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. We observed the single variable variation as well as covariation between variables such as amount of money pledged, the kickstarter goal of projects, project counts by year and country in order to better distinguish influences on the final states (success or failure) of the projects. Finally, a discussion on our interpretations of our results and final conclusions and thoughts for our research question.

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)
```

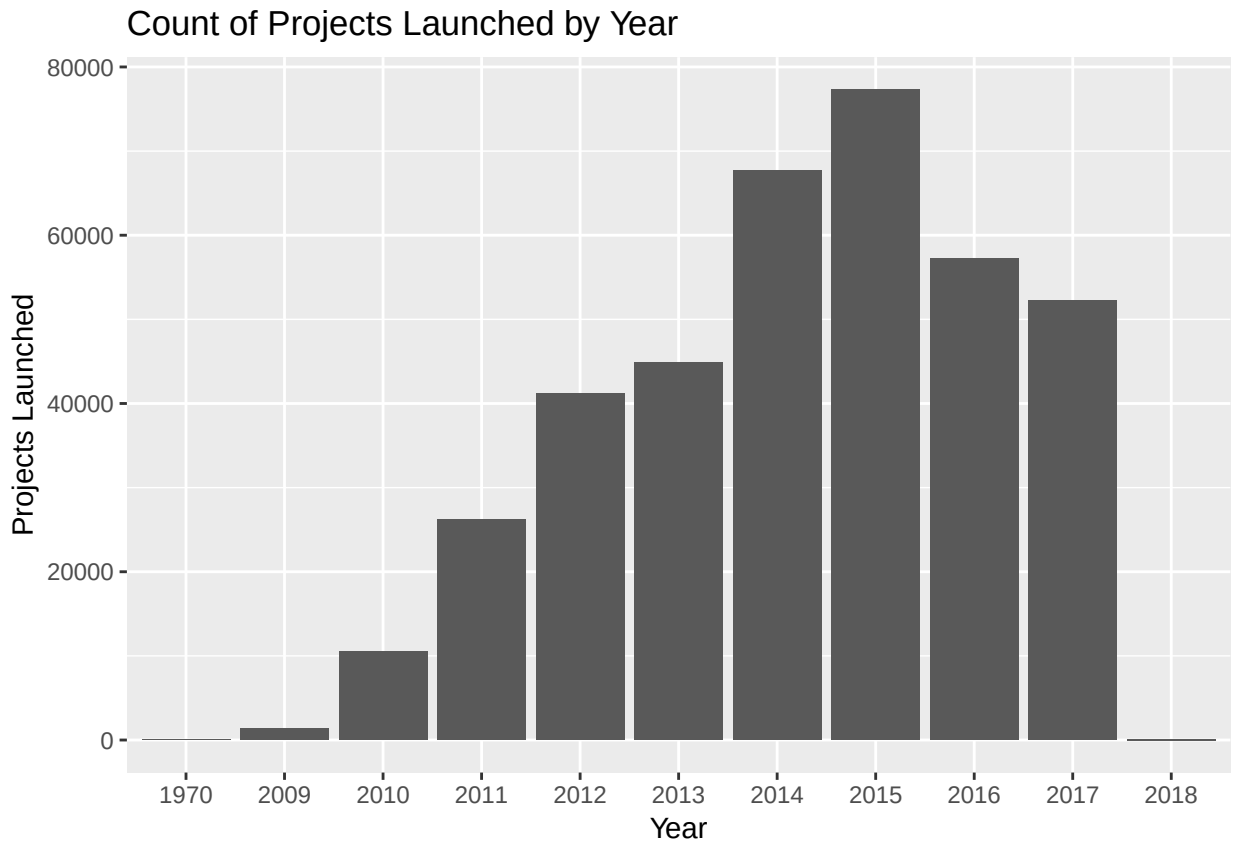
The data import process went smoothly, with no errors found when using problems() to parse the data.

```
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.

