# 데이터과학을 위한 R프로그래밍

## 5주차. 데이터탐색

**이혜선 교수**
포항공과대학교 산업경영공학과

# 목차

## 5주차. 데이터탐색

5주차

3차시

# 데이터시각화를 이용한 데이터탐색

# 데이터시각화를 이용한 데이터탐색

POSTECH

## 데이터 기술통계치요약

☑ stud_math 데이터 : 포루투칼의 고등학생 수학성적(stud_math_desc.doc 참고)

school : 학교이름 (GP, MS)

sex : 성별 (F, M)

age : 나이 (15-22)

address : 주소 (Urban:도심, Rural:외곽)

Medu : 엄마교육수준

famsize : 가족수 (LE3 :≤3, GT3: >3)

Fedu: 아빠교육수준

traveltime: 통학시간: 1(15분이하),2,3,4(1시간이상)

Dalc : 음주(1-5)

studytime: 주중공부시간: 1(<2시간), 2(2-5시간), 3(5-10시간), 4(>10시간)

health : 건강상태 (1(매우나쁨)-5(매우좋음))

activties : 방과후활동(yes, no)

romantic : 이성교제여부(yes, no)

nursery:유치원다녔는지여부(yes,no)

soout : 친구들과 외출 (1-5)

internet : 집에서 인터넷사용(yes,no)

absences : 학교결석 (0-93)

타겟변수 : G3(최종성적, 0-20), G2(2학년), G1(1학년)

**Attribute Information:**

# Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets:
1 school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2 sex - student's sex (binary: 'F' - female or 'M' - male)
3 age - student's age (numeric: from 15 to 22)
4 address - student's home address type (binary: 'U' - urban or 'R' - rural)
5 famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6 Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7 Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary edu
8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary educa
9 Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10 Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
11 reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12 guardian - student's guardian (nominal: 'mother', 'father' or 'other')
13 traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15 failures - number of past class failures (numeric: n if 1<=n<3, else 4)
16 schoolsup - extra educational support (binary: yes or no)
17 famsup - family educational support (binary: yes or no)
18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19 activities - extra-curricular activities (binary: yes or no)
20 nursery - attended nursery school (binary: yes or no)
21 higher - wants to take higher education (binary: yes or no)
22 internet - Internet access at home (binary: yes or no)
23 romantic - with a romantic relationship (binary: yes or no)
24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29 health - current health status (numeric: from 1 - very bad to 5 - very good)
30 absences - number of school absences (numeric: from 0 to 93)

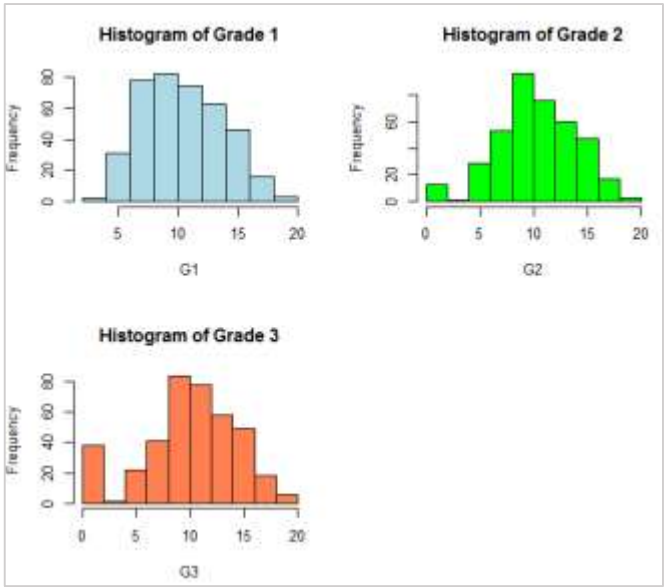
# these grades are related with the course subject, Math or Portuguese:
31 G1 - first period grade (numeric: from 0 to 20)
31 G2 - second period grade (numeric: from 0 to 20)
32 G3 - final grade (numeric: from 0 to 20, output target)

POSTECH

# 그래프를 이용한 데이터탐색

## 히스토그램 (1학년, 2학년, 3학년 성적의 분포)

```
# Graphical analysis
library(dplyr)

# set working directory
setwd("D:/tempstore/moocr")

### student math grade data ####

stud<-read.csv("stud_math.csv")

head(stud)
dim(stud)
str(stud)

attach(stud)

# 1. histogram with color and title, legend
par(mfrow=c(2,2))
hist(G1, breaks = 10, col = "lightblue", main="Histogram of Grade 1" )
hist(G2, breaks = 10, col = "green", main="Histogram of Grade 2" )
hist(G3, breaks = 10, col = "coral", main="Histogram of Grade 3" )
```

