KickStarter Analysis

Alex Flores

October 10, 2018

Contents

What is KickStarter?

Kickstarter is a website which allows users to post their projects online in the hopes of reaching a fund goal via donations. The projects are funded through a crowdfunding campaign which is essentially raising small amounts of money through large amounts of people. If you would like to learn more, please visit the following link: (https://www.kickstarter.com/)

Load the Data

The first thing to do is to load the data. I use the "head" function first to get an idea of what I am working with. From the following table, you are only seeing a tiny sample of the data I am actually working with. If you would like to see the data for yourself, please visit the following link: (https://www.kaggle.com/kemical/kickstarter-projects)

Note: A kickstarter is considered successful if it reaches it fund goal

```
setwd("C:\\Users\\flore\\Desktop\\Job Search Documents\\Project2R")
library("tidyverse")
library("dplyr")
library("stringr")
library("knitr")
library("kableExtra")
library("ggplot2")
```

```
data<-read.csv("KickStarterData.csv")
head(data[1:10,1:4])
```

ID	name	category	main_category
1000002330	The Songs of Adelaide & Abullah	Poetry	Publishing
1000004038	Where is Hank?	Narrative Film	Film & Video
1000007540	ToshiCapital Rekordz Needs Help to Complete Album	Music	Music
1000011046	Community Film Project: The Art of Neighborhood Filmmaking	Film & Video	Film & Video
1000014025	Monarch Espresso Bar	Restaurants	Food
1000023410	Support Solar Roasted Coffee & Green Energy! SolarCoffee.co	Food	Food

Understanding the Data we are working with

From the original data I downloaded, I draw two conclusions. The first is that the data is not complete because we have undefined values and we need to filter this out. The next conclusion I draw is that I'm not going to factor in the "live" kickstarters because the end result of the "live" kickstarters are still unknown.

