Kickstarter Project Analysis

AUTHOR
Phiphat Chayasil

Background

https://www.kickstarter.com Links to an external site. is a crowdfunding site, where people make a pitch for a project and an amount they need to raise to do the project. Then users can pledge support, and if users pledge enough support for a project, then the project gets the pledged money. If users do not pledge enough support, then the users keep their money and the project gets no money. The dataset comes from https://www.kaggle.com/datasets/kemical/kickstarter-projects?select=ks-projects-201801.csv.

Loading Libraries

```
# Load all the necessary libraries
library(tidyverse)
library(dplyr)
library(ggplot2)
library(plotly)
library(scales)
```

Loading Data

As previously stated, the data can be downloaded from https://www.kaggle.com/datasets/kemical/kickstarter-projects?select=ks-projects-201801.csv.

```
ks_data <- read.csv('ks-projects-201801.csv')

# Example of our dataset
glimpse(ks_data)
```

```
Rows: 378,661
Columns: 15
$ ID               <int> 1000002330, 1000003930, 1000004038, 1000007540, 10000…
```

```
$ name             <chr> "The Songs of Adelaide & Abullah", "Greeting From Ear…
$ category         <chr> "Poetry", "Narrative Film", "Narrative Film", "Music"…
$ main_category    <chr> "Publishing", "Film & Video", "Film & Video", "Music"…
$ currency         <chr> "GBP", "USD", "USD", "USD", "USD", "USD", "USD", "USD…
$ deadline         <chr> "2015-10-09", "2017-11-01", "2013-02-26", "2012-04-16…
$ goal             <dbl> 1000, 30000, 45000, 5000, 19500, 50000, 1000, 25000, …
$ launched         <chr> "2015-08-11 12:12:28", "2017-09-02 04:43:57", "2013-0…
$ pledged          <dbl> 0.00, 2421.00, 220.00, 1.00, 1283.00, 52375.00, 1205.…
$ state            <chr> "failed", "failed", "failed", "failed", "canceled", "…
$ backers          <int> 0, 15, 3, 1, 14, 224, 16, 40, 58, 43, 0, 100, 0, 0, 7…
$ country          <chr> "GB", "US", "US", "US", "US", "US", "US", "US", "US",…
$ usd.pledged      <dbl> 0.00, 100.00, 220.00, 1.00, 1283.00, 52375.00, 1205.0…
$ usd_pledged_real <dbl> 0.00, 2421.00, 220.00, 1.00, 1283.00, 52375.00, 1205.…
$ usd_goal_real    <dbl> 1533.95, 30000.00, 45000.00, 5000.00, 19500.00, 50000…
```

Data Processing

```
# add 4 new columns with deadline and launch dates and years
ks_data <- ks_data |>
  mutate(
    deadline_date = as.Date(deadline),
    launched_date = as.Date(launched),
    deadline_year = format(deadline_date, '%Y'),
    launched_year = format(launched_date, '%Y')
  )

# See what the dataset now looks like
glimpse(ks_data)
```

```
Rows: 378,661
Columns: 19
$ ID               <int> 1000002330, 1000003930, 1000004038, 1000007540, 10000…
$ name             <chr> "The Songs of Adelaide & Abullah", "Greeting From Ear…
$ category         <chr> "Poetry", "Narrative Film", "Narrative Film", "Music"…
$ main_category    <chr> "Publishing", "Film & Video", "Film & Video", "Music"…
$ currency         <chr> "GBP", "USD", "USD", "USD", "USD", "USD", "USD", "USD…
$ deadline         <chr> "2015-10-09", "2017-11-01", "2013-02-26", "2012-04-16…
$ goal             <dbl> 1000, 30000, 45000, 5000, 19500, 50000, 1000, 25000, …
$ launched         <chr> "2015-08-11 12:12:28", "2017-09-02 04:43:57", "2013-0…
$ pledged          <dbl> 0.00, 2421.00, 220.00, 1.00, 1283.00, 52375.00, 1205.…
$ state            <chr> "failed", "failed", "failed", "failed", "canceled", "…
$ backers          <int> 0, 15, 3, 1, 14, 224, 16, 40, 58, 43, 0, 100, 0, 0, 7…
$ country          <chr> "GB", "US", "US", "US", "US", "US", "US", "US", "US",…
```

```
$ usd.pledged      <dbl> 0.00, 100.00, 220.00, 1.00, 1283.00, 52375.00, 1205.0…
$ usd_pledged_real <dbl> 0.00, 2421.00, 220.00, 1.00, 1283.00, 52375.00, 1205.…
$ usd_goal_real    <dbl> 1533.95, 30000.00, 45000.00, 5000.00, 19500.00, 50000…
$ deadline_date    <date> 2015-10-09, 2017-11-01, 2013-02-26, 2012-04-16, 2015…
$ launched_date    <date> 2015-08-11, 2017-09-02, 2013-01-12, 2012-03-17, 2015…
$ deadline_year    <chr> "2015", "2017", "2013", "2012", "2015", "2016", "2014…
$ launched_year    <chr> "2015", "2017", "2013", "2012", "2015", "2016", "2014…
```

