Hotel_booking_analysis.R

DELL

2025-05-03

```r
# Load the necessary librarries
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(lubridate)
library(dplyr)
library(ggplot2)
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```r
# Load dataset
hotel_data <- read.csv("C:/Users/DELL/Downloads/hotel_bookings.csv", stringsAsFactors = FALSE)

# View the structure of the data
str(hotel_data)
```

```
## '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" ...
```

```
summary(hotel_data)
```

```
##     hotel            is_canceled       lead_time   arrival_date_year
##  Length:119390      Min.   :0.0000   Min.   :  0   Min.   :2015
##  Class :character   1st Qu.:0.0000   1st Qu.: 18   1st Qu.:2016
##  Mode  :character   Median :0.0000   Median : 69   Median :2016
##                     Mean   :0.3704   Mean   :104   Mean   :2016
##                     3rd Qu.:1.0000   3rd Qu.:160   3rd Qu.:2017
##                     Max.   :1.0000   Max.   :737   Max.   :2017
##
##  arrival_date_month arrival_date_week_number arrival_date_day_of_month
##  Length:119390      Min.   : 1.00            Min.   : 1.0
##  Class :character   1st Qu.:16.00            1st Qu.: 8.0
##  Mode  :character   Median :28.00            Median :16.0
##                     Mean   :27.17            Mean   :15.8
##                     3rd Qu.:38.00            3rd Qu.:23.0
##                     Max.   :53.00            Max.   :31.0
##
##  stays_in_weekend_nights stays_in_week_nights     adults
##  Min.   : 0.0000         Min.   : 0.0         Min.   : 0.000
##  1st Qu.: 0.0000         1st Qu.: 1.0         1st Qu.: 2.000
##  Median : 1.0000         Median : 2.0         Median : 2.000
##  Mean   : 0.9276         Mean   : 2.5         Mean   : 1.856
##  3rd Qu.: 2.0000         3rd Qu.: 3.0         3rd Qu.: 2.000
##  Max.   :19.0000         Max.   :50.0         Max.   :55.000
##
##     children           babies              meal              country
##  Min.   : 0.0000   Min.   : 0.000000   Length:119390      Length:119390
##  1st Qu.: 0.0000   1st Qu.: 0.000000   Class :character   Class :character
##  Median : 0.0000   Median : 0.000000   Mode  :character   Mode  :character
##  Mean   : 0.1039   Mean   : 0.007949
##  3rd Qu.: 0.0000   3rd Qu.: 0.000000
##  Max.   :10.0000   Max.   :10.000000
```

