Final Report -01

KICKSTARTER

A1812485

Introduction and motivation

Kickstarter is one of the most popular crowdfunding websites with more than 6 billion USD in pledged amount and 200k successful projects. Crowdfunding has gained a lot of popularity in recent years. It is a practice where a large number of people invest small amounts of money in a project or a venture, typically via the internet. In 2015, 34 billion dollars were raised by crowdfunding. Even though the amount people invest is small. It still poses the risk of losing money by investing in the wrong venture like many other modes of investments. I was intrigued by this concept of crowdfunding and plan on investing in several projects soon. This fascination and the wish to invest motivated me to analyze different categories of projects to make an educated guess of which projects can become successful. This can help me minimize the risk of losing my money. The analysis of crowdfunding focuses on different categories of projects to identify potential investment prospects. This brings me to my research question:

To ascertain if certain categories of start-ups receive more crowdfunding and are more likely to succeed than others.

The analysis will be conducted on Kickstarter data from 2009 to 2018 and will have two key benefits. First, it helps the investor predict which category of projects can be profitable to get good returns on their investments. Second, It helps motivate aspiring entrepreneurs by conveying that projects in diverse categories like dance, theater, etc can be successful. This report goes through the methods I used to plot the graph from the data available. Also, What conclusions can be drawn from these visualizations.

Methods and dataset

The dataset is from Kaagle. The dataset was uploaded by Mickaël Mouillé and was collected from the Kickstarter Platform. It has more the 370000 projects with a total of 15 variables. The analysis focuses on the variables main_categories which contains one of the fifteen main categories of the projects; launched which container the launch date and time for the project; state which contains the status of the project telling

whether they were successful or failed and other status; backers which is the number of investors and lastly USD.pledged which contains the total amount invested by the backers in USD(Image1). To perform the analysis R version 4.0.0 "Arbor Day." The dataset is divided into smaller data frames to make it easy to work with, plot data easily, and perform calculations without making any changes to the dataset. Additionally, libraries like ggplot2 and dplyr are used to plot the graphs. These libraries make it easy to work with the graphs which help in the construction of graphs that show the data clearly and are simple in design. Most of the analysis is done by correlating different graphs and finding results. The analysis contains the following graphs:

- **Success rates in a different year** - The graph shows the increasing success rate in different years.
- **Total projects in each category**- This graph Shows the popular categories amongst fund seekers.
- **Average funds raised per backers in each category** - The graph shows in which category backers invest more.
- **Success to fail** - the graph shows the relation between the percentage of successful and failed projects in each category.
- **The portion of a category in successful projects each year**- The graph shows which categories are growing more successful each year.

As mentioned above a separate data frame is created for the graphs. (Image 2) Then ggplot is used to plot graphs for each data frame. I have three types of graphs a line graph, a scatter plot and a couple of bar-plot. To create a line graph geom_line function was used. Similarly, for scattering and bar, geom_point and geom_bar were used respectively.

Experimental Setup and Results

To make an educated guess of which projects can become successful. Let's look at the total projects, successful and failed projects for all the given year. This graph will help us to compare the proportion of successful and failed projects and ascertain if the difference between the successful and failed projects is high or low. (Image3) In the graph, we can see that the failed projects follow the trends of the total project pretty closely. However, in the year 2016-2017, the slope is steeper for the failed projects when compared to the total projects. Also, there is comparatively less change in the successful projects. This indicated that the gap between the projects is closing in. Now, let us see Which categories the fundraiser is interested in. This graph shows the diverse categories that are available on the platform. The graph shows different categories

