07 데이터 시각화(2)

학습 내용

- ggplot2 패키지 알아보기
- geom_col(), geom_bar()에 대해 알아본다. (막대 그래프 그리기)
- table() 함수에 대해 알아본다. (각 범주 빈도확인)
- gridExtra 패키지 gridExtra::grid.arrange() 함수에 대해 알아본다.

In [1]:

```
library(dplyr)
titanic <- read.csv("D:WWdatasetWWtitanic_dataWWtr_mod.csv")
head(titanic,10)
```

```
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
```

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabi
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.2500	N
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C8
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.9250	N
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1000	C12
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.0500	N
6	0	3	Moran, Mr. James	male	29	0	0	330877	8.4583	N
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E4
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.0750	N
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333	N
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708	N

7-1 막대 그래프(geom_col)

- 막대 그래프(barplot)은 범주형 변수와 수치형 변수와의 관계를 보여주기 위해 사용되어진다.
- ggplot에서 geom_col() 함수를 사용한다.

In [2]:

```
library(ggplot2)
library(dplyr)
```

In [3]:

```
head(mpg,5)
```

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

어떤 모델이 있을까?

In [4]:

```
table(mpg$model)
```

```
          4runner 4wd                    a4            a4 quattro
                    6                     7                     8
           a6 quattro                altima     c1500 suburban 2wd
                    3                     6                     5
                camry           camry solara           caravan 2wd
                    7                     7                    11
                civic                corolla              corvette
                    9                     5                     5
     dakota pickup 4wd           durango 4wd        expedition 2wd
                    9                     7                     3
         explorer 4wd        f150 pickup 4wd          forester awd
                    6                     7                     6
    grand cherokee 4wd             grand prix                   gti
                    8                     5                     5
          impreza awd                   jetta        k1500 tahoe 4wd
                    8                     9                     4
land cruiser wagon 4wd                 malibu                maxima
                    2                     5                     3
       mountaineer 4wd                mustang         navigator 2wd
                    4                     9                     3
            new beetle                 passat         pathfinder 4wd
                    6                     7                     4
   ram 1500 pickup 4wd           range rover                sonata
                   10                     4                     7
              tiburon      toyota tacoma 4wd
                    7                     7
```

어떤 trans가 있을까?

In [5]:

```
table(mpg$trans)
```

```
  auto(av)   auto(l3)   auto(l4)   auto(l5)   auto(l6)   auto(s4)   auto(s5)
         5          2         83         39          6          3          3
auto(s6) manual(m5) manual(m6)
        16         58         19
```

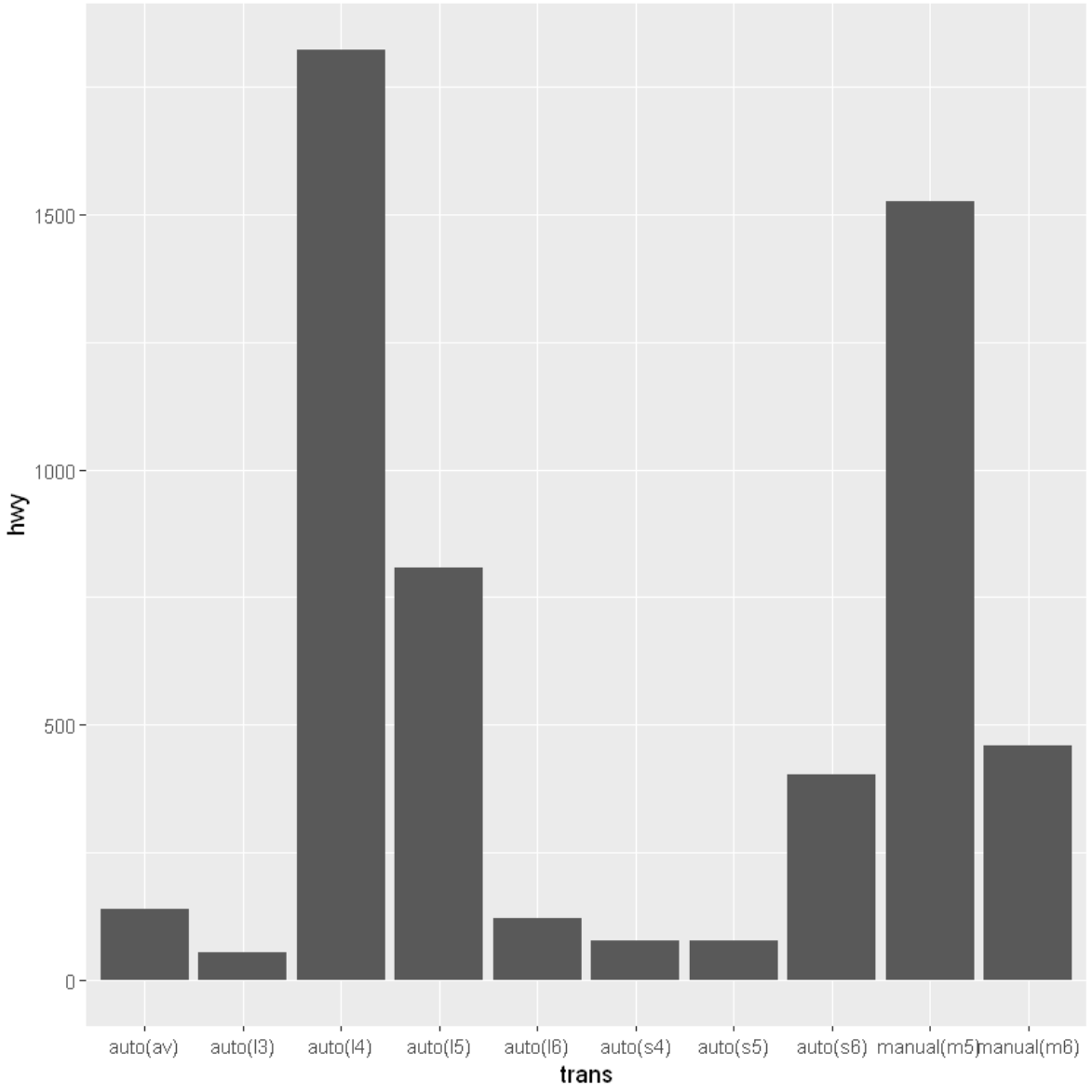
mpg데이터 셋을 이용

- drv(구동방식)별 평균(hwy) 고속도로 연비 막대그래프 그리기
- trans : type of transmission(트랜스미션 타입)
- hwy : 고속도로 연비
- cty : 도시에서의 연비
- fl : 연료타입

transmission의 타입별 고속도로 연비(hwy) 알아보기

In [6]:

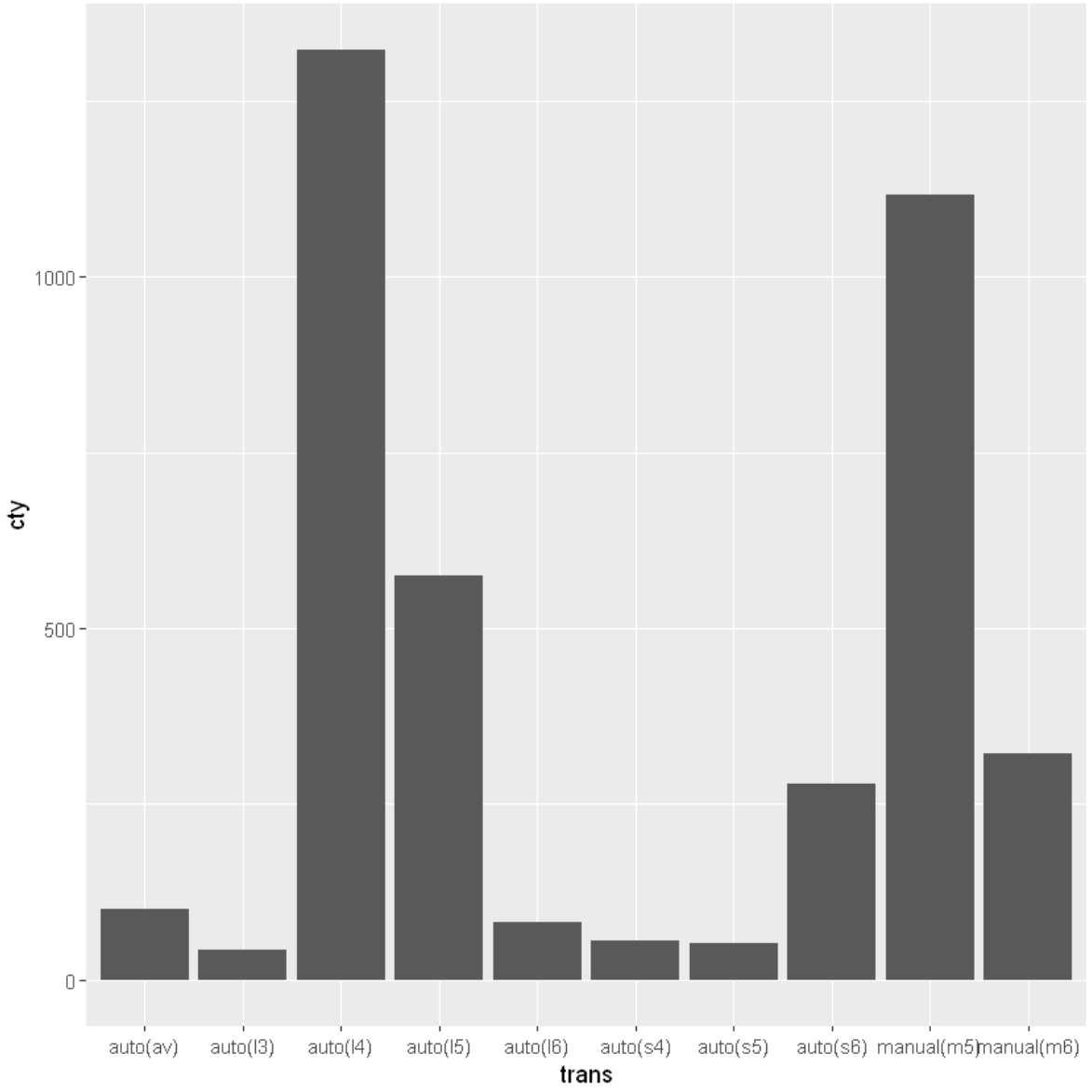
```
ggplot(data=mpg, aes(x=trans, y=hwy)) + geom_col()
```



transmission의 타입별 도시에서의 연비(cty) 알아보기

In [7]:

```
ggplot(data=mpg, aes(x=trans, y=cty)) + geom_col()
```



(실습해보기) 07-01

- Titanic 데이터 셋을 이용하여 PClass별 알아보기 생존자 알아보기

집단별 평균표 만들기

- drv(구동방식별), hwy(고속도로 연비)

In [9]:

```
drv_hwy <- mpg %>%
       group_by(drv) %>%
       summarise(mean_hwy = mean(hwy))
drv_hwy
```

drv	mean_hwy
4	19.17476
f	28.16038
r	21.00000

데이터 시각화

- 구동방식별 평균 고속도로 연비

In [10]:

```
ggplot(data = drv_meanhwy, aes(x=drv, y=mean_hwy)) + geom_col()
```

```
Error in ggplot(data = drv_meanhwy, aes(x = drv, y = mean_hwy)): 객체 'drv_meanhw
y'를 찾을 수 없습니다
Traceback:

1. ggplot(data = drv_meanhwy, aes(x = drv, y = mean_hwy))
```

