Midterm test

Data Science – 2019 Spring Semester

2018-04-16

"**midterm2019_ds.RData**" 파일은 이번 중간고사에서 사용하게 될 dataset을 저장하고 있는 RData 파일입니다. 이 파일을 R에 loading한 후 아래 task를 수행하시오. task를 수행하면서 각 질문의 **답과 답을 도출하게 된 과정 및 설명을 보고서**(문서, word, pdf, 한글, ppt 등)에 작성하여 edmodo에 제출하시오. 문제를 풀기 위해 사용한 **r script 파일**과 결과 **RData 파일**(하나의 RData 파일에 다 넣어서)도 함께 제출하시오. 제출시간이 지나면 edmodo에 제출이 불가능하게 되므로 1시간 20분이 지나기 전에 제출해야 합니다. edmodo에 제출되지 못한 답안은 채점 대상에서 제외됩니다.

Data Exploration and Preparation

"pums.sample" data frame은 2011년 미국 인구조사 데이터의 일부 sample을 저장하고 있는 데이터이다. 이 중에서 미국에서 직업을 가지고 일을 하고 있는 사람을 대상으로 조사한 여러 정보를 담고 있다. 각 변수는 다음과 같은 의미를 지닌다.

PINCP: 개인의 연 수입을 나타낸다. 단위는 US dollar

SEX: 성별을 의미한다. 1은 남성, 2는 여성

AGEP: 나이

MAR: 결혼 상태

COW: 직업의 종류

SCHL: 교육수준

WKHP: 평균 근로시간 (일주일 단위)

FER: 1년동안 자녀를 출산한 적이 있는지

JWMNP: 직장까지의 통근 시간 (단위 분)

DIS: 장애가 있는 지 여부

1. 데이터의 구조를 파악하시오. 데이터의 observation의 수와 variable의 수는 어떻게 되나요? 각 variable의 type은 어떠한가요? 각 variable의 type이 적절하게 설정되어 있는지 그렇지 않은지 설명하시오.
2. 변수 "**SEX**"는 조사 대상자의 성별을 의미한다. 1은 남성, 2는 여성을 의미한다. 의미를 명확히 하게 위해 1과 2대신 Male, Female로 값을 변경하시오. 변수의 type은 적절한 type이 무엇인지 생각하여 해당 type으로 변경하시오. 왜 그 type으로 설정했는지 설명하시오.
3. 변수 "**MAR**"는 marital status를 의미합니다. 1부터 5까지의 숫자로 표시되어 있는데, 각각의 의미를 쉽게 파악할 수 있도록 값을 변경하고 type또한 변환하시오.
   - ✓ 1 .Married
   - ✓ 2 .Widowed
   - ✓ 3 .Divorced
   - ✓ 4 .Separated
   - ✓ 5 .Never married or under 15 years old
4. 각 변수의 Missing Value의 수는 어떻게 되는가? Missing Value가 없는 변수는 어떤 변수이며, Missing Value가 있는 변수는 전체의 몇 %가 Missing 되어 있는지 비율을 계산하시오.
5. 변수 "**FER**"은 과거 12개월 동안 출산여부를 나타낸다. 이 변수의 Missing value는 남성이거나 나이가 15세 미만 또는 50세 초과인 경우 나타난다고 한다. 모든 남성에 대해 FER이 Missing인지 확인하시오. 또한 여성의 경우 15세 미만인 경우 모두 Missing인지 확인하시오. 또 50세 초과 여성이 모두 Missing인지 확인하시오. 여기서 언급된 경우 외에 Missing Value가 발생했는지 했다면 얼마나 발생했는지 확인하시오.
6. 각 변수의 outlier가 존재하는지 확인하시오. 발견한 outlier 중 명백한 오류가 있는지 확인하고, 왜 그 outlier가 명백한 오류인지(혹은 명백한 오류가 아닌지) 설명하시오.
7. 교육수준(SCHL)과 종사업종(COW)를 비교해보시오. 교육수준에 따라 종사업종의 종류나 비율이 어떻게 달라지는지 (혹은 달라지지 않는지) 확인하고 설명하시오.
8. 나이대별 소득을 비교해보시오. 20대, 30대, 40대, 50대, 60대 이상, 5개의 그룹으로 나누어 각 나이 그룹별 평균 소득을 비교해보시오. 가장 소득이 높은 그룹과 가장 소득이 낮은 그룹은 어디인가요?

9. 근무시간(WKHP)과 소득(PINCP)과 어떤 관계가 있는지 그래프를 그려 확인해보시오. 그래프에 관계가 나타난다면 그것을 설명해보시오.

10. 결과 data.frame을 RData 파일로 저장해서 함께 제출하시오. (save 명령 사용)

Tidy Data

"**automobile.long**" data.frame은 396개의 자동차에 대한 여러 정보를 포함하고 있는 dataset이다. 다음 질문에 답하거나 task를 수행하라.

1. 데이터에 표현된 observation은 각각 하나의 자동차 종류를 나타낸다고 한다. 또한 변수는 자동차의 여러 특징(자동차 이름, 연식, 연비, 무게 등)을 표현한다고 하자. 그렇다면 이 데이터는 tidy 데이터인가 그렇지 않은가? 이유를 설명하시오.

2. 1번의 답이 tidy가 아니라고 한다면, tidy 형태로 변환하시오. 결과는 RData에 저장(save 명령 사용)하여 제출하시오.