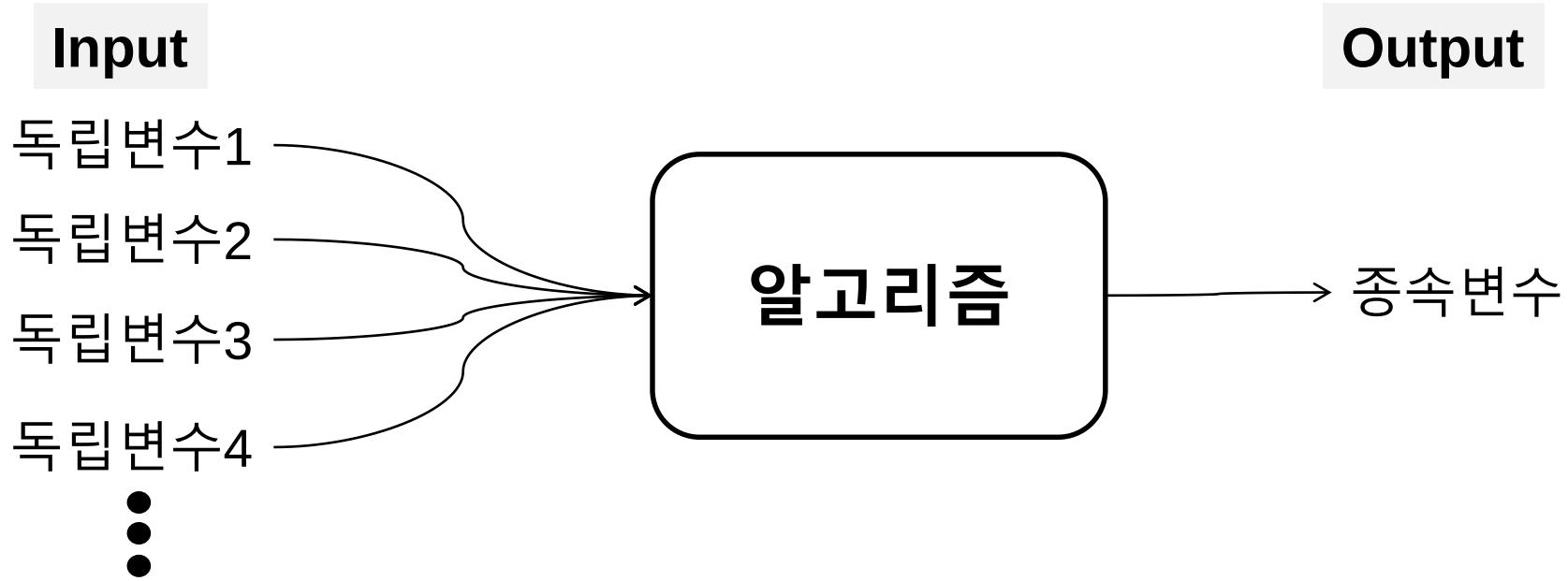
데이터 탐색

D:plus

목차

1. 변수

1.1 종속변수 vs. 독립변수



- ✓ **독립변수** : 특징Feature, 설명변수, 예측변수, 통제변수, 조작변수, 리스크 팩터(risk factor) 등
- ✓ **종속변수** : 반응변수

1.2 변수 유형

❖문자형 / 숫자형

❖범주형

구분	내용
명목형(Nominal)	각 범주간에 순서 없음, 예 : 지역 – 서울/부산/경기
순서형(Ordinal)	순서 있음. 예 : 연령대 – 20대, 30대, 40대…
이항형(Binomial)	두가지 범주. 예 : 성별 – 남,여 / 흡연여부 – 흡연/비흡연

2. R 그래프

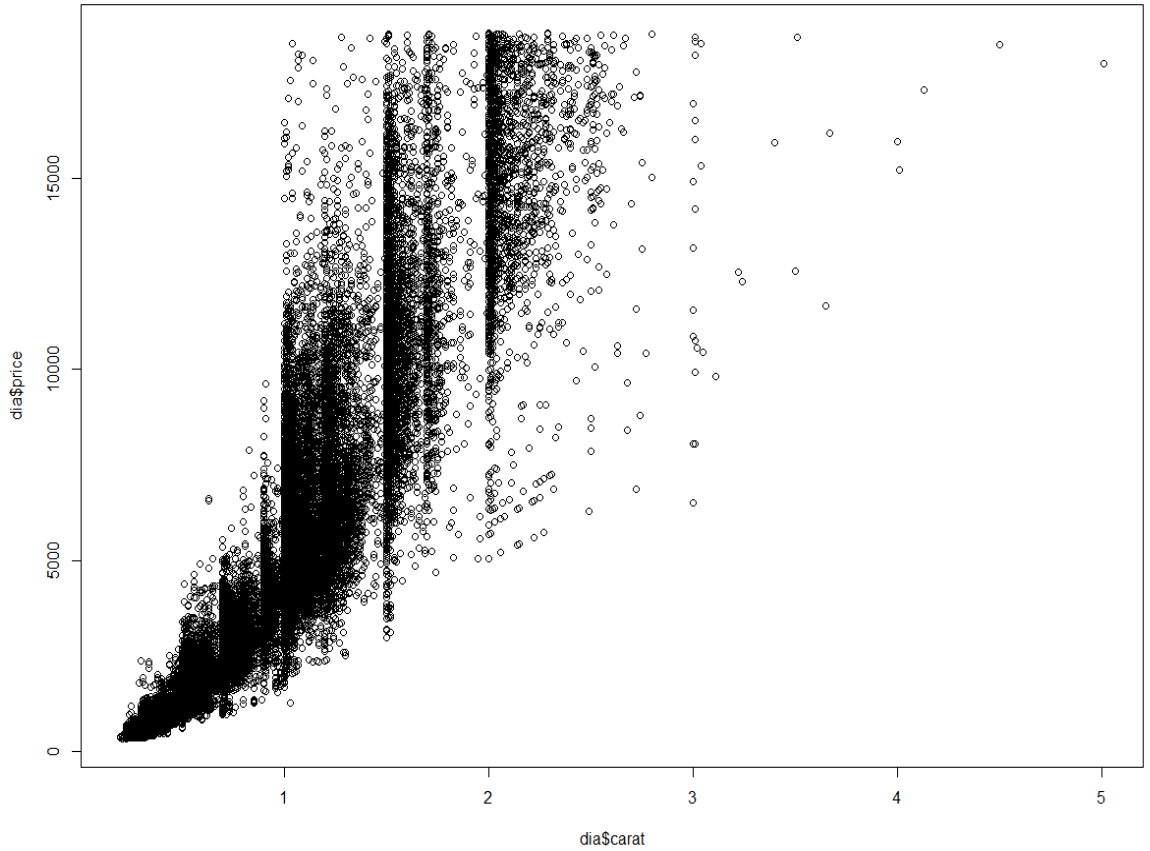
2.1 기본 그래프

Scatter 산점도

- ❖ x축 대비 y축의 값의 분포를 살펴볼 때
- ❖ x축, y축 모두 연속형 데이터
- ❖ 보통 x축은 독립변수, y축은 종속변수
- ❖plot(dia$carat, dia$price)

