Predictive Analytics for Hotel Room Demand & Price Forecasting

Vosnak, Christina

2024-06-01

Introduction

For this project, I will be analyzing various data sets regarding historical information about hotel reservation data to predict room demand & pricing for the future. This topic is important for hotels to maximize their room occupancy rates while also maximizing revenues and profits across the resort. I find this topic to be interesting because it allows hospitality leaders to gain a competitive advantage in the market if they accurately understand their historical data to predict market trends for the future. This information can aid the allocation of resources through various periods of the year in areas such as staffing, inventory, and operational budgets. Overall, this information should be able to assist in minimizing resource costs and maximizing operational revenue and profits.

This is a data science problem as it analyzes large volumes of data regarding historical booking information, customer demographics and market trends. Hotel demand is influenced by a wide range of variables including seasonality, economic indicators, and competitive pricing. Data science models will be able to analyze this data to identify patterns to make accurate decisions for the future.

Research Questions

1. How can historical booking data and market trends be leveraged to develop accurate predictive models for hotel room demand forecasting?
2. What impact do seasonal variations and holidays have on hotel demand?
3. How do economic indicators such as GDP, inflation rates, and consumer spending patterns influence hotel demand?
4. What impact do seasonal variations and holidays have on hotel room demand?
5. What role does customer segmentation based on demographics, preferences, and booking behaviors play in improving the accuracy of demand forecasts?
6. How can social media and review data be leveraged to understand evolving consumer demand?
7. What strategies can hotels use to dynamically adjust pricing and promotional materials based on demand forecasts, competitor pricing, and market trends?
8. How can data-driven approaches enhance collaboration between revenue management teams, marketing departments, and operational staff to optimize resource allocation and maximize revenue opportunities?

Approach

To address predictive analytics for hotel room demand and price forecasting, I will collect information from various data sets that explore historical booking data, market trends, economic indicators and customer demographics. I will perform an exploratory data analysis to visualize the information to understand patterns, correlations, and outliers within the data. I will clean up and condense the data based on it's relativity to the topic and will propse a model to use that will be most accurate for price and demand forecasting in the future.

How your approach addresses the problem

This approach will successfully address the problem by allowing for dynamic research, complex analysis, and adjustable model testing. The results will be seen in the future after the implementation of the model when hotels accurately adjust their resource allocation budgets to fit the model and future data shows a trend of increasing revenue and profits.

Data Overview

Kaggle Hotel Booking Demand Dataset: Mostipak, J. (n.d.). Hotel booking demand [Dataset]. Kaggle. Retrieved from https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand

This data set contains booking information for a city hotel and a resort hotel, and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things. There are quite a few variables to understand room pricing and hotel demand. The data was collected in 2019, and there are 32 columns of data to work with to understand hotel demand. This is the bulk of the data that will be analyzed for this project.

STR Industry Trend Report: STR. (n.d.). Industry Trend Report [Report]. Retrieved from https://str.com/data-solutions/industry-trend-report/

This report provides performance metrics regarding tourism and international travel that should correlate to hotel data.I will be using a sampling of the report to show which destinations are most popular for hotels to capitalize on regarding consumer demand. It contains information relative to the past year.

Kaggle International Tourism Receipts Dataset: Celik, A. (n.d.). International Tourism Receipts [Dataset]. Kaggle. Retrieved from https://www.kaggle.com/datasets/abdulhamitcelik/international-tourism-receipts

This data set contains information about international tourism receipts for various countries. International tourism receipts refer to the expenditures made by international visitors on their trips to a country, including accommodation, food and beverage, transportation, and other tourism-related expenses. This data can help researchers, policymakers, and businesses gain insights into the tourism industry and its impact on economies around the world.

The data set includes information on tourism receipts for over 100 countries from the year 1995 to 2020.

Required Packages

For this project, I will be utilizing various packages to conduct my exploratory data analysis. I will use various packages from the tidyverse including dplyr for data manipulation, ggplot 2 for data visualization and tidyr for data tidying. I will also use the Metrics package to conduct an RMSE analysis.

Plots and Table Needs

For this project I will be using a series of plots to show historical trends in hotel room demand over time highlighting seasonal variations and long-term trends.I will also develop forecast models and compare them to the actual plots through a regression analysis. I will also develop a pricing optimization plot to visualize revenue projections based on different pricing strategies derived from demand forecasts and competitor pricing data.

Questions for future steps

I need further develop my understanding of EDA to clean up my data so that it reports the most accurate results. With having quite a few variables to analyze the data, I want to make sure to remove outliers effectively to not skew any of the information.