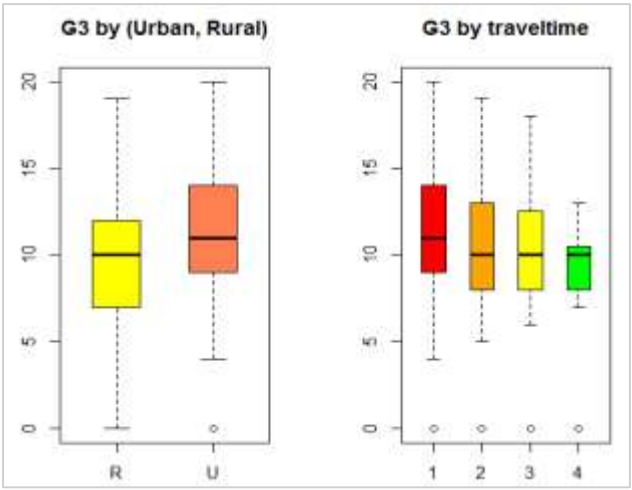


POSTECH

# 그래프를 이용한 데이터탐색

## 상자그림 (거주지역에 따른 G3, 통학시간에 따른 G3)

```
# 2. boxplot
par(mfrow=c(1,2))
boxplot(G3~address, boxwex = 0.5, col = c("yellow", "coral"), main="G3 by
boxplot(G3~traveltime, boxwex = 0.5, col = c("red","orange","yellow","gree
```




예 (1) 도심지역 학생들 성적이 외곽지역 학생들보다 높다

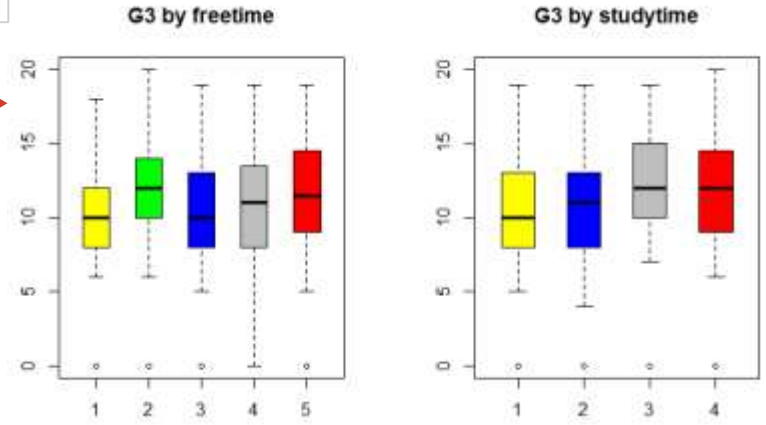
예 (2) 통학시간이 짧은(15분 이내)의 학생들의 성적이 더 높다

POSTECH

# 그래프를 이용한 데이터탐색

## 상자그림 (자유시간에 따른 G3, 공부시간에 따른 G3)

```
# boxplot
par(mfrow=c(1,2))
# academic achievement by freetime
# 1 - very low to 5 - very high
boxplot(G3~freetime, boxwex = 0.5, col = c("yellow", "green", "blue
# academic achievement by studytime
# 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hou
boxplot(G3~studytime, boxwex = 0.5, col = c("yellow", "blue","grey"
```




(1) 방과후 자유시간에 따른 G3의 차이 : 자유시간이 적은 편(low)이라고 응답한 학생들의 성적이 가장 높은데, 다른 요인과 혼합되어 그럴 수 있음..

