Deliverable1

Cathy Kim, Leo McGann, Anastacia Wahl, Stefan Wenc / SWACK

Abstract

The aim of this document is to summarize the early exploration and data analysis performed for our final project. In this project, we seek to investigate the determinants of kickstarter project success. Below, we describe our dataset and variables of interest, state our research question, and describe and summarize the Exploratory Data Analysis (EDA) performed.

Data Description

This dataset is a collection of data on Kickstarter projects, a website that allows companies and individuals to request and receive funding from individuals for their projects or goals. This dataset came from Kaggle, and was gathered by a user named Mickael Mouille. The dataset has information about 378,661 kickstarter projects on 15 variables. Those variables are:

1. ID: a unique ID number for each project (categorical)
2. Name: the name of the project inputted by the person or organization seeking to raise funds (categorical)
3. Category: a categorical variable that places each project into a category of fundraising (categorical)
4. Main_Category: a categorical variable that places each project into a broader category of project type than the Category variable (categorical)
5. Currency: a categorical variable on the type of currency the user is fundraising in and that their goal is measured in (categorical)
6. Deadline: the deadline by which the fundraiser is seeking to meet their fundraising goal (continuous)
7. Goal: the amount of the given currency that the fundraiser is seeking to raise (continuous)
8. Launched: the date and time that the project was posted and began (continuous)
9. Pledged: the amount of money in the given currency the kickstarter project raised between its launch date and deadline (continuous)
10. State: whether or not the project was successful in reaching the goal set by the fundraiser (categorical)
11. Backers: the number of people who donated to the kickstarter project (continuous)
12. Country: the country that the kickstarter project is located in (categorical)
13. USD Pledged: the amount of money pledged to the kickstarter project in US Dollars, as converted by Kickstarter (continuous)
14. USD_pledged_real: the amount of money pledged to the kickstarter project in US Dollars, as converted by the Fixer.io API for currency exchange rates (continuous)
15. USD_goal_real: the amount of money set as the goal for the kickstarter project in US Dollars, as converted by the Fixer.io API for currency exchange rates (continuous)

Research Question

How does success of a kickstarter project depend on duration, goal, and category of project? Additionally, what other variables impact the success of kickstarter projects (ex. Title keywords, month of launch, year of launch)?

This research question is of importance because it will allow us to better understand what leads to success in a kickstarter project, allowing people who want to perform kickstarter projects in the future the opportunity to use these conclusions to optimize their projects.

Data Import

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
kickstarters <- read.csv("kickstarter-projects/ks-projects-201801.csv")
problems(kickstarters)
```

```
## [1] row      col      expected actual
## <0 rows> (or 0-length row.names)
```

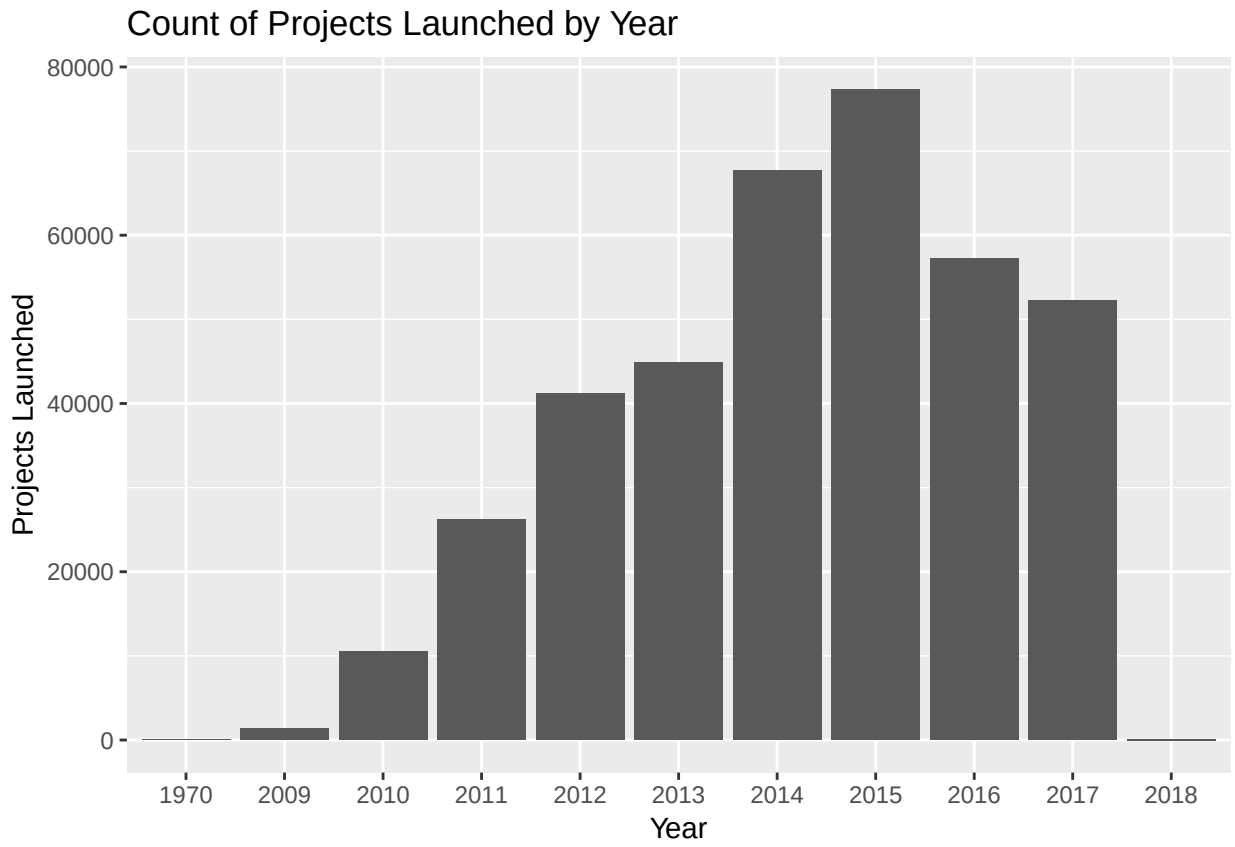
The data import process went smoothly, with no errors found when using problems() to parse the data.