Import and Clean Data with Final Visuals

```
# Set Working Directory
setwd("C:/Users/chris/OneDrive/Documents/DSC 520/Term Project")

# Import Data
hotelBookings <- read.csv("hotel_bookings.csv", header=TRUE, sep=",")
str(hotelBookings)
```

```
## 'data.frame':    119390 obs. of  35 variables:
##  $ hotel                         : chr  "Resort Hotel" "Resort Hotel" "Resort Hotel" "Resort Hotel"
##  $ arrival_full_date             : chr  "2015-July-01" "2015-July-01" "2015-July-06" "2015-July-06"
##  $ total_stay                    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ total_guests                  : int  2 2 2 1 2 2 2 1 2 4 ...
##  $ adr                           : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ is_canceled                   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ lead_time                     : int  342 737 111 0 8 8 6 0 0 16 ...
##  $ arrival_date_year             : int  2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...
##  $ arrival_date_month            : chr  "July" "July" "July" "July" ...
##  $ arrival_date_day_of_month     : int  1 1 6 6 7 7 17 20 20 23 ...
##  $ arrival_date_week_number      : int  27 27 28 28 28 28 29 30 30 30 ...
##  $ stays_in_weekend_nights       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ stays_in_week_nights          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ adults                        : int  2 2 2 1 2 2 2 1 2 2 ...
##  $ children                      : int  0 0 0 0 0 0 0 0 0 2 ...
##  $ babies                        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ meal                          : chr  "BB" "BB" "BB" "BB" ...
##  $ country                       : chr  "PRT" "PRT" "PRT" "PRT" ...
##  $ market_segment                : chr  "Direct" "Direct" "Online TA" "Direct" ...
##  $ distribution_channel          : chr  "Direct" "Direct" "TA/TO" "Direct" ...
##  $ 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" "E" ...
##  $ assigned_room_type            : chr  "C" "C" "H" "H" ...
```

```
##  $ 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" "240" "250" ...
##  $ 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" ...
##  $ required_car_parking_spaces   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ total_of_special_requests     : int  0 0 2 0 1 1 1 0 1 0 ...
##  $ reservation_status            : chr  "Check-Out" "Check-Out" "Check-Out" "Check-Out" ...
##  $ reservation_status_date       : chr  "7/1/2015" "7/1/2015" "7/6/2015" "7/6/2015" ...
```

```r
# Load Librarys
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(tidyr)
library(ggplot2)

# Check for Missing Values and remove them from the Data Set
missingValues <- colSums(is.na(hotelBookings))
print(missingValues)
```

```
##                          hotel              arrival_full_date
##                              0                              0
##                     total_stay                   total_guests
##                              0                              0
##                            adr                    is_canceled
##                              0                              0
##                      lead_time              arrival_date_year
##                              0                              0
##             arrival_date_month      arrival_date_day_of_month
##                              0                              0
##       arrival_date_week_number        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   required_car_parking_spaces
##                              0                              0
##      total_of_special_requests             reservation_status
##                              0                              0
##        reservation_status_date
##                              0
```

```
cleanedData <- na.omit(hotelBookings)

# Remove Duplicates
cleanedData2 <- cleanedData[!duplicated(cleanedData), ]

# Final Visuals
str(cleanedData2)
```

```
## 'data.frame':    87392 obs. of  35 variables:
##  $ hotel                         : chr  "Resort Hotel" "Resort Hotel" "Resort Hotel" "Resort Hotel"
##  $ arrival_full_date             : chr  "2015-July-01" "2015-July-01" "2015-July-06" "2015-July-06"
##  $ total_stay                    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ total_guests                  : int  2 2 2 1 2 2 2 1 2 4 ...
##  $ adr                           : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ is_canceled                   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ lead_time                     : int  342 737 111 0 8 8 6 0 0 16 ...
##  $ arrival_date_year             : int  2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...
##  $ arrival_date_month            : chr  "July" "July" "July" "July" ...
##  $ arrival_date_day_of_month     : int  1 1 6 6 7 7 17 20 20 23 ...
##  $ arrival_date_week_number      : int  27 27 28 28 28 28 29 30 30 30 ...
##  $ stays_in_weekend_nights       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ stays_in_week_nights          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ adults                        : int  2 2 2 1 2 2 2 1 2 2 ...
##  $ children                      : int  0 0 0 0 0 0 0 0 0 2 ...
##  $ babies                        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ meal                          : chr  "BB" "BB" "BB" "BB" ...
##  $ country                       : chr  "PRT" "PRT" "PRT" "PRT" ...
##  $ market_segment                : chr  "Direct" "Direct" "Online TA" "Direct" ...
##  $ distribution_channel          : chr  "Direct" "Direct" "TA/TO" "Direct" ...
##  $ 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" "E" ...
##  $ assigned_room_type            : chr  "C" "C" "H" "H" ...
##  $ 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" "240" "250" ...
```

```
##  $ 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" ...
##  $ required_car_parking_spaces   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ total_of_special_requests     : int  0 0 2 0 1 1 1 0 1 0 ...
##  $ reservation_status            : chr  "Check-Out" "Check-Out" "Check-Out" "Check-Out" ...
##  $ reservation_status_date       : chr  "7/1/2015" "7/1/2015" "7/6/2015" "7/6/2015" ...
##  - attr(*, "na.action")= 'omit' Named int [1:4] 8540 28121 28133 112097
##   ..- attr(*, "names")= chr [1:4] "8540" "28121" "28133" "112097"
```

```
summary(cleanedData2)
```

```
##     hotel           arrival_full_date    total_stay      total_guests
##  Length:87392       Length:87392       Min.   : 0.000   Min.   : 0.000
##  Class :character   Class :character   1st Qu.: 2.000   1st Qu.: 2.000
##  Mode  :character   Mode  :character   Median : 3.000   Median : 2.000
##                                        Mean   : 3.631   Mean   : 2.025
##                                        3rd Qu.: 5.000   3rd Qu.: 2.000
##                                        Max.   :69.000   Max.   :55.000
##       adr           is_canceled       lead_time      arrival_date_year
##  Min.   :  -6.38   Min.   :0.0000   Min.   :  0.00   Min.   :2015
##  1st Qu.:  72.00   1st Qu.:0.0000   1st Qu.: 11.00   1st Qu.:2016
##  Median :  98.10   Median :0.0000   Median : 49.00   Median :2016
##  Mean   : 106.34   Mean   :0.2749   Mean   : 79.89   Mean   :2016
##  3rd Qu.: 134.00   3rd Qu.:1.0000   3rd Qu.:125.00   3rd Qu.:2017
##  Max.   :5400.00   Max.   :1.0000   Max.   :737.00   Max.   :2017
##  arrival_date_month arrival_date_day_of_month arrival_date_week_number
##  Length:87392       Min.   : 1.00             Min.   : 1.00
##  Class :character   1st Qu.: 8.00             1st Qu.:16.00
##  Mode  :character   Median :16.00             Median :27.00
##                     Mean   :15.82             Mean   :26.84
##                     3rd Qu.:23.00             3rd Qu.:37.00
##                     Max.   :31.00             Max.   :53.00
##  stays_in_weekend_nights stays_in_week_nights     adults
##  Min.   : 0.000          Min.   : 0.000       Min.   : 0.000
##  1st Qu.: 0.000          1st Qu.: 1.000       1st Qu.: 2.000
##  Median : 1.000          Median : 2.000       Median : 2.000
##  Mean   : 1.005          Mean   : 2.625       Mean   : 1.876
##  3rd Qu.: 2.000          3rd Qu.: 4.000       3rd Qu.: 2.000
##  Max.   :19.000          Max.   :50.000       Max.   :55.000
##     children           babies             meal              country
##  Min.   : 0.0000   Min.   : 0.00000   Length:87392       Length:87392
##  1st Qu.: 0.0000   1st Qu.: 0.00000   Class :character   Class :character
##  Median : 0.0000   Median : 0.00000   Mode  :character   Mode  :character
##  Mean   : 0.1386   Mean   : 0.01082
##  3rd Qu.: 0.0000   3rd Qu.: 0.00000
##  Max.   :10.0000   Max.   :10.00000
##  market_segment     distribution_channel is_repeated_guest
##  Length:87392       Length:87392         Min.   :0.00000
##  Class :character   Class :character     1st Qu.:0.00000
##  Mode  :character   Mode  :character     Median :0.00000
##                                          Mean   :0.03908
##                                          3rd Qu.:0.00000
##                                          Max.   :1.00000
```

```
##  previous_cancellations previous_bookings_not_canceled reserved_room_type
##  Min.   : 0.00000       Min.   : 0.000                 Length:87392
##  1st Qu.: 0.00000       1st Qu.: 0.000                 Class :character
##  Median : 0.00000       Median : 0.000                 Mode  :character
##  Mean   : 0.03042       Mean   : 0.184
##  3rd Qu.: 0.00000       3rd Qu.: 0.000
##  Max.   :26.00000       Max.   :72.000
##  assigned_room_type booking_changes   deposit_type          agent
##  Length:87392       Min.   : 0.0000   Length:87392       Length:87392
##  Class :character   1st Qu.: 0.0000   Class :character   Class :character
##  Mode  :character   Median : 0.0000   Mode  :character   Mode  :character
##                     Mean   : 0.2716
##                     3rd Qu.: 0.0000
##                     Max.   :21.0000
##    company          days_in_waiting_list customer_type
##  Length:87392       Min.   :  0.0000     Length:87392
##  Class :character   1st Qu.:  0.0000     Class :character
##  Mode  :character   Median :  0.0000     Mode  :character
##                     Mean   :  0.7496
##                     3rd Qu.:  0.0000
##                     Max.   :391.0000
##  required_car_parking_spaces total_of_special_requests reservation_status
##  Min.   :0.00000             Min.   :0.0000            Length:87392
##  1st Qu.:0.00000             1st Qu.:0.0000            Class :character
##  Median :0.00000             Median :0.0000            Mode  :character
##  Mean   :0.08423             Mean   :0.6985
##  3rd Qu.:0.00000             3rd Qu.:1.0000
##  Max.   :8.00000             Max.   :5.0000
##  reservation_status_date
##  Length:87392
##  Class :character
##  Mode  :character
##
##
##
```

```
head(cleanedData2)
```

```
##          hotel arrival_full_date total_stay total_guests adr is_canceled
## 1 Resort Hotel      2015-July-01          0            2   0           0
## 2 Resort Hotel      2015-July-01          0            2   0           0
## 3 Resort Hotel      2015-July-06          0            2   0           0
## 4 Resort Hotel      2015-July-06          0            1   0           0
## 5 Resort Hotel      2015-July-07          0            2   0           0
## 6 Resort Hotel      2015-July-07          0            2   0           0
##   lead_time arrival_date_year arrival_date_month arrival_date_day_of_month
## 1       342              2015               July                         1
## 2       737              2015               July                         1
## 3       111              2015               July                         6
## 4         0              2015               July                         6
## 5         8              2015               July                         7
## 6         8              2015               July                         7
##   arrival_date_week_number stays_in_weekend_nights stays_in_week_nights adults
## 1                       27                       0                    0      2
```

```
## 2                      27                       0                    0      2
## 3                      28                       0                    0      2
## 4                      28                       0                    0      1
## 5                      28                       0                    0      2
## 6                      28                       0                    0      2
##   children babies meal country market_segment distribution_channel
## 1        0      0   BB     PRT         Direct               Direct
## 2        0      0   BB     PRT         Direct               Direct
## 3        0      0   BB     PRT      Online TA                TA/TO
## 4        0      0   BB     PRT         Direct               Direct
## 5        0      0   BB     PRT         Direct               Direct
## 6        0      0   BB     PRT      Online TA                TA/TO
##   is_repeated_guest previous_cancellations previous_bookings_not_canceled
## 1                 0                       0                              0
## 2                 0                       0                              0
## 3                 0                       0                              0
## 4                 0                       0                              0
## 5                 0                       0                              0
## 6                 0                       0                              0
##   reserved_room_type assigned_room_type booking_changes deposit_type agent
## 1                  C                  C               3   No Deposit  NULL
## 2                  C                  C               4   No Deposit  NULL
## 3                  A                  H               0   No Deposit   240
## 4                  E                  H               0   No Deposit   250
## 5                  A                  A               0   No Deposit  NULL
## 6                  A                  E               0   No Deposit   240
##   company days_in_waiting_list customer_type required_car_parking_spaces
## 1    NULL                    0     Transient                            0
## 2    NULL                    0     Transient                            0
## 3    NULL                    0     Transient                            0
## 4    NULL                    0     Transient                            0
## 5    NULL                    0     Transient                            0
## 6    NULL                    0     Transient                            0
##   total_of_special_requests reservation_status reservation_status_date
## 1                         0          Check-Out                7/1/2015
## 2                         0          Check-Out                7/1/2015
## 3                         2          Check-Out                7/6/2015
## 4                         0          Check-Out                7/6/2015
## 5                         1          Check-Out                7/7/2015
## 6                         1          Check-Out                7/7/2015
```

In this section, I imported the data set into R Studio, and printed a Structure Function that shows there are 119386 observations and 32 variables in the original data set. To clean the data, I removed any missing values that were found within the columns. The function above shows there were 4 lines that had missing data in the "Children" column that I removed from the data set. I also removed duplicates which brought the number of observations down from 119386 to 87392. To record the final data, I printed a new Structure analysis and a Summary of each variable indicating the min, max, mean, and median values of the numerical data. I then printed the first 6 lines of the data set using the Head function.

