데이터 사이언스 3조 팀 프로젝트 최종 보고서

3조 이상익, 이진우, 정규형

1. 주제선정

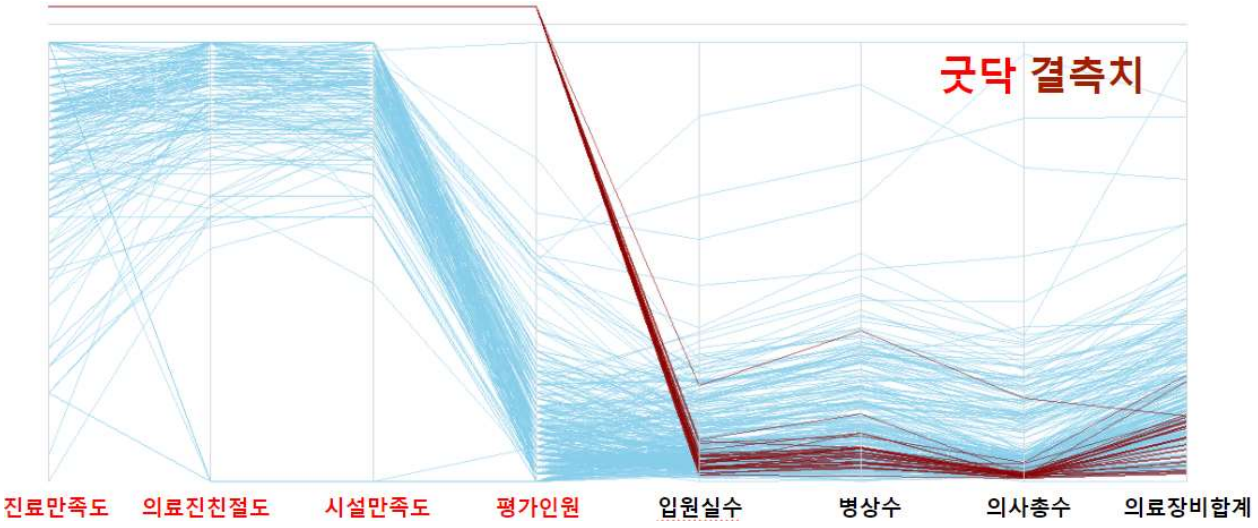
- 일반적인 소비 상품을 구매하기 앞서 타사의 제품과 쉽게 비교할 수 있지만, 종합병원의 경우 어떤 병원이 어떤 진료과목을 주로 가지는지 다른 병원과 비교하며 한 번에 알아보기 쉽지 않다.
- 지방행정 인허가 데이터개방 사이트나 건강보험심사평가원 사이트 등을 통하여 여러 종합병원들의 정량적 정보(진료과, 병상수 등)를 확인할 수 있지만 - 각 사이트가 서로 다른 종류의 정보를 갖고 있어 각 정보들을 서로 통합할 필요가 있다.
- 병원에 대한 온라인 리뷰 또한 하나의 포털사이트가 아닌 여러 사이트에 기록되어있다. 각 포털사이트는 두 가지 유형으로 나뉘는데, 진료만족도/ 의료진친절도/ 시설만족도와 같이 세부 평가항목을 보유하되 리뷰건수가 적거나 아예 없는 사이트와 단순 평점만을 제공하지만 리뷰건수가 많은 사이트로 구분된다.
- 단순 평점만 기록된 정보로는 구체적 사유를 알기 어렵기에 세부 평가항목이 의료서비스 이용자에게 큰 도움이 된다. 하지만 리뷰건수가 적거나 결측치가 많기 때문에 다른 정보들(병원의 정량적 정보, 단순 평점)을 이용하여 보완할 필요가 있다.
- 따라서 병원의 여러 정량적 정보들을 종합적으로, 그리고 타 병원과도 비교할 수 있게 데이터를 통합하고, 보완된 세부항목 평점도 함께 통합하여 - 정보 불균형 상황에 놓이기 쉬운 의료서비스 이용자에게 합리적인 의료기관 선택에 도움을 줄 수 있을 것으로 기대한다. 또한 이용자 뿐 아니라 의료기관 서비스 제공자의 입장에서도 해당 기관이나 타 유사의료기관의 이용자들의 평점을 피드백 지표로 확인하고, 자체 성과 평가지표로 사용하여 의료 시설/기기확충이나 의료진 교육 등 의료기관 운영 기반 자료로 활용한다면 의료서비스의 퀄리티를 높이는데 도움을 줄 수 있을 것으로 보인다.

