesful categories (see above) are including music, theatre, and dance. It looks like these categories along with the word "album" in their title are becoming more successful in kickstarter platform.

```
In [51]: bing <- sentiments %>% filter(lexicon == "bing") %>% dplyr::select(word,
         sentiment)
```

```
In [52]: head(bing)
         sort(table(bing$sentiment))
```

word	sentiment
2-faced	negative
2-faces	negative
a+	positive
abnormal	negative
abolish	negative
abominable	negative

```
positive negative
    2006     4782
```

```
In [53]: tokens <- colnames(names_of_projects_dtm_1)
         tokenMat <- cbind.data.frame(names_of_projects_dtm_1$i,names_of_projects_dtm_1$v,toke
         colnames(tokenMat) <- c("id","freq","word")
         tokenMat$id <- as.character(tokenMat$id)
         tokenMat$freq <- as.numeric(tokenMat$freq)
         tokenMat$word <- as.character(tokenMat$word)
         head(tokenMat)
         dim(tokenMat)
         word_count <- tapply(tokenMat$freq,tokenMat$id,sum)
         tokenMat$n <- word_count[match(tokenMat$id,names(word_count))]
```

id	freq	word
4	1	album
4	1	help
5	1	art
5	1	film
10	1	documentary
10	1	film

1. 130408 2. 3

```
In [54]: by_id_sentiment_bing <- inner_join(bing,tokenMat,by="word")
         by_id_sentiment_bing.unique <- by_id_sentiment_bing[!duplicated(by_id_sentiment_bing$
         dim(by_id_sentiment_bing.unique)
         head(by_id_sentiment_bing.unique)
```

1. 3958 2. 5

word	sentiment	id	freq	n
love	positive	63	1	1
love	positive	69	1	1
love	positive	121	1	1
love	positive	139	1	2
love	positive	318	1	1
love	positive	344	1	2

```
In [55]: count_sentiment_by_id <- matrix(NA,length(kickstarter$name),length(unique(bing$sentim
         colnames(count_sentiment_by_id) <- unique(bing$sentiment)
```

```
for(i in 1:nrow(count_sentiment_by_id)){
    stp <- by_id_sentiment_bing.unique[by_id_sentiment_bing.unique$id == i,]
    sub_mat <- matrix(0,1,ncol(count_sentiment_by_id))
    colnames(sub_mat) <- colnames(count_sentiment_by_id)

  for(j in 1:nrow(stp)){
    sub_mat[1,(stp$sentiment[j-i])] <- sub_mat[1,(stp$sentiment[j-i])] + stp$freq[j]
  }

  count_sentiment_by_id[i,] <- sub_mat[1,]
}
count_sentiment_by_id<- as.data.frame(count_sentiment_by_id)
head(count_sentiment_by_id)
length(count_sentiment_by_id$negative)
c(negative=sum(count_sentiment_by_id$negative),positive=sum(count_sentiment_by_id$pos
```

negative	positive
0	0
0	0
0	0
0	0
0	0
0	0

378661

negative 0 **positive** 4082

```
In [56]: n <- length(kickstarter$name)

         sentiment_hat <- rbinom(n, 1, 0.70) # 1 for positive, 0 for negative

         sentiment_hat <- as.data.frame(sentiment_hat)
         head(sentiment_hat)

         dim(sentiment_hat)
         sum(sentiment_hat$sentiment_hat)
```

sentiment_hat
1
1
1
1
1
0

1. 378661 2. 1

265032

```
In [57]: classification_bing <- ifelse(count_sentiment_by_id$positive > count_sentiment_by_id$r
         confusion_matrix_bing <- table(sentiment_hat$sentiment_hat,classification_bing)
         confusion_matrix_bing
```

```
##Overall Accuracy of Bing Lexicon vs sentiment_hat
accuracy <- sum(diag(confusion_matrix_bing))/sum(confusion_matrix_bing)
print('--Overall Accuracy--')
accuracy
```

```
   classification_bing
         0      1
  0 112470   1159
  1 262233   2799
```

```
[1] "--Overall Accuracy--"
```