R code

```
plot(dia$carat, dia$price)
```



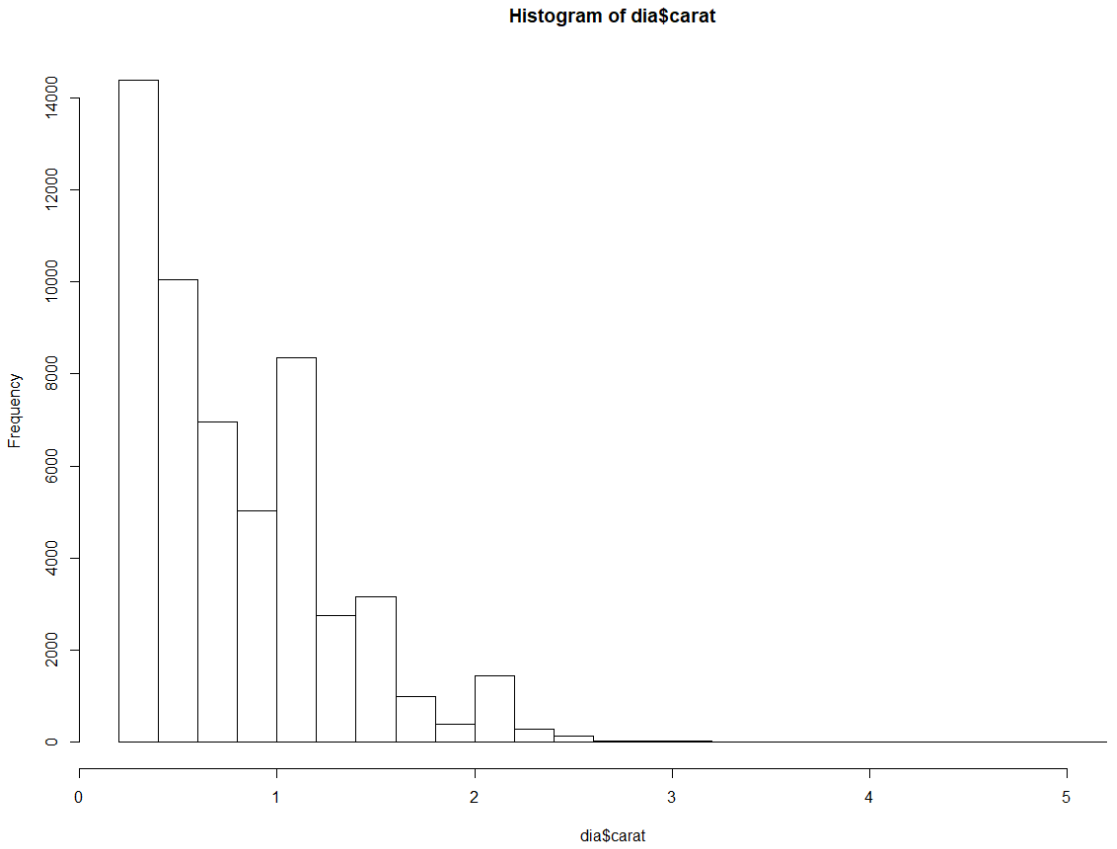
2.1 기본 그래프

Histogram

❖ 특정 변수의 구간별 빈도수 비교
❖ 연속형 변수에서 사용

R code

```
hist(dia$carat, breaks = 20)
```

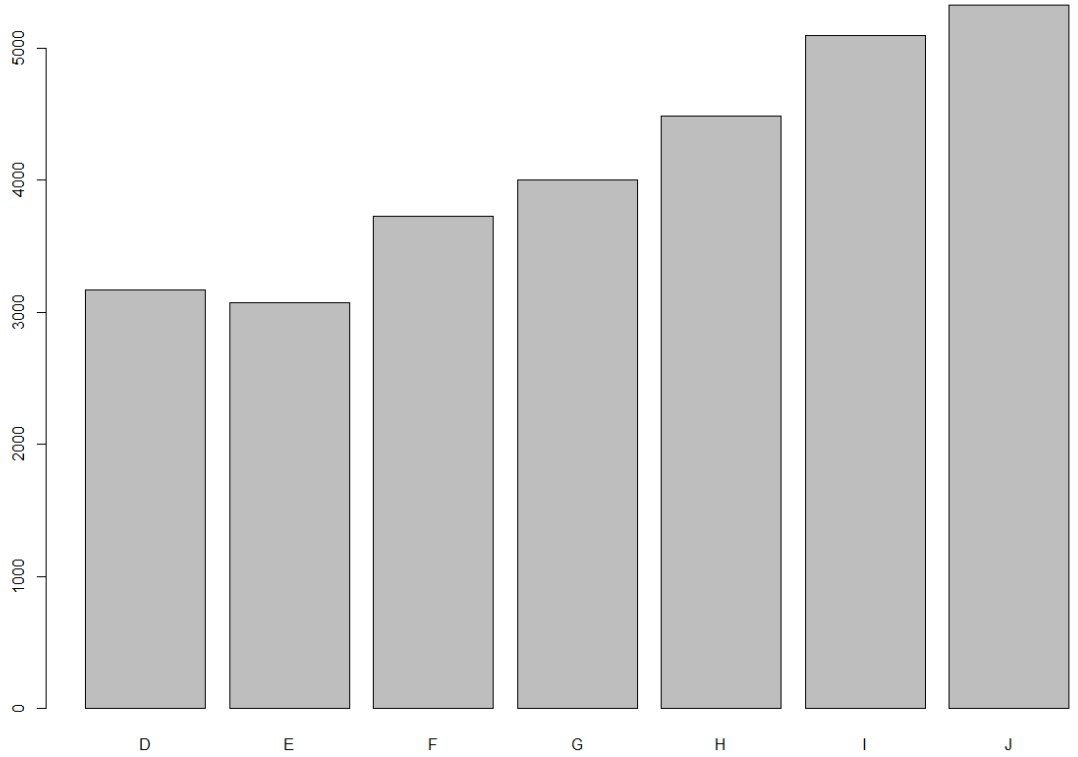


D:plus

2.1 기본 그래프

Barplot

❖ x축 범주형 변수를 대상으로,
y축 값을 비교하기 위해서 사용



R code

```
avgPrice <- aggregate(price ~ color, data=dia, mean)
barplot(avgPrice$price, names.arg=avgPrice$color)
```

