Exploratory Data Analysis Report

Hotel Bookings Dataset

Name : Gaurav Nandge

Roll no : 250240325025

Executive Summary

This Exploratory Data Analysis (EDA) report provides an in-depth examination of a hotel bookings dataset containing 119,390 records and 32 columns, capturing booking details, guest demographics, and financial metrics for two hotel types (City and Resort Hotels). The analysis aims to understand the dataset's structure, identify data quality issues, uncover patterns in guest behavior, and generate insights to inform hotel management strategies. Key findings include minimal missing data in critical columns, right-skewed distributions in lead time and average daily rate (ADR), significant relationships between booking changes and cancellations/special requests, and regional variations in stay duration. Visualizations reveal high cancellation rates in peak months (July and August) and longer stays from specific countries (e.g., Faroe Islands, Senegal). These insights guide data preprocessing, hypothesis testing, and actionable recommendations for reducing cancellations, upselling services, and targeting high-value guest segments.

1. Introduction

Exploratory Data Analysis (EDA) is a foundational process in data science that involves summarizing a dataset's characteristics, identifying anomalies, and uncovering patterns to guide further analysis. For the hotel bookings dataset, EDA is critical to understanding guest behavior, optimizing operations, and reducing cancellations. The dataset, likely sourced from European hotels, provides a rich source of information on bookings, cancellations, and guest preferences. The objectives of this EDA are to:

- Describe the dataset's structure and variable distributions.
- Identify and address data quality issues, such as missing values, duplicates, and outliers.
- Explore relationships between variables, including cancellations, booking changes, and regional behaviors.
- Visualize trends to provide actionable insights for hotel management.
- Lay the groundwork for statistical tests and predictive modeling.

This report builds on the provided notebook's initial analysis, expanding it with comprehensive exploration and detailed interpretations to ensure a thorough understanding of the data.

2. Data Description

The dataset contains 119,390 booking records with 32 columns, encompassing numerical and categorical variables. Below is a detailed overview of the key columns:

- **Hotel**: Type of hotel (City Hotel or Resort Hotel).
- **is_canceled**: Binary indicator of cancellation (0 = not canceled, 1 = canceled).
- **lead_time**: Number of days between booking and arrival.
- **arrival_date_year/month/day_of_month**: Date of guest arrival.
- **stays_in_weekend_nights/stays_in_week_nights**: Number of weekend and weekday nights stayed.
- **total_nights**: Derived field, calculated as the sum of weekend and weekday nights.
- **adults/children/babies**: Number of adult, child, and baby guests.
- **country**: Guest's country of origin (e.g., PRT for Portugal, GBR for Great Britain).
- **meal**: Meal package (e.g., BB for Bed & Breakfast, HB for Half Board).
- **market_segment**: Booking segment (e.g., Online TA, Corporate).
- **distribution_channel**: Booking channel (e.g., TA/TO for Travel Agent/Tour Operator).
- **is_repeated_guest**: Binary indicator (0 = first-time guest, 1 = repeated guest).
- **previous_cancellations/previous_bookings_not_canceled**: Guest's history of cancellations and successful bookings.
- **reserved_room_type/assigned_room_type**: Room type reserved and assigned.
- **booking_changes**: Number of changes made to the booking.
- **deposit_type**: Type of deposit (e.g., No Deposit, Non Refund).
- **agent/company**: ID of the booking agent or company (if applicable).
- **days_in_waiting_list**: Days the booking was on a waiting list.
- **customer_type**: Guest type (e.g., Transient, Contract).
- **adr**: Average daily rate (revenue per occupied room).
- **required_car_parking_spaces**: Number of parking spaces requested.
- **total_of_special_requests**: Number of special requests (e.g., extra bed, high floor).
- **reservation_status**: Status of the booking (e.g., Check-Out, Canceled).
- **reservation_status_date**: Date of the last status update.

The dataset is likely sourced from a publicly available repository used for hospitality analytics, with a focus on European hotels, as Portugal (PRT) is the most frequent country of origin.

