HarvardX: PH125.9x Data Science - Capstone Project
Kickstarter - Crowdfunding Success Prediction

Léo Dange - https://github.com/ldange

12/22/2020

Introduction

Related to the ***Harvardx Data Science Program - Capstone PH125.9x***, this second project of building machine learning algorythm concern the kickstarter platform and it's crowdfunding project. I have been using kickstarter as a "pledger/founder" for many years and I recently helped a friends to launch a project that indeed has not been sucessfully funded.

Kickstarter as a platform is for me an incredible opportunity to help brilliant people to developp a product and launch their company. I like to concept and all it's aspects, but It is sometimes complicate to analyse and understand what determine the success or failure of a project. For some of them it can be obvious but some are more tricky. I often have issue understanding why some projects did got the finance when I though (in my own opinion) they would not have success, and why sometime projects failed when I though they should have succeeded.

This project will help us to determine if we can predict the sucess or failure of a project regarding few variables.

This Report will present you the ***Project Goal, the Data used, The Methods and Analysis of the results*** and at the end some ***Conclusion***.

The project will be delivered of the following files :

- The Report in .pdf
- The Report in .Rmd
- The Code and Script in .R

Project's Goal

As explained in the Introduction, the Project's Goal is to build a machine learning algorythm able to predict the sucess of a project. We will build several models and make them compete to determine which one is the most accurate.

To evaluate our work we are going to use the Root Mean Square Error or RMSE :

$$RMSE = \sqrt{\frac{1}{N}\sum_{u,i}(\hat{y}_{u,i} - y_{u,i})^2}$$

The RMSE is frequently used to measure the difference between values predicted by a model and the observed values, in our case between the ***edx_kickstater set (Train subset) and the Validation_kickstarter subset***.

The best model will be defined as the one with the lowest RMSE, this time we do not have a targeted goal so we will pick the best one of our several models.

Data

Our two sets are based on the Kickstarter project Database available on Kaggle (link in the code bellow) originally contains **378'661 projects, 15 variables/observation**. Before using We will :

1) start by removing renaming and reordering columns.

2) Filtering the status and only keep the "successful" "failed" & "canceled" projects.

After those modifications our dataframe will contain **370'454 observations & 11 variables**.

Then we split this data set in two partitions and i chose to keep the 90/10 as we did in our MovieLens Analysis. the train_kickstarter set (train set) will hold 90% of the data, and the validation_kickstarter 10% left.

For simplicity the Dataset has been download from Kaggle then upload on my github repository dedicated to this project.

```
############################################################
# Create kickstarter train set, validation set (final hold-out test set,
similar to the edx set for the Movielens Project)
############################################################

# Note: this process could take a couple of minutes

if(!require(tidyverse)) install.packages("tidyverse", repos =
"http://cran.us.r-project.org")
if(!require(caret)) install.packages("caret", repos = "http://cran.us.r-
project.org")
```

```r
if(!require(data.table)) install.packages("data.table", repos =
"http://cran.us.r-project.org")
if(!require(stringr)) install.packages("stringr", repos = "http://cran.us.r-
project.org")
if(!require(gridExtra)) install.packages("gridExtra", repos =
"http://cran.us.r-project.org")

library(tidyverse)
library(caret)
library(data.table)
library(stringr)

# Kickstarter's Projects Database:
# https://www.kaggle.com/kemical/kickstarter-projects?select=ks-projects-
201612.csv
# https://www.kaggle.com/kemical/kickstarter-projects/download

# Data can be viewed/downloaded from my github repository
# https://github.com/ldange/HarvardX-Data-Science-Capstone/blob/master/ks-
projects-201801.csv

# Creating the dataframe from csv file
kickstarter <- read.csv("https://raw.githubusercontent.com/ldange/HarvardX-
Data-Science-Capstone/master/ks-projects-201801.csv")

# Removing unecessary column
drops <- c("deadline","usd.pledged","usd_pledged_real","usd_goal_real")
kickstarter <- kickstarter[ , !(names(kickstarter) %in% drops)]

# Renaming and Reordering column
kickstarter <- kickstarter %>%
  rename(
    sub_category = category,
    status = state,
    projectId = ID
  )
kickstarter <- kickstarter[ , c(1,2,4,3,11,5,7,10,8,6,9)]

# Creating subset dataframe without the status "live, suspended & undefined"
kickstarter <- subset(kickstarter, kickstarter$status %in%
c("canceled","failed","successful") )

# Validation set will be 10% of Kickstarter Project data
set.seed(1, sample.kind="Rounding") # if using R 3.5 or earlier, use
`set.seed(1)`
test_index_kickstarter <- createDataPartition(y = kickstarter$status, times =
1, p = 0.1, list = FALSE)
train_kickstarter <- kickstarter[-test_index_kickstarter,]
validation_kickstarter <- kickstarter[test_index_kickstarter,]
```

As stated before, the two sets are now composed of 11 variables instead of the original 15 and we can see that the edx set is, as chosen, containing 90% of the data.

```
str(train_kickstarter)
```

```
## 'data.frame':    333408 obs. of  11 variables:
##  $ projectId    : int  1000002330 1000004038 1000007540 1000011046
1000014025 1000023410 1000030581 1000034518 100004195 100004721 ...
##  $ name         : chr  "The Songs of Adelaide & Abullah" "Where is Hank?"
"ToshiCapital Rekordz Needs Help to Complete Album" "Community Film Project:
The Art of Neighborhood Filmmaking" ...
##  $ main_category: chr  "Publishing" "Film & Video" "Music" "Film & Video"
...
##  $ sub_category : chr  "Poetry" "Narrative Film" "Music" "Film & Video"
...
##  $ country      : chr  "GB" "US" "US" "US" ...
##  $ currency     : chr  "GBP" "USD" "USD" "USD" ...
##  $ launched     : chr  "2015-08-11 12:12:28" "2013-01-12 00:20:50" "2012-
03-17 03:24:11" "2015-07-04 08:35:03" ...
##  $ backers      : int  0 3 1 14 224 16 40 58 43 0 ...
##  $ pledged      : num  0 220 1 1283 52375 ...
##  $ goal         : num  1000 45000 5000 19500 50000 1000 25000 125000 65000
2500 ...
##  $ status       : chr  "failed" "failed" "failed" "canceled" ...
```