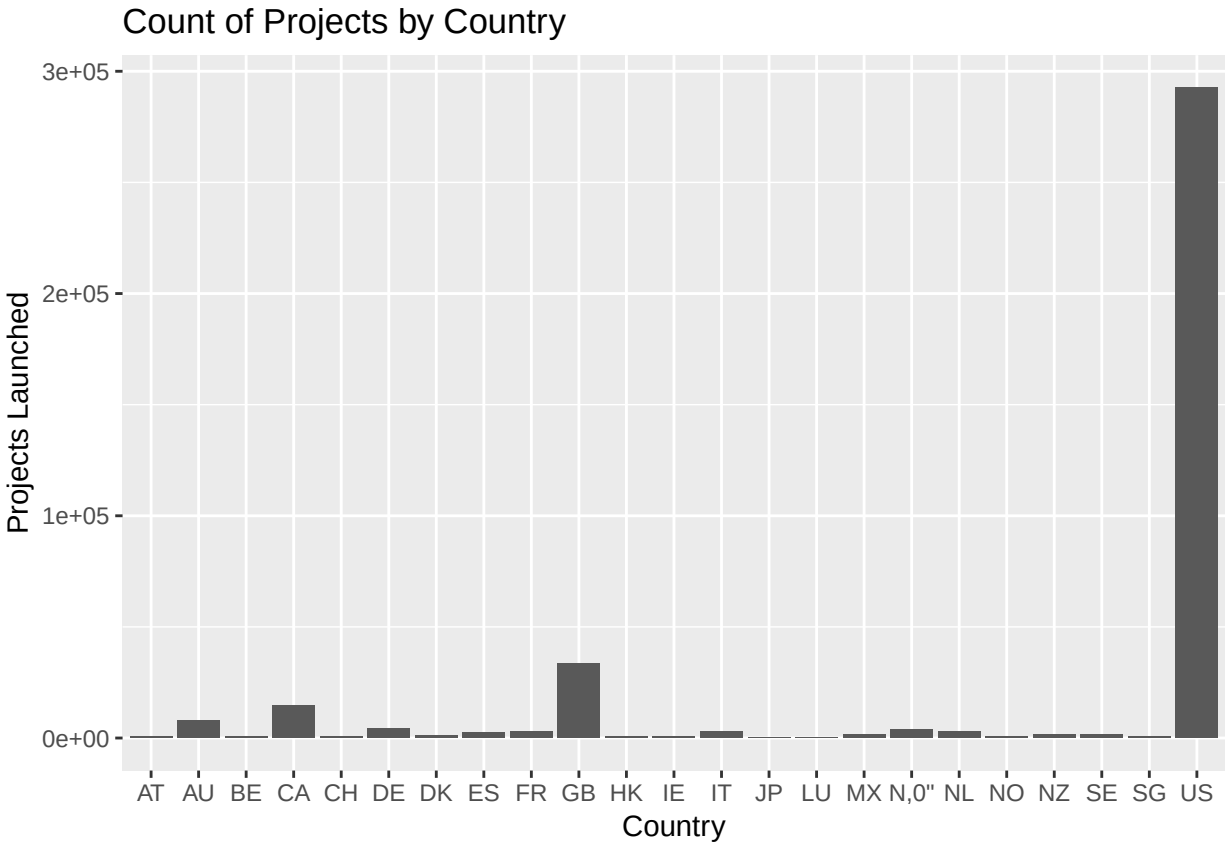
Exploratory Data Analysis: Variation of Single Variables

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



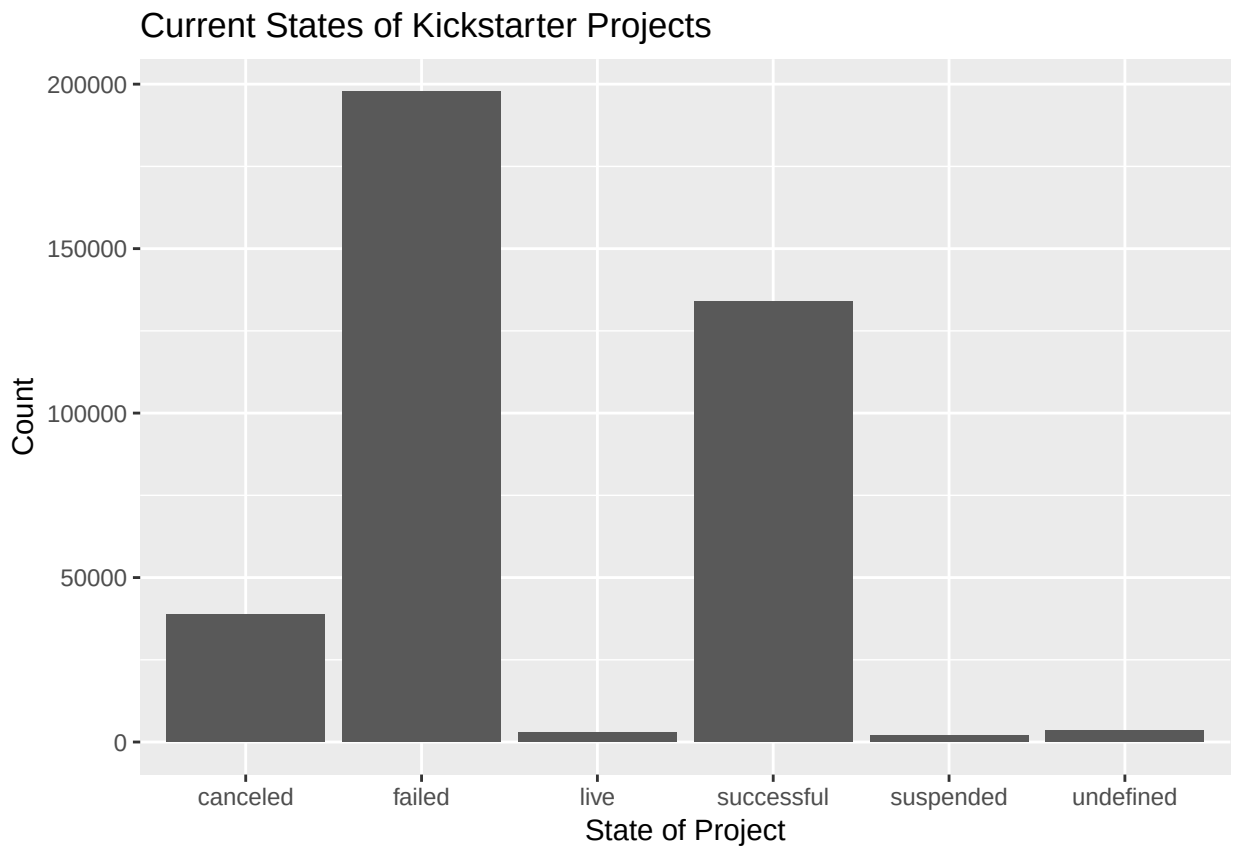
There were seven projects launched in the year 1970. We will opt to exclude these outliers in our analysis, as we feel they do not adequately represent the modern kickstarter climate. Otherwise, the years for which we have data are 2009-2018. We have the largest amount of kickstarter data for the years 2015, 2014, and 2016.

