

2019 Spring Data Analytics

BADA

Budongsan Analytics
Data Analytics

FINAL

X-세권의 영향력 파악 및 프리미엄 정도 예측

201411180 정재민
201311167 이승윤
201411160 송용백
201514181 박영재
201611171 제갈용승

INDEX

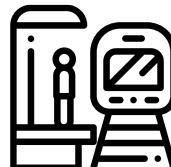
- 0. Project Review
- 1. Project Objective
- 2. Feature Engineering
 - 2.1 Data Collection
 - 2.2 Data Preprocessing
- 3. Data Analysis
 - 3.0 Theoretical Background
 - 3.1 Feature Selection
 - 3.2 Prediction
- 4. Discussion
- 5. Conclusion

0. Project Review



0. 1. 1. 역세권

“역세권의 실질적 범위와 프리미엄 정도 측정”



역세권이란?

지하철역으로의 접근이 용이한 범위



역세권의 실질적 범위

법에서 명시하는 1차 역세권 250m와는 다른 의미
건물의 가격에 영향을 주는 범위를 의미

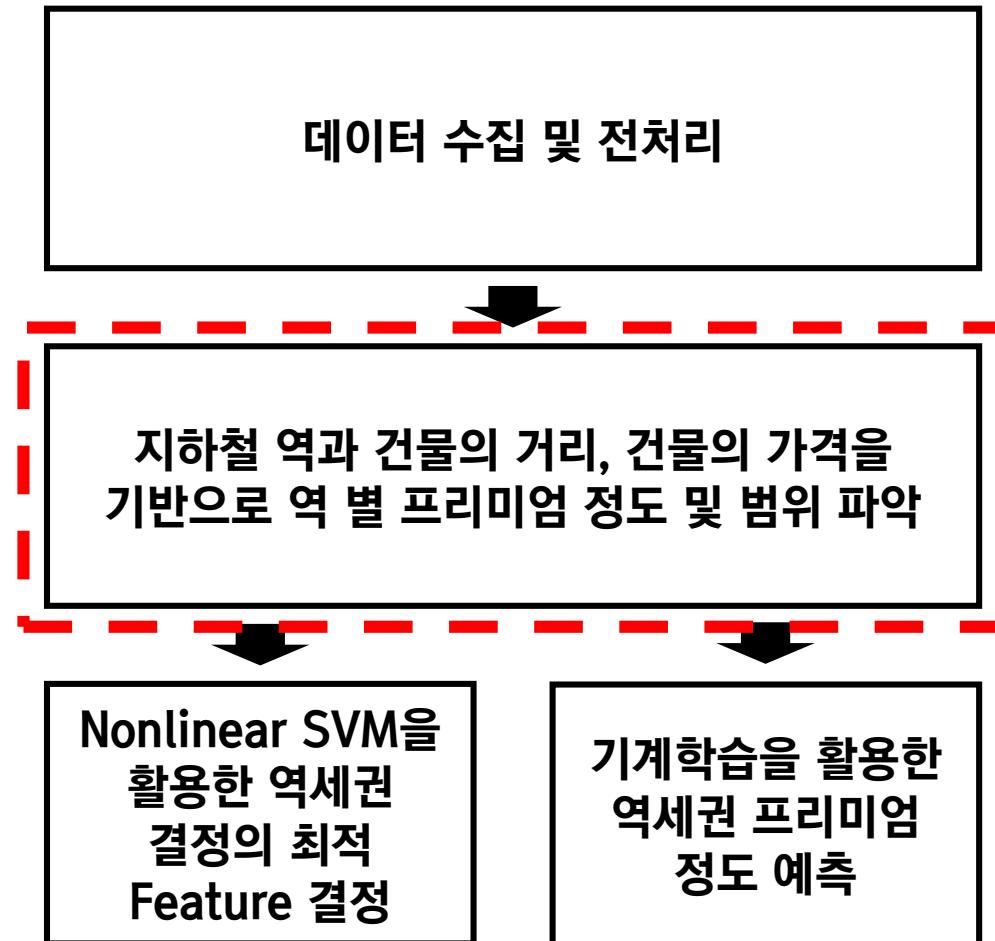


역세권의 프리미엄

역세권에 속함으로써 증가한 비용

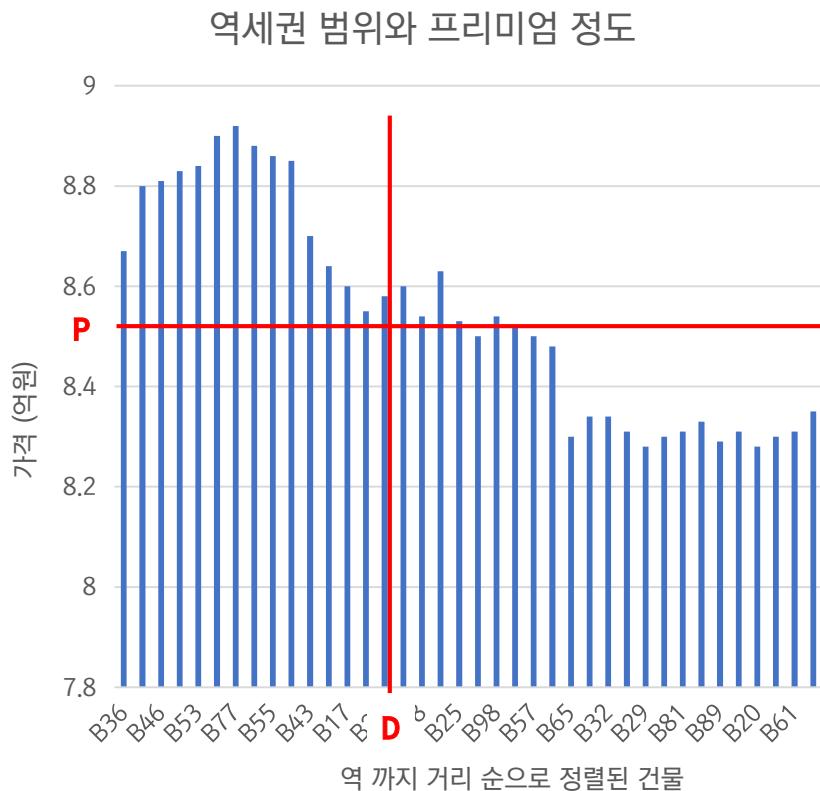


0.1.2. 분석 절차





0. 1. 3. 예상결과



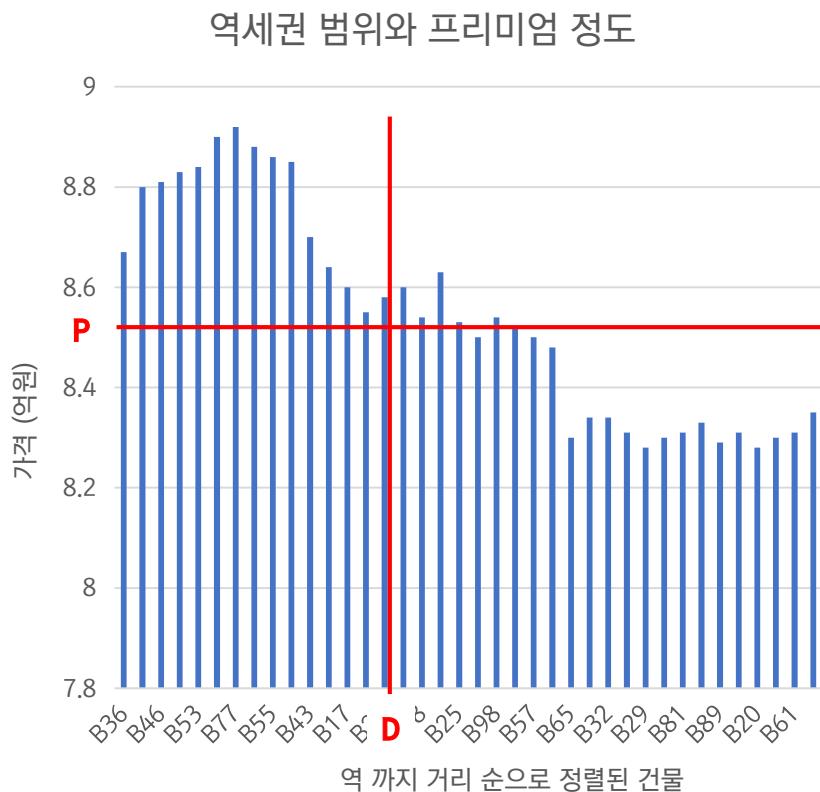
D: 역세권의 실질적 범위

P: 해당 자치구의 평균 매매가

1. 모든 건물들을 가장 가까운 역으로 Labeling
2. 역에 해당하는 건물들을 거리 순으로 정렬
3. 거리 순으로 정렬된 건물들의 가격을 bar graph로 표현



0. 1. 3. 예상결과



D: 역세권의 실질적 범위

P: 해당 자치구의 평균 매매가



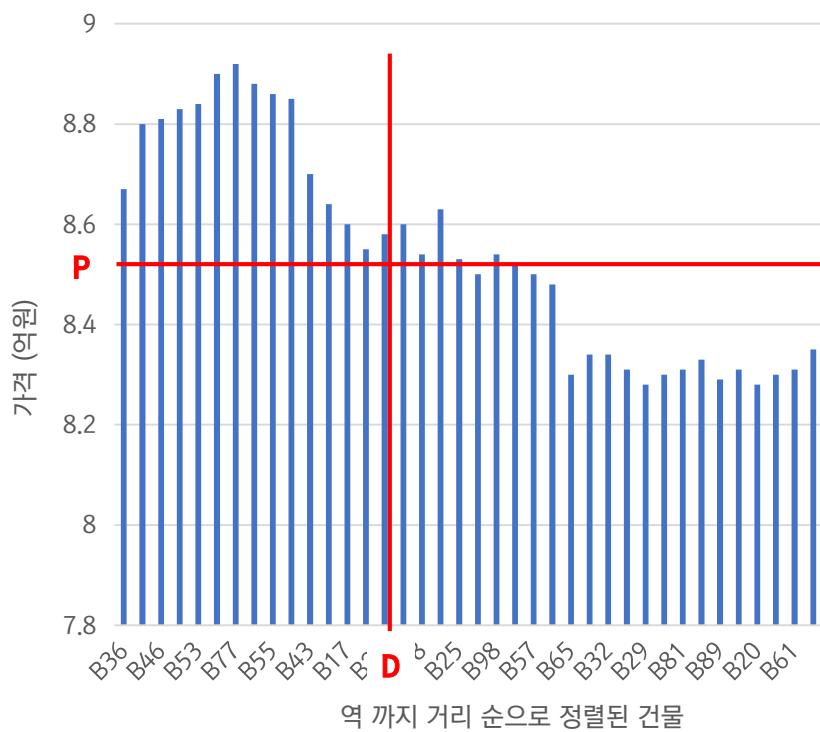
● 250m로 정의된 역세권 범위

● 분석을 통해 정의한 역세권의 실질적 범위



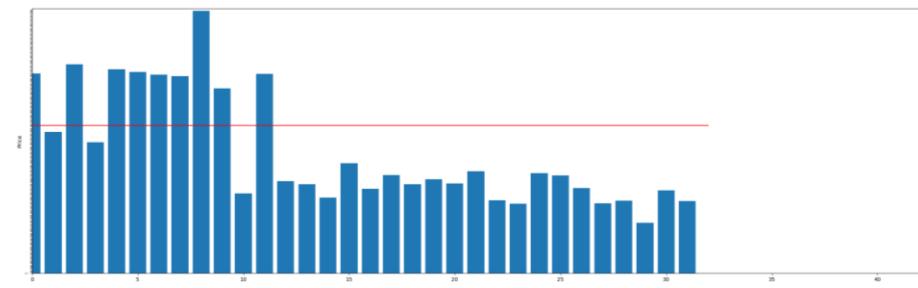
0.1.4. 기초분석 결과

역세권 범위와 프리미엄 정도

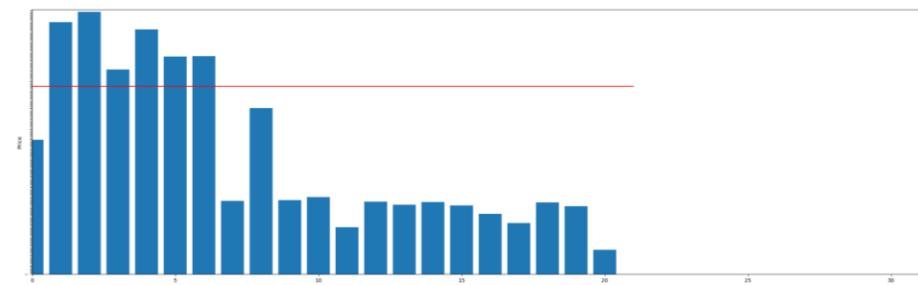


D: 역세권의 실질적 범위

P: 해당 자치구의 평균 매매가



버티고개역



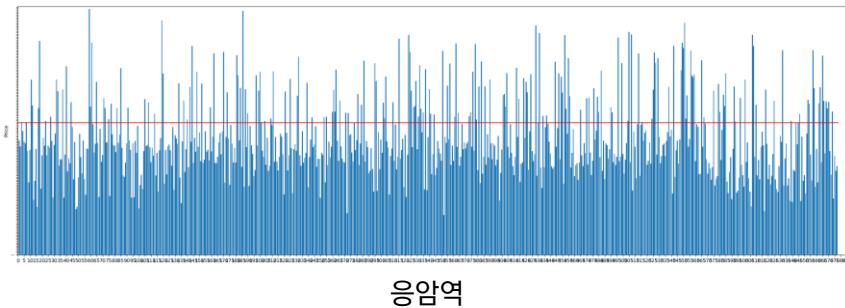
은봉역



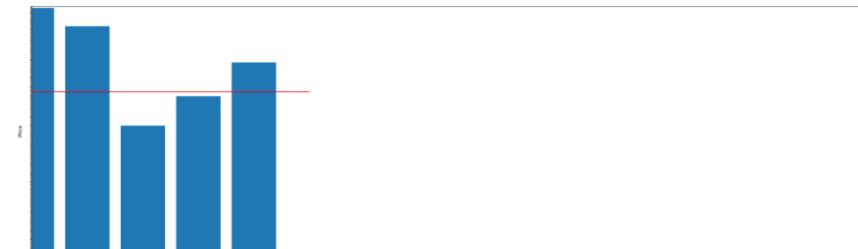
0.1.4. 기초분석 결과

하지만, 대부분의 지하철역이 다음과 같은 추세를 보이기 때문에
역세권의 실질적 범위와 프리미엄 정도를 정의하기가 어려움

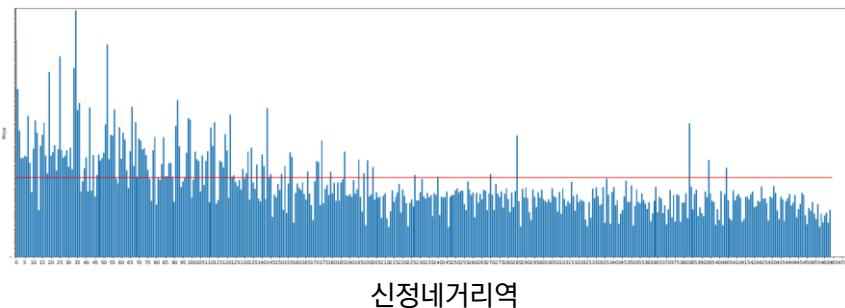
1. 그래프에 추세가 보이지 않음



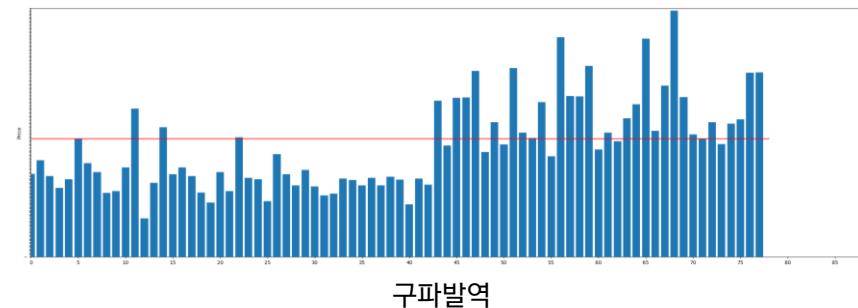
2. 역에 해당하는 건물이 많이 없음



3. 범위를 지정하기 애매함



4. 거리가 멀어질수록 가격이 올라감





0. 1. 4. 기초분석 결과

**역까지의 거리와 건물의 가격만으로는
역세권의 실질적 범위와 프리미엄을 측정할 수 없음**



0. 2. 1. New Approach

**건물의 가격에 영향을 주는
X-세권을 파악하는 분석을 수행하고자 함**



0. 2. 2. 데이터 수집 및 전처리

데이터 전처리 결과	Feature
건물의 특성	평균 거래금액, 건설연도, 평균 전용면적, 평균 대지권면적,
건물에서 역까지의 거리	서울에 존재하는 모든 지하철 역
One-Hot Matrix	자치구, 지하철 노선
최근접 시설 3개와의 거리	유치원, 초등학교, 중학교, 고등학교, 대학교, 주민센터, 치안기관, 영화관, 백화점, 대형마트, 공원, 병원
해당 자치구 소속의 시설과의 거리	구청, 보건소
해당 자치구의 특성	구 범죄율

- 연립주택(건물) 데이터는 주소를 기준으로 group by 하였으며, 이를 기반으로 건물과 시설 사이의 거리를 산출함
- 결과적으로, (26542, 358) 형태의 데이터셋 구축 완료
 - 26542개의 건물 × 건물의 가격을 평가할 수 있는 358개의 Features



0. 2. 3. 상관관계 분석

- 전체 데이터 셋에 대한 상관계수를 도출한 후, 로그로 스케일 된 평균 거래금액에 대해 상관계수가 큰 10개 Feature, 상관계수가 작은 10개 Feature를 추출

양의 상관관계		음의 상관관계	
평균 전용면적	0.533518	잠원	-0.475458
평균 대지권 면적	0.477713	반포	-0.471935
서초구	0.259206	논현	-0.470685
구파발	0.249685	사평	-0.467124
용산구	0.227830	신논현	-0.463791
구 범죄율	0.226564	학동	-0.460123
연신내	0.216137	신사	-0.457365
강남구	0.214923	언주	-0.456791
유치원_3rd	0.211526	한남	-0.454278
개화	0.207907	강남	-0.454188

- 강남, 잠원, 반포, 논현, 신논현, 신사 등 강남구에 위치한 역과의 거리 가까울수록 주택 가격에 긍정적인 영향을 미침
- 구파발, 연신내, 개화 등의 역과의 거리가 가까울수록 주택 가격에 부정적 영향을 미침
- ‘평균 대지권 면적’이 클수록, ‘구 별 범죄율’이 높을수록, ‘유치원’과의 거리가 멀수록 주택 가격에 긍정적인 영향을 미침
- ‘서초구’, ‘강남구’에 속한 주택은 상대적으로 높은 주택 가격을 가짐



0. 2. 4. Machine Learning: Linear Regression

회귀모델 생성에 필요한 데이터 분리

- Y: 평균 거래금액 (로그 스케일)
- X: 전체 데이터 셋에서 평균 거래금액을 제외한 모든 Feature

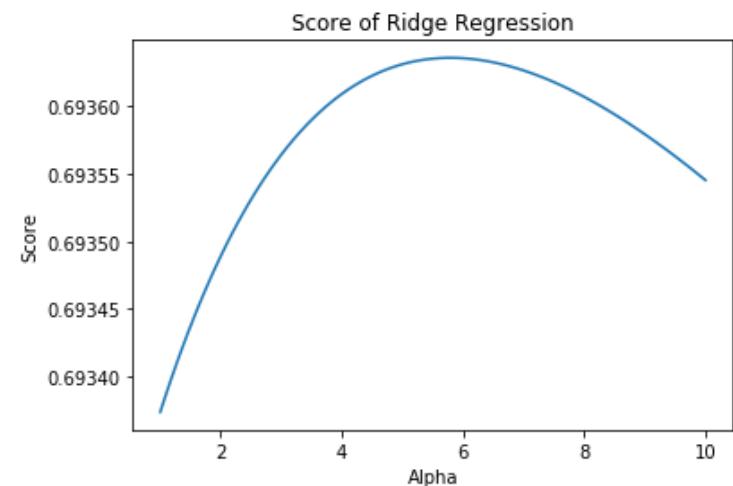
Multi Linear Regression Analysis 수행

- Regularization을 거친 선형회귀 모델인 Ridge Regression의 성능이 0.693636으로, 단순 선형회귀분석보다 미약하게 더 좋은 것을 확인할 수 있음



```
R-square of Linear Regression  
score = 0.693204
```

```
Highest R-square of Ridge Regression  
alpha = 5.818182  
score = 0.693636
```



- 본 프로젝트가 분석하고자 하는 분야는 부동산 역시 경제학 및 사회과학 분야에 속하기 때문에, Ridge 모델을 통해 도출한 R-square값 0.6936는 충분히 유의미함

1. Project Objective



I. 1. 주택 가격에 영향을 미치는 X-세권

“X-세권의 영향력 파악 및 프리미엄 정도 예측”



X-세권이란?

