Assignment: AI & Data Science Intern

Topic: The "Dark Matter" Bridge – Volume Estimation & Probabilistic Attribution

Part 1: Strategic Context

- **The Situation:** We are **Gravton OS**. We build the Operating System for Customer Journeys.
- **The Problem:** Traditional Analytics (Google Analytics) tracks "Visible Light" (Keywords & Clicks). It cannot track "Dark Matter" (AI Prompts & Answers).
  - When a user searches "Creta Price," we see it.
  - When a user asks ChatGPT, *"Compare the safety of Creta vs. Grand Vitara for a family of 4,"* it is invisible.
- **The Consequence:** Brands see massive spikes in "Direct Traffic" (users typing the URL directly or arriving from unknown sources) and have no idea *why* those users arrived. We believe this is actually **AI-Driven Traffic**.
- **The Challenge:** You must build a **Probabilistic Attribution Model** for an SUV client. You will estimate the hidden volume of AI conversations and use that data to "decode" mysterious Direct Traffic on the client's site.

Timeframe: 48 Hours.

Deliverables:

1. **Code:** Colab Notebook/GitHub link containing two distinct modules (Volume Estimation & Attribution).
2. **Memo:** A 2-page PDF explaining your logic, defending your criteria, and proposing innovations.

Part 2: Coding Task 1 – The "Calibration" (Volume Estimation)

Objective: Build a dataset of Intent Clusters and estimate their "Hidden" AI Volume.

Step 1: The "Dark Matter" Generation

- Use an LLM to generate **1,000 unique user prompts** related to "buying a Compact SUV in India" (Tata Nexon, Hyundai Creta, Maruti Grand Vitara, Kia Seltos).
- **Constraint:** Ensure diversity. We need simple queries ("Boot space of Nexon") and complex reasoning ("Which SUV has the best suspension for my parents' back pain?").

Step 2: The "Visible Light" Mapping

- Map each prompt to a **Parent Google Keyword**.
- Assign a simulated **Monthly Search Volume (MSV)** to that keyword (using a log-normal distribution: many low-volume keywords, few high-volume ones).

Step 3: The "Merging" Logic (The Core Task)

- **The Problem:** We know the Google Volume. We *don't* know the AI Volume.
- **The Task:** Create a scoring function that combines **Search Volume** + **Synthetic Signals** to estimate **AI Request Volume** (as realistically close as possible to the actual search).
- **Code Requirement:** Implement a function `estimate_ai_volume(google_msv, complexity_score, ...)` that outputs a predicted volume for every prompt.
  - You must define the "Merging Criteria." How much weight does *Search Volume* carry vs. *Prompt Complexity*?
- **Output:** Cluster the 1,000 prompts into **15–20 Intent Clusters** and calculate the **Total Estimated AI Volume** for each cluster.

Part 3: Coding Task 2 – The "Inference" Engine (Probabilistic Attribution)

Objective: Connect your estimated volume to real-world website behavior using probability.

The Scenario:

You have a "Session Log" of 25 Users who arrived on the client's website via Direct Traffic (Unknown Source). You must infer which Intent Cluster (from Part 2) likely drove them there.

The Data (Use this table for your analysis):

ID	Landing URL	Time on Page	Scroll Depth	Device	Next Action
S01	/blog/safest-cars-india-ncap-2025	5m 12s	90%	Mobile	Click: "Tata Nexon"
S02	/model/hyundai-creta-sx	0m 30s	15%	Desktop	Bounce
S03	/finance/emi-calculator	2m 45s	100%	Desktop	Download Quote
S04	/compare/brezza-cng-vs-petrol	4m 10s	85%	Mobile	Share on WhatsApp
S05	/model/mahindra-thar-rox	0m 10s	5%	Mobile	Bounce
S06	/blog/ev-charging-infrastructure-delhi	6m 20s	95%	Desktop	Click: "Nexon EV"
S07	/model/kia-seltos	4m 50s	80%	Tablet	Check Color Options
S08	/brochure-download/xuv700	1m 05s	N/A	Desktop	Exit
S09	/model/maruti-fronx	0m 05s	0%	Mobile	Bounce
S10	/blog/best-automatic-cars-under-10lakhs	3m 15s	70%	Mobile	Click: "Amaze CVT"
S11	/compare/grand-vitara-vs-hyryder	7m 30s	90%	Desktop	Read Reviews
S12	/showroom-locator/bangalore	0m 45s	40%	Mobile	Click: "Call Now"