and the count with a dashed line showing the mean. (Image 4) It can be assumed that the projects above the mean line are more popular amongst the fundraisers. The category film and video have the largest number of projects and dance are the lowest. The categories technology, Publishing, Music, Games, Design, Art have projects large than the mean projects in all categories. After seeing the fundraiser's interest let's look at the interest of the backers. To show this the graph contains the average funds a backer invests in a particular category. Similar to the previous graph, there is a mean line and it is assumed that the projects above it are popular amongst the backers. (Image 5) The technology receives the most investment per backer. Apart from technology dance, design, film and video, food, photography, and theater are popular. Further, comics have the least average funding per backer. The next graph shows us what we are actually trying to predict. The graph shows the relation between the percentage of the failed projects and the percentage of the successful projects. (Image 6) The project lying on the top left corner has a higher percentage of successful projects than failed. On the other hand, the projects on the left have a higher percentage of failed projects than successful. Dance has the highest percentage of successful projects. Followed by theater, music, and comics. In contrast to this technology, crafts, food, fashion and journalism are the more likely to fail. The graph shows us the cumulative percentage for all the years. This is not enough to make an educated guess of which categories can be more successful. Therefore the last visualization looks at the portion each category has in the total successful projects for a given year. This can help us recognize which categories have an increased rate of success and which are declining. (image 7) Categories like design, technology, Games, Comics, Craft have an increasing number of successful projects. Further, Music, theater, film, and journalism are declining. Dance has slight fluctuation but has maintained the proportion of the successful.

Conclusions and Discussion

The graph showing the count of successful and failed projects indicated that the gap between the successful and failed projects is declining. Which makes 2018 an ideal year to invest. We can assume a larger number of successful projects in all categories. One thing that is worth noting is that Dance, theater, comics, and music have a higher percentage of successful projects. However, when we see the portion of these categories year-wise. We see a sudden drop in the music and theater. In contrast to this comic have an increased rate of success. Further, dance has maintained the rate of successful projects. This shows that dance is a potential investment project. Also, it raises the confidence of the fundraisers and promotes more projects to show up in the category. Further Visualizations indicates that design, technology, and Games have an increased rate of

success which explains the increased interest in the backers for Games, Technology, and Design. These categories have potential and can also be potential investments prospects. At last, the categories more likely to obtain the needed investments and succeed are dance, games, design, and arts.

References

- Holtz, Y., 2021. *The R Graph Gallery – Help and inspiration for R charts*. [online] The R Graph Gallery. Available at: <https://www.r-graph-gallery.com> [Accessed 9 September 2021].
- Investopedia. 2021. *Crowdfunding*. [online] Available at: <https://www.investopedia.com/terms/c/crowdfunding.asp> [Accessed 9 September 2021].
- Kickstarter. 2021. *Kickstarter stats*. [online] Available at: <https://www.kickstarter.com/help/stats> [Accessed 9 September 2021].
- C R, N., 2021. *EDA using Tableau Visualizations*. [online] Kaggle.com. Available at: <https://www.kaggle.com/nikhilncr/eda-using-tableau-visualizations> [Accessed 9 September 2021].
- Mouillé, M., 2021. *Kickstarter Projects*. [online] Kaggle.com. Available at: <https://www.kaggle.com/kemical/kickstarter-projects> [Accessed 9 September 2021].

Appendix

```
      ID                 name              category         main_category       currency
 Min.   :5.971e+03   Length:702411      Length:702411      Length:702411      Length:702411
 1st Qu.:5.380e+08   Class :character   Class :character   Class :character   Class :character
 Median :1.076e+09   Mode  :character   Mode  :character   Mode  :character   Mode  :character
 Mean   :1.075e+09
 3rd Qu.:1.611e+09
 Max.   :2.147e+09

   deadline              goal              launched          pledged            state
 Length:702411      Min.   :       0   Length:702411      Min.   :       0   Length:702411
 Class :character   1st Qu.:    2000   Class :character   1st Qu.:      30   Class :character
 Mode  :character   Median :    5000   Mode  :character   Median :     615   Mode  :character
                    Mean   :   48193                      Mean   :    9229
                    3rd Qu.:   15080                      3rd Qu.:    4021
                    Max.   :100000000                     Max.   :20338986
                    NA's   :632                           NA's   :624
    backers           country           usd.pledged
 Min.   :     0.0   Length:702411      Min.   :       0
 1st Qu.:     2.0   Class :character   1st Qu.:      20
 Median :    12.0   Mode  :character   Median :     456
 Mean   :   103.9                      Mean   :    7410
 3rd Qu.:    55.0                      3rd Qu.:    3262
 Max.   :219382.0                      Max.   :20338986
 NA's   :623                           NA's   :8210
```

Image 1: This image shows some basic statistics performed on different variable from the dataset.

