ANALYZING ADR HOTEL BOOKING DATA

Naufal Dzakia Raiffaza

Introduction

I'm Naufal Dzakia Raiffaza , an Informatics Engineering graduate from Universitas Muhammadiyah Surakarta, with a strong passion for Data Science and Python programming. I combine technical expertise with analytical thinking to solve real-world problems through data.

With experience in international research collaboration , published work in software ecosystems, and hands-on projects in machine learning and sentiment analysis, I thrive in environments that value innovation, teamwork, and continuous learning. Currently expanding my skills through a Data Science Bootcamp , I’m eager to share insights from my recent projects and explore how data can drive impactful decisions.

Project Background

- Purpose:
    - This project focuses on predicting the Average Daily Rate (ADR) for hotel bookings using advanced data analysis and machine learning. ADR is a critical metric for hotel revenue management, reflecting the average income per paid occupied room per day. Accurate ADR prediction enables hotels to optimize pricing, maximize revenue, and plan effective marketing and operational strategies.
- Business Value:
    - Revenue Optimization: By forecasting ADR, hotels can set optimal prices for upcoming periods, especially during high-demand seasons or special events.
    - Proactive Planning: Predictive insights support better inventory management, targeted promotions, and resource allocation.
    - Competitive Advantage: Leveraging ADR predictions helps hotels stay ahead in a dynamic market by responding quickly to demand fluctuations.

- **Stakeholders:**
    - **Hotel managers, revenue analysts, and marketing teams benefit directly from improved ADR prediction and actionable insights.**

Problem Statement

Despite having rich booking data, hotels face several challenges directly related to ADR prediction:

- Uncertainty in Room Pricing
- Ineffective Promotions
- Missed Revenue Opportunities
- Complex Demand Patterns:

Project Goals

This project aims to solve that by cleaning and analyzing the dataset to uncover insights into:

- Develop Accurate ADR Prediction Models:
  - Build and evaluate machine learning models (e.g., Random Forest, XGBoost) to forecast next month's ADR with high accuracy.
- Analyze Key Drivers of ADR Fluctuations:
  - Identify and quantify the main factors influencing monthly ADR changes, such as booking volume, seasonality, and special events.
- Visualize ADR Trends and Predictions:
  - Present ADR trends over time, the relationship between ADR and booking numbers, and the predicted ADR distribution for future periods using clear, impactful charts.
- Enable Data-Driven Pricing Strategies:
  - Use ADR predictions to inform dynamic pricing, targeted promotions, and value-added packages, maximizing revenue across different demand scenarios.
- Support Business Planning:
  - Provide actionable recommendations for pricing tiers, and operational planning based on ADR prediciton and booking patterns.

Problem Definition and Business Objectives

Business Problems

1. **Uncertainty in Room Pricing:**

Difficulty in setting the optimal monthly room price (ADR) due to lack of data-driven forecasts.

1. **Ineffective Promotions:**

Promotions and discounts are often not targeted because seasonal and customer behavior patterns are not well understood.

1. **Missed Revenue Opportunities:**

Without accurate ADR predictions, the hotel risks losing revenue during high demand and facing low occupancy in slow periods.

Analytical Questions

1. What are the main factors influencing monthly ADR fluctuations?
2. How do ADR and booking trends change over time (monthly, seasonally, yearly)?
3. How accurate is the machine learning model (e.g., Random Forest) in predicting next month's ADR, and how can these predictions support business decisions?

Business Objectives

1. **Increase Hotel Revenue:**

The images show how ADR optimization, based on predictions, can maximize revenue.

1. **Proactive Pricing and Promotion:**

Seasonal and demand trends in the images justify the need for proactive strategies.

1. **Improve Planning:**

The predictive distribution and historical trends provide the foundation for better operational and financial planning.

DATA UNDERSTANDING

Data Overview

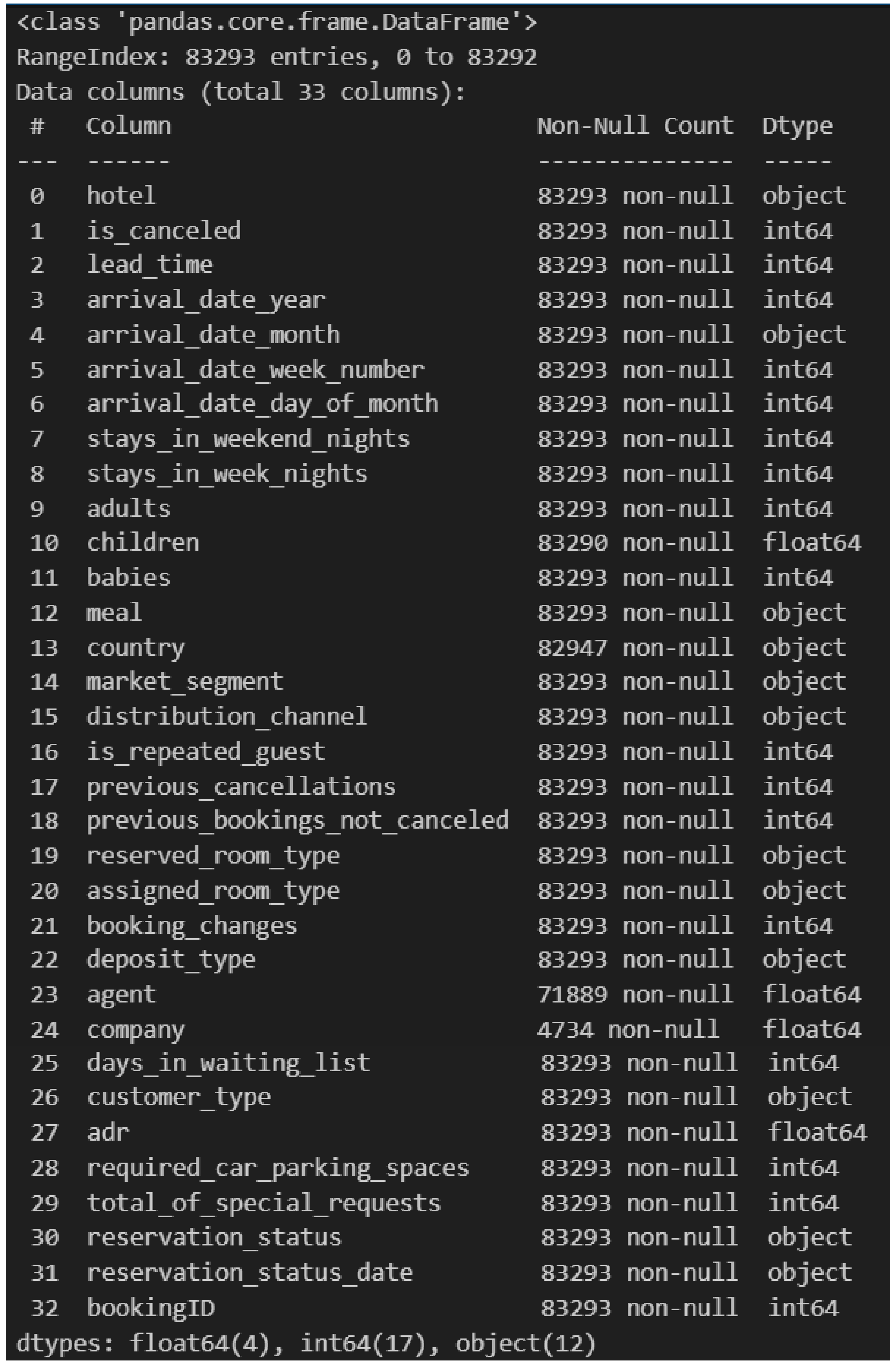
The dataset contains 83,293 entries (representing individual bookings) and includes 33 columns with different types of data. Here's a brief breakdown of the dataset structure:

Hotel Information:

- hotel: The type of hotel (e.g., Resort, City Hotel).