The 11 remaining features/variables/columns in both datasets are as follow :

- **projectId**, ***integer*** containing the identification number of the project.
- **name**, ***character*** containing the name.
- **main-category**, ***character*** containing the main-category (parent of the category).
- **sub-category**, ***character*** containing the "Sub-category".
- **country**, ***character*** containing the country of origin of the project/company behind the project
- **currency**, ***character*** containing the currency related to the country
- **launched**, ***character*** containing the launched date of the project
- **backers**, ***integer*** containing the number of backers of the project
- **pledged**, ***numeric*** containing the amount of money pledged to the project
- **goal**, ***numeric*** containing the target amount of money aimed for the project

Exploratory Data Analysis

```
head(train_kickstarter, n=3)
```

```
##    projectId                                          name
main_category
## 1 1000002330                 The Songs of Adelaide & Abullah
Publishing
## 3 1000004038                                  Where is Hank?  Film &
Video
```

```
## 4 1000007540 ToshiCapital Rekordz Needs Help to Complete Album
Music
##     sub_category country currency            launched backers pledged
goal
## 1         Poetry      GB      GBP 2015-08-11 12:12:28       0       0
1000
## 3 Narrative Film      US      USD 2013-01-12 00:20:50       3     220
45000
## 4          Music      US      USD 2012-03-17 03:24:11       1       1
5000
##   status
## 1 failed
## 3 failed
## 4 failed
```

To be able to conduct an Exploratory Data Analysis those data will need at least one transformations, in addition of cleaning/reorganising we proceed before. We will use an ***eda*** dataframe to do our exploration, without any risk of alteration for the train set. This set will be ***train_eda_kickstarter***.

- Add a **funding_percentage** variable that tell us the percentage of funding of the project

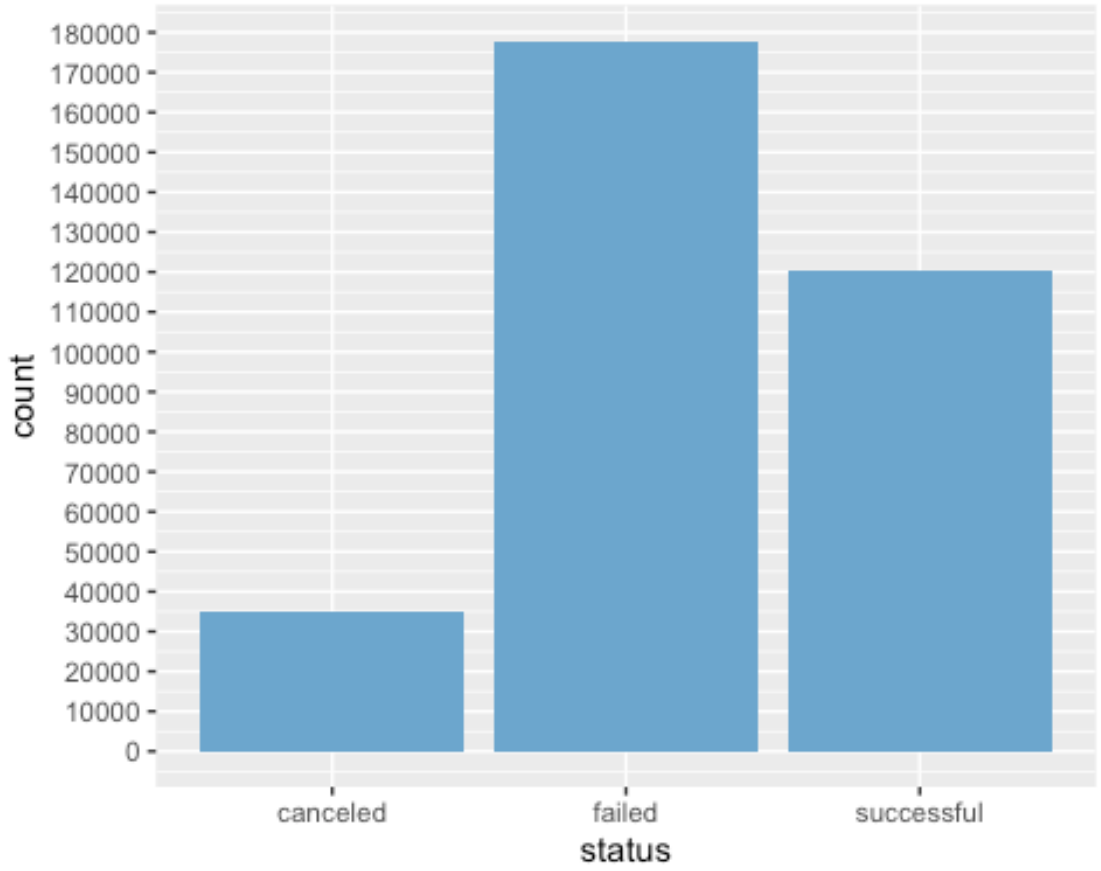
```
# Create a completion percentage for each project and limiting the number to
only two digits (for the whole project).

train_eda_kickstarter <- train_kickstarter %>% mutate(funding_percentage =
as.numeric(pledged / goal)*100)

options(digits=2)
```

Funding Analysis

We will start our analysis by a quick Funding Analysis. This first plot will help us to understand which part of project are successful.

