Julie Son-Garcia

Project 3 Stat 611 Hotel Demands Dataset

Introduction

The objective of this project is to explore Hotels dataset and check out interesting data on which hotel has the most reservations and which hotel is favorable for adults stay. The dataset looks at reservations from July 2015 to August 2017 of 2 different hotels in Portugal. A lot of information can be gathered based on hotel guests' arrivals, cancellations, number of guests. This analysis can help give an idea for vacation planning. I will use some data cleaning to help sort relevant information. Several codes and R packages will be used to gather information from tidyverse.

Packages Required

TidytuesdayR : Tidytuesday is a weekly podcast and community activity to help R learners learn in real-world contexts. This is where we will access our hotel demands dataset.

Github: it is a repository hosting service site. Git is a command line tool but Github also provides web-based graphical interface to control and allow for collaboration features for projects.

Tidyverse: this will load core packages such as ggplot2 for data visualizations, dplyr for data manipulation, tidyr for data tidying, readr for data import, purr for functional programming, tibble for tibbles (modern re-imaging of data frames), stringr for strings, and forcats for factors.

Ggplot2: easy data visualization and plotting

Data Preparation

The open hotel booking demand dataset came from Antonio, Almeida and Nunes, 2019.This dataset provides bookings of two hotels in Portugal with arrival dates from July 1, 2015 to Aug. 31, 2017. These bookings compare H1, a resort hotel and H2, a city hotel with confirmed arrivals and confirmed cancellations. The original purpose of the data collection was for research and education in revenue management, machine learning, or data mining. There are 32 variables.
I used modified dataset that created a new column that combined H1 and H2 and renamed as Resort Hotel and City hotel. The URL for this dataset is https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2020/2020-02-11/hotels.csv

I selected only the variables I was interested in. So I used the filter and select function. I only wanted to see arrival year and month and I wanted to only see adults staying at either resort hotel or city hotel. I filtered the months to the summer months of June through August. I created a new dataframe with only observations from these three months. After that I created a CSV file with the new data I created. Then I wanted to plot the total bookings from 2015 to 2017 from the two hotels. Then I plotted the yearly bookings in each hotel made by adults only.

```
head(yearly_counts,10)
```

A tibble: 6 x 3

arrival_date_year	hotel	n
<dbl>	<chr>	<int>
2015	City Hotel	3874
2015	Resort Hotel	2787
2016	City Hotel	10432
2016	Resort Hotel	4495
2017	City Hotel	10655
2017	Resort Hotel	5230

A tibble: 10 x 4

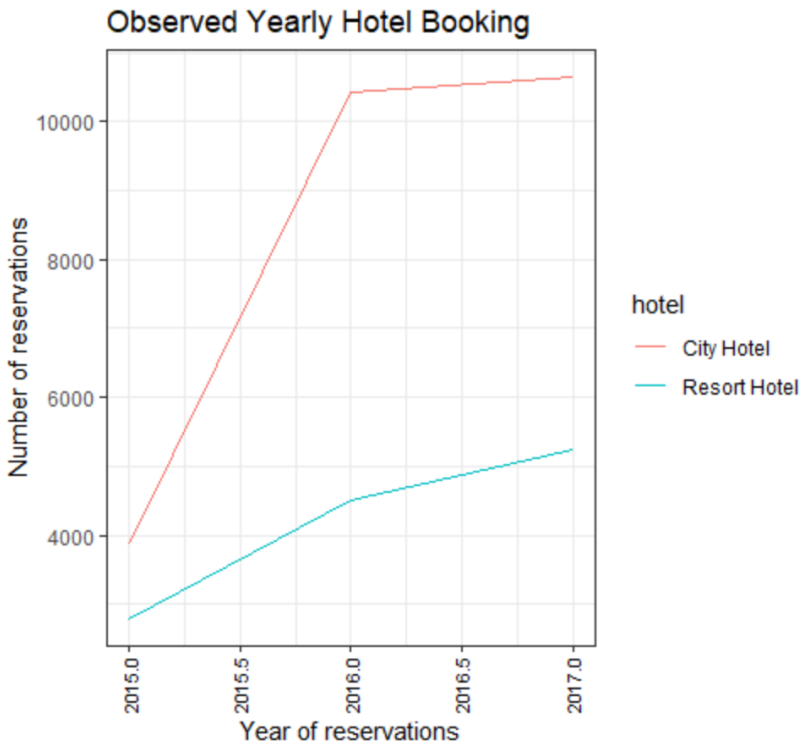
arrival_date_year	hotel	adults	n
<dbl>	<chr>	<dbl>	<int>
2015	City Hotel	0	12
2015	City Hotel	1	436
2015	City Hotel	2	3355
2015	City Hotel	3	71
2015	Resort Hotel	1	153
2015	Resort Hotel	2	2462
2015	Resort Hotel	3	165
2015	Resort Hotel	4	7
2016	City Hotel	0	73
2016	City Hotel	1	1652

