## ****Hotel Review Data Cleaning Documentation****

### ****1. Data Source****

* **Source**: Kaggle - [Hotel Reviews Dataset](https://www.kaggle.com/datasets/jiashenliu/515k-hotel-reviews-data-in-europe)
* **Description**: This dataset contains customer reviews for hotels, with details like review score, text feedback, and hotel information. It’s ideal for sentiment analysis, review classification, and improving customer experiences in the hotel industry.

### ****2. Dataset Attributes****

The dataset contains the following attributes:

* **Hotel\_Address**: Address of the hotel.
* **Additional\_Number\_of\_Scoring**: Additional scoring data for reviews.
* **Review\_Date**: Date when the review was posted.
* **Average\_Score**: Average score given by the reviewer.
* **Hotel\_Name**: Name of the hotel.
* **Reviewer\_Nationality**: Nationality of the reviewer.
* **Negative\_Review**: Text of the negative review.
* **Review\_Total\_Negative\_Word\_Counts**: Total number of negative words in the review.
* **Total\_Number\_of\_Reviews**: Total number of reviews for the hotel.
* **Positive\_Review**: Text of the positive review.
* **Review\_Total\_Positive\_Word\_Counts**: Total number of positive words in the review.
* **Total\_Number\_of\_Reviews\_Reviewer\_Has\_Given**: Total number of reviews the reviewer has given.
* **Reviewer\_Score**: Numerical score (1-10) given by the reviewer.
* **Tags**: Tags associated with the review (e.g., trip type).
* **days\_since\_review**: Number of days since the review was posted.
* **lat**: Latitude of the hotel's location.
* **lng**: Longitude of the hotel's location.

### ****3. Data Cleaning Process****

#### ****3.1. Handling Missing Values****

* **Text Columns**: Missing values in Negative\_Review and Positive\_Review are filled with "No Negative" and "No Positive" placeholders.
* **Numerical Columns**: Missing values in Reviewer\_Score are filled with the column's mean value.

#### ****3.2. Unnecessary Columns Removal****

* The lat and lng columns (latitude and longitude) are dropped as they are irrelevant for review analysis.

#### ****3.3. Combining Reviews****

* The Negative\_Review and Positive\_Review columns are merged into a single review column. This simplifies the data and allows for future use of a large language model (LLM) to classify the sentiment (positive or negative) within the review column itself.

#### ****3.4. Text Cleaning****

* Leading and trailing whitespaces in the review column are removed.
* Empty review entries are removed to ensure consistency and avoid any noise in the data.

#### ****3.5. Dropping Redundant Columns****

* After combining reviews into the review column, the original Negative\_Review and Positive\_Review columns are dropped to avoid redundancy.

### ****4. Final Output****

After cleaning, the dataset consists of:

* **Hotel\_Name**: Hotel name.
* **Reviewer\_Score**: Reviewer score.
* **Review**: A consolidated review column containing both positive and negative feedback, ready for sentiment analysis.

The cleaned dataset is saved as cleaned\_hospitality\_data.csv, ready for further analysis or machine learning tasks.