First major information, we can see that most of projects failed on Kickstarter.

```
# Ploting the Status Distribution
train_eda_kickstarter %>%
    ggplot(aes(status)) +
    geom_bar(fill = "skyblue3") +
    scale_y_continuous(breaks = c(seq(0, 180000, 10000)))
```



Failed project represent 53.37% of all projects, when **sucessful project are only 36.15%**. Finally, **Canceled project represent the remaining 10.48%**.

```
# Looking at the Failing & Canceled ratio of projects
kickstarter_failed <- train_eda_kickstarter %>% filter(status == "failed")
%>% count() / train_eda_kickstarter %>% count
kickstarter_failed
```

```
##      n
## 1 0.53
```

```
kickstarter_successful <- train_eda_kickstarter %>% filter(status ==
"successful") %>% count() / train_eda_kickstarter %>% count
kickstarter_successful

##      n
## 1 0.36
```

Still, one third of the project getting marked as successful seems a lot, comparing to the average success of a company/product "in real life". when we look at the dataframe, lots of successful projects had goal of 1 / 10 or 100 dollars, which is quickly reached but doesn't guarantee any product release.

2'386 projects have a goal inferior to 200 dollars and less than 10 backers in the project. Some even have 0 backers and seems successful because the creator pledge itself.

```
#Counting Sucessfull projects with less than 200 of goals and 10 backers
train_eda_kickstarter %>% filter(status == "successful" & goal <= 200 &
backers <= 10) %>% count()

##      n
## 1 2386
```

Canceled project are often related to insufisant backers/fund, crater rather canceled the project before then end of the financing schedule instead of waiting knowing it won't be successful anyway. In our case we do not have the reason of why the project was canceled but we can analyze what percentage of founding received the canceled project.

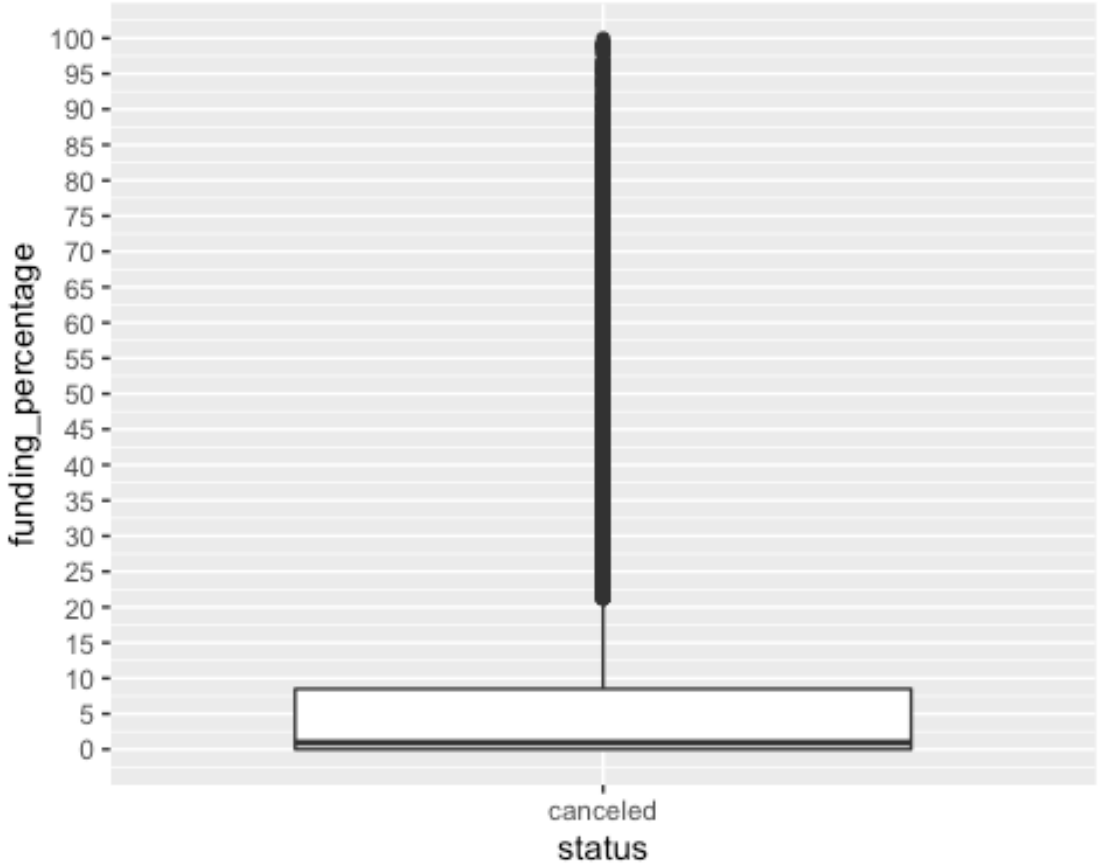
Only 649 of the 34'901 canceled projects reach at least 100% of the funding goal, which mean **34'252 projects where not funded before they were canceled**.

On the 649 funded projects but canceled, **103 projects had goals bellow 200$**, which reduce more the number of projects that reach a real funding goal.

```
# Project funded but canceled
train_eda_kickstarter %>% filter(status == "canceled" & funding_percentage >=
100) %>% count()

##     n
## 1 646

# Project funded with a funding goal bellow 200
train_eda_kickstarter %>% filter(status == "canceled"& goal >= 200 &
funding_percentage >= 100) %>% count()

##     n
## 1 501
```

If we look at the projects that were canceled without been founded, we see that ***most of the project got between 0 (first quartile) and 8% (third quartile) of their funding goal, with a median bellow 2%.*** We exclude canceled project with over 100% of funding on this boxplot

```r
# Boxplot of the Canceled Status regarding the funding_percentage
train_eda_kickstarter %>%
  filter(status == "canceled" & funding_percentage < 100) %>%
  ggplot(aes(status,funding_percentage)) +
  geom_boxplot() +
  scale_y_continuous(breaks = c(seq(0, 100, 5)))
```