Initial Calculations

```
# Display how many projects there are
total_projects <- nrow(ks_data)

# Filter out the successful projects
success_ones <- ks_data |>
  filter(state == 'successful')

# Filter out the failed projects
failed_ones <- ks_data |>
  filter(state == 'failed')

# the ratio of failed projects
success_ratio <- nrow(success_ones)/total_projects
fail_ratio <- nrow(failed_ones)/total_projects
```

As per our initial calculations, there are 378661 projects in total. 133956 are marked successful, and 197719 are marked failed. The ratios of successful and failed projects are 0.35 and 0.52 respectively.

Biggest Non-Success

Now we are interested in finding the biggest non-success project. This is to find the project that is marked anything other than “successful” with the highest “usd_pledged_real” value.

```
# Filter out the row with the most 'usd_pledged_real' with state as anything but
biggest_nonsuccess <- ks_data |>
  filter(state != 'successful') |>
  filter(usd_pledged_real == max(usd_pledged_real))
```

The biggest non-success project is The Skarp Laser Razor: 21st Century Shaving (Suspended).

The Skarp Laser Razor: 21st Century Shaving

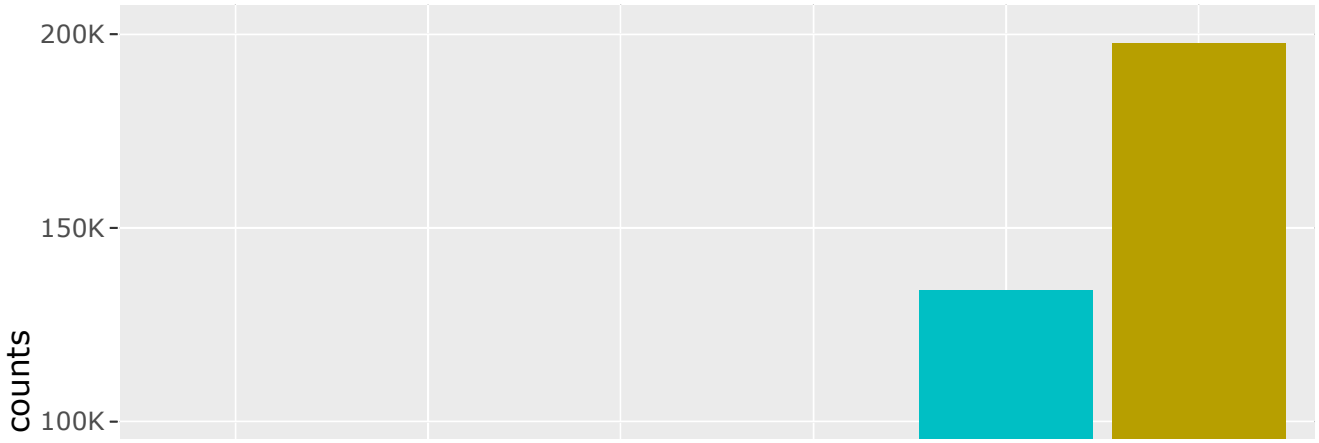
Found by Morgan Gustavasson and Paul Binun, the project was proposed by Skarp Technologies, Inc. in 2025 to raise funding for the 'Skarp Laser Razor' developement. The company claimed that its razor blades was superior as it could cut hair in a cellular level which would cause zero irritation. The razor was expected to be powered only a AAA battery, and the razor blade would be disposable and replaceable. According to our data, the project raised over 4 million dollars before it got suspended by Kickstart, due to not having a working prototype.

