5장. ggplot2에 의한 자료 시각화

그래픽 기법의 활용 예제: 보리자료

- 자료분석과정에서 그래프의 이용이 필수적임을 보여주는 예제

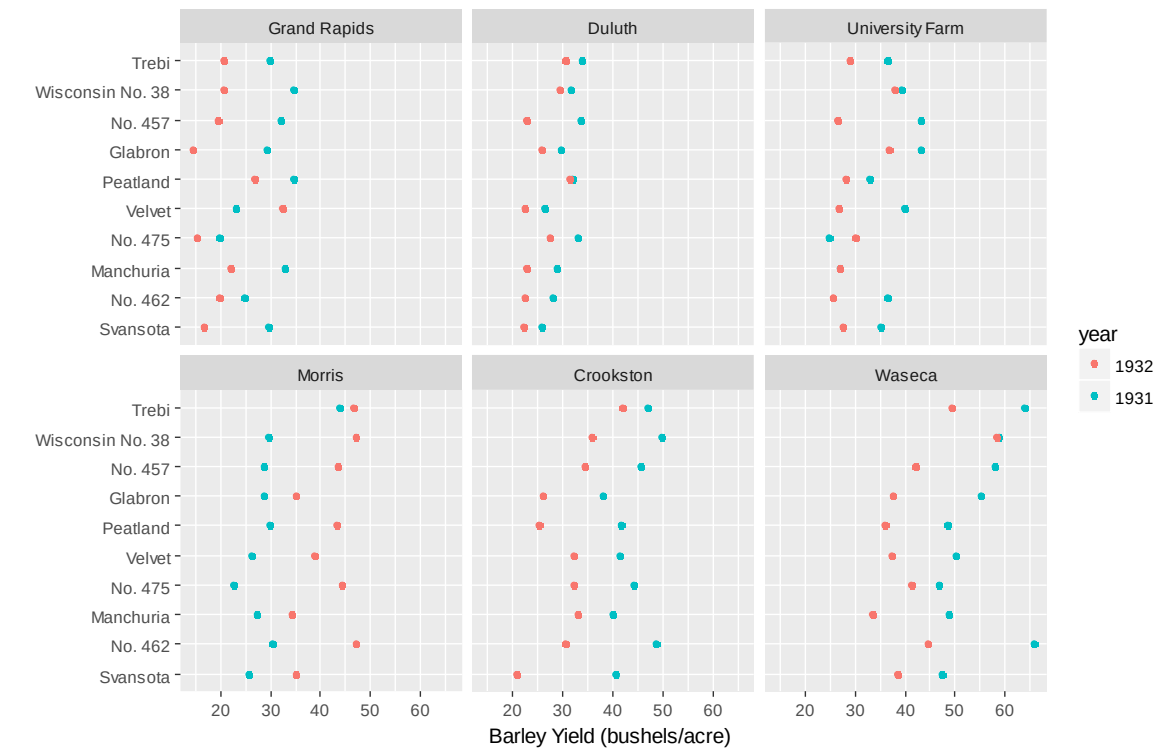
- 보리자료:
  - 1930년대 초 미네소타 주 농경학자들이 보리종류에 따른 수확량의 차이를 비교하기 위해 2년간 경작실험을 실시

  - 설명변수: 6군데 경작지, 10종류의 보리, 2년간의 경작 년도

  - 반응변수: 수확량

- 저명한 학자들에 의해서 여러 번 분석된 자료
- Cleveland가 자료에 있는 문제 발견

- 보리자료의 문제점

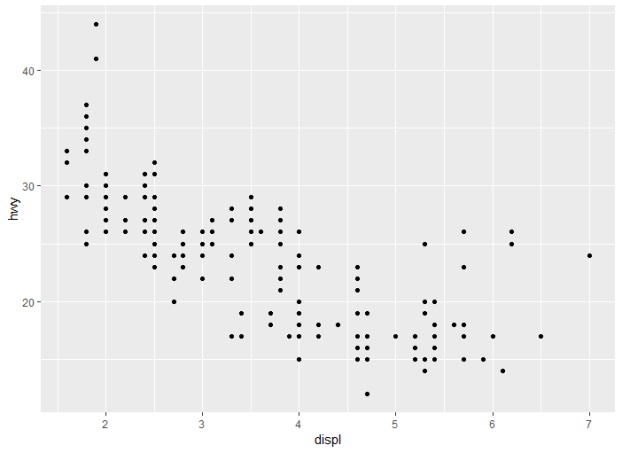


- 자료의 시각화: 효과적인 자료 분석 기법
- ggplot2: 가장 뛰어난 그래프 작성 방식

5.1 ggplot2 시작하기

- 패키지 ggplot2에 있는 데이터 프레임 mpg의 변수 displ과 hwy의 산점도 작성

```
> library(tidyverse)
> ggplot(data=mpg) +
    geom_point(mapping=aes(x=displ, y=hwy))
```



- 함수 ggplot( ): 데이터 프레임 data 지정. 그래프가 작성될 비어있는 영역 확보
- 함수 geom_point( ): 실질적인 그래프를 작성하는 geom 함수
- mapping: geom 함수 내에서 함수 aes( )와 함께 시각적 요소를 데이터와 연결

- ggplot2에서 그래프 작성의 최소 요소

  - 그래프 작성을 위한 법칙이 있음: 그래프의 문법
  - 모든 그래프 작성에 일정하게 적용
  - 익숙해지면 복잡한 형태의 그래프도 어렵지 않게 작성 가능

그래프 작성을 위한 3가지 최소 요소:
<Data>, <Geom_function>, <Mappings>

```
ggplot(data=<Data>) +
        <Geom_function>(mapping=aes(<Mappings>)
```

<Data>: 그래프 작성에 사용될 데이터 프레임 지정

<Geom_function>: 레이어(layer) 작성을 위한 geom 함수.
여러 개의 geom 함수를 덧셈 기호로 연결하면 여러 레이어가 겹쳐진 그래프 작성

<Mappings>: 시각적 요소(점의 크기, 모양, 색깔, ...)와 데이터 연결

5.2 시각적 요소와 연결: Mapping

- 시각적 요소
  - 그래프를 시각적으로 인식할 때 필요한 요소
  - 산점도의 경우, 점의 위치, 크기, 모양 및 색깔 등이 시각적 요소

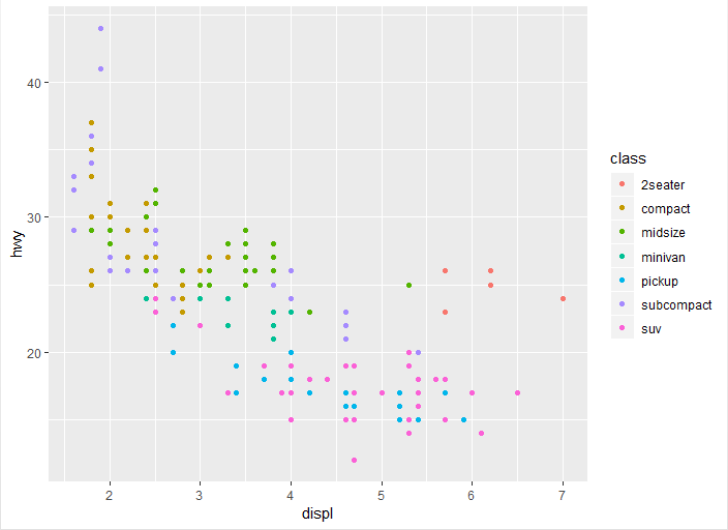
- 시각적 요소의 mapping과 setting
  - mapping: 데이터의 값과 연결되어 결정. 함수 aes( ) 안에서 연결
  - setting: 사용자가 일정한 값을 지정. 함수 aes( ) 밖에서 지정

- 시각적 요소 mapping의 기능:
  - 기존의 그래프에 다른 변수의 정보 추가

- 예제: 데이터 프레임 mpg의 변수 displ과 hwy의 산점도에 시각적 요소와의 mapping으로 다른 변수 정보 추가

  - 변수 class를 시각적 요소 color와 mapping

```
> ggplot(data=mpg) +
    geom_point(mapping=aes(x=displ, y=hwy, color=class))
```

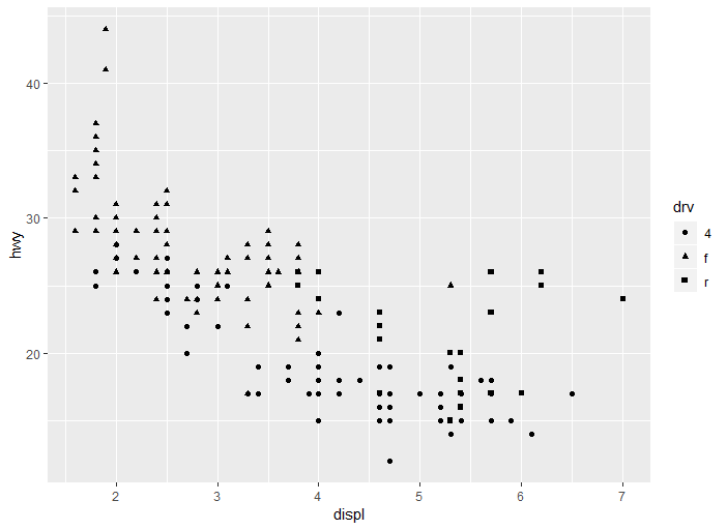


- 색의 종류(범주의 개수) 과다
- 좋은 그래프는 아님

- 변수 class: 문자형 벡터
- 변수 class의 값에 따라 다른 색 사용
- 사용된 색깔에 대한 범례는 자동으로 추가

- 변수 drv를 시각적 요소 shape와 mapping

```
> ggplot(data=mpg) +
    geom_point(mapping=aes(x=displ, y=hwy, shape=drv))
```

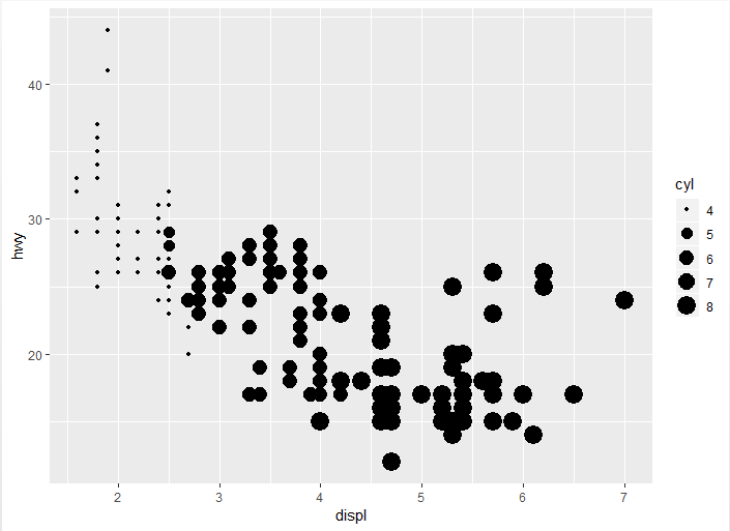


- 구분이 어려운 다른 모양의 점
- 좋은 그래프는 아님

- shape에 mapping되는 변수는 이산형
- 변수 drv: 문자형 벡터
- 변수 drv의 값에 따라 다른 모양의 점 사용
- 범례 자동 추가

- 변수 cyl을 시각적 요소 size와 mapping

```
> ggplot(data=mpg) +
    geom_point(mapping=aes(x=displ, y=hwy, size=cyl))
```

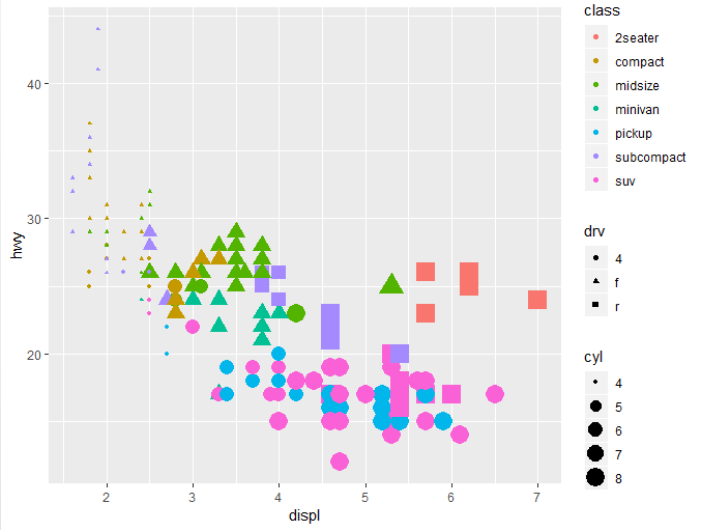


- size에 mapping되는 변수는 연속형이 좋음
- 변수 cyl: 정수형 변수
- cyl의 값에 따라 점의 크기 조절
- 범례 자동 추가

- 구분이 어려운 점 크기
- 범주의 개수 과다 → 좋은 그래프는 아님

- 여러 시각적 요소를 동시에 mapping
  - 변수 class는 color와, drv는 shape와, cyl은 size와 mapping

```
> ggplot(data=mpg) +
    geom_point(mapping=aes(x=displ, y=hwy,
                           color=class, shape=drv, size=cyl))
```



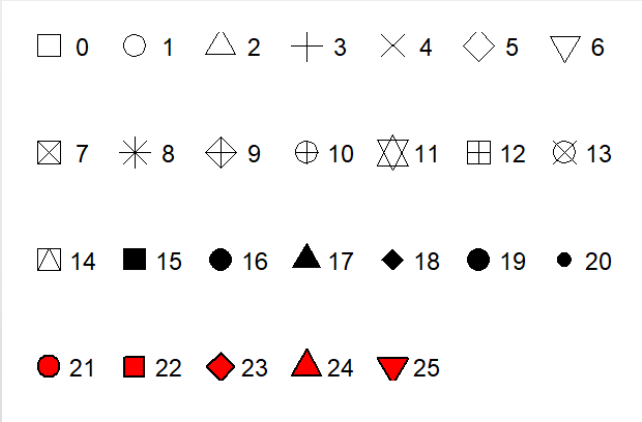
- 너무 많은 정보
- 그래프의 의미가 모호

- 시각적 요소의 setting

  - 함수 aes( ) 밖에서 사용자가 원하는 값으로 지정
  - geom 함수의 입력 요소가 됨

- 시각적 요소 color, size, shape에 값 지정 법칙

  1) color: 색깔을 나타내는 문자열 지정
  2) size: 점 크기를 mm 단위로 지정
  3) shape: 점의 형태를 나타내는 0~26 사이의 숫자



