데이터 사이언스 2차 중간보고서

3조 이상익, 이진우, 정규형

1. 1차 보고 피드백 반영 프로젝트 수정 사항

1) 다양한 정보를 통해 병원평가를 최종목표로 하였으나 병원에 관련한 다양한 데이터를 얻기 어려우며, 의료관련 배경지식이 부족하기에 개인별 적합 의료시설을 제안하는 방향에서 개별 종합병원의 결측 평점을 예측하는 것으로 목표를 변경하였습니다.

2) 결측 평점의 예측으로 목표를 변경하면서 서울 내 종합병원으로만 범위를 한정 했던 것을 확장하여 전국에 있는 종합병원들을 대상으로 확대하였습니다.

2. 데이터 수집 및 전처리/ 통합 과정

1)건강보험심사평가원 병원정보 (병원별 진료과목별 전문의 수, 주요 의료장비현황)

■데이터 수집과정

공공데이터 건강보험심사평가원 병원정보에서 병원별 진료과목과 해당과목 전문의수, 의료장비등 정보를 구할 수 있었습니다. 건강보험심사평가원에서 부여한 병원고유코드를 이용하여 병원의 기본정보와 병원별로 진료과목별 전문의 수, 병원별 의료장비 수를 정리하여 통합하였습니다.

■데이터 전처리/ 통합시 문제 및 해결과정

인허가 데이터와 건강보험심사평가원 자료 간 병원명이나 주소 등이 입력차이(띄어쓰기, 재단표기여부 등)에 따라 차이가 존재하였으며 주소나 번호등 일부 정보가 미입력 된 병원이 존재하는 등의 문제가 있었습니다. 따라서 병원의 고유정보(주소, 전화번호, 기관명, 우편번호) 일치 시 해당병원에 대한 데이터로 보고 추출하여 통합하였습니다.

2) 네이버, 구글, 다음 포탈 병원 평점

■데이터 수집과정

Naver의 경우 사업장명으로 검색시 결과가 안나오는 경우가 빈번하지만 주소로 검색시 검색되어 주소검색에서 병원이외 항목들은 제외하고 평점과 리뷰갯수 정보를 크롤링 하였습니다. 이때 병원 내 부속시설등 관련된 항목들도 모두 크롤링 하였습니다.

Google과 Daum의 경우 Naver와 달리 사업장명으로 검색해야 해당 항목들이 나왔습니다. 크롤링 과정은 Naver와 마찬가지로 병원 관련 모든항목들의 평점과 리뷰갯수를 크롤링 하였습니다.

■데이터 수집시 문제점 및 해결과정

크롤링한 데이터에서 실제 해당 병원명이나 주소가 일치하지 않는 등 잘못 추출된 자료들이 존재하여 이를 제거하고 동일 병원내 부속시설등 한 병원에 여러 평점이 존재한 경우 리뷰갯수를 기준으로 가중평균하여 포탈별 평점을 수집하였습니다.

3)굿닥 평점

■데이터 수집과정

직접 병원명을 검색하여 평점을 추출하는 과정은 복잡하여 병원의 코드를 추출한 뒤 URL에 병원코드를 넣어 평점을 추출하였습니다.

```
for i in tqdm(hosp1['병원코드']):
    url=str('https://www.goodoc.co.kr/hospitals/')+str(i)
    driver.get(url)
    try:
        rate=driver.find_element_by_class_name('review-content')
        rate=rate.text.split('₩n')
        dia=float(rate[1])

    except:
        dia='error'
    diag.append(dia)
time.sleep(0.1)
```

*) 진료만족도 추출 코드이며 다른 속성도 위와 같이 진행함

■크롤링 과정에서 문제점 및 해결과정

코드를 추출하는 과정에서 동명의 타 병원의 정보가 추출됨. 이는 병원코드와 주소, 병원명을 동시에 추출하여 인허가 데이터에 나타난 주소와 일치하는지 확인하고 불일치하는 데이터에 대해 크롤링을 다시 진행했습니다. 띄어쓰기, 특수기호로 인해 병원코드가 추출되지 않는 경우 구글에 '병원이름 + 굿닥' 형식으로 구글링을 하여 평점을 뽑아낼 수 있었습니다. 또한, 굿닥 홈페이지에서 검색했을 때는 결과가 존재하지 않는데 구글 검색으로 접근가능한 경우가 있어 이를 반영했습니다.

