Albukhary International University

SCHOOL OF COMPUTING AND INFORMATICS
Bachelor of Computer Science (Hons)

INDIVIDUAL ASSIGNMENT (20%)
CCS2233
Statistical Programming

City Hotel Reservation Status Analysis

Student Name : Roua Alimam Student_ID: AIU22102291

For Examiner Use Only

Total Marks (40 marks)

Table of Contents

Abstract

This report presents analysis of hotel booking data focusing on how independent variables like: lead time, adr (Average daily price rate for rooms), previous cancelation, deposit type, customer type, and room change requests could potentially affect the reservation status (Canceled, Check-out, No-show). Studying how these independent variables effect the reservation status can significantly positively impact the operational efficiency of the hotels.

1.0 Introduction

Hotel booking cancellation can have a significant impacts on the hospitality industry and the operational efficiency of the hotels. Understanding the factors that contribute to higher cancellation rates is crucial to minimize the negative impacts. In this study we will be analyzing observations collected from hotel to discover the reason behind hotel cancellations and prevent it.

1.1 Dataset

The data set contains 119390 observations collected from two types of hotels (City Hotel & Resort Hotel). Each observation represents a hotel booking over a period spanning from July 2015 till August 2017, including booking that effectively arrived and booking that were canceled.

columns:

Categorical Data:

1. **hotel**: Type of hotel (City Hotel or Resort Hotel)
2. **is_canceled**: Whether the booking was canceled (1) or not (0)
3. **arrival_date_year**: Year of arrival date
4. **arrival_date_month**: Month of arrival date (January to December)
5. **meal**: Type of meal plan
6. **country**: Country of origin of the guest
7. **market_segment**: Market segment designation (e.g., TA – Travel Agents, TO – Tour Operators)
8. **distribution_channel**: Booking distribution channel (e.g., TA – Travel Agents, TO – Tour Operators)
9. **is_repeated_guest**: Whether the guest is a repeated guest (1) or not (0)
10. **reserved_room_type**: Code for the room type reserved (anonymized)
11. **assigned_room_type**: Code for the room type assigned (anonymized)
12. **deposit_type**: Type of deposit (No Deposit, Non-Refund, Refundable)
13. **customer_type**: Type of customer (Group, Transient, Transient-party)
14. **reservation_status**: Status of the reservation (Check-Out, No-Show)
15. **reservation_status_date**: Date when the last status was set

Numerical Data:

1. **lead_time**: Number of days between booking and arrival
2. **arrival_date_week_number**: Week number of the arrival date

3. **arrival_date_day_of_month**: Day of the month of the arrival date
4. **stays_in_weekend_nights**: Number of weekend nights (Saturday or Sunday)
5. **stays_in_week_nights**: Number of weeknights (Monday to Friday)
6. **adults**: Number of adults in the booking
7. **children**: Number of children in the booking
8. **babies**: Number of babies in the booking
9. **previous_cancellations**: Number of previous bookings canceled by the guest
10. **previous_bookings_not_canceled**: Number of previous bookings not canceled by the guest
11. **booking_changes**: Number of changes or amendments made to the booking
12. **days_in_waiting_list**: Number of days the booking was on the waiting list before confirmation
13. **adr**: Average Daily Rate (calculated by dividing the total lodging transactions by the number of staying nights)
14. **required_car_parking_spaces**: Number of car parking spaces required by the guest
15. **total_of_special_requests**: Number of special requests made by the guest (e.g., twin bed, high floor)

1.2 Study Questions

What are the key factors influencing hotel booking cancellations, and how can hotels mitigate these cancellations to improve their operational efficiency and revenue management?

2.0 Data Validation

2.1 Data Loading

```
data <- read.csv("hotel_bookings.csv")

print(data)
```

2.2 Data Summary

```
summary(data)
```

```
     index            hotel            is_canceled       lead_time    arrival_date_year arrival_date_month
 Min.   :     0   Length:119390      Min.   :0.0000   Min.   :  0   Min.   :2015      Length:119390
 1st Qu.: 29847   Class :character   1st Qu.:0.0000   1st Qu.: 18   1st Qu.:2016      Class :character
 Median : 59695   Mode  :character   Median :0.0000   Median : 69   Median :2016      Mode  :character
 Mean   : 59695                      Mean   :0.3704   Mean   :104   Mean   :2016
 3rd Qu.: 89542                      3rd Qu.:1.0000   3rd Qu.:160   3rd Qu.:2017
 Max.   :119389                      Max.   :1.0000   Max.   :737   Max.   :2017

 arrival_date_week_number arrival_date_day_of_month stays_in_weekend_nights stays_in_week_nights
 Min.   : 1.00            Min.   : 1.0              Min.   : 0.0000         Min.   : 0.0
 1st Qu.:16.00            1st Qu.: 8.0              1st Qu.: 0.0000         1st Qu.: 1.0
 Median :28.00            Median :16.0              Median : 1.0000         Median : 2.0
 Mean   :27.17            Mean   :15.8              Mean   : 0.9276         Mean   : 2.5
 3rd Qu.:38.00            3rd Qu.:23.0              3rd Qu.: 2.0000         3rd Qu.: 3.0
 Max.   :53.00            Max.   :31.0              Max.   :19.0000         Max.   :50.0

     adults          children           babies              meal             country
 Min.   : 0.000   Min.   : 0.0000   Min.   : 0.000000   Length:119390      Length:119390
 1st Qu.: 2.000   1st Qu.: 0.0000   1st Qu.: 0.000000   Class :character   Class :character
 Median : 2.000   Median : 0.0000   Median : 0.000000   Mode  :character   Mode  :character
 Mean   : 1.856   Mean   : 0.1039   Mean   : 0.007949
 3rd Qu.: 2.000   3rd Qu.: 0.0000   3rd Qu.: 0.000000
 Max.   :55.000   Max.   :10.0000   Max.   :10.000000
                  NA's   :4
 market_segment     distribution_channel is_repeated_guest previous_cancellations
 Length:119390      Length:119390        Min.   :0.00000   Min.   : 0.00000
 Class :character   Class :character     1st Qu.:0.00000   1st Qu.: 0.00000
 Mode  :character   Mode  :character     Median :0.00000   Median : 0.00000
                                         Mean   :0.03191   Mean   : 0.08712
                                         3rd Qu.:0.00000   3rd Qu.: 0.00000
                                         Max.   :1.00000   Max.   :26.00000

 previous_bookings_not_canceled reserved_room_type assigned_room_type booking_changes   deposit_type
 Min.   : 0.0000                Length:119390      Length:119390      Min.   : 0.0000   Length:119390
 1st Qu.: 0.0000                Class :character   Class :character   1st Qu.: 0.0000   Class :character
 Median : 0.0000                Mode  :character   Mode  :character   Median : 0.0000   Mode  :character
 Mean   : 0.1371                                                      Mean   : 0.2211
 3rd Qu.: 0.0000                                                      3rd Qu.: 0.0000
 Max.   :72.0000                                                      Max.   :21.0000

     agent            company        days_in_waiting_list customer_type           adr
 Min.   :  1.00   Min.   :  6.0    Min.   :  0.000      Length:119390      Min.   :  -6.38
 1st Qu.:  9.00   1st Qu.: 62.0    1st Qu.:  0.000      Class :character   1st Qu.:  69.29
 Median : 14.00   Median :179.0    Median :  0.000      Mode  :character   Median :  94.58
 Mean   : 86.69   Mean   :189.3    Mean   :  2.321                         Mean   : 101.83
 3rd Qu.:229.00   3rd Qu.:270.0    3rd Qu.:  0.000                         3rd Qu.: 126.00
 Max.   :535.00   Max.   :543.0    Max.   :391.000                         Max.   :5400.00
 NA's   :16340    NA's   :112593
 required_car_parking_spaces total_of_special_requests reservation_status reservation_status_date
 Min.   :0.00000             Min.   :0.0000            Length:119390      Length:119390
 1st Qu.:0.00000             1st Qu.:0.0000            Class :character   Class :character
 Median :0.00000             Median :0.0000            Mode  :character   Mode  :character
 Mean   :0.06252             Mean   :0.5714
```

