Task -Data Checking

Python

Before developing the predictive model for hotel cancellations, we will conduct preliminary data analysis. This involves checking for missing values, identifying outliers, and ensuring appropriate data types for each column. Handling missing values and outliers strategically will ensure a reliable dataset for accurate modeling.

- ✓ First, load the dataset and check for:
- ✓ Missing values: Use the appropriate function to check if there are any missing values in the dataset. If there are, decide on the best strategy to handle them based on the nature of the data.
- ✓ Outliers: Check for outliers in the dataset. These can be identified using various techniques, such as boxplots, scatterplots, or Z-scores. If there are any outliers, decide on the best strategy to handle them.
- ✓ Data types: Check the data type of each column. Ensure that the data type is appropriate for the data it represents.

Outcomes:

1. Firstly dataset was loaded and basic assessment of the data was done.
   - With column naming found an inconsistant column named 'Flight ID' which was not in same format as the other columns , fixed that to Flight_ID to standardize the column names
2. Verified data types of each column, all data types were correct
3. Checked for missing values and found no missing values in the dataset
4. Used Boxplots to visually detect outliers and two columns showed outliers named 'Flight_Distance' and "Previous_Flight_Delay_Minutes'
5. Handled Flight_Distance column outliers using Capping method because it reduces the impact of extreme outliers, which can distort the analysis and for Previous_Flight_Delay_Minutes used Log Transformation because data is skewed, compressing the range of delay times will reduce the impact of extreme values.