**Intro**

안녕하세요, 저는 ‘숨겨진 정보를 데이터로! 집단 감염 데이터 분석’ 파트를 맡은 발표자 중앙대 문정과 대학원에 재학중인 박하람입니다. 저희는 집단 감염 데이터를 주제로 데이터 수집 과정부터 분석 결과까지 발표해보려고 합니다.

**목차**

저는 분석 배경 및 목적, 데이터 수집 과정, 결론부를 발표하고, 중간 데이터 분석 파트는 중앙대 문정과 학부생 최은혜 님께서 발표해주실 겁니다.

**분석 배경 및 목적**

저희가 처음 가진 생각은 바로 이러한 기사들을 보면서였습니다. 5인이상 사적 모임 금지가 전격화되기 전에는 업종별 영업금지가 주요한 방역 정책이었습니다. 그런데 특정한 업종의 영업 정지가 지속되면서 방역 정책에 대해 반발의 목소리가 커지게 되었습니다. 만약 코로나19 종식을 위해 영업금지 업종을 선정하는 것이 불가결한 것이라면, 어떤 기준으로 영업 금지 업종을 선택하는지 중요하다고 생각했습니다. 또한 지역별로 집단 감염이 얼마나 일어나는지도 알고 싶었구요.

따라서 ‘집단 감염’ 데이터를 분석하여 어떤 업종에서 집단 감염이 많이 발생하였는지, 지역별 집단 감염 정도는 어떠한지 알아보고자 하였습니다.

**페이지 넘김**

저희가 집단 감염 데이터를 구할 때 중요하게 생각했던 것은 4가지입니다. 첫번째, 이용자가 쉽게 접근할 수 있는 데이터인가? 두번째, 저작권 정책이 명시되어있는 공개된 데이터인가? 세번째, 기계로 처리할 수 있는 데이터인가? 네번째, 코로나 19가 시작됐을 시기부터 현재까지의 데이터가 모두 존재하는가? 였습니다.

그러나 처음부터 난항을 겪었던 것은 제시한 4가지를 모두 충족하는 집단 감염 데이터는 존재하지 않았다는 것입니다. API로도 제공하지 않는 데이터일 뿐만 아니라, 집단 감염 데이터 자체를 찾아보기가 힘들었습니다. 작년 Kaggle에 올라와 있는 데이터가 사용할 수 있는 유일한 데이터였으나 이마저도 8개월째 업데이트되지 않고 있습니다. 그러나 Kaggle 데이터로부터 얻을 수 있었던 단서는 중앙방역대책본부에서 매일 제공하는 코로나19 정례 브리핑 데이터에서 집단 감염 사례를 찾을 수 있다는 것이었습니다.

**페이지 넘김**

왼쪽의 그림을 보면 정례 브리핑 데이터에는 이런 방식으로 구체적인 집단 감염 사례 정보를 체계적으로 기록하고 있었습니다. 집단감염 장소, 첫 집단감염 발생일, 당일 확진자 수, 누적 확진자 수, 일부 감염경로 정보 등을 제공하고 있었습니다. 정례 브리핑 데이터는 누구나 접근할 수 있지만, 데이터를 공공데이터 API와 같이 가공하기 좋은 형태로 제공되지 않는다는 점에서 '숨겨진'이라는 표현을 사용했습니다.

**페이지 넘김**

그래서 저희는 이 정보를 데이터화 해보자고 생각했습니다. 그래서 작년 1월부터 올해 2월까지의 모든 브리핑 데이터를 읽어 이런 방식으로 일일이 엑셀에 기록하는 작업을 진행했습니다.

**데이터 수집과정**

정례 브리핑 데이터에서 가져올 수 있는 데이터와 함께 저희는 몇가지 정보를 더 추가하여 집단감염 데이터를 수집하였습니다. 기존 데이터 수집 기준으로는 행정구역으로 광역시도와 시군자치구, 첫 확진자 발생일, 감염경로, 당일 확진자수와 누적자수, 브리핑 날짜가 있었습니다. 우리 팀은 기존의 정보에 지역 번호와 사건 번호와 분류 기준을 추가하여 2020년 1월부터 2021년 2월까지의 데이터를 수집하였습니다.