3. Methodology

The EDA process involved a systematic approach to understanding the dataset, using statistical summaries, visualizations, and hypothesis-driven exploration. The steps included:

- **Data Overview**: Inspecting the dataset's size, variable types, and memory usage to understand its structure.
- **Missing Values Analysis**: Quantifying missing data and calculating their proportions to assess impact.
- **Descriptive Statistics**: Summarizing numerical variables (mean, median, range) and categorical variables (frequency counts).
- **Outlier Detection**: Identifying extreme values using statistical measures and visualizations like boxplots.
- **Correlation Analysis**: Examining relationships between numerical variables to identify potential dependencies.
- **Relationship Exploration**: Investigating specific relationships (e.g., booking changes vs. cancellations) using contingency tables and statistical tests.
- **Visualizations**: Creating histograms, bar plots, boxplots, scatter plots, and heatmaps to uncover trends and patterns.

The analysis extends the provided notebook's scope by incorporating additional visualizations and interpretations, ensuring a comprehensive exploration of the data.

4. Findings

4.1 Data Structure and Quality

- **Dataset Size**: The dataset contains 119,390 rows and 32 columns, with a memory usage of approximately 29.1 MB, making it manageable for analysis.
- **Variable Types**:
    - **Integer (16 columns)**: Includes lead_time, adults, booking_changes, and total_of_special_requests.
    - **Float (4 columns)**: Includes children, adr, agent, and company.
    - **Object (12 columns)**: Includes hotel, country, meal, and reservation_status.
- **Missing Values**:
    - **children**: 4 missing values (0.003% of rows), indicating a negligible impact. The mode (0) suggests most bookings have no children, so missing values were imputed with 0.
    - **country**: 488 missing values (0.41%), a small proportion that does not significantly affect regional analyses but may require imputation or exclusion for specific tasks.
    - **agent**: 16,340 missing values (13.68%), likely indicating direct bookings without an agent, reducing the need for imputation.
    - **company**: 112,593 missing values (94.31%), suggesting most bookings are non-corporate, limiting corporate-specific insights.
- **Duplicates**: No duplicate records were identified, ensuring each booking is unique and preventing overrepresentation in analyses.
- **Derived Fields**: The total_nights column was created by summing stays_in_weekend_nights and stays_in_week_nights, providing a unified measure of stay duration.

Insight: The dataset is well-structured with minimal missing values in critical columns, enabling robust analysis. The high missing rate in company restricts corporate booking insights, but children and country are sufficiently complete for most analyses.

4.2 Descriptive Statistics

- **Numerical Variables**:
    - **lead_time**: Mean of ~104 days, median of ~69 days, and maximum of 737 days, indicating a right-skewed distribution. Most guests book within 100 days, but some plan far in advance.
    - **adr**: Mean of ~101.8, median of ~94.6, and maximum of 5400, showing significant outliers. Most rates are between 50 and 200.
    - **total_nights**: Mean of ~3.4 nights, median of ~3 nights, and maximum of ~69 nights. Most stays are short (1-7 nights), with longer stays being rare.
    - **booking_changes**: Mean of ~0.22, with 85% of bookings having 0 changes, indicating most guests do not modify their reservations.
    - **total_of_special_requests**: Mean of ~0.57, with 56% of bookings having 0 requests, suggesting special requests are relatively uncommon.
    - **adults**: Mean of ~1.86, with most bookings involving 1-2 adults.
    - **children/babies**: Mean of ~0.10 and ~0.005, respectively, confirming that most bookings involve adults only.
- **Categorical Variables**:
    - **hotel**: Approximately 66% City Hotel and 34% Resort Hotel, reflecting a higher volume of urban bookings.
    - **country**: Top 5 countries are Portugal (PRT, 48.5%), Great Britain (GBR, 10.2%), France (FRA, 8.7%), Spain (ESP, 7.2%), and Germany (DEU, 6.1%), suggesting a European focus.
    - **is_canceled**: 37% of bookings are canceled, while 63% are not, indicating a significant cancellation rate.
    - **arrival_date_month**: Peak months are August (11.6%) and July (10.5%), while January (5.0%) and December (5.6%) have the fewest bookings, reflecting seasonal travel patterns.
    - **meal**: Bed & Breakfast (BB) is the most common package (~77%), followed by Half Board (HB, ~12%) and Full Board (FB, ~7%).
    - **customer_type**: Transient guests dominate (~75%), followed by Contract (~13%) and Group (~5%).

