# Multivariate Data Analysis Assignment #7

Dataset: MOOC Dataset (big_student_clear_third_version.csv)

해당 데이터셋은 MOOC 강좌를 수강한 수강생들에 대한 정보가 포함되어 있는 데이터 셋이다. 다음 각 Instruction에 따라 데이터를 변환하고 연관규칙분석을 수행하여 각 결과물을 제시하고 적절한 해석을 제공하시오.

## [Step 1] 데이터 변환

mooc_dataset <- read.csv("big_student_clear_third_version.csv")

위 스크립트를 사용하여 원 데이터를 불러오면 총 416,921건의 관측치와 22개의 변수가 존재하는 데이터프레임이 생성된다. 이 중에서 아래 그림과 같이 userid_DI (사용자 아이디)를 Transaction ID로 하고, institute (강좌 제공 기관), course_id (강좌코드), final_cc_cname_DI (접속 국가), LoE_DI (학위 과정)을 하나의 string으로 결합하여 Item Name으로 하는 Single 형식의 csv 파일로 저장하고자 한다.

Filter

| | X | institute | course_id | year | semester | userid_DI | viewed | explored | certified | final_cc_cname_DI | LoE_DI | gender | grade |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 4 | HarvardX | PH207x | 2012 | Fall | MHxPC130313697 | 0 | 0 | 0 | India | Bachelor's | m | 0.00 |
| 2 | 6 | HarvardX | PH207x | 2012 | Fall | MHxPC130237753 | 1 | 0 | 0 | United States | Secondary | m | 0.00 |
| 3 | 7 | HarvardX | CS50x | 2012 | Summer | MHxPC130202970 | 1 | 0 | 0 | United States | Bachelor's | m | 0.00 |
| 4 | 20 | HarvardX | CS50x | 2012 | Summer | MHxPC130223941 | 1 | 0 | 0 | Other Middle East/Central Asia | Secondary | m | 0.00 |
| 5 | 22 | HarvardX | PH207x | 2012 | Fall | MHxPC130317399 | 0 | 0 | 0 | Australia | Master's | f | 0.00 |
| 6 | 23 | HarvardX | CS50x | 2012 | Summer | MHxPC130191782 | 1 | 0 | 0 | Pakistan | Bachelor's | m | 0.00 |
| 7 | 24 | HarvardX | ER22x | 2013 | Spring | MHxPC130191782 | 1 | 0 | 0 | Pakistan | Bachelor's | m | 0.00 |
| 8 | 26 | HarvardX | PH207x | 2012 | Fall | MHxPC130267000 | 0 | 0 | 0 | Other South Asia | Master's | f | 0.00 |
| 9 | 27 | HarvardX | CS50x | 2012 | Summer | MHxPC130435800 | 1 | 0 | 0 | India | Bachelor's | m | 0.00 |
| 10 | 28 | HarvardX | PH207x | 2012 | Fall | MHxPC130284813 | 0 | 0 | 0 | United States | Bachelor's | m | 0.00 |
| 11 | 29 | HarvardX | CS50x | 2012 | Summer | MHxPC130235150 | 1 | 1 | 0 | India | Bachelor's | m | 0.00 |
| 12 | 30 | HarvardX | CS50x | 2012 | Summer | MHxPC130001411 | 1 | 1 | 0 | Other Europe | Secondary | m | 0.00 |
| 13 | 31 | HarvardX | PH207x | 2012 | Fall | MHxPC130396873 | 0 | 0 | 0 | United States | Bachelor's | m | 0.00 |
| 14 | 33 | HarvardX | CB22x | 2013 | Spring | MHxPC130469401 | 1 | 0 | 0 | Other Middle East/Central Asia | Bachelor's | | 0.00 |
| 15 | 34 | HarvardX | CS50x | 2012 | Summer | MHxPC130469401 | 1 | 0 | 0 | Other Middle East/Central Asia | Bachelor's | | 0.00 |

Item name　　Transaction ID　　Item name

[Q1] 아래 사항을 수행하는 스크립트와 최종 csv 파일 내용을 캡처하여 제시하시오.

