Kickstarter Analysis Proposal

10/9/2020

The Laureates: Farzeen Najam, Olivia Olsher, Vincent Liu, Emile Therrien

Notes for our group moving forward: to be included along with rough draft comments

Need to change the first data analysis to pledged over real because the "state" variable has too many qualifiers that might neglect certain succesful obs even though they weren't listed as "successful"

In Emile and Olivia's part, exclude the live data

For Farzeen: need to explain why you chose certain variables and how you then incorporated those into your hypo test -why omit tier 7?

Emile and olivia: need to see what category is most likely to be funded what should the size of sample for the bootstraps be? use regression to model this? what sort of regression?

Introduction

Have you ever had an ambitious idea but did not have resources to pursue it? Kickstarter is a crowdfunding platform for creators who need support for their projects. Since launching in 2009, 19 million people have pledged to back up projects, and nearly 190,000 projects have been successfully funded.

However, many more projects failed to reach their goals, and our group is interested in analyzing what made projects succeed and fail. The dataset that we use comes from Kaggle, which contains Kickstarter projects up until January 2018.

We will be exploring which variables influence the success of a Kickstarter project by observing which types of projects are more likely to be funded. Our questions include:

- Does the amount of money a project asks for influence it's chance of success?
- Does the category of project influence it's chance of success?

Our hypotheses include:

- The more money the project asks for, the less successful it will be in terms of getting funding.
- The category "Technology" will be the most likely funded type of projects.

Data description

This data set contains 15 variables and 378,661 observations, where each observation is one kickstarter project.

The categorical variables include the name of each project, the category of each project (music, narrative film, restaurant, etc.), a broader category of each (food, film, publishing, etc.), the crowdsourcing currency, the state of each project (failed, successful, or cancelled), and the country of origin for each project.

The numerical data are the project's ID number, the monetary goal for each, the money pledged to each project, how many backers of each project; there are three numerical variables that are not self-explanatory, usd_pledged: conversion in US dollars of the pledged column (conversion done by kickstarter), usd_pledge_real: conversion in US dollars of the pledged column (conversion from Fixer.io API), and usd_goal_real: conversion in US dollars of the goal column (conversion from Fixer.io API).

There are also two date columns, one for the project launch and the other for the crowdsourcing deadline.

The data were collected from Kickstarter Platform likely using web scraping methods on their own site, to be used by data scientists to model whether or not a project will be successful or not when it is launched.

Glimpse of data

```
library(tidyverse)

kickstarter <- read_csv("data/ks-projects-201801.csv")
glimpse(kickstarter)
```

```
## Observations: 378,661
## Variables: 15
## $ ID               <dbl> 1000002330, 1000003930, 1000004038, 1000007540, 10...
## $ name             <chr> "The Songs of Adelaide & Abullah", "Greeting From ...
## $ category         <chr> "Poetry", "Narrative Film", "Narrative Film", "Mus...
## $ main_category    <chr> "Publishing", "Film & Video", "Film & Video", "Mus...
## $ currency         <chr> "GBP", "USD", "USD", "USD", "USD", "USD", "USD", "...
## $ deadline         <date> 2015-10-09, 2017-11-01, 2013-02-26, 2012-04-16, 2...
## $ goal             <dbl> 1000, 30000, 45000, 5000, 19500, 50000, 1000, 2500...
## $ launched         <dttm> 2015-08-11 12:12:28, 2017-09-02 04:43:57, 2013-01...
## $ pledged          <dbl> 0.00, 2421.00, 220.00, 1.00, 1283.00, 52375.00, 12...
## $ state            <chr> "failed", "failed", "failed", "failed", "canceled"...
## $ backers          <dbl> 0, 15, 3, 1, 14, 224, 16, 40, 58, 43, 0, 100, 0, 0...
## $ country          <chr> "GB", "US", "US", "US", "US", "US", "US", "US", "U...
## $ `usd pledged`    <dbl> 0.00, 100.00, 220.00, 1.00, 1283.00, 52375.00, 120...
## $ usd_pledged_real <dbl> 0.00, 2421.00, 220.00, 1.00, 1283.00, 52375.00, 12...
## $ usd_goal_real    <dbl> 1533.95, 30000.00, 45000.00, 5000.00, 19500.00, 50...
```

