Kick Starters Project Analysis

Team Members:

1. Sai Krishna Mannava - 801136361
2. Kumar Mani Chandra Yelisetty - 801168244
3. Smirthi Meenakshisundaram - 801129947

Programming language: R language

Pros:

- open source
- data wrangling(using dplyr, readr),
- easy plotting using ggplot2
- compatible with other languages
- platform independent
- machine learning operation are easy

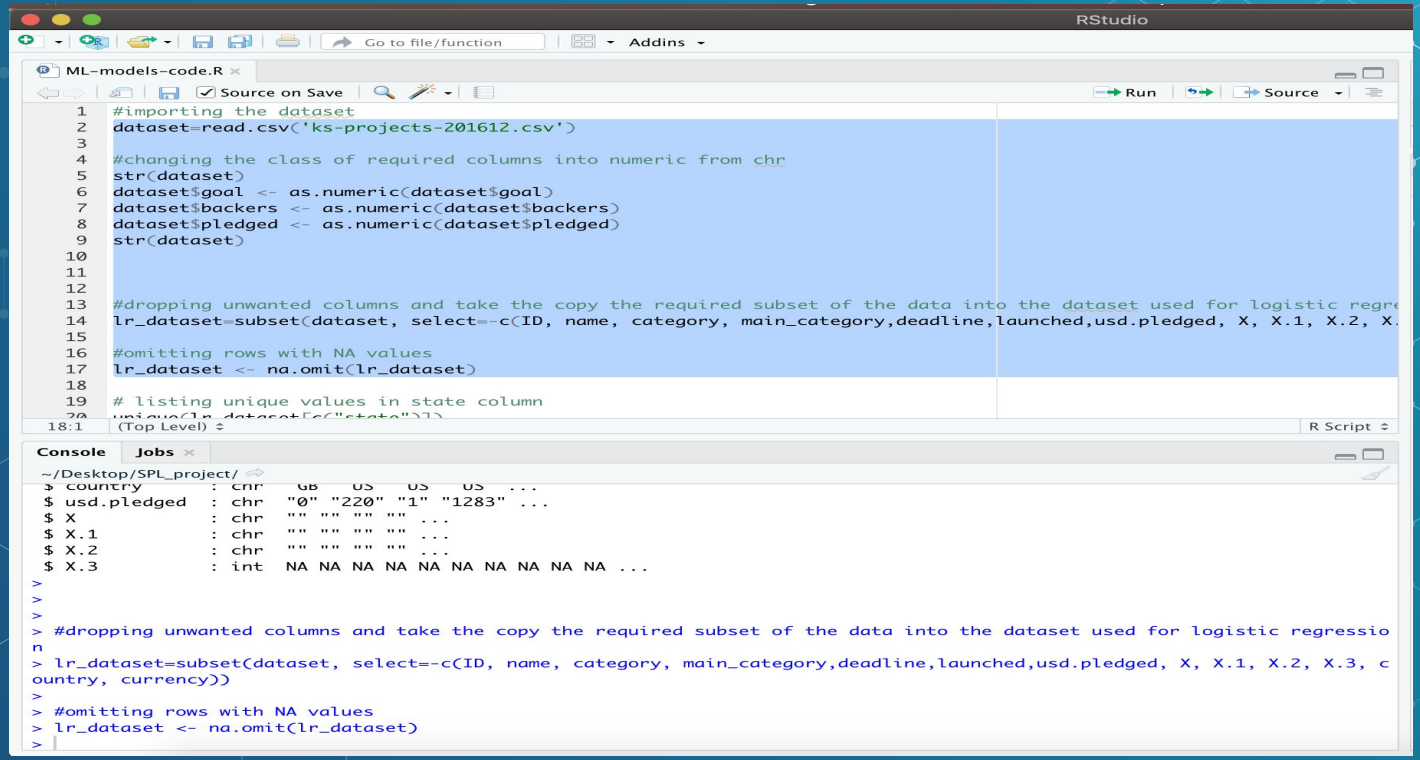
Cons

- data handling is not as easy as in python
- cannot handle big data
- consumes lot of memory
- lacks security
- speed of R is not as much as python or MATLAB

