Comprehensive Analysis of Hotel Booking Trends and Cancellation Patterns

Ehab Nasr Farouk

2024-09-23

Introduction

This report provides a comprehensive analysis of the hotel booking dataset, examining booking trends, cancellations, and customer demographics to derive actionable insights.

Agenda

1. Introduction

- **Objective**: Define the goals of your analysis. What questions are you trying to answer or what insights are you looking for?
- **Dataset Overview**: Brief description of the dataset, including its size, columns, and general contents.

2. Data Preparation

- **Load the Data**: Use read.csv() or other functions to load your dataset into R.
- **Inspect the Data**: Use functions like head(), str(), and summary() to understand the structure and contents of the dataset.
- **Handle Missing Values**: Identify and address missing values using methods like imputation or removal.
- **Convert Data Types**: Ensure that all columns have the correct data types (e.g., factors, dates).
- **Feature Engineering**: Create new variables if needed (e.g., binning continuous variables).

3. Exploratory Data Analysis (EDA)

- **Descriptive Statistics**: Calculate summary statistics such as mean, median, and standard deviation.
- **Data Visualization**:
  - **Bivariate Analysis**: Use scatter plots, line charts

4. Data Analysis

- **Group Analysis**: Aggregate and summarize data based on categorical variables (e.g., total cancellations by country or market segment).
- **Trend Analysis**: Analyze trends over time if applicable (e.g., cancellation rates by month or year).
- **Comparison Analysis**: Compare different groups or categories to identify patterns or significant differences.

Overview of Dataset

1. Hotel: Indicates the type of hotel (e.g., "City Hotel" or "Resort Hotel").

2. Is_canceled: Whether a booking was canceled (1) or not (0).

3. Lead_time: The number of days between the booking date and the arrival date.

4. Arrival_date_year, Arrival_date_month, Arrival_date_day_of_month: Details about the arrival date.

5. Stays_in_weekend_nights, Stays_in_week_nights: The number of nights stayed during the weekend and week, respectively.

6. Adults, Children, Babies: Number of adults, children, and babies in the booking.

7. Meal: Type of meal plan booked.

8. Country: Country of origin of the customer.

9. Market_segment, Distribution_channel: The market segment and channel through which the booking was made.

10. Is_repeated_guest: Indicates if the guest is a returning customer.

11. Previous_cancellations, Previous_bookings_not_canceled: Details about the guest's previous cancellations and bookings.

12. Reserved_room_type, Assigned_room_type: The room types reserved and assigned.

13. Booking_changes: Number of changes made to the booking.

14. Deposit_type: The type of deposit required.

15. Agent, Company: Identifiers for the booking agent or company.

16. Days_in_waiting_list: Number of days the booking was on a waiting list.

17. Customer_type: Type of customer (e.g., "Transient", "Group").

18. ADR: Average Daily Rate, indicating the daily rate paid per room.

19. Required_car_parking_spaces: Number of parking spaces required.

20. Total_of_special_requests: Number of special requests made by the customer.

21. Reservation_status, Reservation_status_date: The reservation's status (e.g., "Canceled", "Check-out") and the date of the status.

Import Necessary Libraries

```
library(tidyverse)

library(ggrepel)

library(dplyr)

library(forcats)
```

Load the dataset

```
df <- read.csv("D:\\DataScience-projects\\R
Project\\Dataset\\hotel_bookings.csv")
```

Display first few rows of the dataset

```
head(df)

##          hotel is_canceled lead_time arrival_date_year
arrival_date_month
## 1 Resort Hotel           0       342              2015
July
## 2 Resort Hotel           0       737              2015
July
##   arrival_date_week_number arrival_date_day_of_month
stays_in_weekend_nights
## 1                       27                         1
0
## 2                       27                         1
0
##   stays_in_week_nights adults children babies meal country
market_segment
## 1                    0      2        0      0   BB     PRT
Direct
## 2                    0      2        0      0   BB     PRT


Direct
##   distribution_channel is_repeated_guest previous_cancellations
## 1               Direct                 0                      0
## 2               Direct                 0                      0
```

```
##   previous_bookings_not_canceled reserved_room_type
assigned_room_type
## 1                              0                  C
C
## 2                              0                  C
C

##   booking_changes deposit_type agent company days_in_waiting_list
customer_type
## 1               3   No Deposit  NULL    NULL                    0
Transient
## 2               4   No Deposit  NULL    NULL                    0
Transient

##   adr required_car_parking_spaces total_of_special_requests
reservation_status
## 1   0                           0                         0
Check-Out
## 2   0                           0                         0
Check-Out

##   reservation_status_date
## 1              2015-07-01
## 2              2015-07-01
```

Dataset dimensions

```r
cat("Number of rows:",nrow(df))

## Number of rows: 119390

cat("Number of columns:",ncol(df))

## Number of columns: 32
```

Summary of columns

```r
str(df)

## 'data.frame':    119390 obs. of  32 variables:
##  $ hotel                         : chr  "Resort Hotel" "Resort
Hotel" "Resort Hotel" "Resort Hotel" ...
##  $ is_canceled                   : int  0 0 0 0 0 0 0 0 1 1 ...
##  $ lead_time                     : int  342 737 7 13 14 14 0 9 85 75
...
##  $ arrival_date_year             : int  2015 2015 2015 2015 2015
2015 2015 2015 2015 2015 ...
##  $ arrival_date_month            : chr  "July" "July" "July" "July"
...
```

```
##  $ arrival_date_week_number      : int  27 27 27 27 27 27 27 27 27
27 ...
##  $ arrival_date_day_of_month     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ stays_in_weekend_nights       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ stays_in_week_nights          : int  0 0 1 1 2 2 2 2 3 3 ...
##  $ adults                        : int  2 2 1 1 2 2 2 2 2 2 ...
##  $ children                      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ babies                        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ meal                          : chr  "BB" "BB" "BB" "BB" ...
##  $ country                       : chr  "PRT" "PRT" "GBR" "GBR" ...
##  $ market_segment                : chr  "Direct" "Direct" "Direct"
"Corporate" ...
##  $ distribution_channel          : chr  "Direct" "Direct" "Direct"
"Corporate" ...
##  $ is_repeated_guest             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ previous_cancellations        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ previous_bookings_not_canceled: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ reserved_room_type            : chr  "C" "C" "A" "A" ...
##  $ assigned_room_type            : chr  "C" "C" "C" "A" ...
##  $ booking_changes               : int  3 4 0 0 0 0 0 0 0 0 ...
##  $ deposit_type                  : chr  "No Deposit" "No Deposit"
"No Deposit" "No Deposit" ...
##  $ agent                         : chr  "NULL" "NULL" "NULL" "304"
...
##  $ company                       : chr  "NULL" "NULL" "NULL" "NULL"
...
##  $ days_in_waiting_list          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ customer_type                 : chr  "Transient" "Transient"
"Transient" "Transient" ...
##  $ adr                           : num  0 0 75 75 98 ...
##  $ required_car_parking_spaces   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ total_of_special_requests     : int  0 0 0 0 1 1 0 1 1 0 ...
##  $ reservation_status            : chr  "Check-Out" "Check-Out"
"Check-Out" "Check-Out" ...
##  $ reservation_status_date       : chr  "2015-07-01" "2015-07-01"
"2015-07-02" "2015-07-02" ...
```

Convert date column to date type

```r
df$reservation_status_date  <-
as.Date(df$reservation_status_date,format="%Y-%m-%d")
```

Check for missing values

```r
missing_values <- sapply(df,function(x) sum(is.na(x)))
missing_values
```

```
##                          hotel                    is_canceled
##                              0                              0
##                      lead_time              arrival_date_year
##                              0                              0
##             arrival_date_month       arrival_date_week_number
##                              0                              0
##      arrival_date_day_of_month        stays_in_weekend_nights
##                              0                              0
##           stays_in_week_nights                         adults
##                              0                              0
##                       children                         babies
##                              4                              0
##                           meal                        country
##                              0                              0
##                 market_segment           distribution_channel
##                              0                              0
##              is_repeated_guest         previous_cancellations
##                              0                              0
## previous_bookings_not_canceled             reserved_room_type
##                              0                              0
##             assigned_room_type                booking_changes
##                              0                              0
##                   deposit_type                          agent
##                              0                              0
##                        company           days_in_waiting_list
##                              0                              0
##                  customer_type                            adr
##                              0                              0
##    required_car_parking_spaces      total_of_special_requests
##                              0                              0


##             reservation_status        reservation_status_date
##                              0                              0
```

Clean column names

```r
df <- df %>% rename_all(~ str_replace_all(.," ","_"))
names(df)
```

```
##  [1] "hotel"                          "is_canceled"
##  [3] "lead_time"                      "arrival_date_year"
##  [5] "arrival_date_month"             "arrival_date_week_number"
##  [7] "arrival_date_day_of_month"      "stays_in_weekend_nights"
##  [9] "stays_in_week_nights"           "adults"
## [11] "children"                       "babies"
```

```
## [13] "meal"                           "country"
## [15] "market_segment"                 "distribution_channel"
## [17] "is_repeated_guest"              "previous_cancellations"
## [19] "previous_bookings_not_canceled" "reserved_room_type"
## [21] "assigned_room_type"             "booking_changes"
## [23] "deposit_type"                   "agent"
## [25] "company"                        "days_in_waiting_list"
## [27] "customer_type"                  "adr"
## [29] "required_car_parking_spaces"    "total_of_special_requests"
## [31] "reservation_status"             "reservation_status_date"
```

Summary statistics for numerical columns

```
summary(df %>% select(lead_time, stays_in_weekend_nights,
stays_in_week_nights, adr))

##    lead_time   stays_in_weekend_nights stays_in_week_nights      adr
##  Min.   :  0   Min.   : 0.0000         Min.   : 0.0          Min.   :
-6.38
##  1st Qu.: 18   1st Qu.: 0.0000         1st Qu.: 1.0          1st Qu.:
69.29
##  Median : 69   Median : 1.0000         Median : 2.0          Median :
94.58
##  Mean   :104   Mean   : 0.9276         Mean   : 2.5          Mean   :
101.83
##  3rd Qu.:160   3rd Qu.: 2.0000         3rd Qu.: 3.0          3rd Qu.:
126.00
##  Max.   :737   Max.   :19.0000         Max.   :50.0          Max.
:5400.00
```

