IE7275: Data Mining in Engineering (Fall 2025)

Project

Sec 04 Group 8: Aarathi Saranya Pandravada & Chirag Mahaveer Chivate

November 8, 2025

Project Overview & Scope

In our proposal, we set out to build two complementary recommendation tasks: **(A) Next Point of Interest (Next-POI)** for one-step-ahead prediction, and **(B) Itinerary Recommendation** for k sequential stops.

1 Task A: Next-POI

1.1 Dataset Description and Preprocessing Steps

What & why. We use the Foursquare NYC & Tokyo check-ins datasets. Each record is an anonymous user's visit to a POI with `userId`, `venueId`, timestamp, latitude/longitude, and category. This structure is ideal for Next-POI because it preserves visit *order*, time patterns, and geographic movement.

Raw data (sanity checks). Table 1 summarizes the core sizes and geo sanity checks we rely on downstream features (time-of-day, distance, short-hop locality).

Preprocessing pipeline (what we did & why).

1. **Canonical local time (DST-aware):** Convert `utcTimestamp` to NYC (America/New_York) or Tokyo (Asia/Tokyo) local time; keep a `tz_offset_min` column. *Why:* correct hour/day/weekend features; avoids session leakage.
2. **Stable venue attributes:** For each `venueId`, compute median `lat_med`/`lon_med` and mode category; attach `dist2med_km` and flag venues with spread >5 km. *Why:* robust coordinates; reliable distance features.

	NYC	**Tokyo**
Check-ins (rows)	227,428	573,703
Unique users	1,083	2,293
Unique venues	38,333	61,858
Lat (1–99%)	40.587–40.964	35.531–35.841
Lon (1–99%)	−74.216–−73.736	139.488–139.873
Points inside 1–99% box	96.00%	96.07%
Per-venue max spread, p99 (km)	0.13	0.05
Venues with spread > 5 km	10 / 38,333	5 / 61,858

Table 1: Core dataset stats and geo sanity checks from our EDA.

3. **Sessionize and label for Next-POI:** Per user, sort by local time; new session if gap >6h; keep sessions with length ≥ 2. For step t, the label is `next_venueId` at t+1. *Why:* turn streams into supervised examples.

4. **Compact features (model-agnostic):** IDs (user, last_venue), time (hour, dayofweek, weekend, offset), geo (last_lat/lon, `dist_to_prev_km`), dynamics (`time_since_prev_h`). *Why:* capture routine, locality, and short-term momentum.

5. **Leakage-free splits:** Chronological 70/10/20 per user into train/val/test. *Why:* avoid look-ahead; fair evaluation.

1.2 EDA Results with Visualizations

Goals. Validate spatial/time quality and understand behavioral patterns that will drive features, candidate generation, and fair evaluation: (i) coordinate sanity and per-venue spatial stability, (ii) user activity distribution, (iii) venue/category concentration and taxonomy reliability.

What we saw & why it matters.

- **Coordinates are clean and compact.** No missing/invalid lat–lon for NYC or Tokyo. In NYC, 96.0% of points fall inside the 1–99% lat×lon box; in Tokyo it's 96.07%. *Why:* confirms city footprints are well-bounded and supports geo-aware features (distance, radius candidates).

- **Per-venue locations are stable (rare wanderers only).** NYC: per-venue max spread p99 = 0.13 km; only 10/38,333 venues exceed 5 km. Tokyo: p99 = 0.05 km; only 5/61,858 exceed 5 km. *Why:* lets us use each venue's median coordinates as its canonical point; a tiny outlier list can be flagged or fixed; distance features are trustworthy.

- **User activity is heavy-tailed.** NYC and Tokyo both show many casual users and a smaller set of power users (thousands of check-ins). *Why:* to avoid dominance, we'll report macro-averaged metrics per user (micro as secondary), and chronological per-user splits are feasible given ample history.

- **Venues are long-concentrated; we inspect tails.** We include a log-x histogram of visits per venue and a "top venues" bar to gauge concentration and check identity consistency (one location/category per ID). *Why:* informs candidate/negative sampling and any min-frequency screen.

- **Categories: use `venueCategoryId` as the key; Tokyo is transit-dominated.** Category IDs map cleanly; some names have variants (e.g., "Airport"). Tokyo shows a mega-class ("Train Station") with extreme skew. *Why:* use `venueCategoryId` for modeling; consider a coarser category view only for diversity diagnostics; ensure macro averaging and include coverage/diversity metrics so big classes don't mask tail performance.

