쌩 초짜를 위한 R

정 태 훈

2014년 6월 9일

- 이제 RStudio를 쓰는 기본적인 방법이나 이런 건 익숙해지셨을테니 (맞나요?) 그래픽으로 어떻게 하는지 이런 부분은 가급적 줄이고 어떤 걸 어떻게 하는지 하는 부분을 좀 더 많이 설명하도록 하겠습니다. 그래도 되겠죠?
- 근데 제가 저번에 올려 놓은 데이터를 보시면서 뭔가 허전함을 못 느끼셨나요? 그러니까 제가 국가별 건강체계, 위험인구, 과학기술투자 뭐 이런 걸 데이터로 준비해 드렸는데 달랑 이것들만 가지고 뭘 본다고 하면 결국 나오는 결론이라는 게 뭘까요? 아무리 생각해도 “아, 이 나라랑 저 나라는 이런 저런 게 서로 다르구나” 정도 밖에는 없을 겁니다, 그죠? 그런데 고작 (누구나 해보지 않고도 아는) “나라들끼리 서로 다르다”는 걸 알자고 이런 걸 하는 건 아닐 거잖아요. 하다 못 해 동네 애들끼리 키를 비교할 때도 나이가 비슷한 애들끼리 비교해서 누구랑 누구는 나이는 같은데 누가 누구보다 크네 작네 이러잖아요. 우린 모두 뭔가 이렇게 서로 같은 어떤 기준에 두고서 봐야 말에 설득력이 있다는 걸 아는데 하물며 나라끼리 비교할 때도 객관적으로 기준이 될 뭔가가 필요하지 않겠어요? 그래야 “이런 이런 게 비슷한 나라들끼리 비교해 보니 이 나라는 저 나라보다 크네, 작네” 이렇게 될 테니까요. 근데 그런 기준이 될 만한 게 없었죠. ㅎㅎ
- 그러니까 “뭔가 데이터가 더 필요하다”는 얘긴데 .. 근데 어떤 데이터가 있으면 될까요? 보통 이럴 땐 뭘 쓰나요? 뭐 사람마다 다를 수도 있겠지만 보통 일인당 GDP를 쓰지 않나요, 그죠? 그래서 GDP 데이터를 올려 놓은 겁니다.

단계 0. 데이터 분석의 목적

- 자 그럼 이제 이들 파일을 일단 읽어 들여 데이터를 분석해 봅시다. 그런데 이것도 뭘 할 건지 목적이 있어야 어떻게 할 건지 구체적인 계획이 설 거 아니겠어요? 그래서 이번엔 이런 걸 한 번 해볼까 합니다.

  도대체 우리나라의 건강관리 관련 전체 지출에서 공공부문이 차지하는 비율은 전세계의 다른 나라와 비교해서 혹은 비슷한 경제 수준의 다른 나라와 비교해서 어떻게 될까요?

  음 .. 눈치 채셨겠지만 이런 걸 하자고 제가 세계은행의 그 수 많은 데이터 가운데 저런 데이터만 뽑은 거겠죠? 그러니까 사실 데이터를 모은다는 건 원래부터 거기에 어떤 목적을 담고 있는 겁니다. ^^ 그런데 사실 여기서 사용한 “비슷한 경제 수준의 다른 나라”라는 표현은 참 애매한 겁니다. 그래서 좀더 구체적으로 일 인 당 GDP가 일정 수준에 다다른 OECD의 다른 나라와 비교해서 어떤지를 비교해 보기로 하겠습니다.

단계 1. 데이터 읽어 들이기.

