What Makes A Successful Startup Company?

Kickstarter Data Analysis

Lab Group - Tessa Grabowski, Himanshu Jain, Paul Noujaim, Tri Truong

Introduction

Our general research question to find out what are the most important factors in contributing to the success of a startup project in the United States. Our dataset is comprised of the statistics from over 300,000 Kickstarter proposals, collected directly from the Kickstarter Platform (found on Kaggle). It includes variables that could be essential to determining the success of a startup such as, the amount of money pledged to a startup, the number of backers the project has, or the industry the company is in. The key variable we are looking at is the binary variable called "state", which shows which startups were successful and which were not. This will be our response variable in the analyses. We think this analysis of this data would be particularly useful if any member of our group wanted to start our own creative project by giving us an idea about which factors are key in indicating the future success of a startup.

Kickstarter is a global crowdfunding platform where different products can be listed in different categories like music, arts, technology etc. Till date, the company has received over $4.6 billion in funding from almost 17.2 million backers. We believe that it would be interesting to analyze the data from this company to recognize the reason behind its success and how useful it might be for upcoming projects. Many new products are launched everyday so it would also provide an insight to the developer as to what products would have a higher rate of success.

The different variables in this dataset are- ID, name (name of the project), category(category of the project), main category, currency, deadline, goal (amount of money required), launched, pledged (amount of money the project got), state, backers, country, and usd pledged. We believe that some of these variables would be really important in providing us an insight about the data/company.

Data Analysis Plan

```
# A tibble: 11,468 x 14
      X1     ID name  category main_category currency deadline             goal
   <dbl>  <dbl> <chr> <chr>    <chr>         <chr>    <dttm>               <dbl>
 1     1 1.46e9 Euro~ Documen~ Film & Video  USD      2014-07-01 06:00:00   700
 2     2 2.29e8 Fant~ Printing Crafts        USD      2015-05-25 22:25:43   777
 3     3 1.51e8 9 st~ Comedy   Film & Video  USD      2015-10-15 16:17:00  1200
 4     4 8.53e8 Casc~ Apparel  Fashion       USD      2016-03-14 20:00:00  4100
 5     5 2.66e8 Wate~ Documen~ Film & Video  USD      2012-11-24 23:58:57  9500
 6     6 5.11e8 My T~ Documen~ Film & Video  USD      2014-09-22 20:41:50 10000
 7     7 7.40e8 Gifw~ Product~ Design        USD      2015-12-09 19:28:40 10000
 8     8 1.81e9 My B~ Video G~ Games         USD      2016-09-16 12:09:38  1000
 9     9 1.40e9 The ~ Hip-Hop  Music         USD      2012-08-31 19:31:37  1750
10    10 1.50e9 Ryan~ Photogr~ Photography   USD      2015-09-03 17:09:46  5000
# ... with 11,458 more rows, and 6 more variables: launched <dttm>,
#   pledged <dbl>, state <fct>, backers <dbl>, country <chr>, usd_pledged <dbl>
```

```
# A tibble: 2 x 3
  state          n  rate
  <fct>      <int> <dbl>
1 successful  4788 0.418
2 failed      6680 0.582
```

From this table, we see that most of the companies have a state of either failed or successful. Thus when we are analyzing our data, we will filter it to only include these companies. We are interested in what makes a project successful versus unsuccessful, so any entry with a different state will be irrelevant to our analysis. We can also see that most of the companies failed so it is really important to understand the problems in order to launch a successful campaign. Hence, we believe that this analysis would be helpful for future entrepreneurs.

```
# A tibble: 15 x 3
   main_category     n share_of_projects
   <chr>         <int>             <dbl>
 1 Film & Video   2148            0.187
 2 Music          1833            0.160
 3 Publishing     1279            0.112
 4 Art             872            0.0760
 5 Games           808            0.0705
 6 Food            787            0.0686
 7 Technology      787            0.0686
 8 Design          753            0.0657
 9 Fashion         634            0.0553
10 Theater         365            0.0318
11 Comics          338            0.0295
12 Photography     331            0.0289
13 Crafts          252            0.0220
14 Dance           149            0.0130
15 Journalism      132            0.0115
```

