# What Makes A Successful Project Idea?

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

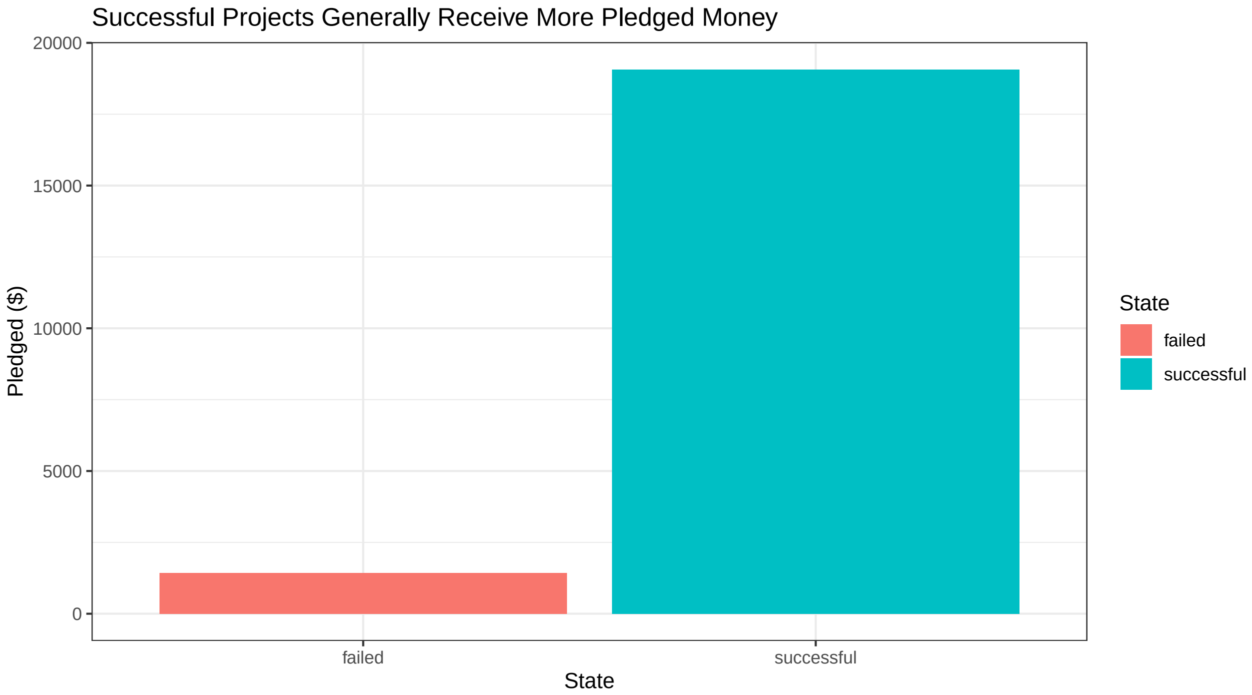
```
# A tibble: 11,468 x 14
      X1     ID name  category main_category currency deadline             goal
   <dbl>  <dbl> <chr> <chr>    <chr>         <chr>    <dttm>              <dbl>
 1     1 1.46e9 Euro~ Documen~ Film & Video  USD      2014-07-01 06:00:00   700
 2     2 2.29e8 Fant~ Printing Crafts        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: 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.

