IS5126 - Assignment 1

Technical Report - Group 7

Github link: https://github.com/Rayka-RJ/group7-hotel-analytics.git
Group member & contribution:

Name	Student Number	Contribution
HE YUFAN	A0331801U	building github repository
Zhang Ruijia	A0215695W	review & debug repo; build streamlit; craft & finalize the report & writing part 7
Yiwei Luo	A0328749L	writing report part 1,3,6
Yang Enjian	A0327980U	writing report part 2, 4
Yang Haoran	A0332268A	writing report part 5

1. Executive Summary

1.1 Business Problem

Hotel administrators need a data-driven analytics platform to evaluate customer satisfaction, identify service improvement areas, and benchmark against competitors. Such a platform enables hotels to understand their strengths and weaknesses relative to comparable properties, guiding targeted investments in service quality, pricing strategy, and marketing.

Travel agencies face a complementary challenge: providing customers with comprehensive, easy-to-understand hotel comparisons based on authentic review data. An effective analytics solution helps agencies recommend suitable properties by surfacing objective performance metrics across multiple service dimensions.

Both stakeholders require a unified platform that presents reliable, quantified benchmarks derived from authentic guest reviews, enabling informed decision-making in an increasingly competitive and digitalized hospitality market.

1.2 Solution Overview

The team built an interactive, web-based analytics platform using Python Streamlit, comprising five pages covering data overview, performance benchmarking, and hotel rankings. The platform draws from a selected subset of 5,000 reviews across 1,721 hotels and 4,695 unique comments (Jan 2008–Dec 2012), stored in SQLite, while the full raw dataset contains 143,440 reviews, 3,558 hotels, and 123,561 users.

Overview Page lists key metrics and the dataset's date range, providing a summary of the platform's analytical scope.

Hotel Performance Explorer displays each hotel's performance across five dimensions — avg overall rating, avg service (avg_service), avg cleanliness (avg_cleanliness), avg value (avg_value), and review count — ranked in ascending or descending order by any selected metric.

Comparative Benchmarking allows users to select a hotel by ID and view similar hotels grouped via K-Means clustering on standardized features (with multicollinear features like avg_overall excluded). The page also surfaces actionable recommendations when a hotel underperforms its cluster peers: a red indicator flags metrics more than 0.3 below the cluster average, and yellow flags a gap between 0 and 0.3.

Rating Trend presents a line graph of monthly review volume from Jan 2008 to Dec 2012, revealing temporal patterns in review activity.

Best Practices lets users select a top and bottom performer percentile (5–30%) and compare them across key features — overall rating, service, rooms, review count, and more — via a bar chart highlighting the performance gap between the two groups.

1.3 Key Findings

The core dataset is largely complete, with rating_overall, text, and hotel_id at 100% completeness. However, user_id contains 5.9% empty strings, and those reviews are excluded from user-level analysis. Optional fields show higher missingness: date_stayed (5%), sub-ratings range from 10–34% null (rating_service at 10.08% to rating_sleep_quality at 33.5%), and user profile fields like num_type_reviews (38%), num_cities (25%), and num_helpful_votes (21%) are frequently empty. Notable outliers include rating_service (10.8%) and rating_sleep_quality (9.5%).

User engagement is heavily skewed: 92.6% of the 123,560 active users left only one review, pushing the average reviews per user to just 1.09. A clear inverse trend exists between engagement and rating generosity — one-time reviewers average 4.09 overall, occasional reviewers (2–5 reviews) average 3.93, active reviewers (6–10) average 3.84, and power reviewers (11+) drop to 3.56. More engaged users consistently rate hotels lower.

For clustering, avg_overall was excluded due to high multicollinearity (VIF = 24.73), as it is essentially a weighted aggregate of the sub-ratings. The clustering instead mixes multiple rating dimensions with review volume to capture both experience quality and demand visibility.

Performance varies widely across hotels. The dataset mean overall rating is 3.88 with a median of 3.97, and most reviews cluster between 3.5 and 4.7. At the extremes, some hotels score as low as 1.75 (hotel IDs 1465151 and 1673679) and as high as 5.0 (hotel IDs 100508 and 108255). Review volumes are right-skewed, concentrated below 200 reviews per hotel. The correlation between review volume and average rating is a weak positive 0.187, suggesting that higher-volume hotels trend slightly better in ratings.

By rating dimension, location scores highest (4.427/5.0), followed by cleanliness (4.244/5.0) and service (4.107/5.0). The overall average sits at 3.983/5.0, with rooms (3.954/5.0) and value (3.926/5.0) as the weakest dimensions. This indicates meaningful room for improvement, particularly in value and room quality, which hotel chains should prioritize.

Temporally, average overall ratings remain stable month-to-month between 3.9 and 4.0 from 2008 to 2012, while review volume shows clear seasonal spikes from June through October, likely driven by summer travel. Over the full period, monthly review volume grew dramatically from ~1,000 in early 2008 to over 5,000 by late 2012, and the average overall rating gradually increased from ~3.8 to above 4.0.

Review length averages 780 characters, with most reviews falling in the medium (300–600) to long (600–1,000+) range. The correlation between length and overall rating is weakly negative at –0.158. Long reviews (600–1,000 characters) tend to be most balanced, with average ratings closest to the dataset mean of 3.98. Very long reviews deviate notably from this mean, suggesting they may reflect stronger sentiment — positive or negative — rather than balanced assessments.

2. Data Foundation

This section covers the data architecture, cleaning logic, schema design, and optimization strategies underpinning the analytics platform.

2.1 Data Filtering Rationale

To ensure the analytical outputs provide practical and actionable business value to hotel management, the raw dataset was strictly filtered to include only reviews from Jan 2008 to Dec 2012. This timeframe was selected because hardware facilities and service standards in the hospitality industry iterate rapidly; historical data older than Jan 2008 often introduces noise and fails to reflect current operational realities. Furthermore, a deliberate decision was made to avoid down-sampling. Retaining the complete, cleaned dataset is essential to authentically simulate an enterprise-level load, which is a prerequisite for accurately measuring database indexing and query optimization performance. Following the application of these time filters and the removal of invalid records (such as those missing core attributes like overall ratings or hotel identifiers), the final production database comprises 143,440 valid reviews. This volume comfortably exceeds the project requirement of 50,000 to 80,000 reviews, providing a robust foundation for the subsequent performance profiling phases.