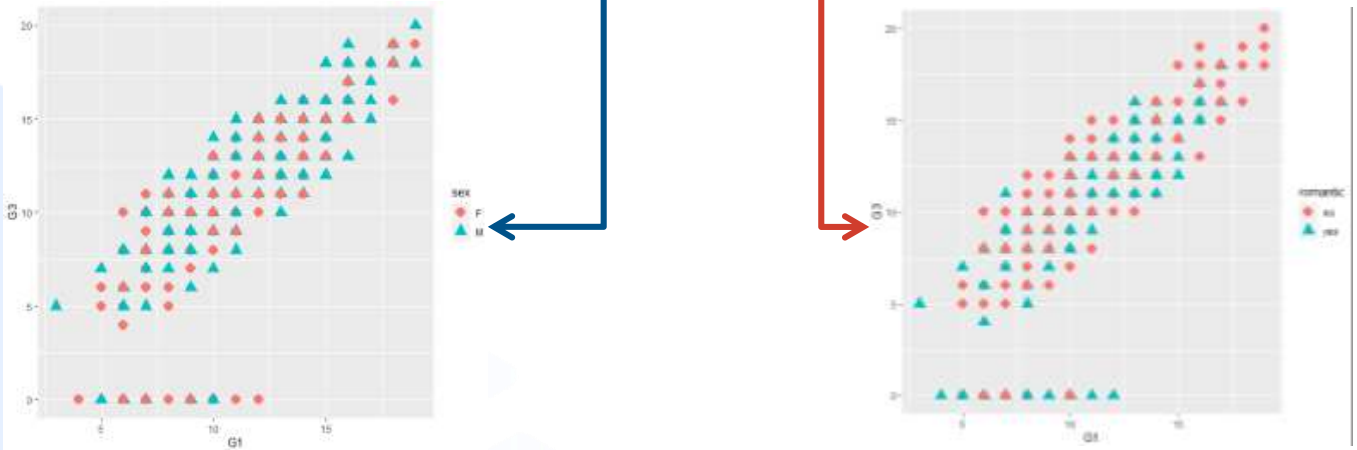
(2) 주중공부시간이 5시간이상 (3: 5-10시간, 4: 10시간이상)인 학생들의 성적이 높은 편

POSTECH

# 그래프를 이용한 데이터탐색

## 산점도 (ggplot2 패키지의 ggplot이용)

```
# ggplot2 package
library(ggplot2)
# 3. scatterplot for G1 and G3 by sex
ggplot(stud, aes(x=G1, y=G3, color=sex, shape=sex)) + geom_point(size=4)
ggplot(stud, aes(x=G1, y=G3, color=romantic, shape=romantic)) + geom_point(size=4)
```



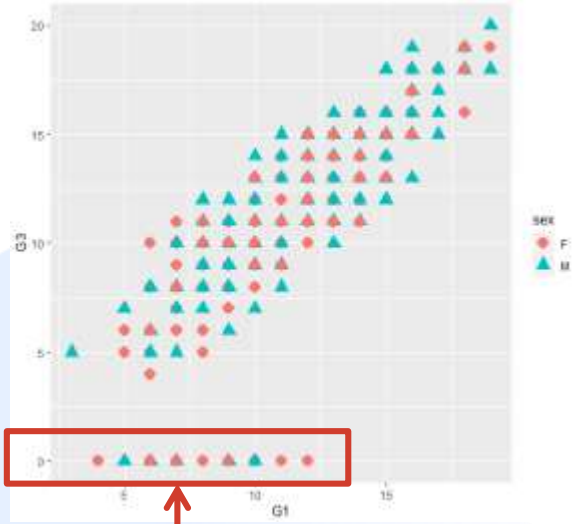
성별이나 연애 경험에 따른 차이는 없음

# 데이터시각화를 이용한 데이터탐색

POSTECH

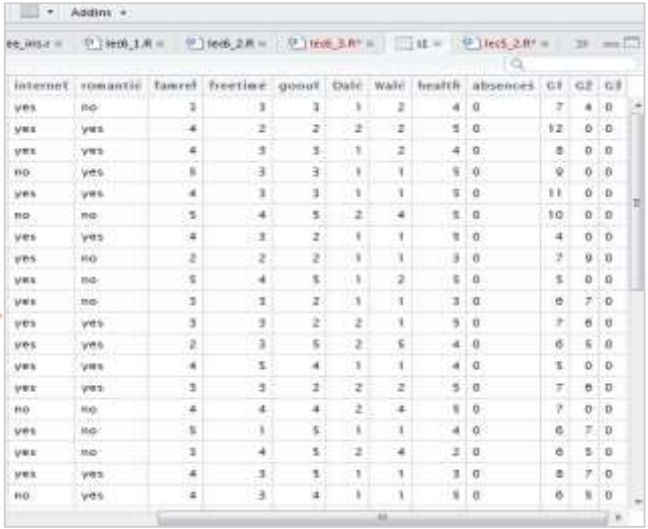
## 그래프를 이용한 데이터탐색

☑ G3=0인 데이터 (n=38명)




0점인 데이터 확인, 점검 필요

```
# data (G3=0)
s1<-subset(stud, G3==0)
#ggplot(data=s1, aes(factor(s1$add
#ggplot(data=s1, aes(factor(s1$int
```



