M23 - More on ggplot

LearningSpoonsR

2018-10-21

GG? - Grammar of Graphics

- Motivation

1. 그래픽스에 대한 원리가 없다면, 그래픽 관련 패키지와 함수는 단지 특수 경우의 모음일 뿐
2. 요리 백과사전을 다 읽는 것 vs. 물과 기름과 불의 작용에 대해서 익히고 백과사전을 **찾아가면서** 요리하는 것

- Advantage

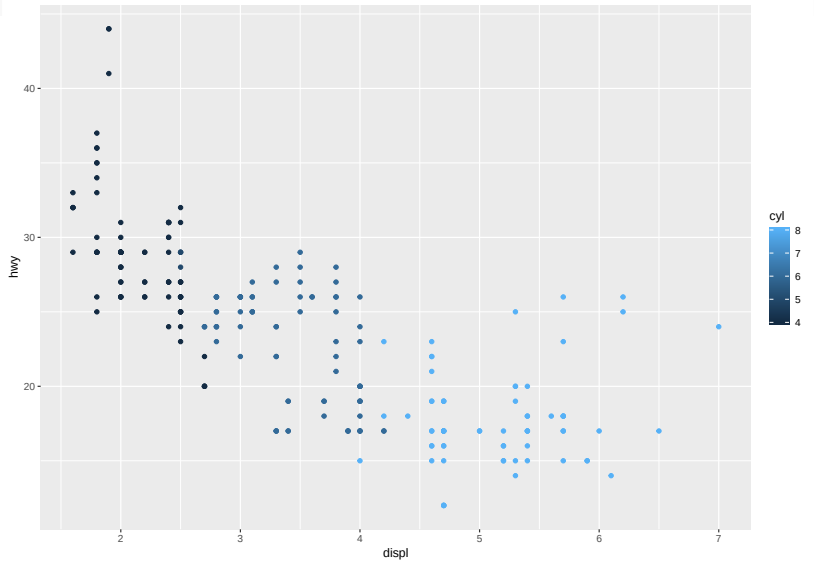
1. 새로운 package나 함수의 등장을 빠르게 흡수
2. 새로운 graphics를 만들어 내는 아이디어가 체계적이 됨

- Features

1. 독립적이고 더할 수 있는 구성 요소들로 그래픽을 표현
2. 개발과정에서 그래프의 특징을 한 가지 씩, 반복적으로 바꾸면서 그래프를 만들어 감
3. 생각의 흐름, 스토리텔링의 흐름과 연계시킬 수 있기에 interactive graphics와 잘 조화됨

구성 요소

```
library(ggplot2)
ggplot(mpg) +
  aes(x = displ, y = hwy, color = cyl) +
  geom_point()
```



- Aesthetics

1. position
2. size
3. color
4. shape

- Geometric Object (**geom_**)

1. Scatterplot - **point**
2. Bubblechart - **point** (size)
3. Barchart - **bar** (frequency)
4. Box-and-whisker plot - **boxplot** (distribution)
5. Line chart - **line**

Behind the scene

data

```
head(mpg[,c("displ", "hwy", "cyl")],10)

## # A tibble: 10 x 3
##    displ   hwy   cyl
##    <dbl> <int> <int>
##  1  1.80    29     4
##  2  1.80    29     4
##  3  2.00    31     4
##  4  2.00    30     4
##  5  2.80    26     6
##  6  2.80    26     6
##  7  3.10    27     6
##  8  1.80    26     4
##  9  1.80    25     4
## 10  2.00    28     4
```

aes

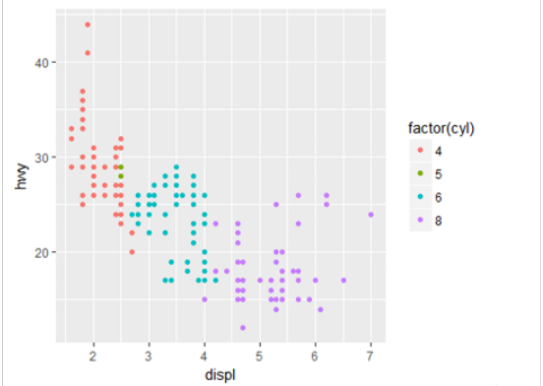
x	y	colour
1.8	29	4
1.8	29	4
2.0	31	4
2.0	30	4
2.8	26	6
2.8	26	6
3.1	27	6
1.8	26	4

geom

x	y	colour	size	shape
0.037	0.531	#F8766D	1	19
0.037	0.531	#F8766D	1	19
0.074	0.594	#F8766D	1	19
0.074	0.562	#F8766D	1	19
0.222	0.438	#00BFC4	1	19
0.222	0.438	#00BFC4	1	19
0.278	0.469	#00BFC4	1	19
0.037	0.438	#F8766D	1	19