Exploratory Data Analysis

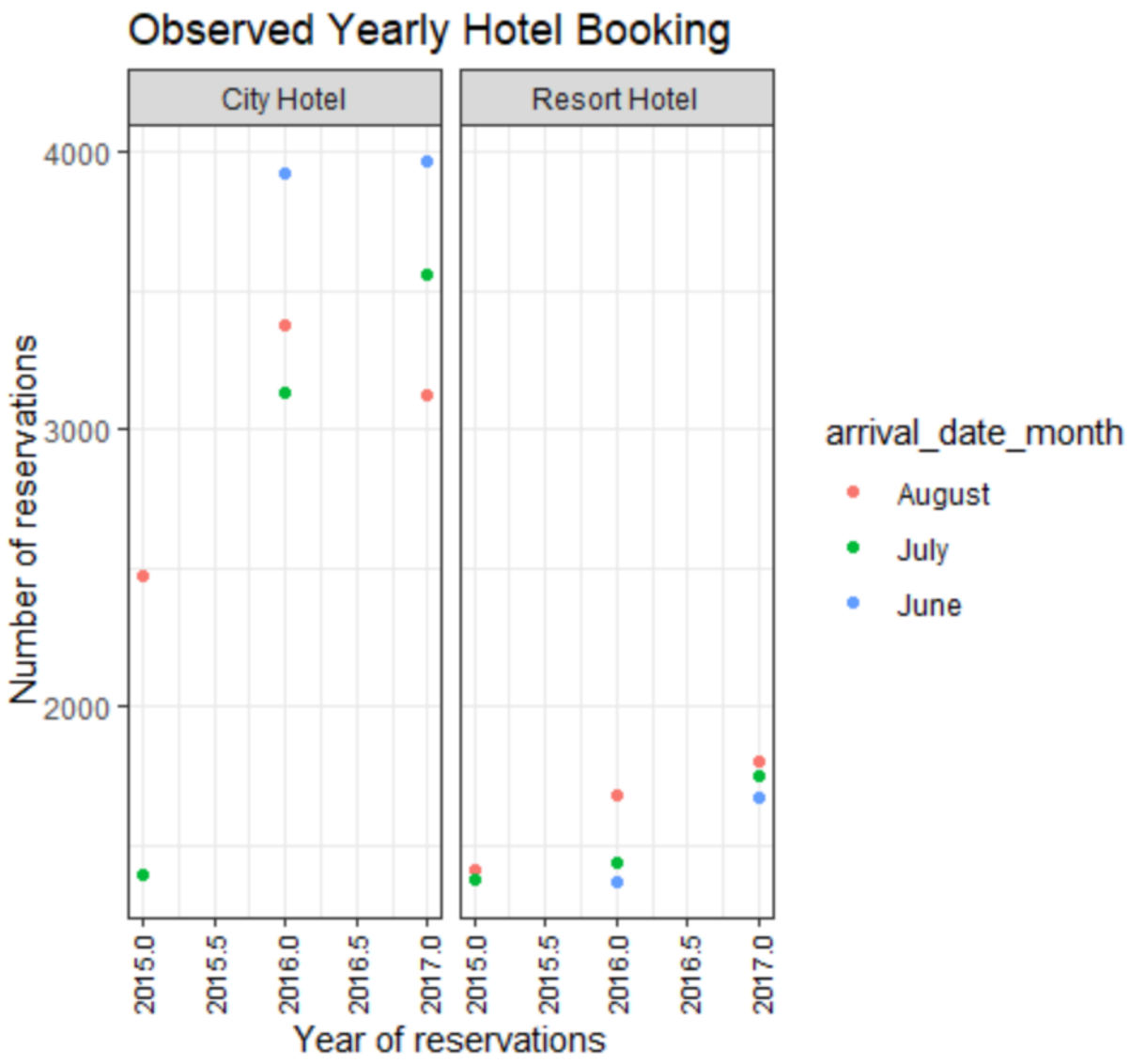
The dataset that was created showed observations that only included summer months of June to August. The first graph shows that between 2015 and 2017 the City hotel had the most bookings and continued to increase throughout the year. The Resort hotel also shows a steady increase in bookings throughout the year, but not nearly as much as City hotel. Since data was collected starting July 1 2015 to August 31,2017, the monthly observations show no June bookings for either hotel until 2016 and 2017.

