# HOTEL BOOKINGS

# DATA ANALYSIS REPORT

*Exploring Demand Patterns, Revenue Potential, and Cancellation Drivers*

Prepared by:

Justine V. Cientos

# EXECUTIVE SUMMARY

This data set contains booking information, from October 2015 to September 2017, for a city hotel and a resort hotel, and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things.

## DATA OVERVIEW:

```
Total Bookings: 119,390
Total Revenue: 25,996,260.41
Average Booking Value: 217.74
Total Cancelled Bookings: 44,224 (37.04%)
Average Waiting Days: 2.32
```

## KEY FINDINGS:

1. **How does seasonality affect booking demand?**
    - Booking demand shows strong **seasonal variation**, peaking between **April and August** and tapering off toward the end of the year.
    - **City Hotels** attract steady weekday demand from business travelers, while **Resort Hotels** experience weekend and summer peaks, consistent with leisure-driven stays.
    - These seasonal patterns highlight the need for **dynamic pricing and resource allocation**, ensuring staffing and inventory align with demand surges.
2. **Which customer segments provide the highest revenue potential?**
    - **Transient customers** generate the highest total revenue potential (≈11M) due to booking volume, but also carry the highest cancellation rates.
    - **Contract customers** contribute fewer bookings but yield the **highest average revenue per booking** and more stable revenue streams.
    - **Online Travel Agents (OTAs)** and the **TA/TO distribution channel** dominate market-driven revenue, accounting for more than half of overall revenue potential, while corporate and group bookings remain marginal contributors.

- This indicates that **volume-driven segments power topline revenue**, while **contract and corporate customers provide strategic stability**.

3. **What factors contribute most to booking cancellations, and how can they be reduced?**
    - **Lead time** is the strongest driver: the longer the gap between booking and arrival, the higher the cancellation risk.
    - Guests with a **history of previous cancellations** are significantly more likely to cancel again, while **repeat guests** show greater loyalty and lower cancellation rates.
    - Price sensitivity is evident: **higher ADR bookings** are slightly more prone to cancellations.
    - Operational signals matter: **special requests** is associated with lower cancellation likelihood, showing stronger guest intent.
    - **Mitigation strategies** include requiring deposits for long lead-time bookings, flagging high-risk customers with past cancellations, and reinforcing loyalty programs for repeat guests.

## STRATEGIES FOR REVENUE GROWTH:

1. **Leverage Seasonality**
    - Apply **dynamic pricing** during peak months to maximize yield.
    - Launch **off-peak promotions and bundled packages** to smooth demand in low months.
    - Align **marketing campaigns and staffing** with seasonal booking cycles.
2. **Optimize Customer Segments**
    - Maintain strong partnerships with **OTAs** for volume but grow **direct booking incentives** to reduce commission costs.
    - Expand **corporate and contract agreements** to secure stable, recurring revenue.
    - Develop tailored offers for **Transient-Party groups** to boost revenue from mid-performing segments.

3. **Reduce Cancellations to Protect Revenue**
    - Introduce **non-refundable deposits or stricter cancellation terms** for long lead-time bookings.
    - Implement a **guest risk scoring system** to flag customers with high cancellation histories.
    - Enhance **loyalty programs and perks** for repeat customers to drive retention.
    - Bundle value-added services (e.g., parking, dining, special requests) to increase guest commitment.

# INTRODUCTION

As a **data analyst for a hotel booking website**, my objective was to help the company understand booking behaviors, improve revenue management, and reduce booking cancellations.

I worked with the **Hotel Booking Demand dataset** (https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand), which contains over 119,000 records with features such as booking lead time, length of stay, cancellation status, market segment, customer type, and pricing.

I explored the data and decided to answer these **key business questions** that would help hotel managers make informed decisions. The questions I focused on were:

1. **How does seasonality affect booking demand?**
2. **Which customer segments provide the highest revenue potential?**
3. **What factors contribute most to booking cancellations, and how can they be reduced?**

Alongside this, I also designed a clear **workflow in Jupyter Notebook** and deliver the results in a way that was meaningful to both technical and business stakeholders, as a **report**.

# METHODOLOGY

To answer the **key business questions**, I structured my Jupyter Notebook into the following workflow:

1. **Data Cleaning & Preparation**

   Tools/libraries: *Python (pandas, numpy)*

   - Loaded the dataset into pandas and handled missing values.
   - Converted dates into seasons, engineered new features such as “overall_stay,” “revenue.”

