Data analytics project

Team members:

1.Shramana Thakur(3169639)

2.Megha Jayakumar(3155859)

3.Rishi Tripathi(3169149)

4.Subbulakshmi Sundaram(3157792)

5.Srinivas Nandagudi Sridharamurthy(3166351)

UNIVERSITY OF BONN SUMMER 2018 BATCH

INTRODUCTION:

The kickstarter dataset contains the information of the startup projects which has been funded globally by crowdsourcing.The source of the dataset which is being used is https://www.kaggle.com/kemical/kickstarter-projects.

```
#importing the dataset
require(gridExtra)
```

```
## Loading required package: gridExtra
```

```
require(pacman)
```

```
## Loading required package: pacman
```

```
require(tidyverse)
```

```
## Loading required package: tidyverse
```

```
## -- Attaching packages ----------------------------------------------------------------------------- ti
dyverse 1.2.1 --
```

```
## v ggplot2 3.0.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.6
## v tidyr   0.8.1     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0
```

```
## -- Conflicts ----------------------------------------------------------------------------------- tidyvers
e_conflicts() --
## x dplyr::combine() masks gridExtra::combine()
## x dplyr::filter()  masks stats::filter()
## x dplyr::lag()     masks stats::lag()
```

```
require(readxl)
```

```
## Loading required package: readxl
```

```
ex_rate_2018 = read_excel("C:/Users/SUBBULAKSHMI/Downloads/2018 Exchange Rate.xlsx")
data2018<-read.csv("C:/Users/SUBBULAKSHMI/Downloads/ks_2018.csv")
```

```
#Description of the dataset
str(data2018)
```

```
## 'data.frame':    378661 obs. of  15 variables:
##  $ ID              : int  1000002330 1000003930 1000004038 1000007540 1000011046 1000014025 1000023410 1
000030581 1000034518 100004195 ...
##  $ name            : Factor w/ 375723 levels "","\177Not Twins - New EP! \"The View from Down Here\"",..
: 332499 135648 364968 344763 77308 206088 293420 69319 284097 290676 ...
##  $ category        : Factor w/ 159 levels "3D Printing",..: 109 94 94 91 56 124 59 42 114 40 ...
##  $ main_category   : Factor w/ 15 levels "Art","Comics",..: 13 7 7 11 7 8 8 8 5 7 ...
##  $ currency        : Factor w/ 14 levels "AUD","CAD","CHF",..: 6 14 14 14 14 14 14 14 14 14 ...
##  $ deadline        : Factor w/ 3164 levels "1/1/2010","1/1/2011",..: 555 566 1272 1670 2818 1618 950 1432
2029 2641 ...
##  $ goal            : num  1000 30000 45000 5000 19500 50000 1000 25000 125000 65000 ...
##  $ launched        : Factor w/ 347035 levels "1/1/1970 1:00",..: 291593 330364 2296 144170 282910 126945
87019 109522 182988 259595 ...
##  $ pledged         : num  0 2421 220 1 1283 ...
##  $ state           : Factor w/ 6 levels "canceled","failed",..: 2 2 2 2 1 4 4 2 1 1 ...
##  $ backers         : int  0 15 3 1 14 224 16 40 58 43 ...
##  $ country         : Factor w/ 23 levels "AT","AU","BE",..: 10 23 23 23 23 23 23 23 23 23 ...
##  $ usd.pledged     : num  0 100 220 1 1283 ...
##  $ usd_pledged_real: num  0 2421 220 1 1283 ...
##  $ usd_goal_real   : num  1534 30000 45000 5000 19500 ...
```

DESCRIPTION OF DATASET:

1.internal kickstarter id (integer variable)

2.name: name of project - A project is a finite work with a clear goal that you'd like to bring to life. Think albums, books, or films.(Factor variable)

3.category of the product (factor variable)

4.main_category : category of campaign (factor variable)

5.currency : currency used to support(factor variable)

6.deadline : deadline for crowdfunding (factor variable)

7.goal: fundraising goal - The funding goal is the amount of money that a creator needs to complete their project(numeric variable)

8.launched :date launched(factor variable)

9.pledged : amount pledged by "crowd"(numeric variable)

10.state : state pledged from(factor variable)

11.backers : number of backers(integer variable)

12.country : country pledged from(factor variable)

13.usd pledged : amount of money pledged(numeric variable)

PREPROCESSING

```
dim(data2018)
```

```
## [1] 378661     15
```

```
data2018<-na.omit(data2018)
dim(data2018)
```

```
## [1] 374864     15
```

The dimension of dataset before omiting the empty and not applicable values is 378661 15 The dimension of dataset after omiting the empty and not applicable values is 374864 15

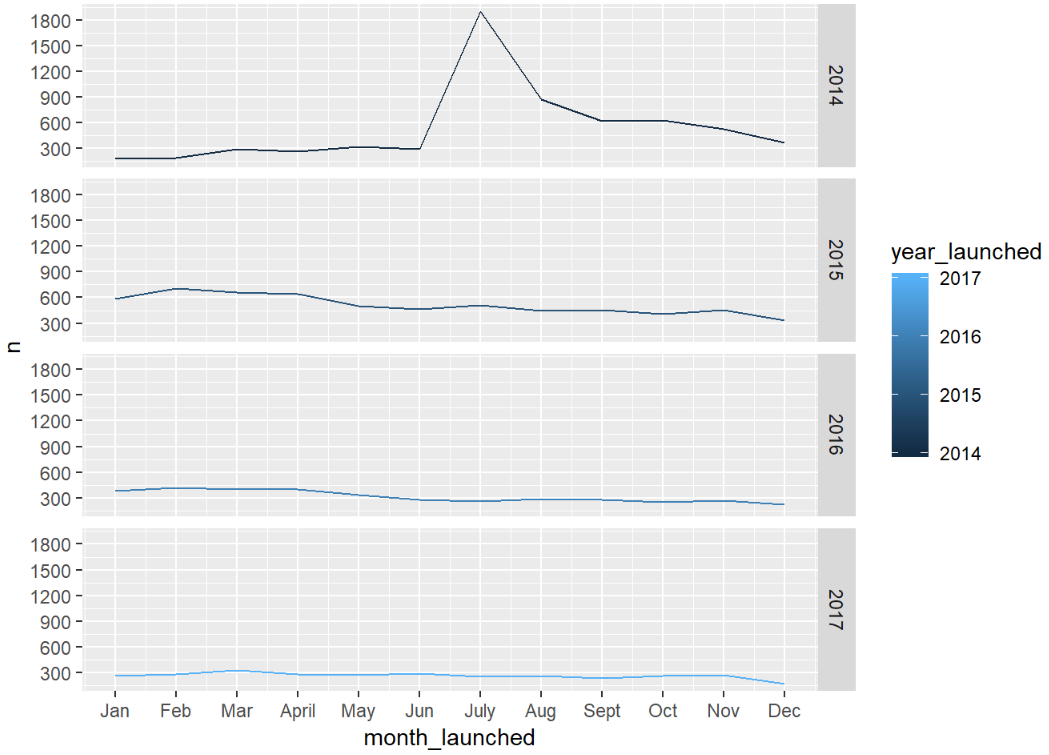
```
#Make new columns to convert all currencies into USD
convert_to_USD_2018 <- function(money,currency) {
  if(currency == "USD") {
    return(money)
  }
  else {
    rate = ex_rate_2018 %>% filter(Code==currency) %>% select(USD_per_Unit)
    return(as.numeric(money*rate))
  }

}

A = mapply(convert_to_USD_2018,data2018$goal,data2018$currency)
B = mapply(convert_to_USD_2018,data2018$pledged,data2018$currency)

data2018$usd_goal_Real = A
data2018$usd_pledged_Real = B
```