S13	/model/tata-safari	0m 20s	10%	Desktop	Bounce
S14	/blog/sunroof-pros-cons	2m 10s	60%	Mobile	Exit
S15	/model/mg-hector	5m 45s	85%	Desktop	Book Test Drive
S16	/service/spare-parts-cost	1m 30s	50%	Mobile	Search: "Wiper Blade"
S17	/model/hyundai-venue	0m 15s	10%	Desktop	Bounce
S18	/blog/diesel-vs-petrol-resale-value	4m 00s	80%	Desktop	Click: "Used Cars"
S19	/model/scorpio-n	0m 12s	5%	Mobile	Bounce
S20	/compare/creta-vs-seltos-vs-elevate	8m 10s	100%	Desktop	Click: "Elevate"
S21	/model/tata-punch	3m 50s	75%	Mobile	View Gallery
S22	/homepage	0m 03s	0%	Desktop	Bounce
S23	/offers/festive-discount	1m 10s	100%	Mobile	Click: "Callback"
S24	/model/toyota-fortuner	0m 40s	20%	Mobile	Bounce
S25	/blog/adas-features-explained	6m 00s	90%	Desktop	Click: "XUV700"

The Logic Challenge:

For a specific session (e.g., S01), calculate the probability that it originated from Cluster X.
P(Cluster | Session)

Your code must implement a logic that weighs two factors:

1. **Your Volume Estimate:** Is *Cluster X* high volume? (If 10,000 people ask AI about "Safety," it is statistically more likely to be the source than a niche topic with 10 users).
2. **Behavioral Fit:** Does the session behavior match the intent? (e.g., A "Deep Research" intent should correlate with *High Time on Page*. A "Price Check" intent might correlate with *Low Time on Page*).

Deliverable:

A final dataframe assigning the Most Likely Intent Cluster to each of the sessions provided above, along with a "Confidence Score."

Part 4: The Memo (The "Why" & "What's Next")

Length: Max 2 Pages.

Focus: Defense of Logic & Innovation.

Section 1: Defending the Merging Criteria (Volume Estimation)

- Explain your formula for estimate_ai_volume.
- Why did you choose the specific weights you did? (e.g., "Why is a complex query with low search volume rated higher than a simple query with high search volume?")
- **Innovation:** What *other* data points (beyond Search Volume and Text Complexity) could we fetch to make this estimation more accurate in the future?

Section 2: Defending the Attribution Logic

- Explain how you mapped "Session Behavior" (Time, Scroll) to "Intent Clusters."
- How did you handle the trade-off between the **Prior** (Volume) and the **Likelihood** (Behavior)? (e.g., What happens if a cluster has huge volume but the behavioral fit is poor?)
- **Innovation:** If you had full access to the client's telemetry, which *new* behavioral signals would you track (and how) to improve this attribution model? (e.g., Copy-paste behavior, specific interaction events, etc.)

Section 3: Action Gap

- Based on your prompt clustering and session logs, identify **ONE content gap**.
- Where do you see high "Prompt Density" (lots of people asking AI about a topic) but poor "Session Performance" (users bouncing or finding no matching landing page)?
- What possible actions can be taken to improve user behaviour or AI search ranking for the product/brand ?

Section 4: (IMP) Future Development (if no time limit)

- If there is no time limit for implementation, what would you do to estimate the ai search prompt volume as close to the actual ai search volume as possible?
- How can you verify the “black box” ai search volume in the real world?
- Discuss in detail in the post assignment interview.

Part 5: Grading Rubric (How to Win)

- **Merging Logic (Part 2):** Did you blindly assume "High Google Volume = High AI Volume" (Fail), or did you build a nuanced model that accounts for complexity and intent?
- **Probabilistic Thinking (Part 3):** Did you successfully treat "Volume" as a *Prior*? Your code should reflect that a high-volume cluster exerts more "gravity" on unknown traffic than a low-volume one.
- **Innovation (Memo):** We are looking for creative thinkers. Did you propose realistic, high-value data sources or signals to improve the model? What would you do different if there is no implementation time limit?
- **Clarity:** Can you explain your complex probabilistic model in simple business terms?

AI Usage Policy

We encourage responsible AI use. You may use AI for research, validation, and brainstorming, and even rapid prototyping if you can manage and generate tasteful output and refine that further until it appears natural and non-GPTish. However, we do not accept rote answers where AI generates the entire strategy. You must be able to justify your decisions at each step and for each word.

- If you use AI, please include a brief note on how it aided your process.
- Requirement: Please share any prompts or chat links in an appendix.