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

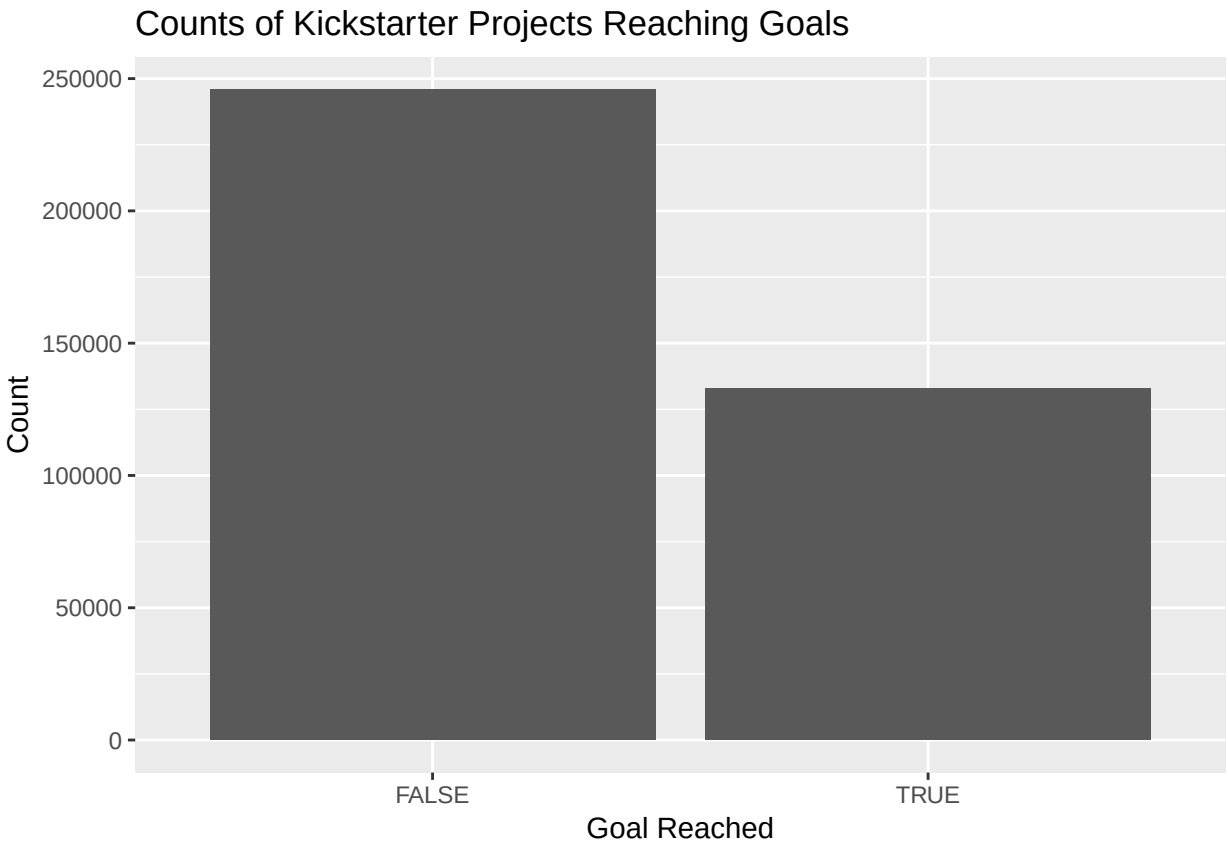


Our dataset includes data for 23 distinct countries. Without our dataset, the greatest number of projects were started in the United States, Great Britain, Canada, and Australia.

