Part 1: Predicting Kickstarter Success

Data Preprocessing

To ensure a robust model, irrelevant post-launch features (e.g., `backers_count`) were removed so that all predictors represented information available at launch. Missing values were handled by removing incomplete rows. This decision was made after verifying that the missing categories were available on the Kickstarter website, meaning imputation would not add meaningful information.

Here are some features I engineered:

Feature Name	**Description**
`goal` (normalized)	Original funding goal, normalized to USD using `static_usd_rate`.
`time_to_launch`	Number of days between project creation and launch, capturing preparation time.
`campaign_duration`	Number of days between launch and deadline, representing the total campaign duration.

Categorical variables, such as `country` and `category`, were encoded as dummy variables. Numerical features were standardized to ensure uniform scaling. To address skewness, log transformation ($\log(1 + x)$) was applied to `goal` and `time_to_launch`, which had high positive skew, ensuring a more normal-like distribution and reducing the impact of extreme values.

Feature Selection

Feature selection methods, including LASSO regression and Random Forest feature importance, were explored to identify key predictors. While both methods highlighted important features such as `goal`, `campaign_duration`, and `video_True`, cross-validation revealed that using the entire feature set resulted in the highest accuracy. Therefore, I opted to include all features in the final model, ensuring no predictive information was omitted.

Model Selection and Justification

For the classification task, I evaluated multiple models, including Logistic Regression, Random Forest, and Gradient Boosting. The final logistic model was trained using all features, and its effectiveness was validated on the test set, achieving an accuracy of **78.65%**. The interpretability of Logistic Regression allows stakeholders to quantify the impact of specific features, such as the campaign duration or funding goal, on project success, so they can adjust their strategy to improve future campaign outcomes.

Part 2: Clustering Analysis and Business Strategy

Data Preprocessing and Feature Selection

To prepare data for clustering, I first selected features based on their relevance to project performance and clustering objectives. `goal` and `campaign_duration` were chosen to represent financial and timeline metrics, respectively, while `main_category` captured the type of project. Indicators such as `video` and `show_feature_image` were included to reflect presentation attributes, which are critical for attracting backers. Additionally, `blurb_len_clean` and `name_len_clean` were selected to assess the quality and length of project descriptions and names, providing insights into project presentation. Post-launch features like `usd_pledged` were incorporated to evaluate the actual financial performance of the projects. All selected features underwent the same preprocessing steps as in Part 1.

Cluster Formation and Business Insights

Principal Component Analysis (PCA) was utilized to reduce dimensionality and improve clustering performance. Various numbers of components (e.g., 2, 5, 10, 15, and 20) were tested, and their corresponding silhouette scores were evaluated. The highest silhouette score of 0.35 was achieved with two components, ensuring optimal compactness and separation among the clusters.

Using the reduced dataset, the K-means algorithm was applied to cluster the data. Using the elbow method and silhouette score analysis, the optimal number of clusters was determined to be 3. These clusters showcased distinct characteristics in terms of funding goals, campaign durations, and category distributions, so that I can use the characteristics of each cluster to provide actionable insights and strategies.

Cluster	**Characteristics**	**Key Business Strategy**
Cluster 1: Premium and Ambitious Projects	High funding goals, long campaign durations, typically in technology and design.	Professional promotional resources
Cluster 2: Mid-Range Projects	Moderate goals, short durations, mostly creative categories (e.g., art, music).	Storytelling workshops
Cluster 3: Small-Scale, Niche Projects	Low funding goals, minimal campaigns, often personal or niche projects.	Budget planning support

Professional Promotional Resources for Cluster 1 Premium and ambitious projects can benefit from tailored resources that refine their campaign presentations. By offering expert consultations and professional design tools, these campaigns can align better with high-value backer expectations and stand out in competitive categories like technology and design.

Storytelling Workshops for Cluster 2 For mid-range projects, creators need support in narrating their vision effectively. Hosting storytelling workshops with step-by-step templates and impactful examples can empower them to connect emotionally with their audience.

Budget Planning Support for Cluster 3 Small-scale and niche projects require efficient resource utilization. Providing structured budget templates along with guidance on leveraging free marketing channels can help creators achieve their goals while resonating with targeted, smaller audiences.