Our Dataset

ID	name	category	main_category	currency	deadline	goal	launched	pledged	state	backers	country	usd pledged	usd_pledged_real	usd_goal_real
1000002330	The Songs of Adelaide & Abullah	Poetry	Publishing	GBP	2015-10-09	1000	2015-08-11 12:12:28	0	failed	0	GB	0	0	1533.95
1000003930	Greeting From Earth: ZGAC Arts Capsule For ET	Narrative Film	Film & Video	USD	2017-11-01	30000	2017-09-02 04:43:57	2421	failed	15	US	100	2421	30000
1000004038	Where is Hank?	Narrative Film	Film & Video	USD	2013-02-26	45000	2013-01-12 00:20:50	220	failed	3	US	220	220	45000
1000007540	ToshiCapital Rekordz Needs Help to Complete Album	Music	Music	USD	2012-04-16	5000	2012-03-17 03:24:11	1	failed	1	US	1	1	5000
1000011046	Community Film Project: The Art of Neighborhood Filmmaking	Film & Video	Film & Video	USD	2015-08-29	19500	2015-07-04 08:35:03	1283	canceled	14	US	1283	1283	19500
1000014025	Monarch Espresso Bar	Restaurants	Food	USD	2016-04-01	50000	2016-02-26 13:38:27	52375	successful	224	US	52375	52375	50000
1000023410	Support Solar Roasted Coffee & Green Energy! SolarCoffee.co	Food	Food	USD	2014-12-21	1000	2014-12-01 18:30:44	1205	successful	16	US	1205	1205	1000
1000030581	Chaser Strips. Our Strips make Shots their B*tch!	Drinks	Food	USD	2016-03-17	25000	2016-02-01 20:05:12	453	failed	40	US	453	453	25000
1000034518	SPIN - Premium Retractable In-Ear Headphones with Mic	Product Design	Design	USD	2014-05-29	125000	2014-04-24 18:14:43	8233	canceled	58	US	8233	8233	125000
100004195	STUDIO IN THE SKY - A Documentary Feature Film (Canceled)	Documentary	Film & Video	USD	2014-08-10	65000	2014-07-11 21:55:48	6240.57	canceled	43	US	6240.57	6240.57	65000
100004721	Of Jesus and Madmen	Nonfiction	Publishing	CAD	2013-10-09	2500	2013-09-09 18:19:37	0	failed	0	CA	0	0	2406.39
100005484	Lisa Lim New CD!	Indie Rock	Music	USD	2013-04-08	12500	2013-03-09 06:42:58	12700	successful	100	US	12700	12700	12500
1000055792	The Cottage Market	Crafts	Crafts	USD	2014-10-02	5000	2014-09-02 17:11:50	0	failed	0	US	0	0	5000
1000056157	G-Spot Place for Gamers to connect with eachother & go pro!	Games	Games	USD	2016-03-25	200000	2016-02-09 23:01:12	0	failed	0	US	0	0	200000
1000057089	Tombstone: Old West tabletop game and miniatures in 32mm.	Tabletop Games	Games	GBP	2017-05-03	5000	2017-04-05 19:44:18	94175	successful	761	GB	57763.78	121857.33	6469.73
1000064368	Survival Rings	Design	Design	USD	2015-02-28	2500	2015-01-29 02:10:53	664	failed	11	US	664	664	2500
1000064918	The Beard	Comic Books	Comics	USD	2014-11-08	1500	2014-10-09 22:27:52	395	failed	16	US	395	395	1500
1000068480	Notes From London: Above & Below	Art Books	Publishing	USD	2015-05-10	3000	2015-04-10 21:20:54	789	failed	20	US	789	789	3000
1000070642	Mike Corey's Darkness & Light Album	Music	Music	USD	2012-08-17	250	2012-08-02 14:11:32	250	successful	7	US	250	250	250
1000071625	Boco Tea	Food	Food	USD	2012-06-02	5000	2012-05-03 17:24:32	1781	failed	40	US	1781	1781	5000
1000072011	CMUK. Shoes: Take on Life Feet First.	Fashion	Fashion	USD	2013-12-30	20000	2013-11-25 07:06:11	34268	successful	624	US	34268	34268	20000
1000081649	MikeyJ clothing brand fundraiser	Childrenswear	Fashion	AUD	2017-09-07	2500	2017-08-08 01:20:20	1	failed	1	AU	0	0.81	2026.1
1000082254	Alice in Wonderland in G Minor	Theater	Theater	USD	2014-06-15	3500	2014-05-16 10:10:38	650	failed	12	US	650	650	3500
1000087442	Mountain brew: A quest for alcohol sustainability	Drinks	Food	NOK	2015-02-25	500	2015-01-26 19:17:33	48	failed	3	NO	6.18	6.29	65.55

