

Data Science

Homework 01

1. 한동 honor code 에 맞게 과제를 진행하여 주세요.
2. 과제의 경우 팀당 1 개의 결과물을 제출하면 됩니다.
3. 과제 제출 기한은 ~3/21 23:59 입니다. (1 분당 0.1 감점)
4. 제출은 LMS>과제 및 평가>Homework01 로 하시면 됩니다. (팀 내 1 명이 제출)
5. LMS 제출이 안되는 경우는 TA 이메일로 제출하시기 바랍니다. (22200479@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 3/21, 23:59. (0.1 points deducted per minute late)
4. Submissions should be made to LMS>Assignments>Homework01. (One member of the team should submit)
5. If you cannot submit via LMS, please submit to the TA email. (22200479@handong.ac.kr)

**모든 학생들은 아래의 링크에 접속하여, 본 과제물에 대한 Peer Evaluation 을 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.

01 분반(KOR) - <https://forms.gle/RKnjY7YNF5vvbmpY7>

2rd Class(ENG) - <https://forms.gle/x23rQsk6nRShTDao9>

<주의>

과제를 진행하며 **경고 메시지(Warning message)**가 출력될 수 있습니다. R에서 경고메시지는 프로그램이 정상적으로 실행되어도 잠재적으로 문제가 발생할 가능성이 있는 경우에 출력되는 메시지입니다.

반면 **에러(Error)**는 프로그램이 정상적으로 실행되지 않을 정도로 심각한 문제를 나타냅니다.

경고메시지는 큰 문제가 아니라면 무시하고 넘어가도 되지만 에러는 대처해야 넘어갈 수 있으므로 경고메시지를 에러로 오해하여 난항을 겪지 않기를 바랍니다.

Data Science – Practice 1

Problem

한동대학교 학생들이 가장 좋아하는 영화는 ‘어벤져스, 해리포터, 나홀로집에, 토이스토리, 겨울왕국, 노트북, 인터스텔라’ 라고 합니다. 이 영화들을 통해서 지금까지 배운 내용들을 실습해보도록 하겠습니다. (단, NA 값이 있다면 제외하고 계산하세요.)

Handong Global University students' favorite films are 'The Avengers, Harry Potter, Home Alone, Toy Story, Frozen, The Notebook, and Interstellar'. We will practice what we have learned so far through the listed films. (If you have NA values, exclude them).

< Question 1 >

위에서 제시된 7 가지 영화들의 이름을 새로운 vector 로 (vector 의 이름은 'movie') 만들어보세요.

Create a new vector (name of vector = 'movie') with the listed seven films.

Sample Result

```
## [1] "The Avengers" "Harry Potter" "Home Alone"   "Toy Story"    "Frozen"
## [6] "The Notebook" "Interstellar"
```

< Question 2 >

각 영화들에 대한 자신의 평점을 5점 만점으로 하여 새로운 vector로 ('my_rating') 만들어보세요. (단, 보지 않았던 영화가 있으면 NA값으로 할당하세요.)

Create your own rating vector ('my_rating') for each movie with a scale of 1 to 5 points. (If there are movies that you have not watched, assign their values to NA values.)

Sample Result

```
## [1] 2.8 3.4 5.0 4.2 1.6 2.7 4.3
```

< Question 3 >

TA 학생은 위 영화들에 대해 '4.3, NA, 3.8, 3, 2.8, NA, 1.6'의 평점을 매겼습니다. 이 값들을 새로운 vector 로 ('TA_rating') 만들어보세요.

The teacher's assistant rated the above films as '4.3, NA, 3.8, 3, 2.8, NA, 1.6' respectively. Create a new vector ('TA_rating') including these values.

Sample Result

```
## [1] 4.3 NA 3.8 3.0 2.8 NA 1.6
```

< Question 4 >

팀원들과 영화 평점을 공유하고, 팀원들의 평점을 담은 vector를 각각 만들어보세요. 그리고 지금까지 만들었던 vector를 모두 합쳐서 하나의 matrix로 ('team_matrix') 만들어보세요.