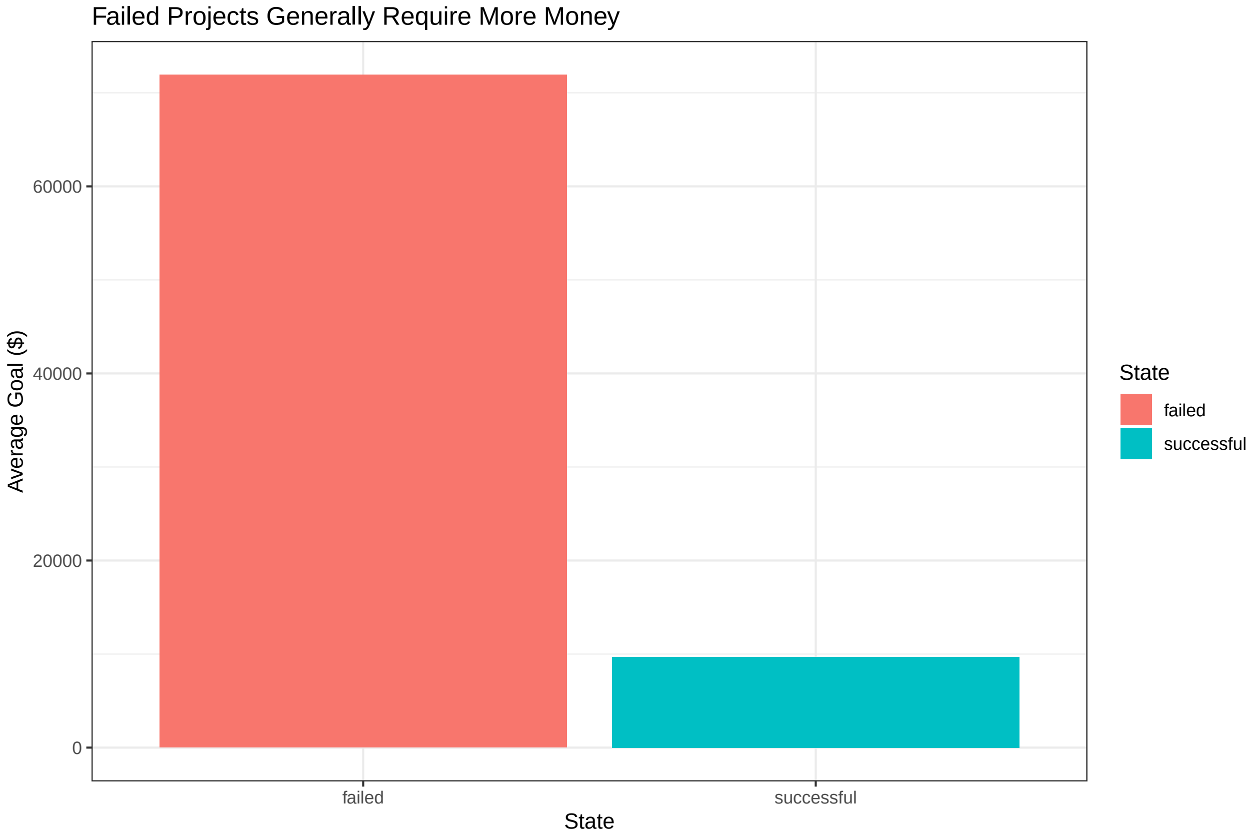
The variable `pledged` tells the amount of money pledged by the crowd, or users of the site.

```
# 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.

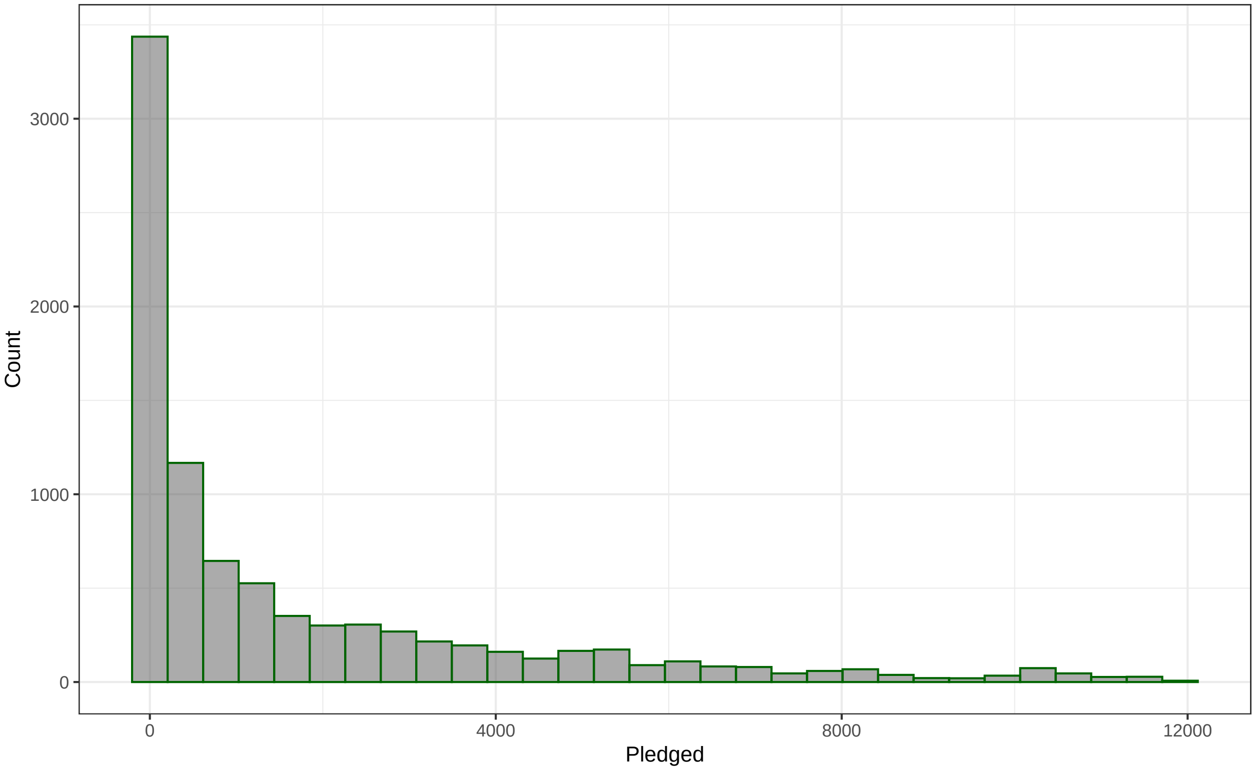
Another variable we are interested in is `goal`. `Goal`gives the amount of money needed by a creator to fund and finish their project.