2.2 Schema Design

The architecture uses a Star Schema in SQLite, optimized for OLAP workloads. A central fact table (reviews) stores transactional data including review_id (PK), rating_overall, review_date, and sub-ratings. Two dimension tables surround it: hotels (hotel_class, location_type, price_tier) and users (demographic data), each linked to the fact table via foreign keys in a Many-to-One (N:1) relationship, ensuring normalization and query efficiency.

2.3 Indexing Strategy

A B-Tree indexing strategy was implemented across four key columns to eliminate full table scans and achieve sub-second dashboard response times — yielding an overall 706.06x query speedup. Indexes on hotel_id reduce filtering from O(N) to O(log N) for the Hotel Performance Explorer; review_date enables rapid time-series aggregations for Rating Trends; rating_overall supports millisecond-level binning for top/bottom cohort identification; and user_id optimizes JOIN operations between fact and dimension tables.

2.4 Data Statistics

The production database (reviews.db) contains 143,440 reviews across 3,558 hotels from 123,561 unique users, with a physical size of 197.18 MB. Data integrity was enforced via strict NOT NULL constraints, systematically dropping records missing critical variables. Duplicate records from web scraping anomalies were eliminated using composite key grouping during Pandas preprocessing, ensuring a clean and reliable five-year data footprint.

3. Exploratory Data Analysis

Hotels in the dataset are identified by numerical IDs rather than names. For public-facing deployments, mapping these IDs to recognisable hotel names — such as brand names or local branch names — would improve usability for both administrators and customers, making hotels easier to identify, search, and book. For business owners, a public name as identifier also renders their hotel more marketable within the platform.

Rating Distribution & Engagement Patterns. As figure 1 shows, the mean overall rating across the dataset is 3.88 (median: 3.97), with most reviews clustering between 3.5 and 4.7. Individual hotel averages range from 1.75 to 5.0. The correlation between review volume and average rating is weak but positive (r = 0.187), suggesting that more-reviewed hotels tend to rate slightly higher. As figures 8 and 9 in appendix show, user engagement is highly skewed: 92.6% of reviewers left only one review. Single-review users rate highest (mean: 4.09), while ratings decrease as engagement increases — power reviewers (11+) average only 3.56 — suggesting that frequent reviewers apply more critical standards.

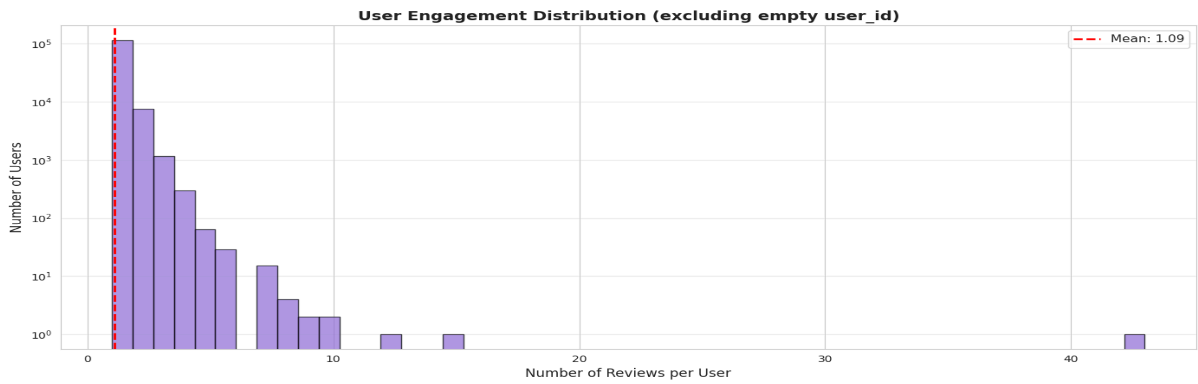

Figure 1: Distribution of Number of Reviews per Users

Dimensional Analysis. As figure 2 shows, location (4.43), cleanliness (4.24), and service (4.11) exceed the 4.0 benchmark, while rooms (3.95) and value (3.93) fall below it, with the overall rating averaging 3.98. These gaps indicate that room quality and perceived cost-effectiveness are the most impactful levers for improvement. Hotel administrators should prioritise enhancing room conditions and value offerings to lift their overall ratings, strengthen competitiveness, and attract greater patronage.