Customer Information:

- country: The country of origin of the guest.
- is_repeated_guest: Indicates if the guest is a repeat customer (1 = yes, 0 = no).
- previous_cancellations: The number of previous cancellations by the guest.
- previous_bookings_not_canceled: The number of previous bookings by the guest that were not canceled.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 83293 entries, 0 to 83292
Data columns (total 33 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   hotel                           83293 non-null  object
 1   is_canceled                     83293 non-null  int64
 2   lead_time                       83293 non-null  int64
 3   arrival_date_year               83293 non-null  int64
 4   arrival_date_month              83293 non-null  object
 5   arrival_date_week_number        83293 non-null  int64
 6   arrival_date_day_of_month       83293 non-null  int64
 7   stays_in_weekend_nights         83293 non-null  int64
 8   stays_in_week_nights            83293 non-null  int64
 9   adults                          83293 non-null  int64
 10  children                        83290 non-null  float64
 11  babies                          83293 non-null  int64
 12  meal                            83293 non-null  object
 13  country                         82947 non-null  object
 14  market_segment                  83293 non-null  object
 15  distribution_channel            83293 non-null  object
 16  is_repeated_guest               83293 non-null  int64
 17  previous_cancellations          83293 non-null  int64
 18  previous_bookings_not_canceled  83293 non-null  int64
 19  reserved_room_type              83293 non-null  object
 20  assigned_room_type              83293 non-null  object
 21  booking_changes                 83293 non-null  int64
 22  deposit_type                    83293 non-null  object
 23  agent                           71889 non-null  float64
 24  company                         4734 non-null   float64
 25  days_in_waiting_list            83293 non-null  int64
 26  customer_type                   83293 non-null  object
 27  adr                             83293 non-null  float64
 28  required_car_parking_spaces     83293 non-null  int64
 29  total_of_special_requests       83293 non-null  int64
 30  reservation_status              83293 non-null  object
 31  reservation_status_date         83293 non-null  object
 32  bookingID                       83293 non-null  int64
dtypes: float64(4), int64(17), object(12)
```

Booking Information:

- **is_canceled:** Indicates if the booking was canceled (1 = canceled, 0 = not canceled).
- **lead_time:** The time (in days) between booking and arrival.
- **arrival_date_year, arrival_date_month, arrival_date_day_of_month:** Date-related columns to track when the booking is scheduled.