(실습해보기) 07-02

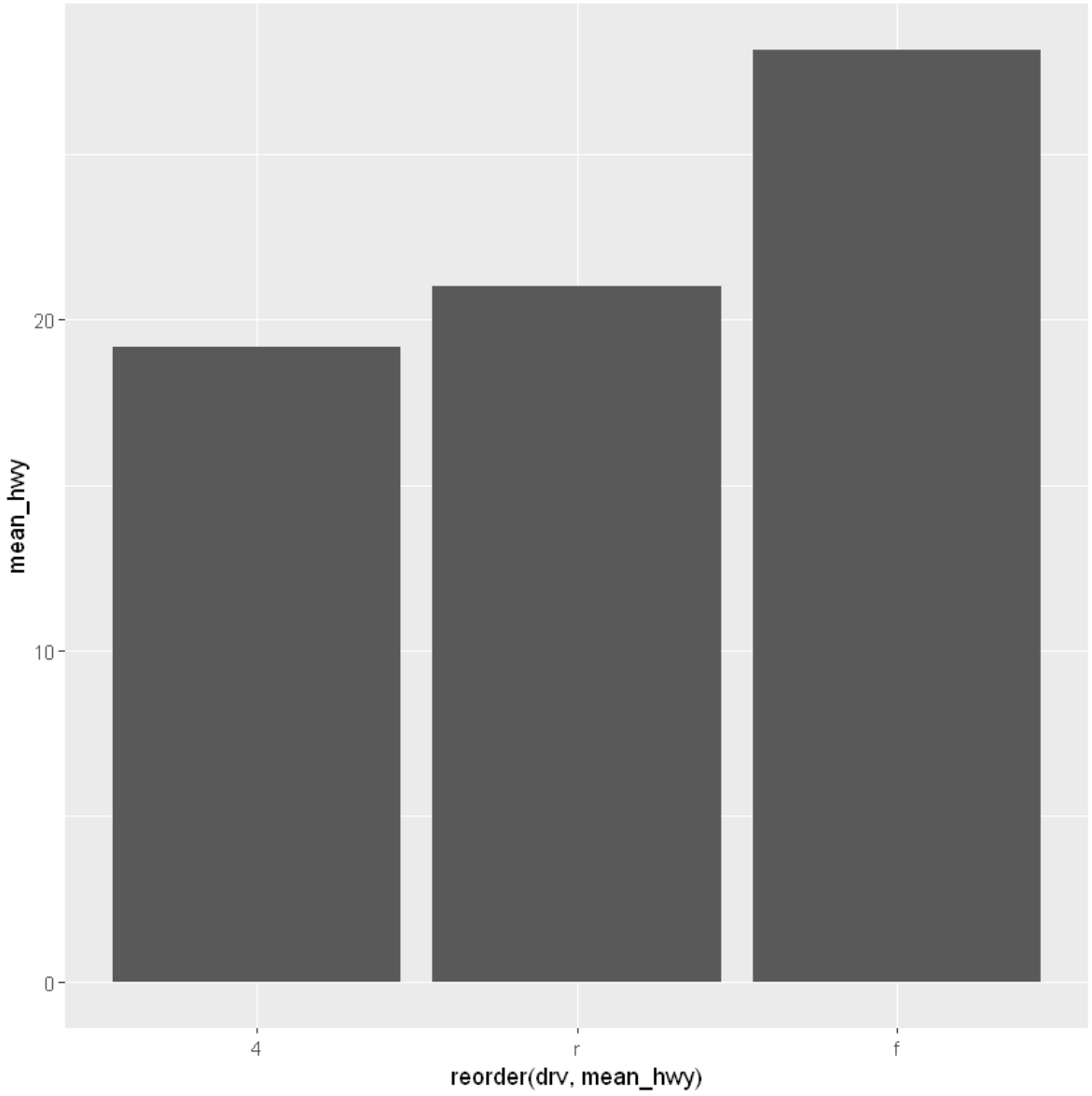
- Titanic 데이터 셋을 이용하여 PClass별 Fare의 평균을 알아보자.

크기순으로 정렬하기

- 범주의 알파벳 순으로 정렬된다.
- reorder(데이터, 정렬할 변수명)
- reorder(데이터, -정렬할 변수명) : '-' 내림차순
- reorder(데이터, 정렬할 변수명) : 없다면 오름차순

In [11]:

```
ggplot(data=drv_hwy, aes(x=reorder(drv, -mean_hwy), y=mean_hwy)) + geom_col()
```



7-2 빈도 막대 그래프

- y축 없이 x축만 지정하고, geom_col() 대신에 geom_bar()를 사용

구동방식(drv)의 데이터 개수

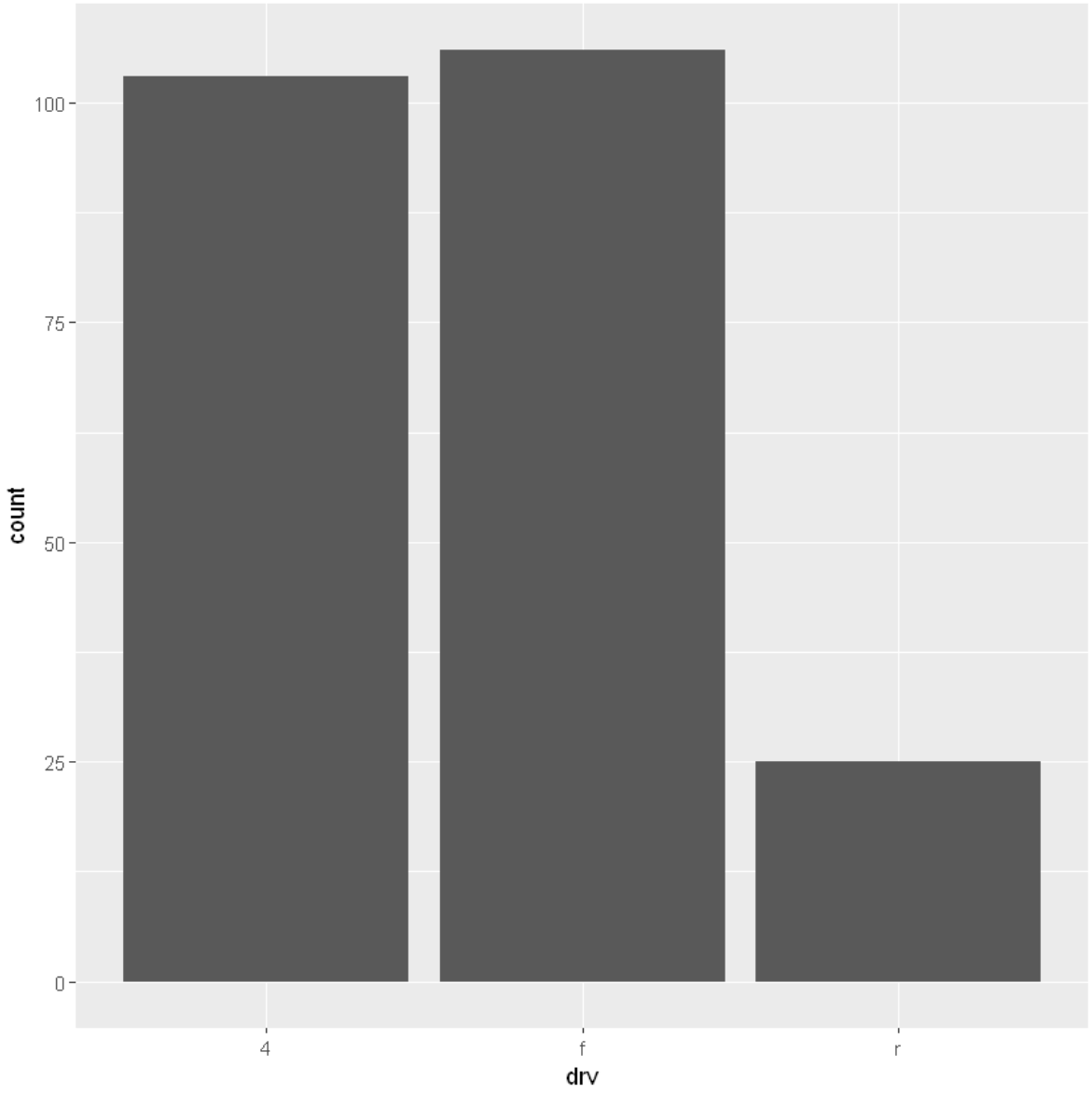
In [66]:

```
table(mpg$drv)
```

```
  4   f   r 
103 106  25 
```