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

   Tools/libraries: *Python (matplotlib, seaborn)*

   - Visualized cancellation patterns by lead time, previous cancellations, etc..
   - Plotted demand trends across seasons.
   - Calculated revenue potential across customer types, market segments and distribution channels to identify revenue drivers.

3. **Statistical Analysis & Modeling**

   Tools/libraries: *Python (scipy, statsmodels)*

   - Used ANOVA and logistic regression to determine significant predictors of cancellations.
   - Built simple forecasting visuals to show demand trends across time.

4. **Deliverables**

   - Created a **business presentation** highlighting key insights, with charts and visuals.

# RESULTS & ANALYSIS

## DATA CLEANING & PREPARATION:

After loading the dataset in Python, upon checking the dataset, a number of columns were observed to have missing values (replaced NaN=0 for children; NaN=N/A for country, agent and company):

| | | | |
|---|---|---|---|
| hotel | 0 | is_repeated_guest | 0 |
| is_canceled | 0 | previous_cancellations | 0 |
| lead_time | 0 | previous_bookings_not_canceled | 0 |
| arrival_date_year | 0 | reserved_room_type | 0 |
| arrival_date_month | 0 | assigned_room_type | 0 |
| arrival_date_week_number | 0 | booking_changes | 0 |
| arrival_date_day_of_month | 0 | deposit_type | 0 |
| stays_in_weekend_nights | 0 | agent | 16340 |
| stays_in_week_nights | 0 | company | 112593 |
| adults | 0 | days_in_waiting_list | 0 |
| children | 4 | customer_type | 0 |
| babies | 0 | adr | 0 |
| meal | 0 | required_car_parking_spaces | 0 |
| country | 488 | total_of_special_requests | 0 |
| market_segment | 0 | reservation_status | 0 |
| distribution_channel | 0 | reservation_status_date | 0 |

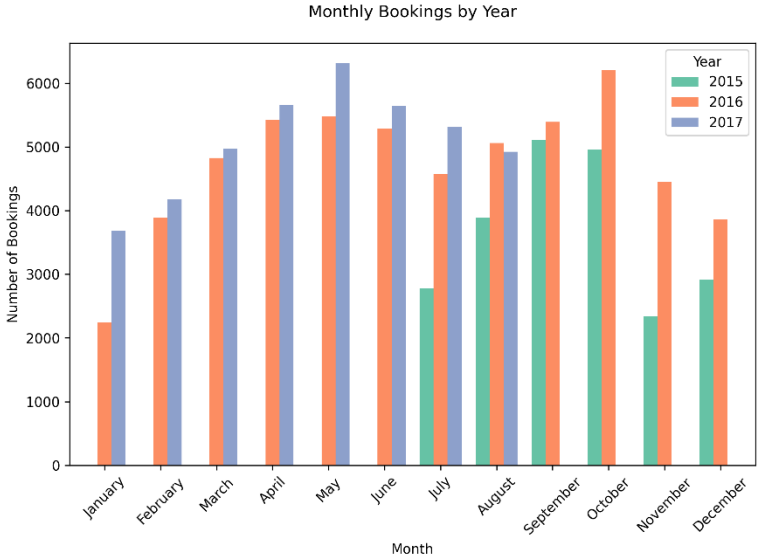
After replacing missing values, additional columns were added such as overall_stay and revenue.

## EXPLORATORY DATA ANALYSIS:

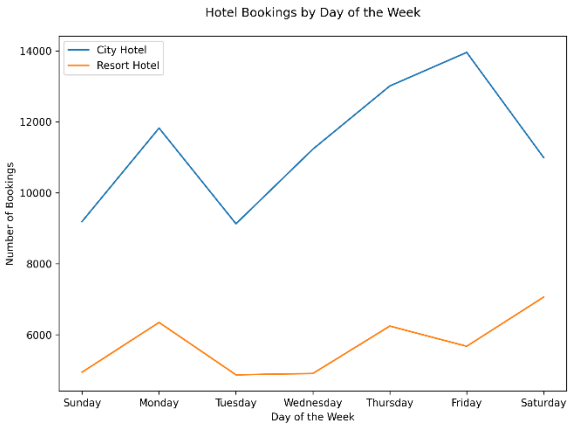
The purpose of this exploratory data analysis (EDA) is to gain a deeper understanding of the hotel booking dataset and uncover meaningful patterns that may influence booking behavior and cancellation trends. By examining key features such as **lead time**, **average daily rate (ADR)**, **repeat guest status**, **previous cancellations**, **booking changes**, and **special requests**, this analysis aims to highlight the most relevant factors that shape customer decisions and overall booking outcomes.

```
Total Bookings: 119,390
Total Revenue: 25,996,260.41
Average Booking Value: 217.74
Total Cancelled Bookings: 44,224 (37.04%)
Average Waiting Days: 2.32
```

