Reducing Hotel Booking Cancellations Using Six Sigma Methodology

A PROJECT REPORT SUBMITTED BY
Group 1

in partial fulfillment of the requirement for the course of
APS4742 - Statistical Applications in Industry and Project Presentation
to the
Department of Environmental and Industrial Sciences
of the
FACULTY OF SCIENCE
UNIVERSITY OF PERADENIYA
SRI LANKA
2025

1. Introduction

In the competitive hospitality industry, hotel booking cancellations are a persistent problem that directly affects revenue, resource allocation, and operational efficiency. Cancellations particularly those made close to the check-in date, lead to empty rooms that are difficult to rebook on short notice, causing lost revenue and inefficiencies in staffing and service planning.

The primary objective of this study is to investigate the key drivers of booking cancellations using a data-driven approach and to provide actionable strategies for reducing the cancellation rate. This analysis leverages the Six Sigma methodology, specifically the DMAIC framework, to structure the investigation, identify root causes, and propose sustainable improvements. A dataset containing over 119,000 hotel bookings is utilized, from which a representative sample of 10,000 records was selected for analysis.

The goal of this project is to reduce the cancellation rate by at least 10%, improve booking reliability, and enhance overall hotel performance through better planning and predictive insights.

2. Methodology

2.1 Understanding the Dataset

The dataset used in this project contains information on 119,390 hotel bookings from both city hotels and resort hotels. It includes 32 variables covering various aspects such as booking status, customer demographics, lead time, length of stay, room types, special requests, deposit types, and booking channels.

For the purpose of efficient analysis under the Six Sigma framework, a stratified random sample of 10,000 records was extracted using R's caret::createDataPartition() function. Stratification was based on the hotel type to ensure proportional representation of both City Hotel and Resort Hotel categories.

2.2 Data Preparation and Preprocessing Steps

- Data Cleaning: Removed duplicates, handled missing values, and standardized formats.
- Variable Selection: Chose key variables for cancellation analysis, such as lead_time, is_canceled, ADR, customer_type, and market_segment.
- Sampling: Applied a sample size determination formula using 95% confidence level and 1% margin of error. This justified the use of 10,000 records for reliable and stable trend analysis.
- Segmentation: Grouped bookings based on hotel type and market segment for targeted analysis.

2.3 Application of Six Sigma DMAIC Framework

The Six Sigma DMAIC methodology was applied as follows:

- Define: Identified the problem (high cancellation rate, especially in City Hotels) and set the project goal (reduce cancellations by 10%). CTQ: Reliable confirmed bookings.
- Measure: Measured key metrics such as cancellation rate (36.84%), average lead time (103 days), ADR, and distribution by customer and market segments.
- Analyze: Used boxplots and Pareto charts to identify that longer lead times and certain booking segments (e.g., OTAs, Groups) were strongly associated with higher cancellations.
- Improve: Proposed solutions included early payment discounts, stricter cancellation policies for OTAs, and targeted retention strategies for transient guests.
- Control: Designed a control plan using P-charts and monthly KPI tracking tools like Excel dashboards or R Shiny apps to monitor and sustain improvements.

3. Results and Discussion

3.1 Descriptive Summary of the Sample

The following table summarizes key statistics calculated from the sample of 10,000 hotel bookings:

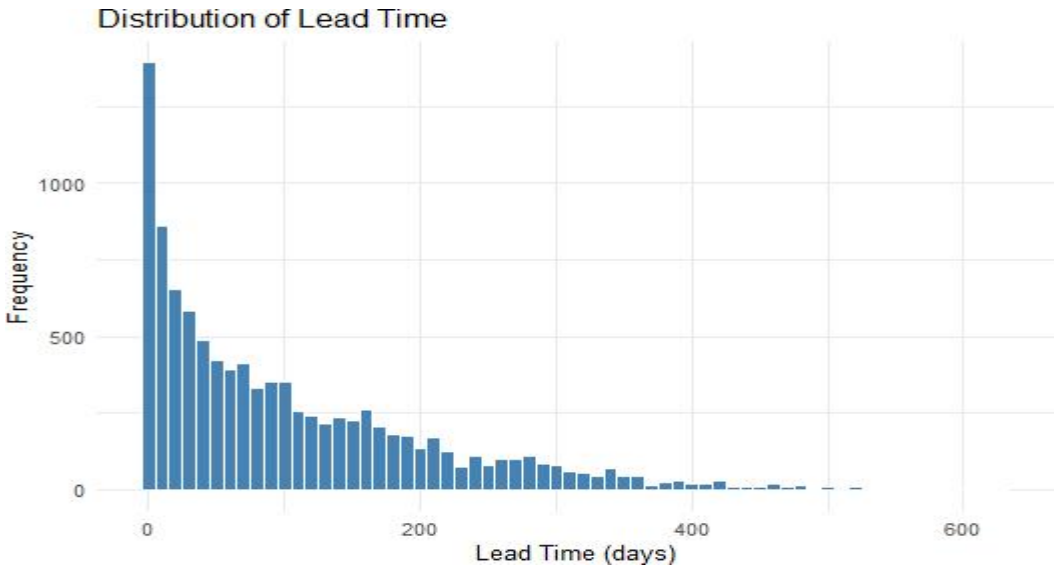
```
> summary_stats
  total_bookings cancellation_rate avg_lead_time median_lead_time  avg_adr median_adr repeated_guest_ratio avg_special_requests
1          10000            0.3684      103.4979               70 101.8764         95              0.0303               0.5618
>
```

