



Final report

*Predicting the Success of a Kickstarter project*

March Continuous DA

Participants: Grigoria Sourri , Ilyes Azzaz, Nizar Marzougui

Team Mentor: Sarah Le Net

September 2023

# An Introduction to Crowdfunding and the Kickstarter Funding Platform

Crowdfunding has emerged as a transformative method for funding projects, particularly in recent years, thanks to technological advancements such as the internet and online payment systems. Instead of relying on traditional lenders like banks or venture capitalists, entrepreneurs can now seek support from a large number of individuals, securing funding through various means, including the promise of future rewards or, in some cases, interest payments.

Among the various crowdfunding platforms, Kickstarter stands out as one of the most prominent and widely utilized. It serves as an online platform where individuals with diverse project ideas, spanning from manufacturing drones to creating documentaries, can solicit financial backing from the general public. As of February 2023, Kickstarter has received US$7 billion in pledges, from 21.7 million backers to fund 233,626 projects. Notably, Kickstarter has provided fertile ground for countless innovative ideas to materialize into reality.

The funding process on Kickstarter typically adheres to a structured procedure. Project creators begin by registering on the platform, where they personalize their profiles and provide pertinent personal information. Subsequently, they craft project pages that include detailed descriptions, visuals, enticing reward systems, and, in some cases, video advertisements. To initiate funding, creators set specific monetary goals and timelines for their projects, making them accessible to the public. At the end of the funding period, the project's fate hinges on whether the funding goal has been met. If the target amount is attained, the project is declared successful, and the creator receives the crowdfunded funds. Conversely, if the funding goal remains unmet, the project is deemed unsuccessful, and the creator does not receive any funds.

# Motivation

Our research is driven by a desire to assist Kickstarter creators in gaining deeper insights into their projects' success dynamics. We aim to empower creators by providing a clear understanding of the crucial factors influencing success. Through predictive models, creators can gauge their projects' chances of success and implement informed strategies for goal attainment.

# Our Aim

As we have already mentioned our main aim is to empower creators with predictive models to assess project success likelihood and make targeted improvements for funding goal achievement.

Our research addresses two key questions:

* Factors for Kickstarter Success: We explore elements contributing to Kickstarter project success, uncovering significant variables.
* Variable Impact on Success: We quantify how project features influence success, identifying critical factors for campaign planning.

# Data Retrieval Methodology

We accessed Kickstarter data from April 2009 to August 2023 using Web Robots (webrobots.io). This involved crafting appropriate URL links to access the desired CSV files containing project details. We utilized Python libraries like Pandas for data handling, requests for file downloads, and zipfile for working with ZIP archives. These steps allowed us to create a unified dataset containing Kickstarter data from 2009 onwards.To create a unified dataset for our analysis, we combined all individual DataFrames into a single DataFrame named "df." The resulting data frame contains 2166290 rows x 39 columns.

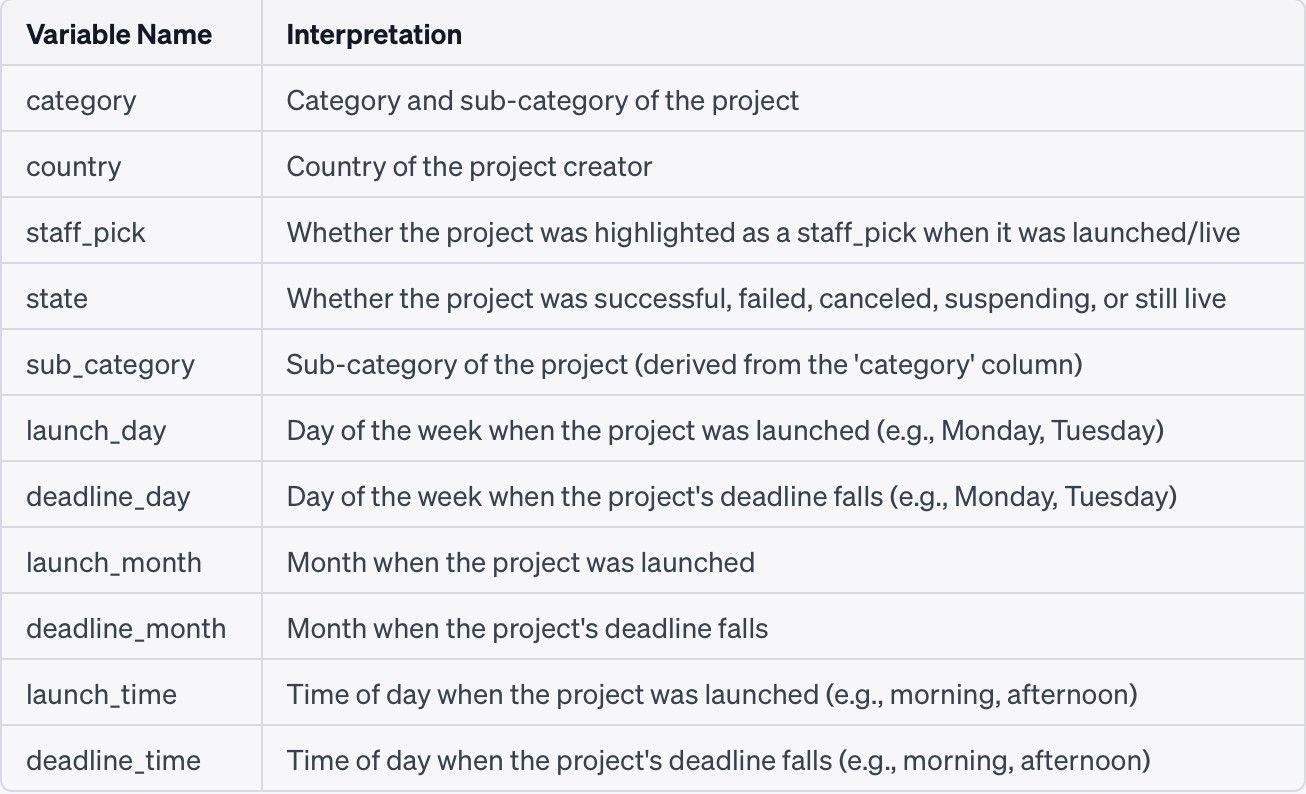
# Data Cleaning and Feature Engineering

In preparing our dataset for analysis and predictive modeling, our team conducted several key steps. We removed columns with limited non-null entries, converted datetime columns for easier analysis, calculated blurb lengths, and extracted category and subcategory information. We standardized currency, dropped less informative variables, and derived additional features. We also ensured data quality by addressing null values and duplicates. Finally, we filtered out irrelevant project states to focus on predicting project success. Removed canceled, live, and suspended projects from the dataset as they are not relevant.