Through descriptive statistics and visualization, the analysis will focus on identifying distribution patterns, relationships between variables, and potential anomalies in the data. These insights are critical for hotels in improving revenue management strategies, forecasting demand, reducing cancellations, and enhancing customer satisfaction.

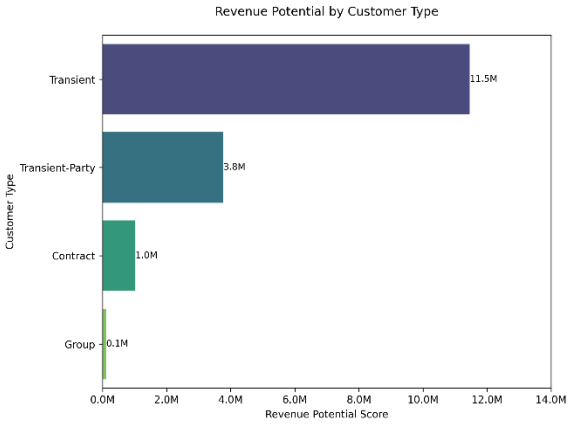


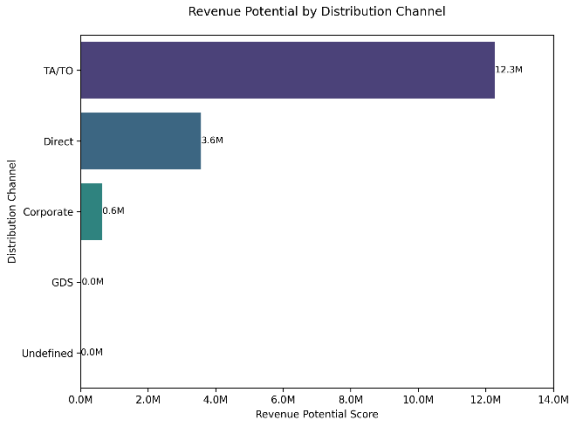

The chart compares hotel bookings across months for three years (2015 – 2017). Booking volumes show clear seasonality, with peaks between April and August and a decline toward the end of the year. Across all years, May registers the highest booking activity, while November and December remain the weakest months. The clustered bars make it easy to see that 2017 consistently outperformed 2016, with higher volumes early in the year before leveling off.

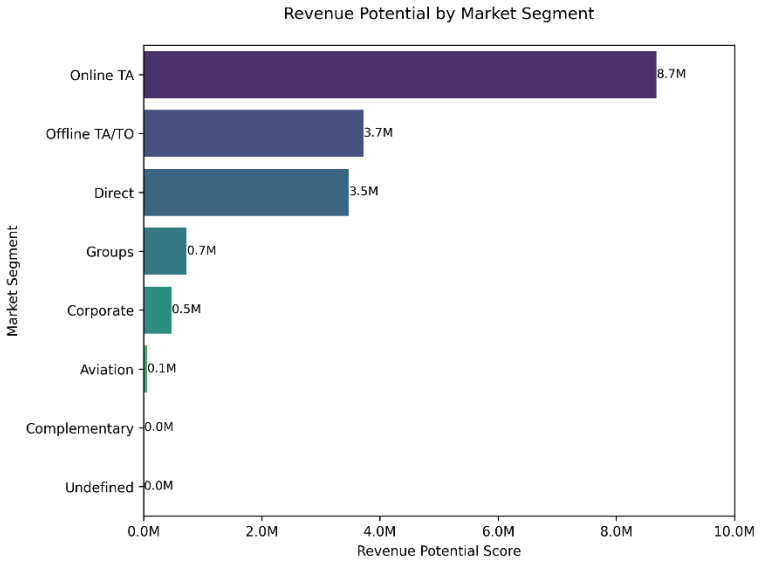


The chart compares hotel bookings by **day of the week** for City Hotels and Resort Hotels. City Hotels show a steady upward trend as the week progresses, peaking on **Friday** with the highest booking volume before dropping on Saturday. This pattern reflects strong demand from weekday business travelers, who typically check in during mid-to-late week. In contrast, Resort Hotels display a different pattern, with relatively balanced bookings throughout the week but noticeable increases on **weekends (Friday and Saturday)**, consistent with leisure and holiday travel behavior. The comparison highlights the contrasting booking dynamics: **City Hotels thrive on weekday demand**, while **Resort Hotels capture more weekend-driven leisure bookings**.