Share the movie ratings with your team members and create new vectors using your team members' ratings. Also, create a new matrix ('team_matrix') including all the rating vectors you have made so far.

Sample Result

```
##           [,1] [,2] [,3] [,4] [,5] [,6] [,7]
## [1,]    4.3   NA  3.8  3.0  2.8   NA  1.6
## [2,]    2.8  3.4  5.0  4.2  1.6  2.7  4.3
## [3,]    1.0  2.0  3.0  4.0  3.8  2.3  4.1
## [4,]    2.0  3.0  4.2   NA  4.7  1.4  4.1
## [5,]    NA  2.0  4.0  1.2  3.0  2.3  4.6
```

< Question 5 >

어떤 학생이 평점을 후하게 주고, 어떤 학생이 짜게 주는 경향이 있는지 파악하고자 합니다. 각 학생 별로 부여한 영화 평점의 평균을 구하여 새로운 vector 로 ('stu_mean') 만들어 보세요.

Each student's tendency to give high or low ratings needs to be figured out. Create a new vector ('stu_mean') by calculating the average movie ratings of each student.

Sample Result

```
## [1] 3.100000 3.428571 2.885714 3.233333 2.850000
```

< Question 6 >

어떤 영화가 전반적으로 관람객이 많고 평점이 좋은지 파악하고자 합니다. 각 영화 별로 받은 평점의 합을 구하여 새로운 vector ('movie_sum') 로 만들어보세요.

The movies' average overall ratings need to be figured out. Hence, create a new vector ('movie sum') by adding the average ratings of the movies given by each student.

Sample Result

```
## [1] 10.1 10.4 20.0 12.4 15.9 8.7 18.7
```

< Question 7 >

문제 5 번에서 만든 학생별 평균 ('stu_mean')을 4 번에서 만든 team_matrix 의 마지막 열에 추가해보세요.

Add the vector ('stu_mean') that you have created in question 5 to the last column of the matrix ('team_matrix') you have created in question 4.

Sample Result

```
##                                     stu_mean
## [1,] 4.3   NA  3.8  3.0  2.8   NA  1.6 3.100000
## [2,] 2.8  3.4  5.0  4.2  1.6  2.7  4.3 3.428571
## [3,] 1.0  2.0  3.0  4.0  3.8  2.3  4.1 2.885714
## [4,] 2.0  3.0  4.2   NA  4.7  1.4  4.1 3.233333
## [5,]  NA  2.0  4.0  1.2  3.0  2.3  4.6 2.850000
```

< Question 8 >

문제 6번에서 만들었던 영화 평점의 합 벡터를 7번에서 만든 matrix의 마지막 행에 추가해보세요. 6번 문제의 벡터의 길이와 7번 matrix의 열의 수가 다른데, **rbind**를 이용해 둘을 합칠 때에 어떤 현상이 발생하는지 설명해보자.

Add the vector ('movie _ sum') that you have created in question 6 to the last row of the matrix ('team _ matrix') you have created in question 7. The length of vector is different from the number of columns of matrix, what happen when you put combine them with **rbind** command?

Sample Result

```
##                                     stu_mean
##          4.3    NA    3.8    3.0    2.8    NA    1.6    3.100000
##          2.8    3.4    5.0    4.2    1.6    2.7    4.3    3.428571
##          1.0    2.0    3.0    4.0    3.8    2.3    4.1    2.885714
##          2.0    3.0    4.2    NA    4.7    1.4    4.1    3.233333
##          NA    2.0    4.0    1.2    3.0    2.3    4.6    2.850000
## movie_sum 10.1  10.4  20.0  12.4  15.9    8.7  18.7  10.100000
```

< Question 9 >

문제 8 번에서 만든 matrix 의 가장 마지막 행의 마지막 열 (예시 답안에서는 '10.1'에 해당하는 값)은 별로 의미가 없는 값입니다. 이 값을 'NA'로 변환하세요.

The matrix's last row's last column you have created in question 8 (the value equivalent to '10.1' in the example) is not a meaningful value. Hence, convert the value to 'N A'.