2. 데이터 수집 이후 첫 번째 데이터 모식도 및 EDA

<table>
<tr><th>사업장명</th><th>병원 기본 속성</th><th>병원 세부 속성</th><th>네이버리뷰</th><th>구글리뷰</th><th>다음리뷰</th><th>진료만족도</th><th>의료진친절도</th><th>시설만족도</th></tr>
<tr><td>비에스종합병원</td><td rowspan="16">병원 인허가 데이터 정보
(도로명전체주소, 입원실 수, 병상 수, 좌표 등)</td><td rowspan="16">건강보험심사평가원 병원정보
(진료과목, 각 과목별 전문의 수, 의료장비 정보 등)</td><td rowspan="15">네이버 평점
크롤링 데이터
(평점, 건수)</td><td rowspan="13">구글 평점
크롤링 데이터
(평점, 건수)</td><td rowspan="13">다음 평점
크롤링 데이터
(평점, 건수)</td><td colspan="3" rowspan="11">굿닥 평점
크롤링 데이터
(평점, 건수)</td></tr>
<tr><td>이화여자대학교의과대학부속서울병원</td></tr>
<tr><td>학교법인연세대학교의과대학세브란스병원</td></tr>
<tr><td>의료법인동신의료재단 동신병원</td></tr>
<tr><td>서울특별시서울의료원</td></tr>
<tr><td>성애의료재단 성애병원</td></tr>
<tr><td>의료법인 청구성심병원</td></tr>
<tr><td>순천향대학교 부속 서울병원</td></tr>
<tr><td>강북삼성병원</td></tr>
<tr><td></td></tr>
<tr><td></td></tr>
<tr><td></td></tr>
<tr><td>의료법인 힘찬의료재단 옥천성모병원</td><td colspan="3">NA</td></tr>
<tr><td>의료법인인화재단한국병원</td><td rowspan="5">NA</td><td rowspan="3"></td></tr>
<tr><td>청주성모병원</td></tr>
<tr><td>의료법인정산의료재단효성병원</td><td colspan="3">NA</td></tr>
<tr><td>한림대학교동탄성심병원</td><td rowspan="2"></td></tr>
<tr><td>당진종합병원</td></tr>
<tr><td>재단법인아산사회복지재단보령아산병원</td><td>NA</td><td></td><td colspan="3">NA</td></tr>
</table>

