Capstone

JP Pugliese

11/20/2019

This is the final assignment for the Harvard Data Science Professional course. The chosen project is using the Kickstarter Kaggle dataset and apply machine learning to estimate the success of a campaign.

Table of Contents

Executive Summary

Using the most recent dataset in Kaggle for the different projects having run on Kickstarter, we look at the data structure, visualizing different trends. Some data cleansing was required to apply some machine learning models in order to forecast any likelihood of a campaign success in Kickstarter.

Introduction

This report uses a dataset of Kickstarter dating from 2018 available on Kaggle. The dataset is loaded into R and the dataset structure is analysed. Some trends can be observed and some conclusions can be derived from the visualization part. Then, some machine learning models are applied to estimate the accuracy of predicting the success or failure of a Kickstarter campaign based on some predicators. The used dataset has been stored in Github.

The dataset

The dataset in a CSV format is first of all loaded and then the dataset structure is looked at. The datset has a size of 55MB with 15 columns and 378661 rows. We can see here below a short view of what is contained in the dataset.

```
#Loading the data from a local HD
setwd("~/projects/Capstone/Capstone")
data <- read_csv("ks-projects-201801.csv")
```

```
## Parsed with column specification:
## cols(
##   ID = col_double(),
##   name = col_character(),
##   category = col_character(),
##   main_category = col_character(),
##   currency = col_character(),
##   deadline = col_date(format = ""),
##   goal = col_double(),
##   launched = col_datetime(format = ""),
##   pledged = col_double(),
##   state = col_character(),
##   backers = col_double(),
##   country = col_character(),
##   `usd pledged` = col_double(),
##   usd_pledged_real = col_double(),
##   usd_goal_real = col_double()
## )
```

```
summary(data)
```

```
##        ID                name             category
##  Min.   :5.971e+03   Length:378661      Length:378661
##  1st Qu.:5.383e+08   Class :character   Class :character
##  Median :1.075e+09   Mode  :character   Mode  :character
##  Mean   :1.075e+09
##  3rd Qu.:1.610e+09
##  Max.   :2.147e+09
##
##  main_category        currency            deadline
##  Length:378661      Length:378661      Min.   :2009-05-03
##  Class :character   Class :character   1st Qu.:2013-06-08
##  Mode  :character   Mode  :character   Median :2015-01-14
##                                        Mean   :2014-11-01
##                                        3rd Qu.:2016-04-28
##                                        Max.   :2018-03-03
##
##       goal             launched                         pledged
##  Min.   :        0   Min.   :1970-01-01 01:00:00   Min.   :       0
##  1st Qu.:     2000   1st Qu.:2013-05-07 22:14:27   1st Qu.:      30
##  Median :     5200   Median :2014-12-10 03:23:41   Median :     620
##  Mean   :    49081   Mean   :2014-09-28 18:06:17   Mean   :    9683
##  3rd Qu.:    16000   3rd Qu.:2016-03-24 10:21:09   3rd Qu.:    4076
##  Max.   :100000000   Max.   :2018-01-02 15:02:31   Max.   :20338986
##
##     state              backers            country
##  Length:378661      Min.   :     0.0   Length:378661
##  Class :character   1st Qu.:     2.0   Class :character
##  Mode  :character   Median :    12.0   Mode  :character
##                     Mean   :   105.6
##                     3rd Qu.:    56.0
##                     Max.   :219382.0
##
##   usd pledged       usd_pledged_real   usd_goal_real
##  Min.   :       0   Min.   :       0   Min.   :        0
##  1st Qu.:      17   1st Qu.:      31   1st Qu.:     2000
##  Median :     395   Median :     624   Median :     5500
##  Mean   :    7037   Mean   :    9059   Mean   :    45454
##  3rd Qu.:    3034   3rd Qu.:    4050   3rd Qu.:    15500
##  Max.   :20338986   Max.   :20338986   Max.   :166361391
##  NA's   :3797

#dimensions of the dataset
dim(data)

## [1] 378661     15

#Information and column's headers
head(data, 5)

## # A tibble: 5 x 15
##        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
## # … with 8 more variables: launched <dttm>, pledged <dbl>, state
<chr>,
## #   backers <dbl>, country <chr>, `usd pledged` <dbl>,
## #   usd_pledged_real <dbl>, usd_goal_real <dbl>

tail(data, 5)

## # A tibble: 5 x 15
##         ID name  category main_category currency deadline    goal
##      <dbl> <chr> <chr>    <chr>         <chr>    <date>     <dbl>
## 1 10.00e8 Chkn… Documen… Film & Video  USD      2014-10-17 50000
## 2 10.00e8 The … Narrati… Film & Video  USD      2011-07-19  1500
## 3 10.00e8 Wall… Narrati… Film & Video  USD      2010-08-16 15000
## 4 10.00e8 BioD… Technol… Technology    USD      2016-02-13 15000
## 5 10.00e8 Nou … Perform… Art           USD      2011-08-16  2000
## # … with 8 more variables: launched <dttm>, pledged <dbl>, state
<chr>,
## #   backers <dbl>, country <chr>, `usd pledged` <dbl>,
## #   usd_pledged_real <dbl>, usd_goal_real <dbl>
```

Data Cleansing

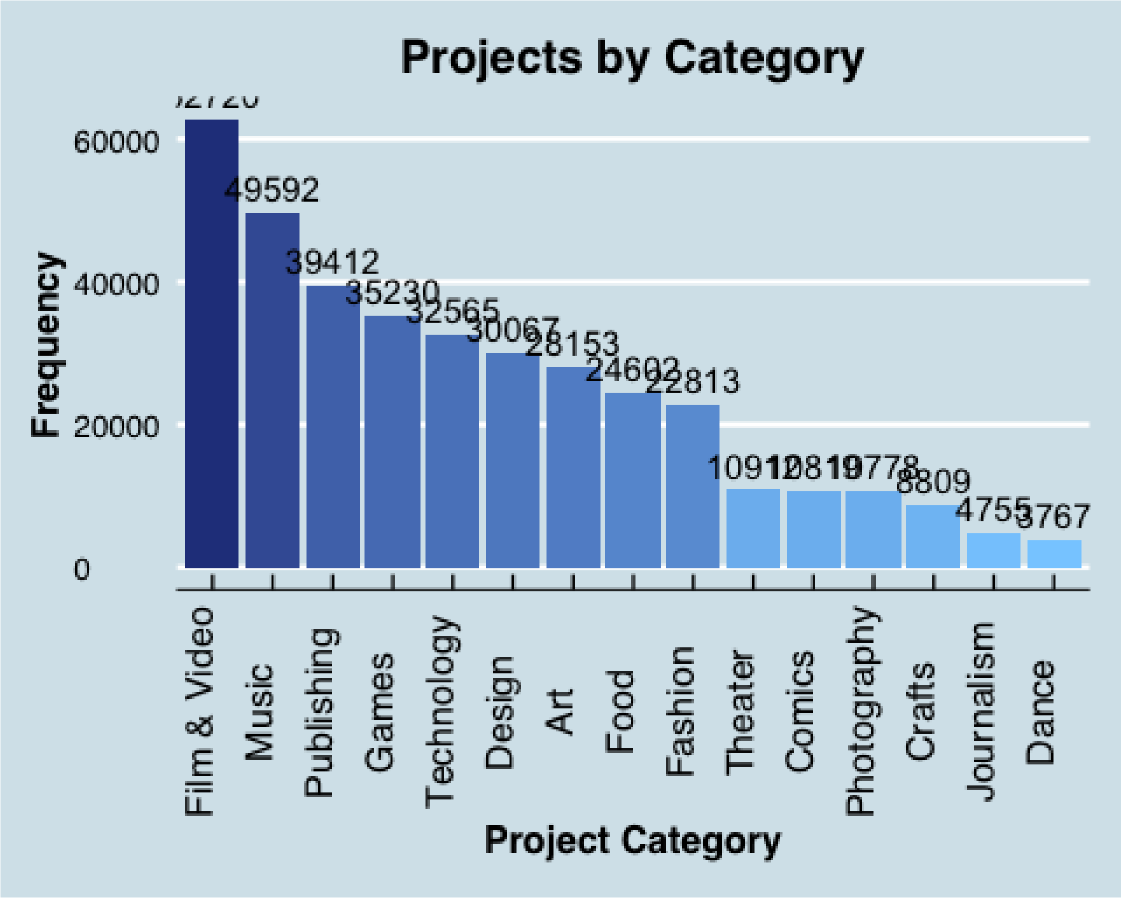
The dataset requires cleansing since some states are set as “undefined” which does not provide any information regarading the success of failure of a project, and this state does not follow the pledges having reached their goals. We are removing a column that does not bring much information and removing any rows where the “undefined” state is set. In the same way we are removing any state set as “successful” which does not have any backers. This does not make sense. We are adding a new column called “launched_year” which uses only the year in the launch column.

```
##              ID            name        category   main_category 
##               0               4               0               0 
##        currency        deadline            goal        launched 
##               0               0               0               0 
##         pledged           state         backers         country 
##               0               0               0               0 
##     usd pledged usd_pledged_real   usd_goal_real 
##            3797               0               0 

## 
##   canceled     failed       live successful  suspended  undefined 
##      38779     197719       2799     133956       1846       3562 

## # A tibble: 6 x 14
##       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
## # … with 7 more variables: launched <dttm>, pledged <dbl>, state
<chr>,
## #   backers <dbl>, country <chr>, usd_pledged <dbl>, usd_goal <dbl>

## [1] 3562
```

