탐색적 자료분석 실습

서울대학교 통계연구소

2023년 2월

Outline

- 본 강의에서는 R의 내장 함수를 이용해 데이터 시각화를 하는 방법과, 기본적인 통계적 추론 방법을 배웠습니다.
- 실습에서는 ggplot2 패키지를 이용해 데이터 시각화를 하는 방법을 익혀보겠습니다.
- 또한 실제 데이터를 사용하여, 데이터를 탐색하고 결론을 도출하는 과정을 살펴보겠습니다.

Contents

Section 1

ggplot2

ggplot2 사용 이유

- ggplot2는 그래프 문법을 통해 더욱 다채로운 시각화 표현을 가능하게 하는 패키지로, 현재 데이터 시각화에 많이 쓰인다.
- tidyverse는 데이터 분석을 쉽게 할 수 있도록 도와주는 패키지 그룹으로, ggplot2, dplyr, tidyr 등의 패키지가 여기에 속한다.
- tidyverse 패키지를 설치/로드하면 ggplot2 패키지가 함께 설치/로드된다.

```
install.packages("tidyverse")
library(tidyverse)
```

ggplot2 기본 문법

- ggplot의 일반적인 형태는 다음과 같다.

```
ggplot(data = <DATA>) +
    <GEOM_FUNCTION>(mapping = aes(<MAPPING>))
```

- ggplot(): 자료의 좌표축 생성
- Function(GEOM_FUNCTION): 좌표축 위에 그래프를 추가
  종류 geom_point(산점도), geom_line(선 그래프), geom_bar(막대 그래프), geom_histogram(히스토그램) 등
- Aesthetic mappings(aes): 그래프의 구성 요소(x축, y축, 색깔/모양/크기/투명도 등)과 변수를 대응

Section 2

Example1: mpg dataset

데이터 소개

```
str(mpg)
```

```
tibble [234 x 11] (S3: tbl_df/tbl/data.frame)
 $ manufacturer: chr [1:234] "audi" "audi" "audi" "audi" ...
 $ model       : chr [1:234] "a4" "a4" "a4" "a4" ...
 $ displ       : num [1:234] 1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
 $ year        : int [1:234] 1999 1999 2008 2008 1999 1999 2008 1999 19
 $ cyl         : int [1:234] 4 4 4 4 6 6 6 4 4 4 ...
 $ trans       : chr [1:234] "auto(l5)" "manual(m5)" "manual(m6)" "auto
 $ drv         : chr [1:234] "f" "f" "f" "f" ...
 $ cty         : int [1:234] 18 21 20 21 16 18 18 18 16 20 ...
 $ hwy         : int [1:234] 29 29 31 30 26 26 27 26 25 28 ...
 $ fl          : chr [1:234] "p" "p" "p" "p" ...
 $ class       : chr [1:234] "compact" "compact" "compact" "compact" ..
```

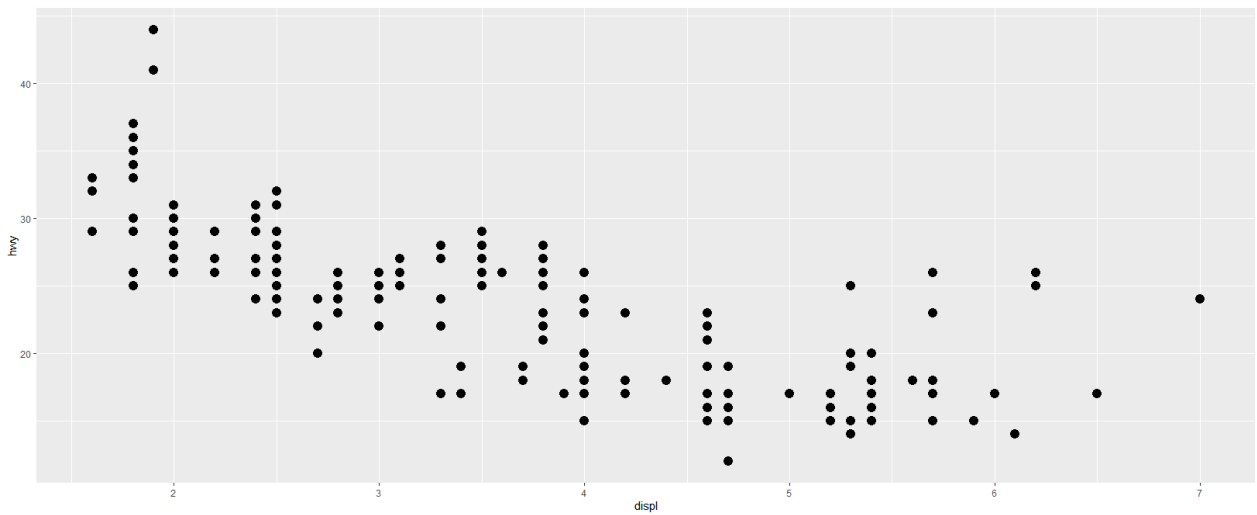
- 1999년부터 2008년까지 미국 EPA에서 조사한 자동차 모델별 연비효율 관련 데이터
- ggplot2 패키지에 내장되어 있음

