

Day31; 20221019

	날짜
	유형
	태그

GitHub - u8yes/R

You can't perform that action at this time. You signed in with another tab or window. You signed out in another tab or window. Reload to refresh your session. Reload to refresh your session.

u8yes/R



<https://github.com/u8yes/R>

As 1 Contributor 0 Issues ⭐ 1 Star 0 Forks

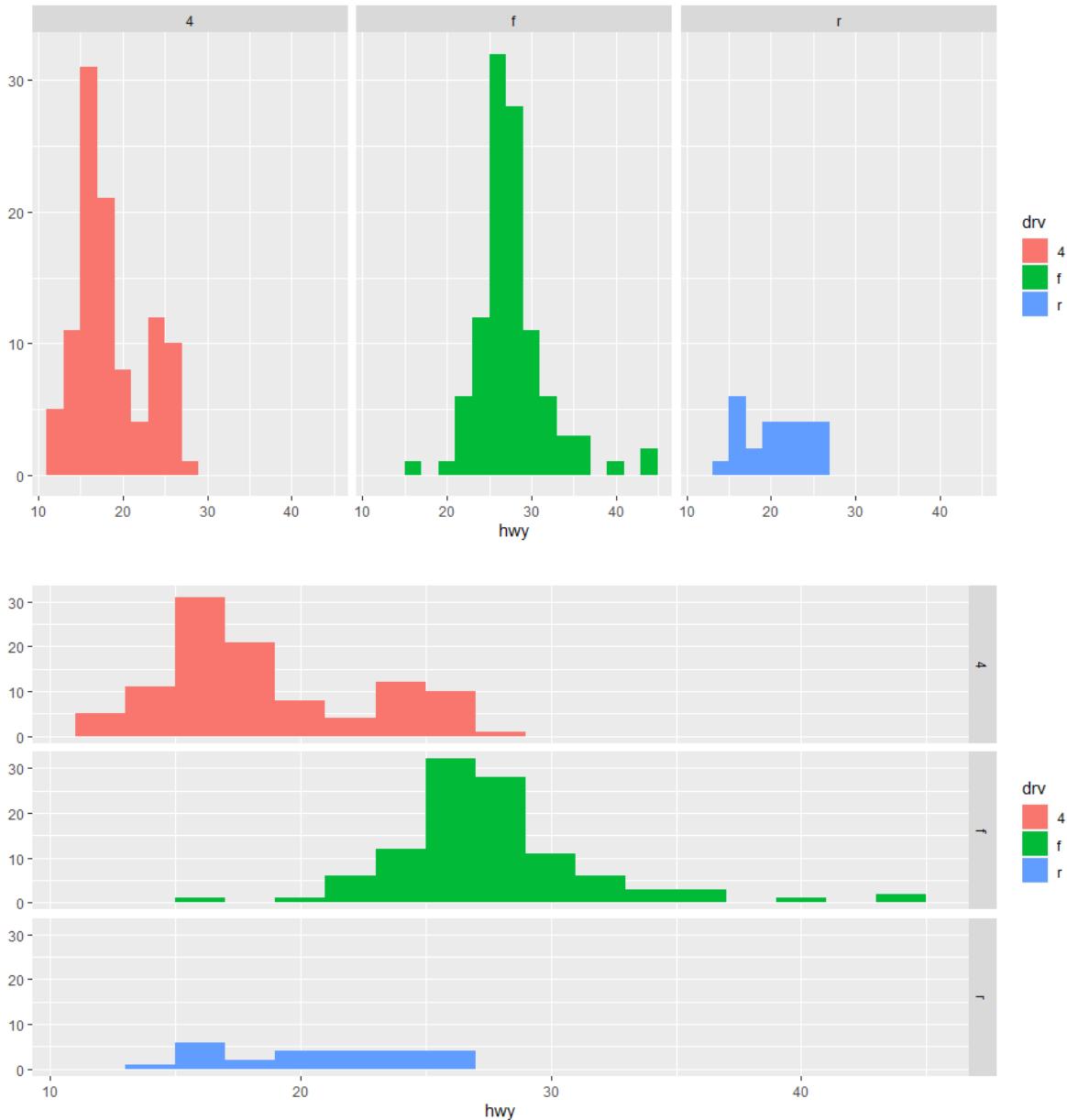
https://s3-us-west-2.amazonaws.com/secure.notion-static.com/9a74104e-fce9-4732-837d-29f202f214f6/chap08_VisualizationAnalysis.r

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/76ac807e-c27e-4136-8f64-85966fe850ee/08_%EA%BA%A0%EA%B8%89_%EC%8B%9C%EA%B0%81%ED%99%94_%EB%B6%84%EC%84%9D.pdf

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/362ad355-189d-459f-808f-1bfeebaf8c7/09_%EC%A0%95%ED%98%95%EA%B3%BC_%EB%B9%84%EC%A0%95%ED%98%95_%EB%8D%B0%EC%9D%B4%ED%84%BO.pdf

