# R Notebook

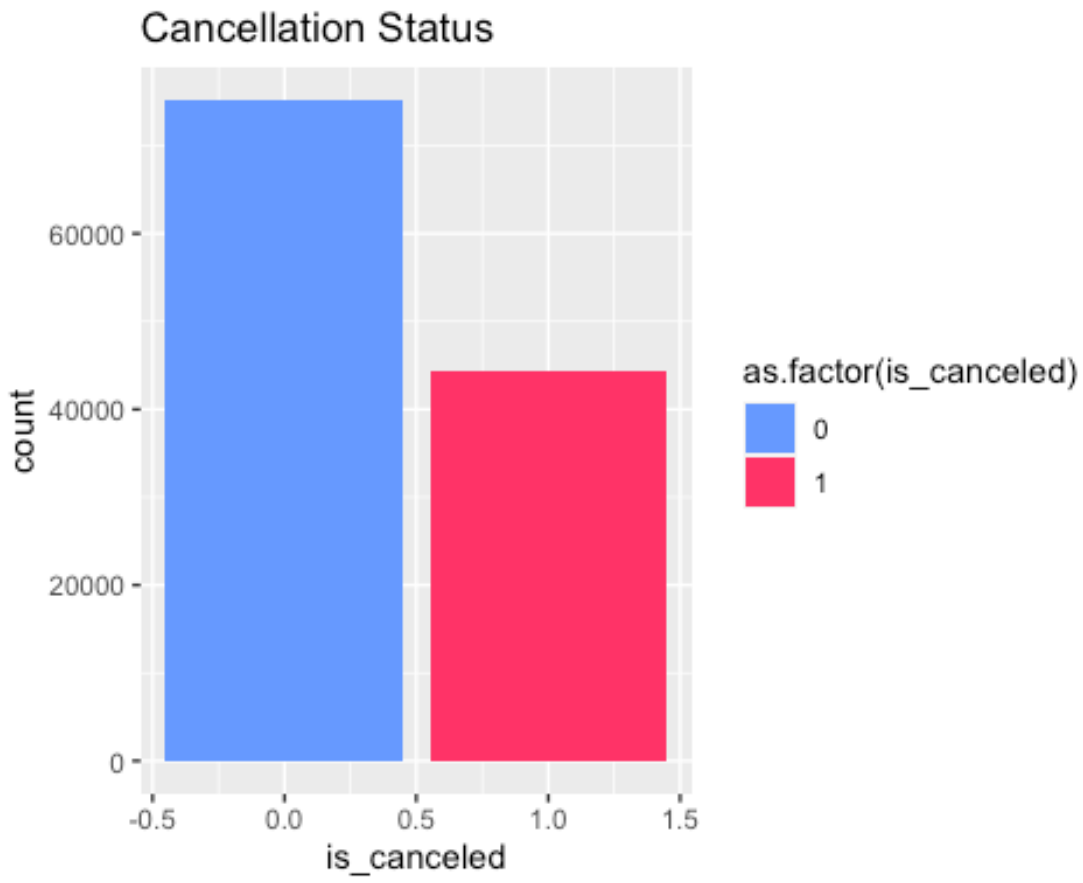
```
library(tidyverse)

## ── Attaching packages
─────────────────────────────────────────────────────────────────────────────── tidyverse 1.3.2 ──
## ✓ ggplot2 3.4.4     ✓ purrr   1.0.2
## ✓ tibble  3.2.1     ✓ dplyr   1.1.3
## ✓ tidyr   1.2.1     ✓ stringr 1.4.0
## ✓ readr   2.1.2     ✓ forcats 1.0.0
## ── Conflicts
─────────────────────────────────────────────────────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()

library(colorspace)

boolean_palette = c("TRUE"="#FF3366",
                    "FALSE"="#6699FF",
                    "1"="#FF3366",
                    "0"="#6699FF")

df = read.csv("/Users/jimin/Desktop/ㅈㅣㅁㅣㄴ/ewha/2024-
1/ㄷㅏㅂㅕㄴㄹㅑㅇㅂㅜㄴㅅㅓㄱ/ㅌㅗㅇㅌㅗㅇㅂㅐ/dataset/hotel_bookings.csv")
head(df)

##          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
```

### 캔슬이 유독 많다.

```
df %>% ggplot(aes(x = is_canceled, fill = as.factor(is_canceled))) +
  geom_bar() +
  ggtitle("Cancellation Status") +
  scale_fill_manual(values = boolean_palette)
```




```
df %>% summarize(canceled_percent = sum(is_canceled) / length(is_canceled))
```

```
##   canceled_percent
## 1        0.3704163
```

## 재방문 / 첫방문에 따른 캔슬 빈도수 차이

```
# 한 번도 방문하지 않은 고객이 캔슬한 횟수
df %>%
  mutate(is_repeated_guest = as.logical(is_repeated_guest)) %>%
  filter(is_repeated_guest == F) %>%
  filter(previous_cancellations != 0) %>%
  count(previous_cancellations, name = "freq") %>%
  summarise(weight_sum = sum(previous_cancellations * freq)) %>% as.numeric()
```

```
## [1] 8611
```

```
#  방문 경험이 있는 사람이 캔슬한 횟수
df %>%
  mutate(is_repeated_guest = as.logical(is_repeated_guest)) %>%
  filter(is_repeated_guest == T) %>%
  filter(previous_cancellations != 0) %>%
  count(previous_cancellations, name = "freq") %>%
  summarise(weight_sum = sum(previous_cancellations * freq))  %>% as.numeric()
```

```
## [1] 1790
```

한 번도 방문하지 않았던 사람들이 이전에 취소한 횟수가 더 많았다.

```
# 한 번도 방문하지 않은 고객이 과거에 캔슬한 횟수
df %>%
    mutate(is_repeated_guest = as.logical(is_repeated_guest)) %>%
    filter(is_repeated_guest == F) %>%
    filter(previous_cancellations != 0) %>%
    select(previous_cancellations) %>%
    table()
```

```
## .
##    1    2    3    6   11   14   19   24   25   26
## 5358   40   13    6    8   14   19   48   25   26
```

```
#  방문 경험이 있는 사람이 과거에 캔슬한 횟수
df %>%
```