- **stays_in_weekend_nights, stays_in_week_nights:** Duration of the stay (in nights).
- **adults, children, babies:** Number of guests (adults, children, babies) in the booking.
- **meal:** Type of meal selected (e.g., bed & breakfast, half board).
- **market_segment:** Customer category (e.g., online travel agents, direct bookings).
- **distribution_channel:** The channel through which the booking was made.

Financial Information:

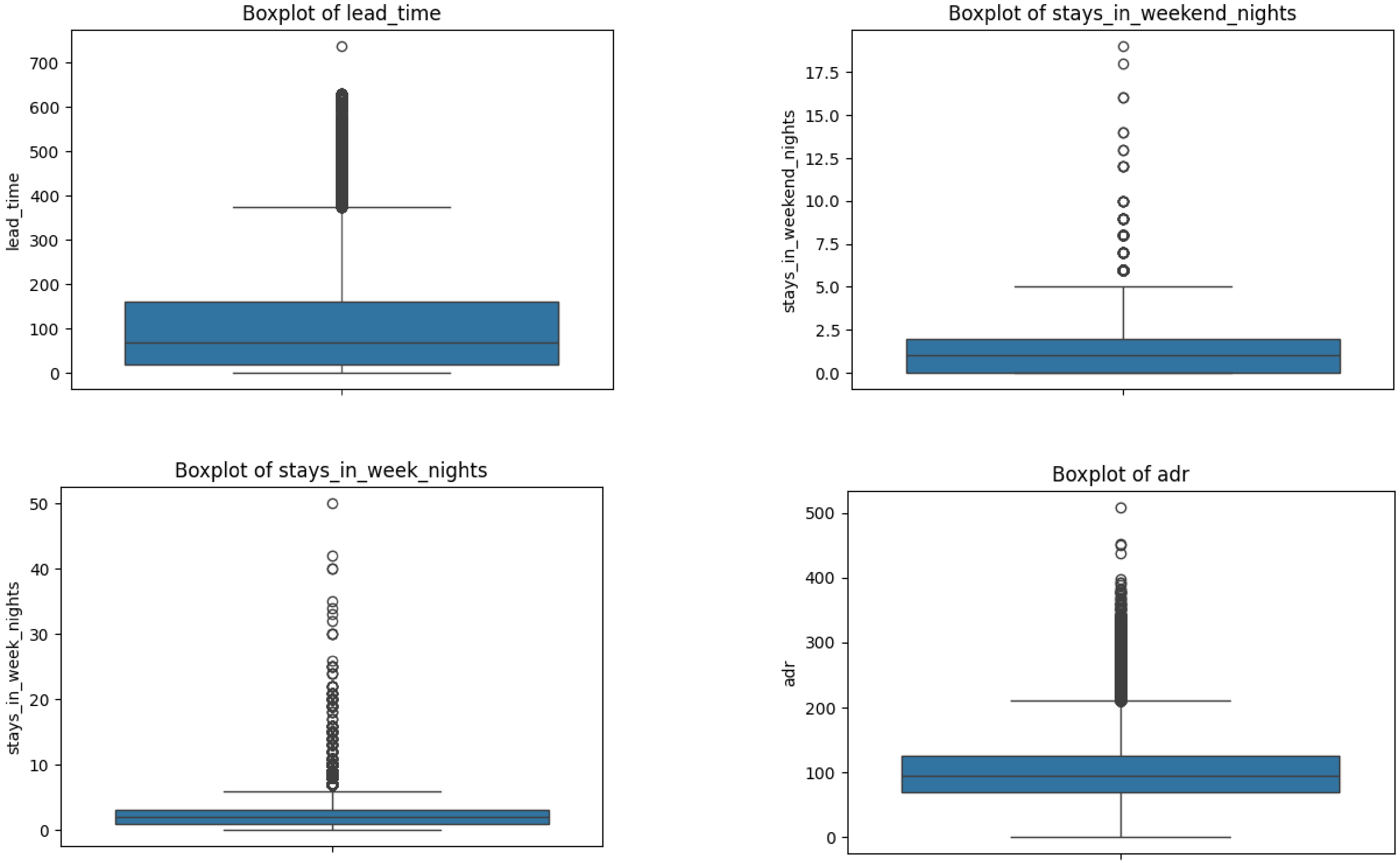
- adr (Average Daily Rate): The average daily rate of the room booked.
- days_in_waiting_list: The number of days the booking was on the waiting list.

DATA CLEANING

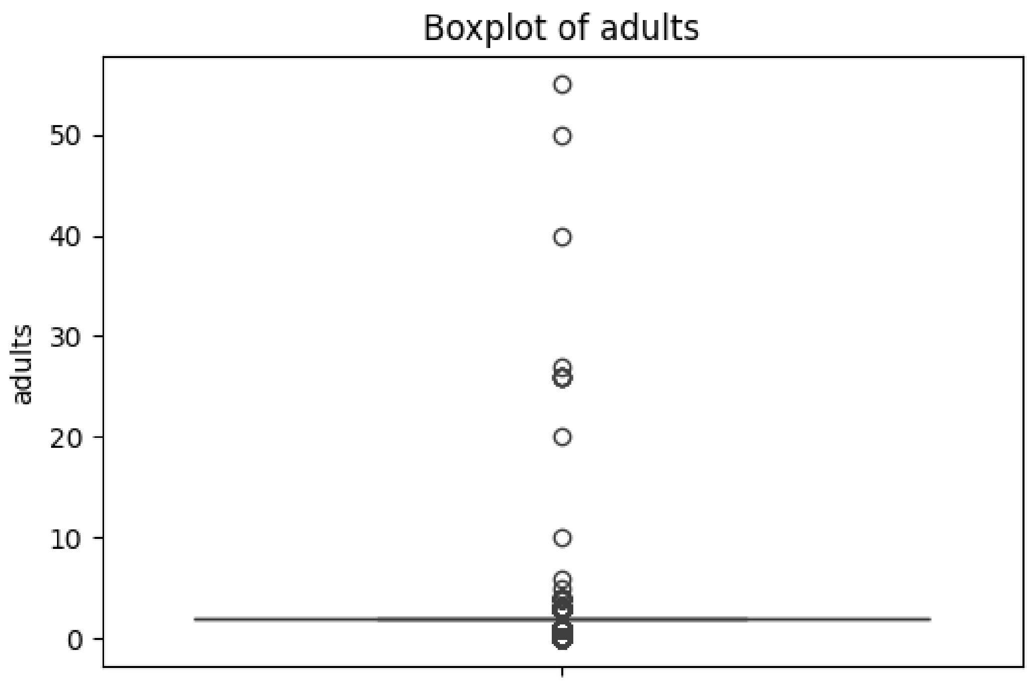
- Check duplicated data:
  - The dataset are safe from duplicated data

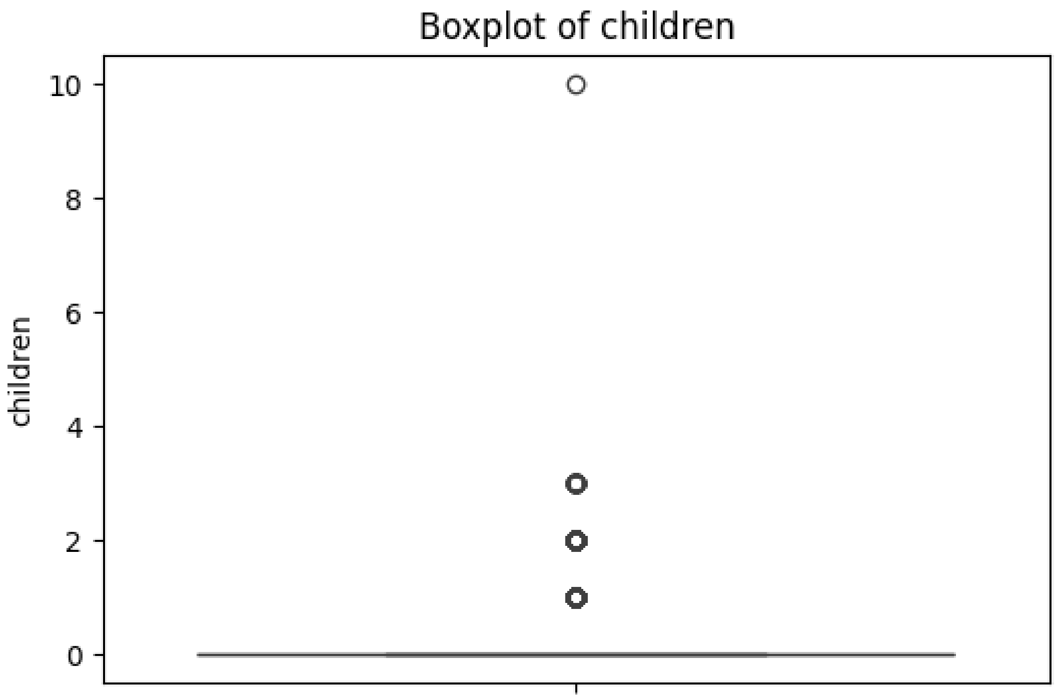
- Check Missing Values:
  - There are missing value on ‘Country’, ‘Children’, ‘Agent’, and ‘Company’
  - The percentage of missing values on each column, Except ‘Company’ that has 94.32%. We handle it because the percentage are 20% [1]
  - The 'Agent' column is filled with 0, indicating bookings without an agent or unknown agent.
  - The 'Country' column is filled with 'Unknown' to preserve all records for analysis.
  - The 'Children' column is filled with 0, assuming missing means no children were present in the booking.

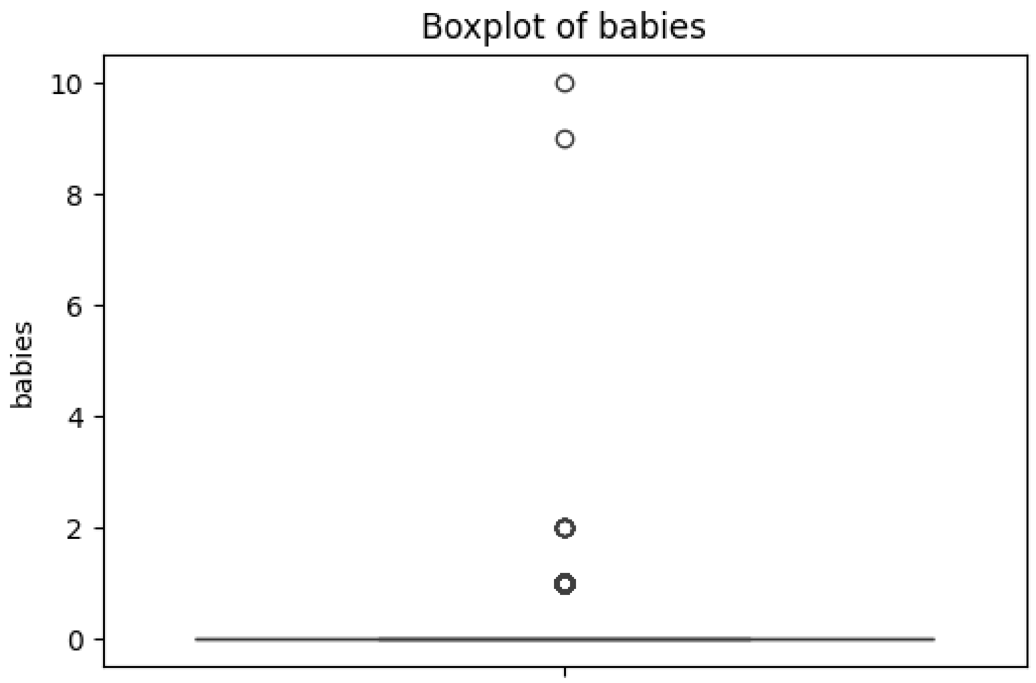
Data Cleaning, Outliers



Data Cleaning, Outliers


Boxplot of adults


Boxplot of children


Boxplot of babies

Data Cleaning, Outliers

- **Outliers Explanation:**
  - **Very Large Number of Adults (`adults > 20`) (Group Bookings)**
    - While most bookings are for individuals, couples, or small families, it is possible to have bookings for large groups, such as tour groups, corporate events, or conferences. These bookings may have a very high number of adults. If these records are labeled as group bookings (e.g., `customer_type = "Group"` and `market_segment` such as "Direct" or "Offline TA/TO"), they are valid and should be kept.
    - **Action:**
      - These records are kept because they are consistent with valid group bookings.

- **Example inside our dataset**

adults	market_segment	customer_type
Direct	Group	4243
Offline TA/TO	Group	28123
Direct	Group	28471
Direct	Group	38012
Offline TA/TO	Group	39004

Notes: It’s same case for Children and Babies

Data Cleaning, Outliers

- **Outliers Explanation:**
  - **Bookings with `adults=0` (Invalid Bookings)**
    - In the hotel business, a booking must include at least one adult. Bookings with zero adults are not logical, as hotels do not allow children or babies to stay without an adult present. These records are considered invalid and should be removed from the dataset.
    - **Action:**
      - All rows with `adults=0` are removed from the dataset.

- **Example inside our dataset**

adults	children	babies
0	3	44
