# **Hospitality Guest Review Dataset Documentation**

## **1. Dataset Overview**

The **Hospitality Guest Review Dataset** consists of guest reviews for hotels, providing insights into customer experiences, feedback, and ratings. It includes a variety of features such as hotel information, reviewer details, review content, and various metrics associated with the reviews. The dataset is designed to be used for analysis and modeling purposes to understand customer sentiments, preferences, and hotel service performance.

## **2. Dataset Structure**

The dataset consists of **X rows** (observations) and **Y columns** (features). The columns represent different attributes related to the hotel reviews and the guests providing those reviews.

### Data Content

The csv file contains 17 fields. The description of each field is as below:

* Hotel\_Address: Address of hotel.
* Review\_Date: Date when reviewer posted the corresponding review.
* Average\_Score: Average Score of the hotel, calculated based on the latest comment in the last year.
* Hotel\_Name: Name of Hotel
* Reviewer\_Nationality: Nationality of Reviewer
* Negative\_Review: Negative Review the reviewer gave to the hotel. If the reviewer does not give the negative review, then it should be: 'No Negative'
* Review\_Total\_Negative\_Word\_Counts: Total number of words in the negative review.
* Positive\_Review: Positive Review the reviewer gave to the hotel. If the reviewer does not give the negative review, then it should be: 'No Positive'
* Review\_Total\_Positive\_Word\_Counts: Total number of words in the positive review.
* Reviewer\_Score: Score the reviewer has given to the hotel, based on his/her experience
* Total\_Number\_of\_Reviews\_Reviewer\_Has\_Given: Number of Reviews the reviewers has given in the past.
* Total\_Number\_of\_Reviews: Total number of valid reviews the hotel has.
* Tags: Tags reviewer gave the hotel.
* days\_since\_review: Duration between the review date and scrape date.
* Additional\_Number\_of\_Scoring: There are also some guests who just made a scoring on the service rather than a review. This number indicates how many valid scores without review in there.
* lat: Latitude of the hotel
* lng: longtitude of the hotel

*In order to keep the text data clean, I removed unicode and punctuation in the text data and transform text into lower case. No other preprocessing was performed.*

## **3. Data Preprocessing and Cleaning Documentation**

This document outlines the steps involved in preprocessing and cleaning the hospitality review dataset.

#### **1. Loading the Dataset**

* The dataset, Hotel\_Reviews.csv, is read into a pandas DataFrame for further manipulation and analysis.

#### **2. Initial Exploration**

* The first few rows of the dataset are examined to get an overview of the data structure.
* The shape of the dataset (i.e., the number of rows and columns) is checked to understand its size.
* The columns and their respective data types are summarized to identify the types of data (e.g., numeric, text) in each column.

#### **3. Identifying and Handling Missing Values**

* Missing values in the dataset are identified by checking each column for null or NaN entries.
* For the **Negative\_Review** and **Positive\_Review** columns, missing values are filled with default text values ("No Negative" and "No Positive") to ensure completeness in the text columns.
* The **Reviewer\_Score** column, which is numeric, has missing values filled with the column's mean value. This is done to avoid data loss while maintaining the integrity of the numerical data.

#### **4. Removing Irrelevant Columns**

* The **lat** and **lng** columns, which represent geographic coordinates, are dropped from the dataset. These columns are deemed irrelevant for the analysis due to a significant number of missing values in those columns.

#### **5. Final Inspection of the Cleaned Data**

* After the missing values are handled and irrelevant columns are removed, the cleaned dataset is reviewed to ensure that the preprocessing steps have been applied correctly. A sample of the data is displayed to verify the changes.

#### **6. Saving the Cleaned Dataset**

* The cleaned dataset is saved to a new CSV file (cleaned\_hospitality\_data.csv). This ensures that the original dataset remains unchanged, and the processed data can be used for further analysis or modeling without the need for repeated preprocessing.

### **Final Columns in the Cleaned Dataset**

After preprocessing and cleaning the dataset, the following columns remain in the final dataset:

1. **Hotel\_Address**: The address of the hotel.
2. **Additional\_Number\_of\_Scoring**: The number of additional scoring items for each review.
3. **Review\_Date**: The date when the review was written.
4. **Average\_Score**: The average score of the hotel based on the review.
5. **Hotel\_Name**: The name of the hotel being reviewed.
6. **Reviewer\_Nationality**: The nationality of the reviewer.
7. **Negative\_Review**: The text of the negative review or "No Negative" if no negative review was provided.
8. **Review\_Total\_Negative\_Word\_Counts**: The total word count of the negative review.
9. **Total\_Number\_of\_Reviews**: The total number of reviews for the hotel.
10. **Positive\_Review**: The text of the positive review or "No Positive" if no positive review was provided.
11. **Review\_Total\_Positive\_Word\_Counts**: The total word count of the positive review.
12. **Total\_Number\_of\_Reviews\_Reviewer\_Has\_Given**: The total number of reviews written by the reviewer.
13. **Reviewer\_Score**: The score given by the reviewer, typically based on their overall experience.
14. **Tags**: A list of tags describing the type of trip or reviewer profile (e.g., "Leisure trip", "Couple").
15. **days\_since\_review**: The number of days since the review was posted.

The columns related to geographic coordinates (**lat**, **lng**) were dropped, as they contained many missing values and were not relevant for the analysis.