칼럼에 세 가지 컬럼을 추가한 이유는 다음과 같습니다.

첫째, 지역별 데이터를 용이하게 정제할 수 있도록 하기 위해서 지역번호를 추가했습니다.

둘째, 집단감염 사건별로 case-ID를 부여하여 감염경로 명칭이 같은 경우, 중복 값 없이 분석을 할 수 있게 했습니다.

셋째, 집단감염경로를 업종별로 분류하기 위해 분류기준을 지정하여 분류하는 작업을 수행하기 위함이었습니다.

**페이지 넘김**

**데이터 수집 과정 오류**

데이터를 수집하는 과정에서 제공하는 정보들의 오류 사항을 비고란에 입력했는데요. 큰 줄기에서는 데이터가 체계적으로 제공되고 있다는 느낌을 받았지만, 하루하루의 확진자수를 계산하면 오류가 생기는 부분이 꽤 존재했습니다. 오류는 전일 누적 및 당일 추가 확진자수를 더한값과 당일 누적 확진자수가 일치하지 않는 경우가 존재했구요. 또한 의료파업이나 국회의원 선거 등과 같은 날에는 구체적 당일 집단감염 확진자 수 정보를 제공하지 않는 경우도 있었습니다.

**페이지 넘김**

**데이터 수집 과정: 데이터의 변화**

또한 데이터를 수집하는 과정에서 초기와 후기의 데이터의 변화를 볼 수 있었습니다. 초기에는 ‘신천지 대구교회’, ‘청도대남병원’ 등과 같이 구체적이고 장소를 특정할 수 있는 집단감염 장소를 기입한 것에 비해, 후기로 갈수록 ‘직장’, ‘요양시설’ 등 특정적이지 않고 포괄적인 이름으로 이름이 변화한 것을 볼 수 있습니다.

**페이지 넘김**

완성된 집단 감염 데이터는 이렇게 스트레드시트에 모두 정리해놓았습니다. 이후 이 데이터는 모두 공개할 예정입니다.

다음은 최은혜 발표자께서 발표해주시겠습니다.

**집단감염 데이터 분석**지금부터는 이렇게 저희가 구축한 집단감염 데이터를 가지고 분석한 내용을 말씀드리겠습니다.  
집단감염 데이터를 업종별로, 지역별로 나누어 크게 2가지 분석을 진행했습니다.

**집단감염 사건 수**

* **월별 집단감염 발생 수**

먼저 전체 집단감염 건수에 대한 월별 통계입니다.

데이터를 집단분류별로 나눠서 보기 전에 전체 집단감염 사건수를 월별로 살펴봤습니다.

그래프를 보시면 직관적으로도 하반기에 집단감염 사건이 많이 발생한 것을 확인할 수 있습니다. 더 자세히 말씀드리면 가장 긴 막대가 작년 12월입니다. 작년년 12월에 215건으로 집단감염이 가장 많이 발생했고, 올해 1월과 2월, 작년 11월이 각각 140건, 122건, 120건으로 그 뒤를 이었습니다..

이는 누적된 방역 피로감, 할로윈 데이와 성탄절, 추석 등의 연말, 연초 행사와 명절 모임이 몰려있던 것의 영향으로 보입니다.

반대로 2020년 상반기 8월 전에는, 집단감염 발생 건수가 각 월 모두 50건 미만으로 전반적으로 적었던 것을 확인했습니다.

위 그래프를 통해 대유행 시기를 언뜻 겹쳐 볼 수 있습니다. 1차 대유행이 나타났던 3월에 45건으로 상반기 중 가장 많은 사건 수를 보였다가 안정되는 모습을 볼 수 있고,

8월 중순, 2차 대유행과 동시에 폭발적으로 집단감염 사건 발생이 늘어났다가 또 다시 약 2개월에 걸쳐 안정되었고,   
11월, 12월에 3차 대유행기로 접어들면서는 집단감염 사건 수가 최고점에 달했다가 현재까지 점차 줄어들고 있는 모습을 보입니다.

