04 데이터 다루기(2) ¶

학습내용

- 데이터 탐색해 보기(head, tail, summary, dim, class, str)
- 변수명 바꿔보기
- 파생변수 만들기

내장 데이터 셋 불러오기

In [1]:

```
data("mtcars")
```

함수	설명
dim()	데이터 셋 객체의 차원을 보기(행,열등)
head()	데이터의 앞에서부터 몇행, 상위6개
tail()	데이터의 뒤에서부터 몇행, 하위6개
str()	데이터 구조, 변수 개수, 변수 명, 관찰치 개수, 관찰치
names() or colnames()	데이터 객체의 컬럼명
class()	데이터 객체의
summary()	요약값
View()	뷰어에서 확인

In [2]:

```
dim(mtcars)
```

32 11

In [3]:

```
head(mtcars)
head(mtcars, 7)
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360	245	3.21	3.570	15.84	0	0	3	4

In [4]:

```
tail(mtcars)
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Porsche 914-2	26.0	4	120.3	91	4.43	2.140	16.7	0	1	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.9	1	1	5	2
Ford Pantera L	15.8	8	351.0	264	4.22	3.170	14.5	0	1	5	4
Ferrari Dino	19.7	6	145.0	175	3.62	2.770	15.5	0	1	5	6
Maserati Bora	15.0	8	301.0	335	3.54	3.570	14.6	0	1	5	8
Volvo 142E	21.4	4	121.0	109	4.11	2.780	18.6	1	1	4	2

In [5]:

```
str(mtcars)
```

```
'data.frame':	32 obs. of  11 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num  16.5 17 18.6 19.4 17 ...
 $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
 $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
 $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

In [6]:

```
names(mtcars)
```

'mpg' 'cyl' 'disp' 'hp' 'drat' 'wt' 'qsec' 'vs' 'am' 'gear' 'carb'

In [7]:

```
class(mtcars)
```

'data.frame'

In [8]:

```
summary(mtcars)
```

```
      mpg             cyl             disp             hp       
 Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
 1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
 Median :19.20   Median :6.000   Median :196.3   Median :123.0  
 Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
 3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
 Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
      drat             wt             qsec             vs        
 Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000  
 1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000  
 Median :3.695   Median :3.325   Median :17.71   Median :0.0000  
 Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375  
 3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000  
 Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000  
       am              gear            carb      
 Min.   :0.0000   Min.   :3.000   Min.   :1.000  
 1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
 Median :0.0000   Median :4.000   Median :2.000  
 Mean   :0.4062   Mean   :3.688   Mean   :2.812  
 3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000  
 Max.   :1.0000   Max.   :5.000   Max.   :8.000  
```

In [9]:

```
# View() : R studio에서 확인 가능
```

(ex) 4-1 실습 해보기

- mpg 데이터 셋에 대한 탐색을 해 보기

dply 패키지 사용해 보기

In [10]:

```
install.packages("dplyr")
```

```
Warning message:
"unable to access index for repository http://www.stats.ox.ac.uk/pub/RWin/bin/window
s/contrib/3.5: (http://www.stats.ox.ac.uk/pub/RWin/bin/windows/contrib/3.5:)
  URL 'http://www.stats.ox.ac.uk/pub/RWin/bin/windows/contrib/3.5/PACKAGES'를 열 수
 없습니다"

package 'dplyr' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:₩Users₩WITHJS₩AppData₩Local₩Temp₩RtmpQ9HUtg₩downloaded_packages
```

In [11]:

```
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

변수명 바꾸기

In [12]:

```
df_new <- mtcars
```

In [13]:

```
colnames(df_new)
```

'mpg' 'cyl' 'disp' 'hp' 'drat' 'wt' 'qsec' 'vs' 'am' 'gear' 'carb'

In [14]:

```
?mtcars
```

In [15]:

```
df_new <- rename(df_new, weight=wt)
names(df_new)
```

'mpg' 'cyl' 'disp' 'hp' 'drat' 'weight' 'qsec' 'vs' 'am' 'gear' 'carb'

(ex) 4-2 mpg 데이터 셋을 불러오기

- cty는 도시의 연비
- hwy는 고속도로 연비를 의미
- cty -> city로
- hwy -> hightway로 바꾸어 보자.

