2019 Spring Data Analytics

BADA

Budongsan Analytics
Data Analytics

INTERIM

X-세권의 실질적 범위와 프리미엄 정도 측정

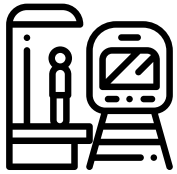
201411180 정재민
201311167 이승윤
201411160 송용백
201514181 박영재
201611171 제갈용승

INDEX

1. Project Review

1. 1. Kick-off: 역세권

“역세권의 실질적 범위와 프리미엄 정도 측정”



역세권이란?

지하철역으로의 접근이 용이한 범위

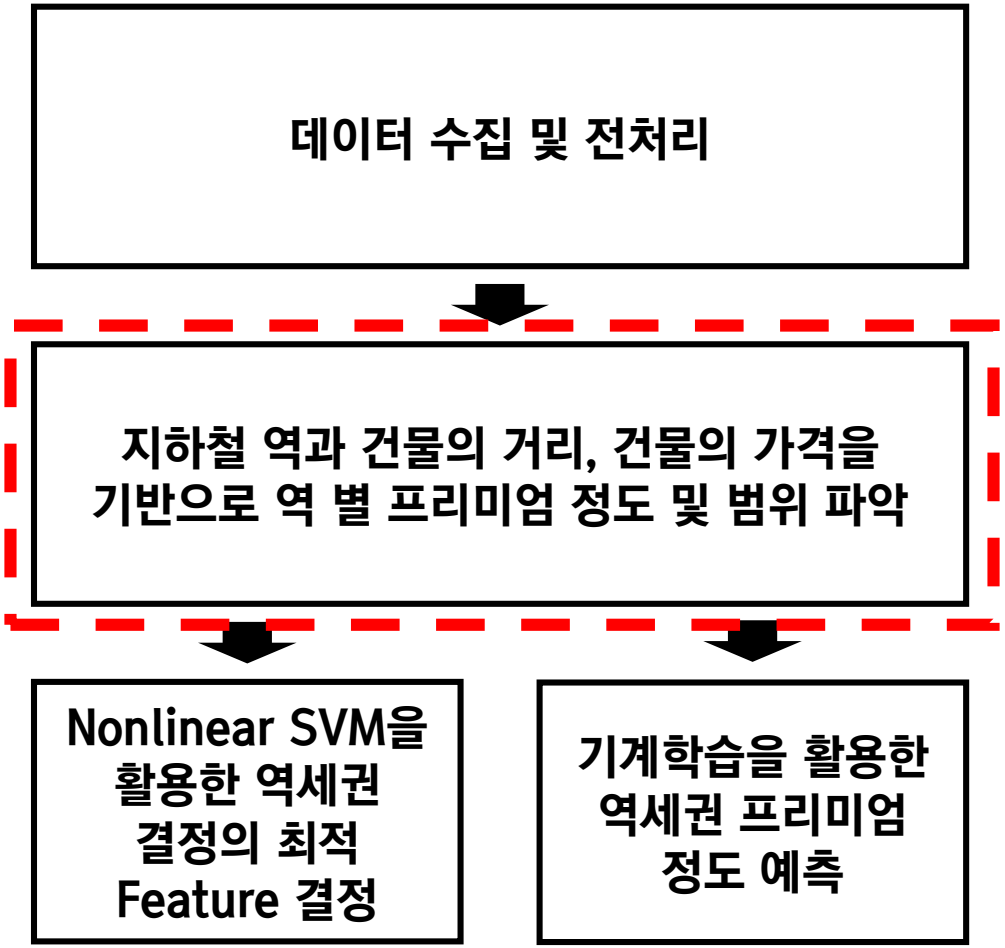
역세권의 실질적 범위

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

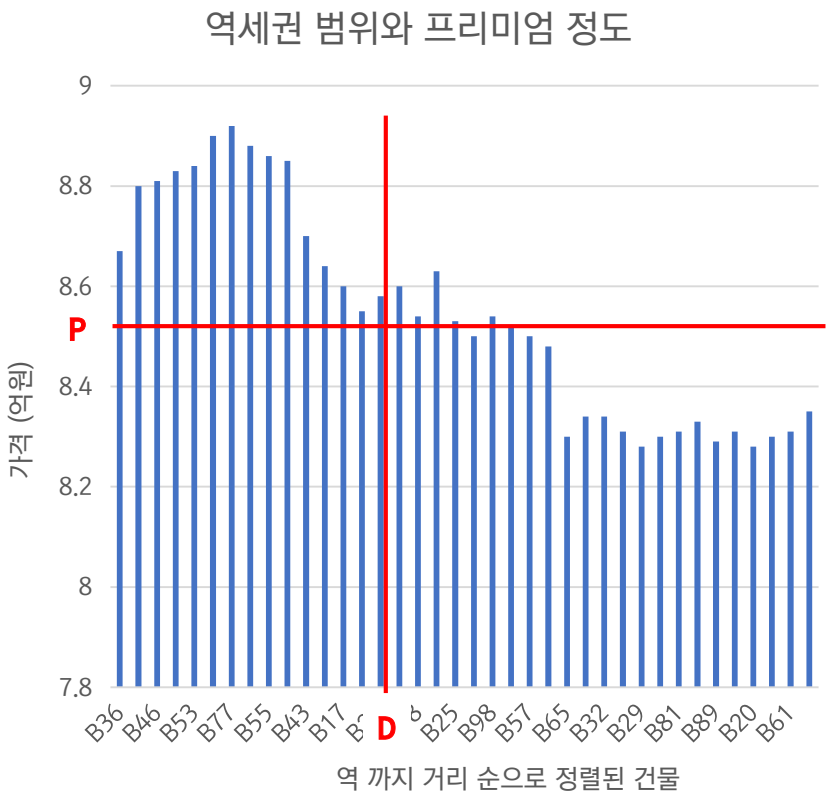
역세권의 프리미엄

역세권에 속함으로써 증가한 비용

1. 1. Kick-off: 분석 절차



1. 1. Kick-off: 예상결과

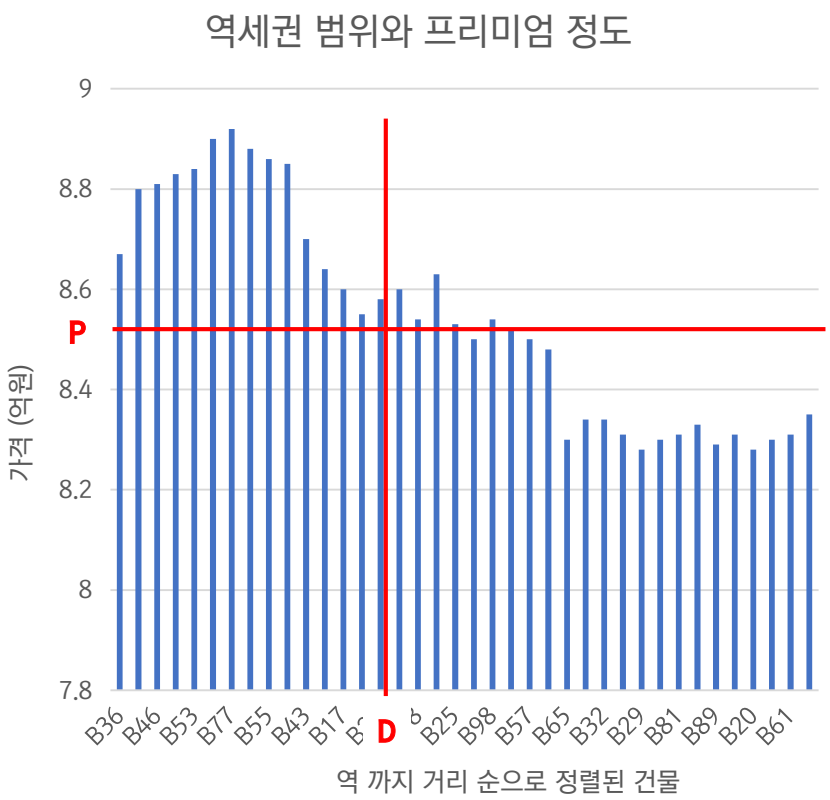


D: 역세권의 실질적 범위

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

1. 1. Kick-off: 예상결과



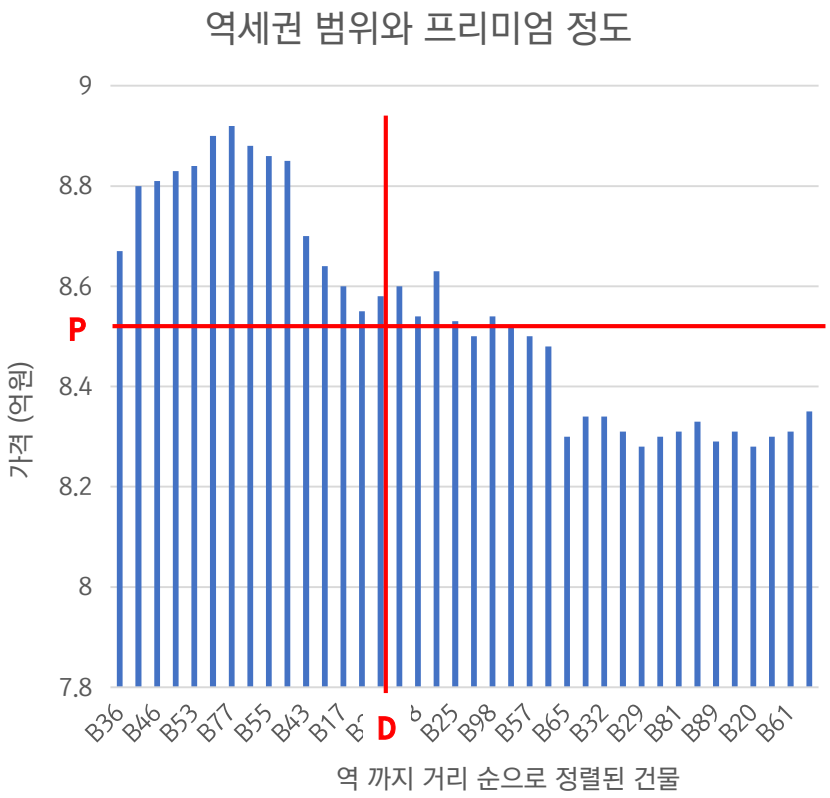
D: 역세권의 실질적 범위

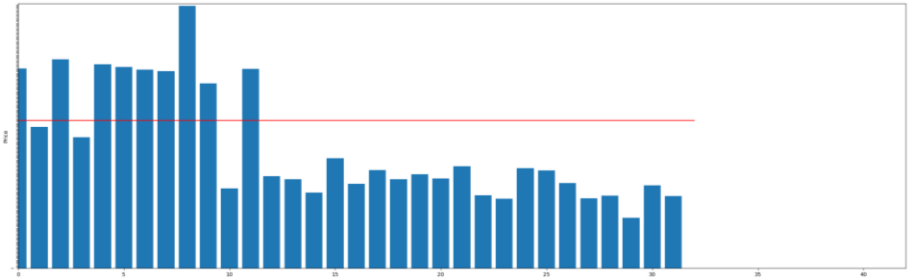
P: 해당 자치구의 평균 매매가

250m로 정의된 역세권 범위

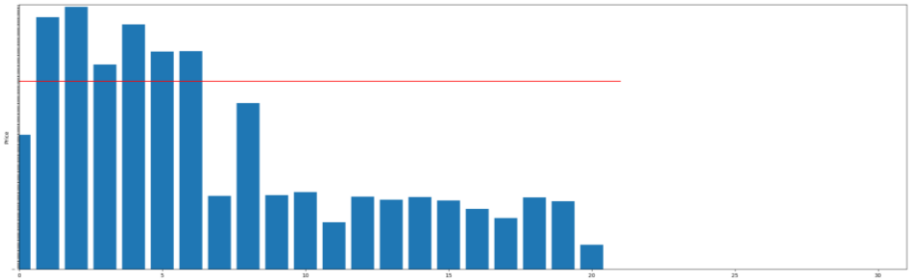
분석을 통해 정의한 역세권의 실질적 범위

1. 1. Kick-off: 기초분석 결과





버티고개역



응봉역

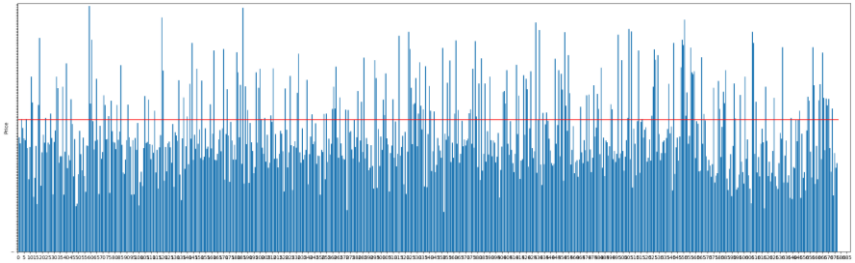
D: 역세권의 실질적 범위

P: 해당 자치구의 평균 매매가

1. 1. Kick-off: 기초분석 결과

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