After implementing all these essential steps to curate a clean and pertinent dataset for our analysis, the focal variables included were:

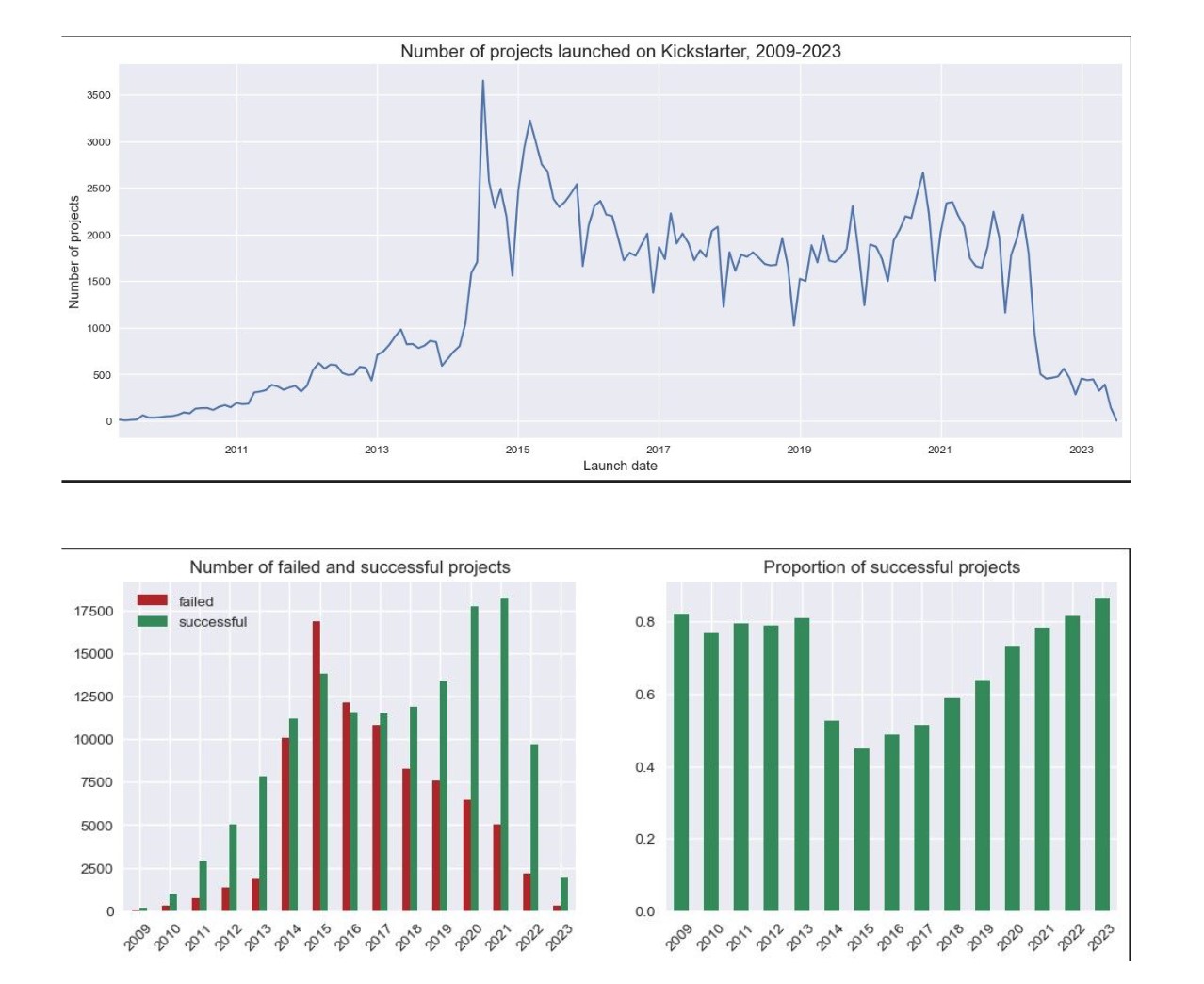
*Table 1.1*

*Numerical Data*



*Table 1.2 Categorical Data*

# Exploratory Analysis and Graphs



*Table 1.1 Statistics for projects launched on Kickstarter (2009-2023)*

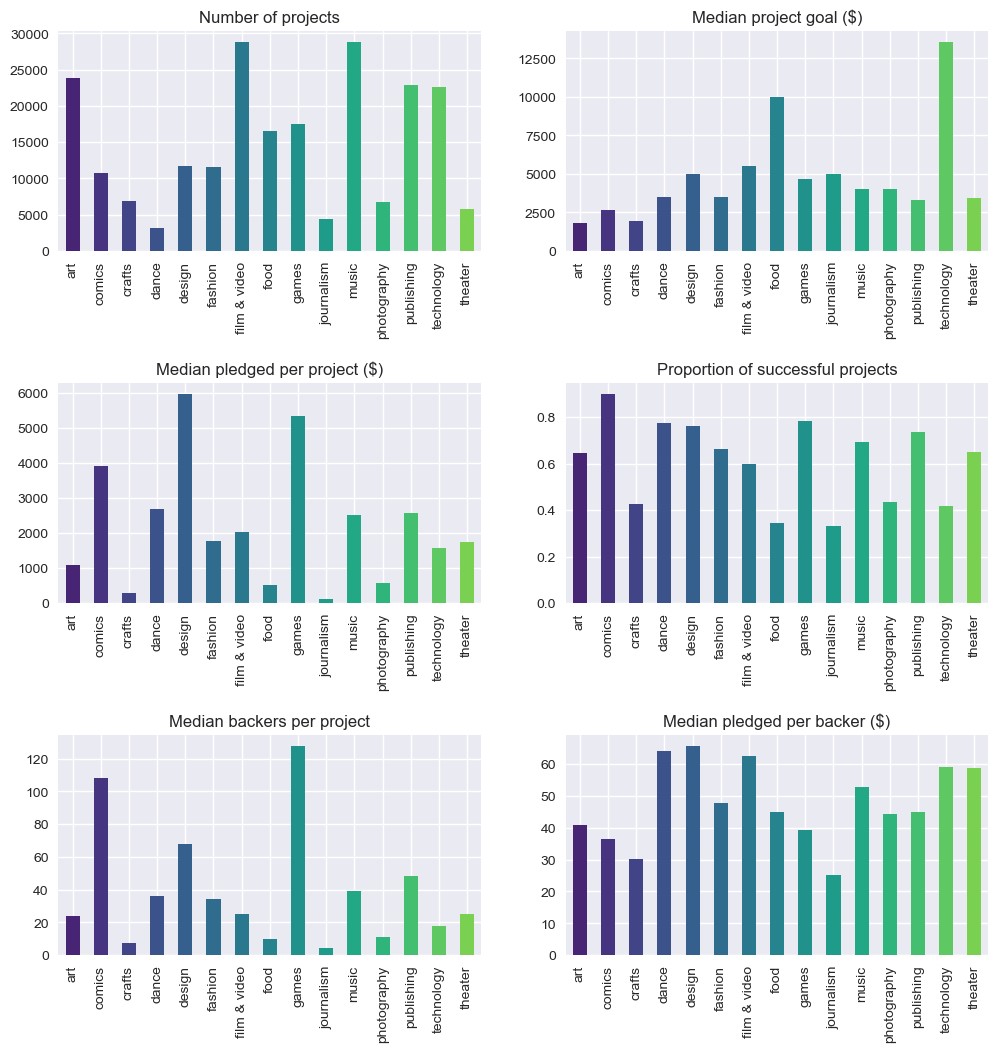
Kickstarter became really popular after it started in 2009, especially in 2014 when it grew a lot. Since 2015 the number of projects is decreasing. However, we can provide some general insights into factors that might have influenced changes in the number of Kickstarter projects, but it is not our objective in this work to analyze why the number of projects in Kickstarter are descending:

* Platform Maturity: By 2015, Kickstarter had already been operating for several years (since its founding in 2009). Over time, crowdfunding platforms tend to evolve and mature. The rapid initial growth of Kickstarter might have stabilized or slowed as the platform matured.
* Market Saturation: As crowdfunding gained popularity, more and more projects were launched on Kickstarter. Eventually, this could lead to market saturation, making it more challenging for new projects to gain visibility and secure funding. Increased competition could result in fewer projects being successful.
* Quality and Curation: Kickstarter has introduced measures to improve project quality and reduce the number of unsuccessful or fraudulent campaigns. This could result in more rigorous project curation and potentially fewer low-quality projects being allowed on the platform.
* Economic Factors: Economic conditions can impact crowdfunding activity.
* Regulatory Changes: Regulatory changes related to crowdfunding, investment, or taxation can impact the number and types of projects on Kickstarter. Such changes might have occurred in 2015 and influenced campaign activity.