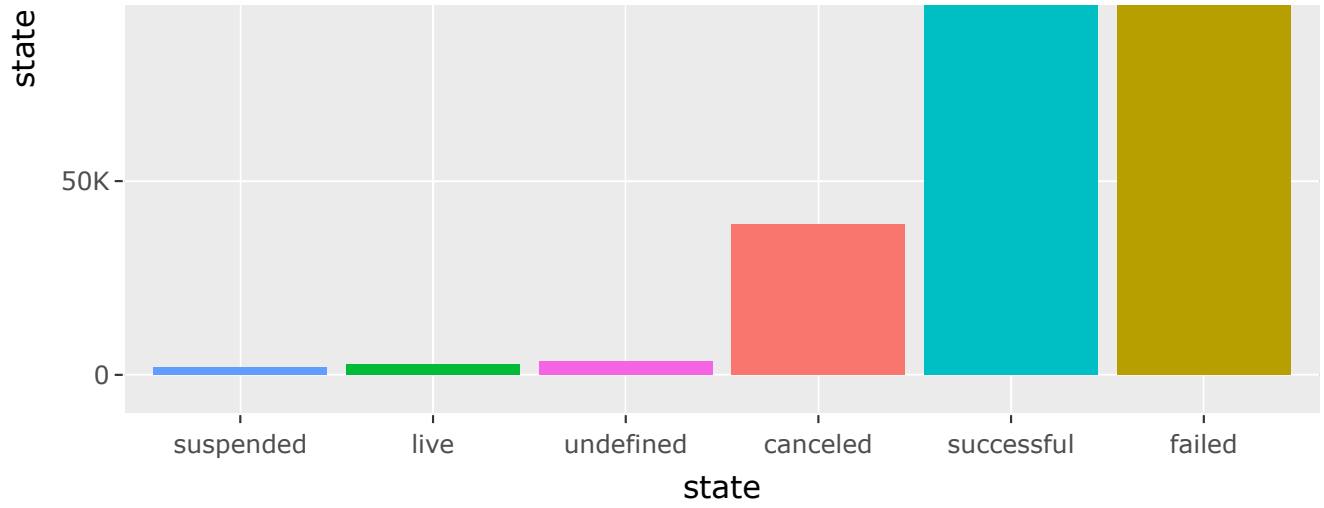
Project State

Now we would like to visualize how many projects have each state.

```
# Group the data by state column and summarize how many there are for each state
project_state <- ks_data |>
  group_by(state) |>
  summarize(state_count = n())
# Plot a bar chart with respect state counts in ascending order
state_plot <- project_state |>
  ggplot() +
  geom_col(aes(x = reorder(state, +state_count),
               y = state_count,
               fill = state,
               text = state)) +
  labs('Kickstart Projects Count', x = 'state', y = 'state counts') +
  scale_y_continuous(labels = label_number(scale_cut = cut_short_scale()))

# Make the plot and interractive one
ggplotly(state_plot, tooltip = 'text')
```





Based on the visualization, it appears that there are more failed projects than successful ones. However, the difference is not significant, compared to the total count of all the projects combined. The bars for live and suspended projects also are almost identical in height. Even though cancelled projects are much taller, but it is still far significant from successful and failed bars. We may conclude that it is not easy for a project to be cancelled, and it is even more unlikely for a project to be suspended.

Yearly Summary

Ultimately, we would like to summarize out data in each year. We will be using deadline_year column in this section. The successful rate will also be determined by projects with successful state versus other projects with otherwise states.

