**코로나19-집단감염현황 ver1**

**데이터 정보**

| 데이터명 | 코로나19-집단감염현황 | 버전 정보 | ver1 |
| --- | --- | --- | --- |
| 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, caseid, infection\_case, confirmed, accumulated, briefingdate, classname, note, keyword ]  12개 항목을 제공한다. | 수집 방법 | 중앙방역대책본부와  중앙사고수습본부 사이트에보도되는 정례브리핑 자료의 집단감염 관련 내용을 정리하여 데이터로 정리함. |

**컬럼 설명**

| 컬럼명 | 타입 | 설명 |
| --- | --- | --- |
| regionid | integer | 각 광역지방자치단체에 부여된 id이다. |
| province | string | 광역지방자치단체명이다. |
| city | string | 기초지방자치단체명이다. |
| date | object | 각 집단감염사건의 최초 발생일이다.  datetime형식으로 기입했지만,  null이 포함되어있어서 object로 인식된다. |
| caseid | integer | 각 집단감염사건의 구분을 위해 임의로 부여한 id이다. |
| infection\_case | string | 정부 정례브리핑자료에 표기된 각 집단감염사건의 이름이다. |
| confirmed | float | 각 집단감염사건의 당일 확진자 수이다. |
| accumulated | float | 각 집단감염사건의 누적 확진자 수이다. |
| briefingdate | datetime | 각 집단감염사건이 정례브리핑자료에 보도된 날짜이다. |
| classname | string | 각 집단감염사건의 집단 분류이다.  집단감염사건의 집단별 특성을 반영하여 만든 집단분류체계를 통해 집단을 분류하였다. |
| note | string | 각 집단감염사건의 데이터에 대한 수정, 참고사항 등을 기재한 주석이다. |
| keyword | string | 각 감염집단의 세부적 특성을 반영한 키워드이다.  빈번히 발생한 집단감염 사건의 추가적인 세부 분석을 위해 부여되었다. |

**case id 부여 기준**



동일한 집단감염사건을 구분하기 위해 각 집단감염사건에 caseid를 부여함.

* 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

# **데이터 분류체계 설명** 집단감염발생지-분류체계 (classname 컬럼에 활용)

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

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

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

**keyword 설명**

각 감염집단의 세부적 특성을 반영할 수 있는 ‘keyword’를 만들어 부여하였다.

classname의 집단분류로는 각 감염집단의 세부적인 특성을 반영하는 데 한계가 있기 때문에 빈번히 발생한 집단감염 사건의 추가적인 세부 분석을 위해 생성했다.

ver2에서 한 집단감염사건에 해시태그를 달듯이 여러 키워드가 부여되는 방향으로 더 추가될 예정이다.

| 키워드 | 예시 |
| --- | --- |
| 설명회 | 설명회, 세미나 등 |
| 집회 | 8.15 집회 등 |
| 콜센터 | 콜센터, 고객센터 등 |
| 목욕장업 | 사우나, 목욕탕, 마사지업 등 |
| 장례식장 | 장례식장 |
| 어린이집/유치원 | 어린이집, 유치원 등 |
| 초중고 | 초/중/고, 특수학교, 대안학교, 중학교 음악캠프 등 |
| 대학교 | 대학교, 대학교 운동부, 직업전문학교 등 |
| 학원 | 어학원, 댄스학원, 태권도학원 등 |
| 마트, 시장 | -시장, -마트, 마트 푸드코트 등 |
| 요양병원/요양시설 | 요양병원, 요양원 등 |
| 일반 음식점 | 식당, 음식점 등 |
| 카페 | 카페 등 |
| 주점 | 유흥시설 분류에 해당하지 않는 일반 주점 |
| 노조 | 항운노조 등 |

# **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 살짝 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
   * 중복인 사례를 모두 기록하며 원 데이터를 수정하지 않고 note에 중복 브리핑 추정으로 설명함.
5. **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.
6. **이전 데이터와의 비교로 추가확진 수 유추하여 기입함**
   * ver1에서만 사용
   * 20.02~21.02 수집 자료에서는 이전 데이터와의 확진자수 비교 계산을 통해 확진자 수를 도출하여 기록한 것이 있음
7. **이전 데이터와의 비교로 누적확진 수 유추하여 기입함**
   * ver1에서만 사용
   * 20.02~21.02 수집 자료에서는 이전 데이터와의 확진자수 비교 계산을 통해 확진자 수를 도출하여 기록한 것이 있음
8. **이전 데이터와의 비교로 추가확진, 누적확진 수 유추하여 기입함**
   * ver1에서만 사용
   * 20.02~21.02 수집 자료에서는 이전 데이터와의 확진자수 비교 계산을 통해 확진자 수를 도출하여 기록한 것이 있음

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