연립 다세대 주택 가격에 영향을 미치는 건물 외부적 특성요인



영향력 파악

371개 Feature 중 영향력 있는 요인 도출 및 영향력 파악



프리미엄 정도 예측

Deep Learning을 포함한 다양한 머신러닝 기법을 통하여 X-세권이 야기하는 가격변화 예측



I. 2. 불분명한 X-세권의 범위와 영향



- ✓ 광고 속 X-세권 건물의 도보 거리와 실제 도보 거리가 다른 경우가 상당히 많음
- ✓ 최근 숲세권, 직세권 등의 무분별한 X-세권 프리미엄 광고가 빈번함
- ✓ 단일 X-세권 하나를 기반으로 주택 가격에 주는 영향을 확인할 수 없음



I. 3. 주택 가격에 영향을 미치는 X-세권

‘숲세권’이 주택의 선택 시 중요한 요인이 된다. 주택산업연구원 「미래 주거 트렌드 연구」

주택시장에서 학교와 인접한 이른 바 ‘학세권’ 단지는 수요가 꾸준하다고 알려져 있다. 매일경제 「학교 인접 학세권 단지 인기.. ‘개교 효과’ 톡톡」

부동산시장에서 메디컬타운, 체육공원, 운동시설 등 건강과 관련된 인프라가 인기다. (중략) ‘의세권’, ‘병세권’ 등 대형병원이 인접한 단지도 각광 받고 있다. 브릿지경제 「부동산도 100세 시대, '의세권 · 공세권' 눈길」





I. 4. 현상황 및 문제점

X-세권 및 주택 가격 관련 이슈 및 분석 필요성

- ✓ 정부는 주택 시장의 안정화를 위한 정책의 구체적인 실현을 위해서는 정확한 데이터를 바탕으로 정책 수립이 필요 (2019.3, 이재삼)
- ✓ 다양한 유형의 주택들이 공존하여 주택마다 대지면적과 건축물 규모 등이 상이하기 때문에 아파트에 비해 주택의 공정한 가치 평가가 어려움 (2019.3, 정우성, 송선주, 신종칠)
- ✓ 주택의 임대료 가격 결정에 대한 객관적인 자료가 부족하여, 어떠한 요인으로 인하여 결정되는지 알지 못하는 경우가 많음 (2018.8 김보찬, 김유현, 김민정, 이종석)
- ✓ 주거 중심 역, 주거 및 상업 중심 역, 주거, 상업, 업무 등 역사의 유형화를 통한 X-세권 분석 필요
- ✓ 자연적, 환경적, 경제적 건축물의 분포 등의 복수의 특성을 분석하여, 각 특성 별 주택 가격에 주는 영향을 주는 다양한 X-세권을 파악 할 필요가 있음





I. 4. 현상황 및 문제점

X-세권의 및 주택 가격 관련 분석을 위한 선행연구

- ✓ 단독주택 재건축지역의 주택가격 영향요인에 관한 연구 (2019.2, 정우성, 송선주, 신종칠)
- ✓ 서울시 지하철 역세권의 공간적 범위 설정과 특성 분석 (2011.12, 이연수, 추상호, 강준모)
- ✓ 주거용과 상업용 부동산의 가격 결정 요인 비교(2012.8, 박성균, 이현석)

X-세권 관련 주택 가격 관련 현 상황의 문제점

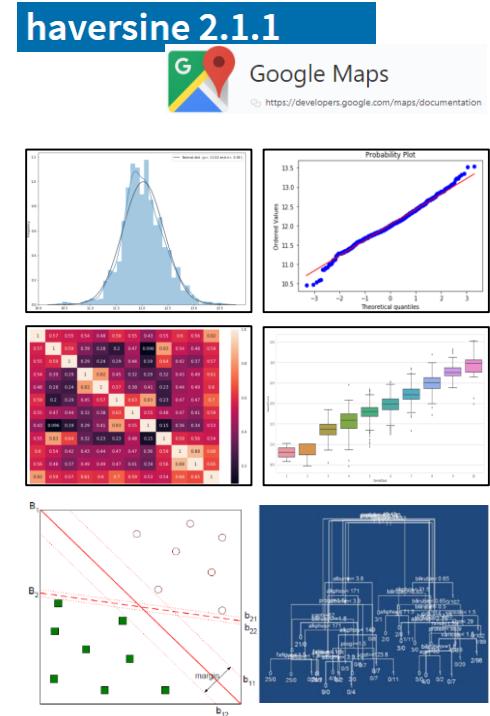
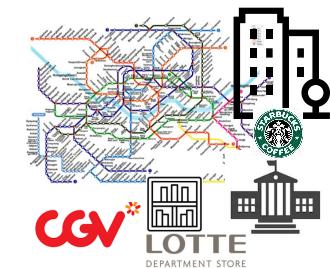
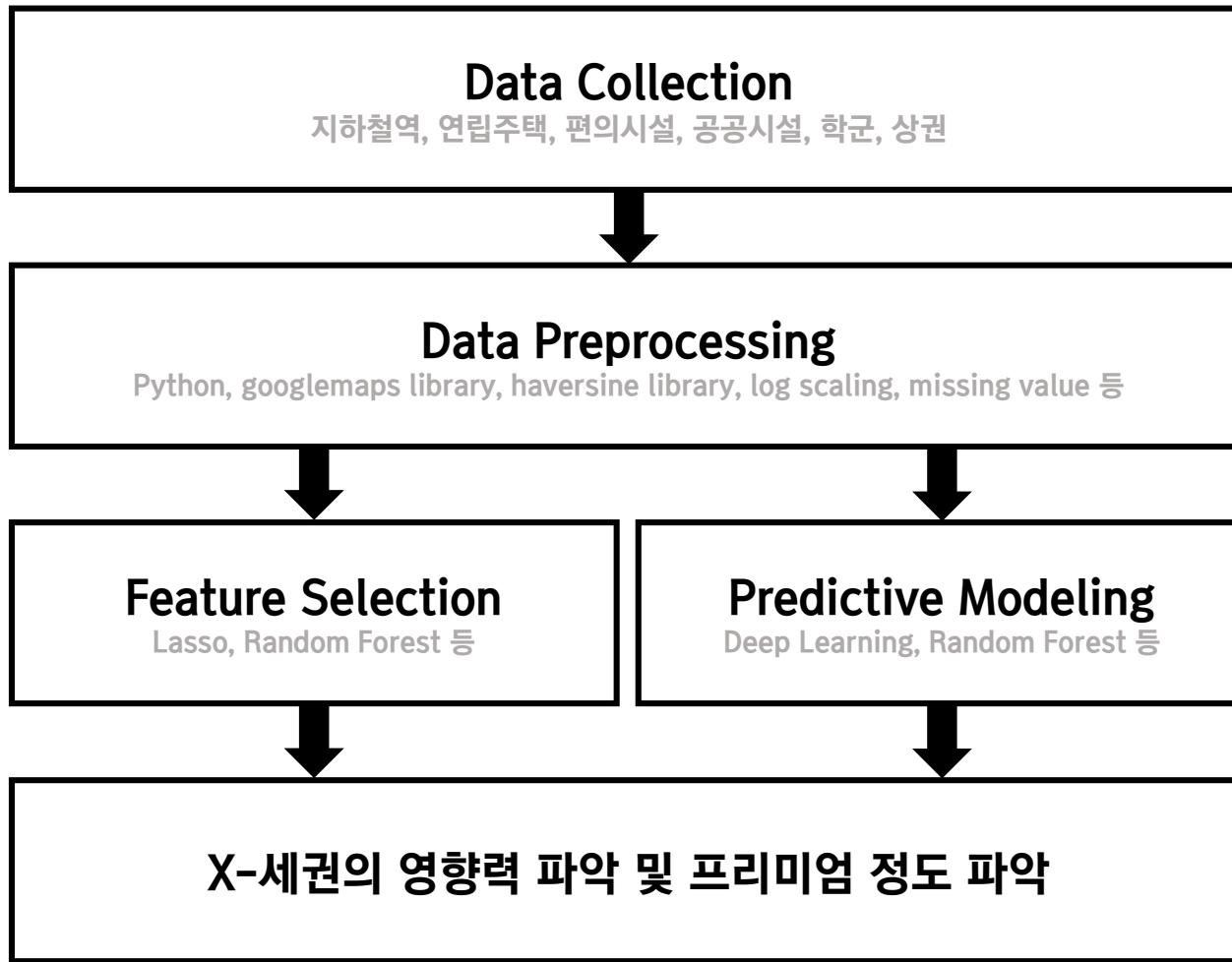
- ✓ 단순히 역과의 거리만으로 주택의 가격을 예측할 수 없음
- ✓ 기존에 진행되었던 연구들은 소량의 데이터를 사용하였기 때문에 일반적 적용이 어려움
- ✓ 최근 역세권을 비롯한 숲세권, 직세권 등의 무분별한 X-세권 프리미엄 광고가 빈번함
- ✓ X-세권의 범위가 불분명하여 정책 개발, 투자 등에 의사결정에 부정적인 영향을 미침

대량의 데이터를 과학적으로 분석함으로써

- 1) X-세권의 영향력 파악
- 2) X-세권의 프리미엄 정도 예측



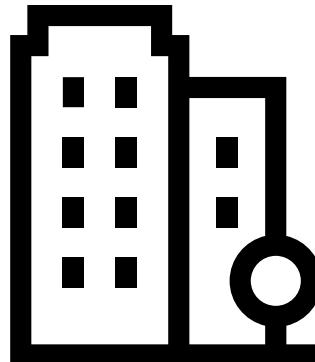
I. 5. 분석절차



2. Feature Engineering



2.1.1. 수집데이터



지하철 역 관련 데이터

- 각 호선 별 정보
- 환승 여부

수집 방법 및 출처

- 공공데이터 포털
- 서울 열린데이터 광장



건물 관련 데이터

- 연립 / 다세대 주택 가격

수집 방법 및 출처

- 국토교통부 실거래가 공개 시스템

상권 및 공공기관 데이터

- 영화관, 백화점, 지하상가
- 유치원, 초, 중, 고등학교
- 구청, 자치센터
- 소상공인

수집 방법 및 출처

- 공공데이터 포털
- 서울 열린데이터 광장
- Web Scrapping



2. 1. 1. 수집데이터

- 서울에 존재하는 연립주택(이하 건물)의 가격에 영향을 줄 수 있는 외부 요인 데이터를 분석에 사용함
- 크게 건물, 지하철 역, 편의시설, 공공시설, 학군, 상권으로 분류되는 데이터를 수집함

데이터	데이터 설명	출처	기간 또는 데이터 생성일
건물	서울시 연립다세대 주택 및 실거래 정보	국토교통부 실거래가 공개시스템	2018.01~2018.12
지하철 역	서울시 지하철 호선별 역별 승하차 인원 정보	서울열린데이터광장	2015.01~2019.03
영화관	서울시 위치, 브랜드 별 영화관 정보	네이버 영화관 정보	2019.05 (웹 크롤링)
백화점	서울시 내 백화점 위치 데이터	네이버 지도	2019.05 (웹 크롤링)
대형마트	서울시 내 대형마트 위치 데이터	네이버 지도	2019.05 (웹 크롤링)
공원	서울시 공원 현황 데이터	서울시 정보소통광장	2018.04.06
치안기관	서울시 각 구별 경찰서, 파출소, 지구대 위치 정보	서울열린데이터광장	2016.03.03
병원	서울시 내 대형병원 데이터	서울열린데이터광장	2018.11.12
소상공인	서울시 내 소상공인 데이터	공공데이터포털	2019.03.12



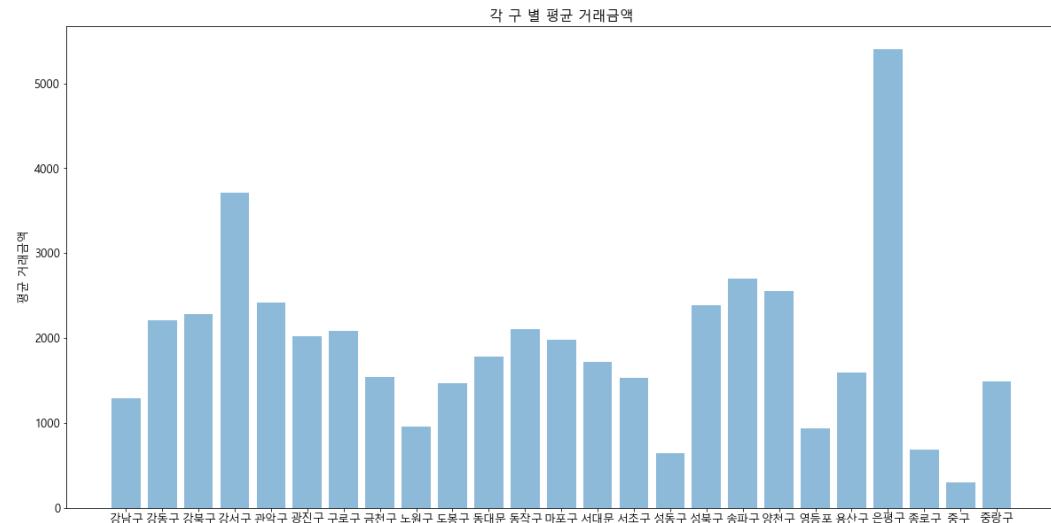
2.1.1. 수집데이터

- 서울에 존재하는 연립주택(이하 건물)의 가격에 영향을 줄 수 있는 외부 요인 데이터를 분석에 사용함
- 크게 건물, 지하철 역, 편의시설, 공공시설, 학군, 상권으로 분류되는 데이터를 수집함

데이터	데이터 설명	출처	기간 또는 데이터 생성일
주민센터	서울시 소재의 주민센터 정보	서울열린데이터광장	2016.02.16
구청	서울시 소재의 구청 정보	서울열린데이터광장	2016.02.16
보건소	서울시 소재의 보건소 정보	서울열린데이터광장	2016.02.16
범죄율	서울시 행정구역별 5대 범죄발생 누적 정보	서울열린데이터광장	2012.03~2018.03
유치원	서울시 구별 유치원 이름, 주소 등 공간 정보	공공데이터포털	2017.04.01
초등학교	서울시 구별 초등학교 이름, 주소 등 공간 정보	공공데이터포털	2017.04.01
중학교	서울시 구별 중학교 이름, 주소 등 공간 정보	공공데이터포털	2017.04.01
고등학교	서울시 구별 고등학교 이름, 주소 등 공간 정보	공공데이터포털	2017.04.01
대학교	서울시 소재 대학교 정보	네이버 지도	2019.05 (웹 크롤링)



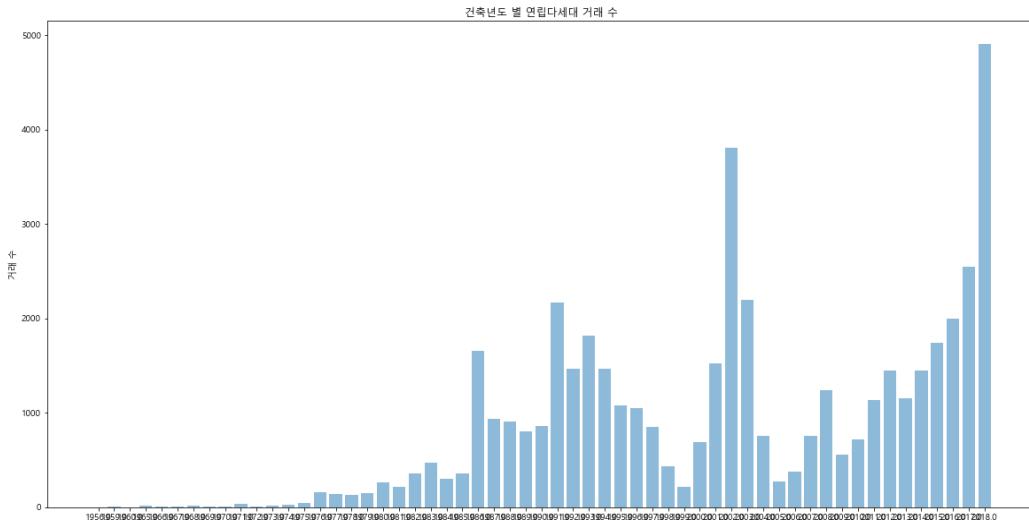
2.1.2. 수집데이터: 건물



- 연립다세대 주택 평균 거래금액
- 2018년 1월 ~ 2018년 12월까지의 구별 연립 다세대 주택 평균 거래 가격
- 작년 서울시 연립 다세대 주택 평균 거래 가격은 2억 6천 8백 80 만원
- Data shape: (47746, 11)



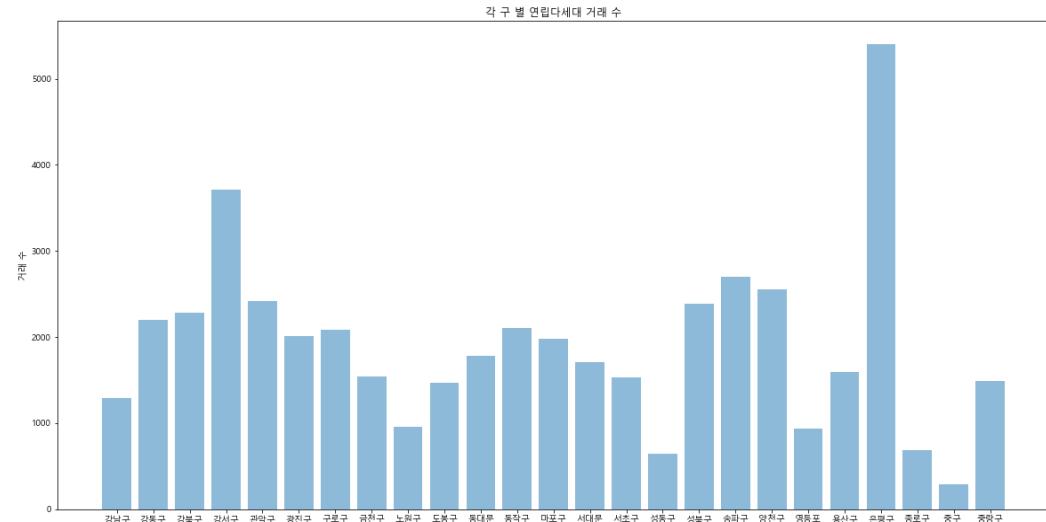
2.1.2. 수집데이터: 건물



- 건축연도 별 연립다세대 주택 거래량
- 1956년 ~ 2018년까지 건축된 연립다세대 주택 건축연도 별 평균 가격
- 가장 최근 지어진 2018년 거래량이 5858건으로 가장 많음
- Data shape: (47746, 11)



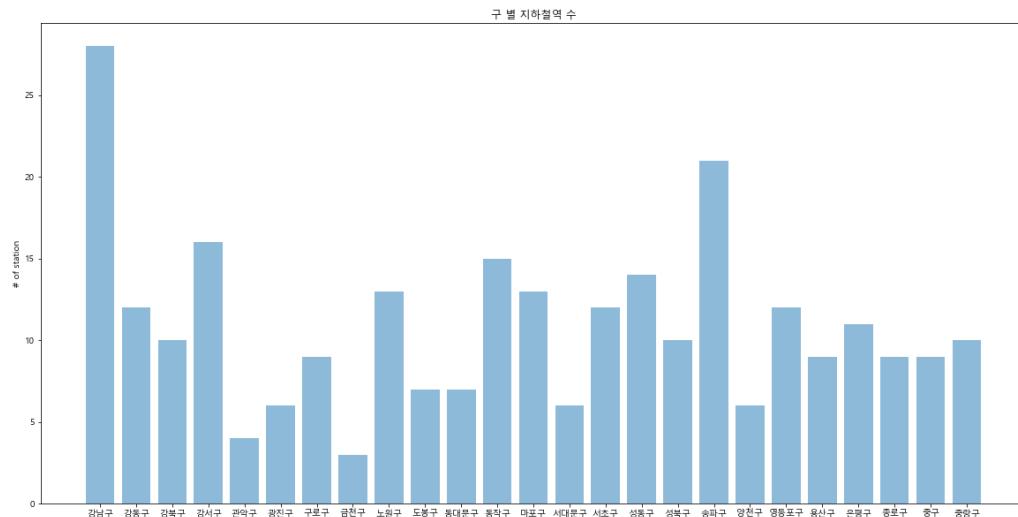
2.1.2. 수집데이터: 건물



- 구 별 연립다세대 주택 거래량
- 1956년 ~ 2018년까지 건축된 연립다세대 주택 건축연도 별 거래량
- 은평구 연립다세대 주택 거래량이 5404건으로 가장 많음
- Data shape: (47746, 11)



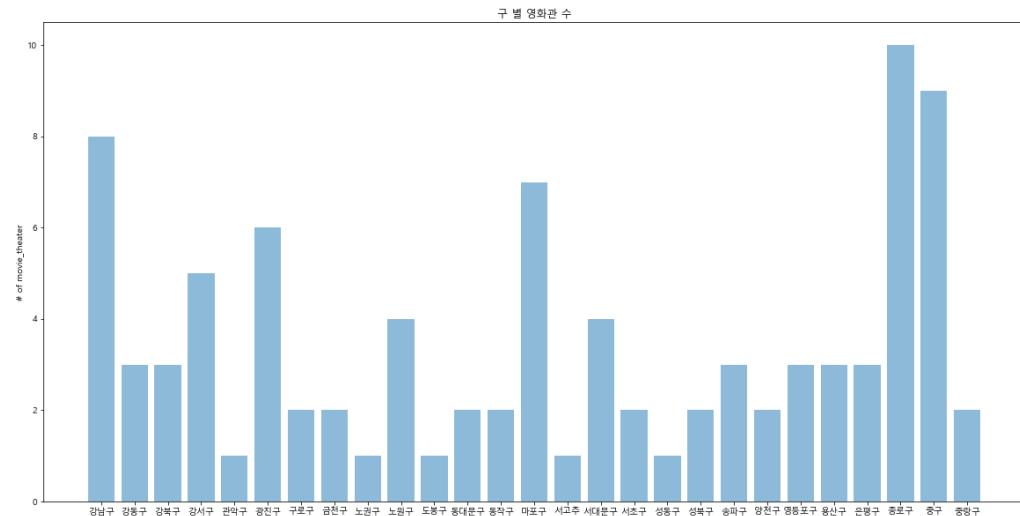
2.1.2. 수집데이터: 지하철 역



- 2015년 1월 ~ 2019년 3월까지의 지하철 역별 승하차 인원 정보
- 지하철 역 이름 전처리 후 역 이름으로 group by한 뒤 서울에 있는 역만 추출
- 지하철 역에 대한 영향을 파악하기 위한 필수적인 자료
- Data shape: (29127, 53)



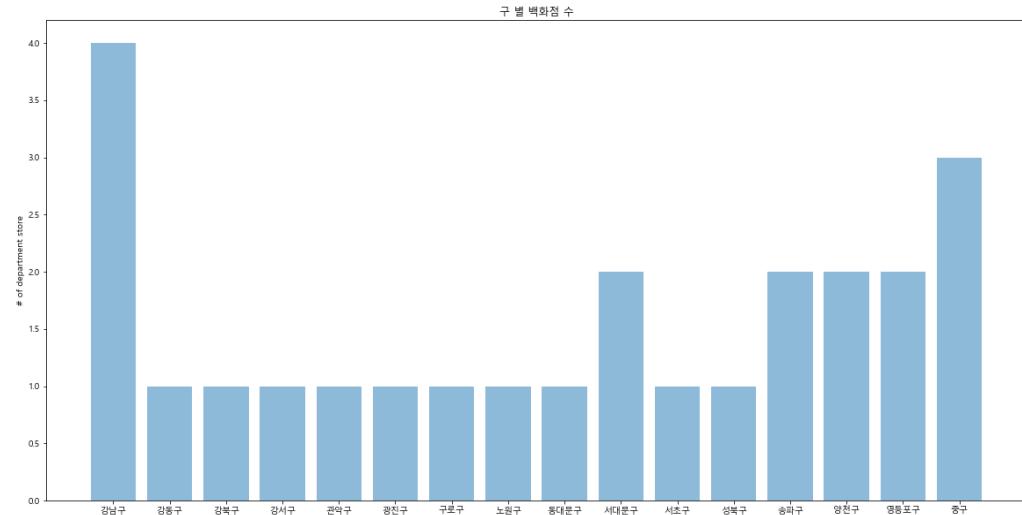
2.1.2. 수집데이터: 영화관



- 2019년도 현재 운영중인 구별 영화관 정보
- ‘네이버 영화관 정보’에서 제공하는 CGV, 메가박스, 롯데시네마와 기타 영화관 포함
- 쇼핑몰, 대형마트와 함께 ‘몰세권’이라 칭해지는 부동산의 가치를 높이는 요소로써 문화생활을 대표함
- Data shape: (92, 5)