산점도 그리기

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

- x축에 displ(엔진 크기) 변수를, y축에 hwy(연비) 변수를 매핑
- 아래와 같이 보다 간결하게 표현할 수 있다

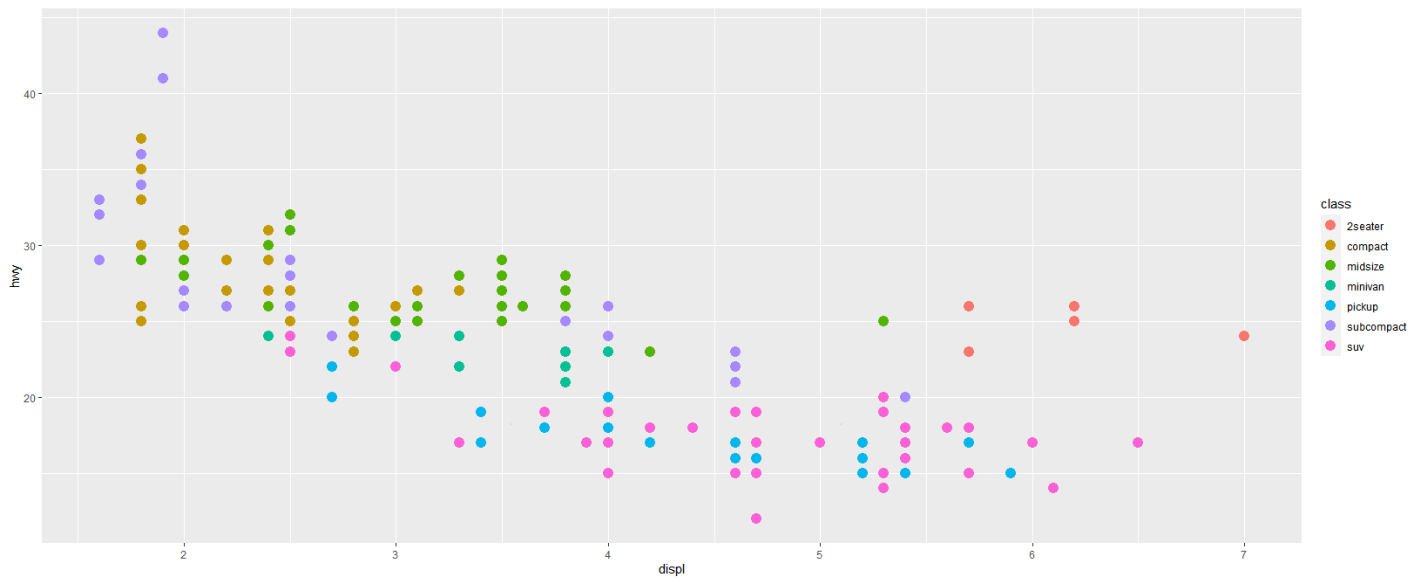
```
ggplot(mpg) +
  geom_point(aes(x = displ, y = hwy))
```



Aesthetic

- 점의 색상(color), 모양(shape), 크기(size) 등에 다른 변수를 매핑함으로써, 데이터로부터 추가적인 정보를 얻을 수 있다.
- 다음은 연비와 엔진크기의 산점도에서 class(차종)에 따라 다른 색상의 점들로 표현한 결과이다.

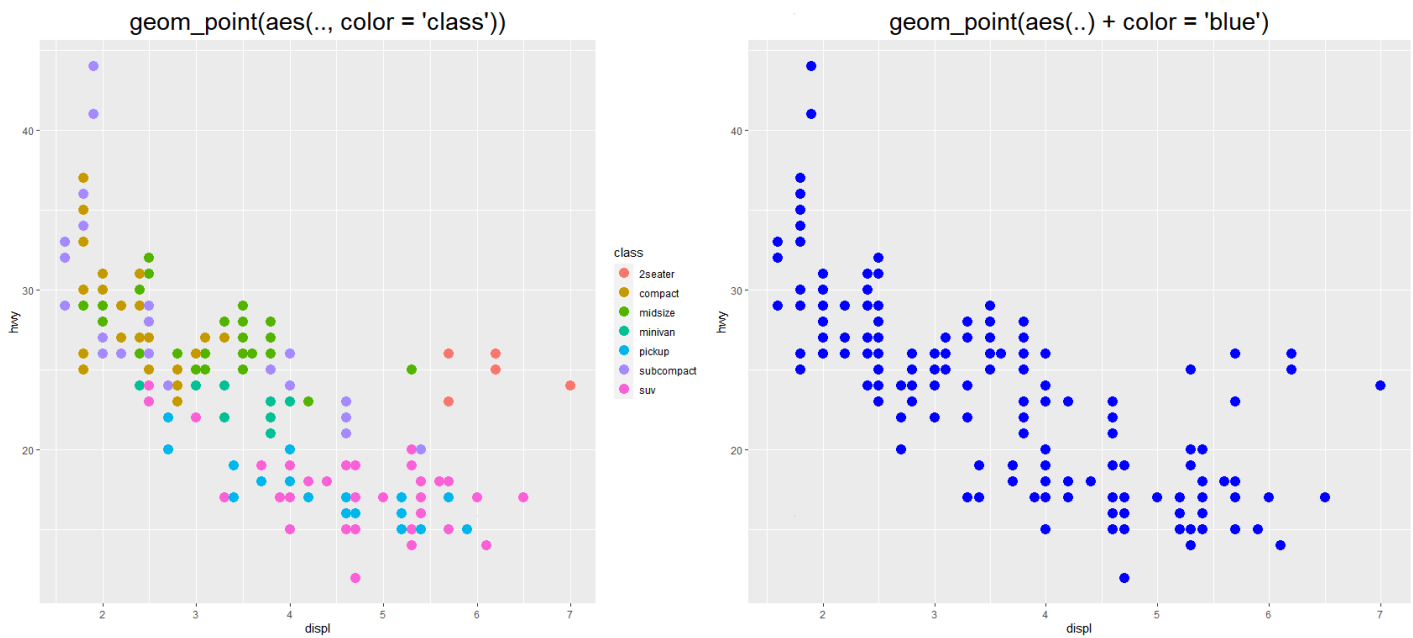
```
ggplot(mpg) +
  geom_point(aes(x = displ, y = hwy, color = class))
```