Explorative Data Analysis

Overview of project state:

```
kickstarter %>%
  group_by(state) %>%
  count()
```

```
## # A tibble: 6 x 2
## # Groups:   state [6]
##   state           n
##   <chr>       <int>
## 1 canceled    38779
## 2 failed     197719
## 3 live         2799
## 4 successful 133956
```

```
## 5 suspended    1846
## 6 undefined    3562
```

```
kickstarter %>%
  filter(state == "successful") %>%
  summarize(success_rate = n()/nrow(kickstarter))
```

```
## # A tibble: 1 x 1
##   success_rate
##          <dbl>
## 1        0.354
```

Overview of usd dollars pledged and goal:

```
overview_usd_pledged_real <- kickstarter %>%
  summarize(mean_usd_pledged_real = mean(usd_pledged_real),
            median_usd_pledged_real = median(usd_pledged_real),
            sd_usd_pledged_real = sd(usd_pledged_real))

overview_usd_goal_real <- kickstarter %>%
  summarize(mean_usd_goal_real = mean(usd_goal_real),
            median_usd_goal_real = median(usd_goal_real),
            sd_usd_goal_real = sd(usd_goal_real))

overview_usd_pledged_real
```

```
## # A tibble: 1 x 3
##   mean_usd_pledged_real median_usd_pledged_real sd_usd_pledged_real
##                   <dbl>                   <dbl>               <dbl>
## 1                 9059.                    624.              90973.
```

```
overview_usd_goal_real
```

```
## # A tibble: 1 x 3
##   mean_usd_goal_real median_usd_goal_real sd_usd_goal_real
##                <dbl>                <dbl>            <dbl>
## 1              45454.                 5500         1152950.
```

Project Goal & Success

Question: Is the project goal (amount of money a project asks for) associated with whether it is successful?

Hypotheses:

- H_0: There is no relationship between a project's goal and whether a project is successful or not.
- H_1: There is a relationship between a project's goal and whether a project is successful or not.

Create a new data frame with projects that are no longer live:

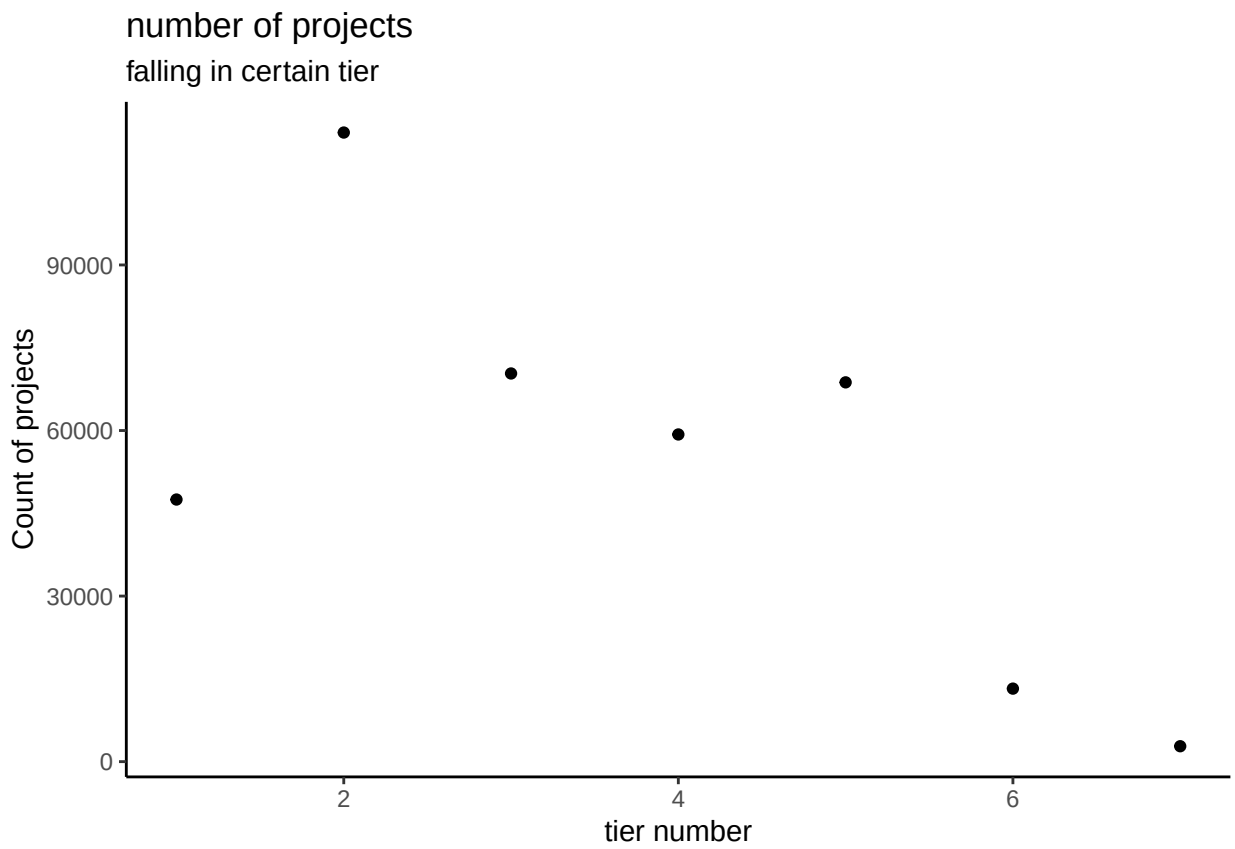
```r
# Exclude ongoing projects
kickstarter_not_live <- kickstarter %>%
  filter(state != "live")
```

