Project Objective: Develop a classification model to predict whether a Kickstarter project will be "successful" or "failed" at the time of its launch. Also, create a clustering model to group similar projects and analyze the characteristics of each cluster to derive meaningful business insights.

TASK 1: Developing a classification model

1. **Relevant pre-processing steps:**

Outlier handling through log transformations (e.g., goal_usd) and anomaly detection (graphical approach and Isolation Forest) ensured data integrity by mitigating the impact of extreme values. Feature engineering added valuable predictors like campaign_duration, time_to_launch, and categorical indicators such as **launched_at_type** (weekend vs. weekday) and **time of day** (e.g., afternoon, evening, night). Scaling normalized features for Logistic Regression, preventing dominance by larger scales. Rare categories like countries were grouped into "Other," and one-hot encoding for categorical variables enabled effective model interpretation.

Note: Feature selection, column decisions (e.g., grouping countries, transforming dates), and various model tunings were explored in separate code files for simplicity. While these trials are excluded from the final code, their insights are fully reflected in choosing the final model.

2. **Justifications of the model:**

Model Name	RF	Gradient Boosting	ANN	KNN	Logistic regression	Ridge-regularized	**Lasso-regularized**
Accuracy score	0.77	0.453	0.77 (hidden_layer=3)	0.73 (n_neighbors=3)	0.7854	0.7859	**0.8016**

Logistic Regression with Lasso Regularization is ideal for binary classification, with Lasso ensuring feature selection by penalizing less important predictors to reduce overfitting. Hyperparameter tuning via GridSearchCV further optimized performance.

3. **Usefulness of the model in the business context:**

Predictive Insights - Identifies projects likely to succeed at launch, allowing Kickstarter to strategically allocate resources and provide targeted support.

Optimizing Backer Engagement - Helps prioritize projects with higher success probabilities, increasing backer trust and encouraging more investments by showcasing well-positioned campaigns.

Data-Driven Strategies - Insights from predictors like ‘campaign_duration’ and ‘goal_usd’ can guide project creators in setting realistic goals and timelines to attract more backers.

Operational Efficiency - Streamlines efforts by highlighting promising projects, enabling Kickstarter to focus editorial features or promotional efforts effectively.

Risk Mitigation - Reduces platform association with failed projects, building confidence among backers and enhancing platform credibility.

TASK 2: Developing a clustering model

1. Clustering Methodology:

Given the nature of the Kickstarter dataset—with numerous quantitative features like funding goals, backer counts, and time-related metrics—K-Means was chosen for its ability to effectively minimize within-cluster variance while maximizing between-cluster separation. **DBSCAN Clustering** was explored but deemed less suitable for this dataset because of its sparsely distributed noise points. However, it effectively identified and removed anomalies (in combination with Isolation Forest), enhancing the overall performance of K-Means clustering. Principal Component Analysis (PCA) reduced the dataset’s complexity while preserving 95% of the original variance.

2. Insights obtained from the results:

Clusters 4, 6, and 3 represent high-success groups, with a focus on categories like Publishing, Film & Video, and Technology. Cluster 4 includes projects with strong engagement, realistic funding goals, and excellent performance. Cluster 6 is characterized by small-scale campaigns with low funding goals and high success rates, supported by focused backer engagement. Cluster 3 consists of niche projects with strong community support, often excelling in more specialized categories.

Clusters 1 and 2 fall into the low-success category. Cluster 1 features high-goal projects with moderate success but struggles to meet ambitious targets, while Cluster 2 comprises small-scale campaigns with below-average success rates and potential for strategy improvements.

Clusters 0 and 5 exhibit average success. Cluster 0 includes mid-range campaigns with moderate performance (Although the cluster centers indicate that Cluster 0 has lower overall success compared to Cluster 1, it performs better in achieving its funding goal amounts) and steady backer engagement, while Cluster 5 contains projects that perform consistently across metrics but lack standout success, requiring better marketing or clearer messaging.

3. Usefulness of the model and potential business impacts:

For **Kickstarter's platform**, it supports targeted resource allocation, promoting high-potential clusters (e.g., Cluster 4) and offering additional support to struggling ones (e.g., Cluster 1). This improves campaign success rates, boosting user satisfaction and retention. For **creators**, the model provides specific guidance, such as aligning funding goals with successful patterns, refining campaign messaging, and optimizing launch timing to maximize success. For **backers**, the model improves project discovery while increasing the chances of supporting successful campaigns, ensuring a more rewarding experience.