파생변수(derived Variable)

In [16]:

```
df <- data.frame(var1 = c(1,3,5), var2=c(2,4,6))
df
```

var1	var2
1	2
3	4
5	6

In [17]:

```
df$sum <- df$var1 + df$var2
df$sum
```

3 7 11

In [18]:

```
df
```

var1	var2	sum
1	2	3
3	4	7
5	6	11

(해보기) 평균 변수 추가하기

조건문을 활용한 파생변수 만들기

In [9]:

```
head(ggplot2::mpg,10)
```

manufacturer	model	displ	year	cyl	trans	drv	cty	hwy	fl	class
audi	a4	1.8	1999	4	auto(l5)	f	18	29	p	compact
audi	a4	1.8	1999	4	manual(m5)	f	21	29	p	compact
audi	a4	2.0	2008	4	manual(m6)	f	20	31	p	compact
audi	a4	2.0	2008	4	auto(av)	f	21	30	p	compact
audi	a4	2.8	1999	6	auto(l5)	f	16	26	p	compact
audi	a4	2.8	1999	6	manual(m5)	f	18	26	p	compact
audi	a4	3.1	2008	6	auto(av)	f	18	27	p	compact
audi	a4 quattro	1.8	1999	4	manual(m5)	4	18	26	p	compact
audi	a4 quattro	1.8	1999	4	auto(l5)	4	16	25	p	compact
audi	a4 quattro	2.0	2008	4	manual(m6)	4	20	28	p	compact

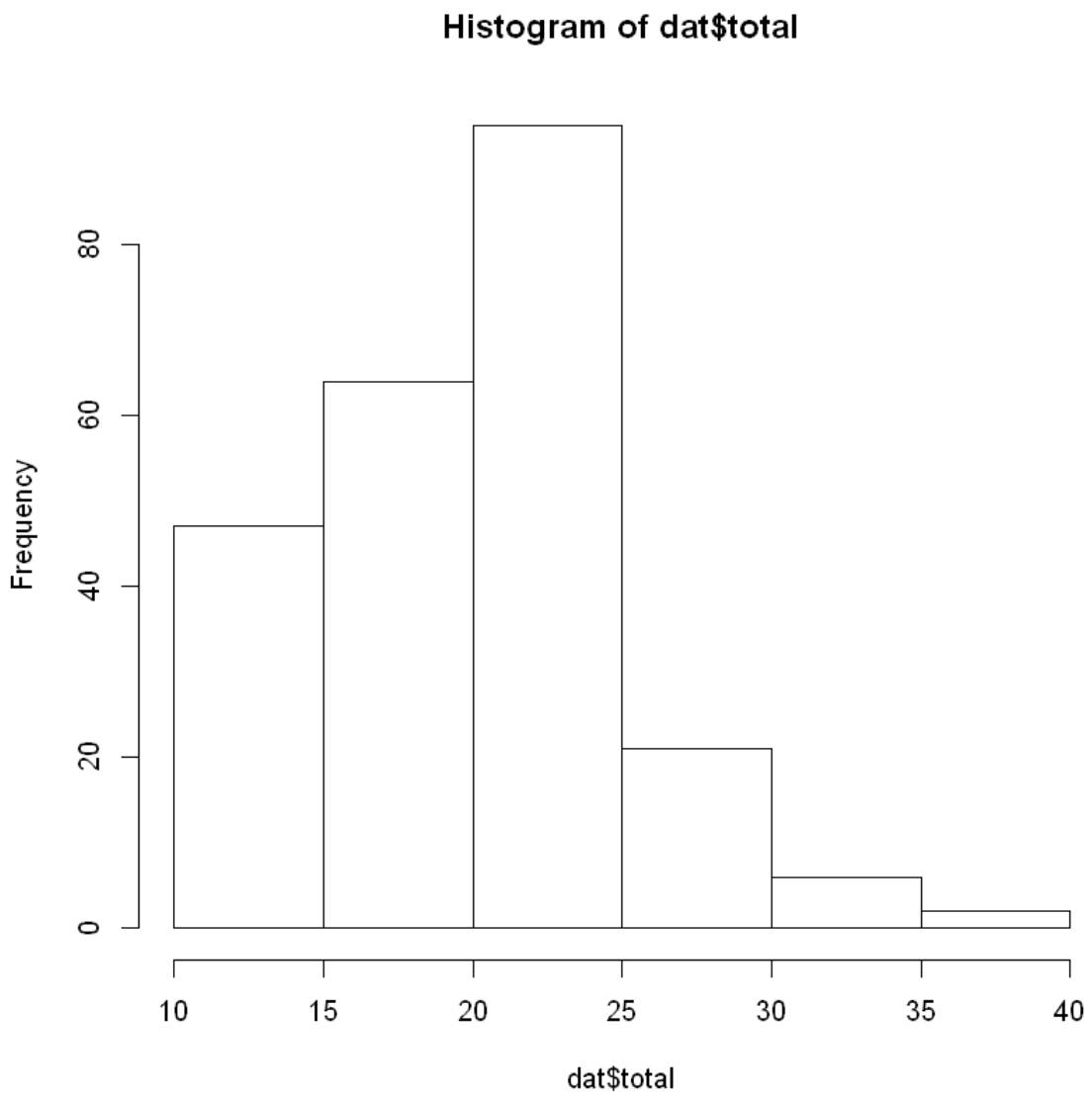
In [20]:

```
dat <- ggplot2::mpg
dat$total <- (dat$cty + dat$hwy) /2  # 통합연비 생성
head(dat)
```