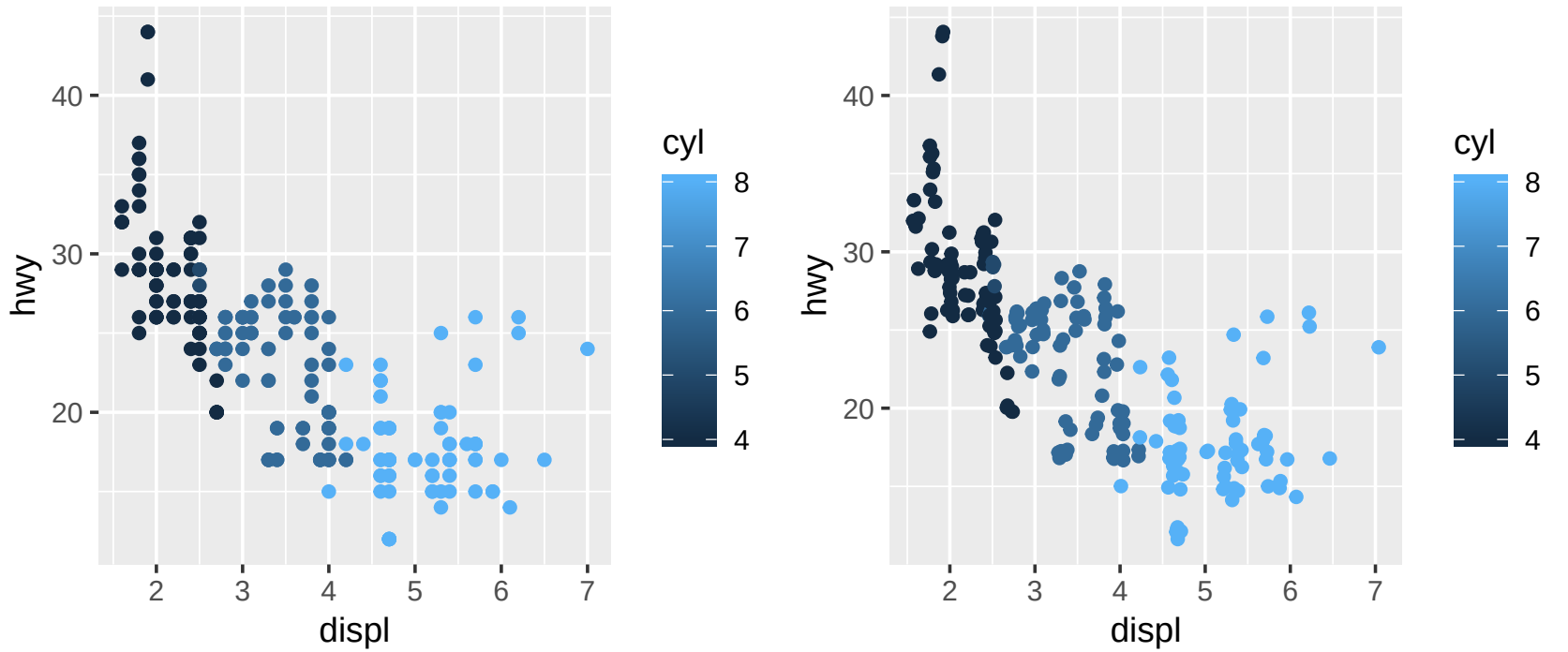
plot

```
library(ggplot2)
ggplot(data = mpg, aes(x = displ, y = hwy, color = factor(cyl))) +
  geom_point()
```



중첩된 관찰값에 노이즈를: position = "jitter"

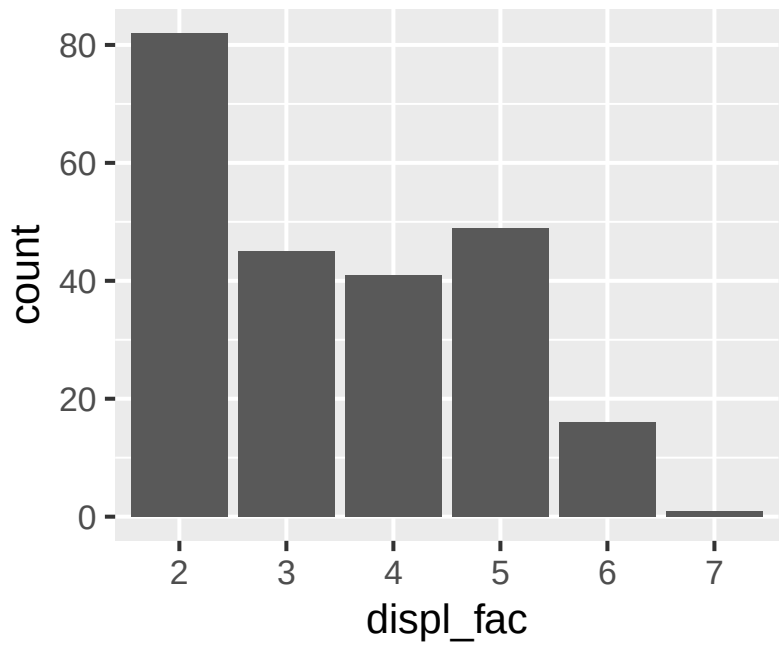
```
library(gridExtra)
a <- ggplot(mpg) + geom_point(aes(displ, hwy, color = cyl))
b <- ggplot(mpg) + geom_point(aes(displ, hwy, color = cyl), position = "jitter")
grid.arrange(a, b, nrow = 1, ncol = 2)
```



