**Analysis Report**

**Hotel Booking Data Analysis**

**Ruchika Mhetre**

**Under the Guidance of**

**Ms. Vineeta Singh**

(Hon. Course coordinator, DBDA, CDAC Mumbai)

**Mr. Nishad Kharote**

(Faculty)

**Mr. Prashant Bhosale**

(Faculty)

**Chapter 1: Introduction**

**Introduction**

In today’s competitive business environment, data-driven insights play a crucial role in shaping strategies and enhancing performance. This report analyzed key aspects of customer behavior, booking patterns, and operational factors to uncover trends that impact revenue and customer satisfaction. Understanding these patterns allows businesses to make informed decisions and stay ahead in the market.

**Objective**

The objective of this chapter is to summarize the critical findings from the analysis, highlight their business implications, and provide actionable recommendations. This helps translate data insights into practical strategies that drive growth, optimize resources, and improve customer experience.

**Chapter 2: Data Cleaning**

**2.1 - Objective**

The primary goal of the data cleaning phase was to ensure the dataset was accurate, consistent, and ready for analysis. This involved handling missing values, correcting data types, eliminating duplicates, and engineering new features to support deeper analytical insights. Clean data is essential to draw valid conclusions and build reliable models.

**2.2 - Steps Performed**

**2.2.1. Initial Inspection**

The dataset was thoroughly inspected to understand its structure, data types, and the presence of missing values or anomalies. Particular attention was paid to fields directly affecting booking behavior and revenue generation, such as lead\_time, adr (Average Daily Rate), booking\_status, and customer demographic attributes.

**2.2.2 Handling Missing Values**

A few key columns contained missing or null values. These were treated as follows:

* **children column:** Missing values were assumed to represent zero and filled accordingly.
* **country column:** Unknown country entries were labeled as 'Unknown' to retain rows for demographic analysis.
* **agent column:** These were filled with a placeholder (e.g., 0 or 'Unknown Agent') indicating no intermediary.
* **company column:** As this field had a large number of missing values and was not significant for analysis, it was dropped from the dataset.

**2.2.3. Date Fields Cleaning**

Date-related fields such as reservation\_status\_date, arrival\_date\_month, and arrival\_date\_year were transformed into standard datetime formats. This facilitated time-series analysis and helped identify seasonal patterns in bookings.

**2.2.4. Feature Engineering**

To enrich the dataset and make it more analytically useful, new features were created:

* **total\_guests**: Calculated by summing the number of adults, children, and babies.
* **total\_stay**: Computed by adding weekday and weekend stay durations.
* **booking\_success**: A binary feature indicating whether a booking was fulfilled or canceled.
* **room\_mismatch**: A new column indicating whether the reserved and assigned room types differed—used to assess operational consistency.

**2.2.5. Removing Duplicates**

Duplicate rows were identified and removed to ensure data integrity and prevent skewed insights. This step helped in maintaining the uniqueness of each booking record.

**2.2.6. Outlier Treatment**

The adr (Average Daily Rate) field showed extreme outliers, with some values exceeding realistic thresholds. To maintain the reliability of trend and correlation analyses, rows with ADR values above 500 were filtered out.

**2.3 - Outcome of Data Cleaning**

After completing the data cleaning process:

* The dataset was free from inconsistencies and missing values.
* Derived features added value to downstream analysis.
* Anomalies and outliers were handled appropriately.
* The data structure was optimized for further exploratory and statistical analysis.