The above code adds two columns usd_goal_real and usd_pledged_real which specify the goal amount and pledged amount which is converted to US dollars from different currencies.

EXPLORATORY DATA ANALYSIS:

Exploratory data analysis is an approach for analysing different data sets to make a summary of their main characteristics.

```
#Create a lineplot for main category "Food" and showing the product launched in each month of the year 2014
to 2017.
data2018$launched <- as.Date(data2018$launched ,"%m/%d/%Y")

data2018 %>%
  filter(main_category == 'Food') %>%
  mutate(year_launched = as.numeric(format(launched, "%Y")), month_launched = as.numeric(format(launched, "%
m"))) %>%
  filter(year_launched > 2013 & year_launched < 2018) %>%
   group_by(year_launched,month_launched) %>%
   summarize(n=n())  %>%
   ggplot(aes(x=month_launched, y = n, color = year_launched)) + geom_line() + scale_x_continuous(breaks=1:1
2, labels =paste0(c("Jan","Feb","Mar","April","May", "Jun","July","Aug","Sept","Oct","Nov","Dec"))) + scale
_y_continuous(breaks = seq(0,1800,300)) + facet_grid(year_launched ~ .)
```



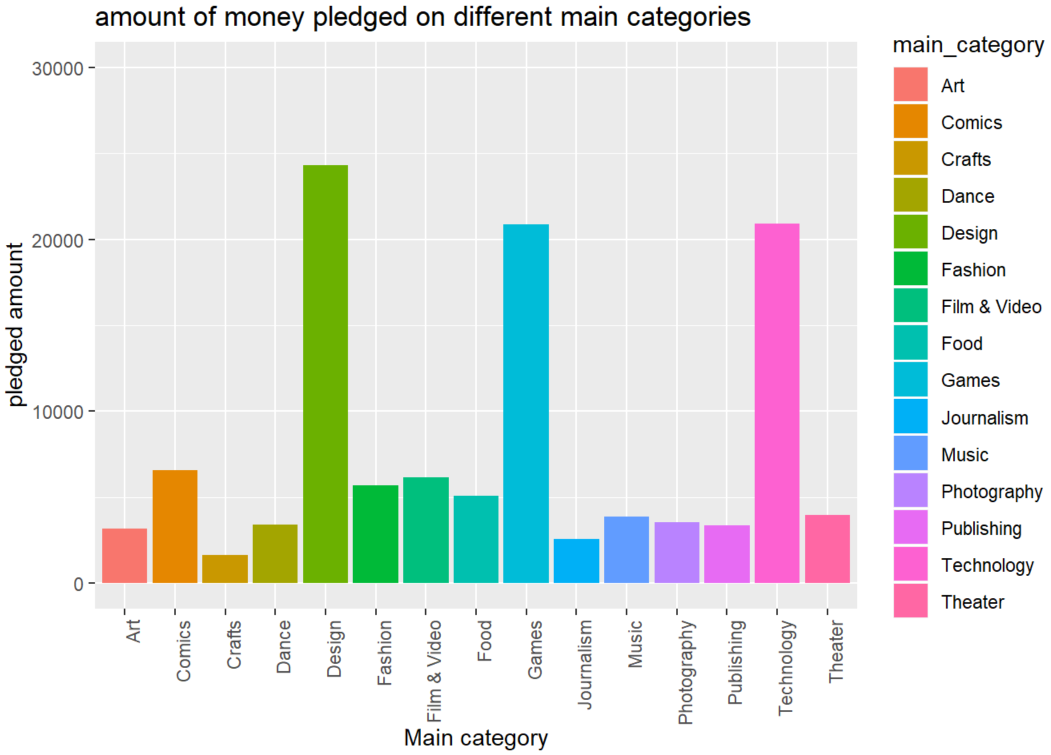
From the analysis of the line plot above we can see that in July 2014, the products launched were maximum. On the other hand in the year 2015 we can observe that the products released are more consistent through out the year. In the years 2016 - 2017 the release of the products were almost same, ranging from 300 to 600

```
#How many projects exceeded their funding goal by 50% or more?

data2018 %>% filter(usd.pledged >= 1.5*usd_goal_real) %>% count
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1 30876
```

