2021 제 1회 데이터과학부 데이터분석 경진대회

# 통계적 분석과 클러스터링을 활용한 농업 손상 패턴 분석

18016019 김수현
19016002 고가연
19016070 이수현

# 0. 목차

# 1. 서론

## ① 농업이 타 산업에 비해 재해 손상율이 2배 높음

- 고용노동부의 산업재해 현황에 따르면 농업인의 **농작업 관련 재해율은 13.8로 전체 산업 재해율 6.9에 비해 약 2배** 높음
- 농촌에서 **유독 재해율이 높은 상황에 대한 정확하고 면밀한 분석이 필요**함
- 출처 : 최근 5년 농업인의 비농업인의 손상률(2020)

## ② 농작업에 대한 손상은 위험의 범위가 광범위함

- 농작업은 일반적인 직업적 손상과 달리 **날씨, 기온, 작업환경 등으로 인한 다양한 변수들**로부터 영향을 받을 수 있는 넓은 야외 공간에서 작업함
- 손상유형에서 **미끄러짐 및 넘어짐, 과도한 힘·동작**에 의한 손상 발생형태가 많으며, 다른 직종에 비해 **위험한 도구나 기계, 화재/전기충격** 등에 노출 위험도 높음
- **근무장소**가 불결하거나 **불편한 자세로 장시간** 일하거나 중량물을 들거나 옮기는 일이 많은 경우가 **손상 발생 위험이 2배** 정도 높음
- 출처 : 우리나라 농업인 손상(2012), 최근 5년 농업인과 비농업인의 손상률(2020)

## ③ 법적인 '근로자'로서의 지위가 없는 농업인

- 농업인 대다수는 법적인 '근로자'가 아니기 때문에 **'산업재해보상보험법'에 따른 보상과 보호의 대상이 되지 못하며** 사회적 관심도 덜한 편임
- 5인 이상의 근로자를 고용하고 있는 사업장에 대해서만 산재보험 의무가입 대상이기 때문에 **농업인의 안전보험의 가입률은 2019년 기준 63.1%**
- 농업인 안전보험의 **급여수준** 역시 산재보험보다 전반적으로 **낮으며** 1년마다 재가입해야 하는 방식도 **보장기간이나 보험가입심사 등과 관련하여 농가에 불리한 요소**가 되고 있음
- 출처 : 농업인 안전보험의 개선 필요성과 향후 과제(2021)

# 1.2 데이터 소개

## 경진대회용데이터.xlsx

**조사기관** 농촌진흥청 국립농업과학원

**조사대상** 전국의 표본 농가 12,000 가구, 19세 이상의 농업인

**주요내용** 농업 관련 손상에 대한 질문

row_id (object) : 인터뷰 대상자 (nrow = 17770)
column_names (feature) : 설문조사 내용 (ncol = 103)

**문항종류**

1. 조사가구의 농업활동 : 주요 농업 종류 등 4항목
2. 가구원 특성 : 가구원 연령 및 성별 등 6항목
3. 가구원의 농업 안전 활동 : 농작업 위험 의식 등 13항목
4. 업무상 손상 : 농작업 관련 손상발생 월, 날씨 등 11항목

| 1 | id | 시도명 | 시군구명 | 읍면동명 | 행정리명 | 조사가구번호 | q1_1 | q1_2 | q2_1_1 | q2_1_2 | q2_2_1 | q2_2_2 | q2_3_1 | q2_3_2 | q2_4_1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 1 | 인천광역시 | 강화군 | 선원면 | 냉정리 | 1 | 고추 | 2-3 | 2 | | 1 | 2645 | 2 | | 2 |
| 3 | 2 | 인천광역시 | 강화군 | 선원면 | 냉정리 | 2 | 마늘 | 2-13 | 2 | | 1 | 1983 | 2 | | 2 |
| 4 | 3 | 인천광역시 | 강화군 | 선원면 | 냉정리 | 3 | 소 | 5-6 | 2 | | 2 | | 2 | | 2 |
| 5 | 4 | 인천광역시 | 강화군 | 선원면 | 냉정리 | 4 | 고구마 | 2-2 | 2 | | 1 | 992 | 2 | | 2 |
| 6 | 5 | 인천광역시 | 강화군 | 선원면 | 냉정리 | 5 | 고추 | 2-3 | 2 | | 1 | 1653 | 2 | | 2 |
| 7 | 6 | 인천광역시 | 강화군 | 선원면 | 냉정리 | 6 | 고추 | 2-3 | 2 | | 1 | 1818 | 2 | | 2 |
| 8 | 7 | 인천광역시 | 강화군 | 선원면 | 냉정리 | 6 | 고추 | 2-3 | 2 | | 1 | 1818 | 2 | | 2 |
| 9 | 8 | 인천광역시 | 강화군 | 선원면 | 냉정리 | 7 | 옥수수 | 2-25 | 2 | | 1 | 992 | 2 | | 2 |
| 10 | 9 | 인천광역시 | 강화군 | 선원면 | 냉정리 | 8 | 고구마 | 2-2 | 2 | | 1 | 3306 | 2 | | 2 |
| 11 | 10 | 인천광역시 | 강화군 | 선원면 | 냉정리 | 9 | 고추 | 2-3 | 2 | | 1 | 991 | 2 | | 2 |
| 12 | 11 | 인천광역시 | 강화군 | 선원면 | 냉정리 | 10 | 벼 | 1-1 | 1 | 9917 | 1 | 1653 | 2 | | 2 |
| 13 | 12 | 인천광역시 | 강화군 | 선원면 | 신정리 | 1 | 고추 | 2-3 | 2 | | 1 | 3306 | 2 | | 2 |
| 14 | 13 | 인천광역시 | 강화군 | 선원면 | 신정리 | 2 | 대파 | 2-7 | 2 | | 1 | 992 | 2 | | 2 |
| 15 | 14 | 인천광역시 | 강화군 | 선원면 | 신정리 | 3 | 고추 | 2-3 | 2 | | 1 | 1653 | 2 | | 2 |
| 16 | 15 | 인천광역시 | 강화군 | 선원면 | 신정리 | 4 | 벼 | 1-1 | 1 | 33058 | 2 | | 2 | | 2 |
| 17 | 16 | 인천광역시 | 강화군 | 선원면 | 신정리 | 5 | 벼 | 1-1 | 1 | 8926 | 1 | 1653 | 2 | | 2 |
| 18 | 17 | 인천광역시 | 강화군 | 선원면 | 신정리 | 6 | 고추 | 2-3 | 2 | | 1 | 992 | 2 | | 2 |
| 19 | 18 | 인천광역시 | 강화군 | 선원면 | 신정리 | 7 | 벼 | 1-1 | 1 | 2810 | 2 | | 2 | | 2 |