**Chapter 3: Exploratory Data Analysis**

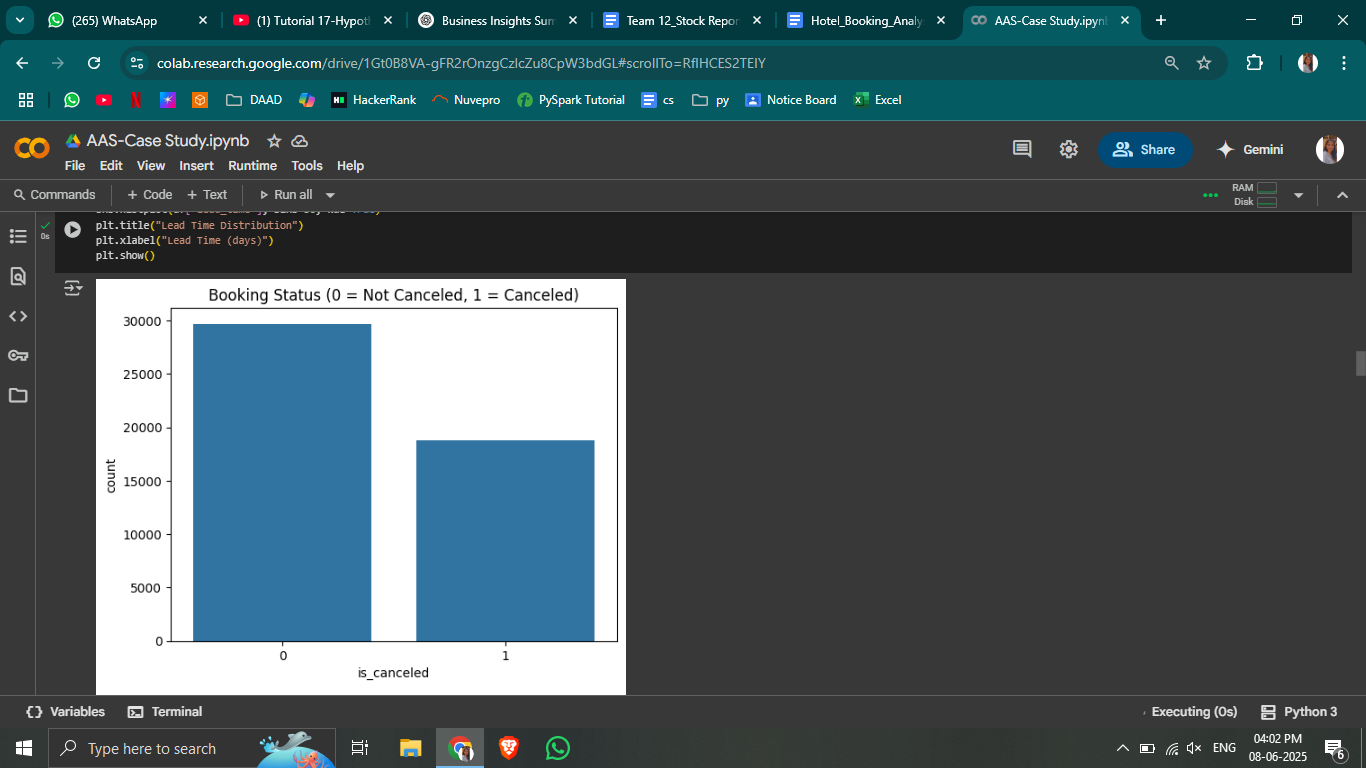
**3.1 - Objective**

The goal of Exploratory Data Analysis (EDA) is to uncover meaningful patterns, detect anomalies, test assumptions, and generate hypotheses for further analysis. EDA bridges the gap between raw data and actionable insights by using statistical summaries and visualizations to understand relationships within the dataset.

**3.2.1 - Univariate Analysis**

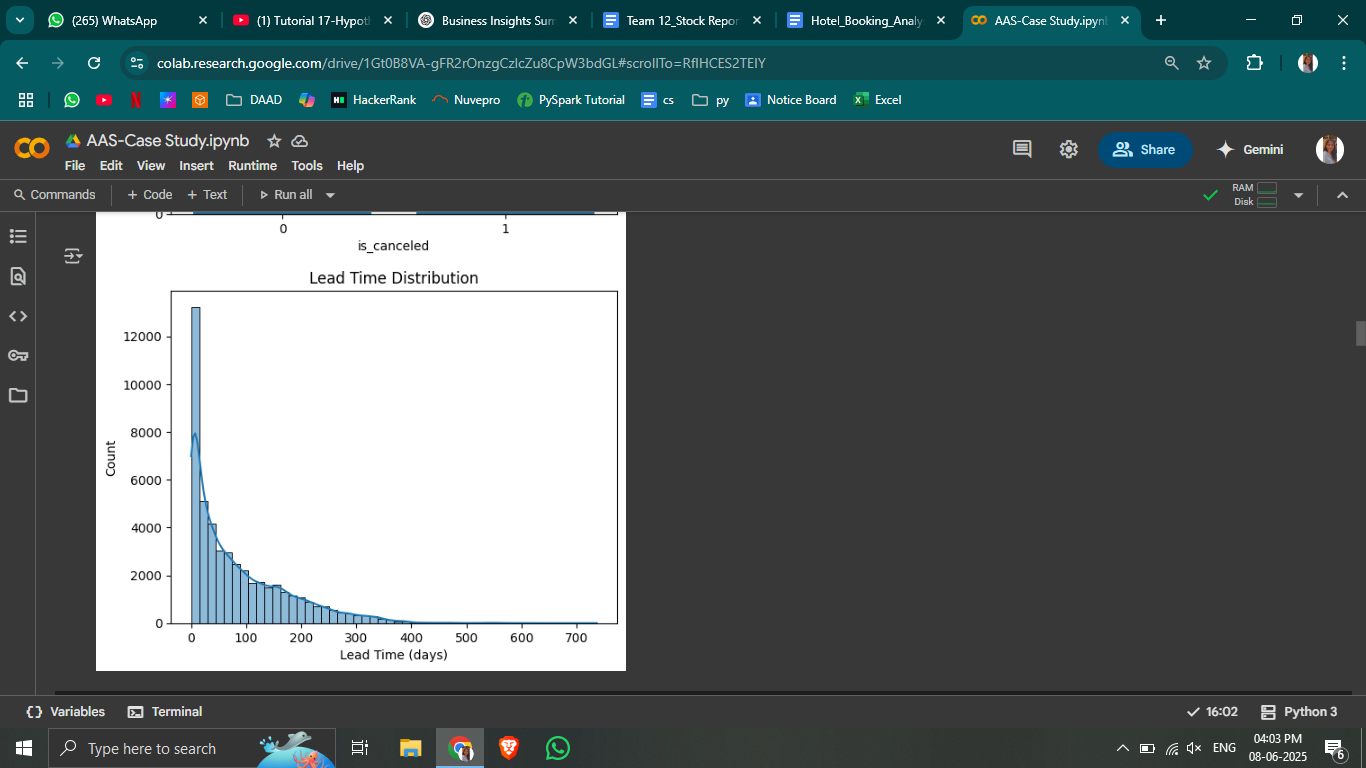
Univariate analysis focuses on examining the distribution of individual variables to understand their behavior and properties.

1. Booking Status

* A **count plot** was used to visualize the distribution between bookings that were honored and those that were canceled.  
    
   **Graph: Booking Status Countplot** **
  + *Insight:* A significant portion of the bookings were canceled, highlighting the need to explore cancellation trends in more depth.

2. Lead Time

* Lead time refers to the number of days between the booking date and the arrival date. A **histogram** of lead time showed a right-skewed distribution.  
    
