1. Landscape of Funding for Kickstarter Campaigns
2. Authors
3. Summary
   1. Kickstarter as a website
   2. Our goal for this project
      1. Understanding the funding landscape of kickstarter
   3. Dataset
      1. Mix from kaggle and scrapped json data
      2. General description of dataset
   4. Brief outline/overview of methods
4. Methods
   1. Data tidying / reformatting
      1. Gathered original dataset from Kaggle. Contained 378661 observations.
      2. Original data had metadata regarding the projects such as number of backers, pledged, success of fail, but lacked information regarding the summaries of each project.
      3. Found JSON representations of the same data from Kaggle with summary data and joined them accordingly.
      4. Feature engineering
         1. Added a days live column (created date - status changed data)
         2. Separated day, month and year into three columns
      5. Splitting of data for exploration and analysis
         1. Resplitting analysis for k fold cross validation
   2. Summary Analysis
      1. Bigram Importance
         1. Explored relevance of words and bigrams by country and main category.
         2. Used TF-IDF of stemmed bigrams to understand differences between successful and failed campaigns.
      2. Sentiment Analysis
         1. Explored how sentiment of english summaries changed over the last decade.
         2. Used afin dataset to attach numerical sentiment values to words. For Bigrams, summed these sentiment values together and took the average of each year and plotted.
   3. Modeling
      1. Linear regression
      2. Logistic regression
5. Results
   1. Exploratory results
      1. General overview of dataset
         1. Anant’s exploratory analysis
      2. Categories and funding
      3. Funding landscape across time
         1. Generally didn’t seem to be seasonal
         2. Some sort of rise in late 2014 - 2015
      4. Summary Analysis Results
         1. Bigram Importance
            1. Found that judging relevance by counts of each word led to some less important words overshadowing words that might show heterogeneity between success and failed campaigns. One word exploration did not expose heterogeneity because less important words overpowered more important words and did not give insight.
            2. Exploring by country showed that words associated with nationality were extremely important, but did not help to differentiate successful and failed campaigns
            3. A combination of looking at categories, TF-IDF and bigrams showed bigrams that differentiate successful and failed campaigns for each category.
         2. Sentiment Analysis
            1. Saw that many campaigns did not have very high or low sentiment scores, possibly due to the fact that words with associated sentiment values were not commonly used.
            2. Saw sentiment overall stayed the same for the last decade.
      5. Model results
         1. Linear
            1. Model chosen
            2. Residuals
            3. RMSE
         2. Logistic
            1. Model chosen
            2. Residuals/error
            3. Accuracy
6. Discussion
   1. What we learned about funding landscape
   2. How effectively can a projects funding be predicted
7. Statement of Contributions
8. References
9. Appendix