What information is not Self-evident?

From the data set and cleaning the data there is no apparent information regarding relationships between the variables. I will conduct an EDA using various plots to further understand the relationships within the data.

From the Summary Function, you can begin to understand whether outliers exist in the data, but further exploration is necessary to understand outliers completely. Any correlations that may exist among the data are also not apparent, but a statistical analysis can be performed to understand any sort of relationships.

What are different ways of looking at the data?

As stated above, summary statistics help to get a beginning understanding of the data set. I plan to incorporate visualization and statistical analysis to further understand the data and how each variable correlates to one another. I hope to use this information to understand some trends within the hotel booking market for these two hotels featured within the data set.

How do you plan to slice and dice the data?

In this data set. I immediate notice that within the 32 variables, there is a variable for arrival year, arrival, month, and arrival date. I am going to combine these 3 variables into 1 singular arrival date variable. Now there are 30 variables to observe within the data.

```
cleanedData2$arrival_date <- as.Date(paste(cleanedData2$arrival_date_year, cleanedData2$arrival_date_mo
cleanedData3 <- cleanedData2[, !(names(cleanedData2) %in% c("arrival_date_year", "arrival_date_month",


# Move the 'arrival_date' column to the 4th position
arrivalDateIndex <- which(names(cleanedData3) == "arrival_date")
cleanedData3 <- cleanedData3[, c(1:3, arrivalDateIndex, (4:(ncol(cleanedData3) - 1)))]
head(cleanedData3)
```

```
##          hotel arrival_full_date total_stay arrival_date total_guests adr
## 1 Resort Hotel      2015-July-01          0   2015-07-01            2   0
## 2 Resort Hotel      2015-July-01          0   2015-07-01            2   0
## 3 Resort Hotel      2015-July-06          0   2015-07-06            2   0
## 4 Resort Hotel      2015-July-06          0   2015-07-06            1   0
## 5 Resort Hotel      2015-July-07          0   2015-07-07            2   0
## 6 Resort Hotel      2015-July-07          0   2015-07-07            2   0
##   is_canceled lead_time arrival_date_week_number stays_in_weekend_nights
## 1           0       342                      27                       0
## 2           0       737                      27                       0
## 3           0       111                      28                       0
## 4           0         0                      28                       0
## 5           0         8                      28                       0
## 6           0         8                      28                       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
## 3                    0      2        0      0   BB     PRT      Online TA
## 4                    0      1        0      0   BB     PRT         Direct
## 5                    0      2        0      0   BB     PRT         Direct
## 6                    0      2        0      0   BB     PRT      Online TA
##   distribution_channel is_repeated_guest previous_cancellations
## 1               Direct                 0                      0
## 2               Direct                 0                      0
## 3                TA/TO                 0                      0
## 4               Direct                 0                      0
```

```
## 5               Direct                0                    0
## 6                TA/TO                0                    0
##   previous_bookings_not_canceled reserved_room_type assigned_room_type
## 1                              0                  C                  C
## 2                              0                  C                  C
## 3                              0                  A                  H
## 4                              0                  E                  H
## 5                              0                  A                  A
## 6                              0                  A                  E
##   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
## 3               0   No Deposit   240    NULL                    0     Transient
## 4               0   No Deposit   250    NULL                    0     Transient
## 5               0   No Deposit  NULL    NULL                    0     Transient
## 6               0   No Deposit   240    NULL                    0     Transient
##   required_car_parking_spaces total_of_special_requests reservation_status
## 1                           0                         0          Check-Out
## 2                           0                         0          Check-Out
## 3                           0                         2          Check-Out
## 4                           0                         0          Check-Out
## 5                           0                         1          Check-Out
## 6                           0                         1          Check-Out
##   reservation_status_date
## 1                7/1/2015
## 2                7/1/2015
## 3                7/6/2015
## 4                7/6/2015
## 5                7/7/2015
## 6                7/7/2015
```

Summarizing Data to Answer Questions

The main question that I want to answer within this project is to understand historical data to forecast future demand for hotels. I first created a time series plot to see how reservations changed over time. There are clear spikes in the data that show when the most reservations occur.

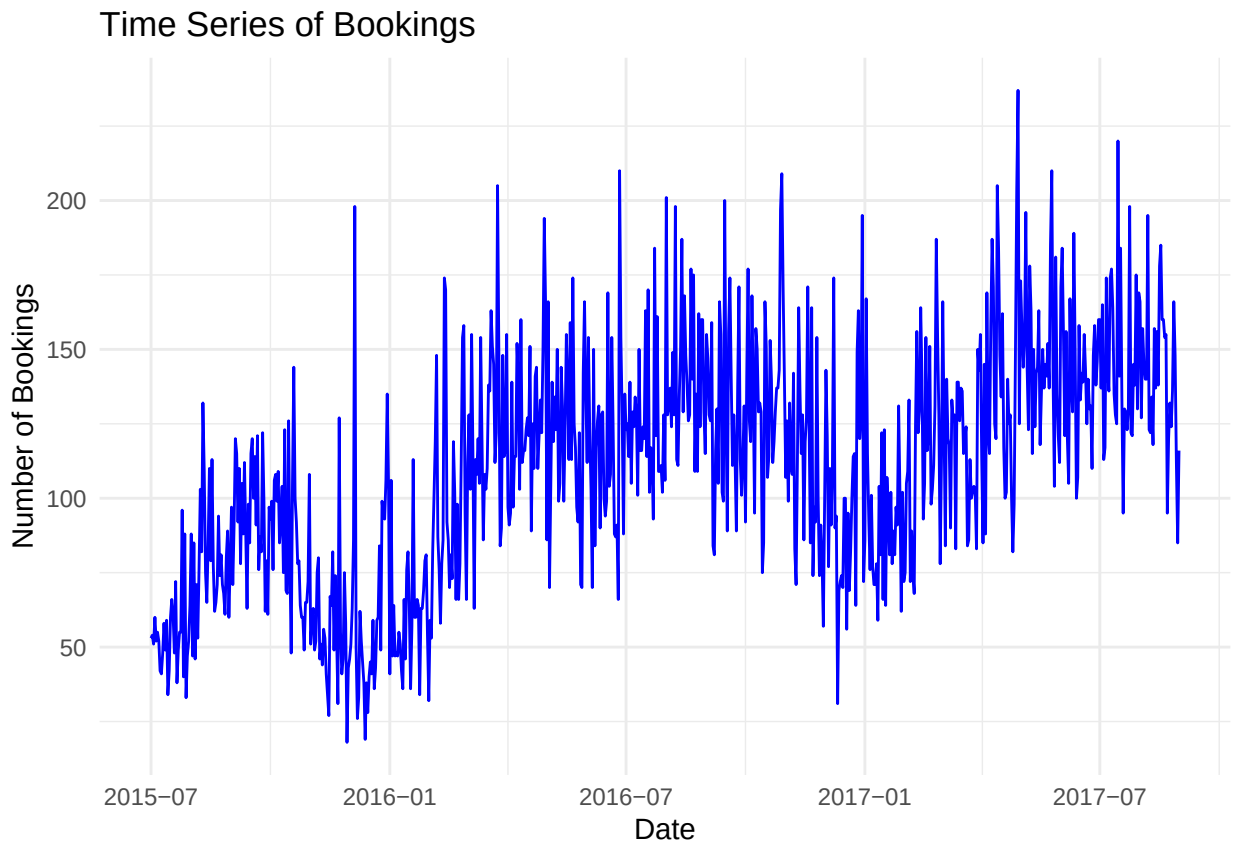
I also created a boxplot so far to analyze the distribution of room types among all the reservations within the data. This will help to understand what types of rooms are most popular at certain times of the year.

