**MODERN DATA SCIENCE – SIT742 – Group 12**

**Question 1.1 Explanation**

We chose to replace null text entries with “no review” and convert Unix timestamps into yyyy-mm-dd dates because this ensures data completeness and readability. Missing values can cause issues in aggregations, while human-readable dates are essential for time-based grouping. Alternative solutions include dropping null reviews or leaving timestamps as raw integers, but those approaches either reduce data richness or limit interpretability. Our method keeps all records intact while still distinguishing between actual text and missing input. It is the most optimal approach here because it is both simple and effective, providing a clean, standardized dataset. This step also prepares the data for downstream tasks like visualization, time series analysis, and categorical exploration without introducing unnecessary preprocessing complexity.

**Question 1.2 Explanation**

We chose to compute the number of reviews per gmap\_id and explore hourly review patterns using PySpark and Pandas because this approach directly answers the question of when and where reviews are most frequent. It highlights activity at both the business and temporal level. Alternative solutions might include normalizing reviews by business size, weighting by star rating, or analyzing sentiment for deeper context. Another option is clustering review times into broader windows (morning, afternoon, evening) to simplify patterns. Our solution is optimal because it focuses precisely on review frequency and timing without overcomplicating analysis. It balances efficiency and interpretability, giving clear visual evidence of customer behavior trends while maintaining computational scalability across large review datasets.

**Question 1.7 Explanation**

We joined reviews with metadata and focused on analyzing ratings by business categories, especially the top 10 most reviewed. This choice was made because examining all categories would overwhelm results, while top ones provide meaningful insights backed by sufficient sample size. Alternatives include clustering categories into broader groups like “food” or “retail,” or applying text-based sentiment analysis for deeper context. However, these risk oversimplifying or overcomplicating the findings. Our approach is optimal because it maintains clarity and detail, highlighting clear rating differences across popular categories. It effectively shows how restaurants and grocery stores receive consistently high ratings, whereas shopping malls and gas stations display more variability. This balance of coverage and interpretability makes it the best fit.

**Question 1.8 Explanation**

We constructed user\_business\_list by ordering visits chronologically and applied array\_distinct to remove duplicate businesses per user. This was chosen because repeated entries can bias similarity comparisons, while unique lists highlight diverse visitation histories. Alternative solutions include keeping duplicates to reflect loyalty or visit frequency, which may be useful in churn prediction or customer engagement studies. Another approach is sessionization, grouping visits within time windows to capture context of repeated interactions. Our solution is optimal for this question because it ensures concise, standardized user histories suited for Jaccard similarity and other comparisons. It avoids skew caused by redundancy and highlights user diversity, enabling more robust and interpretable similarity analysis across 20,000 users without unnecessary computational overhead.

Question 2.1 Explanation

We constructed a daily review time series, filled missing days with the global mean review count, and then applied additive decomposition. This was chosen because time series continuity is required for decomposition, and mean imputation avoids introducing artificial fluctuations that forward/backward fill might cause. Alternative solutions include linear interpolation, moving averages, or advanced imputation methods like ARIMA or Prophet-based forecasting. While those could capture local variation, they add complexity unnecessary for exploratory decomposition. Our solution is optimal here because it balances simplicity, computational efficiency, and interpretability. It successfully decomposed the series into a long-term trend (growth until 2019, decline post-2020), consistent seasonality, and irregular residuals. This provides clear, actionable insights while maintaining analytical transparency.