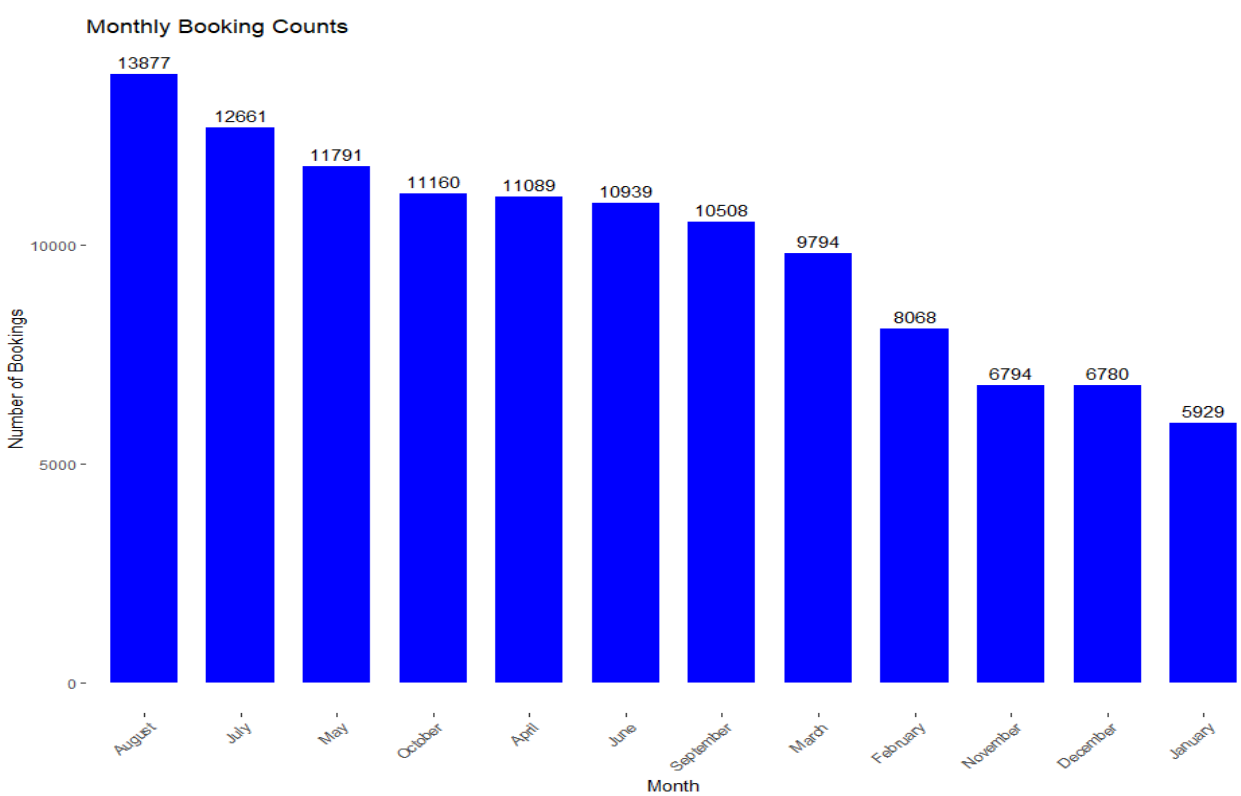
Distribution of cancellation

```
table(df$is_canceled)

##
##     0     1
## 75166 44224
```

Booking Trends

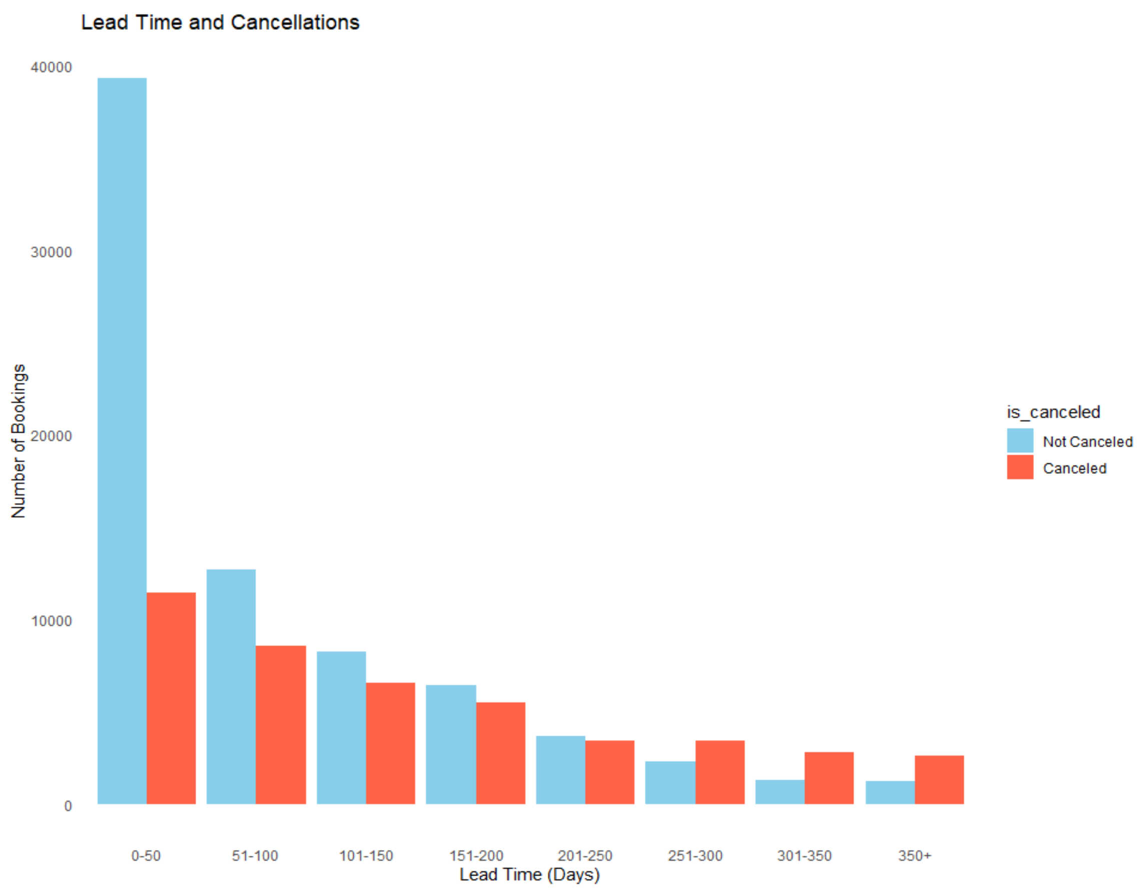
1. Is there a seasonal trend in the booking patterns?

```
library(ggplot2)
library(dplyr)


monthly_booking <- df %>%
  group_by(arrival_date_month) %>%
  summarize(booking_count = n()) %>%
  mutate(arrival_date_month = fct_reorder(arrival_date_month,
booking_count, .desc = TRUE))


ggplot(monthly_booking, aes(x = arrival_date_month, y = booking_count))
+
geom_bar(stat = "identity", fill = "blue", width = 0.7) +
geom_text(aes(label = booking_count), vjust = -0.5) +
labs(title = "Monthly Booking Counts", x = "Month", y = "Number of
Bookings") +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1, vjust = 1),
    panel.background = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    plot.background = element_blank(),
    plot.margin = margin(t = 10, r = 10, b = 10, l = 10)  )
```



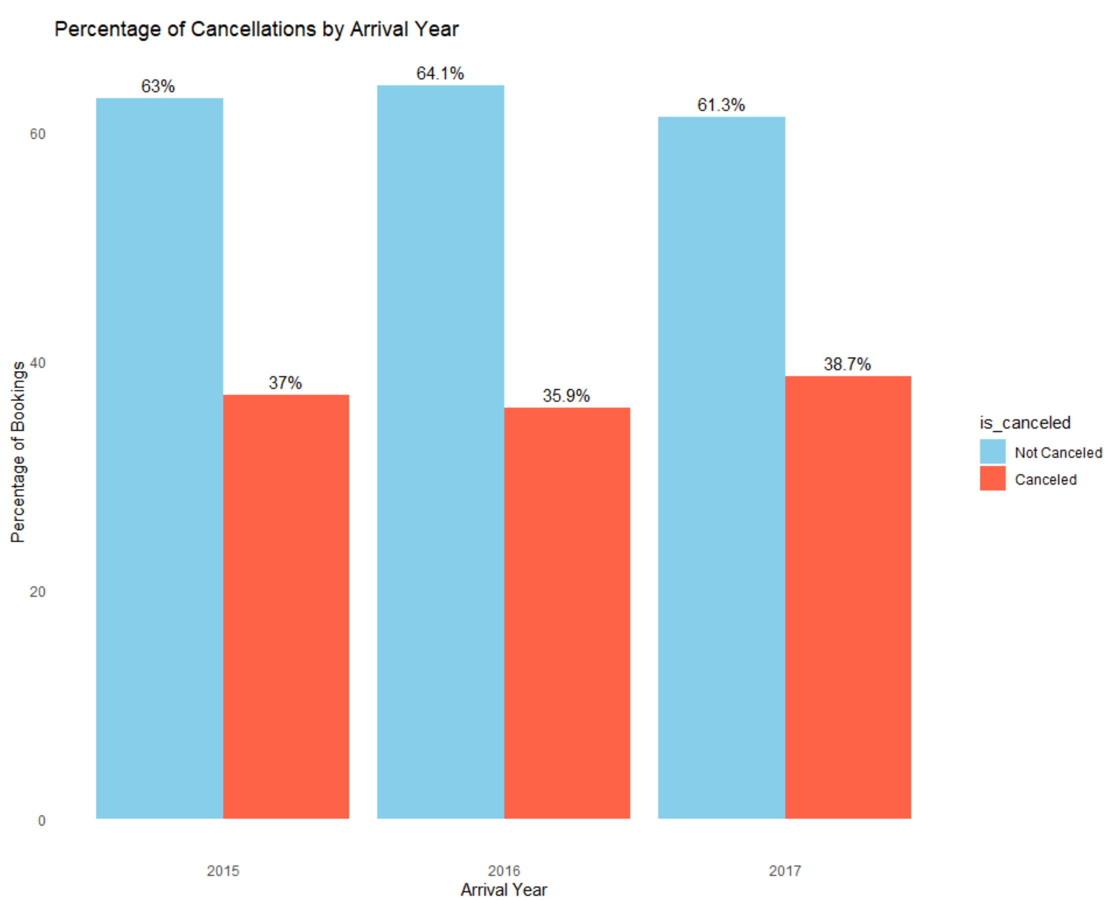
2. Does the booking cancellation rate increase as the lead time increases?

```
df$is_canceled <- factor(df$is_canceled, levels = c(0, 1), labels =
c("Not Canceled", "Canceled"))

df <- df %>%
  mutate(lead_time_bins = cut(lead_time,
          breaks = c(0, 50, 100, 150, 200, 250, 300, 350, Inf),
          labels = c("0-50", "51-100", "101-150", "151-200", "201-250",
"251-300", "301-350", "350+"),
       include.lowest = TRUE))

ggplot(df, aes(x = lead_time_bins, fill = is_canceled)) +
  geom_bar(position = "dodge") +
  labs(title = "Lead Time and Cancellations",
       x = "Lead Time (Days)",
       y = "Number of Bookings") +
  scale_fill_manual(values = c("Not Canceled" = "skyblue", "Canceled" =
"tomato")) +
  theme_minimal() +
  theme(panel.grid = element_blank())
```