2021년농업인의업무상질병및손상조사수집자료 Sheet1

## 데이터 가공

**여러 컬럼으로 나눠져 있는 질문에 대해서 하나의 컬럼으로 병합**

q2의 경우, q2_1_1 ~ a2_6_1으로 1, 2로 표현된 데이터로 나눠져 존재 👉 하나의 컬럼으로 병합

| 농업 종류 | 수행 여부 | | 규 모 |
|---|---|---|---|
| | O (1) | X (2) | |
| 1) 논농사 q2_1_1 | ☐ | ☐ | _____q2_1_2_____ $m^2$ |
| 2) 밭농사(노지) q2_2_1 | ☐ | ☐ | _____q2_2_2_____ $m^2$ |
| 3) 과수원 q2_3_1 | ☐ | ☐ | _____q2_3_2_____ $m^2$ |
| 4) 시설(하우스) q2_4_1 | ☐ | ☐ | _____q2_4_2_____ $m^2$ |
| 5) 축산 q2_5_1 | ☐ | ☐ | ① 소 _____q2_5_2_1_____ 마리<br>② 돼지_____q2_5_2_2_____ 마리<br>③ 닭_____q2_5_2_3_____ 마리<br>④ 기타_____q2_5_2_4_____ 마리 |
| 6) 기타 q2_6_1 | ☐ | ☐ | |

| 농업 종류 | 규모 |
|---|---|
| 0 : 논 | 논 규모 |
| 1 : 밭 | 밭 규모 |
| 2 : 과수원 | 과수원 규모 |
| 3 : 시설(하우스) | 시설(하우스) 규모 |
| 4 : 축산 | 축산 규모 |
| 5 : 기타 | 기타 규모 |

**나이 범주화** - 세분화되어 있는 나이를 연령대별로 범주화 (20, 30, 40, 50, 60, 70, 80, 90)

**필요 없는 컬럼 제거** - 컬럼id('id'), 작목코드(q1_2), 알수없음( 're1','q1_1re','age','age2')

## 데이터 분리 – 의식df 생성

주어진 데이터에서 조사대상자의 안전의식을 알아볼 수 있는 질문을 뽑아내 따로 의식df 생성

**STEP 1**

조사대상자의 안전의식을 파악할 수 있는 질문

((1년간 일출 전/일몰 후에 농기계 사용빈도 등) 19개) 추출

['q20', 'q22', 'q23', 'q24', 'q27_1', 'q27_3', 'q27_4', 'q27_6', 'q27_7', 'q27_8', 'q27_9', 'q27_10', 'q30', 'q30_1', 'q31_1', 'q31_2', 'q31_3', 'q31_4', 'q31_5']

**STEP 2**

안전의식 높다고 판단할 수 있을수록 높은 점수로

(의식 높을 수록 내림차순) 통일

['q20', 'q22', 'q30', 'q30_1']

| | q20 | q22 | q23 | q24 | q27_1 | ... | q31_1 | q31_2 | q31_3 | q31_4 | q31_5 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 4 | 3 | 1 | 1 | 2 | ... | 1 | 1 | 1 | 1 | 1 |
| **1** | 3 | 3 | 1 | 1 | 4 | ... | 3 | 3 | 3 | 3 | 3 |
| **2** | 3 | 3 | 1 | 1 | 3 | ... | 3 | 1 | 3 | 1 | 1 |
| **3** | 3 | 3 | 4 | 1 | 4 | ... | 3 | 1 | 3 | 1 | 1 |
| **4** | 4 | 4 | 4 | 1 | 4 | ... | 1 | 1 | 1 | 1 | 1 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **15982** | 3 | 3 | 1 | 1 | 4 | ... | 1 | 1 | 1 | 1 | 1 |
| **15983** | 1 | 3 | 1 | 1 | 1 | ... | 1 | 1 | 1 | 1 | 1 |
| **15984** | 3 | 4 | 1 | 1 | 4 | ... | 1 | 1 | 1 | 1 | 1 |
| **15985** | 3 | 4 | 1 | 1 | 4 | ... | 1 | 1 | 1 | 1 | 1 |
| **15986** | 4 | 3 | 1 | 1 | 4 | ... | 1 | 1 | 1 | 1 | 1 |

15987 rows × 19 columns

# 2. 본론

2.1 안전의식과 손상의 상관관계

2.2 손상자 및 비손상자 분석

2.3 손상자에 대한 패턴 분석

초기 가설 : **안전의식**이 **낮을수록** **손상율**이 **높을 것**이다

( = **안전의식이 손상경험에 선행**한다고 가정 )

# 2.1 안전의식과 손상의 상관관계

가설 증명을 위해 앞서 생성한 의식df에서 손상자와 비손상자간 안전의식 점수의 차이가 분명한 문항을 선별한다.

