Hotel Reservations: Data Cleaning

# What is Data Cleaning?

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled.

If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset. But it is crucial to establish a template for your data cleaning process so you know you are doing it the right way every time.

# Steps involved in data cleaning:

* Remove duplicate or irrelevant observations
* Fix structural errors
* Filter unwanted outliers
* Handle missing data

# Data Cleaning In Dataset:

**Dealing with Duplicates:**

* Duplicate observations are often encountered during data collection. In our dataset,

consisting of 36275 rows and 18 features, we identified and addressed 10275.

* duplicated rows to ensure data integrity.

## Handling Missing Data:

* + We observed there are no missing values in Dataset.

## Leading Whitespace Removal:

* + In our dataset, 'object' type column names and string values often started with leading whitespace.
  + To enhance data consistency, we systematically removed these leading whitespaces.

## Data Validation and Outlier Handling:

* In the data no null values are observed out of 26000 data points.
* 10275 duplicates are observed in the data and those are removed.
* The values of the column named "avg\_price\_per\_room" is converted to “int”.
* Outliers are observed in the columns “lead\_time”,” avg\_price\_per\_room” and those were treated by using IQR method.

## Feature Encoding:

## In the target variable column named “booking\_status” the features are encoded like cancelled as 1 and not cancelled as 0.

# Conclusion:

* The data cleaning step is crucial for preparing the data for analyzing and modelling, ensuring that the data is accurate, consistent and reliable and it improves the quality of the data that leads to the accurate model building.
* As we performed data cleaning in our data our data quality has improve.