**Report Title** : 강남권의 학업 중단율을 코로나 19, 사교육 시장과 연결하여 인사이트 발견

**DataVisualization**

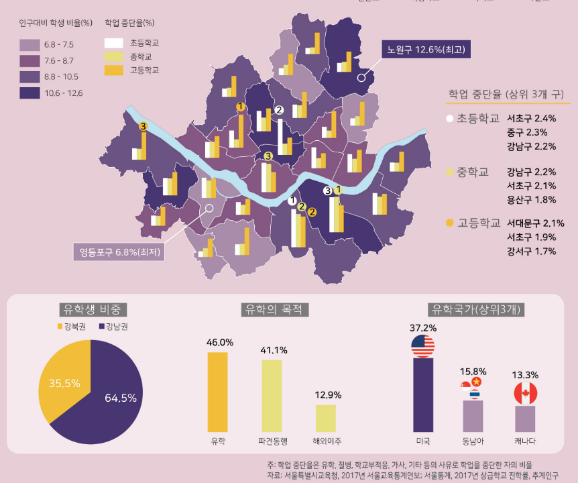
**목차**

1. 문제정의
2. 자료 수집, 전처리 및 탐색
3. 자료 분석
4. 결과 시각화
5. 결론

**1. 문제정의 (**본 프로젝트를 통하여 어떤 문제를 해결하고자 하는지 최대한 구체적으로 기술**)**

뉴스 기사를 보던 중, 서울 서초/강남 등 고소득자가 많고, 교육열이 높으며 사교육 시장이 발달한 지역에서 학교를 그만두는 자퇴학생 비율이 타 지역에 비해 높은 것으로 조사되었다는 것을 보았다.

2017년 당시에 서초/강남의 학업 중단율이 서울에서 가장 높은 것으로 보였음. 교육 전문가들은 “대입에서 학생부 위주 전형 선발비율이 크게 늘면서, 불리해진 교육특구 학생들이 해외로 유학을 하거나, 학교를 자퇴하고 수능을 대비하는 비율이 타지역보다 상대적으로 높기 때문”이라고 밝혔다.



위의 사진에서 보면 유학생의 비율이 강남권에서 높고, 또한 학업중단율 또한 강남권에서 높게 비율을 차지하고 있다는 것을 알 수 있다.

따라서 이번 시각화 프로젝트를 통해

1. 서울시의 학업중단율은 어떤 비율로 나눠지는지 지도 시각화를 통해 알아보고
2. 코로나 19 이후의 서울시의 사교육 시장은 어떻게 흘러가고 있는지
3. 코로나 19 이후 구별 학업 중단율은 어떻게 변화하고, 어떠한 비율로 나눠지고 있는지

알아보고자 하였다.

[참고 : <http://www.edujin.co.kr/news/articleView.html?idxno=19332>]

[사진 : https://www.si.re.kr/node/60720]

**2. 자료수집 (**자료는 어떻게 구하였는지, 어떠한 전처리 과정을 적용하였는지, 그리고 탐색을 통한 자료 이해 과정 기술**)**

**(1) 학업 중단율 데이터 :** [https://data.seoul.go.kr/dataList/10713/C/2/datasetView.do]

-> 서울 열린데이터 광장에서 seoul\_stop\_school.xlsx을 사용

**(2) 지도 시각화를 위한 json 데이터** :

[https://raw.githubusercontent.com/southkorea/seoul-maps/master/kostat/2013/json/seoul\_municipalities\_geo\_simple.json]

**(3) 서울시 사교육비 현황 데이터 :** [ <https://data.seoul.go.kr/dataList/10712/S/2/datasetView.do> ]

-> 서울 열린데이터 광장에서seoul\_private.csv을 사용

전처리 과정의 경우 (1) 학업 중단율 데이터와 (3) 서울시 사교육비 현황 데이터를 전처리 하는 과정을 거쳤다.

대표적으로 (1) 학업 중단율 데이터의 경우 행 삭제와 열 삭제, 그리고 ‘학업중단율’의 단어가 포함된 열만 추출하여 새로 데이터 프레임을 생성하였다.

