데이터 시각화_그래프

숙명여자대학교 경영학부 오중산

그래프 개요

? 목적에 따른 그래프 유형 구분 ex) 코로나 19 확진자수

구성된 결과값들의 비교 trend

	구성비교	변화 추이 확인	분포 확인	연관성 확인
산점도			O	O
막대 그래프	O	O	O	
선 그래프		O		
상자 그래프			O	

: 두 개의 변수 간의 관계

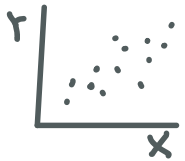
그래프 개요

- ggplot2를 이용한 그래프 레이어 구조
  - 1단계(필수): 데이터 선정 : 어떤 data frame을 대상으로 그릴 것이냐
  - 2단계(필수): X축과 Y축 변수 지정 : aes( )
  - 3단계(필수): 그래프 유형 선정 : ggplot +로 연결
  - 4단계(선택): 옵션(색상/크기 등)

+ 제목. 축제목

산점도(scatter plot)

산점도(scatter plot)란?



=> 양의 관계

- 계량척도로 측정된 두 변수 간의 관계를 이차원 평면에 점으로 표시한 그래프 => 시각적으로 알 수 있음

num/int X.Y

mpg를 이용해서 산점도 그리기

- 배기량(displ, X축)에 따른 고속도로 연비(highway, Y축) 산점도
  - library(ggplot2)
  - ggplot(mpg, aes(displ, highway)) + geom_point()

① ② 배기량 ③

↳ 산점도

    - 1단계: 데이터 선정
    - 2단계: 두 개 축 지정
    - 3단계: 그래프 유형 선정
    - 주의 사항: geom_point 뒤에 () 붙이는 것과 ggplot2의 함수는 +로 연결됨

* ggplot2는 패키지, ggplot은 함수

산점도(scatter plot)

- X축과 Y축 범위 지정하여 산점도 그리기

조건1 조건 2
하한 상한

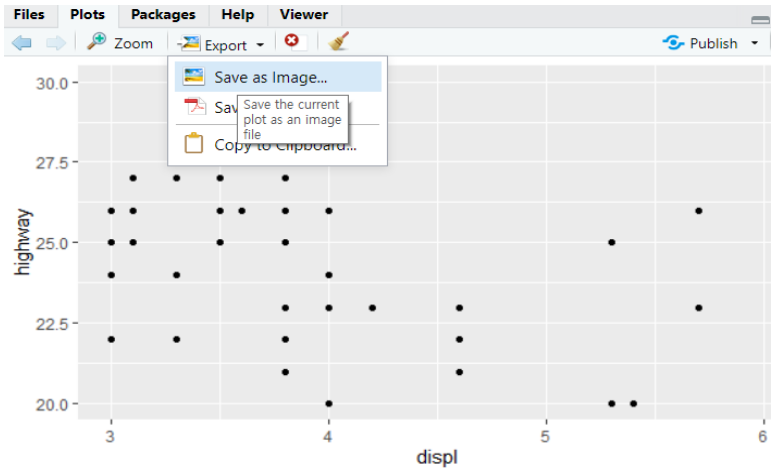
  - ggplot(mpg, aes(displ, highway)) + geom_point() + xlim(3, 6) + ylim(20, 30)

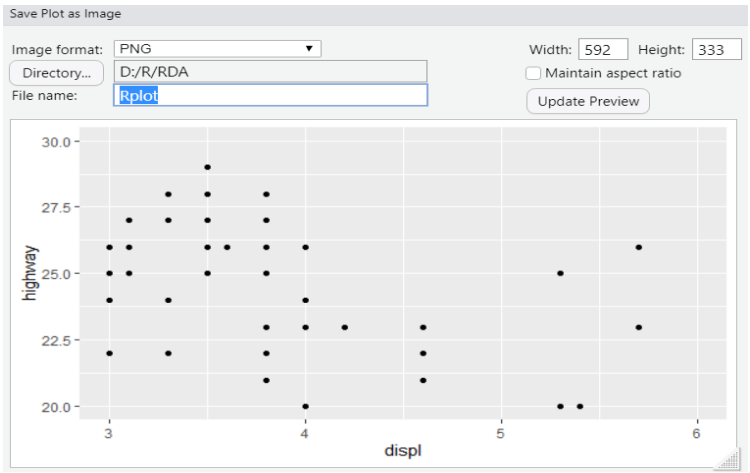
X축 변수에 제약을 가하겠다

배기량 3에서 6 사이 고속도로 연비 20이상 30 이하

    - 4단계: X축은 3~6, Y축은 20~30으로 제약을 가함
    - Console 창에서 경고를 통해 축 범위에 대한 제약으로 인해 제시되지 못하는 사례를 알려줌