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

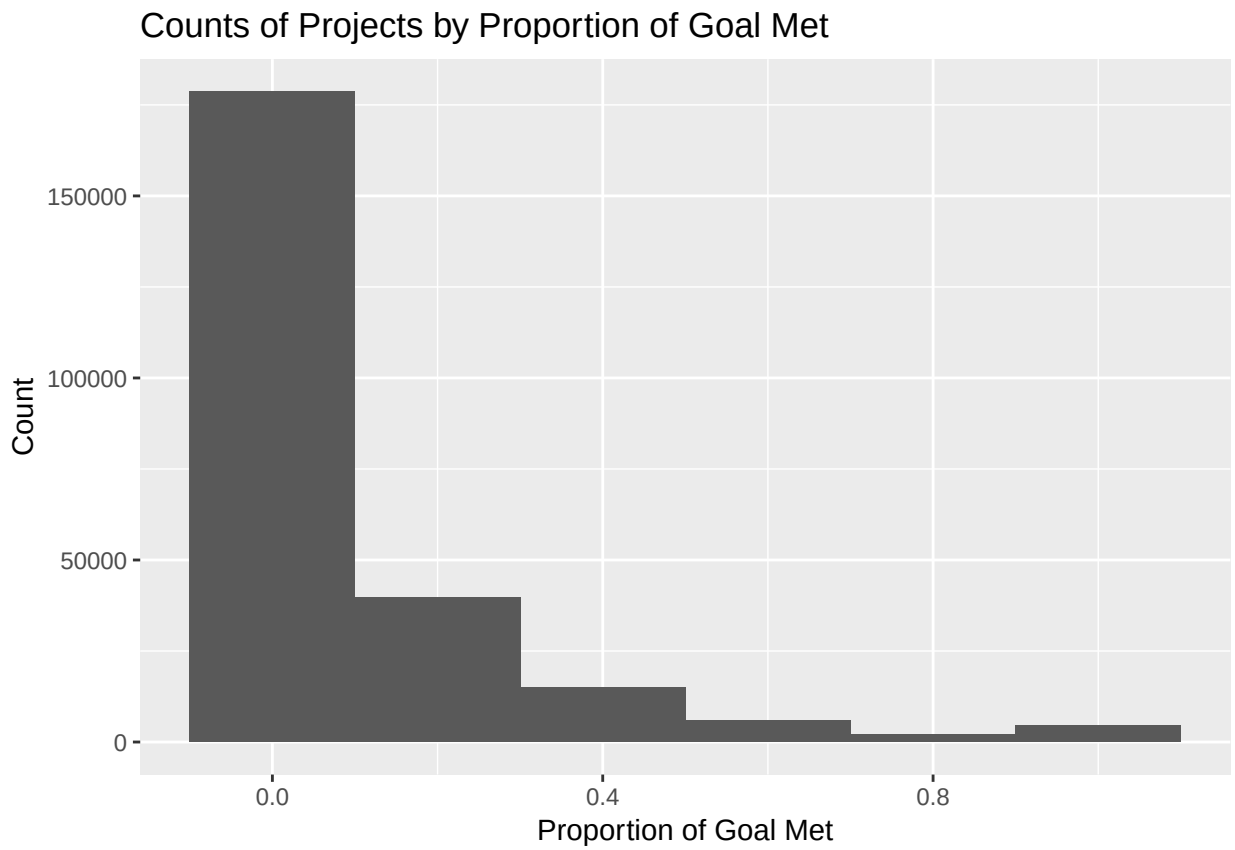



```
kickstarters2 %>% ggplot(aes(x = true_state)) + geom_bar() + labs(x = "Goal Reached", y = "Count", titl
```



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
```

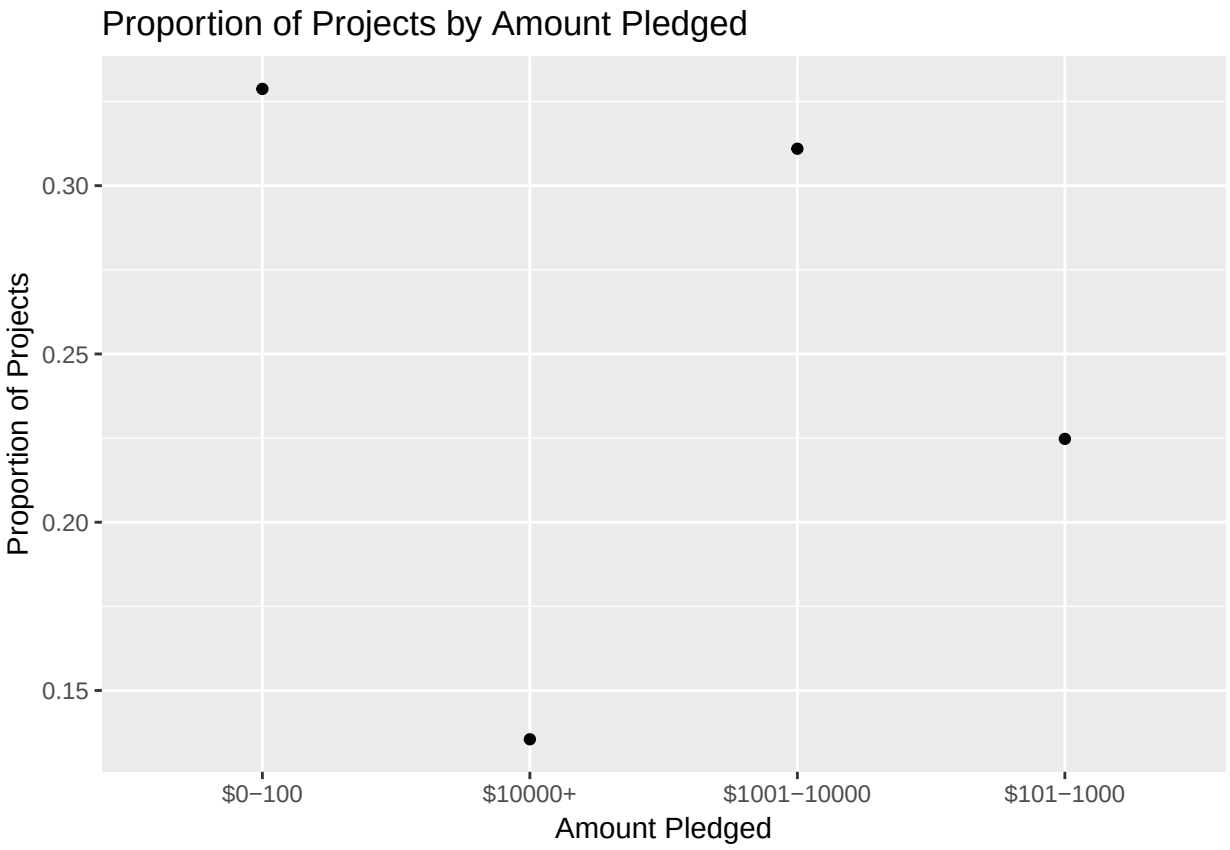
In usd_goals_data, we are observing the variability of the goals set by each kickstarter project in US dollars. We see that over 80% of the projects were set to be over 1,000 USD. The majority at 51% was between 1,000 and 10,000 USD. Furthermore, we found the average to see generally within each grouping, 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
```

```r
#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
```

