데이터 시각화_그래프

숙명여자대학교 경영학부 오중산

그래프 개요

- 목적에 따른 그래프 유형 구분

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

그래프 개요

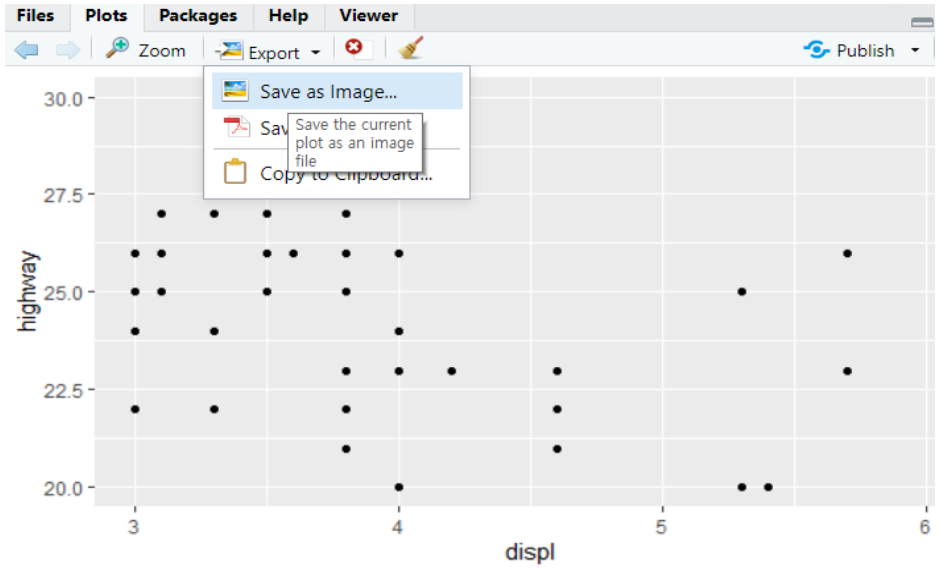
- ggplot2를 이용한 그래프 레이어 구조
  - 1단계(필수): 데이터 선정
  - 2단계(필수): X축과 Y축 변수 지정
  - 3단계(필수): 그래프 유형 선정
  - 4단계(선택): 옵션(색상/크기 등)

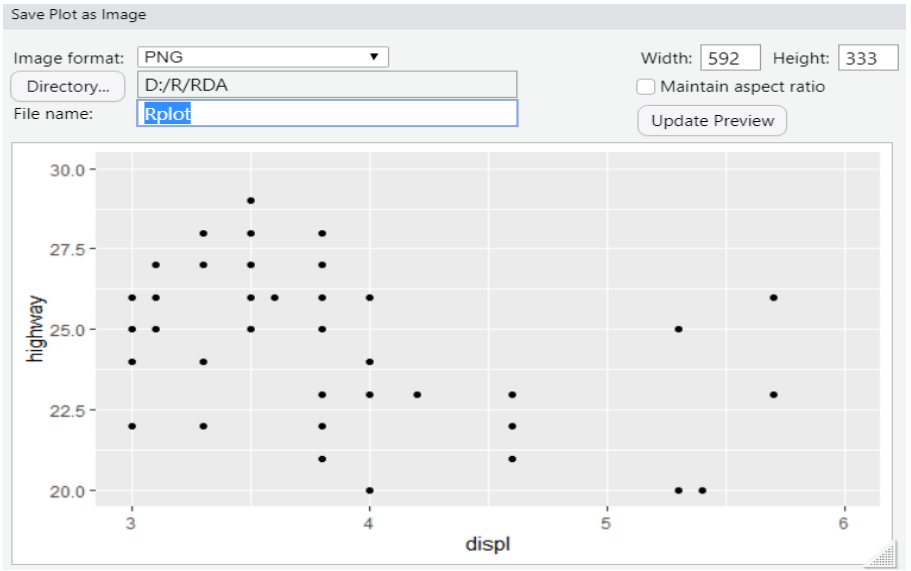
산점도(scatter plot)

