Hotel Booking Demand Dataset Case Study - EDA

Priyadarshini Subramani

05/11/2020

Read the data

```
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------------------------------

## v ggplot2 3.3.0     v purrr   0.3.3
## v tibble  3.0.0     v dplyr   0.8.5
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts ---------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
theme_set(theme_light())

hotels <- readr::read_csv('hotel_bookings.csv')
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   hotel = col_character(),
##   ‘No Response‘ = col_logical(),
##   arrival_date_month = col_character(),
##   meal = col_character(),
##   country = col_character(),
##   market_segment = col_character(),
##   distribution_channel = col_character(),
##   reserved_room_type = col_character(),
##   assigned_room_type = col_character(),
##   deposit_type = col_character(),
##   agent = col_character(),
##   company = col_character(),
##   customer_type = col_character(),
##   reservation_status = col_character(),
##   reservation_status_date = col_character()
## )
## See spec(...) for full column specifications.
```

```
glimpse(hotels)
```

```
## Rows: 119,390
## Columns: 33
## $ hotel                          <chr> "Resort Hotel", "Resort Hotel", "Res...
## $ is_canceled                    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, ...
## $ `No Response`                  <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ lead_time                      <dbl> 342, 737, 7, 13, 14, 14, 0, 9, 85, 7...
## $ arrival_date_year              <dbl> 2015, 2015, 2015, 2015, 2015, 2015, ...
## $ arrival_date_month             <chr> "July", "July", "July", "July", "Jul...
## $ arrival_date_week_number       <dbl> 27, 27, 27, 27, 27, 27, 27, 27, 27, ...
## $ arrival_date_day_of_month      <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ stays_in_weekend_nights        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ stays_in_week_nights           <dbl> 0, 0, 1, 1, 2, 2, 2, 2, 3, 3, 4, 4, ...
## $ adults                         <dbl> 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, ...
## $ children                       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ babies                         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ meal                           <chr> "BB", "BB", "BB", "BB", "BB", "BB", ...
## $ country                        <chr> "PRT", "PRT", "GBR", "GBR", "GBR", "...
## $ market_segment                 <chr> "Direct", "Direct", "Direct", "Corpo...
## $ distribution_channel           <chr> "Direct", "Direct", "Direct", "Corpo...
## $ is_repeated_guest              <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ previous_cancellations         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ previous_bookings_not_canceled <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ reserved_room_type             <chr> "C", "C", "A", "A", "A", "A", "C", "...
## $ assigned_room_type             <chr> "C", "C", "C", "A", "A", "A", "C", "...
## $ booking_changes                <dbl> 3, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ deposit_type                   <chr> "No Deposit", "No Deposit", "No Depo...
## $ agent                          <chr> "NULL", "NULL", "NULL", "304", "240"...
## $ company                        <chr> "NULL", "NULL", "NULL", "NULL", "NUL...
## $ days_in_waiting_list           <dbl> 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....
## $ required_car_parking_spaces    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ total_of_special_requests      <dbl> 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, ...
## $ reservation_status             <chr> "Check-Out", "Check-Out", "Check-Out...
## $ reservation_status_date        <chr> "01-07-2015", "01-07-2015", "02-07-2...
```

```
hotels %>% count(is_canceled)
```

```
## # A tibble: 3 x 2
##   is_canceled     n
##         <dbl> <int>
## 1           0 75166
## 2           1 44046
## 3          NA   178
```

Explore the data

The hotel stay had more reservation without children than with children

```
hotels %>%
  filter(is_canceled == 0) %>%
  mutate(children = case_when(children+babies > 0 ~ "children", TRUE ~ "none")) %>%  count(children)
```

```
## # A tibble: 2 x 2
##   children     n
##   <chr>    <int>
## 1 children  6073
## 2 none     69093
```

Explore the hotel bookings

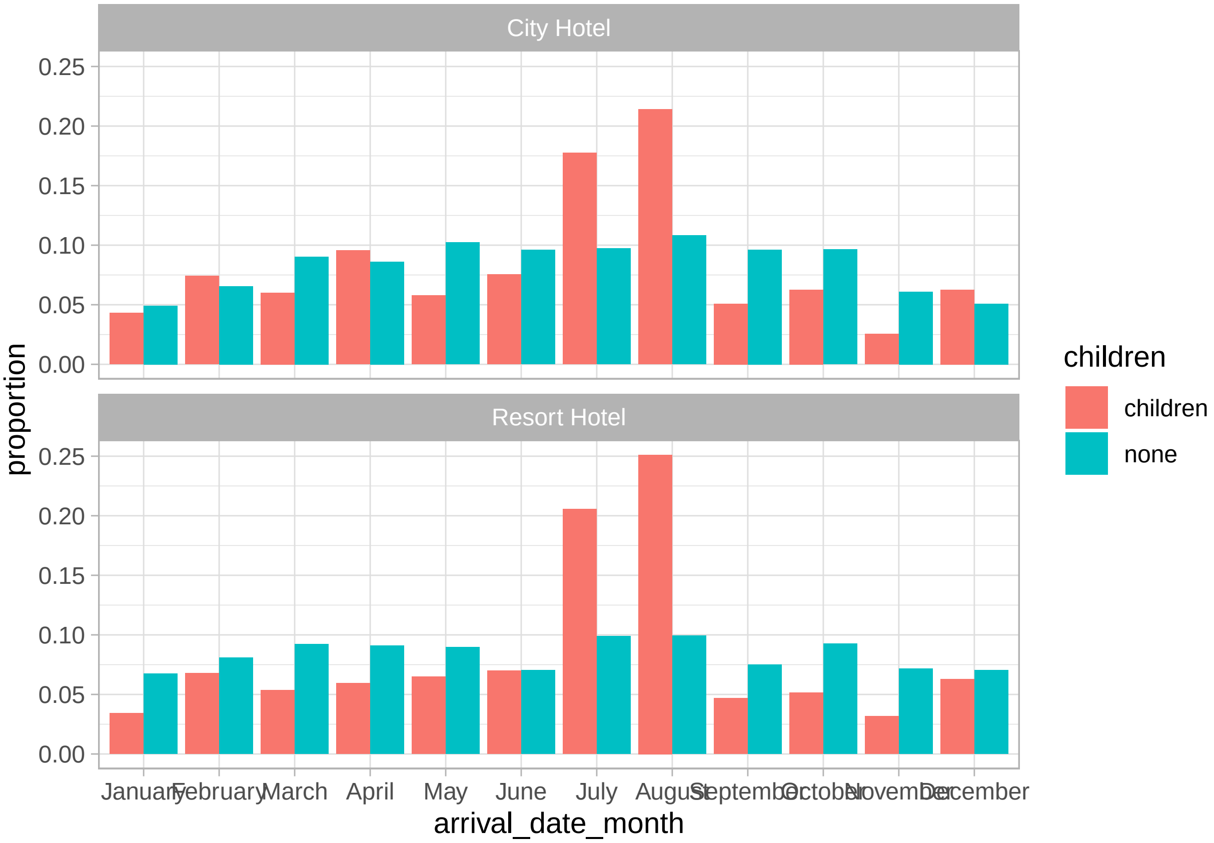
Data cleansing

```
hotel_stays<-hotels %>%
  filter(is_canceled == 0) %>%
  mutate(children = case_when(children+babies > 0 ~ "children", TRUE ~ "none"),
  hotel = recode(hotel,'Citi Hotel'='City Hotel','Reosrt Hotel'='Resort Hotel'),
  required_car_parking_spaces = case_when(required_car_parking_spaces>0~'parking', TRUE~ "none")) %>%

hotel_stays
```

```
## # A tibble: 75,166 x 30
##    hotel ‘No Response‘ lead_time arrival_date_ye~ arrival_date_mo~
##    <chr> <lgl>             <dbl>            <dbl> <chr>
##  1 Reso~ NA                  342             2015 July
##  2 Reso~ NA                  737             2015 July
##  3 Reso~ NA                    7             2015 July
##  4 Reso~ NA                   13             2015 July
##  5 Reso~ NA                   14             2015 July
##  6 Reso~ NA                   14             2015 July
##  7 Reso~ NA                    0             2015 July
##  8 Reso~ NA                    9             2015 July
##  9 Reso~ NA                   35             2015 July
## 10 Reso~ NA                   68             2015 July
## # ... with 75,156 more rows, and 25 more variables:
## #   arrival_date_week_number <dbl>, arrival_date_day_of_month <dbl>,
## #   stays_in_weekend_nights <dbl>, stays_in_week_nights <dbl>, adults <dbl>,
## #   children <chr>, meal <chr>, country <chr>, market_segment <chr>,
## #   distribution_channel <chr>, is_repeated_guest <dbl>,
## #   previous_cancellations <dbl>, previous_bookings_not_canceled <dbl>,
## #   reserved_room_type <chr>, assigned_room_type <chr>, booking_changes <dbl>,
## #   deposit_type <chr>, agent <chr>, company <chr>, days_in_waiting_list <dbl>,
## #   customer_type <chr>, adr <dbl>, required_car_parking_spaces <chr>,
## #   total_of_special_requests <dbl>, reservation_status_date <chr>
```

