Analyzing the Successfulness of Crowdfunding Projects on Kickstarter

Authors: Tianrui Wang & Zi-Qi Liu
Instructor: Prof. David Belanger

STEVENS INSTITUTE of TECHNOLOGY — THE INNOVATION UNIVERSITY®
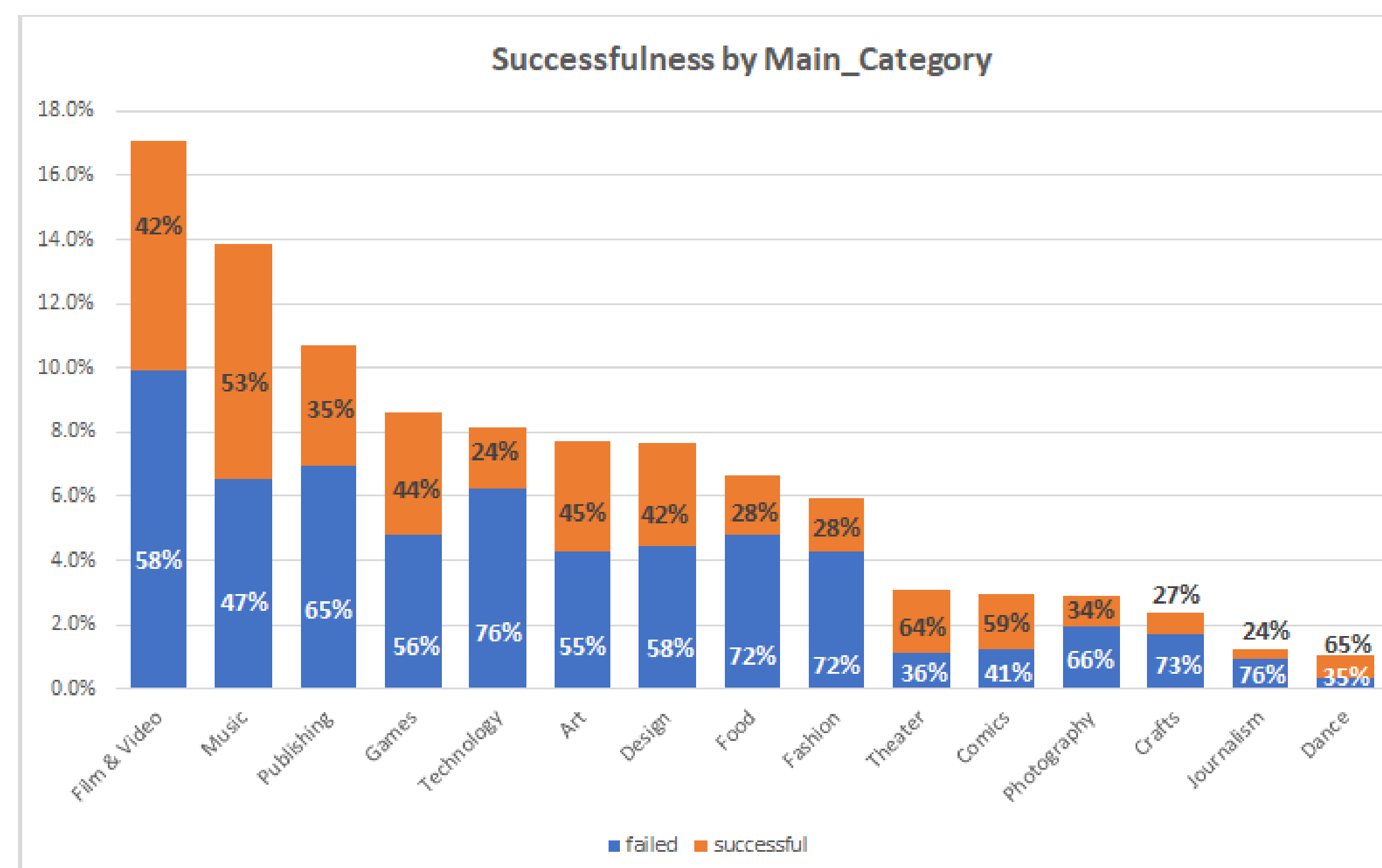
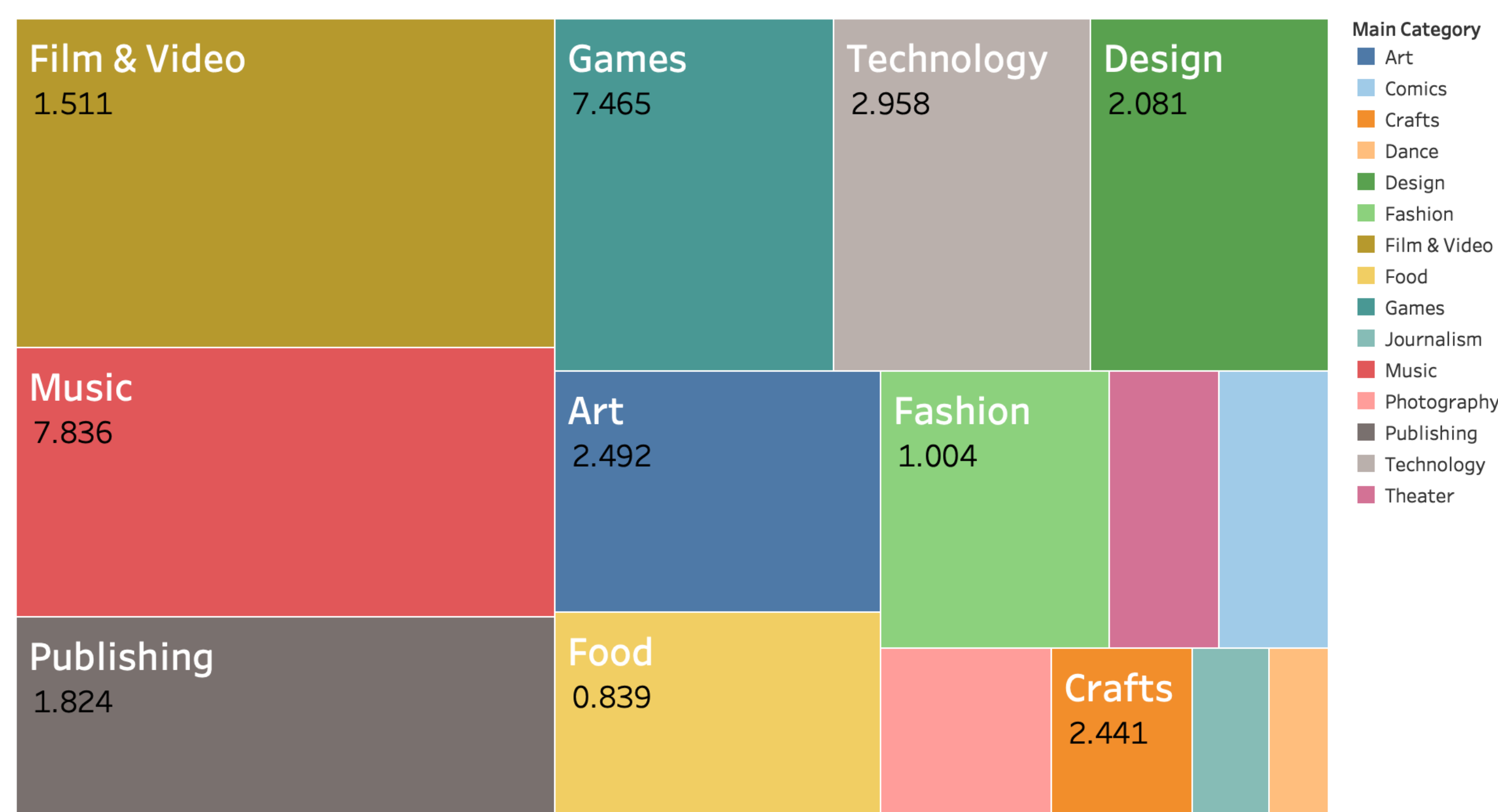Business Intelligence & Analytics

