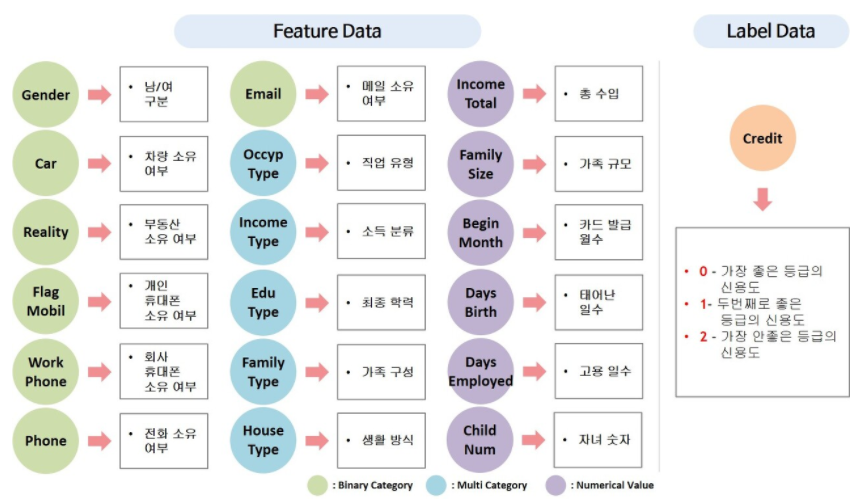
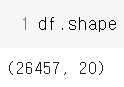
신소율과 권현빈과 아이들2

**임상훈 윤영주**

**조은서 채윤길**

<data.columns>





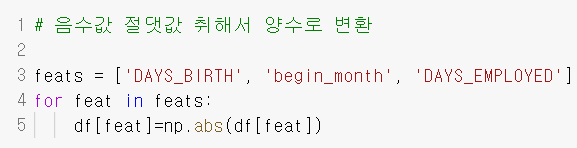
<중복 데이터>

* ‘카드 발급 월 수’ 변수를 제외한 중복 데이터 14358개

=> 모두 제거하는 대신 ‘개설 카드 개수’ 변수 추가

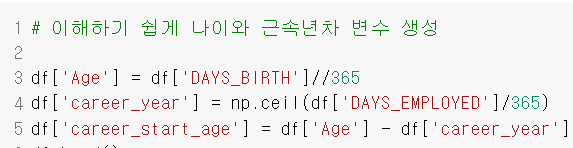
**<전처리>**

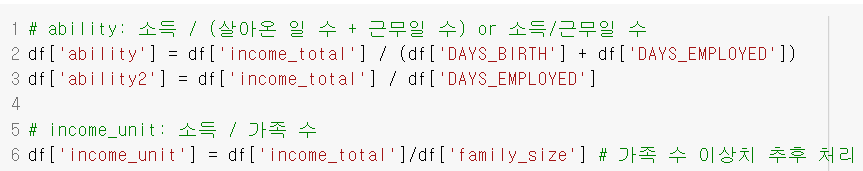
1. 음수값 변수인 ‘DAYS\_BIRTH’, ‘begin\_month’, ‘DAYS\_EMPLOYED’ 양수로 변환