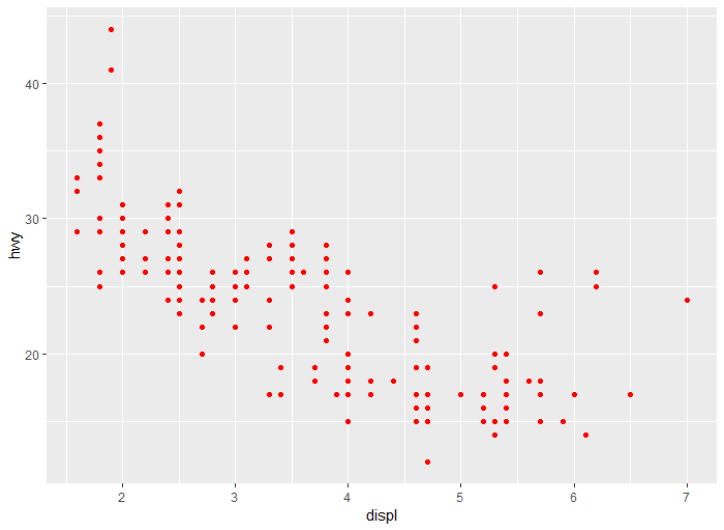
도형에 색깔 지정 방법

1) 0~14의 외곽선 및 15~20의 도형 색: color 사용
2) 21~25의 외곽선: color 사용
3) 21~25의 내부 색: fill 사용

https://www.r-graph-gallery.com/colors

- 시각적 요소 color의 setting: 모든 점을 빨간 색으로

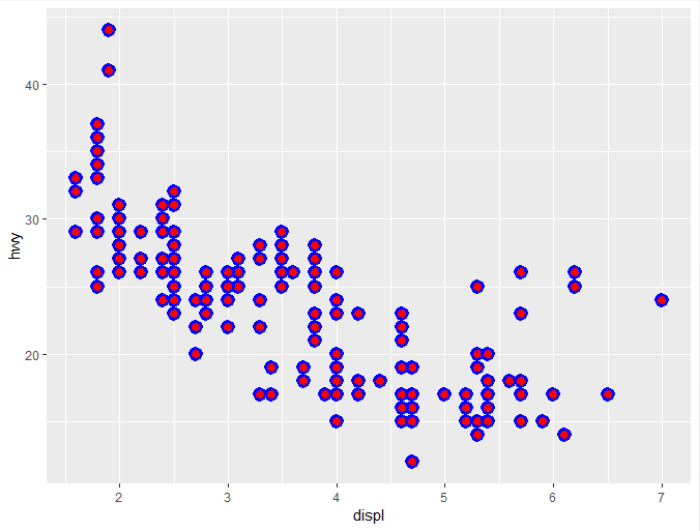
```
> ggplot(data=mpg) +
    geom_point(mapping=aes(x=displ, y=hwy), color="red")
```



- color를 함수 aes( ) 밖에서 지정
- 함수 geom_point( )의 입력 요소

- 여러 시각적 요소를 동시에 setting

```
> ggplot(data=mpg) +
    geom_point(mapping=aes(x=displ, y=hwy), color="blue", size=3,
               shape=21, fill="red", stroke=2)
```



- 점의 모양: shape=21
- 점의 내부 색: 빨간색
- 점의 외곽선 색: 파란색
- 점의 크기 확대: size=3
- 점의 외곽선 두께 조절: stroke=2

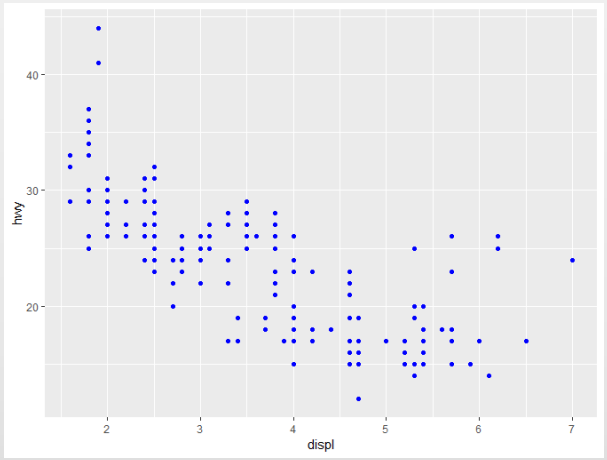
- 함수 aes( ) 안에서 시각적 요소에 특정 값을 setting한 경우의 결과

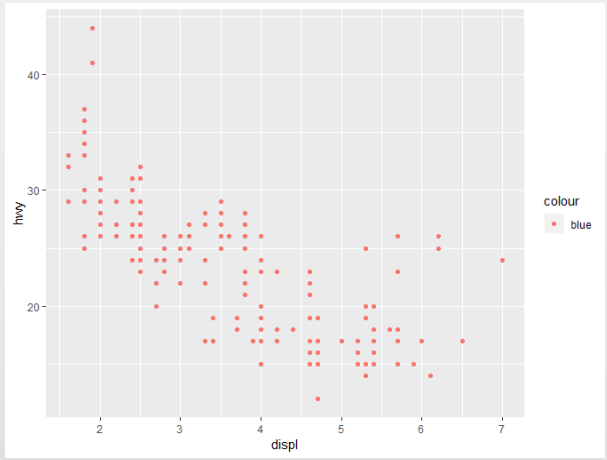
Setting

```
> ggplot(data=mpg)+
    geom_point(mapping=aes(x=displ, y=hwy), color="blue")
```

Mapping

```
> ggplot(data=mpg) +
    geom_point(mapping=aes(x=displ, y=hwy, color="blue"))
```





- mapping은 변수와의 연결을 의미
- "blue"라는 값을 갖는 변수 생성

5.3 그룹별 그래프 작성: Facet

- 범주형 변수가 다른 변수에 미치는 영향력을 그래프로 확인하는 방법

  1) 시각적 요소에 범주형 변수를 mapping
  2) 범주형 변수로 그룹 구성하고, 각 그룹별 그래프 작성: faceting

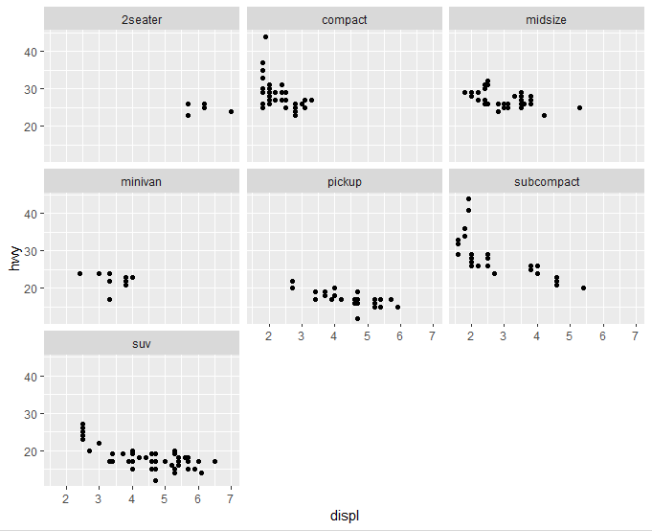
- facet을 적용하기 위한 함수
  ① facet_wrap( ): 한 변수에 의한 facet
  ② facet_grid( ): 한 변수 또는 두 변수에 의한 facet

- 함수 facet_wrap( )에 의한 faceting

  - 데이터를 구분하는 변수가 하나인 경우: facet_wrap(~ x)

• 데이터 프레임 mpg의 변수 displ과 hwy의 산점도를 class의 범주별로 작성

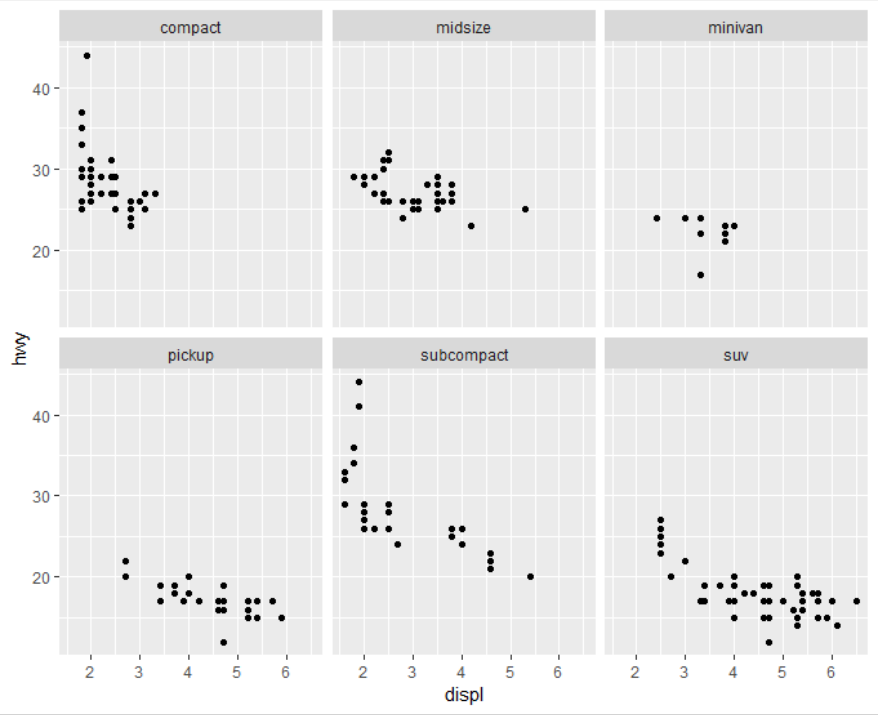
```
> ggplot(data=mpg) +
    geom_point(mapping=aes(x=displ, y=hwy)) +
    facet_wrap(~ class)
```



- 패널 '2seater'에는 적은 수의 데이터 존재
- class가 '2seater'인 케이스 제거 후 다시 작성

- 데이터 프레임 mpg의 변수 displ과 hwy의 산점도를 class의 범주별로 작성 (2seater 케이스 제외)

```
> mpg %>%
    filter(class != "2seater") %>%
    ggplot() +
    geom_point(mapping=aes(x=displ, y=hwy)) +
    facet_wrap(~ class)
```



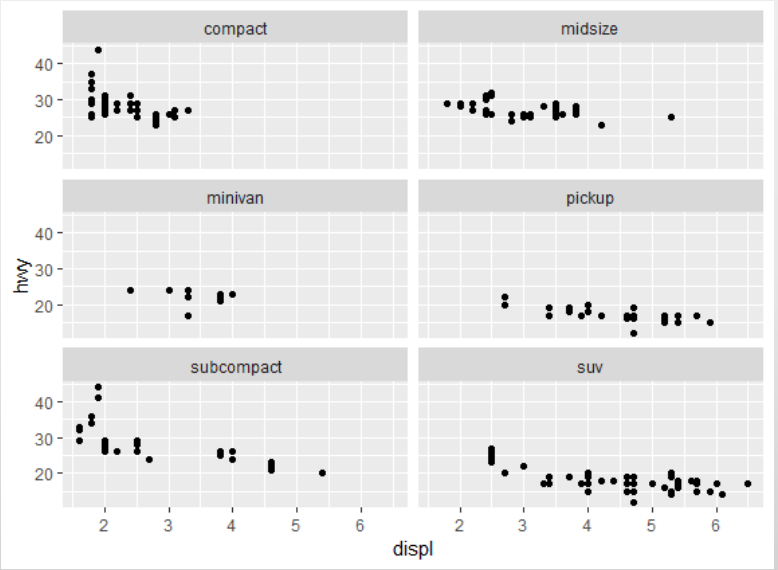
- 패널 배치 조절

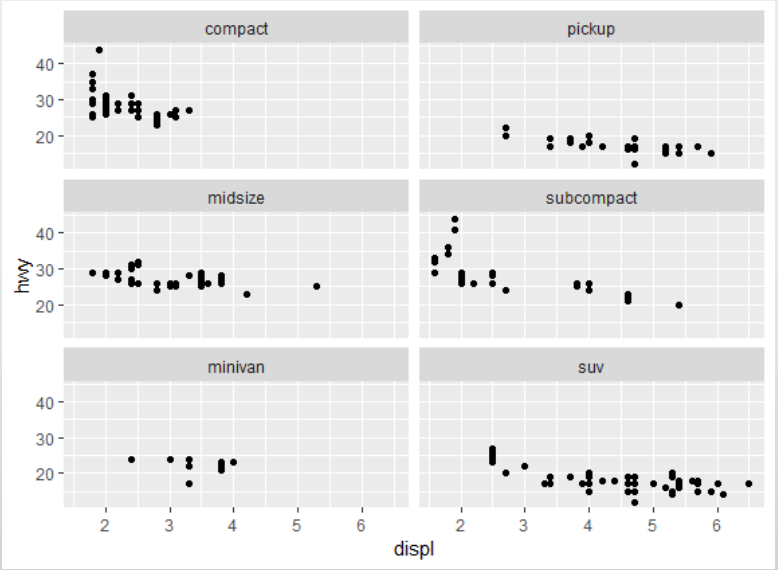
  - 2×3 패널 패치를 3×2 배치로 수정: ncol=2
  - 패널에 그래프 배치 순서를 열 단위로 수정: dir="v"

```
> pp <- mpg %>%
    filter(class != "2seater") %>%
    ggplot() +
    geom_point(mapping=aes(x=displ, y=hwy))
```

```
> pp + facet_wrap(~ class, ncol=2)
> pp + facet_wrap(~ class, ncol=2, dir="v")
```

두 그래프에 공통적으로 적용되는 내용을 개체에 할당





- 함수 facet_grid( )에 의한 faceting

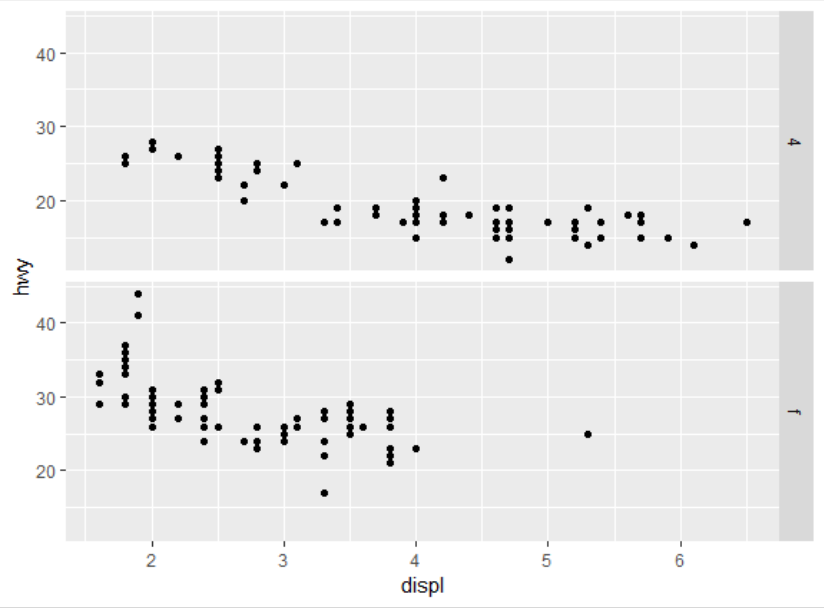
  - 한 변수에 의한 faceting:
    하나의 행으로 패널 배치: `facet_grid(. ~ x)`
    하나의 열로 패널 배치: `facet_grid(x ~ .)`

  - 두 변수에 의한 faceting: `facet_grid(y ~ x)`
    행 범주: 변수 y의 범주
    열 범주: 변수 x의 범주

- 데이터 프레임 mpg에서 변수 drv와 cyl의 범주별로 displ과 hwy의 산점도 작성. 단, drv가 "r"인 자료와 cyl이 5인 자료는 제외

```
> my_plot <- mpg %>%
    filter(cyl != 5, drv != "r") %>%
    ggplot() +
    geom_point(mapping=aes(x=displ, y=hwy))
```

```
> my_plot + facet_grid(drv ~ .)
```

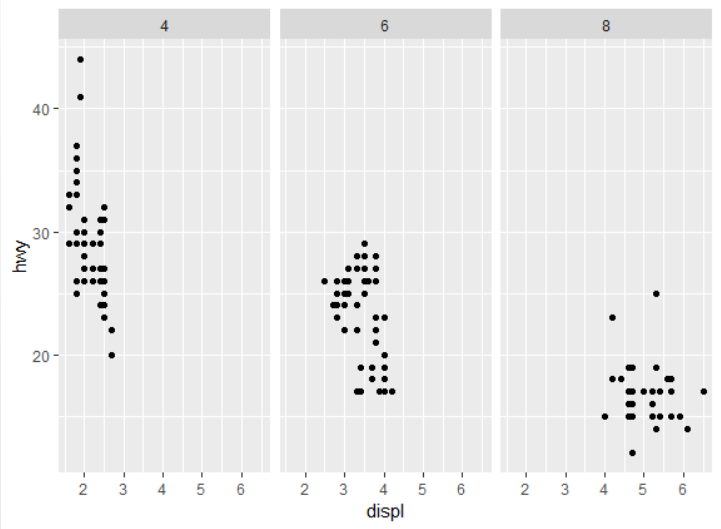


drv='4'

