Hotel_dataset

2229022 Lee Seung Yeon

2024-05-14

```
setwd("/Users/seungyeonlee/Desktop/Rworkspace/MV_teamproject")
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.3     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.3     ✔ tibble    3.2.1
## ✔ lubridate 1.9.2     ✔ tidyr     1.3.0
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts
to become errors
```

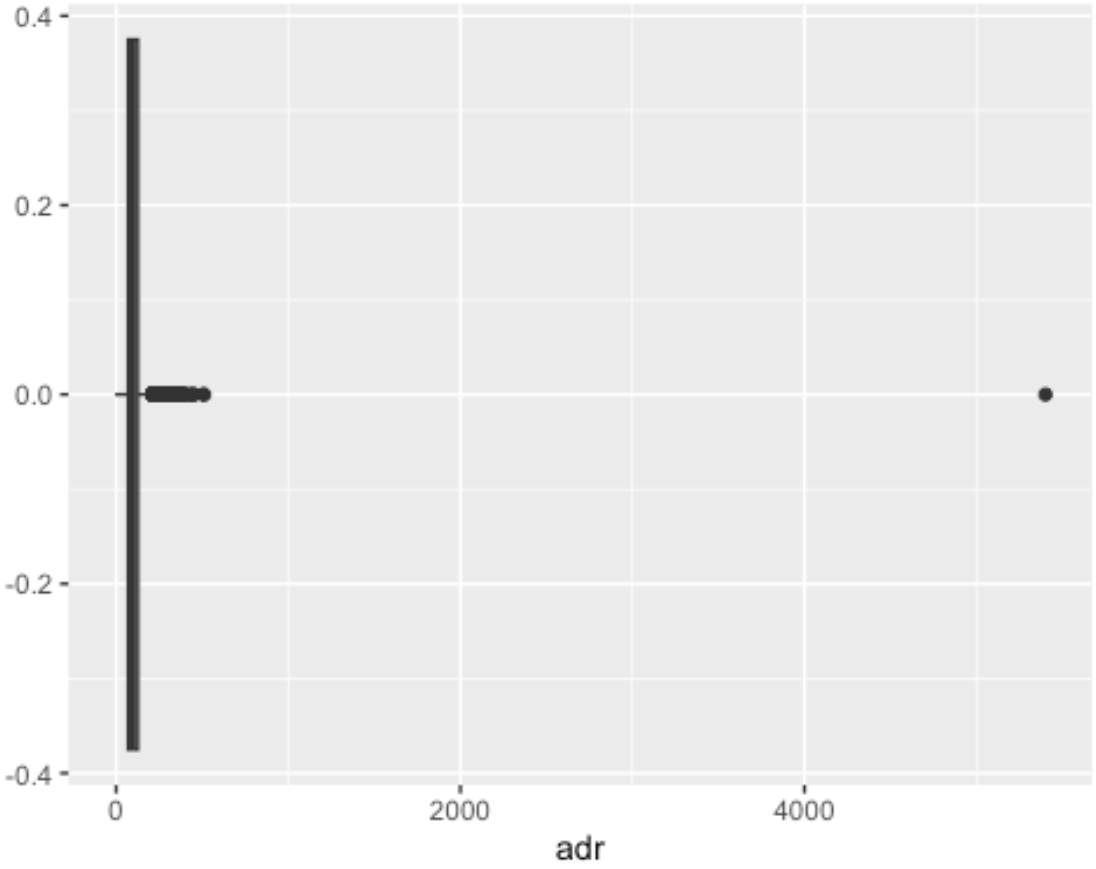
```
data = read.csv("./hotel_bookings.csv")
month_order <- c("January", "February", "March", "April", "May", "June", "July", "August",
 "September", "October", "November", "December")
data$arrival_date_month <- factor(data$arrival_date_month, levels = month_order)
head(data)
```

```
##          hotel is_canceled lead_time arrival_date_year arrival_date_month
## 1 Resort Hotel           0       342              2015               July
## 2 Resort Hotel           0       737              2015               July
## 3 Resort Hotel           0         7              2015               July
## 4 Resort Hotel           0        13              2015               July
## 5 Resort Hotel           0        14              2015               July
## 6 Resort Hotel           0        14              2015               July
##   arrival_date_week_number arrival_date_day_of_month stays_in_weekend_nights
## 1                       27                         1                       0
## 2                       27                         1                       0
## 3                       27                         1                       0
## 4                       27                         1                       0
## 5                       27                         1                       0
## 6                       27                         1                       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                    1      1        0      0   BB     GBR         Direct
## 4                    1      1        0      0   BB     GBR      Corporate
## 5                    2      2        0      0   BB     GBR      Online TA
## 6                    2      2        0      0   BB     GBR      Online TA
##   distribution_channel is_repeated_guest previous_cancellations
## 1               Direct                 0                      0
## 2               Direct                 0                      0
## 3               Direct                 0                      0
## 4            Corporate                 0                      0
```