## 의식df에서 유의미한 문항 재선별 -> 'new의식df'

### STEP 1 의식df의 선지 구성 파악

ⅰ) 1~4점 문항 (3개) : q20, q22, q30

ⅱ) 1~5점 문항 (10개) : q23, q24, q27_1, q27_3, q27_4, q27_6, q27_7, q27_8, q27_9, q27_10

ⅲ) 1, 3점 문항 (5개) : q31_1, q31_2, q31_3, q31_4, q31_5

ⅳ) nan값이 있는 문항 (1개) : q30_1 -> 제외

### STEP 2 문항별 손상자, 비손상자의 안전의식평균점수 비율합의 차이 계산

Ex) q20에 대한 손상자와 비손상자의 안전의식 점수 비율

| | 손상q20 | 비손상q20 |
|---|---|---|
| 1 | 3.366337 | 9.197799 |
| 2 | 14.653465 | 25.305531 |
| 3 | 38.217822 | 36.235158 |
| 4 | 43.762376 | 29.261512 |

손상자 - 1,2점 비율의 합 : 18.01
비손상자 – 1,2점 비율의 합 : 34.5

→ 16.49점 차이
(낮은 점수의 비율합 : 비손상자 > 손상자)

손상자 – 3,4점 비율의 합 : 81.98
비손상자 – 3,4점 비율의 합 : 65.49

→ 16.49점 차이
(높은 점수의 비율합 : 비손상자 < 손상자)

∴ 손상자가 안전의식이 더 높다는 것을 의미한다.

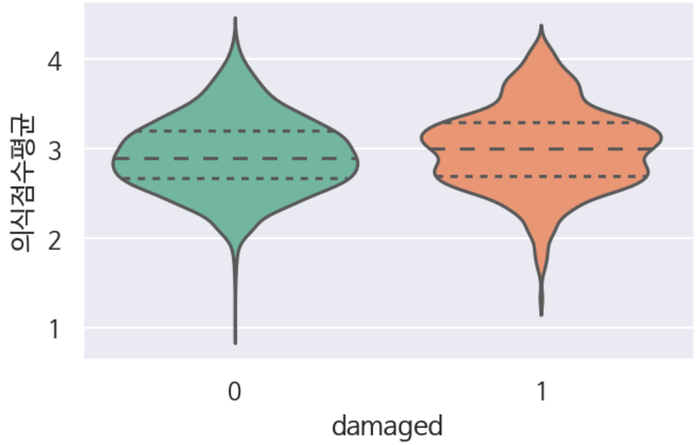
**STEP 3** 점수 비율합 차이가 큰 순서대로 10개 문항 선정

의식df에서 이용할 10개 feature 선정 후 'new의식df'로 저장

| 문항 | 점수 비율합 차이 |
|---|---|
| Q20 | 16.49 |
| Q22 | 3.21 |
| Q30 | 2.03 |
| Q23 | 4.41 |
| Q24 | 5.9 |
| … | … |

→ 상위 10개 컬럼 : 'q20', 'q23', 'q24', 'q27_7','q27_8', 'q27_9', 'q30_1', 'q31_1', 'q31_2', 'q31_3'

## new의식df의 손상여부별 의식점수평균의 유의성 검정




1. 모수적 검정
- Two Sample t-test

2. 비모수적 검정
- Wilcoxon rank sum test
- Mann-Whitney U test

```
Ttest_indResult :  0.00054998
RanksumsResult :  0.00040025
MannwhitneyuResult :  0.0003944
```

→ 유의확률이 매우 작음

∴ 손상과 비손상집단의 의식점수평균의 차이가 유의하다.

초기 가설 : **안전의식**이 **낮을수록** **손상율**이 **높을 것**이다

( = **안전의식이 손상경험에 선행**한다고 가정 )

데이터를 통해 가설과는 반대의 결과를 확인함

## **안전의식**과 **손상율**이 **역상관관계**를 보인다

안전의식과 손상의 인과관계에 대해 다시 생각해보면,

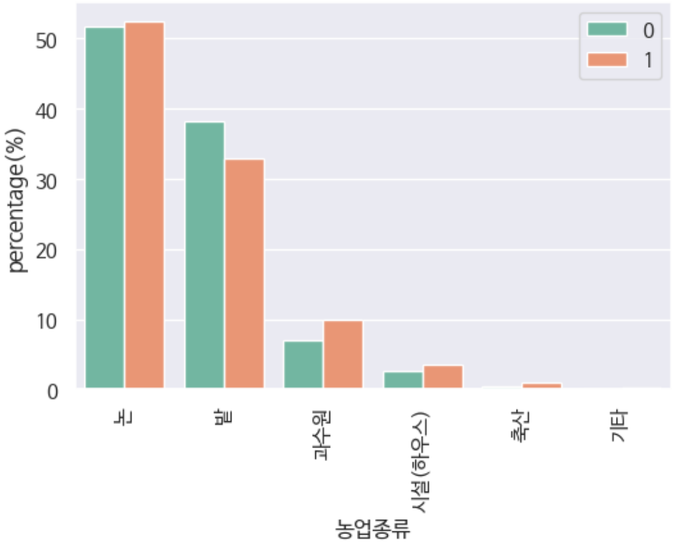
**손상경험이 안전의식에 선행**한다고 볼 수 있다.