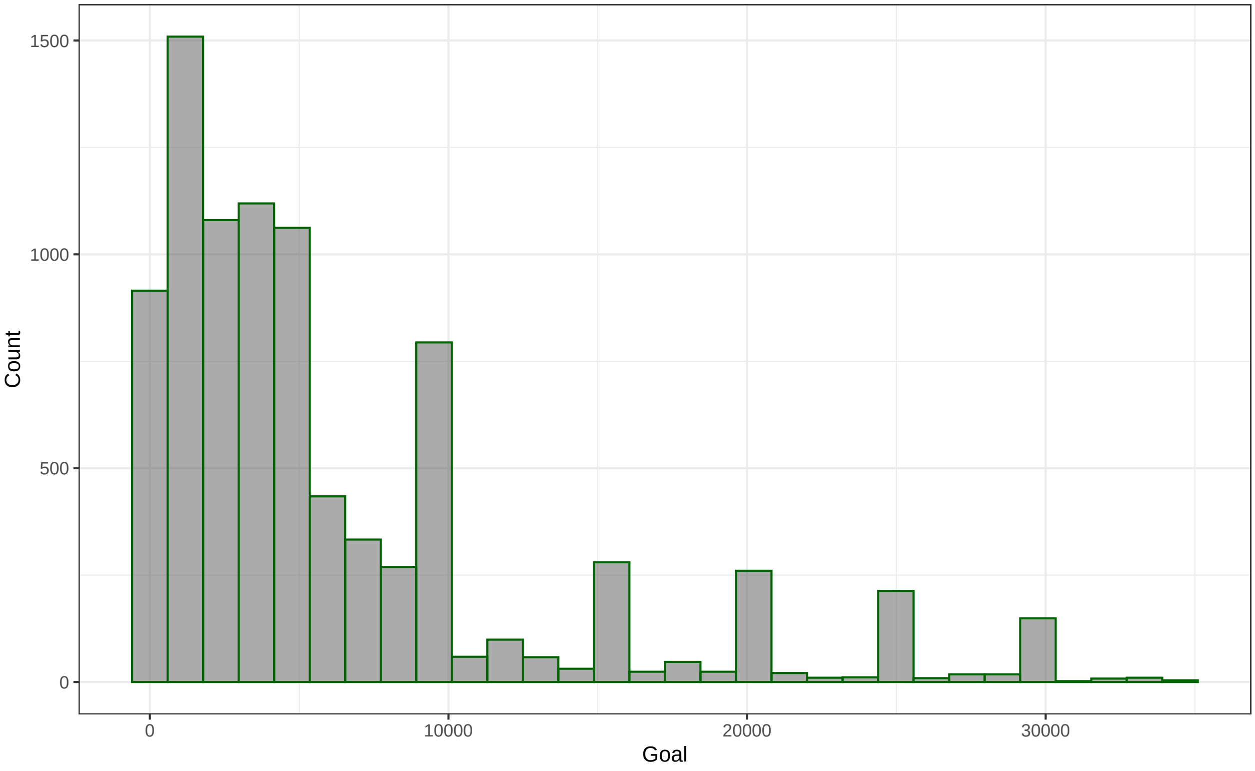
This visualization shows that failed projects generally were much more expensive to create than successful projects. This is very useful in our analysis, since it suggests that there is possibly a threshold to how much money can be required by the creator before the project becomes unreasonable and fails.