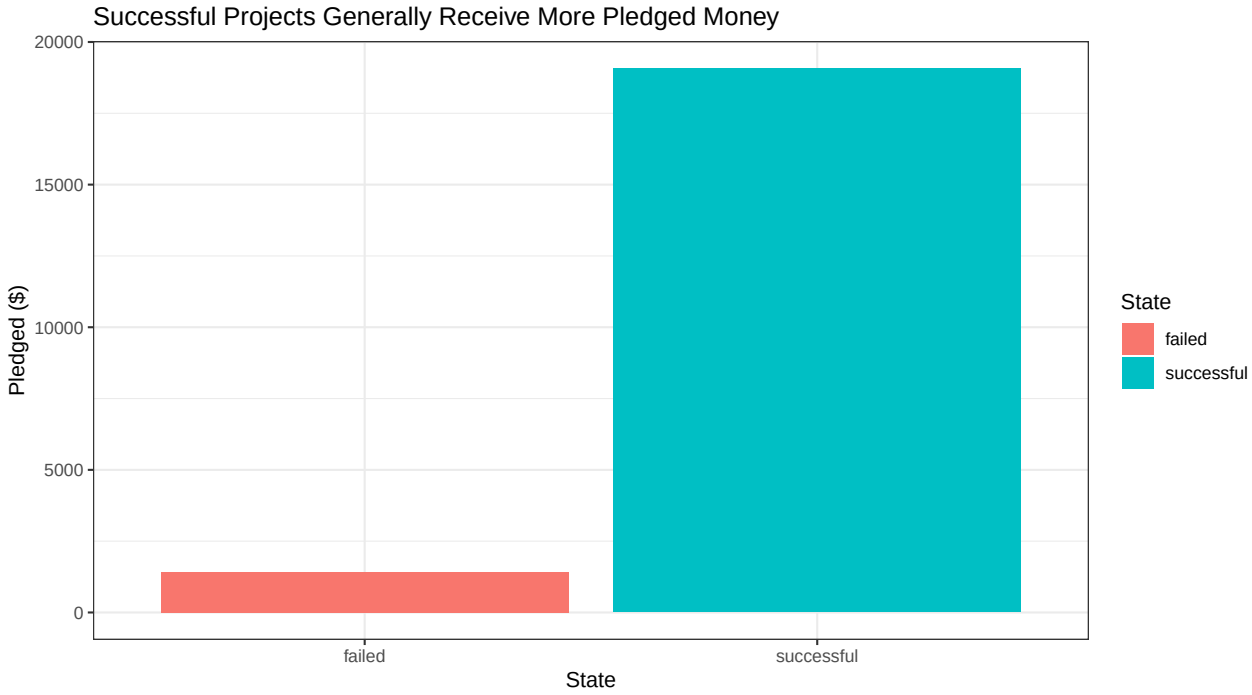
```
# A tibble: 34 x 2
# Groups:   category [34]
   category           n
   <chr>          <int>
 1 Documentary      567
 2 Product Design   549
 3 Shorts           475
 4 Music            465
 5 Food             432
 6 Tabletop Games   342
 7 Film & Video     334
 8 Fiction          329
 9 Nonfiction       301
10 Rock             295
# ... with 24 more rows
```

Looking at the numbers in the categories and main categories, we can get a better idea at what kind of projects are more successful than others. Knowing this will help us narrow down what the crowd is interested in investing in, and what fields people are more likely to succeed in if they were to create their own project. We can see that Film and Video is the most common category and Product Design is the most common sub-category. We also chose to take only those columns that had more than 100 campaigns since the data is not clean and there are a lot of values that would not be allow us to present desired results.

The variable `pledged` tells the amount of money pledged by the crowd, or users of the site.