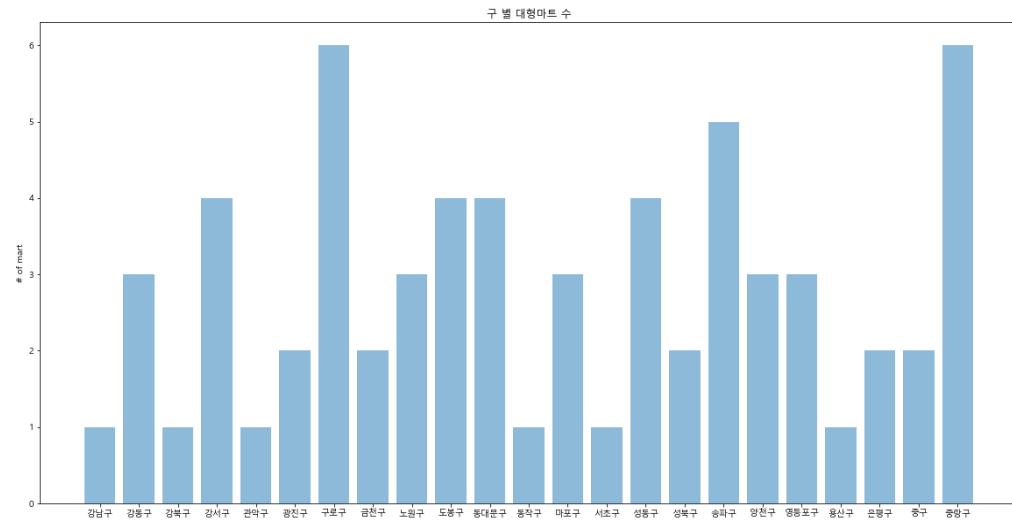
2.1.2. 수집데이터: 백화점



- 2019년 5월 15일 기준 백화점(롯데, 신세계, 현대) 정보
- 강남구에 특히 많은 백화점이 존재하는 것이 확인됨
- ‘백세권’이라 불리는 부동산 가치를 높이는 요소로 문화생활을 대표
- Data shape: (25, 5)



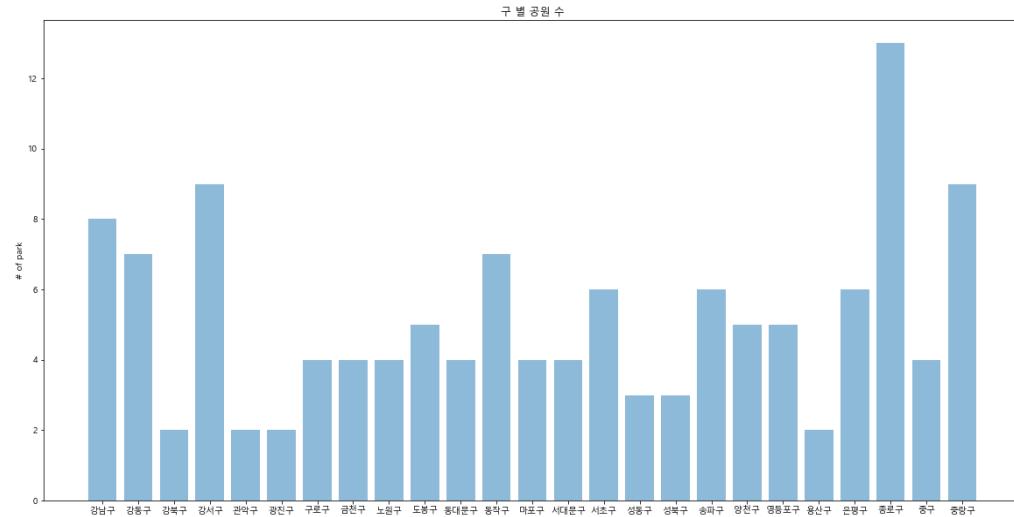
2.1.2. 수집데이터: 대형마트



- 2019년 5월 15일 기준 대형마트(홈플러스, 이마트, 롯데마트) 정보
- 타 구보다 구로구, 중랑구에 많은 대형마트가 존재하는 것으로 보임
- 대형마트 인접성 여부로 집값 상승폭 2배 차이… <건설경제>
- Data shape: (64, 5)



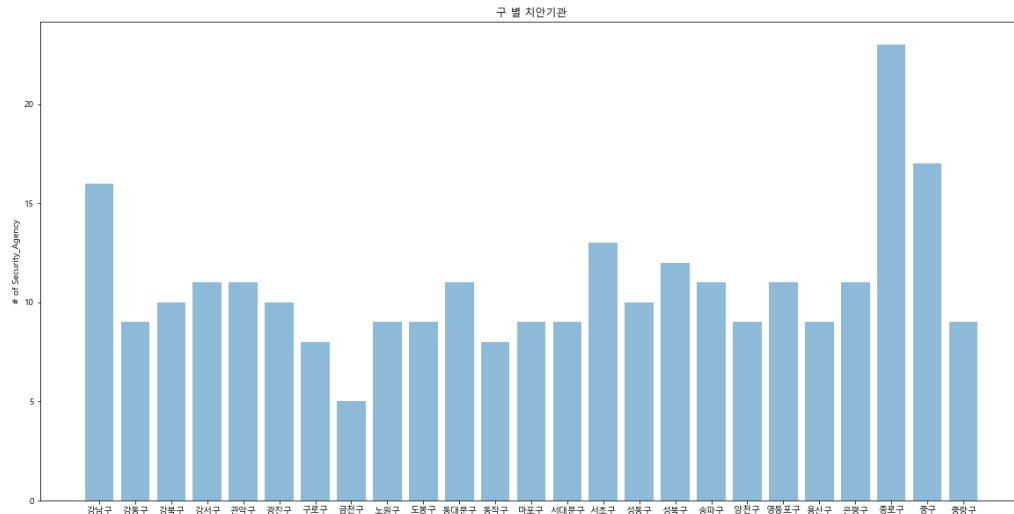
2.1.2 수집데이터: 공원



- 2018년 4월 6일에 생산된 서울시 출처의 공원 현황 데이터
- 공원은 ‘숲세권’이라는 단어를 대표할 수 있음
- 강남, 역세권, 공원은 집값의 3대 키워드… <조선일보>
- Data shape: (126, 5)



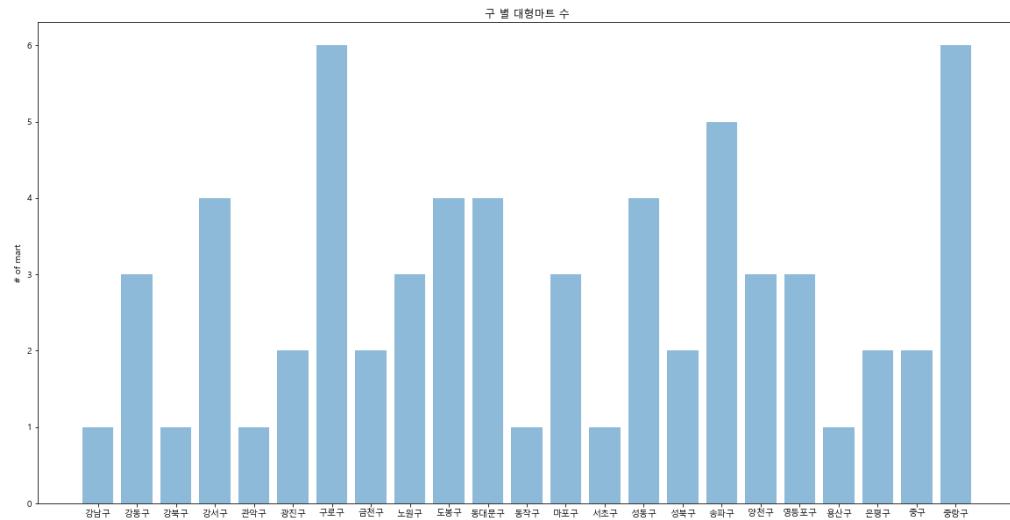
2.1.2. 수집데이터: 치안기관



- 2016년 3월 3일에 제공된 지구대, 파출소, 경찰서의 주소와 관련 정보
- ‘서울 열린 데이터광장’에서 제공하는 공개데이터로써 종로구에 각종 치안기관이 다른 구에 비해 많은 것을 확인
- 경찰서 등의 치안기관의 존재 유무는 일부 주민에게 상권의 형성과 치안강화의 효과 등 긍정적인 영향과 범죄자들이 드나드는 시설로 혐오시설이라는 부정적인 영향 모두 가진 것으로 파악
- Data shape: (270, 5)
http://news.heraldcorp.com/view.php?ud=20131014000847&md=20131017004549_BL



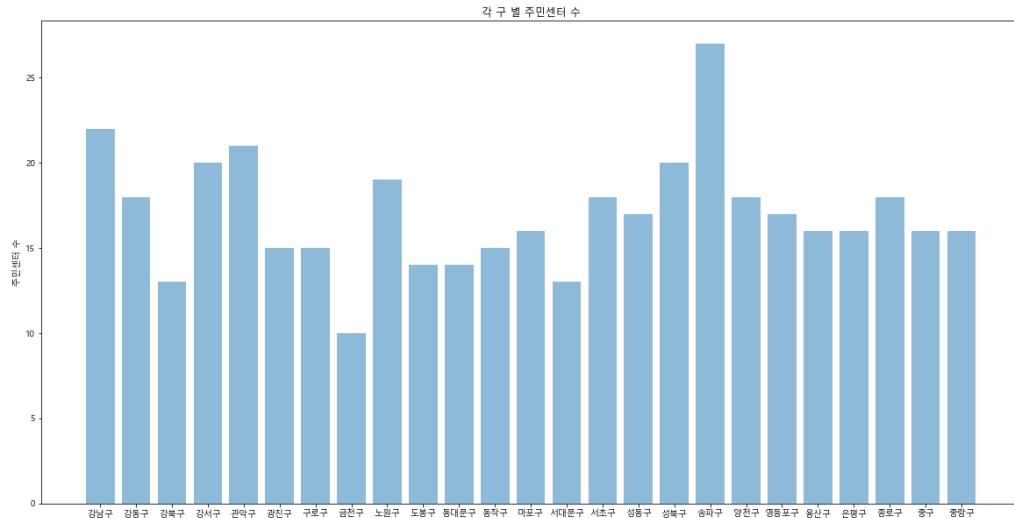
2.1.2. 수집데이터: 대형병원



- 2018년 11월 12일 기준 대형병원 현황 데이터
- 병원 접근성이 집값에 영향.. 5~10분 거리 단지 인기
- Data shape: (57, 5)



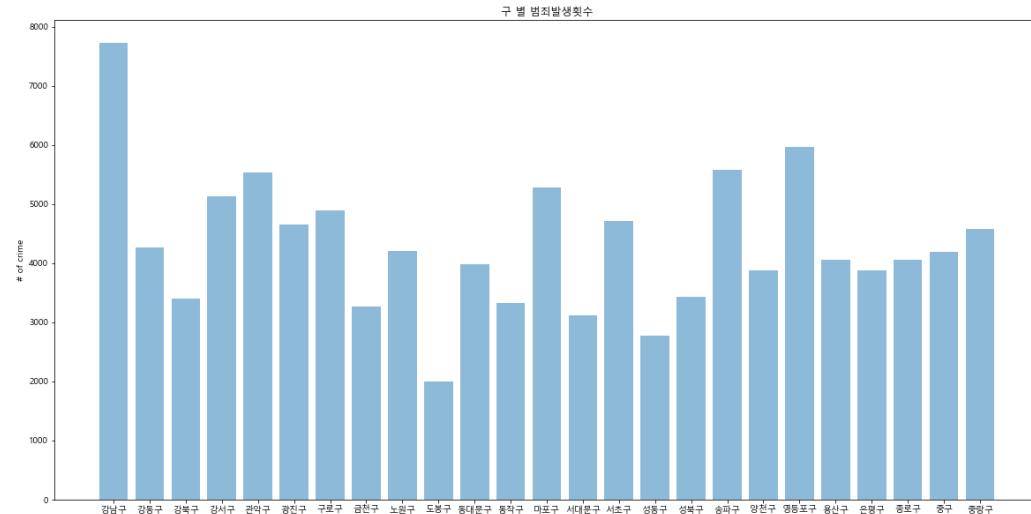
2.1.2. 수집데이터: 주민센터



- 2016년 6월 서울시 주민센터 정보
- 송파구가 27개의 가장 많은 주민센터를 보유하고 있으며, 이는 송파구 총 27개의 행정동 개수와 같음
- 주민센터는 대표적인 생활편의 시설의 일부로서, 집값에 영향을 미침
- Data shape: (512, 14)



2.1.2. 수집데이터: 범죄율

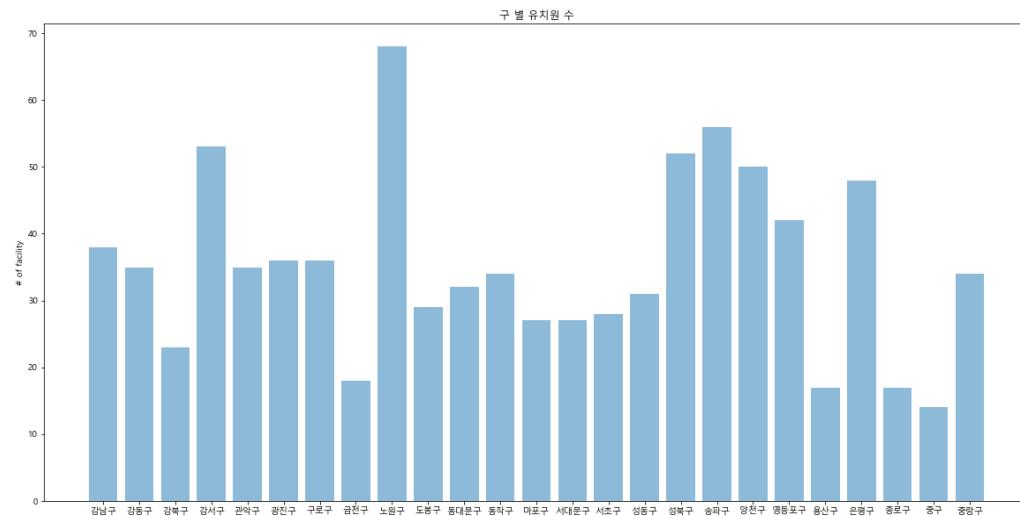


- 2012~2018년도까지 집계된 각 구별 5대 범죄 누적 발생횟수 정보
- 살인, 강도, 강제추행, 절도, 폭력 5대 범죄의 발생횟수를 통해 각 구의 상대적 범죄율을 도출하여 거주지역의 상대적 안전함의 정도로 해석
- 2015년 기준 살인, 강도, 강간 3대 흉악범죄 발생건수와 아파트의 가격의 상관 관계를 분석한 연구에 의하면 범죄율이 낮은 지역의 집값이 낮을 것이라는 상식과는 다르게 양의 상관관계가 있음을 확인
- Data shape: (26, 14)

<https://m.post.naver.com/viewer/postView.nhn?volumeNo=9420125&memberNo=34429994&vType=VERTICAL>



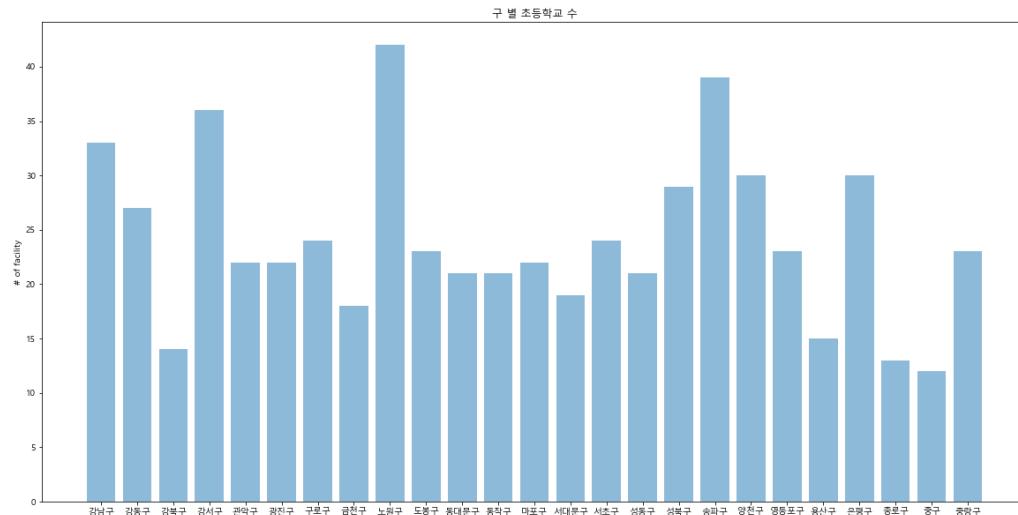
2.1.2. 수집데이터: 유치원



- 2017년 4월 시점의 서울시 구별 유치원 정보
- 서울시 유치원 전체를 소속된 구로 group by한 뒤 분포 확인
- “학부모의 62%가 집 선택 시 유치원과의 거리를 최우선으로 고려함”
- Data shape: (885, 14)



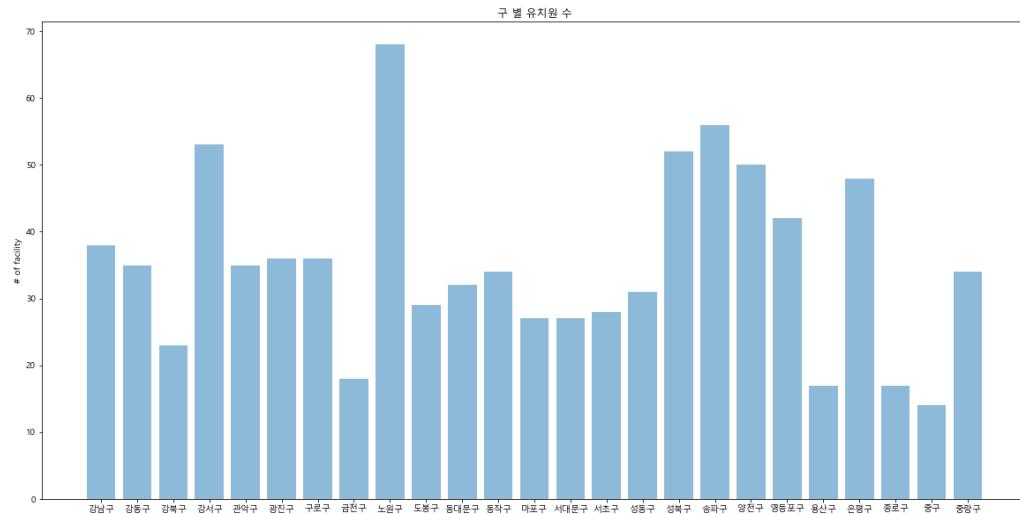
2.1.2. 수집데이터: 초등학교



- 2017년 4월 시점의 서울시 구별 초등학교 정보
- 서울시 초등학교 전체를 소속된 구로 group by한 뒤 분포 확인
- “초등학교와 인접한 집들은 몸 값도 높게 형성되는 것으로 나타났다.”
- Data shape: (602, 42)



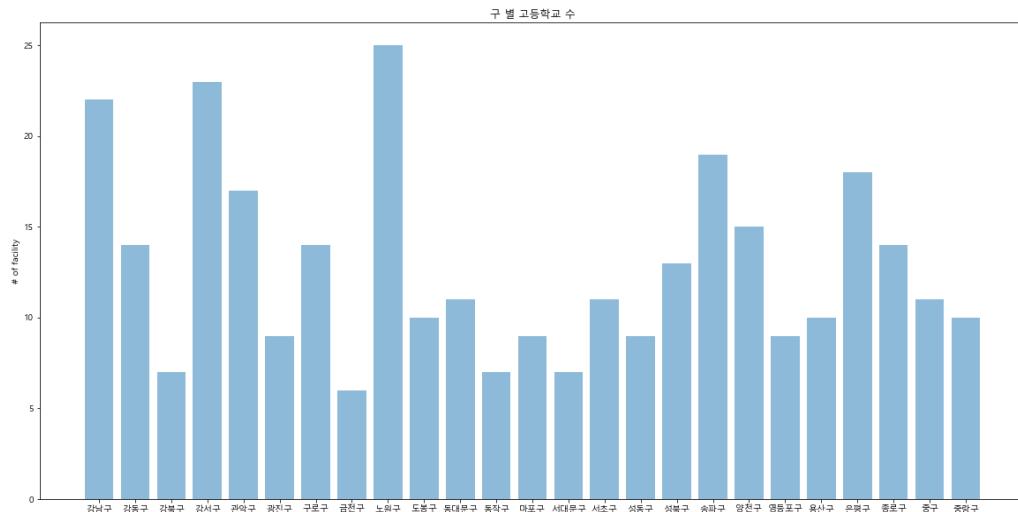
2.1.2. 수집데이터: 중학교



- 2017년 4월 시점의 서울시 구별 중학교 정보
- 서울시 중학교 전체를 소속된 구로 group by한 뒤 분포 확인
- “2019학년도부터 자율형 사립고, 외국어고, 국제고, 일반고의 신입생을 동시에 뽑기로 하면서 명문 학군 인근 단지들이 주목을 받고 있다.”
- Data shape: (384, 24)



