Program1

세션 정보/필요한 패키지

```
sessionInfo()
```

```
## R version 4.0.2 (2020-06-22)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19041)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=Korean_Korea.949  LC_CTYPE=Korean_Korea.949
## [3] LC_MONETARY=Korean_Korea.949 LC_NUMERIC=C
## [5] LC_TIME=Korean_Korea.949
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## loaded via a namespace (and not attached):
##  [1] compiler_4.0.2  magrittr_2.0.1  tools_4.0.2     htmltools_0.5.0
##  [5] yaml_2.2.1      stringi_1.5.3   rmarkdown_2.3   knitr_1.29
##  [9] stringr_1.4.0   xfun_0.17       digest_0.6.25   rlang_0.4.7
## [13] evaluate_0.14
```

```
library(tidyverse)
```

```
## Registered S3 methods overwritten by 'tibble':
##   method     from
##   format.tbl pillar
##   print.tbl  pillar
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## √ ggplot2 3.3.2     √ purrr   0.3.4
## √ tibble  3.0.3     √ dplyr   1.0.2
## √ tidyr   1.1.2     √ stringr 1.4.0
## √ readr   1.4.0     √ forcats 0.5.0
```

```
## Warning: package 'dplyr' was built under R version 4.0.3
```

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

```
#library(tidymodels)
```

읽기

```
DF <- read_csv('C:/Users/nabib/Documents/GitHub/Statics/DataMining/0309/sample.csv') # (기본) r
ead.csv
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   sx = col_character(),
##   ht = col_double(),
##   wt = col_double()
## )
```

```
DF
```

```
## # A tibble: 6 x 3
##   sx       ht    wt
##   <chr> <dbl> <dbl>
## 1 F       159    59
## 2 F       161    61
## 3 F       165    62
## 4 M       173    68
## 5 M       177    72
## 6 M       180    82
```

```
# 변수
# 이산형(셀 수 있는 숫자), 연속형(셀 수 없는 숫자)
# 키, 체중 = 연속형, 성별 = 문자(R에서 처리 못함)
# 성별을 이산형으로 변환 (factor화)
# bmi = wt(kg) / ht(m)^2  변수 추가

DF <- mutate(DF,
       sx = factor(sx),
       bmi = wt / (ht / 100)^2)  # DF 데이터에 있는 성별을 factor로 변환한 후 저장

# 파이프 %>%
# A %>% B => 함수 B의 첫번째 인수로  A를 사용 => B(A)

DF <- DF %>% mutate(sx = factor(sx),
                    bmi = wt / (ht / 100)^2)

# 기초통계
summary(DF)  # factor 변수는 빈도를 알려줌.
```

```
##  sx          ht              wt              bmi
##  F:3   Min.   :159.0   Min.   :59.00   Min.   :22.72
##  M:3   1st Qu.:162.0   1st Qu.:61.25   1st Qu.:22.83
##        Median :169.0   Median :65.00   Median :23.16
##        Mean   :169.2   Mean   :67.33   Mean   :23.44
##        3rd Qu.:176.0   3rd Qu.:71.00   3rd Qu.:23.48
##        Max.   :180.0   Max.   :82.00   Max.   :25.31
```

```
# 성별 기초통계
DF %>%
  group_by(sx) %>%
  summarize(mnht=mean(ht),
            mnwt=mean(wt),
            mnbmi=mean(bmi))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 2 x 4
##   sx     mnht  mnwt mnbmi
##   <fct> <dbl> <dbl> <dbl>
## 1 F      162.  60.7  23.2
## 2 M      177.  74    23.7
```

```
# 성별 표준편차
DF %>%
  group_by(sx) %>%
  summarize(sdht = sd(ht),
            sdwt = sd(wt),
            sdbmi = sd(bmi))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 2 x 4
##   sx     sdht  sdwt sdbmi
##   <fct> <dbl> <dbl> <dbl>
## 1 F      3.06  1.53 0.395
## 2 M      3.51  7.21 1.42
```

```
# 성별 빈도
table(DF$sx)
```

