HAIRCARE PROJECT DATA ANALYSIS CATALOGUE

Study Focus: Scalp Health & Brand Loyalty in Shampoo Usage
Analyst:Aradhya | Brand Studied: Head & Shoulders | Use Case: Haircare Market Strategy Optimization

---

A. TECHNICAL FRAMEWORK – DATA ANALYST VIEW

🔧 MODULE 1: Data Ingestion & Preprocessing

Objective: Create a robust backend pipeline for structured and unstructured data.

- Schema Design: PostgreSQL with 5 main tables — respondent_profile, product_feedback, open_ends, segmentation_labels, tracking_data.
- Python ETL Pipeline: Using Pandas + SQLAlchemy for dynamic ingestion.
- Unique ID Generation: UUID4 applied across all respondent entries to avoid collision.
- Screener Logic Validation:
  - Q1: No shampoo → TERMINATE
  - Q4: If Female → TERMINATE (target: male)
  - Q5: If Age <18 or >45 → TERMINATE
  - Q9: If no brand recall → TERMINATE
- Final Sample Post-Cleaning: N = 1,872 (98.6% retention rate)

🧼 MODULE 2: Structured Cleaning & Categorical Coding

Objective: Standardize and prepare categorical data for analytics.

- Recoding: Gender, Age Groups, NCCS, Shampoo Frequency
- Derived Metrics: Days since last purchase, loyalty scores, repeat usage flags
- Quota Monitoring: Soft (age bands) and Hard (NCCS A1–A3 only)
- Final Variables: Gender (1/0), Age_Group (2/3/4), NCCS (1/2/3)

🧠 MODULE 3: NLP Text Pipeline

Objective: Normalize and structure open-text for analysis.

- Tools Used: spaCy (lemmatization), SymSpell (spell correction), custom dictionaries
- Embeddings: Sentence-BERT vectors stored for further modeling
- Output: 600 cleaned verbatims, indexed by respondent_id

📊 MODULE 4: Descriptive & Inferential Statistics

- Demographics Profiling: Age × NCCS × Frequency
- Chi-Square Tests: Severity × Frequency (Significant at $p < 0.01$)
- ANOVA: Usage frequency ~ Severity Level
- Loyalty Funnel Metrics: Awareness → Usage → Repeat → Advocacy

🤖 MODULE 5: Machine Learning Models

- Segmentation: K-Means (k=5) → Persona clusters like “Frequent Loyal Warriors”
- Predictive Models: XGBoost (AUC: 0.86) to predict purchase intent (Q43)
- Uplift Modeling: CausalML → High uplift in 18–22, NCCS A2 for new matte pack
- Churn: CoxPH → Negative post-wash sentiment = 2.3x churn risk

💬 MODULE 6: NLP Modeling

- Topic Modeling: BERTopic → 10 core themes incl. “cooling,” “residue,” “non-itchy”
- Aspect-Based Sentiment: BERT → Fragrance (+0.62), Residue (–0.31)
- Emotion Detection: GoEmotions → Trust (28%), Joy (23%) drive purchase

📈 MODULE 7: Visualization & Dashboard

Tool: Power BI

- Tabs:
  - Persona Segmentation
  - Brand Funnel Drop-offs
  - Sentiment Heatmaps

    - Churn Predictor
- Visuals: Sankey Journey Map, Radar Charts, UMAP

🧩 MODULE 8: Psychological & Emotional Intelligence

- LIWC: High “Conscientiousness” → Loyalty
- Maslow Mapping: “Esteem + Belonging” drivers
- Personality Trait Inference: Big 5 modeling from open-ends