   **Graph: Lead Time Distribution**

****

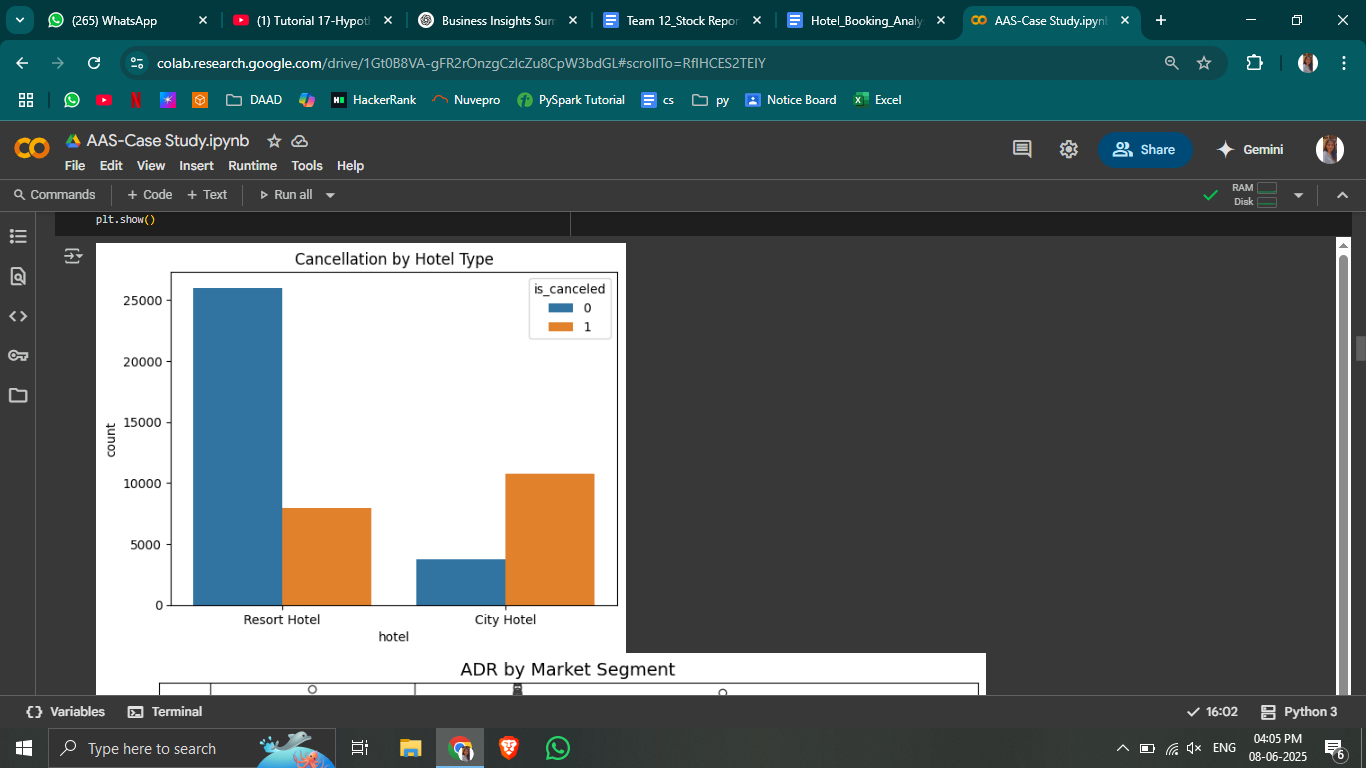
* + *Insight:* Many bookings are made close to the check-in date, but there is a long tail of customers booking well in advance.

**3.2.2 - Bivariate Analysis**

This analysis focuses on relationships between two variables, especially how categorical variables relate to numerical outcomes.

1. Hotel Type vs. Cancellations

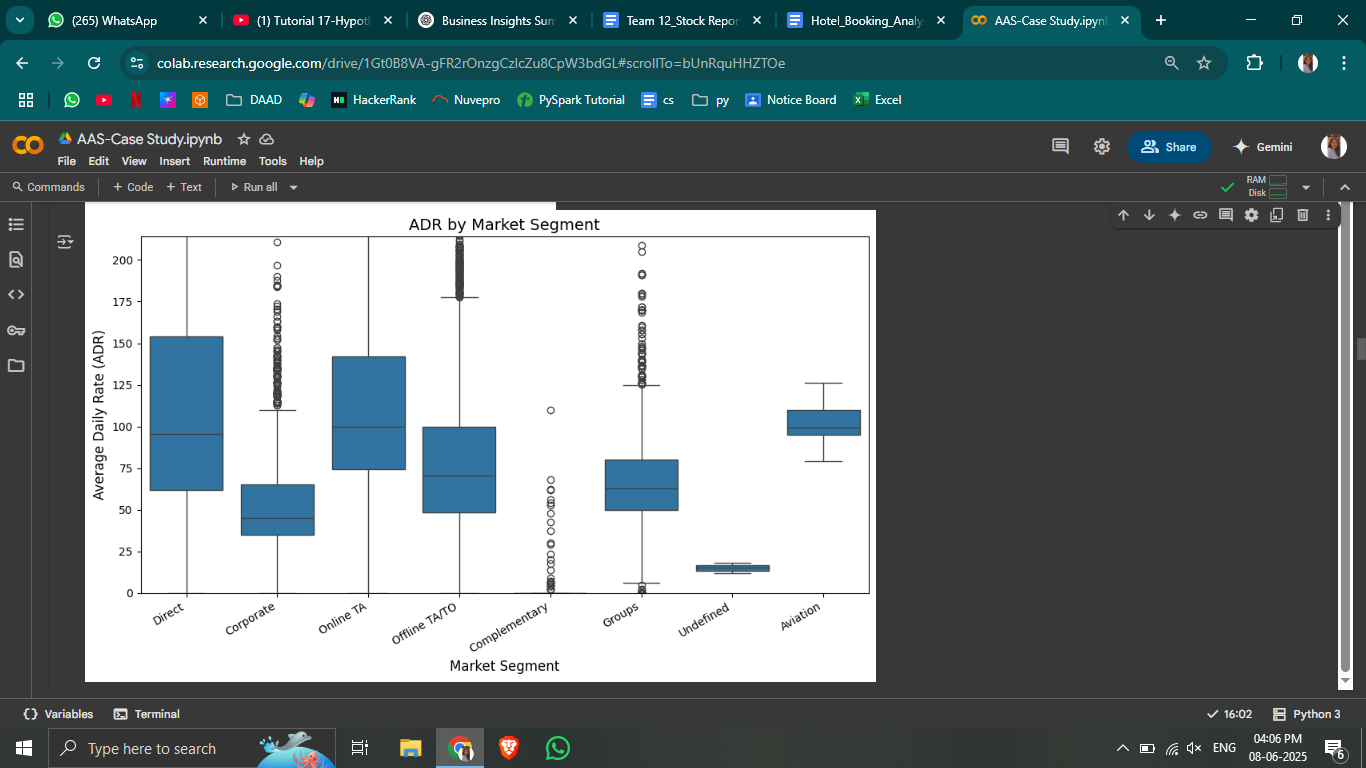
* A **stacked bar chart** was used to show the proportion of cancellations between resort and city hotels.  
    