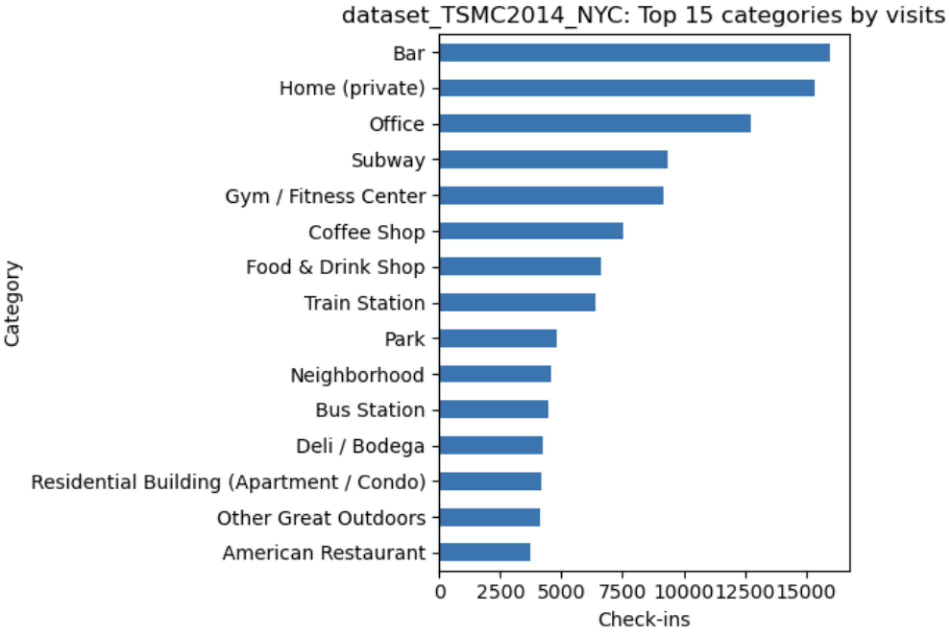

Figure 1: NYC top categories

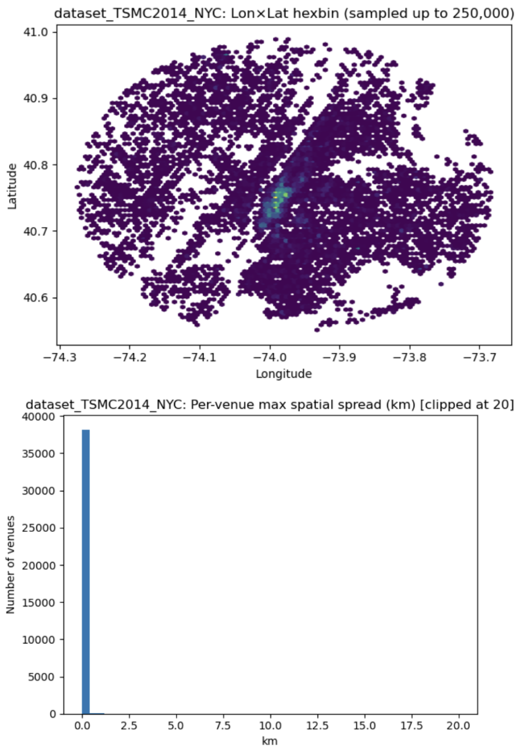

Figure 2: NYC footprint and per-venue spread summary

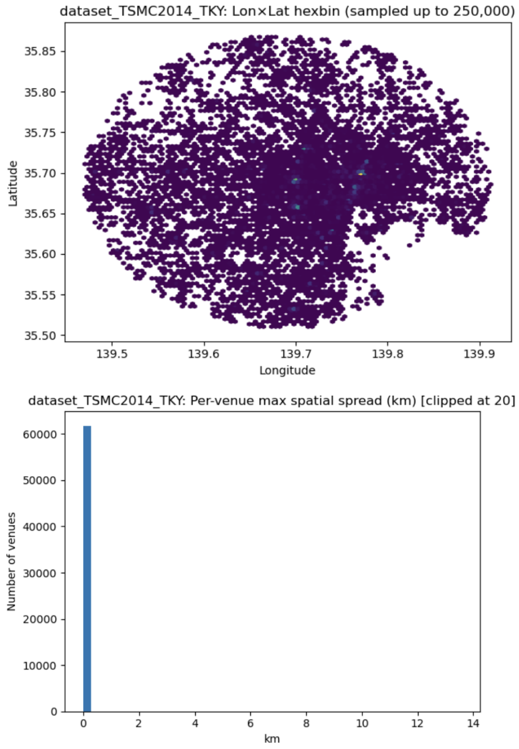

Figure 3: Tokyo footprint and per-venue spread summary.