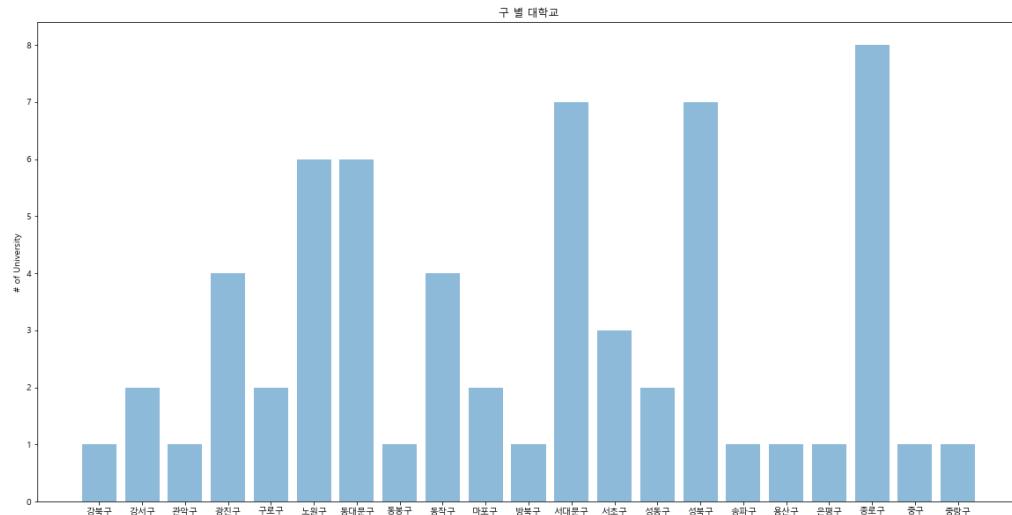
2.1.2. 수집데이터: 고등학교



- 2017년 4월 시점의 서울시 구별 고등학교 정보
- 서울시 고등학교 전체를 소속된 구로 group by한 뒤 분포 확인
- 유치원, 초등학교, 중학교와 같은 맥락으로 현대인 가정에게 자녀의 학교거리는 집 값에 유의한 영향을 주고 있음
- Data shape: (317, 29)



2.1.2. 수집데이터: 대학교



- 2019년 현재 위치상 서울 각 행정구에 속한 대학교 정보
- 서울소재 대학교는 총 62개이며 구별로 상이한 분포를 하고 있으며 종로구에 8개로 가장 많은 대학교가 위치함을 확인
- 대학교의 존재유무는 주변 상권에 강력한 영향을 끼치는 것으로 부동산의 가치에 영향을 줌
- Data shape: (62, 5)



2.1.2. 수집데이터: 소상공인

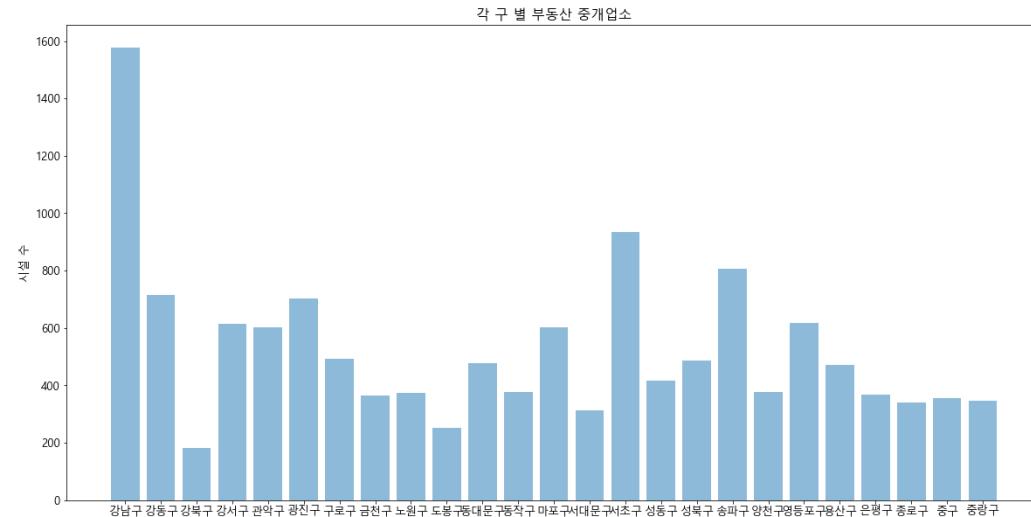
- 소상공인 데이터 약 34만 건은 대분류, 중분류, 세분류로 구분됨
- 본 프로젝트에서는 건물의 가격에 영향을 줄 수 있는 상권을 대분류 또는 중분류 기준으로 나누어 분석에 사용함

대분류	중분류
부동산	-
관광/여가	PC방, 오락, 당구, 무도, 유흥
	안마시술소
	연극, 영화
숙박	모텔
	호텔, 콘도
학문/교육	-

대분류	중분류
의료	수의
	약국
	일반병원
	한의원
생활서비스	-
소매	-
음식	-



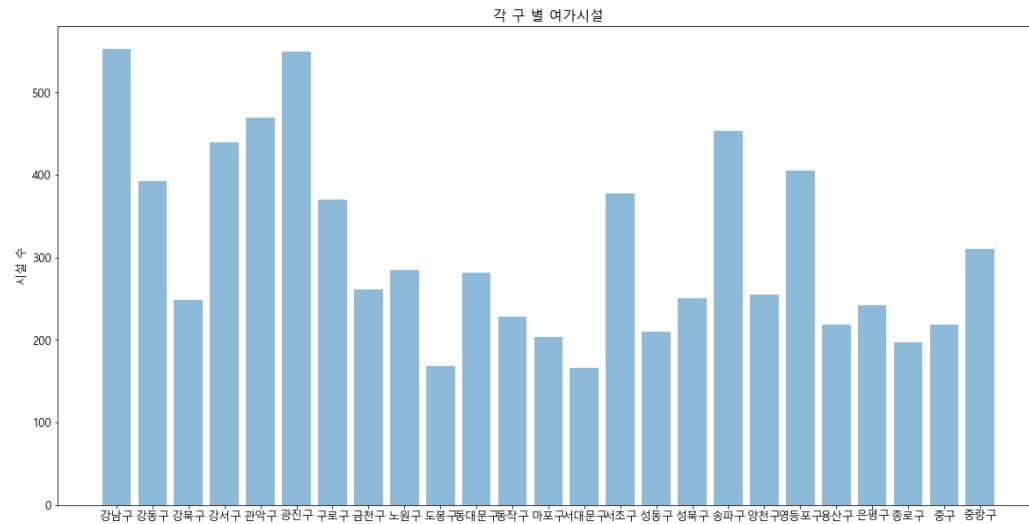
2.1.2. 수집데이터: 부동산



- 2019년 현재 위치상 서울 각 행정구에 속한 대학교 정보
- 서울소재 대학교는 총 62개이며 구별로 상이한 분포를 하고 있으며 종로구에 8개로 가장 많은 대학교가 위치함을 확인
- 대학교의 존재유무는 주변 상권에 강력한 영향을 끼치는 것으로 부동산의 가치에 영향을 줌
- Data shape: (13164, 5)



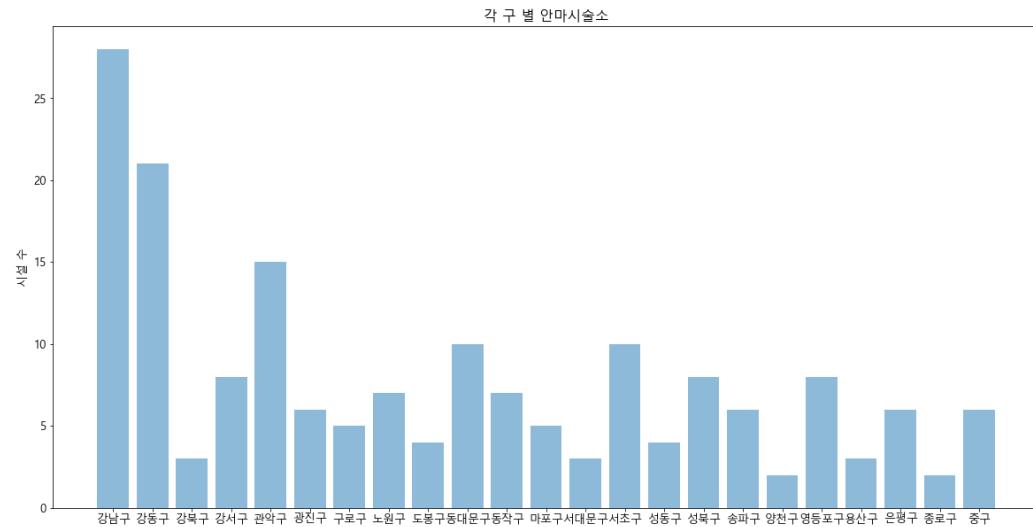
2.1.2. 수집데이터: 관광/여가 1 (PC방, 오락, 당구, 무도, 유흥)



- 2018년 12월 시점의 서울시 구별 PC방, 오락, 당구, 무도, 유흥 정보
- 서울시 해당 상가 전체를 소속된 구로 group by 한 뒤 분포 확인
- “젊은 사람들이 몰려오는 만큼 유흥시설도 들어서고 집값에도 영향이 있을 것”
- Data shape: (7755, 5)



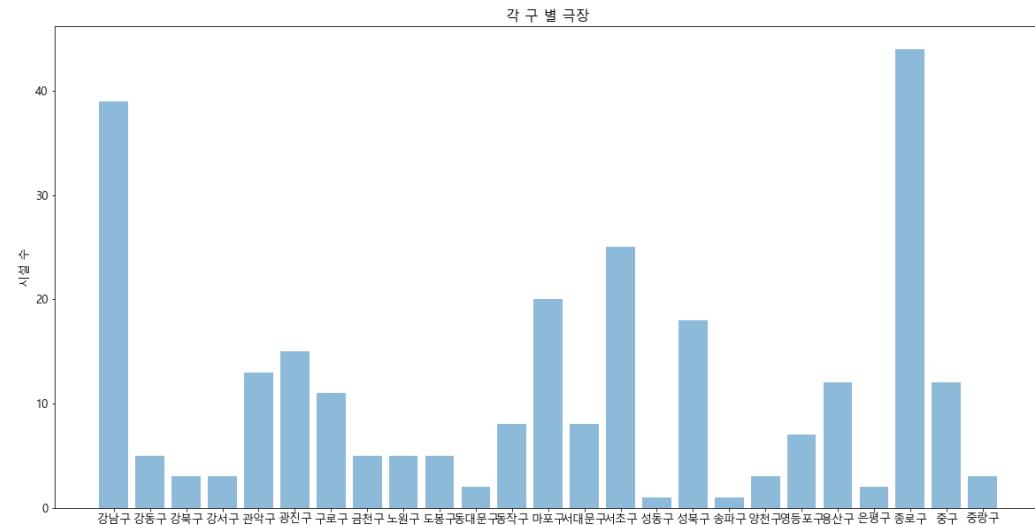
2.1.2. 수집데이터: 관광/여가 2 (안마시술소)



- 2018년 12월 시점의 서울시 구별 안마시술소 정보
- 서울시 안마시술소 전체를 소속된 구로 group by 한 뒤 분포 확인
- “아파트를 구하러 온 사람들이 큰길에 즐비한 안마시술소를 보면 집도 안 보고 그냥 돌아간다.”
- Data shape: (177, 5)



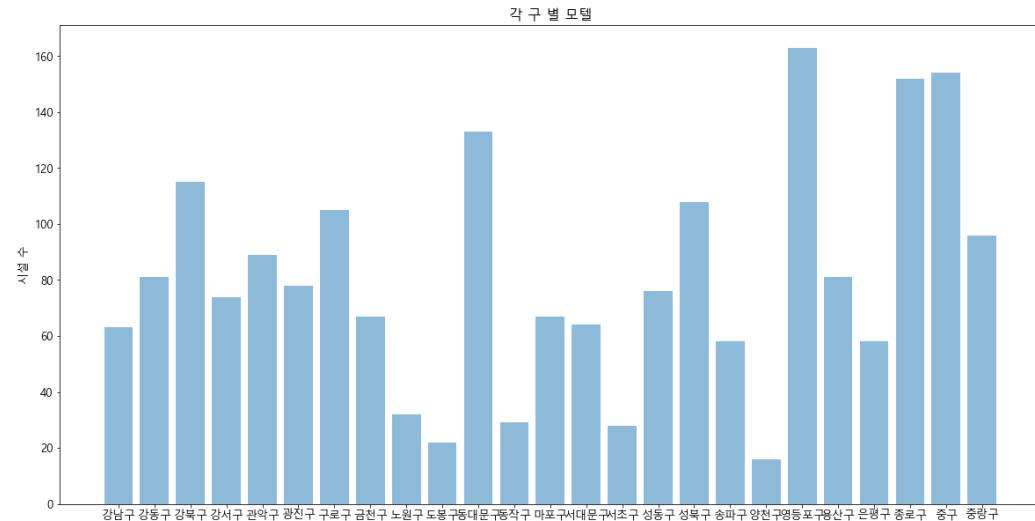
2.1.2. 수집데이터: 관광/여가 3 (연극, 영화관)



- 2018년 12월 시점의 서울시 구별 연극, 영화관 정보
- 종로구, 강남구에 다수의 연극, 영화관 위치
- “신도시나 새롭게 조성되는 택지지구 등에서 귀한 대접을 받는 영화관이지만, 강남 주택가에서는 찬밥신세를 면치 못하고 있다.”
- Data shape: (270, 5)



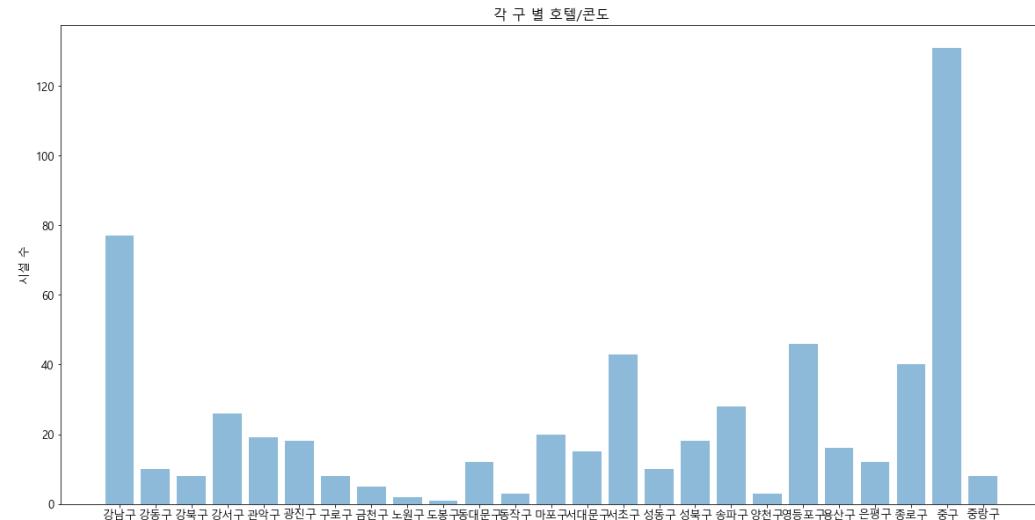
2.1.2. 수집데이터: 숙박 1 (모텔)



- 2018년 12월 시점의 서울시 구별 연극, 영화관 정보
- 서울시 모텔 전체를 소속된 구로 group by 한 뒤 분포 확인
- 호텔과 규모, 서비스 등 여러 측면에서 차이가 많을 것으로 상식적으로 판단하여 분류함
- Data shape: (2009, 5)



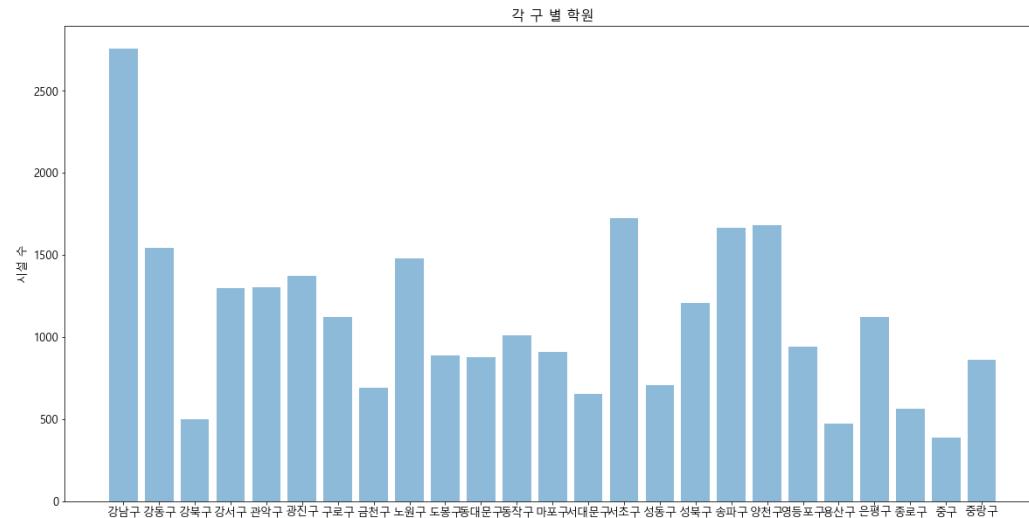
2.1.2. 수집데이터: 숙박 2 (호텔, 콘도)



- 2018년 12월 시점의 서울시 구별 호텔 정보
- 중구, 강남구에 특히 많은 호텔이 위치함
- “기존 거주자들도 강남을 떠나려 하지 않고 있다. (중략) JW메리어트호텔 등 생활편의시설이 잘 돼 있다.”
- Data shape: (579, 5)



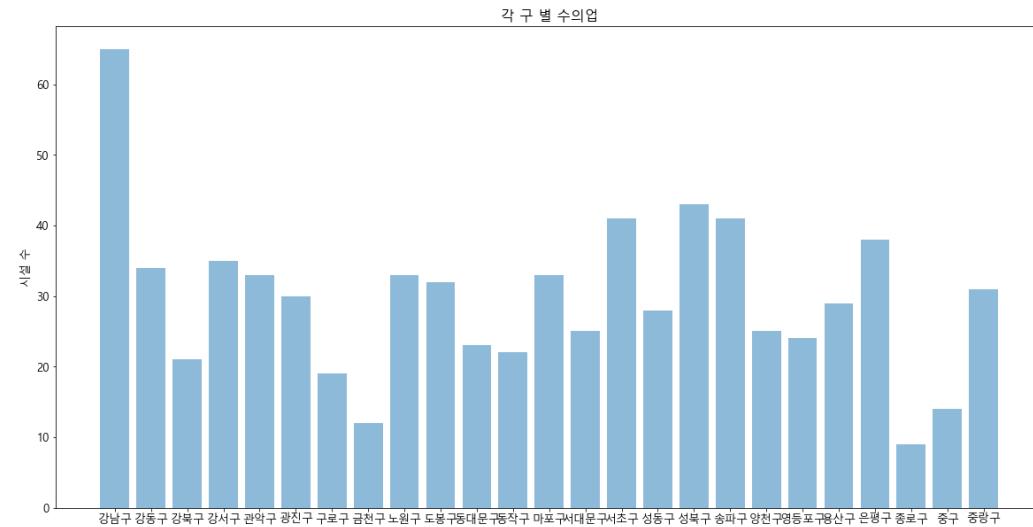
2.1.2. 수집데이터: 학문/교육



- 2018년 12월의 서울시 구별 학문/교육(학원, 도서관, 유아교육 등) 정보
- 서울시 학문/교육 시설 전체를 소속된 구로 group by 한 뒤 분포 확인
- 강남구에 특히 많은 학문/교육 시설이 집중됨
- “300여개 학원들이 집중되면서 지난해 대치동 아파트 값은 폭등했습니다.”
- Data shape: (13858, 5)



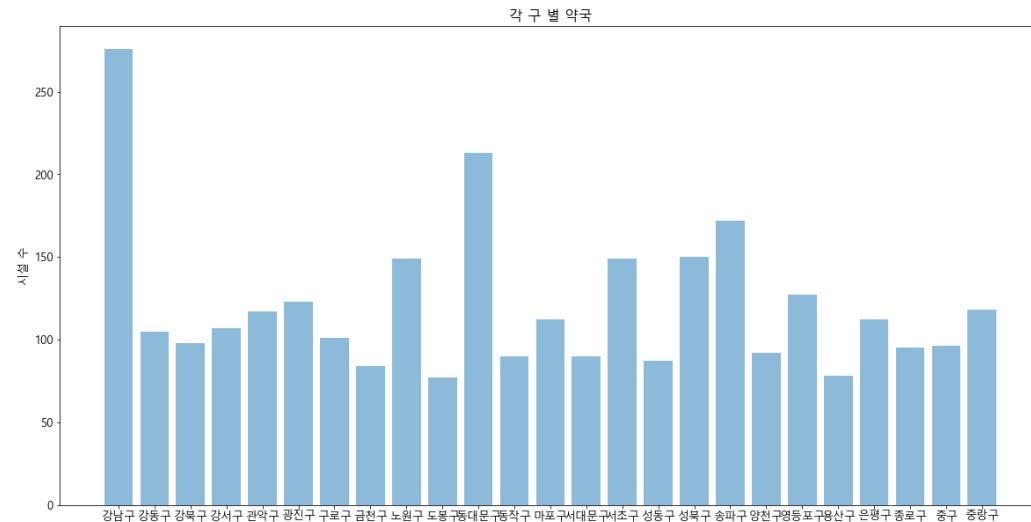
2.1.2. 수집데이터: 의료 1 (수의)



- 2018년 12월 시점의 서울시 구별 수의병원 정보
- 타 자치구보다 강남구에 다수의 동물병원 존재
- “냄새와 분변 등 환경오염과 집값 하락을 우려하는 주민들의 반대에 부딪혀 난항을 겪던 중 시의회마저 나서 반대의견을 피력한 겁니다.”
- Data shape: (740, 5)



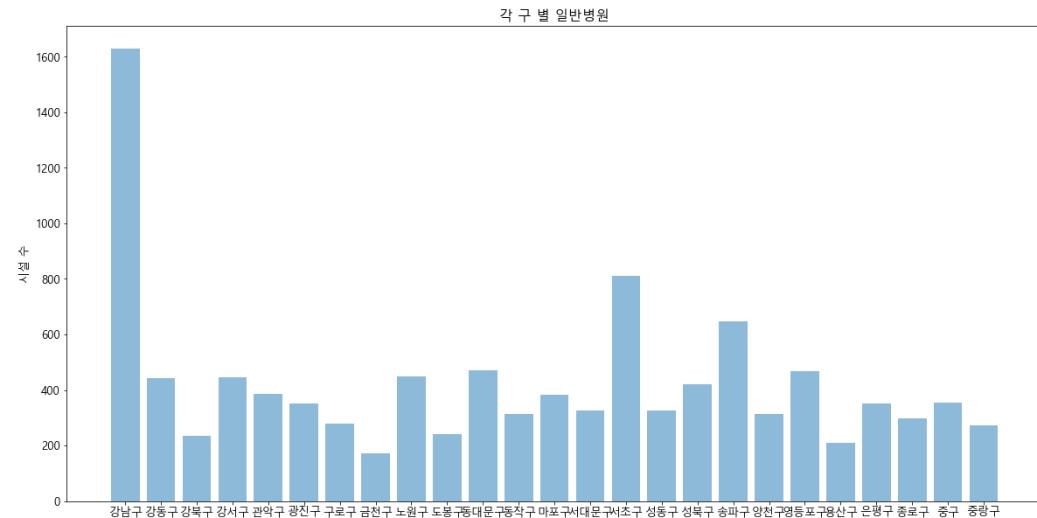
2.1.2. 수집데이터: 의료 2 (약국)



