**Antenna Data Science Exercise** **June 2025**

**Instructions:**

At Antenna, we help well-known brands, studios, and advertisers navigate the competitive streaming landscape. A key area of interest is understanding user engagement with different subscription tiers on major platforms. For instance, Netflix and Hulu both offer **ad-supported plans** (typically lower-priced subscriptions that include commercial interruptions during content playback) and **ad-free plans** (typically higher-priced subscriptions that offer uninterrupted viewing).

Your task is to explore the provided dataset (data.csv), which contains TV viewing sessions from various services, including Netflix and Hulu. Your goal is to develop a robust and well-reasoned solution that leverages this 3rd-party data to understand, infer, or model viewing patterns specifically on **Netflix and Hulu**, with the aim of **distinguishing between ad-supported and ad-free viewership**. This understanding would be valuable to Antenna's clients who want to understand the ad-supported vs. ad-free landscape of these major players.

A core challenge you should consider is that ad-supported viewership will likely exhibit gaps or fragmented sessions in the viewing data. These gaps could represent actual ad breaks. However, they could also be due to natural user pauses (e.g., bathroom breaks, fetching a snack), channel flipping (e.g., switching away from an app temporarily), plan switching (ie. being on ad-free plan and then due to rising costs switching to ad-supported plans), or inherent noise/limitations within TV data such as autoplay in some apps. Your approach should acknowledge this challenge: How might one attempt to distinguish between ad-induced gaps and other types of interruptions or data artifacts using Python-based analysis?

Your goal is to demonstrate through well-structured Python code, thoughtful analysis, and a clear solution design how this 3rd-party TV data can be used to infer ad-supported vs. ad-free viewing on Netflix and Hulu. Remember you are working with 3rd-party TV data, which does not explicitly label subscription plans. Your features should be derived from the session characteristics themselves. The emphasis is on your analytical approach, the quality and maintainability of your Python code, and your problem-solving process, rather than achieving a highly complex or predictive model.

**Files:**

The data required for the task is provided to you in the [same folder](https://drive.google.com/drive/folders/1GcZsAZR7oUMdJqybwXQzb8PGkIHj945a?usp=drive_link) as this document:

* data.csv
* schema

**Requirements:**

1. Propose a solution to understand or model viewing patterns on **Netflix and Hulu** that could help distinguish **ad-supported viewership from ad-free viewership**, using the provided 3rd-party TV data. This will involve analyzing session data for Netflix and Hulu from the provided dataset.
2. Develop a Python-based solution. This could be an analytical framework, a feature engineering pipeline, or a proof-of-concept model, implemented in well-structured, maintainable Python code. The goal is to showcase how 3rd-party data can be used to address the problem, demonstrating good coding practices suitable for a collaborative codebase.
3. Explain why your chosen solution fully solves the problem of identifying Netflix and Hulu's ad-supported viewership from ad-free viewership.
4. Address the challenge of distinguishing ad breaks from other session interruptions, outlining how your Python analysis attempts to tackle this.

**Definitions:**

* **1st-Party Data:** Data that is collected directly by a company about its own customers, users, or services through its own interactions and platforms. For example, Netflix or Hulu would have 1st-party data on which of their subscribers are on ad-supported vs. ad-free plans, their precise ad load, etc. **You do not have this explicit 1st-party plan information or ad-break labels in the provided dataset.**
* **3rd-Party Data:** Data collected by an external entity that does not have a direct relationship with the end-users. In this context, the data.csv provided by Antenna, containing TV viewing sessions for services like Netflix and Hulu, **is the 3rd-party data you will be using.** Antenna collects this data without having a direct relationship with the subscribers of Netflix or Hulu regarding their plan choices or ad interactions. Your task is to use this 3rd-party session data to engineer features from the observed session characteristics and infer patterns that might distinguish ad-supported viewing from ad-free viewing on Netflix and Hulu.

**Technical Approach:**

You may use any of the following tools:

* Python for data exploration, processing, feature creation, and modeling, using standalone Python scripts (.py files).
* SQL for initial data querying if preferred, but the core logic, feature engineering, and any modeling should be implemented and demonstrable in Python scripts.
* Google Slides for presentation building.

