최종작성:

2021. 11. 04(목)

사례연구2

: 세계 인구에 대한 각종 기술통계 분석

B2: 문현진(조장) 남원식 오준서

제출일자:

2021. 11. 08(월)

목 차

서론

월드 뱅크(The World Bank)의 오픈소스 데이터를 통해, R Language를 활용하여 나라별 총 인구수를 기준으로 나라별 가장 큰 도시의 인구 비율을 구하는 방법을 기술한 보고서이다.

본론

환경설정 및 데이터전처리

월드 뱅크 홈페이지(https://www.worldbank.org/en/home)에서 각나라별 총 인구수와 가장 큰 도시의 인구 중 2019년의 데이터만 추출하여 .csv파일로 저장한다.

```
2  # 환경설정
3  install.packages("dplyr")
4  library(dplyr)
5
6  rm(list = ls())
7  getwd()
8  setwd('C:/워킹디렉토리 설정')
9  data <- read.csv("subPopCountry.csv")
10 data2 <- read.csv('subPopCity.csv')
11 country <- data[,c(3,5)] # 전세계 국가와 인구수 데이터
12 city <- data2[,c(3,5)] # 나라별 가장 큰 도시의 인구수 데이터
13
```

data 변수에 받아온 파일들을 불러온다.

각 변수에 불러온 파일에서 나라이름과 인구수 데이터를 추출한다.

(Country: 국가명, 국가 인구수 자료 / City: 국가명, 대도시 인구수 자료)

Country 자료내용

Series Name	Series Code	Country Name	Country Code	2019 [YR2019]
Population, total	SP.POP.TOTL	Afghanistan	AFG	38041757
Population, total	SP.POP.TOTL	Albania	ALB	2854191
Population, total	SP.POP.TOTL	Algeria	DZA	43053054
Population, total	SP.POP.TOTL	American Samoa	ASM	55312
Population, total	SP.POP.TOTL	Andorra	AND	77146
Population, total	SP.POP.TOTL	Angola	AGO	31825299
Population, total	SP.POP.TOTL	Antigua and Barbuda	ATG	97115
Population, total	SP.POP.TOTL	Argentina	ARG	44938712

…(생략)

City 자료내용

Series Name	Series Code	Country Name	Country Code	2019 [YR2019]
Population in largest city	EN.URB.LCTY	Afghanistan	AFG	4114030
Population in largest city	EN.URB.LCTY	Albania	ALB	484624
Population in largest city	EN.URB.LCTY	Algeria	DZA	2729325
Population in largest city	EN.URB.LCTY	American Samoa	ASM	..
Population in largest city	EN.URB.LCTY	Andorra	AND	..
Population in largest city	EN.URB.LCTY	Angola	AGO	8044735
Population in largest city	EN.URB.LCTY	Antigua and Barbuda	ATG	..
Population in largest city	EN.URB.LCTY	Argentina	ARG	15057273

…(생략)

1) 나라별 총 인구수 대비 가장 큰 도시인구의 비율

- 코드

```
15
16 # 1)각 나라에서 가장 큰 도시에 사는 인구의 비율
17 country_name <-list()
18 pop_ratio <-list()
19 for (i in 1:length(country$Country.Name)) {
20   if(country[i,2] > city[i,2]){
21
22     # 각 나라별 인구비율 계산
23     calc <- as.numeric(city[i,2]) / as.numeric(country[i,2]) * 100
24
25     # 각 변수에 나라이름과 계산된 인구비율을 저장
26     country_name <- append(country[i,1], country_name)
27     pop_ratio <- append(calc, pop_ratio)
28
29   }
30 }
31 # 인구비율 계산 결과 도출
32 new_country <- unlist(country_name)
33 new_pop_ratio <- unlist(pop_ratio)
34 result <- data.frame(new_country, new_pop_ratio)
35 result
36
```

(1) 추출된 자료에서 반복문을 이용하여 각 나라별로 총 인구수와 가장 큰 도시인구의 비율을 구한다.

(2) 인구비율이 계산된 나라는 country_name 라는 변수에 저장된다.

(3) 각 나라의 인구비율은 pop_ratio 라는 변수에 저장된다.

