

Predicting honeymoon destinations using Reddit Datasets

Web API and NLP Project

Abirami Rajamanickam

Problem Overview

Title and Objective

- Predicting Top Honeymoon Destination Countries Using Travel and TravelHacks Reddit
- To classify subreddit posts and identify top and bottom honeymoon destinations using sentiment analysis.

Motivation & Problem Statement

- Travel forums are rich sources of real user experiences that can guide honeymoon planning.
- There's limited automated insight into preferred honeymoon destinations based on public sentiment.

Data Collection

Data Collection: Collected subreddit posts from Reddit website using Web API and NLP

Post types: Collected post types like new, top and hot to gather wider data.

Source: Source Reddit website using PRAW Web API

Travel subreddit is a community on Reddit where users discuss topics related to travel.

Travel Hacks is a community on Reddit sharing tips, tricks, and strategies to make travel more affordable, efficient, and enjoyable.

Data Science Steps: Data Collection, Data Cleaning, Data Preprocessing, Data Modeling and Evaluation

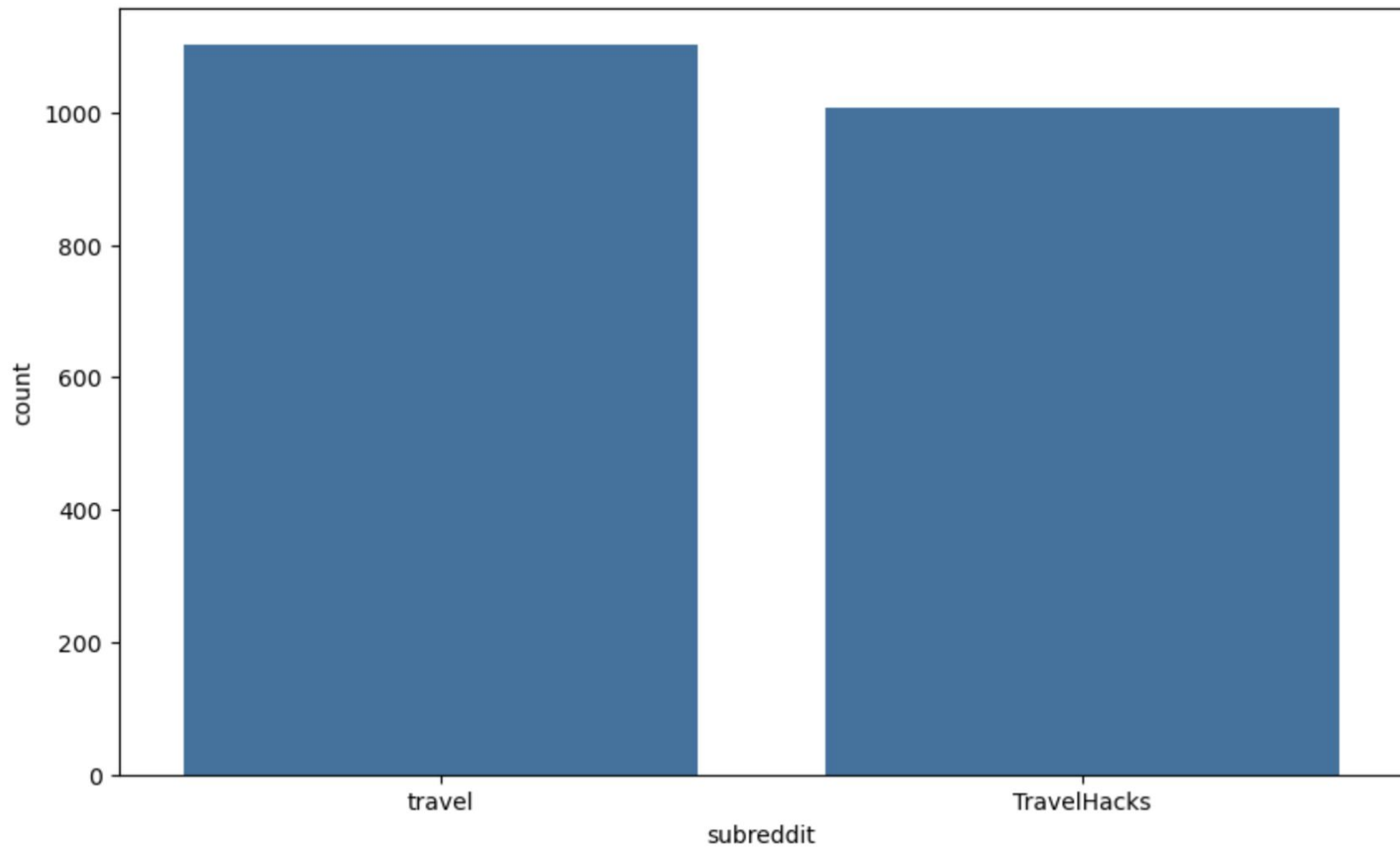
Column	Description
id	unique identifier assigned to each Reddit post
created_utc	column stores the timestamp of when the post was created
title	title of the Reddit post
author	contains the username of the Reddit user who created the post
selftext	column holds the body text of the Reddit post
num_comments	number of comments that the post has received
score	score (upvotes minus downvotes) of the Reddit post
subreddit	the name of the subreddit where the post was submitted

