**Exploratory Data Analysis Report on Hotel Booking Dataset**

**1. Dataset Introduction**

The dataset contains 119,390 hotel booking records across 32 columns. It includes a variety of features about the hotel type, customer demographics, booking preferences, length of stay, pricing, cancellations, and the status of reservations.

Key Feature Categories:

* Date Columns: arrival\_date, reservation\_status\_date
* Categorical Columns: hotel, meal, country, market\_segment, deposit\_type, etc.
* Numerical Columns: lead\_time, adr, stays\_in\_week\_nights, etc.
* Target-related: is\_canceled, reservation\_status, booking\_changes

**2. Data Cleaning & Preprocessing**

**Handling Missing Values:**

* **children**: Had just 4 missing values. Filled using **median** to avoid distortion from outliers.
* **country**: Around 488 missing entries. Imputed using **mode** (most common country).
* **agent**: ~16,340 missing. Replaced with **0** assuming no agent was involved. Converted to **integer**.
* **company**: Over 94% missing values. This column was **dropped** since imputation wouldn’t make sense and its predictive power was doubtful.

**Data Type Fixes:**

* Several object-type columns (like meal, country, etc.) were converted to **category** for memory efficiency and easier analysis.
* The reservation\_status\_date and newly created arrival\_date columns were converted into proper **datetime** formats for time-based analysis.

**Feature Engineering:**

* Created a new column **arrival\_date** by combining **arrival\_date\_year**, **arrival\_date\_month**, and **arrival\_date\_day\_of\_month.**
* Dropped the now redundant columns (**arrival\_date\_year**, **arrival\_date\_month**, **arrival\_date\_day\_of\_month**).
* Introduced a new column **total\_nights** to capture the **total stay duration**, combining **stays\_in\_week\_nights** and **stays\_in\_weekend\_nights.**

These transformations allowed more streamlined analysis and improved interpretability of the dataset.

**3. Univariate Analysis & Observations**

**Hotel Type:**

Most bookings were made at the **City Hotel**, indicating a strong preference among guests for urban accommodations, possibly for business or tourism. Resort Hotels were fewer, often reflecting leisure bookings or seasonal travel.

**Lead Time:**

Bookings varied significantly in lead time — ranging from same-day to over 500 days in advance. The distribution is **right-skewed**, with a long tail. Most customers book within the 0–100 day window, but some bookings are made far in advance, perhaps group or event bookings.

**Cancellation Patterns:**

The dataset revealed that **around 37% of bookings were canceled**. This is a surprisingly high number and suggests:

* Flexible cancellation policies.
* Price shopping behavior.
* Potential for overbooking strategies.

Further analysis showed **City Hotels experienced more cancellations** than Resort Hotels. This might be because city hotel stays are often work-related and can be rescheduled.

**ADR (Average Daily Rate):**

The ADR has significant variability, from very low to extremely high rates. High ADRs are sometimes associated with canceled bookings. This might indicate either luxury bookings that are harder to commit to or issues with pricing strategy.

**Customer Origin:**

The largest share of guests came from **Portugal**, followed by **UK**, **France**, **Spain**, and **Germany**. This suggests a **European traveler base**, and it can help design regional promotions.

**Meal Preferences:**

The most common meal plan was **BB (Bed and Breakfast)**. Other options like **HB (Half Board)** and **FB (Full Board)** were less chosen, possibly due to the hotel types or guest habits.

**➤ Correlation Matrix (Numerical)**

* lead\_time is **moderately correlated** with is\_canceled and booking\_changes.
* special\_requests and booking\_changes have a **slight negative correlation with cancellations** — these guests tend to show up.
* No strong correlation among stay length, guests, or room counts.

**➤ Cancellation Rate by Hotel Type**

* **City Hotel:** ~40% cancellations
* **Resort Hotel:** ~28% cancellations  
  Shows that city hotels are more prone to cancellations — possibly due to flexible travel plans or business reasons.

**➤ Lead Time vs Cancellation**

* **0–30 days**: Lowest cancellation
* **120+ days**: High cancellation (~60% in some bins)  
  Suggests early planners are more likely to cancel — maybe due to better deals later.

**➤ ADR by Market Segment**

* **Corporate and Direct:** Highest ADR
* **TA/TO:** Lowest, due to package deals or discounts  
  Helps hotels design pricing based on segment.

**➤ Special Requests and Cancellation**

* Guests making **more than one request** rarely cancel.
* Could serve as a behavioral predictor for no-show likelihood.

**➤ Reservation Trends by Month**

* **July, August, and October:** Booking peaks
* **January, February:** Cancellation peaks  
  Seasonality impacts both occupancy and business risk.

