

프로젝트 보고서

: 다변량 데이터 분석 - 해법수학

박준식·이찬녕·이형석·임유하

프로젝트 기반 빅데이터 서비스 개발자 양성 과정 (4기)

-----<요약>-----

본 연구에서는 해법수학 학원 입지 정보와 지역별 소득 수준, 지역별 사교육비 지출 현황 및 지역별 인구 데이터를 종합 분석하여, 사교육 프랜차이즈 산업에서의 높은 점유율 차지 요인을 알아내고 이를 적용하여 타 업체보다 우위를 차지하고자 한다.

지역별 인구수가 해법수학의 점유율 우위에 영향을 주었을 것이라는 가설을 설정하고 이를 분석한다. 지역별 인구수 데이터와 지역별 해법수학 지점 수의 데이터의 상관관계를 분석한 결과 둘 사이의 관계가 유의하다는 결과를 얻었다. 경쟁 상대인 웅진씽크빅과 인구수 사이의 상관관계와 비교하여 해법수학이 상대적으로 인구수에 비해 지점을 내고 있다는 결과를 알 수 있었다.

학교와 학원 사이의 거리를 측정하여 둘 사이의 상관관계를 분석했지만, 유의수준이 0.89로 학교와 학원 사이의 거리는 업계 내 점유율과 유의미한 상관관계가 없다는 결과를 얻었다.

따라서 업계 내 점유율에 있어서 인구수가 가장 중요한 요인이라고 할 수 있다.

주제어 : 프랜차이즈, 상권 및 입지분석, 학원 입지요인, 점유율, 상관관계

I. 도입

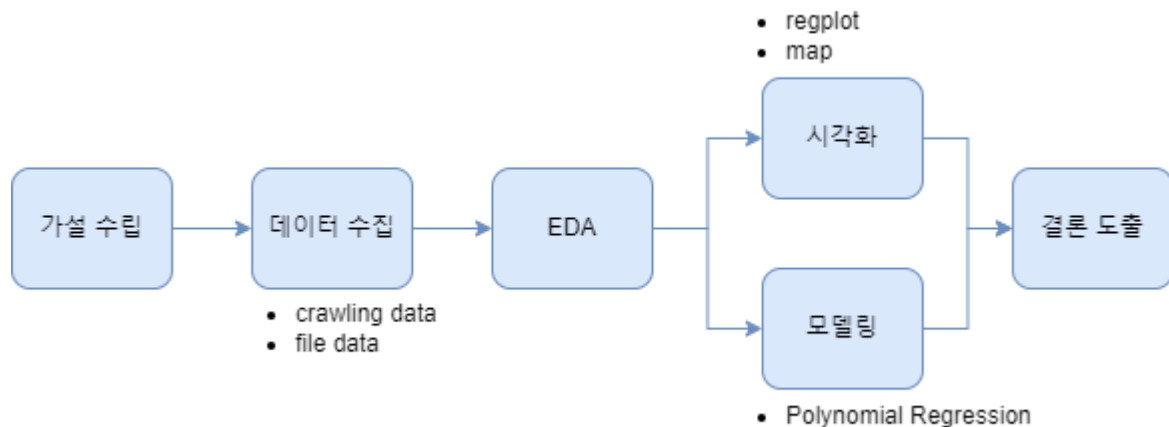
국내 프랜차이즈 산업은 새로운 비즈니스 모델을 개발하고 신규 프랜차이즈 기업으로서 시장에 진출하는 기업체들이 꾸준히 증가함에 따라 양적으로 성장하며 현재까지 지속적인 성장을 거듭하고 있다(강용호, 2009). 이러한 산업의 확대와 환경이 변화하고, 업체간의 치열한 경쟁이 지속됨에 따라 상권과 입지를 정확히 분석하여 더 유리한 위치에 가맹점을 입점시키는 것은 점유율 부분에서 우위를 차지하는데 것이 중요한 사안이 되었다.

우리나라는 현재 사교육이 공교육만큼이나 보편화 되어있어 사교육 여건은 특히 주거지 선택 및 집값 형성의 주요인자로 작용하고, 결과적으로 지역적 격차와 사회 갈등을 조장하는 핵심요소로 인식되어 각종 사회·경제적 쟁점의 이슈가 되는 상황이다(박소현, 이금숙, 2011). 사교육에 대한 많은 관심은 수많은 업체들이 사교육 산업에 뛰어들게끔 하였다. 치열한 경쟁 속에서 다른 업체보다 안정적이고 전략적인 방법을 구축하는 것은 산업에서의 우위를 차지하고 성공으로 이끄는 중요한 방법이 되었다.

그렇다면 어떤 요인이 사교육 프랜차이즈 산업에서의 성공을 이끄는 방법이 되는 것일까. ‘스마트해법수학’은 국내 수학프랜차이즈 점유율 1위를 기록하고 있는 사교육 프랜차이즈 업체 중 하나이다. 이 연구에서는 해법수학 입지요인과 타 데이터와의 상관관계 등을

분석하여 타 업체보다 더 높은 점유율을 차지하는 요인을 알아내고 이를 적용하여 해법수학이 타 업체보다의 점유율 측면에서의 우위를 확고히 하는 전략을 제안하고자 한다.

II. 방법



i) 데이터 수집

통계청 사이트에서 2015년부터 2021년까지의 지역별 개인소득, 1인당 사교육비, 시도별 인구수, 교육 서비스업 관련, 학교 시설관련, 사교육 참여율 데이터를 수집했다. 시군구별 인구 데이터는 2022년의 데이터를 수집했다.