0	2	48
0	2	258
0	2	1378
0	2	1522

Data Cleaning, Outliers

- **Outliers Explanation:**
  - **High ADR (Average Daily Rate) Outliers**
    - The ADR column represents the average daily room rate paid by guests. While high values can occur for luxury rooms, premium dates, or special events, extremely high ADRs may indicate data entry errors. In our dataset, statistical outlier detection (using the IQR method) flags ADR values above 210.70 as outliers. However, the maximum ADR in our data is 508.0, and many of these "outliers" are simply higher-end or premium bookings, not necessarily errors.
    - **Action:**
      - No records need to be removed for high ADR, as values above 210.70 are likely premium bookings or high-demand periods.

- **Example inside our dataset**

Metric	Value
Outliers	2672
Lower Bound	-15.17
Upper Bound	210.7
Min	0
Max	508
Example Outliers	[249.0, 343.0, 335.0, 235.0, 235.45, 228.33, 227.86, 225.0, 217.08, 337.0, 309.0, ...]

Data Cleaning, Outliers

- **Outliers Explanation:**
  - **Inconsistent Lead Time**
    - The `lead_time` column records the number of days between the original booking date and the arrival date. The `reservation_status_date` column indicates the final status date (such as check-out, cancellation, or no-show), not the booking date. Therefore, the difference between `lead_time` and the calculated difference between `arrival_date` and `reservation_status_date` can be large and is not an error.
    - **Action:**
      - **Do not remove these records. Use the `lead_time` column as provided.**

- **Example inside our dataset**

lead_time	arrival_date	reservation_status_date
524	Dec 15, 2018	Oct 21, 2017
395	Mar 21, 2018	Aug 10, 2017
167	Jul 29, 2018	Jul 24, 2018
101	Jun 27, 2018	May 6, 2018
112	Dec 6, 2018	Oct 31, 2018

DATA SCIENCE

MODELLING

Machine Learning Modelling
Data Preprocessing, Feature Engineering

Encoding

Feature	Encoding Method	Reason
Hotel	One-Hot Encoding	Hotel categories are nominal with no order, so One-Hot Encoding avoids implying any ranking.
Country	One-Hot Encoding	Countries have no natural order, so One-Hot Encoding treats each as a separate category.
Meal	One-Hot Encoding	Meal types are nominal and have no order, so One-Hot Encoding is used to prevent implying any ranking.
Market Segment	One-Hot Encoding	Market segments have no inherent order, so One-Hot Encoding is used to represent each separately.
Customer Type	One-Hot Encoding	Nominal Category, there's no inherent order.
Reservation Status	One-Hot Encoding	Reservation statuses are nominal with no order, so One-Hot Encoding treats each status as a separate category.

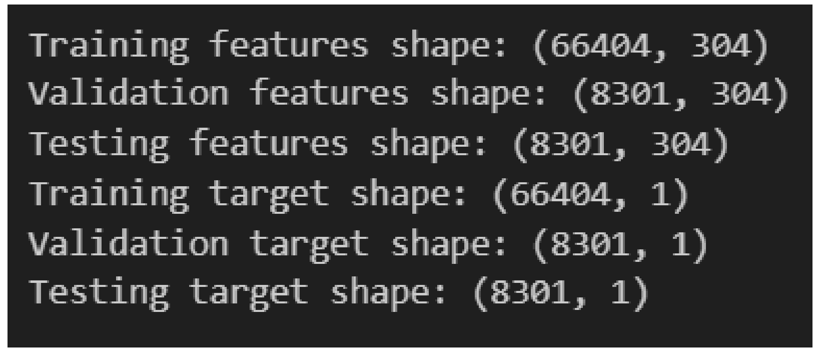
Drop Column

- agent (Unique ID)
- bookingID (Unique ID)
- Company (Because 94.32% data is missing)

Machine Learning
Modelling

Split Data

- Train - Validation- test