GRAPH -1 DATA FRAME

	Var1	Var2	Freq
1970	failed		0
2009	failed		600
2010	failed		4984
2011	failed		11878
2012	failed		20588
2013	failed		21686

GRAPH-2

	Var1	Freq
Art		28153
Comics		10819
Crafts		8809
Dance		3768
Design		30070
Fashion		22816

GRAPH-3

	Var1	average	twodec
Art		60.04150	60.04
Comics		39.65250	39.65
Crafts		43.21962	43.22
Dance		69.20839	69.21
Design		70.58165	70.58
Fashion		69.63603	69.64

GRAPH -4 DATA FRAME

	Var1	success	fail
Art		40.88374	50.19359
Comics		53.99760	37.30474
Crafts		24.00954	64.74061
Dance		62.04883	32.77601
Design		35.08480	49.26505
Fashion		24.51350	62.15813

GRAPH-5

	Var1	categories	Var3	Freq	sum	percentage
2009		Art	successful	77	579	13.298791
2010		Art	successful	382	4593	8.317004
2011		Art	successful	1189	12171	9.769123
2012		Art	successful	1694	17892	9.467919
2013		Art	successful	1684	19415	8.673706
2014		Art	successful	1783	21107	8.447435

Image 2: Data Frames created for different graphs from the main dataset.
•Graph 1- Success rates in a different year.
•Graph 2 - Total projects in each category.
•Graph 3 - Average funds raised per backers in each category.
•Graph 4 - Success to fail
•Graph5 - The portion of a category in successful projects each year

```
# getting the years from the dates
ks2018$dates.launch <- as.POSIXct(ks2018$launched,format = "%Y-%m-%d %H:%M:%S")
ks2018$dates.launch <- format(ks2018$dates.launch, format = "%Y")
# constructing a new data frame for the state count in different years
stateyear <-  data.frame(table(ks2018$dates.launch,ks2018$state))
# Removing the state that are not needed
stateyear <-  subset(stateyear, stateyear$Var2 != "live")
stateyear <-  subset(stateyear, stateyear$Var2 != "undefined")
stateyear <-  subset(stateyear, stateyear$Var2 != "suspended")
stateyear <-  subset(stateyear, stateyear$Var2 != "canceled")
# changing the name of the columns
stateyear <- setNames(stateyear,c("Var1","Legend","Freq"))
totalprojectsyear <- data.frame(table(ks2018$dates.launch))
totalprojectsyear <- data.frame(totalprojectsyear$Var1,c(rep("total",11)),totalprojectsyear$Freq)
totalprojectsyear <- setNames(totalprojectsyear,c("Var1","Legend","Freq"))
stateyear<-rbind(stateyear,totalprojectsyear)
# removing the year with no data
stateyear <-  subset(stateyear, stateyear$Var1 != 1970)
stateyear <-  subset(stateyear, stateyear$Var1 != 2018)
# Ploting the line graph for success and fail each year
ggplot(stateyear, aes(x=Var1, y=Freq , group=Legend, color=Legend))+geom_line()+
  xlab("years")+ylab("Total projects")
```