```
kickstarters2 <- kickstarters %>% mutate(proportion_raised = (usd_pledged_real/usd_goal_real)) %>% muta
```

Added a proportion_raised variable to the data set to display the proportion of the total goal that was met. Added a true_state variable to split projects into success/failure based on the amount of money raised and whether or not the goal was met. For this variable, TRUE is equivalent to successful, and FALSE is equivalent to FAILED.

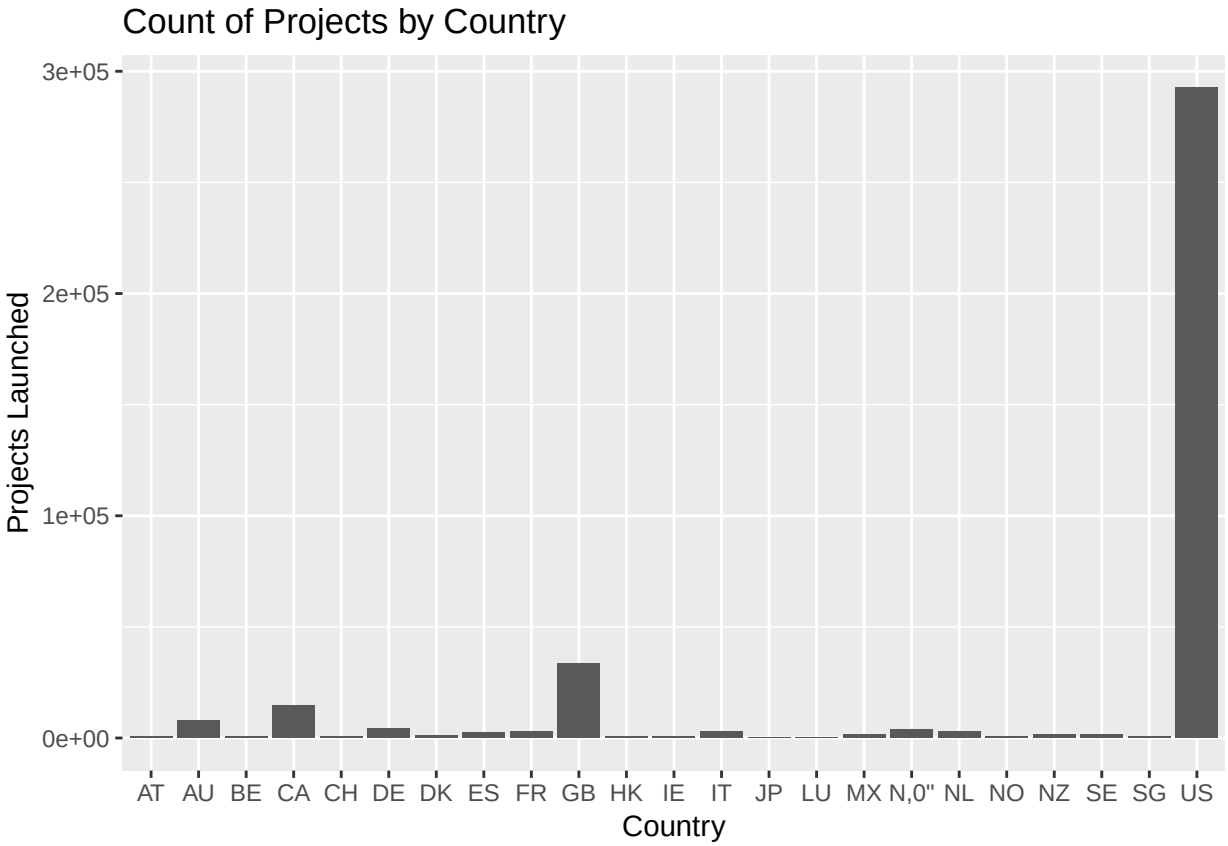
Exploratory Data Analysis: Variation of Single Variables

```
library(stringr)
kickstarters2 %>% mutate(yearlaunched = str_sub(launched, start = 1, end = 4)) %>% ggplot() + geom_bar(a
```



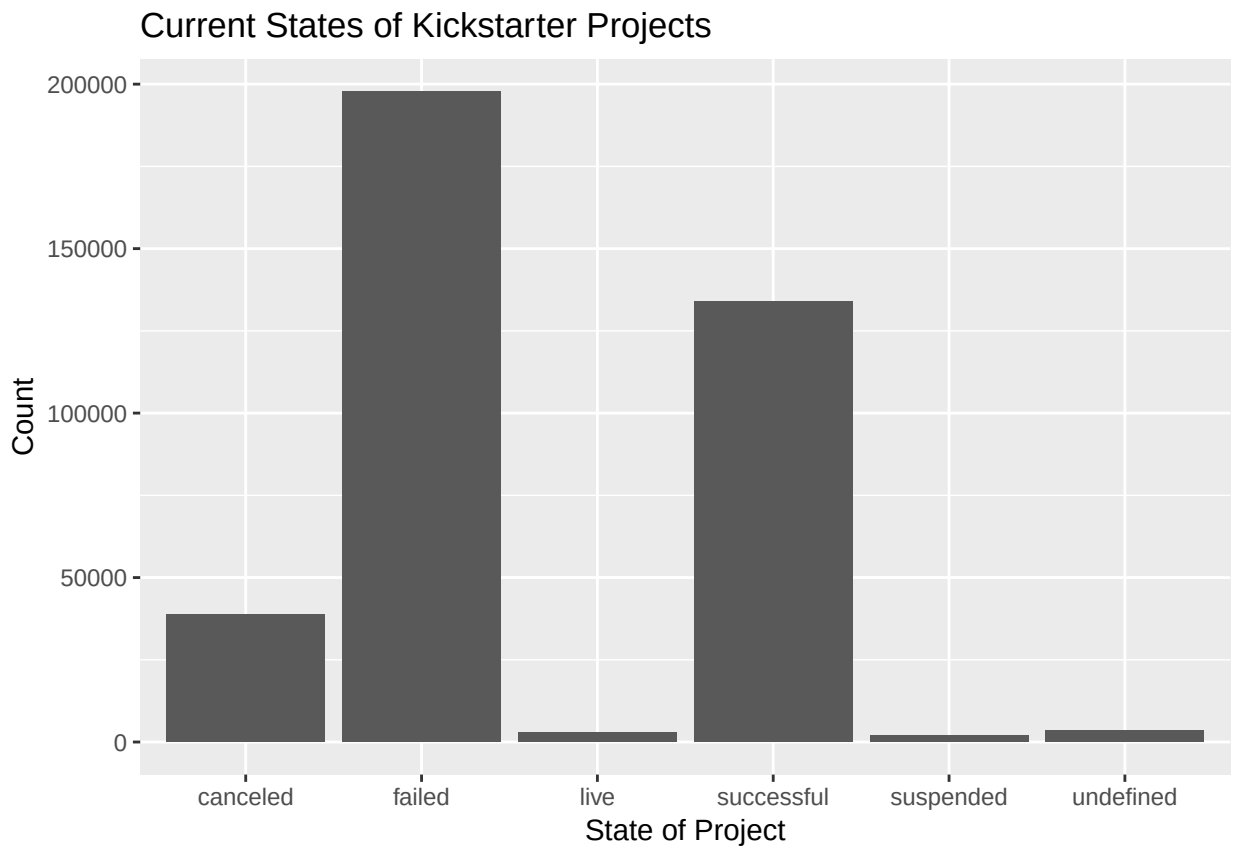
There were 7 projects launched in the year 1970. We opt to exclude these outliers in our analysis, as we feel they do not adequately represent the modern kickstarter climate.