Introduction

- Kickstarter is the world's largest funding platform for creative projects, it works on an all-or-nothing basis i.e if a project doesn't meet its goal by the deadline, the project owner gets nothing.
- Our project digs into which specific characters lead to a successful fundraising for a project, hoping to provide a reference for the Kickstarter community.
- **Data Source:** https://www.kaggle.com/kemical/kickstarter-projects#ks-projects-201801.csv
- **Keywords:** Data Mining, Clustering, Classification, R, Tableau

Project Flow

Research Understanding → Data Understanding (• EDA • K-Means Clustering) → Data Processing (• Standardization • Dummy Variables for categorical data) → Modeling & Validation (• LDA • Random Forest) → Results & Conclusions (• Word Cloud Visualization)

Exploratory Data Analysis

The projects are divided into 15 main categories and 159 sub-categories. Among all the variables, pledged, goal, backers (supporters) and days of duration are the most important numerical variables.

Successfulness by Main_Category

Pledged/Goal Ratio By Size of Main_Category

Main Category	value
Film & Video	1.511
Games	7.465
Technology	2.958
Design	2.081
Music	7.836
Art	2.492
Fashion	1.004
Publishing	1.824
Food	0.839
Crafts	2.441

Main Category and average of Pledged_Goal_Ratio. Color shows details about Main Category. Size shows sum of Number of Records. The marks are labeled by Main Category and average of Pledged_Goal_Ratio.
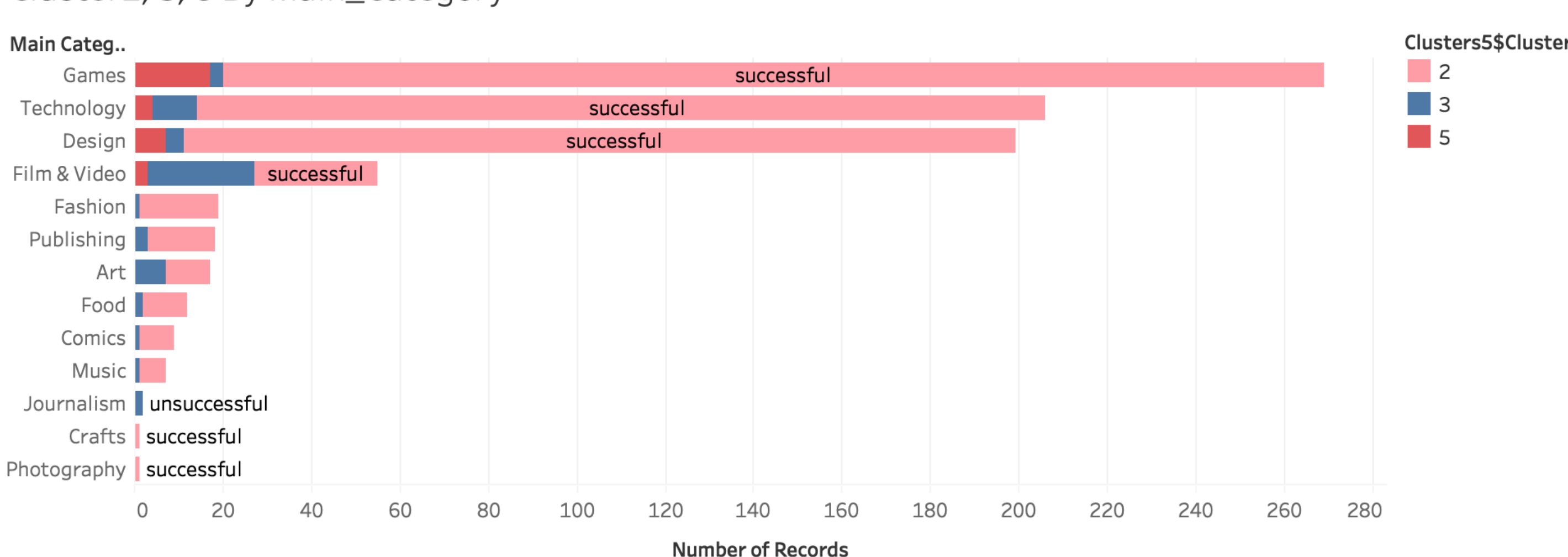
Cluster Analysis: K - Means (k = 5)

While Cluster#1 & #4 represented the crowded majority, Cluster#2,#3, #5 was perfectly separated.