스마트해법수학 사이트에서 지점명과 주소 2296건, 웅진홈스쿨 사이트에서 지점명과 주소 633건을 수집했다. 통계청 사이트에서 시도 및 시군구별 인구 통계를 크롤링해서 데이터를 수집하고, 파이썬 geopy 패키지를 각 가맹점의 주소에 사용하여 위도와 경도를 구한다.

학교알리미 사이트에서 2023년 학교별 공시정보 데이터를 수집하고 위도와 경도 데이터를 사용했다.

ii) 상관관계 분석

시군구별 인구 데이터와 해법수학, 웅진홈스쿨의 지점 수 데이터를 이용해 상관관계를 분석한다. 시군구별 인구수와 지점수로 산점도를 그리고 피어슨 상관계수를 구했다. 그 후에 해법수학의 상관계수가 웅진홈스쿨의 상관계수보다 유의미하게 클 것이라는 가설을 검정했다.

유의수준 0.05 미만의 귀무가설은 상관계수의 유의미한 차이가 없는 것이며, 대립가설은 해법수학의 상관계수가 웅진홈스쿨의 상관계수보다 유의미하게 높다는 것으로 설정했다.

상관계수끼리의 차이는 Z-검정을 통해 알아볼 수 있다. 하지만 상관계수는 -1부터 1 사이의 값이므로 먼저 Fisher transformation을 통해 정규분포의 범위로 늘려주어야 할 필요가 있다. Fisher transformation의 공식은 아래와 같다.

$$r' = 0.5 \times \ln \left| \frac{1+r}{1-r} \right| \quad ※(r = \text{상관계수})$$

이후 transformation된 r_1' 과 r_2' 을 이용해 Z-score를 구한다. Z-score를 구하는 공식은 아래와 같다.

$$Z = \frac{r_1' - r_2'}{\sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}} \quad ※(N = \text{표본 수})$$

마지막으로 Z-score를 정규분포표에 대조해 p-value를 구하고 $p < 0.05$ 이면 귀무가설을 기각한다.

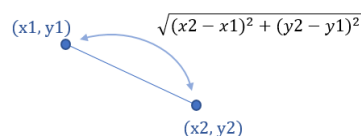
iii) 시각화

파이썬 geopy 패키지를 이용해서 구한 스마트해법과 경쟁사 웅진의 지점에 대한 위도 경도 데이터를 파이썬 folium 패키지를 이용해서 지도위에 나타냈다. 그리고 지도의 각 시도별 위치에 시도별 인구수 데이터를 색의 명암으로 나타내었다.

2021년의 시도별 1인당 사교육 지출, 시도별 개인소득, 시도별 인구수 데이터를 시도별 스마트해법수학의 지점수와 비교하여 산점도를 그리고 상관관계를 구했다. 또한 2022년의 시군구별 인구수와 시군구별 해법수학 및 웅진홍스쿨 지점수를 비교하여 산점도를 그리고 상관관계를 구했다. 마지막으로 2011년 1월부터 2023년 7월까지 월별 인구수 데이터로 산점도를 그리고 polynomial regression을 통해 추세선을 그렸다.

iv) 거리 비교

파이썬 folium 패키지를 이용해서 시각적으로 확인한 인구수와 학원과의 상관관계를 검증하기 위해서 거리를 비교했다. 학원을 주로 이용하는 인구는 학교에 다니는 학생이기 때문에 초등학교와 각 기업별 위치의 거리를 구하기 위해서 두 점 사이의 거리를 구하는 공식을 이용했다.



The diagram illustrates the Euclidean distance between two points in a 2D plane. Point 1 is labeled (x_1, y_1) and Point 2 is labeled (x_2, y_2) . A blue curved line connects the two points, representing the distance. Above the line, the formula $\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$ is written.

하지만 위의 공식은 평면에서의 두 점 사이의 거리를 구하는 경우에 해당하기 때문에 구형체인 지구에서 두 지점 사이의 거리를 구하는 하버사인 공식을 이용해서 실제 거리도

구하였다. ϕ_1 와 ϕ_2 는 초등학교와 학원의 각 위도를 의미하며, λ_1 와 λ_2 는 초등학교와 학원의 각 경도를 의미한다.

$$d = 2r \arcsin \left(\sqrt{\text{hav}(\varphi_2 - \varphi_1) + \cos(\varphi_1) \cos(\varphi_2) \text{hav}(\lambda_2 - \lambda_1)} \right)$$