Data Exploratory

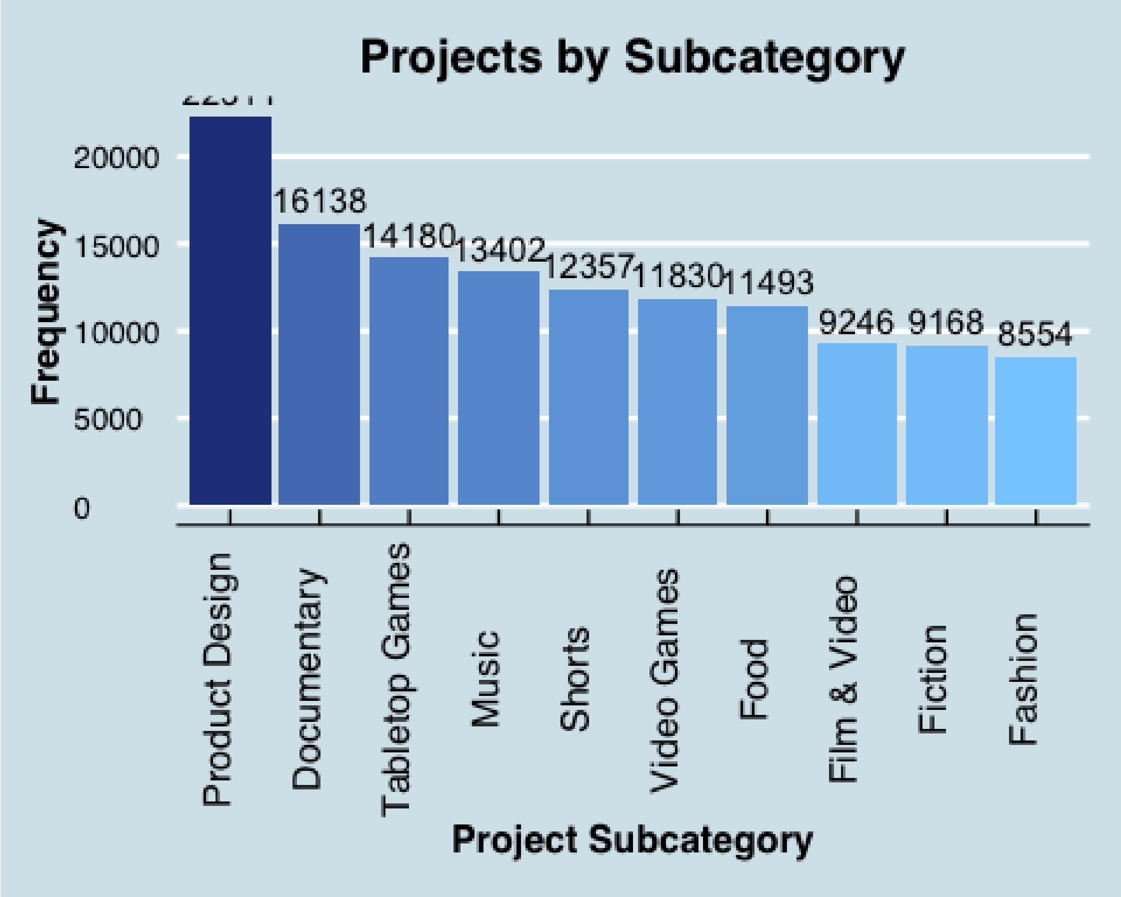
We will look at the projects that are the most popular. The dataset includes 2 columns, “main_category” and “category”. The former provides informtaion in which category a project falls and the latter gives a better insight of the category. The plot here below whows the projects in function of main categories:

```
## # A tibble: 15 x 2
##    main_category count
##    <chr>         <int>
##  1 Film & Video  62720
##  2 Music         49592
##   3 Publishing    39412
##   4 Games         35230
##   5 Technology    32565
##   6 Design        30067
##   7 Art           28153
##   8 Food          24602
##   9 Fashion       22813
##  10 Theater       10912
##  11 Comics        10819
##  12 Photography   10778
##  13 Crafts         8809
##  14 Journalism     4755
##  15 Dance          3767
```



We can see that the most popular Kickstarter campaign falls under the category of "Film & Video" main category. It would be good to get more insight of what this entails exactly. Looking at the projects in function of subcategory we can see, plot below, the "Product Design" is the most popular Kickstarter campaign subcategory.

```
## # A tibble: 159 x 2
##    category       count
##    <chr>          <int>
##  1 Product Design 22311
##  2 Documentary    16138
##  3 Tabletop Games 14180
##  4 Music          13402
##  5 Shorts         12357
##  6 Video Games    11830
##  7 Food           11493
##  8 Film & Video    9246
##  9 Fiction         9168
## 10 Fashion         8554
## # … with 149 more rows
```



Althougt the “Product Design” is the most popular subcategory, we can view the 15 campaigns that are the most pledged:

```
## # A tibble: 15 x 4
##             ID name                                 category
usd_pledged
##          <dbl> <chr>                                <chr>
<dbl>
##  1    1.80e9 Pebble Time - Awesome Smartwatch, No… Product Des…
20338986.
##  2    3.43e8 COOLEST COOLER: 21st Century Cooler … Product Des…
13285226.
##  3    2.10e9 Pebble 2, Time 2 + All-New Pebble Co… Product Des…
```

```
12779843.
##  4    5.45e8 Kingdom Death: Monster 1.5           Tabletop Ga…
12393140.
##  5    5.07e8 Pebble: E-Paper Watch for iPhone and… Product Des…
10266846.
##  6    5.66e8 The World's Best TRAVEL JACKET with … Product Des…
9192056.
##  7    1.96e9 Exploding Kittens                     Tabletop Ga…
8782572.
##  8    1.03e9 OUYA: A New Kind of Video Game Conso… Gaming Hard…
8596475.
##  9    6.47e8 THE 7th CONTINENT – What Goes Up, Mu… Tabletop Ga…
7072757
## 10    4.50e8 The Everyday Backpack, Tote, and Sli… Product Des…
6565782.
## 11    1.39e9 Fidget Cube: A Vinyl Desk Toy         Product Des…
6465690.
## 12    9.48e8 Shenmue 3                             Video Games
6333296.
## 13    1.15e9 Pono Music - Where Your Soul Redisco… Sound
6225355.
## 14    1.45e9 Bring Back MYSTERY SCIENCE THEATER 3… Television
5764229.
## 15    1.76e9 The Veronica Mars Movie Project       Narrative F…
5702153.
```

We can see that that most of those most pledged projects fall under the “Product Design” category.