**집단분류별 집단감염 건수**

* **집단분류별 집단감염 건수(월별)**

다음은 집단분류별로 집단감염 사건 수의 월별 변화를 살펴보도록 하겠습니다.   
그래프의 전반적인 추세를 먼저 보자면, 8월부터 보이는 막대의 수가 많아졌습니다. 많은 분류에서 집단감염 사건이 발생하기 시작한 것입니다. 시간이 더 흘러서 11월, 12월에는 전반적으로 막대가 길어진 것을 확인할 수 있습니다. 대부분의 분류에서 사건 수가 급증한 것입니다.

더 상세히 확인해 봤을 때 한 달 동안 많은 사건이 발생한 케이스를 말씀드리자면, 12월에 ‘종교시설 및 단체’에서 51건, 11월에 ‘가족, 지인 및 소모임’에서 31건의 집단감염이 발생했습니다. 이는 앞서 말했듯이 연말 연초 행사들의 여파로 보입니다.

세번째로 많은 집단감염 발생이 있었던 케이스는 12월에 ‘요양병원 및 시설’에서 27건이 발생한 것이었습니다. 이는 당시의 코호트 격리로 방치된 노인들과 구호물품 부족 등의 요양병원 관련 이슈와 연결지어 생각해볼 수 있습니다.  
참고로 이 ‘요양병원 및 시설’과 ‘종교시설 및 단체’에서는 4월을 제외하고 매월 집단감염이 발생이 있었으며,   
‘의료기관’에서는 7월을 제외하고 매월 집단감염이 발생했습니다.

* **집단 분류별 집단감염 건수(누적)**

다음은 집단감염 발생의 월별 누적 애니메이션입니다.

 7월 전까지는 비교적 작은 폭으로 누적되다가 8월부터 눈에 띄는 변화를 보입니다.   
결과적으로 최종 누적 그래프를 보시면 상위 6개 집단이 직관적으로도 다른 집단에 비해 집단감염 건수가 많음을 알 수 있습니다. 상위 6개 집단 중에서도 특히 사건 수가 많은 이 두 분류는 가족 지인 및 소모임과 종교시설 및 단체입니다. 집단분류를 이런식으로 했을 때 이 두 집단은 집단감염과 관련하여 유의하게 살펴볼 필요가 있습니다.

그리고 ‘집회’에서 발생한 집단감염은 8월 단 한 사례로 가장 적게 집계되었으나, 565명의 확진자를 발생시켰고, 종교시설 분류의 ‘사랑제일교회’와 연관성이 있으며 그 여파가 2차 대유행과 깊게 연관되어 있었습니다. 이러한 점에서, 한 사건의 영향력이 여러 분류에 걸쳐 분산될 수 있으며   
그렇기 때문에 집단감염 건수를 절대적인 심각성과 직결되는 지표로 보는 것은 지양하며 데이터를 참고해야 할 것으로 생각됩니다.

**분류별 집단감염 확진자 수**

* **업종별 집단감염 확진자 수 (월별)**

집단감염 확진자 수 누적값을 그래프로 나타낸 결과, 작년 초반과 8월을 기접으로 종교시설 및 단체에서 확진자 수가 급증한 것을 볼 수 있습니다. 11월쯤부터는 전반적으로 집단감염 확진자가 증가한 것을 확인했습니다. 긴 막대가 비교적 많았던 작년 12월을 다시 짚어보면 요양병원 및 시설에서 한달동안만 1480명, 종교시설 및 단체에서 1382명, 교정시설에서 813명으로 확진자 수가 급증했습니다.

* **업종별 집단감염 확진자 수 (누적)**

다음은 분류별 확진자수 월별 누적 애니메이션입니다. 앞서 말씀드렸던 확진자 증가 양상을 종합적으로 파악할 수 있습니다.

결과적으로 올해 2월 전체 누적값을 확인해보면 종교시설 및 단체에서 발생한 확진자가 10777명으로 가장 많았으며, 2위인 요양병원 및 시설의 확진자보다 약 3배 높은 수치를 기록했습니다. 그리고 가족 및 지인 소모임 분류가 그 뒤를 이었습니다.ㅎ\

확진자가 특히 적게 발생한 분류는 숙박업과 항구/선박업으로 각각 13명 20명의 확진자를 기록했습니다.

**지역별 집단감염 분석**

다음은 집단감염 데이터를 지역별로 분류하여 분석을 진행한 것을 보여드리겠습니다.