Finally, I created a bar chart to show reservations per month. The data shows that this hotel is particularly popular in July and August. This information is very helpful to understand future demand as well.

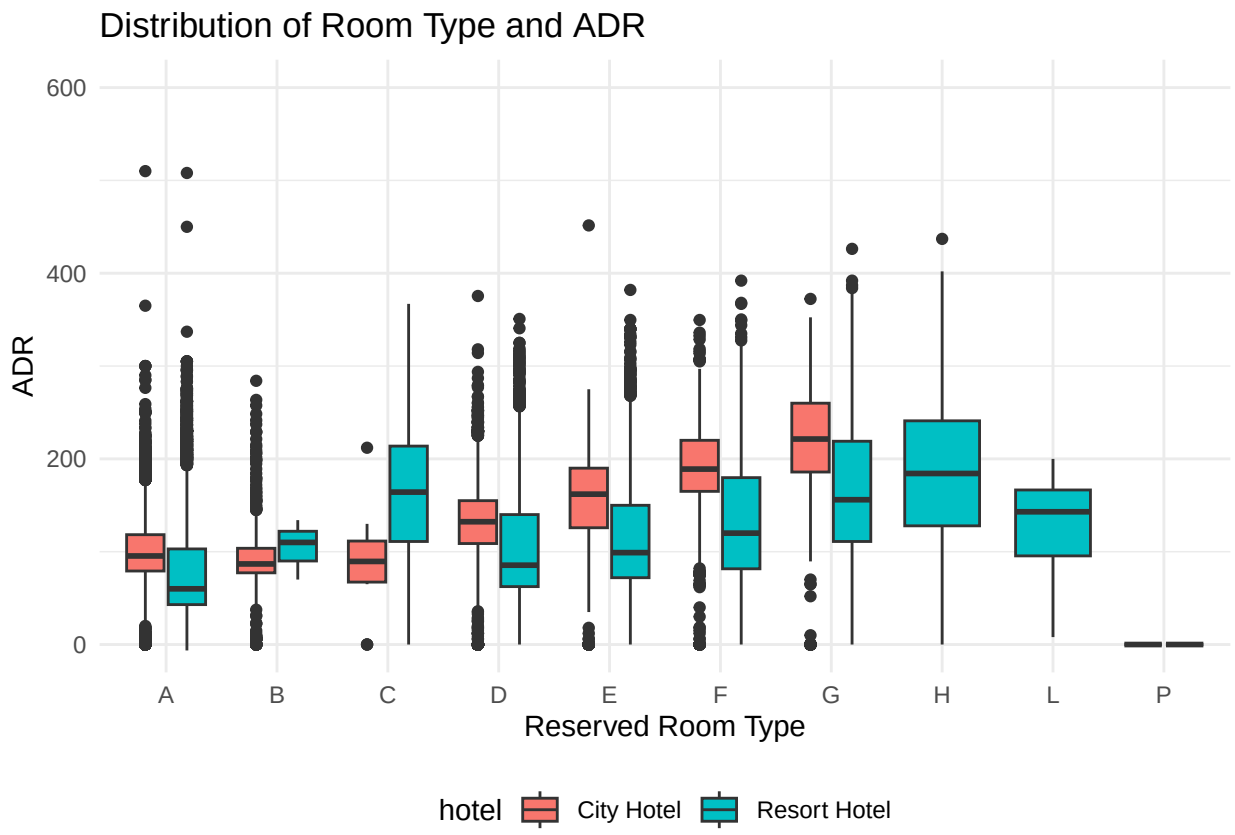
In the final analysis of this project, I plan to create a few more charts to understand the data further. I want to analyze the total amount of guests per month as well. I also want to look at party size to see what types of travelers are utilizing this hotel. Are these families? Larger family reunions? Large travel groups? This information will be useful to understand outliers and who is travelling when and why. ## Plots and Tables

```
# Create a data frame with the date and number of bookings
data <- data.frame(date = cleanedData3$arrival_date,
                   bookings = rep(1, nrow(cleanedData3)))  # Assuming each row represents a booking

# Aggregate the number of bookings by date (in case multiple bookings can occur on the same date)
data <- aggregate(bookings ~ date, data, sum)
```

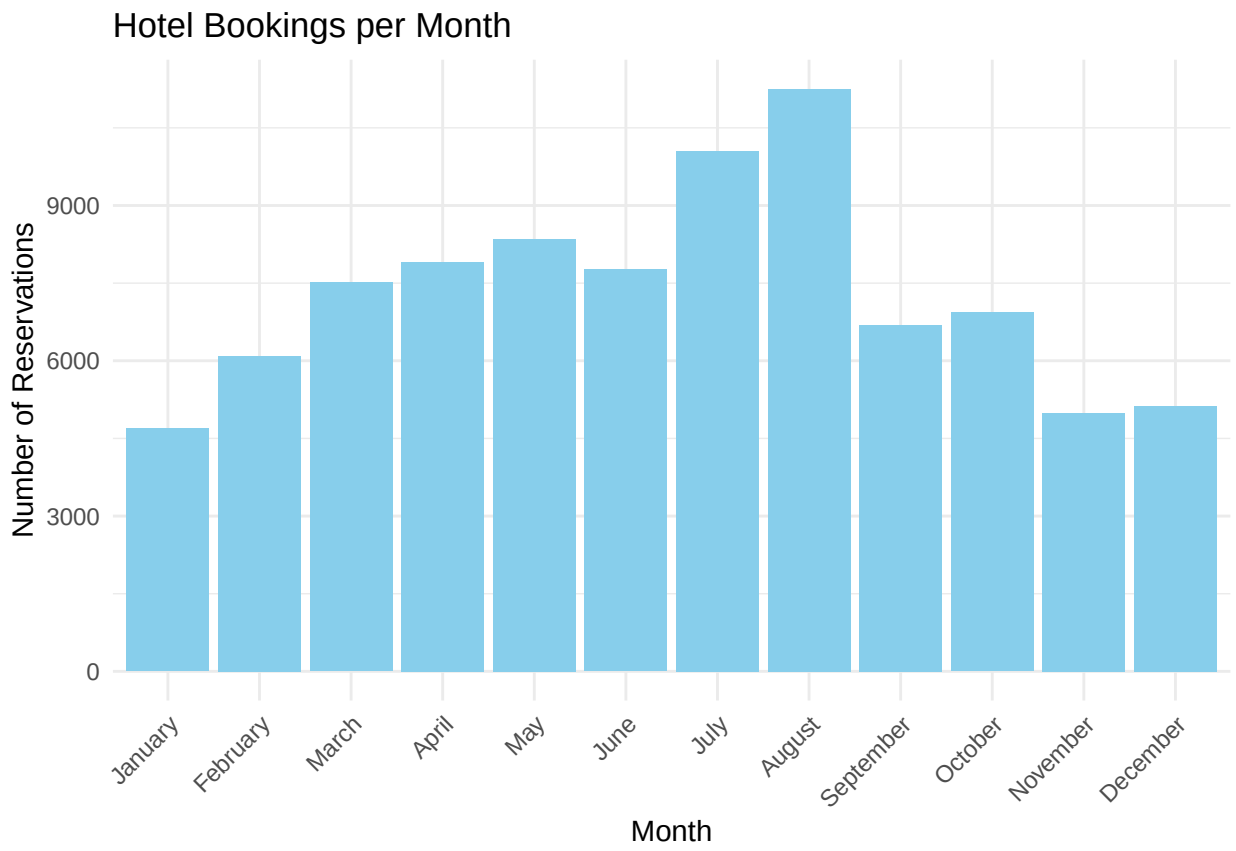
```r
# Plotting the time series
ggplot(data, aes(x = date, y = bookings)) +
  geom_line(color = "blue") +
  labs(title = "Time Series of Bookings", x = "Date", y = "Number of Bookings") +
  theme_minimal()
```