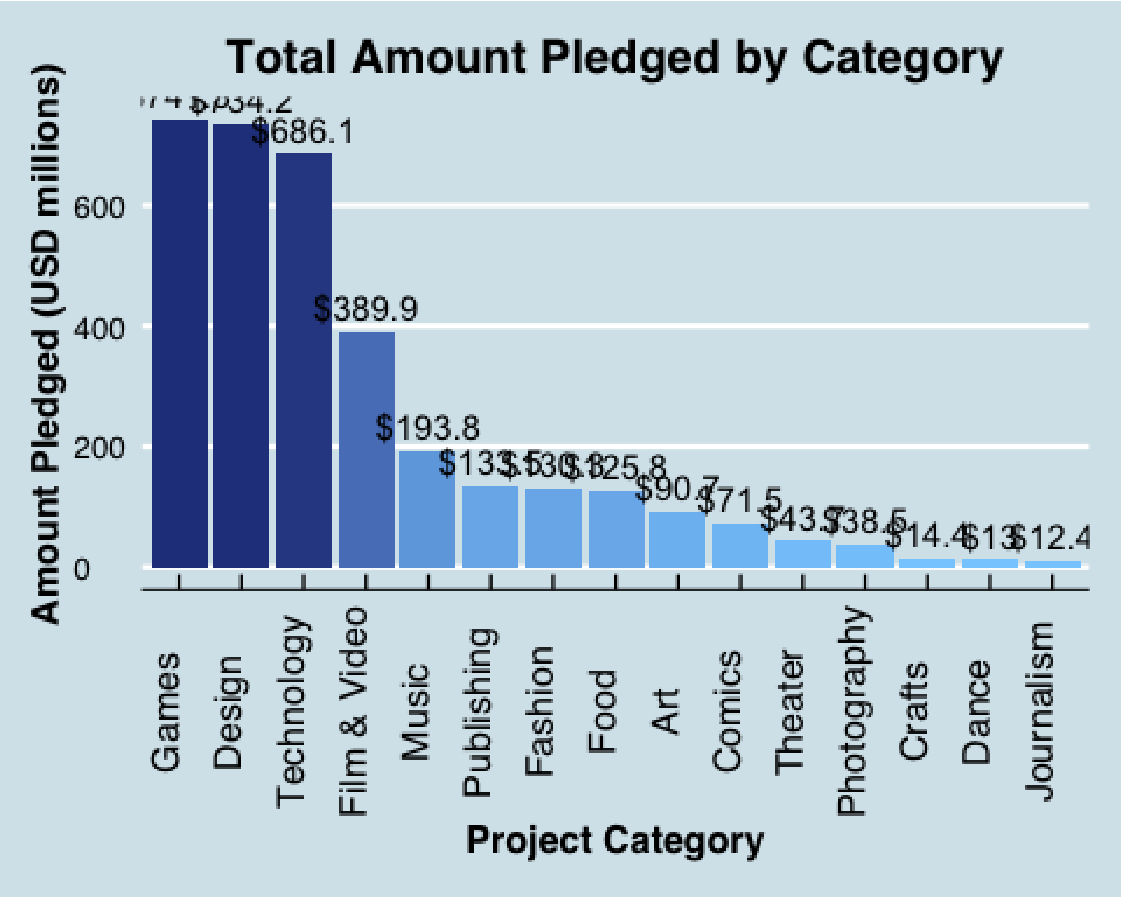
In the same way here below we can see the most popular projects being backed up:

```
## # A tibble: 15 x 3
##    name                                                category
backers
##    <chr>                                               <chr>
<dbl>
##  1 Exploding Kittens                                   Tabletop Gam…
219382
##  2 Fidget Cube: A Vinyl Desk Toy                       Product Desi…
154926
##  3 Bring Reading Rainbow Back for Every Child, Every… Web
105857
##  4 The Veronica Mars Movie Project                     Narrative Fi…
91585
##  5 Double Fine Adventure                               Video Games
87142
##  6 Bears vs Babies - A Card Game                       Tabletop Gam…
85581
##  7 Pebble Time - Awesome Smartwatch, No Compromises   Product Desi…
78471
##  8 Torment: Tides of Numenera                          Video Games
```

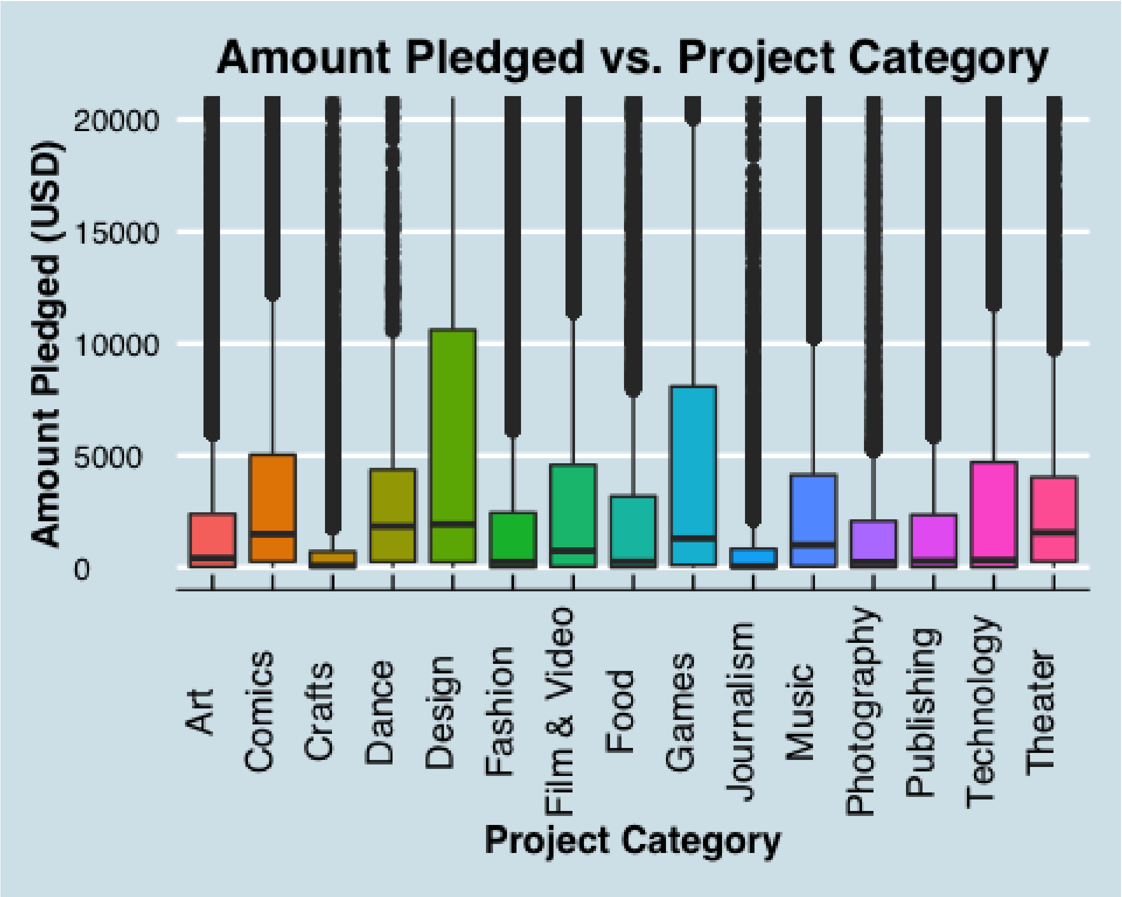
```
74405
##  9 Project Eternity                                      Video Games
73986
## 10 Yooka-Laylee - A 3D Platformer Rare-vival!            Video Games
73206
## 11 ZNAPS -The $9 Magnetic Adapter for your mobile de… Technology
70122
## 12 Shenmue 3                                             Video Games
69320
## 13 Pebble: E-Paper Watch for iPhone and Android          Product Desi…
68929
## 14 Mighty No. 9                                          Video Games
67226
## 15 Pebble 2, Time 2 + All-New Pebble Core                Product Desi…
66673
```

We can see here that a few projects listed here are under the “Video Games” category.

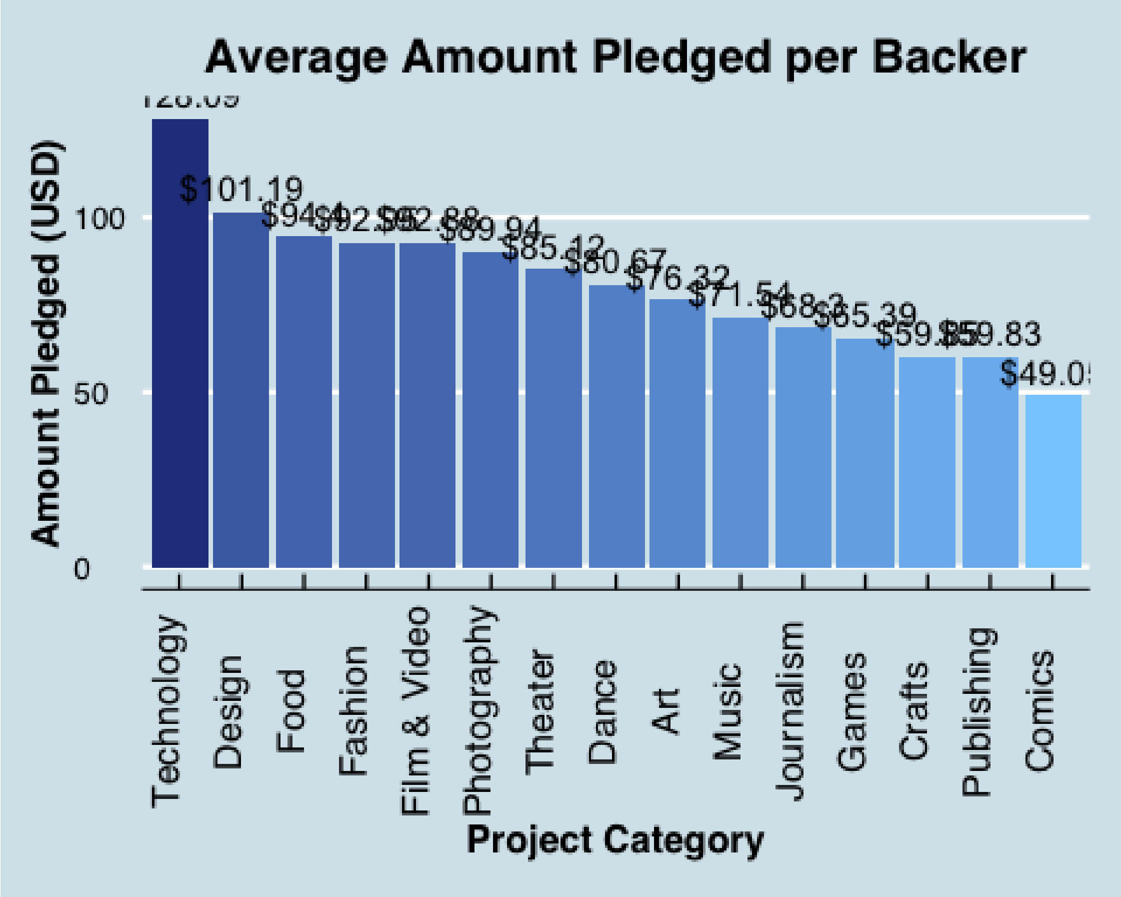
We can now look at the project categories in function of the amount pledged. This represents the categories requiring the most money:



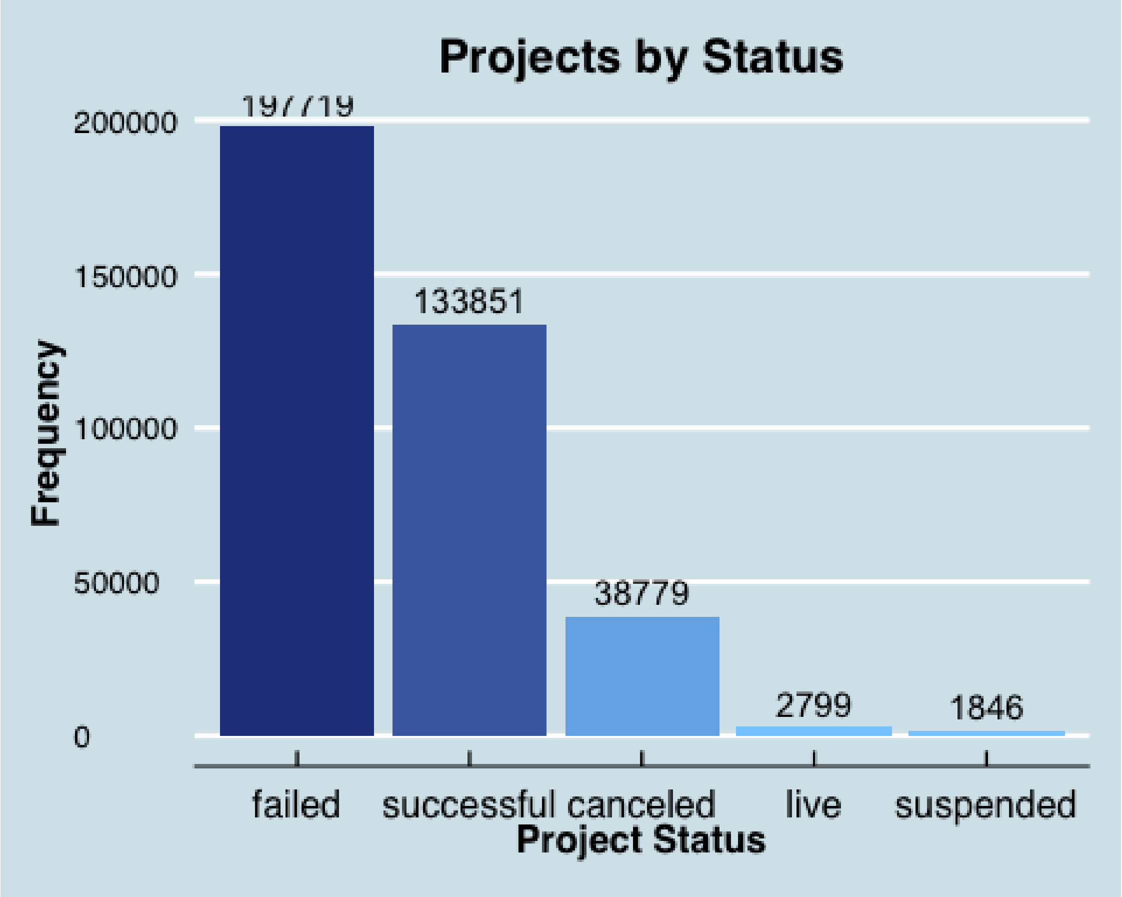
We can see that “Games”, “Design” and “technology” are the categories that total the most revenues. The projects pledged can be now visulaized per category using the boxplot. From this graph, here below, it is possible to see that outliers exist as well as some displersion and skeweness in the data.



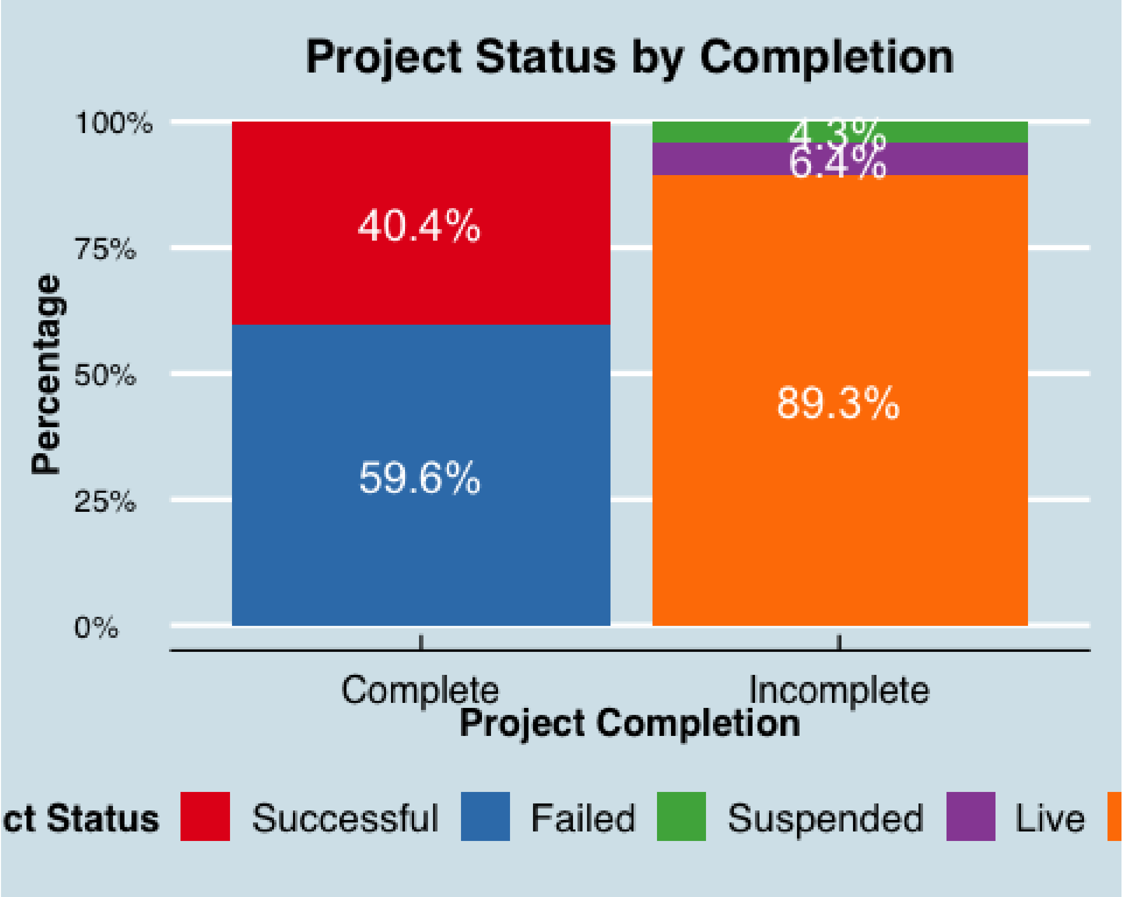
Projects in Design represent the biggest interquantile range followed by Games and technology. Now it will be good to understand which categories are backed the most, in order of preference. The graph below shows that projects falling under "Technology" are the most backed up.