Notes:

- The summary of is_canceled column mean is 0.3704, indicating a quite high calculation rate.
- Lead_time column range widly between 0 to 737, indicating the choice of guest booking early is widely varied.
- The demographic of the guest seems to be mostly adults. Families with children and babies are rare in this dataset.
- Several columns, such as company and agent have large amount of missing values, with the company column having over 112,000 NA values and the agent column over 16,000 NA values.

2.3 Data Profiling

```
library(DataExplorer)
create_report(data)
```

Link: file:///C:/Users/acer/Desktop/Fifth_Semester/Statistical_programming/Hotel_Booking/report.html

2.4 Analysis Specification

2.4.1 Removing Unnecessary columns

```
# Load dplyr library
library(dplyr)

data <- data %>% select(-index, -babies, -children, -arrival_date_year, -arrival_date_month, -agent,
-company, -arrival_date_week_number, -is_repeated_guest, -reservation_status_date, -
previous_bookings_not_canceled)
```

2.4.2 Focusing on City Hotels

```
data <- data %>% filter(hotel == "City Hotel")
```

2.5 Summary and Notes:

After checking the data summary and profiling several columns were dropped for several factors including: having high missing values or being insignificant to the reservation status.

Potentially Insignificant Columns:

These columns might not provide much useful information for the analysis, based on the summary:

1. **Babies**: this variable has very little variance and is unlikely to provide meaningful insights.
2. **Children**: Similar to babies, the number of children in bookings is quite low and might not be a significant factor.
3. **Arrival Date Year and Month**: These columns may not directly affect cancellation, as there isn't much variability across years or months unless analyzing seasonality in cancellations.
4. **Agent** and **Company**: Due to the large amount of missing data.

Potentially Significant Columns:

These columns are likely to be important for understanding the probability of cancellations:

1. **Lead Time**: Longer lead times may result in a higher likelihood of cancellation, as plans can change over time.
2. **Deposit Type**: Guests with no deposits or refundable deposits may be more likely to cancel.
3. **Previous Cancellations**: Guests with a history of cancellations are more likely to cancel again.
4. **Customer type**: Different market segments may show different cancellation behaviors.
5. **ADR (Average Daily Rate)**: The cost per night may affect cancellation rates, particularly for expensive bookings.
6. **Special Requests**: A high number of special requests may correlate with cancellations, especially if guests' needs aren't met.

3.0 Data cleaning and Preprocessing

3.1 checking Duplicates

```
sum(duplicated(data))
```

```
[1] 26184
```

The data set seem to have many duplicated observations. Further examination to the duplicated observation is required to determine whether the duplicates need to be removed or not.

```
#checking some duplicated rows
duplicates <- data[duplicated(data) | duplicated(data, fromLast = TRUE), ]
head(duplicates, 50)
```

	hotel <chr>	is_canceled <int>	lead_time <int>	arrival_date_day_of_month <int>	stays_in_weekend_nights <int>	stays_in_week_nights <int>	adults <int>	meal <chr>	country <chr>	market_segment <chr>
9	City Hotel	1	62	2	2	3	2	BB	PRT	Online TA
10	City Hotel	1	62	2	2	3	2	BB	PRT	Online TA
11	City Hotel	0	43	3	0	2	2	HB	PRT	Groups
13	City Hotel	0	43	3	0	2	2	HB	PRT	Groups
15	City Hotel	1	43	3	0	2	2	HB	PRT	Groups
17	City Hotel	1	43	3	0	2	1	HB	PRT	Groups
18	City Hotel	0	43	3	0	2	2	HB	PRT	Groups
19	City Hotel	0	43	3	0	2	2	HB	PRT	Groups
20	City Hotel	1	43	3	0	2	1	HB	PRT	Groups
21	City Hotel	1	43	3	0	2	2	HB	PRT	Groups
23	City Hotel	0	43	3	0	2	2	HB	PRT	Groups
24	City Hotel	0	43	3	0	2	2	HB	PRT	Groups
55	City Hotel	1	90	7	5	15	1	SC	PRT	Online TA
56	City Hotel	1	90	7	5	15	1	SC	PRT	Online TA
74	City Hotel	1	87	10	3	7	2	BB	PRT	Online TA
75	City Hotel	1	87	10	3	7	2	BB	PRT	Online TA