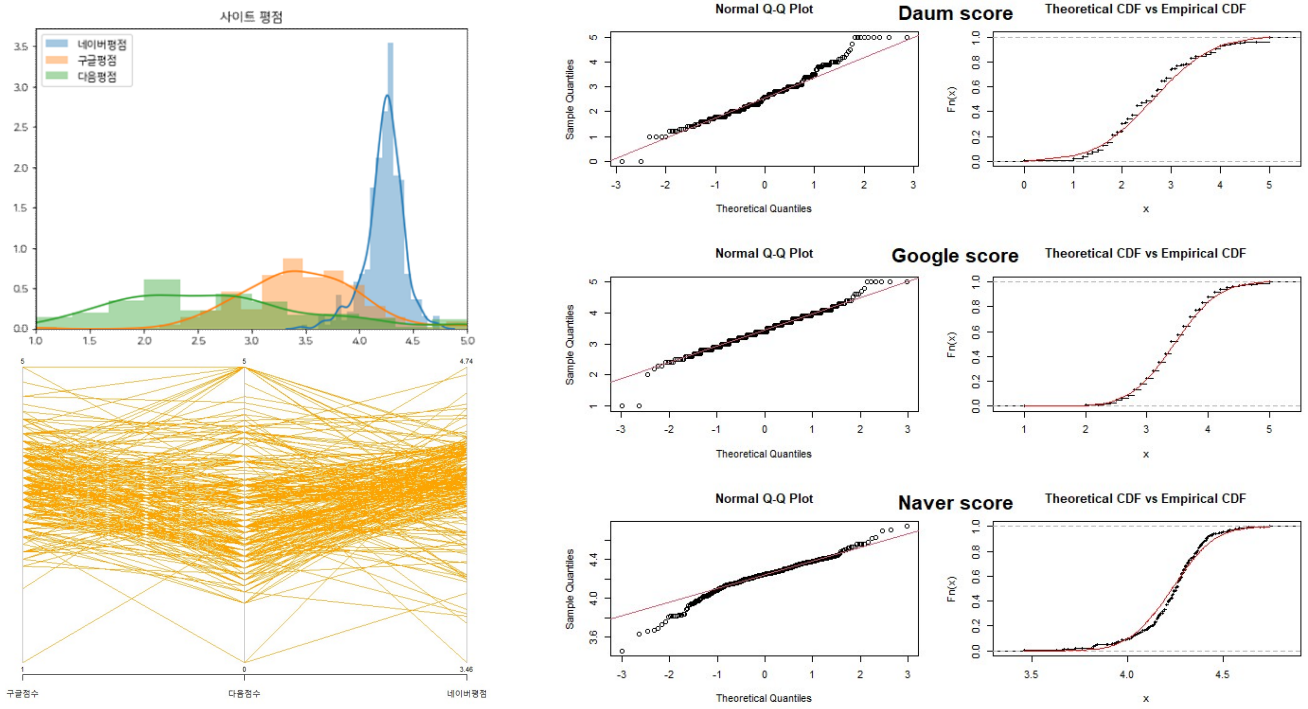
각 지방행정 인허가 데이터개방 사이트와 건강보험심사평가원을 통해 병원의 기본 속성(주소, 입원실수, 총 의료진수 등)과 세부 특징(진료과목과 해당 전문의 수, 특정 의료장비 수 등)을 알 수 있으며, 굿닥 사이트로부터 세부 평가항목의 평점과 건수를, 나머지 세 곳의 포탈사이트로부터는 단순평점과 건수를 알 수 있다.

<그림 1: 몇몇 정량적 특징과 굿닥 결측치의 평행 좌표>



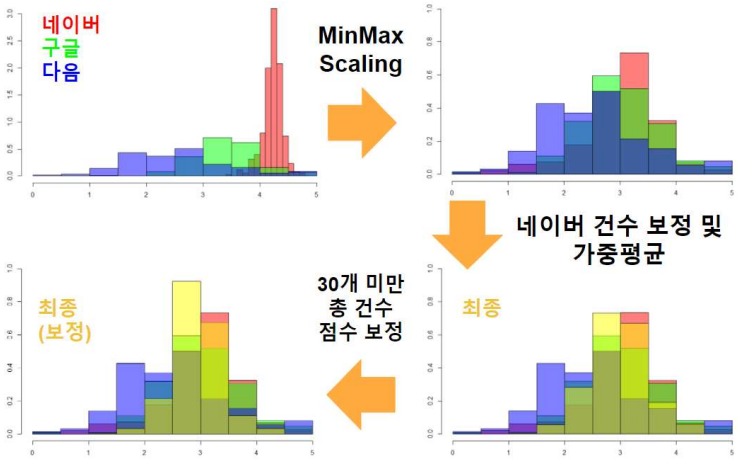
세부 평가항목을 가지는 굿닥 사이트에서 평가가 아예 없는(결측치) 병원들의 특징은 정량적 재원들이 상대적으로 적다는 것이다. 따라서 이후에 굿닥 결측치를 imputation하는 과정에서 이러한 정량적 특징이 크게 반영될 것으로 보인다.