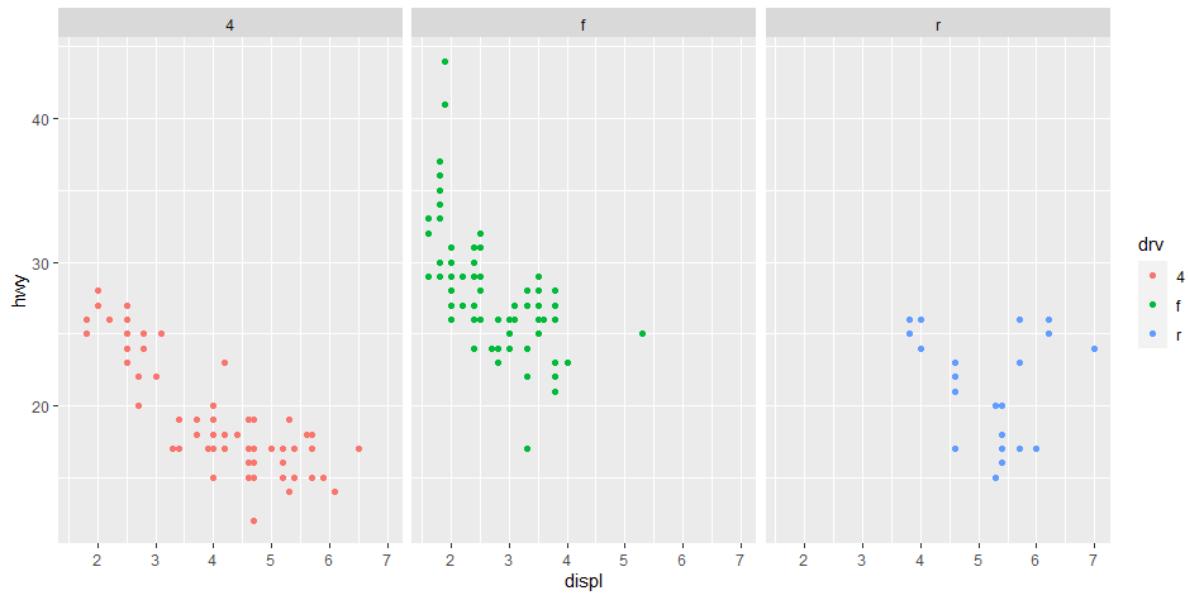
facets = ~ 열단위

facets = 행단위 ~

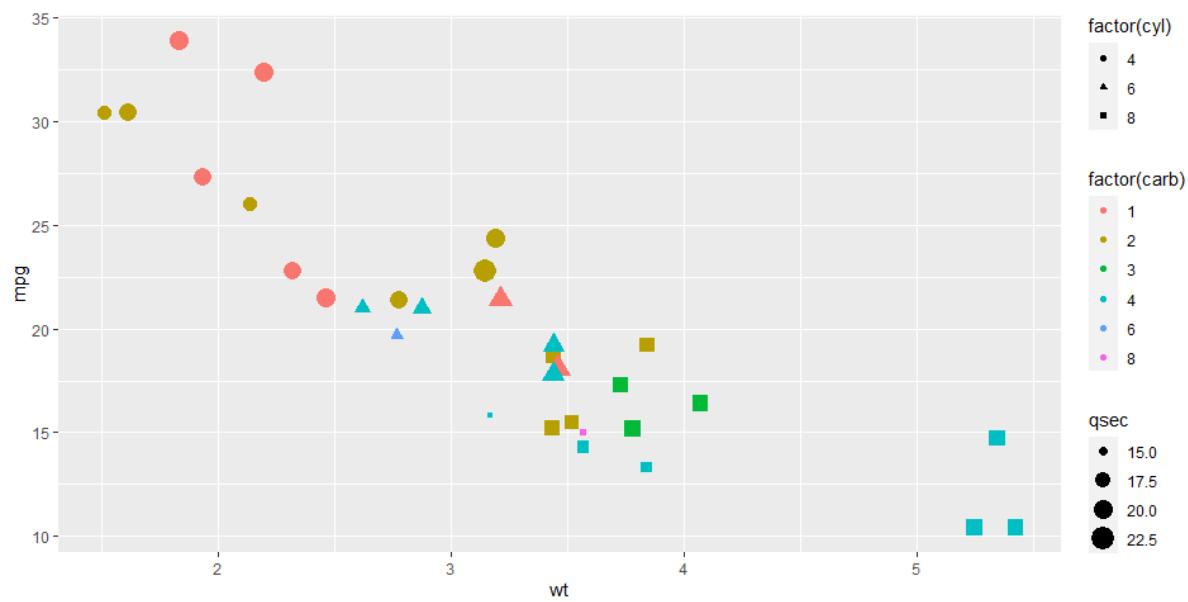


```
# facets 속성:drv 변수 값으로 컬럼 단위와 행 단위로 패널 생성
qplot(hwy, data=mpg, fill=drv, facets = .~ drv, binwidth=2) # 열 단위 패널 생성

qplot(hwy, data=mpg, fill=drv, facets = drv ~. , binwidth=2) # 행 단위 패널 생성
```



```
# displ과 hwy 변수와 관계를 drv로 구분
qplot(displ, hwy, data=mpg, color=drv, facets=.~drv)
```



```
help(mtcars)
qplot(wt,mpg,data=mtcars) # 1개(wt)이면 막대 형태, 2개(mpg)이면 점의 형태
qplot(wt,mpg,data=mtcars,color=factor(carb)) # 색상 적용
qplot(wt,mpg,data=mtcars,color=factor(carb),size=qsec) # 크기 적용
qplot(wt,mpg,data=mtcars,color=factor(carb),size=qsec, shape=factor(cyl)) # 모양 적용
```