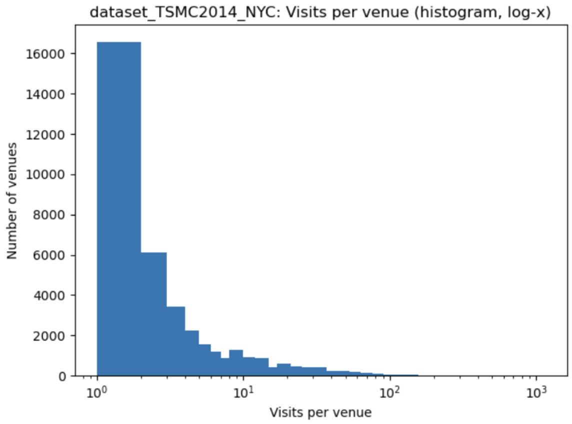

Figure 4: NYC: visits per user.

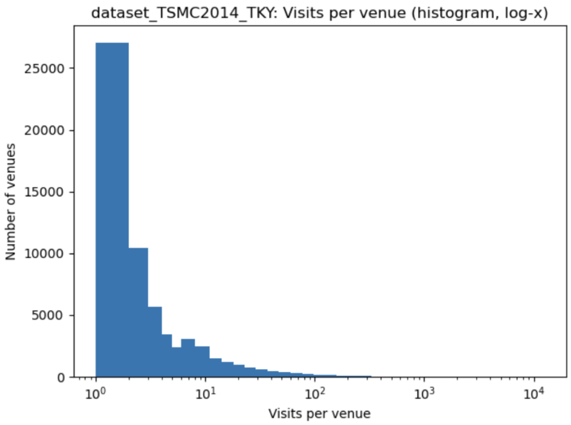

Figure 5: Tokyo: visits per user.

