problm

rhadoop

2019년 3월 13일

### 문제 1. ggplot2 의 midwest 데이터를 데이터 프레임 형태로 불러와서 데이터의 특성을 파악하세요.

library(dplyr)

##   
## 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

library(ggplot2)  
  
df\_midwest = as.data.frame(ggplot2::midwest)  
head(df\_midwest)

## PID county state area poptotal popdensity popwhite popblack  
## 1 561 ADAMS IL 0.052 66090 1270.9615 63917 1702  
## 2 562 ALEXANDER IL 0.014 10626 759.0000 7054 3496  
## 3 563 BOND IL 0.022 14991 681.4091 14477 429  
## 4 564 BOONE IL 0.017 30806 1812.1176 29344 127  
## 5 565 BROWN IL 0.018 5836 324.2222 5264 547  
## 6 566 BUREAU IL 0.050 35688 713.7600 35157 50  
## popamerindian popasian popother percwhite percblack percamerindan  
## 1 98 249 124 96.71206 2.5752761 0.1482826  
## 2 19 48 9 66.38434 32.9004329 0.1788067  
## 3 35 16 34 96.57128 2.8617170 0.2334734  
## 4 46 150 1139 95.25417 0.4122574 0.1493216  
## 5 14 5 6 90.19877 9.3728581 0.2398903  
## 6 65 195 221 98.51210 0.1401031 0.1821340  
## percasian percother popadults perchsd percollege percprof  
## 1 0.37675897 0.18762294 43298 75.10740 19.63139 4.355859  
## 2 0.45172219 0.08469791 6724 59.72635 11.24331 2.870315  
## 3 0.10673071 0.22680275 9669 69.33499 17.03382 4.488572  
## 4 0.48691813 3.69733169 19272 75.47219 17.27895 4.197800  
## 5 0.08567512 0.10281014 3979 68.86152 14.47600 3.367680  
## 6 0.54640215 0.61925577 23444 76.62941 18.90462 3.275891  
## poppovertyknown percpovertyknown percbelowpoverty percchildbelowpovert  
## 1 63628 96.27478 13.151443 18.01172  
## 2 10529 99.08714 32.244278 45.82651  
## 3 14235 94.95697 12.068844 14.03606  
## 4 30337 98.47757 7.209019 11.17954  
## 5 4815 82.50514 13.520249 13.02289  
## 6 35107 98.37200 10.399635 14.15882  
## percadultpoverty percelderlypoverty inmetro category  
## 1 11.009776 12.443812 0 AAR  
## 2 27.385647 25.228976 0 LHR  
## 3 10.852090 12.697410 0 AAR  
## 4 5.536013 6.217047 1 ALU  
## 5 11.143211 19.200000 0 AAR  
## 6 8.179287 11.008586 0 AAR

tail(df\_midwest)

## PID county state area poptotal popdensity popwhite popblack  
## 432 3047 WASHINGTON WI 0.025 95328 3813.1200 94465 125  
## 433 3048 WAUKESHA WI 0.034 304715 8962.2059 298313 1096  
## 434 3049 WAUPACA WI 0.045 46104 1024.5333 45695 22  
## 435 3050 WAUSHARA WI 0.037 19385 523.9189 19094 29  
## 436 3051 WINNEBAGO WI 0.035 140320 4009.1429 136822 697  
## 437 3052 WOOD WI 0.048 73605 1533.4375 72157 90  
## popamerindian popasian popother percwhite percblack percamerindan  
## 432 208 337 193 99.09470 0.1311262 0.2181940  
## 433 672 2699 1935 97.89902 0.3596804 0.2205339  
## 434 125 92 170 99.11288 0.0477182 0.2711262  
## 435 70 43 149 98.49884 0.1496002 0.3611040  
## 436 685 1728 388 97.50713 0.4967218 0.4881699  
## 437 481 722 155 98.03274 0.1222743 0.6534882  
## percasian percother popadults perchsd percollege percprof  
## 432 0.3535163 0.2024589 59583 81.34032 23.39090 4.014568  
## 433 0.8857457 0.6350196 195837 87.98899 35.39678 7.667090  
## 434 0.1995488 0.3687316 30109 72.13790 16.54987 3.138596  
## 435 0.2218210 0.7686355 13316 70.00601 15.06458 2.620907  
## 436 1.2314709 0.2765108 88960 80.61938 24.99550 5.659847  
## 437 0.9809116 0.2105835 46796 78.29515 21.66638 4.583725  
## poppovertyknown percpovertyknown percbelowpoverty percchildbelowpovert  
## 432 94143 98.75692 3.237628 4.069854  
## 433 299802 98.38767 3.121060 3.785820  
## 434 44412 96.33004 8.488697 10.071411  
## 435 19163 98.85478 13.786985 20.050708  
## 436 133950 95.46038 8.804031 10.592031  
## 437 72685 98.75008 8.525831 11.162997  
## percadultpoverty percelderlypoverty inmetro category  
## 432 2.584500 4.280889 1 HLU  
## 433 2.590061 4.085479 1 HLU  
## 434 6.953799 10.338641 0 AAR  
## 435 11.695784 11.804558 0 AAR  
## 436 8.660587 6.661094 1 HAU  
## 437 7.375656 7.882918 0 AAR