These figures provide a baseline for analysis and reinforce that cancellations are a significant concern, with over one-third of bookings not resulting in actual stays. The low repeated guest ratio and minimal special requests suggest a customer base with limited loyalty or personalization.

3.2 Cancellation Rate by Hotel Type

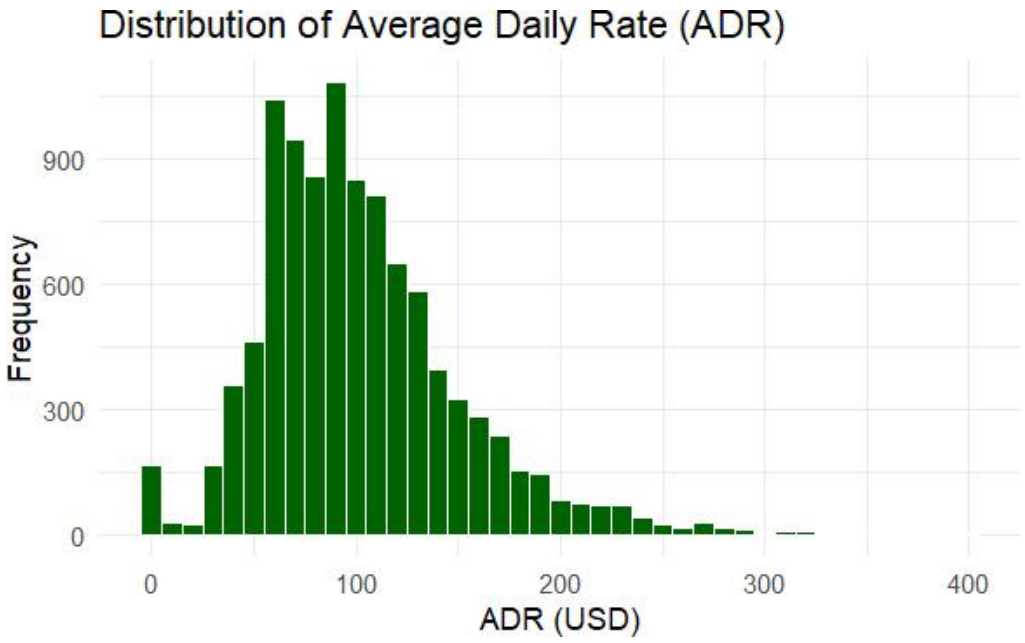
The bar chart comparing cancellation rates by hotel type shows a significantly higher cancellation rate in City Hotels (over 40%) compared to Resort Hotels (approximately 27%). This indicates that City Hotels are more prone to cancellations, likely due to more flexible booking channels or business travelers with frequently changing plans.