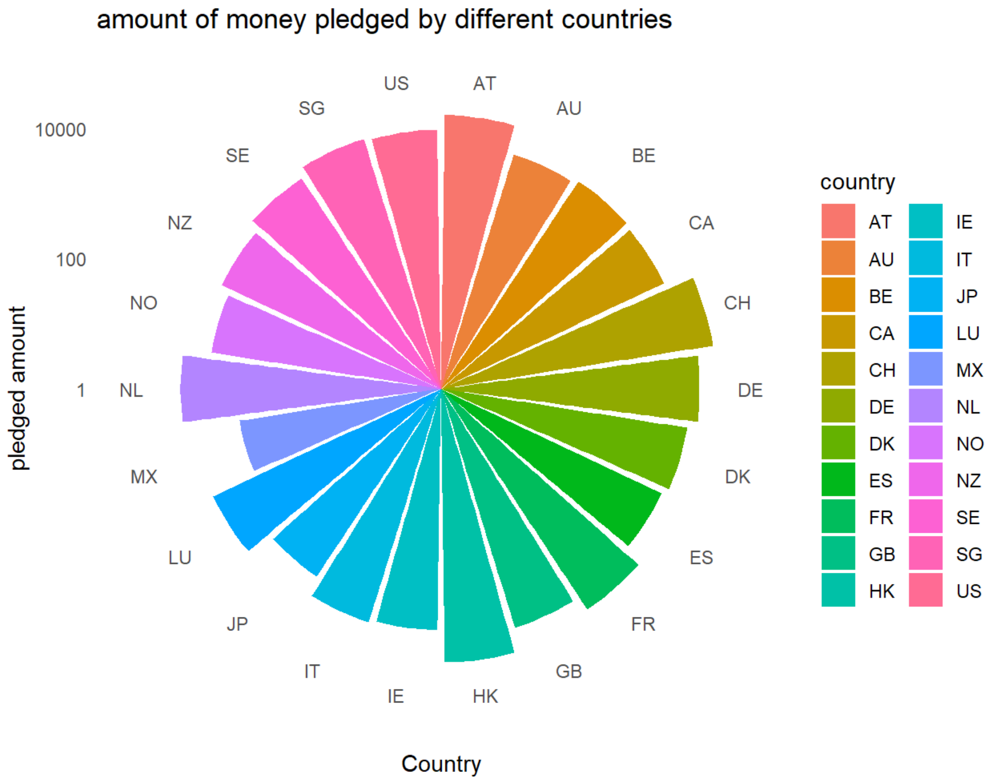
It can be seen that 30876 are the projects are exceeding the funding goal of 50%.

```
## Create a plot of the amount pledged on different main categories
histogram_main_categories<-data2018 %>% group_by(main_category)%>%summarise(Mean=mean(usd_pledged_Real))%>%a
rrange(-Mean)
histogram_country<-data2018%>%group_by(country)%>%summarise(Mean=mean(usd_pledged_real))%>%arrange(-Mean)
plot1<-histogram_main_categories %>%
  ggplot(aes(x=main_category,y=Mean,fill=main_category)) + geom_bar(stat="identity")+ylim(0,30000)+theme(axi
s.text.x = element_text(angle = 90, hjust = 1))+xlab("Main category")+ylab("pledged amount")+ggtitle("amount
of money pledged on different main categories")
plot1
```



The above is a bar plot which shows the amount of money in usd dollars pledged on different main categories and it can be seen that Design and Crafts are the main categories for which maximum and minimum amount has been pledged respectively .

```
## Create a plot of the amount pledged by different countries

histogram_main_categories<-data2018 %>% group_by(main_category)%>%summarise(Mean=mean(usd_pledged_Real))%>%a
rrange(-Mean)
histogram_country<-data2018%>%group_by(country)%>%summarise(Mean=mean(usd_pledged_real))%>%arrange(-Mean)




plot2<-histogram_country %>%
  ggplot(aes(x=country,y=Mean,fill=country)) +geom_bar(stat="identity")+scale_y_log10()+coord_polar()+theme_
minimal()+ggtitle("amount of money pledged by different countries") +xlab("Country")+ylab("pledged amount")+
  theme(
    #axis.text = element_blank(),
    #axis.title = element_blank(),
    panel.grid = element_blank(),
    plot.margin = unit(rep(0,4), "cm")      # This remove unnecessary margin around plot
  )



plot2
```



The above is a cicular barplot which shows the amount of money in usd dollars pledged by different countries and it can be seen that Hongkong,Netherlands and Australia are the countries which have pledged maximum amount of money and Mexico pledged the minimum amount.

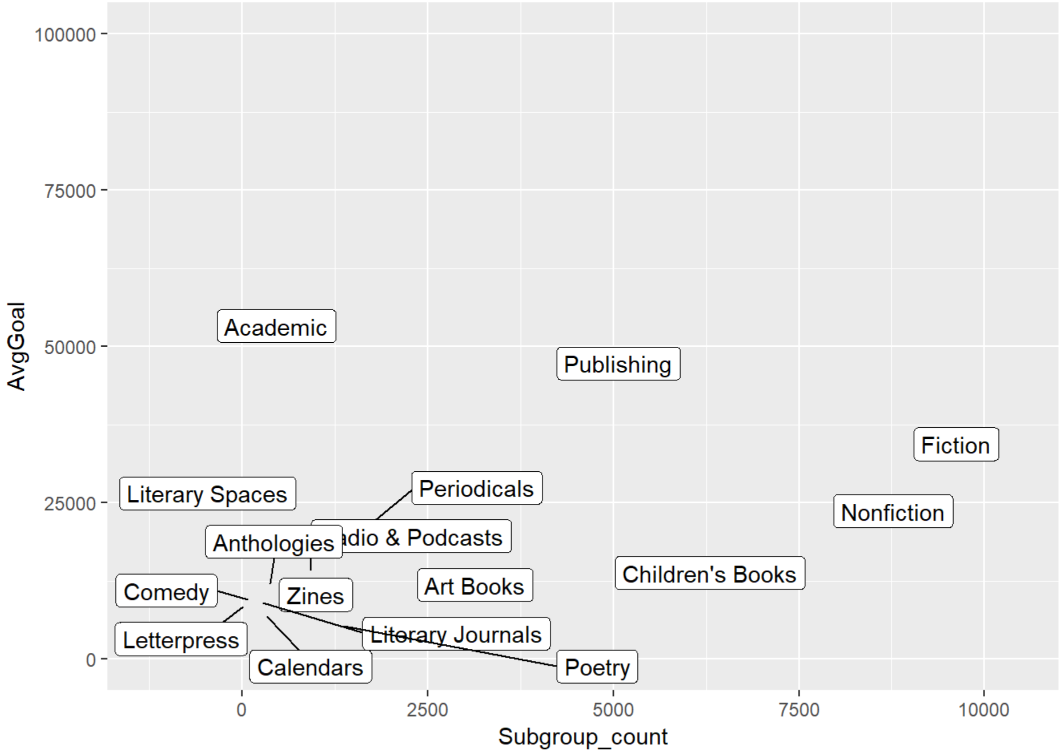
```
#Display the names and category of those projects that has the string "Songs" in it.
data2018 %>% select(name,category) %>% filter((str_detect(data2018$name, "Songs"))==TRUE)%>%head(10)
```

```
##                                                       name   category
## 1                          The Songs of Adelaide & Abullah     Poetry
## 2           Songs of the Damned Jake Brock's Latest Album.       Rock
## 3                                      Songs Without Words Indie Rock
## 4       Songs of Petroleum - autobiography of Jan Lundberg Nonfiction
## 5      Songs of Joy & Pain - Sutter Zachman's Debut Record      Music
## 6                  A Calling . . . Songs Beyond the Traffic      Music
## 7      Keeping Hearts Project: 6 New Songs By Amanda Grace Indie Rock
## 8  Clem Snide -  New Full Length Record - "Songs For Mary" Indie Rock
## 9   Hold On Another Day: "Songs For Project Believe In Me"      Music
## 10        Ambyr D'Amato: Recording New Songs with Friends!      Music
```