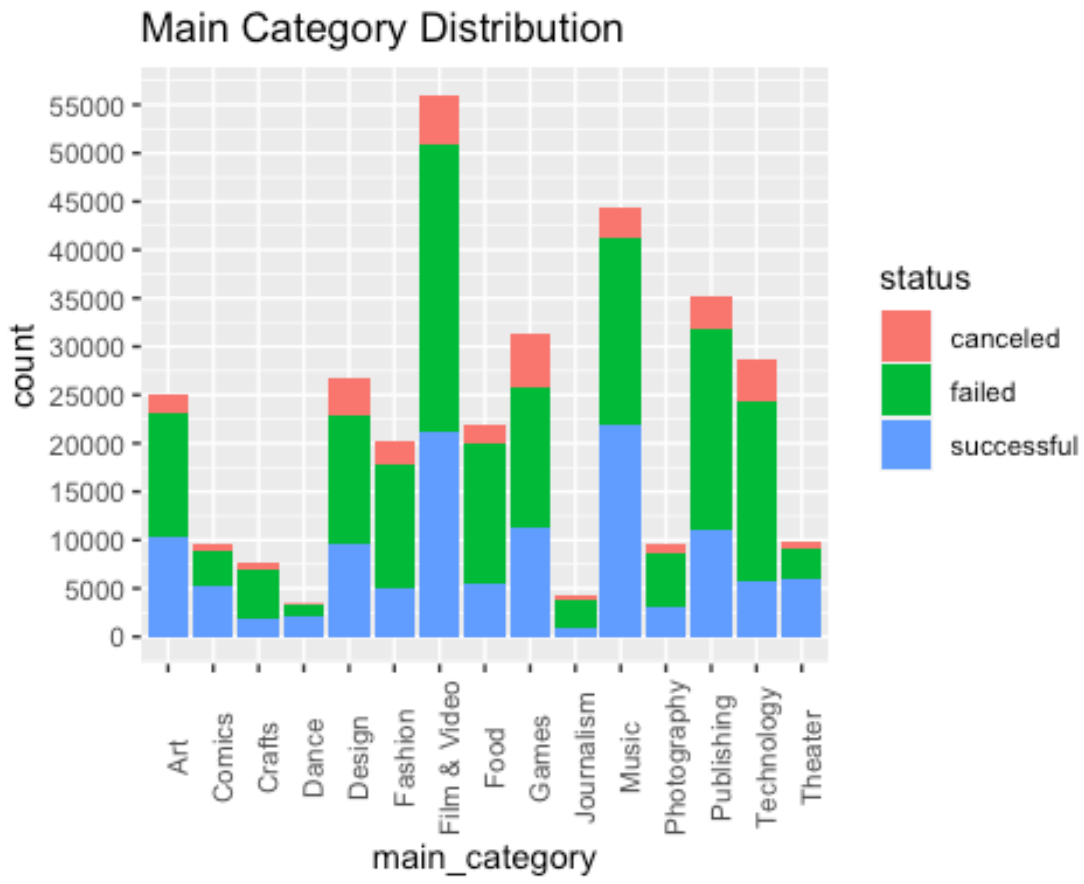


Category Analysis

We will now analyze the category.

By looking at the main category distribution bellow, we can see that most used main category is in order **Film & Video, Music, Publishing, Games & Technology**.

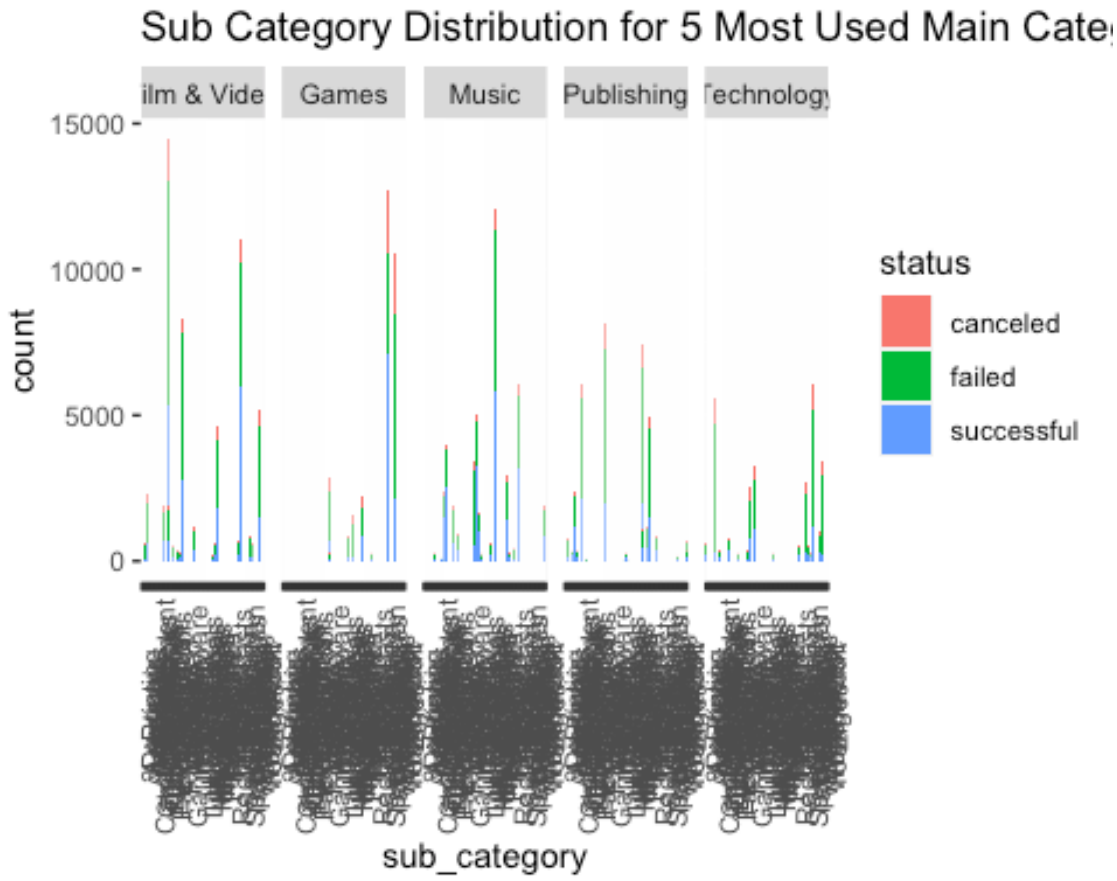
```
# Plot distribution of the Main Category
train_eda_kickstarter %>%
  ggplot(aes(main_category, fill = status)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_y_continuous(breaks = c(seq(0, 60000, 5000))) +
  ggtitle("Main Category Distribution")
```