```
##  NA's   :4
##  market_segment     distribution_channel is_repeated_guest
##  Length:119390      Length:119390        Min.   :0.00000
##  Class :character   Class :character     1st Qu.:0.00000
##  Mode  :character   Mode  :character     Median :0.00000
##                                          Mean   :0.03191
##                                          3rd Qu.:0.00000
##                                          Max.   :1.00000
##
##  previous_cancellations previous_bookings_not_canceled reserved_room_type
##  Min.   : 0.00000       Min.   : 0.0000                Length:119390
##  1st Qu.: 0.00000       1st Qu.: 0.0000                Class :character
##  Median : 0.00000       Median : 0.0000                Mode  :character
##  Mean   : 0.08712       Mean   : 0.1371
##  3rd Qu.: 0.00000       3rd Qu.: 0.0000
##  Max.   :26.00000       Max.   :72.0000
##
##  assigned_room_type booking_changes   deposit_type          agent
##  Length:119390      Min.   : 0.0000   Length:119390      Length:119390
##  Class :character   1st Qu.: 0.0000   Class :character   Class :character
##  Mode  :character   Median : 0.0000   Mode  :character   Mode  :character
##                     Mean   : 0.2211
##                     3rd Qu.: 0.0000
##                     Max.   :21.0000
##
##    company          days_in_waiting_list customer_type           adr
##  Length:119390      Min.   :  0.000      Length:119390      Min.   :  -6.38
##  Class :character   1st Qu.:  0.000      Class :character   1st Qu.:  69.29
##  Mode  :character   Median :  0.000      Mode  :character   Median :  94.58
##                     Mean   :  2.321                         Mean   : 101.83
##                     3rd Qu.:  0.000                         3rd Qu.: 126.00
##                     Max.   :391.000                         Max.   :5400.00
##
##  required_car_parking_spaces total_of_special_requests reservation_status
##  Min.   :0.00000             Min.   :0.0000            Length:119390
##  1st Qu.:0.00000             1st Qu.:0.0000            Class :character
##  Median :0.00000             Median :0.0000            Mode  :character
##  Mean   :0.06252             Mean   :0.5714
##  3rd Qu.:0.00000             3rd Qu.:1.0000
##  Max.   :8.00000             Max.   :5.0000
##
##  reservation_status_date
##  Length:119390
##  Class :character
##  Mode  :character
##
##
##
##
```

