Tourism Reviews Analysis – Methodology, Findings, and Recommendations

This report documents the end-to-end approach, insights, and suggested actions from building and analyzing a tourism-reviews data product. It covers the methodology for data ingestion and transformation, text cleaning and translation, sentiment/aspect extraction, exploratory data analysis (EDA), and strategic recommendations to improve customer experience based on the results.

- Repository: artefact-assessment
- Technology stack: Dagster (orchestration), Supabase/PostgreSQL (storage), Google Cloud Storage (source), FastAPI (microservices), OpenAI & PyABSA (NLP), Docker Compose (local run)

Overview

Business goal: convert raw multi-language tourism reviews into trusted, queryable insight-covering destinations, offerings, and fine-grained aspects and sentiment so operators can see what customers love or struggle with and act quickly.

Key outcomes:

- Repeatable pipelines to ingest, parse, translate, and enrich reviews with sentiment insights
- Aspect-based sentiment at scale (two approaches: LLM and PyABSA) with evidence spans
- Analytics-ready tables, views, and materialized views in Supabase
- EDA of ratings, keywords, and aspect polarity trends to surface opportunities

Methodology

Data ingestion and modeling

Data source: CSV reviews and JSON tag mappings from Google Cloud Storage (GCS).

- Ingestion job (`ingestion_pipeline`):
  - Triggered when new reviews or mappings are uploaded to GCS
  - Load dataset from GCS (`load_dataset_from_gcs`)
  - Transform minimally to review records (`transform_reviews_basic`)
  - Upsert to Supabase PostgreSQL (`load_reviews_to_supabase`)
- Mapping job (`parsing_pipeline`):
  - Load tag-to-destination/offering mappings from GCS (`load_mappings_from_gcs`)

  - Upsert dimensions (`destinations` , `offerings`)
  - Parse raw_tags into `destination_id` and `offering_id` (`parse_review_tags` , `update_parsed_reviews`)
- Schema (see `infra/supabase/init.sql`):
  - `reviews(id BIGINT, content, title, date, language, rating_normalized, rating_raw, destination_id, offering_id, raw_tags, translated_content, translated_title, created_at, updated_at)` with indexes on foreign keys, date, language, and ratings; trigram indexes for full-text are attempted with graceful fallback
  - `destinations(id, name)` and `offerings(id, name)` with uniqueness constraints
  - Analytics view `review_analytics` and materialized view `review_summary_stats` + refresh function

Operational notes:

- IDs: numeric ids derived from CSV id fields (robust to string ids, de-duplicated via upsert)
- Batch inserts via `psycopg2.execute_batch` for performance and idempotency

Text cleaning and translation

Use cases: unify multilingual reviews into English to enable consistent downstream analysis.

- Candidate text for analysis: COALESCE(translated_content, content)
- Translation job (`translation_pipeline`):
  - Load batches of non-English reviews (`load_reviews_for_translation`)
  - Translate content and titles via OpenAI in parallel (`translate_reviews`) with configurable model and thread pool, then update rows (`update_translated_reviews`)
- Cleaning choices in EDA (`submissions/EDA1.ipynb`):
  - Lowercasing, URL removal, punctuation stripping, whitespace normalization
  - Tokenization + English stopword removal
  - Derived features: word/char counts, token counts

Assurances and constraints:

- Rate-limiting and batching to respect API quotas
- Translation results stored to avoid repeated API calls

Aspect-based sentiment extraction

Two interchangeable services expose the same response schema and are integrated via Dagster ops.

1. LLM-based aspect service (`services/aspect-api`)

- FastAPI endpoints: `/health`, `/extract`, `/extract-batch`
- Prompting: system and user instructions bias toward concise, non-overlapping aspects with explicit polarity and evidence spans; strict JSON response enforced
- Resilience: retries with exponential jitter; explicit handling of rate-limit, upstream, and auth errors (Tenacity + OpenAI client)
- Normalization: ensures `category` and `confidence` keys present; returns `meta` with model, prompt_version, latency, tokens estimate
- Security: header key `X-API-Key` required; secrets via environment or Secret Manager

2. PyABSA-based service (`services/pyabsa-api`)

- FastAPI endpoints: `/health`, `/extract`, `/extract-batch`
- Model: multilingual ATEPC; container pre-loads checkpoint to reduce cold-start; CPU by default, GPU if available
- Output mapped to the common schema, including evidence spans reconstructed from token positions and confidence
- Operational sizing guidance included (Cloud Run resources)