* Platform Policy Changes: Kickstarter periodically updates its policies and guidelines for project creators. These changes can affect who can launch projects and how they are run, potentially impacting the number of projects.
* Shift to Other Platforms: Some crowdfunding activity may have shifted to other crowdfunding platforms or methods (e.g., Patreon), affecting the number of projects on Kickstarter.

The decline in the number of projects can be attributed to the trend depicted in the graph, where the number of failed projects surpasses that of successful projects, particularly in 2015. This shift in dynamics may influence backers' behavior, leading to increased skepticism about investing in Kickstarter campaigns.

## Kickstarter Projects for Success Factors

*Table 1.2 Analysis Grouped by Category*

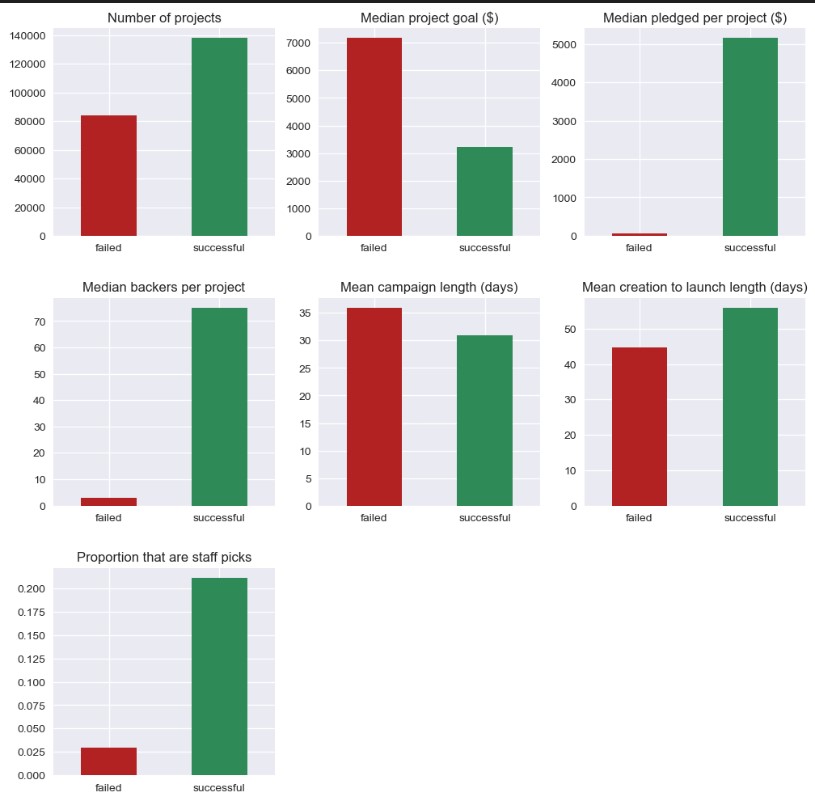


In our exploration of Kickstarter projects, we aimed to understand the diverse landscape of crowdfunding and uncover factors contributing to project success.

Among the 15 project categories, *"Music,"**"Film & Video,"*and*"Art"* emerge as the most prevalent. However, *"Technology"*and *"Food"* projects stand out with ambitious funding goals. Surprisingly, *"Technology"* projects, despite their high goals, receive lower median pledges compared to their substantial targets. On the other hand, *"Games," "Comics," "Dance,"* and*"Design"* projects secure substantial average funding. *"Comics" and "Dance"* categories succeed with modest funding goals, while *"Food," "Journalism," and "Technology"* categories face funding challenges. *"Comics" and "Games"* attract more backers, albeit with smaller individual contributions, while *"Dance" and "Film & Video"* attract generous backers who contribute higher amounts. In essence, Kickstarter's landscape reveals diverse categories with varying funding goals, backer engagement, and project success rates.