Explore the confirmed bookings proportion in City Hotel and Resort Hotel wrt children and without children guests

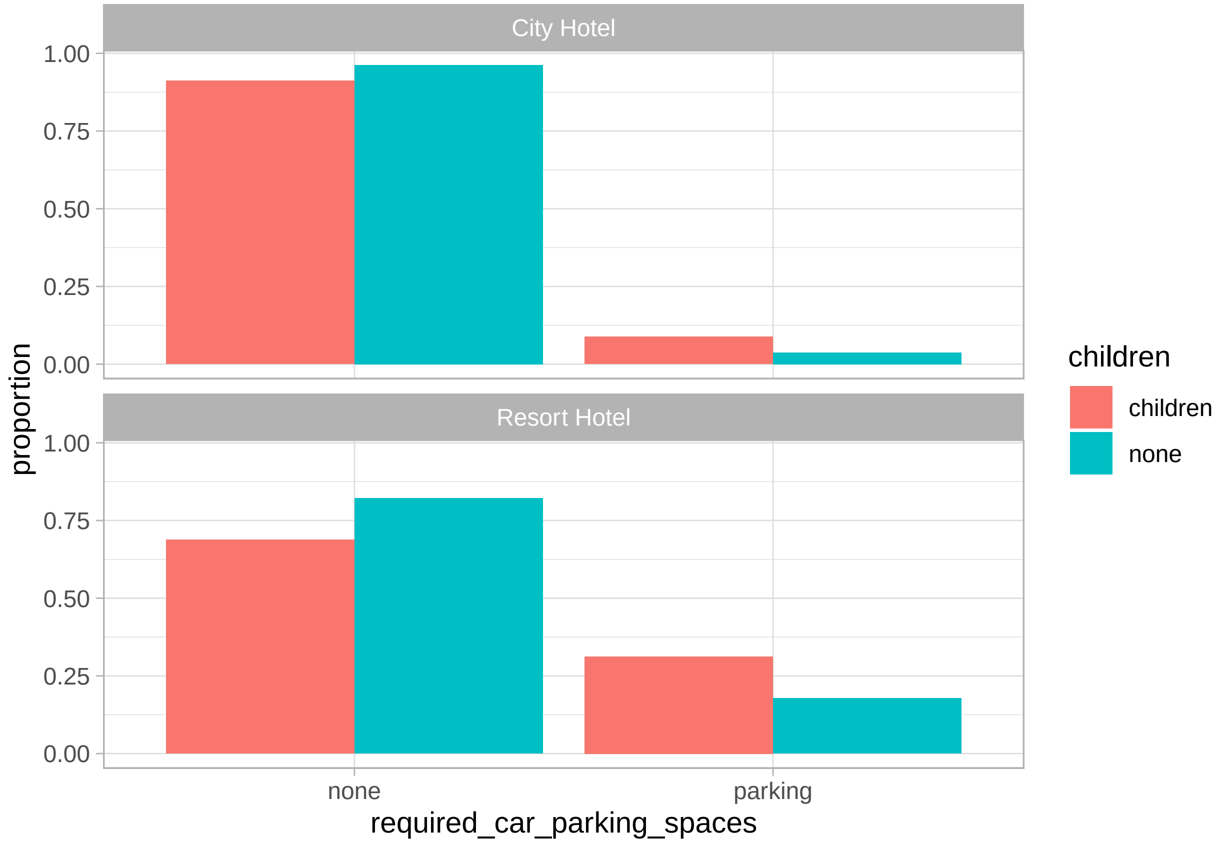
```
hotel_stays %>%
  mutate(arrival_date_month = factor(arrival_date_month,levels = month.name)) %>%
  count(hotel,arrival_date_month,children) %>%
  group_by(hotel,children) %>%
  mutate(proportion = n/sum(n)) %>%
  ggplot(aes(arrival_date_month,proportion,fill=children))+
  geom_col(position="dodge")+
  facet_wrap(~hotel,nrow=2)
```



Observations: We can see that there were more checkins by the guest having children in the month of July and August for both the hotels. However, guest without children we dont notice any considerable booking variations.

Lets compare the car parking spaces for guests with children and without children.

