Business Problem Statement:

A major streaming platform (OTT) is facing increasing competition and rising customer acquisition costs. To maintain profitability, the platform must focus on **viewer retention** and reducing "churn" (when a viewer stops watching a series or cancels their subscription). Management has observed that viewer drop-off is not random but appears linked to specific content attributes such as episode pacing, dialogue density, and "hook strength".

The Problem

While the platform has a vast library of shows across genres like Sci-Fi, Drama, and Crime, a significant number of viewers drop off after only a few episodes. The production team needs to identify which technical and creative factors (e.g., pacing score, visual intensity, or cognitive load) are leading to high **retention risk** and **drop-off probability**. Without understanding these patterns, the platform risks investing millions in content that fails to keep viewers engaged through an entire season.

Key Objectives

1. **Identify Retention Drivers:** Determine which factors (e.g., high hook strength vs. low pacing) most significantly correlate with a viewer completing an episode or season.
2. **Segment High-Risk Content:** Identify specific genres or show types (e.g., Sci-Fi vs. Documentary) that exhibit the highest drop-off probabilities to prioritize for creative intervention.
3. **Optimize Content Design:** Answer the overarching question: *"How can we leverage engagement metrics like 'avg_watch_percentage' and 'skip_intro' behavior to predict and prevent viewer drop-off in future productions?"*

Success Metrics

The project will be considered successful if the model and analysis:

- Accurately identify high-risk content with elevated drop-off probability.
- Achieve meaningful predictive performance (e.g., ≥70% accuracy or AUC) in classifying viewer retention risk.
- Provide actionable insights that can guide content design decisions such as optimal pacing, hook strength, and episode structure.

Project Deliverables

To address this business problem, the following five deliverables are required (modeled after the retail project structure):

1. **Data Preparation & Modeling (Python):** Clean the ott_viewer_dropoff_retention_us_v1.0.csv dataset, handling missing values and engineering features like "Engagement-to-Duration Ratio".
2. **Data Analysis (SQL):** Execute queries to extract insights on average drop-off rates by platform (Netflix vs. HBO Max vs. Hulu) and analyze the impact of "night_watch_safe" status on retention.
3. **Visualization & Insights (Power BI):** Develop an interactive dashboard that maps drop_off_probability against visual_intensity and dialogue_density to help producers see exactly where shows lose their audience.
4. **Report and Presentation:** Summarize findings on what makes a "sticky" show and provide actionable recommendations for showrunners to improve episode hooks.
5. **GitHub Repository:** Compile all analysis scripts, SQL queries, and the dashboard file into a structured repository for the data science team.