Barchart

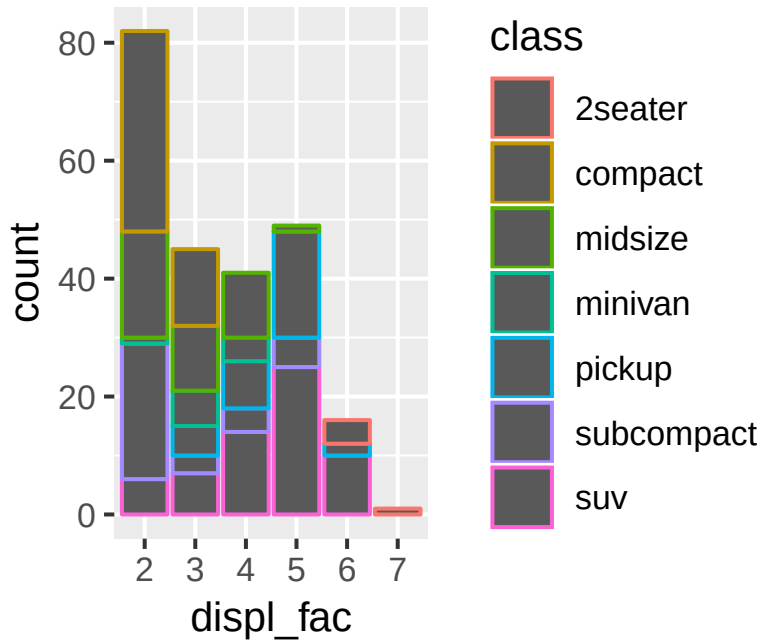
- x축에 이산변수 (discrete value)를 넣고 변수 x의 각각의 값에 대해서 몇 개의 관찰값이 있는지를 보여줌.
- #count #density #distribution

```
mpg$displ_fac <-
  as.factor(round(mpg$displ,0))
ggplot(mpg) +
  geom_bar(aes(x = displ_fac))
```



- x변수 외에도 이산 변수를 추가할 수 있음.

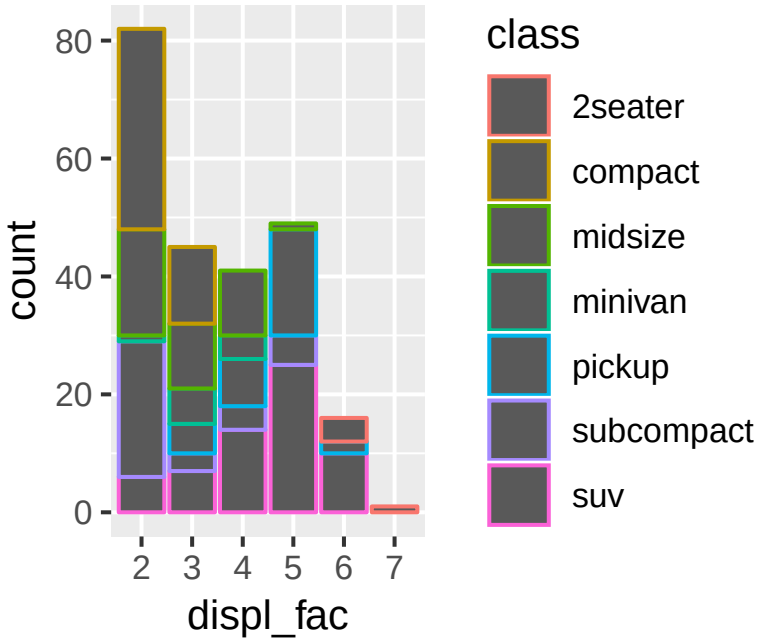
```
ggplot(mpg) +
  geom_bar(aes(x = displ_fac, color = class))
```



Barchart (Color and Fill)