After checking some duplicates, they **seem to be mostly a result of data entry mistakes** and not identical bookings. Nevertheless, there exist some duplicates where the booking are not identical but they share similar entries. Thus, the duplicates will be removed.

```
clean_data <- data[!duplicated(data), ]
```

checking the dimensions of the dataset

```
dim(clean_data)
```

```
[1] 53146    22
```

3.2 Handling Missing Values

3.2 Checking for missing and undefined values

```
colSums(is.na(clean_data))
```

```
                  hotel                 is_canceled                    lead_time 
                      0                           0                            0 
   arrival_date_day_of_month     stays_in_weekend_nights         stays_in_week_nights 
                      0                           0                            0 
                 adults                        meal                      country 
                      0                           0                            0 
         market_segment        distribution_channel       previous_cancellations 
                      0                           0                            0 
     reserved_room_type          assigned_room_type              booking_changes 
                      0                           0                            0 
           deposit_type        days_in_waiting_list                customer_type 
                      0                           0                            0 
                    adr required_car_parking_spaces    total_of_special_requests 
                      0                           0                            0 
     reservation_status 
                      0 
```

There isn't any missing NA values in the dataset

```
# Count occurrences of 'Undefined' per column
sapply(clean_data, function(x) sum(x == 'Undefined'))
```

```
                  hotel                 is_canceled                    lead_time 
                      0                           0                            0 
   arrival_date_day_of_month     stays_in_weekend_nights         stays_in_week_nights 
                      0                           0                            0 
                 adults                        meal                      country 
                      0                           0                            0 
         market_segment        distribution_channel       previous_cancellations 
                      2                           4                            0 
     reserved_room_type          assigned_room_type              booking_changes 
                      0                           0                            0 
           deposit_type        days_in_waiting_list                customer_type 
                      0                           0                            0 
                    adr required_car_parking_spaces    total_of_special_requests 
                      0                           0                            0 
     reservation_status 
                      0 
```

There are some undefined values in some columns like market segment (2 undefined values), and distribution_channel (4 undefined values).

3.2.2 Handling Undefined values

```
# Create a function to calculate the mode
get_mode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}

# Calculate the mode of market_segment and distribution channel and replace 'Undefined' values
modeM <- get_mode(clean_data$market_segment)
clean_data$market_segment <- replace(clean_data$market_segment, clean_data$market_segment ==
"Undefined", modeM)
clean_data$market_segment <- factor(clean_data$market_segment)

modeD <- get_mode(clean_data$distribution_channel)
clean_data$distribution_channel <- replace(clean_data$distribution_channel,
clean_data$distribution_channel == "Undefined", modeD)
clean_data$distribution_channel <- factor(clean_data$distribution_channel)
```

3.3 Feature Engineering and data encoding

```r
#feature engineering

clean_data <- clean_data %>% mutate(total_stay_duration = stays_in_weekend_nights +
stays_in_week_nights)

clean_data <- clean_data %>% mutate(adr_per_night = adr/ total_stay_duration)

clean_data <- clean_data %>% mutate(special_requests = ifelse(total_of_special_requests > 0, 1, 0))

clean_data <- clean_data %>% mutate(previous_cancelation = ifelse(previous_cancellations >0, 1, 0))

clean_data <- clean_data %>% mutate(room_changed = ifelse(reserved_room_type != assigned_room_type,
1, 0))

#Drop Unnecessary columsn to reduce dataset dimentionaltiy
clean_data <- clean_data %>%select(-stays_in_weekend_nights, -stays_in_week_nights, -
total_of_special_requests,-previous_cancellations, -reserved_room_type, -assigned_room_type, -
arrival_date_day_of_month, -adults, -meal, -required_car_parking_spaces, -country)


colnames(clean_data)

#Factor encoding for the rest of categorical data
# Convert categorical variables to factors
clean_data <- clean_data %>%
  mutate(
    market_segment = as.factor(market_segment),
    distribution_channel = as.factor(distribution_channel),
    customer_type = as.factor(customer_type),
    deposit_type = as.factor(deposit_type)
  )

# Check the structure of the updated dataset
Str(clean_data)
```

```
'data.frame':	53146 obs. of  21 variables:
 $ hotel                      : chr  "City Hotel" "City Hotel" "City Hotel" "City Hotel" ...
 $ is_canceled                : int  0 1 1 1 1 1 0 1 1 0 ...
 $ lead_time                  : int  6 88 65 92 100 79 3 63 62 43 ...
 $ arrival_date_day_of_month  : int  1 1 1 1 2 2 2 2 2 3 ...
 $ adults                     : int  1 2 1 2 2 2 1 1 2 2 ...
 $ meal                       : Factor w/ 4 levels "BB","FB","HB",..: 3 1 1 1 1 1 3 1 1 3 ...
 $ country                    : Factor w/ 167 levels "","ABW","AGO",..: 127 127 127 127 127 127 127
127 127 127 ...
 $ market_segment             : Factor w/ 7 levels "Aviation","Complementary",..: 6 7 7 7 7 7 5 7 7
5 ...
 $ distribution_channel       : Factor w/ 4 levels "Corporate","Direct",..: 4 4 4 4 4 4 4 4 4 4 ...
 $ booking_changes            : int  0 0 0 0 0 0 1 0 0 0 ...
 $ deposit_type               : Factor w/ 3 levels "No Deposit","Non Refund",..: 1 1 1 1 1 1 1 1 1 1
...
 $ days_in_waiting_list       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ customer_type              : Factor w/ 4 levels "Contract","Group",..: 3 3 3 3 3 3 4 3 3 4 ...
 $ adr                        : num  0 76.5 68 76.5 76.5 ...
 $ required_car_parking_spaces: int  0 0 0 0 0 0 0 0 0 0 ...
 $ reservation_status         : chr  "Check-Out" "Canceled" "Canceled" "Canceled" ...
 $ total_stay_duration        : int  2 4 4 6 2 3 3 4 5 2 ...
 $ adr_per_night              : num  0 19.1 17 12.8 38.2 ...
 $ special_requests           : num  0 1 1 1 1 1 0 0 1 0 ...
 $ previous_cancelation       : num  0 0 0 0 0 0 0 0 0 0 ...
 $ room_changed               : num  0 0 0 0 0 0 0 0 0 0 ...
```