- 저걸 하려면 어떻게 해야 할까요? 우선 어느 파일에 든 어느 데이터가 저 정보를 줄 수 있는지 알아야겠죠?
- 그럼 데이터 파일을 죽 한 번들 다 둘러보시고 찾아보세요. 혹시 전체 의료비 지출 가운데 공공부문과 개인이 부담하는 비율에 대한 데이터 열 발견하셨나요? 음 .. 이게 각 파일 데이터 열 이름을 저는 알 수 있게 줄여놔서 한 번에 확 눈에 들어오지는 않을 수도 있는데 아무튼 HealthSystem.txt 파일에 있는 두 데이터 열 PublicHealthExpensePercTotal, OutOfPocketPercTotal 이 녀석들이 바로 전체 의료비 지출 가운데 공공부문이 부담하는 퍼센트와 개인이 자기 주머니에서 부담하는 퍼센트입니다. 그러니까 우선 HealthSystem.txt 이 파일을 읽어 들여야 하는군요.
- 그럼 일 인 당 GDP는 어디 있죠? GDP.txt 파일에 GDPperCapita2012 데이터 열 찾으셨나요? 바로 이겁니다. 왜냐 하면 HealthSystem.txt 파일의 데이터 열은 대부분 2012년 자료이기 때문에 2012년의 일 인 당 GDP를 써야 했던 겁니다. 그러니까 GDP.txt 이 파일도 읽어 들여야 하는 거구요.
- 그러니까 우선 이 두 파일을 읽어 들입시다!

- 그런데 여기서 한 가지 간과하기 쉬운 게 있습니다. 지금 계획대로 하자면 우리는 두 개의 데이터 파일을 읽어 들일 겁니다. 이렇게 읽어 들인 두 데이터를 각각 `HealthSys`와 `GDP`라고 부른다고 칩시다. 결국 일 인 당 GDP에 따른 의료비 지출 문제니까 이 두 데이터를 합쳐야 합니다, 그죠? 그럼 과연 이 데이터를 어떻게 합치는 게 맞을까요?

- 이게 무슨 말인지 이해하기 쉽도록 이렇게 한 번 볼까요? `HealthSys`와 `GDP`는 각각 214개 국가의 건강관리체계와 국내 총생산(GDP)에 대한 정보가 들어 있습니다. 과연 두 데이터는 동일한 국가들에 대한 정보일까요? `HealthSys`와 `GDP`에 들어 있는 국가가 다르다면 어떻게 하죠? 이걸 어떻게 확인하죠?

- 그런데 뭐 어차피 같은 시스템에서 제공한 정보니까 양쪽 다 똑같은 국가 데이터라 칩시다. 그럼 과연 이들 국가 정보가 두 파일에 동일한 순서로 들어 있을까요? (이게 뭔 뚱딴지같은 소리냐 싶으십니까? 그럼 GDPTrend.txt 파일을 한 번 열어보세요. 이게 같은 세계은행 자료인데 희한하게 국가 순서가 다르더군요. 하긴 뭐 자료를 정렬하는 방식은 워낙 많으니까요. ^^) 만약 똑같은 국가들에 대한 정보가 똑같은 순서로 들어 있다면 그냥 있는 그대로 나란히 이어 붙이면 되겠지만 (사실 지금 데이터는 제가 다 순서를 맞춰놨다는 거! ㅋㅋ) 만약 그렇지 않다면 어떻게 해야 할까요? (이건 한 번 생각해 보세요!)

- 우선 HealthSys과 GDP는 이제 다들 아시겠지만 이렇게 불러 들입니다. (아래 각 줄의 맨 앞 “>”는 프롬프트입니다. 그러니까 R에서 사용자의 입력을 기다릴 때 그냥 늘 있는 그런 거라서 여러분이 직접 입력하실 필요가 없다는 말입니다.)

```
> HealthSys <- read.table(<…파일…>, header=T, sep=“\t”, quote=“”, as.is=T, check.names=F, comment.char=“#”)
> GDP <- read.table(<…파일…>, header=T, sep=“\t”, quote=“”, as.is=T, check.names=F, comment.char=“#”)
```

여기 T는 TRUE, F는 FALSE를 줄인 겁니다!

- 데이터가 궁금하시면 이걸 쓰시면 되죠? (이건 RStudio에서만 됩니다!)

```
> View(HealthSys)
> View(GDP)
```

- 데이터가 몇 개의 열(이걸 앞으로 **변수**라고 부르겠습니다)과 몇 개의 행(이걸 앞으로 **항목**이라고 부르겠습니다)으로 이뤄졌는지 궁금하시면 이렇게 하시면 됩니다. (밑에 dim은 dimension의 줄임 말입니다.)

```
> dim(HealthSys)
[1] 214  15
> dim(GDP)
[1] 214   8
```

15개 변수에 각각 214개 항목 값이 들었다는 소립니다!

