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** ………………………….………….………………………………………………… 3

**본론** ………………………...…..………………………………………………………… 4

#데이터전처리 및 환경설정 4

1. 나라별 총 인구수 대비 가장 큰 도시인구의 비율 5
2. 가장 큰 도시의 인구비율이 가장 높은 20개국 7
3. 가장 큰 도시의 인구비율이 가장 낮은 20개국 8
4. 2), 3) 항목에 데이터가 없는 나라의 리스트 9

**결론** ………..……………………………………………………………….…………… 11

*참고자료* ………………………………...…………………………….……………… 11

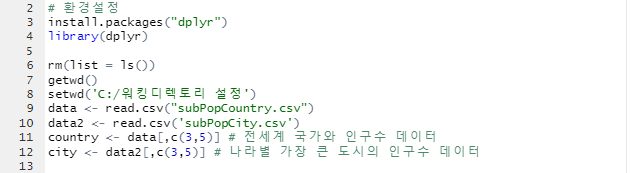
**서론**

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.



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)

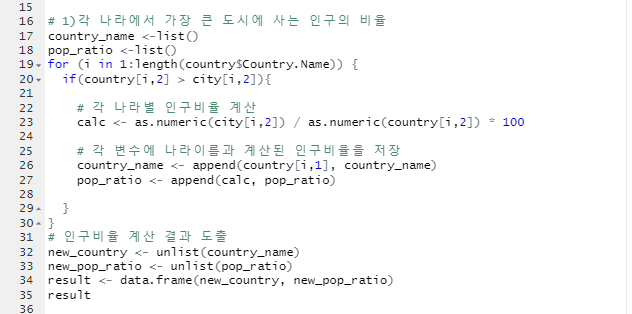
City 자료내용

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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)

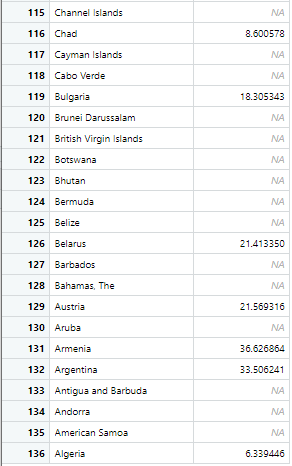
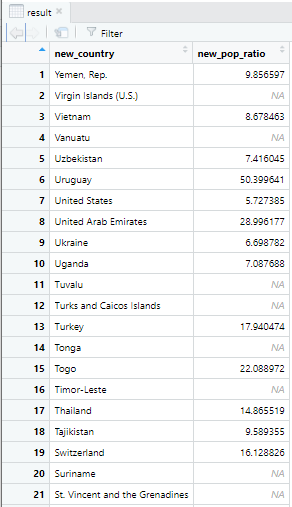
1. **나라별 총 인구수 대비 가장 큰 도시인구의 비율**

* 코드



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.

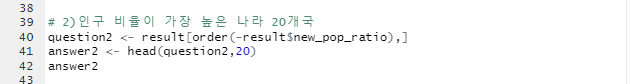
* 결과



(중략)

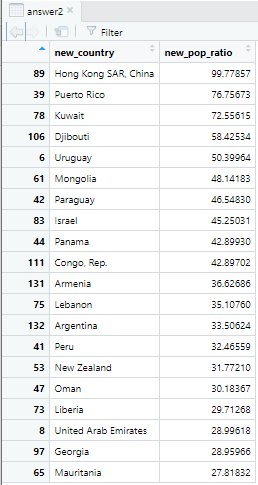
1. **가장 큰 도시의 인구비율이 가장 높은 20개국**

* 코드



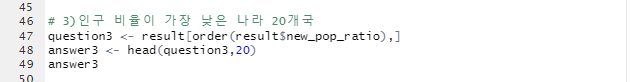
1. 앞서 생성한 데이터프레임(result)에서 인구비율을 내림차순으로 question2 변수에 저장한다.
2. head함수로 question2의 자료 중 20개를 보여준다.

* 결과



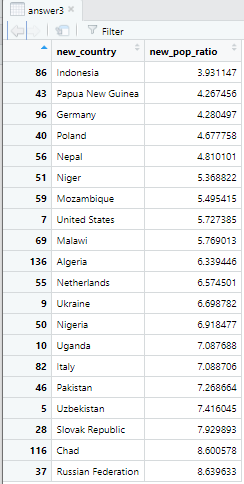
1. **가장 큰 도시의 인구비율이 가장 낮은 20개국**

* 코드



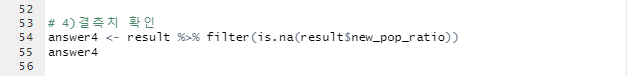
1. 앞서 생성한 데이터프레임(result)에서 인구비율을 오름차순으로 question3 변수에 저장한다.
2. head함수로 question3의 자료 중 20개를 보여준다.

* 결과



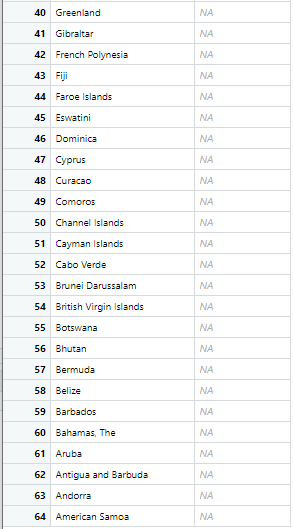
1. **2), 3) 항목에 데이터가 없는 나라의 리스트**

* 코드

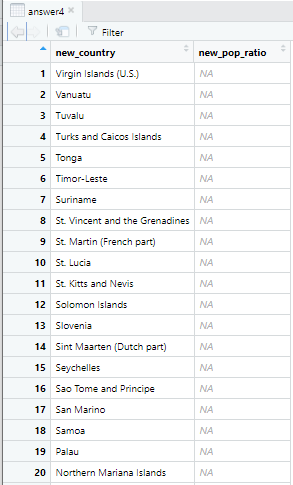


1. result 변수 안에 인구비율이 저장된 컬럼에서 filter()함수를 이용하여 NA값이 저장된 자료들을 answer4 변수에 저장한다.
2. answer4 변수를 출력하여 결과를 확인한다.

(중략)



* 결과



***#전체코드***



**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