1. 그래프에 추세가 보이지 않음

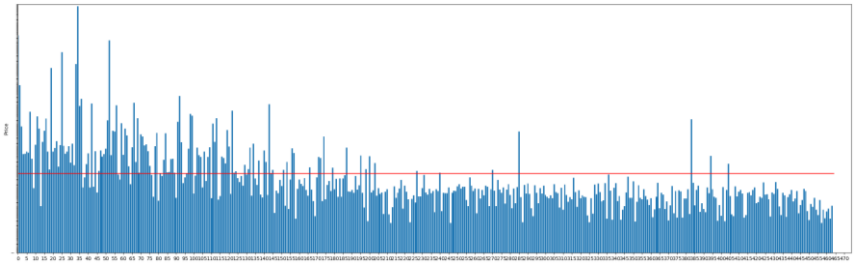


응암역

2. 역에 해당하는 건물이 많이 없음

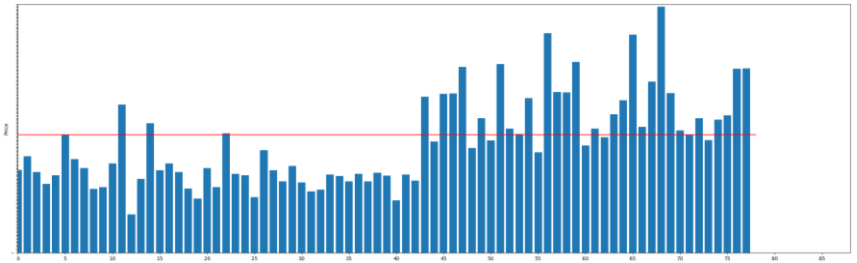
강남역

3. 범위를 지정하기 애매함



신정네거리역

4. 거리가 멀어질수록 가격이 올라감



구파발역

1. 1. Kick-off: 기초분석 결과

역까지의 거리와 건물의 가격만으로는

역세권의 실질적 범위와 프리미엄을 측정할 수 없음

1. 2. New Approach

건물의 가격에 영향을 주는
X-세권을 파악하는 분석을 수행하고자 함

1. 2. New Approach: 분석절차

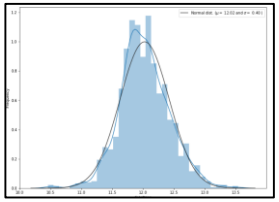
Data Collection, Preprocess

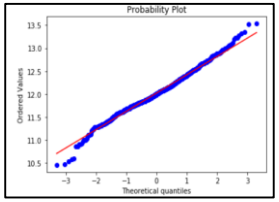
Exploratory Data Analysis

Dimension Reduction

Predictive Modeling

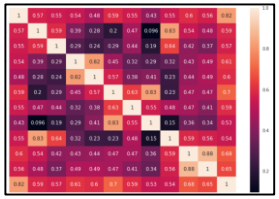
Prediction, Contribution

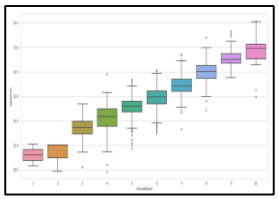




Normal Distribution

Probability Plot

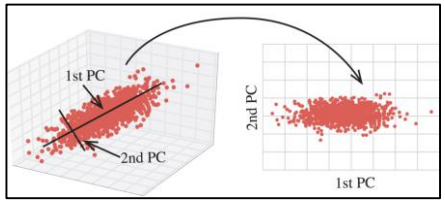




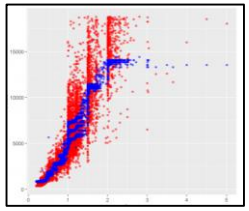
Missing Value

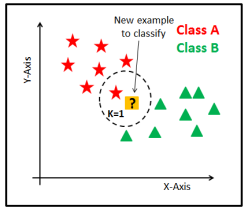
Correlation

Target Feature



PCA, SVD





Linear Regression

SVM

Decision Tree

Random Forest

Deep Learning

1. 2. New Approach: 활용방안

1) 신개발 지구의 집 값 상승폭 예측

- 개발 예정 지역의 Feature 변화를 통한 실거래 상승가격 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

