# Deep dive into ANOVA(Analysis of Variance) Test

**Van Nguyen**
Department of Computer Science
Boise State University
vannguyen599@u.boisestate.edu

## 1 Introduction

Kickstarter is an online platform that enables creators to raise funds for their innovative projects from a community of supporters. The platform has become significantly popular among other crowdfunding platforms due to their user-friendly interface and dynamic ecosystem that fosters diverse creative campaigns. As someone who has backed several Kickstarter campaigns and found the experience satisfying, I at one point have considered launching some experimental project related to keyboards on the platform. Like other creators, the ultimate question is how to ensure their projects to be funded successfully. Given this question, we need to engage in a study to understand what factors can influence the success of a project. The objective for this course project is to analyze all Kickstarter project data to identify patterns in order to predict successful funding campaign. The motivation is to employ data science techniques to analyze and uncover actionable insights that can inform future crowdfunding strategies.

## 2 Methodology

The datasets used will be generated by *https://webrobots.io/kickstarter-datasets/* which a platform where they have a regular scraper robot that crawls all Kickstarter projects every once a month. we used the datasets through 2025-10-13, which have 186,372 data points and 46 columns

### 2.1 Data Exploration

Kickstarter does not provide any public API to retrieve data other then *https://status.kickstarter.com/api/v2/summary.json* which is used to look up the status of the platform. As a result, it is necessary to explore and investigate existing datasets from other platforms such as **Kaggle** to determine whether they are relevant or suitable to use. The selected dataset should be up-to-date and include all key factors required for an examination of funding trends and project success. we have attempted to scrape the data directly from the website but was not able to get all the sufficient data so we have decided to go with *https://webrobots.io/kickstarter-datasets/*. However to make sure that data is relevant and webrobots does not just patch missing data with wrong information, we have matched data we scraped with what currently provided by webrobots on 5-6 projects. **sample.json** is the data that we have pulled down to check for webrobots's validity.

Other than that we need to go through the dataset in order to make sure that we only select our features for data training. After consideration, we decided to go with the below list as features:

- The campaign's financial goal in USD

- The time taken to prepare the campaign (from the creation date to the launch date)

- The duration of the campaign (from the launch date to the deadline)

- The geographical location of the campaign (whether it is based in the US or not)

- The category under which the campaign falls

We will need to do some data preprocessing where we dropped unused columns and check for duplicates or invalidate data points. Other than that for any timestamp we want to convert them to datetime object. In order to accommodate the above feature list, we need to create extra columns that suit our purpose namely *main_category*, *sub_category*, *campaign_goal_USD*, *campaign_location_US*, *campaign_timeline*, *prep_timeline*.

### 2.2 Data Analysis

We will use the preprocess dataset to analyze the trends evaluate how they affect the success of a

campaign. First we need to take a look at the success rate of all the campaigns in Figure 1. We can see that the success rate is roughly 60% while the failed rate is around 40% which is a not a significantly big gap. Next, we take a look at the
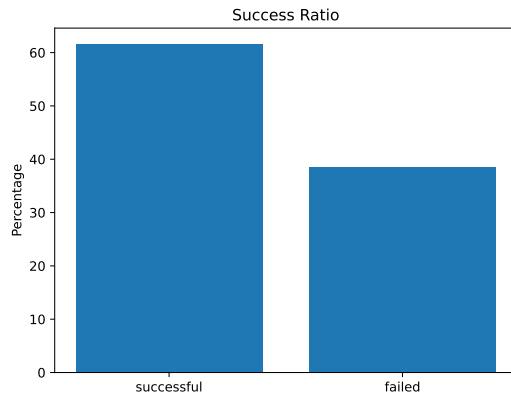


Figure 1: Success Ratio

relationship between the campaign's financial goal and the success rate of it using box plot in Figure 2. Visually looking at the box plot graph, we can see that failed projects shows a massive number of outliers with very high goals up to $140,000,000. The primary takeaway from this plot is the significant difference in the magnitude and frequency of high-goal outliers between failed and successful projects. Projects with extremely large goals are overwhelmingly found in the failed group. This strong concentration of high-goal outliers in the failed state suggests that setting a very high financial goal might be strongly correlated with project failure. Next, we take a look at the relationship
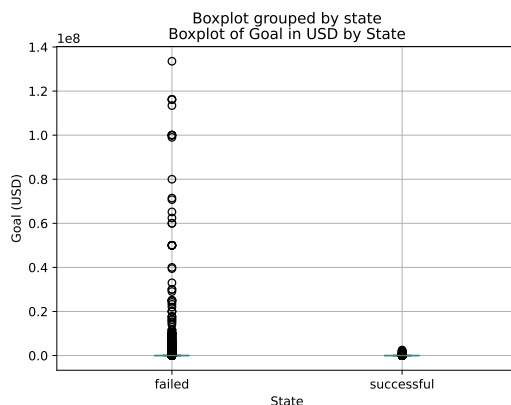