```
kickstarters2 %>% ggplot() + geom_bar(aes(x = country)) + labs (x = 'Country', y = 'Projects Launched',
```

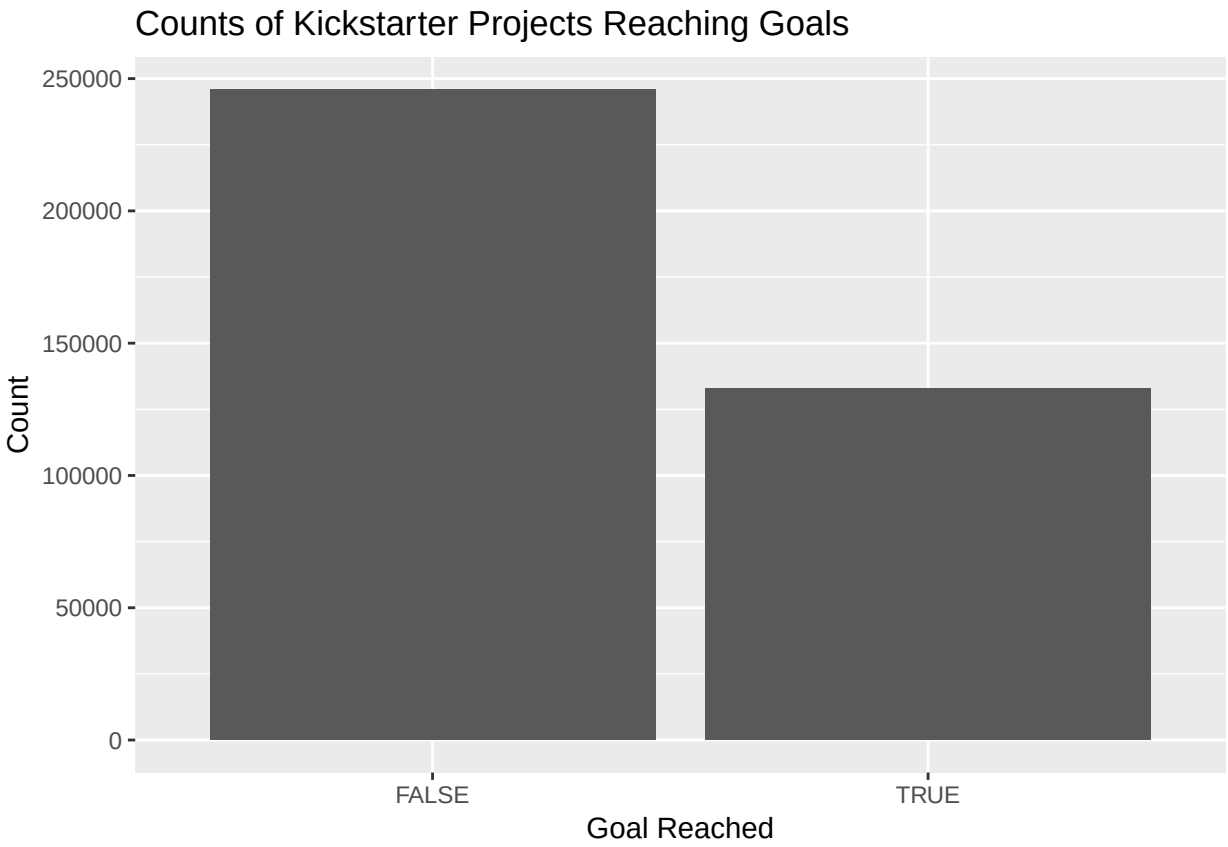


This bar graph indicates that the majority of projects were launched from the US, with the second most projects coming from Great Britain.