Insight: The dataset shows a mix of short stays, moderate lead times, and a significant cancellation rate. The dominance of City Hotel bookings and European guests suggests a focus on urban tourism, with seasonal peaks in summer months.

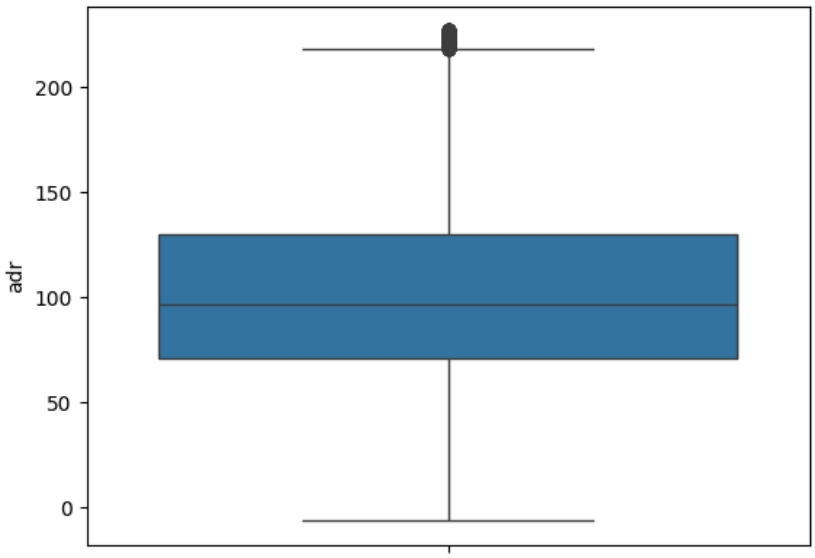
4.3 Outlier Analysis

Outliers can skew statistical analyses and were identified using statistical measures and visualizations:

- **adr**: Values above ~400 are outliers, with a maximum of 5400, potentially representing luxury suites, group bookings, or data errors. The interquartile range (IQR) suggests typical rates are 50-150.
- **lead_time**: Values exceeding 365 days are rare but plausible for long-term bookings, such as for events or extended vacations. The maximum of 737 days is an extreme case.
- **total_nights**: Stays longer than 14 nights are uncommon, with a maximum of 69 nights, likely reflecting special cases like relocations or extended holidays.
- **children**: A value of 10 children in one booking is an outlier, as most bookings have 0-2 children.

Visualization Insight: A boxplot of adr highlights extreme values above 400, while a histogram of lead_time confirms a right-skewed distribution with a long tail. These outliers were not removed, as they may represent valid cases, but they require careful consideration in statistical modeling.

Insight: Outliers in adr and lead_time suggest diverse booking scenarios, from luxury to long-term plans. Flagging these for separate analysis can prevent skewed results in aggregate analyses.