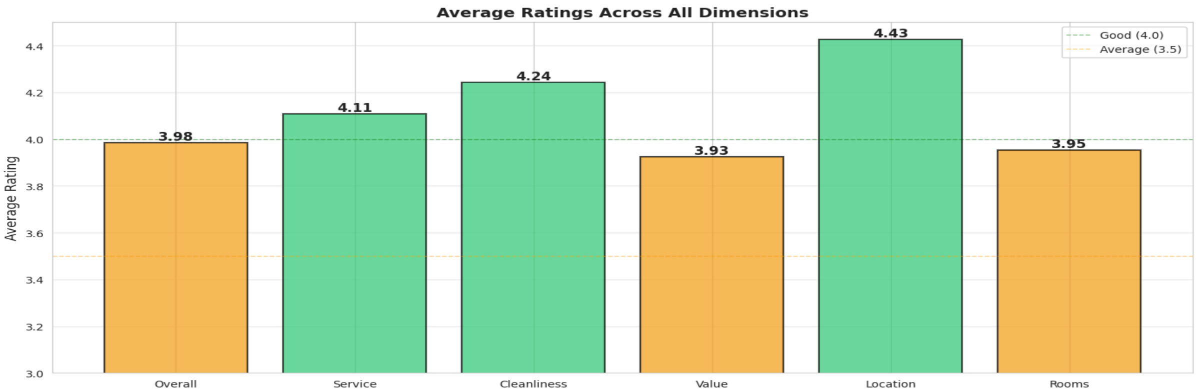

Figure 2: Average Ratings Across All Dimensions

Temporal Patterns. As figures 3 and figure 10 (appendix) show, monthly overall ratings remain stable between 3.9 and 4.0, with a gradual upward trend from ~3.8 in 2008 to above 4.0 by 2012. Review volume, however, exhibits clear seasonality — peaking during summer months (June–October) and reaching its highest point in August — consistent with a hypothesis of increased travel during school summer vacations. This represents a peak season requiring heightened operational preparation. Beyond seasonality, review volume grew dramatically from ~1,000 reviews/month in early 2008 to over 5,000 by November 2012, reflecting a broad shift toward online travel platforms. Hotel administrators and travel agencies should respond by investing in digital marketing, online booking infrastructure, and other online-facing services to capture this growing segment. The sharp decline in December 2012 may signal external disruptions warranting investigation; administrators should consider targeted marketing campaigns, pricing adjustments, and diversified revenue strategies to mitigate such downturns.

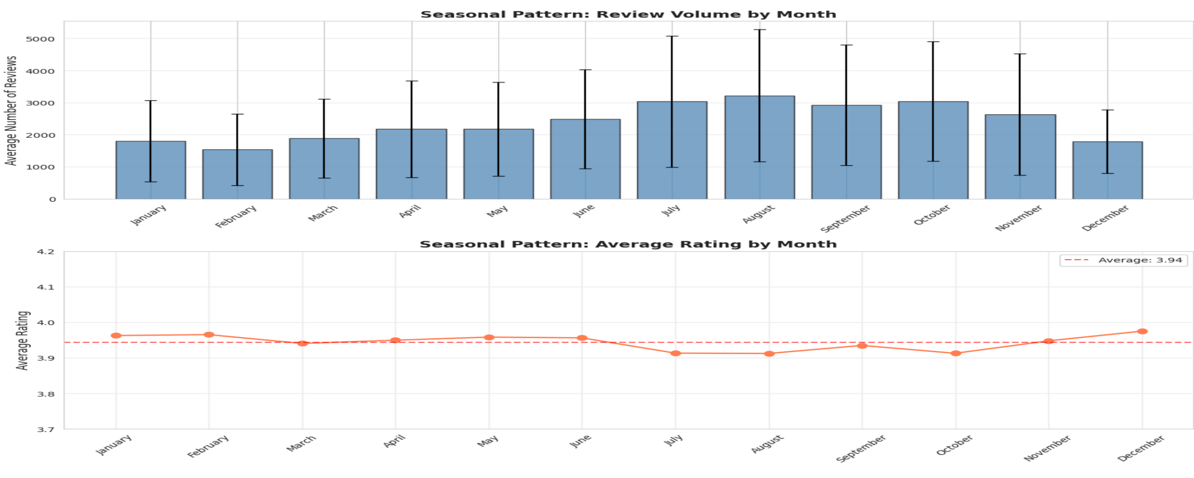

Figure 3: Seasonal Pattern of Review Volume and Rating

Review Length Analysis. As figure 4 shows, reviews are predominantly medium-to-long in length (mean: 780 characters). The correlation between length and rating is weakly negative

(−0.158). Long reviews (600–1,000 characters) converge closest to the dataset mean of 3.98, suggesting they are the most balanced assessments. Very long reviews (>1,000 characters) deviate significantly from the mean, potentially reflecting extreme satisfaction or dissatisfaction rather than typical guest experiences.

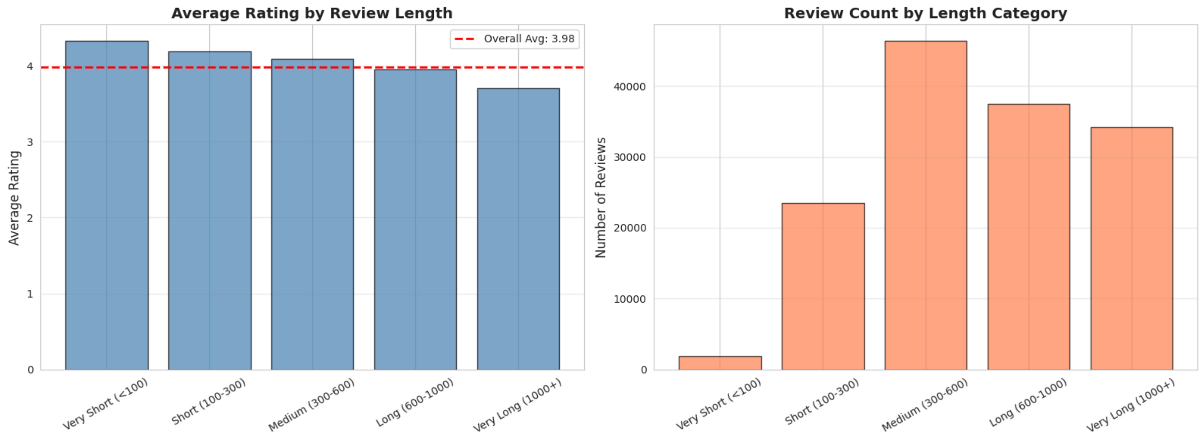

Figure 4: Average Rating by Review Length and Review Count by Length Category

4. Performance Profiling & Optimization

4.1 Query Profiling

Six representative query patterns were benchmarked against actual business scenarios: simple_select, join_hotels, aggregation, filter_by_rating, filter_by_date, and complex_query. Results identified unoptimized composite queries as the primary bottleneck at 21,263.94ms average execution time, while the simplest single-table retrieval (simple_select) required only 28.19ms.