(4) 각 변수안에 문자열로 저장되어 있는 자료들을 리스트 해제한다.

(5) 두 변수를 합하여 데이터프레임을 생성하고 결과를 출력한다.

- 결과

result

Filter

	new_country	new_pop_ratio
1	Yemen, Rep.	9.856597
2	Virgin Islands (U.S.)	NA
3	Vietnam	8.678463
4	Vanuatu	NA
5	Uzbekistan	7.416045
6	Uruguay	50.399641
7	United States	5.727385
8	United Arab Emirates	28.996177
9	Ukraine	6.698782
10	Uganda	7.087688
11	Tuvalu	NA
12	Turks and Caicos Islands	NA
13	Turkey	17.940474
14	Tonga	NA
15	Togo	22.088972
16	Timor-Leste	NA
17	Thailand	14.865519
18	Tajikistan	9.589355
19	Switzerland	16.128826
20	Suriname	NA
21	St. Vincent and the Grenadines	NA

(중략)

115	Channel Islands	NA
116	Chad	8.600578
117	Cayman Islands	NA
118	Cabo Verde	NA
119	Bulgaria	18.305343
120	Brunei Darussalam	NA
121	British Virgin Islands	NA
122	Botswana	NA
123	Bhutan	NA
124	Bermuda	NA
125	Belize	NA
126	Belarus	21.413350
127	Barbados	NA
128	Bahamas, The	NA
129	Austria	21.569316
130	Aruba	NA
131	Armenia	36.626864
132	Argentina	33.506241
133	Antigua and Barbuda	NA
134	Andorra	NA
135	American Samoa	NA
136	Algeria	6.339446

2) 가장 큰 도시의 인구비율이 가장 높은 20개국

- 코드

```
38
39 # 2)인구 비율이 가장 높은 나라 20개국
40 question2 <- result[order(-result$new_pop_ratio),]
41 answer2 <- head(question2,20)
42 answer2
43
```

(1) 앞서 생성한 데이터프레임(result)에서 인구비율을 내림차순으로 question2 변수에 저장한다.

(2) head함수로 question2의 자료 중 20개를 보여준다.

- 결과

answer2

Filter

	new_country	new_pop_ratio
89	Hong Kong SAR, China	99.77857
39	Puerto Rico	76.75673
78	Kuwait	72.55615
106	Djibouti	58.42534
6	Uruguay	50.39964
61	Mongolia	48.14183
42	Paraguay	46.54830
83	Israel	45.25031
44	Panama	42.89930
111	Congo, Rep.	42.89702
131	Armenia	36.62686
75	Lebanon	35.10760
132	Argentina	33.50624
41	Peru	32.46559
53	New Zealand	31.77210
47	Oman	30.18367
73	Liberia	29.71268
8	United Arab Emirates	28.99618
97	Georgia	28.95966
65	Mauritania	27.81832

3) 가장 큰 도시의 인구비율이 가장 낮은 20개국

- 코드

```
45
46 # 3)인구 비율이 가장 낮은 나라 20개국
47 question3 <- result[order(result$new_pop_ratio),]
48 answer3 <- head(question3,20)
49 answer3
50
```

(1) 앞서 생성한 데이터프레임(result)에서 인구비율을 오름차순으로 question3 변수에 저장한다.

(2) head함수로 question3의 자료 중 20개를 보여준다.

- 결과

answer3

Filter

	new_country	new_pop_ratio
86	Indonesia	3.931147
43	Papua New Guinea	4.267456
96	Germany	4.280497
40	Poland	4.677758
56	Nepal	4.810101
51	Niger	5.368822
59	Mozambique	5.495415
7	United States	5.727385
69	Malawi	5.769013
136	Algeria	6.339446
55	Netherlands	6.574501
9	Ukraine	6.698782
50	Nigeria	6.918477
10	Uganda	7.087688
82	Italy	7.088706
46	Pakistan	7.268664
5	Uzbekistan	7.416045
28	Slovak Republic	7.929893
116	Chad	8.600578
37	Russian Federation	8.639633

4) 2), 3) 항목에 데이터가 없는 나라의 리스트

- 코드

```
52
53  # 4)결측치 확인
54  answer4 <- result %>% filter(is.na(result$new_pop_ratio))
55  answer4
56
```

(1) result 변수 안에 인구비율이 저장된 컬럼에서 filter()함수를 이용하여 NA값이 저장된 자료들을 answer4 변수에 저장한다.