```
kickstarters2 %>% ggplot(aes(x = state)) + geom_bar() + labs(x = "State of Project", y = "Count", title
```

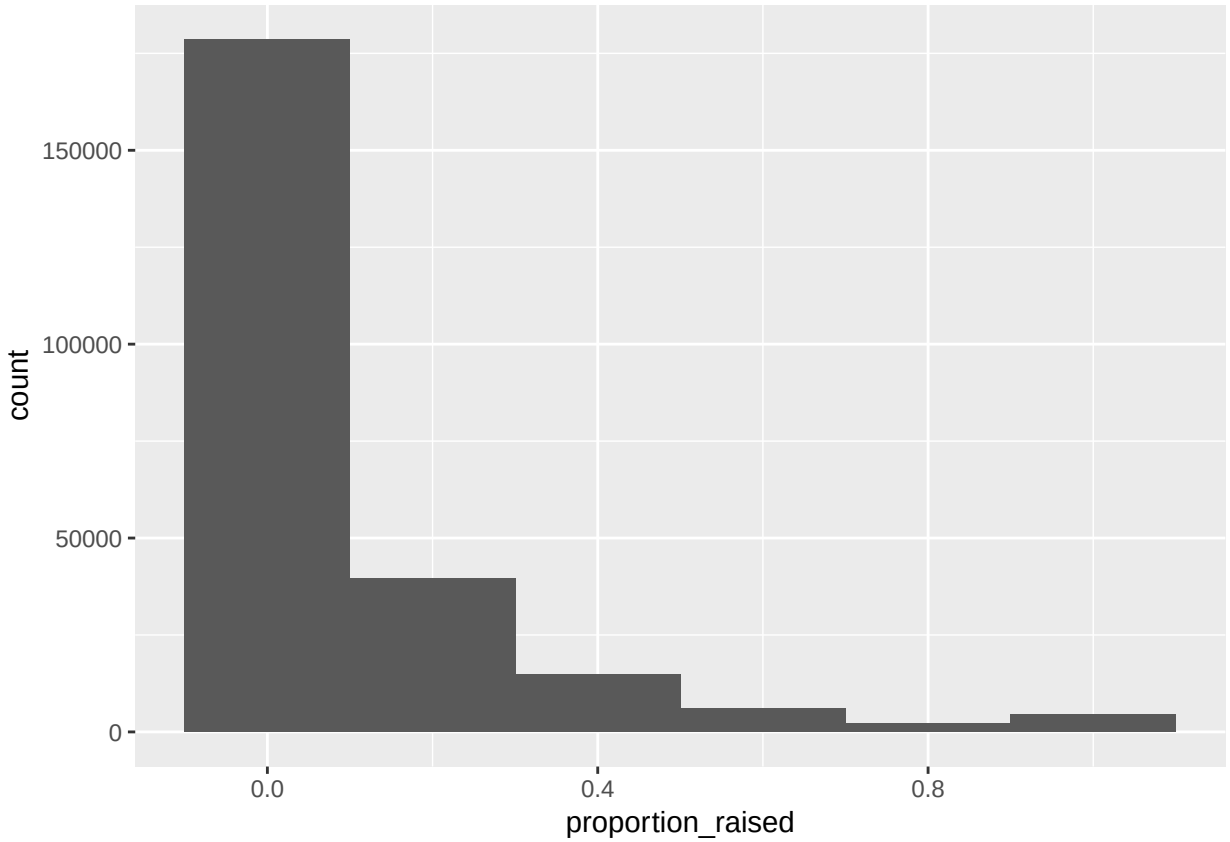



```
kickstarters2 %>% ggplot(aes(x = true_state)) + geom_bar() + labs(x = "Goal Reached", y = "Count", titl
```



These two bar graphs show that overall, the highest number of the projects end up as failed, and the majority of projects do not meet their eventual goals.

```
kickstarters2 %>% filter(proportion_raised <= 1) %>% ggplot(aes(x = proportion_raised)) + geom_histogra
```



This histogram of the proportions raised for projects that did not meet their goals shows that the majority of projects that do not meet their goals don't come close, only reaching less than 20% of their goal.

```
pledgedtable <- kickstarters2 %>% group_by(main_category) %>% summarise(Mean_Raised = mean(usd_pledged_
pledgedtable
```

```
## # A tibble: 15 x 3
##    main_category Mean_Raised Standard_Deviation
##    <fct>               <dbl>              <dbl>
##  1 Crafts              1633.              8087.
##  2 Journalism          2616.             10451.
##  3 Art                 3221.             21855.
##  4 Publishing          3350.             14817.
##  5 Dance               3453.              5769.
##  6 Photography         3572.             19273.
##  7 Music               3858.             13015.
##  8 Theater             4006.             10850.
##  9 Food                5114.             31220.
## 10 Fashion             5712.             29930.
## 11 Film & Video        6158.             41399.
## 12 Comics              6610.             24409.
## 13 Games              21042.            168530.
## 14 Technology         21151.            126726.
## 15 Design             24417.            214977.
```

This data table shows the mean and standard deviation of funds raised in USD in each of the main category groups. From this table we can see that many of the categories have very large standard deviations, indicating

a large amount of variation in the amount raised. Additionally, we can see that Crafts projects raise the least amount of money on average, while design projects raise the most on average.

```
#0-100
group1 <- kickstarters2 %>% filter(usd_goal_real <= 100) %>% select(usd_goal_real)
group1_mean <-sum(group1$usd_goal_real) / 6082
group1_prop <- 6082/378661
#100-1000
group2 <- kickstarters2 %>% filter(usd_goal_real > 100 & usd_goal_real <= 1000) %>% select(usd_goal_real
group2_mean <- sum(group2$usd_goal_real) / 54811
group2_prop <- 54811/378661
#1000-10000
group3 <- kickstarters2 %>% filter(usd_goal_real > 1000 & usd_goal_real <= 10000) %>% select(usd_goal_re
group3_mean <- sum(group3$usd_goal_real) / 193351
group3_prop <- 193351/378661
#10000+
group4 <- kickstarters2 %>% filter(usd_goal_real > 10000) %>% select(usd_goal_real)
group4_mean <- sum(group4$usd_goal_real) / 124417
group4_prop <- 124417/378661
usd_goals_data <- data_frame(
  name = c("$0-100", "$101-1000", "$1001-10000", "$10000+"),
  total_avg = c(group1_mean, group2_mean, group3_mean, group4_mean),
  proportion_of_data = c(group1_prop, group2_prop, group3_prop, group4_prop)
)
```

```
## Warning: `data_frame()` is deprecated, use `tibble()`.
## This warning is displayed once per session.
```