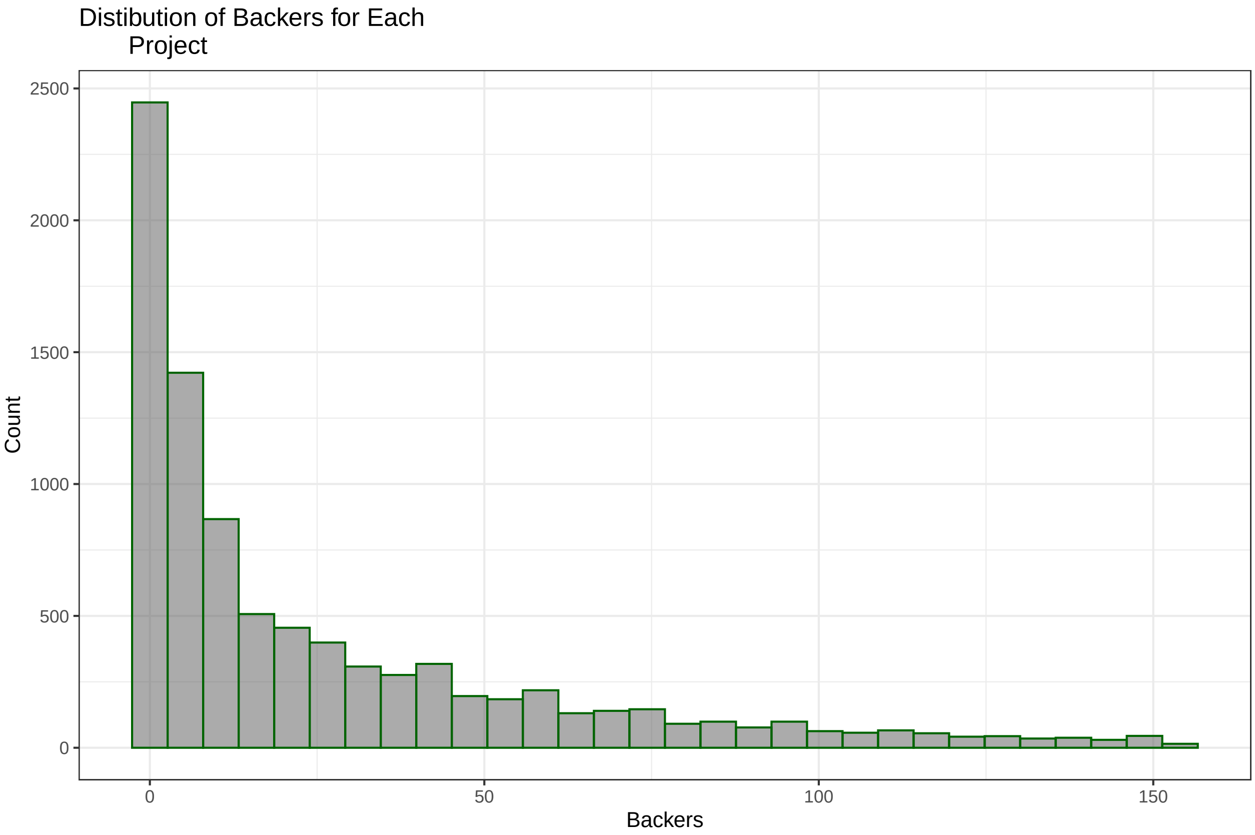
In our analysis it could be interesting to compare the goal of a creator to how much they received from pledged money as well. Looking at these two variables and how they interact could be very telling of the outcome of the project.