# Key Differences between Successful and Unsuccessful Projects

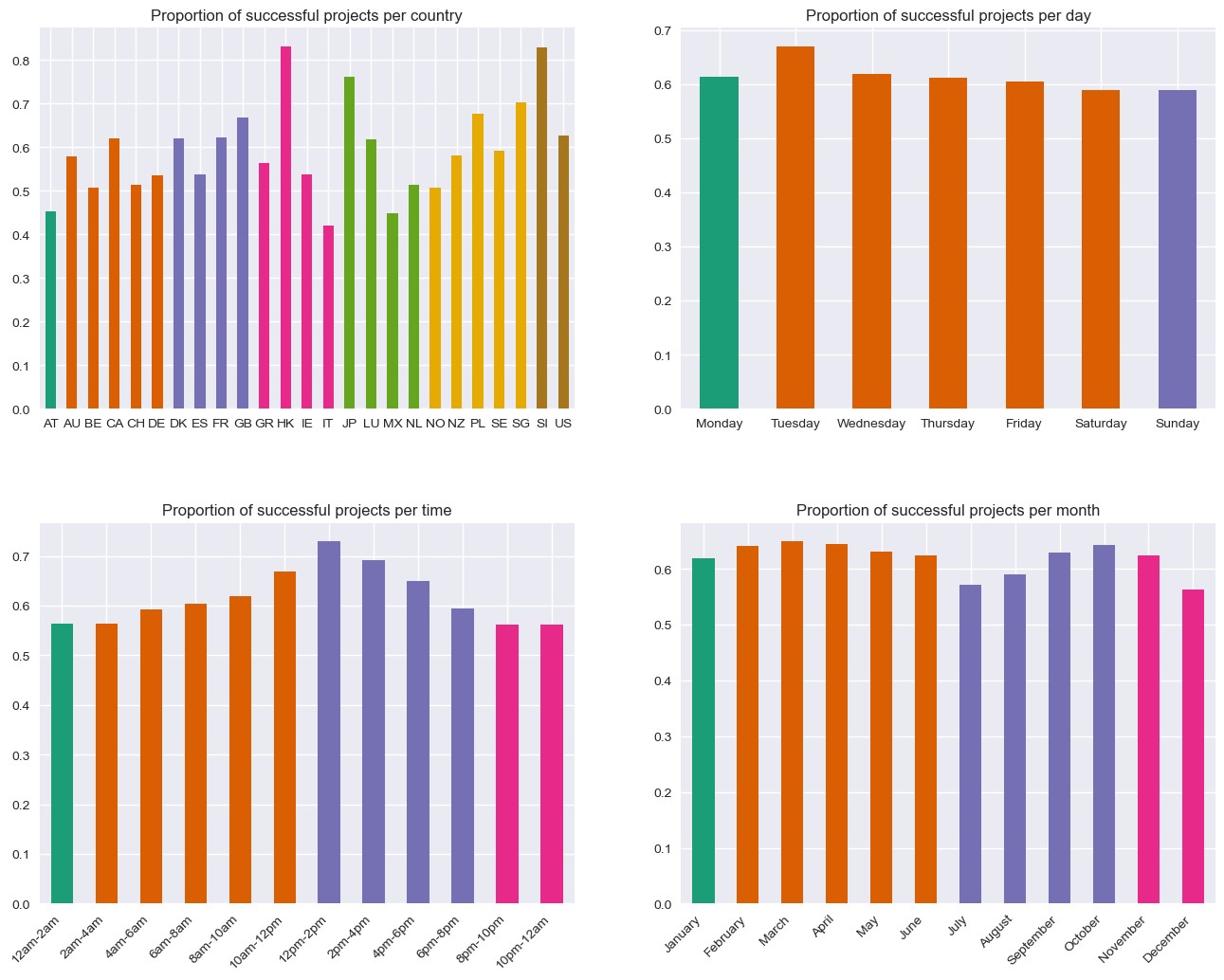
*Table 1.3 Average amount pledged to Successful and Unsuccessful projects*



Approximately 62% of completed projects achieve success on Kickstarter. Successful projects typically set smaller, more attainable funding goals, often around half the median goal of unsuccessful projects. They also receive significantly higher pledges compared to their initial goals, effectively becoming 'over-funded.' Backer engagement, measured by the amount pledged and the number of backers, shows pronounced differences between successful and failed projects, with backers favoring projects they believe will succeed. Additionally, successful projects tend to have slightly shorter campaign durations and usually take a bit more time to launch.

Projects recognized as 'Staff Picks' by Kickstarter enjoy higher success rates, highlighting the potential impact of platform endorsement. These insights offer valuable guidance to creators and backers in the diverse landscape of Kickstarter crowdfunding.

# Success ratio



*Table 1.4 Success Proportions*

Hong Kong and Singapore boast a higher proportion of successful projects.

Tuesday is the optimal day for project launches, while weekends consistently show lower success rates, mirrored in funding and backers.

The best time to initiate a project falls between 12 pm and 2 pm UTC, with the highest median backers and funding. Conversely, launching between 6 pm and 4 am UTC is less advantageous. March stands out as the best month for project launches with April and October being close. The least beneficial month to launch a project is July.