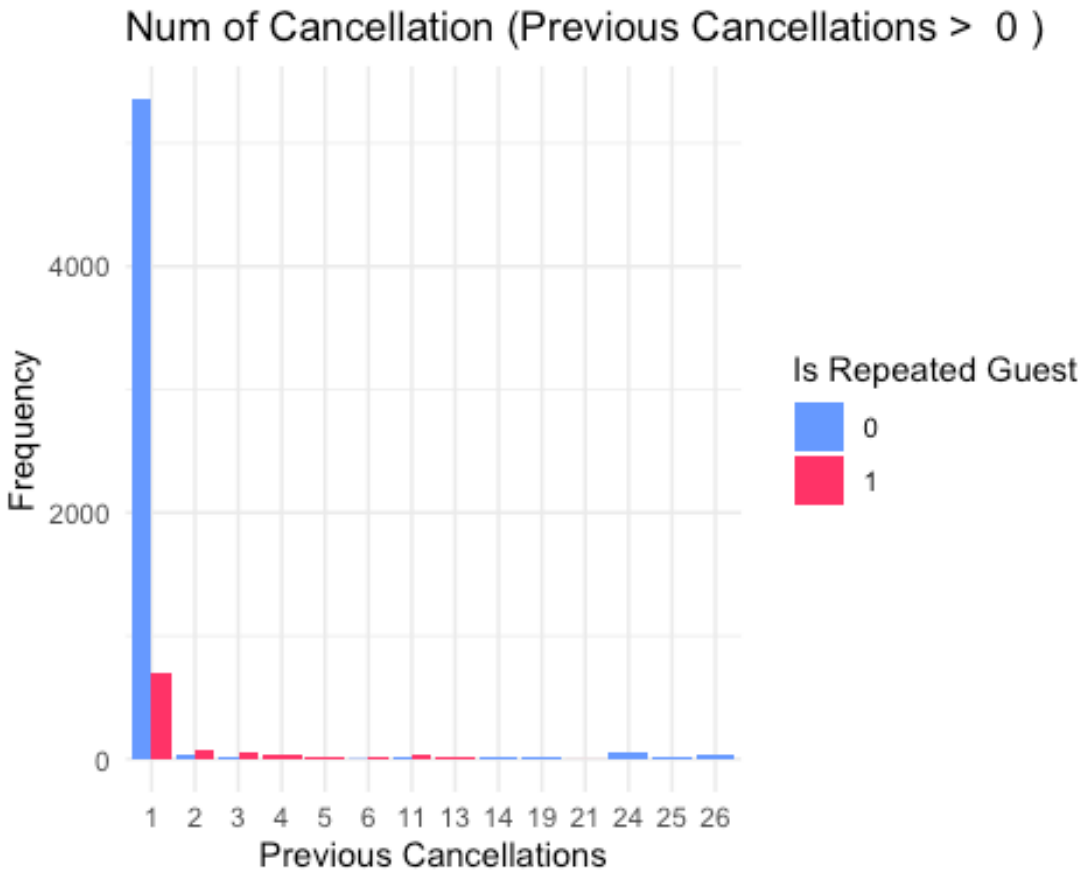
```r
  mutate(is_repeated_guest = as.logical(is_repeated_guest)) %>%
  filter(is_repeated_guest == T) %>%
  filter(previous_cancellations != 0) %>%
  select(previous_cancellations) %>%
  table()
```

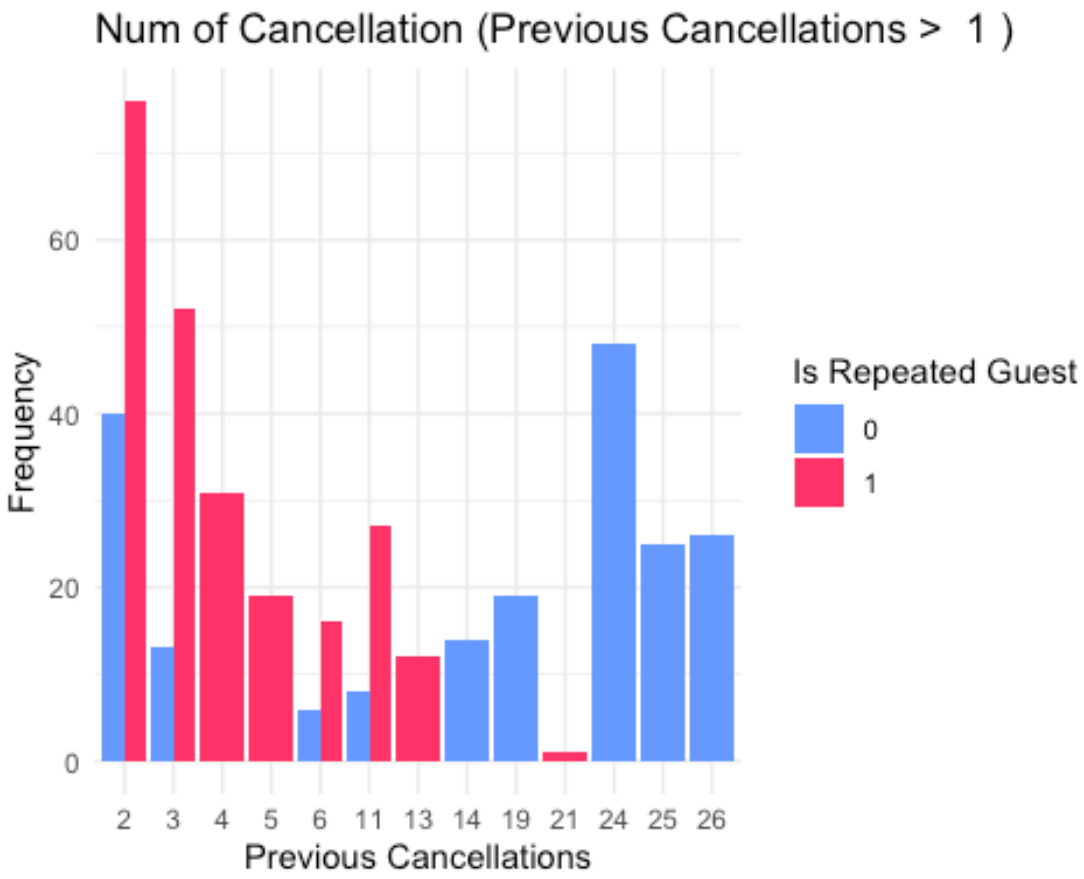
```
## .
##   1   2   3   4   5   6  11  13  21
## 693  76  52  31  19  16  27  12   1
```

```r
# ggplot 객체를 미리 정의합니다.
plot_list <- list()

for (i in 0:1){
 plot_list[[i + 1]] <- df %>%
  group_by(is_repeated_guest) %>%
  filter(previous_cancellations > i) %>%
  select(previous_cancellations) %>%
  count(previous_cancellations, name = "freq") %>%
  ungroup() %>%
  ggplot(aes(x = as.factor(previous_cancellations), y = freq, fill = as.factor(is_repeated_guest))) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Previous Cancellations", y = "Frequency", fill = "Is Repeated Guest") +
  scale_fill_manual(values = boolean_palette) +
  ggtitle(paste("Num of Cancellation (Previous Cancellations > ", i, ")")) +
  theme_minimal()
}
```

```
## Adding missing grouping variables: `is_repeated_guest`
## Adding missing grouping variables: `is_repeated_guest`
```

```r
for (i in 1:length(plot_list)) {
 print(plot_list[[i]])
}
```