The select statement is used to determine the data columns that has to be displayed in the output and filter uses the str_detect function to detect the names that has the string “Songs”

```
#Create a plot of main category "Publishing" showing top 10 categories with respect to Average Goal and Aver
age Pledge.

p_load(ggrepel)
data2018 %>% na.omit() %>% filter(main_category=="Publishing") %>% group_by(category) %>% summarise(Subgroup
_count = n(),AvgGoal = mean(goal)) %>%arrange(-AvgGoal) %>%  head(100) %>%
ggplot(aes(x=Subgroup_count, y=AvgGoal, label=category)) +
geom_label_repel() + ylim(0,100000) + scale_x_continuous(expand = c(0.2,0.2))
```

```
## Warning: Removed 2 rows containing missing values (geom_label_repel).
```



The data with the main_category = “Publishing” is filtered and grouped as per their sub category which has the column name “category”. The number of sub groups and average goal is computed using the mean() function. The data is arranged in descending order of Average goal which is achieved using the arrange() function. For the plot we have used the top 100 data of the ordered data.

The plot is between the Average pleadge and Average goal.

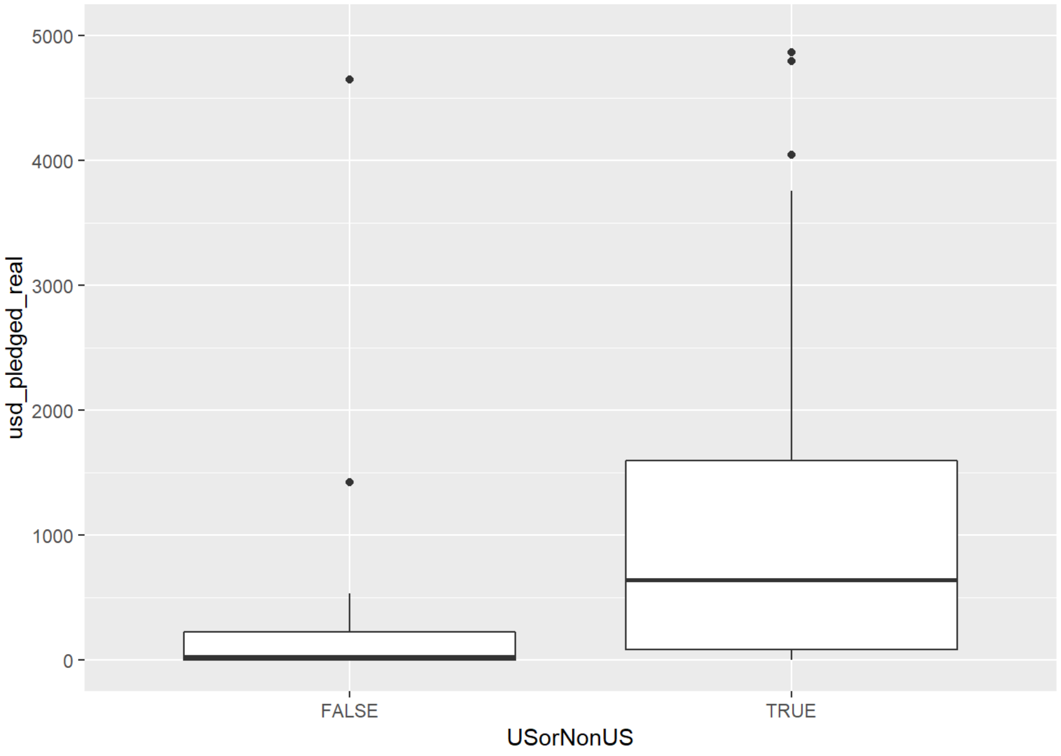
geom_label_repel() is used for Text labels to repel away from each other, away from data points, and away from edges of the plotting area. This helps to give a clear picture of the plotted graph avoiding any overlapping of the points. ylim() methodis used to set the y axis scale.

We can see from the plot that the number of sub categories has no effect in setting a goal value for a particular group.

The warning we obtained is due to how ggplot2 deals with data that are outside the axis ranges of the plot.

```
#Create a box plot against the number of projects pledged by US and non - US states against the Pledged real value

data2018 %>% mutate(USorNonUS = str_detect(country,"US")) %>% head(100) %>%
ggplot(aes(x=USorNonUS, y=usd_pledged_real)) +
geom_boxplot() + ylim(0,5000)
```

```
## Warning: Removed 29 rows containing non-finite values (stat_boxplot).
```