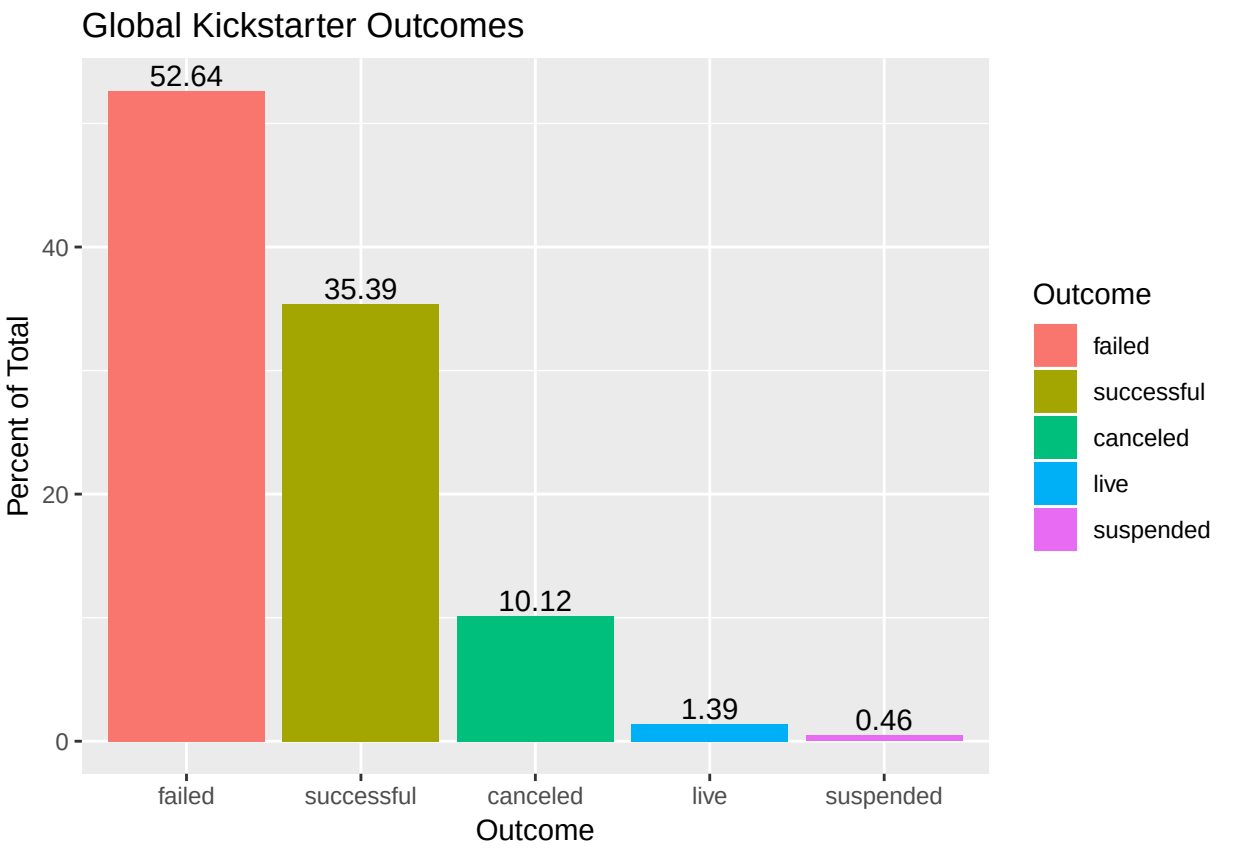
```
data<-data #You need to save the data into an object each time you use it in R
data$launched<- str_sub(data$launched,start=1,end=-16)
data$launched<- as.character(data$launched)
data$launched<- as.integer(data$launched)
data$state<-as.character(data$state)
data<- data %>%
  filter(state %in% c("successful","canceled","failed","live","suspended")) %>%
  filter(launched!="undefined")
```

```
data1<-data
datatable<-table(data1$state)
datatable <-tail(datatable)
dataframe <- data.frame(datatable)
colnames(dataframe) <- c("Outcome", "Count")
dataframe2 <- dataframe %>%
  group_by(Outcome) %>%
  summarise(Percent_of_Total=(round(Count/sum(dataframe$Count),digits=4)*100)) %>%
  arrange(desc(Percent_of_Total))
dataframe2$Outcome <- factor(dataframe2$Outcome,
                             levels=dataframe2$Outcome
                             [order(dataframe2$Percent_of_Total,
                                    decreasing = TRUE)])
Global_Kickstarter_Outcomes<-ggplot(dataframe2,
                                    aes(x=Outcome, y=Percent_of_Total, fill=Outcome)) +
   geom_bar(stat="identity") +
   labs(y="Percent of Total",x="Outcome") +
   ggtitle("Global Kickstarter Outcomes") +
   geom_text(aes(label=Percent_of_Total), vjust=-0.25)
```

Global Kickstarter Outcomes: Bar Graph

Around the world, failed kickstarters outnumber any other outcome. To my suprise though, I see that failed kickstarters account for only ~15% more of the overall data in comparison to successful kickstarters. I originally anticipated that failed kickstarters would take up a much larger chunk of the data.

```
Global_Kickstarter_Outcomes
```



Important Assumption

I don't have a column representing the exact origin of each kickstarter, therefore I am going to be using the currency the kickstarter accepts as a proxy for which country the kickstarter was made in. The following code will change the currency names to the respective countries names to make the following data easier to understand.

```
AdjustedData <- data[c(5,8,10)]
colnames(AdjustedData) <- c("Country", "launched","state")
AdjustedData$Country <- gsub('AUD', 'AU', AdjustedData$Country)
AdjustedData$Country <- gsub('CAD', 'CA', AdjustedData$Country)
AdjustedData$Country <- gsub('GBP', 'GB', AdjustedData$Country)
AdjustedData$Country <- gsub('NZD', 'NZ', AdjustedData$Country)
AdjustedData$Country <- gsub('SEK', 'SW', AdjustedData$Country)
AdjustedData$Country <- gsub('USD', 'US', AdjustedData$Country)
```

Which countries had the highest success ratio?