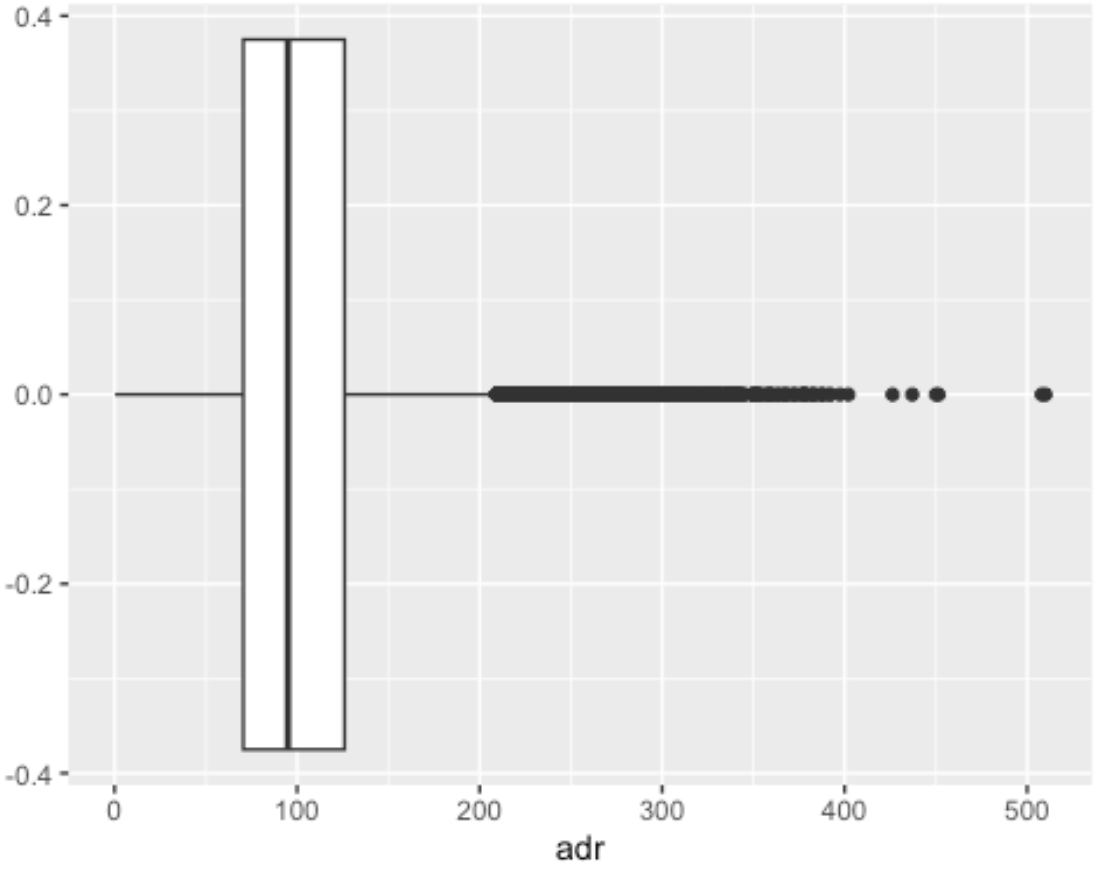
```
## 5                TA/TO                0                     0
## 6                TA/TO                0                     0
##   previous_bookings_not_canceled reserved_room_type assigned_room_type
## 1                              0                  C                  C
## 2                              0                  C                  C
## 3                              0                  A                  C
## 4                              0                  A                  A
## 5                              0                  A                  A
## 6                              0                  A                  A
##   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  NULL    NULL                    0     Transient
## 4               0   No Deposit   304    NULL                    0     Transient
## 5               0   No Deposit   240    NULL                    0     Transient
## 6               0   No Deposit   240    NULL                    0     Transient
##   adr required_car_parking_spaces total_of_special_requests reservation_status
## 1   0                           0                         0          Check-Out
## 2   0                           0                         0          Check-Out
## 3  75                           0                         0          Check-Out
## 4  75                           0                         0          Check-Out
## 5  98                           0                         1          Check-Out
## 6  98                           0                         1          Check-Out
##   reservation_status_date
## 1              2015-07-01
## 2              2015-07-01
## 3              2015-07-02
## 4              2015-07-02
## 5              2015-07-03
## 6              2015-07-03
```