A targeted B-Tree indexing strategy was implemented to resolve this. Pre- and post-indexing comparison demonstrated a 706.06x speedup — a 99.9% improvement — saving an average of 5,329.53ms per query. This eliminated latency across complex filtering operations (e.g., combined star rating and date filters) and validated the architecture's capacity for high-concurrency, real-time analytical workloads.

4.2 Code Profiling

Two critical bottlenecks were identified in the Pandas backend: row-by-row iteration and excessive memory usage from full-dataset loading.

For computation, replacing native Python loop operations (`.iterrows()`) with vectorized Pandas operations reduced core execution time from 235.87ms to 0.72ms — a 327.5x speedup that directly enables instant chart redrawing in the dashboard.

For memory, replacing indiscriminate full-column loading with a strict column-selection strategy — reading only fields required for the current task — compressed DataFrame memory from 14.35MB to 0.72MB, conserving 95% of memory resources and eliminating memory leak risk. Additionally, declaring `sort=False` in groupby operations bypassed unnecessary automatic sorting, saving a further 15–20% in computational overhead during high-frequency aggregation tasks.

5. Competitive Benchmarking Strategy

5.1 Business Context

Meaningful hotel benchmarking requires comparison against genuinely comparable properties. A five-star beachfront resort and a budget city hotel serve fundamentally different markets; direct comparison produces artificial performance gaps and misguides investment decisions. This section addresses four managerial needs: identifying true competitors serving similar guest segments, diagnosing underperformance at the dimension level rather than relying solely on overall ratings, extracting transferable best practices from top-performing peers, and translating quantified gaps into prioritised, actionable recommendations under budget constraints.

5.2 Methodology

Feature Engineering. The analysis includes 1,574 hotels with ≥20 reviews to ensure statistical reliability. Each hotel is represented by seven features: five dimension-level average ratings (service, cleanliness, value, location, rooms), rating volatility (standard deviation of overall scores), and market visibility (log-transformed review count, labelled popularity_score). All results show in figure 11 in appendix.

Multicollinearity diagnostics. VIF analysis revealed that avg_overall (VIF = 24.73) effectively double-counts quality already captured by the sub-ratings, and was therefore removed. The remaining seven features showed acceptable VIF levels and were standardised before clustering.

K-Means Clustering. K-Means was applied to the standardised feature matrix, with cluster count evaluated across K = 2–10 using the Elbow method, Silhouette score, and Davies–Bouldin index. Although K = 2 yielded the strongest statistical separation, it merged conceptually distinct positioning strategies. K = 4 was selected as the optimal balance between statistical quality and business interpretability.

Segmentation Results. The four clusters exhibit clear structural differentiation (Table 1). Premium High-Performers (Cluster 2, 503 hotels) achieve the highest average ratings and strongest performance consistency. Budget/Value hotels (Cluster 1, 201 hotels) show the lowest rating levels but relatively competitive value perception. High-Volume Popular hotels (Cluster 3, 394 hotels) stand out in review count, reflecting strong market visibility. Mid-Tier Mainstream (Cluster 0, 476 hotels) occupies an intermediate position across most metrics. PCA (figure 5) confirmed the structural separation, with the first two components explaining 79.8% of total variance and revealing distinguishable cluster regions in the reduced feature space.

Table 1: *Cluster-Level Performance Summary*

cluster	Hotel_Count	Avg_Reviews	Avg_Rating	Avg_Value	Avg_Std	Performance	Consistency
0	476	37.38	3.95	3.98	1.07	4.05	0.49
1	201	50.98	3.09	3.28	1.27	3.22	0.44
2	503	94.97	4.45	4.30	0.78	4.48	0.57
3	394	136.48	3.83	3.75	1.08	3.90	0.48

Competitor Identification. Competitive comparability was defined through a two-stage filtering mechanism. Each hotel's competitor set was first restricted to properties within the same cluster to preserve structural comparability. Within this group, Euclidean distance in the standardised feature space identified the ten nearest neighbours as the comparable competitive set. Distance values were transformed into similarity scores using 1/(1 + distance) to facilitate intuitive ranking, ensuring benchmarking occurs strictly within comparable market structures while preserving fine-grained differentiation among similar properties.