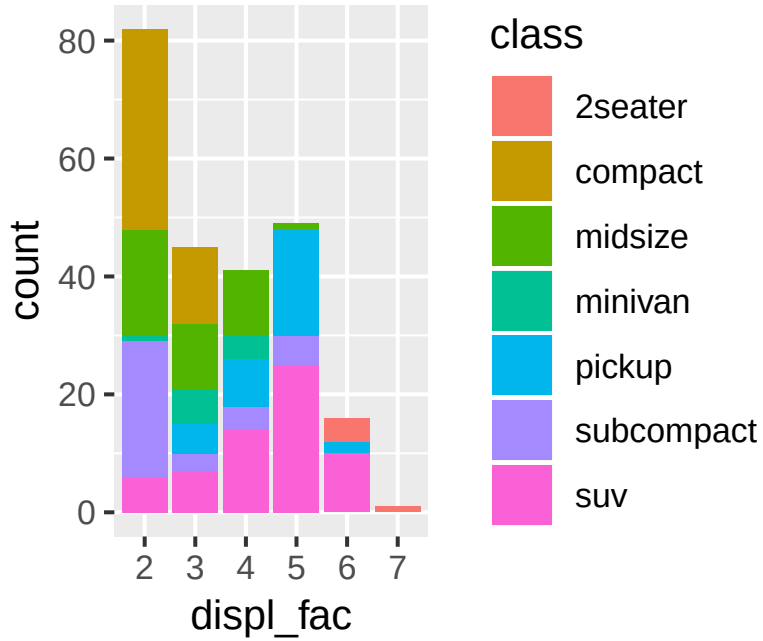
- `color` - 테두리만 됨

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



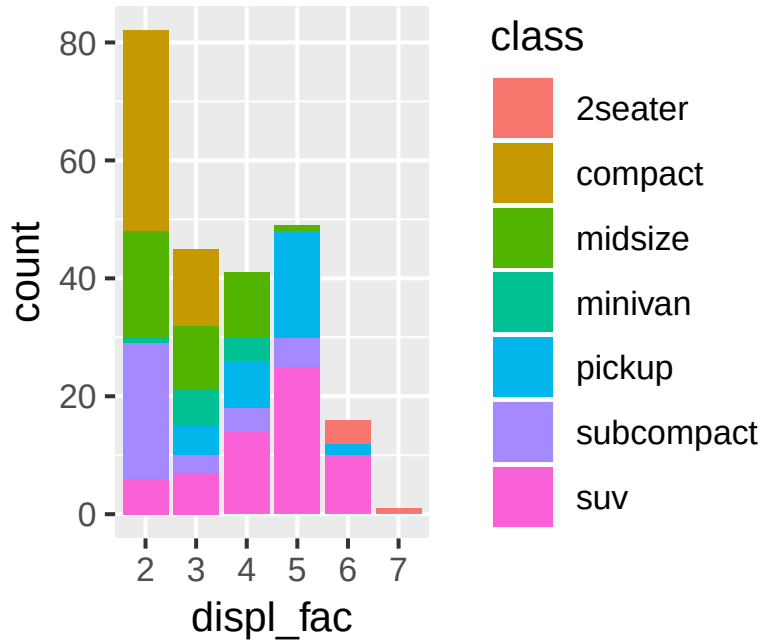
- `fill` - 채워짐

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

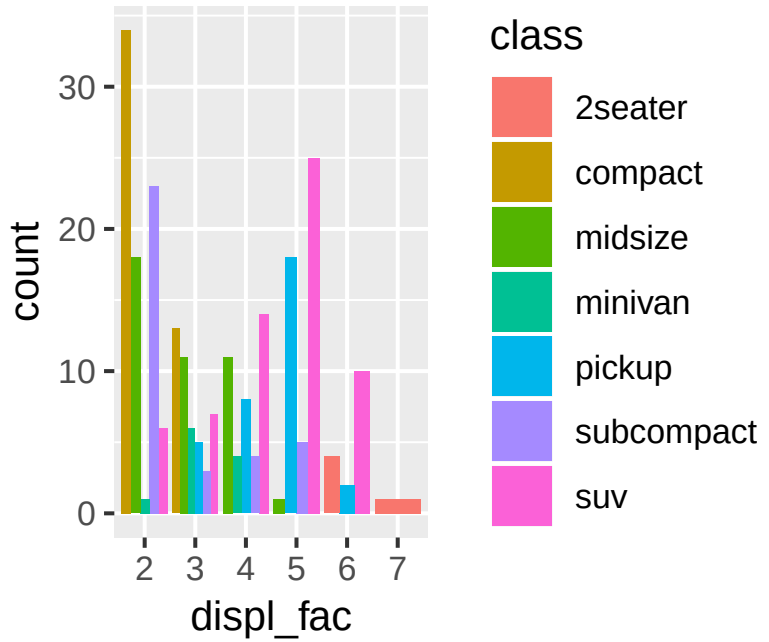


Barchart (position = "dodge")