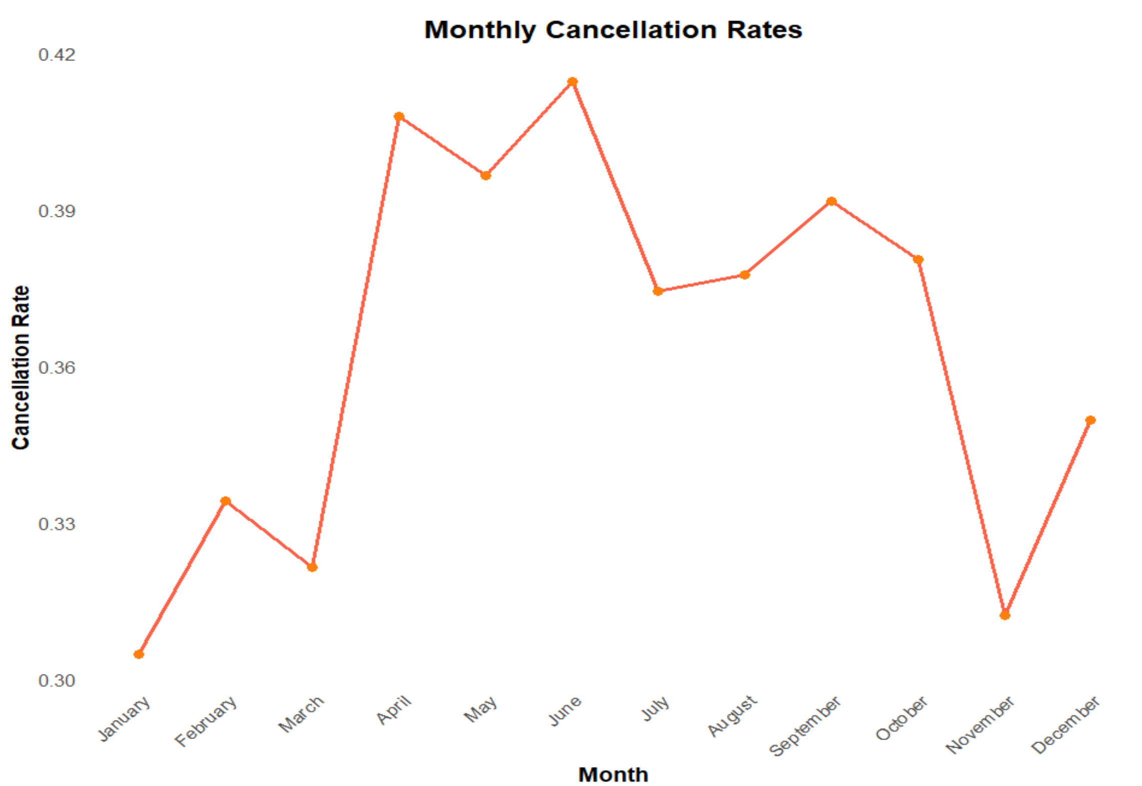
3. How does the percentage of cancellations vary by arrival year?

```
df$is_canceled <- factor(df$is_canceled, levels = c(0, 1), labels =
c("Not Canceled", "Canceled"))

summary_df <- df %>%
  group_by(arrival_date_year, is_canceled) %>%
  summarise(count = n(), .groups = 'drop') %>%
  group_by(arrival_date_year) %>%
  mutate(percentage = count / sum(count) * 100)

ggplot(summary_df, aes(x = factor(arrival_date_year), y = percentage,
fill = is_canceled)) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(aes(label = paste0(round(percentage, 1), "%")),
            position = position_dodge(width = 0.9),
            vjust = -0.5) +
  labs(title = "Percentage of Cancellations by Arrival Year",
       x = "Arrival Year",
       y = "Percentage of Bookings") +
  scale_fill_manual(values = c("Not Canceled" = "skyblue", "Canceled" =
"tomato")) +
  theme_minimal()+
  theme(panel.grid = element_blank())
```



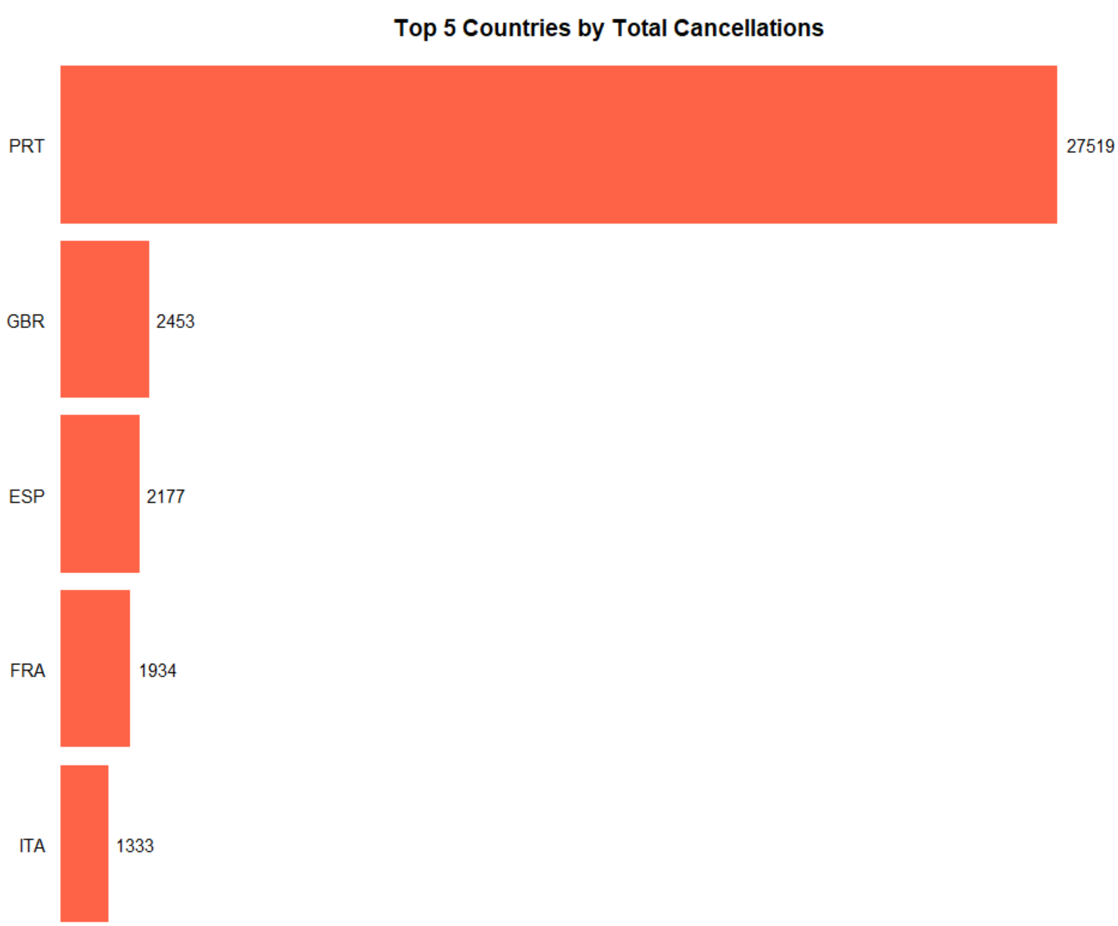
4. How has the cancellation rate changed over the months?

```
df$arrival_date_month <- factor(df$arrival_date_month,
           levels = c("January", "February", "March", "April", "May",
                      "June", "July", "August", "September", "October",
                                          "November", "December"))

cancellation_rates <- df %>%
  group_by(arrival_date_month) %>%
  summarize(cancellation_rate = mean(is_canceled), .groups = 'drop')
%>%
  filter(!is.na(cancellation_rate))

ggplot(cancellation_rates, aes(x = arrival_date_month, y =
cancellation_rate, group = 1)) +
  geom_line(color = "tomato", linewidth = 1.2) +
  geom_point(color = "#ff7f0e", size = 3) +
  labs(x = "Month", y = "Cancellation Rate", title = "Monthly
Cancellation Rates") +
  theme_minimal(base_size = 15) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(hjust = 0.5, size = 18, face="bold"),
        axis.title = element_text(size = 14, face = "bold"),
        axis.text = element_text(size = 12),
       panel.grid.major = element_blank(),
        panel.grid.minor = element_blank())
```