   **Graph: Cancellation Rate by Hotel Type**

****

* + *Insight:* City hotels experienced a higher cancellation rate compared to resort hotels.

2. Market Segment vs. ADR

* A **boxplot** illustrated differences in Average Daily Rate (ADR) across various market segments.  
    
   **Graph: ADR by Market Segment**

****

* + *Insight:* Certain segments like "Online Travel Agents (OTA)" showed a wider spread and higher ADR variability.

**3.2.3 - Multivariate Analysis**

Multivariate analysis explores interactions between three or more variables.

1. Correlation Heatmap

* A **heatmap** was generated using Pearson and Spearman correlations to understand linear and monotonic relationships among numerical variables.  
    
   **Graph: Correlation Heatmap**

****

* + *Insight:* Lead time, number of special requests, and booking changes showed significant correlations with ADR and booking outcomes.

2. Room Type Mismatch vs. Lead Time and Special Requests

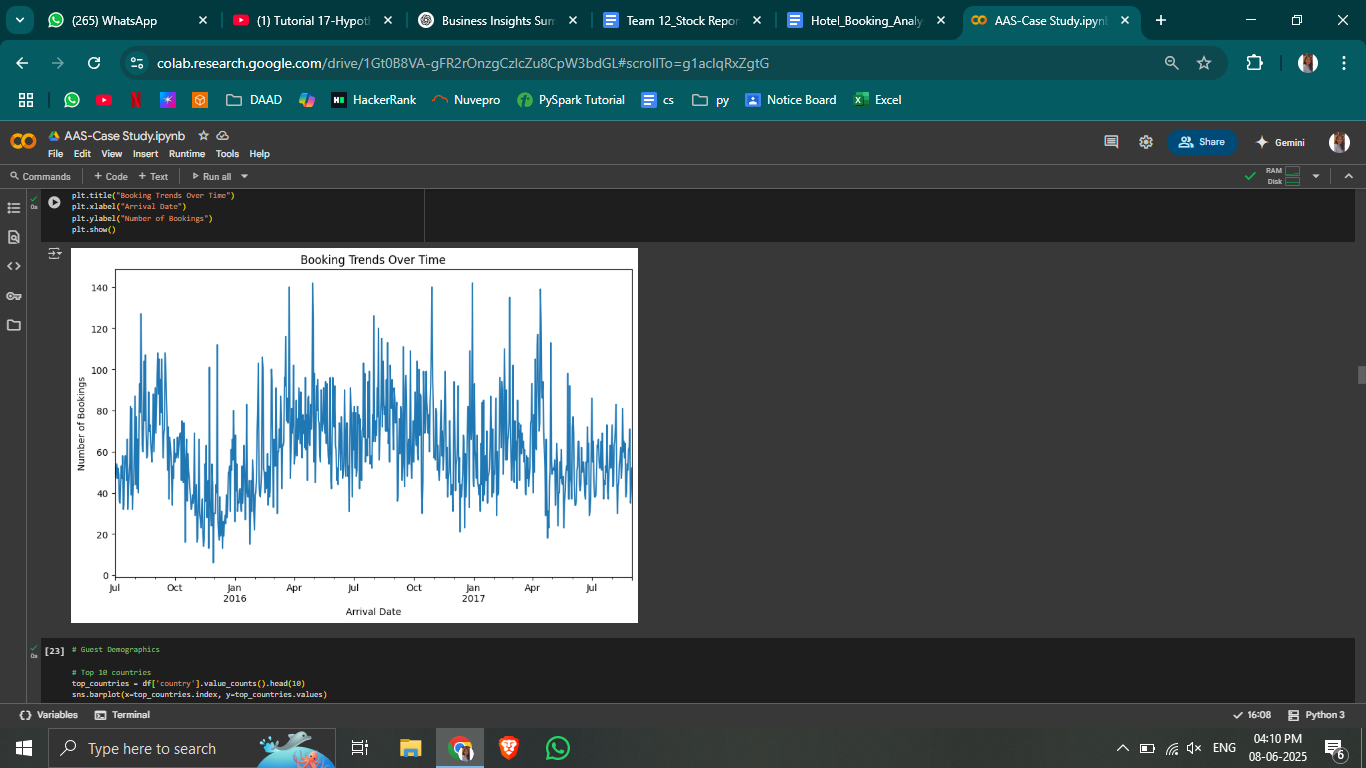
* A **grouped bar chart** revealed that bookings with a longer lead time or more special requests often resulted in room type mismatches.

**3.3 - Time-Series Analysis**

Understanding how bookings fluctuate over time is critical for operations and pricing strategies.