0.304412125885687

6 WILL PROJECT FAIL OR NOT?

```
In [58]: kickstarter2 <- (kickstarter[,c("state_numeric","category_numeric","main_category_nume
                                        "backers","totaldays","launch_month_numeric","launch_
```

```
In [59]: # for logistic regression, make it fail=0, success=1
         # there is 6 states, 4th is success, convert 4 to 1,
         # convert all the other numbers into 0.

         kickstarter2$state_numeric[kickstarter2$state_numeric==1] <- 0
         kickstarter2$state_numeric[kickstarter2$state_numeric==2] <- 0
         kickstarter2$state_numeric[kickstarter2$state_numeric==3] <- 0
         kickstarter2$state_numeric[kickstarter2$state_numeric==4] <- 1
         kickstarter2$state_numeric[kickstarter2$state_numeric==5] <- 0
         kickstarter2$state_numeric[kickstarter2$state_numeric==6] <- 0
```

```
In [60]: head(kickstarter2,1)
```

state_numeric	category_numeric	main_category_numeric	backers	totaldays	launch_month_nume
0	109	13	0	59	8

```
In [61]: str(kickstarter2)
```

```
'data.frame':        378661 obs. of  7 variables:
 $ state_numeric        : num  0 0 0 0 0 1 1 0 0 0 ...
 $ category_numeric     : num  109 94 94 91 56 124 59 42 114 40 ...
 $ main_category_numeric : num  13 7 7 11 7 8 8 8 5 7 ...
 $ backers              : int  0 15 3 1 14 224 16 40 58 43 ...
 $ totaldays            : num  59 60 45 30 56 35 20 45 35 30 ...
 $ launch_month_numeric  : num  8 9 1 3 7 2 12 2 4 7 ...
 $ launch_weekday_numeric: num  6 3 3 3 3 1 2 2 5 1 ...
```

```
In [62]: set.seed(1)
         test_index <- createDataPartition(y = kickstarter2$state_numeric, times = 1, p = 0.1,
         train <- kickstarter2[-test_index,]
         test <- kickstarter2[test_index,]
```

```
In [63]: # LOGISTIC REGRESSION

         # http://r-statistics.co/Logistic-Regression-With-R.html

         logitpredtrain <- glm(state_numeric~category_numeric+backers+totaldays+launch_month_nu
                             train, family=binomial(link="logit"))
```

```
Warning message:
"glm.fit: fitted probabilities numerically 0 or 1 occurred"
```

```
In [64]: predicted <- predict(logitpredtrain, test, type="response")
```

```
In [65]: summary(logitpredtrain)
```

```
Call:
glm(formula = state_numeric ~ category_numeric + backers + totaldays +
    launch_month_numeric + launch_weekday_numeric, family = binomial(link = "logit"),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.4904  -0.7053  -0.6244   0.7053   2.3985

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)            -4.508e-01  1.885e-02 -23.918  < 2e-16 ***
category_numeric       -9.012e-04  9.445e-05  -9.543  < 2e-16 ***
backers                 1.937e-02  9.701e-05 199.706  < 2e-16 ***
totaldays              -2.410e-02  3.643e-04 -66.147  < 2e-16 ***
launch_month_numeric   -2.042e-02  1.276e-03 -15.999  < 2e-16 ***
launch_weekday_numeric -5.409e-03  1.965e-03  -2.753  0.00591 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 442783  on 340793  degrees of freedom
Residual deviance: 330781  on 340788  degrees of freedom
AIC: 330793

Number of Fisher Scoring iterations: 10
```

```
In [66]: AIC(logitpredtrain)
         BIC(logitpredtrain)
```