D:plus

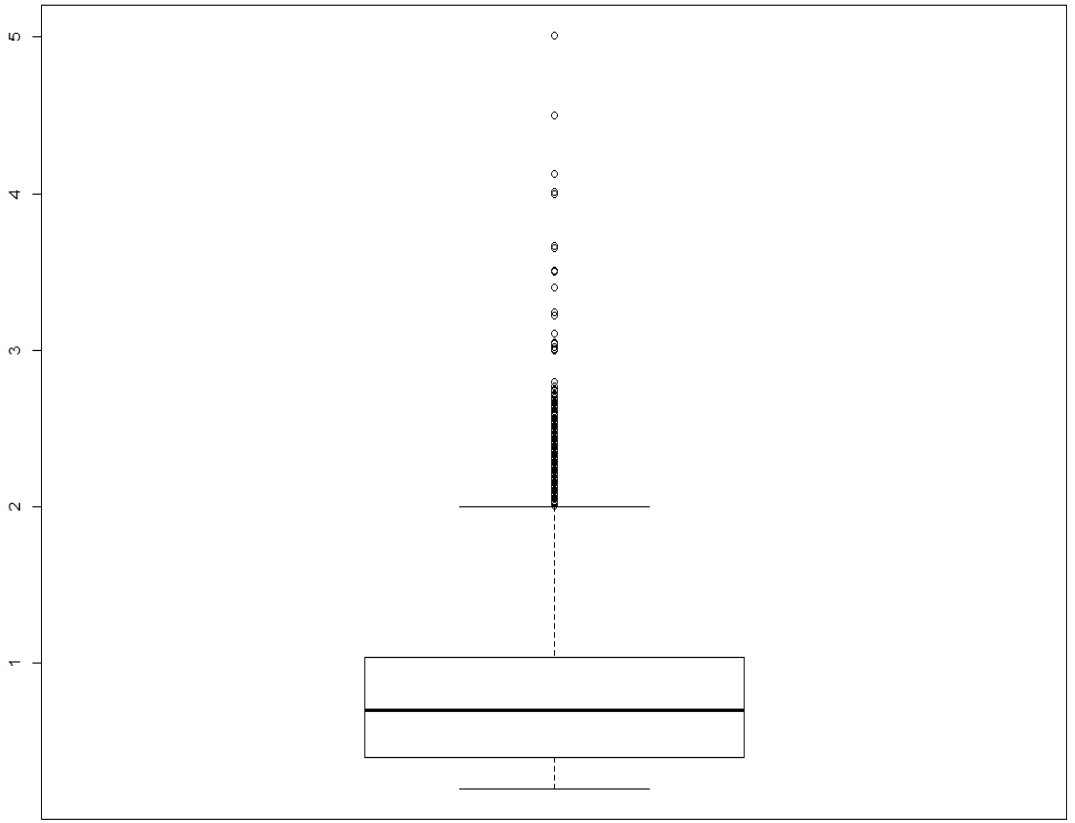
2.1 기본 그래프

Boxplot

❖ 특정 변수의 값의 분포를 확인하기 위해서 사용

R code

```
boxplot(dia$carat)
```

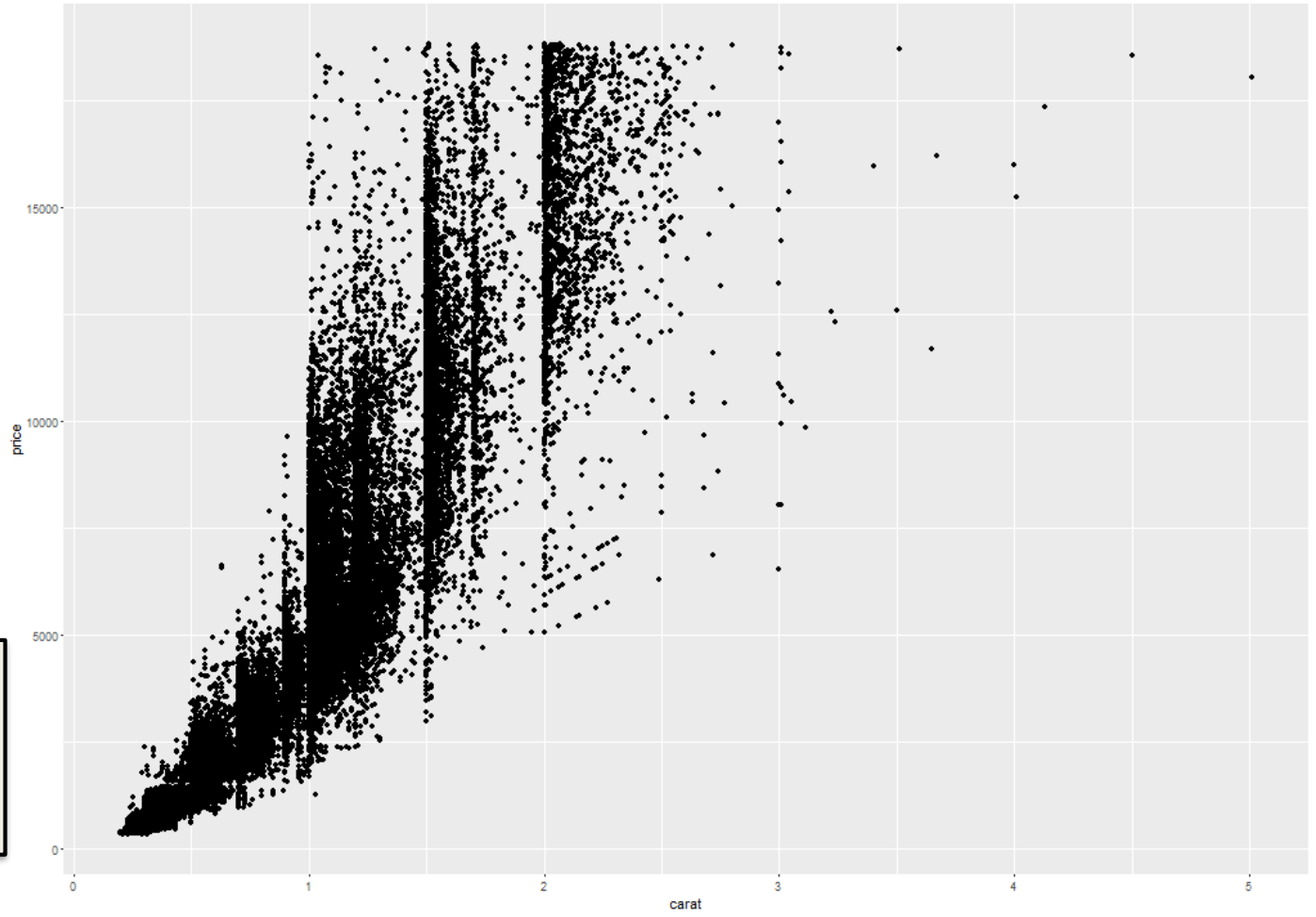


D:plus

2.2 그래프 – ggplot2 패키지 활용

산점도(scatter)

- X에 대한 Y값 점으로 표시
  - X와 Y의 상관관계를 확인
- qplot(x축변수, y축변수, 데이터셋)



R code

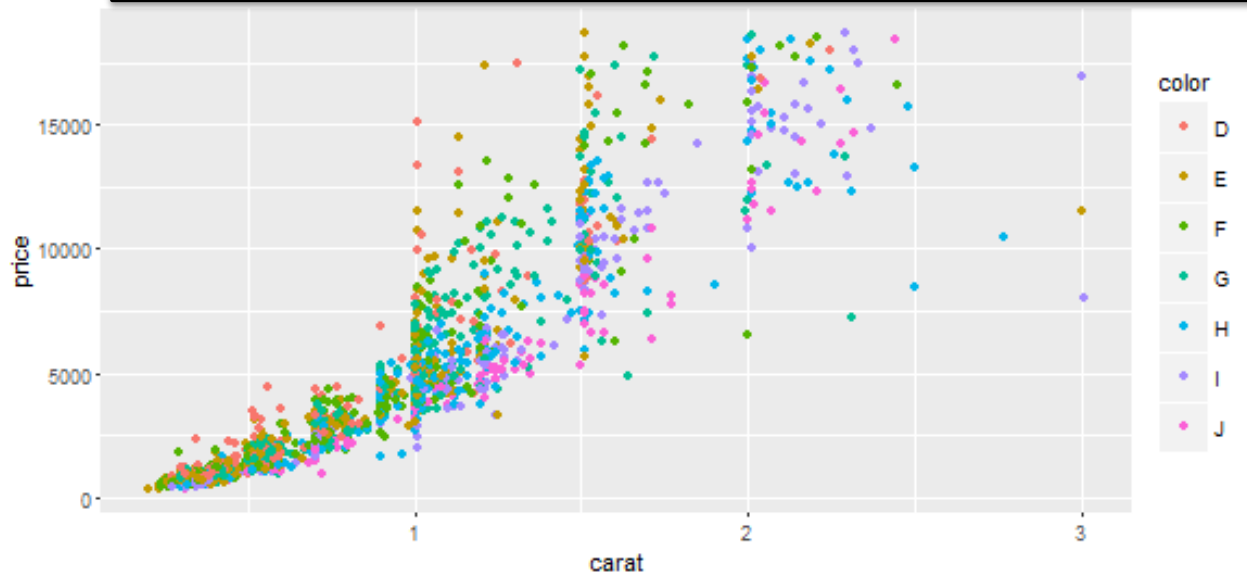
```
qplot(carat, price, data = diamonds)
```

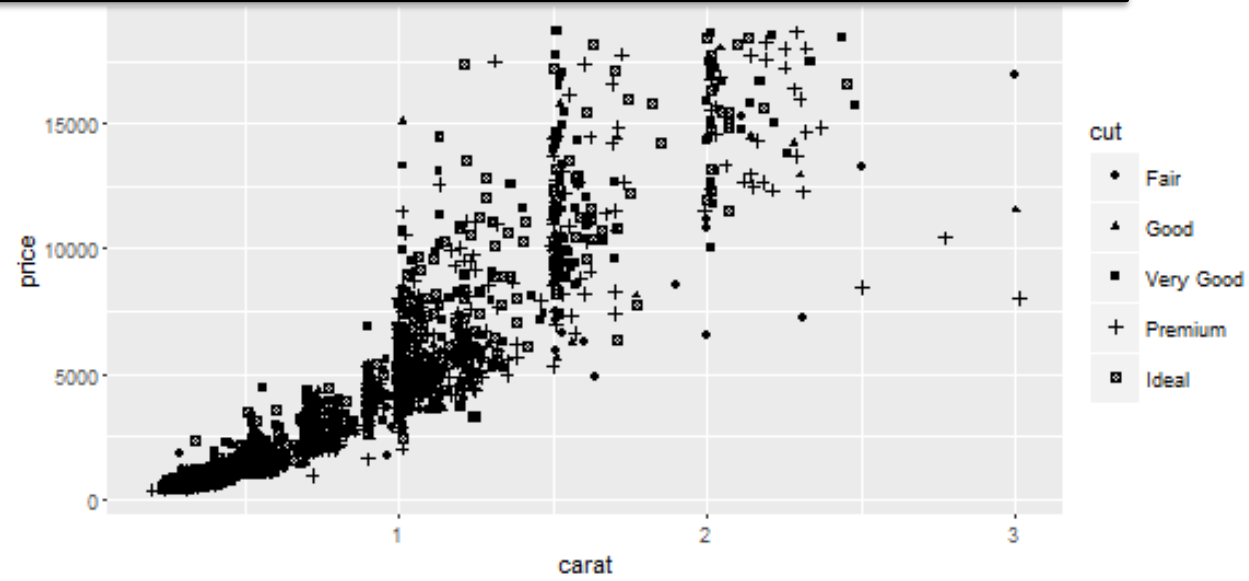
2.2 그래프 – ggplot2 패키지 활용

산점도(scatter) – 몇가지 옵션

R code

```
qplot(carat, price, data = dsmall, colour = color)
qplot(carat, price, data = dsmall, shape = cut)
```





D:plus

2.2 그래프 – ggplot2 패키지 활용

2차원 분석을 위한 geom(geometry) 옵션

geom 옵션	설명
geom = "point"	Scatter plot, 기본값
geom = "smooth"	그래프를 부드럽게 만들어줌
geom = "boxplot"	box-and-whisker plot
geom = "path" / "line"	Line plot

D:plus

2.2 그래프 – ggplot2 패키지 활용

1차원 분석을 위한 geom(geometry) 옵션

구분	geom 옵션	설명
연속형 변수	geom = "histogram"	histogram
	geom = "freqpoly"	a frequency polygon
	geom = "density"	density plot
범주형 변수	geom = "bar"	bar chart