* **지역별 집단감염 건수(누적)**

그래프를 보시면 etc.라는 항목이 존재하는데 etc는 한 사건에 여러 지역의 확진자가 혼재되어 존재하는 경우를 분류한 것입니다.

먼저 애니메이션 그래프를 통해 지역별로 집단감염 사건 수 누적값의 변화를 월별로 살펴봤습니다.

그래프를 보시면 작년 7월부터 서울, 경기 지역과 다른 지역의 격차가 빠르게 벌어지며 결과적으로 전 기간동안 다른 지역의 전체 집단감염 사건 수는 가장 많은 곳이 70건 이하인 반면, 서울과 경기지역에서 발생한 집단감염 사건의 수는 각각 200건 이상을 기록한 결과를 볼 수 있습니다.

유동인구가 많은 수도권 지역의 집단 감염 발생이 다른 지역에 비해 심각했던 것을 데이터를 통해 확인할 수 있었습니다.

작년 12월과 올해 1월에는 서울 경기 지역 뿐만 아니라 전 지역의 막대가 빠르게 길어지는 것을 보실 수 있습니다. 이를 통해 이 시기에 집단감염 사건 발생 수가 전국 전반적으로 증가한 것을 확인할 수 있습니다.

* **지역별 집단감염 확진자 수(누적)**

다음은 지역별 집단감염 ‘확진자 수’를 살펴본 시각화 분석을 보여드리겠습니다.

이 그래프를 통해 지역별로 집단감염 확진자의 전체 누적값 변화를 월별로 살펴봤습니다.

그래프를 재생하면 초반에 대구 확진자 수에 해당하는 막대가 빠르게 길어지는 모습을 볼 수 있습니다. 이후 증가가 거의 멈춘 수준으로 그 속도가 현저히 줄어들고 7월부터 서울과 경기의 확진자 수 막대가 빠르게 길어지는 것을 확인할 수 있습니다.

12월과 1월에는 전국의 확진자 수 막대가 빠르게 길어지며 전국 전반적으로 집단감염 확진 상황이 악화됐었던 것을 다시 한 번 확인했습니다.

결과적으로 봤을 때 서울, 대구, 경기 모두 집단감염 확진자 수가 다른 지역들에 비해 압도적이지만 대구는 2, 3월에 전체 확진자의 대부분이 발생한 후 확진상황은 안정되었고, 서울과, 경기는 7, 8월부터 지금까지 계속 심각한 상황이었습니다. 이런 차이를 그래프를 통해 확인할 수 있었습니다.

* **지역별 인구대비 분석**

앞에서 확인했던 지역별 집단감염 사건 수와 확진자 수를 가지고 지역의 집단감염 확진 상황을 더 정확히 파악하기 위해서는 지역별 인구수도 함께 고려하는 과정이 필요하다고 생각해서 인구수 대비 집단감염 상황을 분석해봤습니다. 지역별 인구수로 행정안전부의 2021년 2월 기준 주민등록 인구 및 세대현황 데이터를 이용하였고, 누적 사건 수, 누적 확진자 수에 각각 인구 수를 나눠서 분석을 진행했습니다.

* **인구수 대비 집단감염 건수**

먼저 인구수 대비 집단감염 ‘사건수’ 분석 결과를 말씀드리겠습니다. 앞에서 인구수를 고려하지 않고 누적 집단감염 사건수를 봤을 때는 서울과, 경기가 압도적으로 사건 수가 많았습니다. 하지만 각 지역의 인구수를 고려더니 양상이 달라졌습니다.

서울은 여전히 가장 높은 수치를 보이지만, 서울 다음으로 경기가 아닌 강원도, 경북, 광주, 대구가 차례로 뒤를 이었고 서울과 다른 지역간의 수치 차이도 작아졌습니다.  
사건 수만 고려했을 때 눈에 띄게 많았던 서울과 경기는 이 그래프에서는 비교적 눈에 띄지 않습니다. 서울과 경기는 인구수가 다른 지역에 비해 훨씬 많기 때문에 전체 누적 사건 수가 비교적 많아 보이는 것이라는 사실을 알 수 있었습니다.