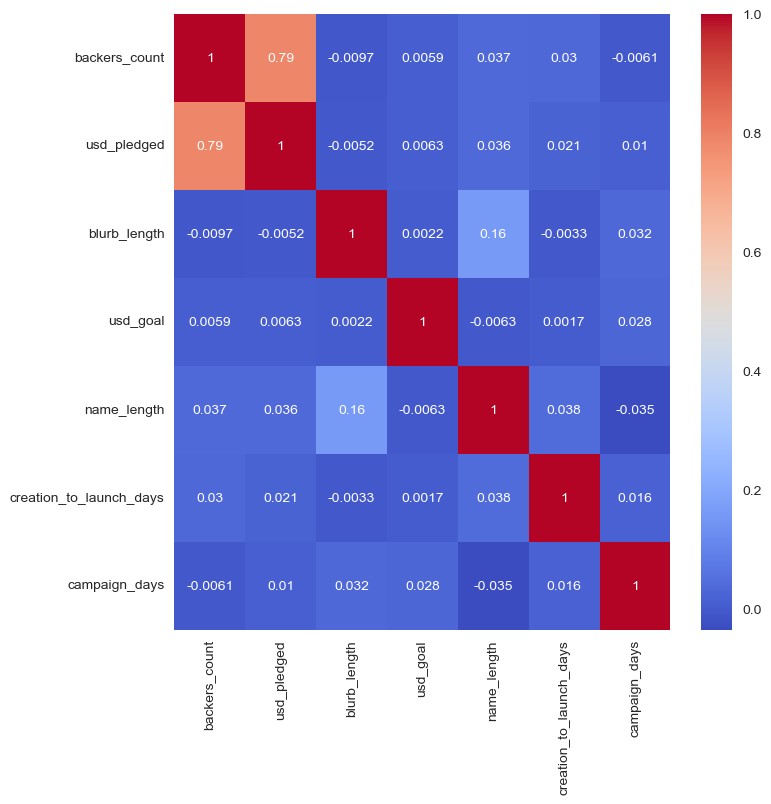
**Preparing the data for Machine Learning**

The primary objective of this project was to develop a model capable of accurately predicting whether a project would succeed or fail. To facilitate the application of machine learning, the following data preparation steps were executed:

## Multicollinearity

Multicollinearity happens when independent variables in a study are closely connected, and it can cause problems for our analysis. To spot multicollinearity, we use a tool called Pearson's correlation coefficient, which measures how strongly two variables are linked. When these correlations are too high, it can make our prediction models unstable.

In our case, we discovered that one variable, "pledge\_per\_backer," wasn't really helping our analysis, so we decided to remove it since it wasn't necessary for our goals.

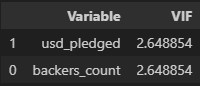


*Table 1.5 Heatmap Correlation*

The heatmap reveals a moderate connection between 'usd\_pledged' and 'backers\_count.' However, we shouldn't hastily remove these features to reduce model variability. Doing so may lead to a loss of crucial information necessary for accurate predictions. To be sure, we'll calculate the VIF (Variance Inflation Factor) to check for multicollinearity.

## Variance Inflation Factor (VIF)

Calculating the VIF (Variance Inflation Factor) for the identified features is essential to assess multicollinearity. The table demonstrates that 'backers\_count' and 'usd\_pledged' have VIF values below 5, indicating no multicollinearity issues. However, for fair and accurate predictions we opted to exclude specific columns like 'usd\_pledged' and 'backers\_count.'



After computing VIF it was necessary to execute the subsequent actions:

**Target Variable Transformation**: The target variable converted into binary values: 1 representing successful projects and 0 representing failed projects. This conversion facilitated the inclusion of boolean features in subsequent encoding.

**Categorical Variable Encoding**: Categorical variables within the dataset underwent encoding. This process involved converting categorical features into numerical format using one-hot encoding.

**Dataset Splitting**: The dataset divided into training and testing subsets using the train\_test\_split function. The split ratio will be 80% for training and 20% for testing, with a random state of 42 for consistent reproducibility.

## Modeling

We employed five different models for our Kickstarter project success prediction:

1. **Logistic Regression** (Model 1): A classic method for binary classification.

1. **Random Forests** (Model 2): Known for their robustness and ability to capture complex relationships.

1. **Decision Tree** (Model 3): Offers transparency and insight into decision-making.

1. **K-Nearest Neighbors** (KNN) Model (Model 4): Utilizes data point similarity for predictions.

1. **XGBoost** (Model 5): A powerful gradient boosting algorithm for exceptional performance.

In our machine learning analysis, we employed accuracy as a crucial metric for evaluating model performance. This metric allows us to assess the models by examining the percentage of correct predictions they make on a validation set, providing a valuable measure of their effectiveness in making accurate predictions.

## The confusion matrix order

True Negative (TN): This value represents the count of correctly predicted failed projects. These are instances that were correctly identified as failures by the model.

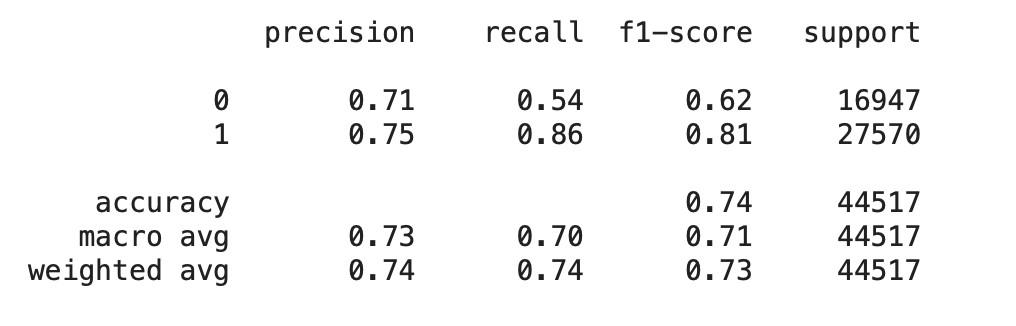
False Positive (FP): The false positive value indicates the number of projects that the model predicted as successful, but in reality, they failed. These are instances where the model made an incorrect positive prediction.

False Negative (FN): This value shows the count of projects that were predicted to fail by the model, but they actually succeeded. These instances were incorrectly predicted as negatives.

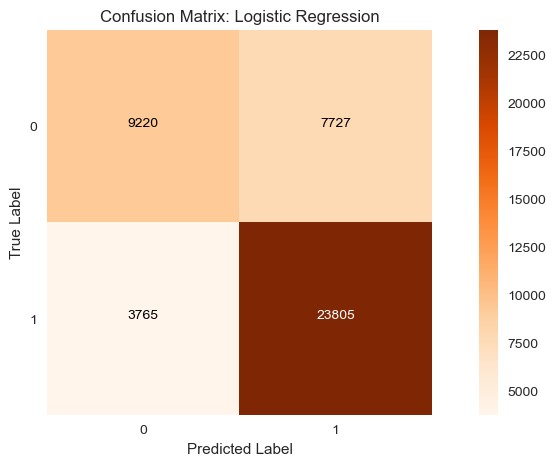
True Positive (TP): The true positive value is the count of correctly predicted successful projects. These are instances that were correctly identified as successes by the model.