Data persistence and de-duplication (`pipelines/dagster/ops/aspect_extraction.py`):

- Table `review_aspect_extractions`: `review_id`, `aspect`, `evidence_span`, optional `category`, `start_idx`, `end_idx`, `confidence`, `model`, `prompt_version`, `polarity`, `approach`
- Unique index to dedupe by `(review_id, aspect, evidence_span, prompt_version, model)`
- Evidence indexing: locate `evidence_span` in source text case-sensitively, then case-insensitively; store start/end positions when found
- Caching: API cache table supported with versioned keys `(model|prompt_version|text)` to skip repeated extraction
- Batch size: up to 32 per HTTP call (service limit), with fault-tolerant fallbacks and per-batch timing telemetry

Orchestration, environments, and runtime

- Jobs are defined in `pipelines/dagster/jobs.py`; repository wires sensors and chaining in `repository.py`
- Optional GCS-triggered Pub/Sub sensors chain ingestion to parsing to translation to aspect extraction (LLM and PyABSA in parallel)
- Local run via Docker Compose (`dagster-daemon` and `dagster-webserver`) pointing to Supabase Cloud; secrets come from `.env` variables mounted into containers (see `docker-compose.yml`)
- All ops produce `AssetMaterialization` events for observability (counts, durations, success rates)

Findings

This section compiles insights from two EDA notebooks and the enriched aspect store. Where relevant, we cite the notebook stage that produced the view.

A. Dataset and rating/volume patterns (EDA1)

- **Language mix**: the dataset contains both Arabic and English; translation coverage is material and required for full-scope analysis
- **Rating distribution**: generally right-skewed (more positive than negative); mean and median skew to the positive range
- **Top destinations**: major religious destinations and large metropolitan cities dominate volume; emerging destinations present meaningful but smaller cohorts
- **Offerings**: “Tourism Attractions / Sites” leads volume; “Accommodation” and other service categories vary in both volume and average rating
- **Review length**: most reviews are concise; negative reviews tend to be more detailed.
- **Common keywords** (post-cleaning, English): quality, service, cleanliness, location, staff, family, price/value-themes consistent with hospitality and attractions
- **Temporal trend**: volume varies month-to-month; identifiable peaks suggest seasonality (holidays, pilgrimage periods) that should inform staffing and operations

Selected highlights:

- Religious sites show highest average ratings; attractions dominate total review count
- Accommodation exhibits greater variance and generally more detailed texts

B. Aspect and sentiment patterns (EDA2)

- **Aspect normalization**: consolidating plural/singular and formatting variants notably reduces unique aspect terms caused by LLM extracted aspects
- **Polarity distribution**: aspect-level mentions are predominantly positive, with neutral and negative providing targeted signals for improvement
- **Most-mentioned aspects (normalized)**: frequently include cleanliness, staff/service, facilities, price/value, crowd/queue, location, signage, accessibility
- **By offering**: attractions typically show high positivity for aesthetics and activities; accommodation sees more negative share on cleanliness, noise, check-in/queue, and amenities consistency
- **By destination**: most high-volume destinations are net-positive, with a few showing higher negative share on crowding, traffic/parking, or facility maintenance
- **Evidence spans**: negative examples often include precise phrases about wait times, cleanliness issues, availability/stock-outs, or unclear guidance (signage/policies)
- **Temporal trend**: aspect volume moves with review volume; negative share can trend upward during peak seasons (crowding-driven) and decline post-peak

Selected Highlights:

- The aspect store provides specific, evidence-backed, and aggregable signals for business reviews and continuous improvement funnels
- Combining `approach` and `model` fields enables A/B comparisons of extraction methods for trust and coverage over time

Recommendations

The actions below prioritize the issues that appear most frequently and/or carry the highest negative share across offerings and destinations. They are framed as concrete playbooks with measurable targets so progress can be tracked week over week.

Priority focus areas (what to fix first)

1. Price perception and transparency (highest negative volume)

- Actions
  - Eliminate surprise fees; publish all-in prices early in the journey (website, ticketing, menus).
  - Introduce value anchors: bundles, off-peak discounts, family passes; show “from” pricing by date.
  - Standardize refund/exchange policies and display them prominently.
- Metrics (targets per quarter)
  - Price-related negative share reduction 30%.
  - Chargeback/refund disputes reduction 25%.
  - Cart abandonment on pricing step reduction 10%.