그리고 (3) 서울시 사교육비 현황 데이터의 경우 컬럼 이름을 변경하고, 시점을 (1) 학업 중단율 데이터와 맞추기 위해 2011~2021의 데이터만 남겨두어 새로 데이터 프레임을 생성하였다.

자료를 탐색하며 분석을 하기 위해서는 같은 시점을 맞춰야 한다는 것을 깨달았고 뿐만 아니라. 퍼센트 율 (%)로 형성된 데이터의 시각화를 할 때 시간이 단축된다는 것을 깨달았다.

**3. 자료분석** (분석 적용 과정),

자료분석의 경우 크게 3단계를 진행했다.

*1step) 구별 학업 중단율을 알기 위한 자료 분석*

구별 학업 중단율은 지도에 시각화를 하기 위해 자료를 분석하였다. 구별 학업 중단율을 시점별로 알기 위해 시점을 index로 설정하였고, ‘학업중단율’의 단어가 포함된 데이터를 추출하여 분석하고자 하였다.

*2step) 서울시 사교육 참여율을 알기 위한 자료 분석*

문제 정의 단계시 ‘서울 서초/강남 등 고소득자가 많고, 교육열이 높으며 사교육 시장이 발달한 지역에서 학업 중단율이 높다’ 라는 것을 보고 서울시 사교육 참여율을 알기 위한 자료 분석을 진행하였다. *1step*와 같은 시점으로 설정하였고, *1step*에서 데이터가 초/중/고 로 나뉘어져 있었기 때문에 *2step*에서도 초/중/고 별 사교육 참여율을 알기 위해 자료 분석을 진행하였다. 이때, 각각의 그래프로 표현하면 각 교육과정별 어느 정도인지 가늠이 안 갈 것이라고 판단하였다. 따라서 하나의 그래프에 표현하고자 생각하였다.

또한, 문제 정의 시 ‘코로나 19 이후의 서울시의 사교육 시장은 어떻게 흘러가고 있는지’를 알기 위해 sns.lineplot의 그래프를 사용하여 그리는 것이 타당하다고 생각해 이의 방법으로 자료 분석을 진행하였다.

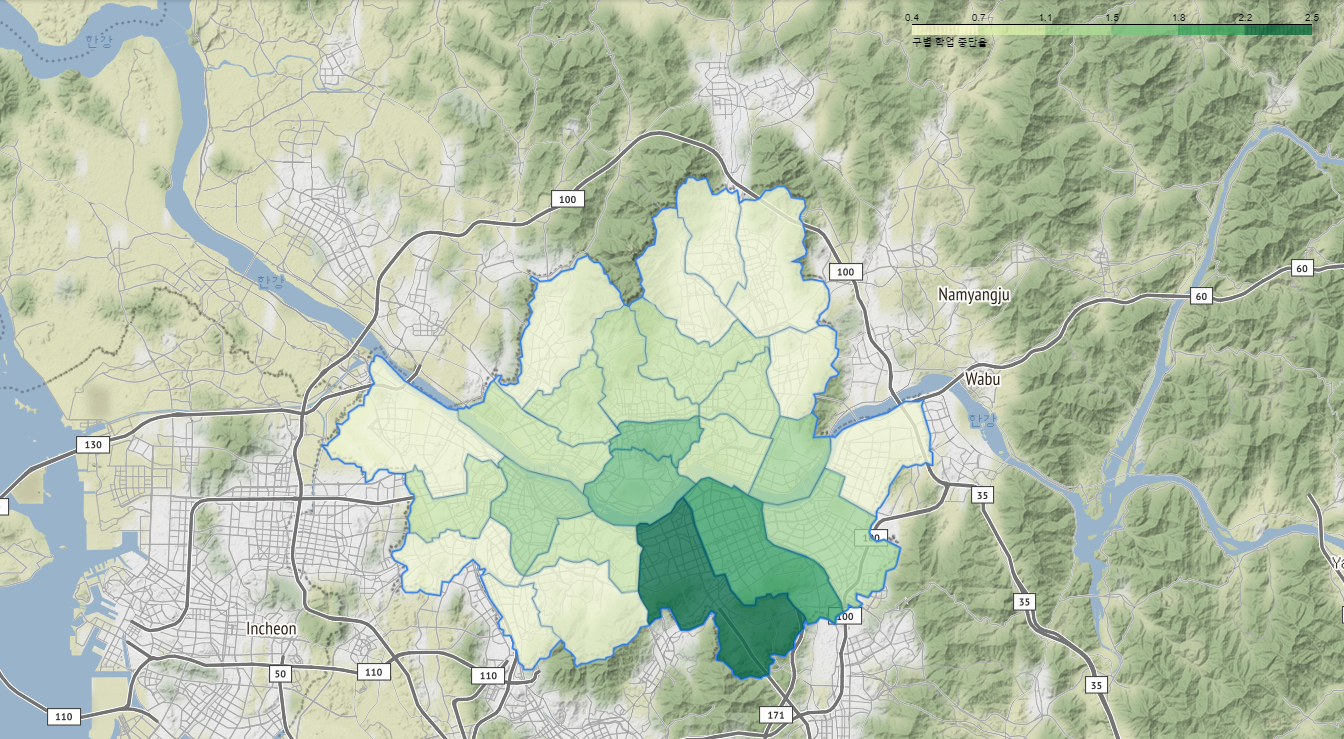
*3step) 코로나 19 이후 구별 학업 중단율의 변화를 알기 위한 자료 분석*

