hw01.비데마

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(corrplot)
```

```
## corrplot 0.91 loaded
```

```
library(skimr)
```

```
## Warning: 패키지 'skimr'는 R 버전 4.1.3에서 작성되었습니다
```

```
library(naniar)
```

```
## Warning: 패키지 'naniar'는 R 버전 4.1.3에서 작성되었습니다
```

```
##
## 다음의 패키지를 부착합니다: 'naniar'
```

```
## The following object is masked from 'package:skimr':
##
##     n_complete
```

```
TIPS <- read.csv("C:₩₩Users₩₩User₩₩Desktop₩₩data₩₩tips.csv",sep =',')
dim(TIPS)
```

```
## [1] 244   7
```

```
str(TIPS) #변수 측도확인
```

```
## 'data.frame':    244 obs. of  7 variables:
##  $ total_bill: num  17 10.3 21 23.7 24.6 ...
##  $ tip       : num  1.01 1.66 3.5 3.31 3.61 4.71 2 3.12 1.96 3.23 ...
##  $ sex       : chr  "Female" "Male" "Male" "Male" ...
##  $ smoker    : chr  "No" "No" "No" "No" ...
##  $ day       : chr  "Sun" "Sun" "Sun" "Sun" ...
##  $ time      : chr  "Dinner" "Dinner" "Dinner" "Dinner" ...
##  $ size      : int  2 3 3 2 4 4 2 4 2 2 ...
```

```
TIPS <- mutate(TIPS,sex=factor(sex),smoker=factor(smoker),day=factor(day),time=factor(time))
TIPS$tiprate= TIPS$tip/TIPS$total_bill
str(TIPS)
```

```
## 'data.frame':    244 obs. of  8 variables:
##  $ total_bill: num  17 10.3 21 23.7 24.6 ...
##  $ tip       : num  1.01 1.66 3.5 3.31 3.61 4.71 2 3.12 1.96 3.23 ...
##  $ sex       : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 2 2 2 2 2 ...
##  $ smoker    : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ day       : Factor w/ 4 levels "Fri","Sat","Sun",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ time      : Factor w/ 2 levels "Dinner","Lunch": 1 1 1 1 1 1 1 1 1 1 ...
##  $ size      : int  2 3 3 2 4 4 2 4 2 2 ...
##  $ tiprate   : num  0.0594 0.1605 0.1666 0.1398 0.1468 ...
```

```
head(TIPS)
```

```
##   total_bill  tip    sex smoker day   time size    tiprate
## 1      16.99 1.01 Female     No Sun Dinner    2 0.05944673
## 2      10.34 1.66   Male     No Sun Dinner    3 0.16054159
## 3      21.01 3.50   Male     No Sun Dinner    3 0.16658734
## 4      23.68 3.31   Male     No Sun Dinner    2 0.13978041
## 5      24.59 3.61 Female     No Sun Dinner    4 0.14680765
## 6      25.29 4.71   Male     No Sun Dinner    4 0.18623962
```

```
sapply(TIPS,function(x)sum(is.na(x))) #결측값 확인
```

```
## total_bill        tip        sex     smoker        day       time       size
##          0          0          0          0          0          0          0
##    tiprate
##          0
```

```
#tips 요약통계량
summary(TIPS)
```

```
##    total_bill         tip             sex      smoker      day         time
##  Min.   : 3.07   Min.   : 1.000   Female: 87   No :151   Fri :19   Dinner:176
##  1st Qu.:13.35   1st Qu.: 2.000   Male  :157   Yes: 93   Sat :87   Lunch : 68
##  Median :17.80   Median : 2.900                          Sun :76
##  Mean   :19.79   Mean   : 2.998                          Thur:62
##  3rd Qu.:24.13   3rd Qu.: 3.562
##  Max.   :50.81   Max.   :10.000
##       size         tiprate
##  Min.   :1.00   Min.   :0.03564
##  1st Qu.:2.00   1st Qu.:0.12913
##  Median :2.00   Median :0.15477
##  Mean   :2.57   Mean   :0.16080
##  3rd Qu.:3.00   3rd Qu.:0.19148
##  Max.   :6.00   Max.   :0.71034
```