- 산점도(scatter plot)란?
  - 계량척도로 측정된 두 변수 간의 관계를 이차원 평면에 점으로 표시한 그래프
- 기존에 만들어 둔 mpg 이용를 해서 산점도 그리기
  - 배기량(X축)에 따른 고속도로 연비(Y축) 산점도
    - library(ggplot2)
    - ggplot(mpg, aes(displ, highway)) + geom_point()
      - 1단계: 데이터 선정
      - 2단계: 두 개 축 지정
      - 3단계: 그래프 유형 선정
      - 주의 사항: geom_point 뒤에 () 붙이는 것과 ggplot2의 함수는 +로 연결됨

산점도(scatter plot)

- X축과 Y축 범위 지정하여 산점도 그리기
  - ggplot(mpg, aes(displ, highway)) + geom_point() + xlim(3, 6) + ylim(20, 30)
    - 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
  - 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
  - 형태와 크기 조정
    - ggplot(mpg, aes(displ, highway, color = drv)) + geom_point(aes(shape = drv, size = fuel)) + xlim(3, 6) + ylim(20, 30)
      - 구동방식에 따라 색상과 점의 형태를 모두 다르게 표현함
      - 연료 종류(5개)에 따라 점의 크기를 구분함

산점도

- (실습문제) mpg 데이터 프레임을 이용한 산점도 그리기
  - 도심연비를 X축에, 고속도로 연비를 Y축에 두고 산점도를 그리시오.
  - 도심연비 상한을 30으로, 고속도로 연비 상한을 40으로 설정하시오.
  - 기통수(cyl)를 기준으로 색상을 구분하시오.
  - 구동방식(drv)을 기준으로 형태를 구분하시오.
  - 추세선을 추가하시오.

산점도

- (실습문제) midwest 데이터 프레임을 이용한 산점도 그리기
  - 전체인구(poptotal)를 X축에, 아시아인구(popasian)를 Y축에 두고 산점도를 그리시오.
  - 전체인구 상한을 35만 명으로, 아시아인구 상한을 5천 명으로 설정하시오.
  - 주(state)를 기준으로 색상과 형태를 구분하시오.
  - 추세선을 넣어 보시오.

막대 그래프

- 유형1: 두 변수 막대 그래프 그리기
  - X축 변수는 문자나 범주형 척도로 측정하고, Y축 변수는 계량형 척도로 측정한 후의 요약결과(예: 평균)
    - 예: 사회복지패널데이터분석에서 ‘성별’에 따른 ‘월급평균’ 막대 그래프 그리기
- 구동방식에 따른 전체 연비평균
  - 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")

막대 그래프

- 구동방식에 따른 전체 연비평균
  - 내림차순으로 막대 그래프 정리하기
    - 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")

막대 그래프

- 유형2: 빈도 막대 그래프 그리기
  - 개별 변수에 대한 빈도수 확인
    - 변수의 측정 척도는 반드시 범주형/문자형일 필요가 없음
    - qplot과 비교했을 때 다양한 옵션을 적용할 수 있음
- 차량등급(class)에 따른 빈도 막대 그래프 그리기
  - 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()
  - 막대 그래프를 거미줄 그래프로 변경하기
    - ggplot(mpg, aes(class, fill = class)) + geom_bar()+ coord_polar()

막대 그래프

- 차량등급(class)에 따른 빈도 막대 그래프 그리기
  - 연료 유형에 따른 색상 구분
    - 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")

막대 그래프

- (실습문제) 회사별로 suv 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!

히스토그램

- 빈도 막대 그래프 vs. 히스토그램
  - 히스토그램은 계량형 척도로 측정된 변수에 대해 구간별 빈도를 구함
- mpg에서 고속도로연비 히스토그램 그리기
  - 기본적인 형태: ggplot(mpg, aes(highway)) + geom_histogram(binwidth = 1)
  - 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")
    - 여러 개 막대에 대해서 각각 색상을 구분하려면 명령문이 복잡해짐!