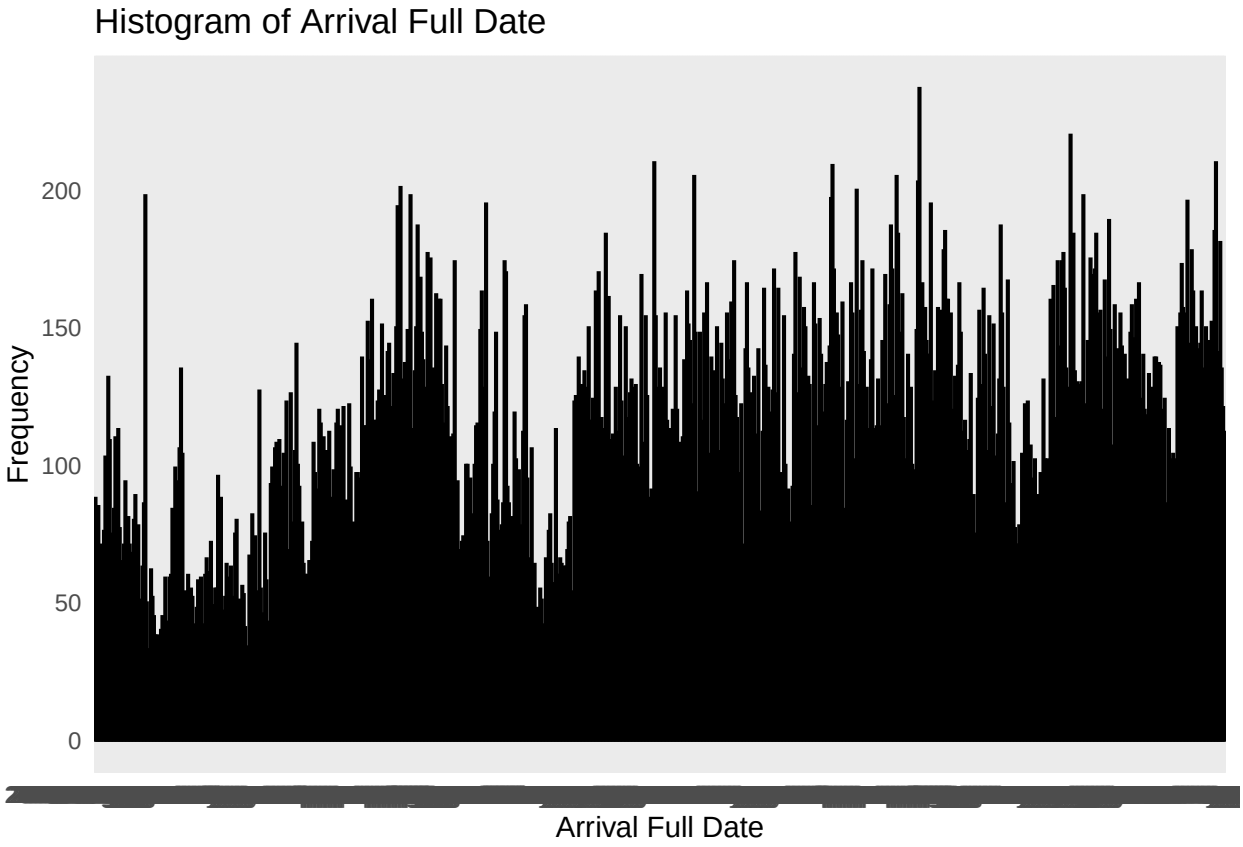

```r
ggplot(cleanedData3, aes(x = reserved_room_type, y = adr, fill = hotel)) +
  geom_boxplot() +
  labs(title = "Distribution of Room Type and ADR",
       x = "Reserved Room Type",
       y = "ADR") +
  theme_minimal() +
  theme(legend.position = "bottom") +  # Move legend to the bottom
  coord_cartesian(ylim = c(0, 600))    # Limit y-axis to 0 to 2000
```




```
cleanedData2$arrival_date_month <- factor(cleanedData2$arrival_date_month,
                                          levels = month.name, ordered = TRUE)

# Count the number of reservations per month
reservations_per_month <- table(cleanedData2$arrival_date_month)

# Convert the result to a data frame
reservations_df <- as.data.frame(reservations_per_month)
names(reservations_df) <- c("Month", "Count")

# Plotting the bar chart
ggplot(reservations_df, aes(x = Month, y = Count)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Hotel Bookings per Month",
       x = "Month",
       y = "Number of Reservations") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate x-axis labels for better readabili
```



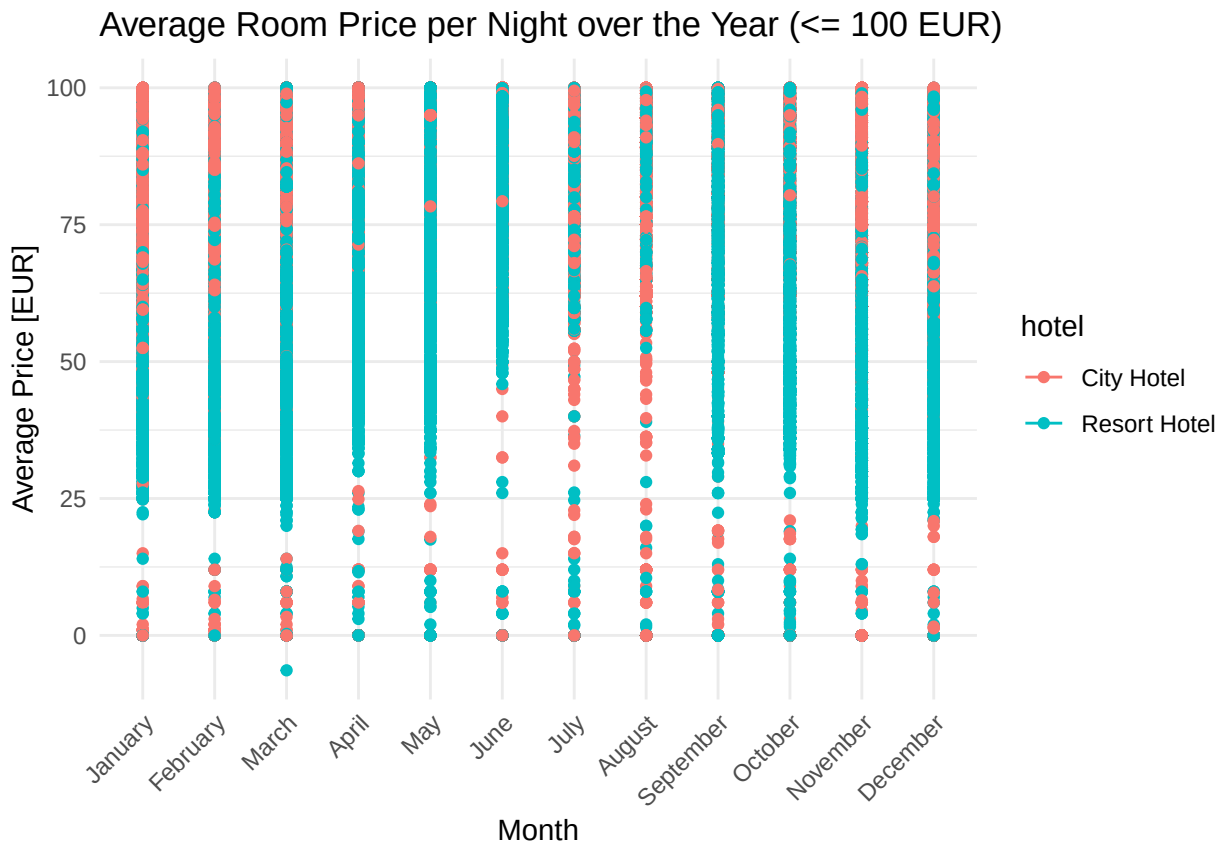

```
# Create a histogram for the arrival_full_date
ggplot(cleanedData3, aes(x = arrival_full_date)) +
  geom_bar(stat = "count", fill = "skyblue", color = "black") +
  labs(title = "Histogram of Arrival Full Date",
       x = "Arrival Full Date",
       y = "Frequency") +
  theme_minimal()
```




```
# Order months
ordered_months <- c("January", "February", "March", "April", "May", "June",
                    "July", "August", "September", "October", "November", "December")

# Convert arrival_date_month to factor with ordered levels
hotelBookings$arrival_date_month <- factor(hotelBookings$arrival_date_month,
                                           levels = ordered_months,
                                           ordered = TRUE)

# Filter the data to include only prices <= 100 EUR
hotelBookings_filtered <- hotelBookings[hotelBookings$adr <= 100, ]

# Line plot with standard deviation
ggplot(hotelBookings_filtered, aes(x = arrival_date_month, y = adr)) +
  geom_line(aes(color = hotel), stat = "summary", fun.y = "mean") +  # Plotting mean prices
  geom_point(aes(color = hotel)) +  # Add points for better visualization
  labs(title = "Average Room Price per Night over the Year (<= 100 EUR)",
       x = "Month",
       y = "Average Price [EUR]") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate x-axis labels for better readabili
```

```
## Warning in geom_line(aes(color = hotel), stat = "summary", fun.y = "mean"):
## Ignoring unknown parameters: `fun.y`

## No summary function supplied, defaulting to `mean_se()`
```

```
## `geom_line()`: Each group consists of only one observation.
## i Do you need to adjust the group aesthetic?
```




```r
# Create a dataframe with the relevant data
resort_guests_monthly <- table(hotelBookings$arrival_date_month)
city_guests_monthly <- table(hotelBookings$arrival_date_month)

resort_guest_data <- data.frame(month = names(resort_guests_monthly),
                                hotel = "Resort hotel",
                                guests = as.numeric(resort_guests_monthly))

city_guest_data <- data.frame(month = names(city_guests_monthly),
                              hotel = "City hotel",
                              guests = as.numeric(city_guests_monthly))

full_guest_data <- rbind(resort_guest_data, city_guest_data)

# Order by month
ordered_months <- c("January", "February", "March", "April", "May", "June",
                    "July", "August", "September", "October", "November", "December")

full_guest_data$month <- factor(full_guest_data$month,
                                levels = ordered_months,
                                ordered = TRUE)

# Normalize data
```