3.3 Distribution of Lead Time



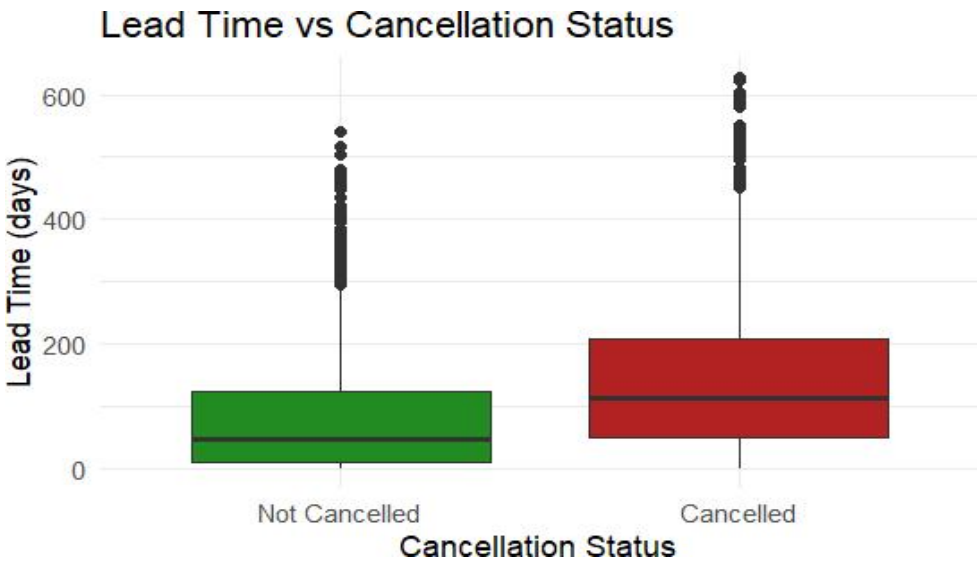
The histogram of lead time reveals that a majority of bookings are made within the first 0–50 days, but some bookings occur as early as 500+ days in advance. This distribution is heavily right-skewed, and longer lead times are associated with higher cancellation risks, as confirmed in later analysis.

3.4 Distribution of Average Daily Rate (ADR)



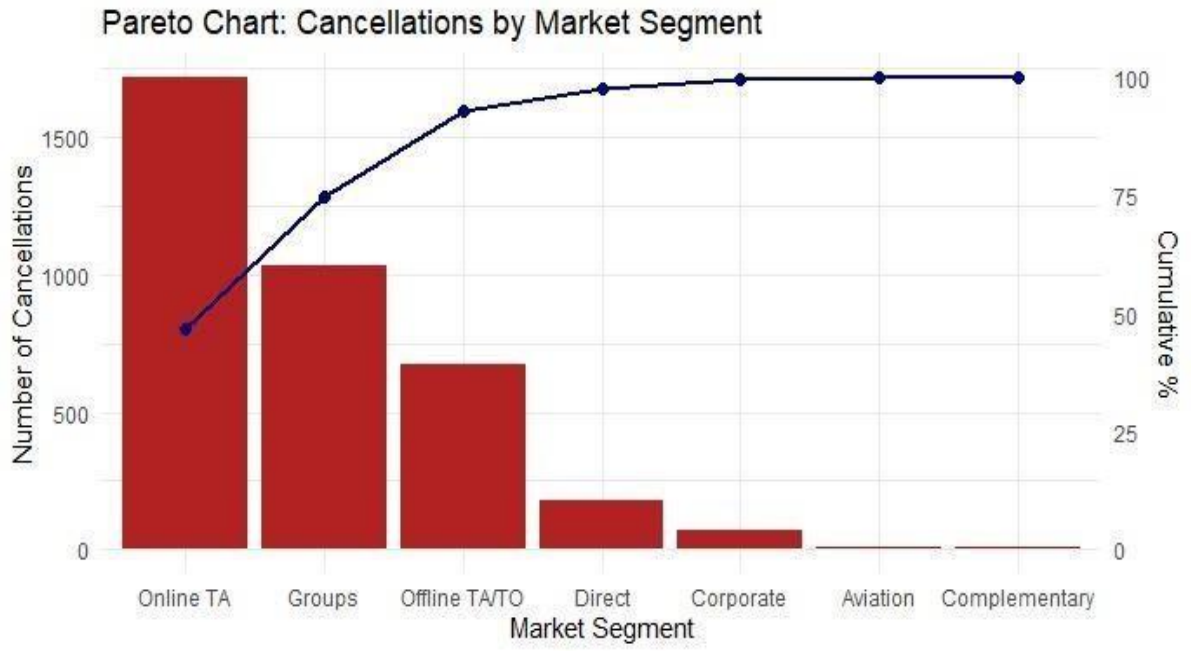
The distribution of the Average Daily Rate (ADR) shows a peak around USD 90–100, with most values falling below USD 150. A few outliers exist above USD 300. This suggests that pricing is clustered in the mid-range, which may correlate with guest expectations and booking behavior.