Num of Cancellation (Previous Cancellations > 0 )
Frequency
4000
2000
0
1 2 3 4 5 6 11 13 14 19 21 24 25 26
Previous Cancellations
Is Repeated Guest
0
1




첫 방문 고객 중에 한 번 캔슬한 경우가 압도적으로 많았다. 시각화에 있어 1이 outlier의 역할을 하기에, 1을 제외한 그래프를 그렸다. 그 결과 고빈도 캔슬 수는 첫 방문인 경우가 많고, 저빈도 캔슬 수는 재방문인 경우가 많았다.

## 과거의 캔슬 여부가 현재 캔슬 여부와 관련 있을까?

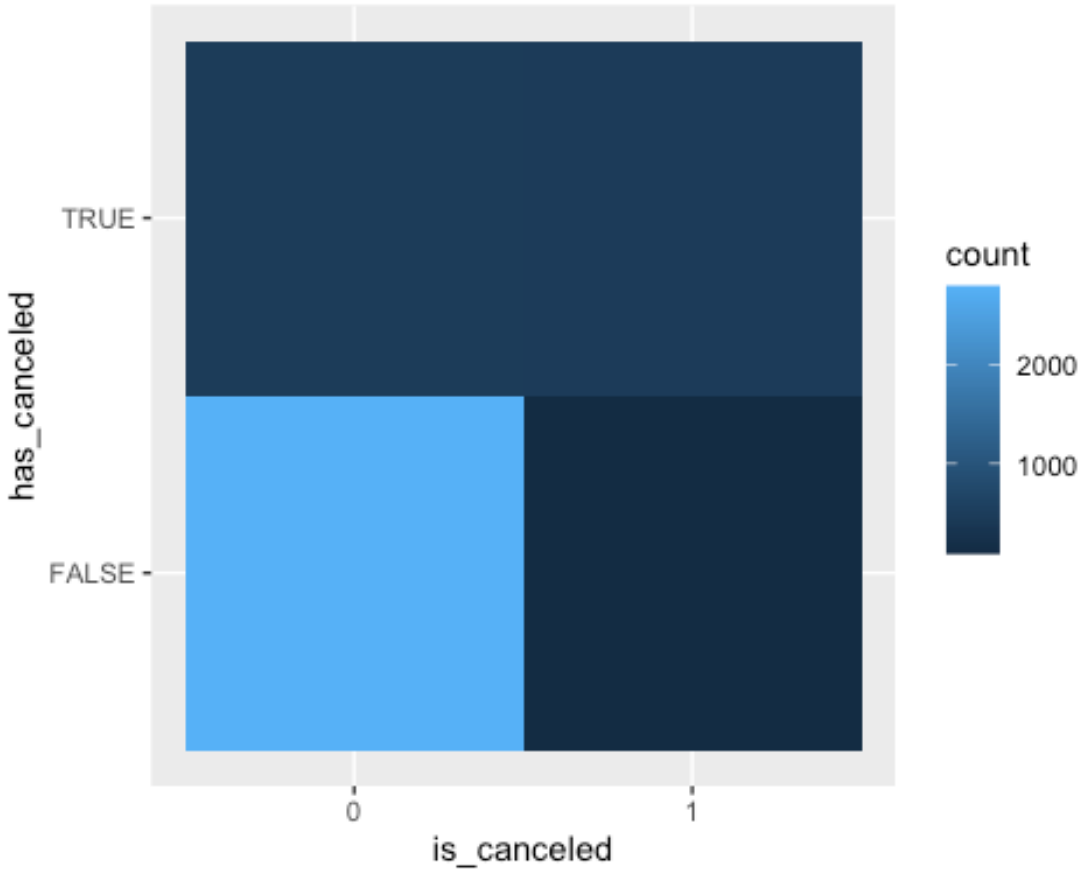
- 재방문 고객 기준 분석

```
# 재방문인 사람들만 뽑는다.
repeated_guest_df = df %>%
  mutate(is_repeated_guest = as.logical(is_repeated_guest)) %>%
  filter(is_repeated_guest)

head(repeated_guest_df)

##          hotel is_canceled lead_time arrival_date_year arrival_date_month
## 1 Resort Hotel           0         6              2016             April
## 2 Resort Hotel           1       202              2015              July
## 3 Resort Hotel           1       187              2015            August
```

