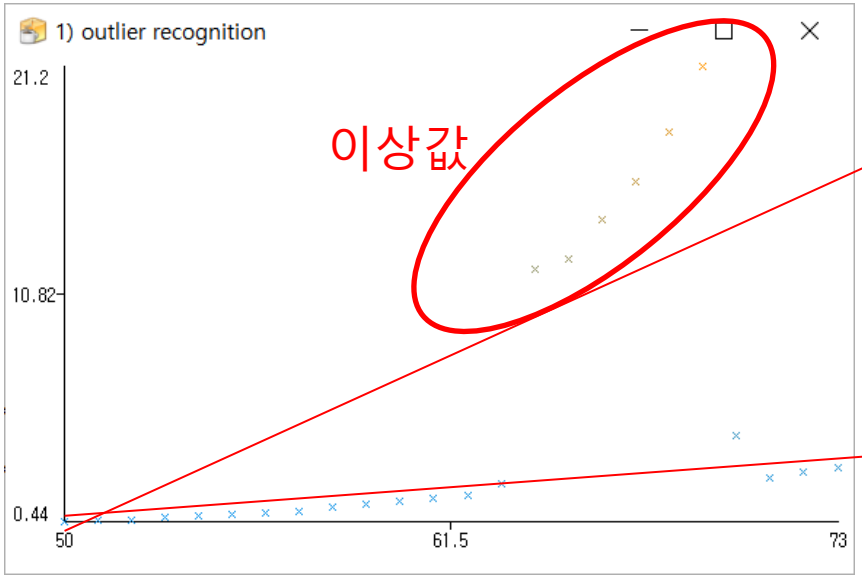
이상값/결측값 처리

Why : 불완전한 데이터세트의 클린징

1. 이상값 처리 결과

- 이상값 포함할 때 회귀식 + 이상값 제외할 때 회귀식 산출
- 회귀식 산출과 상관계수 해석보다는 **상황에 맞는 회귀식 선택**이 중요



회귀식 (Q2,Q4)
classification =
0.5041 * year -26.0059

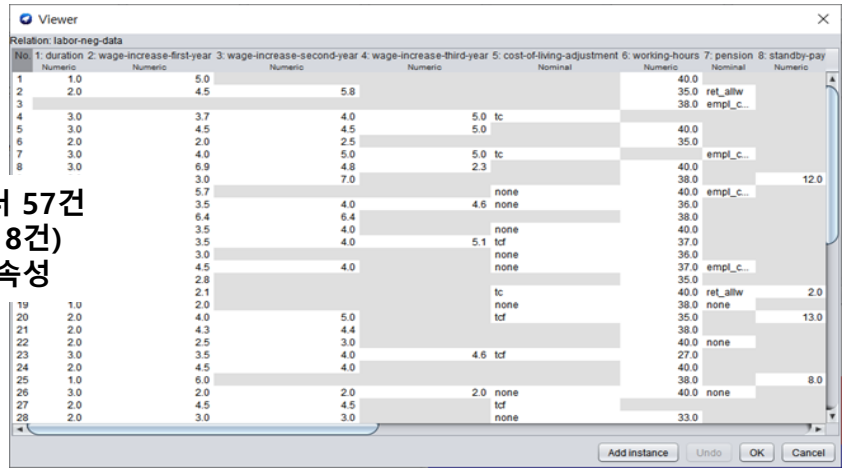
회귀식 (Q5,Q6)
phone calls =
0.1096 * year -5.2386

적용방법 / 회귀식	알고리즘	필터/전처리	상관계수
phone calls =0.5041 * year - 26.0059	LinearRegression	X	0.4452
classification=0.5041 * year -26.0059	LeastMedSq	AddClassification	1
phone calls = 0.1096 * year -5.2386	LeastMedSq	X	0.5447
phone calls = 0.1096 * year -5.2386	LinearRegression	수동삭제	0.9894

2. 결측값

- 결측값 포함 J48 + 33% 결측속성 삭제 J48 + 대체 J48 (평균/임의)
- 정분류율/트리복잡도 기준 선택가능하나 **상황에 맞는 J48 선택**

총 데이터 57건
(33%:18건)
17개 속성



labor.arff 데이터세트에 J48 적용결과

학습방법 \ 지표	결측율	정분류율	트리 복잡도
1) No filtered , No deleted (결측값 있는 그대로 학습)	33.64%	74.60%	고
3) 10 attributes deleted (결측율 33% 이상인 10개 속성 삭제)	7.27%	80.95%	중
4) ReplaceMissingValue filtered (결측값을 평균값으로 대체)	0.0%	80.95%	하