manufacturer	model	displ	year	cyl	trans	drv	cty	hwy	fl	class	total
audi	a4	1.8	1999	4	auto(l5)	f	18	29	p	compact	23.5
audi	a4	1.8	1999	4	manual(m5)	f	21	29	p	compact	25.0
audi	a4	2.0	2008	4	manual(m6)	f	20	31	p	compact	25.5
audi	a4	2.0	2008	4	auto(av)	f	21	30	p	compact	25.5
audi	a4	2.8	1999	6	auto(l5)	f	16	26	p	compact	21.0
audi	a4	2.8	1999	6	manual(m5)	f	18	26	p	compact	22.0

In [21]:

```
hist(dat$total)
```



In [22]:

```
summary(dat$total)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  10.50   15.50   20.50   20.15   23.50   39.50 
```

- total 연비의 평균과 중앙값은 약 20
- total연비가 20~25사이의 해당하는 자동차 모델이 많다.
- 대부분 25이하, 25를 넘기는 자동차는 많지 않음.

ifelse()

- ifelse(조건문, 참일때, 거짓일때)
- (ex) ifelse(dat$total >= 20, "pass", "fail")

In [23]:

```
ifelse(dat$total >= 20, "pass", "fail")
```

'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass'
'pass' 'pass' 'pass' 'fail' 'pass' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail' 'pass' 'fail' 'pass'
'pass' 'fail' 'fail' 'fail' 'fail' 'fail' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass'
'fail' 'fail' 'pass' 'pass' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail'
'fail' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail'
'fail' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail'
'fail' 'fail' 'fail' 'fail' 'pass' 'pass' 'pass' 'pass' 'fail' 'fail' 'fail' 'fail' 'fail' 'pass'
'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass'
'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'fail' 'fail' 'fail'
'fail' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail'
'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'fail' 'fail' 'fail' 'fail'
'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass'
'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail'
'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass'
'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail' 'fail'
'fail' 'fail' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass'
'pass' 'pass' 'fail' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass'
'pass' 'pass' 'pass' 'pass' 'pass'

