**코로나19-집단감염현황 ver2**

1. **데이터 정보**

|  |  |  |  |
| --- | --- | --- | --- |
| 데이터명 | 코로나19-집단감염현황 | 버전 정보 | ver2 |
| 업데이트 주기 | 주간 | 제공기관 | hikelab |
| 데이터 시작일 | 2020-02-18 | 작업자 | 이정윤,설지은,최은혜 |
| 데이터 설명 | 전국 코로나19 집단감염현황에 대한 데이터로 [ regionid,  province, city, date, subdivided\_Y/N, caseId, infection\_case, confirmed, accumulated, briefingdate, classname, note, keyword ]  13개 항목을 제공함. | 수집 방법 | 중앙방역대책본부와  중앙사고수습본부 사이트에보도되는 정례브리핑 자료의 집단감염 관련 내용을 정리하여 데이터로 정리함. |

1. **컬럼 설명**

|  |  |  |
| --- | --- | --- |
| 컬럼명 | 타입 | 설명 |
| regionid | integer | 각 광역지방자치단체에 부여된 id |
| province | string | 광역지방자치단체명 |
| city | string | 기초지방자치단체명 |
| date | object | 각 집단감염사건의 최초 발생일.  datetime형식으로 기입했으나  null이 포함되어있어서 object로 인식됨. |
| subdivided\_Y/N | object | ver2에서 새로 도입한 컬럼. 한 집단감염사건에서 파생되어 보도자료 상에서 묶어서 보도한 것을 세부분류한(쪼갠) 경우 ‘Y’ 표기. (이외 하나의 사건은 ‘N’으로 표기) |
| caseid | integer | 각 집단감염사건의 구분을 위해 임의로 부여한 id |
| infection\_case | string | 정부 정례브리핑자료에 표기된 각 집단감염사건의 이름 |
| confirmed | float | 각 집단감염사건의 당일 확진자 수 |
| accumulated | float | 각 집단감염사건의 누적 확진자 수 |
| briefingdate | datetime | 각 집단감염사건이 정례브리핑 보도 날짜 |
| classname | string | 각 집단감염사건의 집단 분류.  집단감염사건의 집단 별 특성을 반영하여 만든 집단분류체계를 통해 집단을 분류함. |
| note | string | 각 집단감염사건의 데이터에 대한 수정, 참고사항 등을 기재한 주석 |
| keyword | string | 각 감염집단의 세부적 특성을 반영한 키워드.  해시태그의 기능으로, 한 사건에 여러 개의 키워드가 부여될 수 있음. |

1. **subdivided\_Y/N 부여 기준**

한 집단감염사건에서 파생되어 보도자료 상에서 묶어서 보도하지만, 집단 분류(classname)가 다른 사건을 세부 구분할 필요성이 있다고 판단하여 ver2에서 새로 도입한 분류 기준임. 세부분류한 (한 사건을 여러 사건을 쪼갠) 사건은 한 사건에서 파생되었으므로 동일한 caseId를 부여하고, subdivided Y/N 컬럼에서 ‘Y’를 표기함. (이외 하나의 사건은 ’N’으로 표기)

예시)

테이블이(가) 표시된 사진

자동 생성된 설명

10/14에 처음 등장함 ‘잠언의료기기’가 10/18일 이후 ‘잠언의료기기’와 ‘CJ텔레닉스’로 세부분류되어 보도됨. 이런 경우 subdivided\_Y/N에 ‘Y’를 표기함. (동일한 caseId)

테이블이(가) 표시된 사진

자동 생성된 설명

|  |  |  |  |
| --- | --- | --- | --- |
| **종류** | **케이스** | **분류방법** | **기타 설명** |
| 1 | /로 분리 O + 표 O | 그대로 나누어서 기록 |  |
| 2 | /로 분리 O + 표 X | '(구분)'으로 되어 있는 내용 그대로 나누어서 기록. | ‘(구분)’에는 사건 종류보다 보다 상세하게 구분되어 보도하는 경우가 있음. 이 경우 최대한 비슷한 사건끼리 합쳐서 기록하거나, 구분하기 어려운 경우 기록하지 않음. |
| 3 | /로 분리 X + 표 O | ‘00과 관련하여~’ 라고 설명하면서 감염지 별로 일별 확진자 수를 표로 분류하여 기록하는 경우는 구분하여 그대로 나누어서 기록. | 확진자의 정보(ex 의사\*명, 간호사\*명 등)를 표로 구분하는 경우는 제외함. |
| 4 | /로 분리 X + 표 X | 기록하지 않음 | '(구분)'에 추가전파 (ex 학원생→ 가족및지인) 내용이 기록되어 있는 경우가 있으나, 이는 감염발생지(감염원)이 아닌 감염자에 대한 구분이므로 제외함. |
| 4 |  | 발생 지역이 여러 곳인 경우 (ex (지역) 서울\*명, 부산\*명) 비고에 기록하거나province를 ‘etc’로 분류하고 발생 지역을 city에 표기함. 지역별 세부 확진수는 note에 기록. | 발생한 지역의 수가 많은 경우 (임의로 선정)에만 etc로 분류함. |