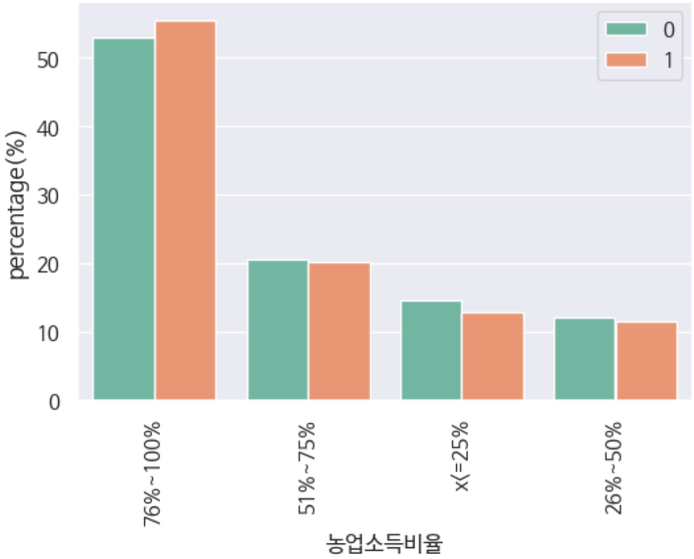
**손상경험을 이미 했기 때문에 안전에 대한 경각심이 생긴 것**으로 해석할 수 있다.

# 2.2 손상자 및 비손상자 분석

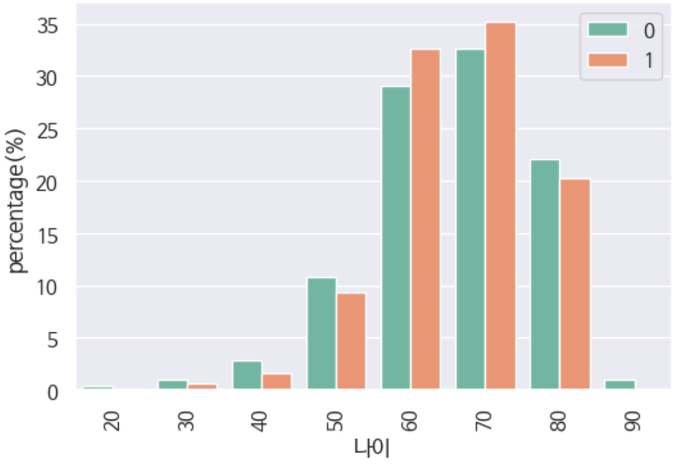
## 손상자/비손상자 분석

먼저, 손상자와 비손상자가 어떠한 피처에서 차이가 두드러지게 나타나는지 알아보기 위해 각 피처별로 손상자/비손상자 비율을 시각화해서 분석했다.