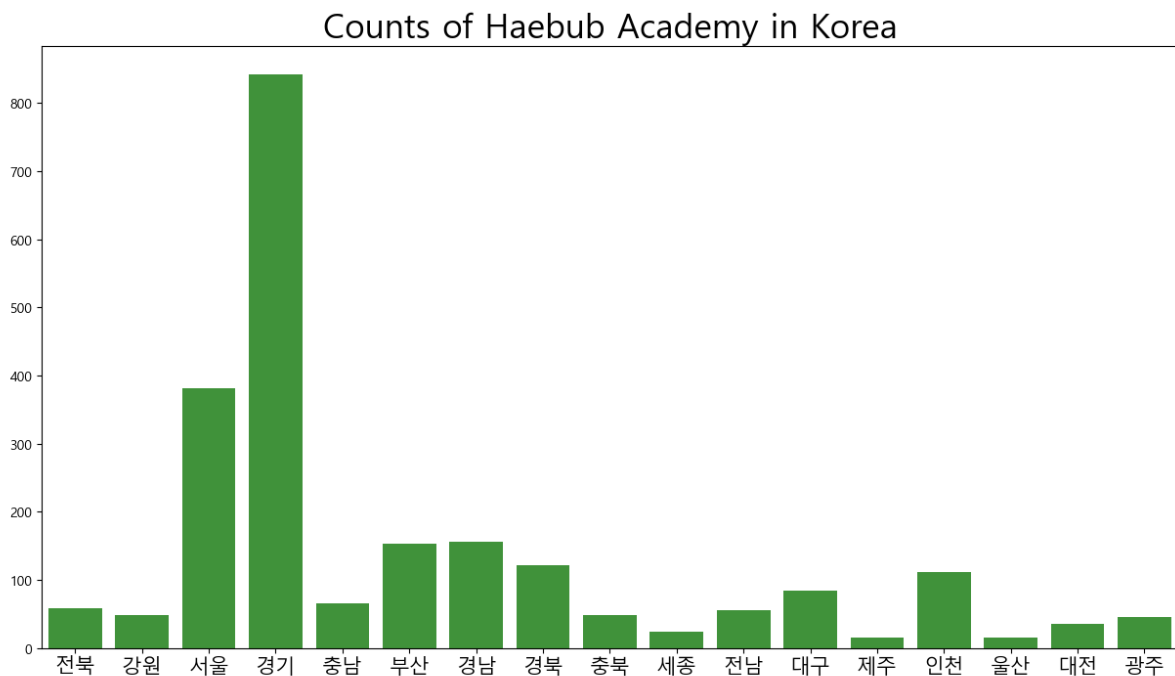
$$= 2r \arcsin \left(\sqrt{\sin^2 \left(\frac{\varphi_2 - \varphi_1}{2} \right) + \cos(\varphi_1) \cos(\varphi_2) \sin^2 \left(\frac{\lambda_2 - \lambda_1}{2} \right)} \right)$$

평면상의 거리와 실제 거리 모두 유의수준 0.05 미만으로 학교와 학원과의 거리가 업계 내의 점유율에 영향을 미친다는 귀무가설과 학교와 학원과의 거리가 업계 내의 점유율에 미치는 영향은 없다는 대립가설로 검정을 진행했다.