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

2. Data Collection & Preprocessing

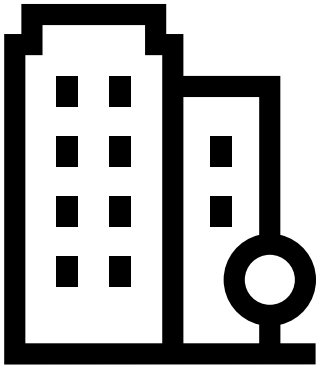
2. 1. 수집데이터

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. 수집데이터

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

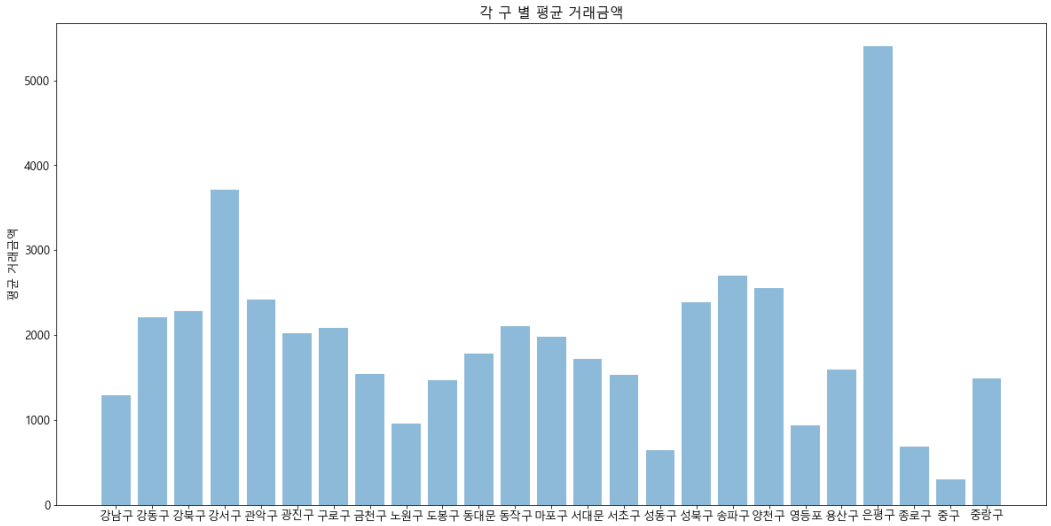
데이터	데이터 설명	출처	기간 또는 데이터 생성일
건물	서울시 연립다세대 주택 및 실거래 정보	국토교통부 실거래가 공개시스템	2018.01~2018.12
지하철 역	서울시 지하철 호선별 역별 승하차 인원 정보	서울열린데이터광장	2015.01~2019.03
영화관	서울시 위치, 브랜드 별 영화관 정보	네이버 영화관 정보	2019.05 (웹 크롤링)
백화점	서울시 내 백화점 위치 데이터	네이버 지도	2019.05 (웹 크롤링)
대형마트	서울시 내 대형마트 위치 데이터	네이버 지도	2019.05 (웹 크롤링)
공원	서울시 공원 현황 데이터	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

2.1. 수집데이터

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

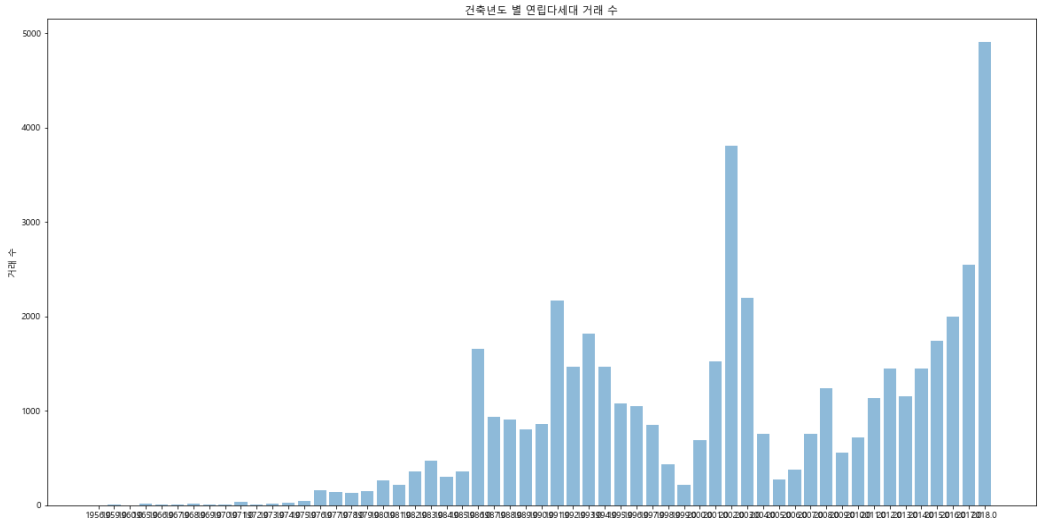
데이터	데이터 설명	출처	기간 또는 데이터 생성일
주민센터	서울시 소재의 주민센터 정보	서울열린데이터광장	2016.02.16
구청	서울시 소재의 구청 정보	서울열린데이터광장	2016.02.16
보건소	서울시 소재의 보건소 정보	서울열린데이터광장	2016.02.16
범죄율	서울시 행정구역별 5대 범죄발생 누적 정보	서울열린데이터광장	2012.03~2018.03
유치원	서울시 구별 유치원 이름, 주소 등 공간 정보	공공데이터포털	2017.04.01
초등학교	서울시 구별 초등학교 이름, 주소 등 공간 정보	공공데이터포털	2017.04.01
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. 수집데이터: 건물



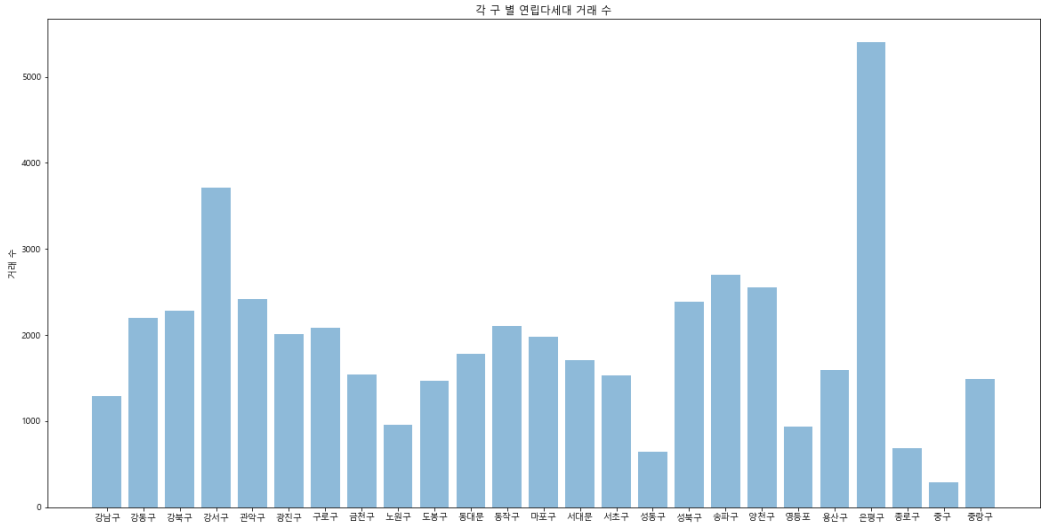
- 연립다세대 주택 평균 거래금액
- 2018년 1월 ~ 2018년 12월까지의 구별 연립 다세대 주택 평균 거래 가격
- 작년 서울시 연립 다세대 주택 평균 거래 가격은 2억 6천 8백 80 만원
- Data shape: (47746, 11)

2.1. 수집데이터: 건물



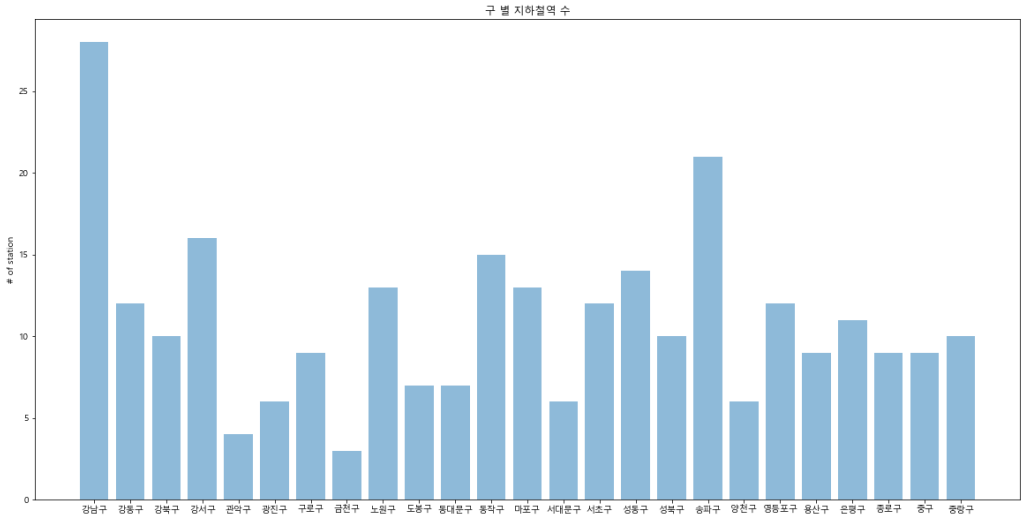
- 건축연도 별 연립다세대 주택 거래량
- 1956년 ~ 2018년까지 건축된 연립다세대 주택 건축연도 별 평균 가격
- 가장 최근 지어진 2018년 거래량이 5858건으로 가장 많음
- Data shape: (47746, 11)