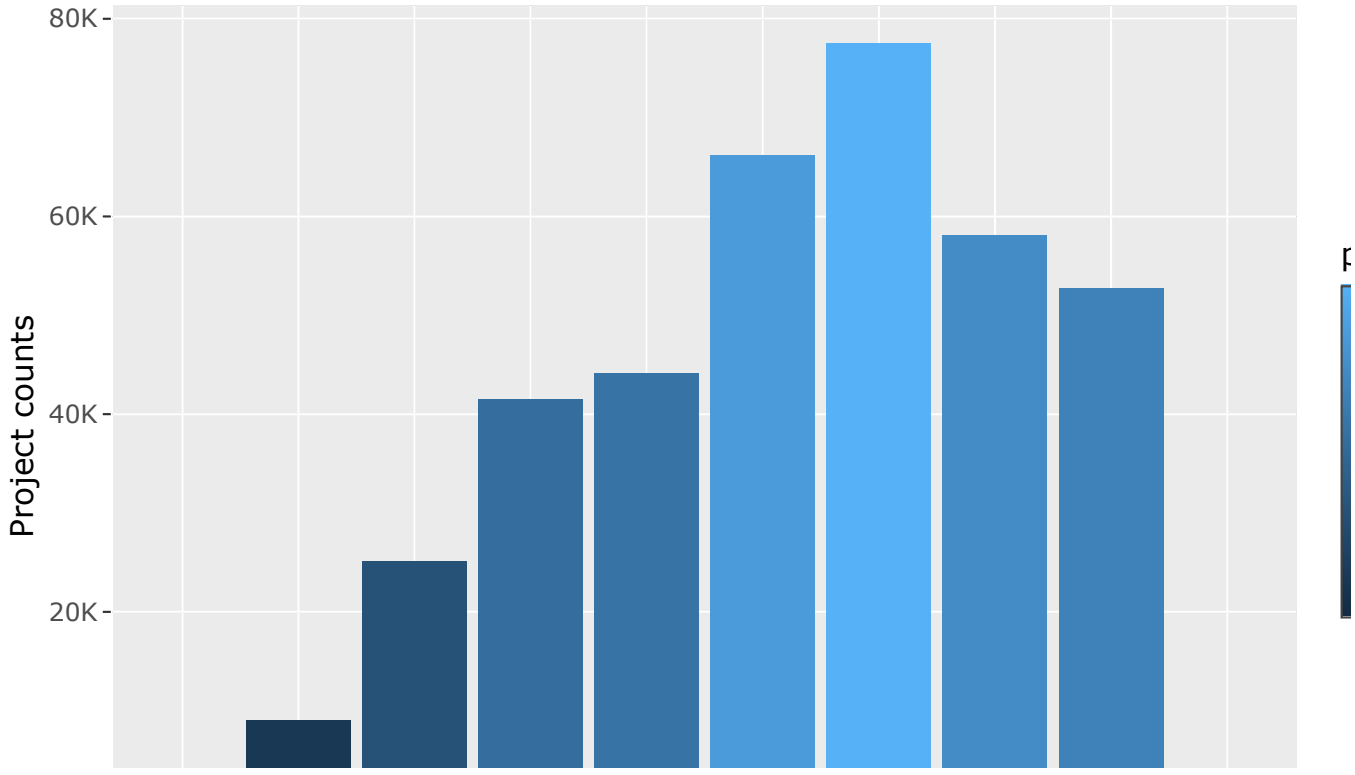
```
# Group the data by deadline_year and make new columns by using summarize()
yearly_summary <- ks_data |>
  group_by(deadline_year) |>
  summarize(project_count = n(),
            percent_success = mean(state == 'successful'),
            avg_fund_raise = mean(usd_pledged_real),
            max_fund_raise = max(usd_pledged_real))
print(yearly_summary)
```

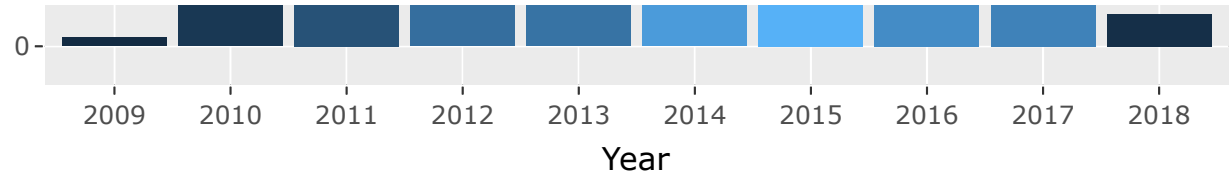
```
# A tibble: 10 × 5
   deadline_year project_count percent_success avg_fund_raise max_fund_raise
   <chr>                <int>           <dbl>          <dbl>          <dbl>
 1 2009                   902           0.426          2039.         84614.
 2 2010                  9098           0.441          2740.        942578.
 3 2011                 25107           0.469          3708.        830828.
 4 2012                 41560           0.435          7551.      10266846.
 5 2013                 44178           0.435         10653.       5702153.
```

6	2014	66231	0.320	7933.	13285226.
7	2015	77498	0.272	8847.	20338986.
8	2016	58074	0.324	11116.	12779843.
9	2017	52741	0.367	12347.	12393140.
10	2018	3272	0.0263	5545.	724424.