<그림 2: 세 포털 사이트의 분포, 평행 좌표, 그리고 Q-Q plot>



반면 단순 평점만을 가지는 세 곳의 포털사이트의 결측치들은 정략적 재원과 비교했을 때 어떠한 패턴을 보이지 않아 결측치를 채우기 쉽지 않다. 대신 세 포털사이트의 평점분포는 모두 정규분포의 형태를 따르고 서로 뚜렷한 연관성을 보이지 않기 때문에 - 세 사이트의 평점을 하나로 통합하여 결측치 문제를 해결할 수 있다.

3. 네이버, 구글, 다음 평점 통합 과정



	Min	Mean	Max
구글 건수	0	31	336
다음 건수	0	14	142
네이버 건수	4	263	2451
네이버 건수 (보정 이후)	0	31	300

세 포털사이트는 다음과 같은 항목에서 서로 차이를 보인다 :

1. 다음과 구글에 비해 네이버의 평균이 월등히 높다. 이는 네이버가 '마이플레이스 영수증 리뷰'라는 서비스로 인센티브를 주기 때문에 그에 영향받은 것으로 보인다.
2. 네이버의 리뷰건수 또한 구글과 다음에 비해 월등하게 많다.

이 상태 그대로 사이트별 가중평균내어 최종점수를 만든다면 네이버 분포에 편향될 수 밖에 없다. 따라서 다음과 같은 과정을 거쳐 최종점수를 만든다 :

1. 세 사이트 모두 [0,5]로 MinMax scaling
2. 네이버 건수를 구글의 건수 분포와 비슷하게 MinMax scaling
   가. 건수의 Standard를 구글로 잡았다.
   나. 다음 건수는 적은 편인데다 억지로 scale을 늘린다면 과대해석이 일어나므로 그대로 둔다.
3. 각 사이트의 건수로 가중평균내어 최종점수를 만든다.

$$ith\ Final\ Score = \frac{Daum\ \#_i}{Total\ \#_i} \cdot Daum\ Score_i + \frac{Google\ \#_i}{Total\ \#_i} \cdot Google\ Score_i + \frac{Naver\ \#_i}{Total\ \#_i} \cdot Naver\ Score_i$$

최종점수의 건수(다음건수+구글건수+네이버건수)범위가 min값 3부터 max값 746까지 매우 크며, 30건수 미만이 약 25%정도 된다. 주관적 판단으로 평점이 신뢰할만한 기준 건수를 30으로 결정하여 그보다 적은 건수의 평점들을 그 건수만큼 보정해준다 :

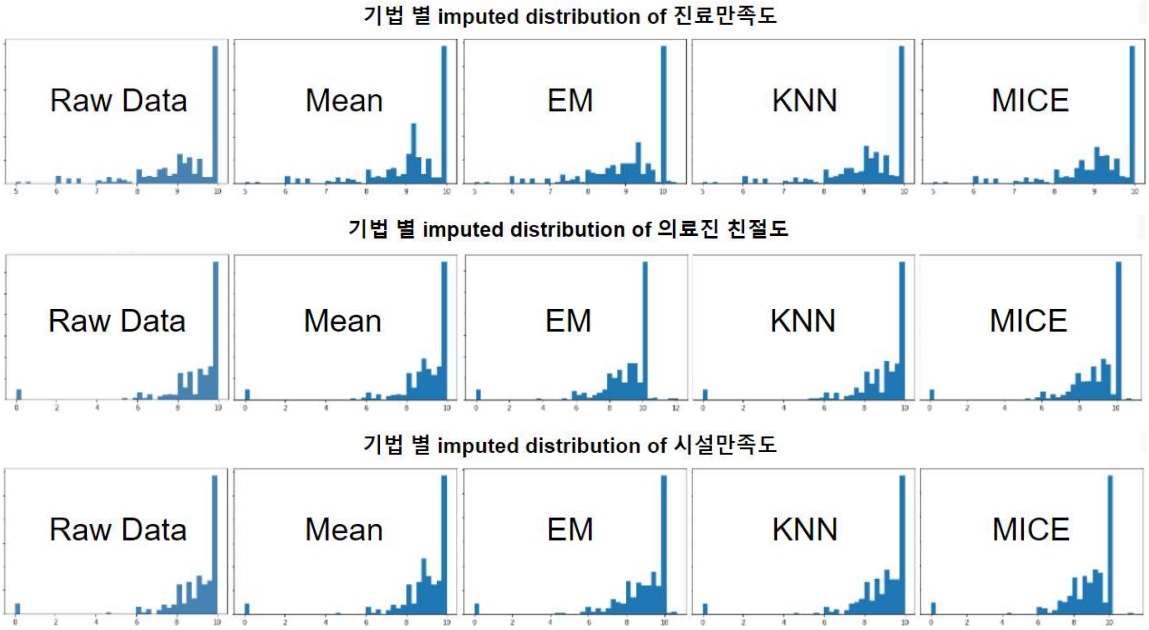
30건 미만 중 3점보다 높은 경우 : $ith\ 최종점수_2 = ith\ 최종점수 \times -\exp(-\frac{ith\ 총건수}{29} - 1.5) + 1$

