R 통계 분석 과제 분석 보고서

60191697 최 솔

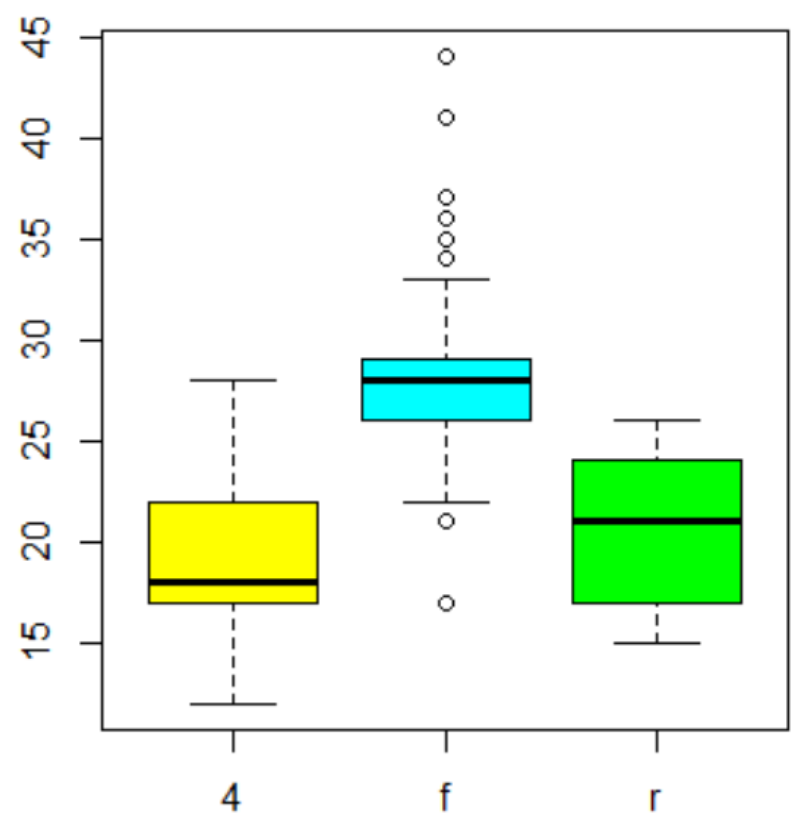
1. 과제 boxplot의 R코드

```r
library(ggplot2)
library(dplyr)
mpg=as.data.frame(ggplot2::mpg)
k = boxplot(mpg$hwy~mpg$drv, data=mpg, xlab="",ylab="",col=c("yellow","cyan","green"))
k
```

2. 과제 boxplot의 출력 (1)

```
> k
$stats
     [,1] [,2] [,3]
[1,]   12   22   15
[2,]   17   26   17
[3,]   18   28   21
[4,]   22   29   24
[5,]   28   33   26
attr(,"class")
        4
"integer"

$n
[1] 103 106  25

$conf
         [,1]     [,2]   [,3]
[1,] 17.22159 27.53961 18.788
[2,] 18.77841 28.46039 23.212

$out
 [1] 17 21 34 36 36 35 37 35 44 44 41

$group
 [1] 2 2 2 2 2 2 2 2 2 2 2

$names
[1] "4" "f" "r"
```

3. 과제 boxplot의 출력 (2)



3. 과제 boxplot 해석

구동 방식에 따른 고속도로 연비 분포를 해석 하기 위해, R통계를 이용해 boxplot 그래프를 생성하였다. 그리고 그래프로 알아낼 수 있는 값을 표로 정리하여 분석했다.

각 boxplot의 범위는 극단치를 포함한 (최댓값 – 최솟값)이다. drv가 4인 경우는 범위가 16, f인 경우는 범위가 24, r인 경우는 11로, 범위가 가장 큰 것은 drv가 f인 것이고, 가장 작은 것은 drv가 r인 것임을 알 수 있다.

하지만 이러한 범위는 극단적인 두 값을 이용해 도출되었기 때문에 수집한 데이터들의 분포를 왜곡할 가능성이 있다.

그래서 극단적인 데이터를 제외한 데이터 분포를 판단하기 위해, 사분 편차가 필요하다. 사분 편차는 3사분위수에서 1사분위수를 뺀 것을 2로 나눈 것으로 전체 데이터의 ¼ 지점과 ¾ 지점 사이의 데이터 분포를 볼 수 있다. drv가 4인 경우는 사분 편차가 2.5, drv가 f 인 경우는 사분 편차가 1.5, drv가 r 인 경우는 사분 편차가 3.5이다. 이렇게 구한 사분 편차를 이용해 r, 4, f 순으로 분포가 큼을 알 수 있다.

마지막으로, 극단치와 2사분위수의 위치가 데이터 분포에 영향을 미치기 때문에 관련 정보를 서술하며 보고서를 마무리하고자 한다. 극단치는 drv가 f인 경우에만 8개가 존재한다. 그리고, 2사분위수가 drv가 4인 경우는 3사분위수보다 1사분위수에 더 가깝고, drv가 f, r인 경우는 1사분위수보다 3사분위수에 더 가까움을 참고해야 한다.