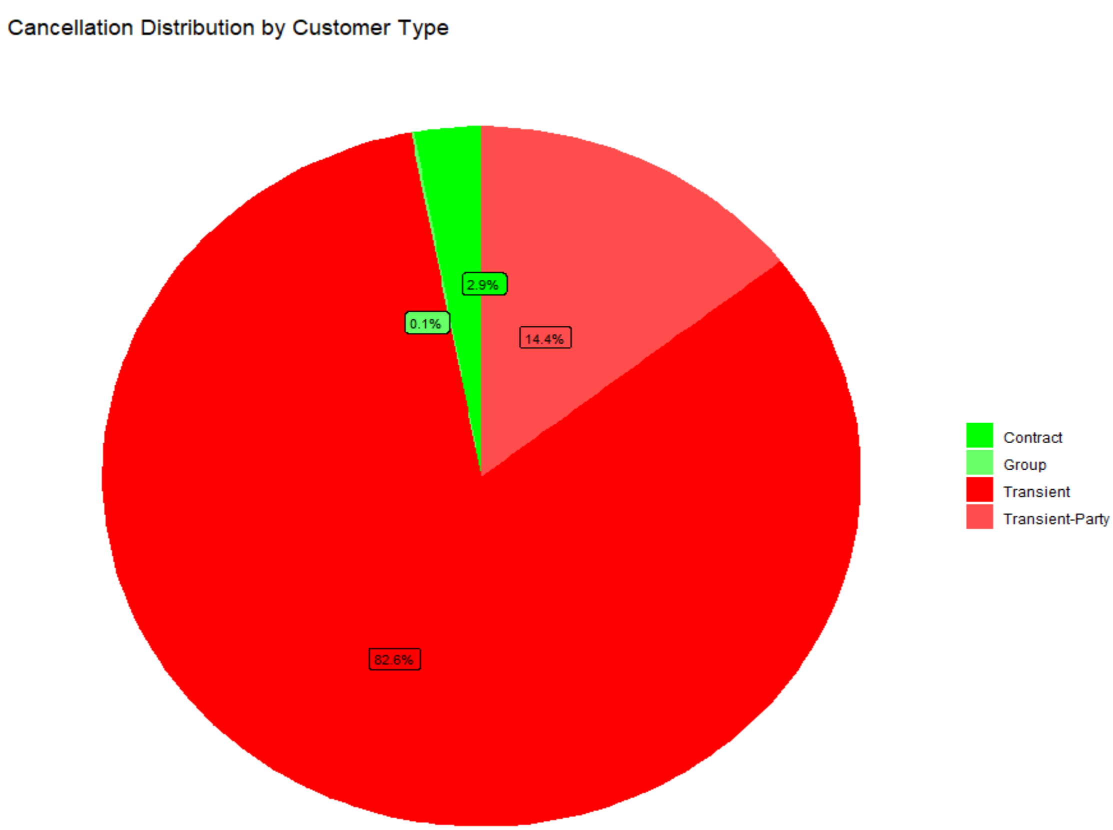
5. Which countries have the highest number of cancellations

```r
df$is_canceled <- as.numeric(as.character(df$is_canceled))

agg_data <- df %>%
  group_by(country) %>%
  summarise(total_cancellations = sum(is_canceled, na.rm = TRUE)) %>%
  slice_max(total_cancellations, n = 5)

ggplot(agg_data, aes(x = total_cancellations, y = reorder(country,
total_cancellations))) +
  geom_bar(stat = "identity", fill = "tomato") +
  geom_text(aes(label = total_cancellations), hjust = -0.2, color =
"black") +
  labs(x = "Total Cancellations", y = "Country", title = "Top 5
Countries by Total Cancellations") +
  theme_void() +
  theme(
    axis.text.y = element_text(angle = 0, hjust = 1, margin = margin(r
= 10)),
    plot.title = element_text(hjust = 0.5, face = "bold", size = 14)
  ) +
  scale_x_continuous(expand = expansion(mult = c(0, 0.1)))
```



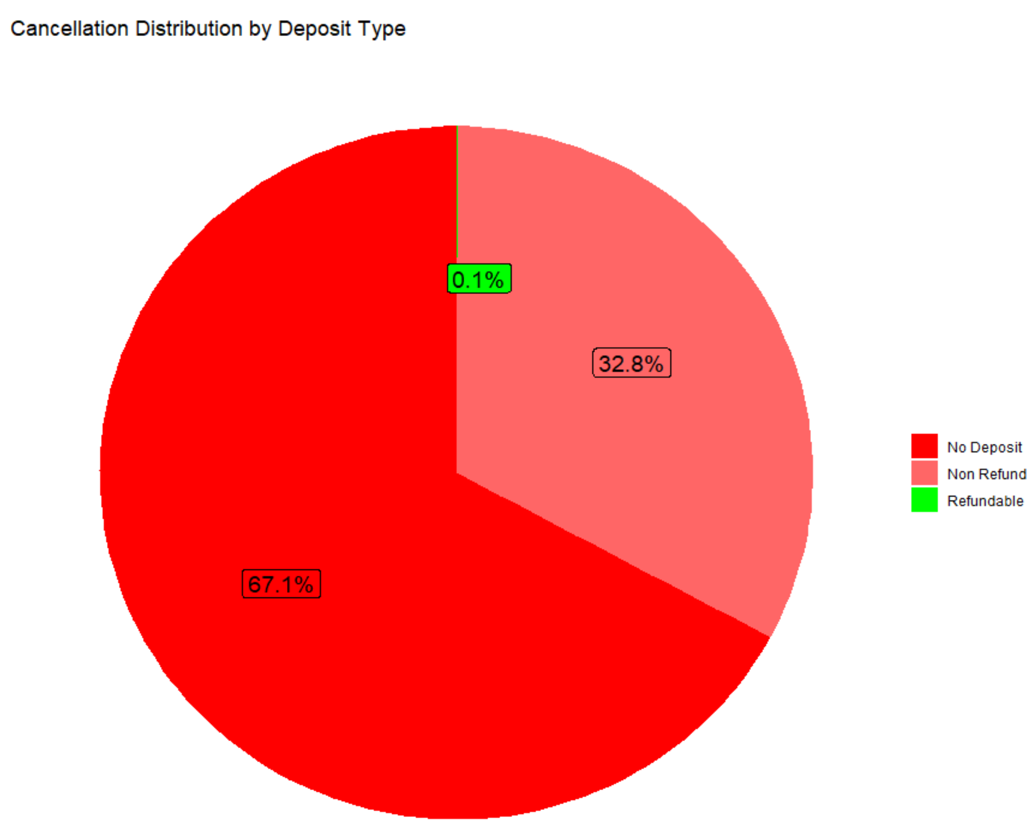
6. Is there a customer type that consistently avoids cancellations?

```
agg_data <- df %>%
  group_by(customer_type) %>%
  summarise(total_cancellations = sum(is_canceled, na.rm = TRUE)) %>%
  mutate(percentage = total_cancellations / sum(total_cancellations) *
100, label = paste0(round(percentage, 1), "%"))

custom_colors <- c("Contract" = "#00ff00",
                   "Group" = "#66FF66",
                   "Transient" = "red",
                   "Transient-Party" = "#FF4C4C")

ggplot(agg_data, aes(x = "", y = percentage, fill = customer_type)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  labs(x = NULL, y = NULL, title = "Cancellation Distribution by
Customer Type") +
  theme_void() +
  theme(legend.title = element_blank()) +
  geom_label_repel(aes(label = label), position = position_stack(vjust
= 0.5), size = 3, show.legend = FALSE) +
  scale_fill_manual(values = custom_colors)
```



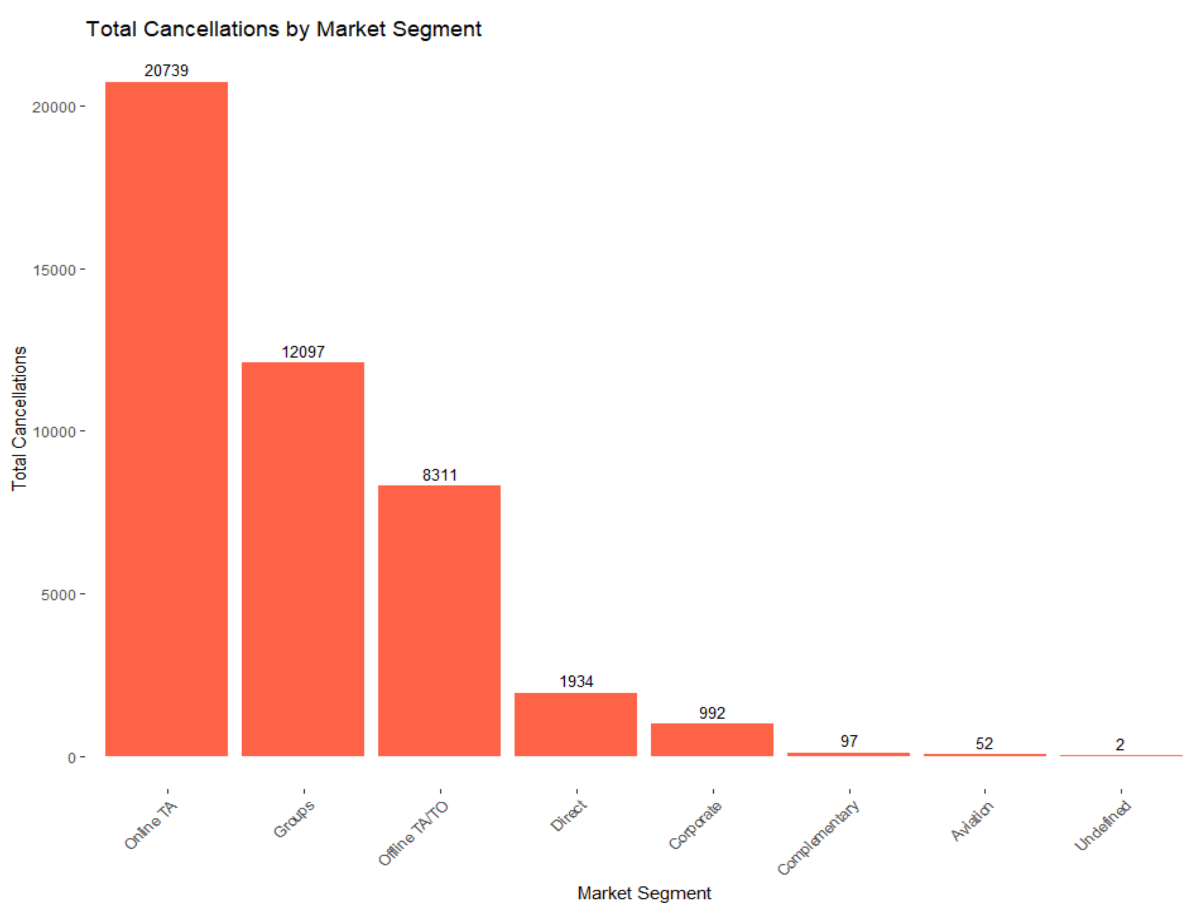
7. How does the deposit policy affect cancellation behavior?

```
agg_data <- df %>%
  group_by(deposit_type) %>%
  summarise(total_cancellations = sum(is_canceled, na.rm = TRUE)) %>%
  mutate(percentage = total_cancellations / sum(total_cancellations) *
100, label = paste0(round(percentage, 1), "%"))

custom_colors <- c("No Deposit" = "red",
                   "Non Refund" = "#FF6666",
                   "Refundable" = "green")

ggplot(agg_data, aes(x = "", y = percentage, fill = deposit_type)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  labs(x = NULL, y = NULL, title = "Cancellation Distribution by
Deposit Type") +
  theme_void() +
  theme(legend.title = element_blank()) +
  geom_label_repel(aes(label = label), position = position_stack(vjust
= 0.5), size = 5, show.legend = FALSE) +
  scale_fill_manual(values = custom_colors)
```



