데이터 분석 기초

데이터 파악하기

메소드

함수	기능
head()	데이터 앞부분 출력
tail()	데이터 뒷부분 출력
View()	뷰어 창에서 데이터 확인
dim()	데이터 차원 출력
str()	데이터 속성 출력
summary()	요약통계량 출력

실습

```
exam <- read.csv("csv_exam.csv")
# 앞에서부터 6행까지 출력
head(exam)
# 앞에서부터 10행까지 출력
head(exam, 10)
# 뒤에서부터 6행까지 출력
tail(exam)
# 뒤에서부터 10행까지 출력
tail(exam, 10)
```

```
View(exam)
# 행, 열 출력
dim(exam)
# 데이터 속성 확인
str(exam)
# 요약통계량 출력
summary(exam)
```

```
# ggplo2의 mpg 데이터를 데이터 프레임 형태로 불러오기
mpg <- as.data.frame(ggplot2::mpg)
```

데이터 수정하기

실습 – 행이름 변경

```
# dplyr 설치
install.packages("dplyr")
# dplyr 로드
library(dplyr)
df_raw <- data.frame(var1 = c(1, 2, 1), var2 = c(2, 3, 2))
df_raw
# 복사본 생성
df_new <- df_raw
# 출력
df_new
# var2를 v2로 수정
df_new <- rename(df_new, v2 = var2)
#두 변수의 차이점 비교
df_new
df_raw
```

Quiz

❖ mpg 데이터의 변수명은 긴 단어를 짧게 줄인 축약어로 되어있습니다. cty 변수는 도시 연비, hwy 변수는 고속도로 연비를 의미합니다. 변수명을 이해하기 쉬운 단어로 바꾸려고 합니다. mpg 데이터를 이용해서 아래 문제를 해결해 보세요

- Q1. ggplot2 패키지의 mpg 데이터를 사용할 수 있도록 불러온 뒤 복사본을 만드세요.
- Q2. 복사본 데이터를 이용해서 cty는 city로, hwy는 highway로 변수명을 수정하세요.
- Q3. 데이터 일부를 출력해서 변수명이 바뀌었는지 확인해 보세요. 아래와 같은 결과물이 출력되어야 합니다.

```
##   manufacturer model displ year cyl      trans drv city highway fl   class
## 1         audi    a4   1.8 1999   4   auto(l5)   f   18      29  p compact
## 2         audi    a4   1.8 1999   4 manual(m5)   f   21      29  p compact
## 3         audi    a4   2.0 2008   4 manual(m6)   f   20      31  p compact
## 4         audi    a4   2.0 2008   4   auto(av)   f   21      30  p compact
## 5         audi    a4   2.8 1999   6   auto(l5)   f   16      26  p compact
## 6         audi    a4   2.8 1999   6 manual(m5)   f   18      26  p compact
```

파생변수 만들기

실습 – 평균, 합계

```
df <- data.frame(var1 = c(4, 3, 8),
                 var2 = c(2, 6, 1))
df
# var_sum 파생변수 생성
df$var_sum <- df$var1 + df$var2
df
# var_mean 파생변수 생성
df$var_mean <- (df$var1 + df$var2)/2
df
```

Quiz

❖ head(mpg)

```
  trans drv cty hwy fl    class
1    auto(l5)   f  18  29   p compact
2 manual(m5)   f  21  29   p compact
3 manual(m6)   f  20  31   p compact
4    auto(av)   f  21  30   p compact
5    auto(l5)   f  16  26   p compact
6 manual(m5)   f  18  26   p compact
```

위의 내용은 mpg의 실행 결과이다. cty는 도심에서의 연비이며 hwy는 고속도로에서의 연비이다.

Q1. 복합연비를 구하시오

Q2. 복합연비의 평균값을 구하시오

mpg 분석

```
#요약 통계량 산출
summary(mpg$total)
#히스토그램을 이용한 데이터 분석
h = hist(mpg$total)
#x좌표 수치
h$breaks
#y좌표 수치
h$counts
#pnorm은 정규분포에서 누적치를 구함
mpgMean = mean(mpg$total)
mpgSd = sd(mpg$total)
pnorm(26, mpgMean, mpgSd)
```

고연비 차량 등록

1. summary로 확인한 결과 평균과 중앙값이 대략 20이다.
2. 히스토그램 분석 결과 가장 많은 모델은 20~25사이이다.
3. Counts확인결과 94종류이다.
4. pnorm을 통해 연비가 26이상인 차량은 전체의 13%정도로 적다

위의 내용을 기반으로 고연비 차량을 선별할 경우 평균값이나 중간값을 활용하는 것이 공평할 것으로 판단됨.

실습 – 고연비 차량 분류

```
#조건문 처리를 통한 고연비 차량 분류
mpg$test=ifelse(mpg$total>=mean(mpg$total), "pass", "fail")
#연비 합격 빈도표 생성
table(mpg$test)
#qplot 사용을 위한 library 등록
library(ggplot2)
#막대 그래프 생성
qplot(mpg$test)
```