adr 관찰

```
data %>% ggplot(aes(adr)) + geom_boxplot()
```

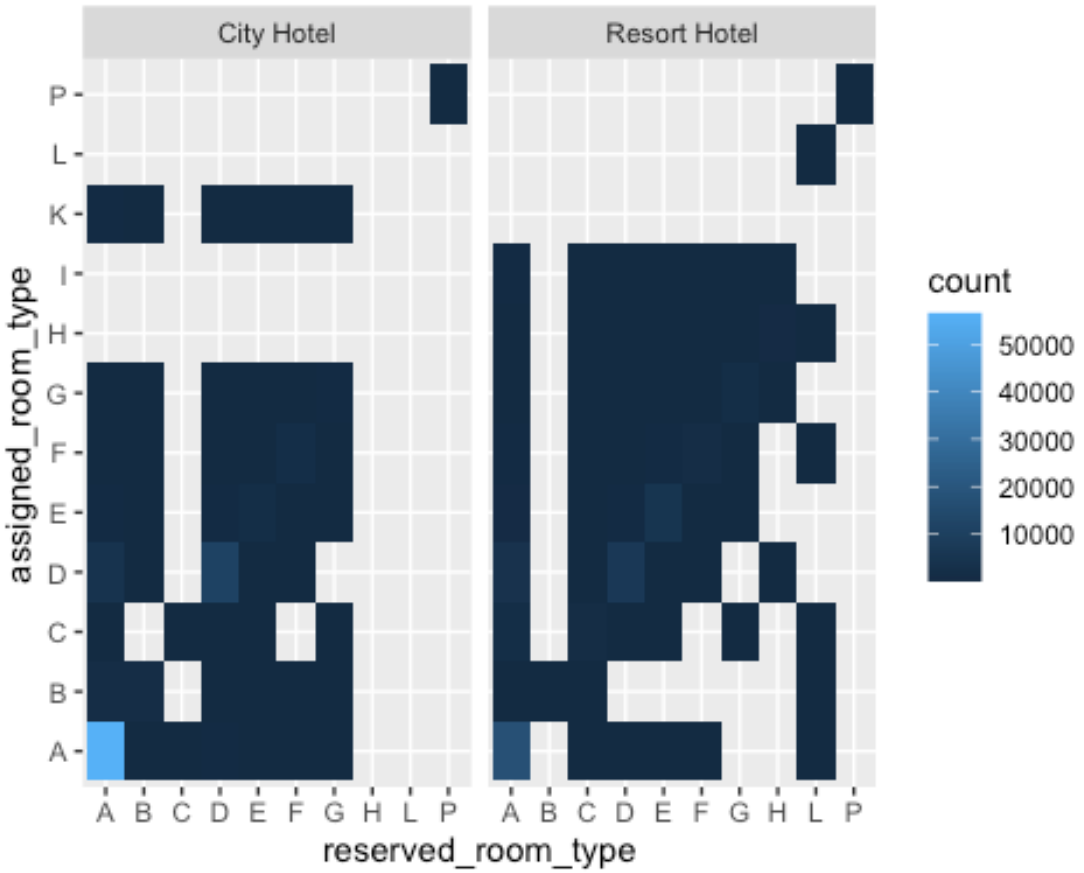



```r
# 너무 큰 값과 음수인 값 제외
data %>% filter(adr < 1000 & adr > 0) %>% ggplot(aes(adr)) + geom_boxplot()
```



방 타입 변경 여부 시각화

```
data %>%
  ggplot(aes(reserved_room_type, assigned_room_type)) + geom_bin2d() + facet_wrap("hotel")
```



hotel - room type 으로 나눠서 adr 관찰하기

room type 별로 adr 에 차이가 있을 것이다.

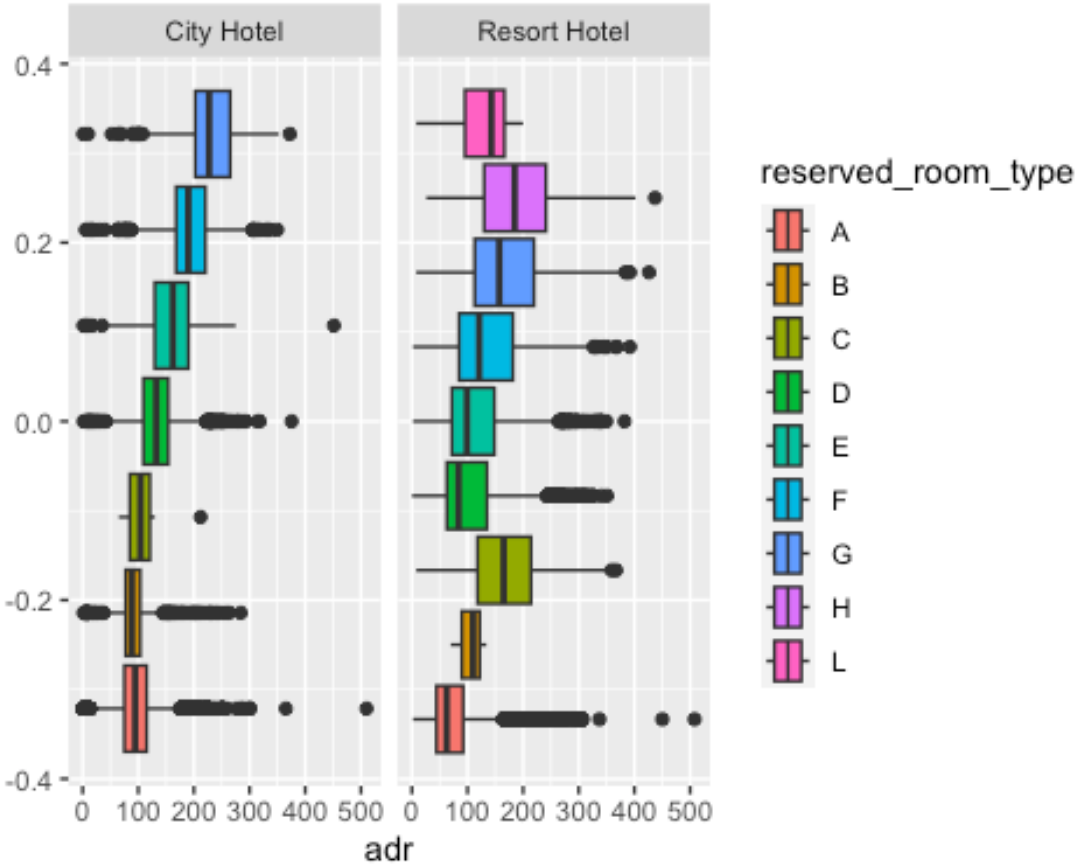
```
data %>% group_by(hotel, reserved_room_type) %>%
  summarise(mean=mean(adr), median = median(adr))
```

```
## `summarise()` has grouped output by 'hotel'. You can override using the
## `.groups` argument.
```