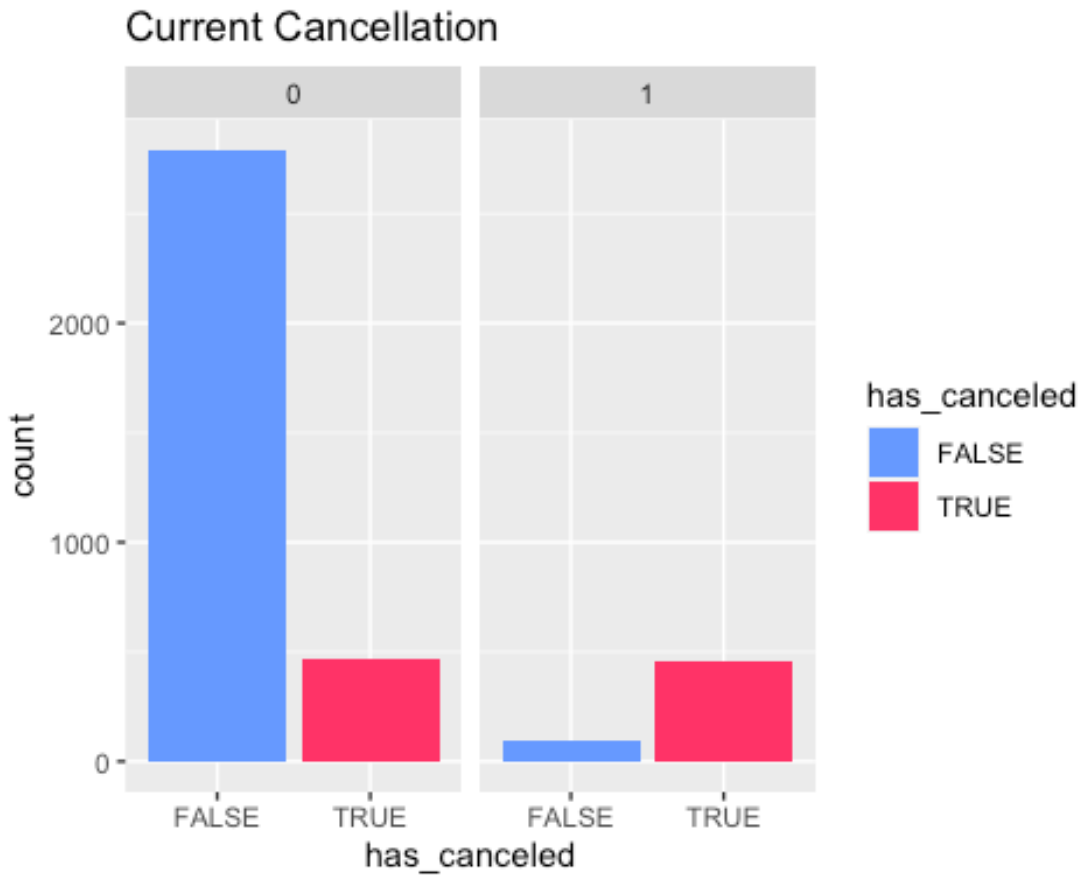
```
## 4 Resort Hotel          1       202             2015         September
## 5 Resort Hotel          1       173             2015            August
## 6 Resort Hotel          1       137             2015              July
##   arrival_date_week_number arrival_date_day_of_month stays_in_weekend_nights
## 1                       17                       20                       0
## 2                       30                       22                       2
## 3                       32                        4                       0
## 4                       36                        1                       2
## 5                       34                       16                       2
## 6                       29                       13                       2
##   stays_in_week_nights adults children babies meal country market_segment
## 1                    2      1        0      0   BB     PRT  Complementary
## 2                    5      2        0      0   BB     PRT  Offline TA/TO
## 3                    5      2        0      0   HB     PRT      Online TA
## 4                    8      2        0      0   BB     PRT  Offline TA/TO
## 5                    5      2        0      0   BB     PRT         Direct
## 6                    5      2        0      0   BB     PRT         Direct
##   distribution_channel is_repeated_guest previous_cancellations
## 1                TA/TO              TRUE                       0
## 2                TA/TO              TRUE                       1
## 3                TA/TO              TRUE                       1
## 4                TA/TO              TRUE                       1
## 5               Direct              TRUE                       1
## 6               Direct              TRUE                       1
##   previous_bookings_not_canceled reserved_room_type assigned_room_type
## 1                              1                  E                  E
## 2                              0                  A                  A
## 3                              0                  E                  E
## 4                              0                  A                  A
## 5                              0                  D                  D
## 6                              0                  A                  A
##   booking_changes deposit_type agent company days_in_waiting_list customer_type
## 1               0   No Deposit     5    NULL                    0     Transient
## 2               0   No Deposit   156    NULL                    0      Contract
## 3               0   No Deposit   240    NULL                    0     Transient
## 4               0   No Deposit   156    NULL                    0      Contract
## 5               0   No Deposit   250    NULL                    0     Transient
```

```
## 6              0   No Deposit   250    NULL                  0     Transient
##      adr required_car_parking_spaces total_of_special_requests
## 1   0.00                          1                         0
## 2  90.95                          0                         0
## 3  66.00                          0                         2
## 4  55.68                          0                         1
## 5 130.90                          0                         0
## 6   8.00                          0                         0
##   reservation_status reservation_status_date
## 1          Check-Out              2016-04-22
## 2           Canceled              2015-01-01
## 3           Canceled              2015-01-29
## 4           Canceled              2015-02-11
## 5           Canceled              2015-02-24
## 6           Canceled              2015-02-26
```

```
repeated_guest_df %>%
  mutate(has_canceled = (previous_cancellations > 0),
         is_canceled = as.factor(is_canceled)) %>%
  ggplot(aes(is_canceled, has_canceled)) + geom_bin2d(binwidth=1)
```

