2017 InfoVis

Visual Exploration Exercise Report

-2015년 진료내역 Data기준-

삼성전자

기여운

Date : 11th, Nov, 2017

Table of contents

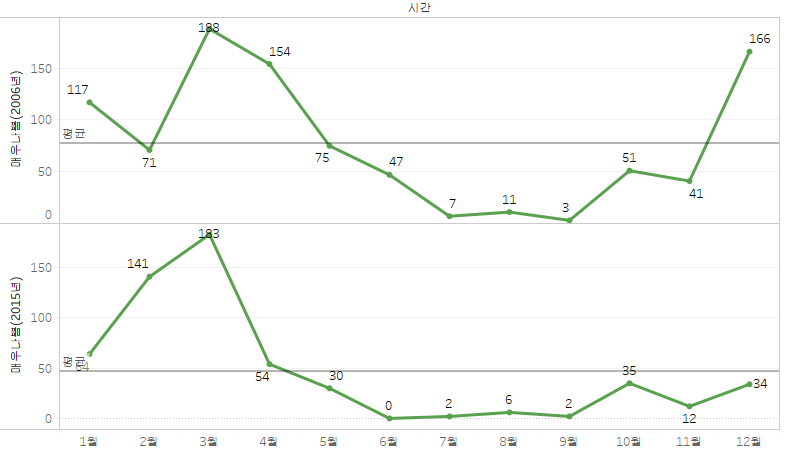
1. 배경
2. 선정 Data 소개
3. 1차 발견
4. 2차 발견
5. 정리
6. 배경

데이터를 시각화하고 인사이트를 얻기 위한 가장 강력한 동기는 데이터에 대한 목적의식과 호기심이라고 생각한다. 본 과제를 통한 결과에의 목적의식(과정이 아닌)은 명확하지 않았으나 그 대신 호기심으로 시작했다. 또한 인사이트를 얻는 과정에서의 통계값 들은 통계학적인 가정과 그에 대한 증명에 기반한 분석이 아님을 밝힌다.

*“미세먼지와 목감기(호흡기감염)는 실제로 관련이 있을까?”*

*[[1]](#footnote-1)*

대한민국 공공데이터포털(<http://data.go.kr>)에서 얻을 수 있는 다양한 데이터 중, 먼저 15년과 06년의 전국 미세먼지 수치 통계데이터를 시각화했다. 이는 10년동안 미세먼지의 수치는 급격하게 증가했을 것이라는 가정하에 이를 확인한 후 호흡기 질환의 발병의 양상을 보고자 함이었다.

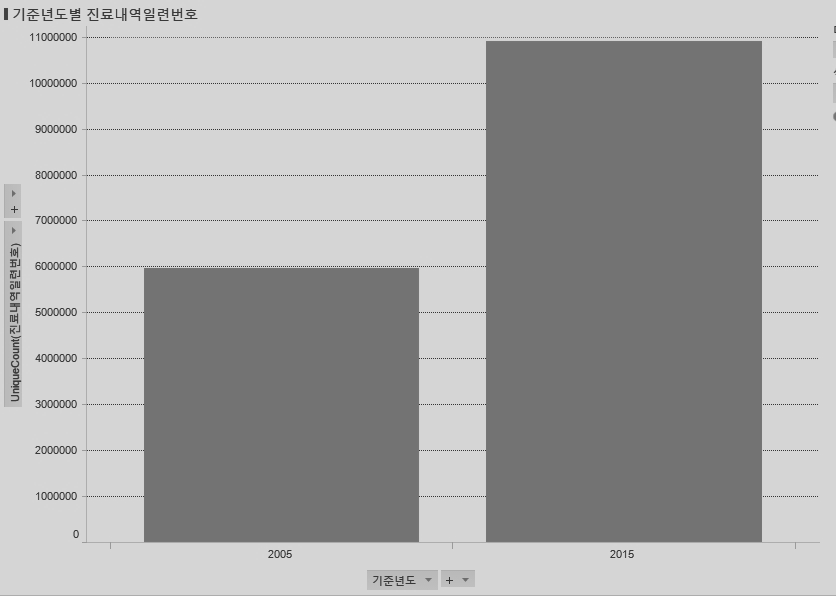


<그림 1>

그림1은 강화도 지역에 대한 06년과 15년도의 시간 별 미세먼지(PM101) 데이터[[2]](#footnote-2)에서 시간당 측정 농도가 100 ㎍/㎥이상(WHO기준 매우나쁨 예보)인 횟수의 합을 월 단위로 보여주고 있다. (전국 전역에 대한 수치를 사용해보고자 했으나 06년에는 수도권에 대한 미세먼지 데이터가 보관되어 있지 않아 두 해에 걸쳐 모두 데이터가 있는 강화 지역을 샘플로 분석해보았다. 전국으로 확대할 경우 경향이 달라질 수 있지만 , 그렇지 않을 것이라는 가정하에 인사이트를 도출하기로 했다.) 예상과는 달리 2월을 제외한 모든 기간에 06년 측정결과가 많게는 3배까지 높았으며 평균 미세먼지 농도 역시 06년은 58.5㎍/㎥, 15년은 46.91㎍/㎥ 이다.