3.5 Lead Time vs Cancellation Status



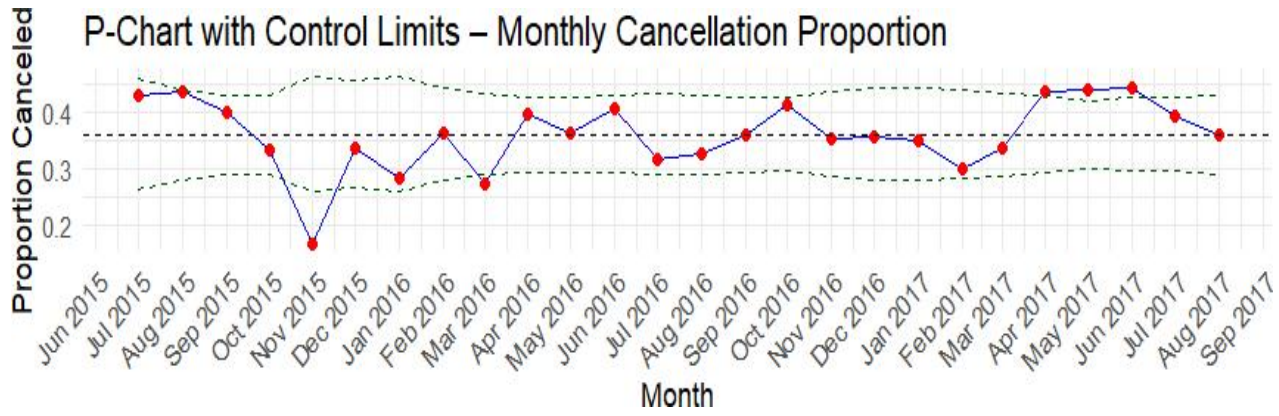
The boxplot comparing lead time by cancellation status indicates that canceled bookings have a higher median lead time and more variability. Many outliers also exist in the canceled group. This suggests that guests who book earlier are more likely to cancel possibly due to changes in plans or lower booking commitment.

3.6 Cancellation by Market Segment (Pareto Analysis)



The Pareto chart shows that the majority of cancellations are driven by a few segments: Online Travel Agents, Groups, and Offline Travel Agents/Tour Operators. These three segments together contribute to over 80% of all cancellations. In contrast, direct and corporate bookings have much lower cancellation rates. This confirms the 80/20 rule and suggests that targeted interventions on these top segments can yield the highest impact.

3.6 Monthly Cancellation Trends (P-Chart)



The P-chart displays monthly cancellation proportions along with control limits. Most of the monthly cancellation rates remain within control limits, indicating a stable process overall. However, there are a few points near the upper control limit (e.g., July–August 2015, May–June 2017), which may require further investigation or indicate seasonal effects. This chart can be used as a control tool for future process monitoring.

3.7 Key Insights Summary

- The overall cancellation rate is high at approximately 36.84%, with City Hotels experiencing a significantly higher rate than Resort Hotels.
- Bookings with longer lead times are more likely to be canceled, suggesting that early planners are less committed or more likely to change their plans.
- Guests booking through Online Travel Agents (OTAs), Groups, and Offline Travel Agents contribute to over 80% of total cancellations, supporting the Pareto Principle.
- Direct and corporate bookings show lower cancellation rates, indicating stronger customer commitment through these channels.
- The average booking price (ADR) clusters between US 80 and USD 150. Bookings with higher ADRs tend to have slightly more variability in cancellation.
- Monthly cancellation behavior is mostly stable, with a few seasonal or irregular spikes, indicating the need for routine monitoring using control charts.

4. Conclusion and Recommendations

4.1 Conclusion

The Six Sigma DMAIC methodology successfully identified the key drivers contributing to the high rate of hotel booking cancellations. Through a structured, data-driven approach, actionable insights were derived using statistical summaries and visual analytics.

The primary goal of the project was to reduce the cancellation rate by 10%, and this goal is supported by the patterns discovered in customer behavior, lead times, booking channels, and market segments. The findings provide a strong foundation for making informed decisions and implementing impactful interventions.

4.2 Limitations

- The dataset does not include customer satisfaction or refund information, which limits the understanding of customer intent and service recovery.
- The study analyzes associations but does not establish causal relationships between variables and cancellations.
- The analysis is based on a single sample in time; future patterns may differ due to seasonality, policy changes, or market shifts.

4.3 Recommendations

- Apply stricter cancellation policies specifically targeting Online Travel Agent (OTA) bookings, which have the highest cancellation rates.
- Introduce early payment incentives or non-refundable deposit schemes for bookings with long lead times to reduce cancellation probability.
- Monitor key performance indicators (KPIs) such as monthly cancellation rates and segment trends using automated dashboards (e.g., Excel or R Shiny) to ensure continuous improvement and long-term control.

5. References

[1] R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.r-project.org/

[2] Montgomery, D. C. (2009). *Introduction to Statistical Quality Control* (6th ed.). John Wiley & Sons.