1. 파생 변수 생성

* 나이, 근속년차, ability, income\_unit 변수 생성



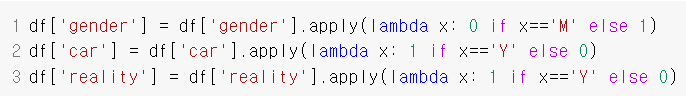


1. 이상치 처리

* Widow의 gender가 M인 경우 -> 이상치로 고려하여 F로 바꿈

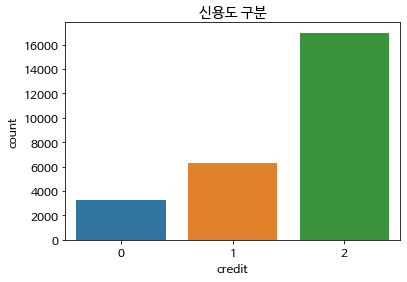


1. 범주형 변수 수치화하기





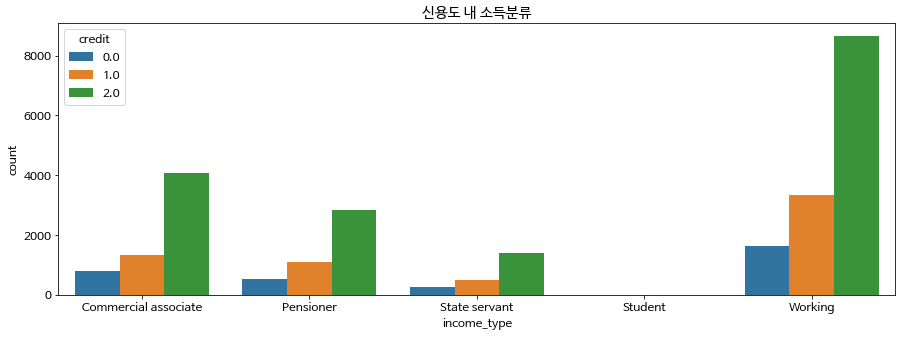
<성별, 신용도 분포>



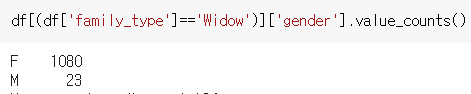
<credit에 따른 income type 분포>

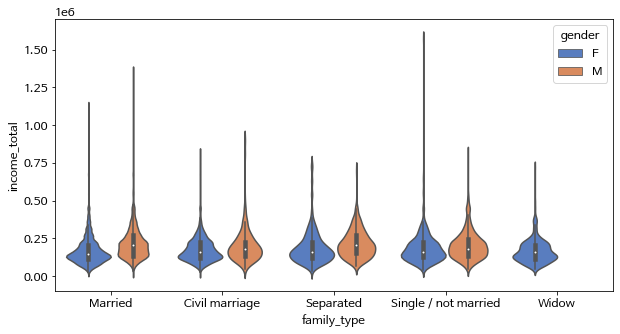


<income type에 따른 credit 분포>

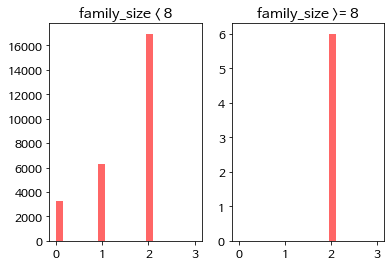
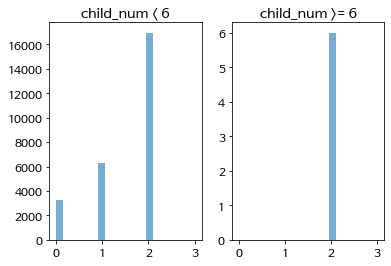


<family type 별 income total>

* widow(과부)인데 gender가 male인 이상치가 23명 존재했음 제거 필요



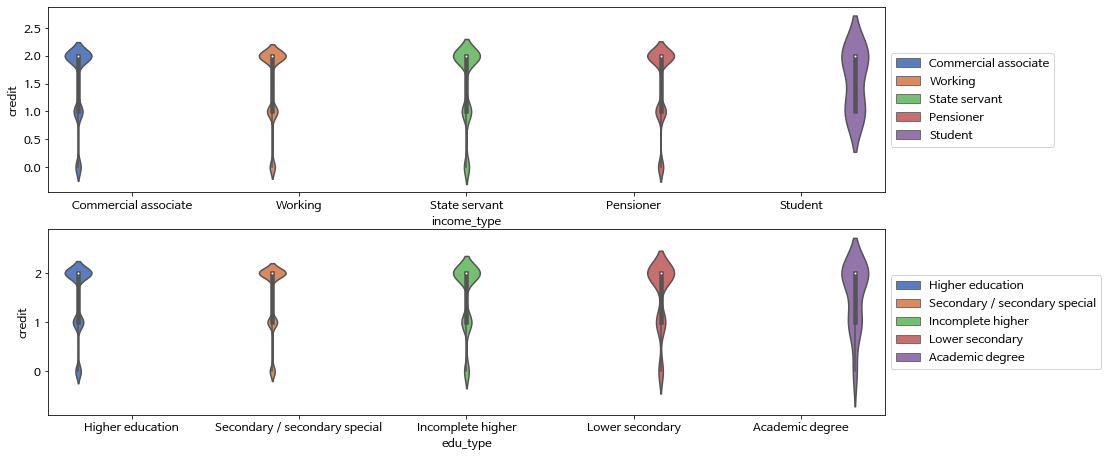
<child num, family size 에 따른 credit 분포>



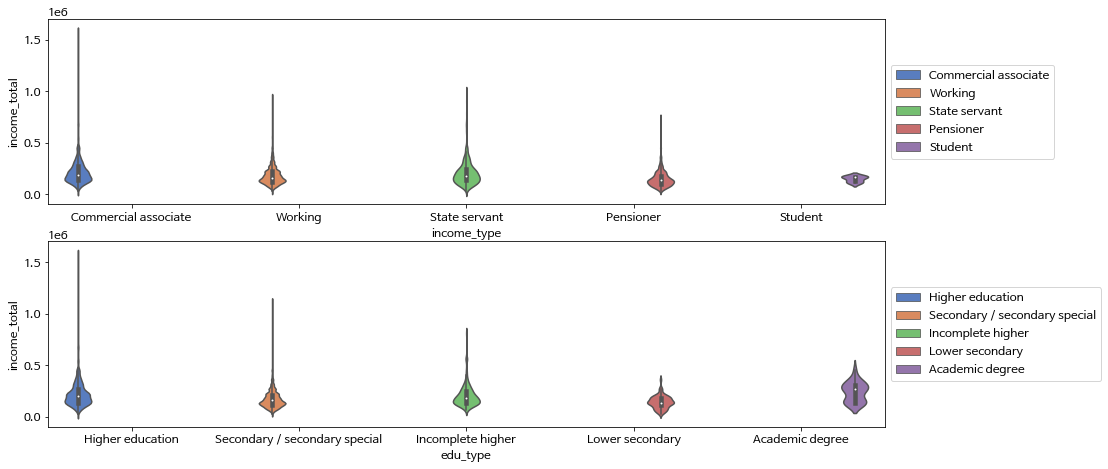
- child\_num이 6명 이상인 경우 무조건 신용도 2

- family\_size가 8명 이상인 경우 무조건 신용도 2

**< income type, edu type에 따른 credit 분포>**

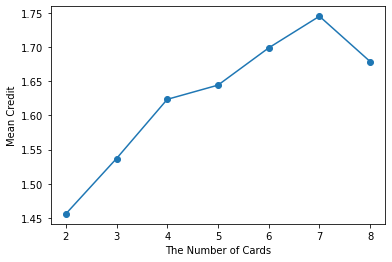


**< income type, edu type에 따른 income total 분포>**



- income\_type의 Student와 Edu\_type의 Academic 값이 다른 값에 비해 현저하게 낮은 수치를 보여 유사 카테고리와 합치려 하였으나 분포가 크게 달라 합치지 못했음

**<카드 개수에 따른 신용도>**

****