Sample Result

```
##                                     stu_mean
##          4.3    NA    3.8    3.0    2.8    NA    1.6    3.100000
##          2.8    3.4    5.0    4.2    1.6    2.7    4.3    3.428571
##          1.0    2.0    3.0    4.0    3.8    2.3    4.1    2.885714
##          2.0    3.0    4.2    NA    4.7    1.4    4.1    3.233333
##          NA    2.0    4.0    1.2    3.0    2.3    4.6    2.850000
## movie_sum 10.1  10.4  20.0  12.4  15.9    8.7  18.7    NA
```

< Question 10 >

matrix 를 직관적으로 이해하기 위해서는 행과 열에 이름을 붙여주는 것이 좋습니다. 행은 학생의 이름으로, 열은 영화의 이름으로 할당해보세요.

For an intuitive understanding of the matrix, it is recommended to name the rows and columns. Hence, assign the rows and columns to the students' names and the movies titles respectively.

Sample Result

```
##          Avengers  H.P Home Alone  T.S Frozen  Notebook  Interstellar  stu_mean
## TA              4.3    NA          3.8    3.0    2.8          NA          1.6 3.100000
## Paul            2.8    3.4          5.0    4.2    1.6          2.7          4.3 3.428571
## John            1.0    2.0          3.0    4.0    3.8          2.3          4.1 2.885714
## Jessie          2.0    3.0          4.2    NA    4.7          1.4          4.1 3.233333
## Mary            NA    2.0          4.0    1.2    3.0          2.3          4.6 2.850000
## movie_sum      10.1  10.4          20.0  12.4    15.9          8.7          18.7    NA
```

< Question 11 >

어떤 사람은 영화에 대한 평점을 평균적으로 후하게 주고, 어떤 사람은 평균적으로 박하게 주기도 한다. 따라서 영화를 감상한 관객의 성향 분포에 따라서 영화평점은 편향되게 나타나기도 한다.

관객들의 영화 평점에서 해당 관객의 평균 평점을 빼주면 이러한 편향을 줄일 수 있다. 이러한 방법을 **centering**이라고 하는데 **centering**을 수행하여보자.

Some audiences tend to rate movie generously, whereas some other are relatively harsh. So according to the group of audience, movie rating can be bias.

By subtracting average rating of certain audience from all his movie ratings, we may reduce bias which is called **centering**. Let us perform **centering** on our rating matrix.

Sample Result

##	Avengers	H.P	Home Alone	T.S	Frozen	Notebook
## TA	1.2000000	NA	0.7000000	-0.1000000	-0.3000000	NA
## Paul	-0.6285714	-0.02857143	1.5714286	0.7714286	-1.8285714	-0.7285714
## John	-1.8857143	-0.88571429	0.1142857	1.1142857	0.9142857	-0.5857143
## Jessie	-1.2333333	-0.23333333	0.9666667	NA	1.4666667	-1.8333333
## Mary	NA	-0.85000000	1.1500000	-1.6500000	0.1500000	-0.5500000
##	Interstellar					
## TA	-1.5000000					
## Paul	0.8714286					
## John	1.2142857					
## Jessie	0.8666667					
## Mary	1.7500000					

< Question 12 >

centering 을 하기 전 영화별 평균 평점과 **centering** 을 수행한 후 영화별 평균 평점은 달라질 수 있다. 각각을 계산해보고 어떻게 달라졌는지 비교하시오.

The average rating of each movie after centering can differ from the average rating before **centering**. Calculate the movie's average rating before and after centering.

Sample Result

•before centering

##	Avengers	H.P	Home Alone	T.S	Frozen	Notebook
##	2.525	2.600	4.000	3.100	3.180	2.175
##	Interstellar					
##	3.740					

•after centering

##	Avengers	H.P	Home Alone	T.S	Frozen	Notebook
##	-0.63690476	-0.49940476	0.90047619	0.03392857	0.08047619	-0.92440476
##	Interstellar					
##	0.64047619					