330792.877621601
330857.31182241

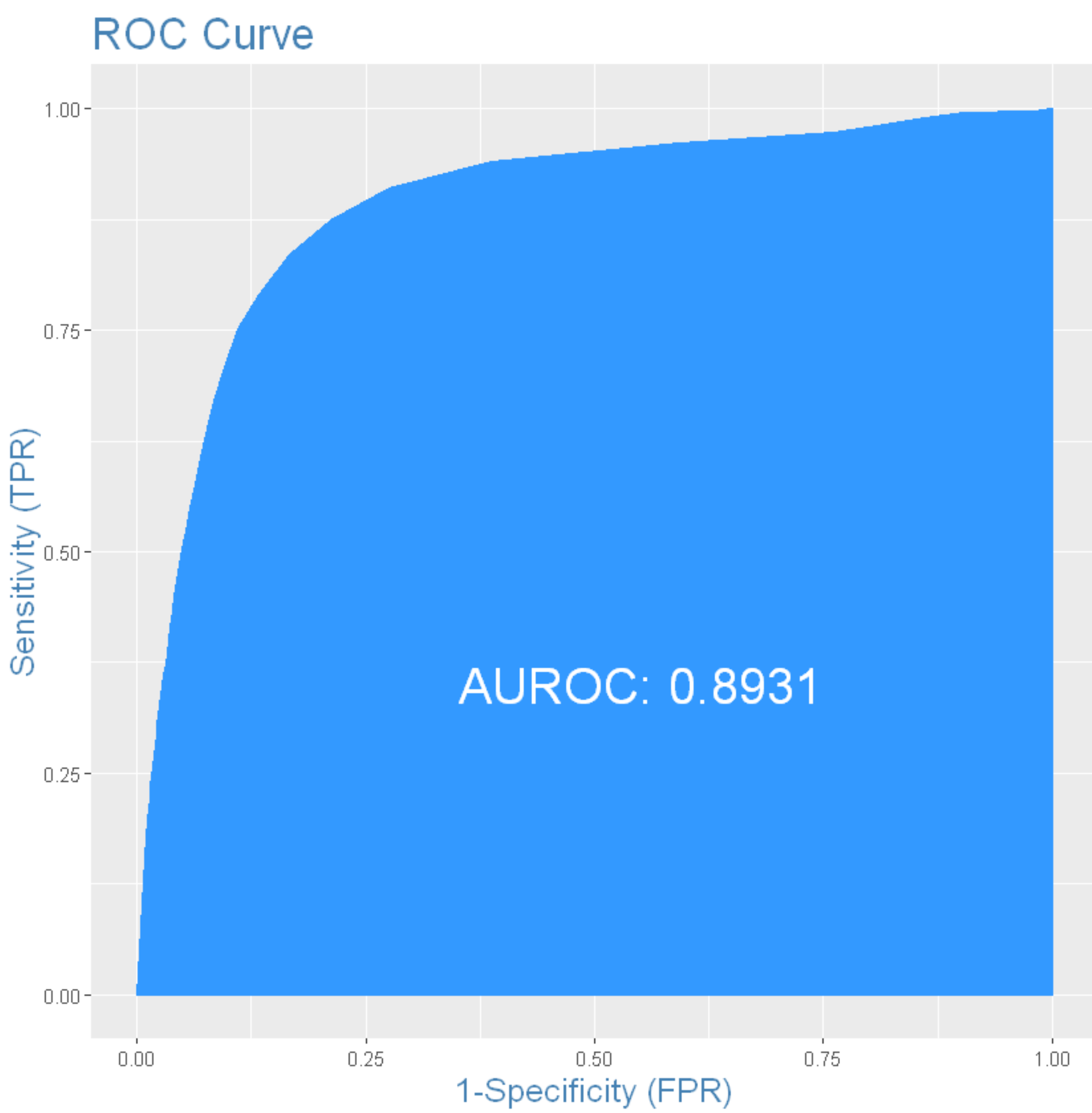
```
In [67]: # check for multicollinearity
         vif(logitpredtrain)
```

category_numeric 1.00200663005289 **backers** 1.00611589899617 **totaldays** 1.00326392920343 **launch_month_numeric** 1.00113798547846 **launch_weekday_numeric** 1.00057758786645

```
In [68]: optCutOff <- optimalCutoff(test$state_numeric, predicted)[1]
         misClassError(test$state_numeric, predicted, threshold = optCutOff)
         # The lower the misclassification error, the better is your model
```

