

Hotels

Christina Liang

```
library(tidyverse)
library(infer)
```

```
hotel_bookings <- read.csv("~/R/DIIG/hotel_bookings.csv")
```

First, I made some new variables and did some data cleaning:

New variable for total amount of nights stayed:

```
hotel_bookings <- hotel_bookings %>%
  mutate(total_nights = stays_in_week_nights + stays_in_weekend_nights)
```

Changing the month of arrival into chronologically-ordered levels:

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

I also changed the is_canceled variable from numeric to categorical, as 0 and 1 represent a booking being cancelled or not.

```
hotel_bookings$is_canceled <- as.factor(hotel_bookings$is_canceled)
```

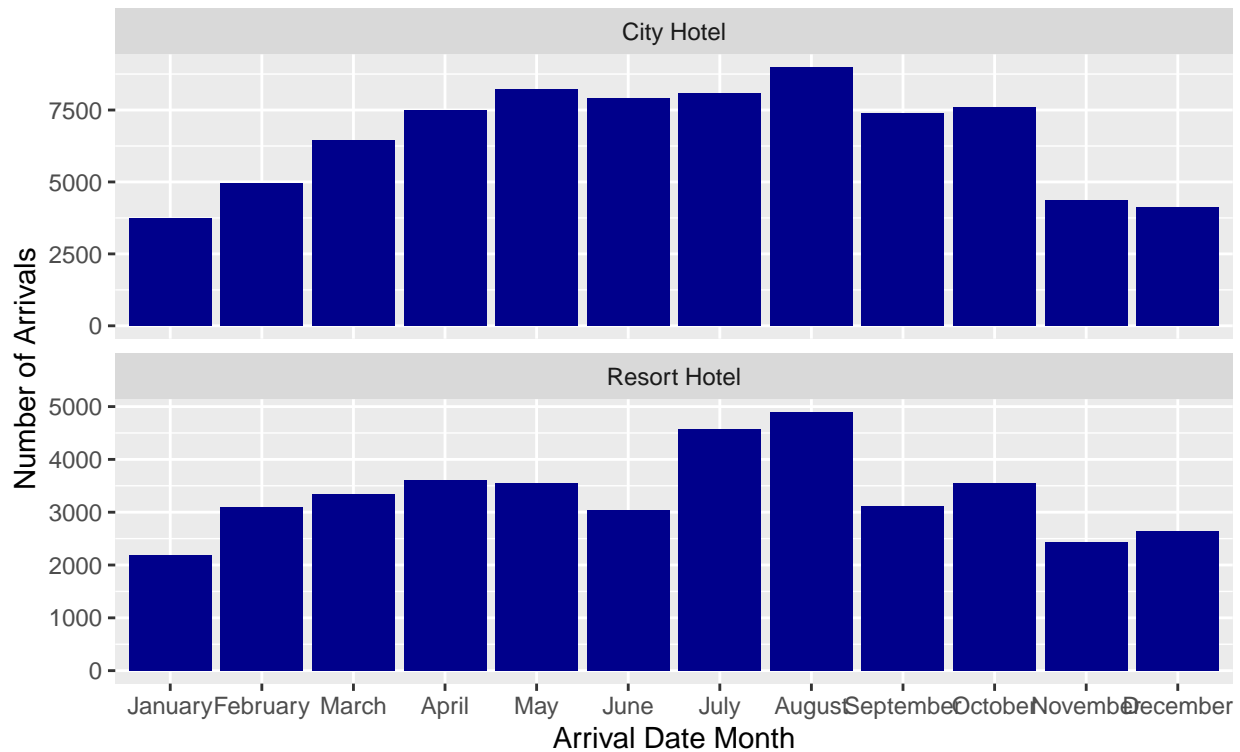
I created a variable for the total number of guests during the duration of the stay:

```
hotel_bookings <- hotel_bookings %>%
  mutate(total_guests = adults + children + babies)
```

First, I visualized the distribution of visits to the hotels based on month of the year, to find that there was an increase in volume of arrivals in the warmer months.

```
hotel_bookings %>%
  group_by(hotel, arrival_date_month) %>%
  ggplot(aes(x = arrival_date_month)) +
  geom_bar(fill = "darkblue") +
  facet_wrap(~ hotel,
             nrow = 2,
             scales = "free_y") +
  labs(title = "Distribution of Arrivals at Hotel by Month of the Year",
       subtitle = "Faceted by City vs. Resort Hotel",
       x = "Arrival Date Month",
       y = "Number of Arrivals")
```

