**Assessment: Data Cleaning & Exploration of Booking Data**

**By: Shubham Rampurkar**

**Objective:**

The goal of this assignment is to assess your data analysis, problem-solving, and deriving helpful insights. The purpose of this report is to document the data cleaning and exploration process for the Booking Dataset. The dataset contains booking information, including Service Name, Facility, Booking Type, Price, Duration, and other relevant details. The goal of this process was to clean the dataset by handling missing values, duplicates, incorrect data formats, and outliers, ensuring it is ready for further analysis and visualization.

**Process:**

**Data Loading and EDA (Exploratory Data Analysis):**

* The dataset was loaded using Pandas from an Excel file.
* The first few rows were displayed using df.head() to understand its structure.
* df.info() was used to check data types and missing values.
* df.describe() was used to analyze numerical statistics.
* The number of unique values in each column was checked using df.nunique().

**Data Cleaning Process:**

* The data cleaning process focused on removing inconsistencies and handling missing values to ensure data quality.
* Columns that did not hold any meaningful values, such as Subscription Type, were removed from the dataset.
* Missing values were carefully addressed based on their relevance to the data.
* In the Class Type column, which contained values associated with class bookings, null values were filled with "Not a Class" to maintain consistency.
* The Facility column was closely related to Class Type, as one was null when the other had a value. This relationship was taken into account while filling missing values to ensure data accuracy.
* The Instructor column had missing values when the booking was not for a class.
  + If the Booking Type was something other than a class, such as a party, the instructor value was filled with "No Need".
  + If the Booking Type was a class, and the instructor value was missing, it was replaced with "Unknown" to maintain consistency.
* For missing values in the Time Slot column, they were replaced with "Not Specified" to indicate that no specific time was assigned to the booking.
* In the Duration and Price columns, missing or ‘0’ values were replaced with the median value to ensure that the data remained balanced without being affected by extreme variations.

**Data Observations:**

* The dataset contained booking information for facility providers and class-based services, such as small events, dance, and gymnastics classes.
* It included details related to the facilities offered, booking charges, and service types.
* Key attributes in the dataset included instructor names, customer details, and contact information, providing insights into customer preferences and service allocation.
* Additionally, the dataset contained themes selected by customers.
* The Duration and Time Slot information specified the length of each booking and the preferred timing selected by customers.

**Data Discrepancies and Inaccuracies:**

* The columns Service Name, Facility, Class Type, Booking Type, and Service Type contained highly similar data.
* These columns could not be removed, as they might hold different values before or after booking confirmation.The column names did not clearly indicate their purpose, which could lead to confusion in analysis.
* The dataset contained missing values in multiple columns, including Facility, Class Type, Instructor, Time Slot, Theme, Duration, and Price.
* Facility and Class Type were interdependent, where one was missing when the other was present.
* The Instructor column had missing values when the booking type was not a class, indicating that an instructor was not needed in those cases.
* Time Slot values were missing, likely due to incomplete data entry, requiring a default "Not Specified" value.
* Duration and Price contained missing values, which were replaced with their respective median values to maintain consistency.