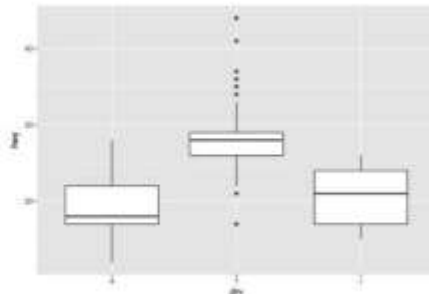
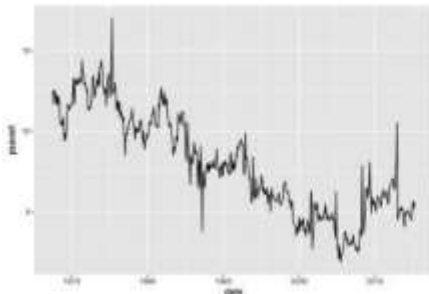
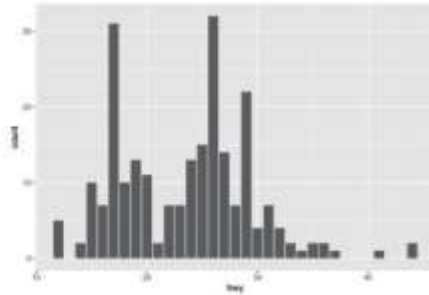
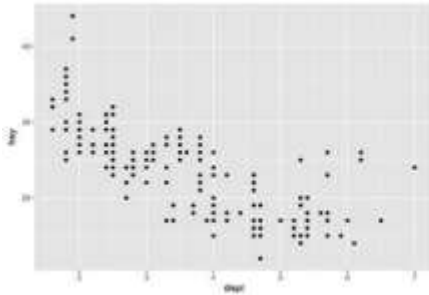


산점도 그래프 막대 그래프(빈도 막대 그래프) 선그래프 상자 그래프



소제목

ggplot2 패키지



값

내용

geom_point()

산점도

geom_col()

막대 그래프 - 요약표

geom_bar()

막대 그래프 - 원자료

geom_line()

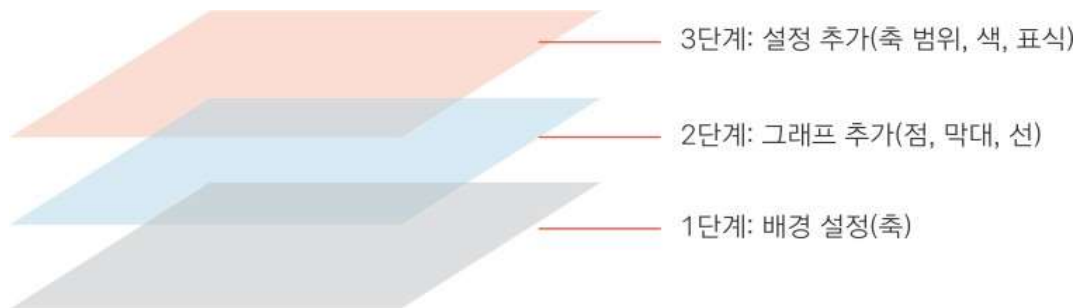
선 그래프

geom_boxplot()

상자 그림

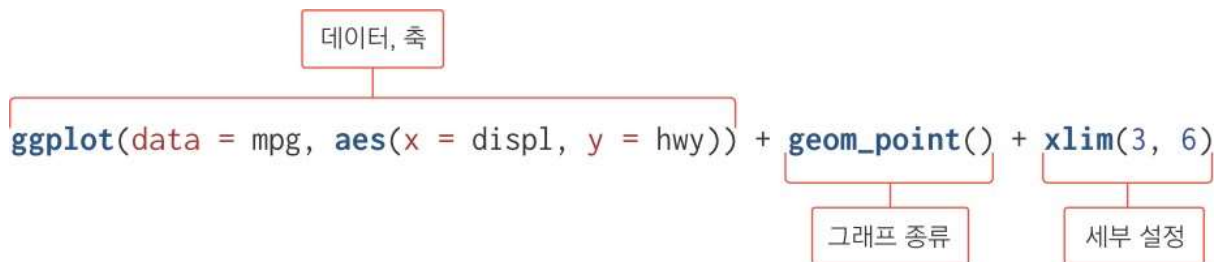
소제목

ggplot2 레이어 구조 이해하기



ggplot2 레이어 구조

- 산점도(Scater Plot) : 데이터를 x축과 y축에 점으로 표현한 그래프
- 나이와 소득처럼, 연속 값으로 된 두 변수의 관계를 표현할 때 사용



ggplot2 로드

```
library(ggplot2)
```

ggplot() vs qplot()