In this section, the deadline_year was selected in the dataset as it indicates when each project would end. It is more intuititive as some projects may have a long active duration. Moreover, we may categorize the states as binomial; the state is either success or not, and that how we can determine the success rate for each year. In the dataset, there are two more columns that help identify the mean and the max of fund raise in that particular year. This is useful as it shows how projects in each year got attentions for funding.

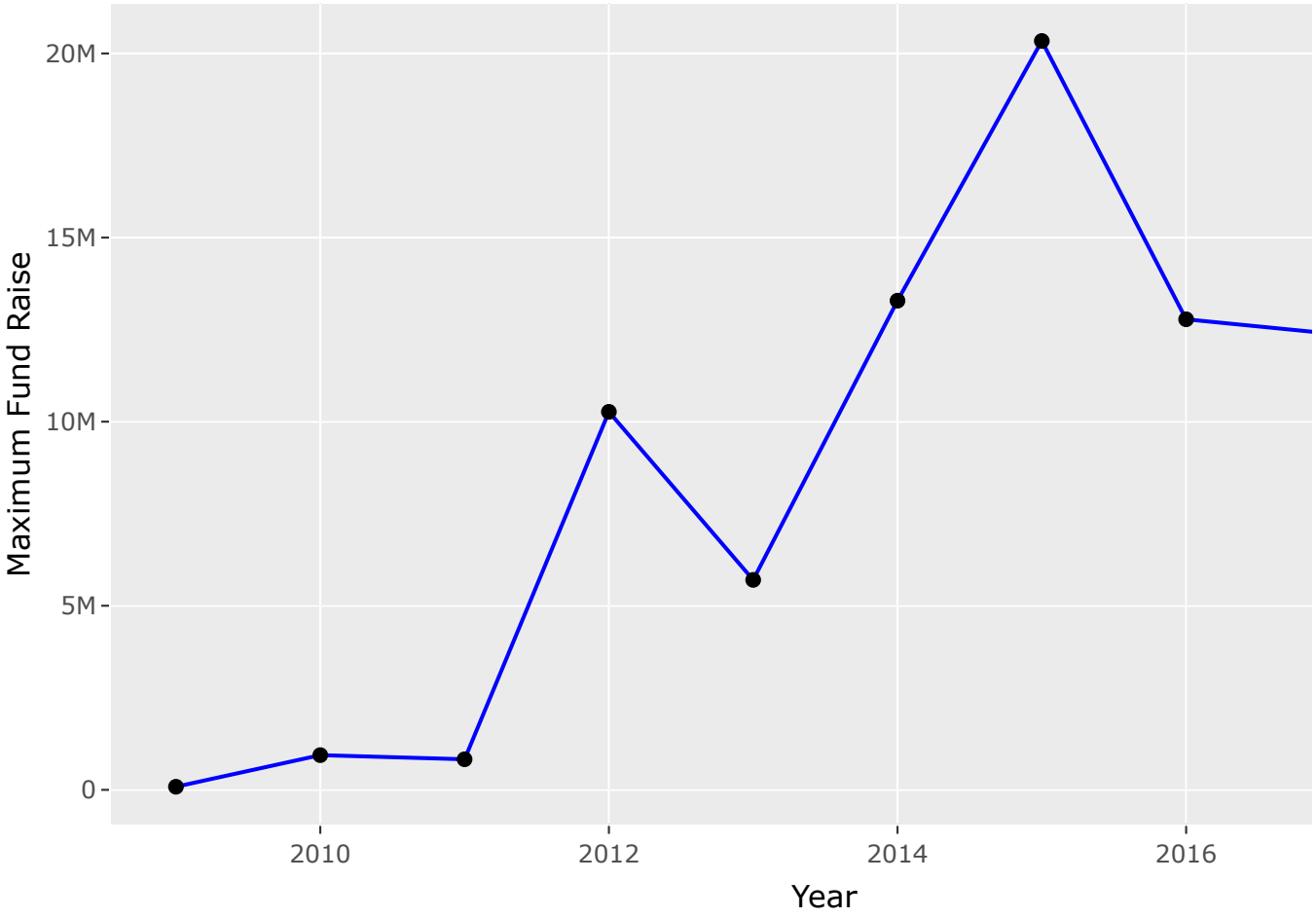
```
# Plot a bar chart with deadline_year on x-axis and project counts on y-axis
project_count <- yearly_summary |>
  ggplot() +
  geom_col(aes(x = deadline_year,
               y = project_count,
               fill = project_count,
               text = project_count)) +
  labs('Projects Count Yearly Summary', x = 'Year', y = 'Project counts') +
  scale_y_continuous(labels = label_number(scale_cut = cut_short_scale()))
ggplotly(project_count, tooltip = 'text')
```