Create a new variable named usd_goal_real_tier to classify usd_goal_real into tiers:

```r
# perhaps need a better way to categorize values in this column
kickstarter_not_live <- kickstarter_not_live %>%
  mutate(usd_goal_real_tier = case_when(
    usd_goal_real < 1000 ~ 1,
    usd_goal_real >= 1000 & usd_goal_real < 5000 ~ 2,
    usd_goal_real >= 5000 & usd_goal_real < 10000 ~ 3,
    usd_goal_real >= 10000 & usd_goal_real < 20000 ~ 4,
    usd_goal_real >= 20000 & usd_goal_real < 100000 ~ 5,
    usd_goal_real >= 100000 & usd_goal_real < 500000 ~ 6,
    usd_goal_real >= 500000 ~ 7,
  ))

chart<- kickstarter_not_live %>%
  group_by(usd_goal_real_tier) %>%
  summarize(numbers = n())

chart
```

```
## # A tibble: 7 x 2
##   usd_goal_real_tier numbers
##                <dbl>   <int>
## 1                  1   47494
## 2                  2  113993
## 3                  3   70338
## 4                  4   59285
## 5                  5   68727
## 6                  6   13231
## 7                  7    2794
```

```r
ggplot(data = chart, mapping = aes((x = usd_goal_real_tier), y = numbers)) +
 geom_point () +
  scale_colour_viridis_d()+
  labs(title = "number of projects",
       subtitle = "falling in certain tier",
       x = "tier number", y = "Count of projects") + theme_classic()
```



Create a new variable for success state

```
kickstarter_not_live <- kickstarter_not_live %>%
  mutate(success_state = if_else(state == "successful", 1, 0))
```

We'll run a chi-square test at alpha = 0.05.

```
chisq.test(table(kickstarter_not_live$success_state, kickstarter_not_live$usd_goal_real_tier))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(kickstarter_not_live$success_state, kickstarter_not_live$usd_goal_real_tier)
## X-squared = 19624, df = 6, p-value < 2.2e-16
```

Our test statistic was 19624, which has a chi-square distribution with 6 degree of freedom under H_0. This corresponds to a p-value less than 2.2e-16. So our decision is to reject H_0, and there is sufficient evidence that there is a relationship between a project's monetary goal and a project's state of success.