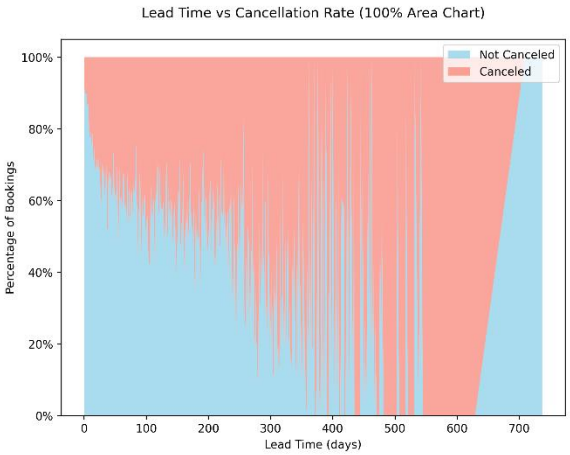
$$\boldsymbol{REVENUE\ POTENTIAL} = Average\ Revenue\ \times Booking\ Volume\ \times Cancellation\ Rate$$

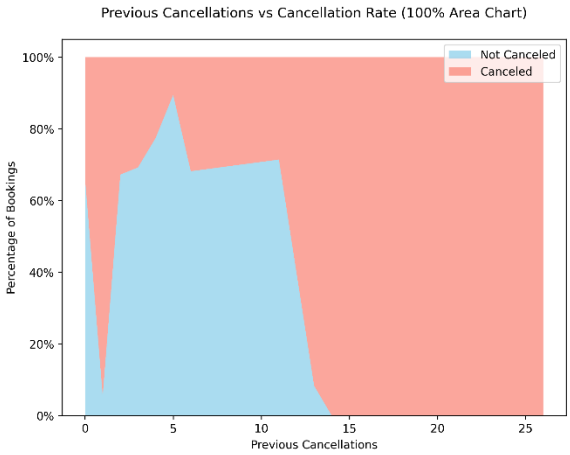


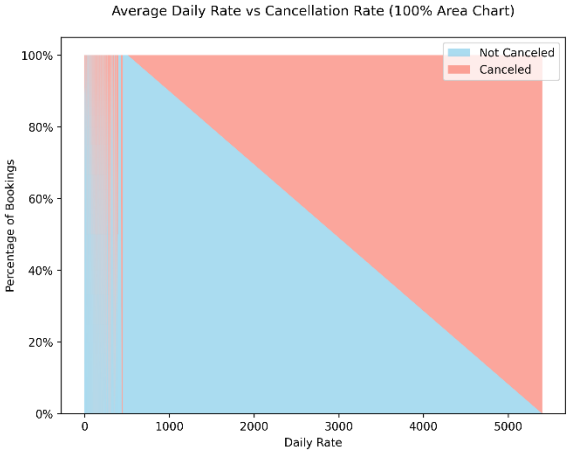



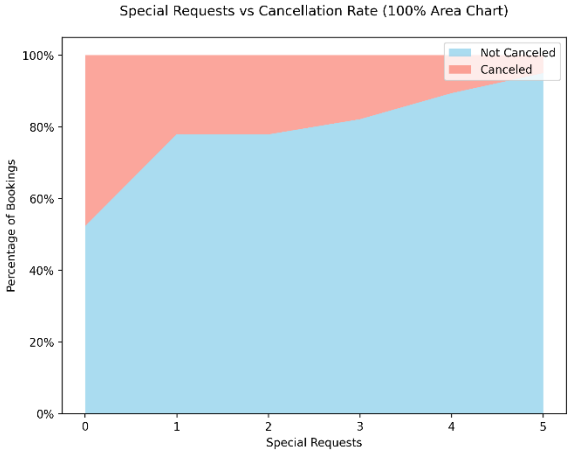



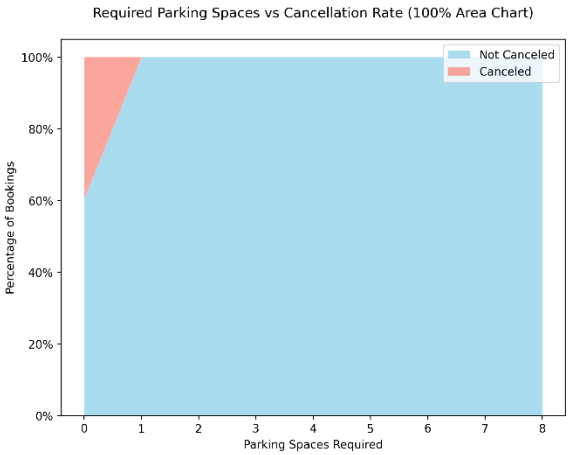
The analysis of revenue potential across customer type, market segment, and distribution channel reveals clear patterns in profitability. Among customer types, **Transient guests dominate with over 11M in revenue potential**, largely due to volume, while Contract customers contribute a smaller but steadier share, and Groups have minimal impact. Looking at market segments, **Online travel agents (OTAs) generate the highest revenue at 8.7M**, followed by offline travel agencies and direct bookings, with corporate and group bookings contributing much less. Finally, distribution channel analysis highlights the overwhelming importance of **TA/TO (12.3M)** as the primary revenue source, compared to the more limited contributions of Direct and Corporate channels. Together, these findings emphasize that while volume-driven segments like Transient guests and OTA channels deliver the bulk of revenue, Contract and Corporate channels—though smaller—offer stability and should not be overlooked in strategic planning.

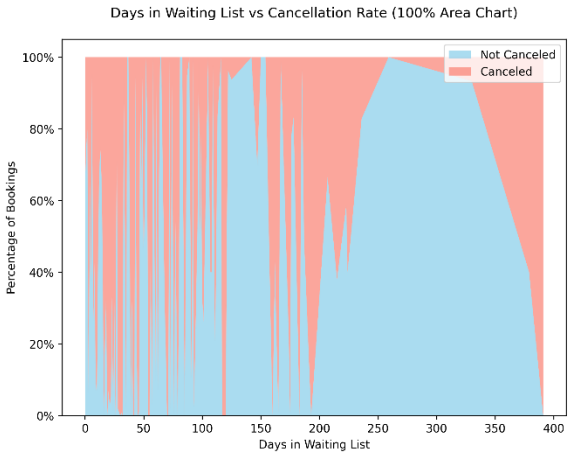