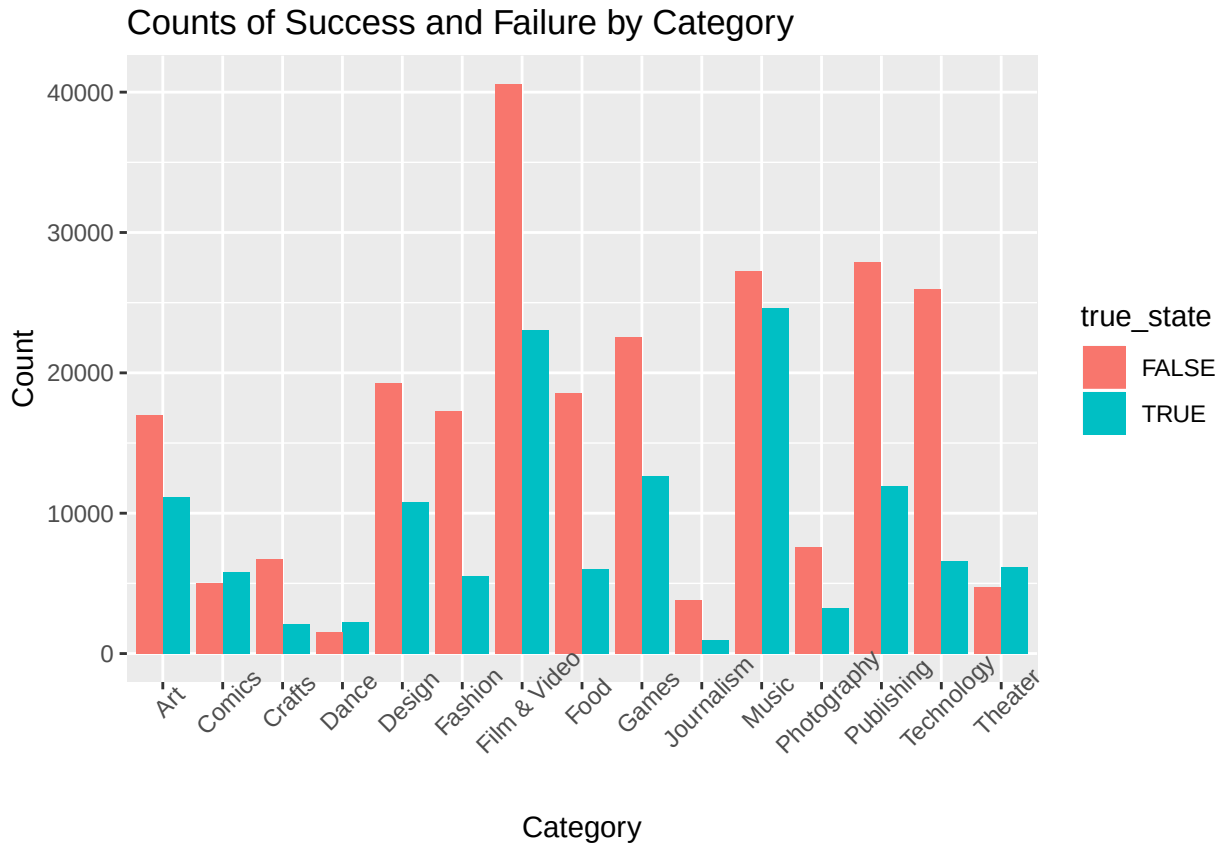
```r
ggplot(usd_pledged_data, (aes(x= name,y = proportion_of_data))) + geom_point() + labs(x = "Amount Pledg
```