- 2018년 12월 시점의 서울시 구별 약국 정보
- 타 자치구보다 강남구에 약국이 다수 존재
- “기존 집값 끌어올리는 효과… 병원 방문객 등 겨냥 상가 투자도”
- Data shape: (3018, 5)



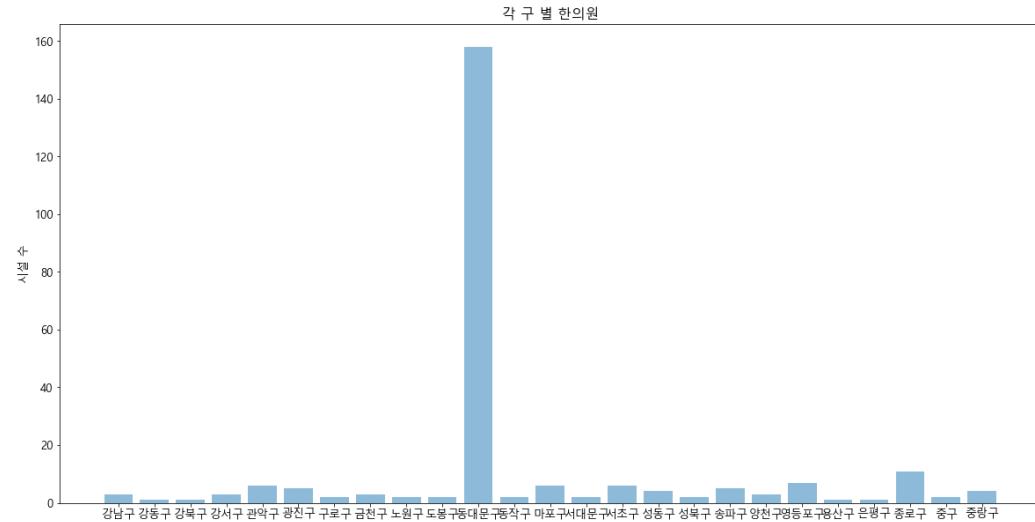
2.1.2. 수집데이터: 의료 3 (일반 병원)



- 2018년 12월 시점의 서울시 구별 일반 병원 정보
- 타 자치구보다 강남구에 다수의 일반 병원이 존재
- Data shape: (10601, 5)



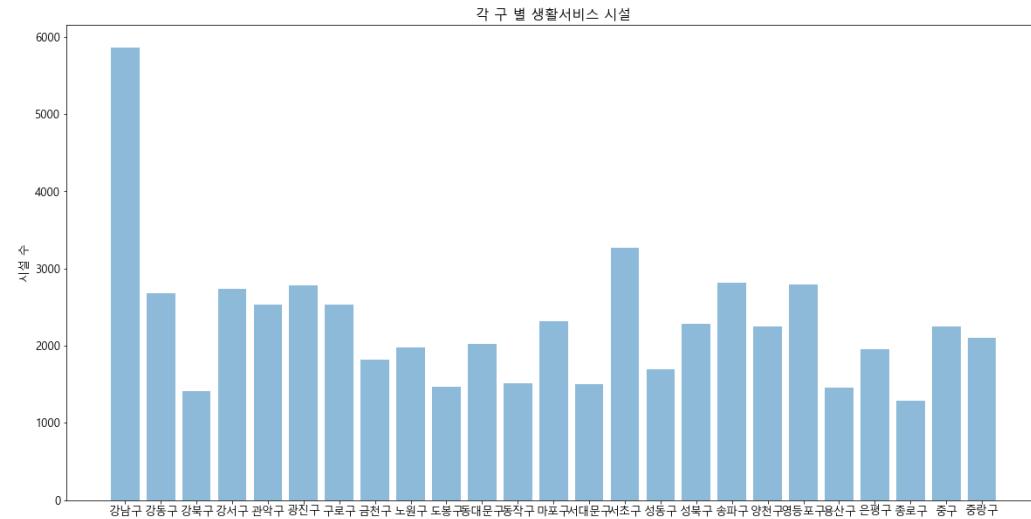
2.1.2. 수집데이터: 의료 4 (한의원)



- 2018년 12월 시점의 서울시 구별 한의원 정보
- 동대문구에 한약 시장이 존재하기 때문에 타 자치구에 비해 특히 많은 한의원 존재
- Data shape: (242, 5)



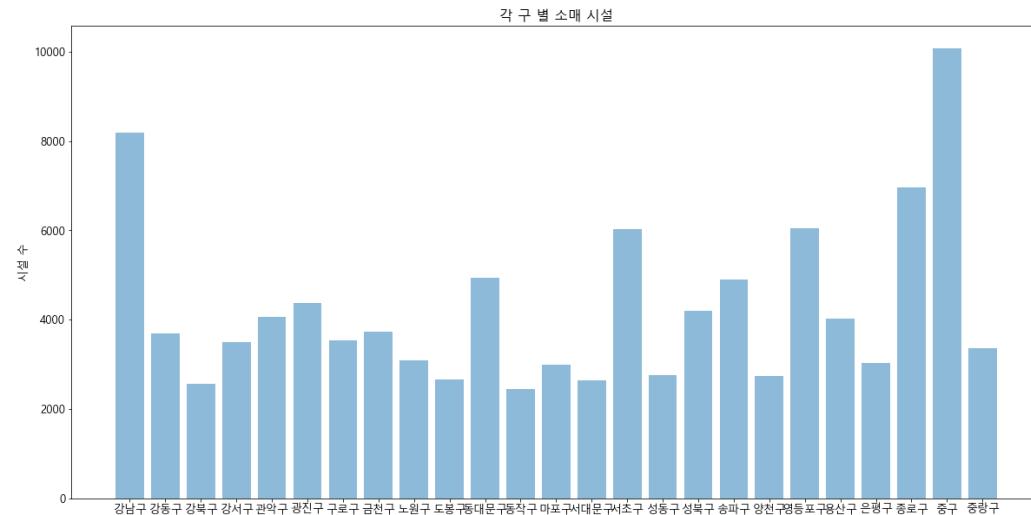
2.1.2. 수집데이터: 생활서비스



- 2018년 12월 시점의 서울시 구별 생활서비스 (세탁, 광고, 자동차 등) 정보
- 서울시 생활서비스 (세탁, 광고, 자동차 등) 전체를 소속된 구로 group by 한 뒤 분포 확인
- “생활편의시설과 더불어 (중략) 이용객이 많은 대단지를 우선적으로 지나는 경우가 많고 단지 안팎으로 대규모 상가나 문화 체육시설이 조성돼 입주민들의 자부심이 높다.”
- Data shape: (57266, 5)



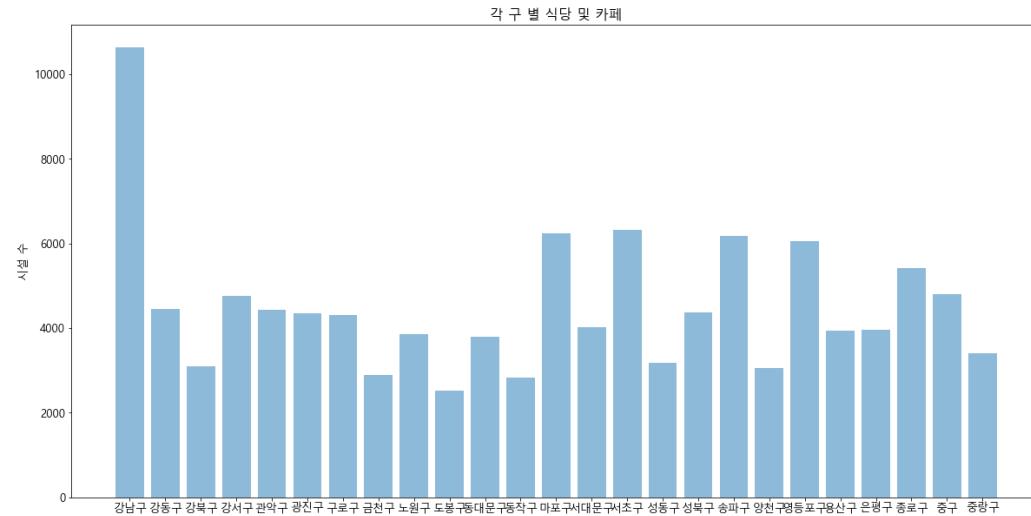
2.1.2. 수집데이터: 소매



- 2018년 12월 시점의 서울시 구별 소매 (건강, 미용, 책, 의류, 문구 등) 정보
- 서울시 소매 (건강, 미용, 책, 의류, 문구 등) 전체를 소속된 구로 group by 한 뒤 분포 확인
- “집값이 최대 20%까지 폭락할 수 있다는 예측이 비현실적이 아닐 수 있다. 이럴 경우 전반적으로 (중략) 소매업계에 큰 타격을 줄 것”
- Data shape: (106490, 5)



2.1.2. 수집데이터: 음식



- 2018년 12월 시점의 서울시 구별 음식 (식당, 카페, 베이커리 등) 정보
- 서울시 음식 (식당, 카페, 베이커리 등) 전체를 소속된 구로 group by 한 뒤 분포 확인
- “교외에 부동산을 매입하고자 하는 구매자들은 ‘라이프스타일’을 고려하기 마련입니다. 그러니 매매자들은 카페와 같이 부동산에 가치를 더해줄 수 있는 이러한 라이프스타일의 매력을 이용하는 것이죠.”
- Data shape: (112894, 5)



2.2.1. 데이터 전처리 과정





2.2.2 데이터 전처리 과정: googlemaps 라이브러리



```
In [7]: gmaps.geocode('구의역', language='ko')
Out[7]:
[{'address_components': [{'long_name': '216-22',
   'short_name': '216-22',
   'types': ['premise']},
  {'long_name': '자양2동',
   'short_name': '자양2동',
   'types': ['sublocality', 'sublocality_level_2']},
  {'long_name': '광진구',
   'short_name': '광진구',
   'types': ['locality', 'sublocality_level_1']},
  {'long_name': '서울특별시',
   'short_name': '서울특별시',
   'types': ['administrative_area_level_1', 'political']},
  {'long_name': '대한민국',
   'short_name': 'KR',
   'types': ['country', 'political']},
  {'long_name': '143-192',
   'short_name': '143-192',
   'types': ['postal_code']}],
 'formatted_address': '대한민국 서울특별시 광진구 자양2동 216-22',
 'geometry': {'location': {'lat': 37.5370464, 'lng': 127.0859404},
   'location_type': 'ROOFTOP'},
 'viewport': {'northeast': {'lat': 37.5383953802915,
   'lng': 127.0872893802915},
   'southwest': {'lat': 37.53569741970851, 'lng': 127.0845914197085}},
 'place_id': 'ChiJG0ullLiClfDURjhyiug0QwhM',
 'plus_code': {'compound_code': 'G3PP+R9 대한민국 서울특별시',
   'global_code': '8Q99G3PP+R9'},
 'types': ['establishment', 'point_of_interest']]}
```

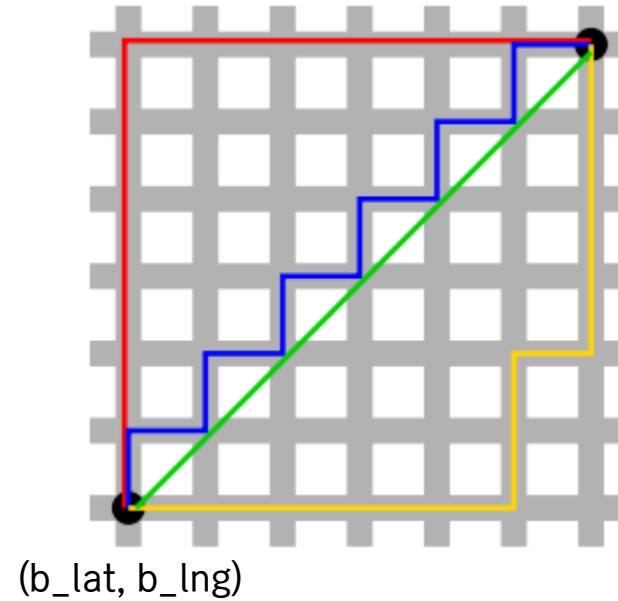


2.2.2 데이터 전처리 과정: haversine 리이브러리

haversine 2.1.1

```
pip install haversine
```

(b_lat, f_lng) (f_lat, f_lng)



f_lat: 시설의 위도

f_lng: 시설의 경도

b_lat: 건물의 위도

b_lng: 건물의 경도

```
# 단위의 manhattan dist
```

```
dist = (haversine((f_lat,f_lng), (b_lat,f_lng)) + haversine((b_lat, f_lng), (b_lat, b_lng))) * 1000
```

대도시 Manhattan에서 비롯된 거리 측정 방법이기 때문에,
도시라는 특성을 가진 서울에서도 Manhattan distance의 활용이 가능할 것으로 판단



2. 2. 3. 데이터 전처리 결과

데이터 전처리 결과	Feature
건물의 특성	평균 거래금액, 건설연도, 평균 전용면적, 평균 대지권면적,
건물에서 역까지의 거리	서울에 존재하는 모든 지하철 역
One-Hot Matrix	자치구, 지하철 노선
최근접 시설 3개와의 거리	유치원, 초등학교, 중학교, 고등학교, 대학교, 주민센터, 치안기관, 영화관, 백화점, 대형마트, 공원, 병원
해당 자치구 소속의 시설과의 거리	구청, 보건소
해당 자치구의 특성	구 범죄율
일정 거리 내에 존재하는 시설 수	소상공인 (관광, 학문, 의료, 생활서비스, 소매, 음식)

- 연립주택(건물) 데이터는 주소를 기준으로 group by 하였으며, 이를 기반으로 건물과 시설 사이의 거리를 산출함
- 결과적으로, (26542, 371) 형태의 데이터셋 구축 완료
 - 26542개의 건물 × 건물 가격 & 371개의 Feature



2. 2. 3. 데이터 전처리 결과

건물의 특성

- 동일한 건물에서 여러 건의 거래가 발생한 경우, 건물 주소를 기준으로 group by시킨 후 평균 수치를 산출함
- 건물번호는 건물 주소를 가나다 순으로 정렬한 후 순서대로 부여

건물번호	평균 거래금액	평균 전용면적	평균 대지권면적	건축년도
0	57250	58.05	31.44	2017
1	48000	54.61	33.8	2015
2	25000	39.93	19.86	1996
3	42500	50.28	41.77	1989
4	44266.66667	59.61	33.05	2015



2. 2. 3. 데이터 전처리 결과

지하철역

- 건물에서 서울에 존재하는 272개 역까지의 Manhattan distance
- 모든 지하철 역까지의 거리를 구함으로써 건물의 가격에 영향을 미치는 지하철 역의 영향을 충분히 반영

건물번호	4.19민주묘지	가락시장	가산디지털단지가양	가오리	강남	강남구청	강동	
0	23054.92376	8161.644546	15873.97968	27343.38762	21893.87944	4969.628104	5945.811394	14255.07758
1	23112.15453	7566.309439	15930.83603	27400.16304	21951.11915	5026.898745	6003.12072	13659.70227
2	23081.12174	7531.544988	15899.75748	27369.07459	21920.08746	4995.870833	5972.097538	13624.9329
3	23038.77916	7617.077443	15857.44495	27326.76857	21877.74415	4953.525042	5929.748642	13710.46858
4	23045.50636	7788.850554	15864.29635	27333.64681	21884.46838	4960.239018	5936.449786	13882.25503



2. 2. 3. 데이터 전처리 결과

One-Hot Matrix

- 자치구, 최근접 3개 역의 호선정보에 대한 One-Hot matrix
- 자치구 또는 역의 호선정보는 범주형 자료이기 때문에, 기계학습 적용을 위하여 One-Hot matrix로 변환
- 서울의 자치구는 25개이며, 호선은 1~9호선, 경인선, 경춘선, 분당선, 공항철도 등을 포함하여 총 18개

건물번호	강서구	양천구	강남구	송파구
0	0	0	1	0
1	0	0	1	0
2	0	0	1	0
3	0	0	1	0
4	0	0	1	0

건물번호	1호선	2호선	3호선	4호선
0	0	0	1	0
1	0	0	1	0
2	0	0	1	0
3	0	0	1	0
4	0	0	1	0



2. 2. 3. 데이터 전처리 결과

최근접 시설 3개와의 거리

- 시설의 유형 별로, 건물에서 Manhattan distance가 가장 가까운 3개 시설과의 거리
- 3개 시설과의 거리를 구함으로써, 다양한 시설에 대한 접근성을 고려할 수 있음

건물번호	유치원_1st	유치원_2nd	유치원_3rd
0	281.6093204	670.2823727	1601.775626
1	727.6128504	876.747589	1006.595653
2	696.5922553	911.487968	971.8501743
3	654.2416613	825.9713102	1057.37018
4	654.2634976	660.935788	1229.091827

건물번호	영화관_1st	영화관_2nd	영화관_3rd
0	4896.582542	5079.27352	5295.547595
1	4301.417369	5136.540951	5352.816374
2	4266.6737	5105.512647	5321.788235
3	4352.192517	5063.167114	5279.442593
4	4523.909255	5069.882154	5286.157187



2. 2. 3. 데이터 전처리 결과

해당 자치구 소속 시설과의 거리

- 건물과 해당 자치구 소속의 시설까지의 Manhattan distance
- 보편적으로 구청과 보건소는 해당 자치구의 기관을 이용한다는 특성을 반영

건물번호	구청	보건소
0	5376.198	5718.455
1	5433.525	5775.767
2	5402.503	5744.745
3	5360.153	5702.395
4	5366.848	5709.096



2. 2. 3. 데이터 전처리 결과

일정거리 내에 존재하는 시설의 수

- 건물에서부터 도보거리 20분 이내에 존재하는 시설의 개수
- 도보 속도를 시속 5km/h로 가정하여 Manhattan Distance 1333m 이내에 존재하는 시설의 개수 계산
- 이를 통해 건물에서부터 접근 가능한 시설의 수를 포괄적으로 고려할 수 있음

	부동산	PC, 오락, 당	안마	시술	연극	영화	숙박_모텔	호텔콘도	의료_수의	의료_약국	의료_일반	의료_한의	생활서비스	소매	음식
0	80	61	1	1	0	2	4	7	38	0	287	337	649		
1	60	41	1	1	1	1	1	4	7	34	0	215	267	471	
2	63	41	1	1	1	1	1	4	7	33	0	221	268	473	
3	62	41	1	1	1	1	1	4	7	33	0	225	278	477	
4	65	41	1	1	1	1	1	4	8	34	0	247	294	503	



2. 2. 4. Feature scaling: 건물 마마마

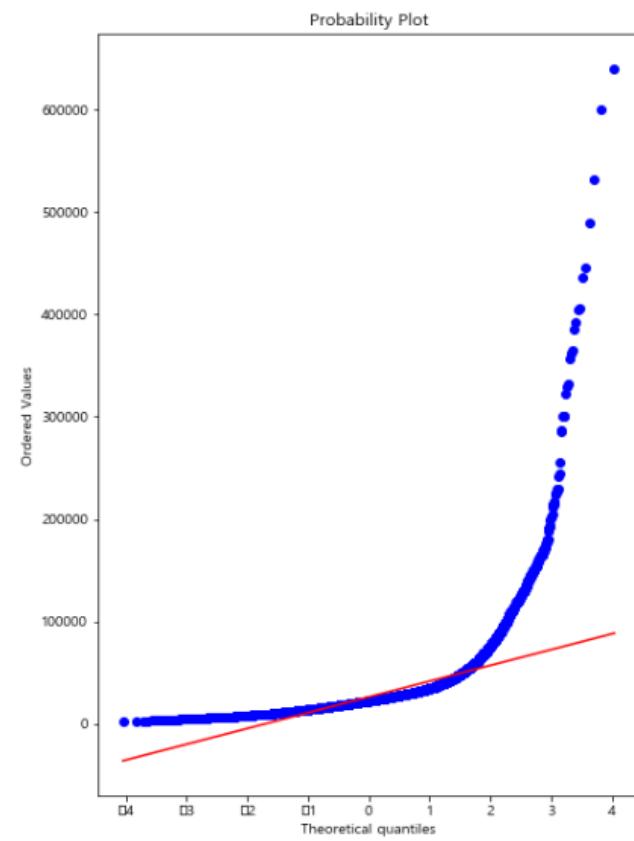
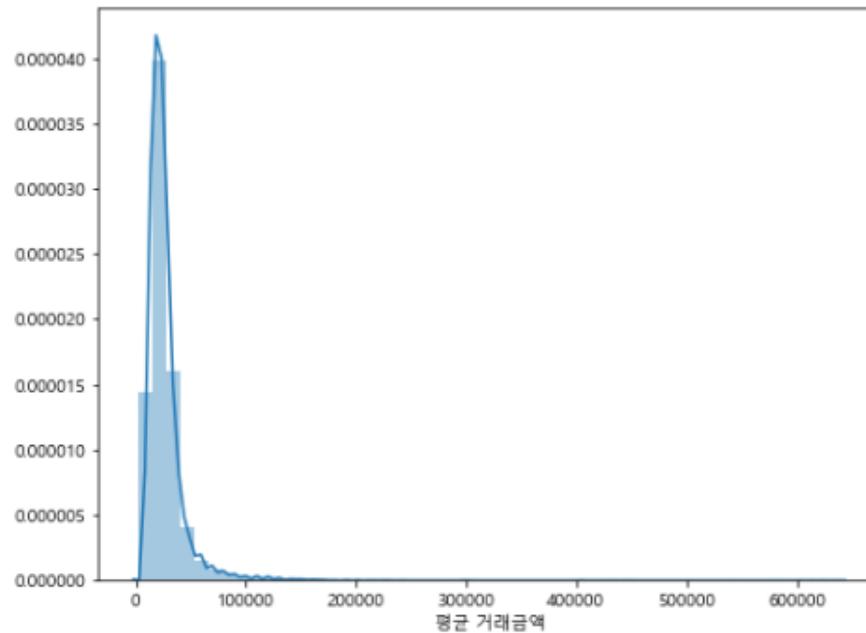
A	B
	평균 거래금액
0	57250
1	48000
2	25000
3	42500
4	44266.66667
5	39066.66667
6	22200
7	38500

In [19]: df['평균 거래금액'].describe()

Out[19]:

```
count    26542.000000
mean    26371.606822
std     20562.587945
min     2200.000000
25%    16625.000000
50%    22500.000000
75%    29800.000000
max    640000.000000
```