8개 변수에 각각 214개 항목 값이 들었다는 소립니다!

- 데이터 내 변수의 이름 (그러니까 열의 이름)은 어떻게 하면 알 수 있을까요?

```
> names(HealthSys)
 [1] "Country"                          "CountryCode"
 [3] "Region"                           "OECD"
 [5] "TotalHealthExpensePercGDP"        "PublicHealthExpensePercTotal"
 [7] "OutOfPocketPercTotal"
"HealthExpenseExternalResources"
 …
> colnames(GDP)
[1] "Country"                  "CountryCode"
[3] "Region"                   "OECD"
[5] "GDPperCapita2012"         "GDP2012"
[7] "GDPperCapitaPPP2011"      "GDPperCapitaPPPCurrent"
```

(데이터프레임의 경우 `names(…)`와 `colnames(…)`는 동일한 결과를 줍니다!)

- 이제 데이터 각 변수를 요약해 보고 싶다면 이렇게 하면 되겠죠?

```
> summary(HealthSys)
> summary(GDP)
```

각각 전체 변수(`HealthSys`는 15개, `GDP`는 8개)에 대한 요약 결과를 열을 맞춰 아주 줄줄 출력할 겁니다.

단계 2. 데이터의 변수.

- 그런데 우린 매번 데이터 전체를 가지고 일하지는 않습니다. 데이터의 특정 변수를 살펴보고 그 가운데 특정한 데이터 항목을 들여다 보기도 합니다. 예를 들어 “2012년 일 인당 GDP 변수”를 들여다 봐야 할 때도 있습니다. 이런 건 어떻게 할 수 있을까요?
- 우선 변수에 접근해 봅시다. 예를 들어 `HealthSys`의 `PublicHealthExpensePercTotal`, `OutOfPocketPercTotal` 이런 변수는 어떻게 다룰 수 있을까요? 제가 잠깐 세어 봤더니 R에는 여섯 가지 정도 방법이 있던데 그 가운데 일단 연관된 세 가지만 보겠습니다.

```
> HealthSys[[6]]
> HealthSys[[“PublicHealthExpensePercTotal”]]
> HealthSys$PublicHealthExpensePercTotal
```

- 왜 이렇게 많은가 싶으시죠? 이게 다 요긴한 상황이 달라서 그런데 예를 들어 우리가 원하는 변수가 몇 번 째 것인지를 아는 경우라면 첫번째 방법이 딱 입니다. 그런데 데이터에 변수가 수십 개 된다면 이 방법은 아주 불편하고 오히려 헷갈립니다. 이럴 땐 변수 이름이 더 찾기 쉽죠. 문제는 변수 이름에 “`Public Health Expense (% Total)`”처럼 특수한 문자나 빈칸 이런 게 뒤섞여 있는 경우인데 이럴 땐 볼 것 없이 두번째 방법입니다. 따옴표로 묶여 있어 통째로 이름이 되니까요. 하지만 지금 같은 상황이면 세번째 방법이 제일 낫습니다.(무엇보다 입력해야 할 문자 수가 달라집니다. “`$`” 하나면, 끝! ^^)

- 그런데 이게 다가 아닙니다. 사실 세번째 방식은 보통은 표준적인 방식처럼 여겨지는데 이 방식을 쓰면 **RStudio가 제공하는 도움 기능**을 쓸 수가 있습니다! 무슨 말인가 하면 변수가 많아 일일이 다 외우기도 어렵고 변수 이름도 너무 길어 헷갈리고 귀찮을 땐 "`<데이터 이름>$`"까지만 쓰고 바로 `탭` 키를 누르는 겁니다. 예를 들어 `HealthSys`의 `PublicHealthExpensePercTotal` 변수를 이용하려 한다면 "`HealthSys$`"까지만 쓰고 `탭` 키를 눌러보는 겁니다. 무슨 일이 생깁니까? 글 상자가 생기면서 관련된 변수 목록이 죽 뜨죠? 바로 그 가운데서 고르면 되는 겁니다! 만세! (이건 RStudio가 제공하는 도움 기능입니다. 프로그램을 해본 분들은 많은 파일 에디터에 이와 비슷한 도움말 기능이 있는 걸 아실 겁니다.)