View(df\_midwest)  
dim(df\_midwest)

## [1] 437 28

str(df\_midwest)

## 'data.frame': 437 obs. of 28 variables:  
## $ PID : int 561 562 563 564 565 566 567 568 569 570 ...  
## $ county : chr "ADAMS" "ALEXANDER" "BOND" "BOONE" ...  
## $ state : chr "IL" "IL" "IL" "IL" ...  
## $ area : num 0.052 0.014 0.022 0.017 0.018 0.05 0.017 0.027 0.024 0.058 ...  
## $ poptotal : int 66090 10626 14991 30806 5836 35688 5322 16805 13437 173025 ...  
## $ popdensity : num 1271 759 681 1812 324 ...  
## $ popwhite : int 63917 7054 14477 29344 5264 35157 5298 16519 13384 146506 ...  
## $ popblack : int 1702 3496 429 127 547 50 1 111 16 16559 ...  
## $ popamerindian : int 98 19 35 46 14 65 8 30 8 331 ...  
## $ popasian : int 249 48 16 150 5 195 15 61 23 8033 ...  
## $ popother : int 124 9 34 1139 6 221 0 84 6 1596 ...  
## $ percwhite : num 96.7 66.4 96.6 95.3 90.2 ...  
## $ percblack : num 2.575 32.9 2.862 0.412 9.373 ...  
## $ percamerindan : num 0.148 0.179 0.233 0.149 0.24 ...  
## $ percasian : num 0.3768 0.4517 0.1067 0.4869 0.0857 ...  
## $ percother : num 0.1876 0.0847 0.2268 3.6973 0.1028 ...  
## $ popadults : int 43298 6724 9669 19272 3979 23444 3583 11323 8825 95971 ...  
## $ perchsd : num 75.1 59.7 69.3 75.5 68.9 ...  
## $ percollege : num 19.6 11.2 17 17.3 14.5 ...  
## $ percprof : num 4.36 2.87 4.49 4.2 3.37 ...  
## $ poppovertyknown : int 63628 10529 14235 30337 4815 35107 5241 16455 13081 154934 ...  
## $ percpovertyknown : num 96.3 99.1 95 98.5 82.5 ...  
## $ percbelowpoverty : num 13.15 32.24 12.07 7.21 13.52 ...  
## $ percchildbelowpovert: num 18 45.8 14 11.2 13 ...  
## $ percadultpoverty : num 11.01 27.39 10.85 5.54 11.14 ...  
## $ percelderlypoverty : num 12.44 25.23 12.7 6.22 19.2 ...  
## $ inmetro : int 0 0 0 1 0 0 0 0 0 1 ...  
## $ category : chr "AAR" "LHR" "AAR" "ALU" ...

summary(df\_midwest)