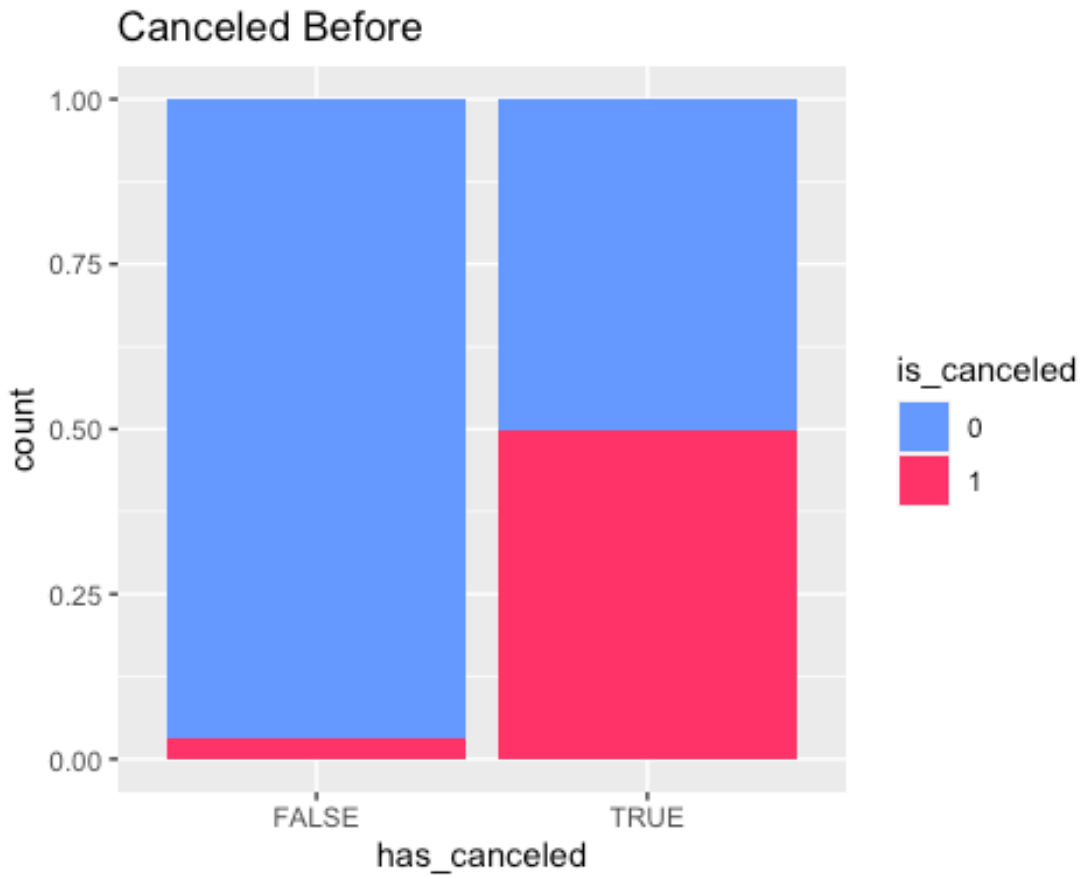



```
repeated_guest_df %>%
  mutate(has_canceled = (previous_cancellations != 0),
         is_canceled = as.factor(is_canceled)) %>%
  ggplot(aes(has_canceled, fill=has_canceled)) +
  geom_bar() +
  scale_fill_manual(values = boolean_palette) +
  facet_wrap(~is_canceled) +
  ggtitle("Current Cancellation")
```




이렇게만 봐서는 제대로 보이지 않는다.

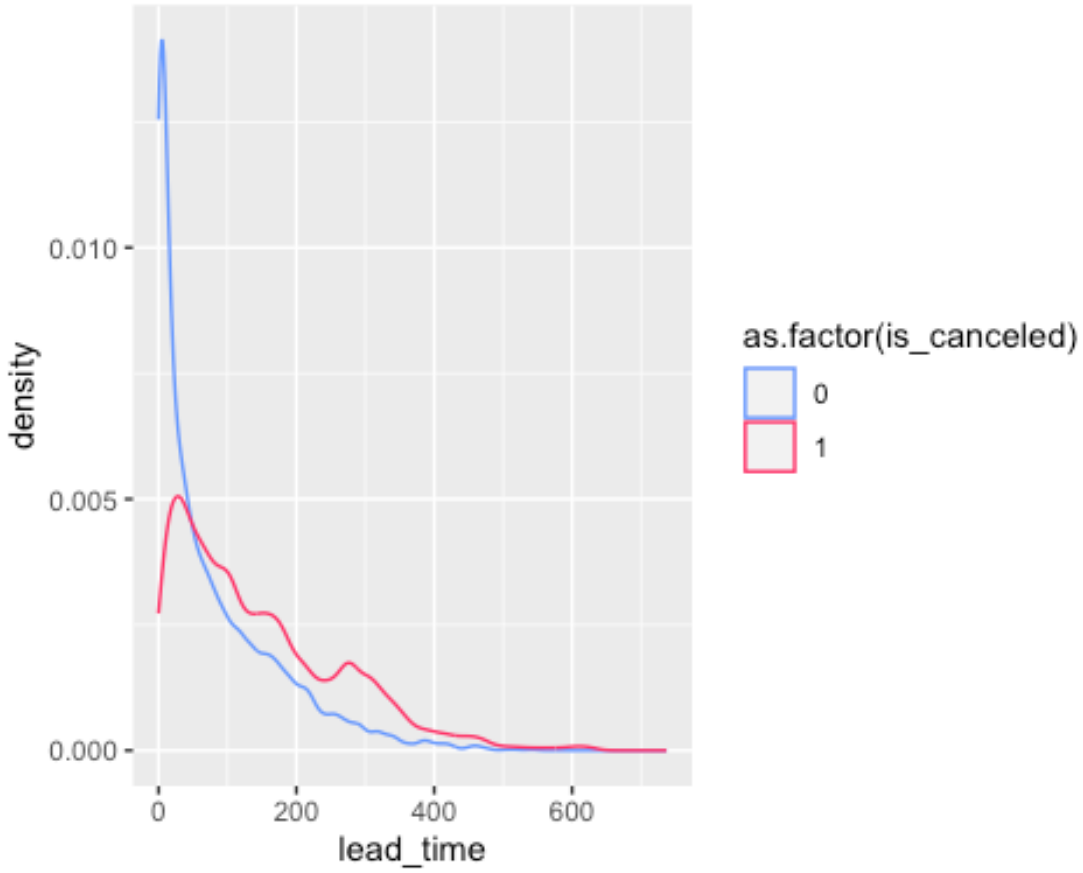
```
# 전체 데이터 중 과거 캔슬을 한 적 있는 사람의 비율
length(df$previous_cancellations[df$previous_cancellations != 0]) /
length(df$previous_cancellations)

## [1] 0.05430941

repeated_guest_df %>%
  mutate(has_canceled = (previous_cancellations != 0),
         is_canceled = as.factor(is_canceled)) %>%
  ggplot(aes(has_canceled, fill=is_canceled)) +
  geom_bar(position = "fill") +
  scale_fill_manual(values = boolean_palette) +
  ggtitle("Canceled Before")
```




과거에 부킹을 취소한 경우가 전체 데이터에 비해 크지 않아, 전체 빈도로 계산했을 때는 과거 캔슬한 사람이 현재도 캔슬을 한 경우가 명확히 보이지 않는다. 이를 보완하기 위해 비율로 데이터를 표준화시켜 시각화했다. 그 결과 과거에 캔슬을 한 사람이 또 캔슬을 하는 비율이 50%였다.