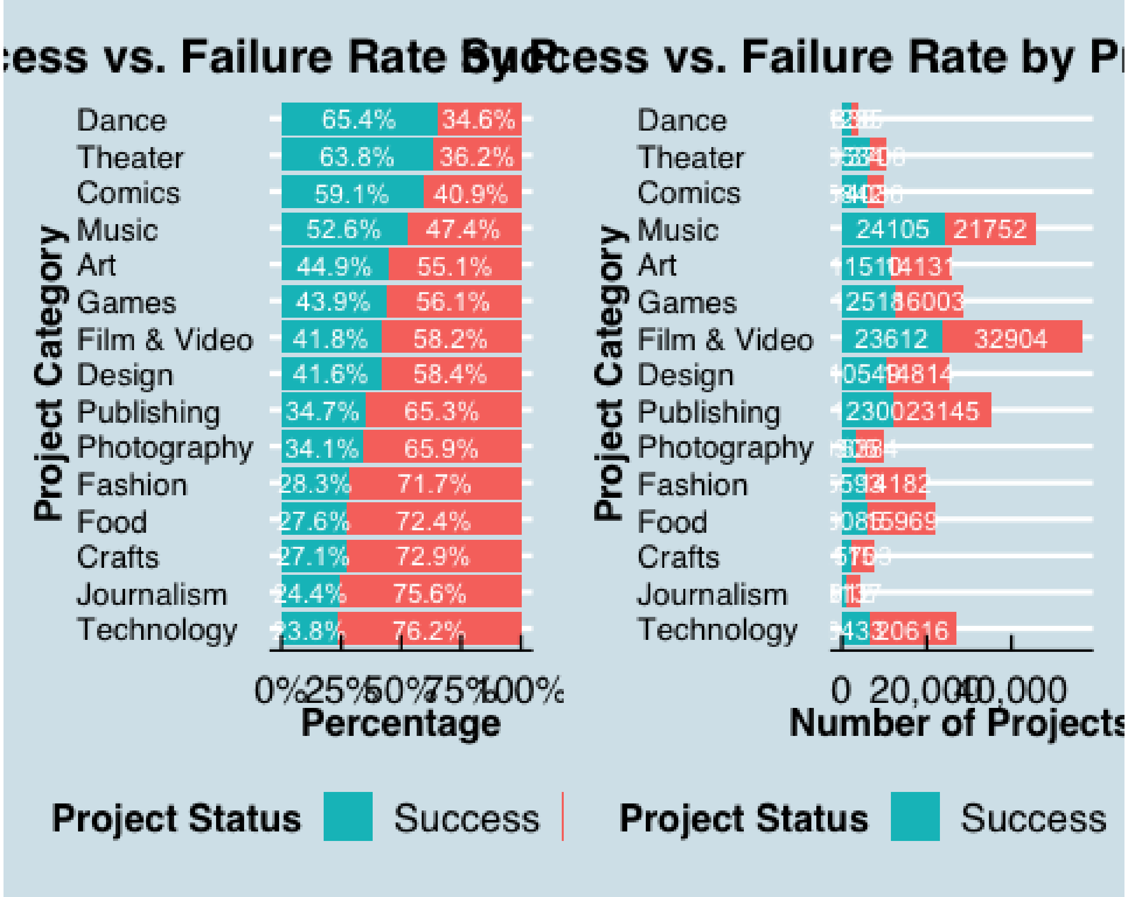
We can now look at the status of the different projects run in Kickstarter, see graph below. The majority of projects have failed, the successful projects are about 67% of the proportion of failed projects.



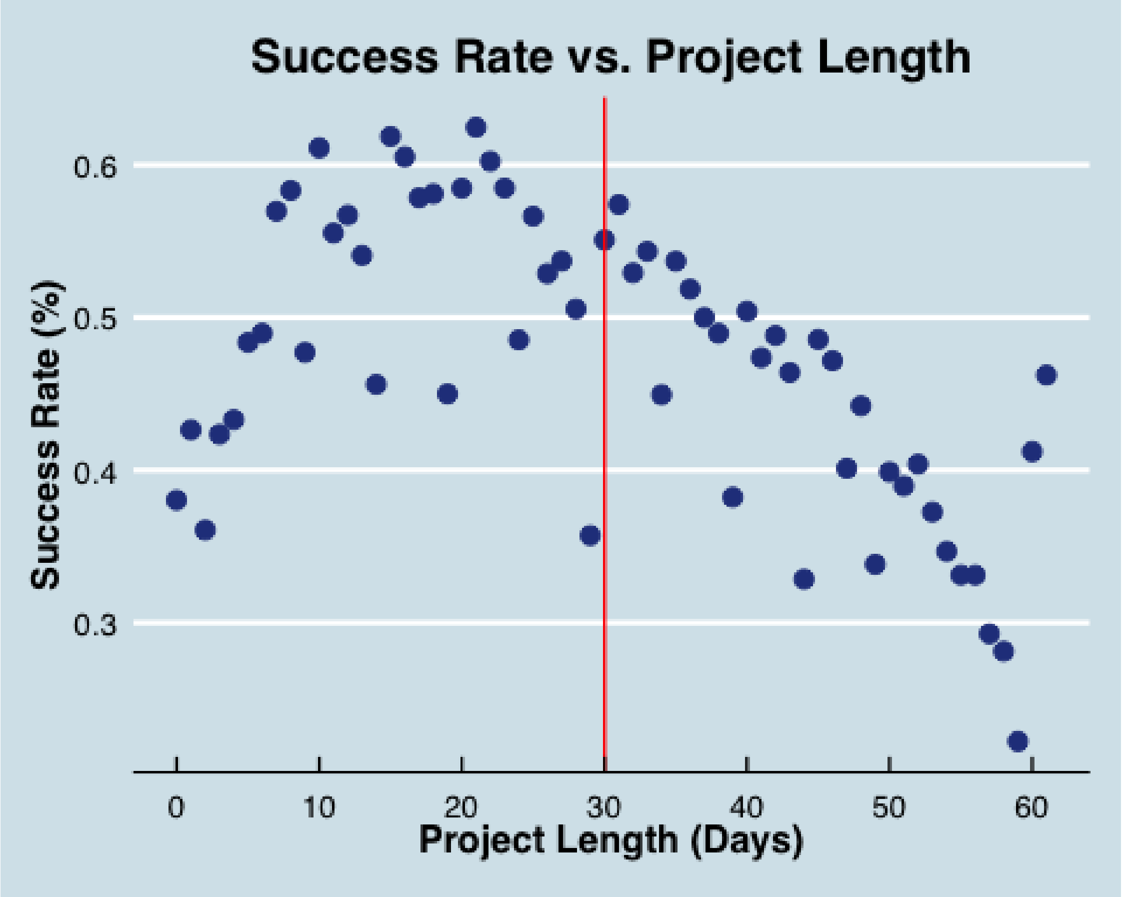
Since there are different status existing in this dataset, we can see the porportion of completed projects versus the ones that have been stopped before.



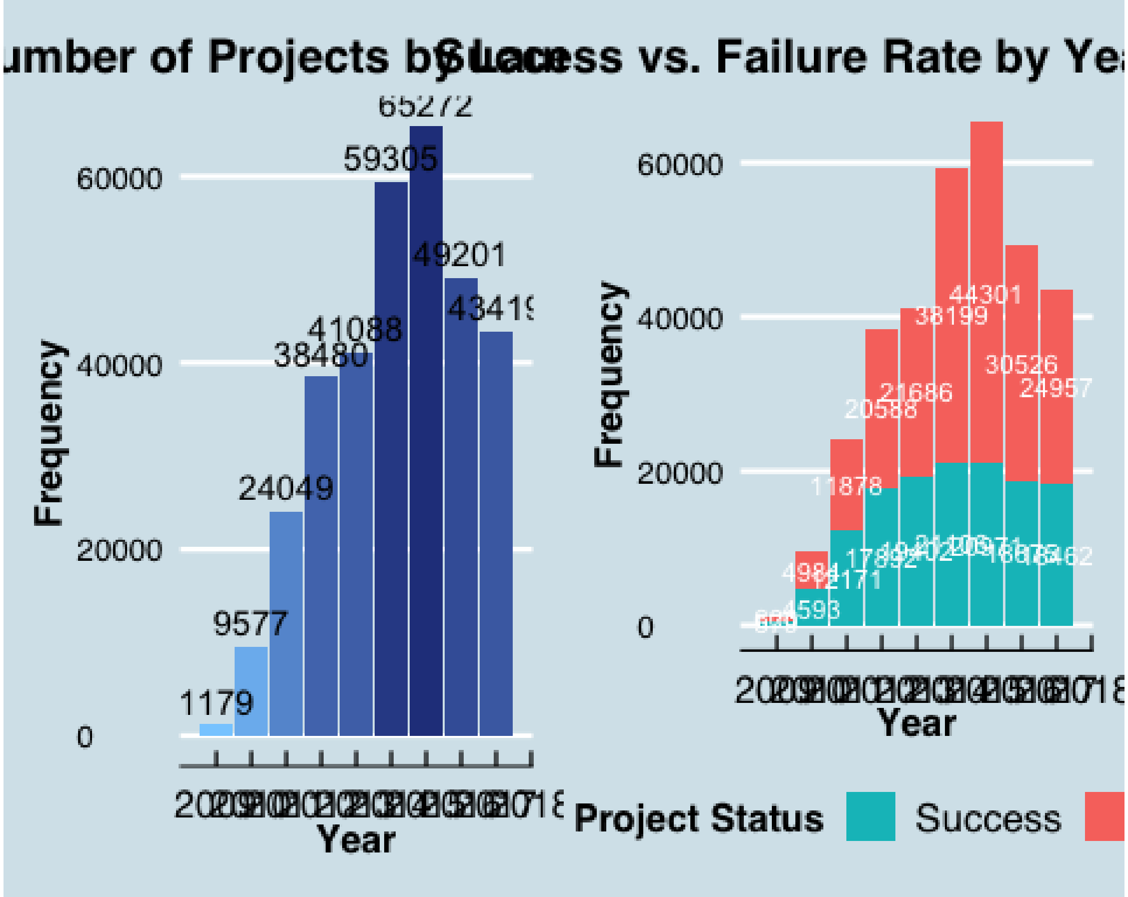
This graph is much clearer in terms of the success rate of a porject in Kickstarter irrespective of the category the project falls into. Only ~40% of the projects are successful and ~90% of the porjects were cancelled. We can now see the porject success ratio per category as well as their numbers. Projects in categories such as Dance, Theater and Comics have a success rate of 60% or above, however their numbers are much lower than other projects.