For all category excepted **Music, Comics, Dance & Theater**, there is more canceled & Failed projects than successful one. Three of those 4 main category ***Comics, Dance & Theater*** are in the & least used main category of the entire platform.

```
# Plotting the Sub_category distribution for the fifth biggest main_category
train_eda_kickstarter %>%
  filter(main_category %in% c("Film & Video", "Music", "Publishing", "Games",
"Technology")) %>%
     ggplot(aes(sub_category, fill = status)) +
```

```
    geom_bar() +
    theme(axis.text.x = element_text(angle = 90)) +
    scale_y_continuous(breaks = c(seq(0, 60000, 5000))) +
    ggtitle("Sub Category Distribution for 5 Most Used Main Category") +
  facet_grid(.~main_category)
```

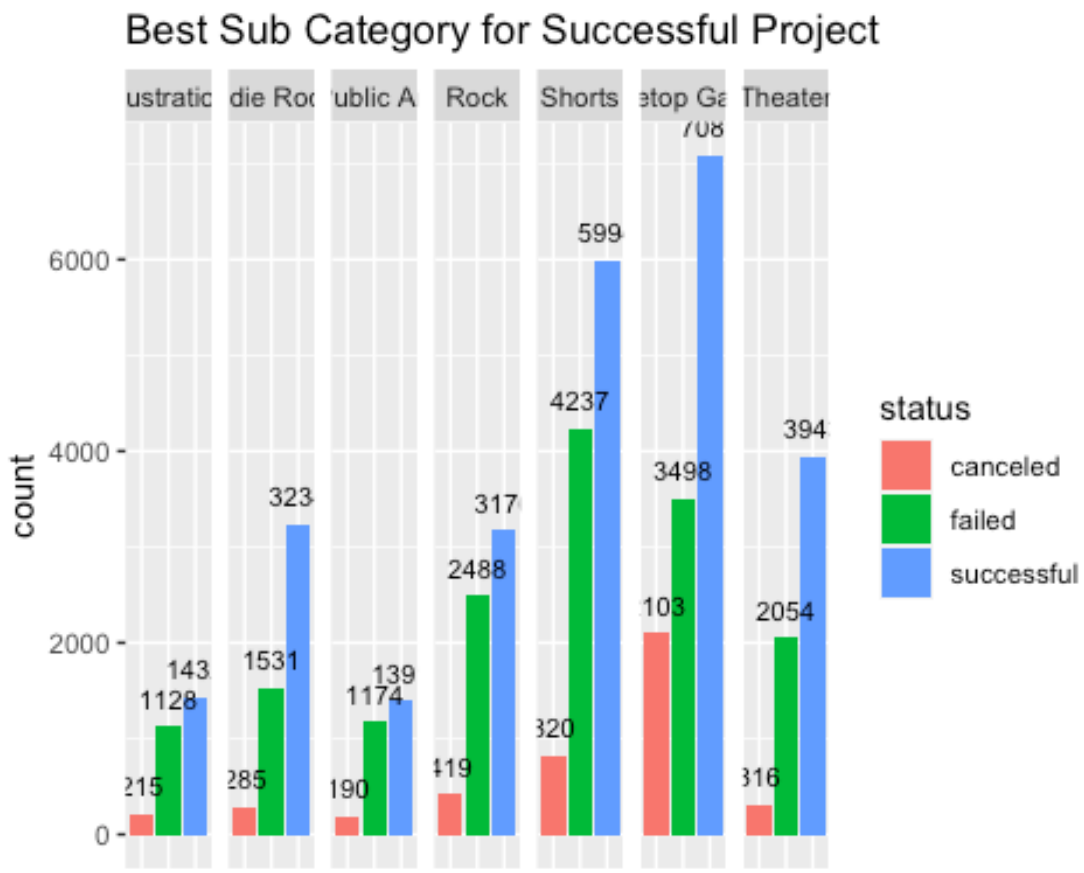


For each of the 5 most used main category, is always two or three sub_category that are way more used than de rest of the entire category.

For our next plot we've looked on all the sub_category and we saw there is 7 sub_category where there is more successful projects than failed and canceled one. We've ploted the status for each of them. This information could help kickstarters to understand where to post their projects to maximize the chance of funding.

```
# Grid Faceting plot of the best sub_category regarding their status

train_eda_kickstarter %>%
    filter(sub_category %in% c("Shorts", "Country", "Indie Rock", "Rock",
"Tabletop Games", "Illustration", "Public Art", "Theater")) %>%
    ggplot(aes(status, fill = status)) +
    geom_bar(position = "dodge") +
    theme(axis.title.x=element_blank(),
                axis.text.x=element_blank(),
                axis.ticks.x=element_blank()) +
```

```
    ggtitle("Best Sub Category for Successful Project") +
    geom_text(stat='count', aes(label=..count..), vjust=-1, size = 3) +
    facet_grid(.~sub_category)
```



We will plot the best and the worst "canceled project" regarding their genre.

```
#Creating dataframes for the highest & lowest funding_percentage but canceled
project

best_canceled <- train_eda_kickstarter %>%
  filter(status == "canceled" & goal >= 200 & funding_percentage >= 100)

best_canceled <- data.frame(main_category =
best_canceled$main_category,funding_percentage =
best_canceled$funding_percentage)

worst_canceled <- train_eda_kickstarter %>%
  filter(status == "canceled" & funding_percentage <= 100)

worst_canceled <- data.frame(main_category =
worst_canceled$main_category,funding_percentage =
worst_canceled$funding_percentage)
```

```r
# Plotting the distribution regarding the category

best_plot <- best_canceled %>% ggplot(aes(main_category)) +
  geom_bar(fill = "skyblue3") +
  theme(axis.text.x = element_text(angle = 90)) +
  ggtitle("genre distribution for best canceled project")

worst_plot <- worst_canceled %>% ggplot(aes(main_category)) +
  geom_bar(fill = "skyblue3") +
  theme(axis.text.x = element_text(angle = 90)) +
  ggtitle("genre distribution for worst canceled project")

grid.arrange(best_plot,worst_plot, ncol = 1)
```

We can see on those two plots that, Design, Games & Technology are vastly present in the canceled projects.Except for the Design category, almost all project's category mostly failed and were canceled, rather than being funded but canceled, which is not surprising.

