# Data Science

## Homework 04

1. 한동 honor code 에 맞게 과제를 진행하여 주세요.
2. 과제의 경우 팀당 1 개의 결과물을 제출하면 됩니다.
3. 과제 제출 기한은 ~4/11 23:59 입니다. (1 분당 0.1 감점)
4. 제출은 LMS>과제 및 평가>Homework04 로 하시면 됩니다. (팀 내 1 명이 제출)
5. LMS 제출이 안되는 경우는 TA 이메일로 제출하시기 바랍니다. (22100733@handong.ac.kr)

1. Please proceed with the assignment following the Handong honor code.
2. For assignments, one submission per team is sufficient.
3. The deadline for assignment submission is until 4/11, 23:59. (0.1 points deducted per minute late)
4. Submissions should be made to LMS>Assignments>Homework04. (One member of the team should submit)
5. If you cannot submit via LMS, please submit to the TA email. (22100733@handong.ac.kr)

모든 학생들은 아래의 링크에 접속하여, 본 과제물에 대한 Peer Evaluation을 진행하여 주시기 바랍니다.
제출시간 마감 이전에 응답하지 않으면 불이익이 있을 수 있습니다.
All students are requested to access the link below and conduct a Peer Evaluation for this assignment.

01분반(KOR) - https://forms.gle/ybauzsr1wPbPuJJy7
2rd Class(ENG) - https://forms.gle/ZFsZ3D24cFK7eEL9A

# Data Science - Practice 4

모든 문제에 대하여 코드만 작성하지 말고 데이터를 해석한 결과를 함께 작성하시오.
Put your explanation of your findings from dataset with your answer in your report as well as R code.

## Loading data into R

```
GDP <- read.csv('GDP.csv')
POP <- read.csv('population.csv')
LIFE_EXP <- read.csv('Life  Expectancy.csv')
```

```
str(GDP)
```
```
##   'data.frame':      204   obs.  of  2 variables:
##   $ Country          :   chr    "Afghanistan" "Albania" "Algeria" "Andorra" ...
##   $ GDP              :   int    1763 12694 13967 53245 5544 24463 17529 9728 45495 46873 ...
```

```
str(POP)
```
```
##   'data.frame':      195   obs.  of  2 variables:
##   $ Country          :   chr    "Afghanistan" "Albania" "Algeria" "Andorra" ...
##   $ X2019            :   int    38000000 2880000 43100000 77100 31800000 97100 44800000 2960000 25200000..
```

```
str(LIFE_EXP)
```
```
##   'data.frame':      186   obs.  of  2 variables:
##   $ Country          :   chr    "Afghanistan" "Albania" "Algeria" "Andorra" ...
##   $ Life_exp         :   num    64.1 78.5 78.1 65 77.3 ...
```

## Data description

**GDP** data frame
2019년 기준 204개 국가별 GDP = GDP of 204 countries(2019)

| variable | 의미 | description |
|---|---|---|
| Country | 국가 | country_name |
| GDP | 1인당 국내총생산(USD) | Gross Domestic Product per capita (USD) |

**POP** data frame
2019년 기준 195개 국가별 인구수 = Population of 195 countries(2019)

| variable | 의미 | description |
|---|---|---|
| Country | 국가 | country_name |
| X2019 | 2019년 인구수 | population in 2019 |

**LIFE_EXP** data frame
2019년 기준 186개 국가별 기대수명 = Life expectancy of 186 countries(2019)

| variable | 의미 | description |
|---|---|---|
| Country | 국가 | country_name |
| Life_exp | 기대수명 | Life_expectancy |

---

**< Question 1 >**

불러온 데이터들 중에서 GDP와 인구수 데이터프레임 열의 이름을 바꿔보자.
GDP는 각각 'Country'와 'GDP'로, 인구수는 'Country'와 'POP'로 바꿔보도록 하자.
Change the column names of GDP and POP data frame. For GDP set the names as 'Country' and 'GDP' and for POP, set them to 'Country' and 'POP'

*Sample Result*

```
##  'data.frame':  204  obs.  of  2 variables:
##  $ Country      :  chr    "Afghanistan" "Albania" "Algeria" "Andorra" ...
##  $ GDP          :  int    1763 12694 13967 53245 5544 24463 17529 9728 45495 46873 ...

##  'data.frame':  195  obs.  of  2 variables:
##  $ Country      :  chr    "Afghanistan" "Albania" "Algeria" "Andorra" ...
##  $ POP          :  int    38000000 2880000 43100000 77100 31800000 97100 44800000 2960000 25200000..
```

### < Question 2 > - ‘merge’ function

GDP와 인구수 데이터프레임을 합쳐서 새로운 데이터프레임을 만들어보고, 적절하게 잘 합쳐졌는지 확인하자.
Merge the 'GDP' dataframe and 'POP' dataframe to one and examine if two dataframes were properly merged.