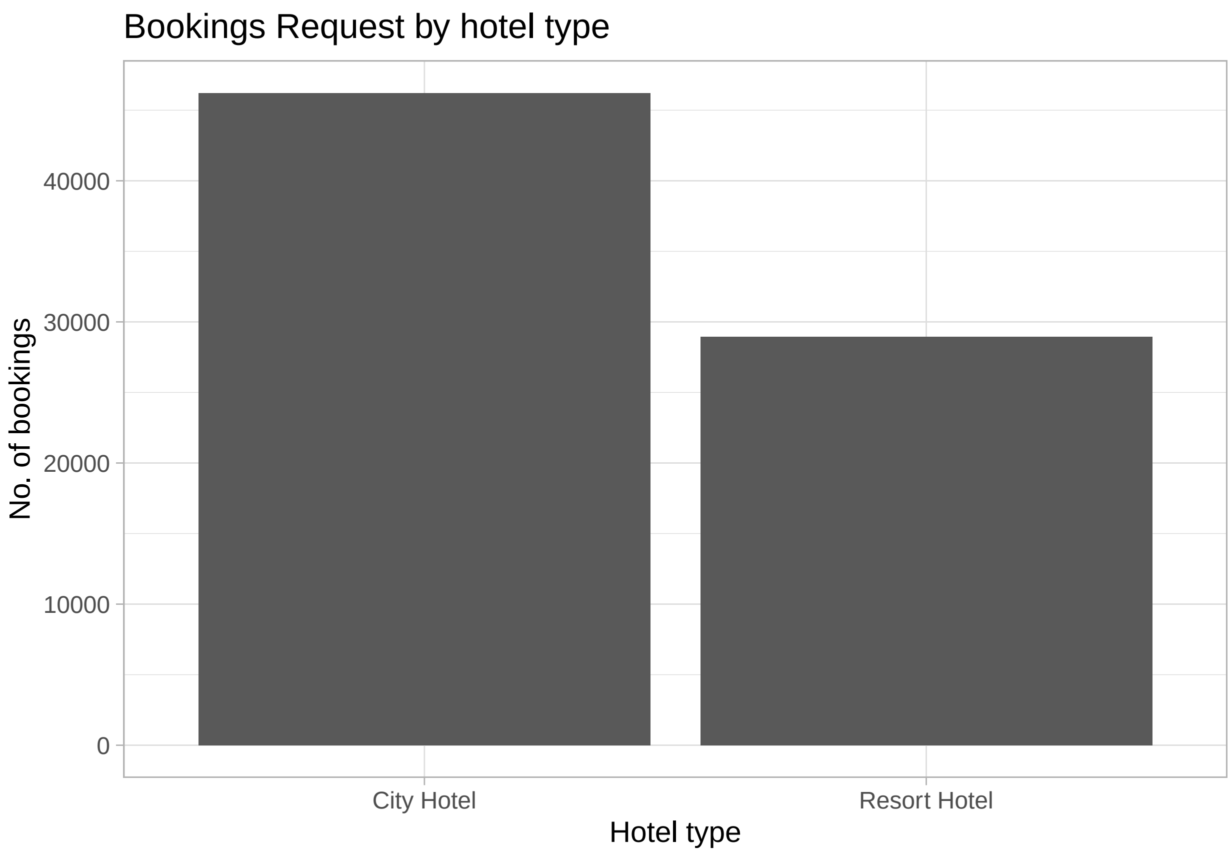
```
hotel_stays %>%
  count(hotel,required_car_parking_spaces,children) %>%
  group_by(hotel,children) %>%
  mutate(proportion = n/sum(n)) %>%
  ggplot(aes(required_car_parking_spaces,proportion,fill=children))+
  geom_col(position="dodge")+
  facet_wrap(~hotel,nrow=2)
```



Obervations: Guests having chidren tend to require car parking space. As the family size would relate to.

Exploring booking request by hotel type

```
hotel_stays %>%
ggplot(aes(x=hotel))+
  geom_bar(stat = "count")+
  labs(title = "Bookings Request by hotel type",
       x = "Hotel type",
       y = "No. of bookings")
```



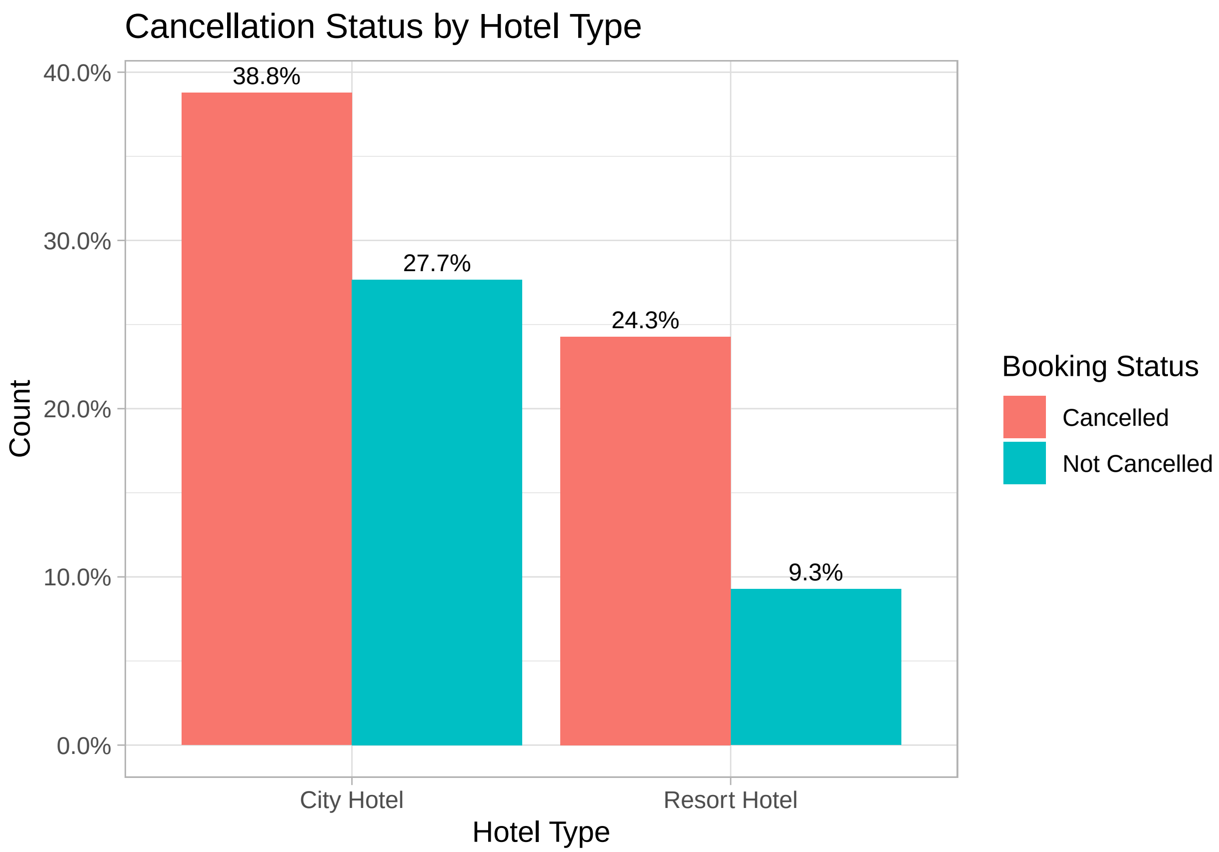
Observation: There is comapritively more booking requests in City Hotel than compared to Resort Hotel.