Exploratory Data Analysis

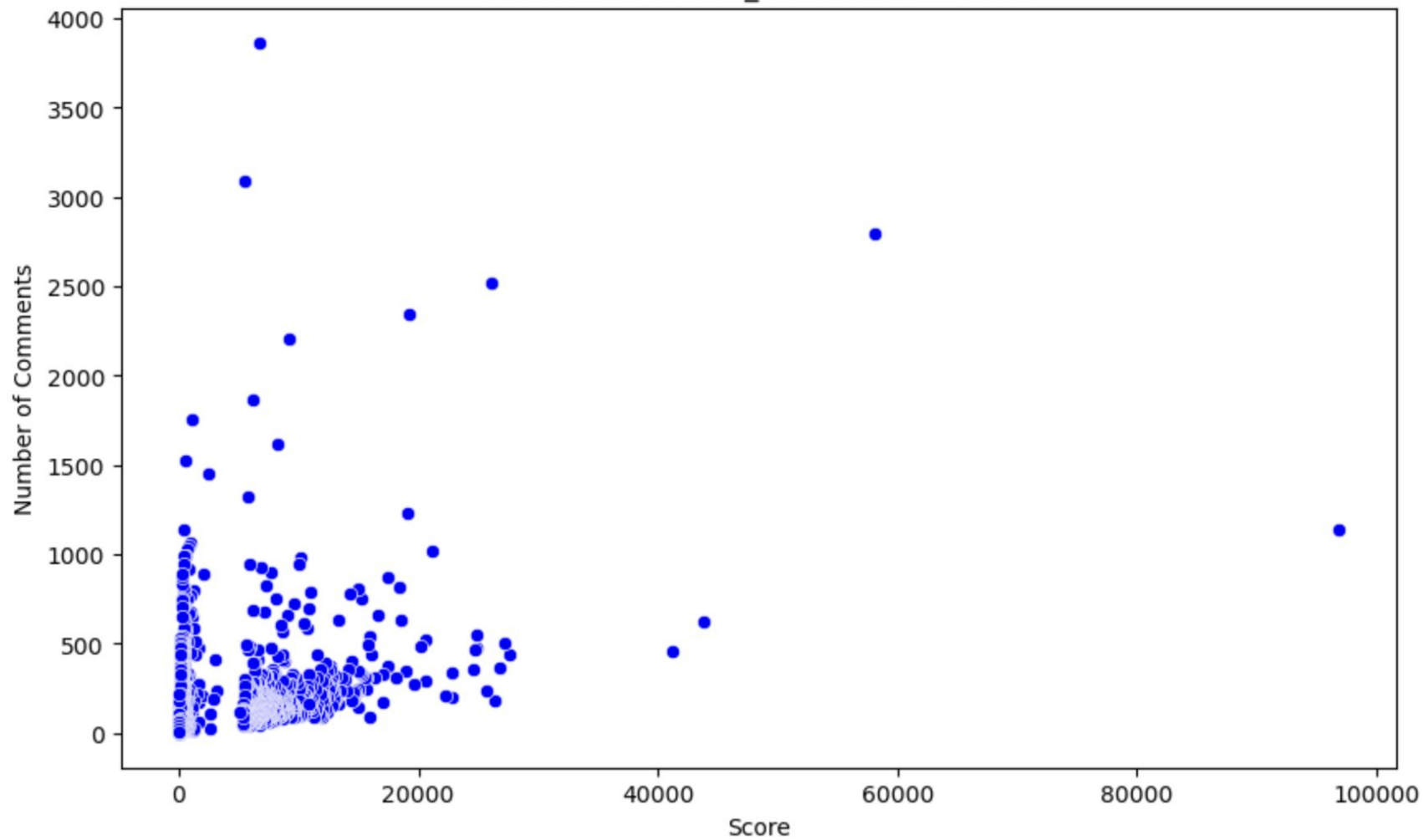
Performed Exploratory Data Analysis separately to identify statistical analysis of both travel and travel hacks subreddits