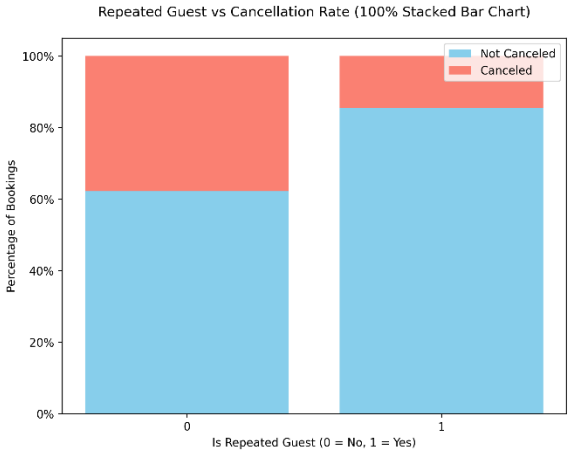






The findings from this EDA will also serve as a foundation for further statistical testing and predictive modeling, ultimately providing data-driven recommendations for operational and strategic decision-making.

## STATISTICAL ANALYSIS & MODELLING:

The cancellation analysis across different booking features reveals consistent behavioral patterns. **Longer lead times** are strongly associated with higher cancellation rates, confirming that early bookings are less reliable. Similarly, guests with **previous cancellations** show a much greater likelihood of canceling again, while **repeated guests** tend to be more loyal with fewer cancellations. Economic factors also matter: bookings with a **higher ADR** (average daily rate) show slightly higher cancellation risk, suggesting sensitivity to price. Operational variables, such as having a **parking space** or making **special requests**, are linked with lower cancellation rates, as these bookings reflect higher customer commitment.

Meanwhile, extended **waiting list durations** are associated with instability and greater risk of cancellation. Overall, the charts highlight that both guest history and booking conditions play a significant role in predicting cancellations, with lead time and prior cancellations being the most decisive indicators.