코로나 19 이후 구별 학업 중단율의 변화를 알기 위해서는 다시 학업 중단율의 데이터를 사용해야 했다. 따라서 seoul\_stop\_school.xlsx의 데이터를 *1step*과 구별하기 위해 새로운 변수에 담고 진행하였다. 이때, y축은 서울시의 구별, x축에는 시점을 나타내고자 하였고 초/중/고 각각 알아보고자 하였다. 따라서 해당 목적을 이루기 위해서는 sns.heatmap의 그래프를 사용하여 그리는 것이 타당하다고 생각해 이의 방법으로 자료 분석을 진행하였다.

뿐만 아니라 , 초/중/고 그래프를 따로 그리기 위해서 자료 중 ‘초등학교’ ‘중학교’ ‘고등학교’의 문자열이 들어간 열을 각각 추출하여 변수에 담았고, 이를 사용하여 서울시 구별 학업 중단율 그래프를 표현하고자 하였다.

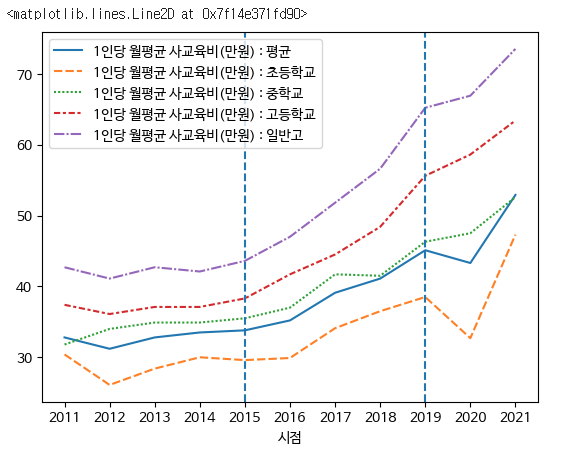
**4. 결과 시각화**

1step) 구별 학업 중단율을 알기 위한 자료 분석

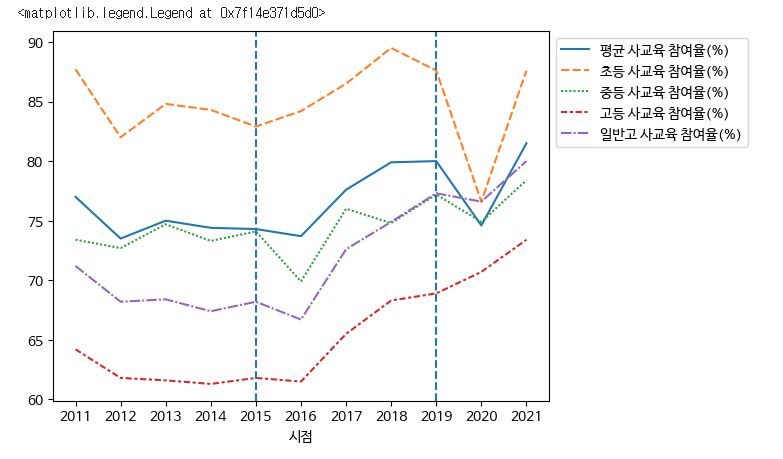


[ 해당 그래프 분석을 통해 문제정의 단계에서 발견한 ‘서울 서초/강남 등 고소득자가 많고, 교육열이 높으며 사교육 시장이 발달한 지역에서 학교를 그만두는 자퇴학생 비율이 타 지역에 비해 높은 것’을 사실로 확인하였고, 이는 현 시점에도 계속해서 나타나는 현상이라는 것을 알게되었다. ]