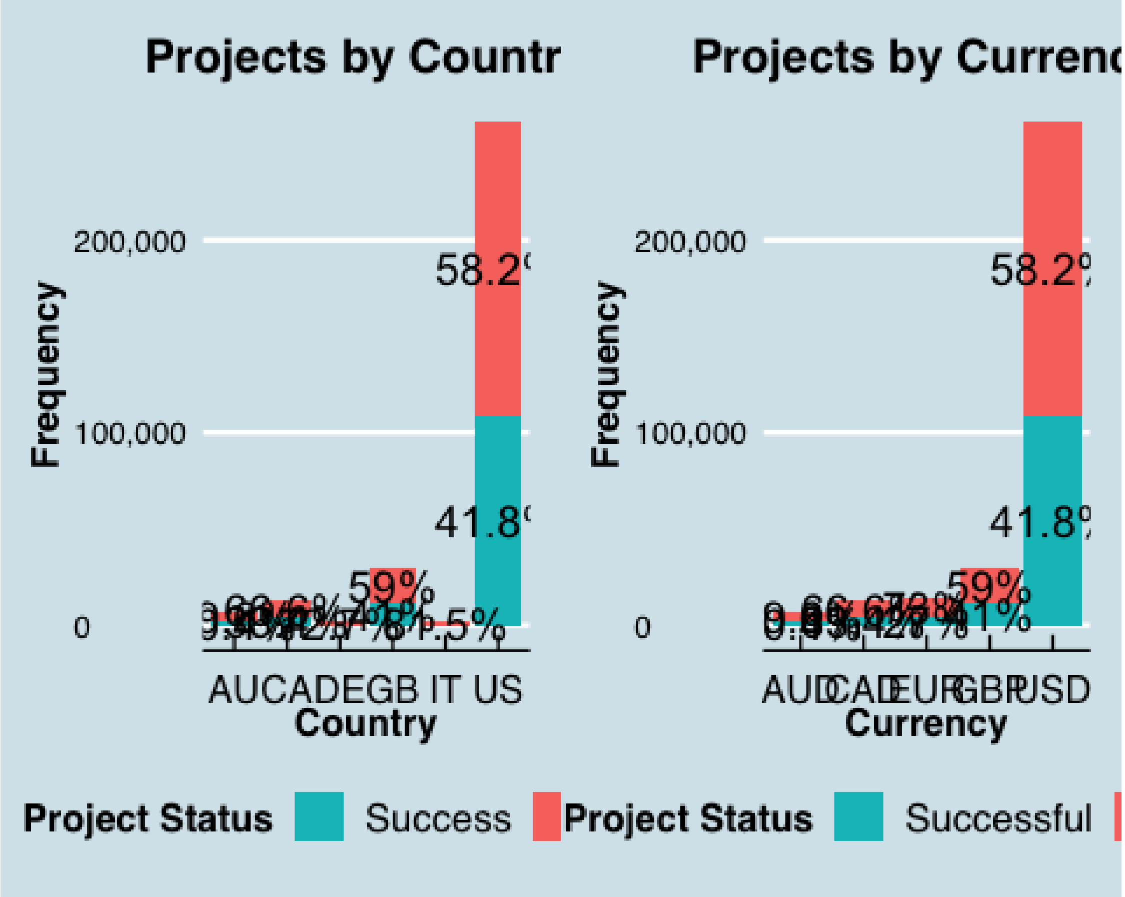
Although Kickstarter rules for the maximum project length is 60 days, they recommend to set the length of time for 30 days. From our dataset we can see that in geenral any project lasting more than 30 days have a lower sucess rate from the ones lasting between 5 to 30 days.



Projects also varies in time and the graph, here below, shows the success rate of the project over time. 2014 and 2015 have been the years where the most porjects were launched however looking at the sucess ratio, we can see that the failure is quite constant over the years.