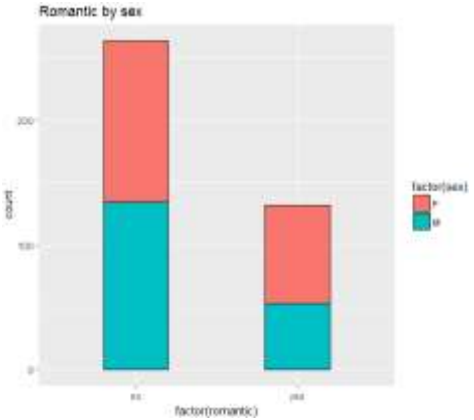
| internet | romantic | famrel | freetime | goout | Dalc | Walc | health | absences | G1 | G2 | G3 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| yes | no | 3 | 3 | 3 | 1 | 2 | 4 | 0 | 7 | 4 | 0 |
| yes | yes | 4 | 2 | 2 | 2 | 2 | 5 | 0 | 12 | 0 | 0 |
| yes | yes | 4 | 3 | 3 | 1 | 2 | 4 | 0 | 8 | 0 | 0 |
| no | yes | 5 | 3 | 3 | 1 | 1 | 5 | 0 | 9 | 0 | 0 |
| yes | yes | 4 | 3 | 3 | 1 | 1 | 5 | 0 | 11 | 0 | 0 |
| no | no | 5 | 4 | 5 | 2 | 4 | 5 | 0 | 10 | 0 | 0 |
| yes | yes | 4 | 3 | 2 | 1 | 1 | 5 | 0 | 4 | 0 | 0 |
| yes | no | 2 | 2 | 2 | 1 | 1 | 3 | 0 | 7 | 9 | 0 |
| yes | no | 5 | 4 | 5 | 1 | 2 | 5 | 0 | 5 | 0 | 0 |
| yes | no | 3 | 3 | 2 | 1 | 1 | 3 | 0 | 0 | 7 | 0 |
| yes | yes | 3 | 3 | 2 | 2 | 1 | 5 | 0 | 7 | 6 | 0 |
| yes | yes | 2 | 3 | 5 | 2 | 5 | 4 | 0 | 6 | 5 | 0 |
| yes | yes | 4 | 5 | 4 | 1 | 1 | 4 | 0 | 5 | 0 | 0 |
| yes | yes | 3 | 3 | 2 | 2 | 2 | 5 | 0 | 7 | 6 | 0 |
| no | no | 4 | 4 | 4 | 2 | 4 | 5 | 0 | 7 | 0 | 0 |
| yes | no | 5 | 1 | 5 | 1 | 1 | 4 | 0 | 6 | 7 | 0 |
| yes | no | 3 | 4 | 5 | 2 | 4 | 2 | 0 | 6 | 5 | 0 |
| yes | yes | 4 | 3 | 3 | 1 | 1 | 3 | 0 | 8 | 7 | 0 |
| no | yes | 4 | 3 | 4 | 1 | 1 | 5 | 0 | 6 | 5 | 0 |

POSTECH

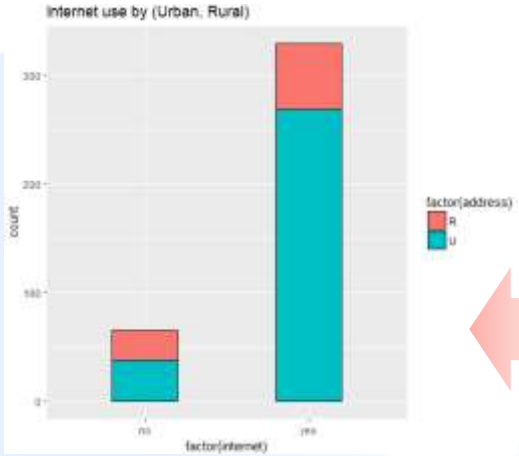
# 그래프를 이용한 데이터탐색

## 막대그림 (ggplot2 패키지의 ggplot이용)

```
# bar chart for romantic by sex
ggplot(data=stud, aes(factor(romantic)))+geom_bar(aes(fill=factor(sex)),
```




예 연애경험 있는 경우 여학생 비율이 높음




```
# bar chart for internet use by (Urban, Rural)
ggplot(data=stud, aes(factor(internet)))+geom_bar(aes(fill=factor(address)
```

