데이터 사이언스 1차 중간보고서

3조 이상익, 이진우, 정규형

1. Team Data 선택 과정

첫 모임에서 데이터 통합을 통해 가치 창출이 가능한 데이터를 탐색했습니다. 범죄데이터, 기업데이터, 병원데이터, 교통데이터 등을 후보군으로 요약하였습니다. 채택되지 못한 데이터의 특징과 한계는 아래와 같습니다.

범죄데이터	특징	1. 지역별 특정 유형의 범죄 빈도를 통해 역사적, 지역적, 경제적 분석 가능 2. 복합적인 범죄요인을 데이터 통합으로 특징 파악 및 적용하여 범죄방지에 이바지
	한계	1. 지역별 범죄유형별 세부적인 데이터 획득 불가 2. 지역별로 29개의 행만 존재
기업데이터	특징	1. 재무정보를 통한 기업 분석 및 전망 예측 2. 모델링을 통해 기업의 파산 여부 예측 가능
	한계	1. 공시서류 이외의 추가적인 데이터 구할 수 없음 2. 관련 지식 부족으로 추가적인 분석의 어려움
교통데이터	특징	1. 방대한 공공데이터 통합 및 분석 가능 2. 네트워크 자료 분석 지식을 보유하여 다양한 분석 가능
	한계	1. 이미 다양하게 활용된 결과가 존재하여 새로운 가치 창출의 어려움 2. 방대한 데이터의 핸들링 능력 부족

최종적으로 병원데이터를 선택했습니다. 코로나19 이후 의료서비스의 중요도가 높아지고 있는 만큼 데이터 통합으로 의미 있는 결과 도출이 기대되기 때문입니다. 병원데이터는 공공데이터뿐 아니라 포털사이트의 병원 정보데이터를 통합할 수 있어 더욱 다양한 시도가 가능할 것입니다. 또한, 모든 조원이 네트워크 자료 분석이 가능하기에 이를 반영한 분석도 가능합니다. 마지막으로 의료서비스 특성상 환자와 병원 간 정보 비대칭으로 피해를 보는 경우가 있습니다. 팀 프로젝트를 통해 본인에게 적합한 의료시설을 선택하도록 이끄는 데이터를 산출하여 국민의 의료서비스 개선에 도움을 주고자 합니다. 혹은 지역구

에 따른 병원의 현황과 질병 취약 계층 분포를 통해 향후 병원 배치에 도움을 주는 데이터 통합을 하고자 합니다.

2. 관련 데이터 파악

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

30명 이상의 환자를 수용할 수 있는 시설을 갖춘 전국 5853개의 의료기관 정보입니다. 데이터의 속성은 아래와 같습니다.

```
Index(['번호', '개방서비스명', '개방서비스ID', '개방자치단체코드', '관리번호', '인허가일자', '인허가취소일자',
       '영업상태구분코드', '영업상태명', '상세영업상태코드', '상세영업상태명', '폐업일자', '휴업시작일자', '휴업종료일자',
       '재개업일자', '소재지전화', '소재지면적', '소재지우편번호', '소재지전체주소', '도로명전체주소', '도로명우편번호',
       '사업장명', '최종수정시점', '데이터갱신구분', '데이터갱신일자', '업태구분명', '좌표정보(X)', '좌표정보(Y)',
       '의료기관종별명', '의료인수', '입원실수', '병상수', '총면적', '진료과목내용', '진료과목내용명', '지정취소일자',
       '완화의료지정형태', '완화의료담당부서명', '구급차특수', '구급차일반', '총인원', '구조사수', '허가병상수',
       '최초지정일자'],
```

2) 서울시 의료기관 (구별) 통계 (https://data.seoul.go.kr/)

2018년 기준 각 자치구의 종합병원, 병원, 의원 등의 의료기관 병원수 및 병상수 현황 정보입니다.

자치구	계		종합병원		•	•	•	의원	
	병원수	병상수	병원수	병상수				병원수	병상수
서울시	17,387	86,536	58	33,318				8,379	10,750
종로구	500	3,408	4	2,970				186	170
중구	562	1,473	3	974				234	185
					•				
					•				
					•				

3) 병원등급정보(건강보험 심사평가원) (http://www.hira.or.kr/main.do)

수술, 질병, 약제사용 등 병원의 의료서비스를 의·약학적 측면과 비용 효과적인 측면에서 평가한 정보입니다. 데이터의 속성은 아래와 같습니다. 추가로 API를 이용한 데이터 추출이 가능합니다.

```
Index(['NO', '병원명', '평가항목', '평가등급', '소재지'],
```

4) 포털사이트의 정보(네이버, 구글 등)

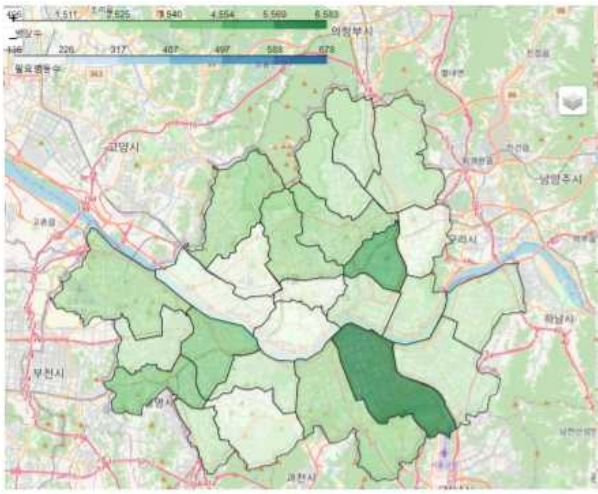
병원명을 검색하였을 때 노출되는 모든 정보입니다. 예시로 리뷰, 평점 등이 있습니다.