Successfulness By Cluster

state_cal	Clusters5$Cluster				
	1	2	3	4	5
successful	15,402 21.53%	707 97.38%		117,816 38.46%	31 100.00%
unsuccessful	56,123 78.47%	19 2.62%	58 100.00%	188,505 61.54%	

% of Total Number of R.. 2.62% — 100.00%

Cluster2, 3, 5 By Main_Category

Sum of Number of Records for each Main Category. Color shows details about Clusters5$Cluster. The marks are labeled by state_cal. The view is filtered on Clusters5$Cluster, which keeps 2, 3 and 5.
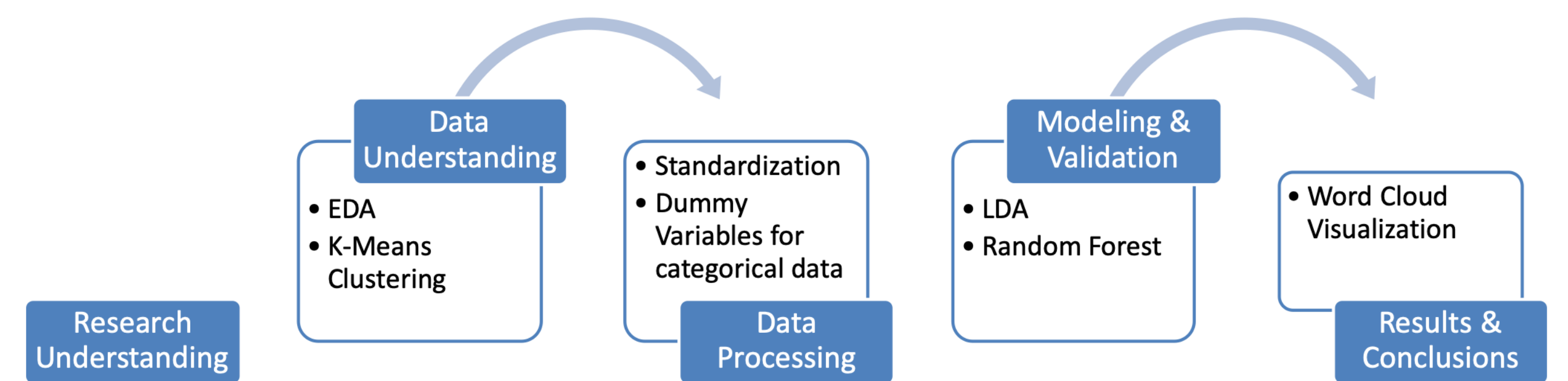