```
usd_goals_data
```

```
## # A tibble: 4 x 3
##   name        total_avg proportion_of_data
##   <chr>           <dbl>              <dbl>
## 1 $0-100           53.8             0.0161
## 2 $101-1000       634.              0.145
## 3 $1001-10000    4826.              0.511
## 4 $10000+      130557.              0.329
```

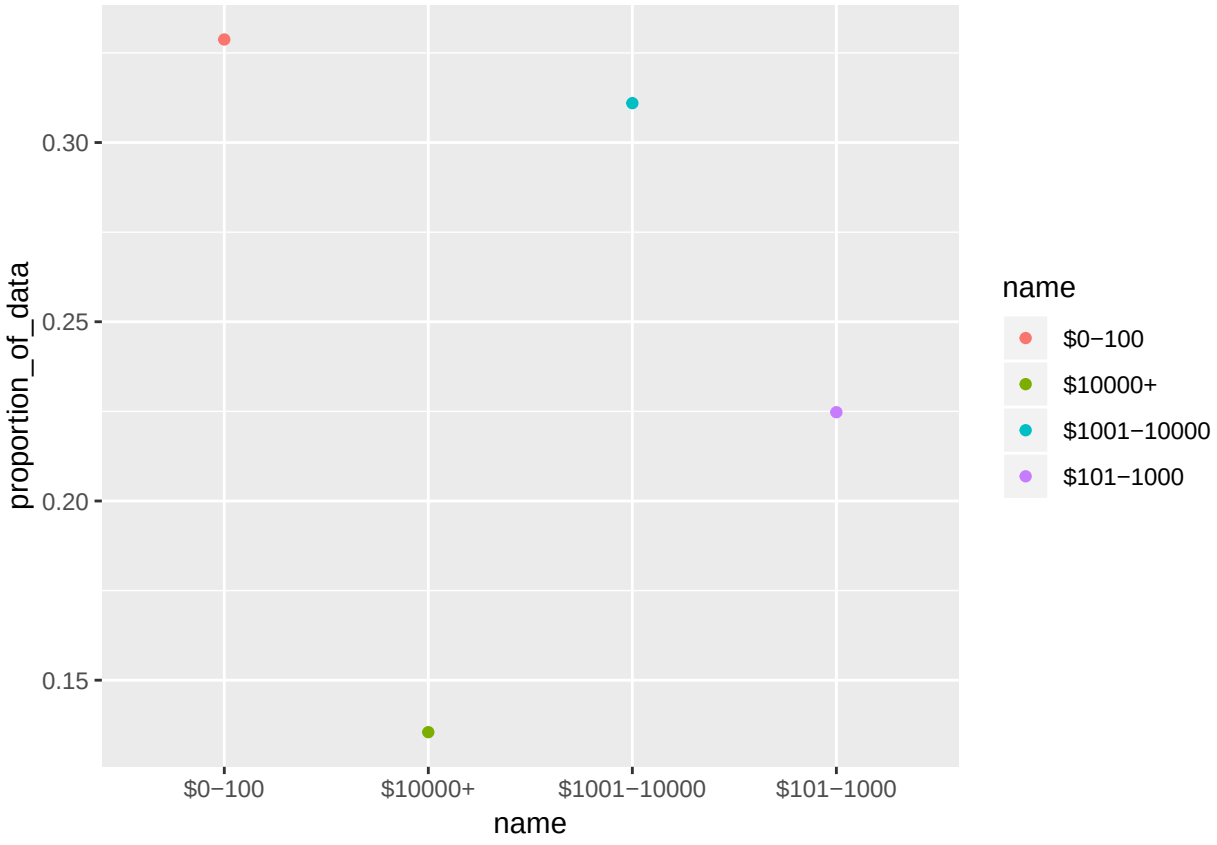
In usd_goals_data, we are observing the variability of the goals set by each kickstarter project in US dollars. We see that over 80% of the projects were set to be over 1,000 USD. The majority at 51% was between 1,000 and 10,000 USD. Furthermore, we found the average to see generally within each grouping, what the average was. For example, between 100-1,000 USD, we see that the average was ~$634, which is very near the midway point of the grouping, so this could suggest that there is not much of a certain bracket but more evenly spread out of number of projects that set their goals between 100-1000.

```
#0-100
group1 <- kickstarters2 %>% filter(usd_pledged_real <= 100) %>% select(usd_pledged_real)
group1_pmean <-sum(group1$usd_pledged_real) / 124486
group1_pprop <- 124486/378661
#100-1000
group2 <- kickstarters2 %>% filter(usd_pledged_real > 100 & usd_pledged_real <= 1000) %>% select(usd_pl
group2_pmean <- sum(group2$usd_pledged_real) / 85108
group2_pprop <- 85108/378661
#1000-10000
group3 <- kickstarters2 %>% filter(usd_pledged_real > 1000 & usd_pledged_real <= 10000) %>% select(usd_
group3_pmean <- sum(group3$usd_pledged_real) / 117758
group3_pprop <- 117758/378661
```

```
#10000+
group4 <- kickstarters2 %>% filter(usd_pledged_real > 10000) %>% select(usd_pledged_real)
group4_pmean <- sum(group4$usd_pledged_real) / 51309
group4_pprop <- 51309/378661
usd_pledged_data <- data_frame(
  name = c("$0-100", "$101-1000", "$1001-10000", "$10000+"),
  total_pledged_avg = c(group1_pmean, group2_pmean, group3_pmean, group4_pmean),
  proportion_of_data = c(group1_pprop, group2_pprop, group3_pprop, group4_pprop)
)
usd_pledged_data
```

