now the cleaning:

1. when i was cleaning i noticed the column country has 488 missing values out of 119,390 values which is approximately 0.41% since my geographical analysis heavily depends on it. I filled it with ‘Unknown’ why:

* To preserve those 488 rows for other analysis
* So I don’t distort country-level statistics by falsely assigning them to any specific country
* So I can analyze the unknown group separately to see if there’s pattern

1. I converted my Date columns to Date Time because they are currently stored as object (i.e just plain text or string format) which limit what I can do with them
2. I checked for Duplicates to ensure there are no repeated records
3. VALIDATING & FIXING MY DATA TYPES
4. **Converted my booking ID to ‘string’:** BecauseI’m not performing math on it, treating it as a string prevents accidental numerical operations
5. **Converted Avg Daily rate, Revenue, Revenue Loss to Float:** Because they are monetary values, keeping them as int64 might hide decimals for price-related analysis, float ensures precision
6. **Lead time, Nights Used and Guests are in integer:** It is Okay but I checked to ensure no negative values
7. **Converted Hotel Distribution channel, Customer Type, Deposit Type, Country, Status to ‘Category’ :** Since it’s a qualitative variable with repeated values, I used Category to tell people it’s not a free form text or numerical data to prevent error in analysis
8. VALIDATING BUSINESS LOGIV

Checking if the data make sense based on real- world rules or expectation

1. If a booking is labeled ‘Cancelled’, it should have the binary cancellation flag set to 1 (checked for mismatch)
2. If Revenue = 0, is Revenue loss >= 0 ?

If the hotel made no revenue, it could mean: The guest cancelled last minute…so you lost money (Revenue Loss > 0) or the booking was valid but the room was free (Revenue loss = 0).

1. Are there rows where guests == 0 ?

Found out there were illogical rows:

* Checked out bookings with 0 guests
* Some have non-zero revenue which should not happen

I dropped it completely because:

* They don’t follow business logic
* They are few (180 out of 119,390 just 0.15%)
* They could mislead my analysis