(2) answer4 변수를 출력하여 결과를 확인한다.

- 결과

answer4

Filter

	new_country	new_pop_ratio
1	Virgin Islands (U.S.)	NA
2	Vanuatu	NA
3	Tuvalu	NA
4	Turks and Caicos Islands	NA
5	Tonga	NA
6	Timor-Leste	NA
7	Suriname	NA
8	St. Vincent and the Grenadines	NA
9	St. Martin (French part)	NA
10	St. Lucia	NA
11	St. Kitts and Nevis	NA
12	Solomon Islands	NA
13	Slovenia	NA
14	Sint Maarten (Dutch part)	NA
15	Seychelles	NA
16	Sao Tome and Principe	NA
17	San Marino	NA
18	Samoa	NA
19	Palau	NA
20	Northern Mariana Islands	NA

(중략)

43	Fiji	NA
44	Faroe Islands	NA
45	Eswatini	NA
46	Dominica	NA
47	Cyprus	NA
48	Curacao	NA
49	Comoros	NA
50	Channel Islands	NA
51	Cayman Islands	NA
52	Cabo Verde	NA
53	Brunei Darussalam	NA
54	British Virgin Islands	NA
55	Botswana	NA
56	Bhutan	NA
57	Bermuda	NA
58	Belize	NA
59	Barbados	NA
60	Bahamas, The	NA
61	Aruba	NA
62	Antigua and Barbuda	NA
63	Andorra	NA
64	American Samoa	NA

#전체코드

study2_1.5.0(submit).r

Source on Save

```r
# 사례연구2 B2조 문현진(조장) 남원식 오준서
# 환경설정
install.packages("dplyr")
library(dplyr)

rm(list = ls())
getwd()
setwd('C:/워킹디렉토리 설정')
data <- read.csv("subPopCountry.csv")
data2 <- read.csv('subPopCity.csv')
country <- data[,c(3,5)] # 전세계 국가와 인구수 데이터
city <- data2[,c(3,5)] # 나라별 가장 큰 도시의 인구수 데이터



# 1)각 나라에서 가장 큰 도시에 사는 인구의 비율
country_name <-list()
pop_ratio <-list()
for (i in 1:length(country$Country.Name)) {
  if(country[i,2] > city[i,2]){

    # 각 나라별 인구비율 계산
    calc <- as.numeric(city[i,2]) / as.numeric(country[i,2]) * 100

    # 각 변수에 나라이름과 계산된 인구비율을 저장
    country_name <- append(country[i,1], country_name)
    pop_ratio <- append(calc, pop_ratio)

  }
}
# 인구비율 계산 결과 도출
new_country <- unlist(country_name)
new_pop_ratio <- unlist(pop_ratio)
result <- data.frame(new_country, new_pop_ratio)
result



# 2)인구 비율이 가장 높은 나라 20개국
question2 <- result[order(-result$new_pop_ratio),]
answer2 <- head(question2,20)
answer2



# 3)인구 비율이 가장 낮은 나라 20개국
question3 <- result[order(result$new_pop_ratio),]
answer3 <- head(question3,20)
answer3



# 4)결측치 확인
answer4 <- result %>% filter(is.na(result$new_pop_ratio))
answer4

```

결론

월드 뱅크(The world bank)의 인구데이터를 R language를 통해서 각 나라별로 가장 큰 도시의 인구비율을 알아보았다.

정형 데이터를 추출하여 계산하고 병합하는 과정에서 컬럼명 encoding 오류가 발생하고, 잘못된 이중 for문의 사용으로 프로그램의 효율성이 지나치게 떨어지는 등의 문제가 발생하였으나 웹 서칭 및 상호 간의 코드 피드백으로 해결하였다.

본 프로젝트를 통하여 .csv파일을 한층 더 능숙하게 핸들링 할 수 있게 되었다.

참고자료

https://www.worldbank.org/en/home
subPopCity.csv
subPopCountry.csv