Pre-processing Data



Encoding values , splitting & Scaling the testing and training dataset

RStudio

Go to file/function | Addins

ML-models-code.R

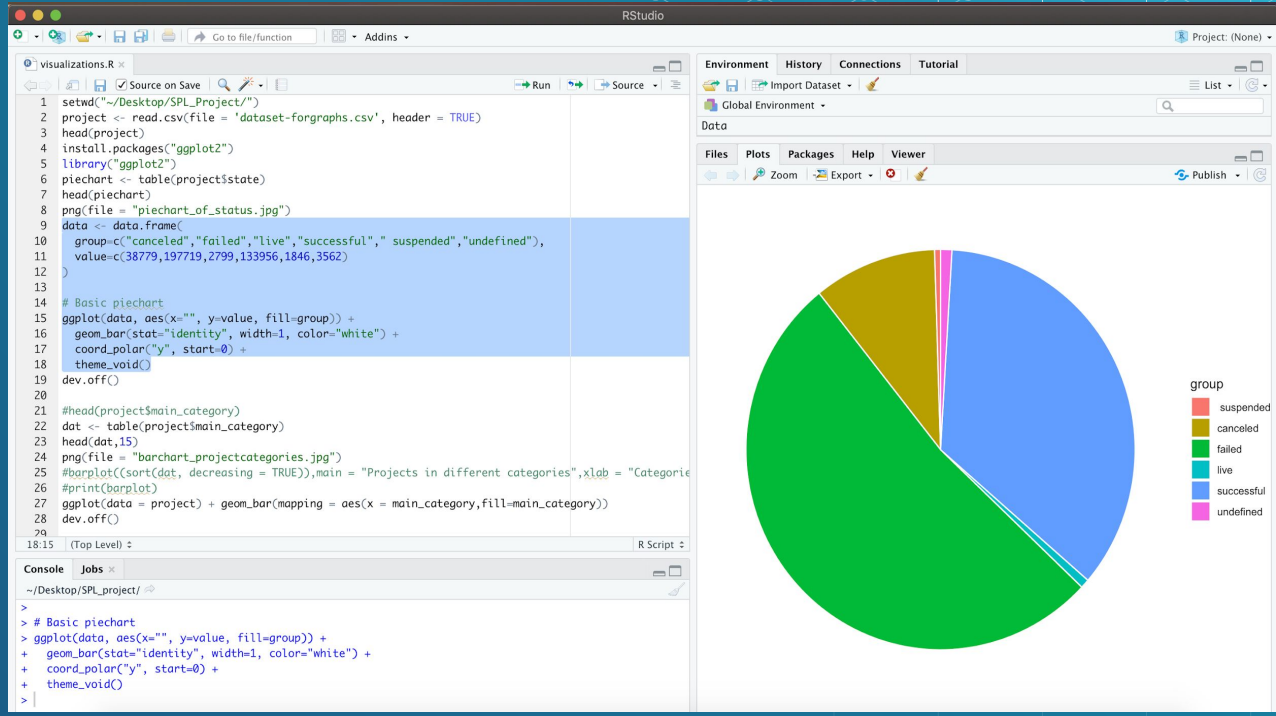
Source on Save | Run | Source

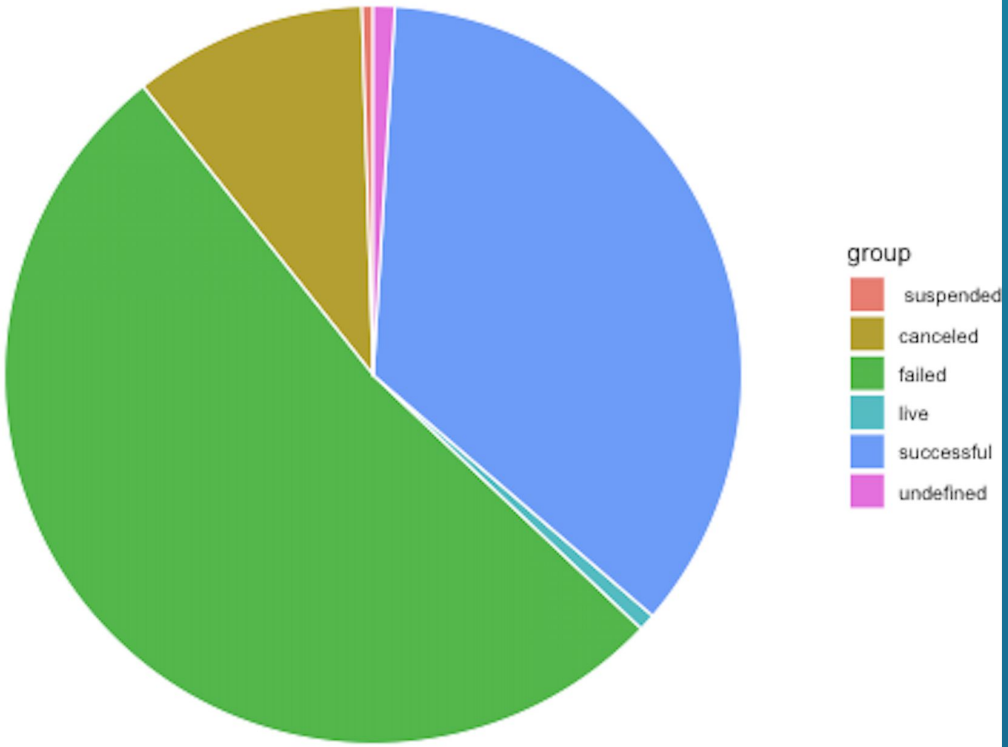
```r
#encoding categorical values of state column into numerical values
lr_dataset$state = factor(lr_dataset$state, levels= c('successful', 'failed', 'canceled', 'live', 'undefined', 'suspe
unique(lr_dataset[c("state")])


unique(dataset[c("currency")])
dataset<- subset(dataset, currency =="GBP" | currency=="USD" | currency == "CAD" | currency =="NOK" | currency=="AUD'
unique(dataset[c("currency")])
#dataset$backers=ifelse(is.na(dataset$backers),ave(dataset$backers, FUN= function(x) mean(x,na.rm=TRUE)),dataset$bacl

#intalling the package used for splitting the dataset
install.packages('caTools')
library(caTools)
set.seed(150)

#splitting the dataset into trainging 80% and remaining test dat
split =sample.split(lr_dataset$state, SplitRatio = 0.8)
lr_training_set=subset(lr_dataset, split==TRUE)
lr_test_set=subset(lr_dataset, split=FALSE)

#loop for Feature Scaling of numeric columns in training_data
performScaling <- TRUE  # Turn it on/off for experimentation.

if (performScaling) {

  # Loop over each column.
  for (colName in names(lr_training_set)) {

    # Check if the column contains numeric data.
    if(class(lr_training_set[,colName]) == 'integer' | class(lr_training_set[,colName]) == 'numeric') {

      # Scale this column (scale() function applies z-scaling).
      lr_training_set[,colName] <- scale(lr_training_set[,colName])
    }
```