Project Goal Amount & Success

Claim: The more money the project asks for, the less successful it will be in terms of getting funding.

H_0: $S_{less-money} \leq S_{more-money}$ H_1: $S_{less-money} > S_{more-money}$

H_0:proportion of people who successful with low money goal - proportion of people who successful with high money <= 0 H_1: proportion of people who successful with low money goal - proportion of people who successful with high money > 0

$\alpha = 0.05$

```
set.seed(2)
     kick1 <- kickstarter_not_live %>%
       filter(usd_goal_real_tier == 6 | usd_goal_real_tier == 5 | usd_goal_real_tier ==4)

     kick2 <- kickstarter_not_live %>%
        filter(usd_goal_real_tier == 1 | usd_goal_real_tier == 2 | usd_goal_real_tier ==3)

n_sims <- 500
boot_dist <- numeric(n_sims)
for(i in 1:n_sims){

  # create indices
  indices_m <- sample(1:nrow(kick1), replace = T)
  indices_l <- sample(1:nrow(kick2), replace = T)


  # bootstrap est. group means
  temp_m <- kick1 %>%
    slice(indices_m) %>%
    filter(success_state == 1) %>%
    summarize(prop_m = n()/ 141243) %>%
    select(prop_m) %>%
    pull()


  temp_l <-kick2  %>%
    slice(indices_l) %>%
    filter(success_state == 1) %>%
    summarize(prop_s= n()/231825) %>%
    select(prop_s) %>%
    pull()

  boot_dist[i] <-   temp_l - temp_m   #diff btw low goal success high goal success
}
boot_diffs <- tibble(diffs = boot_dist)


obs_diff <- boot_diffs %>%
  summarize(obs_diff = mean(diffs)) %>%
  pull()

offset <- boot_diffs %>%
  summarize(offset = 0 - mean(diffs)) %>%
  pull()
null_dist <- boot_diffs %>%
  mutate(centered_diffs = diffs + offset) %>%
  select(centered_diffs)
```

```
null_dist %>%
  mutate(extreme = ifelse(centered_diffs > abs(obs_diff), 1, 0)) %>%
  summarize(p_val = mean(extreme))
```

```
## # A tibble: 1 x 1
##   p_val
##   <dbl>
## 1     0
```

```
boot_diffs %>%
  summarize(lower = quantile(diffs, 0.025),
            upper = quantile(diffs, 0.975))
```

```
## # A tibble: 1 x 2
##   lower upper
##   <dbl> <dbl>
## 1 0.178 0.184
```

Since our p-value is less than our alpha value, we reject our null hypothesis that the difference in the proportion of projects with less money as goal and proportion of projects with more money as goal is less than or equal to 0.This means we reject the null hypothesis that success rate for projects with bigger money is equal or higher than projects with lower money.

Project Category and Success

- Is the category associated with the chance of success a project has?
- The category "Technology" will be the most likely funded type of project.