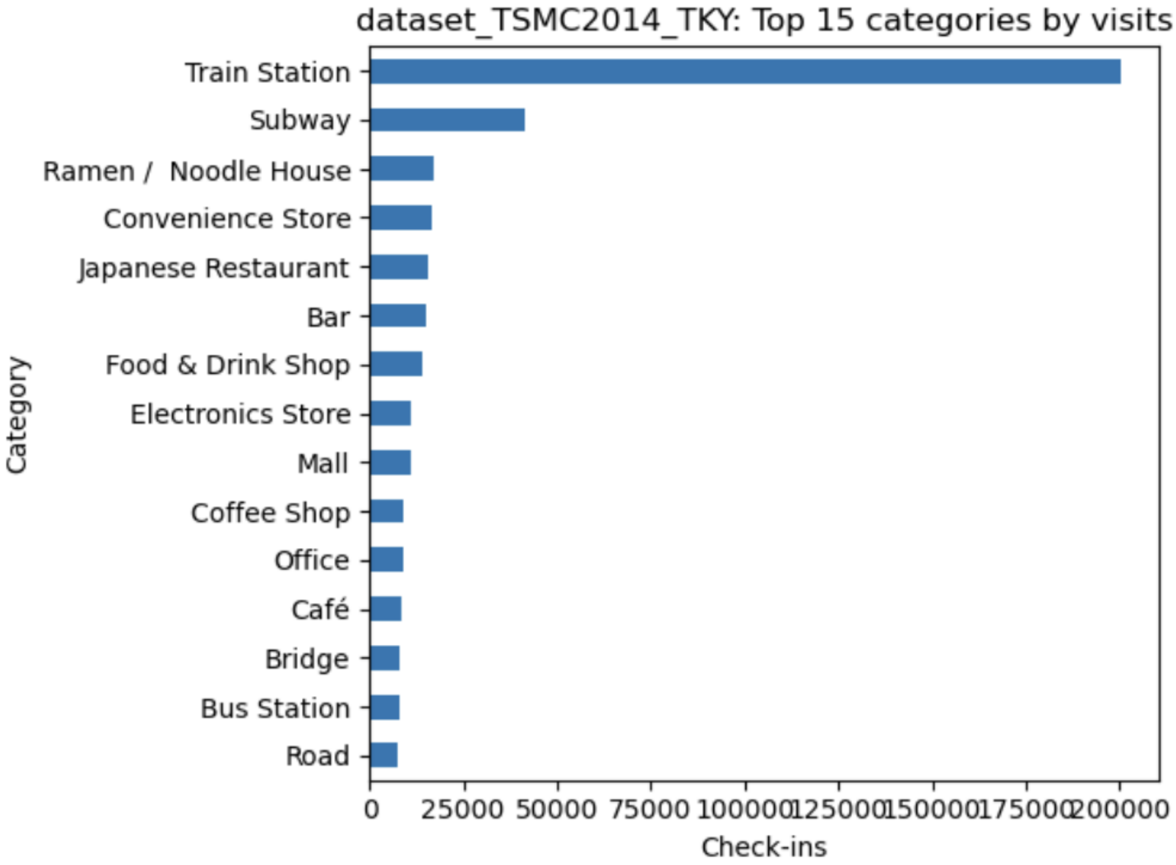

Figure 6: Tokyo top categories

1.3 Model Design and Rationale

Design goals. Task A (Next–POI) is a top-K ranking problem where the next venue should be (i) relevant to the user, (ii) consistent with their immediate sequence of stops, and (iii) geographically plausible. Our proposal committed to a transparent baseline and a sequence-aware advanced model, evaluated on chronological splits with ranking and realism metrics; the final system implements exactly this plan.

Candidate generation (shared for fair comparison). Both models re-rank the *same* per-row candidate sets to keep the comparison apples-to-apples. Candidates blend three signals surfaced by EDA: short-range geo proximity, time-aware co-visitation from the last stop, and time/global popularity. The per-city parameters (radius R_{km}, counts for co-visit and popularity, and list cap) were tuned to balance recall and pool size and then saved for reproducibility (e.g., NYC: R=4 km, M_{covis}=80, P_{pop}=80, cap=240).

Baseline: time-aware heuristic (transparent). The baseline is a fast, interpretable ranker that linearly blends five ingredients for each candidate venue: (a) time-aware co-visit (last → next, by time bin), (b) time-agnostic co-visit, (c) time-aware popularity, (d) city-level popularity (fallback), and (z) a distance penalty from the user's last location. Weights are tuned on the validation split and stored; notably, the same weight vector won in both NYC and TKY, indicating robustness. This baseline intentionally captures the strongest EDA signals while remaining easy to explain and debug.

Advanced model: candidate-aware FPMC+ (sequence-aware re-ranker). To capture "who you are" *and* "where you just were," we use FPMC (Factorizing Personalized Markov Chains) as the core re-ranker. FPMC learns embeddings that combine user–venue affinity with a Markovian transition from the recent venue(s), making it well-suited to next-stop prediction. Our "FPMC+" adds: (1) *candidate-aware* hard negatives (sampled from the same candidate pool you will face at inference), (2) *last-N* context (we pool last 1–3 stops), (3) *category embeddings* to generalize across sparse venues, (4) a *learned* distance feature so the model discovers the right near–vs–good trade-off, and (5) early stopping on user-macro Hit@10. Training is per-city on chronological splits; the objective is masked Bayesian Pairwise Ranking (BPR) over (positive, hard negative) pairs from the candidate set. Hyperparameters searched include embedding dimensions for venues and categories, sequence window (N), negatives per positive, learning rate, and regularization.

Fair comparison protocol. At test time we (i) build one candidate set per example using the saved candidate parameters, (ii) score the *same* candidates with the baseline and with FPMC+, and (iii) report the same macro-averaged ranking and realism metrics. This isolates the re-ranker's contribution from candidate-recall effects and keeps the analysis faithful to the proposal and course brief.

Why this two-model design. The heuristic baseline provides a strong, transparent yardstick rooted in co-visitation, popularity, and distance; FPMC+ injects personalization and short-sequence structure to lift inclusion (Hit@K) while learning a principled distance trade-off, reflecting the real decision a traveler makes: "what's next, near me, that fits my pattern?"