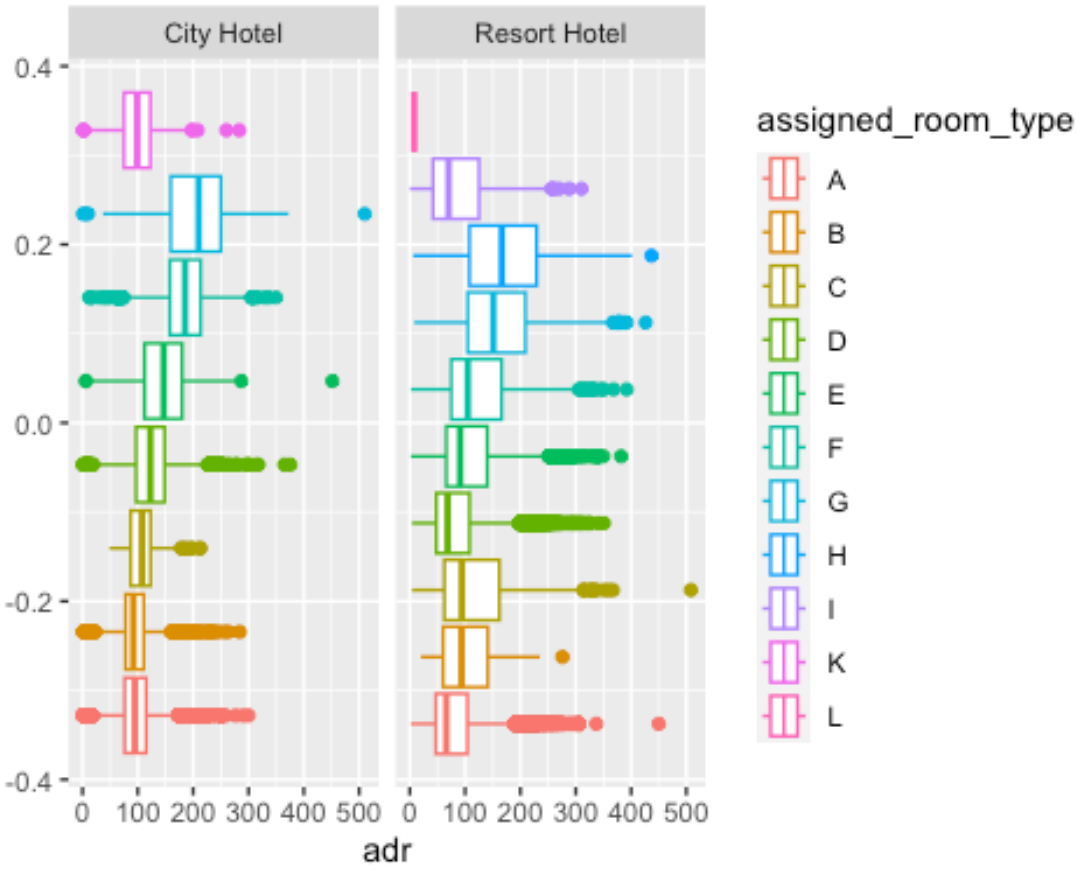
```
## # A tibble: 18 × 4
## # Groups:   hotel [2]
##    hotel        reserved_room_type  mean median
##    <chr>        <chr>              <dbl>  <dbl>
##  1 City Hotel   A                   96.2   93.8
##  2 City Hotel   B                   90.3   87
##  3 City Hotel   C                   85.5   89.5
##  4 City Hotel   D                  131.   132.
##  5 City Hotel   E                  157.   162.
##  6 City Hotel   F                  189.   189.
##  7 City Hotel   G                  202.   222.
##  8 City Hotel   P                    0      0
##  9 Resort Hotel A                   76.2   60.6
## 10 Resort Hotel B                  105.   110
## 11 Resort Hotel C                  161.   163.
## 12 Resort Hotel D                  104.    83
## 13 Resort Hotel E                  114.    98
```

```
## 14 Resort Hotel F                  133.   118.
## 15 Resort Hotel G                  168.   155
## 16 Resort Hotel H                  188.   184
## 17 Resort Hotel L                  125.   143
## 18 Resort Hotel P                    0      0
```

```
data %>% filter(adr < 1000 & adr > 0) %>%
  ggplot() + geom_boxplot(aes(adr, fill= reserved_room_type)) +
  facet_wrap("hotel")
```