D:plus

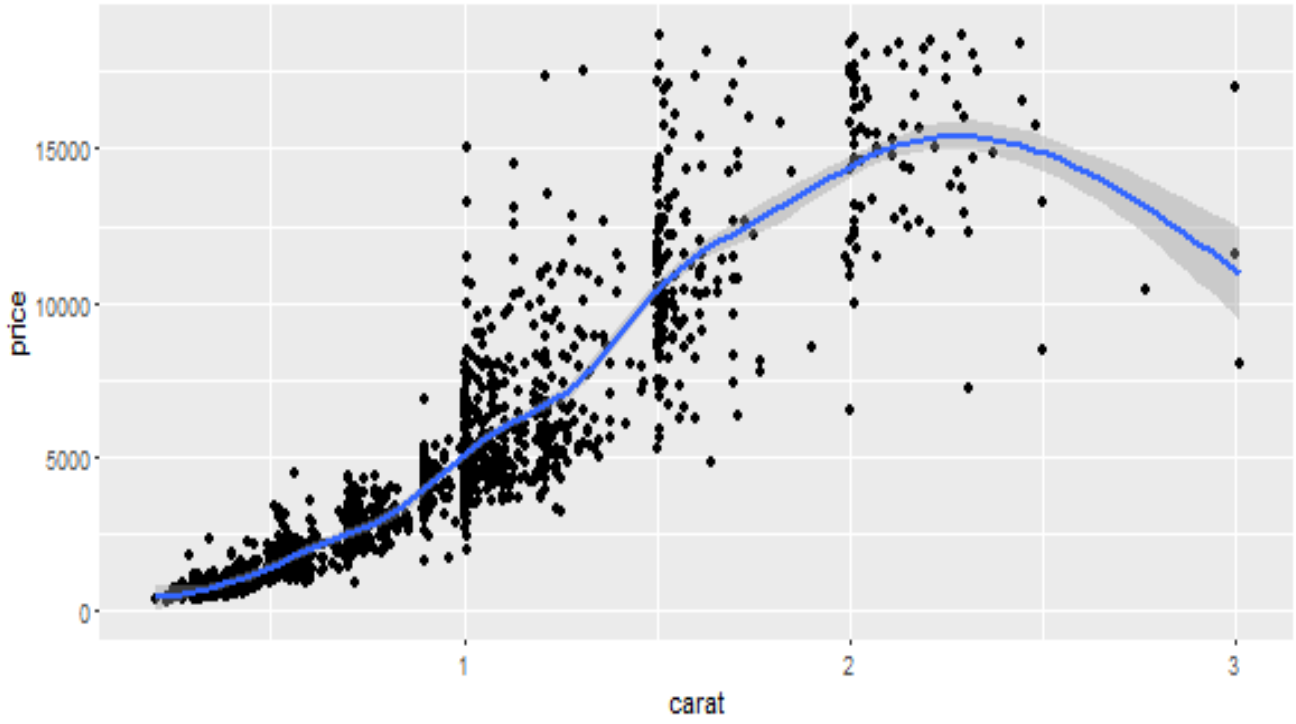
2.2 그래프 – ggplot2 패키지 활용

- Adding a **smoother** to a plot

R code

```
qplot(carat, price, data = dsmall,
geom = c("point", "smooth"))
```

Smoothing method 기본값은 = "loess"

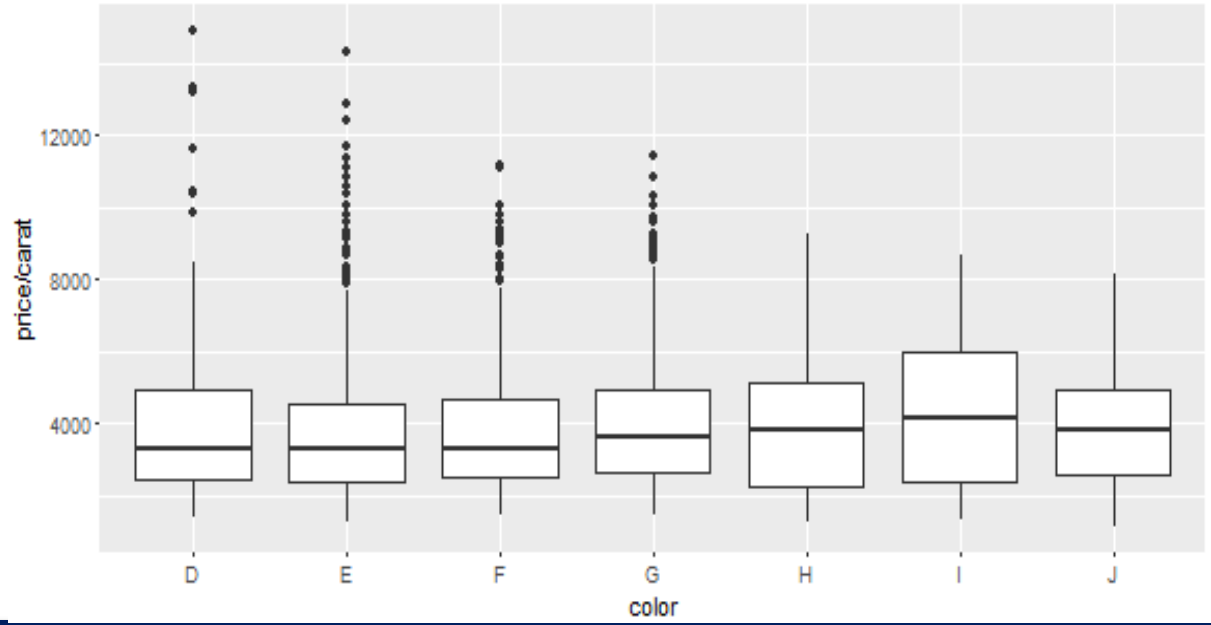


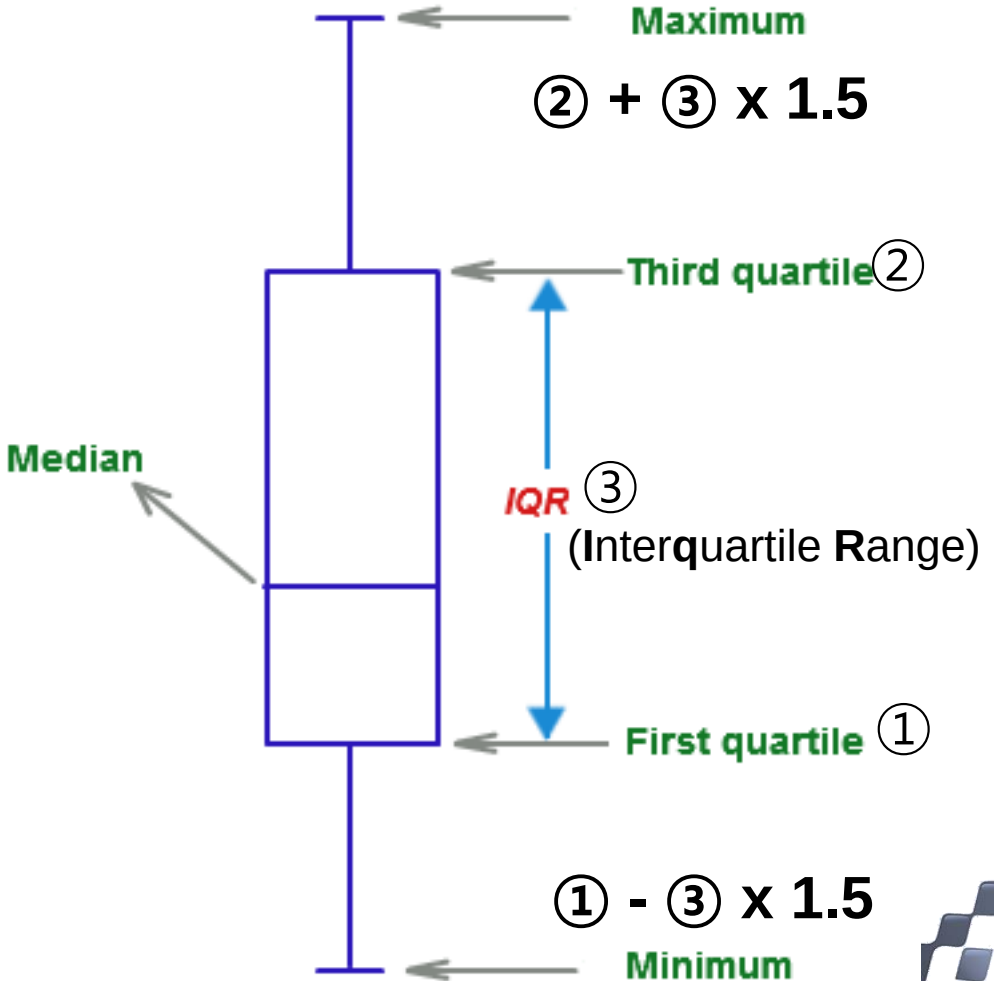
2.2 그래프 – ggplot2 패키지 활용

Boxplot

R code

```
qplot(color, price / carat, data = dsmall
      , geom = "boxplot")
```