*Sample Result*

```
##                 Country   GDP      POP
## 1           Afghanistan  1763 38000000
## 2               Albania 12694  2880000
## 3               Algeria 13967 43100000
## 4               Andorra 53245    77100
## 5                Angola  5544 31800000
## 6  Antigua and Barbuda 24463    97100
```

### < Question 3 >

2번에서 만든 데이터프레임과 기대수명 데이터를 합쳐서 새로운 데이터프레임을 만들어보자. 적절하게 합쳐졌는지 확인하여라.
Merge the 'LIFE EXP' data with data frame that you made in the previous question. See if they are merged as you expected.

*Sample Result*

```
##                 Country   GDP      POP Life_exp
## 1           Afghanistan  1763 38000000    64.08
## 2               Albania 12694  2880000    78.47
## 3               Algeria 13967 43100000    78.12
## 4                Angola  5544 31800000    65.00
## 5  Antigua and Barbuda 24463    97100    77.28
## 6             Argentina 17529 44800000    76.96
```

### < Question 4-1 > - ‘subset’ function

GDP가 대한민국보다 높은 나라들만 subset 함수를 이용하여 추출해보자. 출력은 국가명만 나오도록 하자.
Use 'subset' function to select the country names of GDP higher than GDP of South Korea. Show the list of country names.

*Sample Result*

```
##  [1] "Australia" "Austria"   "Bahrain"   "Belgium"   "Brunei"    "Canada"
```

**< Question 4-2 >**

GDP가 대한민국보다 높으면서 인구가 대한민국보다 적은 국가를 찾고, 출력은 국가명, GDP, 인구수가 보이도록 하자.

Find countries that meet the following conditions.

1. higher GDP than South Korea.
2. lower Population than South Korea.

*Sample Result*

```
##         Country   GDP      POP
## 8     Australia 45495 25200000
## 9       Austria 46873  8960000
## 12      Bahrain 41966  1640000
## 16      Belgium 43517 11500000
## 24       Brunei 72376   433000
## 30       Canada 44181 37400000
```

**< Question 4-3 >**

기존 데이터프레임에 Country_GDP(=GDP * POP/1000) (단위 1000 USD) 라는 변수를 새롭게 추가하고 Country_GDP가 미국보다는 낮고, 대한민국보다는 높은 국가들을 출력하라.

Add new column 'Country_GDP'( = GDP * POP / 1000 )

Find the countries whose Country_GDP is lower than United States and higher than South Korea.

*Sample Result*

```
##                 Country   GDP       POP Life_exp Country_GDP
## 23               Brazil 14307 211000000    75.93  3018777000
## 58               France 39989  65100000    83.07  2603283900
## 62              Germany 46173  83500000    80.92  3855445500
## 74                India  7227 1370000000    69.46  9900990000
## 75            Indonesia 12061 271000000    71.91  3268531000
## 80                Italy 35816  60600000    83.49  2170449600
## 82                Japan 39739 127000000    84.53  5046853000
## 105              Mexico 18002 128000000    75.63  2304256000
## 134              Russia 25654 146000000    72.52  3745484000
## 168              Turkey 25039  83400000    79.48  2088252600
## 173 United Kingdom 40392  67500000    81.12  2726460000
```

**< Question 5-1 > - 'sample' function**

sample 함수를 사용하여 임의로 20개의 국가를 추출하여라. 추출한 국가들의 GDP, POP, Life_exp의 평균을 계산하여라.

Use 'sample' function to randomly extract 20 countries. Calculate the average of those countries' GDP, Population, Life expectancy.

*Sample Result*

```
##          GDP          POP     Life_exp  Country_GDP
##     15519.90 101746700.00        71.94 1131899366.20
```

**< Question 5-2 >**

5-1번의 과정을 10번 반복한 후 그 결과를 새로운 matrix로 저장해라. matrix의 행에는 1번의 시도에 대한 결과를 기록하여 10개의 시도를 10개의 row로 표현한다.
Repeat the previous process 10 times and record each trial as a row of matrix. Since you are asked to perform 10 times, the matrix should have 10 rows.
tip) If the output is different from the example, check the function "t()" which is for transpose.