```
#tiprate의 평균= 16%
```

```
#tiprate의 표준편차
sd(TIPS$tiprate)
```

```
## [1] 0.0610722
```

```
skim(TIPS)
```

Data summary

Name	TIPS
Number of rows	244
Number of columns	8
_____	
Column type frequency:	
factor	4
numeric	4
_____	
Group variables	None

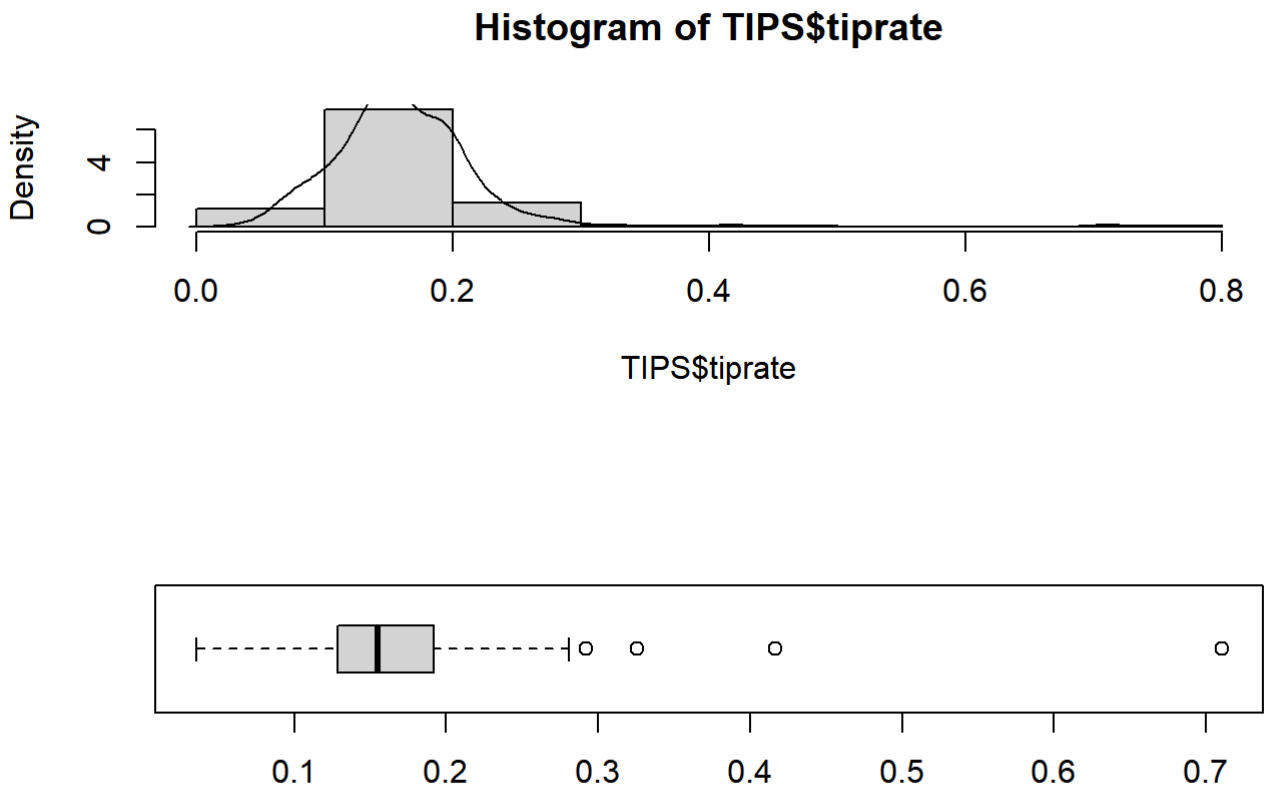
Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
sex	0	1	FALSE	2	Mal: 157, Fem: 87
smoker	0	1	FALSE	2	No: 151, Yes: 93
day	0	1	FALSE	4	Sat: 87, Sun: 76, Thu: 62, Fri: 19
time	0	1	FALSE	2	Din: 176, Lun: 68

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
total_bill	0	1	19.79	8.90	3.07	13.35	17.80	24.13	50.81	
tip	0	1	3.00	1.38	1.00	2.00	2.90	3.56	10.00	
size	0	1	2.57	0.95	1.00	2.00	2.00	3.00	6.00	
tiprate	0	1	0.16	0.06	0.04	0.13	0.15	0.19	0.71	