In [67]:

```
ggplot(data=mpg, aes(x=drv)) + geom_bar()
```



In [68]:

```
names(mpg)
```

'manufacturer' 'model' 'displ' 'year' 'cyl' 'trans' 'drv' 'cty' 'hwy' 'fl' 'class'

class별 빈도수

In [69]:

```
str(mpg)
```

```
Classes 'tbl_df', 'tbl' and 'data.frame':       234 obs. of  11 variables:
 $ manufacturer: chr  "audi" "audi" "audi" "audi" ...
 $ model       : chr  "a4" "a4" "a4" "a4" ...
 $ displ       : num  1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
 $ year        : Factor w/ 2 levels "1999","2008": 1 1 2 2 1 1 2 1 1 2 ...
 $ cyl         : int  4 4 4 4 6 6 6 4 4 4 ...
 $ trans       : chr  "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...
 $ drv         : chr  "f" "f" "f" "f" ...
 $ cty         : int  18 21 20 21 16 18 18 18 16 20 ...
 $ hwy         : int  29 29 31 30 26 26 27 26 25 28 ...
 $ fl          : chr  "p" "p" "p" "p" ...
 $ class       : chr  "compact" "compact" "compact" "compact" ...
```

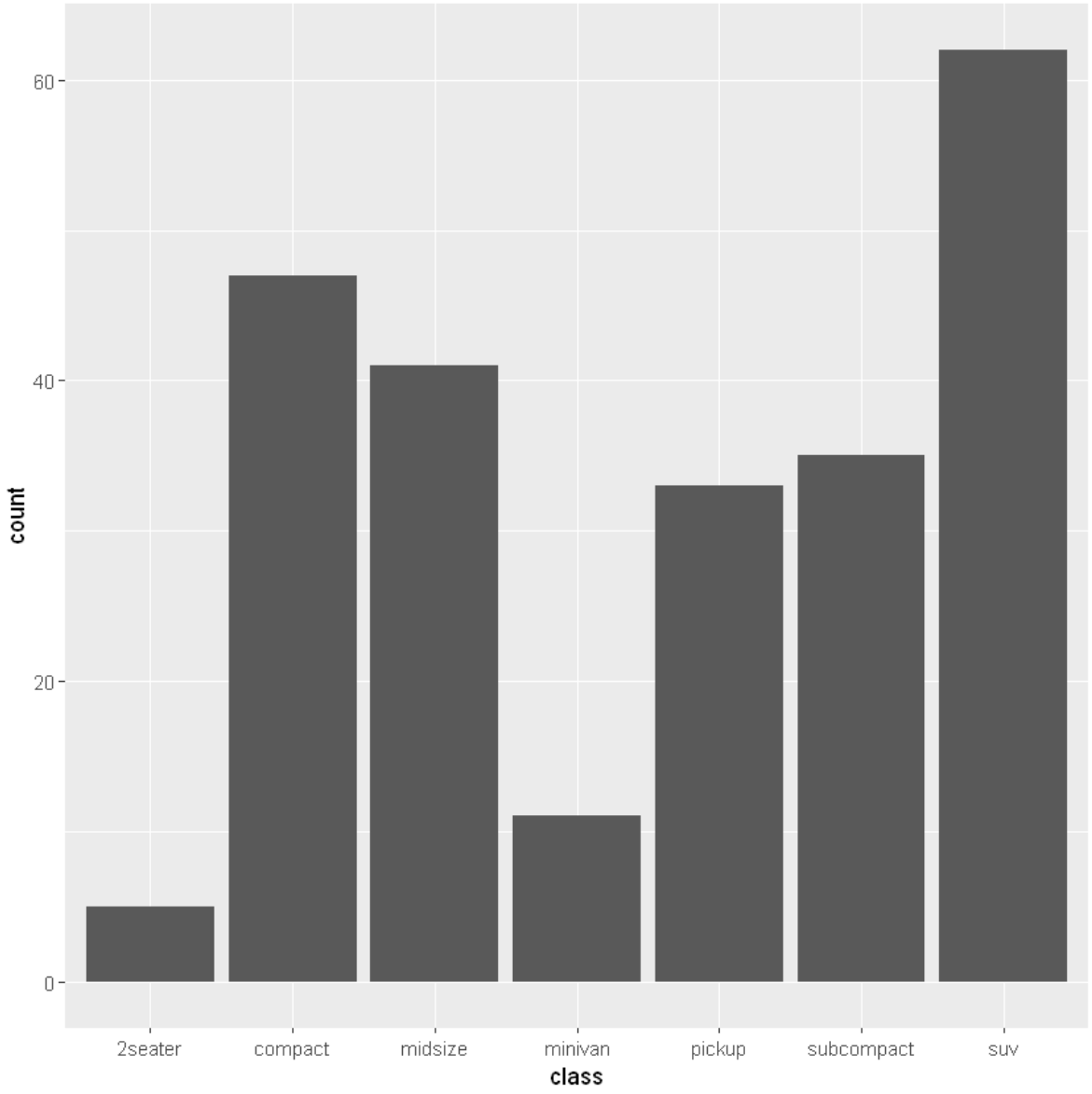
In [70]:

```
table(mpg$year)
table(mpg$fl)
table(mpg$class)
```

```
1999 2008
 117  117


  c   d   e   p   r
  1   5   8  52 168


   2seater     compact     midsize     minivan      pickup subcompact         suv
         5          47          41          11          33          35          62
```