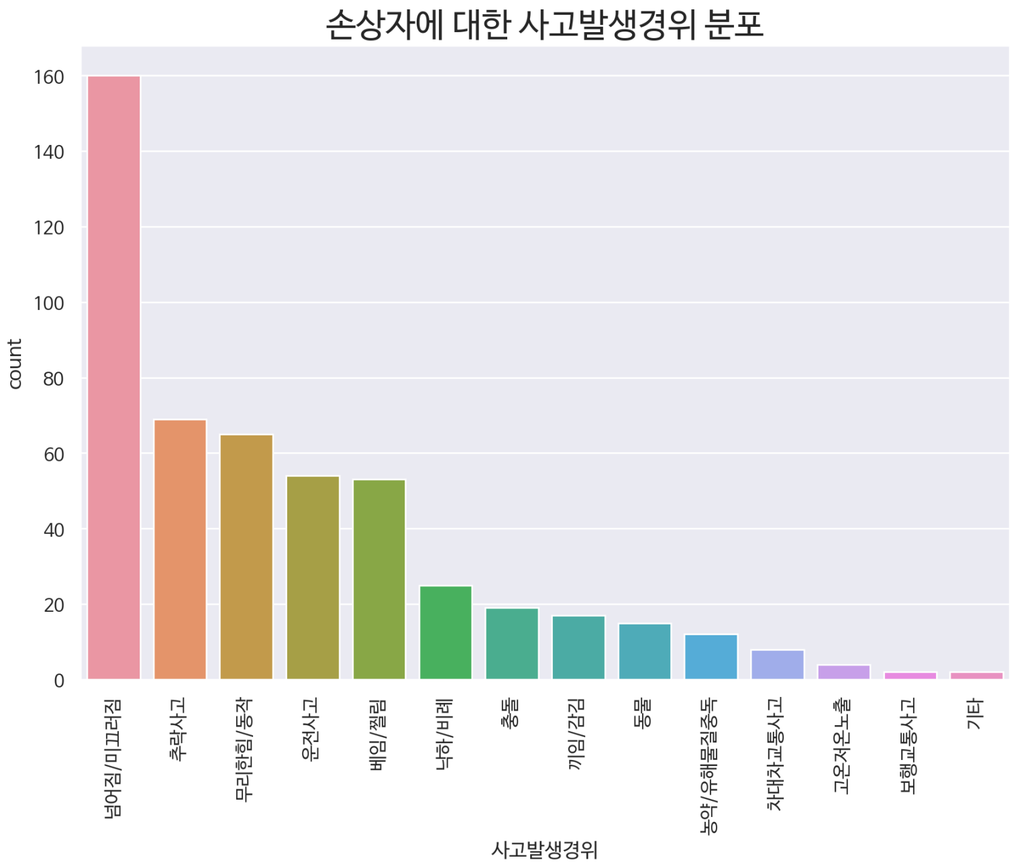








- ✓ 타지역에 비해 경상남도, 충청북도에서 손상을 더 입는 것으로 보인다.
- ✓ **농업종류**에서 두 그룹을 비교했을 때, 밭에서는 비손상자의 비율이 더 높지만 논과 과수원, 시설하우스에서 손상자의 비율이 더 높다.
- ✓ **농업소득비율**에서는 농업소득이 76%이상일 때 손상을 더 입는 것으로 보인다.
- ✓ **성별**에서는 남성이 여성보다 손상을 더 입는 것으로 보인다.
- ✓ **나이**에서는 손상자의 비율이 60-70대에서 더 높은 것으로 보인다.

## Q 어떤 사고가 어떤 농업에서 가장 많이 발생하는가?




1. 미끄러짐 : 논,밭,시설
2. 추락 : 과수원, 시설
3. 무리한 힘/동작 : 시설
4. 밭 베임/찔림 : 시설,과수원,밭

- 미끄러지는 사고를 당한 농업인의 **강제 휴식기간은 30~180일이 50%**를 차지하고 있으며 회복정도는 수행 능력이 약간에서 많이 떨어진다고 답하였다.
- **90%이상이 치료를 받은 경험**이 있는 걸로 보아 미끄러지는 사고를 당하면 농업활동에 지장을 많이 주는 것으로 판단된다.

→ 더 자세히 살펴보기 위해서 다양한 피처별로 클러스터링 분석을 수행한다.

# 2.3 손상자에 대한 패턴 분석

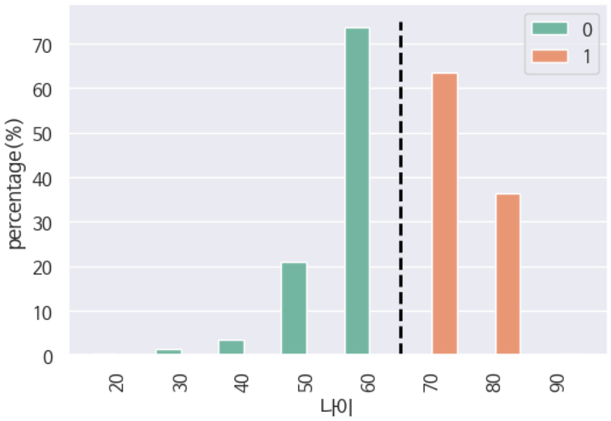
## 손상자 패턴 분석

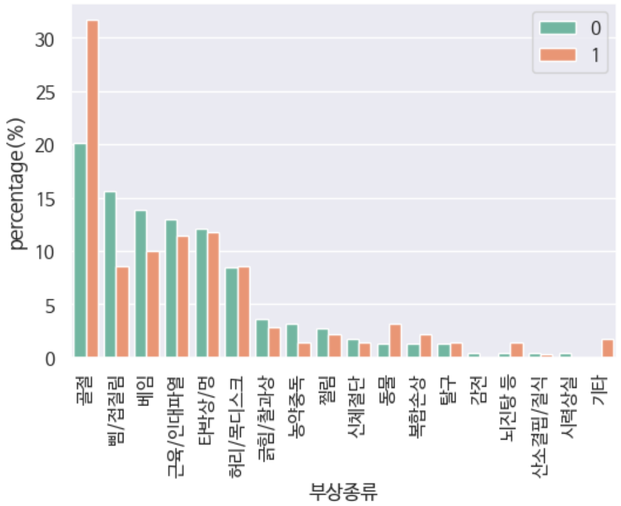
다양한 변수를 사용해서 손상자의 패턴을 분석하기 위해 **클러스터링 분석**을 수행한다.

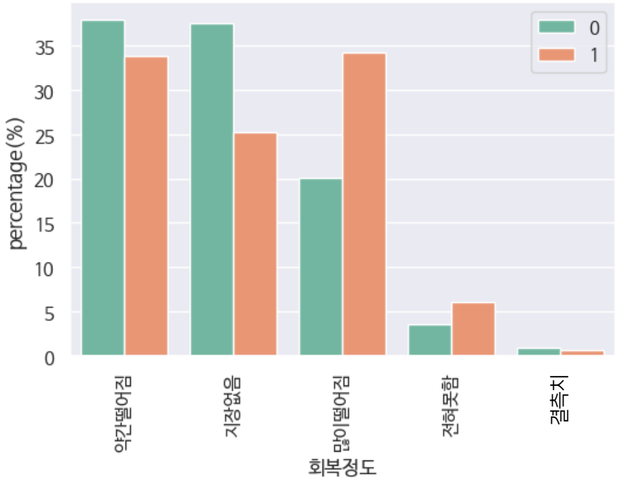
**클러스터 개수를 2개로 지정**해 각 변수에 대해 **두 개의 그룹으로 분리**, 각 그룹에 대한 패턴을 분석하였다.