- qplot() : 전처리 단계 데이터 확인용 문법 간단, 기능 단순
- ggplot() : 최종 보고용. 색, 크기, 폰트 등 세부 조작 가능

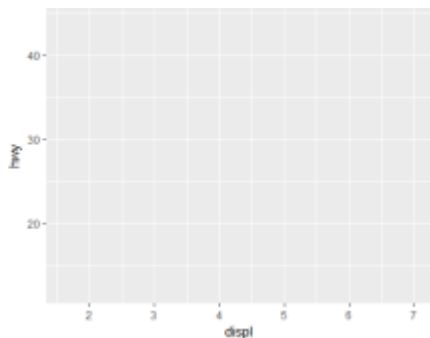
산점도 그래프



소제목

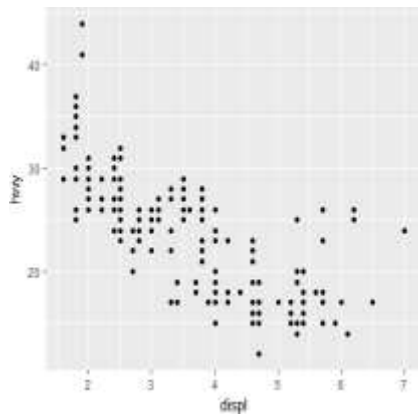
1. 배경 설정하기

```
# x 축 displ, y 축 hwy 로 지정해 배경 생성  
ggplot(data = mpg, aes(x = displ, y = hwy))
```



2. 그래프 추가하기

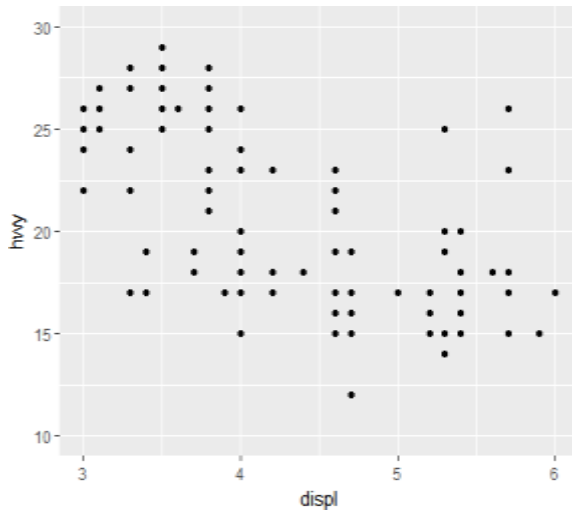
```
# 배경에 산점도 추가  
ggplot(data = mpg, aes(x = displ, y = hwy)) + geom_point()
```



축

3. 축 범위를 조정하는 설정 추가하기, option추가하기

```
# x 축 범위 3~6, y 축 범위 10~30 으로 지정  
ggplot(data = mpg, aes(x = displ, y = hwy)) +  
  geom_point(color = 'red', size = 0.5) +  
  xlim(3, 6) +  
  ylim(10, 30)
```



막대 그래프



막대 그래프

geom_col()

- 각 집단의 평균값을 막대 길이로 표현한 그래프
- 막대그래프를 사용하기 위해서는 일반적으로 **평균값**을 구해서 사용합니다.

1. 집단별 평균표 만들기

```
library(dplyr)
```

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

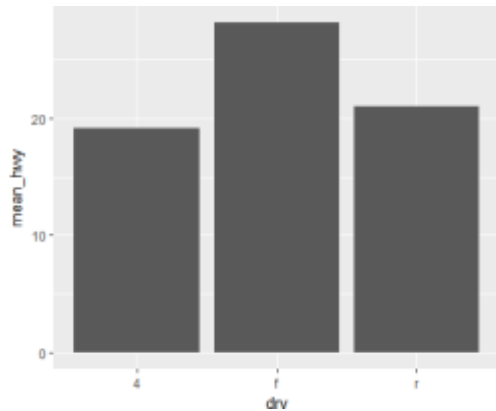
```
df_mpg  
##      drv mean_hwy  
##      <chr>      <dbl>  
## 1      4 19.17476  
## 2      f 28.16038  
## 3      r 21.00000
```



