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.

	type	count	nunique	null	mean	min	max
name	object	411	398	0	NaN	NaN	NaN
lat	float64	410	402	1	7.304668	5.941381	9.820859
lng	float64	410	398	1	80.583211	79.694183	81.859583
formatted_address	object	411	220	0	NaN	NaN	NaN
rating	float64	355	24	56	4.459437	0.900000	5.000000
user_ratings_total	float64	355	316	56	1608.639437	27.000000	26736.000000
latest_reviews	object	411	411	0	NaN	NaN	NaN

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.

	type	count	nunique	null	mean	min	max
User ID	int64	10000	10000	0	5000.5	1.0	10000.0
Name	object	10000	9419	0	NaN	NaN	NaN
Email	object	10000	9369	0	NaN	NaN	NaN
Preferred Activities	object	10000	9830	0	NaN	NaN	NaN
Bucket list destinations Sri Lanka	object	10000	9995	0	NaN	NaN	NaN

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

We then examined the `Preferred Activities` 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

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.

	name	lat	lng	formatted_address	rating	latest_reviews	extracted_activities	activity_scores
0	Aanda Ella Fall	6.712021	80.460996	Ratnapura	4.280646	Aanda Ella Fall is a hidden gem The hike to th...	[hiking, caving, waterfalls]	[3.5, 3.6666666666666665, 3.0]
1	Aberdeen Waterfall	6.949149	80.501514	Ginigathhena	4.790742	Aberdeen Waterfall is a stunning natural wonde...	[hiking, waterfalls]	[4.0, 4.2]
2	Ahangama	5.973975	80.362159	Ahangama	NaN	Ahangama was a bit disappointing for me as a s...	[surfing, beach visits]	[2.1666666666666665, 3.3333333333333335]
3	Ahungalla	6.313278	80.040918	Ahungalla	NaN	Ahungalla seemed promising, but I found it a b...	[beach visits, beachfront dining, arts and cul...	[2.4, 2.5, 2.0, 1.0, 1.0, 2.0, 2.0, 2.0, 2.0, ...
4	Alahana Pirivena	7.961924	81.003995	Polonnaruwa	4.700491	Visiting Alahana Pirivena was a spiritual jour...	[archaeological sites, historic sites, arts an...	[5.0, 5.0, 5.0, 5.0, 4.0, 3.2, 3.2, 4.0, 4.0, ...

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.