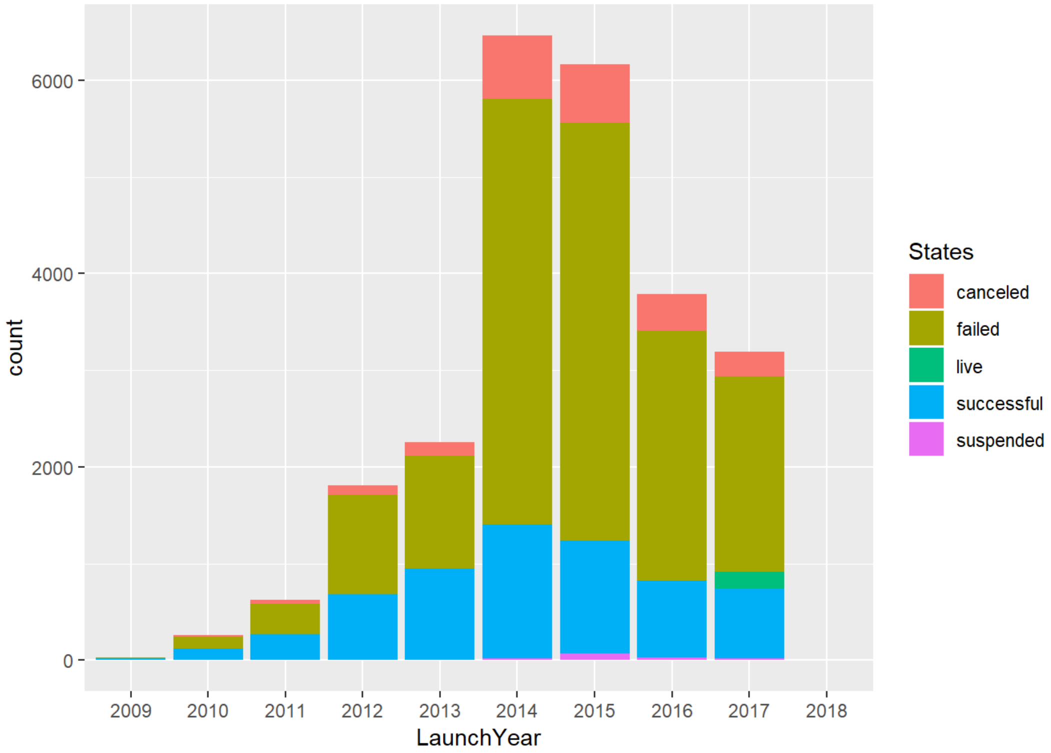


Here we add a new column “USorNonUS” to our data using the mutate() function to determine which are the projects from US. A box plot is plotted for the top 100 data fo the US and Non-Us countries against the usd_pleadged_real values.

The block labelled as False is depicting the Non-US countries and the one labelled as True is for the US. From the plot we can infer that the amount pledged by the US is very high compared to the non-US countries for the 1st 100 data of the data set.

The warning is obtained as some of the data are out of range of the scale specified in the plot. By altering this scale the plot becomes unclear and non readable.

```
#Show the different States of the main category "Food" along the years.
data2018  %>% mutate( LaunchYear = format(as.Date(data2018$launched, format="%d/%m/%Y"),"%Y")) %>% filter(ma
in_category == "Food")  %>% na.omit() %>% arrange(LaunchYear) %>% ggplot(aes(x=LaunchYear, fill=factor(state
))) +
geom_bar() + labs(fill="States")
```



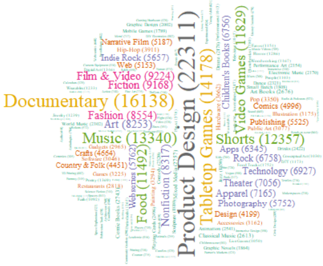
For the above plot we have included a new column LaunchYear to the data which will have the year of the launched year. The bar plot is plotted against the launchYear and count of the projects in the main_category “Food”. The different states are depicted using different colors in the plot which is done using the fill property.

From the plot obtained we can infer that the highest number of projects are launched in the year 2014 and 2015 with the failure rate higher than the success rate.

```
# Make a wordcloud plot of number of projects in each category
library(wordcloud)
```

```
## Loading required package: RColorBrewer
```

```
wordcloud_data = data2018 %>% group_by(category) %>% summarise(n=n())
wordcloud(words = sprintf("%s (%s)", wordcloud_data$category, wordcloud_data$n), freq = wordcloud_data$n, ma
x.words = Inf, scale=c(1,.1), random.order=FALSE, rot.per=0.30, use.r.layout=FALSE,colors=brewer.pal(8, "Dar
k2"),family = "serif", font = 1)
```



The above code snippet creates a word cloud of the different categories .It generates a png image of the world cloud thus formed

PREDICTIVE ANALYSIS

We would like to study how various factors affect the funding of a Kickstarter project. The column “usd_pledged_real” depicts the Pledged money. However, simply predicting the value of the Pledged money will not be of practical importance. Instead, we study the effect of a new mutated variable “Difference” that depicts the difference between the goal money and actual pledged money. Thus, we study the effect of variables “Country (factor)”, “Main category (factor)”, “No. of backers (integer)”, “Goal money (numeric)”, on the success/failure of goal money collection.

```
data2018 = data2018 %>% mutate(Difference = (as.numeric(usd_goal_Real) - as.numeric(usd.pledged)))
```

The above code is to evaluate the difference between the real goal money in US dollars and the pledged money in US dollars.

####1.Difference ~ Backers

```
##Checking for outliers in "backers" and "Difference" column

data2018 %>% group_by(cut(backers, breaks = 10)) %>% summarise(n=n())
```

