Doit\_part06

LEE SANGIN

2017년 9월 27일

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

#### Doit\_part06

##### 06-2 혼자서해보기

Q1. displ(배기량)이 4 이하인 자동차와 5 이상인 자동차 중 어떤 자동차의 hwy(고속도로 연비)가 평균적으로 더 높은지 확인하기

mpg <- ggplot2::mpg  
  
displ4 <- as.data.frame(mpg %>% filter(displ <= 4))  
displ5 <- as.data.frame(mpg %>% filter(displ >= 5))  
  
mean\_hwy <- data.frame(data = c("displ4","displ5"),  
 Mean = c(mean(displ4$hwy),mean(displ5$hwy)))  
mean\_hwy

## data Mean  
## 1 displ4 25.96319  
## 2 displ5 18.07895

* displ이 4 이하인 자동차가 hwy가 평균적으로 더 높다.

Q2. "audi"와 "toyota" 중 어느 manufacturer의 cty가 평균적으로 더 높은지 확인하기

mpg\_audi <- as.data.frame(mpg %>% filter(manufacturer =="audi"))  
mpg\_toyo <- as.data.frame(mpg %>% filter(manufacturer == "toyota"))  
  
mean\_cty <- data.frame(data = c("mpg\_audi","mpg\_toyo"),  
 Mean = c(mean(mpg\_audi$cty),mean(mpg\_toyo$cty)))  
mean\_cty

## data Mean  
## 1 mpg\_audi 17.61111  
## 2 mpg\_toyo 18.52941

* toyota의 cty가 평균적으로 더 높다.

Q3. "chevrolet", "ford", "honda" 자동차의 고속도로 연비 평균을 알아보려고 합니다. 이 회사들의 자동차를 추출한 뒤 hwy 전체 평균을 구하라

mpg3cars <- as.data.frame(mpg %>% filter(manufacturer %in% c("chevrolet", "ford", "honda")))  
  
mean(mpg3cars$hwy)

## [1] 22.50943

#### 06-3 혼자서해보기

Q1. 이 중 일부만 추출해서 분석에 활용하려고 한다. mpg 데이터에서 class(자동차 종류), cty(도시 연비) 변수를 추출해 새로운 데이터를 만들고 출력하라.

mpgcc <- as.data.frame(mpg %>% select("class","cty"))  
head(mpgcc)

## class cty  
## 1 compact 18  
## 2 compact 21  
## 3 compact 20  
## 4 compact 21  
## 5 compact 16  
## 6 compact 18

Q2. 앞에서 추출한 데이터를 이용해서 class(자동차 종류)가 "suv"인 자동차와 "compact"인 자동차 중 어떤 자동차의 cty(도시 연비)가 더 높은지 확인하라

mpg\_suv <- as.data.frame(mpg %>% filter(class == "suv"))  
mpg\_compact <- as.data.frame(mpg %>% filter(class == "compact"))  
  
mean\_cty2 <- data.frame(data = c("mpg\_suv","mpg\_compact"),  
 Mean =c(mean(mpg\_suv$cty),mean(mpg\_compact$cty)))  
mean\_cty2

## data Mean  
## 1 mpg\_suv 13.50000  
## 2 mpg\_compact 20.12766

* 자동차의 class가 compact인 자동차가 cty가 더 높다.

#### 06-4 혼자서해보기

mpg 데이터를 이용해서 "audi"에서 생산한 자동차 중 hwy가 1~5위에 해당하는 자동차의 데이터를 출력하라.

mpg %>% as.data.frame() %>%   
 filter(manufacturer =="audi") %>%  
 arrange(desc(hwy)) %>%  
 head(5)