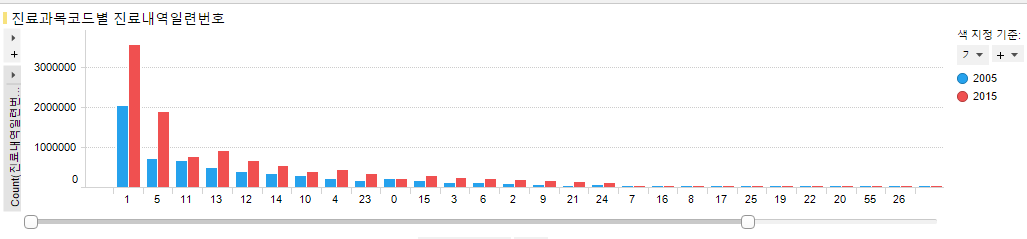
이 결과로 인해, 10년간 미세먼지 수치의 급격한 변화에 대한 가정은 성립되지 않게

되었는데 그렇다면 정말 사람들의 발병 경향 역시 연관성을 얻기가 힘든지 확인해보고자 두 해의 진료내역[[3]](#footnote-3)을 시각화해보고자 했다. 미세먼지는 06년과 15년 데이터를 분석했으나 진료내역은 05년과 15년 데이터를 분석했음을 미리 알린다.



<그림2>

그림2[[4]](#footnote-4)는 대한민국 국민 100만명이 05년, 15년 각 해에 총 몇 번의 진료를 받았는지에 대한 그래프이다. 두 해의 인구수 차이는 약 380만명으로, 비율로 따지면 약 8%정도가 증가[[5]](#footnote-5)한 것인데 샘플링이 고르게 분포된 데이터이기때문에 단순하게 동일한 비율을 적용한다면 진료내역도 유사한 수준이어야 하겠다. 그러나 진료내역수는 약 80%이상 증가했다. 이것은 무엇을 의미하는 것일까? 좀 더 살펴보기 위해 아래와 같이 진료과목 별 진료횟수를 그려보았다<그림3>. 내림차순 정렬 후, 두 해의 진료횟수의 차이가 2배 이상 나는 과목을 발견할 수 있었다.



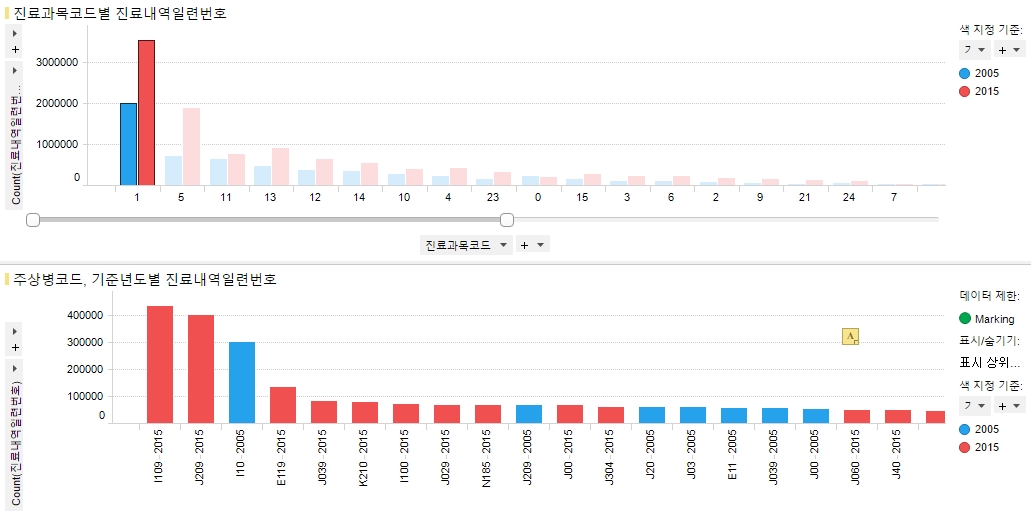
<그림3>

05년 기준으로 상위 5개 진료과목은 내과(1), 정형외과(5), 소아청소년과(11), **이비인후과(13)**, 안과(12) 순인데 15년에는 그 순위가 내과(1), 정형외과(5), **이비인후과(13)**, 소아청소년과(11), 안과(12) 순으로 이비인후과의 순위가 한 단계 높아졌고, 1,2 순위의 상승폭은 압도적이다.

표 1주요 진료과목 코딩표

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **구분** | **의과\_보건기관** | **구분** | **의과\_보건기관** | **구분** | **의과\_보건기관** |
| **1** | **내과** | 6 | 신경외과 | 12 | 안과 |
| 2 | 신경과 | 8 | 성형외과 | **13** | **이비인후과** |
| 3 | 정신과 | 9 | 마취통증의학과 | 14 | 피부과 |
| 4 | 외과 | 10 | 산부인과 | 15 | 비뇨기과 |
| 5 | 정형외과 | **11** | **소아청소년과** | **23** | **가정의학과** |

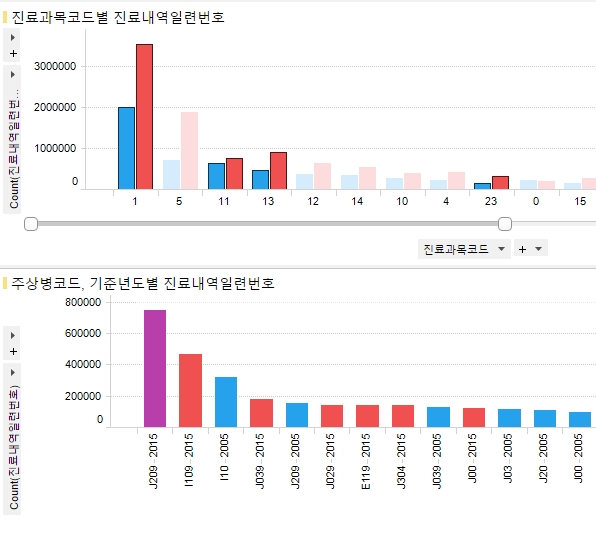
먼저 가장 상승폭이 큰 내과에서 어떤 원인의 질병이 가장 많이 진단되었는지 보자<그림4>. 진료과목 마킹 후에 해당 내용에 대한 세부 시각화를 하여 상위 25개의 진단명[[6]](#footnote-6)을 보여주도록 했다.