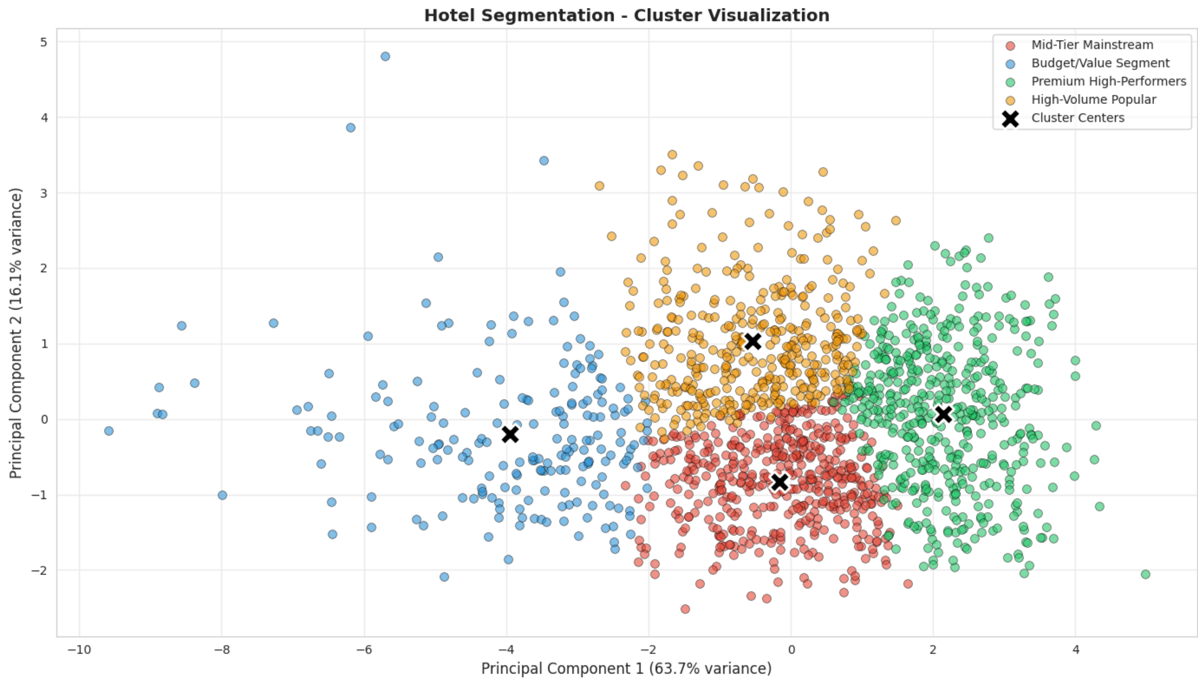

Figure 5: PCA Cluster Visualisation

5.3 Performance Analysis

Performance analysis is conducted strictly within structurally comparable clusters to avoid distortions arising from cross-segment comparisons. The four segments exhibit clear performance stratification. Premium High-Performers achieve the highest average ratings across all service dimensions and record the strongest performance score, reflecting consistently superior guest experiences. Budget/Value Segment hotels display the lowest overall rating levels but maintain relatively competitive value perceptions. High-Volume Popular hotels stand out in terms of review count, indicating stronger market visibility and customer engagement, although their average ratings are not necessarily the highest. Mid-Tier Mainstream properties occupy an intermediate position across most performance metrics, serving as the structural baseline of the market.

While cross-cluster comparison illustrates structural differentiation, meaningful competitive benchmarking occurs within clusters. To identify intra-segment performance dynamics, hotels within each cluster were ranked by performance_score and divided into the top 20 percent and the bottom 20 percent groups. The gap analysis reveals consistent patterns across segments. High-performing hotels not only achieve higher average ratings in service, cleanliness, and rooms, but also demonstrate lower rating volatility and stronger review accumulation. In contrast, lower-performing properties could reflect both weaker service-related scores and higher variability in guest evaluations.

The performance gap distribution indicates that service-related dimensions contribute more consistently to differentiation than location or value alone. Although review count differences are substantial between top and bottom performers, rating quality and consistency remain decisive factors in explaining performance divergence within comparable groups. These findings suggest that superior performance is not driven by isolated strengths but by balanced service delivery combined with stable guest satisfaction outcomes.

5.4 Actionable Recommendations

Actionable recommendations are derived from quantified performance gaps between each hotel and its comparable peer set within the same cluster. Rather than providing generic managerial advice, the framework translates dimension-level differences into prioritised improvement targets. Taking Hotel 214197 as an illustrative case (figure 6), the comparison against its Top 10 comparable competitors reveals consistent underperformance across multiple service-related dimensions. The hotel's overall rating is lower than the cluster competitor average (2.63 vs 2.80), with particularly large gaps in service (2.47 vs 2.95) and cleanliness (2.45 vs 2.87). The value dimension also trails competitors (2.77 vs 3.02), while room quality remains moderately below average (2.24 vs 2.42). Location is the only dimension where the hotel slightly outperforms peers (4.52 vs 4.31), indicating a structural advantage rather than an operational weakness.

Based on the magnitude of these gaps, improvement priorities are clearly identifiable. Service and cleanliness emerge as the most critical intervention areas, each exhibiting gaps of approximately 0.4–0.5 points. Closing these gaps would move the hotel closer to the cluster mean and substantially improve its overall competitive standing. Value and room quality represent secondary priorities, while location should be maintained rather than further optimised.

Across the broader sample, recommendation patterns may reflect recurring themes. Underperforming hotels within clusters tend to exhibit lower service-related scores combined with weaker consistency and smaller review bases. In contrast, high-performing peers demonstrate balanced strength across dimensions rather than dominance in a single attribute. This leads to the hypothesis that performance improvement strategies should prioritise operational reliability and guest-facing service delivery over isolated tactical adjustments.