* **인구 수 대비 집단감염 확진자수**

다음은 인구수 대비 집단감염 ‘확진자 수’ 분석 결과를 보여드리겠습니다.   
지역별로 집단감염 확진자수만 봤을 때는 서울, 대구, 경기의 누적확진자가 눈에 띄게 많았던 것과 달리 인구수를 고려하자 대구의 인구대비 확진자수가 다른 지역들에 비해 수치가 압도적인 것을 확인할 수 있었습니다. 대구에서는 인구 수를 고려했을 때도 집단감염 확진 상황이 특히 심각했던 것으로 보입니다.

누적확진자수만 결과적으로 고려했을 때 수가 눈에 띄게 컸던 서울, 대구, 경기 중에서도 수가 가장 컸던 서울의 경우 대구보다 인구수가 약 4배 많은 것을 고려했을 때 인구대비 누적확진자 수는 대구의 절반도 되지 않는 것을 확인했습니다.

서울 다음으로는 광주와, 충북이 인구대비 확진 상황이 심각했던 것으로 보입니다.

사건 수와 확진자 수 모두 절대적인 수치만 고려되어 서울과 경기의 집단감염 발생이 특별히 심각한 것으로 여겨지고 있지만 인구 수를 고려하여 집단감염 현황을 봤을 때 사건 수는 수도권과 다른 지역들의 차이가 생각보다 크지 않았고, 인구수 대비 확진자 수도 대구는 특히 심각했지만 다른지역과 수도권의 격차가 크지 않았습니다. 서울과 경기 못지않게 다른 몇몇 지역 또한 집단감염상황이 좋지 않았던 것으로 보입니다.

인구수가 많은 곳은 당연히 확진자 수가 많게 보일 수밖에 없습니다. 지역별 전체 인구수를 고려하여 데이터 기반의 지역별 방역조치가 이뤄졌는지, 절대적인 수치만 고려하여 우리가 수도권 외 지역에 대한 집단감염 위험의 심각성을 놓치지 않았는지 되짚어 볼 필요가 있습니다.

* **지역별 집단감염 사건수와 확진자수**

마지막으로 앞에서 살펴본 지역별 집단감염 사건 수와 확진자 수를 한 번에 비교할 수 있도록 X축을 누적 사건 수, y축을 누적확진자수로 하는 애니메이션 그래프를 그려봤습니다. 원의 크기는 y축과 마찬가지로 누적 확진자수를 의미합니다.

그래프를 보시면 초반에 대구의 누적확진자 수가 약 5천명으로 급증한 것을 확인할 수 있고, 이후 대구는 사건 수와 확진자 수의 변화가 비교적 작아진 것에 비해 서울과 경기의 누적 사건 수가 급격히 증가하며 빠르게 오른쪽으로 원이 이동합니다.

이렇게 계속 누적되어서 결과적으로 2021년 2월, 전체 그래프를 봤을 때   
대구와 경기가 원의 크기가 비슷한 것을 확인했습니다. 하지만 대구가 원이 훨씬 왼쪽에 위치해 있어서, 경기보다 누적확진자는 비교적 많고, 누적사건 수는 비교적 적은 것을 확인할 수 있습니다.

대구는 2020년 초반에 확진자 수가 급증했다가 확진 상황이 잠잠해지고, 경기는 이후 계속 지금까지 확진자 수와 사건 수 모두 빠르게 증가했음에도 불구하고 결과적으로 대구의 집단감염 확진자가 경기보다 많습니다.

그만큼 2020년 초반에 대구에서 발생했던 집단감염 사건은 약 5천명의 많은 집단감염 확진자를 만든 심각한 사건이었음을 확인했습니다.

이를 통해 코로나19 발생 초반에 집단감염 발생에 대한 방역체계가 제대로 대비되어 있었는지, 그에 따라 방역조치가 취해진 것이 맞는지, 그것이 적절했는지 재고해볼 필요가 있다고 생각합니다.

지난 기간동안 지역별로 집단감염 상황에 따라 다른 방역대처가 이뤄졌습니다. 지역별로 집단감염 상황이 다르기 때문에 다른 방역 조치가 취해지는 것은 합당한 것으로 받아들여질 수 있습니다. 하지만 이렇게 지역별로 달랐던 방역 대처에 대한 명확한 기준이 존재했는지, 그 내용이 적절했는지 되짚고 넘어갈 필요는 있습니다. 각 방역대처의 기준이 무엇인지 확인하고, 그 대처가 효과적이었는지 평가하기 위해서는 앞으로 지역별 집단감염 데이터와 데이터에 대한 정확한 분석이 수반되어야할 것입니다.