In [71]:

```
ggplot(mpg, aes(x=class)) + geom_bar()
```



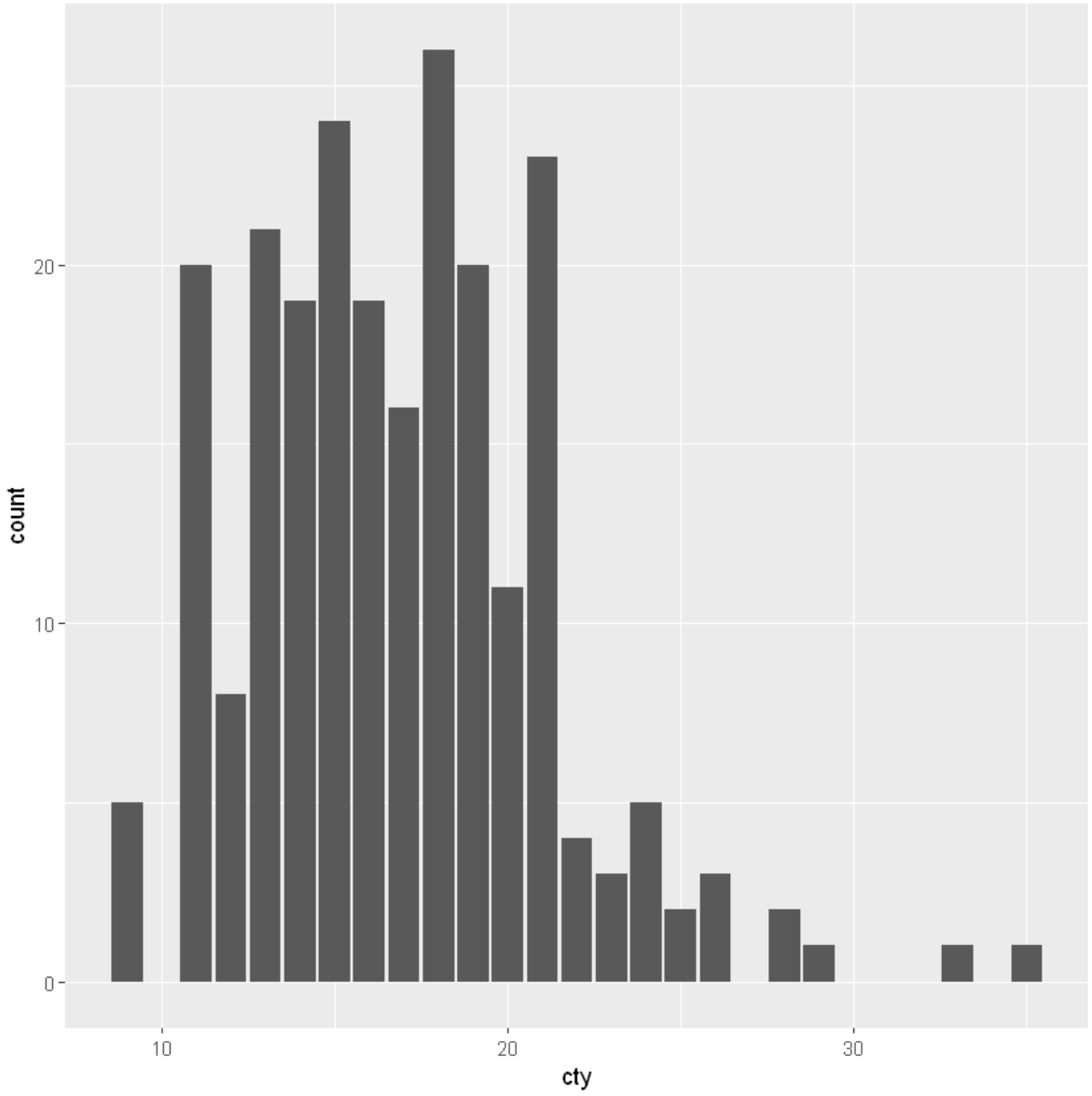
(실습해보기) 07-03

- Titanic 데이터 셋을 이용하여 PClass별 빈도수를 알아보자.

만약 x의 값에 연속 변수 지정시에 값의 분포를 파악이 가능

In [14]:

```
ggplot(mpg, aes(x=cty)) + geom_bar()
```



(직접해보기) 나머지 연속변수에 대해 확인해 보기

여러그래프를 한꺼번에 그려보기

- 패키지 : gridExtra, 라이브러리 : gridExtra
- 함수 : grid.arrange()