<그림4>

두 해 모두 최다진단질병은 고혈압(I10,I109)이며, 그 다음으로 많이 진단받은 것이 상세 불명의 급성기관지염(J209)이다. 그러나 상세 불명의 급성기관지염의 진료내역수는 05년 대비 15년에 4배 가까이 증가했다 상세 불명의 급성기관지염 다음으로 많은 진단명으로, 05년에는 급성기관지염(J20), 급성편도염(J30), 당뇨병( E11), 상세 불명의 급성편도염(J039), 코감기(J00)순이었는데, 15년도에는 당뇨병(E119), 상세 불명의 급성편도염(J039), 역류성 식도염(K210), 상세 불명의 급성인두염(J209) 등의 순으로 식습관 변화에 따른 질병의 증가와 더불어 “상세 불명”의 호흡기 질환의 증가를 알 수 있었다. (원인을 알 수 없는 질병 현상에 대한 진단코드는 상이하며 “상세 불명” 으로 별도 분리되어 구분하고있다.) 이는 미세먼지와 호흡기질병의 연관성을 설명할 수 있는 증거가 될까?

좀 더 확인해 보기 위해, 호흡기질환을 진단할 가능성이 높은 4개 진료 과목인 내과(1), 소아청소년과(11), 이비인후과(13), 가정의학과(23)를 선택하여 발병 현황을 살펴보기로 했다. 4개 진료과목으로 제한하여 본 결과, 그림5와 같이 15년도에 상세 불명의 급성기관지염(J209)이 최다 질병으로 순위가 바뀌는 현상이 나타났다. 가정에 대한 확신을 좀 더 갖게 해주는 사실이었다.



<그림 5>

지금까지 살펴본 내용을 토대로 다음과 같은 정리가 가능하다.

|  |
| --- |
| 1. 강화지역의 10년간 미세먼지(PM10)의 수치의 급상승은 없다. 2. 오히려 10년전 평균 미세먼지(PM10)수치가 더 좋지 않다. 3. 10년간 인구수증가률에 비해 진료횟수 증가률이 급격히 상승했다. 4. 특히 내과의 진료수가 급격히 상승했으며, 05년 대비 이비인후과가 진료수 많은 진료과목의 순위를 4위에서 3위로 올라왔다. 5. 내과에서 05년 대비 15년에 원인을 알 수 없는 다양한 호흡기 질환의 발병이 증가했다. 6. 호흡기질환의 진단이 가능한 4개 진료과목을 필터링해서 봤을 때, 상세 불명의 급성 기관지염이 고혈압보다 더 많은 발병되었다. |

인구수가 증가한 것에 비해 사람들은 더 자주 병원을 가게 되었고, 특히 호흡기계통에 대한 질병 진단이 확연히 증가했다. 이러한 질병을 발생시키는데 우리가 마시는 공기에는 정말 문제가 없었는지 확인하기 위해서는 앞서 강화지역으로 제한했던 데이터 범위를 전역으로 확대하여 분석하는 것이 필요하고 또한 미세먼지 기준 역시 초미세먼지(PM2.5) 등도 함께 분석이 이루어져야 했다. 그러나 데이터 확보의 어려움 및 현재 보유한 HW 및 SW 장비로는 년도별 약 1천만건 이상의 진료내역 데이터 분석을 효율적으로 진행할 수가 없었다. 따라서 앞으로 보게 될 데이터 시각화는 이러한 궁금증을 염두에 두고 15년 진료 내역을 다양한 각도에서 살펴본 결과이다.

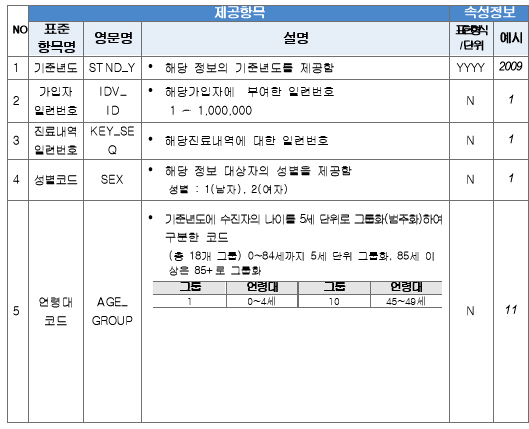
2. 선정 Data 소개

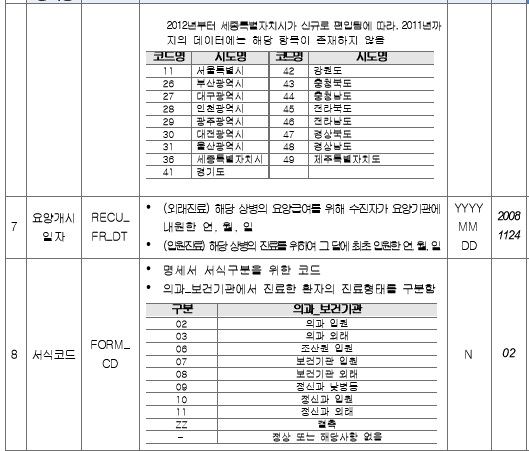
진료내역정보는 국민건강보험 가입자 중 요양기관(병, 의원 등)으로부터의 진료이력이 있는 각 연도별 수진자 100만 명에 대한 기본정보(성, 연령대, 시도코드 등)와 진료내역(진료과목코드, 주상병 코드, 요양일수, 총처방일수 등)으로 구성된 개방데이터이다. 총 19개의 변수로 가입자 일련번호와 진료내역 일련번호 1) 수진자 기본정보 : 성, 연령, 거주지 시도코드와 같은 기본정보 2) 진료상세 정보 : 주상병, 부상병, 요양일수, 입.내원일수, 총 처방일수 등 구체적인 진료내역 정보 3)요양급여 청구 심사 결과 : 요양 급여비용 심사청구 결과 결정된 금액 등에 대한 정보로 구성되어있다.