## Model 1. Logistic Regression

Classification report:

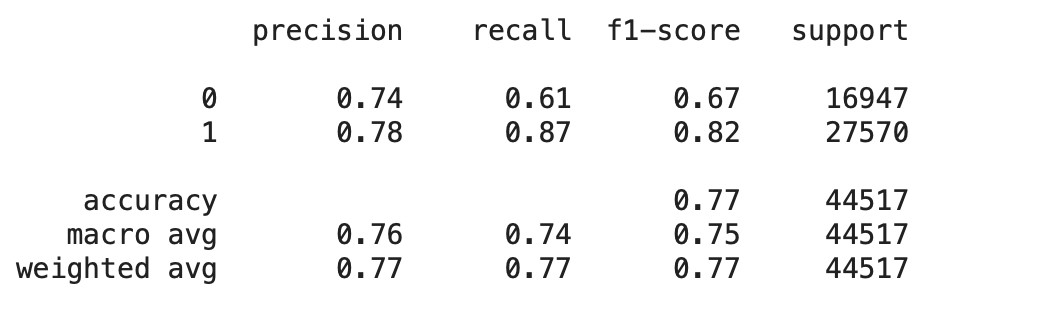


**Accuracy**: Overall accuracy is 74%, signifying 74% correct predictions across both classes.

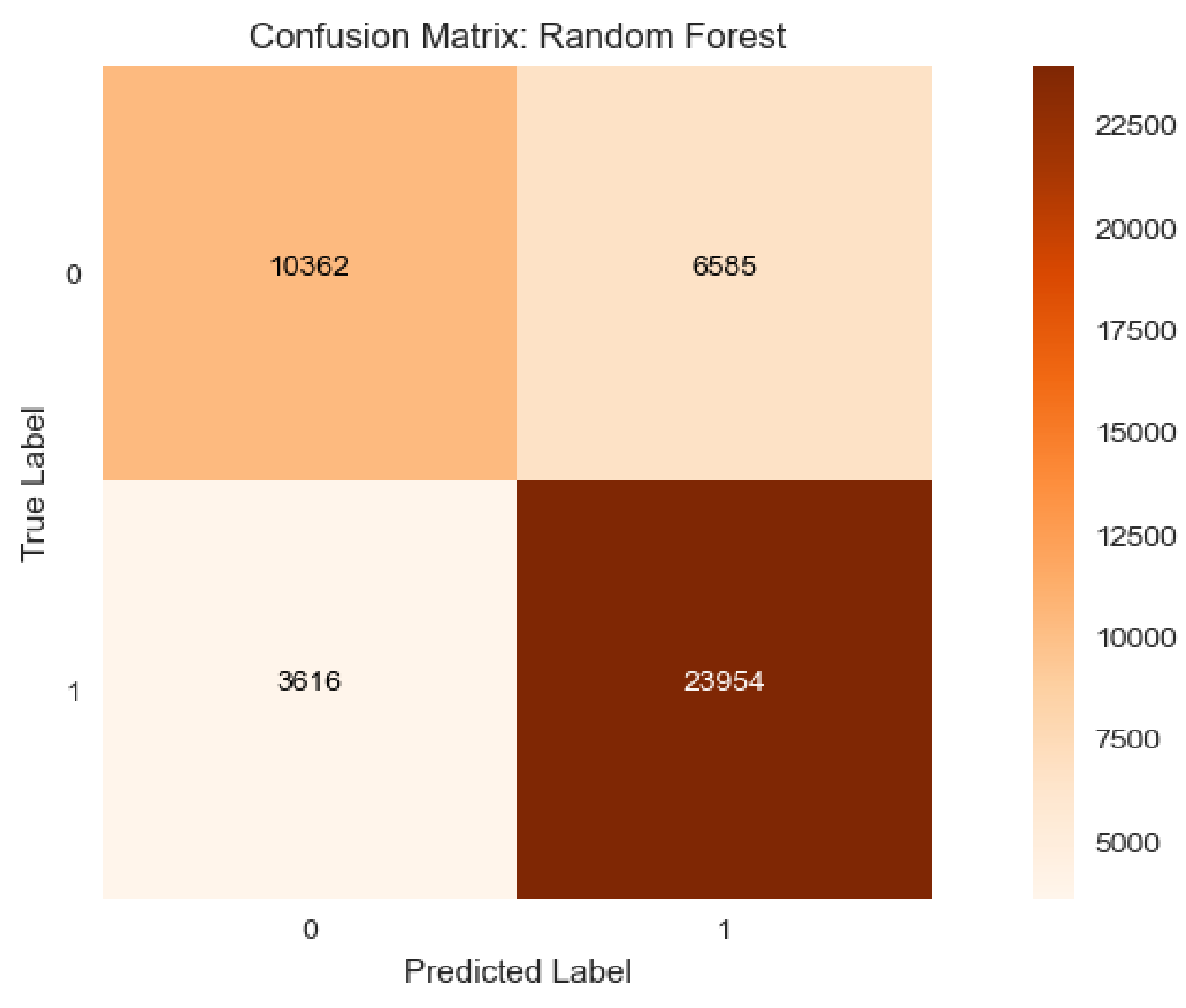


## Model 2: Random Forests

Classification report:

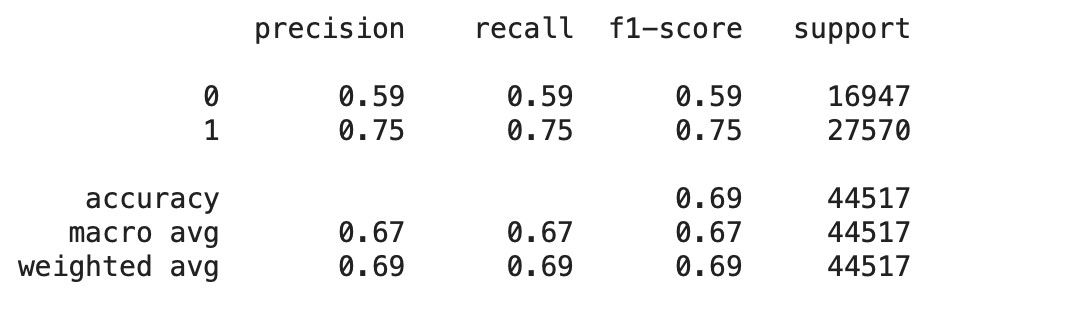


**Accuracy**: Overall accuracy is 77%, indicating 77% correct predictions across both classes.

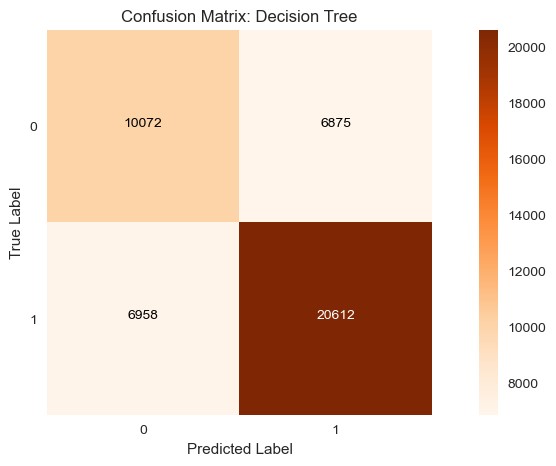


## Model 3: Decision Tree

Classification report:

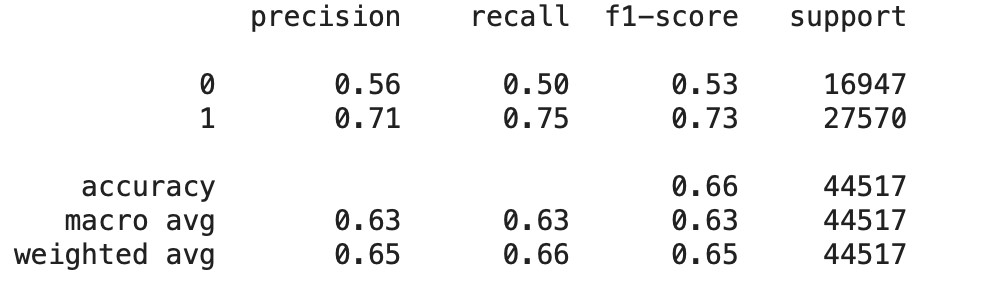


**Accuracy**: The overall accuracy of the model is 69%, implying 69% correct predictions across both classes.

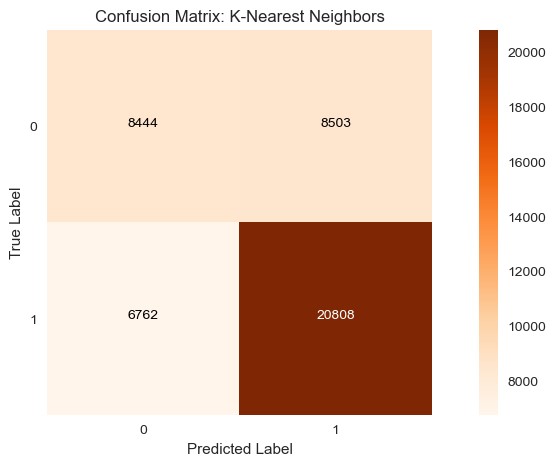


## Model 4: KNN Model

Classification report:



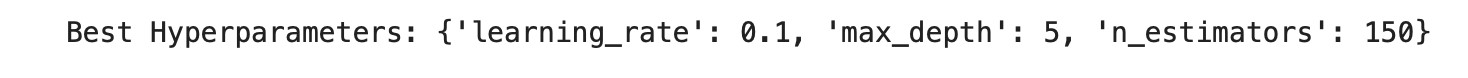
**Accuracy**: The overall accuracy of the model is 66%, implying 66% correct predictions across both classes.



## Model 5: XGBoost

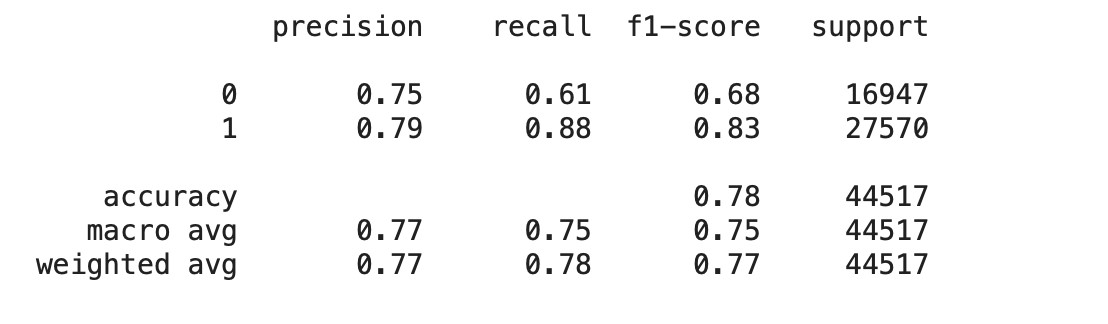
XGBoost is a form of gradient boosting algorithm. Similar to Random Forests, it is an ensemble method that produces multiple decision trees to improve classification of data points.

* Create an XGBoost classifier.
* Define a parameter grid to search.
* Perform grid search using cross-validation.
* Obtain the best hyperparameters from the grid search.



* Build an XGBoost classifier using the best hyperparameters.
* Train the classifier on the training dataset.
* Generate predictions using the trained model on the test dataset.
* Evaluate the model's performance on the test data.