Name: 평균 거래금액, dtype: float64

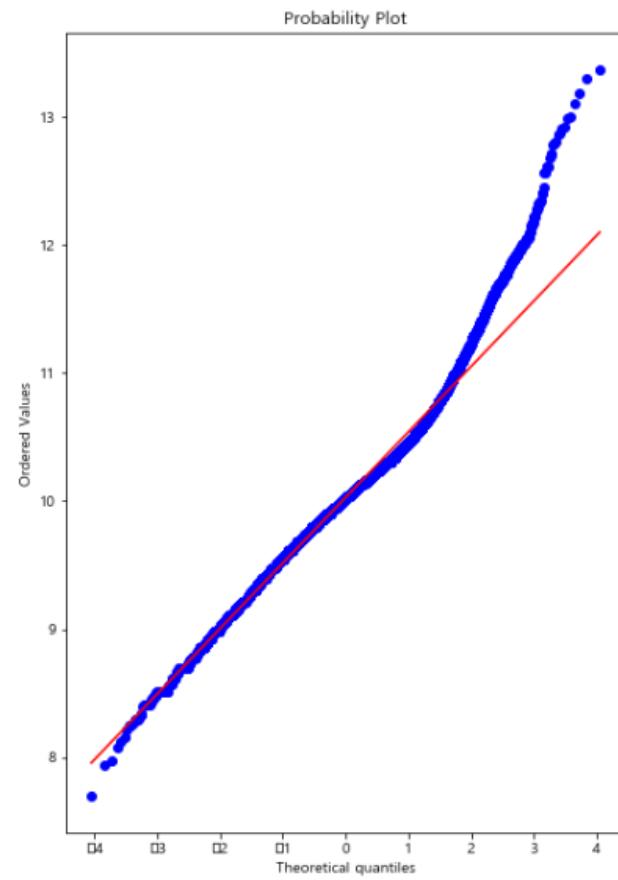
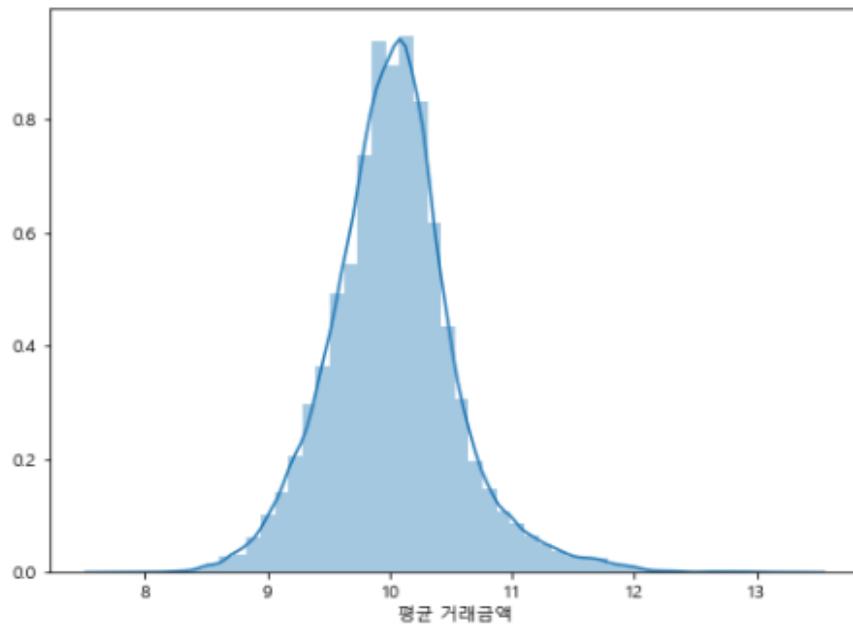




2. 2. 4. Feature scaling: 건물 매매가

A	B
	평균 거래금액
0	10.95520039
1	10.77897712
2	10.12667111
3	10.65728288
4	10.69800982
5	10.57305046
6	10.00789261
7	10.55843949

```
In [9]: df['평균 거래금액 (로그 스케일)'].describe()  
Out[9]:  
count      26542.000000  
mean       10.027472  
std        0.517869  
min        7.696667  
25%        9.718723  
50%       10.021315  
75%       10.302297  
max       13.369225  
Name: 평균 거래금액 (로그 스케일), dtype: float64
```



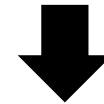


2.2.5. Data Imputation

- 최근점 시설 도출 시, 인접한 구에서만 시설을 찾았기 때문에 특정 최근점 시설이 없는 건물이 존재함

대학교_1st	대학교_2nd	대학교_3rd
7736.718712	0	0
7809.717833	0	0
7883.018659	0	0
7940.126161	0	0
7892.278442	0	0

백화점_1st	백화점_2nd	백화점_3rd
1516.721113	5222.277898	0
1831.060419	5416.155841	0
1844.702595	5434.691203	0
1870.918727	5475.475687	0
1887.844288	5427.966874	0



- 다양한 Missing value 처리 방법 중, 해당 열의 평균값을 투입하는 방법을 통하여 Missing value 처리

```
In [13]: len(df.loc[df['대학교_2nd']==0])
Out[13]: 0
```

```
In [14]: len(df.loc[df['대학교_3rd']==0])
Out[14]: 0
```

```
In [15]: len(df.loc[df['백화점_3rd']==0])
Out[15]: 0
```

3. Data Analysis



3. 0. 1. 선형회귀

1) Multiple Linear Regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

- 가장 적절한 직선을 이용해, 종속변수와 하나 이상의 독립변수 사이의 관계를 찾는 방법
- 실제값과 예측값의 차이 제곱의 합(MSE)를 최소로 하는 모델 생성
- 일반적으로 예측된 변수의 분산 비율의 비(R-square)로 모델의 성능을 측정함
- 일반적인 선형회귀에서는 독립변수들이 서로 독립이라는 가정사항이 있지만, 독립변수의 수가 증가할 수록 변수들간의 상관관계가 강해질 수 있음
 - 예를 들어, 본 프로젝트에서는 ‘강남구’라는 변수와 강남구에 존재하는 ‘강남역’, ‘선릉역’ 등의 지하철 역 관련 변수들에 어떠한 관계가 존재할 수 있음
- 이를 다중공선성(Multicollinearity)이라고 함
- 다중공선성이 존재하면 회귀식의 예측 정확도에 대한 안정성이 떨어지는 문제가 발생하기 때문에, 선형회귀 계수의 크기를 감소시키거나, 계수 자체를 없애는 방식을 선택해야 함

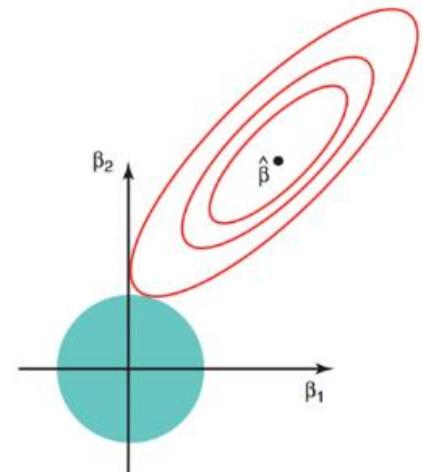


3. 0. 1. 선형회귀

2) Ridge Linear Regression

$$\hat{\beta} = \operatorname{argmin}_{\beta} |Y - X\beta|^2 + \lambda_1 |\beta|^2$$

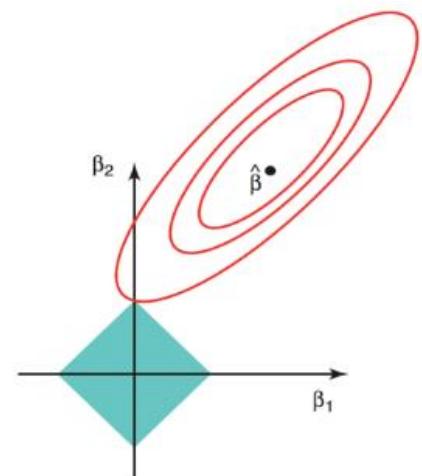
- MSE를 최소화하면서, 회귀계수 벡터의 L2 norm을 제한하는 기법
- 회귀계수의 값을 무한히 작게 하지만, 0으로는 만들 수 없음
- 변수간의 상관관계가 높아도 좋은 성능을 보임
- 크기가 큰 회귀계수를 우선적으로 줄임



2) Lasso Linear Regression

$$\hat{\beta} = \operatorname{argmin}_{\beta} |Y - X\beta|^2 + \lambda_2 |\beta|^1$$

- MSE를 최소화하면서, 회귀계수 벡터의 L1 norm을 제한하는 기법
- 회귀계수를 0으로 만들기 때문에, 변수 선택이 가능함
- 변수간의 상관관계가 높으면 성능이 떨어짐
- 비중요 변수의 회귀계수를 우선적으로 줄임

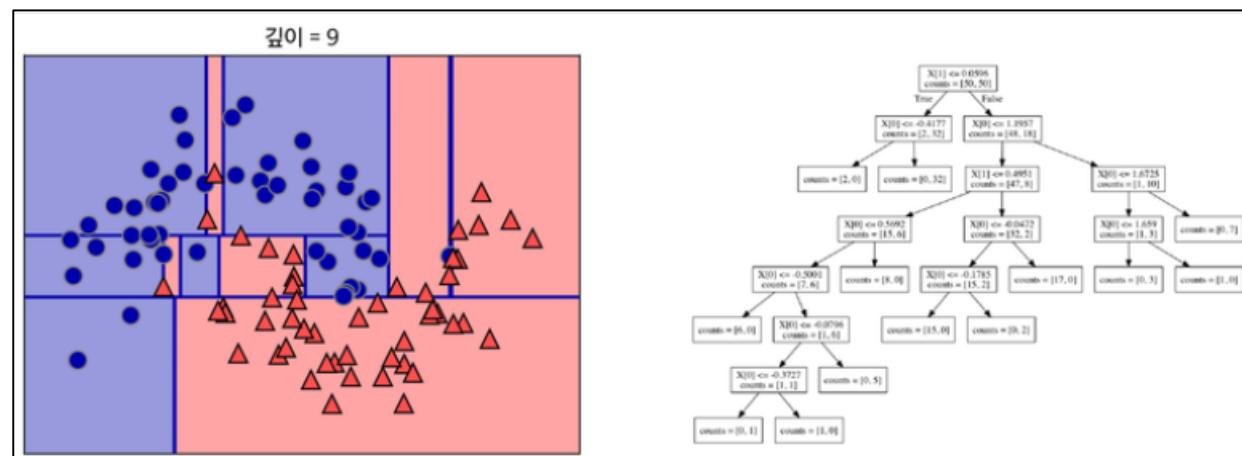




3. 0. 2. Tree 기반 회귀

1) Decision Tree

- Decision Tree는 일련의 조건에 근거하여 데이터를 subset으로 나누며 학습하는 머신러닝 알고리즘
- 데이터를 구분한 후 각 영역의 순도(homogeneity)가 커지도록, 불순도(impurity) 또는 불확실성(uncertainty)가 감소하도록 학습을 진행
- Classification 문제에서는 비슷한 feature를 가진 최빈 데이터 집합을 정답으로,
- Regression 문제에서는 탐색한 leaf node의 평균값을 정답으로 지정함

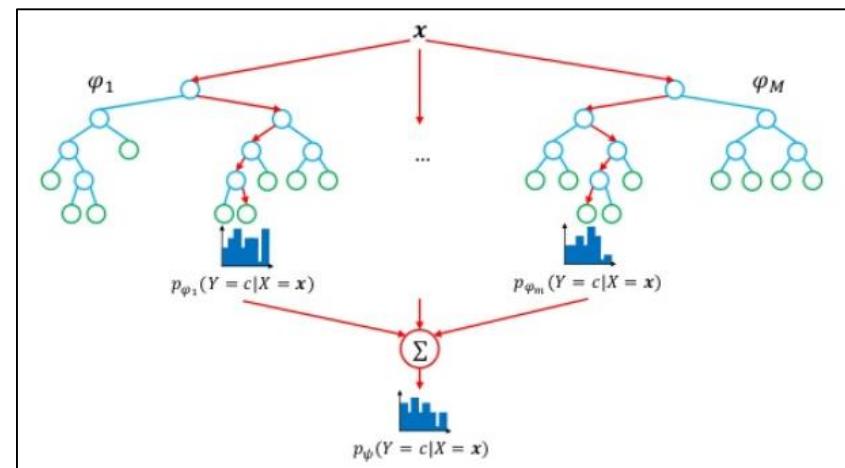




3. 0. 2. Tree 기반 회귀

2) Random Forest

- Random Forest Regressor는 각 node마다 feature를 랜덤하게 추출하여 sub tree를 만들고, 이 중에서 최선의 결과값을 찾는 머신러닝 알고리즘
- 서로 다른 feature로 overfitting된 트리를 양상별 함으로써, Decision Tree의 고유 성질인 overfitting을 회피할 수 있음
- max_feature 파라미터를 통해 랜덤으로 추출한 feature의 개수를 제한함
 - max_feature값이 클수록, 각 sub tree는 서로 비슷해지며, 가장 두드러진 feature를 가진 데이터 예측에 용이
 - max_feature값이 작을수록, sub tree들이 서로 달라지며, 각 트리는 예측을 위해 깊이가 깊어짐

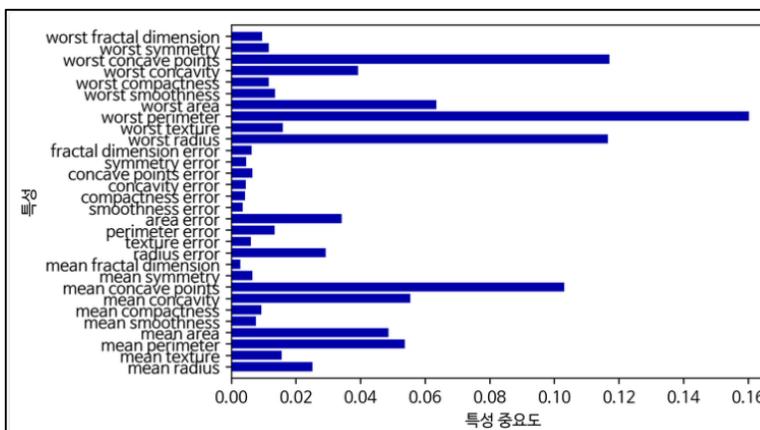




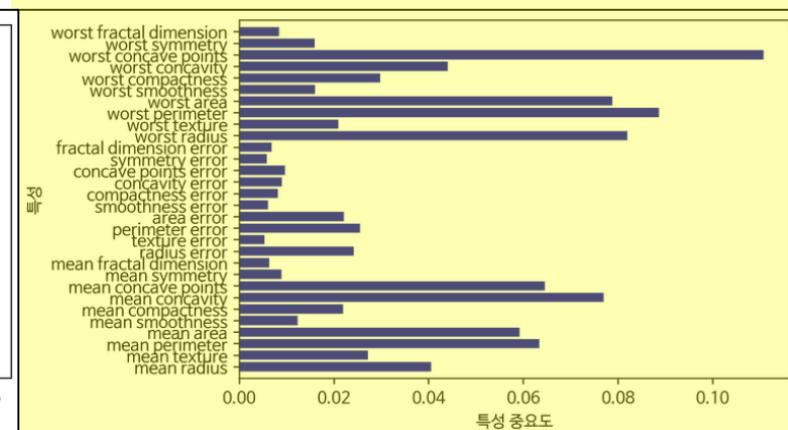
3. 0. 2. Tree 기반 회귀

3) Extra Tree Regression

- Extra Tree Regressor는 node마다 feature를 랜덤하게 분할하고 그 중에서 최상의 분할방식을 선택하여 sub tree를 만들어 최선의 결과값을 찾는 머신러닝 알고리즘
- Random Forest와 유사하지만, 극단적으로 랜덤하게 sub tree를 생성함
- Random Forest에 비해 전반적으로 feature 중요도를 더 높게 평가하는데, 이는 Extra Tree가 더 폭넓은 시각으로 feature를 평가한다는 것을 의미



〈Random Forest Regressor〉



〈Extra Tree Regressor〉



3. 0. 2. Tree 기반 회귀

4) Gradient Boosting Regression

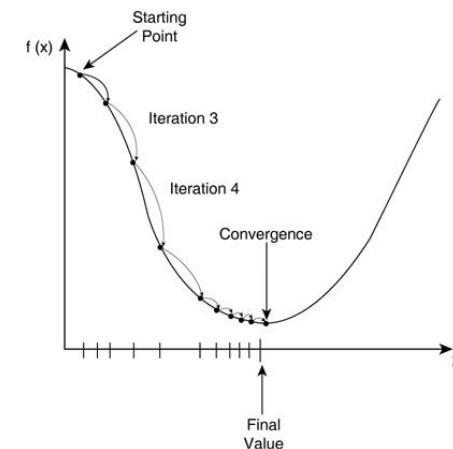
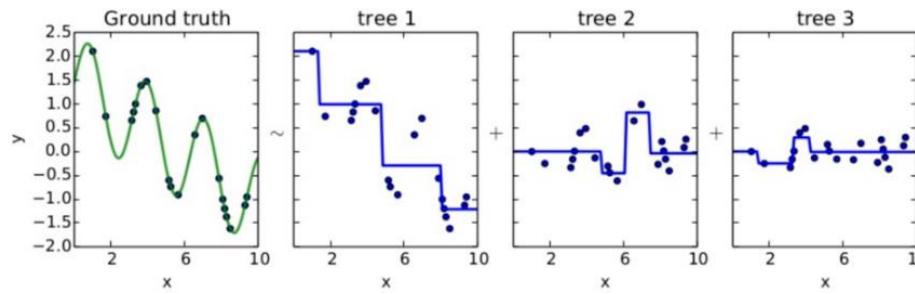
Boosting이란?

- 성능이 약한 모델들을 결합하여 강력한 모델을 만드는 과정
- 예를 들어, 모델 A, B, C의 성능이 각각 0.3정도라고 할 때, Boosting은 A 모델을 만든 후, 그 정보를 바탕으로 B 모델을 만들고, 다시 그 정보를 바탕으로 C 모델을 만드는 방법

Gradient란?

- ‘기울기’를 뜻하는 말로, MSE 등 손실함수(loss function)의 기울기를 통해 손실함수의 값을 최소화하는 기법을 의미함

따라서 Gradient Boosting Regression은 성능이 약한 모델들을 여러 개 결합하여 Gradient를 기반으로 손실함수를 최소로 하는 회귀 모델

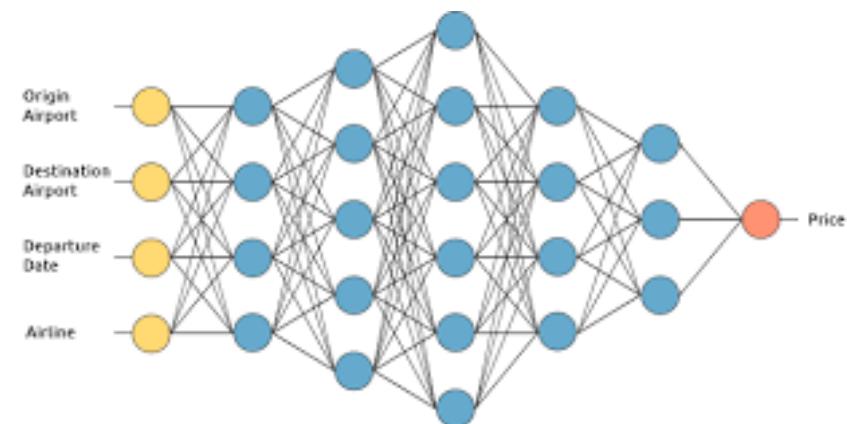
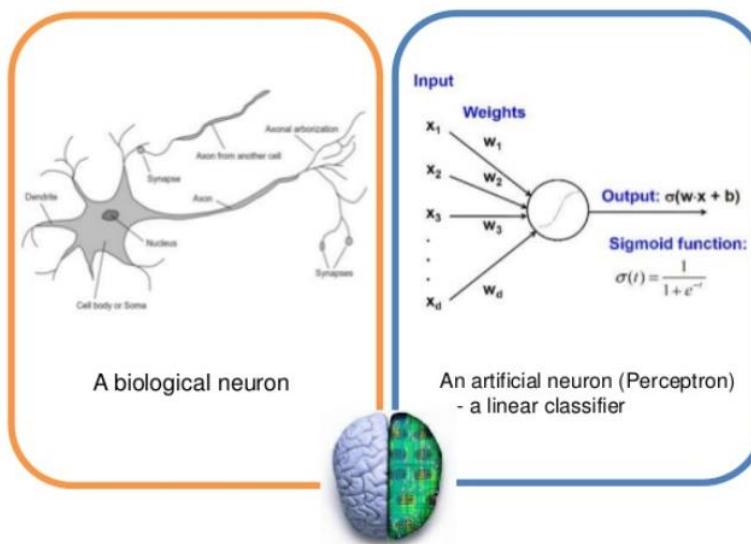




3. 0. 3. Deep Learning

1) Deep Learning

- 머신러닝의 한 부분으로 인공신경망(ANN)에 기반하여 설계된 개념
- 다수의 신호(input)을 입력받아 하나의 신호(output)을 출력하는 퍼셉트론을 다층으로 설계
- 선형 맞춤(LINEAR FITTING) 과 비선형 변환(nonlinear transformation)을 반복해 쌓아 올리는 구조
- 복잡한 공간 속에서 최적의 구분선을 만들어 내는 목적
- 모델을 구성하는 각 계층(layers)에 설계자가 목적하는 바에 적합한 활성함수와 노드의 개수를 설정하여 목적 결과를 도출 하는 것





3. 0. 4. K-fold

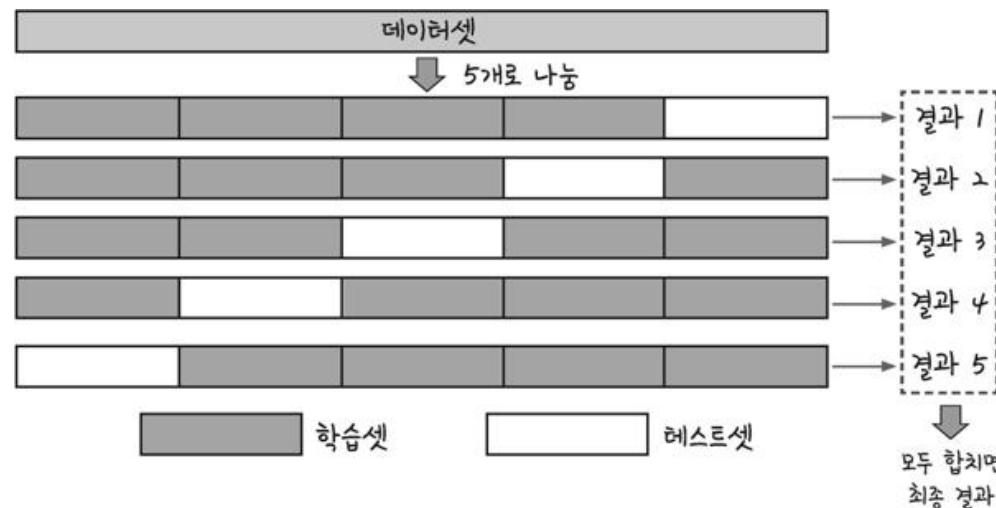
1) K-fold Cross Validation

Cross Validation이란?