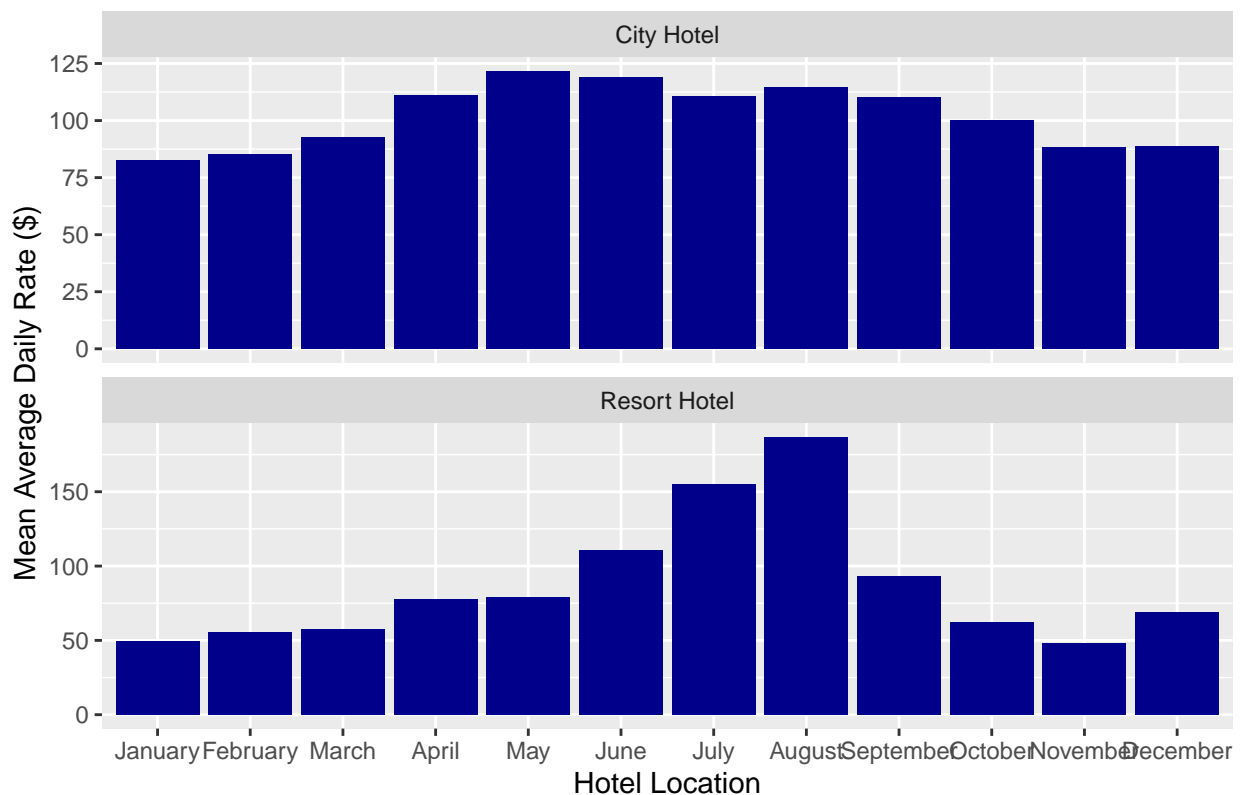
Distribution of Arrivals at Hotel by Month of the Year
Faceted by City vs. Resort Hotel



```
hotel_bookings %>%
  group_by(hotel, arrival_date_month) %>%
  summarise(meanadr = mean(adr)) %>%
  ggplot(aes(x = arrival_date_month, y = meanadr)) +
  geom_col(fill = "darkblue") +
  facet_wrap(~ hotel, nrow = 2, scales = "free_y") +
  labs(title = "Mean Average Daily Rate by Hotel",
       x = "Hotel Location",
       y = "Mean Average Daily Rate ($)")

## `summarise()` regrouping output by 'hotel' (override with `.groups` argument)
```

Mean Average Daily Rate by Hotel



```
hotel_bookings %>%
  group_by(hotel, arrival_date_month) %>%
  summarise(meanadr = mean(adr))

## `summarise()` regrouping output by 'hotel' (override with `.groups` argument)

## # A tibble: 24 x 3
## # Groups:   hotel [2]
##   hotel      arrival_date_month meanadr
##   <chr>      <fct>                <dbl>
## 1 City Hotel January              82.6
## 2 City Hotel February             85.1
## 3 City Hotel March                92.6
## 4 City Hotel April                111.
## 5 City Hotel May                  122.
## 6 City Hotel June                 119.
## 7 City Hotel July                 111.
## 8 City Hotel August               115.
## 9 City Hotel September            110.
## 10 City Hotel October             100.
## # ... with 14 more rows
```

For this data challenge, I'll mainly be focusing on Resort Hotels, so I filtered the "City Hotels" out of my dataset.

```
resort_bookings <- hotel_bookings %>%
  filter(hotel == "Resort Hotel")
```

Question: what predicts cancellations?

How does having kids influence rates?

```
logit_mod1 <- glm(is_canceled ~ arrival_date_month + total_nights,  
                 data = resort_bookings, family = "binomial", maxit = 100)
```

```
logit_mod2 <- glm(is_canceled ~ adults + children + babies + meal,  
                 data = resort_bookings, family = "binomial", maxit = 100)
```

```
logit_mod2
```

```
##  
## Call: glm(formula = is_canceled ~ adults + children + babies + meal,  
##         family = "binomial", data = resort_bookings, maxit = 100)  
##  
## Coefficients:  
##      (Intercept)      adults      children      babies      mealFB  
##      -1.85054      0.41463      0.32996     -0.64143      1.36302  
##      mealHB      mealSC mealUndefined  
##      0.22745     -2.19245     -0.05317  
##  
## Degrees of Freedom: 40059 Total (i.e. Null); 40052 Residual  
## Null Deviance: 47330  
## Residual Deviance: 46310 AIC: 46330
```

```
logit_mod3 <- glm(is_canceled ~ adr + required_car_parking_spaces + total_of_special_requests,  
                 data = resort_bookings, family = "binomial", maxit = 100)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```