Distibution of the Amount of
Money Pledged to Each Project



Distibution of the Goal Amount of
Money Each Project wanted to Raise





By defining a large project with one that has over $4000 pledged to it, the graphs show that big projects are launched less often than smaller projects on kickstarter. It also shows that the pledged amount is typically smaller than the goal amount 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
```

We can see that larger projects tend to have a much higher success rate (85.4%) relative to their smaller counterparts (24.8%).

## Codebook

We can see that there are 11468 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

# Statistical Methods

In considering the state of a project, we hypothesize that goal, pledged, backers, and main category will be the most significant predictors of whether or not a project succeeds.

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.

Finally, we will use a logistic 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 projects.

## Standardize Data

We standardized the variables to improve the accuracy of our K-NN models. K-NN 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 AND RESULTS

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 insignificant values so we filtered the dataset. We are also looking at the 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 orriginate from the same country.

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 visualizations more generic and to get better results. Outliers would have an effect on how the data is displayed.

## Main Category

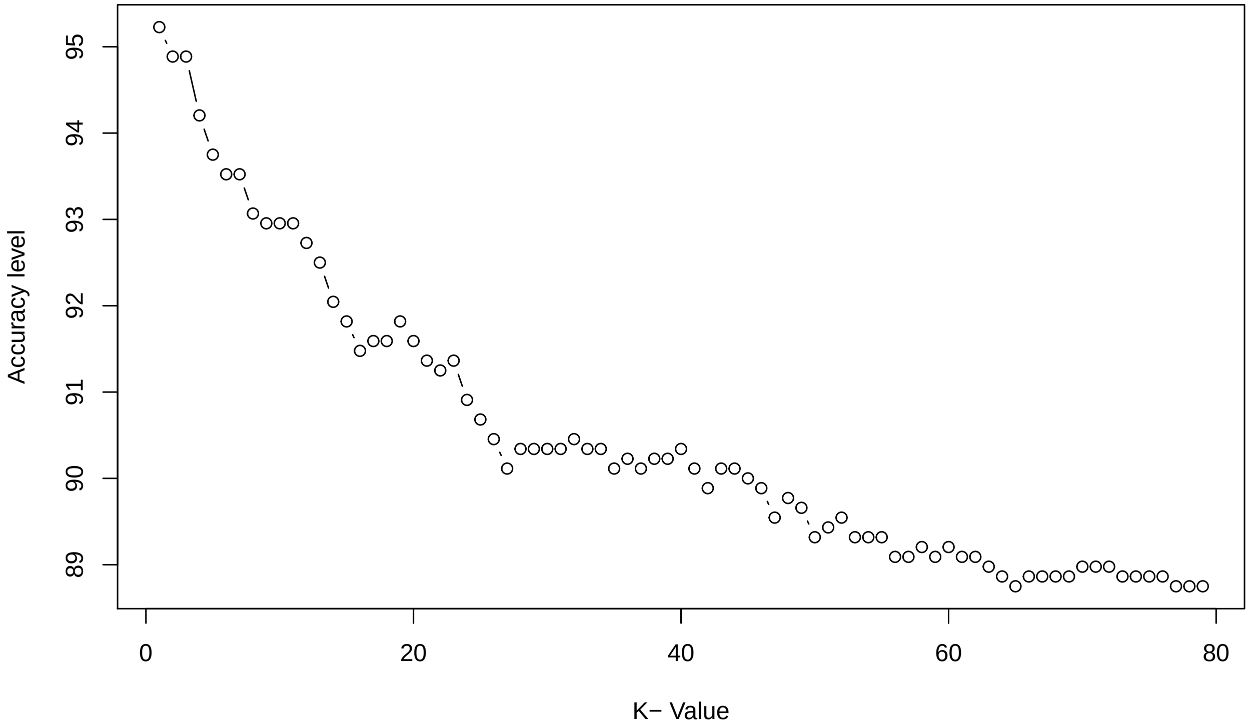
```
# A tibble: 18 x 3
   term                      estimate   p.value
   <chr>                        <dbl>     <dbl>
 1 (Intercept)                10.2    0.184
 2 goal                       -0.846  0.0000740
 3 backers                     0.0597 0.672
 4 pledged                     0.845  0.0000746
 5 main_categoryComics         2.70   0.989
 6 main_categoryCrafts         1.99   0.978
 7 main_categoryDance          3.63   0.990
 8 main_categoryDesign         0.269  0.992
 9 main_categoryFashion        3.83   0.661
10 main_categoryFilm & Video  -3.06   0.714
11 main_categoryFood          -6.85   0.368
12 main_categoryGames         -2.71   0.913
13 main_categoryJournalism    59.1    0.964
14 main_categoryMusic         35.6    0.807
15 main_categoryPhotography   29.4    0.860
16 main_categoryPublishing    -4.96   0.531
17 main_categoryTechnology    13.3    0.995
18 main_categoryTheater       99.3    0.919
```

In order to ensure that, we have the best fit model, we tested to see if it had any effect on the outcome of state. After fitting a linear regression model, it was observed that main category did not have a statistically significant effect, since each category's p-value was greater than our alpha value of 0.05. Therfore, we do not consider the main category variable in further models.

