# 1. Problem Statement: - Uncovering Patterns in Hotel Booking Data for Operational Efficiency and Revenue Growth

# 2. Overview

The goal of this exploratory data analysis is to investigate a Resort Hotel’s booking dataset to identify key patterns, trends, and relationships that can support data-driven decision-making.

This includes analyzing booking behavior, customer demographics, pricing strategies, and operational factors such as room assignments and special requests. The study will involve cleaning and preprocessing the data, examining variable correlations, and validating key business assumptions through hypothesis testing.

# 3. Core Objectives

* Understand how customer attributes and booking behaviors impact revenue.
* Identify trends in lead time, stay duration, and booking channels.
* Detects inconsistencies or anomalies in room allocation and guest handling.
* Explore relationships between booking patterns and customer satisfaction indicators.
* Evaluate whether specific operational or customer variables significantly affect outcomes such as ADR or room upgrades.

# 4. Scope of Work

* Handle missing values, duplicates, and outliers.
* Convert and standardize date and categorical fields.
* Perform univariate, bivariate, and time-series analysis.
* Conduct correlation analysis to identify influential variables.
* Apply hypothesis testing to validate business assumptions.
* Answer key operational and revenue-driving business questions.

1. **Data Cleaning and Preprocessing**

* **Code Actions Explained:**
* **pd.read\_csv('hotel\_bookings.csv'):** Loads the dataset.
* **df.info(), df.isnull().sum():** Inspects missing values and column types.
* **fillna():** Replaces nulls — 0 for numerical, "Unknown" for country.
* **drop\_duplicates():** Ensures no repeated records distort results.
* **astype('category'):** Optimizes memory and processing for categorical columns.

**Derived fields:**

* **total\_stay\_nights:** Combines weekend + weekday stays.
* **total\_guests:** Sums all guest types.
* **pd.to\_datetime(...):** Parses and standardizes date fields.

**Outlier treatment:**

* **df = df[df['adr'] < df['adr'].quantile(0.99)]:** Removes extreme ADRs.
* **df = df[df['lead\_time'] < df['lead\_time'].quantile(0.99)]:** Removes long lead times.

1. **Exploratory Data Analysis (EDA)**

* **Univariate visualizations:**
* **sns.histplot(df['adr']):** Understands ADR spread.
* **sns.histplot(df['lead\_time']):** Shows lead time behavior.
* **sns.countplot(x='customer\_type'):** Shows distribution of customer types.
* **sns.countplot(x='market\_segment'):** Displays channel preferences.
* **Time-Series Analysis:**
* **.dt.to\_period("M"):** Extracts monthly periods to plot booking trends.
* **Bivariate visualizations:**
* **barplot(adr by market\_segment):** Compares ADR across sources.
* **boxplot(lead\_time vs customer\_type):** Shows spread per type.
* **barplot(total\_guests.value\_counts()):** Reveals common group sizes.
* **countplot(country):** Shows top countries.
* **barplot(adr by segment and customer\_type):** Multi-facet comparison.
* **Multi variate visualizations:**

**barplot(adr by market\_segment and customer\_type):** Compares average ADR across booking channels segmented by customer type for multi-dimensional revenue insights.

1. **Correlation Analysis**

* Selected numerical variables like **adr, lead\_time, booking\_changes, special\_requests,** **etc.**
* Calculates strength of linear associations. **corr(method='pearson')**
* Highlights top correlators visually — e.g., ADR is moderately tied to lead time, special requests. **sns.heatmap().**

1. **Hypothesis Testing**

**T-Test (ADR differences):**

* + **ttest\_ind():** Tests ADR difference between Online TA and Direct.
  + Result interprets significance with p-value.

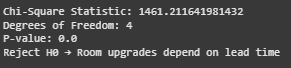
**Output: -**



**Chi-Square Test (Room upgrades vs. Lead time):**

* + **pd.cut():** Buckets lead time.
  + **chi2\_contingency():** Checks if upgrade frequency varies by booking timing.