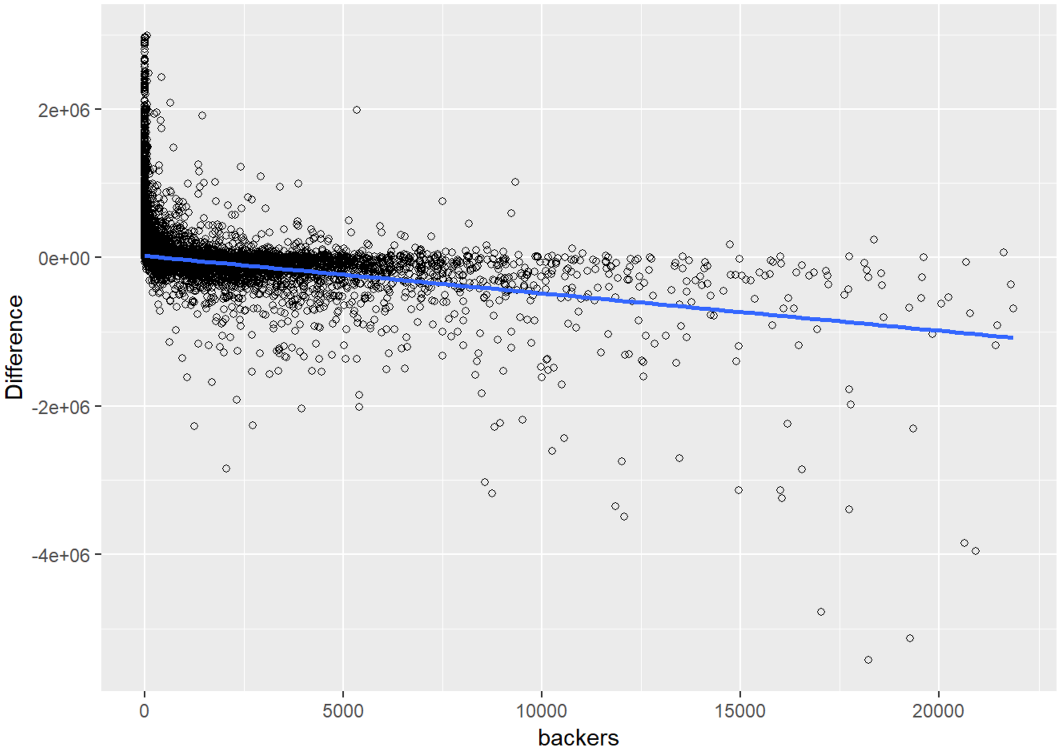
```
## # A tibble: 7 x 2
##   `cut(backers, breaks = 10)`      n
##   <fct>                        <int>
## 1 (-219,2.19e+04]             374796
## 2 (2.19e+04,4.39e+04]             44
## 3 (4.39e+04,6.58e+04]              9
## 4 (6.58e+04,8.78e+04]             11
## 5 (8.78e+04,1.1e+05]               2
## 6 (1.54e+05,1.76e+05]              1
## 7 (1.97e+05,2.2e+05]               1
```

```
data2018 %>% group_by(cut(Difference, breaks = 20)) %>% summarise(n=n())
```

```
## # A tibble: 18 x 2
##    `cut(Difference, breaks = 20)`      n
##    <fct>                           <int>
##  1 (-2e+07,-1.22e+07]                  2
##  2 (-1.22e+07,-4.62e+06]              10
##  3 (-4.62e+06,2.99e+06]           374416
##  4 (2.99e+06,1.06e+07]               294
##  5 (1.06e+07,1.82e+07]                30
##  6 (1.82e+07,2.58e+07]                36
##  7 (2.58e+07,3.34e+07]                10
##  8 (3.34e+07,4.1e+07]                  8
##  9 (4.1e+07,4.86e+07]                  2
## 10 (4.86e+07,5.63e+07]                15
## 11 (5.63e+07,6.39e+07]                 2
## 12 (6.39e+07,7.15e+07]                 1
## 13 (7.15e+07,7.91e+07]                 5
## 14 (7.91e+07,8.67e+07]                 2
## 15 (8.67e+07,9.43e+07]                 1
## 16 (9.43e+07,1.02e+08]                26
## 17 (1.1e+08,1.17e+08]                  2
## 18 (1.25e+08,1.32e+08]                 2
```

We see that more than 3 lakh entries lie within a certain value. It is practical to leave out the outliers in further analysis

```
dataFilteredBackers = data2018 %>% filter(Difference < 2.99e+06 & backers < 2.19e+04)
ggplot(dataFilteredBackers, aes(y=Difference, x=backers)) +
    geom_point(shape=1) +
    geom_smooth(method=lm, se=FALSE)
```



```
ModelDiffBackers = lm(formula = Difference ~ backers ,data = dataFilteredBackers)
summary(ModelDiffBackers)
```

```
##
## Call:
## lm(formula = Difference ~ backers, data = dataFilteredBackers)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4529870   -19859   -16621    -6631  2966450
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21671.0569   163.8814   132.2   <2e-16 ***
## backers       -50.3379     0.3264  -154.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 98330 on 374358 degrees of freedom
## Multiple R-squared:  0.05973,    Adjusted R-squared:  0.05972
## F-statistic: 2.378e+04 on 1 and 374358 DF,  p-value: < 2.2e-16
```

From the plot we can see that with increasing number of backers, the Difference decreases. This conclusion is logical, as with greater number of contributors, the chances of meeting the goal fund should increase.

The fitted linear model is as follows: Difference = 21671.05 - 50.33* backers

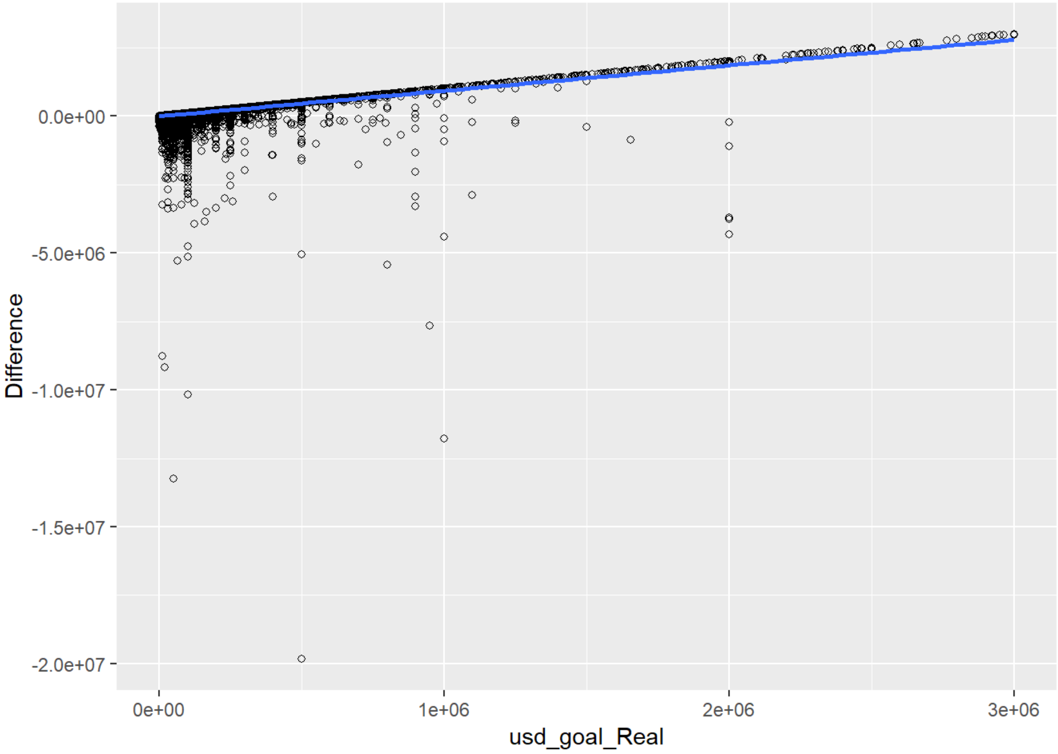
There is a strong negative correlation between Difference and backers. The p-value is < 0.001, suggesting that the model is highly significant. The Adjusted R-squared value however is 0.05972, suggesting the model is unable to explain most of the variance in the data.