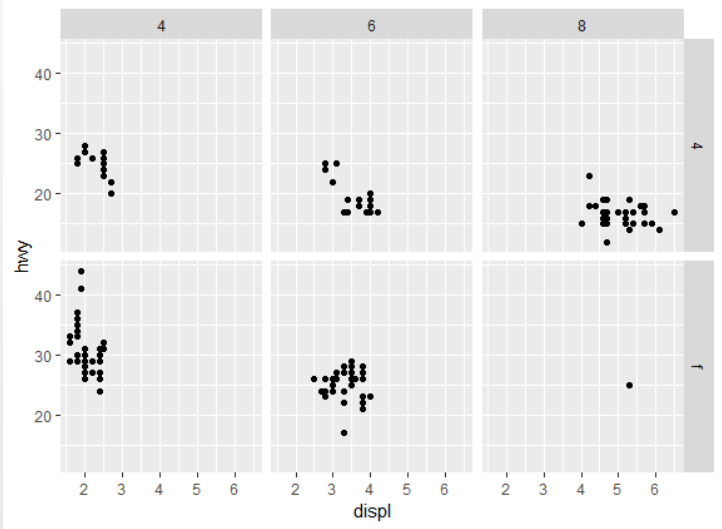
drv='f'

```
> my_plot + facet_grid(. ~ cyl)
```

```
> my_plot + facet_grid(drv ~ cyl)
```



cyl=4 cyl=6 cyl=8



drv='4'

drv='f'

cyl=4 cyl=6 cyl=8

- 연속형 변수에 의한 faceting

  - 연속형 변수를 범주형 변수로 변환 후 faceting

  - 유용한 함수

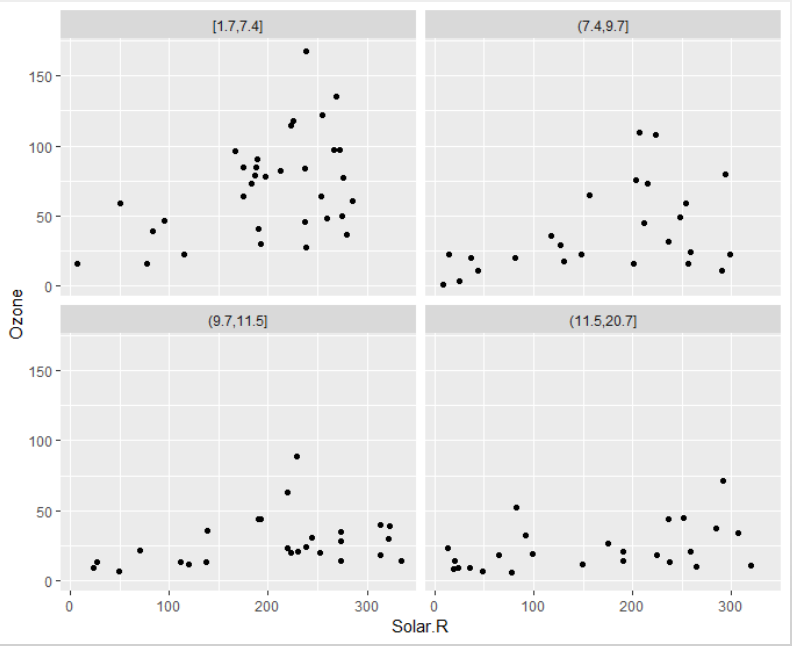
    ① `cut_interval(x, n, length)`: 벡터 x를 길이가 length인 n개의 구간으로 구분

    ② `cut_width(x, width, boundary)`: 벡터 x를 길이가 width인 구간으로 구분. 옵션 boundary는 구간의 시작점 지정.

    ③ `cut_number(x, n)`: 벡터 x를 n개의 구간으로 구분하되 각 구간에 속한 데이터의 개수가 대략 동일하도록 구분

- 데이터 프레임 airquality에서 변수 Ozone, Solar.R, Wind의 관계 탐색
  1) 변수 Wind를 4개의 구간으로 구분하되 속한 자료의 개수가 비슷하도록
  2) 4개의 구간에서 Ozone과 Solor.R의 산점도 작성

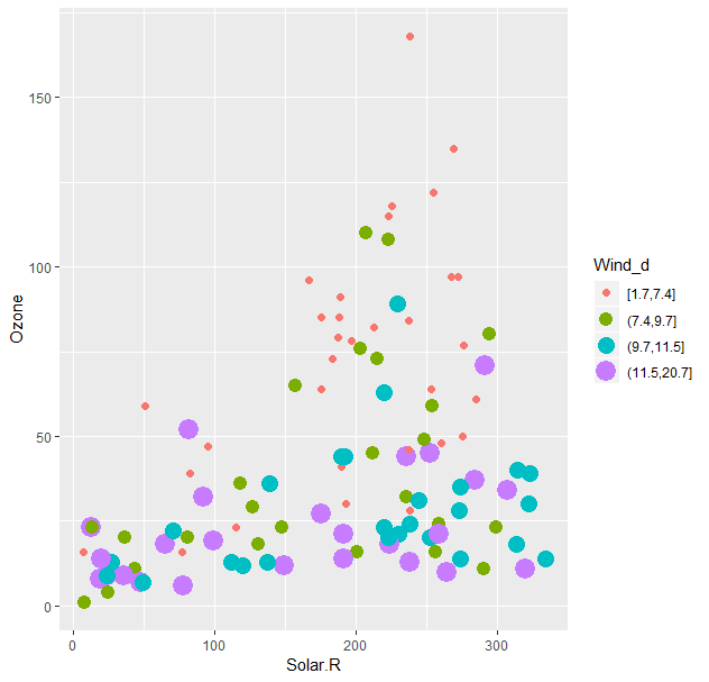
```
> air_1 <- airquality %>%
    mutate(Wind_d=cut_number(Wind, n=4))
> air_1 %>%
    ggplot() +
    geom_point(mapping=aes(x=Solar.R,y=Ozone)) +
    facet_wrap(~Wind_d)
```



- 변수 Wind가 큰 값을 가질수록 두 변수의 관계는 점점 미약해지고 있음
- 세 연속형 변수의 관계 탐색 방법 중 하나

- 한 그래프에 함께 작성

```
> air_1 %>%
    ggplot() +
    geom_point(mapping=aes(x=Solar.R, y=Ozone, color=Wind_d,
                           size=Wind_d))
```



5.4 기하 객체: Geometric object

- Base graphics에서 그래프 작성 방식: pen on paper
  - 높은-수준의 그래프 함수: 좌표축과 주요 그래프 작성
  - 낮은-수준의 그래프 함수: 점, 선, 문자 등을 추가하여 원하는 그래프 작성

- ggplot2에서 그래프 작성 방식
  - 작성하고자 하는 그래프: 몇몇 유형의 그래프(점 그래프, 선 그래프 등등)를 겹쳐 놓은 것
  - 몇몇 유형의 그래프를 각기 따로 작성
  - 작성된 그래프를 겹쳐 놓음으로써 원하는 그래프 작성

- ggplot2 시스템

  원하는 유형의 그래프(점 그래프, 선 그래프 등등) 작성

  ↔ 해당되는 기하 객체(geom)를 사용하여 그래프 작성

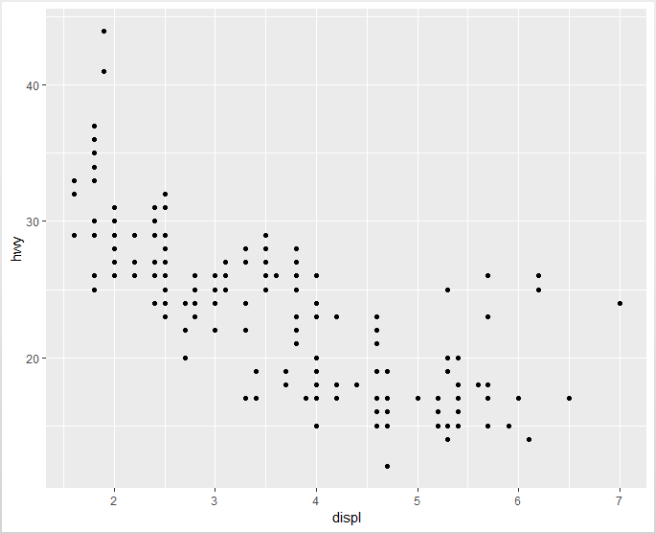
- 기하 객체의 사용
  - 해당되는 geom 함수의 실행
  - geom 함수 실행 → 해당 유형의 그래프가 작성된 layer 생성
  - 여러 개의 geom 함수 실행: 여러 layer 생성되고 이것들이 겹쳐져서 원하는 그래프 완성

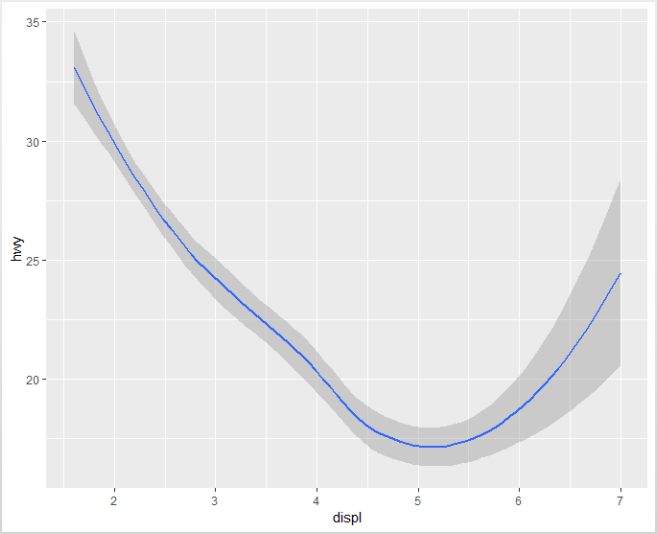
- 동일 자료에 다른 geom 적용

  - mpg의 변수 displ과 hwy를 대상으로 point geom과 smooth geom 적용
    - point geom: 점 그래프 작성
    - smooth geom: 비모수 회귀곡선 작성

```
> ggplot(data=mpg) +
    geom_point(mapping=aes(x=displ, y=hwy))
```

```
> ggplot(data=mpg) +
    geom_smooth(mapping=aes(x=displ, y=hwy))
```





- geom 함수 리스트

  - 현재 대략 30개 이상의 geom 함수가 있음

  - 한 변수에 대한 함수: geom_bar( ), geom_histogram( ), geom_density( ), geom_dotplot( ) 등등

  - 두 변수에 대한 함수: geom_point( ), geom_smooth( ), geom_text( ), geom_line( ), geom_boxplot( ) 등등

  - 세 변수에 대한 함수: geom_contour( ), geom_tile( ) 등등

  - geom 함수의 리스트: R studio의 메뉴에서 'Help > Cheatsheets > Data Visualization with ggplot2' 에서 확인 가능

- 글로벌 매핑과 로컬 매핑

  - 글로벌 매핑: 함수 ggplot( )에서의 매핑. 해당 그래프 작성에 참여한 모든 geom 함수에 적용

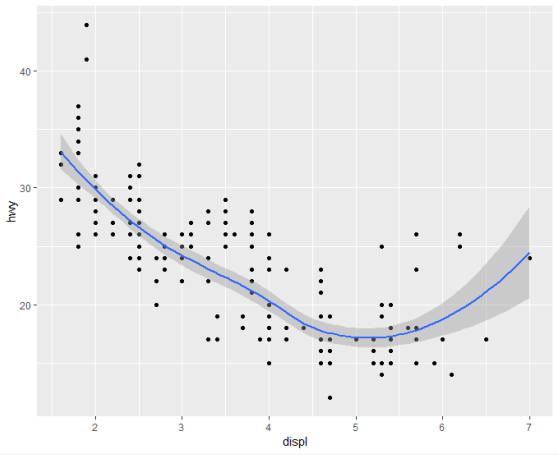
  - 로컬 매핑: geom 함수에서의 매핑. 해당 geom 함수로 작성되는 layer에만 적용. 해당 layer에서는 글로벌 매핑보다 우선해서 적용됨.

```
ggplot(data, mapping=aes( ) ) +
      geom_*(mapping=aes( ) ) +
      geom_*(mapping=aes( ) )
```

글로벌 매핑

로컬 매핑

- 예: mpg의 변수 displ과 hwy의 산점도에 비모수 회귀곡선 추가



- 두 geom 함수에 동일한 내용의 매핑이 중복 입력되는 상황

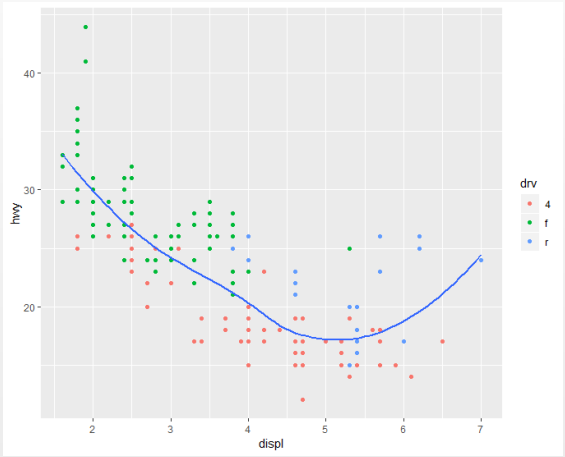
```
> ggplot(data=mpg) +
    geom_point(mapping=aes(x=displ, y=hwy)) +
    geom_smooth(mapping=aes(x=displ, y=hwy))
```

- 글로벌 매핑으로 중복 입력 문제 해결

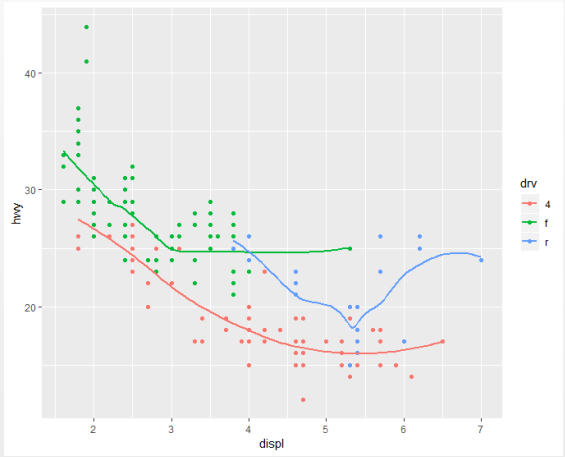
```
> ggplot(data=mpg, mapping=aes(x=displ, y=hwy)) +
    geom_point() +
    geom_smooth()
```

- 예: mpg의 변수 displ과 hwy의 비모수 회귀곡선 작성. 그 위에 산점도 추가하되 drv의 값에 따라 점의 색을 구분.