- 그래프 이미지 저장





산점도(scatter plot)

점의 색상 변경

forward
rear
4

X축 Y축

ggplot(mpg, aes(displ, highway, color = drv)) + geom_point() + xlim(3, 6) + ylim(20, 30)

- 세 가지 구동방식(drv)에 따라 점의 색상을 다르게 표현
- 색상을 구분하기 위한 기준 변수는 반드시 문자나 범주형 척도로 측정되어야 함

=> 비계량형 척도

drv chr factor

형태와 크기 조정

Shape size

① ② ③

ggplot(mpg, aes(displ, highway, color = drv)) + geom_point(aes(shape = drv, size = fuel)) + xlim(3, 6) + ylim(20, 30)

↳ 빨강 ○
초록 △
파랑 □

- 구동방식에 따라 색상과 점의 형태를 모두 다르게 표현함
- 연료 종류(5개)에 따라 점의 크기를 구분함

산점도

(실습문제) mpg 데이터 프레임을 이용한 산점도 그리기

- 도심연비(city)를 X축에, 고속도로연비(highway)를 Y축에 두고 산점도를 그리시오.

```
ggplot(mpg, aes(city, highway)) + geom_point()
```

- 도심연비 상한을 30으로, 고속도로 연비 상한을 40으로 설정하시오.

```
ggplot(mpg, aes(city, highway)) + geom_point() + xlim(0,30) + yjim(0,40
```

- 기통수(cyl)를 기준으로 색상을 구분하시오.

```
ggplot(mpg, aes(city, highway, color = cyl)) + geom_point() + xlim(0,30) + ylim(0,40)
```

- 구동방식(drv)을 기준으로 형태를 구분하시오.

```
ggplot(mpg, aes(city, highway, color = cyl)) + geom_point(aes(shape = drv)) + xlim(0,30) + ylim(0,40)
```

- 추세선을 추가하시오.

```
ggplot(mpg, aes(city, highway, color = cyl))+geom_point(aes(shape = drv))+ xlim(0,30) + ylim(0,40) +geom_smooth()
```

산점도

- (실습문제) midwest 데이터 프레임을 이용한 산점도 그리기
  - 전체인구(poptotal)를 X축에, 아시아인구(popasian)를 Y축에 두고 산점도를 그리시오.

```
ggplot(midwest, aes(poptotal, popasian)) + geom_point()
```

  - 전체인구 상한을 35만 명으로, 아시아인구 상한을 5천 명으로 설정하시오.

```
ggplot(midwest, aes(poptotal, popasian)) + geom_point() + xlim(0, 350000) + ylim(0, 5000)
```

  - 주(state)를 기준으로 색상과 형태를 구분하시오.

```
ggplot(midwest, aes(poptotal, popasian, color = shape)) + geom_point(aes(shape = state)) + xlim(0, 350000) + ylim(0, 5000)
```

  - 추세선을 넣어 보시오.

```
ggplot(midwest, aes(poptotal, popasian, color = shape)) + geom_point(aes(shape = state)) + xlim(0, 350000) + ylim(0, 5000)
+geom_smooth()
```

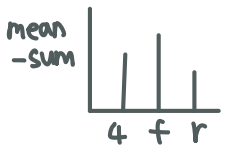
막대 그래프

유형1: 두 변수 막대 그래프 그리기

- X축 변수는 문자나 범주형 척도로 측정하고 , Y축 변수는 계량형 척도로 측정한 후의 요약결과 (예: 평균)
  - ↳ income
  - 예: 사회복지패널데이터분석에서 '성별'에 따른 '월급평균' 막대 그래프 그리기

구동방식에 따른 전체 연비평균

Sum = city + highway



- 새로운 데이터 프레임 만들기
  - df_mpg <- mpg %>% group_by(drv) %>% summarise(mean_sum = mean(sum))
- 평균 막대 그래프 그리기
  - ggplot(df_mpg, aes(drv, mean_sum)) + geom_bar(stat = "identity")