2. Difference ~ usd_goal_real

```
data2018 %>% group_by(cut(usd_goal_Real , breaks = 20)) %>% summarise(n=n())
```

```
## # A tibble: 16 x 2
##    `cut(usd_goal_Real, breaks = 20)`       n
##    <fct>                               <int>
##  1 (-1.32e+05,6.62e+06]               374638
##  2 (6.62e+06,1.32e+07]                    95
##  3 (1.32e+07,1.99e+07]                    22
##  4 (1.99e+07,2.65e+07]                    35
##  5 (2.65e+07,3.31e+07]                     8
##  6 (3.31e+07,3.97e+07]                     3
##  7 (3.97e+07,4.63e+07]                     6
##  8 (4.63e+07,5.29e+07]                    13
##  9 (5.29e+07,5.96e+07]                     4
## 10 (5.96e+07,6.62e+07]                     2
## 11 (7.28e+07,7.94e+07]                     5
## 12 (7.94e+07,8.6e+07]                      2
## 13 (9.26e+07,9.93e+07]                     2
## 14 (9.93e+07,1.06e+08]                    25
## 15 (1.12e+08,1.19e+08]                     2
## 16 (1.26e+08,1.32e+08]                     2
```

```
dataFilteredGoal = data2018 %>% filter(Difference < 2.99e+06 & usd_goal_Real < 6.62e+06)
ggplot(dataFilteredGoal, aes(y=Difference, x=usd_goal_Real)) +
    geom_point(shape=1) +
    geom_smooth(method=lm,  se=FALSE)
```



```
ModelDiffGoal = lm(formula = Difference ~ usd_goal_Real ,data = dataFilteredGoal)
summary(ModelDiffGoal)
```

```
##
## Call:
## lm(formula = Difference ~ usd_goal_Real, data = dataFilteredGoal)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -20297945      2991      5356      5795    216804
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -5.378e+03  1.320e+02  -40.75   <2e-16 ***
## usd_goal_Real  9.287e-01  1.366e-03  679.83   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 78390 on 374426 degrees of freedom
## Multiple R-squared:  0.5524, Adjusted R-squared:  0.5524
## F-statistic: 4.622e+05 on 1 and 374426 DF,  p-value: < 2.2e-16
```

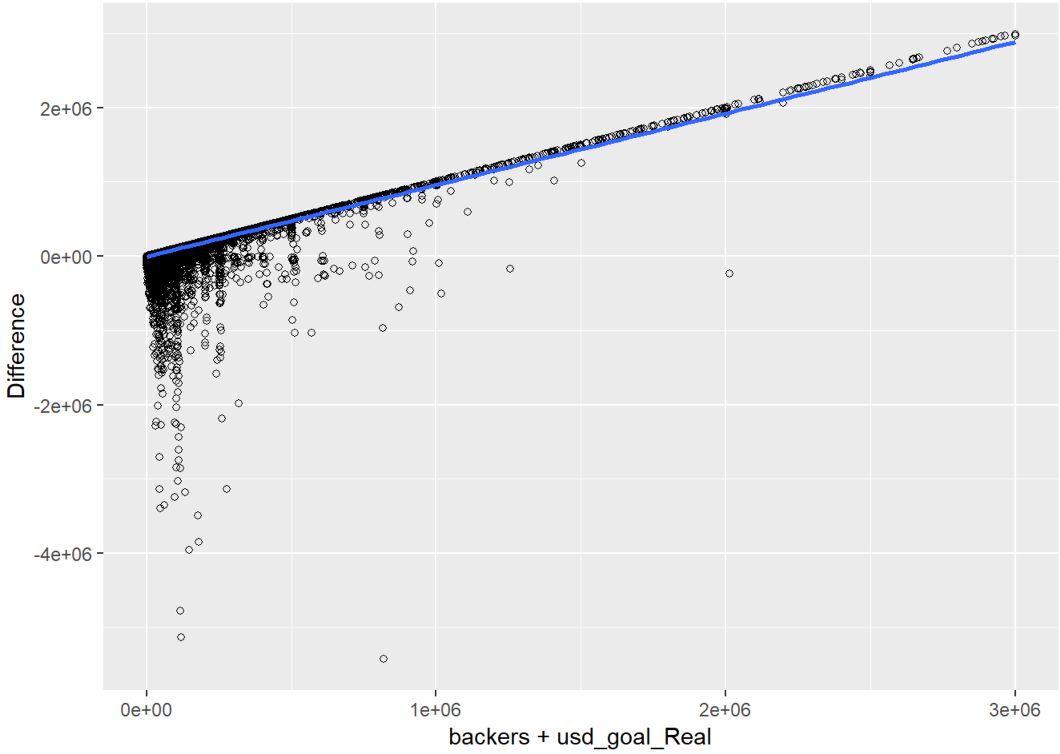
From the plot we can see that with a higher goal value for the project, the Difference increases. This conclusion is understandable, as it should be more difficult to crowdfund large project expenses.

The fitted linear model is as follows: Difference = -5.378e+03 + 0.9287* usd_goal_real

There is a positive correlation between Difference and usd_goal_real. The p-value is < 0.001, suggesting that the model is highly significant. The Adjusted R-squared value is 0.55, suggesting this model able to better explain the variance in data than the previous one.

