**DATA 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 LOGIC

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

**EXPLORATIVE DATA ANALYSIS**

1. **Booking Patterns:**

* **What is the trend in booking patterns over time, and are there specific seasons or months with increased booking activity?**

To Tackle The Booking pattern:

1. I extracted the Month & Year
2. I grouped it by Month-Year to count the number of booking made in each month named it Booking Trend
3. Before Plotting I converted the month-year column to string so that seaborn can interpret it correctly

I did Plot a line Plot Titled **‘Booking Trend Overtime’** with **‘Month-Year’** on X-Axis and **‘Number of Booking’** on Y-Axis

Chart Present in: Folder(**SHG-Booking-EDA\Visuals/Fig1.**)

What the chart shows:

* Bookings were Low & Flat before Late 2014 then started rising
* Two Major Spikes Occurred:

Around December 2015 & December 2016

* There is a Noticeable Drop after Early 2017, Indicating a potential issue or Seasonality shift

**INFERENCE**

From the chart, We Observed a sharp Increase in bookings from Late 2014 with consistent peaks around the end of each year especially in December. This suggests strong Seasonal Demand around Year-End, possibly Due to Holidays or Corporate Events. However, There was a sharp decline after early 2017, which may warrant further investigation into Internal or External Factors affecting Booking Performance.

* **Specific Seasons/Month with Increased booking Activity**

To spot the top booking Month, I plot a Bar Chart named **‘Total Bookings by Month’** with **‘Booking Month Name’** on X-Axis and **‘Number of Booking’** on Y-Axis

Chart Present in: Folder(**SHG-Booking-EDA\Visuals/Fig2.**)

What The Chart Shows:

* January and February had the highest Volume likely from New Year Travel or Corporate Planning
* October to December also shows strong performance indicating Holiday or End of Year Trends
* Mid Year Months like July & March also perform well which could align with school breaks or summer travel

**INFERENCE**

* Strong Seasonality

Bookings Peak Consistently in January, February, October and December.

* December & January spikes may be tied to Holiday Seasons or Promotions.
* There is a clear Growth Trend between 2014 & 2016 followed by a decline in 2017
* July & March also show above average activity potentially linked to school breaks or travel seasons.

**Recommendation to Stakeholders**

* Capitalize on High-Performing Months (Jan, Feb, Oct & Dec) by launching targeted Discount Campaign, early booking offers or Corporate Packages.
* Investigate What Happened Post-2017 e.g Marketing Changes, Booking Platform Issues, Competitor shifts or Economic Factors
* Use this Insight to adjust Staffing, Marketing budget and Inventory to match Demand Cycles more efficiently
* **How does lead time vary across different booking channels, and is there a correlation between lead time and customer type?**

To figure out how lead time vary across different booking channels.

Plotted a Boxplot Titled ‘Lead Time by Booking Channel’ with x-axis as Distribution channel and y-axis as Lead Time.

Chart Present in: Folder(**SHG-Booking-EDA\Visuals/Fig3.**)

|  |  |
| --- | --- |
| Channel | Insight |
| 1. Direct | Most Bookings are made with low to moderate lead time. A few rare cases go beyond 400+ days, but the majority book last-minute. |
| 1. Corporate | It has the shortest lead time overall.  Most booking are within 0-20 days. |
| 1. Online Travel Agent | It has the widest spread a lot of bookings with higher lead time.Many Customer book more than 100 days ahead |
| 1. Offline Travel Agent | It also has a median lead time with ‘booking spread out over a large range-shows a more planned, traditional booking behavior |

* Undefined: Minimal & Scattered, likely noise or data quality issue.

**Recommendation to Stakeholders**

Customers who book through corporate & Direct channels tends to make last minute reservations, while online & offline travel agents attracts more advance planners

This suggests we can:

* Launch Early Bird promotions via travel agents especially for high-season months
* Target Last Minute Upgrade offers or dynamic pricing for direct/corporate customers.

To figure out how lead time vary across Customer type.