Figure 2: Goal in USD

between the campaign's geographical location and the success rate of it using Figure 3, and Figure 4. By observation, the US accounts for the majority of launched campaigns on KickStarted, more

than 60% with over 66% of those are successful projects. The second country that accounts for the most projects launched is Great Britain, even though they are the second one but they only accounts for around 10% with around 70% of those are successful projects. The gap between the number of launched products between US and Great Britain is significantly large.
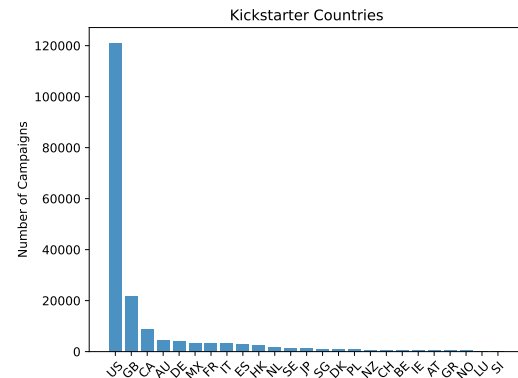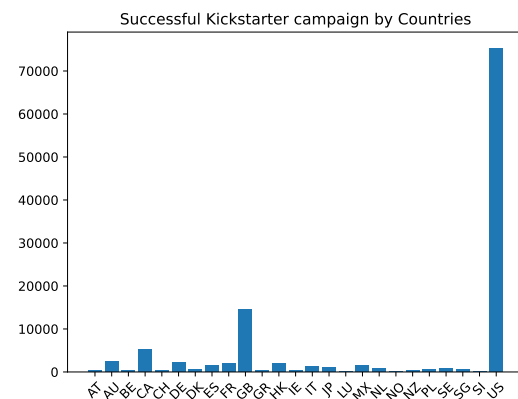


Figure 3: Campaign's country



Figure 4: Successful rate by country

Next, we will observe the relationship between the campaign's duration and the rate of success of those projects in Figure 5. The majority of campaigns fall within the first three duration bins (0-60 days), with very few campaigns lasting longer than 60 days. The 20-40 day bin is overwhelmingly the most common campaign timeline, accounting for over 100,000 projects. Despite having a moderate success rate, the 20-40 day bin also has the highest absolute count of failed campaigns ($\approx 70,000$) due to the sheer volume of projects launched in this window. The data suggests that a slightly longer 40-60 day timeline is associated with the highest likelihood of success Other than that, looking at the preparation in time Figure 6 we can see that
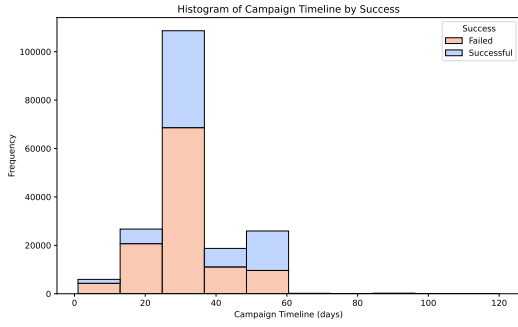
Figure 5: Histogram of Campaign Timeline by Success

the distribution is heavily skewed, with the vast majority of all projects falling into the first bin. The shortest preparation timeline is by far the most popular choice for creators, reflecting a common behavior to launch soon after starting the project creation process. Lastly, we will analyze the rela-
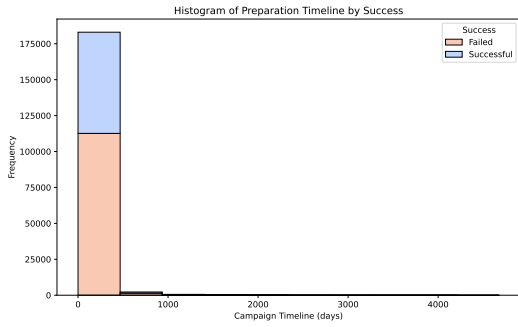