[3] Batista, F., & Moniz, H. (2019). Predicting booking cancellations in hotels. *Tourism Management*, 75, 145–155. https://doi.org/10.1016/j.tourman.2019.05.005

6. Appendices

6.1 Data Set (Link)

https://www.kaggle.com/datasets/muhammaddawood42/hotel-booking-cancelations

6.2 R Codes

Import Libraries

```
library(caret)
library(dplyr)
library(ggplot2)
library(forcats)
```

Select a Data Sample

```
# Create stratified sample (10,000 rows based on hotel)
set.seed(123)
sample_index <- createDataPartition(cleaned_data$hotel, p = 10000 / nrow(cleaned_data), list = FALSE)
hotel_sample <- cleaned_data[sample_index, ]

# Save the stratified sample to CSV
write_csv(hotel_sample[1:10000,], "sample_data.csv")

# Load data
data <- read.csv(file.choose())
```

Data Analysis

```
# Summary Statistics for Key Variables
summary_stats <- data %>%
  summarise(
    total_bookings = n(),
    cancellation_rate = mean(is_canceled),
    avg_lead_time = mean(lead_time),
    median_lead_time = median(lead_time),
    avg_adr = mean(adr),
    median_adr = median(adr),
    repeated_guest_ratio = mean(is_repeated_guest),
    avg_special_requests = mean(total_of_special_requests)
  )
summary_stats
```

```r
# Bar Chart
data %>%
  group_by(hotel) %>%
  summarise(cancellation_rate = mean(is_canceled)) %>%
  ggplot(aes(x = hotel, y = cancellation_rate, fill = hotel)) +
  geom_col() +
  labs(title = "Cancellation Rate by Hotel Type", y = "Cancellation Rate", x = "Hotel Type") +
  theme_minimal()

# Histogram for Lead Time
ggplot(data, aes(x = lead_time)) +
  geom_histogram(binwidth = 10, fill = "steelblue", color = "white") +
  labs(title = "Distribution of Lead Time", x = "Lead Time (days)", y = "Frequency") +
  theme_minimal()

# Histogram for ADR
ggplot(data, aes(x = adr)) +
  geom_histogram(binwidth = 10, fill = "darkgreen", color = "white") +
  labs(title = "Distribution of Average Daily Rate (ADR)", x = "ADR (USD)", y = "Frequency") +
  theme_minimal()
```

```r
#Box Plot
ggplot(data, aes(x = as.factor(is_canceled), y = lead_time, fill = as.factor(is_canceled))) +
  geom_boxplot() +
  labs(title = "Lead Time Distribution by Cancellation Status",
       x = "Is Canceled (0 = No, 1 = Yes)",
       y = "Lead Time (Days)",
       fill = "Canceled") +
  scale_fill_manual(values = c("0" = "green", "1" = "red")) +
  theme_minimal()

# Prepare data for Pareto Chart: cancellations by market segment
cancellation_pareto <- hotel_sample %>%
  filter(is_canceled == 1) %>%
  count(market_segment, sort = TRUE) %>%
  mutate(percent = 100 * n / sum(n),
         cum_percent = cumsum(percent))

# Pareto chart
ggplot(cancellation_pareto, aes(x = fct_reorder(market_segment, -n), y = n)) +
  geom_col(fill = "firebrick") +
  geom_line(aes(y = cum_percent * max(n) / 100, group = 1), color = "darkblue", size = 1) +
  geom_point(aes(y = cum_percent * max(n) / 100), color = "darkblue", size = 2) +
  scale_y_continuous(sec.axis = sec_axis(~ . * 100 / max(cancellation_pareto$n), name = "Cumulative %")) +
  labs(title = "Pareto Chart: Cancellations by Market Segment",
       x = "Market Segment", y = "Number of Cancellations")
```

```r
#Add UCL and LCL (3-sigma control limits)
monthly_data <- monthly_data %>%
  mutate(
    UCL = p_bar + 3 * sqrt((p_bar * (1 - p_bar)) / total_bookings),
    LCL = p_bar - 3 * sqrt((p_bar * (1 - p_bar)) / total_bookings),
    LCL = ifelse(LCL < 0, 0, LCL)  # LCL can't be negative
  )

#PLot P-Chart
ggplot(monthly_data, aes(x = date_label, y = proportion_canceled, group = 1)) +
  geom_line(color = "blue") +
  geom_point(color = "red") +
  geom_line(aes(y = UCL), linetype = "dashed", color = "darkgreen") +
  geom_line(aes(y = LCL), linetype = "dashed", color = "darkgreen") +
  geom_hline(yintercept = p_bar, linetype = "dashed", color = "black") +
  labs(
    title = "P-Chart with Control Limits - Monthly Cancellation Proportion",
    x = "Month", y = "Proportion Canceled"
  ) +
  scale_x_date(date_labels = "%b %Y", date_breaks = "1 month") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```