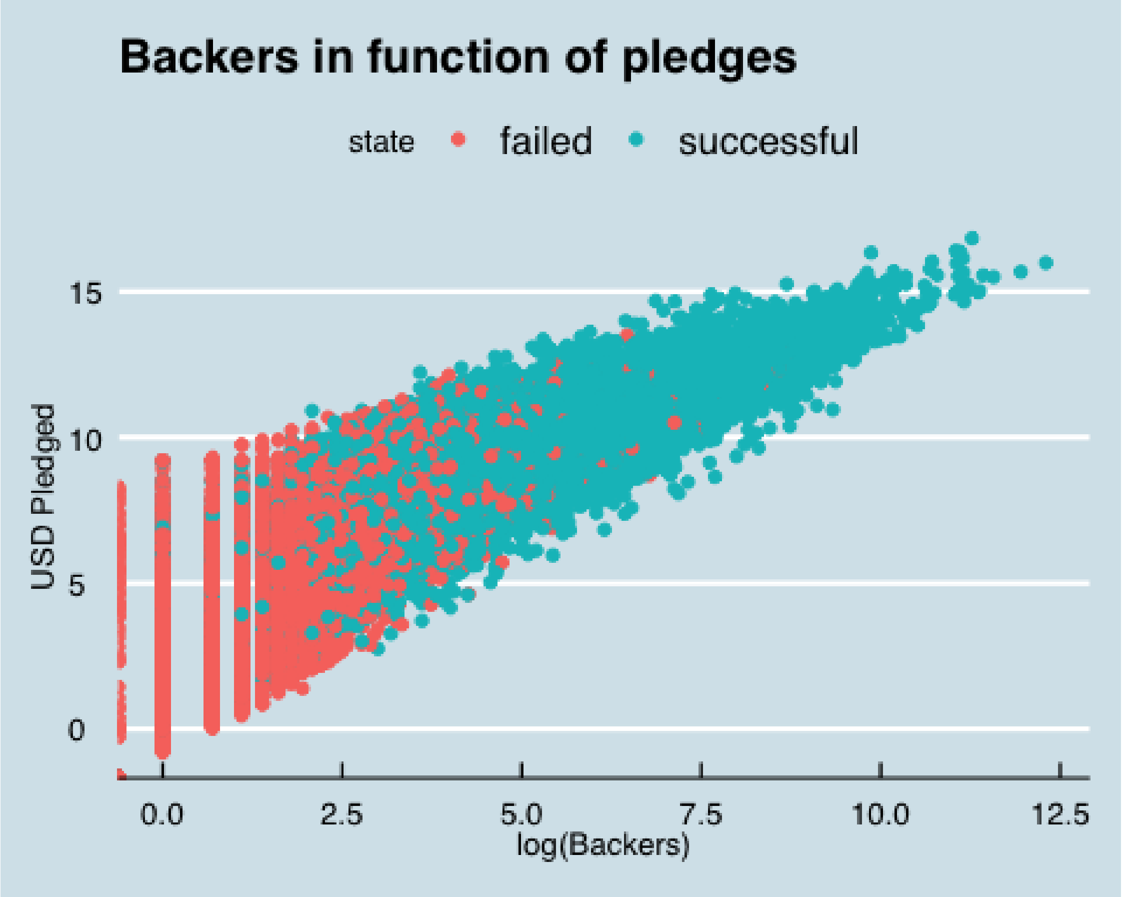


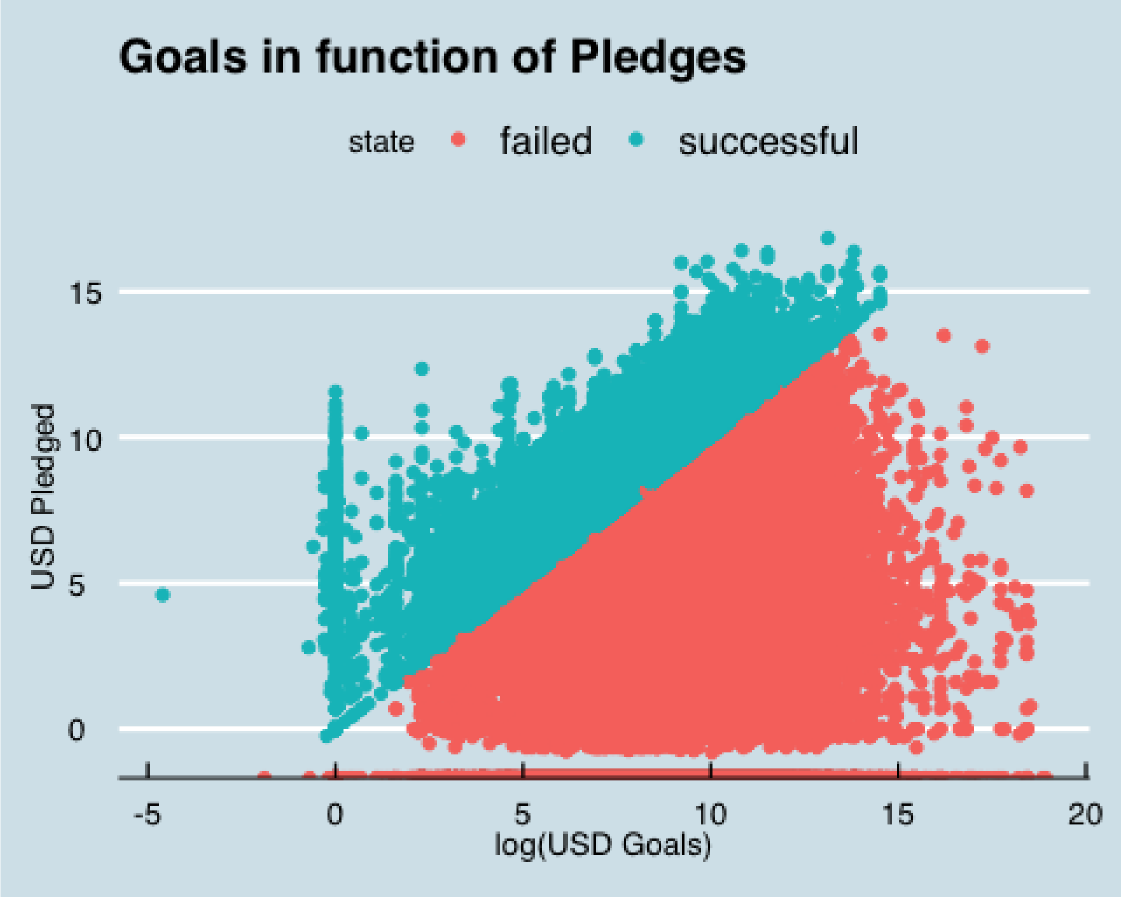
Projects vary also per country and currency. The majority of projects are coming out from the USA followed by Great Britain and Canada. In terms of the currency the most used currency is the US dollar, then quite far behind is the British pounds and then the Euro. The success rate is around 40% for either projects launched in the US and Great Britain or in US dollar and British pounds. Projects launched in Euros have a success rate of only 27% though.