v) 모델링

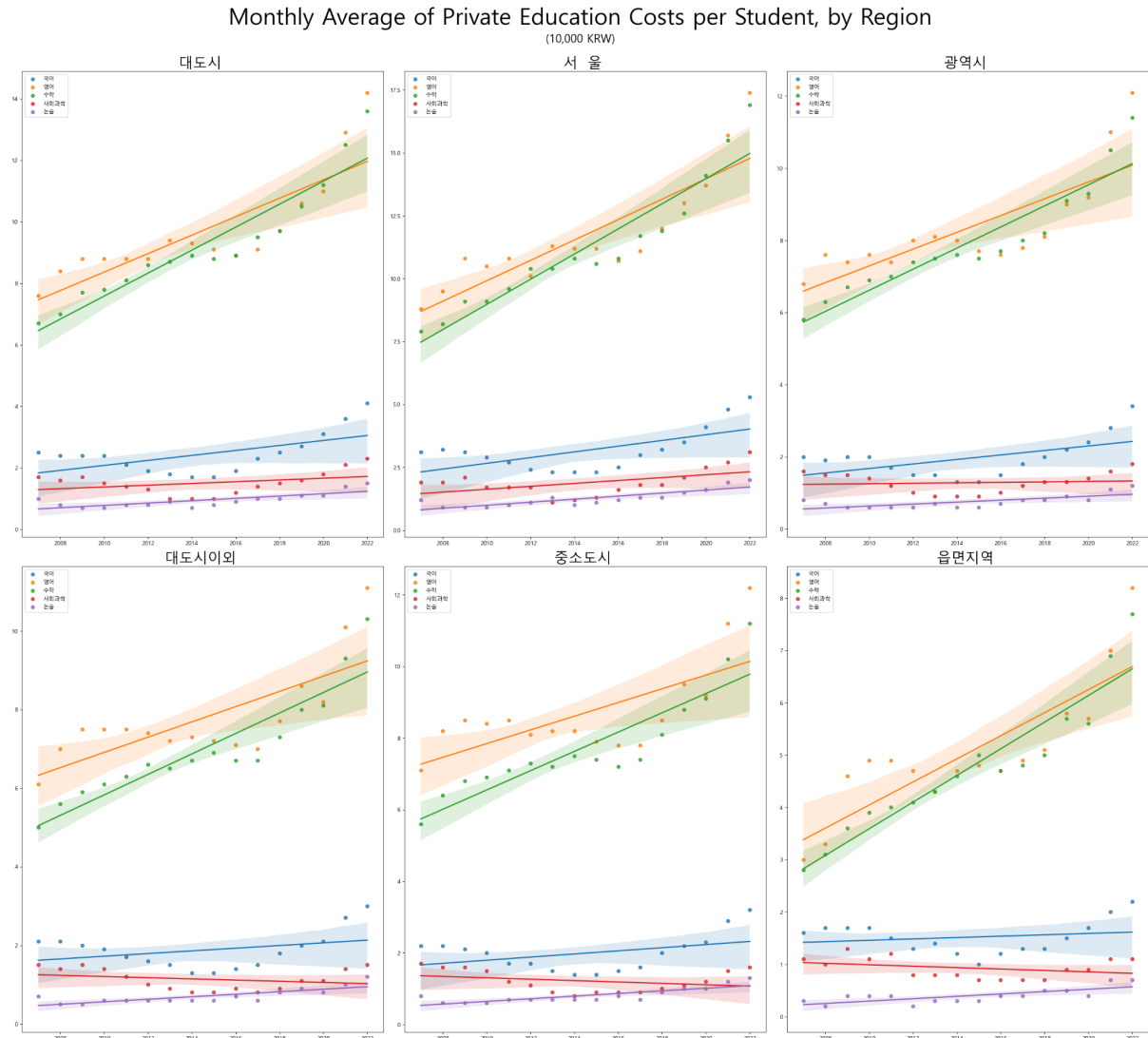
인구추세 모델링

III. 결과



[그림1] 지역별 해법수학 지점수

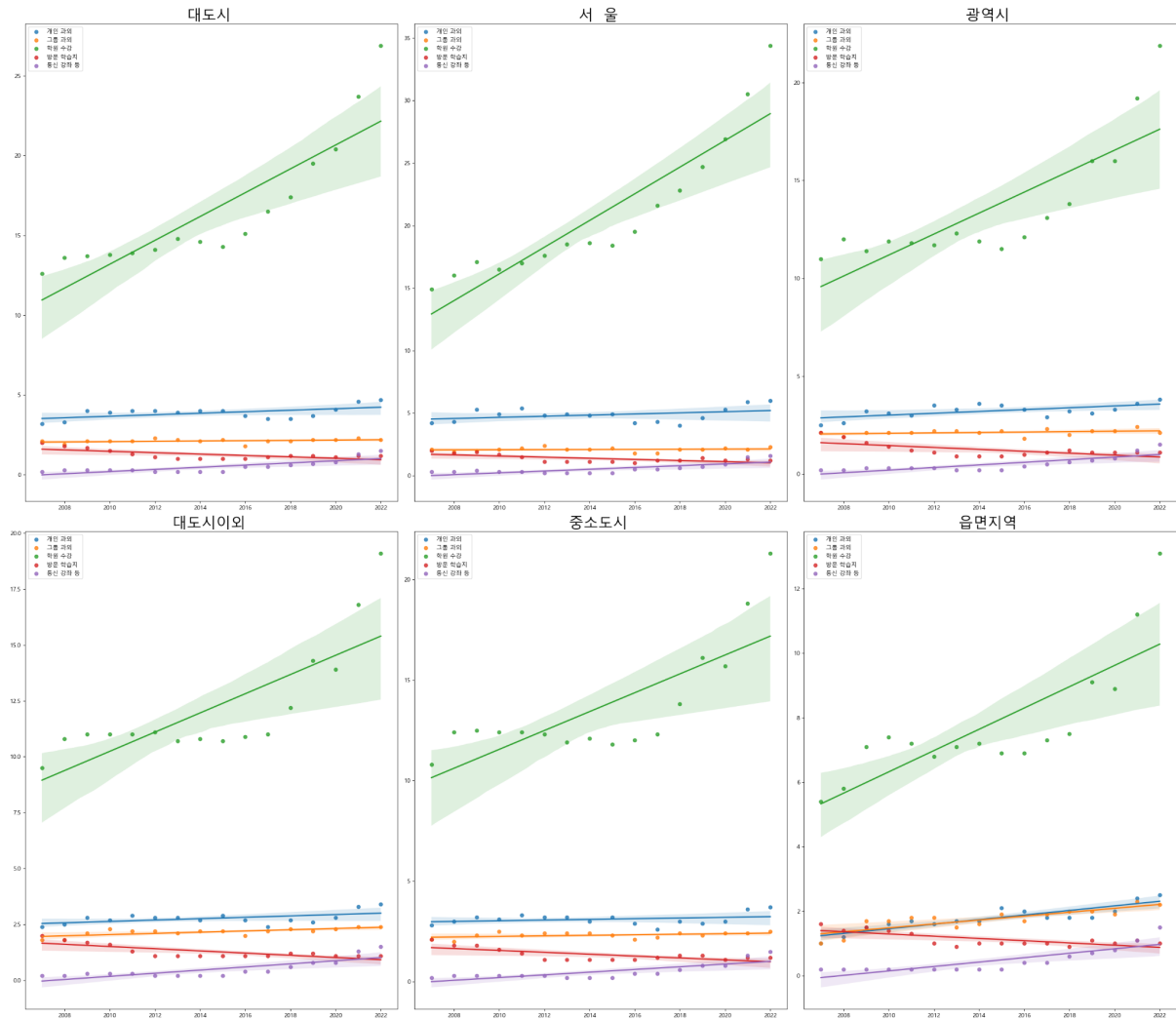
지역별 해법수학 지점수 분포를 나타내는 그래프이다. 경기 지역에서 가장 많은 분포를 보이고 있으며, 서울, 경남 부산 등이 뒤를 따르고, 제주 지역에서 가장 적은 분포를 보이는 것을 알 수 있다.



[그림 2] 지역별 1인당 월 평균 사교육비 추세

대도시와 서울의 월평균 사교육비 추세는 우상향하고 있으며, 특히 영어 수학 과목의 상승세가 가장 가파르다. 영어, 수학의 겨우 대도시 뿐 아니라 중소도시, 읍면지역에서도 가파른 상승세를 보여준다.