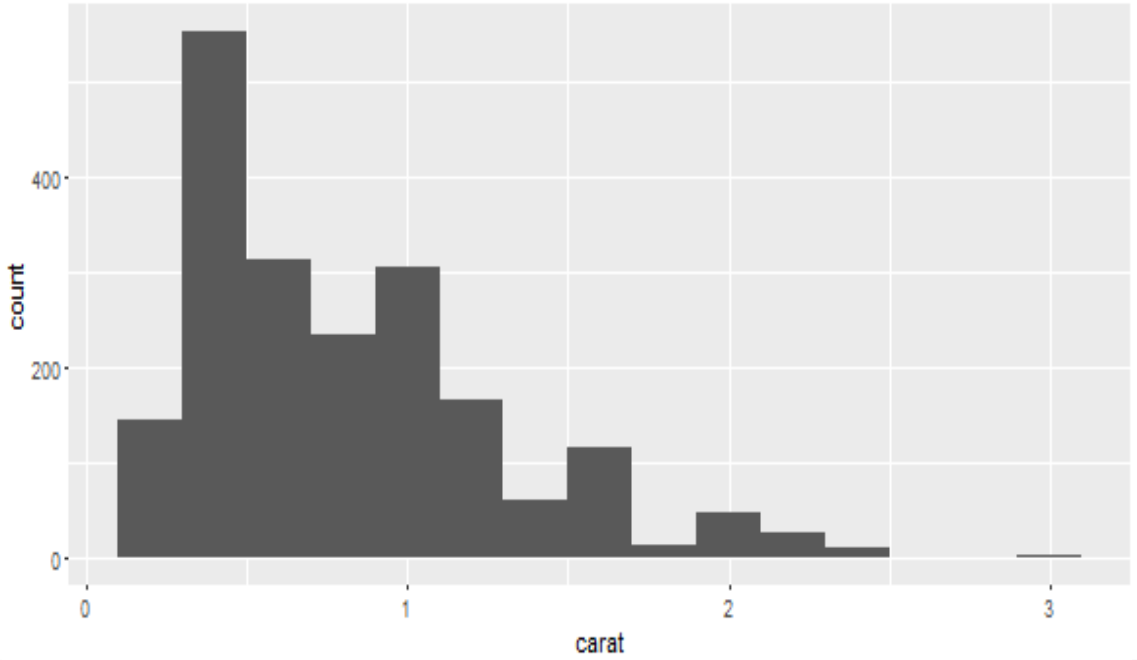


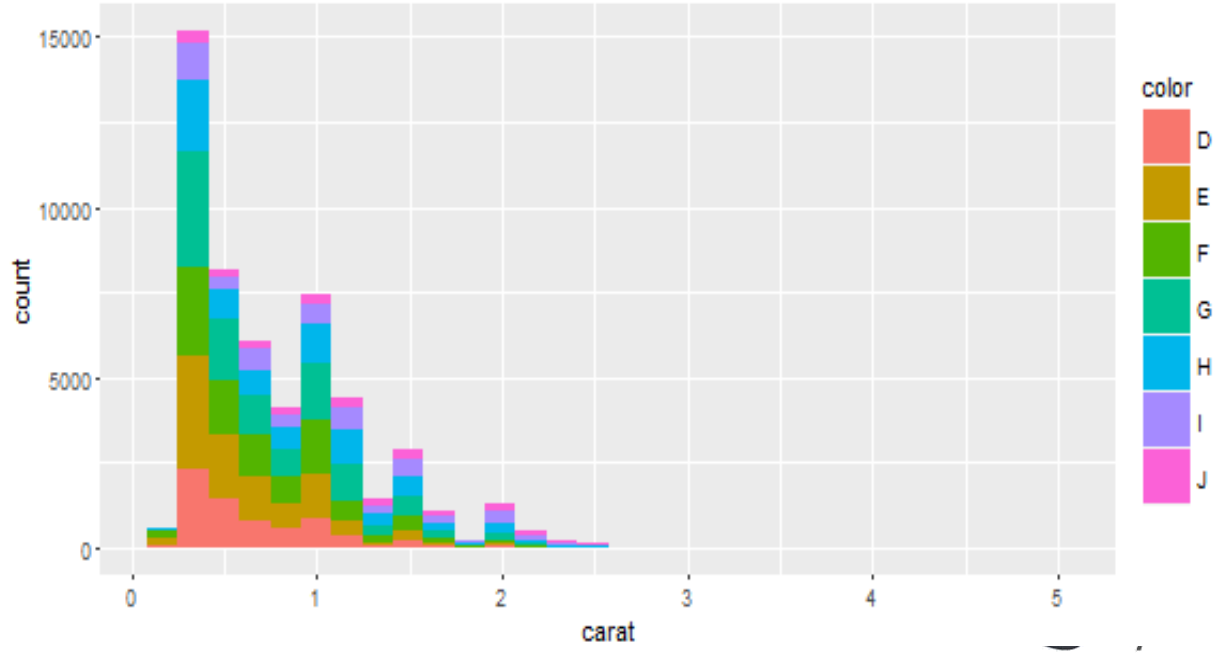
2.2 그래프 – ggplot2 패키지 활용

Histogram

R code

```
qplot(carat, data = dsmall, geom = "histogram“, binwidth = 0.2)
qplot(carat, data = diamonds, geom = "histogram", fill = color)
```