3.4 Outlier Handling

```
{r}
boxplot(clean_data$adr, main= "Boxplot for adr", ylab= "Values", col="lightblue")

boxplot(clean_data$lead_time, main= "Boxplot for lead time", ylab= "Values", col="deepskyblue")

boxplot(clean_data$total_stay_duration, main= "Boxplot for total stay duration", ylab= "Values",
col="darkslategray3")
```

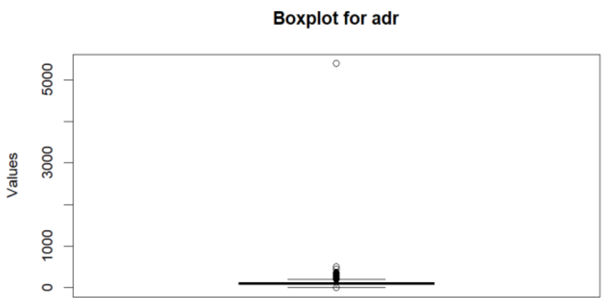

Figure 1a boxplot for adr before removing outliers

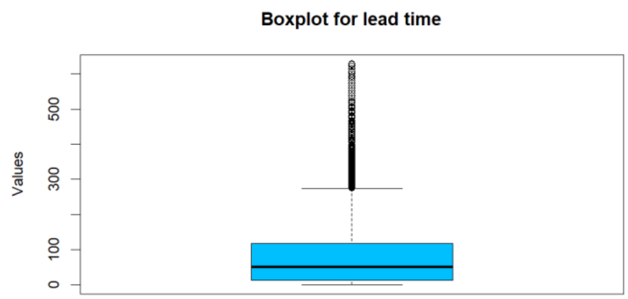

Figure 1b boxplot for lead time before removing outliers

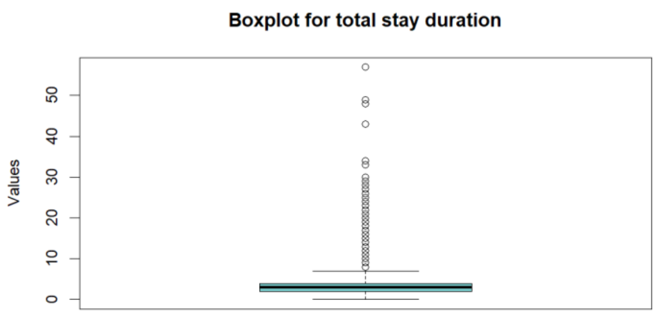

Figure 1c boxplot for total stay duration before removing outliers

There are many outliers in the data. Especially the adr that has a very extreme outlier. Thus, an outlier treatment and removal is necessary in this case.

```
# ADR
Q1 <- quantile(clean_data$adr, 0.25)
Q3 <- quantile(clean_data$adr, 0.75)
IQR_value <- Q3 - Q1

lower_bound <- Q1 - 1.5 * IQR_value
upper_bound <- Q3 + 1.5 * IQR_value


# lead_time
Q1_lead_time <- quantile(clean_data$lead_time, 0.25)
Q3_lead_time <- quantile(clean_data$lead_time, 0.75)
IQR_lead_time <- Q3_lead_time - Q1_lead_time

lower_bound_lead_time <- Q1_lead_time - 1.5 * IQR_lead_time
upper_bound_lead_time <- Q3_lead_time + 1.5 * IQR_lead_time
```

```
# total_stay_duration
Q1_total_stay <- quantile(clean_data$total_stay_duration, 0.25)
Q3_total_stay <- quantile(clean_data$total_stay_duration, 0.75)
IQR_total_stay <- Q3_total_stay - Q1_total_stay

lower_bound_total_stay <- Q1_total_stay - 1.5 * IQR_total_stay
upper_bound_total_stay <- Q3_total_stay + 1.5 * IQR_total_stay

# Filtering the dataset
cleaneast_data <- subset(clean_data, adr > lower_bound & adr < upper_bound &
lead_time > lower_bound_lead_time & lead_time < upper_bound_lead_time & total_stay_duration >
lower_bound_total_stay & total_stay_duration < upper_bound_total_stay)

boxplot(cleanest_data$adr, main= "Boxplot for adr", ylab= "Values", col="lightblue")

boxplot(cleanest_data$lead_time, main= "Boxplot for lead time", ylab= "Values", col="deepskyblue")

boxplot(cleanest_data$total_stay_duration, main= "Boxplot for total stay duration", ylab= "Values",
col="darkslategray3")
```