Check the distribution of hotel type for cancellation/confirmed statuses.

```
hotel_stays_overall<-hotels %>%
  filter(is_canceled %in% c(1,0)) %>%
  mutate(children = case_when(children+babies > 0 ~ "children", TRUE ~ "none"),
  hotel = recode(hotel,'Citi Hotel'='City Hotel','Reosrt Hotel'='Resort Hotel'),
  required_car_parking_spaces = case_when(required_car_parking_spaces>0~'parking', TRUE~ "none")) %>%


hotel_stays_overall %>%
  count(is_canceled)
```

```
## # A tibble: 2 x 2
##   is_canceled     n
##         <dbl> <int>
## 1           0 75166
## 2           1 44046
```

```
ggplot(data = hotel_stays_overall,
       aes(
         x = hotel,
         y = prop.table(stat(count)),
         fill = factor(is_canceled),
         label = scales::percent(prop.table(stat(count)))
       )) +

  geom_bar(position = position_dodge()) +
  geom_text(
    stat = "count",
    position = position_dodge(.9),
    vjust = -0.5,
    size = 3
  ) +
  scale_y_continuous(labels = scales::percent) +
  labs(title = "Cancellation Status by Hotel Type",
       x = "Hotel Type",
       y = "Count") +
  scale_fill_discrete(
    name = "Booking Status",
    breaks = c("0", "1"),
    labels = c("Cancelled", "Not Cancelled")
  )
```



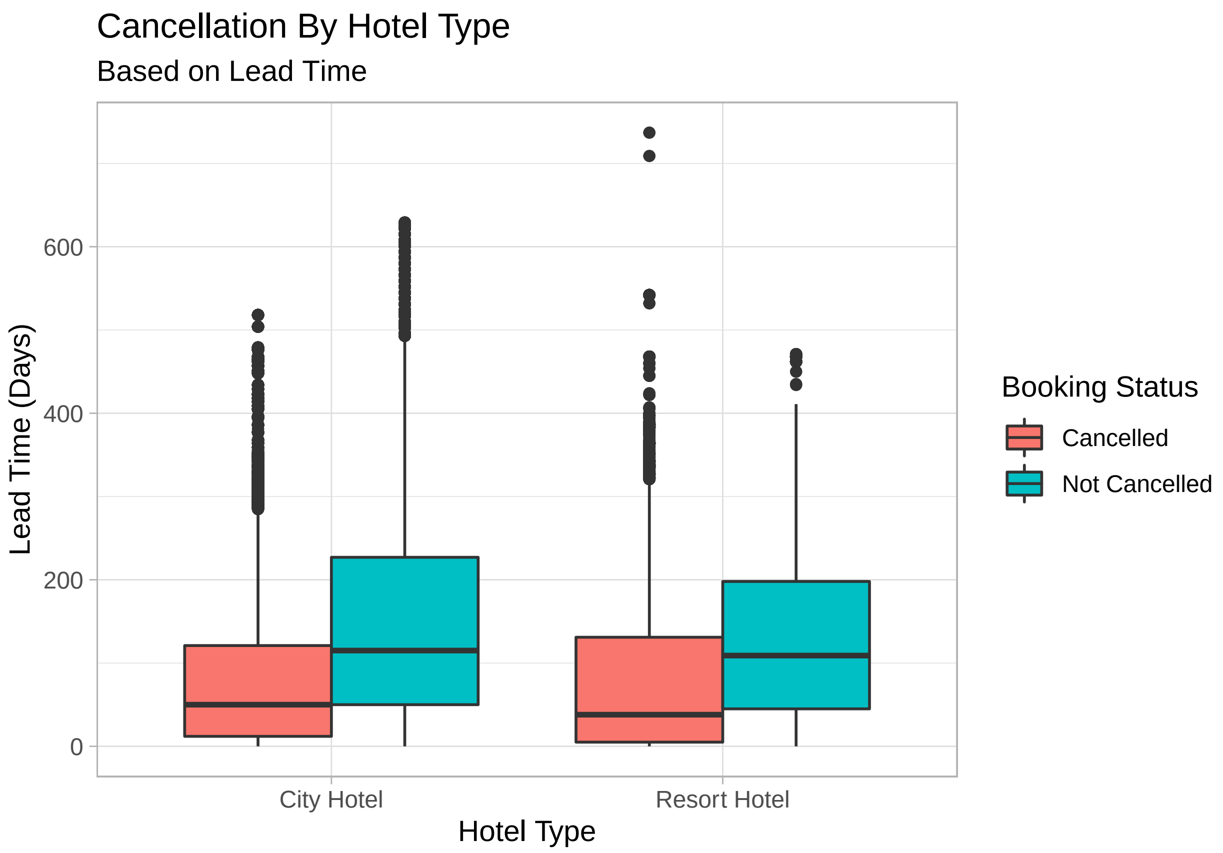
Obervation : Out of the bookings that were made more than 66% were done for City Hotel, and around

34% were for Resort Hotel. However both the hotel types that proportion of cancellation more than the confirmed status.

Cancellation ratio by hotel type based on the lead time.

