**Jing Sun (Description of this Project)**

In the beginning, we didsome research on Kickstarter. After understanding the background, we understood better the meaning of each variable in the dataset, which helped us to make better decisions to analyze the data and the direction of the analysis. When we were working on cleaning data, we were able to figure out which variable could be removed from the dataset, and the rest of the variables can still explain the data information efficiently. After checking on the missing value, we decided to drop the variables location\_country and name. Then we started to manipulate the data, which included converting ‘'True', 'False' values in the data frame as 1 for ‘True’ and 0 for ‘False’, making the columns with date and time into a DateTime, getting dummies for the categorical variables. Then we started doing data summarization and data visualization. We separated the analysis for this part into two groups. We did data summarization and visualization on the whole Kickstarter dataset and then picked the top 15 categories to analyze. We followed the same rules for the rest of the analysis, which are regression analysis and prediction analysis. First, we picked the whole dataset to do regression analysis, and then we picked the top 15 categories to do the regression analysis. We did the same thing on the prediction analysis.

**Rachel Fluegel (Findings from Regression Analysis):**

While doing the regression analysis there were a few interesting findings. I did the regression analysis on some of the key variables that I thought would be the main effect on the success of any Kickstarter. The variables that I thought would be most effective were goal\_USD, usd\_pledged, staff\_pcik, year, days\_to\_deadline, blurb\_length, and location\_type. I made dummy variables for both the year and the location\_type to see the effect of each year and the different location types within the data. The first regression analysis was all these effects on the whole dataset. With the regression analysis, you can see that when staff\_pick would increase to 1 it would increase the chance of success my 0.2857. There was also an interesting trend with the years’ effects on success. In earlier years (2009-2013) and 2019, there is a positive effect on success. In the middle years (2014-2018) there was a negative effect on success. We can see that there was more success in the earlier years and last year than all of the other years. This could be due to Kickstarter falling off a bit in 2014 and making a resurgence in 2019.

I also did a regression analysis of the top 15 categories found in the dataset. With this analysis, staff\_pick was still positive but had a higher coefficient. There was a slight difference in the trend of the years’ effect on success when just looking at the top 15 categories. There are still positive coefficients in the early years (2009-2013) and the year 2018. The negative coefficients are in the years (2014-2017) and 2019. There is a slight difference in the years between the whole data set and the top 15 categories in the years 2018 and 2019. The year 2018 also now has not statistically significant since the p-value is 0.487 which is greater than 0.01. Even though there is a change in 2018, with the top 15 categories there is no sign of the year to the success of Kickstarters.

**Xiaojing Ge: (Findings from prediction analysis)**

In order to keep consistent with regression analysis, we chose the same variables which included goal\_USD, usd\_pledged, staff\_pcik, year, days\_to\_deadline, blurb\_length, and location\_type. Based on the property of this project itself and outcome variable, we decided to choose 6 different classifications and supervised machine learning models which were logistic regression model, k-NN classification model, Naive Bayes classification model, Decision Tree model, and Random Forests model. By evaluating the classification report of each model, both the Decision Tree model and Random Forests model showed good accuracy, precision, recall, and F1 score results. k-NN classification model for k = 5 also showed relatively good classification results. However, the computing time for k-NN classification was much longer and k value tuning would take even more time. Naive Bayes classification did not show good classification results due to the limitation of variable independence. Logistic regression also did not present good classification results. Therefore, the Decision tree and Random Forests model were chosen to do the prediction and classification analysis for this project.

**Zach Brown: (Summary of Key Findings)**

Our findings from the initial data exploration indicated that factors impacting project success varied wildly. The most glaring of these was that the most popular categories had a wide fluctuation of successful projects between categories. From here, it’s easy to see that the popularity of a project doesn’t mean much for success. It’s when we started to break down the different variables that things got interesting. The most consistent variable correlating to success was staff\_picks, which isn’t overly surprising to see, given that a staff pick will have much more exposure than normal projects.

Our regression analysis for all projects supported staff\_picks being a strong factor in project success. We also found that usd\_pledged showed a positive effect on project success. On the other end of things, we see that goal\_USD has a negative effect on project success. The years in which the projects took place to have assorted effects. We see a positive effect from 2009 to 2013 and 2019, but a negative effect from 2014 to 2018. Looking at the top 15 categories, we see similar stories with these variables, but with different significance and effects.