- 이쯤 되면 왜 "`$`" 표시가 표준적인 방식으로 대접 받는지 아실 겁니다. 그리고 바로 이래서 데이터 변수 이름이 다들 저렇게 우스꽝스럽게 긴 겁니다. 변수 이름에 빈칸이나 이상한 글자가 들어가면 "`$`"로 지정하지 못 합니다. 그러면 불편해집니다. 그러니까 변수 이름에 빈칸이나 특수 문자는 쓰지 않습니다. 그러면서도 변수의 의미를 전달할 수 있어야 합니다. 그래서 관련된 단어를 저렇게 그냥 죽 이어 쓰는 겁니다. ^^

- 그럼 이 변수 접근법을 써먹어 봅시다. 예를 들어 변수를 직접 보고 싶다면

```
> HealthSys[[6]]
  [1] 20.8 47.6 84.1   NA 76.6 62.2 75.4 69.2 41.8   NA 66.9 75.5 22.8
 [14] 46.1 71.9 34.4 65.6 77.2 75.9 64.9 51.5   NA 83.9 71.8 71.1 56.4
 [27] 46.4 91.8 56.3 54.3 59.5 77.4 24.7 33.5 70.1   NA 49.7 31.3   NA
 [40] 48.6 56.0   NA   NA 75.8 55.9 51.4 73.9 74.6 27.5 82.3 94.2   NA
 …
> HealthSys$OutOfPocketPercTotal
  [1] 74.4 52.2 15.0   NA 17.5 26.7 22.2 20.1 54.6   NA 18.5 15.2 69.0
 [14] 29.1 16.5 63.3 34.4 19.5 19.7 24.5 44.3   NA 15.2 23.2 27.8  5.5
 [27] 31.0  8.1 42.3 36.4 28.3 21.2 61.7 62.6 15.0   NA 45.6 66.4   NA
 [40] 32.1 34.3   NA   NA 14.8 44.1 32.5 25.1 23.1 55.8 13.9  5.8   NA
 …
```

이렇게 하면 되는데 .. 놀라셨죠? 숫자가 그냥 줄줄 나와서. ㅎㅎ 어쩔 수 없습니다.

- 그렇기 때문에 보통은 요약된 내용을 보는 겁니다. 이렇게 말이죠.

```
> summary(GDP$GDPperCapita2012)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    251    1605    5482   14170   15450  103900      29
```

보니까 최소값은 251, 최대는 103900, 평균값은 14170, 중간값은 5480 뭐 이렇군요. 없는 값도 29개나 되구요.

- 내친 김에 몇 가지 더 해 볼까요? 변수의 최대값, 최소값, 평균값, 표준편차, 변수값의 범위 등만 볼 땐 물론 summary(..) 함수를 쓰진 않습니다. 이럴 땐 max(..), min(..), mean(..), sd(..), range(..) 이런 함수를 씁니다. 그럼 2012년 일인 당 GDP 최대나 최소값을 봅시다.

```
> max(GDP$GDPperCapita2012)  ## 2012년 일인당 GDP 최대값
[1] NA
> min(GDP$GDPperCapita2012)  ## 2012년 일인당 GDP 최소값
[1] NA
```

그런데 .. 뭔가 이상하죠? 결과가 숫자가 아니고 "NA"입니다. 이건 값이 없는 게 있어서 어떻게 해야 할지 모르겠다고 R이 불평하고 있는 겁니다. 그런데 그럼 summary(..) 함수는 도대체 어떻게 숫자를 구한 거죠? 뭔가 방법이 있다는 겁니다! 도움말 한 번 봐야겠습니다. 아하, "na.rm"이 있군요. 설명에 따르면 이건 논리값(logical, 그러니까 참(TRUE) 아니면 거짓(FALSE))을 받아들이는 입력 변수로 "na.rm=TRUE"면 "NA"를 다 빼고 계산한다는군요. 진짜 그런지 한번 해 봅시다.

```
> max(GDP$GDPperCapita2012, na.rm=TRUE) ## 2012년 일인당 GDP 최대값
[1] 103858
> min(GDP$GDPperCapita2012, na.rm=T) ## 2012년 일인당 GDP 최소값
[1] 251.0145
```

단계 3. 데이터의 변수의 항목.