```
glimpse(hotel_data)
```

```
## Rows: 119,390
## Columns: 32
```

```
## $ hotel                           <chr> "Resort Hotel", "Resort Hotel", "Resort~
## $ is_canceled                     <int> 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, ~
## $ lead_time                       <int> 342, 737, 7, 13, 14, 14, 0, 9, 85, 75, ~
## $ arrival_date_year               <int> 2015, 2015, 2015, 2015, 2015, 2015, 201~
## $ arrival_date_month              <chr> "July", "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, 1, 1, 1, ~
## $ stays_in_weekend_nights         <int> 0, 0, 0, 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, 4, 4, 4, ~
## $ adults                          <int> 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, ~
## $ children                        <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ babies                          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ meal                            <chr> "BB", "BB", "BB", "BB", "BB", "BB", "BB~
## $ country                         <chr> "PRT", "PRT", "GBR", "GBR", "GBR", "GBR~
## $ market_segment                  <chr> "Direct", "Direct", "Direct", "Corporat~
## $ distribution_channel            <chr> "Direct", "Direct", "Direct", "Corporat~
## $ is_repeated_guest               <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ previous_cancellations          <int> 0, 0, 0, 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, 0, 0, 0, ~
## $ reserved_room_type              <chr> "C", "C", "A", "A", "A", "A", "C", "C",~
## $ assigned_room_type              <chr> "C", "C", "C", "A", "A", "A", "C", "C",~
## $ booking_changes                 <int> 3, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ deposit_type                    <chr> "No Deposit", "No Deposit", "No Deposit~
## $ agent                           <chr> "NULL", "NULL", "NULL", "304", "240", "~
## $ company                         <chr> "NULL", "NULL", "NULL", "NULL", "NULL",~
## $ days_in_waiting_list            <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ customer_type                   <chr> "Transient", "Transient", "Transient", ~
## $ adr                             <dbl> 0.00, 0.00, 75.00, 75.00, 98.00, 98.00,~
## $ required_car_parking_spaces     <int> 0, 0, 0, 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, 0, 0, 3, ~
## $ reservation_status              <chr> "Check-Out", "Check-Out", "Check-Out", ~
## $ reservation_status_date         <chr> "2015-07-01", "2015-07-01", "2015-07-02~
```

```
colnames(hotel_data)
```

```
##  [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"
```

```r
# Data cleaning
# Check for missing values
colSums(is.na(hotel_data))
```

```
##                          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
```

```r
# Remove the rows with missing data
hotel_data <- na.omit(hotel_data)

# Convert month names to numbers
hotel_data$arrival_date_month <- match(tolower(hotel_data$arrival_date_month), tolower(month.name))

# Combine year, month, and day into a single date
hotel_data$arrival_date <- make_date(
  year = hotel_data$arrival_date_year,
  month = hotel_data$arrival_date_month,
  day = hotel_data$arrival_date_day_of_month
)


# Convert date columns
# arrival date column is converted into format '2015-07-01'
```

```r
hotel_data$arrival_date <- with(hotel_data, paste(arrival_date_year, arrival_date_month, arrival_date_da
hotel_data$arrival_date <- ymd(hotel_data$arrival_date)

# Remove the unnecessary columns
# Because some columns are used to create the arival_date columns so after that those
# columns are removed
hotel_data <- hotel_data %>%
  select(-c(arrival_date_year, arrival_date_month, arrival_date_day_of_month, agent, company,
            reservation_status_date))

# Confirm changes
glimpse(hotel_data)
```

```