Y축 변수 값을 쓰겠다

막대 그래프

구동방식에 따른 전체 연비평균

내림차순으로 막대 그래프 정리하기

X -Y Y

- ggplot(df_mpg, aes(reorder(drv, -mean_sum), mean_sum)) + geom_bar(stat = "identity")
  - reorder 함수 사용: 내림차순일 경우 Y축 변수명 앞에 –를 붙이고, 오름차순일 경우 붙이지 않음

막대 그래프에 색깔 입히기

- ggplot(df_mpg, aes(reorder(drv, -mean_sum), mean_sum, fill = drv)) + geom_bar(stat = "identity")

⇓
X축

막대 그래프

유형2: 빈도 막대 그래프 그리기

- 개별 변수에 대한 빈도수 확인 ↱ 비계량척도일 필요 X
  - 변수의 측정 척도는 반드시 범주형 /문자형일 필요가 없음
  - qplot과 비교했을 때 다양한 옵션을 적용할 수 있음

차량등급(class)에 따른 빈도 막대 그래프 그리기

midsize / suv / compact …

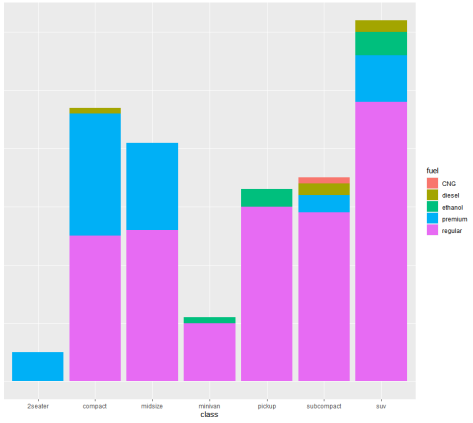
- ggplot(mpg, aes(class)) + geom_bar()
  - stat = "identity“ 불필요함
- 막대 그래프에 색깔 입히기
  - ggplot(mpg, aes(class, fill = class)) + geom_bar()

막대 그래프

- 차량등급(class)에 따른 빈도 막대 그래프 그리기
  - 세 가지 등급(compact, midsize, suv)에 대해서만 빈도 막대 그래프 그리기
    - ggplot(mpg, aes(class, fill = class)) + geom_bar() + xlim(c("compact", "midsize", "suv"))
  - 막대 그래프를 가로로 변경하기
    - ggplot(mpg, aes(class, fill = class)) + geom_bar()+ coord_flip()
      ↳동전 뒤집는
  - 막대 그래프를 거미줄 그래프로 변경하기
    - ggplot(mpg, aes(class, fill = class)) + geom_bar()+ coord_polar()
      ↳ 극

막대 그래프

- 차량등급(class)에 따른 빈도 막대 그래프 그리기



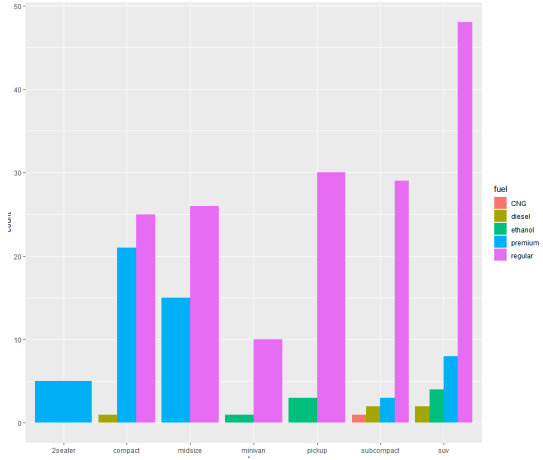
  - 연료 유형에 따른 색상 구분

    ↳ 5개 目 → ㅠ

    - ggplot(mpg, aes(class, fill = fuel)) + geom_bar()

  - 연료 유형에 따라 막대 그래프를 옆으로 쌓기

    - ggplot(mpg, aes(class, fill = fuel)) + geom_bar(position = "dodge")



  - 막대 그래프 크기를 동일하게 조정하기

    - ggplot(mpg, aes(class, fill = fuel)) + geom_bar(position = "fill")