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

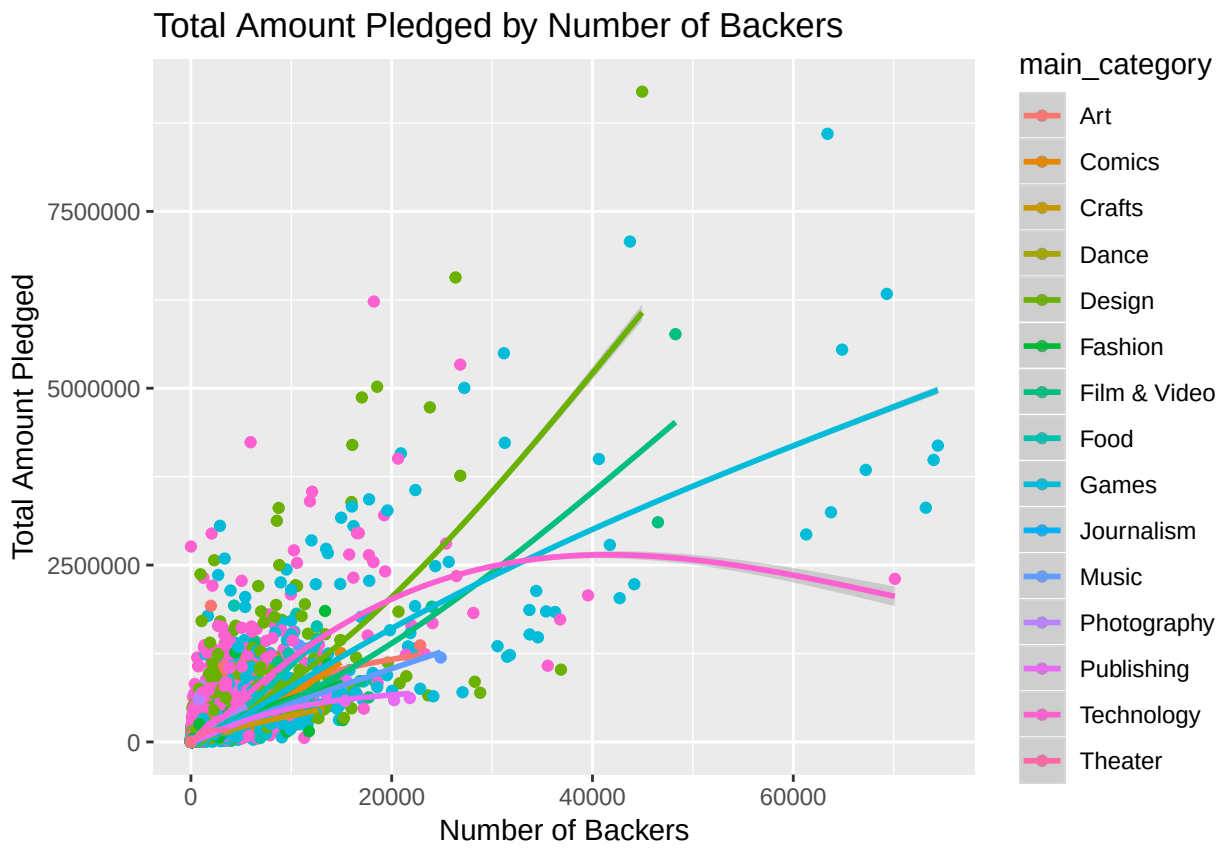
```
kickstarters2 %>% ggplot(aes(fill = true_state, x = main_category)) + geom_bar(position = "dodge") + th
```