```
## # A tibble: 4 x 3
##   name        total_pledged_avg proportion_of_data
##   <chr>                   <dbl>              <dbl>
## 1 $0-100                   18.9              0.329
## 2 $101-1000               420.               0.225
## 3 $1001-10000            3723.               0.311
## 4 $10000+               57566.               0.136
```

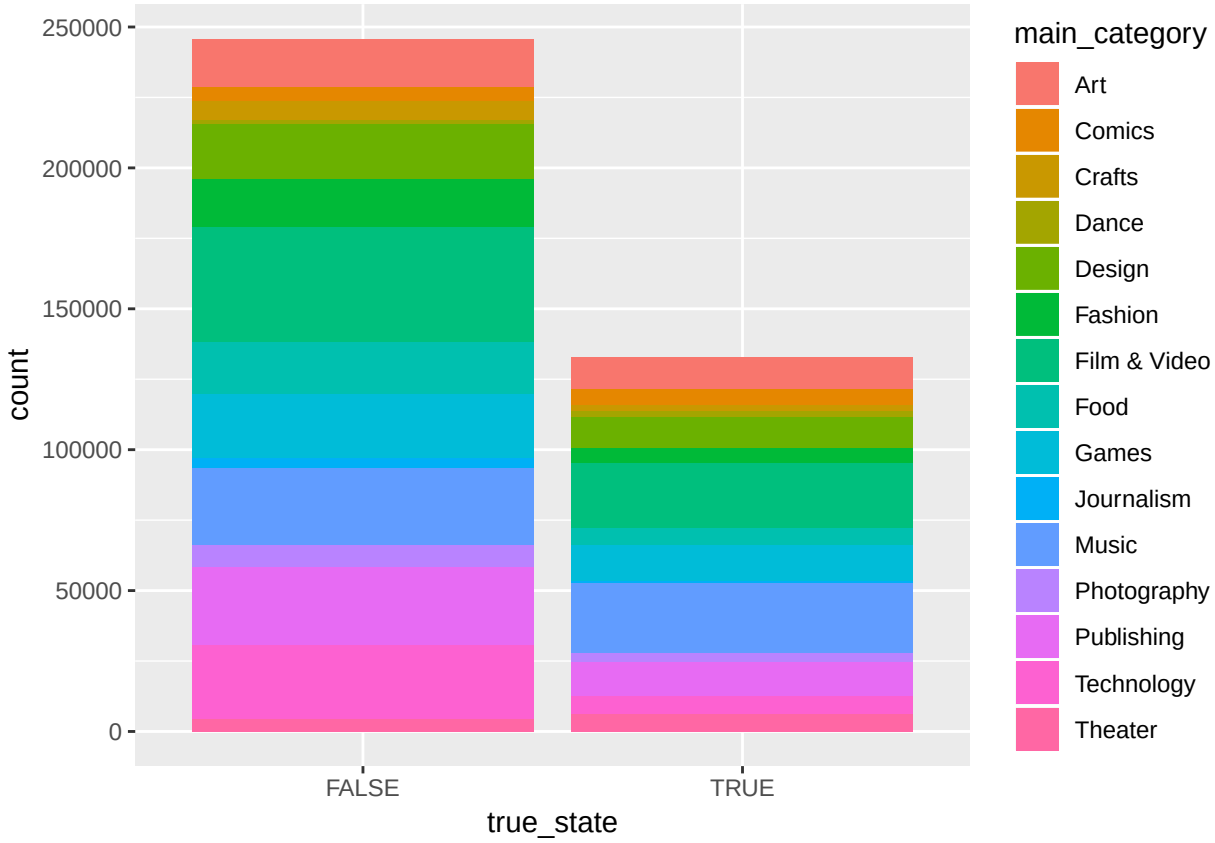
```
ggplot(usd_pledged_data, (aes(x= name,y = proportion_of_data))) + geom_point(aes(color = name))
```



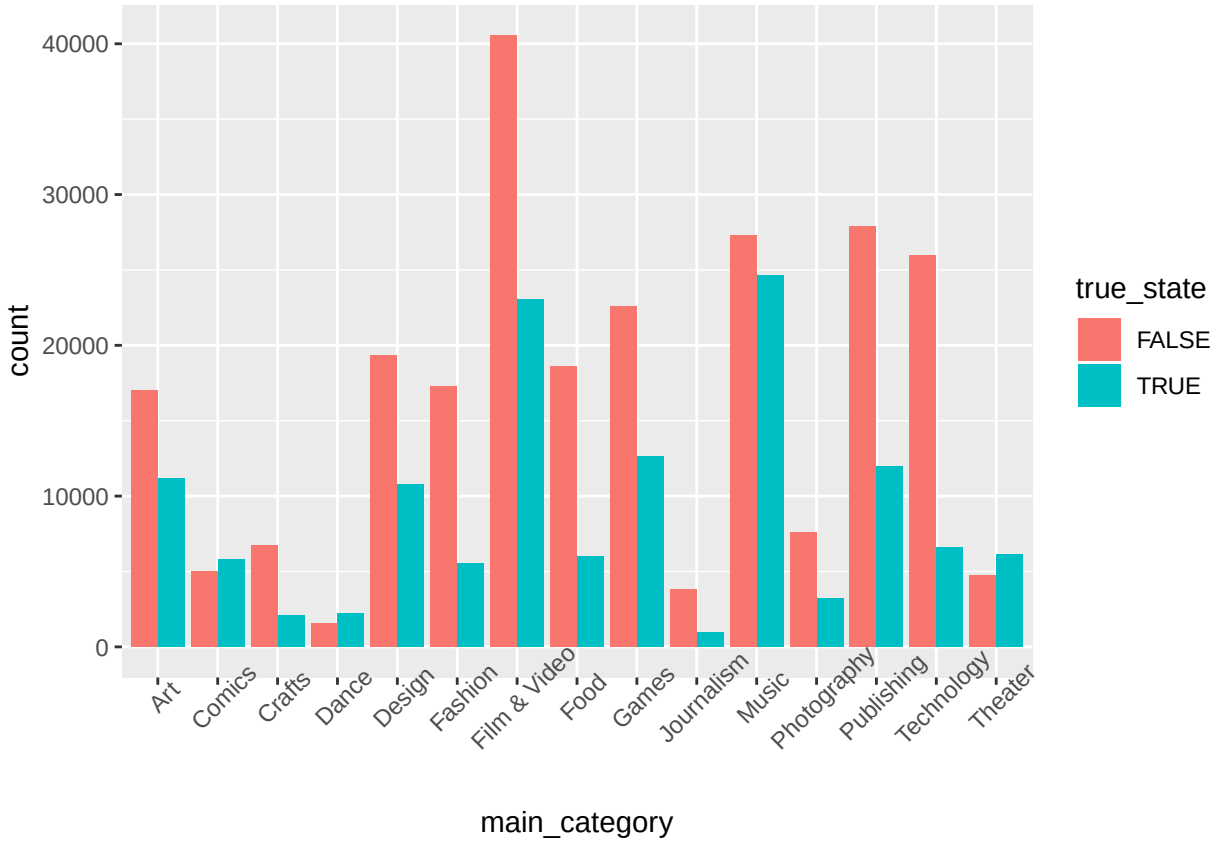
In the usd_pledged, we see that the majority of the projects had between 0 and 100 USD pledged, next between 1000 to 10,000 and then 100-1000 at 22.5% and only about 14% pledged over 10,0000. Furthermore, we can see the averages for USD actually pledged must have some clustering especially in the group 0-100 because the average is 18USD, suggesting a large number of the projects got around $0 pledged.

Exploratory Data Analysis: Covariation of Multiple Variables