Spellcheck

Visualizing the Data

- Ggplot2, dplyr
- Pie chart to show the projects categorized by their state(status of work)

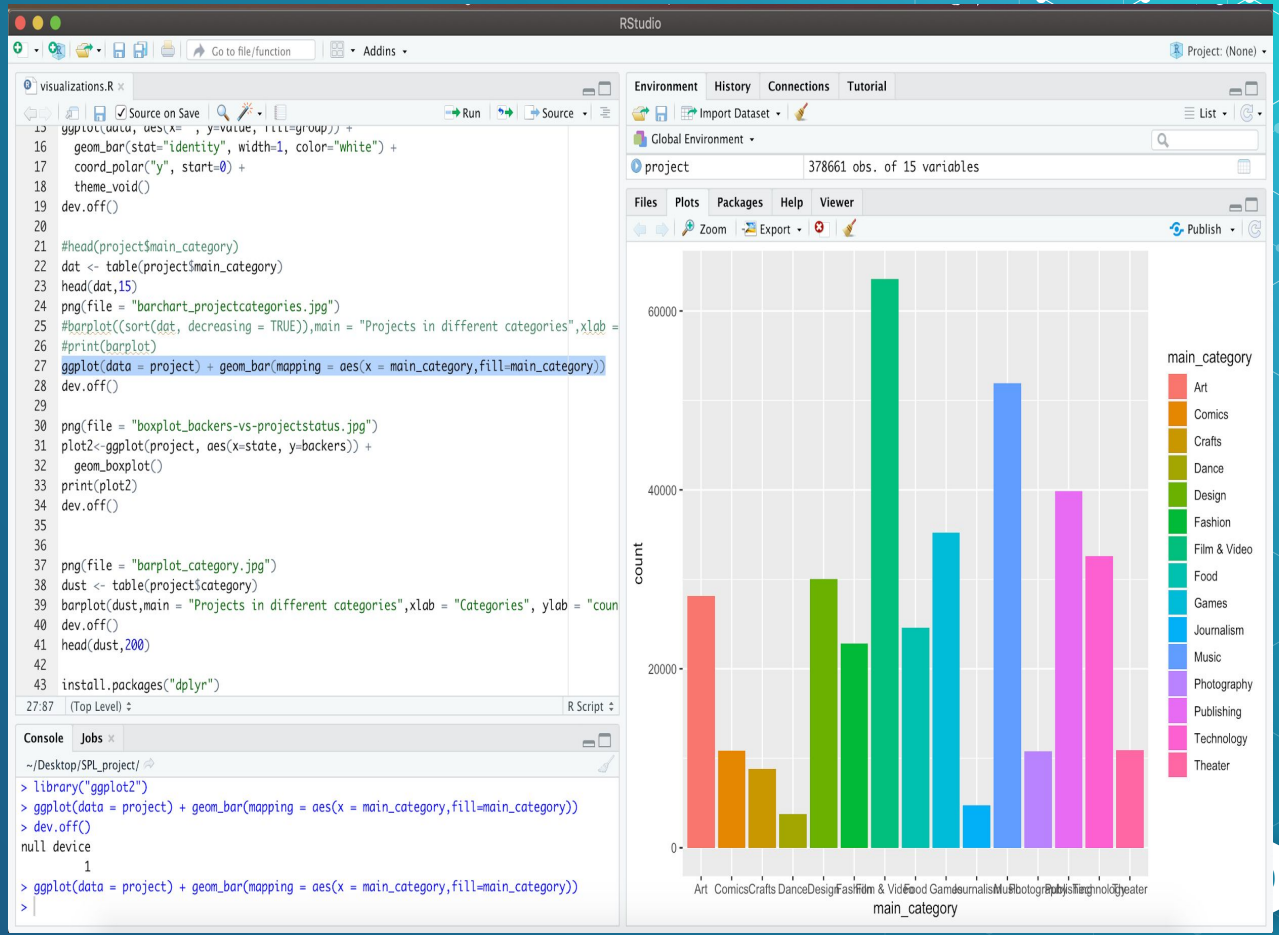