1. **caseId 부여 기준**

동일한 집단감염사건을 구분하기 위해 각 집단감염사건에 caseId를 부여함.

테이블이(가) 표시된 사진

자동 생성된 설명

* infection\_case 동일하고 누적값 일치 or 불일치 → 동일한 caseID
* infection\_case 약간 다르고 누적값 일치 → 동일한 caseID
* infection\_case 약간 다르고 누적값 불일치 → 다른 caseID

테이블이(가) 표시된 사진

자동 생성된 설명

* 한 집단감염사건에서 파생되어 보도자료 상에서 묶어서 보도한 것을 세부분류한(쪼갠) 경우 → 같은 caseID

# **데이터 분류체계 설명**

# 집단감염발생지-분류체계 (classname 컬럼에 활용)

### **[집단감염발생지-분류체계](https://docs.google.com/spreadsheets/d/19i8kUj0NZ8mSGh1DDtOTQypZ8g1MF4hvJVzpPsQXz38/edit" \l "gid=0)**

코로나19 집단감염 발생지를 비슷한 특성을 가진 집단별로 분류하기 위해 분류 체계를 새로 정리함.

한국 표준 산업 분류를 바탕으로 하여, 집단감염이 많은 사례의 분류를 새로 추가하거나, 기존 분류명을 집단을 설명하기에 더 직관적인 단어로 변경함.

\* ver1의 분류 중 일부 수정사항이 있는 경우 새로 분류체계(classname)를 부여한 경우가 존재할 수 있음.

1. **keyword 설명**

classname의 집단분류로는 각 감염집단의 세부적인 특성을 반영하는 데 한계가 있으므로 각 감염집단의 세부적 특성을 반영할 수 있는 ‘keyword’를 부여함.

ver1에서는 빈번히 발생한 집단감염 사건만 추가적인 세부 분석을 위해 keyword를 부여하였음.

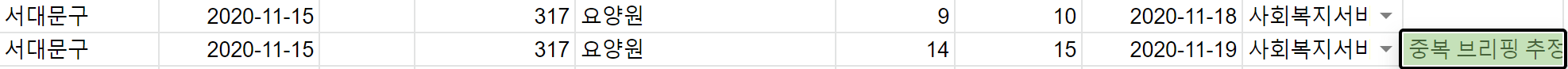
ver2에서는 모든 집단감염 사건에 세부 keyword를 부여함. 이때의 keyword는 ‘해시태크’의 기능으로, 한 사건에 여러 keyword가 부여될 수 있음.

\*\* keyword가 여러개인 경우 ; 로 구분하여 기록.

# **note 설명**

1. **(전일+추가)값과 누적값 불일치**
   * 보도 날짜가 연속되는데 확진자 수 계산이 안 맞는 경우
2. **중간 기록 누락 추정 (누적값 불일치)**
   * 보도 날짜가 연속되지 않고 확진자 수 계산이 안 맞는 경우,   
     중간 기록이 누락된 것으로 추정함.
3. **집단명은 다르지만 누적값 일치로 같은 케이스로 추정됨**
   * infection\_case가 일부 다른데 누적값이 일치하는 경우
   * 집단명은 수정 없이 보도자료 표기 그대로 두고, 같은 케이스로 분류함 (caseID, 최초발생일을 같게함.)
   * infection\_case이 다르고 누적값도 다르면 다른 케이스로 분류함.
4. **중복 브리핑 추정**
   * 전날 기록과 최초발생일, 집단명, 당일 확진 수, 누적 확진 수가 모두 동일한 기록인 경우 중복 브리핑으로 추정함.
   * 전날 기록과 최초발생일, 집단명이 동일하고 당일 확진 수, 누적 확진 수가 1명씩 차이나는 경우도 중복 브리핑으로 추정함.