비모수 회귀곡선: 전체 자료 대상



각 그룹별로 추정된 비모수 회귀




```
> ggplot(data=mpg, mapping=aes(x=displ, y=hwy)) +
    geom_point(mapping=aes(color=drv)) +
    geom_smooth(se=FALSE)
```

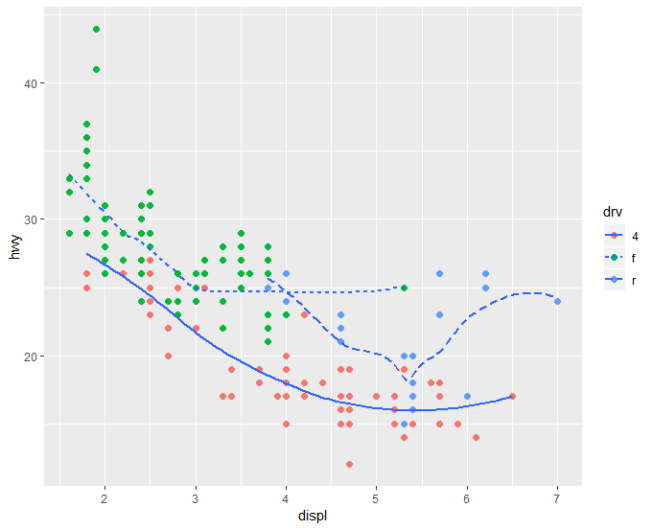
x, y: 글로벌 매핑
color: 로컬 매핑

```
> ggplot(data=mpg, mapping=aes(x=displ, y=hwy, color=drv))+
    geom_point()+
    geom_smooth(se=FALSE)
```

x, y, color: 글로벌 매핑

- 예: mpg의 변수 displ과 hwy의 비모수 회귀곡선 작성하되 drv에 의해 구분되는 그룹별 각각 추정하여 선의 종류를 다르게 표시. 그 위에 산점도 추가하되 drv의 값에 따라 점의 색을 구분, 점의 크기 확대.

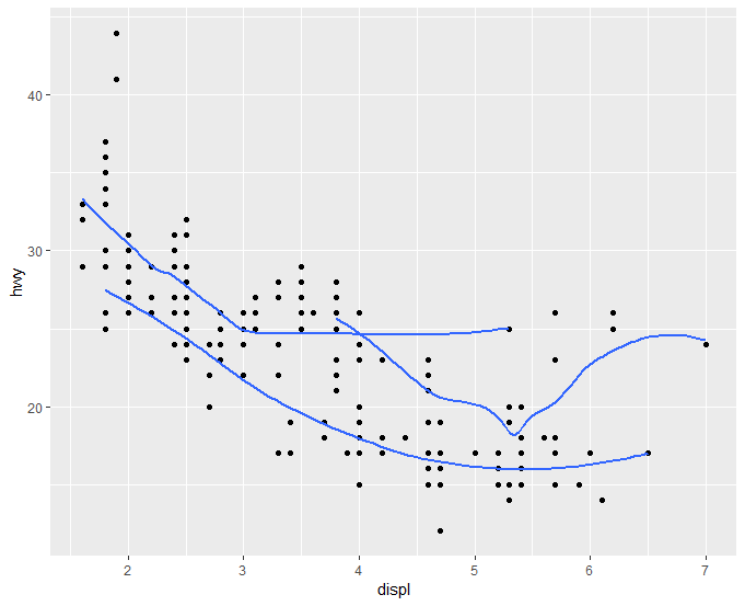
```
> ggplot(data=mpg, mapping=aes(x=displ, y=hwy)) +
    geom_point(mapping=aes(color=drv), size=2) +
    geom_smooth(mapping=aes(linetype=drv), se=FALSE)
```



linetype: 선의 종류를 나타내는 시각적 요소

- 예: 다음의 그래프 작성
  - 변수 drv의 그룹별로 따로 비모수 회귀곡선 작성하되, 선의 색과 종류는 같은 것을 사용

```
> ggplot(data=mpg, mapping=aes(x=displ, y=hwy)) +
     geom_point() +
     geom_smooth(mapping=aes(group=drv), se=FALSE)
```



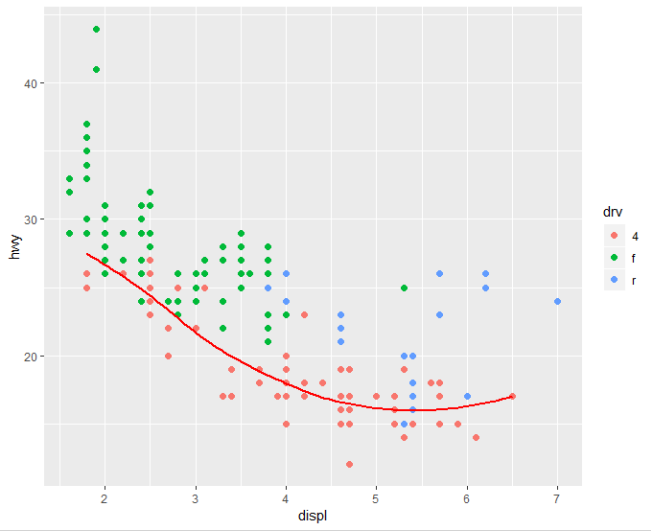
group: 그룹을 구성하는 시각적 요소

- 각 geom 함수에서 다른 데이터 사용

  - 각 geom 함수로 작성되는 layer마다 다른 데이터로 그래프 작성 가능

  - 예: mpg의 변수 displ과 hwy의 산점도. drv에 따라 점의 색 구분. 비모수 회귀곡선 추가하되 drv가 4인 데이터만을 대상으로 추정.

```
> ggplot(data=mpg, mapping=aes(x=displ, y=hwy)) +
    geom_point(mapping=aes(color=drv), size=2) +
    geom_smooth(data=filter(mpg, drv=="4"),
                se=FALSE, color="red")
```



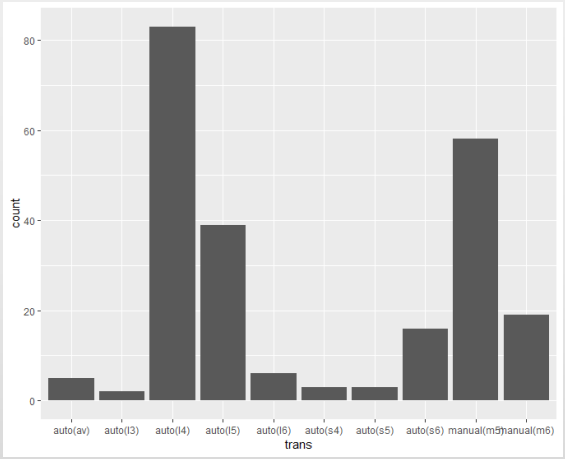
5.5 통계적 변환: Statistical transformation

- 그래프 작성에 사용되는 자료
  1) 입력된 자료: 산점도
  2) 입력된 자료를 대상으로 통계적 변환 과정을 거쳐 생성된 자료: 비모수 회귀곡선 그래프

- 통계적 변환(stat)
  - 입력된 데이터 프레임 자료의 변환을 의미
  - 각 그래프 유형별 대응되는 stat 존재
    - 산점도: stat="identity"
    - 비모수 회귀곡선: stat="smooth"
    - 막대 그래프: stat="count"
  - 각 geom 함수마다 대응되는 디폴트 stat 존재
    - geom_point( ) → geom_point(stat="identity")
    - geom_smooth( ) → geom_smooth(stat="smooth")
    - geom_bar( ) → geom_bar(stat="count")

- stat 함수
  - geom 함수 대신 stat 함수로 그래프 작성 가능
  - 각 stat 함수마다 디폴트 geom 존재
    - stat_identity( ) → stat_identity(geom="point")
    - stat_smooth( ) → stat_smooth(geom="smooth")
    - stat_count( ) → stat_count(geom="bar")

- 예: mpg의 변수 trans의 막대 그래프 작성




```
> ggplot(data=mpg, mapping=aes(x=trans)) +
    geom_bar()
> ggplot(data=mpg, mapping=aes(x=trans)) +
    stat_count()
```

- 그래프 작성은 geom 함수 또는 stat 함수 모두 가능
- 사용자가 선택할 사항
- geom 함수의 사용이 비교적 더 명확함

- stat으로 계산된 변수의 이용

  - stat 함수: 입력된 데이터 프레임을 대상으로 변환을 실시하여 그래프 작성에 필요한 변수로 이루어진 데이터 프레임을 내부적으로 생성

  - 생성된 변수를 사용자가 직접 지정해서 사용 가능

  - 예: 함수 geom_bar( ) 혹은 stat_count( )에서 계산된 변수
    - 변수 count: 각 범주의 빈도
    - 변수 prop: 그룹별 비율

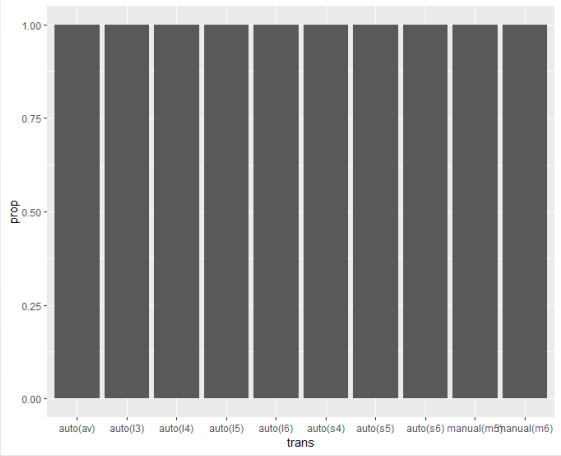
  - 계산된 변수를 사용자가 지정할 때에는 변수를 ".." 기호로 감싸야 함
    - 원래 데이터 프레임에 있는 변수와 혼동 방지
    - 예: `..count..` 또는 `..prop..`

  - 최근에는 ".." 기호로 감싸는 방법 대신 함수 stat( ) 사용
    - `stat(count)` 또는 `stat(prop)`

- 예: mpg의 변수 trans의 막대 그래프를 상대 도수로 작성

  - 변수 stat(prop)를 이용하여 막대 그래프 작성

```
> ggplot(data=mpg) +
    geom_bar(mapping=aes(x=trans, y=stat(prop)))
```



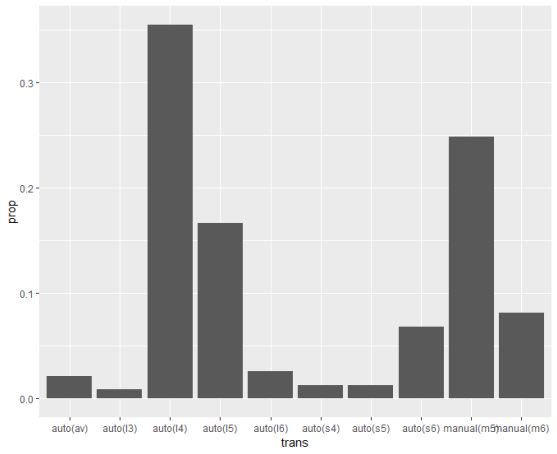
- stat_count( )에서 생성된 자료

	trans	count	prop	group
1	auto(av)	5	1	1
2	auto(l3)	2	1	2
3	auto(l4)	83	1	3
4	auto(l5)	39	1	4
5	auto(l6)	6	1	5
6	auto(s4)	3	1	6
7	auto(s5)	3	1	7
8	auto(s6)	16	1	8
9	manual(m5)	58	1	9
10	manual(m6)	19	1	10

- 변수 stat(prop) : 그룹별 비율
- 모든 범주를 하나의 그룹으로 구성

- 상대 도수 막대 그래프 작성

```
> ggplot(data=mpg) +
    geom_bar(mapping=aes(x=trans, y=stat(prop), group=1))
```



- group에 하나의 값 지정
- 어떤 값도 가능

	trans	count	prop	group
1	auto(av)	5	0.021367521	1
2	auto(l3)	2	0.008547009	1
3	auto(l4)	83	0.354700855	1
4	auto(l5)	39	0.166666667	1
5	auto(l6)	6	0.025641026	1
6	auto(s4)	3	0.012820513	1
7	auto(s5)	3	0.012820513	1
8	auto(s6)	16	0.068376068	1
9	manual(m5)	58	0.247863248	1
10	manual(m6)	19	0.081196581	1

- geom 함수에서 stat을 따로 지정해야 하는 경우

  - geom 함수의 디폴트 stat이 아닌 다른 stat을 사용해야 하는 경우
  - 예: 도수분포표로 막대 그래프를 작성하는 경우

- mpg의 trans를 auto(av)에서 auto(s6)를 auto로, manual(m5)와 manual(m6)를 manual로 통합. 도수분포표 작성. 막대 그래프 작성.

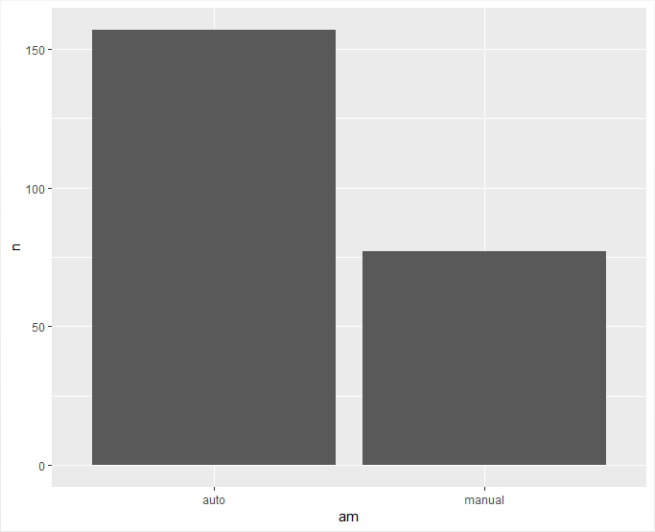
  - 통합된 범주의 도수분포 tibble 작성

```
> mpg_am <- mpg %>%
    mutate(am=substr(trans, 1, nchar(trans)-4)) %>%
    count(am)
```

```
> mpg_am
# A tibble: 2 x 2
  am          n
  <chr>   <int>
1 auto      157
2 manual     77
```

- 도수분포표로 막대 그래프 작성

```
> ggplot(data=mpg_am) +
    geom_bar(mapping=aes(x=am, y=n), stat="identity")
```




```
> mpg_am
# A tibble: 2 x 2
  am          n
  <chr>   <int>
1 auto      157
2 manual     77
```

5.6 위치 조정: Position adjustments

- 그래프 요소들의 위치 조정
  - 연속형 자료: 산점도의 점이 겹쳐지는 경우
  - 범주형 자료: 이변량 막대 그래프 작성

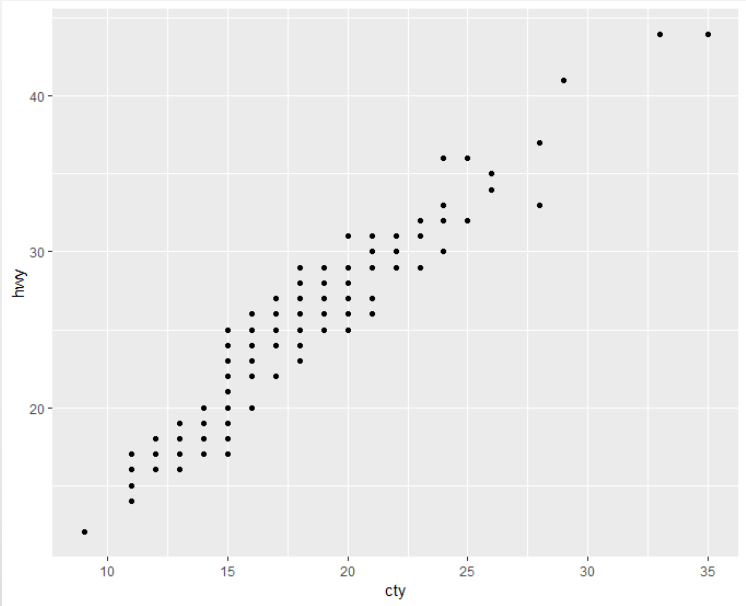
- 산점도의 점이 겹치는 문제
  - 산점도 작성의 가장 큰 문제
  - 해결 방안
    - 반올림 처리 등으로 같은 값이 많은 자료의 경우: 자료에 약간의 난수 추가로 점의 위치 조정(jittering)
    - 대규모의 자료가 좁은 구역에 몰려서 한 무리를 형성하는 경우: 추후에 다룰 예정