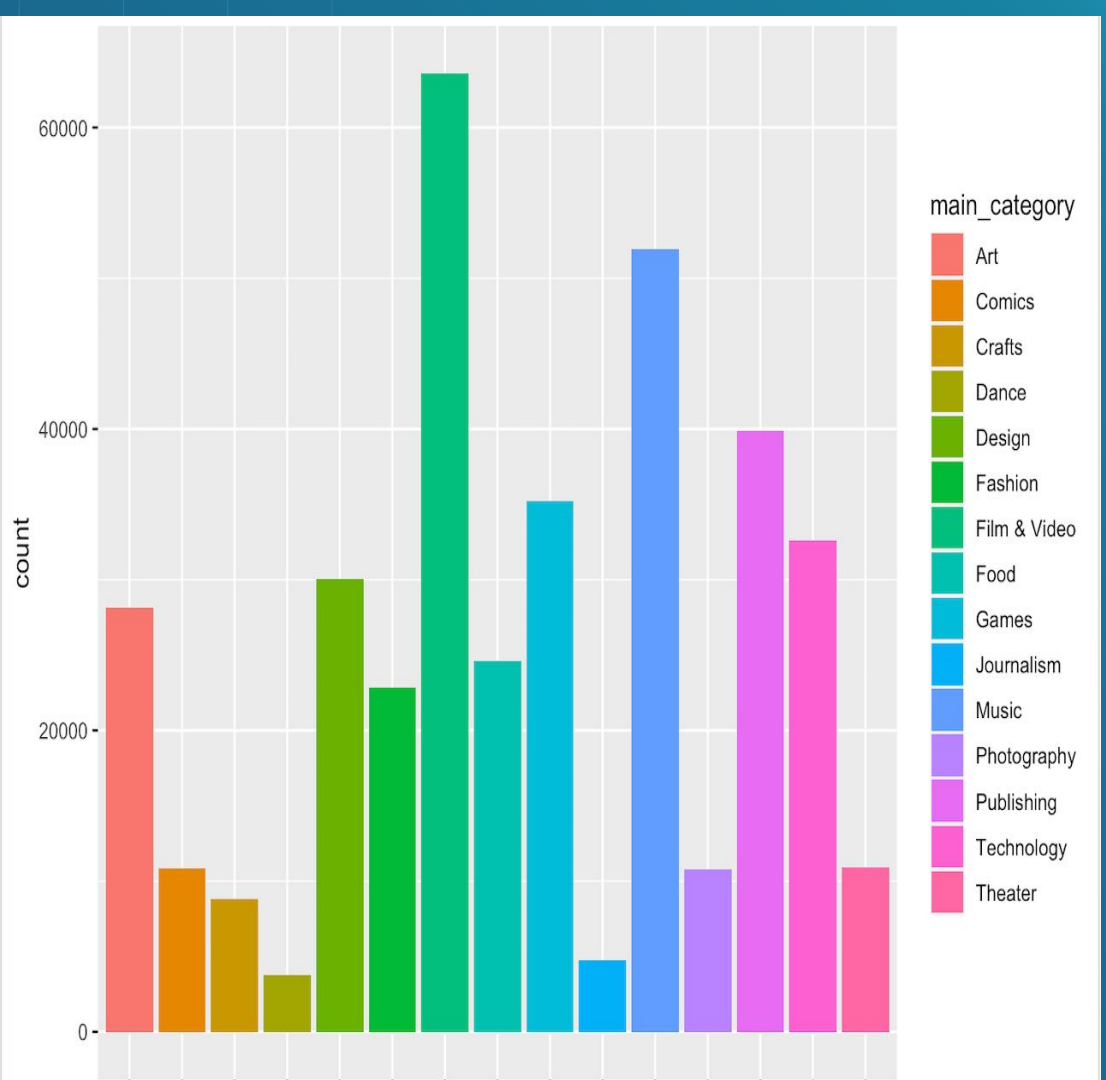


Most of the projects have failed, considerable amount have been successful and some of them have been either canceled, suspended or they are live