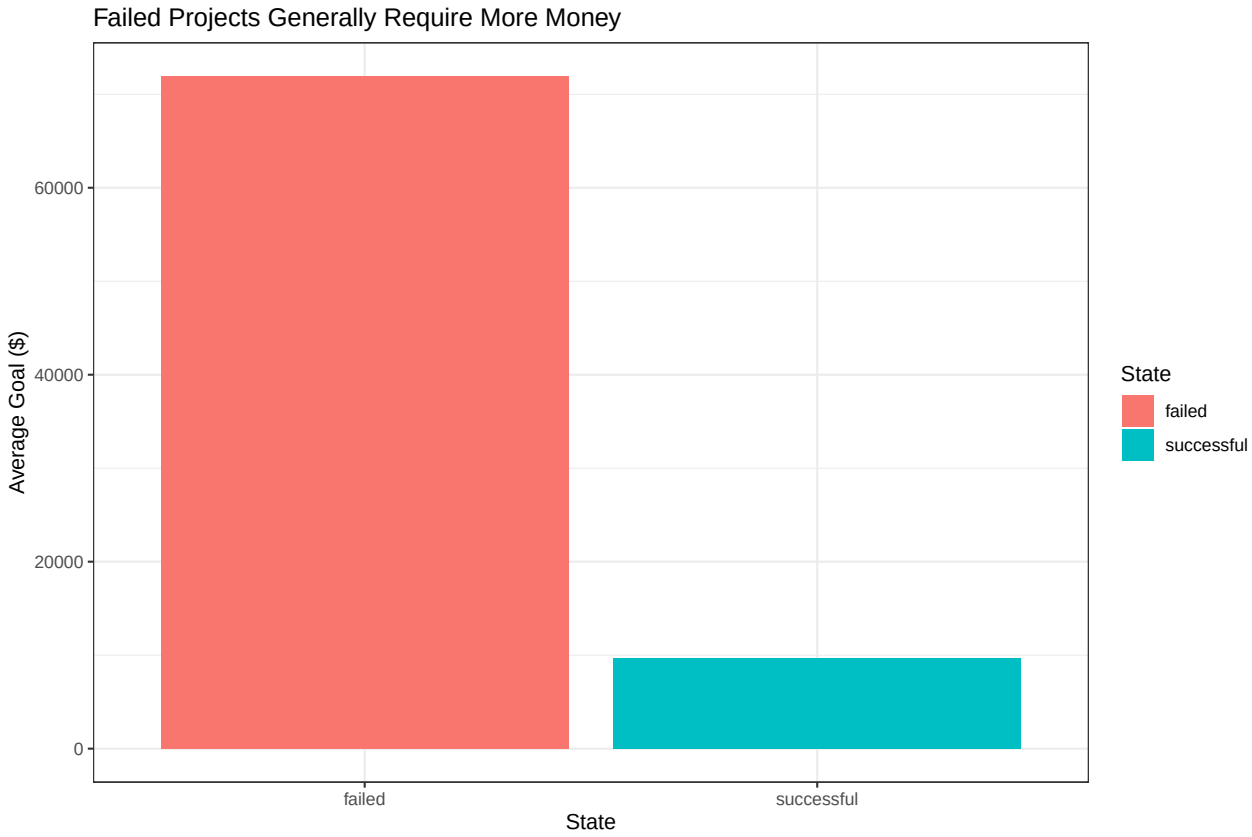
```
# A tibble: 2 x 6
```

```
  state        mean   min      max goal_mean    range
  <fct>       <dbl> <dbl>    <dbl>     <dbl>    <dbl>
1 failed      1420.     0  460657.    71929.  460657.
2 successful 19071.     1 1924018      9710. 1924017
```



This table shows that there is a large gap between the amount of money pledged to projects that failed as opposed to those that were successful. The average amount of money pledged to successful projects is more than ten times greater than that pledged to failed projects. We can also see this in the bar plot above, as the bar for successful projects is significantly higher than the bar for failed. In addition, the range of money pledged to successful projects is significantly greater than the range of failed projects. We can also see that the failed projects required much more money than the successful projects and even then they didn't receive much money. We believe that the large amount of money required makes people not pledge money for that project and hence, we believe that we should be having a realistic goal for the project to be successful on kickstarter. These statistics suggest that the variable `pledged`has an effect on determining the state of a project.

Another variable we are interested in is `goal`. `Goal`gives the amount of money needed by a creator to fund and finish their project.



This visualization shows that failed projects generally were much more expensive to create than successful projects. This is very useful in our analysis, since it suggests that there is possibly a threshold to how much money can be required by the creator before the project becomes unreasonable and fails.

In our analysis it could be interesting to compare the goal of a creator to how much they received from pledged money as well. Looking at these two variables and how they interact could be very telling of the outcome of the project.

Codebook

We can see that there are 323750 rows and 13 relevant columns in the dataset we are using.

Variable-> Label ID -> ID of the project that was listed name-> Name of the Project that was listed category->Category of the project that was listed main_category-> Main catefgory to which the project belonged currency-> The currency funding was requested In deadline-> The deadline to get the required funding goal-> The amount that was requested launched-> The date when the funding was started pledged-> The amount pledged by the backers state-> The final outcome of the project backers-> Number of people who funded the project country-> Country where the project was launched usd pledged-> US Dollars that the project got

Field Name->Value Label currency AUD Australian Dollars CAD Canadian Dollars CHF Swiss Franc DKK Danish Krone EUR Euro GBP Pound Sterling MXN Mexican Peso NOK Norwegian Krone NZD New Zealand Dollar SEK Swedish krona USD US Dollar

country AT Austria AU Australia CA Canada CH Switzerland DE Germany DK Denmark ES Spain FR France GB United Kingdom IE Ireland IT Italy MX Mexico NL Netherlands NO Norway NZ New Zealand SE Sweden SG Singapore US United States

Statistical Methods

Further, in our analysis, we will use other statistical methods to answer our question of what makes a project successful. We will create our own hypothetical projects and determine statistics for the needed variables. Then we will use the knn method to predict if this project will be successful or not based on the data we have.

We also will use confidence intervals for the mean amount of money required, number of backers necessary, amount pledged, etc. that we expect a successful company to have.

Finally, we will use a linear regression model to figure out which variables are the most significant in determining the eventual 'state' of the startup, whether they ended up being successful or unsuccessful.