**4. Bivariate & Multivariate Insights**

**🔹Correlation Matrix:**

* lead\_time ↑ → is\_canceled ↑ (moderate positive).
* special\_requests ↓ → is\_canceled ↓ (committed guests cancel less).
* total\_guests, stays, and adr are largely uncorrelated — behavior-driven.

**🔹 Hotel vs Cancellation Rate:**

* City Hotel: ~40% cancellations
* Resort Hotel: ~28%  
  Business plans are more prone to last-minute change than vacation ones.

**🔹 Lead Time vs Cancellation:**

* 0–30 days: Low cancellation
* 120+ days: Up to 60% cancel  
  Suggests price-sensitive or event-based bookings prone to revision.

**🔹 ADR by Market Segment:**

| Segment | ADR Trend |
| --- | --- |
| Corporate | High ADR (business bookings) |
| Direct | Higher ADR (no commissions) |
| TA/TO | Lowest ADR (bulk deals/packages) |

**🔹 Special Requests vs Cancellation:**

* Guests with >1 request rarely cancel.
* Indicates commitment and potential for loyalty targeting.

**5. Time-Based Trends**

Using the cleaned arrival\_date and reservation\_status\_date, the dataset was analyzed for **seasonality**.

* The **peak months** for bookings were **July, August, and October**, consistent with European summer and early autumn holidays.
* **Off-peak periods** like January and February saw fewer guests, hinting at seasonal business fluctuations.
* For **Resort Hotels**, bookings are highest in summer months, aligning with vacation travel.
* **Cancellations** often occur shortly before the reservation date, especially in off-seasons.

 **High booking months**: July, August, October

 **High cancellation months**: January, February

 **City Hotel** is more consistent year-round.

 **Resort Hotel** spikes in summer — shows **strong seasonality**.

**6. Guest Behavior by Segment & Country**

**Market Segments:**

* **TA/TO (Travel Agent/Tour Operator)** is the largest segment — price-sensitive, lower ADR.
* **Corporate and Direct** are more profitable — fewer cancellations, higher ADR.

**Country Patterns:**

* International guests (e.g., UK, Germany) **book earlier and stay longer**.
* Local guests (e.g., Portugal) have shorter lead times and lower cancellations.

**7. Market Segment & Booking Behavior**

* **Market Segments**: Majority of bookings came through **TA/TO (Travel Agents/Tour Operators)**.
* **Distribution Channels**: Online portals were dominant, underscoring the importance of digital marketing and third-party platforms.
* **Special Requests**: Most guests made **0 or 1 special request**, but those who made more (e.g., 2+) were **less likely to cancel**, suggesting higher intent to follow through with the stay.

**8. Booking Changes & Room Assignments**

* Many bookings underwent changes in the number of guests, rooms, or other preferences.
* Often, the **reserved and assigned room types did not match**, suggesting frequent upgrades or reallocations by the hotel.
* This mismatch can affect satisfaction, and further study could link this to cancellation or review scores if available.

**9. Deposit Type and Payment Behavior**

* **No Deposit** was the most common type. Cancellations were significantly higher in this group.
* **Non-refundable bookings** had much lower cancellation rates, indicating that **financial commitment leads to fewer cancellations**.

| **Deposit Type** | **Cancellation Rate** |
| --- | --- |
| No Deposit | Very high |
| Non-refundable | Very low |
| Refundable | Moderate |

* Guests without financial commitment cancel more.
* Introduce **incentives for prepayment** to reduce volatility.

**10. Final Insights and Recommendations**

1. **Cancellation Modeling Needed**: With such high cancellation rates, a prediction model could help flag risky bookings early. Hotels could then send confirmations, restrict payment options, or offer discounts for commitment.
2. **Pricing Optimization**: ADR shows huge variance. Applying machine learning to optimize price based on season, customer segment, and lead time could improve revenue.
3. **Seasonal Strategy**: Staff planning, marketing, and room upgrades should be adjusted for seasonality. Resort Hotels need extra support in summer, while City Hotels might maintain steadier flows.
4. **Digital Channel Dominance**: Since most bookings are made online, investing in SEO, OTA partnerships, and review management will directly impact bookings.
5. **Special Requests as Loyalty Indicator**: Track and reward customers who make more requests — they’re more likely to follow through and possibly return.
6. **Room Allocation**: The mismatch between reserved and assigned rooms could lead to dissatisfaction. A tracking system could help better manage inventory and guest experience.

