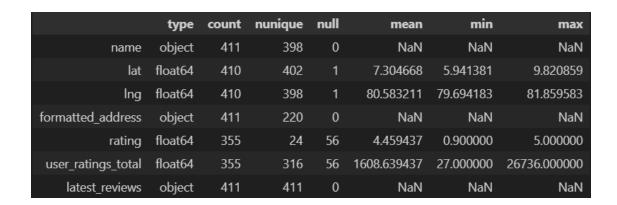
Documentation

1. Data analysis and derived insights

We began by exploring the places dataset to understand its structure, data types, and value distribution. Key features such as name, latest_reviews, activity_scores, rating, lat, lng, and formatted_address were analyzed.



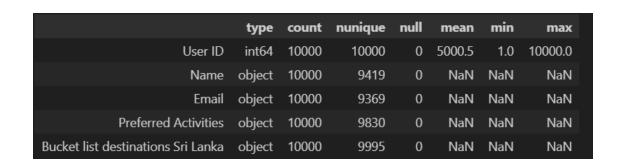
We observed the presence of non-English words in the dataset.

There were some duplicated data on places.

There was only one entry without location details. (Leisure World)

We identified some issues in the data related to the **formatted_address** feature.

Then we explore the Visitors **Visitors Preference** dataset.



Here we identified that the data related to Name and Email are generated. There were many duplicates in this dataset.

We then examined the <u>Preferred Activities</u> and found that they were gathered through some kind of user selection process. There were only 68 unique activities, each different in nature.

We discovered that the **Bucket list destinations Sri Lanka** data was manually added by users. There were many different forms of input to streamline this data.

We found 104 places that appeared in both the places dataset and the visitors Preference dataset.

The number of places only in the 'places' dataset is 291.

The number of places only in the bucket list is 54.

Additionally, there are 39 places in both datasets that are not exactly equal but have a 90% similarity match.

Finally,

Importance of Activity Coverage: Users have diverse interests, so it's crucial to ensure recommendations span the full range of activities. Our analysis revealed that some activities correspond to more places than others, potentially affecting the balance of recommendations.

User Bucket List as a Key Factor: We recognized that incorporating bucket list locations into the scoring process would significantly personalize results. This insight prompted us to add a scoring boost for places on a user's

Documentation

bucket list.

Distance as a Factor: We observed that the geographical spread of recommended places greatly influences their visit feasibility.

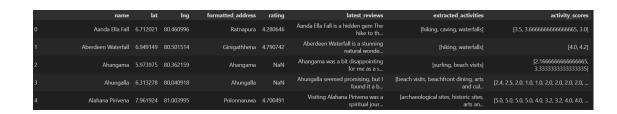
2. Pre-processing methods

Initially, we cleaned all the data by fixing the problems identified in the analysis phase:

- Non-English characters were removed
- Duplicates were merged
- formatted_address Was corrected
- Missing values were filled

Given the presence of both rating and user_ratings_total sections, we implemented a weighted rating system to provide a more balanced and reliable rating for each place.

A notable improvement was the extraction of relevant activities and their corresponding satisfaction scores for each location. This significantly aids our future processes.



We also removed unwanted symbols and characters from the datasets.

Finally, we normalized user input activities to match the terminology used in the dataset. For example, we converted "safaris" to "wildlife safaris" and "hot air ballooning" to "air ballooning." This pre-processing step ensured that our similarity calculations accurately matched user preferences with the corresponding activities in the dataset.

3. Evaluation Metrics and Rationale

Primary Evaluation Metrics:

1. Activity Score Sum:

- What: The sum of scores for the user's selected activities at each place.
- Why: This metric quantifies how well a combination of places matches the user's preferred activities. Higher scores indicate a better alignment between the user's interests and available places.

2. Bucket List Points:

- What: Additional points for places on the user's bucket list.
- **Why:** This personalized scoring prioritizes places the user explicitly wants to visit, enhancing the likelihood of satisfaction with recommendations.

3. Travel Distance (Haversine Formula):

- What: Total travel distance for visiting recommended places, calculated using the Haversine formula.
- **Why:** Minimizing travel distance is crucial for efficient trip planning. This metric ensures recommendations are both relevant and logistically feasible.

4. Location Ratings:

- What: Average rating of the places in the combination.
- Why: Integrating location ratings ensures that higher-rated places are prioritized, enhancing the overall quality of recommendations based on user feedback and satisfaction.

Secondary Evaluation Metric:

1. Normalized Scores:

- What: Average rating of the places in the combination.
- Why: Integrating location ratings ensures that higher-rated places are prioritized, enhancing the overall quality of recommendations based on user feedback and satisfaction.

Rationale Behind Selection of Evaluation Metrics:

Using both activity scores and travel distance as metrics aims to create recommendations that are personalized, relevant, and practical for real-world travel. This balance maximizes user satisfaction while minimizing logistical challenges.

Bucket list points prioritize places explicitly desired by the user, enhancing recommendation personalization.

Generating combinations that cover all user activities and applying specific constraints (e.g., limiting activity occurrences) ensures diverse recommendations, enhancing the user experience.

Documentation

3