Plotted a box-plot of ‘Lead Time by Customer Type’ with x-axis as Customer Type and y-axis as Lead Time

Chart Present in: Folder(**SHG-Booking-EDA\Visuals/Fig4.**)

Observation:

|  |  |
| --- | --- |
| Customer type | insight |
| 1. TransieNt | Median Lead Time is low. Most booking are within 0-50 days, but some book far in advance. |
| 1. Contract | It has consistently high lead time. These are long term pre-arranged bookings likely through agreements. |
| 1. Transient-party | Similar Pattern to Transient, but slightly more spread. Represents small group booking by individuals. |
| 1. Group | It has the shortest & Highest Lead Time. Most group bookings are very close to arrival, suggesting quick arrangements for events or meetings. |

**Stake-Holder Insight**

* ‘Contract Customers’ book well in advance likely due to long term deals with companies or institutions.
* ‘Group Bookings’ are often made close to the arrival date, suggesting they are for events or last minute conferences.
* ‘Transient & Transient-Party’ Bookings show more flexibility in lead time.

**Correlation between lead time and customer type**

1. **Contract Customers Book Earliest**

* Offline Travel agent + Contract (160.8 days) and Online Travel Agent + Contract (130.4 days) tops the list
* Contract bookings involve corporate deals or recurring business, so they are booked in advance.

INFERENCE: Maintain Long term Partnership with Contract Clients & offer them annual bulk-booking options.

1. **Transient party bookings have long lead-time when via agents**

* Through offline (159.8 days ) or online Travel agents (139.5 days), transient-party customers book early-perhaps for planned group vacations or events

STRATEGY : Travel Agents should be supported with group packages well ahead of high seasons.

1. **Corporate Channel = Last Minute Bookings**

* Corporate + Group = 6.9 days (Shortest)
* Corporate + Transient = 25.8 days

Observation: Corporate Clients tends to book last minute for urgent business travel

Action: Offer Flexible rates, last minute availability and express check-in for corporate clients.

1. **Group Customers Book Late**

* Regardless of channel, groups generally have short lead time
* Direct + Group = 69.6 days
* Offline Agent + Group = 68.1 days
* Online Agent + Group = 42.3 days
* Corporate + Group = 6.9 days

**Observation:** Group Booking are often event-driven and arranged quickly, possibly due to logistics or short planning windows

**Action**: Prepare Scalable Group Offers that can be activated quickly e.g Event-hosting Bundles

1. **Direct Channel is used more by last minute bookers**

* Direct +Transient = 49.1 days
* Direct + Contract = 64.3 days
* Direct + Group = 69.6 days

**Observation**: These are all significantly lower than booking through OTA or Offline Agents

**Actions**: Use your website to push last minute promotions, ,mobile-first booking experiences and personalized pricing.

1. **Customer Behavior Analysis:**

* Which distribution channels contribute the most to bookings, and how does the average daily rate (ADR) differ across these channels?

**a(i) Which distribution channels contribute the most to bookings**

First Step: Preprocess my data: I filtered out where booking was not completed to focus on actual (non-canceled booking)

Chart Present in: Folder (**SHG-Booking-EDA\Visuals/Fig5.**)

**Insights**

The Analysis shows that the online Travel Agent distribution channel contributes the most to completed bookings, followed by offline travel agent. This Insight provide a clear picture of where Splendor Hotel Group Customer Engagement is Strongest, helping inform channel optimization strategies.

Corporate Contributes the least in both Volume and Revenue

**a(ii) How does the average daily rate (ADR) differ across these channels?**

Steps:

* I calculated the average daily rate for each channel using all non-cancelled booking
* I filtered out undefined distribution channel (removed rows where the distribution channel was not defined, so your results only reflects meaningful categories)
* I dropped Duplicated Booking ID.

Then I plotted a bar chart to visually compare ADR across Different Distribution Channel

Chart Present in: Folder (**SHG-Booking-EDA\Visuals/Fig6.**)

**Insight**

* Online Travel Agent : Highest Number of Bookings & Highest ADR
* Direct: Close ADR to OTA, despite fewer bookings
* Corporate: Lowest ADR, Confirming it brings Smaller Revenue per booking.

**Business Recommendation**

Most of Splendor Hotel Group’s Revenue Volume comes from online Travel Agents (OTA), but Direct Bookings show cpmpetitive ADR. This suggests an opportunity to strengthen direct sales channels (website/App Bookings) potentially reducing commission fees from OTA platforms.