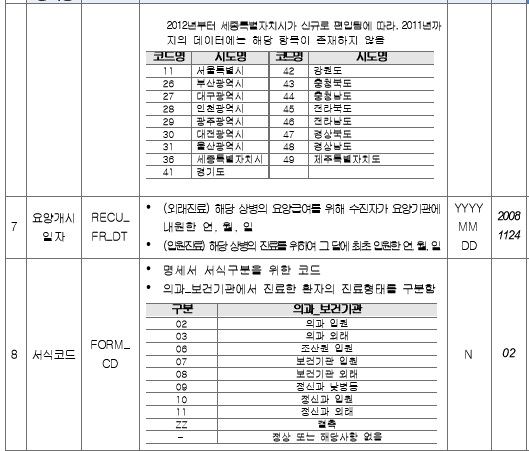
|  |
| --- |
| Data Set in detail   * 진료내역 : 총 11,440,376 items, 19 attributes, * 진료과목 : 총 48 items, 2 attributes * 시도코드 : 총 17 items, 2 attributes |

상세한 내역은 국민건강정보데이터 사용자 메뉴얼의 데이터 설명을 가져왔으며, 본 분석에서는 요양급여 관련 정보는 사용하지않았다. 그리고 일련번호로 이루어진 진료과목코드, 시도코드, 주상병코드 는 별도의 테이블을 만들어 조인 후 텍스트형태의 차원값으로 사용하였다. 더하여 분석의 용이성을 위해 아래와 같은 데이터 재구성을 조치했다.

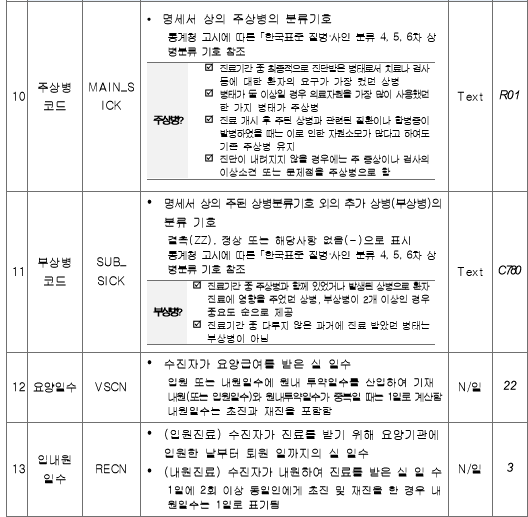
* 5세단위로 그룹화된 연령대코드를 10세 단위로 재그룹화
* 대,중,소,세,세세 단계로 나누어진 주상병코드를 호흡기 질환은 중 혹은 소 단위로 그룹화하고 분석범위 외에 있는 질환들은(정신장애, 선천기형, 근골격계 질환, 비뇨기 생식기 질환 등) 대분류로 통합그룹화