```
Training features shape: (66404, 304)
Validation features shape: (8301, 304)
Testing features shape: (8301, 304)
Training target shape: (66404, 1)
Validation target shape: (8301, 1)
Testing target shape: (8301, 1)
```

Trained Model

- Random Forest
- LightGBM
- Decision Tree
- XG-Boost
- Linear Regression

Comparing Method

- Before tuning
- After tuning

Evaluation Metrics

- R-squared
- MAE
- RMSE

What we do

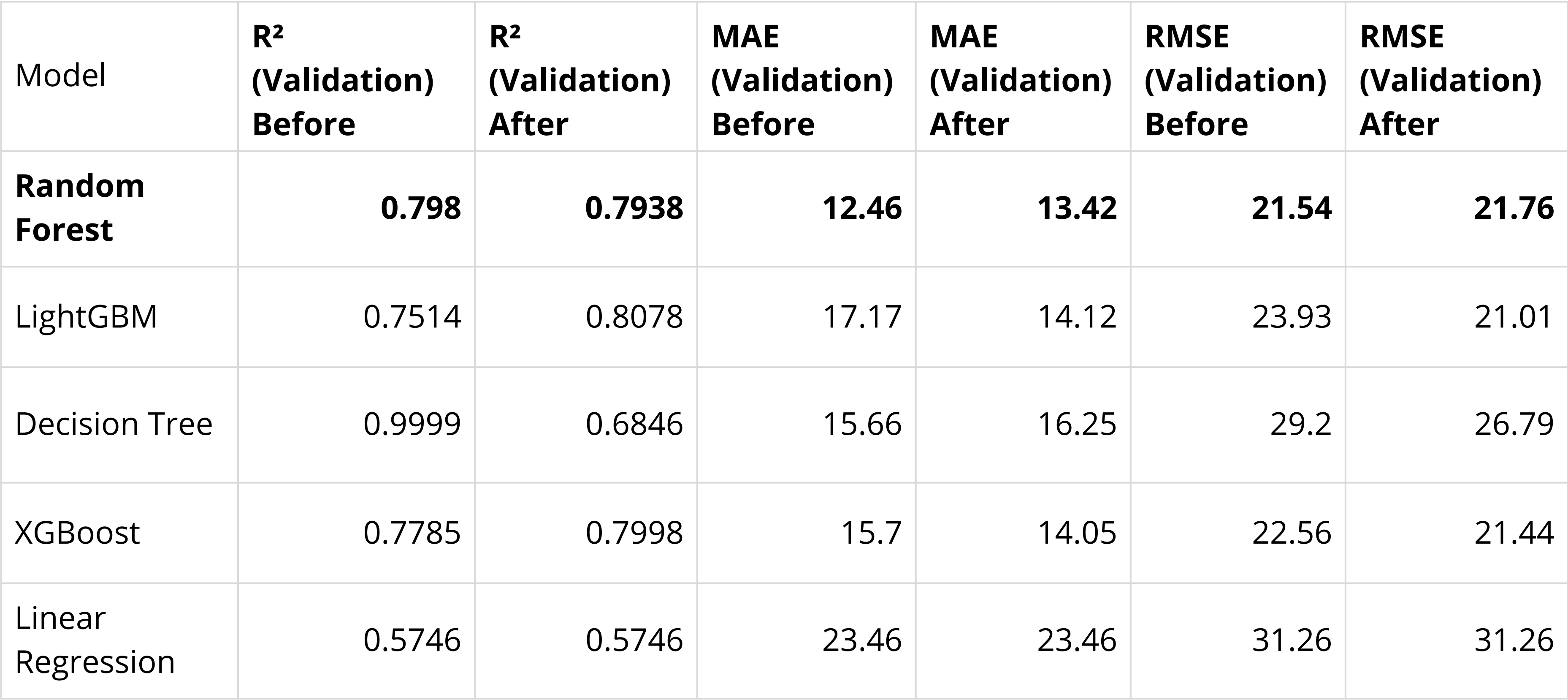
- Tuning
- Validation

Results

- Model by model Performance Analysis
- Create model based on best model which can predict ....

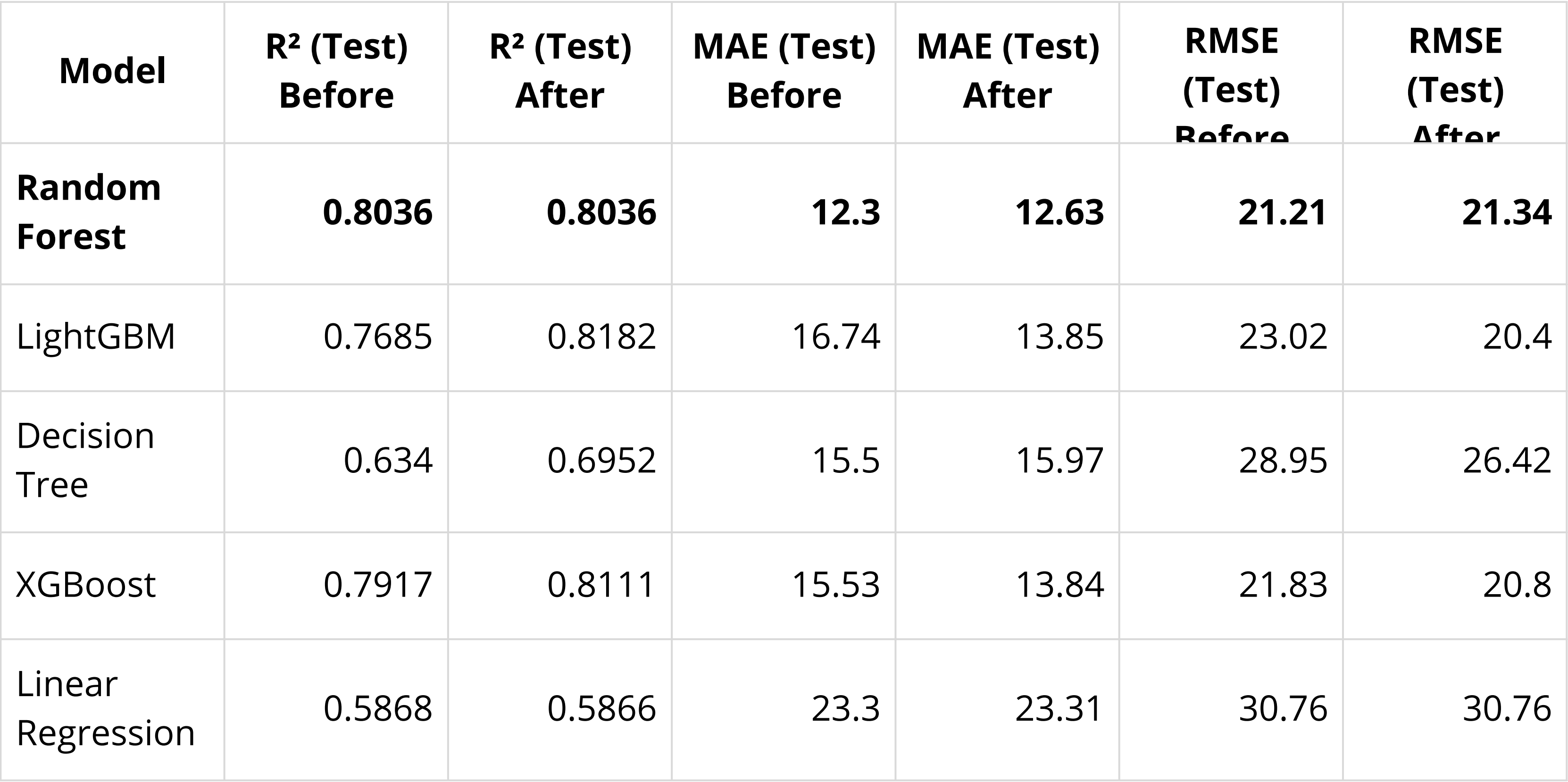
Before vs After Tuning
Model Performance

- **Validation Performance (Before vs After Tuning)**



Model	**R² (Validation) Before**	**R² (Validation) After**	**MAE (Validation) Before**	**MAE (Validation) After**	**RMSE (Validation) Before**	**RMSE (Validation) After**
Random Forest	**0.798**	**0.7938**	**12.46**	**13.42**	**21.54**	**21.76**
LightGBM	0.7514	0.8078	17.17	14.12	23.93	21.01
Decision Tree	0.9999	0.6846	15.66	16.25	29.2	26.79
XGBoost	0.7785	0.7998	15.7	14.05	22.56	21.44
Linear Regression	0.5746	0.5746	23.46	23.46	31.26	31.26

- **Test Performance (Before vs After Tuning)**



Model	**R² (Test) Before**	**R² (Test) After**	**MAE (Test) Before**	**MAE (Test) After**	**RMSE (Test) Before**	**RMSE (Test) After**
Random Forest	**0.8036**	**0.8036**	**12.3**	**12.63**	**21.21**	**21.34**
LightGBM	0.7685	0.8182	16.74	13.85	23.02	20.4
Decision Tree	0.634	0.6952	15.5	15.97	28.95	26.42
XGBoost	0.7917	0.8111	15.53	13.84	21.83	20.8
Linear Regression	0.5868	0.5866	23.3	23.31	30.76	30.76

Final Results

Best Overall Model: Random Forest, Before Tuning

Random Forest before tuning is the best model across all metrics (R^2, MAE, RMSE). Despite a slight drop in performance after hyperparameter tuning, it consistently delivers the best R^2, MAE, and RMSE values compared to other models, particularly XGBoost and LightGBM.

R^2 Comparison:

Random Forest before tuning performs the best in terms of R^2, achieving 0.8036 on the test set. In comparison, XGBoost and LightGBM show lower R^2 values, especially on the test set (XGBoost: 0.7917, LightGBM: 0.7685), indicating that Random Forest explains the variance in the data more effectively.

MAE Comparison:

When comparing Mean Absolute Error (MAE), Random Forest again outperforms the other models with the lowest error values: 12.46 on the validation set, and 12.30 on the test set. XGBoost and LightGBM have higher MAE values, especially on the test set (XGBoost: 15.53, LightGBM: 16.74), demonstrating that Random Forest provides more accurate predictions.

RMSE Comparison:

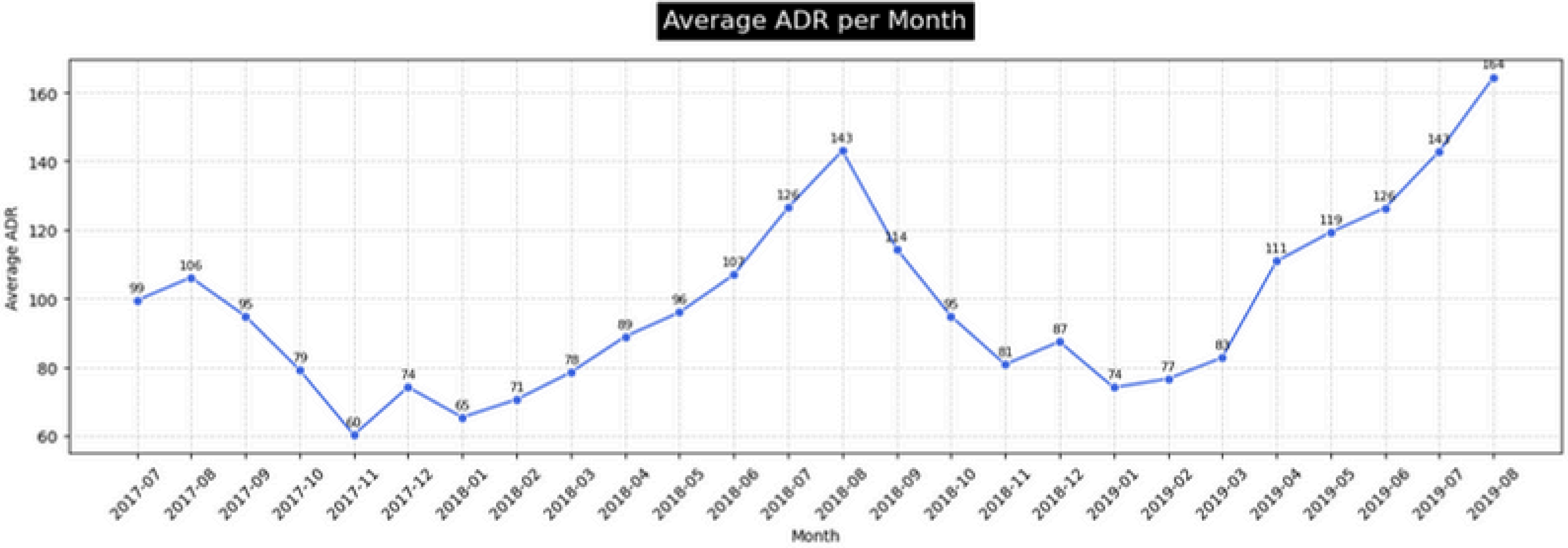
For Root Mean Squared Error (RMSE), Random Forest consistently shows the lowest error across all datasets, with 21.54 on the validation set, and 21.21 on the test set. In comparison, XGBoost (21.83) and LightGBM (23.02) have higher RMSE values, reflecting greater prediction error, especially on the test set.

PROJECT

EDA, GARTNER ANALYTIC ASCENDANCY MODEL

EDA, Gartner Analytic Ascendancy Model

Average ADR Per Month



Descriptive Analytics

Average ADR Per Month

Seasonal Pattern

ADR consistently rises during the middle of the year, peaking in August, and then drops at the end and beginning of each year. This repeating pattern highlights strong seasonality in hotel pricing.

Year-on-Year Growth

There is a clear upward trend in ADR over the years, with the highest value observed in August 2019. This suggests that the hotel has been able to increase its rates over time, possibly due to higher demand, improved services, or market positioning.

Volatility

Some months show sharp increases or decreases in ADR, indicating the influence of special events, holidays, or market changes that temporarily affect pricing.

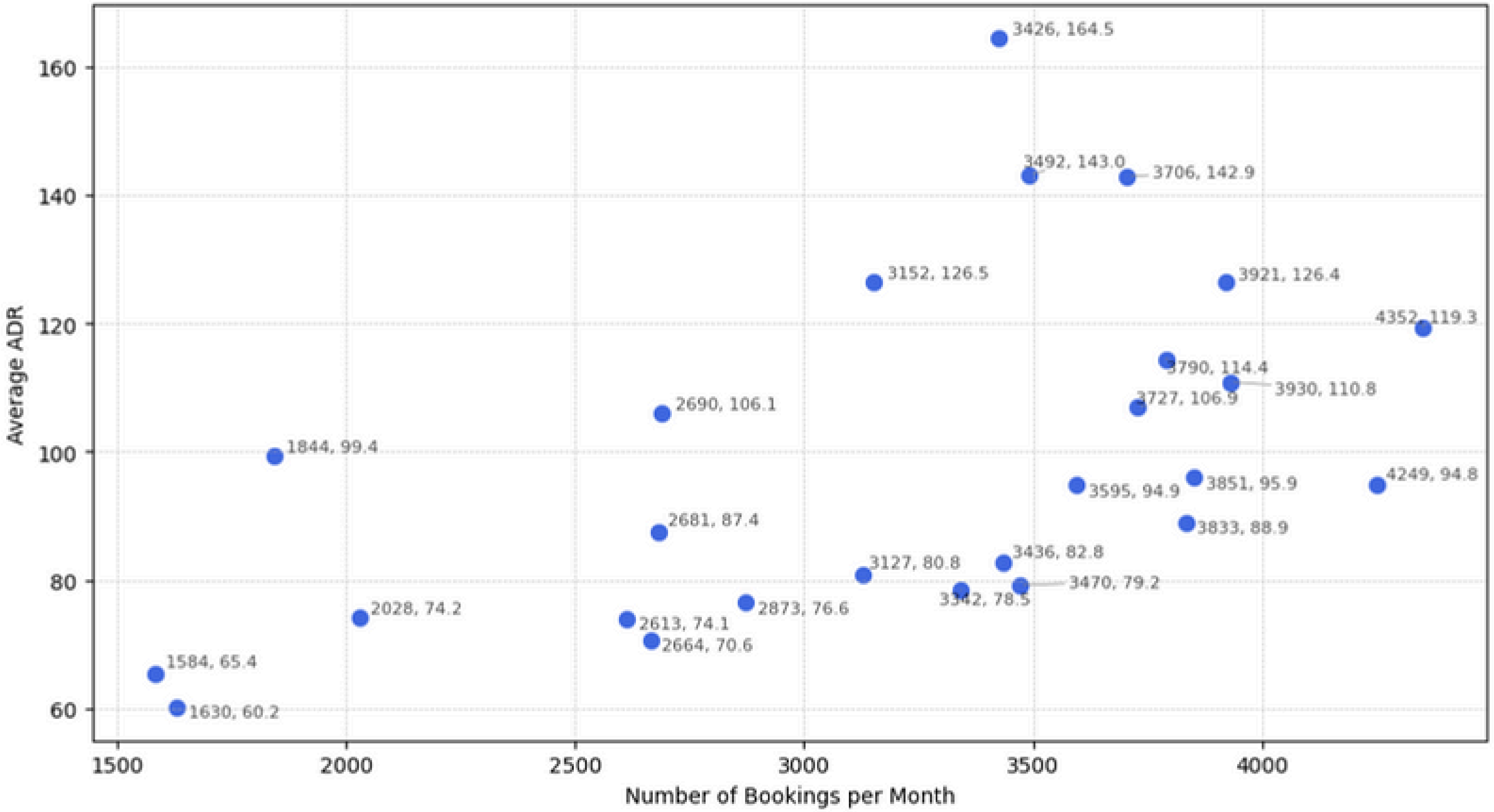
Business Implication

Recognizing these patterns allows hotels to:

- Plan pricing strategies in advance.
- Anticipate high- and low-demand periods.
- Optimize revenue by adjusting rates according to expected demand

EDA, Gartner Analytic Ascendancy Model

Relationship between ADR and Number of Booking per Month



Diagnostic Analytics

Relationship between ADR and Number of Booking per Month

- **Positive Correlation**

  There is a clear positive relationship months with more bookings tend to have higher ADR. This means that when demand increases, hotels are able to charge higher rates, which is typical in the hospitality industry during peak seasons or special events.

- **Clustered Data**

  Most data points are concentrated in the mid-to-high range for both bookings and ADR, indicating that the majority of months experience moderate to high demand and pricing.

- **Outliers and Extremes**

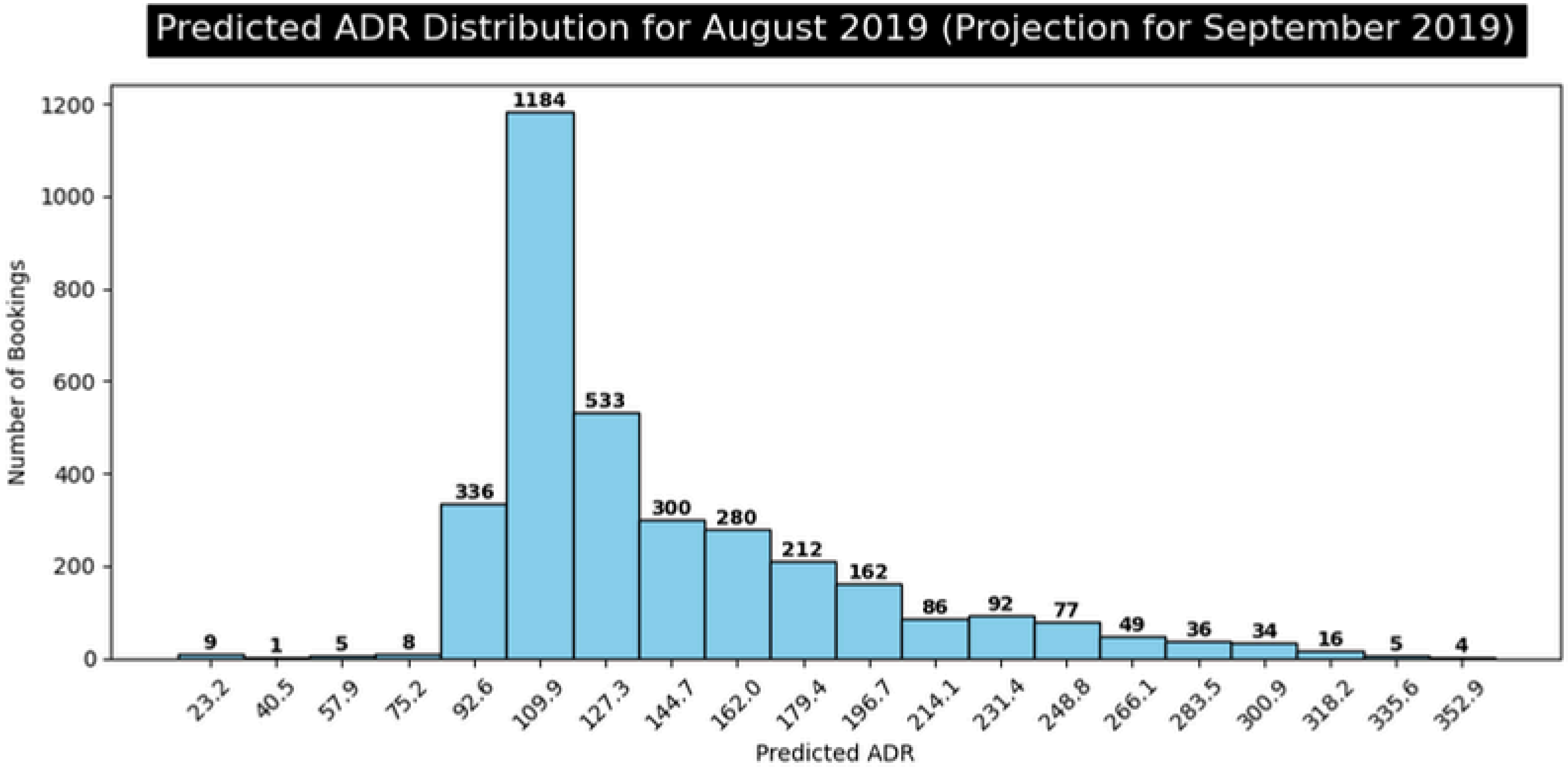
  Some months stand out with exceptionally high ADR and booking numbers, likely corresponding to peak travel periods, holidays, or major events. Conversely, months with low bookings also tend to have lower ADR, reflecting low season or periods of reduced demand.

- **Business Implication**

  This relationship supports the use of dynamic pricing strategies raising rates when demand is high and offering promotions or discounts when demand is low. It also highlights the importance of monitoring booking trends to optimize revenue

