# PROJECT SUMMARY

*The goal of this analysis is to group Kickstarter projects into distinct clusters using K-Means clustering. This would help understand patterns in project features and identify factors contributing to project success.*

## 1. Data Preprocessing Steps, Feature Engineering, and Standardization

For the purpose of this project, only observations where the variable ‘state’ is ‘successful’ or ‘failed’ are included in the analysis. Log transformations were  applied to skewed variables like goals and pledged amounts to mitigate outlier effects. Frequency encoding was applied to ‘country’, ‘main_category’, and ‘category’ variables, as dummifying them would add excessive columns, making clustering less effective. Then, dummification was applied to the rest of the categorical variables.

For feature engineering, below features were added to the dataset:

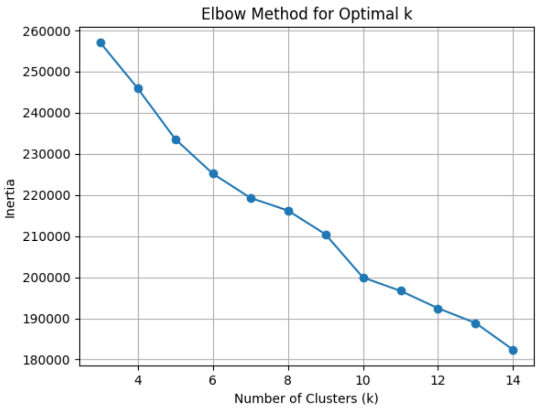
| **Feature** | **Rationale** |
|---|---|
| time_to_launch | Captures how time to prepare for a project can affect its outcome. |
| funding_duration | Captures how time available to fund a project can affect its outcome. |
| Goal_usd | Standardizes the goal values across projects from different countries to allow for fair comparisons. |
| Category_success_rate | Captures historical trends where some categories may have higher success rates based on popularity/demands. |
| Category_frequency_tier | Differentiates between more popular categories with less popular ones, which can potentially affect the success of the project. Category is classified into three tiers (‘High’, ‘Medium’, and ‘Low’) based on its frequency. |
| pledged_goal_difference_usd | Indicates funding progress, i.e. how close a project is to meeting its funding goal (in USD). |
| pledged_goal_ratio | Indicates funding progress, i.e. how close a project is to meeting its funding goal (in %). |

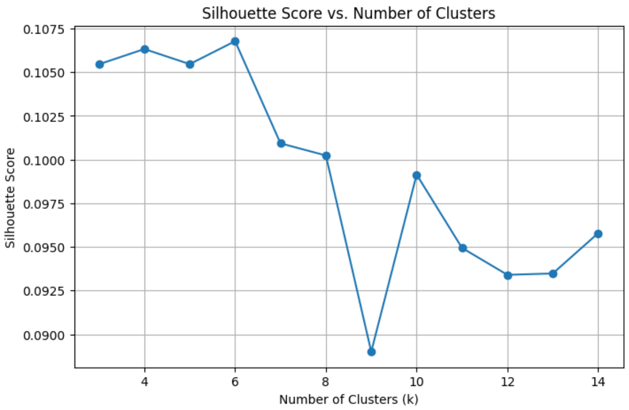
Standardization was applied to ensure all features contributed equally to the K-Means clustering process. Numerical features were scaled to have a mean of 0 and a standard deviation of 1, preventing larger-range features (e.g., pledged amount, goals amount) from dominating the Euclidean distance calculations and improving clustering reliability. To improve clustering performance, I removed 145 anomalies using Isolation Forest and then eliminated highly correlated features through VIF analysis. The final dataset comprises of 14,318 lines, with 23 variables.

## 2. Develop a Clustering Model

K-Means clustering was chosen for this task for its interpretability, its ability to efficiently handle dataset with moderate amount of features, and its ability to create compact clusters, aligning with the goal of identifying similar project groups.

To determine the optimal number of clusters, two methods were employed: the Elbow Method and Silhouette Score. The Elbow Method evaluated the inertia (within-cluster sum of squares) for cluster numbers ranging from 3 to 15, identifying a significant "elbow" at k = 6. The Silhouette Score, calculated for the same range, also peaked at 6. Therefore, the optimal number of clusters is 6. While the low Silhouette score (0.1068) at k = 6 suggests some cluster overlap, the model still provides valuable insights into project groupings.


Elbow Method for Optimal k


Silhouette Score vs. Number of Clusters

K-means clustering with k = 6 is applied to the dataset. For further analysis, the cluster labels and the ‘state’ variable are added to the dataset. The cluster success rate is then derived from the ratio of actual number of successful projects within a cluster over the total number of projects in that cluster.

## 3. Results Discussions

From the clustering result, we can see that each cluster demonstrated unique characteristics, helping distinguish between successful and unsuccessful projects. Cluster characteristics are summarized below:

| | **Success Rate** | **Characteristics** |
|---|---|---|
| 1 | 91.1% | Low goals amount with short funding duration, cover more popular categories with above average category success rate. |
| 2 | 0.0% | Higher goals, focus on less popular categories with a very low success rate. |
| 3 | 99.3% | Lower goals and short funding duration, with a focus on niche categories. |
| 4 | 94.6% | High goals and pledged amounts, cover high-success-rate categories, strong visuals (image and videos) and long descriptions. |
| 5 | 0.0% | High funding goals, focus on niche categories with low success rate. |
| 6 | 98.3% | Moderate goals, cover moderately popular categories with high success rate. |

The results reveal several insights: First, setting realistic goals is crucial, as projects with low or moderate goals consistently perform better. Second, moderately popular and high-success-rate categories yield better outcomes, while niche and low-success-rate categories require extra effort to succeed. Third, campaign quality (strong visuals, detailed descriptions, etc.) can boost success rates. Lastly, niche categories can either perform really well (cluster 3) but also can easily fail without proper campaign and engagement strategy (cluster 5).

## 4. Business Implications

With these insights, Kickstarter can offer tailored support to struggling projects (e.g., clusters 2 and 5) through additional marketing support and project consultations, while promoting high-potential ones to attract more backers and funding. Kickstarter can also leverage these findings to optimize project onboarding processes by offering goal-setting advice for specific categories. To drive engagement and improve backers' experience, Kickstarter can highlight projects with realistic goals and offer personalized recommendations to backers based on the project characteristics that attract them.

In summary, by offering insights into failed and successful projects, this model benefits Kickstarter by allowing the business to allocate resources more effectively. For creators, these insights enable them to improve their project design and goal settings. For backers, these insights ensure better project recommendations and help build greater confidence in the platform. This ultimately enhances satisfaction for all stakeholders involved.