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 = 
```

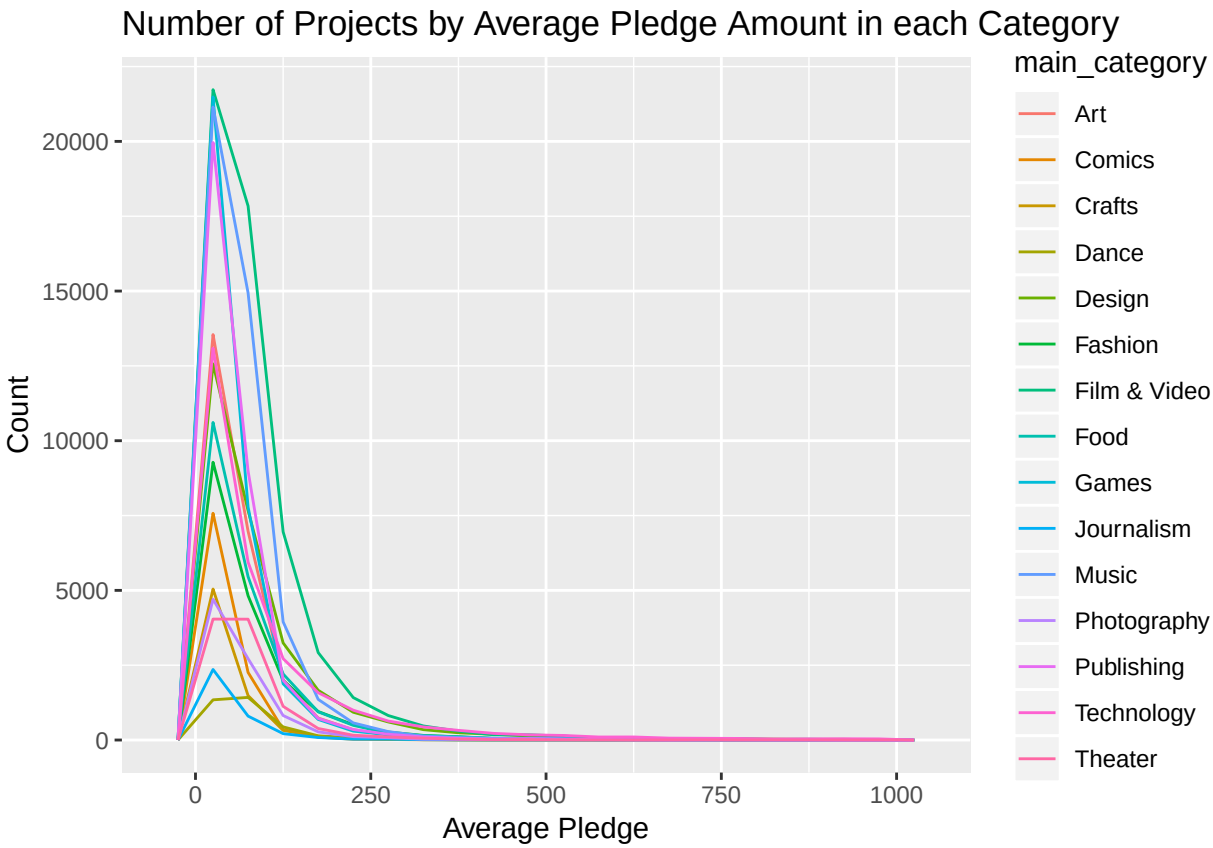
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




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

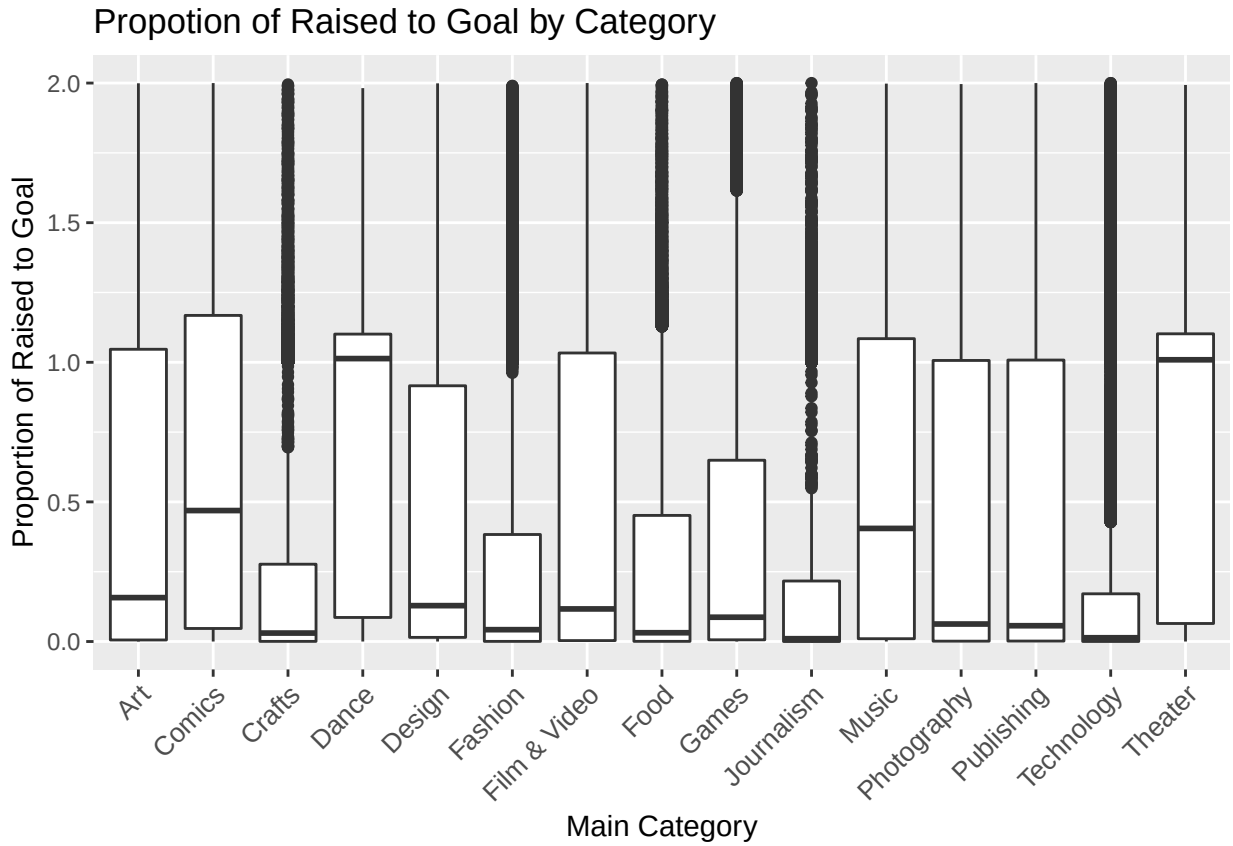
From this scatter plot, we can see that in general there is a positive correlation between the number of backers and the the total amount pledged. Data was filtered for pledged totals less than 10 million total dollars and number of backers less than 75000 to better show data. As shown by these two data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



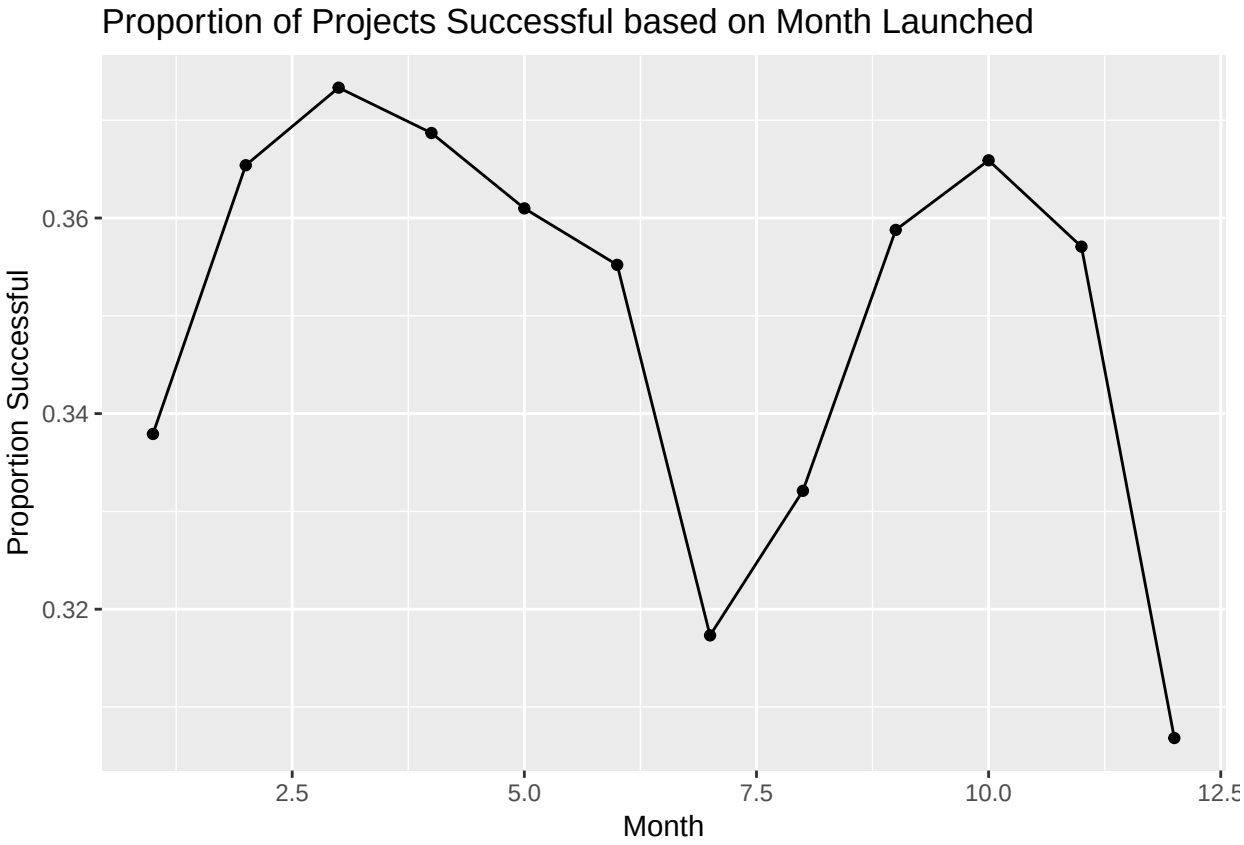
The filters used in creating this graph filtered out all kickstarters that had no backers or no amount pledged. To create the histogram, we also filtered out the 862 rows of data with an average pledge of greater than 1000, to allow us to view the graph better. From this histogram we find that the most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