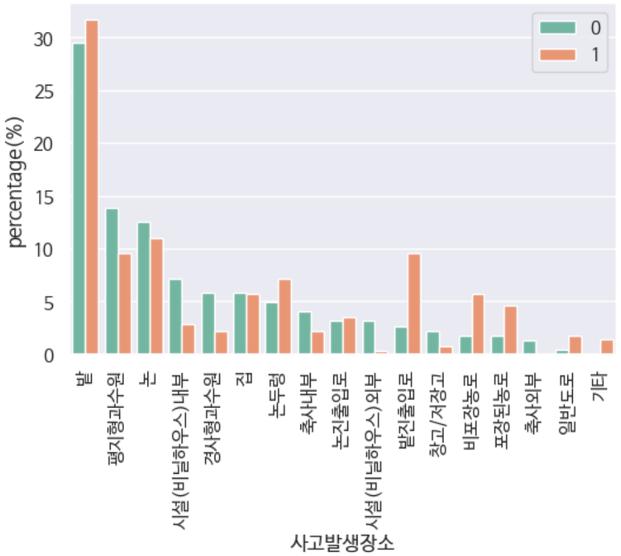
## ① 연령대가 높은 그룹, 낮은 그룹

{Cluster : 데이터 개수} = { 0 : 281, 1 : 224 }  **0** : 연령대 낮은 그룹, **1** : 연령대 높은 그룹



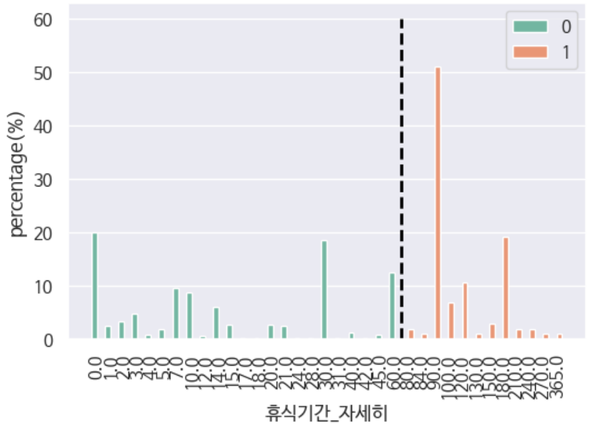

- ✓ 나이 60대 전후로 그룹이 나눠졌다.
- ✓ 연령대가 높은 그룹의 **부상종류**는 골절이 가장 많다.

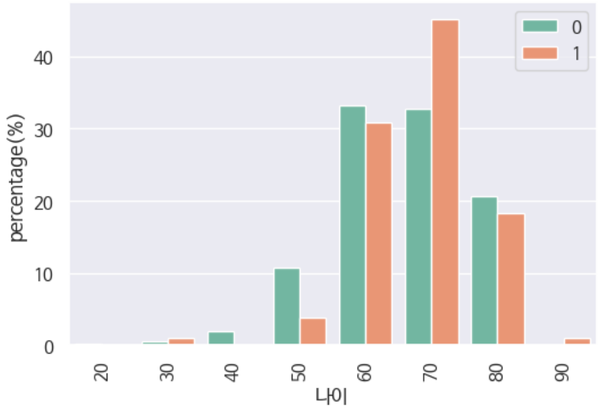
- ✓ 연령대가 높은 그룹은 낮은 그룹보다 **사고 후 회복**이 오래 걸린다.
- ✓ **사고발생 장소**는 두 그룹 비슷하나 밭진출입로, 농로 등에서는 연령대가 높은 그룹의 비율이 더 높다.