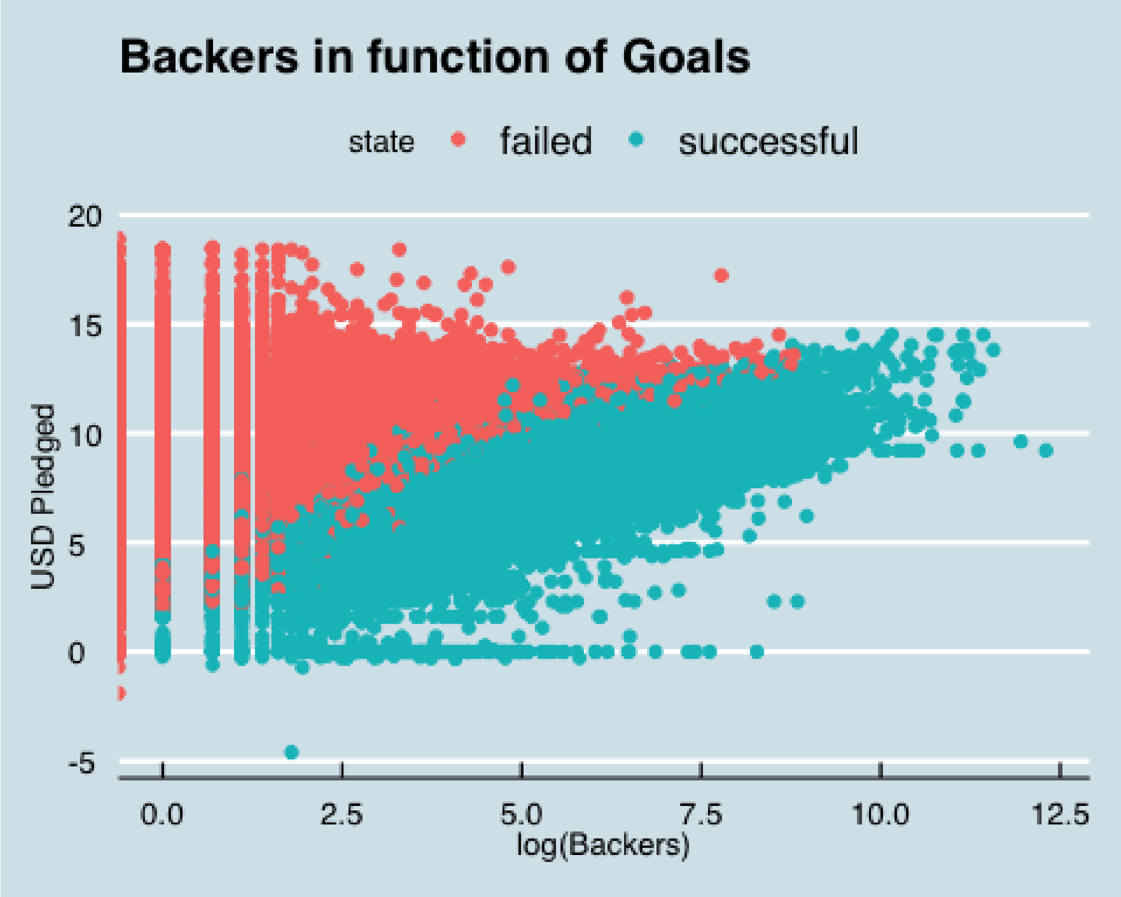


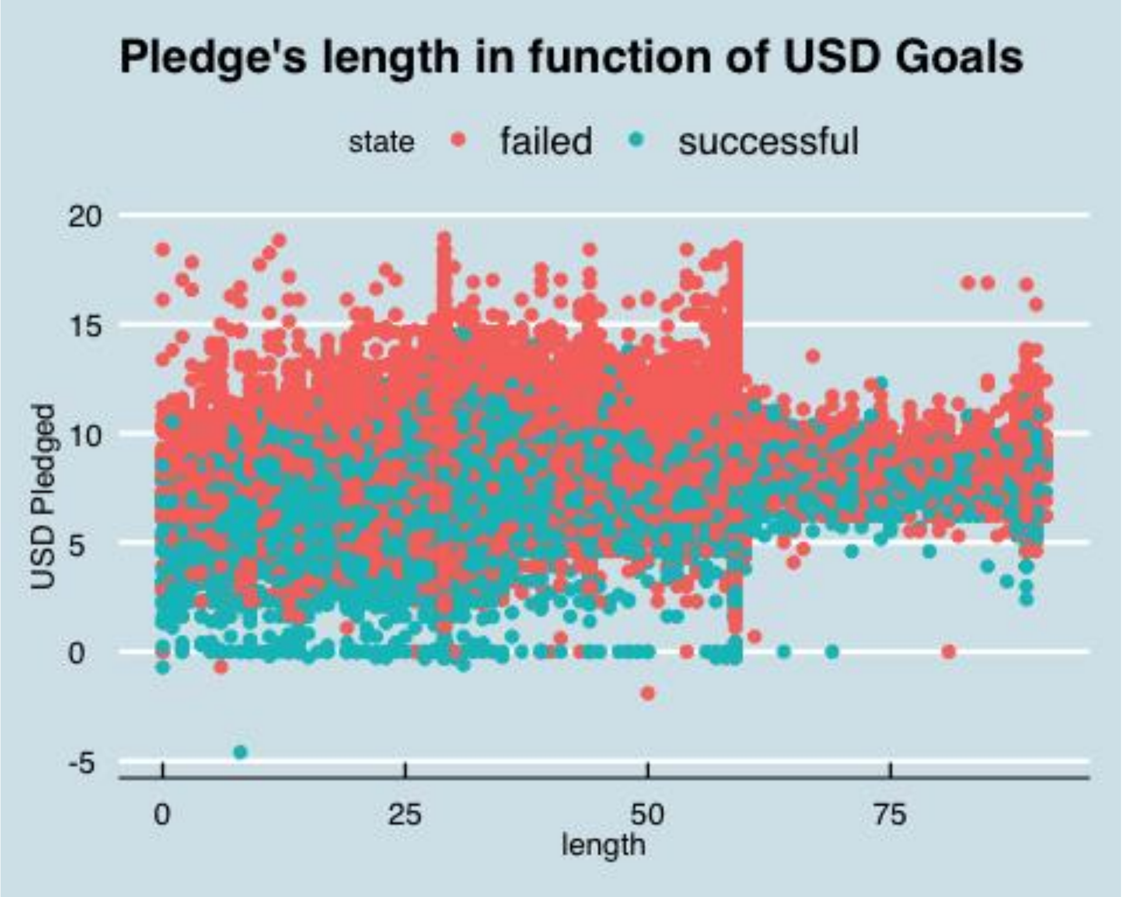
Data correlation

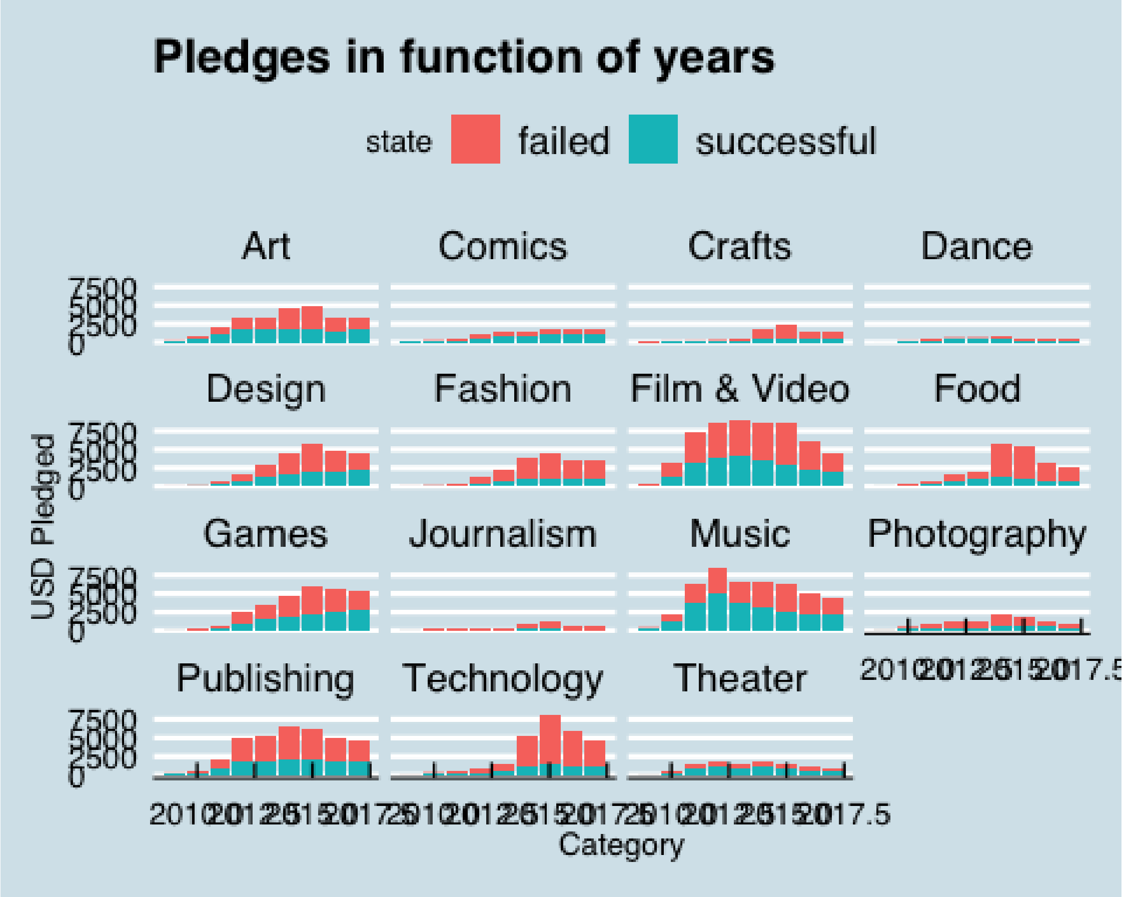
Now that we have view different dataset, it is possible to see whether some data have a high correlation between them of now. THis can be done graphically but also tabularly.


Backers in function of pledges
state
failed
successful
USD Pledged
15
10
5
0
0.0
2.5
5.0
7.5
10.0
12.5
log(Backers)


Goals in function of Pledges
state
failed
successful
USD Pledged
15
10
5
0
-5
0
5
10
15
20
log(USD Goals)


Backers in function of Goals
state
failed
successful
20
15
10
5
0
-5
USD Pledged
0.0
2.5
5.0
7.5
10.0
12.5
log(Backers)


Pledge's length in function of USD Goals
state
failed
successful
USD Pledged
20
15
10
5
0
-5
0
25
50
75
length



As for the graphs we can see that the best correlation is obtained between pledge and backers.

```
##                   backers usd_pledged    usd_goal launched_year
## backers       1.000000000 0.753449599 0.004476723    0.01675119
## usd_pledged   0.753449599 1.000000000 0.005566034    0.02220426
## usd_goal      0.004476723 0.005566034 1.000000000    0.01241036
## launched_year 0.016751194 0.022204257 0.012410361    1.00000000
```

Modelling part

We saw that some correlations exist in the dataset, we are now going to use some models to fit a subset of the dataset. Here we use 10% of the dataset for the training set.

```
## Warning in set.seed(77777777, sample.kind = "Rounding"): non-uniform
## 'Rounding' sampler used
```

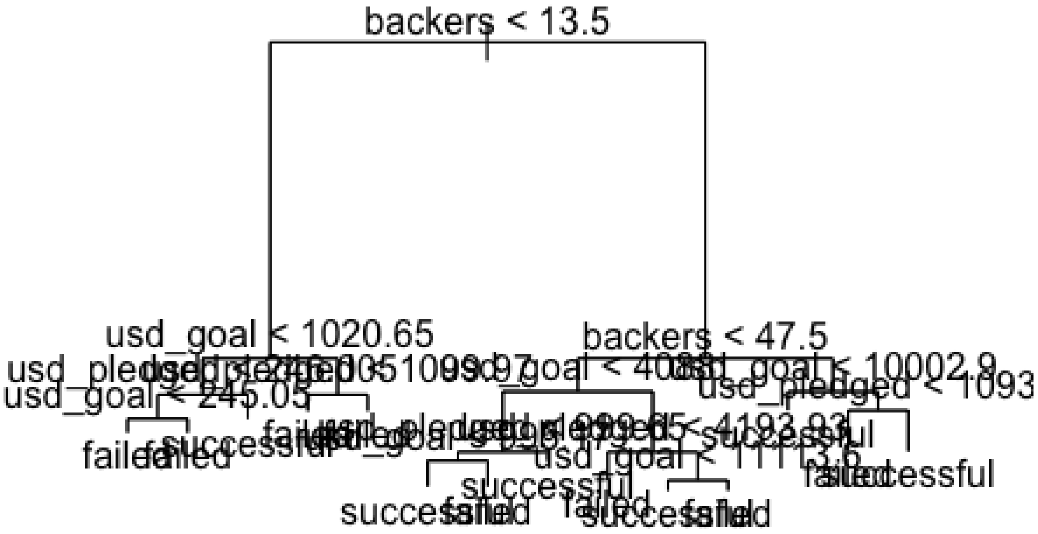
Some models such as K-Neareast Neighbor (KNN), Random Forest (RF), New View (NV) and Support Vector Machine (SVM) take too long to find convergence and we are just using 2 that provide relatively quickly some outcomes. Those models are Logistic Regression (LDA) and Classification And Regression Trees (CART).

```
##
## Call:
## summary.resamples(object = results)
##
## Models: lda, cart
## Number of resamples: 25
##
## Accuracy
##           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## lda  0.6506395 0.6541737 0.6566998 0.6565241 0.6586422 0.6633921    0
## cart 0.9009689 0.9023279 0.9242312 0.9147175 0.9259007 0.9290247    0
##
## Kappa
##           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## lda  0.2220851 0.2279128 0.2329633 0.2337572 0.2397269 0.2476707    0
## cart 0.7944641 0.7981545 0.8419328 0.8226543 0.8449652 0.8519687    0
```

We can see the results of both LDA and CART training models on the training dataset. We can look at the most influencial predictors in this training set using the tree methodology, as see below. The most influencial predictors: - backers - usd goals - usd pledged

We then apply the model to the rest of the dataset to look at the accuraacy of our model. We can see that a good accuracy is obtained, with an area under the curve of above 0.95.

```
##
## Classification tree:
## tree(formula = state ~ ., data = trainset)
## Variables actually used in tree construction:
## [1] "backers"     "usd_goal"    "usd_pledged"
## Number of terminal nodes:  14
## Residual mean deviance:  0.2937 = 78890 / 268600
## Misclassification error rate: 0.0488 = 13109 / 268607
```



```
##
##                  0     1
##   failed     37221  2323
##   successful   913 25859

## Setting levels: control = failed, case = successful

## Setting direction: controls < cases

## Area under the curve: 0.9536
```