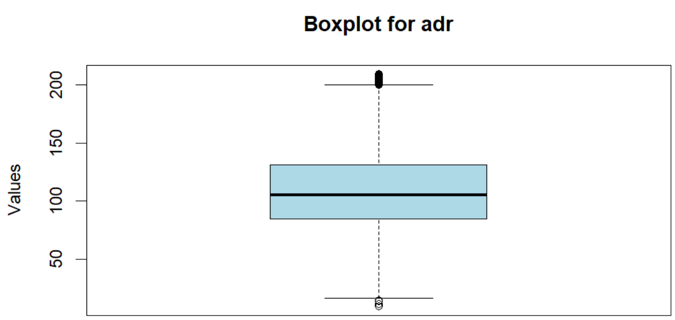

Figure 2a boxplot for adr after removing outliers

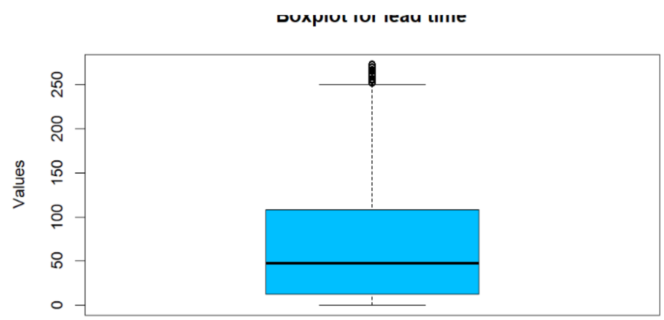

Figure 2b boxplot for lead time after removing outliers

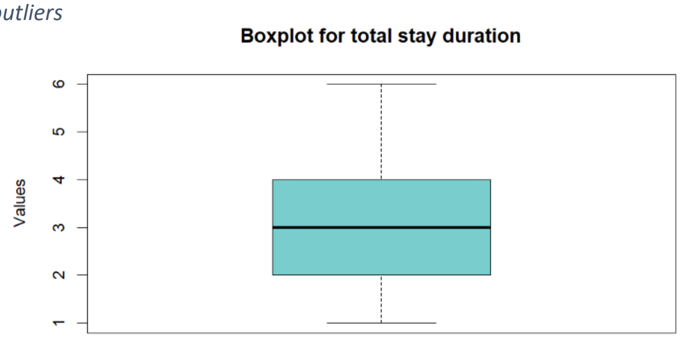

Figure 2c boxplot for total stay duration after removing outliers

These are the boxplots after outliers' removal. Since the columns contain some extreme outliers points, not all outliers were treated. Although the IQR multiplier could be reduced further to eliminate more outliers, but the decision not to remove all outliers was taken. This decision ensures that the dataset reflects the variability in booking behavior and some outliers are kept to ensure fairness in further analysis.

3.5 Data profiling again

Basic Statistics

Raw Counts

Name	Value
Rows	46,205
Columns	16
Discrete columns	6
Continuous columns	10
All missing columns	0
Missing observations	0
Complete Rows	46,205
Total observations	739,280
Memory allocation	4.1 Mb

Figure 3a Basic information

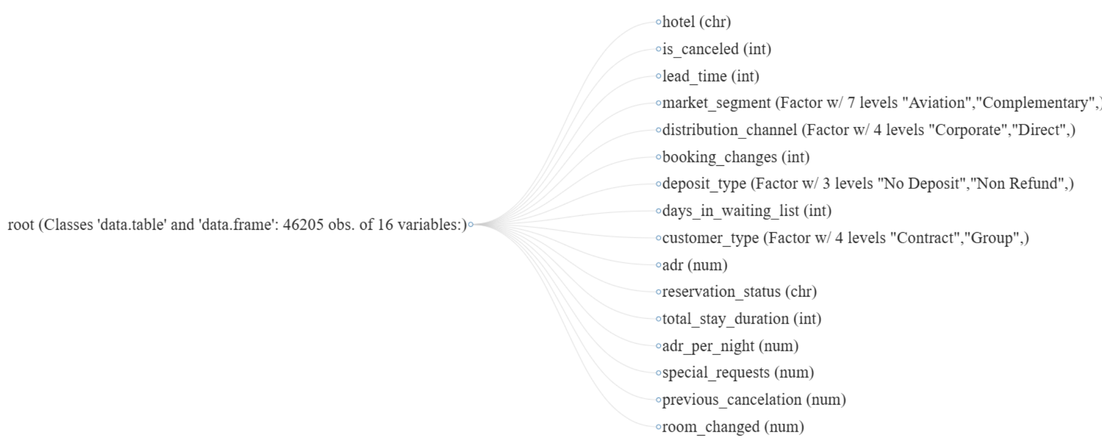

Figure 3b columns name and type

Numerical Variables

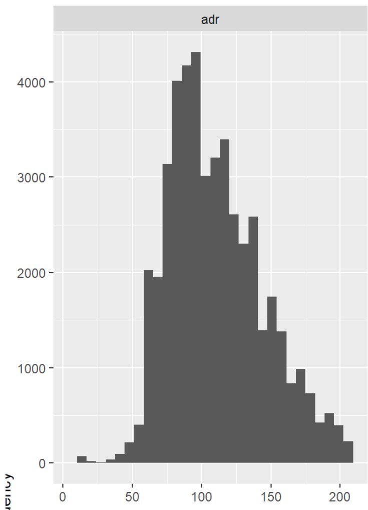

Figure 4a histogram of the adr

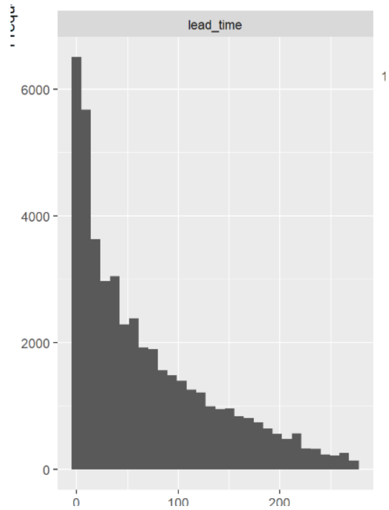

Figure 4b histogram of the lead time

Figure 4a: ADR (Average daily rate)

- The distribution of the ADR data points is normally distributed, with a peak around 100 suggesting high frequency at this point.
- The data points range from 0 to 200, having the majority for the points between 50 to 150.

Figure 4b: lead time

- The distribution of lead_time is right-skewed, with a high concentration on around the 0-value indicating that most bookings have a low lead time.
- The lead times vary widely, with some going over 200, but the frequency of longer lead times is decreasing.