      ↳ Y축 비율

막대 그래프

```
mpg_suv <- mpg %>% group_by(manufacturer) %>% filter(class=="suv")
%>% summarise(mean_city = mean(city)) %>% arrange(-mean_city) %>% head(5)
```

- (실습문제) 회사별로 class가 suv인 차종의 도심연비 평균이 높은 순서대로 5개 회사의 도심연비 평균 막대 그래프를 그리시오
  - 조건1: 내림차순 막대 그래프로 표현할 것
  - 조건2: 회사별로 색상을 구분할 것
  - 조건3: 가로 막대 그래프로 그릴 것
  - 추가문제: 그래프 제목(회사별 suv 도심연비 평균 비교)을 만들고, 축 제목(X축: 제조사, Y축: suv 도심연비 평균)을 만들 것!

41:21

```
ggplot(mpg_suv, aes(reorder(manufacturer, -mean_city), mean_city, fill = manufacturer))
+ geom_bar(stat = "identity") + coord_flip() + labs(title = "회사별 suv 도심연비 평균 비교",
                                                    X = "제조사", Y="suv 도심연비 평균")
```

히스토그램

변수 X 1개 => 계량, 비계량 모두 가능 ↳ 계량형

X : 계량척도 => 3000개
구간 50 => 60개) hist

(보통은 막대그래프 사용)

? 빈도 막대 그래프 vs. 히스토그램

- ? 히스토그램은 계량형 척도로 측정된 변수에 대해 구간별 빈도를 구함

? mpg에서 고속도로연비 히스토그램 그리기

↳ 구간

- ? 기본적인 형태: ggplot(mpg, aes(highway)) + geom_histogram(binwidth = 1)

( ) => default 값

- ? 막대 색상 변경 및 그래프 제목과 축제목 지정
  - ? ggplot(mpg, aes(highway)) + geom_histogram(binwidth = 1, fill = "yellow", colour = "red") +
    labs(title = "고속도로 연비 히스토그램", x = "고속도로 연비", y = "빈도")

↳ 겉에 싸는 색깔

  - ? 여러 개 막대에 대해서 각각 색상을 구분하려면 명령문이 복잡해짐 !

* colors( ) 색 뭐 있는지 알 수 있음

선 그래프

선 그래프의 용도

- 시간의 흐름에 따른 시계열 데이터 (time series data)를 표현하는데 적합

economics 데이터 프레임 이용하여 선 그래프 그리기

↪ mpg, midwest

- ggplot2에 들어 있는 내장 데이터 프레임이며 , 주요 변수는 다음과 같음
  - pce: personal consumption expenditures, in billions of dollars : 개인 소비 지출
  - pop: total population, in thousands : 총 인구수
  - psavert: personal savings rate : 개인 저축률
  - uempmed: median duration of unemployment, in weeks : 실업 지속 기간 중위수
  - unemploy: number of unemployed in thousands : 실업자 수

선 그래프

economics 데이터 프레임 이용하여 선 그래프 그리기

시간에 따른 실업자 수 현황

date(X) unemploy(Y)

ggplot(economics, aes(date, unemploy)) + geom_line() ⇒ 싸이클도 있으며 전반적으로 상승세

df X Y

점(point) 추가하기

ggplot(economics, aes(date, unemploy)) + geom_line() + geom_point()