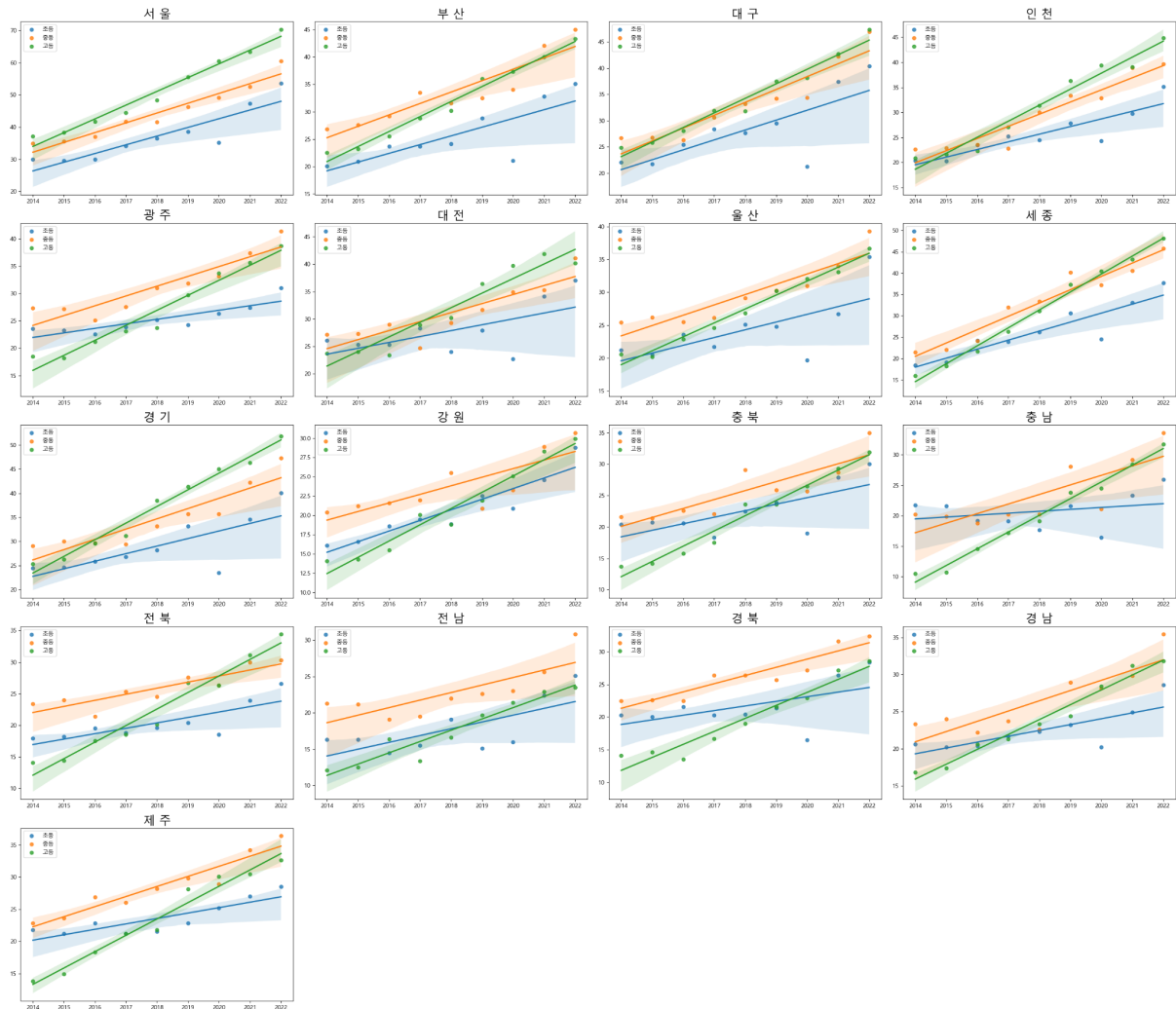
Types of Private Education, by Region
(10,000 KRW)



[그림3] 사교육 경로 추세 (2007-2022)

해법수학이 오프라인 매장임을 고려하여 어떤 경로로 사교육을 주로 소비하는지 분석했다. 학원 수강이 압도적으로 높은 지출을 보여줬고, 방문 학습지의 경우 모든 지역에서 하락하는 추세를 보인다. 통신 학습이 학원 수강과 경쟁구도를 보이지 않을까 예상했지만, 의외로 낮은 소비율과 추세를 보였다.