2. 1. 수집데이터: 건물



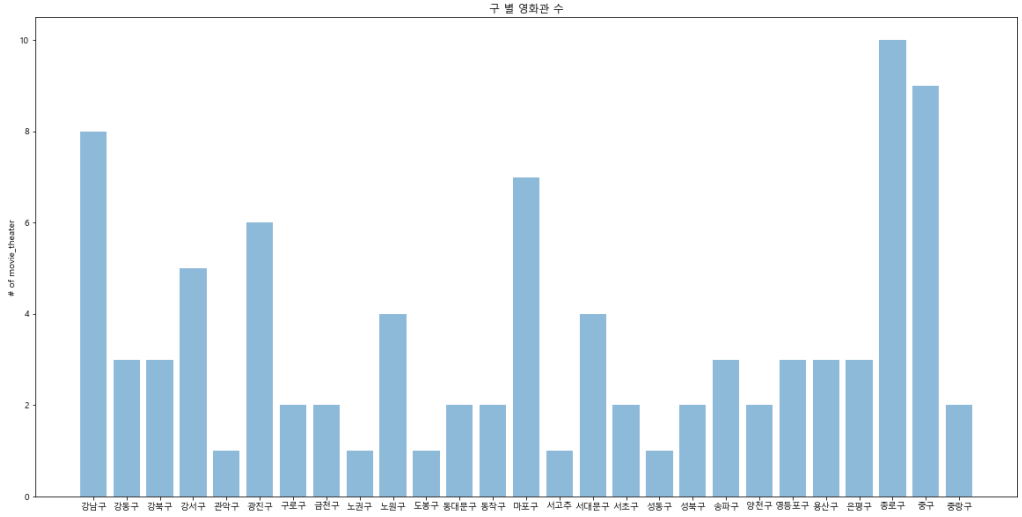
- 구 별 연립다세대 주택 거래량
- 1956년 ~ 2018년까지 건축된 연립다세대 주택 건축연도 별 거래량
- 은평구 연립다세대 주택 거래량이 5404건으로 가장 많음
- Data shape: (47746, 11)

2.1. 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)

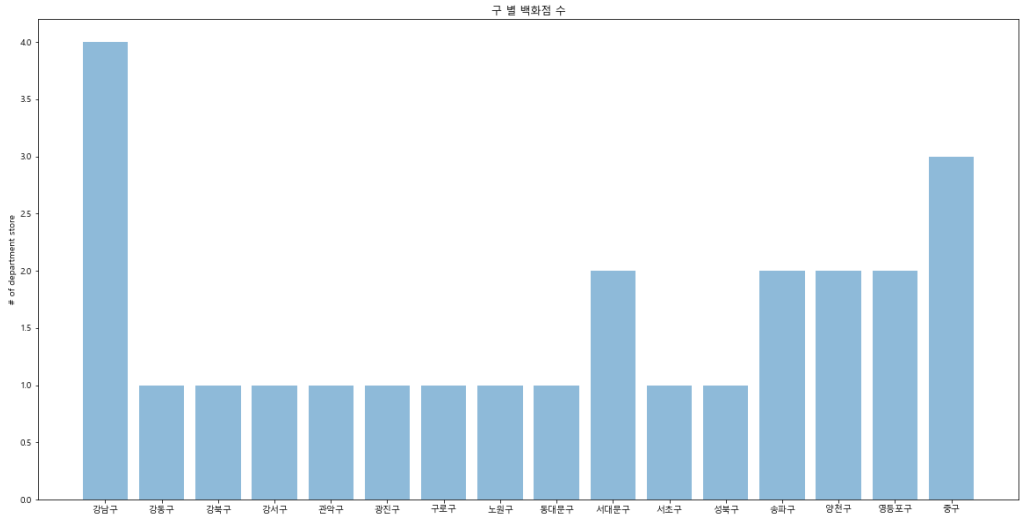
2. 1. 수집데이터: 영화관



- 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)

https://cnbc.sbs.co.kr/article/10000941395?division=NAVER

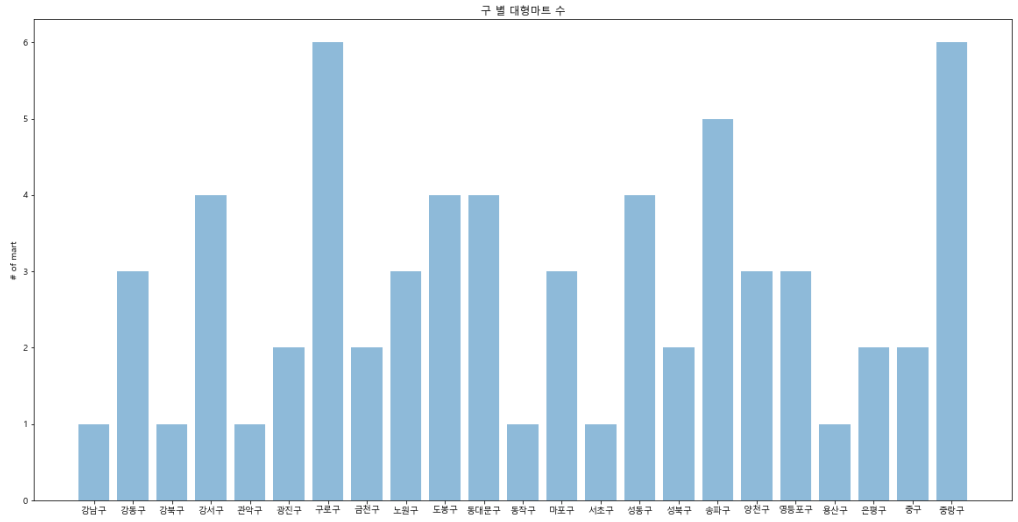
2. 1. 수집데이터: 백화점



- 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)

http://www.fnnews.com/news/201811191614284541

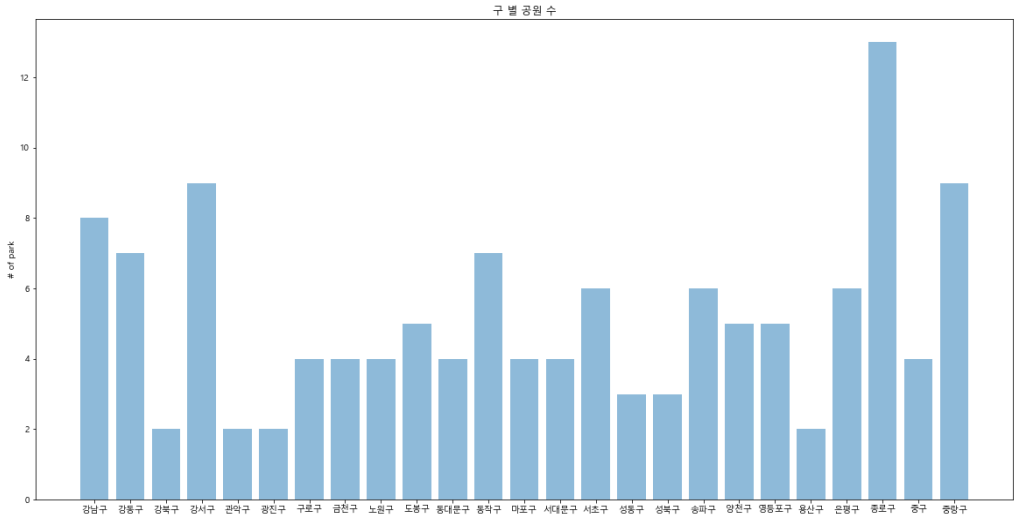
2.1. 수집데이터: 대형마트



- 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)