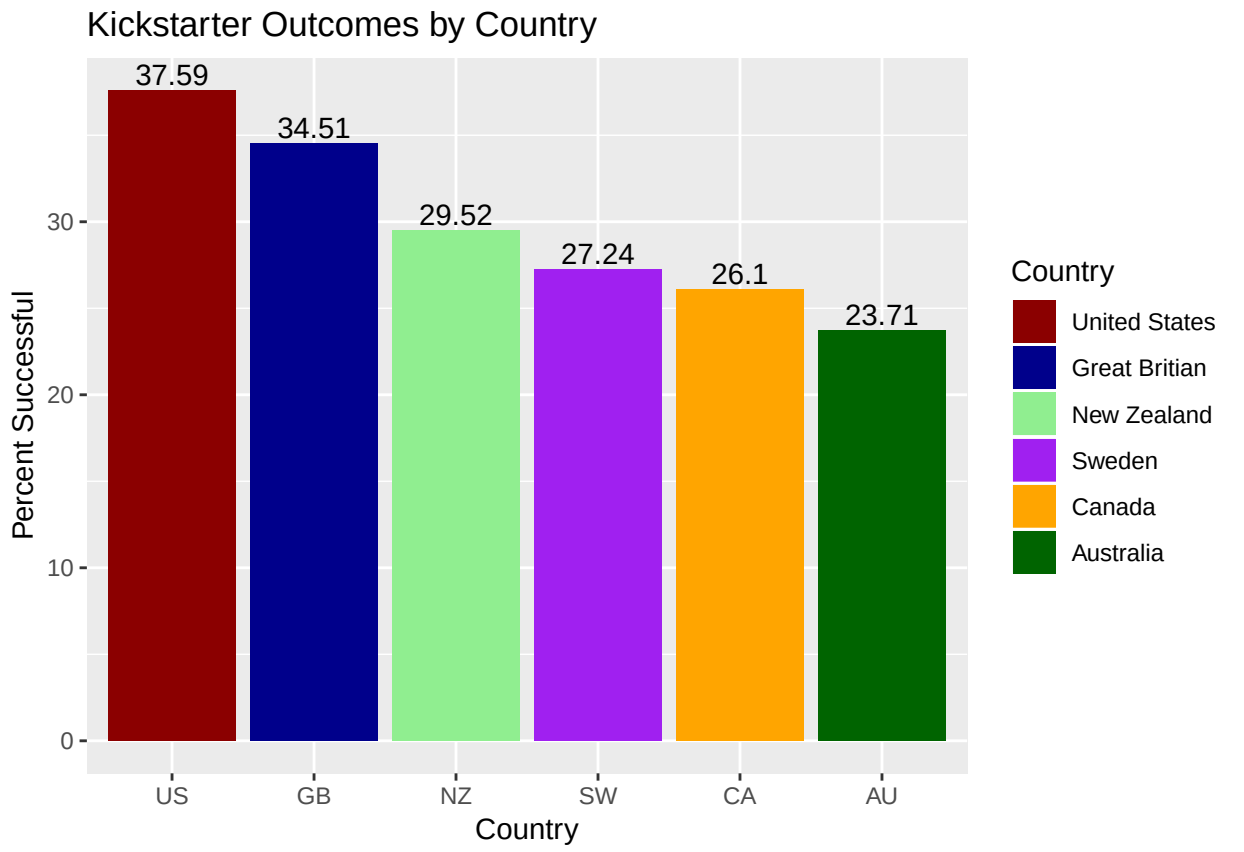
To make this code, I filter out countrys with less than 1000 observations in the dataset because I want to keep my data clear of sampling bias. I also filter out Europe because we are purely looking at countrys.

```
DataCountryState<-AdjustedData %>%
  filter(Country!="EUR") %>% #Europe is a continent, not a country
  filter(launched>2000) %>%
  filter(state %in% c("successful","canceled","failed","suspended")) %>%
  group_by(Country) %>%
  summarise(Percent_Successful=(round((sum(state=="successful")/
                                       sum(state %in% c("successful","canceled",
                                                        "failed","suspended"))),
                                       digits=4)*100),Total_Observations=n()) %>%
  arrange(desc(Total_Observations)) %>%
  filter(Total_Observations > 1000)
DataCountryState$Country <- factor(DataCountryState$Country,
                                   levels=DataCountryState$Country
                                   [order(DataCountryState$Percent_Successful,
                                          decreasing = TRUE)])
Kickstarer_Outcomes_by_Country <- ggplot(DataCountryState,
                                         aes(x=Country, y=Percent_Successful,
                                             fill=Country)) +
  geom_bar(stat="identity") +
  labs(y="Percent Successful",x="Country") +
  ggtitle("Kickstarter Outcomes by Country") +
  scale_fill_manual(labels = c("United States", "Great Britian","New Zealand",
                               "Sweden","Canada","Australia"),
                    values = c("dark red","dark blue","light green",
                               "purple","orange","dark green")) +
   geom_text(aes(label=Percent_Successful), vjust=-0.25)
```