```
par(mfrow=c(2,1))
hist(TIPS$tiprate,prob=TRUE)
lines(density(TIPS$tiprate))
boxplot(TIPS$tiprate,horizontal=TRUE)
```

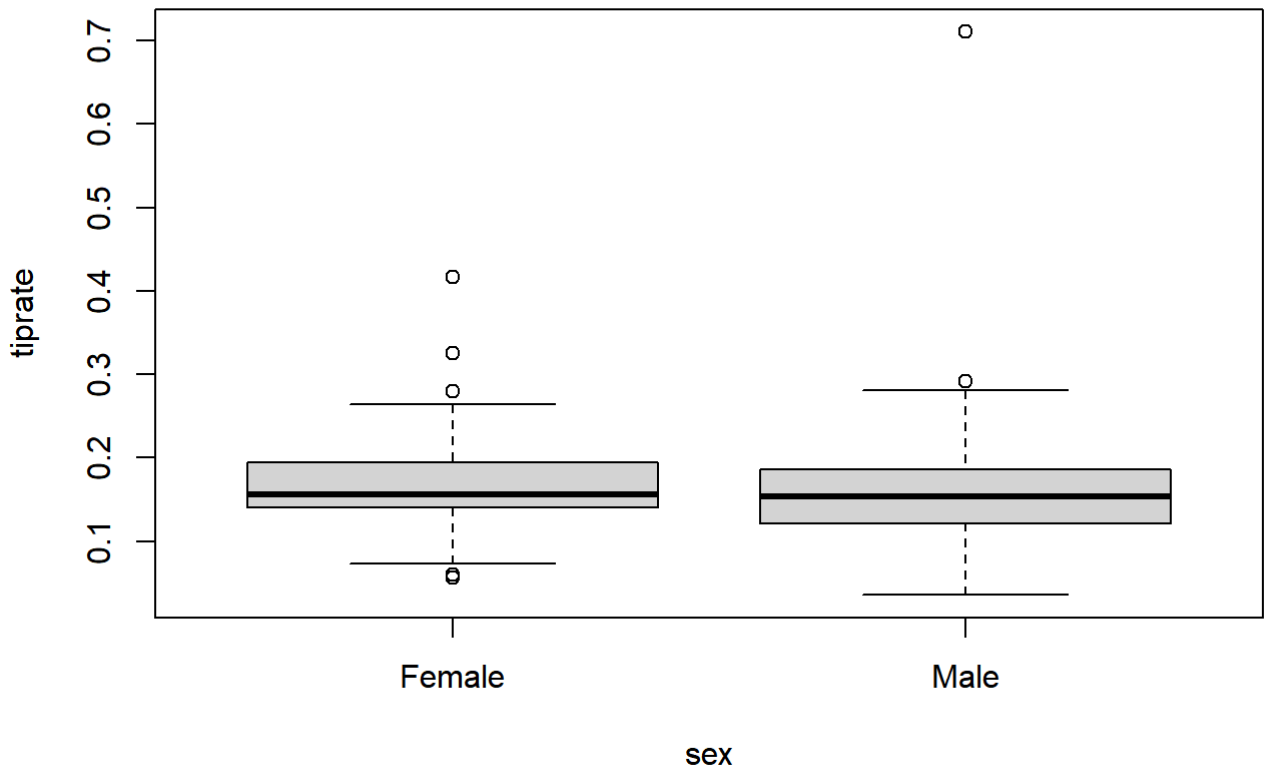



```
TIPS %>% group_by(sex)%>% summarize_at('tiprate',list('mean(tiprate)'=mean, 'sd(tiprate)'=sd))
```

```
## # A tibble: 2 x 3
##   sex    `mean(tiprate)` `sd(tiprate)`
##   <fct>            <dbl>         <dbl>
## 1 Female           0.166        0.0536
## 2 Male             0.158        0.0648
```