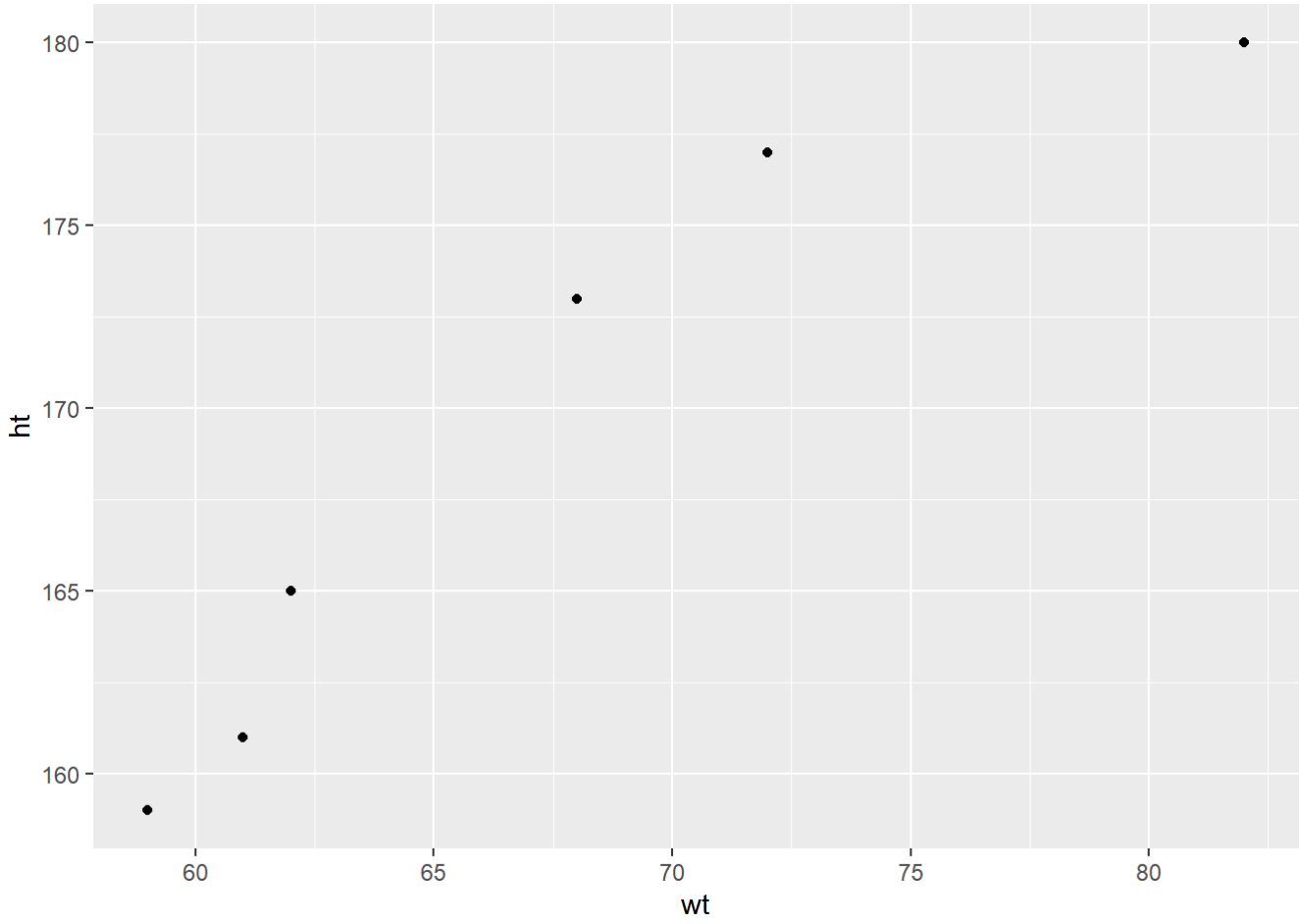
```
##
## F M
## 3 3
```

```
xtabs(~sx, data=DF)
```