Categorial variables

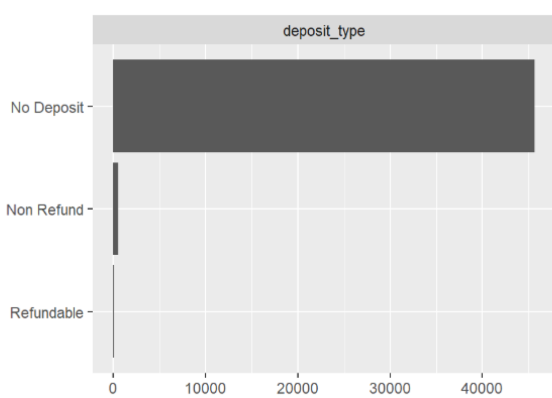

Figure 5a deposit bar chart

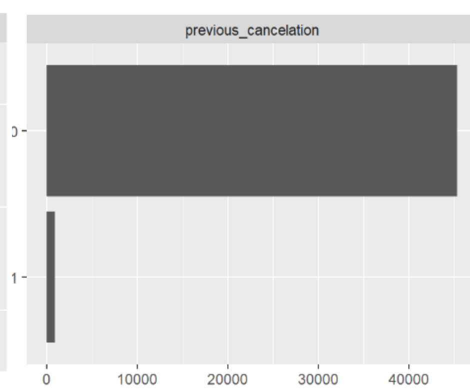

Figure 5b previous cancellation bar chart

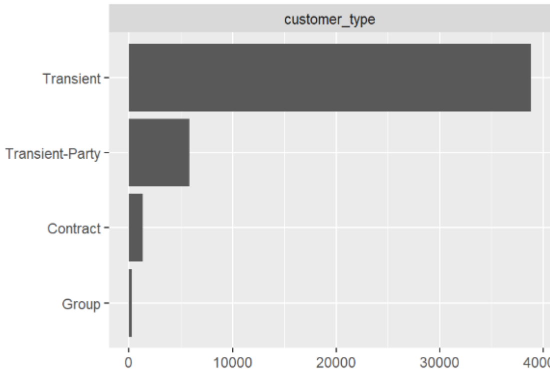

Figure 5c Customer type bar chart

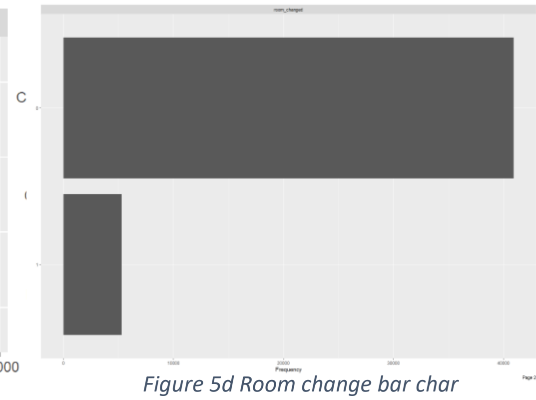

Figure 5d Room change bar char

Figure 5a Deposit type bar chart: "No Deposit", "Non Refund" and "Refundable". The majority of the bookings have “No-deposit” payment type. There are significantly fewer bookings with "Non Refund" and "Refundable" deposit types, with "Refundable" being the least common.

Figure 5b previous cancellation bar chart: "0" (no previous cancellations) and "1" (at least one previous cancellation). The majority of people who booked don’t have previous cancellation.

Figure 5c Customer type bar chart: "Transient", "Transient-Party", "Contract", and "Group". The mijority of the customers fall under the transients’ type, since most of those who visit city hotels visit it for short term to complete a mission they have.

Figure 5d Room change bar chart: "0" (no room change) and "1" (room changed). The majority of customers were satisfied about their rooms and didn’t request to change.

3.6 Summary and Notes

- The dataset contained 26,184 duplicated observations. After inspection, most duplicates appeared to be data entry errors. For this reason, duplicates in this dataset were removed.
- There were no missing values in the dataset; however, 'Undefined' values were identified in certain columns, specifically: market_segment (2 undefined values) and distribution_channel (4 undefined values). The undefined values were replaced by the mode (most frequent value).
- Several features were engineered to enhance the dataset structure and reduce the dataset dimensionality.
- Categorical columns were encoded as factors, including meal, country, market_segment, distribution_channel, customer_type, and deposit_type.
- Boxplots were used to identify outliers in key numerical columns, such as adr, lead_time, and total_stay_duration. After examining the outliers, the decision to remove some outliers was taken.

4.0 Statistical Data Analysis

4.1 Lead time stats

```
lead_time_stats <- cleanest_data %>%
  group_by(reservation_status) %>%
  summarise(
    mean = mean(lead_time),
    median = median(lead_time),
    mode = get_mode(lead_time),
    range = max(lead_time) - min(lead_time),
    variance = var(lead_time),
    sd = sd(lead_time),
    iqr = IQR(lead_time)
  )

print(lead_time_stats)
```

reservation_status	mean	median	mode	range	variance	sd	iqr
<chr>	<dbl>	<dbl>	<int>	<int>	<dbl>	<dbl>	<dbl>
Canceled	88.45352	71.0	18	273	4729.617	68.77221	103.0
Check-Out	61.39924	39.0	0	273	4100.305	64.03363	88.0
No-Show	51.23746	29.5	0	272	3383.753	58.17003	78.5

Key Points:

- **Canceled reservations** have the highest mean and median lead times, indicating that these bookings are made farthest in advance.
- **Check-out reservations** generally have shorter lead times than canceled ones, but there is still significant variation in booking behavior.
- **No-show reservations** tend to have the shortest lead times, with many bookings made at the last minute or even for the same day, possibly indicating less commitment from these guests.

- All three categories show substantial variability, as indicated by high standard deviations and wide ranges. However, canceled reservations have the largest spread. The smaller variability in no-show reservations may indicate that guests who book at the last minute are less committed, leading to more no-shows.

Tips to mitigate cancellation and no show-off issue:

- For bookings with high lead time (showed high cancellation), offer more flexible cancellation terms but gradually tighten the policy as the check-in date approaches.
- For bookings with low lead time (which have higher no-show rates), consider requiring non-refundable deposits or limiting flexibility.
- For bookings made well in advance, remind guests a month, two weeks, and a few days before their stay.

4.2 ADR (Average Daily Rate) Stats:

```
adr_stats <- clean_data %>%
  group_by(reservation_status) %>%
  summarise(
    mean = mean(adr),
    median = median(adr),
    mode = get_mode(adr),
    range = max(adr) - min(adr),
    variance = var(adr),
    sd = sd(adr),
    iqr = IQR(adr)
  )

print(adr_stats)
```

reservation_status <chr>	mean <dbl>	median <dbl>	mo... <dbl>	range <dbl>	variance <dbl>	sd <dbl>	iqr <dbl>
Canceled	116.3098	112.59	126	196.93	1105.635	33.25109	48.5000
Check-Out	108.6258	103.00	75	196.83	1148.669	33.89202	47.5500
No-Show	104.8178	96.30	65	198.00	1187.332	34.45768	47.0075

Key Points

- **Canceled bookings** have the highest ADR and the most variability, suggesting that higher-priced bookings are at greater risk of being canceled.
- **Check-out bookings** have a moderately high ADR with more consistency, indicating that guests are more likely to complete their stays in rooms with stable pricing. The mode for check-out reservations is relatively low (75), suggesting that many bookings are made at more affordable rates.
- **No-shows** tend to happen more frequently for lower-priced bookings, and the consistency in ADR suggests that guests booking cheaper rooms are more likely to fail to show up.