```
#여성의 봉사료비율이 더 높음, 평균이 더 높기 때문
```

```
boxplot(tiprate~sex,data=TIPS)
```

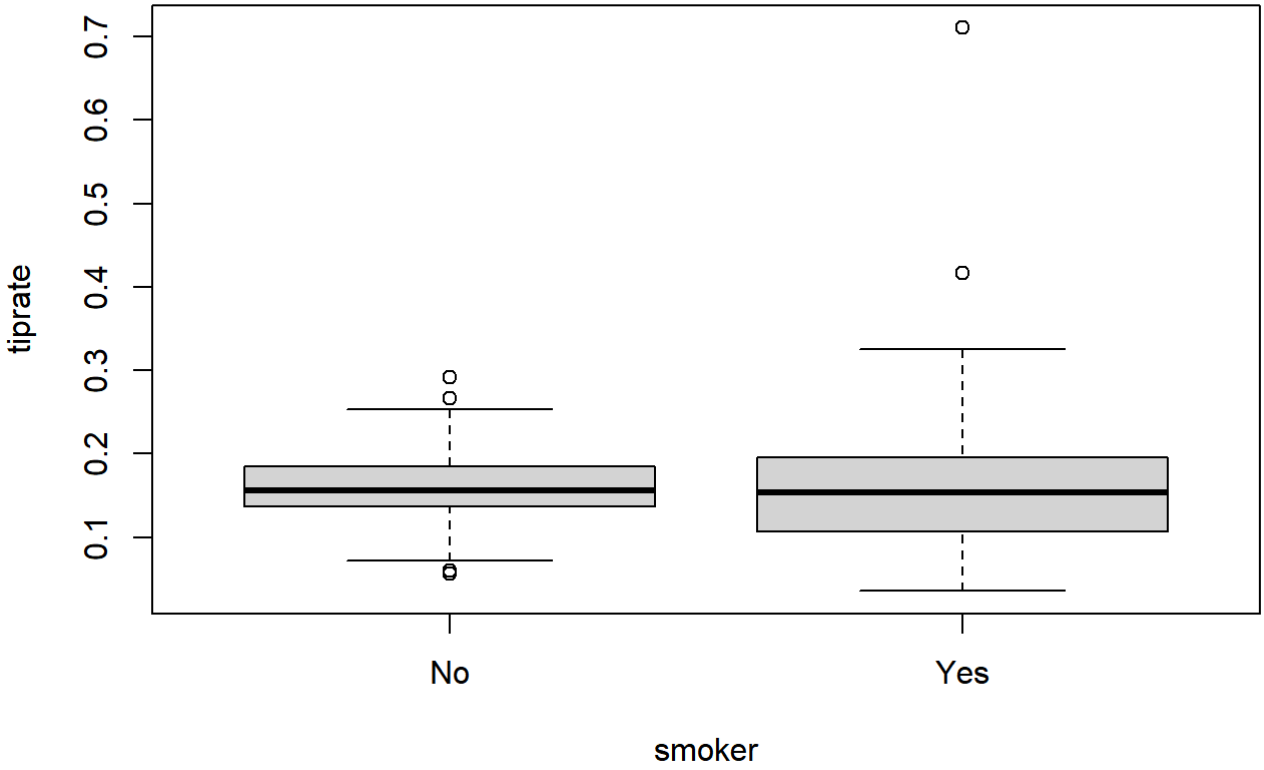



```
TIPS %>% group_by(smoker)%>% summarize_at('tiprate',list('mean(tiprate)'=mean, 'sd(tiprate)'=s
d))
```

```
## # A tibble: 2 x 3
##   smoker `mean(tiprate)` `sd(tiprate)`
##   <fct>            <dbl>         <dbl>
## 1 No               0.159        0.0399
## 2 Yes              0.163        0.0851
```

```
#흡연자가 봉사료비율이 더 높음
```

```
boxplot(tiprate~smoker,data=TIPS)
```



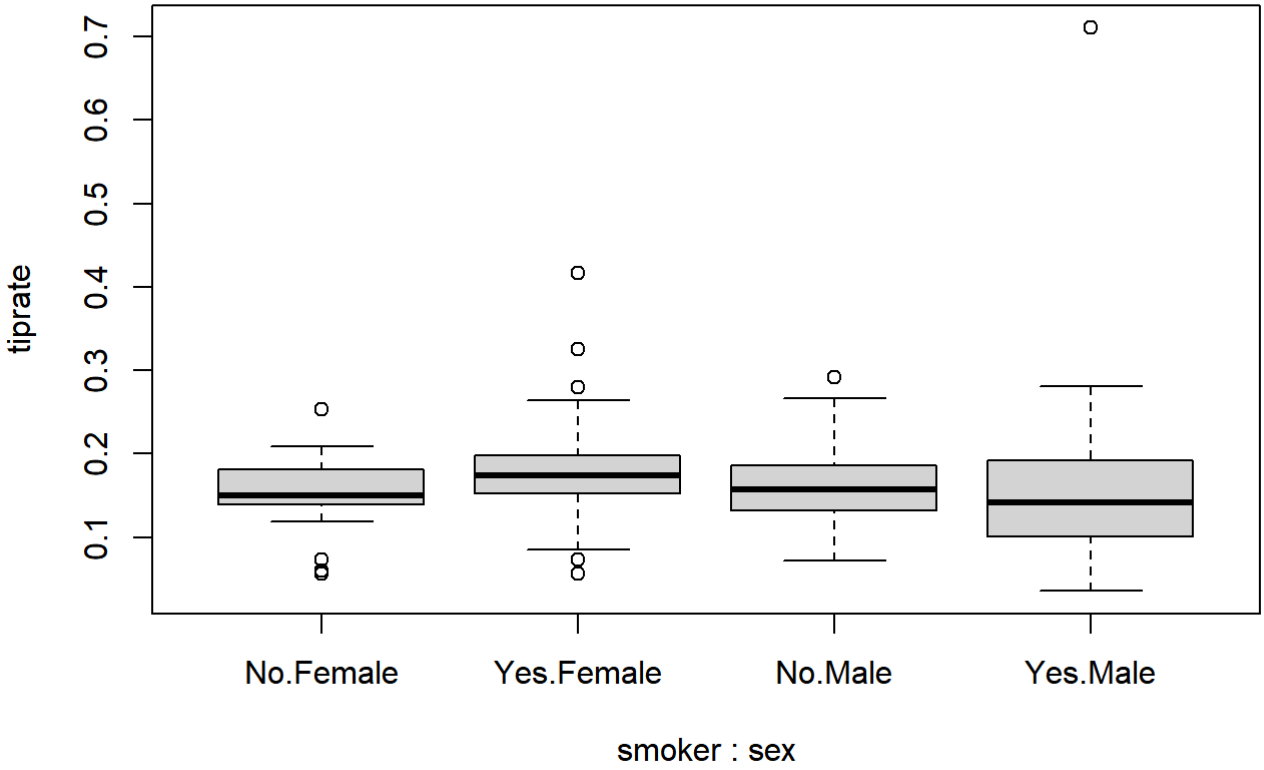