2step) 서울시 사교육 참여율을 알기 위한 자료 분석

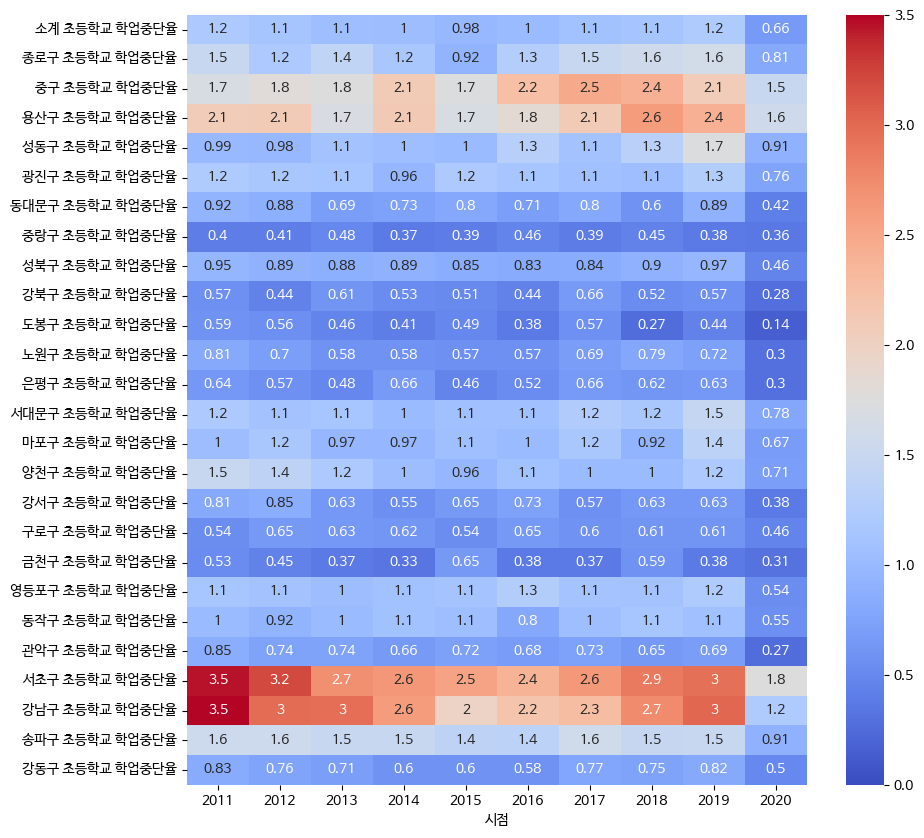


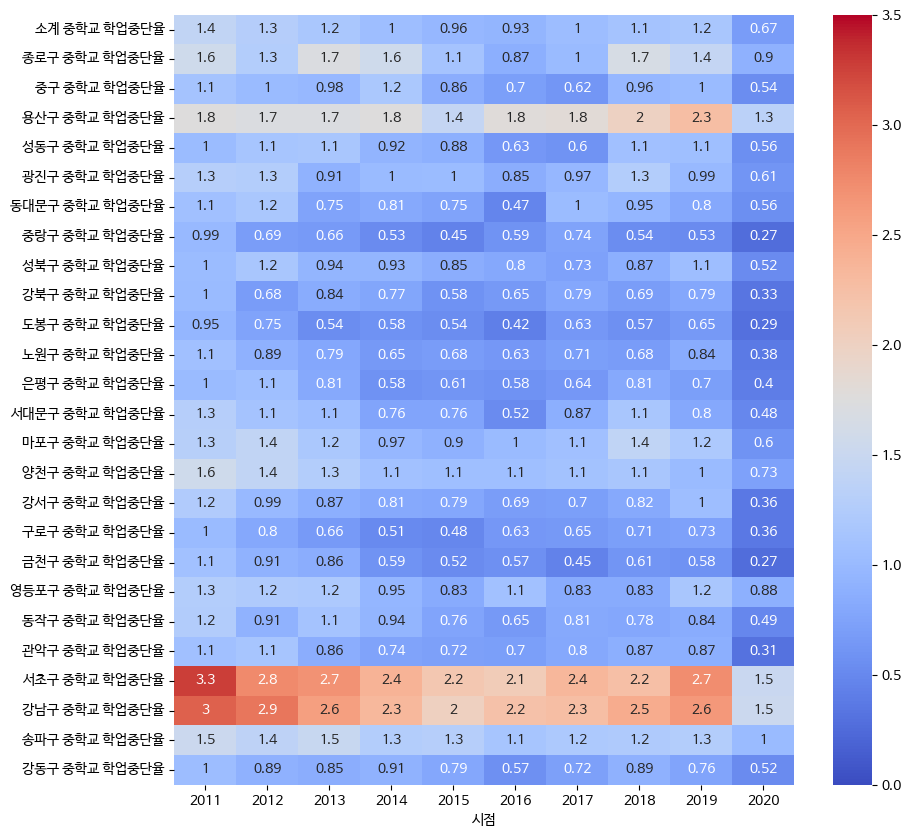
[ 해당 그래프를 통해 2011~2021로 갈수록 서울 1인당 월평균 사교육비가 높아졌고, 평균보다 일반고, 고등학교, 중학교의 사교육비가 높음을 알게 되었다. ]

