**비즈니스 모델링 1**

**Assignment 5**

**Due: 5/17, 2020 (일요일), 11:59PM**

**Instructions:** 아래의 문제를 잘 읽고 해당 데이터분석을 실행하는 R 코드를 작성하여 그 스크립트 파일을 제출하시오. 파일명은 ‘HW6\_김일주.r’로 저장하여 제출하시오. 데이터의 특징 등 분석에 관한 설명을 요하는 문제의 경우에는 현재 MS Word파일을 통해 제출하도록 하시오. Word 파일의 이름은 ‘숙제6\_김일주(본인이름).docx’로 하시오. R 코드파일에서, 각 문제의 번호를 주석으로 명시한 후 해당 코드를 그 아래에 위치시키고, 추가적으로 설명해야 부분이 있다면 그 역시 주석으로 제출하시오. 늦은 제출은 받지 않으니, 제출시한에 맞추도록 유의하시오. 문제 번호(와 설명)을 담은 주석이 없을시 감점됨에 주의하시오. 질문 있을 시 [ijkim@mju.ac.kr](mailto:ijkim@mju.ac.kr)로 연락하시오.

1. **(5 pts) ggplot2 패키지에 미국 동북중부 437개 지역의 인구통계 정보를 담은 midwest라는 데이터가 있습니다. 이 데이터를 사용해 아래의 데이터 분석을 하시오.**
2. **Midwest 데이터를 데이터 프레임 형태로 불러온 다음 수업시간에 배운 여러 함수들을 사용해 데이터의 특징을 파악하고, 파악된 특징들을 이곳에 설명해 보시오.**

* **전체 행렬은 437, 28로 이루어져 있다. midwest에는 아시아 인구, 흑인 인구 백인 인구의 통계를 보여주며 각각의 인구들을 살펴보았을 때 백인의 인구가 가장 많고, 흑인과 아시아 인구는 매우 적은 것을 알 수 있다.**
* **Summary함수를 통해 아시아 인구는 평균적으로 1310명, 흑인 인구는 11024명, 백인 인구는 81840명으로 백인이 가장 많은 인구 수를 차지하는 것을 확인할 수 있다.**
* **아시아 인구 백분율값으로 summary를 하면 평균은 0.48 %이고, 지역의 아시아 인구 비율은 중간값(Median)인0.2972%이라는 것을 알 수 있다.**

**약 0.3%을 중심으로,0.17%에서 0.52% 사이(1분기, 3분기)에 몰려있고, 최고 값으로는 약 5% 정도의 비율을 차지하는 지역이 있다는 것을 알 수 있다.**

1. **Poptotal(전체인구) 변수를 total로, popasian(아시아 인구) 변수를 asian으로 수정하시오.**
2. **total, asian 변수를 이용해 ‘전체 인구 대비 아시아 인구 백분율’ 파생변수를 만들고 히스토그램을 만들어 도시들이 어떻게 분포하는지 살펴보고 이곳에 설명해 보시오.**

* **미국에서 아시아 인구는 대체로 적은 편이다. 미국 동북중부 437개 지역중에서 300개 지역 이상의 지역이 1%의 아시아 인이 존재하고 있음을 알 수 있다. 1%이상 아시아 인구가 존재하는 도시는 100개도 되지 않는 것을 알 수 있다.**

1. **아시아 인구 백분율 전체 평균을 구하고, 평균을 초과하면 “large”, 그 외에는 “small” 을 부여하는 파생변수를 만들어 보시오.**
2. **“large” 와 “small”에 해당하는 지역이 얼마나 되는지 빈도표와 빈도 막대 그래프를 만들어 확인해 보시오. (참고로, 빈도표는 table(midwest$group) 과 같이 생성)**