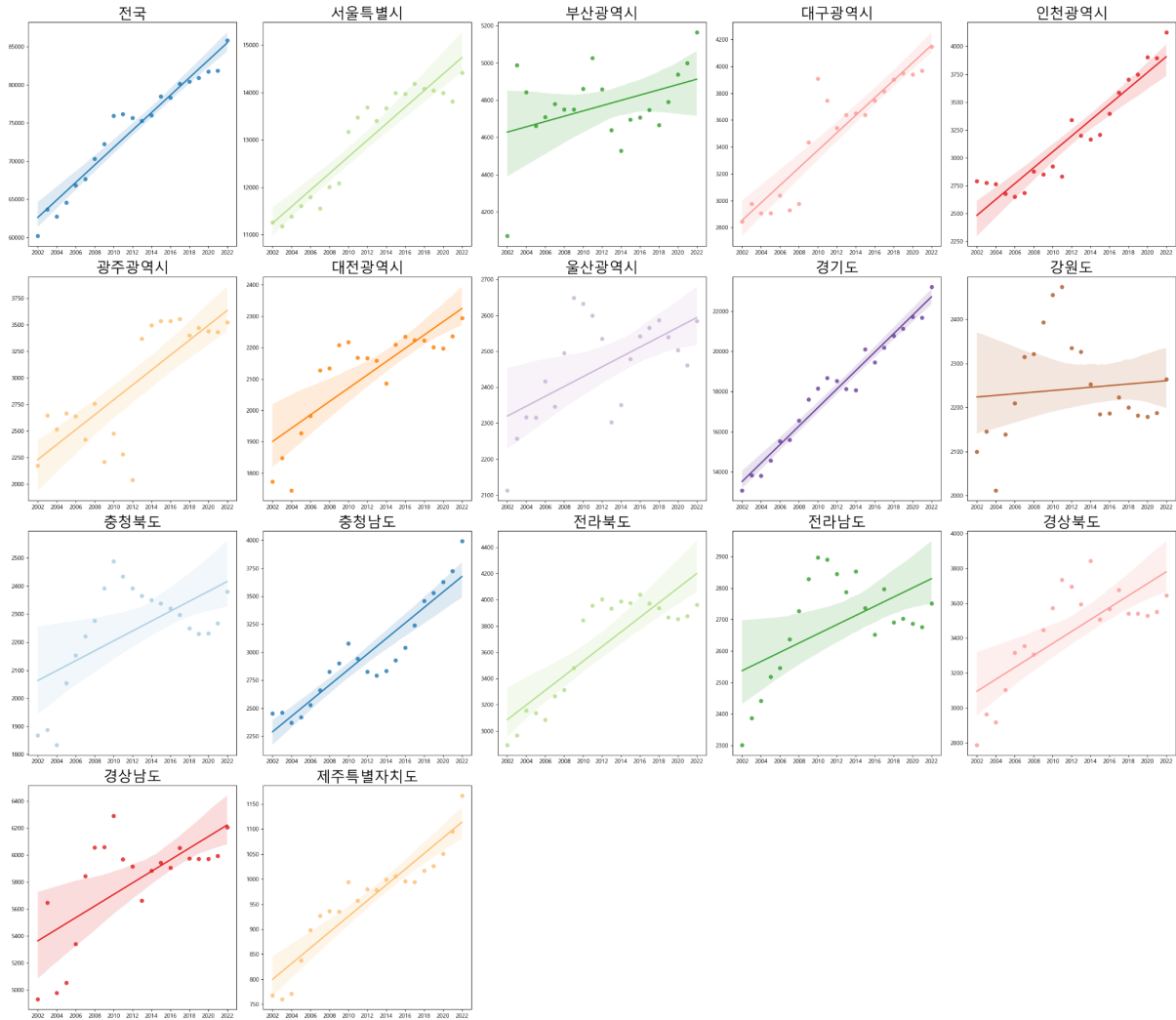
Trend of Private Education Costs, by School Class (10,000 KRW)



[그림4] 초, 중, 고 지역별 사교육비 추세(2014~2022)

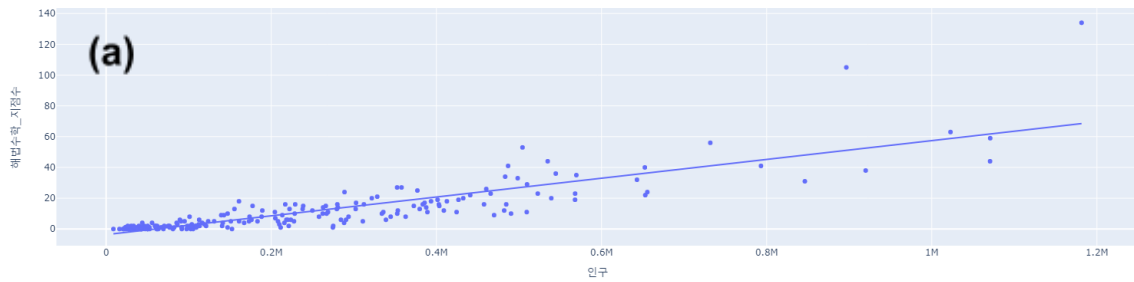
대학 진학을 위한 고등학교에서 가장 높은 사교육비가 나가지 않을까 예상했지만, 의외로 중학교 시기 사교육비 지출이 높은 경우가 많았다. 한 가지 특이한 점은 2014년 대비 고등학교 사교육비 추세가 모든 지역에서 가장 가파른 상승세를 보이고 있다는 것이다.

Number & Trend of All Academy

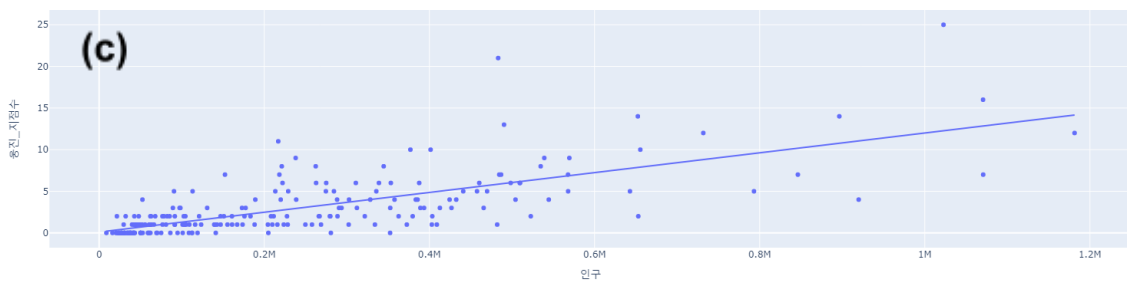


[그림5] 국내 지역별 학원 수 및 추세 (2002-2022)

위에서 정의한 학원은 일반 교과 뿐 아니라 다른 분야의 학원도 포함되어 있으며, 한국교육개발원에서 조사한 데이터를 기준으로 하고 있다. 대한민국의 높은 학구열과 일반 교과 학원 중 업계 상위권을 차지하고 있는 해법수학이 전체 학원 중에서도 높은 비율을 차지하지 않을까 예상했지만, 전체 학원수에 비하면 아주 낮은 비율의 수준이라 그래프에 표시하지 못했다. 편차가 큰 지역도 있지만 대부분의 지역에서 우상향하는 추세를 보인다.