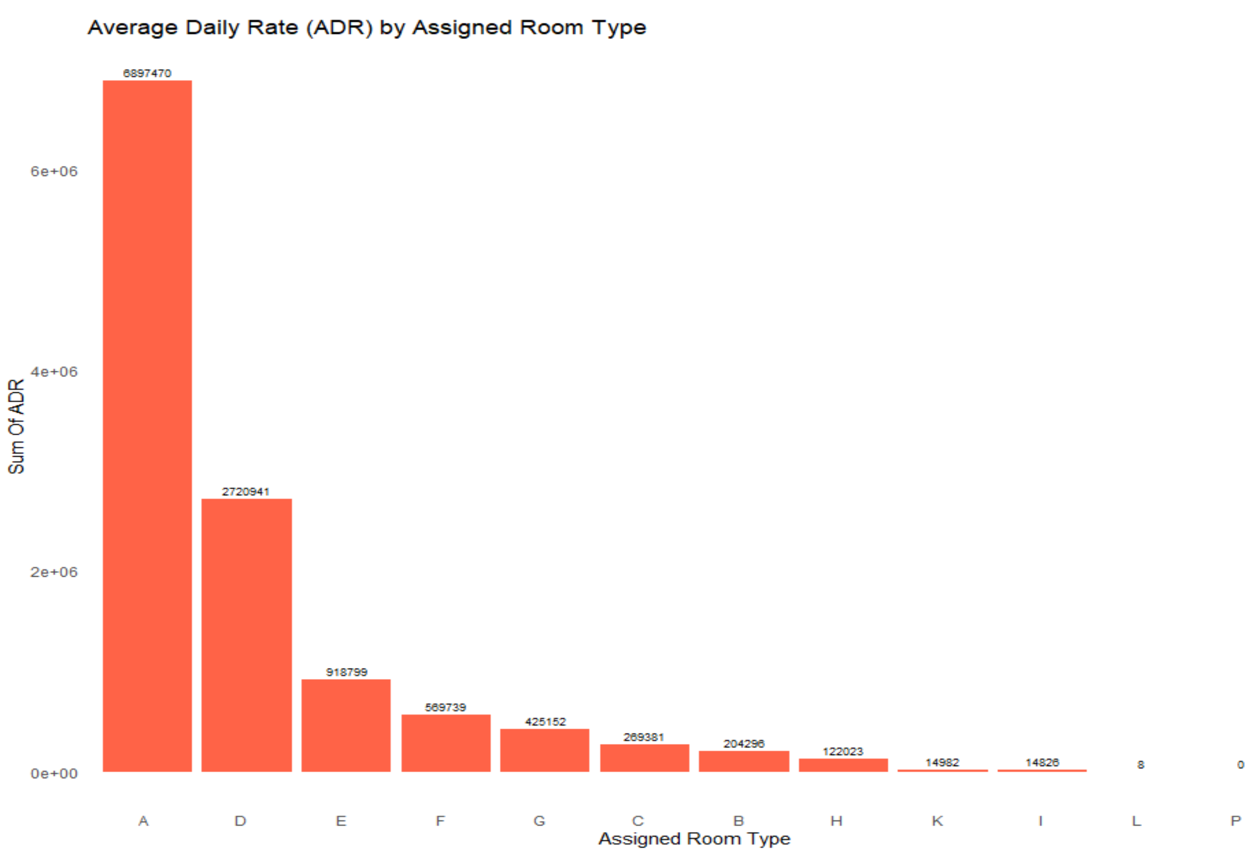
8. How do different market segments compare in terms of cancellation rates?

```r
agg_data <- df %>%
  group_by(market_segment) %>%
  summarise(total_cancellations = sum(is_canceled, na.rm = TRUE))
ggplot(agg_data, aes(x = reorder(market_segment, -total_cancellations),
y = total_cancellations)) +
  geom_bar(stat = "identity", fill = "tomato") +
  geom_text(aes(label = total_cancellations), vjust = -0.5, size = 3.5)
+
  labs(x = "Market Segment", y = "Total Cancellations", title = "Total
Cancellations by Market Segment") +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    plot.background = element_blank()
  )
```



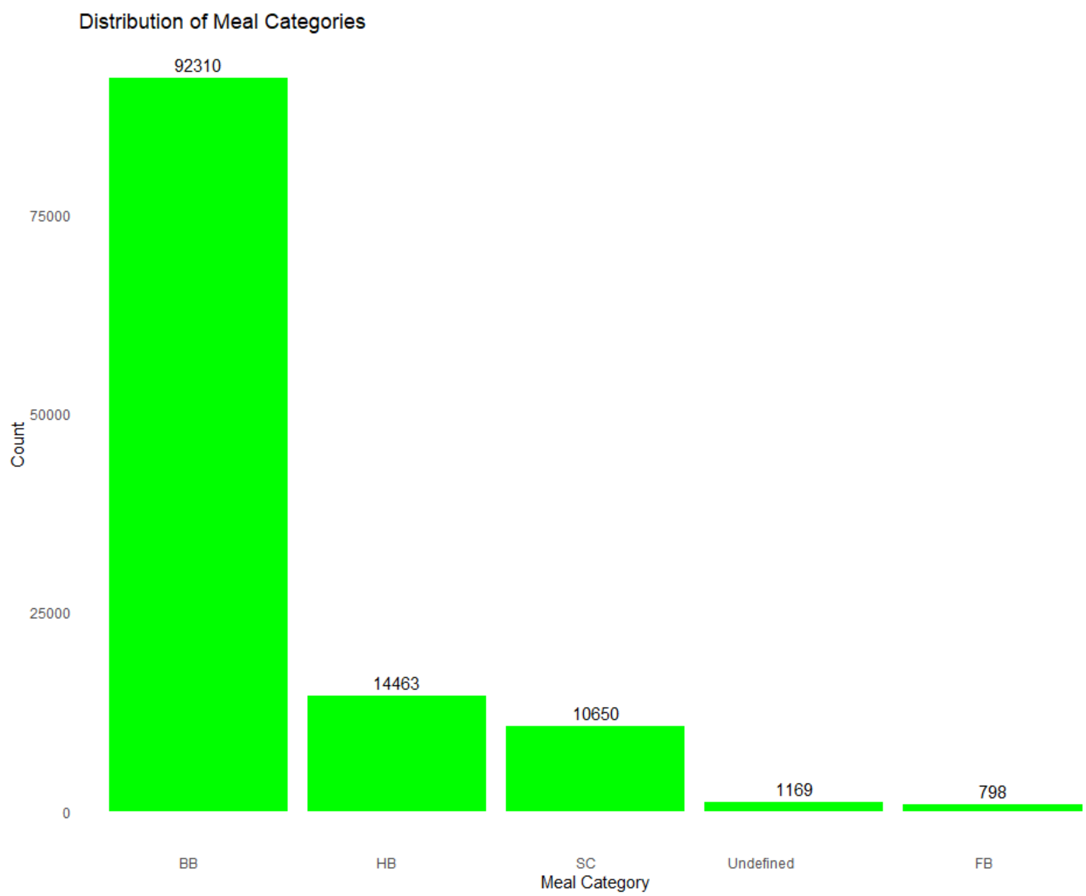
9. Which assigned room type generates the highest sum of ADR?

```
agg_data <- df %>%
  group_by(assigned_room_type) %>%
  summarise(average_adr = sum(adr, na.rm = TRUE))

agg_data <- df %>%
  group_by(assigned_room_type) %>%
  summarise(average_adr = sum(adr, na.rm = TRUE))

ggplot(agg_data, aes(x = reorder(assigned_room_type, -average_adr), y =
average_adr)) +
  geom_bar(stat = "identity", fill = "tomato", width = 0.9) +
  geom_text(aes(label = round(average_adr, 0)), vjust = -0.5, size =
2.5) +
  labs(x = "Assigned Room Type", y = "Sum Of ADR", title = "Average
Daily Rate (ADR) by Assigned Room Type") +
  theme_minimal() +
  theme(
    axis.text.x = element_text(hjust = 1),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    plot.background = element_blank()
  )
```