2. Service consistency (second highest negative volume; notably weaker at attractions/activities)

- Actions
  - Define a simple “Service Charter” (greeting, wait-time updates, resolution path in ≤10 minutes).
  - Staff-to-crowd surge playbook: pull flex staff when queue > threshold; rotate “floor captains.”
  - Empower quick recovery: comp small items or upgrades when delays occur; capture recovery outcomes.
- Metrics
  - Service-related negative share reduction 25% overall; reduction 35% at attractions.
  - Average queue-time perceived in reviews reduction 20% (via aspect mentions around waiting/attention).

3. Cleanliness and facilities (high mention volume; restrooms/bathrooms/rooms recur in negatives)

- Actions
  - Peak-aware cleaning SLAs: restrooms every 20–30 minutes during peak; public areas hourly.
  - “Last cleaned at” signage with QR to report issues; fast triage queue for incidents.
  - Room and bathroom checklists (corners, linens, fixtures) with digital verification.
- Metrics
  - Cleanliness-related negative share reduction 30%.
  - Time-to-clean incident (90th percentile) ≤ 25 minutes at peak.

4. Parking, crowding, and wayfinding (mid/high negative volume and very high negative share when present)

- Actions
  - Real-time parking guidance (signage/app) + overflow/shuttle plans on peak days.
  - Timed entry and/or virtual queues at high-demand sites; cap capacity per slot.
  - Wayfinding uplift: more multilingual signs; pre-arrival “how to navigate” tips; staff at choke points.
- Metrics
  - Parking/crowding negative mentions reduction 30%.
  - Average waits at peak reduction 20%; missed-entry complaints reduction 40%.

5. Furniture/amenities maintenance (concentrated negatives in accommodation and F&B)

- Actions
  - 90-day refresh plan for high-wear items (beds, seating, lighting, AC noise, breakfast variety).
  - Close-the-loop work orders: from guest complaints to ticket to resolution proof.
- Metrics
  - Amenities-related negative share reduction 25% in accommodation and F&B.

6. Organization/signage in retail and busy venues (elevated negative share in retail; disorientation complaints)

- Actions
  - Clear queue lines, mobile POS at peaks, staff marshals for flow.
  - A/B test in-venue floor plans and sign density; optimize for dwell time and frictionless exits.
- Metrics
  - Organization-related negatives reduction 25% in retail; conversion rate ↑ 5% at peaks.

Offering-specific playbooks

- Accommodation (highest overall negative share)
  - Deepen housekeeping QA; publish micro-scorecards at front desks.
  - Accelerate check-in: pre-ID capture, mobile keys, separate desk for issues vs. standard check-ins.
  - Tackle “furniture/room size/bathroom” complaints with a rolling room-refresh program.
- Food & beverage
  - Menu price clarity and “value bundles”; manage peak seating with table timers only if needed and compassionately.
  - Back-of-house SLAs for cleanliness; rapid bussing; visible standards in dining areas.
- Retail
  - Floor captains at peaks; additional mobile checkout; tidy cadence; clearer signage to reduce confusion.
- Religious sites (lowest negative share)
  - Preserve strengths: maintain cleanliness cadence, crowd stewardship, and multilingual guidance.

Measurement and operating rhythm

- Aspect Health dashboard (weekly)
  - Track: aspect volume, positive/negative share, top 10 negative aspects, negative share by offering/destination, and example evidence.
  - Alert on spikes: any aspect–destination combo crossing a negative-share threshold with $n \geq 30$.
- Seasonal readiness checklist (monthly in peak quarters)
  - Staffing models, queue thresholds, restroom cadence, parking overflow, signage audits.
- Continuous extraction QA
  - Compare multiple extraction approaches for coverage and polarity agreement; sample 50 negatives monthly for human QA; iterate prompts/config.

Delivery timeline

- 0–30 days (quick wins)
  - Publish price inclusions and refund policy; add “last cleaned at” restroom signage with QR; enable queue-time announcements; stand up Aspect Health dashboard.
- 30–90 days (core systems)
  - Timed entry/virtual queue pilots; parking guidance + overflow shuttles; room-refresh wave 1; frontline service charter roll-out; mobile check-in.

- 90+ days (structural)
  - Facility upgrades (lighting, furniture, wayfinding), bundle pricing strategy at scale, automated negative-spike alerting across all destinations.