(b) $\text{PearsonRResult}(\text{statistic}=0.8599797252129884, \text{pvalue}=2.9522098292060247\text{e-}68)$

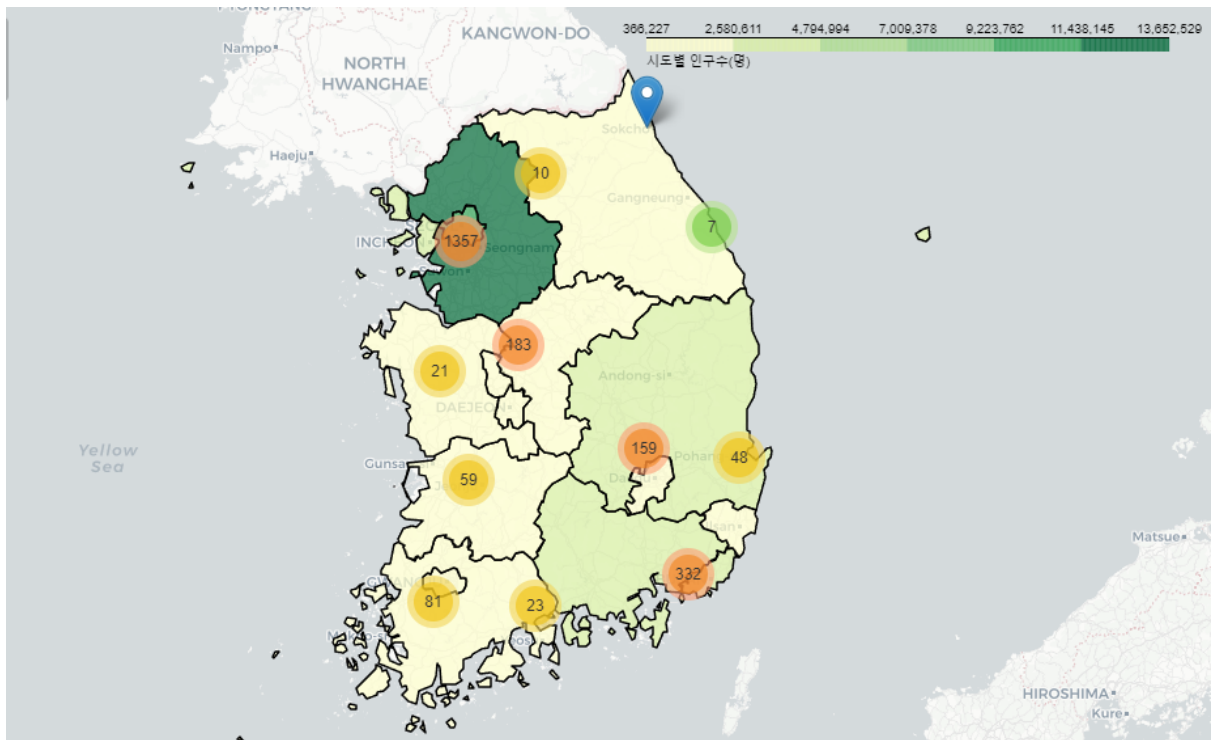


(d) $\text{PearsonRResult}(\text{statistic}=0.7314040762920714, \text{pvalue}=1.3142086621347002\text{e-}39)$

(e) $\text{Z-score} = 3.7910400052349207, \text{p-value} = 7.500889277026627\text{e-}05$

[그림 6] (a) 시군구별 인구수에 따른 해법수학 지점 수 (b) 시군구별 인구수와 해법수학 지점 수 사이의 상관계수 및 p-value (c) 시군구별 인구수에 따른 웅진홈스쿨 지점 수 (d) 시군구별 인구수와 웅진홈스쿨 지점 수 사이의 상관계수 및 p-value (e) 해법수학과 웅진홈스쿨 사이의 상관계수 차이 검정 결과

해법수학과 웅진홈스쿨 모두 인구수와 지점수 사이의 강력한 상관관계를 나타냈다. 이후 두 상관관계 중 해법수학이 더 강력한 상관관계를 나타낸다는 가설을 증명하기 위해 상관계수의 차이를 검정했다. Z-score는 3.79로 나타났고 유의확률이 유의수준 0.05보다 작기 때문에 귀무가설을 기각하고 대립가설을 채택해서 해법수학과 인구수의 상관계수가 웅진홈스쿨과 인구수의 상관계수보다 높다고 주장할 수 있다.