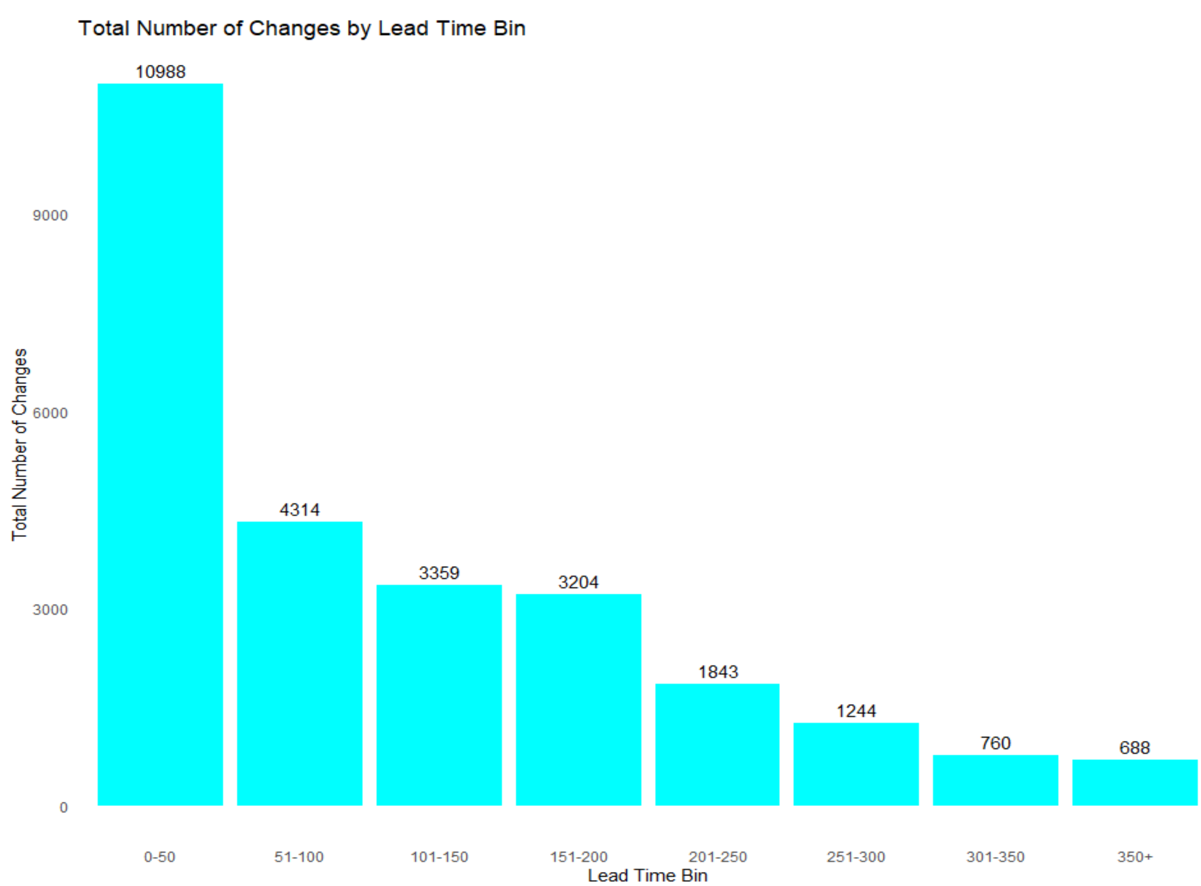
10. Which meal category is the most frequently selected?

```
meal_data <- df %>%
  group_by(meal) %>%
  summarise(count = n())

ggplot(meal_data, aes(x = reorder(meal, -count), y = count)) +
  geom_bar(stat = "identity", fill = "green") +
  geom_text(aes(label = count), vjust = -0.5) +
  labs(x = "Meal Category", y = "Count", title = "Distribution of Meal
Categories") +
  theme_minimal() +
  theme(axis.text.x = element_text(hjust = 1),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        plot.background = element_blank())
```



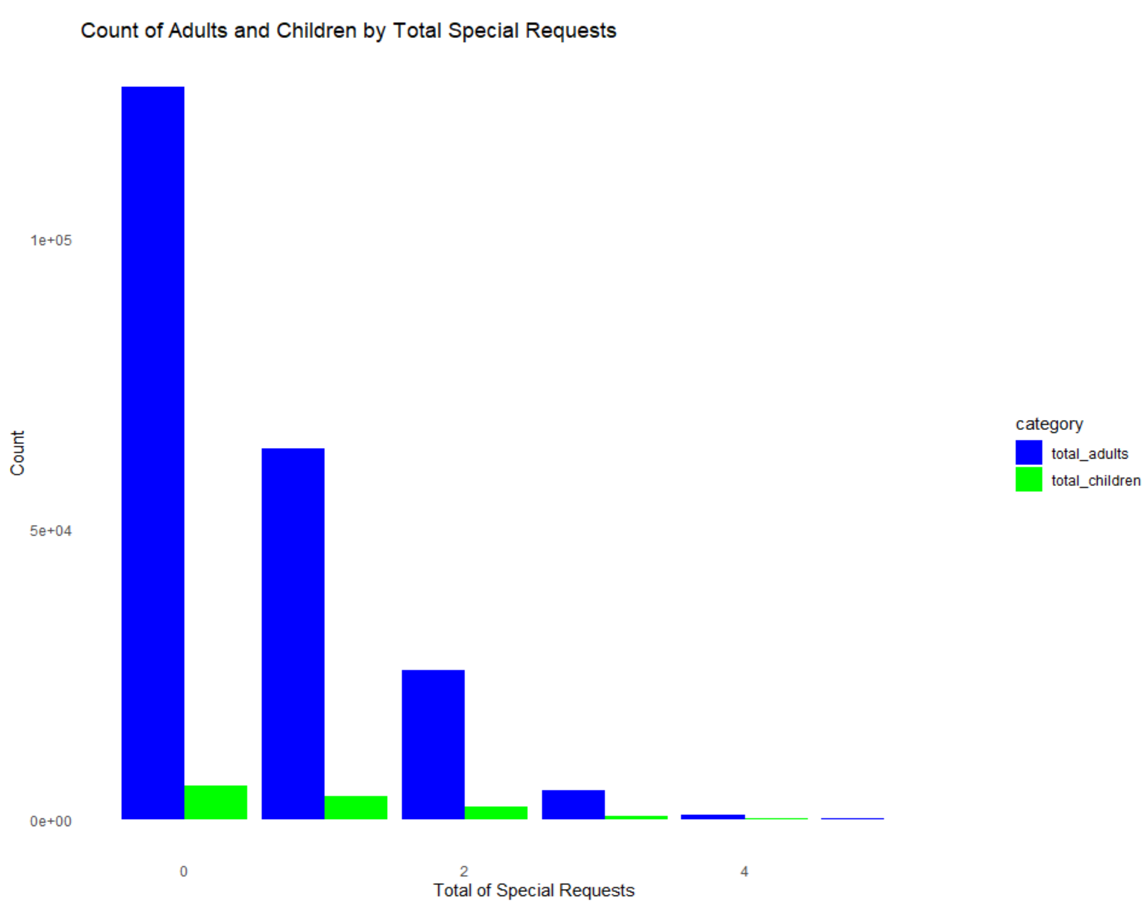
11. How does the total number of booking changes vary across different lead time bins?

```
df <- df %>%
   mutate(lead_time_bin = cut(lead_time,
   breaks = c(0, 50, 100, 150, 200, 250, 300, 350, Inf),
   labels = c("0-50", "51-100", "101-150", "151-200", "201-250",
"251-300", "301-350", "350+"),
       include.lowest = TRUE))

agg_data10 <- df %>%
  group_by(lead_time_bin) %>%
  summarise(total_changes = sum(booking_changes, na.rm = TRUE))

ggplot(agg_data10, aes(x = lead_time_bin, y = total_changes)) +
  geom_bar(stat = "identity", fill = "#00FFFF") +
  geom_text(aes(label = total_changes), vjust = -0.5) +
  labs(x = "Lead Time Bin", y = "Total Number of Changes", title =
"Total Number of Changes by Lead Time Bin") +
  theme_minimal() +
  theme(axis.text.x = element_text(hjust = 0.5),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        plot.background = element_blank())
```



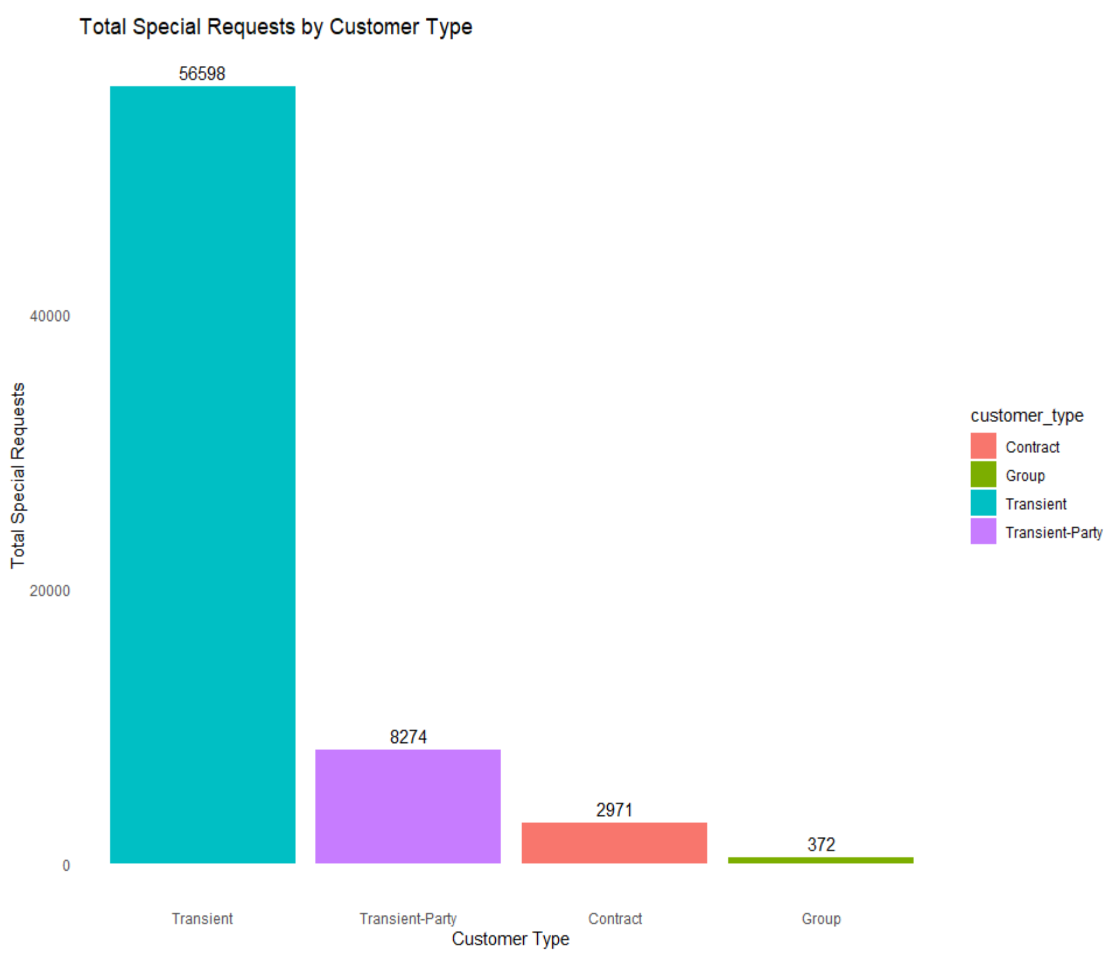
12. How does the number of special requests relate to the number of adults and children?

```
agg_data <- df %>%
  group_by(total_of_special_requests) %>%
  summarise(
    total_adults = sum(adults, na.rm = TRUE),
    total_children = sum(children, na.rm = TRUE)
  )

agg_data_long <- agg_data %>%
  pivot_longer(cols = c(total_adults, total_children),
               names_to = "category",
               values_to = "count")

ggplot(agg_data_long, aes(x = total_of_special_requests, y = count,
fill = category)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Total of Special Requests", y = "Count", title = "Count of
Adults and Children by Total Special Requests") +
  scale_fill_manual(values = c("blue", "green")) +
  theme_minimal() +
  theme(
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank()
  )
```