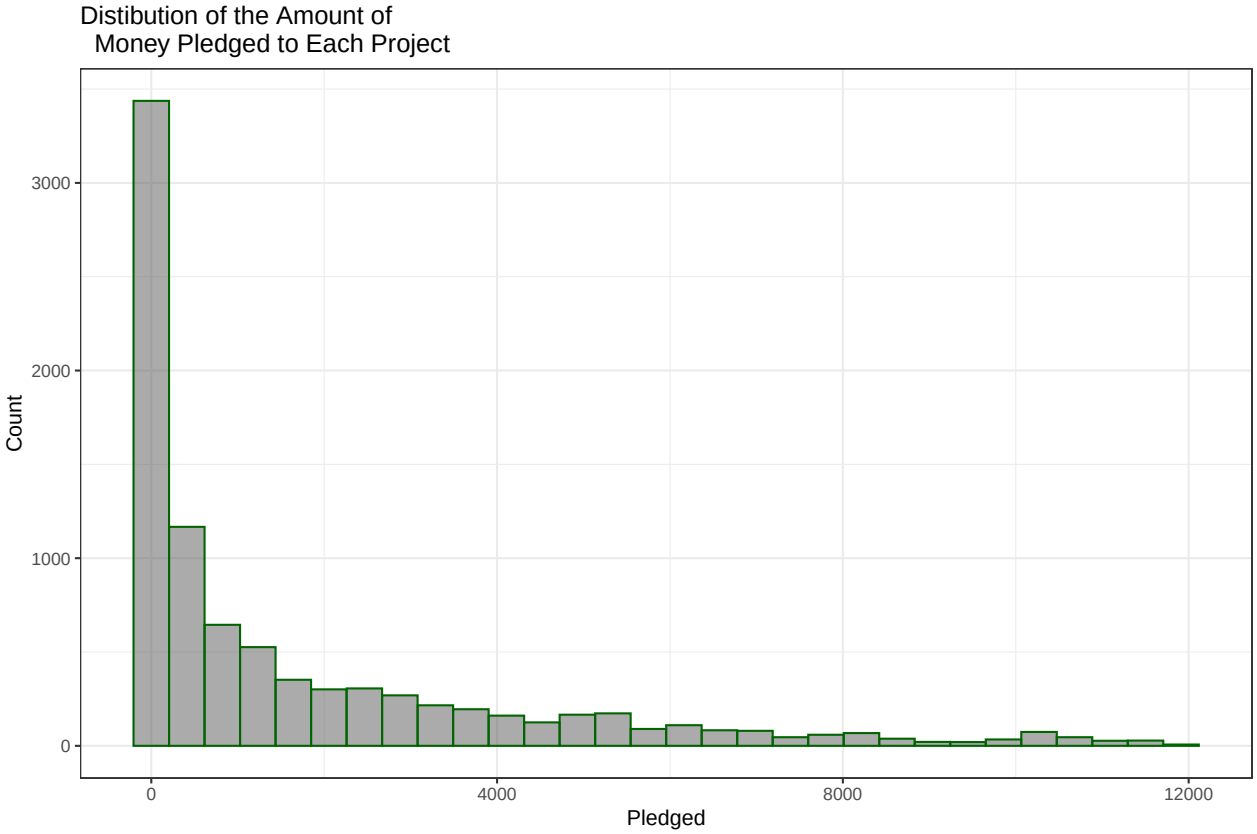
Using these methods, we will hopefully be able to determine why certain projects are successful while others are not and use this information to help us in the future if we ever want to start our own companies.