1. Monthly Booking Trend

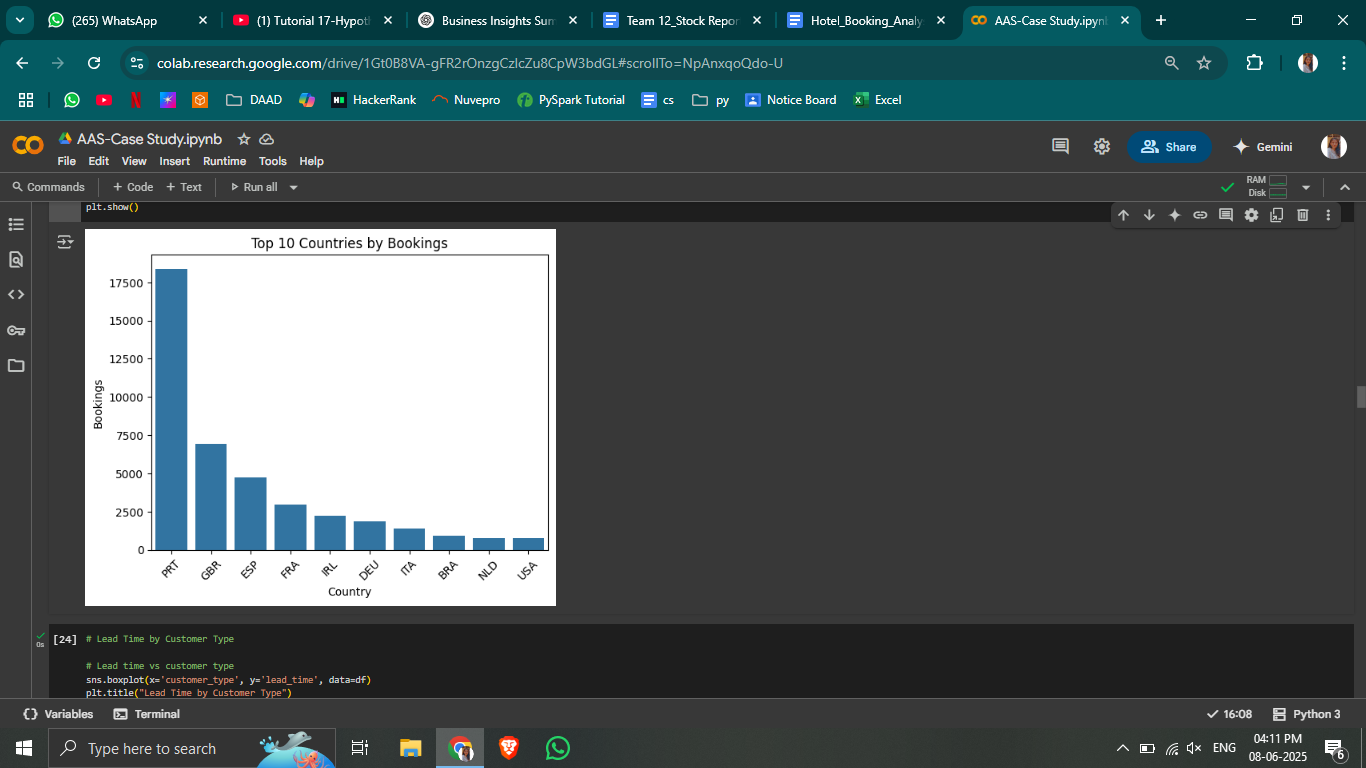
* A **line chart** plotted the number of bookings per month across the dataset timeline.  
    
   **Graph: Monthly Booking Trend**

****

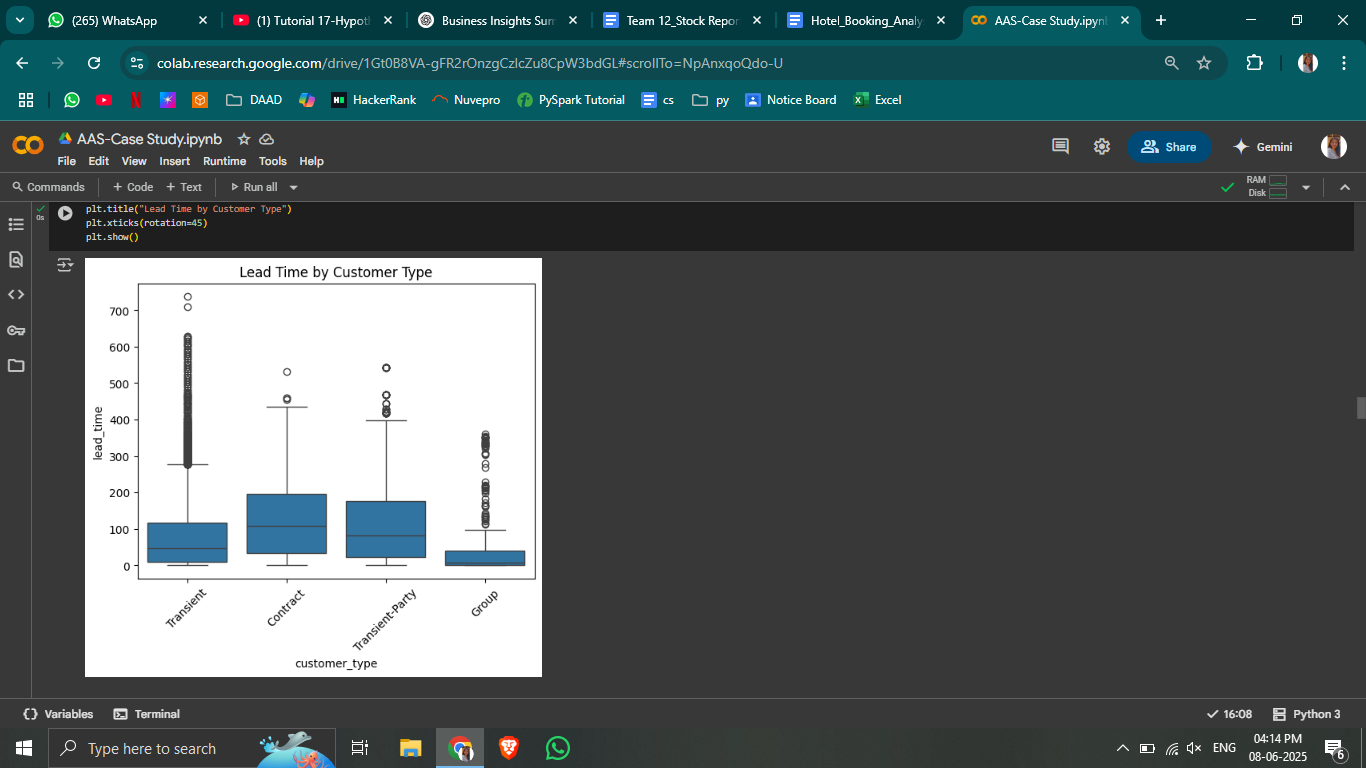
* + *Insight:* Booking activity showed seasonality, with peaks around summer months and dips in the off-season.