3. 통합데이터 snapshot

인허가데이터+의사/장비 데이터+구글+다음+네이버+굿닥

C	D	E	F	G	H	I	J	Q	R	S	T	U	V	W	X	Y	BS	BU	BV	CF	CG	CH	CI	CJ	CK	CL	CM	CN	CO
인허가일자	도로명전체주소	사업장명	의료인수	입원실수	병상수	총면적	진료과목내용명	X좌표	Y좌표	의사총수	일반의 의사수	인턴 의사수	레지던트 의사수	전문의합계	내과	신경과	의료장비합계	CT	MRI	구글점수	구글건수	다음점수	다음건수	네이버점수	네이버건수	진료만족도	대표진찰료	시설만족도	평가인원
20181106	인천광역시 강화군 강화비에스종합병원		14	55	213	12110.61		126.48599	37.736048	20	1	0	0	18	4	0	33	1	1	3.1	14	4.1	7	3.95	47				
20181226	서울특별시 강서구 공항이화여자대학교의과대		796	303	653	205543.13	내과, 신경과, 정	126.83627	37.557263	182	1	0	0	181	43	8	107	4	2	3.6	48	3.5	57	4.20	341	10	10	10	3
19940622	서울특별시 서대문구 연세로병원등신의료재단		112	59	207	8493.73	내과, 신경과, 외	126.93643	37.581332	21	3	0	0	18	4	0	22	1	1	3.1	22	3.9	8	3.97	78	10	10	10	1
20110309	서울특별시 중랑구 신내서울특별시서울의료원		840	168	623	100120.53	내과, 신경과, 정	127.09783	37.613276	224	1	26	71	124	35	5	104	3	3	3.7	75	2.6	41	4.40	257	8.2	7.8	7.8	6
19820930	서울특별시 영등포구 의성애의료재단 성애병원		336	85	275	20380	내과, 신경과, 정	126.9225	37.512027	83	2	7	11	63	17	2	68	2	2	3.5	31	2.6	41	4.24	201	10	10	10	1
19800128	서울특별시 은평구 홍은의료법인 청구성심병원		118	63	210	5894.91	내과, 신경과, 외	126.9197	37.620862	24	1	0	0	23	6	1	11	1	1	3.8	34	2.8	32	4.20	83	7.4	5.8	6.5	6
19780223	서울특별시 용산구 대사순천향대학교 부속 서울		917	175	725	46974.38	내과, 신경과, 정	127.00426	37.533804	355	2	38	119	197	59	5	110	4	3	3.3	66	2.2	40	4.16	348	9.1	8.8	8.8	8

*)일부 주요 열만 표시

```
data.shape
```

(358, 93)

총358행, 93열의데이터 (data shape)

```
data.columns
```

```
Index(['개방서비스명', '개방자치단체코드', '인허가일자', '도로명전체주소', '사업장명', '의료인수', '입원실수', '병상수',
       '총면적', '진료과목내용명', '종별코드', '종별코드명', '시도코드', '시도명', '시군구코드', '시군구명',
       'X좌표', 'Y좌표', '의사총수', '일반의의사수', '인턴의사수', '레지던트의사수', '전문의합계', '내과',
       '신경과', '정신건강의학과', '외과', '정형외과', '신경외과', '흉부외과', '성형외과', '마취통증의학과',
       '산부인과', '소아청소년과', '안과', '이비인후과', '피부과', '비뇨의학과', '영상의학과', '방사선종양학과',
       '병리과', '진단검사의학과', '결핵과', '재활의학과', '핵의학과', '가정의학과', '응급의학과', '직업환경의학과',
       '예방의학과', '치과', '구강악안면외과', '치과보철과', '치과교정과', '소아치과', '치주과', '치과보존과',
       '구강내과', '영상치의학과', '구강병리과', '예방치과', '통합치의학과', '한방내과', '한방부인과', '한방소아과',
       '한방안이비인후피부과', '한방신경정신과', '침구과', '한방재활의학과', '사상체질과', '한방응급', '의료장비합계',
       '유방촬영장치', 'CT', '콘빔CT', '양전자단층촬영기 (PET)', '골밀도검사기', 'MRI', '초음파영상진단기',
       '종양치료기(Gamma Knife)', '종양치료기(Cyber Knife)', '종양치료기(양성자치료기)', '체외충격파쇄석기',
       '혈액투석을위한인공신장기', '구글점수', '구글건수', '다음점수', '다음건수', '네이버평점', '네이버건수',
       '진료만족도', '의료진친절도', '시설만족도', '굿닥건수'],
      dtype='object')
```