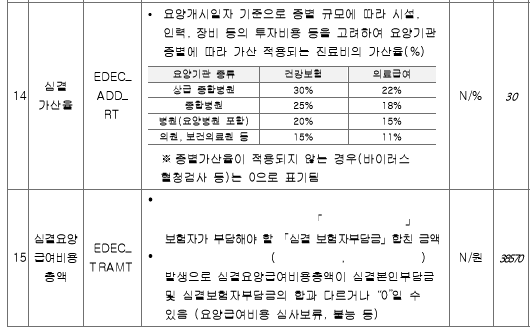


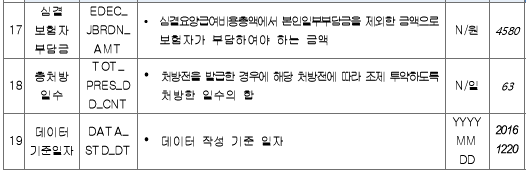








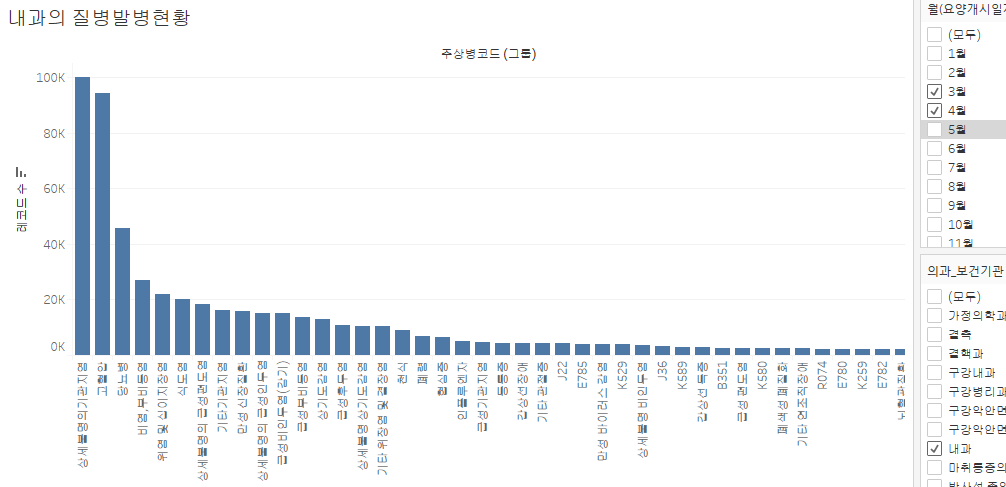




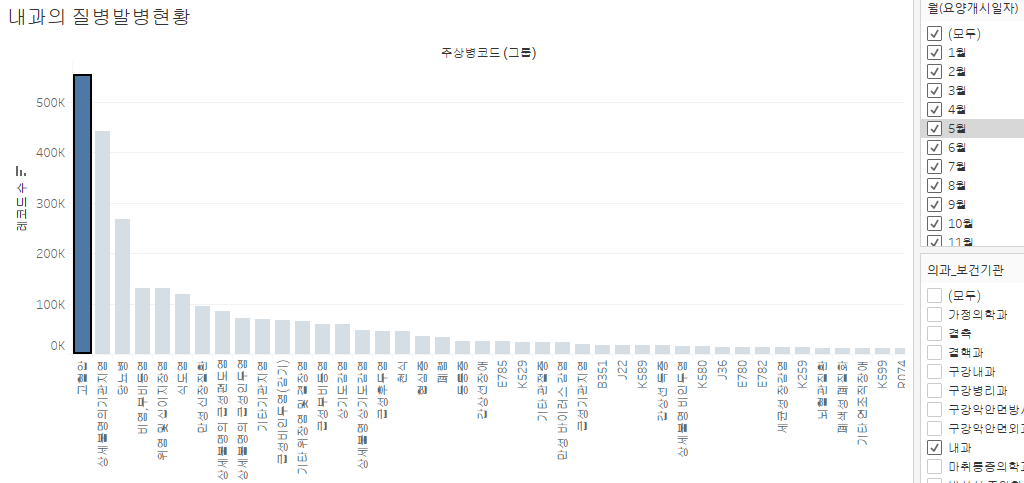
3. 1차 발견

1. **“15년 3~4월에는 내과의 최다질병인 고혈압보다 상세불명의 기관지염으로 병원을 더 많이 찾았다.”**

가장 진료내역수가 많은 내과에 대한 분석 중, 15년 기간전체<그림7>를 보면 고혈압이 1위이나, 황사가 많이 발생한다고 생각되는 3,4월(요양개시일기준)을 기준으로 봤을 때는 그림 6과 같이 상세불명의 급성기관지염이 고혈압보다 더 많이 발병했음을 알 수 있다.



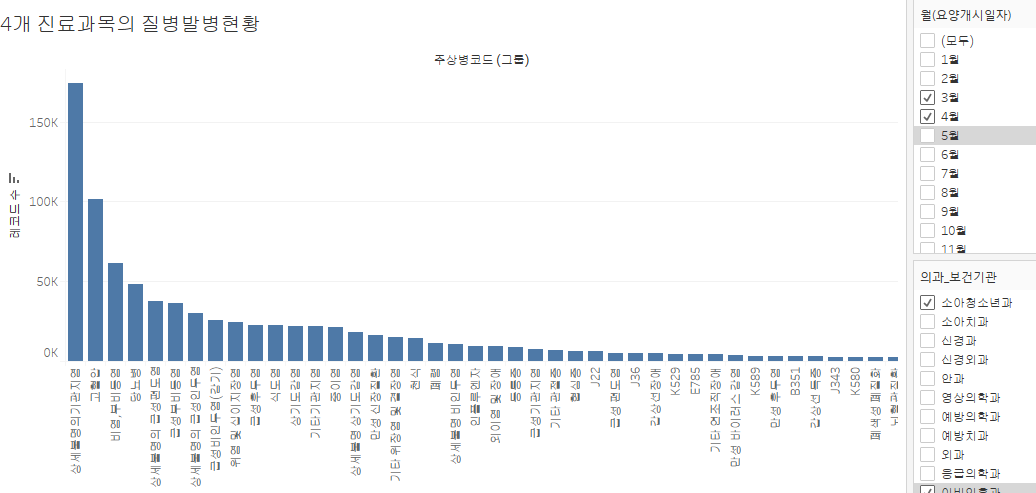
<그림 6>



<그림7>

**2) “호흡기계통의 진단이 가능한 4개 진료과목(내과, 이비인후과, 소아청소년과, 가정의학과) 에서 상세불명의 기관지염에 대한 3~4월 사이의 발병률은 고혈압보다 45% 이상 증가했다.”**

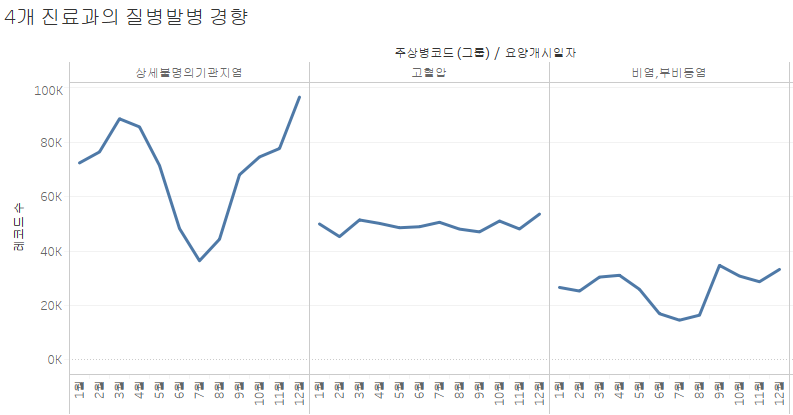
그렇다면 내과 외에 호흡기 계통에 대한 질병진단이 가능한 주요 3개 진료과목 (이비인후과, 소아청소년과, 가정의학과)을 추가하여 해당 경향에 변함이 없는지 살펴보았다. 내과 1개 과목에 대한 그래프에 비해 1위와 2위의 차이가 매우 벌어졌음을 알 수 있다.