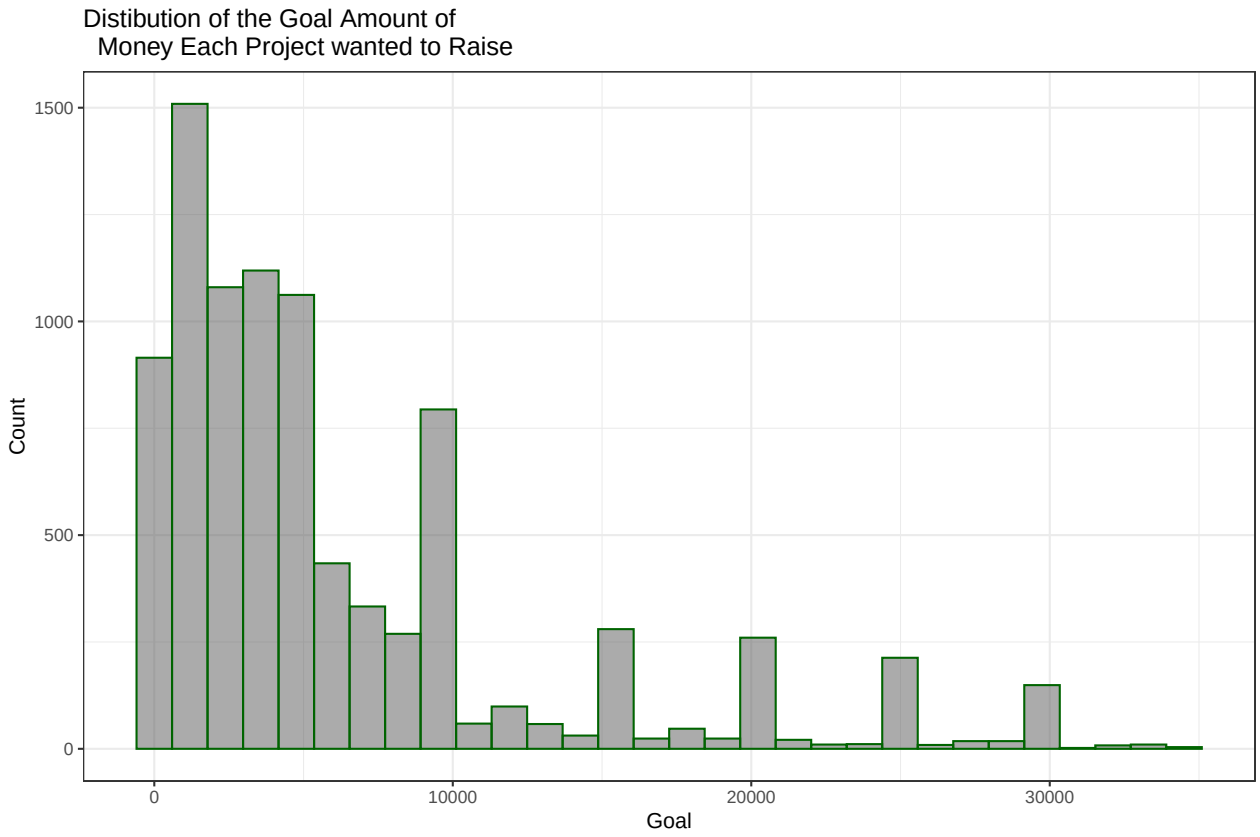
Clean Data

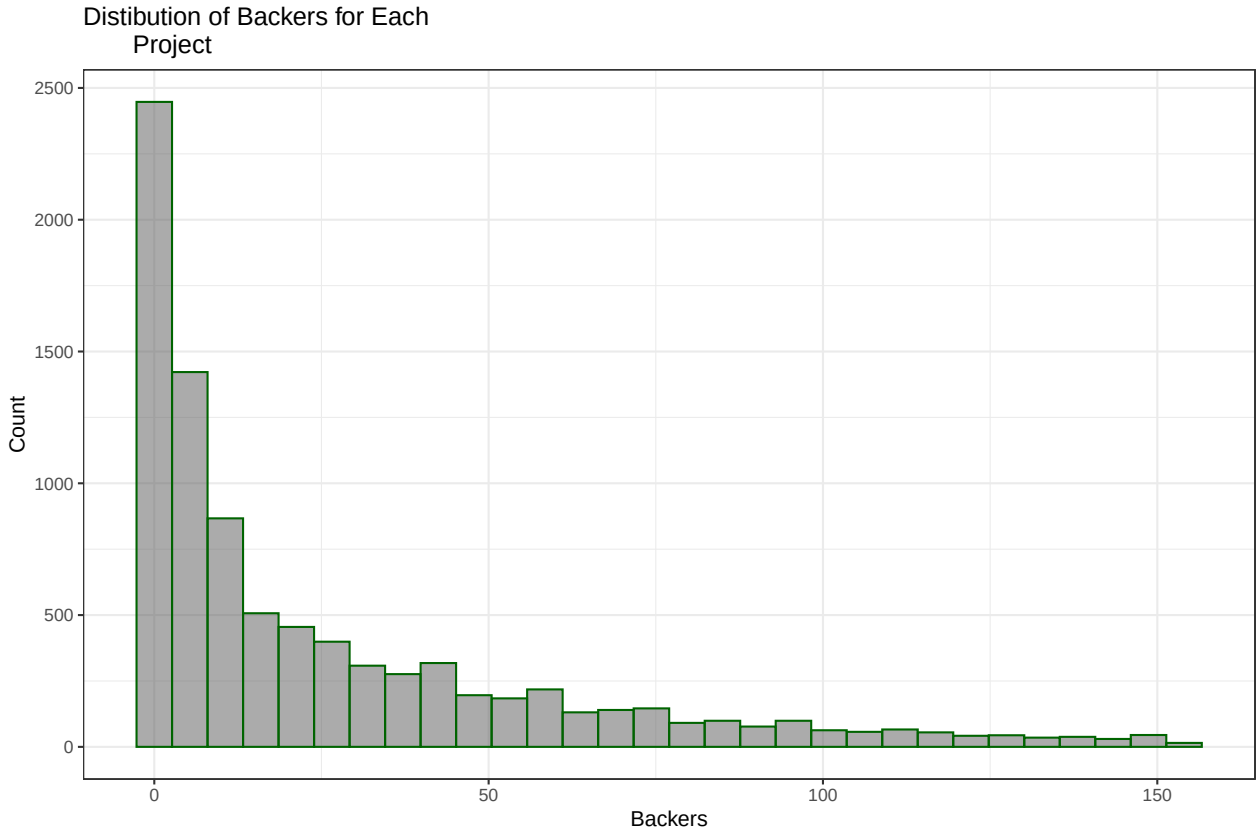
We filtered out the data since we want to concentrate on the startups that either failed or were successful. There are many other states but some of them have garbage values so we filtered the dataset. We are also looking at US specific startups since other countries have different startup ecosystem and just looking at this dataset and making conclusions for them would give us inaccurate results. We are basically using this to make sure that all the observations used have things in common.