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



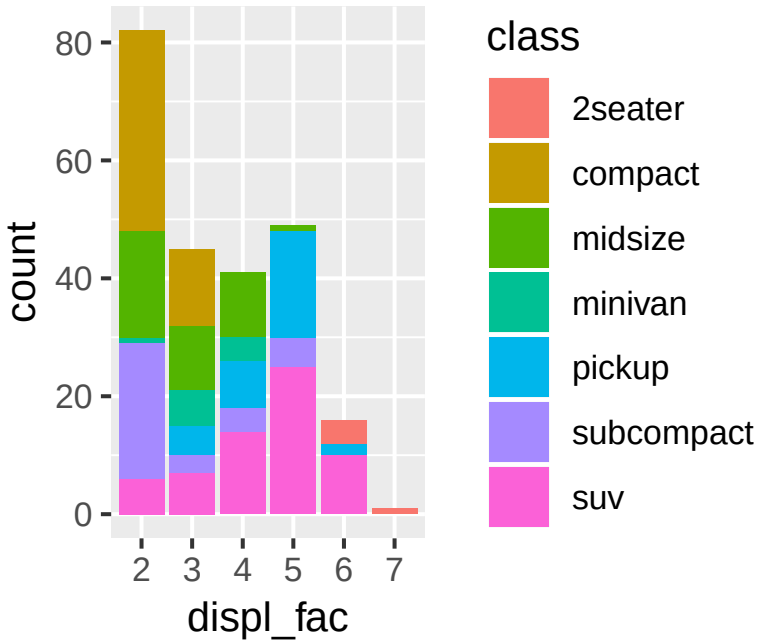

```
ggplot(mpg) +
  geom_bar(aes(x = displ_fac, fill = class),
           position = "dodge")
```



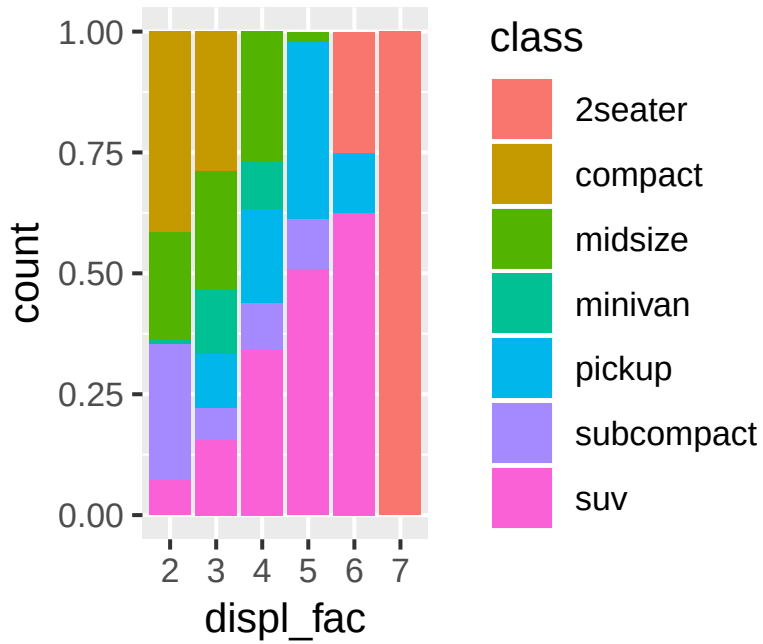
- 2개의 discrete 변수를 잘 처리하는 법?

Barchart (position = "fill")

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




```
ggplot(mpg) +
  geom_bar(aes(x = displ_fac, fill = class),
           position = "fill")
```

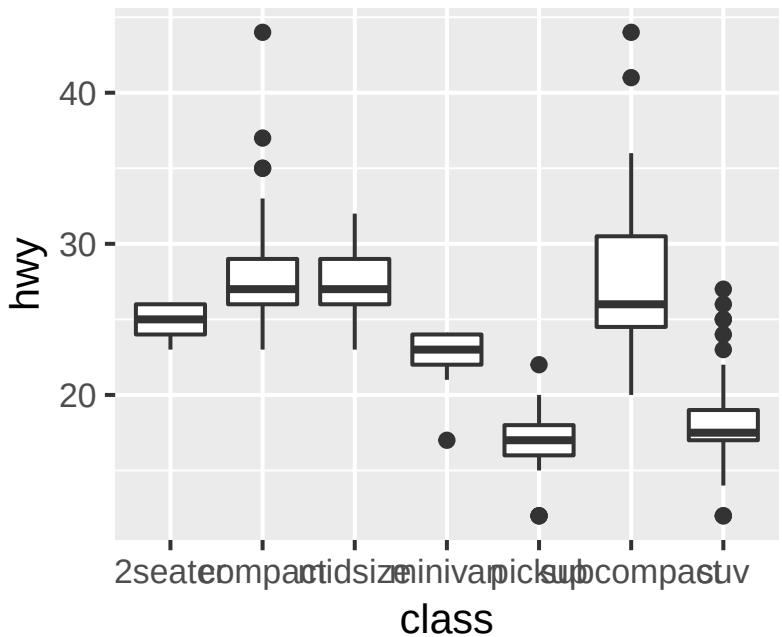


- `class`의 각각의 `displ_fac` 값에서의 분포?

Boxplot

- x 변수는 이산 변수이고 x 변수의 각각의 값에 대해서 연속 변수인 y의 분포를 보기 위함.