In [15]:

```
# 처음 설치시에 주석 없애고 설치
# install.packages("gridExtra")
```

In [16]:

```
?grid.arrange    # 도움말 살펴보기
```

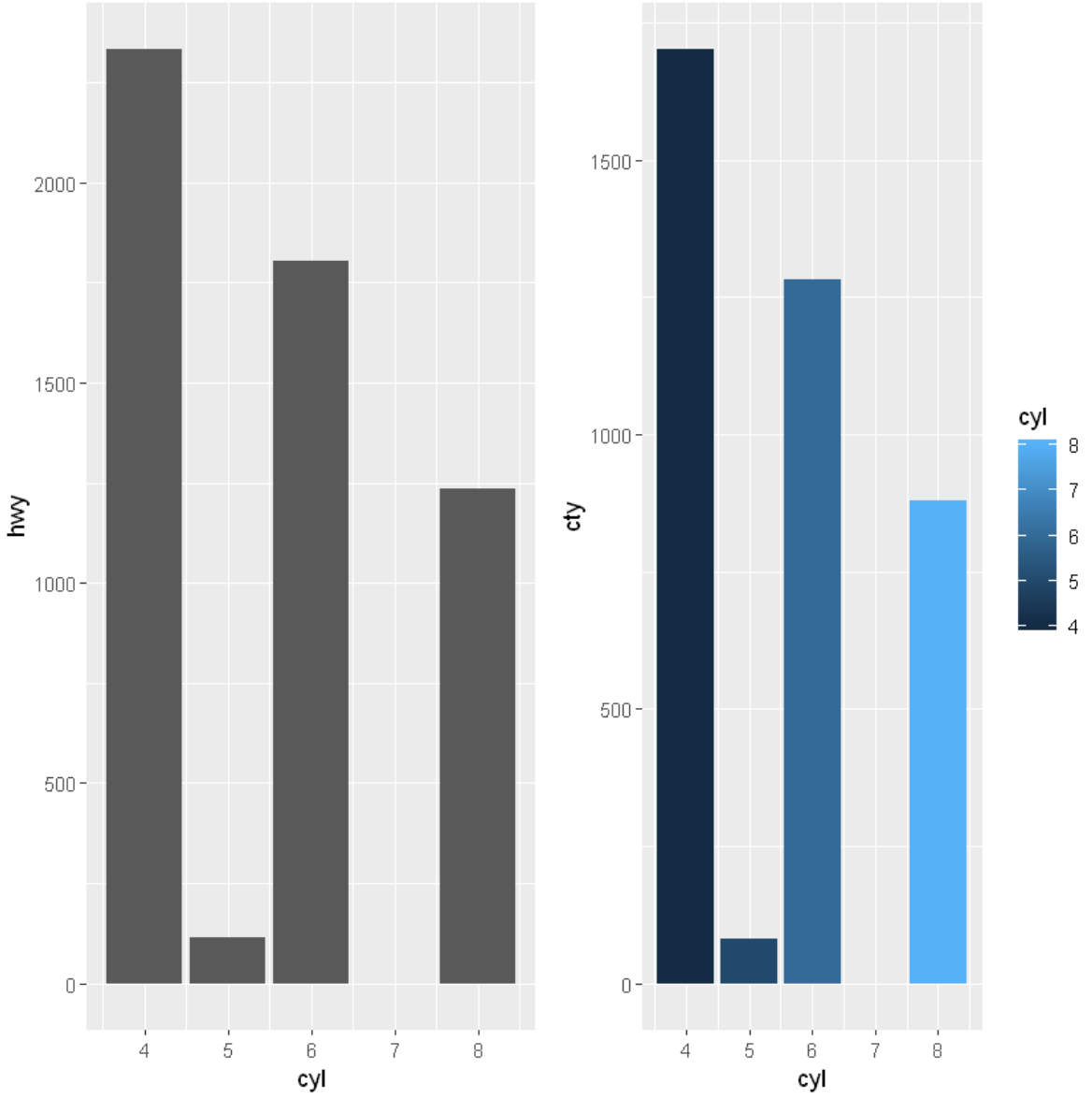
In [17]:

```
library(ggplot2)
library(gridExtra)
```

```
Attaching package: 'gridExtra'

The following object is masked from 'package:dplyr':

    combine
```

In [18]:

```
p1 <- ggplot(data=mpg, aes(x=cyl, y=hwy)) + geom_col()
p2 <- ggplot(data=mpg, aes(x=cyl, y=cty, fill=cyl)) + geom_col()

grid.arrange(p1, p2, ncol=2)
```

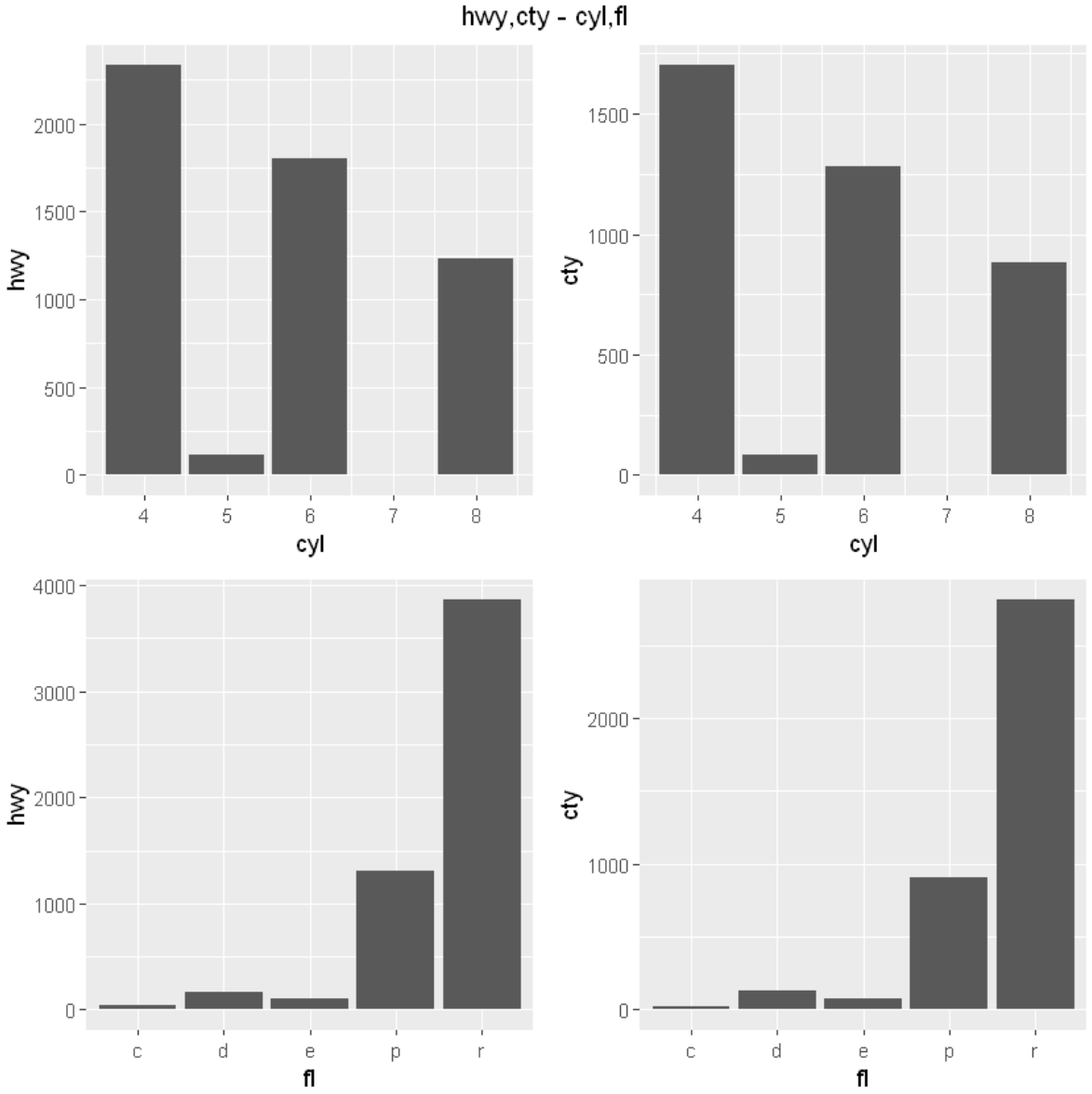


연료 타입 별, cyl(실린더)별 고속도로, 도시 연비 확인

- grid.arrange(그래프1, 그래프2,.., nrow=[숫자], ncol=[숫자],..)