With the visualization, it can be observed that the distribution of project counts is close to normal distribution, but a little skewed-left. This means that, between 2009-2018, there were more projects toward the end of duration rather than the beginning. It can be seen that the project counts peaked at about 2015 before it dropped down dramastically in 2018.

```
# Plot a bar chart with deadline_year on x-axis and project counts on y-axis
max_fund_raise <- yearly_summary |>
  ggplot(aes(x = as.numeric(deadline_year),
             y = max_fund_raise)) +
  geom_line(color = 'blue') +
  geom_point() +
  labs('Kickstart Maximum Fund Raise', x = 'Year', y = 'Maximum Fund Raise') +
  scale_y_continuous(labels = label_number(scale_cut = cut_short_scale()))
ggplotly(max_fund_raise)
```



The line graph compliments the previous bar chart as, not only the total project counts, it also shows that the maximum fund raise peaked in 2015 with a project that could raised over 20 million dollars, and then dropped down significantly in 2018.

Unusual data values

Another interesting about observation about our data is that some projects can have very small values of 'usd_goal_real'.

```
unusual <- ks_data |>
  filter(usd_goal_real < 1)
head(unusual)
```

```
          ID                                                name       category
1 1002571103                           3mm Aluminum Pick Plectrum          Music
2 1050732941                                 LENDEMAIN DE VEILLE Country & Folk
3 1081928099 $9 Clearest Phone Hologram (Pepper's Ghost Pyramid)     Technology
4  117369169                                          Platformer37    Video Games
5 1202361966                                 Eyes Without A Face         Horror
6 1379346088                 Bug's Big Adventure: Mexico Edition            Art
  main_category currency   deadline goal            launched pledged      state
1         Music      CAD 2014-03-24    1 2014-02-22 23:28:33   20.00 successful
2         Music      CAD 2017-03-29    1 2017-01-30 05:12:44 5307.87 successful
3    Technology      CAD 2015-09-06    1 2015-08-07 02:45:35 7669.00 successful
4         Games      CAD 2015-04-24    1 2015-03-23 19:05:05   31.00   canceled
5  Film & Video      CAD 2016-09-08    1 2016-07-10 00:38:48   24.00 successful
6           Art      MXN 2016-11-12   10 2016-11-11 16:30:00  335.00 successful
  backers country usd.pledged usd_pledged_real usd_goal_real deadline_date
1       9      CA       17.99            18.12          0.91    2014-03-24
2     118      CA        0.00          3978.02          0.75    2017-03-29
3     331      CA     5814.05          5782.25          0.75    2015-09-06
4       4      CA       24.68            25.58          0.83    2015-04-24
5       6      CA       15.39            18.52          0.77    2016-09-08
6       7      MX       18.05            16.41          0.49    2016-11-12
  launched_date deadline_year launched_year
1    2014-02-22          2014          2014
2    2017-01-30          2017          2017
3    2015-08-07          2015          2015
4    2015-03-23          2015          2015
5    2016-07-10          2016          2016
6    2016-11-11          2016          2016
```

It can be seen that there were many projects that were asking for less than $1. As the amount

of fund is way too low to be realistic, it can be assumed that some projects might have incorrect data. Another possible cause could be that some projects were raising funds for more symbolic purposes. A question that can be passed on to the creator in regards to this unusual if it could be a case that some projects were created to raise more awareness rather than raising funds, and if it is possible to recognize and add an indicator into the dataset.