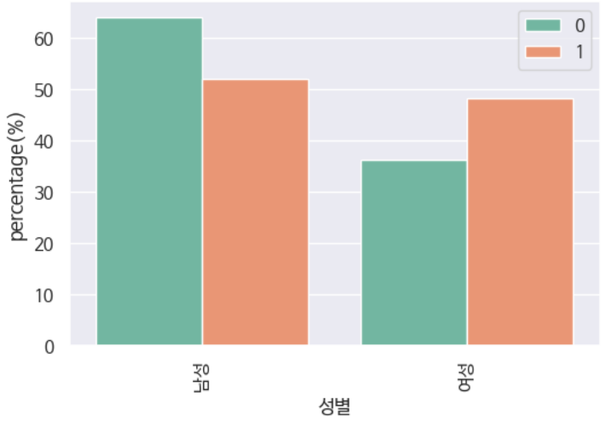
# 2.3 손상자에 대한 패턴 분석

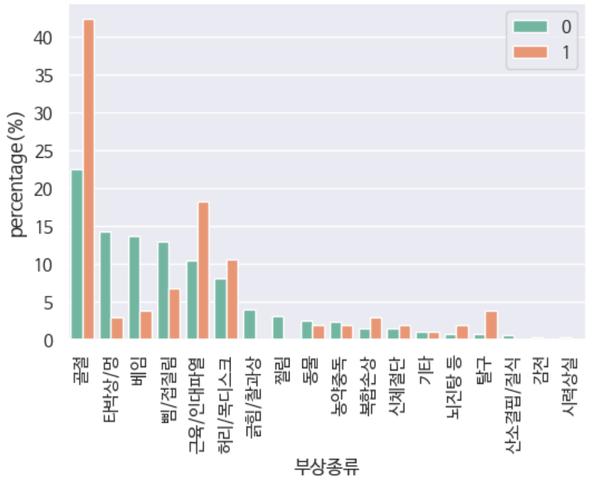
## ② 손상 후 휴식기간이 긴 그룹, 짧은 그룹

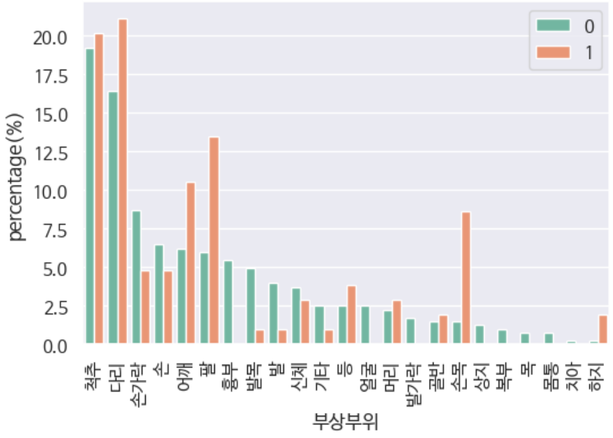
{Cluster : 데이터 개수} = { 0 : 401, 1 : 104 } **0** : 휴식기간 짧은 그룹, **1** : 휴식기간 긴 그룹

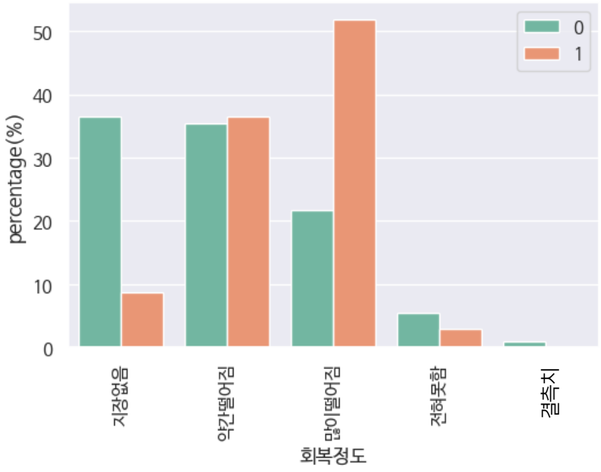


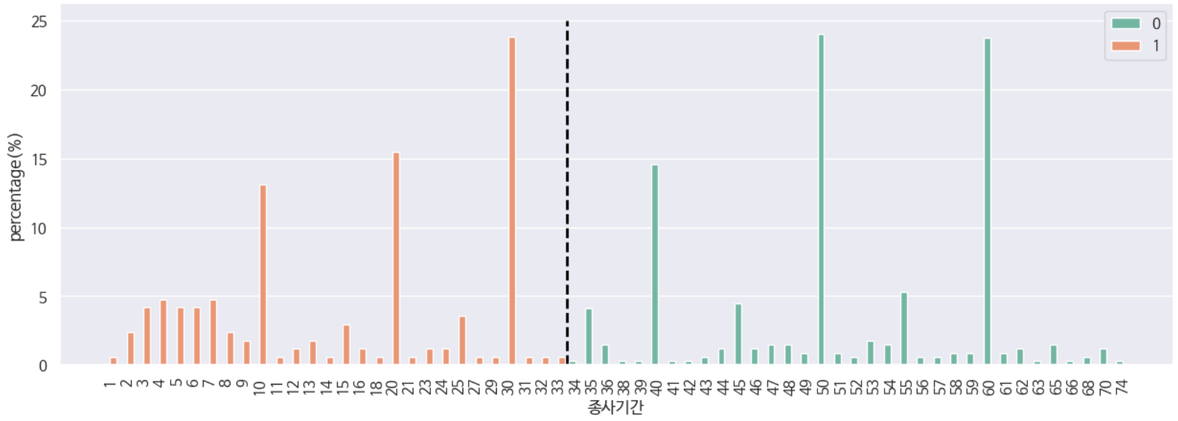



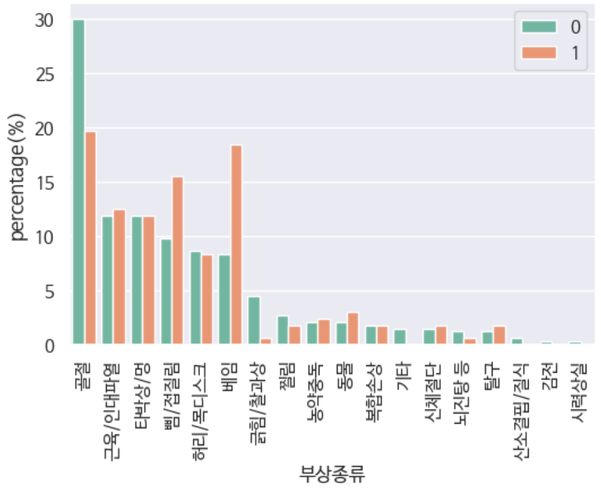

- ✓ 휴식기간 60일 전후로 그룹이 나눠졌다.
- ✓ **나이** 70대에서는 손상 후 휴식기간이 긴 그룹의 비율이 더 많다.
- ✓ **여성**의 경우 부상 시 휴식기간이 긴 경우가 더 많다.
- ✓ **부상종류**에서 두 그룹을 비교했을 때, 골절과 근육/인대파열에서 휴식기간이 더 길고 타박상/멍과 베임 등에서는 휴식기간이 비교적 짧다.
- ✓ **부상부위**에서는 어깨와 팔, 손목을 부상당했을 때 휴식기간이 더 길고 흉부, 발목 등의 부상의 경우는 휴식기간이 비교적 짧다.
- ✓ 휴식기간이 길면, **회복 정도**가 더 떨어진다.