price : 다이아몬드 가격(\$326 ~ \$18,823)

carat : 다이아몬드 무게(0.2 ~ 5.01)

cut : 컷의 품질(Fair, Good, Very Good, Premium, Ideal)

color : 색상(J : 가장 나쁨 ~ D : 가장 좋음)

clarity : 선명도(I1 : 가장 나쁨, SI2, SI1, VS2, VS1, VVS2, VVS1, IF : 가장 좋음)

x : 길이(0~10.74mm), y : 폭(0~58.9mm), z : 깊이(0~31.8mm),

depth : 깊이 비율 = z / mean(x, y)

clarity

미국·영국[ˈklærəti] ↗ 영국식 ↗

(영사)

1 (표현의) 명료성

a lack of clarity in the law ↗

그 법률의 명료성 부족

2 (사고력·이해력 등의) 명확성

clarity of thought/purpose/vision ↗

사고/목적/비전의 명확성

3 (사진·소리·물질의) 선명도[투명도]

the clarity of sound on a CD ↗

시디 음질의 투명도

[영어사전 결과 더보기](#)

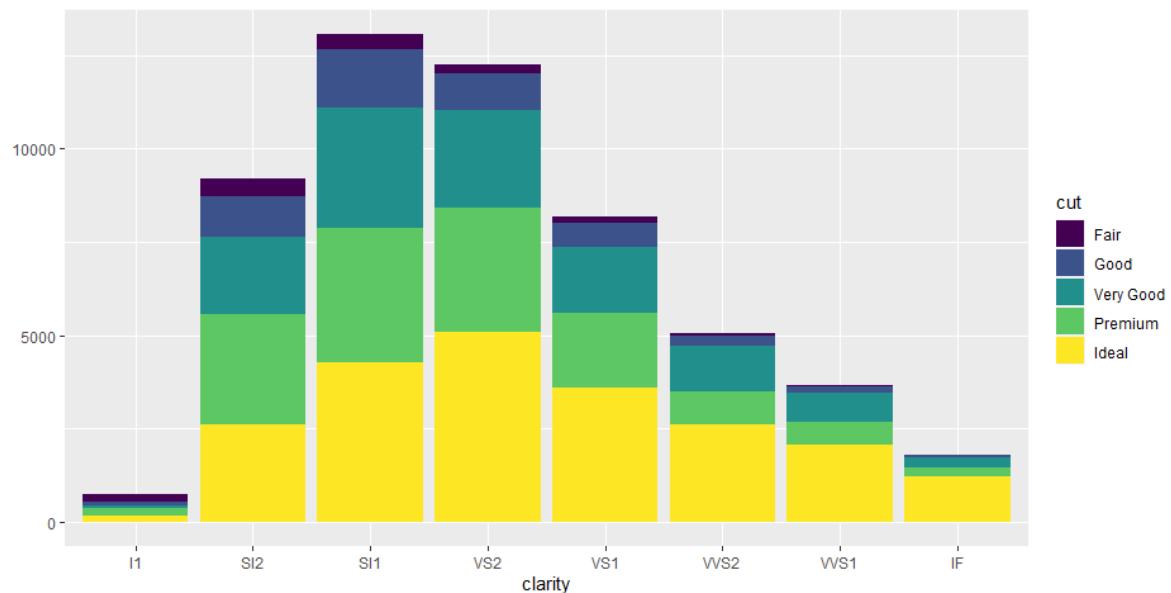
geometry

미국식[dʒiˈɑ:mətri] ↗ 영국식 ↗

1 기하학

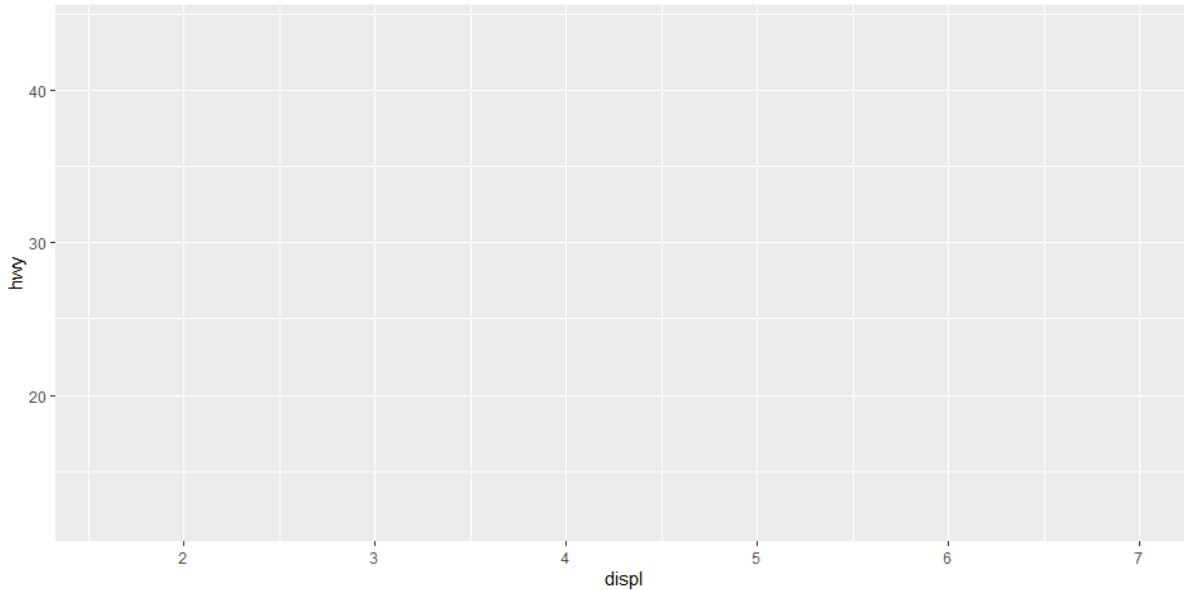
2 기하학적 구조

[영어사전 결과 더보기](#)



```
# geom="bar"(속성으로 막대그래프 그리기)
# -> clarity 변수 대상 cut 변수로 색 채우기
qplot(clarity,data=diamonds,fill=cut, geom="bar")# 레이아웃에 색 채우기
```

기하학 - 점, 선, 면 등의 형태



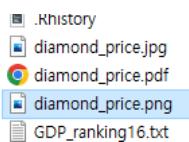
```
# 3.2 ggplot() 함수
# 단계1(layer1): 배경 설정하기.
# x축은 displ, y축은 hwy로 지정해 배경 생성
ggplot(data=mpg,aes(x=displ,y=hwy)) # aesthetics(미학)
```