Since our dataset had too many observations, we decided to stick to using a smaller fraction of it in order to not break our program.

The outliers were observed and removed to make the dataset more generic and to get better results. Outliers would have an effect on our results.




Distibution of the Goal Amount of
Money Each Project wanted to Raise
Count
1500
1000
500
0
0
10000
20000
30000
Goal


Distibution of Backers for Each
Project
Count
2500
2000
1500
1000
500
0
0
50
100
150
Backers

The graphs clearly shows that big projects are hardly launched on kickstarter. It also shows that the pledged amount is typically smaller than the goal amount typically for larger projects. Additionally, the number of backers is only above 100 for only a small number of projects.

```
# A tibble: 2 x 3
  state          n  freq
  <fct>      <int> <dbl>
1 failed       465 0.145
2 successful  2738 0.855
```

```
# A tibble: 2 x 3
  state          n  freq
  <fct>      <int> <dbl>
1 failed      6215 0.752
2 successful  2050 0.248
```

By defining a large project with one that has over $4000 pledged to it, we can see that larger projects tend to have a much higher success rate relative to their smaller counterparts.

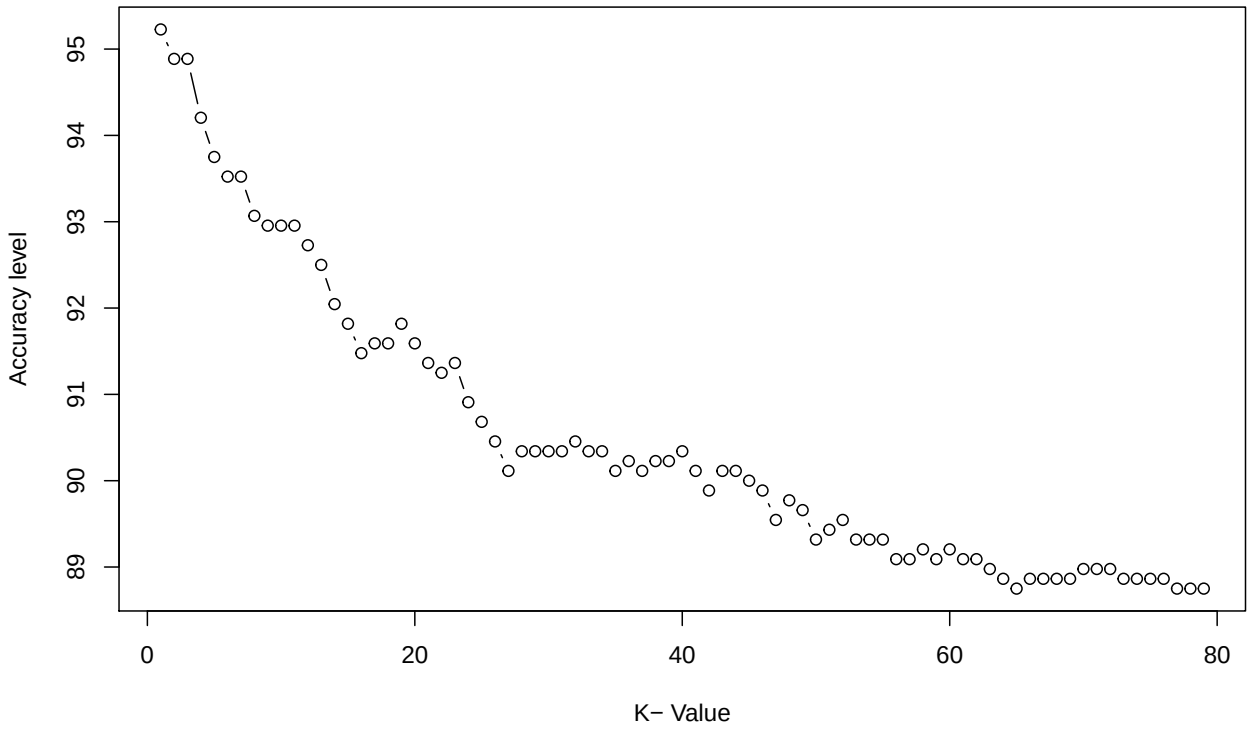
We standardised the variables to improve the accuracy of our KNN models. KNN is a distance based algorithm which is affected by the scale of the variables. Hence, we need to make sure that each feature has 0 mean and 1 standard deviation.