```
TIPS %>% group_by(smoker, sex) %>% summarize(mean(tiprate), sd(tiprate))
```

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

```
## # A tibble: 4 x 4
## # Groups:   smoker [2]
##   smoker sex    `mean(tiprate)` `sd(tiprate)`
##   <fct>  <fct>            <dbl>         <dbl>
## 1 No     Female           0.157        0.0364
## 2 No     Male             0.161        0.0418
## 3 Yes    Female           0.182        0.0716
## 4 Yes    Male             0.153        0.0906
```

```
#여흡연자>남비흡연자>여비흡연자>남흡연자 순으로 높음
```

```
boxplot(tiprate~smoker+sex, data=TIPS)
```

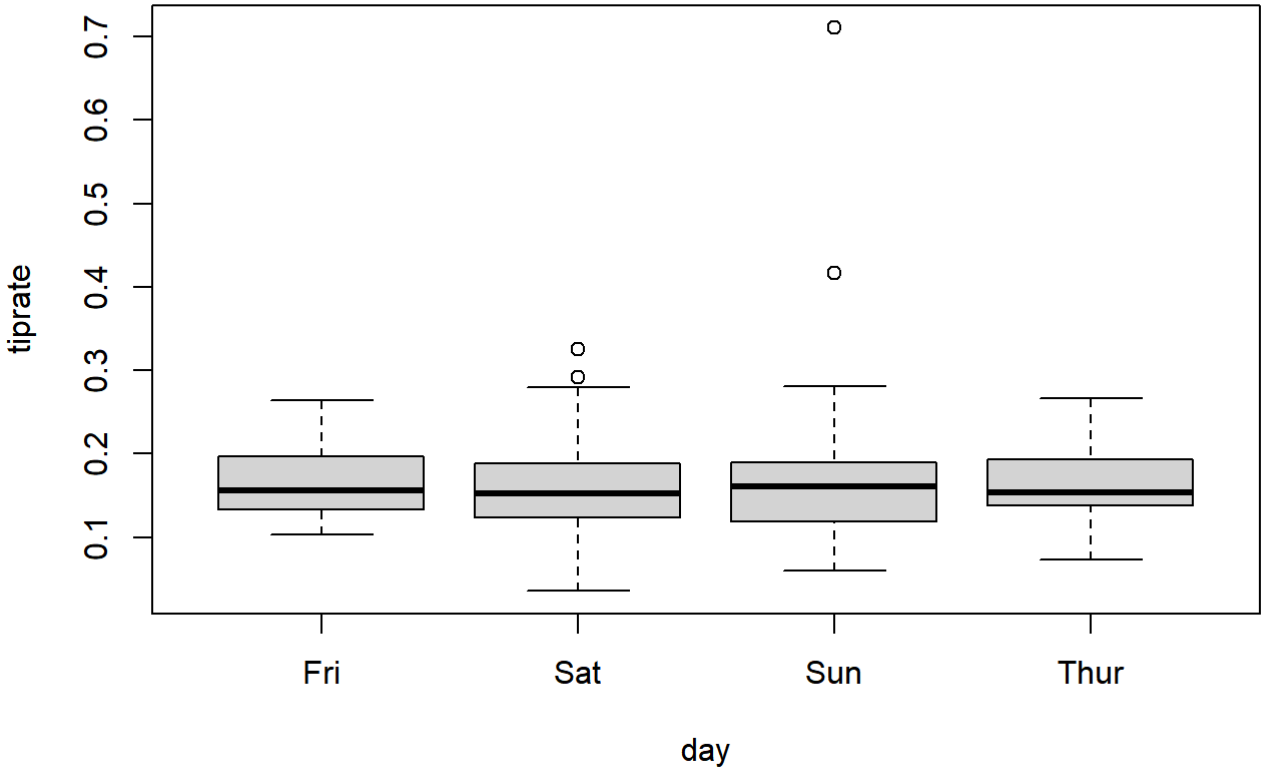



```
TIPS %>% group_by(day)%>% summarize_at('tiprate',list('mean(tiprate)'=mean, 'sd(tiprate)'=sd))
```

```
## # A tibble: 4 x 3
##   day   `mean(tiprate)` `sd(tiprate)`
##   <fct>           <dbl>         <dbl>
## 1 Fri             0.170        0.0477
## 2 Sat             0.153        0.0513
## 3 Sun             0.167        0.0847
## 4 Thur            0.161        0.0387
```