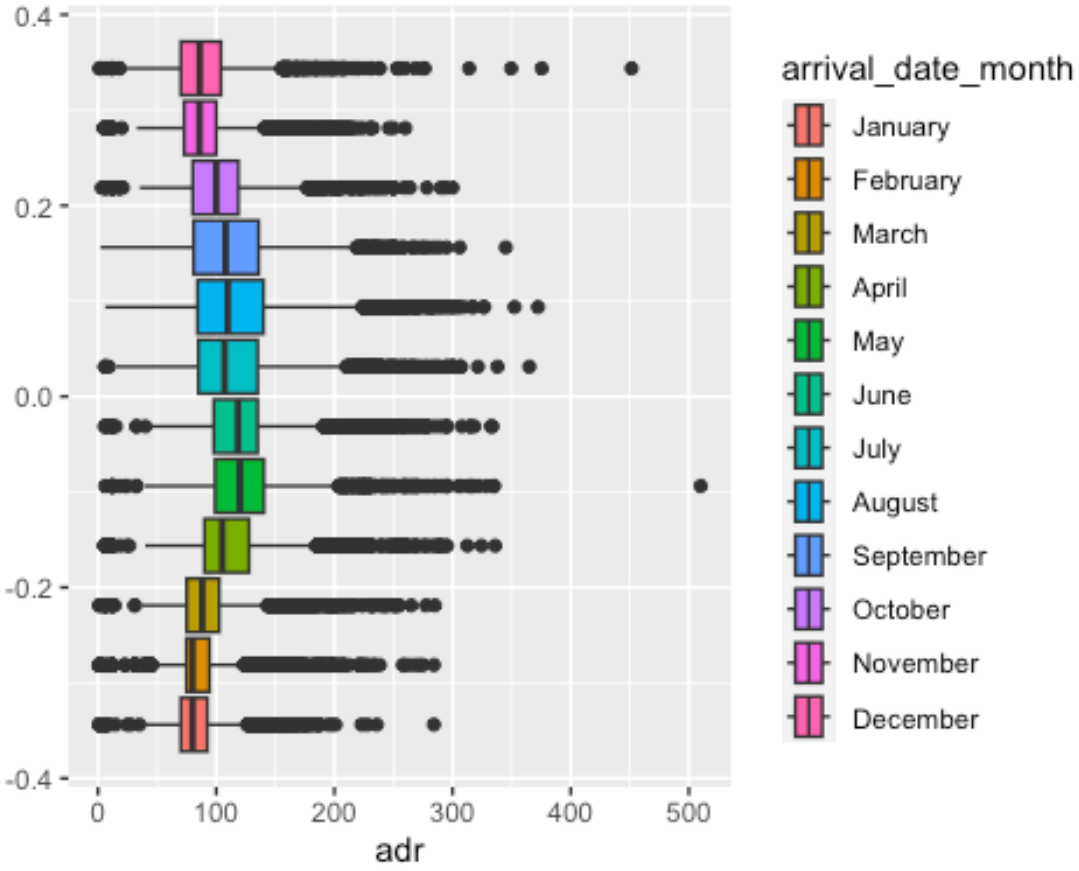



```
data %>% filter(adr < 1000 & adr > 0) %>%
  ggplot(aes(adr, color = assigned_room_type)) + geom_boxplot() +
  facet_wrap("hotel")
```