5) 전국병원현황 (http://www.kha.or.kr/)

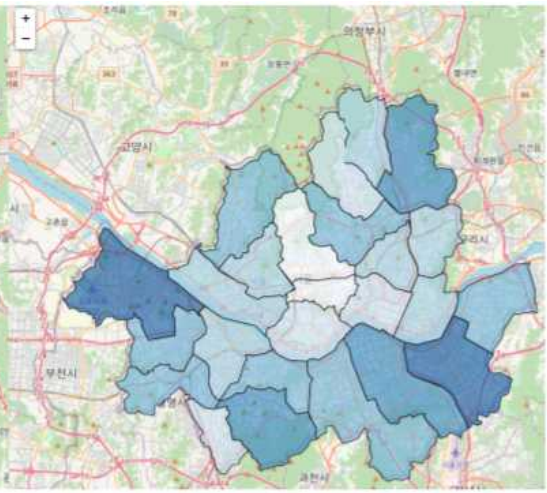
전국의 병원명, 지역, 전화번호와 권역별 진료과목집계 자료 정보입니다.

현재까지 찾아본 관련 데이터는 위와 같습니다. 1) 데이터가 많은 속성을 가지고 있는 만큼 주요 데이터로 쓰일 것이라 예상되며 나머지 데이터와 추가 수집을 통해 프로젝트를 진행할 것입니다.

3. 추후 프로젝트 계획



<그림1> 지역 내 정상운영 병원들의 총 병상수



<그림2> 지역 내 필요 병상수

먼저 서울의 29개 자치구에 대해 의료서비스 공급이 충분한지 확인하고자 위와 같은 시각화를 진행했습니다. 그 결과 필요 병상수에 비해 총 병상수가 압도적으로 높아 의료공급 자체는 크게 부족하지 않다는 것을 확인하였습니다. 위 주제로 새로운 가치를 도출하는 것은 어려울 것으로 판단했습니다. 이러한 이유로 적합한 의료시설을 제안하는 데이터를 산출하는 것에 초점을 맞추었습니다.

<그림3> 데이터 통합 스케치

<그림3>과 같이 데이터 통합과정을 거쳐 최종적으로 개인별 병원평가를 통해 적합한 의료시설을 제안하고자 합니다. 현재 위치 특징은 좌표정보만 간략하게 기재하였지만 충분한 논의를 통해 다양한 변수를 적용할 것입니다. 지하철역과의 접근성, 주변 버스 정류장의 수 등을 고려하고 있습니다. 다음으로 포털사이트의 병원 평점, 리뷰에 나타난 긍·부정적 단어를 반영할 것입니다. 이 과정에서 평점이 존재하지 않는 병원의 경우 다른 변수와 인근 병원의 특징을 반영한 결측치 처리 작업이 진행될 것입니다. 또한, '모두닥'이라는 앱에 반영된 구체적인 고객평가(자세한 설명, 적절한 금액)를 추가하고자 합니다. 마지막으로 개인에 따라 병원의 평가가 달라지는 만큼 진료과목과 의료인 수의 특징을 최대한 반영할 것입니다. 만약 이 속성이 무시된다면 질병이 고려되지 않아 그저 단순한 데이터 통합에 그칠 수 있기 때문입니다.

4. 데이터 통합과정에서의 예상 문제점

1) 인허가 데이터에서 '의료진 수' 속성은 존재하지만, 진료과목에 따른 의료진 수를 알 수 없다는 것입니다. 환자에 따른 병원평가를 하기 위해 병원의 구체적인 의료진 수는 중요한 요소지만 이것을 반영할 수 없어 대체할 방법을 찾아야 합니다.

2) 모든 병원의 등급이 존재하지 않아 많은 결측치가 생기는 것입니다. 등급이 각 질병에 따라 다르게 매겨져 있어 종합병원이나 수술이 이루어지는 병원에만 자료가 존재하고 있습니다. 많은 의료기관의 등급이 존재하지 않아 방향성이 분산될 가능성이 큽니다. 현재 범위를 줄여 서울의 종합병원급에 집중하여 데이터 통합을 진행하는 것으로 수정하였습니다. 다만 행의 개수가 약 60개로 줄어 유의미한 결과를 도출해낼지 의문이 듭니다.

3) 병원별 포털사이트와 '모두닥' 데이터 편차가 심하다는 것입니다. 예시로 세브란스 병원은 228개의 구글 리뷰가 존재하지만 대림성모병원은 29개의 리뷰가 존재합니다. 또한, '모두닥'에서는 차이가 더 심하게 나타납니다. 환자의 입장이 직접 반영된 자료인 만큼 이 문제점은 반드시 해결하고자 합니다.

4) 어느 분야에서든 나타나는 공공데이터의 품질문제입니다. 인허가 데이터에서 위치 정보, 의료진 수가 타 데이터와 다른 것을 확인하였습니다. 이는 'gg map'과 같은 위치 정보 추출 패키지를 이용하여 해결할 것입니다. 다만 실제 값 파악이 불가능한 경우 우선 공공데이터 문의 및 답변을 통해 해결책을 얻을 것입니다. 또한, 데이터마다 기준년이 다르게 나타나고 있어 기준점을 잡아야 합니다.