Methods of Statistical Analysis and Results

K-NN

```
1 = 95.22727 2 = 94.88636 3 = 94.88636 4 = 94.20455 5 = 93.75 6 = 93.52273 7 = 93.52273 8 = 93.06818 9 =
```

```
# A tibble: 1 x 1
  values
   <dbl>
1  0.952
```




```
[1] 1
```

We set the indices, the number of observations to be tested on, to be 880.

We make the vectors required for training and testing the model. We set the indices to be 880. The model is tested on the number of indices observations that we get from the dataset.

We create two new datasets, kickstart_train and kickstart_test. kickstart_train contains only observations that are not in the row indices created, and kickstart_test contains only observations that are in the row indices created.

We create a vector of the class labels for the training dataset and call it train_state. We create a vector of true kickstart state in our test dataset and call it true_state.

We fit the k-nearest neighbors model on our raw training dataset. We let k varying from 1 to 79 and calculate the value of k that results in the greatest prediction accuracy in our dataset along with its associated prediction accuracy. We then plot all the values of k and its associated prediction accuracy to examine the trend.

We get the highest accuracy of 95.23% when the value of K is 1. The graph also displays this outcome and shows that prediction accuracy tends to decrease as k increases.

Logistic Regression Model

```
# A tibble: 4 x 5
  term           estimate std.error statistic    p.value
  <chr>             <dbl>     <dbl>     <dbl>      <dbl>
1 (Intercept)   -39628.      8251.     -4.80  0.00000156
2 goal        -1449204.    301660.     -4.80  0.00000155
3 backers           7.61      29.6     0.257 0.797
4 pledged       53238.     11085.      4.80  0.00000156
```

$$\widehat{\text{state}} = (intercept) + (goal_num) \text{ goal} + (backers_num) \text{ backers} + (pledged_num) \text{ pledged}$$

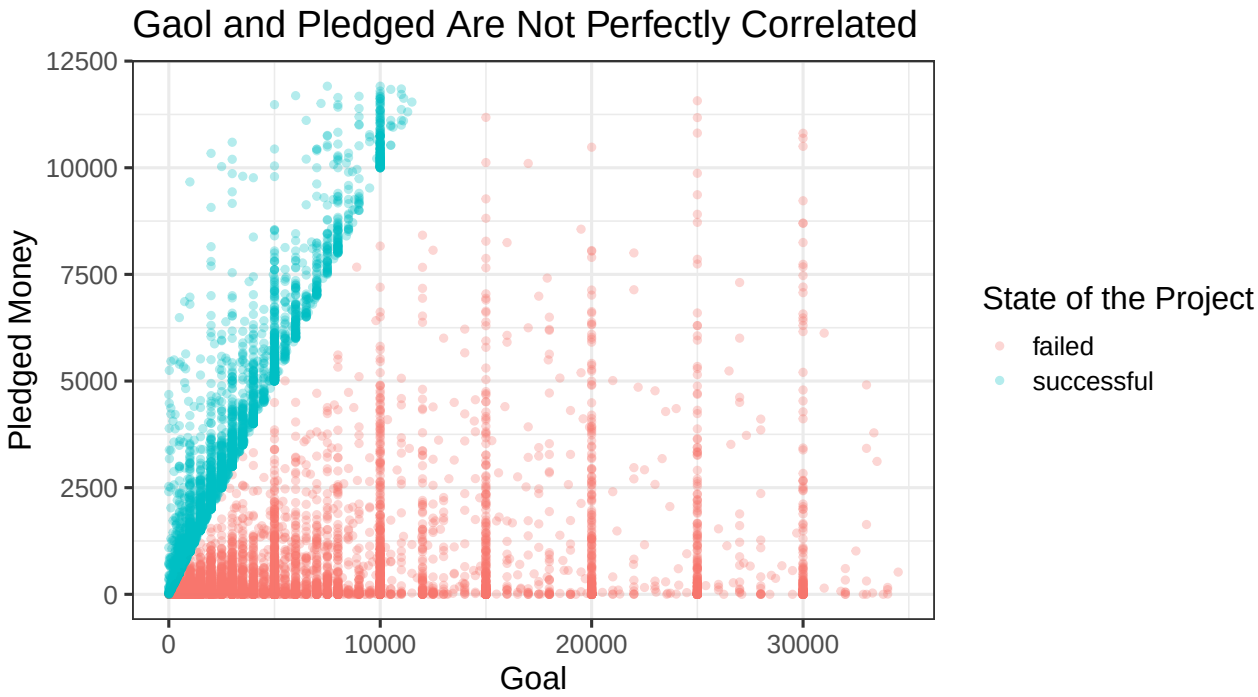
```
# A tibble: 3 x 5
  term         estimate std.error statistic    p.value
  <chr>           <dbl>     <dbl>     <dbl>      <dbl>
1 (Intercept)   -48503.    10441.     -4.65 0.00000339
2 goal        -1773677.   381771.     -4.65 0.00000339
3 pledged        65162.    14027.      4.65 0.00000339
```