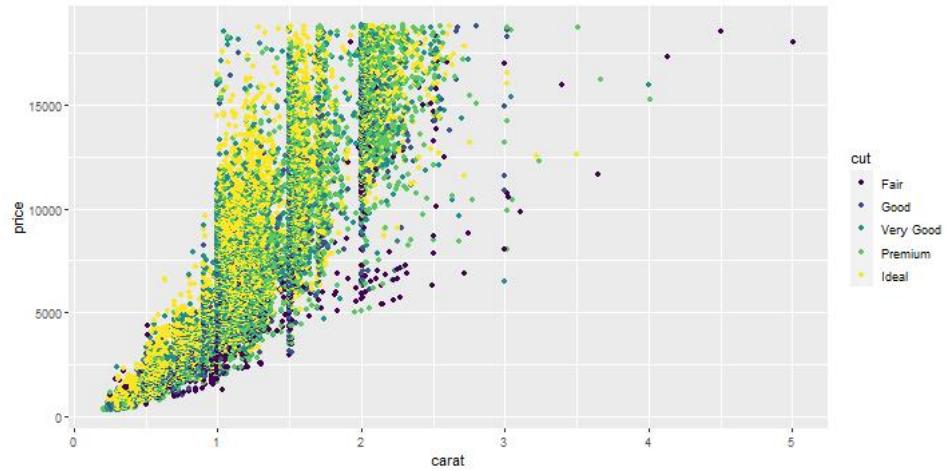
```
# 3.3 ggsave() 함수
p <- ggplot(diamonds, aes(carat,price,color=cut))
p + geom_point()

# 가장 최근 그래프 저장
ggsave(file="D:/heaven_dev/workspaces/R/output/diamond_price.pdf")
ggsave(file="D:/heaven_dev/workspaces/R/output/diamond_price.jpg", dpi=72)

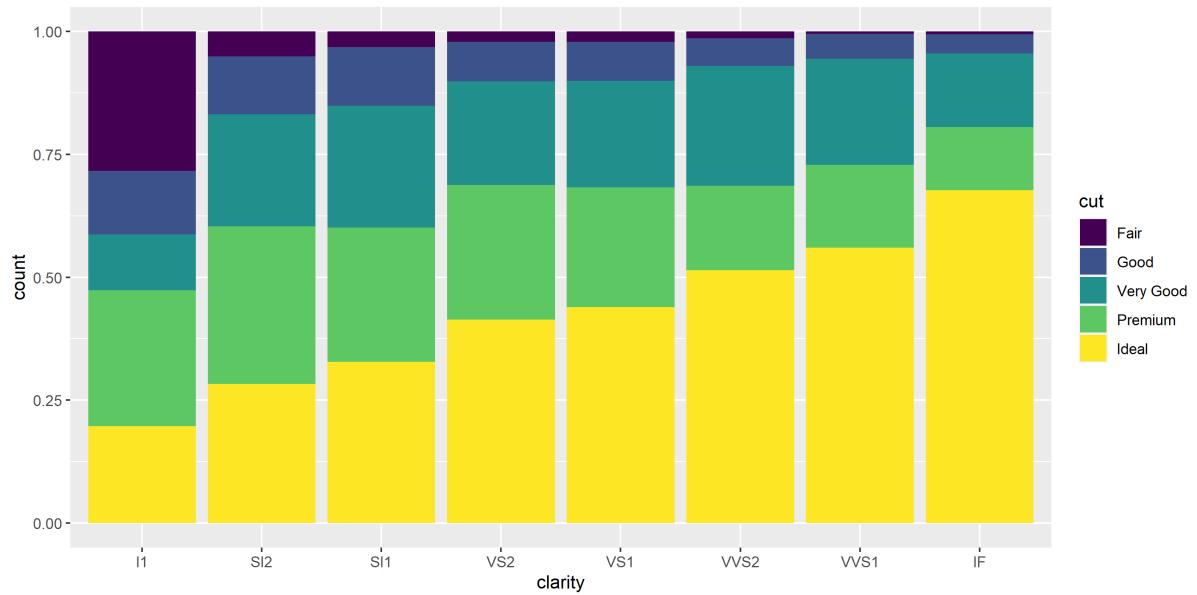
# 변수에 저장된 그래프 저장
p <- ggplot(diamonds, aes(clarity)) # 선명도
p <- p + geom_bar(aes(fill=cut), position="fill") # bar 추가
p

ggsave(file="D:/heaven_dev/workspaces/R/output/diamond_price.png",plot=p,width=10, height=5)
```

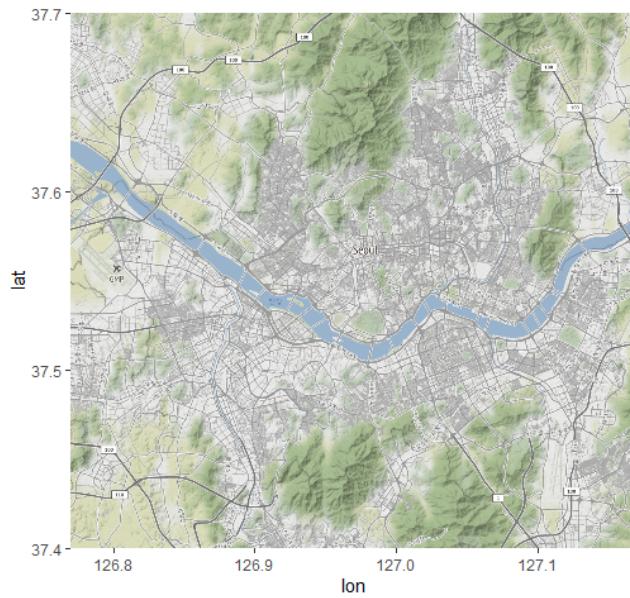




https://s3-us-west-2.amazonaws.com/secure.notion-static.com/d53de105-2481-4574-ba74-a9437e0ea4b9/diamond_price.pdf