## EQUATIONS USED:

$$Probability\ of\ Cancellation = f(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \times 100\%$$

$$Odds - Ratio = g(x) = e^{\beta_1 x}$$

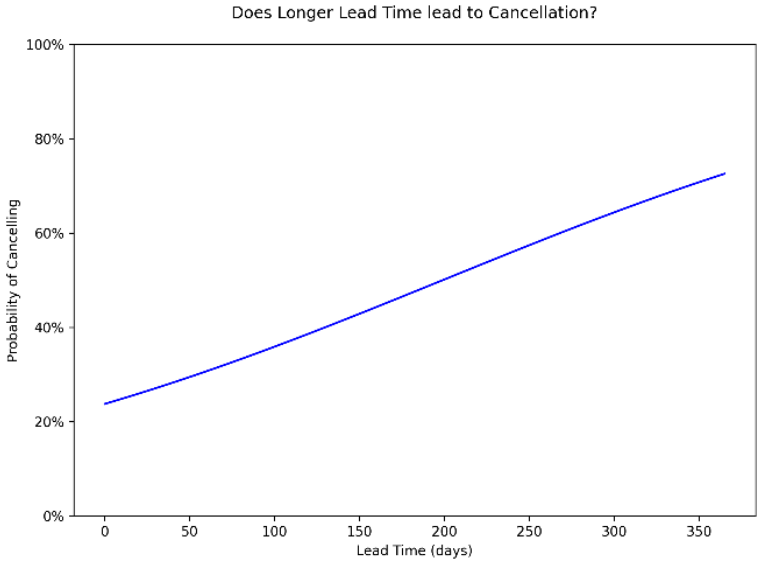
```
Optimization terminated successfully.
         Current function value: 0.616290
         Iterations 5
                           Logit Regression Results
==============================================================================
Dep. Variable:            is_canceled   No. Observations:               119390
Model:                          Logit   Df Residuals:                   119388
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Sep 2025   Pseudo R-squ.:                0.06506
Time:                        11:38:26   Log-Likelihood:                -73579.
converged:                       True   LL-Null:                       -78699.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -1.1656      0.009   -126.940      0.000      -1.184      -1.148
lead_time      0.0059   6.14e-05     95.404      0.000       0.006       0.006
==============================================================================
```

B0

B1

Having a p-value of 0.000, there is a significant relationship between Booking Lead Time and the likelihood of Cancellation!

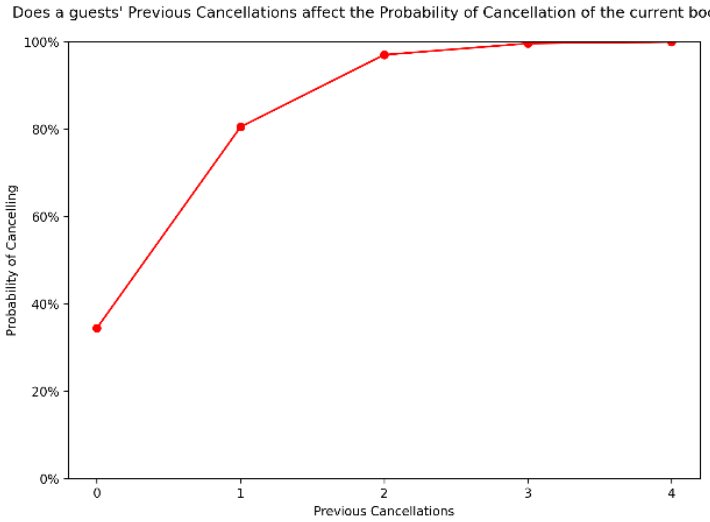
The probability of cancellation for a lead time of zero days is 23.77% while for every additional day, the odds of cancellation increases by 1.01 times.



```
Optimization terminated successfully.
         Current function value: 0.634290
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:            is_canceled   No. Observations:               119390
Model:                          Logit   Df Residuals:                   119388
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Sep 2025   Pseudo R-squ.:                0.03775
Time:                        11:38:49   Log-Likelihood:                -75728.
converged:                       True   LL-Null:                       -78699.
Covariance Type:            nonrobust   LLR p-value:                     0.000
================================================================================================
                                   coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                           -0.6423      0.006   -102.578      0.000      -0.655      -0.630
previous_cancellations           2.0608      0.033     62.875      0.000       1.997       2.125
================================================================================================
```

Having a p-value of 0.000, there is a significant relationship between guest's Previous Cancellation and the likelihood of Cancellation!