Corporate Bookings Contribute Minimally to revenue.

b) Can we Identify any patterns in the distribution of guest based on their country of origin and how does this Impact revenue?

b(i) Identifying patterns in the distribution of guests based on country

Steps:

1. Preprocess the data

I checked for missing country data first

1. Checked basic stats of the guests e.g mean, standard deviation, 25%, 50% e.t.c

Analysis

Grouped guests by country

Then plotted a barplot for Top 10 country by guest with Country on x axis and Guests on y axis

Chart Present in: Folder (**SHG-Booking-EDA\Visuals/Fig7.**)

**Observation**

* Portugal dominates in guest volume, contributing almost 4x more than the next country
* The top 5 countries are all in Europe, suggesting a strong regional customer base.

b(ii) How it does impact revenue ?

**Steps**

1. Grouped revenue by country
2. Merge guests & revenue data on country
3. Calculate revenue per guest

Then plotted a bar plot of Top 10 countries by total revenue with country on x and revenue on y

Chart Present in: Folder (**SHG-Booking-EDA\Visuals/Fig8.**)

**Observation**

* While Portugal leads in total revenue, it’s revenue per guest is the lowest among the top contributors.
* Ireland & United Kingdom generate significantly more revenue per guest compared to Portugal
* High guest count doesn’t automatically mean high value per customer

Then to identify high-value(premium) guest countries then I plotted a bar chart of Top Countries by revenue per guest.

With country on x-axis and revenue per guest on y-axis

Chart Present in: Folder (**SHG-Booking-EDA\Visuals/Fig9.**)

**Observation**

* These countries have very few bookings but exceptionally high revenue per guest
* These are likely VIP, Long-stay or luxury guests

**Recommendation**

1. Portugal – Despite having the highest guest count, Portugal has the lowest revenue per guest.

**Recommendation**

* Introduce upselling strategies for Portugese guests (e.g spa, dinning packages, premium room upgrades)
* Segment & target frequent domestic travelers with loyalty rewards or tiered pricing

1. UK, Ireland & Belgium – Prioritize premium services

* These countries have fewer guests than Portugal but significantly higher revenue per guest.

**Recommendation**

* Market Premium offerings (executive suites, experiences, all-inclusive plant) to UK & Ireland Visitors
* Launch geo-targeted ads showing luxury experiences or business-friendly features

1. France, Spain, Germany – Maintain Consistency

* These countries show a balanced performance in guest volume & revenue.

**Recommendation**

* Consider cultural-specific amenities or services to drive services to drive satisfaction and repeat bookings
* Maintain current engagement levels but test-cross selling and partnership opportunities with airlines or OTAs

1. High Value Countries – Develop exclusive programs

* Countries like palau, Togo etc contribute very high revenue per guest despite low visit frequency

**Recommendation**

* Identify booking channels or customer profiles for these guest
* Create an “Elite Guest Program” or Luxury Concierge to attract & retain these high value customers.

1. Overall Guest Distributionn Strategy

* Guest & revenue concentration is heavily Europe based

**Recommendation**

* Diversify market reach through campaigns in under-represented high-income regions (e.g North Africa, Middle East)
* Monitor Travel trends & Visa Policy changes that could open up emerging markets.

1. **Cancellation Analysis:**

* What factors are most strongly correlated with cancellations, and can we predict potential cancellations based on certain variables?

Step 1: I used correlation heat map to identify numeric features associated with cancellations. I plot a correlation heat map of:

* **Lead Time**
* **Guests**
* **Nights**
* **Revenue**
* **Avg Daily Rate**
* **Revenue Loss**

Chart Present in : Folder (**SHG-Booking-EDA\Visuals/Fig10.**)

From the heat map I discovered that:

* **Lead Time vs. Cancellation = 0.29 → Moderate Positive Correlation**  
  → Longer lead time (long gap between booking and arrival) is associated with a higher chance of cancellation.
* **Other features (Nights, Guests, Avg Daily Rate, Revenue, Revenue Loss)** show very low correlation with Canceled (0/1) (between -0.06 and 0.07).