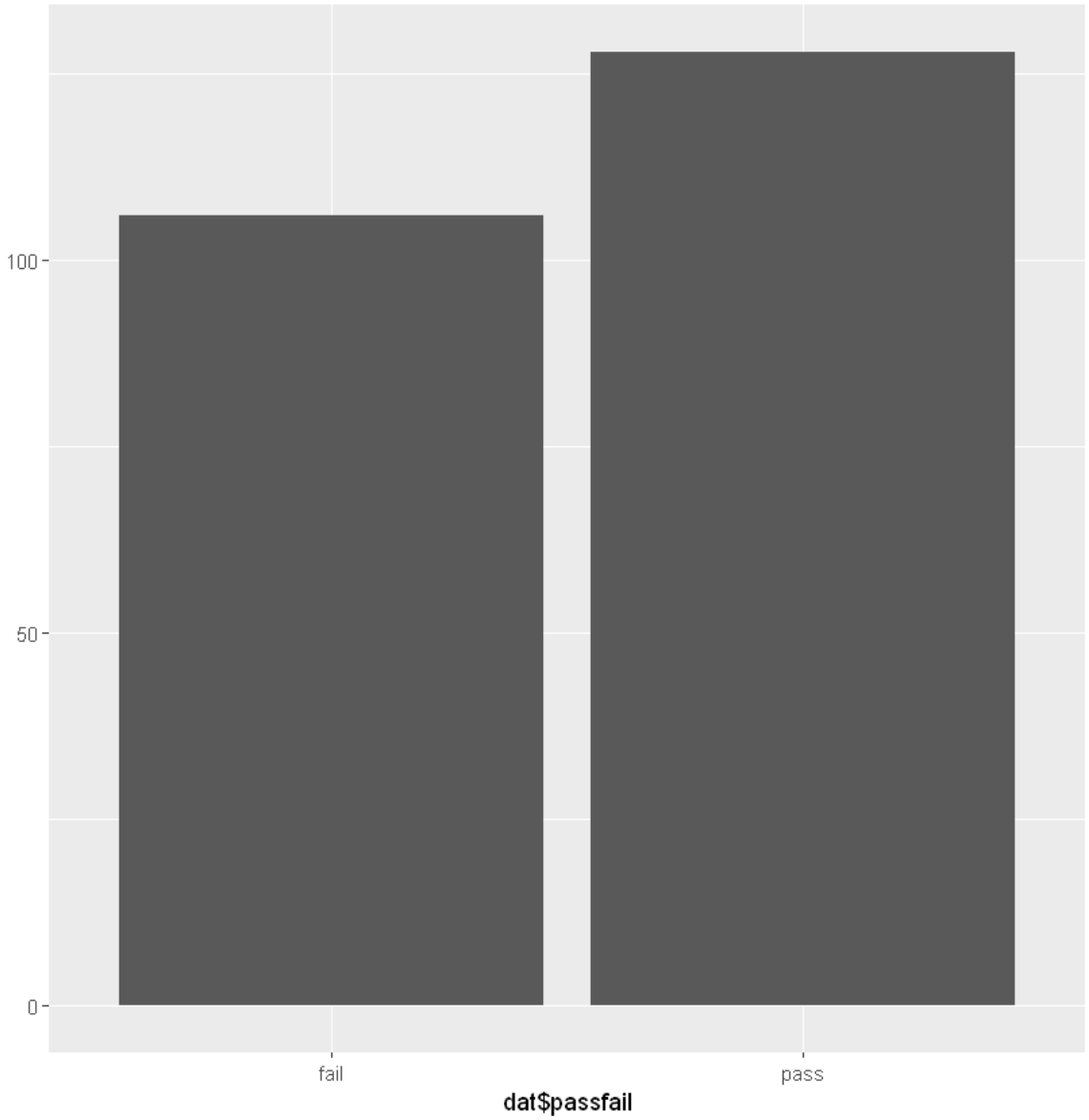
In [24]:

```
### 위의 내용을 갖는 변수 추가
dat$passfail <- ifelse(dat$total >= 20, "pass", "fail")
head(dat$passfail,15)
```

'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass' 'pass'
'pass' 'pass' 'pass'

In [25]:

```
library(ggplot2)
qplot(dat$passfail)
```



실습과제 4-3

- total을 이용하여 A, B, C 등급 부여하기

In [26]:

```
head(dat)
```

manufacturer	model	displ	year	cyl	trans	drv	cty	hwy	fl	class	total	passfail
audi	a4	1.8	1999	4	auto(l5)	f	18	29	p	compact	23.5	pass
audi	a4	1.8	1999	4	manual(m5)	f	21	29	p	compact	25.0	pass
audi	a4	2.0	2008	4	manual(m6)	f	20	31	p	compact	25.5	pass
audi	a4	2.0	2008	4	auto(av)	f	21	30	p	compact	25.5	pass
audi	a4	2.8	1999	6	auto(l5)	f	16	26	p	compact	21.0	pass
audi	a4	2.8	1999	6	manual(m5)	f	18	26	p	compact	22.0	pass

In [27]:

