ALY6110：Data Management and Big Data

Module 1 Assignment：Data Set Selection

Group members：Chen-Yu Chu，Sean Mclean，Yin Tang，Jingyi Wang，Runcheng Yang

Instructor：Ajit Appari

May 28，2024

Dataset: Hotel Booking Demand

Dataset Overview: The dataset used for this analysis is the Hotel Booking Demand dataset. It contains booking information for a city hotel and a resort hotel and includes the following details:

Rows: 119,390

Columns: 32

- Hotel Type: Indicates whether the booking is for a city hotel (H2) or a resort hotel (H1).
- Is Canceled: Indicates if the booking was canceled (1) or not (0).
- Lead Time: Number of days between booking and arrival.
- Arrival Date: Includes the year, month, week number, and day of the month.
- Stays in Weekend Nights: Number of weekend nights booked.
- Stays in Week Nights: Number of week nights booked.
- Number of Adults, Children, and Babies: Indicates the number of guests.
- Booking Changes: Number of changes to the booking.
- Deposit Type: No deposit, non-refundable, or refundable deposit types.
- Special Requests: Number of special requests made.

The dataset is substantial with over 500,000 observations, making it suitable for our analysis and predictive modeling.

Problem/Question:
Hotel booking cancellations are a significant issue for the hospitality industry, leading to lost revenue and inefficient resource allocation. Predicting cancellations effectively can help hotels optimize their bookings, adjust pricing dynamically, and improve overall customer service.

Methodology:

Data Collection and Preparation： The dataset will be cleaned and preprocessed to handle missing values and ensure data consistency. Exploratory Data Analysis (EDA) will be conducted to understand the distribution of variables and identify any patterns or anomalies.

Feature Engineering： Relevant features will be engineered from the existing variables to enhance the predictive power of the model. For instance, calculating the lead time, creating categorical variables for different customer types, and aggregating stay details.

Modeling： Various machine learning models, such as logistic regression, decision trees, and random forests, will be applied to predict the likelihood of booking cancellations. Model performance will be evaluated using metrics such as accuracy, precision, recall, and F1 score. **Interpretation and Insights**： The results will be interpreted to identify the most significant factors contributing to cancellations. These insights will help in formulating strategies to reduce cancellations, such as targeted marketing campaigns for high-risk customers or adjusting booking policies.

Reference：

Mostipak, J. (2019). Hotel booking demand. Kaggle. https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand/data