데이터 열 설명

1)인허가 데이터 (http://localdata.go.kr/main.do)

'개방자치단체코드','인허가일자','도로명전체주소','사업장명','의료인수','입원실수','병상수','총면적', '진료과목내용명', '종별코드', '종별코드명', '시도코드', '시도명', '시군구코드', '시군구명', 'X좌표', 'Y좌표', '의사총수'

개별 병원의 기본적인 설명 데이터로 병원의 이름, 위치 등의 속성을 가지고 있습니다. 또한, 입원실 수, 병상 수 등의 병원 별 세부사항이 기록되어 있습니다. 추가적으로 좌표정보가 있어 시각화에 유용하게 쓰일 것입니다.

2)건강보험심사평가원 데이터 (https://www.data.go.kr/data/15051059/fileData.do)

'일반의의사수', '인턴의사수', '레지던트의사수', '전문의합계', '내과', '신경과',.....'의료장비합계', '유방촬영장치', 'CT',.....'혈액투석을위한인공신장기'

개별병원별 진료과목별 의사수와 보유 주요 의료장비수가 기록되어 있습니다. 구체적인 진료과목별 전문의 현황과 의료기기현황을 확인할 수 있습니다.

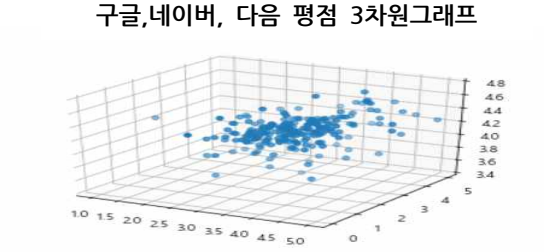
3)평점데이터(네이버, 구글, 다음, 굿닥 리뷰 정보)

'구글점수;, '구글건수', '다음점수', '다음건수', '네이버평점', '네이버건수', '진료만족도','의료진친절도', '시설만족도', '굿닥건수'

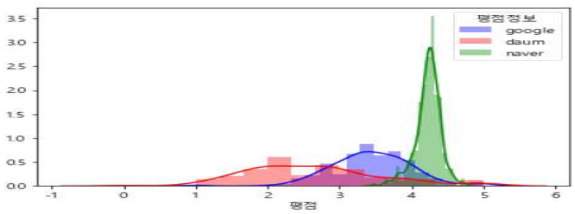
각 포털사이트와 굿닥에서 개별 병원의 평점과 평가인원을 크롤링한 정보입니다. 굿닥의 경우 진료만족도, 의료진친절도, 시설만족도로 세분화 되어있습니다.

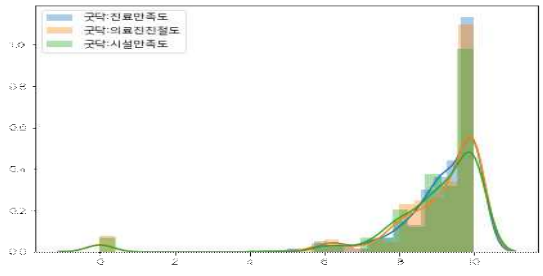
4. 데이터에 대한 간략한 분석 및 결과

구글,네이버, 다음 평점 3차원그래프



각 포털사이트의 점수분포





굿닥 점수분포

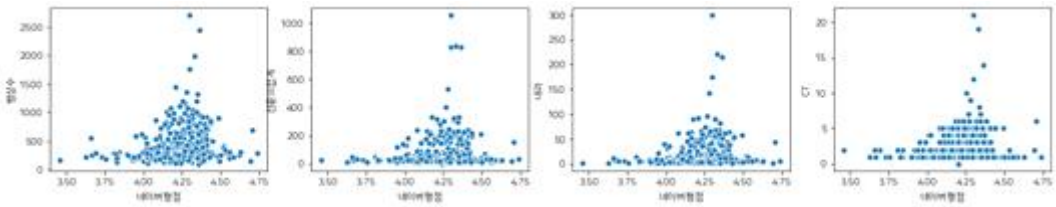
종합병원 위치 시각화자료

포털 평점 간에 큰 선형관계는 보이지 않으며, 네이버 평점이 높은 쪽에 분포하고 있으며 구글, 다음 순으로 이루어져 있습니다. 굿닥 만족도의 경우 어느정도 비슷한 분포를 띄고 있습니다. 시각화 자료를 보면 대부분의 종합병원이 수도권에 밀집해 있음을 확인할 수 있습니다.