Aesthetic (참고)

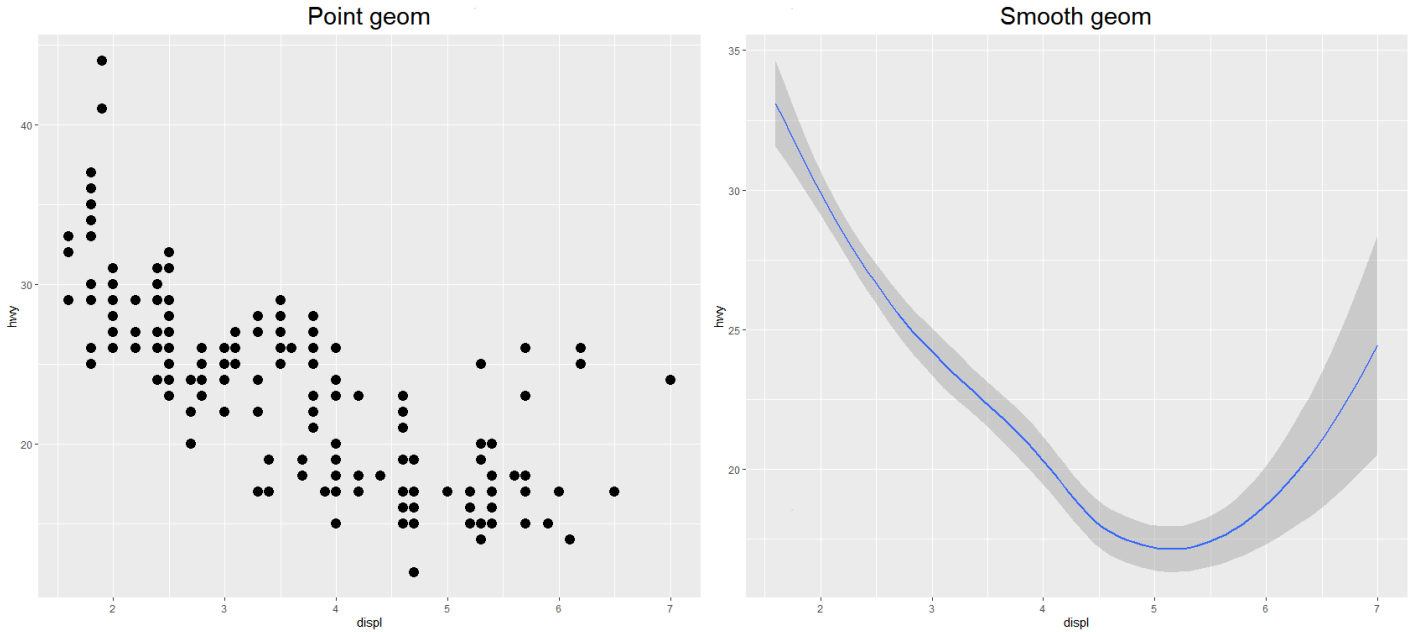
- aes 바깥에서 시각적 속성을 특정 값으로 지정하면, 그래프의 모든 점의 속성이 그 값으로 동일하게 바뀐다.

```
# left plot
ggplot(mpg) +
  geom_point(aes(x = displ, y = hwy, color = class))
# right plot
ggplot(mpg) +
  geom_point(aes(x = displ, y = hwy), color = 'blue')
```



Geometric object(Geom)

- 아래는 두 가지 다른 geom을 사용하여 그린 그래프이다.




```
# point geom (left plot)
ggplot(mpg) +
  geom_point(aes(x = displ, y = hwy))
# smooth geom (right plot)
ggplot(mpg) +
  geom_smooth(aes(x = displ, y = hwy))
```

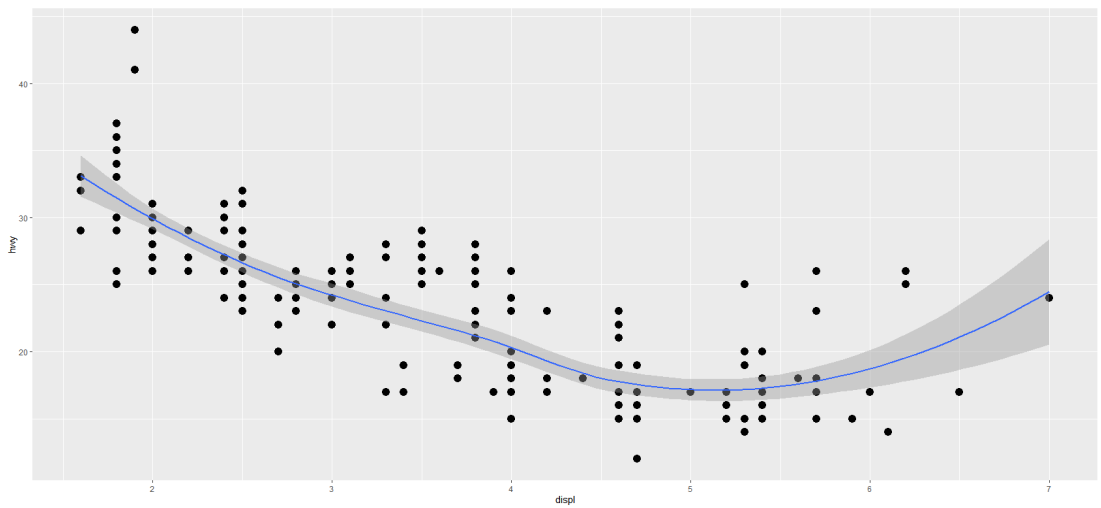
Geometric object(Geom)

- 여러 개의 geom을 동시에 사용할 수도 있다.

