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)

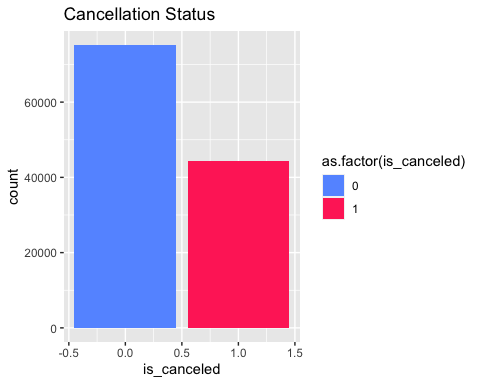
boolean\_palette = c("TRUE"="#FF3366",  
 "FALSE"="#6699FF",  
 "1"="#FF3366",  
 "0"="#6699FF")

df = read.csv("/Users/jimin/Desktop/지민/ewha/2024-1/다변량분석/통통배/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

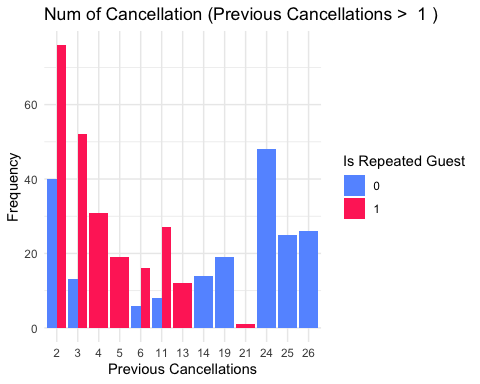
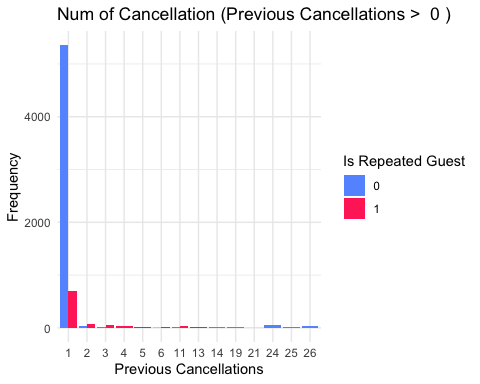
# 방문 경험이 있는 사람이 과거에 캔슬한 횟수   
df %>%   
 mutate(is\_repeated\_guest = as.logical(is\_repeated\_guest)) %>%  
 filter(is\_repeated\_guest == T) %>%  
 filter(previous\_cancellations != 0) %>%   
 select(previous\_cancellations) %>%  
 table()

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

# 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`

for (i in 1:length(plot\_list)) {  
 print(plot\_list[[i]])  
}



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.

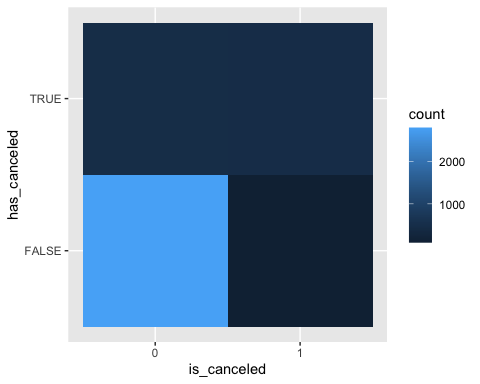
## 과거의 캔슬 여부가 현재 캔슬 여부와 관련 있을까?