## PID county state area   
## Min. : 561 Length:437 Length:437 Min. :0.00500   
## 1st Qu.: 670 Class :character Class :character 1st Qu.:0.02400   
## Median :1221 Mode :character Mode :character Median :0.03000   
## Mean :1437 Mean :0.03317   
## 3rd Qu.:2059 3rd Qu.:0.03800   
## Max. :3052 Max. :0.11000   
## poptotal popdensity popwhite popblack   
## Min. : 1701 Min. : 85.05 Min. : 416 Min. : 0   
## 1st Qu.: 18840 1st Qu.: 622.41 1st Qu.: 18630 1st Qu.: 29   
## Median : 35324 Median : 1156.21 Median : 34471 Median : 201   
## Mean : 96130 Mean : 3097.74 Mean : 81840 Mean : 11024   
## 3rd Qu.: 75651 3rd Qu.: 2330.00 3rd Qu.: 72968 3rd Qu.: 1291   
## Max. :5105067 Max. :88018.40 Max. :3204947 Max. :1317147   
## popamerindian popasian popother percwhite   
## Min. : 4.0 Min. : 0 Min. : 0 Min. :10.69   
## 1st Qu.: 44.0 1st Qu.: 35 1st Qu.: 20 1st Qu.:94.89   
## Median : 94.0 Median : 102 Median : 66 Median :98.03   
## Mean : 343.1 Mean : 1310 Mean : 1613 Mean :95.56   
## 3rd Qu.: 288.0 3rd Qu.: 401 3rd Qu.: 345 3rd Qu.:99.07   
## Max. :10289.0 Max. :188565 Max. :384119 Max. :99.82   
## percblack percamerindan percasian percother   
## Min. : 0.0000 Min. : 0.05623 Min. :0.0000 Min. :0.00000   
## 1st Qu.: 0.1157 1st Qu.: 0.15793 1st Qu.:0.1737 1st Qu.:0.09102   
## Median : 0.5390 Median : 0.21502 Median :0.2972 Median :0.17844   
## Mean : 2.6763 Mean : 0.79894 Mean :0.4872 Mean :0.47906   
## 3rd Qu.: 2.6014 3rd Qu.: 0.38362 3rd Qu.:0.5212 3rd Qu.:0.48050   
## Max. :40.2100 Max. :89.17738 Max. :5.0705 Max. :7.52427   
## popadults perchsd percollege percprof   
## Min. : 1287 Min. :46.91 Min. : 7.336 Min. : 0.5203   
## 1st Qu.: 12271 1st Qu.:71.33 1st Qu.:14.114 1st Qu.: 2.9980   
## Median : 22188 Median :74.25 Median :16.798 Median : 3.8142   
## Mean : 60973 Mean :73.97 Mean :18.273 Mean : 4.4473   
## 3rd Qu.: 47541 3rd Qu.:77.20 3rd Qu.:20.550 3rd Qu.: 4.9493   
## Max. :3291995 Max. :88.90 Max. :48.079 Max. :20.7913   
## poppovertyknown percpovertyknown percbelowpoverty percchildbelowpovert  
## Min. : 1696 Min. :80.90 Min. : 2.180 Min. : 1.919   
## 1st Qu.: 18364 1st Qu.:96.89 1st Qu.: 9.199 1st Qu.:11.624   
## Median : 33788 Median :98.17 Median :11.822 Median :15.270   
## Mean : 93642 Mean :97.11 Mean :12.511 Mean :16.447   
## 3rd Qu.: 72840 3rd Qu.:98.60 3rd Qu.:15.133 3rd Qu.:20.352   
## Max. :5023523 Max. :99.86 Max. :48.691 Max. :64.308   
## percadultpoverty percelderlypoverty inmetro category   
## Min. : 1.938 Min. : 3.547 Min. :0.0000 Length:437   
## 1st Qu.: 7.668 1st Qu.: 8.912 1st Qu.:0.0000 Class :character   
## Median :10.008 Median :10.869 Median :0.0000 Mode :character   
## Mean :10.919 Mean :11.389 Mean :0.3432   
## 3rd Qu.:13.182 3rd Qu.:13.412 3rd Qu.:1.0000   
## Max. :43.312 Max. :31.162 Max. :1.0000