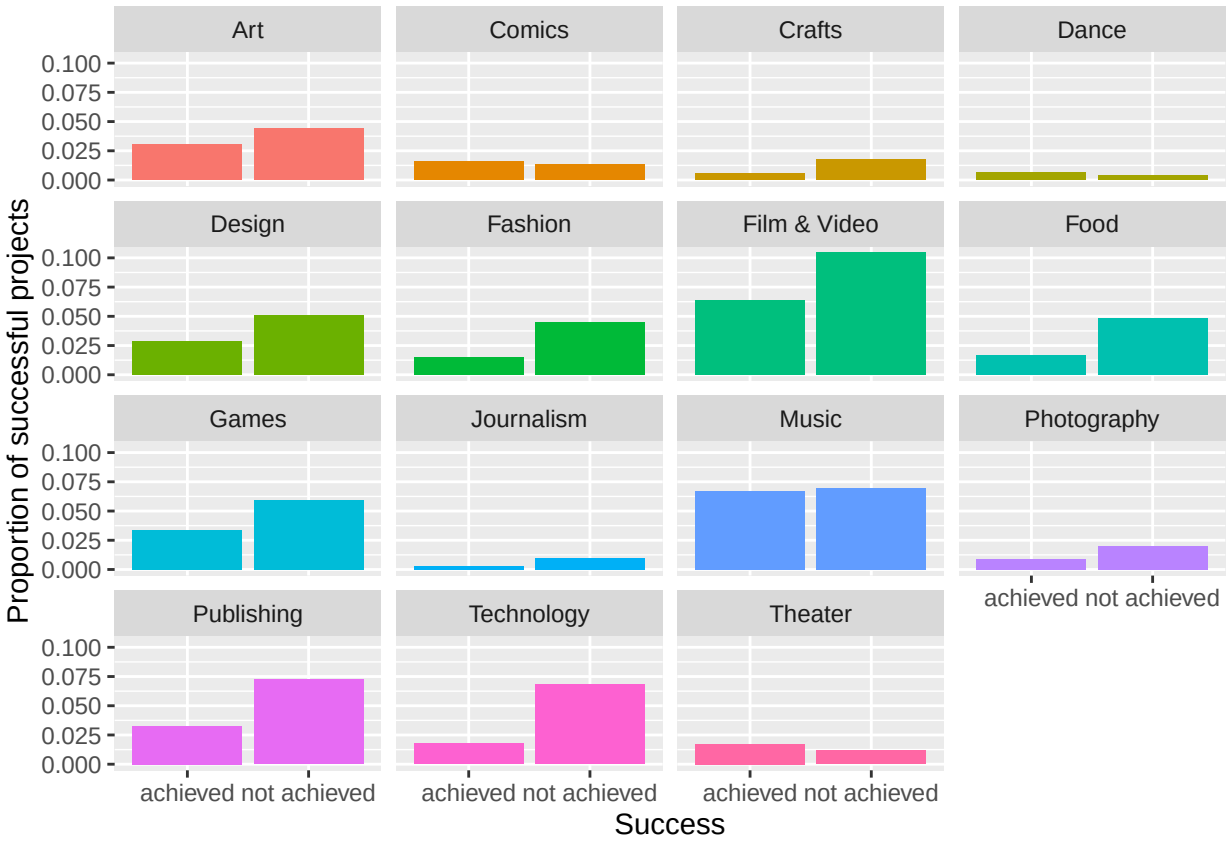
limitations: possible crossover in the main_categories that we cannot screen for

questions for Yue: * how to weight each bar graph the same to show proportionality between main categories? * how not to hard code sum in kickstarter_success_ratio pipeline

```
kickstart_success = kickstarter %>%
  mutate(
    project_funding_ratio = pledged/goal,
    success = case_when(
      project_funding_ratio >= 1 ~ "achieved",
      project_funding_ratio < 1 ~ "not achieved"
    )
  )

kickstater_success_ratios = kickstart_success %>%
  group_by(success, main_category) %>%
  count(success) %>%
  mutate(
    sum = 137042+241619
  )%>%
  mutate(
    proportion_achieved = n/sum
  )

kickstater_success_ratios
```

```
## # A tibble: 30 x 5
## # Groups:   success, main_category [30]
##    success  main_category     n    sum proportion_achieved
##    <chr>    <chr>         <int>  <dbl>               <dbl>
##  1 achieved Art           11601 378661             0.0306
##  2 achieved Comics         5875 378661             0.0155
##  3 achieved Crafts         2138 378661             0.00565
##  4 achieved Dance          2340 378661             0.00618
##  5 achieved Design        10870 378661             0.0287
##  6 achieved Fashion        5690 378661             0.0150
##  7 achieved Film & Video  24063 378661             0.0635
##  8 achieved Food           6178 378661             0.0163
##  9 achieved Games         12747 378661             0.0337
## 10 achieved Journalism     1026 378661             0.00271
## # ... with 20 more rows
```

```
ggplot(data = kickstater_success_ratios, aes(x = success,
                                             y = proportion_achieved,
                                             fill = main_category)) +
  facet_wrap(. ~ main_category) +
  geom_col() +
  theme(legend.position = "none") +
  labs(x = "Success",
       y = "Proportion of successful projects")
```



main_category vs. success

H_0: There is no relationship between main_category and success.