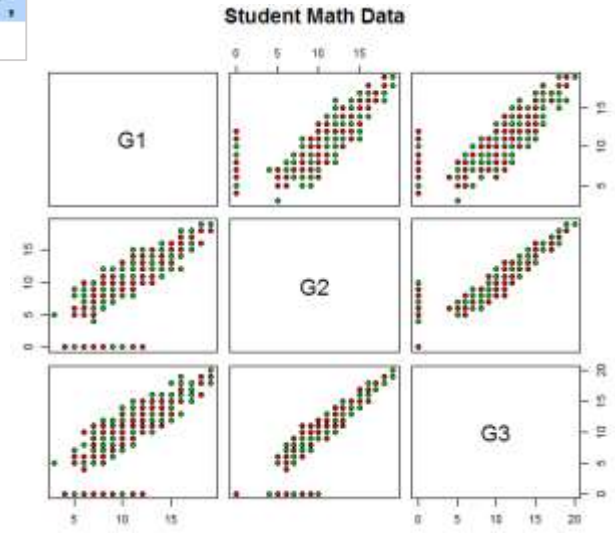
예 인터넷사용자 중에는 도심지역에 사는 경우가 훨씬 높음

POSTECH

# 그래프를 이용한 데이터탐색

## pairwise scatterplot : pairs(변수리스트)

```
# 5. pariwise plot
# new variable lists
vars1<-c("G1", "G2", "G3")
# pariwise plot
pairs(stud[vars1], main = "Student Math Data",
      pch = 21, bg = c ("red","green3"))
```




(1) G1, G2, G3간의 상관성은 매우 높다

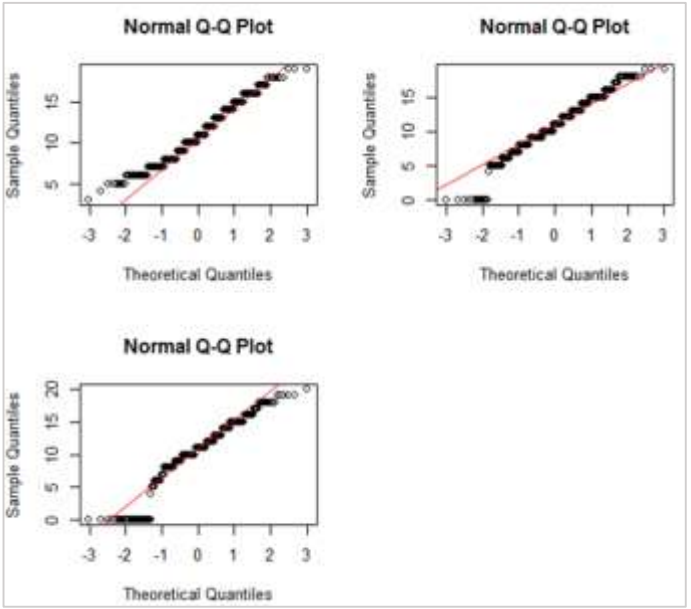
(2) 성별 간 차이는 없다

POSTECH

# 데이터의 정규성검정

☑ 정규확률도 (Normal Q-Q plot) : 데이터가 정규분포하는가?

```
# multiple plot (2 by 2)
par(mfrow=c(2,2))
#Quantile plot
qqnorm(G1)
qqline(G1, col = 2, cex=7)

qqnorm(G2)
qqline(G2, col = 2, cex=7)

qqnorm(G3)
qqline(G3, col = 2, cex=7)
```




예 qqline의 디폴트는 정규분포의 1사분위, 3사분위를 직선

**qqline(y, distribution = qqnorm, probs = c(0.25, 0.75))**

POSTECH

# 데이터의 정규성검정

## 정규분포 적합성검정 : 데이터가 정규분포 하는지에 대한 검정

### (1) Shapiro-Wilks검정

```
#Shapiro-Wilks test
shapiro.test(G3)
```

```
> shapiro.test(G3)

        Shapiro-Wilk normality test

data:  G3
W = 0.92873, p-value = 8.836e-13
```

G3는 정규분포한다고 볼 수 없다 (p-value~0)

### (2) Anderson-Darling검정 (추가패키지 필요)

```
#Anderson-Darling test require installing package "nortest"
install.packages('nortest')
library(nortest)
ad.test(G3)
```

```
> ad.test(G3)

        Anderson-Darling normality test

data:  G3
A = 8.3032, p-value < 2.2e-16
```

G3는 정규분포한다고 볼 수 없다 (p-value~0)