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

- 이 때 중복되는 코드를 묶어 간결하게 표현할 수 있다.

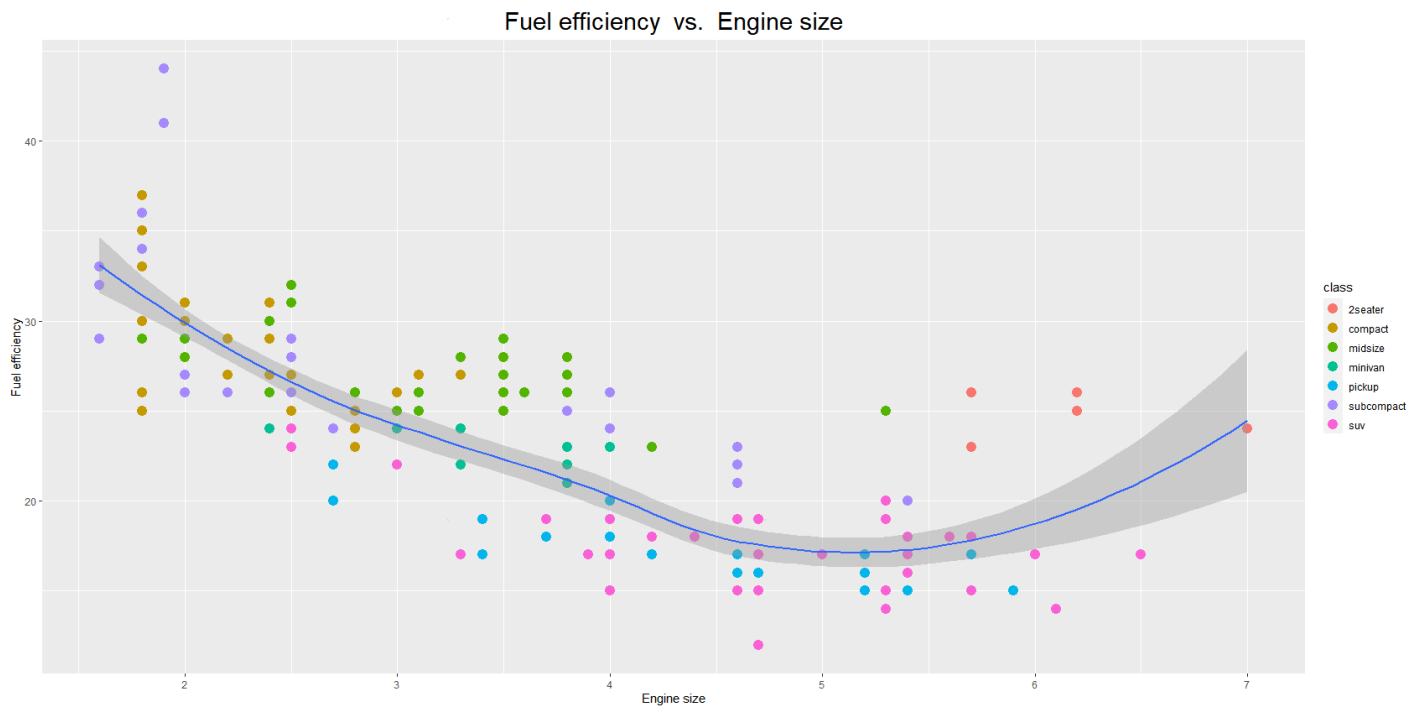
```
ggplot(mpg, aes(x = displ, y = hwy)) +
  geom_point() + geom_smooth()
```



Label

- 그래프 제목(title), 소제목(subtitle), x축 이름(x), y축 이름(y), 캡션(caption) 등을 지정할 수 있다.

```
ggplot(mpg, aes(x = displ, y = hwy)) +
  geom_point(aes(color = class), size = 4) + geom_smooth() +
  labs(title = "Fuel efficiency  vs.  Engine size",
       x = "Engine size", y = "Fuel efficiency") +
  theme(plot.title = element_text(size = 22, hjust = 0.5))
```