1.4 Evaluation Results with Tables

Metrics (macro-averaged). We report per-user (macro) averages so heavy users don't dominate. **Hit@K**: whether the true next POI is in the top K. **MRR@10**: rewards higher rank of the true POI within top 10. **nDCG@10**: position-weighted relevance within top 10.

coverage@10: distinct POIs that appear across users' top-10 (catalog breadth). **Dist@1**: Haversine distance (km) from last stop to the model's top-1 (lower is better). Values are rounded to three decimals.

NYC (users = 1083)	**Hit@5**	**Hit@10**	**Hit@20**	**MRR@10**	**nDCG@10**	**coverage@10**	**Dist@1 (km)**
Baseline (heuristic)	0.150	0.177	0.198	0.113	0.128	22,294	3.30
FPMC+ (advanced)	0.143	0.185	0.227	0.092	0.114	19,008	2.92

Table 2: Test results on NYC (macro per-user).

Tokyo (users = 2293)	**Hit@5**	**Hit@10**	**Hit@20**	**MRR@10**	**nDCG@10**	**coverage@10**	**Dist@1 (km)**
Baseline (heuristic)	0.193	0.243	0.287	0.134	0.160	36,150	3.59
FPMC+ (advanced)	0.192	0.253	0.316	0.128	0.158	23,895	3.32

Table 3: Test results on Tokyo (macro per-user).

1.5 Key Insights, Limitations, and Future Improvements

Key insights (from the test tables).

- **Advanced model improves inclusion and realism.** FPMC+ gets the correct POI into the list more often (NYC: Hit@10 +0.009, Hit@20 +0.029; TKY: Hit@10 +0.010, Hit@20 +0.029) and proposes closer top-1 suggestions (NYC: Dist@1 3.30 → 2.92 km; TKY: 3.59 → 3.32 km). This matches our design goal of learning a distance-aware, sequence-aware re-ranker.

- **Baseline is a very strong top-ranker with broad catalog reach.** The heuristic slightly outperforms on rank-sensitive metrics (MRR@10, nDCG@10) and touches more unique venues (higher coverage@10), showing that time-aware co-visitation + popularity + distance remains a high-quality, interpretable yardstick.

- **City effect.** Both models score higher in Tokyo than NYC across Hit@K/nDCG, consistent with stronger sequential regularities and denser POI flows observed there; the baseline itself performs better on TKY than NYC.

- **Fair comparison preserved.** All results come from re-ranking the *same* candidate sets per example, making the comparison apples-to-apples.

Limitations.

- **Candidate-recall ceiling.** The true next POI is only in our candidate lists about 45.1% (NYC) and 58.3% (TKY) of the time, which hard-limits achievable Hit@K regardless of the re-ranker; test recall was a bit below validation, explaining part of the val → test drop.

- **Coverage vs. precision trade-off.** FPMC+ improves Hit@K but reduces coverage relative to the heuristic (NYC: 22,294 → 19,008; TKY: 36,150 → 23,895), indicating a tendency to concentrate on a smaller subset of venues.

- **Early-rank placement.** The baseline places the truth higher within top-10 (better MRR/nDCG) when it succeeds; FPMC+ often includes the truth but sometimes lower in the list. This affects user experience if only top-few are visible.

- **Scope of realism.** Dist@1 checks spatial plausibility for top-1 but does not account for walking time, road networks, or transit; nor do we model transient context (weather, events).

- **Modeling scope.** We use short history (last 1–3) and city-specific models; long-range dependencies, cross-city transfer, and true cold-start handling (new users/venues) remain open.

Future improvements.

- **Stronger candidate generation.** Learn a retrieval model (graph or dual-encoder) and/or adapt the geo radius by local density to raise candidate recall, since it sets the upper bound.

- **Multi-objective, route-aware re-ranking.** Add explicit terms for distance/time and category diversity in the scoring function (or train a multi-task head) so we jointly optimize inclusion and itinerary realism—aligned with the proposal's "route-aware re-ranking" and diversity goals.

- **Better top-rank placement.** Combine the heuristic's time-aware co-vis/popularity signals with FPMC+ via blending or a shallow "learning-to-rank" layer targeted at MRR/nDCG@10.