http://m.cnews.co.kr/m_home/view.jsp?idxno=201703151036023590136#cb

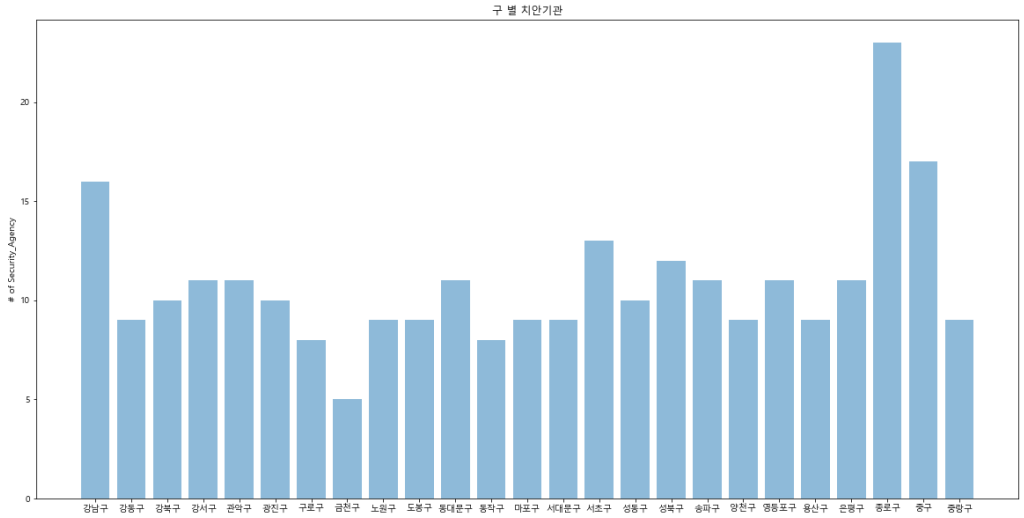
2.1. 수집데이터: 공원



- 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)

http://www.kapanet.or.kr/kapabook/78/PDF/22_Think%20Tank.pdf

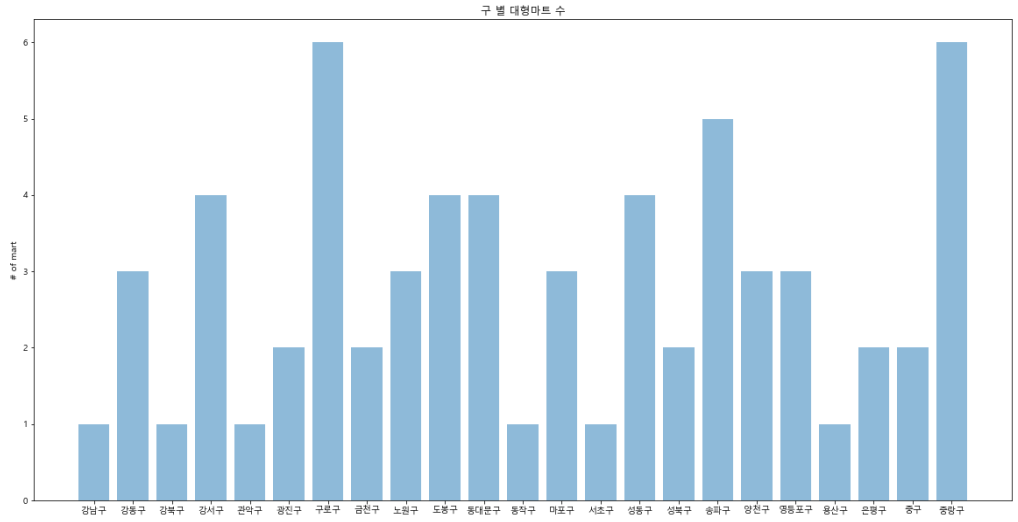
2.1. 수집데이터: 치안기관



- 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

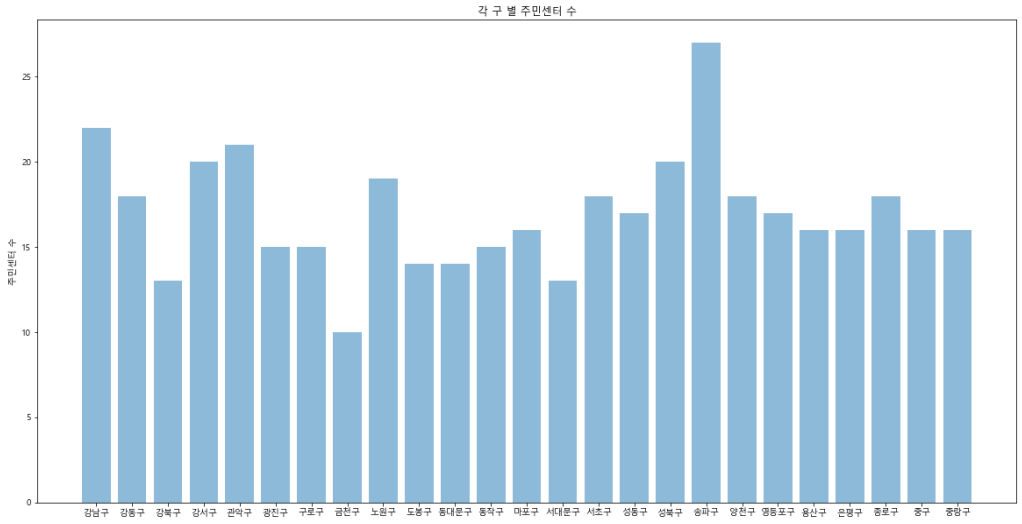
2. 1. 수집데이터: 대형병원



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

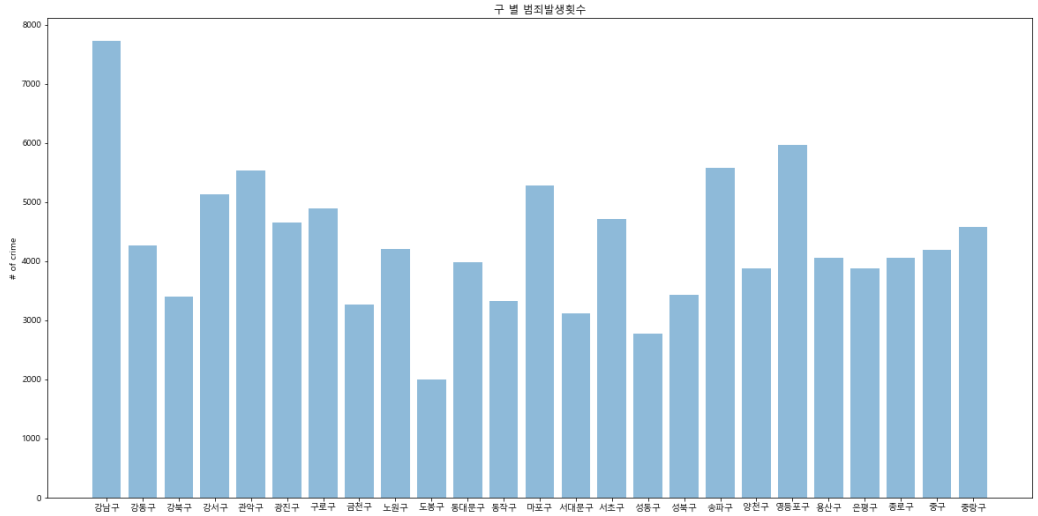
http://www.fnnews.com/news/201708041728501795

2.1. 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)

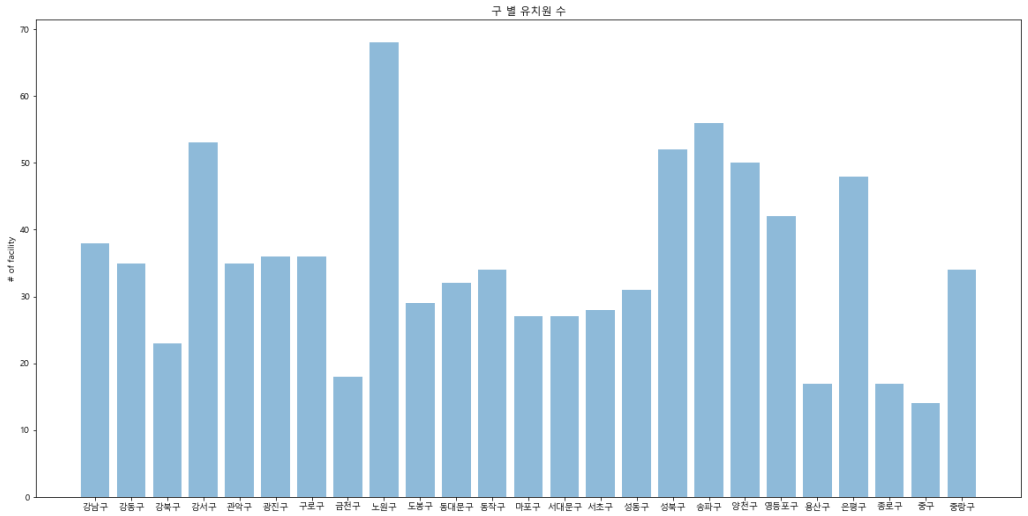
2.1. 수집데이터: 범죄율



- 2012~2018년도까지 집계된 각 구별 5대 범죄 누적 발생횟수 정보
- 살인, 강도, 강제추행, 절도, 폭력 5대 범죄의 발생횟수를 통해 각 구의 상대적 범죄율을 도출하며 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