<그림8>

**3) “ 4개 진료과목에서 가장 많이 발생한 상세불명의 기관지염은 봄(뿐 아니라)과 가을에 많이 발생하는 경향이 있다.”**

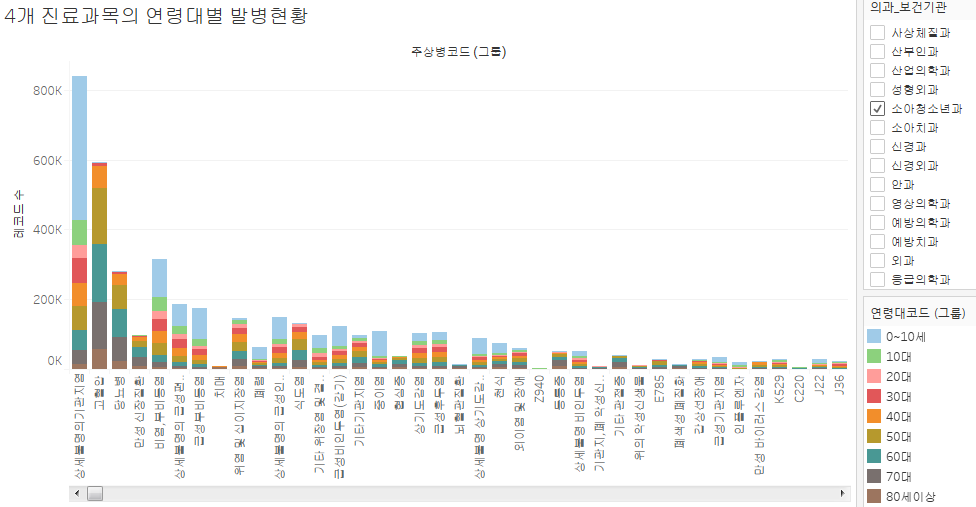
4개 과목에 대한 질병발병경향을 3~4월에 제한하지 않고 전체 기간으로 했을 때 그림9와 같이 봄뿐 아니라 가을에도 상세불명의 급성기관지염에 대한 발병이 증가했음을 확인할 수 있었다. 전체적으로 진단횟수가 고르게 분포하는 고혈압과 달리 호흡기 질환은 계절을 많이 타고있다. 혹 미세먼지뿐 아니라 이는 환절기영향(심한 온도차이)도 있는 것은 아닌지 추가 분석을 요하는 대목이다.



<그림9>

**4) “4개 진료과목에서 가장 많이 발병한 상세불명의 기관지염으로 10세미만의 연령대의 아이들이 가장 많이 치료를 받았다**”

4개 진료과목에 대해 연령대별 분포를 확인해보았다. 1위인 상세불명의 기관지염에 대해서 10세 미만의 질병 발병횟수가 압도적임을 알 수 있다<그림10>. 다른 해의 데이터를 분석해보지는 않았지만 문득 가습기 살균제 사건을 떠올리게 하는 결과였다.

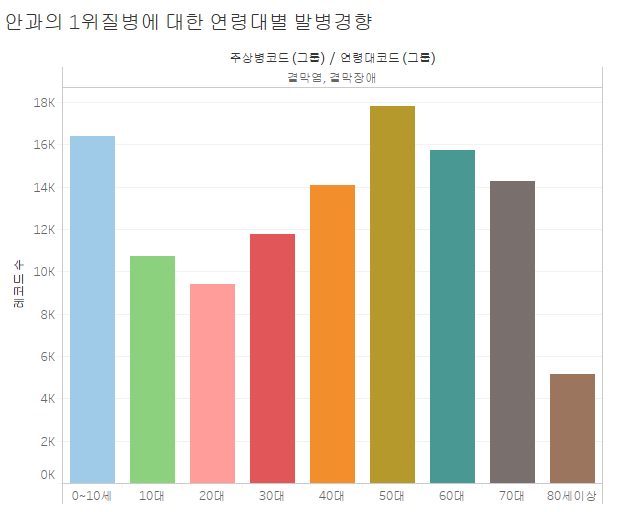


<그림10>

4. 2차 분석

1. **“호흡기외 미세먼지의 직접적인 영향을 받는 안구관련 질환 중 최다 발병은 결막염이며, 미세먼지(황사 등)가 많이 발생하는 봄,가을에 10세 미만의 아이들이 이 질병으로 인해 급격히 병원을 많이 찾았다”**

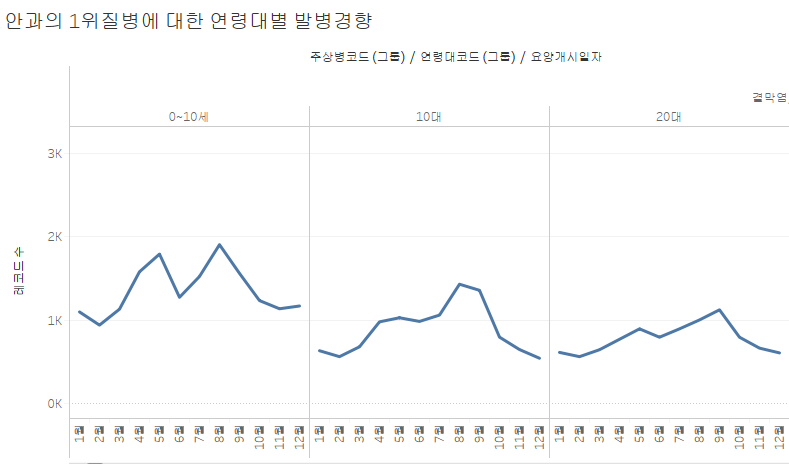
미세먼지가 직접적으로 영향을 미치는 신체기관 중 호흡기 외 안구관련 질환에 대한 데이터를 살펴보았다. 먼저 안과에서 진단된 질환 중 가장 많은 것은 결막염 류의 질환이다. 이를 연령대별로 발병경향을 살펴보기위해 아래와 같은 그래프를 그려보았다. 50대 , 10세미만, 60대 순으로 발병횟수가 가장 많다.