- **Richer context.** Incorporate travel time on the street network, public-transit reachability, and transient signals (hourly weather, events) to refine both candidate generation and re-ranking.

- **Robustness and coverage.** Add coverage/diversity regularizers or re-rankers (e.g., category-aware MMR) to prevent over-concentration while preserving Hit@K.
- **Ablations & error taxonomy.** Quantify the contribution of last-N, category embeddings, and learned distance; categorize misses by (i) candidate-not-found vs. (ii) rank-too-low to guide the next iteration.

2 Task B: Itinerary Recommendation

This part of the project implements a travel recommendation system that helps users discover places in Boston and Miami (more cities and data can be added to the current system) based on their preferences, budget, and trip duration. The system integrates data from Google Places API and Yelp Fusion API, providing comprehensive information about restaurants, attractions, and activities with ratings and reviews.

2.1 Key Metrics

- Cities Supported: 2 (Boston, Miami)
- Categories Available: 9
- Budget Tiers: 4
- Total Places Collected: 500
- APIs Integrated: 2 (Google Places, Yelp)

2.2 Key Features

- Personalized recommendations based on 9 different activity categories
- Budget-aware filtering with 4-tier system (Budget to Luxury)
- Smart caching mechanism that reduces API costs by 100% on subsequent runs
- Multi-day itinerary generation
- Top famous places identification based on popularity and ratings

2.3 Technical Approach

The system uses content-based filtering with a weighted scoring algorithm that combines place ratings (70% weight) and popularity metrics (30% weight) to generate high-quality recommendations. All API responses are cached using Parquet format for efficient storage and retrieval.

2.3.1 Recommendation Algorithm: Content-Based Filtering with Feature Weighting

1. Filter places by user-selected categories and budget level
2. Normalize review counts using logarithmic scale to handle wide range

$$\text{review\_score} = \frac{\log(\text{review\_count} + 1)}{\text{max\_log\_reviews}}$$

3. Calculate weighted score combining rating quality (70%) and popularity (30%)
4. Sort places by final score and return top N results

$$\text{score} = (\text{rating} \times 0.7) + (\text{review\_score} \times 0.3)$$

2.4 Data Collection Statistics

2.4.1 Boston, MA

Overall Statistics

- Total Places: ~ 245
- Average Rating: 4.3/5.0
- Total Reviews: $\sim 185{,}000$
- Real Price Data Coverage: 45%

Places by Category

- Food: 85 places
- Cultural: 42 places
- Nightlife: 38 places

- Nature: 28 places
- Shopping: 22 places
- Leisure: 15 places
- Adventure: 8 places
- Family: 5 places
- Physical Activity: 2 places

Data Sources

- Google Places: 122 places
- Yelp: 123 places

2.4.2 Miami, FL

Overall Statistics

- Total Places: ~ 240
- Average Rating: 4.2/5.0
- Total Reviews: $\sim 175{,}000$
- Real Price Data Coverage: 48%

Places by Category

- Food: 82 places
- Nightlife: 45 places
- Cultural: 38 places
- Nature: 30 places
- Leisure: 20 places
- Shopping: 15 places
- Adventure: 6 places
- Family: 3 places

- Physical Activity: 1 place

Data Sources

- Google Places: 118 places
- Yelp: 122 places

2.5 Price Data Strategy

2.5.1 Real Price Data

- Restaurants: 80–90% coverage from Yelp
- Bars/Nightlife: 70–80% coverage from Yelp
- Used when available

2.5.2 Price Estimation Fallback

When real price data unavailable, category-based defaults are used:

- Cultural (museums): Level 2 ($$20–30)
- Adventure (tours): Level 3 ($50–100)
- Nature (parks): Level 1 (often free)
- Food/Nightlife: Uses real Yelp data

This hybrid approach ensures all places have price information for budget filtering.

2.6 Available Categories

The system supports 9 activity categories:

- **Leisure**: Spa, beach, resort, relaxation, massage
- **Adventure**: Hiking, water sports, kayaking, adventure tours, snorkeling
- **Physical Activity**: Gym, yoga studio, fitness, cycling, running trail
- **Nightlife**: Bar, nightclub, lounge, live music, dance club
- **Cultural**: Museum, art gallery, theater, historical site, cultural center

- **Food**: Restaurant, cafe, food tour, brunch, dining
- **Nature**: Park, botanical garden, nature reserve, beach, hiking trail
- **Shopping**: Shopping mall, market, boutique, shopping district
- **Family**: Zoo, aquarium, children's museum, family activities, amusement park

Users can select multiple categories for mixed-interest trips.