선 그래프

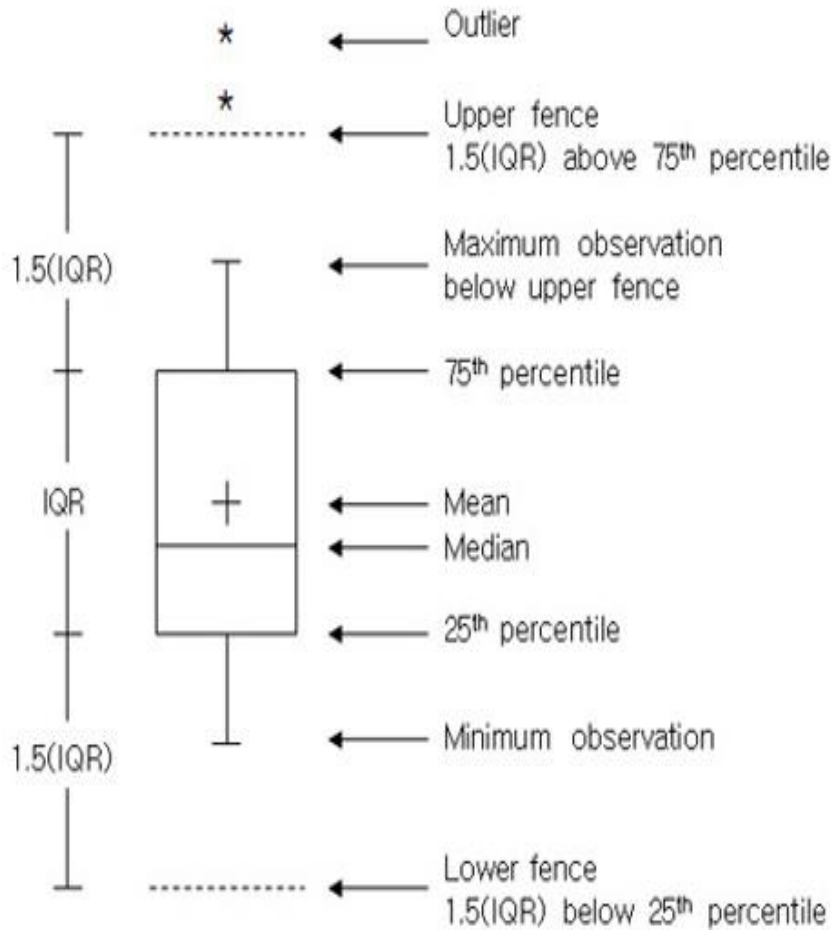
- 선 그래프의 용도
  - 시간의 흐름에 따른 시계열 데이터(time series data)를 표현하는데 적합
- economics 데이터 프레임 이용하여 선 그래프 그리기
  - ggplot2에 들어 있는 내장 데이터 프레임이며, 주요 변수는 다음과 같음
    - pce: personal consumption expenditures, in billions of dollars
    - pop: total population, in thousands
    - psavert: personal savings rate
    - uempmed: median duration of unemployment, in weeks
    - unemploy: number of unemployed in thousands

선 그래프

- economics 데이터 프레임 이용하여 선 그래프 그리기
  - 시간에 따른 실업자 수 현황
    - ggplot(economics, aes(date, unemploy)) + geom_line()
  - 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)
    - ggplot(mpg, aes(drv, highway, fill = drv)) + geom_boxplot()
  - 이상치를 빨간색으로 표시하기
    - ggplot(mpg, aes(drv, highway, fill = drv)) + geom_boxplot(outlier.colour ="red")
  - 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.csv 데이터 불러오고, date 척도 변경
  - 산점도: X축(new_tests)과 Y축(new_cases)
  - 막대 그래프: 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
people_fully_vaccinated	누적 백산접종완료자수