- 회귀분석 모형을 만드는 목적 중 하나는 종속 변수의 값을 예측하는 것
- 학습에 쓰이지 않은 표본 데이터 집합의 종속 변수 값을 얼마나 잘 예측하는가를 검사하는 것

K-fold 란?

- 데이터의 수가 적은 경우 일부를 추출하여 검증데이터로 사용했을 경우 검증데이터의 수가 적어 검증 성능의 신뢰도가 떨어진다. 하지만 학습 데이터를 줄이면 학습이 정상적이지 않다.
- 검증데이터의 수를 증가시킬 수도, 그대로 활용할 수도 없는 딜레마를 해결하는 검증 방법
- 과적합을 막을 수 있는 방법 중 하나

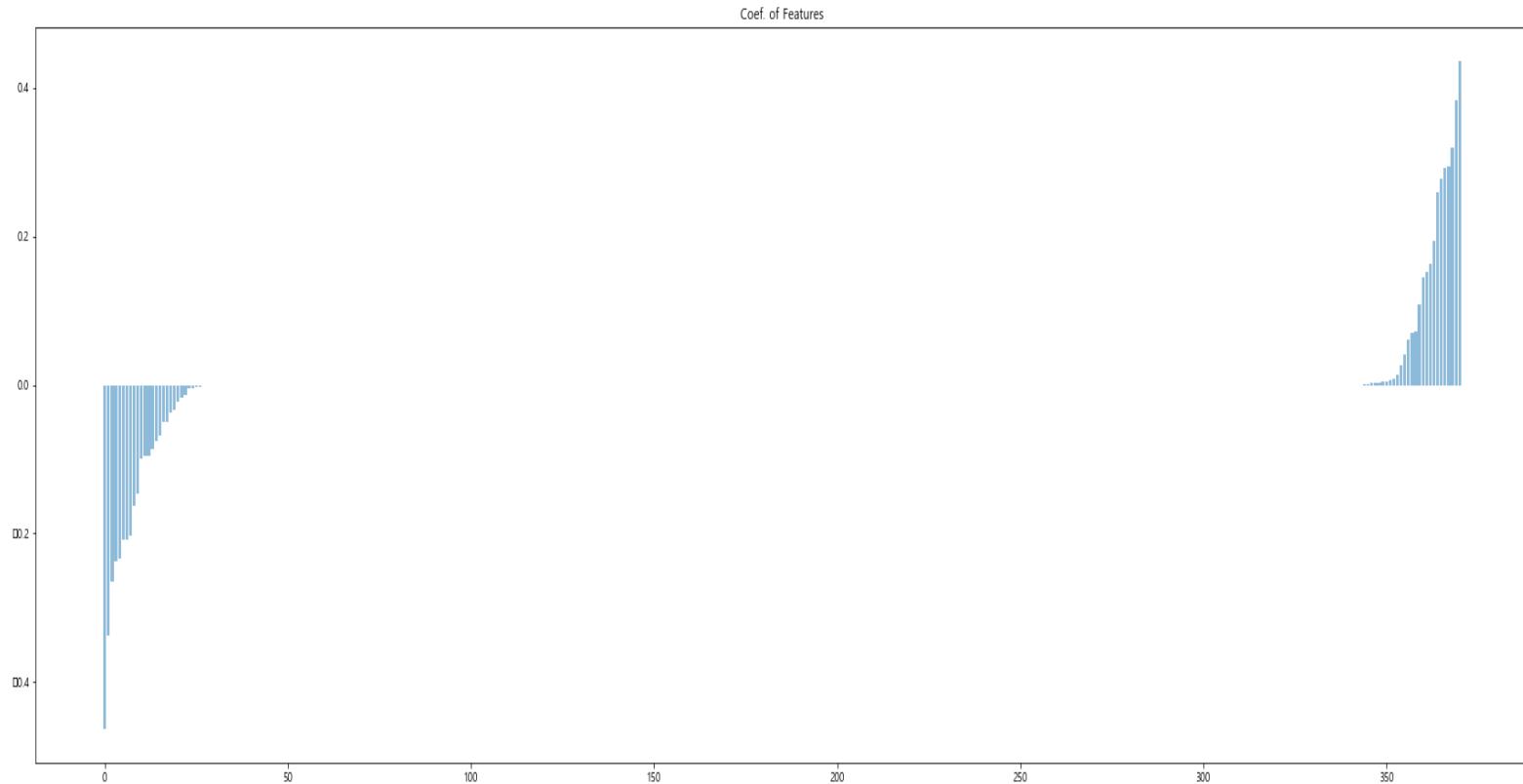




3. 1. 1. Feature 선택

Lasso Regression은 L1 norm에 정규화를 진행하여 회귀 분석

- Feature 별로 0 또는 0 이상의 회귀계수가 도출되기 때문에 feature 선택 가능
- Lasso 객체의 `coef_` 필드를 통해 feature들의 회귀계수 확인 가능

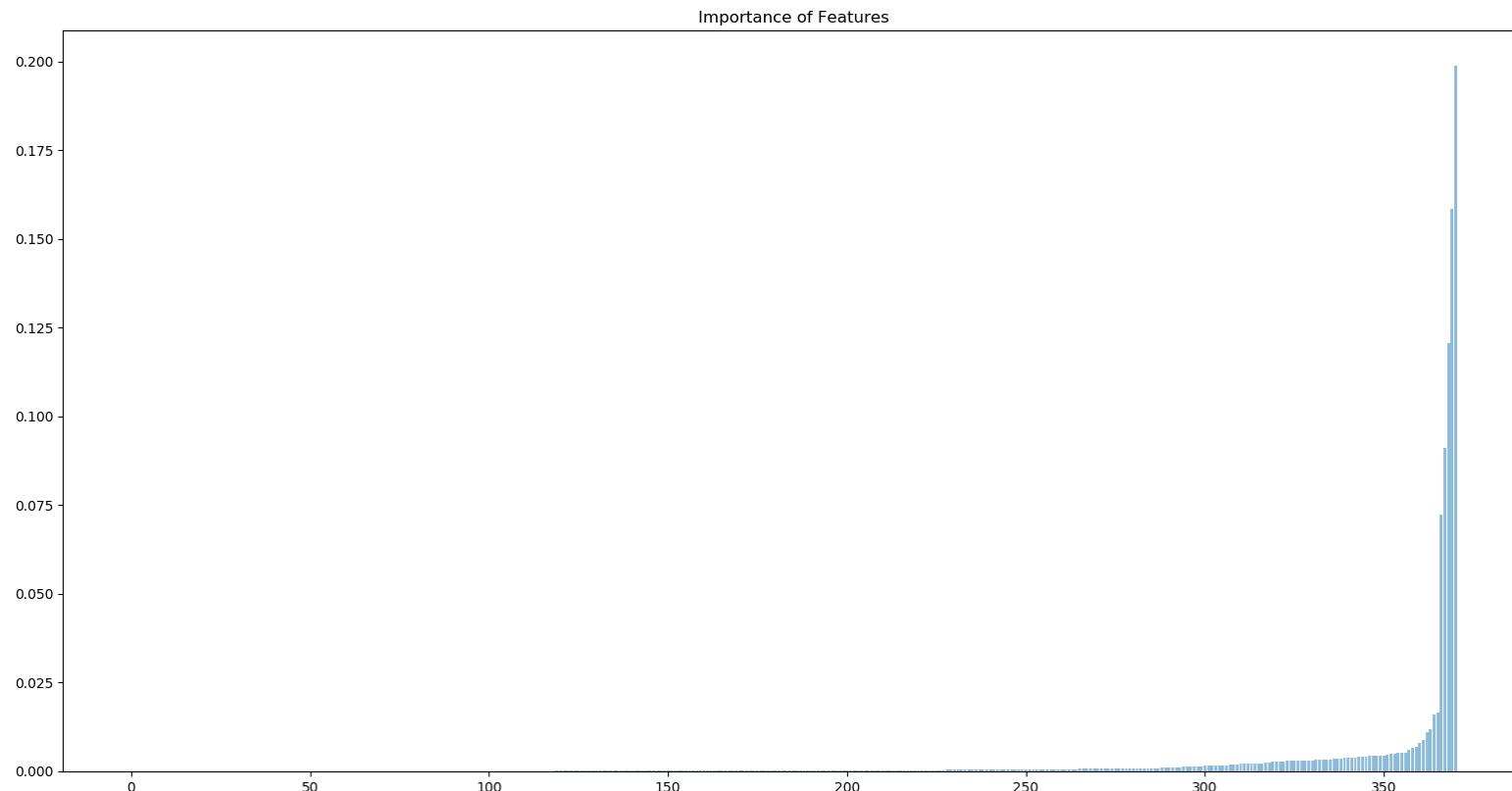




3. 1. 1. Feature 선택

Tree 기반의 모델들은 모델 생성에 있어서 feature들의 중요도를 결과로 도출해줌

- Tree 객체의 `feature_importance_` 필드를 통해 feature들의 중요도 확인 가능





3. 1. 2. Feature 선택: TOPSIS

본 프로젝트에서는 target data인 건물의 평균 매매가에 큰 영향을 미치는 주요 feature를 선정하기 위해 MCDM기법 중 하나인 TOPSIS를 사용함

- 1) 모델을 통해 feature 선택에 영향을 미치는 feature의 계수 또는 feature의 중요도 벡터 추출
 - Lasso Regression, Decision Tree, Random Forest, Extra Tree Regression, Gradient Boosting Regression 사용
- 2) min-max normalization을 통하여 모델 별 회귀계수 또는 feature 중요도 벡터 정규화
- 3) 모델 별 정규화된 벡터에 각각 모델의 MAE의 역수를 곱하여 가중치를 반영한 매트릭스 생성
- 4) 모델 별로 긍정적으로 이상적인 대안(PIS; positive ideal solution)과 부정적으로 이상적인 대안(NIS; negative ideal solution) 도출
- 5) 각각의 feature가 가지는 값과 PIS와 NIS의 거리를 기반으로 closeness 도출



3. 1. 2. TOPSIS 결과

〈우선순위 상위 20개 feature〉

순위	feature
1	평균 전용면적
2	평균 대지권면적
3	건축년도
4	신사
5	서초구
6	강동구
7	경인선
8	강남구
9	사평
10	압구정

순위	feature
11	언주
12	용산구
13	구로구
14	마포구
15	도봉구
16	은평구
17	서대문구
18	노원구
19	금천구
20	양천구



3. 1. 2. TOPSIS 결과

〈우선순위 하위 20개 feature〉

순위	feature
352	신대방
353	종로5가
354	종로3가
355	상도
356	잠실새내
357	남구로
358	금천구청
359	대모산입구
360	을지로3가
361	장승배기

순위	feature
362	장승배기
363	경찰병원
364	신설동
365	개화산
366	사가정
367	거여
368	동대입구
369	독산
370	충무로
371	강서구



3. 2 1. 예측모델 생성

선형회귀 기반의 회귀, Tree 기반의 회귀와 Deep Learning 기반의 회귀의 성능을 비교함

정확한 성능을 도출하기 위해 K-Fold 방식을 사용하였으며, 본 프로젝트에서는 K=5로 설정하여, 5개의 train, test set을 만들고 모델 생성을 반복

성능 비교는 Deep Learning 회귀 문제에서 자주 사용되는 MAE(mean absolute error)를 공통적으로 사용

- MAE는 예측 값과 실제 값의 차이의 절대값의 평균을 의미하는 수치로, MAE가 낮을수록 좋은 예측 성능을 가진 모델임을 의미함

$$\text{MAE} = \frac{\sum_{i=1}^n |y_i - x_i|}{n} = \frac{\sum_{i=1}^n |e_i|}{n}$$

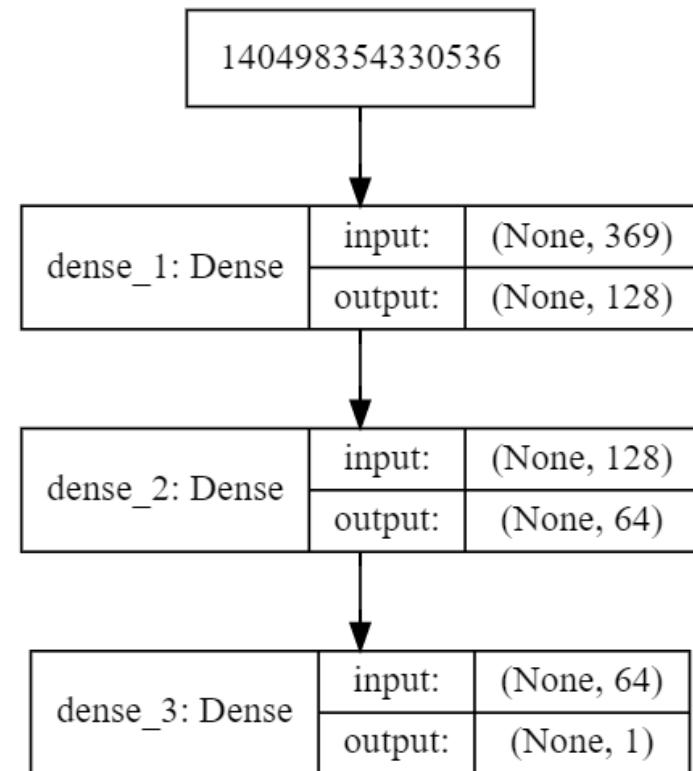
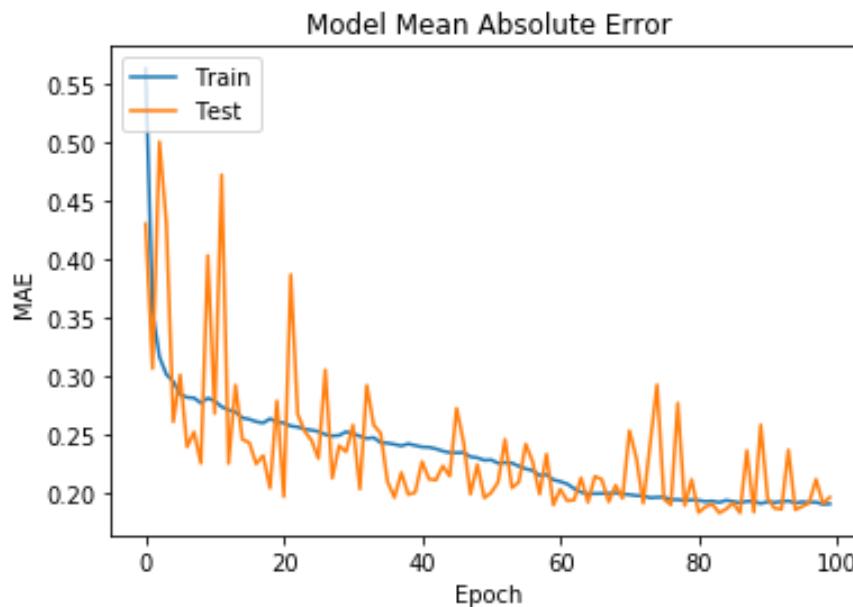


3. 2. 2. Deep Learning 예측모델

본 연구에서는 3개의 layer를 사용하여 모델을 생성함

- 첫번째 layer와 두번째 layer는 활성함수로 ReLu를 사용함
- 세번째 layer는 회귀를 위한 layer

MAE가 최소가 되는 Epoch=100을 사용





3. 2. 2. 예측모델 성능 비교

모델 명		MAE
Linear-based	Linear Regression	0.214259
	Ridge	0.214221
	Lasso	0.215788
Tree-based	Decision Tree	0.227487
	Random Forest	0.156876
	<u>Extra Tree Regressor</u>	<u>0.153623</u>
	Gradient Boosting Regressor	0.177074
Deep Learning		0.194739

4. Discussion



4. 1. 주요 X-세권 파악

TOPSIS 결과로 도출된 371개의 feature 우선순위 중, 파레토 법칙에 따라 상위 20%인 74개 feature들을 크게 7개의 X-세권으로 분류

X-세권	특성	데이터 설명
역세권	지하철 역	지하철 역(이름), 지하철 호선(번호) 관련 역세권을 의미하는 지하철 관련 특성
학세권	학군	학문/교육, 유치원, 초등학교, 중학교, 고등학교 등 학군에 대한 접근성을 의미하는 특성
문세권	문화생활	영화관, 문화/예술 관련 시설과의 접근성을 의미하는 특성
구세권	자치구	자치구(이름)를 포함한 지역별 영향력을 의미하는 특성
몰세권	쇼핑	대형마트, 백화점과 같은 대형 쇼핑몰과 같은 시설과의 접근성을 의미하는 특성
안세권	안전	치안기관, 범죄율과 같이 안전 및 치안과 관련 특성
주세권	주민편의시설	주민센터, 구청, 보건소와 같은 주민 편의시설과의 접근성을 의미하는 특성



4. 1. 주요 X-세권 파악

1) 역세권(지하철 역)

도출된 Feature의 순위 중 상위 20%에 역세권 관련 특성이 절대 다수를 차지함
역 관련 데이터는 각 지하철 역(이름), 각 지하철 호선을 포함하고 있음

- 역세권은 다른 X-세권에 비해 주택 가격에 상대적으로 큰 영향을 주는 것으로 분석
- 신사역과의 거리는 주택 가격에 큰 영향을 미치는데, 이는 신사역 주변 역세권의 프리미엄이 크며, 가로수길 주변 지역으로서 흔히 알려진 인식과 비슷한 수준임
- 반면, 충무로역, 동대입구역, 개화산역 등은 역세권의 영향을 확인하기 어려움
- 1호선은 총 18개의 노선 중 주택 가격에 가장 큰 영향을 미치는 것으로 분석됨
- 3호선은 주택 가격과 상대적으로 무관한 것으로 보여, 역세권으로서 프리미엄이 미미함

Rank	Feature	Closeness
4	신사	0.421461695
10	압구정	0.208100394
27	논현	0.102786159

Rank	Feature	Closeness
22	1호선	0.121869105
153	3호선	0.006370683



4. 1. 주요 X-세권 피약

2) 학세권(학문/교육)

도출된 Feature의 순위 중 상위 20%에 학업과 관련 특성이 상당 수를 차지함

학군과 관련된 유치원, 초,중,고등학교 및 학문/교육 관련 소상공인 데이터를 포함하고 있음

- 유치원은 66위, 초등학교는 71위, 고등학교는 73위로는 상대적으로 주택 가격에 큰 영향을 미침
- 학문/교육은 63위로서 상대적으로 주택 가격과 상관 관계가 존재함
- 중학교는 상위 20% 안에 포함되지 않았음
따라서 이는 주택 가격에 유치원, 초등학교, 고등학교에 비해 중학교는 상대적으로 주택 가격에 큰 영향을 미치지 않는 것으로 분석됨
- 유치원, 초등학교, 고등학교는 맹모삼천지교와 같이 학군이 매우 중요하다는 사회적 통념을 잘 반영하고 있지만, 중학교는 상식과 다르게 다양한 학세권 중에서 상대적으로 중요성이 떨어진다는 것을 확인 할 수 있음

Rank	Feature	Closeness
63	학문/교육	0.019810767
66	유치원_1st	0.019337941
71	초등학교_3rd	0.016062696
73	고등학교_1st	0.01589276
83	중학교_1st	0.014878779



4. 1. 주요 X-세권 파악

3) 문세권(문화생활)

도출된 Feature의 순위 중 상위 20%에 문화 생활 관련 특성 데이터가 포함되어 있음
문세권 데이터는 주택과의 거리 기준 영화관 정보, 문화/예술 시설 정보를 포함하고 있음

- 영화관은 51위로서 상대적으로 주택 가격에 큰 영향을 미침
- 이는 영화관이 들어섬과 동시에 교통의 발달 및 유동 인구 증가를 통하여, 인근 주택 가격에 영향을 미치는 것을 잘 드러내고 있는 것으로 판단
- 따라서 영화관은 문세권으로서 프리미엄이 크다고 판단함

Rank	Feature	Closeness
51	영화관 _2nd	0.028459994
143	연극/영화	0.007472905



4. 1. 주요 X-세션 피약

4) 구세권(자치구)

도출된 Feature의 순위 중 상위 20%에 주택의 자치구 관련 특성이 상당 수를 차지함
이는 주택이 어떤 자치구에 포함되어 있는지에 대한 데이터를 포함하고 있음

- 서초구는 전체 feature 중 우선순위가 5위로서 상대적으로 주택 가격에 큰 영향을 미침
- 강서구, 동작구, 종로구는 상위 20%개 안에 포함되지 않았으며, 따라서 이들은 서초구에 비해 상대적으로 구세권의 프리미엄이 상대적으로 부족함
- 강서구는 특히 서울시 25개 구 중에서 주택가격에 가장 작은 영향을 미침
- 서초구는 강남 지역 자체가 주택 가격이 높다는 상식을 잘 반영하고 있음
- 반면 강서구는 예상과 다르게 자치구 중에서 상대적으로 주택 가격에 프리미엄을 제공하지 않는다는 것을 확인 할 수 있음

Rank	Feature	Closeness
5	서초구	0.286334689
141	종로구	0.007653649
335	동작구	0.00203149
371	강서구	0.000677448



4. 1. 주요 X-세권 파악

5) 안세권(범죄율, 치안기관)

도출된 Feature의 순위 중 상위 20% 내에 범죄 및 치안 관련 데이터가 분포함
이는 해당 지역의 범죄율, 치안기관에 대한 데이터를 포함하고 있음

- 구별 범죄율은 39위, 치안 기관은 73위로서 상대적으로 주택 가격에 큰 영향을 미침
- 특히 연립 주택과 치안기관과의 거리는 주택 가격에 큰 영향을 미치는 것으로 파악됨
- 반면, 범죄율은 예상과 다르게 주택 가격에 유의미한 영향을 미치는 것으로 분석
- 이에 따라, 범죄율 및 치안기관은 X-세권 중 안세권(안전)으로서 프리미엄이 있는 것으로 보임

Rank	Feature	Closeness
39	구 범죄율	0.051629107
73	치안기관	0.015847509



4. 1. 주요 X-세션 파악

6) 몰세권(소비생활)