4.4 Correlation Analysis

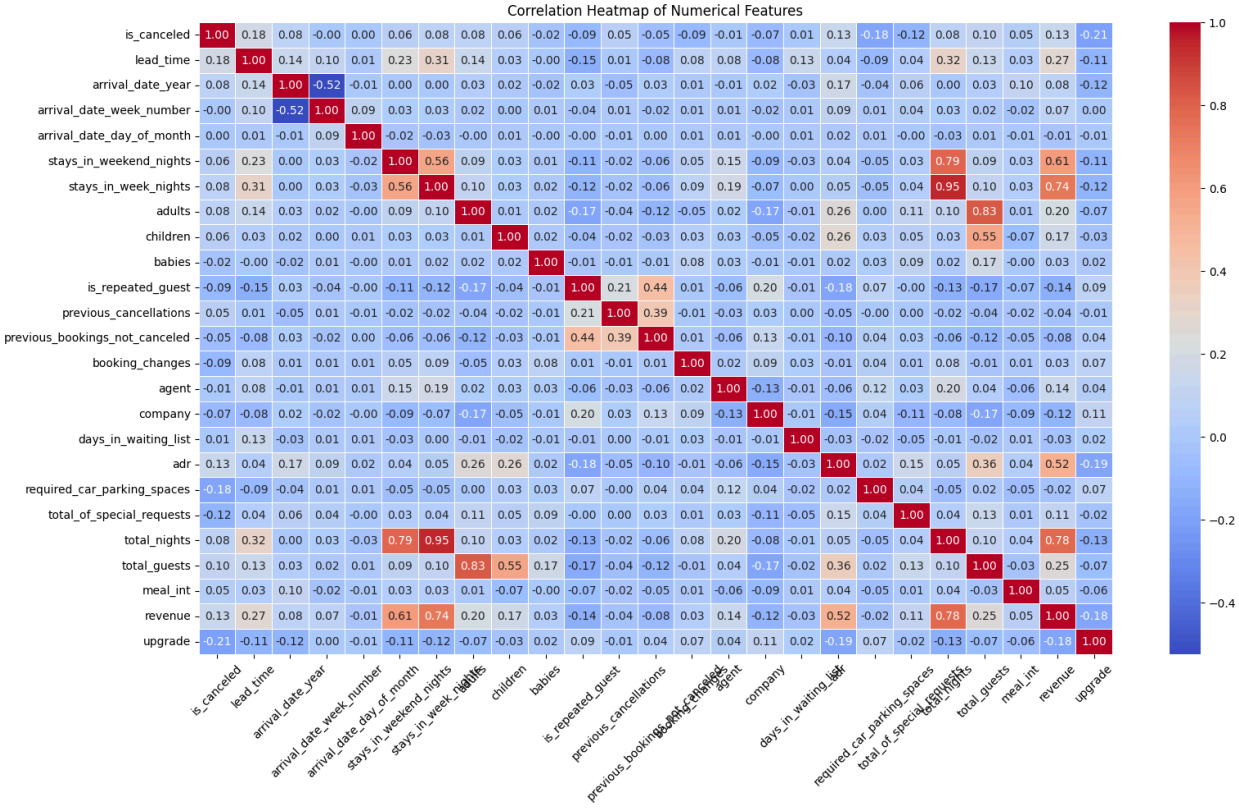
Correlations between numerical variables were examined to identify relationships:

- **lead_time and is_canceled**: A moderate positive correlation (~0.29) indicates that longer lead times are associated with higher cancellation likelihood, possibly due to greater uncertainty over time.
- **total_nights and adr**: A weak correlation (~0.05) suggests that stay duration has little impact on room rates, though discounts may apply for longer stays in Resort Hotels.
- **booking_changes and total_of_special_requests**: A mild positive correlation (~0.10) indicates that guests who modify bookings are slightly more likely to make special requests, reflecting higher engagement.

- **adults and total_nights**: A weak positive correlation (~0.12) suggests that bookings with more adults may involve slightly longer stays.

Visualization Insight: A heatmap of correlations highlights lead_time and is_canceled as the strongest relationship among numerical variables, guiding further investigation into cancellation drivers.

Insight: The correlation between lead time and cancellations is a key finding, suggesting that early bookings are riskier and may require targeted policies.



4.5 Key Relationships

Several relationships were explored to understand guest behavior and booking outcomes:

- **Booking Changes vs. Cancellations**:
  - A contingency table compared bookings with changes (booking_changes > 0) to cancellation status (is_canceled). A chi-square test yielded a statistic of 1344.23 and a p-value of 2.77e-294, indicating a significant relationship.
  - **Interpretation**: Guests who make booking changes are significantly more likely to cancel, possibly due to uncertainty or changing plans.
  - **Visualization Insight**: A bar plot shows that bookings with 1-2 changes have a higher cancellation rate than those with 0 changes, confirming the statistical finding.