- 이변량 막대 그래프
  - 쌓아 올린 막대 그래프 / 옆으로 붙여 놓은 막대 그래프

- 산점도에서 점이 겹쳐지는 문제 해결

  - 예: mpg에서 변수 cty와 hwy의 산점도 작성

```
> ggplot(data=mpg, mapping=aes(x=cty, y=hwy)) +
    geom_point()
```



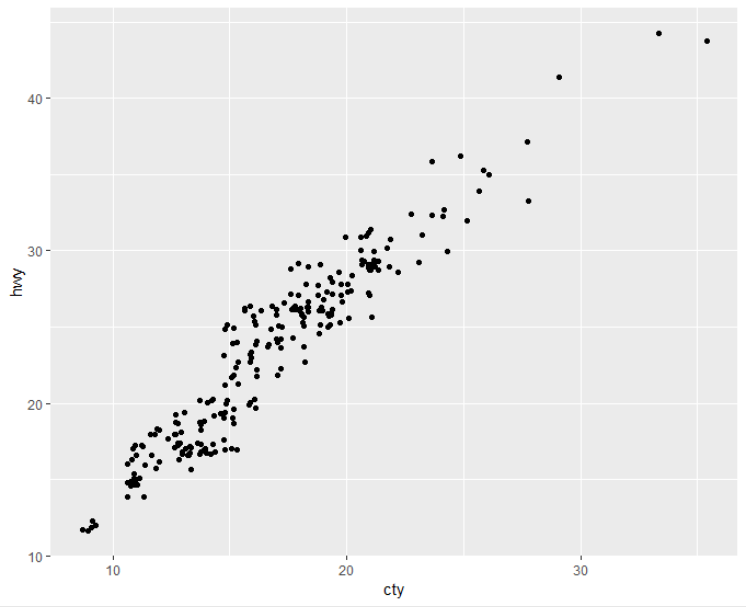
- 산점도에 나타난 점의 개수가 전체 데이터 개수인 234개에 훨씬 못 미쳐 보임

- 두 변수의 값이 반올림 처리되어 같은 값이 많아짐

jittering
- 자료에 약간의 난수 추가
- $(x, y) \rightarrow (x + \varepsilon,\ y + \varepsilon)$
  $\varepsilon \sim Unif(-\alpha, \alpha)$

- jittering 실시

```
> ggplot(data=mpg, mapping=aes(x=cty, y=hwy)) +
    geom_point(position="jitter")
```

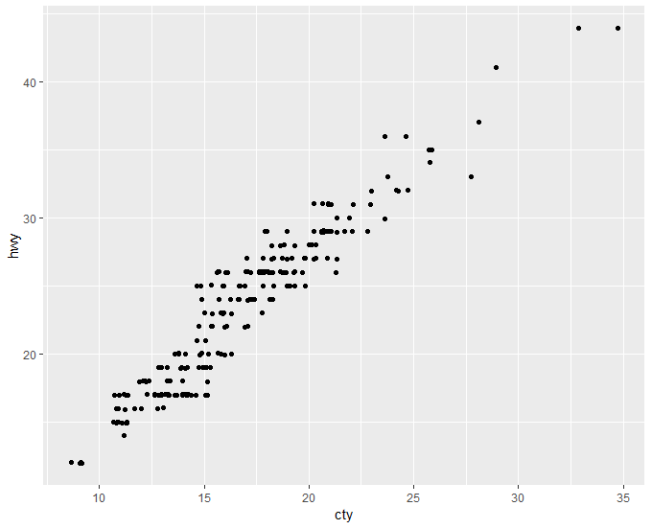


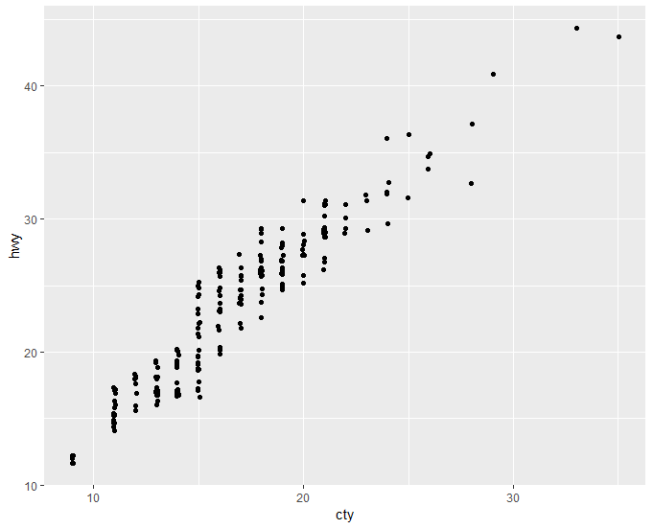
- 작성되는 그래프마다 미세한 차이 발생
- 추가되는 난수의 크기를 조절하고자 하는 경우에는 함수 geom_jitter( ) 사용

- 함수 geom_jitter()

```
> ggplot(data=mpg, mapping=aes(x=cty, y=hwy))+
    geom_jitter(width=0.4, height=0.05)
```

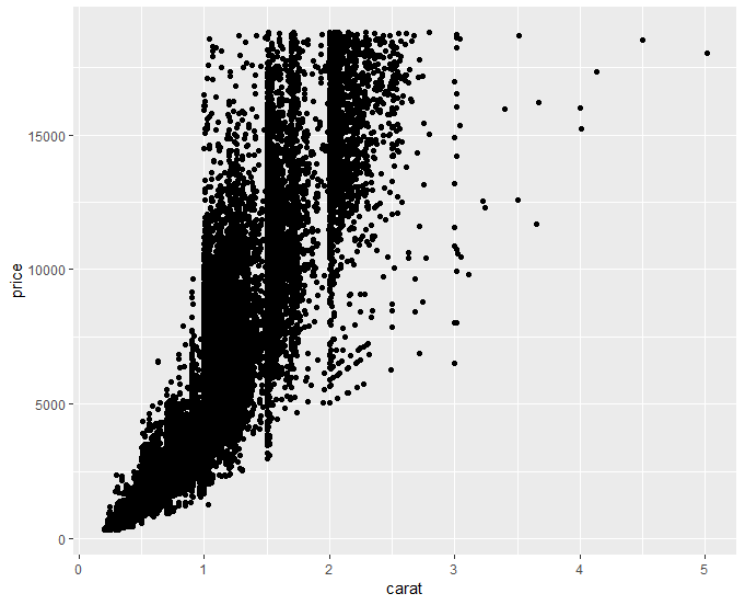
```
> ggplot(data=mpg, mapping=aes(x=cty, y=hwy))+
    geom_jitter(width=0.05, height=0.4)
```





- 예: diamonds에서 변수 carat과 price의 산점도

```
> ggplot(data=diamonds) +
    geom_point(mapping=aes(x=carat, y=price))
```



- 너무 많은 점들이 밀집되어 있는 상황
- 두 변수의 정확한 관계 파악이 어려운 그래프
- Jittering으로는 문제 해결이 불가능
- 다른 방법이 필요함(9장에서 살펴볼 예정)

- 이변량 막대 그래프 작성
  - 막대 그래프 작성: geom_bar( )
  - 이변량 막대 그래프: 함수 geom_bar( )에 시각적 요소 x와 fill, position 사용

• 예제: mpg에서 trans의 범주를 auto와 manual로 통합한 변수 am 생성
변수 cyl이 5인 케이스 제거 후 am과 cyl의 이변량 막대 그래프 작성

- 자료 준비

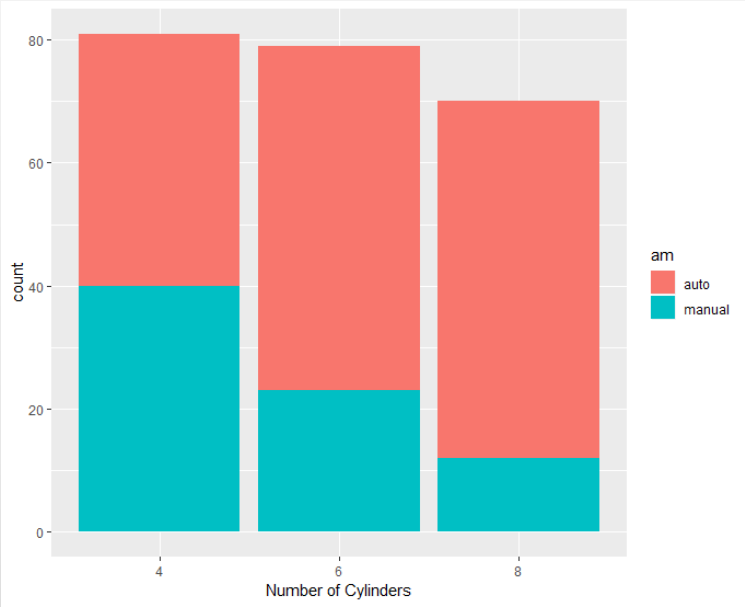
```
> mpg_1 <- mpg %>%
    mutate(am=substr(trans,1,nchar(trans)-4)) %>%
    filter(cyl!=5)
```

- 쌓아 올린 막대 그래프와 옆으로 붙여 놓은 막대 그래프 작성

```
> p_1 <- ggplot(data=mpg_1,
          mapping=aes(x=as.factor(cyl), fill=am)) +
    xlab("Number of Cylinders")
```

디폴트
position="stack"

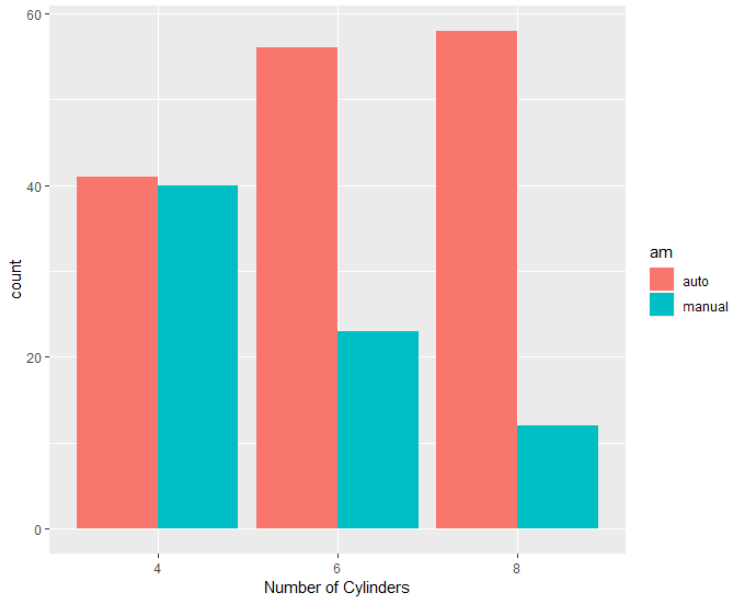
```
> p_1 + geom_bar()
```



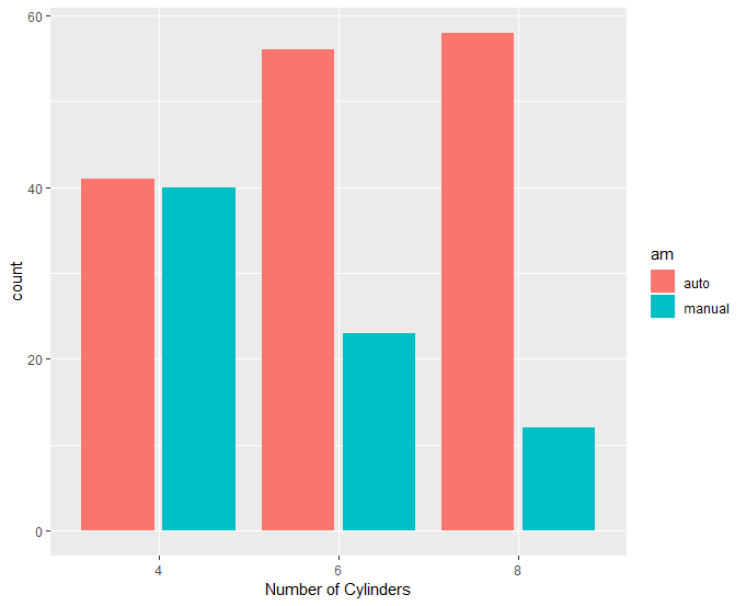
ggplot(mpg_1, aes(x=cyl, fill=am))을
실행하면 어떤 문제가 있는가?

- 옆으로 붙여 놓은 막대 그래프

```
> p_1 + geom_bar(position="dodge")
```

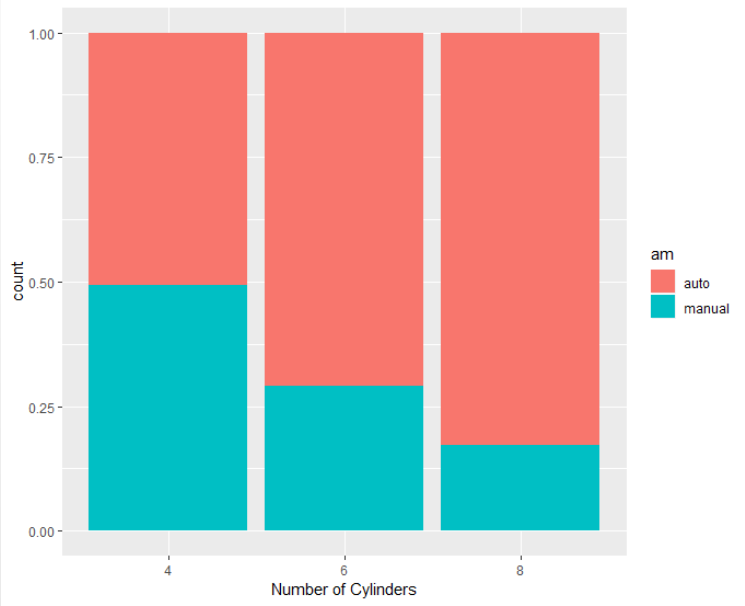



```
> p_1 + geom_bar(position="dodge2")
```



- 조건부 확률로 쌓아 올린 막대 그래프

```
> p_1 + geom_bar(position="fill")
```



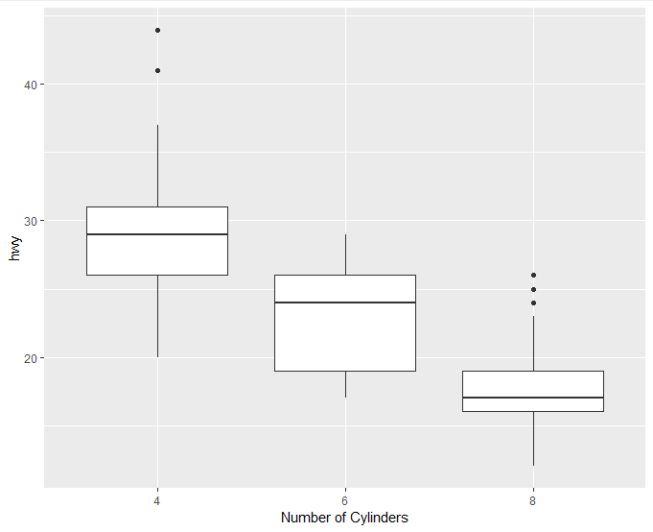
cyl을 조건으로 하는 cyl과 am의 조건부 확률

```
    am
cyl      auto    manual
  4 0.5061728 0.4938272
  6 0.7088608 0.2911392
  8 0.8285714 0.1714286
```

- 나란히 서 있는 상자그림

  - geom_boxplot( )
  - 필요한 시각적 요소: x=그룹을 구성하는 변수(요인)
    y=연속형 변수

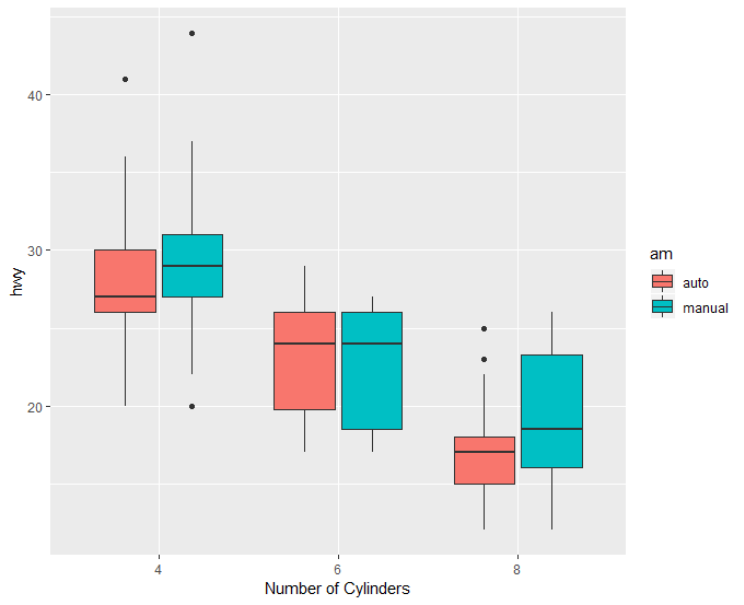
```
> ggplot(data=mpg_1, mapping=aes(x=as.factor(cyl), y=hwy)) +
    geom_boxplot() +
    xlab("Number of Cylinders")
```



- 시각적 요소 x에 요인이 아닌 숫자형 변수인 cyl을 매핑하면 어떤 그래프가 작성되는가?