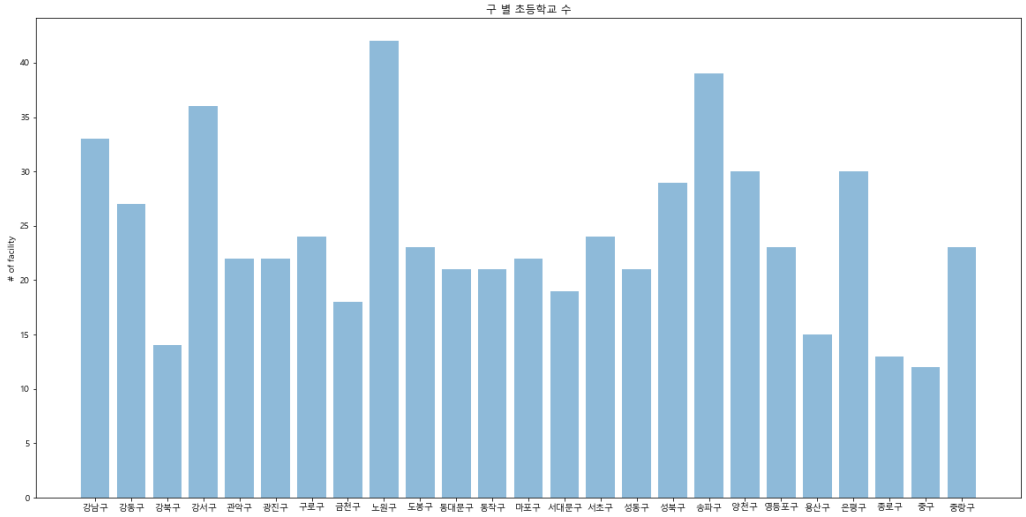
2. 1. 수집데이터: 유치원



- 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)

https://www.sedaily.com/NewsView/1S79GMF6YP

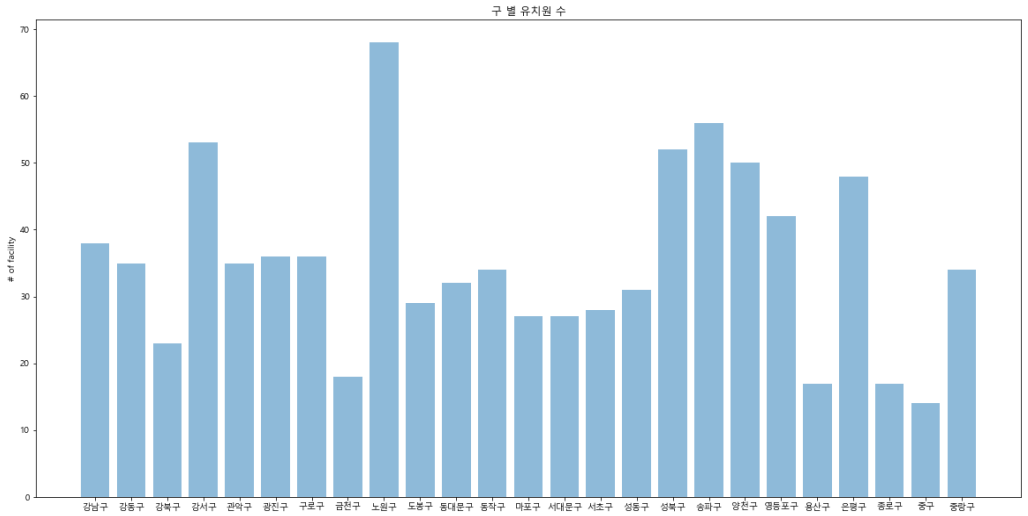
2. 1. 수집데이터: 초등학교



- 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)

https://www.mk.co.kr/news/realestate/view/2016/09/683007/

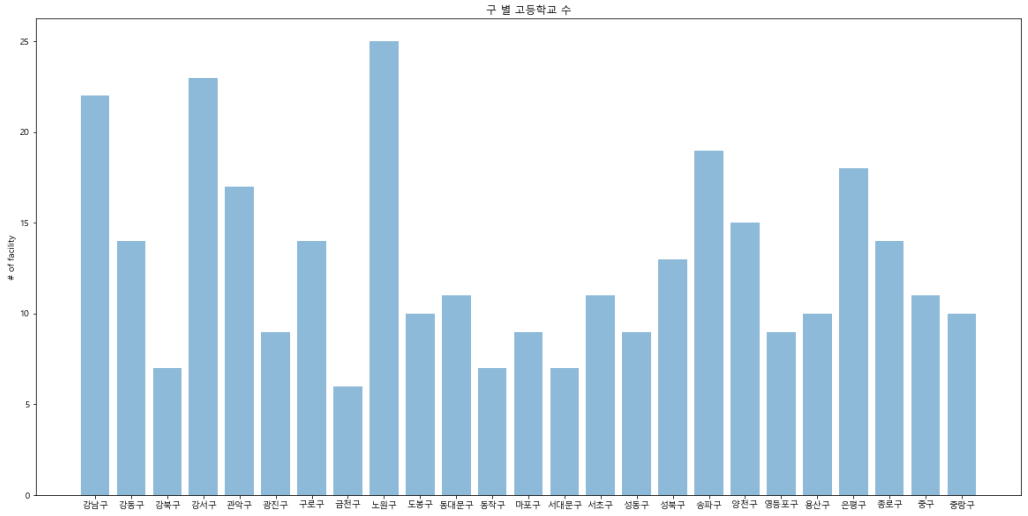
2. 1. 수집데이터: 중학교



- 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)

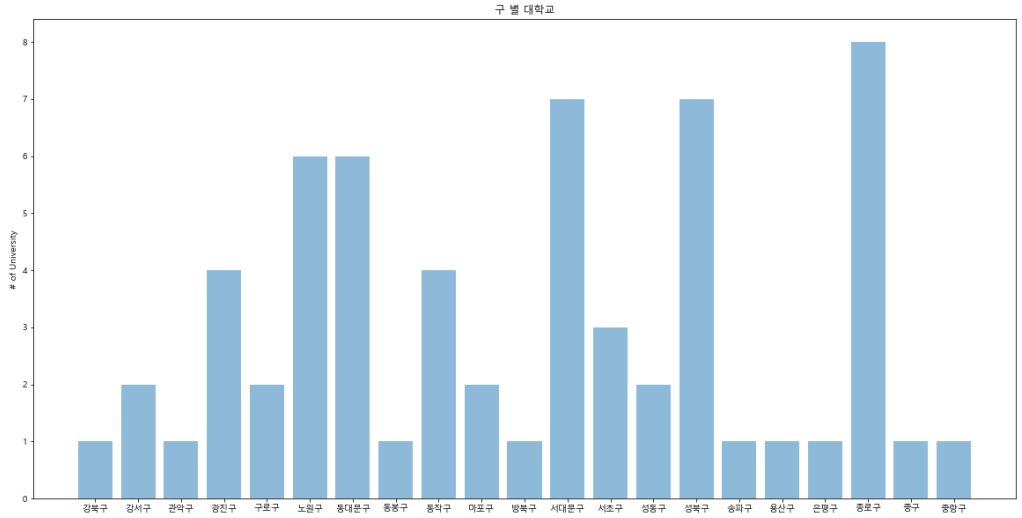
https://www.hankyung.com/realestate/article/201804043228a

2. 1. 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. 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)