30건 미만 중 3점보다 낮은 경우 : $ith\ 최종점수_2 = ith\ 최종점수 \times \exp(-\frac{ith\ 총건수}{29}) + 0.7$

지수함수의 강점은 건수가 적을수록 보정을 강하게 줄 수 있으며, 지수함수 이외의 숫자들은 상한, 하한선, 그리고 보정 속도를 정해준 것이다. 이때 threshold를 3으로 잡아주므로써 높은 점수가 심하게 패널티 되어도 적어도 3점 아래로 떨어지지 않게, 반대로 낮은 점수가 심하게 올라가더라도 3점 위로 올라가지 않게 해준 것이다.

4. 굿닥 결측치 imputation 과정

34개의 종합병원이 굿닥에서 결측값으로 존재하며 이는 약 10%에 해당되는 데이터이다. 이 결측값들을 병원의 기본 속성과 세부 특징, 그리고 앞서 구한 최종점수와 총건수를 이용하여 4가지 NA imputation기법으로 채워보았다 :



- 평균 기법은 Raw data의 분포와 달리 특정 값에서 확 티며, 이는 결측치들을 모두 같은 값으로 대체하기 때문이다. 따라서 개별병원의 특징을 잘 반영하지 못한다.
- MICE의 경우 파라미터 조절을 하여도 종종 10점(상한선)을 넘는 경우가 발생했고 seed값에 따라 분포 변화가 너무 크게 나타나 적절치 못하다고 판단하였다.
- EM의 경우 또한 10점을 넘는 경우가 존재한다.
- KNN의 경우 K값을 8로 정하였을 때 Raw data의 분포와 어느 정도 일치된 분포를 띄고 있다. 이는 결측값 병원들이 피쳐 스페이스에서 자기 자신과 가장 가까운 객체들의 평균으로 imput되기 때문에 - Raw data 분포와 가깝다. 따라서 KNN으로 imput한 값을 사용하기로 하였다.
- 이외에도 고려할 수 있는 imputation method로 Fractional Imputation을 생각해 볼 수 있으나, 단순히 병원의 '정량적' 피쳐들과 평점과의 직접적 연관성을 가정하는 것이 쉽지 않기 때문에 힘들 것으로 보인다. 따라서 보다 소극적인 방법인 - KNN과 같이 Similarity Measure에 기반한 또다른 기법들도 생각해볼 수 있을 것이다.

5. 최종 데이터 모식도 및 앞으로의 계획

최종적으로 통합된 데이터 자체로 앞서 선정한 주제와 부합하지만, 조원들이 이 데이터에 대해 각자 가지는 Research Question을 해결해가는 과정에서 추가 인사이트를 도출하여 주제를 더 보완할 수 있을 것이다. 혹은 이후에 더 나은 NA imputation method나 평점 보정법을 찾아볼 수도 있을 것이다.