Lead time is the gap between Booking made and actual date check in date

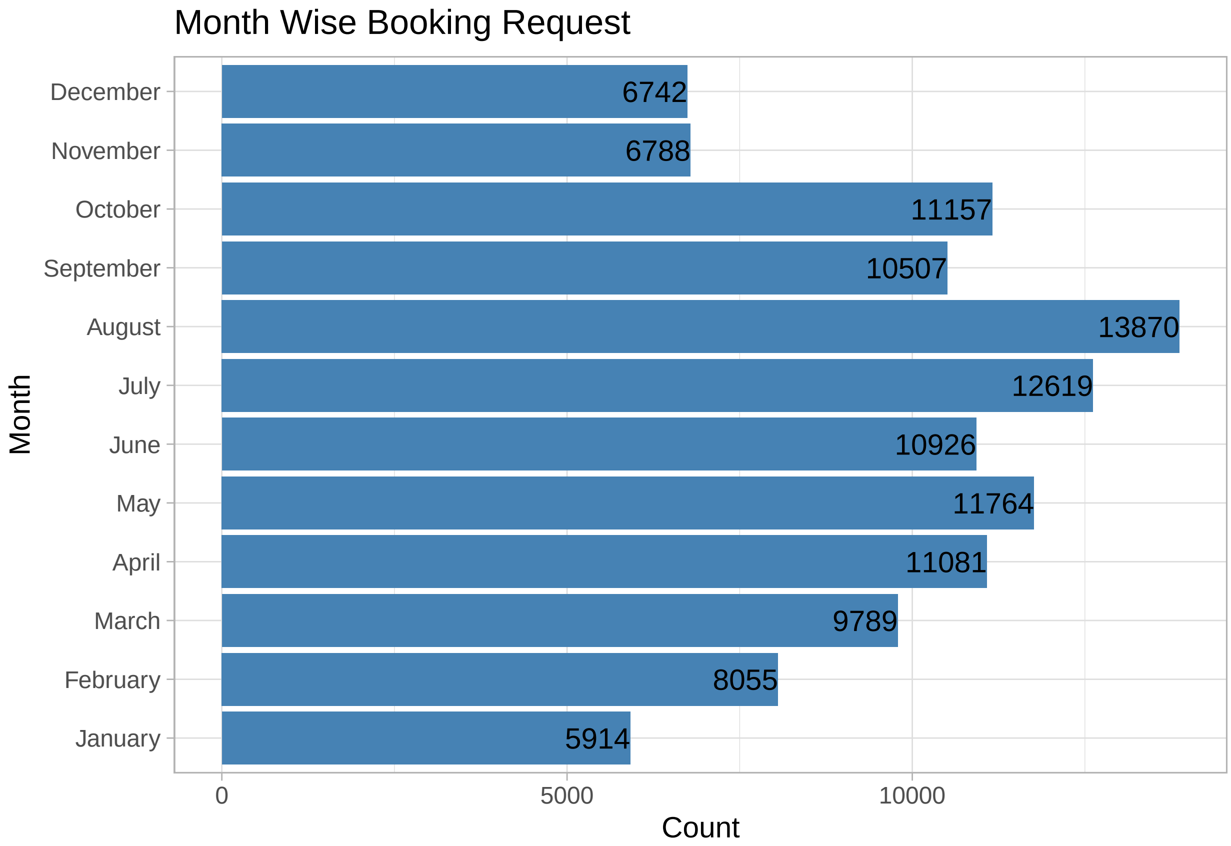
```r
ggplot(data = hotel_stays_overall, aes(
  x = hotel,
  y = lead_time,
  fill = factor(is_canceled)
)) +
  geom_boxplot(position = position_dodge()) +
  labs(
    title = "Cancellation By Hotel Type",
    subtitle = "Based on Lead Time",
    x = "Hotel Type",
    y = "Lead Time (Days)"
  ) +
  scale_fill_discrete(
    name = "Booking Status",
    breaks = c("0", "1"),
    labels = c("Cancelled", "Not Cancelled")
  )
```



Observation: We can see that most booking was cancelled very near to booked date as compared to non cancelled.

Explore month favorable for hotels when they can expect maximum demand.

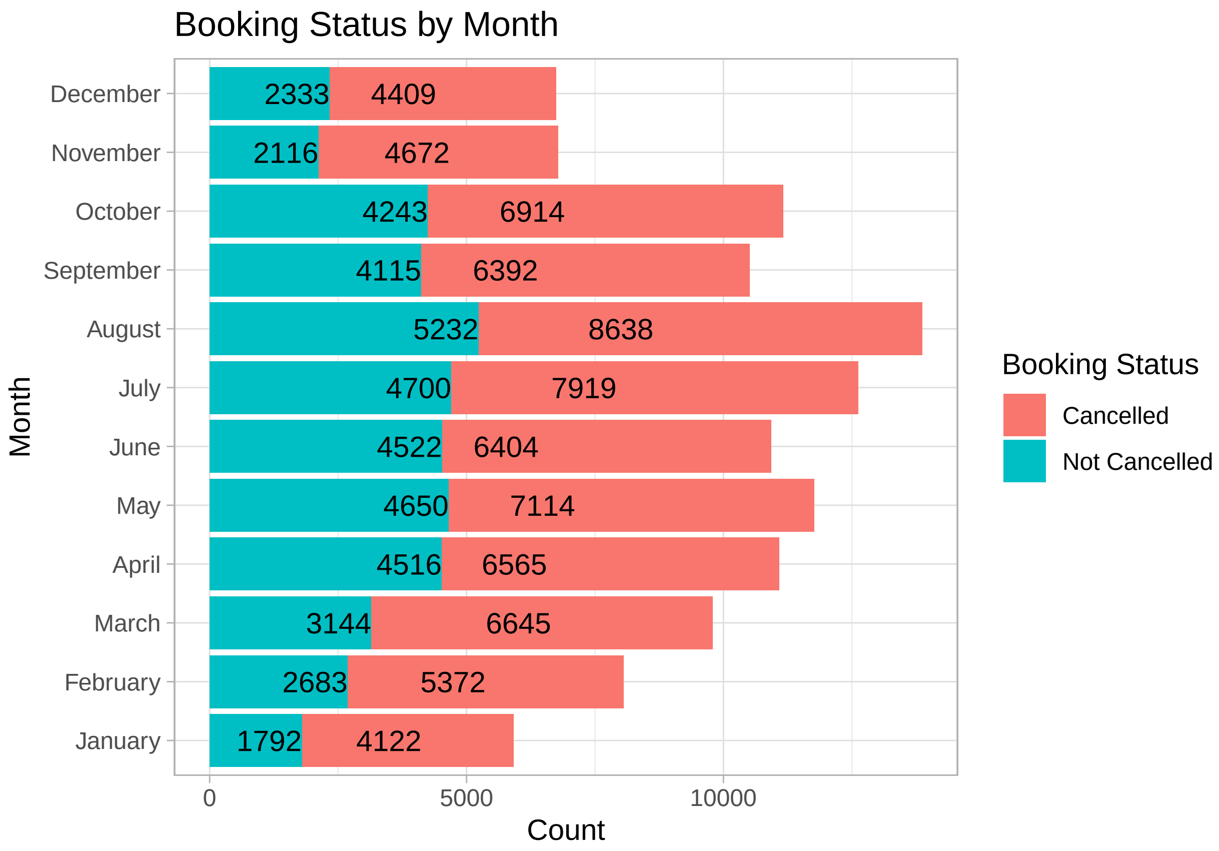
```r
hotel_stays_overall$arrival_date_month <-
  factor(hotel_stays_overall$arrival_date_month, levels = month.name)

ggplot(data = hotel_stays_overall, aes(x = arrival_date_month)) +
  geom_bar(fill = "steelblue") +
  geom_text(stat = "count", aes(label = ..count..), hjust = 1) +
  coord_flip() + labs(title = "Month Wise Booking Request",
                      x = "Month",
                      y = "Count")
```



Obervation: Month wise booking analysis depicts that more number of booking request is in July ,August respectively.