```
KSDataFiltered2 <- kickstarters2 %>% mutate(prop_raised_goal = usd_pledged_real/usd_goal_real) %>% filt
#KSDataFiltered2

ggplot(KSDataFiltered2) + geom_boxplot(aes(x = main_category, y = prop_raised_goal)) + theme(axis.text.
  labs(title = "Propotion of Raised to Goal by Category", x = "Main Category", y = "Proportion of Raise
```



We chose to filter out data points with a proportion of raised to goal greater than 2 as there were many outliers with a proportion reaching as high as 10^8. Upon inspection we found these data points were from Kickstarters where the project owner set the goal to $1 and proceeded to raise millions. Doing this filtered out ~20,000 entries of the 378,654 entries. The majority have a mean proportion of success less than 25%, however some groups, such as Dance, Theater, Comics, and Music have a higher proportion of successess. Most of the categories have a 75th quartile greater than 1.0, indicating that at least 25% of the projects succeed.

```
month<- kickstarters %>% mutate(monthlaunched = str_sub(launched, start = 6, end = 7)) %>% mutate(true_

month_prop <- month %>% mutate(true_state = (usd_pledged_real / usd_goal_real)>1)%>% group_by(monthlaun
month_prop2 <- month %>% group_by(monthlaunched) %>% count(monthlaunched)
month_prop3 <- merge(x = month_prop, y = month_prop2, by = "monthlaunched", all = TRUE)
month_prop3 <- month_prop3 %>% mutate(proportion_success = n.x/n.y) %>% filter(true_state == TRUE)
month_prop3 %>% ggplot(aes(x=as.integer(monthlaunched),y=proportion_success)) + geom_line() + geom_poin
```



By observing the proportions of successes by month we can see a significant drop in July and December/January. Thus could suggest more failures around holiday/vacation months for kickstarter projects that were launched during those months.

```
year_prop <- kickstarters2 %>% mutate(yearlaunched = str_sub(launched, start = 1, end = 4)) %>% group_by
year_prop2 <- kickstarters2 %>% mutate(yearlaunched = str_sub(launched, start = 1, end = 4)) %>% group_b
year_prop3 <- merge(x = year_prop, y = year_prop2, by = "yearlaunched", all = TRUE)
year_prop3 <- year_prop3 %>% mutate(proportion_success = n.x/n.y) %>% filter(true_state == TRUE)

year_prop3 %>% ggplot(aes(x=as.integer(yearlaunched),y=proportion_success, color = true_state)) + geom_l
```

There is an overall trend of decreased proportion project success with increased year of launch. The launch year with highest proportion of success in our dataset is 2011. The launch year with lowest proportion of success in our dataset is 2018. This could be indicative of a decreased public interest in kickstarter projects with time.

Discussion