Bar chart

Category vs projects

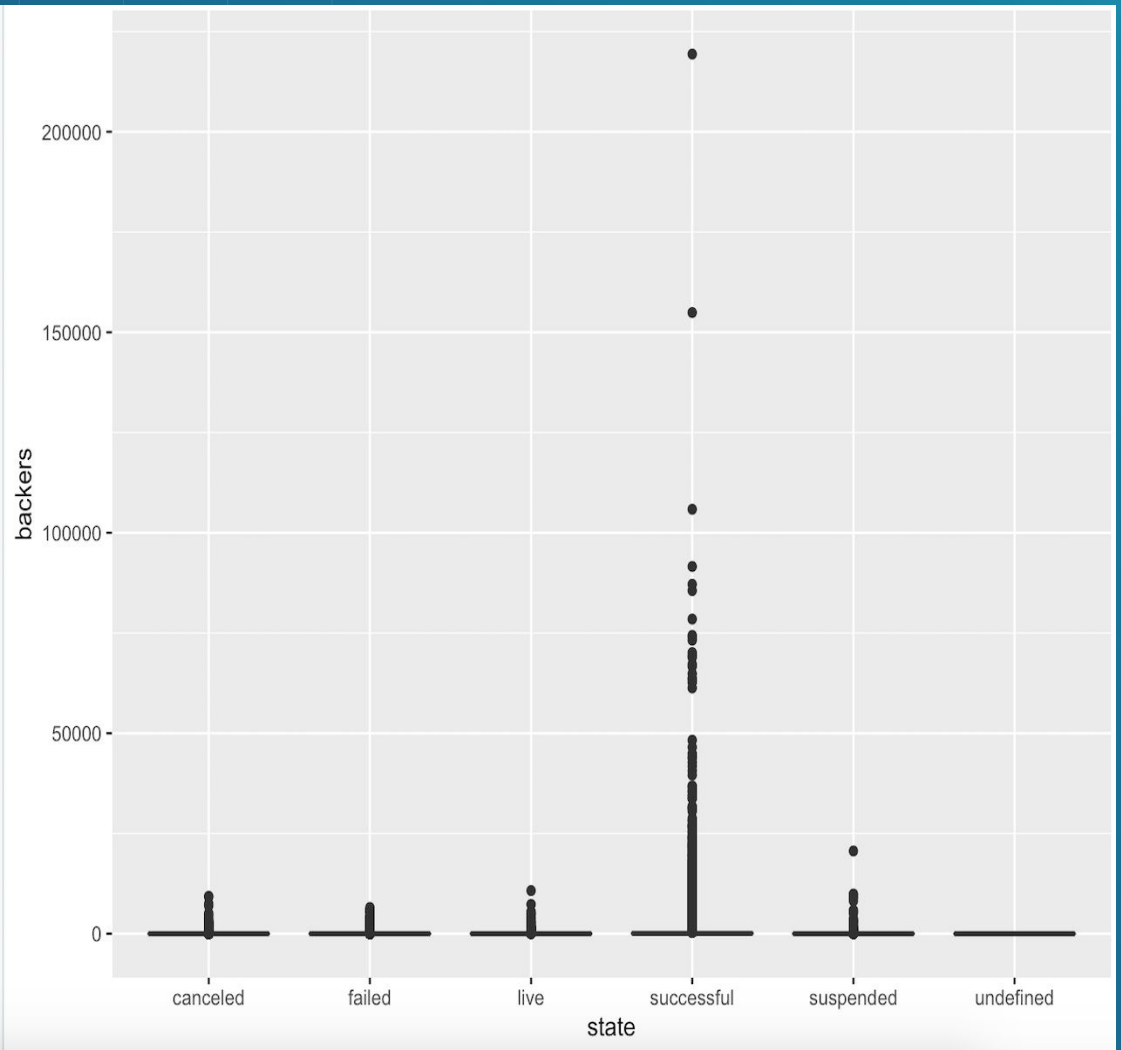




x -axis: category

Y -axis: count of the projects

Film and Video is the category which has the maximum number of projects, followed by Music, Publishing and other Categories.



Box Plot Backers vs state

In this plot we see that as the number of backers that a project gets increases there is a greater probability of the project to be successful

Predicting the probability of success using Machine Learning Models

1. Pre-processing data
2. Selecting Features for testing and Training the Dataset
3. Encoding values , splitting & Scaling the testing and training dataset
4. Predicting using Linear Regression, Random Forest and Decision Tree Model