We will construct a kick dataframe regarding the amount of project for each main_category and the ratio of successful project to see if there is any correlation.

```r
# create dataframes for the total projects and sucessful project for each
main_category
```

```
total_maincategory <- train_eda_kickstarter %>%
  group_by(main_category) %>%
  count()

total_maincategory <- data.frame(main_category =
total_maincategory$main_category,
                                 total = as.numeric(total_maincategory$n))

successful_maincategory <- train_eda_kickstarter %>%
  group_by(main_category) %>%
  filter(status == "successful") %>%
  count()

successful_maincategory <- data.frame(main_category =
successful_maincategory$main_category,
                                      successful =
as.numeric(successful_maincategory$n))

projects_maincategory <-
left_join(total_maincategory,successful_maincategory)

## Joining, by = "main_category"

projects_maincategory <- projects_maincategory[order(-
projects_maincategory$total),]

success_ratio <- data.frame(main_category =
projects_maincategory$main_category,
                            success_ratio =
as.numeric(projects_maincategory$successful / projects_maincategory$total))

projects_maincategory <- left_join(projects_maincategory,success_ratio)

## Joining, by = "main_category"
```

A quick look at the values

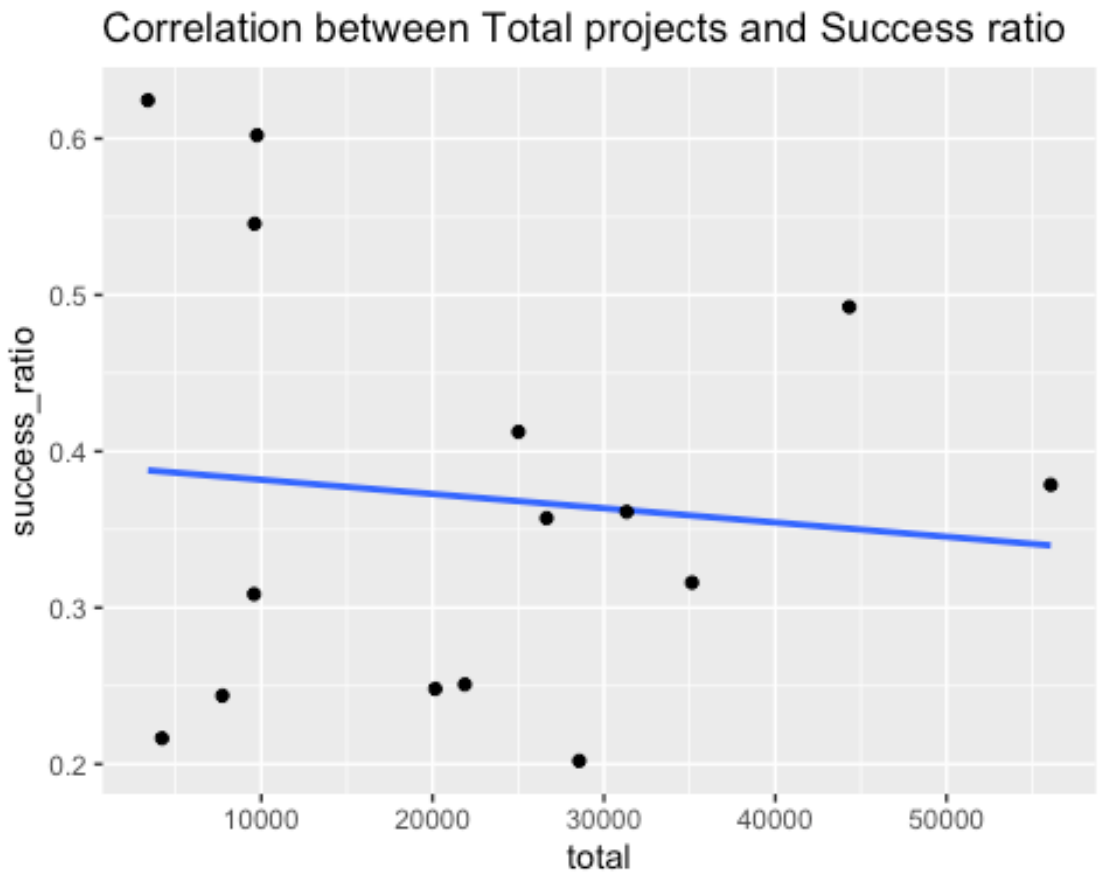
```
# Looking at the data
head(projects_maincategory, n=5)

##   main_category total successful success_ratio
## 1  Film & Video 56080      21232          0.38
## 2         Music 44321      21819          0.49
## 3    Publishing 35134      11104          0.32
## 4         Games 31329      11319          0.36
## 5    Technology 28559       5770          0.20

# Correlation test between total project and success ratio
projects_maincategory %>% summarize(r = cor(total,success_ratio))
```

```
##      r
## 1 -0.1
```