- 그룹을 구성하는 변수가 두 개인 경우의 상자그림

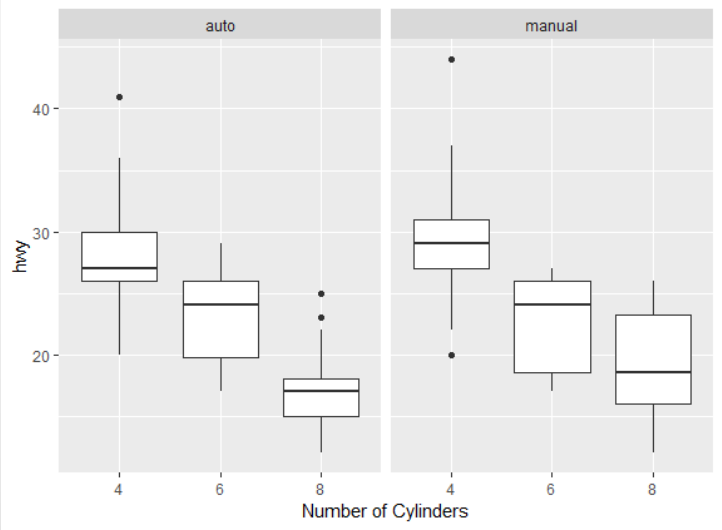
```
> ggplot(data=mpg_1, mapping=aes(x=as.factor(cyl), y=hwy)) +
    geom_boxplot(mapping=aes(fill=am)) +
    xlab("Number of Cylinders")
```



- 변수 am에 따라 다른 색이 채워져 있고 두 상자그림이 옆에 붙어 있다
- 필요한 시각적 요소: x, y, fill, position
- position="dodge"가 디폴트로 적용됨

- 그룹을 구성하는 변수가 두 개인 경우의 상자그림

```
> ggplot(data=mpg_1, mapping=aes(x=as.factor(cyl), y=hwy)) +
    geom_boxplot() +
    xlab("Number of Cylinders") +
    facet_wrap(~am)
```



5.7 좌표계: Coordinate system

- 좌표계: 시각적 요소 x와 y를 근거로 그래프의 각 요소의 2차원 위치를 결정하는 체계

- 좌표계의 종류
  - coord_cartesian( ): 디폴트
  - coord_flip( )
  - coord_polar( )

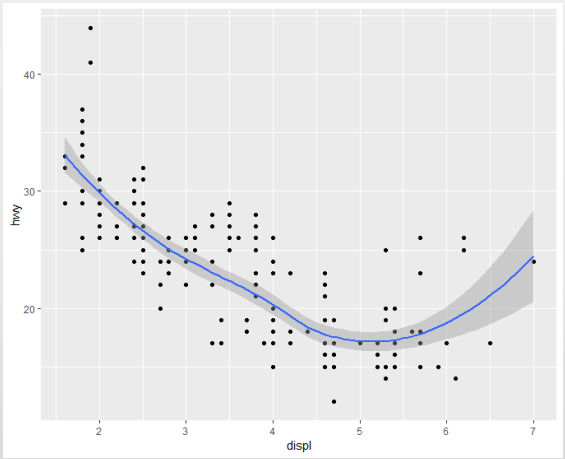
- coord_cartesian( )의 활용: XY축 범위 조정

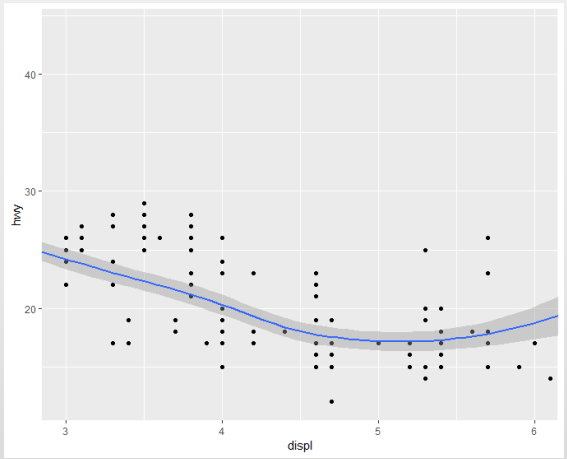
  - 예: mpg에서 displ과 hwy의 산점도에 비모수 회귀곡선 추가한 그래프 작성. X축의 범위를 (3,6)으로 축소한 그래프 작성.

```
> p <- ggplot(data=mpg, mapping=aes(x=displ,y=hwy)) +
    geom_point() + geom_smooth()
```

```
> p
```

```
> p + coord_cartesian(xlim=c(3,6))
```





- scale에 의한 XY축 범위 조정

  - scale: 자료와 시각적 요소의 매핑 및 XY축과 범례 등의 내용 조정을 의미

  - 대부분의 경우 디폴트 상태에서 그래프 작성

  - XY축 범위 조정, XY축 라벨 변경이 필요한 경우에는 scale 함수를 사용하여 조정

  - scale 함수의 일반적인 형태: `scale_*1*_*2*()`
    - *1*: 수정하고자 하는 시각적 요소; color, x, y, fill 등등
    - *2*: 적용되는 scale 지칭; discrete, continuous 등등

  - 예: 연속형 X 변수의 범위 (3,6)으로 수정: scale_x_continuous(limits=c(3,6))
    연속형 X축의 라벨 변경: scale_x_continuous(name="Engine")

  - 간편 함수:
    - XY 축 범위 조정: xlim(3,6), ylim(0,1)
    - XY축 라벨 변경: xlab("Engine"), ylab( ), labs(x="Engine")

- 예: XY축 조정 비교

1) 함수 xlim( )에 의한 조정

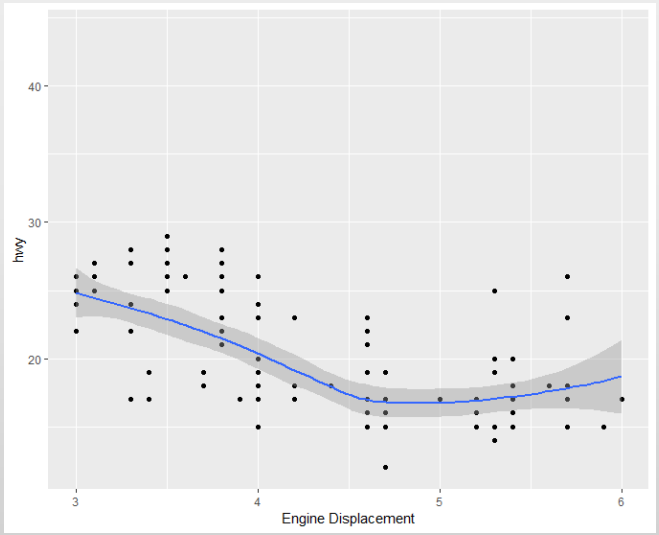
범위를 벗어난 자료: 삭제

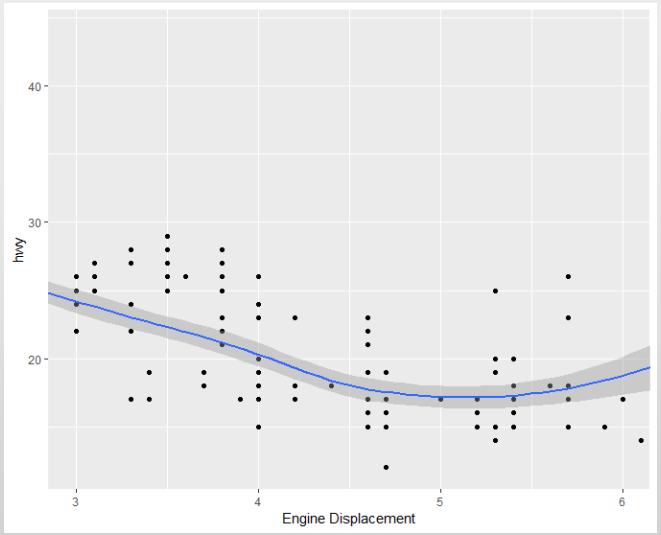
```
> p + xlim(3,6) + xlab("Engine Displacement")

Warning messages:
1: Removed 105 rows containing non-finite values (stat_smooth).
2: Removed 105 rows containing missing values (geom_point).
```

2) 함수 coord_cartesian( )에 의한 조정

```
> p + coord_cartesian(xlim=c(3,6)) +
    xlab("Engine Displacement")
```





- 함수 coord_flip( )의 활용: 평행한 상자그림 작성

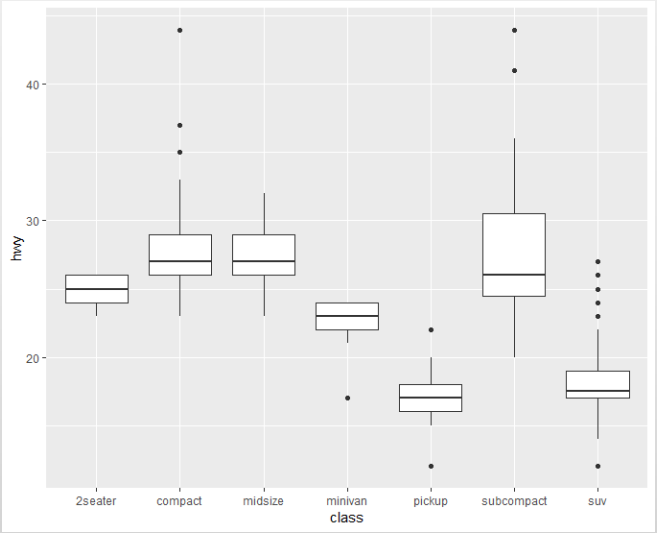
  - 대부분의 geom 함수: 주어진 x 값에 대한 y의 분포 표현
  - 상자그림: 수직 방향의 작성되는 것이 디폴트
  - 수평 방향 상자그림: 디폴트 방향으로 작성하고, 그래프의 좌표를 90° 회전
  - 함수 coord_flip( ): 작성된 그래프의 좌표 회전

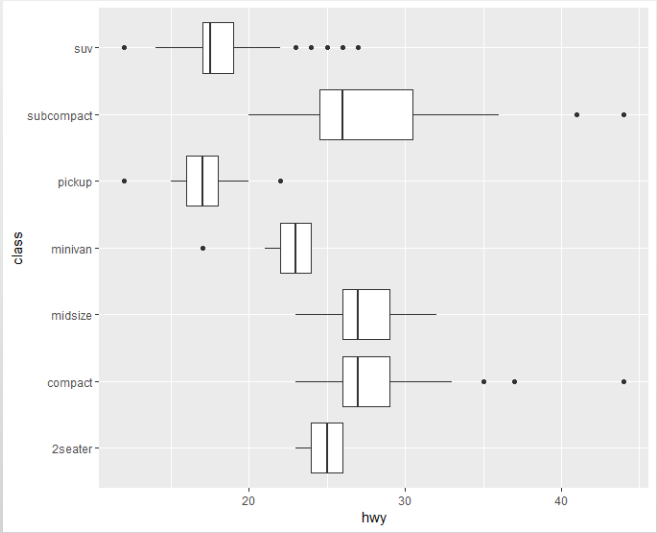
- 예: mpg에서 class의 그룹별로 hwy의 상자그림 작성

  - 상자그림: geom_boxplot( )
  - x 변수=class, y 변수=hwy

```
> ggplot(data=mpg, mapping=aes(x=class, y=hwy)) +
    geom_boxplot()
```

```
> ggplot(data=mpg, mapping=aes(x=class, y=hwy)) +
    geom_boxplot() +
    coord_flip()
```



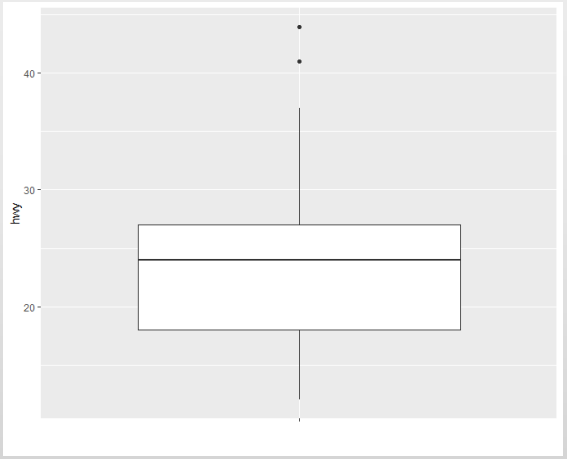


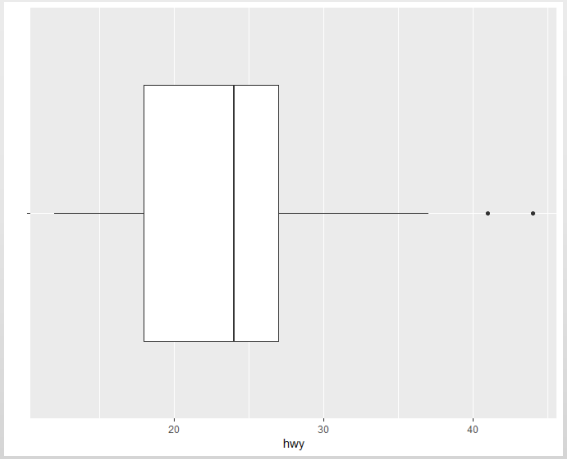
- 예: 한 변수(hwy)의 상자그림 작성

  - 함수 geom_boxplot( )에는 x와 y 모두 필요
  - x에는 하나의 값, y에는 연속형 변수 매핑

```
> ggplot(data=mpg, mapping=aes(x="", y=hwy)) +
    geom_boxplot() +
    xlab("")
```

```
> ggplot(data=mpg, mapping=aes(x="", y=hwy)) +
    geom_boxplot() +
    xlab("") +
    coord_flip()
```