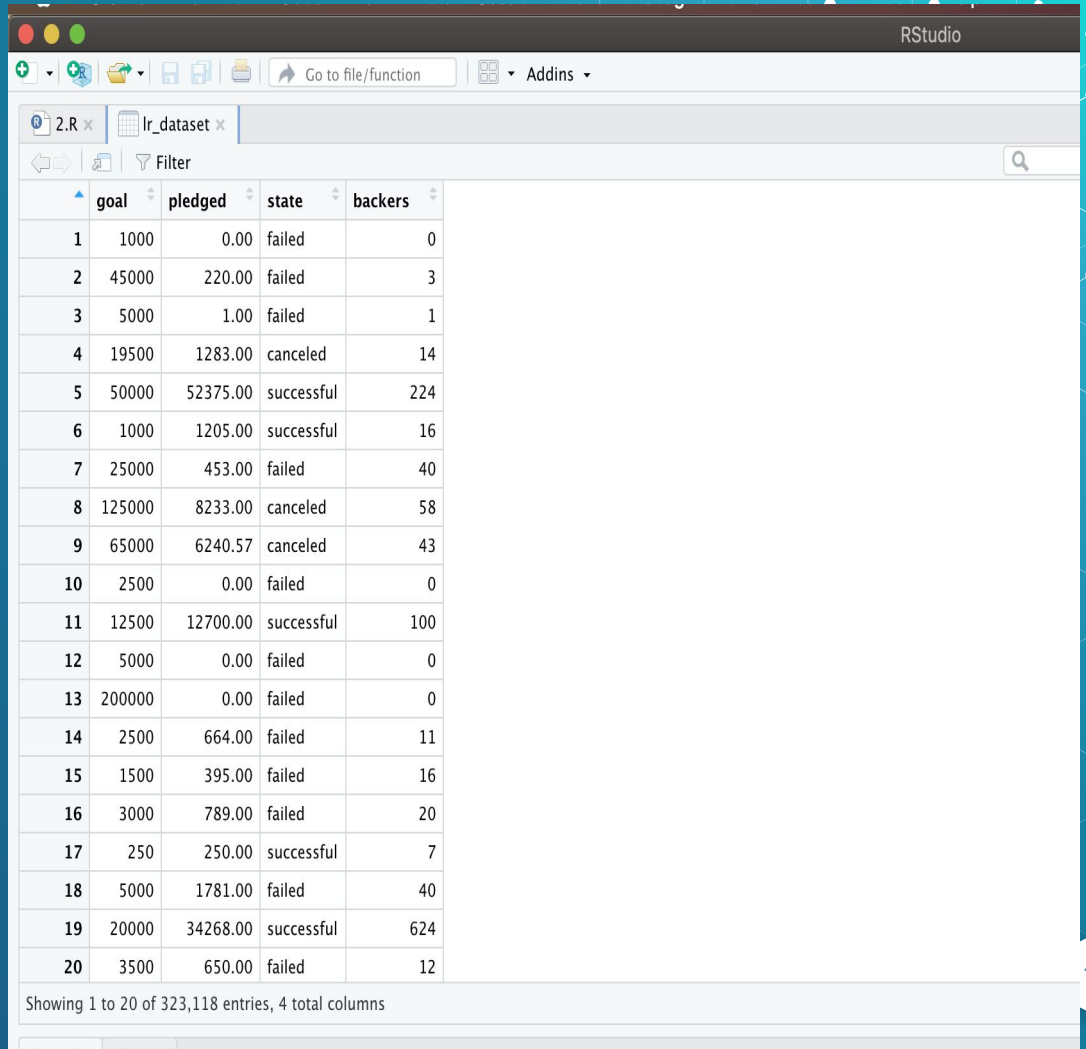
Selecting Features

RStudio

ML-models-code.R

```
#encoding categorical values of state column into numerical values
lr_dataset$state = factor(lr_dataset$state, levels= c('successful', 'failed', 'canceled', 'live', 'undefined', 'suspe
unique(lr_dataset[c("state")])


unique(dataset[c("currency")])
dataset<- subset(dataset, currency =="GBP" | currency=="USD" | currency == "CAD" | currency =="NOK" | currency=="AUD'
unique(dataset[c("currency")])
#dataset$backers=ifelse(is.na(dataset$backers),ave(dataset$backers, FUN= function(x) mean(x,na.rm=TRUE)),dataset$bacl

#intalling the package used for splitting the dataset
install.packages('caTools')
library(caTools)
set.seed(150)

#splitting the dataset into trainging 80% and remaining test dat
split =sample.split(lr_dataset$state, SplitRatio = 0.8)
lr_training_set=subset(lr_dataset, split==TRUE)
lr_test_set=subset(lr_dataset, split=FALSE)

#loop for Feature Scaling of numeric columns in training_data
performScaling <- TRUE  # Turn it on/off for experimentation.

if (performScaling) {

  # Loop over each column.
  for (colName in names(lr_training_set)) {

    # Check if the column contains numeric data.
    if(class(lr_training_set[,colName]) == 'integer' | class(lr_training_set[,colName]) == 'numeric') {

      # Scale this column (scale() function applies z-scaling).
      lr_training_set[,colName] <- scale(lr_training_set[,colName])
    }
```

Features selected for Testing and Training



	goal	pledged	state	backers
1	1000	0.00	failed	0
2	45000	220.00	failed	3
3	5000	1.00	failed	1
4	19500	1283.00	canceled	14
5	50000	52375.00	successful	224
6	1000	1205.00	successful	16
7	25000	453.00	failed	40
8	125000	8233.00	canceled	58
9	65000	6240.57	canceled	43
10	2500	0.00	failed	0
11	12500	12700.00	successful	100
12	5000	0.00	failed	0
13	200000	0.00	failed	0
14	2500	664.00	failed	11
15	1500	395.00	failed	16
16	3000	789.00	failed	20
17	250	250.00	successful	7
18	5000	1781.00	failed	40
19	20000	34268.00	successful	624
20	3500	650.00	failed	12

Showing 1 to 20 of 323,118 entries, 4 total columns

ML-models-code.R

```
81
82  #Fitting the logistic regression to the training set
83  classifier =glm(formula=state ~ ., family =binomial, data=lr_training_set, na.action = na.pass )
84
85  lr_test_set
86  lr_test_set[3]
87  unique(lr_test_set[3])
88
89  #Predicting the test set results
90  prob_pred =predict(classifier, type='response', newdata =lr_test_set[-3] )
91
92  #probability results
93  prob_pred
94  min(prob_pred)
95  max(prob_pred)
96  #converting the propabiltiy results to 0 to 1
97  #y_pred =ifelse(prob_pred < 0.167, 0, ifelse (0.167<=prob_pred < 0.33, 1,ifelse(0.33<=prob_pred<0.5,2,ifelse(0.5<= pr
98  y_pred <- ifelse(prob_pred<0.5,0,1)
99  unique(y_pred)
100 unique(lr_test_set[,3])
101 #making the confusion matrix
102 cm =table(y_pred,lr_test_set[,3])
103 cm
104 install.packages('e1071')
105 library(e1071)
106
107
108 result1 <- confusionMatrix(cm)
109 result1
```

109:8 (Top Level) R Script

Console Jobs

~/Desktop/SPL_project/

```
y_pred      0      1
     0 111762   4872
     1   1319 205165

               Accuracy : 0.9808
                 95% CI : (0.9804, 0.9813)
    No Information Rate : 0.65
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9582

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9883
```