## manufacturer model displ year cyl trans drv cty hwy fl class  
## 1 audi a4 2.0 2008 4 manual(m6) f 20 31 p compact  
## 2 audi a4 2.0 2008 4 auto(av) f 21 30 p compact  
## 3 audi a4 1.8 1999 4 auto(l5) f 18 29 p compact  
## 4 audi a4 1.8 1999 4 manual(m5) f 21 29 p compact  
## 5 audi a4 quattro 2.0 2008 4 manual(m6) 4 20 28 p compact

#### 06-5 혼자서해보기

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

mpg\_copy2 <- mpg %>% as.data.frame() %>%  
 mutate(total = cty + hwy)

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

mpg\_copy2 <- mpg\_copy2 %>% mutate(Mean = total/2)

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

mpg\_copy2 %>% arrange(desc(Mean)) %>%  
 head(3)

## manufacturer model displ year cyl trans drv cty hwy fl  
## 1 volkswagen new beetle 1.9 1999 4 manual(m5) f 35 44 d  
## 2 volkswagen jetta 1.9 1999 4 manual(m5) f 33 44 d  
## 3 volkswagen new beetle 1.9 1999 4 auto(l4) f 29 41 d  
## class total Mean  
## 1 subcompact 79 39.5  
## 2 compact 77 38.5  
## 3 subcompact 70 35.0

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

mpg %>% as.data.frame() %>%  
 mutate(total = cty + hwy, Mean = total/2) %>%   
 arrange(desc(Mean)) %>%   
 head(3)

## manufacturer model displ year cyl trans drv cty hwy fl  
## 1 volkswagen new beetle 1.9 1999 4 manual(m5) f 35 44 d  
## 2 volkswagen jetta 1.9 1999 4 manual(m5) f 33 44 d  
## 3 volkswagen new beetle 1.9 1999 4 auto(l4) f 29 41 d  
## class total Mean  
## 1 subcompact 79 39.5  
## 2 compact 77 38.5  
## 3 subcompact 70 35.0

#### 06-6 혼자서해보기

Q1. mpg 데이터의 class는 "suv","compact"등 자동차를 특징에 따라 일곱종류로 분류한 변수이다. 어떤 차종의 연비가 높은지 비교하려고 할 때 class별 cty 평균을 구하라

mpg %>% as.data.frame() %>%  
 group\_by(class) %>%  
 summarise(cty\_mean = mean(cty))

## # A tibble: 7 x 2  
## class cty\_mean  
## <chr> <dbl>  
## 1 2seater 15.40000  
## 2 compact 20.12766  
## 3 midsize 18.75610  
## 4 minivan 15.81818  
## 5 pickup 13.00000  
## 6 subcompact 20.37143  
## 7 suv 13.50000

Q2. 어떤 차종의 도시 연비가 높은지 알아볼 수 있도록 cty 평균이 높은 순으로 정렬해 출력하라

mpg %>% as.data.frame() %>%  
 group\_by(class) %>%  
 summarise(cty\_mean = mean(cty)) %>%  
 arrange(desc(cty\_mean))

## # A tibble: 7 x 2  
## class cty\_mean  
## <chr> <dbl>  
## 1 subcompact 20.37143  
## 2 compact 20.12766  
## 3 midsize 18.75610  
## 4 minivan 15.81818  
## 5 2seater 15.40000  
## 6 suv 13.50000  
## 7 pickup 13.00000

Q3. 어떤 회사 자동차의 hwy(고속도로 연비)가 가장 높은지 알아보려한다. hwy 평균이 가장 높은 회사 세곳을 출력하라

mpg %>% as.data.frame() %>%  
 group\_by(manufacturer) %>%  
 summarise(hwy\_mean = mean(hwy)) %>%  
 arrange(desc(hwy\_mean)) %>%  
 head(3)

## # A tibble: 3 x 2  
## manufacturer hwy\_mean  
## <chr> <dbl>  
## 1 honda 32.55556  
## 2 volkswagen 29.22222  
## 3 hyundai 26.85714