3.Difference ~ usd_goal_real + backers

```
dataFilteredGoalBackers = data2018 %>% filter(Difference < 2.99e+06 & usd_goal_Real < 6.62e+06 & backers < 2
.19e+04)
ggplot(dataFilteredGoalBackers, aes(y=Difference, x=backers+usd_goal_Real)) + geom_point(shape=1) +   geom_s
mooth(method=lm,  se=FALSE)
```



```
ModelDiffBackersGoal = lm(formula = Difference ~ backers + usd_goal_Real ,data = dataFilteredGoalBackers)
summary(ModelDiffBackersGoal)
```

```
##
## Call:
## lm(formula = Difference ~ backers + usd_goal_Real, data = dataFilteredGoalBackers)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5090408       85      295      950  1193210
##
## Coefficients:
##                 Estimate Std. Error  t value Pr(>|t|)
## (Intercept)   -7.269e+01  6.086e+01   -1.194    0.232
## backers       -6.160e+01  1.183e-01 -520.769   <2e-16 ***
## usd_goal_Real  9.844e-01  6.242e-04 1577.058   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35570 on 374357 degrees of freedom
## Multiple R-squared:  0.877,  Adjusted R-squared:  0.877
## F-statistic: 1.334e+06 on 2 and 374357 DF,  p-value: < 2.2e-16
```

The fitted linear model is as follows: Difference = -7.269e+01 -6.160e+01* backers + 9.844e-01* usd_goal_real

The p-value is < 0.001, suggesting that the model is highly significant. The Adjusted R-squared value is 0.877, suggesting the model is able to explain the variance in the data very well.

4.Difference ~ usd_goal_real + backers + country + main_category

Country and Main_category being factor variables, causes the linear regression to treat levels in these variables as a separate coefficient.

```
summary(lm(formula = Difference ~ backers + usd_goal_Real + country + main_category ,data = dataFilteredGoal
Backers))
```

```
##
## Call:
## lm(formula = Difference ~ backers + usd_goal_Real + country +
##     main_category, data = dataFilteredGoalBackers)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5084465     -873      375     1504  1194471
##
## Coefficients:
##                              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                 3.488e+02  1.469e+03    0.237 0.812363
## backers                    -6.172e+01  1.195e-01 -516.292  < 2e-16 ***
## usd_goal_Real               9.862e-01  6.296e-04 1566.317  < 2e-16 ***
## countryAU                   1.756e+03  1.509e+03    1.164 0.244535
## countryBE                   3.648e+02  2.039e+03    0.179 0.858048
## countryCA                   9.417e+02  1.484e+03    0.635 0.525652
## countryCH                   8.183e+02  1.945e+03    0.421 0.673981
## countryDE                   1.734e+03  1.555e+03    1.115 0.264817
## countryDK                   7.203e+02  1.803e+03    0.400 0.689476
## countryES                   2.175e+03  1.634e+03    1.331 0.183298
## countryFR                   1.112e+03  1.595e+03    0.697 0.485692
## countryGB                   1.298e+02  1.468e+03    0.088 0.929538
## countryHK                   6.902e+03  2.039e+03    3.385 0.000713 ***
## countryIE                   1.058e+03  1.916e+03    0.552 0.580714
## countryIT                   1.576e+03  1.598e+03    0.986 0.324151
## countryJP                   2.575e+03  5.800e+03    0.444 0.656999
## countryLU                   6.981e+02  4.738e+03    0.147 0.882864
## countryMX                   1.764e+03  1.684e+03    1.047 0.294903
## countryNL                   1.447e+03  1.599e+03    0.905 0.365457
## countryNO                   3.687e+02  1.976e+03    0.187 0.851965
## countryNZ                   1.343e+03  1.729e+03    0.777 0.437198
## countrySE                   2.243e+03  1.684e+03    1.332 0.182848
## countrySG                   5.816e+03  2.095e+03    2.776 0.005501 **
## countryUS                  -4.075e+02  1.457e+03   -0.280 0.779648
## main_categoryComics         2.939e+03  4.020e+02    7.311 2.66e-13 ***
## main_categoryCrafts         3.718e+02  4.336e+02    0.857 0.391193
## main_categoryDance         -2.910e+02  6.161e+02   -0.472 0.636718
## main_categoryDesign        -1.175e+03  2.961e+02   -3.968 7.26e-05 ***
## main_categoryFashion       -6.037e+02  3.167e+02   -1.906 0.056606 .
## main_categoryFilm & Video -1.035e+03  2.553e+02   -4.054 5.03e-05 ***
## main_categoryFood          -6.277e+02  3.103e+02   -2.023 0.043116 *
## main_categoryGames          3.296e+03  2.860e+02   11.525  < 2e-16 ***
## main_categoryJournalism     2.203e+02  5.575e+02    0.395 0.692705
## main_categoryMusic          6.379e+01  2.654e+02    0.240 0.810045
## main_categoryPhotography   -4.778e+02  4.024e+02   -1.187 0.235183
## main_categoryPublishing     8.360e+02  2.773e+02    3.015 0.002572 **
## main_categoryTechnology    -5.303e+03  2.918e+02  -18.173  < 2e-16 ***
## main_categoryTheater       -5.791e+02  4.009e+02   -1.444 0.148638
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35510 on 374322 degrees of freedom
## Multiple R-squared:  0.8774, Adjusted R-squared:  0.8774
## F-statistic: 7.241e+04 on 37 and 374322 DF,  p-value: < 2.2e-16
```

The fitted linear model is as follows: Difference = 3.488e+02 - 6.172e+01* backers + 9.862e-01 *usd_goal_real + 1.756e+03* countryAU + 3.648e+02*countryBE +.. + 2.939e+03* main_categoryComics + 3.718e+02*main_categoryCrafts

For each factor variable, the value can be either 0/1, reducing the entire linear model to be an equation in “backers” and “usd_goal_real”. The addition of these columns do not change the Adjusted R-squared value, suggesting no value addition in the linear regression model upon their inclusion.

CONCLUSION:

1. It can be seen that only about 8 percent of the projects have exceeded their funding by more than 50 percent of the goal.

2.After having observed multiple plot of different subcategories of main catogeries it could be observed that the goal doesnt depend on the number of sub categories of a main category.For example we have plotted for the main category “publishing”.

3.After having observed multiple categories ,taking for example the “food” category ,the failure rate is higher than the success rate after launch of the projects and also that for the “food” category in July 2014 ,the products launched were maximum compared to other months and also between the months of the years 2014-2017.

4.It can be observed that the product design has the highest number of projects followed by documentary and table top games.

5.Upon studying the effects of different factors on the success/failure of generating the goal money through linear regression models, we can say that failure increases linearly with the target goal amount and decreases linearly with the number of backers. There was no apparent effect of the country and category the project belonged to.