The probability of cancellation with zero previous cancellations is 34.47% while for every extra previous cancellation multiplies the odds by 7.85 times.

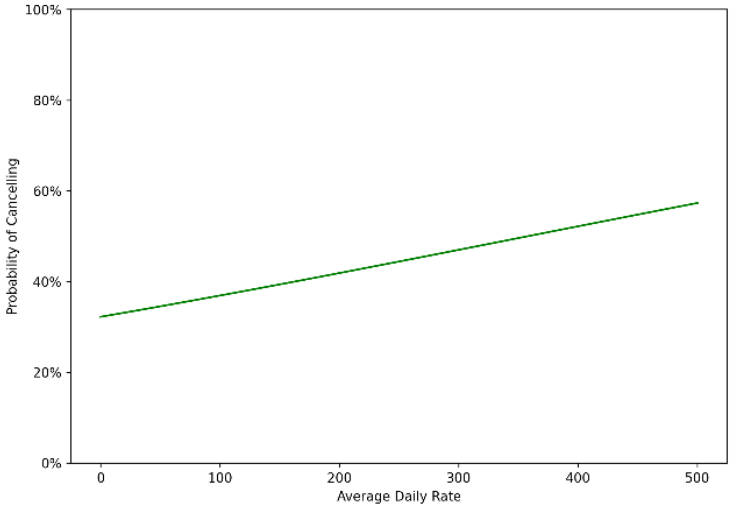



```
Optimization terminated successfully.
         Current function value: 0.657990
         Iterations 4
                           Logit Regression Results
==============================================================================
Dep. Variable:            is_canceled   No. Observations:               119390
Model:                          Logit   Df Residuals:                   119388
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Sep 2025   Pseudo R-squ.:                0.001800
Time:                        11:39:06   Log-Likelihood:                -78557.
converged:                       True   LL-Null:                       -78699.
Covariance Type:            nonrobust   LLR p-value:                 1.424e-63
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.7429      0.014    -52.829      0.000      -0.770      -0.715
adr            0.0021      0.000     16.796      0.000       0.002       0.002
==============================================================================
```

Having a p-value of 0.000, there is a significant relationship between the room's Average Daily Rate and the likelihood of Cancellation!

The probability of cancellation at ~0 daily rate is 32.24% while for every 100 unit increase multiplies the odds by 100.21 times.

What effect does increasing the daily rate of booking have on the probability of cancelling?

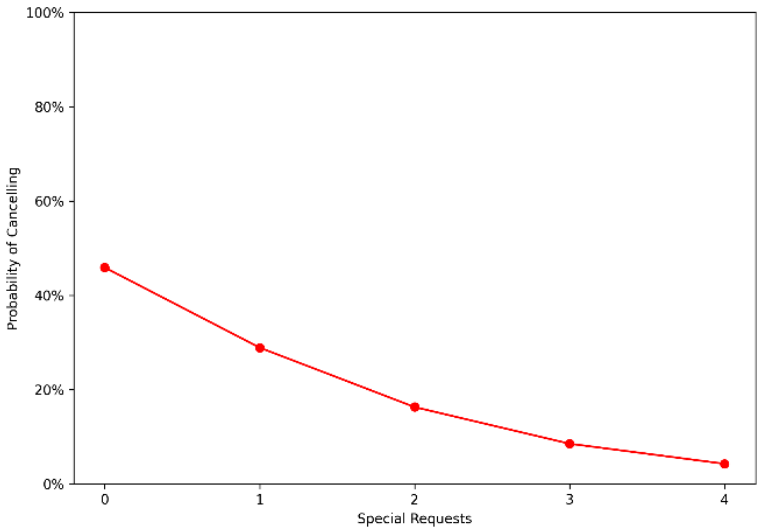



```
Optimization terminated successfully.
         Current function value: 0.628594
         Iterations 5
                           Logit Regression Results
==============================================================================
Dep. Variable:            is_canceled   No. Observations:               119390
Model:                          Logit   Df Residuals:                   119388
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Sep 2025   Pseudo R-squ.:                0.04640
Time:                        11:39:22   Log-Likelihood:                -75048.
converged:                       True   LL-Null:                       -78699.
Covariance Type:            nonrobust   LLR p-value:                     0.000
================================================================================================
                                   coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                          -0.1646      0.007    -22.471      0.000      -0.179      -0.150
total_of_special_requests      -0.7351      0.009    -78.121      0.000      -0.754      -0.717
================================================================================================
```

Having a p-value of 0.000, there is a significant relationship between guests' Total Special Requests and the likelihood of Cancellation!