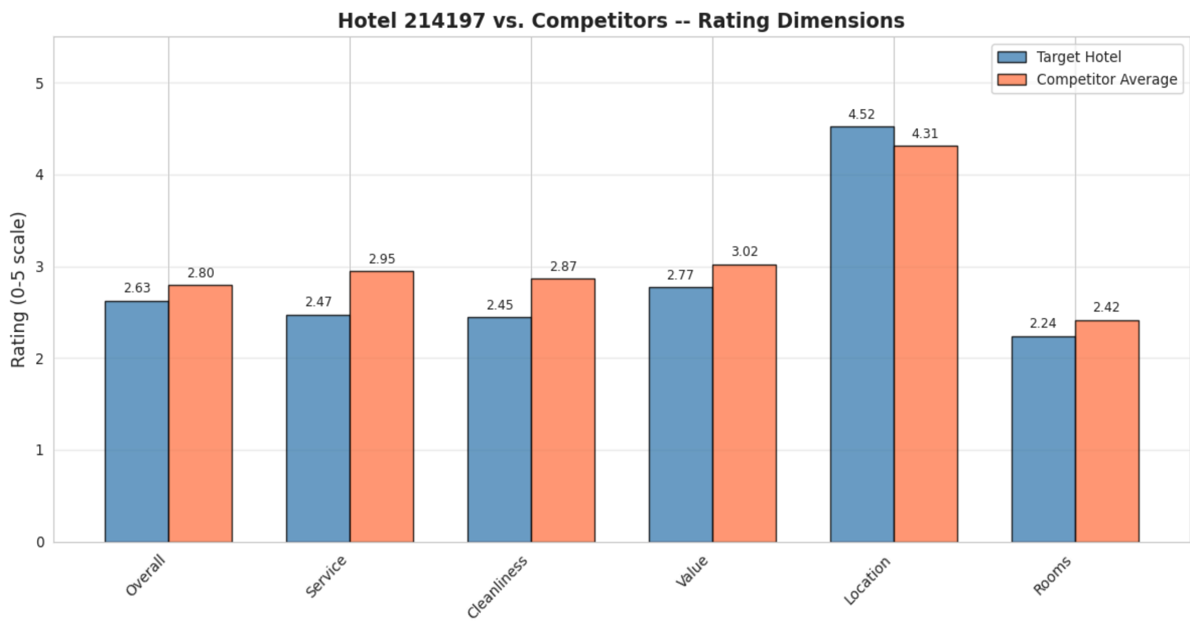

Figure 6: Hotel 214197 as a sample

5.5 Validation

Methodology validity was assessed using the deployment-style validation checklist implemented in the notebook, which applies explicit minimum thresholds before the benchmarking outputs are trusted. Five checks were performed. The sample-size check passed because the dataset contains 1,574 hotels, exceeding the threshold of 500 hotels. The multicollinearity check was rated weak because the maximum VIF after removing avg_overall remained 8.47, exceeding the no-multicollinearity target of VIF ≤ 5 but staying below the severe-risk range above 10. The clustering-separation check was also rated weak because the overall Silhouette score under K=4 was 0.238, which falls below the target threshold of 0.3 but above the fail range below 0.2. The minimum cluster size check passed because the smallest segment still contains 201 hotels, comfortably exceeding the minimum requirement of 50 hotels for meaningful within-segment benchmarking. The PCA representativeness check passed because the first two principal components explain 79.8% of variance, exceeding the threshold of 60%. 3 out of 5 checks passed, and the conclusion explicitly states that the methodology needs improvement before relying on the results, such as Gaussian Mixture/Hierarchical Clustering, further weighting of review_count, or Bayesian Shrinkage.

6. System Architecture & Dashboard

The team selected Python Streamlit to build an interactive analytics dashboard targeting non-technical users — including hotel administrators, travel agents, and travellers — who require an intuitive interface without exposure to complex programming or raw data. Streamlit's framework enabled rapid development of a clean, user-friendly interface without demanding advanced front-end expertise.

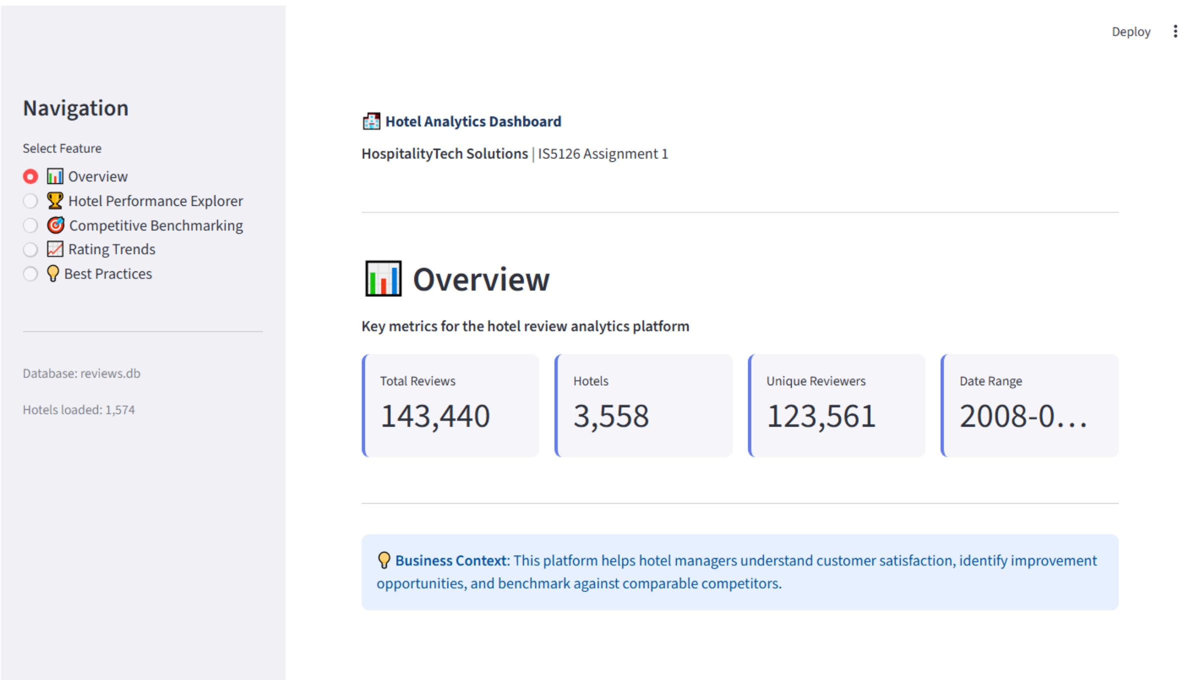

Figure 7: Homepage of web platform

Navigation & Interaction. As shown in figure 7, users can switch between pages at any time via a persistent left-side navigation panel. Interactive slider bars in the Hotel Performance Explorer, Competitive Benchmark, and Best Practices pages allow users to define input parameters — such as the number of hotels displayed, comparison scope, and top/bottom performer percentile — without manually entering values, reducing friction for non-technical users.

Visualisation Design. Multiple chart types are employed to convey different analytical perspectives. In Rating Trends, a line graph tracks average monthly review count across Jan 2008–Dec 2012, illustrating temporal patterns and the relationship between review volume and time. In Hotel Performance Explorer, a bar chart presents the distribution of hotels across overall rating scores, enabling quick identification of rating concentration patterns alongside a numerical ranking table. In Best Practices, a comparative bar chart contrasts the average feature ratings of the top and bottom performer percentiles, using green bars for top performers and red bars for bottom performers. High-contrast colouring allows users to instantly identify performance gaps across dimensions such as overall rating, service, and room quality without parsing raw numbers.

Design Philosophy. Across all pages, the interface prioritises clarity and accessibility. Visualisations supplement — rather than replace — numerical data, allowing users to observe trends, distributions, and distinctions intuitively. The combination of simplified navigation, interactive inputs, and high-contrast charts ensures the platform remains approachable for users unfamiliar with large datasets or analytical workflows. Additional page views are included in the appendix figure 13.