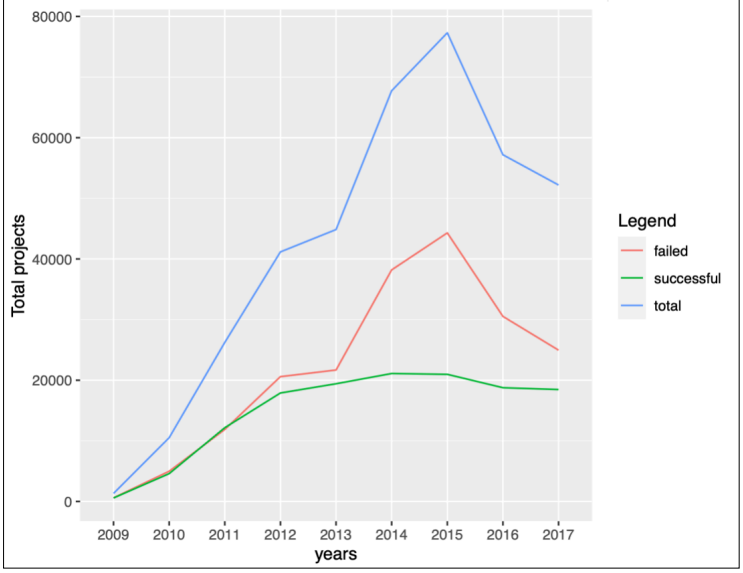

Image 3: Graph 1- Success rates in a different year. This image shows the code and the output generated. The graph shows the total number of successful and failed projects over the year

```
# forming a data frame of the main category
main_category <- data.frame(table(ks2018$main_category))
# plotting the data
ggplot(main_category,aes(x=Var1,y=Freq)) +
geom_bar(stat="identity",fill="cadetblue2") +
coord_flip()+ylab("total projects") + xlab("category")+
geom_text(aes(x = Var1,    y = Freq+30, label = Freq))+
geom_hline(aes(yintercept=mean(Freq)),color="gray70",
size=0.5, linetype="dashed")+geom_text(aes(x="Theater",
label=paste0("Mean ",floor(mean(Freq))), y=29000),size = 3)
```

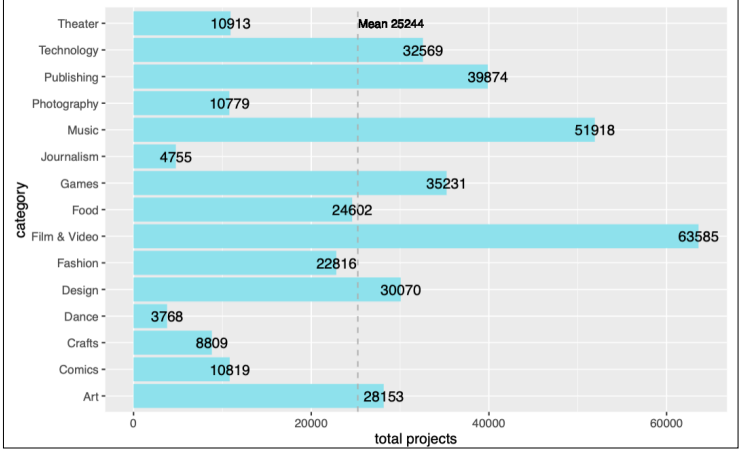

Image 4: Graph 2 - Total projects in each category. This image shows the code and the output generated. The graph shows the total number of projects in each category and a dashed mean line.

```
# getting the sum of backer in different categories
backers <- data.frame(aggregate(ks2018$backers, by=list(categories=ks2018$main_category), FUN=sum))
backers <- setNames(backers,c("categories","backers"))
# getting total investments
investment <- data.frame(aggregate(ks2018$usd.pledged, by=list(categories=ks2018$main_category), FUN=sum))
avgfb <- data.frame(backers$categories)
# calculating average
avgfb$average <- (investment$x/backers$backers)
avgfb <- setNames(avgfb,c("Var1","average"))
#converting to two decimals
avgfb$twodec <- format(round(avgfb$average, 2), nsmall = 2)
# plotting the graph for avg investment per backer in each category
ggplot(avgfb,aes(x=Var1,y=average)) +
  geom_bar(stat="identity",fill="coral") +
  ylab("average funds invested by a backer") + xlab("categories")+
  geom_text(aes(x = Var1,    y = average+1.5, label = twodec))+
  geom_hline(aes(yintercept=mean(average)),color="gray70", size=0.5, linetype="dashed")+
  geom_text(aes(x="Art", label=paste0("Mean: ",floor(mean(average))), y=66),size = 3)
```