Explore the booking made in not confirmed statuses month wise.

```r
ggplot(hotel_stays_overall, aes(arrival_date_month, fill = factor(is_canceled))) +
  geom_bar() + geom_text(stat = "count", aes(label = ..count..), hjust = 1) +
  coord_flip() + scale_fill_discrete(
    name = "Booking Status",
    breaks = c("0", "1"),
    label = c("Cancelled", "Not Cancelled")
  ) +
  labs(title = "Booking Status by Month",
       x = "Month",
       y = "Count")
```

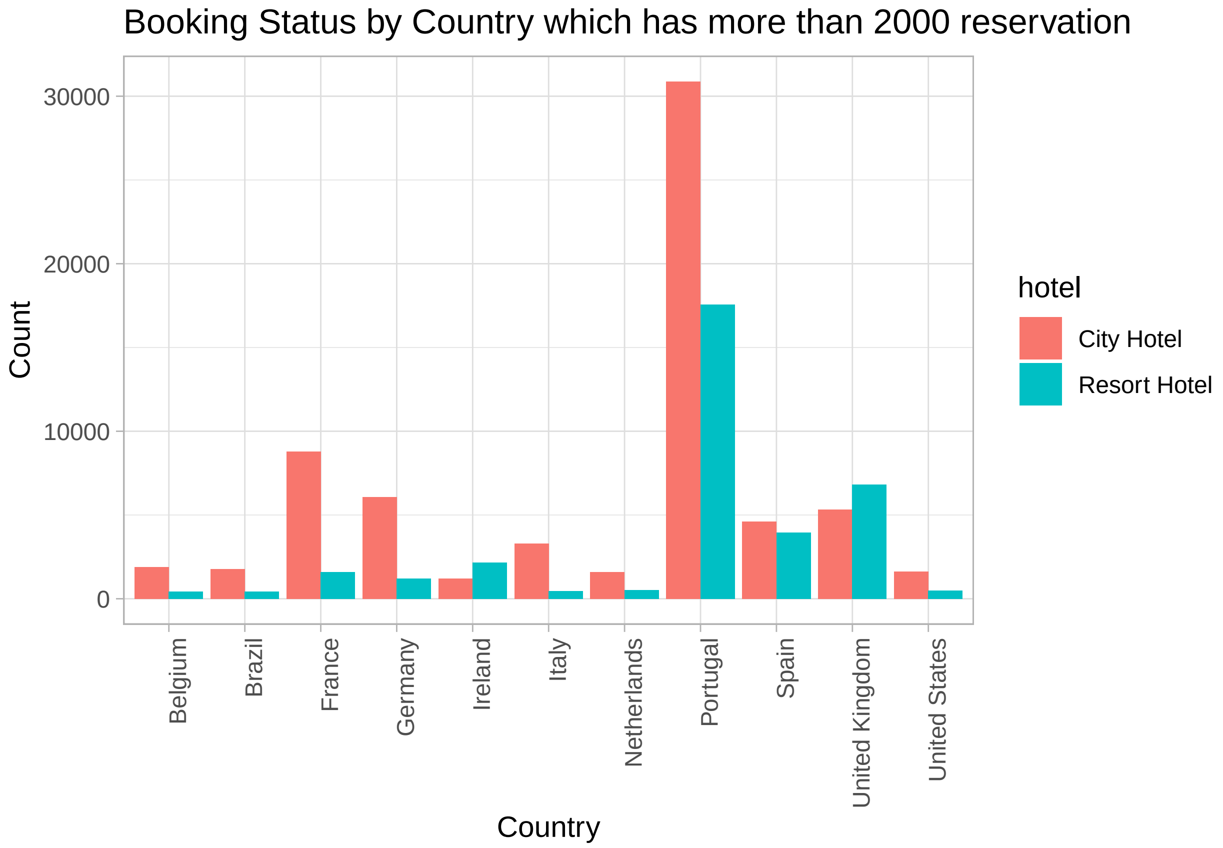


Explore guest country and booking status country wise

```r
library(countrycode)
hotel_stays_overall$country_name <- countrycode(hotel_stays_overall$country,
                                     origin = "iso3c",
                                     destination = "country.name")
```

```
## Warning in countrycode(hotel_stays_overall$country, origin = "iso3c", destination = "country.name"):
```

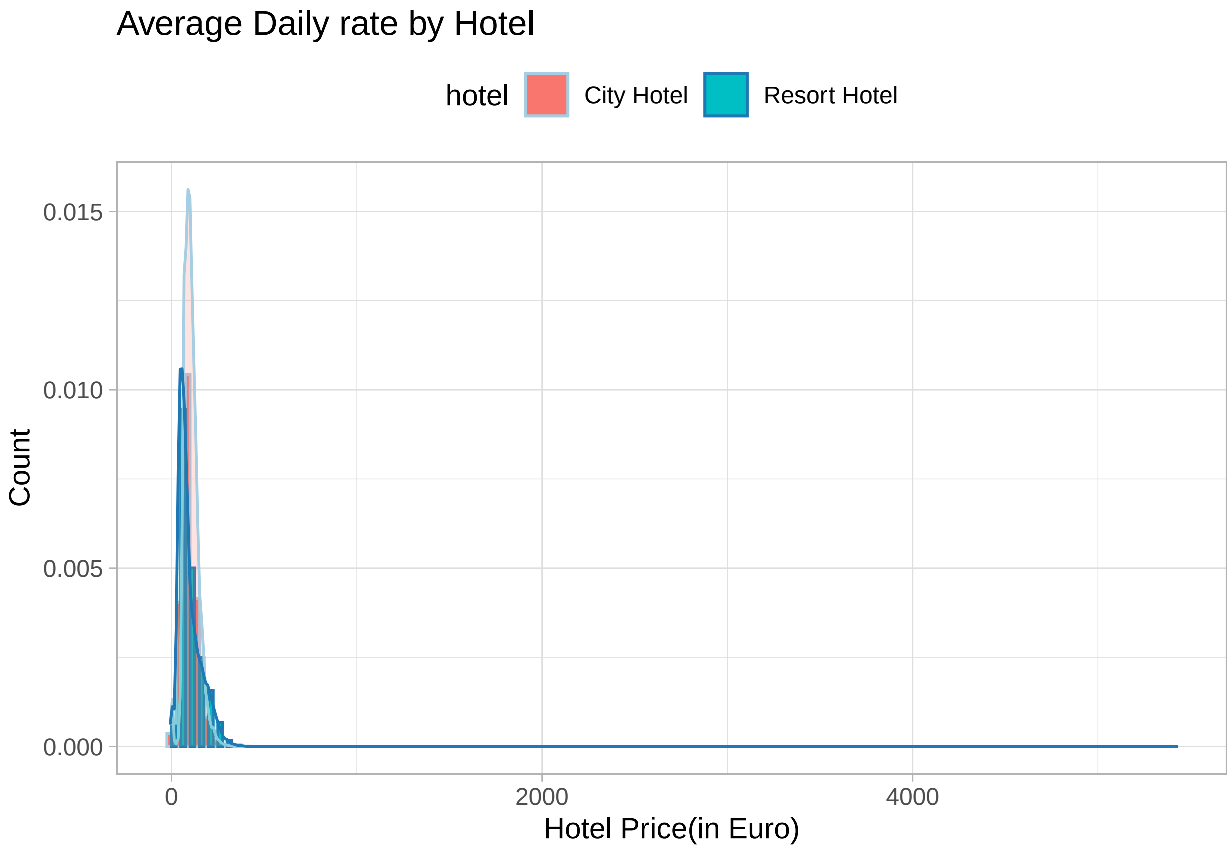
```
hotel_stays_overall %>%
  group_by(country) %>%
  filter(n()>2000) %>%
ggplot(aes(country_name, fill = hotel)) +
  geom_bar(stat = "count", position = position_dodge()) +
  labs(title = "Booking Status by Country which has more than 2000 reservation",
       x = "Country",
       y = "Count") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        panel.background = element_blank())
```