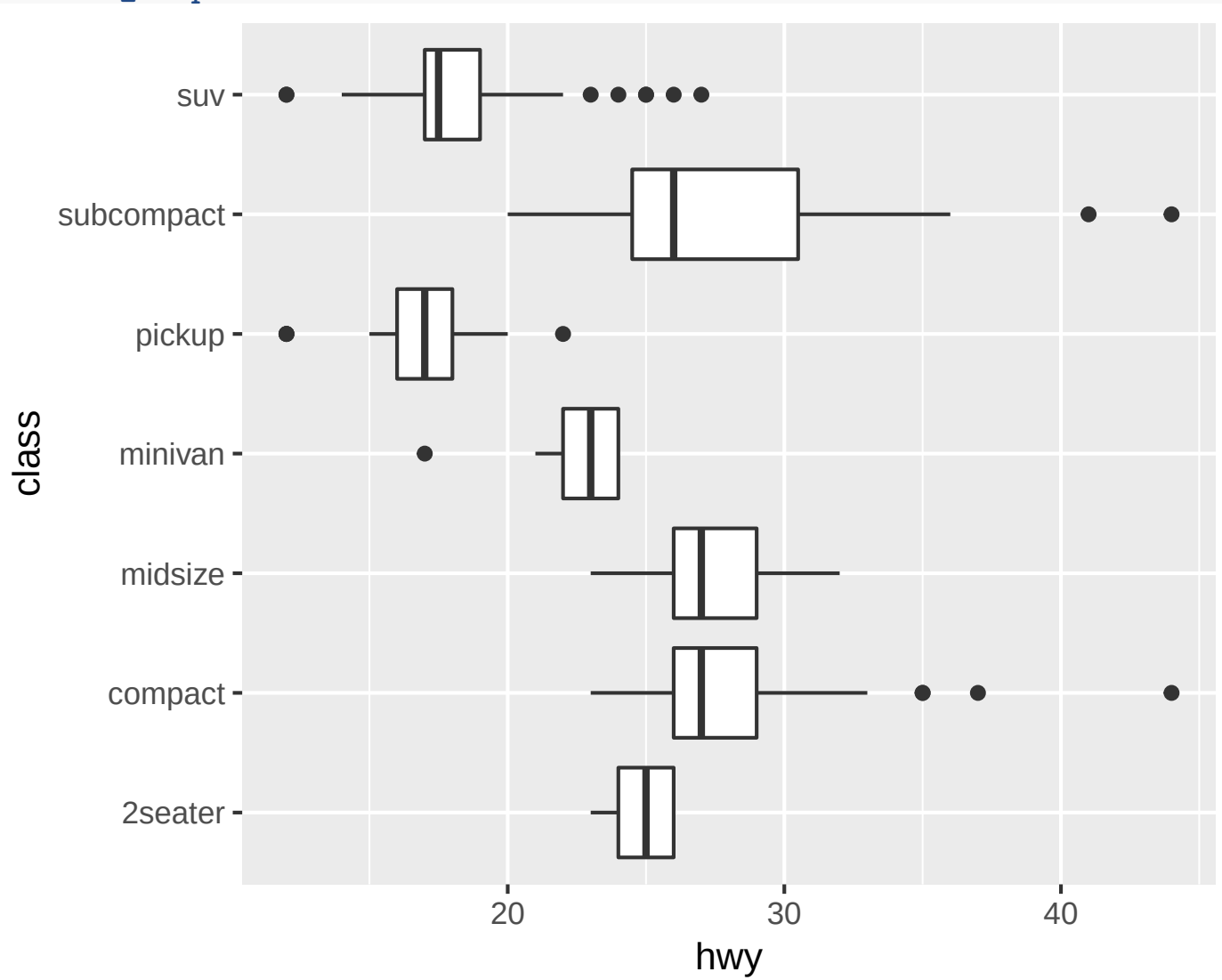
```
ggplot(mpg) +
  geom_boxplot(aes(x = class, y = hwy))
```



- 점으로 표현된 것은 이상치(outlier)로서 이상하게 높거나 낮은 값
- 박스의 상단은 상위 25%, 하단은 하위 25%
- 박스의 가운데 직선은 중간값
- x변수가 이산 변수이면서 **factor**라면, 변수의 값이 **character**라서 display가 복잡할 가능성이 높음
- 이럴때는?