* 재방문 고객 기준 분석

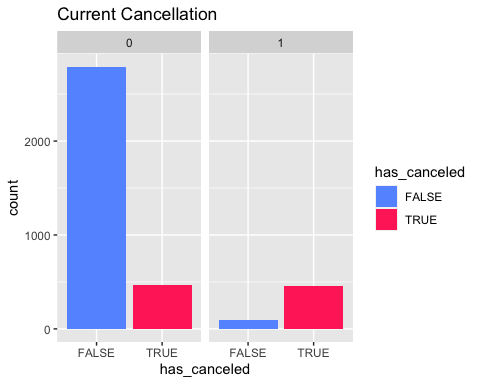
# 재방문인 사람들만 뽑는다.   
repeated\_guest\_df = df %>%   
 mutate(is\_repeated\_guest = as.logical(is\_repeated\_guest)) %>%  
 filter(is\_repeated\_guest)   
  
head(repeated\_guest\_df)

## 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)



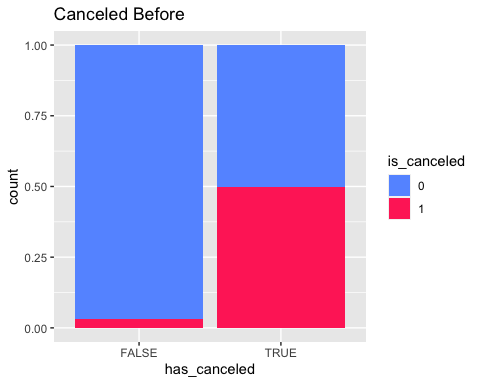
repeated\_guest\_df %>%   
 mutate(has\_canceled = (previous\_cancellations != 0),  
 is\_canceled = as.factor(is\_canceled)) %>%  
 ggplot(aes(has\_canceled, fill=has\_canceled)) +   
 geom\_bar() +   
 scale\_fill\_manual(values = boolean\_palette) +   
 facet\_wrap(~is\_canceled) +   
 ggtitle("Current 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

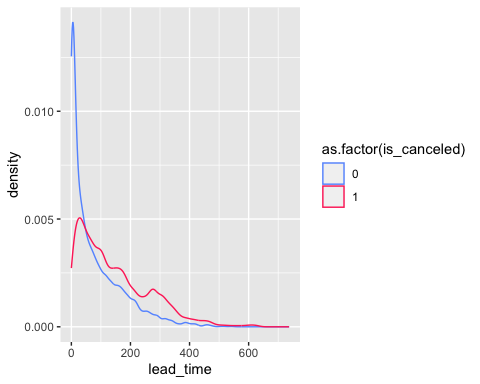
repeated\_guest\_df %>%   
 mutate(has\_canceled = (previous\_cancellations != 0),  
 is\_canceled = as.factor(is\_canceled)) %>%  
 ggplot(aes(has\_canceled, fill=is\_canceled)) +   
 geom\_bar(position = "fill") +   
 scale\_fill\_manual(values = boolean\_palette) +  
 ggtitle("Canceled Before")



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

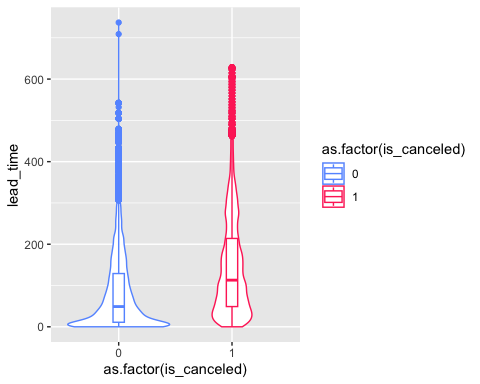
df %>%   
 ggplot() +   
 geom\_density(aes(lead\_time, color=as.factor(is\_canceled))) +   
 scale\_color\_manual(values = boolean\_palette)