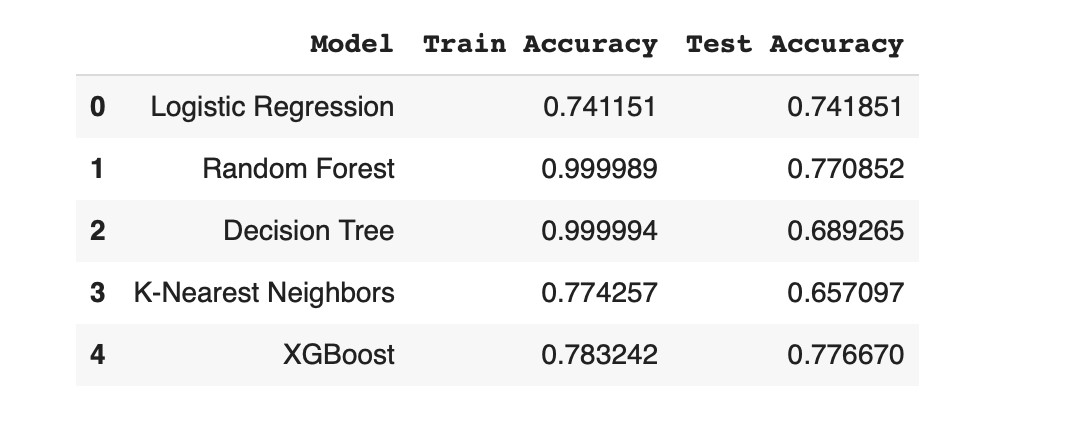
Classification report:



**Accuracy**: The overall accuracy of the model is 78%, indicating 78% correct predictions across both classes.

### Model Selection

In this section we were required to construct a data frame containing the models along with their corresponding training and testing scores

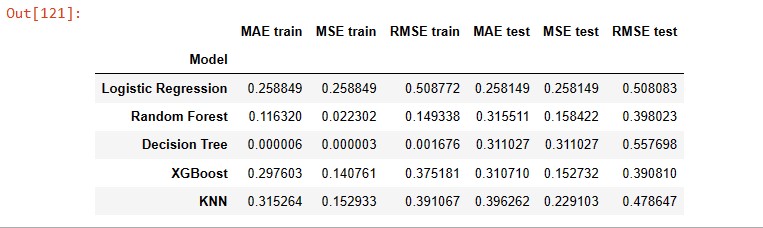


### Observations

Among the machine learning models assessed, Logistic Regression, Random Forest, and XGBoost demonstrated relatively balanced training and test accuracy. Decision Tree showed a significant gap between training and test accuracy, indicating potential overfitting. K-Nearest Neighbors exhibited lower accuracy in both training and test sets compared to the other models. Based on these results, both Random Forest and XGBoost appear promising, showing high test accuracy and relatively balanced performance between training and test sets.

Comparison of metrics

:



### Evaluation

**Random Forest** exhibits strong performance in the training set, indicated by consistently low MAE, MSE, and RMSE values. However, its performance on the test set displays slightly increased error, implying some level of overfitting.

**Decision Tree** yields extremely low error on the training set, suggesting a perfect fit to the training data. However, its test set performance showcases substantially higher error, indicating poor generalization capability.

**XGBoost** demonstrates a balanced performance between the training and test sets, reflecting good generalization ability.

**Logistic Regression** maintains consistent performance across both the training and test sets, with relatively low error metrics.

**KNN** registers comparatively higher errors in both training and test sets, suggesting potential limitations for this specific problem.

In conclusion, based on these evaluation metrics, Random Forest and XGBoost emerge as promising contenders for our application.

### Recommendations

Based on the analysis of the Kickstarter dataset and the model evaluation results, here are some key recommendations for launching successful Kickstarter campaigns:

Factors with Positive Effects on Success Rate and Funding Amount:

1. **Set Smaller Project Goals**: Campaigns with smaller funding goals tend to have higher success rates. Consider breaking down larger projects into smaller, achievable goals.
2. **Strive for Quality and Staff Picks**: Aim to create high-quality projects that may attract attention from Kickstarter staff picks. This can positively impact your project's success.
3. **Opt for Shorter Campaigns**: Shorter campaign durations, especially around 30 days, tend to perform better. It keeps the momentum and urgency high.
4. **Plan Adequate Pre-launch Time**: Allow sufficient time between creating and launching your campaign. This preparation phase can contribute to better results.
5. **Explore Comics, Dance, and Games Categories**: Consider launching projects in categories like comics, dance, and games, as they have shown a higher likelihood of success.

Factors with Negative Effects on Success Rate and Funding Amount:

1. **Avoid Large Funding Goals**: Setting excessively large funding goals can deter backers. Focus on realistic and achievable targets.
2. **Keep Campaigns Short:** Longer campaign durations, especially well beyond 30 days, may lead to campaign fatigue and lower success rates.
3. **Be Cautious with Food and Journalism Projects:** Food and journalism projects may face challenges on Kickstarter. Ensure your project stands out in these categories.
4. **Consider Alternative Launch Locations:** If possible, explore options beyond Italy for launching your Kickstarter project to maximize success.

**Additional Insights:**

**Launch Strategically**: Launch your campaign on a Tuesday, as it has shown some positive results. Be mindful of the competition on this popular launch day.

**Choose March, April and October for launch**: These months tend to be a favorable month for campaign launches, offering better chances of success.

**Consider Launch Timing**: Launch your campaign between 12pm and 2pm UTC, taking into account the global audience on Kickstarter.

**Optimize Project Names and Blurbs**: Craft shorter and compelling blurbs, and consider longer project names, as these factors have shown preferences among backers. **Other Geographic Considerations**: Pay attention to geographic factors such as project location and its potential impact on success.

### Conclusion

In our fascinating exploration of Kickstarter projects, we set out to unravel the intricacies of crowdfunding and uncover the secrets to success. Along the way, we made some remarkable discoveries.

Among the 15 diverse project categories, "Music," "Film & Video," and "Art" emerged as the most prominent. However, "Technology" and "Food" projects caught our attention with their ambitious funding goals. Interestingly, "Technology" projects, despite their grand aspirations, often received pledges that fell short of their colossal targets.

Amidst this landscape, we encountered thriving genres such as "Games," "Comics," "Dance," and "Design," all boasting substantial average funding. "Comics" and "Dance" genres, in particular, showcased how dreams could take flight with modest funding goals, achieving remarkable success. On the flip side, "Food," "Journalism," and "Technology" projects faced the challenge of meeting their high funding needs.

Within the realm of backers, "Comics" and "Games" attracted a multitude of enthusiasts, each contributing their part to support these creative endeavors. In contrast, "Dance" and "Film & Video" were favored by generous patrons who made substantial contributions, highlighting the diverse nature of Kickstarter backers.

As we conclude this chapter of our journey, remember that these insights are guiding stars for aspiring creators. They offer valuable directions to navigate the world of Kickstarter. Yet, always keep in mind that your project is unique and has its path to follow. Adapt your strategy, stay true to your vision, and may your Kickstarter journey be as bright as the stars themselves.