Tips to mitigate Canceled booking and no show off:

- for high ADR consider making offers that can attract the customers and reduce the probabitly of cancellation because of high priced bookings.
- Set non-refundable fees or some penalty for no show off.

5.0 Data Visualization

5.1 Categorical data visualization

Key points for deposit type:

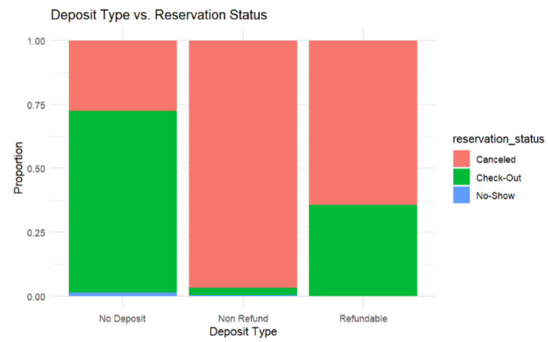

Figure 6a deposit type vs Reservation Status bar chart

No deposit options, seem to be the most effective deposit type, with the highest check-out rates.

Non-refundable deposits seem to be linked to high cancellation rates.

Refundable deposits do not appear to encourage guests to follow through with their stays, which could suggest that the option to cancel without penalty diminishes the commitment to the booked stay.

```
ggplot(cleanest_data, aes(x = deposit_type, fill = reservation_status)) +
  geom_bar(position = "fill") +
  labs(title = "Deposit Type vs. Reservation Status", x = "Deposit Type", y = "Proportion") +
  theme_minimal()
```

Key points for previous cancellation:

Figure 6b Previous Cancellation vs Reservation Status bar chart

Customers with previous cancellations are significantly more likely to cancel again compared to those with no previous cancellations. This suggests that past behavior can be a strong predictor of future cancellation.

```
ggplot(cleanest_data, aes(x = as.factor(previous_cancelation), fill = reservation_status)) +
  geom_bar(position = "fill") +
  labs(title = "Previous Cancellations vs. Reservation Status", x = "Previous Cancellations", y =
"Proportion") + theme_minimal()
```

Key points for room changes:

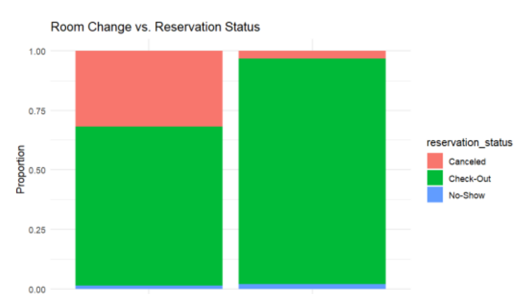

Figure 6c Room Change vs Reservation Status bar chart

Customers who requested or were granted room changes are highly likely to check out successfully and are much less likely to cancel compared to those who did not change rooms.

Allowing room changes could potentially improve customer satisfaction.

```
ggplot(cleanest_data, aes(x = as.factor(room_changed), fill = reservation_status)) +
  geom_bar(position = "fill") +
  labs(title = "Room Change vs. Reservation Status", x = "Room Changed", y = "Proportion") +
  theme_minimal()
```

Key points for customer types:

Transit customers showed the highest cancelation rate. This is most probably because of their need for immediate accommodation on the last minutes. Unlike the contract or group customer types were their plans can be more predictable and have higher lead time.

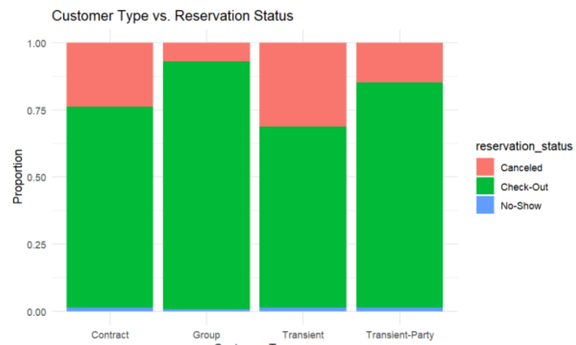

Figure 6d Customer Type vs Reservation Status bar chart

```
# Customer Type vs. Reservation Status
ggplot(cleanest_data, aes(x = customer_type, fill = reservation_status)) +
  geom_bar(position = "fill") +
  labs(title = "Customer Type vs. Reservation Status", x = "Customer Type", y = "Proportion") +
  theme_minimal()
```