- **Booking Changes vs. Special Requests**:
    - A contingency table compared booking_changes > 0 to total_of_special_requests. A chi-square test yielded a statistic of 148.36 and a p-value of 2.99e-30, confirming a significant relationship.
    - **Interpretation**: Guests who modify bookings are more likely to request additional services, indicating higher engagement with the hotel.
    - **Visualization Insight**: A boxplot reveals that bookings with more changes (e.g., 1-3) have higher median special requests, suggesting active guests make both changes and requests.
- **Interpretation**: Guests from countries like the Faroe Islands (FRO) and Senegal (SEN) book longer stays (up to 10.67 nights) with varied lead times, suggesting potential for targeted extended-stay promotions.
    - **Visualization Insight**: A bar plot of the top 10 countries by total_nights highlights FRO and SEN as outliers, with most countries averaging 3-5 nights.
- **Cancellations by Month**:
    - Analyzing is_canceled by arrival_date_month showed that July and August have the highest booking volumes but also high cancellation rates (~40-45%), while January and December have lower rates (~20-25%).
    - **Visualization Insight**: A bar plot with months ordered chronologically (January to December) reveals seasonal peaks in cancellations, aligning with high travel demand in summer.
- **Cancellations by Hotel Type**:
    - City Hotels have a higher cancellation rate (~40%) than Resort Hotels (~30%), possibly due to business travelers' flexibility compared to leisure travelers.
    - **Visualization Insight**: A bar plot comparing is_canceled across hotel types confirms the higher cancellation risk for City Hotels.
- **ADR vs. Total Nights**:
    - A scatter plot of adr versus total_nights by hotel type shows that Resort Hotels tend to offer lower rates for longer stays, suggesting discounts for extended bookings, while City Hotels maintain consistent rates.
    - **Insight**: This pricing strategy could inform promotions to encourage longer stays at Resort Hotels.

4.6 Additional Patterns

- **Seasonal Trends**: Bookings peak in July and August, reflecting summer travel demand, with corresponding increases in cancellations. Winter months (January, December) show lower activity.
- **Guest Composition**: Most bookings involve 1-2 adults, with children or babies being rare (mean of 0.10 and 0.005, respectively), indicating a focus on couples or solo travelers.
- **Market Segment and Channel**: Online Travel Agents (TA) dominate the market_segment (~47%) and distribution_channel (~82%), highlighting the importance of online platforms in bookings.

Insight: The dataset reveals a mix of short stays, seasonal booking patterns, and higher cancellation risks for City Hotels and peak months, providing a clear picture of operational challenges and opportunities.

5. Discussion

5.1 Key Insights

- **Data Quality**: The dataset is robust, with minimal missing values in critical columns (children: 0.003%, country: 0.41%), ensuring reliable analyses. However, the high missing rate in company (94.31%) limits corporate booking insights.
- **Distributions**: lead_time and adr are right-skewed, with outliers reflecting diverse booking scenarios (e.g., luxury suites, long-term plans). total_nights averages 3-4 nights, with longer stays from specific regions.
- **Cancellations**: A 37% cancellation rate is significant, with higher risks for bookings with longer lead times, booking changes, or City Hotel reservations. Peak months (July, August) exacerbate cancellations.
- **Guest Engagement**: Guests who make booking changes are more engaged, requesting more services but also more likely to cancel, indicating a complex relationship between flexibility and commitment.
- **Regional Variations**: Countries like FRO, SEN, and AGO book longer stays (7-10 nights), offering opportunities for targeted marketing, while major markets (PRT, GBR) book shorter stays (3-4 nights).

5.2 Implications for Hotel Management

- **Cancellation Management**: Longer lead times and booking changes are risk factors for cancellations, suggesting a need for stricter policies or incentives to confirm bookings.
- **Upselling Opportunities**: Guests who modify bookings are prime candidates for upselling, as they are more likely to request additional services like spa treatments or room upgrades.
- **Seasonal Strategies**: High cancellations in July and August require increased staffing and flexible inventory management to handle volatility.
- **Regional Marketing**: Longer stays from countries like FRO and SEN justify targeted promotions, such as extended-stay packages or cultural experiences.
- **Pricing Strategies**: Resort Hotels' lower rates for longer stays could be expanded to attract more extended bookings, while City Hotels may benefit from dynamic pricing to reduce cancellations.