**Deliverables:**

* **Code Repository (Required):**
  + **A private** GitHub repository shared with the interview team.
  + Please share with alilakda, geogria, salomevianaAntenna, ssanchezantenna.
  + Include all standalone Python scripts (.py files) used for data exploration, processing, feature engineering, and any modeling. Jupyter Notebooks (.ipynb files) should not be submitted.
  + Ensure code is well-commented, organized, and demonstrates good software development practices (e.g., functions, classes where appropriate, considerations for testability, readability, and maintainability).
* **Presentation (Required, to be included in GitHub Repository):**
  + At most 5 slides (PDF or Google Slides link in your repository's README).
  + 10-minute presentation time.
  + Content should cover these themes:
    - Your proposed solution to identify/model ad-supported vs. ad-free viewership on Netflix and Hulu, and its value to Antenna's clients.
    - Your approach to distinguishing ad breaks from other interruptions, supported by your Python analysis.
    - Data visualization (generated via Python) that supports your hypothesis or approach.
    - The analytical or modeling approach used (or proposed in detail), implemented, or prototyped in Python.
    - Potential evaluation methods (if applicable) and expected impact of the insights.
    - Next steps or future work, including how your solution might be operationalized or integrated into a larger system.

**Evaluation Criteria:**

To ensure you are well-prepared for the team presentation, we are sharing the evaluation criteria that will be used during your interview. Below are the key criteria based on which the team will be evaluating your presentation and asking questions.

* **Emphasis on Problem-Solving, Analytical Thinking, and Solution Design:** Our primary interest lies in your thought process, problem-solving skills, and the clarity of your proposed solution. We want to see how you break down the problem, reason about potential approaches, and address challenges using the available data and Python. A highly complex or predictive model is not the main goal; a well-reasoned analytical framework or a proof-of-concept demonstrating your approach is more valuable.
* **Python Coding and Software Engineering Practices:** We will heavily evaluate the quality of your Python code. Is it well-structured, readable, commented, and maintainable? Does it show an understanding of good software development principles relevant to a collaborative data science environment (e.g., modularity via functions/classes, considerations for testing, reusability, and potential deployment)? This is critical as the role involves maintaining and enhancing a core Python codebase.
* **Effective Use of 3rd-Party Data & Python for Feature Creation:** We will evaluate your ability to creatively engineer features from the provided 3rd-party TV data using Python. Focus on extracting meaningful insights from session characteristics on Netflix and Hulu that could help infer ad-supported vs. ad-free viewing patterns.
* **Thoughtful Approach to the Ad-Break vs. Pause/Noise Problem:** We will assess how you conceptualize and propose to tackle the ambiguity of gaps in session data using analytical techniques implemented in Python. This includes your reasoning for why certain patterns might indicate ad breaks versus other interruptions.
* **Awareness of Data Leakage and Model Integrity:** Be prepared to discuss your approach, specifically how you have ensured that there is no data leakage if a model is proposed (e.g., using future information to predict past/current states). We will ask questions to understand your strategies for maintaining analytical or model integrity.
* **Effective Communication and Presentation Skills:** We will assess your ability to clearly present your findings, explain your methodology (including your Python code and its logic), and articulate the value of your proposed solution. Clarity, organization, and the ability to defend your approach are key components we will be looking for. Be prepared to answer questions about your work and discuss your decisions in depth.

For the 'Next steps or future work' part of your presentation, consider how your solution would be validated in a real-world scenario. At Antenna, we would typically evaluate such a model by comparing its inferred percentages against our ground truth data, which is derived from our proprietary transaction dataset. Discussing how you would use this known data to measure accuracy and improve your model would be a strong addition to your proposal.

**Antenna ground-truth data for plan distribution:**

| **Time Period** | **Netflix Ad-Supported** | **Netflix Ad-Free** | **Hulu Ad-Supported** | **Hulu Ad-Free** |
| --- | --- | --- | --- | --- |
| 2023-02 | 1.40% | 98.60% | 61.90% | 38.10% |
| 2023-05 | 2.30% | 97.70% | 62.90% | 37.10% |
| 2023-07 | 4.00% | 96.00% | 61.70% | 38.30% |
| 2023-10 | 6.40% | 93.60% | 61.20% | 38.80% |
| 2024-01 | 10.00% | 90.00% | 62.80% | 37.20% |
| 2024-04 | 13.00% | 87.00% | 62.60% | 37.40% |
| 2024-08 | 19.60% | 80.40% | 61.80% | 38.20% |
| 2024-12 | 27.80% | 72.20% | 62.90% | 37.10% |
| 2025-03 | 29.10% | 70.90% | 63.90% | 36.10% |

**Important Notes:**

* This work should be your original ideas. We repeat, we know when your ideas are not your own and generated by AI.
* That said, please make full use of Generative AI tools (e.g., ChatGPT, Gemini, Claude), but credit their usage in your code comments or README.
* You don't need to build a full, production-ready model; a proof-of-concept implemented in well-structured, maintainable Python scripts, a detailed analytical framework, or a well-reasoned proposal with supporting code is sufficient. The emphasis is on demonstrating your Python skills, analytical thinking, and code quality in the context of the problem.

Good luck, and we look forward to your innovative ideas!