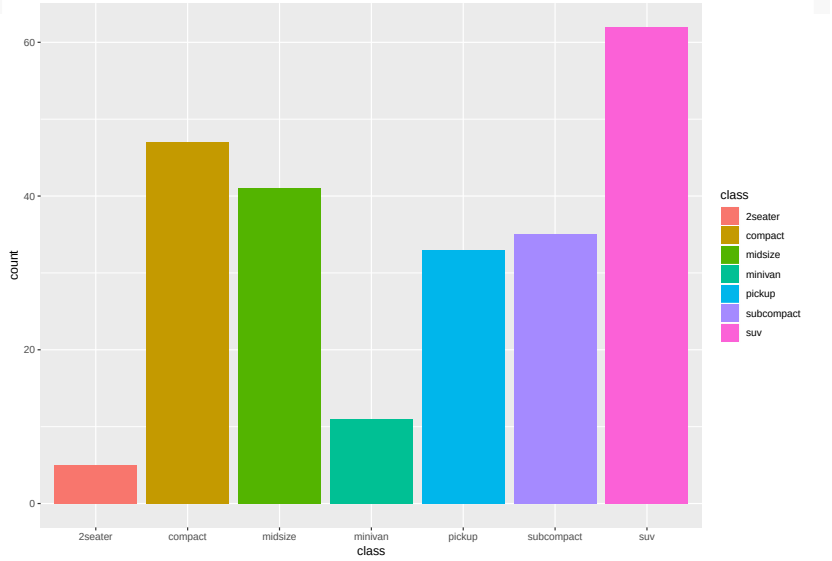
Boxplot (coord_flip())

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

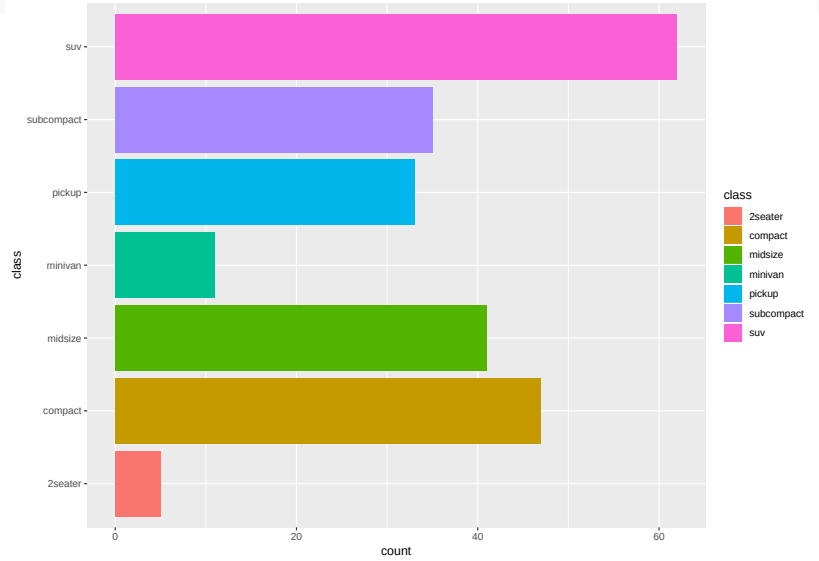


Barplot (coord_flip())

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

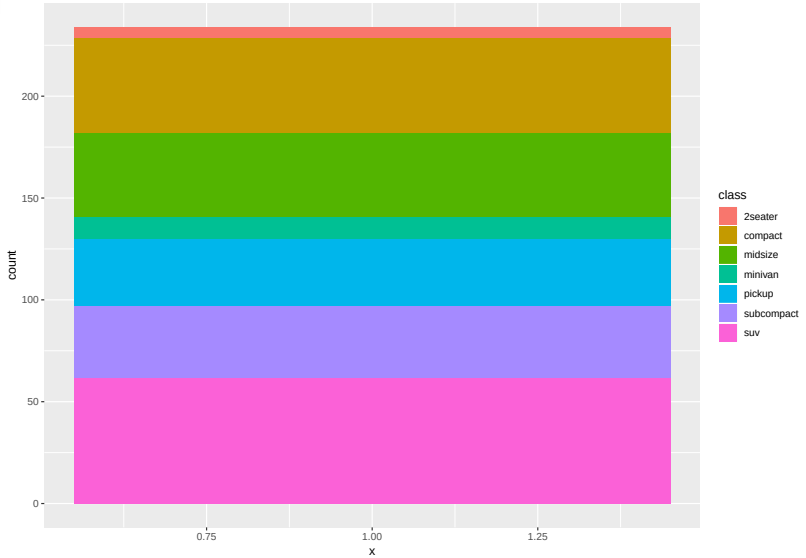



```
ggplot(mpg, aes(x = class, fill = class)) +
  geom_bar() + coord_flip()
```