```
kickstarters2 %>% ggplot(aes(x = true_state, fill = main_category)) + geom_bar()
```



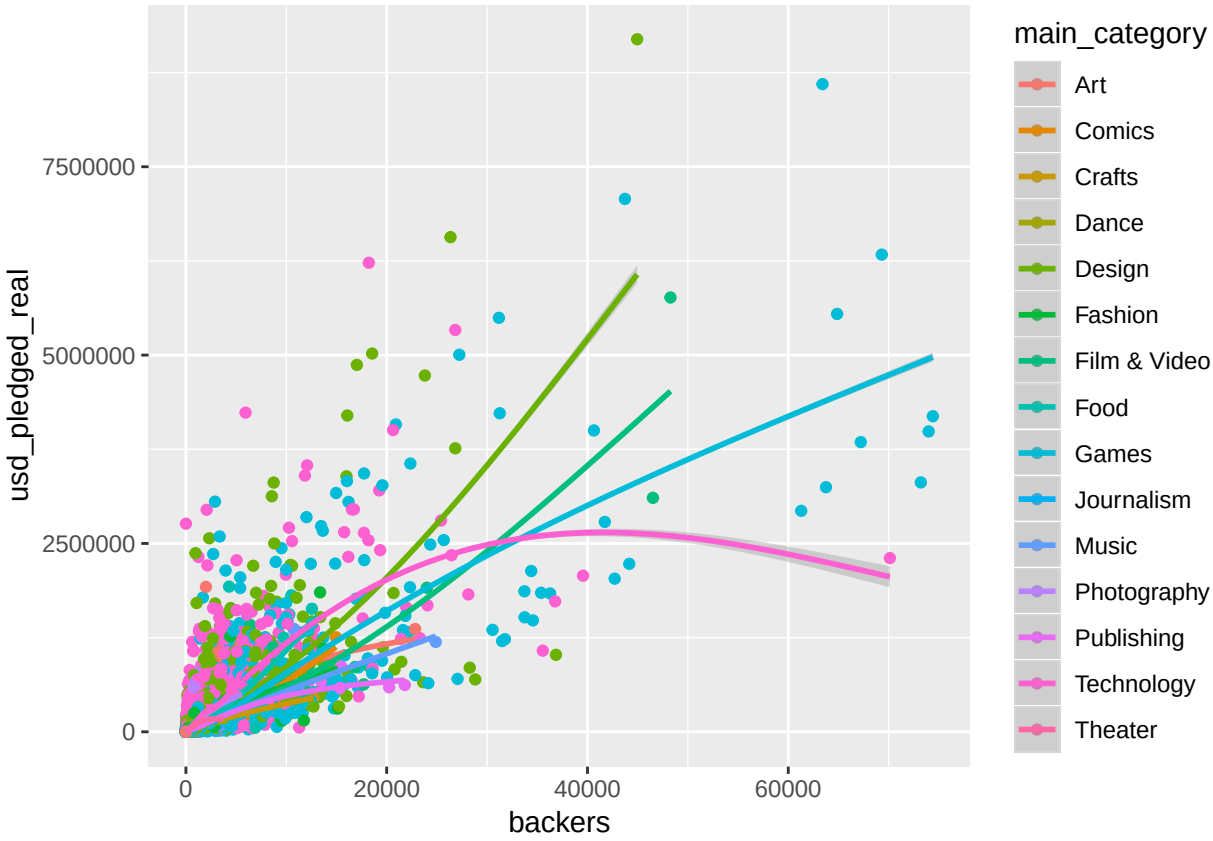

```
kickstarters2 %>% ggplot(aes(fill = true_state, x = main_category)) + geom_bar(position = "dodge") + th
```



From this graph we can see that kickstarters in the Comics, Dance, and Theater categories have more successes than failures. This is an interesting observation, because overall kickstarters tend to fail more often than they succeed as our previous graphs have shown.

```
kickstarters2 %>% filter(usd_pledged_real < 10000000, backers < 75000) %>% ggplot(aes(x = backers, y = u
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
kickstarters3 <- kickstarters2 %>% filter(usd_pledged_real > 10000000)
kickstarters4 <- kickstarters2 %>% filter(backers > 75000)
kickstarters3
```