EDA, Gartner Analytic Ascendancy Model

Predicted ADR Distribution



EDA, Gartner Analytic Ascendancy Model

Predicted ADR Distribution

Distribution Shape

The distribution is skewed to the right, meaning most bookings are concentrated in the lower to mid-range ADRs, with a long tail of fewer bookings at higher prices.

Peak Price Range

The most frequent predicted ADR is centered around $115.1 , with 887 bookings . This indicates that this price range is the most likely for the upcoming month.

Primary Booking Zone

The majority of bookings are predicted to have an ADR between $86.9 and $157.5 . This range represents the core market for the hotel's pricing strategy.

Low and High Extremes

There are very few bookings predicted at the extremely low end (e.g., ADR below $72.7) or the very high end (e.g., ADR above $228.1). This suggests that deep discounts or premium luxury rates are not common for the projected period.

PROJECT

RECOMMENDATION
&
ACTION

Predictive Analytics: Predicted ADR Distribution

Key Takeaways

- Seasonal Pattern: ADR peaks mid-year (especially August), dips at the start/end of the year—use this for forecasting [1].
- Demand & Price: More bookings = higher ADR. Raise rates when demand is high, lower when demand is low[2],[3].
- Next Month: Average ADR forecast is $107.36; most bookings expected in the $92.6-$162.0 range.

Simple Recommendations

- Segment our Prices
  - Core: $92.6-$162.0 (main target for most bookings)
  - Discount: $75-$92.5 (for last-minute/low demand)
  - Premium: $162.1-$200+ (special events/high demand)
  - void offering standard rates below $75, as this can devalue the brand and rarely boosts revenue proportionally [3].
- Automate Rate Changes
  - Use a dynamic pricing system to adjust rates daily based on bookings and inventory [2],[4].
  - Set alerts for slow bookings to trigger flash sales or promos.
- Promote Value Packages
  - Offer bundles (e.g., breakfast, late checkout) instead of just discounts [5].
  - Bundling increases perceived value and keeps ADR strong [6].
- Focus on the Right Channels
  - Invest marketing in channels and segments that book in your core ADR range.
  - Review channel performance weekly and adjust as needed.
- Clean Data, Watch Outliers
  - Exclude very low ADR bookings from main calculations to avoid skewed averages.
  - Regularly check for pricing errors or data issues.

Predictive Analytics

Predicted ADR Distribution

Scenario Table

Scenario	ADR	Booking Change	Revenue Impact
10% Price Increase	$118	-5%	4%
10% Discount (Low Season)	$97	8%	2%
Bundling Promotion	$110	Stable	+2–3%

Explanation.

1. 10% Price Increase, Bookings drop 3-8%, revenue rises 2-5% [3],[7],[8].