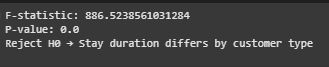
**Output: -**

****

**ANOVA (Stay duration vs. customer type):**

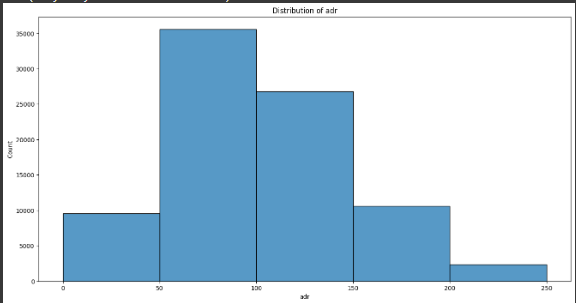
* + **f\_oneway():** Compares average stay across all types.

**Output: -**



**Graphs Summary: -**

**Graph 1: Distribution of ADR (Average Daily Rate)**



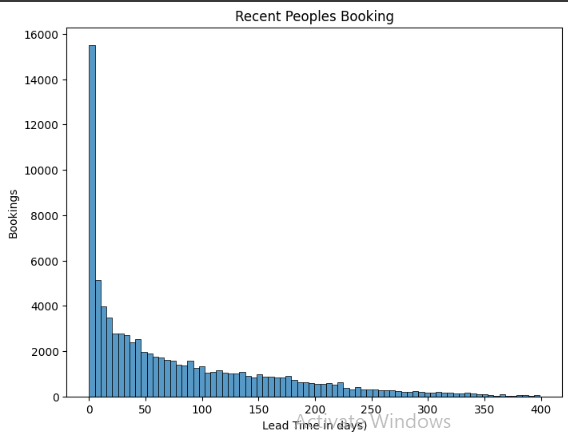
**Insights:**

* The ADR distribution is **right-skewed**, meaning most bookings are concentrated in the lower to mid-price ranges.
* The **peak frequency** occurs around the **€75–€100** range.
* There are a few **very high ADR values** (above €200), but these are rare.
* A notable **small spike** at €0 suggests either complimentary stays or data anomalies that may require review.

**Business Takeaway:**

* The majority of hotel stays are moderately priced.
* A long tail on the right indicates the presence of premium bookings or outliers.
* Understanding ADR segments can help tailor pricing strategies and revenue optimization.

**Graph 2: Distribution of Lead Time**



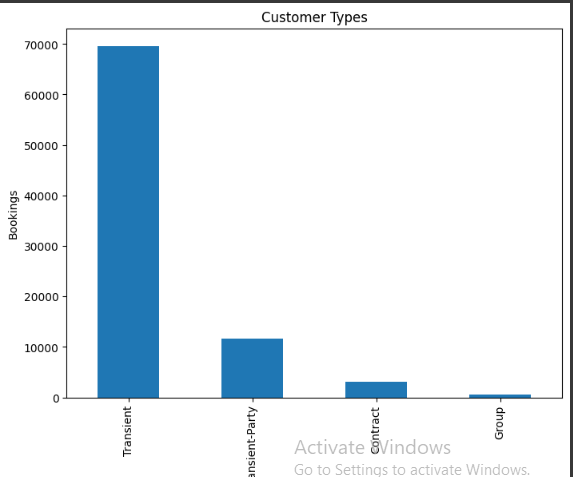
**Insights:**

* The distribution is **heavily right-skewed** with a sharp peak at **0–10 days**, especially at **0**, indicating many last-minute bookings.
* As lead time increases, booking frequency drops off rapidly.
* A long tail extends out to nearly 350 days, though these are very infrequet.

**Business Takeaway**:

* A **large proportion of bookings are made very close to the check-in date**, suggesting spontaneous travel or short decision windows.
* The presence of long lead times suggests some bookings are planned far in advance, possibly for events or group bookings.
* This insight is crucial for forecasting and managing inventory availability across different time windows.

**Graph 3: Count of Customer Types**



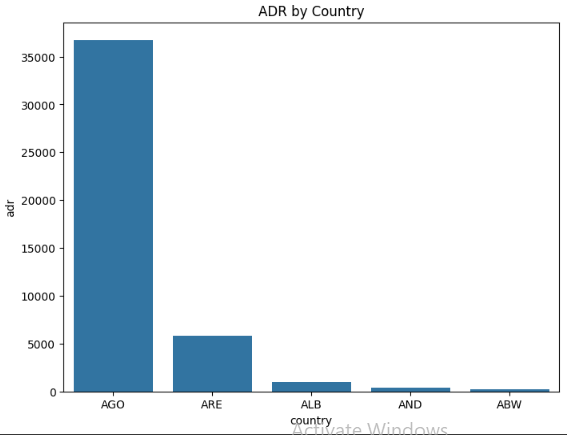
**Insight:**

* The majority of bookings come from **Transient** customers.
* **Transient-Party** is the second-highest, followed by **Contract** customers.
* **Group** bookings are minimal.