- 그런데 앞에서 데이터에 대해서 얘기했던 것과 마찬가지로 우린 매번 변수 전체를 가지고 일하지는 않습니다. 어떨 때는 변수의 특정 항목을 살펴봐야 할 때가 생깁니다. 예를 들어 "2012년 일 인 당 GDP 변수" 가운데 "한국의 항목"은 어떤지 궁금할 때도 있습니다. 그죠? 그럼 이런 건 어떻게 할 수 있을까요?
- 그런데 이 면에서 우린 상황이 좀 불리합니다. 왜냐하면 지금 한국의 2012년 일 인 당 GDP를 알아볼 방법은 딱 하나 밖에 없기 때문입니다. 그건 바로 한국이 몇 번 째 항목인지 찾아보고 (살펴보니 103번째로군요) 그 값을 정해 보는 겁니다.

```
> GDP$GDPperCapita2012[103]
[1] 22589.96  ## 헉, 2012년 우리나라 일인당 GDP가 2만 2천 불이 넘었군요!!
```

- 차이를 찾으셨나요? 변수를 지정할 때는 "`[[..]]`"를 (혹은 "`$`"를) 썼는데 항목을 지정할 때는 "`[..]`"를 썼습니다. 뭐 그건 그런데 이 상황은 영 그렇습니다. 그럼 우리가 다른 나라를 찾으려면 다시 그 나라를 뒤져서 그 순서를 찾아본 다음에 다시 그 번호를 매번 써야 하는 거잖아요! 아니 변수를 지정할 때는 변수 이름을 썼었는데 (그러니까 위에서처럼 "`GDPperCapita2012`" 같은 문자로 썼었는데) 지금은 그럴 수가 없잖아요.
- 그런데 뭔가 이렇게 할 수 있을 것 같지 않으세요? 변수도 이름을 썼었는데 항목이라고 쓰지 말라는 법이 없을 것 같잖아요. 그죠?

- 물론 됩니다! 그런데 그럴려면 좀 준비를 해야 합니다. 무슨 준비가 필요할까요? 뭐 볼 것도 없죠. 당연히 항목 각각에 이름을 부여해야 합니다. 그죠? 그럼 해 봅시다!
- 2012년 일 인 당 GDP 변수는 아시는 것처럼 `HealthSys` 데이터 안에 있습니다. 그런데 애석하게도 데이터 안에 들어 있는 변수는 각각 이름을 지정하지 못 하고 전체 데이터의 "행 이름"을 지정해 쓰는 수 밖에 없더라구요. (이 부분은 설명이 제법 복잡해 지니까 나중에 중급 강좌 뭐 이런 거 할 때 자세히 설명드린다고 합시다요! ^^ 그런데 뭔가 R 시스템의 버그(?) 같은 느낌이 스멀스멀 ..)
- 어쨌거나 이제 데이터의 행 이름을 붙여 보겠습니다. (여기에 붙이는 이름은 "`Country`"가 말고 "`CountryCode`"를 쓰겠습니다. 왜 그런지는 아시겠죠? ^^)

  ```
  > rownames(GDP) <- GDP$CountryCode
  ```

  이렇게 해놓고 나면 한국의 2012년 일 인 당 GDP는 이렇게 알아볼 수 있습니다.

  ```
  > GDP["KOR", "GDPperCapita2012"]
  [1] 22589.96
  ```

- 차이를 찾으셨나요? 변수를 지정할 때는 "`[[..]]`"를 (혹은 "`$`"를) 썼는데 항목을 지정할 때는 "`[a,b]`" 형태를 썼습니다. 짐작 하시겠지만 이건 행렬을 표현하는 방식이죠. `a`는 행의 번호나 이름, `b`는 열의 번호나 이름이구요.