We could have think that the bigger the category is, to bigger the chance of funding is (due to traffic on this category), but we can see that there is strictly no correlation between the amount of projects and the ratio of sucessfully funded project.

```
# Plotting the total project and success ratio with a regression line
projects_maincategory %>%
  ggplot(aes(total, success_ratio)) + geom_smooth(method=lm, se=FALSE) +
  geom_point() +
  ggtitle("Correlation between Total projects and Success ratio")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

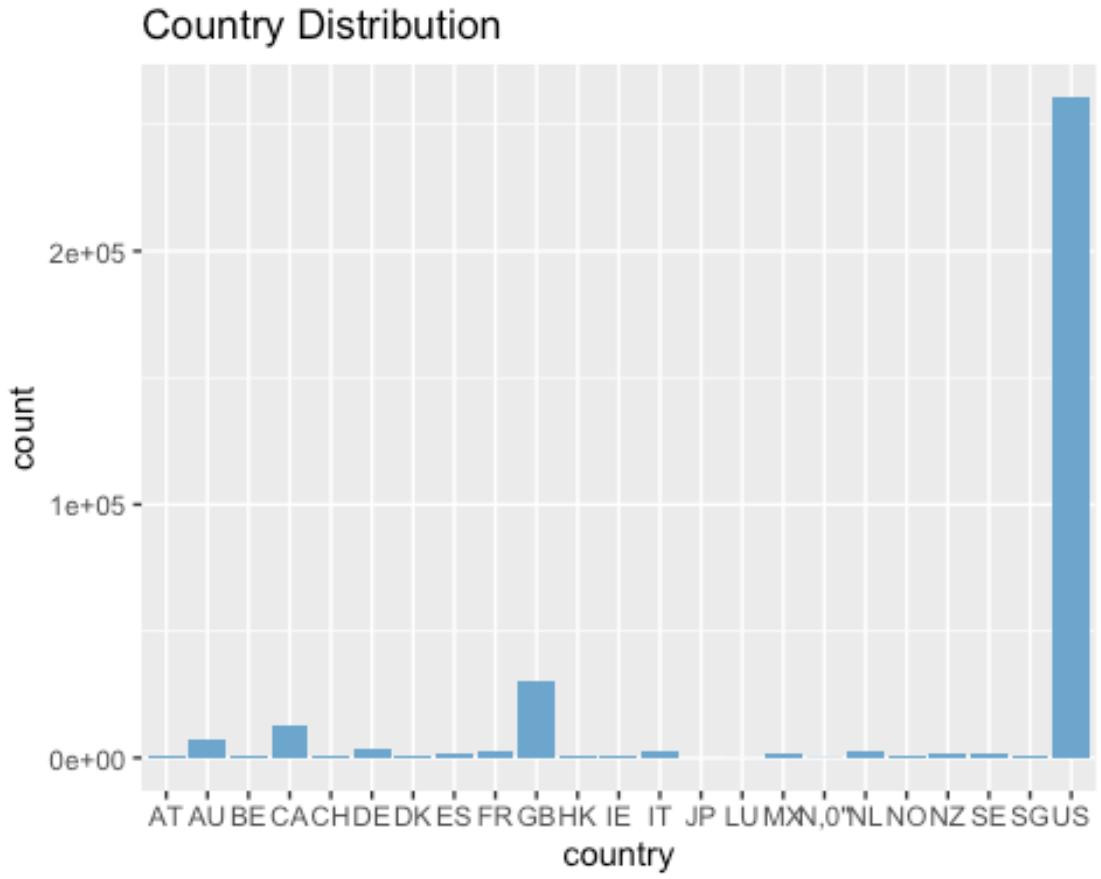


Country Analysis

Lastly we will conduct an analysis to analyse trend regarding the different countries.

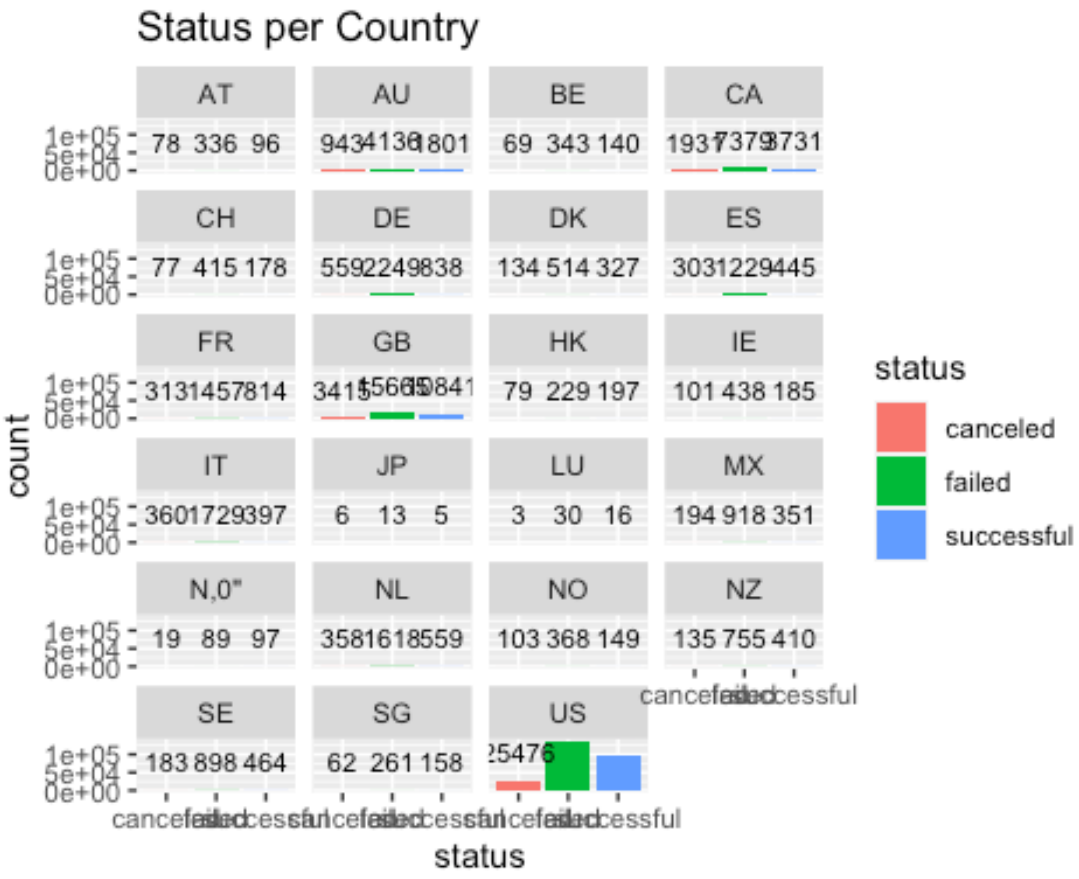
We start our analysis with a country distribution that show US as the biggest provider of projects, followed by Great Britain, Canada and Australia. All English speaker country.