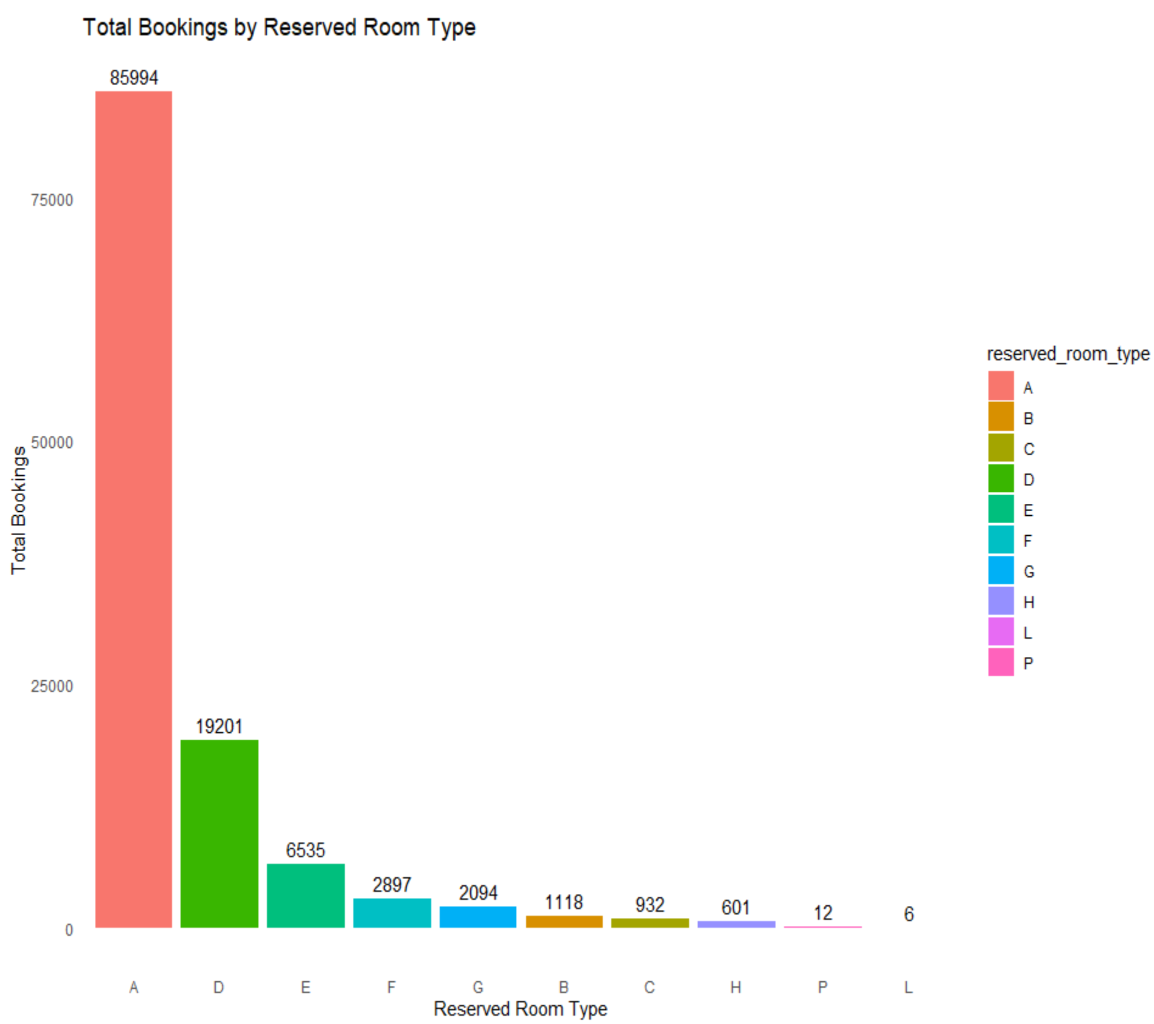
13. Which customer type makes the highest number of special requests?

```
agg_data <- df %>%
  group_by(customer_type) %>%
  summarise(total_special_requests = sum(total_of_special_requests,
na.rm = TRUE))

ggplot(agg_data, aes(x = reorder(customer_type, -
total_special_requests), y = total_special_requests, fill =
customer_type)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = total_special_requests), vjust = -0.5) +
  labs(title = "Total Special Requests by Customer Type",
       x = "Customer Type",
       y = "Total Special Requests") +
  theme_minimal() +
  theme(panel.grid = element_blank())
```



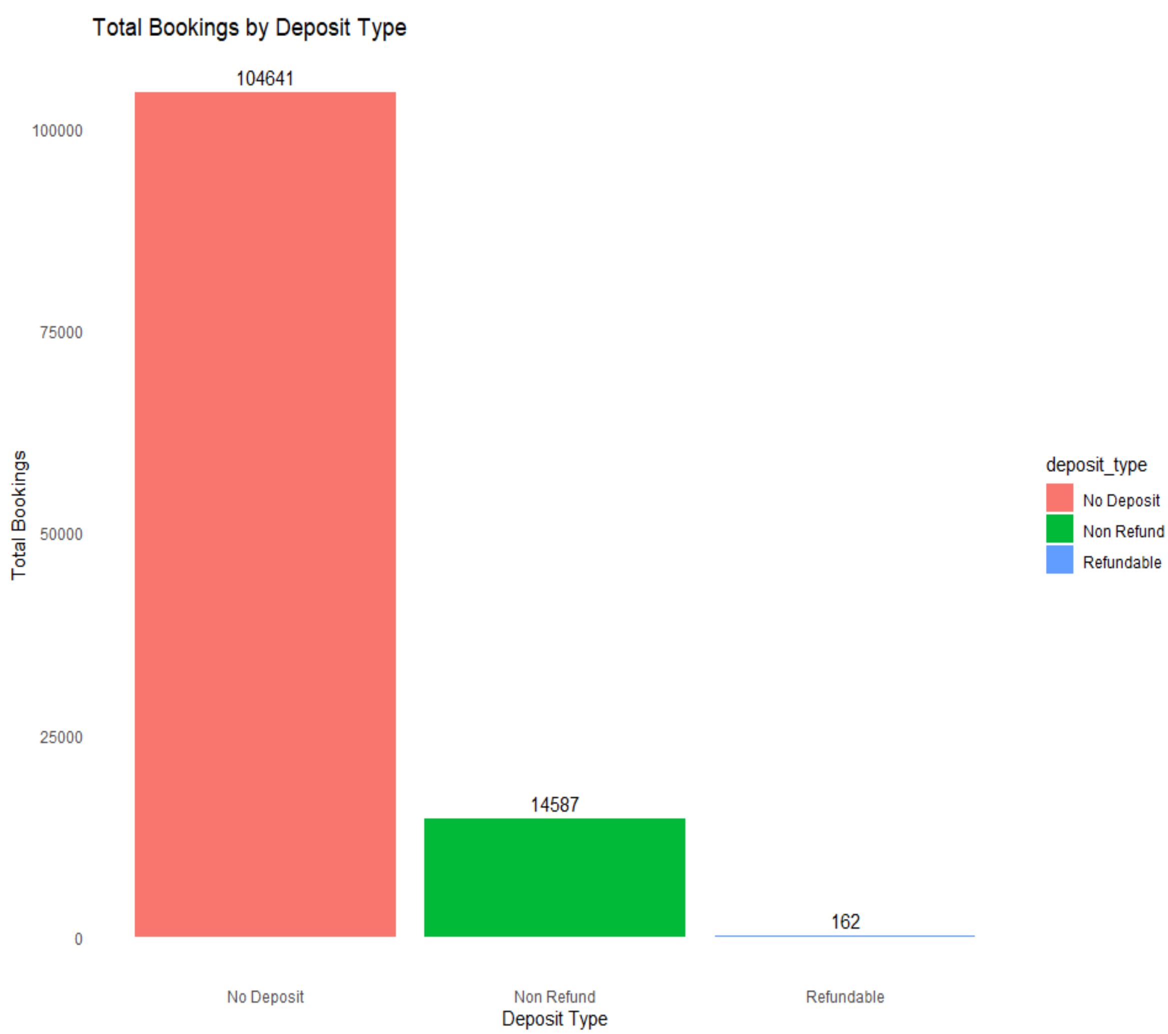
14. Which reserved room type has the highest number of bookings?

```
agg_data <- df %>%
  group_by(reserved_room_type) %>%
  summarise(total_bookings = n())

ggplot(agg_data, aes(x = reorder(reserved_room_type, -total_bookings),
y = total_bookings, fill = reserved_room_type)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = total_bookings), vjust = -0.5) +
  labs(title = "Total Bookings by Reserved Room Type",
       x = "Reserved Room Type",
       y = "Total Bookings") +
  theme_minimal() +
  theme(panel.grid = element_blank())
```