막대 그래프

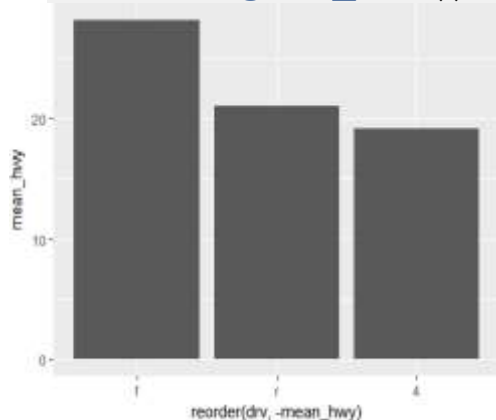
2. 그래프 생성하기

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



3. 크기 순으로 정렬하기

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



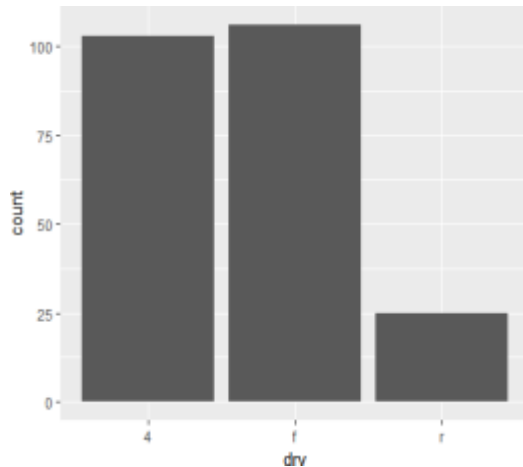
한줄로 줄바꿈 시 (+) 기호를 반드시 마지막에 적습니다

막대 빈도 그래프

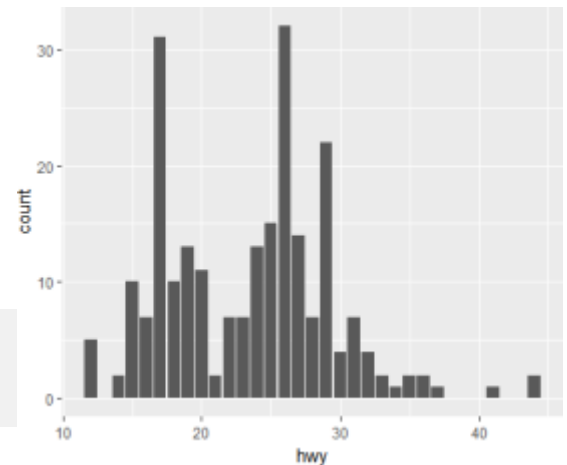
geom_bar()

- 값의 개수(빈도)로 막대의 길이를 표현한 그래프
- 빈도만으로 그래프를 확인합니다.
- X축 값의 한가지의 값을 지정합니다

```
# x 축 범주 변수, y 축 빈도  
ggplot(data = mpg, aes(x = drv)) + geom_bar()
```



```
# x 축 연속 변수, y 축 빈도  
ggplot(data = mpg, aes(x = hwy)) + geom_bar()
```





소제목

geom_col() VS geom_bar()

- 평균 막대 그래프 :
 - 데이터를 요약한 평균표를 먼저 만든 후 평균표를 이용해 그래프 생성 - geom_col()
- 빈도 막대 그래프 :
 - 별도로 표를 만들지 않고 원자료를 이용해 바로 그래프 생성 - geom_bar()

Chapter 8

선 그래프



선그래프

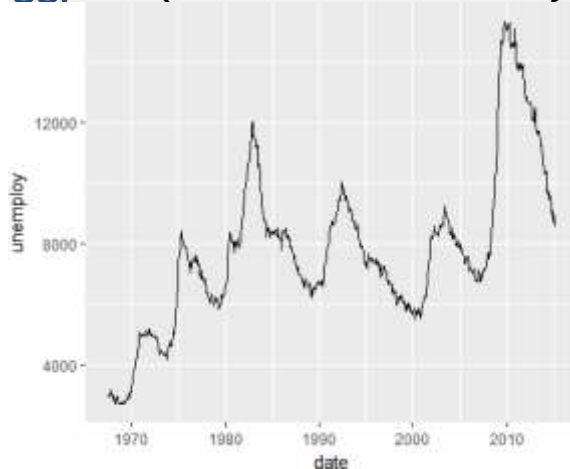
geom_line()

- 선 그래프(Line Chart) : 데이터를 선으로 표현한 그래프
- 시계열 그래프(Time Series Chart) : 일정 시간 간격을 두고 나열된 시계열 데이터(Time Series Data)를 선으로 표현한 그래프. 환율, 주가지수 등 경제 지표가 시간에 따라 어떻게 변하는지 표현할 때 활용