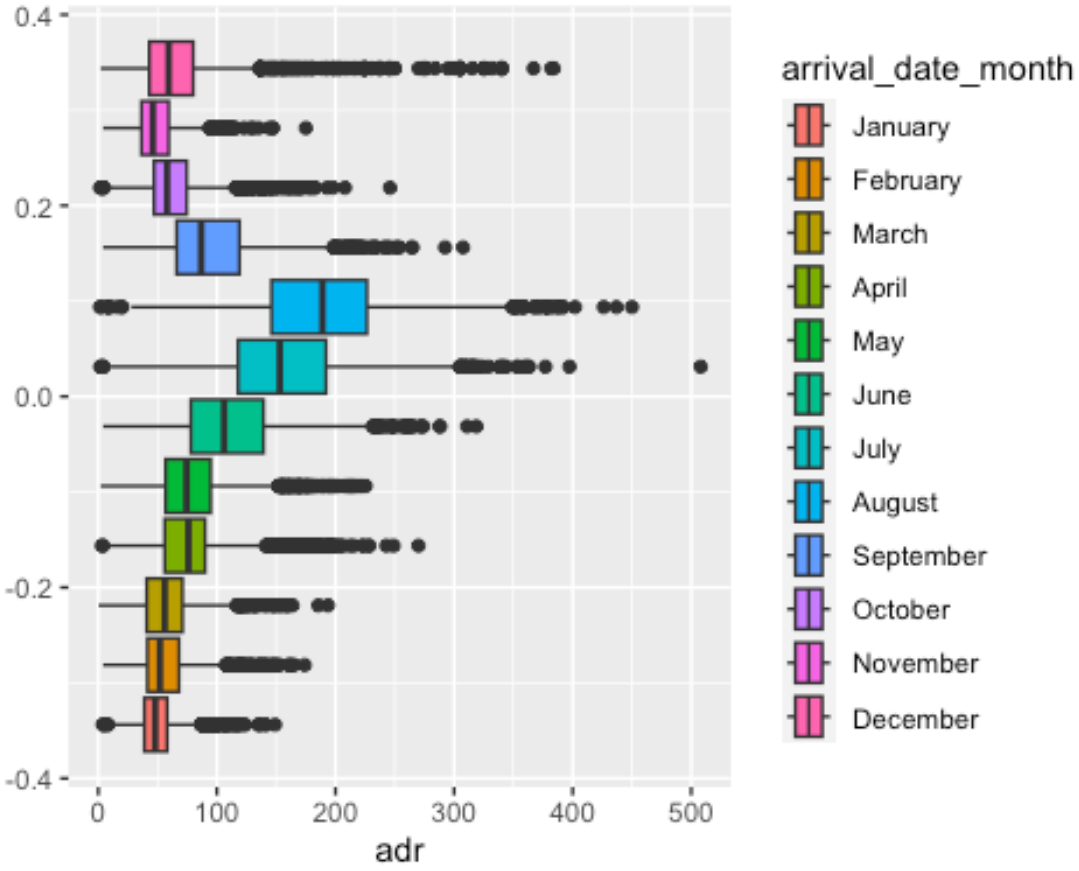


hotel 별로 월별 adr 관찰

```
data %>% filter(adr < 1000 & adr > 0) %>%
  filter(hotel == "City Hotel") %>%
  ggplot(aes(adr, fill=arrival_date_month)) + geom_boxplot()
```



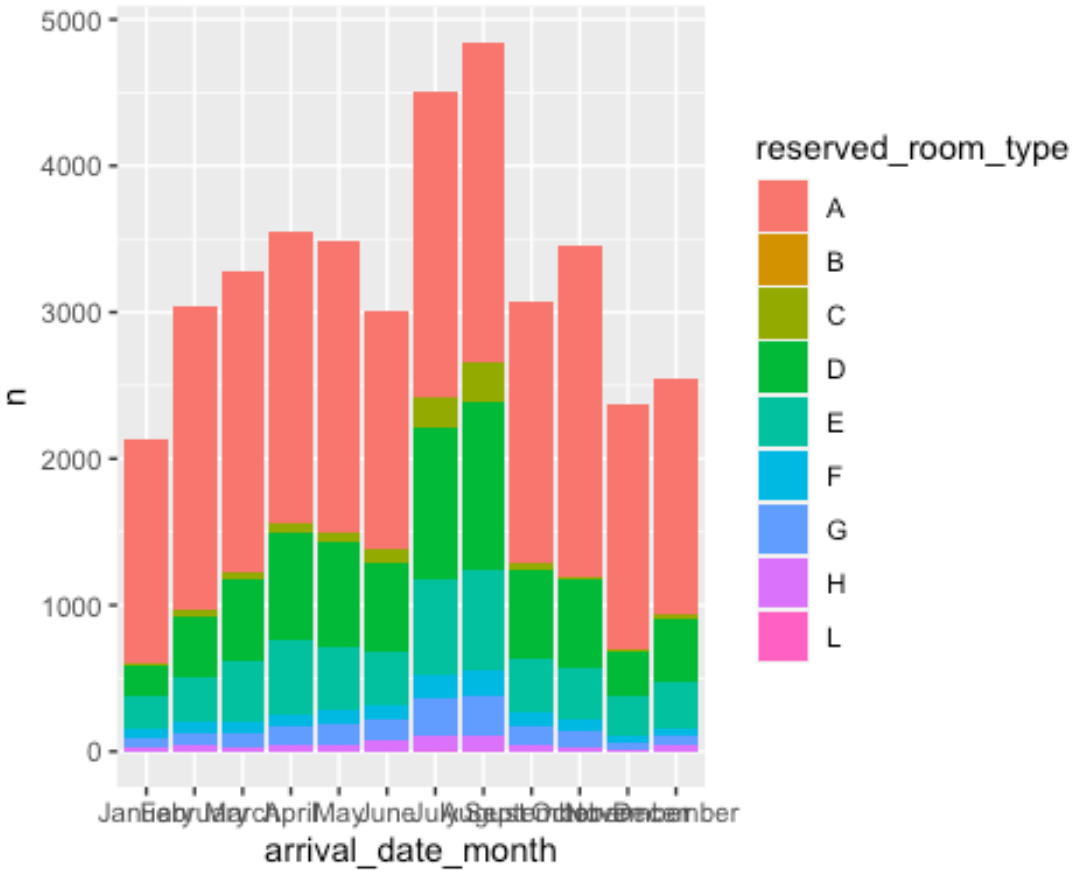

```
data %>% filter(adr < 1000 & adr > 0) %>%
  filter(hotel == "Resort Hotel") %>%
  ggplot(aes(adr, fill=arrival_date_month)) + geom_boxplot()
```