## 예약과 방문의 시간차와 취소의 관계

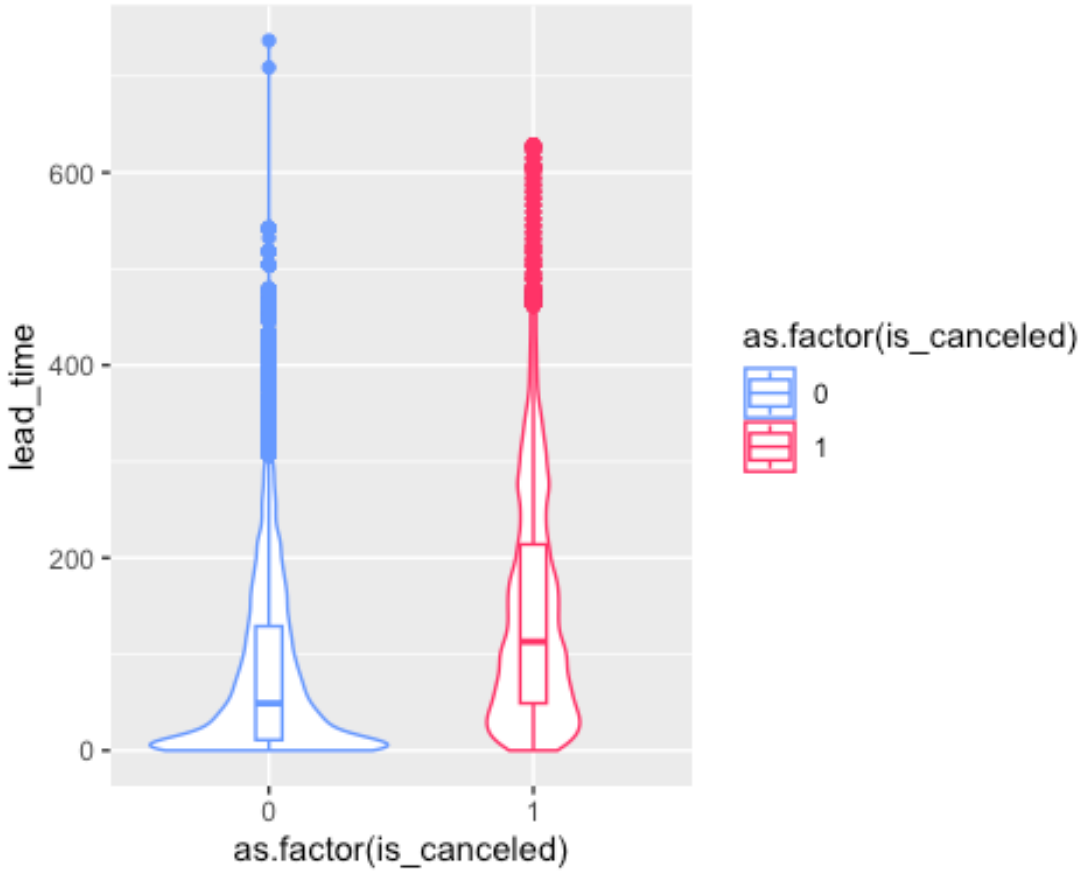
```
df %>%
  ggplot() +
  geom_density(aes(lead_time, color=as.factor(is_canceled))) +
  scale_color_manual(values = boolean_palette)
```




방문한 손님은 직전에 예약하는 경우가 많다. 대략 1달 전에 예약한 경우 취소하는 건수가 더 많다.

- 첫 번째 방문 손님

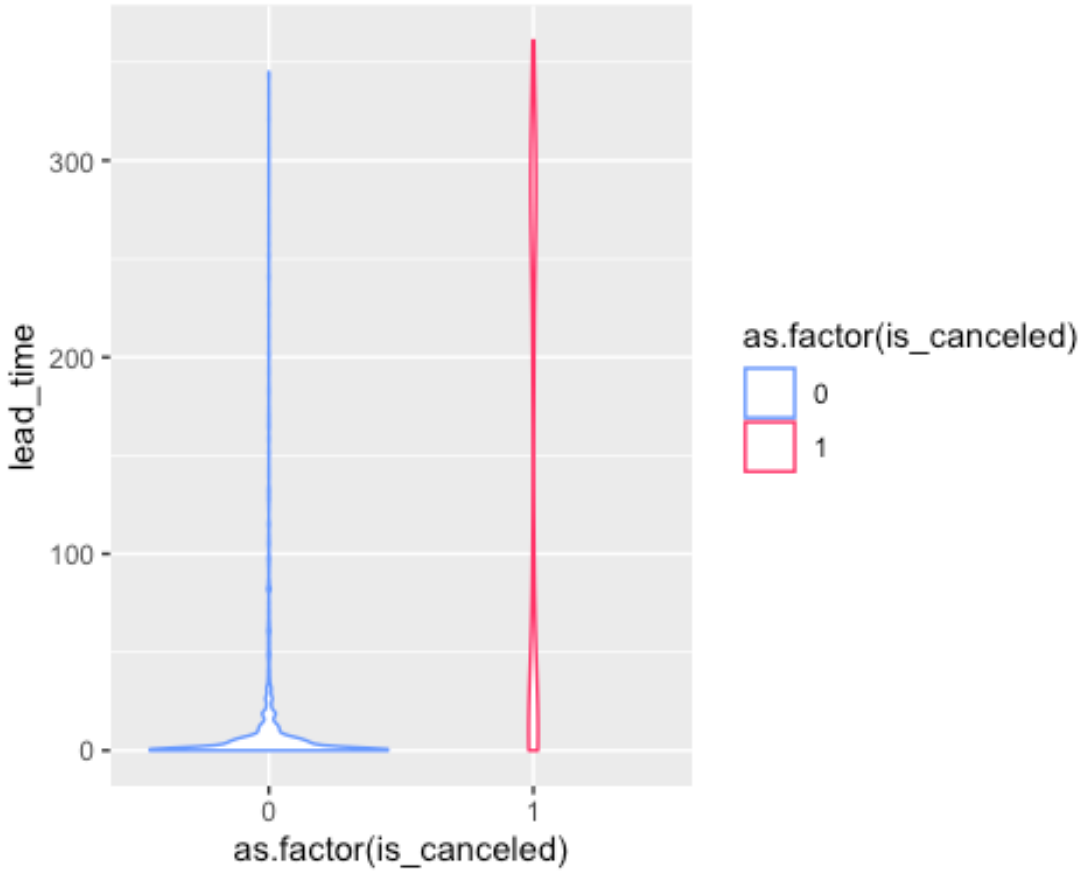
```
df %>%
  mutate(is_repeated_guest = as.logical(is_repeated_guest)) %>%
  filter(is_repeated_guest == F) %>%
  ggplot(aes(as.factor(is_canceled), lead_time, color=as.factor(is_canceled))) +
  geom_violin() +
  geom_boxplot(width=.1) +
   scale_color_manual(values = boolean_palette)
```