그래프 생성하기 (ggplot2의 economics데이터 가져오기)

```
library(ggplot2)
economics <- data.frame(economics)
```

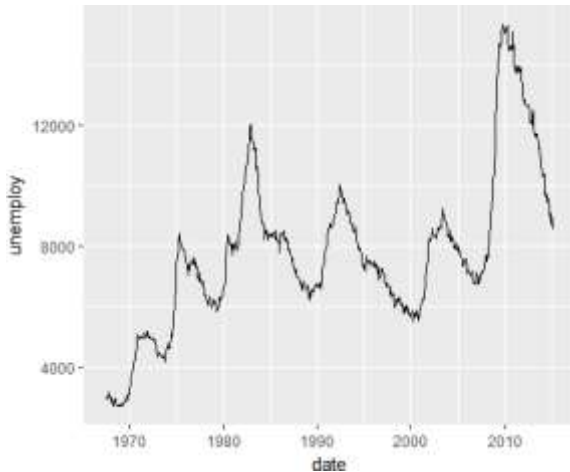
```
ggplot(data = economics, aes(x = date, y = unemploy)) + geom_line()
```



선그래프

그래프 색상 지정하기

```
ggplot(data = economics, aes(x = date, y = unemploy)) +  
  geom_line(color = 'red', size = 0.5)
```



상자 그래프



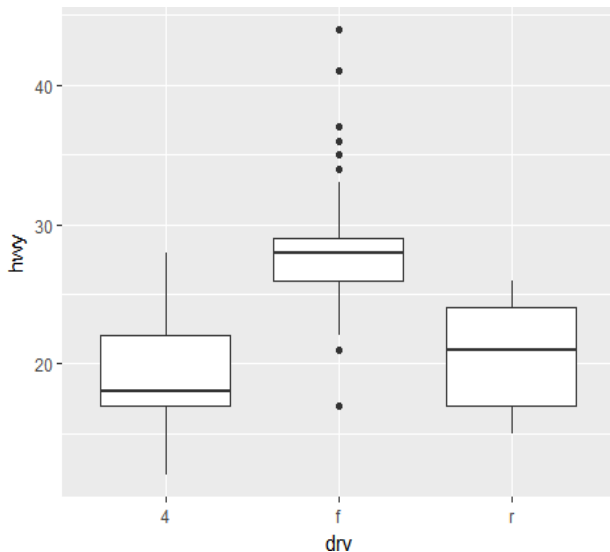
상자 그래프

geom_boxplot()

- 상자 그림(Box Plot) : 데이터의 분포(퍼져 있는 형태)를 직사각형 상자 모양으로 표현한 그래프
- 분포를 알 수 있기 때문에 평균만 볼 때보다 데이터의 특성을 좀 더 자세히 이해할 수 있음

그래프 생성하기

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



문제

Q1

mpg 데이터의 cty(도시 연비) hwy(고속연비)를 이용해서 x 축은 cty, y 축은 hwy 로 된 산점도를 생성하세요.

Q2

미국 지역별 인구통계 정보를 담은 ggplot2 패키지의 midwest 데이터를 이용합니다

x 축은 poptotal(전체 인구), y 축은 popasian(아시아인 인구)으로 된 산점도를 만들어 보세요.

전체 인구는 50 만 명 이하, 아시아인 인구는 1 만 명 이하인 지역만 산점도에 표시되게 설정하세요.

Q3

mpg데이터에서 어떤 회사별 suv 차종의 도시 연비가 높은지 확인하려 합니다.

suv차종을 대상으로 **평균** cty(도시 연비)가 가장 높은 회사 다섯 곳을 막대 그래프로 표현해 보세요.

막대는 연비 가 높은 순으로 reorder하세요.

Q4

자동차 중에서 어떤 class(자동차 종류)가 가장 많은지 알아보려고 합니다.

자동차 종류별 빈도를 표현한 막대 그래프를 만들어 보세요.



문제

Q5

ggplot2패키지 economics데이터를 이용합니다.

psavert(개인 저축률)가 시간에 따라서 어떻게 변해왔는지 알아보려고 합니다. 시간에 따른 psavert(개인 저축률)의 변화를 나타낸 시계열 그래프를 만들어 보세요.

Q6

mpg데이터를 이용합니다.

class(자동차 종류)가 "compact", "subcompact", "suv"인 자동차의 cty(도시 연비)가 어떻게 다른지 비교해보려고 합니다. 세 차종의 cty를 나타낸 상자 그림을 만들어 보세요.

힌트: 파이프라인 추출



Chapter 8

수고하셨습니다