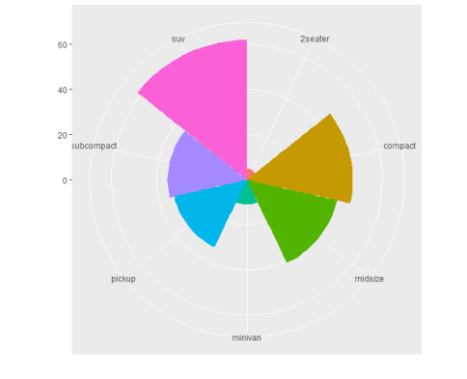


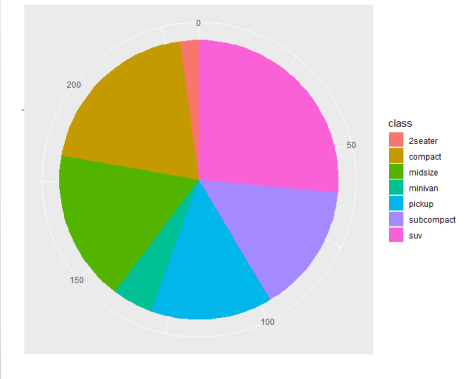


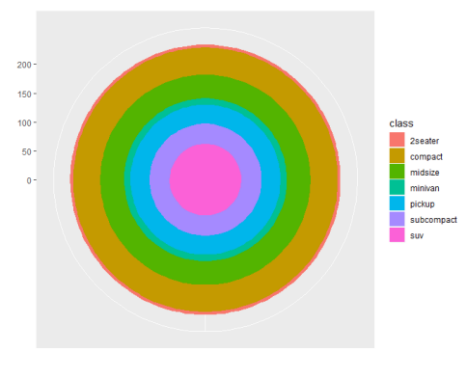
- 함수 coord_polar( )의 활용: 파이 그래프 작성

  - 극좌표(polar coordinate): 2차원 공간의 어느 한 점의 위치를 원점에서의 거리와 각도로 표현
  - 함수 coord_polar( ): 데카르트 좌표를 극좌표로 전환
  - 변수 theta: 시각적 요소 x와 y 중 각도로 전환할 요소 지정(디폴트는 theta="x")

  • 함수 coord_polar( )를 활용하여 막대 그래프에서 변형된 그래프
    1) Coxcomb 또는 Wind rose 그래프
    2) 파이 그래프
    3) Bullseye 그래프



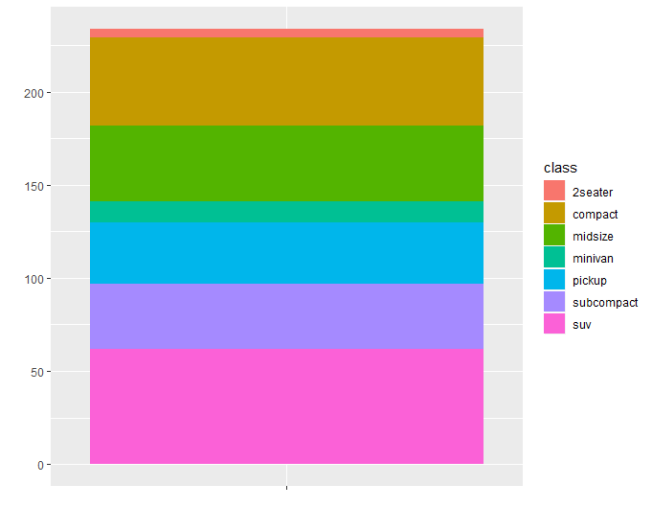




- 예제: mpg의 변수 class의 파이 그래프 작성

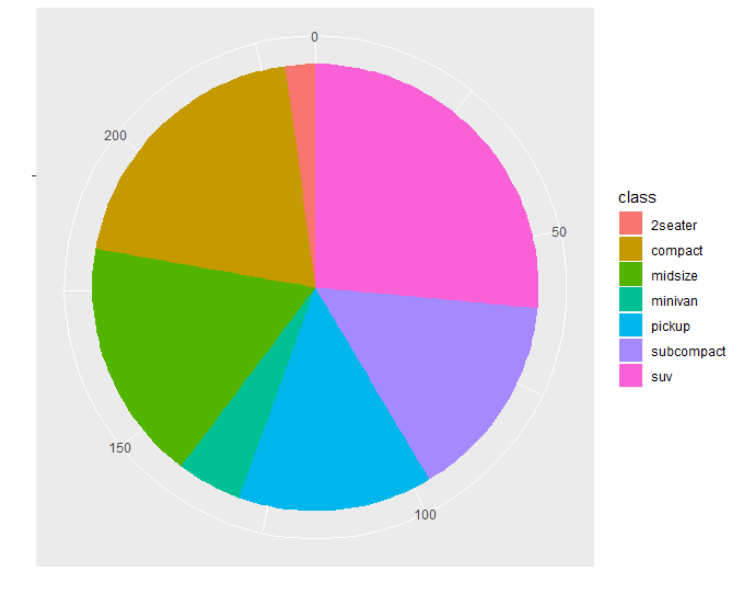
1) 막대 그래프 작성
   - 각 막대마다 다른 색 사용
   - 쌓아 올린 형태로 작성
   - 막대의 폭을 X축 전체 구간으로 확장

```
> b2 <- ggplot(data=mpg, mapping=aes(x="", fill=class)) +
    geom_bar(width=1) +
    labs(x="",y="")
> b2
```

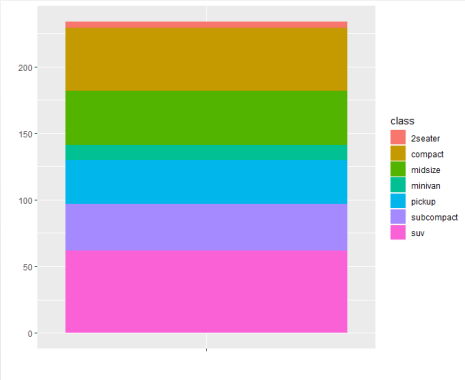


2) theta="y"로 함수 coord_polar( ) 실행: 파이 그래프 작성

```
> b2 + coord_polar(theta="y")
```



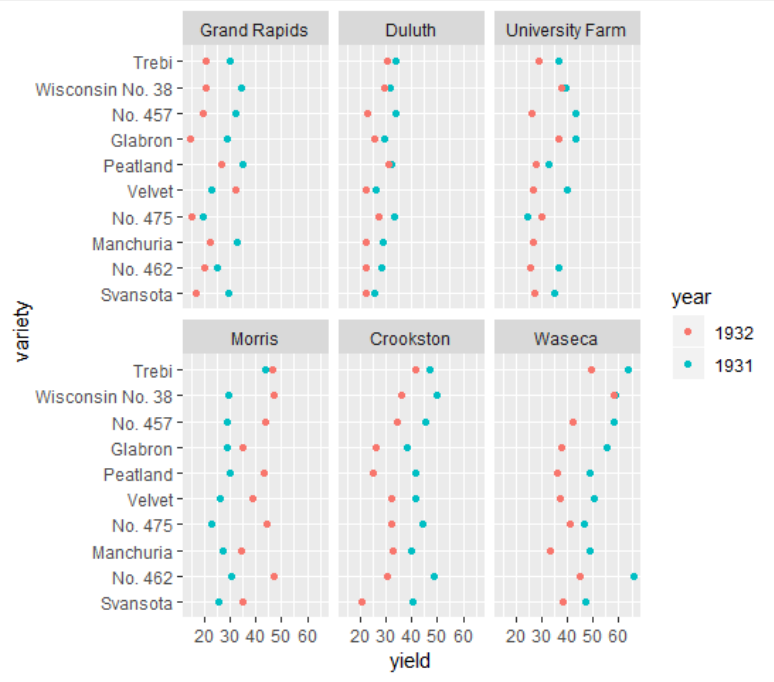
- 각 조각의 각도: 막대의 높이에 비례
- 각 조각의 반지름: 막대의 폭에 비례



- 연습문제

1. 패키지 `lattice`에 있는 데이터 프레임 `barley`는 미네소타 주 농경학자들이 보리 종류에 따른 수확량의 차이를 비교하기 위해 2년간 경작하여 얻은 자료이다. 설명변수로는 6군데 경작지(`site`), 10종류의 보리(`variety`), 경작 년도(`year`)이고 반응변수는 수확량(`yield`)이다.

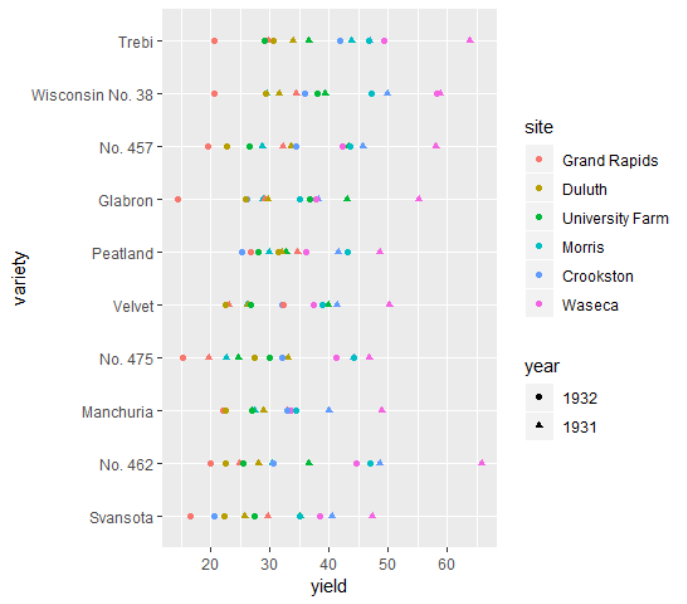
   1) 세 설명변수의 조합에 따른 수확량의 분포를 알아보는 다음의 그래프를 작성해 보자.



```
> data(barley, package="lattice")

> barley %>%
    ggplot(mapping=aes(x=yield, y=variety, color=year)) +
    geom_point() +
    facet_wrap(~ site)
```

2) 보리 종류(variety)에 따른 수확량(yield)의 비교분석에서 경작지(site)와 년도(year)의 효과를 단순 반복으로 처리한 다음의 그래프를 작성해 보자.




```
> barley %>%
   ggplot(mapping=aes(x=yield, y=variety, color=site, shape=year)) +
   geom_point()
```

3) 각 보리종류(variety)의 평균 수확량을 계산하여 다음과 같이 크기 순으로 나타내자.

```
> barley %>%
    group_by(variety) %>%
    summarise(mean_yield=mean(yield)) %>%
    arrange(desc(mean_yield))
```

```
# A tibble: 10 x 2
   variety           mean_yield
   <fct>                  <dbl>
 1 Trebi                   39.4
 2 Wisconsin No. 38        39.4
 3 No. 457                 35.8
 4 No. 462                 35.4
 5 Peatland                34.2
 6 Glabron                 33.3
 7 Velvet                  33.1
 8 No. 475                 31.8
 9 Manchuria               31.5
10 Svansota                30.4
```

2. 데이터 프레임 mpg의 변수 hwy는 자동차의 고속도로 연비를 나타낸다. 범주형 변수인 fl(연료 종류), trans(변속기 종류)에 따른 hwy의 분포를 알아보자.

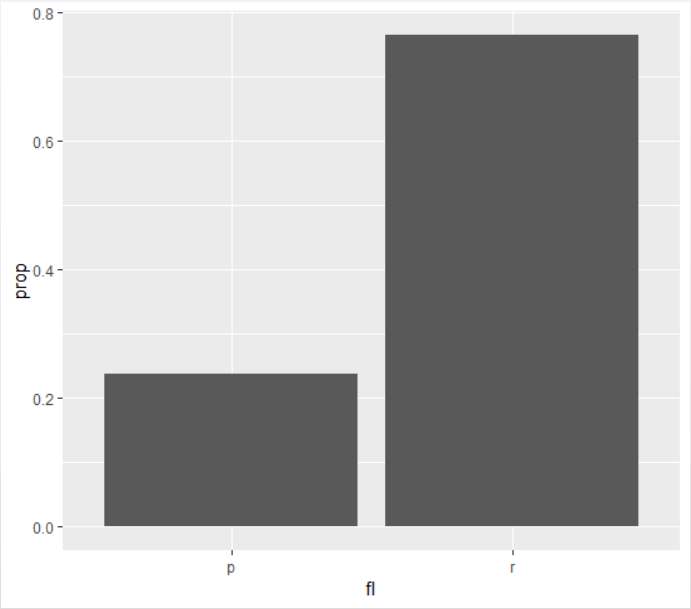
   1) 변수 fl의 종류별 빈도를 다음과 같이 구하라.

```
> mpg %>% count(fl)
```

```
# A tibble: 5 x 2
  fl        n
  <chr> <int>
1 c         1
2 d         5
3 e         8
4 p        52
5 r       168
```

2) 변수 fl에서 c(CNG), d(Diesel), e(Ethanol)는 제외하고 p(Premium)와 r(Regular)만을 대상으로 상대도수에 의한 막대 그래프를 작성해 보자.

```
> mpg_1 <- mpg %>% filter(fl %in% c("p","r"))
> mpg_1 %>%
    ggplot() +
    geom_bar(mapping=aes(x=fl, y=stat(prop), group=1))
```



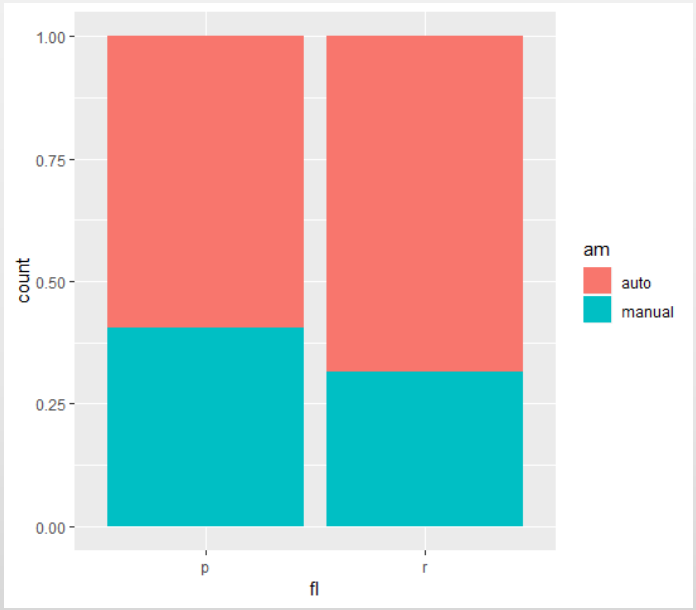
3) 2)에서 사용된 데이터 프레임을 대상으로 변수 trans의 종류별 빈도를 다음과 같이 구하라.

```
> mpg_1 %>% count(trans)
```

```
# A tibble: 10 x 2
   trans           n
   <chr>       <int>
 1 auto(av)        5
 2 auto(l3)        2
 3 auto(l4)       78
 4 auto(l5)       33
 5 auto(l6)        6
 6 auto(s4)        3
 7 auto(s5)        3
 8 auto(s6)       16
 9 manual(m5)     56
10 manual(m6)     18
```