## 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 made the vectors required for training and testing the model. The model is tested on the number of observations that we get from the dataset.

We created two new datasets, kickstart_train and kickstart_test using the three predicting numeric variables (goal, pledged, backers) for the categorical variable state. 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 created a vector of the class labels for the training dataset and called it train_state. We created a vector of true kickstart state in our test dataset and called it true_state.

We fit the k-nearest neighbors model on our raw training dataset. We let k vary from 1 to 79 and calculated the value of k that results in the greatest prediction accuracy in our dataset along with its associated prediction accuracy. We then plotted 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, suggesting that backers, pledged, and goal are effective predictors of state according to the k-NN model.

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.

## Logistic Regression Model

Let $\alpha = 0.05$ :

```
# A tibble: 4 x 3
  term           estimate    p.value
  <chr>             <dbl>      <dbl>
1 (Intercept)   -39628.   0.00000156
2 goal        -1449204.   0.00000155
3 backers            7.61 0.797
4 pledged        53238.   0.00000156
```

The coefficient for pledged means that, holding all other variables constant, for each dollar increase in money pledged, we would expect the log-odds of a start-up being successful to increase by approximately 5.32*10ˆ4.

The coefficient for goal means that, holding all other variables constant, for each dollar increase in the goal amount of money a company would like to receive, we would expect the log-odds of a start-up being successful to decrease by approximately 1.45*10ˆ6.

Since, the p-values for money pledged and goal are 1.56*10ˆ-6 and 1.55*10ˆ-6 respectively, that means both variables are significant in predicting the state of a start-up at the alpha = .05 level.

However, since the p-value for backers is 7.97*10ˆ-1, the variable is not a significant predictor of the state of a start-up at the alpha = .05 level.

It is important to note that these two variables share an opposite relationship with the state variable. Pledged has a positive relationship with state, meaning that with each increase in the amount of dollars pledged to a start-up will increase the probability that it will become successful. Whereas, goal shares a negative relationship with state, meaning that with each increase in the amount of dollars a start-up requires will decrease the probability that it will become successful.