2.7 Sample Recommendations

2.7.1 Example 1

Input Parameters:

- City: Boston, MA
- Categories: Food, Cultural
- Budget Level: 2 (Moderate, $50–150/day)
- Trip Duration: 3 days
- Requested: Top 10 recommendations

Sample Output (Top 5):

1. Museum of Fine Arts
   Category: Cultural; Rating: 4.7; Reviews: 15,234; Price: $$; Score: 4.52
2. Union Oyster House
   Category: Food; Rating: 4.5; Reviews: 8,932; Price: $$; Score: 4.38
3. Isabella Stewart Gardner Museum
   Category: Cultural; Rating: 4.6; Reviews: 9,123; Price: $$; Score: 4.35
4. Neptune Oyster
   Category: Food; Rating: 4.8; Reviews: 5,234; Price: $$; Score: 4.47
5. Boston Tea Party Ships & Museum
   Category: Cultural; Rating: 4.5; Reviews: 6,892; Price: $$; Score: 4.29

2.7.2 Example 2

Input Parameters:

- City: Miami, FL
- Categories: Nightlife, Leisure
- Budget Level: 3 (Comfortable, $150–300/day)
- Trip Duration: 2 days
- Requested: Top 8 recommendations

Sample Output (Top 5):

1. South Beach
   Category: Leisure; Rating: 4.6; Reviews: 25,123; Price: Free; Score: 4.58
2. LIV
   Category: Nightlife; Rating: 4.3; Reviews: 12,456; Price: $$$; Score: 4.18
3. The Standard Spa
   Category: Leisure; Rating: 4.5; Reviews: 3,892; Price: $$$; Score: 4.12
4. Ball & Chain
   Category: Nightlife; Rating: 4.6; Reviews: 5,234; Price: $$; Score: 4.35
5. Nikki Beach
   Category: Leisure; Rating: 4.4; Reviews: 8,123; Price: $$$; Score: 4.22

2.8 Top Famous Places Feature

Boston's Most Famous Places:

1. Fenway Park – 45,234 reviews, 4.7 rating
2. Museum of Fine Arts – 15,234 reviews, 4.7 rating
3. Boston Common – 12,450 reviews, 4.6 rating
4. New England Aquarium – 11,892 reviews, 4.5 rating
5. Freedom Trail – 10,234 reviews, 4.8 rating

2.9 Example: 3-Day Boston Trip

2.9.1 Day 1 (Cultural Focus)

Morning:

- Museum of Fine Arts Boston (4.7, Cultural)
- Boston Tea Party Ships & Museum (4.5, Cultural)

Afternoon:

- Union Oyster House (4.5, Food)
- Quincy Market (4.3, Food)

Evening:

- North End Italian Restaurant (4.6, Food)

2.9.2 Day 2 (Mix of Activities)

Morning:

- Isabella Stewart Gardner Museum (4.6, Cultural)
- Boston Public Garden (4.6, Nature)

Afternoon:

- Neptune Oyster (4.8, Food)
- Newbury Street Shopping (4.4, Shopping)

Evening:

- Legal Sea Foods (4.3, Food)

2.9.3 Day 3 (Exploration)

Morning:

- Harvard University Tour (4.7, Cultural)
- Harvard Square Cafes (4.5, Food)

Afternoon:

- MIT Museum (4.4, Cultural)
- Central Square Dining (4.5, Food)

Evening:

- Cambridge Nightlife (4.3, Nightlife)

2.10 Conclusion

This travel recommendation system successfully demonstrates the integration of multiple data sources, intelligent caching, and content-based filtering to provide personalized travel recommendations. The project achieves its core objectives to integrate two major APIs (Google Places and Yelp) with proper error handling and rate limiting awareness.

Collected and processed 500+ places across Boston and Miami with reviews, ratings, and categorization showing comprehensive data collection. Created flexible recommendation engine that balances quality (ratings) with popularity (review count). Designed modular, maintainable codebase following best practices and separation of concerns. Built system that can easily expand to more cities, categories, and features without major refactoring.