2.2. 데이터 전처리 과정

역 데이터
지하철 역 데이터 수집

→

googlemaps 라이브러리

→

데이터 필터링
서울에 존재하는 역 선별
및 소속 구 정보 추출

수집 데이터
역, 건물, 시설의 주소 데이터

→

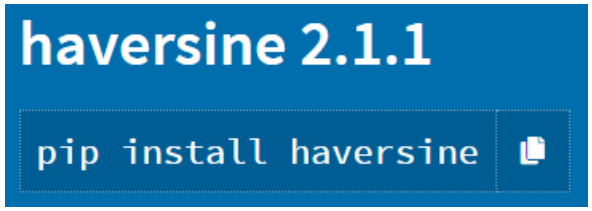
googlemaps 라이브러리

→

위치 데이터
역, 건물, 시설의 위도, 경도

위치 데이터
역, 건물, 시설의 위도, 경도

→



haversine 라이브러리

→

거리 데이터
건물과 역, 시설 사이의 거리

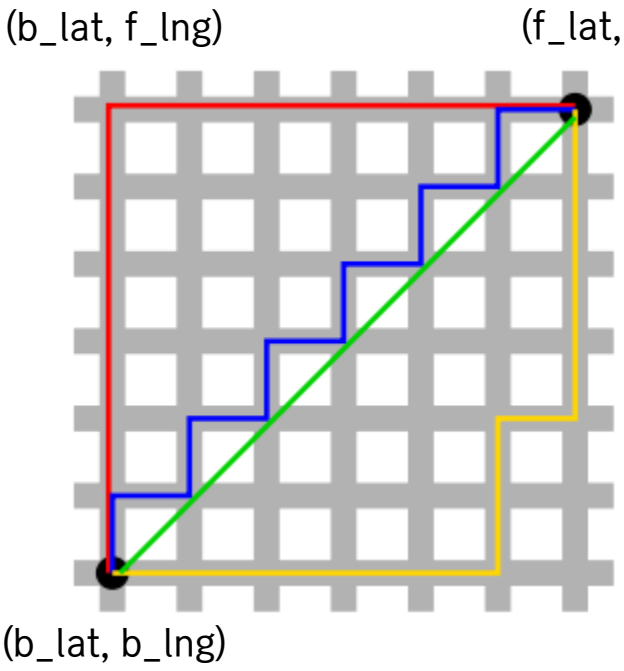
2. 2. 데이터 전처리 과정: googlemaps 라이브러리


```
In [7]: gmaps.geocode('구의역', language='ko')
Out[7]:
[{'address_components': [{'long_name': '2 1 6-2 2',
    'short_name': '2 1 6-2 2',
    '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': 'ChIJG0ulLiClfDURjhyiugOQwhM',
  'plus_code': {'compound_code': 'G3PP+R9 대한민국 서울특별시',
   'global_code': '8Q99G3PP+R9'},
  'types': ['establishment', 'point_of_interest']}]
```

2. 2. 데이터 전처리 과정: haversine 라이브러리

haversine 2.1.1

```
pip install haversine
```



f_lat: 시설의 위도
f_lng: 시설의 경도
b_lat: 건물의 위도
b_lng: 건물의 경도

```
# m단위의 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. 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

- 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

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. 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	가산디지털단지	가양	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. 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. 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. 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

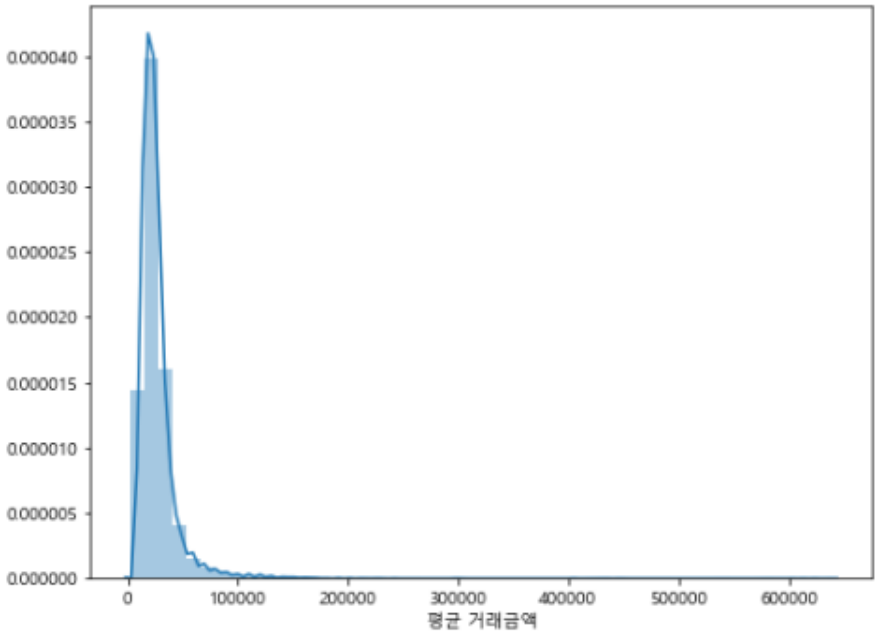
건물번호	구청	보건소
0	5376.198	5718.455
1	5433.525	5775.767
2	5402.503	5744.745
3	5360.153	5702.395
4	5366.848	5709.096

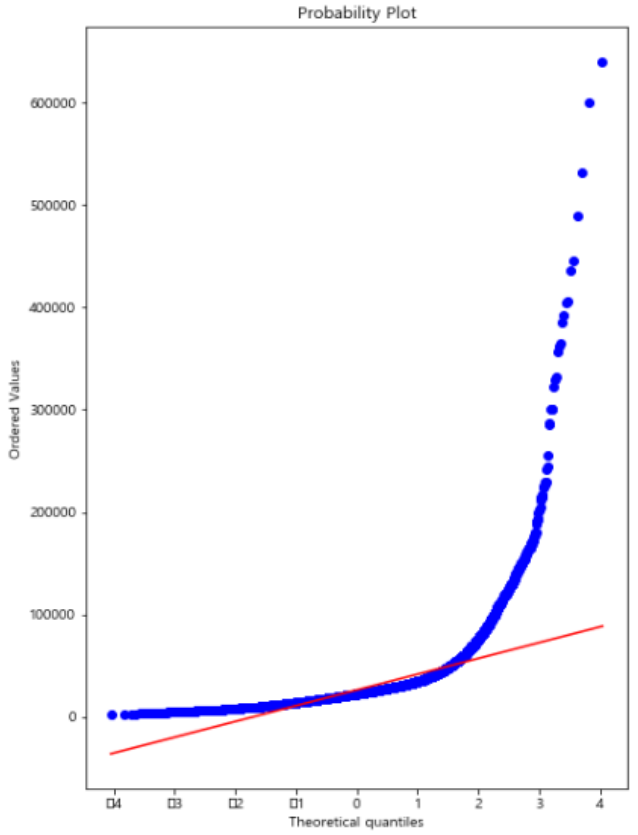
3. Exploratory Data Analysis

3. 1. Target Data: 건물 매매가

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
```

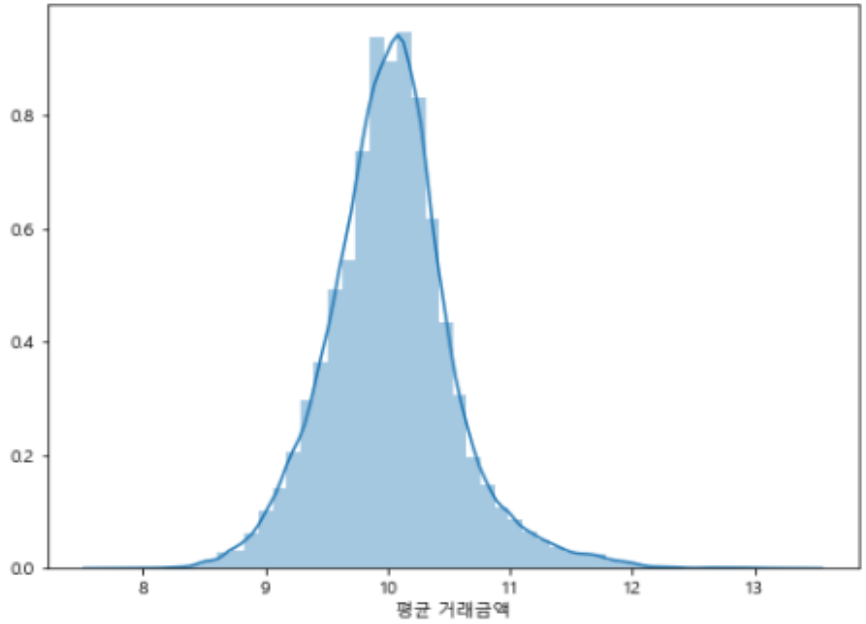


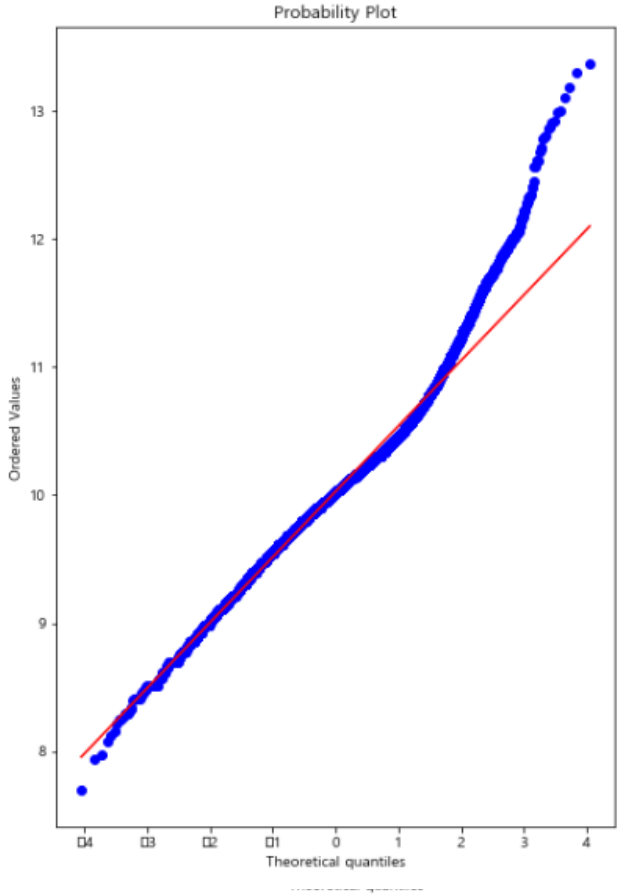


3. 1. Target Data: 건물 매매가

A	B
	평균 거래금액
0	10.95520039
1	10.77897712
2	10.1266711
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
```





3. 2. Missing value 처리

- 최근접 시설 도출 시, 인접한 구에서만 시설을 찾았기 때문에 특정 최근접 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
```