도출된 Feature의 순위 중 상위 20% 중에 쇼핑 및 소비 생활 관련 특성이 도출됨
몰세권 데이터는 백화점, 대형마트 정보를 포함하고 있음

- 대형마트는 59위로서 상대적으로 주택 가격에 큰 영향을 미침
- 백화점은 상위 20%에 속하지 않아 상대적으로 주택 가격에 영향에 적음
- 이는 대형마트가 들어섬과 동시에 편리성의 증가, 교통의 발달 및 유동 인구 증가를 통하여, 인근 주택 가격에 영향을 미치는 것을 잘 드러내고 있는 것으로 판단
- 한편, 백화점은 일반적인 예상과 달리 주택 가격에 상대적으로 영향이 적은 것으로 분석
- 따라서 대형마트는 연립 주택 가격에 대한 몰세권으로서 프리미엄이 충분하나
백화점은 프리미엄이 충분치 않다고 판단함

Rank	Feature	Closeness
59	대형마트_2nd	0.023595668
120	백화점_1st	0.010887807



4. 1. 주요 X-세권 파악

7) 이외의 도출된 주요 Feature

- 도출된 Feature의 순위 중 상위 20% 중에 보건소, 구청이 포함됨
- 따라서 보건소, 구청은 주세권(주민 편의시설) 주택 가격에 큰 영향을 미치는 것으로 파악됨
- 반면, 주민센터는 보건소와 구청에 비해 상대적으로 주세권으로서 프리미엄이 적은 것으로 판단됨

Rank	Feature	Closeness
49	보건소	0.032818125
64	구청	0.019724075
102	주민센터	0.01327498

8) X-세권으로서 영향력이 적은 Feature

- 한의원, 동물병원, 약국, 일반 병원과 같은 병원 시설 Feature의 영향력이 전반적으로 낮음
- 따라서 의세권(의료 관련 시설)은 X-세권으로 분류되기에 전반적으로 프리미엄 정도가 부족한 것으로 판단

Rank	Feature	Closeness
126	일반병원	0.010290571
135	약국	0.008608182
139	동물병원	0.008263576
166	한의원	0.005257586



4. 2. 프리미엄 정도 예측

본 프로젝트에서는 앞서 추출한 주요 feature들에 의한 프리미엄 정도를 측정하기 위해 해당 feature의 실제 데이터를 수정하여 주택 가격의 변화 정도를 파악하는 분석을 진행함

- 지하철역이 가격에 미치는 영향을 파악하기 위해, 주택 하나의 모든 feature는 고정시킨 채 특정 지하철역까지의 거리를 등차수열로 증가시켜 100개의 데이터를 생성하고 3.2 절에서 생성한 모델에 투입하여 변화된 값의 추세를 파악함
- 자치구가 주택 가격에 미치는 영향을 파악하기 위해, 모든 자치구 column의 성분을 0으로 지정한 뒤, 분석하고자 하는 자치구에 해당하는 column의 성분을 1로 변환하여 3.2 절에서 생성한 모델에 투입하여 변화된 값을 예측함
- 시설 feature가 주택 가격에 미치는 영향을 파악하기 위해, 해당 feature의 column 값을 해당 column의 최대값, 최소값으로 변화시킨 데이터를 생성한 뒤 각각의 데이터를 3.2 절에서 생성한 모델에 투입하여 변화된 값을 예측함
- 데이터 변경은 test data에 대하여 수행하였으며, 변경하기 전의 test data와의 비교를 진행함

신사역
100
200
300
400
500

타 데이터는 고정하고 역만 변환

서초구	강남구	송파구
0	1	0
0	0	1
0	0	1
0	1	0
0	1	0

여러 자치구를 하나의 자치구로 통합

서초구	강남구	송파구
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0

영화관_2nd	영화관_2nd	영화관_2nd
121.664785	5079.27352	10040.19687
121.664785	5136.540951	10040.19687
121.664785	5105.512647	10040.19687
121.664785	5063.167114	10040.19687
121.664785	5069.882154	10040.19687

열 최소값, 열 최대값으로 변환



4. 2. 브리미엄 정도 예측

데이터셋	데이터셋 설명
X_test	전체 X 데이터를 0.8 : 0.2로 나눈 test data
y_test	전체 y 데이터를 0.8 : 0.2로 나눈 test data
y_predict	X_test 데이터를 통해 모델이 예측한 주택 가격 데이터
X_station	주택 하나의 타 데이터는 고정하고, 특정 역과의 거리를 등차수열로 변환한 test data
y_station	X_station 데이터를 통해 모델이 예측한 주택 가격 데이터
X_test_gu	X_test 데이터의 모든 자치구 정보를 특정 자치구로 변경시킨 test data
y_test_gu	X_test_gu 데이터를 통해 모델이 예측한 주택 가격 데이터
X_test_max	X_test 데이터 중, 특정 column의 값을 해당 column의 최대값으로 변환한 test data
X_test_min	X_test 데이터 중, 특정 column의 값을 해당 column의 최소값으로 변환한 test data
y_predict_max	X_test_max 데이터를 통해 모델이 예측한 주택 가격 데이터
y_predict_min	X_test_min 데이터를 통해 모델이 예측한 주택 가격 데이터

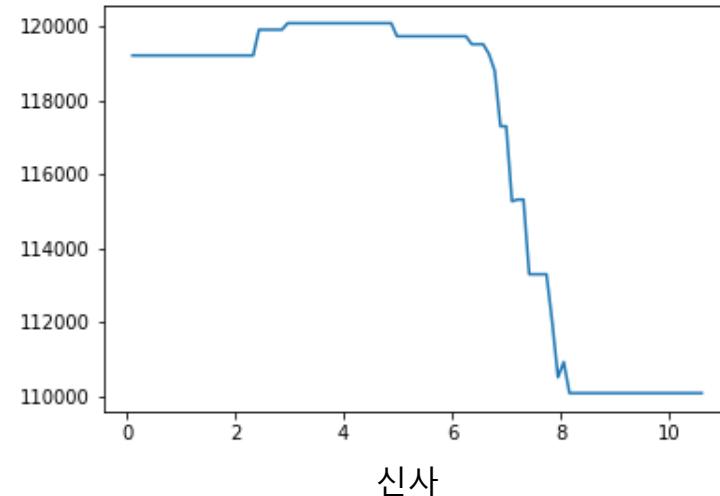
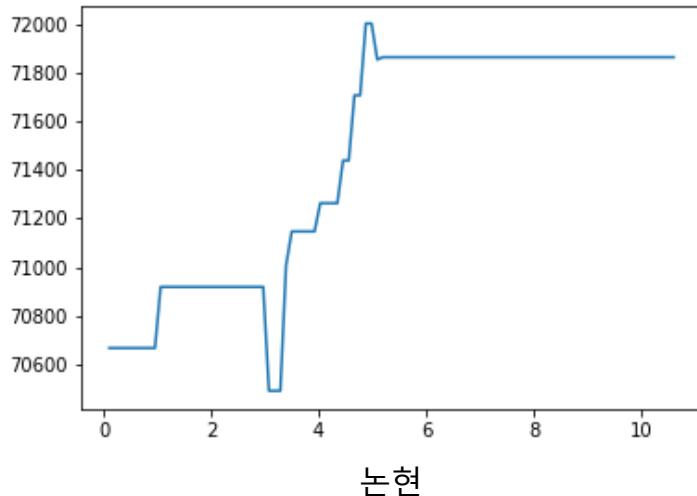
현재 y_test 데이터는 로그 스케일링된 상태이기 때문에, 변형 후 예측한 값 역시 로그 스케일링되어 있음
따라서 직관적인 비교를 위해 y 데이터들은 자연상수에 대한 거듭제곱 과정을 거친 뒤 비교에 사용됨
이후 등장하는 y 데이터들은 자연상수에 대한 거듭제곱 과정을 거친 수치임



4. 2. 브리미엄 정도 예측

1) 역세권(지하철 역)

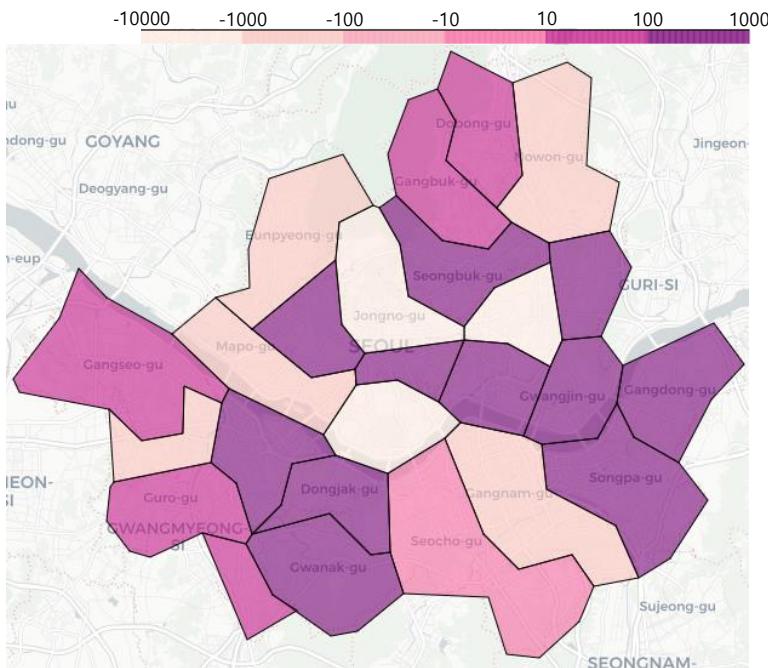
- TOPSIS 결과로 상위 우선 순위를 받은 지하철역들까지의 거리가 특정 건물에 어떠한 영향을 미치는지 파악하기 위함
- 건물 하나를 지정하여, 특정 역까지의 Manhattan distance를 점진적으로 증가시켜 증가된 거리에 따른 주택 가격 변화를 확인함
- 논현역의 경우 역에서부터 일정 거리를 넘어서면 가격이 상승하고, 신사역의 경우에는 일정 거리를 넘어서면 가격이 하락함
- 사용된 모델이 Extra Tree Regressor기 때문에 결과값이 연속적이지 않음





4. 2. 프리미엄 정도 예측

2) 구세권(자치구)



- Test 데이터의 모든 자치구를 서초구로 변경
- 서초구는 TOPSIS 결과 우선 순위 5위로, 전체 자치구 중 가장 높은 우선 순위를 받음
- $y_{\text{test}}_{\text{gu}}$ 와 y_{predict} 값을 구별로 비교
- 모든 구를 서초구로 바꾸었을 때, 기존 가격 평균과 바뀐 가격 평균의 차이를 지도에 나타냄
- 서초구가 절반 이상의 자치구 보다 예측 가격이 높은 것으로 보아, 서초구에 속한다는 것만으로도 구세권 프리미엄이 존재하는 것이 확인됨



4. 2. 프리미엄 정도 예측

3) 주요 X-세권의 단일 Feature 프리미엄 정도 예측

- 학세권, 문세권, 안세권, 몰세권에 대한 개별 Feature의 프리미엄 정도 측정
- 해당 Feature의 실제 데이터를 수정하여 주택 가격의 변화 정도를 파악
- 구별 주택 가격의 기초 통계량을 확인해본 결과, 금천구에 위치한 주택들의 표준편차가 가장 작음
- 따라서 주택 가격이라는 우연변동이 최소화된 금천구의 X_test 데이터의 최소, 최대값으로 조정하여 가격 예측 확인

[X 단위 : 미터]
[y 단위 : 만원]

X-세권	Feature	X_test_min mean	X_test_max mean	y_test_min mean	y_test_max mean	y_pred mean
학세권	유치원_1st	53.64503	1419.33164	19364.90869	19409.81437	19414.88911
문세권	영화관_2nd	801.53445	6337.27141	19607.82345	19311.90189	19414.88911
안세권	구 범죄율	1.85375	7.15908	19401.79199	19496.99251	19414.88911
몰세권	대형마트_2nd	863.75846	4692.14628	19500.73057	19480.60393	19414.88911



4. 2. 프리미엄 정도 예측

4) 주요 X-세권의 Feature 간의 상관관계



- 현실에서 주택 가격에 영향을 주는 것은 단일 Feature가 아닌, 복수의 X-세권 Feature임
- Feature 간의 상관관계를 확인하여, 하나의 Set의 X-세권 Feature를 확인하고자 함
- 학세권이(초등학교_3rd, 유치원_3rd, 유치원_2nd, 초등학교_2nd, 중학교_3rd) 높은 상관관계를 보임
- 위 5가지 Feature를 적용하여 복합적인 학세권의 프리미엄을 분석하고자 함



4. 2. 프리미엄 정도 예측

5) 학세권 주요 Feature들의 복합적 프리미엄 정도 예측

[X 단위 : 미터]
[y 단위 : 만원]

학세권_Feature	X_test_min mean	X_test_max mean	y_test_min mean	y_test_max mean	y_pred mean
초등학교_2nd	70.548695	2125.215855	19284.99696	19465.56722	19414.88911
초등학교_3rd	74.930969	2994.865493			
유치원_2nd	161.573162	2361.578453			
유치원_3rd	353.159167	2377.118342			
중학교_3rd	357.203282	3530.138337			

- 높은 상관관계를 보인 학세권 (초등학교_3rd, 유치원_3rd, 유치원_2nd, 초등학교_2nd, 중학교_3rd) 분석
- 현실에서는 학세권의 단일 Feature가 아닌 복수의 Feature가 주택 가격에 복합적으로 작용함
- 각 Column의 최소값, 최대값을 넣어 모델로 예측하였을 시, 위와 같은 결과를 확인 할 수 있음
- 학세권 내에서 다양한 Feature들의 복합적으로 작용함으로써 종합적 학세권의 프리미엄 파악
- 위와 같은 방식으로 수많은 조합의 Feature 및 X-세권에 대한 분석이 가능함

5. Conclusion



5. 1. 프로젝트 정리

1. Project Objective

- 주택 가격은 평수, 건축년도 이외에도 지하철 역, 학군, 상권 등 다양한 요소들에 영향을 받음
- 하지만, 이러한 요소들을 대량의 데이터를 과학적으로 분석하여 주택 가격에 미치는 영향을 파악한 연구들이 진행되지 않음
- 따라서 본 프로젝트는 주택 가격에 영향을 미치는 X-세권의 영향력 파악 및 프리미엄 정도 예측을 위한 분석을 진행함

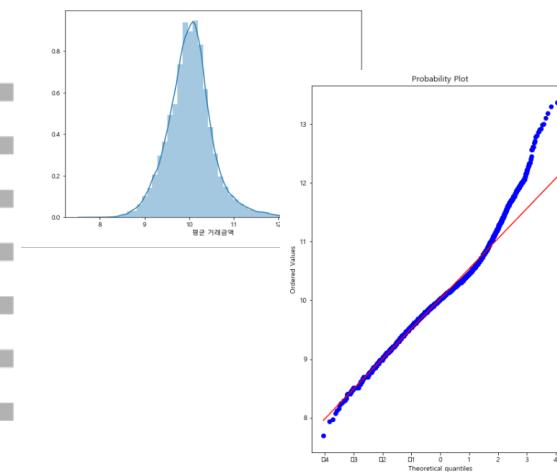
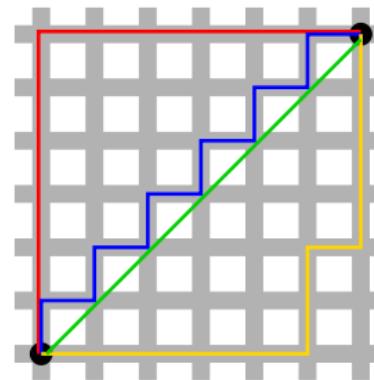
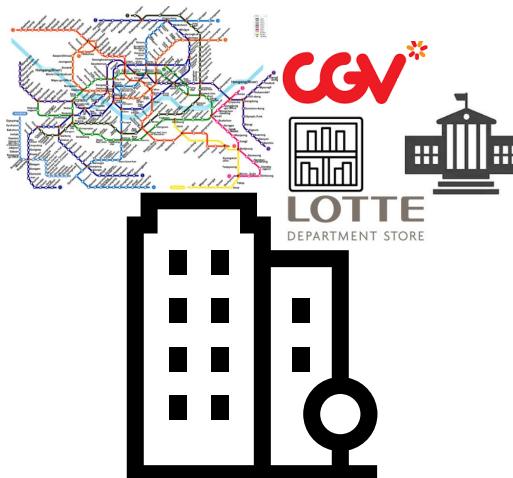




5. 1. 프로젝트 정리

2. Feature Engineering

- 본 프로젝트에서는 분석을 위해 서울에 존재하는 연립주택의 매매가, 평수, 건축년도 등의 데이터를 수집하였으며, 가격에 영향을 줄 수 있는 요인을 찾기 위해 지하철역, 공공시설, 학군, 상권으로 분류되는 17개의 시설 데이터를 수집함
- Python의 googlemaps 라이브러리를 통해 수집한 데이터의 위도, 경도를 추출한 뒤, haversine 라이브러리를 통해 건물과 시설과의 Manhattan Distance를 계산함
- 이후, 시설의 특징에 맞게 다양한 형태로 데이터를 변환 및 스케일링하여 총 26542×371 형태의 데이터셋 구축

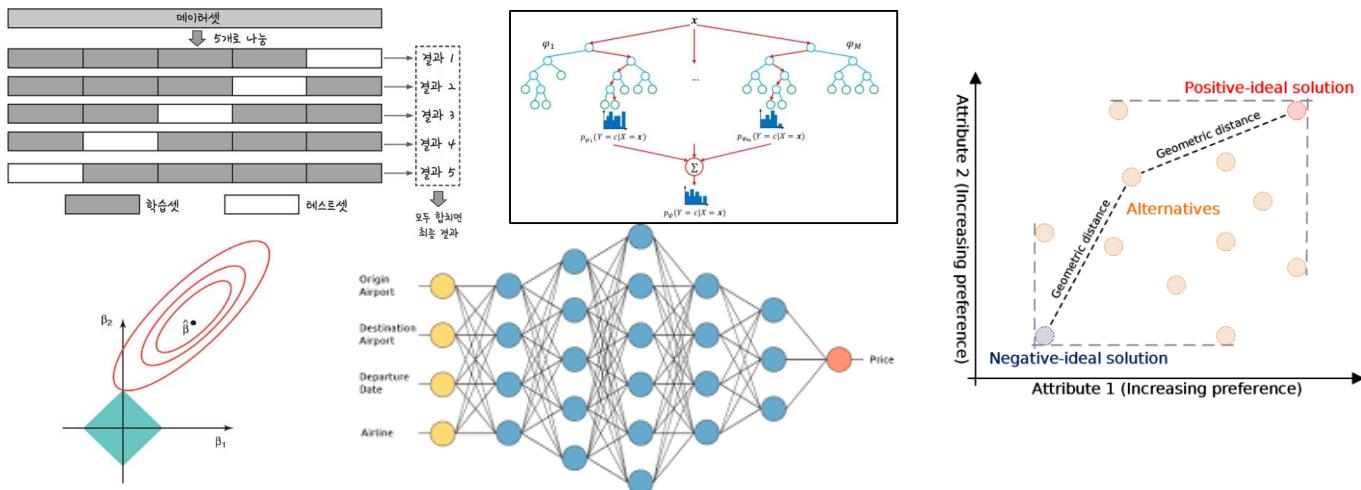




5. 1. 프로젝트 정리

3. Data Analysis

- 본 프로젝트에서는 가격에 영향을 미치는 주요 feature를 선정하기 위해 Lasso, Decision Tree, Random Forest, Extra Tree Regression, Gradient Boosting Regression의 결과를 기반으로 TOPSIS를 수행함
- 또한, 본 프로젝트에서는 가격을 예측할 수 있는 모델을 구축하기 위해 선형회귀 기반의 모델, Tree 기반의 모델, 그리고 Deep Learning 모델을 구축하고 성능을 평가함
 - 평가 결과 Extra Tree Regression이 가장 좋은 성능을 보임





5. 1. 프로젝트 정리

4. Discussion

1) 주요 X-세권 파악

- TOPSIS로 도출한 371개의 Feature 기반으로 7개의 주요 X-세권 분류
역세권, 학세권, 문세권, 구세권, 몰세권, 안세권, 주세권으로 분류함
- 7개의 주요 X-세권으로서 역할을 하는 Feature 파악

2) 프리미엄 정도 예측

- 주출한 주요 Feature들에 의한 프리미엄 정도를 측정하기 위해
해당 Feature의 실제 데이터를 수정하여 주택 가격의 변화 정도를 파악하는 분석을 진행
- 단일 Feature로서 학세권의 유치원_1st, 문세권의 영화관_2nd 등에서
유의미한 가격 차이를 확인
- 복합적인 Feature로서 학세권의 초등학교_3rd, 유치원_3rd, 유치원_2nd, 초등학교_2nd,
중학교_3rd를 동시 수정하여 프리미엄 정도 예측
- 학세권 내에서 위 Feature들이 복합적으로 작용함으로써 종합적 학세권의 프리미엄 파악
- 위와 같은 방식으로 수많은 조합의 Feature 및 X-세권에 대한 분석이 가능함



5. 2. 기여점 및 활용방안

1) 신개발 지구의 주택 가격 상승폭 예측

- 개발 예정 지역의 Feature 변화를 통한 실거래가 변동폭 예측
- 학교, 백화점, 편의시설 등 새로운 시설이 생겼을 때의 가격 변동 여부 파악 가능

2) 실거래 데이터 기반으로 부동산 가치에 영향을 주는 주요 Feature 식별

- 지역 이해관계자들의 사업 투자 지원 정보 제공
- X-세권의 범위와 프리미엄 정도를 정의할 수 있는 표준 생성

3) 다각적 부동산 정책 지원을 위한 부동산 가치 예측 모델 생성

- 빅데이터 학습 모델에 기반하여 정책 결정자의 의사결정 지원
- 무분별한 X-세권 프리미엄 광고 분별



5. 3. 한계점 및 추후연구

1) 주택 가격에 영향을 줄 수 있는 Feature 추가

- 본 프로젝트에서는 주택 가격에 영향을 미치는 요인으로 주변 시설만 고려하였는데, 인구통계학적 지표와, 정책적 지표를 추가로 반영하여 주택 가격에 대한 추가적인 고려가 필요할 것으로 예상됨

2) 아파트, 오피스텔에 대한 모델 생성

- 본 프로젝트에서는 연립 다세대 주택에 대해서 분석을 실시하였으나, 추후 아파트와 오피스텔의 브랜드 가치를 고려한 모델을 구축 할 수 있을 것으로 보임

3) 모델 성능 개선

- 파라미터 최적화에 대한 추가적인 학습을 통해 Extra Tree Regression보다 우수한 성능을 보유한 Deep Learning 모델 구축 필요

THANK YOU

2019 Spring Data Analytics

BADA