$$\widehat{\text{state}} = -4121.634 - 98315.777 \text{ goal} + 17313.216 \text{ pledged}$$

We believe that goal and the pledged amount are the two most important variables responsible for predicting the startups success. We started with a model that had backers as one of the factors but from our backward eliminiation process we can conclude that only goal and pledged are the necessary variables.

```
[1] 1
```

```
# A tibble: 2 x 3
  term        estimate p.value
  <chr>          <dbl>   <dbl>
1 (Intercept) 0.000152   0.988
2 goal        0.0139     0.290
```



discuss how compares to knn test We can see that the accuracy for a regression model is 99.88%, which is much more than the accuracy we got from our KNN model. The KNN model had the highest accuracy at 99%.

Does category change anything?

```
# A tibble: 18 x 5
   term                      estimate std.error statistic   p.value
```

```
   <chr>                        <dbl>     <dbl>     <dbl>     <dbl>
 1 (Intercept)                10.2        7.65    1.33    0.184
 2 goal                       -0.846      0.213  -3.96    0.0000740
 3 backers                     0.0597     0.141   0.423   0.672
 4 pledged                     0.845      0.213   3.96    0.0000746
 5 main_categoryComics         2.70     191.      0.0142  0.989
 6 main_categoryCrafts         1.99      72.7     0.0274  0.978
 7 main_categoryDance          3.63     280.      0.0130  0.990
 8 main_categoryDesign         0.269     28.5     0.00941 0.992
 9 main_categoryFashion        3.83       8.72    0.439   0.661
10 main_categoryFilm & Video  -3.06       8.36   -0.366   0.714
11 main_categoryFood          -6.85       7.61   -0.901   0.368
12 main_categoryGames         -2.71      24.7    -0.110   0.913
13 main_categoryJournalism    59.1     1303.      0.0453  0.964
14 main_categoryMusic         35.6      145.      0.245   0.807
15 main_categoryPhotography   29.4      166.      0.177   0.860
16 main_categoryPublishing    -4.96       7.92   -0.626   0.531
17 main_categoryTechnology    13.3     2078.      0.00642 0.995
18 main_categoryTheater       99.3      972.      0.102   0.919
```

```
# A tibble: 1 x 7
  null.deviance df.null logLik   AIC   BIC deviance df.residual
          <dbl>   <int>  <dbl> <dbl> <dbl>    <dbl>       <int>
1        15584.   11467 -0.904  37.8  170.     1.81        11450
```

We included the categorical variable main category to check if it has any influence over the state of the startup. From the table above we can see that the p-values are generally above the significance level of 0.05, so we do not consider main category as a major predictor of success.

Predict Potential Project Success

```
# A tibble: 1 x 2
  state      med_goal
  <fct>         <dbl>
1 successful  -0.0309
```

```
# A tibble: 1 x 2
  state      med_pledged
  <fct>            <dbl>
1 successful     -0.0739
```

```
# A tibble: 1 x 2
  state      med_backers
  <fct>            <dbl>
1 successful     -0.0698
```

```
[1] failed     failed     successful successful failed
Levels: failed successful
```

ANALYSIS

We use this prediction accuracy to predict the project's success. Given a new start-up project, we plot that project acordingly to its attributes, check for the nearest neightbor of the project, and assign that project to the according state. Using this method, the prediction accuracy is 95.23%, which is the highest prediction accuracy we can get.

-discuss outliers and need to shrink data -discuss k-nn test and 99% accuracy result -discuss logistic regression model result and how backers and main_category are not significant predictors -compare knn and glm accuracy (99 vs 99.88) -used glm because more accurate to predict potential project success -explain why we chose the values of new project - explain results of that and how we could use these to our benefit in the future -discuss what we could do differently - data set huge - had to make changes, cannot predict for bigger values after we removed outliers - usd_pledged - used knn instead - etc.

- success rates of big projects vs. smaller ones
- remove outliers for visualization purposes

further analysis: - use interaction between pledged and goal

THINGS TO GO BACK OVER: -organize cleaning data code -do we need to use outlier set for everything or can they be considered -knitting -VISUALIZATIONS???