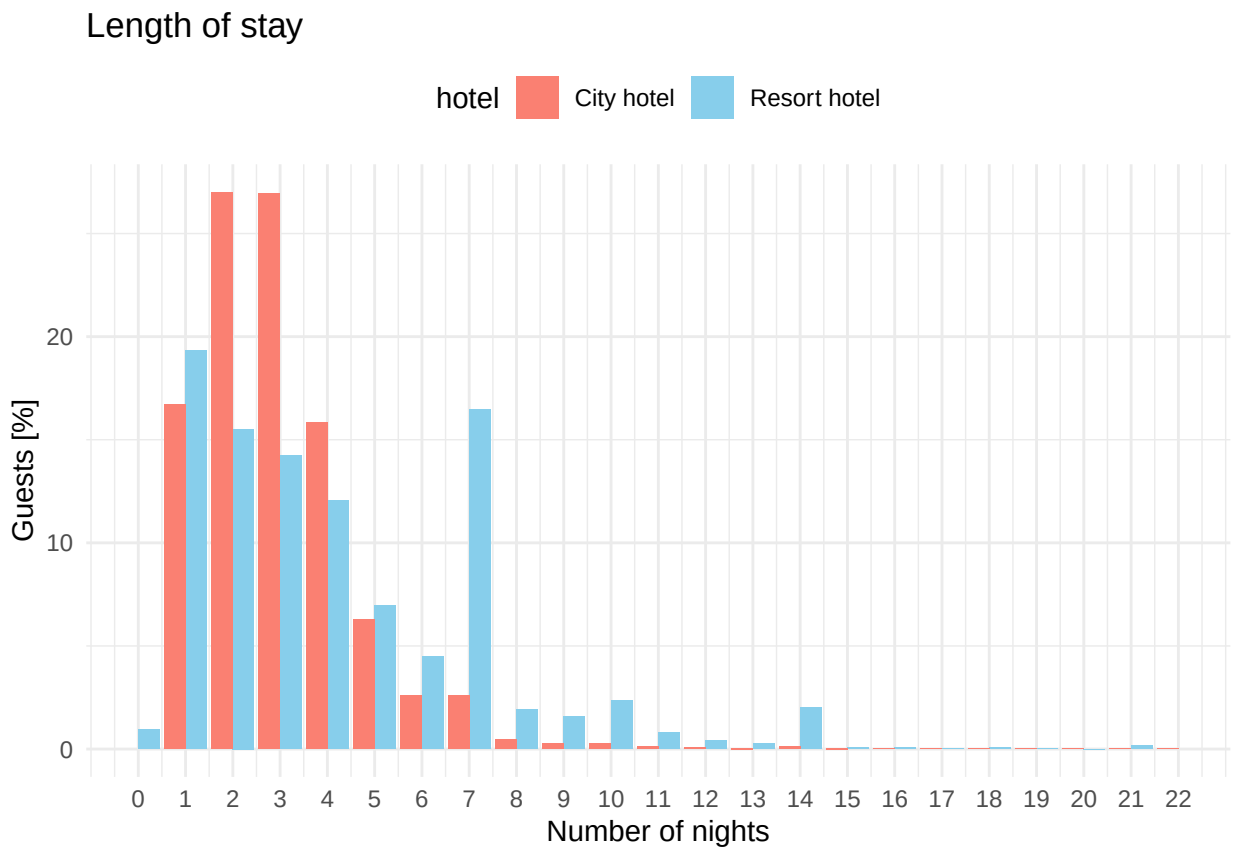
```r
full_guest_data$guests <- ifelse(full_guest_data$month %in% c("July", "August"),
                                 full_guest_data$guests / 3,
                                 full_guest_data$guests / 2)

# Plot line graph
library(ggplot2)

ggplot(full_guest_data, aes(x = month, y = guests, color = hotel, group = hotel)) +
  geom_line(size = 1) +
  geom_point(size = 3) +
  labs(title = "Average number of hotel guests per month",
       x = "Month",
       y = "Number of guests") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate x-axis labels for better readabili
```

```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```


```r
# Calculate total nights for each hotel
hotelBookings$total_nights <- hotelBookings$stays_in_weekend_nights + hotelBookings$stays_in_week_night
```

```r
# Create a dataframe with the relevant data
res_nights <- table(hotelBookings[hotelBookings$hotel == "Resort Hotel", "total_nights"])
cty_nights <- table(hotelBookings[hotelBookings$hotel == "City Hotel", "total_nights"])

res_nights_data <- data.frame(hotel = "Resort hotel",
                              num_nights = as.numeric(names(res_nights)),
                              rel_num_bookings = (as.numeric(res_nights) / sum(res_nights)) * 100)

cty_nights_data <- data.frame(hotel = "City hotel",
                              num_nights = as.numeric(names(cty_nights)),
                              rel_num_bookings = (as.numeric(cty_nights) / sum(cty_nights)) * 100)

nights_data <- rbind(res_nights_data, cty_nights_data)

# Plot bar graph
library(ggplot2)

ggplot(nights_data, aes(x = num_nights, y = rel_num_bookings, fill = hotel)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Length of stay",
       x = "Number of nights",
       y = "Guests [%]") +
  theme_minimal() +
  scale_x_continuous(limits = c(0, 22), breaks = seq(0, 22, 1)) +
  scale_fill_manual(values = c("Resort hotel" = "skyblue", "City hotel" = "salmon")) +
  theme(legend.position = "top")
```

```
## Warning: Removed 32 rows containing missing values or values outside the scale range
## (`geom_bar()`).
```



Machine Learning Techniques

Yes, I will be utlizing a regression model to forecast future demand. In the domain of hotel demand forecasting and revenue management, the integration of regression models provides a powerful means to address critical research inquiries. By harnessing historical booking data and market trends, regression models offer an effective tool to accurately predict hotel room demand. Through sophisticated algorithms such as linear regression, decision trees, and gradient boosting, these models can discern patterns and relationships between various factors influencing demand, enabling precise forecasts. Regression analysis allows for the exploration of seasonal variations and the impact of holidays on hotel occupancy, providing insights into demand fluctuations over time. Additionally, regression models facilitate the analysis of economic indicators' effects on demand dynamics, helping identify key drivers of variations in hotel bookings. By focusing on regression techniques, hotels can leverage advanced analytics to make informed decisions, optimize pricing strategies, and enhance revenue generation capabilities in a competitive market environment.