The probability of cancellation with zero special requests is 45.89% while for every additional request multiplies the odds by 0.48 times.

Does multiple special requests decreases the likelihood of cancelling?



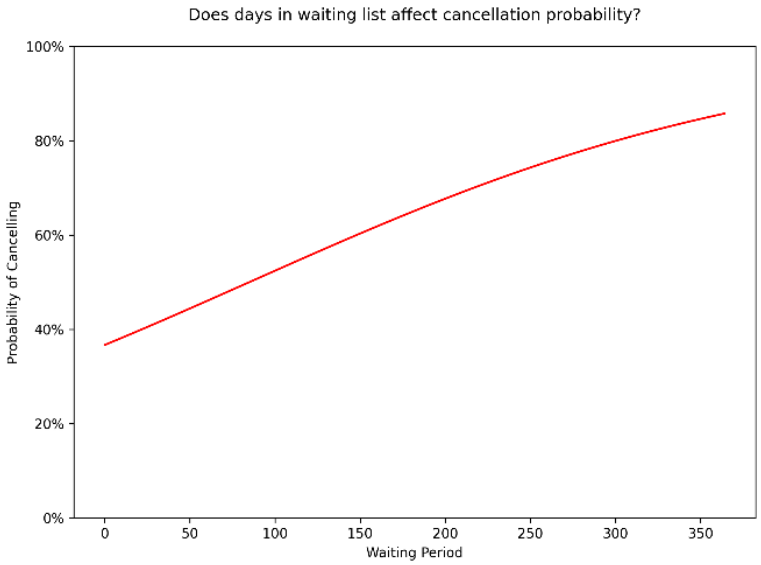
```
Warning: Maximum number of iterations has been exceeded.
         Current function value: 0.629236
         Iterations: 35
                           Logit Regression Results
==============================================================================
Dep. Variable:            is_canceled   No. Observations:               119390
Model:                          Logit   Df Residuals:                   119388
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Sep 2025   Pseudo R-squ.:                0.04542
Time:                        11:39:41   Log-Likelihood:                -75125.
converged:                      False   LL-Null:                       -78699.
Covariance Type:            nonrobust   LLR p-value:                     0.000
================================================================================================
                                   coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                           -0.4266      0.006    -69.775      0.000      -0.439      -0.415
required_car_parking_spaces    -24.5542   3093.019     -0.008      0.994   -6086.759    6037.651
================================================================================================
```

There is no significant relationship!

```
Optimization terminated successfully.
         Current function value: 0.657742
         Iterations 4
                           Logit Regression Results
==============================================================================
Dep. Variable:            is_canceled   No. Observations:               119390
Model:                          Logit   Df Residuals:                   119388
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Sep 2025   Pseudo R-squ.:                0.002176
Time:                        13:31:55   Log-Likelihood:                -78528.
converged:                       True   LL-Null:                       -78699.
Covariance Type:            nonrobust   LLR p-value:                 1.815e-76
============================================================================================
                               coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                   -0.5457      0.006    -90.052      0.000      -0.558      -0.534
days_in_waiting_list     0.0064      0.000     17.356      0.000       0.006       0.007
============================================================================================
```

Having a p-value of 0.000, there is a significant relationship between how long the guest was in the waiting list and the likelihood of Cancellation!

The probability of cancellation with zero wait is 36.69% while for every extra day multiplies the odds by 1.01 times.



```
Optimization terminated successfully.
         Current function value: 0.655047
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:            is_canceled   No. Observations:               119390
Model:                          Logit   Df Residuals:                   119388
Method:                           MLE   Df Model:                            1
Date:                Fri, 26 Sep 2025   Pseudo R-squ.:                0.006265
Time:                        19:44:42   Log-Likelihood:                -78206.
converged:                       True   LL-Null:                       -78699.
Covariance Type:            nonrobust   LLR p-value:                1.925e-216
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                -0.4987      0.006    -82.200      0.000      -0.511      -0.487
is_repeated_guest    -1.2766      0.046    -27.499      0.000      -1.368      -1.186
=====================================================================================
```

Having a p-value of 0.000, there is a significant relationship between Repeating Guests and the likelihood of Cancellation!

The probability of cancellation for a non-repeat guest is 37.79% while for a repeat guest it is only 14.49%.