Exploratory Data Analysis (EDA) Ideas for Hacker News Upvote Prediction

# 📈 Upvote Distribution & Temporal Patterns

* Histogram of upvotes (with log scale) — Since upvotes are likely heavily skewed.
* Cumulative distribution function (CDF) of upvotes — Helps assess thresholds (e.g. what % of posts get over 100 upvotes?).
* Time series plot of average upvotes per day — Useful for identifying long-term trends (e.g. HN popularity increasing?).
* Upvotes by time-of-day + day-of-week — Use a heatmap to show how post timing affects upvotes.

# 👥 User-Based Patterns

* Top users by number of posts / average upvotes per post / total karma.
* Distribution of user karma (if available).
* Relationship between user karma and upvotes per post — Maybe high-karma users get more attention.
* New users vs. established users — Are newer users treated differently?

# 📝 Text-Based Patterns

* Title length vs. upvotes — Consider both char and word count.
* Title lexical richness — Unique word ratio, readability (e.g. Flesch score).
* Top/bottom scoring n-grams — Which phrases correlate with high/low upvotes.
* Punctuation/symbols — Do titles with “?” or “!” perform differently?
* Stopword density / capitalisation patterns — May hint at spam or formal/informal tone.
* TF-IDF outlier detection — Find weird/off-topic titles.

# 🌐 URL and Source Domain

* Top linked domains and their average upvotes (e.g. github.com, nytimes.com).
* Self-posts vs. external links — Is there a difference in engagement?
* Domain categories — Tech vs. news vs. blogs etc. (use regex or a lookup to group domains).

# 🧵 Post Structure / Metadata

* Story vs. comment — Compare distributions.
* Descendants (comment count) vs. upvotes — Do more-discussed posts get more upvotes?
* Post age when scraped / time until peak upvotes — If you have timestamps.
* Deleted or dead posts — Check for anomalies.

# 🧪 Data Quality Checks

* Missing values — Especially in URL, title, user, etc.
* Duplicated posts or titles — Are there reposts?
* Title anomalies — Extremely short, empty, or repeated titles.

# 📦 SQL-Driven Ideas

* SELECT COUNT(\*) GROUP BY user — User post activity distribution.
* SELECT AVG(score), HOUR(timestamp) — Temporal upvote patterns.
* SELECT domain, COUNT(\*) — Most common external sources.
* SELECT title FROM posts WHERE score > X — Manual inspection of viral posts.