```
##           ID                                                    name
## 1 1799979574           Pebble Time - Awesome Smartwatch, No Compromises
## 2 2103598555                      Pebble 2, Time 2 + All-New Pebble Core
## 3  342886736 COOLEST COOLER: 21st Century Cooler that's Actually Cooler
## 4  506924864               Pebble: E-Paper Watch for iPhone and Android
## 5  545070200                                  Kingdom Death: Monster 1.5
##         category main_category currency   deadline  goal
## 1 Product Design        Design      USD 2015-03-28 5e+05
## 2 Product Design        Design      USD 2016-06-30 1e+06
## 3 Product Design        Design      USD 2014-08-30 5e+04
## 4 Product Design        Design      USD 2012-05-19 1e+05
## 5 Tabletop Games         Games      USD 2017-01-08 1e+05
##              launched  pledged      state backers country usd.pledged
## 1 2015-02-24 15:44:42 20338986 successful   78471      US    20338986
## 2 2016-05-24 15:49:52 12779843 successful   66673      US    12779843
## 3 2014-07-08 10:14:37 13285226 successful   62642      US    13285226
## 4 2012-04-11 06:59:04 10266846 successful   68929      US    10266846
## 5 2016-11-25 06:01:41 12393140 successful   19264      US     5228482
##   usd_pledged_real usd_goal_real proportion_raised true_state
## 1         20338986         5e+05          40.67797       TRUE
## 2         12779843         1e+06          12.77984       TRUE
## 3         13285226         5e+04         265.70453       TRUE
## 4         10266846         1e+05         102.66846       TRUE
```

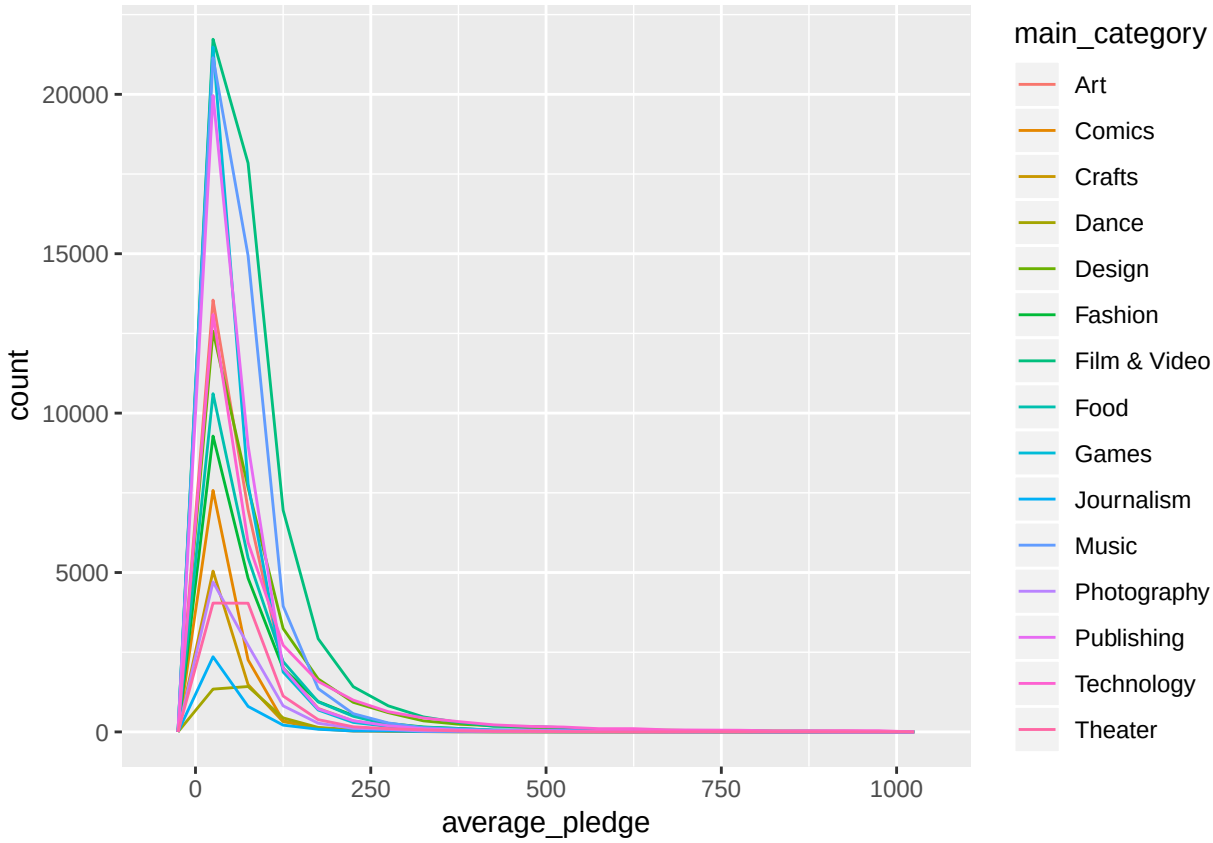
```
## 5         12393140         1e+05         123.93140       TRUE
```

```
kickstarters4
```

```
##           ID                                                    name
## 1 1118803016                           Bears vs Babies - A Card Game
## 2 1386523707                           Fidget Cube: A Vinyl Desk Toy
## 3 1755266685                         The Veronica Mars Movie Project
## 4 1799979574        Pebble Time - Awesome Smartwatch, No Compromises
## 5 1929840910                                   Double Fine Adventure
## 6 1955357092                                       Exploding Kittens
## 7  557230947 Bring Reading Rainbow Back for Every Child, Everywhere!
##         category main_category currency   deadline    goal
## 1 Tabletop Games         Games      USD 2016-11-18   10000
## 2 Product Design        Design      USD 2016-10-20   15000
## 3 Narrative Film  Film & Video      USD 2013-04-13 2000000
## 4 Product Design        Design      USD 2015-03-28  500000
## 5    Video Games         Games      USD 2012-03-14  400000
## 6 Tabletop Games         Games      USD 2015-02-20   10000
## 7            Web    Technology      USD 2014-07-02 1000000
##              launched  pledged      state backers country usd.pledged
## 1 2016-10-18 18:59:32  3215680 successful   85581      US     1231456
## 2 2016-08-30 22:02:09  6465690 successful  154926      US       13770
## 3 2013-03-13 15:42:22  5702153 successful   91585      US     5702153
## 4 2015-02-24 15:44:42 20338986 successful   78471      US    20338986
## 5 2012-02-09 02:52:52  3336372 successful   87142      US     3336372
## 6 2015-01-20 19:00:19  8782572 successful  219382      US     8782572
## 7 2014-05-28 15:05:45  5408917 successful  105857      US     5408917
##   usd_pledged_real usd_goal_real proportion_raised true_state
## 1          3215680         10000        321.567979       TRUE
## 2          6465690         15000        431.046020       TRUE
## 3          5702153       2000000          2.851077       TRUE
## 4         20338986        500000         40.677973       TRUE
## 5          3336372        400000          8.340930       TRUE
## 6          8782572         10000        878.257199       TRUE
## 7          5408917       1000000          5.408917       TRUE
```

From this scatter plot, we can see that in general there is a positive correlation between the number of backers and the the total amount pledged. Data was filtered for pledged totals less than 10 million total dollars and number of backers less than 75000 to better show data. As shown by these two data tables, this filters out 11 rows of data.

```
kickstarters_avpledge <- kickstarters2 %>% filter(backers > 0, usd_pledged_real > 0) %>% mutate(average_
kickstarters_highavg <- kickstarters_avpledge %>% filter(average_pledge > 1000)
kickstarters_avpledge %>% filter(average_pledge < 1000, average_pledge > 0) %>% ggplot(aes(x = average_p
```



The filters used in creating this graph filtered out all kickstarters that had no backers or no amount pledged. To create the histogram, we also filtered out the 862 rows of data with an average pledge of greater than 1000, to allow us to view the graph better. From this histogram we find that the most frequent average pledge is actually the second bin, with a range of 50 to 100 dollars, for most categories.