**Business Takeaway:**

* Focus marketing efforts on individual travellers (Transient), as they dominate the customer base.
* Design tailored loyalty programs and personalized offers to retain transient customers.
* Consider whether group sales strategies need improvement or are intentionally deprioritized.

**Graph 4: Distribution of Market Segment by Country**



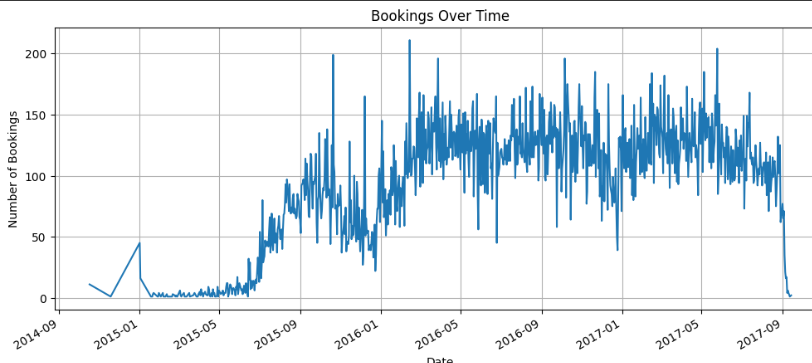
**Insight:**

* **Online Travel Agencies (OTA)** generate the highest volume of bookings by far.
* **Offline TA/TO** and **Direct bookings** also contribute significantly.
* Segments like **Corporate**, **Complementary**, and **Aviation** are relatively small.

**Business Takeaway:**

* Invest in OTA partnerships and improve visibility on platforms like Booking.com or Expedia.
* Promote direct booking benefits (e.g., discounts, free perks) to reduce commission costs.
* Evaluate lower-performing segments for growth potential or resource reallocation.

**Graph 5: Monthly Booking Trend**



**Insight:**

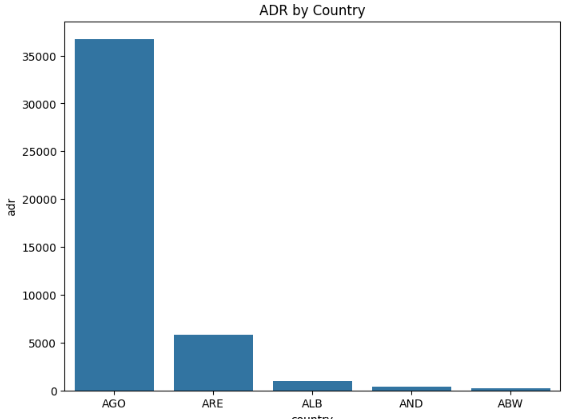
* Booking volumes show **seasonal trends**, with **peaks around mid-year** and **dips in late-year and early-year months**.
* There's a steady growth pattern across **2014 to 2017** despite fluctuations.

**Business Takeaway:**

This chart shows **hotel bookings over time** from **2014 to 2017**:

* **Rise**: Bookings increased steadily from mid-2015.
* **Peak**: High activity during 2016–2017.
* **Drop**: Sharp fall after mid-2017.
* Likely shows **seasonal trends** and growth in popularity.

**Graph 6: Average ADR by Market Segment**



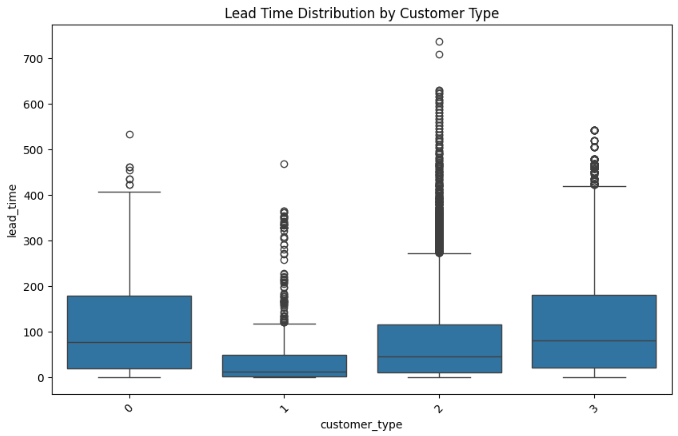
**Insight:**

* **Online TA** and **Direct** bookings yield the **highest ADR (Average Daily Rate)**.
* **Complementary** and **Undefined** bookings generate the **lowest ADR**, adding minimal revenue.
* **Corporate** and **Groups** have moderate ADRs.