	구글점수	구글건수	다음점수	다음건수	네이버평점	네이버건수	진료만족도	의료진친절도	시설만족도	굿닥건수
구글점수	1.000000	0.046359	0.353544	-0.029893	0.162336	0.018220	0.103401	-0.004656	-0.014544	-0.022574
구글건수	0.046359	1.000000	-0.154584	0.705079	0.091894	0.701847	0.080301	0.115033	0.137140	0.706139
다음점수	0.353544	-0.154584	1.000000	-0.179569	0.073167	-0.163566	-0.072611	-0.068538	-0.054743	-0.105668
다음건수	-0.029893	0.705079	-0.179569	1.000000	0.039931	0.667633	0.051050	0.126147	0.148213	0.751267
네이버평점	0.162336	0.091894	0.073167	0.039931	1.000000	0.116223	0.149837	0.137791	0.130688	0.087890
네이버건수	0.018220	0.701847	-0.163566	0.667633	0.116223	1.000000	0.098192	0.142164	0.150816	0.670350
진료만족도	0.103401	0.080301	-0.072611	0.051050	0.149837	0.098192	1.000000	0.400834	0.351497	0.050258
의료진친절도	-0.004656	0.115033	-0.068538	0.126147	0.137791	0.142164	0.400834	1.000000	0.952277	0.151626
시설만족도	-0.014544	0.137140	-0.054743	0.148213	0.130688	0.150816	0.351497	0.952277	1.000000	0.162041
굿닥건수	-0.022574	0.706139	-0.105668	0.751267	0.087890	0.670350	0.050258	0.151626	0.162041	1.000000

각 사이트간 평점과 건수 상관계수 표

각 사이트간 건수는 상관관계가 있어 보이지만 점수 간의 상관관계는 보이지 않습니다. 굿닥 자료에는 어느정도 상관관계를 보이고 있습니다.



병원의 여러 피쳐들과 평점간 관계

병상수나 전문의수, 특수장비수등 병원의 여러 피쳐들과 평점간 관계도 명확하지 않았습니다.

수집하고 통합한 데이터에서 확인한 종합병원의 평점 분포가 포탈마다 다른 양상을 띠는 이유는 평점에 영향을 미치는 요인이 많기 때문인 것으로 보입니다. 다양한 과의 많은 의료진이 존재하는 만큼 환자 개개인이 제공받는 의료서비스에서 경험이 다양하기 때문에 환자들의 평가가 일관성이 떨어지는 것 같습니다. 아직 평점 관련 변수만을 확인하였기에 더욱 구체적인 분석과 관계를 찾으려면 세분화 등을 통해 추가적인 분석이 필요할 것으로 보입니다.

5. 추후 프로젝트 계획

아직 평점관련 변수만 구체적으로 확인했기 때문에 다른 변수들의 특징을 파악하고 평점관련 변수와 관계를 확인하고자 합니다. 만약 관계를 찾기 힘들다면 사람들의 평가 중 수많은 영향요인들이 평균되어 만들어진게 평점이므로 평점을 세분화할 필요가 있을것 같습니다. 평점을 세분화하면 그래도 병원의 피쳐들과 상관관계를 찾아볼 여지가 더 생길 것 같습니다. 따라서 평점과 리뷰갯수만 크롤링 했던 것에서 더 나아가 모든 리뷰를 크롤링하여 평가기준을 세분화하는등 (ex. Daum 평점 => 시설평점, 의료평점, 직원평점, 가격평점 등) 여러 시도를 통해 보다 유의미한 결과를 도출하여 최종적으로 비어있는 병원별 진료만족도/ 의료진 친절도/ 시설만족도 점수들을 채우는 것이 최종 목표입니다.