## Rows: 119,386
## Columns: 27
## $ hotel                         <chr> "Resort Hotel", "Resort Hotel", "Resort~
## $ is_canceled                   <int> 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, ~
## $ lead_time                     <int> 342, 737, 7, 13, 14, 14, 0, 9, 85, 75, ~
## $ arrival_date_week_number      <int> 27, 27, 27, 27, 27, 27, 27, 27, 27, 27,~
## $ stays_in_weekend_nights       <int> 0, 0, 0, 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, 4, 4, 4, ~
## $ adults                        <int> 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, ~
## $ children                      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ babies                        <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ meal                          <chr> "BB", "BB", "BB", "BB", "BB", "BB", "BB~
## $ country                       <chr> "PRT", "PRT", "GBR", "GBR", "GBR", "GBR~
## $ market_segment                <chr> "Direct", "Direct", "Direct", "Corporat~
## $ distribution_channel          <chr> "Direct", "Direct", "Direct", "Corporat~
## $ is_repeated_guest             <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ previous_cancellations        <int> 0, 0, 0, 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, 0, 0, 0, ~
## $ reserved_room_type            <chr> "C", "C", "A", "A", "A", "A", "C", "C",~
## $ assigned_room_type            <chr> "C", "C", "C", "A", "A", "A", "C", "C",~
## $ booking_changes               <int> 3, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ deposit_type                  <chr> "No Deposit", "No Deposit", "No Deposit~
## $ days_in_waiting_list          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ customer_type                 <chr> "Transient", "Transient", "Transient", ~
## $ adr                           <dbl> 0.00, 0.00, 75.00, 75.00, 98.00, 98.00,~
## $ required_car_parking_spaces   <int> 0, 0, 0, 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, 0, 0, 3, ~
## $ reservation_status            <chr> "Check-Out", "Check-Out", "Check-Out", ~
## $ arrival_date                  <date> 2015-07-01, 2015-07-01, 2015-07-01, 20~
```

```r
# Check the unique values in some categorical columns
# Check unique values in some categorical columns
unique(hotel_data$meal)
```

```
## [1] "BB"        "FB"        "HB"        "SC"        "Undefined"
```

```r
unique(hotel_data$market_segment)
```

```
## [1] "Direct"        "Corporate"     "Online TA"     "Offline TA/TO"
## [5] "Complementary" "Groups"        "Aviation"
```