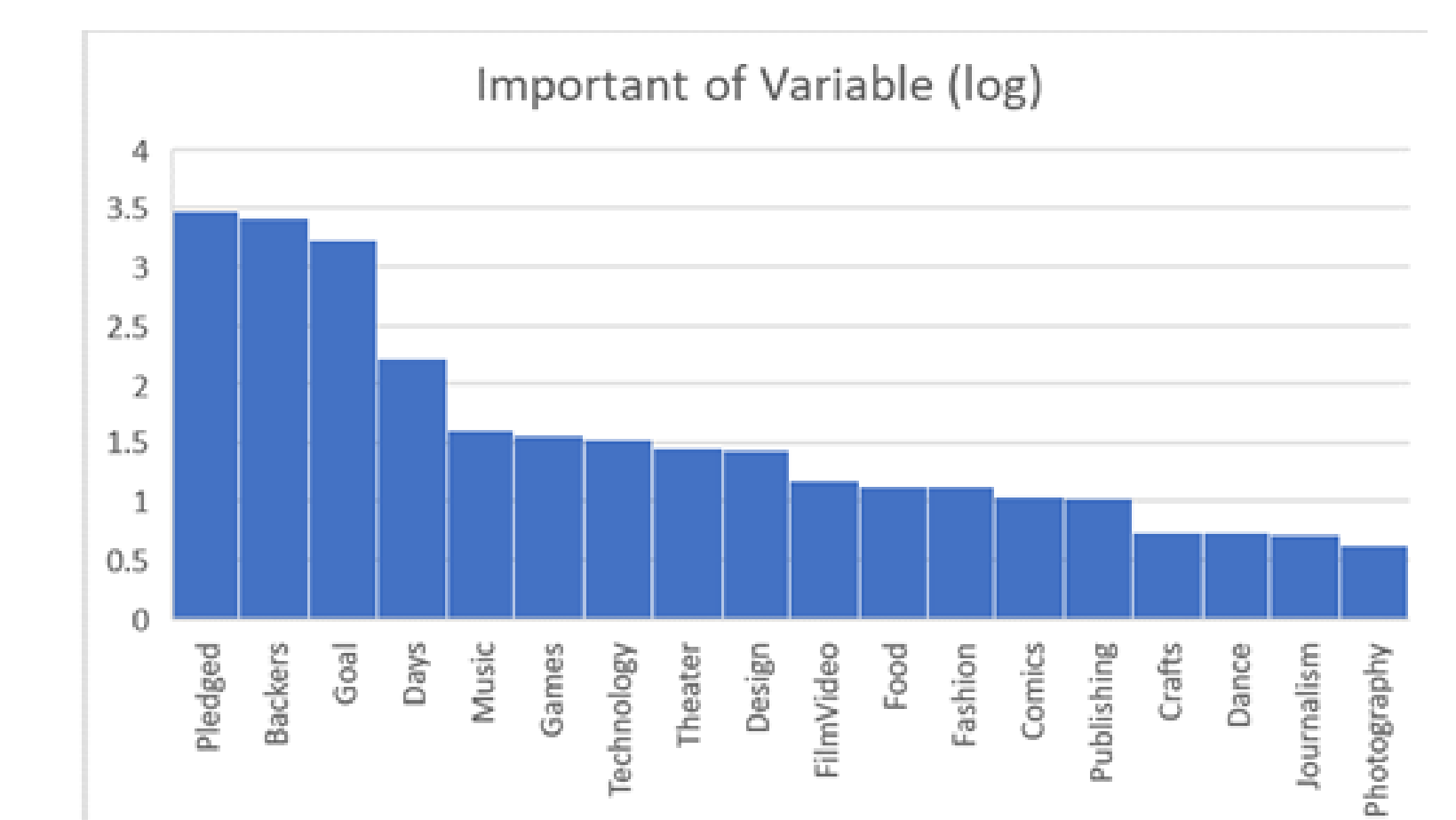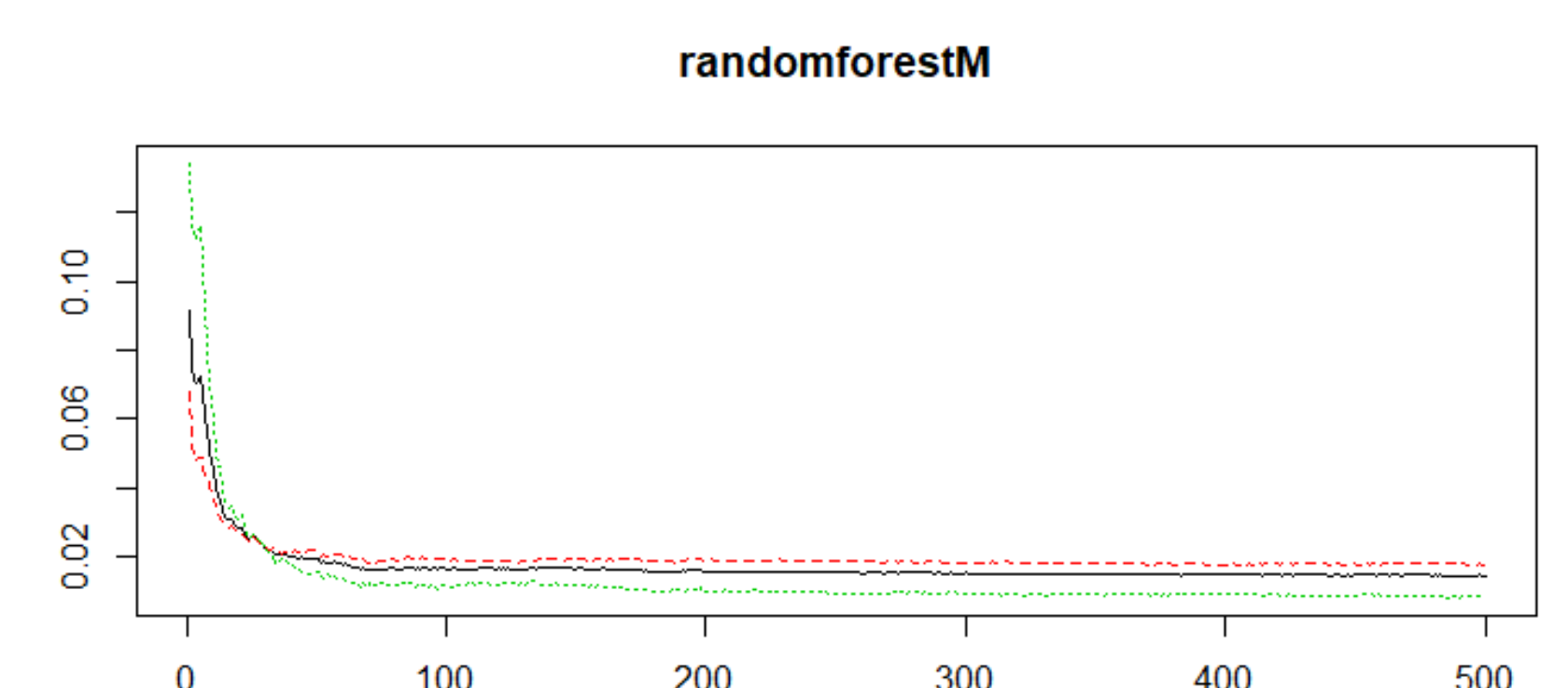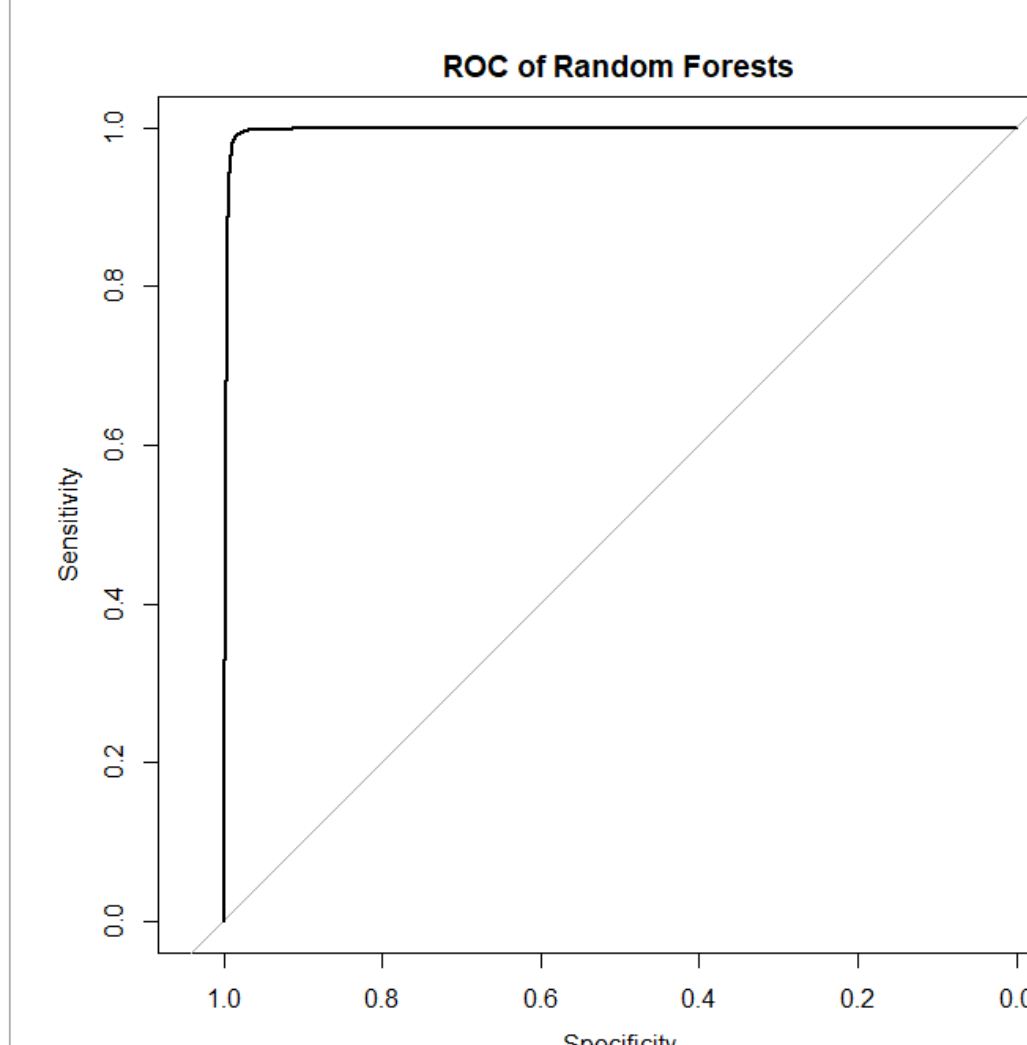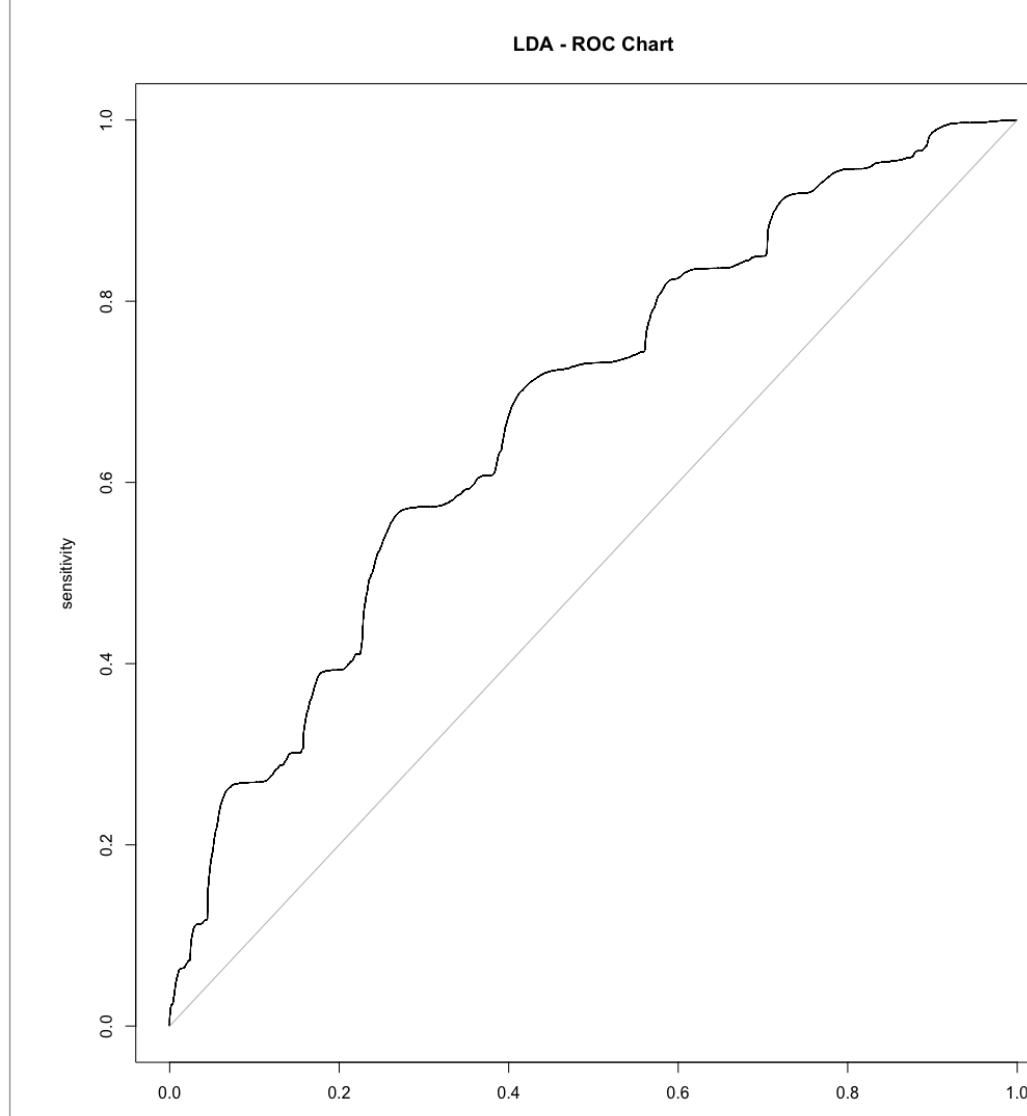
Results