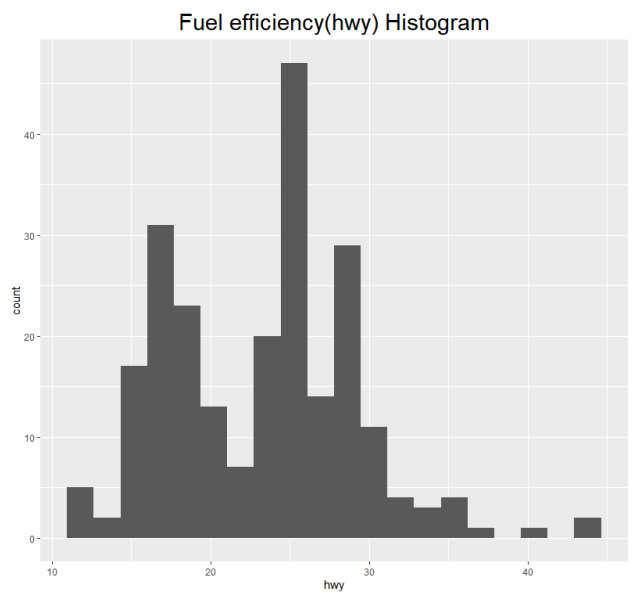
연속형 자료의 요약

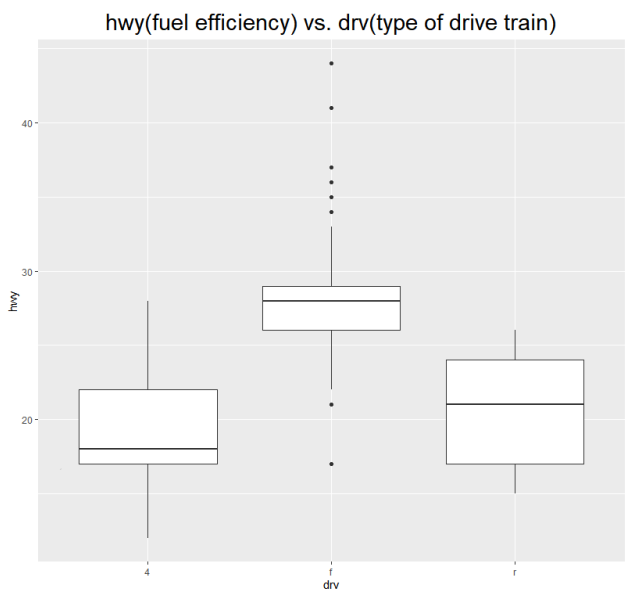
- 히스토그램 (Histogram)

```
ggplot(mpg, aes(x = hwy)) +
  geom_histogram(bins = 20) +
  labs(title = "Fuel efficiency(hwy) Histogram")
```

- 상자 그림 (Box plot)

```
ggplot(mpg, aes(x = drv, y = hwy)) +
  geom_boxplot() +
  labs(title = "hwy(fuel efficiency) vs. drv(type of drive train)")
```





이산형 자료의 요약

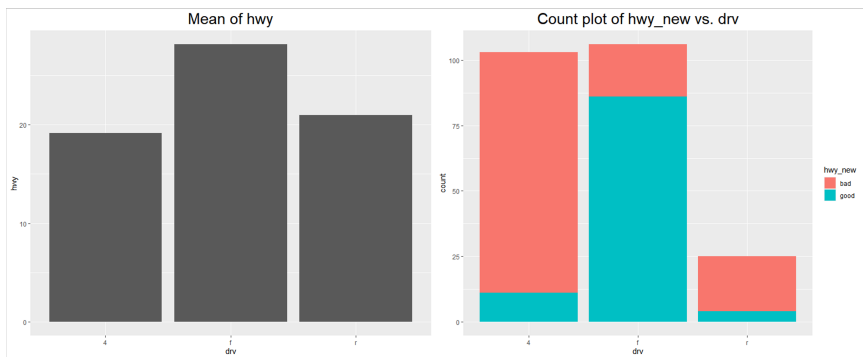
- 도수분포표

```
mpg$hwy_new = ifelse(mpg$hwy <= 25, "bad", "good")
table(mpg$drv, mpg$hwy_new)
```

- 막대 그래프

```
ggplot(mpg) +
  geom_bar(aes(x = drv, y = hwy), stat = 'summary', fun = 'mean') +
  labs(title = 'Mean of hwy')
ggplot(mpg) +
  geom_bar(aes(x = drv, fill = hwy_new))+
  labs(title = 'Count plot of hwy_new vs. drv')
```

	bad	good
4	92	11
f	20	86
r	21	4



도수분포표 | 막대 그래프

Section 3

Example2: Iris dataset

데이터 소개

```
data(iris)
str(iris)
```

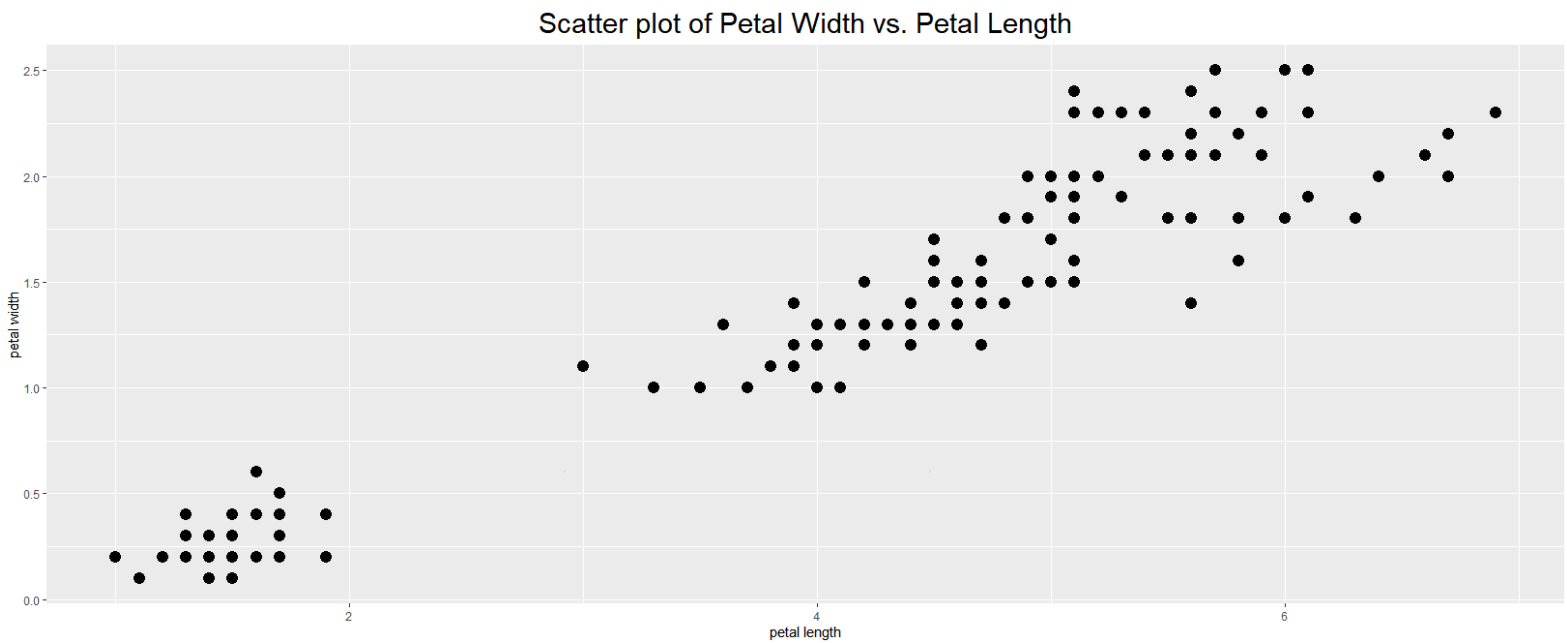
```
'data.frame':        150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1
```