```r
unique(hotel_data$distribution_channel)
```

```
## [1] "Direct"    "Corporate" "TA/TO"     "Undefined" "GDS"
```

```r
# Fix meal type 'Undefined'
hotel_data$meal[hotel_data$meal == "Undefined"] <- "SC"

# Fix inconsistent customer type and  market_segment
hotel_data$market_segment <- trimws(tolower(hotel_data$market_segment))
hotel_data$distribution_channel <- trimws(tolower(hotel_data$distribution_channel))
hotel_data$distribution_channel[hotel_data$distribution_channel == "undefined"] <- "unknown"

# Create some new features from existing columns
# Total guests
hotel_data$total_guests <- hotel_data$adults + hotel_data$children + hotel_data$babies

# Length of stay
hotel_data$total_nights <- hotel_data$stays_in_weekend_nights + hotel_data$stays_in_week_nights

# Average revenue per stay
hotel_data$revenue_per_stay <- hotel_data$adr * hotel_data$total_nights

# Convert categorical variables to factor
hotel_data <- hotel_data %>%
  mutate(across(c(hotel, meal, market_segment, distribution_channel,
                  customer_type, deposit_type, reserved_room_type,
                  assigned_room_type, reservation_status), as.factor))

# Remove invalid or impossible values
# Remove bookings with zero guests
hotel_data <- hotel_data %>% filter(total_guests > 0)

# Remove negative or unrealistic 'adr'
hotel_data <- hotel_data %>% filter(adr >= 0)

# Give columns meaningful names

colnames(hotel_data) <- c(
  "Hotel", "IsCanceled", "LeadTime","ArrivalDateWeekNumber", "StaysInWeekendNights", "StaysInWeekNights
  "Adults", "Children", "Babies", "Meal", "Country", "MarketSegment",
  "DistributionChannel", "IsRepeatedGuest", "PreviousCancellations",
  "PreviousBookingsNotCanceled", "ReservedRoomType", "AssignedRoomType",
  "BookingChanges", "DepositType", "DaysInWaitingList",
  "CustomerType", "ADR", "RequiredCarParkingSpaces", "TotalOfSpecialRequests",
  "ReservationStatus", "ArrivalDate", "TotalGuests", "TotalNights", "RevenuePerStay"
)


# Confirm changes
colnames(hotel_data)
```

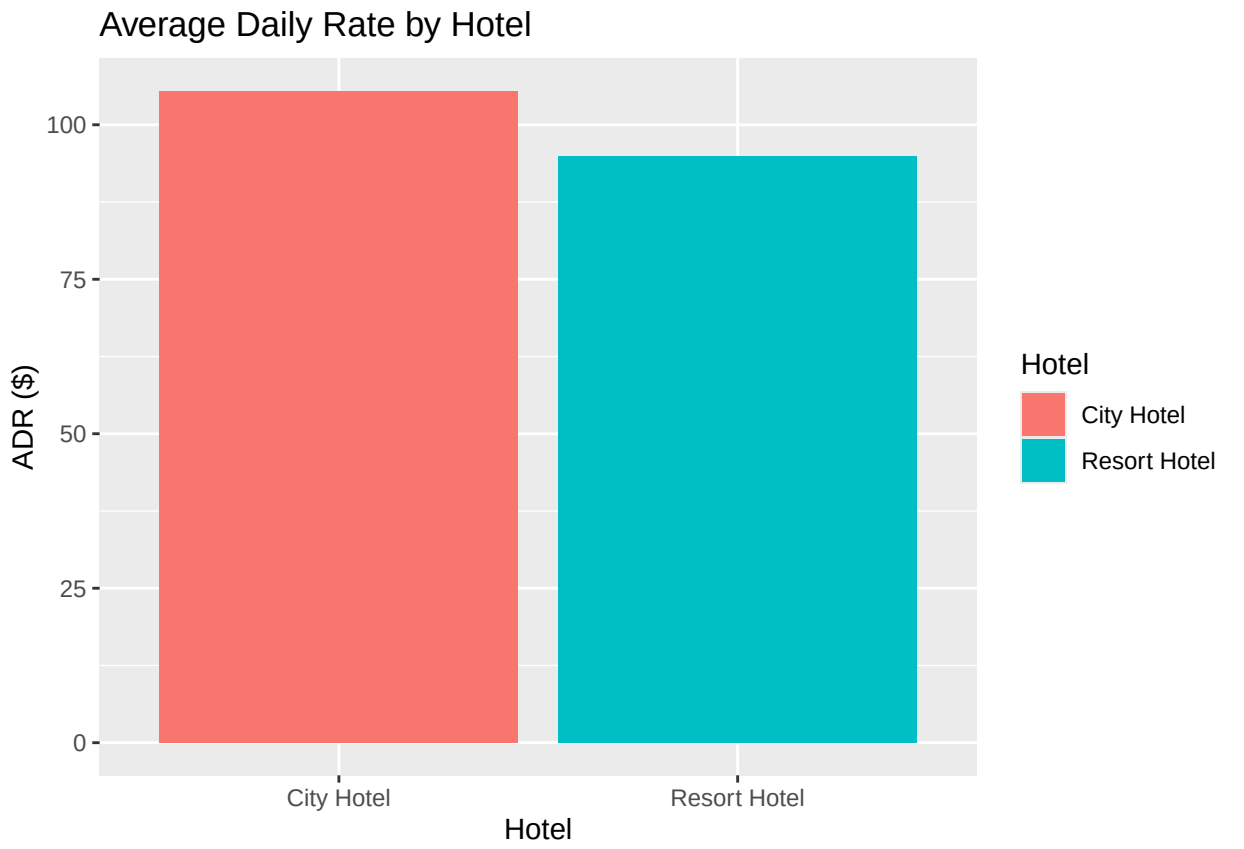
```
##  [1] "Hotel"                      "IsCanceled"
##  [3] "LeadTime"                   "ArrivalDateWeekNumber"
```

```