Q4. 어떤 회사에서 "compact"(경차) 차종을 가장 많이 생산하는지 알아보려고 한다. 각 회사별 "compact" 차종 수를 내림차순으로 정렬해 출력하라

mpg %>% as.data.frame() %>%  
 group\_by(manufacturer) %>%  
 filter(class == "compact") %>%  
 summarise(compact\_n = n()) %>%  
 arrange(desc(compact\_n))

## # A tibble: 5 x 2  
## manufacturer compact\_n  
## <chr> <int>  
## 1 audi 15  
## 2 volkswagen 14  
## 3 toyota 12  
## 4 subaru 4  
## 5 nissan 2

#### 06-7 혼자서해보기

Q. fuel data frame

fuel <- data.frame(fl = c("c", "d", "e", "p", "r"),  
 price\_fl = c(2.35, 2.38, 2.11, 2.76, 2.22),  
 stringsAsFactors = F)  
fuel

## fl price\_fl  
## 1 c 2.35  
## 2 d 2.38  
## 3 e 2.11  
## 4 p 2.76  
## 5 r 2.22

Q1. fuel 데이터를 이용해서 mpg 데이터에 price\_fl(연료 가격) 변수를 추가하라

mpg <- as.data.frame(mpg)  
mpg <- left\_join(mpg,fuel,by = "fl")

Q2. 연료 가격 변수가 잘 추가됐는지 확인하기 위해 model, fl, price\_fl 변수를 추출해 앞부분 5행을 출력하라

mpg %>% select(model, fl, price\_fl) %>%  
 head(5)

## model fl price\_fl  
## 1 a4 p 2.76  
## 2 a4 p 2.76  
## 3 a4 p 2.76  
## 4 a4 p 2.76  
## 5 a4 p 2.76

#### 06-7 분석 도전

Q1. popadults는 해당 지역의 성인 인구, poptotal은 전체 인구를 나타낸다. midest 데이터에 전체 인구 대비 미성년 인구 백분율 변수를 추가하라.

midwest <- as.data.frame(ggplot2::midwest)  
  
midwest <- midwest %>%   
 mutate(perchild = ((poptotal - popadults) / poptotal) \* 100)

Q2. 미성년 인구 백분율이 가장 높은 상위 5개 county(지역)의 미성년 인구 백분율을 출력하라

midwest %>%   
 group\_by(county) %>%  
 arrange(desc(perchild)) %>%  
 select(county, perchild) %>%  
 head(5)

## # A tibble: 5 x 2  
## # Groups: county [5]  
## county perchild  
## <chr> <dbl>  
## 1 ISABELLA 51.50117  
## 2 MENOMINEE 50.59126  
## 3 ATHENS 49.32073  
## 4 MECOSTA 49.05918  
## 5 MONROE 47.35818

Q3. 분류표의 기준에 따라 미성년 비율 등급 변수를 추가하고 각 등급에 몇 개의 지역이 있는지 확인하라

midwest <- midwest %>%  
 mutate(childLevel = ifelse(perchild >= 40, "large",  
 ifelse(perchild >= 30, "middle",  
 "small")))  
table(midwest$child)

##   
## large middle small   
## 32 396 9

Q4. 전체 인구 대비 아시아인 인구 백분율 변수를 추가하고 하위 10개 지역의 state, county, 아시아인 인구 백분율을 출력하라

midwest %>% mutate(perAsian = (popasian/poptotal) \* 100) %>%  
 arrange(perAsian) %>%  
 select(state, county, perAsian) %>%   
 head(10)

## state county perAsian  
## 1 WI MENOMINEE 0.00000000  
## 2 IN BENTON 0.01059210  
## 3 IN CARROLL 0.01594981  
## 4 OH VINTON 0.02703190  
## 5 WI IRON 0.03250447  
## 6 IL SCOTT 0.05315379  
## 7 IN CLAY 0.06071645  
## 8 MI OSCODA 0.06375925  
## 9 OH PERRY 0.06654625  
## 10 IL PIATT 0.07074865