```
# Ploting the country Distribution
train_eda_kickstarter %>%
    ggplot(aes(country)) +
    geom_bar(fill = "skyblue3") +
    ggtitle("Country Distribution")
```



If we look at the status for each country, we see that no country as more sucessful projects than failed/conceled projects.

```
train_eda_kickstarter %>%
  group_by(country) %>%
  ggplot(aes(status, fill = status)) +
  geom_bar(position = "dodge") +
  geom_text(stat='count', aes(label=..count..), vjust=-1, size = 3) +
  facet_wrap(country~., ncol = 4) +
  ggtitle("Status per Country")
```



This is comfirmed by the sucess rate for each country.

```
# Creating dataframe regarding sucess ratio per country
country_total_success <- train_eda_kickstarter %>%
     group_by(country) %>%
     count()
country_total_success <- data.frame(country = country_total_success$country,
                                    total =
as.numeric(country_total_success$n))

country_success <- train_eda_kickstarter %>%
     group_by(country) %>%
     filter(status == "successful") %>%
     count()
country_success <- data.frame(country = country_success$country,
                              successful = as.numeric(country_success$n))

country_ratio <- data.frame(country = country_total_success$country,
                            success_ratio =
as.numeric(country_success$successful / country_total_success$total))

country_total_success <- left_join(country_total_success,country_success)
```

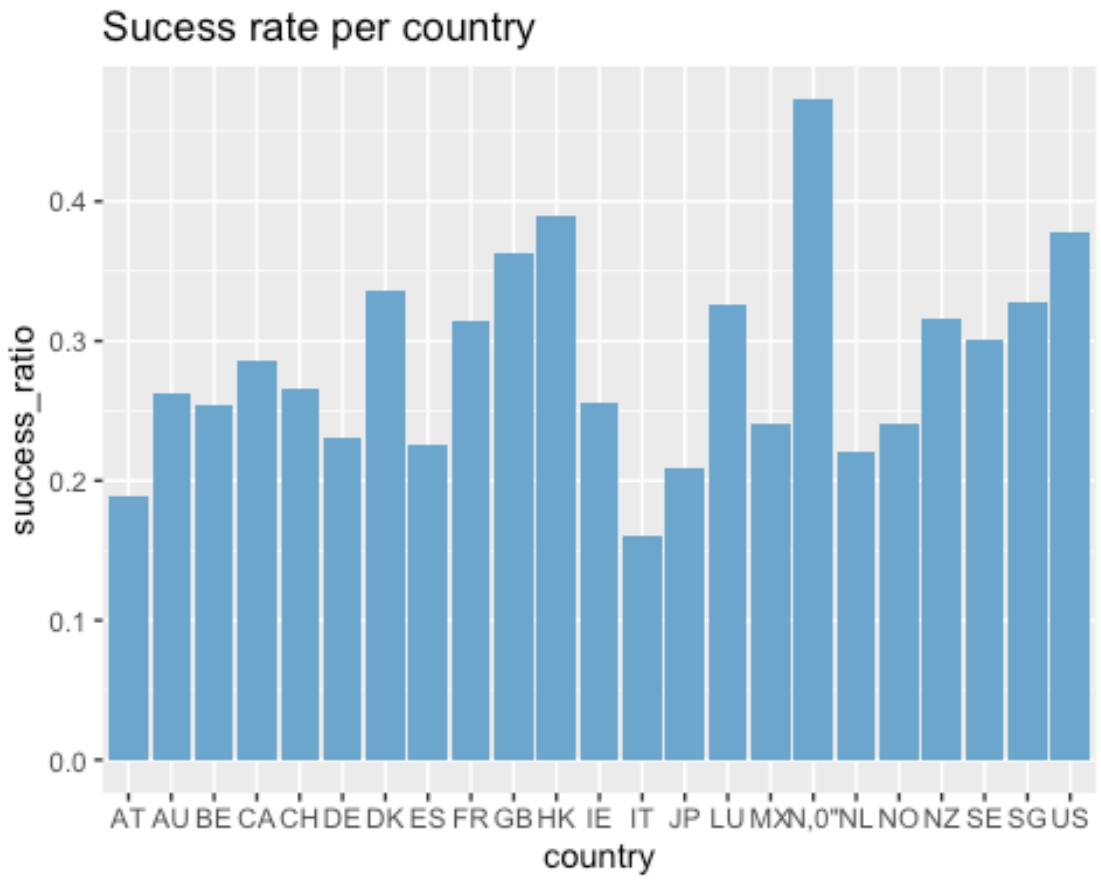
```
## Joining, by = "country"
```

```
country_total_success <- left_join(country_total_success,country_ratio)
```

```
## Joining, by = "country"
```

Here is the plot of the sucess rate. Wee quickly see that the amount of projects (previous graph) have zero impact on the sucess ratio for each country.

```
country_total_success %>%
    ggplot(aes(country, success_ratio)) +
    geom_col(fill = "skyblue3") +
    ggtitle("Sucess rate per country")
```



Now we are interested of the amount of backers regarding the origins/country of the projects.

EDA Analysis Summary

Methods & Analysis

Results

Conclusion