H_1: There is a relationship between main_category and success.

Where the $alpha/ = 0.05$

```
chisq.test(table(kickstart_success$main_category, kickstart_success$success))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(kickstart_success$main_category, kickstart_success$success)
## X-squared = 17099, df = 14, p-value < 2.2e-16
```

Given that the p-value is less than $alpha/ = 0.05$, we can reject the null hypothesis. Thus, there is sufficient evidence to suggest that there is a relationship between main category and success.

Simulation, tech is the most likely to be funded

H_0: $p_{s-tech} \leq p_{s-other}$ H_1: $p_{s-tech} > p_{s-other}$

```
set.seed(1)

kickstart_boot = kickstart_success %>%
  mutate(
    TECH = case_when(
      main_category == "Technology" ~ "TECH",
      main_category != "Technology" ~ "other"
    )
  )
kickstart_boot
```

```
## # A tibble: 378,661 x 18
##        ID name  category main_category currency deadline     goal
##     <dbl> <chr> <chr>    <chr>         <chr>    <date>      <dbl>
##  1 1.00e9 The ~ Poetry   Publishing    GBP      2015-10-09   1000
##  2 1.00e9 Gree~ Narrati~ Film & Video  USD      2017-11-01  30000
##  3 1.00e9 Wher~ Narrati~ Film & Video  USD      2013-02-26  45000
##  4 1.00e9 Tosh~ Music    Music         USD      2012-04-16   5000
##  5 1.00e9 Comm~ Film & ~ Film & Video  USD      2015-08-29  19500
##  6 1.00e9 Mona~ Restaur~ Food          USD      2016-04-01  50000
##  7 1.00e9 Supp~ Food     Food          USD      2014-12-21   1000
##  8 1.00e9 Chas~ Drinks   Food          USD      2016-03-17  25000
##  9 1.00e9 SPIN~ Product~ Design        USD      2014-05-29 125000
## 10 1.00e8 STUD~ Documen~ Film & Video  USD      2014-08-10  65000
## # ... with 378,651 more rows, and 11 more variables: launched <dttm>,
## #   pledged <dbl>, state <chr>, backers <dbl>, country <chr>, `usd
## #   pledged` <dbl>, usd_pledged_real <dbl>, usd_goal_real <dbl>,
## #   project_funding_ratio <dbl>, success <chr>, TECH <chr>
```

```
obs_diff = kickstart_boot %>%
  group_by(TECH) %>%
  count(success) %>%
```

```
  mutate(prop = n/sum(n)) %>%
  filter(success == "achieved") %>%
  select(TECH, prop)
obs_diff
```

```
## # A tibble: 2 x 2
## # Groups:   TECH [2]
##   TECH   prop
##   <chr> <dbl>
## 1 other 0.377
## 2 TECH  0.206
```

```
diff = obs_diff[2,2] - obs_diff[1,2] #take second row and second column - ...
diff
```

```
##         prop
## 1 -0.1708602
```