0.1592

```
In [69]: plotROC(test$state_numeric, predicted)
         #Receiver Operating Characteristics Curve traces the percentage of true positives acc
         #predicted by a given logit model as the prediction probability cutoff is lowered fro
         #For a good model, as the cutoff is lowered,
         #it should mark more of actual 1s as positives and lesser of actual 0s as 1s.
         #So for a good model, the curve should rise steeply,
         #indicating that the TPR (Y-Axis) increases faster than the FPR (X-Axis) as the cutof
         #Greater the area under the ROC curve, better the predictive ability of the model.
```




```
In [70]: Concordance(test$state_numeric, predicted)
         # the higher the concordance, the better is the quality of model
```

$Concordance 0.894044539713294

$Discordance 0.105955460286706

$Tied -4.16333634234434e-17

$Pairs 328496322

```
In [71]: sensitivity(test$state_numeric, predicted, threshold = optCutOff)
         # Sensitivity (or True Positive Rate) is the percentage of 1s (actuals) correctly pre
         specificity(test$state_numeric, predicted, threshold = optCutOff)
```

```
# specificity is the percentage of 0s (actuals) correctly predicted
# Specificity can also be calculated as 1False Positive Rate.
```

0.752414920493387
0.889507968372322

```
In [72]: confusMat <- confusionMatrix(test$state_numeric, predicted, threshold = optCutOff)
         confusMat
```

	0	1
0	21712	3332
1	2697	10126

```
In [73]: accuracy <- sum(diag(as.matrix(confusMat)))/sum(confusMat)
         print('--Overall Accuracy--')
         accuracy

[1] "--Overall Accuracy--"
```

0.840784852246019

```
In [74]: # RANDOMFOREST

In [75]: pred_randomforest <- randomForest(state_numeric~category_numeric+backers+totaldays+lau
                                 data=train,
                                 importance=TRUE,
                                 ntree=5)

Warning message in randomForest.default(m, y, ...):
"The response has five or fewer unique values.  Are you sure you want to do regression?"

In [76]: predicted_randomforest <- predict(pred_randomforest, test)

In [77]: confusMat_rs <- confusionMatrix(test$state_numeric, predicted_randomforest)
         confusMat_rs
```

	0	1
0	21060	2172
1	3349	11286

```
In [78]: accuracy <- sum(diag(as.matrix(confusMat_rs)))/sum(confusMat_rs)
         print('--Overall Accuracy--')
         accuracy

[1] "--Overall Accuracy--"
```

0.854200227110677

7 CONCLUSION

I, first, explored the data and created variables that I thought could be useful. Then, I created some visualizations to better grasp the intuition of the dataset. In addition, I created a wordcloud for the project titles whether there is a common theme. And lastly, I tried to predict the success / fail likelihood of a project.

I am using logistic model here by using the variables that I believe to be correcting whether the project will go through. First, I converted successful attempts in the kickstarter dataset to 1, and all the others to 0. Then, I set my equation with the independent variables that I found, in this case, to better predict. These variables do not carry multicollinearity and all statistically significant. Therefore, I can move on with them. Of course, this is a preliminary result but to give an idea of what can be done, this is a good approach.

Although the model I created is far from perfect, I can predict whether the project will be successful at 84% of the time by using logistic regression.

I also tried randomforest, which improves the prediction accuracy just a bit. It's stable at 85.4%.

This prediction can be extended predicting a project from wordcloud base. By working with only the titles, no description, 30% accuracy can be obtained. I assume that including a detailed description of each project and breaking it down to wordcloud analysis, this accuracy can be improved.

```
In [ ]:
```