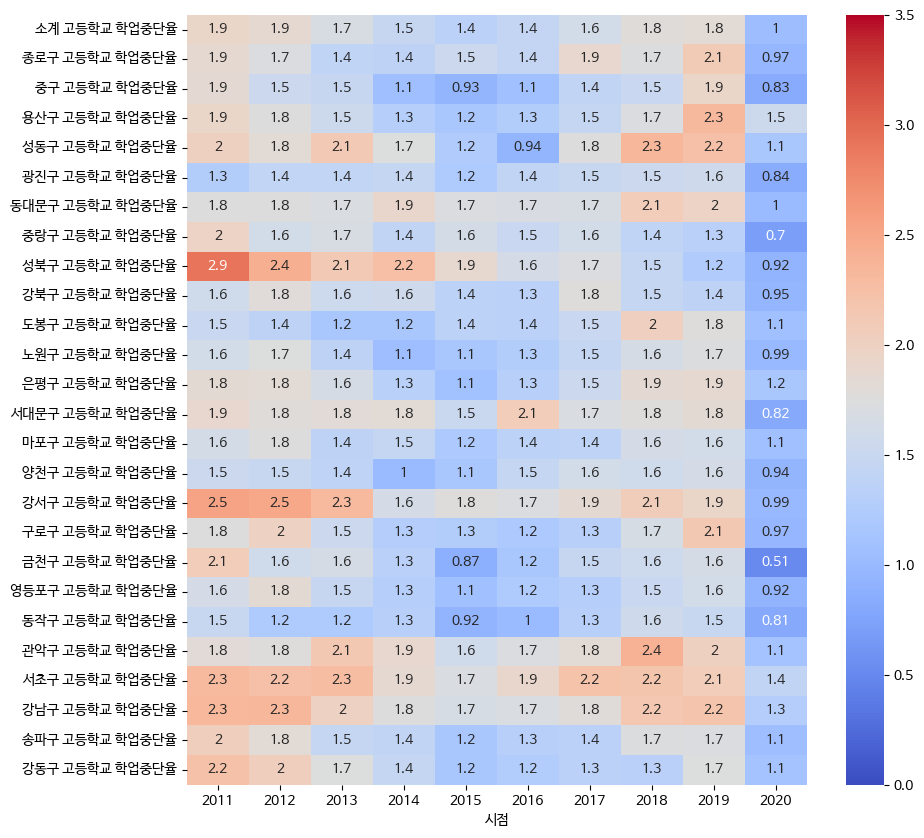


[ 해당 그래프를 통해 참여율은 초등이 높지만, 초등을 제외한 순서로 보면 일반고-> 중등-> 고등의 참여율순으로 흘러가고, 2011~2021로 갈수록 참여율 또한 점점 증가하는 추세를 보임을 알게 되었다 ]