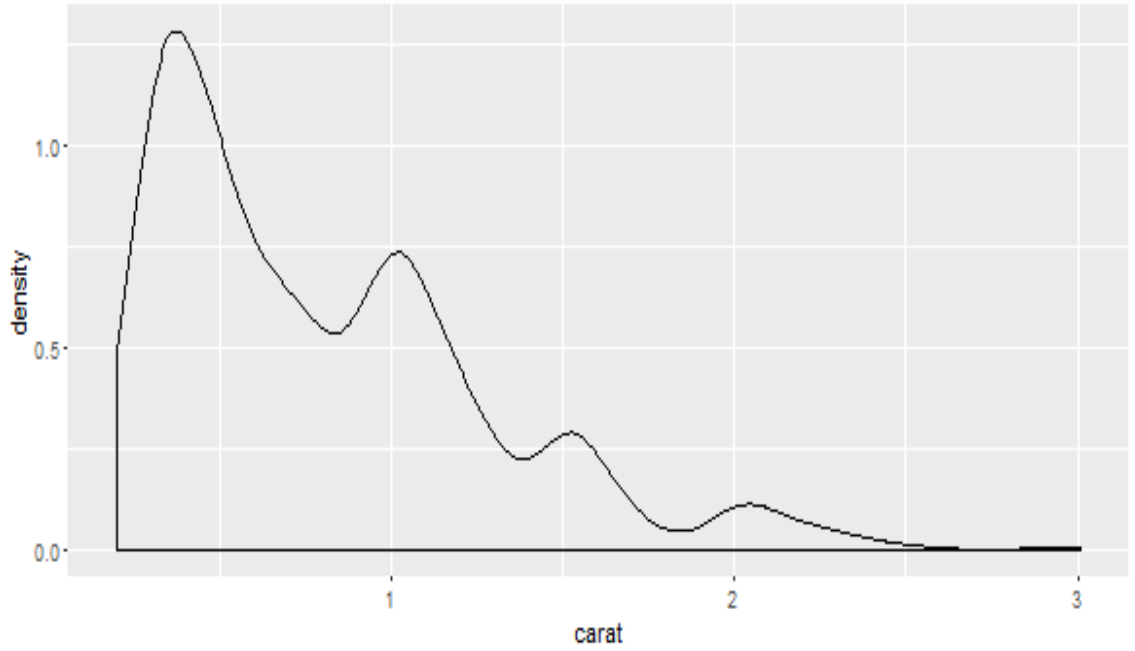


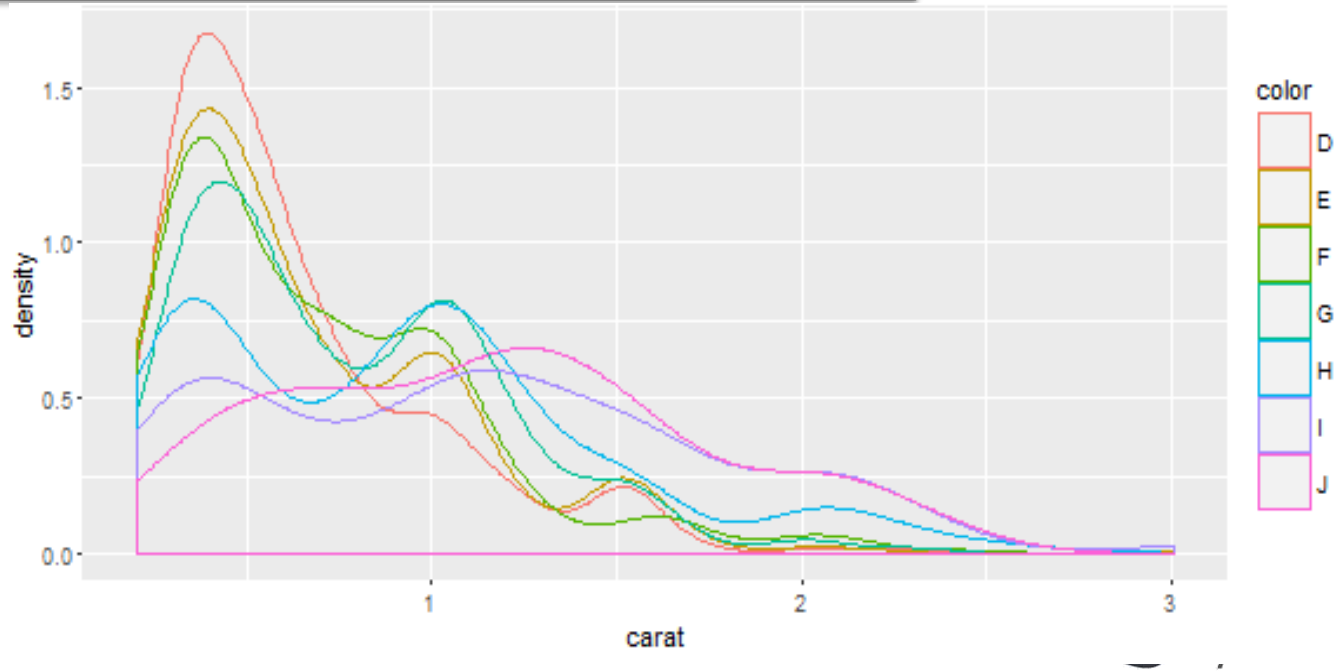
2.2 그래프 – ggplot2 패키지 활용

density plot

R code

```
qplot(carat, data = diamonds, geom = "density")
qplot(carat, data = diamonds, geom = "density", colour = color)
```





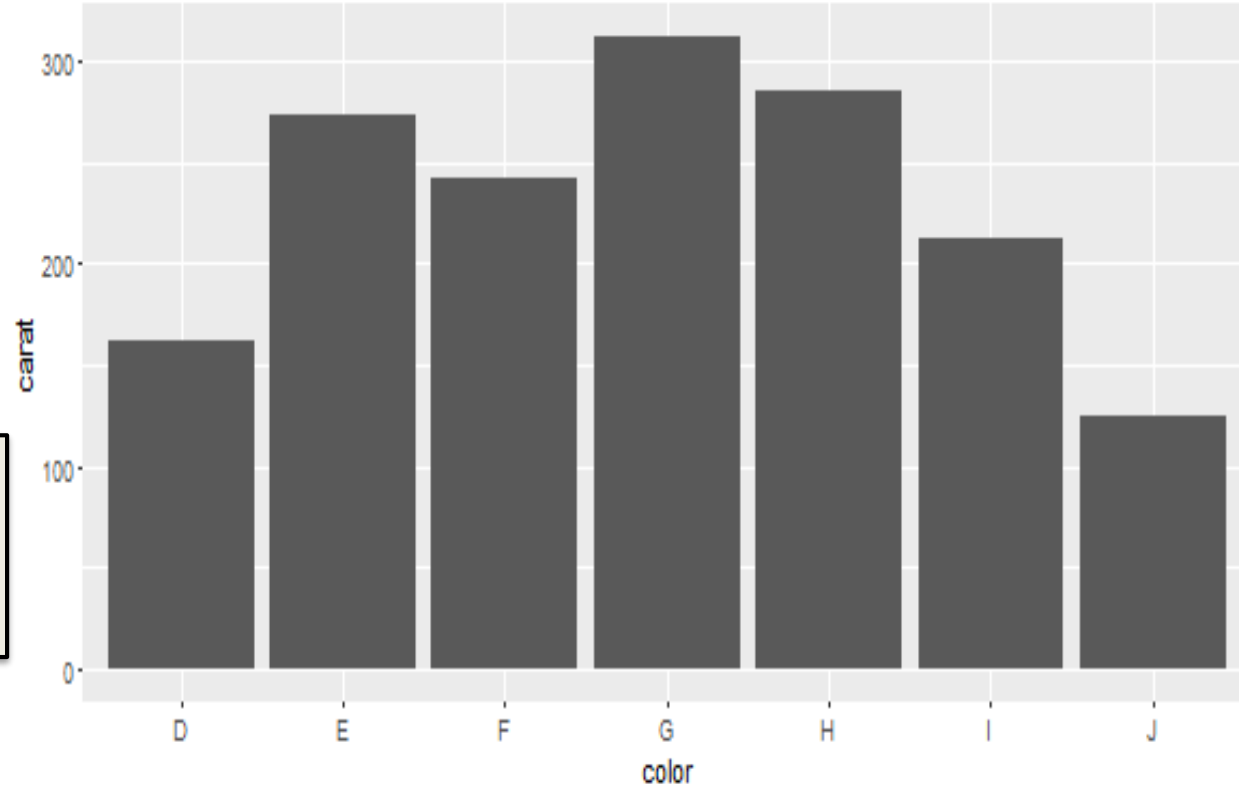
2.2 그래프 – ggplot2 패키지 활용

Bar chart

- 범주형 변수에 대해서 비교하고자 할 때 사용

R code

```
qplot(color, data = dsmall, geom = "bar")
```



D:plus

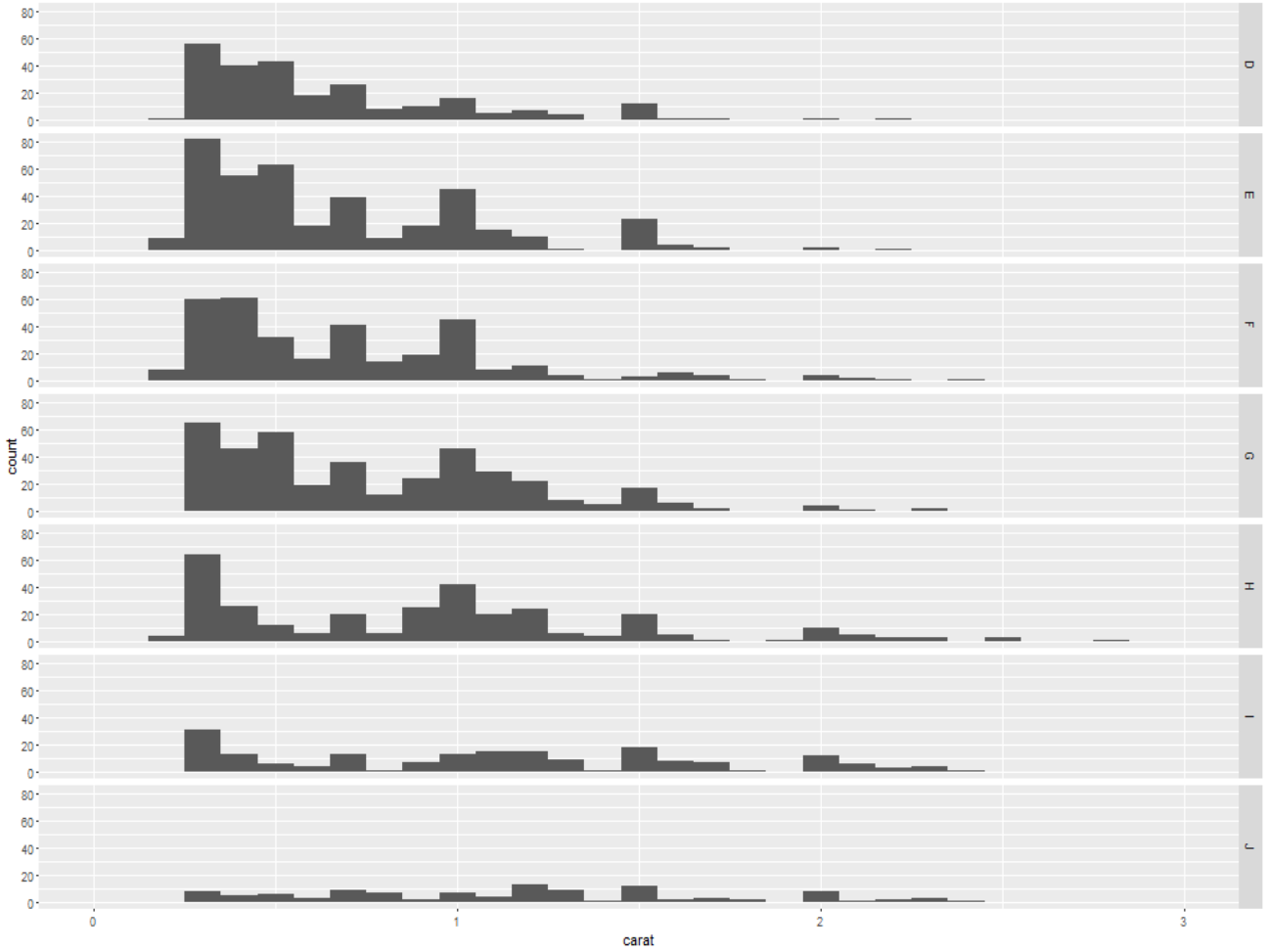
2.2 그래프 – ggplot2 패키지 활용

facet 옵션

- 범주형 변수 별 히스토그램으로 비교할 때.