**Business Takeaway:**

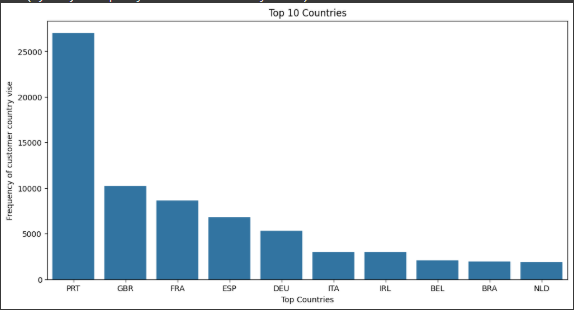
* Encourage Direct bookings through incentives—this channel is both high-volume and high-value.
* Review policies around low-ADR segments (e.g., Complementary stays) to ensure they are strategically justified.
* Target Online TA customers with upselling strategies to further maximize revenue.

**Graph 7:**  **Lead Time Distribution by Customer Type**



* **Insight**:
  + **Contract** and **Transient-Party** customers have the **longest lead times**, with median values around 70–80 days.
  + **Group** customers show a **very short lead time**, indicating last-minute bookings (median < 20 days).
  + **Transient** customers vary widely in lead time, including many long-lead bookings and outliers (up to 350+ days).
* **Business Takeaway**:
  + Contract bookings are planned well in advance—ideal for forecasting and revenue management.
  + Group bookings often come close to stay dates, implying limited time to upsell or reallocate resources.

**Graph 8: Top 10 Countries by Number of Bookings**



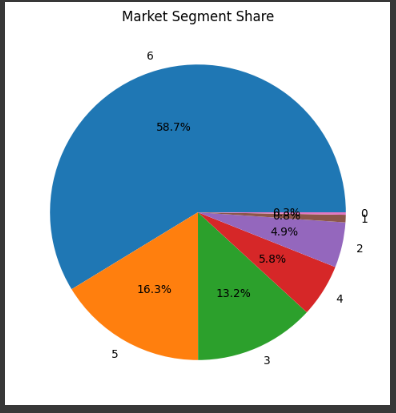
**Insight**:

* **Portugal (PRT)** dominates with the highest number of bookings, followed by **UK (GBR)**, **France (FRA)**, and **Spain (ESP)**.
* A sharp drop is seen after the top few countries.

**Business Takeaway**:

* Marketing efforts should prioritize PRT, GBR, FRA, and ESP.
* Language localization, pricing, and offers for these top countries could boost engagement further.

**Graph 9: Mean ADR by Market Segment and Customer Type**



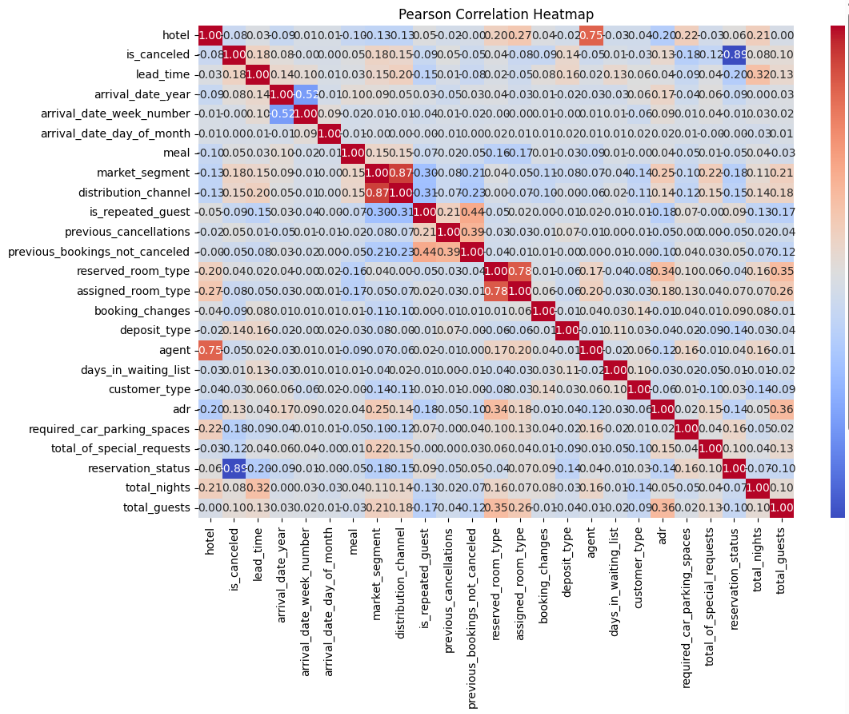
**Insight**:

* **Online TA** and **Aviation** segments generate the **highest ADRs**, especially for **Transient** and **Transient-Party** customers.
* **Complementary** and **Undefined** segments yield the lowest ADRs (as expected).
* **Group** customer type tends to show **lower ADRs** across segments.

**Business Takeaway**:

* Direct bookings (high ADR) and Online TA channels are profitable—invest in direct channel optimization and OTA partnerships.
* Consider yield strategies for group bookings to improve ADR without affecting volume.

**Graph 10: Correlation Heatmap (Pearson)**



**Insight**:

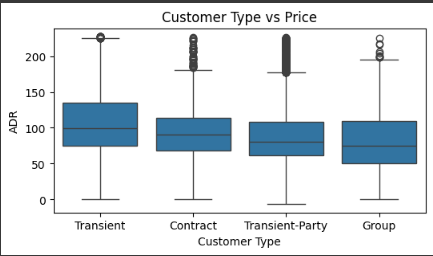
* **Lead time** has a moderate positive correlation with **is\_canceled** (0.19): longer lead times are more likely to cancel.
* **Stays in week and weekend nights** are strongly correlated (0.55), as expected.
* **Special requests** are **negatively correlated with cancellations** (-0.12): more requests = more committed guests.
* **Children and ADR** show a mild positive correlation (0.30): bookings with children may tend to cost more.

**Business Takeaway**:

* Cancellation policies may need to be stricter for long-lead bookings.
* Guests with special requests are more likely to follow through—consider these for upselling and loyalty strategies.
* Families (children) may represent higher-spending segments; target with tailored packages.

# Business Question

## What influences ADR the most?



Used **df.corr()['adr']** to find features most correlated with ADR, identifying key revenue drivers like lead time and special requests.

## Do guests who book earlier tend to request more changes?

Used **.corr()** to check if lead time positively relates to booking\_changes, showing planning vs. flexibility.

## Are there pricing or booking differences across countries?

Grouped ADR and lead time by country using **groupby().mean()** to identify high-paying or early-booking markets.

## Is there a pattern in room upgrades or reassignment?

Compared **reserved\_room\_type** and **assigned\_room\_type** to flag reassignments, and calculated upgrade rate.

## Are reserved room types consistently matched with assigned room types?

Used equality check and **.mean()** to determine how often booked room types match assigned ones.

## What are the most common guest demographics?

Used **value\_counts()** on group size and country to profile typical guests and top nationalities.

## Are there guest-type patterns in booking behavior?

Grouped ADR and lead time by **customer\_type** to compare planning and spend levels across guest types.

## How does booking lead time vary across customer types and countries?

Used **groupby().mean()** on lead\_time to explore differences in advance planning by group.

## Are longer lead times associated with fewer changes or cancellations?

Checked correlation between **lead\_time** and **booking\_changes**, **is\_canceled** to detect patterns of stability.

## What is the typical duration of stay, and how does it vary?

Created **stay\_duration** and grouped it by customer type and segment to compare visit lengths.

## How often are guests upgraded or reassigned?

Used boolean flag from room mismatch and calculated mean to get upgrade frequency.

## Are guests who make special requests more likely to change bookings or stay longer?

Used correlation of special requests with **booking\_changes** and stay duration to check behavior trends.