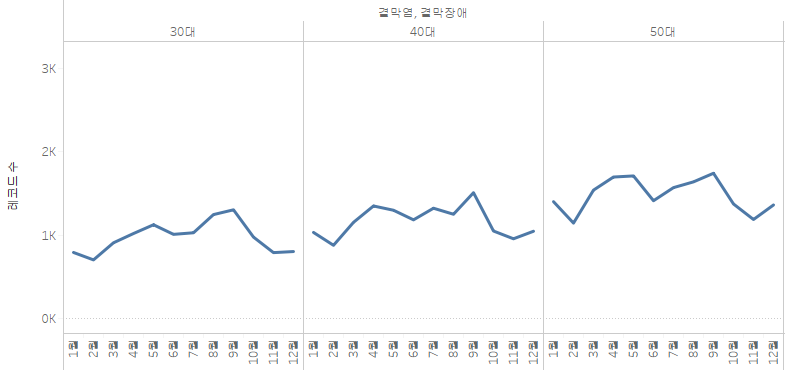


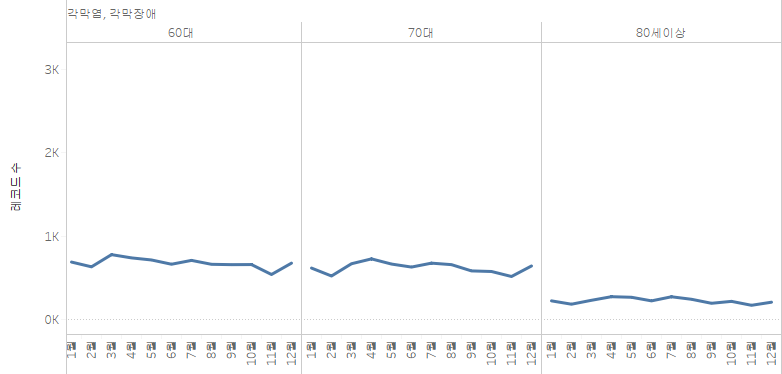
<그림11>

월별 발병 경향도 살펴보았다<그림12>. 안구질환과 환절기의 영향이 적다고 가정할 때, 봄 가을에 황사 등이 주로 많이 발생하는 현상과 연관하면 이에 가장 많은 영향을 받는 연령대는 10세미만이다. 이는 상세불명의 급성기관지염의 절반 이상을 차지하는 10세 미만의 발병횟수와도 연관이 있을 듯 싶다.

60세 이상의 어르신들의 발병경향이 낮은 이유는 안구기능의 저하로 불편함을 덜 느끼거나 기타 질병에 대한 치료로 인해 불편함이 상대적으로 낮은 질병에 대해서는 치료에 덜 신경을(병원을 덜 찾는) 쓰는 것은 아닐지도 생각해보았다.





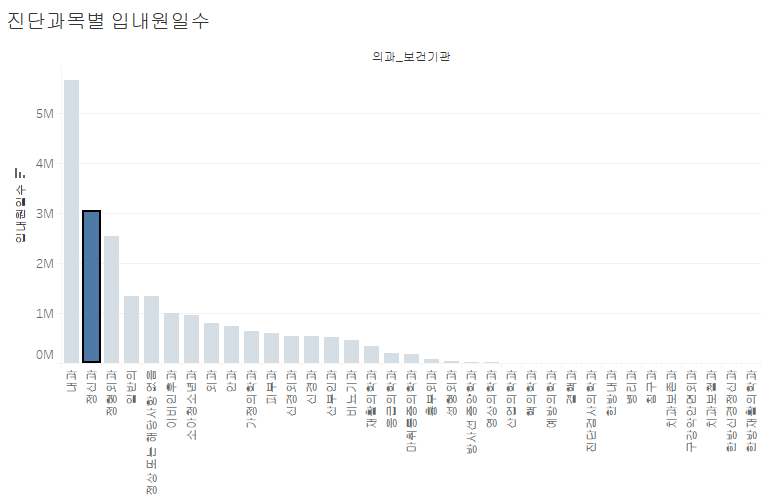


<그림12>

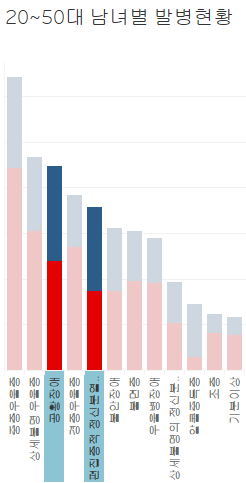
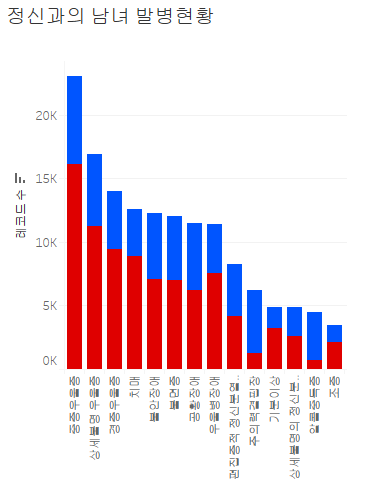
1. **“실제 총 진료일수를 기준으로는(입원일 포함) 정신과 질병이 급격한 순위 변화를 보인다(12->2위).**

진료내역횟수가 아닌 진료후 입원을 했을 경우 치료기간이 늘어남을 고려해 실제로 해당 질병을 통해 진료를 받은 입내원일수을 비교해보았다. 진료내역수를 기준으로 한 그림3과 비교하면 정신과가 12위에서 2위로 급격히 상승함을 알 수 있다<그림 13>.