- 붓꽃의 세 품종 setosa, versicolor, virginica에 대하여, 각 품종당 50개 개체의 꽃받침 길이/너비, 꽃잎 길이/너비를 측정한 데이터
- 붓꽃의 품종에 따라 꽃의 크기에 차이가 있는가?

데이터 시각화

꽃잎의 크기와 너비 산점도

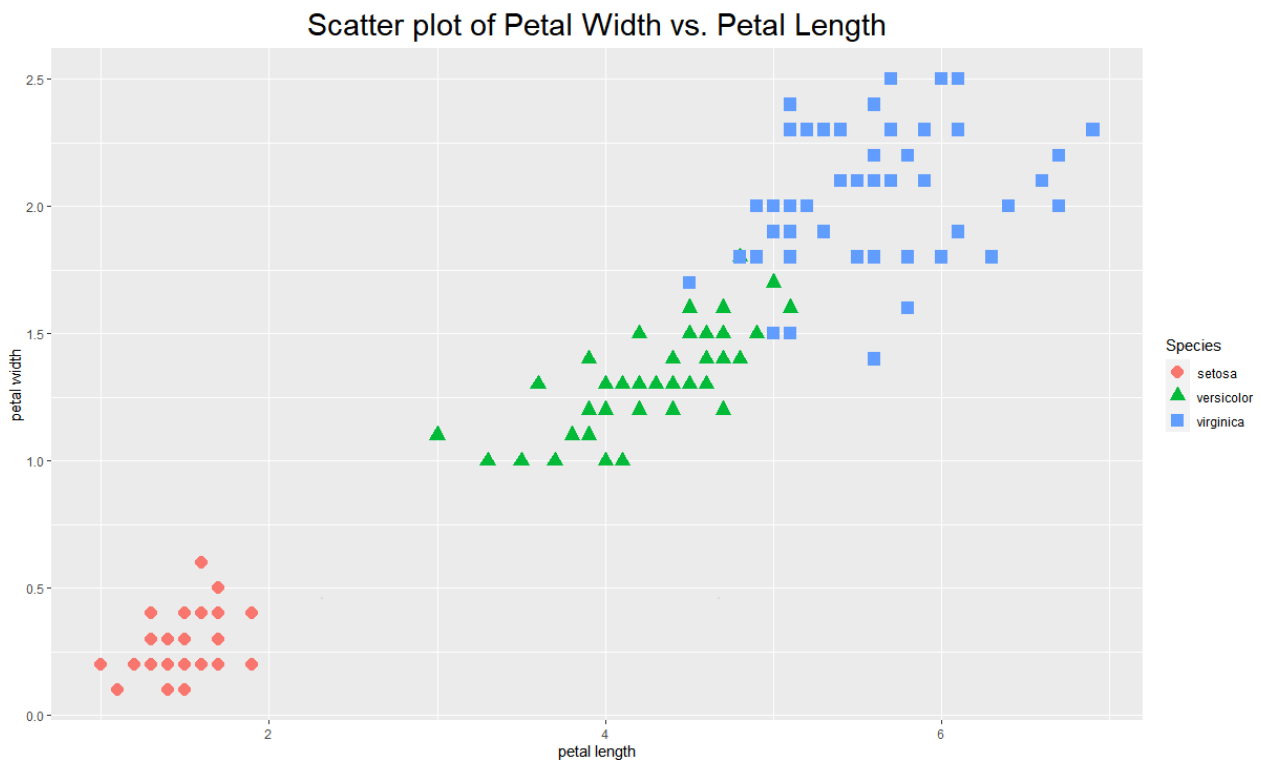
```
ggplot(iris) +
  geom_point(aes(x = Petal.Length, y = Petal.Width)) +
  labs(title = "Scatter plot of Petal Width vs. Petal Length",
       x = "petal length", y = "petal width") +
  theme(plot.title = element_text(size = 22, hjust = 0.5))
```