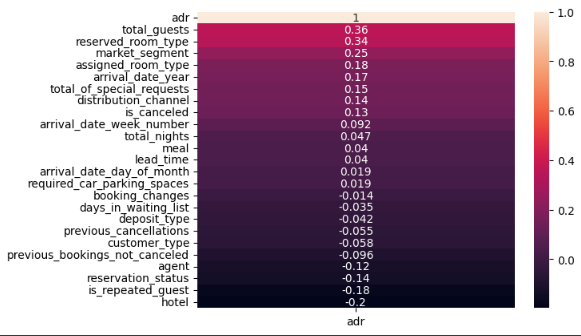
## Do some segments show higher consistency or revenue?

Compared mean ADR and cancellation rate by market\_segment to rank channel reliability.

## What factors are most strongly associated with higher ADR?

Visualized top ADR correlations from .**corr()** using bar plots to highlight influential variables.

**Graph 11: Top Factors Correlated with Higher ADR (Average Daily Rate)**



**Insight:**

* The strongest positive correlation with ADR is the **total number of guests**, followed by **number of children**, **number of adults**, and **arrival year**.
* **Special requests** also show a modest positive correlation with ADR.

**Business Takeaway:**

* Target larger groups and families (more guests) as they tend to book rooms with higher rates.
* Packages or upsells for families or guests making special requests can increase revenue.
* Focus marketing on time periods (e.g., specific years) when higher ADRs are observed.

## Which customer types contribute most to revenue?

Created revenue as **adr \*** nights` and grouped it by customer type to rank revenue contribution.

## Do long-lead bookings or certain countries yield higher ADR?

Checked correlation between **lead\_time** and ADR, plus group ADR by country to find profitable sources.

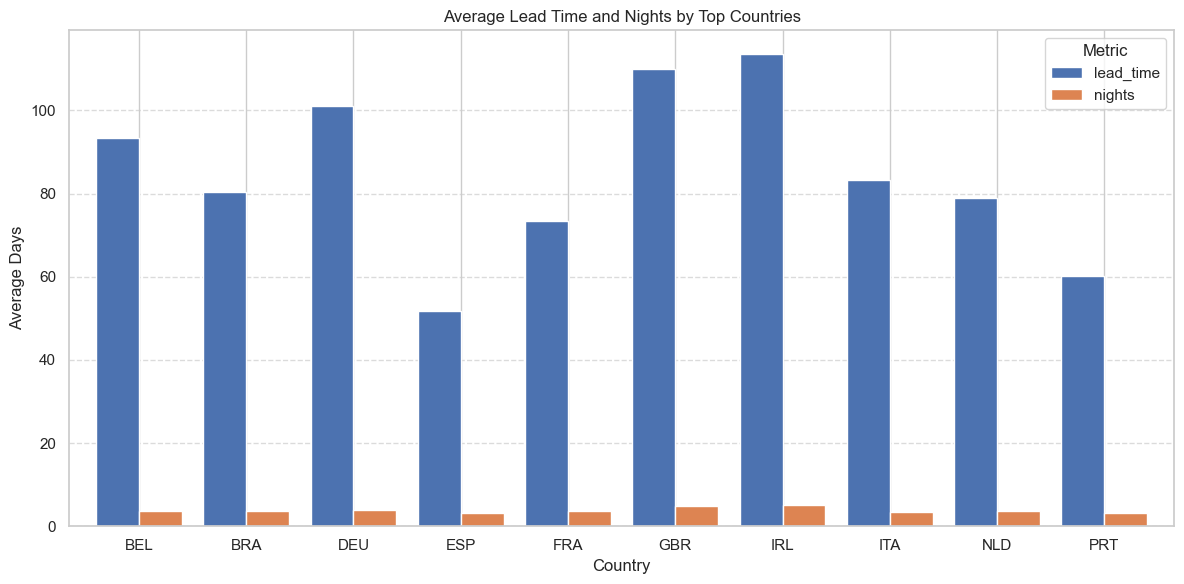
## Do high ADR guests request more or modify bookings?

Grouped ADR by guests with special requests or room changes to detect behavior of high spenders.

## Do guests from different countries behave differently?

Grouped lead time and nights by country to analyze nationality-specific patterns.