선과 점에 색상 입히기

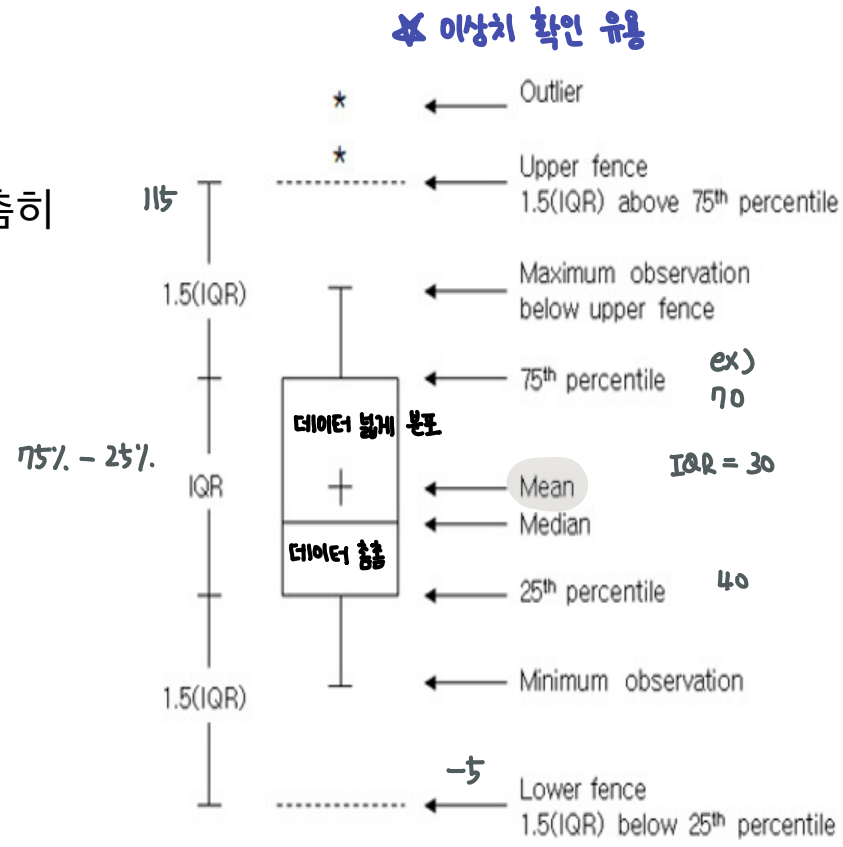
ggplot(economics, aes(date, unemploy)) + geom_line(color = “red”) + geom_point(color = “darkred”)

참고: R에서 제공하는 색상 종류를 확인하려면 colors() 실행

상자 그래프

상자 그래프 설명

- 데이터의 분포에 대해 확인할 수 있음
- 중앙값(median)이 아래에 있으면 하위 25~50%가 촘촘히 분포하고, 위에 있으면 상위 50~75%가 촘촘히 분포
- IQR이 크면(상자가 크면) 데이터가 넓게 분포
- 이상치(outlier)에 대해서도 판단할 수 있음



상자 그래프

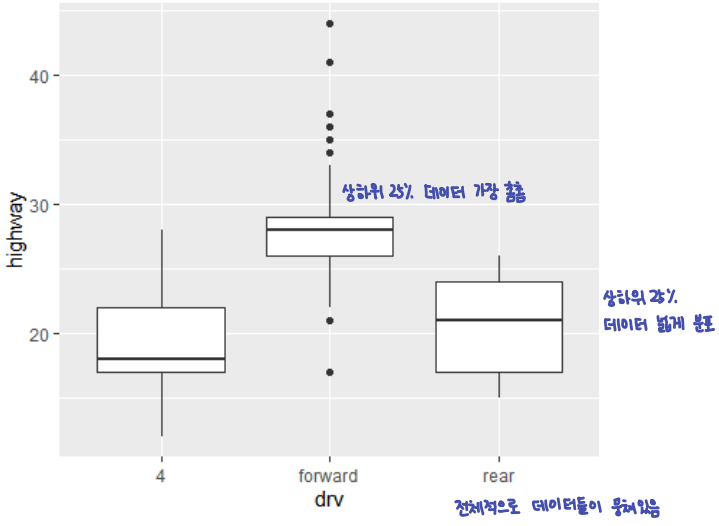
mpg를 이용한 상자 그래프 그리기

구동방식별 고속도로 연비 상자 그래프 (색상 추가)

drv highway

ggplot(mpg, aes(drv, highway, fill = drv)) + geom_boxplot()

↳ 구동방식 별로 색깔 달리해라



이상치를 빨간색으로 표시하기

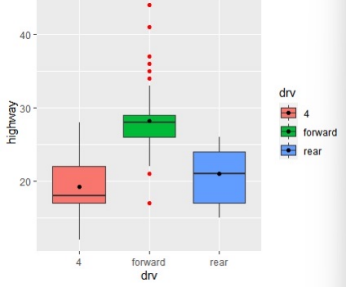
ggplot(mpg, aes(drv, highway, fill = drv)) + geom_boxplot(outlier.colour ="red")

Color

평균을 점의 형태로 추가하기

ggplot(mpg, aes(drv, highway, fill = drv)) + geom_boxplot(outlier.colour ="red") +

stat_summary(fun = "mean", geom = "point")