- 행렬의 원소를 지정하는 방식에 문자도 되고 숫자도 되니까 사실 다음과 같은 조합이 다 됩니다.

```
> GDP[103, 5]
[1] 22589.96
> GDP['KOR', 5]
[1] 22589.96
> GDP[103, 'GDPperCapita2012']
[1] 22589.96
```

- 한 가지 팁을 드리면 맨처음에 파일을 읽을 때 행이름을 지정할 수도 있습니다.

```
> GDP <- read.table(<…파일…>, header=T, sep="\t", quote="", as.is=T,
check.names=F, comment.char="#", row.names=2)
```

여기서 "row.names=2"는 "두번째 열을 행 이름으로 쓸테니 대신 걔는 불러 들일 때 변수에서 넣지마라" 이런 뜻입니다. 해보시면 아시겠지만 이렇게 데이터를 불러들이면 두번재 열은 변수에 들어 있지 않습니다!

- 또 한 가지 팁을 드리자면 (첫번째 팁의 방식을 쓰지 않는 경우에 한 해) "**'CountryCode'가 'KOR'인 녀석**"이라는 뜻으로 이렇게 항목의 위치를 정할 수도 있다는 겁니다. (실제 이 방식은 R에서 두고두고 요긴합니다!!)

```
> GDP$GDPperCapita2012[GDP$CountryCode == "KOR"]
[1] 22589.96
```

저 밑줄 그은 부분이 바로 "'CountryCode'가 'KOR'인 녀석"을 나타내고 그 외적인 부분이 바로 "그 때의 2012년 일 인 당 GDP"를 나타냅니다.

- 앞 슬라이드의 두번째 팁은 워낙 활용도가 높은 녀석이니까 좀더 설명을 드리죠. 예를 들어 OECD 국가들의 2012년 1인당 GDP를 다 보려면 이렇게 하시면 됩니다.

```
> GDP$GDPperCapita2012[GDP$OECD == "Y"]
 [1]  67441.593  46791.781  43398.691  51206.156  15452.174  18689.967
 [7]  56363.945  16832.801  45694.115  39746.293  42597.367  22441.521
[13]  12560.075  42339.463  45921.316  32567.089  33815.668  46730.918
…
```

- 여기서 "=="는 R에서 소위 논리연산자라고 부르는 겁니다. 그러니까 참(T), 거짓(F)을 결과값으로 주는 그런 계산이라는 건데 "=="는 "같다"를 나타내고 "다르다"는 "!="로 표현됩니다. 비슷하게 두 숫자 사이에서 "크다", "크거나 같다", "작다", "작거나 같다" 등을 표현할 때 는 각각 ">", ">=", "<", "<=" 등을 씁니다. 그렇기 때문에 똑같은 것도 몇 가지로 표현할 수도 있는데 예를 들어 "OECD가 아닌 나라"를 찾으려면 `GDP$Country[GDP$OECD != "Y"]` 이렇게 할 수도 있고 아니면 `GDP$Country[GDP$OECD == ""]` 이렇게 할 수도 있는 겁니다.

- 논리적으로 따질 땐 당연히 여러 가지 논리값을 합칠 것인자를 고민해야 합니다. 말 그대로 "그리고"나 "또는"이 필요하다는 거죠. R에서 얘들은 각각 "&", "|" (이 문자는 보통 리턴 키 바로 위에 뒤로 누운 해시 "\"키 위에 그려져 있습니다. 그러니까 시프트+"\" 를 하면 나오는 문자죠 ^^)입니다. 그러니까 예를 들어 "100보다 크고 200보다 작은 x"라고 한다면 `x[x > 100 & x < 200]` 이렇게 되는 겁니다.

- 얘네들을 쓰면 재미난 걸 금방 할 수 있습니다. 예를 들어 2012년 일 인 당 GDP가 10000이 넘는 나라는 어떤 나라들이 있을까요?

```
> GDP$Country[GDP$GDPperCapita2012 > 10000]
 [1] NA                          NA
 [3] "Antigua and Barbuda"       "Argentina"
 [5] NA                          "Australia"
 [7] "Austria"                   "Bahamas, The"
 …
```

음 .. 데이터가 없는 녀석들이 또 말썽이군요. 이 녀석들만 빼고 보고 싶으시다면 이렇게 하십시오. (이번엔 "na.rm=T"로 안 됩니다. ㅠㅠ)

```
> GDP$Country[which(GDP$GDPperCapita2012 > 10000)]
 [1] "Antigua and Barbuda"  "Argentina"
 [3] "Australia"            "Austria"
 [5] "Bahamas, The"         "Bahrain"
 …
```

which(A) 이렇게 하면 A에서 (NA나 FALSE는 말고) 참(TRUE)인 녀석만 골라줍니다. ㅎㅎ 그래서 위의 표현은 "2012년 일 인 당 GDP가 10000이 넘지만 NA가 아닌 나라의 이름만 출력하라"는 뜻이 되는 겁니다.