5.2 Numerical Variables

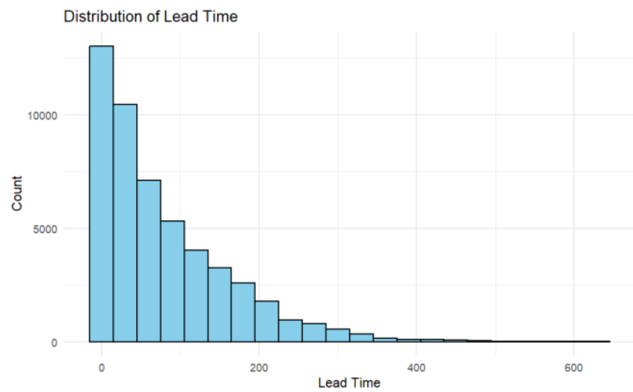

Figure 7a Histogram of lead time

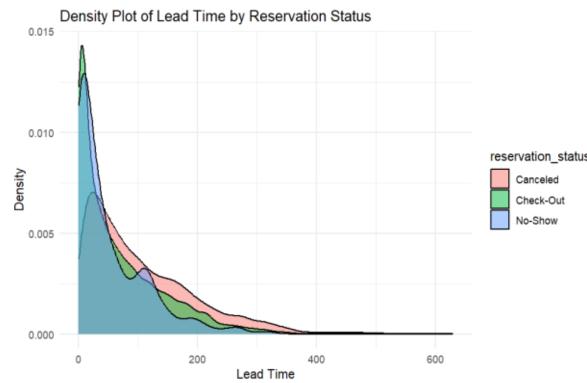

Figure 7b Density Plot of lead time

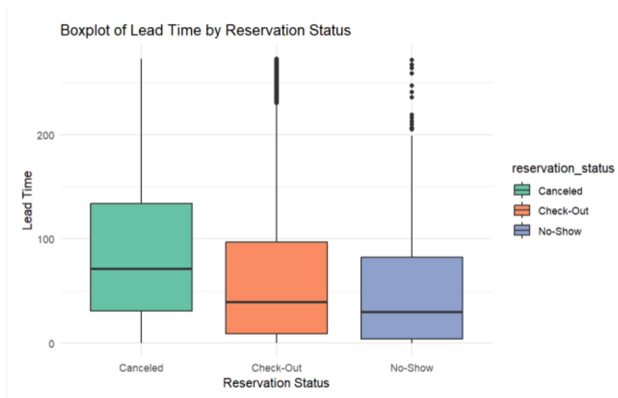

Figure 7c Boxplot for lead time by Reservation Status groups

Figure 7a: Histogram of Lead Time

- Most reservations have short lead times, meaning lower lead time values (close to 0).
- The distribution is skewed right, meaning the majority of lead times are low, but there are still a few reservations with higher lead times.

Figure 7b: Density Plot of Lead Time by Reservation Status

- Canceled reservations have longer lead times compared to other statuses.

Figure 7c: Boxplot of Lead Time by Reservation Status Groups

- Canceled reservations have the largest variance of lead times, with a median lead time around 80 days.
- Check-Out reservations have a shorter median lead time, around 40 days, and less variability compared to the canceled group.
- No-Show reservations have the lowest median indicating that no-show behavior tend to appear with those who book few days before their stay.

Distribution of ADR (Average Daily Rate):

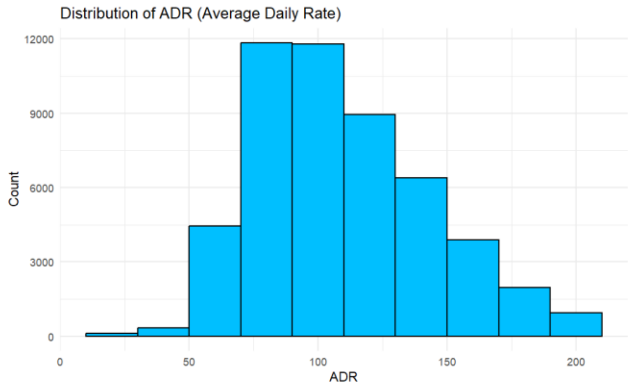

Figure 8a histogram of adr

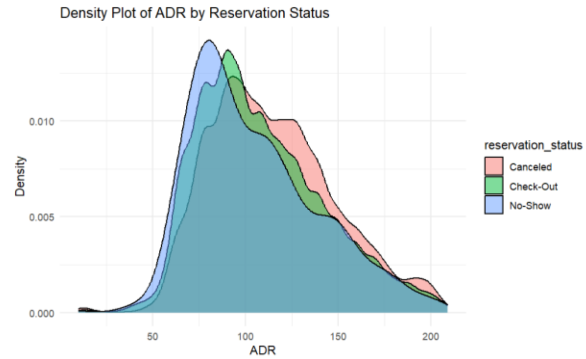

Figure 8b Density plot of adr

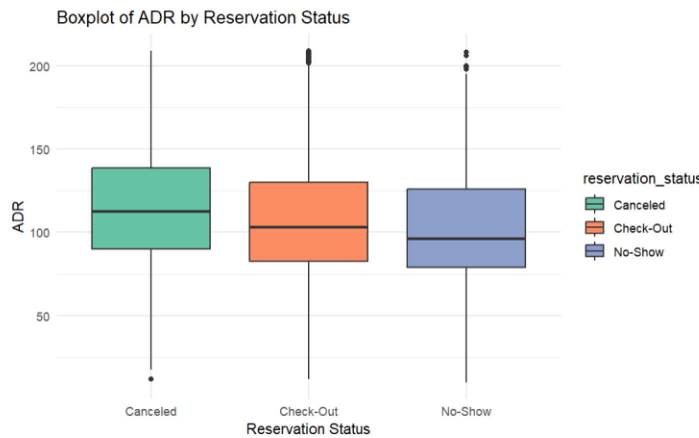

Figure 8c Boxplot of ADR by reservation Status

Figure 8a: Histogram of ADR (Average Daily Rate)

- Most ADR values cluster around the center, with a peak between 100 and 150.
- The distribution is approximately symmetric, but slightly skewed to the right.

Figure 8b: Density Plot of ADR by Reservation Status

- High ADR seem to correlate with higher chance of booking cancelation.
- Lower ADR seems to correlate with higher chances of no-show

Figure 8c: Boxplot of ADR by Reservation Status

- Canceled reservations tend to have a higher median compared to the other two groups.
- No-Show reservations have the lowest median ADR around 96, showing that bookings with lower daily rates tend to result in no-shows.
- The spread for No-Show reservations is similar to Check-Out.

5.3 Summary and notes:

- **Categorical Data**: Variables such as deposit type, previous cancellations, room changes, and customer type provide significant insight into reservation outcomes.
  - non-refundable deposits correlate with higher cancellation rates, and transit customers show the highest cancellation rates.
  - Those who previously cancelled are more likely to cancel future bookings.
  - Those who were granted room request change were less likely to cancel bookings.
  - Transit customers showed the highest cancelation rate.
- **Numerical Data:** Lead time and ADR have distinct distribution patterns provided valuable insights into customer behavior.
  - Longer lead times tend to result in higher cancellation rates.
  - Bookings with high ADR showed higher cancelation rate.

6.0 Conclusion

In conclusion, the analysis of this hotel booking data shows that hotels can increase its operational efficiency in a lot of ways. Key examples are: implementing flexible yet firm policies for bookings with high lead times, offering incentives for high ADR bookings, adopting stricter cancellation terms for transit customers and guests with previous cancellations, and meeting customer requirement.