   - Why?
     - In high-demand periods, hotel bookings are typically inelastic—meaning guests are less sensitive to price increases. Research and industry benchmarks show that a 10% price hike usually leads to a 3-8% drop in bookings, but the higher rate per room means total revenue still increases by 2-5%.
     - Example: If you raise prices by 10%, even if you lose a few bookings, the extra income per booking more than compensates for the loss in volume.
2. 10% Discount (Low Season),Bookings up 5-10%, revenue up 1-3% [2],[3].
   - Why?
     - In low season, demand is elastic—guests are more price-sensitive. A 10% discount can attract 5-10% more bookings, but the lower price means the revenue increase is modest, typically 1-3%.
     - Example: Lowering prices fills more rooms, but the gain in volume only slightly outweighs the loss in rate, so revenue grows slowly.
3. Bundling/Targeted Promotion, Revenue up 2-3% by adding value, not just cutting price [5],[6],[9].
   - Why?
     - Bundling (offering packages like breakfast or late checkout) adds perceived value, encouraging guests to spend more without lowering the base rate. Studies show this approach can increase total revenue per guest by 2-3% while keeping booking volume stable.
     - Example: Guests choose a slightly higher-priced package for added benefits, boosting overall revenue.

Predictive Analytics

Predicted ADR Distribution

Recommendation Implementation Steps

- Week 1: Set up pricing tiers in our system.
- Week 2: Launch 2–3 value package promos for different guest types.
- Weekly: Monitor bookings, ADR, and revenue; adjust as needed.
- Monthly: Review outlier data and competitor prices.

References

[1] https://pmc.ncbi.nlm.nih.gov/articles/PMC8183332/
[2] https://www.mews.com/en/blog/dynamic-pricing-hotels
[3] https://www.rateboard.io/en/blog/details/price-elasticity-in-the-hotel-industry
[4] https://openreview.net/forum?id=vxyv7E5j9A
[5] https://rategain.com/blog/walking-the-tight-rope-high-adr-and-low-distribution-cost/
[6] https://www.hotelmanagement.net/operate/new-expedia-white-paper-finds-package-bookings-drive-higher-adr
[7] https://journal.uii.ac.id/JKEK/article/download/27592/14818/84948
[8] https://scholarworks.umass.edu/bitstreams/9e733007-dfc1-493d-a393-5ee6d747ffc5/download
[9] https://journals.sagepub.com/doi/10.1177/14727978241298467

THANK YOU

Wanna discuss it together?
You can contact me whenever you want.

https://www.linkedin.com/in/naufal-dzakia-raiffaza/

https://github.com/raiffaza

Biological Science
Graduate School of Science and Technology
バイオサイエンス棟
Biological Science Complex
遺伝子教育研究センター
Research and Education Center for Genetic Information
高圧ガス貯蔵所