[그림 7] 시도별 인구수(명) 및 시도별 스마트해법 지점 수

각 시도별 인구수(명)와 스마트해법의 지점 수를 지도로 확인한 결과, 인구가 밀집되어 있는 지역에 스마트해법의 지점이 많이 분포되어 있는 것을 확인할 수 있었다. 그래서 인구 중에서도 학원을 이용하는 인구는 주로 학생이기 때문에 학원과 학교와의 거리를 비교했다. 우선, 위도와 경도 데이터를 이용해서 두 점 사이의 거리를 측정했지만, p-value가 약 0.59로 유의하지 않게 나왔다.

```
Ttest_indResult(statistic=0.537360869425262, pvalue=0.5910592150105263)
```

다음으로는 단순히 위도와 경도 값을 점으로 보고 거리를 구하는 것이 아니라 지구가 구의 형태인 점을 감안해서 실제 거리를 구할 수 있는 하버사인 공식을 이용해서 학교와 학원사이의 거리를 측정했지만, 이도 유의수준이 약 0.89로 유의하지 않게 나왔기 때문에 귀무가설을 채택하고, 학원과 학교와의 거리는 업계 내의 점유율과 유의미한 영향을 미치지 않는다고 주장할 수 있다.

```
Ttest_indResult(statistic=-0.12683361932218112, pvalue=0.8990808100115038)
```

IV. 논의

i) [그림 2]부터 [그림 5]까지의 전체적인 추세를 보면 1인당 사교육비 지출은 증가하고 있고 그 중에서도 수학과 영어 과목이 현재도 높고 상승 추세도 높았다. [그림 3]의 사교육 유형으로 보자면 학원을 통한 지출이 가장 높았고 상승 추세도 가파르다. [그림

4]의 초·중·고별 사교육비 추세를 보면 초등학생의 사교육비 지출 증가는 중, 고등에 비해 상대적으로 완만하다. 그에 비해 고등학생의 사교육비는 가파르게 증가하고 있다. [그림 5]의 지역별 학원 수 추세 역시 전체적으로 증가하는 것을 볼 수 있으나 수도권에 비해 지방은 변동이 크다.

해법수학의 입장에서 보면 수학과 영어 과목, 그리고 학원의 사교육비 상승 추세가 가파른 것은 긍정적이다. 하지만 초등학생의 사교육비 지출 증가가 중, 고등학교에 비해 높지 않은 점 그리고 학원 즉 경쟁자가 많아지고 있는 점은 부정적으로 볼 수 있다.

ii) [그림 6]을 보면 해법수학과 경쟁사인 웅진홍스쿨 모두 지역별 인구수와 지점 수 사이에 강한 상관관계가 있는 것을 확인할 수 있다. 그리고 이 두 상관관계수 사이의 차이를 알아보기 위해 Z-검정을 한 결과 $p < .0001$ 로 해법수학이 훨씬 더 강력한 상관관계를 나타낸다는 결과를 볼 수 있다. 이를 통해 해법수학이 수학 프랜차이즈 점유율 기준 1위인 이유 중 하나를 알 수 있는데 적절한 수요가 있는 지역에 가맹점을 개설함으로써 알맞은 공급을 하고 있는 것이다.

iii) 학교와의 거리는 해법수학과 웅진홍스쿨 사이에 유의미한 차이가 없는 것으로 나타났다. 그러나 이 결과를 가지고 학교와의 거리가 중요하지 않다고 해석하는 것은 적절하지 않다. 학교와의 거리가 중요한 만큼 경쟁사도 이를 1차적으로 고려하기 때문인 것으로 생각하는 것이 적절할 것이다. 다만 학교와의 거리는 학부모들이 다른 학원이 아닌 해법수학을 선택해야만 하는 요인은 아니라고 볼 수 있다.

iv) 분석결과 해법수학 학원 지점은 인구 수가 많은 지역에 비례하여 분포하고 있다는 사실을 알 수 있다. 점유율 측면에서 우위를 차지하기 위해서는 인구수가 많고, 학생이 많이 분포하는 지역에 지점을 내는 것이 필요하다. 인구수가 지속적으로 성장하여 학생수가 많이 분포할 가능성이 있는 지역은 경기, 세종, 충북 지역으로 이 지역에 지점을 낸다면 타 업체보다 우위를 차지할 수 있을 것이다.

V. 코드

프로젝트를 수행하기 위해 만든 데이터 및 코드는 다음의 Github에서 확인할 수 있다.

(URL: <https://github.com/bigdata4th-first-line/Smarthb.git>)

VI. 사용 라이브러리

pandas Python package (version 1.5.3)

plotly Python package (version 5.9.0)

selenium Python package (version 4.10.0)

Matplotlib Python package (version 3.7.0)

geopy Python package (version 2.3.0)

folium Python package (version 0.14.0)

json Python package (version 2.0.9)

dash Python package (version 2.9.3)
scipy Python package (version 1.10.0)
statsmodels Python package (version 0.14.0)

VII. 참고문헌

한영은, 이승철 "안양시 평촌 학원가의 교육 서비스 실태 및 공간 범위에 관한 연구"
한국경제지리학회지 15.4 pp.721-734 (2012) : 721.

김태환, 김은란, 신휴석, 이혜민, 박미래, 이혜진 "지역별 소득 격차와 불균형" 국토연구원
균형발전 모니터링 & 이슈 Brief, 제7호 (2021)

강용호. "프랜차이즈 시스템 확장을 위한 상권 및 입지분석 모형설정에 관한 연구."
국내석사학위논문 漢陽大學校 都市大學院, 2009. 서울

박소현, 이금숙 "사교육 시설의 수요와 공급에 나타나는 공간적 특성: 수도권 지역
사설학원을 중심으로" 한국경제지리학회지 14.1 pp.33-51 (2011) : 33.

Fisher, Ronald Aylmer. "014: On the " Probable Error" of a Coefficient of Correlation
Deduced from a Small Sample." (1921).