**Summary Insight:**  
Lead time stands out as the numeric feature most correlated with cancellation in your data.

Step 2: To check if categorical features like deposit type, customer type & Distribution Channel

Distribution channel Chart Present in : Folder (**SHG-Booking-EDA\Visuals/Fig11.**)

Deposit Type Chart Present in : Folder (**SHG-Booking-EDA\Visuals/Fig12.**)

Customer Type Chart Present in : Folder (**SHG-Booking-EDA\Visuals/Fig13.**)

**Observations from Distribution channel:**

* Bookings made through Online Travel Agents have the highest cancellation rate.
* Direct and Corporate channels show more stable, lower cancellation rates

**Observations from Deposit Type:**

* Customers without financial commitment tend to cancel more.
* Refundable deposits help balance flexibility and commitment.

**Observations from Customer Type:**

* One-off individual bookings (Transient) are the most unpredictable.
* Groups provide more reliable, lower cancellation rates.

**Recommendations**

1. **Prioritize Direct and Corporate Booking Channels**  
   → Invest in marketing and loyalty strategies encouraging customers to book directly or through corporate arrangements.
2. **Refine Deposit Policies**  
   → Review how non-refundable bookings are handled to avoid system misreporting.  
   → Promote refundable deposit options over no-deposit offers to reduce cancellations while maintaining flexibility.
3. **Focus on Customer Segmentation**  
   → Develop targeted retention strategies for Transient customers, such as:
   * Early check-in offers
   * Flexible date change options
   * Reminder notifications
4. **Promote Group and Contract Bookings**  
   → Highlight group packages and corporate deals in marketing campaigns as these have lower cancellation rates.

* How does the revenue loss from cancellations compare across different customer segments and distribution channels?

Step 1: I focused only on cancelled booking

So i filter my DataFrame where Canceled == 1

## Step 2: Group and Calculate Average Revenue Loss by Customer Type

**Why:**  
This tell which customer segment’s cancellations hit revenue the hardest on average.

Plotted a bar plot with x='Customer Type', y='Revenue Loss' titled: ‘Average Revenue Loss from Cancellations by Customer Type’

Chart Present in : Folder (**SHG-Booking-EDA\Visuals/Fig14.**)

**Key Observations:**

* **Transient Customers** cause the highest average revenue loss per canceled booking.
* **Group Cancellations** follow closely, though group bookings tend to have a lower cancellation rate.
* **Contract and Transient-Party** customers result in comparatively lower revenue loss when canceled.

## Step 3: Group and Calculate Average Revenue Loss by Distribution Channel

**Why:**  
I want to check which channel’s cancellations affect revenue the most — not just how often they cancel, but how expensive those cancellations are

Plotted a bar plot with x='Distribution Channel', y='Revenue Loss' titled: 'Average Revenue Loss from Cancellations by Distribution Channel'

Chart Present in : Folder (**SHG-Booking-EDA\Visuals/Fig15.**)

**Key Observations:**

* **Direct Bookings**, despite having a lower cancellation rate, generate the largest financial loss when canceled. This suggests that direct bookings are usually higher-value transactions.
* **Online Travel Agents** contribute significantly to revenue loss, aligning with both a high cancellation rate and moderately high revenue loss.
* **Offline Travel Agent and Corporate channels** produce much smaller average revenue losses per cancellation.

## Combined Strategic Recommendations:

1. **Prioritize Safeguarding High-Value Direct Bookings:**
   * Implement stricter cancellation policies or partial deposit requirements for direct bookings.
   * Offer rescheduling incentives instead of cancellations.
2. **Target Transient Customers with Proactive Retention Strategies:**
   * Send reminder emails or SMS follow-ups.
   * Offer flexible rescheduling or booking change options for transient customers with long lead times.
3. **Review Pricing and Policy Alignment on Online Travel Agent Platforms:**
   * Consider adjusting non-refundable pricing or requiring deposits for OTA bookings.
   * Collaborate with OTA partners on cancellation control strategies.
4. **Focus on Strengthening Group and Corporate Customer Retention:**
   * Although revenue loss per cancellation is lower here, these are stable customer types. Strengthening long-term contracts or loyalty programs would protect consistent revenue streams.