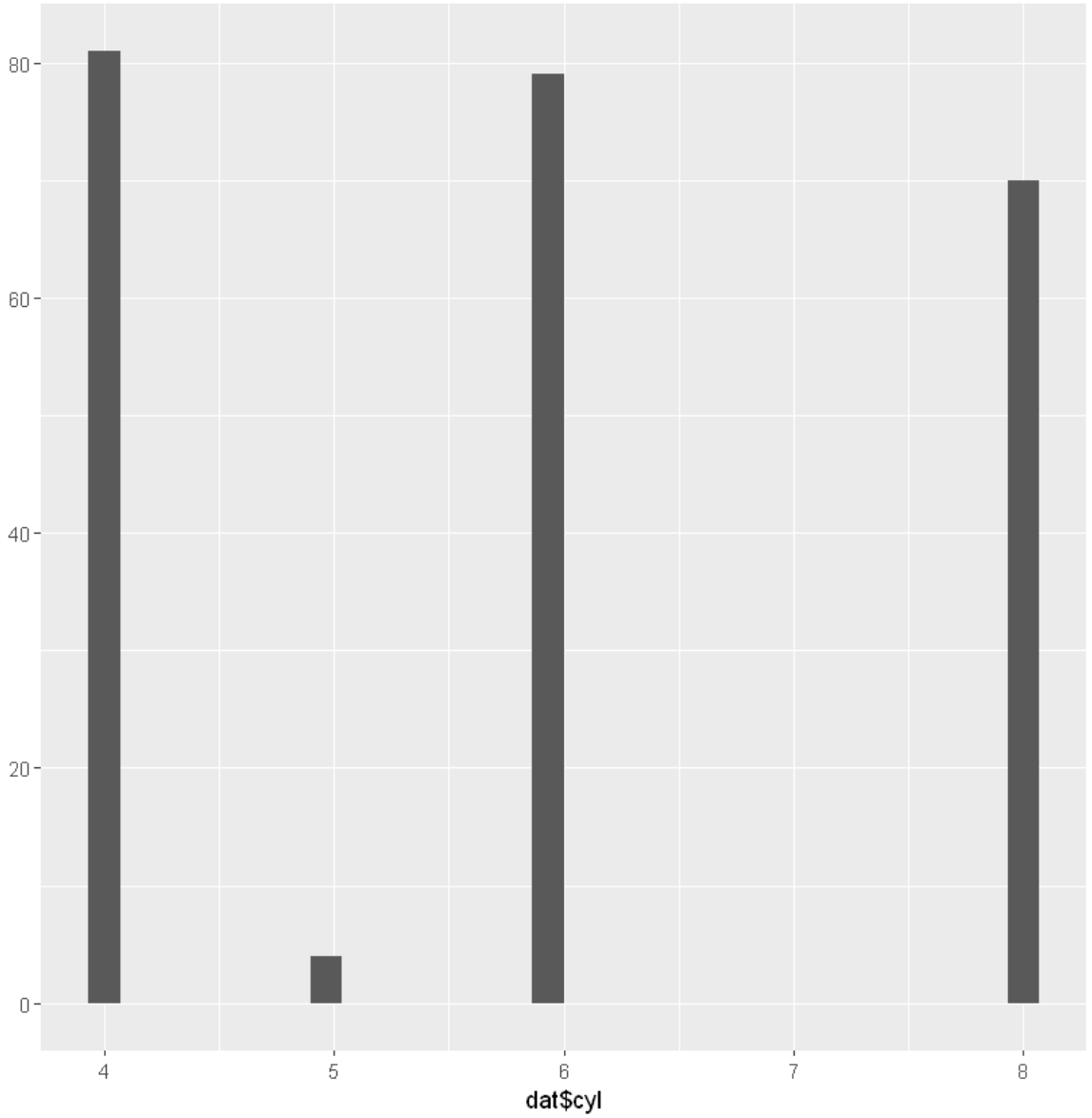
```
table(dat$cyl)
```

```
 4  5  6  8 
81  4 79 70 
```

In [28]:

```
qplot(dat$cyl)
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



실습과제 4-4 (p123)

- ggplot2 패키지의 미국 동북중부 437개 지역의 인구 통계 정보를 담은 midwest를 데이터 셋을 이용하여 분석 문제 해결해 보기