```
# Assign the dataframe 'df'
df <- hotelBookings

# Descriptive characteristics for 'hotel' (Categorical Variable)
hotel_mode <- names(sort(table(df$hotel), decreasing = TRUE))[1]
hotel_mode_count <- max(table(df$hotel))
hotel_total_count <- length(df$hotel)

# Descriptive characteristics for 'arrival_full_date' (Temporal Variable)
arrival_date_mode <- names(sort(table(df$arrival_full_date), decreasing = TRUE))[1]
arrival_date_mode_count <- max(table(df$arrival_full_date))
```

```r
arrival_date_total_count <- length(df$arrival_full_date)

# Descriptive characteristics for 'total_stay' (Numerical Variable)
total_stay_mean <- mean(df$total_stay)
total_stay_mode <- names(sort(table(df$total_stay), decreasing = TRUE))[1]
total_stay_std <- sd(df$total_stay)

# Descriptive characteristics for 'total_guests' (Numerical Variable)
total_guests_mean <- mean(df$total_guests)
total_guests_mode <- names(sort(table(df$total_guests), decreasing = TRUE))[1]
total_guests_std <- sd(df$total_guests)

# Descriptive characteristics for 'adr' (Numerical Variable)
adr_mean <- mean(df$adr)
adr_mode <- names(sort(table(df$adr), decreasing = TRUE))[1]
adr_std <- sd(df$adr)

# Output results
cat("Descriptive characteristics for 'hotel' (Categorical Variable):\n")
```

```
## Descriptive characteristics for 'hotel' (Categorical Variable):
```

```r
cat(paste("Mode:", hotel_mode, "(Appears", hotel_mode_count, "times)\n"))
```

```
## Mode: City Hotel (Appears 79330 times)
```

```r
cat(paste("Total Count:", hotel_total_count, "\n\n"))
```

```
## Total Count: 119390
```

```r
cat("Descriptive characteristics for 'arrival_full_date' (Temporal Variable):\n")
```

```
## Descriptive characteristics for 'arrival_full_date' (Temporal Variable):
```

```r
cat(paste("Mode:", arrival_date_mode, "(Appears", arrival_date_mode_count, "times)\n"))
```

```
## Mode: 2015-December-05 (Appears 448 times)
```

```r
cat(paste("Total Count:", arrival_date_total_count, "\n\n"))
```

```
## Total Count: 119390
```

```r
cat("Descriptive characteristics for 'total_stay' (Numerical Variable):\n")
```

```
## Descriptive characteristics for 'total_stay' (Numerical Variable):
```

```r
cat(paste("Mean:", total_stay_mean, "\n"))
```

```
## Mean: 3.42790015914231
```

```r
cat(paste("Mode:", total_stay_mode, "\n"))
```

```
## Mode: 2
```

```r
cat(paste("Standard Deviation:", total_stay_std, "\n\n"))
```

```
## Standard Deviation: 2.55743866905197
```

```r
cat("Descriptive characteristics for 'total_guests' (Numerical Variable):\n")
```

```
## Descriptive characteristics for 'total_guests' (Numerical Variable):
```

```r
cat(paste("Mean:", total_guests_mean, "\n"))
```

```
## Mean: 1.96823854594187
```

```r
cat(paste("Mode:", total_guests_mode, "\n"))
```

```
## Mode: 2
```

```r
cat(paste("Standard Deviation:", total_guests_std, "\n\n"))
```

```
## Standard Deviation: 0.722394242658672
```

```r
cat("Descriptive characteristics for 'adr' (Numerical Variable):\n")
```

```
## Descriptive characteristics for 'adr' (Numerical Variable):
```

```r
cat(paste("Mean:", adr_mean, "\n"))
```

```
## Mean: 101.831121534467
```

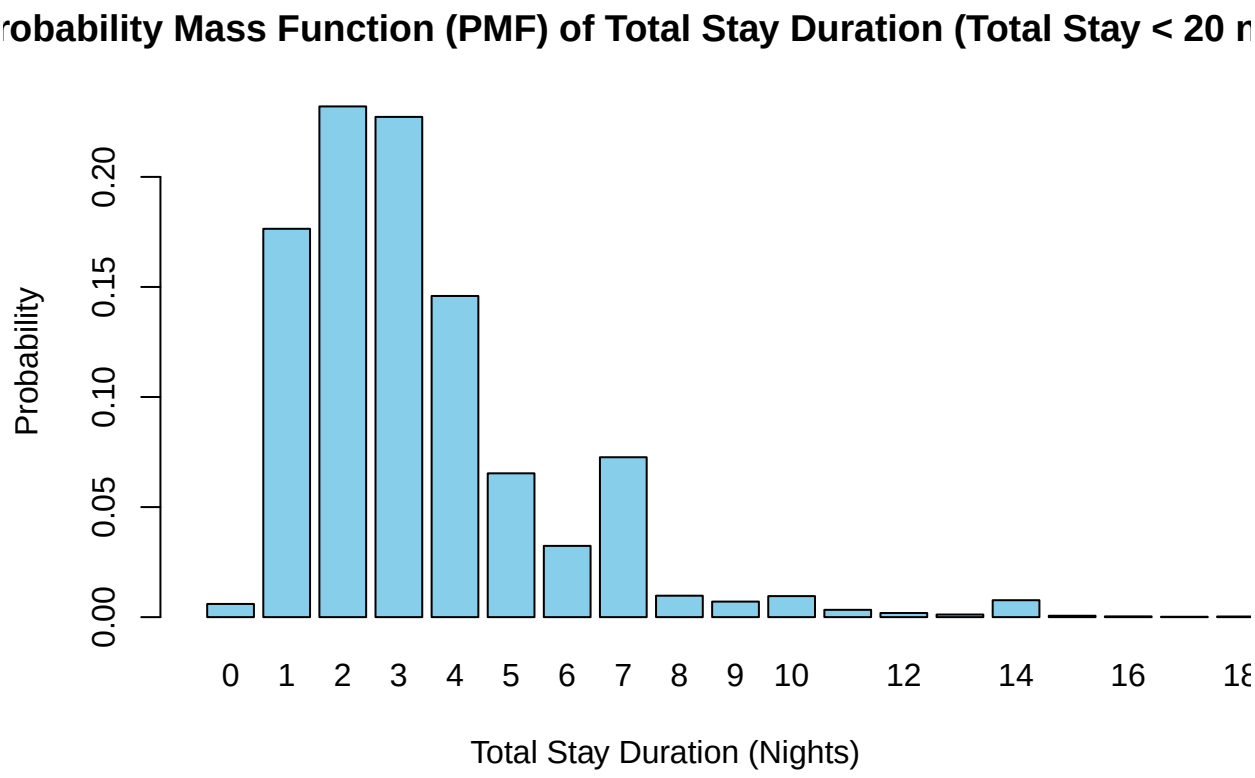
```r
cat(paste("Mode:", adr_mode, "\n"))
```

```
## Mode: 62
```

```r
cat(paste("Standard Deviation:", adr_std, "\n"))
```

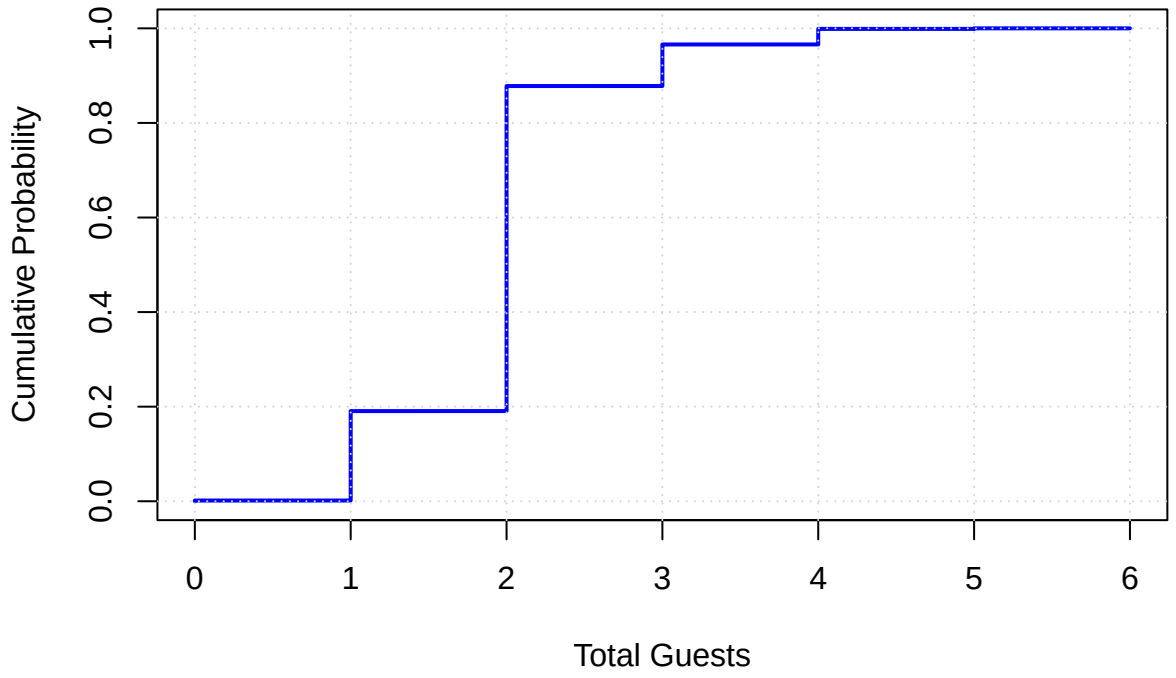
```
## Standard Deviation: 50.5357902855487
```

```r
# Filter the dataset to include only records where total_stay is less than 20 nights
df_filtered <- subset(df, total_stay < 20)

# Compute the PMF for total_stay
pmf <- table(df_filtered$total_stay) / nrow(df_filtered)

# Plot the PMF
barplot(pmf, main = "Probability Mass Function (PMF) of Total Stay Duration (Total Stay < 20 nights)",
        xlab = "Total Stay Duration (Nights)", ylab = "Probability", col = "skyblue",
        ylim = c(0, max(pmf)), xlim = c(0, 20))
```



This PMF function shows that guests are most likely to book a stay between 1-5 nights at either the resort hotel or city hotel. While guests have reservations stays beyond 5 nights, they are significantly more uncommon.