**3.4 - Guest Demographics**

1. Country-Wise Distribution

* A **bar chart** displayed the top 10 countries by the number of bookings.  
    
   **Graph: Top 10 Countries by Booking Volume  
  **
  + *Insight:* Portugal had the highest number of bookings, followed by the UK and France.

2. Customer Type vs. Lead Time

* A **boxplot** compared lead times across customer types (e.g., transient, contract, group).  
    
   **Graph: Lead Time by Customer Type  
  **
  + *Insight:* Contract customers typically booked with shorter lead times, while transient guests booked further in advance.

**Chapter 4: Correlation Analysis & Hypothesis Testing**

**4.1 - Objective**

This chapter aims to uncover statistical relationships between key variables and validate assumptions using formal hypothesis tests. Correlation analysis helps understand how variables move together, while hypothesis testing determines if observed differences are statistically significant or due to random chance.

**4.2 - Correlation Analysis**

Correlation analysis helps identify and quantify the strength and direction of relationships between continuous variables.

1. Pearson Correlation

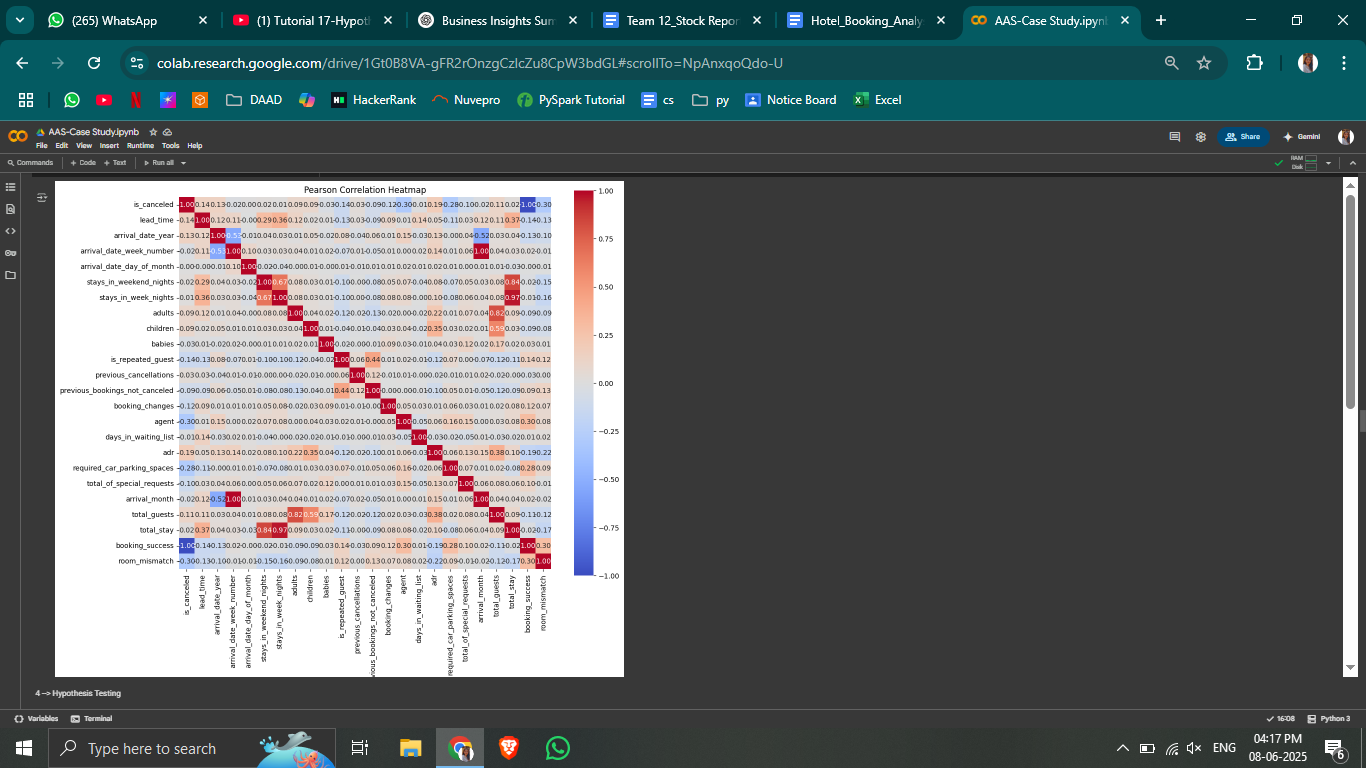
* Measures linear relationships between pairs of variables.
* Best used when data is normally distributed.

2. Spearman Correlation

* A rank-based method that captures monotonic (but not necessarily linear) relationships.
* Useful when variables are not normally distributed or have outliers.