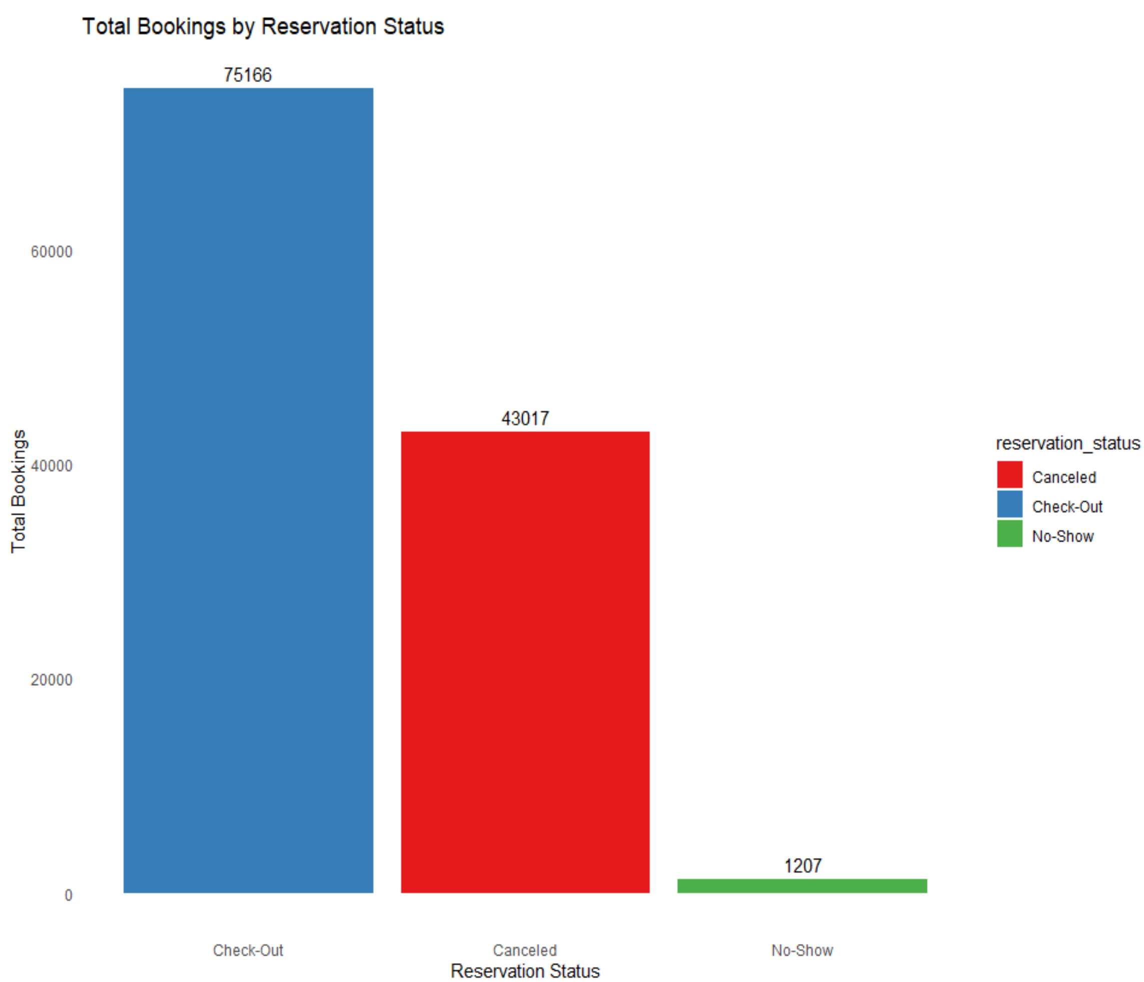
15. Which deposit type is associated with the highest number of bookings?

```
agg_data <- df %>%
  group_by(deposit_type) %>%
  summarise(total_bookings = n())

ggplot(agg_data, aes(x = deposit_type, y = total_bookings, fill =
deposit_type)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = total_bookings), vjust = -0.5) +
  labs(title = "Total Bookings by Deposit Type",
       x = "Deposit Type",
       y = "Total Bookings") +
  theme_minimal() +
  theme(panel.grid = element_blank())
```



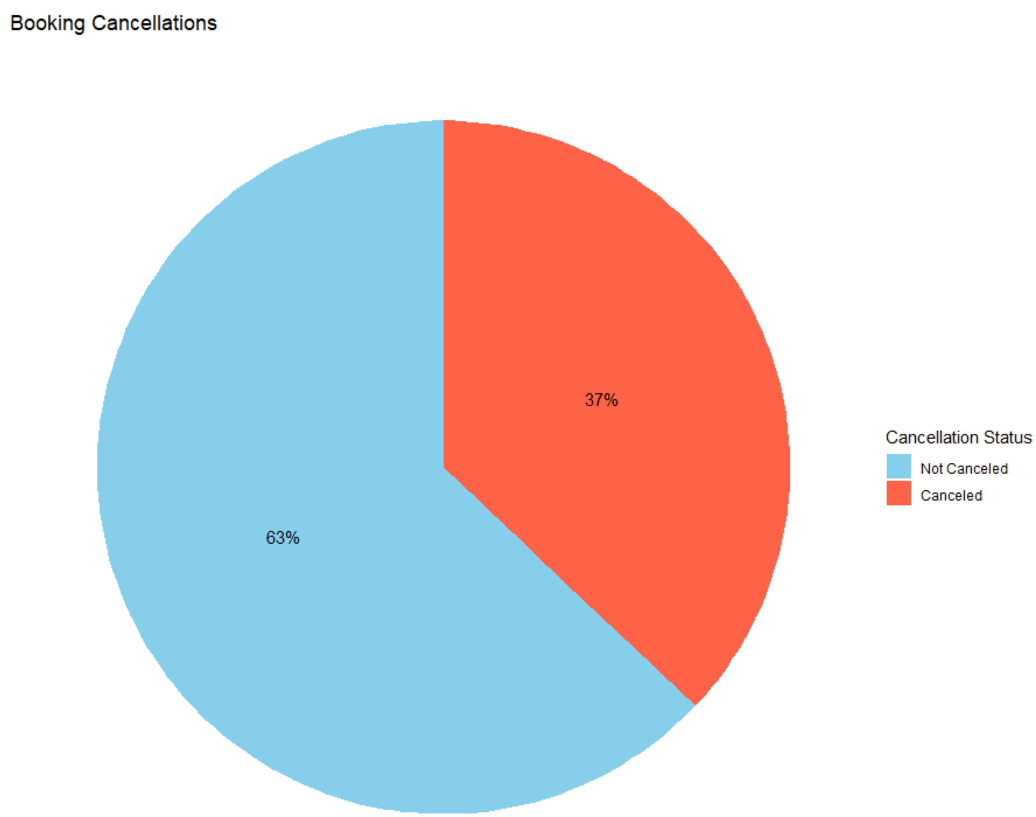
16. Which reservation status category has the highest number of bookings?

```
agg_data <- df %>%
  group_by(reservation_status) %>%
  summarise(total_bookings = n())

ggplot(agg_data, aes(x = reorder(reservation_status, -total_bookings),
y = total_bookings, fill = reservation_status)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = total_bookings), vjust = -0.5) +
  labs(title = "Total Bookings by Reservation Status",
       x = "Reservation Status",
       y = "Total Bookings") +
  theme_minimal() +
  theme(panel.grid = element_blank()) +
  scale_fill_brewer(palette = "Set1")
```



17. What percentage of total bookings are canceled versus not canceled?

```
agg_data <- df %>%
  group_by(is_canceled) %>%
  summarise(total_bookings = n()) %>%
  mutate(percentage = total_bookings / sum(total_bookings) * 100,
         is_canceled_label = factor(is_canceled, levels = c(0, 1),
labels = c("Not Canceled", "Canceled")))

ggplot(agg_data, aes(x = "", y = percentage, fill = is_canceled_label))
+
 geom_bar(stat = "identity", width = 1) +
 coord_polar("y") +
 geom_text(aes(label = paste0(round(percentage, 1), "%")), position =
position_stack(vjust = 0.5)) +
  labs(title = "Booking Cancellations", fill = "Cancellation Status") +
  theme_void() +
  theme(legend.position = "right") +
  scale_fill_manual(values = c("Not Canceled" = "skyblue", "Canceled" =
"tomato"))
```



Conclusion for the Hotel Bookings Data Analysis

After performing a detailed analysis of the hotel bookings dataset, the following key insights have emerged

Cancellation Patterns

A significant portion of bookings were canceled, highlighting a potential issue with booking reliability. Understanding the reasons behind cancellations (e.g., long lead times or booking policies) could help in mitigating future cancellations. The cancellation rate fluctuates across different years, with certain years showing a higher tendency for cancellations. This trend may be influenced by external factors such as economic conditions or travel restrictions.

Lead Time Impact

Bookings with longer lead times (more days between booking and arrival) have a higher likelihood of being canceled. The majority of cancellations happen when the lead time is beyond 100 days. This suggests that people who book further in advance are more likely to cancel. Conversely, bookings with shorter lead times (e.g., less than 50 days) show fewer cancellations, indicating that last-minute bookings tend to be more reliable.

Seasonality of Cancellations

The month of arrival has a noticeable impact on cancellation rates. Certain months, particularly the peak vacation seasons (e.g., July and August), show higher cancellation rates. This could be due to overbooking or fluctuating vacation plans. Hotels might benefit from implementing more flexible booking or cancellation policies during these high-cancellation periods to reduce booking loss.

Impact of Hotel Type

Resort hotels experience higher lead times and slightly higher cancellation rates compared to city hotels. This is likely because resort bookings are made well in advance for holidays or special events, which are more prone to cancellation as plans change. City hotels tend to attract more last-minute bookings, leading to lower cancellation rates and quicker turnover of rooms.

Yearly Booking Trends

The number of bookings has grown consistently over the years, indicating increasing popularity or market demand for these hotels. Despite this growth, the proportion of cancellations has remained relatively stable, suggesting that the hotels are maintaining a balance between their booking and cancellation rates over time.

Recommendations

Cancellation Policy Adjustments

Consider offering incentives for early bookings with shorter lead times or penalties for cancellations closer to the booking date. Flexible booking options could help reduce cancellations while maintaining customer satisfaction. Marketing Strategies: Target customers who tend to make last-minute bookings, especially during off-peak months, to improve occupancy rates and reduce reliance on bookings with long lead times.

Seasonal Preparations

Be proactive in adjusting pricing and availability during high-cancellation periods such as summer holidays. Overbooking strategies or stricter cancellation policies may be necessary during these months to avoid revenue loss.