*Sample Result*

```
##              GDP      POP Life_exp Country_GDP
##  [1,] 12839.10 37697850    71.61  1192863147
##  [2,] 20519.55 42037450    74.31   607208683
##  [3,] 19193.30 35868900    73.90  1300765794
##  [4,] 15994.10 29196000    74.23   420455662
##  [5,] 20865.70 14230450    72.94   171653395
##  [6,] 22539.75 22978885    74.99   370606577
##  [7,] 18261.00 91732500    71.58   776907126
##  [8,] 13853.40 17478150    72.16   341582384
##  [9,] 16397.85 34888000    73.14   755744287
## [10,] 17252.60 32958000    74.78   630229224
```

**< Question 5-3 >**

방금 구한 matrix로부터 GDP, POP, Life_exp의 평균을 각각 구하고, 그것을 전체 국가의 dataset의 GDP, POP Life_exp의 평균과 비교해보고 비슷한 값을 얻었는지 보라.
Calculate the average of GDP, POP, Life_exp from the matrix. Compare them with the average from the original dataframe of all countries.

*Sample Result*

```
##          GDP          POP     Life_exp  Country_GDP
##    17771.635 35906618.500       73.364 656801627.708
```

```
##          GDP          POP     Life_exp  Country_GDP
##        18271     42147164           73    674061178
```

**< Question 5-4 >**

set.seed(2024)이라는 함수를 입력한 이후에 5-2번 문제를 다시 수행하라. 그 결과를 팀 동료들과 비교해보고 어떠한 차이가 있는지 이야기해봐라. set.seed의 역할이 무엇이고 언제 사용할 수 있는지 생각해보아라.
(R버전에 따라서 결과가 다를 수 있습니다.)
Run the code 'set.seed(2024)' and try question 5-2 again. Compare your result with your team members'.
Explain what this function "set.seed" does and when to use.
(depending on the version of R, the result could be somewhat different)

*Sample Result*

```
##          GDP  Population     Life_exp    GDPbyPOP
##    16747.875 28915655.500      72.861 353006337.954
```

**< Question 6 >**

Country_GDP 변수는 숫자의 단위가 너무 그다. 1,000,000으로 나누고, 소수점 2자리에서 반올림한 뒤 B(Bi11ion)을 붙여서 표기하라. (단위 USD)
Since Country_GDP has very large values that is not easy for us to read through, Divide Country_GDP by 1,000,000, round off at the 2 decimal places and put B(BiIlion) at the tail.
example) 66994000 -> 66.99B

*Sample Result*

```
##                 Country Country_GDP
##  1          Afghanistan      66.99B
##  2              Albania      36.56B
##  3              Algeria     601.98B
##  4               Angola      176.3B
##  5  Antigua and Barbuda       2.38B
##  6            Argentina      785.3B
##  7              Armenia      28.79B
##  8            Australia    1146.47B
##  9              Austria      419.98B
##  10          Azerbaijan      162.48B
```

**< Question 7 > - 'which' function**

which 함수를 이용하여 GDP가 평균보다 높은 국가의 index, 인구수가 평균보다 높은 국가의 index, 기대수명이 평균보다 높은 국가의 index를 각각 찾고 intersect 함수를 사용해서 index의 교집합을 찾아라.

조건에 해당하는 국가가 몇개나 되는지, 그 국가들의 이름을 출력해보아라.

Using the which function, find the index of the countries of GDP higher than average, the index of the countries of population higher than average, and the index of the countries of life expectancy higher than average.
Find the intersection of those countries using "intersect" function to the indices.
How many countries are there?

*Sample Result*

```
##[1] "France"         "Germany"    "Italy"          "Japan"
##[5] "South Korea"    "Spain"      "Turkey"         "United Kingdom"
##[9] "United States"
```

### < Question 8 > - ‘quantile, cut’ function

GDP의 크기에 따라 국가를 총 4개의 그룹으로 분류하려고 한다. (Very Low, Low, High, Very High)
각 그룹에 속하는 나라의 수를 최대한 동등하게 나눠보려고 한다.
quantile과 cut 함수를 이용해서 'GDP_group'이라는 새로운 변수를 만들어보자. table 함수를 이용하여 잘 나눠졌는지 확인하라.
Split counties into four groups of VeryLow, Low, High, and VeryHigh according to their GDP.
Use quantile and cut function to add new columns named 'GD_ group'.
compare the number of countries of each group using table function to see if they are equally distributed.

*Sample Result*

```
##        0%       25%       50%       75%      100% 
##     631.0    3891.5   11849.0   26877.5  113331.0 
##  Very low       Low      High Very High           
##        46        45        45        46
```

### < Question 9 > - ‘aggregate’ function

8번 문제에서 나눈 그룹을 기준으로 인구, 기대수명의 그룹별 평균을 구하여 비교해보아라.
경제수준(GDP)와 인구, 기대 수명이 상관관계가 있다고 생각되는가?
Based on the GDP_group, find average of population and life expectancy.
Can we say that population and life expectancy are related to GDP?

*Sample Result*

```
##   GDP_group      POP   Life_exp
## 1  Very Low 18862870   65.09978
## 2       Low 61238622   71.94244
## 3      High 64017758   75.63356
## 4 Very High 25860549   80.25600
```