MV\_teamprj1\_EDA

1973046 윤여름

2024-05-21

# data importing & analysis

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.1 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
## ✔ 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

library(ggplot2)  
library(ggmosaic)  
library(GGally)

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2  
##   
## Attaching package: 'GGally'  
##   
## The following object is masked from 'package:ggmosaic':  
##   
## happy

hotel = read.csv("./source/hotel\_bookings.csv")  
head(hotel)

## 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

dim(hotel)

## [1] 119390 32

summary(hotel)

## hotel is\_canceled lead\_time arrival\_date\_year  
## Length:119390 Min. :0.0000 Min. : 0 Min. :2015   
## Class :character 1st Qu.:0.0000 1st Qu.: 18 1st Qu.:2016   
## Mode :character Median :0.0000 Median : 69 Median :2016   
## Mean :0.3704 Mean :104 Mean :2016   
## 3rd Qu.:1.0000 3rd Qu.:160 3rd Qu.:2017   
## Max. :1.0000 Max. :737 Max. :2017   
##   
## arrival\_date\_month arrival\_date\_week\_number arrival\_date\_day\_of\_month  
## Length:119390 Min. : 1.00 Min. : 1.0   
## Class :character 1st Qu.:16.00 1st Qu.: 8.0   
## Mode :character Median :28.00 Median :16.0   
## Mean :27.17 Mean :15.8   
## 3rd Qu.:38.00 3rd Qu.:23.0   
## Max. :53.00 Max. :31.0   
##   
## stays\_in\_weekend\_nights stays\_in\_week\_nights adults   
## Min. : 0.0000 Min. : 0.0 Min. : 0.000   
## 1st Qu.: 0.0000 1st Qu.: 1.0 1st Qu.: 2.000   
## Median : 1.0000 Median : 2.0 Median : 2.000   
## Mean : 0.9276 Mean : 2.5 Mean : 1.856   
## 3rd Qu.: 2.0000 3rd Qu.: 3.0 3rd Qu.: 2.000   
## Max. :19.0000 Max. :50.0 Max. :55.000   
##   
## children babies meal country   
## Min. : 0.0000 Min. : 0.000000 Length:119390 Length:119390   
## 1st Qu.: 0.0000 1st Qu.: 0.000000 Class :character Class :character   
## Median : 0.0000 Median : 0.000000 Mode :character Mode :character   
## Mean : 0.1039 Mean : 0.007949   
## 3rd Qu.: 0.0000 3rd Qu.: 0.000000   
## Max. :10.0000 Max. :10.000000   
## NA's :4   
## market\_segment distribution\_channel is\_repeated\_guest  
## Length:119390 Length:119390 Min. :0.00000   
## Class :character Class :character 1st Qu.:0.00000   
## Mode :character Mode :character Median :0.00000   
## Mean :0.03191   
## 3rd Qu.:0.00000   
## Max. :1.00000   
##   
## previous\_cancellations previous\_bookings\_not\_canceled reserved\_room\_type  
## Min. : 0.00000 Min. : 0.0000 Length:119390   
## 1st Qu.: 0.00000 1st Qu.: 0.0000 Class :character   
## Median : 0.00000 Median : 0.0000 Mode :character   
## Mean : 0.08712 Mean : 0.1371   
## 3rd Qu.: 0.00000 3rd Qu.: 0.0000   
## Max. :26.00000 Max. :72.0000   
##   
## assigned\_room\_type booking\_changes deposit\_type agent   
## Length:119390 Min. : 0.0000 Length:119390 Length:119390   
## Class :character 1st Qu.: 0.0000 Class :character Class :character   
## Mode :character Median : 0.0000 Mode :character Mode :character   
## Mean : 0.2211   
## 3rd Qu.: 0.0000   
## Max. :21.0000   
##   
## company days\_in\_waiting\_list customer\_type adr   
## Length:119390 Min. : 0.000 Length:119390 Min. : -6.38   
## Class :character 1st Qu.: 0.000 Class :character 1st Qu.: 69.29   
## Mode :character Median : 0.000 Mode :character Median : 94.58   
## Mean : 2.321 Mean : 101.83   
## 3rd Qu.: 0.000 3rd Qu.: 126.00   
## Max. :391.000 Max. :5400.00   
##   
## required\_car\_parking\_spaces total\_of\_special\_requests reservation\_status  
## Min. :0.00000 Min. :0.0000 Length:119390   
## 1st Qu.:0.00000 1st Qu.:0.0000 Class :character   
## Median :0.00000 Median :0.0000 Mode :character   
## Mean :0.06252 Mean :0.5714   
## 3rd Qu.:0.00000 3rd Qu.:1.0000   
## Max. :8.00000 Max. :5.0000   
##   
## reservation\_status\_date  
## Length:119390   
## Class :character   
## Mode :character   
##   
##   
##   
##

str(hotel)

