Team 3 - Kickstarter dataset for Survival Analysis

1. Name of the data set and its source

We – team 3 will be using the **"Kickstarter"** dataset, originally curated by the Virginia Tech DMKD lab. It can be downloaded from:

- https://dmkd.cs.vt.edu/projects/survival/data/18k_Projects.csv (raw project data)

- https://dmkd.cs.vt.edu/projects/survival/data/kickstarter.zip (pre-processed survival table) - this file will be used for analysis as it already **one-hot encoded** its categorical variables

2. Number of observations and number of variables

- **Observations:** 4 175 campaigns (rows)

- **Variables:** 56 columns, including

    o 2 survival core fields

    o 15 category dummies

    o 6 currency dummies

    o 27 campaign- and creator-level features

    o 6 network aggregates

The description for variables will be included at the end of this file

3. Specify the starting event

The time-zero for every campaign is its **launch date** on Kickstarter, when the project first becomes publicly available for backing.

4. Determine the elimination ("failure") event

The "event" is defined as **the first day the campaign's pledged amount meets or exceeds its funding goal**. Once that threshold is crossed, the campaign is considered "successful" and exits the risk set.

5. Provide the primary feature

- **Primary time-to-event variable:**

    o **day_succ** is the number of calendar days from launch until that first-goal-met day.

- **Primary event indicator:**

- **Status** equals 1 if the project succeeded within the observed window, and 0 if it never reached its goal (right-censored).

6. Examples of secondary characteristics

These are static covariates measured at launch and used to explain variation in survival times. Examples include:

- **Campaign design & engagement**: updates, comments, rewards, videos, images, words_description, words_risks_and_challenges, faqs

- **Funding targets & outcome**: goal, pledged, backers, duration

- **Social-media footprint**: facebook_connected, facebook_friends, facebook_shares, has_video

- **Creator track record**: creator_projects_created, creator_projects_backed, crt_suc_perc

- **Promotion badge**: prj_prom

- **Social-network aggregates**: e.g. 1f_plg, 2f_bck, usr_net_pg_rnk, usr_net_bic

7. How the time-until-event variable is determined

day_succ was computed as

$$day\_succ_i = \text{(the calendar date on which project i first reaches its funding goal)} - \text{(the calendar date the project was launched)}$$

in days. If a project never hit its goal, day_succ records the full campaign length (capped at 60 days) or the data-collection cutoff.

8. How the censoring variable is determined

Status is set to:

- **1** if the project's pledged total reached or exceeded the goal on or before day 60

- **0** otherwise (i.e. the project closed without ever hitting the goal).
  Those with Status = 0 are treated as right-censored at their day_succ value.

Variable Descriptions

A. Survival Analysis Core

day_succ (Integer): Number of days from project launch until the funding goal was met (or the full observed time if goal not met).

Status (Binary 0/1): Event indicator (1 if project succeeded within the window, 0 if right-censored).

B. Category Dummies (cat1–cat15)

- **cat1**: Art
- **cat2**: Comics
- **cat3**: Crafts
- **cat4**: Dance
- **cat5**: Design
- **cat6**: Fashion
- **cat7**: Film & Video
- **cat8**: Food
- **cat9**: Games
- **cat10**: Journalism
- **cat11**: Music
- **cat12**: Photography
- **cat13**: Publishing
- **cat14**: Technology
- **cat15**: Theater

C. Currency Dummies (curr1–curr6)

- **curr1**: AUD
- **curr2**: CAD
- **curr3**: EUR
- **curr4**: GBP
- **curr5**: NZD
- **curr6**: USD

D. Core Campaign Stats (z-scored)

- **goal**: Funding target
- **pledged**: Total pledged amount
- **backers**: Number of backers
- **duration**: Campaign length in days
- **updates**: Number of creator updates
- **comments**: Number of backer comments
- **rewards**: Number of reward tiers
- **videos**: Number of embedded videos
- **images**: Number of images
- **words_description**: Word count of description
- **words_risks_and_challenges**: Word count of Risks & Challenges section
- **faqs**: Number of FAQs

E. Social-Media & Engagement
- **facebook_connected**: 0/1 whether Facebook account is linked
- **facebook_friends**: Number of Facebook friends
- **facebook_shares**: Number of public shares
- **has_video**: 0/1 whether campaign includes a video

F. Creator History
- **creator_projects_created**: Previous projects launched
- **creator_projects_backed**: Projects backed by creator
- **crt_suc_perc**: Historical success rate of creator

G. Promotion Badge
- **prj_prom**: 0/1 Kickstarter promotion badge
- **1prj_prom**: Proportion of 1-hop neighbors promoted
- **2prj_prom**: Proportion of 2-hop neighbors promoted
- **3prj_prom**: Proportion of 3-hop neighbors promoted

H. Network Structural Features
- **usr_net_pg_rnk**: PageRank centrality
- **usr_net_nodes**: Ego-network size
- **usr_net_tie**: Average tie strength
- **usr_net_bic**: Betweenness-in-community centrality

I. Friends-of-Friends Aggregates
- **1f_plg**: Sum pledged by 1-hop friends
- **1f_bck**: Sum backers of 1-hop friends
- **2f_plg**: Sum pledged by 2-hop friends
- **2f_bck**: Sum backers of 2-hop friends
- **3f_plg**: Sum pledged by 3-hop friends
- **3f_bck**: Sum backers of 3-hop friends