- Displayed sample data using `df.head()` method
- Performed summary statistics using `df.describe`
- Visualize the Distribution of features
- Explored correlation between features
- Text Data Exploration
- Tokenization, Lemmitizing
- Bigrams and Trigrams analysis
- Checks for balance of class/target feature
- Explored score and number of posts

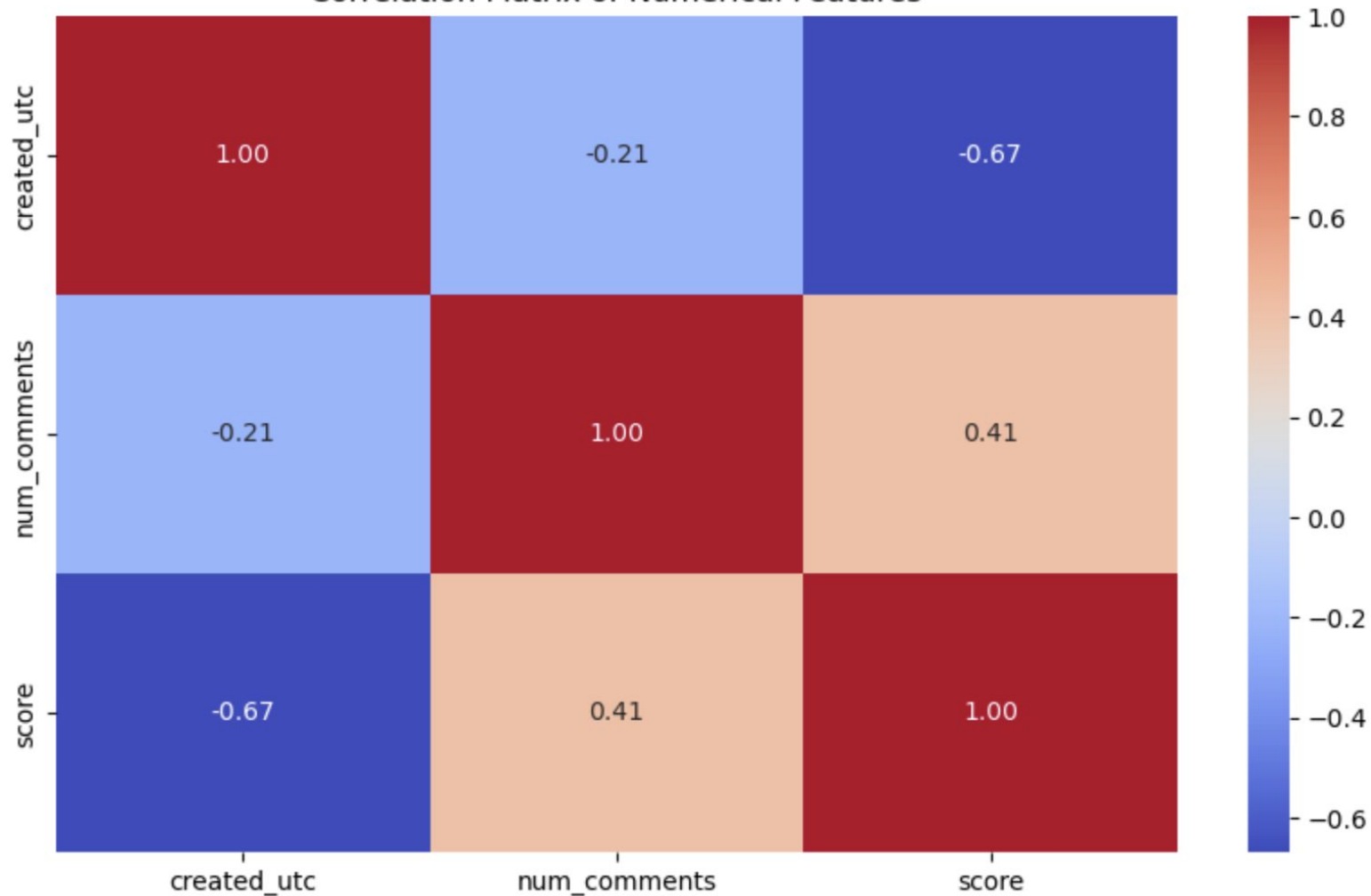
Count of Posts in Each Subreddit



Scatter Plot of num_comments vs. Score



Correlation Matrix of Numerical Features



Preprocessing

- **Stop Words Removal:** Removed unnecessary words.
- **Tokenization:** Converted text into a processable format.
- **Baseline Model:** Implemented Logistic Regression with TF-IDF for initial predictions.
- **TF-IDF Vectorization:** Used `TfidfVectorizer(stop_words='english', ngram_range=(1, 1))` to convert text to numerical vectors, capturing word importance.

Sentimental Analysis

Sentiment Analysis: Applied VADER to score text sentiment.

Honeymoon Pattern Identification: Created a new column, `is_honeymoon`, mapped to 1 if the text contains "honeymoon," else 0.

spaCy Installation: Installed spaCy (pip install spacy) for advanced NLP to recognize locations and places.

Place Identification: Cleaned and visualized place data, comparing location with sentiment score to predict positive reviews.

Model Finalization: Deployed an optimized ensemble model based on accuracy score using test data.

Model Evaluation

Baseline and Final Model Performance

Baseline Model: Logistic Regression

- Achieved an accuracy of approximately 81.75%, demonstrating a solid initial classification of subreddit posts.
- Precision, recall, and f1-scores were balanced across classes, indicating consistent model performance.

Final Model: Voting Classifier

- Improved accuracy to about 86.5% by combining Logistic Regression, Random Forest, and SVM classifiers.
- Achieved higher precision and recall, especially improving the model's ability to correctly identify Travel and TravelHacks posts.