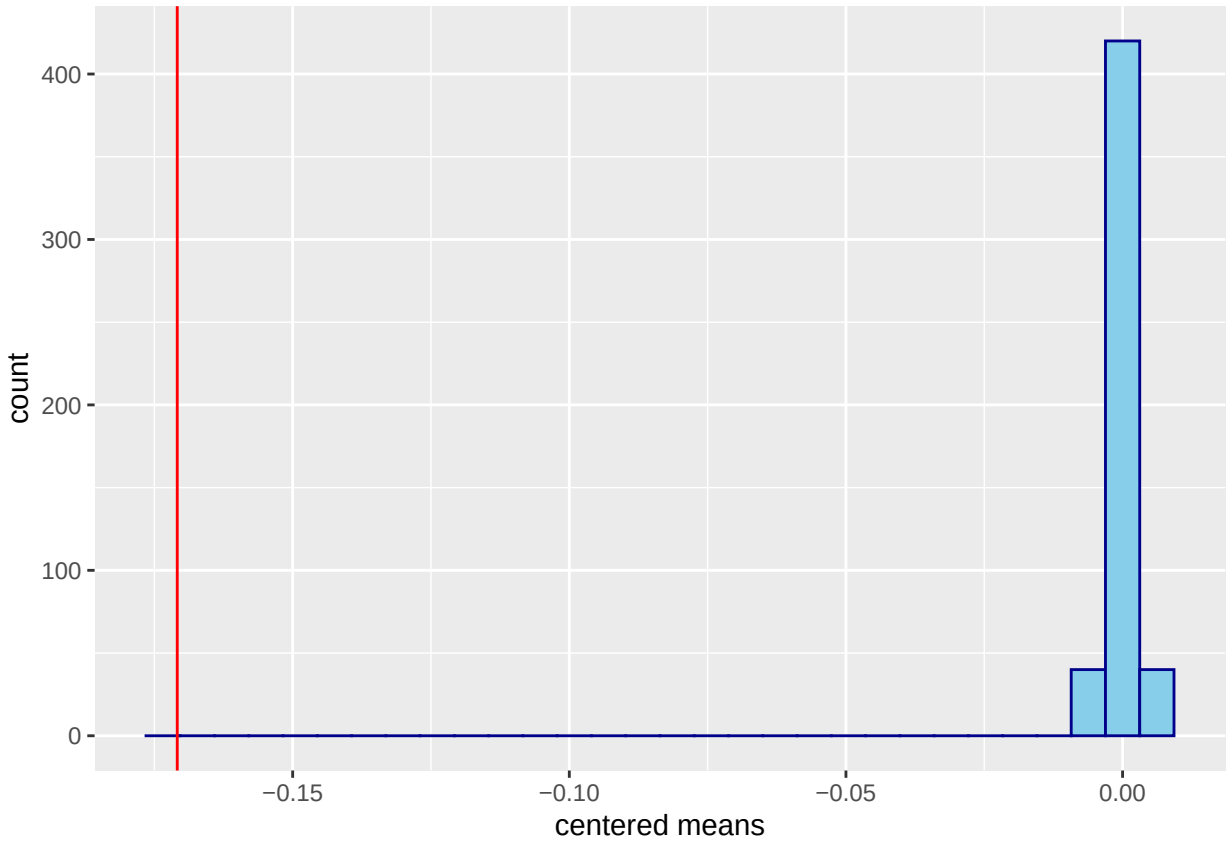
```
tech = kickstart_boot %>%
  filter(TECH == "TECH")
other = kickstart_boot %>%
  filter(TECH != "TECH")

boot_samp <- numeric(500)

for(i in 1:500){
  boot_tech <- tech %>%
    slice(sample(1:nrow(tech), replace = T)) %>%
    count(success) %>%
    mutate(prop = n/sum(n)) %>%
    select(prop) %>%
    slice(1) %>%
    pull()
  boot_other<- other %>%
    slice(sample(1:nrow(other), replace = T)) %>%
    count(success) %>%
    mutate(prop = n/sum(n)) %>%
    select(prop) %>%
    slice(1) %>%
    pull()

  boot_samp[i] <- boot_tech - boot_other
}

#mean of bootstrap samples:
mean_boot = mean(boot_samp)
boot_samp = tibble(boot_samp)

boot_samp <- boot_samp %>%
  mutate(centered = boot_samp + abs(mean_boot))

boot_samp %>%
```

```
  mutate(extreme = if_else(centered > abs(diff), 0, 1)
  ) %>%
  summarize(p_val = mean(extreme))
```

```
## # A tibble: 1 x 1
##   p_val
##   <dbl>
## 1     1
```

```
ggplot(data = boot_samp, aes(x = centered)) +
  geom_histogram(color = "darkblue", fill = "skyblue") +
  labs(x = "centered means", y = "count") +
  geom_vline(xintercept = diff$prop, color = "red")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



We have a p-value of 1, which means there is not sufficient evidence at the $\alpha = 0.05$ level to reject the null hypothesis that the proportion of success for main categories other than Technology is the same or greater than the proportion of sucess for Technology.