그래프 실습

- corona19 데이터 프레임 그래프 그리기

corona19 <- read.csv("corona19.csv", stringsAsFactors = F)

  - corona19.csv 데이터 불러오고, date 척도 변경
  - 산점도: X축(new_tests)과 Y축(new_cases)
  - 막대그래프: X축(date)과 Y축(new_cases)
  - 선 그래프: X축(date)과 다양한 Y축 변수
    - new_daths / total_deaths
    - positive rate / reproduction rate
    - total_vaccinations / people_fully_vaccinated

date	일자
total_cases	누적 확진자수
new_cases	신규 확진자수
total_deaths	누적 사망자수
new_deaths	신규 사망자수
new_tests	신규 검사자수
total_tests	누적 검사자수
positive rate	확진율
reproduction rate	재생산지수(전염력)
total_cases_per_million	백만명당 누적 확진자수
new_cases_per_million	백만명당 신규 확진자수
total_deaths_per_million	백만명당 누적 사망자수
new_deaths_per_million	백만명당 신규 사망자수
total_vaccinations	누적 백신접종자수 1차접종 + 2차접종
people_fully_vaccinated	누적 백산접종완료자수

2차접종까지 끝낸 사람

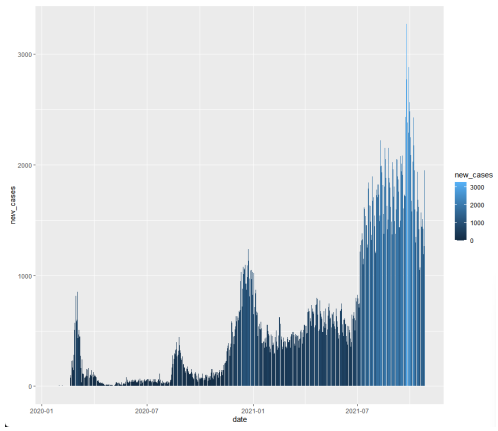
산점도

```
ggplot(corona19, aes(new_tests, new_cases)) + geom_point() + xlim(10000, 60000) + ylim(0,3000)
                                                          + geom_smooth()
```

↳ 추세

막대그래프

```
ggplot(corona19, aes(date, new_cases, fill = new_cases))
                                  + geom_bar(stat = "identity")
```



선그래프

```
ggplot(corona19, aes(date, new_cases)) + geom_line(color = "red")
                                       + geom_point(color = "blue")
```

* 선그래프 두개 같이 안그리는 이유

  최댓값 scale이 다름

  ⇒ scale이 비슷해야 함

  ex) 1. 0 ~ 100

  2. 150 ~ 300 이러면 안됨

```
ggplot(corona19_new, aes(date, number, color = type)) + geom_line() + geom_point()
```

지도시각화

숙명여자대학교 경영학부 오중산

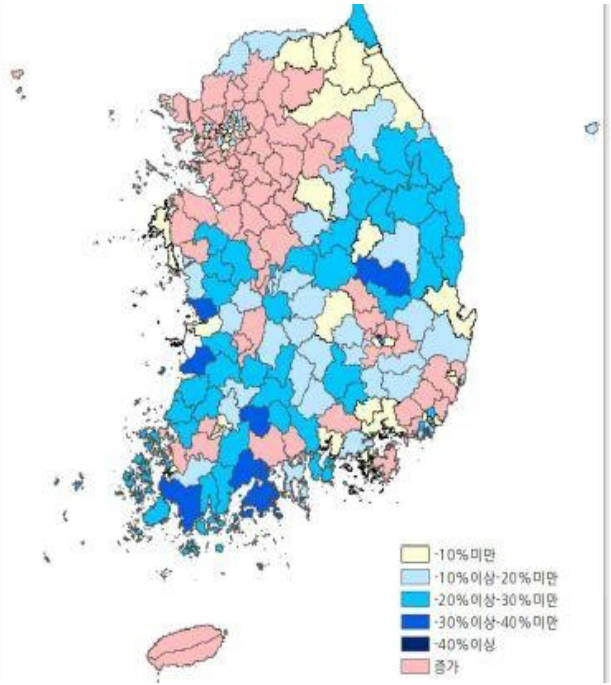
단계구분도(choropleth map) 소개

- 단계구분도의 정의

  지도 상에서 통계치 결과에 따라 음영 /색상/패턴별 차이를 다르게 보여주는 지도

〈2000~2020년 간 인구감소율 분포〉

인구감소율	지자체 수	인구수
-10% 미만	42개	1,243만명
-10~-20% 미만	52개	801만명
-20~-30% 미만	46개	286만명
-30~-40% 미만	10개	67만명
-40 이상	1개	17만명
계	151개	2,416만명