월별 인기있는 room_type

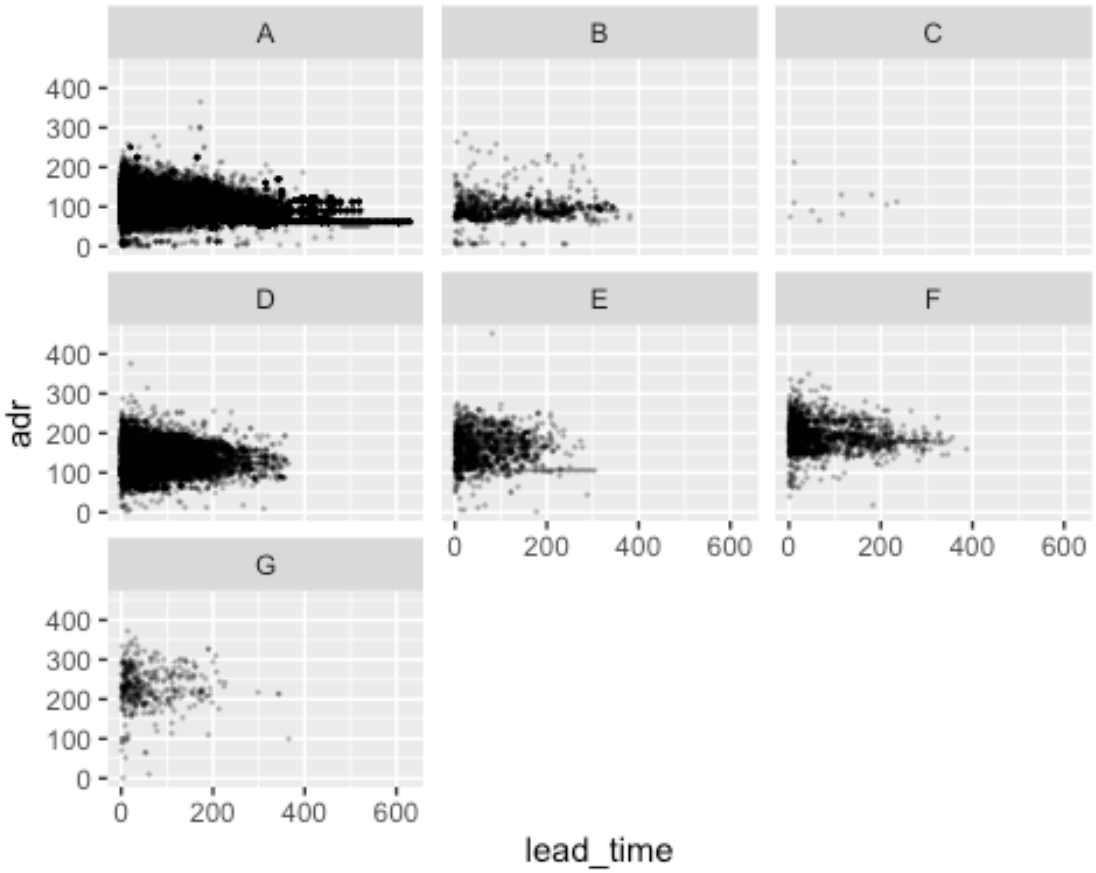
```
data %>% filter(adr < 1000 & adr > 0) %>%
  group_by(hotel, arrival_date_month, reserved_room_type) %>%
  summarise(n=n()) %>% filter(hotel=="Resort Hotel") %>%
  ggplot(aes(arrival_date_month, n, fill=reserved_room_type)) + geom_bar(stat="identity")
```

```
## `summarise()` has grouped output by 'hotel', 'arrival_date_month'. You can
## override using the `.groups` argument.
```