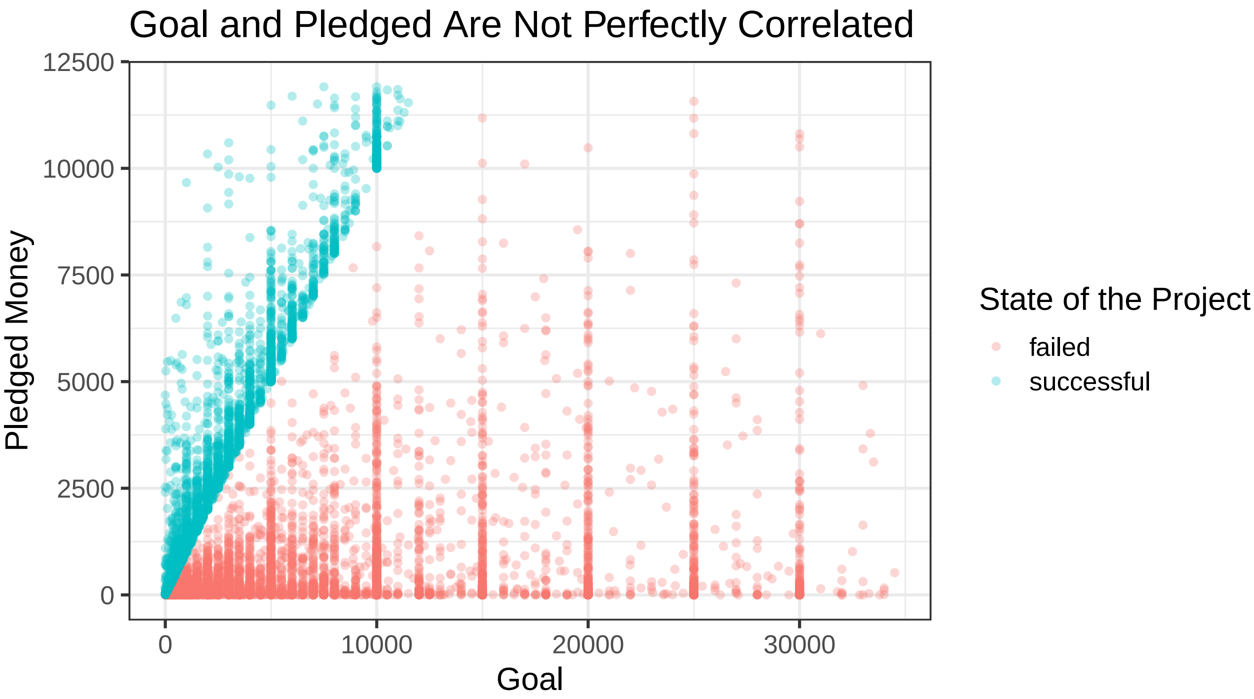
```
# A tibble: 3 x 3
  term         estimate    p.value
  <chr>           <dbl>      <dbl>
1 (Intercept)   -48503. 0.00000339
2 goal        -1773677. 0.00000339
3 pledged        65162. 0.00000339
```

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
```

Using the same variables as our k-NN model, we can see that the accuracy for the regression model predicting state is 100%, which is slightly higher than the accuracy we got from our K-NN model, which had the highest accuracy at 95.23%. This suggests that the regression model is better at predicting state given the same variables.

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



A linear model predicting the amount of money pledged from the goal, shows that goal is not a strong predictor of pledged. Just because a project demands a certain sum of money does not imply that they will receive it, as seen in the visualization above in most failed projects.

## 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
```

We will use these values as a benchmark when testing potential hypothetical projects.

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

We wanted to predict a potential project's success, simulating what it would be like if we were to create our own individual projects.

We proposed five possible projects with different values for goal, pledged, and backers. Our third project used the median values for each of the three variables to see how the average project proposed on kickstarter would play out. We used the median value instead of the mean since our data had outliers present, which would skew the mean value. All the other proposed projects had considerably random values chosen.

We used the k-NN test to predict the state of these new projects, and the median project resulted in success. This means that the average project compared to the 1 nearest neighbor is 95.23 % likely to be successful.

The first two projects had the same random values for goal and pledged, both being lower than the respective median value. However, the backers for both was varied to see if the number of backers would change the outcome with all else constant. The second project had a much larger value for backers, but it still resulted in failure like the first. This is consistent with what we observed in our logistic regression model, which showed that backers does not have a significant effect in predicting the outcome of a project.

