데이터 사이언스 개인 중간보고서

2015122055 응용통계학과 정규형

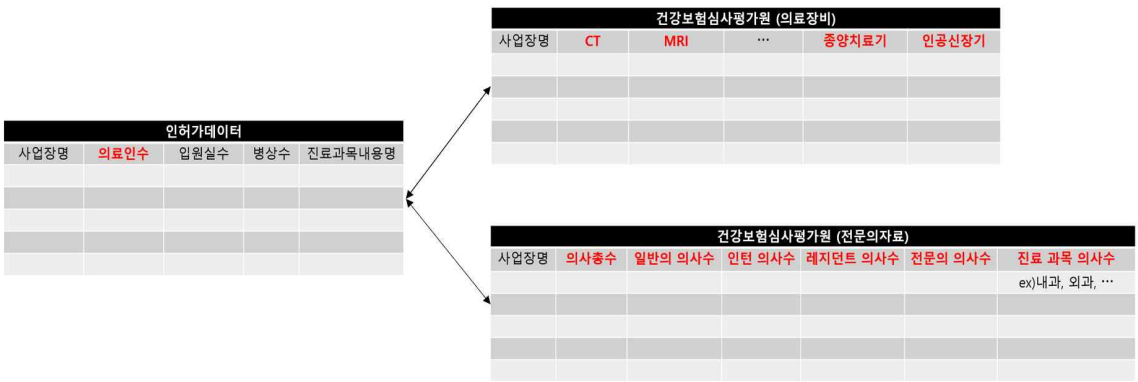
1. 개요

팀 프로젝트의 경우 병원 관련 데이터와 평점 데이터 통합 과정을 진행했다. 현재 평점 관련 변수 EDA에 진행 중이며 최종적으로 다양한 변수를 통해 진료만족도, 의료진 친절도 등의 점수를 채운 완전한 데이터를 생성하는 것이 목표이다.

개인 프로젝트에서는 평점을 제외한 두 데이터 간의 관계를 파악하고 오류가 존재하지 않는지 확인해보고자 한다. 공공데이터의 경우 데이터 탐색을 진행하지 않고 분석을 진행한 결과 잘못 입력된 값들로 예상치 못한 결과를 얻은 경험이 있다. 이번 기회를 통해 공공데이터의 품질을 높이는 방안에 대해 고민해 볼 것이다.

2. Research Question

1) 의료 인력 속성값의 차이



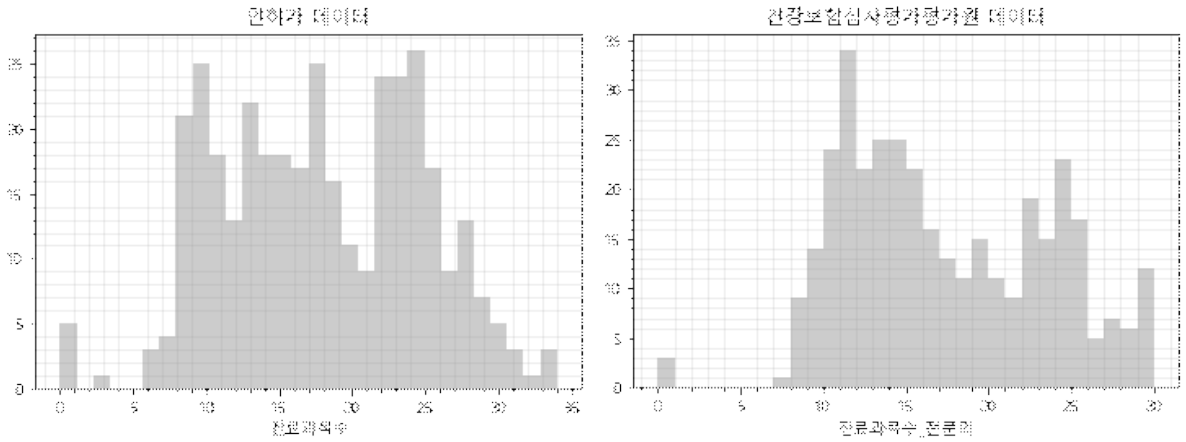
그림[1]

평점 데이터를 제외한 인허가 데이터(http://localdata.go.kr/main.do)와 건강보험심사평가원(https://www.data.go.kr/data/15051059/fileData.do)의 통합 과정은 그림[1]과 같다. 사업장명을 기준으로 통합하였으며 이 과정에서 의료인 관련 중복 속성(붉은색 강조 부분)이 존재하게 되었다.