**Graph 12: Average Lead Time and Nights by Top Countries**



**Insight:**

* Guests from countries like **Ireland (IRL), Great Britain (GBR), and Germany (DEU)** tend to book **well in advance (100+ days)**.
* **Portugal (PRT)** and **Spain (ESP)** show the **shortest lead times**, typically under 60 days.
* Despite variation in booking lead times, the **average stay duration (nights)** remains **consistently short (3–5 nights)** across all countries.

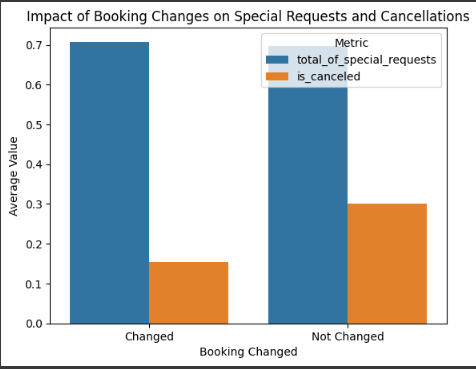
**Business Takeaway:**

* **Target early-bird offers** to countries with long lead times (e.g., IRL, GBR, DEU) to secure high-value future bookings.
* **Promote last-minute deals and flash sales** in regions like PRT and ESP to attract spontaneous travellers.
* **Standardize operational resources** for short stays, as trip length does not vary significantly by country.

## Are guests who make booking changes more likely to cancel or request extras?

Grouped **is\_canceled** and special requests by **booking\_changed** to measure risk and service demand.

**Graph 13: Impact of Booking Changes on Special Requests and Cancellations**



**Insight:**

* Bookings that were changed had a **higher average number of special requests**.
* **Cancellation rate** is significantly **lower** for bookings that were changed compared to those that were not.

**Business Takeaway:**

* Allowing and facilitating **flexible changes** in bookings can **reduce cancellations**.
* Customers who modify bookings may be more engaged and willing to spend (as shown by high special requests).
* Implement user-friendly change options and possibly incentivize changes over cancellations.

**Conclusion: -**

This exploratory analysis of the Hotel’s booking dataset has revealed several important patterns and actionable insights that can significantly inform operational strategy and revenue optimization.

**Key Findings:**

1. **Revenue Drivers:**
   * **Average Daily Rate (ADR)** is positively influenced by **group size**, **presence of children**, and **special requests**.
   * High ADR is more common in bookings from **Online Travel Agencies (OTA)** and **Direct channels**, particularly among **Transient** customer types.
2. **Booking Behavior & Lead Time:**
   * Bookings exhibit a **bimodal lead time distribution**, with clusters of **last-minute** and **long-term planners**.
   * Countries like **Ireland, UK, and Germany** tend to book well in advance, whereas **Portugal and Spain** often book closer to check-in.
3. **Customer Segmentation:**
   * **Transient guests** dominate the customer base, suggesting a need for personalized, short-stay promotions.
   * **Group** and **Contract bookings** are fewer but may require distinct pricing and lead time strategies.
4. **Operational Patterns:**
   * **Special requests** correlate with lower cancellations, indicating that such guests are more committed and potentially more profitable.
   * **Room assignment mismatches** and upgrades, while infrequent, should be monitored for operational efficiency and guest satisfaction.
5. **Channel Performance:**
   * **OTAs bring volume**, but **Direct bookings** offer strong revenue potential with lower acquisition costs.
   * **Complementary bookings** offer minimal revenue and should be evaluated for strategic justification.
6. **Cancellation Risk Factors:**
   * Longer lead times show a higher risk of cancellation.
   * Guests who make **booking changes** are **less likely to cancel**, highlighting the value of **flexibility** in booking management.

**Strategic Recommendations:**

* Develop **targeted pricing and upselling strategies** for high-ADR guest profiles (e.g., families, early planners).
* **Optimize OTA visibility** while **promoting direct booking incentives** to improve profit margins.
* Leverage **booking patterns by country** to tailor promotions (e.g., early-bird offers in Germany, flash sales in Portugal).
* Use **special requests and booking modifications** as signals for high-value, engaged customers.
* Enhance **inventory and staffing forecasts** around peak months and group arrival patterns to maximize occupancy and service quality.

**Final Thoughts**

This case study underscores the value of **data-driven decision-making** in hospitality management. By aligning marketing, operations, and pricing strategies with observable guest behaviours, hotels can not only improve guest satisfaction but also maximize revenue and resource efficiency.