Under the graph of observed yearly adult bookings, the City hotel still had the most adult bookings compared to the Resort hotel.

Julie Son-Garcia



Julie Son-Garcia



Julie Son-Garcia

```r
# Clean workspace
rm(list=ls())
## Load datasets

hotels <-
readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2020/2020-02-11/hotels.csv')
View(hotels) #this author changed the original file by combining H1 and H2 data together into one file.

library(tidyverse)
library(dplyr)
library(tidyr)
library(tibble)
glimpse(hotels)

How many adult bookings in the summer from 2015 to 2017?
  #first remove data

## Extract the most common hotel
hotel_counts <- hotels %>%
  count(arrival_date_month) %>%
  filter(arrival_date_month %in% c("June","July","August"))
view(hotel_counts)

##Extract the most common bookings from the months of June, July, August


        hotels_complete <- hotels_complete%>%
          filter(arrival_date_month %in% hotel_counts$arrival_date_month)
        dim(hotels_complete)
        write_csv(hotels_complete, file = "hotels_complete.csv")

    #plotting
        library(ggplot2)
        dev.off()
    #not relevant data
    hotels_complete<-read_csv("hotels_complete.csv")
    new.data<-ggplot(data=hotels_complete, aes(x=arrival_date_month,y=arrival_date_year)) +
      geom_point(aes(color=adults))
    plot(new.data)

 #good plot   the yearly counts of books from 2015 to 2017 for each hotel
    yearly_counts<-hotels_complete%>%
      count(arrival_date_year,hotel)
    view(yearly_counts)
    yearly_counts%>%
      ggplot(aes(x=arrival_date_year,y=n, color=hotel)) +
      geom_line()+
      labs(title = "Observed Yearly Hotel Booking",
           x = "Year of reservations",
           y = "Number of reservations") +
      theme_bw() +
      theme(axis.text.x = element_text(colour = "black", size = 8, angle = 90, hjust = 0.5, vjust = 0.5))
    head(yearly_counts,10)

    #plot of counts in the data frame grouped by year, adults, counts
    yearly_adult_counts<-hotels_complete%>%
      count(arrival_date_year, hotel, adults)
    yearly_adult_counts%>%
      ggplot(aes(x=arrival_date_year,y=n,color=adults)) +
      geom_point() +
      facet_wrap(facets=vars(hotel))+
      labs(title = "Observed adult hotel bookings",
           x = "Year of reservations",
           y = "Number of reservations") +
      theme_bw() +
      theme(axis.text.x = element_text(colour = "black", size = 8, angle = 90, hjust = 0.5, vjust = 0.5))

    head(yearly_adult_counts,10)

    month_counts<-hotels_complete%>%
      count(arrival_date_year,hotel, arrival_date_month)
    view(month_counts)
    month_counts%>%
      ggplot(aes(x=arrival_date_year,y=n, color=arrival_date_month)) +
      geom_point()+
      facet_wrap(facets=vars(hotel))+
      labs(title = "Observed Yearly Hotel Booking",
           x = "Year of reservations",
           y = "Number of reservations") +
      theme_bw() +
      theme(axis.text.x = element_text(colour = "black", size = 8, angle = 90, hjust = 0.5, vjust = 0.5))
    head(yearly_counts,10)
```

Julie Son-Garcia

Summary

In summary, if a couple of adults wanted to plan a vacation in Portugal and wanted to know which hotel to book then this data will tell you that over the 3 year span, the City Hotel seemed to be the most popular one. More single adults and up booked at the City hotel compared to the Resort Hotel. The consumer can also see that in 2016 and 2017 the month of June was had the most bookings in City Hotel and in the Resort Hotel the most bookings occurred in August. Some limitations to the dataset is that sometimes, you cannot tell based on graphs whether or not there were cancellations. It just shows observed bookings. Also, since the data started in July, we couldn't tell if there was a trend of which summer months is the peak time of reservations.