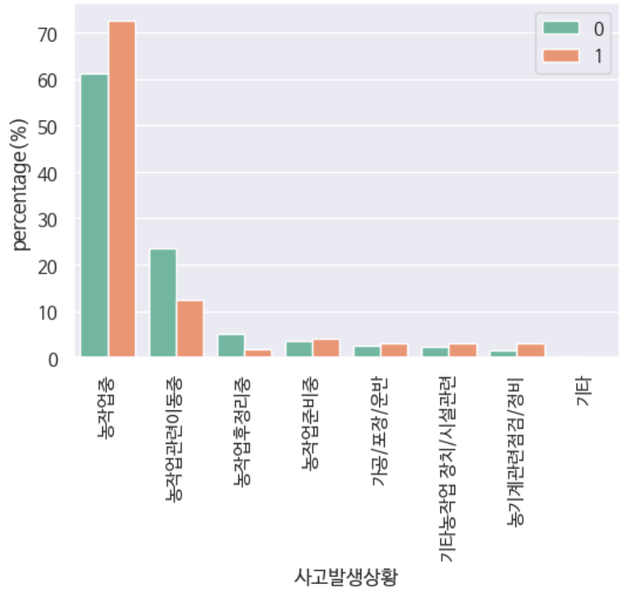
# 2.3 손상자에 대한 패턴 분석

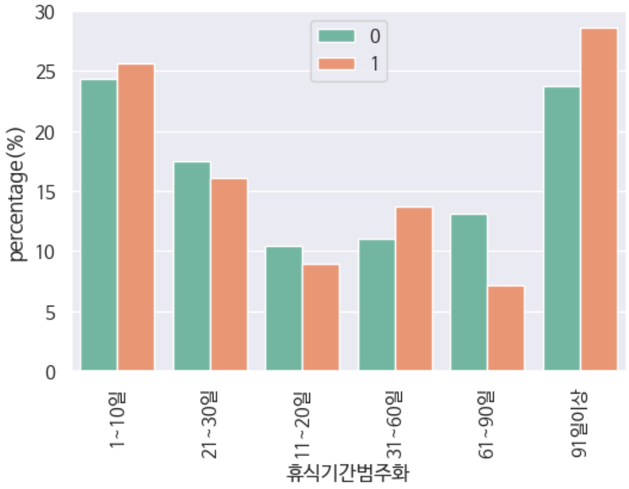
## ③ 종사기간이 긴 그룹, 짧은 그룹

{Cluster : 데이터 개수} = { 0 : 337, 1 : 168 } **0** : 종사기간 긴 그룹, **1** : 종사기간 짧은 그룹




- ✓ 종사기간 34년 전후로 그룹이 나눠졌다.
- ✓ **부상종류**에서 두 그룹을 비교했을 때, 가장 많은 부상종류인 골절에서는 종사기간이 긴 그룹의 비율이 더 높고, 삠/접질림과 베임에서는 종사기간이 짧은 그룹의 비율이 더 높다.
- ✓ **사고 발생 상황**에서는 농작업 중에는 종사기간이 짧은 그룹의 비율이 더 높고, 농업관련 이동중에는 종사기간이 긴 그룹의 비율이 더 높다.

# 3. 결론

# 3. 결론

## 요약 및 기대효과

- 전국의 표본 농가 12,000 가구(19세 이상의 농업인 17,770명)을 대상으로 실시한 조사 데이터를 이용한 분석에서 **통계적 가설검증**을 통해 농업인의 **안전의식과 손상률이 역상관관계**를 나타낸다는 결론을 도출했다. 또한 **비손상자 및 손상자 패턴을 시각화, 클러스터링**을 통해 분석하여 각 특성마다 **손상과의 관계를 도출**할 수 있었다.
- 농촌손상자들은 다른 직업적 손상과 달리 광범위한 위험요소에 노출되어있다. 이러한 **광범위한 위험요인을 손상자 패턴 분석을 통해 빈번한 특정 유형으로 좁혀**서 보다 체계적이고 효율적으로 손상율을 줄일 수 있을 것으로 기대된다.
- 손상을 이미 겪어본 자들에 대한 패턴 분석에 기반하여 손상경험이 없어서 오히려 **안전에 대한 경각심이 적은 농업인들을 위한** 교육자료 제작, 손상 예방 프로그램 기획, 정책 등의 **기초자료로 활용**할 수 있다.

## 한계점

- 지난 1년간의 손상 경험과 이후 주관적 의견을 평가하는 면접 조사였기에 기억의 의존한 자료 수집 과정에서 **기억 편차가 발생**할 수 있다.
- 전체 조사 대상(17,770명)에 비해 추가손상조사표의 **낮은 응답률(약 0.03%)**로 인해 **손상자의 표본집단이 매우 작아** 손상의 특성에 대한 답변과의 **상관관계를 일반화하기 어렵다**.
- 따라서 추후 이러한 한계점들을 보완하여 분석한다면 보다 다양한 인사이트와 정확한 결과를 도출할 수 있을 것이다.