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

* 첫 번째 방문 손님

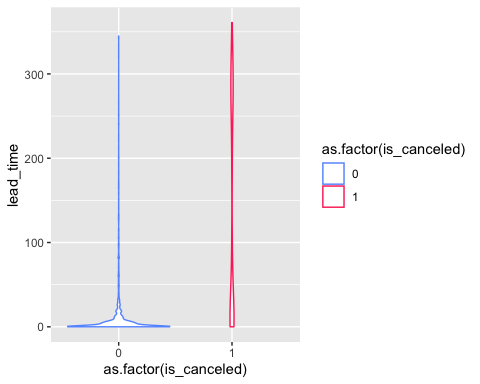
df %>%   
 mutate(is\_repeated\_guest = as.logical(is\_repeated\_guest)) %>%  
 filter(is\_repeated\_guest == F) %>%  
 ggplot(aes(as.factor(is\_canceled), lead\_time, color=as.factor(is\_canceled))) +  
 geom\_violin() +  
 geom\_boxplot(width=.1) +   
 scale\_color\_manual(values = boolean\_palette)



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

* 재방문 손님인 경우

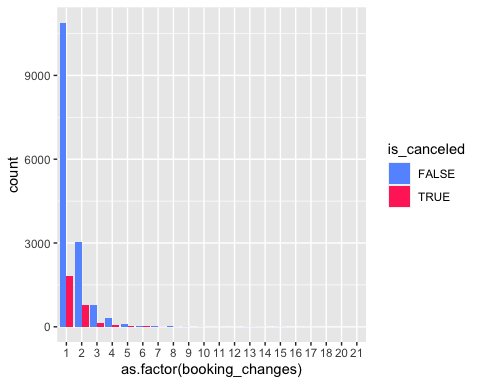
df %>%   
 mutate(is\_repeated\_guest = as.logical(is\_repeated\_guest)) %>%  
 filter(is\_repeated\_guest == T) %>%  
 ggplot(aes(as.factor(is\_canceled), lead\_time, color=as.factor(is\_canceled))) +  
 geom\_violin() +  
 scale\_color\_manual(values = boolean\_palette)



재방문의 경우도 첫방문과 마찬가지로 직전 예약이 가장 많다.

## 예약 변경과 취소

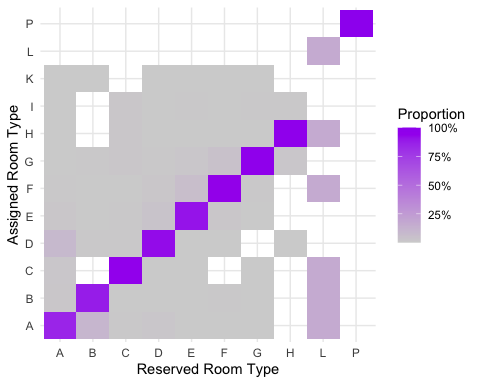
df %>%   
 mutate(is\_canceled = as.logical(is\_canceled)) %>%  
 filter(booking\_changes != 0) %>%  
 ggplot(aes(as.factor(booking\_changes), fill=is\_canceled)) +   
 geom\_bar( position = "dodge") +   
 scale\_fill\_manual(values = boolean\_palette)



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

## 예약한 방 - 받은 방

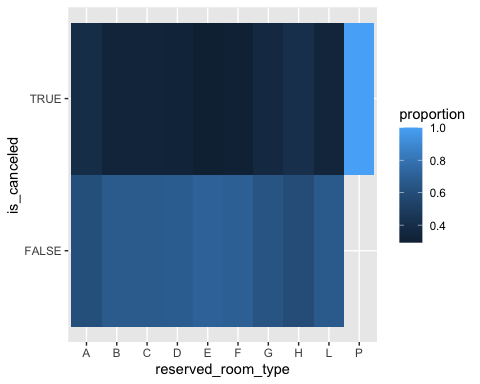
df %>%  
 group\_by(reserved\_room\_type) %>%  
 count(assigned\_room\_type) %>%  
 mutate(proportion = n / sum(n)) %>%  
 ungroup() %>%  
 ggplot(aes(x = reserved\_room\_type, y = assigned\_room\_type, fill = proportion)) +  
 geom\_tile() +  
 scale\_fill\_continuous(name = "Proportion", labels = scales::percent, , 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.