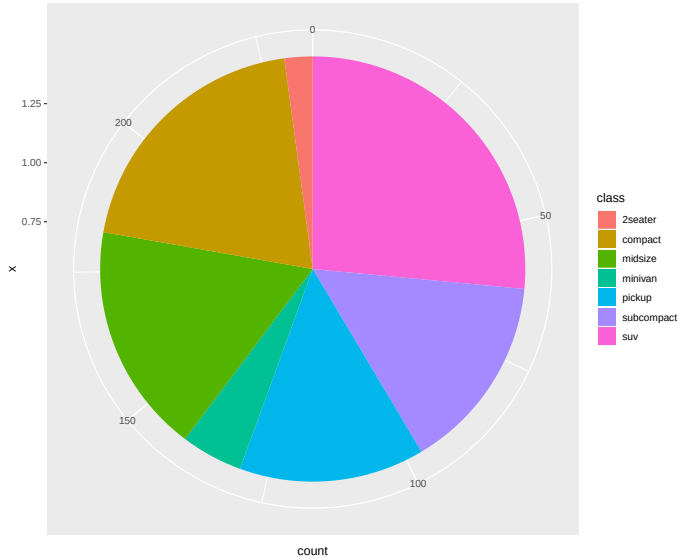


Pie-chart (coord_polar())

```
ggplot(mpg, aes(x = 1, fill = class)) +
  geom_bar()
```

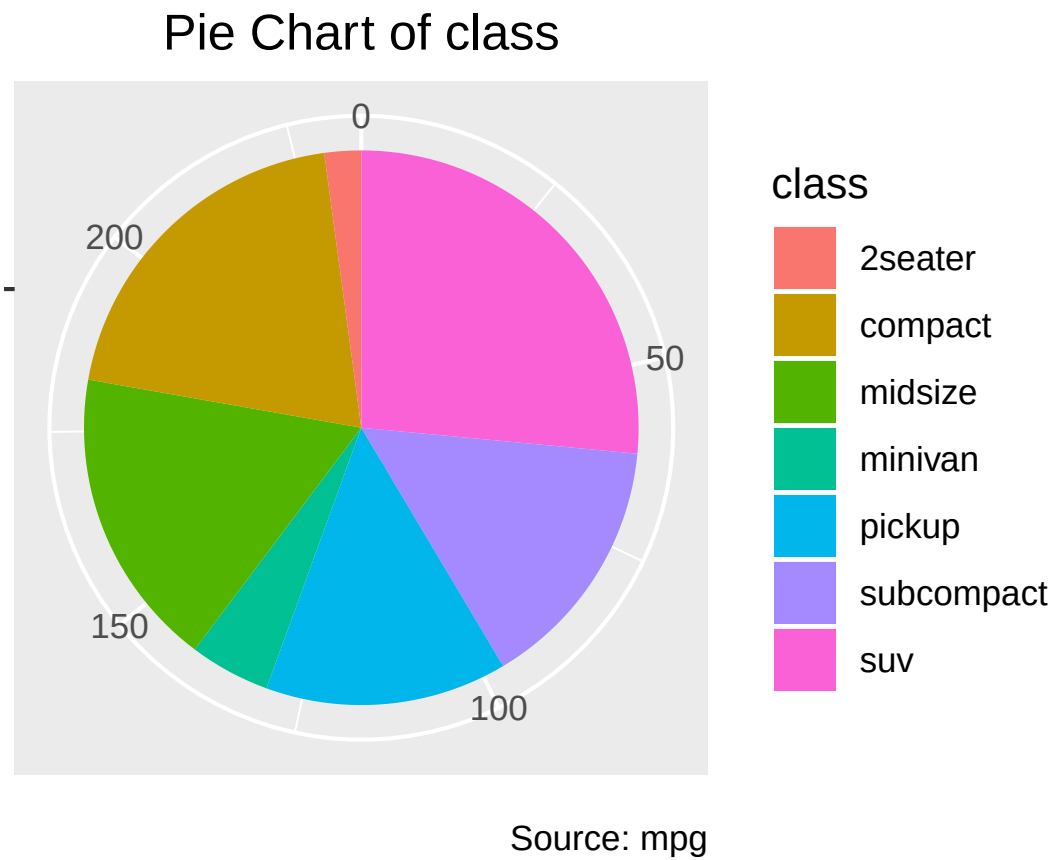



```
ggplot(mpg, aes(x = 1, fill = class)) +
  geom_bar() + coord_polar(theta = "y")
```



Pie Chart (from M91-ggplot2-50examples.pdf)

```
ggplot(mpg, aes(x = "", fill = factor(class))) +
  geom_bar(width = 1) +
  theme(axis.line = element_blank(), plot.title = element_text(hjust=0.5)) +
  labs(fill = "class", x = NULL, y = NULL,
       title = "Pie Chart of class", caption = "Source: mpg") +
  coord_polar(theta = "y", start=0)
```



ggplot 기타 기능 (저장 & plotly)

- ggplot객체를 png 파일로 저장

```
png("out_file.png") # initiate
ggplot(mpg) + geom_bar(aes(x=class)) +
  coord_flip() # save
dev.off() # finish
```

- plotly - html에서 각종 추가 기능 제공

```
library(plotly)
a <- ggplot(mpg) + geom_bar(aes(x=class))
ggplotly(a)
```

continuous = numeric

discrete, categorical = factor

	x	y	Note.
scatter	conti	conti	line, bubble chart
barplot	discrete	frequency distribution	
boxplot	discrete	conti	
piechart	discrete	proportion (bis 2)	