1. mpg 데이터를 이용해서 분석 문제를 해결해보세요.

mpg 데이터는 연비를 나타내는 변수가 hwy(고속도로 연비), cty(도시 연비) 두 종류로 분리되어 있습니다. 두 변수를 각각 활용하는 대신 하나의 통합 연비 변수를 만들어 분석하려고 합니다.

• Q1. mpg 데이터 복사본을 만들고, cty와 hwy를 더한 '합산 연비 변수'를 추가하세요.

• Q2. 앞에서 만든 '합산 연비 변수'를 2로 나눠 '평균 연비 변수'를 추가세요.

• Q3. '평균 연비 변수'가 가장 높은 자동차 3종의 데이터를 출력하세요.

• Q4. 1~3번 문제를 해결할 수 있는 하나로 연결된 dplyr 구문을 만들어 출력하세요. 데이터는 복사본 대신 mpg 원본을 이용하세요.

• Q5. mpg 데이터의 class는 "suv", "compact" 등 자동차를 특징에 따라 일곱 종류로 분류한 변수입니다. 어떤 차종의 연비가 높은지 비교해보려고 합니다. class별 cty 평균을 구해보세요.

• Q6. 앞 문제의 출력 결과는 class 값 알파벳 순으로 정렬되어 있습니다. 어떤 차종의 도시 연비가 높은지 쉽게 알아볼 수 있도록 cty 평균이 높은 순으로 정렬해 출력하세요.

• Q7. 어떤 회사 자동차의 hwy(고속도로 연비)가 가장 높은지 알아보려고 합니다. hwy 평균이 가장 높은 회사 세 곳을 출력하세요.

• Q8. 어떤 회사에서 "compact"(경차) 차종을 가장 많이 생산하는지 알아보려고 합니다. 각 회사별 "compact" 차종 수를 내림차순으로 정렬해 출력하세요.

1. 미국 동북중부 437개 지역의 인구통계 정보를 담고 있는 midwest 데이터를 사용해 데이터 분석 문제를 해결해 보세요. midwest는 ggplot2 패키지에 들어 있습니다.

• 문제1. popadults는 해당 지역의 성인 인구, poptotal은 전체 인구를 나타냅니다. midwest 데이터에 '전체 인구 대비 미성년 인구 백분율' 변수를 추가하세요.

• 문제2. 미성년 인구 백분율이 가장 높은 상위 5개 county(지역)의 미성년 인구 백분율을 출력하세요.

• 문제3. 분류표의 기준에 따라 미성년 비율 등급 변수를 추가하고, 각 등급에 몇 개의 지역이 있는지 알아보세요.

|  |  |
| --- | --- |
| 분류 | 기준 |
| large | 40% 이상 |
| middle | 30% ~ 40% 미만 |
| small | 30% 미만 |

• 문제4. popasian은 해당 지역의 아시아인 인구를 나타냅니다. '전체 인구 대비 아시아인 인구 백분율' 변수를 추가하고, 하위 10개 지역의 state(주), county(지역명), 아시아인 인구 백분율을 출력하세요

1. mpg 데이터 원본에는 결측치가 없습니다. 우선 mpg 데이터를 불러와 몇 개의 값을 결측치로 만들겠습니다. 아래 코드를 실행하면 다섯 행의 hwy 변수에 NA가 할당됩니다.

mpg <- **as.data.frame**(ggplot2::mpg) # mpg 데이터 불러오기

mpg[**c**(65, 124, 131, 153, 212), "hwy"] <- NA # NA 할당하기

결측치가 들어있는 mpg 데이터를 활용해서 문제를 해결해보세요.

• Q1. drv(구동방식)별로 hwy(고속도로 연비) 평균이 어떻게 다른지 알아보려고 합니다. 분석을 하기 전에 우선 두 변수에 결측치가 있는지 확인해야 합니다. drv 변수와 hwy 변수에 결측치가 몇 개 있는지 알아보세요.

• Q2. filter()를 이용해 hwy 변수의 결측치를 제외하고, 어떤 구동방식의 hwy 평균이 높은지 알아보세요. 하나의 dplyr 구문으로 만들어야 합니다.

1. 우선 mpg 데이터를 불러와서 일부러 이상치를 만들겠습니다. drv(구동방식) 변수의 값은 4(사륜구동), f(전륜구동), r(후륜구동) 세 종류로 되어있습니다. 몇 개의 행에 존재할 수 없는 값 k를 할당하겠습니다. cty(도시 연비) 변수도 몇 개의 행에 극단적으로 크거나 작은 값을 할당하겠습니다.

mpg <- **as.data.frame**(ggplot2::mpg) # mpg 데이터 불러오기

mpg[**c**(10, 14, 58, 93), "drv"] <- "k" # drv 이상치 할당

mpg[**c**(29, 43, 129, 203), "cty"] <- **c**(3, 4, 39, 42) # cty 이상치 할당

이상치가 들어있는 mpg 데이터를 활용해서 문제를 해결해보세요.

구동방식별로 도시 연비가 다른지 알아보려고 합니다. 분석을 하려면 우선 두 변수에 이상치가 있는지 확인하려고 합니다.

• Q1. drv에 이상치가 있는지 확인하세요. 이상치를 결측 처리한 다음 이상치가 사라졌는지 확인하세요. 결측 처리 할 때는 %in% 기호를 활용하세요.

• Q2. 상자 그림을 이용해서 cty에 이상치가 있는지 확인하세요. 상자 그림의 통계치를 이용해 정상 범위를 벗어난 값을 결측 처리한 후 다시 상자 그림을 만들어 이상치가 사라졌는지 확인하세요.

• Q3. 두 변수의 이상치를 결측처리 했으니 이제 분석할 차례입니다. 이상치를 제외한 다음 drv별로 cty 평균이 어떻게 다른지 알아보세요. 하나의 dplyr 구문으로 만들어야 합니다.