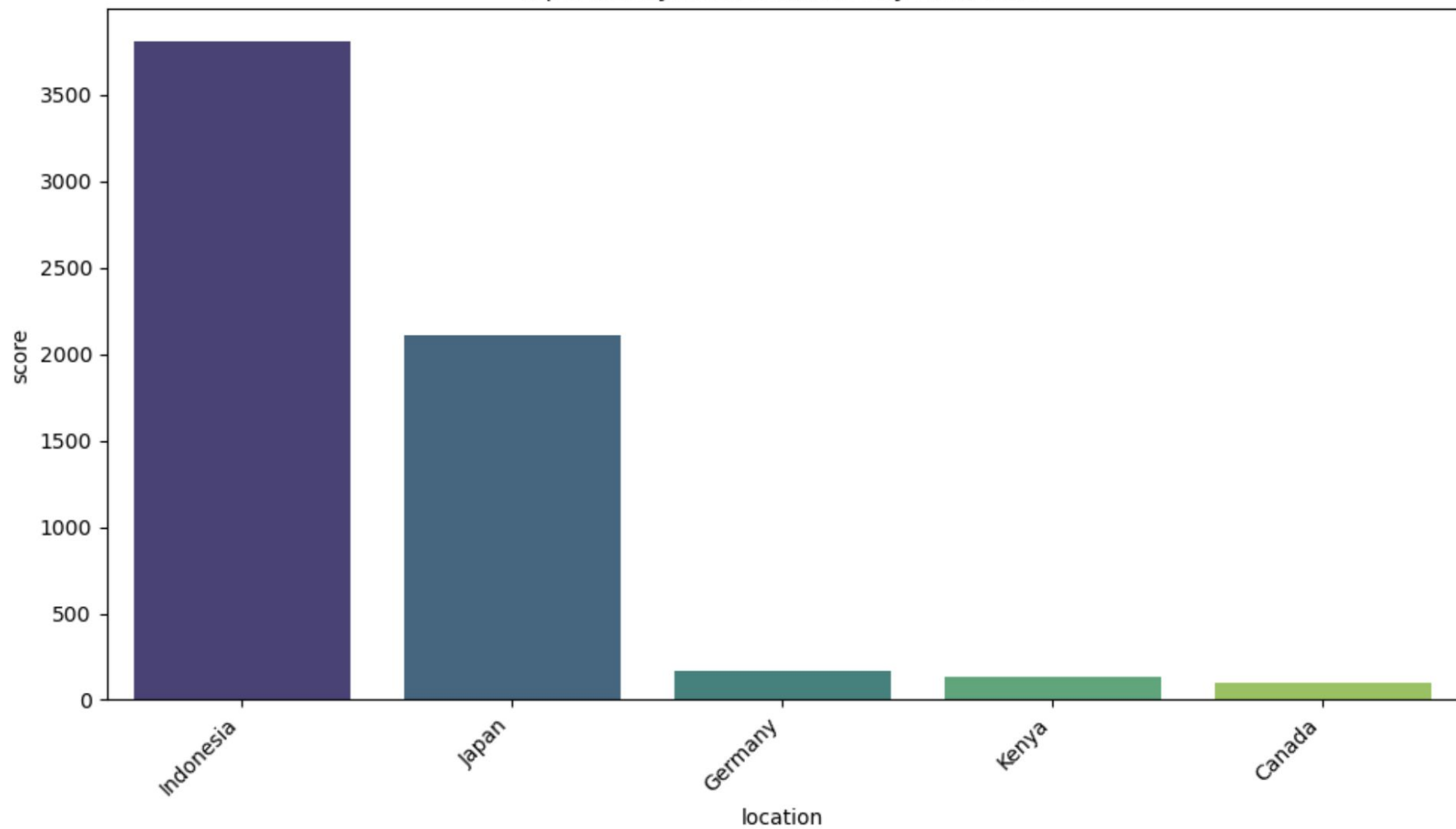
Honeymoon Destination Extraction

- Extracted location entities using spaCy and matched them to valid countries using `pycountry`.
- Aggregated sentiment scores by country to determine top and bottom honeymoon destinations.

Results & Visualizations

- Identified top 5 countries based on average sentiment.
- Bar plots highlighted the most loved appreciated destinations among users.

Top 5 Honeymoon Countries by Sentiment



Recommendations

- Optimize model performance through hyperparameter tuning like grid search, Bayesian optimization, with k-fold cross-validation for better generalization.
- Group by continent, country, cities and other famous tourist spots
- Gather more data to include features to predict 'Travel Mode', 'Mode of Staying' and many more
- Enhance predictions by analyzing user preferences and trending destinations.

Thank you