##  [5] "StaysInWeekendNights"        "StaysInWeekNights"
##  [7] "Adults"                      "Children"
##  [9] "Babies"                      "Meal"
## [11] "Country"                     "MarketSegment"
## [13] "DistributionChannel"         "IsRepeatedGuest"
## [15] "PreviousCancellations"       "PreviousBookingsNotCanceled"
## [17] "ReservedRoomType"            "AssignedRoomType"
## [19] "BookingChanges"              "DepositType"
## [21] "DaysInWaitingList"           "CustomerType"
## [23] "ADR"                         "RequiredCarParkingSpaces"
## [25] "TotalOfSpecialRequests"      "ReservationStatus"
## [27] "ArrivalDate"                 "TotalGuests"
## [29] "TotalNights"                 "RevenuePerStay"
```

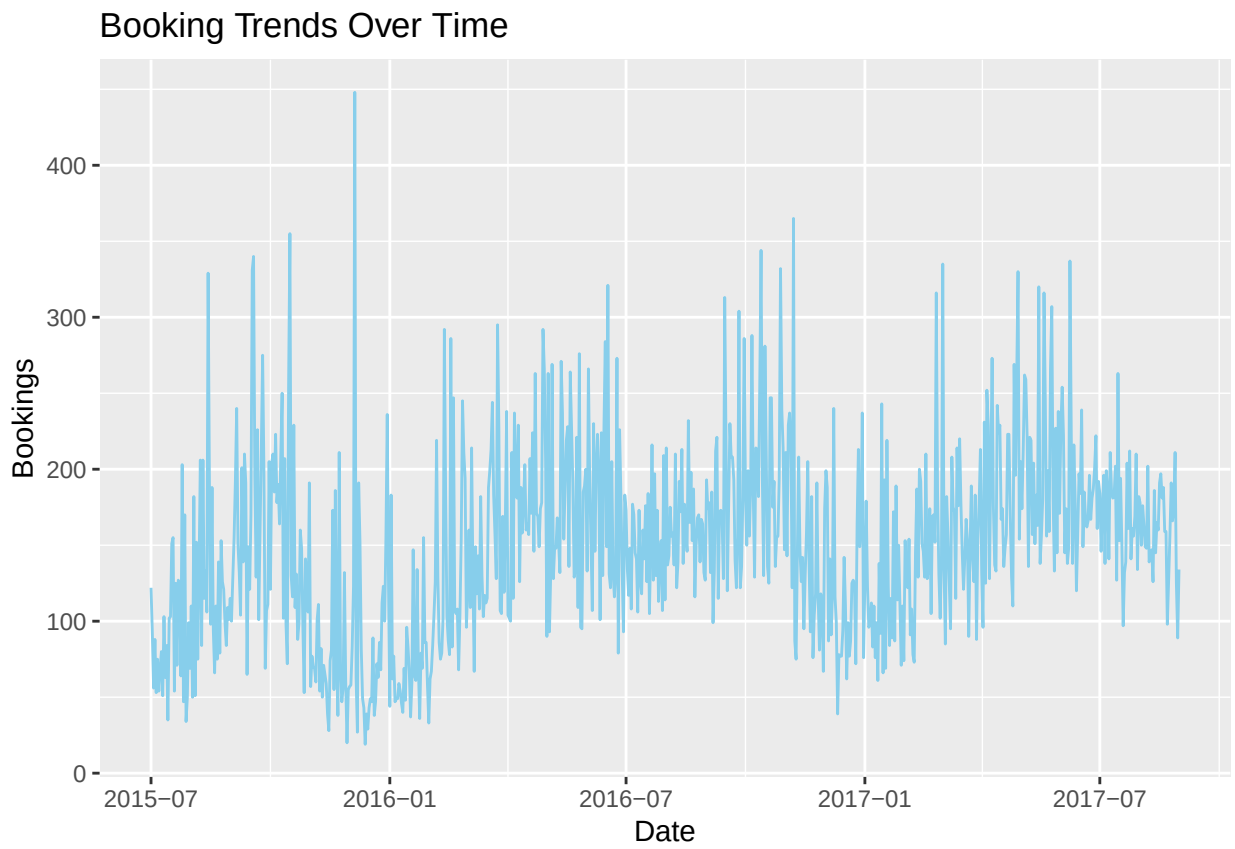
```r
# Now perform data analysis
# Booking distribution by hotel
hotel_data %>%
  count(Hotel) %>%
  ggplot(aes(x = Hotel, y = n, fill = Hotel)) +
  geom_col() +
  labs(title = "Number of Bookings by Hotel", y = "Bookings")
```


```r
# Average Daily Rate by Hotel type
hotel_data %>%
  group_by(Hotel) %>%
```

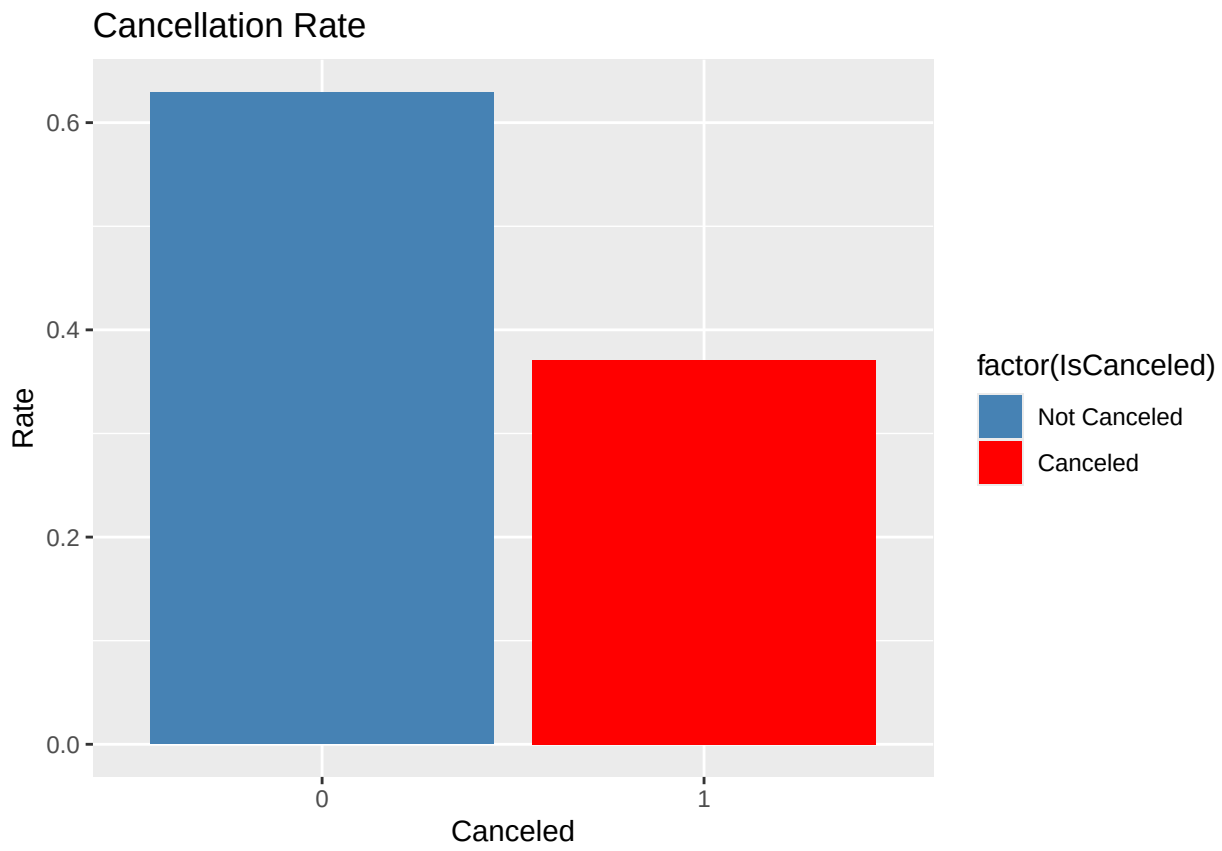
```
  summarise(avg_adr = mean(ADR, na.rm = TRUE)) %>%
  ggplot(aes(x = Hotel, y = avg_adr, fill = Hotel)) +
  geom_col() +
  labs(title = "Average Daily Rate by Hotel", y = "ADR ($)")
```




```
# Booking trends over time
hotel_data %>%
  group_by(ArrivalDate) %>%
  summarise(bookings = n()) %>%
  ggplot(aes(x = ArrivalDate, y = bookings)) +
  geom_line(color = "skyblue") +
  labs(title = "Booking Trends Over Time", x = "Date", y = "Bookings")
```



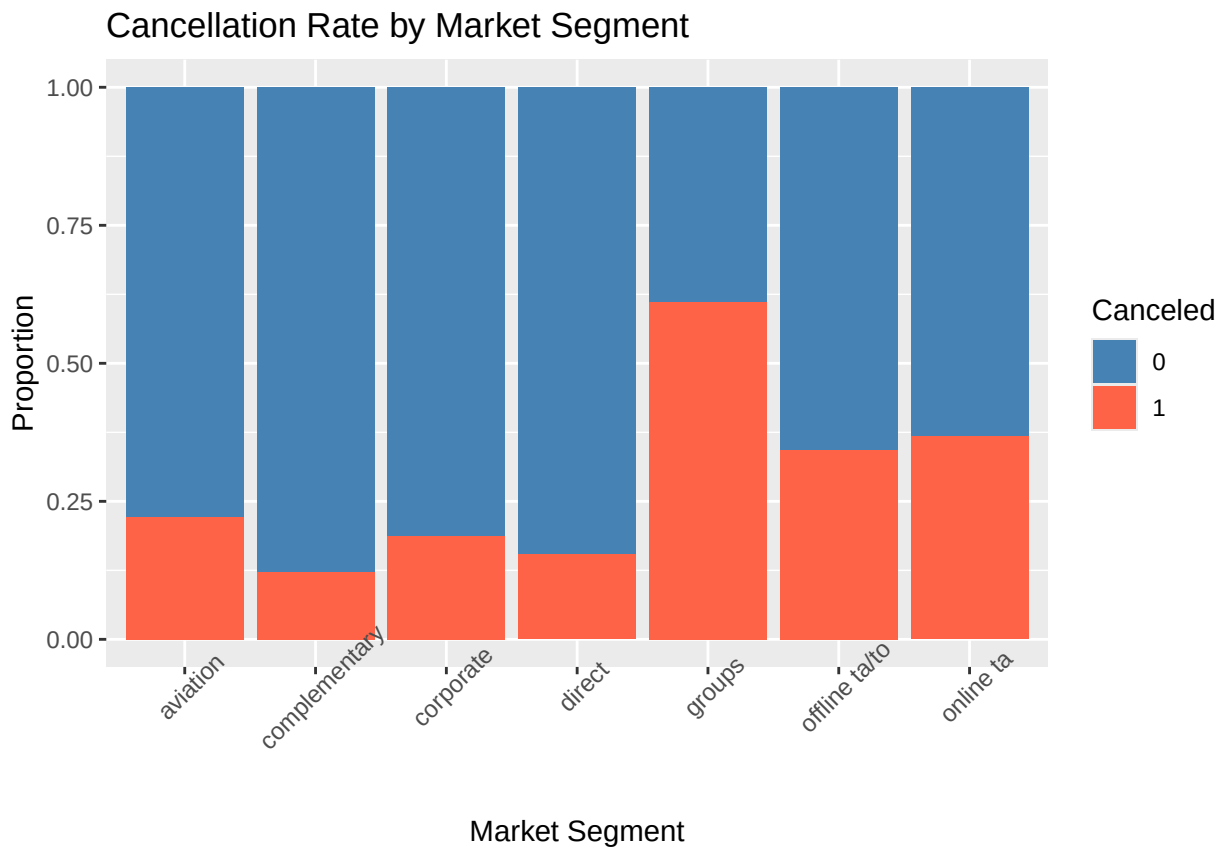

```r
# Cancellation rate
hotel_data %>%
  count(IsCanceled) %>%
  mutate(rate = n / sum(n)) %>%
  ggplot(aes(x = factor(IsCanceled), y = rate, fill = factor(IsCanceled))) +
  geom_col() +
  scale_fill_manual(values = c("steelblue", "red"), labels = c("Not Canceled", "Canceled")) +
  labs(title = "Cancellation Rate", x = "Canceled", y = "Rate")
```




```
# Cancellation rate by segment
hotel_data %>%
  group_by(MarketSegment) %>%
  summarise(
    CancellationRate = mean(IsCanceled),
    Bookings = n()
  ) %>%
  arrange(desc(CancellationRate))
```

```