```
#금요일이 가장 봉사료비율이 높음
```

```
boxplot(tiprate~day, data=TIPS)
```

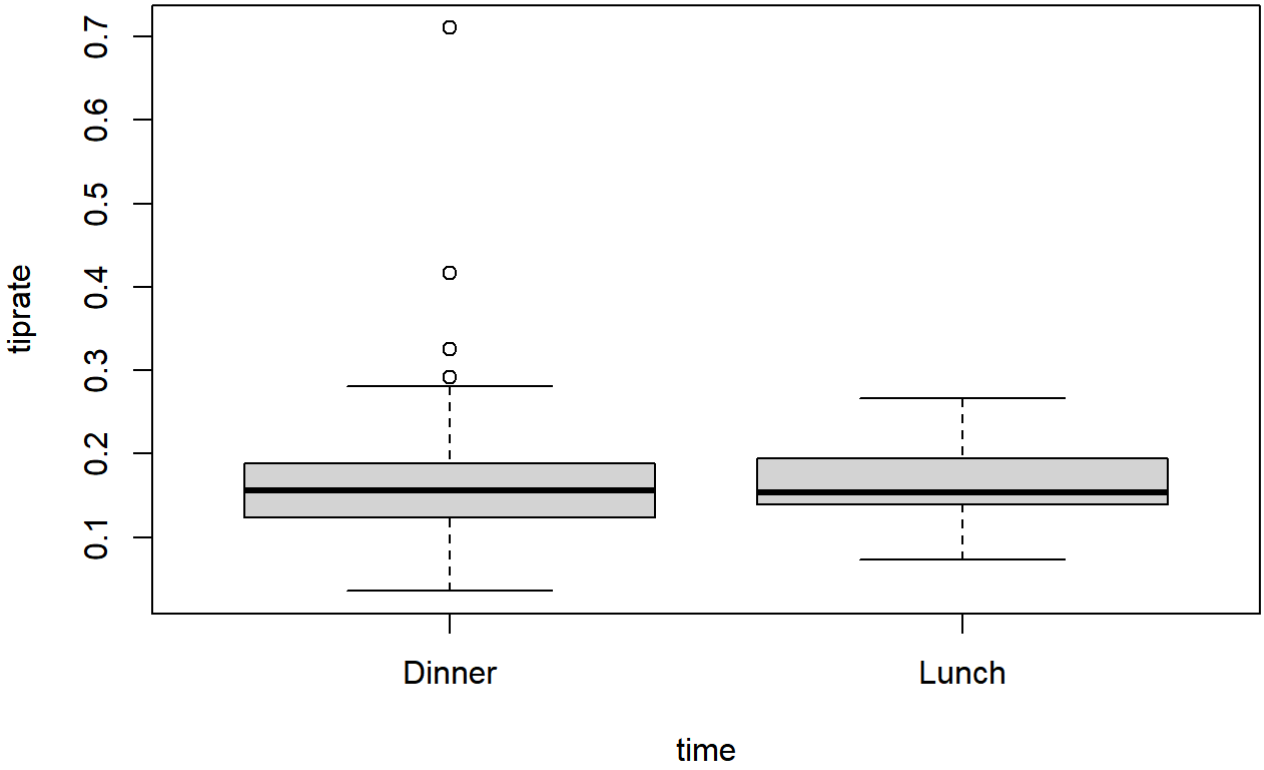



```
TIPS %>% group_by(time)%>% summarize_at('tiprate',list('mean(tiprate)'=mean, 'sd(tiprate)'=sd))
```

```
## # A tibble: 2 x 3
##   time   `mean(tiprate)` `sd(tiprate)`
##   <fct>            <dbl>         <dbl>
## 1 Dinner           0.160        0.0675
## 2 Lunch            0.164        0.0402
```