첫 손님인 경우 방문한 손님보다 취소한 손님이 더 방문과 예약 사이의 기간이 길다.

- 재방문 손님인 경우

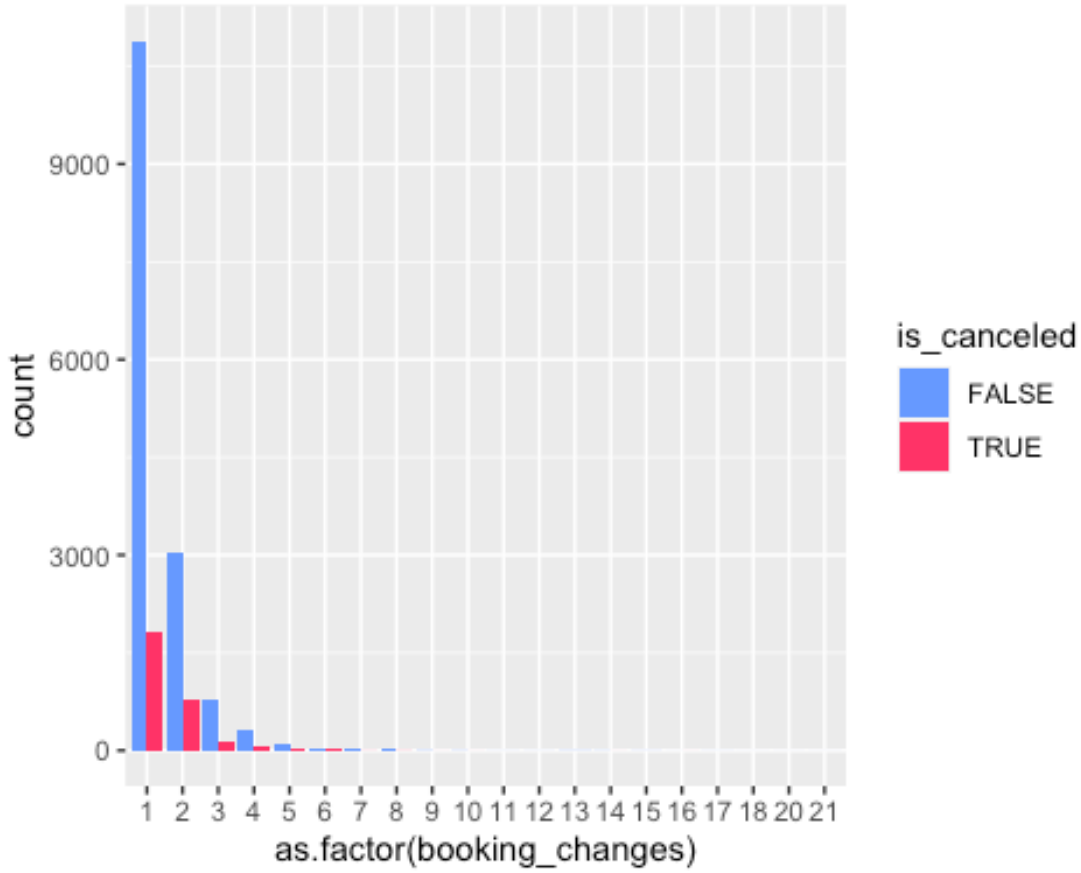
```
df %>%
  mutate(is_repeated_guest = as.logical(is_repeated_guest)) %>%
  filter(is_repeated_guest == T) %>%
  ggplot(aes(as.factor(is_canceled), lead_time, color=as.factor(is_canceled))) +
  geom_violin() +
  scale_color_manual(values = boolean_palette)
```




재방문의 경우도 첫방문과 마찬가지로 직전 예약이 가장 많다.