20년 동안 인구감소와 관련된 단계구분도 (출처: 국토연구원)

단계구분도 데이터: USArrests

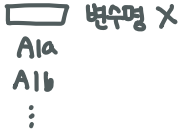
- USArrests 데이터

내장데이터

- 1973년 미국 주(state)별 강력 범죄 용의자 체포 관련 데이터
  - crime <- USArrests
    - 변수: Murder, Assault, Urbanpop, Rape
      살인 폭행 도시인구비율 성범죄
- 변수명이 없는 주(state)에 대해 변수명 state 부여하고, 주 명칭을 소문자로 바꾸기
  - library(tibble) : dplyr에 내장된 패키지
  - crime <- rownames_to_column(crime, var = "state")
    : 변수명 부여
    : 행의 이름을 행에게
  - crime$state <- tolower(crime$state)

ex) Alabama → alabama

ex) 히개



USArrests 단계구분도 실습

↳ 내장 되어있는 지도 데이터

- state 데이터 활용하기

  - maps 패키지 내장 데이터(지도)이며, 지역별 위도와 경도 정보 등을 담고 있음 (미국)

    - install.packages("maps") / library(maps)

  - ggplot2 패키지의 map_data 함수를 이용하여 state를 데이터 프레임 형태로 저장

    - library(ggplot2)

    - states_map <- map_data("state")

      ↳ maps 패키지 내장 데이터

      - states_map에는 주별로 소속된 15,000개가 넘는 여러 지역의 위치 정보가 담겨 있음

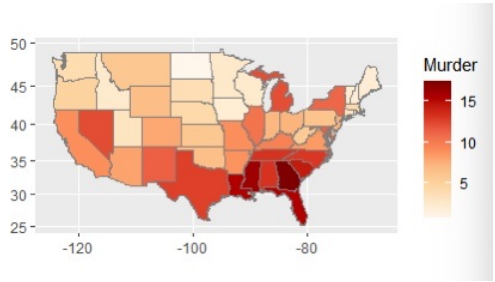
      - states_map은 형태상 데이터 프레임이지만 내용상 지도로 볼 수 있으며, 이것을 플랫폼으로 삼아 다양한 정보 (예: 범죄율, 소득, 교육수준 등…)를 반영하여 단계구분도로 표현 가능

USArrests 단계구분도 실습

- 강력범죄 단계구분도 만들기

① install.packages ("mapproj")

- map_id 명령어 사용을 위해 mapproj 패키지 설치 및 로딩
- ggChoropleth 함수 사용을 위해 ggiraphExtra 패키지 설치 및 로딩
  - ggChoropleth 함수는 단계구분도를 그려주는 함수



- ggChoropleth를 이용한 단계 구분도 만들기 (ex 시·군·구)
  ↳ 주를 기준으로
  ↳ map은 형태상 df, 내용상 지도
  - ggChoropleth(data = crime, aes(fill = Murder, map_id = state), map = states_map)
    ↳ ggplot2 불러와야함
    - crime 데이터 프레임의 Murder 측정결과를 states_map에 주(state)를 기준으로 반영하여 단계구분도로 표현하라는 의미
    - 주의) aes를 활용하려면 ggplot2 패키지 불러와야 함
  - ggChoropleth(data = crime, aes(fill = Assault, map_id = state), map = states_map, interactive = T)
    - Export에서 'Save As Web Page…" 선택하여 html 형식으로 저장