R code

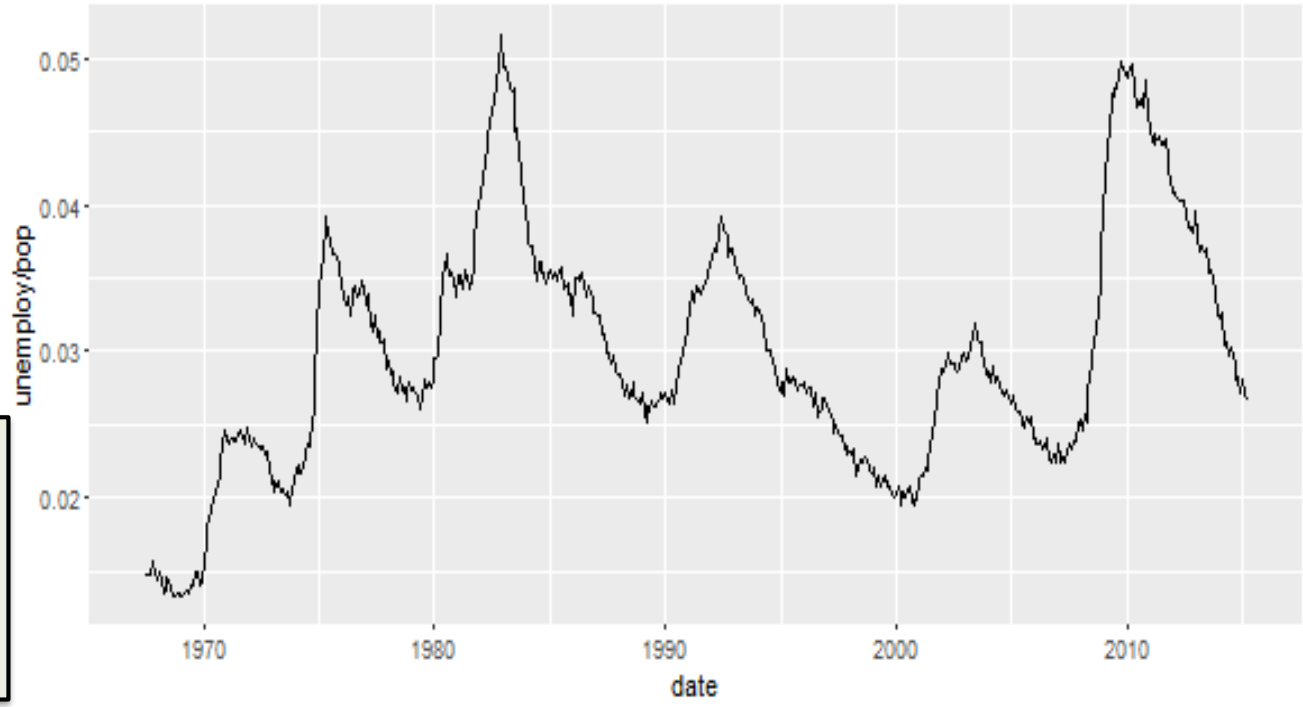
```
qplot(carat, data = dsmall
        , facets = color ~ .
        , geom = "histogram“
    , binwidth = 0.1, xlim = c(0, 3))
```



2.2 그래프 – ggplot2 패키지 활용

Line chart

- 경향성 확인하기 위해
  - –X축은 시간,날짜 축



R code

```
qplot(date, unemploy / pop, data =
economics, geom = "line")
```

실습

D:plus

3. 통계 기초

D:plus

3.1 집단의 대표값

- 집단을 대표하는 수?

평균?

3.1 집단의 대표값

집단을 대표하는 수?

- MIN, MAX, MEAN
- 중위수(median)
- 최빈값(mode)

R code

```
min(cars$dist)
max(cars$dist)
mean(cars$dist)
median(cars$dist)

mode <- function(x) {
    t <- table(x)
    names(t)[which.max(t)]
}
mode(cars$dist)
```

D:plus

3.2 데이터의 분포

- 다음 경기에 내보낼 선수가 한 명 필요합니다.
- 누구를 내보내야 할까요?

Player 1

7	9	10	11	13
1	2	4	2	1

경기당 점수

Player 2

7	8	9	10	11	12	13
1	1	2	2	2	1	1

(빈)도수(frequency)

Player 3

3	6	7	10	11	13	30
2	1	2	3	1	1	1

※ Reference : head first statistics

3.2 데이터의 분포

데이터의 분포를 알아야 한다.

- 평균과의 차이

–분산(Variation) $= \frac{\sum(x - \mu)^2}{n}$

–표준편차(Standard Deviation)

$$= \sqrt{\frac{\sum(x - \mu)^2}{n}}$$

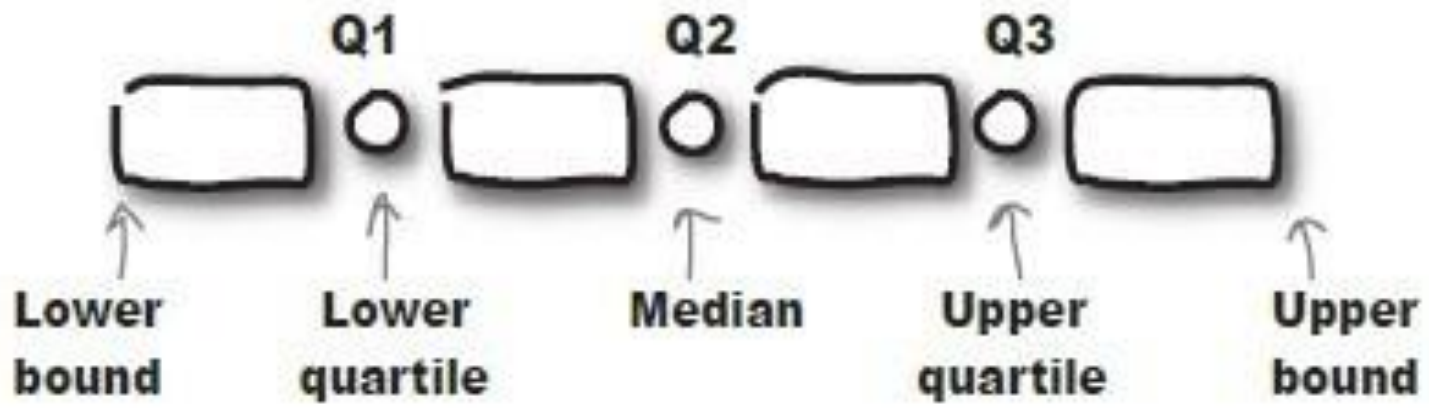
R code

```
p1 <- c(7,9,9,10,10,10,10,11,11,13)
p2 <- c(7,8,9,9,10,10,11,11,12,13)
p3 <- c(3,3,6,7,7,10,10,10,11,13,30)

mean(p1); median(p1)
mean(p2); median(p2)
mean(p3); median(p3)

sd(p1) ; var(p1)
sd(p2) ; var(p2)
sd(p3) ; var(p3)
```

3.2 데이터의 분포

데이터의 분포를 알아야 한다.

- 4분위수(quartile)



R code

```
summary(p1)
summary(p2)
summary(p3)
```

D:plus

실습

D:plus

4. 데이터 탐색으로 분석하기

D:plus

4.1 탐색적 데이터 분석

- **Exploratory Data Analysis**
  - 통계와 그래프를 이용해서 대상 데이터를 파악하는 것.
  - 본격적인 분석에 들어가기 전에 반드시 거쳐야 할 단계

- **EDA를 통해 무엇을 파악해야 하는가?**
  - 각 변수들의 분포(결측치, 이상치 포함)
  - 종속변수와 독립변수의 관계
  - 독립변수들 간의 관계

D:plus

4.2 기초 통계량① : 숫자형 변수들

	변수 명	최소값	1사분위수	중위수	평균	3사분위수	최대값	범위
1								
2								

R code

```
stat_fn <- function(x) {
        c(n = length(x),     na.count = sum(is.na(x))
          , min = min(x, na.rm = T) , qt1st = quantile(x, 0.25,na.rm = T)
          , median = median(x, na.rm = T) , mean = mean(x, na.rm = T)
          , qt3st = quantile(x, 0.75,na.rm = T)  ,      max = max(x, na.rm = T)
          , range = max(x, na.rm = T) - min(x, na.rm = T))
}
```