3step) 코로나 19 이후 구별 학업 중단율의 변화를 알기 위한 자료 분석



****

****

[ 순서대로 초등-> 중등 -> 고등의 구, 시점 별 학업 중단율 heatmap이다. 해당 그래프를 통해 두 가지의 인사이트를 얻을 수 있었다. 이 두가지 인사이트는 초등학교,중학교, 고등학교 모든 결과에 적용 가능하다.

(1) 서초구,강남구,용산구가 다른 구에 비해 월등히 높은 비율로 학업 중단 비율이 높았다. 이는 2011~2020 현 시점까지 계속해서 이어지고 있는 추세이다.

(2) 하지만 2020 즉, 코로나가 발생한 이후 오히려 학업 중단율의 비율이 다른 시점에 비해 많이 줄어들었다.]

**5. 결론** (본 프로젝트를 통해 얻은 insight 및 의견 등)

문제정의에서 생각한 문제는 다음과 같았다.

1. 서울시의 학업중단율은 어떤 비율로 나눠지는지 지도 시각화를 통해 알아보고
2. 코로나 19 이후의 서울시의 사교육 시장은 어떻게 흘러가고 있는지
3. 코로나 19 이후 구별 학업 중단율은 어떻게 변화하고, 어떠한 비율로 나눠지고 있는지

이는 데이터 분석과 시각화를 통해 다음과 같은 insight를 얻을 수 있었다.

(1) 서울시의 학업중단율은 서초, 강남 즉 강남권 위주로 높은 비율을 보인다. 즉, 다른 구에 비해 높은 학업 중단율을 보이는 것은 사실이다. 그리고 이는 현재까지 이어지고 있는 추세이다.

(2) 사교육 참여율과 사교육비는 매년 증가하고 있다. 특히 초등학교의 경우에는, 코로나 19전까지 감소추세를 보이다가 코로나19 발생 이후 급속한 증가율을 보이며 높은 사교육 참여율 상승세를 보였다. 또한 초,중,고 모두 코로나 19 발생 이후 사교육비 증가율이 다른 년도에 비해 급한 성장 (기울기가 가파름)을 보이고 있다는 것을 알게 되었다. 이를 통해 서울시의 사교육 참여율은 계속해서 증가하고 있으며, 코로나 19로 인해 사교육계가 더 큰 성장세를 보였다는 것을 알게 되었다.

(3) 여전히 코로나19 전후로 강남권의 학업 중단율이 높은 것은 맞았다. 하지만, heatmap의 결과 코로나19 발생 이후 강남권을 위주로 보았을 때 학업 중단율이 다른 시점보다 급격하게 하락되고 있는 것을 발견하였다.

특히 여기서 (3)의 결과에 주목하였다.

코로나 19발생은 해외 출국 (유학, 여행)이 제한되는 시기가 되었었다. 다시 문제정의로 돌아가면, <2017년 당시에 서초/강남의 학업 중단율이 서울에서 가장 높은 것으로 보였음. 교육 전문가들은 “대입에서 학생부 위주 전형 선발비율이 크게 늘면서, 불리해진 교육특구 학생들이 해외로 유학을 하거나, 학교를 자퇴하고 수능을 대비하는 비율이 타지역보다 상대적으로 높기 때문”이라고 밝혔다.> 의 내용이 있었다.

즉, 정리하자면 코로나 19 발생으로 인해 해외 출국에 제한이 생겼고, 이는 강남권에서 학업 중단율이 눈에 띄게 감소하는 것으로 발견되었다. 하지만 동시에 이는 코로나 발생 이후 서울시의 사교육비가 증가하는 추세임을 알 수 있었다. 따라서 비대면 학습과 해외 출국 제한으로 인해 학업 중단율이 감소되었지만, 동시에 사교육 시장이 증가하는 경향을 보였음을 추측할 수 있다.