**결론과 제언**

다시 저희의 분석 목적으로 돌아가보면, 과연 정부의 업종별 영업금지 방역 정책과 지역별 방역 정책은 효과적이었을까요?

**페이지 넘김**

사실 이 데이터만으로는 충분히 알 수 없습니다. 그러나 분명한 것은 집단감염 데이터는 업종별, 지역별 방역 정책에 대한 효과를 알기 위해서 기본적으로 필요한 데이터라는 것입니다.

**페이지 넘김**

그래서 저희 팀에서 강조하고 싶은 것은 잘 정제된 데이터가 정부 차원에서 제공되어야 한다는 것입니다. 코로나가 시작된지 1년이 지난 후에도 집단감염 데이터는 이용자가 사용하기 편한 형태로 제공되고 있지 않습니다. 정례 브리핑 문서를 보면 체계적으로 집단 감염 사례가 기록되고 있다고 생각되지만, 실질적인 분석을 위해서는 table 형태나 공공데이터 API과 같은 형태로 제공하는 것이 필수적입니다. 더 나아가 machine-readable한 형태로 읽을 수 있게 된다면 더 활용성이 높아지겠죠.

더불어 데이터를 제공할 때에 더욱 풍부한 데이터를 제공하는 것이 필요해보입니다. 이용자가 이를 가공하기 위해서는 첫번째로 집단을 구분할 수 있는 집단 ID를 제공할 수 있어야 합니다. 사실 정부 브리핑 일부에 존재하는 관리번호 즉 집단 ID의 데이터는 방역 당국에서 기록하지 않는 정보가 아닙니다. 이미 가지고 있는 데이터이기 때문에 쉽게 공개할 수 있다고 생각하구요.

두번째로 집단명 부여 기준을 수립하여 일관적인 집단명이 기록될 수 있도록 해야 하고 체계적인 집단 분류 체계 또한 세워야할 필요가 있습니다. 최근의 사례들에서 더욱 애매한 명칭들이 많이 존재했는데요, 가령 종교단체의 경우 교회와 불교, 성당 등을 모두 포괄하여 종교단체로만 기입했기 때문에 정확한 분류를 하기 어렵습니다. 따라서 일관적인 집단명과 함께 체계적인 분류기준를 제공해주는 것이 필요합니다

덧붙여 집단감염에 대한 정의 또한 부재하다고 볼 수 있습니다. 어떤 것이 집단 감염인지 명확히 알기 어려웠는데요, 브리핑 자료에서는 국내 주요 감염 사례 혹은 주요 집단 발생 현황으로 기록하고 있습니다. 저희는 이것을 집단감염 데이터라고 생각했지만, 이를 집단 감염 데이터로 볼 수 있는지, 볼 수 있다면 집단 감염이라 생각하는 기준은 무엇인지 등 명확한 정의를 내릴 필요가 있습니다.

**페이지 넘김**

**발표를 맺으며**

업종별 영업금지 방역정책은 많은 사람들의 생업에 영향을 미치는 만큼, 충분히 설득력 있는 근거 하에 영업정지 업종이 선택되어야 할 필요가 있습니다. 지역별 방역 정책 또한 마찬가지이구요. 이를 위해 ‘집단 감염 데이터’는 방역정책 기준 선정의 기본적인 데이터가 될 수 있습니다. 이를 정부 차원에서 공개한다면, 방역 정책에 대한 기준을 설득력 있게 제시할 수 있을 뿐만 아니라, 사회적으로 중요한 의미를 가지는 데이터인만큼 더욱 유의미한 가치를 창출할 수 있을 것으로 생각합니다.

마지막으로 이번 분석을 통해 지금까지 앞서 집단감염 취약집단으로 언급됐던 업종들에 대해서 코로나 방역 대처가 적절히 이루어졌는지를 되살펴보고 앞으로의 방역 대책을 고민하는 계기가 되었으면 합니다.