## 'data.frame': 119390 obs. of 32 variables:  
## $ hotel : chr "Resort Hotel" "Resort Hotel" "Resort Hotel" "Resort Hotel" ...  
## $ is\_canceled : int 0 0 0 0 0 0 0 0 1 1 ...  
## $ lead\_time : int 342 737 7 13 14 14 0 9 85 75 ...  
## $ arrival\_date\_year : int 2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...  
## $ arrival\_date\_month : chr "July" "July" "July" "July" ...  
## $ arrival\_date\_week\_number : int 27 27 27 27 27 27 27 27 27 27 ...  
## $ arrival\_date\_day\_of\_month : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ stays\_in\_weekend\_nights : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ stays\_in\_week\_nights : int 0 0 1 1 2 2 2 2 3 3 ...  
## $ adults : int 2 2 1 1 2 2 2 2 2 2 ...  
## $ children : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ babies : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ meal : chr "BB" "BB" "BB" "BB" ...  
## $ country : chr "PRT" "PRT" "GBR" "GBR" ...  
## $ market\_segment : chr "Direct" "Direct" "Direct" "Corporate" ...  
## $ distribution\_channel : chr "Direct" "Direct" "Direct" "Corporate" ...  
## $ is\_repeated\_guest : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ previous\_cancellations : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ previous\_bookings\_not\_canceled: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ reserved\_room\_type : chr "C" "C" "A" "A" ...  
## $ assigned\_room\_type : chr "C" "C" "C" "A" ...  
## $ booking\_changes : int 3 4 0 0 0 0 0 0 0 0 ...  
## $ deposit\_type : chr "No Deposit" "No Deposit" "No Deposit" "No Deposit" ...  
## $ agent : chr "NULL" "NULL" "NULL" "304" ...  
## $ company : chr "NULL" "NULL" "NULL" "NULL" ...  
## $ days\_in\_waiting\_list : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ customer\_type : chr "Transient" "Transient" "Transient" "Transient" ...  
## $ adr : num 0 0 75 75 98 ...  
## $ required\_car\_parking\_spaces : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ total\_of\_special\_requests : int 0 0 0 0 1 1 0 1 1 0 ...  
## $ reservation\_status : chr "Check-Out" "Check-Out" "Check-Out" "Check-Out" ...  
## $ reservation\_status\_date : chr "2015-07-01" "2015-07-01" "2015-07-02" "2015-07-02" ...

# 소주제 1 : baby의 수/children의 수와 연관된 변수 파악

* meal type (식사타입)
* total\_of\_special\_requests (특별 요청 수)
* 가설 1 : baby의 유무가 meal type에 영향을 미칠 것이다.
* 가설 2 : baby의 유무가 total\_of\_special\_requests(특별 요청 수)에 영향을 미칠 것이다.
* 가설 3: baby의 유무가 meal type과 total\_of\_special\_requests(특별 요청 수)에 영향을 미칠 것이다. (one-way manova)

#babies  
summary(hotel$babies)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000000 0.000000 0.000000 0.007949 0.000000 10.000000

sum(is.na(hotel$babies))

## [1] 0

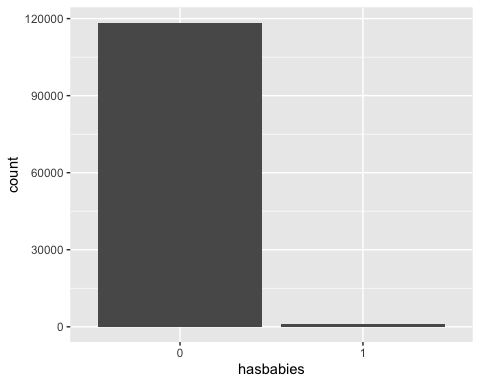
sum(ifelse(hotel$babies != 0, 1, 0))

## [1] 917

table(hotel$babies)

##   
## 0 1 2 9 10   
## 118473 900 15 1 1

#baby 유무로 전처리  
hotel$hasbabies = ifelse(hotel$babies != "0", 1, 0)  
hotel$hasbabies = factor(hotel$hasbabies, levels = c(0, 1), labels = c("0", "1"))  
ggplot(hotel, aes(hasbabies)) + geom\_bar()