Kickstarter success by Country: Bar Graph

The USA comes in first place with the best chance of a kickstarter succeeding at 37.59%. Great Britian comes in at a close second while Australia falls far behind in overall kickstarter success.

```
Kickstarer_Outcomes_by_Country
```



What was the success to failure ratio over time?

In order to write this code, I will first create a dataset that sums the total amount of successful kickstarters then divides it by the total amount of kickstarters for each country. Then I will group by the Year the kickstarter was launched and summarize the data with the success ratios over time for each country. From the following data table that is created, I can make line graphs using the ggplot package.

```
newData <- AdjustedData[c(2,3,1)]
colnames(newData) <- c("Year", "state","Country")
newData$Country<- as.character(newData$Country)
newData$state<- as.character(newData$state)
newData<-newData %>%
  filter(Year>2000) %>%
  group_by(Year) %>%
  filter(state %in% c("successful","canceled","failed","suspended"))
over_time<-newData %>%
  filter(Country %in% c("US","GB","NZ","SW","CA","AU")) %>%
  group_by(Country,Year) %>%
  summarise(Percent_Successful=(round((sum(state=="successful")/
                                        sum(state %in% c("successful","canceled",
                                                         "failed","suspended"))),
                                      digits=4)*100))
graph_data<-over_time
over_time$Percent_Successful <- (paste0(over_time$Percent_Successful,"%"))
over_time <- over_time %>%
  kable() %>%
  kable_styling(c("striped", "bordered"))
```

Kickstarter success over time: Data Table

Looking at the outliers, we can see that the best year to have launched a kickstarter was in the USA, in 2011 with an impressive 46.43% success rate. The worst time to have launched a kickstarter was in Sweden, in 2014 with only a 19.64% success rate. Our next step should be to make a graph of the data because it will allow us to easily spot the trends over time.