투명도(clarity) IF 쪽이 품질이 더 좋다.



```
# 4. 지도 공간 기법 시각화(ggmap package)

# stamen Maps API 이용

# 지도 관련 패키지 설치
library(ggplot2) # ggplot2 패키지 로딩
install.packages("ggmap") # ggmap 패키지 설치
library(ggmap)

# 위도와 경도 중심으로 지도 시각화
# 실습: 서울을 중심으로 지도 시각화하기
# 단계 1: 서울 지역의 중심 좌표 설정
seoul <- c(left = 126.77, bottom = 37.40,
           right = 127.17, top = 37.70)

# 단계 2: zoom, maptype으로 정적 지도 가져오기
map <- get_stamenmap(seoul, zoom = 12, maptype = 'terrain')
ggmap(map)
```



```

# 실습 : 2019년도 1월 대한민국 인구수를 기준으로 지역별 인구수 표시하기
# 단계 1: 데이터 셋 가져오기
pop <- read.csv(file.choose(), header = T)
View(pop)

library(stringr)

region <- pop$'지역명'
lon <- pop$LON
lat <- pop$LAT
tot_pop <- as.numeric(str_replace_all(pop$'총인구수', ',', '')) # 완전 숫자로 강제 형변환

df <- data.frame(region, lon, lat, tot_pop)
df
df <- df[1:17, ]
df

# 단계 2: 정적 지도 가져오기
daegu <- c(left = 123.4423013, bottom = 32.8528306,
            right = 131.601445, top = 38.8714354)
map <- get_stamenmap(daegu, zoom = 7, maptype = 'watercolor')

# 단계 3: 지도 시각화하기
layer1 <- ggmap(map)

# 단계 4: 포인트 추가
layer2 <- layer1 + geom_point(data = df,
                                aes(x = lon, y = lat,
                                    color = factor(tot_pop),
                                    size = factor(tot_pop)))
layer2

# 단계 5: 텍스트 추가
layer3 <- layer2 + geom_text(data = df,
                               aes(x = lon + 0.01, y = lat + 0.08,
                                   label = region), size = 3)
layer3

# 단계 6: 크기를 지정하여 파일로 저장하기
ggsave("D:/heaven_dev/workspaces/R/output/pop201901.png", scale = 1, width = 10.24, height = 7.68)

```

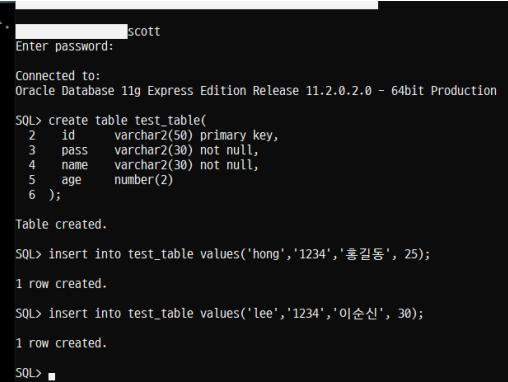
https://s3-us-west-2.amazonaws.com/secure.notion-static.com/cf9566db-1361-4dc8-8566-e2dda378faa0/09_%EC%A0%95%ED%98%95%EA%B3%BC_%EB%B9%84%EC%A0%95%ED%98%95_%EB%8D%B0%EC%9D%B4%ED%84%BO.pdf

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/058d5390-f04d-4c03-8b4c-1417d3dba1e2/chap09_Form_allinformal.r

정형 - 일정 형식을 갖춰서 저장돼있는 데이터

비정형 - text, 영상, 음원 파일 등 데이터

```
12 # cmd 창 들어가서 sqlplus로 scott계정 접속해서 sql 작성해주고 create해줌.
13 """
14
15 SQL>
16 create table test_table(
17   id      varchar2(50) primary key,
18   pass    varchar2(30) not null,
19   name    varchar2(30) not null,
20   age     number(2)
21 );
22
23 """
24
25 # 단계2 : 레코드 추가와 조회하기.
26 # SQL>insert into test_table values('hong','1234','홍길동', 25);
27 # SQL>insert into test_table values('lee ','1234','이순신', 30);
28
29 # 단계3 : transaction 처리 - commit;
30 # SQL>commit;
```



```
C:\WINDOWS\system32>sqlplus

SQL*Plus: Release 11.2.0.2.0 Production on 수 10월 19 12:43:19 2022

Copyright (c) 1982, 2014, Oracle. All rights reserved.

Enter user-name: scott
Enter password: tiger

Connected to:
Oracle Database 11g Express Edition Release 11.2.0.2.0 - 64bit Production

SQL> create table test_table(
2   id      varchar2(50) primary key,
3   pass    varchar2(30) not null,
4   name    varchar2(30) not null,
5   age     number(2)
6 );

Table created.

SQL> insert into test_table values('hong','1234','홍길동', 25);

1 row created.

SQL> insert into test_table values('lee ','1234','이순신', 30);

1 row created.
```

```
SQL> commit;  
Commit complete.  
SQL>
```

```
# Oracle 연동을 위한 R 패키지 설치.  
  
# 1) 패키지 설치  
#   - RJDBC 패키지를 사용하기 위해서는 우선 java를 설치해야 한다.  
install.packages("rJava")  
install.packages("DBI")  
install.packages("RJDBC")  
  
# 2) 패키지 로딩  
Sys.setenv(JAVA_HOME='C:/Program Files/Java/jdk-11.0.16.1')  
library(DBI)  
library(rJava)  
library(RJDBC) # rJava에 의존적이다(rJava 먼저 로딩돼있어야 한다).  
  
# 3) Oracle 연동  
  
### Oracle 11g Ex.  
# driver  
drv <- JDBC("oracle.jdbc.driver.OracleDriver",  
            "C:/oraclexe/app/oracle/product/11.2.0/server/jdbc/lib/ojdbc6.jar") # driver 이름, 경로  
  
# db 연동(driver, url, id, pwd)  
conn <- dbConnect(drv, "jdbc:oracle:thin:@//localhost:1521/xe", "scott", "tiger")  
  
# (1) 모든 레코드 검색  
query <- "select * from test_table"  
dbGetQuery(conn, query)
```

```
> dbGetQuery(conn, query)  
  ID PASS  NAME AGE  
1 hong 1234 홍길동 25  
2 lee 1234 이순신 30
```