In [77]:

```
p1 <- ggplot(data=mpg, aes(x=cyl, y=hwy)) + geom_col()
p2 <- ggplot(data=mpg, aes(x=cyl, y=cty)) + geom_col()

p3 <- ggplot(data=mpg, aes(x=fl, y=hwy)) + geom_col()
p4 <- ggplot(data=mpg, aes(x=fl, y=cty)) + geom_col()
grid.arrange(p1,p2,p3,p4, ncol=2, nrow=2, top = "hwy,cty - cyl,fl")
```



실습과제

- mpg 데이터 셋을 활용.
- (1) "compact" 차종 대상으로 평균 cty(도시연비)가 가장 높은 다섯 곳 막대그래프 표시
- (2) 막대는 연비가 높은 순으로 정렬
- (3) 자동차 중에서 어떤 연료 타입(fl)가 가장 많은지 연료타입(fl)별 빈도를 표시해 보자.

7-3 선 그래프

- 시간에 따라 달라지는 데이터를 표현할 때, 주로 선 그래프를 이용.
- 환율, 주가지수 등 경제지표가 시간에 따라 어떻게 달라지는지를 표현

데이터 셋(economics : ggplot2의 안의 데이터 셋)

- economics : 미국의 경제 지표들을 월별로 나타냄.

US economic time series

- US 경제 시계열 데이터
- url : http://research.stlouisfed.org/fred2 (http://research.stlouisfed.org/fred2)
- 478개의 행, 6개의 변수
- data format(데이터 형태)
  - date : 월별(날짜)
  - psavert : 개인 저축률, http://research.stlouisfed.org/fred2/series/PSAVERT/ (http://research.stlouisfed.org/fred2/series/PSAVERT/)
  - pce : 개인소비 지출, 수십억 달러, http://research.stlouisfed.org/fred2/series/PCE (http://research.stlouisfed.org/fred2/series/PCE)
  - unemploy : 실업자수(수천), http://research.stlouisfed.org/fred2/series/UNEMPLOY (http://research.stlouisfed.org/fred2/series/UNEMPLOY)
  - uempmed : 평균 실업 기간,(주별), http://research.stlouisfed.org/fred2/series/UEMPMED (http://research.stlouisfed.org/fred2/series/UEMPMED)
  - pop : 총 인구, (수천), http://research.stlouisfed.org/fred2/series/POP (http://research.stlouisfed.org/fred2/series/POP)

In [86]:

```
names(economics)
str(economics)
head(economics,3)
```

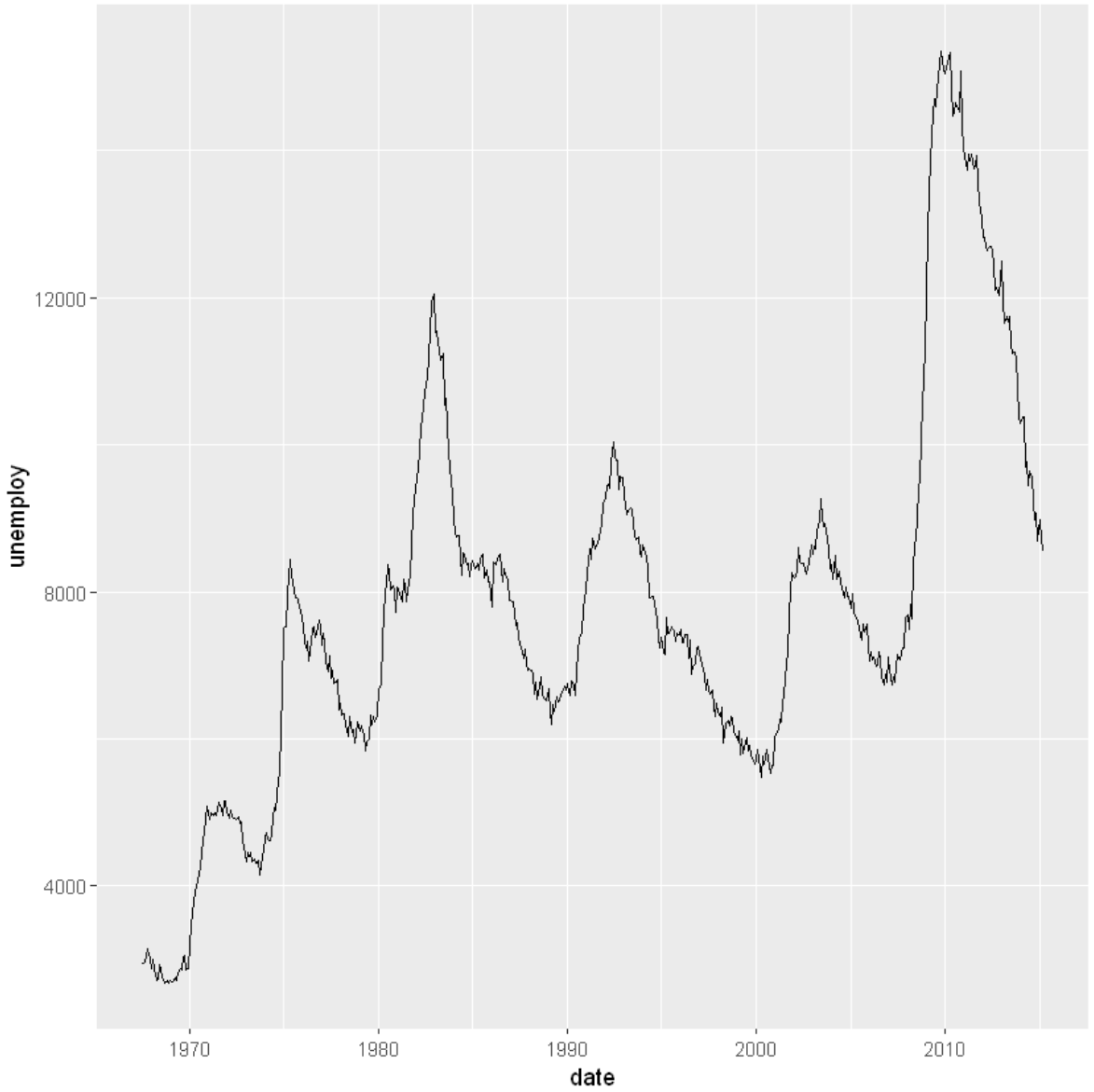
'date' 'pce' 'pop' 'psavert' 'uempmed' 'unemploy'

```
Classes 'tbl_df', 'tbl' and 'data.frame':       574 obs. of  6 variables:
 $ date    : Date, format: "1967-07-01" "1967-08-01" ...
 $ pce     : num  507 510 516 513 518 ...
 $ pop     : int  198712 198911 199113 199311 199498 199657 199808 199920 200056 200
208 ...
 $ psavert : num  12.5 12.5 11.7 12.5 12.5 12.1 11.7 12.2 11.6 12.2 ...
 $ uempmed : num  4.5 4.7 4.6 4.9 4.7 4.8 5.1 4.5 4.1 4.6 ...
 $ unemploy: int  2944 2945 2958 3143 3066 3018 2878 3001 2877 2709 ...
```

date	pce	pop	psavert	uempmed	unemploy
1967-07-01	507.4	198712	12.5	4.5	2944
1967-08-01	510.5	198911	12.5	4.7	2945
1967-09-01	516.3	199113	11.7	4.6	2958