- 2012년 일 인 당 GDP가 10000불에서 20000불 사이인 나라는 어디가 있을까요?

```
> GDP$Country[which(GDP$GDPperCapita2012 >= 10000 & GDP
$GDPperCapita2012 < 20000)]
 [1] "Antigua and Barbuda" "Argentina"           "Barbados"
 [4] "Brazil"              "Chile"               "Croatia"
 [7] "Czech Republic"      "Estonia"             "Gabon"
…
```

총 24개 나라 중에 9개만 보여드렸습니다.

- 마지막으로 한국보다 일인당 GDP가 더 높았던 나라는 어디가 있을까요?

```
> GDP$Country[which(GDP$GDPperCapita2012 >
GDP$GDPperCapita2012[GDP$CountryCode == 'KOR'])]
 [1] "Australia"              "Austria"
 [3] "Bahrain"                "Belgium"
 [5] "Bermuda"                "Brunei Darussalam"
 …
```

총 36개 나라가 있는데 그 가운데 6개만 보여드렸습니다. 여기서 한국의 일 인 당 GDP가 어떻게 표현되었는지만 보시면 되겠죠? 위에서 밑줄 그은 부분 보이시죠? 바로 그 부분이 한국의 일인당 GDP를 나타냅니다.

- 참고로 앞의 예와 같은 경우 저렇게 죽 늘어놓으면 읽기도 불편하고 그렇습니다, 그죠? 이럴 땐 한국의 GDP를 새로운 변수로 지정해 읽기 편하게 만들기도 합니다.

```
> KORGDP <- GDP$GDPperCapita2012[GDP$CountryCode == 'KOR']
> GDP$Country[which(GDP$GDPperCapita2012 > KORGDP)]
 [1] "Australia"              "Austria"
 [3] "Bahrain"                "Belgium"
 [5] "Bermuda"                "Brunei Darussalam"
 …
```

- 마지막 참고로 이제까지는 변수만 고르는데 논리 연산을 썼지만 실제로는 그 쓰임새가 더 광범위합니다. 예를 들어 데이터 가운데도 일부를 고르는데 쓸 수도 있습니다. 즉, HealthSys 가운데 OECD의 것만 골라서 쓰고 싶다고 한다면 이렇게 하시면 됩니다.

```
> ind <- which(HealthSys$OECD == "Y")
> HealthSysOECD <- HealthSys[ind,]
> dim(HealthSysOECD)
[1] 34 15
> HealthSysOECD$Country
 [1] "Australia"       "Austria"         "Belgium"
 [4] "Cameroon"        "Chile"           "Czech Republic"
 [7] "Denmark"         "Estonia"         "Finland"
 …
```

HealthSys 행 가운데 ind에 있는 녀석들만 골라서 쓰겠다는 뜻입니다.

OECD 34개 회원국으로만 구성된 데이터가 되었습니다!

단계 4. 문제 풀이.

- 우리가 풀려고 했던 문제는 “도대체 우리나라의 건강관리 관련 전체 지출에서 공공부문이 차지하는 비율은 전세계의 다른 나라와 비교해서 혹은 비슷한 경제 수준의 다른 나라와 비교해서 어떻게 될까요?”였습니다. 이걸 이제 다뤄 봅시다.
- 이 문제를 여기선 일단 “전체 국가 가운데 우리나라의 공공부문 지출은 순위가 몇 위인가”와 “OECD 국가 가운데 우리나라의 공공부문 지출 순위는 몇 위인가”로 나눠서 풀어보겠습니다. 보시는 것처럼 한 문제를 풀면 다른 문제는 금방 풀립니다, 그죠?
- 우선 우리나라의 공공부문 지출 퍼센트를 구해보겠습니다.

```
> indkor <- which(HealthSys$CountryCode == “KOR”)
> pubkor <- HealthSys$PublicHealthExpensePercTotal[indkor]
> pubkor
[1] 54.4
```

그러니까 우리나라는 공공부문이 전체 지출의 54.4%를 부담하고 있군요.