간략하게 나타난 인허가 데이터의 의료인 수[1]에 비해 건강보험심사평가평가원의 전문의 자료에서는 세부적인 진료 과목별 의사 수 확인이 가능하다.

데이터 통합 과정에서 의료인 수가 가장 큰 집단임에도 불구하고 전문의 자료의 의사 총수가 더 큰 경우가 존재했다. 심지어 의료인 수 일부 값이 0으로 나타나고 있어 수정될 필요가 있다. 또한, 전문의 의사 수 합계와 진료 과목별 의사 수의 합이 일치하지 않는 병원이 발견되었다.

추가로 건강보험심사평가원의 자체 데이터에서도 의사 총수와 세부 과목별 의사 수가 일치하지 않는 병원이 존재한다. 의사 수가 실제보다 적은 인원이 입력된 것도 데이터의 품질 저하를 일으키지만, 만약 실제 존재하지 않는 의사가 특정 과목의 전문의로 기재되어 있다면 이는 큰 문제일 것이다. 진료과목은 상급종합병원[2], 종합병원 요건에 명시되어 있을 만큼 중요한 부분이기 때문이다.



종합병원 요건에 의하면 100개 이상의 병상을 갖추고 최소 7개 이상의 진료과목을 갖추고 각 진료 과목별 전문의를 두어야 하지만 각각의 데이터에서 5개 이하의 값이 나타나고 있으며 두 분포 또한 다르게 나타나고 있다. 우선으로 잘못된 값에 대한 수정을 진행할 것이며 두 데이터를 통해 실제 진료과목 수를 알아보고자 한다.

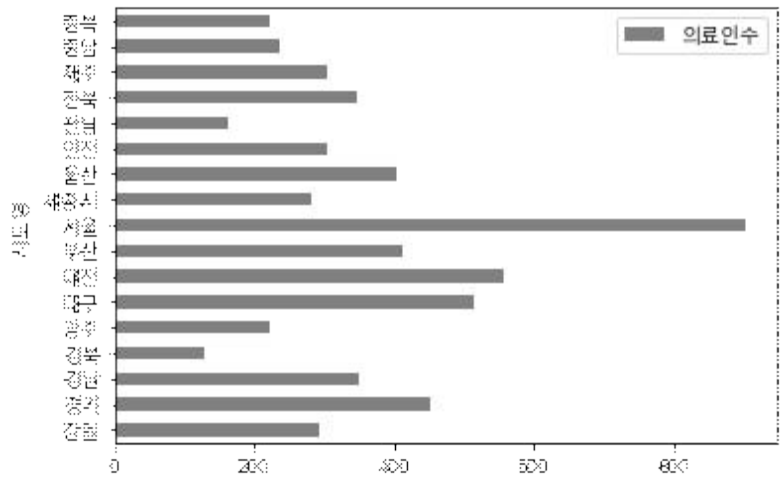
---

1) 해당 병원에서 입력한 값이며 의료법상 의사, 치과의사, 한의사. 조산사, 간호사가 의료인에 해당

2) 의료법 제3조의4 제1항) 보건복지부령으로 정하는 20개 이상의 진료과목을 갖추고 각 진료 과목별 전속하는 전문의를 둘 것, 제3조의3(종합병원)의 내용은 생략

2) 간호인 수 예측 및 정원 준수 현황

팀 프로젝트의 첫 목표는 시도별 인구와 병상 수의 관계를 통해 부족한 곳을 파악하고 해결방안을 탐색하는 것으로 했지만 이미 병상 수는 포화상태로 나타나 목표를 변경했다. 하지만 기준 인구 1000명당 간호 인력은 6.9명으로 OECD 평균(9명)보다 23% 낮은 수치이다. '활동' 간호사 수 인구는 이보다 더 적을 것이다.