Linear Regression Model predicts with an accuracy of 98.08%

ML-models-code.R

```
#install.packages('randomForest')
library(randomForest)

RF_classifier = randomForest(x= RF_training_set[-3], y=RF_training_set$state, ntree= 100 )

RF_prob_pred =predict(RF_classifier, type='response', newdata =RF_test_set[-3] )
unique(RF_prob_pred)

RF_cm =table(RF_prob_pred,RF_test_set[,3])
RF_cm
install.packages('caret')
library(caret)
install.packages('e1071')
library(e1071)

result <- confusionMatrix(RF_cm)
result

#install.packages('rpart')
library(rpart)
```

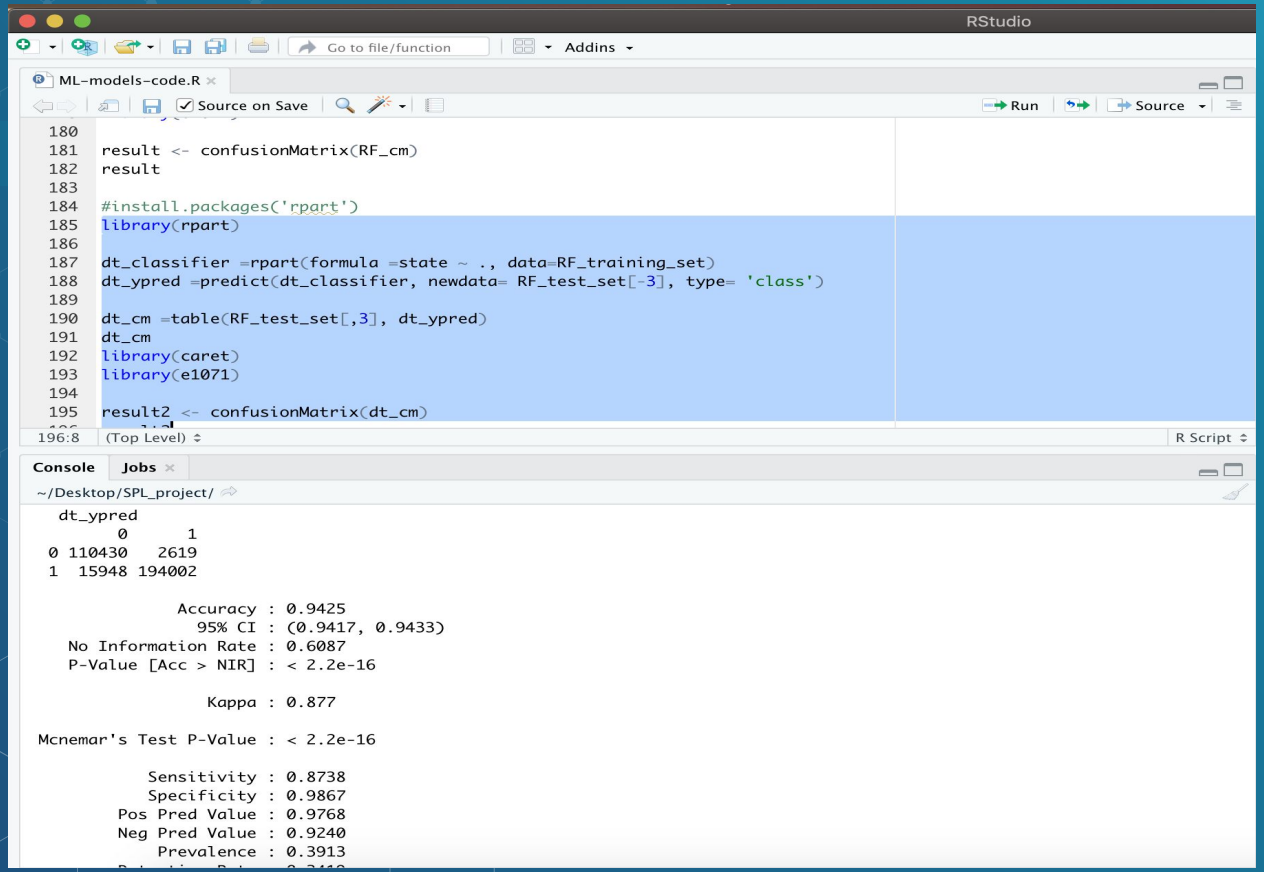
182:7 (Top Level) R Script

Console Jobs

~/Desktop/SPL_project/

```
RF_prob_pred      0      1
           0 113032  33286
           1     17 176664

               Accuracy : 0.8969
                 95% CI : (0.8958, 0.8979)
    No Information Rate : 0.65
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7878

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9998
            Specificity : 0.8415
         Pos Pred Value : 0.7725
         Neg Pred Value : 0.9999
             Prevalence : 0.3500
         Detection Rate : 0.3499
   Detection Prevalence : 0.4530
      Balanced Accuracy : 0.9207

       'Positive' Class : 0
```

Random Forest Model predicts with an accuracy of 89.69%



Decision Tree Model predicts with an accuracy of 94.25%

Conclusion

1. Linear Regression model has given the best results with an accuracy of 98.08% compared to the Random Forest model of accuracy 89.69% and Decision tree model with accuracy 94.35%

2. The visualizations plotted with libraries shows the most endorsed categories, effect of backers on the success of a project and the status of various projects

THANK YOU