인터랙티브 그래프

숙명여자대학교 경영학부 오중산

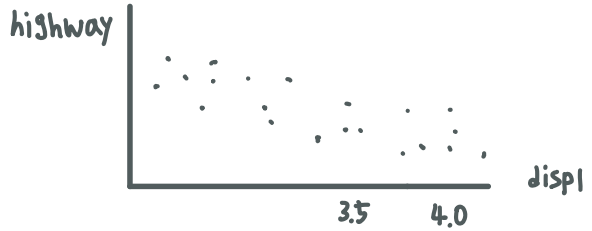
인터랙티브 산점도와 막대그래프 그리기

- 관련 패키지 plotly 설치 및 불러오기
  - install.packages("plotly")
  - library(plotly)
- 인터랙티브 산점도와 막대그래프 그리기 과정
  - Step1: ggplot을 이용하여 그래프 객체 만들기
  - Step2: 만들어진 그래프 객체를 ggplotly함수로 실행하기

list 형태

인터랙티브 산점도와 막대그래프 그리기

- mpg를 이용한 인터랙티브 그래프 그리기



인터랙티브 산점도 그리기

↱ 리스트 형태의 객체

```
p1 <- ggplot(data = mpg, aes(displ, highway, col = drv)) + geom_point()
```

배기량 고속도로연비 4.f.r

```
ggplotly(p1)
```

인터랙티브 막대그래프 그리기 변수 1개 빈도

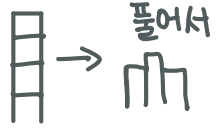
```
p2 <- ggplot(mpg, aes(class, fill = class)) + geom_bar()+ coord_flip()
```

기기

```
ggplotly(p2)
```

```
p3 <- ggplot(mpg, aes(class, fill = fuel)) + geom_bar(position = "dodge")
```

```
ggplotly(p3)
```



① ggplot 함수 통해 객체에 저장 ② ggplotly로 불러오기

인터랙티브 산점도와 막대그래프 그리기

- diamonds를 이용한 인터랙티브 막대그래프 그리기
  - diamonds 데이터 소개
    - ggplot2에 내장된 데이터
    - str()를 이용한 diamonds 데이터 검토
  - p4 <-ggplot(data = diamonds, aes(cut, fill = clarity)) + geom_bar(position = "dodge")

    둘 다 범주형
  - ggplotly(p4)
  - p5 <- ggplot(data = diamonds, aes(cut, fill = color)) + geom_bar(position = "dodge")
  - ggplotly(p5)

인터랙티브 시계열 그래프 그리기

- 관련 패키지 dygraphs 설치 및 불러오기
  - install.packages("dygraphs")
  - library(dygraphs)
- economics를 이용한 인터랙티브 시계열 그래프 그리기
  - 필요한 xts 내장패키지 불러오기
  - 패키지와 동명의 xts 함수로 시계열 그래프 그리기 위한 객체 만든 후 실행하기

    ↳ 시계열 그래프를 그리기 위한 객체를 그리는 함수

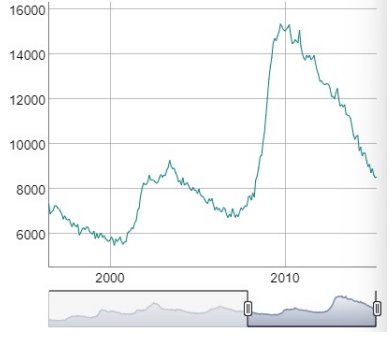
    - eco <- xts(economics$unemploy, order.by = economics$date)

      : 시간 순서에 따라서 실업자 수를 보여줘

      - date에 따라 unemploy에 시간속성을 부여한 후 정렬

        => date 속성을 반영하겠다
    - dygraph(eco) %>% dyRangeSelector()

      구간을 선택한다



인터랙티브 시계열 그래프 그리기

- corona19를 이용한 복수의 인터랙티브 시계열 그래프 그리기

  - 일자(date)별 누적 확진자와 누적 백신접종회수 시계열 그래프 함께 그리기

    - eco_a <- xts(corona19$total_cases, order.by = corona19$date)

    - eco_b <- xts(corona19$total_vaccinations/100, order.by = corona19$date)

      - scale을 비슷하게 하기 위해 100으로 나눔 ※ 해석할 때만 주의하면 됨

    - eco_c <- cbind(eco_a, eco_b)
      : column을 합친다

      ex) | eco_a
          | eco_b

      - cbind는 left_join과 기능은 동일한데, 위에서와 같이 동일한 변수 없을 때 사용이 편함

    - colnames(eco_c) <- c("total_cases", "total_vaccinations") V1으로 되어있는 것 이름 변경

      - colnames는 변수명을 변경하는, rename과 동일한 기능의 함수

    - dygraph(eco_c) %>% dyRangeSelector()