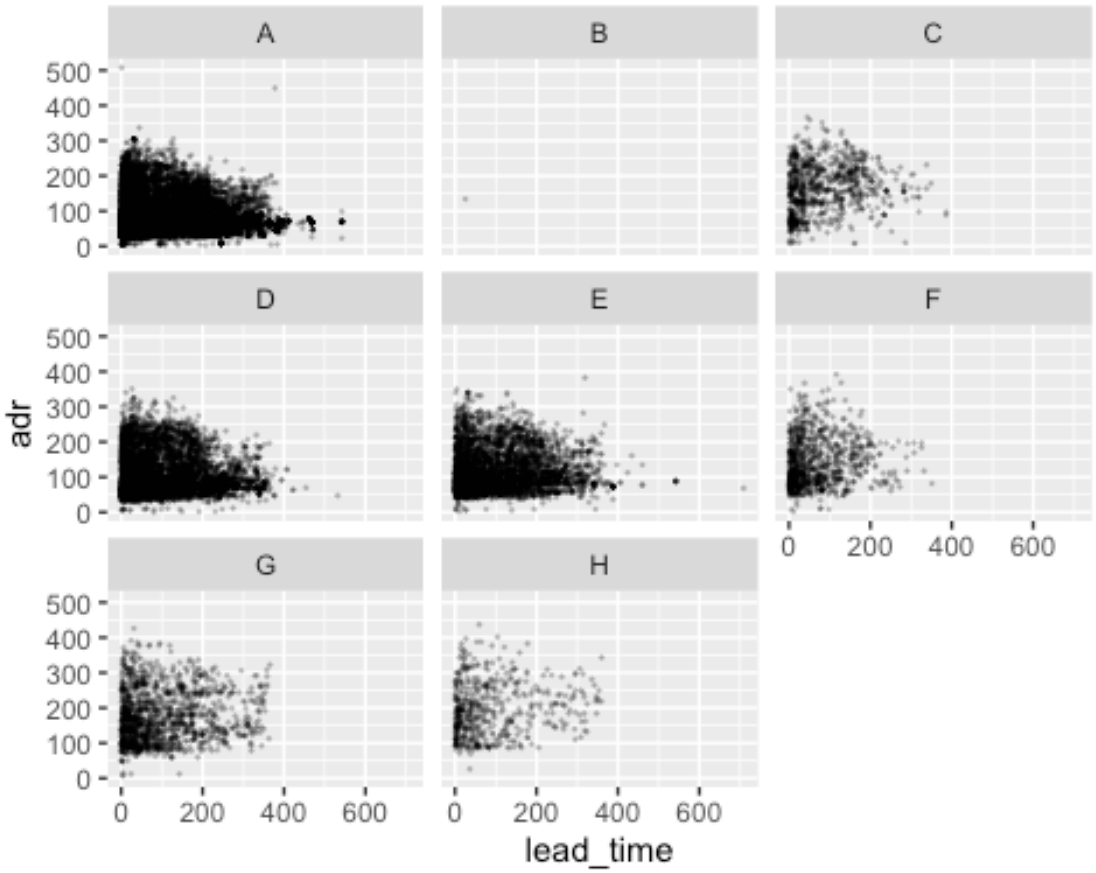


adr 과 lead_time 의 관계

```
data %>% filter(adr < 1000 & adr > 0, lead_time > 0) %>%
  filter(hotel == "City Hotel") %>%
  ggplot(aes(lead_time, adr)) + geom_point(alpha=0.2, size=0.2) + facet_wrap("reserved_ro
om_type")
```

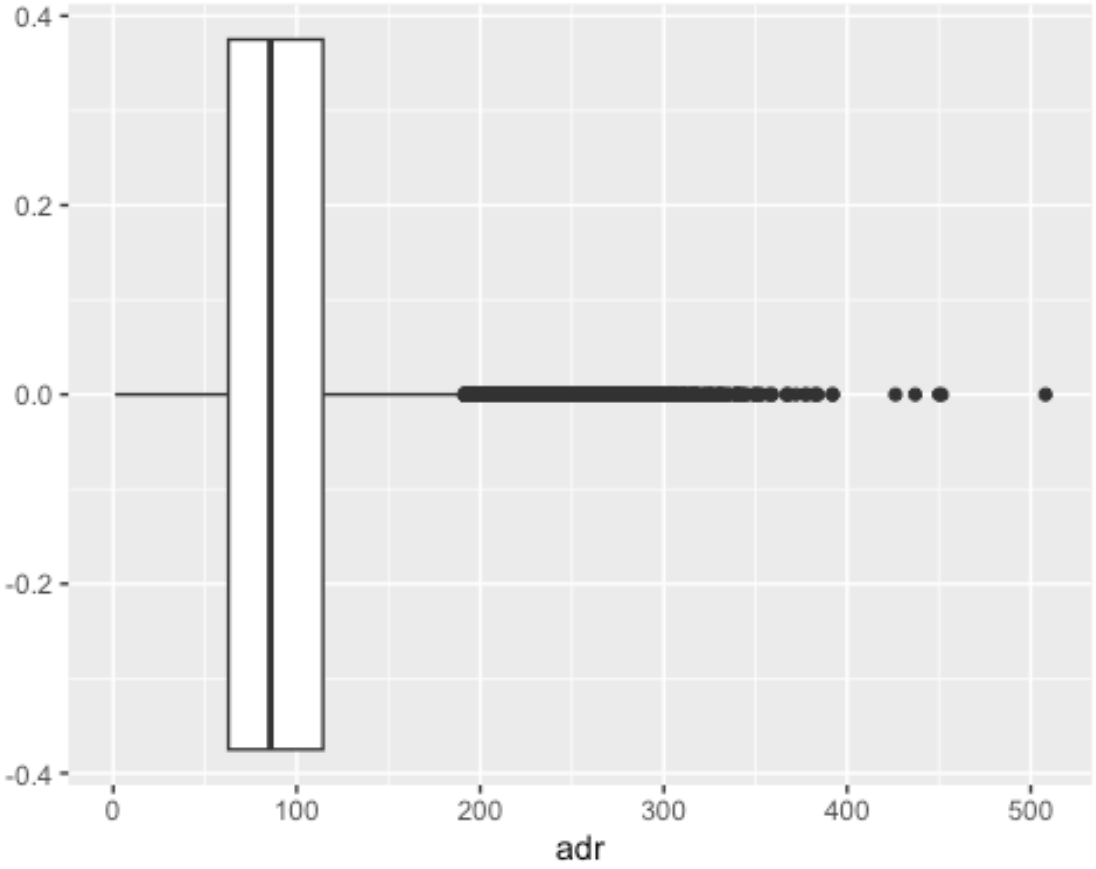



```
data %>% filter(adr < 1000 & adr > 0, lead_time > 0) %>%
  filter(hotel == "Resort Hotel") %>%
  ggplot(aes(lead_time, adr)) + geom_point(alpha=0.2, size=0.2) + facet_wrap("reserved_ro
om_type")
```



포르투갈 현지 사람들의 adr 과 그 외 국가에서 온 사람들의 adr 비교

```
data %>% filter(country == "PRT", adr >0, adr<1000) %>% ggplot(aes(adr)) + geom_boxplot()
```




```
data %>% filter(country != "PRT", adr >0, adr<1000) %>% ggplot(aes(adr)) + geom_boxplot()
```