Figure 6: Histogram of Preparation Timeline by Success

tionship between category and the rate of success Figure 7. The chart demonstrates a clear concentration of successful campaigns in the Arts, Media, and Entertainment sectors, with Music and Film & Video leading by a significant margin.
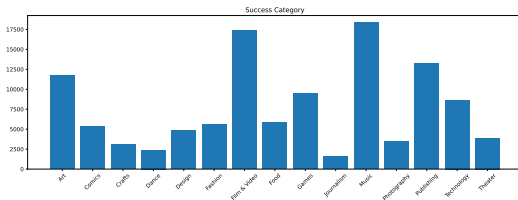


Figure 7: Success Category

## 3 Experiments

In this section, we will be using the preprocessed data and then we will train them using 3 different model namely Logistic Regression, Decision Tree, and Random Forest to select the best model that yields that best performance. All three models are

typically evaluated using the same common classification performance metrics, which are designed to compare their predictions against the true labels. We will initially train the models using their default parameters to evaluate their performance, and then proceed with hyperparameter tuning on the best-performing model. The table 1 summarized all the training metrics after model training process. The table shows that Logistic Regression and Random Forest have only slight differences across all evaluated metrics. However, Random Forest performs slightly better in both Accuracy and ROC. In order

| Model type | accuracy | precision | recall | f1 | roc |
|---|---|---|---|---|---|
| Logistic | 0.690 | 0.701 | 0.866 | 0.774 | 0.712 |
| Decision Tree | 0.643 | 0.717 | 0.693 | 0.705 | 0.631 |
| Random Forest | 0.695 | 0.734 | 0.790 | 0.761 | 0.730 |

Table 1: Summary of chosen model's performance before tuning

to see it better, we also have a ROC curve of all the three models in Figure 8. From the graph we can see that Random Forest model has the highest AUC at 0.734, indicating it is the overall best classifier among the three, as it has the best ability to discriminate between the positive and negative classes
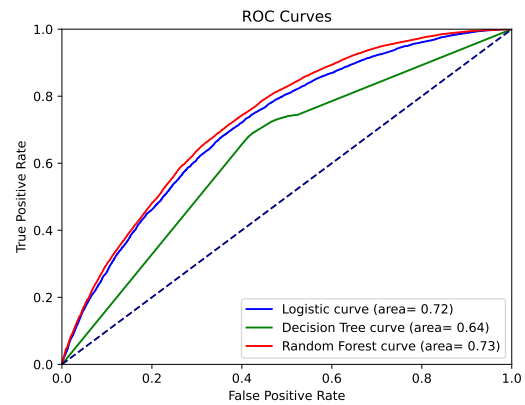


Figure 8: ROC Curves

### 3.1 Hyperparam tuning

We opted to fine-tune the Random Forest parameters using GridSearchCV due to its superior overall performance. Our goal was to reduce false positives and enhance precision. After param tuning,

the best parameter that we got is criterion: entrophy and n_estimator: 500 which yields the following improvement in table 2

| accuracy | precision | recall | f1 | roc |
|----------|-----------|--------|----|-----|
| 0.7007 | 0.7379 | 0.7968 | 0.76626 | 0.735 |

Table 2: Summary of chosen model's performance of Random Forest after tuning

Commonly used All of the steps used in this paper used the guidance from (Thevapalan, 2024) post-hoc tests include Tukey's Honestly Significant Difference (HSD) (Howell, 2010) and the Bonferroni Correction (Thevapalan, 2024). Commonly used post-hoc tests include Tukey's Honestly Significant Difference (HSD) (Armstrong, 2014) and the Bonferroni Correction (Thevapalan, 2024).

## 4  Conclusion

In this paper, we explored the fundamental concepts and applications of the Analysis of Variance (ANOVA) test especially one-way ANOVA. ANOVA is a powerful statistical tool used to determine whether there are significant differences between the means of multiple groups. We discussed the core concepts, including getting all the assumptions required for valid results, and interpreted the final result( including whether or not the null hypothesis is accepted). Additionally, we emphasized the importance of post-hoc tests, such as Tukey's HSD and the Bonferroni Correction, in identifying specific group differences when the ANOVA results are significant. Overall, ANOVA remains an essential technique in experimental design and data analysis, providing researchers with a robust method for comparing group means and uncovering insights in diverse fields such as education, agriculture, and healthcare. By understanding its principles and limitations, practitioners can effectively apply ANOVA to real-world problems and make data-driven decisions.

## References

Robert A Armstrong. 2014. When to use the bonferroni correction. *Ophthalmic and Physiological Optics*, 34(5):502–508.

D. C. Howell. 2010. *Statistical Methods for Psychology*, 7th edition. Wadsworth, Cengage Learning.

Arunn Thevapalan. 2024. Anova test: Definition, types, examples, and assumptions. Accessed: 2025-10-27.