데이터 시각화

꽃잎의 크기와 너비 산점도 - 점의 색상과 모양으로 품종 구별

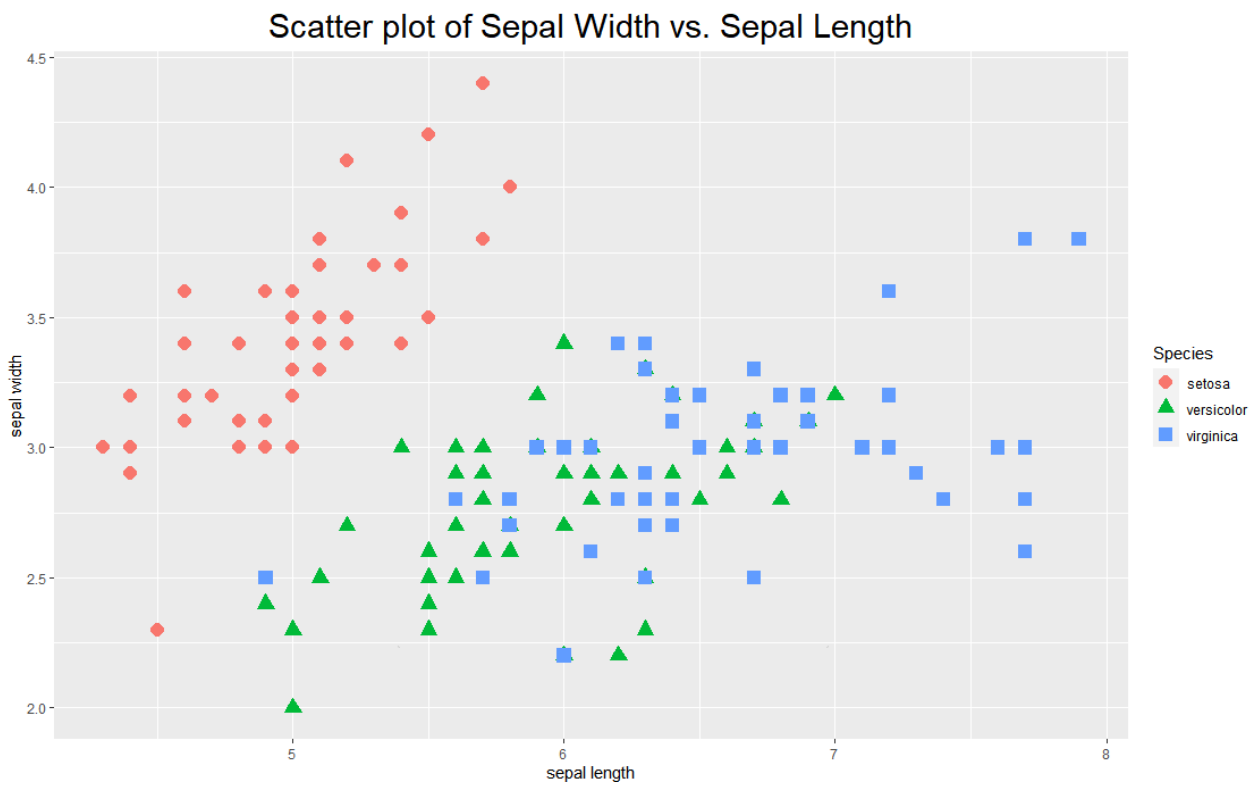
```
ggplot(iris) +
  geom_point(aes(x = Petal.Length, y = Petal.Width,
                 color = Species, shape = Species), size = 4) +
  labs(title = "Scatter plot of Petal Width vs. Petal Length",
       x = "petal length", y = "petal width") +
  theme(plot.title = element_text(size = 22, hjust = 0.5))
```



데이터 시각화

꽃받침의 크기와 너비 산점도 - 점의 색상과 모양으로 품종 구별

```
ggplot(iris) +
  geom_point(aes(x = Sepal.Length, y = Sepal.Width,
                 color = Species, shape = Species), size = 4) +
  labs(title = "Scatter plot of Sepal Width vs. Sepal Length",
       x = "sepal length", y = "sepal width") +
  theme(plot.title = element_text(size = 22, hjust = 0.5))
```



