Assumptions and Exploratory Data Analysis (EDA)

1. Data Cleaning Assumptions and Handling Logic

A. Date Handling

- Swapping `checkInDate` and `checkOutDate` if checkOutDate < checkInDate
  - Reason: 19,556 rows had inverted dates, likely due to data entry errors.
  - Action: Swapped values to preserve data integrity.
- Removing stays longer than 90 days
  - Reason: Some rows had unreasonably long stays (>300 days), assumed to be outliers or errors.
  - Action: Rows with `stayLength > 90` days were removed.
- Ensuring `searchDate < checkInDate`
  - Reason: Prevent logically impossible searches that happen after a check-in.
  - Action: If found, `searchDate` was set to `checkInDate - 1 day`.

B. Missing Values and Invalid Inputs

- Missing `destinationName`
  - 18,730 missing rows were filled with "UNKNOWN".
- Negative or Placeholder Values (-1)
  - `starLevel`, `customerReviewScore`, and `reviewCount` had -1 values assumed to represent missing data.
  - Replaced with `NaN`, and rows with critical missing values were later dropped.
- Outlier Price Capping
  - `minPrice` and `minStrikePrice` capped at 99th percentile to reduce skew.
- Ensuring `rank >= 1`
  - Minimum rank should start from 1, clipped values accordingly.

2. EDA Summary Insights

A. General Dataset Overview

- Rows: 769,109 initially, ~734,000 after cleaning.
- Columns: Dates, categorical fields, numeric pricing and review metrics, binary behavior labels.

B. Booking and Click Statistics

- Booking Rate: ~3.72%
- Click Rate: ~4.58%

C. Impact of Hotel Features

- Free Breakfast Effect:
    - Booking rate: 3.87% (vs 3.60% without)
- Free Internet Effect:
    - Booking rate: 3.73% (vs 3.57% without)

D. User Behavior

- Signed-in users click more: ~4.65% click rate
- Most searches use MOBILEWEB and DESKTOP

E. Top Destinations

- Las Vegas, Orlando, New York, Atlanta, and Miami top the list by volume.

F. Data Quality Concerns (Addressed)

- Garbled IDs (searchId, userId, brandId) retained as-is but treated categorically.
- Negative review scores and star levels removed or imputed.

G. Feature Engineering Conducted

- `stayLength`, `leadTime`, `weekendStay`, `highRank`, `discountPercent`, `logMinPrice`
- Frequency encoding of destination popularity.
- One-hot encoding of deviceCode and vipTier.