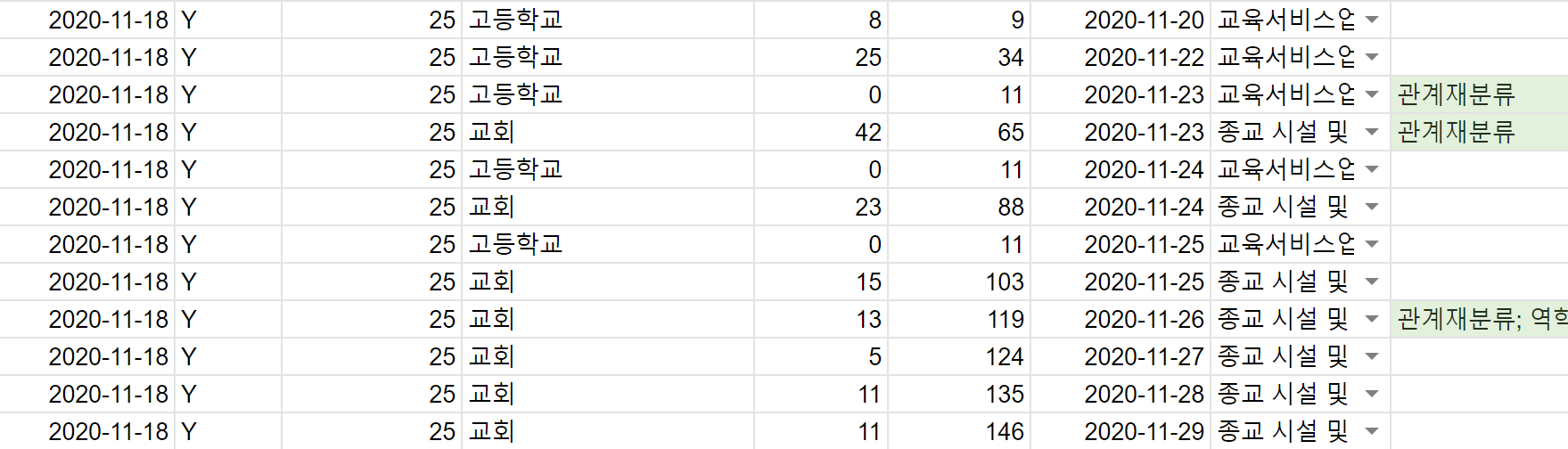
예시)



* + 중복인 사례를 모두 기록하며 원 데이터를 수정하지 않고 note에 중복 브리핑 추정으로 설명함.

1. **브리핑자료 지역명 오기로 [기존 표기 지역 > 수정한 지역] 수정**
   * 지역명 오기로 추정되는 경우 뉴스 기사 확인 후 임의로 지역명 수정함.
   * [충남>충북]은 충남에서 충북으로 수정된 것을 의미함.
2. **이전 데이터와의 비교로 추가확진 수 유추하여 기입함**
   * 보도자료상 누적확진 수만 기록되어 있고 추가확진(당일 신규확진) 수가 기록되어 있지 않은 경우, 이전 데이터와 비교하여 유추할 수 있는 경우에는 유추하여 기록함.
   * 유추할 수 없는 경우에는 공란으로 유지.
3. **이전 데이터와의 비교로 누적확진 수 유추하여 기입함**
   * 초기(2020년 4월 이전) 보도자료에는 추가확진 수, 누적확진 수를 명확하게 표기하지 않으므로, 이 경우 이전 데이터와의 확진자수 비교 계산을 통해 누적확 수를 도출하여 기록함.
4. **관계재분류**
   * ver2에서 진행한 세부분류 작업시, 일자별로 사건 분류가 변경된 경우가 있음. 이 경우, 사건 분류가 변경된 최초 보도일의 note에 ‘관계재분류’를 표기함.

예시)



* + 사건이 세부분류로 추가된 경우 혹은 세부분류되어 있던 사건이 하나의 사건의 축소된 경우를 모두 포함함.
  + 보도자료 상에서 ‘역학조사결과 관계가 재분류됨’ 이라고 명시한 경우에도 표기함.

\*\* 이외의 이슈는 각 케이스에 따라서 note에 기록.  
\*\* 노트 내용이 여러 개인 경우 ; 로 구분하여 기록.