Obervations: We have filtered the reservation counts greater than 2000, to accomodate our visualization. More reservations are done by Portugal guests than any other country.

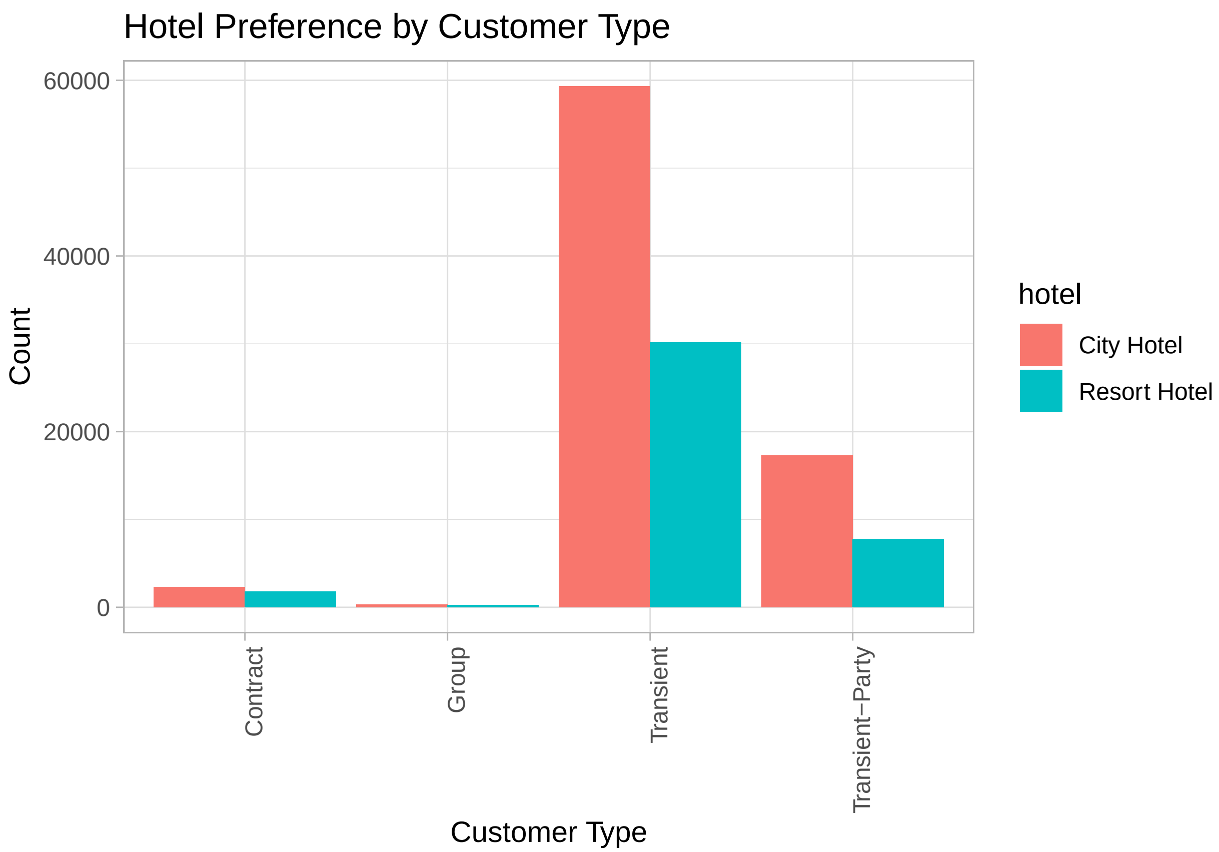
Average daily rate by hotel type

```
hotel_stays_overall %>%
ggplot(aes(x = adr, fill = hotel, color = hotel)) +
  geom_histogram(aes(y = ..density..), position = position_dodge(), binwidth = 50 ) +
  geom_density(alpha = 0.2) +
  labs(title = "Average Daily rate by Hotel",
       x = "Hotel Price(in Euro)",
      y = "Count") + scale_color_brewer(palette = "Paired") +
    theme(legend.position = "top")
```



Hotel prefernce by customer type

```
hotel_stays_overall %>%
ggplot(aes(customer_type, fill = hotel)) +
  geom_bar(stat = "count", position = position_dodge()) +
  labs(title = "Hotel Preference by Customer Type",
      x = "Customer Type",
      y = "Count") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        panel.background = element_blank())
```