- ✓ 1단계: 제공된 csv파일을 읽어온 뒤 Item Name에 해당하는 네 개의 변수를 각각 Institute, Course, Region, Degree로 정의한 변수명에 저장하시오.
- ✓ 2단계: Region에 해당하는 변수의 경우 한 칸 공백(" ")을 제거하는 전처리를 수행하시오(hint: gsub( ) 함수 사용).
- ✓ 3단계: 네 변수를 밑줄(_)로 연결하여 RawTransactions로 정의된 하나의 변수에 저장하시오 (hint: paste( ) 함수 사용).
- ✓ 4단계: Transaction ID에 해당하는 변수와 3단계의 결과물을 한 칸 공백(" ")으로 연결하여

MOOC_transactions로 정의된 변수에 저장하시오.

- ✓ 5단계: 4단계의 MOOC_transactions 변수를 MOOC_User_Course.csv라는 파일명으로 저장하고 저장된 파일을 열어 내용을 확인하시오.

**[Step 2] 데이터 불러오기 및 기초 통계량 확인**

[Q2-1] [Q1]에서 생성된 single format의 데이터를 read.transactions( )함수를 이용하여 읽어들이고 summary( )함수를 사용하여 해당 데이터의 속성을 파악해보시오.

[Q2-2] 아이템 이름과 아이템 카운트를 이용하여 워드클라우드를 생성해 보시오. 워드클라우드 생성 시 Color는 실습 자료로 제시된 것과 다른 색상을 사용하여 생성하며, min.freq는 최소 100 이상의 값을 직접 지정해서 생성하시오.

[Q2-3] itemFrequencyPlot( )함수를 사용하여 최소 빈도 1% 이상 등장한 Items들의 Bar Chart를 도시하시오. 상위 5개의 Item에 대해 접속 국가는 각각 어느 국가인지 확인하시오.

**[Step 3] 규칙 생성 및 결과 해석**

[Q3-1] 최소 10개 이상의 규칙이 생성될 수 있도록 support와 confidence의 값을 조정해 가면서 각 support-confidence 조합에 대해 총 몇 가지의 규칙이 생성되는지 확인하고 그 결과를 아래 표와 같은 형태로 제시하시오. 최소한 3개 이상의 support, 3개 이상의 confidence, 총 9개 이상의 조합에 대한 규칙 생성을 수행하시오.

| Number of rules | Confidence = 0.XXX | Confidence = 0.XXX | … |
|---|---|---|---|
| Support = 0.XXX | | | |
| Support = 0.XXX | | | |
| … | | | |

[Q3-2] support = 0.001, confidence = 0.05로 지정하여 생성된 연관규칙분석들에 대해 다음 질문에 대한 답과 본인의 생각을 서술하시오.

- ✓ Support가 가장 높은 규칙은 무엇인가?
- ✓ Confidence가 가장 높은 규칙은 무엇인가?
- ✓ Lift가 가장 높은 규칙은 무엇인가?
- ✓ 만일 하나의 규칙에 대한 효용성 지표를 Support × Confidence × Lift로 정의한다면 효용성이 가장 높은 규칙 1위~3위는 어떤 것들인가?

- ✓ 생성된 규칙을 plot( )함수의 “graph” method를 이용하여 도시할 경우 두 아이템이 서로 조건절/결과절을 달리해서 생성되는 경우가 존재함을 확인할 수 있다(X → Y 규칙과 Y → X 규칙이 함께 존재한다는 뜻). 이 중에서 세 가지 규칙을 선택하여 각 규칙들에 대한 Support/Confidence/Lift 값을 확인해보고 조건절과 결과절의 위치에 따라서 어떤 지표 값들이 차이가 나는지와 왜 그러한 상황이 발생하는지 서술하시오.

**[Extra Question] 이 외 수업시간에 다루지 않은 연관규칙분석 시각화 및 해석을 시도해 보시오.**