## 예약 변경과 취소

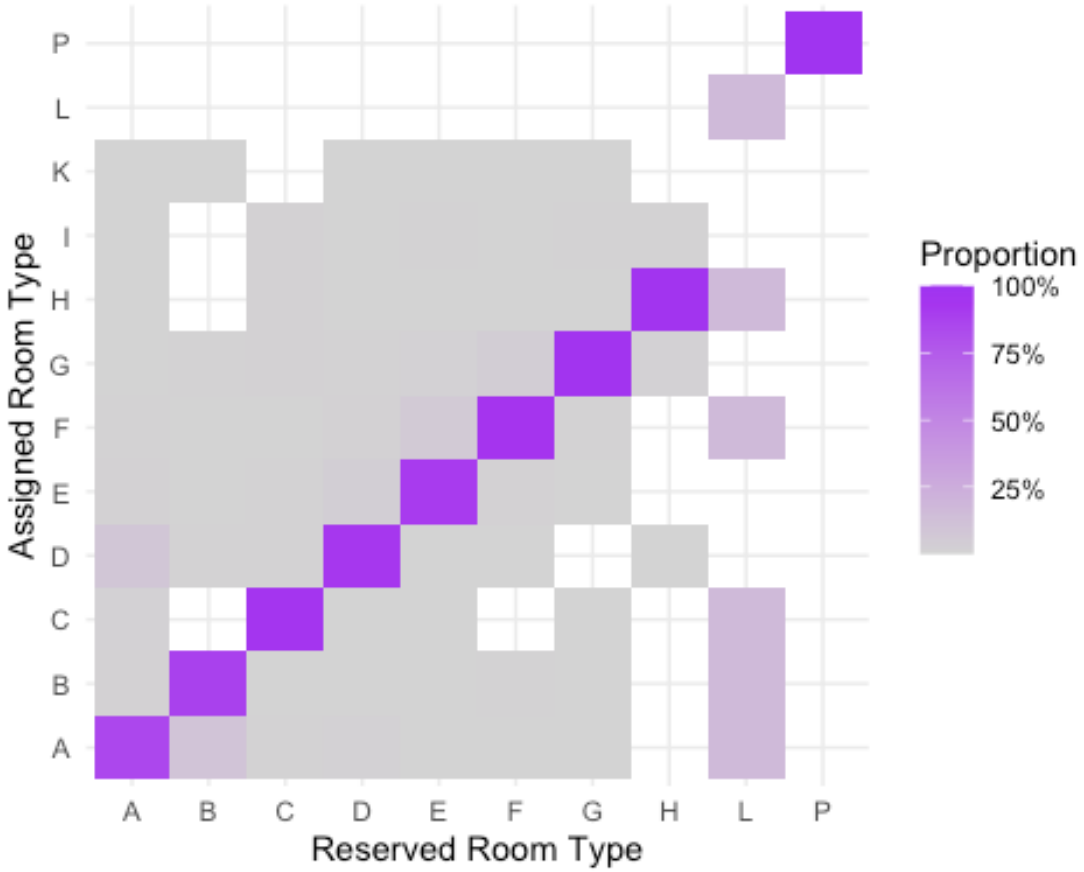
```
df %>%
  mutate(is_canceled = as.logical(is_canceled)) %>%
  filter(booking_changes != 0) %>%
  ggplot(aes(as.factor(booking_changes), fill=is_canceled)) +
  geom_bar( position = "dodge") +
  scale_fill_manual(values = boolean_palette)
```




대부분 5회 이내로 예약을 변경하며, 취소하지 않은 사람이 예약을 바꾼 경우가 더 많다.

## 예약한 방 - 받은 방

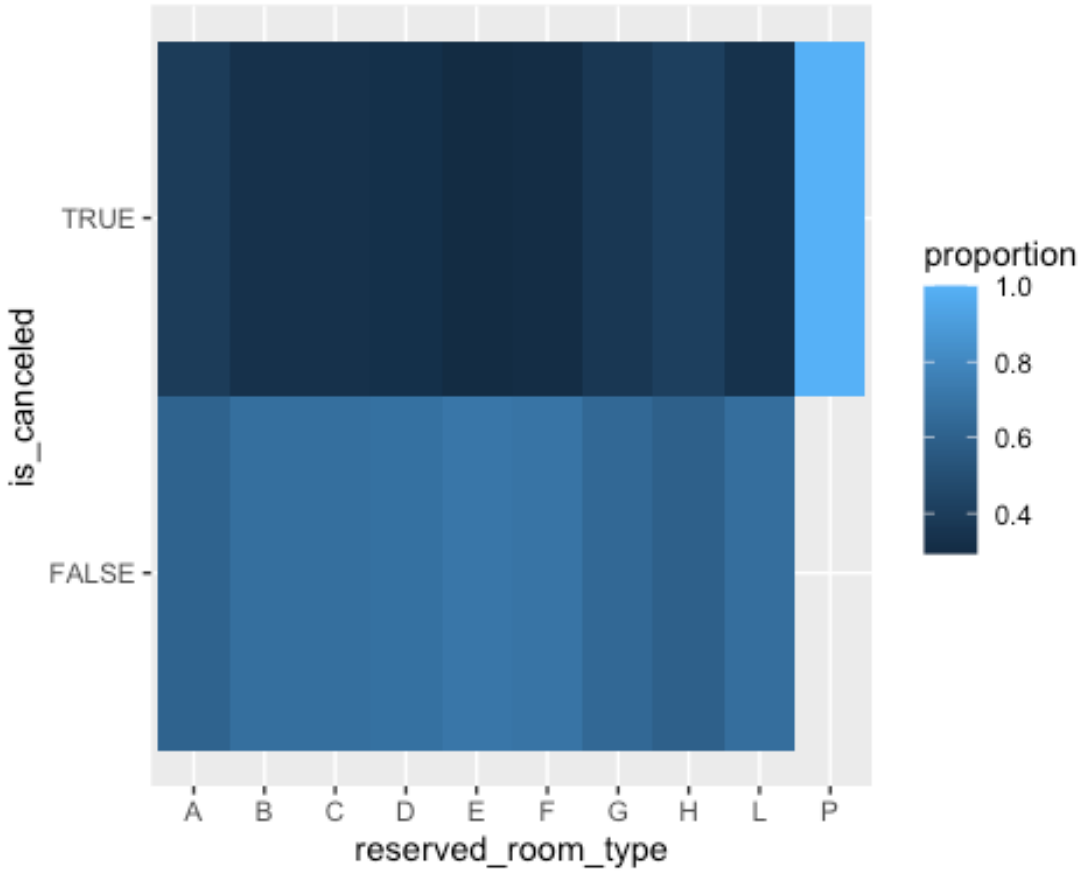
```
df %>%
  group_by(reserved_room_type) %>%
  count(assigned_room_type) %>%
  mutate(proportion = n / sum(n)) %>%
  ungroup() %>%
    ggplot(aes(x = reserved_room_type, y = assigned_room_type, fill = proportion)) +
  geom_tile() +
  scale_fill_continuous(name = "Proportion", labels = scales::percent, , low = "lightgrey", high =
"purple") +
  labs(x = "Reserved Room Type", y = "Assigned Room Type") +
  theme_minimal()
```




I,K 타입은 예약한 케이스는 없는데 배정받은 경우만 있으며, A-H 방까지만 I,K 타입 방을 배정받는다. 대부분 자기가 예약한 방을 갖는다. L 타입이 유독 예약한 것과 다른 방을 배정받는 경우가 많다.

## 예약한 방과 취소

```
df %>%
  mutate(is_canceled = as.logical(is_canceled)) %>%
  group_by(reserved_room_type) %>%
  count(is_canceled) %>%
  mutate(proportion = n / sum(n)) %>%
  ungroup() %>%
  ggplot() + geom_bin_2d(aes(reserved_room_type, is_canceled, fill=proportion))
```




P 타입 방을 예약한 사람이 전부 예약을 취소했다.