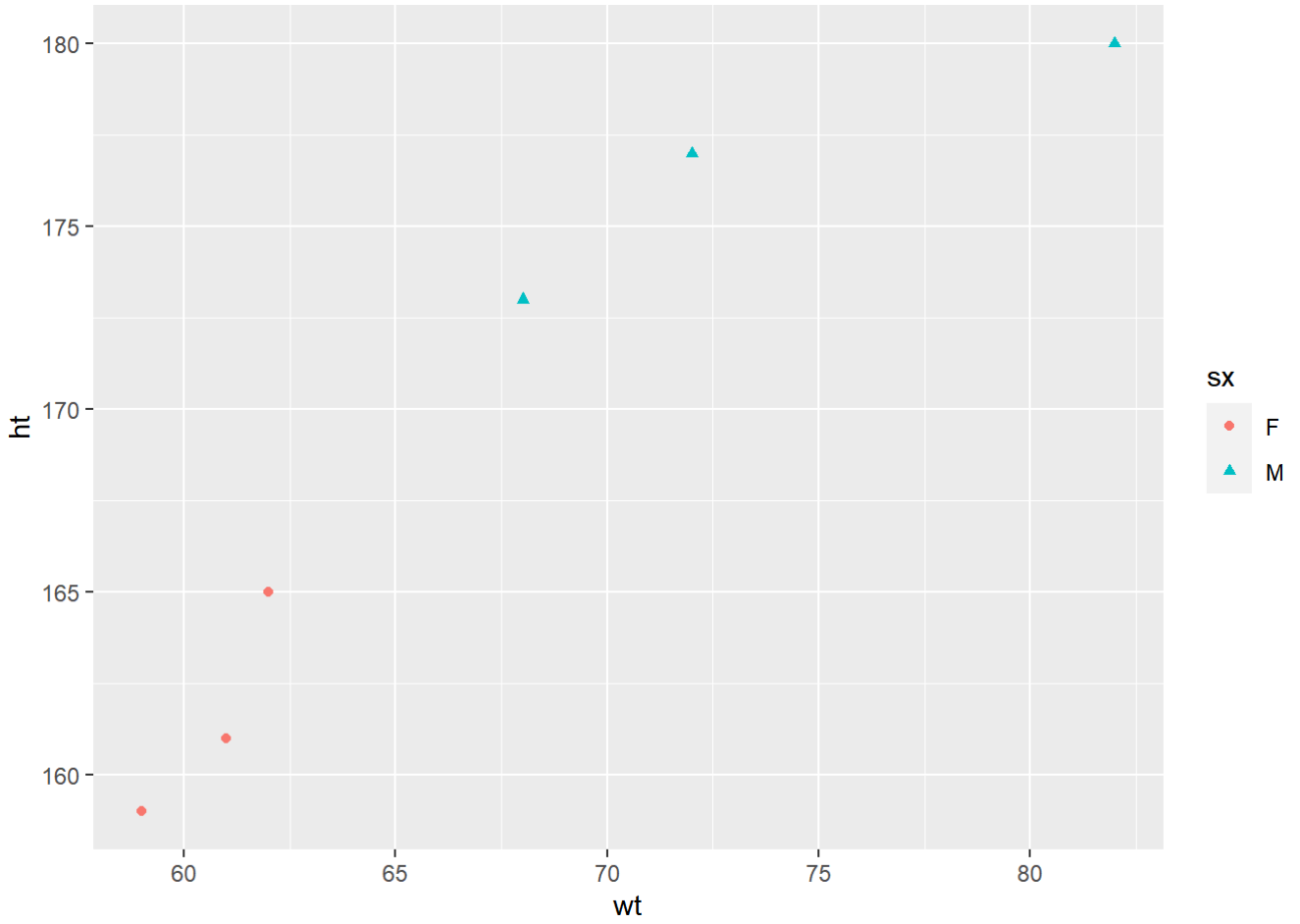
```
## sx
## F M
## 3 3
```

```
# 시각화
# 쓰는 방식: ggplot() + geom_????()
# ggplot: 사용할 데이터 명시, geom_????(): 그림 모양 지정

# 산점도
ggplot(DF, aes(x=wt, y=ht)) + geom_point()
```




```
ggplot(DF, aes(x=wt, y=ht, col=sx, shape=sx)) + geom_point() #성별에 따라 색, 모양 바꿈
```



```
# 상자그림
ggplot(DF, aes(x=sx, y=ht)) + geom_boxplot()
```