**4.3 - Graph: Correlation Heatmap (Pearson & Spearman)**

* A dual **heatmap** was generated to visualize Pearson and Spearman correlation coefficients for numerical variables like:  
  + lead\_time
  + adr
  + special\_requests
  + booking\_changes
  + stays\_in\_weekend\_nights, stays\_in\_week\_nights
  + is\_canceled

**

* **Insights:**
  + lead\_time showed a **positive correlation** with cancellation rate.
  + adr correlated moderately with the number of special requests and booking success.
  + total\_guests was positively related to total\_stay, as expected.
  + Some relationships were weak or insignificant, indicating a need for deeper analysis through hypothesis testing.



**4.4 - Hypothesis Testing**

Hypothesis testing was performed to validate assumptions and assess the significance of observed patterns in the data.

1. T-Test: ADR for 'TA/TO' vs 'Direct' Bookings

* **Hypothesis:**
  + **Null (H₀):** The average ADR is the same for bookings made through Travel Agents (TA/TO) and Direct customers.
  + **Alternate (H₁):** The average ADR differs significantly between these groups.
* **Result:**
  + The T-test showed a statistically significant difference in ADR between the two booking types.

**4.5. Chi-Square Test: Lead Time vs. Room Upgrades**

* **Hypothesis:**
  + **Null (H₀):** Lead time and room upgrade frequency are independent.
  + **Alternate (H₁):** Lead time and room upgrades are associated.
* **Result:**
  + The Chi-Square test indicated a significant association between lead time and the likelihood of room upgrades.

**4.6. ANOVA: Customer Type vs. Stay Duration**

* **Hypothesis:**
  + **Null (H₀):** Mean stay duration is equal across all customer types.
  + **Alternate (H₁):** At least one customer type has a significantly different average stay.
* **Result:**
  + ANOVA results confirmed a statistically significant difference in stay duration across customer types.

**Chapter 5: Key Business Questions**

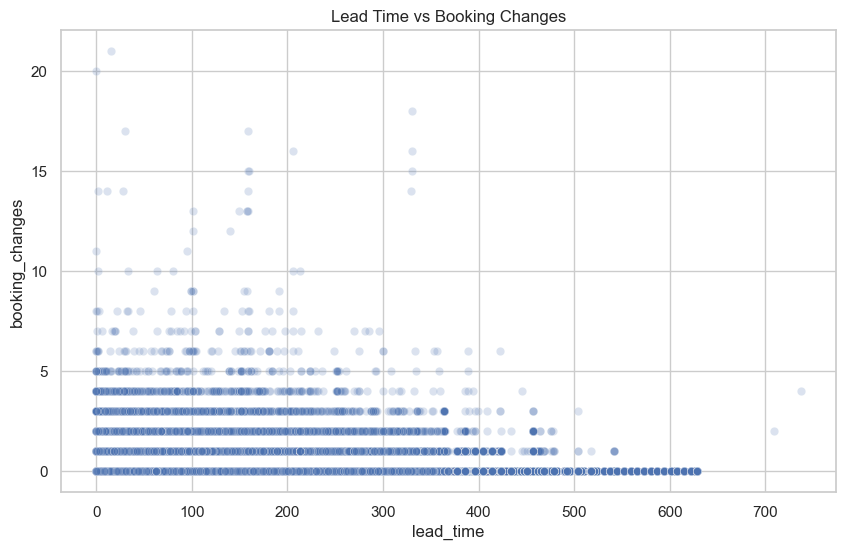
**5.1 - Objective**

The purpose of this chapter is to answer crucial business questions derived from stakeholder interests. These questions focus on understanding guest behavior, pricing dynamics, booking trends, and operational inconsistencies, all of which influence hotel performance and profitability.

1. What Factors Influence ADR (Average Daily Rate)?

**Analysis:**

* Investigated the relationship between ADR and variables like market segment, booking lead time, room type, customer type, and special requests.



**Conclusion:** ADR is heavily influenced by **market segment**, **booking channel**, **seasonality**, and **booking lead time**.

2. Do Guests Who Book Earlier Tend to Request More Changes or Upgrades?

**Analysis:**

* Explored the relationship between lead time and the number of special requests or booking changes.

**Conclusion:** Yes, guests who book earlier tend to **request more changes** and **make more special requests**, potentially due to more time to plan or change plans.

3. Are There Pricing or Booking Differences Across Countries?

**Analysis:**

* Compared average ADR and booking frequencies across top 10 source countries.

**Conclusion:** Yes, there are significant **differences in both pricing and volume** based on the country of origin. This can inform **localized pricing strategies**.

4. Is There a Pattern in Room Upgrades or Reassignments?

**Analysis:**

* Focused on comparing reserved room types with assigned room types.
* Created a feature for **room mismatch** (when the reserved and assigned room types don’t match).

**Conclusion:** Room mismatches are **not random**—they correlate with **booking pressure**, **lead time**, and **special request load**, suggesting a need for **better inventory planning**.

5. Are Reserved Room Types Consistently Matched with Assigned Room Types?

**Analysis:**

* A confusion matrix or mismatch frequency table was generated.

**Conclusion:** There are **frequent mismatches** between reserved and assigned rooms, indicating either overbooking or manual reassignment practices, which could impact guest satisfaction.