## # A tibble: 7 x 3
##   MarketSegment CancellationRate Bookings
##   <fct>                    <dbl>    <int>
## 1 groups                   0.611    19790
## 2 online ta                0.368    56407
## 3 offline ta/to            0.343    24182
## 4 aviation                 0.221      235
## 5 corporate                0.188     5282
## 6 direct                   0.154    12581
## 7 complementary            0.122      728
```

```
ggplot(hotel_data, aes(x = MarketSegment, fill = as.factor(IsCanceled))) +
  geom_bar(position = "fill") +
  ylab("Proportion") + xlab("Market Segment") +
  ggtitle("Cancellation Rate by Market Segment") +
  scale_fill_manual(values = c("0" = "steelblue", "1" = "tomato"), name = "Canceled") +
  theme(axis.text.x = element_text(angle = 45))
```




```
# Average revenue per customer type
hotel_data %>%
  group_by(CustomerType) %>%
  summarise(
    AvgADR = mean(ADR, na.rm = TRUE),
    AvgNights = mean(TotalNights, na.rm = TRUE),
    AvgRevenue = mean(RevenuePerStay, na.rm = TRUE),
    Count = n()
  ) %>%
  arrange(desc(AvgRevenue))
```

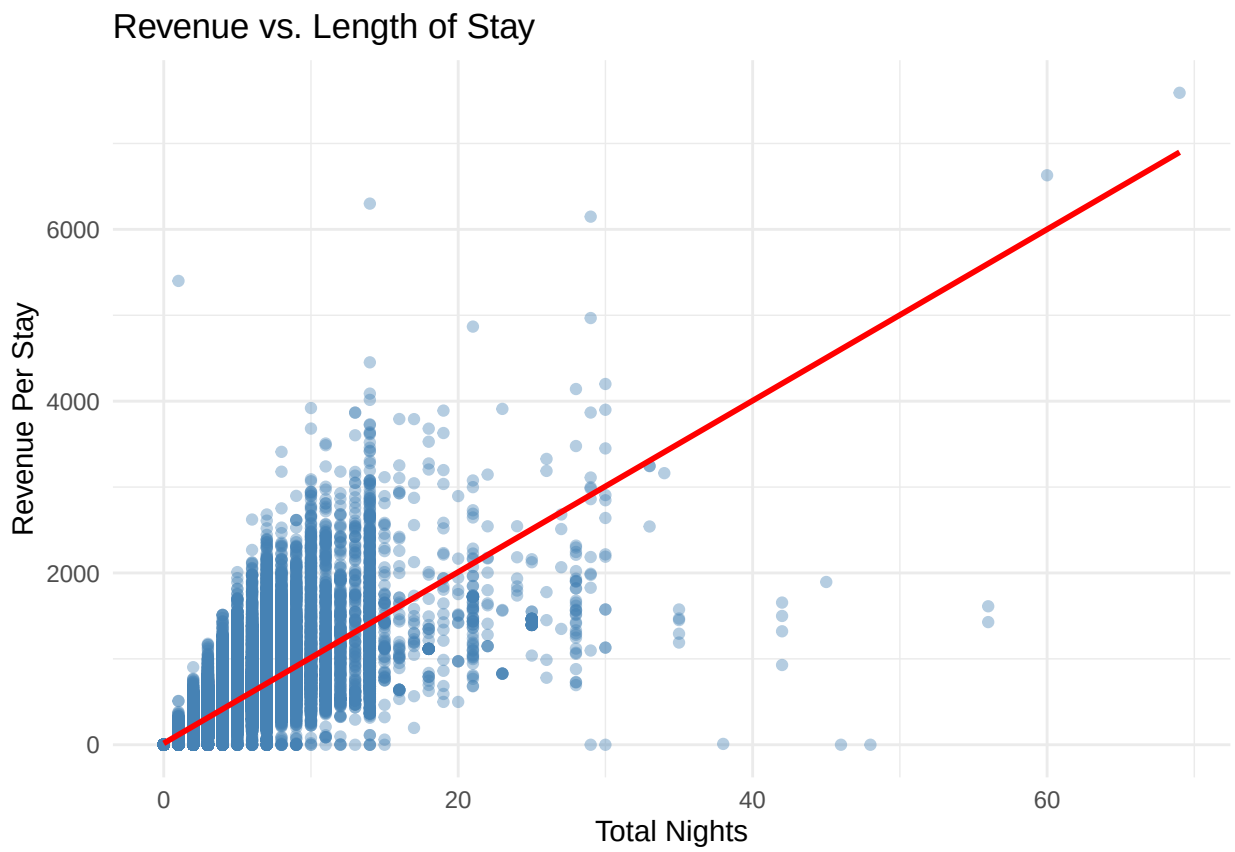
```
## # A tibble: 4 x 5
##   CustomerType    AvgADR AvgNights AvgRevenue Count
##   <fct>            <dbl>     <dbl>      <dbl> <int>
## 1 Contract          87.6      5.32       452.  4072
## 2 Transient        107.       3.45       382. 89476
## 3 Transient-Party   86.2      3.06       261. 25083
## 4 Group             83.9      2.86       245.   574
```

```
# Does Longer Stay leads to higher revenue

ggplot(hotel_data, aes(x = TotalNights, y = RevenuePerStay)) +
  geom_point(alpha = 0.4, color = "steelblue") +
```

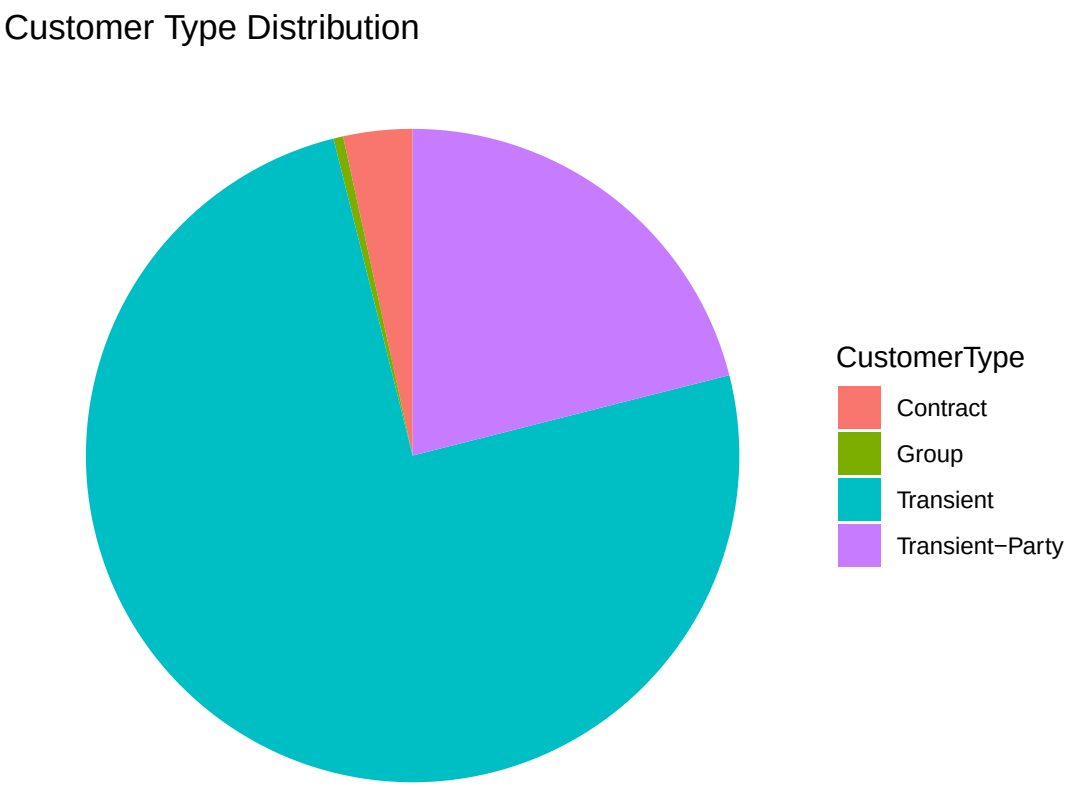
```
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Revenue vs. Length of Stay", x = "Total Nights", y = "Revenue Per Stay") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# What proportion of customers are transient, group etc.
customer_dist <- hotel_data %>%
  group_by(CustomerType) %>%
  summarise(Count = n())

ggplot(customer_dist, aes(x = "", y = Count, fill = CustomerType)) +
  geom_col(width = 1) +
  coord_polar(theta = "y") +
  labs(title = "Customer Type Distribution") +
  theme_void()
```


Customer Type Distribution
CustomerType
Contract
Group
Transient
Transient−Party