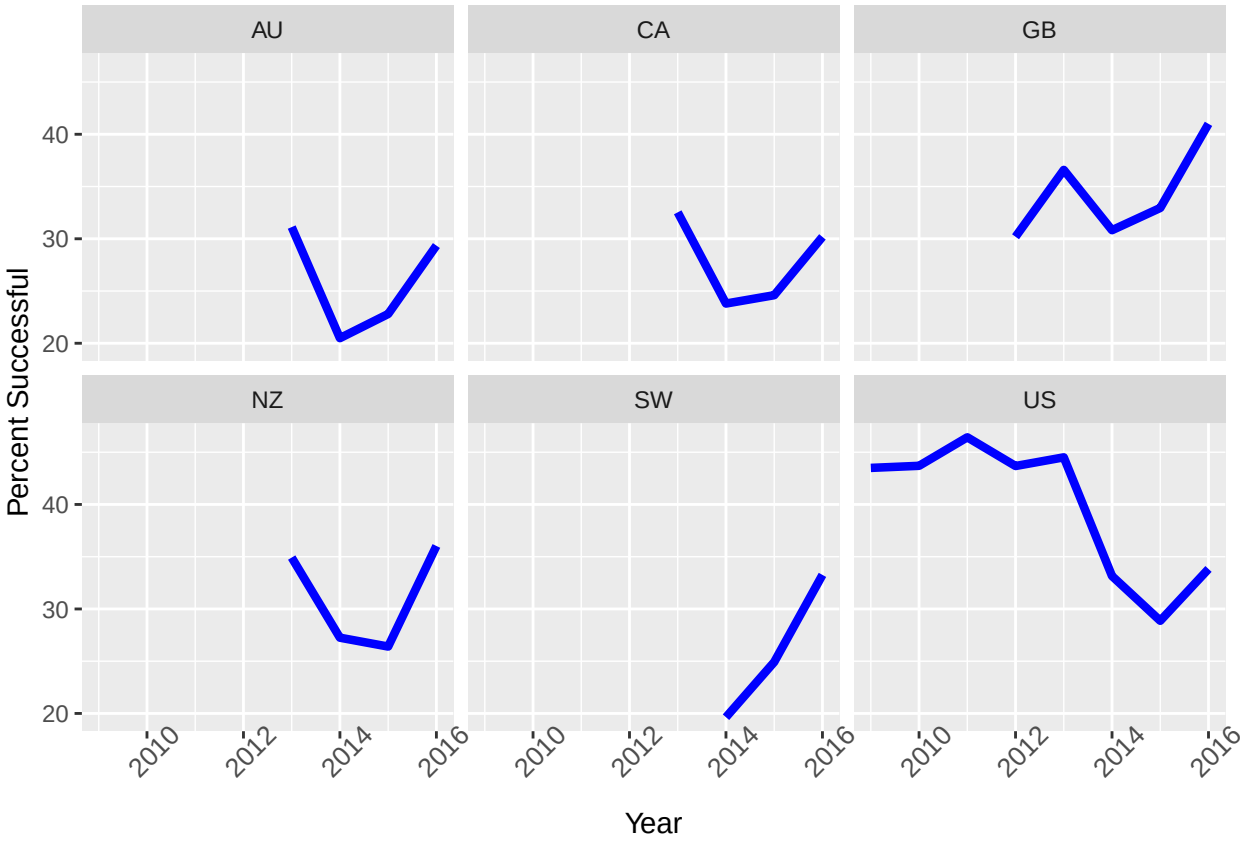
```
over_time
```

Country	Year	Percent_Successful
AU	2013	31.12%
AU	2014	20.49%
AU	2015	22.8%
AU	2016	29.35%
CA	2013	32.53%
CA	2014	23.8%
CA	2015	24.6%
CA	2016	30.18%
GB	2012	30.19%
GB	2013	36.59%
GB	2014	30.82%
GB	2015	32.95%
GB	2016	41%
NZ	2013	34.92%
NZ	2014	27.25%
NZ	2015	26.4%
NZ	2016	36.03%
SW	2014	19.64%
SW	2015	24.92%
SW	2016	33.26%
US	2009	43.5%
US	2010	43.7%
US	2011	46.43%
US	2012	43.68%
US	2013	44.51%
US	2014	33.15%
US	2015	28.87%
US	2016	33.83%

Kickstarter success over time: Line Graphs

The following graphs show us kickstarter success rates over time, by country. I find it interesting that every country increased its success rates during the years 2015-2016.

```
ggplot(graph_data, aes(x=Year,y=Percent_Successful)) +
  geom_line(size=1.5,color="blue") +
  labs(y="Percent Successful",x="Year") +
  facet_wrap(.~Country) +
  theme(axis.text.x = element_text(size=10,angle=45))
```



Which category's are the most successful?

There are two main variables in this equation for success. The first is how much money does each kickstarter raise? Not just total but averaged for the number of observations. The next is how often does that category succeed? The success of a kickstarter is very important because kickstater is an all or nothing ordeal. If one category has the highest total cash raised but none of the projects succeeded then it still is considered a failure of a category. Therefore my goal in the next table will be to figure out which categorys raised the most cash on average, filtered for the observations that were successful. Then we will see exactly what ratio of those categorys were successful in raising their fund goal.

Which categorys raised the most money on average?