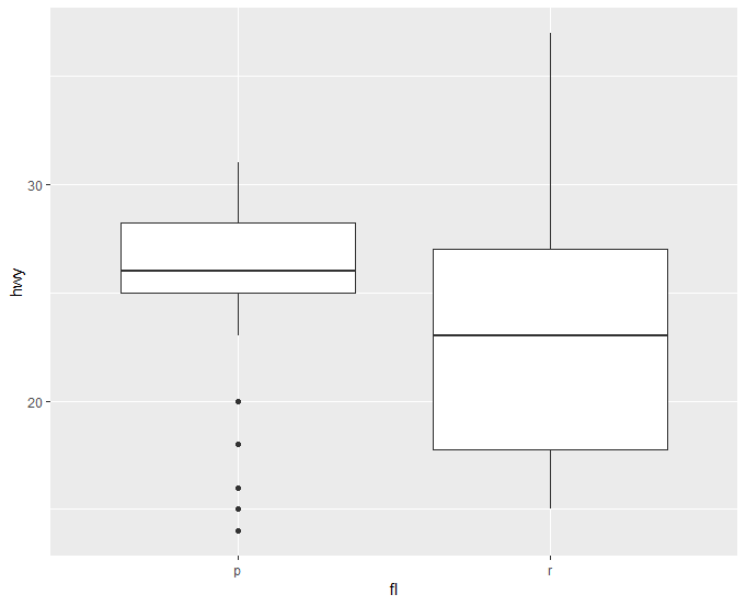
4) 2)에서 사용된 데이터 프레임을 대상으로 변수 trans의 범주를 'auto'와 'manual'로 통합한 변수 am을 생성하고 변수 fl과의 관계를 다음과 같은 막대 그래프로 나타내 보자.

```
> mpg_2 <- mpg_1 %>%
    mutate(am=substr(trans,1,nchar(trans)-4))
> mpg_2 %>%
    ggplot() +
    geom_bar(mapping=aes(x=fl,fill=am), position="fill")
```



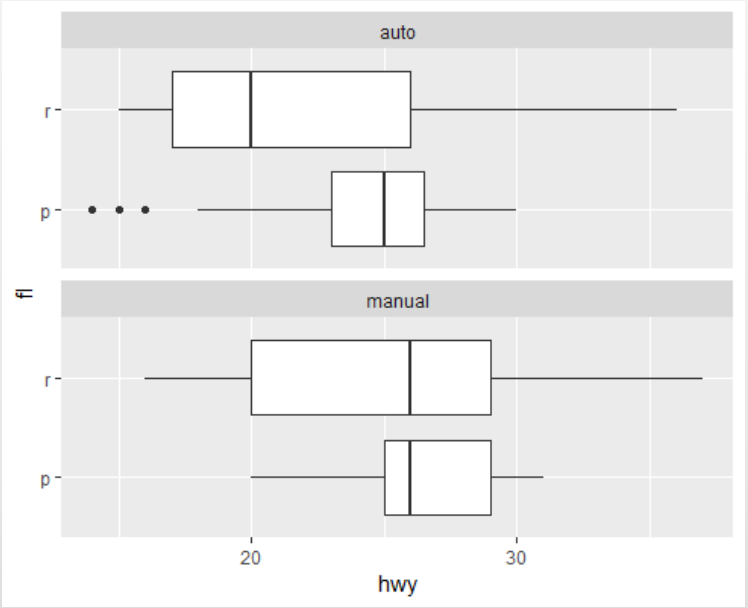
5) 2)에서 사용된 데이터 프레임을 대상으로 변수 fl에 따른 hwy의 분포를 상자그림으로 나타내 보자.

```
> mpg_2 %>%
    ggplot() +
    geom_boxplot(mapping=aes(x=fl,y=hwy))
```



6) 변수 am과 fl에 따른 hwy의 분포 비교를 위해 상자그림을 작성해 보자.

```
> mpg_2 %>%
    ggplot() +
    geom_boxplot(mapping=aes(x=fl,y=hwy)) +
    facet_wrap(~am, ncol=1) +
    coord_flip()
```

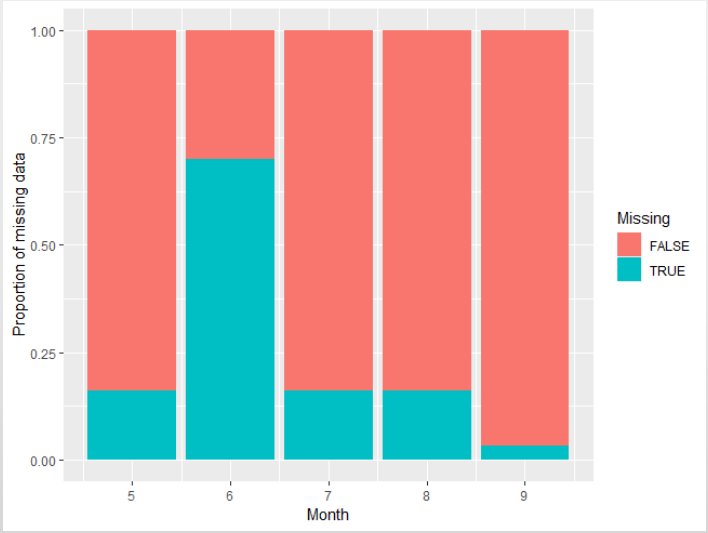


3. airquality

1) 월별 Ozone의 결측값 비율 그래프

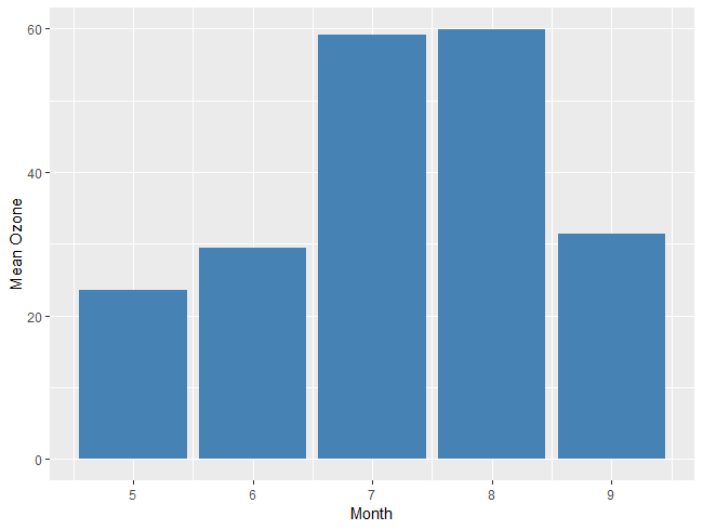
```
> airs <- as_tibble(airquality)

> airs %>%
    mutate(Missing=is.na(Ozone)) %>%
    ggplot(aes(x=Month, fill=Missing)) +
    geom_bar(position="fill") +
    labs(y="Proportion of missing data")
```



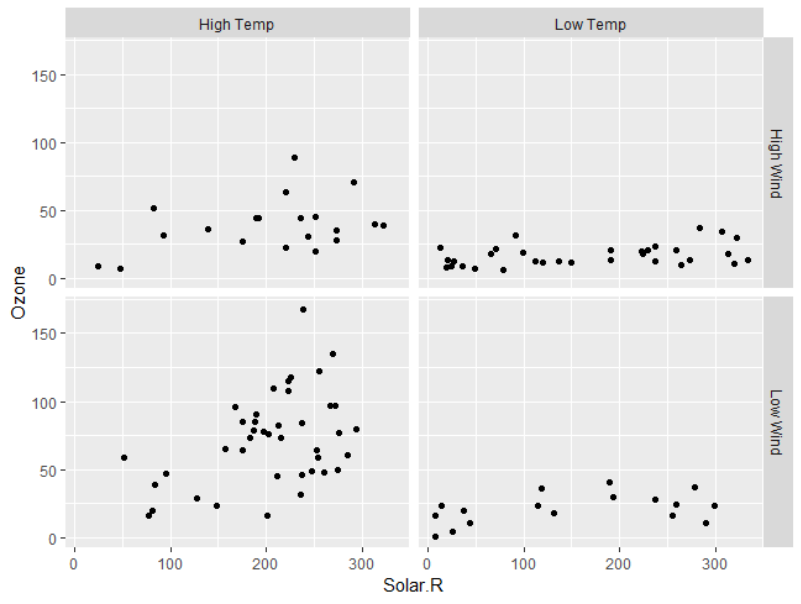
2) Ozone의 월별 평균값 막대 그래프

```
> airs %>%
    group_by(Month) %>%
    summarise(m.Oz=mean(Ozone, na.rm=TRUE)) %>%
    ggplot(aes(x=Month, y=m.Oz)) +
    geom_bar(stat="identity", fill="steelblue") +
    labs(y="Mean Ozone")
```



3) Ozone과 Solar.R의 그룹별 산점도
   - Wind의 값이 평균값 이상인 그룹과 이하인 그룹
   - Temp의 값이 평균값 이상인 그룹과 이하인 그룹

```
> airs %>%
    mutate(gp_Wind=if_else(Wind>=mean(Wind),
                           "High Wind","Low Wind"),
           gp_Temp=if_else(Temp>=mean(Temp),
                           "High Temp","Low Temp")) %>%
    ggplot(aes(x=Solar.R, y=Ozone)) +
    geom_point() +
    facet_grid(gp_Wind~gp_Temp)
```



4. mtcars

1) 변수 mpg와 disp, wt와 행의 이름을 변수로 변환한 변수 model만으로 이루어진 mtcars_t를 생성하고 처음 5개 케이스를 다음과 같이 출력하라. 단, 변수 mpg의 값이 가장 작은 자동차부터 출력하되 mpg의 값이 같은 케이스는 변수 disp의 값이 큰 자동차가 먼저 나타나도록 하자.

```
> mtcars_t <- mtcars %>%
    rownames_to_column(var="model") %>%
    as_tibble() %>%
    select(model,mpg,disp,wt) %>%
    arrange(mpg, desc(disp)) %>%
    print(n=5)
```

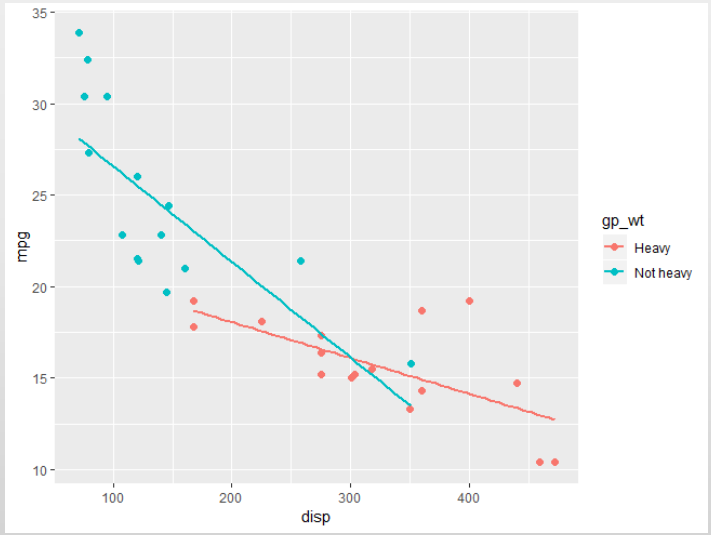
```
# A tibble: 32 x 4
  model                 mpg  disp    wt
  <chr>               <dbl> <dbl> <dbl>
1 Cadillac Fleetwood   10.4   472  5.25
2 Lincoln Continental  10.4   460  5.42
3 Camaro Z28           13.3   350  3.84
4 Duster 360           14.3   360  3.57
5 Chrysler Imperial    14.7   440  5.34
# ... with 27 more rows
```

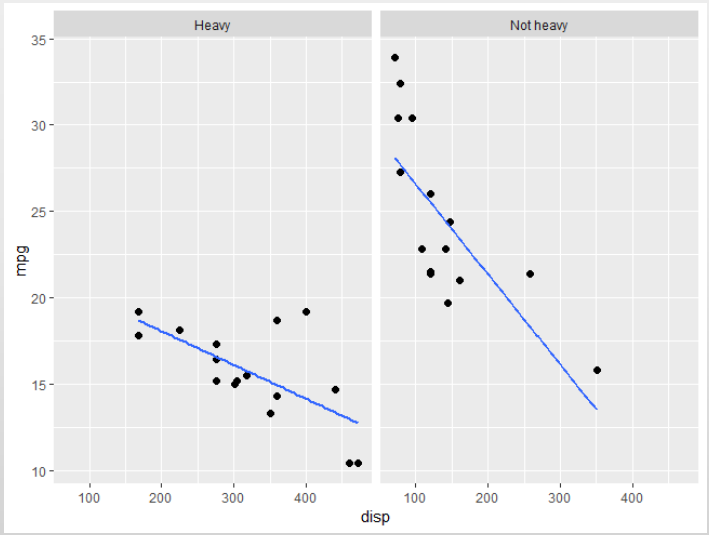
2) disp와 mpg의 그룹별 산점도 작성
   - wt가 중앙값 이상인 그룹과 이하인 그룹

```
> p1 <- mtcars_t %>%
    mutate(gp_wt=if_else(wt>=median(wt),"Heavy","Not heavy")) %>%
    ggplot(aes(x=disp, y=mpg))
```

```
> p1 + geom_point(aes(color=gp_wt), size=2) +
    geom_smooth(aes(color=gp_wt), se=FALSE, method="lm")
```

```
> p1 + geom_point(size=2) + geom_smooth(se=FALSE, method="lm") +
    facet_wrap(~gp_wt)
```





5. mpg

1) 다음의 조건을 만족하는 데이터 프레임 mpg_1을 생성하고, 처음 다섯 케이스를 출력해 보자.

- 변수 cyl이 4인 케이스만 선택하고, 변수 year는 요인으로 변환
- 변수는 model, year, displ, cty, hwy만으로 이루어져 있으며, 케이스는 year가 작은 값부터 배열하되, year의 값이 같은 케이스는 displ이 큰 값부터 배열하고, 또한 displ의 값이 같은 케이스는 cty가 작은 값부터 배열해 보자.

```
# A tibble: 81 x 5
  model              year  displ   cty   hwy
  <chr>              <fct> <dbl> <int> <int>
1 4runner 4wd        1999    2.7    15    20
2 toyota tacoma 4wd 1999    2.7    15    20
3 4runner 4wd        1999    2.7    16    20
4 toyota tacoma 4wd 1999    2.7    16    20
5 forester awd       1999    2.5    18    25
# ... with 76 more rows
```

```
> mpg_1 <- mpg %>%
    filter(cyl==4) %>%
    mutate(year=factor(year)) %>%
    select(model, year, displ, cty, hwy) %>%
    arrange(year, desc(displ), cty) %>%
    print(n=5)
```

2) 변수 displ이 2.0 이상인 그룹과 미만인 그룹으로 구분하고 각 그룹별로 다른 색을 사용하여 변수 cty와 hwy의 산점도를 작성하되, year에 따라 두 그룹으로 구분하여 따로 작성하고 위아래로 배치해 보자. 단, 두 변수는 같은 값은 갖는 케이스가 많기 때문에 jittering을 실시하고 산점도를 작성해야 한다.

```
> p2 <- mpg_1 %>%
    mutate(gp_displ=if_else(displ>=2.0, "Large","Small")) %>%
    ggplot()
```

```
> p2 + geom_jitter(aes(x=cty, y=hwy, color=gp_displ), size=2) +
    facet_wrap(~year, ncol=1)
```

3) 변수 displ이 2.0 이상인 그룹과 미만인 그룹으로 구분하여 변수 hwy의 상자그림을 작성하되, year에 따라 두 그룹으로 구분하여 따로 작성하고 옆으로 배치해 보자

```
> p2 + geom_boxplot(aes(x=gp_displ, y=hwy)) +
    facet_wrap(~year)
```