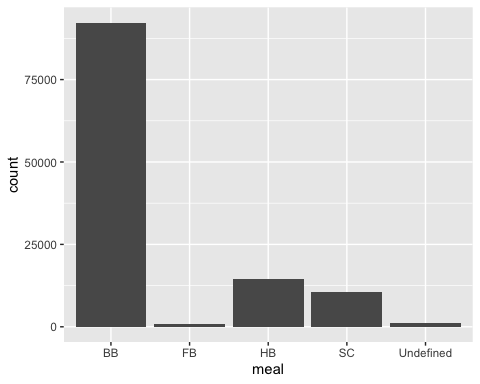
#meal type  
sum(is.na(hotel$meal))

## [1] 0

table(hotel$meal) #BB : breakfast&Bed / FB : full-board(breakfast, lunch & dinner) / HB : Half Board (Breakfast and Dinner normally) / SC : self-catering / undefined = sc로 통합하거나 삭제

##   
## BB FB HB SC Undefined   
## 92310 798 14463 10650 1169

ggplot(hotel, aes(meal)) + geom\_bar()



#total\_of\_special\_requests  
summary(hotel$total\_of\_special\_requests)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.0000 0.0000 0.5714 1.0000 5.0000

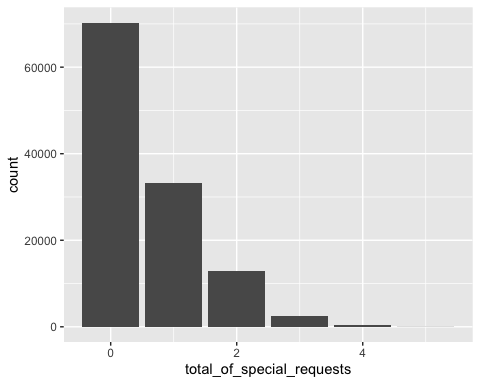
sum(is.na(hotel$total\_of\_special\_requests))

## [1] 0

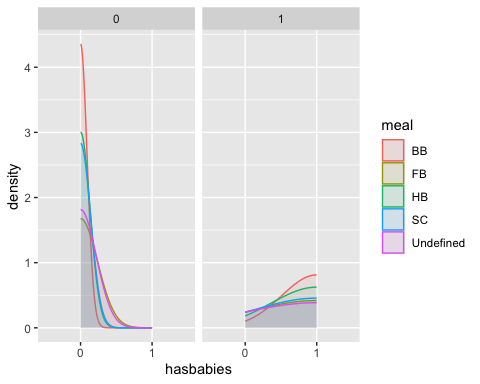
table(hotel$total\_of\_special\_requests)

##   
## 0 1 2 3 4 5   
## 70318 33226 12969 2497 340 40

ggplot(hotel, aes(total\_of\_special\_requests)) + geom\_bar()

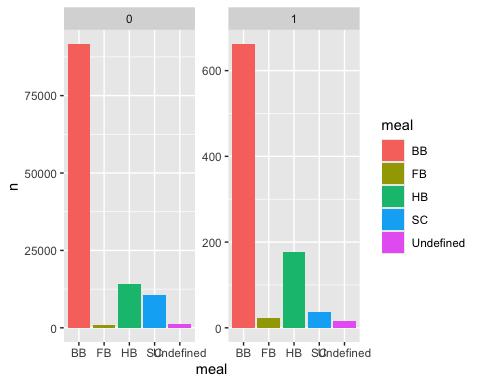


#hyp 1 : babies - meal type   
ggplot(hotel, aes(x = hasbabies, colour = meal, fill = meal)) + geom\_density(alpha = 0.1) + facet\_wrap(~hasbabies)

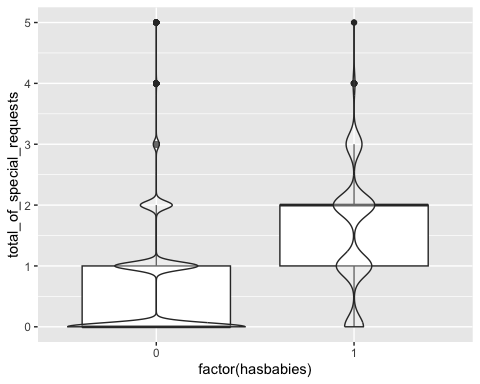


hotel %>% group\_by(meal, hasbabies) %>% summarise(n=n()) %>% ggplot(aes(meal,n,fill=meal)) + geom\_bar(stat = 'identity', position='dodge') + facet\_wrap(~hasbabies, scales = "free\_y")

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



#hyp 2 : babies - total\_of\_special\_requests  
ggplot(hotel, aes(x = factor(hasbabies), y = total\_of\_special\_requests)) + geom\_boxplot() + geom\_violin(alpha = 0.3)



#hyp 3 : babies - meal type & total\_of\_special\_requests  
  
ggpairs(hotel[,c(13,30,33)])

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