Index of /bin/windows/contrib/3.4
 <https://cran.rstudio.com/bin/windows/contrib/3.4/>

분석절차와 통계지식

② 연구가설(대립가설)

'차이가 있다.' 또는 '효과가 있다.'

- 긍정적 형태 진술(예, H_1 : 영양소별 효과의 차이는 있다.)

※ 논문에서 **연구가설 제시**, 귀무가설을 통해서 가설 검정

논문에서 주장하고 싶은 가설을 연구가설(대립가설)이라고 한다.

연구 논문, 약 등의 효과가 없다는 것이 귀무가설

• 가설의 유형

① 귀무가설(영가설)

'두 변수간의 관계가 없다.' 또는 '차이가 없다.'

- 부정적 형태 진술(예, H_0 : 교육수준에 따라서 사회 정책에 대한 비판적 태도에서 차이가 없다.)

● 사회과학분야 임계값 : $\alpha=0.05(p<0.05(5\% \text{미만}))$

➢ 적어도 96마리 이상 효과

● 의.생명분야 임계값 : $\alpha=0.01(99\% \text{ 신뢰도 보장})$

➢ 적어도 99마리 이상 효과

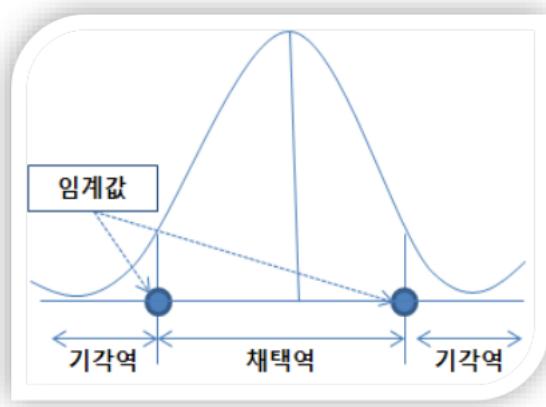
의학에서는 더 높은 기준을 부여하고 있다. 0.01%(99%신뢰도를 기준값으로)

통계적으로 유의하다 = 대립가설 채택 = 귀무가설 기각

$p(0.04) < \alpha(0.05) \rightarrow$ 귀무가설(영가설) 기각

- 영양소별 효과의 차이가 있을 확률이 높기 때문에 연구가설 채택
- 이때 "통계적으로 유의하다"라고 해석, $p<0.01$ 이면 매우 유의하다.

통계 사전 지식



- 1) 통계학 개요
- 2) 모집단과 표본
- 3) 추정과 검정
- 4) 가설검정 오류
- 5) 검정통계량
- 6) 정규분포
- 7) 모수 & 비모수

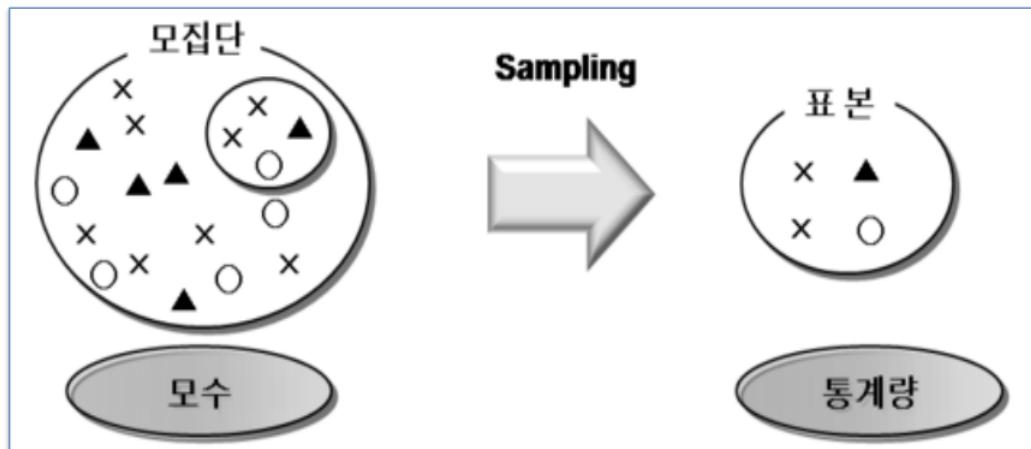
통계학 - 인과관계 규명

- 통계학(Statistics)?
 - ✓ 논리적 사고와 객관적인 사실에 의거, 확률 기반 인과관계 규명
 - ✓ 통계 모형에 의해 추정과 검정에 대한 통계적 방법을 사용

인구조사는 전수조사한다.