Our last two projects had random values for goal and backers, both being greater than the respective median value. However, the pledged for both was different to see if the amount pledged would change the outcome with all else constant. The last project had a lower value for amount pledged and resulted in a failure. The fourth project resulted in success though, which implies that the amount pledged is important to determining the state of the project.

## DISCUSSION

Through our analysis, we were able to determine that the variables `goal` and `pledged` are the most essential factors in deciding the outcome of a project, either successful or failed. This is different than what our hypothesis stated, which predicted that `goal`, `pledged`, `backers`, and `main_category` would all be predictors of the outcome of state.

First, we eliminated `main_category` as a predictor by fitting a linear regression model and observing the p-values of each main category to be insignificant at our alpha level of 0.05.

Next, using our k-NN model, we determined that we could predict with a 95.23 % accuracy the state of a project given only the variables `goal`, `pledged`, and `backers` when the number of nearest neighbors is equal to 1. While 1 seems like a very small number of observations to be compared to, this can be explained due to our high number of observations in general and the trends detected. Given the visualization that plots the relationship between `goal` and `pledged`, we see that there is a very sharp distinction between the projects that have succeeded or failed based on their pledged and goal values. Therefore, a project's nearest neighbor most likely has a high chance of having the same state if pledged and goal values are very similar. We found out through our logistic model next that `backers` is less essential to predicting state.

The logistic model further reinforced the value of the variables `pledged` and `goal` at predicting the state of a project. It revealed the variables pledged and goal to be extremely significant in predicting the state of a given project at the alpha = .05 level. Their p-values being 1.56*10ˆ-6 and 1.55*10ˆ-6 respectively, meaning that these two variables are essential in determining if your project will be a success or a failure, while backers is not. Backers was shown to be an insignificant predictor of state with a p-value of only 7.97 *10ˆ-1 meaning the variable has limited capacity to predict the state of a company. In essence, when beginning a start-up the amount of money pledged to one's startup in relation to its goal amount will be the ultimate determinate of success.

Despite the ineffectiveness of backers as a predictor of state, when using the logistic model to predict the success of a project, we had to consider `backers` still in order to provide an equal comparison between the k-NN model and the logistic model. According to these three variables, the logistic regression model predicts the state of a project with 100 % accuracy, further proving the importance of `goal` and `pledged` as predictors, while also suggesting that `backers` may have some effect even if it may be small.

Finally, we created a model using the k-NN test to determine the success of hypothetical projects. In doing so, we further disproved the value of `backers` as significant in predicting state, since when both pledged and goal were held constant in two different potential projects, a large increase in backers did not cause the state to change to be successful. However, when goal and backers were held constant and even a relatively small increase in pledged occurred between two projects, the state changed from failed to successful.

In retrospect, there are a few things we could have done differently to improve the results of our analysis. First, our original data set included 280,000 plus observations, which caused us problems in trying to perform tests. Although we took a smaller sample of this data, it might have been better to have chosen a dataset that was smaller from the start that would be easier to use in computation. Additionally, the data spans over the years of 2010 to 2016, which may be slightly outdated for predicting a project that would be successful in the present, since it is possible that trends have changed since then. Finally, in predicting the outcome of our hypothetical projects, we decided to use the k-NN model over the logistic regression model even though the logistic model had a higher prediction accuracy. This was due to coding errors in R. When the model tried to predict the outcome of the inputted values, we continuously got values of zero percent success rate even when we knew this was not the case. Thus, using the logistic regression model would have in theory been more accurate, but we were unable to do so. The k-NN model, however, still had a high prediction accuracy, so our predicted project outcomes are still likely valid.

If we wanted to further explore this analysis, we could take into account projects proposed outside of the US and see if our prediction variables still hold the same value in determining state. In this situation, we would need to use the usd_pledged variable instead of the pledged variable in order standardize the currency in US dollars. Additionally, we could make use of the two variables deadline and launched to determine whether or not time elapsed from start to deadline is a determining factor of success. For example, if projects with a short time between launch and deadline tend to have a higher chance of success compared to those with longer elapsed time.