R Notebook

전처리

```
library(tidyverse)

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


dim(df)
```

```
## [1] 119390     32
```

중복 데이터 삭제

```
nrow(df[duplicated(df), ]) # 중복 열의 개수
```

```
## [1] 31994
```

```
df = df %>% distinct() # 중복 데이터 삭제
dim(df)
```

```
## [1] 87396    32
```

어른이 없는 데이터 삭제

```
df = df %>% filter(adults != 0)
dim(df)
```

```
## [1] 87011    32
```

결측치 처리

NA case

```
# 결측치 확인
df %>% filter(rowSums(is.na(df)) == 1)
```

```
##        hotel is_canceled lead_time arrival_date_year arrival_date_month
## 1 City Hotel           1         2              2015              August
## 2 City Hotel           1         1              2015              August
## 3 City Hotel           1         1              2015              August
## 4 City Hotel           1         8              2015              August
##   arrival_date_week_number arrival_date_day_of_month
stays_in_weekend_nights
## 1                       32                         3
1
## 2                       32                         5
```

```
0
## 3                       32                        5
0
## 4                       33                       13
2
##   stays_in_week_nights adults children babies meal country market_segment
## 1                    0      2       NA      0   BB     PRT      Undefined
## 2                    2      2       NA      0   BB     PRT         Direct
## 3                    2      3       NA      0   BB     PRT      Undefined
## 4                    5      2       NA      0   BB     PRT      Online TA
##   distribution_channel is_repeated_guest previous_cancellations
## 1            Undefined                 0                      0
## 2            Undefined                 0                      0
## 3            Undefined                 0                      0
## 4            Undefined                 0                      0
##   previous_bookings_not_canceled reserved_room_type assigned_room_type
## 1                              0                  B                  B
## 2                              0                  B                  B
## 3                              0                  B                  B
## 4                              0                  B                  B
##   booking_changes deposit_type agent company days_in_waiting_list
## 1               0   No Deposit  NULL    NULL                    0
## 2               0   No Deposit    14    NULL                    0
## 3               0   No Deposit  NULL    NULL                    0
## 4               0   No Deposit     9    NULL                    0
##     customer_type  adr required_car_parking_spaces
total_of_special_requests
## 1 Transient-Party 12.0                           0
1
## 2 Transient-Party 12.0                           0
1
## 3 Transient-Party 18.0                           0
2
## 4 Transient-Party 76.5                           0
1
##   reservation_status reservation_status_date
## 1           Canceled              2015-08-01
## 2           Canceled              2015-08-04
## 3           Canceled              2015-08-04
## 4           Canceled              2015-08-09

df = df %>% filter(rowSums(is.na(df)) == 0)
dim(df)

## [1] 87007    32
```

NULL case

```
df %>% filter(agent == "NULL") %>% dim()

## [1] 12122    32
```

```
# 그냥 날리기엔 NULL 값이 너무 많다.
# agent가 중요한 정보가 아니라서 agent를 날려버리는 쪽으로 선택
```

한 방에 처리하는 코드(팀원 배포용)

```
df = df %>%
        filter(rowSums(is.na(df)) == 0) %>% # 결측치 삭제(NA)
        distinct() %>% # 중복 데이터 삭제
        filter(adults != 0) %>% # 어른이 없는 데이터 삭제
        select(-agent,-company) %>% # 사용하지 않을 agent, company 삭제
        mutate(reservation_status_date = as.Date(reservation_status_date)) #
reservation_status_date dtype 변환

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         0              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     PRT         Direct
##   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               Direct                 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                  C                  C
##   booking_changes deposit_type days_in_waiting_list customer_type adr
## 1               3   No Deposit                    0     Transient   0
## 2               4   No Deposit                    0     Transient   0
## 3               0   No Deposit                    0     Transient  75
## 4               0   No Deposit                    0     Transient  75
## 5               0   No Deposit                    0     Transient  98
## 6               0   No Deposit                    0     Transient 107
##   required_car_parking_spaces total_of_special_requests reservation_status
## 1                           0                         0          Check-Out
## 2                           0                         0          Check-Out
## 3                           0                         0          Check-Out
## 4                           0                         0          Check-Out
## 5                           0                         1          Check-Out
## 6                           0                         0          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
```

```
dim(df)
```

```
## [1] 87007    30
```