```r
# Filter the dataset to include only records where total_guests is less than 10
df_filtered <- subset(df, total_guests < 10)

# Plot the CDF of total guests with the filtered dataset
sorted_data_filtered <- sort(df_filtered$total_guests)
cumulative_probability_filtered <- (1:length(sorted_data_filtered)) / length(sorted_data_filtered)

plot(sorted_data_filtered, cumulative_probability_filtered, type = "l",
     main = "Cumulative Distribution Function (CDF) of Total Guests (Total Guests < 10)",
     xlab = "Total Guests", ylab = "Cumulative Probability", col = "blue", lwd = 2)
grid()
```



Most guests travel in parties of 2-6. There is a high probability that most reservations will be for couples (parties of 2.)

```r
# Filter the dataset to include only records where total_guests is less than 10 and adr is less than 10
df_filtered <- subset(df, total_guests < 10 & adr < 1000)

# Fit the linear regression model
model <- lm(adr ~ total_stay, data = df_filtered)

# Print the model summary
summary(model)
```

```
##
## Call:
## lm(formula = adr ~ total_stay, data = df_filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -166.46  -33.22   -7.49   23.91  411.51
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 97.12730    0.23242  417.90   <2e-16 ***
## total_stay   1.36260    0.05435   25.07   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 48.02 on 119371 degrees of freedom
## Multiple R-squared:  0.005238,   Adjusted R-squared:  0.00523
## F-statistic: 628.6 on 1 and 119371 DF,  p-value: < 2.2e-16
```

What do you not know how to do right now that you need to learn to answer your questions?

While I do understand major concepts relating to R and R Studio, I do not know how to incorporate all 3 of my data sets to formulate a clear conclusion in the next step. I will need to look into my other 2 data sets to conduct an analysis of the information and integrate it in the analysis I've already started with this comprehensive set of hotel data. I will be spending the next 2 weeks pulling together relevant information to draw conclusions from all the data available for this project.

Term Project Part 3

Introduction

In this analysis, I introduce the topic of predictive analytics for hotel room demand and price forecasting. I highlight the importance of this analysis for maximizing hotel occupancy rates, revenues, and profits, emphasizing its relevance for hospitality leaders in gaining a competitive advantage in the market. Additionally, I express my personal interest in this topic due to its potential to optimize resource allocation and operational revenue, ultimately leading to cost minimization and profit maximization for hotels.

Problem Statement

Ultimately, hotels want to maximize revenue. To do so, they need to understand historical data to predict future demand. How can historical data be leveraged to develop predictive models regarding hotel room demand forecasting?

Analysis

I used a variety of plots to conduct my exploratory data analysis of the information in this dataset between two hotels, a resort hotel and city hotel. For this analysis, I focused on 5 key factors that influence hotel reservations. First, I looked at reservation dates to understand the most popular times of the year to travel. One plot shows that for the city hotel, peak periods with the most number of guests include May and September/October while the resort hotel sees a consistent flow of guests throughout the year, mainly spiking in late winter through early spring (February – May) as well as a peak in October. I also looked at average room prices per night throughout the year, finding that the city hotel is most expensive in May and September while the resort hotel is most expensive in August. Throughout this analysis it is important to understand city hotel is significantly more popular than the resort hotel and that information may skew data conclusions that are found throughout the project. Finally, I analyzed the average length of stay in relationship to the amount of guests for each hotel finding that most reservations for the city hotel range between 1-4 nights. For the resort hotel, I the chart shows that smaller parties tend to stay at the hotel longer, while larger parties tend to book reservations for shorter amount of time.

The regression results indicate that the model explains only a small portion of the variance in the dependent variable (ADR), with an R-squared value of 0.005. The coefficient for the total_stay variable is statistically significant ($p < 0.05$), suggesting that there is a significant relationship between total stay duration and average daily rate. Specifically, for each additional night of stay, the average daily rate increases by approximately 1.36 units. The F-statistic is highly significant ($p < 0.05$), indicating that the overall regression model is statistically significant in predicting the dependent variable. Additionally, the coefficients for both the intercept (const) and the explanatory variable (total_stay) are statistically significant, as their p-values are less than 0.05. The Durbin-Watson statistic of 0.827 suggests the presence of some positive autocorrelation in the residuals, and the Jarque-Bera test indicates that the residuals are not normally distributed, suggesting potential violations of the assumption of normality. Overall, the model provides some predictive power, but additional factors beyond total stay duration may be needed to better explain variations in the average daily rate.

To conclude, based on these two hotels in this particular area, between the resort hotel and city hotel combined, reservations can be expected to be made at peak travel times in the year for parties of 2 with a length of stay of 2-3 nights.

Implications

The implications of this project are manifold, offering significant value to the hospitality industry and beyond. By harnessing predictive analytics and data-driven insights, hotels can enhance their operational efficiency,

optimize resource allocation, and maximize revenue generation. Accurate demand forecasting enables hotels to strategically plan staffing, inventory management, and marketing efforts, ensuring optimal utilization of resources and enhancing overall guest experience. Moreover, the ability to dynamically adjust pricing and promotional strategies based on demand forecasts and market trends empowers hotels to remain competitive in a rapidly evolving market landscape. Beyond the hospitality sector, the methodologies and insights derived from this project can be applied to various other industries, highlighting the broader impact of data science and predictive analytics in driving informed decision-making and sustainable business growth.

Limitations

Analyzing hotel data presents several limitations that need to be considered when interpreting the findings. Firstly, the dataset may not capture all relevant factors influencing hotel bookings and guest behavior. For instance, it might lack detailed information on customer preferences, such as specific amenities desired or preferred room types. Additionally, the dataset may not fully represent the diversity of hotel guests, potentially overlooking important demographic or cultural factors that could affect booking patterns. Moreover, the data's temporal scope may be limited, covering only a specific time period or excluding certain seasons or events that could impact hotel occupancy and pricing dynamics. Furthermore, data quality issues such as missing values, inaccuracies, or inconsistencies could affect the reliability of analyses and conclusions drawn from the dataset. Finally, the dataset's representativeness may vary across different types of hotels (e.g., luxury resorts vs. budget accommodations), potentially limiting the generalizability of findings to the broader hospitality industry. Overall, while hotel data offers valuable insights, researchers and analysts should be mindful of these limitations and interpret results with caution.

Concluding Remarks

This data is extremely useful in understanding trends within the hotel industry. However, the information contained is only surface level. While this analysis is helpful in predicting reservations for the future within these two hotels, the results may not translate to other companies around the world. I suggest a further analysis with a larger sample of hotel data or a specific sample in the field future researchers want to analyze. This project helps to explore concepts learned in this course and apply them to real world information.

In conclusion, this project provides valuable insights into hotel bookings, guest behavior, and pricing dynamics based on the analysis of a comprehensive hotel dataset. Through exploratory data analysis and statistical modeling, we gained a deeper understanding of factors influencing hotel occupancy rates, average daily rates, and length of stay. The findings highlight the importance of considering various factors, including seasonality, booking lead time, and guest demographics, in predicting hotel demand and pricing. Moreover, the project identified key trends and patterns in hotel booking behavior, such as the popularity of certain room types, preferred lengths of stay, and common booking channels. However, it's essential to recognize the limitations of the dataset and analyses conducted, including data quality issues and potential biases. Moving forward, further research could explore additional factors impacting hotel performance, such as guest reviews, marketing strategies, and economic indicators. Overall, this project serves as a foundation for future research and decision-making in the hospitality industry, providing valuable insights for hotel management, marketing professionals, and policymakers alike.