```
#점심식사 봉사료비율이 더 높음
```

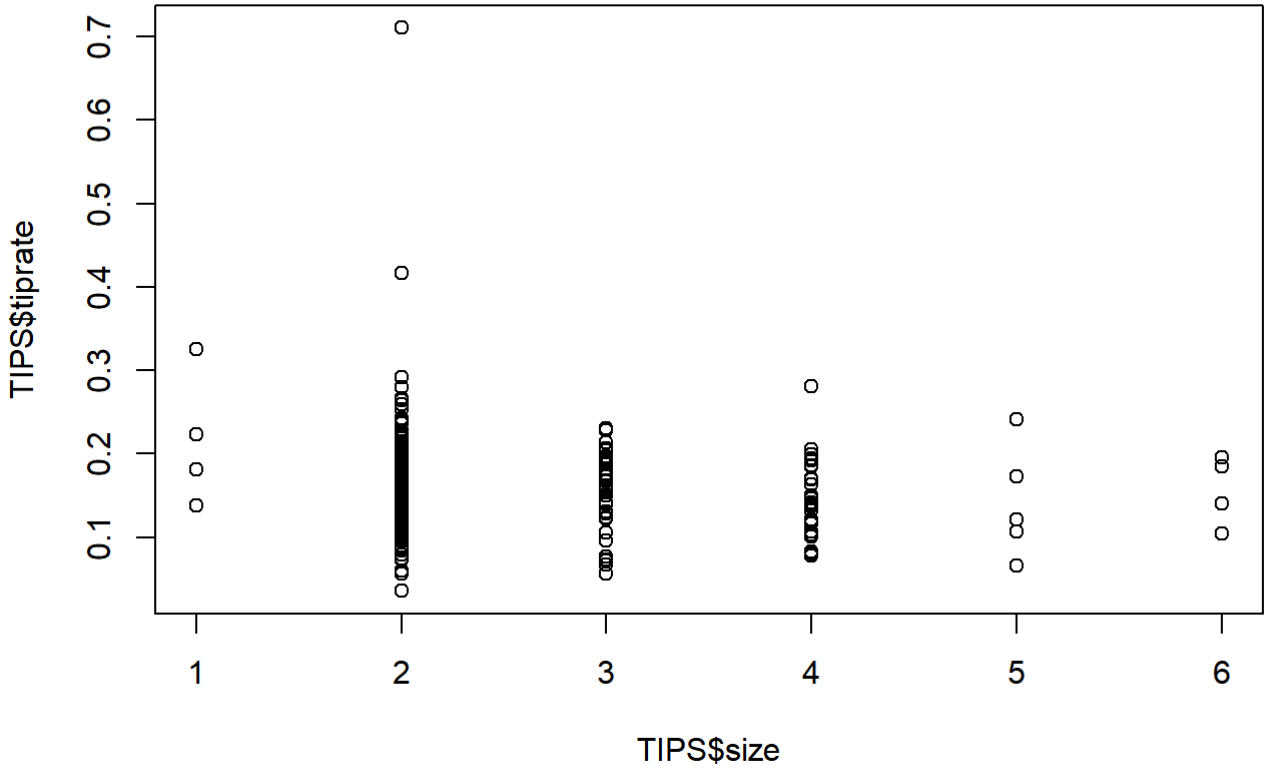
```
boxplot(tiprate~time, data=TIPS)
```




```
#size와 tiprate 상관계수
cor(TIPS$size, TIPS$tiprate)
```

```
## [1] -0.1428596
```

```
#size와 tiprate 산점도
plot(x=TIPS$size,y=TIPS$tiprate)
```



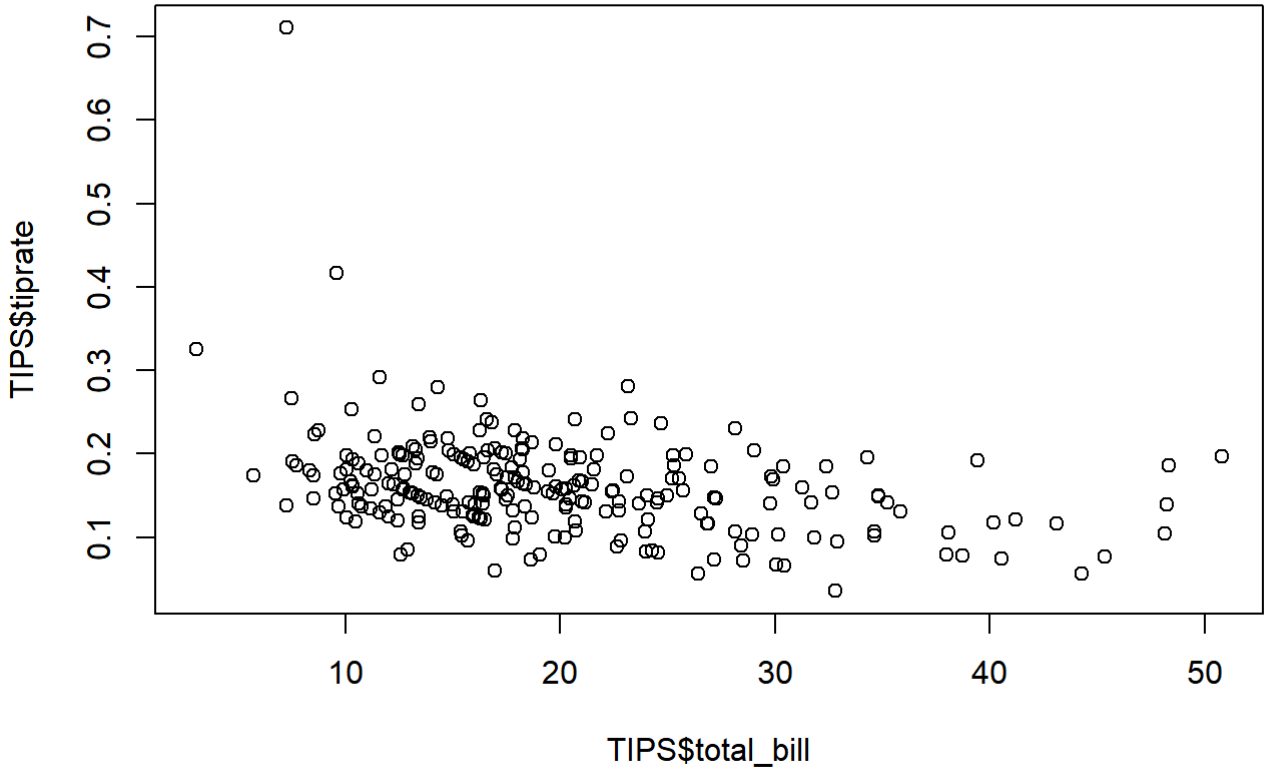