6. What Are the Most Common Guest Demographics?

**Analysis:**

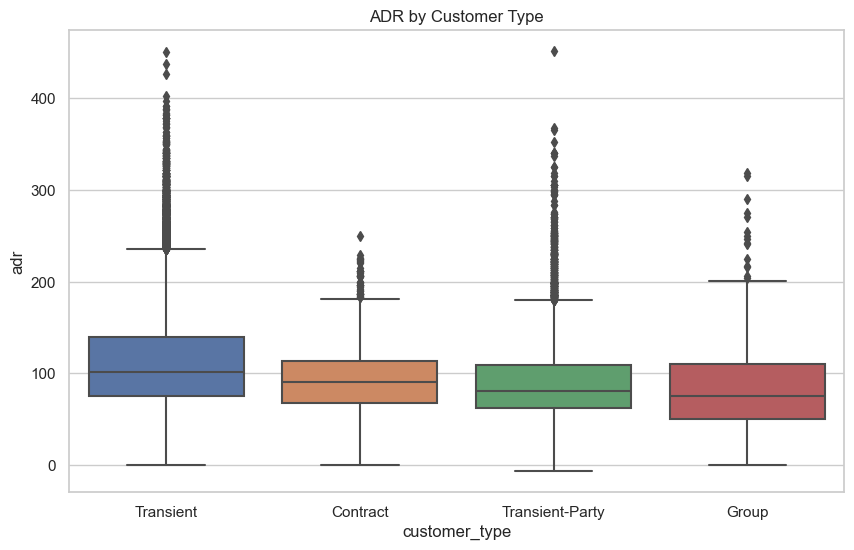
* Evaluated guest origin, customer type, number of guests, and presence of children or babies.

**Conclusion:** Most guests are **individual travelers or small groups**, coming from a **handful of key countries**, and booking through **online channels**.

7. How Does Booking Lead Time Vary Across Customer Types?

**Analysis:**

* Compared lead time across customer types using descriptive statistics and visualizations.



**Conclusion:** Lead time **varies widely** by customer type, influencing **forecasting**, **pricing**, and **staffing decisions**.

**Chapter 6: Business Recommendations**

**6.1 - Objective**

Based on the data-driven insights from exploratory analysis, correlation tests, and business question assessments, this chapter outlines targeted recommendations that hotels can implement to optimize operations, pricing, customer satisfaction, and revenue.

**6.2. Implement Dynamic Pricing Based on Lead Time and Market Segment**

**Insight:**

* ADR varies significantly across booking lead times and market segments.
* OTA customers book earlier but often receive lower ADRs.

**Recommendation:**

* Introduce a **dynamic pricing model** that adjusts ADR based on:  
  + Lead time windows (e.g., 0–7 days, 8–30 days, 31+ days)
  + Segment type (Direct, OTA, Corporate)
  + Seasonality and demand forecast

Suggested Graph: Heatmap of ADR by Lead Time and Segment

This can highlight which combinations yield optimal rates and where discounting or upcharging may be strategic.

**6.3. Predict Cancellations Early Using Machine Learning**

**Insight:**

* Cancellations are linked with high lead time, special requests, and booking changes.

**Recommendation:**

* Develop a **cancellation prediction model** to identify high-risk bookings early.
* Use predictive alerts to:  
  + Overbook cautiously in high-cancellation slots.
  + Proactively reach out to at-risk customers with confirmations or incentives to retain them.

Suggested Graph: Lead Time vs. Cancellation Rate Trendline

Shows how probability of cancellation rises with lead time.

**6.4. Tailor Marketing Campaigns Based on Geography**

**Insight:**

* Booking behaviors and ADRs differ across countries.

**Recommendation:**

* Launch **geo-targeted marketing campaigns** focusing on:  
  + Countries with high booking volume (e.g., Portugal, UK)
  + Countries with high ADRs for profitability
  + Low-volume but high-potential countries with special offers or cultural packages

Suggested Graph: Country-wise ADR vs Booking Volume Bubble Chart

Enables segmentation by both profitability and market size.

**6.5. Improve Room Assignment Accuracy to Reduce Mismatches**

**Insight:**

* Room mismatches are frequent, especially with longer lead times or during high-demand periods.

**Recommendation:**

* Introduce an **intelligent room allocation system** that:  
  + Prioritizes room assignment accuracy at reservation stage
  + Locks high-demand room types earlier for loyal or high-value guests
  + Flags likely mismatches during overbooking or reassignment

Suggested Graph: Room Mismatch Rate by Month or Occupancy Level

Highlights operational lapses or peak pressure periods.

**6.6. Personalize Offers Based on Customer Type and Behavior**

**Insight:**

* Transient and group guests show distinct behavior in stay duration, lead time, and ADR.

**Recommendation:**

* Use past behavior to tailor offers:  
  + **Transient guests:** Personalized packages or upgrades
  + **Group bookings:** Loyalty points, repeat discounts
  + **Corporate clients:** Streamlined booking experience, flat-rate stays

Suggested Graph: Customer Type vs. Stay Duration and ADR Comparison

Shows how revenue opportunities can be maximized per segment.

**6.7. Upsell Services Based on Special Requests**

**Insight:**

* Special requests indicate guest intent or preferences (e.g., crib, late check-in).

**Recommendation:**

* Train staff or automate systems to **suggest relevant upsells** like:  
  + Premium rooms
  + Food packages
  + Early check-in/late checkout offers
* Integrate with CRM to deliver automated yet personal messages pre-arrival

**Chapter 7: Conclusion & Business Insights**

**7.1 Summary of Key Findings**

The analysis revealed important patterns in customer behavior, booking trends, and pricing strategies. Key insights include distinct customer segments with varying booking habits, the influence of lead time on pricing, and noticeable seasonal demand fluctuations. Additionally, mismatches between reserved and assigned room types highlight operational challenges.

**7.2 Business Implications**

These findings suggest opportunities to optimize revenue through dynamic pricing, target marketing efforts towards profitable customer groups, and improve operational efficiency by reducing booking errors. Seasonal trends can guide resource allocation and staffing decisions, while channel performance insights help in managing partnerships effectively.

**7.3 Recommendations**

* Adopt predictive pricing to better match demand.
* Enhance customer data collection for personalized marketing.
* Upgrade booking systems to minimize errors.
* Develop loyalty programs for repeat customers.
* Use ongoing data analysis to refine strategies regularly.

**7.4 Conclusion**

Overall, leveraging these insights enables smarter decision-making, improved customer satisfaction, and stronger business performance. A data-driven approach will support sustainable growth amid changing market conditions.