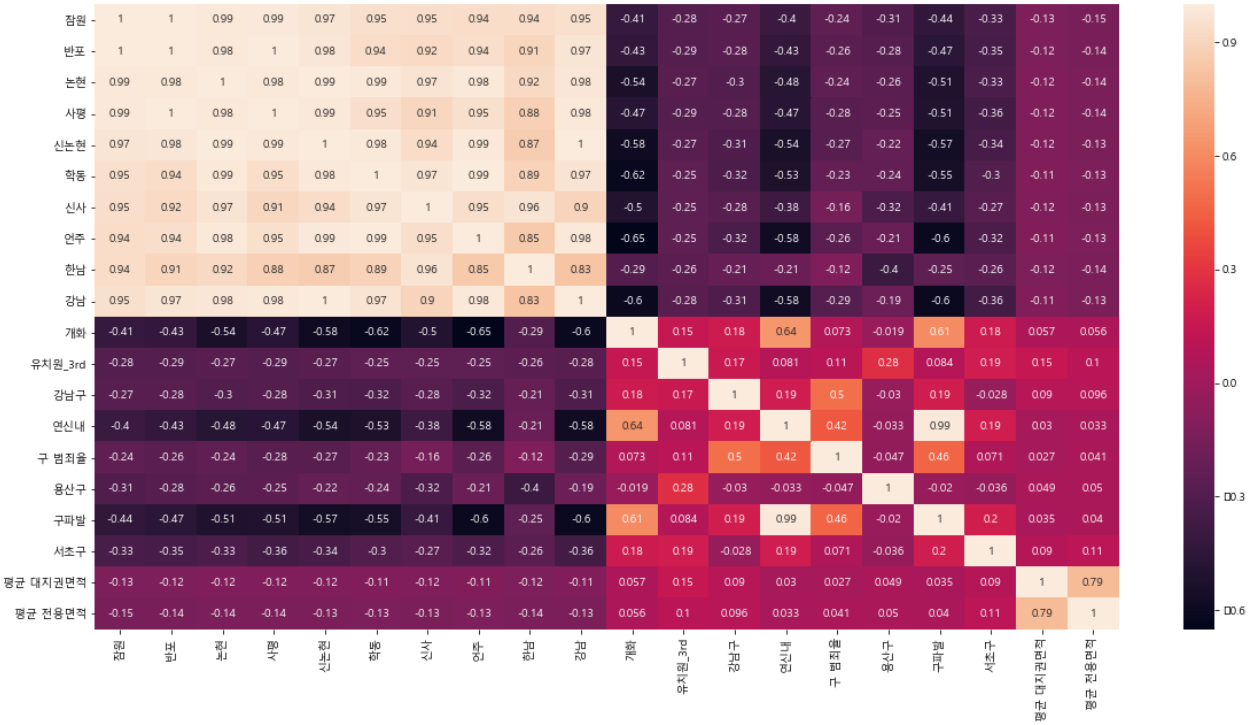
3. 3. 상관관계 분석

- 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

3. 3. 상관관계 분석



- 위 Hitmap은 상관계수가 양으로 높은 10개 feature와 음으로 높은 10개 feature를 대상으로 서로의 상관관계를 표현
- 2와 4사분면은 각 양과 음의 상관계수를 가진 feature들의 상관관계를 표현하며 대체적으로 양의 상관관계를 가짐
- 1과 3사분면을 통해 '평균 거래가격' 에 양과 음의 상관관계를 주는 feature들은 음의 상관관계를 가짐을 확인

4. Pilot Study

4. 1. Machine Learning: Linear Regression

1) 회귀모델 생성에 필요한 변수 분리

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

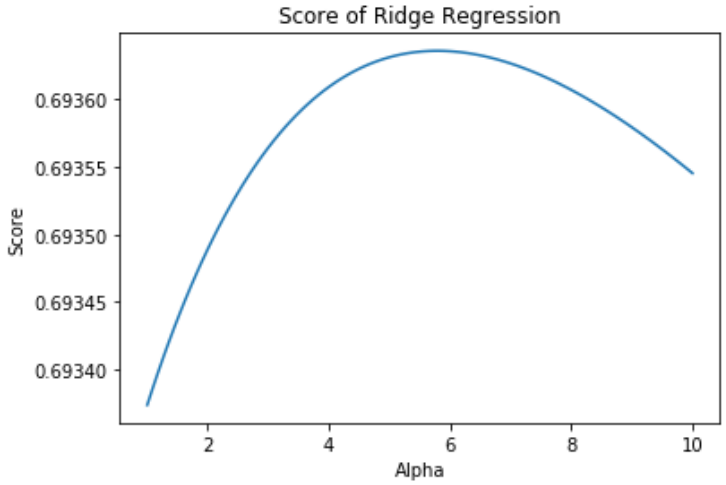
2) Test, Train data를 25:75로 분리

3) 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
```



4. 1. Machine Learning: Linear Regression

경제학 및 사회과학 분야에서는 수많은 Variable과 Feature들 사이에서 정성적으로 지정한 범위 내에서 모델을 생성함

위 분야에서는 비교적 실험자의 Variable 통제의 자유도가 높은 컴퓨터 공학 관련 분야와는 다르게, R-square값이 0.6정도 되어도 유의미한 모델임

- 오윤경, 강정규, 김종민,(2014).지리가중회귀모델을 이용한 주택가격 결정요인의 지역별 특성에 관한 연구 -부산광역시를 중심으로-.,40(),1-17.

구 분	최소값	최대값	평 균	표준편차
Local R^2	0.43	0.86	0.65	.099
절 편	407222.75	3458094.94	1324982.31	782486.18
노령화지수	−13396.37	−1659.88	−6143.63	3259.89
경제인구	1.40	11.25	6.89	2.84
개별공시지가	1.34	8378.80	1635.95	2129.44
사용승인년도	1301.83	3036.64	2651.75	403.19

- 본 프로젝트가 분석하고자 하는 분야인 부동산 역시 경제학 및 사회과학 분야에 속하기 때문에, Ridge 모델을 통해 도출한 R-square값 0.6936는 충분히 유의미함
- 따라서 이후 진행하고자 하는 Deep Learning을 포함한 다양한 회귀 모형을 통한 분석도 충분히 유의할 것으로 예상됨

5. Further Steps

5. 1. 추후 프로젝트 진행 예정 사항

1) 상권 관련 Feature 추가

- 대기업 프랜차이즈 이외의 영세 상권이 주택 매매가에 큰 영향을 줄 것이라 판단하여, 학문/교육, 의료, 음식, 스포츠, 숙박, 소매, 생활서비스, 부동산, 관광/여가/오락 관련 데이터 포함 예정
- 소상공인 진흥회에서 제공하는 34만개의 소상공인 데이터 활용 예정

2) Deep Learning을 포함한 다양한 회귀모형을 통한 예측 실험

- Deep Learning을 통하여 새로 형성되는 상권의 건물 가격 예측
- Regression Model을 통하여 각 Feature의 영향을 다양한 시각에서 분석
- MSE, R-square 등의 지표를 활용한 성능 검증

3) Dimension Reduction 기법을 활용한 주요 Feature 추출

- PCA를 통한 주성분 분석 및 주성분을 이용한 딥러닝 모델 생성
- SVD를 통해 차원축소 및 주요 Feature에 대한 딥러닝 모델 생성

5. 2. 프로젝트 일정

3월		4월				5월					6월
3주차	4주차	1주차	2주차	3주차	4주차	1주차	2주차	3주차	4주차	5주차	1주차
주제 선정											
			데이터 수집 및 전처리								
						데이터 분석					
							기계학습 모델 구축				
										모델 평가	
						Kickoff		Interim			Final

THANK YOU

2019 Spring Data Analytics

BADA