LDA

The result of LDA yields an accuracy of 66.25%. Among all other variables, the coefficient of Technology, Dance, Theater & Journalism ranked top.

Random Forest

As the amount of trees increased, the error of the model decreased and stabilized at 0.02%. The most important variables are Pledged, Backers, Goal, and Days. Moreover, some categories have larger impact than others. Accuracy = 98.5%

LDA - ROC Chart

ROC of Random Forests

randomforestM

Important of Variable (log)

Word Cloud For Top 6 Categories (from the result of Random Forest)
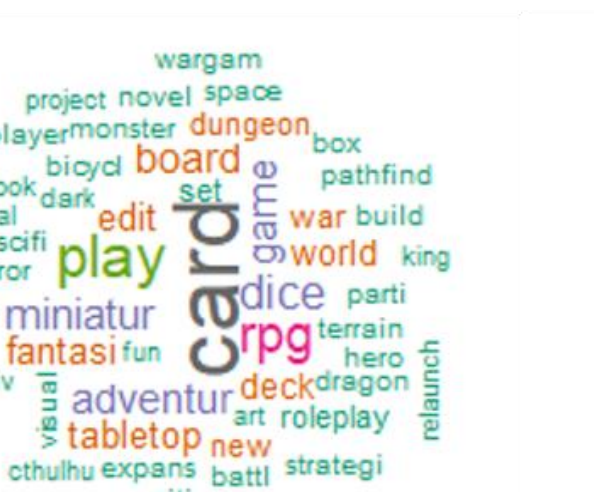

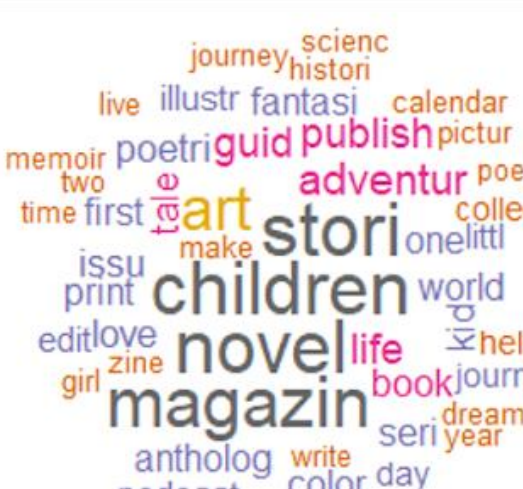Music — Film & Video — Games — Publishing — Art — Design

Conclusion & Future Work

- **The result for EDA, Cluster Analysis, LDA and Random Forest are similar.** They revealed the same group of categories that attracted a majority of the backers in the Kickstarter Community.
- **Games, Technology, Design, Film & Video** are the majority industries among the most mature projects. These industries have higher barrier-to-entry and require more technology inputs. On one hand, this resulted in many well-prepared projects and their successful fundraising; on the other hand, immature project also exists, which could be a reason balancing out the successfulness in such categories.
- **The World Cloud** helps summarized some featured keywords for the top 6 categories, implying majority of the topics within each category. It could be developed to a future research in NLP.
- **Fun Fact:** Cluster#2 & #5 are mainly made up of "outliers" with high goal & pledged. These data are in fact our focus – the ones that "ought to be successful".