다른 과목보다도 정신과가 내과 다음으로 위치한 사실이 놀라웠다. 정신과 질병의 특성상 타 질환에 비해 장기간의 치료기간이 필요한 것이다. 좀 더 상세히 어떤 질병이 많이 발병했는지, 성별(푸른색은 남, 붉은색은 여)의 차이는 없는지 나타내보았다<그림14>.



<그림13>



<그림14> <그림15>

그림14를 보면 전반적으로 여성이 더 많은 진단을 받은 것으로 보인다. 여성이 남성보다 더 적극적으로 병원을 찾았을 수도 있고 혹은 실제로 여성의 우울증 발병이 남성보다 많을 수도 있겠다. 이 중 주의력 결핍장애와 알콜중독증은 남성에게서 더 많이 발생했음을 알 수 있다.

활발히 경제활동을 해야하는 20~50대로 필터링하여 그래프를 그리자<그림15>, 공황장애, 편집증적 정신분열증이 상위로 이동함을 볼 수 있었는데 사회생활 및 생계를 꾸리기 위한 다양한 스트레스로 인해 남성의 정신과 질병이 늘어나고 사회생활에 연관된 정신과 질병이 증가하는 것은 아닐지 추측해보았다.

5. 정리

시각화와 데이터분석의 그 사이를 넘나드는 것은 매우 스릴있었다. 시각화과목에서 시각화에 집중해야함에도 불구하고, 엔지니어적인 습성을 버리지못하고 숫자에 매달리되는 것을 경험하며 이후 데이터에 대한 본인의 역할을 명확히 해야 효율적이고 효과적인 task수행결과를 얻을 수 있다는 것을 간접적으로 알게 된 과제 수행이었다. 시각화를 위해 데이터분석을 하는 것인지 데이터분석을 잘하기위해 시각화를 하는 것인지 그 선후관계를 명확히 알기는 어려웠지만 tableau와 spotfire 두 tool을 써보면서 조금은 비교할 수 있었다.

**[성능]**

먼저 raw data를 처음 분석하기 시작한 것은 spotfire였다. Dataset의 크기가 1천만건 이상이었는데 tableau는 추출모드로 하고 분석을 했음에도 매 시각화 시도보다 query시간이 매우 느렸다. Spotfire의 경우 한번 원본 데이터를 로딩하고 난 이후에는 채널이나 마크를 수시로 바꿔가며 분석을 해도 기다리는데 크게 문제가 없었다(10s이하). 때문에 기본적인 초벌 분석 즉 인사이트를 얻기위한 준비단계에 대한 다량의 데이터 분석에는 spotfire가 더 적합하다고 생각된다.

**[사용성]**

방금 언급한 것과 같이 spotfire는 시각화보다는 데이터분석에 좀 더 초점을 맞춘 느낌이 강했다. Tableau의 경우 데이터에 oriented된 분석보다는 시각화에서 인사이트를 바로 찾기위해 도움을 준다는 느낌이 강했는데, 특히 UI구성자체가 마크라던지 채널이라던지 시각화를 염두에 두고 구성이 되어있어 다양한 표현을 시도해보기에 용이했다. Spotfire는 그런 점에 있어 메뉴기반의 도구이다보니 사용자입장에서 어디에 어떤 메뉴가 있는지 능숙하지 않은 상태에서는 활용하기가 매우 어려운 점이 있었다.

**[데이터조작]**

Tableu는 그룹화, 필터 등을 통해 데이터를 조작하는 방법을 바로 인지할 수 있도록 구성되어 있었다. 또한 마킹된 데이터만 바로 필터링하여 볼 수 있는 부분이 편리했다. 그에 비해 spotfire 는 그룹화 메뉴를 찾기 어려웠고, 원하는 필터만 조정하기에는 조잡스러운 면이 있었다. 그러나 마킹된 데이터를 상세시각화하고 데이터간 링크를 맺어 인터랙티브한 시각화를 볼 수 있는 것은 spotfire가 간편했다. tableu에서는 spotfire와 같이 한 페이지에 여러 그래프를 놓고 연관관계를 볼 수 있는 기능을 찾지못했다.

이후 시간이 된다면 미세먼지 pm2.5이하 농도에 대한 측정데이터와 이번 분석과의 연관관계나 기타 진료과목에 대한 연령대별 특이사항, 연령대별 라이프스타일 등에 대한 데이터와의 연계분석등을 해보면 재미있겠다는 생각이 든다. 본 과제의 경험을 바탕으로 기말프로젝트 시 데이터 분석에 도구사용에 많은 도움이 될 것이다. 또한 업무 상 이러한 도구를 사용하여 의미있는 인사이트를 얻을 수 있는 기회를 찾고자 한다.

1. 공기 중의 부유하는 공기를 흡입하여 직경이 10㎛ 이하인 먼지(황사 포함). 초미세먼지는 PM2.5 [↑](#footnote-ref-1)
2. 기상자료개방포털 https://data.kma.go.kr/data/climate/selectDustList.do?pgmNo=67 [↑](#footnote-ref-2)
3. 공공데이터포털 https://www.data.go.kr/dataset/15007115/fileData.do [↑](#footnote-ref-3)
4. 진료내역일련번호는 중복되지않는 값이다. [↑](#footnote-ref-4)
5. 05년 인구수 47,278,951명, 15년 인구수 51,069,375 명 국가통계포털(http://www.kosis.kr) [↑](#footnote-ref-5)
6. 제6차 한국표준질병사인분류표 https://kssc.kostat.go.kr:8443/ksscNew\_web/index.jsp# [↑](#footnote-ref-6)