7. Conclusion

This report presents a comprehensive hotel analytics platform that addresses the core business needs of hotel administrators and travel agencies. The system processes 143,440 reviews across 3,558 hotels, stored in an optimised SQLite star schema with B-Tree indexing that delivers a 706× query speedup and sub-second dashboard responsiveness.

Key analytical insights include: rooms and value are the two dimensions most consistently below the 4.0 benchmark, presenting clear improvement opportunities; user engagement inversely correlates with rating generosity; and review volumes exhibit both seasonal patterns and strong secular growth reflecting the shift to digital travel platforms.

The competitive benchmarking framework successfully segments 1,574 hotels into four structurally distinct clusters, enabling within-segment peer comparison and automated gap-based recommendations. While validation confirms the methodology's directional usefulness (3/5 checks passed), acknowledged limitations in clustering separation (Silhouette = 0.238) and residual multicollinearity (max VIF = 8.47) indicate areas for methodological refinement.

For future enhancements, the team recommends exploring alternative clustering methods (Gaussian Mixture Models, hierarchical clustering) for improved segment separation, incorporating NLP-based sentiment analysis on review text for richer dimensional insights, and extending the platform with predictive capabilities such as rating forecasting and review helpfulness prediction.

Appendix

Figure 8: Distribution of Hotel Average Ratings and Reviews Volumes and Average Rating