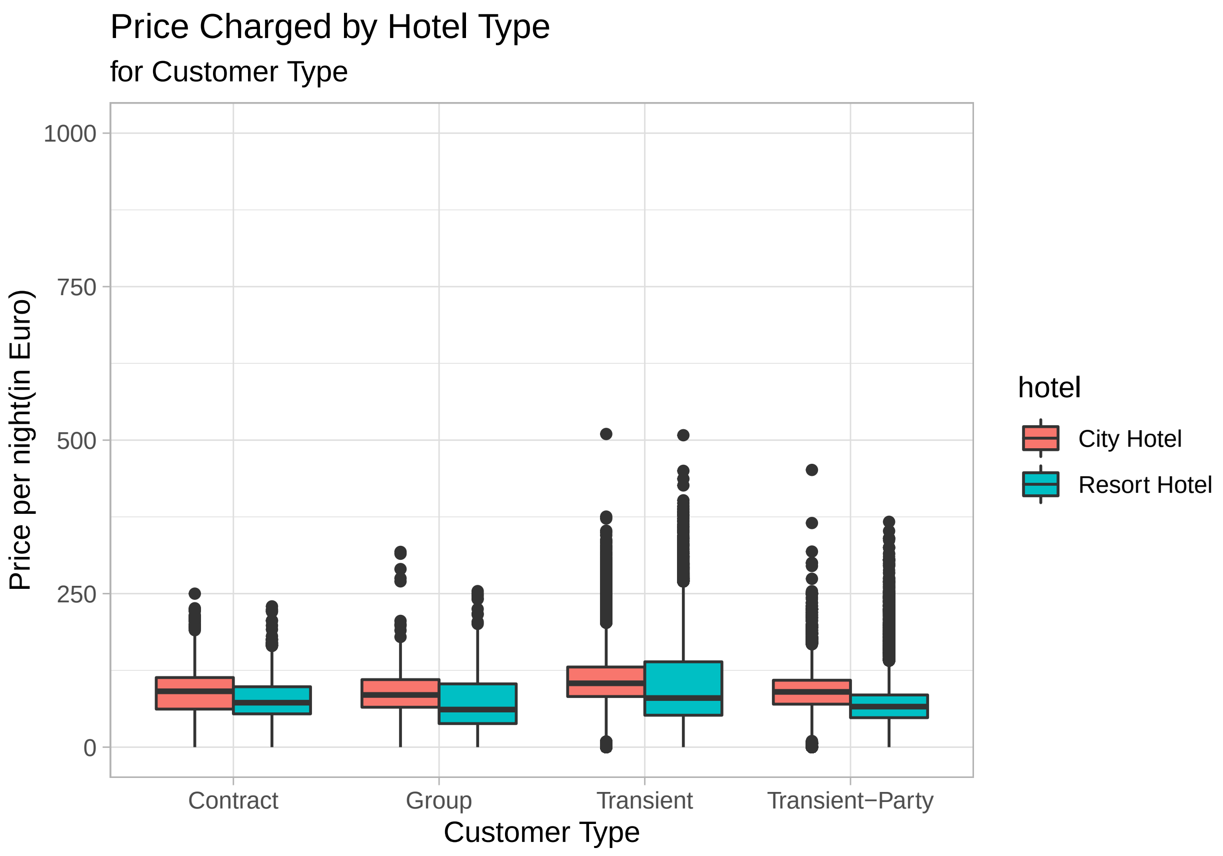


Observation : There are more Transiet type of customer in both the hotels

Does the hotel charge differently for different customer type

```
hotel_stays_overall %>%
ggplot(aes(x = customer_type, y = adr, fill = hotel)) +
  geom_boxplot(position = position_dodge()) +
  ylim(0,1000)+
  labs(title = "Price Charged by Hotel Type",
       subtitle = "for Customer Type",
       x = "Customer Type",
       y = "Price per night(in Euro)")
```

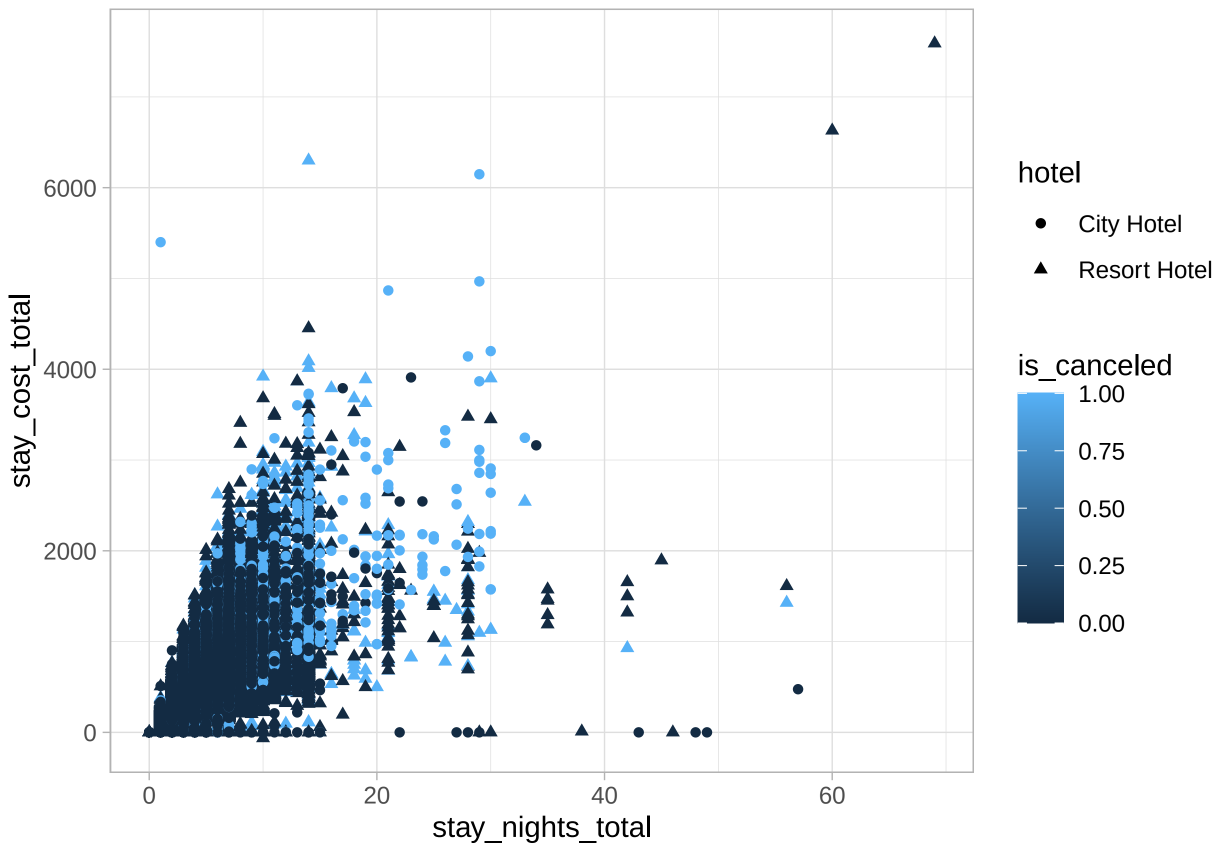
```
## Warning: Removed 2 rows containing non-finite values (stat_boxplot).
```



Scatter plots with total nights and total cost

Creating two new columns to calculate total number of days stayed and total cost

```
hotel_stays_overall %>%
mutate(stay_nights_total = stays_in_weekend_nights + stays_in_week_nights,
stay_cost_total = adr * stay_nights_total) %>%
ggplot(aes(x=stay_nights_total,y=stay_cost_total,shape=hotel,color=is_canceled))+
  geom_point(alpha=1)
```



Date of completion : 05/11/2020 Author: Priyadarshini Subramani