In [82]:

```
### x : 날짜, y : 실업자수
ggplot(data=economics, aes(x=date, y=unemploy)) + geom_line()
```



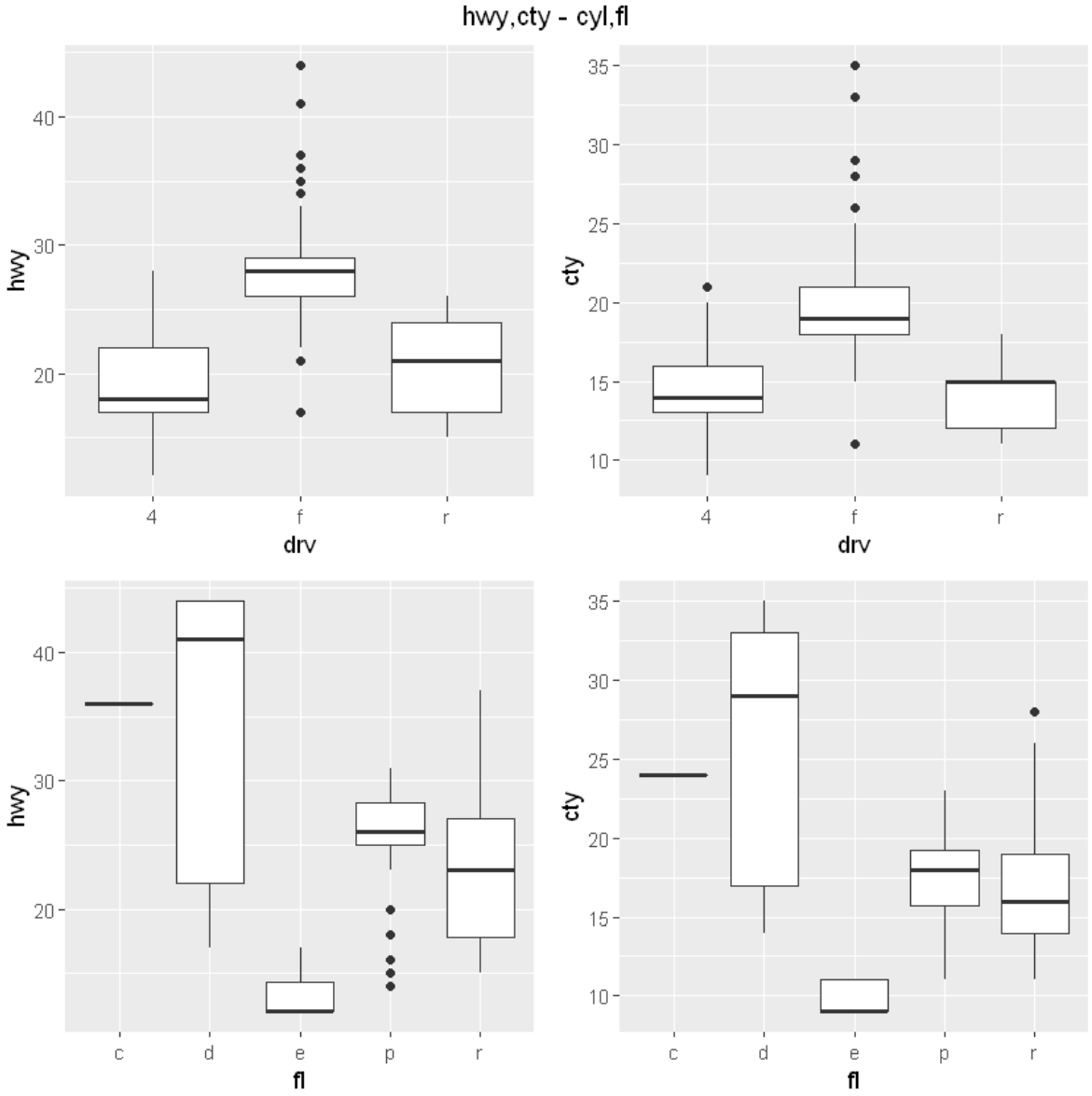
해석

- 실업자 수가 5년 주기로 등락을 반복
- 2005년 이후 급격하게 증가 후, 2010년 이후 다시 감소하는 추세

(실습해보기) 시간에 따른 개인 소비 지출의 변화(pcc) 를 알아보기

In [85]:

```
p1 <- ggplot(data=mpg, aes(x=drv, y=hwy)) + geom_boxplot()
p2 <- ggplot(data=mpg, aes(x=drv, y=cty)) + geom_boxplot()

p3 <- ggplot(data=mpg, aes(x=fl, y=hwy)) + geom_boxplot()
p4 <- ggplot(data=mpg, aes(x=fl, y=cty)) + geom_boxplot()
grid.arrange(p1,p2,p3,p4, ncol=2, nrow=2, top = "hwy,cty - cyl,fl")
```



(실습해보기) class가 'compact', 'subcompact', 'suv'인 자동차의 hwy(고속도로 연비)의 차이 비교.

REF

- 데이터 셋(economics) : http://research.stlouisfed.org/fred2 (http://research.stlouisfed.org/fred2)
- The R Graph Gallery : https://www.r-graph-gallery.com/ (https://www.r-graph-gallery.com/)
- ggplot 확장 패키지 : http://www.ggplot2-exts.org/gallery/ (http://www.ggplot2-exts.org/gallery/)
- 기타 ggplot : http://www.sthda.com/english/wiki/ggplot2-barplots-quick-start-guide-r-software-and-data-visualization (http://www.sthda.com/english/wiki/ggplot2-barplots-quick-start-guide-r-software-and-data-visualization)