From the following graph, we see that the Technology raised the most cash on average however Games raised the most amount of cash, total. To the individual who is just looking to maximize the total amount of money their kickstarter will generate, Technology is your best bet. However, just because it makes the most money on average, does not mean that it is the safest bet. This leads to my next question...

```
categorys <- data[c(4,10,13)]
categorys$usd.pledged <- as.character(categorys$usd.pledged)
categorys$usd.pledged <- as.integer(categorys$usd.pledged)
categorys <- categorys %>%
  group_by(main_category) %>%
  mutate(Percent_Successful=(round(sum(state %in% "successful")/
                                     sum(state %in% c("successful","canceled",
                                                      "failed",
                                                      "suspended")),digits=4))*100) %>%

  filter(state %in% ("successful")) %>%
  group_by(main_category) %>%
  summarize(Average_Cash_Rasied=mean(usd.pledged,na.rm=TRUE),
              Total_Cash_Raised=sum(usd.pledged,na.rm=TRUE),
     arrange(desc(Average_Cash_Rasied))

categorys$Average_Cash_Rasied<- format(categorys$Average_Cash_Rasied,big.mark=",",digits=2)
categorys$Total_Cash_Raised<- format(categorys$Total_Cash_Raised,big.mark=",")
categorys$Total_Observations<- format(categorys$Total_Observations,big.mark=",")
categorys %>% kable() %>% kable_styling(c("striped", "bordered"))
```

main_category	Average_Cash_Rasied	Total_Cash_Raised	Total_Observations
Technology	80,706	408,451,755	5,062
Design	54,076	430,332,830	7,959
Games	50,844	477,166,251	9,385
Fashion	18,664	80,441,974	4,310
Food	16,012	84,157,230	5,256
Film & Video	13,488	288,542,832	21,404
Comics	11,239	50,564,769	4,499
Journalism	9,772	8,491,977	869
Photography	8,546	24,809,787	2,903
Publishing	8,265	84,760,841	10,255
Music	6,857	148,603,284	21,763
Art	6,134	59,214,907	9,654
Theater	5,714	34,226,004	5,990
Dance	4,880	10,253,130	2,101
Crafts	4,879	8,152,782	1,671

Success Rate of each category: Data Table

Just because Technology raised the most money on average does not necessarily mean its the best investment. If only 19.87% of the technology kickstarters succeeds, then a safe investor might not want to launch a technology kickstarter. Of course, this all depends on your risk preferences. The following table lists the percent of each category that raised their fund goal in the "Percent_Successful" column.

```
category_success <- data[c(4,10)]
category_success$main_category <- as.character(category_success$main_category)
success_summary <- category_success %>%
   group_by(main_category) %>%
   summarize(Percent_Successful=(round(sum(state %in% "successful")/
                                       sum(state %in% c("successful","canceled",
                                                        "failed",
                                                        "suspended")),digits=4))*100) %>%
   arrange(desc(Percent_Successful))
success_summary$Percent_Successful <- (paste0(success_summary$Percent_Successful,"%"))
category_data<-right_join(success_summary,
                          categorys[,c("main_category","Average_Cash_Rasied")],
                          by = "main_category") %>%
   arrange(desc(Percent_Successful))
category_data %>% kable() %>% kable_styling(c("striped", "bordered"))
```

main_category	Percent_Successful	Average_Cash_Rasied
Dance	62.92%	4,880
Theater	60.54%	5,714
Comics	52.16%	11,239
Music	49.44%	6,857
Art	40.73%	6,134
Film & Video	38.01%	13,488
Design	34.14%	54,076
Games	34.08%	50,844
Publishing	30.79%	8,265
Photography	30.31%	8,546
Food	25.13%	16,012
Fashion	23.88%	18,664
Crafts	23.73%	4,879
Journalism	21.68%	9,772
Technology	19.87%	80,706

Summary of Interesting Findings

- **The USA has the highest success rate between 2009-2013**
  - These success rates may be skewed upwards because less people may have know about kickstarter at this time in comparison to 2014 and so on.(Kickstarter launched in 2009) If less people know about kickstarter, we have less kickstarters in total and therefore we are dividing by a smaller observation set during averaging, in effect increasing the success ratio. This is one possible explanation.
- **All 6 countries increased their percent of successful kickstarters between 2015 and 2016**
- **Successful Technology kickstarters raised the most money on average**
  - This number may be so high because we are only accounting for the **19.87%** of technology kickstarters that actually reached their fund goal(selection bias). To say that technology would not be your best bet because it has a low success rate is a fallacy. Investments are determined by average return and risk preference. The best way to determine your best kickstarter investment would be to average the two values.
- **Dance has the highest percentage of successful kickstarters**
  - Dance almost ties for 1st for lowest amount of average cash raised.