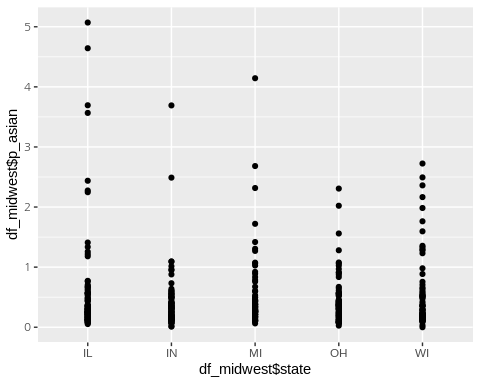
### 문제 2. poptotal(전체 인구)을 total 로, popasian(아시아 인구)을 asian 으로 변수명을 수정하세요.

df\_midwest <- rename(df\_midwest, total = poptotal, asian=popasian)  
head(df\_midwest)

## PID county state area total popdensity popwhite popblack  
## 1 561 ADAMS IL 0.052 66090 1270.9615 63917 1702  
## 2 562 ALEXANDER IL 0.014 10626 759.0000 7054 3496  
## 3 563 BOND IL 0.022 14991 681.4091 14477 429  
## 4 564 BOONE IL 0.017 30806 1812.1176 29344 127  
## 5 565 BROWN IL 0.018 5836 324.2222 5264 547  
## 6 566 BUREAU IL 0.050 35688 713.7600 35157 50  
## popamerindian asian popother percwhite percblack percamerindan  
## 1 98 249 124 96.71206 2.5752761 0.1482826  
## 2 19 48 9 66.38434 32.9004329 0.1788067  
## 3 35 16 34 96.57128 2.8617170 0.2334734  
## 4 46 150 1139 95.25417 0.4122574 0.1493216  
## 5 14 5 6 90.19877 9.3728581 0.2398903  
## 6 65 195 221 98.51210 0.1401031 0.1821340  
## percasian percother popadults perchsd percollege percprof  
## 1 0.37675897 0.18762294 43298 75.10740 19.63139 4.355859  
## 2 0.45172219 0.08469791 6724 59.72635 11.24331 2.870315  
## 3 0.10673071 0.22680275 9669 69.33499 17.03382 4.488572  
## 4 0.48691813 3.69733169 19272 75.47219 17.27895 4.197800  
## 5 0.08567512 0.10281014 3979 68.86152 14.47600 3.367680  
## 6 0.54640215 0.61925577 23444 76.62941 18.90462 3.275891  
## poppovertyknown percpovertyknown percbelowpoverty percchildbelowpovert  
## 1 63628 96.27478 13.151443 18.01172  
## 2 10529 99.08714 32.244278 45.82651  
## 3 14235 94.95697 12.068844 14.03606  
## 4 30337 98.47757 7.209019 11.17954  
## 5 4815 82.50514 13.520249 13.02289  
## 6 35107 98.37200 10.399635 14.15882  
## percadultpoverty percelderlypoverty inmetro category  
## 1 11.009776 12.443812 0 AAR  
## 2 27.385647 25.228976 0 LHR  
## 3 10.852090 12.697410 0 AAR  
## 4 5.536013 6.217047 1 ALU  
## 5 11.143211 19.200000 0 AAR  
## 6 8.179287 11.008586 0 AAR

### 문제 3. total, asian 변수를 이용해 ‘전체 인구 대비 아시아 인구 백분율’ 파생변수를 만들고, 히스토그램을 만들어 도시들이 어떻게 분포하는지 살펴보세요.

df\_midwest$p\_asian = (df\_midwest$asian /df\_midwest$total) \*100  
qplot(df\_midwest$state,df\_midwest$p\_asian, data=df\_midwest, geom="auto")

 ### 문제 4. 아시아 인구 백분율 전체 평균을 구하고, 평균을 초과하면 “large”, 그 외에는 “small”을 부여하는 파생변수를 만들어 보세요.

avg=mean(df\_midwest$p\_asian)  
df\_midwest$grade = ifelse(df\_midwest$p\_asian>=avg,"large","small")

### 문제 5. “large”와 “small”에 해당하는 지역이 얼마나 되는지, 빈도표와 빈도 막대 그래프를 만들어 확인해 보세요

table(df\_midwest$grade)

##   
## large small   
## 119 318

qplot(df\_midwest$grade)