```
#일행이 적을수록 봉사료비율이 높아짐 상관계수 음수이기 때문
```

```
#total_bill과 tiprate 상관계수
cor(TIPS$total_bill, TIPS$tiprate)
```

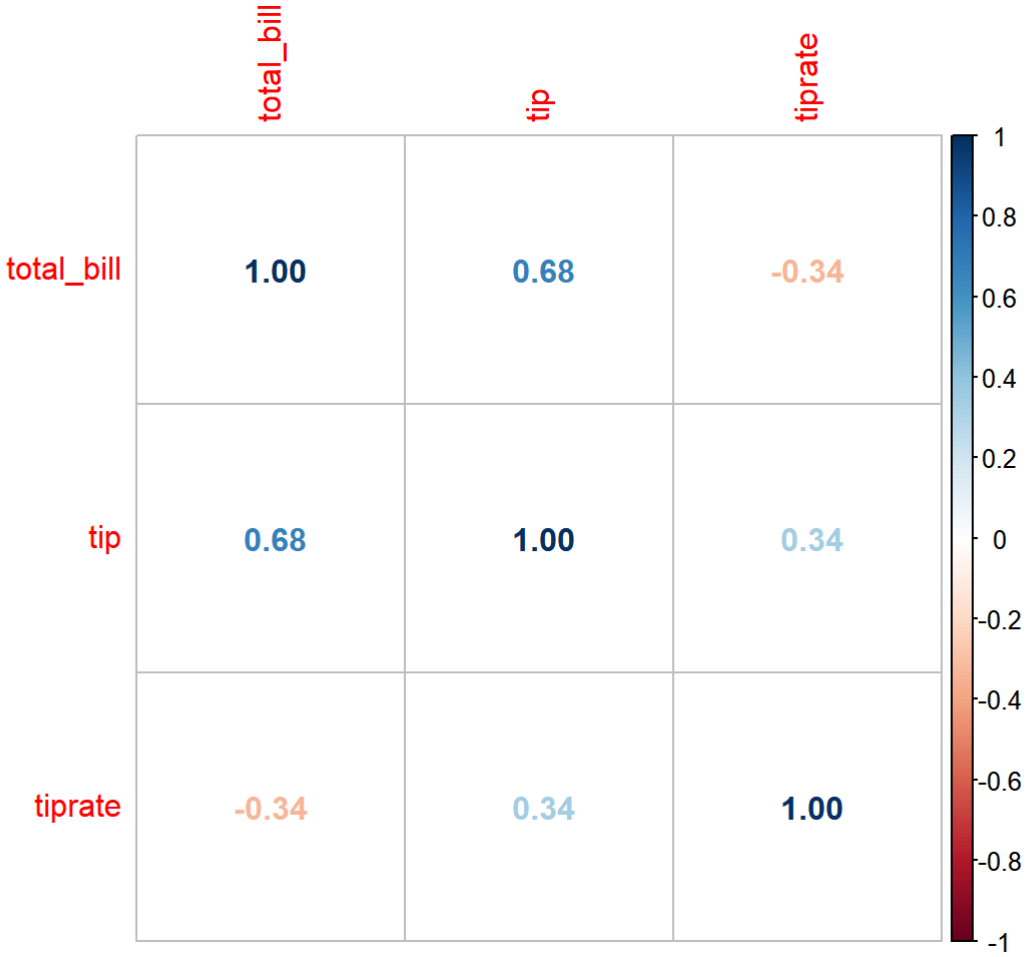
```
## [1] -0.3386241
```

```
#총지불 금액이 많을수록 봉사료비율이 낮아짐
```

```
#size와 tiprate 산점도
plot(x=TIPS$total_bill,y=TIPS$tiprate)
```




```
library(corrplot)
tipsi <- TIPS[,c(1,2,8)]
r<- cor(tipsi,method="pearson")
corrplot(r,method="number")
```


total_bill
tip
tiprate
total_bill
tip
tiprate
1.00
0.68
-0.34
0.68
1.00
0.34
-0.34
0.34
1.00
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

```
#이상치 없다.?
```