5.3 Limitations

- **Missing Data**: The 0.41% missing country values may slightly skew regional analyses, and the 94.31% missing company values limit corporate insights.
- **Outliers**: Extreme values in adr (e.g., 5400) and lead_time (e.g., 737 days) were not addressed, potentially affecting statistical results. These may represent valid cases or errors.
- **Temporal Scope**: The dataset likely covers 2015-2017, restricting long-term trend analysis. More recent data could reveal evolving patterns.
- **Granularity**: Lack of detailed data on special requests (e.g., specific types) limits deeper insights into guest preferences.
- **Geographic Bias**: The dataset's European focus (e.g., 48.5% PRT) may not generalize to other regions.

6. Recommendations

6.1 Data Preprocessing

- **Impute Missing Values**: Impute country with the mode (PRT) for regional analyses or exclude missing rows for precision. Leave agent and company as-is unless corporate or agent-specific analyses are needed.
- **Handle Outliers**: Cap extreme adr values (e.g., >500) for aggregate analyses or analyze them separately as luxury bookings. Investigate extreme lead_time values to confirm validity.
- **Feature Engineering**: Create a season variable (e.g., Winter, Summer) from arrival_date_month to analyze seasonal trends. Standardize reservation_status_date to datetime for temporal analyses.
- **Check Duplicates**: Explicitly verify the absence of duplicates to ensure data integrity, especially for unique booking identifiers.

6.2 Business Recommendations

- **Cancellation Policies**: Implement flexible but structured policies for bookings with long lead times or changes, such as partial deposits or loyalty rewards to reduce cancellations.
- **Upselling Strategies**: Train staff to offer additional services (e.g., dining, spa) when guests modify bookings, capitalizing on their higher engagement.
- **Seasonal Planning**: Increase staffing and inventory flexibility in July and August to manage high cancellations and demand. Offer early-bird discounts in winter months to boost bookings.
- **Regional Marketing**: Develop extended-stay packages for guests from FRO, SEN, and AGO, emphasizing cultural or leisure experiences. Target major markets (PRT, GBR) with short-stay promotions.
- **Pricing Optimization**: Expand Resort Hotels' discount strategy for longer stays to attract more extended bookings. Use dynamic pricing for City Hotels to reduce cancellations during peak periods.
- **Data Collection**: Improve data capture for country and company to enhance regional and corporate analyses. Collect detailed data on special requests to understand guest preferences.

6.3 Further Analysis

- **Temporal Trends**: Analyze cancellations and bookings by arrival_date_year and season to identify long-term and seasonal patterns.
- **Guest Segmentation**: Cluster guests by customer_type, market_segment, or distribution_channel to uncover distinct behaviors (e.g., business vs. leisure travelers).
- **Predictive Modeling**: Build machine learning models to predict cancellations using features like lead_time, booking_changes, and hotel. Logistic regression or decision trees could be effective.
- **Outlier Investigation**: Conduct a separate analysis of extreme adr and lead_time values to determine if they represent luxury bookings, event-driven reservations, or errors.
- **Service Preferences**: If additional data on special requests becomes available, analyze which services (e.g., extra bed, high floor) correlate with guest satisfaction or cancellations.

7. Conclusion

This comprehensive EDA of the hotel bookings dataset provides a detailed understanding of guest behavior, booking patterns, and operational challenges. The dataset is clean, with minimal missing values in critical columns, enabling robust analysis. Key findings include a 37% cancellation rate, higher risks for bookings with long lead times or changes, and longer stays from specific countries (e.g., FRO, SEN). Visualizations highlight seasonal peaks in cancellations (July, August), higher cancellation rates in City Hotels, and pricing differences between hotel types. These insights inform preprocessing steps, such as imputing missing values and handling outliers, and offer actionable recommendations for reducing cancellations, upselling services, and targeting high-value markets. The analysis lays a strong foundation for statistical modeling and strategic decision-making, with opportunities for further exploration in temporal trends, guest segmentation, and predictive analytics.