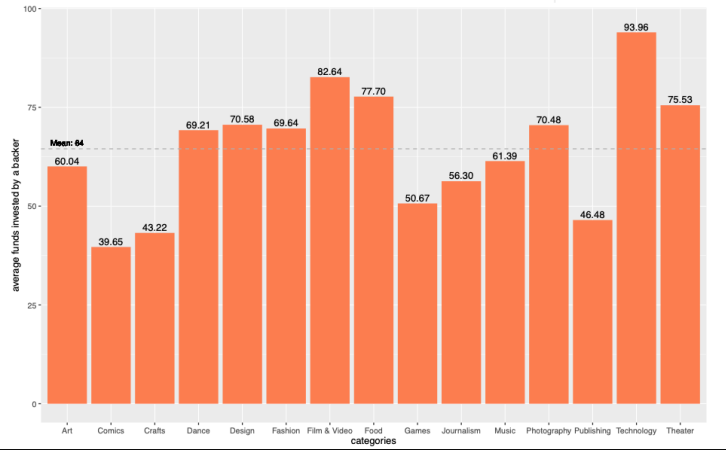

Image 5: Graph 3 - Average funds raised per backers in each category.his image shows the code and the output generated. The graph shows the average funds raised by backers in each category.

```
#dividing categories based on succcesss and failure
status_categories <-  data.frame(table(ks2018$main_category,ks2018$state))
#removing all the other status from the data frame
status_categories <-  subset(status_categories, status_categories$Var2 != "live")
status_categories <-  subset(status_categories, status_categories$Var2 != "undefined")
status_categories <-  subset(status_categories, status_categories$Var2 != "suspended")
status_categories <-  subset(status_categories, status_categories$Var2 != "canceled")
# creating two different categories for success and failure
categories_fail <- subset(status_categories, status_categories$Var2 != "successful")
categories_pass <- subset(status_categories, status_categories$Var2 != "failed")
#performing calculationsan renaming the columns
categories_pass <- data.frame(categories_pass,main_category$Freq)
categories_pass$success_percentage <- ((categories_pass$Freq / categories_pass$main_category.Freq)*100)
categories_fail <- data.frame(categories_fail,main_category$Freq)
categories_fail$fail_percentage <- ((categories_fail$Freq / categories_fail$main_category.Freq)*100)
names(categories_fail)[5] <- 'percentage'
names(categories_pass)[5] <- 'percentage'
# creating a data with both percentages and the categories
pass_fail <- data.frame(categories_fail$Var1,categories_fail$percentage,categories_pass$percentage)
pass_fail <- setNames(test,c("Var1","fail","success"))
#plotting the graph for fail success
ggplot(pass_fail, aes(x=fail, y=success)) +
  geom_point() + # Show dots
  geom_label(
    label=pass_fail$Var1,
    nudge_x = 0.25, nudge_y = 0.25,
    check_overlap = T
  )+xlab("percentage of projects failed")+ylab("percentage of projects successful")
```