표본을 잘 추출하는 것이 핵심.

- 모집단과 표본
 - Sampling : 표본추출



- 모수와 통계량 표현

구분	모수(모집단)	통계량(표본)
의미	모집단의 특성을 나타내는 수치	표본의 특성을 나타내는 수치
표기	그리스, 로마자	영문 알파벳
평균	μ (모평균)	\bar{x} (표본의 평균)
표준편차	σ (모표준편차)	S (표본의 표준편차)
분산	σ^2 (모분산)	S^2 (표본의 분산)
대상 수	N(사례수)	n(표본수)

1000명 정도면 유한모집단의 경우

- 표본크기 결정
 - 유한모집단의 경우

$$n \geq \frac{N}{\left(\frac{e}{k}\right)^2 \frac{N-1}{P(1-P)} + 1}$$

- 무한모집단의 경우

$$n \geq \frac{1}{\left(\frac{e}{k}\right)^2 \frac{1}{P(1-P)}}$$

N : 모집단의 크기

e : 요구정밀도

P : 모집단의 비율

k : 신뢰수준($\alpha=0.05$ 일 때 $k=1.96$)

100%는 모든 가능성을 열어두는 것이라서 당연히 통계 낼 이유가 없다.

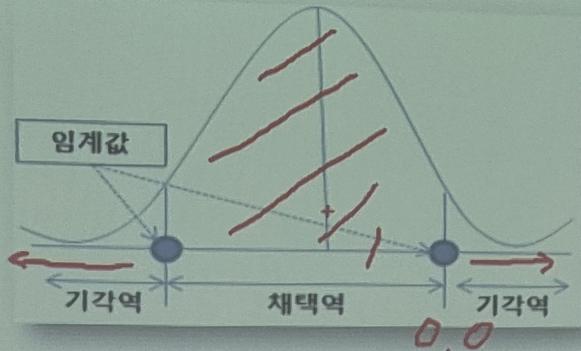
-3%는 29.4%, 32.4%로 조사, +3%는 35.4% 신뢰구간

신뢰수준 95%

예)) 대통령 후보의 지지율 여론조사에서 특정 후보의 지지율이 95% 신뢰수준에서 표본오차 $\pm 3\%$ 범위에서 32.4%로 조사 되었다고 가정한다면 실제 지지율은 29.4%~35.4%(-3%~+3%)사이에 나타날 수 있다는 의미이다. 여기서 95% 정도는 이 범위의 지지율을 신뢰할 수 있지만 5% 수준에서는 틀릴 수도 있는 의미이다.

→ 신뢰수준 95%, 신뢰구간 29.4%~35.4%, 표본오차 $\pm 3\%$

- 임계값에 따른 기각역과 채택역
 - 임계값(Critical value) : 귀무가설 채택 or 기각 기준점
 - 채택역(Acceptance region) : 임계값 기준 채택(귀무가설) 범위
 - 기각역(Critical region) : 기각 범위



가운데가 채택영역.

평균을 제곱 - 분산

분산에 루트 - 표준편차

```
# 중앙값(median): 중위수 - 모든 데이터를 크기 순서대로 정렬시킨 후 가운데 있는 값을 의미.
# ex) 100, 100, 54, 50, 52 : 중앙값 -> 54
median(score3) # 54

# ex) 6, 6, 7, 8, 9, 10
num <- c(6, 6, 7, 8, 9, 10)
median(num) # 7.5=(7+8)/2

# 편차(Deviation) : 평균값을 기준으로 각 값의 차이. 즉, 평균과 해당값의 차이.

# 분산(Variance) : 편차 값을 제곱해서 마이너스 값을 플러스 값으로 바꾼 후 평균을 구하는 방법.
# ex) ((100-71.2)^2+(100-71.2)^2+(54-71.2)^2+(50-71.2)^2+(52-71.2)^2) / 5 = 554.56
score <- c(100, 100, 54, 50, 52)
mean(score) # 71.2
var(score) # 693.2 # 5개가 아닌 5-1을 해서 나누어줌.
sd(score) # 표준편차

# 표준편차(Standard Deviation:SD) : 분산 값에 루트를 적용해서 제곱을 제거한 값. ex) 23.55(sqrt(554.56))

# 자유도(degree of freedom) : 표본의 분산과 표준편차를 계산할 때 나누는 분모의 수를 (모집단-1)개로 계산하여 주어진 데이터에서 표본을 자유롭게 뽑을 수 있는 경우의 수를
```