Inital Data Cleaning:

Here we preformed some initil data cleaning to remove columns that are not helpful in our EDA or models. We also added a column called Goal rate which is just the pledged divided by the goal set. We also changed our classification column to be either 0 or 1, 0 if it failed or was canceled and 1 if it was successful or live. We also grouped the countries that had less than 1% of projects to avoid having so many different countries.

EDA:

First let’s explain what each of our columns means before getting into the EDA. Our "main_category" column tells us the type of or category kickstarter project it is. Our “country” column tells us the country that the kickstarter projected was created in. Our “state” tells us whether the kickstarter project failed or was successful. In this section we created a bar graph of projects failed based on their "main_category" and then created another bar graph for projects that succeeded based on their "main_category". We then created a bar graph of projects failed but based on "country" rather than "main_category" and also created a bar graph of projects that succeeded based on their country. However, this did not show us as much as we had hoped, so we tried to analyze further by looking at the success rate based on each category in "main_catergory" and "state" by using a cross tab and then making a stacked bar graph. So we created one stacked bar graph of the cross tab between country and state and another stacked bar graph of the cross tab between main category and state. This showed us more of what we are looking for because the data didn't look as skewed when we had a large majority of projects being from the US or being a Film and Video project. We found that most of the countries had a similar success to fail ratio. We also saw that technology had the highest fail rate in main category and theater and dance had the highest success rate in "main_category".

More Data Cleaning:

We changed our categorical features into numeric representations so that our data will work with our models. We then split our training and testing data 80/20

Logistic Regression Analysis with Goal Rate

Here we performed logistic regression with "goal rate", "country", "main_category", and "backers". We tested 74669 data points and our prediction came out 24.8% of the projects succeeded and 75.2% failed. We then calculated the accuracy of our predicted with the actual classification. We got a 83% accuracy score

Logistic Regression analysis without Goal Rate included

Here we performed Logistic Regression without Goal Rate, because we hypothesize that Goal Rate will be the most influential feature. This was also confirmed in our previous logistic regression section by looking at the coefficients. We tested 74669 data points and our prediction came out 38.8% of the projects succeeded and 61.2% failed. We then calculated the accuracy of our predicted with the actual classification. We got a 78% accuracy score on our logistic regression model when using "country", "main_category" and "backers"

KNN using Goal and Backers:

We performed KNN on our data set using the features "goal rate" and "backers". We used these two features first because they have been our most useful features as shown in our coefficients array in Logistic Regression. We first graphed the relationship between goal rate and backers. Once we ran KNN on our data using the selected features and a K of 5 we got 36.2% classified as successful and 63.8% classified as failed. We then calculated our accuracy score and got 98.8% correctly predicted. We decided to stick with K = 5 because our accuracy was so high. This further confirms that our 2 best features are goal_rate and backers.

KNN using Main_Category and Country:

We performed KNN on our data set using the features "main_category" and "country". We used these two features next because they have are our worst features as shown in our coefficients array in Logistic Regression. We first graphed the relationship between main_category and country. We then ran KNN on our data set 3 different times using K = 3, K = 5, and K = 20. For K = 3 we get an accuracy of 60.3%, for K = 5 we get an accuracy of 60.8% and K = 20 we get an accuracy of 63.1%. Based on the Elbow method, after 3 we do not get a much better score for accuracy, so we decide K = 3 is the best K.

Contributions:

Ryan: EDA, KNN
Steven: EDA, Data Cleaning
Angelica: EDA, Data Cleaning
Jinseok: Data Cleaning, Logistic Regression

```
In [ ]:
```