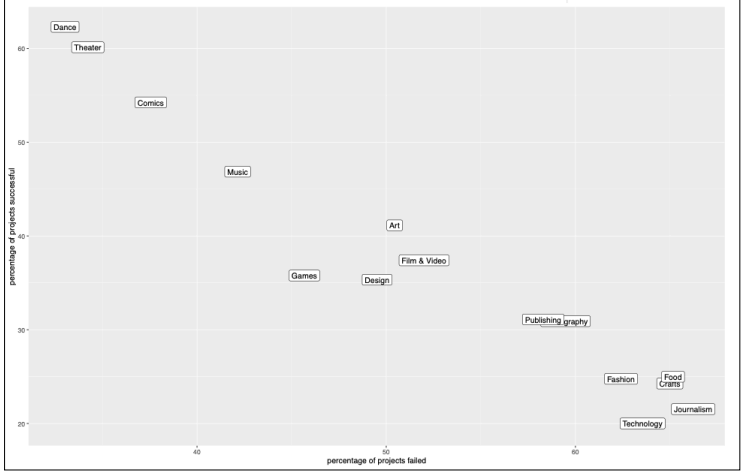

Image 6: Graph 4 - Success to fail. This image shows the code and the output generated. The graph shows the percentage of successful and failed projects. The project lying on the top left corner has a higher percentage of successful projects than failed.

```
successyear <- subset(stateyear, stateyear$Var2 != "failed")
successyear <-  subset(successyear, successyear$Var1 != 1970)
successyear <-  subset(successyear, successyear$Var1 != 2018)
# creating a new dataset
catyearsuccess<- data.frame(table(ks2018$dates.launch,ks2018$main_category,ks2018$state))
catyearsuccess <-  subset(catyearsuccess, catyearsuccess$Var3 != "live")
catyearsuccess <-  subset(catyearsuccess, catyearsuccess$Var3 != "undefined")
catyearsuccess <-  subset(catyearsuccess, catyearsuccess$Var3 != "suspended")
catyearsuccess <-  subset(catyearsuccess, catyearsuccess$Var3 != "canceled")
catyearsuccess <-  subset(catyearsuccess, catyearsuccess$Var3 != "failed")
catyearsuccess <-  subset(catyearsuccess, catyearsuccess$Var1 != 1970)
catyearsuccess <-  subset(catyearsuccess, catyearsuccess$Var1 != 2018)
catyearsuccess$sum <- c(rep(successyear$Freq,3))
catyearsuccess$percentage <- (catyearsuccess$Freq/catyearsuccess$sum)*100
catyearsuccess<-setNames(catyearsuccess,c("Var1","categories","Var3","Freq","sum","percentage"))
#The portion of a category in successful projects each year
ggplot(catyearsuccess, aes(fill=categories, y=percentage, x=Var1)) +
  geom_bar(position="fill", stat="identity",color = "black")+xlab("Years")+ylab("% total number c
  scale_fill_manual(values = c("aliceblue","brown1","burlywood1","cadetblue1","chocolate","coral'
```

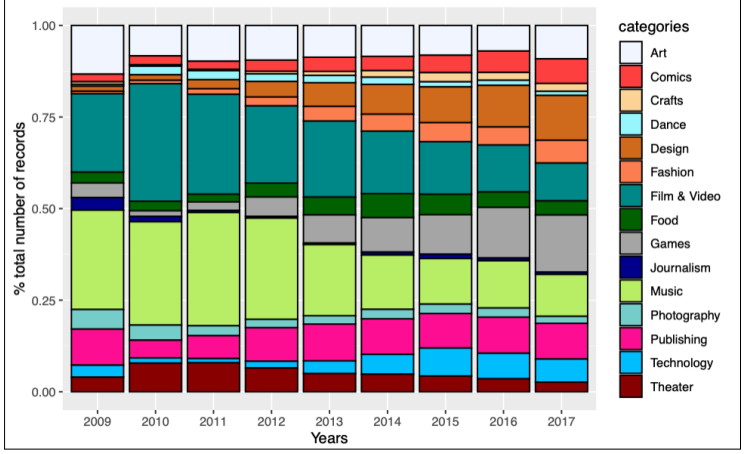

Image 7: Graph5 - The portion of a category in successful projects each year. This image shows the code and the output generated.The graph shows the portion of each category in the total successful projects for a given year. The graph helps in understanding the change in the portion of successful projects and helps us predict future trends.