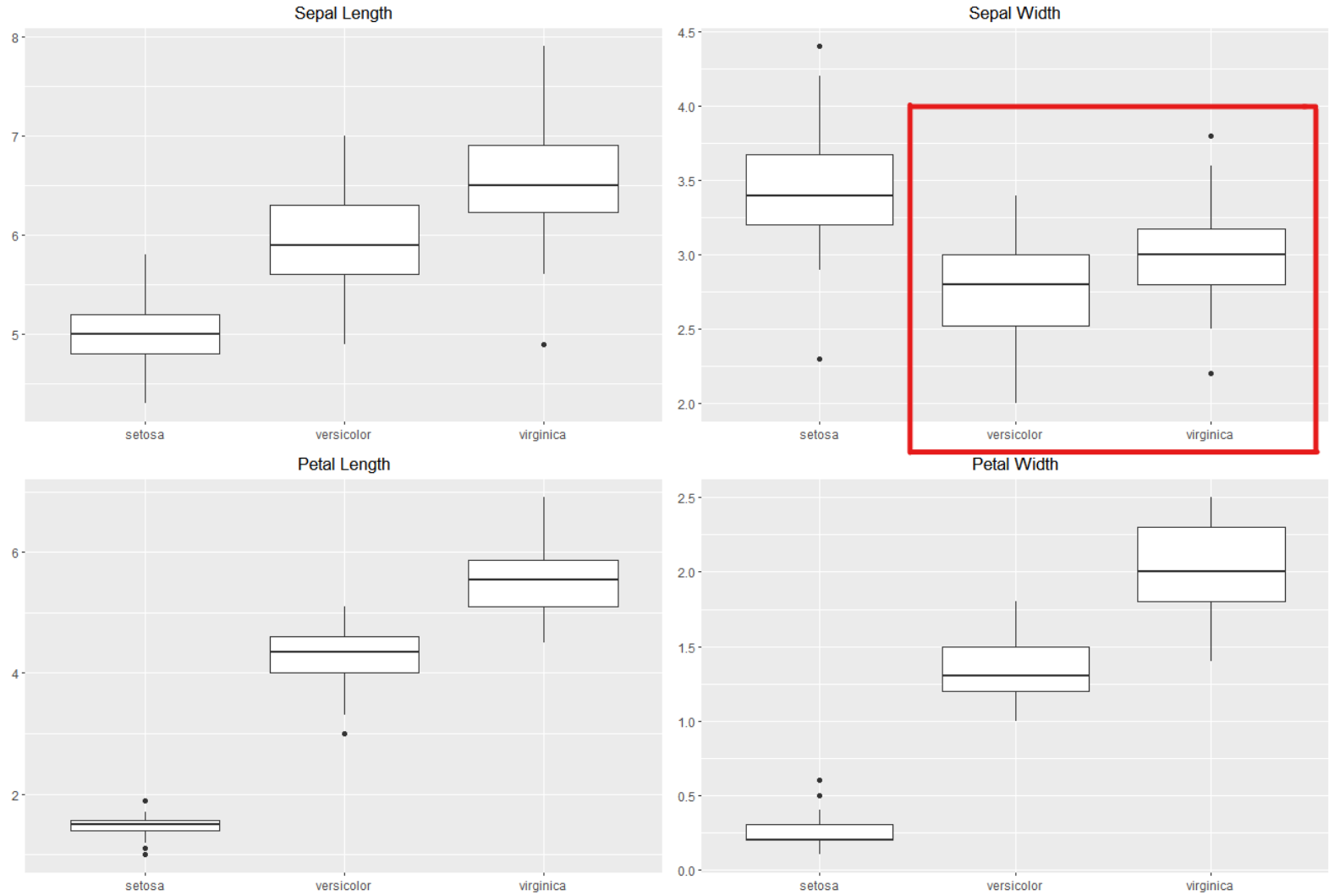
데이터 시각화

상자그림을 이용한 탐색

```
library(gridExtra)
p1 = ggplot(iris) +
  geom_boxplot(aes(x = Species, y = Sepal.Length)) +
  labs(title = "Sepal Length") +
  theme(axis.title.x = element_blank(),
        axis.title.y = element_blank())
p2 = ggplot(iris) +
  geom_boxplot(aes(x = Species, y = Sepal.Width)) +
  labs(title = "Sepal Width") +
  theme(axis.title.x = element_blank(),
        axis.title.y = element_blank())
p3 = ggplot(iris) +
  geom_boxplot(aes(x = Species, y = Petal.Length)) +
  labs(title = "Petal Length") +
  theme(axis.title.x = element_blank(),
        axis.title.y = element_blank())
p4 = ggplot(iris) +
  geom_boxplot(aes(x = Species, y = Petal.Width)) +
  labs(title = "Petal Width") +
  theme(axis.title.x = element_blank(),
        axis.title.y = element_blank())
grid.arrange(p1,p2,p3,p4, nrow = 2)
```

데이터 시각화

상자그림을 이용한 탐색



수치 요약을 통한 탐색

```
tapply(iris$Sepal.Width, iris$Species, summary)
```

```
$setosa
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.300   3.200   3.400   3.428   3.675   4.400

$versicolor
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.525   2.800   2.770   3.000   3.400

$virginica
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.200   2.800   3.000   2.974   3.175   3.800
```

가설 검정

Q : 유의수준 5%에서 versicolor와 virginica의 평균 꽃받침 너비(Sepal Width)가 다르다고 할 수 있는가?

- $H_0 : \mu_{vs} = \mu_{vg}$ vs. $H_1 : \mu_{vs} \neq \mu_{vg}$
  μ_{vs}: versicolor 품종의 평균 꽃받침 너비
  μ_{vg}: virginica 품종의 평균 꽃받침 너비

A : 독립 이표본 t 검정을 적용할 수 있다.
'등분산 검정'을 먼저 실시하고, 그 결과에 따라
'등분산(이분산) 독립 이표본 t 검정'을 진행한다.

가설 검정

(Step 1) 등분산 검정

$$H_0 : \sigma_{vs}^2 = \sigma_{vg}^2 \quad \text{vs.} \quad H_1 : \sigma_{vs}^2 \neq \sigma_{vg}^2$$

```
vs = iris$Sepal.Width[iris$Species == 'versicolor']
vg = iris$Sepal.Width[iris$Species == 'virginica']
var.test(vs, vg)
```

```
        F test to compare two variances

data:  vs and vg
F = 0.94678, num df = 49, denom df = 49, p-value = 0.849
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5372773 1.6684117
sample estimates:
ratio of variances
         0.9467839
```

- p-value > 0.05 이므로 유의수준 5%에서 귀무가설을 기각할 수 없다.
- 등분산을 가정한 독립 이표본 t 검정을 진행한다.

가설 검정

(Step 2) 등분산 독립 이표본 t 검정

$H_0 : \mu_{vs} = \mu_{vg}$ vs. $H_1 : \mu_{vs} \neq \mu_{vg}$

```
t.test(vs, vg, var.equal = TRUE)
```

```
        Two Sample t-test

data:  vs and vg
t = -3.2058, df = 98, p-value = 0.001819
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.33028246 -0.07771754
sample estimates:
mean of x mean of y
    2.770     2.974
```

- p-value < 0.05 이므로 유의수준 5%에서 귀무가설을 기각할 수 있다.
- 유의수준 5%에서 versicolor와 virginica의 평균 꽃받침 너비는 다르다고 할 수 있다.