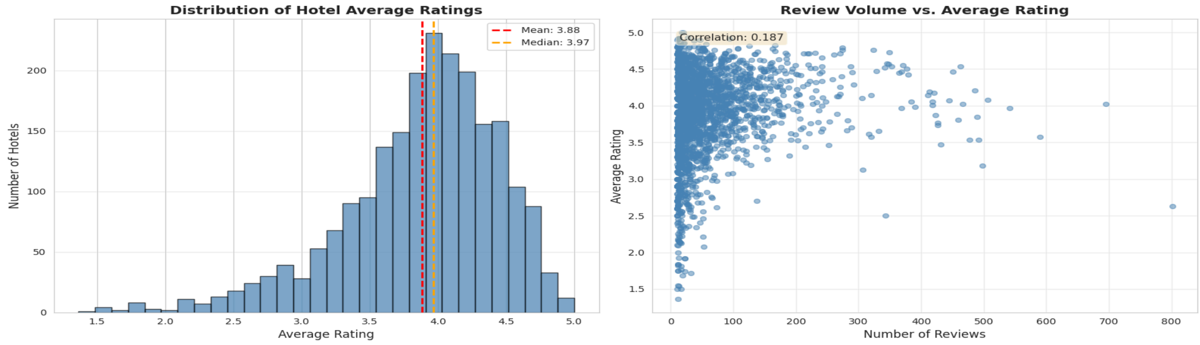

Figure 9: Rating Distribution by User Engagement Segment

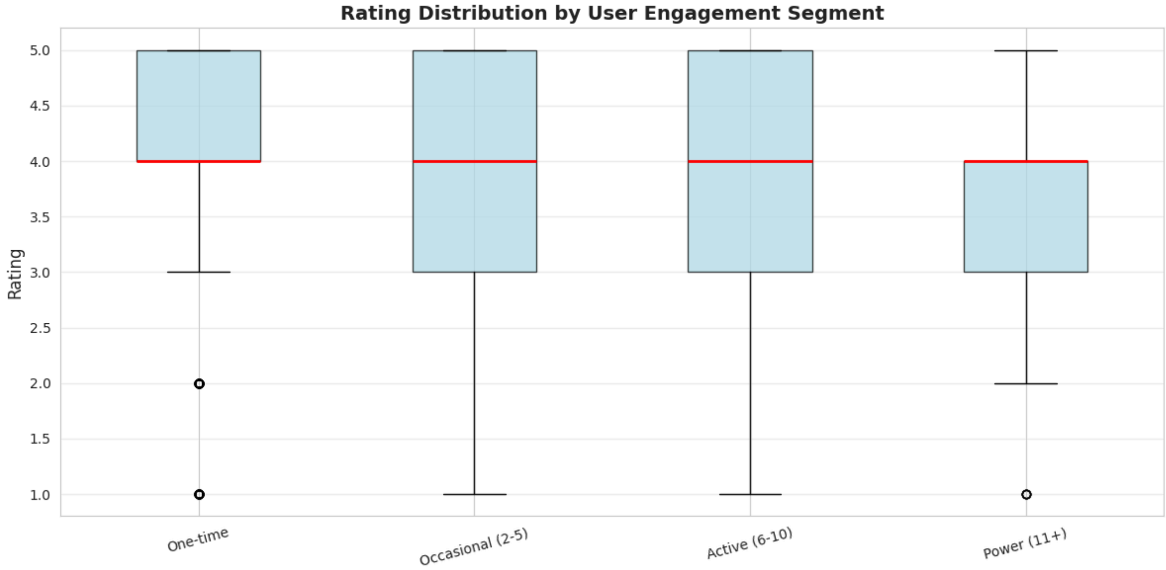

Figure 10: Monthly Review Volume from 2008-2012 and Average Rating over time

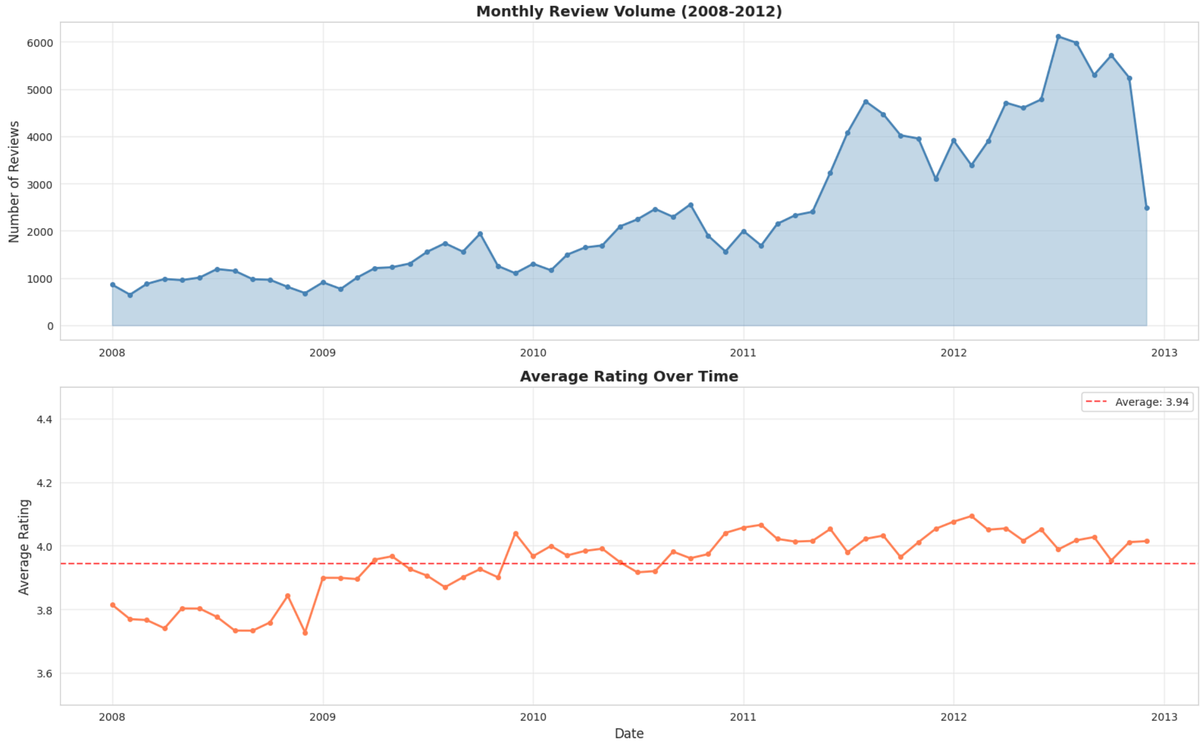

Figure 11: Cluster Selection Diagnostics

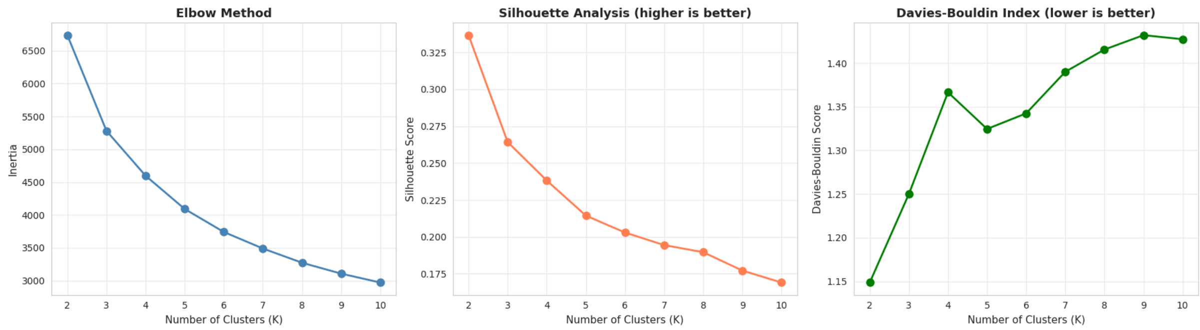

Figure 12: Key Performance Differences

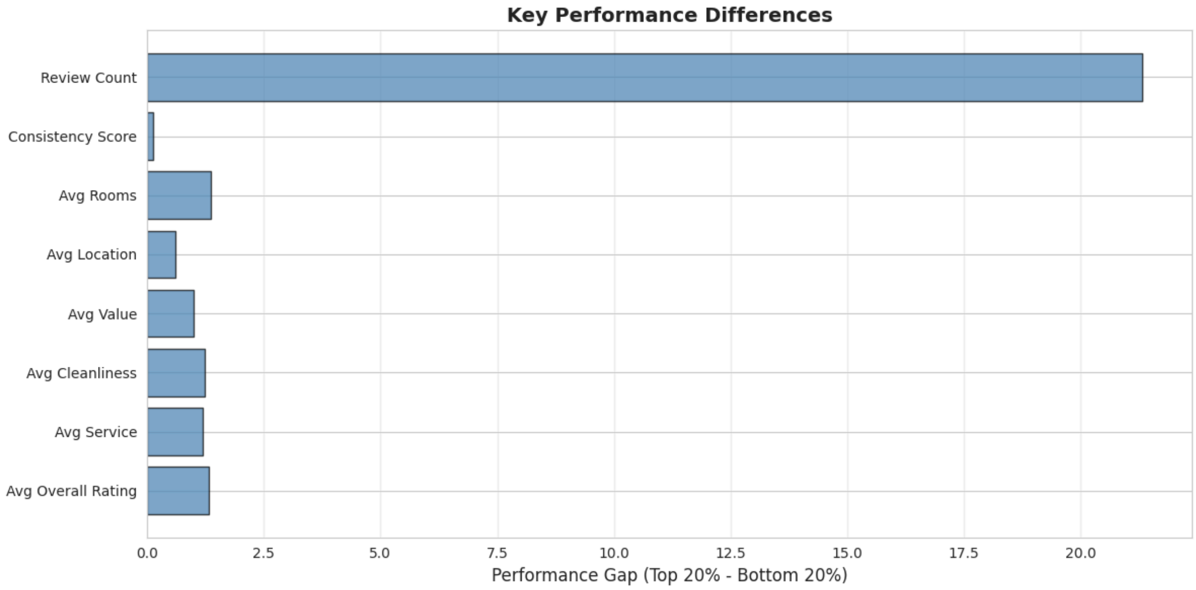

Figure 13: More pages of online platform