- 근데 이건 도대체 큰 건가요 작은 건가요? 그걸 위해 공공부문 지출을 전체적으로 요약해 볼까요.

```
> summary(HealthSys$PublicHealthExpensePercTotal)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 16.60   45.95   60.45   59.26   75.43   99.90      26
```

흐억 .. 우리나라 공공부문 지출은 중간값은 물론 전세계 조사대상국의 평균에도 미치지 못 하는군요. 저기엔 우리보다 못 사는 나라도 수두룩한데 …

- 자 이걸 좀더 객관적으로 확인해 보겠습니다. 어떻게 하면 좋을까요? 음 .. 이런 걸 한 번 봅시다. 우리나라의 공공부문 지출은 전세계 국가를 대상으로 했을 때 도대체 몇 번째 정도 될까요?
- 이걸 알아보는데 쓸 함수는 `rank(..)`라는 녀석입니다. `?rank` 해보시면 설명이 죽 나올텐데 변수값을 넣어주면 각 값에 대해 전체 변수 범위에서의 등수를 보여줍니다. 헷갈리지 말아야 하는 건 이 등수가 작은 값에서부터 시작한다는 거죠.

```
> rank(HealthSys$PublicHealthExpensePercTotal)[indkor]
[1] 74
```

헉! 그러니까 우리나라의 공공부문 지출은 아래에서부터 즉 공공부문 지출이 작은 순서로 74번째라는 겁니다. 그럼에도 불구하고 전체 214개 국가 가운데 밑에서 74번째는 아닙니다. 왜냐하면 `rank(..)`의 설명에 따르면 기본적으로 `na.last=TRUE`라고 되어 있으니까 데이터 값이 없는 26개 국가의 등수는 뒤에 둔 것이죠. 그러니까 214개 국가 가운데 아래에서 74번째가 아니라 188개 국가 가운데 아래에서 74번째인 것이죠. ㅠㅠ 그러니까 전체 국가의 분포에서 하위 39% (= 74 / 188)에 해당하는 만큼의 돈만 공공부문이 쓰고 있다는 겁니다. 우리 생활이 달리 빡빡한 게 아니었군요. 다 데이터에 들어 있었습니다.

- 공공부문 지출이 우리보다 적은 나라 가운데 우리나라보다 일 인 당 GDP가 큰 나라는 얼마나 될까요?

```
> HealthSys$Country[HealthSys$PublicHealthExpensePercTotal < pubkor &
GDP$GDPperCapita2012 > KORGDP]
character(0)
```

불행히도 없군요! ㅠㅠ

- 그럼 우리나라의 공공부문 지출은 OECD를 대상으로 했을 때 몇 번째 정도 될까요?

```
> indkoroecd <- which(HealthSysOECD$CountryCode == “KOR”)
> rank(HealthSysOECD$PublicHealthExpensePercTotal)[indkoroecd]
[1] 5
> dim(HealthSysOECD)
[1] 34 15
> summary(HealthSysOECD$PublicHealthExpensePercTotal)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 33.50   63.80   74.65   70.95   80.50   85.50
```

역시 이 부분에서는 정말 초라하기 짝이 없군요. 아래에서부터 공공부문 지출이 작은 순서로 5번째라는 겁니다. 전체 OECD 국가가 34개이고 이들 국가에서는 누락된 데이터가 없으니까 결국 34개 국가 가운데 밑에서 5번째라는 겁니다. ㅎㅎ 게다가 OECD 전체의 평균이나 중간값에도 미치지 못 할 뿐 아니라 OECD 국가 전체의 공공부문 지출 분포를 0 ~ 100%로 봤을 때 25% (1st Qu.)에도 미치지 못 합니다. 정말 가슴 답답한 수준이군요. ㅠㅠ