**Business Questions & Insights :**

1. **What influences ADR the most?**
   * **Hotel type**, **market segment**, and **lead time** heavily influence ADR.
   * City hotels usually charge more than resort hotels, and bookings made via corporate or direct channels have higher ADRs.
   * Lead time has a moderate positive correlation — last-minute bookings often have lower ADRs due to discounts.
2. **Do guests who book earlier tend to request more changes?**
   * Yes, guests with longer **lead times** show higher **booking changes** on average.
   * Longer planning windows allow more time for changes — possibly reflecting uncertainty or extended planning.
3. **Are there pricing or booking differences across countries?**
   * Definitely. Guests from Western European countries (like UK, Germany, France) tend to have higher ADRs.
   * Some countries (e.g., Portugal) dominate in volume but not in high revenue.
   * Cultural or economic differences can impact spending behavior and length of stay.
4. **Is there a pattern in room upgrades or reassignment?**
   * Yes, many bookings show **reserved room type ≠ assigned room type**.
   * Upgrades may occur when better rooms are available, or due to overbooking.
   * Reassignments are more frequent in City Hotels.
5. **Are reserved room types consistently matched with assigned room types?**
   * No. Around **20–25%** of bookings show mismatches.
   * Most common when original room type is overbooked or guest is upgraded.
   * Tracking this helps understand inventory issues or upgrade strategies.
6. **What are the most common guest demographics (e.g., group size, nationality)?**
   * Most bookings are for **2 adults** with **no children**.
   * **Portugal**, **UK**, and **France** top the nationality list.
   * This suggests a strong domestic market and short-haul European tourism.
7. **Are there patterns in guest types (e.g., transient vs. corporate) that influence booking behavior?**
   * Yes. **Corporate guests** rarely cancel and often book fewer nights at higher ADRs.
   * **Transient or TA/TO guests** are more likely to cancel and stay longer.
   * Behavioral patterns vary strongly by segment.
8. **How does booking lead time vary across customer types and countries?**
   * Corporate guests tend to book at shorter notice.
   * Guests from far countries (USA, Australia) tend to book earlier than domestic travelers.
   * Longer lead times are more common for TA/TO segments.
9. **Are longer lead times associated with fewer booking changes or cancellations?**
   * No. Surprisingly, **longer lead times show higher cancellations** and more booking changes.
   * This is possibly due to more planning uncertainty or availability of better deals later.
10. **What is the typical duration of stay, and how does it vary by customer type or segment?**

* Most stays are **2–5 nights**.
* Corporate guests often stay 1–2 nights, while leisure guests (especially resort ones) stay longer.
* TA/TO guests have more extended stays, especially in Resort Hotels.

1. **How often are guests upgraded or reassigned to a different room type?**

* About **20–25% of guests** are reassigned or upgraded.
* Upgrades are more frequent for loyal or special-request guests, or in case of overbooked standard rooms.

1. **Are guests who make special requests more likely to experience booking changes or longer stays?**

* Yes. Special-request guests usually plan longer stays and occasionally change their bookings.
* These guests are also **less likely to cancel**, indicating stronger intent to complete the trip.

1. **Do certain market segments or distribution channels show higher booking consistency or revenue?**

* **Corporate and Direct** channels are the most consistent and profitable.
* OTA and TA/TO channels show high volume but lower revenue and higher cancellation rates.

1. **What factors are most strongly associated with higher ADR?**

* Key influencers: **hotel type**, **market segment**, **distribution channel**, and **lead time**.
* Bookings through direct and corporate channels, especially in City Hotels, yield higher ADR.

1. **Are there customer types or segments consistently contributing to higher revenue?**

* Yes. Corporate guests and direct online bookers contribute higher revenue despite lower booking volume.
* TA/TO segments bring more volume but lower revenue per booking.

1. **Do bookings with more lead time or from specific countries yield higher ADR?**

* Bookings from **longer distances or earlier planners** tend to have **moderate to high ADR**.
* Countries like **USA, Germany, and UK** have above-average ADR values.

1. **Are guests with higher ADR more likely to request special services or make booking modifications?**

* Yes, high-ADR guests often make **more special requests**.
* This implies they expect better service or personalization.
* However, they **rarely cancel**, showing reliability.

1. **Do guests from different countries behave differently in terms of booking timing or stay length?**

* Yes.
  + **Portuguese guests** often book closer to the date.
  + **UK/USA guests** book earlier and stay longer.
  + **Domestic guests** often stay shorter and book last-minute.

1. **Are guests who make booking changes more likely to request additional services or cancel?**

* Partially.
  + Guests with **more booking changes** sometimes request services, especially for longer stays.
  + However, **not all booking modifiers cancel** — it depends on lead time and guest type.

1. **Are guests who make special requests more loyal or committed to their bookings?**

* Yes.
  + These guests almost never cancel, making them **ideal for upselling or loyalty programs**.
  + They are more engaged and possibly returning customers.
* **Summary**

The EDA uncovered deep insights into customer behavior, revenue patterns, and hotel operations. High cancellations, variable pricing, and strong seasonality were key patterns.

A smart mix of **data-driven pricing**, **targeted customer retention**, and **proactive inventory management** can turn these insights into action for improved profitability and guest satisfaction.