D:plus

4.2 기초 통계량① : 숫자형 변수들

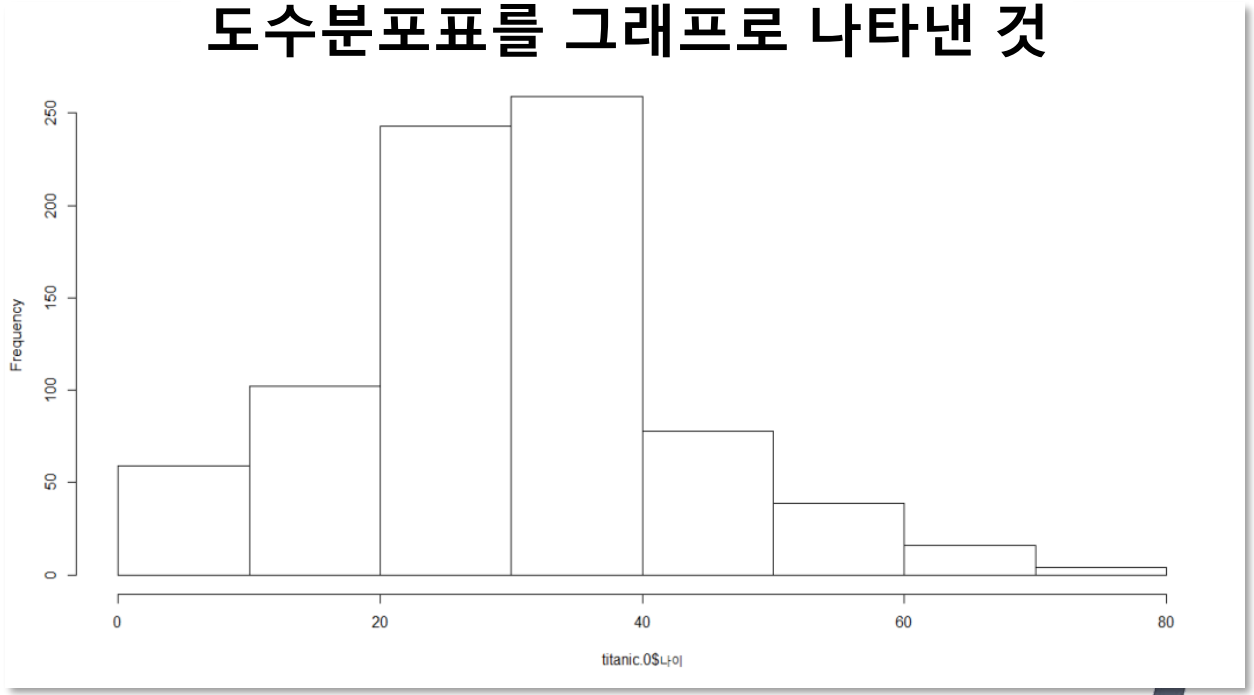
- **값의 분포 확인하기**
  - 밀도함수, 히스토그램

R code

```
hist(titanic.0$Fare)
hist(titanic.0$Age)

plot(density(titanic.0$Fare))
plot(density(titanic.0$Age, na.rm = T))
```

히스토그램

도수분포표를 그래프로 나타낸 것



D:plus

4.2 기초 통계량② : 범주형 변수들

- table, prop.table 함수 사용하여 값 찾기.

R code

```
table(titanic.0$Sex)
table(titanic.0$Sex,titanic.0$Survived)
prop.table(table(titanic.0$Sex,titanic.0$Survived))
```

4.2 기초 통계량② : 범주형 변수들

- 분포 비교 그래프

R code

```
install.packages("mosaic")
library(mosaic)
mosaicplot(Sex ~ Survived + Pclass, data = titanic.0, color = TRUE)
```

4.3 결측치와 이상치(Outlier)

- 결측치 찾기

R code

```
colSums(is.na(titanic.0))

library(Amelia)
missmap(titanic.0, col=c("yellow", "black"), legend=FALSE)
```

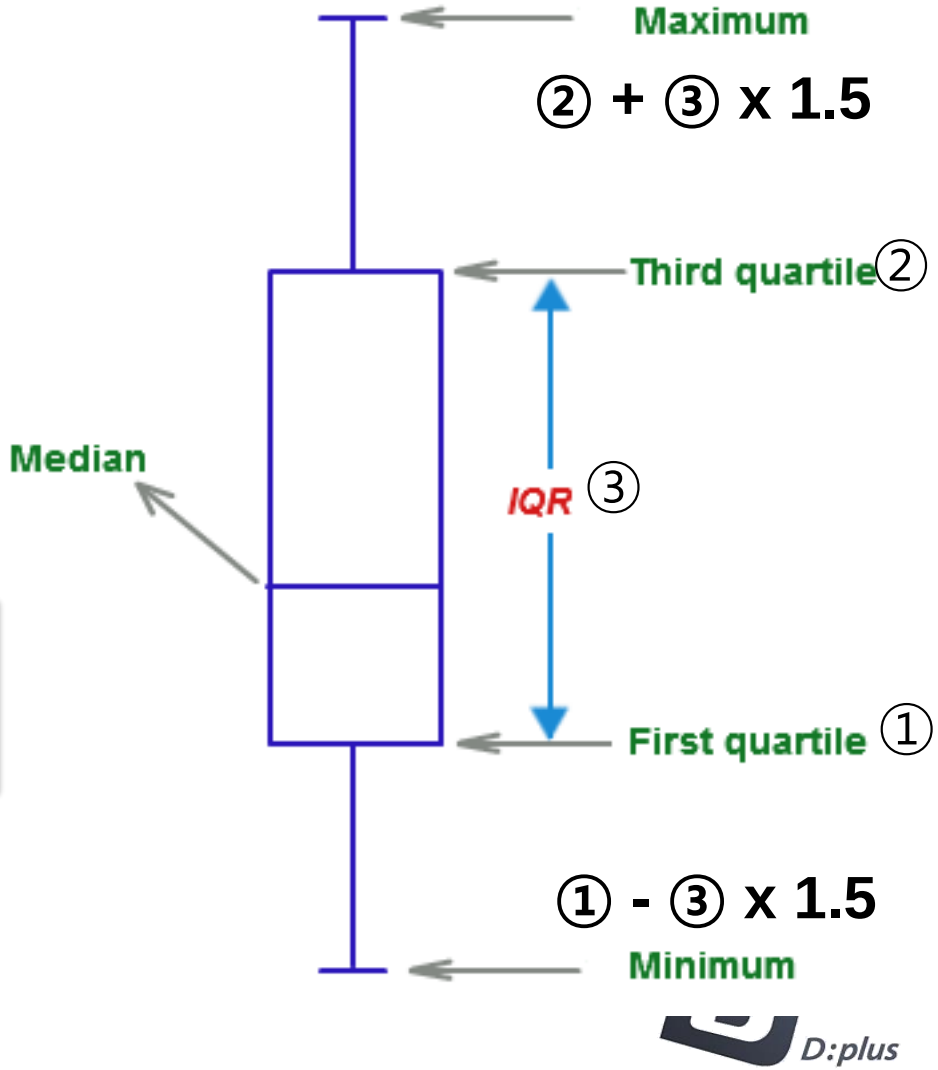
4.3 결측치와 이상치(Outlier)

- **이상치**
  - 다른 데이터들과 비교하여 유달리 높거나 낮은 값을 보이는 것

- **Boxplot으로 이상치 찾기**

R code

```
boxplot(titanic.0$Age)
```



4.4 결측치와 이상치 데이터 다루기

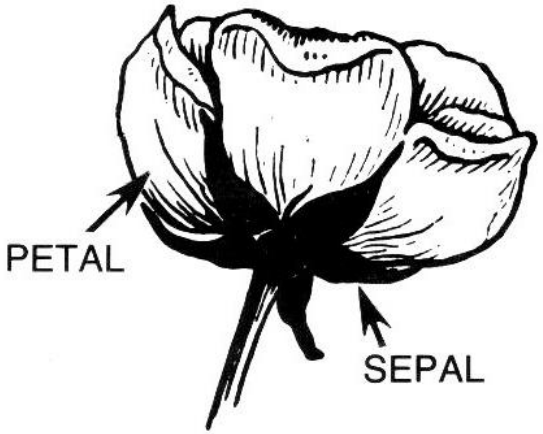
- 데이터 분석 전에 반드시 결측치와 이상치를 처리해줘야 한다.

구분	제거	대체
이상치	제거는 권장하지 않음. 특히 자료가 많지 않은 경우	▪ 자료의 하한/ 상한 값으로 대체
결측치		▪ 시계열 데이터 : 같은(비슷한) 시기의 데이터 ▪ 최빈값/ 평균값으로 대체

D:plus

4.5 상관 분석(Correlation)

- 두 변수 간의 선형관계가 존재하는지 판단하기 위한 분석 방법
- 상관계수(Correlation coefficient)
  - 상관관계의 정도를 나타내는 단위 : -1 ~ 1.
    - 1에 가까울 수록 : 양의 상관성이 높다.
    - -1에 가까울 수록 : 음의 상관성이 높다.



R code

```
iris[1:3,1:2]
cor(iris[,1:4])
pairs(iris[,1:4])
plot(iris$Petal.Length, iris$Petal.Width)
qplot(Petal.Length, Petal.Width, data = iris, geom = c("point", "smooth"))
```

D:plus

실습

D:plus