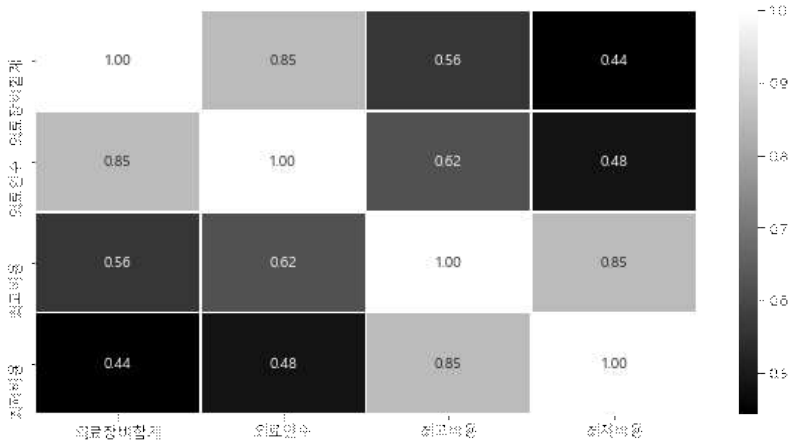
특히, 코로나 19라는 악성 바이러스가 나타난 현시점에서 이 데이터를 통해 위 궁금증을 해결해보고자 한다. 한 기사에 의하면 지방으로 갈수록 간호사의 처우는 더욱 열악하다고 한다. 인허가 데이터의 의료인 수를 확인해본 결과 실제로 서울시의 의료인 수가 가장 많다. 지방의 의료인 수는 서울시보다 현저히 낮게 나타나고 있다. 하지만 의료인 수가 어느 범위까지의 의료인을 포함하고 있는지 알 수 없기에 이는 추가적인 데이터를 통해 의료인 수의 기준을 잡아야 한다.

이는 의료 인력 현황(http://opendata.hira.or.kr/)을 통해 해결할 것이다. 분기별, 요양 기관 종별, 지역별로 의료 인력을 확인할 수 있다. 이를 통해 의료인 범위를 어느 정도 추정할 수 있을 것이다.

최종적으로 인허가 데이터에 나타난 병상 수와 간호인 수, 의사 수 간의 관계를 파악하고 정원이 적절하게 배정되어 있는지 판단해볼 것이다. 또한, 역으로 종합병원에서 간호 인력을 지나치게 갖추고 있는 것은 아닌지도 확인해볼 요소 중 하나이다.

3) 병원별 상급병실료(1인실)의 적절 여부 판단

건강보험심사평가원 비급여진료비정보에 의하면 통합데이터에 나타난 병원들의 1인실 최저비용은 205,936원, 최고비용은 234,228원이다. 병실료는 간호등급제[3], 의료장비, 의사 수 등의 다양한 속성을 통해 정해진다.



통합데이터를 활용한 상관관계는 위와 같다. 예상대로 대부분 높게 나타나고 있는 것을 볼 수 있다. 이는 병원별 상급병실료가 적절하게 책정되어 있는지 판단하는 기준이 될 수 있을 것이다. 상급병실료의 비용이 적지 않은 만큼 위 기준뿐 아니라 통합데이터에 주어진 다양한 속성을 파악하여 가장 적절한 기준을 세워보고자 한다.

3. 방법론 탐색

의료 인력 속성값 차이 문제를 해결하는 과정에서 '데이터 통합과 정보보호' 수업에서 배운 다양한 방법을 적용할 것이다. 이를 통해 값 차이 문제를 해결하고 더 나아가 NA값을 채우는 작업을 진행할 것이다.

완전하게 완성된 데이터를 통해 간호인 수 정원 준수 현황 및 상급병실료 적절 여부 문제를 해결해보고자 한다. 이 과정에서는 '회귀분석'의 선형회귀 모델, '딥러닝' 강의 내용 중 일부인 CNN, 마지막으로 머신러닝 기법(XGBoost, LightGBM and CatBoost)을 적용하여 가장 우수한 모델을 택할 것이다.

---

3)http://www.mohw.go.kr/upload/viewer/skin/doc.html?fn=1245291534078_20090618111854.hwp&rs=/upload/viewer/result/202011/ (붙임1 참고)